NYC Flight Predictive Project

Ismail Olasege

11/24/2020

Data Exploration and Visualization

This project is about exploring and visualizing flights data that departed from New York City

Import the csv file and explore it using str and summary functions.

```
flights <- read.csv("NYC_Flights.csv")
head(flights)
```

```
##   year month day dep_time dep_delay arr_time arr_delay cancelled carrier origin
## 1 2014     1   1      914        14     1238        13         0      AA    JFK
## 2 2014     1   1     1157        -3     1523        13         0      AA    JFK
## 3 2014     1   1     1902         2     2224         9         0      AA    JFK
## 4 2014     1   1      722        -8     1014       -26         0      AA    LGA
## 5 2014     1   1     1347         2     1706         1         0      AA    JFK
## 6 2014     1   1     1824         4     2145         0         0      AA    EWR
##   dest distance
## 1  LAX     2475
## 2  LAX     2475
## 3  LAX     2475
## 4  PBI     1035
## 5  LAX     2475
## 6  LAX     2454
```

```
str(flights)
```

```
## 'data.frame':    253316 obs. of  12 variables:
##  $ year     : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
##  $ month    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time : int  914 1157 1902 722 1347 1824 2133 1542 1509 1848 ...
##  $ dep_delay: int  14 -3 2 -8 2 4 -2 -3 -1 -2 ...
##  $ arr_time : int  1238 1523 2224 1014 1706 2145 37 1906 1828 2206 ...
##  $ arr_delay: int  13 13 9 -26 1 0 -18 -14 -17 -14 ...
##  $ cancelled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ carrier  : Factor w/ 14 levels "AA","AS","B6",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ origin   : Factor w/ 3 levels "EWR","JFK","LGA": 2 2 2 3 2 1 2 2 2 2 ...
##  $ dest     : Factor w/ 109 levels "ABQ","ACK","AGS",..: 53 53 53 75 53 53 53 53 62 94 ...
##  $ distance : int  2475 2475 2475 1035 2475 2454 2475 2475 1089 2422 ...
```

```
summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2014   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2014   1st Qu.: 3.000   1st Qu.: 8.00   1st Qu.: 902
##  Median :2014   Median : 6.000   Median :16.00   Median :1347
##  Mean   :2014   Mean   : 5.639   Mean   :15.89   Mean   :1338
##  3rd Qu.:2014   3rd Qu.: 8.000   3rd Qu.:23.00   3rd Qu.:1734
##  Max.   :2014   Max.   :10.000   Max.   :31.00   Max.   :2400
##
##    dep_delay          arr_time      arr_delay          cancelled
##  Min.   :-112.00   Min.   :   1   Min.   :-112.000   Min.   :0
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.: -15.000   1st Qu.:0
##  Median :  -1.00   Median :1519   Median :  -4.000   Median :0
##  Mean   :  12.47   Mean   :1494   Mean   :   8.147   Mean   :0
##  3rd Qu.:  11.00   3rd Qu.:1934   3rd Qu.:  15.000   3rd Qu.:0
##  Max.   :1498.00   Max.   :2400   Max.   :1494.000   Max.   :0
##
##     carrier      origin           dest           distance
##  UA     :46267   EWR:87400   LAX    : 14434   Min.   :  80
##  B6     :44479   JFK:81483   ATL    : 12808   1st Qu.: 533
##  DL     :41683   LGA:84433   SFO    : 11907   Median : 944
##  EV     :39819               MCO    : 11709   Mean   :1099
##  AA     :26302               BOS    : 11609   3rd Qu.:1416
##  MQ     :18559               ORD    : 11589   Max.   :4983
##  (Other):36207               (Other):179260
```

A new variable was added called the total_delay, which is the sum of dep_delay and arr_delay.

The columns “year” and “cancelled” were removed from the data frame because it is all 2014 and there were no cancelled flights.

A table was made containing the number of flights by carrier, and then the carriers with less than 1000 flights in 2014 were removed.

```
flights$total_delay <- flights$dep_delay + flights$arr_delay

flights1 <- flights[, c(-1, -8)]

table(flights$carrier)
```

```
##
##    AA    AS    B6    DL    EV    F9    FL    HA    MQ    OO    UA    US    VX
## 26302   574 44479 41683 39819   473  1251   260 18559   200 46267 16750  4797
##    WN
## 11902
```

```
flights1 <- subset(flights, carrier!="AS" & carrier!="F9"& carrier!="HA"& carrier!="OO")

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
```

```
flights1 <- flights %>%
  mutate(total_delay = dep_delay + arr_delay) %>%
  select(-year, -cancelled) %>%
  group_by(carrier) %>%
  mutate(carrier_count = n()) %>%
  filter(carrier_count>1000) %>%
  ungroup()
head(flights1)
```

```
## # A tibble: 6 x 12
##   month   day dep_time dep_delay arr_time arr_delay carrier origin dest
##   <int> <int>    <int>     <int>    <int>     <int> <fct>   <fct>  <fct>
## 1     1     1      914        14     1238        13 AA      JFK    LAX
## 2     1     1     1157        -3     1523        13 AA      JFK    LAX
## 3     1     1     1902         2     2224         9 AA      JFK    LAX
## 4     1     1      722        -8     1014       -26 AA      LGA    PBI
## 5     1     1     1347         2     1706         1 AA      JFK    LAX
## 6     1     1     1824         4     2145         0 AA      EWR    LAX
## # ... with 3 more variables: distance <int>, total_delay <int>,
## #   carrier_count <int>
```

The flights was sorted by the newly created column called “total_delay” in descending order and the average of flight disctance was calculated among the top 10 flights with the longest total delay.

```
flights2 <- flights1[order(flights1$total_delay, decreasing = TRUE), ]
head(flights2)
```

```
## # A tibble: 6 x 12
##   month   day dep_time dep_delay arr_time arr_delay carrier origin dest
##   <int> <int>    <int>     <int>    <int>     <int> <fct>   <fct>  <fct>
## 1    10     4      727      1498     1008      1494 AA      EWR    DFW
## 2     4    15     1341      1241     1443      1223 AA      JFK    BOS
## 3     7    14      823      1087     1046      1090 DL      EWR    ATL
## 4     9    12      636      1056     1015      1115 AA      EWR    DFW
## 5     6    13     1046      1071     1329      1064 AA      EWR    DFW
## 6     6    16      731      1022     1057      1073 AA      EWR    DFW
## # ... with 3 more variables: distance <int>, total_delay <int>,
## #   carrier_count <int>
```

```
mean(flights2$distance[1:10])
```

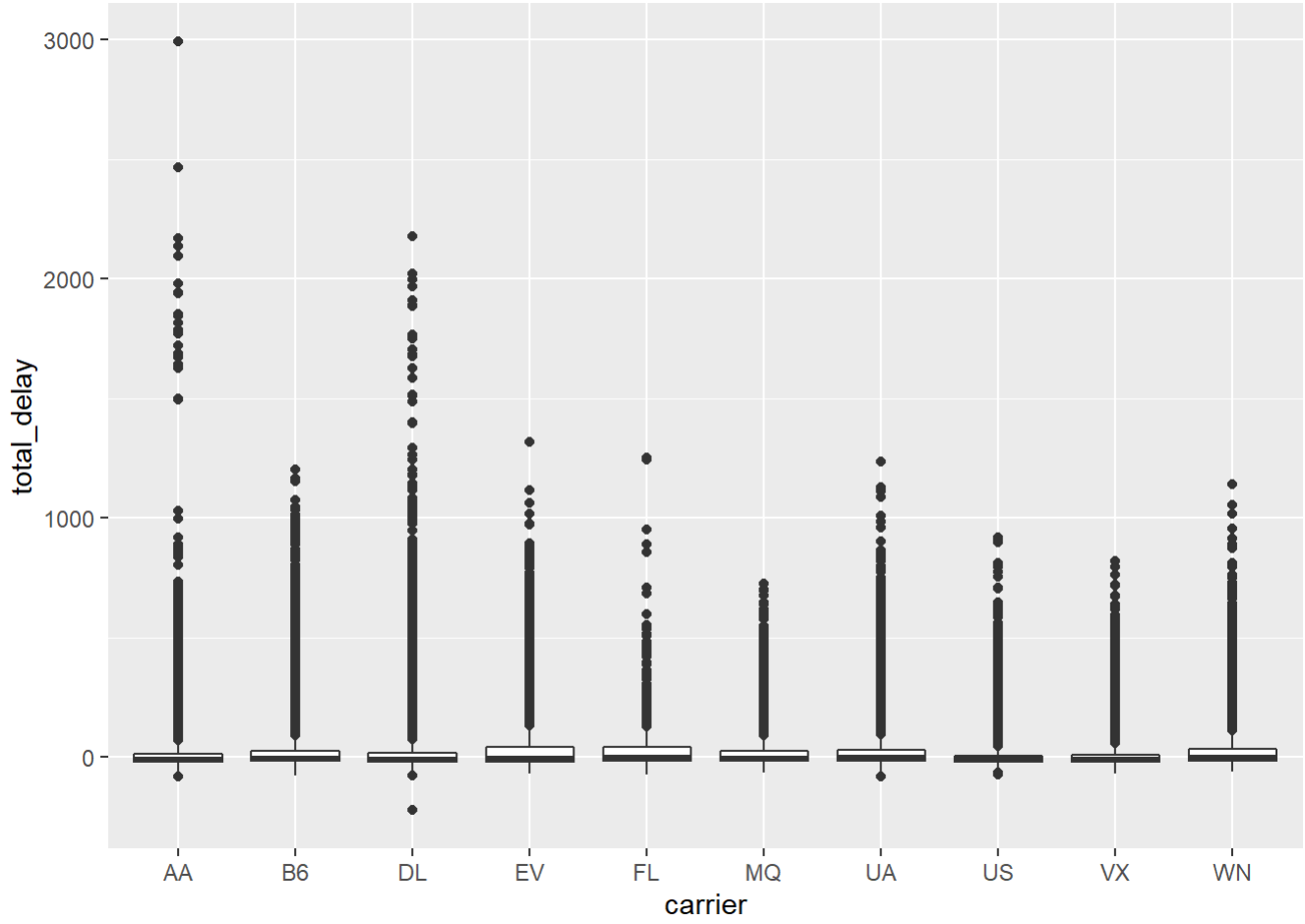
```
## [1] 1145.9
```

```
flights2 <- flights1 %>%
  arrange(desc(total_delay))

mean(flights2$distance[1:10])
```

```
## [1] 1145.9
```

The graph below shows a boxplot of the total delay by each carrier

```
library(ggplot2)

ggplot(flights2, aes(x=carrier, y=total_delay)) +
  geom_boxplot()
```

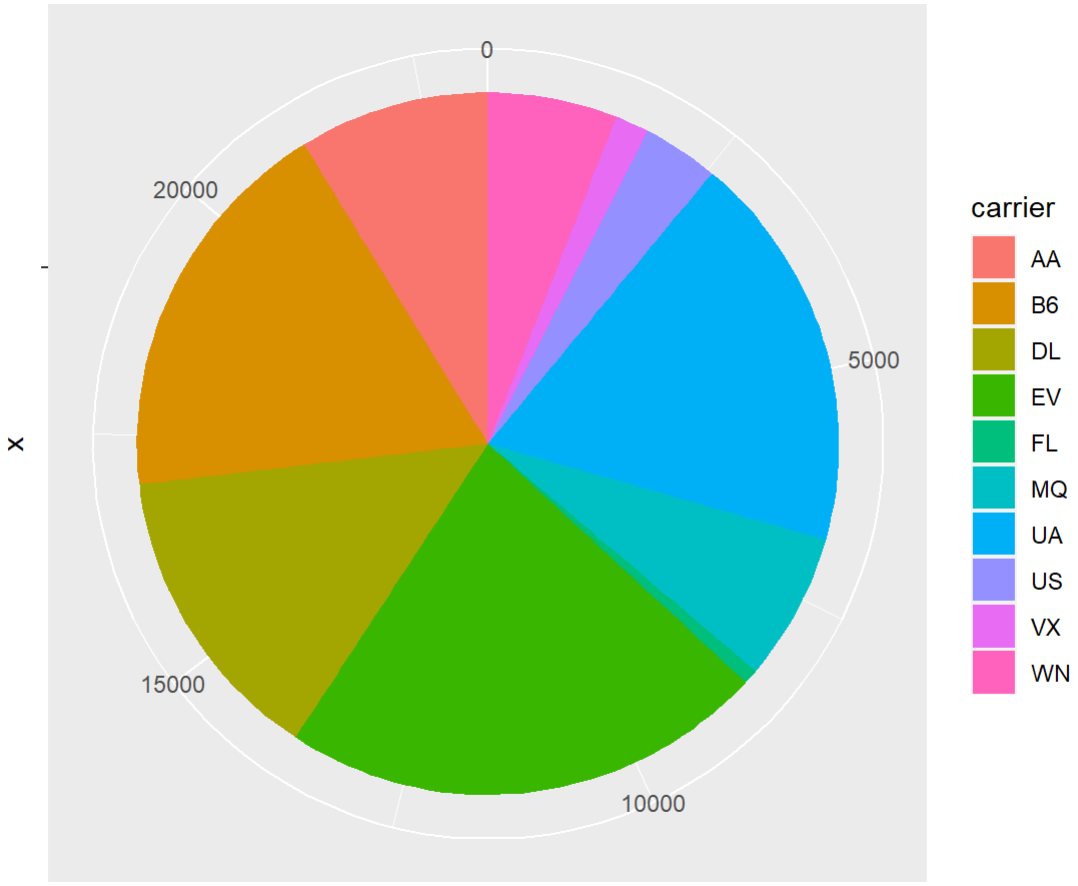


Flights with more than 60 minutes delay time are regarded as major delay. A pie chart was further plotted to show the most frequently delayed carrier in NYC.

```
major_delay <- flights2 %>%
  filter(dep_delay > 60 | arr_delay > 60)

carrier_table <- major_delay %>%
  group_by(carrier) %>%
  summarize(count = n())
```

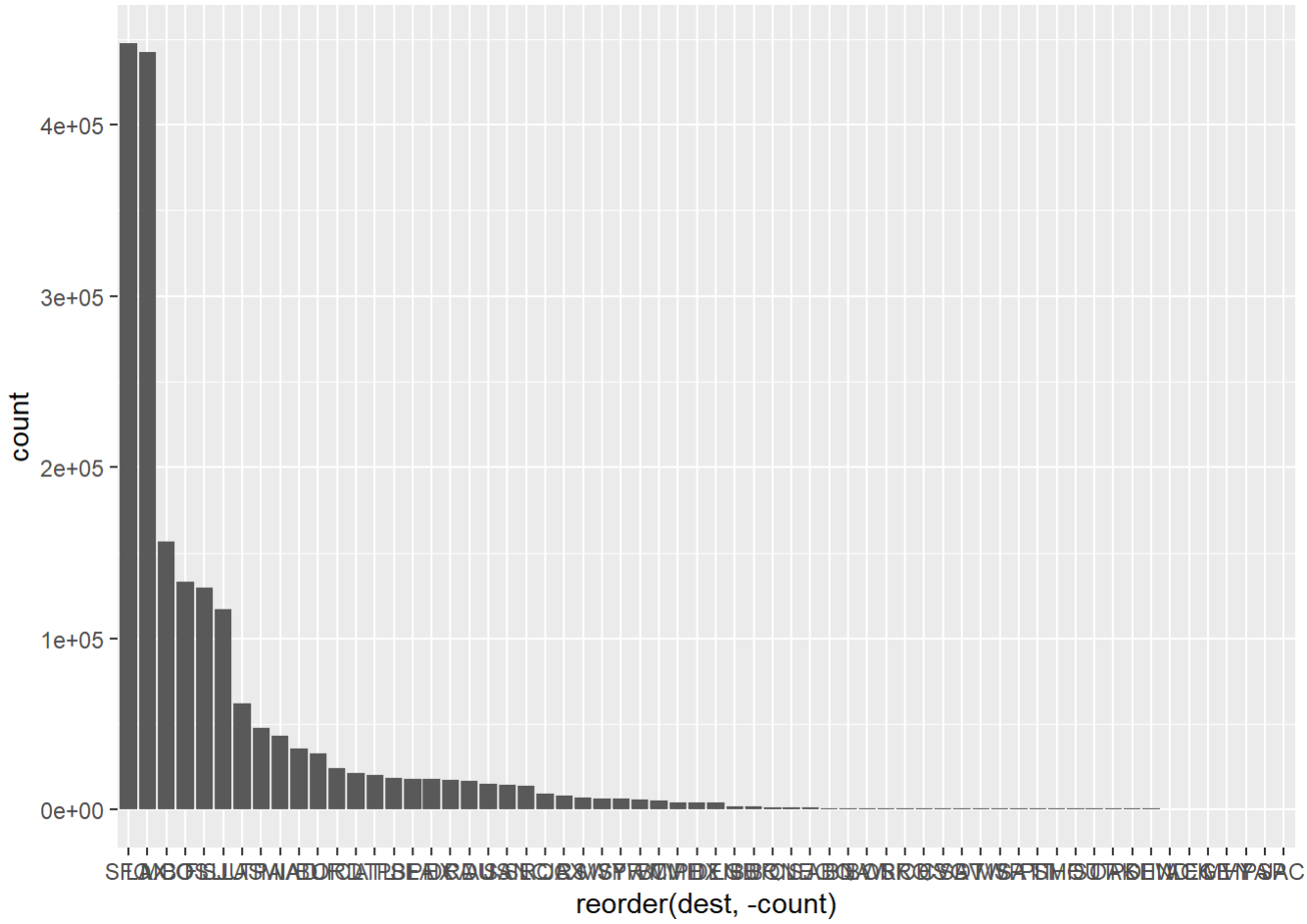
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(carrier_table, aes(x="", y=count, fill = carrier)) +
  geom_col() +
  coord_polar("y", start=0)
```



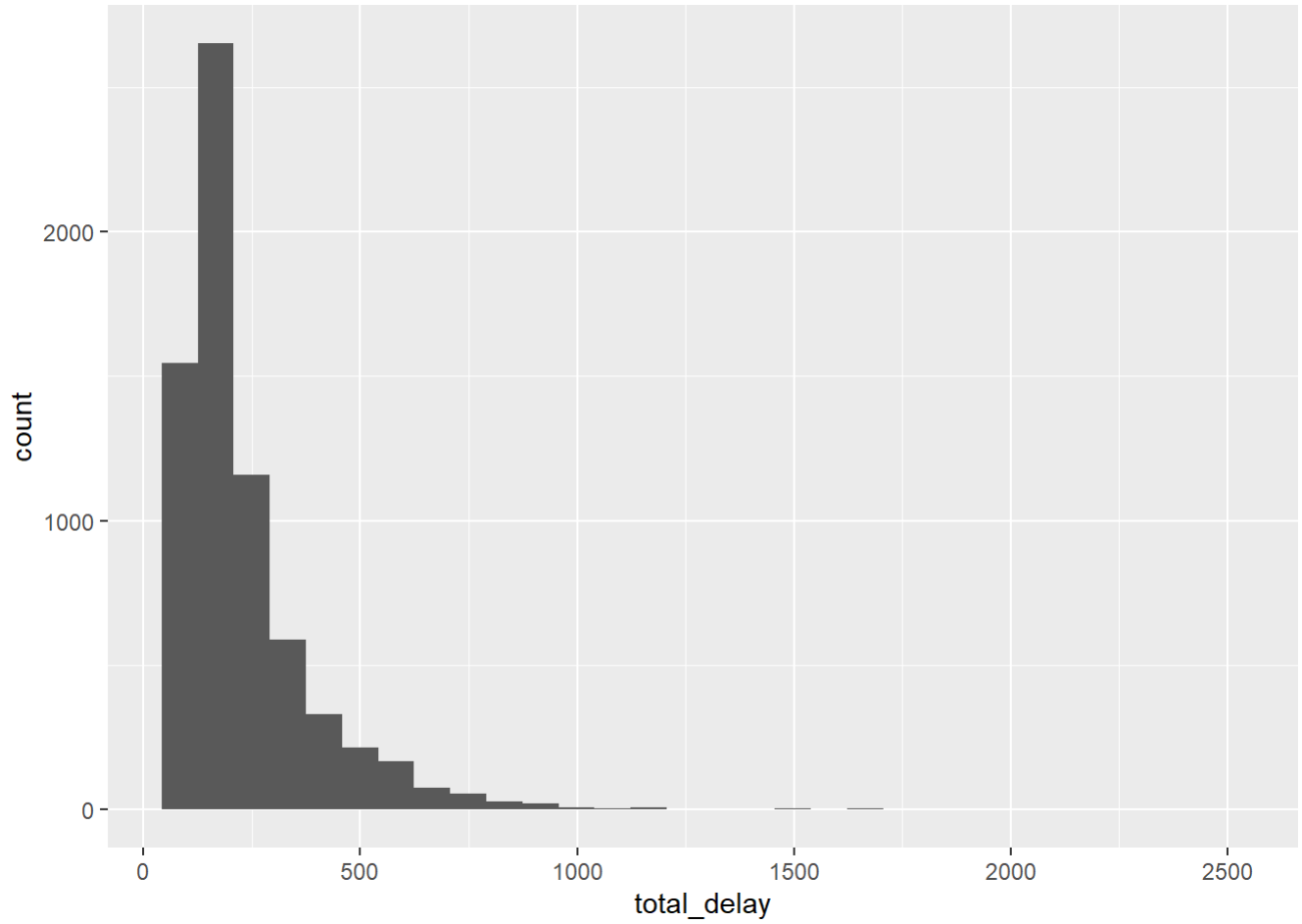
Using the major delayed flights above, the number of flight by destination is shown in decreading order that departed from JFK airport and the total delay of flights that departed from JFK airport can be show with the use of histogram

```
JFK_flights <- major_delay %>%
  filter(origin=="JFK")

dest_table <- JFK_flights %>%
  group_by(dest) %>%
  mutate(count = n()) %>%
  arrange(desc(count))

ggplot(dest_table, aes(x=reorder(dest, -count), y=count)) +
  geom_col()
```




```
ggplot(JFK_flights, aes(x=total_delay)) +
  geom_histogram()
```

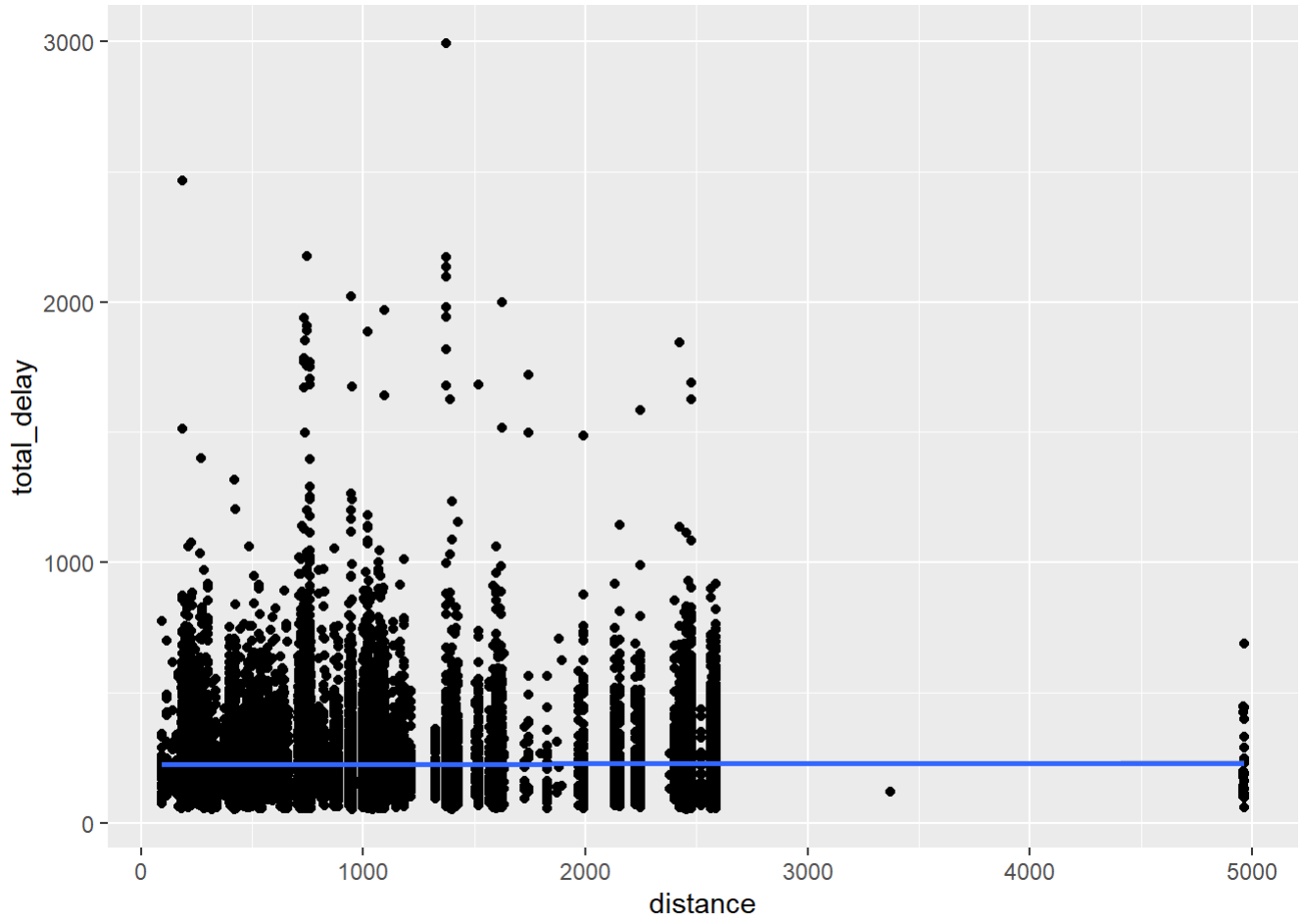
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The relationship between the total delayed time and the flight distance can be shown in the visualization below.

```
ggplot(major_delay, aes(x=distance, y=total_delay)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(major_delay$distance, major_delay$total_delay)
```

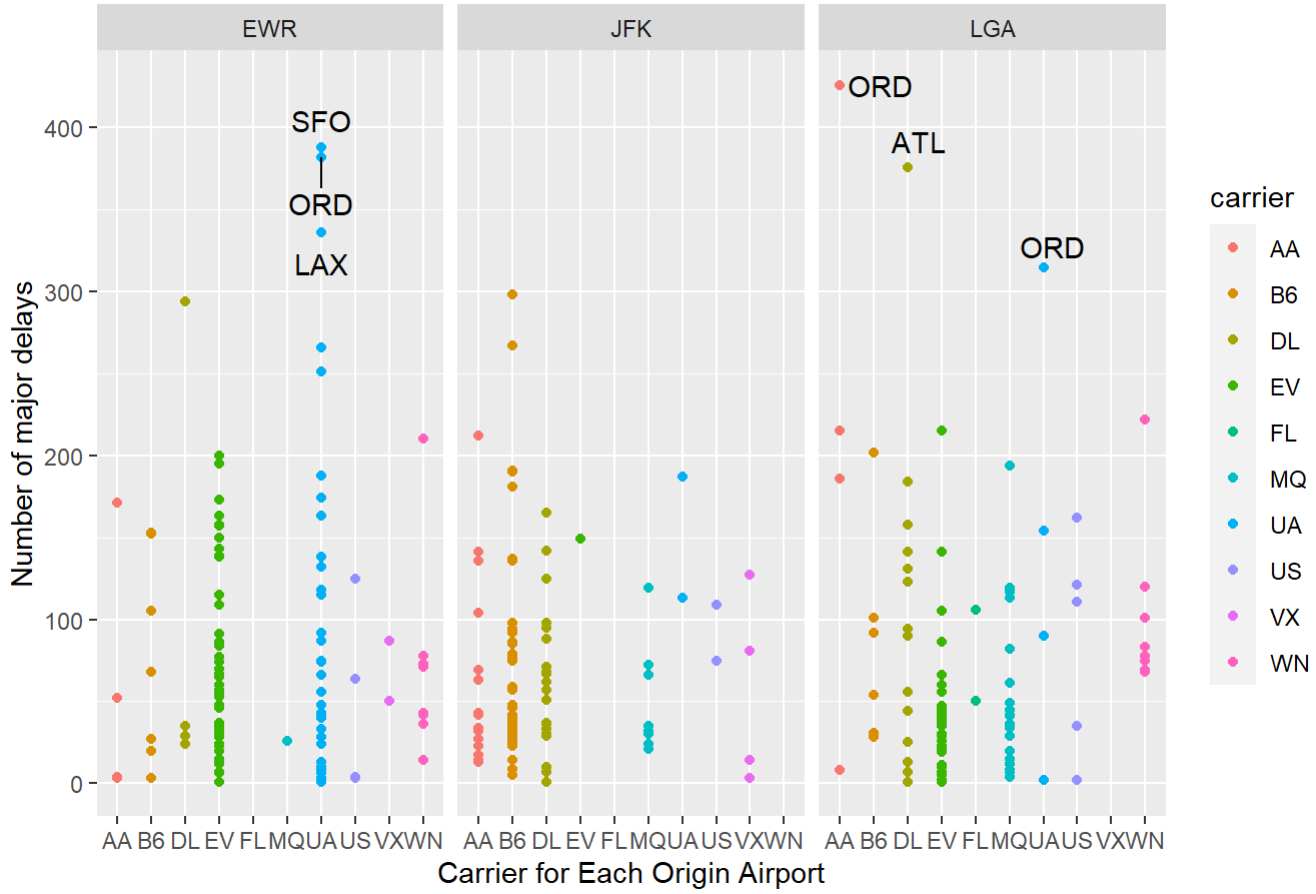
```
## [1] 0.003605182
```

```
#detach("package:plyr", unload=TRUE) # Don't use both dplyr and plyr at the same time.
library(ggrepel)
count_flights <- major_delay %>%
  group_by(origin, dest, carrier) %>%
  summarize(freq = n())
```

```
## `summarise()` regrouping output by 'origin', 'dest' (override with `.groups` argument)
```

```
ggplot(count_flights, aes(x=carrier, y=freq)) + geom_point(aes(color=carrier)) +
  geom_text_repel(aes(label=dest), force=10, data=count_flights[count_flights$freq>300, ]) +
  labs(title="Visual", x="Carrier for Each Origin Airport", y="Number of major delays") +
  facet_wrap(.~origin)
```

Visual


EWR
JFK
LGA
SFO
ORD
LAX
ORD
ATL
ORD
Number of major delays
0
100
200
300
400
AA B6 DL EV FL MQ UA US VX WN
Carrier for Each Origin Airport
carrier
AA
B6
DL
EV
FL
MQ
UA
US
VX
WN