Group 2 Codebook – Budweiser EDA

Machine:

The required machine needed to run the code is the RStudio software tool. The latest version RStudio 2023.12.1+402 is advised to be used.

Operating System:

The required software needed to run the code is R. The latest R software, R version 3.3.0 + is required.

Data Dictionary:

There were two datasets, Beers and Breweries used during this analysis. The Beers datasets contains a list of 2,410 craft beers and the Breweries dataset contains the list of Budweisers 558 US breweries. Below are the details of the each dataset:

Beers.csv:

Name: Name of the beer.

Beer_ID: Unique identifier of the beer.

ABV: Alcohol by volume of the beer.

IBU: International Bitterness Units of the beer.

Brewery_ID: Brewery id associated with the beer.

Style: Style of the beer.

Ounces: Ounces of beer.

Breweries.csv:

Brew_ID: Unique identifier of the brewery.

Name: Name of the brewery.

City: City where the brewery is located.

State: U.S. State where the brewery is located.

Transformations/Feature Creation:

- **“merged_df”:** initial data frame merging the two data sets Beers and Breweries.
- **“Budweiser_df**”: a filtered version of the ‘merged_df’ variable. It filters in strings in the Style column that contains the letters “IPA” and “Ale”.
- **“Budweiser_df2”**: derived from the dataset “Budweiser_df”, it renames the strings in the Style column in to 2 different strings; “IPA” & “Ale”
- **“trainData”**: Contains 70% of the “Budweiser_df2” dataset. Needed for training the kNN classification.
- **“testData”**: Contains 30% of the “Budweiser_df2” dataset. Needed to test the kNN.
- **“Budweiser_IPA”:** filtered version of the “Budweiser_df2” dataset, it contains only “IPA” strings from the “Style” column.
- **“Budweiser_Ale”:** filtered version of the “Budweiser_df2” dataset, it contains only “Ale” strings from the “Style” column
- **“lookup”:** a data frame to bring in the state abbreviation and full state name to be used in a merge for the heat map.
- **“Breweries2”:** a data frame merging Breweries and lookup to pull in the state name for the heat map.
- **“map.df”:** a data frame merging states and BreweriesMapdata2 by region.

Merging Details:

The Breweries data was merged into the Beers data to create the merged_df data frame.

For the heat map a data frame with the state abbreviation and state name was added as lookup, to merge with the Breweries state abbreviation data to bring in the state name. The state name is then merged as region to be used in map.df to create the visualization.

General Information:

In order to get the merge to work for the heat map, the Breweries State column needs to have leading and trailing spaces removed. This can be done in the csv file itseflt with the trim() function. If this step is not done, the map will appear greyed out as it does not count the state with spaces.