Week 2 Lab

Arin Parsa

5/20/2021

```
library(statsr)
```

```
## Loading required package: BayesFactor

## Loading required package: coda

## Loading required package: Matrix

## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

```
library(dplyr)
```

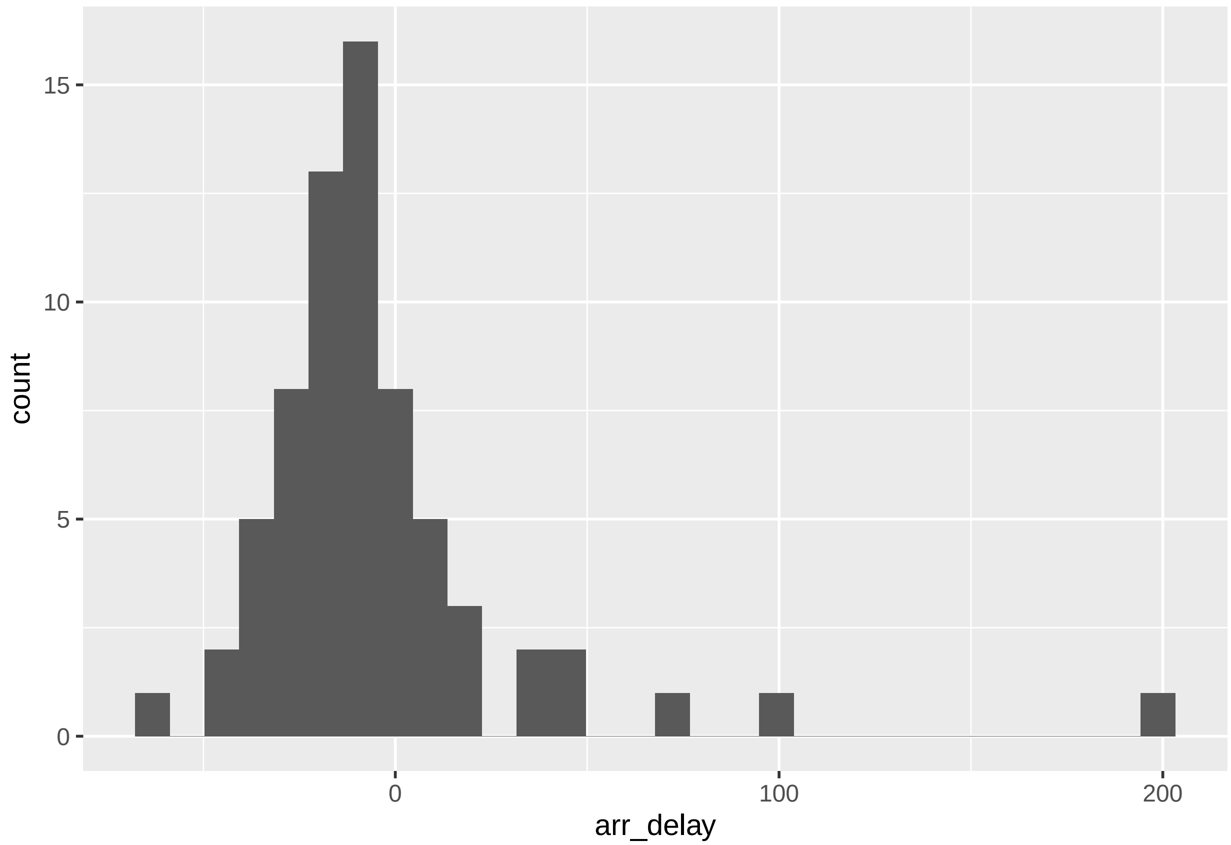
```
## Warning: package 'dplyr' was built under R version 4.0.5

##
## 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)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
data("nycflights")

#Question 1: Create a new data frame that includes flights headed to SFO in February, and save this dat
#How many flights meet these criteria?
```

```
sfo_feb_flights <- nycflights %>% filter(dest=="SFO", month==2)

#Question 2: Make a histogram and calculate appropriate summary statistics for arrival delays of sfo_fe

ggplot(sfo_feb_flights, aes(x=arr_delay)) + geom_histogram()
```

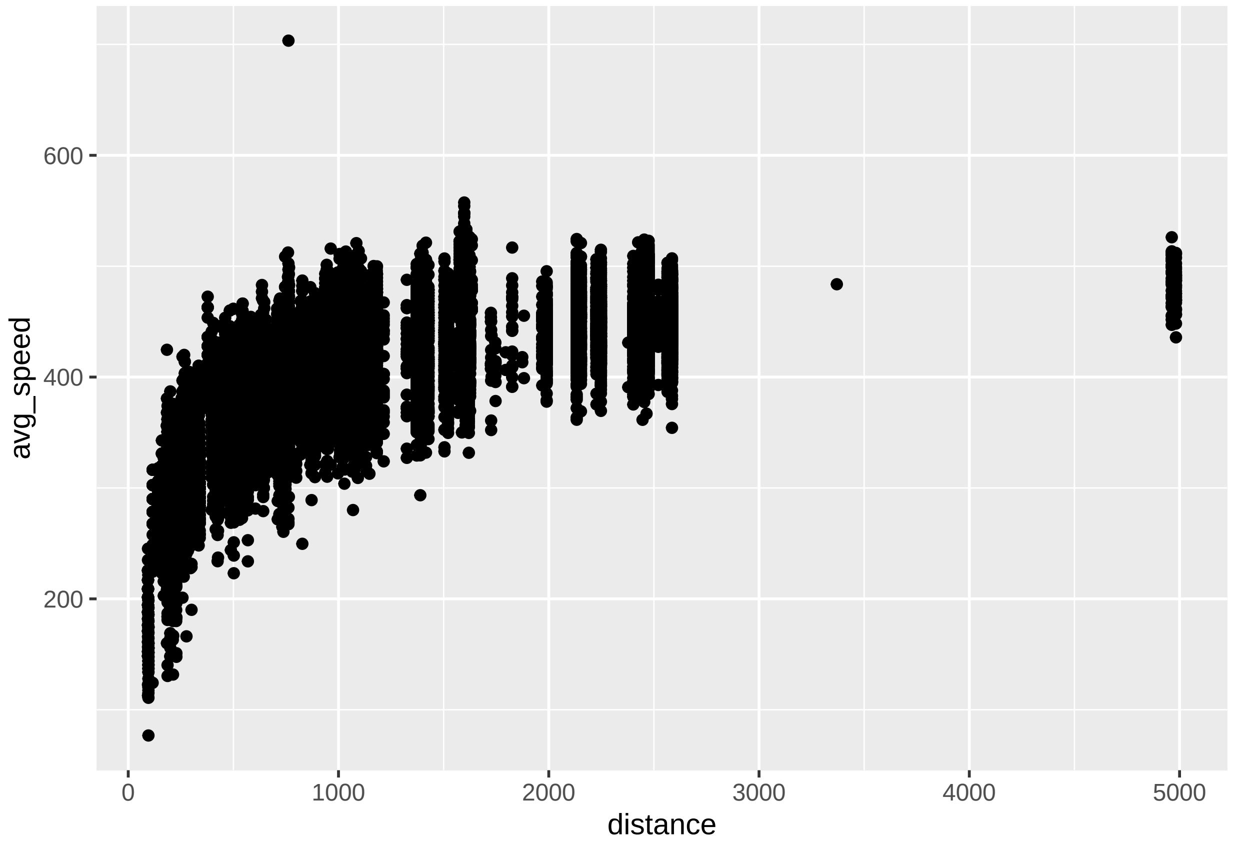


```
summary(sfo_feb_flights$arr_delay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -66.00  -21.25  -11.00   -4.50    2.00  196.00
```

```
#No more flights is delayed more than 2 hours

#Question 3: Calculate the median and interquartile range for arr_delays of flights in the sfo_feb_flig
sfo_feb_flights_delays <- sfo_feb_flights %>% group_by(carrier) %>% summarise(median_column = median(ar

#Question 4: Considering the data from all the NYC airports, which month has the highest average depart
nycflights_mean_median <- nycflights %>% group_by(month) %>% summarise(mean_column = mean(dep_delay), me

#Question 5: #Which month has the highest median departure delay from an NYC airport?
nycflights_mean_median <- nycflights_mean_median %>% arrange(desc(median_column))

#Question 7: If you were selecting an airport simply based on on time departure percentage, which NYC a
nycflights <- nycflights %>% mutate(dep_type = ifelse(dep_delay < 5, "on time", "delayed"))
```

```
nycflight_ot_percent <- nycflights %>% group_by(origin) %>% summarise(on_time_percent = sum(dep_type ==

#Question 8: What is the tail number of the plane with the fastest avg_speed?
nycflights <- nycflights %>% mutate(avg_speed = distance/(air_time/60))
nycflights_tailnum <- nycflights %>% select(avg_speed, tailnum) %>% arrange(desc(avg_speed))

#Question 9: Make a scatterplot of avg_speed vs. distance. Which of the following is true about the rel

ggplot(nycflights, aes(distance, avg_speed)) + geom_point()
```



```
#There is a overall positive association between distance and average speed

#Question 10: What fraction of flights that were "delayed" departing arrive "on time"?
nycflights <- nycflights %>% mutate(arr_type = ifelse(arr_delay <= 0, "on time", "delayed")) %>% filter
on_time_fraction <- sum(nycflights$arr_type == "on time")/nrow(nycflights)
```