Blue Beer
Expansion Project
Picasso
Codebook

Karl Jurek
Travis Daun
Joe Jiang

January 7, 2019

Contents

Chapter 1

Executive Summary

We wish to develop a strategy for the deployment of a new beer, Blue Beer's Picasso. Blue Beer's Picasso has higher than typical alcohol content (ABV= 0.06) while having a lower than typical bitterness rating (IBU = 30). Market research has shown that Millennials and iGen (>= 21) prefer locally brewed beers, that have a high alcohol content but are not drawn to highly bitter beers. Blue Beer's Picasso is looking to launch their brewery production in an area with fewer than 10 breweries in the state, has a supporting demographic, and prefers beers with Blue Beer's ABV and IBU characteristics.

The data analysis performed can be found in *JKT_Final.Rmd*. This codebook will provide information on the datasets used, the methods employed, and the pertinent information on how results were obtained to aid in the reproduction of our results.

Chapter 2

R and RStudio

2.1 Version

```
R version 3.5.1 (2018-07-02)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Sierra 10.12.6
```

2.2 Base Packages

```
attached base packages:
grid
stats
graphics
grDevices
utils
datasets
methods
base
```

2.3 Other Packages

Package	Version	Description
forcats	0.3.0	Tools for working with categorical variables (factors)
stringr	1.3.1	Simple, consistent wrappers for common string operations
purrr	0.2.5	Functional programming tools (required for tidyr)
tibble	1.4.2	simple data frames (required for ggplot2, dplyr, and tidyr)
spData	0.2.9.6	Diverse spatial datasets for spatial data analysis
sf	0.7-2	Standardized way to encode spatial vector data
tmap	2.1-1	Flexible, layer-based, and easy to use approach to create thematic maps
bindrcpp	0.2.2	An 'Rcpp' interface to active bindings (required for dplyr)
ggplot2	3.1.0	Create elegant data visualizations using the grammar of graphics
sp	1.3-1	Classes and methods for spatial data
tidyr	0.8.2	Easily tidy data with 'spread()' and 'gather()' functions
dplyr	0.7.8	A grammar of data manipulations

Chapter 3

Datasets

3.1 beers.csv

This dataset contains 2410 observations with 7 variables each. Each observation is for a different beer. The variables associated with each observation are as follows:

Variable	Description	Class
Name	Name of the beer	Factor
Beer_ID	Unique identifier for the beer	Integer
ABV	Alcohol by volume	Number
IBU	The International Bittering Units scale (IBU) scale, is used to approximately quantify the bitterness of beer. This scale is not measured on the perceived bitterness of the beer, but rather the amount of iso-alpha acids.	Integer
Brewery_id	Unique identifier for the brewery where the beer is brewed	Integer
Style	Label given to a beer that describes its overall character and, oftentimes, its place of origin. It's a name that has been broadly accepted by brewers and consumers after years or even centuries of trial and error, scientific research, and marketing.	Factor
Ounces	Volume in ounces of bottled beer	Number

3.2 breweries.csv

This dataset contains 558 observations with 5 variables each. Each observation is for a different beer brewery. The variables associated with each observation are as follows:

Variable	Description	Class
Brew_ID	Unique identifier for the brewery where the beer is brewed	Integer
Name	Name of brewery	Factor
City	City location of brewery	Factor
State	Two letter abbreviation for the State location of the brewery	Factor

3.3 us_states{spData}

US states polygons is a sf object containing the contiguous United States data from the US Census Bureau with a few variables from American Community Survey (ACS). The data contains a data.frame with 49 obs. of 7 variables:

Variable	Description
GEOID	character vector of geographic identifiers
NAME	character vector of state names
REGION	character vector of region names
AREA	area in square kilometers of units class
total_pop_10	numerical vector of total population in 2010
total_pop_15	numerical vector of total population in 2015
geometry	sfc_MULTIPOLYGON The object is in geographical coordinates using the NAD83 datum.

This dataset was used to create map geometry for the US. To use equal area projection us_states must be re-projected as follows:

```
us_states2163 = st_transform(us_states, 2163)
```

Chapter 4

Data Wrangling

4.1 Convert State abbreviations to names

Since the geometry polygons dataset contains a NAME variable that is a character vector of the state names, a method of converting the state abbreviations to names was employed. The following code created a listing of state names and abbreviations. The District of Columbia was also added in.

```
st_crosswalk <- tibble(state = state.name) %>%
  bind_cols(tibble(abb = state.abb)) %>%
  bind_rows(tibble(state = "District of Columbia", abb = "DC"))
colnames(st_crosswalk) <- c("State", "Abb")
```

White space was removed from the State variable in breweries.csv and then renamed to Abb. breweries.csv was then left-joined with the st_crosswalk so that the name of the state was included in the breweries data frame.

```
#Remove any whitespace in state abbreviations
breweries$State <- gsub("\\s+", "", breweries$State)
#Give known column names
colnames(breweries) <- c("Brew_ID", "Name", "City", "Abb")

#Join data frames so state name is available for use
breweries <- left_join(breweries, st_crosswalk, by = "Abb")
```

4.2 Merge beers.csv with breweries.csv

The two dataframes, beers and breweries, were merged into a single dataframe named breweriesdf. The unique brewery ID variable in each dataset was used to perform this merge. Column names for the new breweriesdf dataframe were then set to ensure consistency.

```
breweriesdf <- merge(breweries, beers, by.x = "Brew_ID", by.y ="Brewery_id")

colnames(breweriesdf) <- c("Brewery_ID", "Brewery Name", "City", "Abb", "State", "Beer Name", "Beer_ID"
                           "Ounces")
```

4.3 mapdata

A copy of the breweriesdf dataframe was joined with the us_states2163 dataframe for use in interactive mapping. The State variable in breweriesdf was renamed NAME in mapdata df for the inner join for the map_and_data dataframe.

```
mapdata <- breweriesdf
str(mapdata)
colnames(mapdata) <- c("Brewery_ID", "Brewery Name", "City", "Abb", "NAME",
"Beer Name", "Beer_ID", "ABV", "IBU", "Style", "Ounces")
map_and_data <- inner_join(us_states2163, mapdata)
```

Chapter 5

Data Analysis

5.1 Data Issues

There were some data issues that were identified in the two datasets that were provided.

1. Duplicate beer names with different beer IDs but identical IBU, ABV, Ounces: We assumed that these were in fact different brews of the same beer and they were treated as different beers.
2. Same beer names with different beer IDs but different volumes: We assumed that these were in fact different brews of the same beer and they were treated as different beers.
3. Missing data:

```
Name    Beer_ID        ABV        IBU Brewery_id      Style     Ounces
   0          0         62       1005          0          0          0
```

   There were 62 missing ABV values and 1005 missing IBU values for beers. For the purposes of computing medians we elected to omit these values since it seemed representative of the data that remained. The NAs were spread across breweries in different states and different styles of beers. The only significant impact this had was on South Dakota which did not have a median IBU rating as the result of missing data for the one brewery that is located there.

5.2 Final Processed Data Definitions

Data	Observations	Variables	Description
beers	2410	7	dataframe of beers from beers.csv
breweries	558	5	dataframe of breweries (with state names) and abb from breweries.csv
breweriesdf	2410	11	merged dataframe of beers and breweries
by_state	51	2	dataframe of states and number of breweries in each state
map_and_data	2358	17	joined dataframe of state map geometry and mapdata
mapdata	2410	11	Copy of breweriesdf with State variable renamed to NAME
Medians	51	3	dataframe of median ABV and IBU by state
Medians_map_data	49	4	dataframe of median ABV and IBU by state with state geometry (omitted Alaska and Hawaii)
st_crosswalk	51	2	dataframe of States and Abbreviations (includes DC)
state_map_and_data	49	8	joined dataframe of us_states2163 and by_state
us_states2163	49	7	reprojection to use equal area projection of us_states (provided by spData)

5.3 Other Data Sources

Data for specific areas of exploration were obtained from the following websites:

https://suburbanstats.org/population/how-many-people-live-in-connecticut

https://en.wikipedia.org/wiki/List_of_cities_in_Connecticut

https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_Connecticut

https://suburbanstats.org/population/how-many-people-live-in-maryland

https://en.wikipedia.org/wiki/List_of_municipalities_in_Maryland

https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_Maryland

https://www.google.com/search?q=population+of+chicago+2017&rlz=
1C1CHFX_enUS749US749&oq=population+of+chicago+2017&aqs=
chrome.0.69i59j69i60l3j0l2.2727j1j7&sourceid=chrome&ie=UTF-8

https://en.wikipedia.org/wiki/List_of_cities_in_Indiana

https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_Indiana

https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_Chicago