# STAA57 W21 - Project Proposal

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

https://rstudio.cloud/spaces/115177/project/2202992

### Analysis Plan

*Specify the questions you will address. Describe the general topic of your investigation, and state specific questions you will address. Include any relevant information.*

General topic: analysis trends in average aircraft operating costs and the total aircraft operating costs for DFC

We will be examining the average aircraft operating costs and the total aircraft operating costs for DFC. The factors that influence the cost are fuel, weather and COVID-19. Fuel and weather affect the average aircraft operating costs while COVID-19 affects total demand and indirectly affects fuel thus the average aircraft operating costs as well. Due to the multidimensional impact of COVID-19 affecting fuel and total demand, there is going to be a comparison between the year 2020 and the average of the previous years. In this paper, demand is defined as the number of sessions.

Questions:
1. What are the trends in demand over different time units (months/season/year)?
2. What are the trends in jet fuel costs over different time units (months/season)?
3. What kind of impact does weather have on estimated aircraft operating cost?
4. Considering the previous questions, what are the estimated average aircraft operating over different time units (months/season)?
5. What are the total aircraft operating costs over month/season/year?

*Specify your data analysis plan. Describe (in words) how you will address these questions using data.*

Data analysis plan per question:

1. a. Yearly demand trend: aggregate the duration data yearly and observe the trend in the demand.
   b. Average demand for each month/season using all 2016-2020 data: see the trend over the months/seasons as an indication of total costs for each month/season.
   c. Average demand for each month/season using 2016-2019 data vs 2020 data (COVID-19 comparison): comparison of trends over the months/seasons vs COVID-19 as an indication of the impact COVID-19 on total costs.
2. a. Average fuel costs for each month/season using all 2016-2020 data: see the trend over the months/seasons as an indication of fuel costs fluctuations for each month/season.
   b. Average fuel costs for each month/season using 2016-2019 data vs 2020 data (COVID-19 comparison): comparison of trends over the months/seasons vs COVID-19 as an indication of the impact COVID-19 on fuel costs.
3. a. Understand how percipitation affects the average aircraft operating cost per hour (Cost on y-axis; precipitation level on x-axis).

   b. Understand how the average percipitation fluctuates on a monthly/seasonal level (Precipitation on y-axis and x-axis will months/seasons).

c. Combine a and b to estimate average aircraft operating cost per hour due to pericipitation level on a monthly/seasonal basis (Cost on y-axis and x-axis months/seasons).

4. combining calculations made in 2, 3 and an estimated satitic aircraft repair operating cost found online

   a. the total average aircraft operating cost per hour considering fuel and weather fluctuations on a monthly/seasonal basis using all 2016-2020 data.
   b. the total average aircraft operating cost per hour considering fuel and weather fluctuations on a monthly/seasonal using 2016-2019 data vs 2020 data (COVID-19 comparison).

5. combining calculations made in 4 and 1

   a. the total aircraft operating costs calculated by obtaining the total average aircraft operating cost per hour from 4 multiplied by the total demand from 1 on a monthly basis, which can then be combined to form seasonal and yearly.
   b. in the same fashion as before, we are going to split the data to 2016-2019 and 2020 to see the overall impact on COVID-19 on the total aircraft operating costs on a monthly basis (seasonal and yearly as well).

### Data

*Specify your data sources, and the type of information you will use (for external data, provide links/references). For each data source, describe the variables and observations used in your analysis. Identify any potential issues (e.g. bias).*

Within the internal data source (original clean data):
- Duration (measurement of demand)
- Session_ID/Training_Type (unique identifier for each duration within a session)
- Aircraft (breakdown of duration by type of aircraft)
- Year/Month/Day (breakdown of duration by season/month/year and used for time series of external factors)

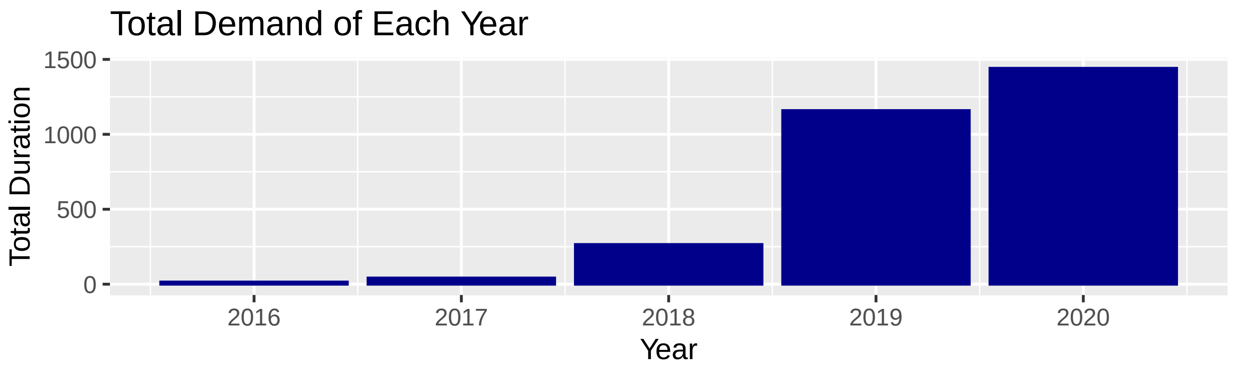
We plan on using the following external data sources:

1. Weather (https://climate.weather.gc.ca/historical_data/search_historic_data_e.html)
   - Historical weather data by day, including temp, precipitation, wind, etc.
   - No data for oshawa. We are assuming that data from another Toronto location (TORONTO BUTTONVILLE A) is accurate enough.
2. Fuel (https://www.indexmundi.com/commodities/?commodity=jet-fuel&months=240&currency=cad)
   - Historical fuel data by month, including the cost/gallon and %change
   - Assuming that all planes used in the flight school use the same fuel
3. Operating costs and repairs (https://cessna150152club.org/Costs)
   - The annual operating cost of a cessna plane model.
   - A static value, since the models being used are very similar

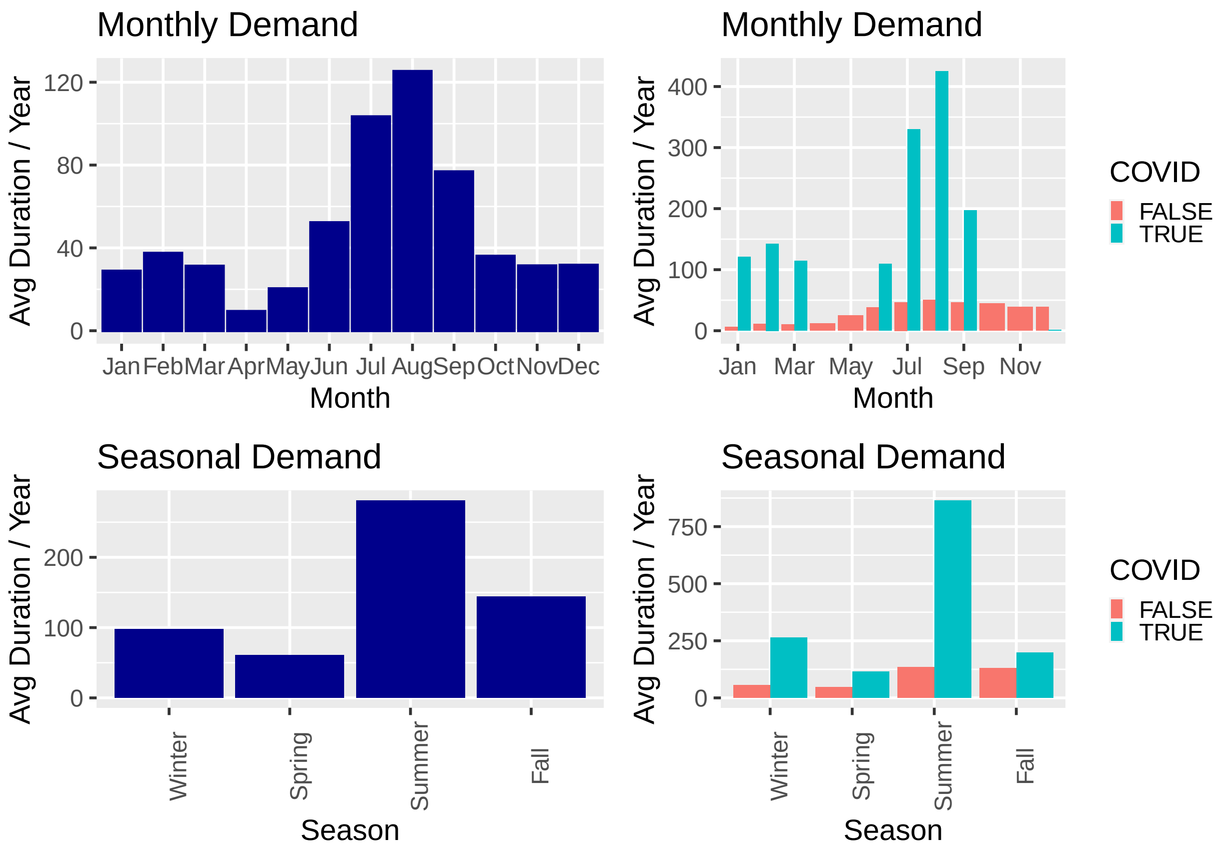
*Provide the R code that imports the data into R, and formats them appropriately (this can go to the appendix if it is too lengthy). Submit a copy of your external data files, if any.* See appendix.

**Preliminary Results**

*Create at least three data summaries/visualisations which are relevant to your questions, and comment on your results.*

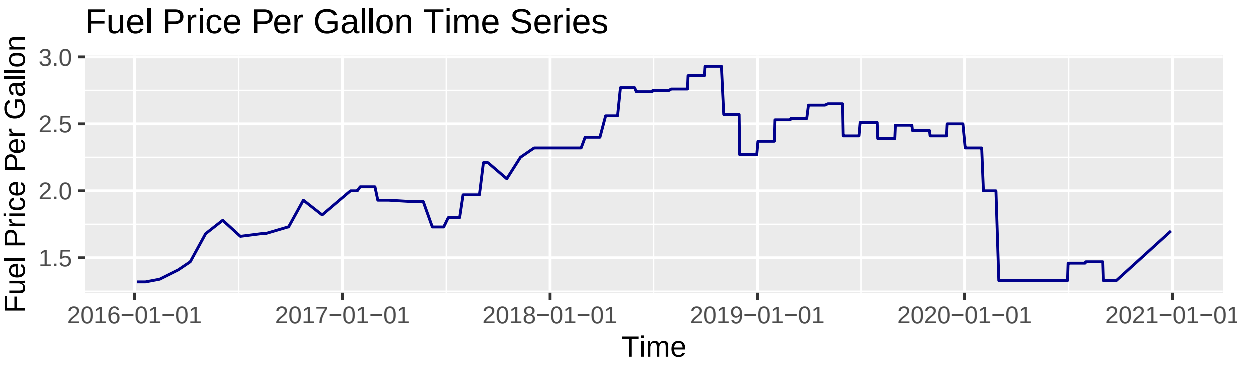


By the chart above, the total demand increased with COVID-19 there we expect that this will lead to higher total aircraft operating costs. We need to find the optimal conditions that will minimize aircraft operating costs during this spike.



We can see from the graphs that the month with the most demand is August.

Within the COVID-19, the pre-2020 data can actually be interpreted as skewed towards lower values since the overall demand in 2016-2018 is relatively low. Regardless of COVID-19, the month of most demand was August.



From the above graph, it can be seen there was a significant decrease in the jet fuel price per gallon, leading to lower aircraft operating costs. Therefore, we expect that the average and total estimated aircraft costs to be lower in 2020.

# Appendix

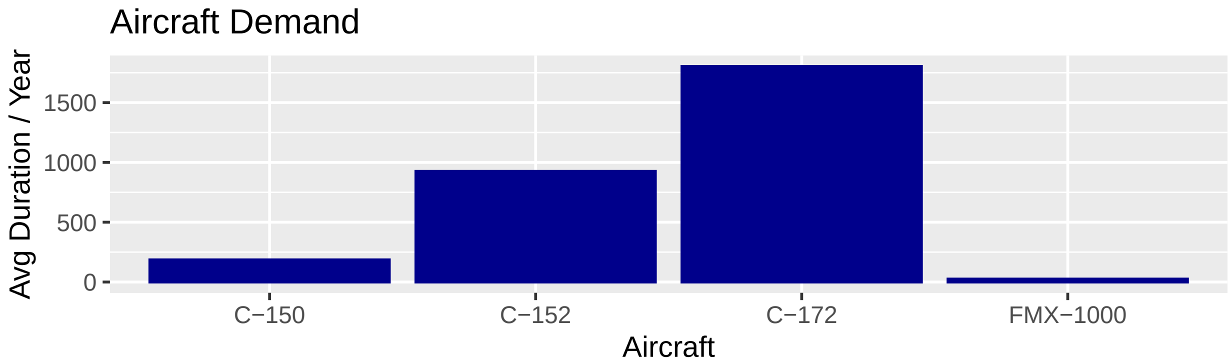
## Import/Format Data

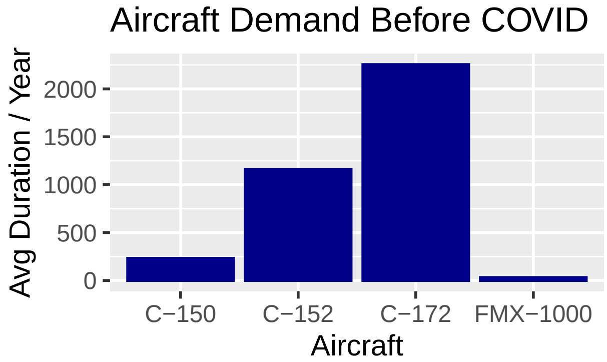
```
library(tidyverse)
library(lubridate)
library(ggpubr)
library(rvest)
rm(list = ls())

raw_data = NULL
for( i in 1:17){
  tmp = readxl::read_xlsx( "data/UofT Data Set.xlsx", skip = 1, sheet = i,
                           col_names =  paste( "X", 1:12, sep="" ) ) %>%
    mutate( Instructor_ID = i,
            PPL = X1,
            X1 = replace( X1, !str_detect(X1, "Student"), NA ),
            PPL = zoo::na.locf( PPL ),
            X1 = zoo::na.locf(X1) )
  raw_data = bind_rows( raw_data, tmp )
}
rm(tmp,i)

names( raw_data  ) = c( "Student", "Year", "Month", "Day", "Aircraft", "LF_dual",
                        "LF_solo", "Instrument_AC",  "Instrument_Sim", "CC_dual", "CC_solo", "Exercises
                        "Instructor_ID", "Licence")

head(raw_data)

raw_data %>%
  filter( !is.na(Year), Year != "Year",
          Year >= 2016, Year <= 2020,) %>%
  mutate_at( .vars = c(2:4), .funs = as.integer ) %>%
  mutate_at( .vars = c(6:11), .funs = as.numeric ) %>%
  mutate( Aircraft = str_to_upper(Aircraft),
          Aircraft = replace( Aircraft, str_detect(Aircraft, "GROUND"), "GROUND"),
          Aircraft = replace_na( Aircraft, "NA"),
          Aircraft = replace(Aircraft, Aircraft=="C152", "C-152"),
          Month = replace(Month, Month==111, 11),
          Other = ifelse( str_detect(Aircraft,"GROUND|NA"), -1, NA ),
          Student_ID = as.numeric( factor( paste( Student, Instructor_ID) ) ),
          Session_ID = row_number() ) %>%
  gather( key = "Training_Type", value = "Duration", 6:11, Other) %>%
  filter( !is.na(Duration) ) %>%
  mutate( Duration  = na_if(Duration, -1),
          Aircraft = na_if(Aircraft, "NA")) %>%
  select( Instructor_ID, Student_ID, Session_ID, Year, Month, Day,
          Aircraft, Duration, Training_Type, Exercises, Licence ) -> clean_data

getSeason <- function(month) {
  ifelse(month >= 3 & month <= 5, "Spring",
         ifelse(month >= 6 & month <= 8, "Summer",
                ifelse(month >= 9 & month <= 11, "Fall", "Winter")))
}
```

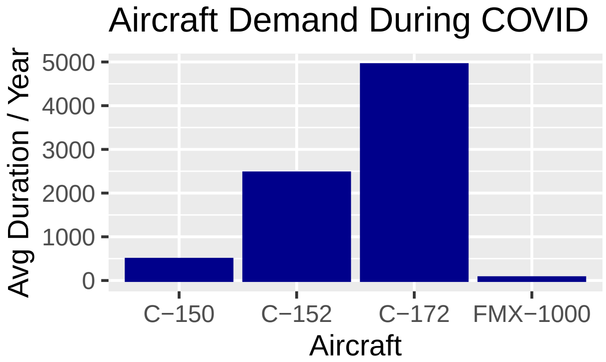
```
clean_data_processed = clean_data %>%
  distinct( Session_ID, .keep_all = T) %>%
  # split the exercises string into a "list" column w str_split()
  mutate( Exercises = str_split(Exercises, ",") ) %>%
  # and expand list contents into multiple rows w/ unnest()
  unnest(Exercises) %>%
  # remove invalid exercises
  mutate(Exercises = as.integer(Exercises)) %>%
  filter(Exercises >= 1 & Exercises <= 30) %>%
  distinct_all() %>%
  mutate(
    Season = getSeason(Month),
    Date = make_date(Year, Month, Day),
    COVID = Year >= 2020
  )

# reading in the fuel per gallon price
webpage <- read_html("https://www.indexmundi.com/commodities/?commodity=jet-fuel&months=240&currency=ca
tbls <- html_nodes(webpage, "table") %>%
  html_table(fill = TRUE)
fuel_prices <- as.data.frame(tbls[2])
colnames(fuel_prices) <- c("Month_Year", "Price","Change")
clean_data_processed$Month_Words <- month.abb[clean_data_processed$Month]
clean_data_processed$Month_Year <- (str_c(clean_data_processed$Month_Words, clean_data_processed$Year, 

clean_data_processed <- left_join(clean_data_processed,fuel_prices,by="Month_Year",copy = TRUE)
clean_data_processed <- select(clean_data_processed, -c("Month_Year","Change","Month_Words"))
colnames(clean_data_processed)[colnames(clean_data_processed) == 'Price'] <- 'Fuel Price Per Gallon'

process_weather <- function(data) {
  data %>%
    select(
      'Date/Time',
      "Mean Temp (°C)",
      "Total Precip (mm)",
      "Total Rain (mm)",
      "Total Snow (cm)",
      "Spd of Max Gust (km/h)"
    ) %>%
    mutate(wind = as.character(`Spd of Max Gust (km/h)`)) %>%
    select(-c("Spd of Max Gust (km/h)")) %>%
    mutate(wind = ifelse(wind == '<31', 31, wind))
}

climate_data_2016 <- read_csv('data/en_climate_daily_ON_6158410_2016_P1D.csv') %>% process_weather()
climate_data_2017 <- read_csv('data/en_climate_daily_ON_6158410_2017_P1D.csv') %>% process_weather()
climate_data_2018 <- read_csv('data/en_climate_daily_ON_6158410_2018_P1D.csv') %>% process_weather()
climate_data_2019 <- read_csv('data/en_climate_daily_ON_6158410_2019_P1D.csv') %>% process_weather()
climate_data_2020 <- read_csv('data/en_climate_daily_ON_6158410_2020_P1D.csv') %>% process_weather()
climate_data <- bind_rows(climate_data_2016,
                          climate_data_2017,
                          climate_data_2018,
                          climate_data_2019,
                         climate_data_2020)

clean_data_processed <- left_join(clean_data_processed, climate_data, by=c("Date" = "Date/Time"), copy=
```

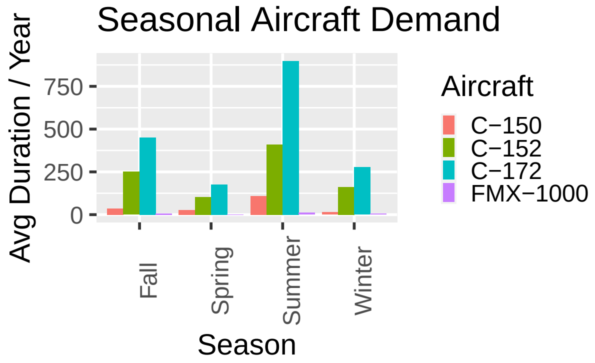
## Demand per aircraft

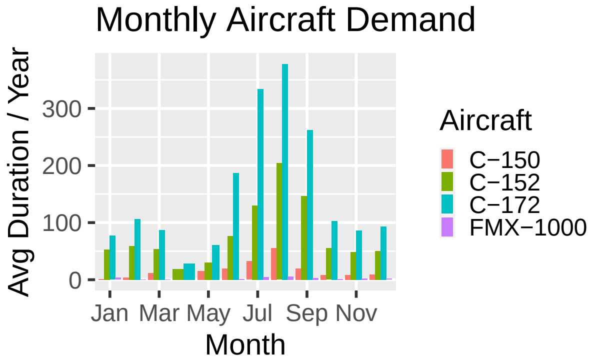
Below is an attempt to see if an interesting analysis can be made from examing the demand per aircraft and hence the aircraft operating costs per airctaft.


Aircraft Demand
Avg Duration / Year
C−150
C−152
C−172
FMX−1000
Aircraft


Aircraft Demand Before COVID
Avg Duration / Year
C−150
C−152
C−172
FMX−1000
Aircraft


Aircraft Demand During COVID
Avg Duration / Year
C−150
C−152
C−172
FMX−1000
Aircraft


Seasonal Aircraft Demand
Avg Duration / Year
Fall
Spring
Summer
Winter
Season
Aircraft
C−150
C−152
C−172
FMX−1000


Monthly Aircraft Demand
Avg Duration / Year
Jan
Mar
May
Jul
Sep
Nov
Month
Aircraft
C−150
C−152
C−172
FMX−1000