STAT 515 FINAL PROJECT
ANALYSIS OF CHOCOLATE BAR RATINGS
Swathi Pollishetty
G01242838

1 .Project description

Chocolate is one of the most popular candies in the world.Every second, Americans collectively eat 100 pounds of chocolate[1].This analysis will bring some insight into consumer patterns in the chocolate industry. It will be possible to see a pattern in the data that could lead us to observe the factors and how they will affect consumer rating.

2 .Data set

2.1 Source

The dataset has been obtained from Kaggle Repository.
Source: https://www.kaggle.com/rtatman/chocolate-bar-ratings

2.2 Context

Each chocolate is evaluated from a combination of both objective qualities and subjective interpretation. A rating here only represents an experience with one bar from one batch.[2]
The Rating Scale is described as follow :
4.0 - 5.0 = Outstanding
3.5 - 3.9 = Highly Recommended
3.0 - 3.49 = Recommended
2.0 - 2.9 = Disappointing
1.0 - 1.9 = Unpleasant
The components that contribute to the Cacao Flavour system are Flavour,Texture,Aftermelt and the Overall opinion.

2.3 Data inspection

The dataset comprises 1795 rows and 9 columns.
The 9 features include:

Feature name	Description
Company (Maker-if known)	Name of the company manufacturing the bar.
Specific Bean Origin or Bar Name	The specific geo-region of origin for the bar.
REF	A value linked to when the review was entered in the database; Higher = more recent.
Review Date	Year of publication of the review.

[1] "Every Second, Americans Collectively Eat 100 Pounds Of ...." 27 Oct. 2018, https://southfloridareporter.com/every-second-americans-collectively-eat-100-pounds-of-chocolate/. Accessed 24 Nov. 2020.
[2] "Review Guide - Flavors of Cacao." http://flavorsofcacao.com/review_guide.html. Accessed 24 Nov. 2020.

Cocoa Percent	Cocoa percentage (darkness) of the chocolate bar being reviewed.
Company Location	Manufacturer base country.
Rating	Expert rating for the bar.
Bean Type	The variety (breed) of bean used, if provided
Broad Bean Origin	The broad geo-region of origin for the bean.

2.4 Preparation of analysis data set

The data set obtained needs to be preprocessed for further analysis.

Initially the column names were changed for easier legibility.The new names do not have spaces and line breaks between the columns and are simple.

I have dropped the Bean Type and Broad Bean Origin columns as they have more missing values and do not contribute much to our analysis.

Further, the '%' symbol in the Cocoa Percentage column has been removed and the variable has been converted to a numeric type.

Maker variable has been converted to a factor variable.

Added a variable RatingLevel which equals "High" if Rating >3.5 and "Low" if Rating <= 3.5.

```
> #displaying top 6 rows data
> head(flavours)
     Maker  BeanOrigin  REF ReviewDate CocoaPercentage CompanyLocation Rating
1 A. Morin Agua Grande 1876       2016              63          France   3.75
2 A. Morin       Kpime 1676       2015              70          France   2.75
3 A. Morin      Atsane 1676       2015              70          France   3.00
4 A. Morin       Akata 1680       2015              70          France   3.50
5 A. Morin      Quilla 1704       2015              70          France   3.50
6 A. Morin    Carenero 1315       2014              70          France   2.75
```

Figure 1. Displaying top 6 rows of the flavours dataset

```
> #checking summary of the data
> summary(flavours)
    Maker           BeanOrigin            REF         ReviewDate   CocoaPercentage CompanyLocation
 Length:1795        Length:1795        Min.   :   5   Min.   :2006   Min.   : 42.0   Length:1795
 Class :character   Class :character   1st Qu.: 576   1st Qu.:2010   1st Qu.: 70.0   Class :character
 Mode  :character   Mode  :character   Median :1069   Median :2013   Median : 70.0   Mode  :character
                                       Mean   :1036   Mean   :2012   Mean   : 71.7
                                       3rd Qu.:1502   3rd Qu.:2015   3rd Qu.: 75.0
                                       Max.   :1952   Max.   :2017   Max.   :100.0

     Rating
 Min.   :1.000
 1st Qu.:2.875
 Median :3.250
 Mean   :3.186
 3rd Qu.:3.500
 Max.   :5.000
```

Figure 2.Displaying summary of the flavours dataset

3. Exploratory analysis and research questions

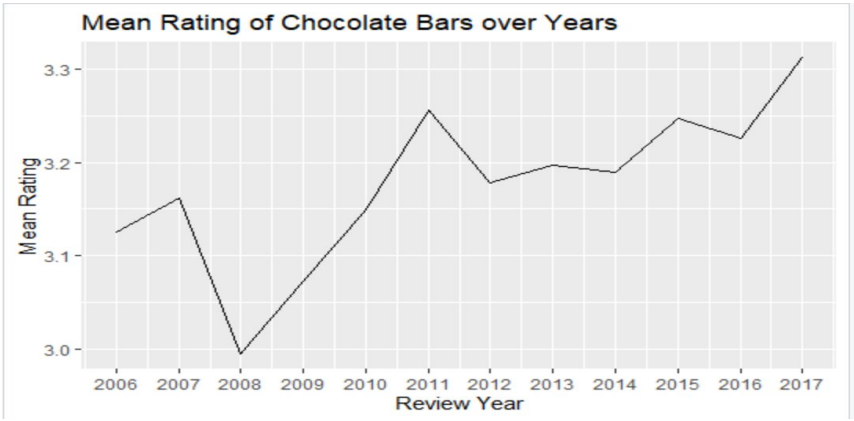

Figure 3.Plot for Mean Rating of Chocolate Bars over Years

From the above visualisation,we can see that the Average Rating of the chocolate bars has increased with time.

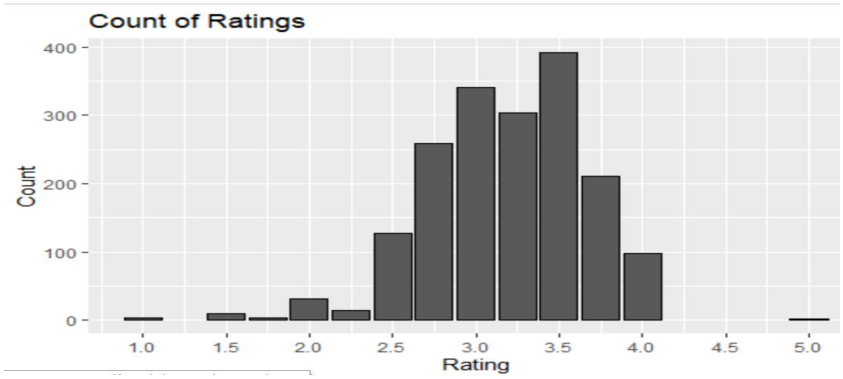

Figure 4.Bar chart showing Count of Ratings

From the above visualisation,we can see that the count of ratings is maximum when the rating is 3.5.

Tableau Visualisations:

After initial preprocessing, I have written the data frame data to a csv file 'finalflavours.csv' and taken it as an input to Tableau for constructing few visualisations.

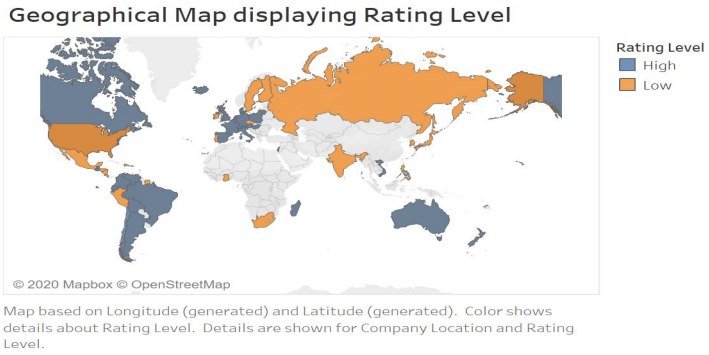

Figure 5.Tableau Visualisation showing Rating Level

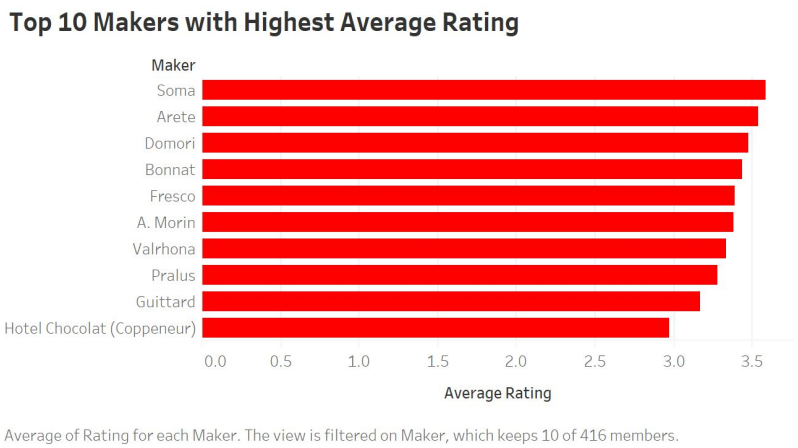

Figure 6. Tableau Visualisation showing Top 10 Makers as per Average Rating

Research Questions:

1. What are the most important variables that contribute to the Chocolate Bar Rating?
2. Are the chocolate ratings getting better with time?
3. How does the CocoaPercentage affect the rating of the chocolate?

4. Data analysis

4.1. Methods and software used

I have used R studio and Tableau for my Data Analysis.

4.2. Results

Linear Regression Model:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.[3]

I have reproduced two models to predict Rating using ReviewDate and CocoaPercentage as the predictors.

For Model 1, a linear regression fit has been done with the formula in the figure below.

```
> sm1

Call:
lm(formula = Rating ~ ReviewDate, data = flavours)

Residuals:
     Min       1Q   Median       3Q      Max
-2.11540 -0.29747  0.03676  0.31937  1.91721

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.628257   7.722599  -3.837 0.000129 ***
ReviewDate    0.016307   0.003838   4.249 2.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4758 on 1793 degrees of freedom
Multiple R-squared:  0.009969,  Adjusted R-squared:  0.009417
F-statistic: 18.05 on 1 and 1793 DF,  p-value: 2.256e-05
```

Figure 7.Displaying summary of Linear Regression Model 1

[3] "Linear Regression." http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm. Accessed 25 Nov. 2020.

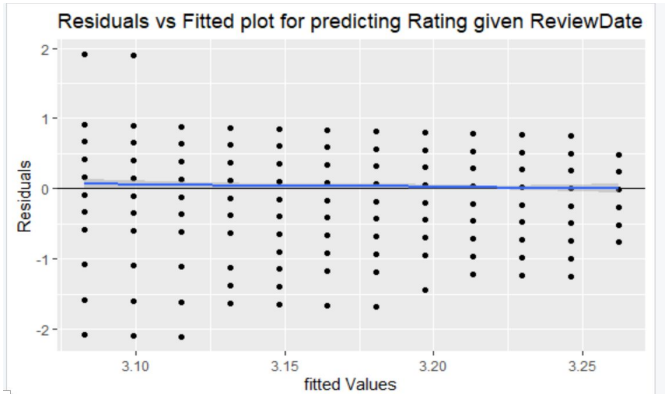

Figure 8.Residual vs Fitted Plot for Rating,ReviewDate Linear Model

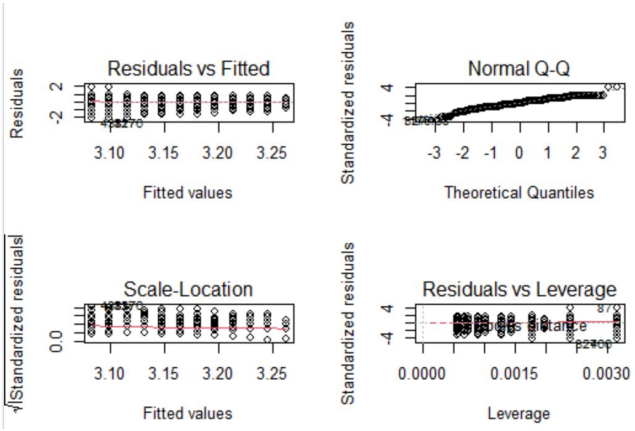

Figure 9. Residual Plots for Model 1

From the model summary, we can see that the Adjusted R-square value is very less;however the p-values are very less which suggests that the coefficients are significant which means an increase in 0.016 units of ReviewDate increases the Rating by one unit.
The grey region of Residual vs Fitted plot has Y=0 axis and indicates linearity.
The Normal Q-Q plot indicates normality and the Scale Location plot indicates homoscedasticity.
The RMSE value for Model 1 is 0.2261392.
For model 2, the predictor variable is CocoaPercentage and formula is as below.

```
> sm2

Call:
lm(formula = Rating ~ CocoaPercentage, data = flavours)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2071 -0.3196  0.0429  0.3178  1.7929

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.079388   0.126757  32.183  < 2e-16 ***
CocoaPercentage -0.012461   0.001761  -7.076 2.12e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4717 on 1793 degrees of freedom
Multiple R-squared:  0.02717,   Adjusted R-squared:  0.02662
F-statistic: 50.07 on 1 and 1793 DF,  p-value: 2.122e-12
```

Figure 10.Displaying summary of Linear Regression Model 2

From the model summary, we can see that the Adjusted R-square value is less;however the p-values are very less which suggests that the coefficients are significant which means an increase in 0.012 units of CocoaPercentage decreases the Rating by one unit.

The grey region of the Residuals vs Fitted plot doesn't include Y=0 axis which suggests there is departure from linearity.
The RMSE for Model 2 is 0.2222112.

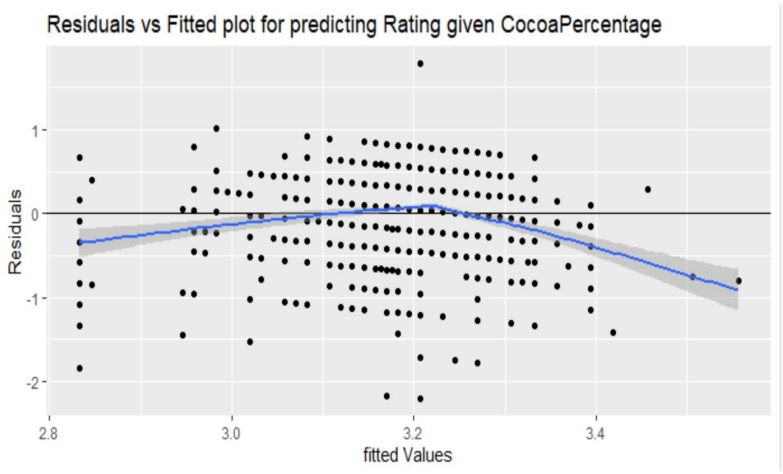

Figure 11.Residual vs Fitted Plot for Rating,CocoaPercentage Linear Model

So we further try to fit a linear regression model by including a polynomial of CocoaPercentage which shows a slight increase in Adjusted R-squared value which suggests a better model.
The grey region of the Residual vs Fitted plot includes the Y=0 axis which suggests linearity.The Normal Q-Q plot indicates normality and the Scale location plot is pipe shaped which indicates homoscedasticity.
The RMSE value for Model 3 is 0.2115356 which makes it better than Model 2.

```
> sm3

Call:
lm(formula = Rating ~ CocoaPercentage + I(CocoaPercentage^2),
    data = flavours)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23467 -0.23467  0.02292  0.28388  1.76533

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.9034092  0.6411626  -2.969  0.00303 **
CocoaPercentage      0.1484531  0.0170079   8.729  < 2e-16 ***
I(CocoaPercentage^2) -0.0010722  0.0001127  -9.510  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4603 on 1792 degrees of freedom
Multiple R-squared:  0.0739,    Adjusted R-squared:  0.07287
F-statistic:  71.5 on 2 and 1792 DF,  p-value: < 2.2e-16
```

Figure 12.Displaying summary of Linear Regression Model 3

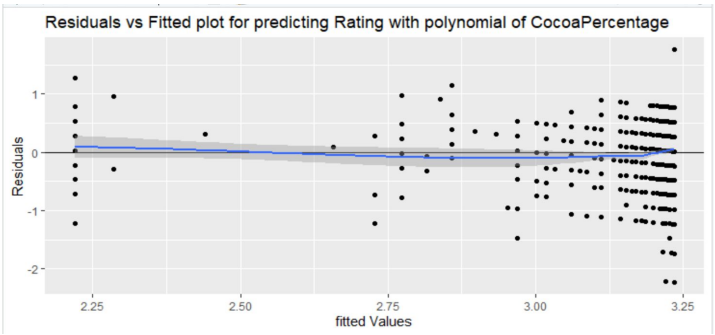

Figure 13.Residual vs Fitted Plot for Rating,CocoaPercentage Linear Model

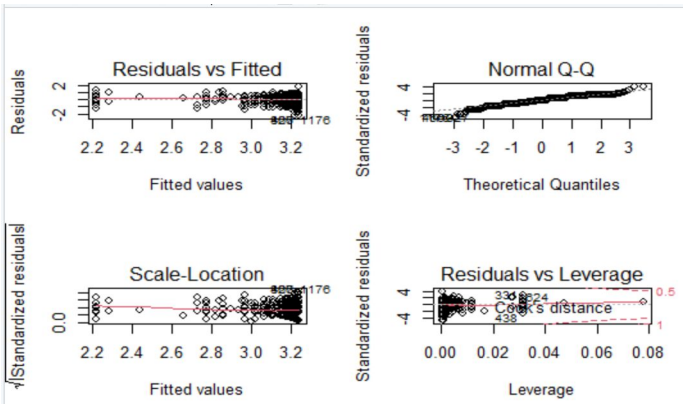

Figure 14.Residual Plots for Model 3

Random Forest Model:

Decision trees are prone to overfitting, and many times the resulting tree is overcrowded. This is usually due to too many variables taking up nodes. As such, much like regression, decision trees need to be 'pruned' to best fit only the important variables. However, deciding the best number of nodes and variables can be difficult. The R package Random Forests best addresses this issue. Random Forests take a set of data and break it up using a user-defined number of nodes a total of a user-defined number of times. Afterwards, a summary of the random forest can be used to determine the variables with the greatest effect on the dependent variable.

For establishing a Random Forest Model, the variable to be predicted is Rating and the predictor variables are BeanOrigin,REF,ReviewDate,CocoaPercentage and CompanyLocation.The variable Importance plot below gives information

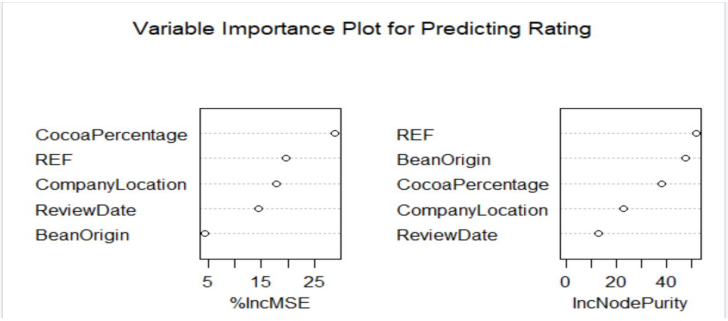

Figure 15.Random Forest evaluation of variables

```
Call:
 randomForest(formula = Rating ~ ., data = flavours_rf, mtry = 2,      importance = TRUE, ntr
ees = 500, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.2045677
                    % Var explained: 7.07
```

Figure 16.Summary of Random Forest Call

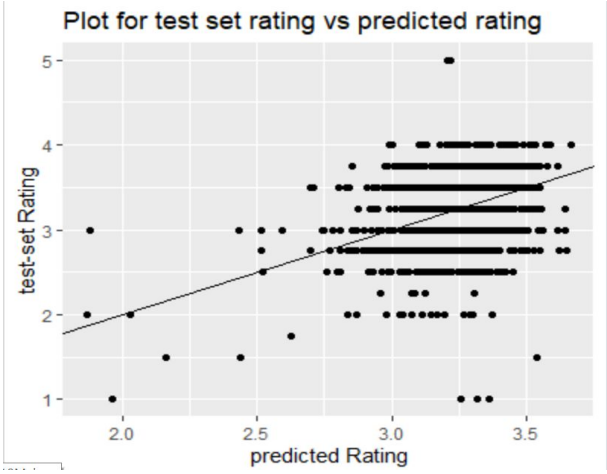

Figure 17. Random Forest Model Test set vs Predicted Rating Plot

The test set RMSE value for Random Forest Model is 0.2185883.

```
> #RMSE for Random Forest model
> mean((yhat.bag-flavours_rf.test)^2)
[1] 0.2185883
```

K-means Clustering

The K-means algorithm identifies the k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.[4]

For my analysis,I have performed K-means clustering on Rating,ReviewDate,REF and CocoaPercentage. The optimal number of clusters has been obtained using the silhouette method and is equal to 2.Value of nstart has been assigned to 30.

The results of the K-means clustering are in the figure below:

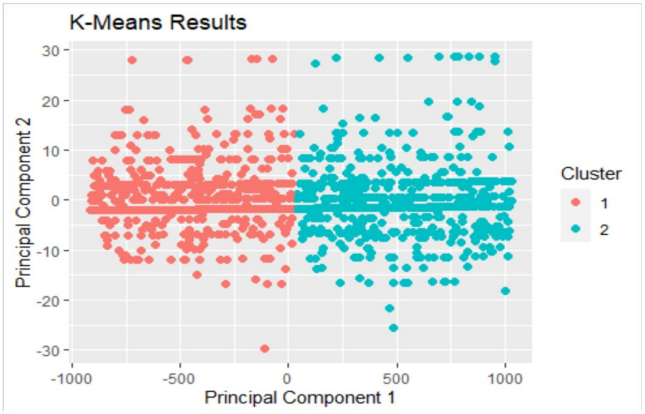

Figure 18. Plot of K-means Results

[4] "Understanding K-means Clustering in Machine Learning | by ...." 12 Sep. 2018, https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. Accessed 29 Nov. 2020.

A better illustration of the clusters is obtained using the fviz_cluster as below:

Figure 19. Fviz cluster visualisation for k-means

5. Conclusions

While the Linear Regression model aims to find relationships between two variables, the Random Forest model tries to fit all the variables.
From the results we can draw conclusions about the research questions.

1. What are the most important variables that contribute to the Chocolate Bar Rating?
From the Random Forest model,we can see that CocoaPercentage is an important variable with respect to percentage increase in MSE and REF is an important variable with respect to increase in purity measure.

2. Are the chocolate ratings getting better with time?
From the Mean Rating vs Time plot we can see that the rating is getting better with time.
We can also see that the increase in time contributes to an increase in rating from the Linear Regression fit model and establishes a linear relationship between them.

3. How does the CocoaPercentage affect the rating of the chocolate?
From the Linear Regression model, we see that an increase in Cocoa Percentage leads to a decrease in rating.

6.References

Flavors of Cacao - Flavor. (2020). Retrieved 27 November 2020, from http://flavorsofcacao.com/flavor.html

Understanding Random Forest. (2020). Retrieved 27 November 2020, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Tableau. 2020. Tableau: Business Intelligence And Analytics Software. [online] Available at: <https://www.tableau.com/>

7. Appendix
R Code

```
#loading required libraries
library(tidyverse)
library(useful)
library(dplyr)
library(cluster)
library(randomForest)
library(GGally)
library(corrplot)
library(DMwR2)
library(factoextra)
#read csv file
flavours <- read.csv("C:\\Users\\sravan\\Documents\\STAT515\\Final\\flavors_of_cacao.csv")
names(flavours)

#examining the data
str(flavours)
dim(flavours)

#remove spaces and line brakes from column names
names(flavours) <- tolower(gsub(pattern = '[[:space:]+]', '_', names(flavours)))
names(flavours)

#renaming columns for easier legibility
flavours<-dplyr::rename(flavours,Maker=company
,REF=ref,Rating=rating,BeanOrigin=specific.bean.origin.or.bar.name,ReviewDate=review.date,Company
Location=company.location,BroadBeanOrigin=broad.bean.origin,BeanType=bean.type,CocoaPercentage
=cocoa.percent)

#displaying column names
names(flavours)

#checking for null values for each column
lapply(flavours,function(x) { length(which(is.na(x)))})

#Removing Bean Type since it has a lot of missing values
#We do not use Broad Bean Origin,Bean Type in our Analysis
flavours <- subset(flavours,select=c(-BeanType,-BroadBeanOrigin))

#checking for null values again
lapply(flavours,function(x) { length(which(is.na(x)))})

#checking size of flavours
dim(flavours)


#remove % from CocoaPercentage and change it to numeric type
```

```
flavours$CocoaPercentage <- lapply(flavours$CocoaPercentage, function(x) gsub("%", "", x))
flavours$CocoaPercentage <- as.numeric(flavours$CocoaPercentage)

#checking summary of the data
summary(flavours)

#checking new column names
colnames(flavours)

#factor
flavours$Maker <- as.factor(flavours$Maker)

#Adding Rating level variable
RatingLevel=ifelse(flavours$Rating<=3.5,"Low","High")
flavours <- data.frame(flavours,RatingLevel)
write.csv(flavours,"C:\\Users\\sravan\\Documents\\STAT515\\Final\\finalflavours.csv",row.names=FALSE)


# get the mean chocolate rating by year
meanRatingByYear <- flavours %>%
  group_by(ReviewDate) %>%
  summarise(Rating = mean(Rating))
print(meanRatingByYear)

# line chart showing change in reviews over years
ggplot(data = meanRatingByYear, aes(x = ReviewDate, y = Rating)) +
  geom_line() +
  scale_x_continuous(breaks = c(2006:2017)) +
  labs(title = "Mean Rating of Chocolate Bars over Years",
       x = "Review Year",
       y = "Mean Rating")

#Histogram showing Count of Ratings
ggplot(flavours, aes(x=Rating, fill=Rating,color=Blue)) +
  geom_bar(color="black") +
  scale_x_continuous(breaks=seq(0,5,0.5))+
  labs (x = "Rating",
        y = "Count",
        title = "Count of Ratings")
#-------------------------------------------------------------

#fitting linear regression model for Rating given ReviewDate
model1 <- lm(Rating~ReviewDate,data=flavours)
sm1<-summary(model1)
sm1
#printing RMSE for model 1
mean(sm1$residuals^2)
#plotting residual plots for model 1
par(mfrow=c(2,2))
```

```
plot(model1)

#Residuals vs Fitted values plot for predicting Rating given ReviewDate
ggplot(data=model1,
       aes(x=.fitted, y=.resid)) +
  geom_point( ) +
  geom_hline(yintercept=0) +
  geom_smooth(se=TRUE, method="loess",
              method.args=list(degree=1, family="symmetric")) +
  labs(x="fitted Values", y="Residuals",title="Residuals vs Fitted plot for predicting Rating given
ReviewDate")


#-----------------------------------------------

#Linear Regression Model for predicting Rating given CocoaPercentage
model2 <- lm(Rating~CocoaPercentage,data=flavours)
sm2<-summary(model2)
sm2

#plotting Residual plots for model 2
par(mfrow=c(2,2))
plot(model2)

#printing RMSE for model 2
mean(sm2$residuals^2)

#Residuals vs Fitted values plot for predicting Rating given CocoaPercentage
ggplot(data=model2,
       aes(x=.fitted, y=.resid)) +
  geom_point( ) +
  geom_hline(yintercept=0) +
  geom_smooth(se=TRUE, method="loess",
              method.args=list(degree=1, family="symmetric")) +
  labs(x="fitted Values", y="Residuals",title="Residuals vs Fitted plot for predicting Rating given
CocoaPercentage")

#fitting using polynomial of CocoaPerecentage
model3 <- lm(Rating~CocoaPercentage+I(CocoaPercentage^2),data=flavours)
sm3 <- summary(model3)
sm3

#plotting residual plots for Model 3
par(mfrow=c(2,2))
plot(model3)

#printing RMSE for Model 3
mean(sm3$residuals^2)
```

```
ggplot(data=model3,
       aes(x=.fitted, y=.resid)) +
  geom_point( ) +
  geom_hline(yintercept=0) +
  geom_smooth(se=TRUE, method="loess",
              method.args=list(degree=1, family="symmetric")) +
  labs(x="fitted Values", y="Residuals",title="Residuals vs Fitted plot for predicting Rating with polynomial
of CocoaPercentage")

#Plot for CocoaPercentage vs Rating
ggplot(aes(x=CocoaPercentage, y=Rating), data=flavours) +
  geom_point() +
  geom_smooth(se=TRUE, method="loess",
              method.args=list(degree=1, family="symmetric")) +
  labs( x= "CocoaPercentage", y="Rating")

#-----------------------------------------------

#perform k-means
flavours_kmeans <- dplyr::select(flavours,Rating,CocoaPercentage,ReviewDate,REF)


#Optimal number of clusters using silhouette method
fviz_nbclust(flavours_kmeans, kmeans, method = "silhouette")


#perform k-means
CompanyK2N5<- kmeans(x=flavours_kmeans, centers=2, nstart=30)

#centers
CompanyK2N5$centers

#plotting k means plot
plot.kmeans(CompanyK2N5, data=flavours_kmeans)
fviz_cluster(CompanyK2N5,data=flavours_kmeans)

#Random Forests
set.seed(599)
flavours_rf <-
select(flavours,c(BeanOrigin,Rating,CocoaPercentage,REF,ReviewDate,CompanyLocation))
head(flavours_rf)
train = sample(1:nrow(flavours_rf), nrow(flavours_rf)/2)
head(train)
set.seed(111)
bag.flavours=randomForest(Rating~., data=flavours_rf, subset=train,mtry=2,
                          importance=TRUE,ntrees=500)

bag.flavours
```

```
# plot OOB-estimated MSE vs # trees
plot(bag.flavours, main="Bagged trees", mtry=2)

# Compute test MSE for  bagged trees
flavours_rf.test=flavours[-train,"Rating"]
yhat.bag = predict(bag.flavours,newdata=flavours_rf[-train,],na.action = na.pass)

#RMSE for Random Forest model
mean((yhat.bag-flavours_rf.test)^2)

#plotting test set rating vs predicted rating
ggplot(data.frame(yhat.bag, flavours_rf.test),
       aes(x=yhat.bag ,y=flavours_rf.test)) +
  geom_point() +
  geom_abline(slope=1,intercept=0) +
  labs(x="predicted Rating",
       y="test-set Rating",
       title="Plot for test set rating vs predicted rating")

#finding important variables using importance function
importance(bag.flavours)
varImpPlot(bag.flavours,main = 'Variable Importance Plot for Predicting Rating')

#printing Confusion matrix
table(flavours_rf.test,yhat.bag)
```