Title: "Beer and Breweries data analysis"

author: "Khawulani Thebe

date: "June 2020"

Introduction and Overview

In this report, we explore datasets for `beers` and the `breweries` that are produced in the US. The steps and procedures taken for this analysis are detailed below.

The initial data provided were in two separate datasets, one each for beers and breweries, which were combined into a single dataset.

Analysis

We loaded various libraries needed for our analysis, and reading in our data.

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, warning = FALSE)
```

```{r , include=FALSE}
library(tidyr)
library(ggthemes) # for themes in ggplot
library(knitr)
library(ggplot2) # Data visualization
library(readr) # CSV file I/O, e.g. the read_csv function
library(dplyr) # Heavy use of this library
library(kableExtra) # library to make the document more presentable
```

```{r}

# Read in beer and breweries data set using _csv for more tidy output

Beers <- read_csv('Beers.csv', col_types = cols())

Breweries <- read_csv('Breweries.csv', col_types = cols())

```

We discovered that field Brewery_id and Brew_ID are the same. We corrected this by renaming the column in Beers.csv.

Research Questions

We can now address the research questions against the dataset.

1. How many breweries are present in each state?

We answered this question by retrieving the value of `State` from the `Breweries` data.

```{r}

BrewPerState <-table(Breweries$State)

BrewPerState

```

```{r}

# Renamed Brewery_id to Brew_ID to allow merging of the two data sets

Beers <- rename(Beers, Brew_ID = Brewery_id)

```

2. Merge beer data with the breweries data. Print the first 6 observations and the last six observations to check the merged file.

As part of our analysis, merged the 2 data sets using a full join.

```{r}

BrewBeer <- full_join(Beers, Breweries, by="Brew_ID")

```

```{r}

# We changed the variable names

BrewP <- rename(BrewBeer, Brewery = Name.y, Beer = Name.x,

                OZ = Ounces)

```

To retrieve the first and last six observations from the combined data, we run `head` and `tail` on `BrewBeer`, our combined dataset.

```{r}

kable(BrewBeer %>% head())

```

```{r}

kable(BrewBeer %>% tail())

```

3. Address the missing values in each column.

Use functions `is.na` and `sapply` to determine the number of missing values for each column within `BrewBeer`.

62 ABV values missing, and 1005 IBU values missing.

```{r}

MissingValues <- sapply(BrewBeer, function(x)sum(is.na(x)))

# Used kable library to make document more presentable

MissingValues %>%

  kable("html") %>%

  kable_styling()

```

4. Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.

Create a new data set, from `BrewBeer` dataset, omitting NA values. We then group the relevant values by state, and summarise by the median of the relevant value.

```{r}
# 4. Compute the median alcohol content unit for each state.
 <- BrewBeer%>%
  na.omit() %>%
group_by(State) %>%
summarise(Median = median(ABV)) %>%
arrange(Median)
```

```{r}
# 4. a Compute the median international bitterness unit for each state.
Bitter <- Brewbeer %>%
  na.omit() %>%
group_by(State) %>%
summarise(Median = median(IBU)) %>%
arrange(Median)
```

```{r}
# 4. b Plot a bar chart to compare ABV by state
library(ggthemes)
ggplot(data=, aes(x=State, y=Median)) +
  geom_bar(stat="identity")+
  theme_economist() +
  scale_color_economist()+
  theme(axis.text.x=element_text(size=rel(0.8), angle=90)) +
```

```
ggtitle("Median ABV by State") +
  labs(x="State",y="ABV")
```

```{r}
# 4. c Plot a bar chart to compare IBU by state
ggplot(data=Bitter, aes(x=State, y=Median)) +
  geom_bar(stat="identity")+
  theme_economist() +
  scale_color_economist()+
  theme(axis.text.x=element_text(size=rel(0.8), angle=90))+
  ggtitle("Median IBU by State") +
  labs(x="State",y="IBU")
```

5. Which state has the maximum alcoholic (ABV) beer? Which state has the most bitter (IBU) beer?

We use `which.max` on our `BrewBeer` dataset, seeking maximum value in the `ABV` and `IBU` columns.

From this, we identify Colorado as having the beer with the highest ABV, at `.128`; and we identify Oregon has having the beer with the highest IBU, at `138`.

```{r}
kable(BrewPub[which.max(BrewBeer$ABV),])
```

```{r}
kable(BrewPub[which.max(BrewBeer$IBU),])
```

6. Comment on the summary statistics for the ABV variable.

We do this by calling `summary` on the `ABV` column in our `BrewBeer` dataset.

```{r}
BeerSummary <- (summary(BrewBeer$ABV))
print(BeerSummary)
```

```

### 7. Is there an apparent relationship between the bitterness of the beer and its alcoholic content? Draw a scatter plot. Make your best judgement of a relationship and EXPLAIN  your answer

Use ggplot to plot a scatter plot of the data, using `IBU` and `ABV` as our variables.

Examination: The scatter plot and the regression line suggest that there is a positive, linear relationship between `IBU` and `ABV`.

```{r}
# 7. Draw a scatter plot to compare relationship between beer
# bitterness and alcohol content
ggplot(BrewBeer, aes(x=IBU, y= ABV)) +
  geom_point(shape=1) +
   geom_smooth(method=lm) + # add linear regression line
  theme_economist() +
  scale_color_economist()+
  theme(axis.text.x=element_text(size=rel(1.0)))+
ggtitle("Correlation between IBU and ABV ") +
  labs(x="IBU",y="ABV")
```

Review and Conclusion

We examined the structure of this data, cleaned, and merged them. Various analyses were performed, such as calculating median values for IBU and ABV by state, and determining which states had the beers with the highest ABV and IBU values. We finished by looking for a potential relationship between IBU and ABV, and found that there is evidence to suggest that a positive correlation exists between the two.
```