**Data Preparation and Descriptive Analytics**

1-

In the beer: I had to first rearrange the columns to show the ID, Name, brewer\_ID , abv, ibu, style and ounces.

There are numbers in the name of the beer, there are also nulls in the abv (62 rows) and the ibu (1005 rows) and in style (5 rows) also the ID for beers were not arranged from smaller to bigger.

In the brewery: the IDs needed to be arranged again from smaller to bigger. And the brewery id needed to add to it underscore sign and the states were not matching a same format and had different style.

2-

In the beer: there are 7 attributes  
in the brewery: there are 3 attributes

It seems that they are connected by the brewery\_id

3-

For the brewery:

Had to split the location to city and state, also had to group some of the states together as there were shorten and the name was full so had to group them by the short term for states.

For the beer I had to remove the punctuation and extra spaces from the name field

4-

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

I Have used join on berewery id and its useful link SQL to bring the 2 database with a relation to get more info and meaning from the data.

5-

I think I have used the sum of rows aggregated with the group to be beer name

A better measure to use when removing duplicates is typically the "COUNT" aggregation. This will count the number of occurrences of each unique beer in your dataset. By using "COUNT,"

6-

For me I have some duplicates I have 11 aggregated fields with 262 rows

For example dales pale ale beer with number of aggregated rows 26

I have now 1908 unique values

8-

2907 rows

9-

Avg across: 28.39246543

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Average of abv** | **Average of ibu** |
| American IPA | 0.072738292 | 54.95592287 |
| Hefeweizen | 0.057342857 | 13.57142857 |
| Oatmeal Stout | 0.067125 | 19.3125 |
| Wheat Ale | 0.06 | 24 |
| **Grand Total** | **0.071192771** | **50.01686747** |

It appears that the IBU in American IPA is extremely higher while its lower in Hefeweizen

Its also appears that the IBU is the significant factor more than the ABV

10-

|  |  |  |
| --- | --- | --- |
|  | *abv* | *ibu* |
| abv | 1 |  |
| ibu | 0.602331 | 1 |

It seams there is a medium correlation between ABV and IBU with 0.6 score from 1.0

11-

I think I have deleted some of the nulls not in the IBU or the ABV as I have left them but I think during the prosses I got rid of some of the null rows for different attributes and that of course will impact the averages results and the correlation as well. How big is the impact I am not sure depending on how many entries deleted and there score for both attributes.

I also tried to fill in the excel sheet the nulls in the IBU with 0’s instead.