SYS 5581 Project - Exploratory Data Analysis

Nick Coronato

Version of 2021-03-14 | Due 2021-03-10

Step 1: Identify a Time Series data set that you want to work with

For this project, I will be analyzing a set of airline passenger data from the Bureau of Transportation Statistics (BTS). The data set includes air travel data from *January 2015 - August 2020*.

Step 2: Acquire the data from its source location, reproducibly.

For this project, my data is stored on my local machine. The file name is "5year_t100_data.csv"

Data was obtained from the BTS website: https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FIM or https://www.transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FIM

Note: Ideally the data will be stored at and read from a Github repository.

Note 2: The BTS recently released 2020 data through November, so I updated the .csv file on my local machine with the new data. Also, I decided to go back 5 years instead of 2.

```
library("googledrive")
(sharedrive <- drive_get("https://drive.google.com/file/d/1Z41tBkWTtdVNUJyfhLEVjeDn7f_CBI9x/view?usp=sha
```

```
## Using an auto-discovered, cached token.
## To suppress this message, modify your code or options to clearly consent to the use of a cached toke
## See gargle's "Non-interactive auth" vignette for more details:
## https://gargle.r-lib.org/articles/non-interactive-auth.html

## The googledrive package is using a cached token for njc7f@virginia.edu.

## # A tibble: 1 x 3
##   name                 id                                drive_resource
## * <chr>                <chr>                             <list>
## 1 5year_t100_data.csv 1Z41tBkWTtdVNUJyfhLEVjeDn7f_CBI9x <named list [39]>
```

```
# WARNING : this is a large file (about 118MB). The googledrive package should install this to your cur

t100_raw <- drive_download(
  "5year_t100_data.csv",
  path = "5year_t100_data.csv",
  overwrite = TRUE
)
```

```
## File downloaded:
##   * 5year_t100_data.csv
## Saved locally as:
##   * 5year_t100_data.csv
```

```
t100_raw  <-  read_delim("5year_t100_data.csv",",",
                      col_types = cols(.default = col_character(),
                                       "YEAR" = col_integer(),
                                       "MONTH" = col_integer(),
                                       "UNIQUE_CARRIER" = col_factor(),
                                       "PASSENGERS" = col_integer(),
                                       "DISTANCE" = col_integer(),
                                       "ORIGIN_AIRPORT_ID" = col_factor(),
                                       "ORIGIN_STATE_NM" = col_factor(),
                                       "DEST_AIRPORT_ID" = col_factor(),
                                       "DEST_STATE_NM" = col_factor(),
                                       "SEATS" = col_integer(),
                                       "AIR_TIME" = col_integer()
                                               ))
```

Step 3: Organize your data into a *tidy* data frame.

Organize by taking out the non-useful variables. Make another new variable called RPM (revenue passenger miles), which is simply Passengers X Miles for each observation. Put the Index variable in the first column (eventually it will be 4dig year, 2dig month). Put the Key variables next (unique_carrier, Passengers, Distance, RPM). Then all the other (possibly) relevant variables.

```
t100_raw %>%
  mutate(., RPM = PASSENGERS * DISTANCE) %>%
  select(YEAR, MONTH, UNIQUE_CARRIER, PASSENGERS, DISTANCE, RPM, -ORIGIN_AIRPORT_ID, -ORIGIN_CITY_NAME,
```

Make a new variable called YRMO that concatenates Year and Month into a YR-MO format. Make it a time series column.

```
# Put together YEAR and MONTH into a single column called YRMO
t100_raw %>%
  mutate(YRMO = paste(t100_raw$YEAR, t100_raw$MONTH)) ->t100_raw

# Convert YRMO to a time series object column (year month) and re-arrange
t100_raw %>%
  mutate(YRMO = yearmonth(YRMO))  %>%
  select(YRMO, everything())-> t100_raw2

# Create summary values for Passengers, RPM, Distance, Seats, and Air_time that will prevent us from ha
t100_raw2 %>%
  group_by(YRMO, YEAR, MONTH, UNIQUE_CARRIER) %>%
  summarize(TotalPax=sum(PASSENGERS), TotalRPM = sum(RPM), TotalDistance = sum(DISTANCE), TotalSeats =

# convert to tsibble
as_tsibble(t100_raw3, index = YRMO, key = UNIQUE_CARRIER) -> t100_ts_tbl
```

Some quality control checks.

```
t100_ts_tbl %>% distinct() ->t100_ts_tbl  #remove duplicated rows, if any exist
head(t100_ts_tbl)  #Check that each column got the right Type.
```

```
## # A tibble: 6 x 9
## # Groups:   YRMO, YEAR, MONTH [6]
##       YRMO  YEAR MONTH UNIQUE_CARRIER TotalPax TotalRPM TotalDistance TotalSeats
##      <mth> <int> <int> <fct>             <int>    <dbl>         <int>      <int>
## 1 2015 Jan  2015     1 UA              4847338   6.54e9       2446772    5956785
## 2 2015 Feb  2015     2 UA              4639362   6.08e9       2316983    5590926
## 3 2015 Mar  2015     3 UA              5823398   7.74e9       2474956    6831596
## 4 2015 Apr  2015     4 UA              5556445   7.43e9       2549360    6531908
## 5 2015 May  2015     5 UA              6048718   8.07e9       2585598    7051073
## 6 2015 Jun  2015     6 UA              6296079   8.62e9       2655756    7228780
## # ... with 1 more variable: TotalAIR_TIME <int>
```

Generate and print the tsibble.

```
## # A tsibble: 213 x 9 [1M]
## # Key:       UNIQUE_CARRIER [3]
## # Groups:    YEAR, MONTH @ YRMO [71]
##        YRMO  YEAR MONTH UNIQUE_CARRIER TotalPax TotalRPM TotalDistance
##       <mth> <int> <int> <fct>             <int>    <dbl>         <int>
##  1 2015 Jan  2015     1 UA              4847338   6.54e9       2446772
##  2 2015 Feb  2015     2 UA              4639362   6.08e9       2316983
##  3 2015 Mar  2015     3 UA              5823398   7.74e9       2474956
##  4 2015 Apr  2015     4 UA              5556445   7.43e9       2549360
##  5 2015 May  2015     5 UA              6048718   8.07e9       2585598
##  6 2015 Jun  2015     6 UA              6296079   8.62e9       2655756
##  7 2015 Jul  2015     7 UA              6501688   9.00e9       2538269
##  8 2015 Aug  2015     8 UA              6335476   8.63e9       2688078
##  9 2015 Sep  2015     9 UA              5711433   7.43e9       2617487
## 10 2015 Oct  2015    10 UA              6386700   8.20e9       2541631
## # ... with 203 more rows, and 2 more variables: TotalSeats <int>,
## #   TotalAIR_TIME <int>
```

Exploratory Data Analysis

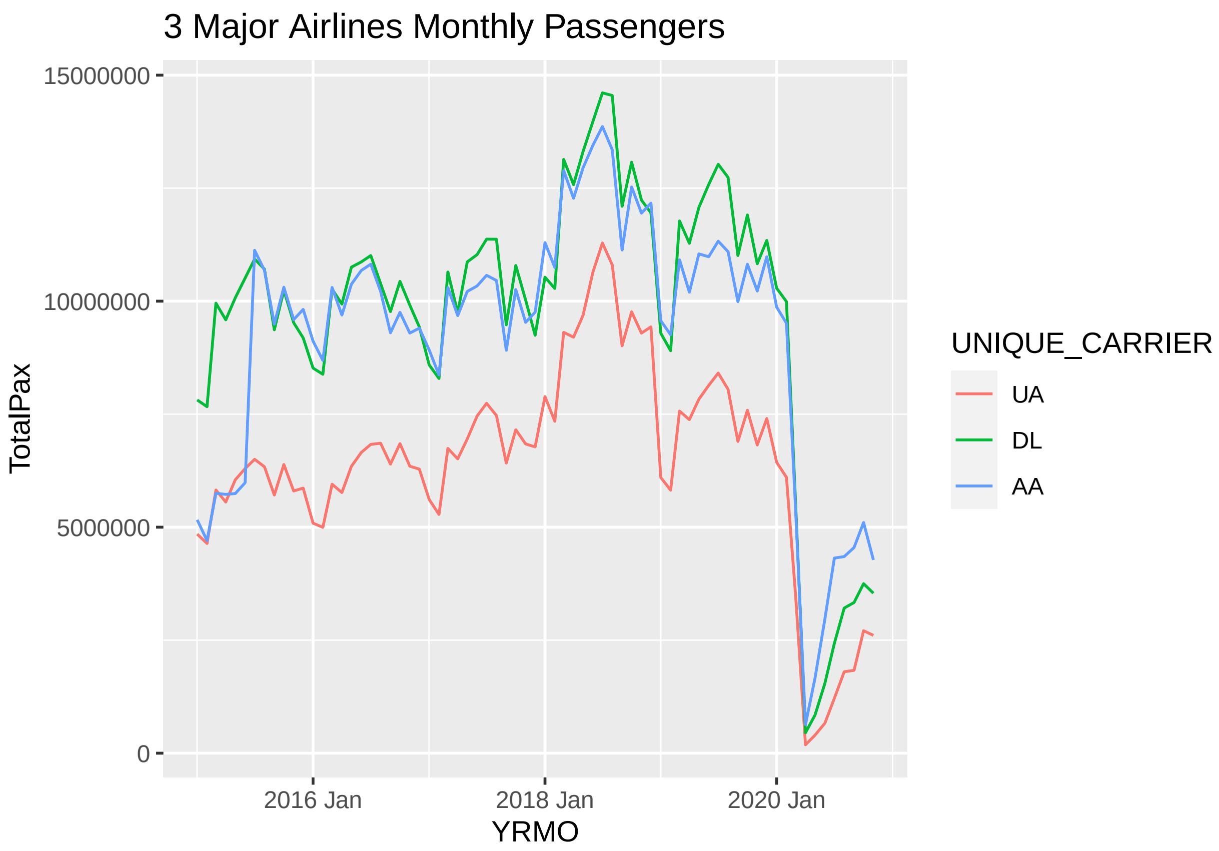
Briefly characterize the data set.

Provide a brief example of the data, showing how they are structured.

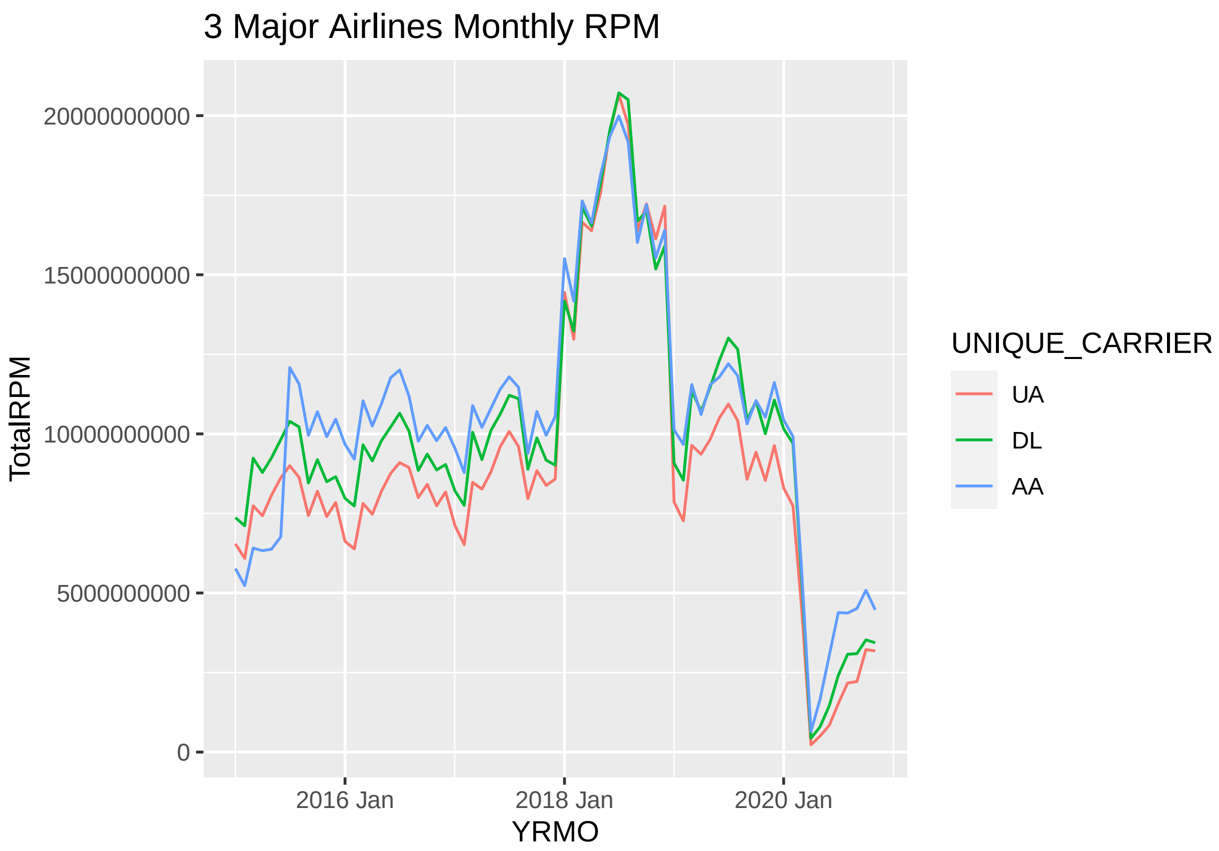
After extracting and transforming my data, I have a useful tibble with 204 observations. Each observation is a Monthly report of 9 variables. The time-series tibble called "big3" is the object I will be using for most analysis: it highlights the 3 U.S.-based airline giants that could give useful information about passenger trends in the domestic market.

I can view any of my trends for (Passengers, Seats, Miles Flown, Hours of Air Time) through ggplots like seen below.

```
#Show total passengers trend since 2015 Jan
ggplot(data = big3, mapping = aes(x = YRMO, y = TotalPax, color = UNIQUE_CARRIER)) + geom_line() + ggti
```



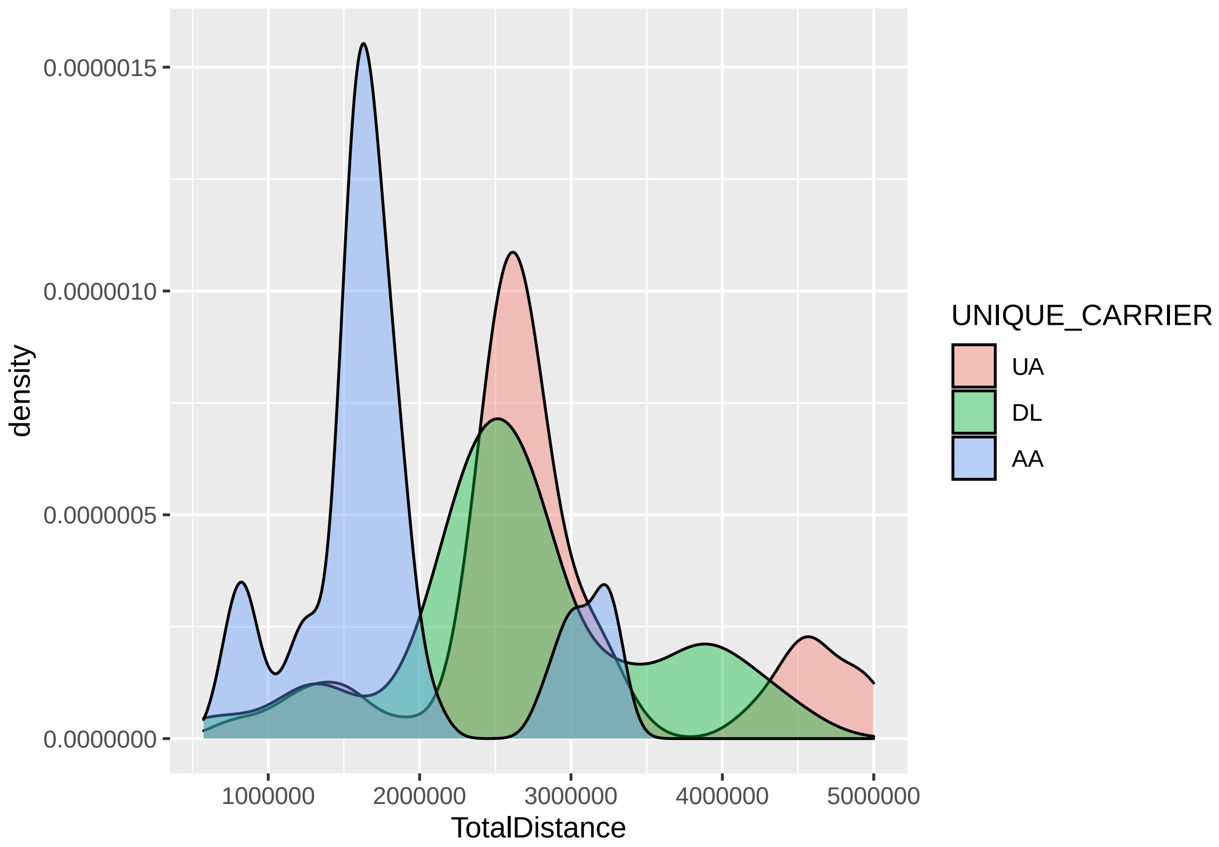
```
#Using the Feasts package (optional)
# feasts::autoplot(big3, TotalPax) +
#   ylab("Total Passengers per Month") +
#   xlab("")
#


#Show total Revenue Passenger Miles (RPM) trend since 2015 Jan
ggplot(data = big3, mapping = aes(x = YRMO, y = TotalRPM, color = UNIQUE_CARRIER)) + geom_line() + ggti
```



Interesting note: United Airlines produces a consistently lower TotalPax count. The red line is a similar shape to DL and UA, but shifted down by about 5 million passengers per month. However, I noticed that they are near the same level of production of Revenue Passenger Miles. This makes me think that United Airlines moves way fewer passengers, but possibly over longer domestic routes, therefore producing a comparable RPM value every month.

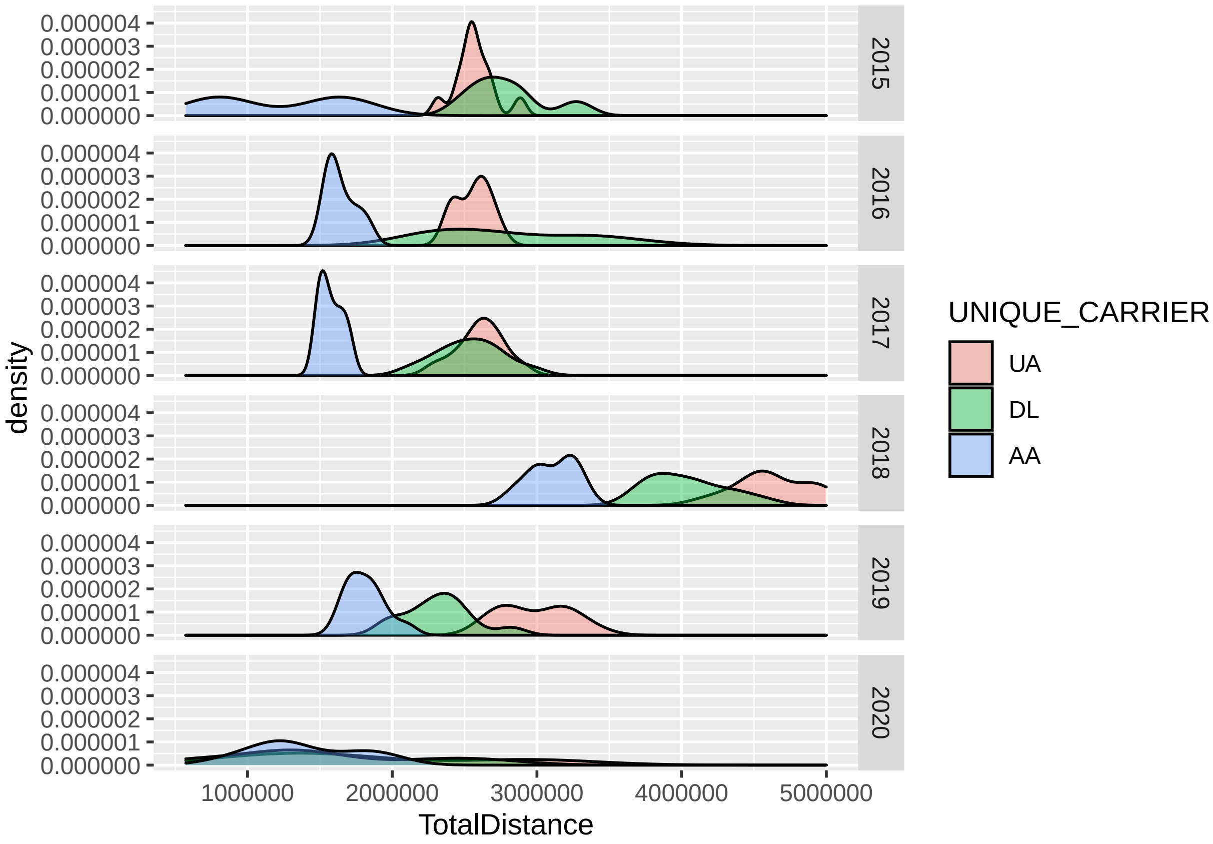
So I checked the distance flown statistics across the airlines.

```
#Display density plots of distance flown between the 3 carriers
ggplot(big3, aes(x = TotalDistance, fill = UNIQUE_CARRIER)) +
  geom_density(position = "identity", alpha = 0.4, color = "black")
```

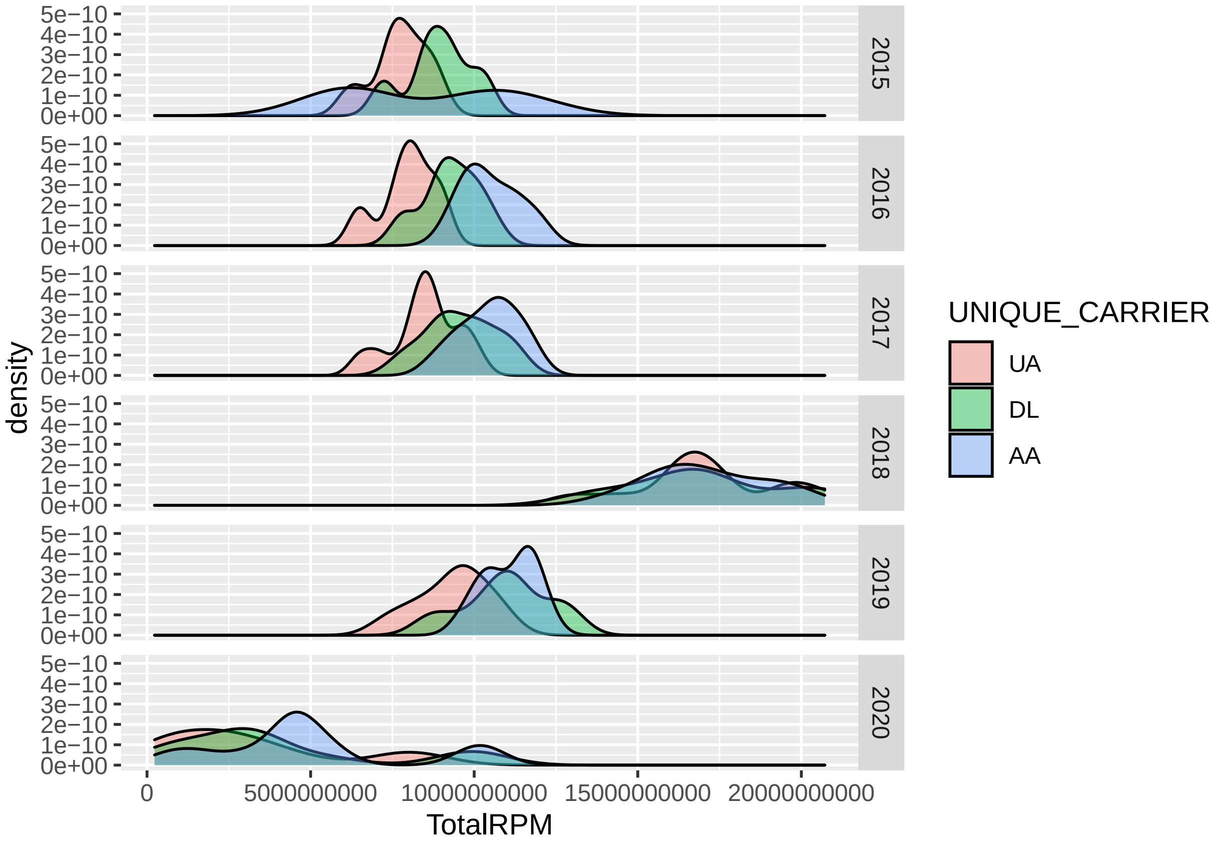


Yes, there seems to be evidence that the 3 airlines under investigation do have different strategies in terms of route lengths/distance flown in this data set. This could be a function of their actual corporate strategy, or just the nature of their "hub" locations, or even totally random. The chart below breaks it out by year and shows that United Airlines seems to fly longest distance routes (on average), American Airlines flies the shortest routes, and Delta's flight distance distribution is widely spread. Interesting stuff; maybe useful.

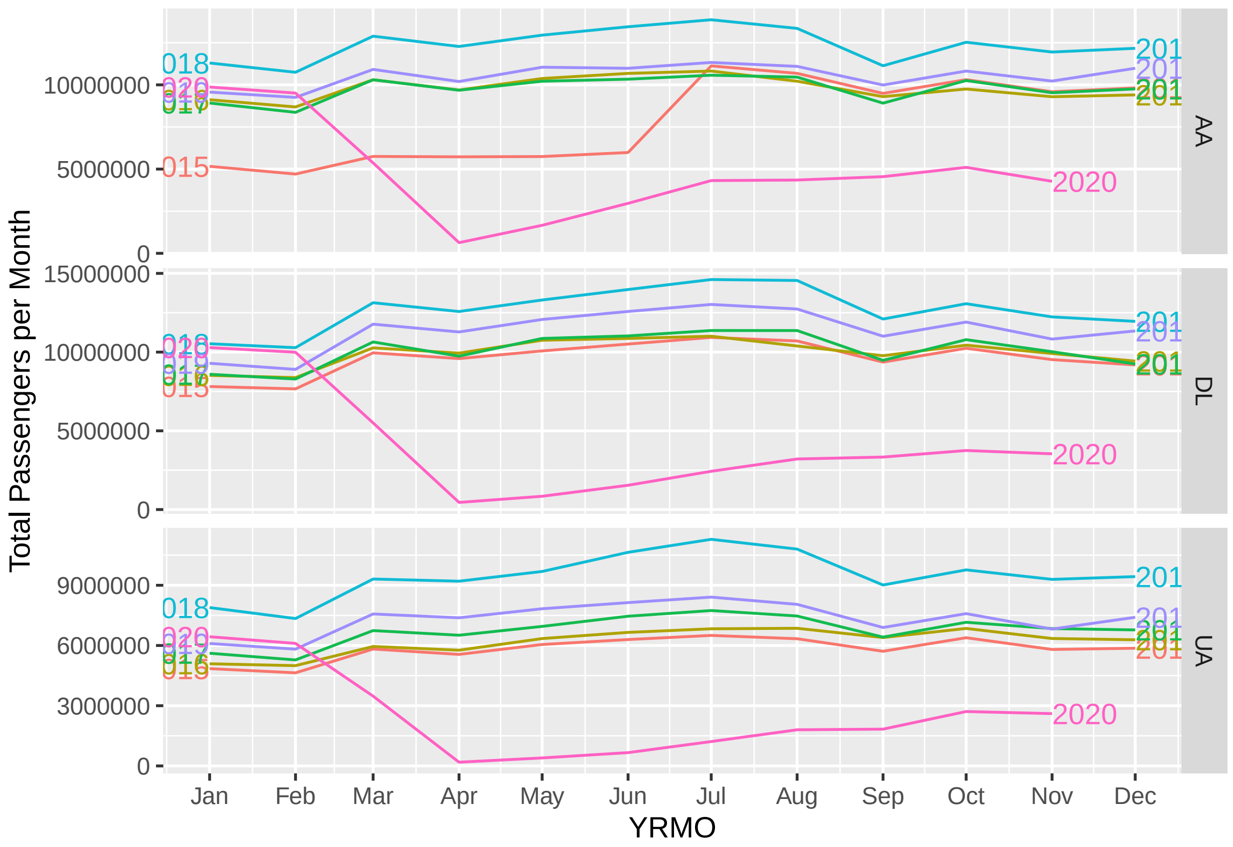
```
#Display density plots of distance flown between the 3 carriers
ggplot(big3, aes(x = TotalDistance, fill = UNIQUE_CARRIER)) +
  geom_density(position = "identity", alpha = 0.4, color = "black") +
  facet_grid(YEAR ~ .)
```



```r
ggplot(big3, aes(x = TotalRPM, fill = UNIQUE_CARRIER)) +
  geom_density(position = "identity", alpha = 0.4, color = "black") +
  facet_grid(YEAR ~ .)
```



```r
#Seasonal plot
big3 %>%
  feasts::gg_season(TotalPax, labels = "both") + ylab("Total Passengers per Month")
```



The Seasonal Plot above shows a pretty clear representation of the seasonal trends in air travel amongst these three airlines. (Note: Y-axis is not consistent; adjusted for each airline).

Prior to 2020, the most popular year for air travel was 2018 (according to most metrics of interest here).

Seasonal trends are somewhat as one would expect for air travel. We can see that February and September are the low points for passengers being transported, and the summer and holiday months create a spike for these 3 airlines.

There was an interesting phenomenon for American Airlines in summer 2015, where their Total Passenger count increased greatly. This could be due to general corporate expansion, acquisition of new (or bigger) airplanes, or initiation of new flight routes.

(A little research showed that AA did place "the 'largest aircraft order in history' in July 2011, purchasing 460 'next generation' Boeing 737 and Airbus A320 aircraft for delivery between 2013 and 2022. Also, 2015 was the year they completed a merger with US Airways: On April 8, 2015, the Federal Aviation Administration awarded American Airlines and US Airways a single operating certificate." At some point in 2015, I'm assuming the Transportation Bureau started counting the former-US Airways flights as part of American Airlines, and it was probably during the June or July timeframe.)

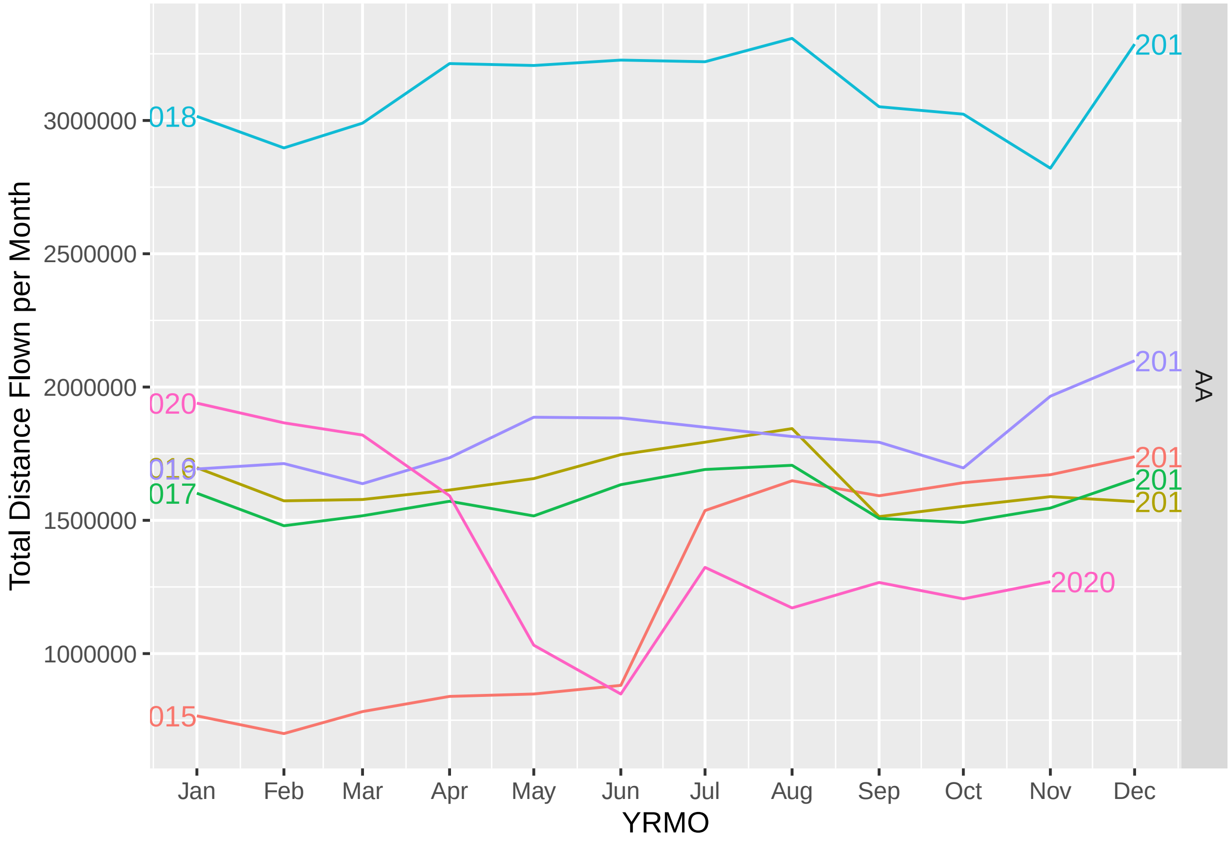
[https://en.wikipedia.org/wiki/History_of_American_Airlines]

I was interested to see what happened with the other metrics for AA during summer 2015.

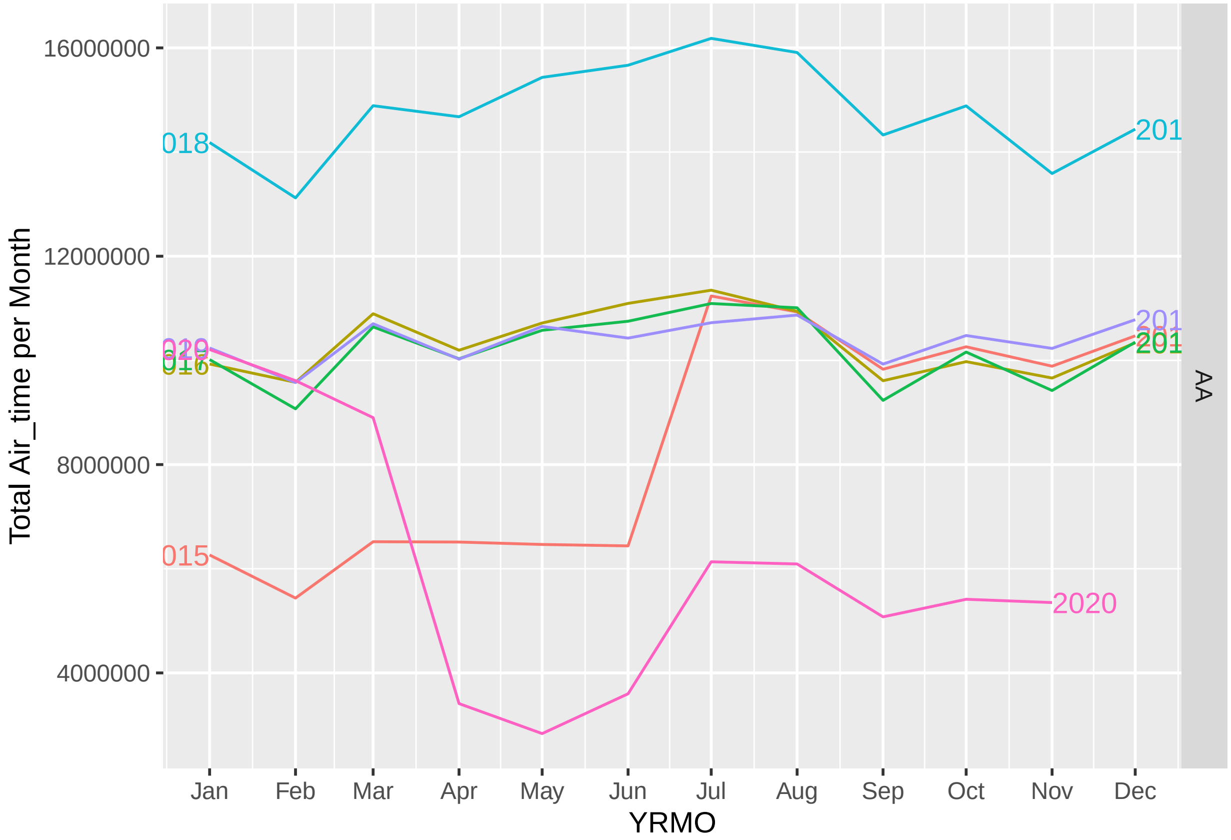
The first chart depicts "Total Distance flown" per month during 2015.

The second chart depicts "Total air time flown" per month during 2015.

```
big3 %>%
  filter(UNIQUE_CARRIER == "AA") %>%
  feasts::gg_season(TotalDistance, labels = "both") + ylab("Total Distance Flown per Month")
```



```
big3 %>%
  filter(UNIQUE_CARRIER == "AA") %>%
  feasts::gg_season(TotalAIR_TIME, labels = "both") + ylab("Total Air_time per Month")
```

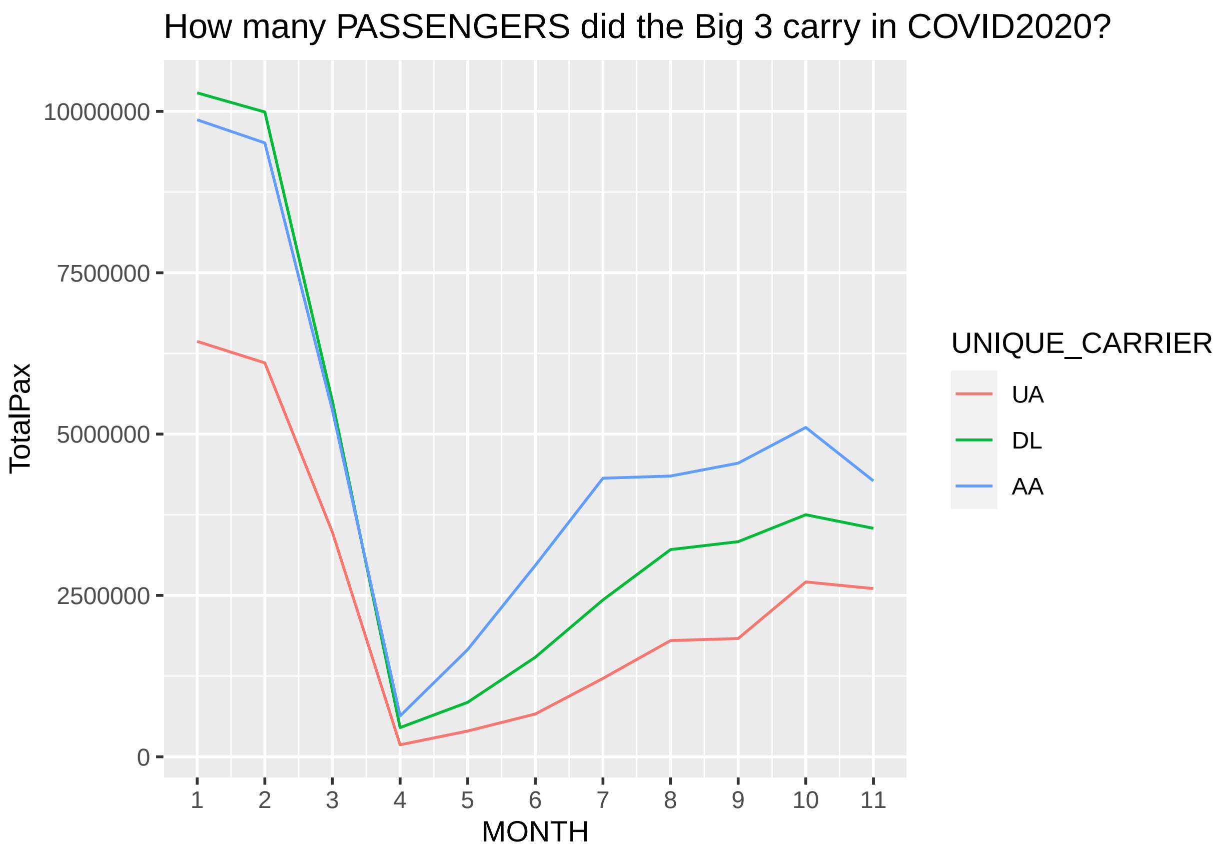


Sure enough, American Airlines experienced a huge jump in summer '15, to where Distance and Air Time flown became consistent with 2016-2019 levels. This leads me to believe that 2015 was a year in which external factors were at play – important to note but probably will not play into my analysis any further.

More directly to the questions I'm trying to answer:

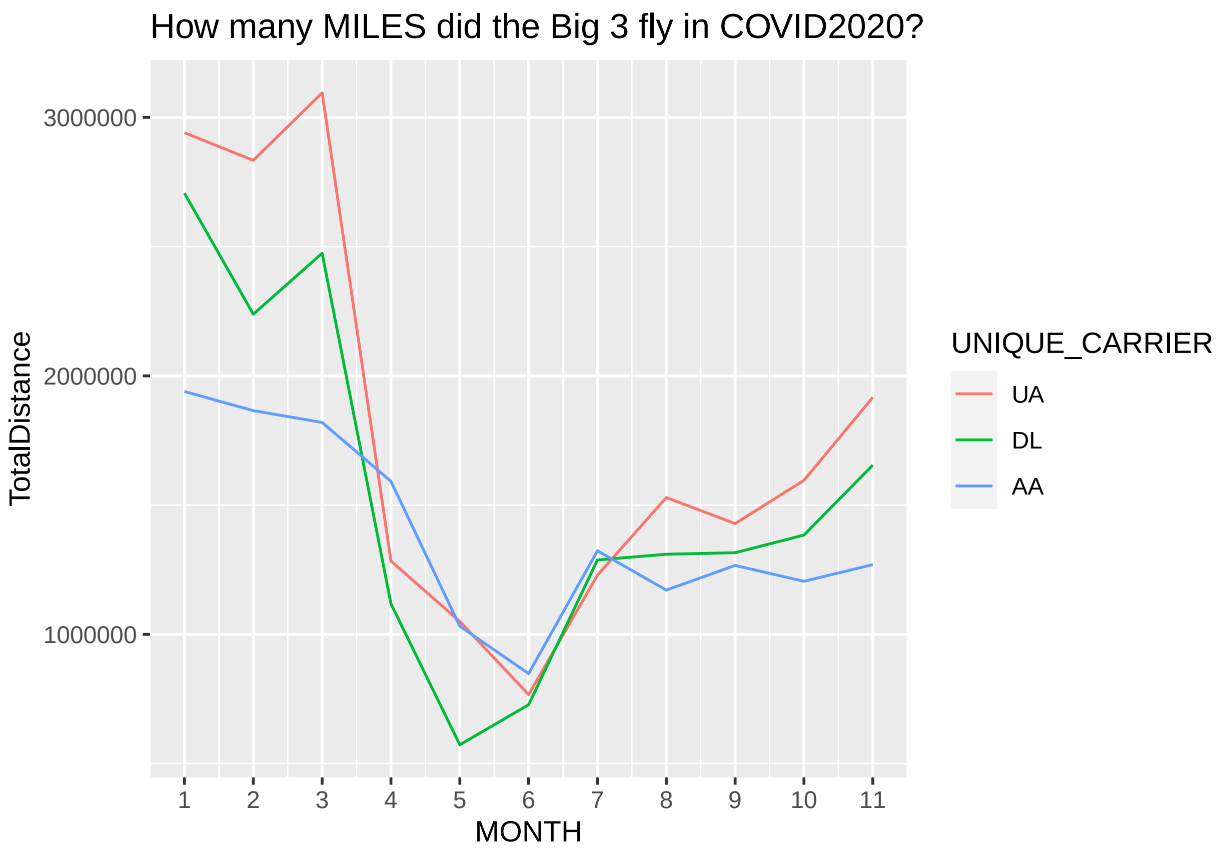
Another aspect of these gg_season plots that I see is the stabilization of Passengers flown as the year 2020 proceeded. The latest data in this set is from November 2020. We can see below how each airline performed through November 2020. Their monthly passenger numbers were essentially cut to near 500,000 by April 2020, and as of November they did climb back towards 5 million (with American Airlines moving the most pax). This is roughly half the pre-COVID19 passenger count for these airlines. The question we can address is how long it may take to recover to pre-COVID19 numbers.

```
#Show total passengers trend since 2020 Jan
filter(big3, YEAR == "2020") -> big32020
big32020 %>%
  ggplot(mapping = aes(x = MONTH, y = TotalPax, color = UNIQUE_CARRIER)) + geom_line() + ggtitle("How ma
```



This one shows a similar story, in terms of Distance Flown over 2020.

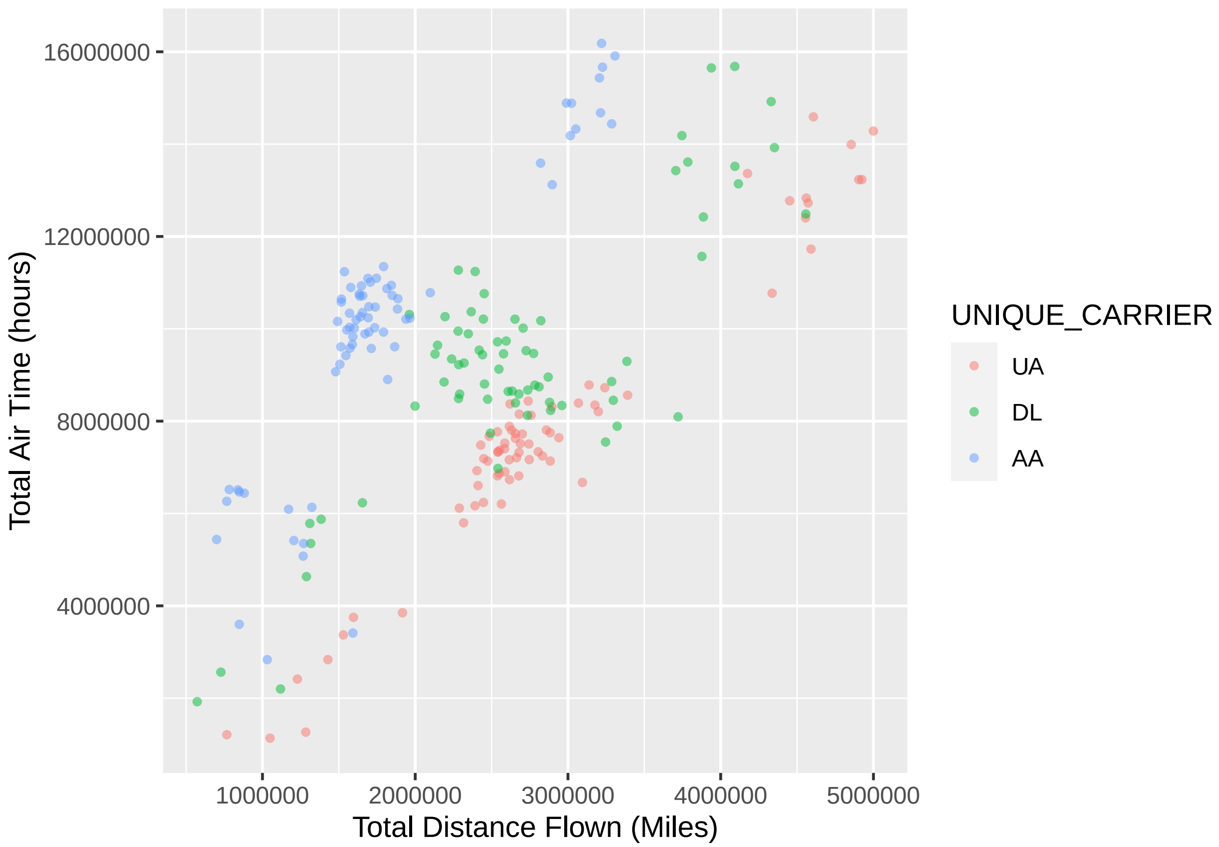
```
big32020 %>%
  ggplot(mapping = aes(x = MONTH, y = TotalDistance, color = UNIQUE_CARRIER)) + geom_line() + ggtitle("I
```



I also wanted to plot some possibly related variables in a scatter plot, to see if there was anything to discover.

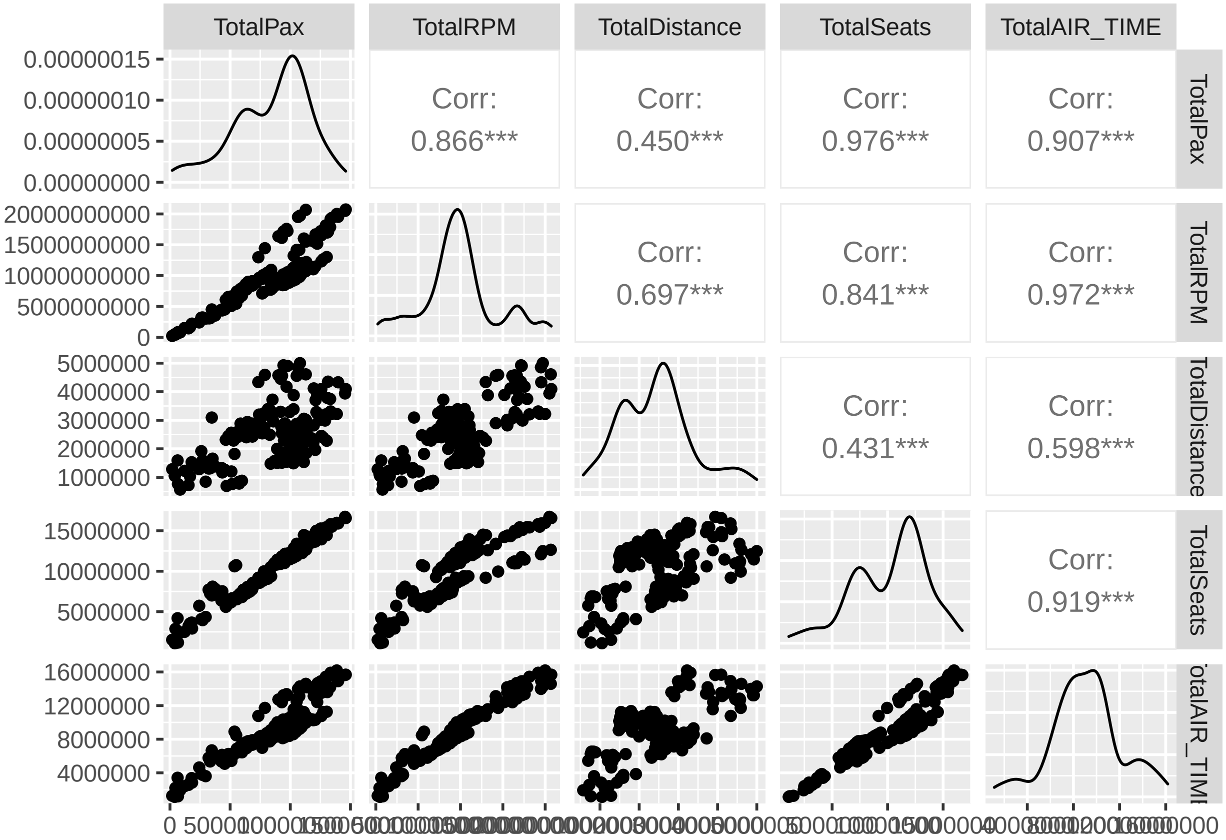
Below is Air Time (hours spent in flight for the month) vs. Distance Flown (miles). Of course, as distance flown increases, we expect to see an increase in air time, but is it linear?

```
#Is there a relationship between Total Air_Time and Distance Flown?
big3 %>%
  ggplot(aes(x = TotalDistance, y = TotalAIR_TIME))+
  geom_point(size=1, aes(colour = UNIQUE_CARRIER), alpha =0.5)+
  labs(y = "Total Air Time (hours)", x = "Total Distance Flown (Miles)")
```



```
timedistancemodel <- lm(TotalAIR_TIME ~ TotalDistance, data = big3)
summary(timedistancemodel)
```

```
##
## Call:
## lm(formula = TotalAIR_TIME ~ TotalDistance, data = big3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5594531 -1910676  -276750  2167982  5627942
##
## Coefficients:
##                   Estimate   Std. Error t value Pr(>|t|)
## (Intercept)   4409794.1067  462991.2702   9.525   <2e-16 ***
## TotalDistance       1.9082       0.1763  10.826   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2424000 on 211 degrees of freedom
## Multiple R-squared:  0.3571, Adjusted R-squared:  0.3541
## F-statistic: 117.2 on 1 and 211 DF,  p-value: < 2.2e-16
```



The Ggally relationship plots above show correlation amongst features. Highly linear relationships seem to exist between many of these.