Dataset Used:
Coffee Quality database
The datasets are gathered from Coffee Quality Institute (CQI) in January, 2018. The data contain reviews of 1312 arabica and 28 robusta coffee beans from the Coffee Quality Institute's trained reviewers.
The datasets can be found using this link: https://www.kaggle.com/ankurchavda/coffee-beans-reviews-by-coffee-quality-institute

**Background Information & Definitions**

The dataset as mentioned contains reviews from the trained reviewers of the Coffee Quality Institute. It contains the data from 1340 coffee beans from across the world. 1312 of them belong to the Arabica specie while 28 are of the robusta variety.

The data set records the metadata about the quality measure, bean itself and its growing. The quality measures include aroma, flavour, aftertaste, acidity, body, balance, uniformity, cup cleanliness, sweetness, moisture and defects.

The defects are described as quality one defects, quality two defects and quakers. A primary defect would be something like a 'full sour' bean, where the coffee has been overly fermented, or organic matter (like sticks) mixed with in the coffee, while a secondary defect would be something like a broken or insect-damaged bean. Generally, specialty Grade Coffee Beans has no primary defects, 0-3 full defects. Premium Grade Coffee Beans: Same as Grade 1 except maximum of 3 quakers, 0-8 full defects. Quakers are "immature beans, unripe.

The growing metadata records the region in which it is grown, the names of the farms, the altitude at which it is grown and in country partners among others.

Coffee brewing methods can have a great effect on the overall quality of the brew but as always, the most important factor is the raw material, more specifically in this case the coffee beans. This analysis attempts to empirically suggest the better coffee beans from a larger set that grantee a better brew.

**Business Understanding**

This research is (hypothetically) being conducted by a group of individuals who plan on starting a Gourmet Coffee House in a Hoboken. The city itself has a reputation for catering to gourmands for various culinary disciplines and coffee being one of the more critiqued areas, it is important that the selection of the beans which in turn affect the quality of the coffee is immaculate.

The paper is (hypothetically) presented by the group of individuals to investors to justify their coffee selection process backed by data science driven visualisations.

The dataset gives a good idea of the fact that most of the coffee produced in the world is of the Arabica specie. But any coffee house that caters to gourmands needs to have a variety that offers different flavour profiles to the customers. The dataset quantifies this by attaching a number value to the quality measure and then adding them up to get a Total Cupper Point score. The score itself may not be indicative of the quality of the coffee as the quality measure of one bean may skew the results. A more through analysis is done and presented.

The main questions that are being answered are:

- What kinds of Coffee beans are guaranteed to give a better brew?

- What kind of a processing method is preferred?
- Which Coffee beans must be ordered?
- Recommendations

**Data Preparation**

The dataset itself is a pretty concise one but it has a large number of unwanted columns that are not relevant for our analysis. These are manually removed. The datasets with the columns removed are submitted with the report. Another point in the preparation in that the datasets did not have a serial number column which was added to make the modelling steps easier to perform.

A range of operations were performed on the dataset to derive a variety of plots; the specifics of which will be discussed in the Modeling phase.

From the lack of a lot of steps in the preparation phase, we can conclude that the datasets were quite clean save for a few issues, and as always, a **na.omit** function was use to ensure that the dataset being loaded into the program was complete.

**Modelling**

First to visualise the data we need to put it into the appropriate formats. For the scatterplots that plotted datapoints based on the mean growing altitude, there was a single point at 19000mts that was definitely a faulty value. To remove such erratic values, **mutate** function was used to only use datapoints which featured a range below 5000mts. This chart will be used to give insight about the preference of growing altitude.

For the choropleth plot, we did not have any data about the concentration of bean sources in each country. A separate data frame was prepared that involved counting the number of times each country featured in the dataset and then relating the number to the name of the country. This data frame was plotted using **plotly** and an interactive map was generated. The file is attached with the submission.

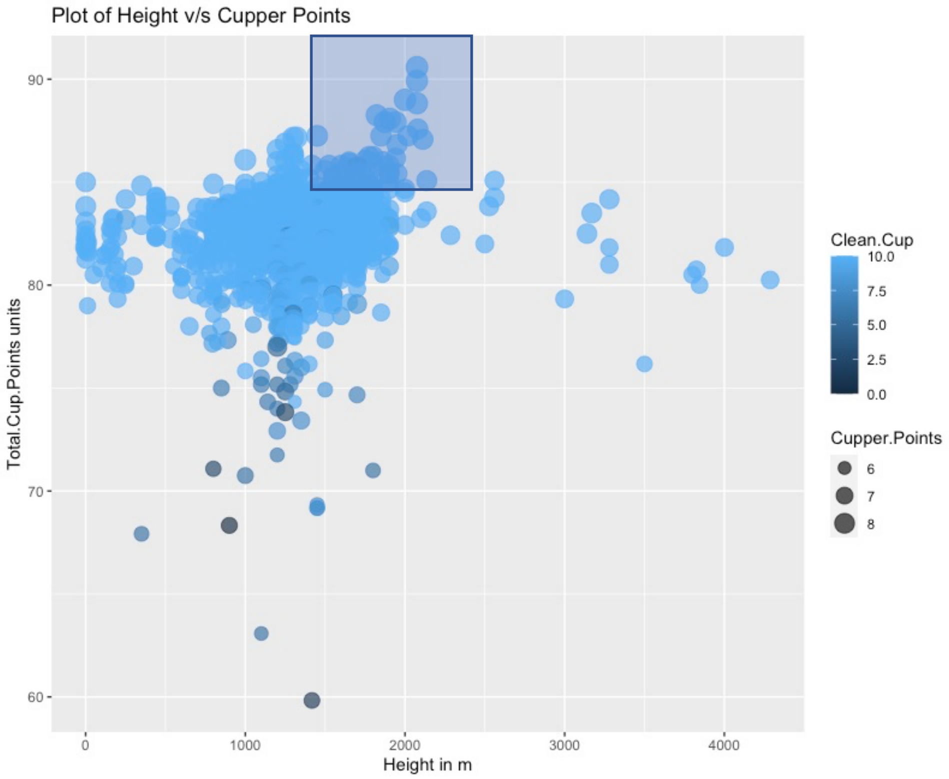
For the doughnut graph, which was used to contextualise the share of each processing method and also to make a decision on the kind of method to be used, table is generated to give the times each processing method is featured. The data from the table is used to manually prepare a data frame to plot the doughnut chart. From the same processing method, we generate a second scatterplot to give a preference for processing method.

Once the altitude and method have been assessed, the quality measure of the beans need to be analysed. For this, a subset is made of only the numerical quality measure and a correlation graph using **ggcorr** is generated.

Finally, to quantify the individual flavour profiles each of the bean species, radar plots are generated to provide a visual representation of what one can expect from each bean type.

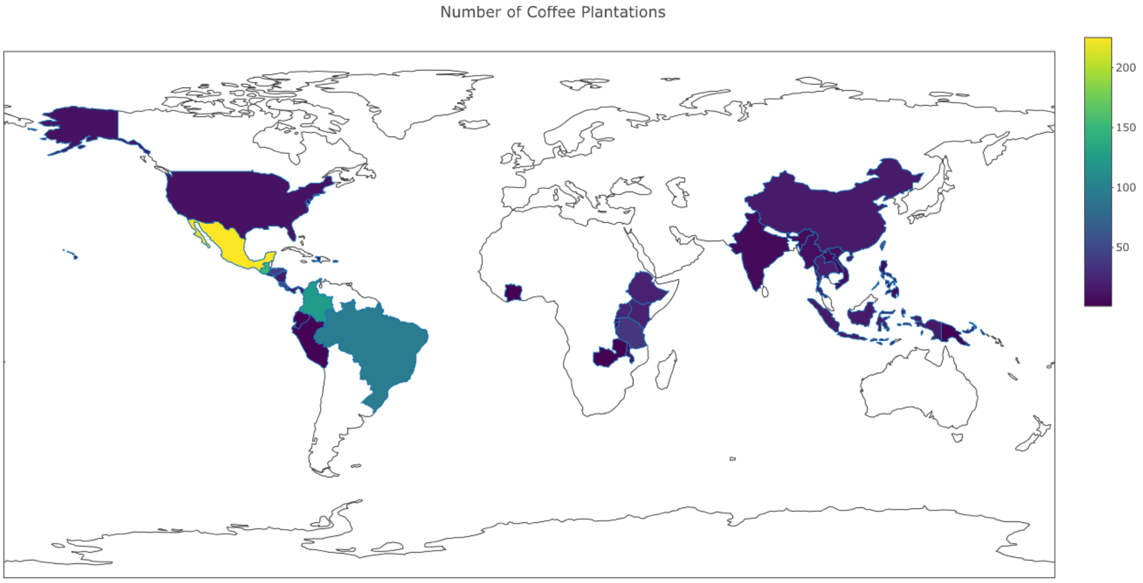
**Evaluation**

First the scatterplot was generated. It puts into context the height at which the coffee beans are grown against the Total cup points. The total cup points are the net indicator of quality of each bean. The size is a function of the cupper points and the color is a function of the clean cup.



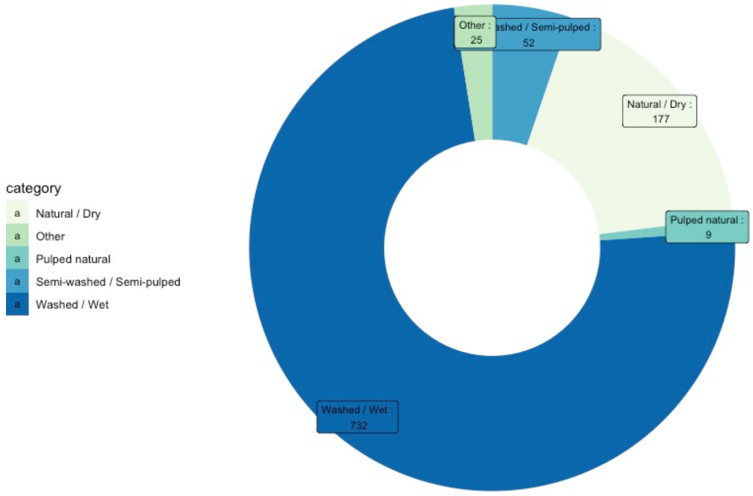
The graph above indicates that although was see a lot of beans grown at very high altitudes, it does not necessarily have an impact on the quality of coffee. From the above visualisation, the chosen beans must have an elevation in the range of 1500 and 2250 mts, while having a total cup points exceeding 85.

For the choropleth, we simply have a map that indicates the number of beans that can be sourced from each country of the world.



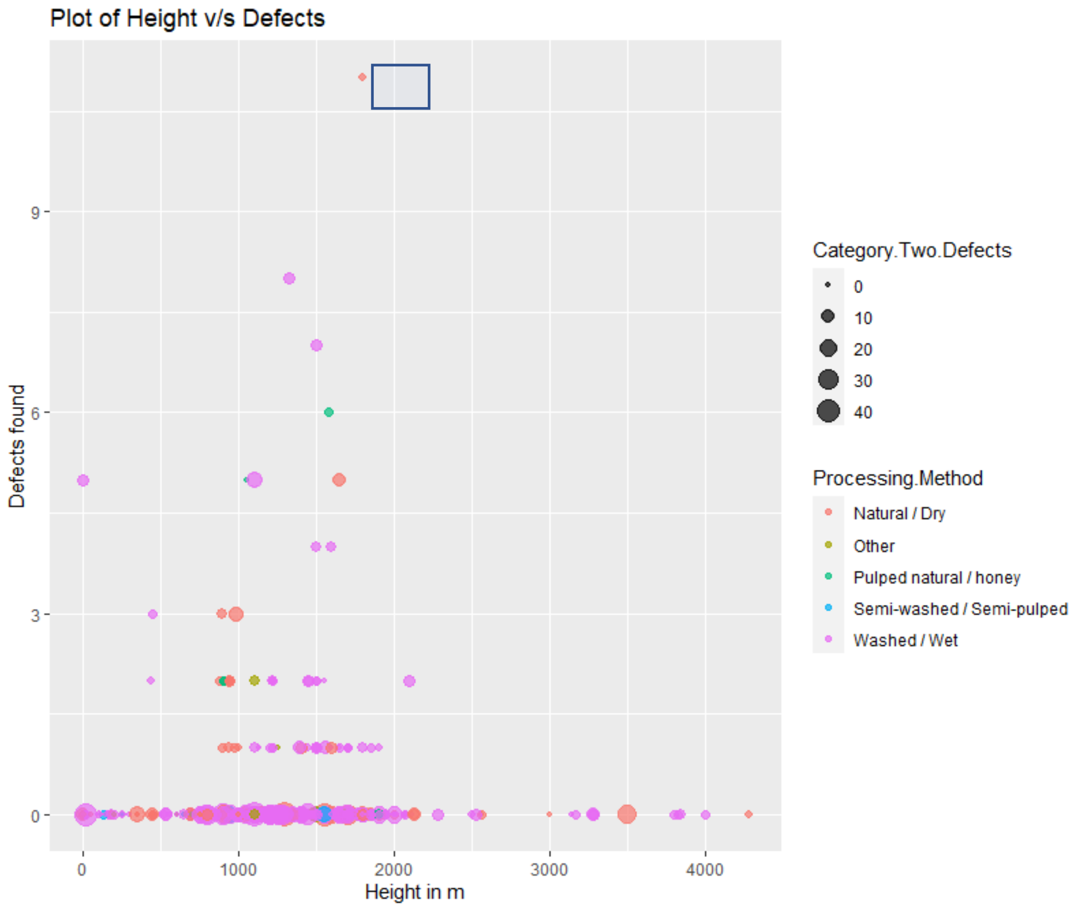
From the map above, we see that Mexico, Guatemala and Colombia source the largest number of beans, and it would be most effective to select a bean coming from these regions for economic reasons as they happen to be the closest geographically. Although thee countries are the largest producers, from a qualitative perspective, the country from which the beans are sourced are secondary. (Please do find a more interactive version of the map in the submitted folder).

For the doughnut map, a table of the processing methods involved were prepared and this was then used to make a data frame that was used to make the doughnut graph.



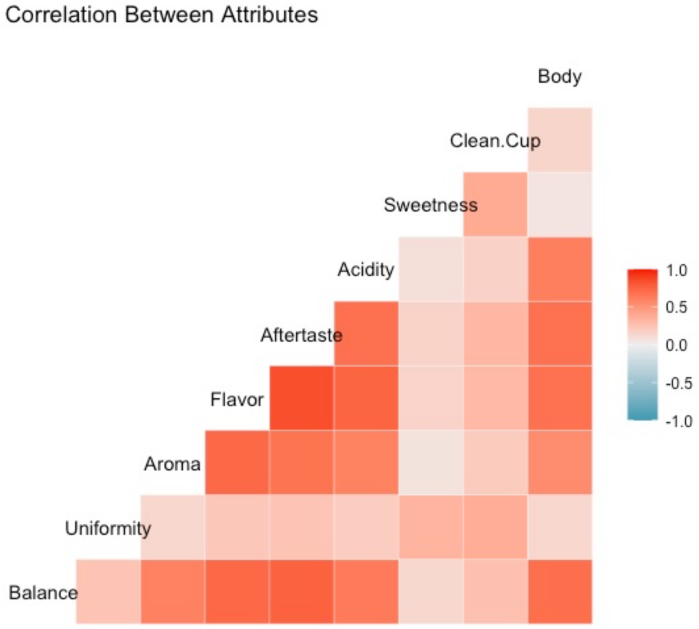
From the graph above, we can clearly see that a washed method is used most often. Using this data, we can further generate a scatterplot to indicate if the processing method does indeed have an effect on the overall produce quality.

For the second scatterplot, we use the height at which the beans are grown and the Quakers. The quakers are an umbrella term for any defects present in the bean. The colour is a function of the processing method and the size is a function of the category two defects.



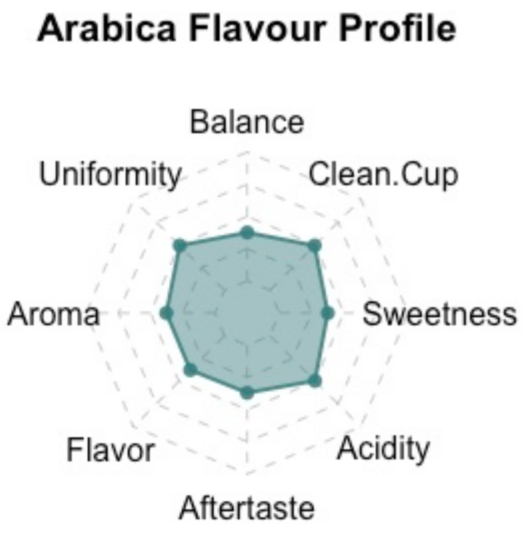
From the graph above, we can see that, the natural/dry method have the largest number of quakers. The washed method seems to have the largest number of zero defects and although the it seems to have a number of category two defects; this can be attributed to the larger number of beans that use this method. That being said, it is evident that washed/wet is the processing method that should be preferred.

To visualise which of the factors impact the overall score the most, we can choose a correlation plot to indicate the most closely related quality measure.

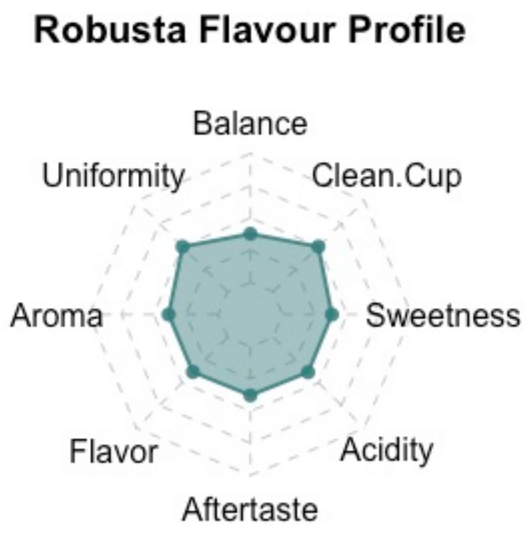


A number closer to 1 indicates that an increase in the value of one attribute will result in an increase in the other attribute too. In case of all of the factors in this correlation chart, it is clear that all of the quality measures have a positive correlation. More importantly, we see that “Flavour” and “Aftertaste” happen to be the most closely related factors. Hence beans with the highest Flavour and Aftertaste values must be preferred.

Finally, we can assess the flavour profiles of the beans themselves. The best chart for this visualisation is the radar or the spider chart. To generate these charts, we need to prepare the data frame of a matrix with the values we want to visualise. Once this is done, we simply need to use the **radarchart** to plot the figures we need.



Arabica Flavour Profile



Robusta Flavour Profile

From the charts above we see that arabica has a clearly more acidic flavour profile while robusta seems to show more sweetness. The display similar values for the flavour and aftertaste hence for the purpose of catering to different tastes, it is preferred that both beans are a part of the inventory. And simply due to the large number of arabica beans produced, it would be prudent to stock the inventory with that in mind.

**Result**

Finally after performing all the aforementioned analyses, we can make some conclusions on the nature of beans we must select.

- The beans selected must ideally be grown between an altitude of 1500 mts and 2250 mts,
- Must have a total cupper point of at least 85,
- Sourced from Mexico, Guatemala, Colombia or Ethiopia,
- Must be processed using a wet/washed method and have zero defects,
- Should have the highest possible Flavour and Aftertaste values.
- A number of arabica and a kind of robusta bean must be selected.

With all these requirements, we can use the mutate function to run a query to generate a set of recommendations.

After running these queries, we come up with around 3 recommendations to make that satisfy the Quality control requirements.

- Arabica beans from Ethiopia through METAD Agricultural Development plc which features a Total cup points of 90.58,
- Arabica beans from United States through Almacaf√© which features a Total cup points of 87.92 and
- Arabica beans from Uganda through Africa Fine Coffee Association which features a Total cup points of 86.00.

All of these beans happen to be Arabica. For the sake of variety, a robusta bean can be considered that has the highest possible Flavour and Aftertaste values.

- Robusta beans from Uganda through Ankole coffee producers coop which feature a total cupper point of 83.75.

Among these 4 it is clear that the robusta bean falls very short in terms of quality measure and it can be dropped in case the procurement is a bad choice to make.