Key Drivers for Quality Coffee

Bosniack, Lesley, Bowen, Janet, Higgins, Thomas, & Palacios, Gerardo

[1] DSC 424 - Advanced Data Analysis

[2] DePaul University

Key Drivers for Quality Coffee

Domain Knowledge and Data Gathering

According to the Coffee Quality Institute (CQI), coffee quality is one of the most important variables that influence a coffee's value. One way that coffee quality can be measured is through a blind tasting, also known as cupping, by certified coffee analysts using the SCAA Cupping Protocol. This protocol gives guidelines for evaluation ranging from necessary equipment to preparation of the coffee. Ratings are given on a scale of 6-point scale in various categories such as aroma, flavor, aftertaste, acidity, body, balance, sweetness, clarity, consistency, and overall impression. The final grade sums the ratings against a total of 100 points, similar to the scale that exists for wine. Anything rated over 80 points is considered a premium coffee. This method of evaluation provides a consistent and objective methodology for capturing some of the beans' sensory aspects and for evaluating quality.

The CQI's goal is to improve the quality of coffee and the lives of coffee producers. As a result, the CQI has provided opportunities to individuals to become certified as coffee graders, analysts among other experts to support and educate current and future generations of coffee growers. The data-set selected and described below was compiled by the CQI from samples submitted for evaluation. It provides a profile of coffee growers and coffee beans and the quality of the coffee grown, as measured according to the SCAA categories. Analysis of this data-set could identify valuable findings that would strengthen growing practices and knowledge which may favorably affect coffee producers . The final data-set was gathered from multiple sources and combined into a single data-set. The parent data was gathered from *https://www.kaggle.com/volpatto/coffee-quality-database-from-cqi* representing 1339 observations and 43 variables from the CQI database of coffee ratings from 2010 through 2018. Each observation represents a coffee grower and 43 variables. The relevant weather and geographical data associated with the growing regions of the parent data-set was gathered using two APIs to capture their potential impact on coffee quality. Isolating this data for the coffee growing season of each region is necessary for the weather data to be useful. Because growing season was not easily identified, data on harvest season was collected by country and if available, by region from a number of sources. Harvest time frame will be utilized to back into growing season.

API Methodologies

Google Place API. The Google Places API is a service used to gather data from any given location. The available request, "Place Search" , was used to determine the Latitude and Longitude for each of the given countries regions and countries. A custom user function was created using Excel VBA to build an http request based on the name of the given region in the Kaggle data-set. The GET request returns a JSON file which was then parsed to focus on the central coordinates of any given region in the data-set.

DarkSkies API. The Dark sky API is a comprehensive weather API that allows users to look up weather data from anywhere in the world. The API is very simple to use and only requires 2 or 3 parameters (Lat|Long and/or Time). The weather data was gathered based on the estimated harvest season months for each of the regions. The return is beginning of month weather measurements for that particular region and time. The readings associated then is calculated to determine an average measure for the harvest season.

Data Pre-processing

The data gathered from the API's came in a clean format which did not need further processing. The parent data-set, however, required the following steps in order to be correctly processed by R.

- The numerical variable "Weight", came in a "String" format containing mixed unit and measurements for coffee (kg or lbs).The variable was then separated so that all numerical units were separated and converted under a unified unit (kg).

- The numerical variable "Harvest Year", came in a "String" format containing mixed start | end harvest years. To standardize the column, only the end Harvest year was taken.

- The numerical variable "Grading Date" and "Expiration" came in a "String" format containing the date that the coffee was graded or expired. In order for R to correctly process that variable it needed to be converted into a time variable. The year, day and month for each row was extracted to build the correct date value in order for R to interpret.

- Data needed to be reduced to represent only complete data.

- Values of each variable were reviewed. Data adjustments were identified for potentially erroneous values.

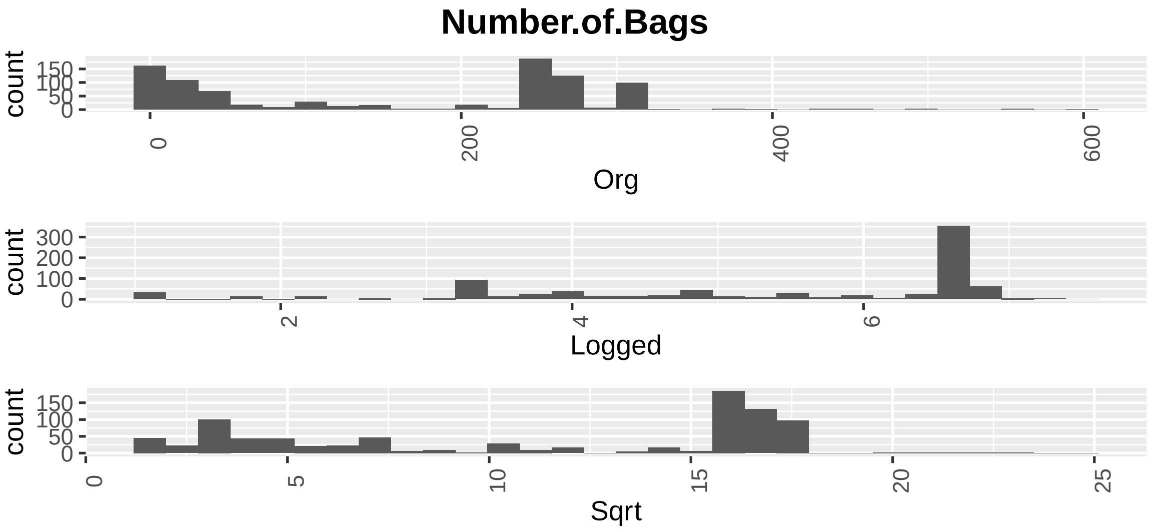
Variable Survey

The main purpose with this data-set is to find the *key drivers* for well rated coffee or quality coffee. The predictors generally fall under 5 categories that portray different characteristics for each of the coffee producers.

Category	Description
Grader scores	This is a set of scores given to a sample by a certified Q grader, each corresponding with different aspects of quality in the coffee. Cupper.Points is a subjective quality evaluation made by the grader, and Total.Cup.Points is the sum total of all of these. These are all numeric variables, and most of them are normally distributed, with the exception of Uniformity, Clean.Cup, and Sweetness, which are strongly skewed left - most values are at 10 for those.
Geographical data	These are traits of the coffee beans noted during inspection; Color is a categorical variable, the others are numeric (although treating them as categorical might be an option). Category One defects are considered more undesirable than Category Two defects. For color, most of the data falls into two categories; for the others, the data is skewed right, with most values at zero (in other words, no defects of that sort were observed).
Production characteristics	Data associated with where the graded coffee sample was produced; used to get weather information, or possibly as a variable of its own, depending on analysis methodology. Categorical variables; some countries/regions are represented much more than others.
Weather data	Variables that are in some way associated with the production process of the coffee. Mix of categorical and numerical variables.
Other quality indicators	Variables describing the weather conditions in the region the coffee was produced. Mostly numeric; the categorical variables might not be usable. In general, the idea here is that growing conditions are known to affect the quality of farm products

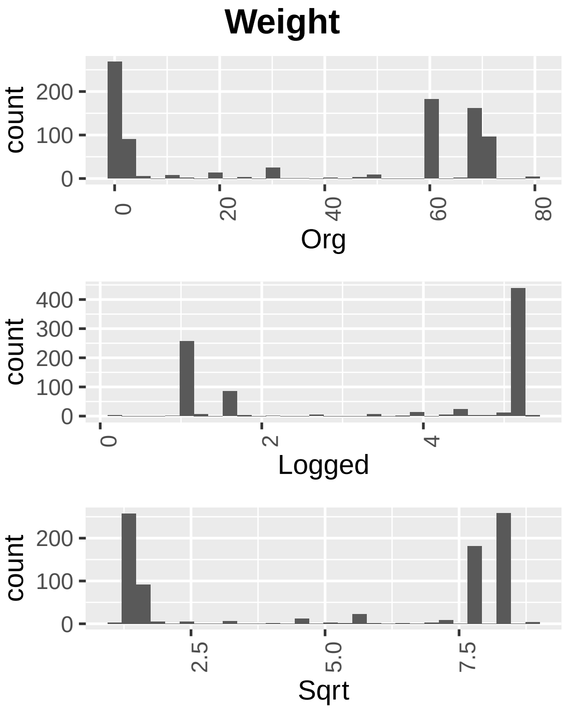
Notable Distributions and Transformations. The majority of the variables were normally distributed with the exception of a few that may need special consideration when running regressions or conducting PCA analysis. 6 variables were shown to have non-normal distributions.

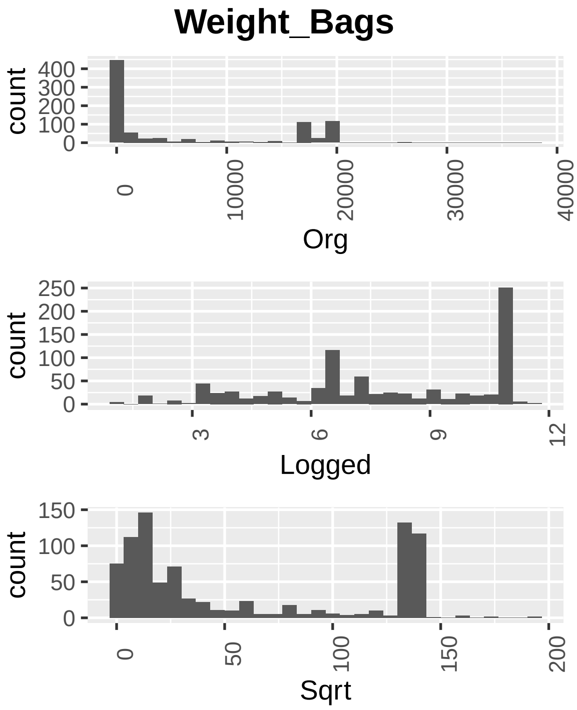
1. Number of Bags



The number of bags has a highly skewed-right bi-modal distribution at zero and 250 bags. A square root transformation is shown to improve the normality of the distribution.

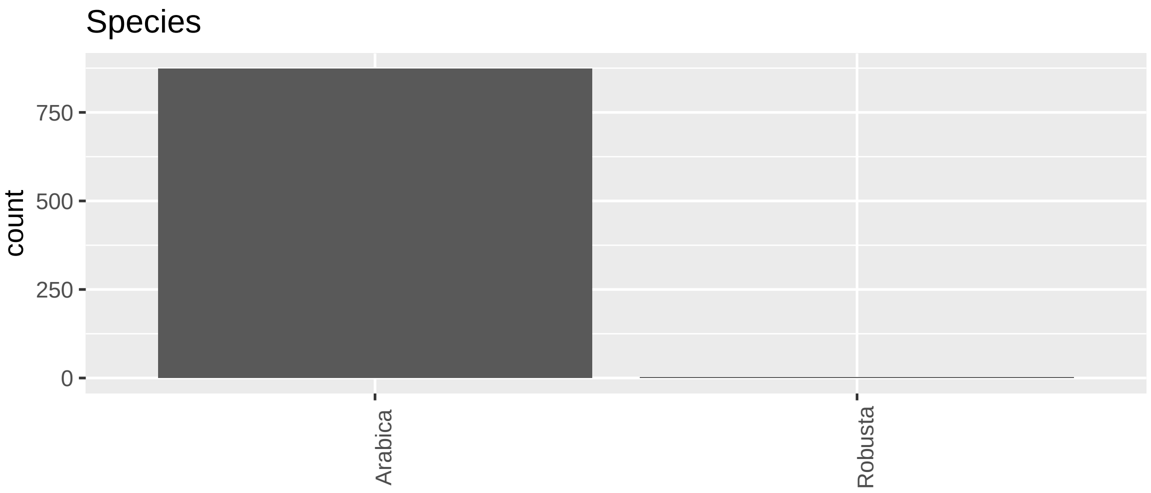
2. Weight and Num of Bags x Weight





Both variables are highly skewed-right with most of the observations generally are generally light. A logarithmic transformation is shown to have the most improvement.

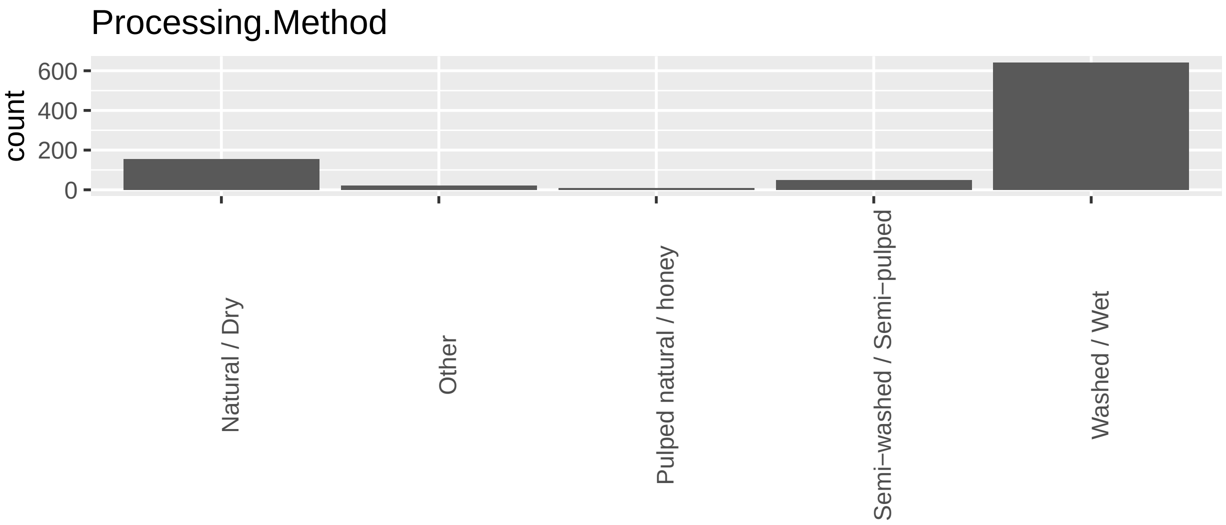
3. Species



Species is highly disproportionate with most of the observations falling under *Arabica*. Only 33 observations from the final data-set are categorized as *Robusta*.

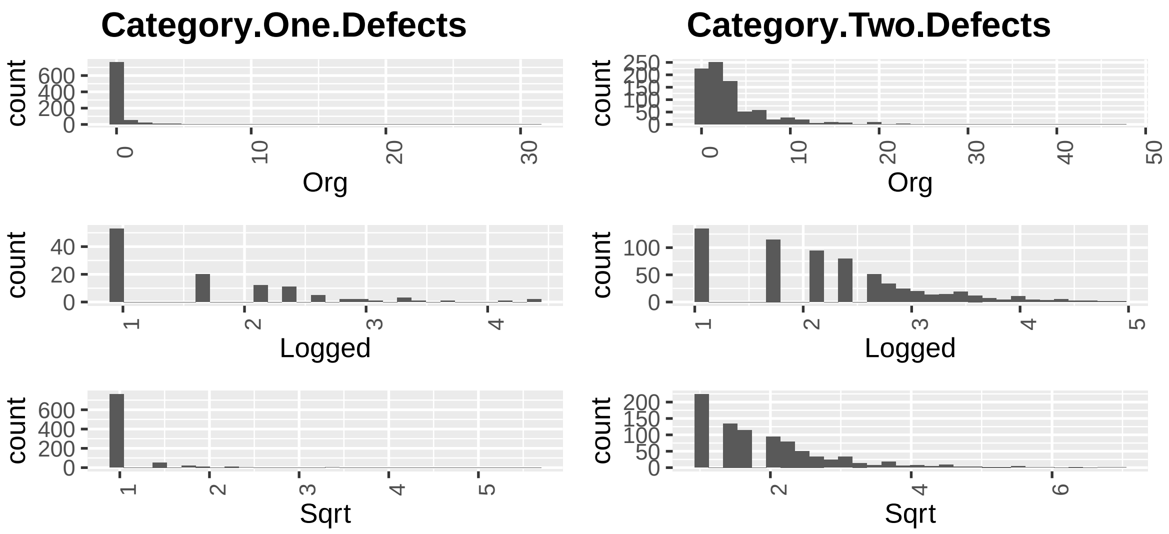
4. Color

```
## [[1]]
```



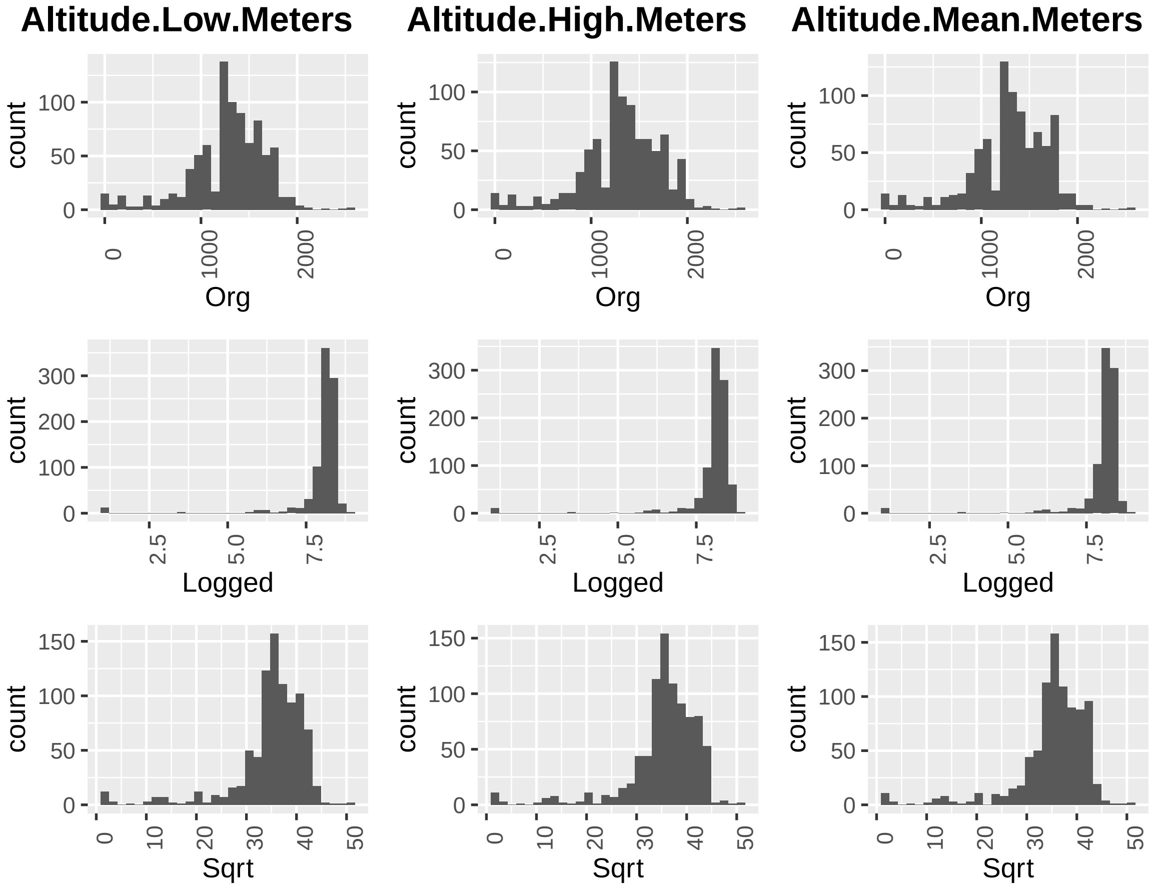
The color of the coffee largely falls into a single category making it a highly uneven distribution between the rest of the categories. 638 observations fell under “Green” while the remaining 167 observations were spread evenly between 3 categories.

5. Category One and Two Defects



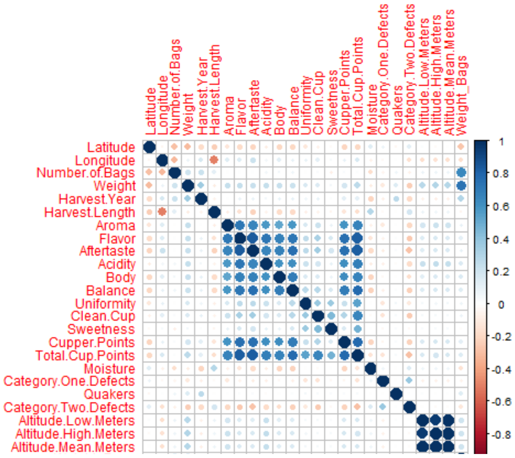
Category one and two defects are strongly skewed to the right with the majority of the values at or near 0. This distribution is anticipated as we would expect most coffees would be devoid of any defects. Both the logarithmic and square root transformations improve the distributions by reducing skew but the improvement is only nominal.

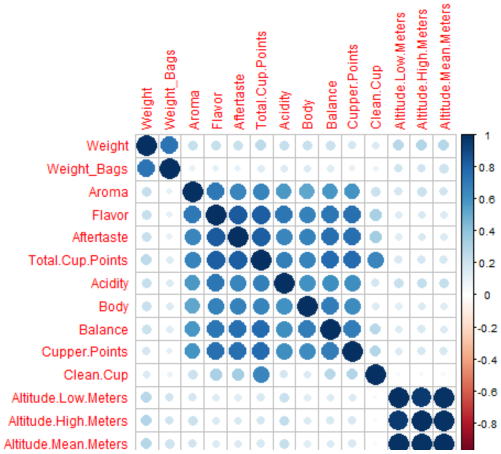
6. Altitudes (Low, High and Mean)

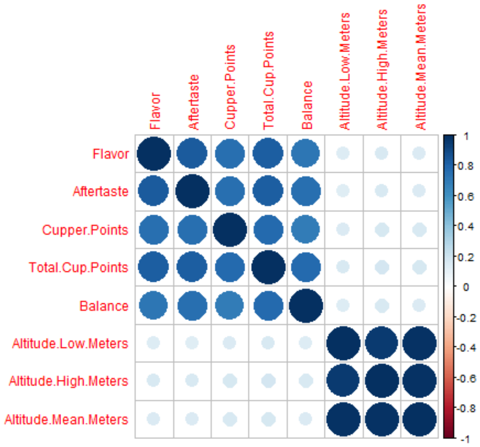


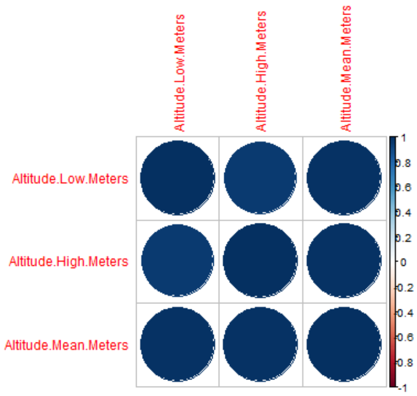
All three altitude variables were strongly skewed right. A logarithmic transformation was shown to have the biggest improvement on the distribution. It is worth noting that the most of the "High" and "Low" values are identical. With the "Mean" variable depicting the most variance between the three. Going forward the mean will be the sole variable to use relating to the altitude.

Notable Correlations. Correlations were reviewed across all variables. Below are initial plots against all variables as well as whose correlations of above 0.65 or below -0.65. All of these variables exhibit a positive correlations. It can be noted that altitude variables all show perfect multi-collinearity suggesting repetitive data. This supports our earlier suggestion on the altitude variables. The Specialty Coffee Association of America ("SCAA") quality scores are all positively correlated with the lowest correlation of 0.51 between aroma and body. This suggests that if quality is high on some criteria, it is likely that quality will be high for other criteria and for the overall quality score, Total.Cup.Points. Similarly, where low quality scores exist on some criteria, one would expect low quality scores on other criteria and on the combined score. Defects 1 and 2 and moisture levels exhibit mild negative correlations to the SCAA scores. Altitude does not appear to be correlated with SCAA scores.

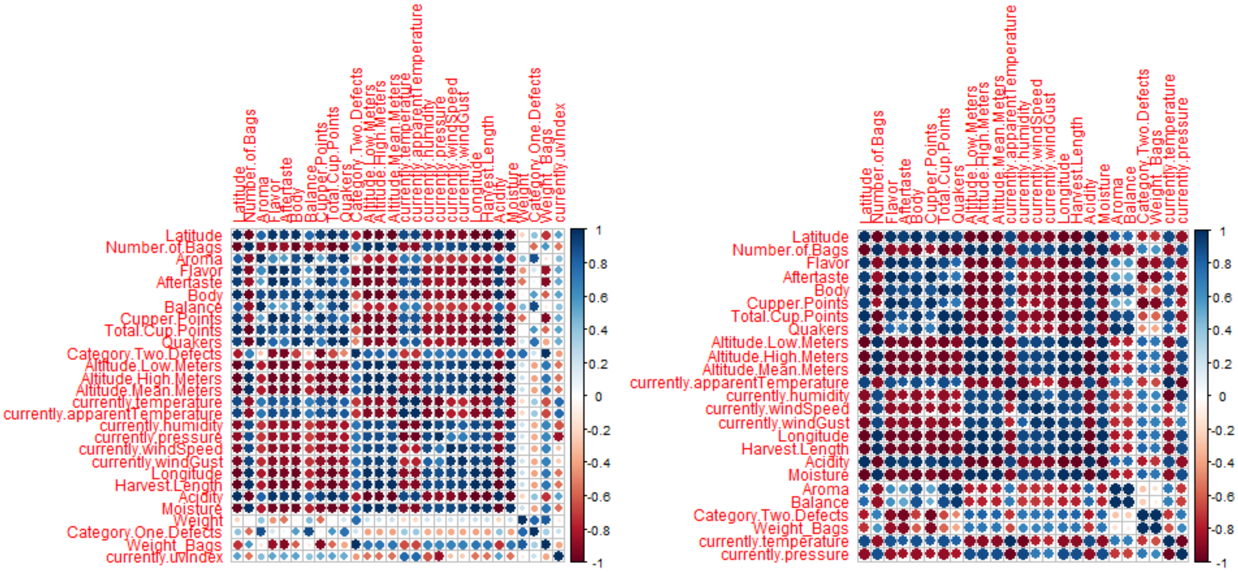








The weather data-set consist of three categories of measurements. Hourly, daily and currently. The most complete section of the data consisted of the "daily" measurements This consisted of 39 additional variables related to the geographical location of any given coffee producer. Nearly all of the variables are correlated with each other while a few are negatively correlated. The variables humidity, pressure and visibility all tend to decrease in value while all other variables increase. The correlation groupings of at least 0.75 display a similar pattern. The strongest group of correlated variables are all positive with only one,(pressure) that is negative. From the graph, you can see two groupings of strongly correlated variables, the moon phases and the daily temperatures, and the other daily apparent temperatures and daily pressure.



Tables

Variable	Type	Description	Relevance
Species	Categorical		Different species of coffee cherry may have different qualities that impact the score
Number.of.Bags	Numeric	Number of Bags of Coffee in a Growing Season	Small batches are often more premium
Wt	Numeric	Average weight of bags in kilograms	Small batches are often more premium
Harvest.Year.Fixed	Date	Year the coffee was harvested	For similar products (like wine), some years are known to be good/bad for quality historically
Grading.Date.Fixed	Date	Month, day and year coffee was graded	Probably not relevant
Variety	Categorical	Variety of coffee plant	Different varieties sometimes have unique characteristics
Processing.Method	Categorical	Method of processing (wet/washed, dry, pulped and semi-washed)	Processing method can alter taste
Aroma	Numeric	SCAA aroma (dry fragrance) quality score ranging from 6-10	Higher values have a better aroma
Flavor	Numeric	SCAA flavor quality score ranging from 6-10	Higher values have a better flavor
Aftertaste	Numeric	SCAA aftertaste quality score ranging from 6-10	Higher values have a better aftertaste
Acidity	Numeric	SCAA acidity quality score ranging from 6-10	Higher values have a better flavor profile
Body	Numeric	SCAA body quality score ranging from 6-10	Higher values have better body
Balance	Numeric	SCAA balance quality score ranging from 6-10	Higher values have better balance
Uniformity	Numeric	SCAA uniformity quality score ranging from 6-10	Higher values have better unifomity
Clean.Cup	Numeric	SCAA clean cup quality score ranging from 6-10	Higher values have better clen up

Variable	Type	Description	Relevance
Sweetness	Numeric	SCAA sweetness quality score ranging from 6-10	Higher values have better sweetness
Cupper.Points	Numeric	Correction points added by grader for more appealing coffees not reflected in other scores	Grader's personal evaluation
Total.Cup.Points	Numeric	Sum of Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup and Sweetness scores	Total score
Moisture	Numeric	Moisture percentage	Important to roasters for quality control. Contributes to acidity and aroma.
Category.One.Defects	Numeric	Number of primary defects	Catch-all for detrimental bean traits noticed by grader
Quakers	Numeric	Quantity of unripe or poorly roasted beans	Unripe beans, sign of bad nutrition/picking practices/etc
Color	Categorical	Coffee color ranges from green to blue	Factors during cultivation, picking, drying, and milling can impact bean color and give clues about the characteristics a particular bean will take on during roasting and brewing
Category.Two.Defects	Numeric	Number of secondary defect	Less severe detrimental bean traits
Expiration.Fixed	Date	One year post grade date	Probably not relevant
altitude_low_meters	Numeric	Distance from Sea Level - minimum for grower	
altitude_high_meters	Numeric	Distance from Sea Level - maximum for grower	Higher altitudes are often associated with better quality coffee
altitude_mean_meters	Numeric	Distance from Sea Level - average for grower	Mean altitude range of coffee farmers
Regional Longitude	N/A	Longitude coordinate of coffee producer	Provides primary key to merge weather data
Regional Latitude	N/A	Latitude coordinate of coffee producer	Provides primary key to merge weather data
Apparent (feels-like) temperature	Numeric	How weather conditions feel to bare skin	Weather patterns can affect crop production and quality

Variable	Type	Description	Relevance
Atmospheric pressure	Numeric	Pressure due to weight of atmosphere	Weather patterns can affect crop production and quality
Cloud cover			Weather patterns can affect crop production and quality
Dew point	Numeric	Atmospheric temperature at which dew will form	Weather patterns can affect crop production and quality
Humidity	Numeric	Concentration of water vapor in air	Weather patterns can affect crop production and quality
Liquid precipitation rate	Numeric		Weather patterns can affect crop production and quality
Moon phase			Gravitational force of the moon can stimulate plant growth and crop quality
Nearest storm distance	Numeric		Weather patterns can affect crop production and quality
Nearest storm direction	Numeric		Weather patterns can affect crop production and quality
Ozone			Ozone can damage plants and reduce their survival
Precipitation type			Weather patterns can affect crop production and quality
Snowfall			Weather patterns can affect crop production and quality
Sun rise/set			Weather patterns can affect crop production and quality
Temperature			Weather patterns can affect crop production and quality
Text summaries	N/A		Not relevant - text blurb about daily weather
UV index			Weather patterns can affect crop production and quality

Variable	Type	Description	Relevance
Atmospheric pressure	Numeric	Pressure due to weight of atmosphere	Weather patterns can affect crop production and quality
Cloud cover			Weather patterns can affect crop production and quality
Dew point	Numeric	Atmospheric temperature at which dew will form	Weather patterns can affect crop production and quality
Humidity	Numeric	Concentration of water vapor in air	Weather patterns can affect crop production and quality
Liquid precipitation rate	Numeric		Weather patterns can affect crop production and quality
Moon phase			Gravitational force of the moon can stimulate plant growth and crop quality
Nearest storm distance	Numeric		Weather patterns can affect crop production and quality
Nearest storm direction	Numeric		Weather patterns can affect crop production and quality
Ozone			Ozone can damage plants and reduce their survival
Precipitation type			Weather patterns can affect crop production and quality
Snowfall			Weather patterns can affect crop production and quality
Sun rise/set			Weather patterns can affect crop production and quality
Temperature			Weather patterns can affect crop production and quality
Text summaries	N/A		Not relevant - text blurb about daily weather
UV index			Weather patterns can affect crop production and quality

Statistic	Latitude	Longitude	Number.of.Bags	Weight	Harvest.Year	Harvest.Length	Aroma	Flavor	Aftertaste
1st Qu.	2.536	-91.47	20.0	1.0000	2012	92.0	7.420	7.330	7.170
3rd Qu.	17.059	-44.56	275.0	69.0000	2015	152.0	7.750	7.670	7.580
Max.	37.090	126.09	600.0	80.0000	2018	214.0	8.750	8.670	8.500
Mean	9.174	-44.41	159.2	36.1845	2014	133.3	7.561	7.505	7.376
Median	14.518	-84.04	200.0	60.0000	2014	151.0	7.580	7.500	7.420
Min.	-23.563	-118.28	1.0	0.4536	2011	31.0	5.080	6.170	6.170

Statistic	Acidity	Body	Balance	Uniformity	Clean.Cup	Sweetness	Cupper.Points	Total.Cup.Points	Moisture
1st Qu.	7.330	7.330	7.330	10.00	10.000	10.000	7.250	81.17	0.10000
3rd Qu.	7.670	7.670	7.670	10.00	10.000	10.000	7.670	83.50	0.12000
Max.	8.580	8.420	8.580	10.00	10.000	10.000	8.580	89.92	0.17000
Mean	7.517	7.494	7.492	9.87	9.846	9.933	7.461	82.05	0.09807
Median	7.500	7.500	7.500	10.00	10.000	10.000	7.500	82.42	0.11000
Min.	5.250	6.330	6.080	6.00	0.000	1.330	5.170	59.83	0.00000

Statistic	Category.One.Defects	Quakers	Category.Two.Defects	Altitude.Mean.Meters	Weight_Bags
1st Qu.	0.0000	0.0000	0.000	1100	180
3rd Qu.	0.0000	0.0000	5.000	1550	17250
Max.	31.0000	11.0000	47.000	2560	37950
Mean	0.4263	0.1463	3.822	1293	6558
Median	0.0000	0.0000	2.000	1311	640
Min.	0.0000	0.0000	0.000	1	1