**CHOCOLATE BAR RATINGS**

The assignment involved exploratory data analysis – drawing conclusions from existing data based on different factors. The data set we chose described different types of chocolate bars and their ratings. The ratings contained in the data set were based on things like the flavor, texture, and aftermelt of the chocolate. However, the purpose of the dataset was to determine what factors affected the ratings – the location of the company, the type of bean used, the cocoa percent in the chocolate, and so on – and what combination of these factors resulted in the best chocolate.

**Data Set**

The data set, “Chocolate Bar Ratings”, was selected from the website Kaggle.com. It consists of 9 columns and 1795 rows. The columns specify the company making the chocolate, the REF number, the date of review of the chocolate, the location of the company, the origin of the cocoa bean used (specific and broad), the cocoa percent in the chocolate, the bean type, and the rating. It is a large dataset, with specific properties in each row.

**Introduction**

The main purpose of performing analysis on this data set was to determine which types of chocolate were the best. Questions that we asked were as follows:

*Where are the best cocoa beans grown?*

*Which is the best bean type for quality chocolate?*

*Where are the companies with the best ratings located? What are the reasons for this?*

*What is the relationship between the cocoa percent and the quality of the chocolate?*

It is also possible to draw conclusions about how the rating was affected by the date of review, but since that is not directly relevant to the conclusion on the quality of the chocolate, we have not considered it. Further, we could draw conclusions related to the company and the ratings – but the reasons for one company having better ratings than another is related to the factors mentioned above, which we will already have dealt with.

**Data Processing**

The data set had, for the most part, two columns which required cleaning – the bean type, and the broad bean origin. Both these columns had missing values. To deal with broad bean origin column, we looked at the company that made the bars in these rows, and found the most frequently occurring broad bean origin for that company, and entered them into the rows. In case that company had only one entry, we entered the most frequently occurring value from the entire set. We did the same with the bean type, but instead of the company, we compared it with the broad bean origin, as it is more likely that the bean type will be influenced by its place of origin than the company that uses it. We discovered through this exercise that most beans used originate from Venezuela.

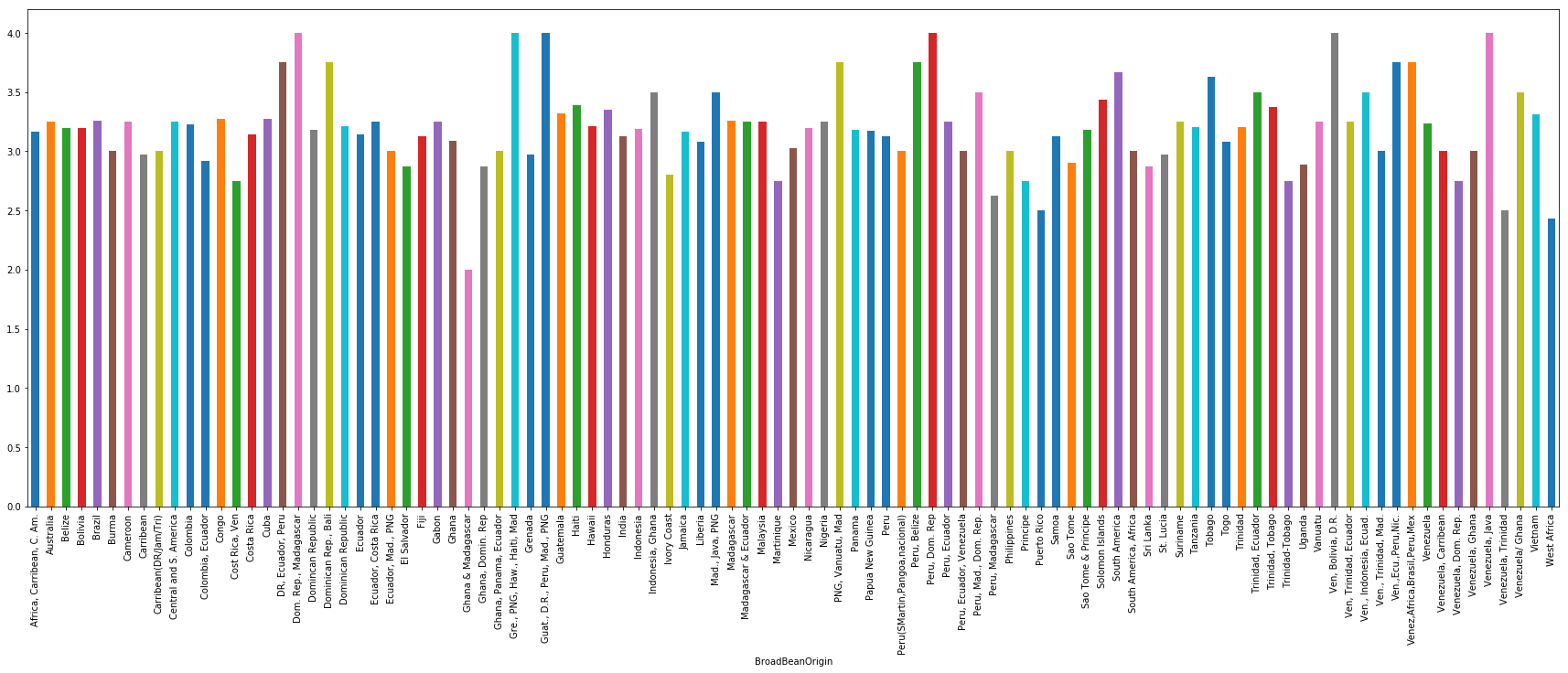
There was another column that required cleaning- the cocoa percent column. It consisted of outliers. So we found the values lying outside the interval for two standard deviations, and replaced those with the median of the data.

**ANALYSIS OF DATA**

We started off our analysis by comparing the origin of the bean to the rating. To do this, we plotted a graph between the broad bean origin to the rating (Graph on next page).

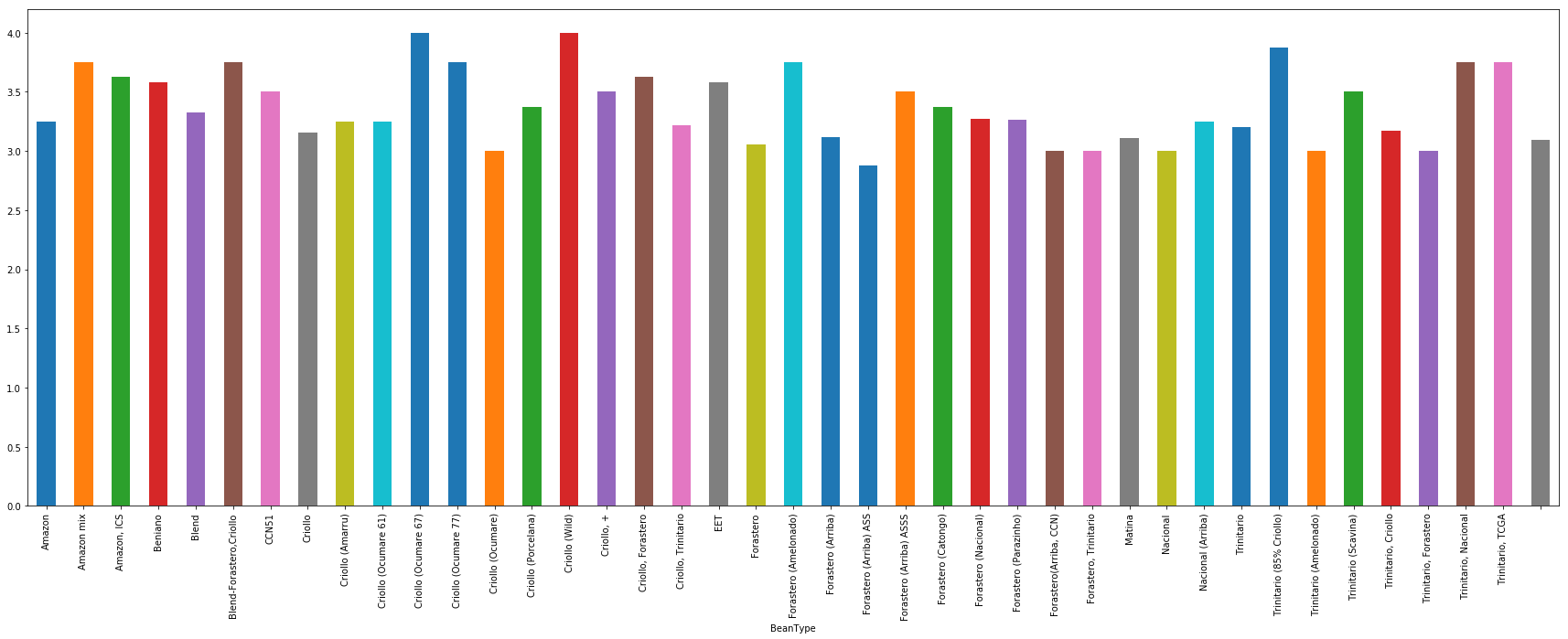
There were six types of origins that gave better ratings than the rest. These beans were from Venezuela and Java, Venezuela, Bolivia, and the Dominican Republic, Peru and the Dominican Republic, Madagascar and the Dominican Republic, and blends from regions that included Madagascar and the Dominican Republic. Upon analyzing further, we noted that when the origin included Venezuela, the ratings were relatively high. This led us to conclude that when the chocolate bars were made from beans that came from Venezuela, the Dominican Republican or Madagascar, or blends of these, they were superior.

We decided to see if there was any relationship between the specific region the bean originated from and its rating. However, there was no discernible pattern, as the data was too spread out over all the categories.

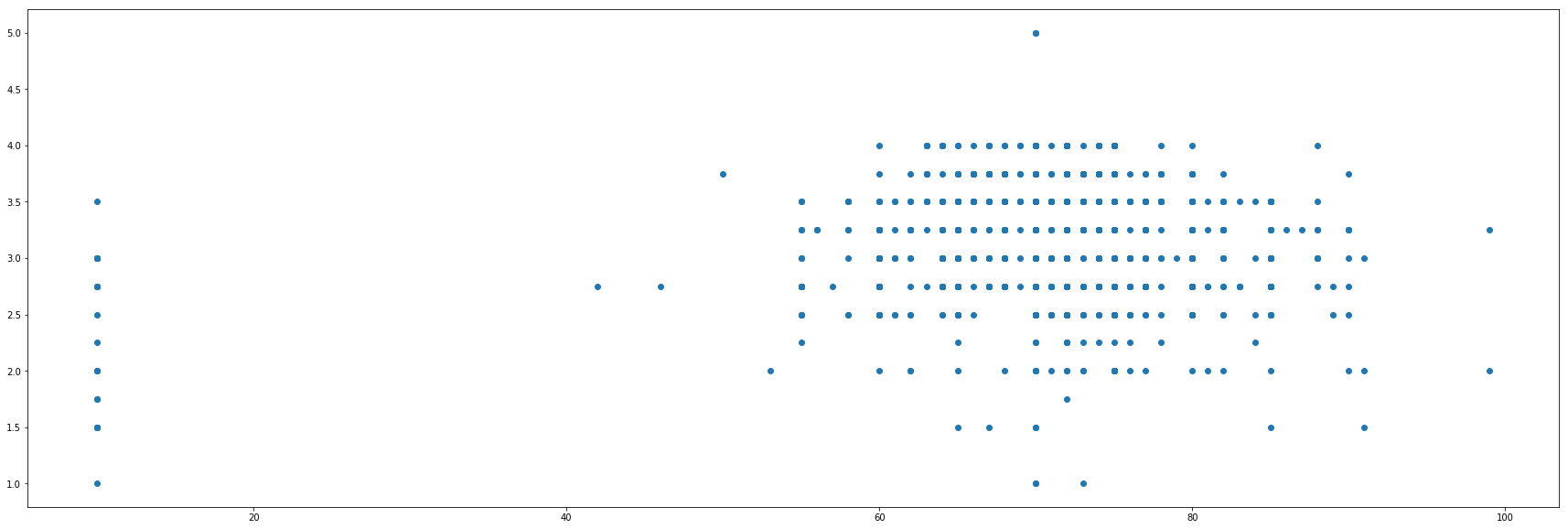
**

The origin of a bean is not the only way of differentiating it from all the others. The bean type is also important. In general, cocoa beans are of three types – criollo, forastero, and trinitario. There are further subgroups as well, which were detailed in the data obtained. To find out the the best type of bean, we plotted a graph between the bean type and the ratings.

We found that that there were mainly two bean types that produced good quality chocolate. These were both variants of the criollo bean – the wild variety, and ocumare 67.



Next, we analysed the relationship between the cocoa solids percentage in a chcolate bar, and its rating. The graph is as shown.



It is clear that most makers of chocolate prefer to use 50-90% cocoa in their bars. However, the highest rated bars have an average of 70% cocoa. In general, bars with 60% to 70% cocoa have high ratings.

Does the cocoa percent used have any relationship with a company’s ratings? We found that the company with the highest ratings average(3.84) was Amedei, and decided to take a look at their way of making chocolate in general (we took into account the average of numerical values, and the mode of categorical).

Bean Origin - Venezuela

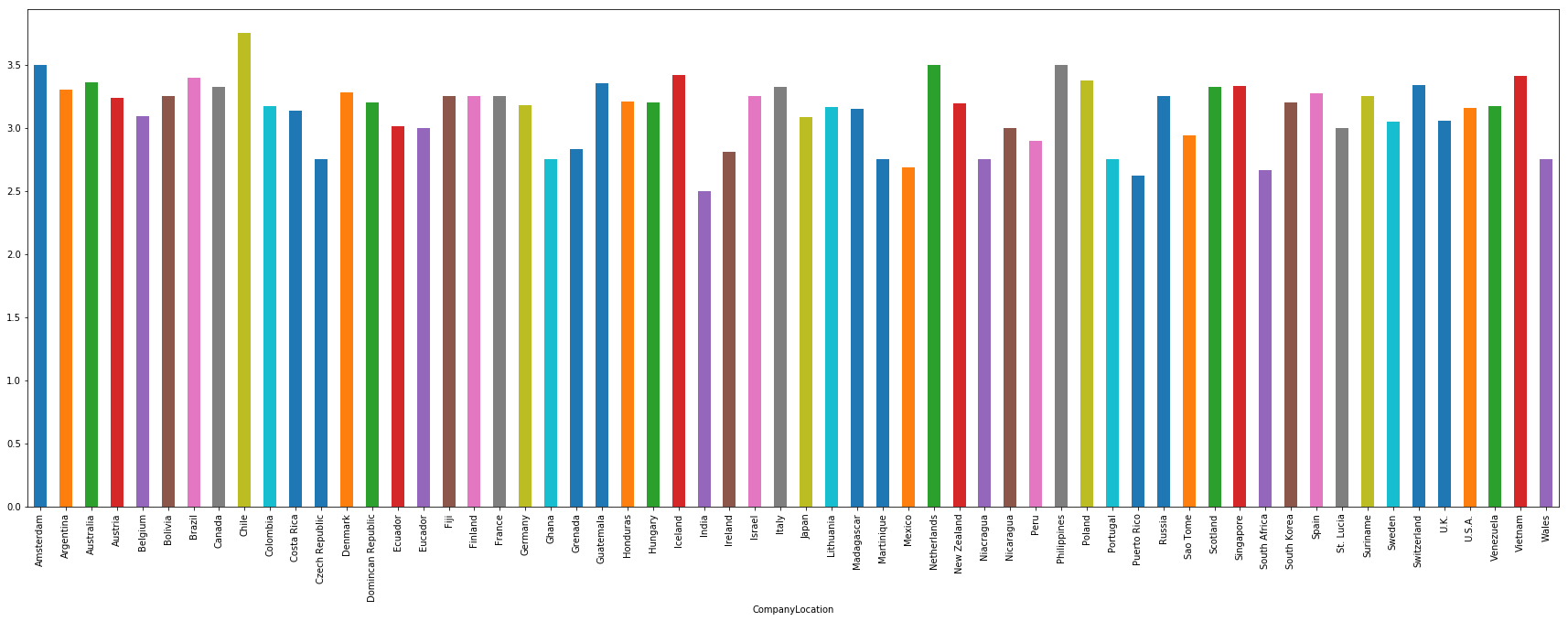
Cocoa Percent – 69.53

Bean Type – Trinitario

Location of company – Italy

We see that the average cocoa percent in their chocolate bars is very close to 70%, which seems to be the ideal quantity for good chocolate. Their beans, although not criollo, are from Venezuela, which again is a prime location for the growing of good quality cocoa beans.

The other criterion on that list the location of their company -Italy. Is Italy a good place for chocolate manufacturers? The following graph shows the relationship between the company location and the rating.



We can see that companies located in Chile seem to have the highest ratings, followed by those in Amsterdam, the Phillipines, and the Netherlands, in no particular order. Chile has an average rating of 3.75, and most companies there use criollo beans from Peru. Italy, however, does have high ratings, even if it isnt’t the highest.

What could be the reason for the discrepancy? When we searched for the location that was most used by chocolate manufacturers, we found that it was the U.S.A. However, the companies located in the U.S don’t have a high ratings average (only 3.15). This led us to conclude that the location does not play that important of a role in the quality of chocolate produced. Companies choose their locations based on business prospects, and how easy it is to import or export from there. Therefore, while the location does play a small part (it could influence from where cocoa beans are imported, for instance), it does not contribute directly to the quality of the chocolate.

**CONCLUSION**

After analysing the influence of various factors on the ratings of chocolate bars, we concluded that there is a pattern, and a relationship, between the origin of the bean and the ratings, the type of bean and the ratings, and the cocoa percent in the chocolate bars. Beans of the criollo type and originating in the Dominican Republic, Venezuela or Madagascar contribute to making superior chocolate. Further, having the chocolate be made of 60-70 percent of cocoa improves its quality. The location of the company plays a minor role in the quality of the chocolate produced, and in general, companies located in Chile are able to produce better chocolate bars. The highest rated company, i.e, the company which makes the best chocolate, according to this data set, is Amedei, located in Italy.

This assignment was useful in understanding the procedures involved in coming to conclusion about a data set. It is necessary to understand the data, what it represents, and the ways in which it could be interpreted. We must manipulate the data precisely, so that we obtain the results we require. Most data is not clean- we must keep in mind all the anomalies, the discrepancies, and the outliers while forming conclusion. Informed, cautious and adequate analysis of data is essential to learning about trends, and thereby improving the quality of business and life in a fruitful manner.