**Faculty of Business and Hospitality**

**2019-2020 Academic Year**

**Semester 1**

****

**Student Name: Yash Sinojia**

**Student ID: A00268852**

**Course: Master of Science in Data Analytics**

**Date: 18-Nov-2019**

**Title: Part 1, Client - Analyst - Observer Interview**

* **The Business Problem: Draft**

When it comes to beer in our current globalized world, we can find exported beer brands from all over the world in most of the bars. People are gradually forgetting their traditional taste, aroma and experience of indigenous beers, the characteristic of a particular region where it originated. My business proposal is to open a franchise of beer bars that serves homebrewed beers instead of exported brands. The assumption being, homebrewed beers have lesser production costs and hence can be served cheaper than beers in other bars. Hence, significant amount of people may choose this traditional bar over the others. My problem is to select the best set of homebrewed beer recipes to be included in these traditional bars.

* **The Interview**
  + **Background**

Team members:

1. **Y**ash Sinojia
2. **F**redrick Sigalla
3. **C**hinonyerem Onyeonoru

In our discussion meetings, we divided our roles using Latin Squares randomization technique as follows:

For the default case, we sorted with the initial of our first names

(i.e. C – F – Y)

Then we formed the Latin Squares as:

Client – Analyst – Observer

C F Y

F Y C

Y C F

And hence, for this report, I am the Client, Chinonyerem is the Analyst and Fredrick is the Observer.

* **Discussion**

**Client**: As per as my business proposal, I have two datasets as described in data dictionary. [1] consists of reviews of expert opinion on beers. This dataset consists of 1.5m rows and 13 columns. [2] consists of various home brewed recipes, its techniques and features. This dataset consists of 7.5k rows and 21 columns. I want you to skim through it and come up with some factors.

**Analyst:** I see a diverse and highly insightful data but with null values, outliers and other issues. Before the analytics part, let me delve deeper into the business understanding of your problem. So, what is the geographical scope of your business?

**Client:** The preference for geography will be the major cities of Ireland. And I want the set of recipes with Irish origin only.

**Observer:** This preference is reflected in the datasets, as there are diverse categories of beer that have the string ‘Irish’ in it.

**Analyst:** What is the objective behind analyzing two datasets?

**Client:** In the current market scenario, the exported beers are more famous than the homebrewed beers. So the target is to gather the valuable insights from [1] and filter [2] on the basis of same beer styles in it. Further filtering on the significance of alcohol by volume in that particular beer styles.

So what insights do you see in [1]?

**Observer:** The diverse user reviews in [1], if analyzed properly can aid the shortlisting to a great extent.

**Analyst:** My perception would be to calculate the impact of the ratings of aroma, palate, taste and appearance in determining the overall rating in [1]. And then, considering the categories which best matches the calculated impacts.

**Client:** Interesting.

**Analyst:** Then I need to clean up the data by getting rid of missing values, duplicate data and out-of-context data. Which is straightforward in [1] but in [2], it gets a bit complex. So please throw some light on what kind of nature are you expecting to find in the results?

**Client:** Definitely, there are some standard protocols for homebrewed beers which I expect to adhere to, as expert opinion is always the best when money is on the roll.

**Observer:**

(*Quick visualization & analysis of Homebrew Recipes | Kaggle*, no date)

(*Beer Judge Certification Program (BJCP)*, no date)

and (*American Craft Beer Reviews and Ratings*, no date)

The references provide information on the practical factors, i.e. the possible range of values, time efficiency, resource optimization and other data sanity checks.

**Analyst:** Alright, we’re good to go on resolving all the discussed issues with the dataset.

**Client**: Perfect!

* **The issues and solutions**

1. **Filtering Beer Styles:**

As the geographical focus of the business is Ireland, only the beer style types having the string ‘Irish’ in their category remains in the dataset.

As a result, the dataset is shrunk to 36982 rows in [1] and 3382 rows in [2].

And the resulting categories are as follows:

* Irish Dry Stout
* Irish Extra Stout
* Irish Red Ale
* Irish Pale Ale
* Irish Cream Ale
* Irish Lager

1. **Handling Missing Values:**

In [1],

All the missing values are in ‘beer\_abv’ column, which only consists of 5.22% of the entire dataset. So it is better to drop all rows with missing values rather than replacing it with central measures as this column is most important for analysis so values can’t be manipulated.

Now, the dataset is shrunk to 35051 rows.

In [2],

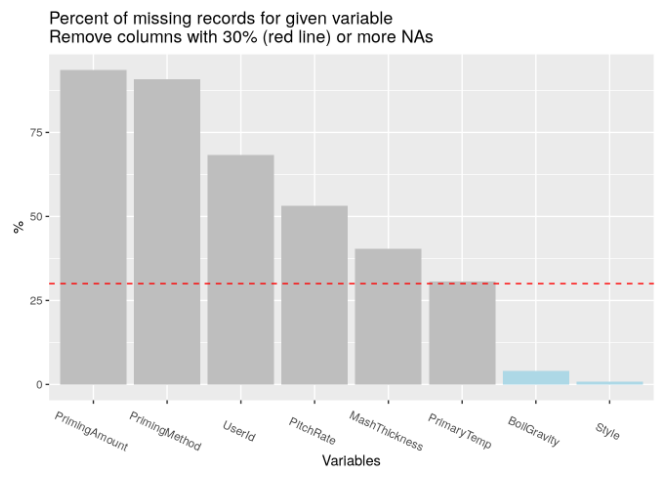
There are a lot of fields termed as N/A, so the first part would be to convert N/A values to NA values.

Replaced 10350 values.

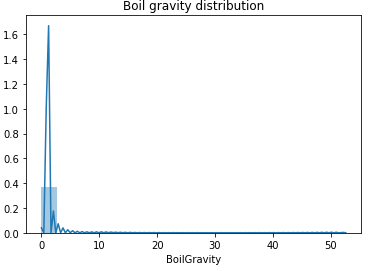
There are 7 columns with null values in the entire dataset. It is statistically better to drop those columns who have more than 30% of its values as NA to avoid biases.

The columns ‘PrimingAmount’, ‘PrimingMethod’, ‘Userid’, ‘PitchRate’, ‘MashThickness’ and ‘PrimaryTemp’ have been dropped. None of these columns were essential for analysis.

17 columns left in the dataset.



Still, NA values in ‘BoilGravity’ needs to be fixed.



Boil Gravity is heavily left skewed with a lot of outliers.

The ‘SugarScale’ column mentions the units of ‘OG’, ‘FG’ and ‘BoilGravity’ columns. It can be seen that ‘SugarScale’ has 3% of values in Plato units. So, this step converts it into Specific Gravity units as: SG = 1 + (4P/1000), which will normalize the outliers from all ‘OG’, ‘FG’ and ‘BoilGravity’ columns.

Now, as there is still significant amount of outliers, we replace the NA values with median as the measure of centrality.

1. NA values are replaced by 1.042 as the median.
2. **Handling out-of-range values**

There are a set of protocols for homebrewed beers in [2] as described in

(*Quick visualization & analysis of Homebrew Recipes | Kaggle*, no date)

(*Beer Judge Certification Program (BJCP)*, no date)

and (*American Craft Beer Reviews and Ratings*, no date)

From these, the datasets can be normalized to realistic values and even further reduce outliers.

1. IBU values should fall between 0 and 120 to avoid excessive bitterness in beers.

(7 rows dropped)

1. Color of a beer should not be higher than 50 as per Standard Reference Method cutoff.

(0 rows dropped)

1. As yeast converts sugar to alcohol gravity decreases. If FG > OG, something would be very wrong indeed.

(0 rows dropped)

1. ABV < 25% to keep alcohol contents practical for a beer.

(1 row dropped)

1. ABV can be calculated as (OG - FG) \* 131.25. So the calculated ABV can be compared with Listed ABV.

Listed ABV: Mean = 5.3265, Median = 5.24

Calculated ABV: Mean = 5.3217, Median = 5.25

Very minute differences, the data is spot on.

(0 rows dropped)

1. **Handling other lurking outliers**
2. Boil Time

Boil Time in the range of 60 – 96 can normalize and introduce time optimization too.

(76 rows dropped)

1. Efficiency

Practically, the efficiency in extracting mash can be never > 85% and <= 60% seems too inefficient.

(731 rows dropped)

1. Size(L) and BoilSize

Handling more than 500 Litres of single vessel size for brewing a beer can need too much capacity.

(33 rows dropped)

1. **Removing Duplicate Ratings**
2. In [1], the rating scale is in 1-5, all the zero ratings need to be removed.
3. row dropped)
4. In [1], there are duplicate values of same user rating for the same beer more than once, here the rating with highest number is considered and other rows are pruned.

(363 rows dropped)

1. **Dropping Redundant Columns**

There are some columns in the dataset that cannot be used for analysis, so better get rid of it.

In [1], ‘brewery\_id’, ‘brewery\_name’, ‘review\_time’, ‘review\_profilename’ and ‘beer\_beerid’ are dropped.

In [2], ‘URL’, ‘StyleID’ and ‘SugarScale’ are dropped.

After all these imputations,

[1] has 34687 Rows x 8 Columns

[2] has 2533 Rows x 14 Columns

Now, the datasets are clean ready for the statistical analysis.

* **Business Problem: Refined**

The clearly defined business proposal pertaining from the above discussions stands as: To start a franchise of beer bars in the major cities of Ireland that serves traditional Irish homebrewed beers, with practical and efficient production techniques. These homebrewed beers shall act as alternatives to exported beers in terms of similar beer style in lesser price. These homebrewed beers adhere to standard protocols defined by experts in the subject and also mimics characteristics of exported beers to compete within current market scenario.

* **Data Dictionary**

1. Beer Reviews Dataset

(*Beer Reviews | Kaggle*, no date)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Measurement Units** | **Allowed Values** | **Description** |
| brewery\_id | ID | Numeric | 0 – 28003 | Unique Identification for a brewery |
| brewery\_name | Category | - | Any | Full name of the brewery |
| review\_time | Integer | Seconds | Numeric | Time in number of seconds |
| review\_overall | Decimal | Numeric | 1 – 5  (Step 0.5) | Overall review of beer |
| review\_aroma | Decimal | Numeric | 1 – 5  (Step 0.5) | Review on aroma of beer |
| review\_appearance | Decimal | Numeric | 1 – 5  (Step 0.5) | Review on appearance of beer |
| review\_profilename | String | - | Any | Reviewer user identification |
| beer\_style | Category | - | Any | Style of the beer |
| review\_palate | Decimal | Numeric | 1 – 5  (Step 0.5) | Review on palate of beer |
| review\_taste | Decimal | Numeric | 1 – 5  (Step 0.5) | Review on taste of beer |
| beer\_name | Category | - | Any | Name of the beer |
| beer\_abv | Decimal | V/V% | 1 – 50  (Step 0.1) | Alcohol content by volume |
| beer\_beerid | ID | Numeric | Numeric | Unique Identification of a beer |

1. Brewer’s Friend Beer Recipes Dataset

(*Brewer’s Friend Beer Recipes | Kaggle*, no date)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Measurement Units** | **Allowed Values** | **Description** |
| BeerID | ID | Numeric | 1 - 73861 | Unique Identification of a beer |
| Name | String | - | Any | Name of recipe provider |
| URL | Weblink | - | Web address | Website for the recipe |
| Style | Category | - | Any | Type of Brew |
| Size(L) | Integer | Litres | 1 - 9200 | Amount brewed for recipe listed |
| OG | Decimal | Unitless | 1 - 35  (4 Deci) | Specific gravity of wort before fermentation |
| FG | Decimal | Unitless | 1 - 35  (4 Deci) | Specific gravity of wort after fermentation |
| ABV | Decimal | V/V% | 1 - 60  (2 Deci) | % alcohol by volume |
| IBU | Decimal | Numeric | 0 - 120  (2 Deci) | International bittering units |
| Color | Decimal | Numeric | 0 - 50  (2 Deci) | Color units by Standard Reference Method |
| BoilSize | Decimal | Litres | 1 - 9700  (2 Deci) | Fluid at beginning of boil |
| BoilTime | Integer |  | 0 - 240 | Time wort is boiled |
| BoilGravity | Decimal | Unitless | 1 - 100  (4 Deci) | Specific gravity of wort before boil |
| Efficiency | Integer | Numeric | 1 - 100 | Efficiency in extracting sugars from the grain during mash |
| Mash Thickness | Decimal | Numeric | 1 - 100  (4 Deci) | Amount of water per pound of grain |
| SugarScale | Category | - | SG / Plato | Scale to determine concentration of dissolved solids in wort |
| BrewMethod | Category | - | Any | Various techniques for brewing |
| PitchRate | Decimal | M cells/ml/deg P | 0 - 2  (Step .25) | Yeast added to the fermenter per gravity unit |
| PrimaryTemp | Decimal | deg C | -20 - 120  (2 Deci) | Temperature at the fermenting stage |
| PrimingMethod | Category | - | Any | Method used for priming sugar |
| PrimingAmount | String | Oz/tsp/bottle/ml | Any | Amount of priming sugar used |

* **References**

*American Craft Beer Reviews and Ratings* (no date). Available at: https://www.twobeerdudes.com/ (Accessed: 17 November 2019).

*Beer Judge Certification Program (BJCP)* (no date). Available at: https://www.bjcp.org/ (Accessed: 17 November 2019).

*Beer Reviews | Kaggle* (no date). Available at: https://www.kaggle.com/rdoume/beerreviews (Accessed: 17 November 2019).

*Brewer’s Friend Beer Recipes | Kaggle* (no date). Available at: https://www.kaggle.com/jtrofe/beer-recipes (Accessed: 17 November 2019).

*Quick visualization & analysis of Homebrew Recipes | Kaggle* (no date). Available at: https://www.kaggle.com/blasterbrewmaster/quick-visualization-analysis-of-homebrew-recipes (Accessed: 17 November 2019).