## 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)

'data.frame': 558 obs. of 4 variables:

$ Brew\_ID: int 1 2 3 4 5 6 7 8 9 10 ...

$ Name : Factor w/ 551 levels "10 Barrel Brewing Company",..: 355 12 266 319 201 136 227 477 59 491 ...

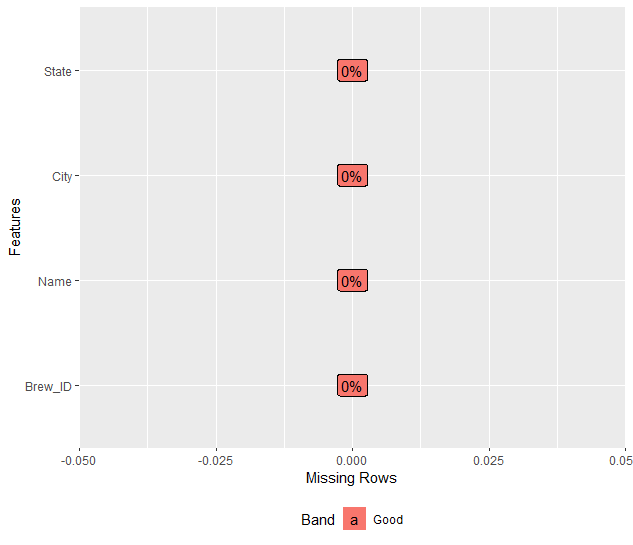
$ City : Factor w/ 384 levels "Abingdon","Abita Springs",..: 228 200 122 299 300 62 91 48 152 136 ...

$ State : Factor w/ 51 levels " AK"," AL"," AR",..: 24 18 20 5 5 41 6 23 23 23 ...

dim(Breweries)

[1] 558 4

plot\_missing(Breweries)



summary(Breweries)

Brew\_ID Name City State

Min. : 1.0 Blackrocks Brewery : 2 Portland: 17 CO : 47

1st Qu.:140.2 Blue Mountain Brewery : 2 Boulder : 9 CA : 39

Median :279.5 Lucette Brewing Company: 2 Chicago : 9 MI : 32

Mean :279.5 Oskar Blues Brewery : 2 Seattle : 9 OR : 29

3rd Qu.:418.8 Otter Creek Brewing : 2 Austin : 8 TX : 28

Max. :558.0 Sly Fox Brewing Company: 2 Denver : 8 PA : 25

(Other):546 (Other) :498 (Other):358

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)

'data.frame': 51 obs. of 2 variables:

$ State : Factor w/ 51 levels " AK"," AL"," AR",..: 1 2 3 4 5 6 7 8 9 10 ...

$ Tot\_Breweries: int 7 3 2 11 39 46 8 1 2 15 ...

arrange(Sum\_Brew,desc(Tot\_Breweries))

State Tot\_Breweries

1 CO 46

2 CA 39

3 MI 32

4 OR 29

5 TX 28

6 PA 24

7 MA 23

8 WA 23

9 IN 22

10 NC 19

11 WI 19

12 IL 18

13 NY 16

14 FL 15

15 OH 15

16 VA 15

17 AZ 11

18 MN 11

19 ME 9

20 MO 9

21 MT 9

22 VT 9

23 CT 8

24 AK 7

25 GA 7

26 MD 7

27 OK 6

28 IA 5

29 ID 5

30 LA 5

31 NE 5

32 RI 5

33 HI 4

34 KY 4

35 NM 4

36 SC 4

37 UT 4

38 WY 4

39 AL 3

40 KS 3

41 NH 3

42 NJ 3

43 TN 3

44 AR 2

45 DE 2

46 MS 2

47 NV 2

48 DC 1

49 ND 1

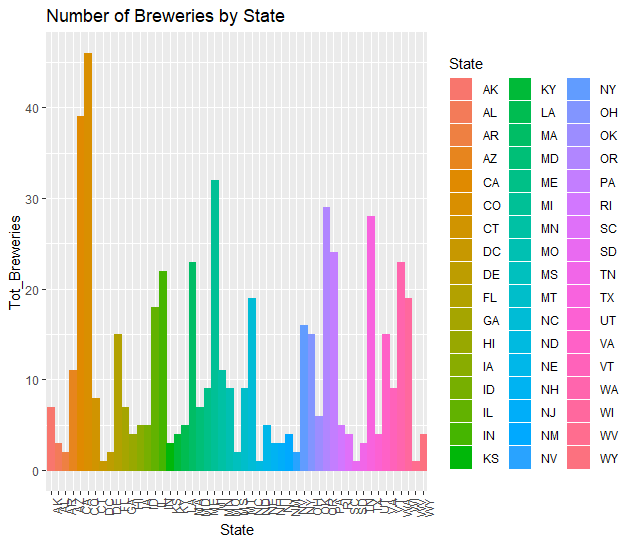
50 SD 1

51 WV 1

Sum\_Brew %>% 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)

'data.frame': 2410 obs. of 7 variables:

$ Name : Factor w/ 2305 levels "#001 Golden Amber Lager",..: 1638 577 1705 1842 1819 268 1160 758 1093 486 ...

$ Beer\_ID : int 1436 2265 2264 2263 2262 2261 2260 2259 2258 2131 ...

$ ABV : num 0.05 0.066 0.071 0.09 0.075 0.077 0.045 0.065 0.055 0.086 ...

$ IBU : int NA NA NA NA NA NA NA NA NA NA ...

$ Brewery\_id: int 409 178 178 178 178 178 178 178 178 178 ...

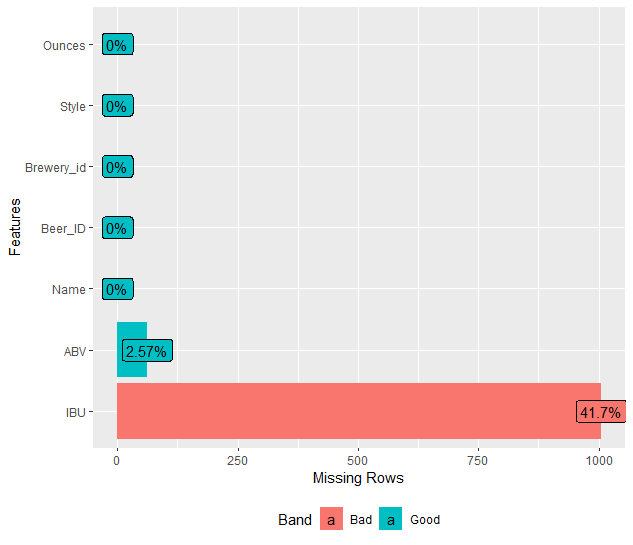
$ Style : Factor w/ 100 levels "","Abbey Single Ale",..: 19 18 16 12 16 80 18 22 18 12 ...

$ Ounces : num 12 12 12 12 12 12 12 12 12 12 ...

> dim(Beers)

[1] 2410 7

plot\_missing(Beers)



summary(Beers)

Name Beer\_ID ABV IBU Brewery\_id Style

Nonstop Hef Hop : 12 Min. : 1.0 Min. :0.00100 Min. : 4.00 Min. : 1.0 American IPA : 424

Dale's Pale Ale : 6 1st Qu.: 808.2 1st Qu.:0.05000 1st Qu.: 21.00 1st Qu.: 94.0 American Pale Ale (APA) : 245

Oktoberfest : 6 Median :1453.5 Median :0.05600 Median : 35.00 Median :206.0 American Amber / Red Ale : 133

Longboard Island Lager: 4 Mean :1431.1 Mean :0.05977 Mean : 42.71 Mean :232.7 American Blonde Ale : 108

1327 Pod's ESB : 3 3rd Qu.:2075.8 3rd Qu.:0.06700 3rd Qu.: 64.00 3rd Qu.:367.0 American Double / Imperial IPA: 105

Boston Lager : 3 Max. :2692.0 Max. :0.12800 Max. :138.00 Max. :558.0 American Pale Wheat Ale : 97

(Other) :2376 NA's :62 NA's :1005 (Other) :1298

Ounces

Min. : 8.40

1st Qu.:12.00

Median :12.00

Mean :13.59

3rd Qu.:16.00

Max. :32.00

## 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 Beer and 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)

'data.frame': 2410 obs. of 10 variables:

$ Name.x : Factor w/ 2305 levels "#001 Golden Amber Lager",..: 1638 577 1705 1842 1819 268 1160 758 1093 486 ...

$ Beer\_ID : int 1436 2265 2264 2263 2262 2261 2260 2259 2258 2131 ...

$ ABV : num 0.05 0.066 0.071 0.09 0.075 0.077 0.045 0.065 0.055 0.086 ...

$ IBU : int NA NA NA NA NA NA NA NA NA NA ...

$ Brewery\_id: int 409 178 178 178 178 178 178 178 178 178 ...

$ Style : Factor w/ 100 levels "","Abbey Single Ale",..: 19 18 16 12 16 80 18 22 18 12 ...

$ Ounces : num 12 12 12 12 12 12 12 12 12 12 ...

$ Name.y : Factor w/ 551 levels "10 Barrel Brewing Company",..: 1 2 2 2 2 2 2 2 2 2 ...

$ City : Factor w/ 384 levels "Abingdon","Abita Springs",..: 32 131 131 131 131 131 131 131 131 131 ...

$ State : Factor w/ 51 levels " AK"," AL"," AR",..: 38 16 16 16 16 16 16 16 16 16 ...

head(AllAboutBeer)

Name.x Beer\_ID ABV IBU Brewery\_id Style Ounces Name.y City State

1 Pub Beer 1436 0.050 NA 409 American Pale Lager 12 10 Barrel Brewing Company Bend OR

2 Devil's Cup 2265 0.066 NA 178 American Pale Ale (APA) 12 18th Street Brewery Gary IN

3 Rise of the Phoenix 2264 0.071 NA 178 American IPA 12 18th Street Brewery Gary IN

4 Sinister 2263 0.090 NA 178 American Double / Imperial IPA 12 18th Street Brewery Gary IN

5 Sex and Candy 2262 0.075 NA 178 American IPA 12 18th Street Brewery Gary IN

6 Black Exodus 2261 0.077 NA 178 Oatmeal Stout 12 18th Street Brewery Gary IN

tail(AllAboutBeer)

Name.x Beer\_ID ABV IBU Brewery\_id Style Ounces Name.y City State

2405 Rocky Mountain Oyster Stout 1035 0.075 NA 425 American Stout 12 Wynkoop Brewing Company Denver CO

2406 Belgorado 928 0.067 45 425 Belgian IPA 12 Wynkoop Brewing Company Denver CO

2407 Rail Yard Ale 807 0.052 NA 425 American Amber / Red Ale 12 Wynkoop Brewing Company Denver CO

2408 B3K Black Lager 620 0.055 NA 425 Schwarzbier 12 Wynkoop Brewing Company Denver CO

2409 Silverback Pale Ale 145 0.055 40 425 American Pale Ale (APA) 12 Wynkoop Brewing Company Denver CO

2410 Rail Yard Ale (2009) 84 0.052 NA 425 American Amber / Red Ale 12 Wynkoop Brewing Company Denver CO

names(AllAboutBeer)[names(AllAboutBeer)=="Name.x"] <- "Beer\_Name"

names(AllAboutBeer)[names(AllAboutBeer)=="Name.y"] <- "Brewery\_Name"

str(AllAboutBeer)

'data.frame': 2410 obs. of 10 variables:

$ Beer\_Name : Factor w/ 2305 levels "#001 Golden Amber Lager",..: 1638 577 1705 1842 1819 268 1160 758 1093 486 ...

$ Beer\_ID : int 1436 2265 2264 2263 2262 2261 2260 2259 2258 2131 ...

$ ABV : num 0.05 0.066 0.071 0.09 0.075 0.077 0.045 0.065 0.055 0.086 ...

$ IBU : int NA NA NA NA NA NA NA NA NA NA ...

$ Brewery\_id : int 409 178 178 178 178 178 178 178 178 178 ...

$ Style : Factor w/ 100 levels "","Abbey Single Ale",..: 19 18 16 12 16 80 18 22 18 12 ...

$ Ounces : num 12 12 12 12 12 12 12 12 12 12 ...

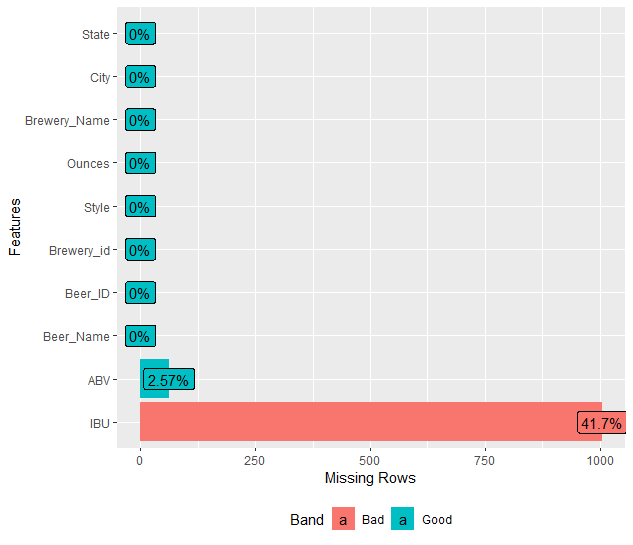
$ Brewery\_Name: Factor w/ 551 levels "10 Barrel Brewing Company",..: 1 2 2 2 2 2 2 2 2 2 ...

$ City : Factor w/ 384 levels "Abingdon","Abita Springs",..: 32 131 131 131 131 131 131 131 131 131 ...

$ State : Factor w/ 51 levels " AK"," AL"," AR",..: 38 16 16 16 16 16 16 16 16 16 ...

### 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))

# A tibble: 51 x 4

State count median\_ABV median\_IBU

<fct> <int> <dbl> <dbl>

1 " DC" 8 0.0625 47.5

2 " KY" 21 0.0625 31.5

3 " MI" 162 0.062 35

4 " NM" 14 0.062 51

5 " WV" 2 0.062 57.5

6 " CO" 265 0.0605 40

7 " AL" 10 0.06 43

8 " CT" 27 0.06 29

9 " NV" 11 0.06 41

10 " OK" 19 0.06 35

# ... with 41 more rows

> arrange(Medians,desc(median\_IBU))

# A tibble: 51 x 4

State count median\_ABV median\_IBU

<fct> <int> <dbl> <dbl>

1 " ME" 27 0.051 61

2 " WV" 2 0.062 57.5

3 " FL" 58 0.057 55

4 " GA" 16 0.055 55

5 " DE" 2 0.055 52

6 " NM" 14 0.062 51

7 " NH" 8 0.055 48.5

8 " DC" 8 0.0625 47.5

9 " NY" 74 0.055 47

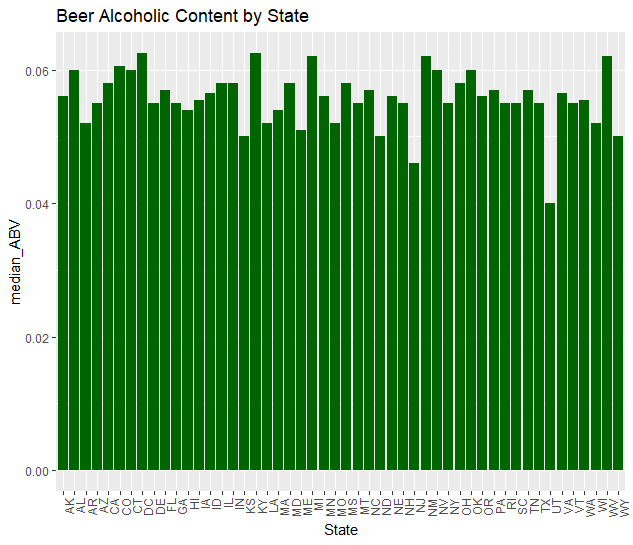
10 " AK" 25 0.056 46

# ... with 41 more rows

Medians %>% drop\_na() %>% ggplot(aes(x = State, y = median\_ABV)) + geom\_bar(stat="identity", fill = "dark green") +

+ labs(title = "Beer Alcoholic Content by State") +

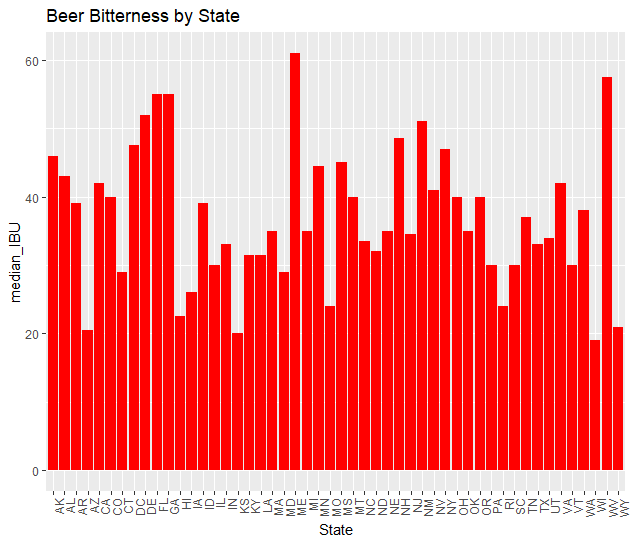
+ theme(axis.text.x = element\_text(angle = 90, hjust = 1))



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),]

# A tibble: 1 x 4

State count median\_ABV median\_IBU

<fct> <int> <dbl> <dbl>

1 " DC" 8 0.0625 47.5

Medians[which.max(Medians$median\_IBU),]

# A tibble: 1 x 4

State count median\_ABV median\_IBU

<fct> <int> <dbl> <dbl>

1 " ME" 27 0.051 61

## 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),]

Beer\_Name Beer\_ID ABV IBU Brewery\_id Style Ounces Brewery\_Name City State

2279 Lee Hill Series Vol. 5 - Belgian Style Quadrupel Ale 2565 0.128 NA 52 Quadrupel (Quad) 19.2 Upslope Brewing Company Boulder CO

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

Beer\_Name Beer\_ID ABV IBU Brewery\_id Style Ounces Brewery\_Name City State

148 Bitter Bitch Imperial IPA 980 0.082 138 375 American Double / Imperial IPA 12 Astoria Brewing Company Astoria OR

## 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

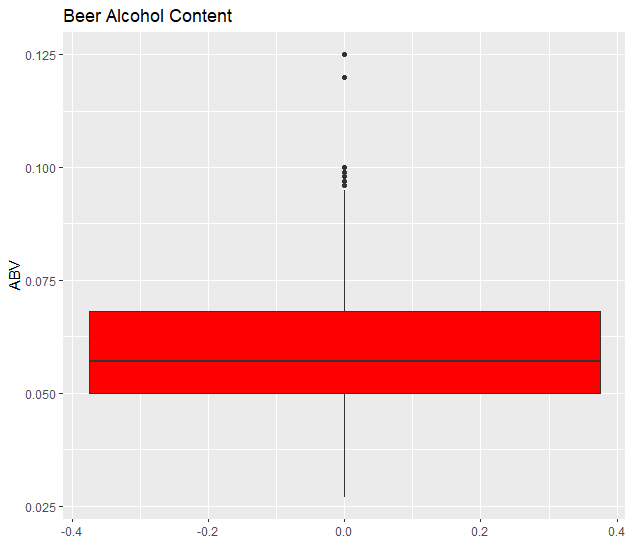
## Question 6 Comment on the summary statistics and distribution of the ABV variable

summary(AllAboutBeer$ABV,digits = 3)

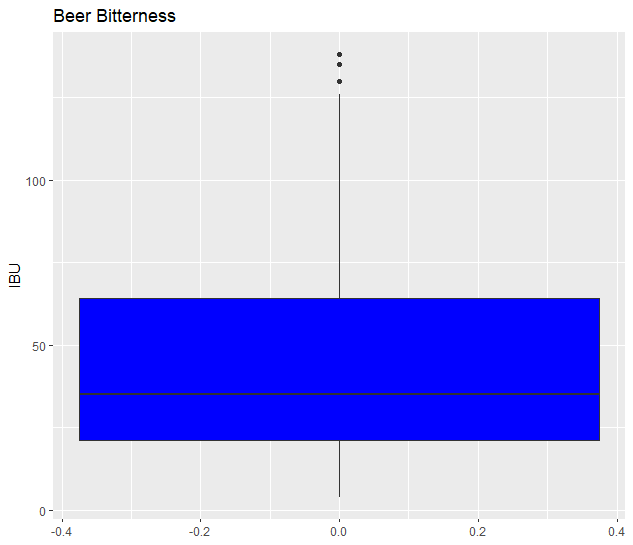
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.0010 0.0500 0.0560 0.0598 0.0670 0.1280 62

AllAboutBeer %>% drop\_na() %>% ggplot(aes(y=ABV)) + geom\_boxplot(fill = "Red") + ggtitle("Beer Alcohol Content")

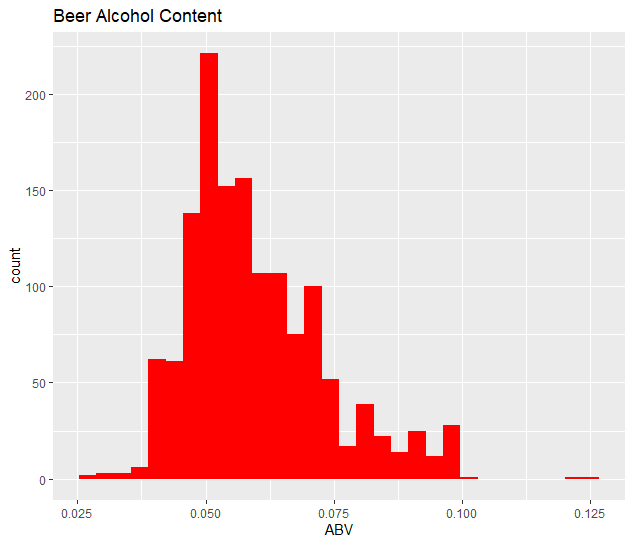


AllAboutBeer %>% drop\_na() %>% ggplot(aes(y=IBU)) + geom\_boxplot(fill = "Blue")+ ggtitle("Beer Bitterness")



AllAboutBeer %>% drop\_na() %>% ggplot(aes(x=ABV)) + geom\_histogram(fill = "Red") + ggtitle("Beer Alcohol Content")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



##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)

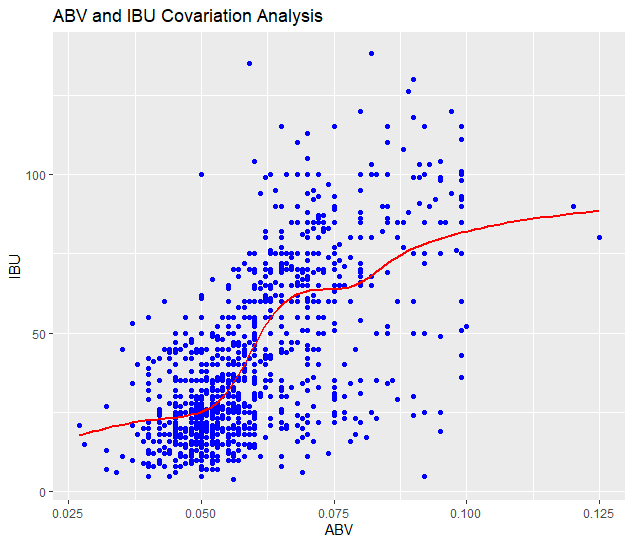
###The following section explores the relationship between beer alcohol content (ABV) and beer bitterness (IBU)

## 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")

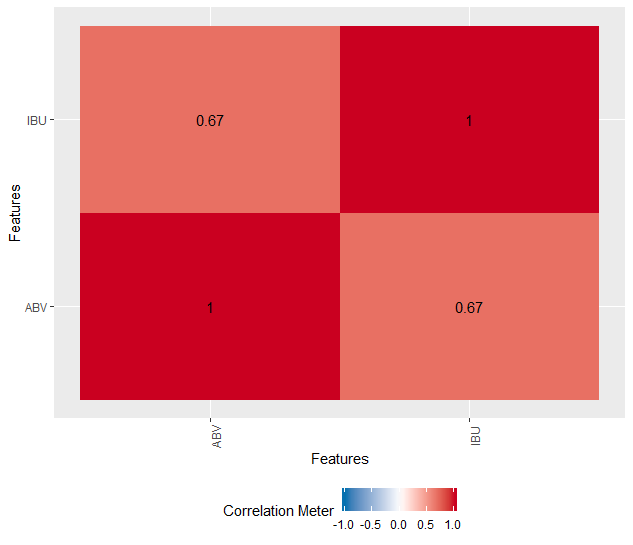
`geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



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)

'data.frame': 1534 obs. of 10 variables:

$ Beer\_Name : Factor w/ 2305 levels "#001 Golden Amber Lager",..: 1705 1842 1819 486 2099 30 352 353 987 351 ...

$ Beer\_ID : int 2264 2263 2262 2131 1980 799 797 796 432 10 ...

$ ABV : num 0.071 0.09 0.075 0.086 0.085 0.07 0.07 0.07 0.097 0.07 ...

$ IBU : int NA NA NA NA NA 70 70 70 94 65 ...

$ Brewery\_id : int 178 178 178 178 178 369 369 369 369 369 ...

$ Style : chr "IPA" "IPA" "IPA" "IPA" ...

$ Ounces : num 12 12 12 12 12 12 12 12 12 12 ...

$ Brewery\_Name: Factor w/ 551 levels "10 Barrel Brewing Company",..: 2 2 2 2 2 4 4 4 4 4 ...

$ City : Factor w/ 384 levels "Abingdon","Abita Springs",..: 131 131 131 131 131 300 300 300 300 300 ...

$ State : Factor w/ 51 levels " AK"," AL"," AR",..: 16 16 16 16 16 5 5 5 5 5 ...

summary(IPA\_and\_Only\_Ale\_Comb)

Beer\_Name Beer\_ID ABV IBU Brewery\_id Style Ounces

Nonstop Hef Hop : 12 Min. : 1.0 Min. :0.02700 Min. : 4.00 Min. : 1.0 Length:1534 Min. :12.00

Dale's Pale Ale : 6 1st Qu.: 783.2 1st Qu.:0.05100 1st Qu.: 27.00 1st Qu.: 92.0 Class :character 1st Qu.:12.00

Dagger Falls IPA : 3 Median :1450.0 Median :0.06000 Median : 45.00 Median :205.5 Mode :character Median :12.00

312 Urban Pale Ale : 2 Mean :1421.1 Mean :0.06131 Mean : 49.95 Mean :231.8 Mean :13.58

312 Urban Wheat Ale: 2 3rd Qu.:2071.2 3rd Qu.:0.07000 3rd Qu.: 70.00 3rd Qu.:363.0 3rd Qu.:16.00

Beach Blonde : 2 Max. :2692.0 Max. :0.09900 Max. :138.00 Max. :558.0 Max. :32.00

(Other) :1507 NA's :42 NA's :590

Brewery\_Name City State

Oskar Blues Brewery : 39 Chicago : 43 CO :178

Brewery Vivant : 35 Portland : 41 CA :124

Sun King Brewing Company : 23 Grand Rapids: 39 MI : 94

Cigar City Brewing Company: 19 San Diego : 33 IN : 88

Hopworks Urban Brewery : 19 Denver : 29 OR : 84

Sixpoint Craft Ales : 15 Longmont : 28 TX : 77

(Other) :1384 (Other) :1321 (Other):889

#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)

`geom\_smooth()` using method = 'loess' and formula 'y ~ x'

Warning messages:

1: Removed 590 rows containing non-finite values (stat\_smooth).

2: Removed 590 rows containing missing values (geom\_point).



#Delete all missing observations

CleanData\_IPA\_and\_Ale <- na.omit(IPA\_and\_Only\_Ale\_Comb)

#Change Style to a factor variable

CleanData\_IPA\_and\_Ale$Style = factor(CleanData\_IPA\_and\_Ale$Style)

#Internal cross Validation, unstandardized with k=3

classifications = knn.cv(CleanData\_IPA\_and\_Ale[,c(3,4)], CleanData\_IPA\_and\_Ale$Style, k = 3)

data.frame(classifications = classifications, true = CleanData\_IPA\_and\_Ale$Style)

classifications true

1 IPA IPA

2 IPA IPA

3 IPA IPA

4 IPA IPA

5 IPA IPA

6 IPA IPA

7 Ale IPA

8 IPA IPA

9 IPA IPA

10 IPA IPA

11 IPA IPA

12 IPA IPA

13 IPA IPA

14 IPA IPA

15 IPA IPA

16 IPA IPA

17 Ale IPA

18 IPA IPA

19 IPA IPA

20 IPA IPA

21 IPA IPA

22 IPA IPA

23 IPA IPA

24 IPA IPA

25 IPA IPA

26 IPA IPA

27 IPA IPA

28 IPA IPA

29 Ale IPA

30 IPA IPA

31 IPA IPA

32 Ale IPA

33 IPA IPA

34 IPA IPA

35 IPA IPA

36 IPA IPA

37 IPA IPA

38 IPA IPA

39 IPA IPA

40 IPA IPA

41 IPA IPA

42 Ale IPA

43 IPA IPA

44 IPA IPA

45 Ale IPA

46 IPA IPA

47 IPA IPA

48 IPA IPA

49 IPA IPA

50 Ale IPA

51 Ale IPA

52 Ale IPA

53 IPA IPA

54 Ale IPA

55 Ale IPA

56 Ale IPA

57 IPA IPA

58 IPA IPA

59 IPA IPA

60 Ale IPA

61 IPA IPA

62 IPA IPA

63 IPA IPA

64 IPA IPA

65 IPA IPA

66 IPA IPA

67 IPA IPA

68 IPA IPA

69 IPA IPA

70 IPA IPA

71 IPA IPA

72 IPA IPA

73 IPA IPA

74 IPA IPA

75 IPA IPA

76 IPA IPA

77 IPA IPA

78 Ale IPA

79 IPA IPA

80 Ale IPA

81 Ale IPA

82 IPA IPA

83 IPA IPA

84 IPA IPA

85 IPA IPA

86 IPA IPA

87 Ale IPA

88 IPA IPA

89 IPA IPA

90 IPA IPA

91 IPA IPA

92 IPA IPA

93 IPA IPA

94 IPA IPA

95 IPA IPA

96 IPA IPA

97 IPA IPA

98 IPA IPA

99 IPA IPA

100 IPA IPA

101 Ale IPA

102 IPA IPA

103 Ale IPA

104 IPA IPA

105 IPA IPA

106 IPA IPA

107 IPA IPA

108 IPA IPA

109 Ale IPA

110 IPA IPA

111 IPA IPA

112 IPA IPA

113 IPA IPA

114 IPA IPA

115 IPA IPA

116 Ale IPA

117 IPA IPA

118 IPA IPA

119 IPA IPA

120 IPA IPA

121 IPA IPA

122 Ale IPA

123 Ale IPA

124 IPA IPA

125 IPA IPA

126 IPA IPA

127 IPA IPA

128 IPA IPA

129 IPA IPA

130 IPA IPA

131 IPA IPA

132 Ale IPA

133 IPA IPA

134 IPA IPA

135 IPA IPA

136 IPA IPA

137 IPA IPA

138 IPA IPA

139 Ale IPA

140 IPA IPA

141 IPA IPA

142 IPA IPA

143 IPA IPA

144 IPA IPA

145 IPA IPA

146 Ale IPA

147 IPA IPA

148 Ale IPA

149 IPA IPA

150 IPA IPA

151 IPA IPA

152 IPA IPA

153 IPA IPA

154 IPA IPA

155 IPA IPA

156 IPA IPA

157 IPA IPA

158 IPA IPA

159 IPA IPA

160 Ale IPA

161 IPA IPA

162 Ale IPA

163 IPA IPA

164 Ale IPA

165 IPA IPA

166 IPA IPA

167 IPA IPA

168 IPA IPA

169 IPA IPA

170 IPA IPA

171 IPA IPA

172 Ale IPA

173 IPA IPA

174 IPA IPA

175 IPA IPA

176 IPA IPA

177 IPA IPA

178 IPA IPA

179 IPA IPA

180 IPA IPA

181 Ale IPA

182 IPA IPA

183 Ale IPA

184 IPA IPA

185 IPA IPA

186 Ale IPA

187 Ale IPA

188 Ale IPA

189 IPA IPA

190 IPA IPA

191 IPA IPA

192 IPA IPA

193 IPA IPA

194 IPA IPA

195 IPA IPA

196 IPA IPA

197 IPA IPA

198 IPA IPA

199 IPA IPA

200 IPA IPA

201 IPA IPA

202 IPA IPA

203 IPA IPA

204 Ale IPA

205 IPA IPA

206 IPA IPA

207 IPA IPA

208 IPA IPA

209 IPA IPA

210 IPA IPA

211 IPA IPA

212 IPA IPA

213 IPA IPA

214 IPA IPA

215 IPA IPA

216 IPA IPA

217 IPA IPA

218 IPA IPA

219 IPA IPA

220 IPA IPA

221 IPA IPA

222 IPA IPA

223 IPA IPA

224 Ale IPA

225 Ale IPA

226 IPA IPA

227 IPA IPA

228 IPA IPA

229 IPA IPA

230 IPA IPA

231 IPA IPA

232 Ale IPA

233 IPA IPA

234 IPA IPA

235 IPA IPA

236 Ale IPA

237 IPA IPA

238 IPA IPA

239 IPA IPA

240 IPA IPA

241 IPA IPA

242 IPA IPA

243 IPA IPA

244 IPA IPA

245 IPA IPA

246 IPA IPA

247 Ale IPA

248 IPA IPA

249 IPA IPA

250 IPA IPA

251 Ale IPA

252 Ale IPA

253 IPA IPA

254 IPA IPA

255 IPA IPA

256 IPA IPA

257 IPA IPA

258 Ale IPA

259 Ale IPA

260 Ale IPA

261 Ale IPA

262 IPA IPA

263 IPA IPA

264 IPA IPA

265 IPA IPA

266 IPA IPA

267 Ale IPA

268 IPA IPA

269 IPA IPA

270 IPA IPA

271 IPA IPA

272 IPA IPA

273 IPA IPA

274 IPA IPA

275 Ale IPA

276 IPA IPA

277 IPA IPA

278 Ale IPA

279 IPA IPA

280 IPA IPA

281 IPA IPA

282 IPA IPA

283 IPA IPA

284 IPA IPA

285 IPA IPA

286 IPA IPA

287 IPA IPA

288 IPA IPA

289 IPA IPA

290 IPA IPA

291 IPA IPA

292 IPA IPA

293 IPA IPA

294 IPA IPA

295 IPA IPA

296 IPA IPA

297 IPA IPA

298 IPA IPA

299 Ale IPA

300 IPA IPA

301 IPA IPA

302 IPA IPA

303 IPA IPA

304 IPA IPA

305 IPA IPA

306 IPA IPA

307 Ale IPA

308 IPA IPA

309 IPA IPA

310 Ale IPA

311 IPA IPA

312 IPA IPA

313 IPA IPA

314 IPA IPA

315 IPA IPA

316 IPA IPA

317 IPA IPA

318 IPA IPA

319 IPA IPA

320 IPA IPA

321 IPA IPA

322 IPA IPA

323 IPA IPA

324 IPA IPA

325 IPA IPA

326 IPA IPA

327 IPA IPA

328 IPA IPA

329 IPA IPA

330 IPA IPA

331 IPA IPA

332 IPA IPA

333 IPA IPA

334 IPA IPA

335 IPA IPA

336 IPA IPA

337 IPA IPA

338 Ale IPA

339 IPA IPA

340 Ale IPA

341 IPA IPA

342 Ale IPA

343 IPA IPA

344 IPA IPA

345 IPA IPA

346 IPA IPA

347 IPA IPA

348 IPA IPA

349 Ale IPA

350 Ale IPA

351 IPA IPA

352 IPA IPA

353 IPA IPA

354 Ale IPA

355 IPA IPA

356 Ale IPA

357 IPA IPA

358 IPA IPA

359 IPA IPA

360 IPA IPA

361 Ale IPA

362 Ale IPA

363 IPA IPA

364 IPA IPA

365 IPA IPA

366 IPA IPA

367 IPA IPA

368 Ale IPA

369 IPA IPA

370 Ale IPA

371 IPA IPA

372 Ale IPA

373 Ale IPA

374 IPA IPA

375 IPA IPA

376 IPA IPA

377 IPA IPA

378 IPA IPA

379 IPA IPA

380 IPA IPA

381 IPA IPA

382 IPA IPA

383 Ale IPA

384 IPA IPA

385 IPA IPA

386 IPA IPA

387 IPA IPA

388 IPA IPA

389 IPA IPA

390 Ale IPA

391 IPA IPA

392 Ale IPA

393 IPA Ale

394 Ale Ale

395 IPA Ale

396 Ale Ale

397 IPA Ale

398 Ale Ale

399 Ale Ale

400 Ale Ale

401 IPA Ale

402 Ale Ale

403 Ale Ale

404 Ale Ale

405 Ale Ale

406 Ale Ale

407 Ale Ale

408 Ale Ale

409 Ale Ale

410 Ale Ale

411 Ale Ale

412 Ale Ale

413 Ale Ale

414 Ale Ale

415 Ale Ale

416 Ale Ale

417 Ale Ale

418 Ale Ale

419 Ale Ale

420 Ale Ale

421 Ale Ale

422 IPA Ale

423 Ale Ale

424 Ale Ale

425 Ale Ale

426 Ale Ale

427 Ale Ale

428 Ale Ale

429 Ale Ale

430 Ale Ale

431 IPA Ale

432 Ale Ale

433 Ale Ale

434 Ale Ale

435 Ale Ale

436 Ale Ale

437 Ale Ale

438 Ale Ale

439 Ale Ale

440 Ale Ale

441 Ale Ale

442 IPA Ale

443 Ale Ale

444 Ale Ale

445 Ale Ale

446 IPA Ale

447 Ale Ale

448 Ale Ale

449 Ale Ale

450 Ale Ale

451 Ale Ale

452 Ale Ale

453 Ale Ale

454 Ale Ale

455 Ale Ale

456 IPA Ale

457 IPA Ale

458 Ale Ale

459 IPA Ale

460 Ale Ale

461 Ale Ale

462 Ale Ale

463 IPA Ale

464 IPA Ale

465 Ale Ale

466 Ale Ale

467 Ale Ale

468 Ale Ale

469 Ale Ale

470 Ale Ale

471 Ale Ale

472 Ale Ale

473 Ale Ale

474 Ale Ale

475 Ale Ale

476 Ale Ale

477 Ale Ale

478 Ale Ale

479 Ale Ale

480 IPA Ale

481 IPA Ale

482 Ale Ale

483 Ale Ale

484 Ale Ale

485 Ale Ale

486 Ale Ale

487 Ale Ale

488 Ale Ale

489 Ale Ale

490 Ale Ale

491 Ale Ale

492 Ale Ale

493 Ale Ale

494 Ale Ale

495 Ale Ale

496 Ale Ale

497 Ale Ale

498 Ale Ale

499 Ale Ale

500 Ale Ale

[ reached 'max' / getOption("max.print") -- omitted 444 rows ]

> confusionMatrix(classifications,CleanData\_IPA\_and\_Ale$Style)

Confusion Matrix and Statistics

Reference

Prediction Ale IPA

Ale 486 70

IPA 66 322

Accuracy : 0.8559

95% CI : (0.8319, 0.8777)

No Information Rate : 0.5847

P-Value [Acc > NIR] : <2e-16

Kappa : 0.7029

Mcnemar's Test P-Value : 0.797

Sensitivity : 0.8804

Specificity : 0.8214

Pos Pred Value : 0.8741

Neg Pred Value : 0.8299

Prevalence : 0.5847

Detection Rate : 0.5148

Detection Prevalence : 0.5890

Balanced Accuracy : 0.8509

'Positive' Class : Ale

#Internal cross Validation, standardized with k = 3

Standard\_CleanData\_IPA\_and\_Ale <- data.frame(ZABV = scale(CleanData\_IPA\_and\_Ale$ABV),

+ ZIBU = scale(CleanData\_IPA\_and\_Ale$IBU),

+ Style = CleanData\_IPA\_and\_Ale$Style)

> classifications = knn.cv(Standard\_CleanData\_IPA\_and\_Ale[,c(1,2)],Standard\_CleanData\_IPA\_and\_Ale$Style, k=3)

> confusionMatrix(classifications,Standard\_CleanData\_IPA\_and\_Ale$Style)

Confusion Matrix and Statistics

Reference

Prediction Ale IPA

Ale 487 68

IPA 65 324

Accuracy : 0.8591

95% CI : (0.8353, 0.8807)

No Information Rate : 0.5847

P-Value [Acc > NIR] : <2e-16

Kappa : 0.7096

Mcnemar's Test P-Value : 0.8623

Sensitivity : 0.8822

Specificity : 0.8265

Pos Pred Value : 0.8775

Neg Pred Value : 0.8329

Prevalence : 0.5847

Detection Rate : 0.5159

Detection Prevalence : 0.5879

Balanced Accuracy : 0.8544

'Positive' Class : Ale

## 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)

BeerDemo$St\_Abbrev = as.character(BeerDemo$St\_Abbrev)

str(BeerDemo)

'data.frame': 50 obs. of 5 variables:

$ St\_Abbrev : chr "AL" "AK" "AZ" "AR" ...

$ Region : Factor w/ 6 levels "Midwest","Northeast",..: 4 3 5 4 6 6 2 4 4 4 ...

$ Beer\_Consump: num 28.9 26 26.6 23.7 25.1 28.3 20.2 28.7 26.3 24 ...

$ MedHHIncome : int 46472 76114 53510 43813 67169 65458 73781 63036 50883 52977 ...

$ Income\_Rank : int 48 4 31 50 10 13 7 15 39 33 ...

## Merge demographics with ales data

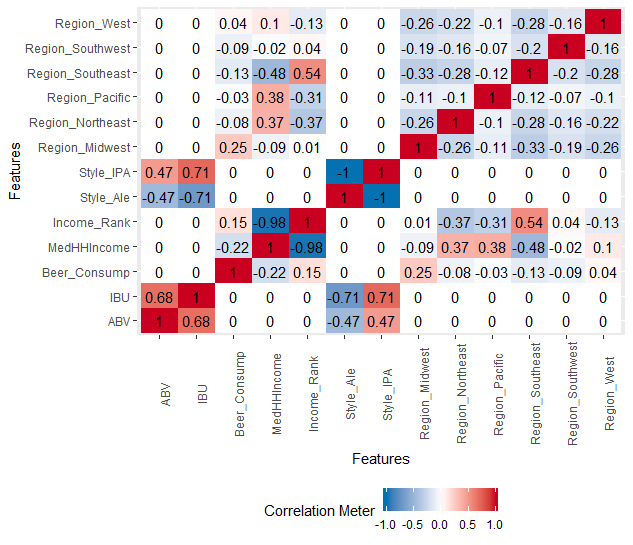
CleanData\_IPA\_and\_Ale$State = as.character(CleanData\_IPA\_and\_Ale$State)

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

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.