1.) Problem Statement

Analyzing customer satisfaction data of Airlines to generate actionable insights to increase customer satisfaction and provide feedback and suggestions to the client.

2.) Business Questions

1. Which airlines do we select to analyze?
2. What are the most important factors that influence customer satisfaction?
3. Which factors when associated together are best improving customer satisfaction?
4. What are some trends or patterns we notice?
5. How do airlines with high satisfaction compare to ones with lower satisfaction?
6. What are some rules that are generating common trends in satisfaction?

3.) Data Munging / Cleaning and Preparation

1. Import the csv file into our system and viewing the data.

Code Snippet

```
airlines<-read.csv("spring19survey.csv")
View(airlines)
```

2. Analyzing the structure of the data

Code Snippet

```
str(airlines)
```

Code Output

```
'data.frame':   194833 obs. of  29 variables:
 $ Satisfaction                 : num  4 4 4 5 4 3 3 4 5 3 ...
 $ Airline.Status               : Factor w/ 4 levels "Blue","Gold",..: 1 4 1 3 2 4 1 1 2 4 ...
 $ Age                          : int  39 61 64 49 57 34 27 55 54 22 ...
 $ Gender                       : Factor w/ 2 levels "Female","Male": 2 1 1 2 2 1 1 2 1 2 ...
 $ Price.Sensitivity            : int  1 2 1 1 1 1 1 1 1 1 ...
 $ Year.of.First.Flight         : int  2006 2005 2004 2011 2011 2009 2003 2010 2007 2012 ...
 $ Flights.Per.Year             : int  22 39 24 37 0 6 24 6 68 12 ...
 $ Loyalty                      : num  -0.517 -0.733 -0.6 -0.396 1 ...
 $ Type.of.Travel               : Factor w/ 3 levels "Business travel",..: 1 1 1 1 1 3 2 1 1 1 ...
 $ Total.Freq.Flyer.Accts       : int  2 0 0 4 1 3 2 0 0 1 ...
 $ Shopping.Amount.at.Airport   : int  0 0 5 0 0 0 10 275 0 0 ...
 $ Eating.and.Drinking.at.Airport: int  60 90 70 20 30 90 110 105 10 90 ...
 $ Class                        : Factor w/ 3 levels "Business","Eco",..: 2 2 2 1 2 2 3 3 2 2 ...
 $ Day.of.Month                 : int  29 30 3 14 1 19 17 16 13 1 ...
 $ Flight.date                  : Factor w/ 90 levels "1/1/2014","1/10/2014",..: 22 24 23 6 1 70 68 67 5 1 ...
 $ Partner.Code                 : Factor w/ 14 levels "AA","AS","B6",..: 4 14 10 1 10 11 9 1 11 5 ...
 $ Partner.Name                 : Factor w/ 14 levels "Cheapseats Airlines Inc.",..: 12 1 8 11 8 10 3 11 10 4 ...
 $ Orgin.City                   : Factor w/ 295 levels "Aberdeen, SD",..: 238 12 166 73 166 126 165 198 199 126 ...
 $ Origin.State                 : Factor w/ 52 levels "Alabama","Alaska",..: 46 44 5 44 5 44 44 32 30 44 ...
 $ Destination.City             : Factor w/ 296 levels "Aberdeen, SD",..: 157 72 157 98 102 36 73 245 126 79 ...
 $ Destination.State            : Factor w/ 52 levels "Alabama","Alaska",..: 28 44 28 9 5 21 44 39 44 22 ...
 $ Scheduled.Departure.Hour     : int  8 11 16 19 14 12 11 19 20 15 ...
 $ Departure.Delay.in.Minutes   : int  0 0 0 0 0 0 0 0 1 32 ...
 $ Arrival.Delay.in.Minutes     : int  0 0 0 2 4 0 0 0 0 9 ...
 $ Flight.cancelled             : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Flight.time.in.minutes       : int  55 50 47 141 39 170 27 188 207 118 ...
 $ Flight.Distance              : int  368 323 236 1119 209 1597 140 1598 1400 1075 ...
 $ Arrival.Delay.greater.5.Mins : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 2 ...
 $ Long.Duration.Trip           : logi  FALSE FALSE FALSE TRUE FALSE TRUE ...
```

3. Dealing with missing values

As the variables that contained NA were numeric in nature, NA values were imputed with the mean.

Code Snippet

```
#Looking for missing values
sapply(airlines,function(x)  sum(is.na(x)))
```

Code Output

```
                  Satisfaction                 Airline.Status                           Age
                             8                              0                             0
                        Gender                Price.Sensitivity          Year.of.First.Flight
                             0                              0                             0
               Flights.Per.Year                      Loyalty                Type.of.Travel
                             0                              0                             0
       Total.Freq.Flyer.Accts    Shopping.Amount.at.Airport Eating.and.Drinking.at.Airport
                             0                              0                             0
                         Class                  Day.of.Month                   Flight.date
                             0                              0                             0
                  Partner.Code                  Partner.Name                   Orgin.City
                             0                              0                             0
                  Origin.State             Destination.City             Destination.State
                             0                              0                             0
      Scheduled.Departure.Hour    Departure.Delay.in.Minutes      Arrival.Delay.in.Minutes
                             0                           3538                          4105
             Flight.cancelled       Flight.time.in.minutes               Flight.Distance
                             0                           4105                             0
 Arrival.Delay.greater.5.Mins            Long.Duration.Trip
                             0                              0
```

Code Snippet

```
#Replacing NA values with mean
airlines[is.na(airlines$Flight.time.in.minutes),]$Flight.time.in.minutes<-
mean(airlines$Flight.time.in.minutes,na.rm = TRUE)
```

```
airlines[is.na(airlines$Departure.Delay.in.Minutes),]$Departure.Delay.in.Minutes<-
mean(airlines$Departure.Delay.in.Minutes,na.rm = TRUE)
airlines[is.na(airlines$Arrival.Delay.in.Minutes),]$Arrival.Delay.in.Minutes<-
mean(airlines$Arrival.Delay.in.Minutes,na.rm = TRUE)
airlines<-airlines[!is.na(airlines$Satisfaction),]
```

Code Output

```
                  Satisfaction                    Airline.Status                               Age
                             0                                 0                                 0
                        Gender                 Price.Sensitivity            Year.of.First.Flight
                             0                                 0                                 0
             Flights.Per.Year                           Loyalty                    Type.of.Travel
                             0                                 0                                 0
      Total.Freq.Flyer.Accts       Shopping.Amount.at.Airport Eating.and.Drinking.at.Airport
                             0                                 0                                 0
                         Class                     Day.of.Month                       Flight.date
                             0                                 0                                 0
                  Partner.Code                     Partner.Name                       Orgin.City
                             0                                 0                                 0
                  Origin.State                   Destination.City                Destination.State
                             0                                 0                                 0
     Scheduled.Departure.Hour          Departure.Delay.in.Minutes         Arrival.Delay.in.Minutes
                             0                                 0                                 0
             Flight.cancelled             Flight.time.in.minutes                   Flight.Distance
                             0                                 0                                 0
 Arrival.Delay.greater.5.Mins                  Long.Duration.Trip
                             0                                 0
```

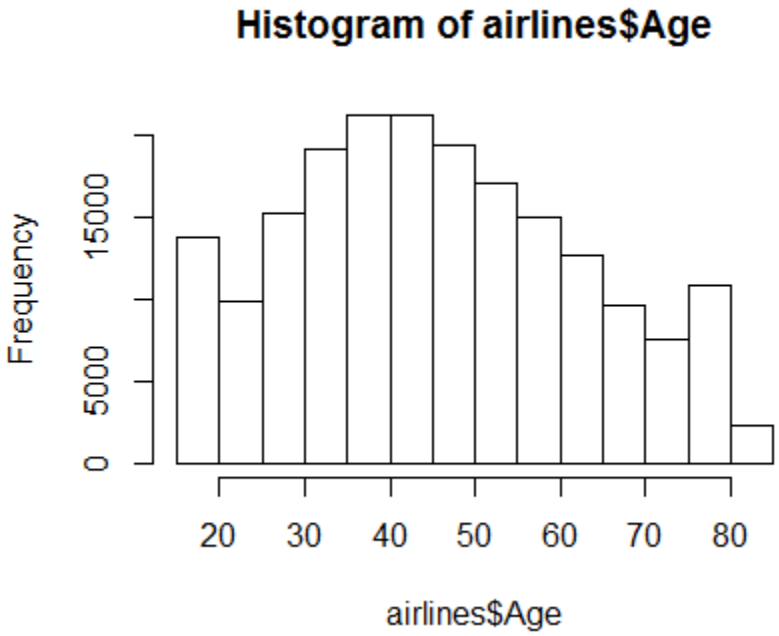
4.) Descriptive Statistics

Statistical summary of the dataset to understand the attributes of the dataset

```
> summary(cheapseats)
  Satisfaction    Airline.Status       Age            Gender      Price.Sensitivity Year.of.First.Flight Flights.Per.Year
Min.   :1.000   Blue    :27143   Min.   :15.00   Female:22076   Min.   :0.000     Min.   :2003         Min.   : 0.00
1st Qu.:3.000   Gold    : 3159   1st Qu.:33.00   Male  :17365   1st Qu.:1.000     1st Qu.:2004         1st Qu.: 9.00
Median :3.000   Platinum: 1240   Median :45.00                  Median :1.000     Median :2007         Median :17.00
Mean   :3.353   Silver  : 7899   Mean   :46.18                  Mean   :1.282     Mean   :2007         Mean   :20.05
3rd Qu.:4.000                    3rd Qu.:59.00                  3rd Qu.:2.000     3rd Qu.:2010         3rd Qu.:29.00
Max.   :5.000                    Max.   :85.00                  Max.   :4.000     Max.   :2012         Max.   :93.00

   Loyalty               Type.of.Travel  Total.Freq.Flyer.Accts Shopping.Amount.at.Airport Eating.and.Drinking.at.Airport
Min.   :-0.97619   Business travel:24099   Min.   :0.0000         Min.   :  0.00            Min.   :  0.00
1st Qu.:-0.70000   Mileage tickets: 3135   1st Qu.:0.0000         1st Qu.:  0.00            1st Qu.: 30.00
Median :-0.42857   Personal Travel:12207   Median :0.0000         Median :  0.00            Median : 60.00
Mean   :-0.27735                           Mean   :0.8997         Mean   : 26.59            Mean   : 67.65
3rd Qu.: 0.04762                           3rd Qu.:2.0000         3rd Qu.: 30.00            3rd Qu.: 90.00
Max.   : 1.00000                           Max.   :8.0000         Max.   :745.00            Max.   :765.00

     Class        Day.of.Month      Flight.date     Partner.Code                    Partner.Name          Orgin.City
Business: 3284   Min.   : 1.00   3/28/2014:  573   WN      :39441   Cheapseats Airlines Inc.:39441   Chicago, IL   : 2563
Eco     :32082   1st Qu.: 9.00   3/13/2014:  537   AA      :    0   Cool&Young Airlines Inc.:    0   Las Vegas, NV: 2535
Eco Plus: 4075   Median :16.00   3/17/2014:  533   AS      :    0   EnjoyFlying Air Services:    0   Baltimore, MD: 2110
                 Mean   :15.88   2/26/2014:  532   B6      :    0   FlyFast Airways Inc.     :    0   Phoenix, AZ   : 1982
                 3rd Qu.:23.00   1/13/2014:  521   DL      :    0   FlyHere Airways          :    0   Denver, CO    : 1800
                 Max.   :31.00   3/24/2014:  521   EV      :    0   FlyToSun Airlines Inc.   :    0   Houston, TX   : 1783
                                 (Other)  :36224   (Other):    0   (Other)                  :    0   (Other)       :26668
    Origin.State        Destination.City  Destination.State Scheduled.Departure.Hour Departure.Delay.in.Minutes
California: 7061   Las Vegas, NV: 2591   California: 6994   Min.   : 5.00            Min.   :  0.00
Texas     : 5144   Chicago, IL  : 2422   Texas     : 5162   1st Qu.: 9.00            1st Qu.:  0.00
Florida   : 3681   Phoenix, AZ  : 2017   Florida   : 3740   Median :13.00            Median :  4.00
Nevada    : 2791   Baltimore, MD: 1978   Nevada    : 2816   Mean   :13.04            Mean   : 17.96
Illinois  : 2563   Denver, CO   : 1973   Illinois  : 2422   3rd Qu.:17.00            3rd Qu.: 21.00
Arizona   : 2185   Houston, TX  : 1754   Arizona   : 2197   Max.   :22.00            Max.   :503.00
(Other)   :16016   (Other)      :26706   (Other)   :16110                            NA's   :493
Arrival.Delay.in.Minutes Flight.cancelled Flight.time.in.minutes Flight.Distance  Arrival.Delay.greater.5.Mins
Min.   :  0.00           No :38948        Min.   : 28.0         Min.   : 148.0   no :23222
1st Qu.:  0.00           Yes:  493        1st Qu.: 58.0         1st Qu.: 362.0   yes:16219
Median :  2.00                            Median : 85.0         Median : 583.0
Mean   : 16.36                            Mean   :100.5         Mean   : 704.2
3rd Qu.: 18.00                            3rd Qu.:130.0         3rd Qu.: 945.0
```

Code Snippet

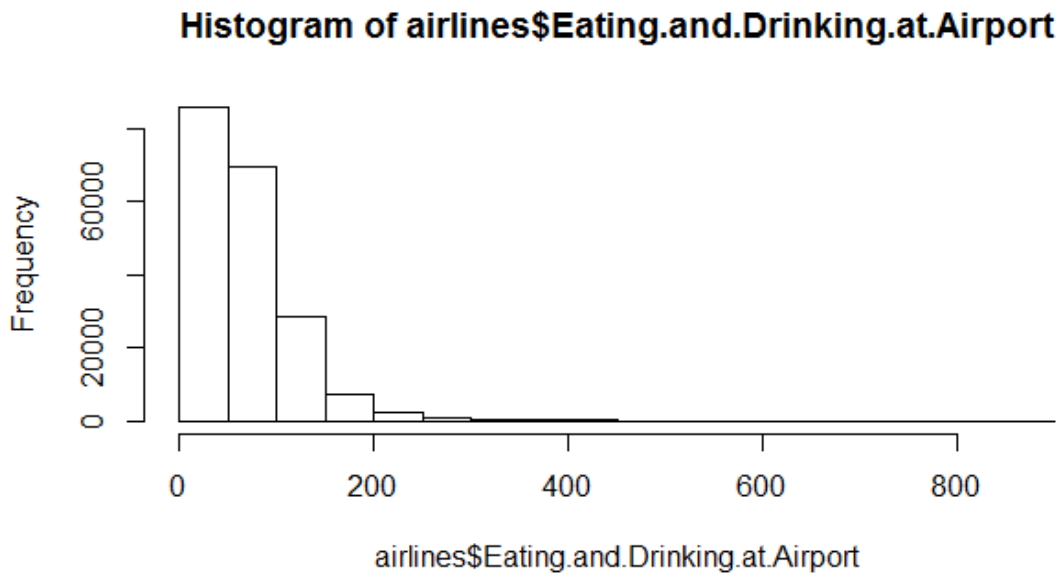
```
mean(airlines$Age)
hist(airlines$Age)
```



The mean age of the travelling passengers 46.

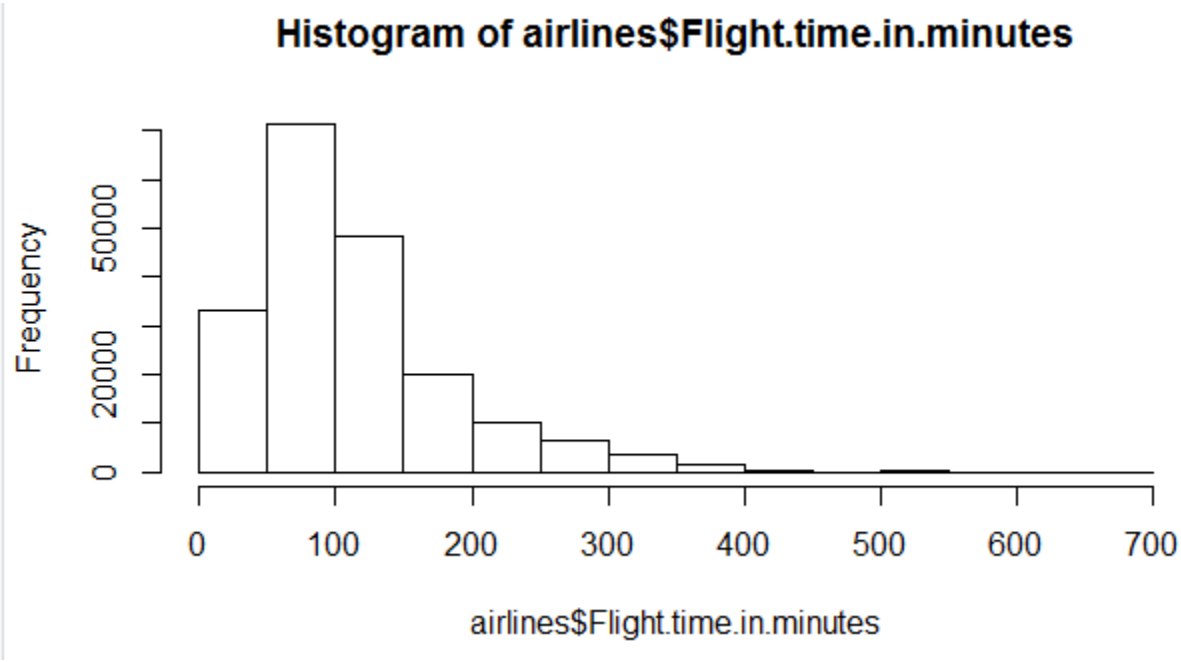
Code Snippet

```
mean(airlines$Eating.and.Drinking.at.Airport)
hist(airlines$Eating.and.Drinking.at.Airport)
```



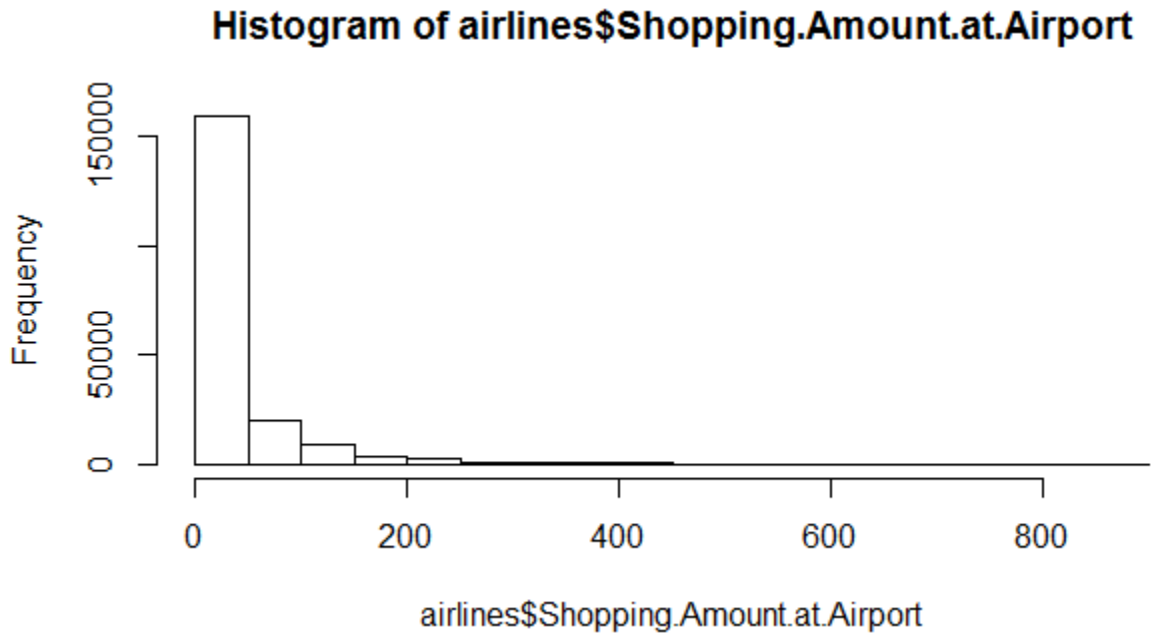
Code Snippet

```
mean(airlines$Flight.time.in.minutes)
hist(airlines$Flight.time.in.minutes)
```



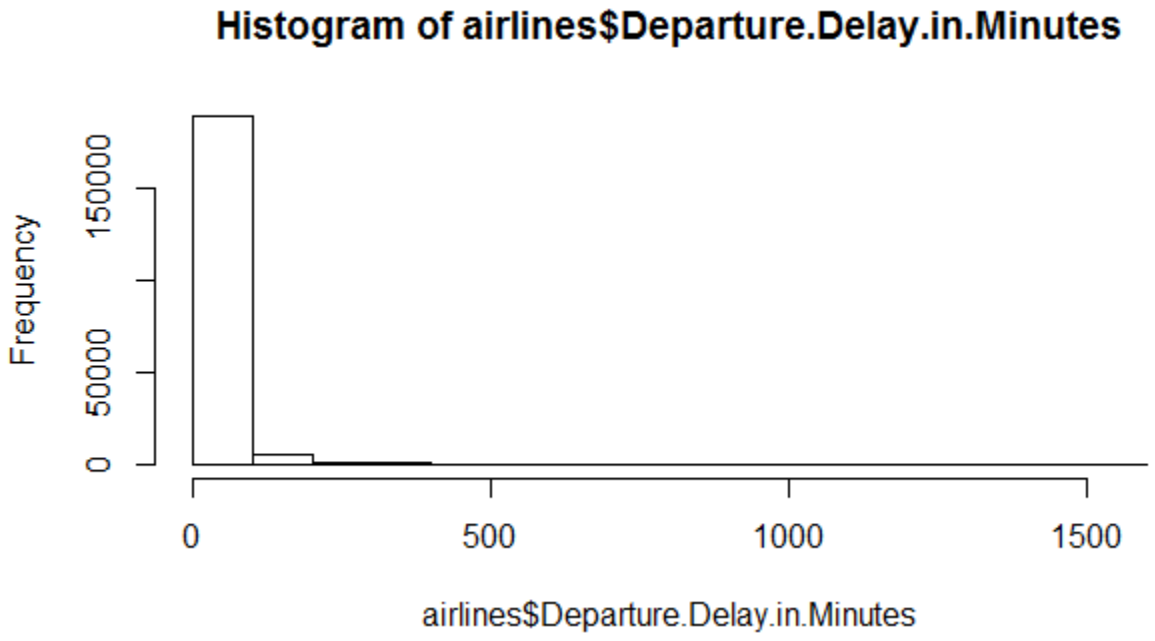
Code Snippet

```
mean(airlines$Shopping.Amount.at.Airport)
hist(airlines$Shopping.Amount.at.Airport)
```



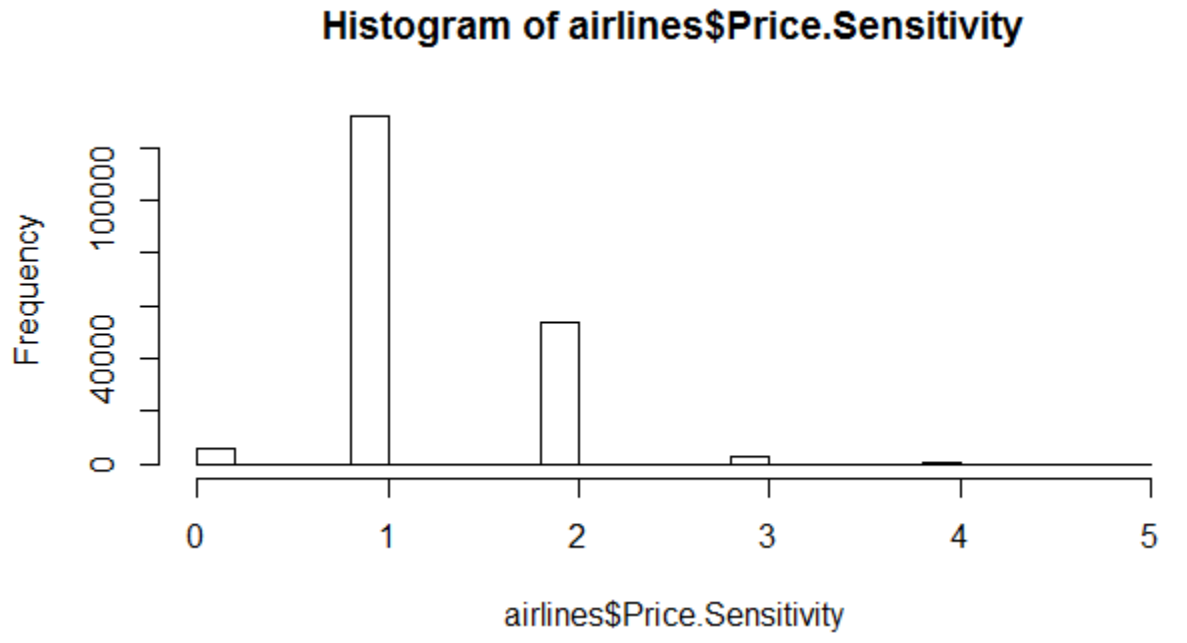
Code Snippet

```
mean(airlines$Departure.Delay.in.Minutes)
hist(airlines$Departure.Delay.in.Minutes)
```



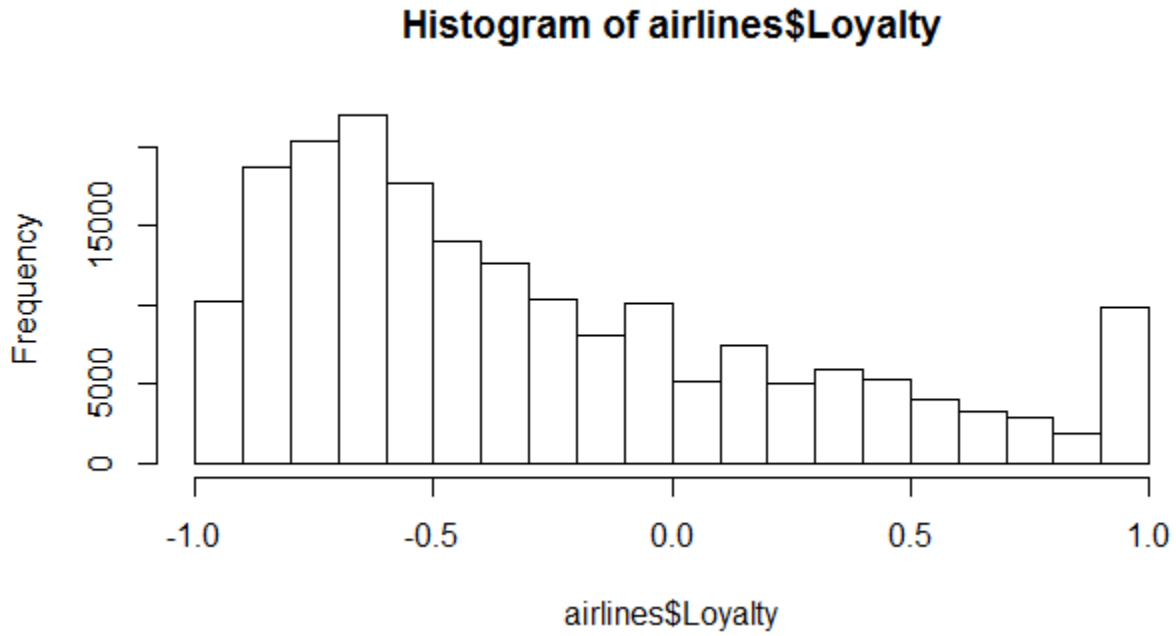
Code Snippet

```
mean(airlines$Price.Sensitivity)
hist(airlines$Price.Sensitivity)
```



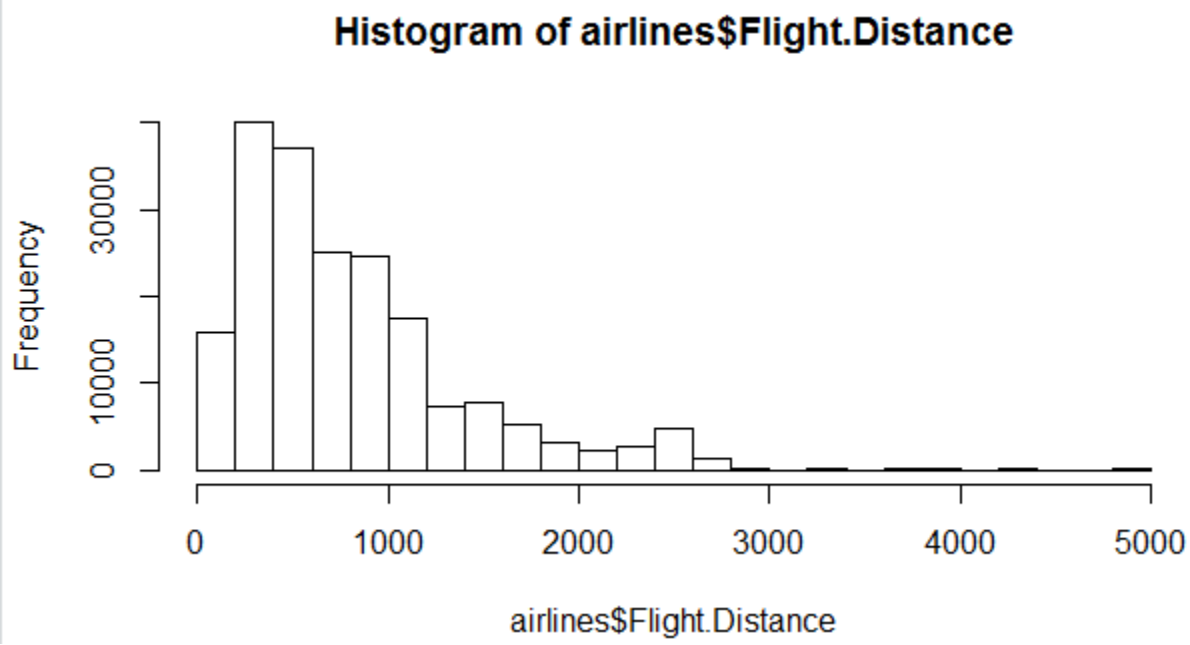
Code Snippet

```
mean(airlines$Loyalty)
hist(airlines$Loyalty)
```



Code Snippet

```
mean(airlines$Flight.Distance)
hist(airlines$Flight.Distance)
```



1. Creating Buckets of low, average and high according to the satisfaction

Code Snippet

```
colnames(airlines)[colnames(airlines)=="Partner.Name"] <- "Partner_Name"
View(airlines)

createBucketSurvey <- function(vec){              #Created a function
  vBuckets <- replicate(length(vec), "Average")      #Calculates the average
  vBuckets[vec >= 4] <- "High"                    #Value with greater than 7 will be denoted as high
  vBuckets[vec < 4] <- "Low"                  #Value with less than 7 will be denoted as low

  return(vBuckets) #Returns the value
}

airlines$happyCust <- createBucketSurvey(airlines$Satisfaction) #Call the function
createBucketSurvey and pass overallCustSat as the argument
View(airlines)

dim(airlines)
```

2. Categorizing the unhappy and happy customers according to each airline

Code Snippet

```
#sqldf("Select Partner_Name from airlines where happyCust = 'Low' '')
a <-data.frame(sqldf("Select Count(happyCust),Partner_Name from airlines where happyCust =
'High' group by Partner_Name "))
View(a)
colnames(a) <- c("Happy_Customers","Airlines")
View(a)

b <- data.frame(sqldf("Select Count(happyCust),Partner_Name from airlines where happyCust =
'Low' group by Partner_Name "))
View(b)
colnames(b) <- c("UnHappy_Customers","Airlines")
View(b)

Happy_Unhappy_Customers <- merge (a,b,by="Airlines")
View(Happy_Unhappy_Customers)
Happy_Unhappy_Customers$ratio_Un <-
1000*(Happy_Unhappy_Customers$UnHappy_Customers/(Happy_Unhappy_Customers$Happy
_Customers+Happy_Unhappy_Customers$UnHappy_Customers))
```

```
Happy_Unhappy_Customers$ratio_HA <-
1000*(Happy_Unhappy_Customers$Happy_Customers/(Happy_Unhappy_Customers$Happy_C
ustomers+Happy_Unhappy_Customers$UnHappy_Customers))
View(Happy_Unhappy_Customers)
#boxplot(Happy_Unhappy_Customers$Airlines, Happy_Unhappy_Customers$ratio_Un)
```

Code Output:

Count of Happy Customers for each Airline:

	Happy_Customers	Airlines
1	19685	Cheapseats Airlines Inc.
12	13181	Sigma Airlines Inc.
4	11663	FlyFast Airways Inc.
8	10918	Northwest Business Airlines Inc.
11	9499	Paul Smith Airlines Inc.
10	8426	Oursin Airlines Inc.
13	7351	Southeast Airlines Co.
3	6738	EnjoyFlying Air Services
9	3972	OnlyJets Airlines Inc.
6	2725	FlyToSun Airlines Inc.
5	1960	FlyHere Airways
14	1461	West Airways Inc.
7	1062	GoingNorth Airlines Inc.
2	1037	Cool&Young Airlines Inc.

Count of Unhappy Customers of each Airline:

	UnHappy_Customers	Airlines
1	19756	Cheapseats Airlines Inc.
2	880	Cool&Young Airlines Inc.
3	6631	EnjoyFlying Air Services
4	11395	FlyFast Airways Inc.
5	1759	FlyHere Airways
6	2345	FlyToSun Airlines Inc.
7	1240	GoingNorth Airlines Inc.
8	9925	Northwest Business Airlines Inc.
9	4174	OnlyJets Airlines Inc.
10	7961	Oursin Airlines Inc.
11	8759	Paul Smith Airlines Inc.
12	12183	Sigma Airlines Inc.
13	7028	Southeast Airlines Co.
14	1111	West Airways Inc.

3. Merging the tables and calculating the ratio of the same

Code Snippet

```
Happy_Unhappy_Customers <- merge (a,b,by="Airlines")
View(Happy_Unhappy_Customers)
Happy_Unhappy_Customers$ratio_Un <-
1000*(Happy_Unhappy_Customers$UnHappy_Customers/(Happy_Unhappy_Customers$Happy_Customers+Happy_Unhappy_Customers$UnHappy_Customers))
Happy_Unhappy_Customers$ratio_HA <-
1000*(Happy_Unhappy_Customers$Happy_Customers/(Happy_Unhappy_Customers$Happy_Customers+Happy_Unhappy_Customers$UnHappy_Customers))
View(Happy_Unhappy_Customers)
#boxplot(Happy_Unhappy_Customers$Airlines, Happy_Unhappy_Customers$ratio_Un)
```

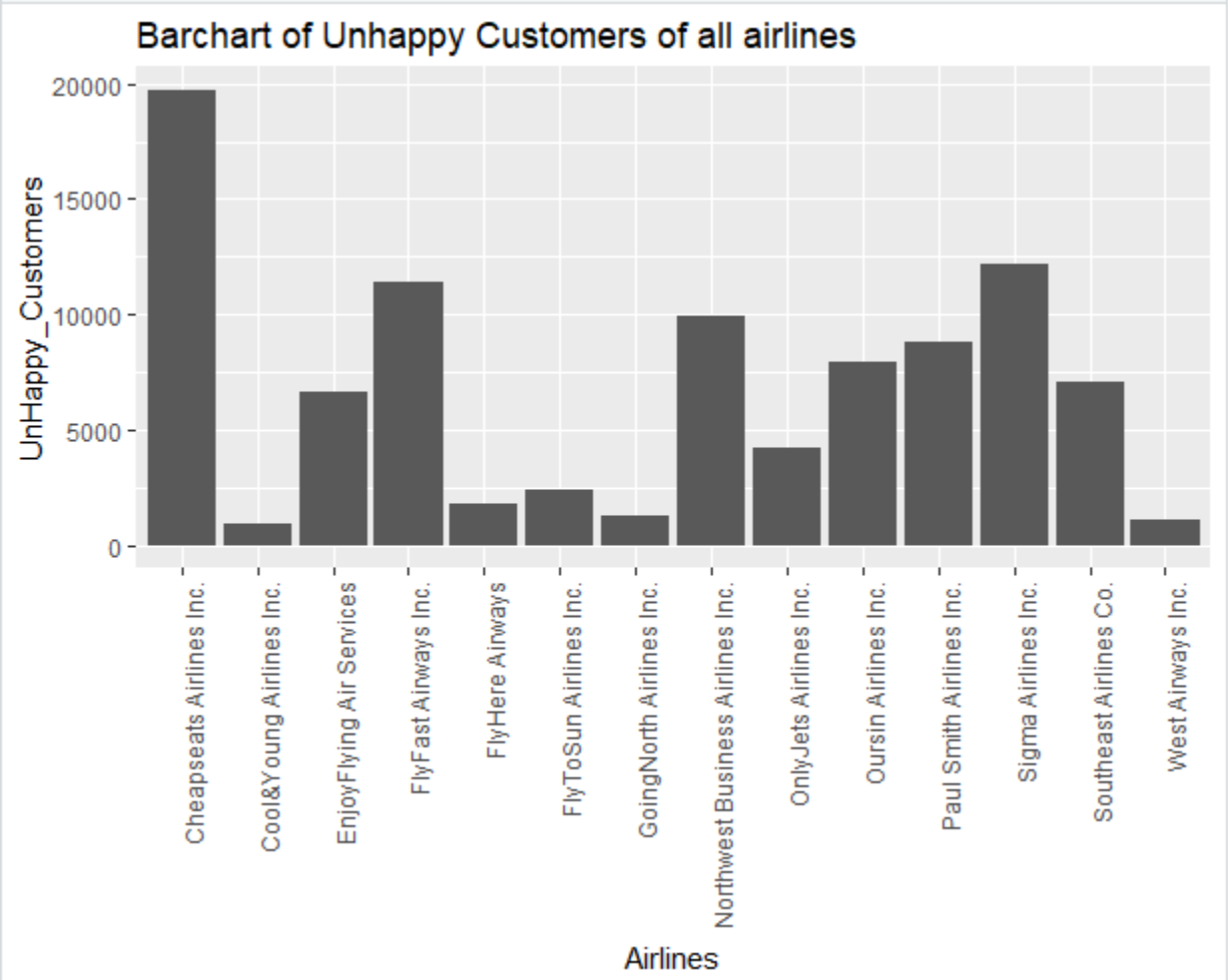
	Airlines	Happy_Customers	UnHappy_Customers	ratio_Un	ratio_HA
1	Cheapseats Airlines Inc.	19685	19756	500.9001	499.0999
2	Cool&Young Airlines Inc.	1037	880	459.0506	540.9494
3	EnjoyFlying Air Services	6738	6631	495.9982	504.0018
4	FlyFast Airways Inc.	11663	11395	494.1886	505.8114
5	FlyHere Airways	1960	1759	472.9766	527.0234
6	FlyToSun Airlines Inc.	2725	2345	462.5247	537.4753
7	GoingNorth Airlines Inc.	1062	1240	538.6620	461.3380
8	Northwest Business Airlines Inc.	10918	9925	476.1791	523.8209
9	OnlyJets Airlines Inc.	3972	4174	512.3987	487.6013
10	Oursin Airlines Inc.	8426	7961	485.8119	514.1881
11	Paul Smith Airlines Inc.	9499	8759	479.7349	520.2651
12	Sigma Airlines Inc.	13181	12183	480.3264	519.6736
13	Southeast Airlines Co.	7351	7028	488.7683	511.2317
14	West Airways Inc.	1461	1111	431.9596	568.0404

3. Creating bar graphs of the unhappy customers of all airlines

Code Snippet:

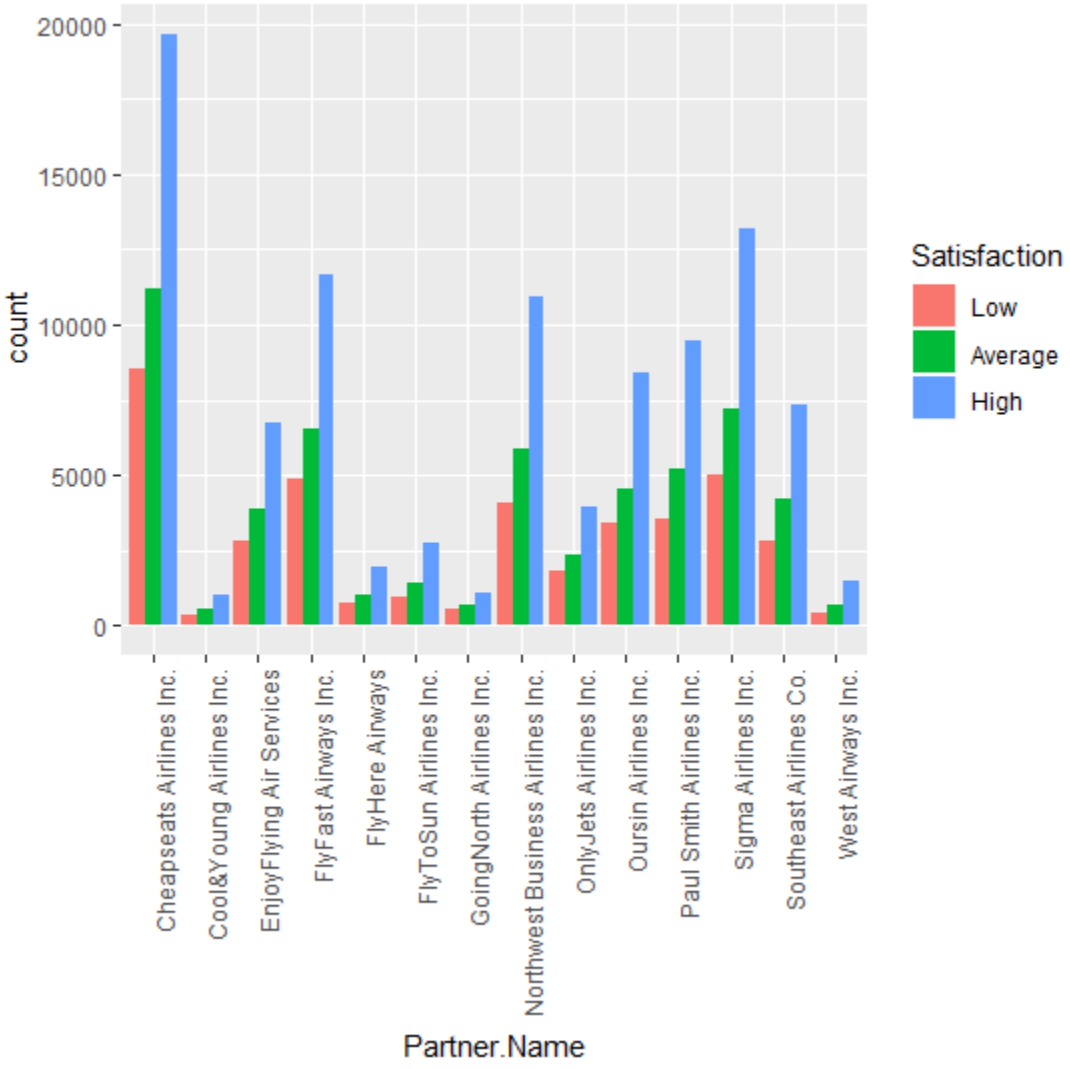
```
barchart1 <- ggplot(Happy_Unhappy_Customers,aes(x=Airlines,y=UnHappy_Customers)) +
geom_col() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) + ggtitle("Barchart of Unhappy Customers of all airlines")
barchart1
```

Code Output:


Barchart of Unhappy Customers of all airlines
UnHappy_Customers
20000
15000
10000
5000
0
Cheapseats Airlines Inc.
Cool&Young Airlines Inc.
EnjoyFlying Air Services
FlyFast Airways Inc.
FlyHere Airways
FlyToSun Airlines Inc.
GoingNorth Airlines Inc.
Northwest Business Airlines Inc.
OnlyJets Airlines Inc.
Oursin Airlines Inc.
Paul Smith Airlines Inc.
Sigma Airlines Inc.
Southeast Airlines Co.
West Airways Inc.
Airlines

Code Snippet

```
ggplot(airlines,aes(x=Partner.Name,fill=Satisfaction))+geom_bar(position='dodge')+theme(axis.text.x = element_text(angle=90,hjust=1))
```



It was analyzed from the above 2 graphs that the number of unhappy customers for the Cheapseats Airlines is the most and the satisfaction for the Cheapseats Airlines is the lowest. Also, there are a lot of customers who are travelling by Cheapseats according to the above graph. Therefore, Cheapseats would be the best Airline for carrying out the analysis.

Descriptive Statistics of Airlines:

Code Snippet:

```
summary(cheapseats)
```

Code Output:

```
> summary(cheapseats)
  Satisfaction    Airline.Status       Age           Gender      Price.Sensitivity
 Min.   :1.000   Blue    :27143   Min.   :15.00   Female:22076   Min.   :0.000
 1st Qu.:3.000   Gold    : 3159   1st Qu.:33.00   Male  :17365   1st Qu.:1.000
 Median :3.000   Platinum: 1240   Median :45.00                  Median :1.000
 Mean   :3.353   Silver  : 7899   Mean   :46.18                  Mean   :1.282
 3rd Qu.:4.000                    3rd Qu.:59.00                  3rd Qu.:2.000
 Max.   :5.000                    Max.   :85.00                  Max.   :4.000

 Year.of.First.Flight Flights.Per.Year     Loyalty                Type.of.Travel
 Min.   :2003         Min.   : 0.00    Min.   :-0.97619   Business travel:24099
 1st Qu.:2004         1st Qu.: 9.00    1st Qu.:-0.70000   Mileage tickets: 3135
 Median :2007         Median :17.00    Median :-0.42857   Personal Travel:12207
 Mean   :2007         Mean   :20.05    Mean   :-0.27735
 3rd Qu.:2010         3rd Qu.:29.00    3rd Qu.: 0.04762
 Max.   :2012         Max.   :93.00    Max.   : 1.00000

 Total.Freq.Flyer.Accts Shopping.Amount.at.Airport Eating.and.Drinking.at.Airport
 Min.   :0.0000         Min.   :  0.00             Min.   :  0.00
 1st Qu.:0.0000         1st Qu.:  0.00             1st Qu.: 30.00
 Median :0.0000         Median :  0.00             Median : 60.00
 Mean   :0.8997         Mean   : 26.59             Mean   : 67.65
 3rd Qu.:2.0000         3rd Qu.: 30.00             3rd Qu.: 90.00
 Max.   :8.0000         Max.   :745.00             Max.   :765.00

     Class        Day.of.Month      Flight.date     Partner.Code
 Business: 3284   Min.   : 1.00   3/28/2014:  573   WN     :39441
 Eco     :32082   1st Qu.: 9.00   3/13/2014:  537   AA     :    0
 Eco Plus: 4075   Median :16.00   3/17/2014:  533   AS     :    0
                  Mean   :15.88   2/26/2014:  532   B6     :    0
                  3rd Qu.:23.00   1/13/2014:  521   DL     :    0
                  Max.   :31.00   3/24/2014:  521   EV     :    0
                                  (Other)  :36224   (Other):    0
                  Partner_Name          Orgin.City        Origin.State
 Cheapseats Airlines Inc.:39441   Chicago, IL  : 2563   California: 7061
 Cool&Young Airlines Inc.:    0   Las Vegas, NV: 2535   Texas     : 5144
 EnjoyFlying Air Services:    0   Baltimore, MD: 2110   Florida   : 3681
```

Code Snippet:

```
table(cheapseats$Satisfaction)
```

Code Output:

```
> table(cheapseats$Satisfaction)

    1     2     3     4     5
  952  7579 11225 15979  3706
```

The above statistics show the number of votes for each satisfaction.

Code snippet:

```
mean(cheapseats$Age)
```

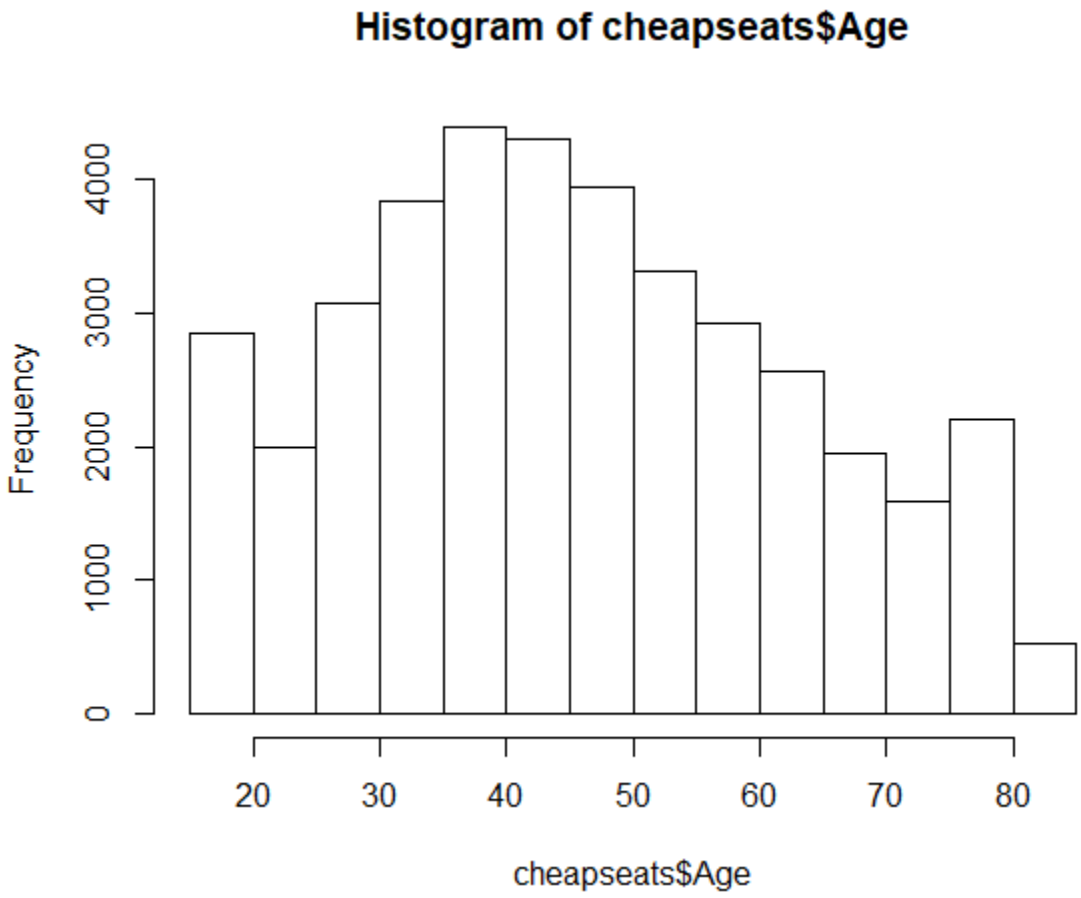
Code Output:

```
> mean(cheapseats$Age)
[1] 46.17984
```

Code Snippet:

```
hist(cheapseats$Age)
```

Code Output:



Code Snippet:

```
mean(cheapseats$Eating.and.Drinking.at.Airport)
```

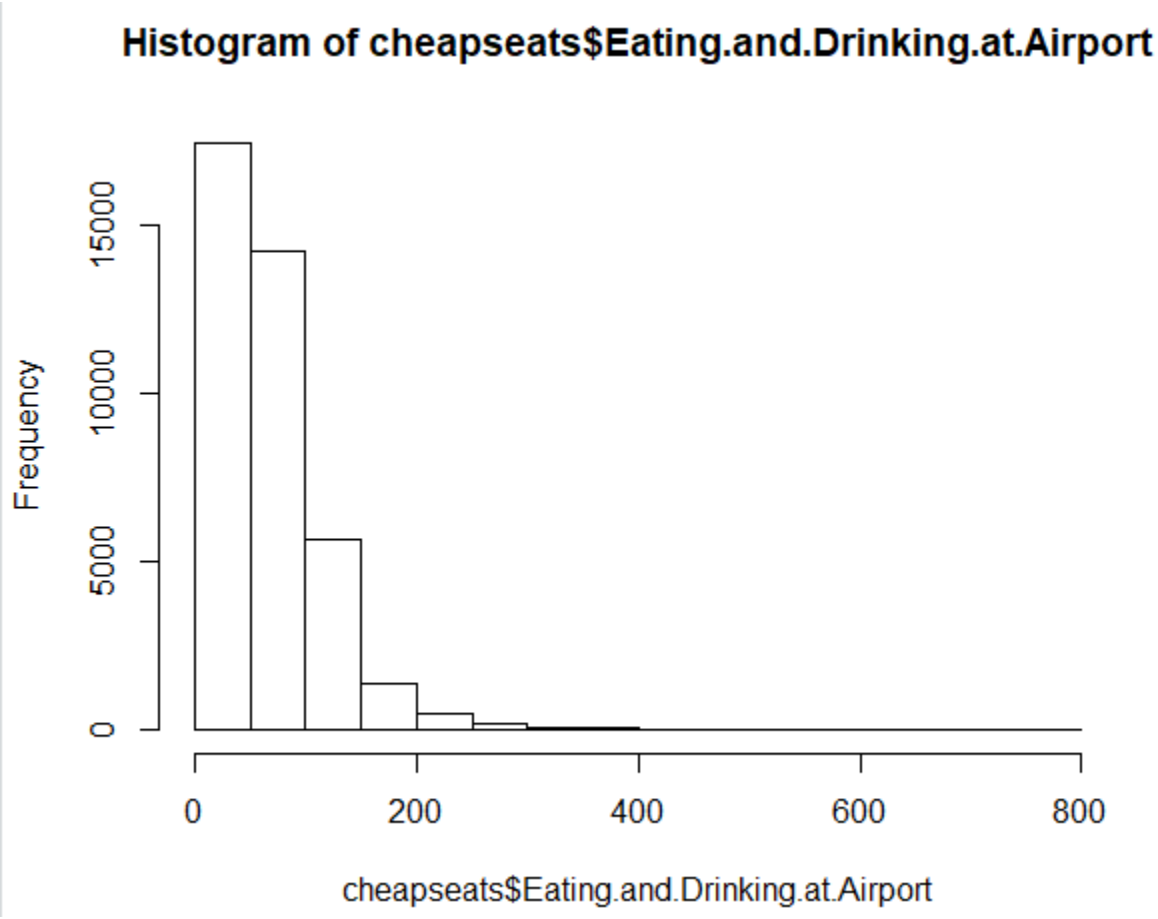
Code Output:

```
> mean(cheapseats$Eating.and.Drinking.at.Airport)
[1] 67.64988
```

Code Snippet:

```
hist(cheapseats$Eating.and.Drinking.at.Airport)
```

Code Output:



Code Snippet:

mean(cheapseats$Flight.time.in.minutes)

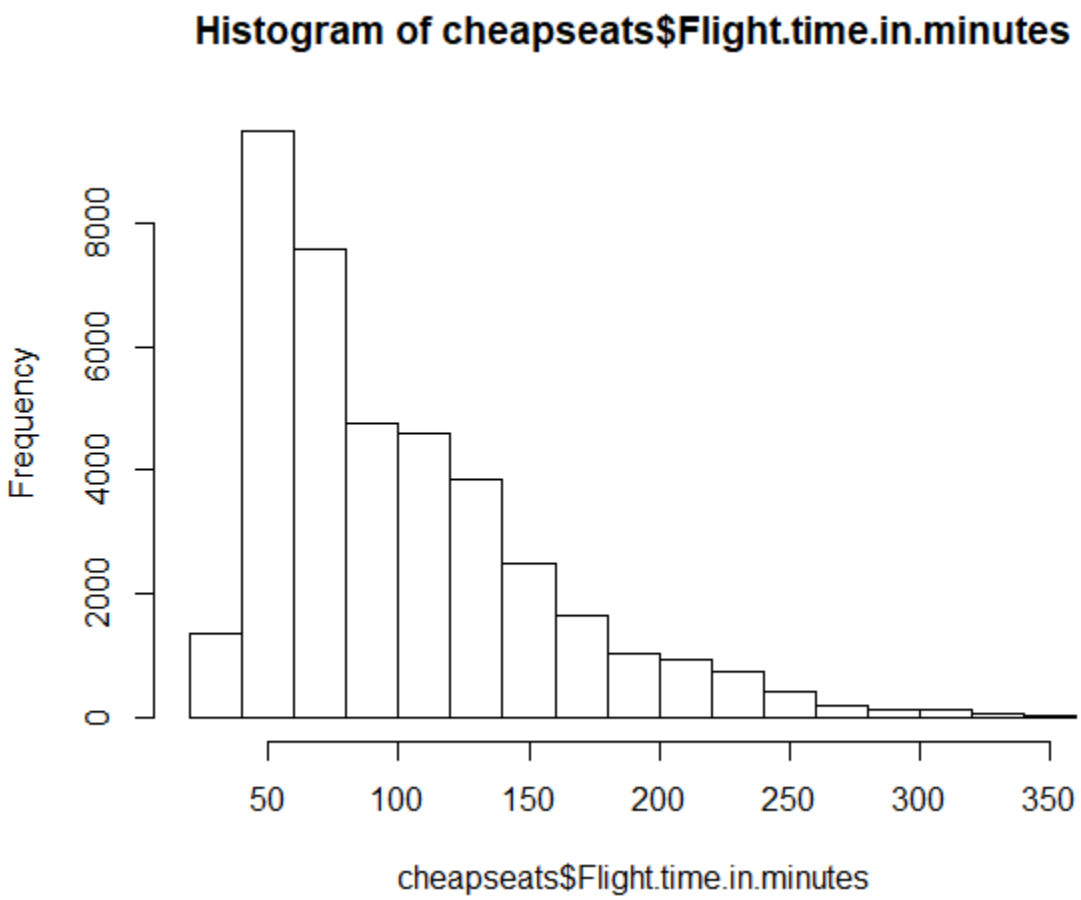
Code Output:

```
> mean(cheapseats$Flight.time.in.minutes)
[1] 100.6489
```

Code Snippet:

hist(cheapseats$Flight.time.in.minutes)

Code Output:



Code Snippet:

```
mean(cheapseats$Shopping.Amount.at.Airport)
```

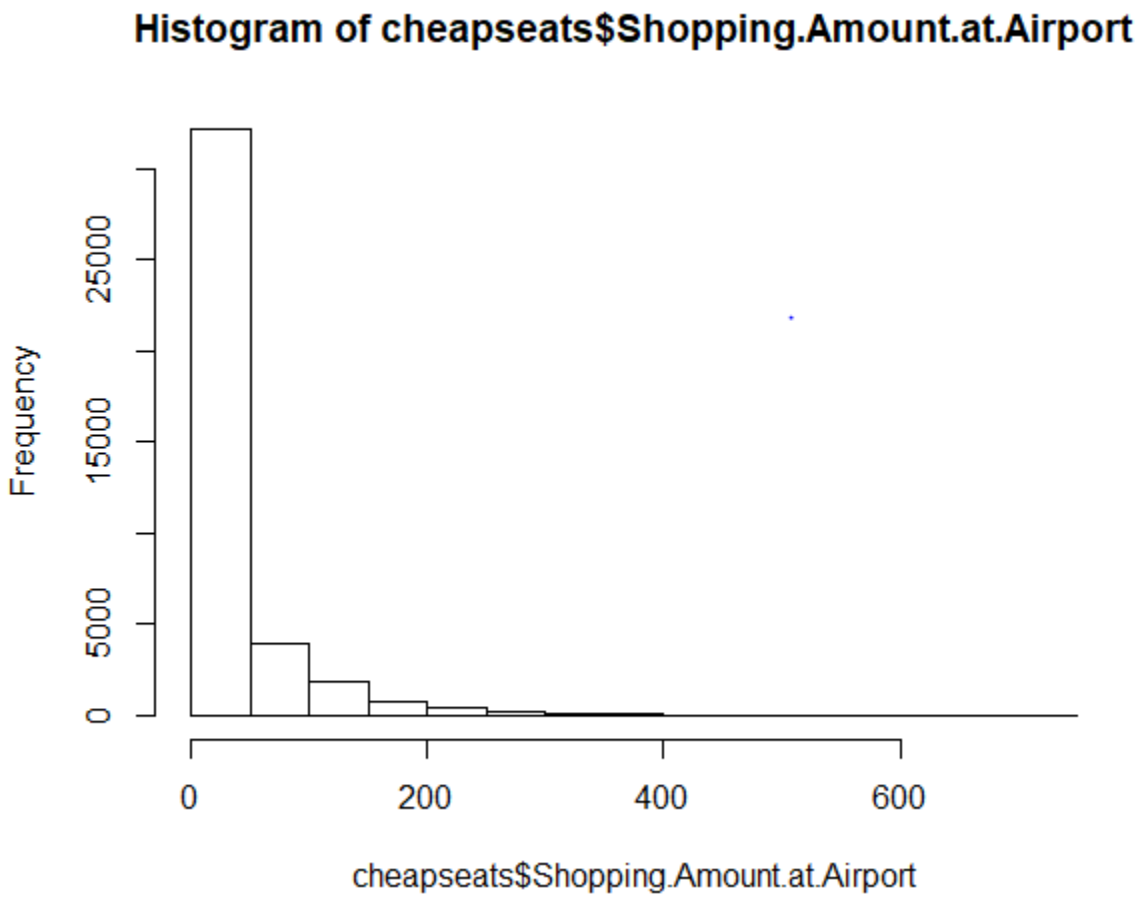
Code Output:

```
> mean(cheapseats$Shopping.Amount.at.Airport)
[1] 26.58771
```

Code Snippet:

```
hist(cheapseats$Shopping.Amount.at.Airport)
```

Code Output:



Code Snippet:

```
mean(cheapseats$Departure.Delay.in.Minutes)
```

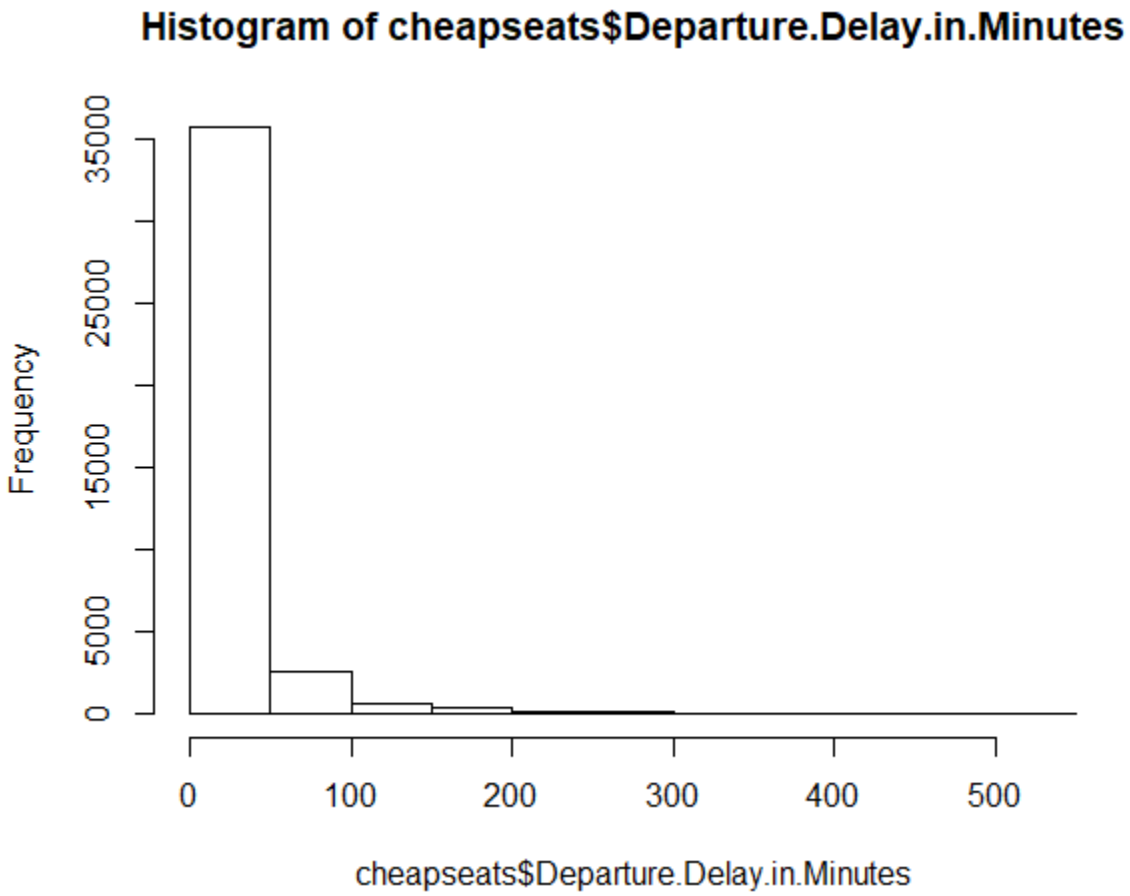
Code Output:

```
> mean(cheapseats$Departure.Delay.in.Minutes)
[1] 17.91763
```

Code Snippet:

```
hist(cheapseats$Departure.Delay.in.Minutes)
```

Code Output:



Code Snippet:

```
mean(cheapseats$Price.Sensitivity)
```

Code Output:

```
> mean(cheapseats$Price.Sensitivity)
[1] 1.281687
```

Code Snippet:

```
mean(cheapseats$Arrival.Delay.in.Minutes)
```

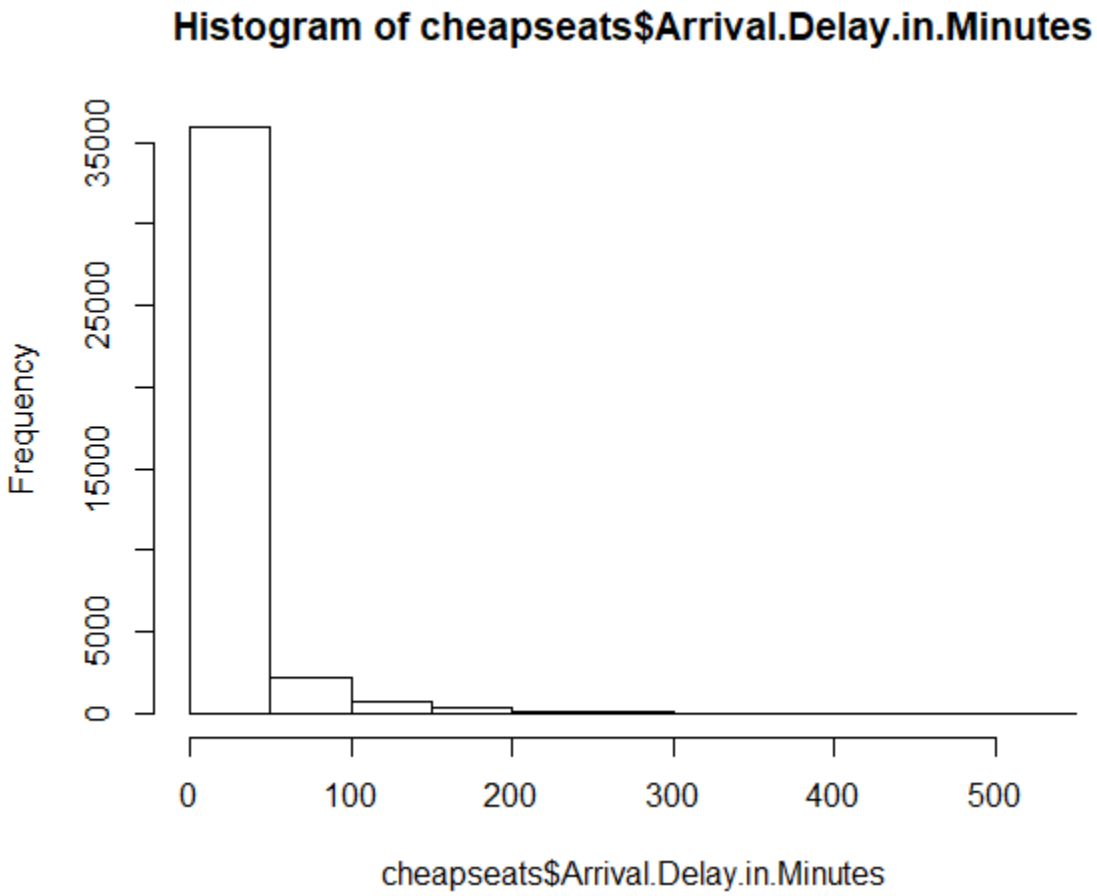
Code Output:

```
> mean(cheapseats$Arrival.Delay.in.Minutes)
[1] 16.33955
```

Code Snippet:

```
hist(cheapseats$Arrival.Delay.in.Minutes)
```

Code Output:



Code Snippet:

```
mean(cheapseats$Loyalty)
```

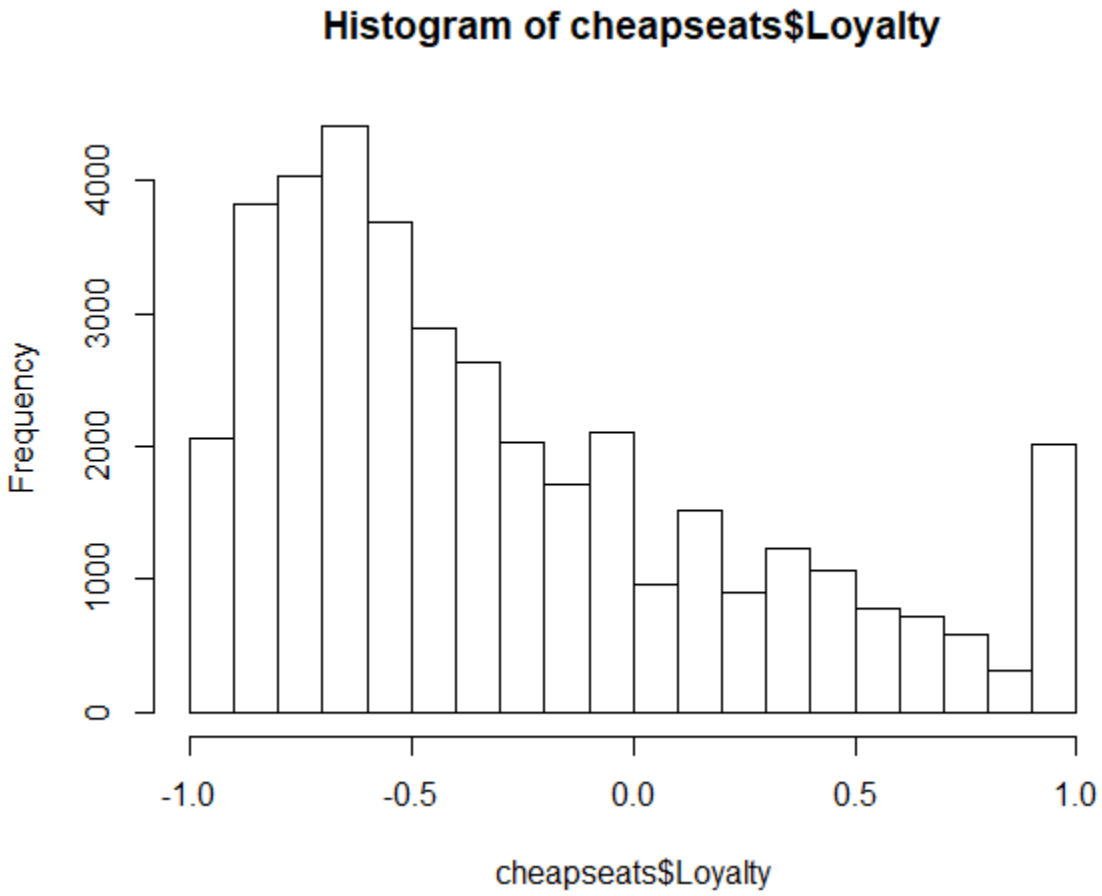
Code Output:

```
> mean(cheapseats$Loyalty)
[1] -0.277346
```

Code Snippet:

```
hist(cheapseats$Loyalty)
```

Code Output:



Code Snippet:

```
mean(cheapseats$Flight.Distance)
```

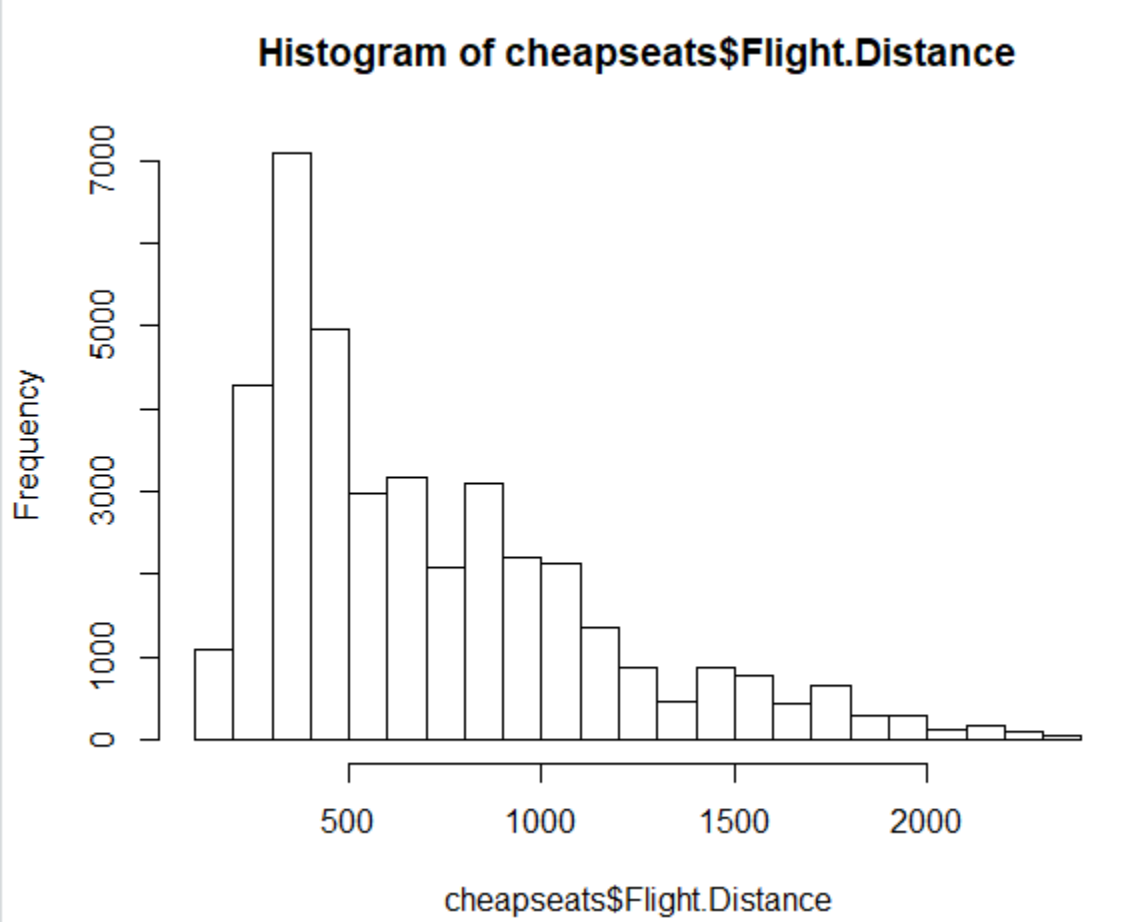
Code Output:

```
> mean(cheapseats$Flight.Distance)
[1] 704.1948
```

Code Snippet:

```
hist(cheapseats$Flight.Distance)
```

Code Output:



5. Modeling Techniques Used

i. Linear Modeling

Code Snippet:

```
CheapseatsDf <- airlinedfCleaned
CheapseatsDf <- subset(CheapseatsDf,select=-
c(Gender,Year.of.First.Flight,Type.of.Travel,Partner.Name, Orgin.City,Origin.State,
Destination.City, Destination.State, Long.Duration.Trip, Flight.date, Partner.Code))

View(head(CheapseatsDf))
str(CheapseatsDf)
model1 <- lm(formula=Satisfaction ~ ., data = CheapseatsDf)
summary(model1)
#Summary of Linkear model for Cheapsets

createBucketSurvey <- function(vec){              #Created a function
  vBuckets <- replicate(length(vec), "Average")      #Calculates the average
```

```
 vBuckets[vec >= 3] <- "High"   #Value with greater than 7 will be denoted as high
 vBuckets[vec < 3] <- "Low"          #Value with less than 7 will be denoted as low

 return(vBuckets) #Returns the value
}

airlinedfCleaned$happyCust <-createBucketSurvey(airlinedfCleaned$Satisfaction)
str(airlinedfCleaned)
require(dplyr)
?count
count(airlinedfCleaned, c("airlinedfCleaned$Partner.Name", "airlinedfCleaned$happyCust"))
```

Code Output

```
Call:
lm(formula = Satisfaction ~ ., data = CheapseatsDf)

Residuals:
     Min       1Q   Median       3Q      Max
-3.07188 -0.57289  0.02366  0.63747  2.64157

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.364e+00  1.354e-02 322.363  < 2e-16 ***
Airline.StatusGold             5.581e-01  7.096e-03  78.650  < 2e-16 ***
Airline.StatusPlatinum         4.435e-01  1.098e-02  40.373  < 2e-16 ***
Airline.StatusSilver           7.279e-01  4.949e-03 147.094  < 2e-16 ***
Age                           -1.155e-02  1.278e-04 -90.361  < 2e-16 ***
Price.Sensitivity             -1.404e-01  3.559e-03 -39.458  < 2e-16 ***
Flights.Per.Year              -1.221e-02  1.939e-04 -62.977  < 2e-16 ***
Loyalty                       -8.194e-02  5.370e-03 -15.259  < 2e-16 ***
Total.Freq.Flyer.Accts        -2.718e-02  1.971e-03 -13.792  < 2e-16 ***
Shopping.Amount.at.Airport    -9.050e-05  3.626e-05  -2.496  0.01257 *
Eating.and.Drinking.at.Airport -4.585e-04  3.767e-05 -12.172  < 2e-16 ***
ClassEco                      -1.358e-01  7.018e-03 -19.351  < 2e-16 ***
ClassEco Plus                 -1.996e-01  9.003e-03 -22.173  < 2e-16 ***
Day.of.Month                   3.916e-05  2.228e-04   0.176  0.86050
Scheduled.Departure.Hour       3.950e-03  4.208e-04   9.387  < 2e-16 ***
Departure.Delay.in.Minutes     7.009e-05  1.904e-04   0.368  0.71274
Arrival.Delay.in.Minutes       3.954e-05  1.954e-04   0.202  0.83960
Flight.cancelledYes           -3.087e-01  1.451e-02 -21.280  < 2e-16 ***
Flight.time.in.minutes         3.223e-04  1.136e-04   2.836  0.00457 **
Flight.Distance               -3.604e-05  1.364e-05  -2.643  0.00821 **
Arrival.Delay.greater.5.Minsyes -3.507e-01  4.889e-03 -71.734  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8474 on 194804 degrees of freedom
Multiple R-squared:  0.2273,    Adjusted R-squared:  0.2272
F-statistic:  2865 on 20 and 194804 DF,  p-value: < 2.2e-16
```

ii. Association Rule Mining

Association rules mining was used to identify strong rules to predict Satisfaction. Since arules apriori does not accept numeric variables, we had to discretize the continuous numeric variables into discrete buckets. Discretize function from arules and cut function on base R was used.

Satisfaction was bucketed into 3 parts, 1,2 being Low, 2.5,3,3.5 being Average and 4,5 being High.

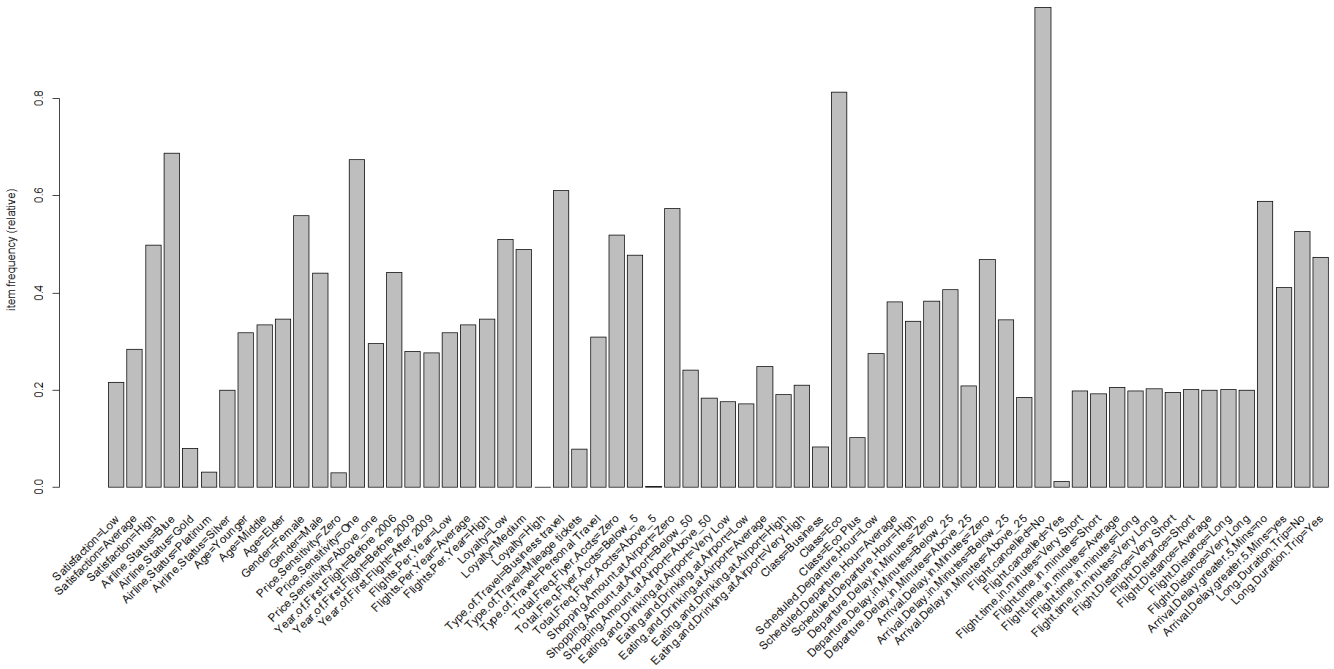
Different numeric variables were inspected to check for distributions and were discretized accordingly

The transformed data was then converted into transactional data to be used for association rule mining.

Item Frequency plot was obtained

```
cheapseats$Flight.time.in.minutes<-discretize(cheapseats$Flight.time.in.minutes,method =
"frequency",breaks=5,labels = c("Very Short","Short","Average","Long","Very Long"),order=T)
cheapseats$Flight.Distance<-discretize(cheapseats$Flight.Distance,method =
"frequency",breaks=5,labels = c("Very Short","Short","Average","Long","Very Long"),order=T)
cheapseats$Flights.Per.Year<-discretize(cheapseats$Flights.Per.Year,method =
"frequency",breaks=3,labels = c("Low","Average","High"),order=T)

cheapseats$Eating.and.Drinking.at.Airport<-
discretize(cheapseats$Eating.and.Drinking.at.Airport,method = "frequency",breaks=5,labels =
c("Very Low","Low","Average","High","Very High"),order=T)
cheapseats$Arrival.Delay.in.Minutes<-cut(cheapseats$Arrival.Delay.in.Minutes,c(-
Inf,0,25,Inf),labels=c("Zero","Below_25","Above_25"))
cheapseats$Scheduled.Departure.Hour<-
discretize(cheapseats$Scheduled.Departure.Hour,method = "frequency",breaks=3,labels =
c("Low","Average","High"),order=T)
cheapseats$Age<-discretize(cheapseats$Age,method = "frequency",breaks=3,labels =
c("Younger","Middle","Elder"),order=T)

cheapseats$Departure.Delay.in.Minutes<-cut(cheapseats$Departure.Delay.in.Minutes,c(-
Inf,0,25,Inf),labels=c("Zero","Below_25","Above_25"))
cheapseats$Shopping.Amount.at.Airport<-cut(cheapseats$Shopping.Amount.at.Airport,c(-
Inf,0,50,Inf),labels=c("Zero","Below_50","Above_50"))

cheapseats$Total.Freq.Flyer.Accts<-cut(cheapseats$Total.Freq.Flyer.Accts,c(-
Inf,0,5,Inf),labels=c("Zero","Below_5","Above_5"))
cheapseats$Price.Sensitivity<-cut(cheapseats$Price.Sensitivity,c(-
Inf,0,1,5),labels=c("Zero","One","Above_one"))

cheapseats$Loyalty<-cut(cheapseats$Loyalty,c(-Inf,-
0.428,1,Inf),labels=c("Low","Medium","High"))
```

```
cheapseats$Long.Duration.Trip<-ifelse(cheapseats$Long.Duration.Trip==0,"No","Yes")

cheapseats$Year.of.First.Flight<-cut(cheapseats$Year.of.First.Flight,c(-
Inf,2006,2009,Inf),labels=c("Before 2006","Before 2009","After 2009"))

cheapseats$Satisfaction<-as.numeric(cheapseats$Satisfaction)

cheapseats$Satisfaction<-
cut(cheapseats$Satisfaction,c(0,2,4,Inf),labels=c("Low","Average","High"))

cheapseats<-cheapseats[,!names(cheapseats) %in%
c("Day.of.Month","Partner.Name","Partner.Code","Destination.State","Destination.City","Origi
n.State","Orgin.City","Flight.date")]
cheapseats$Long.Duration.Trip<-as.factor(cheapseats$Long.Duration.Trip)


cheapseats_trans<-as(cheapseats,"transactions")
```



Flight Cancelled=No, Satisfaction=High ,class=Eco,Arrival.Delay.Greater.Than.5.mins=Yes are some of the frequent items in the transactions dataset.

Wrote a function to subset redundant rules as most of the rules are subset of other rules

```r
#subset rules function
subset_rules<-function(Rules){
  #Removing redundant rules
  sub_rules <- which(colSums(is.subset(Rules,Rules)) > 1)
  Rules <- sort(Rules[-sub_rules], by = "lift", descreasing = T)
  return(Rules)

}
```

This functions takes the rules as the parameter, checks for the rules that are a subset of other rules, removes them and sorts the remaining rules in the decreasing order of lift.

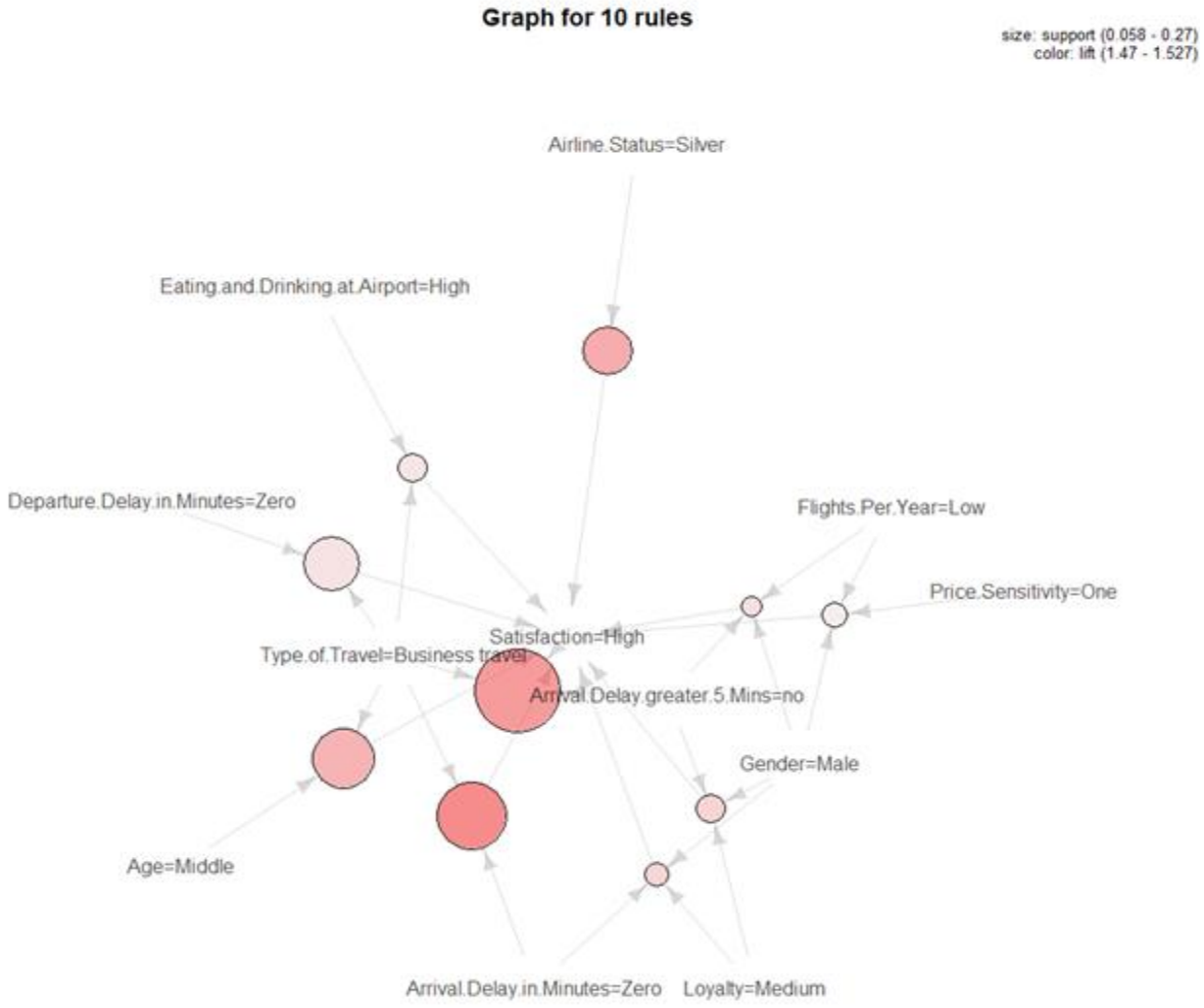
Rules to get predict High Satisfaction

```r
#rules to predict high satisfaction


high_sat_rules<-apriori(cheapseats_trans,parameter = list(support = 0.05, confidence = 0.7,maxlen=6),appearance = list(default="lhs",rhs="Satisfaction=High"))
high_sat_rules<-subset_rules(high_sat_rules)

inspect(high_sat_rules[1:10])
plot(high_sat_rules)
plot(high_sat_rules[1:10],method='graph')
```

We can obtain the rules to get high satisfaction by setting the rhs value in appearance to "Satisfaction=High"
Support was set to 0.05 and confidence to 0.7

lhs		rhs	support	confidence	lift	count
{Type.of.Travel=Business travel,Arrival.Delay.in.Minutes=Zero}	=>	{Satisfaction=High}	0.21789508	0.7620821	1.526913	8594
{Type.of.Travel=Business travel,Arrival.Delay.greater.5.Mins=no}	=>	{Satisfaction=High}	0.27025177	0.7587557	1.520248	10659
{Airline.Status=Silver}	=>	{Satisfaction=High}	0.15111179	0.7545259	1.511773	5960
{Age=Middle,Type.of.Travel=Business travel}	=>	{Satisfaction=High}	0.19236328	0.7532764	1.509270	7587
{Gender=Male,Loyalty=Medium,Arrival.Delay.greater.5.Mins=no}	=>	{Satisfaction=High}	0.08572298	0.7435672	1.489816	3381
{Gender=Male,Loyalty=Medium,Arrival.Delay.in.Minutes=Zero}	=>	{Satisfaction=High}	0.06873558	0.7427397	1.488158	2711
{Gender=Male,Flights.Per.Year=Low,Arrival.Delay.greater.5.Mins=no}	=>	{Satisfaction=High}	0.05785857	0.7389896	1.480645	2282
{Type.of.Travel=Business travel,Departure.Delay.in.Minutes=Zero}	=>	{Satisfaction=High}	0.17213052	0.7376942	1.478049	6789
{Type.of.Travel=Business travel,Eating.and.Drinking.at.Airport=High}	=>	{Satisfaction=High}	0.08693999	0.7359948	1.474644	3429
{Gender=Male,Price.Sensitivity=One,Flights.Per.Year=Low}	=>	{Satisfaction=High}	0.07307117	0.7337067	1.470060	2882

We find that Business Travelers have High satisfaction. Also, if delay in arrival or departure is less than 5 minutes or zero, the satisfaction tends to be high. Also, When Loyalty is medium, it is typically associated with high satisfaction. Also, when airline status is Silver, Satisfaction tends to be high



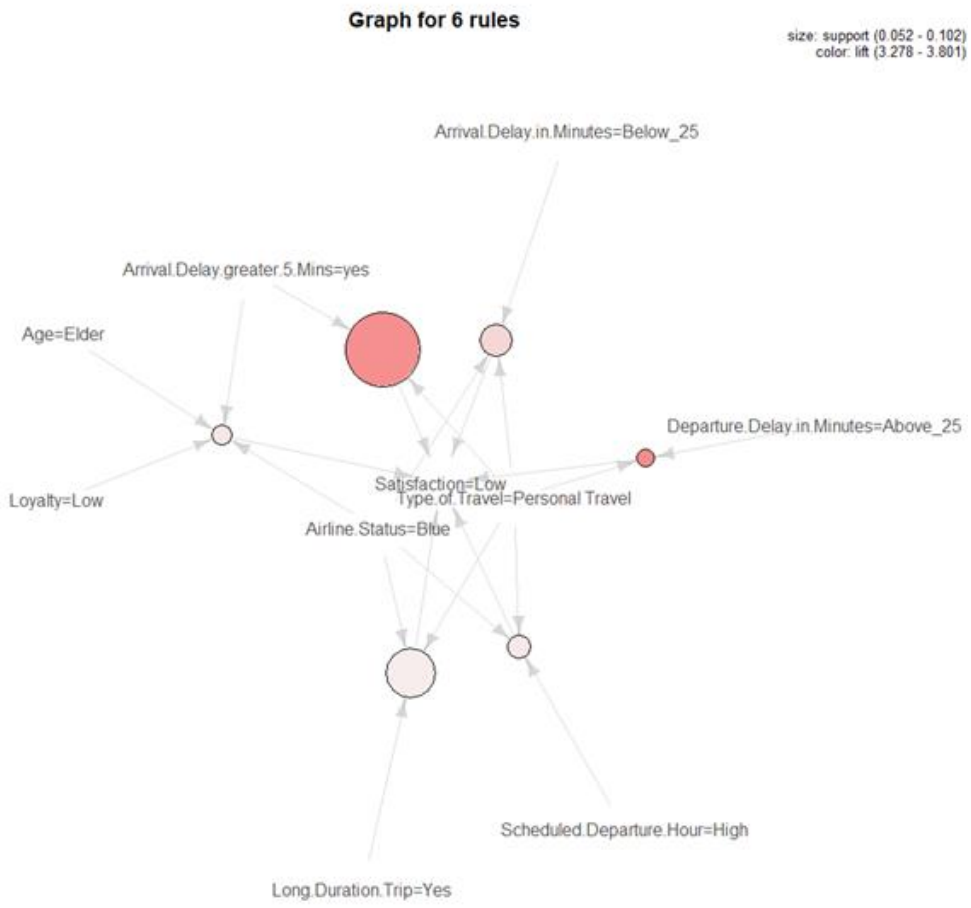
Rules to get Low Satisfaction

```
lowsat_rules<-apriori(cheapseats_trans,parameter = list(support = 0.05, confidence = 0.7,maxlen=6),appearance = list(default="lhs",rhs="Satisfaction=Low"))

lowsat_rules<-subset_rules(lowsat_rules)
inspect(lowsat_rules)
plot(lowsat_rules)
plot(lowsat_rules,method="graph")
```

We can obtain the rules to get high satisfaction by setting the rhs value in appearance to "Satisfaction=Low"
Support was set to 0.05 and confidence to 0.7

```
lhs                                      rhs                  support confidence     lift count
{Type.of.Travel=Personal Travel,
 Departure.Delay.in.Minutes=Above_25} => {Satisfaction=Low} 0.05240739  0.8221957 3.801222  2067
{Type.of.Travel=Personal Travel,
 Arrival.Delay.greater.5.Mins=yes}    => {Satisfaction=Low} 0.10200046  0.8193483 3.788057  4023
{Airline.Status=Blue,
 Type.of.Travel=Personal Travel,
 Arrival.Delay.in.Minutes=Below_25}   => {Satisfaction=Low} 0.06447605  0.7468429 3.452846  2543
{Airline.Status=Blue,
 Age=Elder,
 Loyalty=Low,
 Arrival.Delay.greater.5.Mins=yes}    => {Satisfaction=Low} 0.05443574  0.7166222 3.313128  2147
{Airline.Status=Blue,
 Type.of.Travel=Personal Travel,
 Scheduled.Departure.Hour=High}       => {Satisfaction=Low} 0.05707259  0.7141497 3.301697  2251
{Airline.Status=Blue,
 Type.of.Travel=Personal Travel,
 Long.Duration.Trip=Yes}              => {Satisfaction=Low} 0.08006896  0.7090256 3.278007  3158
```

It is apparent that when the type of travel is personal travel, the satisfaction tends to be low.
Also, Airline Status Blue is a good indicator of low satisfaction.
When the Age is Elder, typically has resulted in low satisfaction.



---

iii. Support Vector Machine Model

For the predictive Analytics part, we split the cheapseats data into training and testing sets. Did the splitting using CreateDataPartition in caret

```
index<-createDataPartition(cheapseats1$Satisfaction,p=0.7,list=FALSE)
index<-sample(index)

cs_train<-cheapseats1[index,]
cs_test<-cheapseats1[-index,]
```

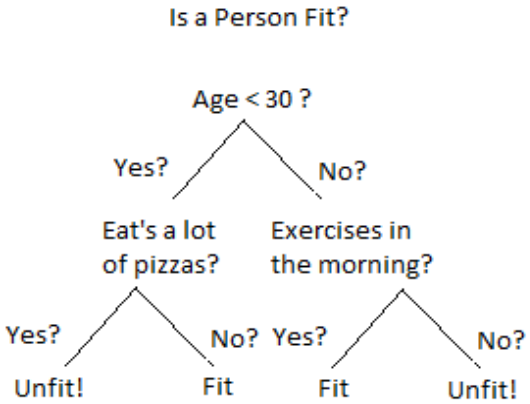
We used 70% of the data points for training and 30% for testing. For the model we used caret library which is one of the most widely used machine learning packages in R.

```
model_svm_rbf <- train(Satisfaction ~ ., data = cs_train,
                       preProcess = c("center", "scale"),
                       tuneGrid = expand.grid(sigma = seq(0, 1, 0.1),
                                              C = seq(0, 1, 0.1)),
                       method = "svmRadial",
                       trControl = trainControl(method = "cv",
                                                number = 3),allowParallel=TRUE
                       )
```

This SVM model uses RBF method to train the model to predict the Satisfaction with the remaining variables. Some variables like origin city,state , destination city, state and date etc were removed as these had many levels within them. The preprocess parameter is used to tell caret to preprocess the data. In this case, all the numeric variables will be centered and scaled which will ensure standardization amongst variables. The Tunegrid basically is the values of the hyper parameters we want caret to consider. For the C(cost) and Sigma(RBF) caret will start at 0.1 and step by 0.1 till it reached 1. It will test all the combinations of these parameters and choose the set of parameters that give the best validation accuracy. Final values of the parameters chosen were sigma=0.4 and C=1. The traincontrol parameter helps us to choose the cross validation technique for model evaluation. The default is bootstrap validation with 25 repeats. In this case, we have chosen 3 fold cross validation. Allow Parallel argument just tells caret to allow parallel processing using all cores of the machine, this has to be used in combination with the doParallel library. The testing accuracy achieved by the SVM model was 82.79%.

iv. Random Forest Model

Random Forest is a bagging technique within the family of ensemble models. It trains a bunch of decision tree models and get the predictions through majority voting. A decision tree model is a data mining technique that tries to mimic human decision making process by splitting the decision space at each level according to the best condition. Here is a visual representation of how a decision tree model works.



In addition to the predictive analytics, the random forest model also ranks the variable in the order of their relative importance. This feature of the random forest model is important to us as it will allow us to validate our findings from other techniques and form conclusions.

For our problem, we used the caret library which has a very good implementation of the random forest algorithm.

```
model_rf<- train(Satisfaction ~ ., data = cs_train,
                 preProcess = c("center", "scale"),
                 method = "rf",
                 trControl = trainControl(method = "cv",
                                          number = 3),allowParallel=TRUE
)
```

We have used similar approach of splitting our original cheapseats dataset into training and testing sets. We used the training set for model building and the testing set for model evaluation. The preprocess parameter is used to tell caret to preprocess the data. In this case, all the numeric variables will be centered and scaled which will ensure standardization amongst variables. Setting the method to 'rf' will tell caret to use the random forest algorithm for model building. Again, we have used a 3-fold cross validation technique. This is important as we should have similar cross validation techniques to compare the models. The testing accuracy achieved by the random forest model was 90.18% which is a very good result.

Variable importance:-

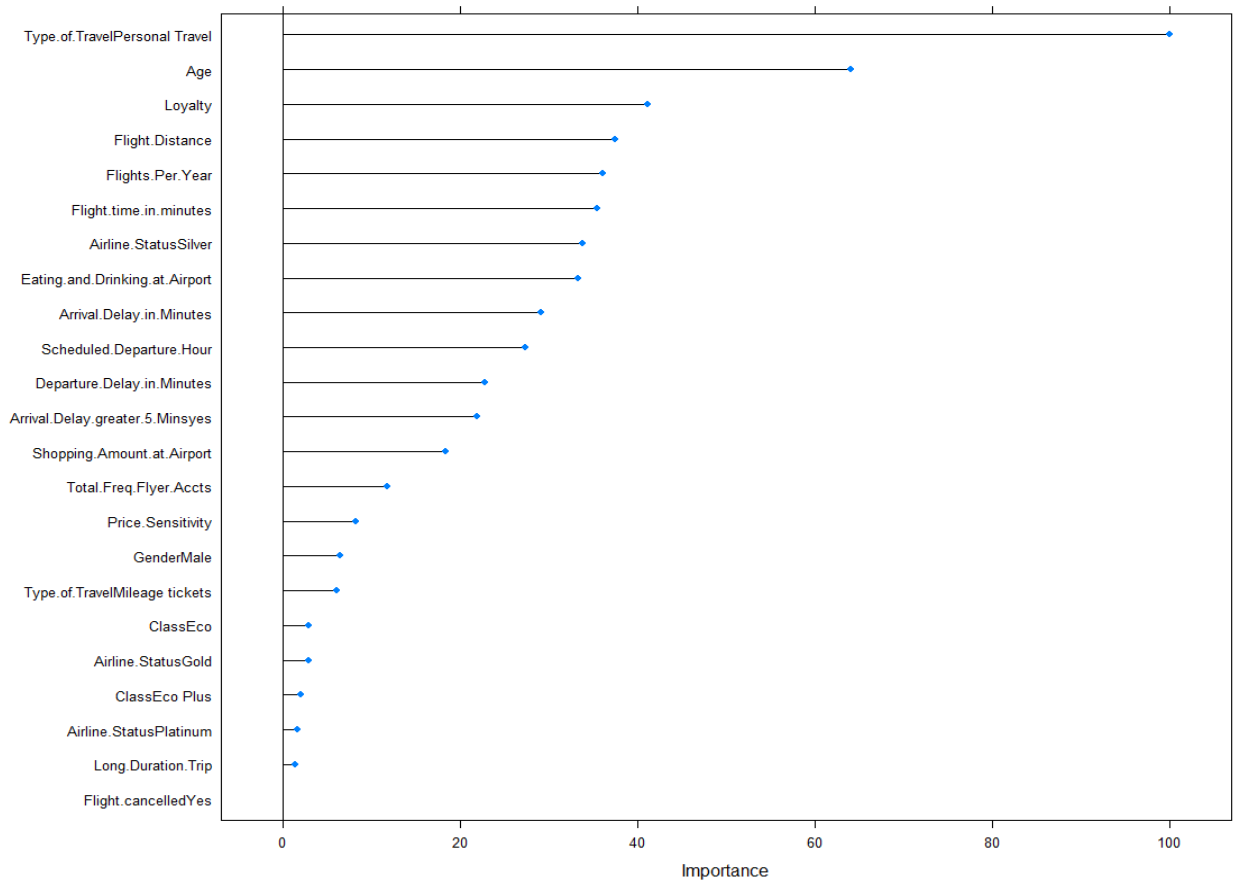
Caret package has a predefined function 'varImp' for getting the variable importance rankings . The input parameter to this function is the trained random Forest model.

varImp(model_rf)

This function gives us the following output.

```
Type.of.TravelPersonal Travel   100.000
Age                              64.038
Loyalty                          41.109
Flight.Distance                  37.464
Flights.Per.Year                 36.056
Flight.time.in.minutes           35.415
Airline.StatusSilver             33.759
Eating.and.Drinking.at.Airport   33.321
Arrival.Delay.in.Minutes         29.098
Scheduled.Departure.Hour         27.328
Departure.Delay.in.Minutes       22.849
Arrival.Delay.greater.5.Minsyes  21.907
Shopping.Amount.at.Airport       18.401
Total.Freq.Flyer.Accts           11.837
Price.Sensitivity                 8.288
GenderMale                        6.404
Type.of.TravelMileage tickets     6.082
ClassEco                          2.954
Airline.StatusGold                2.886
ClassEco Plus                     2.008
```

We can also plot these results for better intuition using plot()



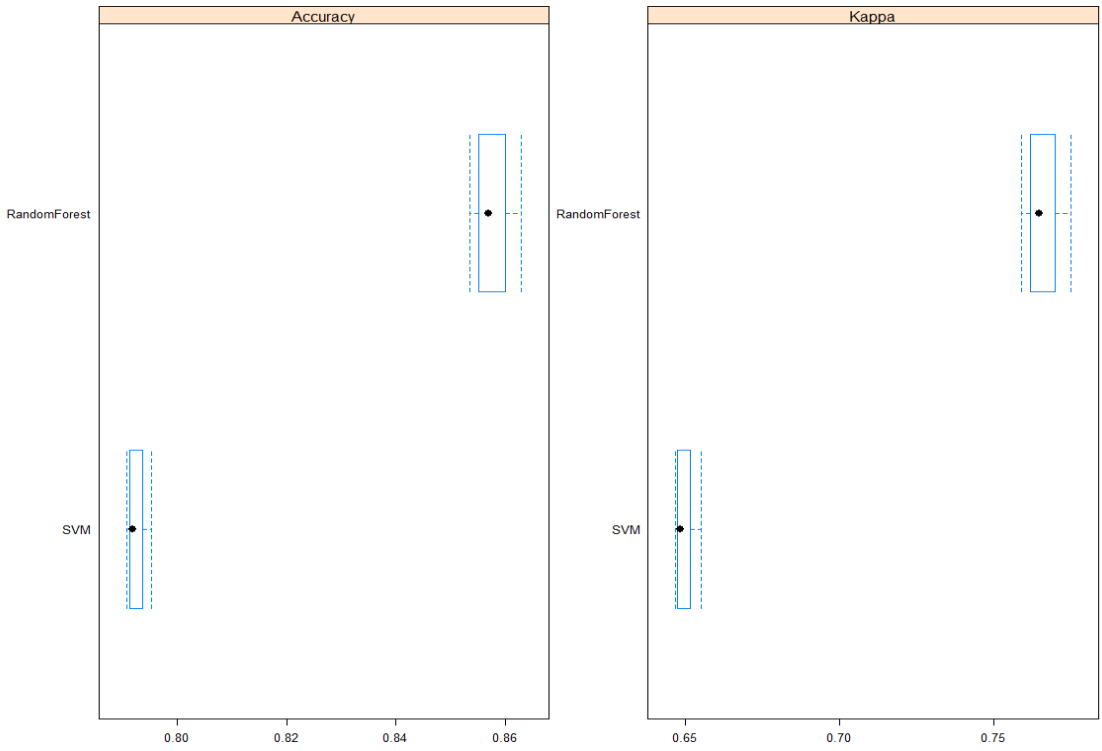
From this it is evident that, Personal travel is the one of the important variable in determining satisfaction. Age is also an important variable, so is Loyalty, Airline status and the delays. These results are consistent with our findings from the association rules and different visual charts.

Model Comparison:

We compared the two models using resamples function from caret and plotted the results.

```
scales <- list(x = list(relation = "free"),
               y = list(relation = "free"))


model_comparison <-resamples(list(SVM = model_svm,RandomForest=model_rf))
summary(model_comparison)
bwplot(model_comparison,scales=scales)
```

This will give us boxplots to compare the two models visually based on accuracy and kappa metrics.



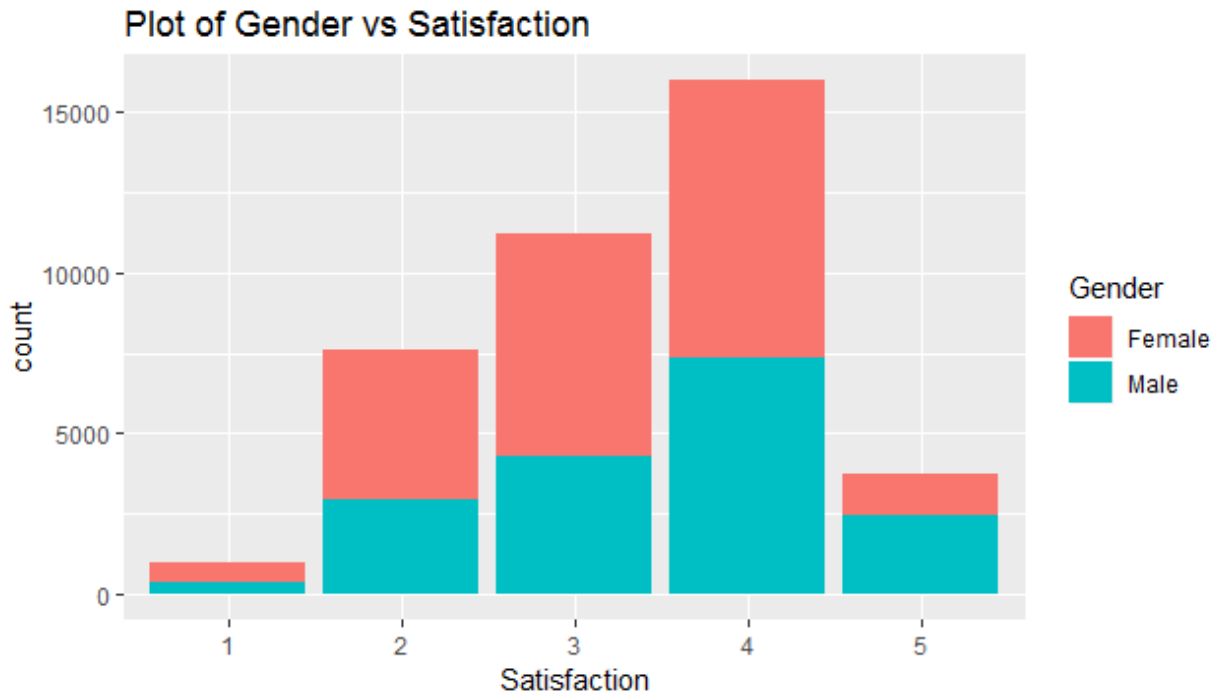
From these plots, it is evident that the random forest model is more accurate than the SVM model.But a point of note that the SVM model is more stable than the random forest model which is evident from the range of cross validation accuracies. The SVM model has a less range of variance of accuracies than the random forest model and therefore is more stable. But the difference of the difference of accuracies in the two models is too large and hence the random forest model is our best pick for predictive analytics. This model can be used to predict customer satisfaction going forward.

6.) Visualization

a. Mapping Gender vs Satisfaction

Code Snippet

```
by_gen<-cheapseats%>%group_by(Satisfaction,Gender)%>%summarise(count=n())
ggplot(data = by_gen,aes(x=Satisfaction,y=count,fill=Gender))+geom_bar(stat
='identity')+ggtitle("Plot of Gender vs Satisfaction")
```



It can be analyzed from the above graph that as compared to males, the number of females are less satisfied.

b. Plot of Type of Travel vs Satisfaction

Code Snippet

```
by_type<-cheapseats%>%group_by(Satisfaction,Type.of.Travel)%>%summarise(count=n())
ggplot(data = by_type,aes(x=Satisfaction,y=count,fill=Type.of.Travel))+geom_bar(stat
='identity')+ggtitle("Plot of Type of Travel vs Satisfaction for CheapSeats")
```

It can be analyzed from the above graph that the people travelling by personal travel are the least happy.

c. Plot of Gender vs Satisfaction for personal travel

Code Snippet

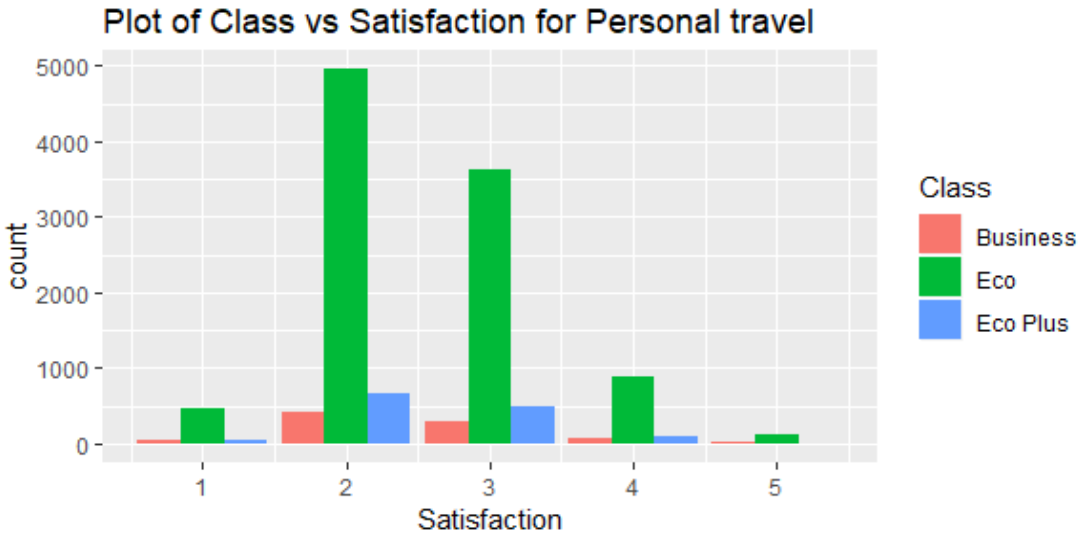
```
myplot <- ggplot(cheapseats_personal,aes(x=Satisfaction,fill=Gender)) + geom_bar(position= 'dodge')
myplot
```

From the above graph, we understand that the number of females travelling by personal travel are the least happy.

d. Class vs Satisfaction or Personal Travel

Code Snippet

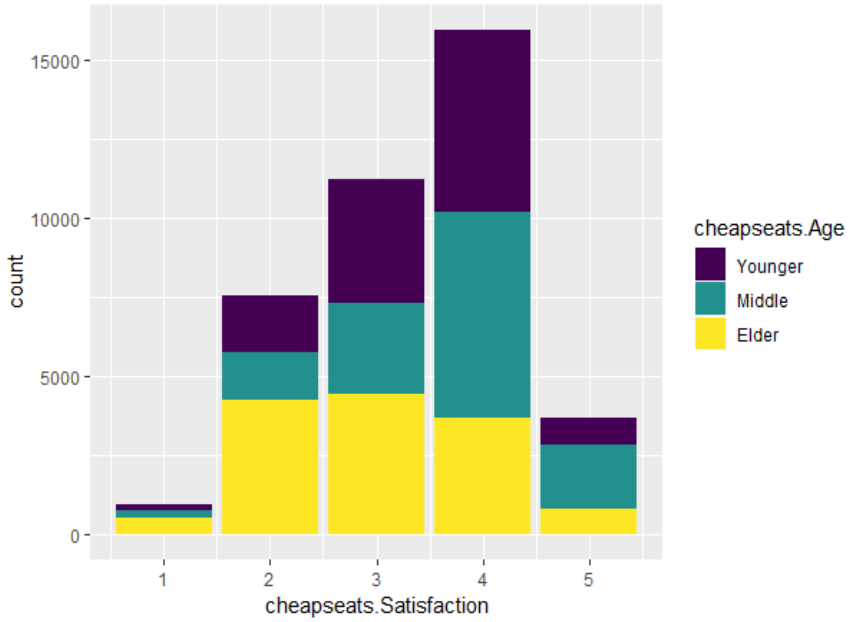
```
myplot11 <- ggplot(cheapseats_personal,aes(x=Satisfaction,fill=Class)) + geom_bar(position=
'dodge') + ggtitle("Plot of Class vs Satisfaction for Personal travel")
myplot11
```



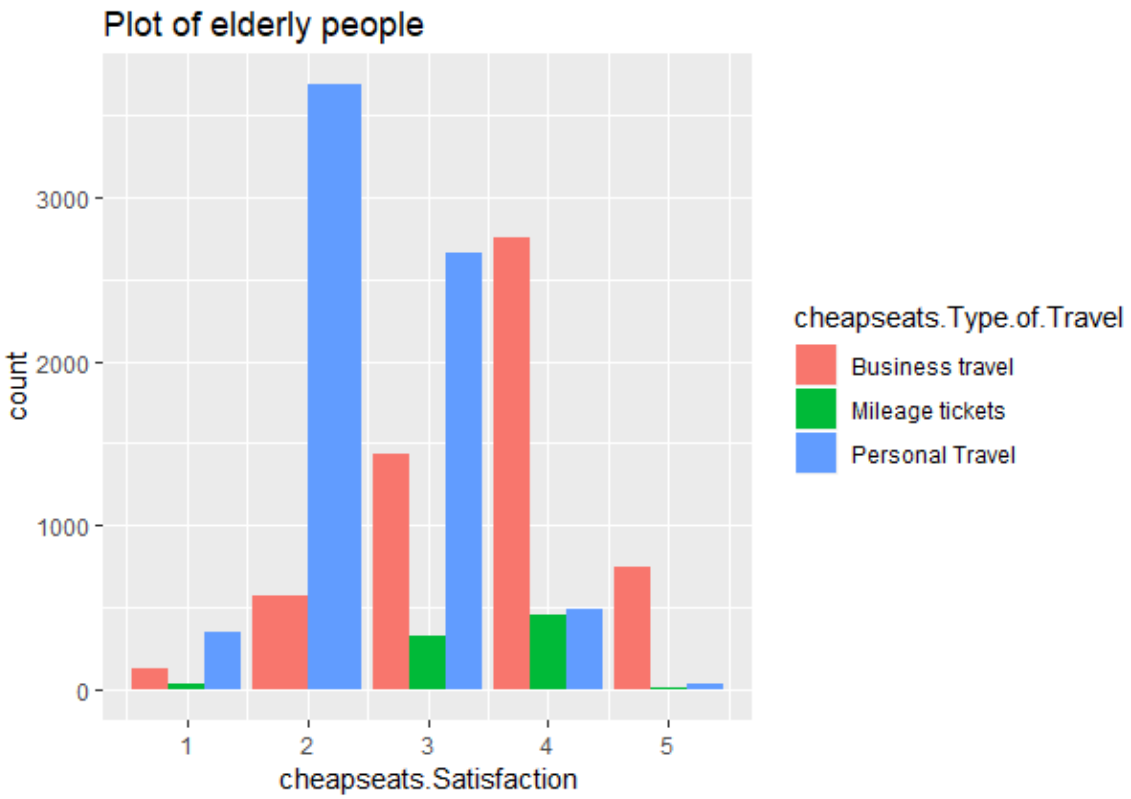
e. Age vs Satisfaction

Code Snippet

```
myplot1 <- ggplot(Age, aes(x=cheapseats.Satisfaction, fill=cheapseats.Age)) + geom_bar()
myplot1
```



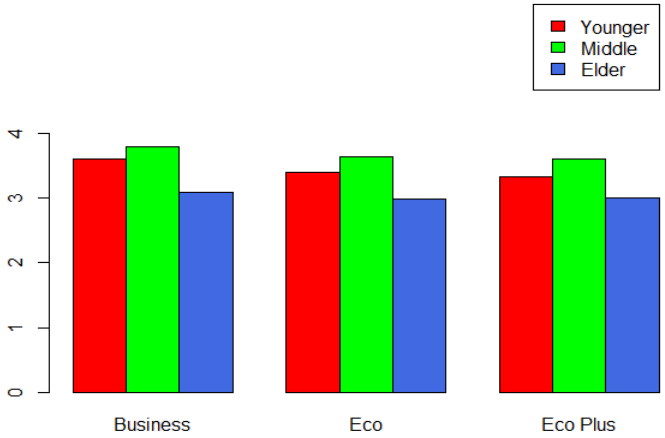
f. Plot of elderly people travelling by Personal Travel



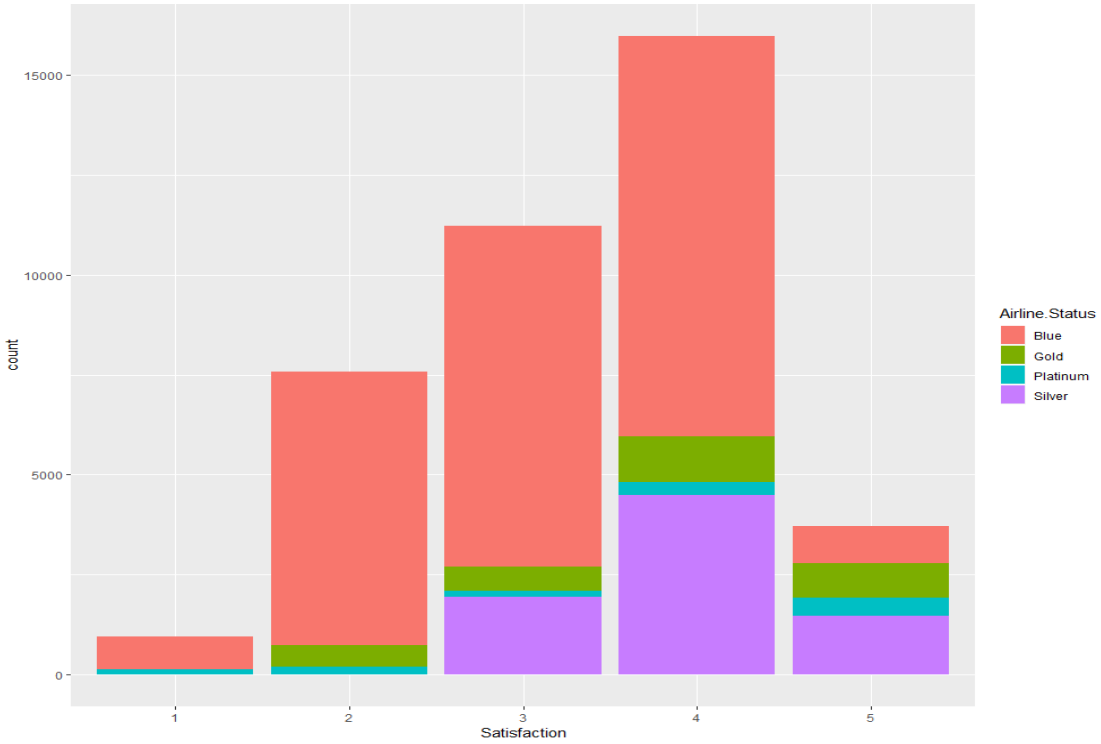
g. Satisfaction vs Classes for different age groups

Code Snippet

```
barplot(a,  beside=T,  legend.text = T, col = c("red","green","royalblue"),ylim = c(0,max(a)+1),
args.legend=list(x=12, y =6))
text(1.5, 4, labels = round(a[1,1],2))
```



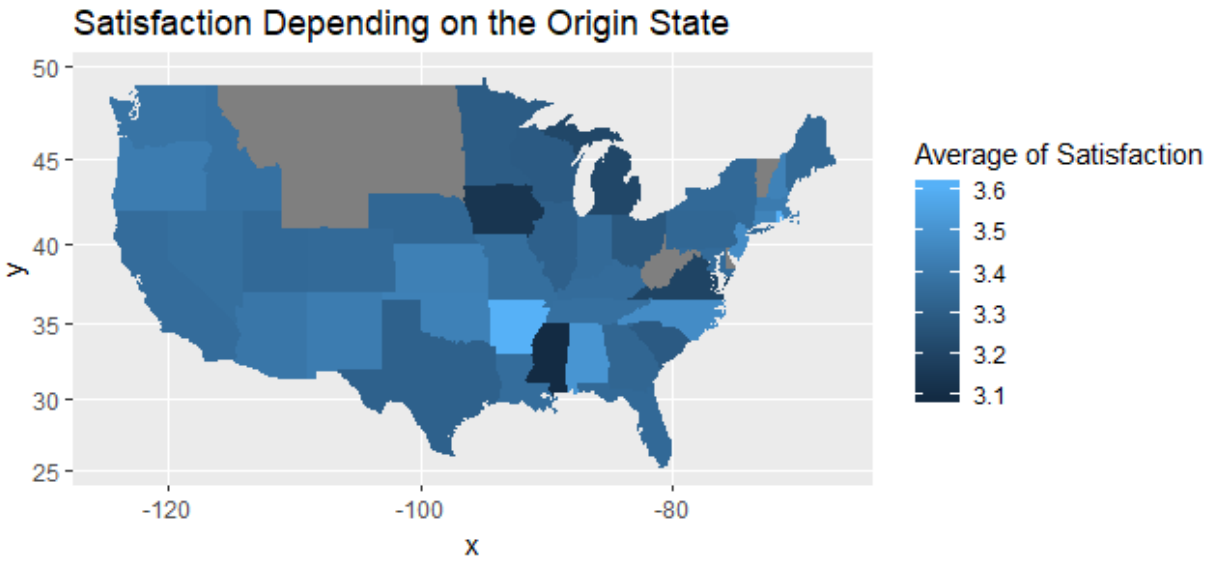
h. Satisfaction vs Classes for different age groups



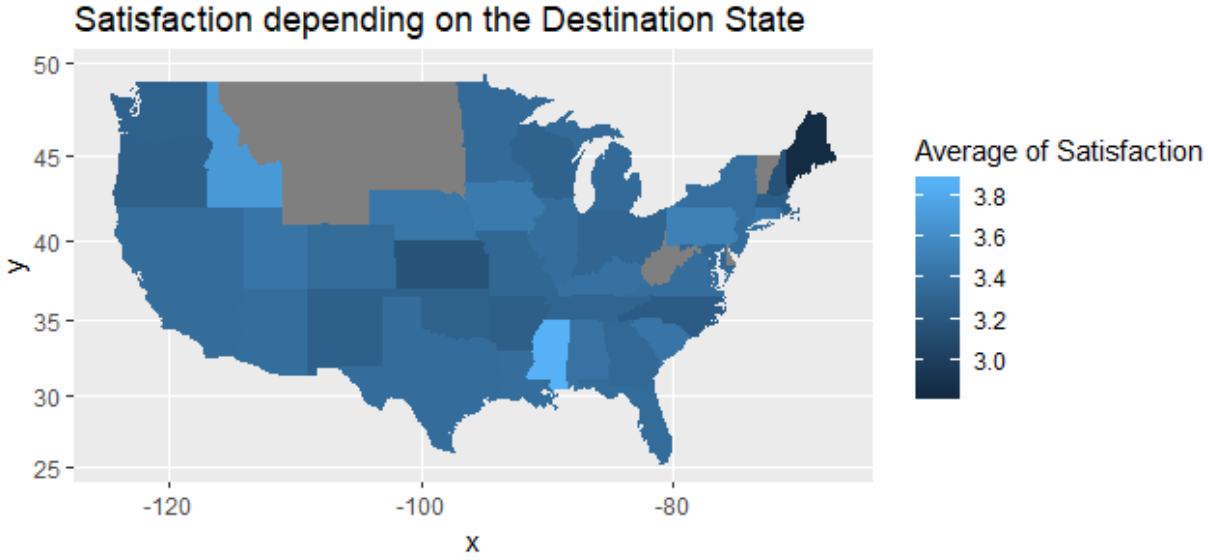
i. Map of the area where the satisfaction is the lowest depending on the origin state of the flight

Code Snippet

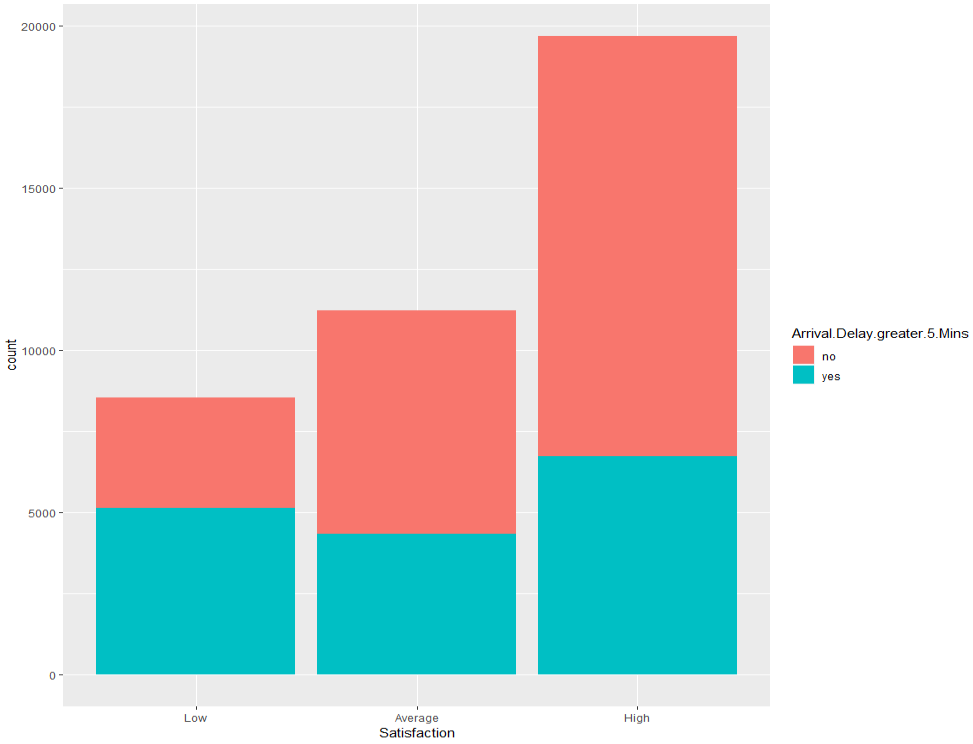
```
mapArea <- ggplot(mergeDF, aes(map_id = stateNames)) #create a ggplot specifying the dataframe and map we want to use
mapArea
mapArea <- mapArea + geom_map(map=us,aes(fill=mergeDF$b1))  #uses the US map and fills the area
mapArea
mapArea <-mapArea + expand_limits(x = us$long, y = us$lat) #sets the x and y for map
mapArea
mapArea <- mapArea + coord_map() + ggtitle("Origin State Map Analysis")+ labs(fill = "Average of Satisfaction") #gives title
mapArea
```



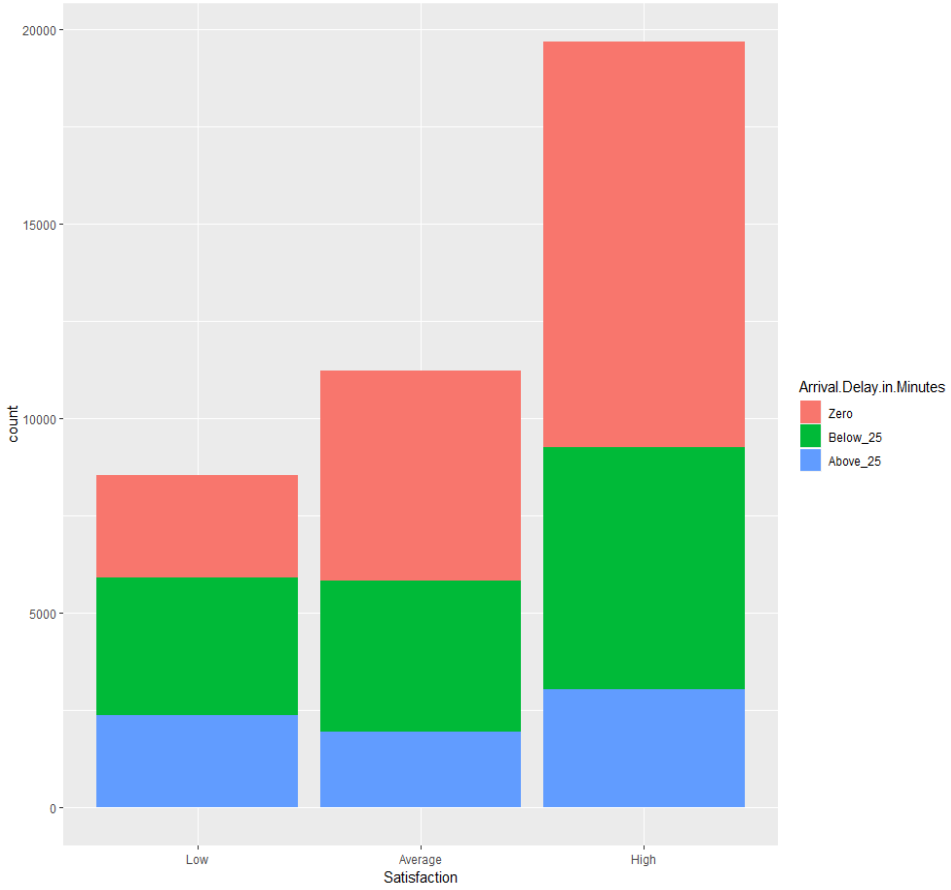
j. Map of the area where the satisfaction is the lowest depending on the origin state of the flight



k. Arrival delay greater than 5 minutes versus satisfaction



l. Arrival delay in minutes versus satisfaction



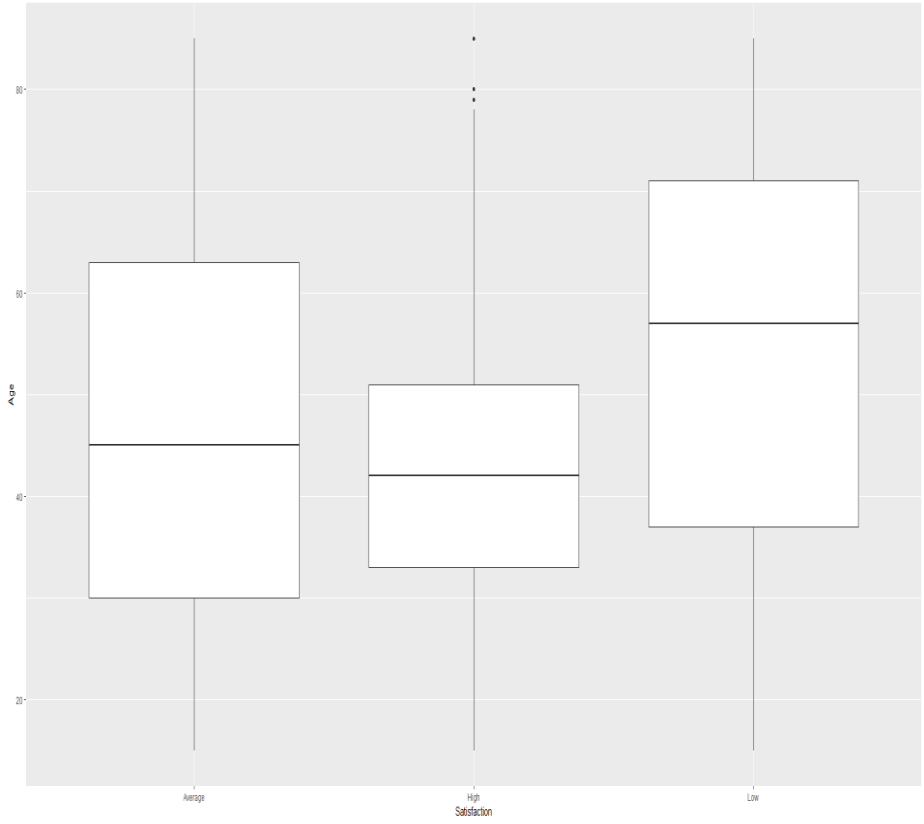
7.) Actionable Insights:

Insights based on Type of Travel:

- As per the association rules, we can see that personal travel within the type of travels has been appearing in all the top rules that are associated with low satisfaction
- Also, from the variable importance plot that we derived from the Random Forest Model, Personal travel appears on the top
- Therefore, personal travel is one of the most important factors for low satisfaction among customers
- Conversely, people travelling by business class are the most happy as per the association rules as well the barchart

Insights based on Age:

- The barchart of Age vs Satisfaction shows that majority of the elderly people are unhappy
- The barchart of Elderly people vs Type of Travel shows that the Elderly people travelling by Personal Class are least happy
- The barchart of Elderly people vs the Class in which they are travelling shows that the Elderly people travelling by Economy class are least happy
- The Boxplot below validates our findings as it shows that the median age for low satisfaction is higher than the median age for average and high satisfaction



Insights based on Gender:

- The barchart of Gender vs Satisfaction gave insights regarding the females being Unhappy as compared to the males
- The barchart of Type of travel vs Satisfaction shows that people travelling by personal travel are the least happy
- Among the people travelling by personal travel, majority of the females are unhappy
- Validation:

	Female	Male
Average	6965	4260
High	9847	9838
Low	5264	3267

Insights based on Loyalty:

- The association rules show that when the loyalty is low, along with age equals elder and airline status blue, the satisfaction tends to be low
- This finding is further validated by the variable importance plot

Insights based on Departure and Arrival Delays:

- From our association rule analysis, we found that for higher satisfaction, the arrival time delay and the departure time delay is either 0 or less than 5 minutes
- Conversely, when arrival and departure delay is above 5 minutes and type of travel is personal travel, satisfaction tends to be low
- This is further validated from our variable importance plot

Insights based on Airline Status:

- From our association rules as well as the barcharts, when our airline status is blue, the satisfaction tends to be on the lower side and when the airline status is silver satisfaction tends to be high
- This is further validated from our variable importance plot

8.) Recommendations

1. Vouchers or discount coupons for female travelers in personal travels.
2. Surprise Upgrades for females travelling by Economy Class.
3. Seat reservations for elderly people travelling in personal class near the aisle and restroom.
4. Separate queues and airport escorts for Elderly People.
5. Extra Loyalty points for people travelling by personal travel.
6. Upgrade facilities for Blue status customers to Silver status.
7. Free lounge/ meal for customers who have delayed flight.