```
#=============================================================================
#              Using R: Kaggle.com (San Francisco Crime Classification)
#=============================================================================
#Title:   San Francisco Crime Classification
#Url: https://www.kaggle.com/c/sf-crime
#From 1934 to 1963, San Francisco was infamous for housing some of the world's most
notorious criminals on the
#inescapable island of Alcatraz. Today, the city is known more for its tech scene than
its criminal past. But,
#with rising wealth inequality, housing shortages, and a proliferation of expensive
digital toys riding BART to
#work, there is no scarcity of crime in the city by the bay. ...
#This dataset contains incidents derived from SFPD Crime Incident Reporting system. It
contains the following fields:
#The data ranges from 1/1/2003 to 5/13/2015.

#Dates      - timestamp of the crime incident
#Category   - category of the crime incident (only in train.csv). This is the target
variable you are going to predict.
#Descript   - detailed description of the crime incident (only in train.csv)
#DayOfWeek  - the day of the week
#PdDistrict - name of the Police Department District
#Resolution - how the crime incident was resolved (only in train.csv)
#Address    - the approximate street address of the crime incident
#X          - Longitude
#Y          - Latitude

#-----------------------------------------------------------------------------
#(b) Perform an initial basic exploratory analysis of the data which includes at
minimum: the number of
#rows, number of variables, descriptive statistics, a selection of visualizations,
information on missing
#value counts, and some form of outlier labeling/detection.
SFCC.Data <- read.csv(file="train.csv", head=TRUE, sep=",")
length(SFCC.Data$Dates)
```

```
## [1] 878049
```

```
#There are 878049 observation from 2003 to 2015. The data as presented above contains
9 variables. By a quick
#observation of the data, it can be seen that the data is pretty clean. The dates are
in the same format. There
#is no missing in the dates, category, etc. There is only missing information about
how the crime resolved
#which is presented by "NONE". Actually the number of the missing seems high.
SFCC.Data[SFCC.Data$Resolution=="NONE","Resolution"] <- NA
str(SFCC.Data)
```

```
## 'data.frame':    878049 obs. of  9 variables:
##  $ Dates     : Factor w/ 389257 levels "2003-01-06 00:01:00",..: 389257 389257
389256 389255 389255 389255 389255 389255 389254 389254 ...
##  $ Category  : Factor w/ 39 levels "ARSON","ASSAULT",..: 38 22 22 17 17 17 37 37 17
17 ...
##  $ Descript  : Factor w/ 879 levels "ABANDONMENT OF CHILD",..: 867 811 811 405 405
407 740 740 405 405 ...
##  $ DayOfWeek : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ PdDistrict: Factor w/ 10 levels "BAYVIEW","CENTRAL",..: 5 5 5 5 6 3 3 1 7 2 ...
##  $ Resolution: Factor w/ 17 levels "ARREST, BOOKED",..: 1 1 1 NA NA NA NA NA NA NA
...
```

```
##  $ Address   : Factor w/ 23228 levels "0 Block of  HARRISON ST",..: 19791 19791
22698 4267 1844 1506 13323 18055 11385 17659 ...
##  $ X         : num  -122 -122 -122 -122 -122 ...
##  $ Y         : num  37.8 37.8 37.8 37.8 37.8 ...
```

```
mean(is.na(SFCC.Data$Resolution))
```

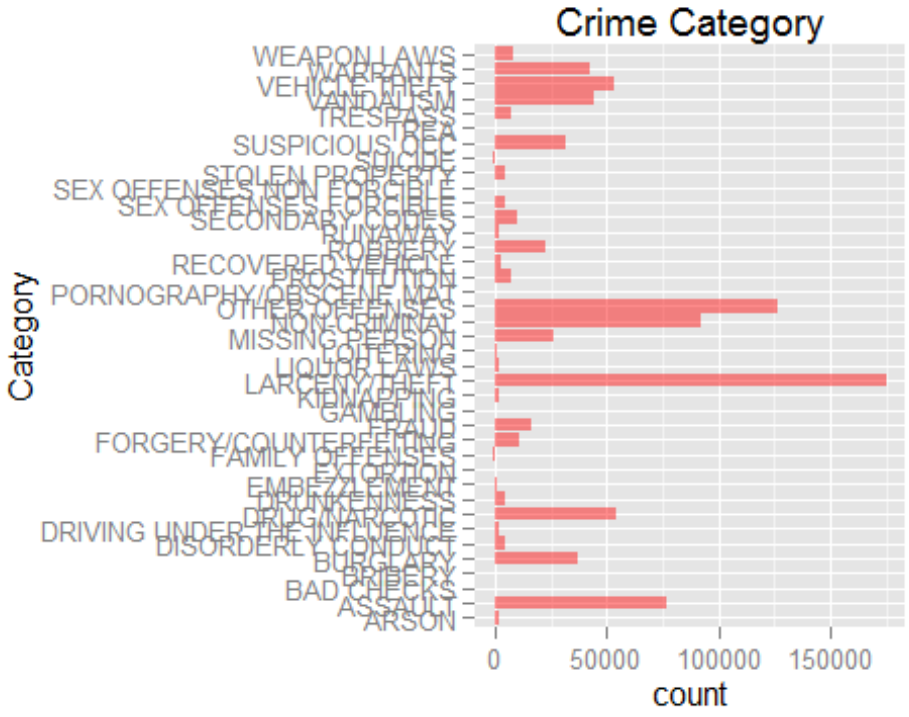
```
## [1] 0.5999551
```

```
#It can be observed that almost 60% of the data from the resolution is missing. So the
number of observation
#contributes here is almost small and make the imputation so difficult or almost
impossible.
aggregate(SFCC.Data, by=list(SFCC.Data$Category), function(x) mean(is.na(x)))
```

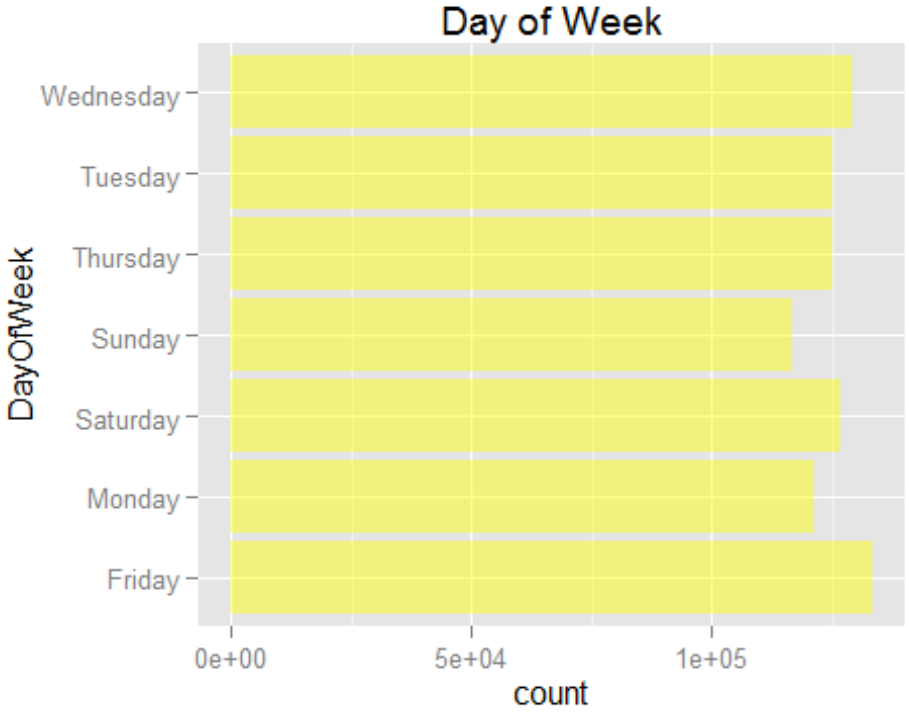
```
##                       Group.1 Dates Category Descript DayOfWeek
## 1                       ARSON     0        0        0         0
## 2                     ASSAULT     0        0        0         0
## 3                   BAD CHECKS     0        0        0         0
## 4                     BRIBERY     0        0        0         0
## 5                    BURGLARY     0        0        0         0
## 6           DISORDERLY CONDUCT     0        0        0         0
## 7  DRIVING UNDER THE INFLUENCE     0        0        0         0
## 8                DRUG/NARCOTIC     0        0        0         0
## 9                  DRUNKENNESS     0        0        0         0
## 10                EMBEZZLEMENT     0        0        0         0
## 11                   EXTORTION     0        0        0         0
## 12             FAMILY OFFENSES     0        0        0         0
## 13      FORGERY/COUNTERFEITING     0        0        0         0
## 14                       FRAUD     0        0        0         0
## 15                   GAMBLING     0        0        0         0
## 16                  KIDNAPPING     0        0        0         0
## 17               LARCENY/THEFT     0        0        0         0
## 18                 LIQUOR LAWS     0        0        0         0
## 19                   LOITERING     0        0        0         0
## 20              MISSING PERSON     0        0        0         0
## 21                NON-CRIMINAL     0        0        0         0
## 22              OTHER OFFENSES     0        0        0         0
## 23     PORNOGRAPHY/OBSCENE MAT     0        0        0         0
## 24                PROSTITUTION     0        0        0         0
## 25           RECOVERED VEHICLE     0        0        0         0
## 26                     ROBBERY     0        0        0         0
## 27                     RUNAWAY     0        0        0         0
## 28             SECONDARY CODES     0        0        0         0
## 29       SEX OFFENSES FORCIBLE     0        0        0         0
## 30   SEX OFFENSES NON FORCIBLE     0        0        0         0
## 31             STOLEN PROPERTY     0        0        0         0
## 32                     SUICIDE     0        0        0         0
## 33              SUSPICIOUS OCC     0        0        0         0
## 34                        TREA     0        0        0         0
## 35                   TRESPASS     0        0        0         0
## 36                  VANDALISM     0        0        0         0
## 37               VEHICLE THEFT     0        0        0         0
## 38                    WARRANTS     0        0        0         0
## 39                WEAPON LAWS     0        0        0         0
##    PdDistrict Resolution Address X Y
## 1           0 0.82485129       0 0 0
## 2           0 0.58185910       0 0 0
## 3           0 0.79064039       0 0 0
## 4           0 0.37370242       0 0 0
```

```
## 5           0 0.83936879       0 0 0
## 6           0 0.31550926       0 0 0
## 7           0 0.05687831       0 0 0
## 8           0 0.08587945       0 0 0
## 9           0 0.16308411       0 0 0
## 10          0 0.71183533       0 0 0
## 11          0 0.74218750       0 0 0
## 12          0 0.57841141       0 0 0
## 13          0 0.62739184       0 0 0
## 14          0 0.75412195       0 0 0
## 15          0 0.35616438       0 0 0
## 16          0 0.46945750       0 0 0
## 17          0 0.89528874       0 0 0
## 18          0 0.10509721       0 0 0
## 19          0 0.10775510       0 0 0
## 20          0 0.36088345       0 0 0
## 21          0 0.74235136       0 0 0
## 22          0 0.26331014       0 0 0
## 23          0 0.45454545       0 0 0
## 24          0 0.03968466       0 0 0
## 25          0 0.93275972       0 0 0
## 26          0 0.76643478       0 0 0
## 27          0 0.33761562       0 0 0
## 28          0 0.54381572       0 0 0
## 29          0 0.52324521       0 0 0
## 30          0 0.46621622       0 0 0
## 31          0 0.12444934       0 0 0
## 32          0 0.72047244       0 0 0
## 33          0 0.87674285       0 0 0
## 34          0 0.50000000       0 0 0
## 35          0 0.29565930       0 0 0
## 36          0 0.87707099       0 0 0
## 37          0 0.91565795       0 0 0
## 38          0 0.05197328       0 0 0
## 39          0 0.27481005       0 0 0
```

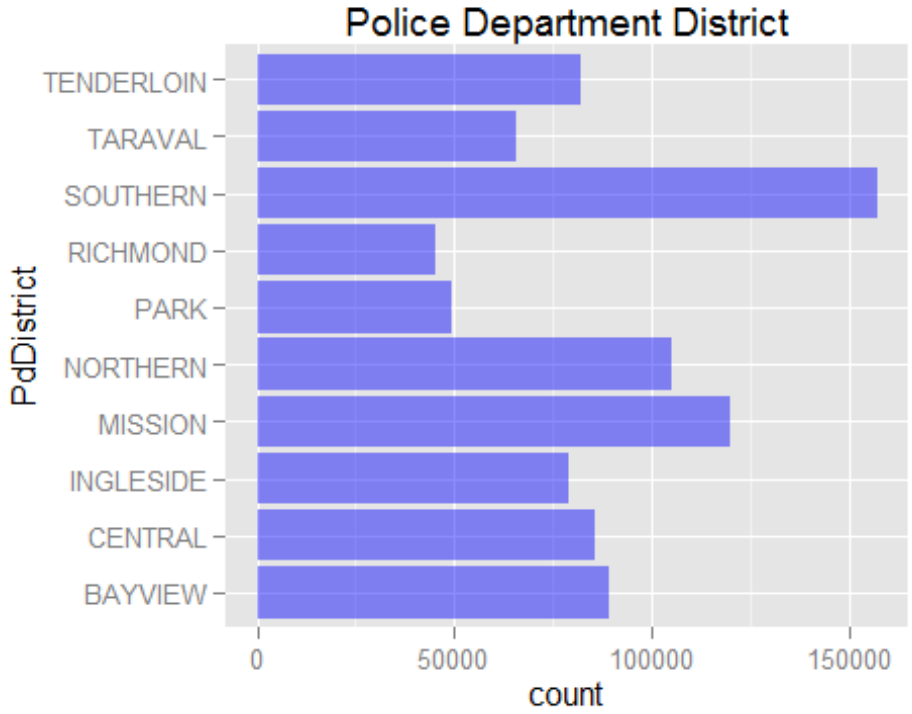
```
#There are 39 crime category and the highest percentage of resolution missing
associates with RECOVERED VEHICLE,
#VEHICLE THEFT and LARCENY/THEFT. By quick look at the data some of them are highly
correlated for example "DayOfWeek"
#can be extracted by dates or the "Address" can be obtained from the Latitude and
Longitude. So it seems they do not
#give us much information and seems redundant.
#From hereforward, for visualization purpose we eliminate Address.
SFCC.Data[,"Address"] <- NULL
SFCC.Data$Dates <- as.Date(strtrim(as.character(SFCC.Data$Dates), 10))

#Here we can limit our data to specific time. However, we use all of them.
Recent.SFCC.Data <- SFCC.Data[SFCC.Data$Dates>as.Date("2003-01-01"),]


library(ggplot2)
qplot(data=Recent.SFCC.Data, Category, geom="histogram", binwidth=1.5,
      main= "Crime Category", fill=I("red"), alpha=I(0.45))+ coord_flip()
```



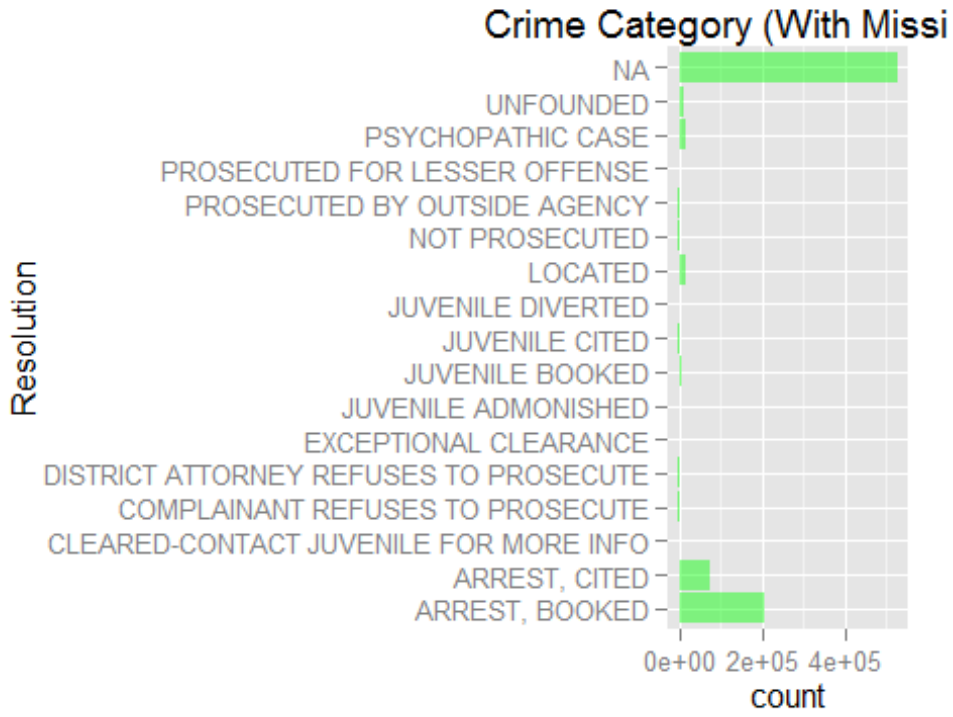

```
qplot(data=Recent.SFCC.Data, DayOfWeek, geom="histogram", binwidth=1.5,
      main= "Day of Week", fill=I("yellow"), alpha=I(0.45))+ coord_flip()
```



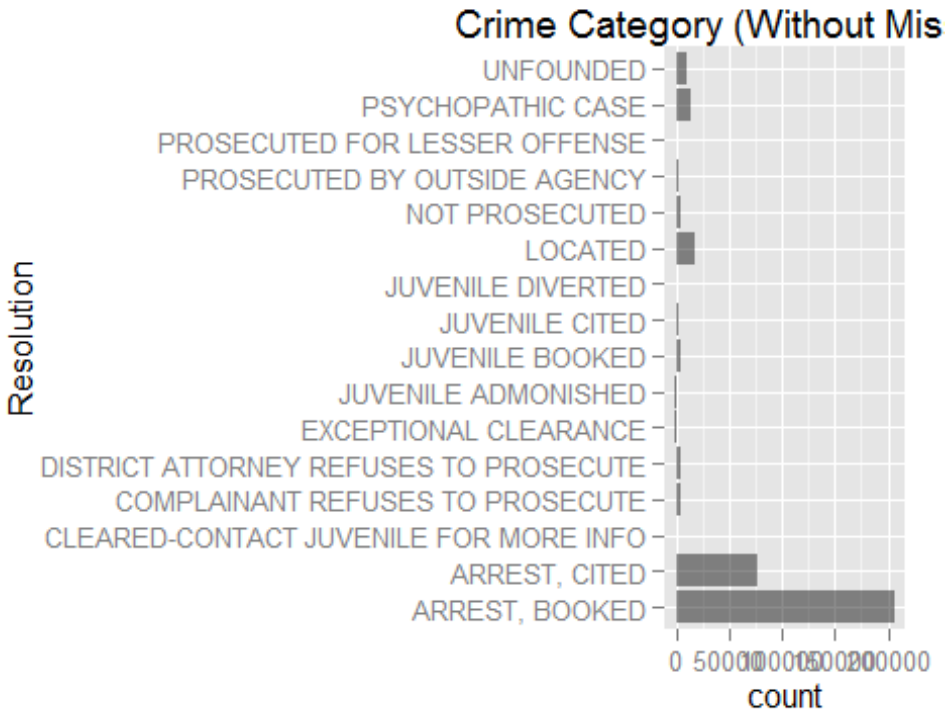

```
qplot(data=Recent.SFCC.Data, PdDistrict, geom="histogram", binwidth=1.5,
      main= "Police Department District", fill=I("blue"), alpha=I(0.45))+ coord_flip()
```



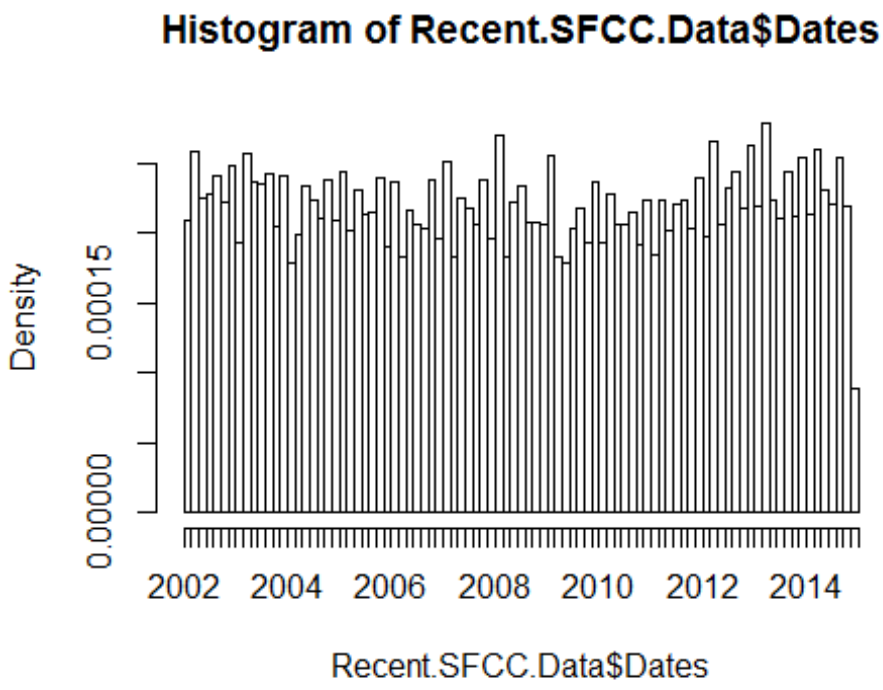

```
qplot(data=Recent.SFCC.Data, Resolution, geom="histogram", binwidth=1.5,
      main= "Crime Category (With Missing Data)", fill=I("green"), alpha=I(0.45))+
coord_flip()
```



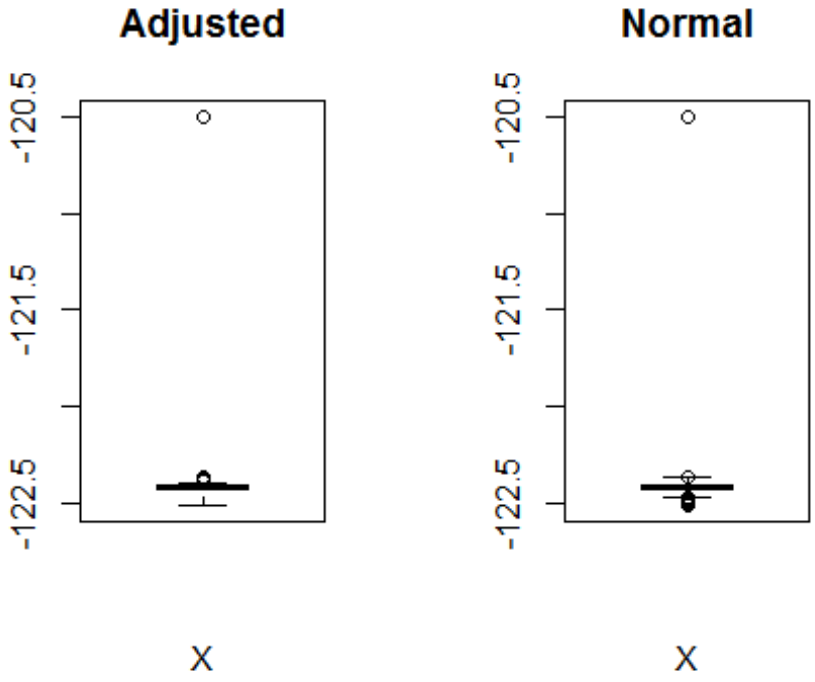

```
Recent2 <- na.omit(Recent.SFCC.Data)
qplot(data=Recent2, Resolution, geom="histogram", binwidth=1.5,
      main= "Crime Category (Without Missing Data)", fill=I("black"), alpha=I(0.45))+
coord_flip()
```



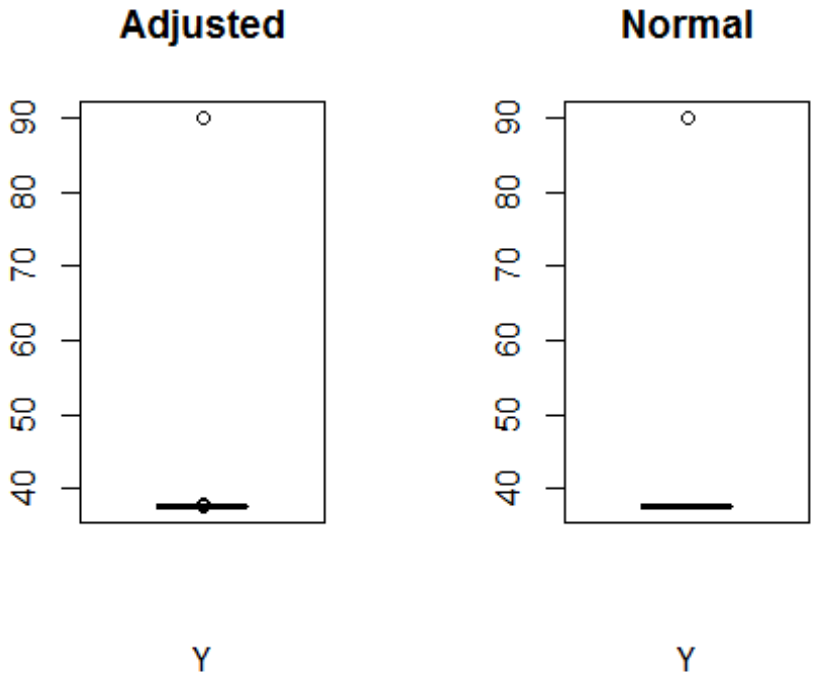

```r
hist(Recent.SFCC.Data$Dates, breaks=100)
```




```r
library(robustbase)

par(mfrow=c(1,2))
adjbox(Recent.SFCC.Data$X, main="Adjusted", xlab="X")
boxplot(Recent.SFCC.Data$X, main="Normal", xlab="X")
```




```
adjbox(Recent.SFCC.Data$Y, main="Adjusted", xlab="Y")
boxplot(Recent.SFCC.Data$Y, main="Normal", xlab="Y")
```

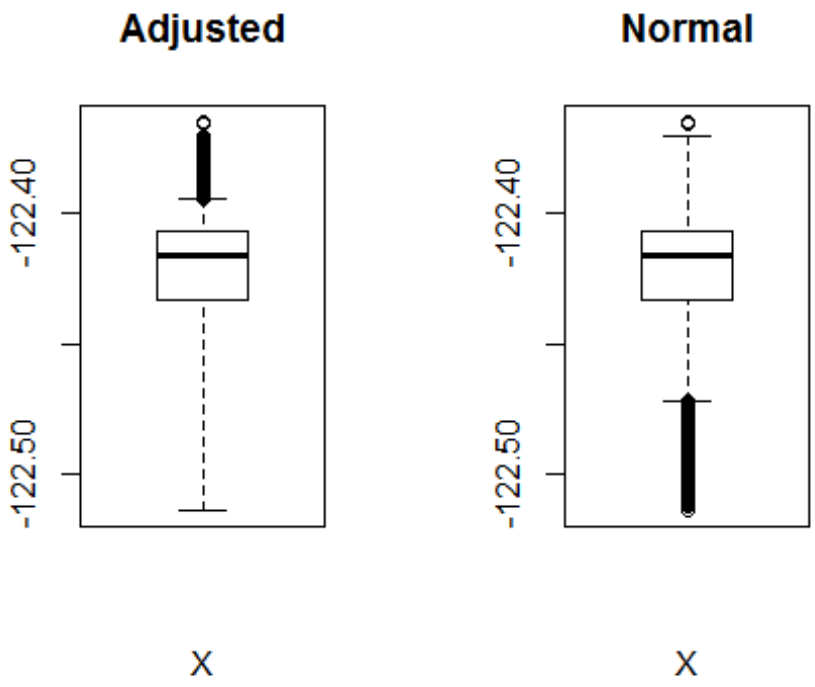



```
par(mfrow=c(1,1))

#From the boxplot above it can be seen that there is an outlier in the data. It is observed that there
#are more than one pair of outliers so we try to eliminate by range of Longitude and Latitude.
#These coordinates are from another place than San Francisco.
Recent.SFCC.Data <- Recent.SFCC.Data[Recent.SFCC.Data$Y<40,]
Recent.SFCC.Data <- Recent.SFCC.Data[Recent.SFCC.Data$X<(-122),]
```

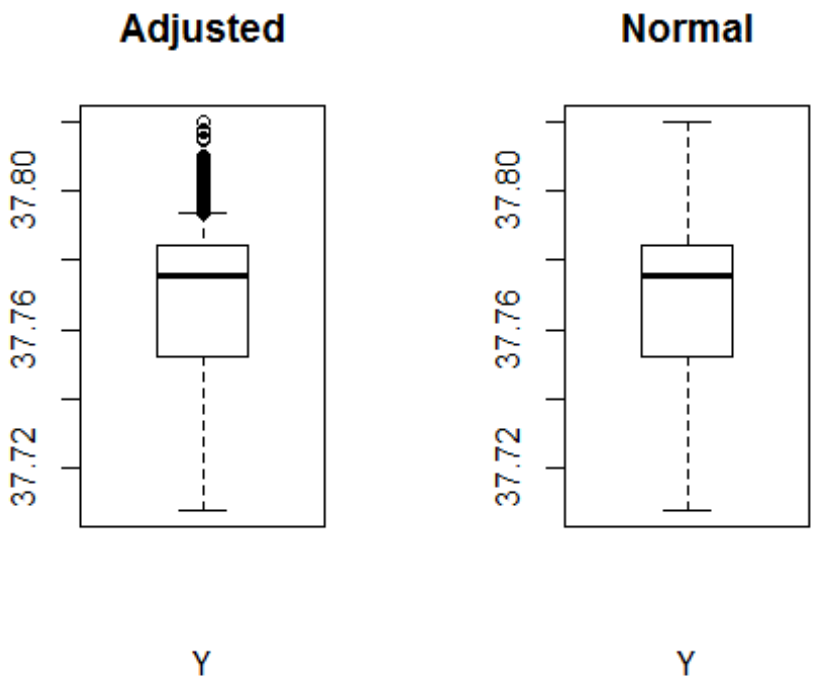
```
# Number of outliers = 67
length(SFCC.Data$X)-length(Recent.SFCC.Data$X)
```

```
## [1] 67
```

```
#The same process as above to remove outlier can be used with the "outlier" command
and can get the same
#result. Here we identified the outliers by understanding the data.
par(mfrow=c(1,2))
adjbox(Recent.SFCC.Data$X, main="Adjusted", xlab="X")
boxplot(Recent.SFCC.Data$X, main="Normal", xlab="X")
```




```
adjbox(Recent.SFCC.Data$Y, main="Adjusted", xlab="Y")
boxplot(Recent.SFCC.Data$Y, main="Normal", xlab="Y")
```



```r
par(mfrow=c(1,1))
#San Francisco is a very large city and the reminded range seems reasonable for such a large city.
```