Homework 2

Anishka Chauhan

2022-07-04

1)

a)

```
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
```

```
data(diamonds, package = "ggplot2")
glimpse(diamonds)
```

```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, ~
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64~
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58~
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34~
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.~
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.~
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.~
```

b)

```
library("ggplot2")
linreg = lm(price ~ carat, data = diamonds)
b = coef(linreg)

ggplot(diamonds) + aes(x = carat, y = price) + geom_point() +
  geom_abline(intercept = b[1], slope = b[2], col = "red") +
labs(x = "Weight of Diamond", y = "Price", title = "Price of Diamond vs Weight")
```

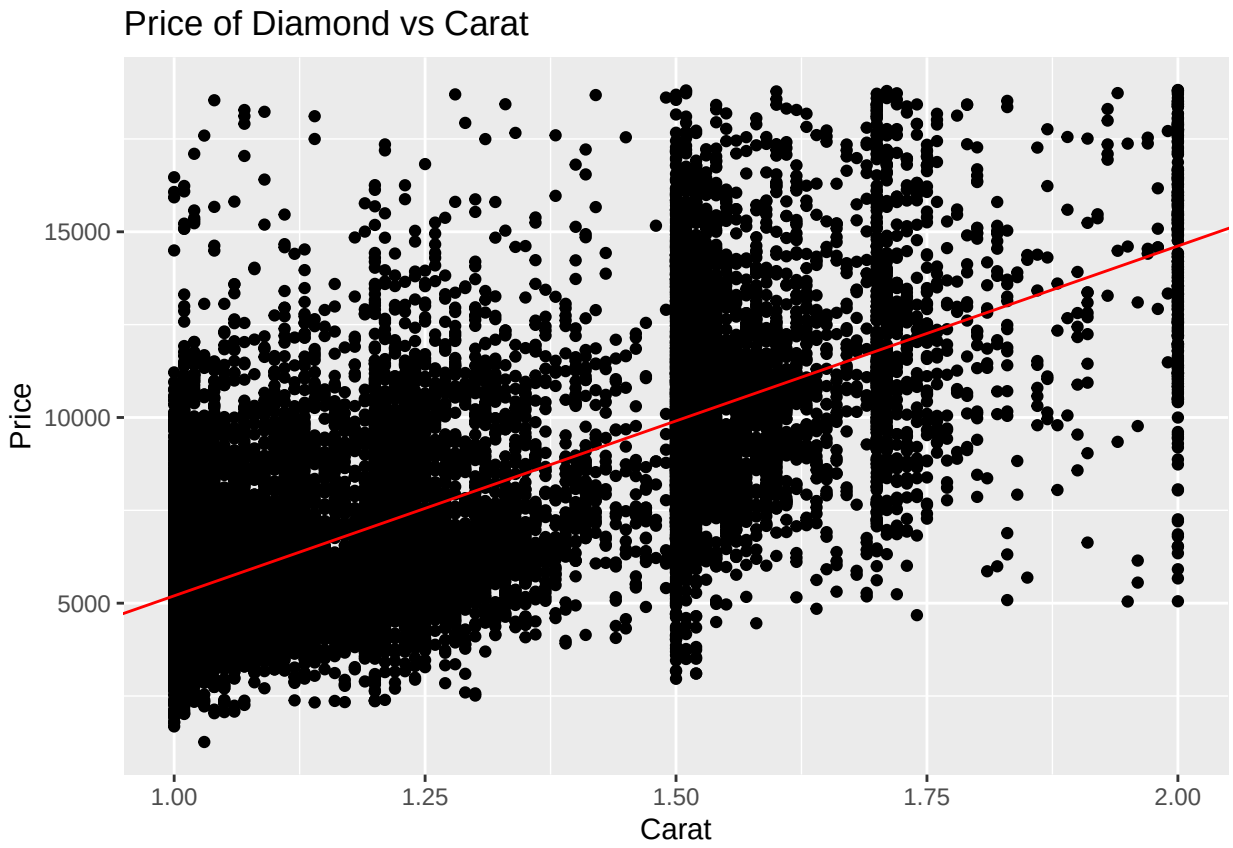
There is a positive corellation between the weight of a diamond and its price. It is also notable to mention that the same diamond weight has different prices, indicating there may be a relationship between the price of a diamond at a certain weight and time. The data is also concentrated more around diamonds with carats between 1 and 2.

c)

```
sub_carat = subset(diamonds, carat >= 1 & carat <= 2)

sub_linreg = lm(price ~ carat, data = sub_carat)
b = coef(sub_linreg)

carat_plot = ggplot(sub_carat) + aes(x = carat, y = price) + geom_point() +
  geom_abline(intercept = b[1], slope = b[2], col = "red") +
  labs(x = "Carat", y = "Price", title = "Price of Diamond vs Carat")
print(carat_plot)
```

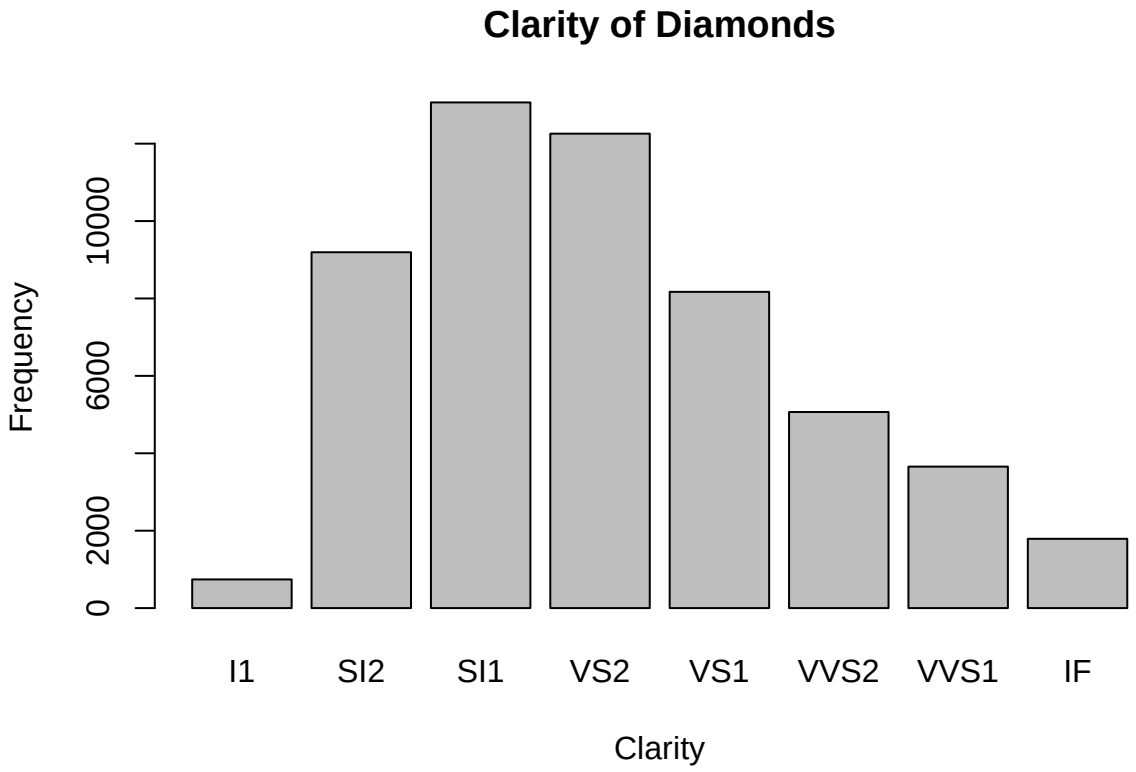


There is a slight positive corellation between the weight of the diamond and its price. There is also a greater variance in price of diamonds with weight between 1 and 1.5 carat, and there is a high concentration of diamonds with a weight between 1 and 1.5 carats.

d)

```
x = table(diamonds$clarity)

barplot(x, main = "Clarity of Diamonds", xlab = "Clarity", ylab = "Frequency")
```

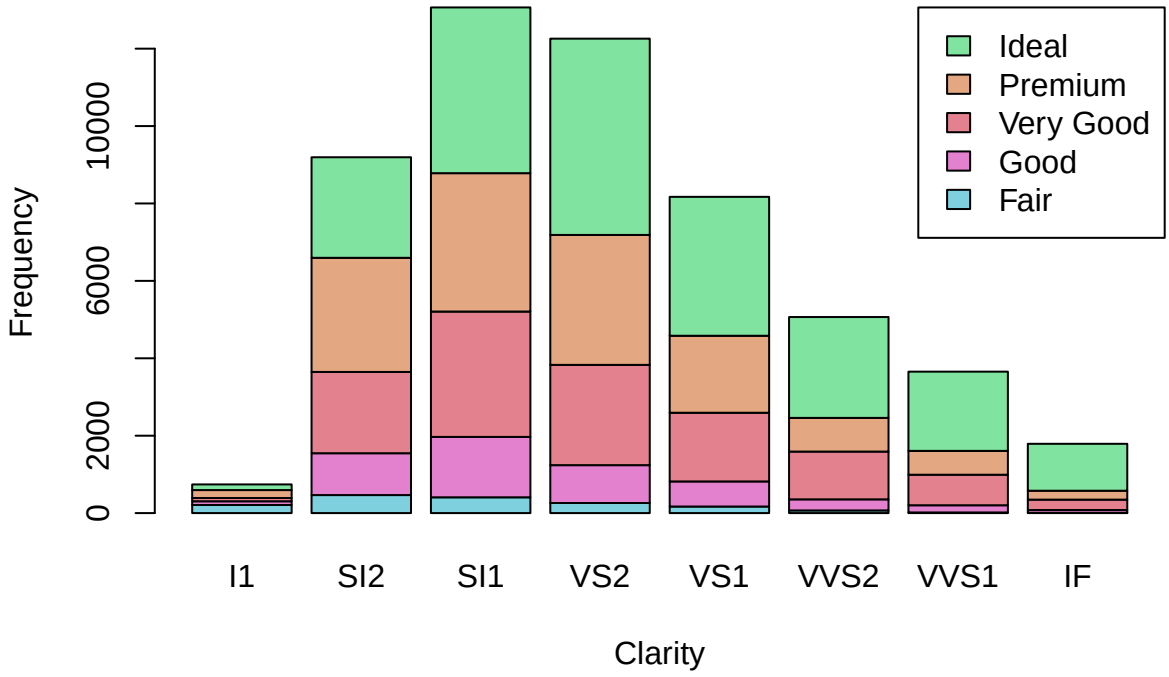


The data is skewed to the right, with most diamonds clustered around clarities of SI2 and VS1. SI1 diamonds have the highest frequency and I1 diamonds have the lowest frequency.

e)

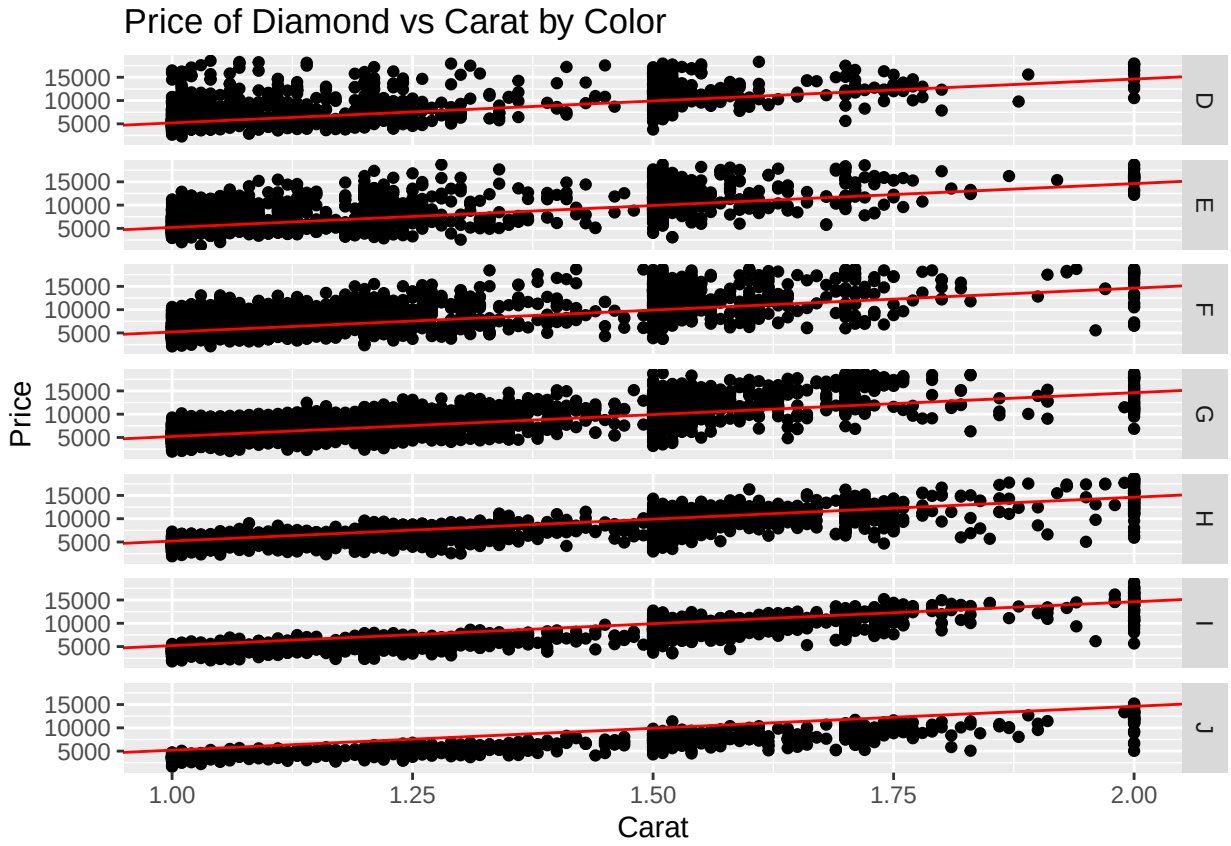
```
x = table(diamonds$cut, diamonds$clarity)
colors = c("#7ed0de", "#e381ce", "#e3818e", "#e3a781", "#81e3a0")

barplot(x, main = "Clarity of Diamonds",
        xlab = "Clarity", ylab = "Frequency",
        col = colors, legend = TRUE, args.legend = list(x = "topright"))
```



f)

```
carat_plot +  labs(x = "Carat", y = "Price",
                   title = "Price of Diamond vs Carat by Color") +
  facet_grid(color ~ .)
```



There is a positive correlation between carat and diamond prices for all diamond colors. For color D, E, F, and G, a considerable amount of diamonds with carats between 1-1.25 had prices that were higher than $10,000, whereas diamonds of the same carats with colors H, I, and J barely crossed $10,000 in price. Diamonds for colors F and G with a carat of around 1.5 had the highest variation in prices, ranging from $5,000 to $20,000. The variability of diamond prices decreases as the color changes from D to J.

2)

a) Dataset used: Police Department Incident Reports: 2018 to Present via DataSF (https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783)

b)

```
crimereports_SF = read.csv("./SFcrimereport.csv",
                           stringsAsFactors = FALSE, sep=",")
```

c)

```
glimpse(crimereports_SF)
```

```
## Rows: 608,258
## Columns: 34
## $ Incident.Datetime                                    <chr> "2021/09/29 12:59~
## $ Incident.Date                                        <chr> "2021/09/29", "20~
## $ Incident.Time                                        <chr> "12:59", "01:51",~
## $ Incident.Year                                        <int> 2021, 2021, 2021,~
## $ Incident.Day.of.Week                                 <chr> "Wednesday", "Fri~
## $ Report.Datetime                                      <chr> "2021/09/29 06:48~
## $ Row.ID                                               <dbl> 107597928150, 103~
## $ Incident.ID                                          <int> 1075979, 1030103,~
## $ Incident.Number                                      <int> 216138427, 210295~
```

```
## $ CAD.Number                                                <int> NA, 211340138, NA~
## $ Report.Type.Code                                          <chr> "II", "II", "II",~
## $ Report.Type.Description                                   <chr> "Coplogic Initial~
## $ Filed.Online                                              <chr> "true", "", "true~
## $ Incident.Code                                             <int> 28150, 26030, 710~
## $ Incident.Category                                         <chr> "Malicious Mischi~
## $ Incident.Subcategory                                      <chr> "Vandalism", "Ars~
## $ Incident.Description                                      <chr> "Malicious Mischi~
## $ Resolution                                                <chr> "Open or Active",~
## $ Intersection                                              <chr> "", "03RD ST \\ C~
## $ CNN                                                       <dbl> NA, 20240000, NA,~
## $ Police.District                                           <chr> "Ingleside", "Bay~
## $ Analysis.Neighborhood                                     <chr> "", "Bayview Hunt~
## $ Supervisor.District                                       <int> NA, 10, NA, NA, N~
## $ Latitude                                                  <dbl> NA, 37.74426, NA,~
## $ Longitude                                                 <dbl> NA, -122.3874, NA~
## $ Point                                                     <chr> "", "POINT (-122.~
## $ Neighborhoods                                             <int> NA, 56, NA, NA, N~
## $ ESNCAG...Boundary.File                                    <int> NA, NA, NA, NA, N~
## $ Central.Market.Tenderloin.Boundary.Polygon...Updated      <int> NA, NA, NA, NA, N~
## $ Civic.Center.Harm.Reduction.Project.Boundary              <int> NA, NA, NA, NA, N~
## $ HSOC.Zones.as.of.2018.06.05                               <int> NA, NA, NA, NA, N~
## $ Invest.In.Neighborhoods..IIN..Areas                       <lgl> NA, NA, NA, NA, N~
## $ Current.Supervisor.Districts                              <int> NA, 9, NA, NA, NA~
## $ Current.Police.Districts                                  <int> NA, 2, NA, NA, NA~
```

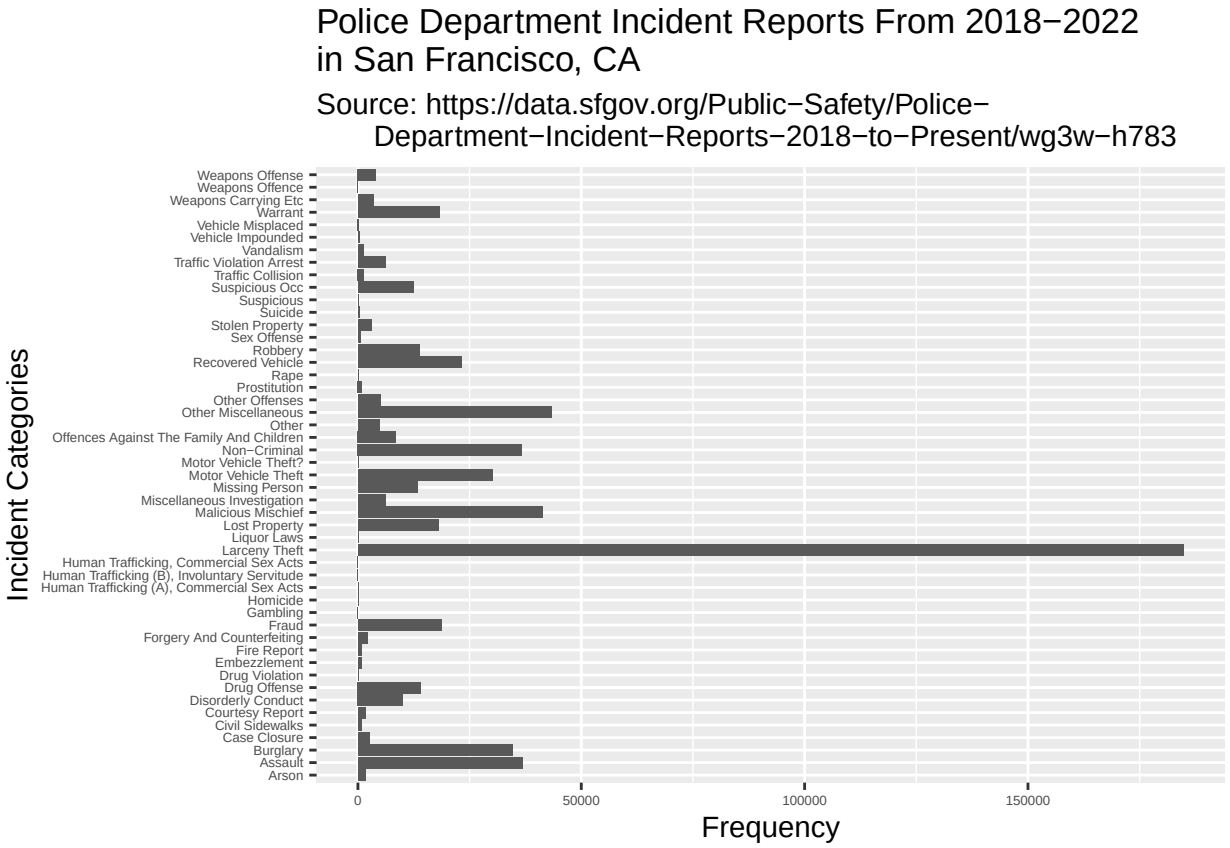
The variables of interest are the date of incidents and the incident categories.

```
require("ggplot2")

incidentCategory = as.data.frame(table(crimereports_SF$Incident.Category))
colnames(incidentCategory)[1] = "Incidents"


incidentCategory = incidentCategory[!incidentCategory$Incidents=="",]

incidentPlot = ggplot(incidentCategory, aes(x = Incidents, y = Freq)) +
  geom_bar(stat = "Identity") +
  coord_flip() +
  theme(axis.text = element_text(size = 5)) +
  labs(x = "Incident Categories", y = "Frequency",
  title = "Police Department Incident Reports From 2018-2022 \nin San Francisco, CA",
  subtitle = "Source: https://data.sfgov.org/Public-Safety/Police-
       Department-Incident-Reports-2018-to-Present/wg3w-h783")
incidentPlot
```




```
sqrrelative_freq = c()
for (item in incidentCategory$Freq)
{
  sqrrelative_freq = c(sqrrelative_freq, (item/sum(incidentCategory$Freq))^2)
}

D = 1 - sum(sqrrelative_freq)
print(c("The diversity index is: ", D))
```

```
## [1] "The diversity index is: " "0.877620706903801"
```

Plotting incident categories in a barplot reveals that larceny theft is the most reported category of police department incident reports, making it the mode of the dataset. The Simpson's diversity index for this data is 0.878, indicating there is little diversity in the category of police incident reports.

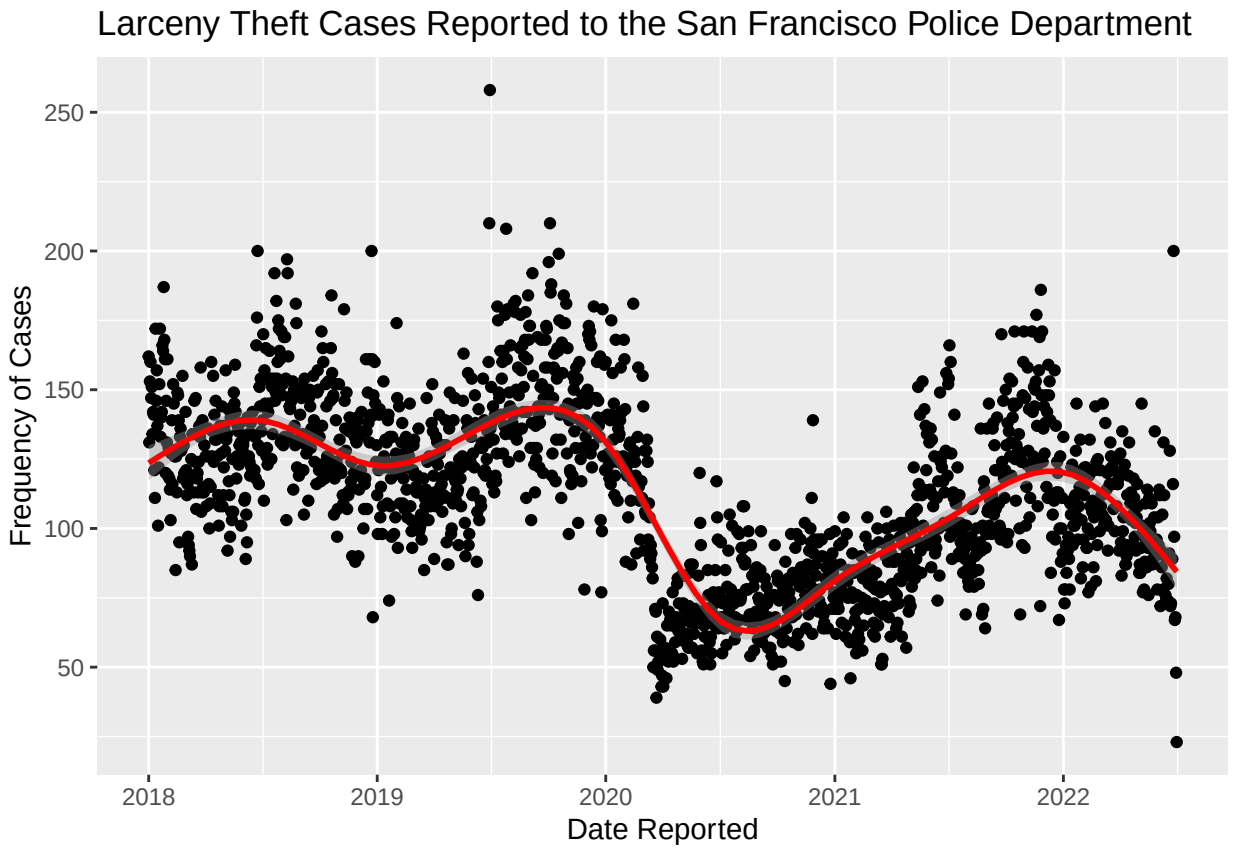
```
tempCase = crimereports_SF[!crimereports_SF$Incident.Category == "",]
tempCase = tempCase[tempCase$Incident.Category=="Larceny Theft",]
larcenyCases = as.data.frame(table(tempCase$Incident.Date))
colnames(larcenyCases)[1] = "Date"

larcenyCases$Date = as.Date(larcenyCases$Date)

larceny_plot = ggplot(larcenyCases) + aes(x = Date, y = Freq) + geom_point() +
  geom_smooth(col = "red")

larceny_plot + labs(x = "Date Reported",
  y = "Frequency of Cases",
  title = "Larceny Theft Cases Reported to the San Francisco Police Department")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
mean(larcenyCases$Freq)
```

```
## [1] 112.5332
```

Plotting the frequency of larceny theft cases and the dates they were reported indicates that larceny theft cases have overall decreased over time, with the mean number of cases reported being approximately 113 per day.

d) How effective is the San Francisco Police Department at handling cases of larceny theft?

3)

a) The plot with the lowest bandwith

b)

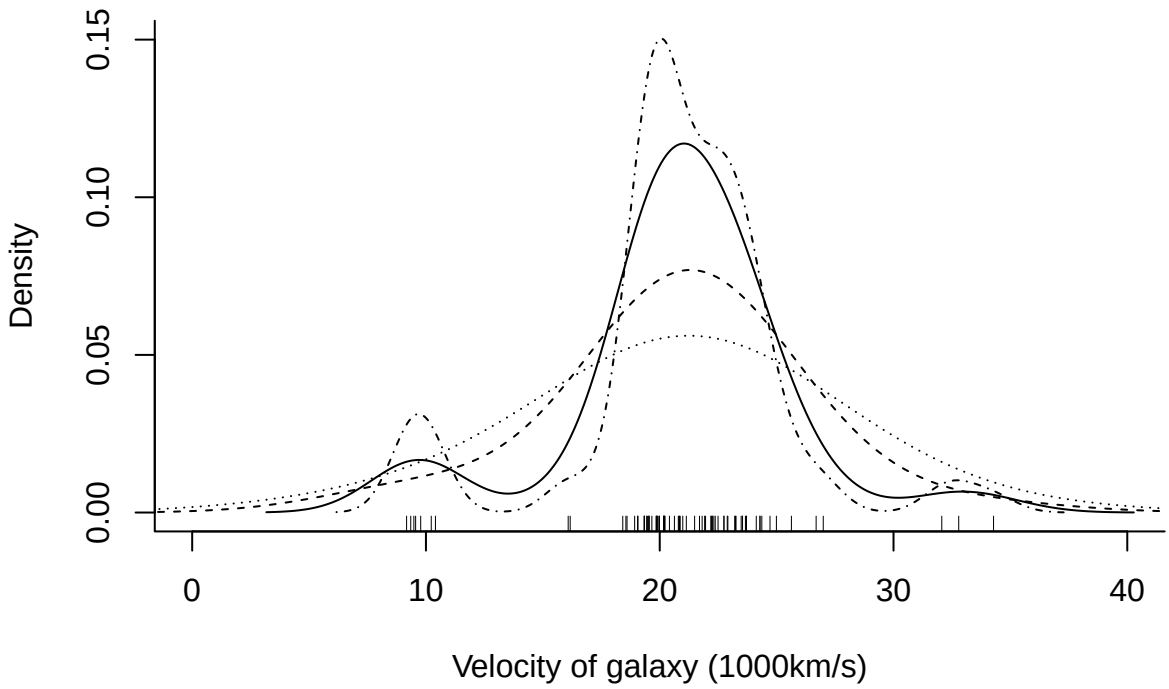
```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
gal <- galaxies/1000
plot(x = c(0, 40), y = c(0, 0.15), type = "n", bty = "l",
xlab = "Velocity of galaxy (1000km/s)", ylab = "Density")
rug(gal) # add a 'rug' (ticks along x axis)
lines(density(gal, bw = 6), lty = 3) # bw means bandwidth
```

```
lines(density(gal, bw = 4), lty = 2)
lines(density(gal, bw = 2), lty = 1) # lty controls y's appearance
lines(density(gal, bw = 1), lty = 4)
```



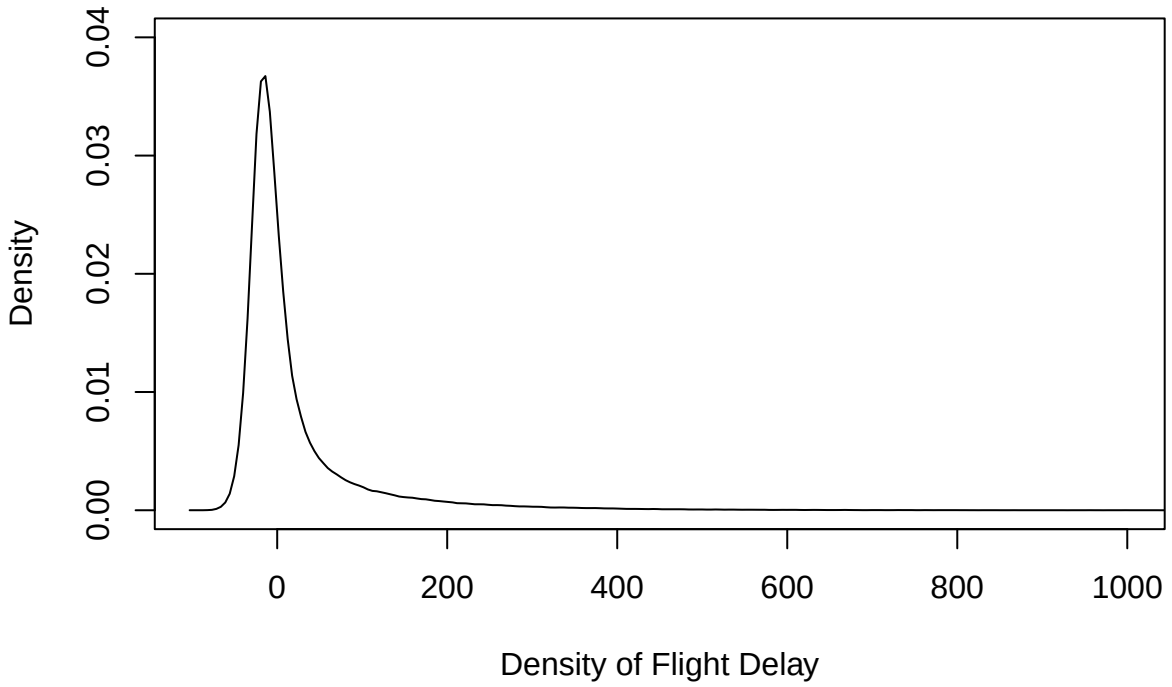
c) I would choose a bandwith of 1 because there are more local maxima in this line than the others, therefore strongly supporting the theory that multimodal data indicates the galaxy is clustered.

4)

```
library(pacman)
p_load(nycflights13, dplyr)
flights <- inner_join(flights, airlines, by="carrier")
flights <- flights[,-20] # drop extra copy of 'name'
colnames(flights)[19] <- "name" # simplify variable name
```

a)

```
flights$delay = flights$dep_delay + flights$arr_delay
delay = na.omit(flights$delay)
plot(x = c(-100, 1000), y = c(0, 0.04), type = "n",
     xlab = "Density of Flight Delay",
     ylab = "Density")
lines(density(delay, bw = 1), lty = 1)
```

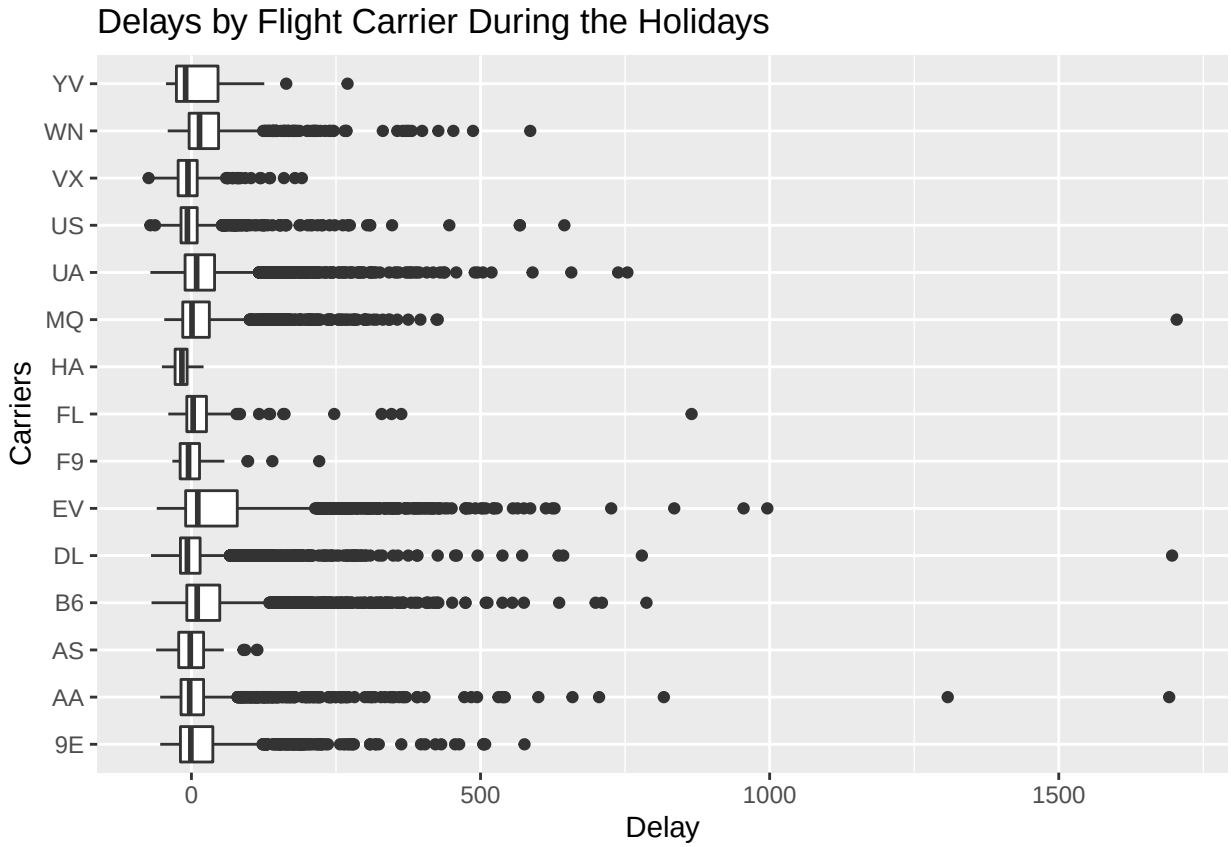


b)

```
holiday_flights = subset(flights, (month == 12 & day >= 15)| (month == 1 & day <=5))
holiday_flights$delay = holiday_flights$dep_delay + holiday_flights$arr_delay

ggplot(holiday_flights, aes(x = carrier, y = delay)) + geom_boxplot() +
  coord_flip() + labs(x = "Carriers",
  y = "Delay",
  title = "Delays by Flight Carrier During the Holidays")
```

```
## Warning: Removed 385 rows containing non-finite values (stat_boxplot).
```



As seen by the median delay of each airline, Hawaiian Airlines (HA) has the least median delay during the holiday months.

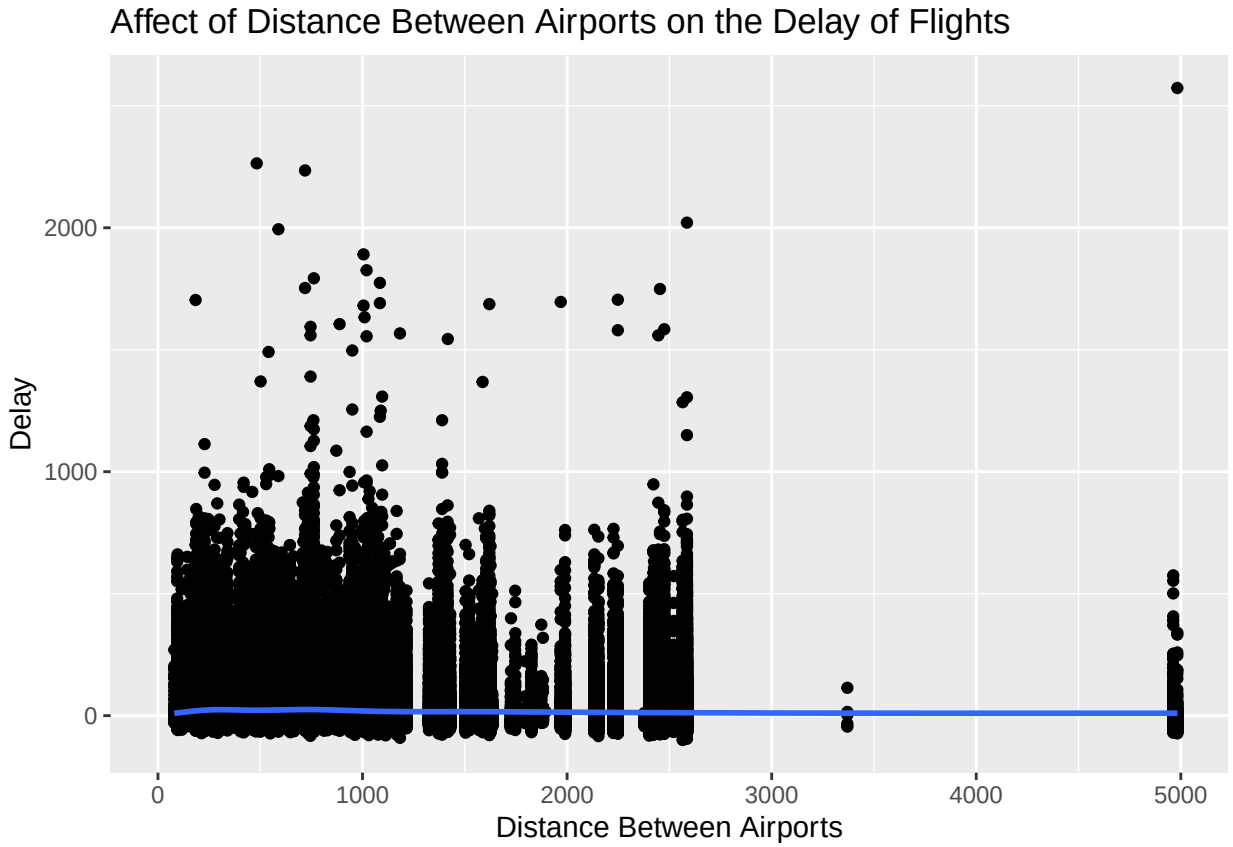
c)

```
ggplot(flights, aes(x = distance, y = delay)) + geom_point() +
  geom_smooth() + labs(x = "Distance Between Airports", y = "Delay",
  title = "Affect of Distance Between Airports on the Delay of Flights")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 9430 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9430 rows containing missing values (geom_point).
```

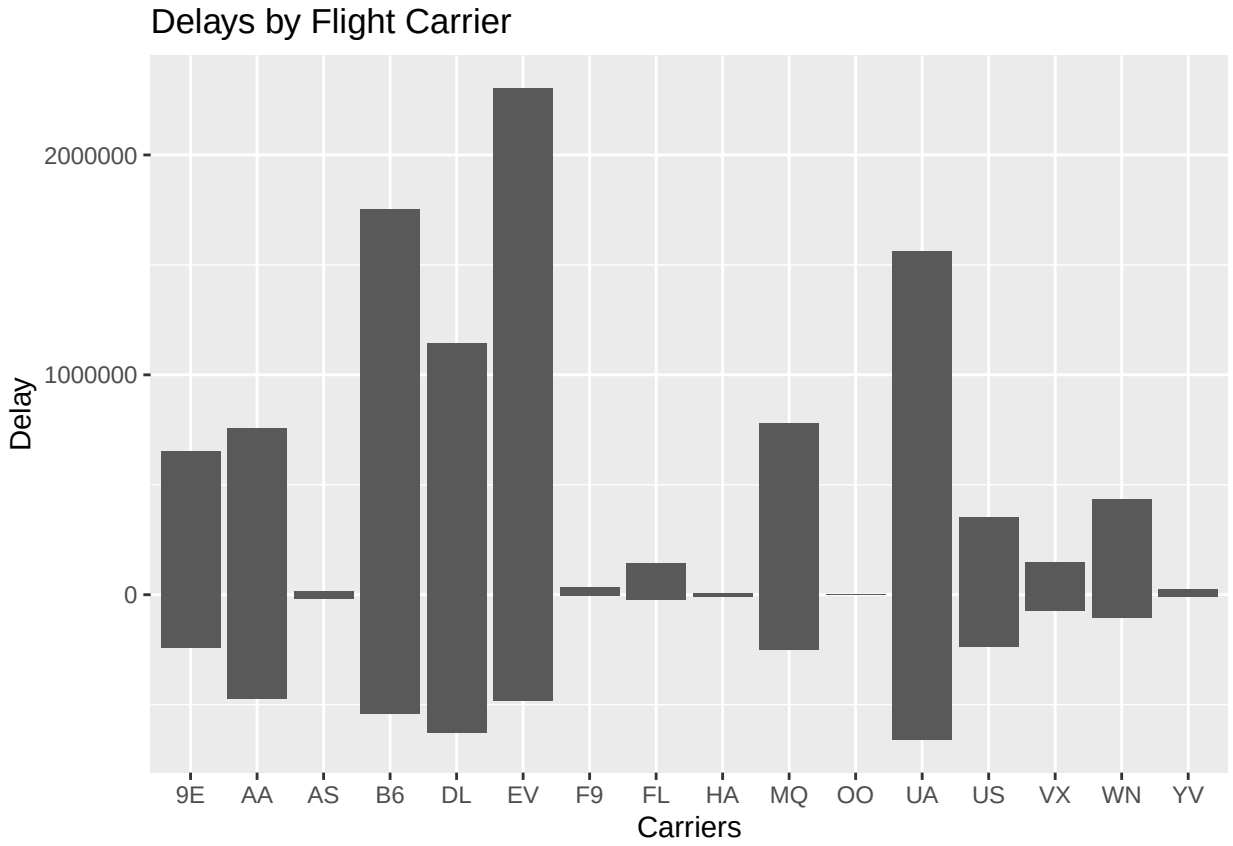


There seems to be no effect on flight delay by the distance between airports.

d)

```
options("scipen"= 100, "digits" = 4)
ggplot(flights, aes(x = carrier, y = delay)) + geom_col() +
  labs(x = "Carriers",
       y = "Delay",
       title = "Delays by Flight Carrier")
```

```
## Warning: Removed 9430 rows containing missing values (position_stack).
```



Flights from SkyWest airlines (OO) are the most likely to be on time. The barchart above shows how the bar for SkyWest is the shortest, indicating majority of the flights from SkyWest have the least delays