Investigate flight performance from 2008-2018

March 11, 2019

INTRODUCTION

In this project, we will study the flight performance of all U.S. domestic carriers and build a linear regression to predict number of delays. The dataset for this study contains information of all U.S. carriers regarding flight delays and performace. This dataset was obtained from the Bureau of Transportation Statistics, which includes data collected from December 2008 to December 2018. An data exploratory study will be focused on the top ten carriers with the largest number of on-time flights and top ten airlines with the largest number of delayed flights caused by different reasons. Finally, we will predict number of delays using linear regression model.

DATA EXPLORATORY SECTION

```r
#Import dataset as a dataframe
flight <- read.csv("airline_delay_causes.csv", header=T, check.names=F)
```

Data Structure

The dataset includes 26 different airline carriers with 21 different variables with 155,317 observations for each variable.

```r
#Get a summary of datatype and data info using summary and str
str(flight)
```

```
## 'data.frame':    155317 obs. of  21 variables:
##  $ year               : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
##  $ month              : int  12 12 12 12 12 12 12 12 12 12 ...
##  $ carrier            : Factor w/ 23 levels "9E","AA","AS",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ carrier_name       : Factor w/ 26 levels "AirTran Airways Corporation",..: 18 18 18 18 18 18 18 18
##  $ airport            : Factor w/ 382 levels "ABE","ABI","ABQ",..: 1 16 22 23 24 25 26 28 29 33 ...
##  $ airport_name       : Factor w/ 382 levels "Aberdeen, SD: Aberdeen Regional",..: 10 367 19 14 22 16
##  $ arr_flights        : int  81 27 888 91 128 91 59 79 54 59 ...
##  $ arr_del15          : int  26 8 352 35 33 31 22 34 5 18 ...
##  $ carrier_ct         : num  8.5 2.93 55.12 14.65 9.92 ...
##  $ weather_ct         : num  2.29 0.16 8.77 0 2.19 0 1.37 2.1 0 1 ...
##  $ nas_ct             : num  10.9 4.91 164.03 15.49 16.56 ...
##  $ security_ct        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ late_aircraft_ct   : num  4.3 0 124.08 4.86 4.32 ...
##  $ arr_cancelled      : int  5 10 22 5 5 3 0 13 3 6 ...
##  $ arr_diverted       : int  0 0 0 3 0 2 0 0 0 0 ...
##  $ arr_delay          : int  1729 472 19902 1853 1607 2107 2112 1896 220 1127 ...
##  $ carrier_delay      : int  308 141 4775 908 525 1283 360 829 51 543 ...
##  $ weather_delay      : int  409 9 972 0 129 0 72 55 0 37 ...
##  $ nas_delay          : int  514 322 7263 531 721 444 385 652 79 357 ...
##  $ security_delay     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ late_aircraft_delay: int  498 0 6892 414 232 380 1295 360 90 190 ...
```

```r
summary(flight)
```

```
##       year          month           carrier     
##  Min.   :2008   Min.   : 1.000   OO     :20553  
##  1st Qu.:2011   1st Qu.: 4.000   EV     :17045  
##  Median :2013   Median : 7.000   DL     :15742  
##  Mean   :2013   Mean   : 6.564   MQ     :12267  
##  3rd Qu.:2016   3rd Qu.:10.000   AA     :10240  
##  Max.   :2018   Max.   :12.000   UA     : 9893  
##                                  (Other):69577  
##                   carrier_name      airport      
##  SkyWest Airlines Inc.   :20553   LAX    :  1538  
##  ExpressJet Airlines Inc.:16792   DTW    :  1499  
##  Delta Air Lines Inc.    :15742   ATL    :  1494  
##  American Airlines Inc.  :10240   MSY    :  1487  
##  United Air Lines Inc.   : 9893   PHX    :  1486  
##  Southwest Airlines Co.  : 9630   DCA    :  1484  
##  (Other)                 :72467   (Other):146329  
##                                                  airport_name   
##  Los Angeles, CA: Los Angeles International               :  1538  
##  Detroit, MI: Detroit Metro Wayne County                  :  1499  
##  Atlanta, GA: Hartsfield-Jackson Atlanta International    :  1494  
##  New Orleans, LA: Louis Armstrong New Orleans International:  1487  
##  Phoenix, AZ: Phoenix Sky Harbor International            :  1486  
##  Washington, DC: Ronald Reagan Washington National        :  1484  
##  (Other)                                                  :146329  
##   arr_flights        arr_del15         carrier_ct        weather_ct    
##  Min.   :    1.0   Min.   :   0.00   Min.   :   0.00   Min.   :  0.00  
##  1st Qu.:   60.0   1st Qu.:   9.00   1st Qu.:   3.12   1st Qu.:  0.00  
##  Median :  123.0   Median :  23.00   Median :   8.13   Median :  0.48  
##  Mean   :  400.6   Mean   :  74.84   Mean   :  21.24   Mean   :  2.31  
##  3rd Qu.:  289.0   3rd Qu.:  57.00   3rd Qu.:  19.62   3rd Qu.:  1.93  
##  Max.   :21977.0   Max.   :5268.00   Max.   :1242.16   Max.   :298.62  
##  NA's   :192       NA's   :226       NA's   :192       NA's   :192     
##      nas_ct         security_ct      late_aircraft_ct  arr_cancelled    
##  Min.   :   0.00   Min.   : 0.0000   Min.   :   0.00   Min.   :   0.000  
##  1st Qu.:   1.98   1st Qu.: 0.0000   1st Qu.:   2.11   1st Qu.:   0.000  
##  Median :   5.66   Median : 0.0000   Median :   6.79   Median :   1.000  
##  Mean   :  23.71   Mean   : 0.1424   Mean   :  27.41   Mean   :   6.403  
##  3rd Qu.:  15.33   3rd Qu.: 0.0000   3rd Qu.:  18.54   3rd Qu.:   4.000  
##  Max.   :2401.79   Max.   :19.5300   Max.   :1885.47   Max.   :1389.000  
##  NA's   :192       NA's   :192       NA's   :192       NA's   :192       
##   arr_diverted         arr_delay      carrier_delay    weather_delay    
##  Min.   :  0.0000   Min.   :     0   Min.   :     0   Min.   :    0.0  
##  1st Qu.:  0.0000   1st Qu.:   463   1st Qu.:   155   1st Qu.:    0.0  
##  Median :  0.0000   Median :  1232   Median :   445   Median :   21.0  
##  Mean   :  0.9549   Mean   :  4361   Mean   :  1328   Mean   :  200.2  
##  3rd Qu.:  1.0000   3rd Qu.:  3189   3rd Qu.:  1142   3rd Qu.:  146.0  
##  Max.   :256.0000   Max.   :429194   Max.   :196944   Max.   :31960.0  
##  NA's   :192        NA's   :192      NA's   :192      NA's   :192      
##    nas_delay      security_delay     late_aircraft_delay
##  Min.   :     0   Min.   :   0.000   Min.   :     0     
##  1st Qu.:    65   1st Qu.:   0.000   1st Qu.:   113     
```

```
##  Median :   208   Median :   0.000   Median :   413
##  Mean   :  1083   Mean   :   5.837   Mean   :  1745
##  3rd Qu.:   600   3rd Qu.:   0.000   3rd Qu.:  1205
##  Max.   :137443   Max.   :2897.000   Max.   :148181
##  NA's   :192      NA's   :192        NA's   :192
```

```r
#Inspect the structure of the data using head(data)
head(flight)
```

```
##   year month carrier           carrier_name airport
## 1 2008    12      9E Pinnacle Airlines Inc.     ABE
## 2 2008    12      9E Pinnacle Airlines Inc.     ALO
## 3 2008    12      9E Pinnacle Airlines Inc.     ATL
## 4 2008    12      9E Pinnacle Airlines Inc.     ATW
## 5 2008    12      9E Pinnacle Airlines Inc.     AUS
## 6 2008    12      9E Pinnacle Airlines Inc.     AVL
##                                                airport_name arr_flights
## 1 Allentown/Bethlehem/Easton, PA: Lehigh Valley International          81
## 2                            Waterloo, IA: Waterloo Regional          27
## 3       Atlanta, GA: Hartsfield-Jackson Atlanta International         888
## 4                        Appleton, WI: Appleton International          91
## 5                 Austin, TX: Austin - Bergstrom International         128
## 6                          Asheville, NC: Asheville Regional          91
##   arr_del15 carrier_ct weather_ct nas_ct security_ct late_aircraft_ct
## 1        26       8.50       2.29  10.90           0             4.30
## 2         8       2.93       0.16   4.91           0             0.00
## 3       352      55.12       8.77 164.03           0           124.08
## 4        35      14.65       0.00  15.49           0             4.86
## 5        33       9.92       2.19  16.56           0             4.32
## 6        31      12.25       0.00  12.30           0             6.46
##   arr_cancelled arr_diverted arr_delay carrier_delay weather_delay
## 1             5            0      1729           308           409
## 2            10            0       472           141             9
## 3            22            0     19902          4775           972
## 4             5            3      1853           908             0
## 5             5            0      1607           525           129
## 6             3            2      2107          1283             0
##   nas_delay security_delay late_aircraft_delay
## 1       514              0                 498
## 2       322              0                   0
## 3      7263              0                6892
## 4       531              0                 414
## 5       721              0                 232
## 6       444              0                 380
```

```r
#Remove Column with NA values
flight <- flight[,colSums(is.na(flight))<nrow(flight)]
```

```r
#Check how many carriers in this dataset
print(paste("There are", length(unique(flight$carrier_name)), "carriers in this dataset."))
```

```
## [1] "There are 26 carriers in this dataset."
```

```r
###Load necessary packages for data exploration and analysis###
require(ggplot2)
require(grid)
require(scales)
require(dplyr)
require(gridExtra)
library(RColorBrewer)
library(ggthemes)
library(ggrepel)
library(rmarkdown)
library(knitr)
```

Generate New Summary Dataset

The chunk below will produce a new summary dataframe, which includes the information regarding the total number of arrivals, delayed flights, cancelled flights, and on-time flights that each carrier has by year.

```r
flight_summary <- flight %>%
  group_by(year, carrier_name) %>%
  summarize(arrivals = sum(arr_flights),
            delayed = sum(arr_del15),
            cancelled = sum(arr_cancelled),
            diverted = sum(arr_diverted)) %>%
  transform(on_time = 1 - delayed/arrivals) %>%
  transform(delayed_percent = delayed/arrivals)
```

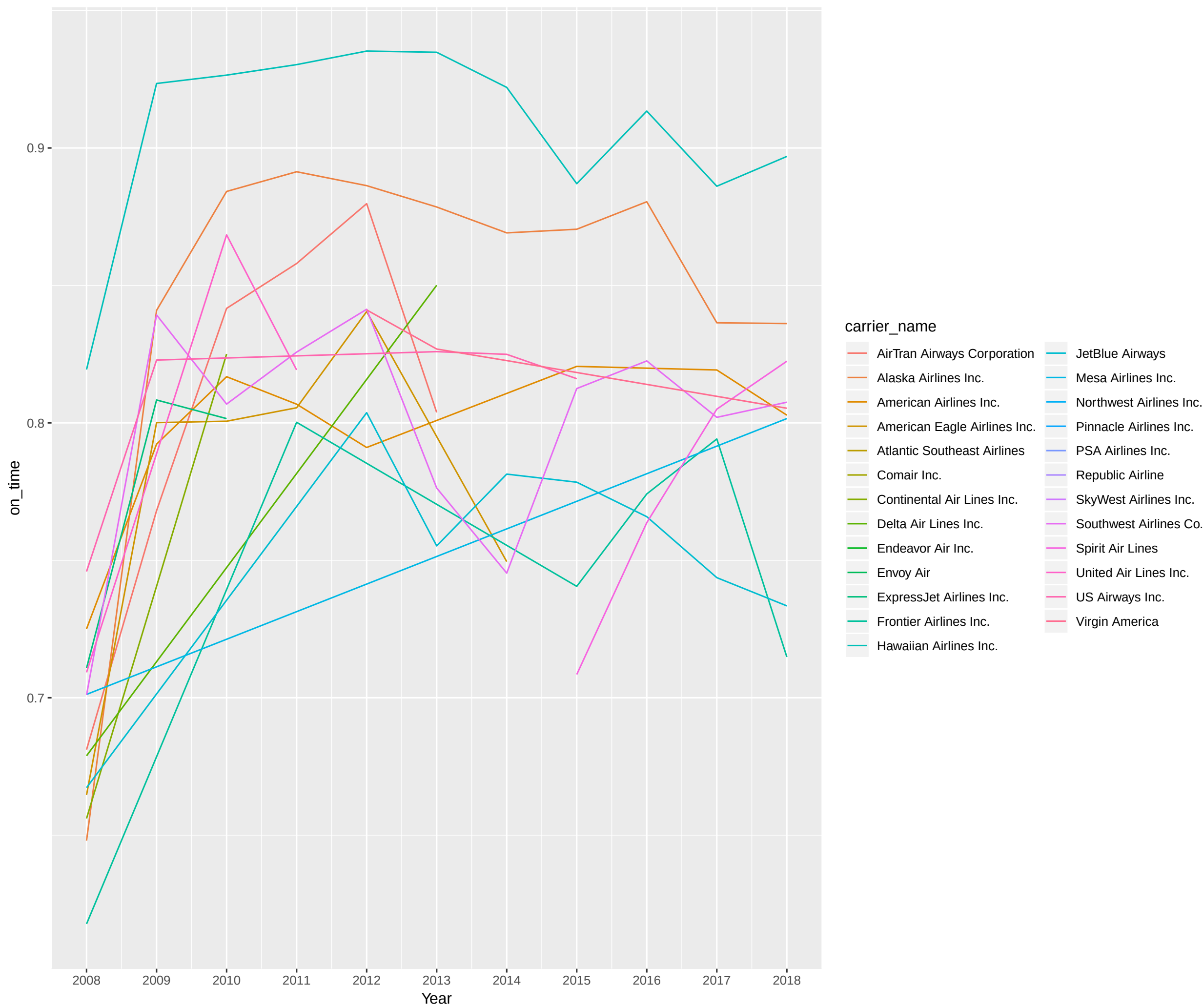
Then, we can remove all rows with NA values since they do not have any information for evaluation.

```r
#Remove all row with NA values from the flight_summary dataframe
flight_summary <- na.omit(flight_summary)

head(flight_summary)
```

```
##   year                 carrier_name arrivals delayed cancelled diverted
## 1 2008  AirTran Airways Corporation    20628    6578       292       58
## 2 2008         Alaska Airlines Inc.    11330    3988       627       95
## 3 2008       American Airlines Inc.    47329   13013      1058      193
## 4 2008 American Eagle Airlines Inc.    35024   11746      2362      141
## 5 2008  Atlantic Southeast Airlines    23474    8017       788       92
## 6 2008                 Comair Inc.    13711    5247       863       53
##     on_time delayed_percent
## 1 0.6811131       0.3188869
## 2 0.6480141       0.3519859
## 3 0.7250523       0.2749477
## 4 0.6646300       0.3353700
## 5 0.6584732       0.3415268
## 6 0.6173146       0.3826854
```

Line Plot by Year for Each Carrier

```r
ggplot(data = flight_summary, aes(x=year, y=on_time))+
  scale_x_continuous(name = "Year",
                     breaks = seq(2008, 2018, 1))+
  geom_line(aes(color=carrier_name))
```



The above line plot looks very busy and hard to follow, so we will only focus on primarily two sets of top ten airlines: + Top 10 airlines that have the average largest number of delayed flights in the past 10 years + Top 10 airlines that have the average largest number of on-time flights in the past 10 years

Beyond that, we will also focus on evaluating the performance of the top ten airlines with the largest number of delayed flights in the past 10 years (2008-2018). we will make a new summary table which includes the average number of arrivals, cancelled flights, diverted flights, delayed flights for each carrier in the last 10 years.

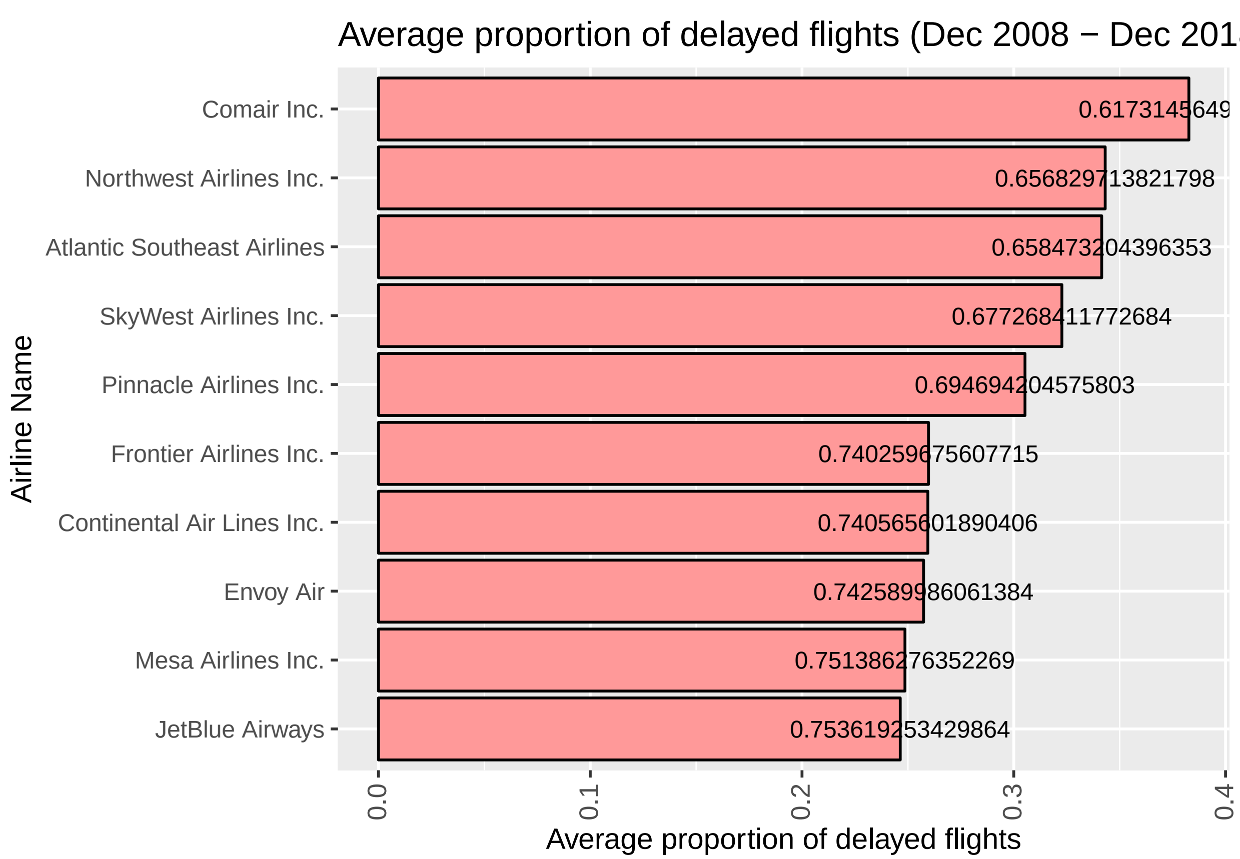
```r
#Make new dataset that includes the average number of arrivals
#delayed flights, cancelled flights, and diverted flights and the proportion of
#on_time flights in the last 10 years by carrier

flight_summary_average <- flight_summary %>%
  group_by(carrier_name) %>%
```

```r
  summarize(ave_arrivals = mean(arrivals),
            ave_delayed = mean(delayed_percent),
            ave_cancelled = mean(cancelled),
            ave_diverted = mean(diverted),
            ave_ontime = mean(on_time))
```

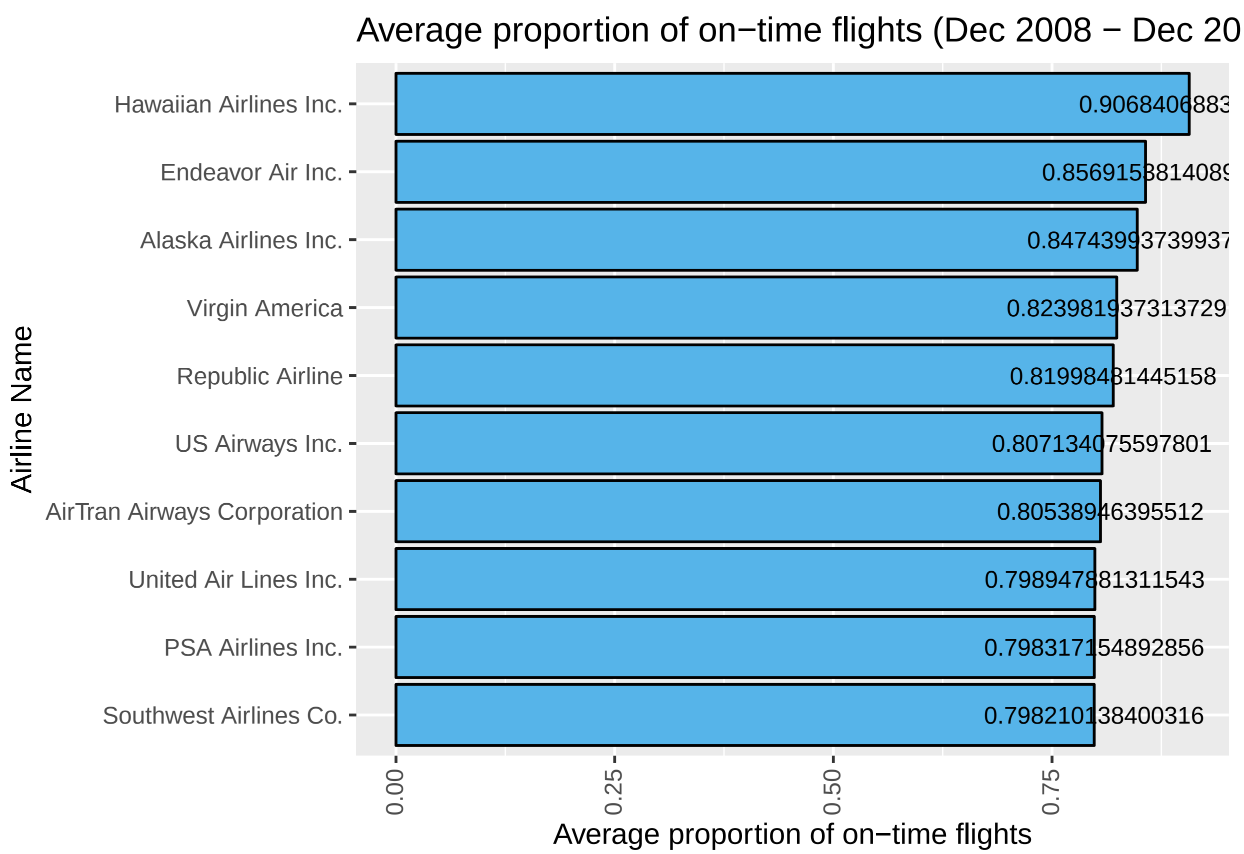
Bar Plots for Top Airlines

```r
top_ten_delayed <- flight_summary_average%>%
  arrange(desc(ave_delayed))%>%
  top_n(10, ave_delayed)

top_ten_ontime <- flight_summary_average%>%
  arrange((desc(ave_ontime)))%>%
  top_n(10, ave_ontime)
```

```r
#Average proportion of delayed flights from Dec 2008 to Dec 2018
ggplot(data = top_ten_delayed, aes(x=reorder(carrier_name,ave_delayed), ave_delayed))+
 geom_bar(stat = 'identity', position = 'dodge', fill="#FF9999", colour="black")+
 geom_text(mapping = aes(label = ave_ontime), size = 3) +
 theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust=.4,size=10))+
 labs(x="Airline Name", y='Average proportion of delayed flights')+
 ggtitle("Average proportion of delayed flights (Dec 2008 - Dec 2018)") + coord_flip()
```



```r
#Average proportion of on-time flights from Dec 2008 to Dec 2018
ggplot(data = top_ten_ontime, aes(x=reorder(carrier_name, ave_ontime), y=ave_ontime))+
   geom_bar(stat = 'identity', position = 'dodge', fill="#56B4E9", colour="black")+
   geom_text(mapping = aes(label = ave_ontime), size = 3) +
   labs(x='Airline Name', y='Average proportion of on-time flights')+
   theme(axis.text.x = element_text(angle = 90, hjust=1, vjust=.4))+
   ggtitle('Average proportion of on-time flights (Dec 2008 - Dec 2018)')+ coord_flip()
```

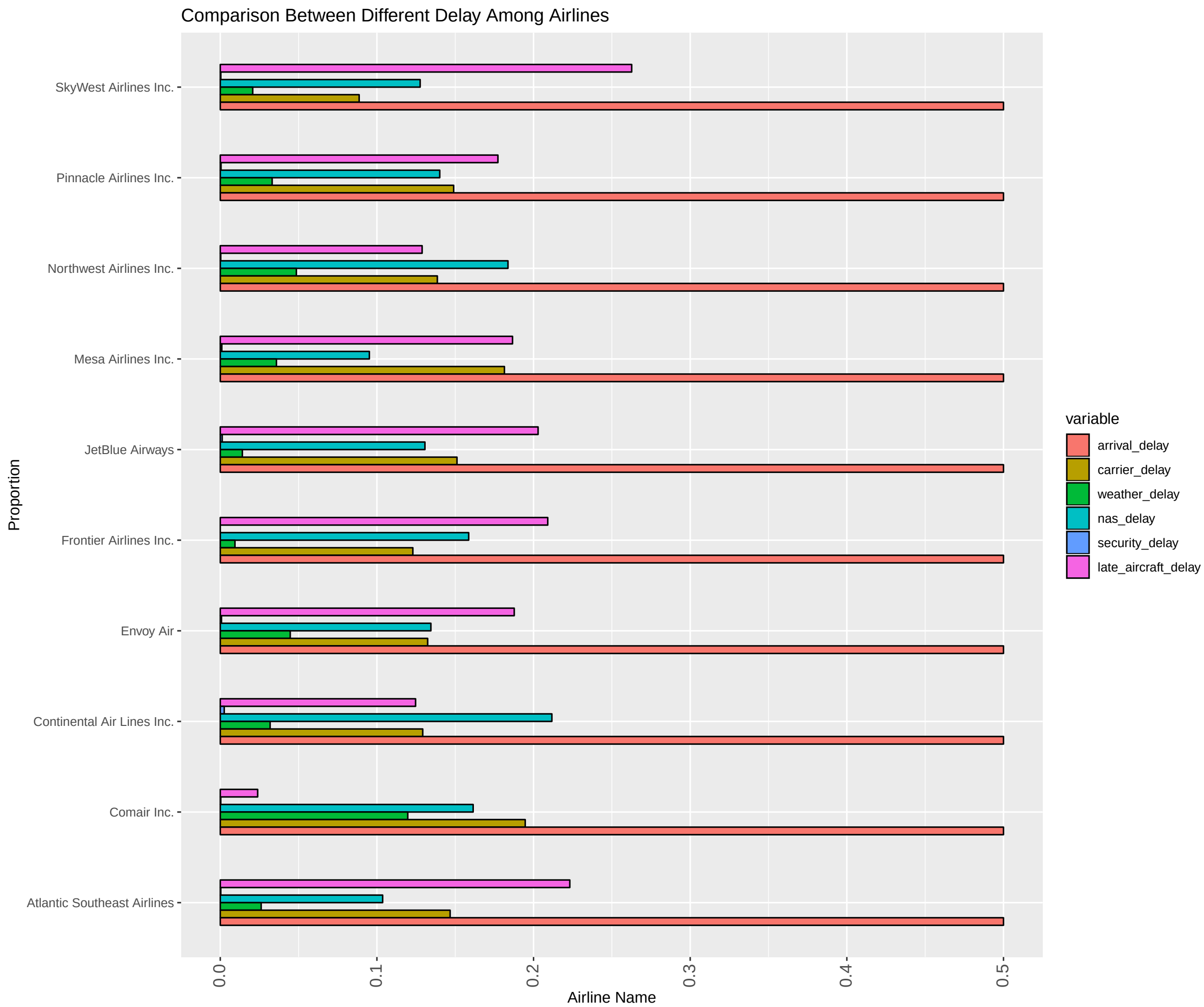


Comair Inc. airlines has the highest average number of delayed flights from 2008-2018. Hawaiian Airlines has high on-time proportion of 90.7%.

Finding The Most Common Delay Cause

```r
#Subseting all data point related to airlines belonging to the top_ten_delay list
subset_delay <- filter(flight,
                        carrier_name %in%
                           top_ten_delayed[['carrier_name']])

#Create new summary for each delay cause for each carrier
delay_summary <- subset_delay %>%
  group_by(carrier_name, year) %>%
  summarize(arr_delay = sum(`arr_delay`),
            carrier_delay = sum(`carrier_delay`),
            weather_delay = sum(weather_delay),
```

```r
            nas_delay = sum(nas_delay),
            security_delay = sum(security_delay),
            late_aircraft_delay = sum(late_aircraft_delay),
            sum_delay = sum(`arr_delay`,`carrier_delay`,weather_delay,
                           nas_delay,security_delay,late_aircraft_delay))%>%
  transform(arr_delay_per = arr_delay/sum_delay,
            carrier_delay_per = carrier_delay/sum_delay,
            weather_delay_per = weather_delay/sum_delay,
            nas_delay_per = nas_delay/sum_delay,
            security_delay_per = security_delay/sum_delay,
            late_aircraft_delay_per = late_aircraft_delay/sum_delay)

#Remove NA rows from delay_summary dataset
delay_summary <- na.omit(delay_summary)

#Calcuate the average number of delayed flight by each category from 2005-2017
average_delay_summary <- delay_summary %>%
  group_by(carrier_name)%>%
  summarize(arrival_delay = mean(arr_delay_per),
            carrier_delay = mean(carrier_delay_per),
            weather_delay = mean(weather_delay_per),
            nas_delay = mean(nas_delay_per),
            security_delay = mean(security_delay_per),
            late_aircraft_delay = mean(late_aircraft_delay_per))

#Create grouped bar plots
library(reshape2)

average_delay_summary <- data.frame(average_delay_summary)
average_delay_summary <- melt(average_delay_summary,
                              id.vars = "carrier_name")

ggplot(data = average_delay_summary,
       aes(x=carrier_name, y=value, fill=variable,width=.5))+
  geom_bar(stat = 'identity',
           colour="black",
           width = 2,
           position = 'dodge',
           aes(color = variable))+
  theme(axis.text.x = element_text(angle = 90, hjust=1, vjust=.4, size=12))+
  ggtitle("Comparison Between Different Delay Among Airlines")+
  labs(x="Proportion",y="Airline Name") + coord_flip()
```



It appears that arrival delay and late aircraft delay are the two most common cause among these top ten delayed airlines. Weather does not have a severe impact on delay for all airlines.

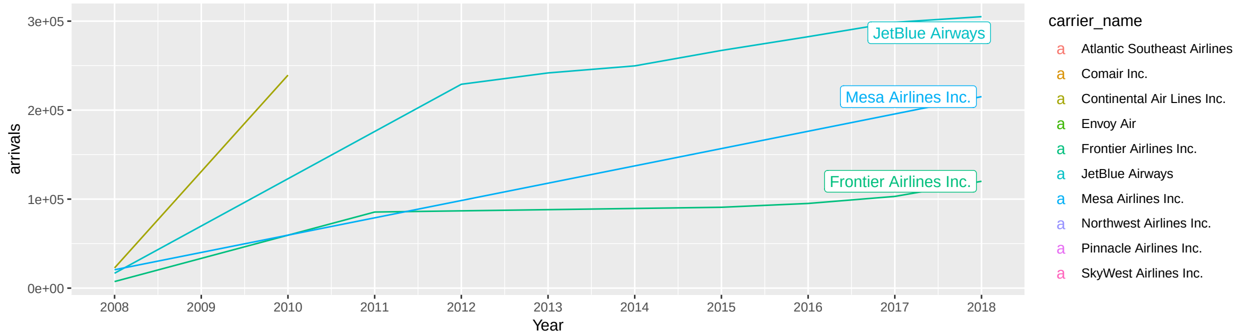
Evaluating The Performance of all Arlines in the `top_ten_delay` List

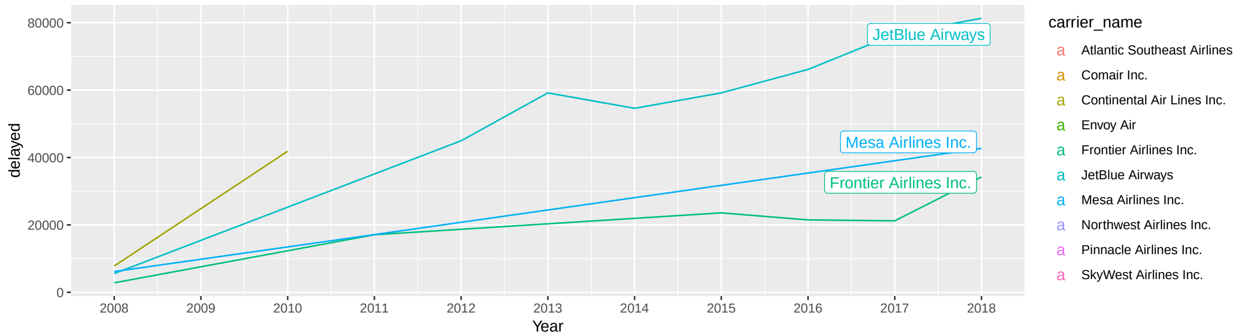
To further evaluate the performance of all airlines in the `top_ten_delayed` list, I have subset all their info from the `flight_summary dataset` before generating any plots for the analysis.

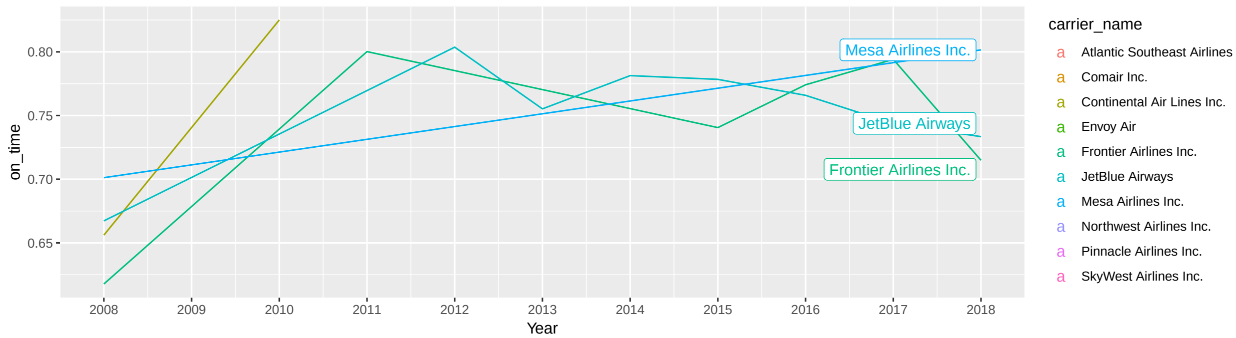
```
#Subseting info for the top ten airlines with the highest average number of
#delayed flights
filter(flight_summary,carrier_name %in% top_ten_delayed[['carrier_name']]) %>%
  head(5) %>% knitr::kable()
```

year	carrier_name	arrivals	delayed	cancelled	diverted	on_time	delayed_percent
2008	Atlantic Southeast Airlines	23474	8017	788	92	0.6584732	0.3415268
2008	Comair Inc.	13711	5247	863	53	0.6173146	0.3826854
2008	Continental Air Lines Inc.	22691	7804	436	83	0.6560751	0.3439249
2008	Frontier Airlines Inc.	7366	2816	71	5	0.6177030	0.3822970
2008	JetBlue Airways	16707	5559	485	177	0.6672652	0.3327348

```r
p1 <- filter(flight_summary,carrier_name %in% top_ten_delayed[['carrier_name']]) %>%
  mutate(label = if_else(year == max(year), as.character(carrier_name), NA_character_)) %>%
  ggplot(aes(x=year, y=arrivals, color=carrier_name))+
  scale_x_continuous(name = "Year",
                     breaks = seq(2008, 2018, 1))+
  geom_line()+
  geom_label_repel(aes(label = label), nudge_x = 0.5, na.rm = TRUE)

p2 <- filter(flight_summary,carrier_name %in% top_ten_delayed[['carrier_name']]) %>%
  mutate(label = if_else(year == max(year), as.character(carrier_name), NA_character_)) %>%
  ggplot(aes(x=year, y=delayed, color=carrier_name))+
  scale_x_continuous(name = "Year",
                     breaks = seq(2008, 2018, 1))+
  geom_line()+
  geom_label_repel(aes(label = label), nudge_x = 0.5, na.rm = TRUE)

p3 <- filter(flight_summary,carrier_name %in% top_ten_delayed[['carrier_name']]) %>%
  mutate(label = if_else(year == max(year), as.character(carrier_name), NA_character_)) %>%
  ggplot(aes(x=year, y=on_time, color=carrier_name))+
  scale_x_continuous(name = "Year",
                     breaks = seq(2008, 2018, 1))+
  geom_line()+
  geom_label_repel(aes(label = label), nudge_x = 0.5, na.rm = TRUE)

grid.arrange(p1, p2, p3)
```







We don't have a lot of information regarding these airlines and thus final conclusion cannot be draw from here. However, this missing-data fact could potentially explain their performance since it could mean that these airlines are still at their early stage of development. It's also noticeable that Mesa Airlines has increasing number of delays relative to increasing number of on-time flights. It's in opposite with JetBlue Airways and Frontier Airlines, which have rapidly decresing number of on-time flights in recent year with increasing number of flights.

```
##Create new CSV
write.csv(top_ten_delayed, file="top_ten_delay.csv", row.names = FALSE)
write.csv(top_ten_ontime, file="top_ten_on_time.csv", row.names = FALSE)
write.csv(average_delay_summary, file="most_common_delay_cause.csv", row.names = FALSE)
```

BUILD LINEAR REGRESSION MODEL

We randomly split it into a training set (70% of the data) and testing set (30% of the data). Since our dependent variable is a continuous one, we cannot use the sample.split function.

```
#Copy the data
Airlines <- flight
```

```r
# Train set (%70) and test set (30%)
set.seed(15071)
spl <- sample(nrow(Airlines), 0.7*nrow(Airlines))
AirlinesTrain <- Airlines[spl, ]
temp <- Airlines[-spl, ]

#Ensure trainset contains carrier_name and airport from testset
#with no NA values
temp <- temp%>%
  semi_join(AirlinesTrain, by = "carrier_name")%>%
  semi_join(AirlinesTrain, by = "airport")%>%
  na.omit(cols = c("arr_del15"))

AirlinesTest <- temp
rm(temp)
```

Linear regression

Build a linear regression model to predict `arr_del15` variable (total number of delays) using all of the other variables as independent variables.

```r
#train model using all of the other variables as independent variables
delayLR <- lm(arr_del15 ~., data = AirlinesTrain)
```

The Residual Mean Squared Error (RMSE)

The RMSE is then defined as below, with N being the number of samples and the sum occurring over all these combinations. This number in our case should be less than 1.

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(\hat{y}_{u,i} - y_{u,i})^2}$$

```r
#The root mean squared error
RMSE <- function(true_values, predicted_values){
            sqrt(mean((true_values - predicted_values)^2))}
```

The R-squared and Adjusted R-squared

R-squared is the proportional improvement in prediction from the regression model, compared to the mean model. It indicates the goodness of fit of the model. R-squared takes values from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

Adjusted R-squared incorporates the model's degrees of freedom. It will increase as predictors are added if the increase in model fit is worthwhile. It is interpreted as the proportion of total variance that is explained by the model.

Prediction on Test Set

Using the function `predict` to predict the total number of delays. Then, we can calculate the R-squared (which is better when higher) values that tell us how reliable our model is.

```r
#Predict number of delays on Test set
delayLRpred <- predict.lm(delayLR, newdata = AirlinesTest)
```

```r
#Get R-squared
Rsquared <-summary(delayLR)$r.squared

#Get Adjusted R-squared
AdjRsquared <-summary(delayLR)$adj.r.squared

# Results for RMSE, R-squared, Adj R-squared
rmse <- RMSE(AirlinesTest$arr_del15, delayLRpred)
results <- tibble(statistics = "RMSE", results = rmse)
results <- bind_rows(results, tibble(statistics="R-squared",
                                     results = Rsquared ))
results <- bind_rows(results, tibble(statistics="Adjusted R-squared",
                                      results = AdjRsquared ))

results %>% knitr::kable()
```

statistics	results
RMSE	0.0050841
R-squared	1.0000000
Adjusted R-squared	1.0000000

We have R-squared and Adjusted R-squared of 1 and RMSE of 0.0050841, which means that this linear regression has high accuracy. Below, we can compare predicted number of delays for each airline in each year with the known number of delays. Even though this practice must be avoided for any bias, but the purpose of this comparison is for us to have a picture of how the prediction work.

```r
#combine Test set and predicted number of delays
output <- cbind(AirlinesTest, delayLRpred)

#Save file
write.csv(output,"Flights_with_predicted_delays.csv", na = "", row.names=FALSE)

#Show top 15 preditec scores
output %>% select(c(year ,carrier_name,airport,arr_del15,delayLRpred)) %>%
  head(15) %>% knitr::kable()
```

year	carrier_name	airport	arr_del15	delayLRpred
2008	Pinnacle Airlines Inc.	AUS	33	32.989849
2008	Pinnacle Airlines Inc.	BGM	18	18.010279
2008	Pinnacle Airlines Inc.	BGR	21	20.989385
2008	Pinnacle Airlines Inc.	BNA	66	66.009972
2008	Pinnacle Airlines Inc.	BOS	28	27.999904
2008	Pinnacle Airlines Inc.	CAE	30	29.999891
2008	Pinnacle Airlines Inc.	CHA	23	23.000265
2008	Pinnacle Airlines Inc.	CMX	12	11.999954
2008	Pinnacle Airlines Inc.	DEN	39	39.009701
2008	Pinnacle Airlines Inc.	EVV	48	47.989727

year	carrier_name	airport	arr_del15	delayLRpred
2008	Pinnacle Airlines Inc.	FLL	7	6.999462
2008	Pinnacle Airlines Inc.	FSM	7	6.999806
2008	Pinnacle Airlines Inc.	FWA	55	54.999920
2008	Pinnacle Airlines Inc.	GFK	9	8.999977
2008	Pinnacle Airlines Inc.	GRB	25	24.999519

CONCLUSION

Through this data analysis, the following points can be made for the flight performance of year period from Dec 2008 to Dec 2018:

- The top five airlines with the largest proportion of on-time flights are: Hawaiian Airlines, Endeavor Air, Alaska Airlines, Virgin America, Republic Airline.
- The top five airlines with the largest proportion of delayed flights are: Comair , Northwest Airlines, Atlantic Southeast Airlines, SkyWest Airlines, Pinnacle Airlines.

We were able to predict the number of delays with high accuracy (RMSE=0.0050841, R-squared=Adj. R-squared==1) with linear regression model using all of the other variables as independent variables. In this case, we was able to use all variable because our dataset is not large. In the cause that our dataset is extremely large, we will have to conduct a more thorough study.