STAA57 W21 Draft Report

Group 4 (Adham F, Jason Y, Mohamed T, Wesley M)

Link to RStudio Cloud shared project: https://rstudio.cloud/spaces/115177/project/2202760

Introduction

The topic of this paper is to investigate the aircraft operating cost of DFC, particularly the Cost per Session. Two main factors affect the aircraft operating cost which is fuel price (we are using gasoline) and the duration of the session. It is intuitive to think that the duration of a flight will be affected by the weather. Factors such as cloud coverage, wind and precipitation can affect an aircraft/pilot's performance, thus the duration of flight/lesson. With an increase in flight duration comes an increase in maintenance and fuel costs. Therefore, we should expect that there will be weather factors in our data that can be significant predictors for the aircraft operating cost to a certain extent. It is also known that fuel price does change monthly/seasonally depending on seasonal travel, such as during Christmas and Summer. Therefore, we should expect that there are seasonal changes to the aircraft operating costs. Since seasonal changes in fuel costs are fairly obvious and weather data is closely correlated to seasonal weather changes, we expect the weather factors that we are using to be able to predict the aircraft operating cost. Also, since COVID-19 has dramatically impacted the Airline industry, which is closely related to fuel, we will also be investigating COVID-19 being a factor in predicting the aircraft operating costs.

Questions:

1. What are the trends in fuel costs over different time units (daily/seasonally)?
    - For this question, web scrapping was used to see how fuel price changes over time and attempted to see how these changes affect the Cost per Session for DFC.
    - A monthly analysis was considered, however, due to the time period, it would result in a relatively hard to interpret analysis of the fuel price fluctuations that can also be deduced using a seasonal time period analysis.

For the next two questions, we will use Linear Regression models. For context:

This is a method to predict numerical values like cost/duration given new input data (exercise plan, season, weather, fuel, etc.). The model is essentially a math equation generated using the given historical data plus external factors we add (e.g. weather, fuel cost).

From this model, we can also see which factors have a significant effect on what we are trying to predict.

2. What factors affect the duration of the sessions?
    - We will use a linear regression model to assess the significant factors associated with the duration of the sessions.
3. What factors affect cost in a significant matter?
    - We will repeat the same process as the previous question, however predicting cost instead of duration.

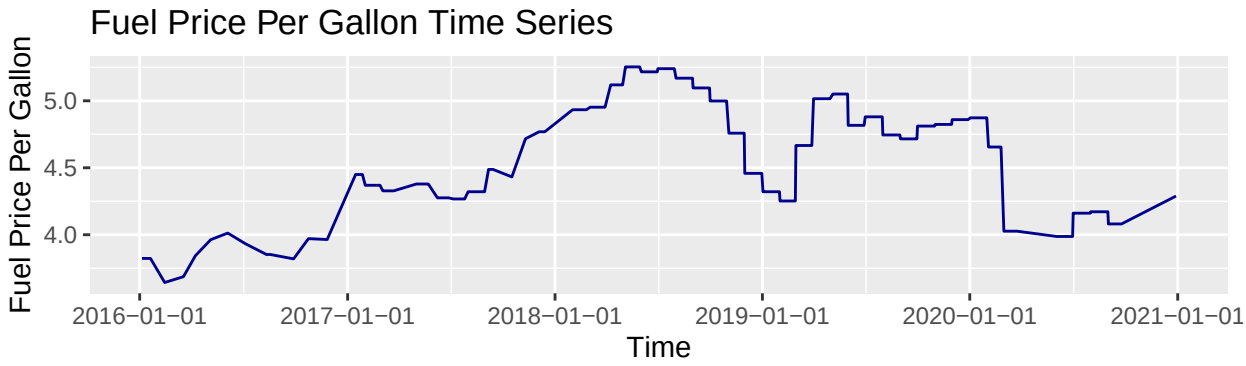
Data

The external data that is used in this project were gathered from the following sources:

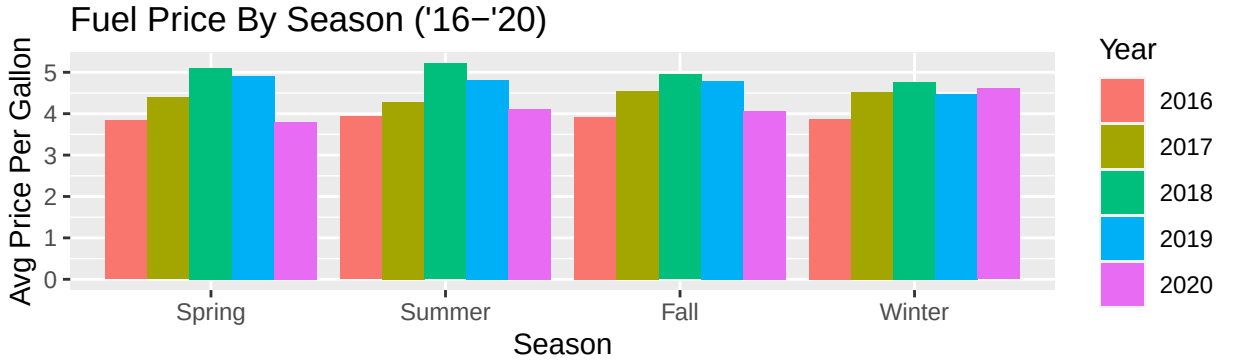
- Fuel (https://open.canada.ca/data/en/dataset/cbb9182b-df94-4ae6-8a85-b75a34b2f990)
  - Historical fuel data by month, including the cost/gallon
  - Assuming that all planes used in the flight school use the same fuel
  - From preliminary research, it was found that C-172, C-150 and C-152 aircraft models only consume AVGas and gasoline. We decided to use gasoline data for our analysis.
- Maintenance costs and repairs (https://cessna150152club.org/Costs):
  - The annual operating cost of a Cessna plane model for maintenance and repairs
  - Based on this research, an assumed 15 dollars static cost per hour was added to each training session cost
  - Also, the fuel consumption per hour is estimated to be 6 gallons
- Weather (https://oshawa.weatherstats.ca/download.html):
  - Historical weather data by day, including temp, precipitation, wind, etc.
  - Gathered data from Oshawa location

Analysis

Question 1: What are the trends in fuel costs over different time units (daily/season)?



From the above graph, it can be seen there was a drop in the fuel price per gallon in 2020, leading to lower aircraft operating costs. Therefore, we expect that the Cost per Session to be lower during 2020, i.e. during the pandemic.



The above graph shows the average price per season from 2016 to 2020. There is no clear pattern to

the fluctuations of the gasoline prices over the season. It is interesting to note that the gasoline prices during the pandemic are not at their lowest as for almost all the seasons 2016 had the lowest average price per gallon, therefore that is an indicator that the pandemic may not be an important factor when analyzing the Cost per Session model.

Question 2: What factors affect the duration of sessions?

How was the data prepared?

For this model, we combined the given flight data with Oshawa weather data. In particular, we referenced https://silicium.dk/fsx/FSWeb/LearningCenter/TheWorld/WeatherAviationWeatherBasics.htm, and identified a set of weather factors that were deemed to have a strong effect on the duration of a session.

- `Avg relative humidity`
- `Avg dew point`
- `Avg sea pressure`
- `Avg visibility`
- `Avg health index` (air quality)
- `Precipitation`
- `Avg cloud cover`
- `Avg temperature`

In total, our factors consist of the selected weather factors, the exercises conducted, the season, and whether or not the session took place in 2020 (Covid year).

We also split our data into two sets at an 8:2 ratio:

1. Training (80% of data) set is used to build the equation and determine the significant factors.
2. Test (20% of data) set is used to test accuracy of the model.

Linear regression model for Duration per Session:

Table 1: Most Significant Factors

term	minutes
Exercises_12	5.0007432
Exercises_16	11.7017440
Exercises_19	-44.6552631
Exercises_23	27.8280248
Exercises_24	29.9557905
Exercises_30	-3.1160614
precipitation	-0.4242294
SeasonSummer	4.7112552

Analysis for Duration per Session Model

From our model, we can see that the most significant factors for predicting duration are:

- Exercises_12: Stall

- Exercises_16: Take-Off
- Exercises_19: First Solo
- Exercises_23: Pilot Navigation
- Exercises_24: Instrument Flying
- Exercises_30: Radio Communications
- precipitation: How much avg precipitation received that day
- SeasonSummer: Whether or not the session took place in the summer

The model predicts that on average:

- Having Exercise 19 (first solo) decreases the duration by 44min
- Every mm of precipitation decreases the duration by 0.4min
- The session being in the summer increases the duration by 4.7min
- etc.

For the remaining factors refer to *Table 3* in the appendix.

Our model has an R^2 value of 0.88 (*Table 4*), meaning that about 88% of the changes in the duration can be explained by changes in our factors. To test the model's accuracy, we tried predicting the duration by inputting the test data and compared the results to the actual duration of the test sessions. We then took the average of the differences (Mean Absolute Error or MAE) and got an average error of ~14min. This means that using our model to predict new data will result in a value that's off by 14min on average.

Question 3: What factors affect cost in a significant manner?

How was the data prepared?

We repeated the same steps as was done in question 2, however, we added an extra column for the estimated cost per session.

The aircraft operating cost per session was calculated using the equation below:

```
CostPerSession = (Duration * (FuelPricePerGallon*GallonsPerHr + OperatingCostPerHr))
```

The aircraft operating cost equation uses duration, fuel price per gallon, and an average number of gallons consumed per hour (6) as well as an estimated maintenance and repairs cost per hour (15).

Table 2: Most Significant Factors

term	cost
Exercises_19	-29.505372
Exercises_23	21.958909
Exercises_24	19.249699
Exercises_30	-2.211338
SeasonWinter	3.457657

Analysis of the Cost Per Session Model From this model, we can see that the most significant factors for predicting cost are:

- Exercises_19: First Solo
- Exercises_23: Pilot Navigation
- Exercises_24: Instrument Flying
- Exercises_30: Radio Communications
- SeasonWinter: Whether or not the session took place in the summer

The model predicts that on average:

- Having Exercise 19 (first solo) decreases the cost by $29.50 per session
- The session being in the winter increases the cost by $3.45 per session
- etc.

For the remaining factors refer to *Table 5* in the appendix.

Our model has an R^2 value of 0.91 (*Table 6*), meaning that about 91% of the changes in the duration can be explained by changes in our factors. We did the same testing method as the above and got an average error of $11.21.

Summary

Overall, we can see that Exercises 19, 23, 24, and 30 seem to have a significant effect on the duration. Surprisingly, aside from precipitation, most weather conditions were not strong predictors for duration or cost. Summer sessions seem to be longer, whereas winter sessions seem to cost more, likely due to increased fuel costs in 2020.

Appendix

Table 3: Most Significant Factors

term	minutes
avg_dew_point	-0.0084790
avg_pressure_sea	1.1377986
avg_visibility	-0.0000316
avg_health_index	2.0676638
avg_cloud_cover_4	0.5210600
avg_temperature	0.0884531
COVIDFALSE	-61.1036012
COVIDTRUE	-56.9494041
Exercises_1	1.7740817
Exercises_2	-5.2066987
Exercises_3	-4.0609103
Exercises_4	2.4035876
Exercises_5	3.5304633
Exercises_6	-2.4802650
Exercises_7	3.2900842

term	minutes
Exercises_8	-0.6099830
Exercises_9	1.5760447
Exercises_10	2.0295462
Exercises_11	1.3150231
Exercises_12	5.0007432
Exercises_13	3.1599839
Exercises_14	1.5986114
Exercises_15	2.1307731
Exercises_16	11.7017440
Exercises_17	-2.4302614
Exercises_18	-5.6741241
Exercises_19	-44.6552631
Exercises_20	1.9623418
Exercises_21	-2.8778298
Exercises_22	-0.0995179
Exercises_23	27.8280248
Exercises_24	29.9557905
Exercises_29	0.1643555
Exercises_30	-3.1160614
precipitation	-0.4242294
SeasonSpring	1.5820476
SeasonSummer	4.7112552
SeasonWinter	0.7761457

Table 4: Duration Model Summary

adj.r.squared	r.squared
0.883447	0.885597

Table 5: Most Significant Factors

term	cost
avg_dew_point	-0.0940106
avg_pressure_sea	-0.4175415
avg_visibility	0.0002016
avg_health_index	0.8597489
avg_cloud_cover_4	-0.3243971
avg_temperature	0.1642045
COVIDFALSE	80.8050449
COVIDTRUE	79.0196440
Exercises_1	-0.1266052
Exercises_2	-2.2589053
Exercises_3	-3.1788506

term	cost
Exercises_4	1.2884758
Exercises_5	4.2610117
Exercises_6	-1.2327129
Exercises_7	0.8771994
Exercises_8	1.8808227
Exercises_9	0.0105638
Exercises_10	1.3093868
Exercises_11	2.3499087
Exercises_12	2.6172431
Exercises_13	3.2478026
Exercises_14	1.3602220
Exercises_15	1.7209143
Exercises_16	2.2193036
Exercises_17	-0.6803567
Exercises_18	1.7604284
Exercises_19	-29.5053718
Exercises_20	0.9892118
Exercises_21	-2.0662619
Exercises_22	0.5648340
Exercises_23	21.9589087
Exercises_24	19.2496985
Exercises_29	-0.6055349
Exercises_30	-2.2113376
precipitation	-0.1306294
SeasonSpring	1.7963173
SeasonSummer	1.2937247
SeasonWinter	3.4576565

Table 6: Cost Model Summary

adj.r.squared	r.squared
0.9087474	0.9104274