STAA57 W21 Draft Report

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

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

Introduction

(*Description of questions that are being investigated*)

1. What are the trends in jet fuel costs over different time units (months/season)?

- For this question we use the web scrapped to see how fuel price changes during time and we try to see how these changes affect the cost of flight or session for the company

2. What factors affect the duration of the sessions (model 1)

- we decide here to use a linear regression model to assess significant facotrs to how duration of session is affected

3. What factors affect cost in a significant matter (model 2)

- we decide here to use a linear regression model to assess significant facotrs to how cost is affected

4. What are the trends in demand over different time units (months/season/year)?

- we find how the demand in flight training changes and was season is more significant

5. Considering the previous questions, what are the estimated average aircraft operating over different time units (months/season)?
6. What are the total aircraft operating costs over month/season/year?

- assuming we know total number of planes

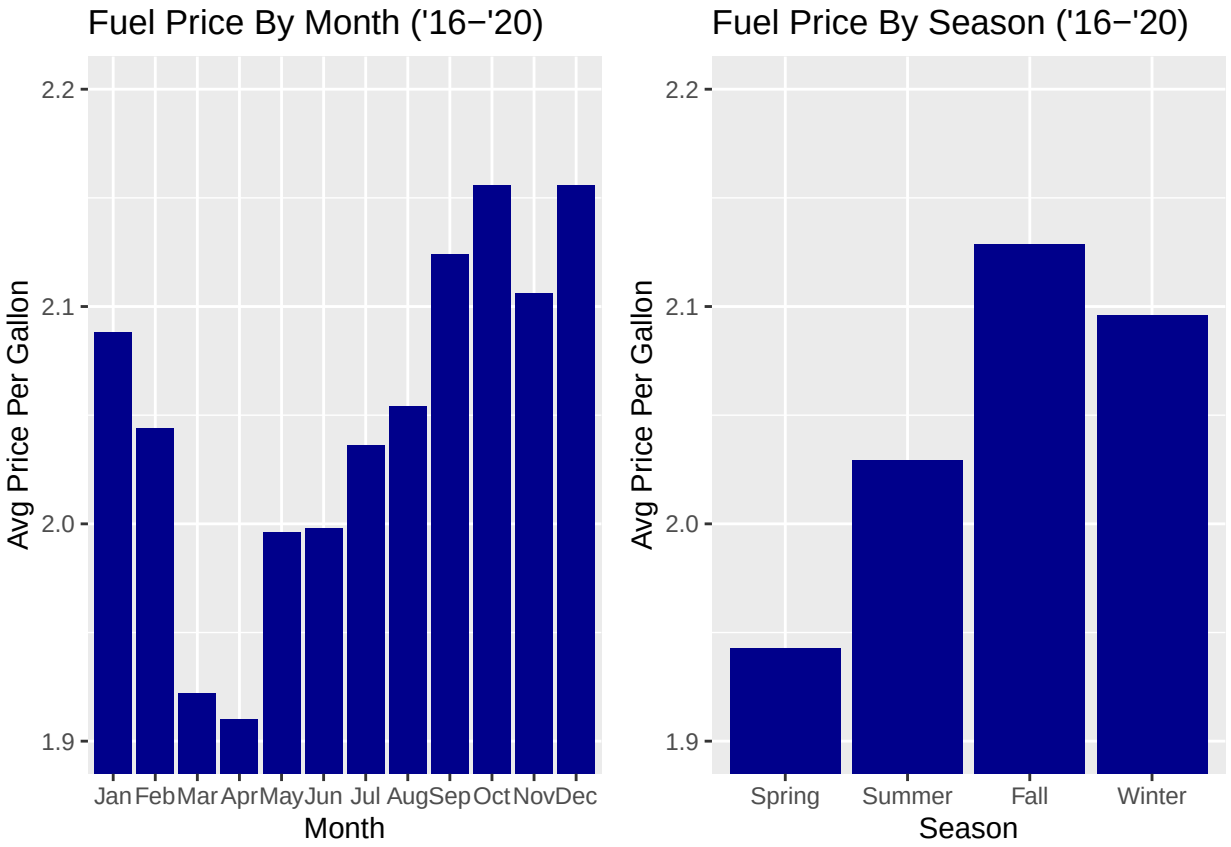
Data

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

- AVGAS: (NEW LINK)
  - From preliminary research, it was found that C-172, C-150 and C-152 aircraft models only consume jet fuel (https://airshare.com/blog/cessna-172/#:~:text=and%20accessible%20training.-,What%20type%20of%20fuel%20does%20it%20use%3F,to%20power%20it%27s%20piston%20engine)
  - The AVGAS historical data from BLANK is given by month, including the cost per gallon and percentage change
- 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 dollar static cost per hour was added to each training session cost
  - In addition, the fuel consumption per hour is estimated to be 6
- Weather (https://climate.weather.gc.ca/historical_data/search_historic_data_e.html): - Historical weather data by day, including temp, precipitation, wind, etc. - Gathered data from Oshawa location

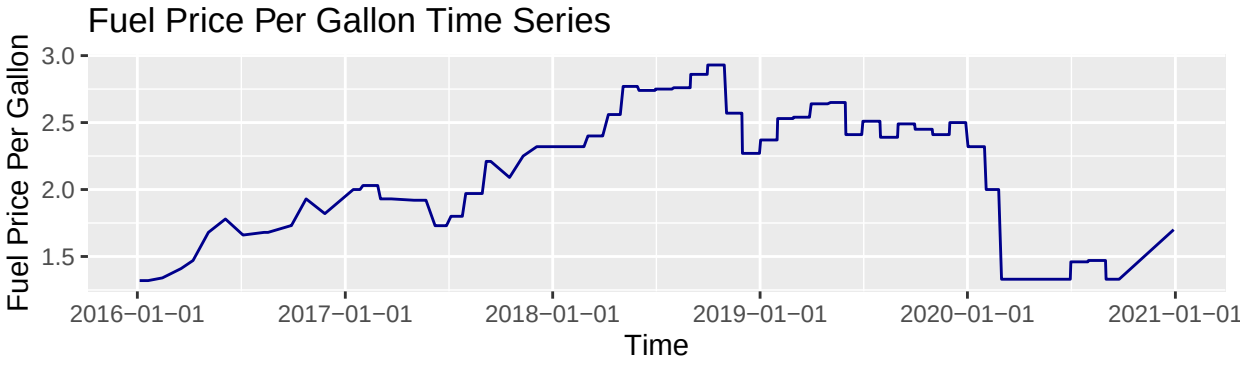
Analysis

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



From the monthly data on the left, we can see that in March and April the fuel price is significantly cheaper than any other month at a price around $1.91/gallon. Following that, it steadily increases and peaks around $2.15/gallon around October and December.

For the seasonal data on the right, we can see that spring is significantly cheaper than any other season with a price at around $1.94/gallon. On the other hand, fuel price is most expensive in the fall season with a price of about $2.13/gallon



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.

Question 3: What factors affect the duration of sessions?

How was the data prepared?

For sessions that perform a particular exercise set, we suspect that their total duration should be around the average total duration for that same exercise set, unless there are external factors that affect it. Thus we can utilize a multiple linear regression model to show any significance for external factors against duration.

To prepare the data for our model, we first have to group the data by exercise set. Noting that there are a lot of unique exercise combinations as we deemed that to have enough data for this analysis.

We also noted that we have to compute the total duration for a single session since a single session might have multiple training types resulting in multiple rows. Therefore we aggregated the rows to get the total duration for a particular session. To do this we can simply grouped by `Session_ID` and added the duration.

As a minor note, we also removed sessions that occurred on the ground as we wanted our model to only measure duration with real planes against external factors.

Lastly, to add on the external factors we joined our weather data to the processed data set. We identified an initial set of weather factors that we believed could have a strong effect on the duration of a session. These were:

- avg_relative_humidity
- avg_dew_point
- avg_pressure_sea
- avg_pressure_station
- avg_visibility
- avg_health_index
- precipitation
- avg_cloud_cover_4
- avg_temperature

Selecting the most popular exercise routine

Table 1: Most Frequest Group of Exercises

Exercises	count
16,17,18,30	664
16,17,18	136
16,17,18,23,30	89

Linear regression model for Duration per Session:

```
##
## Call:
## lm(formula = Total_Duration ~ avg_relative_humidity + avg_dew_point +
##     avg_pressure_sea + avg_pressure_station + avg_visibility +
##     avg_health_index + avg_cloud_cover_4 + avg_temperature +
##     precipitation + Season - 1, data = duration.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90956 -0.12692  0.02309  0.15726  1.30163
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## avg_relative_humidity -1.721e-03  3.209e-03  -0.536    0.592
## avg_dew_point          8.800e-03  1.224e-02   0.719    0.473
## avg_pressure_sea       2.599e+00  1.804e+00   1.441    0.150
## avg_pressure_station  -2.610e+00  1.837e+00  -1.421    0.156
## avg_visibility         3.576e-06  5.475e-06   0.653    0.514
## avg_health_index       3.860e-03  2.885e-02   0.134    0.894
## avg_cloud_cover_4      1.641e-02  2.066e-02   0.795    0.427
## avg_temperature        1.060e-03  1.168e-02   0.091    0.928
## precipitation          3.558e-04  3.885e-03   0.092    0.927
## SeasonFall            -2.415e+00  2.261e+00  -1.068    0.286
## SeasonSpring          -2.413e+00  2.256e+00  -1.069    0.285
## SeasonSummer          -2.277e+00  2.254e+00  -1.010    0.313
## SeasonWinter          -2.384e+00  2.261e+00  -1.054    0.292
##
## Residual standard error: 0.2968 on 642 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.9224, Adjusted R-squared:  0.9208
## F-statistic: 586.9 on 13 and 642 DF,  p-value: < 2.2e-16
```

Analysis for Duration per Session linear regression model

It can be seen from the summary function that the p-values for all the covariates are high, despite having a high R-squared of 0.89. Having a high R-squared is good for a model since most of the variation in the response variable can be explained through the model. However coupled with high p-values for all covariates renders the model useless as it implies that none of the covariates have a satistical significance in association with the total duration. Intuitively since we are multiple covariates from weather data, it was thought that multiple collinearity might be the cause of the high p-values. Hence, a correlation matrix for the weather data was calculated.

Correlation matrix for the weather covariates

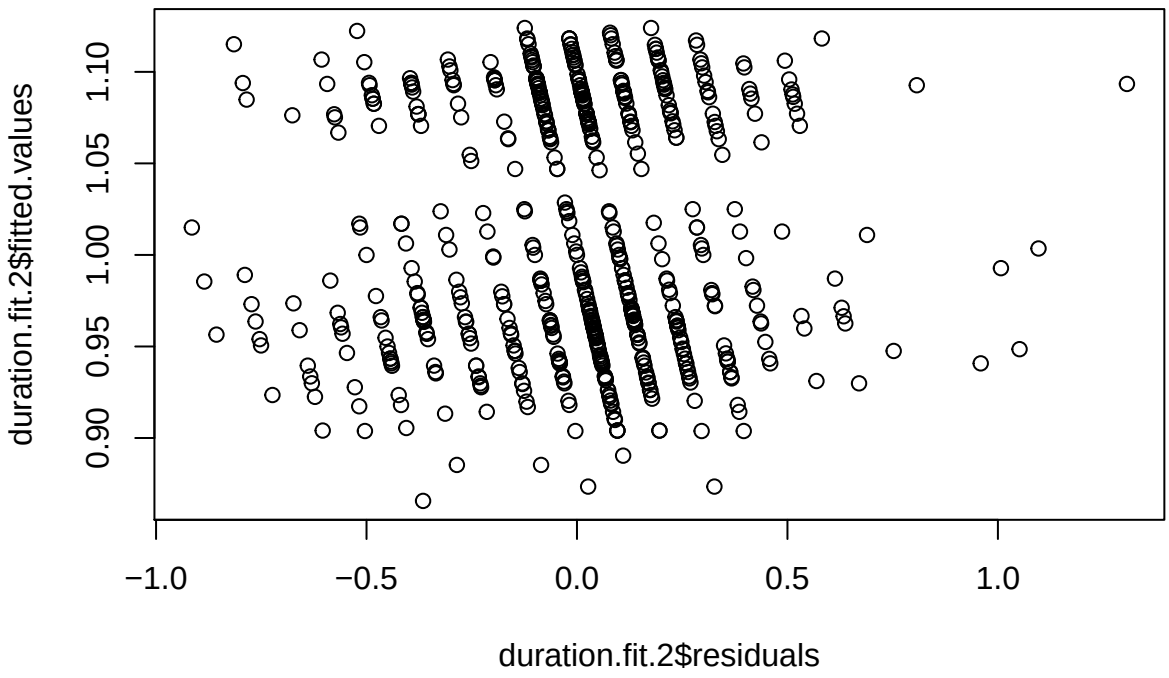
```
##                       avg_relative_humidity avg_dew_point avg_pressure_sea
```

```
## avg_relative_humidity             1.0000000     0.5205644      -0.31752124
## avg_dew_point                     0.5205644     1.0000000      -0.38949256
## avg_pressure_sea                 -0.3175212    -0.3894926       1.00000000
## avg_pressure_station             -0.3166981    -0.3712001       0.99971608
## avg_visibility                   -0.5786371    -0.1708754       0.32336137
## avg_health_index                 -0.1063215    -0.1195438       0.07779705
## precipitation                     0.3246735     0.2533825      -0.26405725
## avg_cloud_cover_4                 0.4869239     0.2761234      -0.34845257
## avg_temperature                   0.1628771     0.9076234      -0.31691614
##                       avg_pressure_station avg_visibility avg_health_index
## avg_relative_humidity          -0.31669809    -0.57863712      -0.10632147
## avg_dew_point                  -0.37120007    -0.17087539      -0.11954378
## avg_pressure_sea                0.99971608     0.32336137       0.07779705
## avg_pressure_station            1.00000000     0.32768446       0.07597322
## avg_visibility                  0.32768446     1.00000000      -0.06999009
## avg_health_index                0.07597322    -0.06999009       1.00000000
## precipitation                  -0.26253135    -0.32768768      -0.01866383
## avg_cloud_cover_4              -0.34900166    -0.35403107      -0.10594446
## avg_temperature                -0.29615833     0.05126322      -0.09290758
##                       precipitation avg_cloud_cover_4 avg_temperature
## avg_relative_humidity    0.32467352         0.4869239      0.16287706
## avg_dew_point            0.25338250         0.2761234      0.90762337
## avg_pressure_sea        -0.26405725        -0.3484526     -0.31691614
## avg_pressure_station    -0.26253135        -0.3490017     -0.29615833
## avg_visibility          -0.32768768        -0.3540311      0.05126322
## avg_health_index        -0.01866383        -0.1059445     -0.09290758
## precipitation            1.00000000         0.3321391      0.15428709
## avg_cloud_cover_4        0.33213906         1.0000000      0.11097086
## avg_temperature          0.15428709         0.1109709      1.00000000
```

From the correlation matrix, it can be seen that avg_pressure_sea and avg_pressure_station have a correlation of almost 1. avg_pressure_station was chosen to be removed as avg_pressure_sea measures are standardly used during aircraft take offs. Also, avg_dew_point and avg_temperature have a correlation of 0.9. From background research, it was found that dew_point is used as an indication of humidity and visibility, thus it was chosen to be removed from model since humidity and visibility are both given. A new linear regression model was fit disregarding these two covariates.

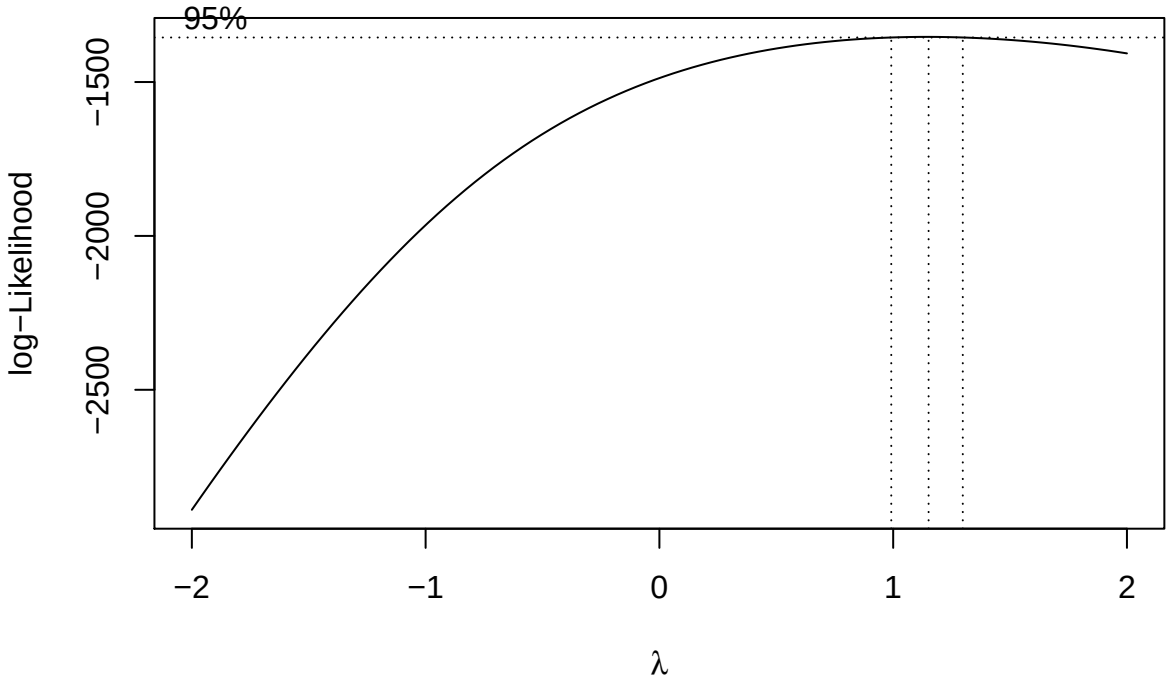
```
##
## Call:
## lm(formula = Total_Duration ~ avg_relative_humidity + avg_pressure_sea +
##     avg_visibility + avg_health_index + avg_cloud_cover_4 + avg_temperature +
##     precipitation + Season - 1, data = duration.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91498 -0.12538  0.02799  0.15767  1.30660
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## avg_relative_humidity  3.320e-04  1.829e-03   0.181   0.8560
## avg_pressure_sea       3.567e-02  2.162e-02   1.650   0.0995 .
## avg_visibility         3.153e-06  5.452e-06   0.578   0.5632
## avg_health_index       1.140e-02  2.831e-02   0.403   0.6873
## avg_cloud_cover_4      1.765e-02  2.064e-02   0.855   0.3929
## avg_temperature        1.357e-03  2.282e-03   0.595   0.5521
## precipitation          2.878e-04  3.809e-03   0.076   0.9398
## SeasonFall            -2.821e+00  2.236e+00  -1.262   0.2075
## SeasonSpring          -2.817e+00  2.231e+00  -1.262   0.2073
## SeasonSummer          -2.678e+00  2.230e+00  -1.201   0.2301
## SeasonWinter          -2.788e+00  2.235e+00  -1.247   0.2127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2969 on 644 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.9221, Adjusted R-squared:  0.9208
## F-statistic: 693.2 on 11 and 644 DF,  p-value: < 2.2e-16
```



Unfortunately, despite attempting to remove variables with high correlation, the model still performed poorly with high p-values. A further look into the residual plot gave a clear indication

that the linear regression model assumption of having a constant variation among residuals was being violated, giving a further indication that a linear model is not a good fit for the data. To remidy this a boxcox transformation was attempted to see if a power of the response variable would improve the linear model fit.



Unfortunately, it seemed that the power of 1 is still one of the best, therefore putting the response variable to a different power wouldn't improve the model. Another attempt at taking log function of the response variable was undertaken instead.

```
##
## Call:
## lm(formula = log(Total_Duration) ~ avg_relative_humidity + avg_pressure_sea +
##     avg_visibility + avg_health_index + avg_cloud_cover_4 + avg_temperature +
##     precipitation + Season - 1, data = duration.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22727 -0.07802  0.08466  0.20658  0.81557
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## avg_relative_humidity -5.716e-04  2.351e-03  -0.243    0.808
## avg_pressure_sea       3.788e-02  2.780e-02   1.363    0.173
## avg_visibility        -2.386e-07  7.009e-06  -0.034    0.973
```

```
## avg_health_index       1.535e-03  3.639e-02   0.042    0.966
## avg_cloud_cover_4      2.252e-02  2.653e-02   0.849    0.396
## avg_temperature        1.437e-03  2.933e-03   0.490    0.624
## precipitation         -1.428e-03  4.896e-03  -0.292    0.771
## SeasonFall            -3.996e+00  2.874e+00  -1.390    0.165
## SeasonSpring          -4.000e+00  2.868e+00  -1.395    0.164
## SeasonSummer          -3.820e+00  2.866e+00  -1.333    0.183
## SeasonWinter          -3.954e+00  2.874e+00  -1.376    0.169
##
## Residual standard error: 0.3816 on 644 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.0609, Adjusted R-squared:  0.04486
## F-statistic: 3.797 on 11 and 644 DF,  p-value: 2.667e-05
```

However, the R-squared decreased to 0.061 and the p-values were still high therefore the model is still insufficient.

Summary of Duration Per Session Model

Overall, it seemed that using the linear regression model to

Plan:

1. try to fit and analyze R squared and pvalues
2. high R squared and pvalues, therefore we suspected that it might be multi-collinearity of the data hence calculated a correlation matrix for the weather data (culprit)
3. there was high correlation between some values, which we decided to remove
4. model was still performing poorly
5. decided to take a power of the y and log of y to see if it will change the model, but it was the same
6. since we didnt find any significant factors that can be used, we will create a cost model with all covariates

From these results we could not identify any common features between the three models. Additionally, for all models, all factors have a p-value > 0.05 and thus individually are not statistically significant. We can conclude that

Some Sub-section Title

(*R output tables can be formatted nicely in .html with* `knitr::kable()`)

Table 2: Table Caption

X	Y
this	is
my	table

Summary

(*Recap of your findings and conclusions*)

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", "I
                        "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&cu
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_processe

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 Ga

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_weat
climate_data_2017 <- read_csv('data/en_climate_daily_ON_6158410_2017_P1D.csv') %>% process_weat
climate_data_2018 <- read_csv('data/en_climate_daily_ON_6158410_2018_P1D.csv') %>% process_weat
climate_data_2019 <- read_csv('data/en_climate_daily_ON_6158410_2019_P1D.csv') %>% process_weat
climate_data_2020 <- read_csv('data/en_climate_daily_ON_6158410_2020_P1D.csv') %>% process_weat
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"
```