

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_Draft_Submitted.Rmd file.

Introduction

We are interested in efficient pricing of flight training courses and factors associated with student success. 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.

We define students as ‘successful’ if they have flown solo at least once during the course of their training.

To analyze the costs of Durham Flight Centre, we would like to estimate:

1. How much money the training centre has spent on fuel for different aircraft models over the past few years.
 - From Durham Flight Centre’s website, we know the hourly rental rate for each plane. 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 demand for flights varies 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 training with different aircrafts 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 visualizing our findings.
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. It may not be suitable for generalizing trends to longer periods of time. We decided there was more reliable data for 2018-2020 than the other years so we focused our analysis on these years.

We used the Durham Flight Centre's website: <https://durhamflightcentre.com/> to get the hourly aircraft rates, flight training package rates, and instructor rates.

We gathered data on the aircrafts' ownership costs 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.

We used the Tillsonburg airport website (<https://www.tillsonburg.ca/en/live-and-play/Fuel.aspx#>) to find the cost of Avgas (the same type of gas the training centre's aircraft use). In April 2021, it said Avgas was \$1.69/L. We believed this was an accurate estimate for fuel pricing in\$/L, since Durham Flight Centre is in the same province as Tillsonburg airport so region and market conditions affecting fuel prices should be similar. 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.

See Appendix for a more detailed discussion of potential biases in the data obtained from these sources.

Importing and Formatting the Data in R

Brief explanation of changes made to data We made changes to the data, with the goal of amending input errors in the data, and making our analyses and visualizations more accurate. For example, we removed entries where the Session_ID's were repeated, which would have gone against the intention of having them be unique identifiers. We also decided to remove rows in which Aircraft was blank, since much of our analyses relied on this variable. Finally,

we modified any entries that were obvious typos, like “C152” and “111” which were meant to be “C-152” and “11” respectively, as well as many individual typo cases.

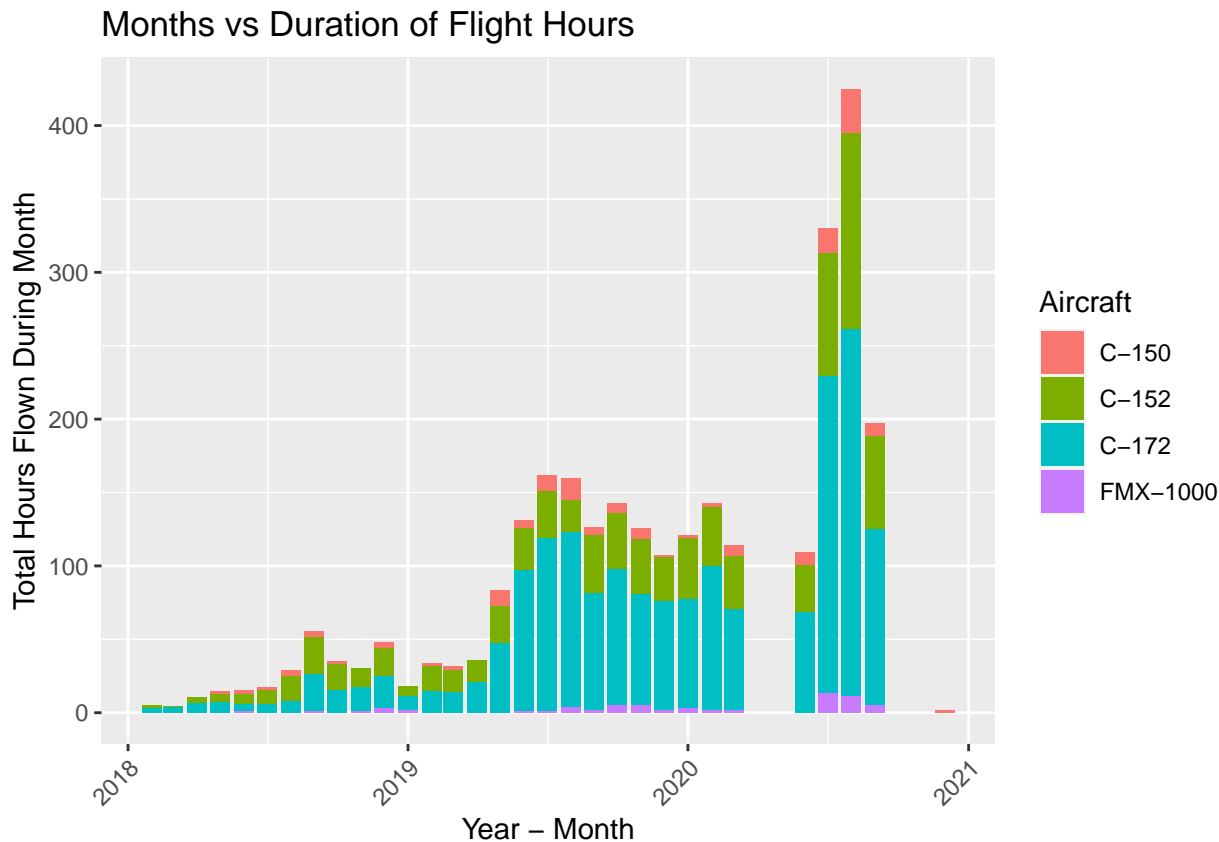
Analysis

We wish to estimate the operating and maintenance cost the training centre has spent on each type of aircraft over time. Thus we are interested in the total amount of time students spent training with different types of aircrafts.

Training Duration

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 2018-2020, and analyzed them.

`## `summarise()` has grouped output by 'Aircraft', 'Year_Month'. You can override using the `.groups``



Note that data for certain months is missing. 2020 only has data for 7 months because of the lockdown from the pandemic and 2018 has no data for January. So we thought that finding the aggregate duration for each month across all years (eg. add up duration for 2018 January, 2019 January, etc) would yield skewed results. So, we considered representing the information as separate graphs from each year.

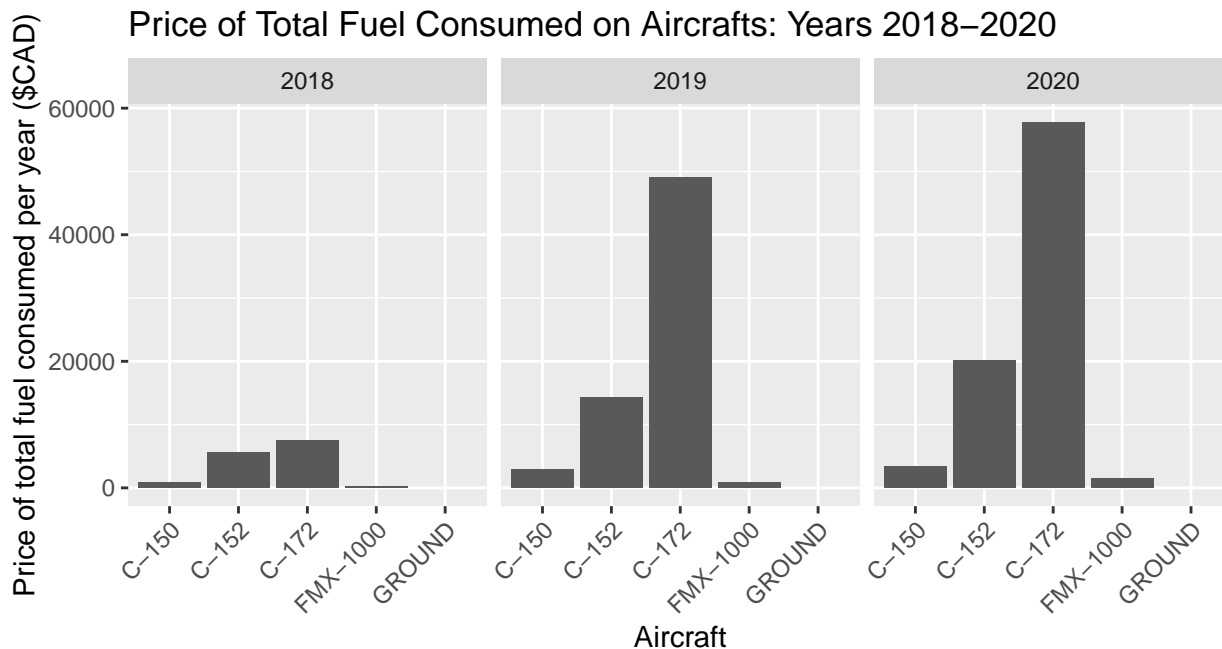
We notice that the training duration generally increases during the months of June, July, August, September, and October, usually peaking around August. This is most likely due to the fact that those months tend to be warmer, and invite clearer skies which are an important consideration when flying.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

Moreover, we see that the hours peak around 60 in 2018, around 180 in 2019, and oddly enough, 2020 seems to have the highest peak of over 400 flight hours. So even during a pandemic, the flight center was able to maintain growth and bring in more students to log flight hours. 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 or spring, they could consider decreasing prices for the training courses or give

special package rates (of course while staying above the fixed + variable costs), which in turn could increase demand and possibly increase growth over the years.

Plane Fuel Consumption

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.



We found that over the years 2018-2020, the training centre's Cessna-172 aircrafts consumed around \$677.40 worth of fuel (14900.76 gallons) in total. The Cessna-152 aircrafts consumed around \$238.13 worth of fuel (5238.07 gallons) in total. The Cessna-150 aircrafts consumed around \$42.59 worth of fuel (936.88 gallons) in total.

C-172 C-152 C-150 \$677.40 \$238.13 \$42.59 14900.76 gallons 5238.07 gallons 936.88 gallons

This is consistent with the fact that over the span of those 3 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. One can infer that the C-172 is a more expensive plane to operate, and that the flight centre could consider offering more services with the C-152 and C-150, thereby decreasing the hourly rates for students and increasing demand.

When we looked at the fuel consumption per year, we found 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 while the C-152 consumed around 2009.95 gallons of fuel. In January-March, June-September 2020, the C-172 consumed approximately 8106.84 gallons 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 cost of fuel consumed per hour remained constant over 2018-2020 regardless of aircraft model, age and condition, nature of exercise, market conditions for fuel pricing (especially given COVID-19 pandemic), and other factors.

Monthly Fuel Costs

We wanted to compute the profit of the centre so we wanted to estimate various costs, one of those being fuel costs. We first computed the cost of fuel in terms of gallons per hour and then multiplied it by the cost of 1 gallon of Avgas.

We found the cost of Avgas (the same type of gas the training centre's aircraft use) in \$CAD/L from the Town of Tillsonburg's website (<https://www.tillsonburg.ca/en/live-and-play/Fuel.aspx#>) under the sections Airport > Services > Fuel. It said Avgas Fuel Prices were \$1.69 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 province as this airport so region and market conditions affecting fuel prices should be similar.

We used the conversion factor 1 Litre = 0.2199692 gallons. So \$1.69/litre is approximately \$7.68/gallon. Multiplying gallons of Avgas consumed for each session by the price of gas per gallon gave us the price of gas consumed for that session. We made a column that multiplies the number of gallons consumed by its corresponding price depending on plane used. We put a corresponding graph in the appendix for the sake of space.

Monthly Instructor Costs

Another cost incurred by the training centre is that of paying instructors. 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 "GROUND" Aircraft type, we noticed that the duration was entered as NA (not given). Initially we thought of completely ignoring the Grounded Aircrafts and the instructor fee for these sessions, but ground training makes up a significant cost of the packages that students pay 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 created 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, the hourly rates for instructor fee are listed as \$65 (CAD). 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. We put a corresponding graph in the appendix to save space.

The graph of instructor cost (from the Appendix) is in line with the graph of fuel cost. This makes sense because the cost is synonymous with the duration the planes are flown for.

Monthly Revenue

We wanted to estimate the revenue with the ultimate goal of estimating the flight centre's profits. 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 estimate revenue using the data available from the flight centre's provided dataset and the flight centre's website showing the different pricing options. We assumed these prices were constant throughout 2018-2020.

We assumed each GROUND student purchased one package because we thought each student would only do preparatory GROUND training when they are first learning to fly, and not 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.

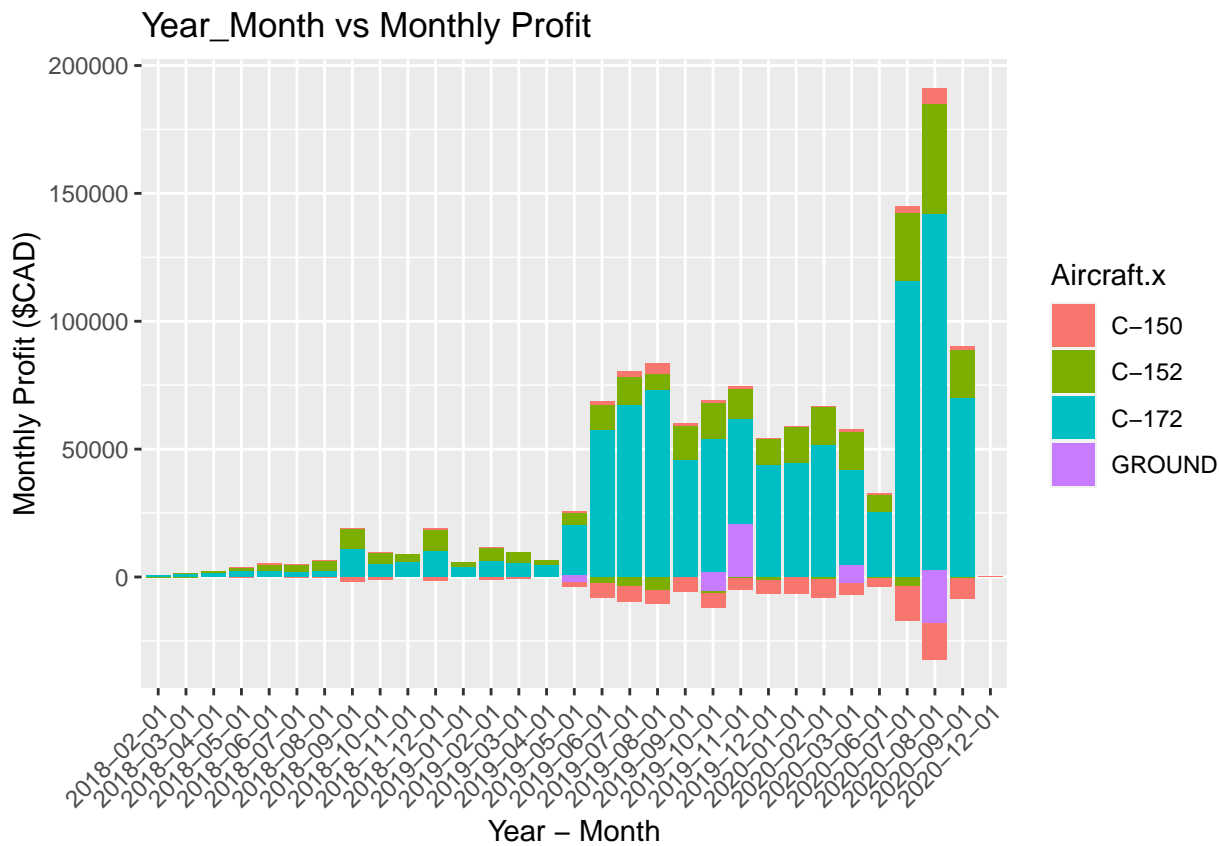
We assumed the revenue earned from GROUND training was equal to the amount of revenue earned from a package minus the amount of revenue earned from flying in the aircraft. We assumed that the revenue earned from students flying in aircraft on hourly rental and students flying in aircraft on package deals 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 a 10 hour package probably won't do more ground school than a person who does a 5-hour package if the end goal is to fly aircraft solo. 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 assumed 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.

Monthly Profit

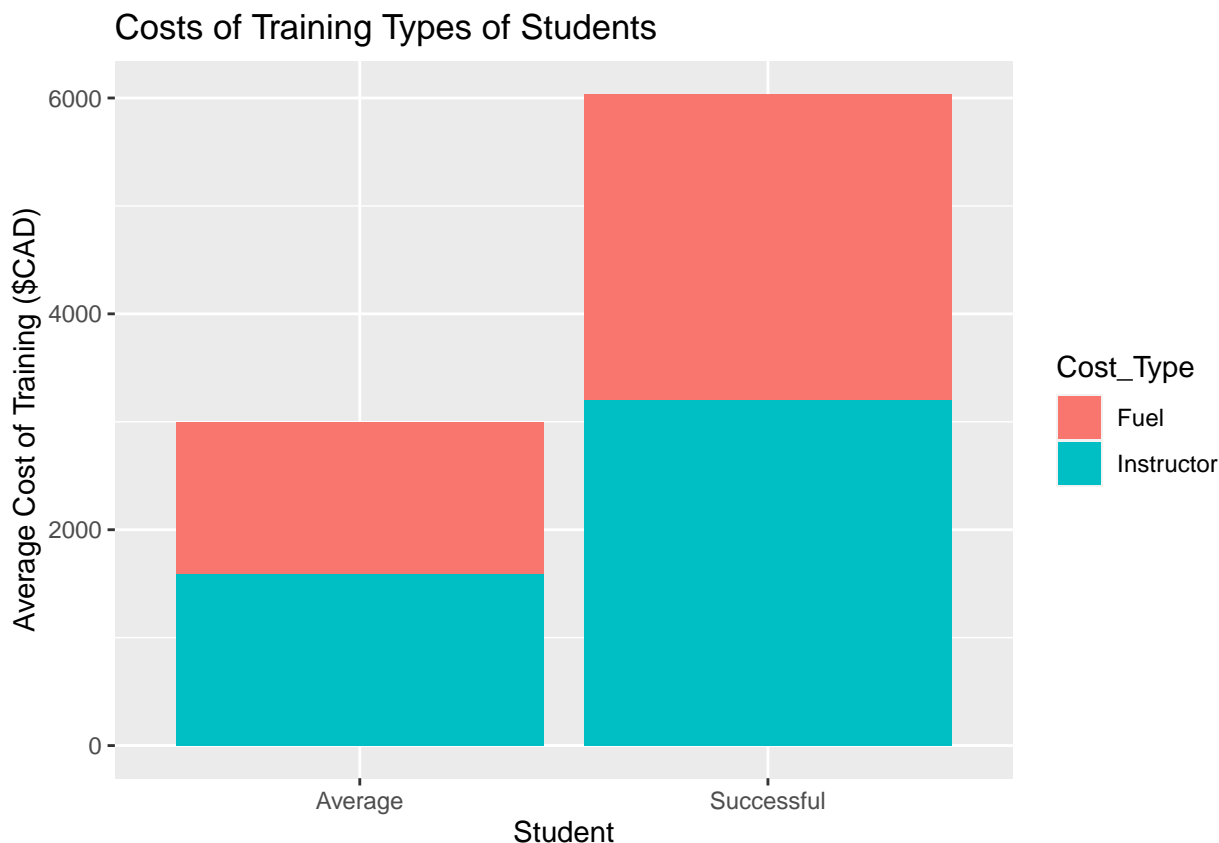
The two main costs we looked at were the fuel costs and the instruction costs. Computing these costs seemed more accessible given the flight centre's dataset and website pricing information. Operating a flight centre and offering flight lessons would certainly incur other costs, however we didn't have enough reliable information for these costs. If in the future we acquire more accurate information for the costs Durham Flight Centre faces, we would certainly like to update our profit estimates. We computed the profit by subtracting the fuel and instruction costs from the revenue, and graphed it.



It seems like most of the flight centre's profit came from renting the C-172 planes to students during the months of July, August, and September, even though Instruction Costs and Fuel Costs were also higher during these times. This is likely because students spent more time flying C-172 planes and the C-172 planes were more expensive to rent.

Cost of Training Average Student

We wanted to estimate the fuel cost and instruction cost of the average student at the flight centre. We were also curious, out of the students who successfully flew solo, what was the average amount of money the flight centre spent on their fuel and instruction?



We estimated that the flight centre spends \$1406.75 on fuel and \$1590.28 on instruction for the average student in their program. The flight centre spends \$2843.00 on fuel and \$3197.82 on instruction for the average successful student in their program. The average successful student requires nearly 2 times the fuel and instructor costs than the average student. This makes sense, since the average successful student likely spends more time training than the average student. The average successful student will likely require more fuel for more aircraft practice, and more Instruction time.

Factors Common To Successful Students

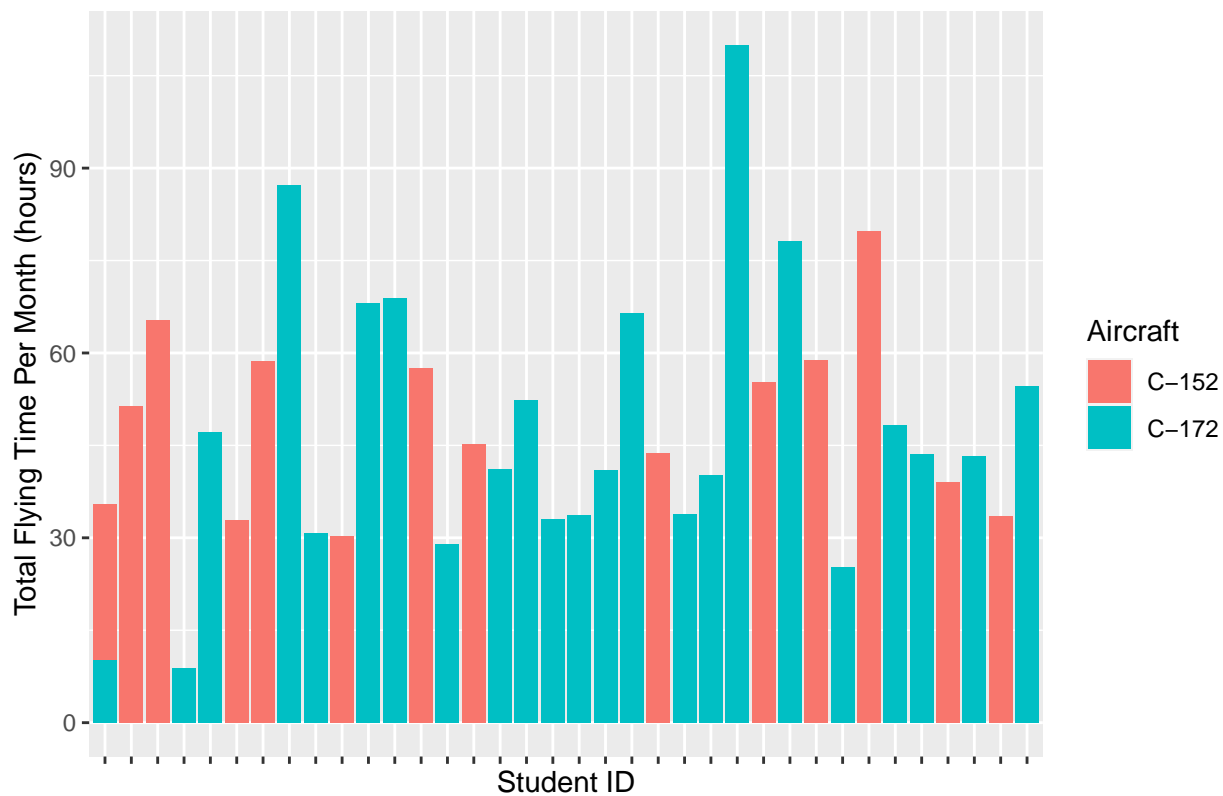
We made two bar graphs. The first shows the students who were able to fly solo at least once, their aircraft choice, and the total time flown on those aircrafts. The second shows the same information for students who were NOT able to fly solo at least once. These graphs show us the “successful” student’s 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 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.

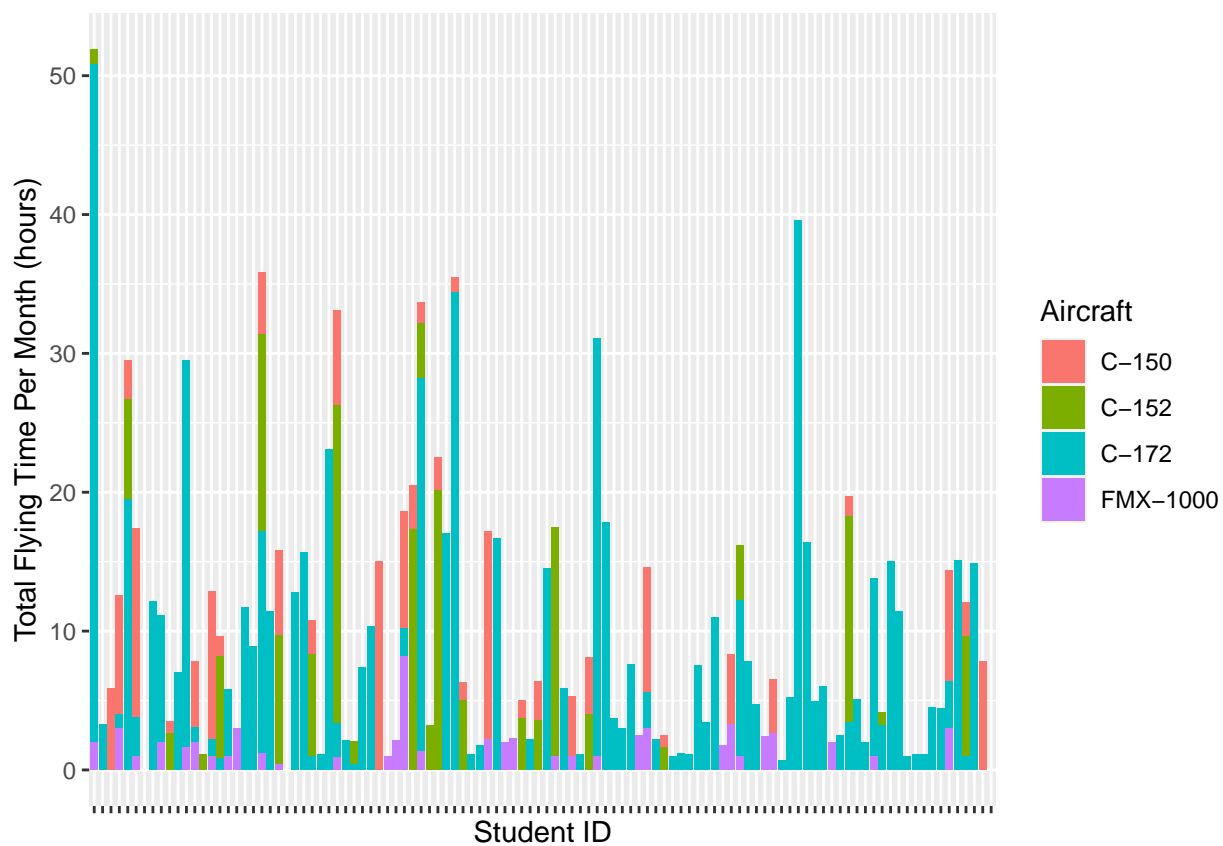
Our goal is to try to find a benchmark of exercises that all “successful” students do that might show types of exercises that lead to greater student success. We are still working on a graph that shows the exercises on the x axis and their frequency on the y axis for successful students.

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

Students Who Have Flown Solo vs. Flight Time on C-152, C-172

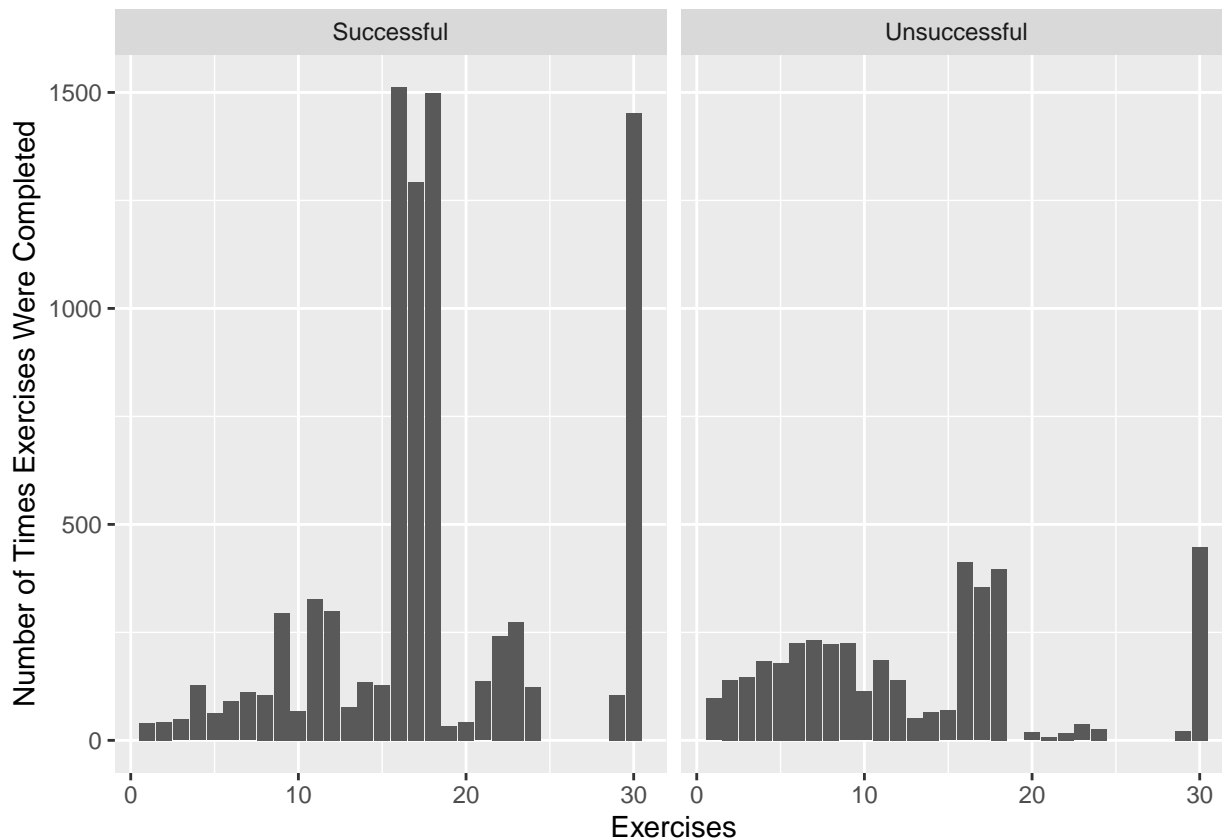


`summarise()` has grouped output by 'Student_ID'. You can override using the `.groups` argument.



Successful Students Exercises

We were interested in what successful students do that sets them apart from unsuccessful students. On the left, we have a plot showing the frequency of Exercises completed by all successful students combined. On the right we have a plot showing the frequency of Exercises completed by all unsuccessful students combined.



We see that the successful students and unsuccessful students completed Exercises 16, 17, 18 and 30 the most often. Exercises 16, 17, 18, and 30 respectively correspond to Take-off, Circuit, Approach and Landing, and Radio Communications. All the successful students combined completed Exercises 16, 17, 18, and 30 1512 times, 1293 times, 1499 times, and 1453 times respectively. All the unsuccessful students combined completed Exercises 16, 17, 18, and 30 412 times, 354 times, 396 times, and 448 times respectively. Successful students practiced these exercises over three times as much as the unsuccessful students. This suggests that Take-off, Circuit, Approach and Landing, and Radio Communications are key components of pilot training. We believe that the number of times a student practices these exercises may be correlated with ability to fly solo.

We also noticed that successful students completed Exercise 29 (Emergency Procedures) nearly 5 times as often as unsuccessful students. Successful students also completed Exercises 21-24 (Precautionary Landing, Forced Landing, Pilot Navigation, Instrument Flying) more often than unsuccessful students. Successful students likely have more practice doing different types of landing and navigation than unsuccessful students.

Progression of Training

We were also interested in finding out what the progression of a successful student's training looks like. Here is a graph showing the frequency of exercises successful students complete as their training progresses.

```
my_cd %>% group_by(Student_ID) %>%  
  mutate(Day_Since_First_Session = rank(Date, ties.method = "first")) -> my_cd  
  
my_cd %>% group_by(Student_ID) %>%  
  summarize(Training_Length = max(Day_Since_First_Session)) %>%
```

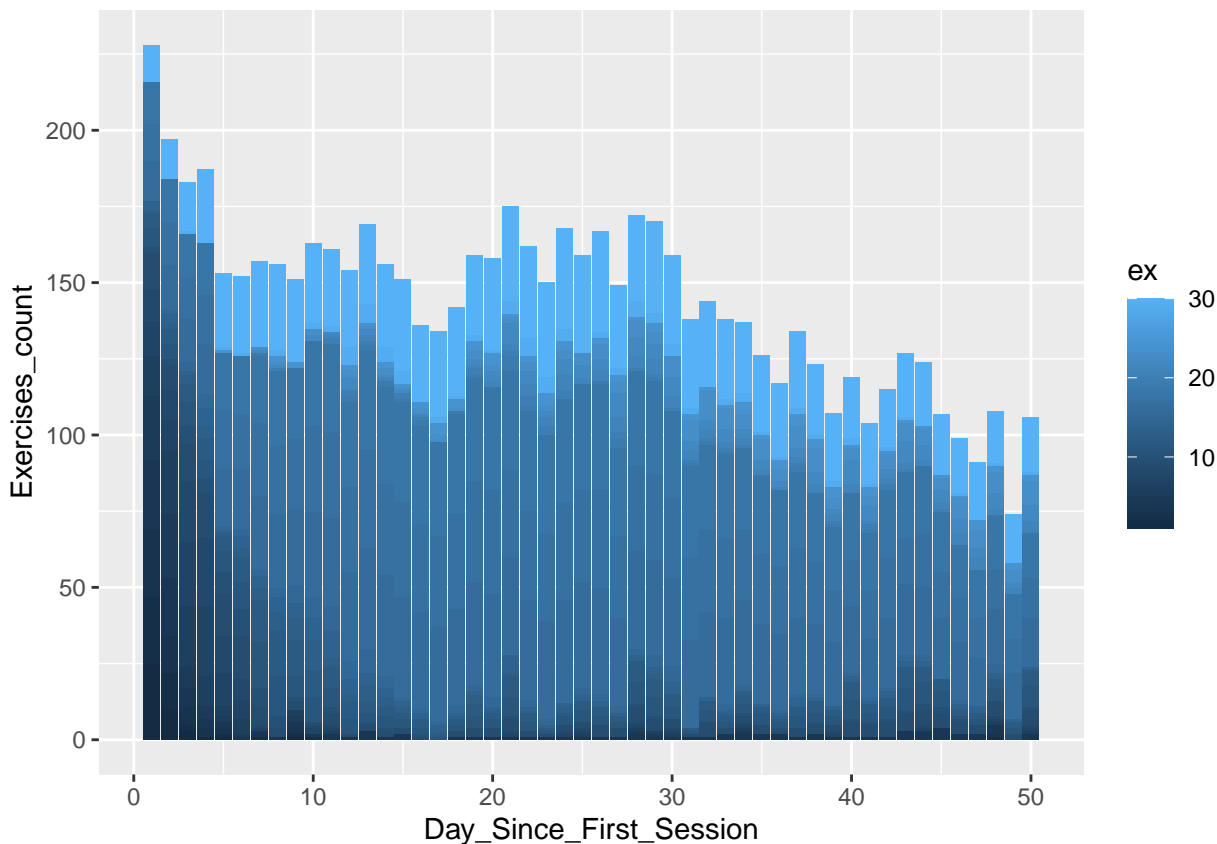
```

inner_join(successfulstudents, by = "Student_ID") %>%
  summarize(mean_training_length = mean(Training_Length))

my_cd2 %>% left_join(my_cd, by="Session_ID") %>%
  mutate(ex = as.numeric(Exercises.x)) %>%
  inner_join(successfulstudents, by = "Student_ID.x") %>%
  group_by(Day_Since_First_Session, ex) %>%
  summarize(Exercises_count = n()) -> training_progression

## `summarise()` has grouped output by 'Day_Since_First_Session'. You can override using the `.groups`
training_progression %>%
  filter(Day_Since_First_Session <= 50) %>%
  ggplot(aes(x=Day_Since_First_Session, y=Exercises_count, fill = ex)) + geom_bar(stat="identity")

```



We see the progression of training, as successful students mostly focus on familiarisation and preparation for flights (through exercises 1-5) for the first couple of days, and shift to slightly more advanced training with maneuvering and controls (through exercises 6-10) in the next few days. From Day 11 onwards, we observe that exercises 16, 17, 18, 30 are practiced often and continue to be practised throughout the training. This is indicated by the medium blue color that dominates a large area of the graph.

On Days 1-2 of training, students mostly focus on Exercises 1-5, which involve Familiarization, Preparation for Flight, Ancillary Controls, Taxiing, and Attitudes & Movements. Students spend some time practicing Exercises 6-10, which involve Straight-and-Level Flight, Climbing, Descending, Turn, and Maximum Range & Endurance.

After the first two days, students hardly practice Exercises 1-5.

On Days 3-4, students spend more time practicing Exercises 6-10. Students occasionally practice Exercises 11-15, which involve Slow Flight, Stall, Spin, Spiral and Sideslip. Exercises 19-29 are rarely practiced.

After the first four days, Exercises 6-10 are rarely practiced.

On Days 5-10 , students practice Exercises 11-15 more often.

On Days 11 onward, students practice Exercises 11-15 less often than they did during days 5-10. At this stage, the focus seems to be on Exercises 16, 17, 18, 30. Some may occasionally practice Exercises 20-24, and 29, which involve Illusions Created by Drift, Precautionary Landing, Forced Landing, Pilot Navigation, Instrument Flying, and Emergency Procedures.

Need to write about how this data can help the centre design training routines to cater towards students who want to be successful.

Summary

Overall, we found that the flight centre's students log more flight hours during late summer to early fall months, and prefer to do so on the C-172. The centre earns more revenue and profit during these months on C-172 aircrafts, but also spends more on fuel and instruction. We computed fuel and instruction costs of training the average student and compared them with the same costs for the average "successful" student. We explored common traits of successful students and determined that successful ones seemed to stick to one aircraft consistently. We hope our findings are interesting and relevant to Durham Flight Centre.

References

We used the official R Documentation from RStudio help section to learn how to process certain types of data. <https://e2a207d9ab5c4c3d9007aa1ad8e21ccd.app.rstudio.cloud/help/doc/html/packages.html> We saw how to use "ifelse" from another course at UTSC and consulted R Documentation for more help. We learned how to change margins, format R Markdown files from this source: <http://www.stat.cmu.edu/~cshalizi/rmarkdown/>

Appendix

More Information About Data Sources

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.

Data Processing Methods and Details

Data processing explanation: The goal of this script was to make changes to the main data object that we will all be working with. We call the object `my_cd`. The first step is to remove rows with duplicated "Session_ID", since these are meant to be unique. This was done using the `distinct` method. Next, we changed the exercise column from being string into a list of character vectors using `str_split`. This was done so that we could perform counting and sorting operations, as well as catching typos, easier.

Next, we used standard subsetting to change typos “C152” in Aircraft, and “111” in Month, into “C-152” and “11”. We were able to determine that “111” was meant to be “11”, since the original data on the spreadsheet was ordered; what appears to have happened was a “1” was taken from the previous entry and added to this one. We ensured all Cessna-152 planes were entered as “C-152” for consistency. Also, the choice to use standard subsetting was appropriate since the change would not be misapplied to other data; the search condition only applies to what we want. In the next step however, we needed to use more specific criteria, so we used the “replace” function within “mutate” to fix the affected Date columns, and the “1” in month that was meant to be “11”.

Next, we removed all entries in which Aircraft wasn’t given, since much of our analyses rely on this information. This was done by filtering the complement of rows for which `is.na` evaluated true. Lastly, the bulk of errors and typos seemed to have come from the exercises column. This step was particularly tricky, since our instinctive approaches did not work. Beginning with empty string which appeared as `c(“ ”)`, and mostly arose from double commas in the original data. Initially, we tried doing: `my_cd %>% filter(str_detect(Exercises, “^$”))` to find instances of empty string, and then replacing them using something like `x[!(x %in% “ ”)]`. In the end, the method that worked was combining the `lapply` function, which applies a function to each element in a list, along with the `stri_remove_empty` function that removes empty string from a character vector.

After that we used conditional subsetting again to replace individual occurrences of errors. `str_detect` generates a logical vector for each error, and `which` returns the indices of rows that evaluated TRUE. We then used those indices to manually replace each character vector in the list. Finally, the row with “40” was removed, simply because it didn’t resemble anything valid and it was the only one.

Airplane Costs Data 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.

Figure 1:

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

Monthly Fuel Costs for Different Months in Years 2018–2020

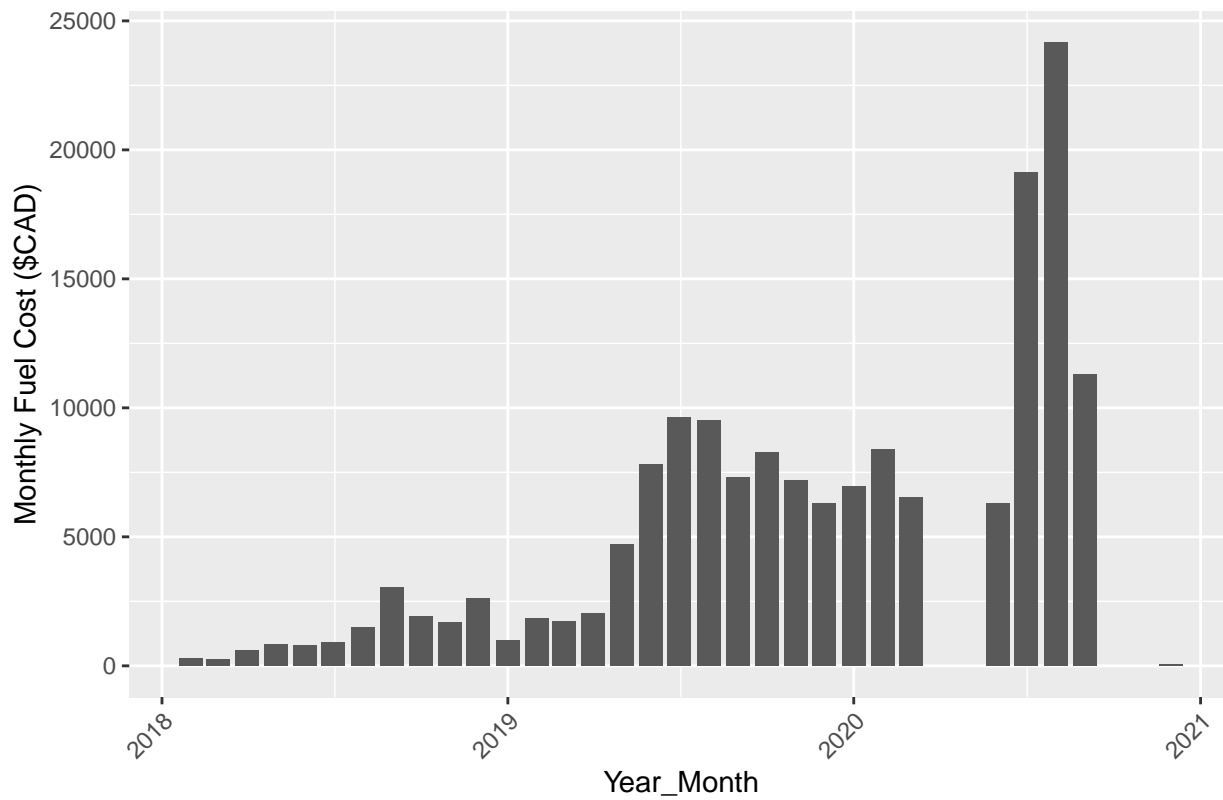


Figure 2:

Year_Month vs Monthly Instructor Cost

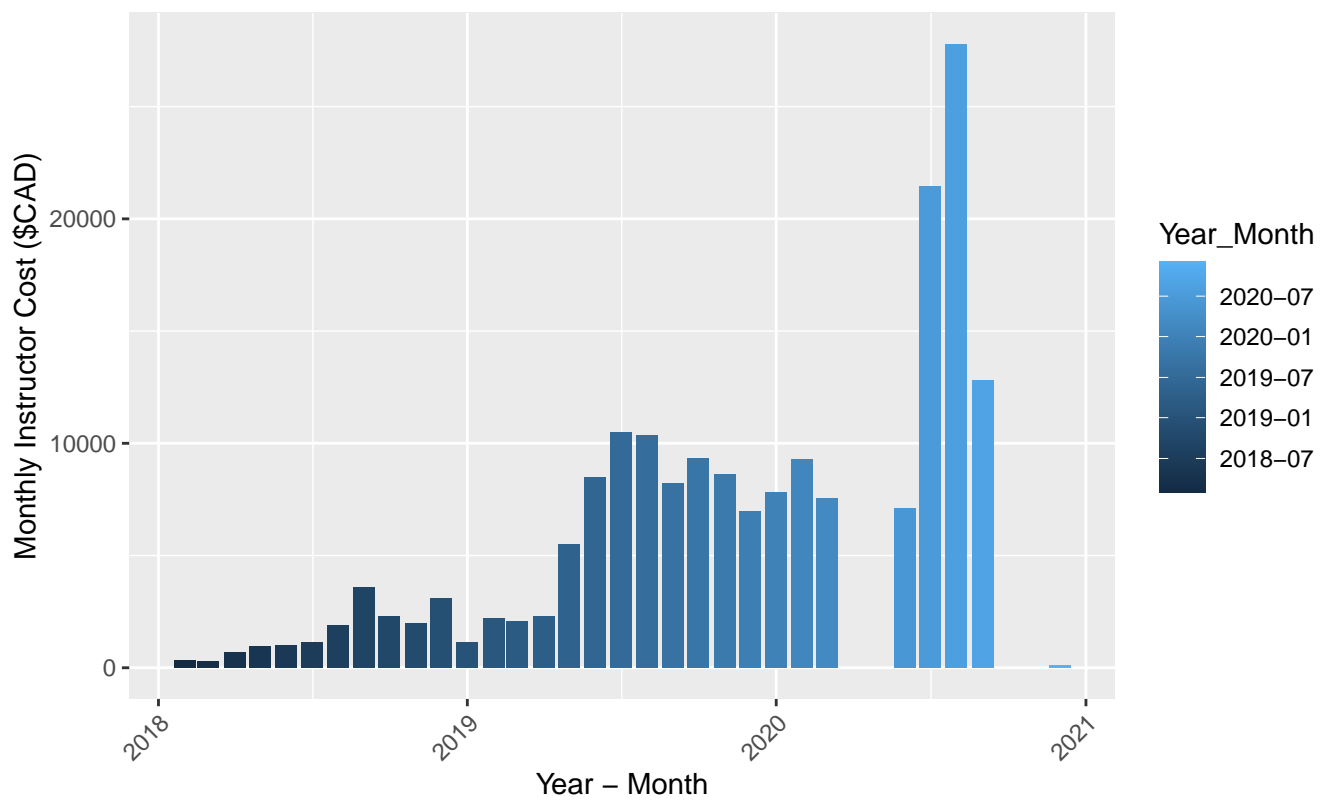


Figure 3:

Year_Month vs Monthly Revenue

