****

|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Starting Homebrewed Beer Bars that stands out in current market of Exported Beers  Data Analytics Project, Part 2, Statistical Analysis Plan |
| |  |  |  | | --- | --- | --- | | **Yash Sinojia** | **11/28/19** | **Data Analytics** | |

**Faculty of Business and Hospitality**

**2019-2020 Academic Year**

**Semester 1**

**Subject: Data Analytics**

**Student Name: Yash Sinojia**

**Student ID: A00268852**

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

**Date: 28-Nov-2019**

**Part 2, Statistical Analysis Plan**

**TABLE OF CONTENTS**

[**1.** **Revision History** 4](#_Toc25770657)

[**2.**  **Abbreviations** 4](#_Toc25770658)

[**3.** **Introduction** 5](#_Toc25770659)

[**4.**  **Purpose of SAP** 5](#_Toc25770660)

[**5**. **Project Objectives and Endpoints** 6](#_Toc25770661)

[**5.1**  **Project Objectives** 6](#_Toc25770662)

[5.1.1 **Pre-Process Objective** 6](#_Toc25770663)

[5.1.2 **Primary Objective** 6](#_Toc25770664)

[5.1.3 **Secondary Objective** 6](#_Toc25770665)

[**5.2**  **Project Endpoints** 7](#_Toc25770666)

[5.2.1 **Pre-Process Endpoint** 7](#_Toc25770667)

[5.2.2 **Primary Endpoint** 7](#_Toc25770668)

[5.2.3 **Secondary Endpoint** 8](#_Toc25770669)

[**6.** **Study Methods** 8](#_Toc25770670)

[**6.1** **General Study Design and Plan** 8](#_Toc25770671)

[**6.2** **General Study Populations** 8](#_Toc25770672)

[**7.** **General Issues for Statistical Analysis** 8](#_Toc25770673)

[**7.1**  **Software Used for the Analysis** 8](#_Toc25770674)

[**7.2** **Issues Planning the Analysis** 9](#_Toc25770675)

[7.2.1 **Unnormalized Data** 9](#_Toc25770676)

[7.2.2  **Prerequisites for Statistical Tests** 10](#_Toc25770677)

[**7.3**  **Multicenter Studies** 10](#_Toc25770678)

[**7.4**  **Descriptive Statistics** 10](#_Toc25770679)

[**8.**  **Analyses** 11](#_Toc25770680)

[**8.1**  **Primary Analysis** 11](#_Toc25770681)

[**8.2**  **Secondary Analysis** 13](#_Toc25770682)

[**9.**  **Reporting Conventions** 13](#_Toc25770683)

[**10.**  **References** 13](#_Toc25770684)

[**Appendix A** 14](#_Toc25770685)

[**Appendix B** 15](#_Toc25770686)

[**Appendix C** 16](#_Toc25770687)

# 1. Revision History

|  |  |
| --- | --- |
| Revision | Description |
| Version 1 | First Release |

Looking back to Part 1 of the project in retrospection, the business problem was defined and refined by insights gathered from brainstorming amongst the Client, Analyst and Observer. The discussion also led to several issues in the dataset and was significant in pointing out several gaps that kept the dataset a step away from the analysis. These problems were handled and as a result the dataset was cleaned and readied to be taken to next level.

Following points were taken into consideration for upgradation of dataset:

1. Filtering by Irish Beer Styles
2. Handling Missing Values
3. Handling Out of Range Values
4. Handling Outliers Lurking Yet
5. Removing Duplicates
6. Dropping Redundant Columns

# 2. Abbreviations

|  |  |
| --- | --- |
| Abbreviation | Stands for |
| SAP | Statistical Analysis Plan |
| [1] | Brewer’s Friend Homebrewed Beer Recipes Dataset  (Brewer’s Friend Beer Recipes | Kaggle, no date) |
| [2] | Beer Reviews Dataset  (Beer Reviews | Kaggle, no date) |

# 3. Introduction

Pertaining from the Client – Analyst – Observer Interview, the business problem stood 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.*” This report defines a statistical plan from it to be analyzed and interpreted in next and final part of the project.

This report also revolves around issues on the nature and format of data which inevitably arises when subjected to a particular kind of statistical test even though the data is already cleaned and hence pre-processed. Also there can be several restrictions that a test can impose on the dataset which will be highlighted here.

# 4. Purpose of SAP

This Statistical Analysis Plan (SAP) describes the planned analyses and reporting of insights that makes traditional beer bars stand out in current market of exported beers. A comparison among several different factors of homebrewed beer recipes will be checked and selected on the basis of insights from similar beer types in reviews. This is a project to make sure the shortlisted homebrewed beers adhere the market preferences (*Beer Judge Certification Program (BJCP)*, no date) and practices (*Brewer’s Friend - Homebrew Beer Recipes, Calculators, and Forum*, no date).

The first part of this SAP is to ensure the data of homebrewed beers are relevant in terms of statistics and practical science with respect to the relationship among several factors, which will be tested by hypotheses.

The second part of this SAP is to analyze trends in beer reviews which can be adopted to find the one best beer for each beer style in homebrewed recipes. This would be possible by gathering correlations among reviews.

The structure and content of this SAP provides sufficient detail to allow a full understanding of the problem. The following documents were reviewed in the preparation of this SAP:

Client – Analyst – Observer Interview, 18th November, 2019.

The reader of this SAP is encouraged to also read the Client – Analyst – Observer Interview for complete clarification and details on Business Problem Statement.

# **5. Project Objectives and Endpoints**

## **5.1 Project Objectives**

### **5.1.1 Pre-Process Objective**

To check whether the data to be analyzed by statistical tests is normally distributed. And if not, to calibrate that data to fit the normal curve. So that the data qualifies for tests that demand normality.

### **5.1.2 Primary Objective**

To verify that there is a consistency in the quality of the brewing process in [1]. And also to ensure the distinctiveness of styles in [1].

### **5.1.3 Secondary Objective**

To determine correlation factors amongst ratings in [2]. To generate a new metric from the determined correlations in reviews of famous exported beers in [2] that would catalyze the shortlisting process of similar and definitive styles in [2].

## **5.2 Project Endpoints**

### **5.2.1 Pre-Process Endpoint**

To apply normality tests on ‘OG’, ‘FG’, ‘BoilGravity’, ‘ABV’, ‘Size(L)’ and ‘BoilSize’ columns in [1] and ‘beer\_abv’ column in [2].

### **5.2.2 Primary Endpoint**

1. To verify if there is a significant difference among ‘OG’, ‘BoilGravity’ and ‘FG’ columns of [1] as readings of fluid density in three stages of brewing.
2. To verify if there isn’t any remarkable reduction of fluid sizes measured initially in ‘Size(L)’ to later boil stage in ‘BoilSize’ column in [1] as parts of the brewing process.
3. To check whether there is any difference in the distribution of % Alcohol by Volume between both datasets [1] and [2].
4. To check the association among various beer styles on the basis of their brewing methods in [1].
5. To check if the IBU units between Irish Red Ale and Irish Pale Ale are different in [1].
6. To check if the alcohol content distribution is different between Irish Dry Stout and Irish Extra Stout in [1].

### **5.2.3 Secondary Endpoint**

To find the degrees of correlation between each feature ratings data and overall ratings. And to determine the impact of feature ratings on the overall ratings. For that, a derived weighted average that identifies the weights of each feature rating in determining the overall rating.

# **6. Study Methods**

## **6.1 General Study Design and Plan**

Involves two major studies: one being the extensive details on home brewing several kinds of recipes on the basis of beer style and the other being the rating details for myriad of user reviews on beers, categorized on the basis of beer features by styles. And henceforth it makes this study a multi-group, parallel double-centered study.

## **General Study Populations**

This study consists of two populations: one from the recipe data provided by Brewer’s Friend which has a specialization in studying and supporting the business or household interests in small scale brewing across the globe. The another being a large collection of general user-oriented preferences for beers across the globe according to features.

This study considers only the sample population for Irish Beer Styles extracted from both of the study populations of recipes and reviews.

# **7. General Issues for Statistical Analysis**

## **7.1 Software Used for the Analysis**

All the statistical analyses will be carried out in Minitab (v19) statistical analysis software.

The data-preprocessing, i.e. calibrating the data to fit the normal curve was carried out in Microsoft Excel.

## **7.2 Issues Planning the Analysis**

### **7.2.1 Unnormalized Data**

As mentioned in the pre-process objective, much of the statistical tests that are considered for analysis requires the data be normalized to fit a bell shaped curve. In raw state, only a few data come naturally normalized. The data in this study wasn’t in normalized form as:

1. [1] consists of distinctively selected beer recipes from plethora of sources all across the globe. Thus, there is high probability that a permutation and combination of such a gathering of data may result into an un-normalized distribution.
2. [2] consists of a large collection of random user reviews from all across the globe. Thus, there is high probability that such random selection of reviews may result into an un-normalized distribution.

To tackle this issue, the data was exposed to pruning, filtering as well as several data points were manipulated. This data was subjected in such a way that it doesn’t lose its essence but may deviate a bit from its descriptive statistics and significantly in its nature of distribution from the pre-calibrated state.

### **7.2.2 Prerequisites for Statistical Tests**

There are certain prerequisites for Statistical Tests, i.e. the data has to be structured and categorized separately in order for the test to recognize it for the analysis.

Here, for a data to be subjected to Chi-square test for association, it is essential for the data to be categorized. So the continuous data points from the data set were counted and categorized accordingly in a separate table for the test.

Also, the count for some categories in separate table were less than 5. Chi-square test doesn’t work if any expected value calculated during the test is less than 1, the test stops execution and throws an error. So it is a statistical rule that if any computed value is less than 5, the entire category has to be merged with the previous category. As a result, some categories were merged for that to work in the test but those merged categories actually lost its distinctiveness as a category.

## 7.3 Multicenter Studies

This is a double center study. The study in both the centers are connected by the same beer styles (i.e. Traditional Irish only) and the content of alcohol data according to beers. Both of the datasets have ‘Beer Style’ and ‘%ABV’ columns and the analysis can be bridged by these columns.

## 7.4 Descriptive Statistics

Descriptive Statistics will be used to perform a preliminary analysis before subjecting the study to any statistical hypotheses. It will also be used to describe the characteristics of the study sample.

Categorical variables will be described by using frequencies and percentages and visualized by Barplots. Scatterplots and Boxplots will be used to evaluate the distribution of continuous variables and to identify any outliers or potential errors in the data.

For continuous variables, the mean and standard deviation will be used to describe the typical value and spread in the sample. For continuous variables that are not normally distributed, the median and interquartile range will be used to describe the typical value and spread in the sample.

The statistical significance level of all tests will be 95%, the nature of tests whether single-tailed or two-tailed will depend upon the significance for directions in a test.

The pooled estimate of variances will be considered whenever the standard deviation differences pertaining the mean difference is not much for two sampled hypothesis estimations.

# **8.** **Analyses**

## **8.1 Pre-Process Analysis**

The normality of ‘ABV’, ‘OG’, ‘FG’, ‘BoilGravity’, ‘Size(L)’ and ‘BoilSize’ of [1] and ‘beer\_abv’ of [2] is tested by Anderson - Darling test of data post-calibration.

## **8.2 Primary Analysis**

The primary outcomes will be analyzed as follows:

1. In the brewing process, the density of the fluid is measured in three different stages:
2. Original Gravity: the density before brewing
3. Boil Gravity: the density after boil stage
4. Final Gravity: the density after brewing

The Columns ‘OG’, ‘BoilGravity’ and ‘FG’ in [1] provides readings as mentioned above. It is essential to note that the density has to reduce significantly in brewing process. The significance of differences is tested by One-way ANOVA. The hypothesis is bounded within 95% of the confidence interval.

1. In the brewing process, the size of the container is measured initially and then it is noted after the addition of other ingredients before the boil stage. To check if there is significant gain in fluid size before boil, a calculated column from the difference in values of ‘Size(L)’ and ‘BoilSize’ in [1] as ‘Difference\_Size’ is analyzed by Paired t test. The hypothesis is bounded within 95% of the confidence interval.
2. Alcohol by Volume is the link between [1] and [2]. It needs to be verified that the distribution of alcohol content, i.e. ‘ABV’ in [1] and ‘beer\_abv’ in [2] are different to distinguish them as homebrewed and commercial beer data by using two sample z test (large population). The hypothesis is bounded within 95% of the confidence interval.
3. All Grain method is the major technique used for home brewing. But does each beer type have equal preference of choosing All Grain method instead of other methods? This has to be tested by Chi-square test of association for each beer style in [1]. All the other brewing methods have been merged into a single category called ‘Other’ as some counts were less than 5 in each method, so it has to be merged to qualify for the test. The hypothesis is bounded within 95% of the confidence interval.
4. Ales are harder to distinguish on the basis of its bitterness. So to check if there is any significant difference in the bitterness units i.e. ‘IBU’ in [1] by using Chi-square test of association for Irish Red Ale and Irish Pale Ale. To make this data qualify for Chi-square the continuous values are categorized into count intervals of ‘1 to 20’, ’21 to 40’, ’41 to 60’ and ’61 to 80.’. The hypothesis is bounded within 95% of the confidence interval.
5. Dry Stouts and Extra Stouts have very minute differences in their features generally. So to check if there is any significant difference in the alcohol by volume content i.e. ‘ABV’ in [1] by using 2-sample t test for Irish Dry Stout and Irish Extra Stout. The hypothesis is bounded within 95% of the confidence interval.

## **8.3 Secondary Analysis**

To determine the Pearson correlation scores among each feature ratings with the overall ratings in [2] to identify the impact of each feature rating in the overall rating. This score would be used to generate calculated weighted averages which along with the overall ratings will help to find the best alcohol content to find the best beer in each beer style of [1] to be finally included in the homebrewed beer bars.

# **9. Reporting Conventions**

P-values ≥ 0.001 will be reported to 3 decimal places; p-values less than 0.001 will be reported as “< 0.001”. The mean, standard deviation, and any other statistics other than quantiles, will be reported to one decimal place greater than the original data. Quantiles, such as median, or minimum and maximum will use the same number of decimal places as the original data.

# **10. References**

*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 - Homebrew Beer Recipes, Calculators, and Forum* (no date). Available at: https://www.brewersfriend.com/ (Accessed: 28 November 2019).

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

# **Appendix A**

**Listing of Tables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table\_ID** | **Table Title** | **Dataset** | **Endpoints** | **Summary Statistics** |
| 1 | Anderson – Darling Normality test on ‘ABV’, ‘OG’, ‘BoilGravity’, ‘FG’, ‘Size(L)’, ‘BoilSize’ and ‘beer\_abv’ | [1] and [2] | pre-process | Probability Plot |
| 2 | One-Way ANOVA test on ‘OG’, ‘FG’ and ‘BoilGravity’ | [1] | i. | Box Plot |
| 3 | Paired t test on ‘Difference Size’ | [1] | ii. | Box Plot |
| 4 | Two sample z test between ‘ABV’ and ‘beer\_abv’ | [1] and [2] | iii. | Graphical Summary |
| 5 | Chi-square test of association on ‘BrewMethods’ by Beer Styles | [1] | iv. | Pie Chart |
| 6 | Chi-square test of association on ‘IBU’ by Ales | [1] | v. | Bar Chart |
| 7 | 2-sample t test on ‘ABV’ by Stouts | [1] | vi. | Scatter Plot |
| 8 | Pearson correlation of ‘review\_taste’, ‘review\_palate’, ‘review\_appearance’ and ‘review\_aroma’ with ‘review\_overall’ | [2] | secondary | Matrix Plot |

# **Appendix B**

**Derived and Computed Variables**

1. In [1], a new calculated column has been derived as the difference between ‘Size(L)’ and ‘BoilSize’ columns as follows:

**Difference\_Size = ‘BoilSize’ – ‘Size(L)’**

1. In [2], a new calculated column has been derived as the weighted average of correlation coefficient of ‘review\_taste’, ‘review\_palate’, ‘review\_taste’ and ‘review\_appearance’ with ‘review\_overall’ as follows:

Let ‘C\_t’ be correlation factor between ‘review\_taste’ and ‘review\_overall’

Let ‘C\_ap’ be correlation factor between ‘review\_apperance’ and ‘review\_overall’

Let ‘C\_ar’ be correlation factor between ‘review\_aroma’ and ‘review\_overall’

Let ‘C\_p’ be correlation factor between ‘review\_palate’ and ‘review\_overall’

Let ‘C+’ be the sum of all correlation scores

So,

**review\_average\_weighted**

**= ((C\_t\*’review\_taste’) + (C\_ap\*’review\_appearance’) + (C\_ar\*’review\_aroma’) + (C\_p\*’review\_palate’))**

**/ C+**

# **Appendix C**

**Data Dictionary**

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 |
| Style | Category | - | 6 Irish Styles | Type of Brew |
| Size(L) | Integer | Litres | 1 - 200 | Amount brewed for recipe listed |
| OG | Decimal | Unitless | 1.0 - 1.1  (4 Deci) | Specific gravity of wort before fermentation |
| FG | Decimal | Unitless | 1.0 - 1.02  (4 Deci) | Specific gravity of wort after fermentation |
| ABV | Decimal | V/V% | 1 - 10  (2 Deci) | % alcohol by volume |
| IBU | Decimal | Numeric | 1 - 80  (2 Deci) | International bittering units |
| Color | Decimal | Numeric | 1 - 50  (2 Deci) | Color units by Standard Reference Method |
| BoilSize | Decimal | Litres | 10 - 250  (2 Deci) | Fluid at beginning of boil |
| BoilTime | Integer |  | 50 - 100 | Time wort is boiled |
| BoilGravity | Decimal | Unitless | 1.02 -1.06  (4 Deci) | Specific gravity of wort before boil |
| Efficiency | Decimal | Percentage | 60 - 85 | Efficiency in extracting sugars from the grain during mash |
| BrewMethod | Category | - | 4 Methods | Various techniques for brewing |
| Difference\_Size | Decimal | Litres | Any | Calculated difference of ‘BoilSize’ – ‘Size(L)’ |

**ii.** **Beer Reviews Dataset**

(*Beer Reviews | Kaggle*, no date)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Measurement Units** | **Allowed Values** | **Description** |
| 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 |
| 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 |
| review\_average\_weighted | Decimal | Numeric | 1 – 5 | Calculated average from the Pearson correlation among each feature to overall |