**Faculty of Business and Hospitality**

|  |
| --- |
| 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 3, Final Report** |
| |  |  |  | | --- | --- | --- | | **Yash Sinojia** | **11/29/19** | **Data Analytics** | |

**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 3, Final Report**

**CONTENTS**

[**1.** **Revision History** 3](#_Toc25949103)

[**2.** **Abbreviations** 4](#_Toc25949104)

[**3.** **Introduction** 4](#_Toc25949105)

[**4.** **Normality Tests** 5](#_Toc25949106)

[**5.** **Descriptive Statistics** 8](#_Toc25949107)

[**5.1** **Boxplot of ‘OG’, ‘FG’ and ‘BoilGravity’** 8](#_Toc25949108)

[**5.2** **Boxplot of ‘Difference\_Size’** 8](#_Toc25949109)

[**5.3** **Graphical Summary for ‘beer\_abv’ and ‘ABV’** 9](#_Toc25949110)

[**5.4** **Pie Chart on ‘BrewMethods’** 10](#_Toc25949111)

[**5.5** **Stacked Bar Chart on ‘IBU’** 11](#_Toc25949112)

[**5.6** **Boxplot on ‘ABV’** 12](#_Toc25949113)

[**6.** **Statistical Tests** 12](#_Toc25949114)

[**6.1** **One-Way ANOVA Test** 12](#_Toc25949115)

[**6.2** **Paired-t Test** 14](#_Toc25949116)

[**6.3** **2-Sample z Test** 15](#_Toc25949117)

[**6.4** **Chi-square test for Association** 16](#_Toc25949118)

[**6.4.1**  **On Brew Methods** 16](#_Toc25949119)

[**6.4.2**  **On IBU** 16](#_Toc25949120)

[**6.5** **2-Sample t Test** 17](#_Toc25949121)

[**6.6** **Person Correlations** 18](#_Toc25949122)

[**7.** **Interpretation of Results** 21](#_Toc25949123)

[**7.1** **One-Way ANOVA Test** 21](#_Toc25949124)

[**7.2** **Paired-t Test** 21](#_Toc25949125)

[**7.3** **2-Sample z Test** 22](#_Toc25949126)

[**7.4** **Chi-square test for Association** 23](#_Toc25949127)

[**7.4.1** **On Brew Methods** 23](#_Toc25949128)

[**7.4.2** **On IBU** 23](#_Toc25949129)

[**7.5** **2-Sample t Test** 24](#_Toc25949130)

[**7.6** **Pearson Correlations** 24](#_Toc25949131)

[**8.** **References** 25](#_Toc25949132)

[**Appendix A** 26](#_Toc25949133)

[**Appendix B** 27](#_Toc25949134)

[**Appendix C** 28](#_Toc25949135)

# 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

Looking back to Part 2 of the project in retrospection, the statistical analysis plan was formulated. In this plan several statistical tests were defined as the established endpoints to meet the objectives of this project as follows:

1. Anderson - Darling Normality tests

2. One-way ANOVA test

3. Paired t test

4. 2-Sample z test

5. Chi-square tests for Association

6. 2-Sample t test

7. Pearson Correlations

Please find the Listing of Table in Appendix A for more details and changes.

# Abbreviations

|  |  |
| --- | --- |
| Abbreviation | Stands for |
| [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) |

# **Introduction**

Pertaining from the Statistical Analysis Plan, a listing of table had been described at the end part of the previous report that proposed a visual structure for this final part. That included various hypotheses tests, a normality test and a correlation test. Each variable involved in these tests will be subjected to a descriptive statistic elaborated by a suitable visualization as a preliminary analysis before running the tests. And at the end of all the analyses from tests, the interpretation of all the results will be briefed.

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

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

Statistical Analysis Plan, 28th November, 2019.

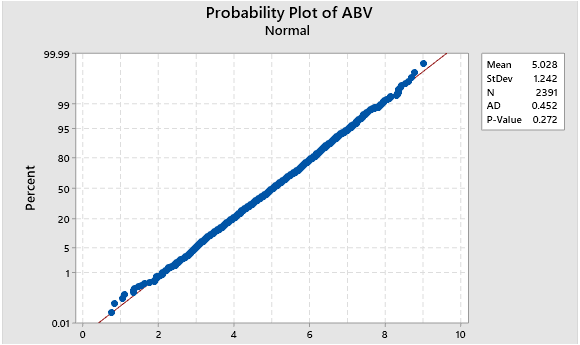
The reader of this report is encouraged to also read the Client – Analyst – Observer Interview and Statistical Analysis Plan for complete clarification and details on Business Problem Statement and the entire statistical plan for this report.

# **Normality Tests**

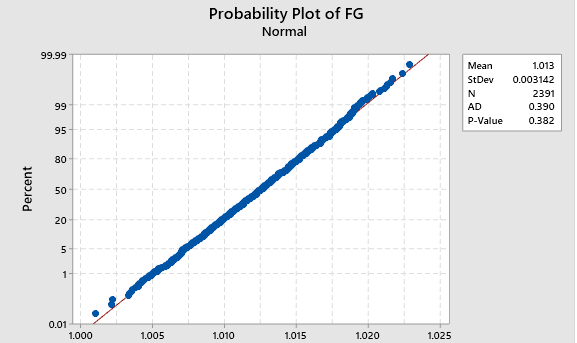
It is essential to start the analyses by doing the Anderson - Darling Normality test in prior on the variables involved in later stage of the analyses.

Anderson - Darling test assumes that there are not any parameters to be estimated in the distribution being tested, in short, the test and its set of critical values is distribution-free (*Anderson–Darling test - Wikipedia*, no date).

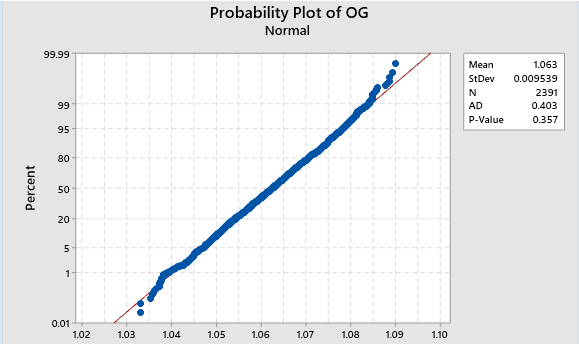
The results of this test on the datapoints of ‘ABV’, ‘OG’, ‘FG’, ‘BoilGravity’, ‘Size(L)’, ‘BoilSize’, and ‘beer\_abv’ as mentioned in Table 1 are as follows:



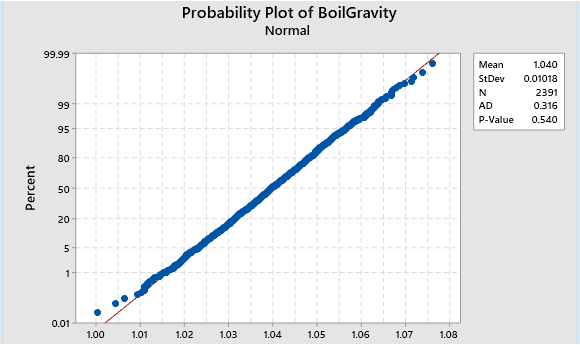
***Fig 1: Probability Plot of ABV***



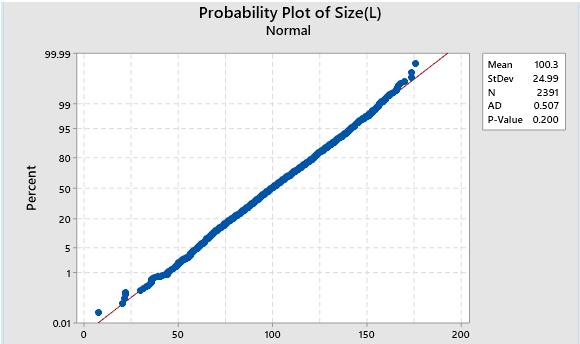
***Fig 2: Probability Plot of FG***



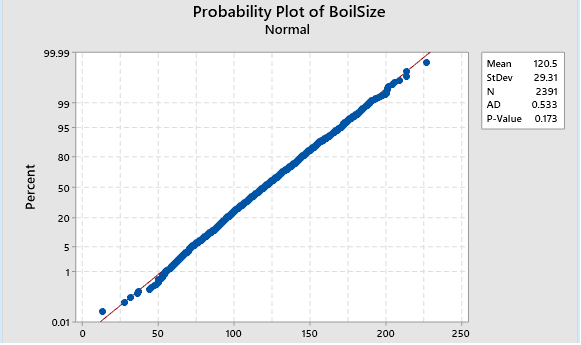
***Fig 3: Probability Plot of OG***



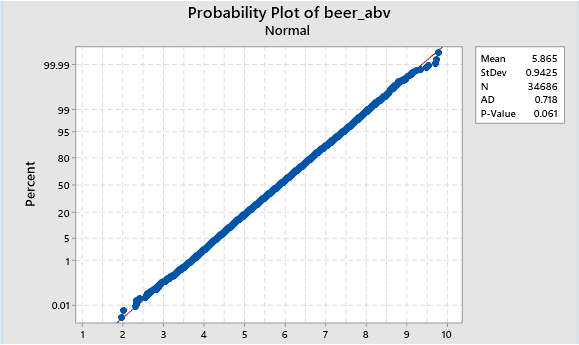
***Fig 4: Probability Plot of BoilGravity***



***Fig 5: Probability Plot of Size(L)***



***Fig 6: Probability Plot of BoilSize***



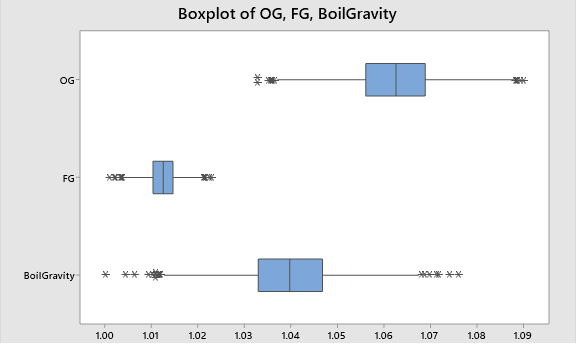
***Fig 7: Probability Plot of beer\_abv***

The above test outputs clearly describe that the P-values of each variable are greater than 0.05 which proves that the data of those variables pass the Anderson – Darling Normality test and their data points fit the normal curve. Henceforth the data in these variables qualify for the further statistical tests.

# **Descriptive Statistics**

*(See Appendix A)*

## **Boxplot of ‘OG’, ‘FG’ and ‘BoilGravity’**



***Fig 8: Boxplot of ‘OG’, ‘FG’ and ‘BoilGravity’***

‘OG’, ‘FG’ and ‘BoilGravity’ columns have continuous values best described by boxplot above. The major range of values seems fairly scattered enough among the three columns. However, there are a set of outliers visible as star marks in each column. Also, the range of values are very narrow, i.e. from 1.0 to 1.1 and hence the difference though appearing significant here, may be actually minute. So, to compensate that cynicism, it has to be verified further from a statistical test.

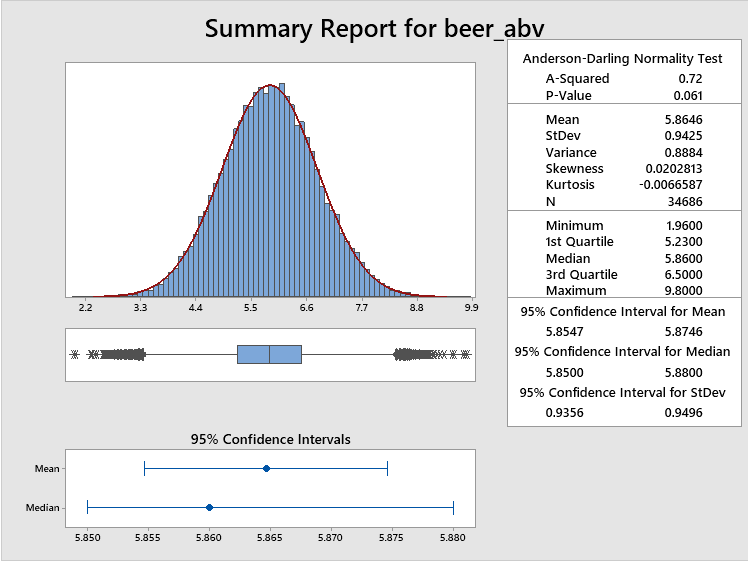
## **5.2 Boxplot of ‘Difference\_Size’**



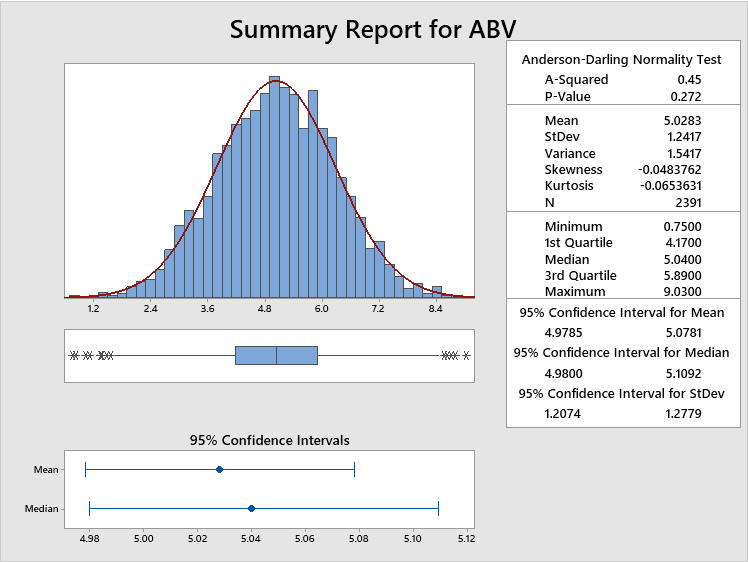
***Fig 9: Boxplot of Difference\_Size***

It is clearly noticeable that the boxplot of difference between ‘BoilSize’ and ‘Size(L)’, i.e. ‘Difference\_Size’ is completely towards higher positive values. That proves that there has been an increase in size from ‘Size(L)’ to ‘BoilSize’. The magnitude of this change will be elaborated further from the results of a statistical test.

## **5.3** **Graphical Summary for ‘beer\_abv’ and ‘ABV’**



***Fig 10: Graphical Summary for ‘beer\_abv’***



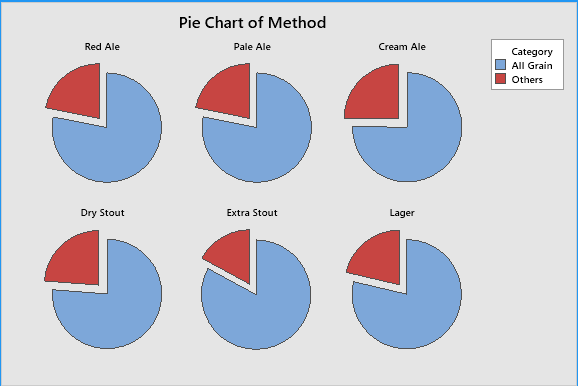
***Fig 11: Graphical Summary for ‘ABV’***

As there is a huge difference in the population of ‘beer\_abv’ and ‘ABV’ columns, it can be best described here by comparing the graphical summaries of both.

Though both the graphs are normally distributed, it can be noticed that the data in ‘beer\_abv’ is a bit right skewed while the data in ‘ABV’ is a bit left skewed. But the magnitude of this skewedness is minute.

The mean and median values in ‘beer\_abv’ are slightly higher than in ‘ABV’. This difference will be tested further by a statistical test.

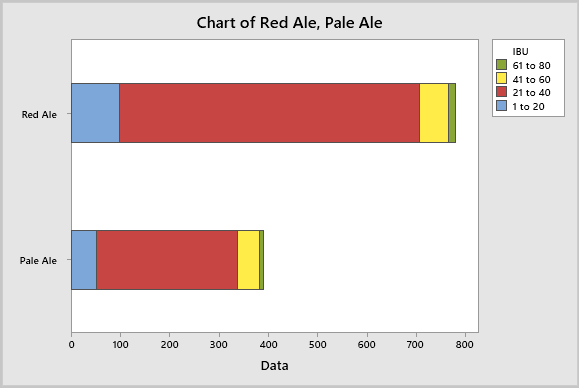
## **5.4 Pie Chart of ‘BrewMethods’**



***Fig 12: Pie Charts of Brew Methods by Beer Styles***

These pie-chart shows that there is some similarity in the ratios of ‘All Grain’ method vs. Other methods among each beer style. However, the significance of this similarity has to be backed up by a statistical test.

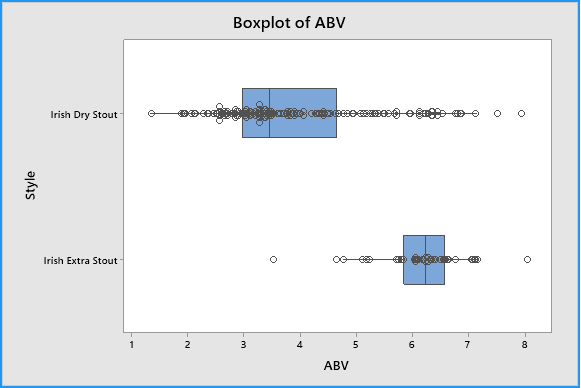
## **5.5 Stacked Bar Chart of ‘IBU’**



***Fig 13: Stacked Bar Chart of IBU by Ales***

It is noticeable with the stacked bar plots that there is some similarity in distribution of data ranges ‘1 to 20’, ’41 to 60’ and ’61 to 80’ between Red and Pale Ale. However, the ratio of range ’21 to 40’ seems different. This similarity and difference will be further evaluated by a statistical test.

## **5.6 Boxplot of ‘ABV’**



***Fig 14: Boxplot of ABV by Stouts***

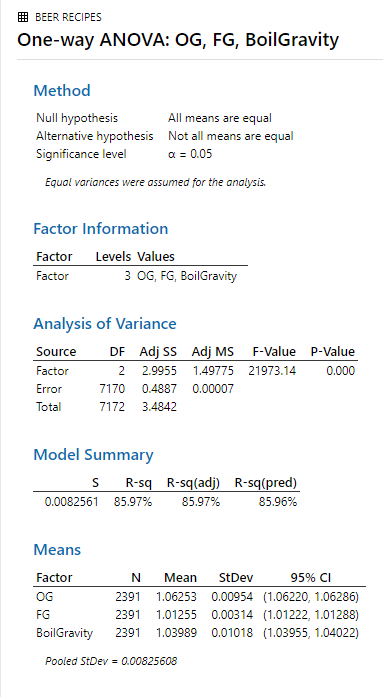
The box plots clearly describe that there is a difference in the ‘ABV’ between Dry Stouts and Extra Stouts. However, there being a lot of outliers, this hypothesis will be tested further with a statistical test.

# **Statistical Tests**

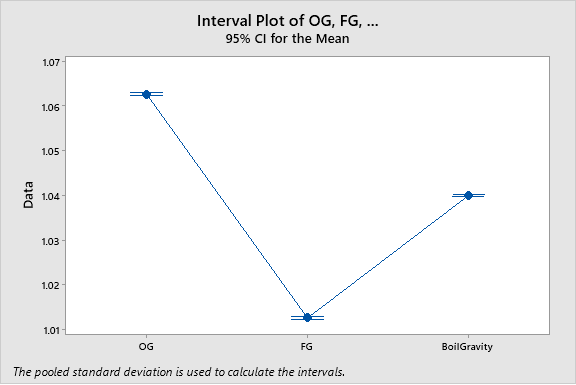
(*See Appendix A*)

## **One-Way ANOVA Test**

It is essential to note that the density in ‘OG’, ‘BoilGravity’ and ‘FG’ respectively 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.



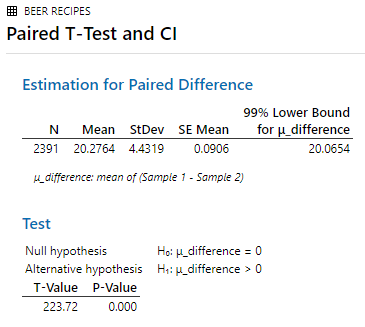
***Fig 15: One-way ANOVA: OG, FG, BoilGravity***



***Fig 16: Interval Plot with ANOVA test***

## **Paired-t Test**

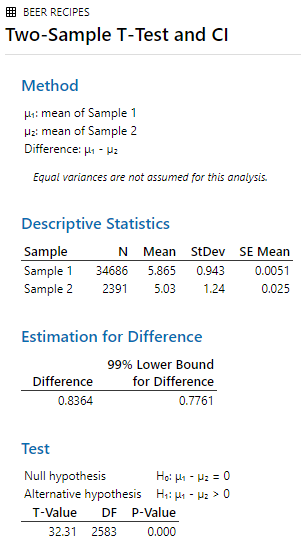
To check if there is significant gain in fluid size before boil, ‘Difference\_Size’ from the difference in values of ‘Size(L)’ and ‘BoilSize’ in [1] as is analyzed by Paired t test. The hypothesis is bounded within 95% of the confidence interval.



***Fig 17: Paired T-Test and CI***

## **6.3 2-Sample z Test**

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.



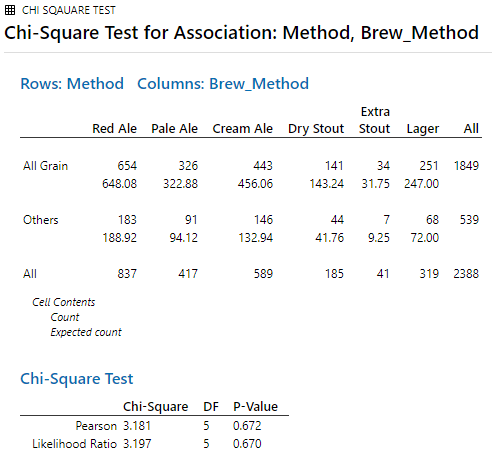
***Fig 18: Two-Sample T-Test and CI***

There is no separate test for 2-sample z in Minitab. Minitab naturally assumes the 2-sample t test with population greater than 30 in each sample as a 2-sample z test.

## **6.4 Chi-square test for Association**

### **6.4.1 On Brew Methods**

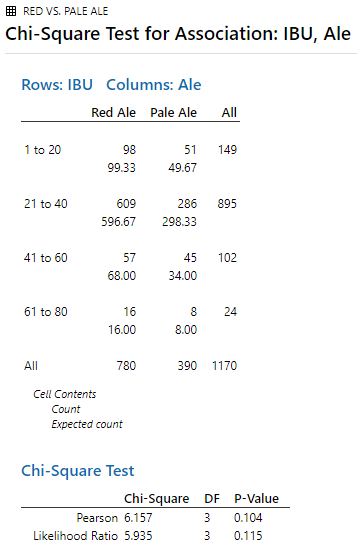
To compare the ratio of ‘All Grain’ method as the major brewing technique for homebrewed beers with all the other methods, it is tested with Chi-square test for association. The hypothesis is bounded by 95% confidence interval.



***Fig 19: Chi-square Test for Association: Brew\_Method, Styles***

### **6.4.2 On IBU**

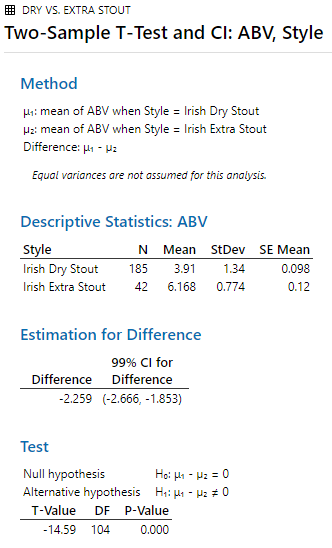
To check if there is any significant difference in the bitterness units i.e. ‘IBU’ in [1] by using Chi-square test of association for Irish Red Ale and Irish Pale Ale. The hypothesis is bounded within 95% of the confidence interval.



***Fig 20: Chi-square Test for Association: IBU, Ale***

## **6.5 2-Sample t Test**

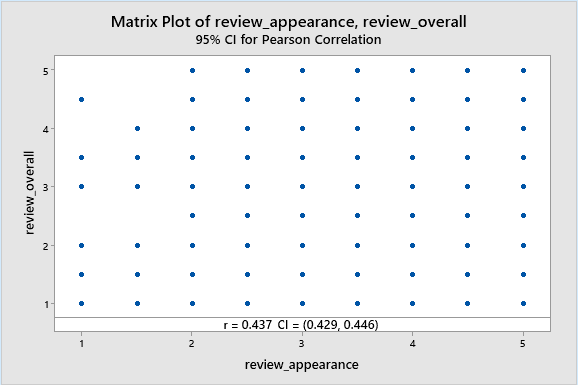
To check if there is any significant difference in the alcohol by volume content i.e. ‘ABV’ in [1] by using 2-sample t test for Irish Dry Stout and Irish Extra Stout. The hypothesis is bounded within 95% of the confidence interval.



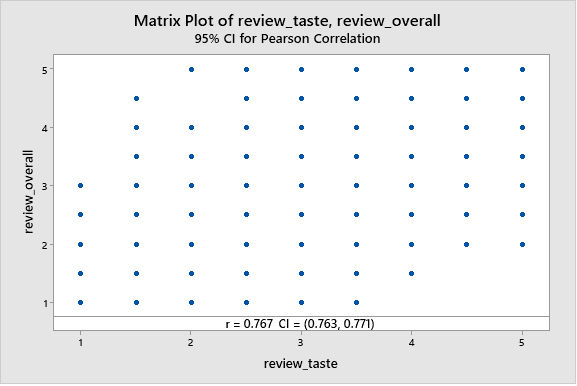
***Fig 21: Two-Sample T-Test and CI: ABV, Style***

## **6.6 Person Correlations**

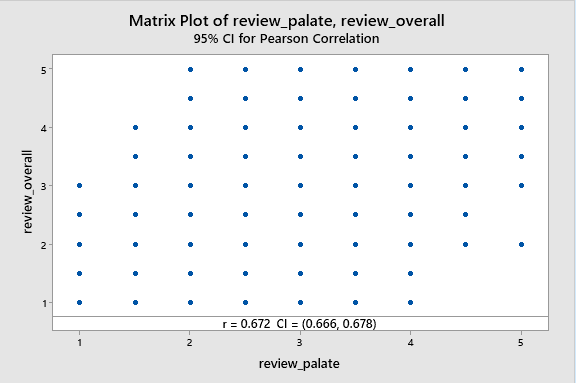
To determine the Pearson correlation scores among each feature ratings with the overall ratings in [2] to identify the impact of each feature rating in the overall rating. This score would be used to generate calculated weighted averages which along with the overall ratings will help to find the best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



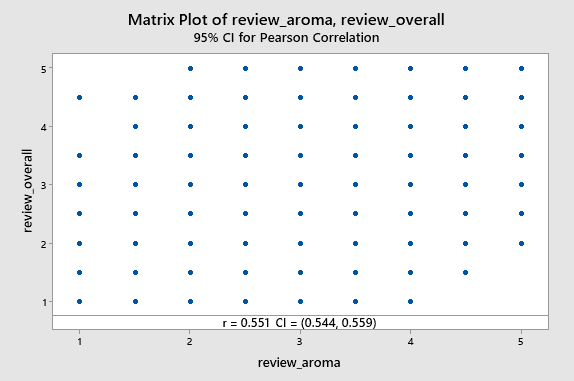
***Fig 22: Pearson correlation for review\_appearance with review\_overall***



***Fig 23: Pearson correlation for review\_taste with review\_overall***



***Fig 24: Pearson correlation for review\_palate with review\_overall***



***Fig 24: Pearson correlation for review\_aroma with review\_overall***

# **Interpretation of Results**

*(See Appendix A)*

## **One-Way ANOVA Test**

**H0**: Means of ‘OG’, ‘FG’ and ‘BoilGravity’ are equal (*Null hypothesis*)

**HA**: Any two means among ‘OG’, ‘FG’ and ‘BoilGravity’ are different (*Alternative hypothesis*)

Noting from the Analysis of Variance table, the P-Value is less than 0.05

Here, Fstatistic is not in range (Fcritical- ,Fcritical+)

Therefore, the statistic value falls in the rejection region of 95% significance

And hence we reject the null.

So there is no evidence that means of ‘OG’, ‘FG’ and ‘BoilGravity’ are equal.

Which proves that there is a significant difference among values in ‘OG’, ‘FG’ and ‘BoilGravity’ Column by two-tailed One-Way ANOVA test. The means table shows the 95% of confidence interval for each variable proving that the range of values are far from any intersection of bounds and hence different.

Moreover, the interval plot at the end supports the alternative hypothesis. It also supports the scientific and practical fact that the density of liquid reduces in these three stages of the brewing process.

## **7.2 Paired-t Test**

**H0**: Mean difference between the sizes is 0 (*Null hypothesis*)

**HA**: Mean difference between the sizes is greater than zero (*Alternative hypothesis*)

Noting from the test table, the P-Value is less than 0.05

Here, tstatistic > tcritical

Therefore, the statistic value falls in the rejection region of 99% significance

And hence we reject the null.

So there is sufficient evidence that the mean difference in ‘Difference\_Size’ is greater than zero.

The t-value is a lot higher. Which proves that there is a significant increase of the fluid size in litre from initial stage to boil stage as per right-tailed Paired-t test.

## **7.3 2-Sample z Test**

**H0:** Mean ’beer\_abv’ is less than or equal to Mean 'ABV’ (*Null hypothesis*)

**HA:** Mean ‘beer\_abv’ is greater than Mean ‘ABV’ (*Alternative hypothesis*)

Noting from the test table, the P-Value is less than 0.05

Here, tstatistic > tcritical (actually, zstatistic > zcritical)

Therefore, the statistic value falls in the rejection region of 99% significance

And hence we reject the null.

Though the standard deviations are higher than the difference, the standard error of the mean is much lower comparatively. And from the analysis of right-tailed two-sampled z test, the values in ‘ABV’ on average, are higher than values in ‘beer\_abv’.

## **7.4 Chi-square test for Association**

### **7.4.1 On Brew Methods**

**H0:** There is no difference in the ratios of Brew Method among Beer Styles (*Null hypothesis*)

**HA:** There is a difference in at least two of the ratios of Brew Method among Beer Styles (*Alternative hypothesis*)

Noting from the Chi-square test table, the P-Value is much greater than 0.05

Here, ꭓ2- < ꭓ2statistic  < ꭓ2+

Therefore, the statistic value falls in the non-rejection region of 95% significance

And hence we do not reject the null.

Henceforth there is no evidence that there is any association of Brew Methods on the basis of Beer Styles by two-tailed Chi-square test. And hence all the Beer Styles have similar ratio for brewing on ‘All Grain’ method vs. any other method.

### **7.4.2 On IBU**

**H0:** There is no difference in the distribution of IBU among Ale Styles (*Null hypothesis*)

**HA:** There is a difference in the distribution of IBU among Ale Styles (*Alternative hypothesis*)

Noting from the Chi-square test table, the P-Value is much greater than 0.05

Here, ꭓ2- < ꭓ2statistic  < ꭓ2+

Therefore, the statistic value falls in the non-rejection region of 95% significance

And hence we do not reject the null.

Henceforth there is no evidence that there is any association of IBU distribution on the basis of Ale Styles by two-tailed Chi-square test. And hence it is proved that on average, the IBU values are similar in Red Ale and Pale Ale.

## **7.5 2-Sample t Test**

**H0:** Mean ’ABV by Dry Stouts’ is equal to Mean 'ABV by Extra Stouts’ (*Null hypothesis*)

**HA:** Mean ‘ABV by Dry Stouts’ is not equal to Mean ‘ABV by Extra Stouts’ (*Alternative hypothesis*)

Noting from the test table, the P-Value is less than 0.05

Here, tstatistic is not in range (tcritical- , tcritical+)

Therefore, the statistic value falls in the rejection region of 99% significance

And hence we reject the null.

The estimation of difference is negative and the value zero does not fall in the 99% confidence interval. And henceforth, there is a significant difference of alcohol contents in Dry Stouts and Extra Stouts.

## **7.6 Pearson Correlations**

The Pearson correlation weights have been used to generate a calculated column called ‘review\_average\_weighted’ as described in Statistical Analysis Plan.

The following points were followed to select the best seven beer recipes from [1]:

1. Only 5-star reviews from ‘review\_average\_weighted’ and ‘review\_overall’ are extracted in [2].
2. The data is then divided on the basis of each Beer Styles in [2].
3. For each beer style the mean of ‘beer\_abv’ is calculated in [2].
4. These mean values describe the best abv values to be considered in [1].
5. From these mean values these six beer recipes are shortlisted to be finally included in the business

The seven recipes are as follows:

* **Rublin Ale (Irish Red Ale)**
* **Friends of Pale (Irish Pale Ale)**
* **Smithwick’s Cream Ale (Irish Cream Ale)**
* **Sore Hands Milk Stout (Irish Dry Stout)**
* **Jabeerini’s Irish Stout (Irish Dry Stout)**
* **Irish Black Hag (Irish Extra Stout)**
* **Yardsman’s Lager (Irish Lager)**

# **References**

*Anderson–Darling test - Wikipedia* (no date). Available at: https://en.wikipedia.org/wiki/Anderson–Darling\_test (Accessed: 29 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).

# **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. | Box 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 |