

## ATPA Sample Assessment – Model Solution

*This sample assessment represents the scope of a typical assessment and the types of tasks that may be included. This and the actual assessment will not cover every topic in the syllabus and may emphasize different learning objectives in the context of a real-life business problem and data. Tasks on the assessment may be structured differently than shown in this sample.*

*This model solution is provided so that candidates may better prepare for future sittings of Exam ATPA. It includes both a sample solution, in plain text, preceded by commentary on the task, in italics. In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.*

### General Information for Candidates

This examination has 7 tasks numbered 1 through 7 with a total of 40 points. The points for each task are indicated at the beginning of the task.

Each task pertains to the business problem and related data files and data dictionary. An .Rmd file with some initial data work accompanies this exam. Unless otherwise specified, each task builds upon the work and conclusions from prior tasks. Due to the nature of predictive modeling, work on later tasks may influence responses on earlier tasks.

The responses to each specific task should be written after the task response header in this Word document. Where code, tables, or graphs from your own work is required, it should be copied and pasted into this Word document.

You may use resources such as textbooks and the internet. You may use any analytics software you wish to perform the analysis directed by the tasks. You may not consult with other individuals about the specific business problem, data, and tasks.

Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. The answer should be confined to the prompt as set and written for the audience specified in the prompt. In tasks 1-6, various portions of a technical report are written but these do not comprise an entire report, e.g., a statement of the business problem is not asked for in these tasks. Only write the sections requested.

Your completed Word file should be the only file submitted and will be the only file graded. If any part of your exam was answered in French, include “French” in the file name.

## Business Problem

The following business problem, while using actual data and referring to actual entities, is entirely fictional.

*You have recently started a consulting firm specializing in predictive analytics in the rural western state of Idaho, USA. Your firm consists of you and an assistant you are mentoring. After reading about a dispute between the airport authority in Boise, the state capital and largest city with about 225,000 residents, and airlines servicing the city, you decide to offer your services to the head of the airport authority despite having almost no knowledge of the aviation industry.*

*The dispute centers on a planned increase in airport passenger capacity. Boise, the largest city in a multi-state region, would like to increase its attractiveness for business and pleasure by upgrading its airport, including adding additional passenger gates, but the mountainous location of the airport makes adding runway capacity prohibitively expensive. Two large and dominant freight carriers, FedEx and UPS, are threatening to severely reduce flights to and from Boise, opting to drive freight to and from Salt Lake City, Utah, several hours of driving distance away, because their profit margin for flying to and from Boise is already thin and they believe additional passenger traffic will increase congestion on the runways. The increase in ground time, during which the airplane jet engines are running, will increase their expenses to the point where rerouting will be less costly. If the freight carriers reduce their flights substantially, airport revenue will decrease to a point where adding the passenger gates may not be viable.*

*The head of airport authority accepts your call and is willing to speak with a supportive voice. “That the airlines have made this dispute public has been particularly stressful,” the head explains, “as their real aim is to not to change their usage of the airport but negotiate lower airport usage fees. Retail freight service has been especially valued in this rural area since the COVID-19 pandemic began in March 2020, and so the airlines figured public pressure would help force a steeper compromise. But I do not believe their story about increased ground time from the additional traffic we are proposing makes any sense—ground time is more about weather and delays at other airports than how much traffic we have here. I just downloaded some public FAA data to try to prove this out, but I don’t really have time to do this analysis with press briefings already adding to my day job of running a large airport!”*

*Sensing an opportunity, you explain how predictive analytics, your specialty, would help verify what the underlying factors on ground time really are and ask the head to send over the data so you can consider further. They send the Federal Aviation Administration (FAA) data, and it is in worse shape than you had hoped. The data, which covers all U.S. domestic flights from 2016 to 2021, does not directly have ground time, instead having ramp-to-ramp time and air time. Instead of the data being provided by flight, it is aggregated into monthly totals by route, airline, and other flight plan characteristics. You find the data dictionary for what the head sent you to confirm that the data is what it is and start searching for better or more helpful data, finding a slightly helpful decoder for airport codes along the way. Suddenly, you receive a call from the head of the airport authority.*

*“Actually, I could really use your help immediately. One of the freight airlines called and demanded a meeting a week from today to finalize negotiations of their fees for the next five years. I cannot stall them any longer and could use any leverage I could get from that predictive analytics thing you were talking about. Can you do analysis on the data I sent you and send me a report in four days?”*

*You realize that there wouldn't be time to find better data and would have to work with the current data to meet this deadline. You explain that you have no experience with the aviation industry and would be relying on data that is not fully fit for the purpose of the analysis and ask whether the head would be comfortable with a report given those circumstances. "I can help determine what makes sense once I have the report, but I cannot do the analysis you are talking about. I agree to use the report for the stated purpose recognizing your limitations."*

*Having no other clients, you also agree to this work and direct your assistant to start working on the data while you draft a consulting agreement for the work. After sending this to the head of the airport authority, you follow up with a phone call to confirm receipt and ask a couple questions about the data. The head confirms that ground time is the difference between ramp-to-ramp time and air time but then says, "I appreciate you getting in touch with me today and helping with this analysis. I'll send back the signed agreement today, but I need to ask now that you do the best you can with the information you have. BOI are the call letters for our airport. I won't be available for more information about the data or airline industry until after you finish your report. We can discuss the matter further then."*

*Just as that call ends, your assistant contacts you, sounding less than well. "I'm so sorry...I appear to have caught a nasty illness. I'm sending you the data work I've done so far, but my mind wasn't working very clearly and I thought I better stop. I filtered the data and dealt with the aggregation but didn't get to joining the data. I hope you can manage from here without me for the next few days."*

*You wish the assistant well and then size up the situation. With less than ideal (or clean) data, you have less than ideal time to analyze and report on a problem where you have less than ideal background knowledge and no one you can talk to for help. However, you feel you've represented your situation fairly and expect that those reading your report will try to be sympathetic to your position.*

#### File List

- Six .csv files labeled T\_T100D\_SEGMENT\_US\_CARRIER\_ONLY\_####, where #### is the year: the FAA data<sup>1</sup> provided with flight statistics aggregated by month, flight path, carrier, and other fields.
- One .xls file called DataDictionary: a data dictionary<sup>2</sup> including descriptions of which fields are totaled by month and several tabs that translate various FAA codes
- One .csv file called airport: a separate data file<sup>3</sup> with more information on airports
- Two .Rmd file called FlightsPrep and FlightsPrep\_python: the assistant's work (in R or Python) to prepare the data before becoming ill

---

<sup>1</sup> Retrieved from table T-100 Domestic Segment (U.S. Carriers) ([https://www.transtats.bts.gov/Fields.asp?gnoyr\\_VQ=FIM](https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FIM)), collected and published by the Office of Airline Information, Bureau of Transportation Statistics (BTS), United States Department of Transportation.

<sup>2</sup> Ibid, under Table Info and other tables also housed on BTS.

<sup>3</sup> Retrieved from OurAirports (<https://ourairports.com/data/>), using public domain data collected from FAA for U.S. airports by David Megginson.

## Task 1 (10 points)

Perform the following data preparation tasks:

- Review your assistant's work on the FAA data and modify it to better address the business problem.
  - Retain the definition of ground time as the difference between ramp-to-ramp time and air time—its average will be the target variable for later modeling.
  - Retain some calculation that unitizes the aggregate monthly data to representative data per flight.
  - Develop a new feature to indicate, for each record, how busy the airport was that month.
  - Modify the filtering and other choices made by the assistant as needed
- Carry out additional data cleaning and validation, include a) filtering observations and b) removing, transforming, and adding fields to improve the modeling.
  - Join the following tables to the revised FAA data:
    - The airports data, by IATA code, to add at least airport type, including other fields as desirable.
    - At least one of the tables in the DataDictionary file.
  - Remove some observations to reduce the number of unique carriers to between 8 and 20, resulting in a manageable but still informative number of levels to later investigate differences in ground time among unique carriers. Each remaining carrier should appear at least once in both 2016-2020 and 2021.
- Explore the target variable and no more than two relationships between it and other types of predictors after your preparation work to prepare for modeling steps.
  - Further in-depth data exploration typical of a predictive analytics project is not required.

Then, write the technical data preparation section of your report below. Because none of your code or the transformed data itself will be available to the reader, all evidence of the data preparation tasks will be contained in your report. This may include written descriptions as well as charts, however the work may be most effectively conveyed. Be sure that evidence of the joins is included in your write-up.

## Task 1 Response

*One of the issues is that the assistant only retained departures from Boise and not arrivals. Some records do not have the number of scheduled departures for averaging. The new feature requires aggregating data and putting the result back on each record, essentially a join. There are other missing data issues to contend with, and multiple collinearity issues among fields. The split of airport data among origin and destination is an unusual situation that should be addressed. The reduction to the number of unique carriers has a significant impact on the later tasks but the candidate is left to decide on how many carriers to include, with the freight carriers being important to retain. The data exploration should mostly be focused on modeling but still be looking for overall insights. In general, the candidate gets some guidance but not much, and different candidates may follow markedly different paths. Because code is not submitted, the level of detail in the report writing should allow for near-replication of results.*

## Data Sources

The principal data source, a series of .csv files called T\_T100D\_SEGMENT\_US\_CARRIER\_ONLY\_####, where #### is the year, was retrieved by the head of the Boise airport authority from the online table T-

100 Domestic Segment (U.S. Carriers) ([https://www.transtats.bts.gov/Fields.asp?gnoyr\\_VQ=FIM](https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FIM)), collected and published by the Office of Airline Information, Bureau of Transportation Statistics (BTS), United States Department of Transportation. The six files cover the years 2016-2021 and were downloaded from site linked above. As the head of the airport authority, who is familiar with aviation data, selected the field to download, I am relying on them to have chosen fields appropriate for investigating factors affecting ground time. The data includes all domestic U.S. flights to and from all domestic airports. I have not determined whether international flights, not included in this data, would have a meaningful impact on this analysis but suspect that these are few compared to domestic flights. I also cannot account for the impact of the airport being a joint civil-military airport, as noted here ([https://en.wikipedia.org/wiki/Boise\\_Airport](https://en.wikipedia.org/wiki/Boise_Airport)). In the next section, I describe this data further, including filtering done to greatly reduce the data used to just flights involving the Boise airport, known as BOI. This flight data is accompanied by a data dictionary, a file called DataDictionary.xls, that defines all the fields present in the data and provides several tables decoding encoded fields.

I found a secondary file called airports.csv from OurAirports (<https://ourairports.com/data/>), using public domain data collected from FAA for U.S. airports by David Megginson. As this data ultimately comes from the U.S. government as well, it is assumed to be reliable. No data dictionary is provided. I describe this data further in the Addition of Airports Data section below.

### Raw Data Summary

To conserve memory, only the flight data involving the Boise airport (BOI) is downloaded, specifically where either ORIGIN or DEST (destination) equals BOI. This greatly reduces the rows involved, for example, the 2016 file is reduced from 379,279 records to 1,913 records. Removing records not involving BOI represents a minor loss of relevant information, as ground time is expected to be highly dependent on specific airports.

The following summarizes the data involving BOI from all six years:

DEPARTURES_PERFORMED	PASSENGERS	FREIGHT	MAIL	DISTANCE
Min. : 0	Min. : 0	Min. : 0	Min. : 0.0	Min. : 40.0
1st Qu.: 2	1st Qu.: 51	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 399.0
Median : 12	Median : 706	Median : 0	Median : 0.0	Median : 523.0
Mean : 24	Mean : 1674	Mean : 46915	Mean : 345.7	Mean : 676.5
3rd Qu.: 31	3rd Qu.: 2479	3rd Qu.: 1575	3rd Qu.: 0.0	3rd Qu.: 749.0
Max. : 256	Max. : 16707	Max. : 2059452	Max. : 200687.0	Max. : 2836.0

RAMP_TO_RAMP	AIR_TIME	UNIQUE_CARRIER	ORIGIN	ORIGIN_STATE_NM	DEST
Min. : 0.0	Min. : 0.0	OO : 3026	BOI : 6268	Idaho : 6540	BOI : 6102
1st Qu.: 235.2	1st Qu.: 195.2	WN : 1988	DEN : 644	California:1306	DEN : 643
Median : 1266.0	Median : 994.0	QX : 1235	SLC : 559	Washington: 726	SLC : 617
Mean : 2617.5	Mean : 2124.4	DL : 875	SEA : 475	Colorado : 675	SEA : 494
3rd Qu.: 3672.8	3rd Qu.: 3026.0	FX : 866	PHX : 315	Utah : 583	PHX : 314
Max. : 26817.0	Max. : 21912.0	5X : 861	LAX : 268	Arizona : 393	LAX : 297
		(Other):3519	(Other):3841	(Other) : 2147	(Other):3903

DEST_STATE_NM	AIRCRAFT_TYPE	AIRCRAFT_CONFIG	YEAR	QUARTER	MONTH	DISTANCE_GROUP
Idaho : 6369	673 : 2288	1:10636	2016:1913	1:2999	12 : 1194	1:4739
California:1421	614 : 1255	2: 1734	2017:2058	2:2861	11 : 1173	2:5077
Washington: 744	612 : 1205		2018:2106	3:3125	9 : 1084	3:1788
Colorado : 689	482 : 917		2019:2121	4:3385	8 : 1052	4: 692
Utah : 648	691 : 853		2020:1677		10 : 1018	5: 70
Arizona : 398	698 : 852		2021:2495		1 : 1017	6: 4
(Other) : 2101	(Other):5000				(Other):5832	

CLASS				
F:10229	G: 1727	L: 407	P: 7	

It is evident when inspecting the data that it is aggregate data. The aggregated measures are DEPARTURES\_PERFORMED, PASSENGERS, FREIGHT, MAIL, RAMP\_TO\_RAMP, and AIR\_TIME, where the remaining variables define the group for aggregation, essentially which carrier flew from where to where in which type of aircraft and in which month. The implications of using grouped data for addressing this business problem are discussed further in Task 2.

Immediate observations prior to further translating and transforming the data include:

- There is no missing data;
- DEPARTURES\_PERFORMED of 0 does not seem helpful and does not make sense in some records where AIR\_TIME > 0;
- RAMP\_TO\_RAMP time is never less than AIR\_TIME, so GROUND\_TIME can be the difference as instructed without being negative;
- Many of the grouping fields require translation to validate them;
- Many of the grouping fields have high numbers of categories, potentially leading to overfitting that is hard to control; and
- The splitting of data about the airport other than BOI among multiple fields like ORIGIN and DEST complicates using the other airport as a predictor.

### *Data Transformation and Validation*

#### *Unitize, then Calculate Ground Time*

Ground time, defined as the difference between ramp-to-ramp time and air time, can only be meaningfully compared across groups on a per-flight basis. The aggregated measures are to be divided by DEPARTURES\_PERFORMED to calculate average passengers, freight, mail, ramp-to-ramp time, and air time. To avoid division by zero error values, the 79 records with DEPARTURES\_PERFORMED = 0 (0.64% of the 12,370 raw records) are removed without further inspection.

The unitized measures retain their same field names, and  $GROUND\_TIME = RAMP\_TO\_RAMP - AIR\_TIME$  is calculated, after which RAMP\_TO\_RAMP time is dropped as a variable as it no longer adds information to the over data.

#### *Class and Aircraft Configuration*

The four levels of CLASS correspond to each combination of scheduled/unscheduled and passenger/freight, but the two levels of AIRCRAFT\_CONFIG correspond exactly with the passenger/freight split of CLASS:

	1	2
F 10150	0	
G 0	1727	
L 407	0	
P 0	7	

To address the duplication of information, the levels of CLASS are reduced and renamed SCHEDULED (F/G) and UNSCHEDULED (L/P), with the variable itself renamed SCHEDULED. In addition, the levels of AIRCRAFT\_CONFIG are renamed PASSENGER (1) and FREIGHT (2).

#### *Passenger/Freight/Mail Indicators*

The passenger and freight configurations in AIRCRAFT\_CONFIG are then checked against the average values per flight for PASSENGER, FREIGHT, and MAIL:

AIRCRAFT_CONFIG <fctr>	MIN_PASSENGERS <dbl>	MAX_PASSENGERS <dbl>	MIN_FREIGHT <dbl>	MAX_FREIGHT <dbl>	MIN_MAIL <dbl>	MAX_MAIL <dbl>
PASSENGER	0	226	0	74962	0	2610
FREIGHT	0	0	0	87309	0	52959

Freight configurations never have passengers (crew are not counted as passengers), but 548 records (representing 631 flights) with passenger configurations have no passengers, sometimes having freight and mail and sometimes not. There are times when an airline needs to reposition a plane due to weather or maintenance issues. 423 of the records are with SkyWest Airlines (OO), indicating that this may be a common practice for them. These records are retained. Similarly, 2 records with freight configuration but no freight or mail (or passengers) are retained.

As the presence or absence of passengers, freight, or mail may be significant for predicting ground time, indicator variables for each (HAS\_PASSENGERS, HAS\_FREIGHT, and HAS\_MAIL) are created with YES if more than 0 and NO if 0.

#### Additional Validation and Removal

DISTANCE\_GROUP has levels based on 500-mile bins, and these are validated against DISTANCE, confirming that it was not an aggregate measure in the raw data. However, more useful binning can be derived from DISTANCE, so DISTANCE\_GROUP is removed.

QUARTER is a less granular variable than MONTH, and these validate perfectly against each other. As seasons are likely to be more useful groups than financial quarters, QUARTER is removed.

#### Departures by Month

To indicate how busy the Boise airport is, the total number of departures per month was calculated, summing DEPARTURES\_PERFORMED by groups defined by YEAR and MONTH. This result was then joined back into each record based on YEAR and MONTH, creating a field called TOT\_DEPARTURES (including both departures and arrivals as defined later). The values joined are as follows:

BOI	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<b>2016</b>	3876	3521	3855	3656	3857	4064	4150	4146	3948	4129	3905	4132
<b>2017</b>	3876	3554	4106	3918	4043	4060	4245	4169	4006	4110	4092	4300
<b>2018</b>	4193	3840	4414	4303	4493	4529	4776	4710	4115	4363	4258	4491
<b>2019</b>	4339	3844	4361	4185	4421	4518	4715	4907	4543	4795	4599	4995
<b>2020</b>	4619	4271	4248	1648	1609	2194	2936	3488	3419	3904	3942	4139
<b>2021</b>	4072	3456	4122	3936	4200	4798	5530	5427	4998	4718	4454	4286

The impact of the COVID-19 pandemic is clear beginning in April 2020, with flights per month reaching pre-pandemic levels towards the end of 2020. Except for the disruption caused by the pandemic, the number of flights at BOI has slowly been increasing overall.

The TOT\_DEPARTURES variable represents a potential timing issue, where the remaining count of future flights in a month is not available to predict ground time for a current flight. However, this variable is being used just as a proxy for how busy the airport is, and, if found significant, a more appropriate predictor could be sought.

#### Code Translation

As the two-letter codes of UNIQUE\_CARRIER are unfamiliar, these are replaced by the text in the data dictionary, using a join directly on the sheet called L\_UNIQUE\_CARRIERS. All records matched.



Similarly, three-digit codes of AIRCRAFT\_TYPE are replaced via joining to L\_AIRCRAFT\_TYPE in the data dictionary. All records matched.

#### Refactoring Airport Information

GROUND\_TIME represents time spent on the ground at both BOI and another airport, but that other airport appears in DEST or ORIGIN based on whether the flight is a departure or arrival respectively. It would be more helpful to have information on the other airport in one field rather than split in two. To accomplish this, a new variable called DIRECTION is given two levels, DEPARTURE if ORIGIN = BOI and ARRIVAL otherwise. There are no flights from BOI to BOI.

Then, DEST and ORIGIN are replaced by a new variable AIRPORT that has the airport code that is not BOI based on DIRECTION. Similarly, AIRPORT\_STATE\_NM replaces DEST\_STATE\_NM and ORIGIN\_STATE\_NM.

As three-letter airport codes are less familiar, descriptions of the airports are desirable to supplement these codes. This data is available in both L\_AIRPORT of the data dictionary and the name field in airports.csv. The latter has difficult to read entries such as Minneapolisâ€”Saint Paul International Airport / Woldâ€”Chamberlain Field for MSP and does not match perfectly, so the join with the data dictionary table is performed to add a new field, AIRPORT\_DESC, retaining AIRPORT with the three-letter code.

#### Addition of Airports Data

The data in airports.csv is considered for possible predictors of ground time. The variable “type” reduces the many airports into relatively few categories, and it is sensible that larger airports may have different ground times than smaller ones. This is joined in and called AIRPORT\_TYPE. The variable “elevation” may also be helpful because inclement weather may be more common at high or low elevations, but this may also vary by region and geography, so it is not expected to be helpful without more information to clarify these other factors.

Joining to airports.csv is not trivial as there are several potential fields to use for joining. The field iata\_code is best aligned with AIRPORT in the main data, but local code also has many matches once filtered to the US. Using the data dictionary to compare airport names, examples such as Scottsdale indicate that iata\_code is more appropriate, but this join leaves two codes, F70 and CHD, unmatched. These both belong to records involving Scott Aviation, a small carrier, whose records are filtered out as described in the next section.

The field AIRPORT\_TYPE has four levels, three for large, medium, and small airports and one for closed airports. There is just 1 record with a closed airport (0.01% of data), where Williston, ND replaced its airport with a larger field soon afterwards, so this record with a single flight is removed without material harm, bringing the record count to 12,290.

#### Reduction in Number of Carriers

There are 40 levels in UNIQUE\_CARRIER, 198 levels in AIRPORT, and 55 levels in AIRCRAFT\_TYPE, rendering many of these nearly useless as predictors due to the curse of dimensionality. To address this, the sum of NUM\_DEPARTURES is calculated for each level in UNIQUE\_CARRIER. There are natural breaks around 1000 and 50 total flights over the six-year period. Retaining only carriers above 1000 total flights would only reduce the 296,838 total flights by 1,261 (0.42%), a small loss of data.



This filtering is carried out, resulting in 553 (4.50%) fewer records and resulting in a final record count of 11,737. All factor variables are releveled to eliminate null levels, resulting in 14 unique carriers, 149 airports, and 34 aircraft types. Airports have other variables describing them with fewer levels. Each of the 14 carriers has activity in at least five of the six years of data, a good level of consistency. Further reduction may be done in the modeling stage to improve predictive performance as needed.

### Final Data Summary

After data transformation and validation as described above, the final data summary is

DEPARTURES_PERFORMED	PASSENGERS	FREIGHT	MAIL	DISTANCE
Min. : 1.00	Min. : 0.0	Min. : 0.00	Min. : 0.00	Min. : 40.0
1st Qu.: 3.00	1st Qu.: 33.0	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 399.0
Median : 15.00	Median : 64.5	Median : 1.72	Median : 0.00	Median : 522.0
Mean : 25.18	Mean : 69.6	Mean : 4664.15	Mean : 62.91	Mean : 670.4
3rd Qu.: 31.00	3rd Qu.:116.0	3rd Qu.: 81.07	3rd Qu.: 0.00	3rd Qu.: 735.0
Max. :256.00	Max. :195.0	Max. :87309.00	Max. :52959.00	Max. :2836.0

AIR_TIME	UNIQUE_CARRIER	AIRCRAFT_TYPE
Min. : 14.00	SkyWest Airlines Inc. :3019	Embraer ERJ-175 :2253
1st Qu.: 62.00	Southwest Airlines Co. :1982	Boeing 737-700/700LR/Max 7 :1196
Median : 84.65	Horizon Air :1235	Boeing 737-800 :1058
Mean : 97.85	Delta Air Lines Inc. : 873	De Havilland DHC8-400 Dash-8 : 917
3rd Qu.:109.50	Federal Express Corporation: 866	Airbus Industrie A300-600/R/CF/RCF: 853
Max. :372.00	United Parcel Service : 861	Airbus Industrie A319 : 846
	(Other) :2901	(Other) :4614

AIRCRAFT_CONFIG	YEAR	MONTH	GROUND_TIME	SCHEDULED	HAS_PASSENGERS
PASSENGER:10010	2016:1830	12 :1117	Min. : 1.00	NOT SCHEDULED: 55	NO :2182
FREIGHT : 1727	2017:1956	11 :1085	1st Qu.: 15.64	SCHEDULED :11682	YES:9555
	2018:1947	8 :1005	Median : 19.00		
	2019:2005	9 :1000	Mean : 20.48		
	2020:1618	1 : 995	3rd Qu.: 23.83		
	2021:2381	3 : 955	Max. :313.00		
	(Other):5580				

HAS_FREIGHT	HAS_MAIL	DIRECTION	TOT_DEPARTURES	AIRPORT	AIRPORT_STATE_NM
NO :5602	NO :11095	ARRIVAL :5779	Min. :1609	SLC :1169	California:2662
YES:6135	YES: 642	DEPARTURE:5958	1st Qu.:3942	DEN :1156	Washington:1452
			Median :4185	SEA : 965	Colorado :1201
			Mean :4201	PHX : 620	Utah :1177
			3rd Qu.:4493	LAX : 559	Arizona : 767
			Max. :5530	LAS : 499	Nevada : 697
				(Other):6769	(Other) :3781

AIRPORT_DESC	LARGE_AIRPORT
Salt Lake City, UT: Salt Lake City International:1169	NO :2091
Denver, CO: Denver International :1156	YES:9646
Seattle, WA: Seattle/Tacoma International : 965	
Phoenix, AZ: Phoenix Sky Harbor International : 620	
Los Angeles, CA: Los Angeles International : 559	
Las Vegas, NV: McCarran International : 499	
(Other) :6769	

Several observations follow, keeping in mind that records do not indicate the number of flights, which will be used as a weight when possible for modeling:

- The two freight carriers at the heart of the dispute, FedEx and UPS, have a significant number of records and may be isolated as a predictor.
- Some levels have again few records, such as unscheduled flights, and these will need to be monitored.

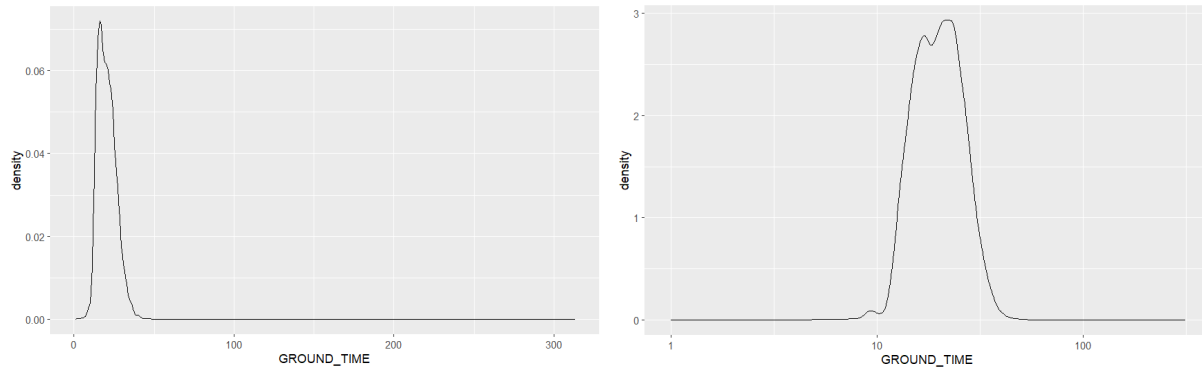
- The average ground time, the target variable, is typically 20 minutes (split between two airports, BOI and AIRPORT) with 50% of records in a smallish range from 16 to 24 minutes.

### Data Exploration

Brief data exploration of GROUND\_TIME and predictors against this variable may reveal additional trends to consider when modeling.

### Ground Time, Univariate

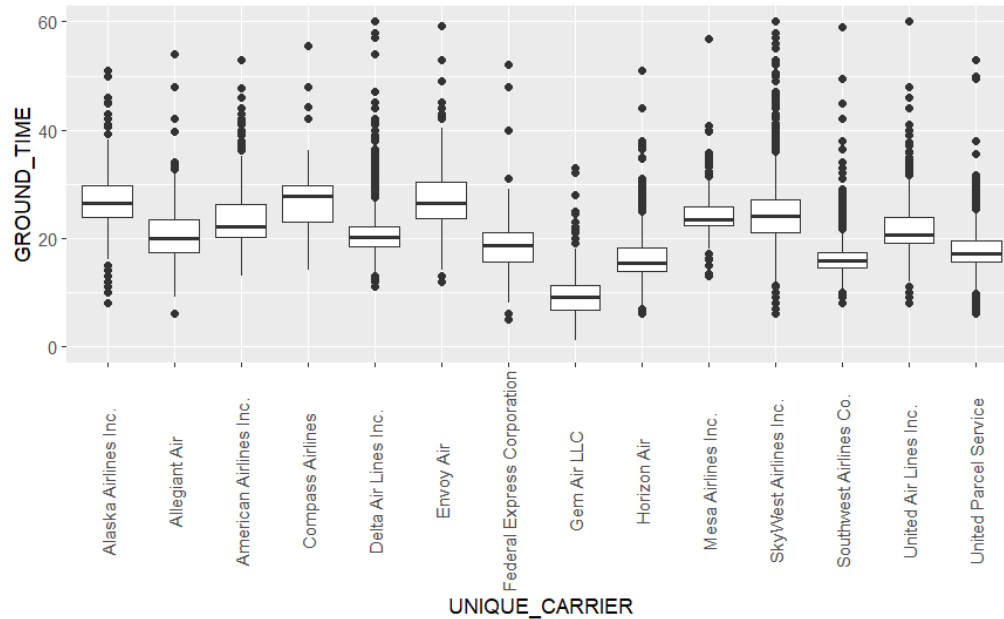
The univariate density graphs of GROUND\_TIME on a linear and log scale are as follows, using NUM\_DEPARTURES as a weight so that extreme values on rare flights are not given too much weight:



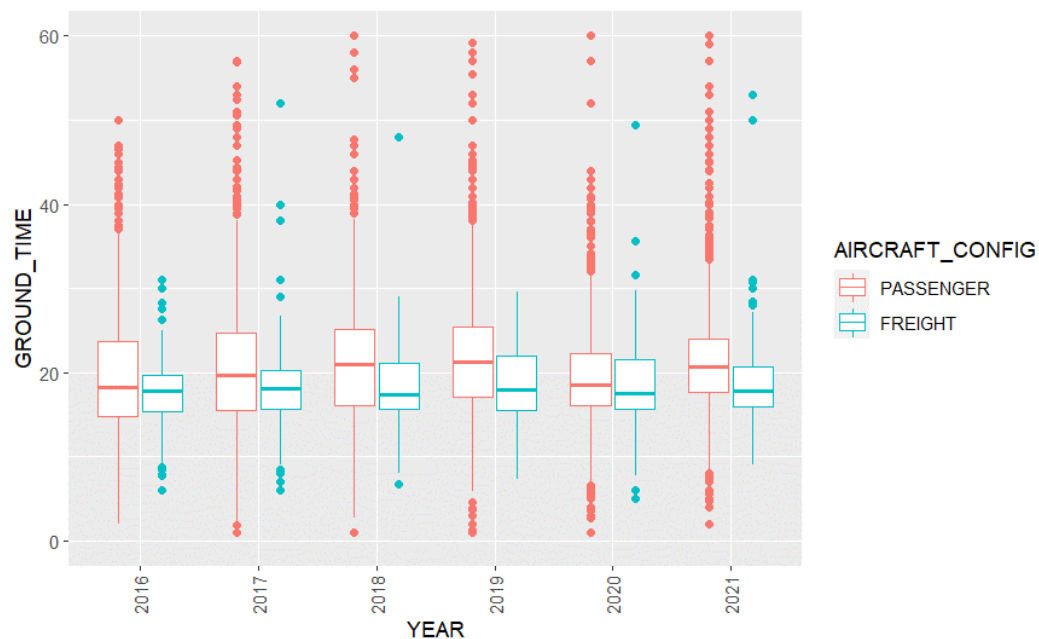
The GROUND\_TIME variable has a moderate right skew due to extraordinarily long ground times not being balanced by negative ground times, which would not make sense. The log of GROUND\_TIME appears to have little skew but has thin tails due to GROUND\_TIME being an average value with varying weights. That is, records with many flights will have extreme ground times on some flights averaged with typical ground times, and those more extreme times will not be preserved as outliers. Only the records with few flights, typically 1, will have extreme averages. For example, of the 61 records with GROUND\_TIME over an hour, all but 5 are aggregations of unique flights for the flight plan and month, and the remaining 5 each include two flights. The extremes in this right tail do not appear to be erroneous; only five flights have ground time over two hours, and the maximum value of 313 minutes, or just over five hours, is conceivable.

### Ground Time, Bivariate

In each of the following graphs, with subsequent commentary, weights based on NUM\_DEPARTURES continue to be used. Ground times over an hour are not shown to enhance visibility.



There is significant variation in ground time by carrier. The two large freight carriers, FedEx and UPS, have relatively low ground time, though not as low as Gem Air, which offers scheduled summer passenger service only between Boise and rural Salmon, Idaho, in this data.



The bivariate graph of interest is aircraft configuration, where freight has a slightly lower ground time than passenger overall. Importantly, the ground time for freight appears to have remained steady over the time period while ground time for passenger flights increased as the airport got busier. While not overly significant, illustrating this in some form may be helpful for the head of the Boise airport authority to dispel the assumption by the freight carriers that increased flights from more passengers will cause increased ground time for freight flights as well.

## Task 2 (3 points)

Write a separate section of your technical report discussing ethical concerns on the use of this data for this business problem, considering selection, measurement, and omitted variable biases.

### Task 2 Response

*Perhaps the biggest issue to recognize is that ground time incorporates what happens both at Boise and at other airports. Because the dispute is about ground time at Boise, the data provided is only partially relevant to the analysis. Another issue is the aggregate data, which can mask variability and make rarer segments seem better or worse by random chance and also may not be de-aggregated correctly. The lack of knowledge regarding aviation statistics is another potential issue. While not described in the following sample responses, COVID-19 pandemic may have affected the data during 2020 and 2021 and may drive data bias issue.*

### Selection Bias

At first glance, the wide range of carriers, flights, and years, using mandatory reporting to a government agency, suggests that little selection bias is present, but some aspects of the data are concerning on how well they represent what is to be predicted, including the use of average ground time over the month for the same flight plan and other characteristics.

Large airlines with many identical flights each month for a particular flight plan will have natural extremes in ground time averaged out, but small airlines with perhaps just one flight in a month for a particular flight plan is more prone to extreme ground times due to random events. If a threshold is set above median ground time as a maximum desirable ground time, the averages for more frequent flights will have a better chance of meeting this threshold than averages for less frequent flights even if the carriers had identical distributions of ground time per flight.

The ideal remedy would be to have data for individual flights, but that is not available in the assigned timeframe for this project. Two less effective remedies are put in place. First, the target variable will be average ground time itself and not a threshold above this. Second, only carriers with more than 1000 flights over the six-year period are included, indirectly reducing the number of infrequently used flight plans. The second remedy comes at the cost of underrepresenting smaller airlines, but for the head of the Boise airport authority, the most impactful decisions concern the larger airlines.

A separate selection bias issue is that flights with such excessive ground time that they never took off are not represented in this data, which only captures completed flights. The inability for a flight to take off due to issues on the ground is far costlier and more disruptive to airlines than short delays on the ground, but the data does not allow consideration of this.

### Measurement Bias

A significant measurement bias issue is that ground time as reported in the data is the sum of ground time at two airports, BOI and another airport. The dispute regarding the increase in airport passenger capacity is directly related only to ground time at BOI. The inclusion of ground time at the other airport, often a larger airport than BOI, makes the target variable only partially reliable for addressing the business problem even before the first measurement bias is considered. That predictive analytics is being applied may help separate out the impact of the other airport, but having only ground time in BOI would be far better.

### *Omitted Variable Bias*

I do not have adequate knowledge about aviation to ascertain whether the data provided include the most significant drivers of ground time, and I do not have direct access to individuals who can provide that knowledge prior to my submitting this report. While the recipient has agreed to discuss additional information about aviation after the report has been submitted, there will be a temptation to use the report as is due to convenience. Having the distance from the hangar or terminal to the end of the runway for a given flight would allow a fixed portion of ground time (assuming standard taxiing speed) to be removed as an offset, leaving the predictors to consider the more variable portion of ground time due to congestion on the runway, weather, and other factors. But someone with more aviation experience could confirm or improve upon this theory and provide more applicable and precise data and relationships to address the business problem.

### Task 3 (6 points)

Using representative ground time as the target variable, fit the following models to perform well on unseen data using unique carrier and other variables you select as predictors:

- GLM, where unique carrier and other variables have a fixed effect
- GLMM, where at least unique carrier and one other predictor have random effects and other variables have fixed or random effects as appropriate

The GLMM should not remove any predictors from the GLM but may transform them and add new predictors. Each model fitting should try to isolate the effect of unique carrier from related predictors in the data as much as possible. The variable selection for this and future models should use 2016-2020 data as training and 2021 data as validation.

Write a technical report section discussing the impact, all else equal, of unique carrier based on these two fitted and validated models. Be sure to give particular attention to the two freight airlines discussed in the business problem. Include detailed results from your model fitting and variable selection to provide sufficient support for your findings.

### Task 3 Response

*Given that it is emphasized in the task, unique carrier should be expected to be interesting and should receive additional attention. The decision to use weights may not occur to all candidates and an unweighted approach produces similar though not identical results. Candidates are left to choose what kind of distribution and link function to use as well as what test metric to use for out-of-sample validation. Many choices are defensible—it is important to document the choice and reasoning. Differing data and model design decisions may lead to a substantially different GLM. The GLMM modeling should take advantage of its ability to handle factor variables where new levels may appear.*

#### General Modeling Decisions

The 11,737 observations are split into train (9,356 observations) and test (2,381 observations) based on YEAR, with 2021 becoming test data and the rest train data. Using year to partition the data allows the modeling process to verify that it will reasonably predict unseen data in the future.

The test metric used throughout the modeling is root mean square error (RMSE) as calculated on the unseen test data. This is chosen for its simplicity compared to other metrics like loglikelihood and is supported by the nearly normal shape of the target variable.

Where possible, NUM\_DEPARTURES, the number of flights represented by an observation, will be used as weight when fitting models and applying the test metric. When a modeling technique does not easily accept weights, an unweighted test metric will be used to compare models, including to ones that accept weights.

#### GLM Fitting and Variable Selection

The shape of the target variable appears normal or lognormal, so these distributions were tested. Based on comparative test results after adding UNIQUE\_CARRIER and DISTANCE to a null GLM to predict GROUND\_TIME, a normal distribution (Gaussian with identity link function) is favored over a lognormal distribution (Gaussian with log link). The normal distribution has an RMSE of 4.16 while the lognormal had an RMSE of 4.18. This choice is likely influenced by the selection of RMSE as the test metric, as GLM

fitting for normal, but not lognormal, optimizes RMSE. That aggregate data is being used also suggests a normal shape due to the central limit theorem. No other distributions or link functions were considered.

The predictors of the final GLM model are presented below, in decreasing order of impact on model fit, with the RMSE that occurs when dropping only that predictor, adding it back when dropping the next predictor. Additional commentary on the GLM fitting and transformed predictors follows. Weighted models are used for GLM and GLMM.

Only Predictor Dropped	RMSE (Weighted)	Difference from Full Model
All are dropped (null model)	5.0729	1.4955
UNIQUE_CARRIER	3.8688	0.2914
AIRPORT_16	3.8026	0.2252
DIRECTION	3.6884	0.1110
WINTER	3.6882	0.1108
PASSENGERS	3.6537	0.0763
AIRCRAFT_TYPE_3	3.6443	0.0669
DISTANCE	3.6042	0.0268
HAS_PASSENGERS	3.5861	0.0087
None are dropped (full model)	3.5774	0.0000

For each test of dropping one variable, the larger the difference from the full model, the more impactful that predictor is in total. For example, UNIQUE\_CARRIER in total is the most impactful predictor.

Commentary on each predictor:

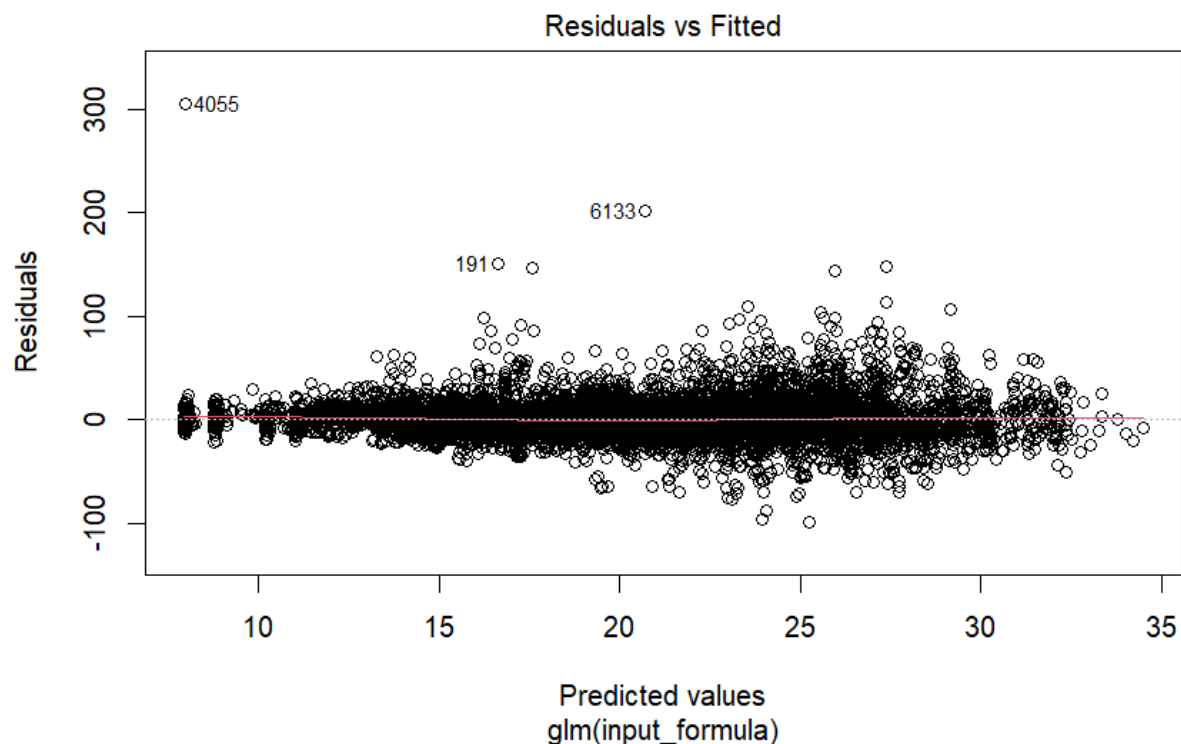
- **UNIQUE\_CARRIER:** 13 impacts (all compared to Alaska Airlines Inc.) varying from -9.9 minutes for Southwest Airlines Inc. to 4.2 minutes for Compass Airlines. Additional discussion on unique carrier is below.
- **AIRPORT\_16:** includes top 16 airports by number of flights and “other” for all 133 other airports in the data. Using other airports as the baseline, impacts vary from -1.0 minutes for MSP (Minneapolis-St. Paul) to 6.2 minutes to ORD (O’Hare near Chicago), with only two airports having negative impacts compared to other, typically smaller, airports.
- **DIRECTION:** predicts that departures from BOI add 0.8 minutes of ground time compared to arrivals, perhaps because clearance is needed from the other airport to take off and BOI is quicker to give that clearance.
- **WINTER:** consists of December through February, where ground time is 2.2 minutes longer than in other months. This reduction of MONTH performs better than November through February, the next best.
- **PASSENGERS:** along with **HAS\_PASSENGERS**, predicts the ground time is longer than no passengers when more than 97 passengers and shorter otherwise, with an impact of 0.4 minutes per 10 passengers.
- **AIRCRAFT\_TYPE\_3:** includes top 3 aircraft types by number of flights and “other” for all 31 other types in the data. Most airlines only use 0 or 1 of the most common aircraft types, where De Havilland DHC8-400 Dash-8 had the greatest impact at -3.1 minutes, but enough use them to justify the distinction.



- **DISTANCE:** increase of 0.4 minutes of ground time for every 100 miles, even after factoring in impact of AIRPORT\_16 and UNIQUE\_CARRIER.

All other reasonable predictors were tried and found to have minimal or no improvement to model fit. For reducing AIRPORT, separating out just the top 1, 4, and 8 airports were also tried based on natural breakpoints, but using 16 was a reduction of 0.12 in RMSE compared to using 4, the next best alternative. For reducing AIRCRAFT\_TYPE, separating out the top 8, 12, and 16 aircraft types were also tried based on natural breakpoints, but only using 3 provided a substantial reduction in RMSE.

The residuals plot (unweighted) indicates that the GLM has similarly distributed errors across the range of predictions from 8 to 34 minutes:



### *GLMM Fitting and Random Effects*

The final GLM is used as a starting point for a GLMM, using the same parameters at the GLM. Predictions made using the mixed effects models included the random effects as the goal is to validate the inference regarding the effects rather than make a prediction where those effects are absent. Only random intercepts were considered, as the GLM has few continuous predictors and random slopes would be distracting and complicated to explain for those predictors.

Considering which predictors to convert to random intercepts, the greatest benefit comes from being able to incorporate new factor levels that have not been previously observed. It is conceivable with this data that new unique carriers, airports, and aircraft types will be encountered in the future, but the GLM will not be able to handle these. Due to the lack of other data on impacts that may be changing with time, including the pandemic, year is also a potentially helpful random effect to allow for unusual years in the data that are not expected to repeat, as seen in the final graph in Task 1.

The following random effects were tried:

Random Intercepts(s)	RMSE (Weighted)
None	3.5774
UNIQUE_CARRIER	3.5776
UNIQUE_CARRIER, AIRPORT_16	3.5762
UNIQUE_CARRIER, AIRPORT	<b>3.5559 (lowest)</b>
UNIQUE_CARRIER, AIRPORT, AIRCRAFT_TYPE_3	3.5587
UNIQUE_CARRIER, AIRPORT, AIRCRAFT_TYPE	3.5788
UNIQUE_CARRIER, AIRPORT, YEAR	3.5780

The final GLMM uses UNIQUE\_CARRIER (14 levels) and AIRPORT (149 levels among all data, a little less in the train data) as random intercepts. While there is little difference from converting UNIQUE\_CARRIER to a random effect, the ability of the mixed effects model to include all levels of airport as random effects, and not just the most frequent airports, provides a slight improvement to the prediction despite the large number of effects carried over to the prediction. The unweighted standard deviation of these effects is 3.3 minutes.

#### *Impact of Unique Carrier*

For the GLM, all impacts of unique carrier on ground time are relative to Alaska Airlines Inc., while, in the GLMM, all impacts are centered, relative to an unknown carrier. To make these comparable, it is assumed that Alaska Airlines, Inc. has the same impact in the GLM as it does in the GLMM and all other GLM impacts are shifted, resulting in the following:

Unique Carrier	Number of Flights	GLMM Random Intercept	GLM Relative Coefficient
SkyWest Airlines Inc.	109,624	4.2	4.1
Horizon Air	67,566	-0.7	-0.8
Southwest Airlines Co.	45,558	-7.0	-6.9
Delta Air Lines Inc.	14,391	-0.9	-1.1
American Airlines Inc.	11,337	0.0	0.2
Federal Express Corporation	<b>9,953</b>	<b>-3.4</b>	<b>-3.5</b>
United Air Lines Inc.	8,791	-4.0	-4.0
United Parcel Service	<b>6,627</b>	<b>-0.3</b>	<b>-4.3</b>
<i>Alaska Airlines Inc.</i>	<i>6,278</i>	<i>3.0</i>	<i>3.0</i>
Mesa Airlines Inc.	4,328	4.3	4.7
Envoy Air	4,089	3.8	3.9
Allegiant Air	3,633	-3.4	-3.6
Compass Airlines	2,023	7.2	7.2
Gem Air LLC	1,379	-2.8	-5.1

In most cases, the impact of unique carrier on ground time is similar between GLMM and GLM, but two large differences stand out for United Parcel Service (UPS) and Gem Air. For UPS in the training data, 445

of the 641 records are to airports that are not in the top 16 airports. Most common among these are 182 records for SDF, the airport in Louisville, Kentucky, which the GLMM model gave a random intercept of -7.1 minutes.

To distinguish this as being an effect of airport and not one of carrier, the model compares UPS ground times to SDF (mean 16.7 minutes) to UPS ground times not to SDF (mean 16.9 minutes, varies by airport) and non-UPS ground times to SDF (single record of 11 minutes). To test whether the single Southwest Airlines flight to SDF may be skewing the carrier impact for UPS, the GLMM is fit without this record, but there is no material difference in results. So, instead, the comparative impact to UPS in the GLMM must be based on the total effect of other shared airports like BIL where other airlines also fly more often.

All else equal, FedEx has 3.4 minutes less ground time than a typical carrier and UPS 0.3 minutes ground time—both are on the better end of airline performance, a result worth pointing out. Of course, all else is not equal, and UPS appears to benefit from its selection of airports outside the top 16 from BOI.

#### Task 4 (5 points)

Using the same model form as the GLM in the previous task, fit a Bayesian model on the same training data (2016-2010), adjusting the parameters of the fitting function as needed to manage runtime while still achieving convergence in the fit.

Write a technical report section on the uncertainty of predicted ground time, relating it to uncertainty around the impacts of individual predictors and model parameters, backing your discussion with evidence from the fitted Bayesian model, with particular focus on unique carrier.

#### Task 4 Response

*The main point is to be able to discuss what the Bayesian model shows about these predictors. In this case, these correspond quite well to the linear model errors, but the Bayesian framework makes it easier to compare the sources of uncertainty.*

#### Application of Bayesian Model

A Bayesian model is fit to the same predictors as presented in Task 3 for the GLM using the `brm()` function in R with its default settings. Due to runtime considerations, an unweighted Bayesian model is fit and its results compared to an unweighted version of the GLM. The Bayesian model has four chains with 1000 warm-up iterations and 1000 sampling iterations. Standard flat improper priors are used and the alternative of a horseshoe prior not considered after seeing the similarity of parameter distributions to those assumed in the GLM fitting.

#### Uncertainty of Predictions and Predictors

The summary of the fitted Bayesian model is as follows:

```
Family: gaussian
Links: mu = identity; sigma = identity
Formula: GROUND_TIME ~ UNIQUE_CARRIER + DISTANCE + DIRECTION + HAS_PASSENGERS + PASSENGERS + WINTER +
AIRPORT_16 + AIRCRAFT_TYPE_3
Data: df.train (Number of observations: 9356)
Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
       total post-warmup draws = 4000
```

#### Population-Level Effects:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	19.72	0.76	18.28	21.20	1.00	950	1879
UNIQUE_CARRIERAllegiantAir	-6.83	0.65	-8.08	-5.52	1.00	1117	2204
UNIQUE_CARRIERAmericanAirlinesInc.	0.12	0.76	-1.42	1.57	1.00	1219	2486
UNIQUE_CARRIERCompassAirlines	2.44	0.96	0.60	4.33	1.00	1549	2652
UNIQUE_CARRIERDeltaAirLinesInc.	-3.38	0.62	-4.62	-2.19	1.00	1085	2093
UNIQUE_CARRIEREnvoyAir	2.39	1.07	0.25	4.54	1.00	1438	2568
UNIQUE_CARRIERFederalExpressCorporation	-6.61	0.75	-8.13	-5.12	1.00	1026	1904
UNIQUE_CARRIERGemAirLLC	-8.97	0.86	-10.69	-7.25	1.00	945	2104
UNIQUE_CARRIERHorizonAir	-2.80	0.86	-4.50	-1.16	1.00	1081	1978
UNIQUE_CARRIERMesaAirlinesInc.	1.66	0.94	-0.21	3.47	1.00	1465	2336
UNIQUE_CARRIERSkyWestAirlinesInc.	1.92	0.59	0.73	3.10	1.00	724	1543
UNIQUE_CARRIERSouthwestAirlinesCo.	-9.09	0.59	-10.22	-7.92	1.00	1084	1999
UNIQUE_CARRIERUnitedAirLinesInc.	-3.42	0.66	-4.71	-2.08	1.00	1198	2157
UNIQUE_CARRIERUnitedParcelService	-7.68	0.77	-9.22	-6.15	1.00	1036	2102
DISTANCE	0.00	0.00	0.00	0.00	1.00	2594	2900
DIRECTIONDEPARTURE	0.79	0.16	0.47	1.11	1.00	5875	2903
HAS_PASSENGERSYES	-2.19	0.50	-3.15	-1.21	1.00	3251	3071
PASSENGERS	0.03	0.00	0.02	0.04	1.00	3282	2662
WINTERYES	2.17	0.18	1.84	2.52	1.00	6820	2855
AIRPORT_16SEA	3.08	0.45	2.23	3.97	1.00	1374	2364
AIRPORT_16SLC	1.92	0.42	1.12	2.75	1.00	1720	2479
AIRPORT_16PDX	-0.18	0.50	-1.18	0.82	1.00	2247	2534
AIRPORT_16DEN	1.51	0.41	0.71	2.32	1.00	1797	2735

AIRPORT_16SFO	2.66	0.51	1.64	3.65	1.00	2354	2710
AIRPORT_16GEG	1.50	0.55	0.44	2.60	1.00	2136	2760
AIRPORT_16PHX	-1.58	0.51	-2.55	-0.61	1.00	2565	3162
AIRPORT_16LAX	3.58	0.46	2.69	4.47	1.00	2355	2309
AIRPORT_16ORD	4.18	0.60	3.04	5.35	1.00	3816	2757
AIRPORT_16OAK	1.54	0.54	0.45	2.58	1.00	2078	2921
AIRPORT_16LAS	2.31	0.52	1.29	3.36	1.00	2343	2858
AIRPORT_16MSP	-2.25	0.52	-3.25	-1.23	1.00	3554	3000
AIRPORT_16SMF	0.21	0.55	-0.88	1.31	1.00	2298	2540
AIRPORT_16SAN	-1.37	0.56	-2.49	-0.32	1.00	3137	3042
AIRPORT_16SJC	1.17	0.60	-0.01	2.32	1.00	2108	2559
AIRPORT_16DFW	0.29	0.80	-1.28	1.88	1.00	4020	3216
AIRCRAFT_TYPE_3EmbraerERJM175	-2.22	0.30	-2.81	-1.63	1.00	3253	3047
AIRCRAFT_TYPE_3DeHavillandDHC8M400DashM8	-4.62	0.75	-6.09	-3.13	1.00	3149	2904
AIRCRAFT_TYPE_3Boeing737M700D700LRDMax7	-0.25	0.37	-0.98	0.46	1.00	5659	3485

Family Specific Parameters:

	Estimate	Est.Error	1-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sigma	7.71	0.06	7.60	7.82	1.00	8653	2651

Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS and Tail\_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

The Rhat values are all 1.00, indicating appropriate convergence of the Bayesian fitting process. Less rounded values for DISTANCE are 0.0036 estimate, 0.0003 error, and (0.0029, 0.0042) confidence interval.

Inspection of the plots of parameters over the sampling iterations shows that each of the parameters appears to have a normal error distribution that is well described by the summary above. Correlations among the predictor distributions were not explored.

To illustrate the variability of predictions from the Bayesian model, predictions based on the 1000 draws from each chain were calculated on the test data. Then, the mean and variance of these predictions were calculated for every observation in the test data, and these were summarized by unique carrier, resulting in the following:

Carrier	Number of Flights	Mean Predicted Ground Time	Std. Dev. Predicted Ground Time
SkyWest Airlines Inc.	19,610	24.6	7.73
Horizon Air	10,661	17.1	7.74
Southwest Airlines Co.	7,375	15.9	7.74
Delta Air Lines Inc.	3,024	21.7	7.71
Alaska Airlines Inc.	2,355	25.3	7.72
American Airlines Inc.	2,234	24.8	7.76
United Parcel Service	2,045	16.4	7.73
United Air Lines Inc.	1,776	24.1	7.73
Federal Express Corporation	1,681	17.8	7.73
Envoy Air	1,302	26.8	7.74
Allegiant Air	951	18.8	7.73
Mesa Airlines Inc.	568	24.9	7.76
Gem Air LLC	158	10.0	7.73

Compass Airlines did not fly to or from BOI in 2021 and thus does not appear in the test data.

The standard deviation of the predicted ground time combines the uncertainty from what the parameters both explain and do not explain for ground time. The first is the parameter uncertainty from the model fitting, and the second comes from the residuals after the model fitting. The Bayesian point estimate for the latter,  $\sigma$ , is 7.71 and far larger than the estimated errors for all parameters, where the greatest value is 1.07. When the corresponding variances are merged together, the parameter certainty adds very little to the overall uncertainty of the predictions, made more explicit by the Bayesian model than seen in the GLM.

While this model is poor predictor of ground time for an individual observation, the relatively small error estimates for unique carriers are evidence that they have markedly different ground times exceeding their uncertainty.

### Task 5 (6 points)

Fit a random forest to the same target variable as the previous models using the same training and test data but independently select the significant variables that improve the predictive error on the test data. Then, use predictions from the fitted GLM from Task 3 and the random forest from this task as inputs to a stacked model, optimizing the form of the stage-1 model using the same test data as previously specified.

Continue the modeling section of your technical report by documenting the above modeling work in Task 5 and showing improvement in the performance of the model compared to the models in Task 3. Also, discuss strengths and weaknesses of the overall modeling process and its impact on addressing the business problem.

### Task 5 Response

*Candidates should read all tasks before starting work to see how particular models will be used and expanded. Task 6 makes clear that a distinction in predictors between random forest and GLM is really desired, and there is ample opportunity for this. The two challenges with the stacked model are managing the various training and test sets for the two stages and tamping down overfitting enough to get a benefit from stacking. Finally, the tasks have forced a variety of good and not so good modeling decisions, and the candidate has an opportunity to opine on those here. There are other strengths and weaknesses to consider besides those discussed here.*

### Random Forest Fitting

The random forest fitting, using the same train and test data, starts with the same predictors as the final GLM in Task 3. The random forest itself is unweighted due to technical constraints, but the test metric may continue to be weighted for optimizing the predictors. The initial random forest has an RMSE of 3.8456, distinctly worse performance compared to the GLM value of 3.5774.

Alternative parameters are tested for the number of trees, the number of variables used at each split (mtry), and whether to sample observations with or without replacement, all using the existing partition but varying the random seed for different hyperparameter tests. The default values of 500 trees, 2 variables, and sampling with replacement are affirmed.

Starting with a new random seed that produces a baseline RMSE of 3.8393, each variable is dropped independently and the resulting model tested. The only helpful move is to drop AIRPORT\_16, which leads to a vast improvement in RMSE, 3.5304. Other versions of airport, with 1, 4, or 8 specified airports are then tried from the GLM development work, and the best RMSE (3.5026) is with one airport (SEA - Seattle, Washington) as a binary predictor; including 4 airports is a close second at 3.5075 but not as good as the simpler model.

Additional variable rejected in the GLM fitting are then tried. Total airport departures does not help, but the number of flights on a particular flight plan (DEPARTURES\_PERFORMED) moves the RMSE down to 3.3843. Adding large airport from here does not help further, so the final random forest uses default hyperparameters and the same predictors as the GLM except to replace AIRPORT\_16 with AIRPORT\_1 and add DEPARTURES\_PERFORMED.



### *Stacked Model Fitting*

A variety of stacked models are built using the forms of the final GLM and random forest models. However, these stage-0 models should not be trained on the same training data as the stage-1 stacked model because predictions on the stage-0 models will have already seen data that the stage-1 model is training itself on. To remedy this, the predictions for the stage-0 models that will act as inputs to the stage-1 model are made on a five-fold random partition of the train data. Each of five separate stage-0 models (for each of GLM and random forest) is trained on four folds and make predictions on the fifth fold, like cross validation. These predictions, all on unseen data for the temporary stage-0 models, because the inputs for the stage-1 model.

Three stage-1 models are tried using now the final GLM and random forest models trained on all training data as before, since the test data will be unseen. All are weighted linear models, but their inputs vary, with the following results:

<b>Stage-1 Inputs</b>	<b>RMSE (Weighted)</b>
<b>Stage-0 Predictions, No Intercept</b>	3.3688
<b>Stage-0 Predictions, Yes Intercept</b>	3.4329
<b>Stage-0 Predictions, Yes Intercept, Common Predictors</b>	3.4692

Stacked modeled are prone to overfitting, and so the simplest stage-1 model, with just the two stage-0 predictions and without an intercept, has the lowest RMSE and is the only one to outperform the random forest. The stage-1 model gives a roughly 8%/92% weight to the GLM and random forest predictions respectively.

### *Modeling Strengths and Weaknesses*

The modeling overall has taken the following course in seeking out a progressively more accurate model, where candidate models were trained on 2016-2020 data and tested on 2021 data using RMSE weighted on the number of flights per month for a given flight plan.

This progression of modeling steps has all been on the same test data, a potentially dangerous practice chosen here due to time constraints. When the same partition of train and test data is repeatedly used to select predictors and model forms, the decisions become increasingly particular to that partition and less likely to generalize as well. While the year-based partition for test data is sensible in that it simulates the forward-looking performance of the model, further work would need to use a variety of partitions to affirm the modeling decisions.

A strength of the model is its ability to isolate the impacts of individual carriers from several other factors that also influence ground time, including the other airport, distance to that airport, the direction of travel to or from BOI, the presence and number of passengers, the type of aircraft used, and the frequency of the flight plan. Many of these are confounding factors in that certain carriers only serve certain airports or have certain types of flights. This predictive model isolates these impacts, with the ability to reliably isolate carrier affirmed by the GLMM. It provides a clearer view of what impacts ground time compared to descriptive analytics.

But this view is hampered, as discussed earlier, by working with average ground time for each flight plan instead of individual flights. The resulting residuals are underestimates of the variability of ground time

for individual instances, which are of greater interest to all parties involved in the dispute at the BOI airport. In addition, the inconsistency of the number of flights leads to a fat-tailed distribution of ground time where particular data points, like the 5-hour ground time some years ago, can have an outsized effect of the resulting model.

### Task 6 (4 points)

Using the random forest from Task 5 trained on all 2016-2020 data, apply partial dependence plots to explain the # most important predictors of ground time, based on variable importance, for flights departing from or arriving at the Boise airport in 2016-2020. Where plots are difficult to read, turn plots off to get a data frame of underlying values.

Write a section of the technical report to explain the most important predictors of the random forest model using the partial dependence plots. Include a comparison of the partial dependence plot results with the coefficients of the GLM from Task 3 where applicable and discuss the differences, with particular attention on unique carrier. Do not interpret findings in this task—just focus on model explanations.

### Task 6 Response

*Partial dependence plots are relatively easy to produce but can easily be misconstrued due to their global approach. Both general and specific differences will appear between the random forest and GLM, the specific differences being related to which predictors are in each model, choices that will vary by candidate. The distinction between model explanation and model interpretation is intentional and expected to be followed.*

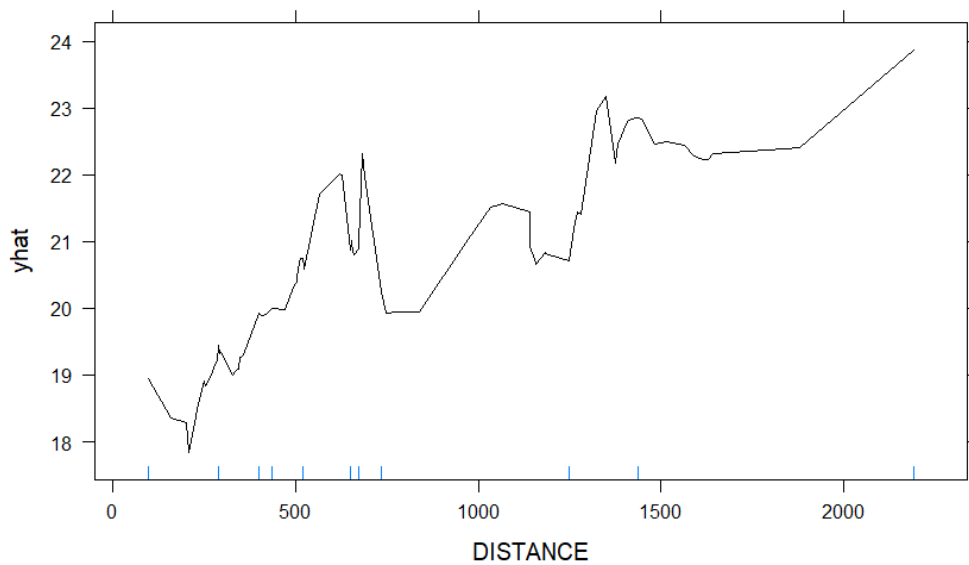
### Random Forest Explanation

Partial dependence plots are used to explain model predictions whose precise calculations are complex, such as in a random forest with 500 trees each contributing to the prediction. This global method estimates the impact of a variable on predictions by cycling through values of a variable, setting all observations for that variable to that value, and calculating the average prediction. These averages are unweighted and may include predictions on combinations of data that have not been encountered in the actual data, though care is taken to only predict actual values for the given variable.

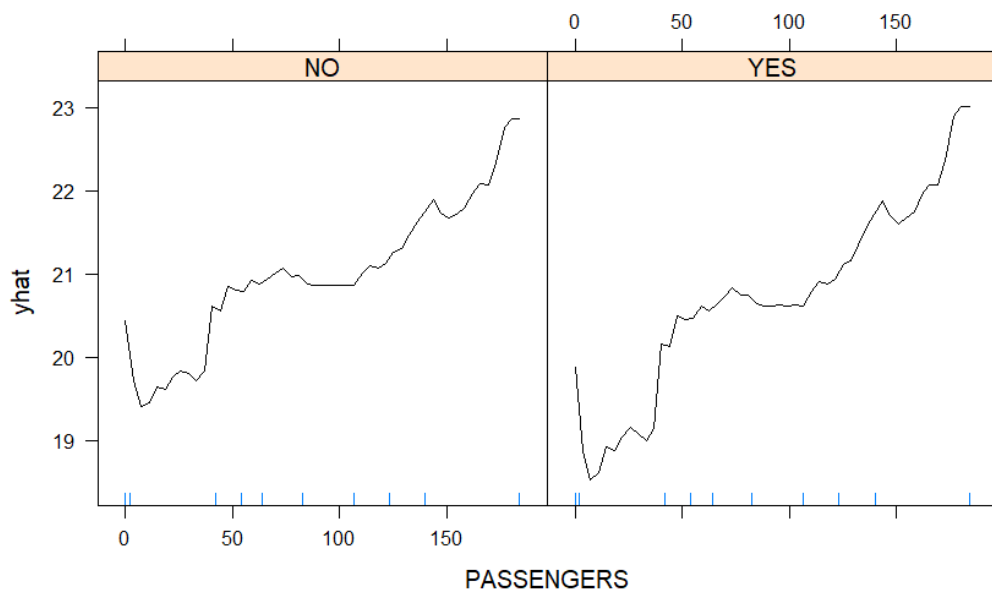
Variable importance based on the relative impact on mean squared error from scrambling each variable in isolation is as follows:

Variable	Importance (% Increase MSE)
UNIQUE_CARRIER	47.1
DISTANCE	29.4
PASSENGERS	24.6
AIRCRAFT_TYPE_3	20.4
DIRECTION	16.8
AIRPORT_1	15.9
WINTER	13.1
DEPARTURES_PERFORMED	13.0
HAS_PASSENGERS	9.6

The most important variable is UNIQUE\_CARRIER, followed by DISTANCE and PASSENGERS. As unique carrier will receive additional analysis, the other two variables have their partial dependence plots presented first.



The random forest predictions for ground time are generally increasing with distance through 1400 miles and then are flat. Some of the predictions forced into the partial dependence plot will be incongruous with other predictors, as such varying the distance to SEA, which is 399 miles away. The GLM has a slope of 4.1 minutes per 1000 miles, producing a similar slope at the more frequent shorter distances but extrapolating higher at longer distances.



The NO and YES boxes correspond to HAS\_PASSENGERS, which is included in the discussion with PASSENGERS due to its close relationship. This illustrates a shortcoming of partial dependence plots. The only realistic value for PASSENGERS on the NO plot is 0, where the average predicted ground time is 20.4 minutes. For YES, it is about a half minute shorter at 0 passengers (not possible) but then the trend

is increasing, crossing the breakeven point (20.4) with no 0 passengers around 45 passengers. The GLM has a steeper overall slope of 4 more minutes per 100 more passengers, corresponding to the more frequent lower part of the curve, but the breakeven point with 0 passengers is at 97 passengers.

#### *Focus on Unique Carrier*

Unique carrier is compared between the random forest and GLM based on their partial dependence results. For the GLM, partial dependence results have the same differences as the differences in coefficients between predictors because it is a linear model with no interactions. To focus on the relative differences between models without a biased intercept as seen in the GLM, Alaska Airlines is set to a value of 3.0 and others adjusted relatively to this. The actual partial dependence mean predictions are 24.2 (GLM) and 24.0 (Random Forest), a small difference that is ignored here in deference to more significant model differences among carriers within each model.

Unique Carrier	Number of Flights	Random Forest Relative Partial Dependence	GLM Relative Coefficient
<b>SkyWest Airlines Inc.</b>	109624	3.3	4.1
<b>Horizon Air</b>	67566	-2.3	-0.8
<b>Southwest Airlines Co.</b>	45558	-4.2	-6.9
<b>Delta Air Lines Inc.</b>	14391	2.3	-1.1
<b>American Airlines Inc.</b>	11337	3.6	0.2
<b>Federal Express Corporation</b>	<b>9953</b>	<b>-2.2</b>	<b>-3.5</b>
<b>United Air Lines Inc.</b>	8791	2.3	-4.0
<b>United Parcel Service</b>	<b>6627</b>	<b>-2.4</b>	<b>-4.3</b>
<i><b>Alaska Airlines Inc.</b></i>	6278	3.0	3.0
<b>Mesa Airlines Inc.</b>	4328	2.1	4.7
<b>Envoy Air</b>	4089	3.0	3.9
<b>Allegiant Air</b>	3633	-1.4	-3.6
<b>Compass Airlines</b>	2023	2.5	7.2
<b>Gem Air LLC</b>	1379	-3.2	-5.1

As a general observation, the random forest impacts vary less than do the GLM impacts. This is because the GLM forces a fit on any predictor no matter how sparse it is. It is sparse for the carriers low on the list, where the GLM has more extreme values (except Southwest Airlines, which appears to be truly more efficient on the ground). The random forest finds value in including carrier but does not necessarily set separate predictions for each level. Instead, the variations may arise in this view from interactions with other predictors where values for those predictors more heavily one airline or another. The GLM prediction has no such interactions.

Some values have changed dramatically. For example, United Air Lines goes from -4.0 in the GLM to 2.3 in the random forest, a six-minute increase. This can be attributed to the change in which airports are separated out as predictors. Most of United Air's flights in the test data go to DEN or ORD, which in the GLM have among the highest ground time adjustments. The random forest does not include either of

these airports as predictors, as multiple airports led to overfitting in the model selection process, and so the impact is transferred largely to United Air.

A similarity between the two models is that the two freight carriers have significantly negative effects, among the lowest 5 in each model. Across all models, the freight carriers have exhibited lower ground times than many other airlines once other effects on ground time are accounted for, to the extent each model can do so.

## Task 7 (6 points)

Write an executive summary to the head of the airport authority incorporating your analysis in tasks 1-6 but also including sections, such as statement of the business problem, not included in those tasks. The executive summary should not exceed 2000 words in length but may be considerably shorter.

### Task 7 Response

*Success from this task comes from being able to guide the reader, based on their experience and knowledge, through what is most important to know regarding the predictive analytics work that they can then apply. While the order can strictly follow that of the technical work, other methods of presenting the information can be as effective or more so. One challenge is, unlike the technical writing, remain concise and editing explanations carefully to communicate accurately what can and cannot be concluded regarding this business problem.*

### Business Problem

As head of the Boise airport authority, you have asked me to produce insights on what factors impact ground time at BOI despite my inexperience with aviation data. The context for this request is a public dispute between the airport authority and two prominent freight airlines, FedEx and UPS, who are negotiating better terms from the airport and leveraging fears regarding a planned expansion of passenger capacity. I will clean and analyze monthly federal aviation data you have collected using a variety of predictive analytics models to isolate the impacts of airline and other factors where possible, delivering this report after four days of work. You will use these impacts to counter potentially inaccurate claims made by the freight airlines in an upcoming negotiation meeting.

### Data Overview

You provided data on all domestic flights from 2016-2021 from the Bureau of Transportation Statistics table [T-100 Domestic Segment \(U.S. Carriers\)](#). I selected from this data all flights arriving or departing BOI during 2016-2021, comprising 12,370 records. Each record in this data totals the ramp-to-ramp time, air time, and other statistics for all flights in a calendar month corresponding to a particular flight plan, my term for the combination of airline, airports, and aircraft involved. The records notes the number of departures performed, but 79 records recorded zero for this and were removed.

All variables in the data were inspected for issues and new variables were created to improve the ability to isolate which variable impact ground time. Notable data preparation actions included:

- Ground time was calculated, per your specification, as ramp-to-ramp time less air time, and average ground time each month for a particular flight plan was what was to be predicted.
- Rather than have separate predictions by origin and destination, one of which is always BOI, the non-BOI airport was used for prediction and the direction of travel also noted. Additional data about the relative size of each airport was added from a public source.
- The data provided indicated 41 unique carriers had flown to or from BOI, a large number that could hamper effective predictions, so the data was reduced to only include 14 carriers who had logged at least 1000 flights over the six-year period. Each of these flew at least five of those years.

The prepared data included 11,737 records and 22 variables, including ground time, though some of these variables overlap and cannot be used as independent predictors of ground time.



The data provided has major limitations that inhibit my ability to make accurate predictions regarding ground time in BOI for the purpose stated above:

- 1) The ground time measurement includes ground time at both BOI and the other airport, with no reliable way to separate these,
- 2) The monthly average of ground time suppresses more extreme (and thus more informative) single-flight ground times by combining them with typical ground times,
- 3) Ground time is being used as a proxy for financial costs from runway and other delays, but other factors than just the amount of time on the ground affect these costs, and
- 4) My inexperience with details of the data, such as aircraft type, may cause me to overlook important impacts or overemphasize immaterial effects on ground time.

The insights shared in this document are subject to and may be nullified by these limitations.

### *Modeling Overview*

I applied a series of predictive models to the prepared data. Predictive modeling can help separate out the overlapping contributions to a given outcome. In this case, the outcome is average ground time and the contributors to be separated out include airline, airport, type of flight, and so on.

The following models were independently fitted to the data, choosing the contributing impacts that most accurately predicted ground time:

- **Generalized linear model (GLM):** model provided straightforward impacts but has difficulty sorting out complex combinations of impacts
- **Generalized linear mixed model (GLMM):** similar to GLM but can more readily accommodate changing contributors, such as new airlines
- **Bayesian model:** measures the amount of uncertainty in the estimated impacts of the above
- **Random forest model:** more readily handles complex combinations of impacts but more difficult to explain
- **Stacked model:** a combination of the GLM and random forest model to improve accuracy

The predictive power of each model was tested by training it only on flights data from 2016-2020 and assessing how well it predicted 2021 ground time. Models with the smallest root mean square error (RMSE), a “typical error” with low tolerance for predictions being very far from actual average ground time, were selected.

A variety of models was fit as they vary in accuracy, interpretability, and other aspects. However, all models used were found to be similar in accuracy. Where predicting that all flights would have ground time of 20.4 minutes in 2021 produces a typical error of 5.1 minutes, all the models used reduced this error to 3.4 - 3.6 minutes. For ease of interpretation, the GLM and GLMM are used below when discussing impacts on ground time, supplemented by findings from other models.

### *Model Results: Impacts on Ground Time*

Several contributors make a meaningful contribution to ground time, using impacts from the GLM with notes from additional modeling (see next page):

Contributor	Impact on Average Ground Time	Notes
Distance	+1 minute per 250 miles	Impact does not extend beyond 1000 miles
Passengers	+1 minutes per 25 passengers	Freight only equivalent to 100 passengers
Winter	+2 minutes if in Dec/Jan/Feb	November partially elevated as well
Direction	+1 minute if departing BOI	As compared to arriving at BOI
Aircraft Type	Varies – see below	Identified separately for 3 most used types
Airport	Varies – see below	Less confidence for rarer airports
Airline	Varies – see below	Significant interaction with airport

On aircraft type, two stood out. Flights using Embraer ERJ-175 had a -2 minute impact and those using De Havilland DHC8-400 Dash-8 had -3 minute impact. Only some airlines use these types on certain flights, and your expertise is needed to further interpret the significance of this finding.

#### Airport and Airline

The GLMM provides an estimate of the distinct impact flying to or from each airport has on average ground time. It similarly provides estimates for each airline. Each of these sets of estimates are impacted by the other, as not all airlines fly equally often to and from each airport. The top 16 airports, considered a reliable set of estimates, and the remaining airlines after filtering out those with less than 1000 flights over 2016-2021, have the following estimated impacts, in minutes, on average ground time:

Airport	Impact
SEA	3
SLC	1
PDX	-1
DEN	3
SFO	3
GEG	0
PHX	-1
LAX	3
ORD	5
OAK	1
LAS	1
MSP	-3
SMF	0
SAN	-2
SJC	0
DFW	2

Airline	Impact
SkyWest Airlines Inc.	4
Horizon Air	-1
Southwest Airlines Co.	-7
Delta Air Lines Inc.	-1
American Airlines Inc.	0
Federal Express Corporation	-3
United Air Lines Inc.	-4
United Parcel Service	-0
Alaska Airlines Inc.	3
Mesa Airlines Inc.	4
Envoy Air	4
Allegiant Air	-3
Compass Airlines	7
Gem Air LLC	-3

Each column is in descending order of number of flights to and from BOI during 2016-2020. The impacts are relative to an airport or airline not on the list, though for the latter, a new airline is assumed to have performance equal the average of the 14 carriers listed. The airports listed mostly have positive impacts

increasing airtime because they are larger airports than most others on the list. Smaller airports tend to have lower ground time.

The freight airlines, FedEx and UPS, have relatively good ground time compared to other airlines flying to and from BOI after isolating the impacts from which airports they fly to and other factors. When using the GLMM trained on 2016-2020 data to predict unseen 2021 data, the actual average ground time was only 0.3 (FedEx) and 0.7 (UPS) minutes over the prediction, more accurate than most airlines. There is reason to believe these trends for the freight airlines will continue beyond 2021.

Variables in the data not included in this section were not found to have a reliable and significant impact on ground time once the above contributors were considered. Notably, the number of flights per month involving BOI was not found to have a significant impact when considering 2016-2020 data on any of the models, countering the claim made by the freight carriers. This evidence is, however, limited to the data available and cannot anticipate what would happen if the number of flights significantly exceeds the 54K observed in both 2019 and 2021, in addition to other limitations described earlier.

#### *Next Steps*

I look forward to discussing this analysis with you soon after receipt in preparation for your meeting with the freight carriers. As this report was assembled in a short period of time with limited data and aviation knowledge, I ask that you consider the following next steps:

- Discussing these results and isolating key points of interest where more explanation of the results would be helpful
- Seeking out reliable data with ground time by individual flight, which would sharpen the analysis
- Expanding the analysis to include flights not including BOI so that the relative performance of BOI compared to other airports may be ascertained