STAA57 W21 - Project Proposal

Group 11 (Alice Huang, Dominic Ma, Jalal Kassab, Vanshika Virmani)

Link to the shared RStudio Cloud project that created this report: https://rstudio.cloud/spaces/115177/project/2132391
The final R code can be found in the Proposal_Template_FINAL.Rmd file.

Introduction

We are interested in efficient pricing of flight training courses. So, we plan to analyse information useful to the Durham Flight Training Centre, in terms of the costs that they must incur for offering the training, and how they should conduct training with the exercises and aircrafts that will maximize the chances for success in training aviation students.

To do so, we are defining student 'success' by determining whether he/she can fly solo at least once during the course of their training.

For the cost analysis for the Durham Flight Centre, we would like to estimate:

1. How much money the training centre has spent on fuel and maintenance for different aircraft models over the past few years.

- From Durham Flight Centre's website, we know on average how much the centre was charging for flying each plane per hour. This will help us determine the profit margins for the flight centre.

2. Whether the number of flight hours have differed across different times of the year. And whether some planes are more in demand during certain months.

- This will help us determine how the demand varies for the flights throughout the year. If we're able to see that certain planes are more in demand during certain months, then we can analyse how the flight centre can adjust the prices for different types of trainings to increase profit.

3. How much money it costs to train the average student until they can safely take off, land and fly solo.

For answering questions in regards to the training schedule for students, we would like to build a "profile" for students. In order to determine how the centre can maximize student success, we want to know:

4. What are the types of exercises that students who were able to fly solo at least once completed that set them apart from students who weren't able to fly solo at all?
5. Which types of planes did the 'successful' students use to complete certain exercises? This will help us map out information for the Durham Training Centre, as to which planes they should allocate for which exercises to maximise student success in training.

By further analysing the types of exercises and the aircraft models suited for those exercises, we'd be able to determine how the training centre should schedule certain exercises for students. More student success can in return help the centre attract more customers.

Data Analysis Plan

1. To estimate how much money the flight training centre has spent on fuel and maintenance costs, we will acquire a range of estimates for the hourly fuel consumption and maintenance fees of each aircraft and we will add up the duration of time spent on training with each type of plane. This will be done by summing up the duration column for each type of plane. This will provide us with a graph that has the x axis being the types of planes, and the y axis being the total cost of flying that plane. To estimate revenue from the trainings, we will find values of how much they charge each student for flying the different types of planes from the training centre's website, and multiply the values with the total duration of flights for each plane. We will then subtract the cost values for each plane for the revenue values to estimate the flight centre's profit margin on each type of plane.

2. We will group the duration of flights by month for each year, and sum up the number of hours. This gives us the plots with how long the students trained for in each month of the corresponding years.

3. To estimate the cost for the "average student" in the training centre, we will group the data by different student IDs and plane types and then find the average number of hours spent in flight training by each student for each plane type. We will then multiply these values with the average costs from above, and create a graph with the x axis as the types of planes and y values as the average cost for training a student.
4. We will find the students who flew solo at least once by finding which students' training types contained the substring "solo". And we will filter out the students who didn't fit this criteria. We will then map the frequency of exercises completed by each student, and add the frequencies together for all exercises. This graph will show us the most frequent exercises that "successful" students completed.
5. Repeat the same process with filtering out students who got to fly "solo". We will then add up the duration of hours for these students with the types of planes used to see if they used one plane more frequently than the other. We will also group the most frequency exercises completed from above by the type of airplane to see which planes are better suited for which exercise.

Data

We intend to use a dataset provided by our client, the Durham Flight Training Centre. This dataset logs the exercises students completed with different instructors on different dates. It also specifies the durations of the training sessions, whether the exercises were completed solo or with an instructor, and which aircrafts were used. The dataset only spans the years 2002, 2015-2020, so we can only observe patterns and compare trends over the span of five years, and it may not be the most suitable for generalizing trends to longer periods of time. Since there is a gap of 13 years between 2002 and 2015, we didn't think it'd make sense to do a lot of analysis on the data from 2002 as a lot of factors such as aircraft condition, market conditions, etc. could have changed in the meantime.

The data we gathered on the aircrafts' ownership costs (C-152, C-172, and C-150) was obtained from aopa.org, the website of the "Aircraft Owners and Pilots Association", and https://cessna150152club.org/, the website of the "Cessna 150-152 Club", a membership club and nonprofit dedicated to educating prospective and current pilots about the Cessna-152 and Cessna-150. We believe these organizations' numbers should be reliable since these organizations and communities have access to active users of Cessna-172, Cessna-152, Cessna-150 planes.

However, the data gathered from the "Cessna 150-152 Club" may be a little biased since the entire organization is based on those two planes and the content on the website seems to feature more positive anecdotal experiences with the two planes. Furthermore, the ownership club may not include as much information for users who did not pay their club membership fee, and others who are not active on their internet forums. Due to these concerns, we checked if their numbers for ownership costs were consistent with other websites, and they were, so we chose this as a data source.

The website for the "Aircraft Owners and Pilots Association" helped us find out the fuel efficiency and price for all three planes. But we had to decide whether we should use the average used price of the planes or the price of the "reimagined" planes, which have been overhauled, repainted and serviced. We ended up using the price of the "reimagined" planes since it gave us a better idea about the actual price difference without any external variables.

Both the "Aircraft Owners and Pilots Association" and "Cessna 150-152 Club" are based in the United States so costs, regulations and experiences may be biased towards English-speaking American users of Cessna-172, Cessna-152 and Cessna-150 planes. Our client is based in Southern Canada, close to the US border, but we believe certain costs like a plane's fuel burned per hour, should not depend on region. Insurance and inspection costs may vary by region, however.

Importing and Formatting the Data in R

We found estimates for fuel and operating costs from the "Aircraft Owners and Pilots Association'' and the "Cessna 150-152 Club". These seemed to be the most accurate numbers we could find. These websites did not post their numbers as datasets with csv, txt, xlsx, xml, or other formats that are convenient to work with. In particular, on the "Cessna 150-152 Club" website's "Members Only" pages, it says that servers containing their data were hacked, so the organization had to pull their data offline and is now in the process of "building a modern and secure database structure". Since we could not find a pre-made dataset, we made a csv file, and entered the hourly estimates of various costs for operating and maintaining the aircraft in the tidy data format. We put this csv file in our data folder with the rest of our project. Wherever the sites provided a range of values, we just took the median for our computations as the median is a good indicator of the "middle value" in a distribution and is less likely to be affected by outliers.

An issue that we ran into while trying to organize exercise data was the fact that the elements were stored as string rather than list objects. Judging from the last few lines of the preprocessed data script, this was not meant to be the case. Although we could not figure out why those lines didn't work, we managed to work around it by creating a copied dataframe with the changes we

wanted. In hindsight, it's likely because dataframes are immutable. This will be useful for future analyses on the frequency of Exercises completed.

Furthermore, when we were plotting the months of the year vs the duration of training hours, we found a year corresponding to the value "201", which we speculated was a typing error. So when graphing the data, we only graphed for years past "2010". We also found an instance where a plane was entered as "C152" while its counterparts were entered as "C-152". We ensured all Cessna-152 planes were entered as "C-152" for consistency. If we run into more typos or strange data values, we will communicate with Durham Flight Training Centre, and change or filter values as appropriate.

Analysis

We wish to estimate the operating and maintenance cost the training centre has spent on each type of aircraft. Thus we are interested in the total amount of time students spent with each of the different types of aircrafts. We found estimates for how much fuel each plane burns, in gallons per hour, the unit pilots conventionally use to measure hourly costs. Computing the hours spent on flying the different aircrafts will help us estimate how much fuel was consumed by each type of aircraft, and in turn, how much money was spent on fuel for each aircraft.

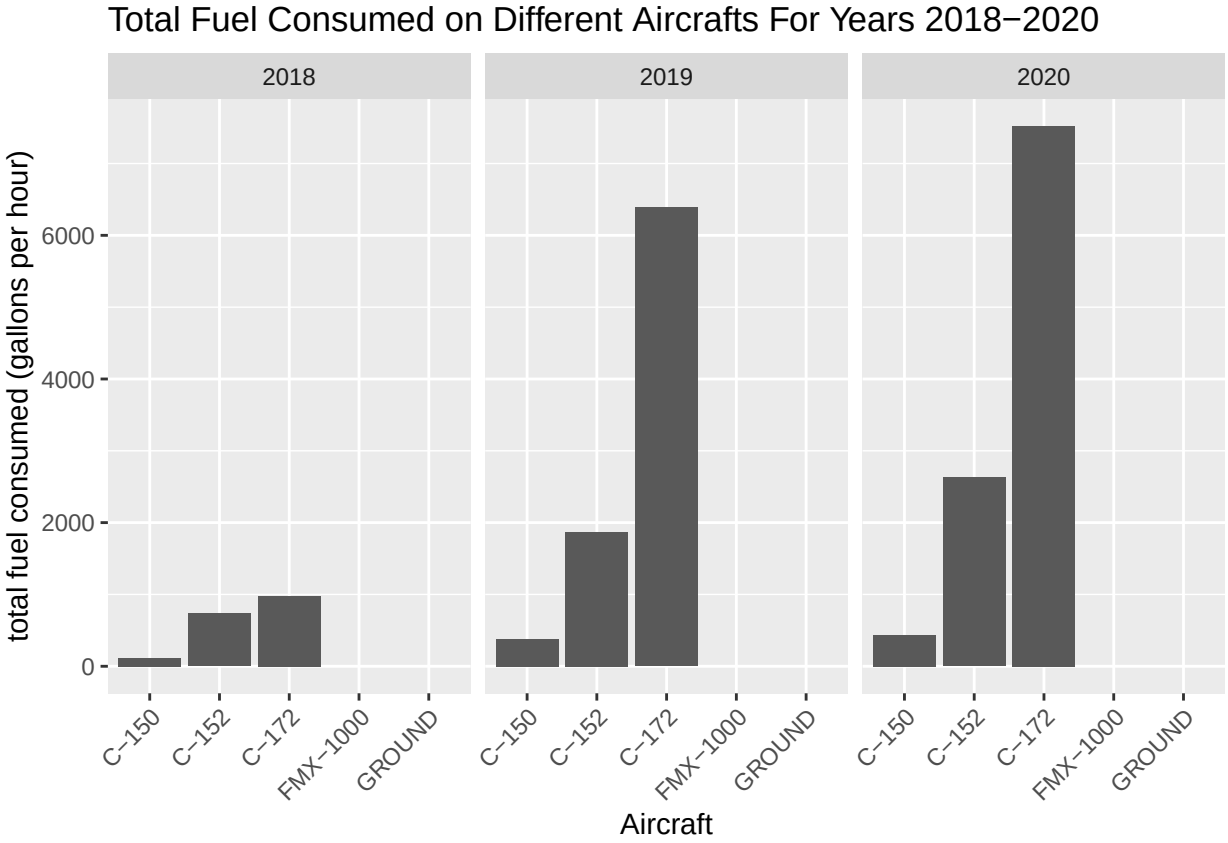
```
#Here is a table that shows the total time the centre spent on each type of aircraft from 2018-2020.
my_cd %>%
  filter(Year >= 2018) %>%
  group_by(Year, Aircraft) %>%
  select(Aircraft, Duration, Training_Type) %>%
  summarize(total_time_aircraft = sum(Duration))
```

```
## Adding missing grouping variables: `Year`
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

```
# Here is a table that shows the total fuel consumed per aircraft over the years 2018-2020.
# Fuel_Cost_Gallons_Session computes the cost of fuel for that session
my_cd %>% left_join(planecosts, by="Aircraft") %>%
  mutate(Fuel_Cost_Gallons_Session = Duration*Fuel_Cost_Per_Hour) -> clean_data_withcosts

clean_data_withcosts %>%
  filter(Year >= 2018) %>%
  group_by(Aircraft) %>%
  summarise(total_fuel_consumed = sum(Fuel_Cost_Gallons_Session)) -> total_aircraft_costs

total_aircraft_costs
```

Here are some graphs that show the total fuel consumed per aircraft over each year in the range 2018-2020.



We found that over the years 2018-2020, the training centre's Cessna-172 aircrafts consumed around 15719.76 gallons of fuel in total. The Cessna-152 aircrafts consumed around 5733.39 gallons of fuel in total. The Cessna-150 aircrafts consumed around 997.92 gallons of fuel in total. This is consistent with the fact that over the span of those 5 years, the Cessna-172 was used for training more often, and its hourly fuel cost is more expensive than that of the Cessna-152 and Cessna-150. It is interesting to note that the training centre spent nearly 2 times the amount of time on Cessna-172 than the Cessna-152, but it spent close to 3 times the amount of fuel on Cessna-172 than the Cessna-152.

Now when we looked at the fuel consumption per year, we saw that in 2019-2020, the Cessna-172 consumed much more fuel than the Cessna-152. In 2019, the C-172 consumed approximately 6622.56 gallons of fuel while the C-152 consumed approximately 2009.95 gallons of fuel. In January-March, June-September 2020, the C-172 consumed approximately 8106.84 gallons of fuel while the C-152 consumed approximately 2940.20 gallons of fuel. Again, this is consistent with the frequency of Cessna-172 training sessions during those years.

A limitation of our current analysis is that we assumed the price of fuel consumed per hour remained constant over 2018-2020 regardless of aircraft age and condition, nature of Exercise, market conditions for fuel pricing (especially given COVID-19 pandemic), and other factors.

Monthly Fuel Costs

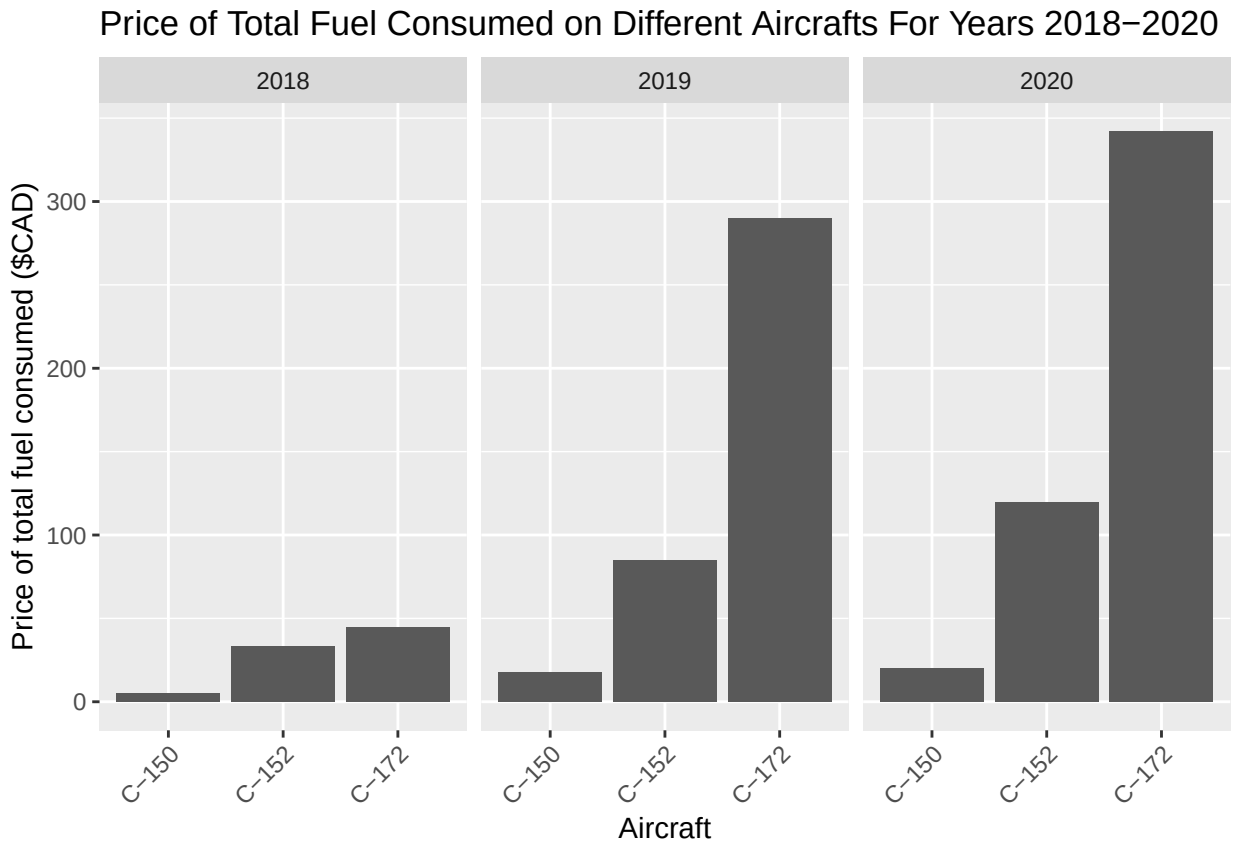
Previously, we computed cost of fuel in terms of gallons per hour, which is the standard unit that pilots use. Now we want to translate our findings towards monetary costs.

We found cost of aviation gas in $CAD/L from the City of Oshawa's site which lists Oshawa Executive Airport's Improvement Fees. Under Airport Improvement Fees, it says Avgas Fuel Sales are $0.01 per litre. We thought this was the most accurate estimate we could get for fuel pricing in dollars, since Durham Flight Centre is in the same city as this airport (and actually uses this airport for training) so region and market conditions affecting fuel prices should be the same.

1 Litre = 0.2199692 gallons. So $0.01/litre is approximately $0.04546/gallon. Multiplying gallons of avgas consumed for each session by the price of gas per gallon gives price of gas consumed for that session.

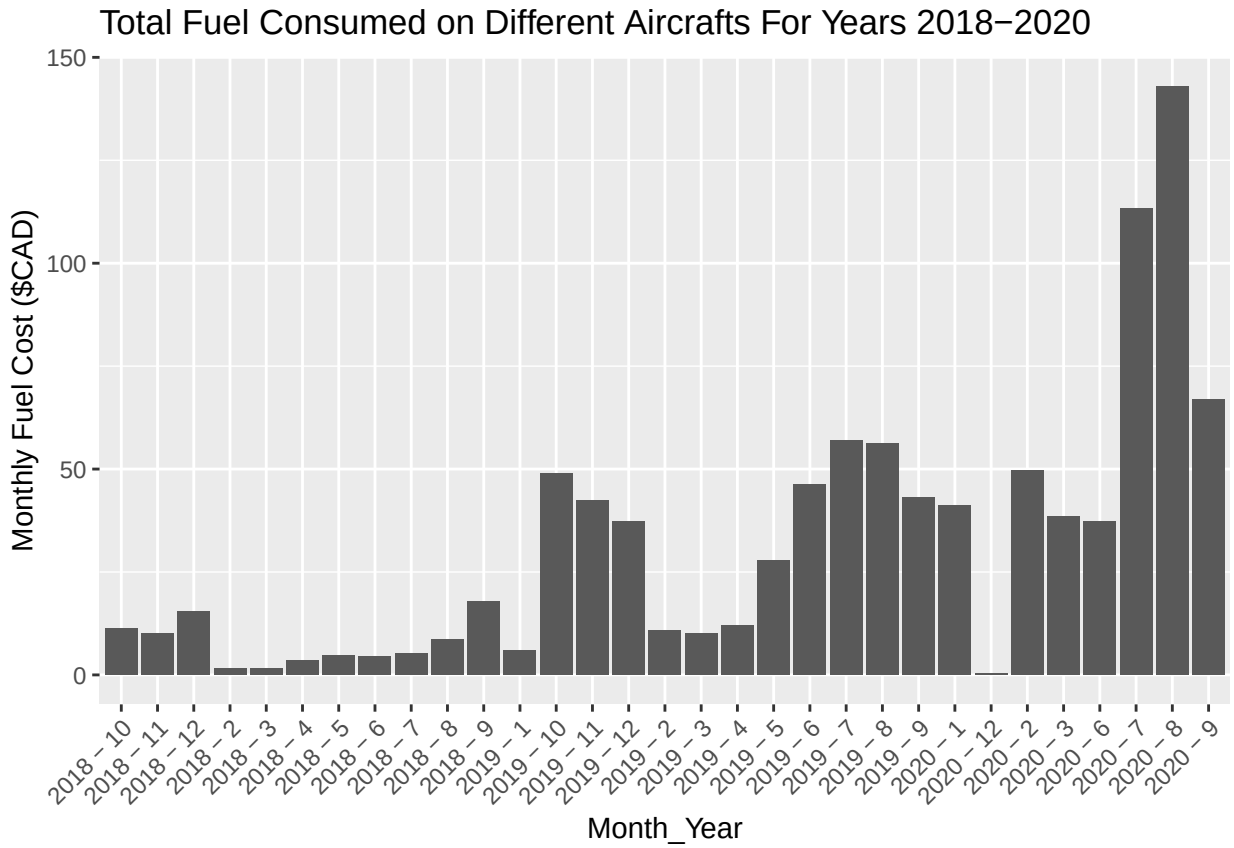
Let's mutate a new column that multiplies the number of gallons consumed by its corresponding price depending on plane used.

```
## `summarise()` has grouped output by 'Month_Year'. You can override using the `.groups` argument.
```




```
fuel %>% ggplot(aes(x = Month_Year, y = total_monthly_fuel)) + geom_bar(stat = "identity") + theme(legend.positi
```

```
## Warning: Removed 8 rows containing missing values (position_stack).
```



Monthly Instructor Costs

For calculating the instructor cost, we found the duration of "LF-dual" training type and "Grounded" Aircraft type. Both of these correspond to when the instructor was involved in either pre-flight or flight training with the students. For the "Grounded" Aircraft type, we noticed that the duration was N/A (not given). Initially we thought of completely ignoring the Grounded Aircrafts and the instructor fee for these sessions, but we realised that ground training makes up a significant cost of the packages that students end up paying for. And completely disregarding this fee would highly skew the data. So we approximated the duration of each session of ground trainings by comparing it to the durations of other sessions for which the students completed the exact same combination of exercises. We then averaged those durations to approximate the hours of ground training each student trained for. From here, we mutated a column corresponding to the approximated ground training hours for Grounded Aircraft type. And we combined the two tables (original and grounded) to have this approximated duration for grounded planes (on the rows that originally showed N/A). On the Durham flight centre website, they published the hourly rates for instructor fee, being CAD 65. So we multiplied the duration of these ground trainings and LF duals with the instructor fee and grouped the data by Month_Year and Aircrafts to show the monthly cost of instructors for each type of Aircraft.

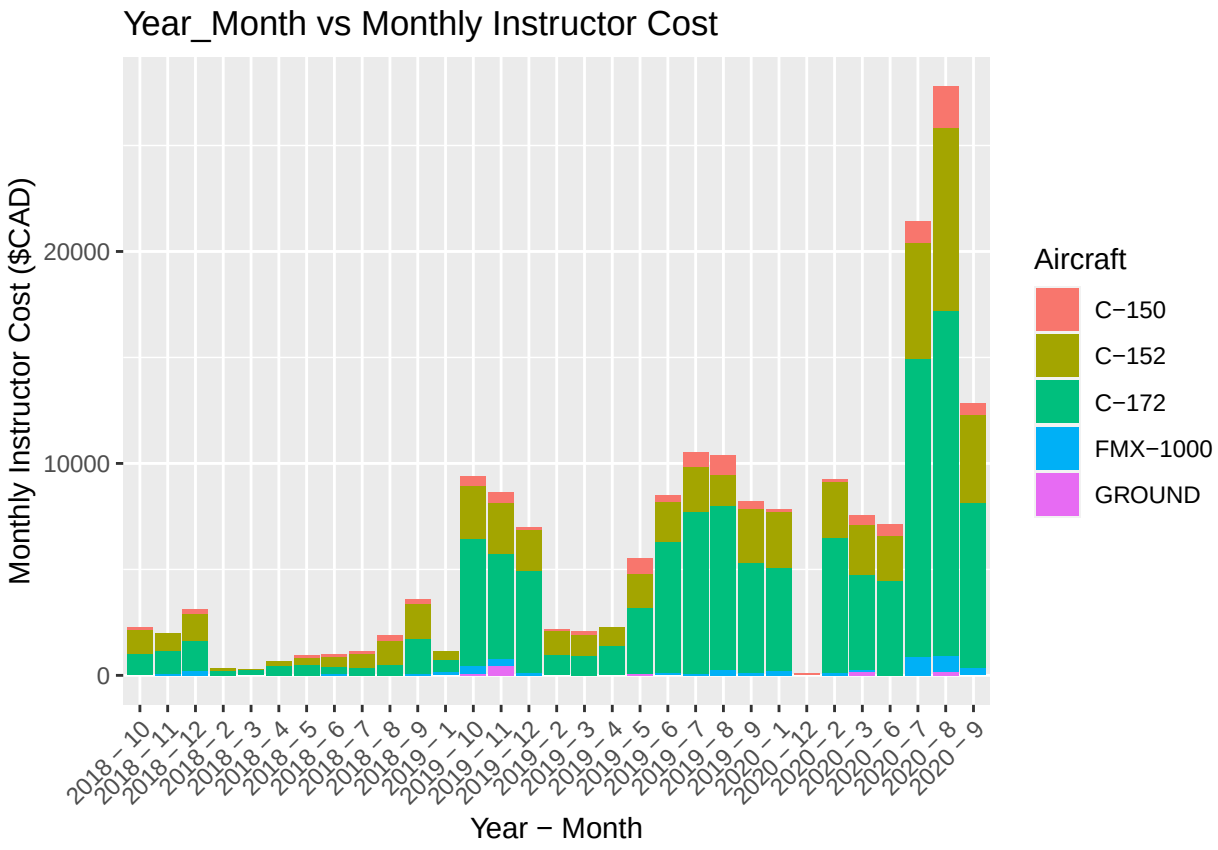
Here's the cost for instructors:

```
ground = my_cd%>%
  filter(Aircraft == "GROUND") %>%
  select(Exercises)

avg = my_cd %>%
  inner_join(ground, by = "Exercises") %>%
  drop_na() %>%
  group_by(Exercises) %>%
  summarise(Ground_Exercises_Duration = mean(Duration))

instructor_cost = my_cd %>%
  left_join(avg, by = "Exercises") %>%
   filter(Year >= 2018) %>%
  mutate(New_Duration = ifelse(Aircraft == "GROUND", Ground_Exercises_Duration, Duration)) %>%
    #drop_na() %>%
  mutate(Instructor_Cost = 65 * New_Duration) %>%
  mutate( Month_Year = str_c(Year, Month, sep = " - ")) %>%
  group_by(Month_Year, Aircraft) %>%
  summarize(instructor_cost_per_month = sum(Instructor_Cost))
```

```
## `summarise()` has grouped output by 'Month_Year'. You can override using the `.groups` argument.
```

```
instructor_cost %>%
  ggplot(aes(x = factor(Month_Year), y = instructor_cost_per_month, fill = Aircraft)) + geom_bar(stat="identity'
  ggtitle("Year_Month vs Monthly Instructor Cost") + ylab("Monthly Instructor Cost ($CAD)") + xlab("Year - Month
```

```
## Warning: Removed 3 rows containing missing values (position_stack).
```



Monthly Revenue

We didn't have a lot of information for computing the revenue earned from Ground School. We found that the Duration of the Ground School Sessions was always entered as NA so we didn't know how many hours the instructors were being paid to teach Ground School sessions. Due to COVID-19 pandemic, the flight centre also stopped offering in-person ground school sessions.

Since there is ambiguity around the revenue due to lack of information, we decided to compute a range of values for estimated revenue. We decided to compute a lower bound, and an upper bound using the data that we had available from the flight centre's dataset and the flight centre's website showing the different pricing options.

For our lower bound, we assumed that the flight centre wasn't earning money from Ground School sessions. We just considered the revenue from the students renting the flight centre's airplanes per hour.

For upper bound, we assumed each GROUND student did one package. We assumed just one package makes sense because we thought students would only need to do GROUND training when they are first learning to fly, eg they would not do GROUND training all throughout their pilot education career. When we looked at the data, we saw that students who did GROUND training had only one or two rows corresponding to GROUND training with dates very close to each other.

We assumed the revenue earned from GROUND training was equal to the amount of revenue earned from a package minus the amount of revenue earned from flying in the aircraft. We assumed that the revenue earned from students flying in aircraft on hourly rental and students flying in aircraft on package deals was the same. We compared the revenue earned from GROUND training for the 5-hour and 10-hour package, and took the higher of the two to get a higher upper bound. We also assumed a person who does 10 hour package probably won't do more ground school than a person who does 5-hour package if the end goal is to get in the air and fly. For students who chose the 10-hour package, we assumed the ground school cost the same for students who flew with C-152 or C-172. Then for the rest of the training of students who took the packages, we decided to check what plane they used and assume the revenue earned was proportional to the flight rental hourly rates.

During our revenue calculations, we rounded up the duration of hours flown to a whole number. For example, if a student flew 2.3 hours, we assumed that the flight centre charged them for 3 hours.

```
# cost for ground session

hourly_c150c152_rental_rate = 135
hourly_c172_rental_rate = 155

package_5hr_revenue = 1575
```

```
(ground_revenue5hr = package_5hr_revenue - 5*hourly_c150c152_rental_rate)
```

```
## [1] 900
```

```
package_10hr_revenue = 2575

(ground_revenue10hr = package_10hr_revenue - 10*hourly_c150c152_rental_rate)
```

```
## [1] 1225
```

```
my_cd%>%
  filter(Aircraft %in% c("C-152", "C-172", "C-150", "GROUND"))%>%
  mutate(Revenue_Session = ifelse(Aircraft =="C-172", hourly_c172_rental_rate*ceiling(Duration), ifelse(Aircraft
  filter(Year >= 2018) %>%
  mutate( Month_Year = str_c(Year, Month, sep = " - ")) %>%
  group_by(Month_Year, Aircraft) %>%
  summarize(total_revenue_per_month = sum(Revenue_Session)) -> upper_bound_month_revenue
```

```
## `summarise()` has grouped output by 'Month_Year'. You can override using the `.groups` argument.
```

```
(upper_bound_month_revenue %>% mutate(Lower_or_Upper = "Upper") -> upper_bound_month_revenue)
```

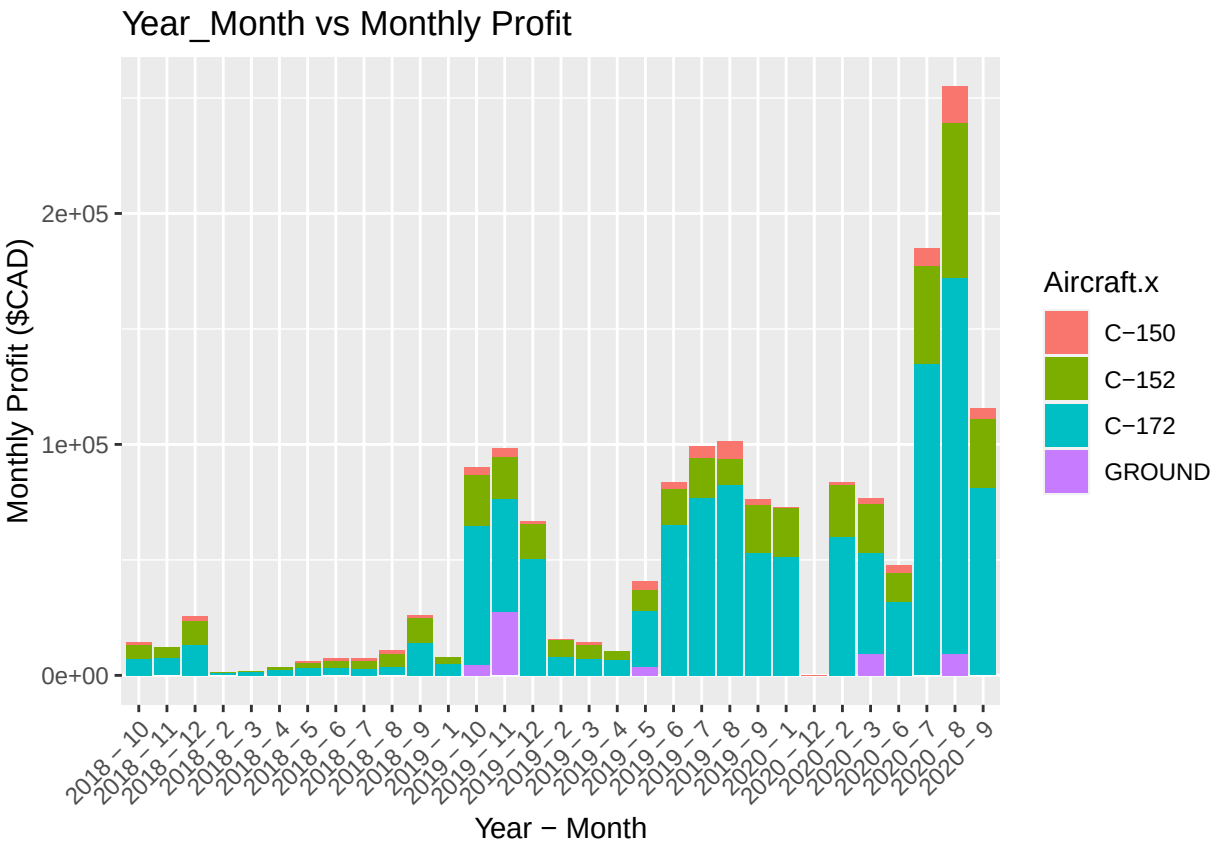
Monthly Profit

```
cost1 = instructor_cost %>%
  inner_join(upper_bound_month_revenue, by = "Month_Year")

(cost1)
```

```
cost2 = cost1 %>%
  inner_join(fuel, by = "Month_Year")
(cost2)
```

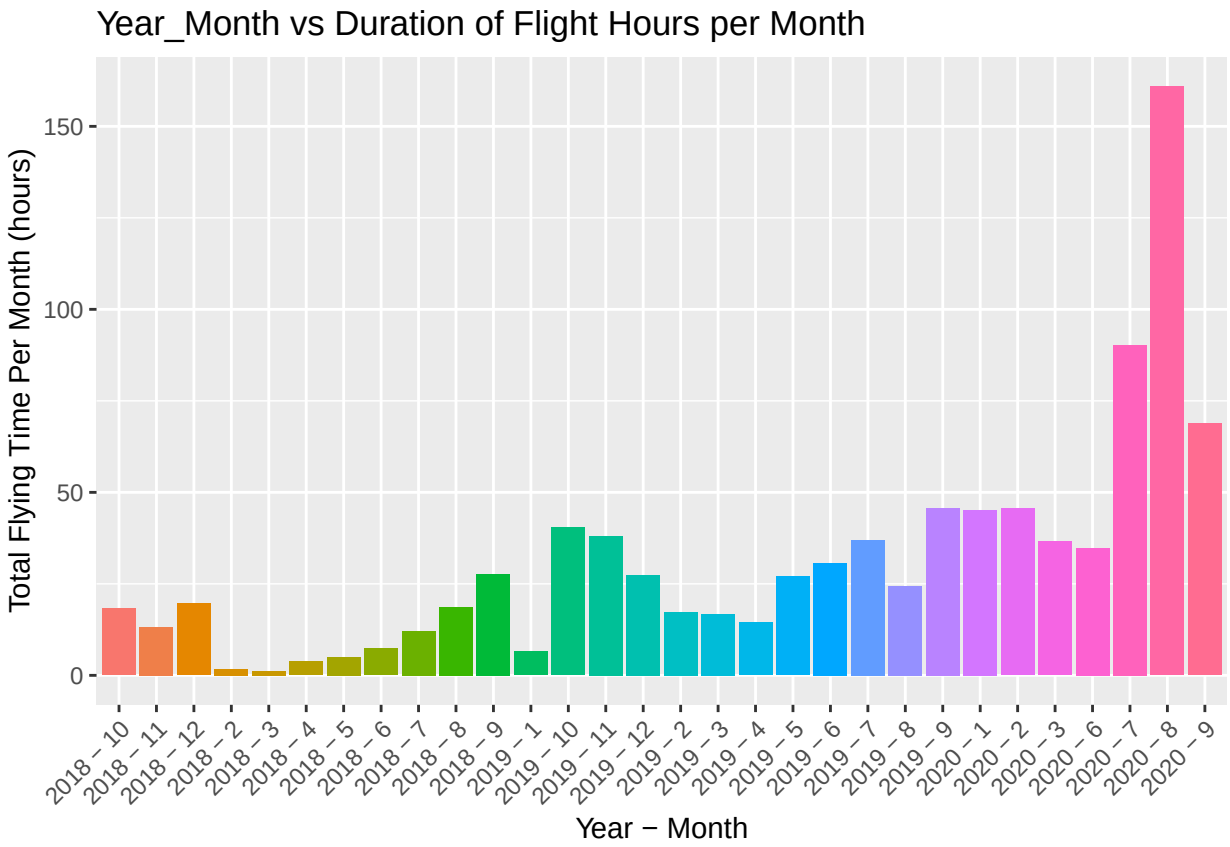
```
monthly_profit = cost2 %>%
  mutate(profit = total_revenue_per_month - instructor_cost_per_month - total_monthly_fuel) %>%
  filter(Aircraft.x == Aircraft.y) %>%
  group_by(Month_Year, Aircraft.x) %>%
  ggplot(aes(x = factor(Month_Year), y = profit, fill = Aircraft.x)) + geom_bar(stat="identity") + theme(axis.te
  ggtitle("Year_Month vs Monthly Profit") + ylab("Monthly Profit ($CAD)") + xlab("Year - Month")

(monthly_profit)
```

```
## Warning: Removed 42 rows containing missing values (position_stack).
```



We made some graphs graphing the months of the year against the total duration of training session for those months. We made separate graphs for each year from 2016-2020, and analyzed them. The code for the separate graphs can be seen in our R Studio Cloud Project R Code. Due to space limitations, we'll just show the graph with the different months of all years in range 2016-2020.

```
clean_data %>% select(Aircraft, Duration, Training_Type, Month, Year) %>%
  filter(Aircraft == "C-152") %>%
  filter(Year>= 2018) %>%
  mutate( Month_Year = str_c(Year, Month, sep = " - ")) %>%
  group_by(Month_Year) %>%
  summarize(total_time_per_month = sum(Duration)) %>%
  ggplot(aes(x = factor(Month_Year), y = total_time_per_month, fill = Month_Year)) + geom_bar(stat="identity") +
```



Note that 2020 only has data for 7 months because of the lockdown from the pandemic. Also, 2018 has no data for January. That is why we thought it would be unfair to find the aggregate of duration for each month across all years (eg. add up duration for 2016 January, 2017 January, etc), because the data would be slightly skewed. So, we thought about representing the information as 5 separate graphs from each year.

We notice that the training duration generally increases during the months of July, August, September, October and November, usually peaking around August. This may be due to the fact that those months tend to be warmer, and invite clearer skies which are an important consideration when flying. We think it'd be interesting to compare the training duration with more detailed weather data on temperatures and precipitation.

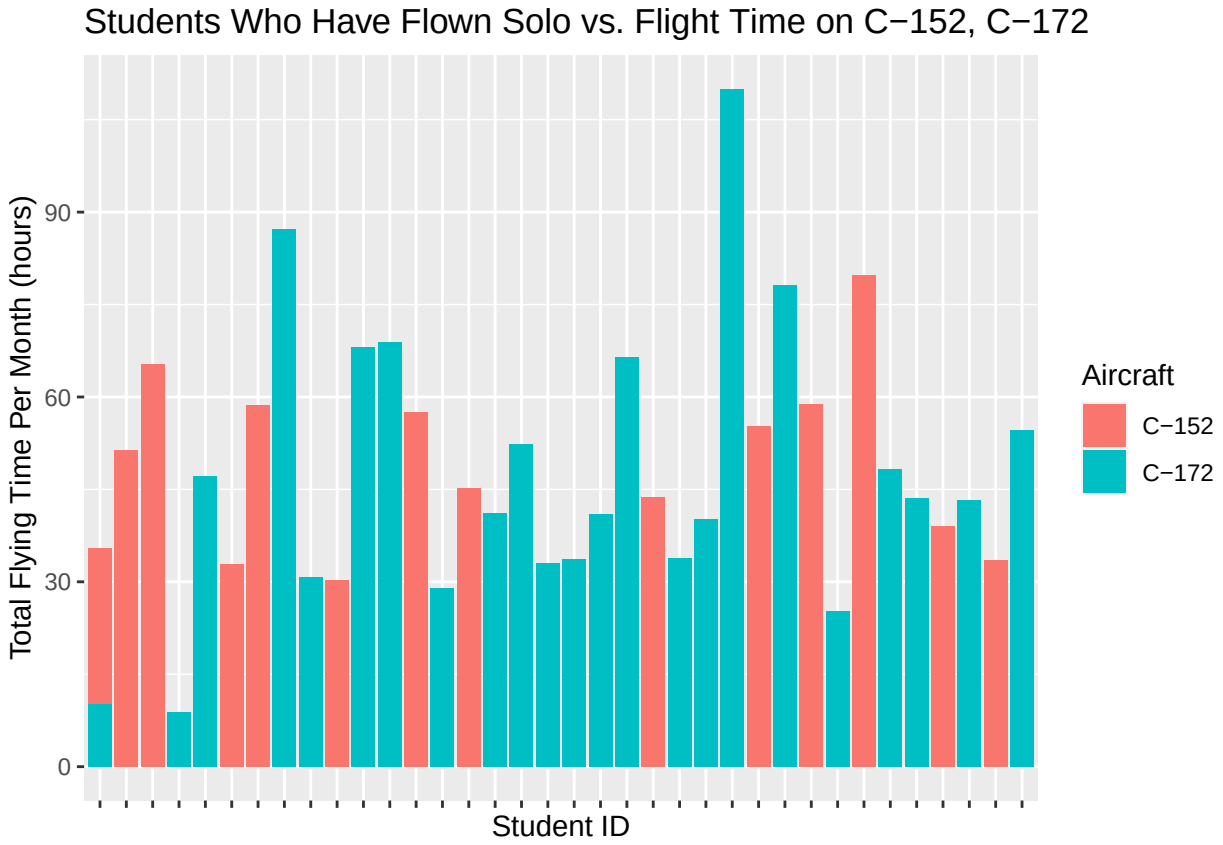
Moreover, we see that from 2016 to 2017 the hours rarely surpass 10, and increase to peak around 60 in 2018. In 2019 it reaches up to 180, but oddly enough 2020 seems to have the highest peak of above 400 flight hours. So even during a pandemic, the flight center was able to maintain growth and bring in more customers. This steady growth in flight hours, which is synonymous to the demand for training, has been increasing over the years, as shown in our graphs. But to help tackle the slump in demand that the flight center sees every winter/spring, they could decrease prices for the training courses (of course while staying above the fixed + variable costs), which in turn could increase demand and possibly increase growth over the years.

We made two bar graphs which can be seen on the next page. The first shows the students who were able to fly solo at least once, their aircraft choice, and the total time flown on those aircrafts. The second shows the same information for students who were NOT able to fly solo at least once. These graphs show us the "successful" student's behaviour and their aircraft of choice.

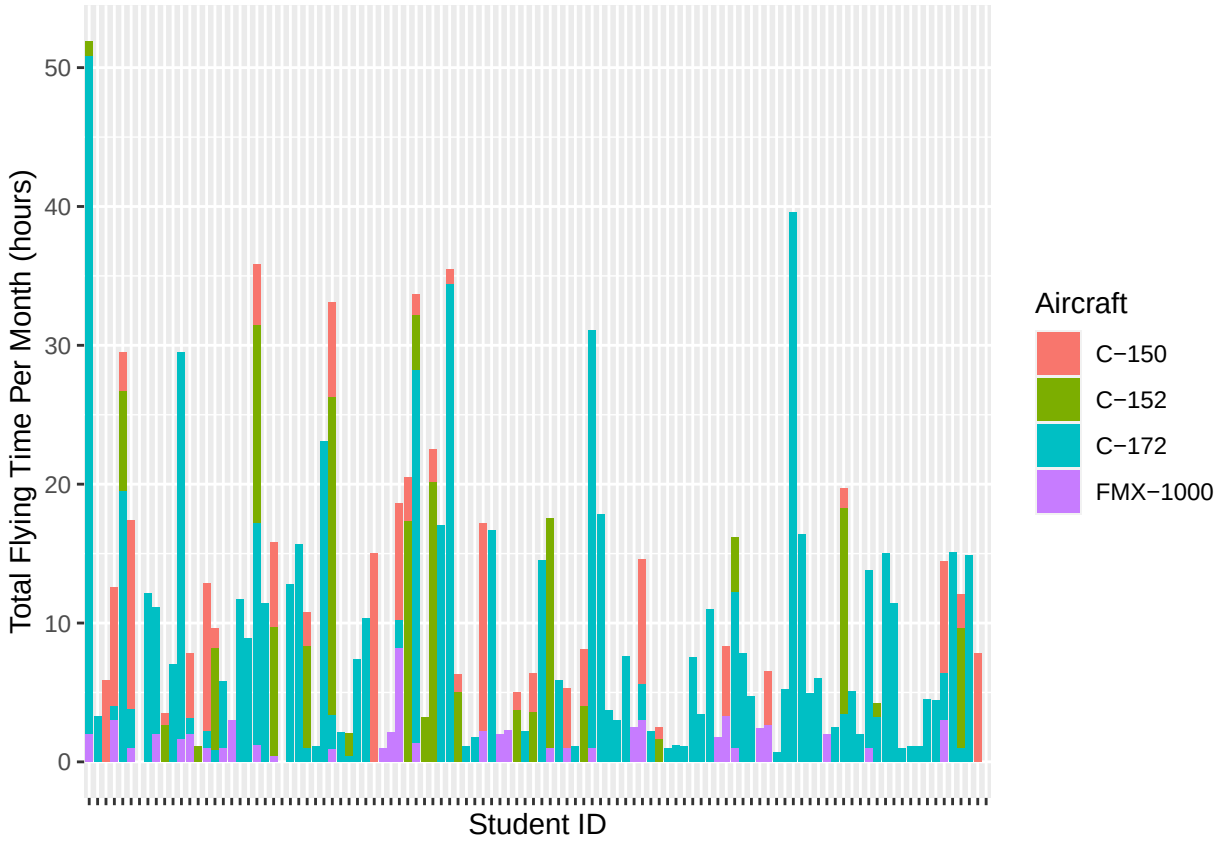
When we compared the graphs for students who have flown solo and those who did not, we saw that those who have flown solo prefer sticking to one aircraft and rarely switch between the two. We believe that that students who stick to their preferred aircraft end up becoming more comfortable with it and mastering it at a faster pace. Meanwhile students who tend to switch between 2 or 3 aircrafts might struggle with adjusting to different aircraft, and end up taking a longer time to perfect their craft.

Our goal is to try to find a benchmark of exercises that all "successful" students do that might show types of exercises that lead to greater student success. We are still working on a graph that shows the exercises on the x axis and their frequency on the y axis for successful students.

```
## `summarise()` has grouped output by 'Student_ID'. You can override using the `.groups` argument.
```




```
## `summarise()` has grouped output by 'Student_ID'. You can override using the `.groups` argument.
```



Summary