Departure Delay Analysis on NYC flights in 2013

Introduction

The data set nycflights13 collects 3,273,346 flights details from three NYC airports - John F. Kennedy International Airport(JFK), Newark Liberty International Airport(EWR) and LaGuardia Airport(LGA). It also captures information about weather, airports, airlines and planes. Departure delay is one of the most addressed problems faced by all stakeholders within the supply chain.

This report aims at using R to visualize the nycflights13 data and gaining insights regarding how to improve departure delay. We believe the four most critical factors - **time, weather, planes (age and type) and carrier** - entail departure delays. To quantify, we choose different indicators of departure delay to visualize, in terms of **average delay time (in minutes)**, **delay occurrences**, and **departure delay percentage**. Delay occurrence is the indicator of departure delay in an absolute term (i.e. how many times the delay occurs); while for departure delay percentage, we mainly focus on the proportion of delays over numbers of flights. Further more, we apply a filtering criteria of 15 minutes when visualising the departure delay.

The remainder of this report is going to provide the answers to the following questions: - Question 1: What is the time pattern of departure delay? - Question 2: What is the effect of weather on departure delay? - Question 3: What is the impact of carrier performance on departure delay? - Question 4: How does the plane type and age affect the departure delay?

Assumption

Before our visualization, we assume the following:

* For calculating average departure delay time and numbers of departure delays, we **remove the 1% quantile largest observations** when filtering data sets that seem to have outliers.
* We apply the threshold of **15 minutes** for departure delay when filtering. After applying distribution() for departure delay, we found the data mostly gather around 0 minutes. Specifically, we would like to investigate the larger departure delay time, hence filtering 15 minutes as criteria.
* Assume the removal of missing values under 2.5% will not affect the analysis result while imputation techniques will be used for missing values larger than 9%.
* We assume that extreme weather conditions or airline operation issues will lead to **flight cancellation**. The below analysis will only focus on non-canceled flights with positive departure delays.
* The carrier performance is solely based on evaluating the departure delay by each carrier. **Those flights with early departure will not be considered in the below analysis**.
* Assume EWR, JFK and LGA airports comply with the same set of air traffic regulations whereas the time interval required for flight take-offs is the same.
* Recommendations are made based on the below assumptions:

1. Assume passenger volume is higher during the holiday period.
2. Assume the precipitation level is higher during the summer period.
3. Assume a preceding flight delay will directly affect the departure delay of a subsequent flight.
4. Extreme weather conditions (e.g. hurricanes, tornadoes, blizzards, etc) will not be considered in the weather analysis session.

Data integration

Load packages

*# Libraries required for this activity*

lapply(c("tidyverse", "dplyr","lubridate","dlookr","prettydoc",

"xray","RColorBrewer","cowplot","gridExtra","hexbin"),

**library**, character.only = TRUE)

Data cleaning and inspection

Basically, after scrutiny of all the data sets, we chose flights, weather, and planes as our targets. The variables in other two data sets - airlines and airports - are included in the three target data sets. When reading the csv, we add new variables which are used in the following visualization part, and select the variables we want in order to simplify the data sets via %>%.

For flights data set, we construct the variables needed in the following part: day\_of\_week, period\_dep\_time, and delay\_or\_not. For day\_of\_week, we define each day with day of week (Monday, Tuesday, etc.). Apart from that, we need to investigate the departure delay in each time period, say morning, afternoon and evening. After applying the distribution() for flights data set, it is obvious that the distribution of departure delay is significantly right-skewed. Furthermore, we would like to focus on delay time with longer duration. On top of that, we apply a threshold of 15 minutes in terms of departure delay time to provide a better picture, as addressed in Assumption. To realize, we set up a binary variable to indicate whether the departure delay time is above 15 minutes. By summing up the delay\_or\_not, we can get the total count of delays which are longer than 15 minutes. On the other hand, the mean value of delay\_or\_not defines the proportion of delays (longer than 15 minutes).

*# Reading flights data*

flights <- read\_csv("flights.csv") %>%

*#Create day of week for calendar heat map and make the variables appear in the right order*

mutate(date=paste(year,month,day,sep="-"),

date=as.Date(date,"%Y-%m-%d"),

dayofweek=weekdays(date))%>%

mutate(day\_of\_week=factor(dayofweek,levels = c("Sunday","Monday","Tuesday","Wednesday",

"Thursday","Friday","Saturday"),

labels=c("Sun","Mon","Tue","Wed","Thu","Fri","Sat"),

ordered=TRUE)) %>%

mutate(Month=factor(month,levels=1:12,labels=c("Jan","Feb","Mar","Apr","May",

"Jun","Jul","Aug","Sep","Oct","Nov","Dec"),

ordered=TRUE))%>%

*#Create time period for barchart*

mutate(period\_dep\_time = ifelse(between(hour, 5, 11), "5-11",

ifelse(between(hour, 12, 18), "12-18",

ifelse(between(hour, 19, 23), "19-23","unmatched")))) %>%

*#Create dummy variable to ascertain whether the departure delay is above 15 minutes*

mutate(delay\_or\_not = ifelse(dep\_delay > 15, 1, 0)) %>%

select(origin,time\_hour,month,day,Month,dep\_delay,hour,carrier,tailnum,

day\_of\_week,delay\_or\_not,period\_dep\_time)

*# Data skimming*

summary(flights)

anomalies(flights)

distributions(flights)

For weather data set, we decided to remove the N.A. after applying anomalies(), from which we can find that the null values only account for 0.02%. Hence, we decide to remove the observations with null values. Apart from that, we select the variables needed to visualize and merge by in the following part.

From distribution(), precipitation, visibility and wind speed are not normally distributed. Hence, it is likely that the visualization of weather factors are not obvious and reliable.

*# Reading weather*

weather <- read\_csv("weather.csv",) %>%

*# 0.02% of N.A., not affect too much*

filter(!is.na(wind\_speed),!is.na(humid))%>%

select(time\_hour,origin,humid,wind\_speed,precip,visib,month)

*# Data skimming*

summary(weather)

anomalies(weather)

distributions(weather)

For planes, we choose the variables we want. After inspecting the anomalies, the null values of the variable year account for 2.11%.

*# Reading planes: removing null values and selecting the variables we want*

planes <- read\_csv("planes.csv") %>%

select(tailnum,year,type)

*# Data skimming*

summary(planes)

anomalies(planes)

distributions(planes)

After loading weather and planes data sets, we merge them into a intergrated data frame data, in order to simplify the process and codes. After applying anomalies() on data, and there are 16.35% of NA in variable of year and model. In order to keep all the observations when visualizing other variables and avoid over tailing the data, we decide to remove the NA when it is necessary. To deal with the null values of year, we decide to apply **imputation**, through replacing the null values with the median of year. When it comes to the null values of type, we will remove them when we use the variable of type for visualization.

data<-flights%>%

left\_join(planes, by = "tailnum")%>%

inner\_join(weather, by = c("origin", "time\_hour","month"))

*# Data skimming*

anomalies(data)

## $variables

## Variable q qNA pNA qZero pZero qBlank pBlank qInf pInf

## 1 precip 325724 0 - 304799 93.58% 0 - 0 -

## 2 delay\_or\_not 325724 0 - 255811 78.54% 0 - 0 -

## 3 year 325724 53257 16.35% 0 - 0 - 0 -

## 4 type 325724 48123 14.77% 0 - 0 - 0 -

## 5 dep\_delay 325724 0 - 16379 5.03% 0 - 0 -

## 6 wind\_speed 325724 0 - 11684 3.59% 0 - 0 -

## 7 visib 325724 0 - 87 0.03% 0 - 0 -

## 8 origin 325724 0 - 0 - 0 - 0 -

## 9 period\_dep\_time 325724 0 - 0 - 0 - 0 -

## 10 day\_of\_week 325724 0 - 0 - 0 - 0 -

## 11 month 325724 0 - 0 - 0 - 0 -

## 12 Month 325724 0 - 0 - 0 - 0 -

## 13 carrier 325724 0 - 0 - 0 - 0 -

## 14 hour 325724 0 - 0 - 0 - 0 -

## 15 day 325724 0 - 0 - 0 - 0 -

## 16 humid 325724 0 - 0 - 0 - 0 -

## 17 tailnum 325724 0 - 0 - 0 - 0 -

## 18 time\_hour 325724 0 - 0 - 0 - 0 -

## qDistinct type anomalous\_percent

## 1 55 Numeric 93.58%

## 2 2 Numeric 78.54%

## 3 47 Numeric 16.35%

## 4 4 Character 14.77%

## 5 526 Numeric 5.03%

## 6 34 Numeric 3.59%

## 7 20 Numeric 0.03%

## 8 3 Character -

## 9 3 Character -

## 10 7 Unknown -

## 11 12 Numeric -

## 12 12 Unknown -

## 13 16 Character -

## 14 19 Numeric -

## 15 31 Numeric -

## 16 2440 Numeric -

## 17 4037 Character -

## 18 6885 Timestamp -

##

## $problem\_variables

## Variable q qNA pNA qZero pZero qBlank pBlank qInf pInf qDistinct

## 1 precip 325724 0 - 304799 93.58% 0 - 0 - 55

## type anomalous\_percent problems

## 1 Numeric 93.58% Anomalies present in 93.58% of the rows.

Visualisation

Question 1 - What is the time pattern of departure delay?

In this part, we investigate the time pattern (Month, Day of week and Hour) that departure delay follows using three different indicators: mean departure delay, count of departure delay, and delay percentage. For average delay time, we remove the outliers; whilst for count and percentage, we apply a threshold of 15 minutes, as referred to the assumption part.

Generally, during April, summer time (June and July), and December, the delay is more serious in terms of duration, absolute and relative frequency. Thursday has the largest number of above-average delay time and percentage. Furthermore, even though the count of delays is the highest during 12:00 to 18:59, the average delay minutes are longest during night time (19:00 to 23:59)

Calendar heat map of average departure delay in 2013

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99)) %>%

group\_by(Month,day\_of\_week,hour)%>%

summarise(delay\_mean=mean(dep\_delay))%>%

ggplot(aes(x=hour,y=day\_of\_week,fill=delay\_mean)) +

geom\_tile(color = "white",size=0.1) +

facet\_wrap(~Month,nrow = 4) +

scale\_fill\_binned(breaks = c(15,30,45,60,75,90),type = "viridis",name="Mean delay(min)")+

labs(x=NULL,y=NULL, title="Average departure delay calandar")+

theme(plot.title = element\_text(hjust = 0.6),

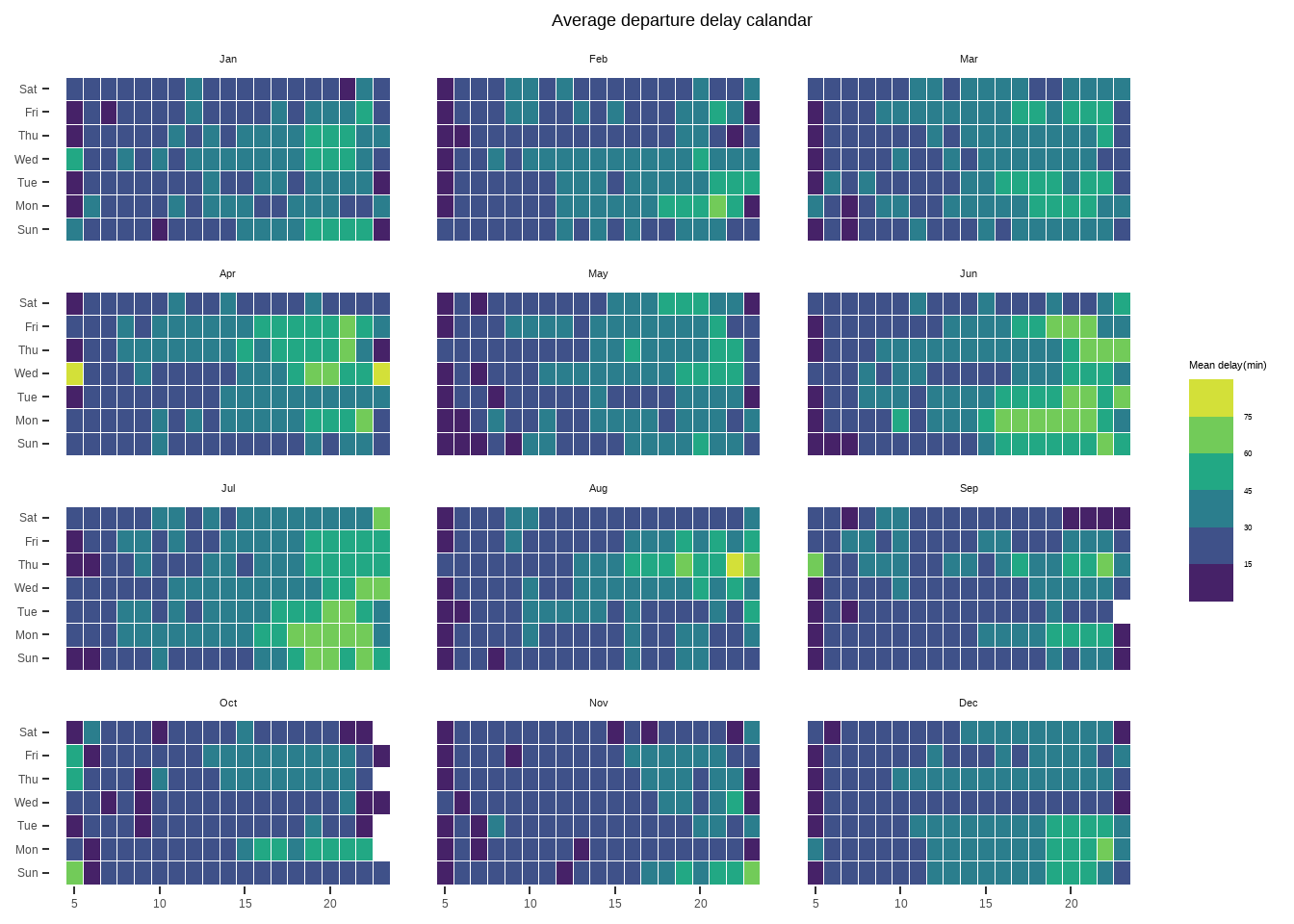
legend.title = element\_text(size = 8),

legend.text = element\_text(size = 6),

strip.text.x=element\_text(size=8),

strip.background = element\_blank(),

panel.background = element\_blank())



Despite lacking the flights information from 0:00 a.m.to 4:59 a.m., this calendar heat map depicts obvious time pattern for average departure delay time. Starting from the combined data set data and grouping by month, day of a week, and hour, the calendar heat map shows that throughout 2013, April, June and July entail more delay time; and the delay which has a duration above 45 minutes occur quite often in the late afternoon and evening.

Bar chart of average departure delay VS hours by day of week

*# Calculate the average value of the filtered delay*

paste(data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

summarise(mean(dep\_delay)))

*#33.44*

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99)) %>%

group\_by(day\_of\_week,hour) %>%

summarise(delay\_mean\_dayhour=mean(dep\_delay))%>%

*# comparison with the mean value*

ggplot(aes(x=hour,y=delay\_mean\_dayhour,fill=delay\_mean\_dayhour>=33.44)) +

geom\_col(alpha=0.7) +

facet\_wrap(~day\_of\_week, nrow=2) +

labs(x="Hour",y="Average Departure Delay (in minutes)",

title="Average delay for each hour by day of week") +

geom\_hline(aes(yintercept=33.44),linetype = 2)+

scale\_fill\_manual(name = "",values = c("#00AFBB", "#E7B800"))+

theme(plot.title = element\_text(hjust = 0.45),

strip.background = element\_blank(),

panel.background = element\_blank(),

legend.position = "none")



Apart from monthly and hourly pattern, the day of week is another factor. Apparently, the departure delay time in Saturdays is shorter than others. There are more above-average departure delays in Mondays and Thursdays. As illustrated in plot 4.1.1, the departure delay time is longer in night time compared to that in day time.

Calendar heat map of percent departure delay in 2013

data%>%

group\_by(Month,day\_of\_week,hour) %>%

summarise(percent\_delay = 100 \* mean(delay\_or\_not))%>%

ggplot(aes(x=hour,y=day\_of\_week,fill=percent\_delay)) +

geom\_tile(color = "white",size=0.1) +

facet\_wrap(~Month,nrow = 4) +

scale\_fill\_viridis\_b(name="Delay in %")+

labs(x=NULL,y=NULL, title="Percentage departure delay calandar")+

theme(plot.title = element\_text(hjust = 0.6),

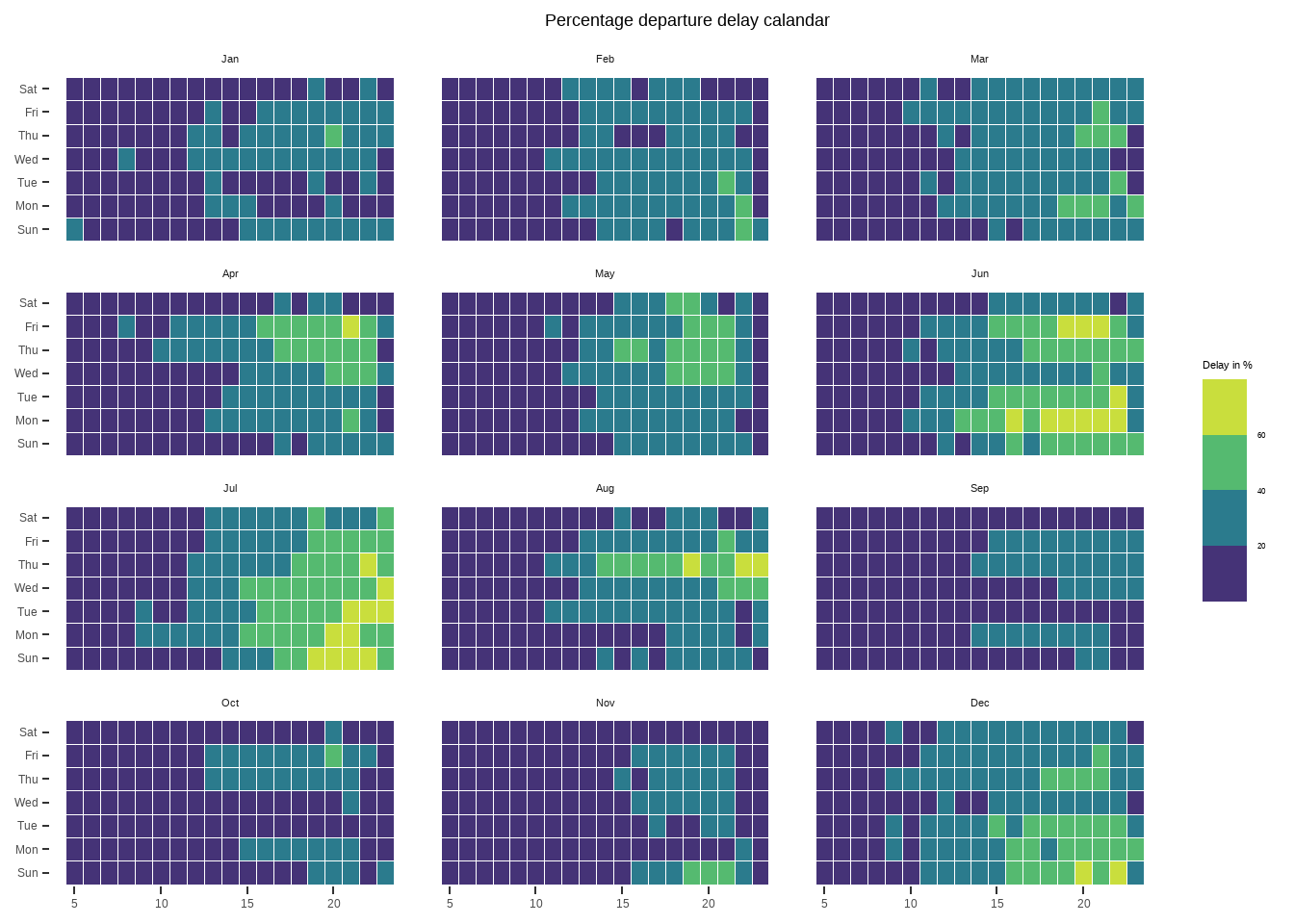
legend.title = element\_text(size = 8),

legend.text = element\_text(size = 6),

strip.text.x=element\_text(size=8),

strip.background = element\_blank(),

panel.background = element\_blank())



To further investigate other indicators of departure delay, we use percentage of delays respective to the number of flights. Specifically, the variable delay\_or\_not is a dummy variables indicating whether or not >15, hence we only focus on the percentage of delays the time of which is longer than 15 minutes. The mean of 100 \* mean(delay\_or\_not) depicts the percetage of delays (>15 minutes) over the number of flights. Similar to the heat map of 4.1.1, in April and summer time, there are more delays in relative term. Interestingly, in December, there are also some serious delays, which is due to Christmas holiday.

Bar chart of percent deprture delay VS hours by day of week

*# Calculate the average percentage of the filtered delay*

paste(flights %>%

mutate(delay\_or\_not = ifelse(dep\_delay > 15 ,1, 0)) %>%

summarise(mean\_percent\_delay = 100 \* mean(delay\_or\_not)))

*## 21.5%*

data%>%

group\_by(day\_of\_week,hour) %>%

summarise(percent\_delay = 100 \* mean(delay\_or\_not))%>%

*# comparison with the mean value*

ggplot(aes(x=hour,y=percent\_delay,fill=percent\_delay>=21.5)) +

geom\_col(alpha=0.7) +

facet\_wrap(~day\_of\_week, nrow=2) +

labs(x="Hour",y="Percentage Departure Delay (%)", title="Percentage delay for each hour by day of week") +

geom\_hline(aes(yintercept=21.5),linetype = 2)+

scale\_fill\_manual(name = "",values = c("#00AFBB", "#E7B800"))+

theme(plot.title = element\_text(hjust = 0.45),

strip.background = element\_blank(),

panel.background = element\_blank(),

legend.position = "none")



Again, Saturday delays are substantially less than that of others. On average, 21.5% of the flights are delayed above 15 minutes. After 8:00 p.m., the percentage delay decreases, which means the peak time for delays is during the afternoon.

Line graph of delay count VS month

data%>%

filter(dep\_delay>15)%>%

group\_by(month) %>%

summarise(count=n())%>%

ggplot(aes(x=month,y=count))+

geom\_line(linetype=2,size=1,color="#00AFBB")+

geom\_point(size=3)+

theme\_bw()+

scale\_x\_discrete(limits=1:12)+

labs(x="Month",y="Number of departure delays",title="Time series of delay count")+

theme(plot.title = element\_text(hjust = 0.5),

plot.background = element\_blank())

Chart, line chart

Description automatically generated

Still, the delay is filtered with the threshold of 15 minutes. The monthly time series line graph demonstrates that summer time (June and July) and December have the largest number of delays. During summer time, the flight schedules might be affected by the weather, which will be detailed discussed in Part 4.2. For December, it is reasonable to have more passengers during the Christmas holiday, hence generating more delays.

Bar chart of departure delay count for three time period by day of week and month

data%>%

filter(dep\_delay>15)%>%

group\_by(Month,day\_of\_week,period\_dep\_time) %>%

summarise(count\_period=n())%>%

ggplot(aes(x=count\_period,y=day\_of\_week,fill=period\_dep\_time))+

geom\_col(alpha=0.7)+

facet\_wrap(~Month)+

labs(x="Delay count",y=NULL, title="Departure delay count for three period by day of week and month")+

scale\_fill\_manual(name = "Time period",values = c("#00AFBB", "#E7B800", "#FC4E07"))+

scale\_x\_continuous(breaks = seq(0,2500,length=3))+

theme(plot.title = element\_text(hjust = 0.6),

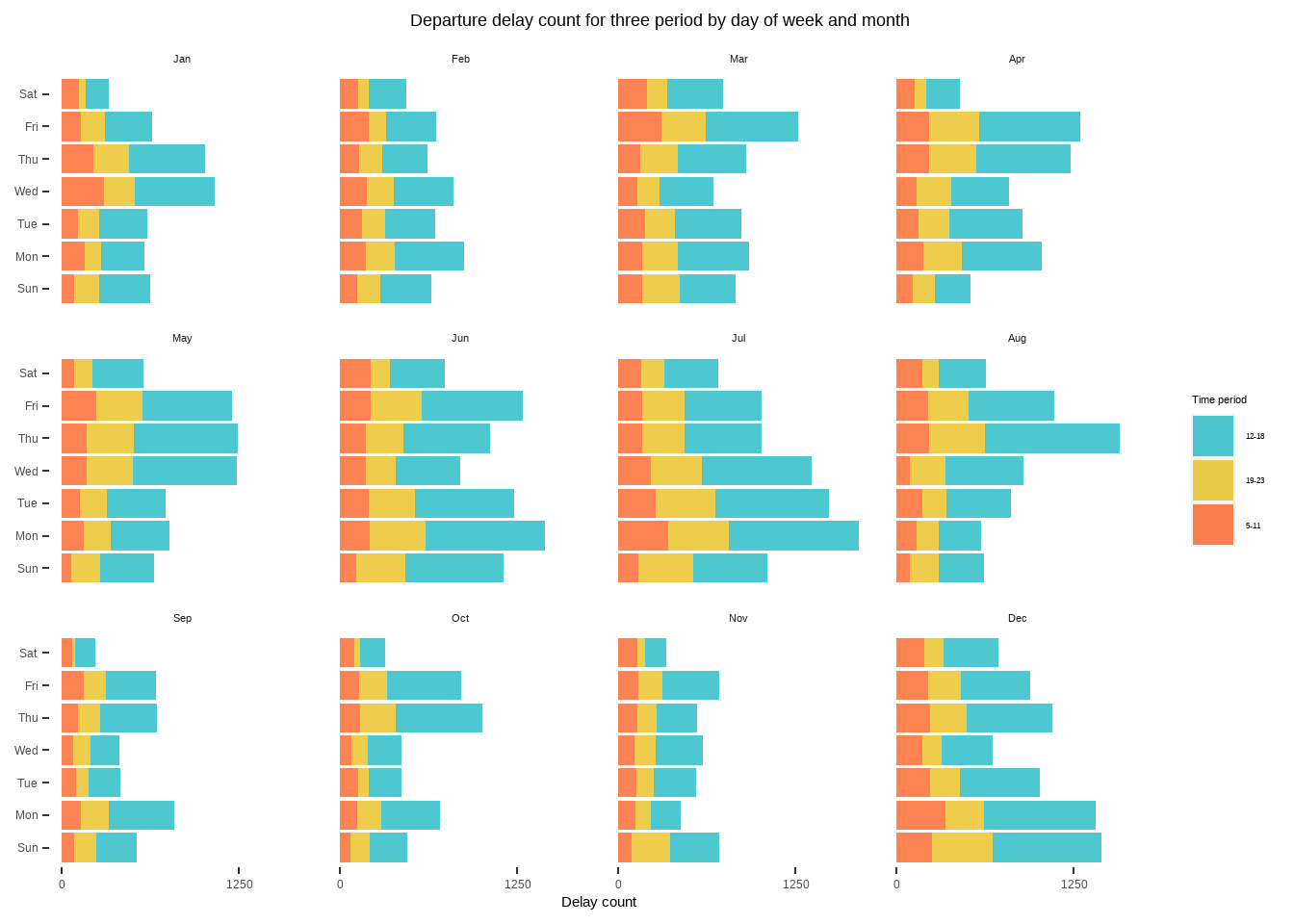
legend.title = element\_text(size = 8),

legend.text = element\_text(size = 6),

strip.text.x=element\_text(size=8),

strip.background = element\_blank(),

panel.background = element\_blank())



In both absolute and relative term, as depicted in 4.1.4, there are more delays during afternoon, especially in Thursdays.

Question 2 - What is the effect of weather on departure delay?

To further investigate the relationship between flight delays and weather, we plan to explore the average delay time versus different levels of weather factors (**visibility, relative humidity, wind speed, and precipitation**) to understand the overall tendency of departure delays. Considering there are some early departure flights in the data set, there might be a risk of averaging out when we carry out the mean calculation. Thus, the below analysis will only focus on flights that have a departure delay time larger than 0 minutes and within the 99% quantile.

Boxplot for 4 weather factors (Precip, Visib, Wind speed and Humid) VS average delay

Before getting into details of each weather factor, we planned to build a box-plot for all weather factors to obtain a general weather phenomenon. We decided to summarize the average delay data for the 4 weather factors and group by each time hour. The average delay column is mutated to cut the bin width of the box-plot to 30 mins intervals for better visualization. Each box-plot for visibility, precipitation, wind speed and humidity is created and further arranged in grids. For better visualization, we apply different filtering criteria on different weather factors, which is explained in the comments.

*#create a box-plot for visibility*

*#Removed visibility equals to 10 to better analyse the effect of low visibility*

p1<- data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

group\_by(time\_hour)%>%

summarise\_at(vars(visib,humid,wind\_speed,precip,avg\_delay = dep\_delay),mean,na.rm=TRUE)%>%

mutate(avg\_delay = cut\_width(avg\_delay, 30))%>%

filter(visib < 10)%>%

ggplot(aes(avg\_delay, visib))+

geom\_boxplot()+

theme(axis.title.x=element\_blank())+

ylab("Visibility (miles)")+

scale\_x\_discrete(labels=c("0","30","60","90","120","150", "180"))

*#create a box-plot for precipitation*

*#Removed precipitation equals to 0 to better analyse the rainfall trend*

p2<-data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

group\_by(time\_hour)%>%

summarise\_at(vars(visib,humid,wind\_speed,precip,avg\_delay = dep\_delay),mean,na.rm=TRUE)%>%

mutate(avg\_delay = cut\_width(avg\_delay, 30))%>%

filter(precip > 0, precip < quantile(precip, 0.99, na.rm = TRUE))%>%

ggplot(aes(avg\_delay, precip))+

geom\_boxplot()+

theme(axis.title.x=element\_blank())+

ylab("Precipitation (inches)")+

scale\_x\_discrete(labels=c("0","30","60","90","120","150"))

*#create a box-plot for wind speed*

p3<-data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

group\_by(time\_hour)%>%

summarise\_at(vars(visib,humid,wind\_speed,precip,avg\_delay = dep\_delay),mean,na.rm=TRUE)%>%

mutate(avg\_delay = cut\_width(avg\_delay, 30))%>%

ggplot(aes(avg\_delay, wind\_speed))+

geom\_boxplot()+

theme(axis.title.x=element\_blank())+

ylab("Wind Speed (mph)")+

scale\_x\_discrete(labels=c("0","30","60","90","120","150","180"))

*#create a box-plot for relative humidity*

p4<-data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

group\_by(time\_hour)%>%

summarise\_at(vars(visib,humid,wind\_speed,precip,avg\_delay = dep\_delay),mean,na.rm=TRUE)%>%

mutate(avg\_delay = cut\_width(avg\_delay, 30))%>%

ggplot(aes(avg\_delay, humid))+

geom\_boxplot()+

theme(axis.title.x=element\_blank())+

ylab("Relative Humidity")+

scale\_x\_discrete(labels=c("0","30","60","90","120","150","180"))

*#arrange the 4 box-plots*

grid.arrange(p1, p2, p3, p4, nrow = 2,

top = "Average delay VS weather factors",

bottom = "Average delay (in minutes)")

Chart, box and whisker chart

Description automatically generated

The graph shows that precipitation and relative humidity factors are mostly correlated with the major departure delays. The higher the precipitation and relative humidity levels, the longer the delay duration of the flight. The two factors are likely correlated since a higher level of precipitation will lead to a higher level of humidity, which will cause a similar effect on flight delays. Furthermore, for visibility, the graph indicates a great increase in departure delay duration when the visibility level drops to 3 miles or below. Lastly, we can observe the wind speed factor causes minor delay as the level increases from 0 mph to 15 mph, yet there are still many long delays do not explain by the wind speed factors.

Monthly trend of precipitation

As we assumed that there is a seasonal pattern for precipitation, we will first using line charts to explore the average level of precipitation versus time series in different NYC airports. We firstly grouped the data by origin and month then calculated the average precipitation level. A vertical line is added to indicate the months having the maximum level of precipitation in different airports.

weather%>%

group\_by(origin, month)%>%

summarise(average\_precip = mean(precip, na.rm = TRUE))%>%

ggplot(aes(x=month, y=average\_precip))+

geom\_line()+

scale\_x\_continuous(breaks = seq(1,12,by=1))+

facet\_wrap(~origin)+

labs(title = "Monthly trend of precipitation in 2013",

y = "Average Precipitation (inches)",

x = "Month")+

geom\_vline(xintercept = 6, col="#FC4E07", lwd = 0.8)+

theme\_bw()+

theme(plot.title = element\_text(hjust = 0.5))

Chart, line chart, histogram

Description automatically generated

Not surprisingly, precipitation is highest during the summer time, in which 3 airports both have the heaviest rainfall in June. By comparing this graph with figure 4.1.5, we can observe a similar trend for each month. This may further indicate that the weather factor of precipitation has a high correlation with departure delays.

Column chart for precipitation VS percentage delay with delay count

To investigate the relationship between precipitation and percentage of delay, we generate a column chart with a smoothing curve to observe the trend. Again, the percentage is based on the filtering of above 15 minutes. Additionally, we remove the zeros and outliers of precipitation, since the zero precipitation accounts for 93% of the data (see by 3.2). With the filling of counts of delays which are above 15 minutes, we can gain deeper insights.

data%>%

filter(precip > 0, precip < quantile(precip, 0.99)) %>% *# remove 0 and outliers*

group\_by(precip) %>%

summarise(percent\_delay = mean(delay\_or\_not)\*100,

count\_delay = sum(delay\_or\_not),

count = n())%>%

ggplot(aes(x = precip, y = percent\_delay, fill = count)) +

geom\_col(alpha = 0.6) +

geom\_smooth(se=F,color="#FC4E07") +

theme\_bw() +

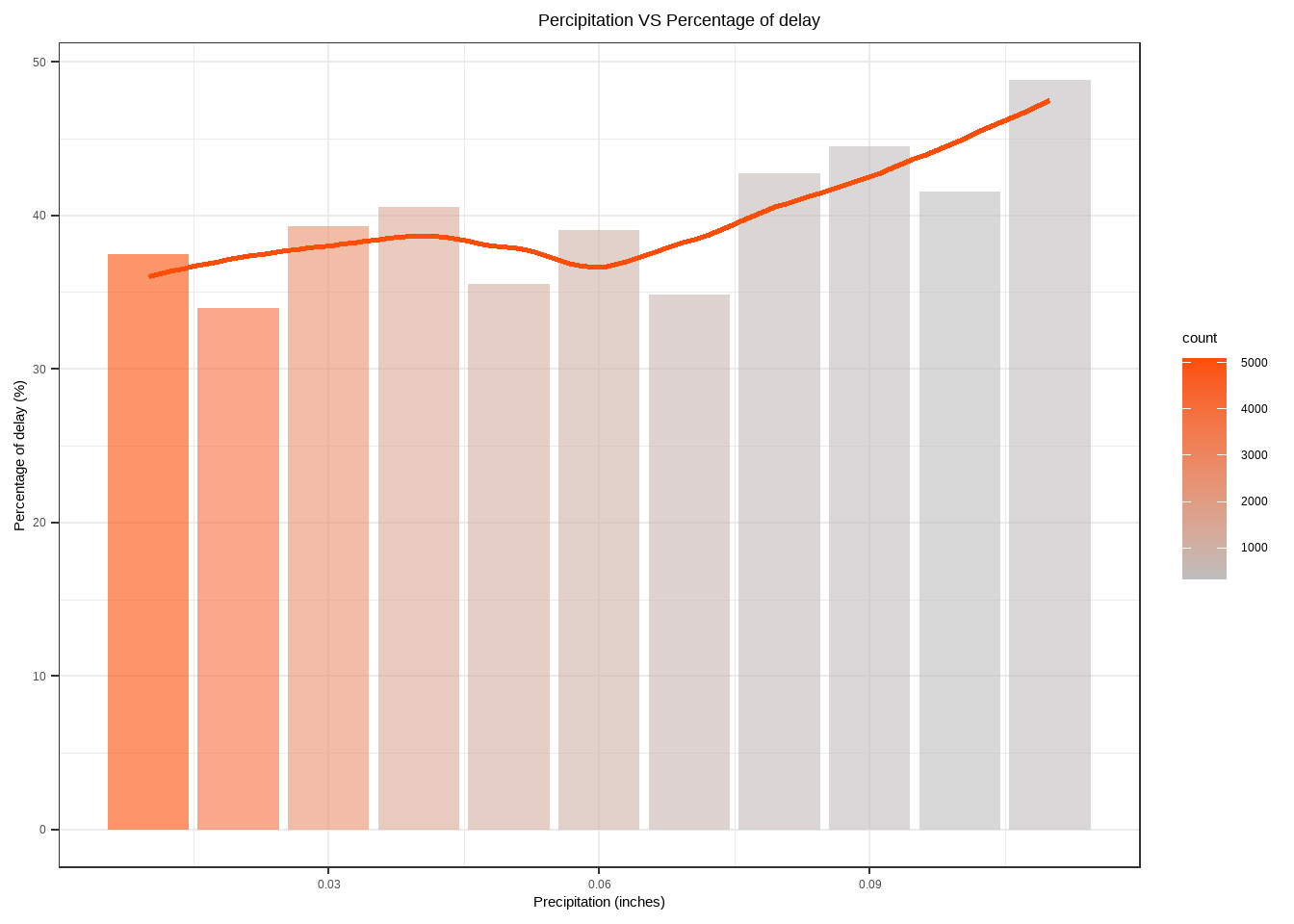
labs(title = "Percipitation VS Percentage of delay",

x = "Precipitation (inches)",

y = "Percentage of delay (%)")+

scale\_fill\_gradient(low="grey", high="#FC4E07")+

theme(plot.title = element\_text(hjust = 0.6))



From the smoothing curve, the overall trend of precipitation and percentage delay is increased. In other words, the larger the precipitation, the more delays in the relative term. Whilst, the counting of delays does not follow the same pattern, for which the lack of data during high precipitation rationalizes.

Scatter plot of visibility VS average departure delay

In order to better analyse the effect of visibility on average departure delay, we have grouped the data by each level of visibility and calculated the average departure delay duration per visibility mile. A scatter plot is plotted as below with a smoothed line for better visualization.

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

group\_by(visib)%>%

*# summarises the data by visib and calculate the average delay per each visibility miles*

summarise(average\_delay = mean(dep\_delay), count = n())%>%

ggplot(aes(x=visib, y=average\_delay)) +

geom\_point()+

stat\_smooth(se=FALSE, color = "#00AFBB")+

labs(title = "Level of visibility v.s. Average departure delay",

y = "Average departure delay (minutes)",

x = "Visibility (miles)")+

theme\_bw()+

theme(plot.title = element\_text(hjust = 0.5))

Chart, line chart

Description automatically generated

As observing the smoothing line, we can see that the lower the visibility higher the average departure delay time. It indicates that the visibility adversely impacts the flight departure in NYC. There is also a significant increase in departure delay duration when visibility drops from 2.5 miles to 0 miles. Yet the line started to flatten out when the visibility level is higher than 2.5 miles. It may indicate that flights are severely affected by visibility levels from 0 to 2.5 miles.

Scatterplot of Visibility VS percent delay with delay count

Besides the mean delay time, we want to focus on the frequency of delays, in both absolute and relative terms. Since the data distribution of delays VS visibility is not normal, we generate bins for visibility to distinguish. Combined with the smoothing line, better visualization is generated.

visib\_flights<-

data%>%

mutate(visib = as.numeric(format(visib, digits = 3, format="f")))%>%

mutate(visib\_bin\_temp1 = cut(visib, breaks = seq(from = -0.5, to = 10.5, by = 1), right = FALSE),

*# sub to choose string before and after the pattern*

visib\_bin\_temp2 = as.numeric(sub("\\,.\*", "", sub(".\*\\[", "", visib\_bin\_temp1))),

visib\_bin = visib\_bin\_temp2 + 0.5) %>%

group\_by(visib\_bin) %>%

summarise(percent\_delay = mean(delay\_or\_not)\*100,

count\_delay = sum(delay\_or\_not),

count = n())

ggplot(visib\_flights,aes(x = visib\_bin, y = percent\_delay)) +

geom\_point(alpha = 0.6, aes(size = count\_delay)) +

geom\_smooth(

data = visib\_flights %>% filter(visib\_bin > 0, visib\_bin < 10), se=F,

method = "lm",color="#00AFBB", fill=NA,

formula=y ~ poly(x, 3, raw=TRUE) )+

scale\_size\_continuous(range = c(1, 30)) +

theme\_bw() +

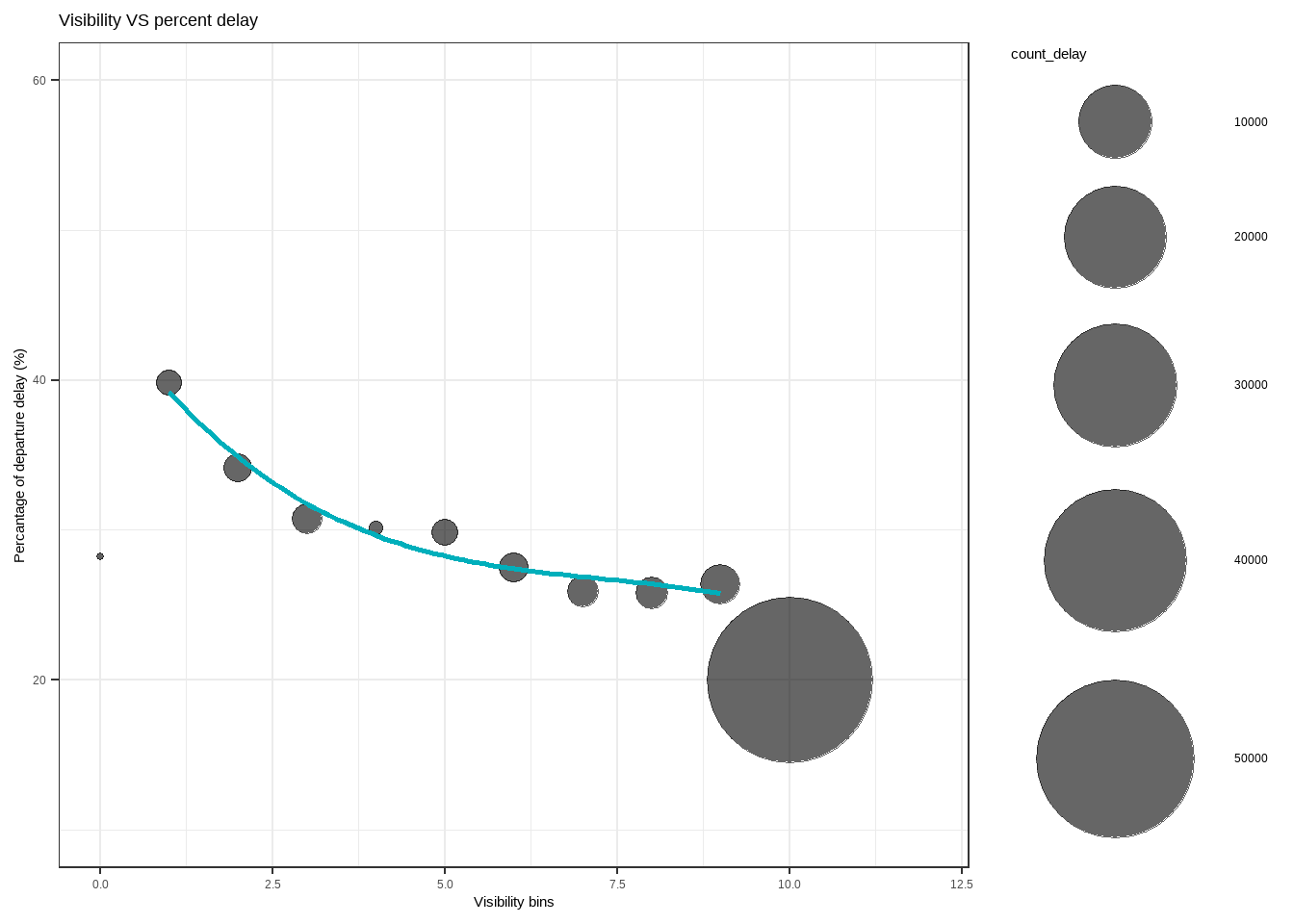
scale\_y\_continuous(limits = c(10,60)) +

scale\_x\_continuous(limits = c(0,12))+

labs(title = "Visibility VS percent delay",

y = "Percantage of departure delay (%)",

x = "Visibility bins")



Due to the nonequivalent data distribution, there are substantially more data when the visibility is around 10 miles. Apart from that, the decreasing trend of percentage delay in terms of visibility level is obvious.

Column chart of Visibility VS percent delay with delay count per orgin

Furthermore, we divide the group into three different origins and generate a column chart with smoothing curve, with both absolute and relative delay occurrence.

data%>%

mutate(visib = as.numeric(format(visib, digits = 3, format="f"))) %>%

filter(!(visib %**in**% c(1,10)))%>%

mutate(visib\_bin\_temp1 = cut(visib, breaks = seq(from = 0.5, to = 10.5, by = 1), right = FALSE),

*# sub to choose string before and after the pattern*

visib\_bin\_temp2 = as.numeric(sub("\\,.\*", "", sub(".\*\\[", "", visib\_bin\_temp1))),

visib\_bin = visib\_bin\_temp2 + 0.5) %>%

group\_by(visib\_bin, origin) %>%

summarise(percent\_delay = mean(delay\_or\_not)\*100,

count\_delay = sum(delay\_or\_not),

count = n())%>%

ggplot(aes(x = visib\_bin, y = percent\_delay)) +

geom\_col(alpha = 1, aes(fill = count\_delay)) +

scale\_size\_continuous(range = c(1, 10)) +

facet\_wrap(~ origin) +

stat\_smooth(se = FALSE, color = "#00AFBB")+

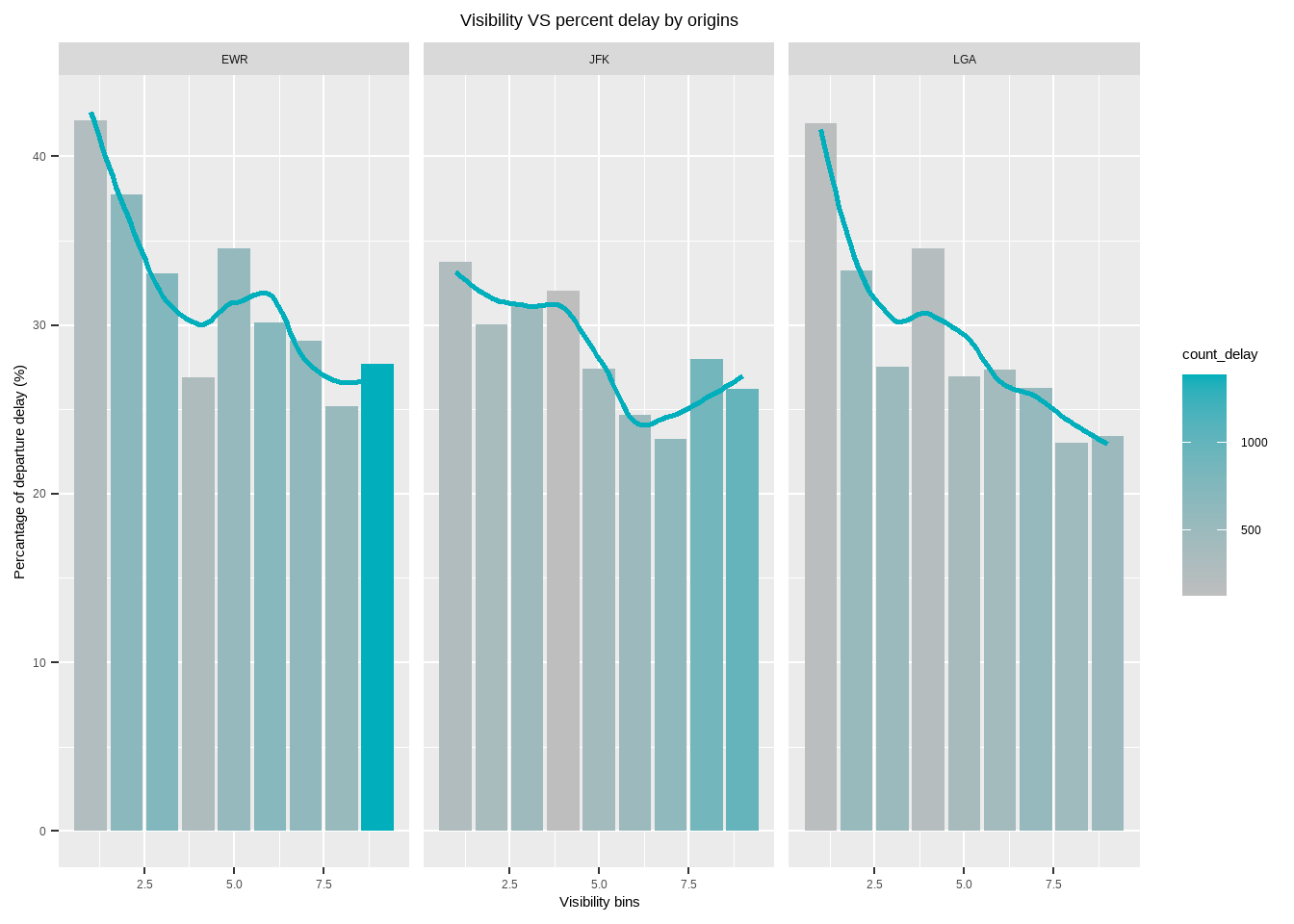
scale\_fill\_gradient(low="grey", high="#00AFBB")+

labs(title = "Visibility VS percent delay by origins",

y = "Percantage of departure delay (%)",

x = "Visibility bins")+

theme(plot.title = element\_text(hjust = 0.5))



Generally, the negative relationship between visibility and percentage delay is more observable in EWR and LGA compared to that in JFK. When the visibility is around 6 miles, the increase of visibility leads to more percentage delays. Surprisingly, there are more delays in EWR and JFK when the visibility is in good condition.

Scatterplot of windspeed VS average departure delay with delay count

Similarly, we use the same combined data set to generate a summarised table for wind speed analysis. The wind\_speed data is grouped by each level and the average departure delay is calculated per each level of wind speed. As scatter plot is plotted and below with a smooth line. In addition, the size of the data point indicates the count of delayed flights.

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

filter(!is.na(wind\_speed)) %>%

group\_by(wind\_speed)%>%

summarise(average\_delay = mean(dep\_delay),

count = n())%>%

ggplot(aes(x=wind\_speed, y=average\_delay)) +

geom\_point(aes(size = count))+

theme\_bw()+

geom\_smooth(se = FALSE, color = "#E7B800")+

labs(title = "Level of windspeed VS average departure delay",

y = "Average departure delay (minutes)",

x = "Wind speed (mph)")+

theme(plot.title = element\_text(hjust = 0.5))

Chart, scatter chart

Description automatically generated

The above graph indicates that the average departure delay duration increases as the wind speed increases from 0 mph to 30 mph. Although we can see that the departure delay time decreases sharply from 30 mph wind speed to 40 mph wind speed, the standard error of those data points is high as we can see the count of data is relatively low which affects the accuracy of the smoothing line.

Column chart for wind speed VS percent delay with delay count

data%>%

group\_by(wind\_speed) %>%

summarise(percent\_delay = mean(delay\_or\_not)\*100,

count\_delay = sum(delay\_or\_not),

count = n())%>%

ggplot(aes(x = wind\_speed, y = percent\_delay, fill = count\_delay)) +

geom\_col(alpha = 0.6) +

geom\_smooth(se=F, color="#E7B800") +

theme\_bw() +

scale\_fill\_gradient(low="grey", high="#E7B800")+

labs(title = "Windspeed VS Departure delay",

y = "Percantage of departure delay (%)",

x = "Wind speed (mph)")+

theme(plot.title = element\_text(hjust = 0.5))

Chart, histogram

Description automatically generated

Generally, more data is gathered around when the wind speed is equal to 10 mph. The smoothing curve depicts a trend of first increasing then decreasing in percentage delay when the wind speed becomes larger. Similar to figure 4.2.7, when speed is above 30 mph, the overall delay seems to be improved in terms of time and occurrence. Whilst, as analysed above, the pattern is because of the high variation of the observations in the decreasing phase of the graph.

Column for Wind speed VS percent delay with count delay per orgin

data%>%

group\_by(wind\_speed, origin) %>%

summarise(percent\_delay = mean(delay\_or\_not)\*100,

count\_delay = sum(delay\_or\_not),

count = n())%>%

ggplot(aes(x = wind\_speed, y = percent\_delay, fill = count)) +

geom\_col(alpha = 0.6) +

scale\_fill\_gradient(low="grey", high="#E7B800")+

geom\_smooth(se=F, color="#E7B800")+

theme\_bw() +

facet\_wrap(~origin)+

labs(title = "Windspeed VS Departure delay by origin",

y = "Percantage of departure delay (%)",

x = "Wind speed (mph)")+

theme(plot.title = element\_text(hjust = 0.5))

Chart, histogram

Description automatically generated

Separating by different origins, the graph could give us a better insight. For EWR and JFK, the trend is similar to the general one. However, LGA does not have a decreasing trend. Instead, the percentage of delays keeps increasing even though the wind speed is higher than 30 mph. Simultaneously, the non-equal data distribution can be observed through the columns (count of delays).

Hexbin for humid VS delay

Lastly, we explored the relationship between humidity and actual departure delay time with a hexbin density plot. The bin size is set to 10 for better visualization.

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

ggplot(aes(x = humid, y = dep\_delay))+

ylim(0,150)+

stat\_binhex(bins= 10, color=c("#D7DADB"))+

theme\_classic() +

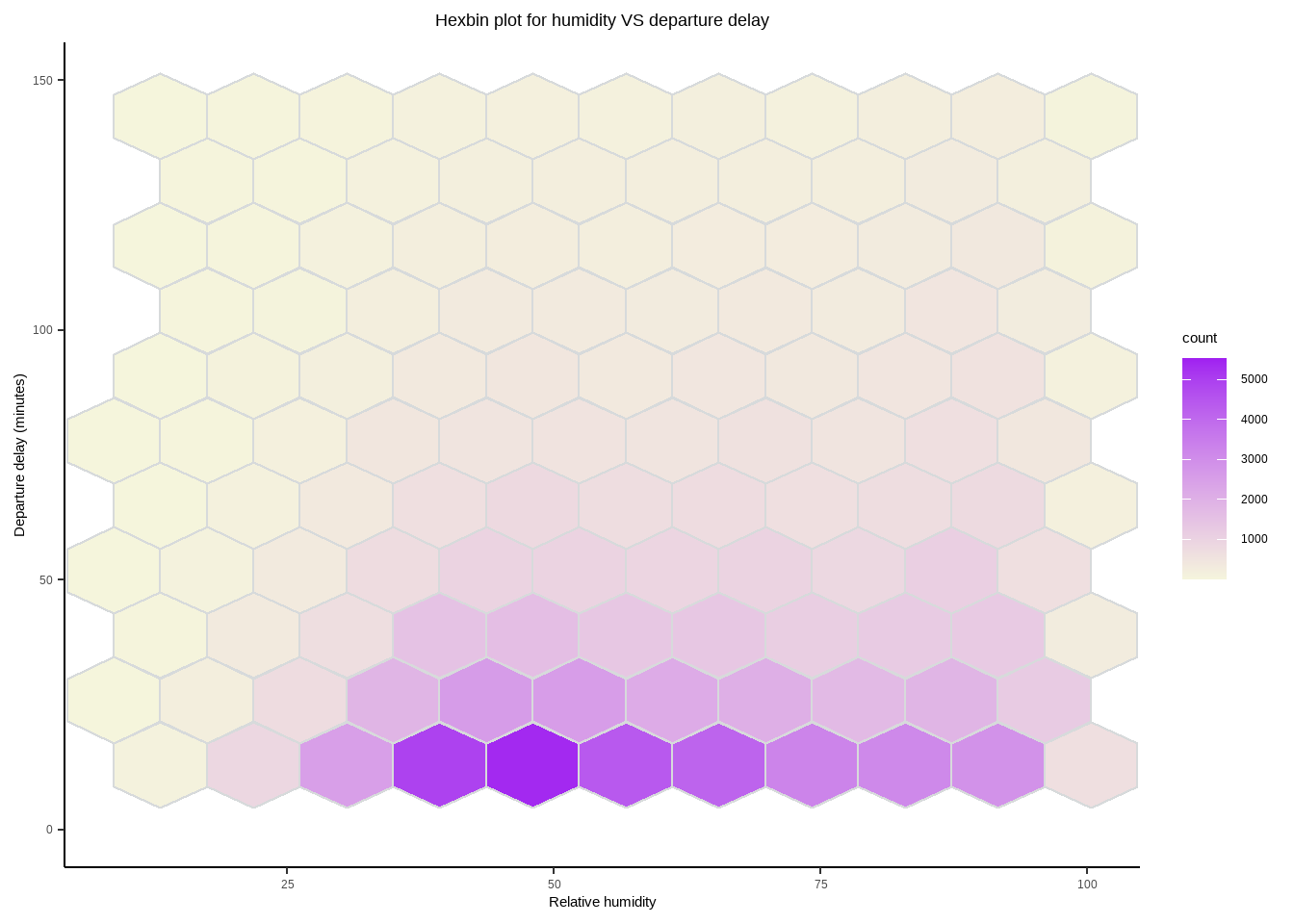
labs(x = "Relative humidity",

y = "Departure delay (minutes)",

title = "Hexbin plot for humidity VS departure delay")+

scale\_fill\_gradient(low = "beige", high = "purple")+

theme(plot.title = element\_text(hjust = 0.5))



As observing the purple color trend in the above graph, we can see a minor tendency indicating the departure delay increases as relative humidity increases. It is believed that humidity affects departure along side high level of precipitation, which may lead to delays for safety reasons.

Column chart for humidity VS percent delay with count delay

Similar to what we did to the other three weather indicator, we generate a graph regarding relative and absolute delays with a smoothing curve.

humid\_flights<-data%>%

mutate(humid\_bin\_temp1 = cut(humid, breaks = seq(from = -1, to = 110, by = 2), right = FALSE),

*# sub to choose string before and after the pattern*

humid\_bin\_temp2 = as.numeric(sub("\\,.\*", "", sub(".\*\\[", "", humid\_bin\_temp1))),

humid\_bin = humid\_bin\_temp2 + 1) %>%

group\_by(humid\_bin) %>%

summarise(percent\_delay = mean(delay\_or\_not)\*100,

count\_delay = sum(delay\_or\_not),

count = n())

ggplot(humid\_flights,aes(x = humid\_bin, y = percent\_delay, fill = count)) +

geom\_col(alpha = 0.6) +

geom\_smooth(se=F, data = humid\_flights %>% filter(humid\_bin != 12), color="purple") +

theme\_bw() +

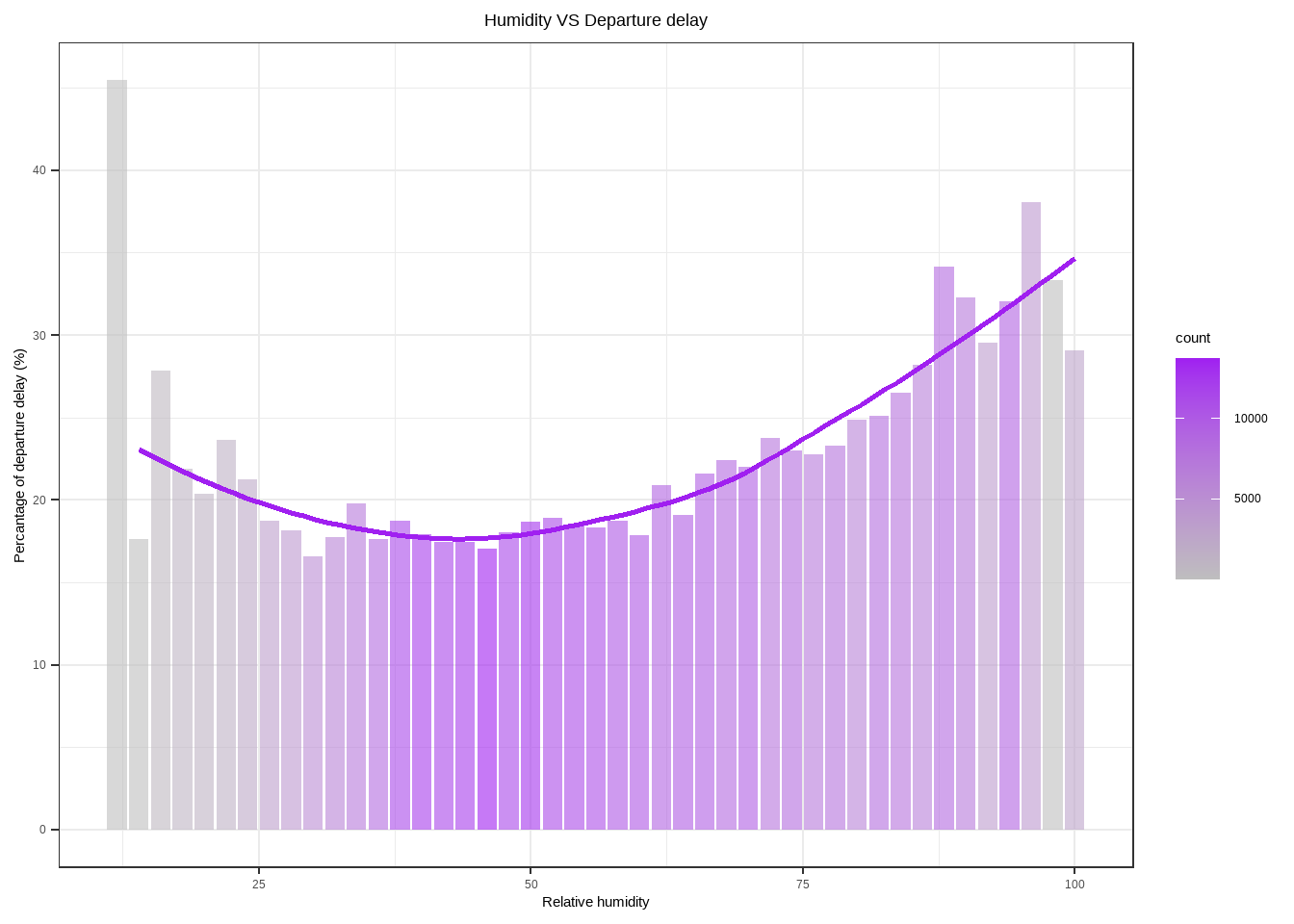
scale\_fill\_gradient(low="grey", high="purple")+

labs(title = "Humidity VS Departure delay",

y = "Percantage of departure delay (%)",

x = "Relative humidity")+

theme(plot.title = element\_text(hjust = 0.5))



Surprisingly, the smoothing depicts a convex pattern, the turning point of which is when the humidity equals to 50. As demonstrated with the absolute number of delays, there are more delays around that point, hence affecting the reliability of the result.

Question 3 - What is the impact of carrier performance on departure delay?

In this section, we are deep into another factor: carrier. Different airlines have different levels of management and assets, contributing to different performance. To address, percentage of flights delays respective to the number of flights in each carrier as well as the percentage of flights respective to total number of flights are used as performance indicators.

Percentage delay pattern for each carrier

In this graph, we use two indicators, each of which is demonstrated using different visualizations. For the percentage of flight delays respective to the number of flights in each carrier, we use colored bars with text; whilst for the percentage of flights respective to the total number of flights in NYC, grey bars are chosen. The column chart in the down-right corner depicts exactly the same as the grey bar. Again, the percentage is calculated based on the filtering of above 15 minutes of departure delay.

carriers<-data%>%

group\_by(carrier) %>%

summarise(

percent\_delay = 100 \* mean(delay\_or\_not) ,

percent\_contribution = 100 \* n() / nrow(flights))

main.plot <- ggplot(carriers, aes(x = percent\_delay,

y = reorder(carrier, percent\_delay),

color = percent\_delay)) +

geom\_point(size = 8) +

geom\_col(aes(x = percent\_contribution, y = carrier), alpha = 0.5) +

geom\_segment(aes(xend = 0, yend = carrier), size = 4) +

geom\_text(aes(label = format(percent\_delay, digits = 2)), color = "white", size = 2.5) +

scale\_color\_gradientn(colors = rev(brewer.pal(5, "RdYlBu")[-(2:4)])) +

labs(

title = "Performance of carriers",

subtitle = "-Colored Bars: Percentage of flight delays respective to the number of flights in each carrier

\n-Grey Bars: Percentage of flights respective to total number of flights in NYC",

x = "Percentage(%)",

y = "Carrier",

color = "Proportion of departure delay (%)") +

theme\_bw() +

theme(legend.position = "none") +

scale\_fill\_gradientn(colors = rev(brewer.pal(5, "RdYlBu")[-(2:4)]))

inset.plot <- ggplot(carriers, aes(x = reorder(carrier, percent\_contribution),

y = percent\_contribution)) +

geom\_col() +

labs(

title = "Percentage of flights respective to total number of flights in NYC",

x = "Carrier",

y = "(%)") +

theme\_bw() +

scale\_fill\_gradientn(colors = rev(brewer.pal(5, "RdYlBu")[-(2:4)])) +

theme(legend.position = "none",

plot.title = element\_text(size=10),

plot.subtitle = element\_text(size=20),

title = element\_text(size = 10),

axis.title = element\_text(size = 15))

ggdraw() +

draw\_plot(main.plot) +

draw\_plot(inset.plot, x = 0.64, y = 0.08, width = 0.33, height = 0.27 )

Chart, bar chart

Description automatically generated

Apparently, EV, YV, F9, WN, FL, and 9E have relatively more delays compared to other carriers. However, the flights from YV, F9, and FL contribute tiny proportions with respect to the total number of flights in NYC. DL and AA performed relatively well since they contribute relatively high percentage of flits whilst with a low percentage of delays. Even though flights from UA are in the largest quantity, the delay percentage is around the median. To further investigate and compare, the following graph (4.3.3) is provided.

Percentage pattern of each carrier by origins

Since it is normal for different carriers to deploy resources according to different airports, the following graph shows the carrier performance indicators separated by origins.

data%>%

group\_by(carrier, origin) %>%

summarise(

percent\_delay = 100 \* mean(delay\_or\_not) ,

percent\_contribution = 100 \* n() / nrow(flights))%>%

ggplot(aes(x = percent\_delay,y = reorder(carrier, percent\_delay), color = percent\_delay)) +

geom\_point(size = 8) +

geom\_col(aes(x = percent\_contribution, y = carrier), alpha = 0.5) +

geom\_segment(aes(xend = 0, yend = carrier), size = 4) +

geom\_text(aes(label = format(percent\_delay, digits = 2)), color = "white", size = 2.5) +

scale\_color\_gradientn(colors = rev(brewer.pal(5, "RdYlBu")[-(2:4)])) +

labs(

title = "Performance of carriers per origin",

subtitle = "-Colored Bars: Percentage of flight delays respective to the number of flights in each carrier per origin \n-Grey Bars: Percentage of flights respective to total number of flights in NYC per origin",

x = "Percentage(%)",

y = "Carrier",

color = "Proportion of departure delay (%)") +

theme\_bw() +

theme(legend.position = "none",

plot.title = element\_text(size=20),

plot.subtitle = element\_text(size=15)) +

scale\_fill\_gradientn(colors = rev(brewer.pal(5, "RdYlBu")[-(2:4)])) +

facet\_wrap(~origin)

Chart, bar chart

Description automatically generated

Overall, the delay in LGA is more serious than the other two, since LGA is a domestic airport. To be specific, EV and UA in EWR contribute about 15% of the flights, but have a relatively high percentage rate. Even though EV performed badly in JFK, while EV accounts for a mere 1% of the flights in JKF. B6 in JFK and EV in LGA needs improvement on how to decrease departure delays.

Calendar heat map of average departure delay-UA from EWR

After painting the general picture of the percentage of each carrier and origin, we need to further investigate the time pattern - during which period the carrier performs badly in specific origin. To deal with this, we pick four pairs of carrier and origin: UA - EWR, EV - EWR, EV - LGA, and B6 - JFK according to figure 4.3.2.

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99)) %>%

group\_by(Month,day\_of\_week,hour,carrier,origin)%>%

summarise(delay\_mean=mean(dep\_delay))%>%

filter(carrier=="UA" & origin == "EWR")%>%

ggplot(aes(x=hour,y=day\_of\_week,fill=delay\_mean)) +

geom\_tile(color = "white",size=0.1) +

facet\_wrap(~Month,nrow = 4) +

scale\_fill\_binned(breaks = c(30,60),type = "viridis",name="UA-EWR\nMean delay(min)",)+

labs(x=NULL,y=NULL, title="Average departure delay calandar of UA from EWR")+

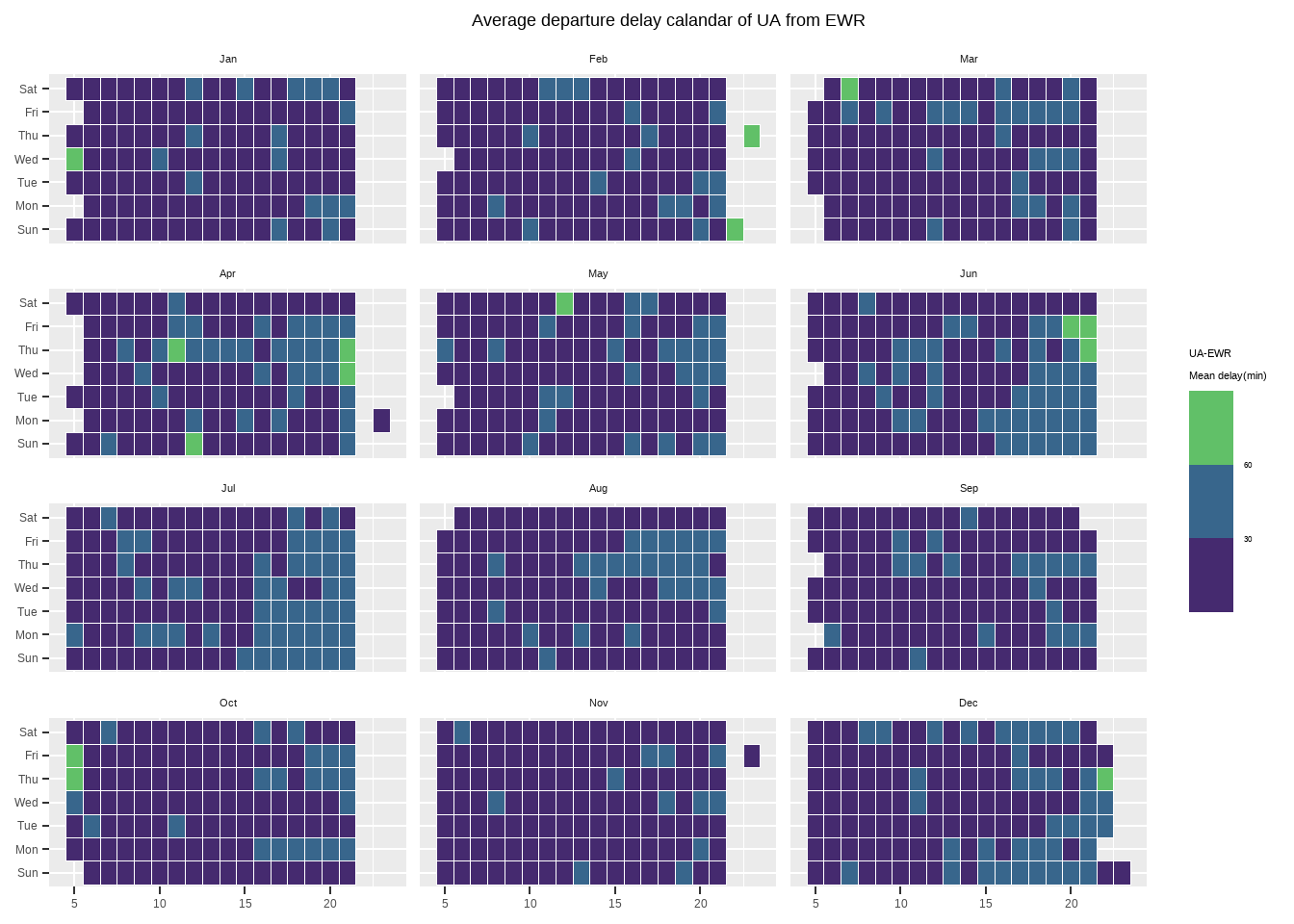
theme(plot.title = element\_text(hjust = 0.6),

legend.title = element\_text(size = 8),

legend.text = element\_text(size = 6),

strip.text.x=element\_text(size=8),

strip.background = element\_blank())



Even though UA contributes to about 15% of the flights in EWR, the delay time mapping is not bad. Throughout 2013, there are few cases when the average delay time is above an hour; and most of the delay time is below 30 minutes.

Calendar heat map of average departure delay-EV from EWR

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99)) %>%

group\_by(Month,day\_of\_week,hour,carrier,origin)%>%

summarise(delay\_mean=mean(dep\_delay))%>%

filter(carrier=="EV" & origin == "EWR")%>%

ggplot(aes(x=hour,y=day\_of\_week,fill=delay\_mean)) +

geom\_tile(color = "white",size=0.1) +

facet\_wrap(~Month,nrow = 4) +

scale\_fill\_binned(breaks = c(30,60),type = "viridis",name="EV-EWR\nMean delay(min)",)+

labs(x=NULL,y=NULL, title="Average departure delay calandar of EV from EWR")+

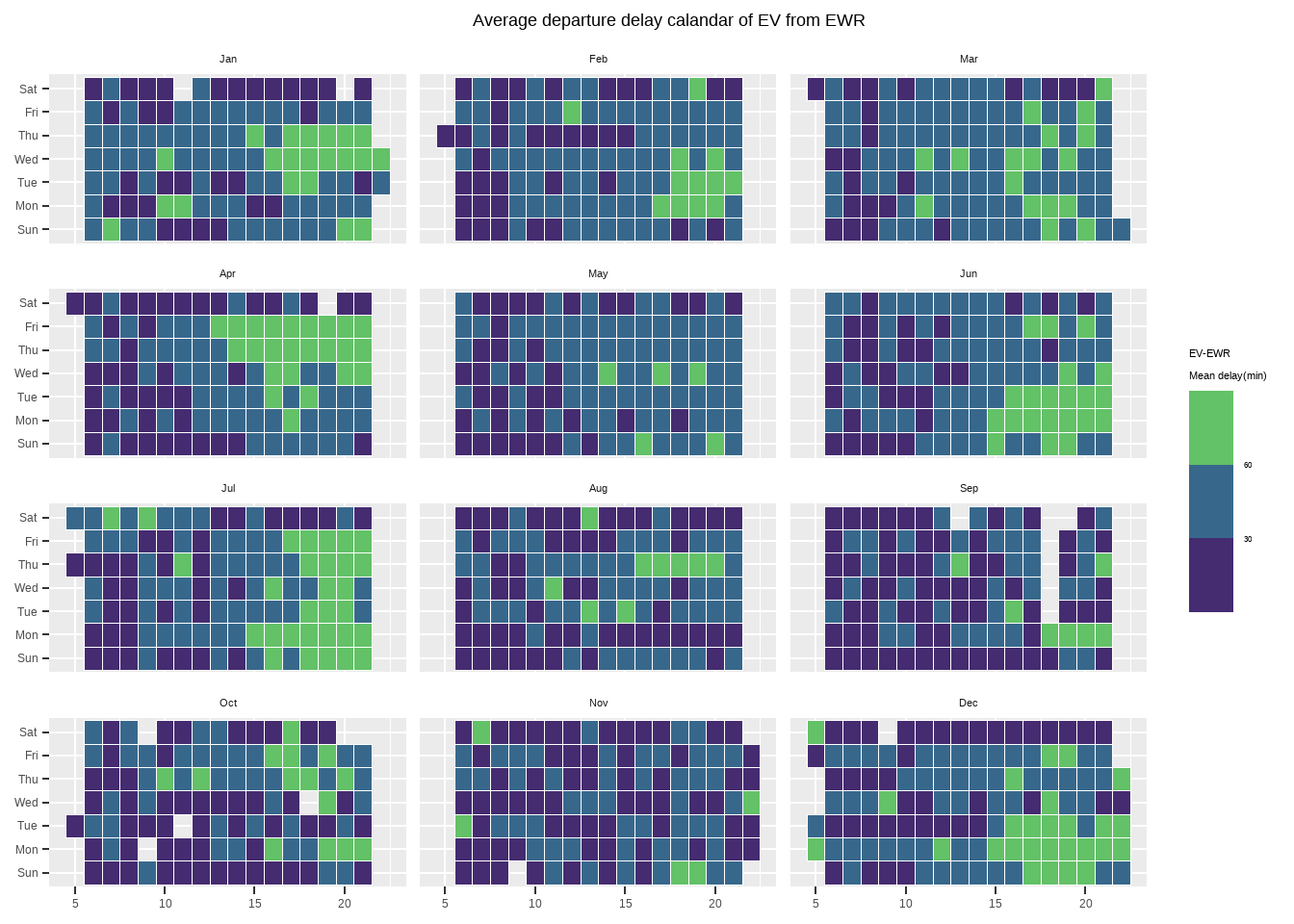
theme(plot.title = element\_text(hjust = 0.6),

legend.title = element\_text(size = 8),

legend.text = element\_text(size = 6),

strip.text.x=element\_text(size=8),

strip.background = element\_blank())



From figure 4.3.2, even though EV contributes about 2% of flights than UA in EWR, the overall delay percentage and time series pattern indicate that EV performs worse than UA. EV has serious delays in summer and winter, especially in the late afternoon and evening.

Calandar heatmap for average delay of EV from LGA

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99)) %>%

group\_by(Month,day\_of\_week,hour,carrier,origin)%>%

summarise(delay\_mean=mean(dep\_delay))%>%

filter(carrier=="EV" & origin == "LGA")%>%

ggplot(aes(x=hour,y=day\_of\_week,fill=delay\_mean)) +

geom\_tile(color = "white",size=0.1) +

facet\_wrap(~Month,nrow = 4) +

scale\_fill\_binned(breaks = c(30,60),type = "viridis",name="EV-LGA\nMean delay(min)",)+

labs(x=NULL,y=NULL, title="Average departure delay calandar of EV from LGA")+

theme(plot.title = element\_text(hjust = 0.6),

legend.title = element\_text(size = 8),

legend.text = element\_text(size = 6),

strip.text.x=element\_text(size=8),

strip.background = element\_blank())

Calendar

Description automatically generated

Compared to the performance in EWR, as shown in 4.3.4, EV performs worse in LGA, with more times of long delays throughout the year. One of the possible reasons is that LGA is a domestic airport, whilst the EWR is an international one; there are fewer delays in international airports compared to that in the domestic one.

Calendar heat map of average departure delay-B6 from JFK

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99)) %>%

group\_by(Month,day\_of\_week,hour,carrier,origin)%>%

summarise(delay\_mean=mean(dep\_delay))%>%

filter(carrier=="B6" & origin == "JFK")%>%

ggplot(aes(x=hour,y=day\_of\_week,fill=delay\_mean)) +

geom\_tile(color = "white",size=0.1) +

facet\_wrap(~Month,nrow = 4) +

scale\_fill\_binned(breaks = c(30,60),type = "viridis",name="B6-JFK\nMean delay(min)",)+

labs(x=NULL,y=NULL, title="Average departure delay calandar of B6 from JFK")+

theme(plot.title = element\_text(hjust = 0.6),

legend.title = element\_text(size = 8),

legend.text = element\_text(size = 6),

strip.text.x=element\_text(size=8),

strip.background = element\_blank())

Diagram

Description automatically generated

B6 performs fairly in JFK, with less extreme delays in contrast to EV. During June and July, B6 has long delays with over 60 minutes. One possible way to improve is to apply more robust and advanced weather forecast techniques to better predict the weather condition. And at the same time, to ensure the efficiency of preparation of re-departure, better optimization regarding staff management is necessary.

Question 4 - How does the plane type and age affect the departure delay?

Plane relating factors are the last one to address. Within the data set of planes, we choose type and year (which is used to calculate the age of planes). To investigate, we need to combine the age and type variables with other factors. For usage, we assume the larger the age is, the worse performance of the carrier; while for type, we suppose that there are some differences between types against wind speed.

Boxplot for carriers VS plane age

To investigate plane age, we use box plot, which could clearly depict the distribution of data. With the order from the largest percentage delay to the smallest, the plane age of each carrier is demonstrated in the following:

summary\_percentage<-data%>%

*# Sort by percent\_contribution: levels =*

mutate(carrier\_ordered=factor(carrier,levels = c("HA","US","AS","AA","DL","VX","OO","UA","MQ","B6",

"9E","FL","WN","F9","YV","EV"),ordered=TRUE))%>%

group\_by(carrier\_ordered) %>%

summarise(percent\_delay = 100 \* mean(delay\_or\_not))

data%>%

*# with only one left\_join we find that there 13.77% of NA - we need imputation technique*

*# Imputation: year with median because it is high left skewed and a lot of outlier*

mutate(year\_replace\_with\_median = ifelse(is.na(year), median(year, na.rm = TRUE), year))%>%

*# create new variable year\_usage*

mutate(year\_usage = 2013-year\_replace\_with\_median) %>%

group\_by(carrier)%>%

*# Sort by percent\_contribution: levels =*

mutate(carrier\_ordered=factor(carrier,levels = c("HA", "US", "AS","AA","DL","VX","OO","UA","MQ","B6",

"9E","FL","WN","F9","YV","EV"),ordered=TRUE)) %>%

ggplot(aes(x = carrier\_ordered, y = year\_usage)) +

geom\_col(data= summary\_percentage,aes(x = carrier\_ordered, y = percent\_delay),fill="black",alpha=0.1 )+

geom\_boxplot(aes(x = carrier\_ordered, y = year\_usage)) +

coord\_flip() +

labs(

title = "The age of airplanes in service by carriers",

x = "Carrier",

y = "Age of airplanes (year)",) +

theme\_bw() +

geom\_hline(aes(yintercept = mean(year\_usage),

linetype = "Average of airplanes age in NYC"),

color = 'red') +

geom\_hline(aes(yintercept = 20,

linetype = "Average insdustrial airplane retirement"),

color = 'blue') +

scale\_linetype\_manual(name = "", values = c(2, 2),

guide = guide\_legend(override.aes = list(color = c("blue", "red")))) +

scale\_y\_continuous(sec.axis = sec\_axis(~., name = "percent delay (%)"))

Chart

Description automatically generated

The red and blue lines are the average life span of planes from the industry average and from all NYC planes. Unfortunately, those carriers which perform badly do not have old planes as we supposed. Even the worst performed carriers have rather new planes.

Density plot for wind speed VS departure delay per plane type

By curiosity, we would like to know whether different types of plane can cope with different levels of wind speed. Hence, we planned to look further into the wind speed factor together with the plane data set. Starting from the very beginning, we merge the three data set into data, which includes plane type, wind speed and departure delay we need.

data%>%

filter(dep\_delay > 0, dep\_delay < quantile(dep\_delay, 0.99))%>%

filter(!is.na(type))%>%

ggplot(aes(x=wind\_speed, y=dep\_delay))+

*#Density plot for windspeed VS departure delay per plane type*

stat\_density\_2d(aes(fill = ..level..), geom = "polygon") +

scale\_fill\_viridis\_c() +

facet\_wrap(~type)+

labs(title = "Density plot for windspeed VS departure delay per plane type",

x = "Wind speed (mph)",

y = "Departure delay (minutes)") +

theme\_bw()+

theme(plot.title = element\_text(hjust = 0.5))

Chart

Description automatically generated

The density plot indicates that the majority of fixed wing single engine planes and rotorcraft planes have 10 minutes departure delay at 10 mph wind speed, whereas the majority of fixed wing multi engine planes only have 5 minutes departure delay at 10 mph. It appears that the fixed wing multi engine plane might have a higher ability to cope with wind speed compared to the other 2 types of plane.

Limitations

* Flight data between 0:00 to 04:59 is missing. The absence of data might cause bias in the analysis.
* Recommendations made are based on the exploratory analysis of historical data in 2013, which might not fully reflect the current conditions of departure delays in the NYC airports.
* The exploratory analysis is focused on visualising the main characteristics of datasets and cannot explain the underlying causality of departure delays. Further data analysis and algorithms should be done to achieve a more accurate conclusion.
* Inaccuracy in the analysis may occur since most of the data are not normally distributed (or skewed) which will lead to a higher standard error of observation trends.
* Inaccuracy in the analysis may occur since weather factors are likely to correlate with each other, resulting in misinterpretation.
* Inaccuracy in the analysis may occur since there might be other factors that are not studies but influencing the departure delays.
* The exploratory analysis is focused on visualising the main characteristics of datasets and cannot explain the underlying causality of departure delays.

Summary and recommandations

The Port Authority of New York and New Jersey operates the 3 major airports including JFK, EWR and LGA in NYC. The analysis is focused on the historical data in 2013 and is aimed at understanding the general trends in flights and performance of NYC airports as to examine different factors related to departure delays. With all careful analysis and data exploration, recommendations are provided in three aspects to address the long departure delay issue in NYC airports.

**Delay pattern throughout the year of 2013**

Diagram

Description automatically generated with medium confidence

To answer whether there is any particular departure delay pattern in NYC, it is worth exploring the delay pattern in terms of time series. We gathered all departing flights details in 3 airports throughout the year 2013 and created a calendar for percentage departure delay. By observing the density of yellow grids, it is obvious that April, June, July and December have a higher percentage of departure delay. The delayed pattern seems to be following the holiday occasions in the United States in which Easter, Summer Holiday and Christmas are in April, June and July, December respectively. It is also clear that a high percentage of delays are observed at night, which is possibly caused by the propagation delay – a subsequent flight might be delayed because it is awaiting the takeoff of the previous flight.

Based on the above findings in time series analysis, it is suggested that the port authority should modify the scheduled departure time for those busy months. A flexible slack time should be re-allocated in the flight schedule in which the ability of departure delay absorption during peak hours should be stronger than that of other months. The enhanced flight schedule could reduce the propagation delay caused by the preceding flight delays and reduce the phenomenon of flights are likely to delay at night. On the other hand, it is foreseeable that passenger volume will substantially increase during holiday periods and often resulting in long queues for flight check-in and security check. It is suggested to increase the air carrier crew during the holiday period to withstand the high pressure of check-in volume and baggage loading problems. More check-in counters should be opened for passengers to increase efficiency and ultimately prevent the departure delay caused by ground staff.

**Carrier performance**



Undoubtedly, aircraft departure delay is often associated with the carrier performance, whereas different carriers might have various performances in different airports. The cause of delay might be due to many circumstances within the airline’s control such as crew onboard problems, catering arrangements, aircraft cleaning, etc. It is important to know which carriers have been the bottom performers by looking into the percentage of flight delays per each carrier (colored bar) and the proportion of flights provided by the carrier (grey bar).

As indicated by the graph, for EWR airport, EV and UA had the worst performance with 31% and 22% of their flights were delayed respectively. Since both carriers provided a high number of flights, they contributed a lot of flight delays in NYC. For JFK airport, B6 and 9E were the bad performers with 23% and 28% of their flights being delayed respectively. Whereas B6 contributed 13% of flights and 9E contributed 4% of flights in NYC. Considering EWR and JFK airport serves as major international hubs, the delays of those airlines are probably caused by waiting for connecting passengers, bags or crews that are arriving from another destination. It is suggested that a clear guideline should be provided for any connecting passengers to make sure they follow the flight transit time straightly. In addition, those airlines should ensure enough or even additional crew on staff during peak hours in case of unexpected situations that the original crew is delayed by other flights.

For LGA, although we observed a lot of red bars with high percentages, those carriers did not contribute much to the total departure delays as they only offer very few flights in LGA. Instead, we can conclude that EV was the bad performer with 16% of its flights being delayed and contributing a relatively high number of flights. As LGA is the flagship domestic airport in NYC, a domestic aircraft will typically be used for more flights per day compared to ones used internationally. Again, a ripple effect might occur if the preceding domestic flight has a significant delay. In consequence, domestic flights are likely to have less time for preparing the aircraft such as cleaning, loading and unloading catering equipment and supplies, etc. Therefore, it is suggested that the EV airline should collaborate with the airline catering agent and cleaning agent by sending more staff for aircraft preparation during the peak hours in LGA airport. The route for catering trucks transportation should also be optimized to efficiently decrease the unnecessary travel time within the airport ground.

**Weather**

Chart, box and whisker chart

Description automatically generated

Significant meteorological conditions might cause departure delays. As indicated by the graph, flights are mostly affected by high precipitation and high relative humidity level which shows that heavy rainfall will contribute to departure delays. On the other hand, visibility also adversely affects the flight departure when it drops to 3 miles or below.

Although it is impossible to alter the weather, an upgrade in the runway lighting system might be useful to withstand the low visibility conditions resulting from fog or heavy rainfall. Furthermore, it is always better to get ready for service resumption when the weather becomes normal again. It is suggested that Airlines should always update their weather forecasting techniques or tools so that they can deal with the aftermath as soon as possible to minimize the flight departure delay time.