Beer EDA

Quynh Chau

October 26, 2019

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```## Beer and Brewery Trends - Budweiser ## by Quynh Chau in collaboration with Antonio Debouse ## This is an exploratory data analysis (EDA) on Beer Styles and Breweries in the US done for the purposes of preparing ## executive management at Budweiser for the organization’s annual strategic planning. ## This EDA is to be used to inform executive leaderships on potential market expansions to be considered in the company’s ## strategic plan ## A reference list of data sources and assumptions included in the EDA for the purposes of data validation and also updating trending ## purposes in future years. ## Statistical analysis will be done using R.

## The following R libraries are needed for analytical purposes

library(dplyr) library(ggplot2) library(GGally) library(class) library(caret) library(stringr) library(DataExplorer) library(magrittr) library(tidyverse) library(lattice) library(knitr) library(corrplot)

### The following section reads the Breweries data file containing information breweries by US states, and quantifies missing values

Breweries = read.csv(“C:/Users/Owner/Documents/SMU/Doing Data Sci/Unit 8 & 9/Breweries.csv”, sep = “,”, header = TRUE) str(Breweries) dim(Breweries) plot\_missing(Breweries) summary(Breweries) Tot\_Brew <- Breweries %>% group\_by(State) %>% summarise(Tot\_Breweries = n\_distinct(Name, na.rm = TRUE)) # this code counts the number of breweries by state Sum\_Brew = data.frame(Tot\_Brew) str(Sum\_Brew) arrange(Sum\_Brew,desc(Tot\_Breweries))  
SumBrew %>% ggplot(aes(x = State, y = Tot\_Breweries, color = State, fill = State, label = Tot\_Breweries)) + geom\_bar(stat=“identity”) + labs(title = “Number of Breweries by State”) + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

## There are 558 breweries by brewery ID, name,city, and state (4 columns) in this data file

## Question 1. Colorado (CO): 46, California(CA): 39, Michigan (MI): 32, Oregon (OR):29 and Texas (TX): 28 are the states with the highest number of breweries, respectively

### The following section reads data files containing information on Beer Bitterness, Alcohol Content, and Breweries by State

Beers = read.csv(“C:/Users/Owner/Documents/SMU/Doing Data Sci/Unit 8 & 9/Beers.csv”, sep = “,”, header = TRUE) str(Beers) dim(Beers) plot\_missing(Beers) summary(Beers)

## Beers has 2,410 rows and 7 columns; Alcoholic Content (ABV) has 62 missing values; Beeer Bitterness (IBU) has 1,005 missing values

## These values will be omitted in specific analyses.

### Merge Beerand Breweries Data using unique Brewery ID and full join

## Question 2: Merge and check beer and breweries data by reviewing first 6 (head) and last six (tail) observations of merged file

## Renames joined columns with Beer\_Name and Brewery\_Name so that they are descriptive

AllAboutBeer <- full\_join(Beers,Breweries,by =c(“Brewery\_id” = “Brew\_ID”))  
str(AllAboutBeer)  
head(AllAboutBeer) tail(AllAboutBeer) names(AllAboutBeer)[names(AllAboutBeer)==“Name.x”] <- “Beer\_Name” names(AllAboutBeer)[names(AllAboutBeer)==“Name.y”] <- “Brewery\_Name” str(AllAboutBeer)

### Question 3. Check for missing values, which will be excluded depending in specific analyses using na.rm/drop.na/na.omit as appropriate

plot\_missing(AllAboutBeer)

## Question 3. IBU seems to have the most missing values at 41.7% and ABV at 2.57%. These rows will be omitted or deleted during specific analyses

### Question 4. Compute the median alcohol content (ABV) amnd international bitterness (IBU) and plot using bar charts to compare

## rename merged columns to more descriptive names

Medians <- AllAboutBeer %>% group\_by(State) %>% summarise(count = n(), median\_ABV= median(ABV,na.rm = TRUE), median\_IBU = median(IBU,na.rm = TRUE)) arrange(Medians,desc(median\_ABV)) arrange(Medians,desc(median\_IBU))

Medians %>% drop\_na() %>% ggplot(aes(x = State, y = median\_ABV)) + geom\_bar(stat=“identity”, fill = “dark green”) + labs(title = “Beer Alcoholic Content by State”) + theme(axis.text.x = element\_text(angle = 90, hjust = 1)) Medians %>% drop\_na() %>% ggplot(aes(x = State, y = median\_IBU)) + geom\_bar(stat=“identity”, fill = “red”) + labs(title = “Beer Bitterness by State”) + theme(axis.text.x = element\_text(angle = 90, hjust = 1)) Medians[which.max(Medians$median\_ABV),] Medians[which.max(Medians$median\_IBU),]

## Maine has the most bitter beer; District of Columbia (DC) the most alcoholic content in beers as measured by median values by state

### Question 5. States with maximum alcoholic beer and state with the most bitter beer

AllAboutBeer[which.max(AllAboutBeer$ABV),]  
AllAboutBeer[which.max(AllAboutBeer$IBU),]

## Boulder, Colorado has the beer with the highest alcohol content - Lee Hill Series Vol. 5 - Belgian Style Quadrupel Ale, produced by

## the Upslope Brewing Company

## Astoria, Oregon produces the most bitter beer - Bitter Bitch Imperial IPA by the Astoria Brewing Company

### This section describes the relationship between ABV and IBU using the AllAboutBeer merged data set

## Question 6 Comment on the summary statistics and distribution of the ABV variable

summary(AllAboutBeer$ABV,digits = 3) AllAboutBeer %>% drop\_na() %>% ggplot(aes(y=ABV)) + geom\_boxplot(fill = “Red”) + ggtitle(“Beer Alcohol Content”) AllAboutBeer %>% drop\_na() %>% ggplot(aes(y=IBU)) + geom\_boxplot(fill = “Blue”)+ ggtitle(“Beer Bitterness”) AllAboutBeer %>% drop\_na() %>% ggplot(aes(x=ABV)) + geom\_histogram(fill = “Red”) + ggtitle(“Beer Alcohol Content”)

##Beer Alcohol content is right skewed. This is supported by the wide range of ABV from 0.01 to 0.128 and the mean (0.0598) ##being greater than the median (0.056)

###The following section explores the relationship between beer alcohol content (ABV) and beer bitterness (IBU) ## Question 7. Describe the relationship between ABV and IBU, provide scatterplot

AllAboutBeer %>% drop\_na() %>% ggplot(aes(x=ABV, y=IBU)) + geom\_point(color=“blue”)+ geom\_smooth(se=FALSE, color = “red”)+ ggtitle(“ABV and IBU Covariation Analysis”)

ABV\_IBU <- AllAboutBeer[,c(3,4)] ABV\_IBU = na.omit(ABV\_IBU) plot\_correlation(ABV\_IBU, type = ‘continuous’,‘Review.Date’)

## There is a strong correlation between beer bittnerness and alcohol content - correlation coefficient of 0.67. The graph

## of covariation between IBU and ABV suggests that the higher the beer bitterness, the higher the alcohol content.

###This section focuses on ales : India Pale Ales (IPA) and other types of ales (Non-IPAs) using knn classifications ## Question 8 Budweiser would like to investigate the difference between IPAs and nonIPAs

All\_Ale <- filter(AllAboutBeer, grepl(“Ale”, Style)) Only\_Ale <- filter(All\_Ale, !grepl(“IPA”,Style)) Only\_Ale$Style = “Ale”

IPA <- filter(AllAboutBeer, grepl(“IPA”,Style)) IPA$Style = “IPA”

IPA\_and\_Only\_Ale\_Comb <- rbind(IPA,Only\_Ale) str(IPA\_and\_Only\_Ale\_Comb) summary(IPA\_and\_Only\_Ale\_Comb)

#Plot relationship between IBU and AVV IPA\_and\_Only\_Ale\_Comb %>% ggplot(aes(x=ABV, y=IBU, color= Style))+geom\_point()+geom\_smooth(aes(linetype = Style),color=“Black”,se=FALSE)

#Delete all missing observations CleanData\_IPA\_and\_Ale <- na.omit(IPA\_and\_Only\_Ale\_Comb)

#Change Style to a factor variable CleanData\_IPA\_and\_AleStyle)

#Internal cross Validation, unstandardized with k=3 classifications = knn.cv(CleanData\_IPA\_and\_Ale[,c(3,4)], CleanData\_IPA\_and\_AleStyle) confusionMatrix(classifications,CleanData\_IPA\_and\_Ale$Style)

#Internal cross Validation, standardized with k = 3 Standard\_CleanData\_IPA\_and\_Ale <- data.frame(ZABV = scale(CleanData\_IPA\_and\_AleIBU), Style = CleanData\_IPA\_and\_Ale$Style)

classifications = knn.cv(Standard\_CleanData\_IPA\_and\_Ale[,c(1,2)],Standard\_CleanData\_IPA\_and\_AleStyle)

## There is a positive correlation between ABV and IBU for all beer styles. In particular to Ales - between IPAs and

## non-IPA ales, IPAs tend to have higher IBU and ABV. This distinction is supported by the high accuracy of prediction

## for IPAs and non-IPA ales using KNN classification modeling with IBU and ABV with 86% accuracy rate.

### This section combines beer styles, breweries, median household income from US Census Bureau, and beer consumption via gallons consumed per capita per state region

## Question 9 Use combined demographics data to make an recommendation on which US region should be targeted for expansion for ales

BeerDemo = read.csv(“C:/Users/Owner/Documents/SMU/Doing Data Sci/Unit 8 & 9/BeerDemo.csv”, sep = “,”, header = TRUE) BeerDemoSt\_Abbrev) str(BeerDemo)

## Merge demographics with ales data

CleanData\_IPA\_and\_AleState) AlesDemo <- merge(CleanData\_IPA\_and\_Ale,BeerDemo, by=NULL) head(AlesDemo) tail(AlesDemo)

AleRegion <- select(AlesDemo,ABV, IBU,Style,Region,Beer\_Consump,MedHHIncome,Income\_Rank) ##Plot correlation matrix for ales with demographics plot\_correlation(drop\_na(AleRegion), type = “all”,‘Review.Date’)

## The correlation Ales plot shows a positive .25 correlation between the Midwest region and beer consumption. This ## represents a potential underserved market for Budweiser’s consideration to expand in the ales consumer market.

```