Assignment(Exploratory analysis of Flight data)

MD300-0006/2020-Adama Dembele

10/10/2021

In this note we are analyzing $566,996$ observations data entries from the flights FAA data describing every commercial flight during the month of December 2009.The data comes from the Research and Innovation Technology Administration at the Bureau of Transportation statistics. # PERFORMANCE OF AIRLINE COMPAGNY ## Loading of data

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(scales)
library(highcharter)
```

```
## Warning: package 'highcharter' was built under R version 4.1.1
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.1.1
```

```
fly_data=read.csv("C:/Users/user/Documents/36169029_T_ONTIME.csv")
fly_data=fly_data[,-c(27,31,32,33,34,35,36)]
```

Exploratory Data Analysis

Our objectives is to check which airlines compagny perform the most in terms of departure time, and less cancelled flight. ### Understanding the data contains This dataset is composed by the following variables: \

```
colnames(fly_data)
```

```
##  [1] "YEAR"              "MONTH"             "DAY_OF_MONTH"      "DAY_OF_WEEK"
##  [5] "UNIQUE_CARRIER"    "AIRLINE_ID"        "CARRIER"           "TAIL_NUM"
##  [9] "FL_NUM"            "ORIGIN"            "ORIGIN_CITY_NAME" "ORIGIN_STATE_ABR"
## [13] "ORIGIN_STATE_NM"  "ORIGIN_WAC"        "DEST"              "DEST_CITY_NAME"
## [17] "DEST_STATE_ABR"    "DEST_STATE_NM"     "DEST_WAC"          "CRS_DEP_TIME"
## [21] "DEP_TIME"          "DEP_DELAY"         "CRS_ARR_TIME"      "ARR_TIME"
## [25] "ARR_DELAY"         "CANCELLED"         "DIVERTED"          "AIR_TIME"
## [29] "DISTANCE"
```

```
head(fly_data,4)
```

```
##   YEAR MONTH DAY_OF_MONTH DAY_OF_WEEK UNIQUE_CARRIER AIRLINE_ID CARRIER
## 1 2009    12            2           3             9E      20363      9E
## 2 2009    12            3           4             9E      20363      9E
## 3 2009    12            4           5             9E      20363      9E
## 4 2009    12            6           7             9E      20363      9E
##   TAIL_NUM FL_NUM ORIGIN ORIGIN_CITY_NAME ORIGIN_STATE_ABR ORIGIN_STATE_NM
## 1   91879E    850    ATL      Atlanta, GA               GA         Georgia
## 2   92289E    850    ATL      Atlanta, GA               GA         Georgia
## 3   91629E    850    ATL      Atlanta, GA               GA         Georgia
## 4   91709E    850    ATL      Atlanta, GA               GA         Georgia
##   ORIGIN_WAC DEST     DEST_CITY_NAME DEST_STATE_ABR  DEST_STATE_NM DEST_WAC
## 1         34  RDU Raleigh/Durham, NC             NC North Carolina       36
## 2         34  RDU Raleigh/Durham, NC             NC North Carolina       36
## 3         34  RDU Raleigh/Durham, NC             NC North Carolina       36
## 4         34  RDU Raleigh/Durham, NC             NC North Carolina       36
##   CRS_DEP_TIME DEP_TIME DEP_DELAY CRS_ARR_TIME ARR_TIME ARR_DELAY CANCELLED
## 1         1755     1803         8         1929     1943        14         0
## 2         1755     1752        -3         1929     1924        -5         0
## 3         1755     1754        -1         1929     1920        -9         0
## 4         1755     1750        -5         1929     1917       -12         0
##   DIVERTED AIR_TIME DISTANCE
## 1        0       61      356
## 2        0       49      356
## 3        0       53      356
## 4        0       51      356
```

A. Data Preproccessing

1) Encoding categorical features

```
fly_data$UNIQUE_CARRIER=factor(fly_data$UNIQUE_CARRIER)
fly_data$ORIGIN_CITY_NAME=factor(fly_data$ORIGIN_CITY_NAME)
fly_data$CANCELLED=ifelse(fly_data$CANCELLED==1,"Yes","No")
fly_data$DAY_OF_WEEK[fly_data$DAY_OF_WEEK==1]="Monday"
fly_data$DAY_OF_WEEK[fly_data$DAY_OF_WEEK==2]="Tuesday"
fly_data$DAY_OF_WEEK[fly_data$DAY_OF_WEEK==3]="Wednesday"
fly_data$DAY_OF_WEEK[fly_data$DAY_OF_WEEK==4]="Thursday"
fly_data$DAY_OF_WEEK[fly_data$DAY_OF_WEEK==5]="Friday"
fly_data$DAY_OF_WEEK[fly_data$DAY_OF_WEEK==6]="Saturday"
fly_data$DAY_OF_WEEK[fly_data$DAY_OF_WEEK==7]="Sunday"
```

2) Checking missing values in dataset

```
missing_check=function(dat){
  sum(is.na(dat))
}
```

```
apply(fly_data,2,missing_check)
```

```
##             YEAR            MONTH     DAY_OF_MONTH      DAY_OF_WEEK
##                0                0                0                0
##   UNIQUE_CARRIER       AIRLINE_ID          CARRIER         TAIL_NUM
##                0                0                0                0
##           FL_NUM           ORIGIN ORIGIN_CITY_NAME ORIGIN_STATE_ABR
##                0                0                0                0
##  ORIGIN_STATE_NM       ORIGIN_WAC             DEST   DEST_CITY_NAME
##                0                0                0                0
##   DEST_STATE_ABR    DEST_STATE_NM         DEST_WAC     CRS_DEP_TIME
##                0                0                0                0
##         DEP_TIME        DEP_DELAY     CRS_ARR_TIME         ARR_TIME
##            14193            14193                0            15178
##        ARR_DELAY        CANCELLED         DIVERTED         AIR_TIME
##            16218                0                0            16218
##         DISTANCE
##                0
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.1.1
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
dep_delay_impute=with(fly_data,impute(DEP_DELAY,mean))
```

```
fly_data["dep_delay_impute"]=dep_delay_impute
```

B. Questions

1.When does airline cancellation happens?

1.1 What is the most popular day of the week that most flight are cancelled?

```
tbles=fly_data%>%group_by(DAY_OF_WEEK)%>%count(CANCELLED)%>%filter(CANCELLED=="Yes")%>%arrange(desc(cancelled_num
ber_by_days=n))
tables=as.data.frame(tbles[,c(1,3)])
colnames(tables)=c("Day_of_Weeks","cancelled_number_by_days")
knitr::kable(tables,caption = "Table of the number of airline cancellation days")
```

Table of the number of airline cancellation days

Day_of_Weeks	cancelled_number_by_days
Saturday	3670
Sunday	2625
Tuesday	2037
Wednesday	1890
Friday	1682
Thursday	1433
Monday	1393

1.2 How likely a flight cancellation would happen?

```
statistic_details=describe(fly_data)
#statistic_details$CANCELLED$values
```

```
knitr::kable(as.data.frame(statistic_details$CANCELLED$values),caption = "Table of cancelled frequency")
```

Table of cancelled frequency

value	frequency
No	514539
Yes	14730

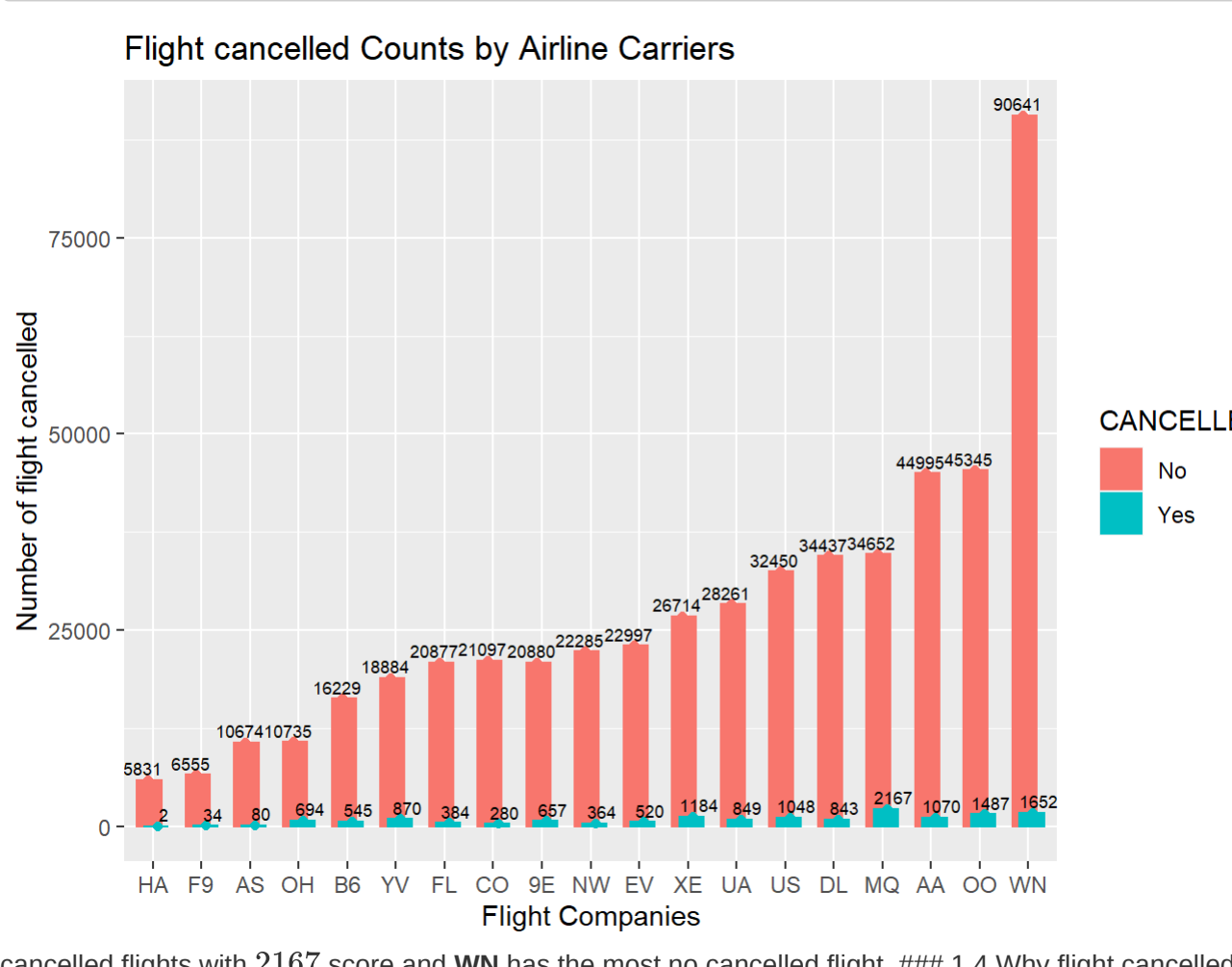
probability_of_cancelled=$14730/514539 = 0.028$ It is very unlikely because all of airline has 0.028 chance to have a flight cancelled.

1.3 What are the flight number cancelled and no cancelled by airline?

```
cancel=fly_data%>%group_by(UNIQUE_CARRIER)%>%count(CANCELLED)%>%arrange(desc(CANCELLED_sum=n()))
```

The most common day for flight cancel is on Saturday.

```
CANCELLED_fly=ggplot(cancel, aes(x = reorder(UNIQUE_CARRIER,n), y = n))+
geom_bar(
aes(color = CANCELLED, fill = CANCELLED),
stat = "identity", position = position_dodge(0.3),
width = 1
)+
  xlab("Flight Companies")+
  ylab("Number of flight cancelled")+
geom_point(
aes(color = CANCELLED),
position = position_dodge(0.4), size = 1.5
)+labs(title = "Flight cancelled Counts by Airline Carriers")+
  geom_text(
aes(label = n, group = CANCELLED),
position = position_dodge(0.9),
vjust = -0.4, size = 2.4
)+theme_set(theme_gray())
CANCELLED_fly
```

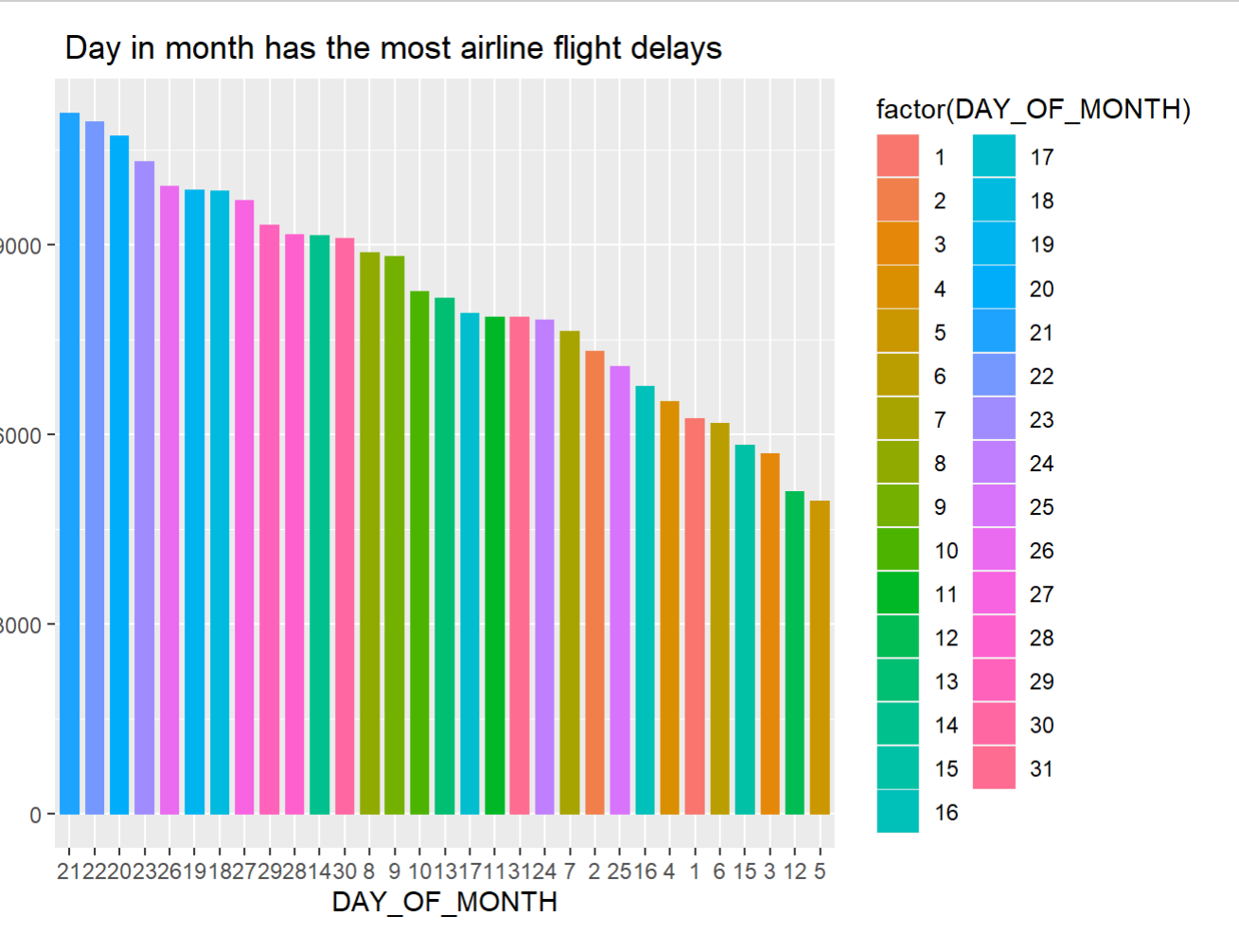


The **MQ** airline has the most cancelled flights with 2167 score and **WN** has the most no cancelled flight. ### 1.4 Why flight cancelled? We can not answer this question because the dataset don't allow it, many features which are related to explain that relationship, have too many missing values. ## 2. When does airline delays happens? ### 2.1 Which data in month has the most airline flight delays?

```
fly_data$dep_delay_impute=ifelse(fly_data$dep_delay_impute>0,"Delay","No Delay")
```

```
month_day=fly_data%>%group_by(DAY_OF_MONTH)%>%count(dep_delay_impute)%>%filter(dep_delay_impute=="Delay")%>%arran
ge(desc(n))
#ggplot(datas,aes(x=factor(DAY_OF_MONTH),y=n))+geom_bar()
```

```
ggplot(month_day, aes(x = reorder(factor(DAY_OF_MONTH),-n), y = n))+
geom_bar(
aes(color =factor(DAY_OF_MONTH) , fill =factor(DAY_OF_MONTH)),
stat = "identity", position = position_dodge(0.3),
width = 0.7
)+labs(x="DAY_OF_MONTH",y="data counts in month",title = " Day in month has the most airline flight delays")
```



The most common data in a month for flight delays is on 21nd day.

2.3 Which airline has the most delay flights?

```
delay_days=fly_data%>%group_by(UNIQUE_CARRIER)%>%count(dep_delay_impute)%>%arrange(n())%>%mutate(Flight_proportio
n = round(n*100/sum(n), 1),
lab.ypos = cumsum(Flight_proportion) - 0.5*Flight_proportion)
head(delay_days,8)
```

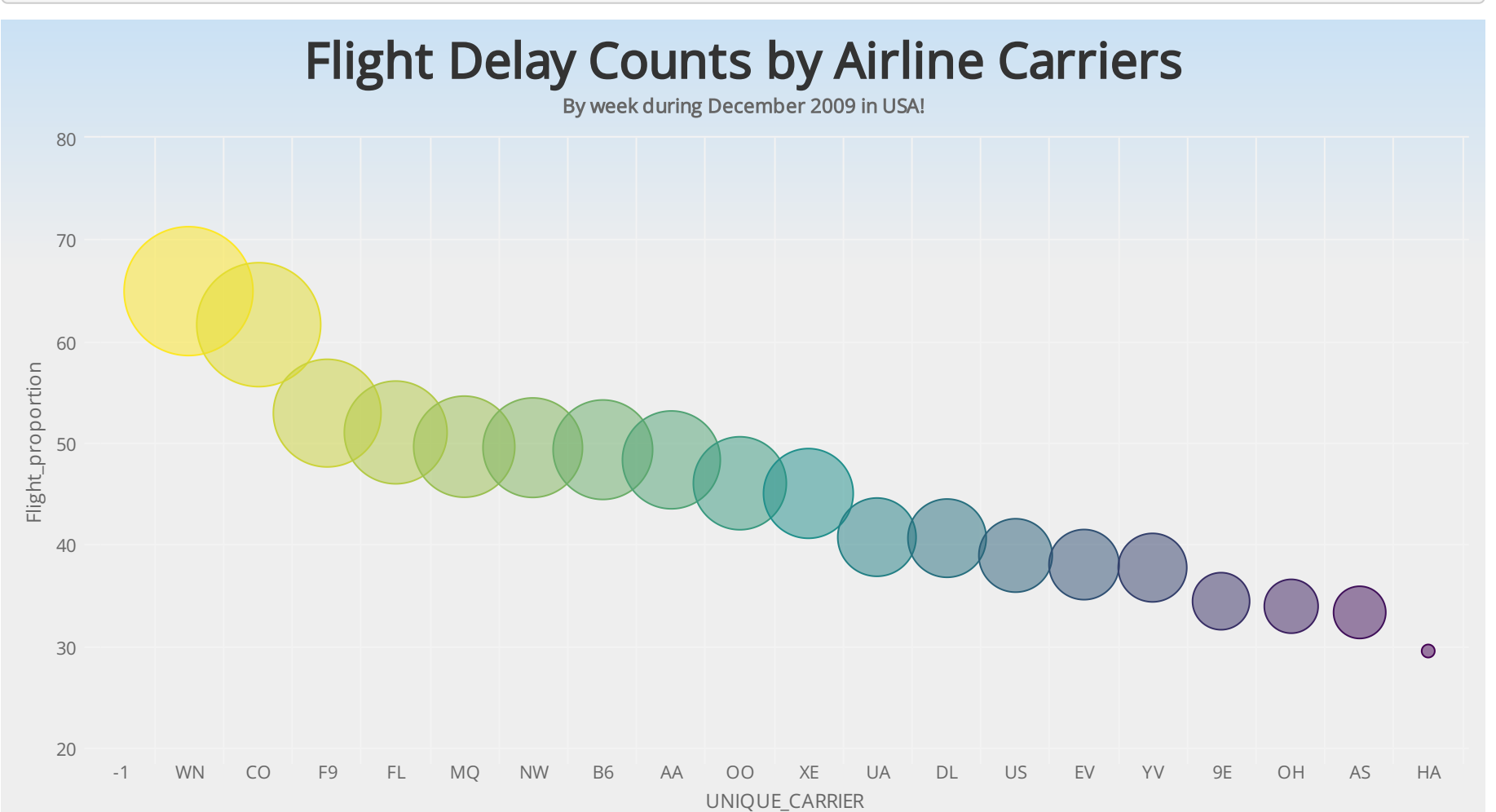
```
## # A tibble: 8 x 5
## # Groups:   UNIQUE_CARRIER [4]
##   UNIQUE_CARRIER dep_delay_impute     n Flight_proportion lab.ypos
##   <fct>          <chr>            <int>             <dbl>    <dbl>
## 1 9E             Delay             7403              34.4     17.2
## 2 9E             No Delay         14134              65.6     67.2
## 3 AA             Delay            22262              48.3     24.2
## 4 AA             No Delay         23803              51.7     74.2
## 5 AS             Delay             3579              33.3     16.6
## 6 AS             No Delay          7175              66.7     66.6
## 7 B6             Delay             8276              49.3     24.6
## 8 B6             No Delay          8498              50.7     74.6
```

```
delay_airline=delay_days%>%group_by(UNIQUE_CARRIER,Flight_proportion)%>%filter(dep_delay_impute=="Delay")%>%arran
ge(desc(Flight_proportion))
head(delay_airline,5)
```

```
## # A tibble: 5 x 5
## # Groups:   UNIQUE_CARRIER, Flight_proportion [5]
##   UNIQUE_CARRIER dep_delay_impute     n Flight_proportion lab.ypos
##   <fct>          <chr>            <int>             <dbl>    <dbl>
## 1 WN             Delay            59931              64.9     32.4
## 2 CO             Delay            13171              61.6     30.8
## 3 F9             Delay             3488              52.9     26.4
## 4 FL             Delay            10842              51       25.5
## 5 MQ             Delay            18260              49.6     24.8
```

```
Nodelay_airline=delay_days%>%group_by(UNIQUE_CARRIER,Flight_proportion)%>%filter(dep_delay_impute=="No Delay")%>%
arrange(desc(Flight_proportion))
```

```
hc <- delay_airline%>%
  hchart(
    "bubble", hcaes(x =UNIQUE_CARRIER , y =Flight_proportion,size=Flight_proportion,color =
 Flight_proportion), maxSize = "20%"
    )%>%
  hc_xAxis(text="Airline company")%>%
 hc_title(text ="Flight Delay Counts by Airline Carriers",
 align = "center",
 style = list(fontWeight = "bold", fontSize = "30px")) %>%
  hc_legend(enabled = T)%>%
 hc_tooltip(pointFormat='{point.y:.2f}%',enabled = T) %>%
 hc_subtitle(text = "By week during December 2009 in USA!",
 align = "center",
 style = list(fontWeight = "bold")) %>%
 hc_add_theme(hc_theme_ffx())
hc
```



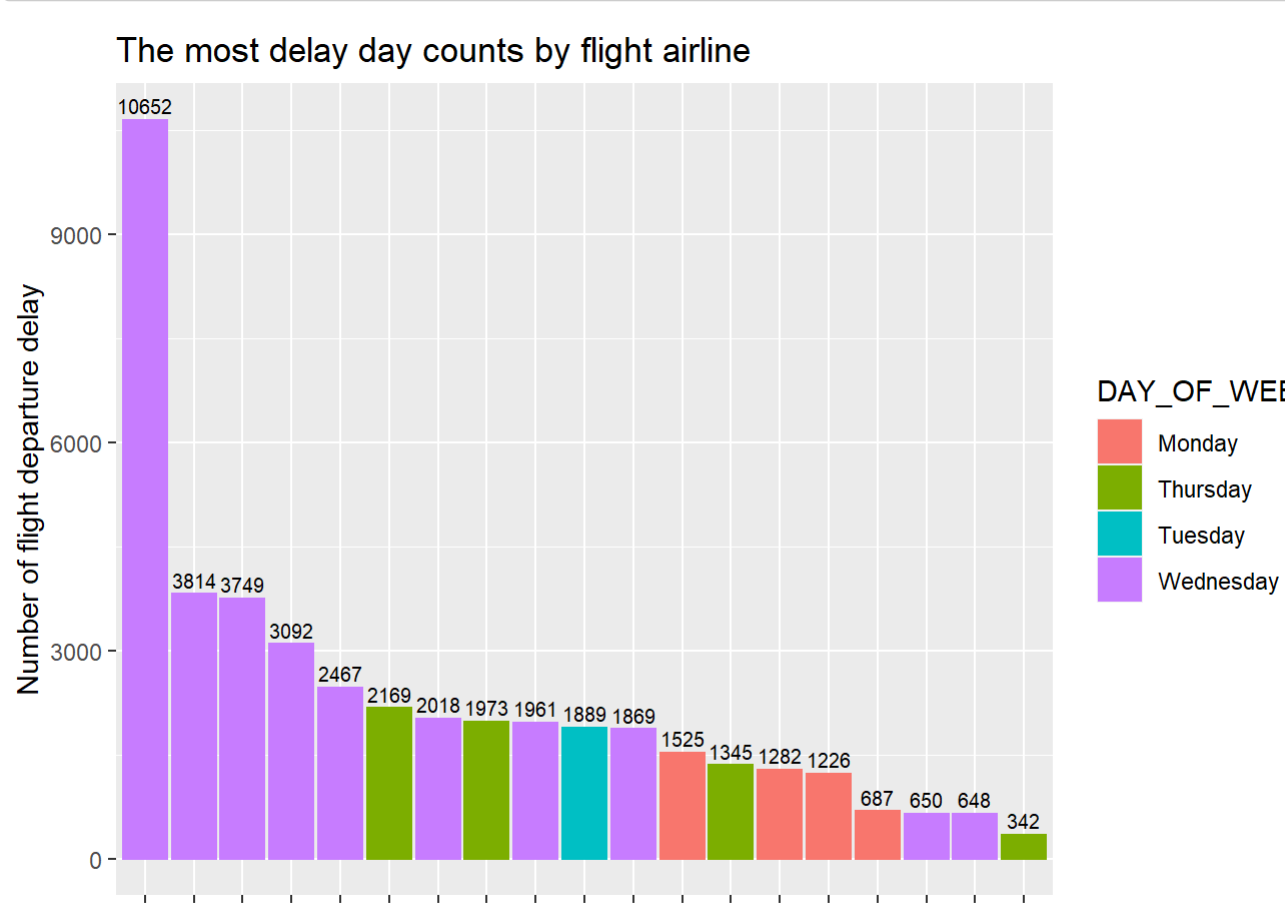
The **WN** airline has the most delay flights.

2.2 The most delay day by flight airline

```
day_delays=fly_data%>%group_by(UNIQUE_CARRIER,DAY_OF_WEEK)%>%count(dep_delay_impute)%>%filter(dep_delay_impute=
="Delay")
```

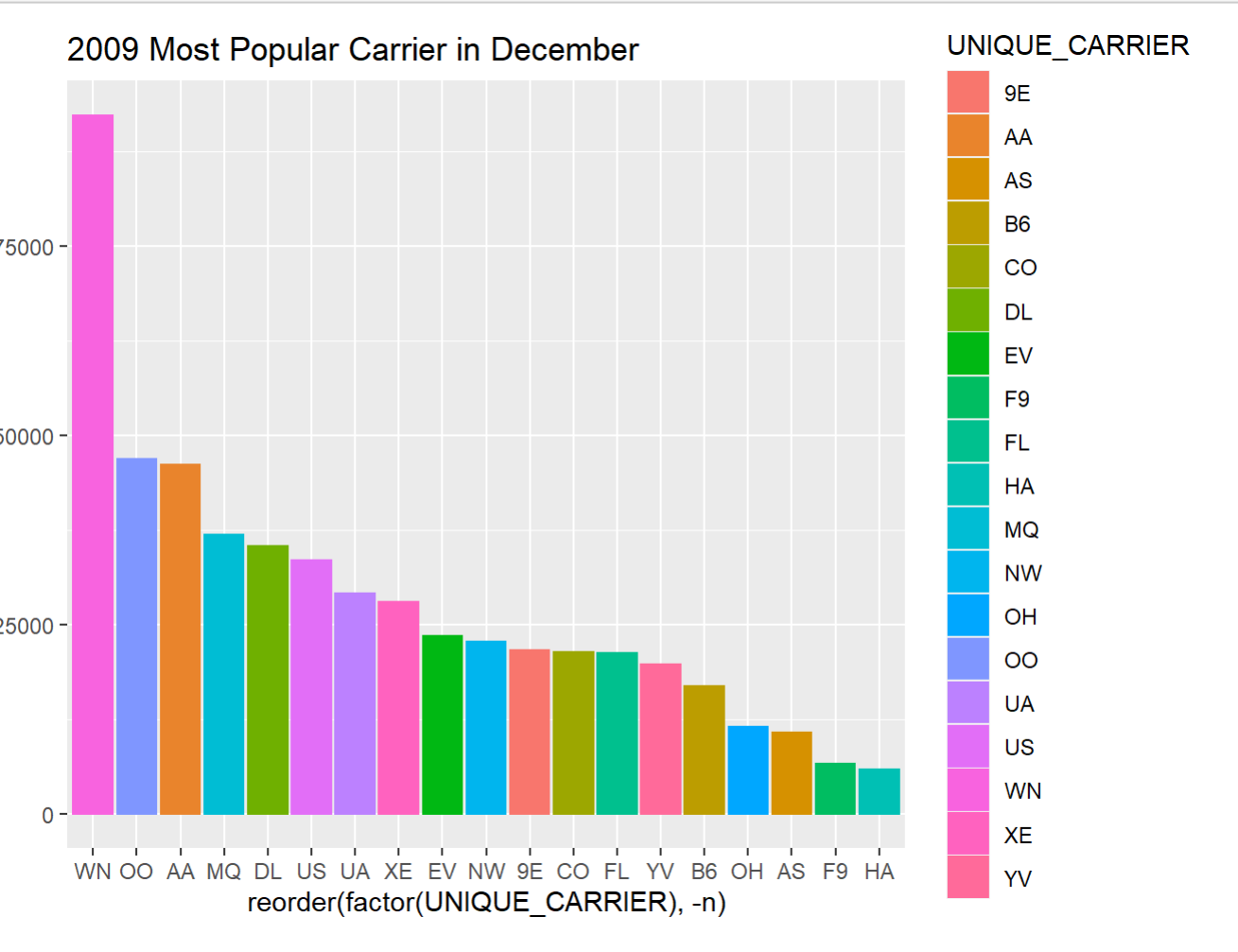
```
most_delay_dayss=day_delays%>%group_by(UNIQUE_CARRIER)%>%filter(n==max(n))%>%arrange(desc(n))
most_delay_days=ggplot(most_delay_dayss, aes(x =reorder(UNIQUE_CARRIER,-n), y = n))+
geom_bar(
aes(color = DAY_OF_WEEK, fill = DAY_OF_WEEK),stat = "identity", position = position_dodge(width=1),
width = 0.9
)+geom_text(
aes(label = n),
position = position_dodge(0.8),
vjust = -0.4, size = 2.6
)+
  xlab("Flight Companies")+
  ylab("Number of flight departure delay")+
  labs(title = "The most delay day counts by flight airline")

most_delay_days
```



The **WN** airline has the most delay flights on **Wednesday**. ### Which is the most Popular airline Carrier 2009?

```
popular_fly=fly_data%>%group_by(UNIQUE_CARRIER)%>%count(counts=n())%>%arrange(desc(n))
f=ggplot(popular_fly,aes(x=reorder(factor(UNIQUE_CARRIER),-n),y=n))+geom_bar(aes(color = UNIQUE_CARRIER, fill = U
NIQUE_CARRIER),stat = "identity", position = position_dodge(width=1),
width = 0.9)+labs("Airline carrier","Counts number carrier",title="2009 Most Popular Carrier in December")
f
```



Conclusion

-In our analysis, the worst day for not traveling is Saturday, December and 21nd to travel and avoid **MQ** airline. -Flight cancelation is very unlikely. -The **WN** airline is the most efficient in terms of departure delays and no cancelled flight in December 2009.