Matthew Roumelis
matthewroumelis@gmail.com
210-421-2531

Predicting Coffee Grade Final Report

I. Context and Problem Statement

Coffee has become such an integral part of contemporary culture that it's hard to imagine a life (or a US workforce) without it. According to the National Coffee Association, coffee related economic activity accounted for 1.6% of the total U.S. GDP in 2015. In the same year, consumers in America spent a whopping $74 billion on coffee. Needless to say, Americans love coffee and are willing to pay for good quality. The good news is that distinguishing 'good quality' coffee is relatively straightforward because of a very standardized way to grade it: The Specialty Coffee Association (SCA) Coffee Beans Classification method. This method judges the relationship between the amount of defective coffee beans and the overall cup quality as evaluated and scored by a panel of judges. The coffee is then assigned a grade based on the results:

- GRADE 1: Specialty Grade Coffee Beans
- GRADE 2: Premium Grade
- GRADE 3: Exchange Grade Coffee Beans
- GRADE 4: Standard Grade Coffee Beans
- GRADE 5: Off Grade Coffee Beans

Many coffee shops pride themselves on their coffee quality and are willing to share information about their coffee such as the region the coffee is from, the altitude at which it was grown, how it was pre processed, etc. but rarely include the official grade of the coffee. Whether this is because the coffee hasn't been officially graded by the SCA or it's simply not advertised for some other reason, businesses and consumers could benefit by knowing the quality of coffee they are buying or selling. **A model that can accurately predict coffee grade based on the known factors of when, where, and how the coffee was grown and processed would help businesses select better quality coffee producers and market/sell their products appropriately.**

II. Data Acquisition & Wrangling

The Data

The data used for this project came from the Coffee Quality Institute website. This website contains a database with recent coffee that has been officially graded using the SCA Coffee Beans Classification method. This data includes relevant information such as altitude and region the coffee was grown, the cupping scores and number of defects used to assign the coffee a grade, as well as other useful information. The data extracted was limited to just the arabica coffee (excluding robusta) because of the small amount of robusta data. Though the data did not include the coffee grade which is the feature we will be predicting, this could be calculated using the official grading standards given the number of quakers and defects provided. Because the data scraped was limited to coffee dating back to 2019, the dataset was small, constituting only **150 coffee samples**. This size of the data set had to be taken into consideration while performing analysis, modeling, and in preparing the final results.

Data Wrangling

This section describes the various data cleaning and data wrangling methods used on the data. It is separated into numerical and categorical features.

Numerical Features

- Altitude - This column contained a messy assortment of values. Most of the altitude values were given as ranges. The average between the high and low values of the range were calculated and a new column created to house these means. There were also some outliers that needed to be addressed such as an altitude of 15,002,100 meters which I concluded was supposed to be the range 1500-2100 meters.
- Moisture - This column was given as a percentage with the percent sign included in the data. The percent sign was removed.
- Category 1 and Category 2 Defects - These columns were inputted as strings and were converted to integers. This would allow us to create our Coffee Grade column later.

Categorical Features

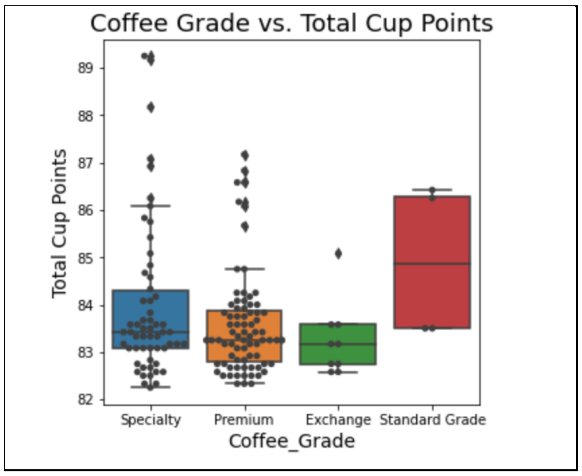
- Harvest Year - The data with “2020/2021” as their harvest year were changed to 2021.
- Color - similar color categories that were only different because of verbage or hyphens were combined ie. Yellow Green and yellow-green were combined to be Yellow-Green.

Lastly, a new column was made which assigned each coffee a SCA Coffee Bean Classification Grade based on their number of quakers and defects. This will be the target variable for our modeling. From our data of 150 coffees, there were 58 Specialty (38.7%), 79 Premium (52.7%), 9 Exchange (6.0%), and 4 Standard Grade (2.7%). The **highly unbalanced** nature of this data set is something important to consider when modeling and analyzing the results.

III. Exploratory Data Analysis

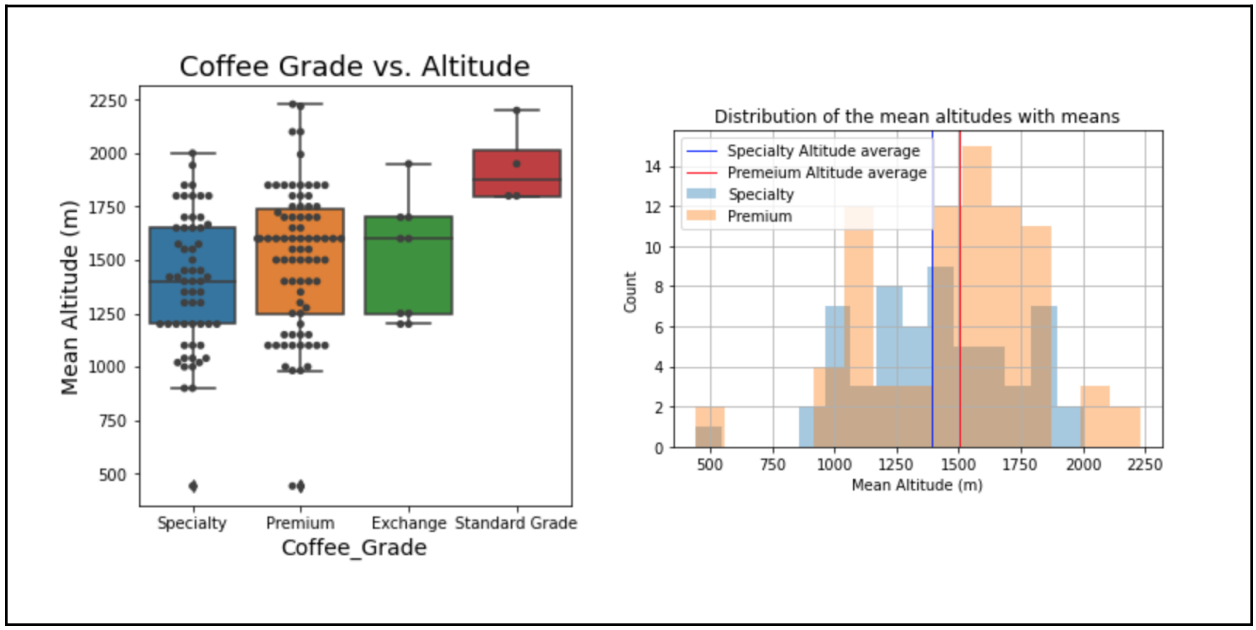
In this section, the various insights produced through descriptive statistics and data visualization is presented. The main goal was to see how our features affected or related to our target variable, Coffee Grade.

Coffee Grade and Total Cup Points

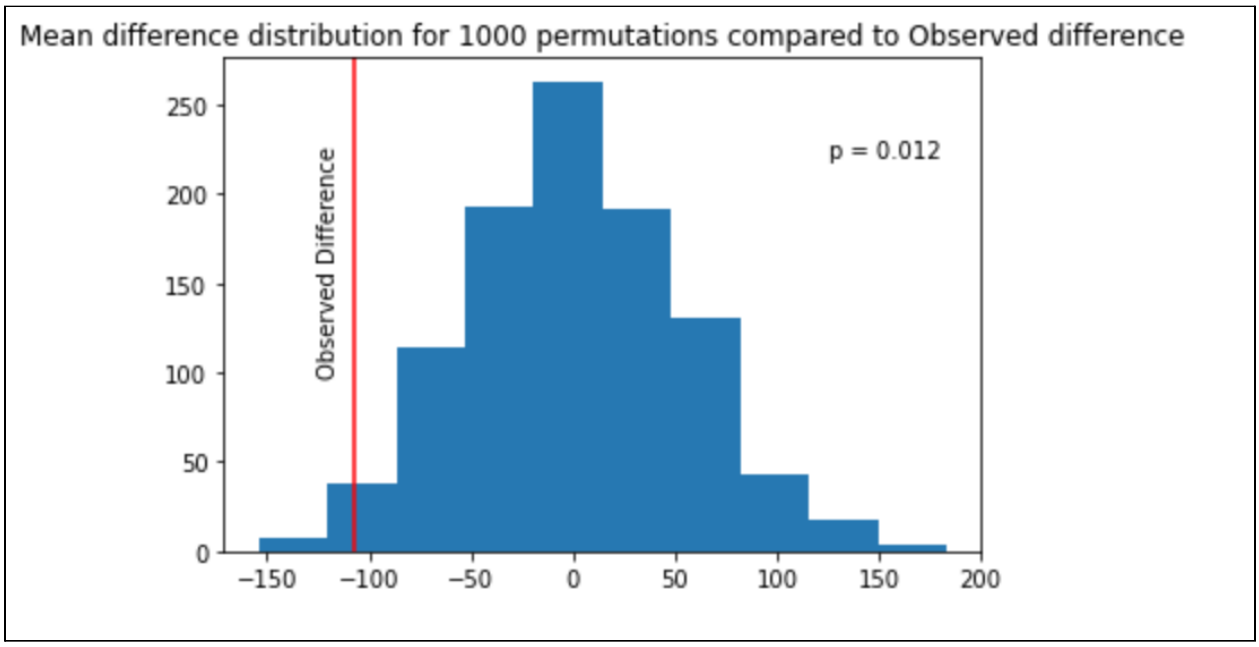


Total Cup Points is the cumulative measure given to the coffee based on the graders' scores while performing the official coffee tasting for grading. As seen by the box plot and swarm plot, Coffee Grade vs. Total Cup Points, our coffee does seem to follow a trend: the higher the Total Cup Points, the better the coffee grade (except for the Standard Grade which is the lowest grade of our data). The Standard Grade results can probably be attributed to random error given that we only have 4 Standard Grade coffees in our entire dataset. Though this was an interesting analysis, the goal is to predict coffee grade *before* it's officially graded and so the Total Cup Points will be taken out of the list of features with which to run our future modes.

Coffee Grade and Altitude

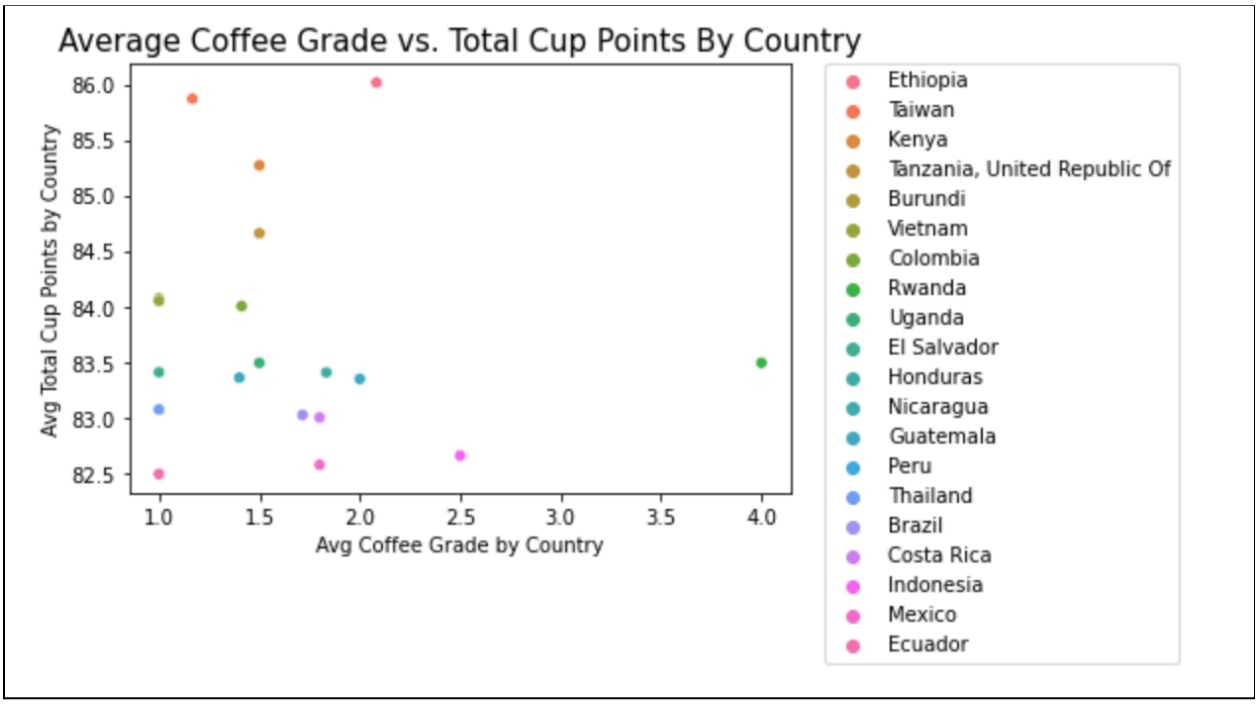


The null hypothesis was that altitude had no effect on coffee Grade. As you can see by the box plot, the initial trend suggested that the **lower altitude coffees in this data set had better Coffee Grades** in general. This observation was confirmed by conducting a permutation test (p = 0.012) to see if there was a statistically significant difference in the altitudes of the Specialty coffees and Premium coffees. (We limited our permutation test to these two grades because of the availability of data). Thus we could reasonably reject the null hypothesis.

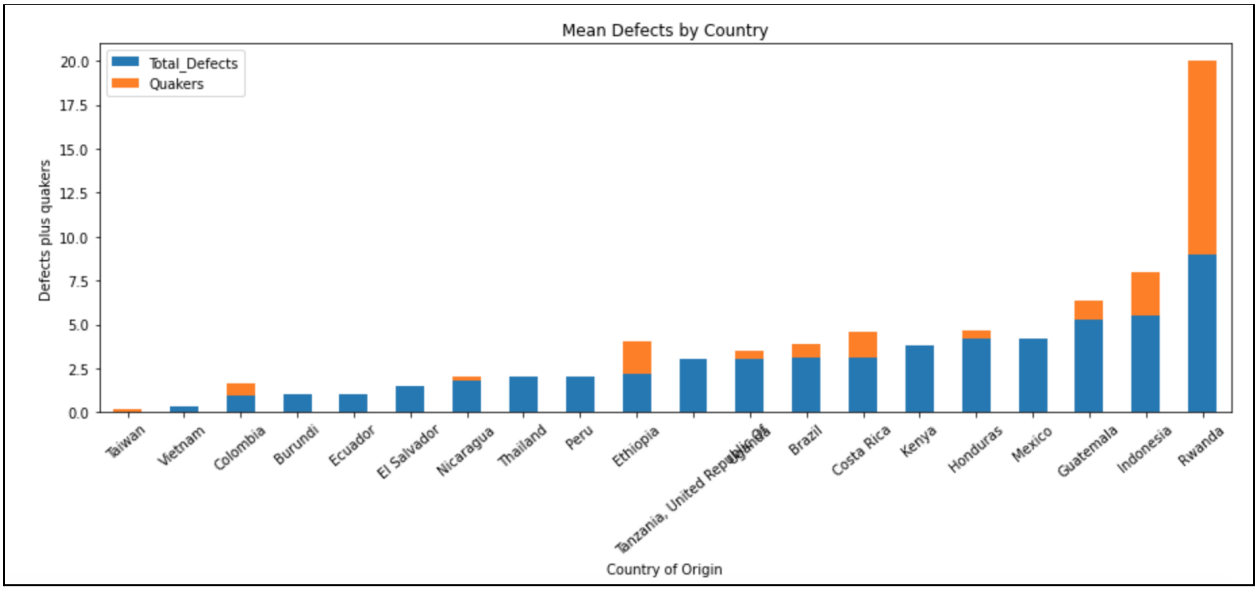


Coffee Grade and Country

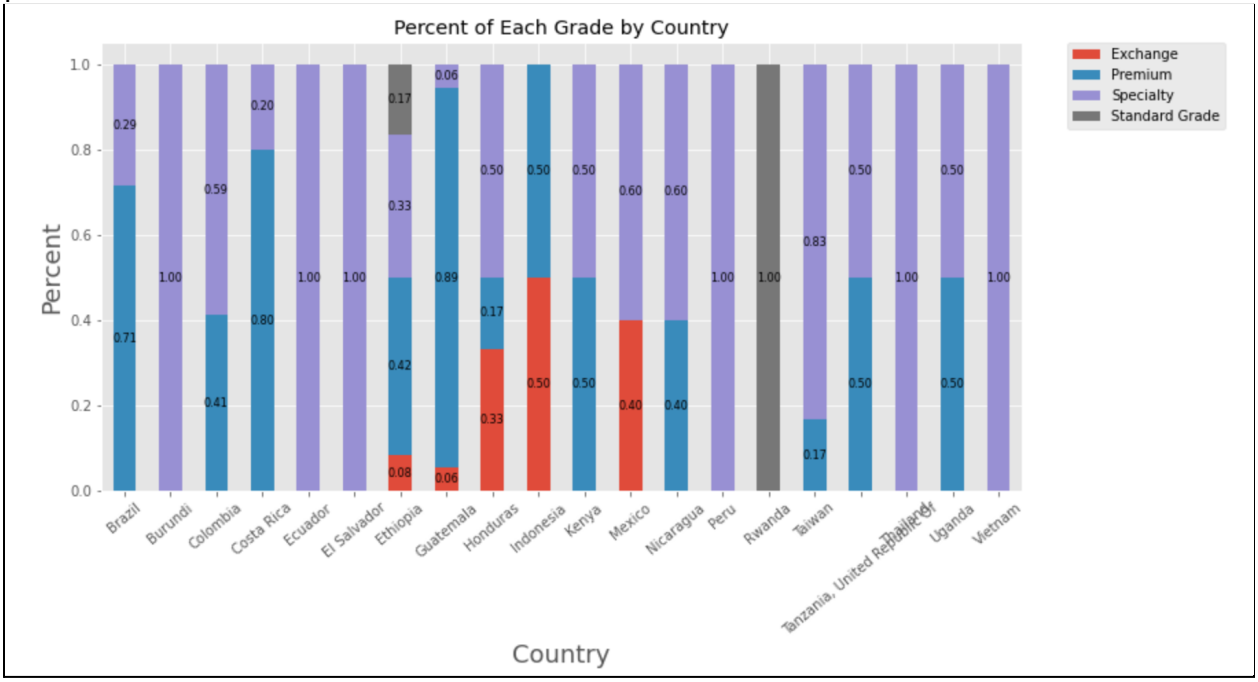
To compare the coffee grade and the country where the coffee was grown, a couple visualizations were created to get a better feel for our data and the general trends. First, each coffee grade was converted to a numerical value. To keep it consistent and interpretable, Specialty coffee (grade 1) was assigned the number 1, Premium coffee (grade 2) was assigned the number 2, Exchange (grade 3) was assigned the number 3, and Specialty Grade (grade 4) was assigned the number 4.



If we imagine the scatter plot above divided into four quadrants, our best coffee would come from the countries in the upper left quadrant (high Total Cup Points and better Coffee Grade). For this data that would include Taiwan, Ethiopia, Kenya, and Tanzania.

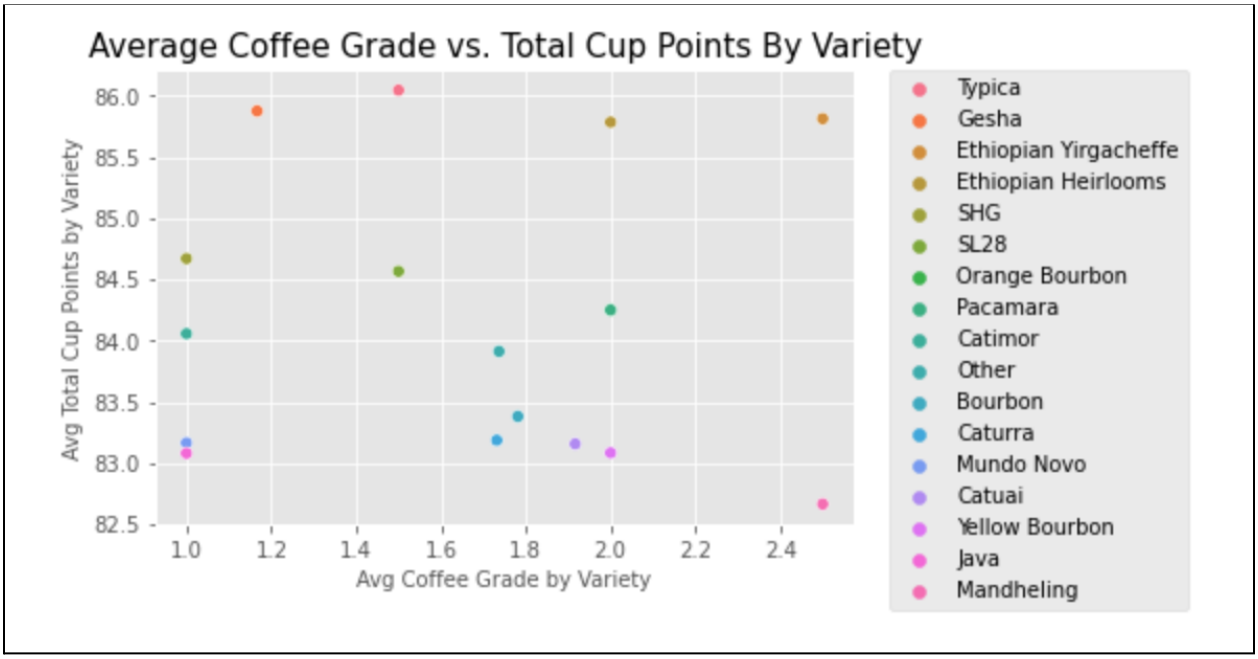


Here we can see that Taiwan, Vietnam, and Colombia have the coffees with the lowest number of defects on average

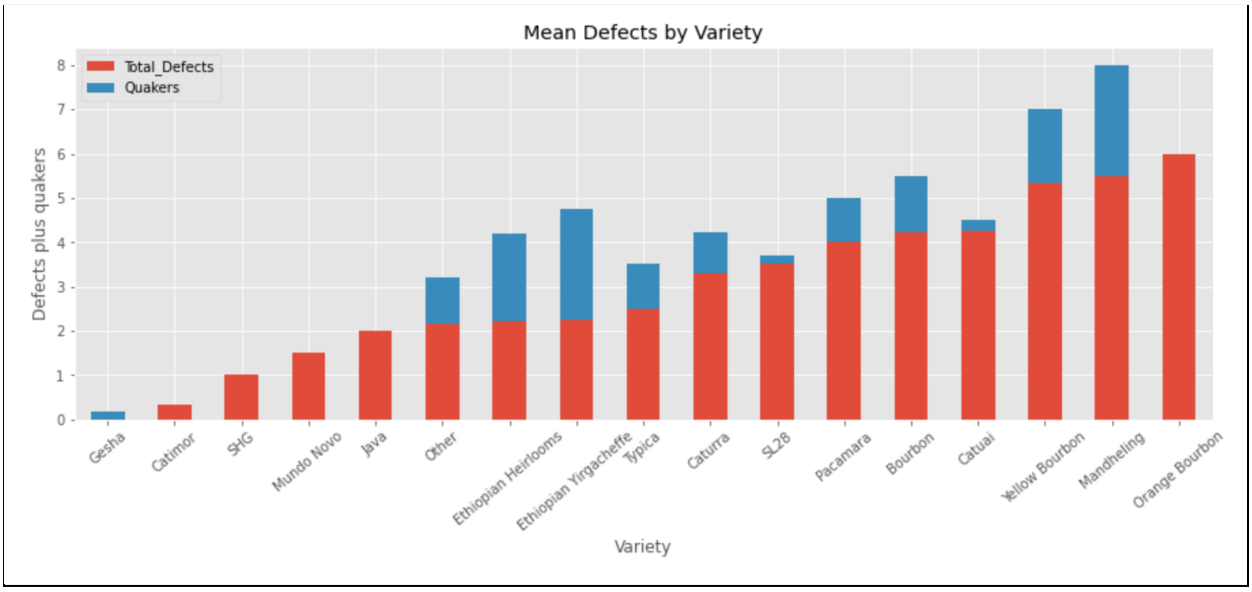


Our final graph shows us which countries have the largest percentages of Specialty and Premium coffees. In our data, the coffees from Burundi, Ecuador, El Salvador, Peru, Thailand, and Vietnam are all Specialty grade.

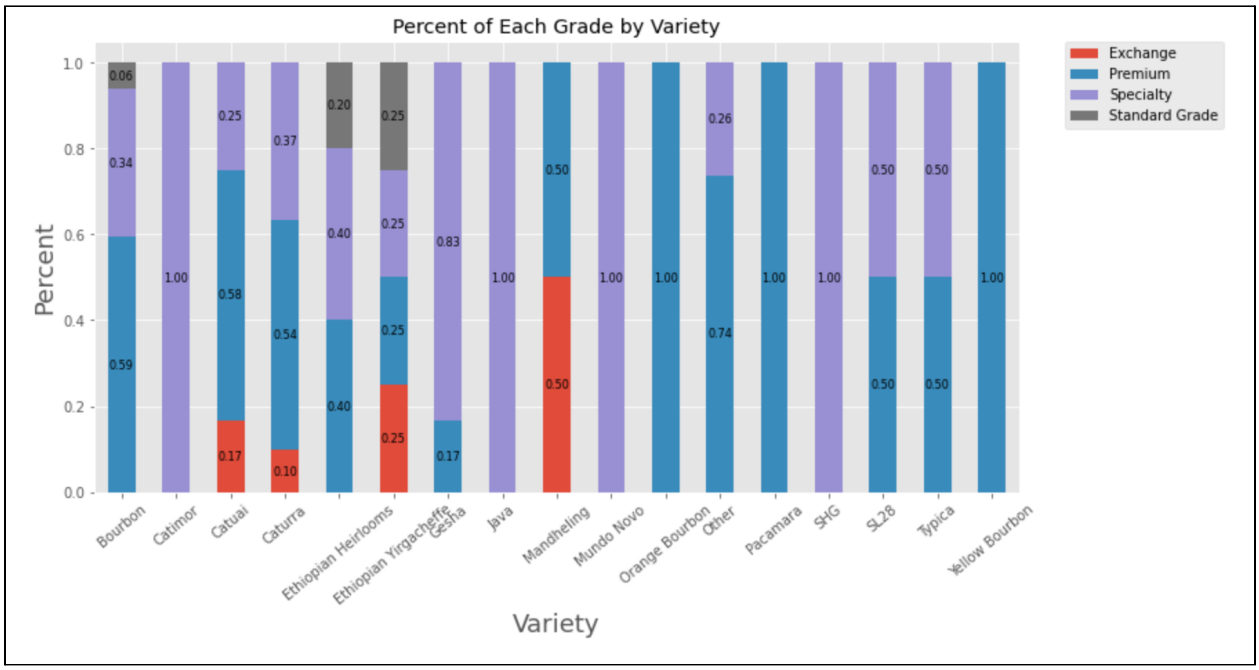
Coffee Grade and Variety



If we imagine the scatter plot above divided into four quadrants, our best coffee would come from the countries in the upper left quadrant (high Total Cup Points and better Coffee Grade). For this data, that would include Typica, Geisha, SHG, and SL28.



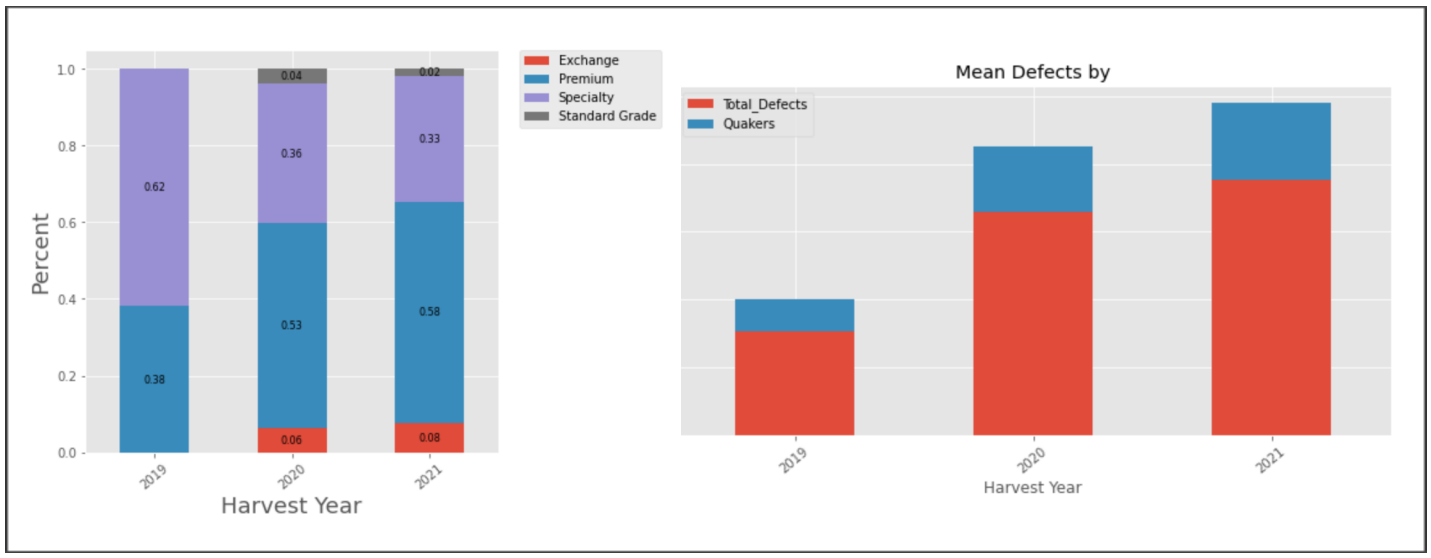
Here we can see that Gesha, Catimor, and SHG have the coffees with the lowest number of defects on average.



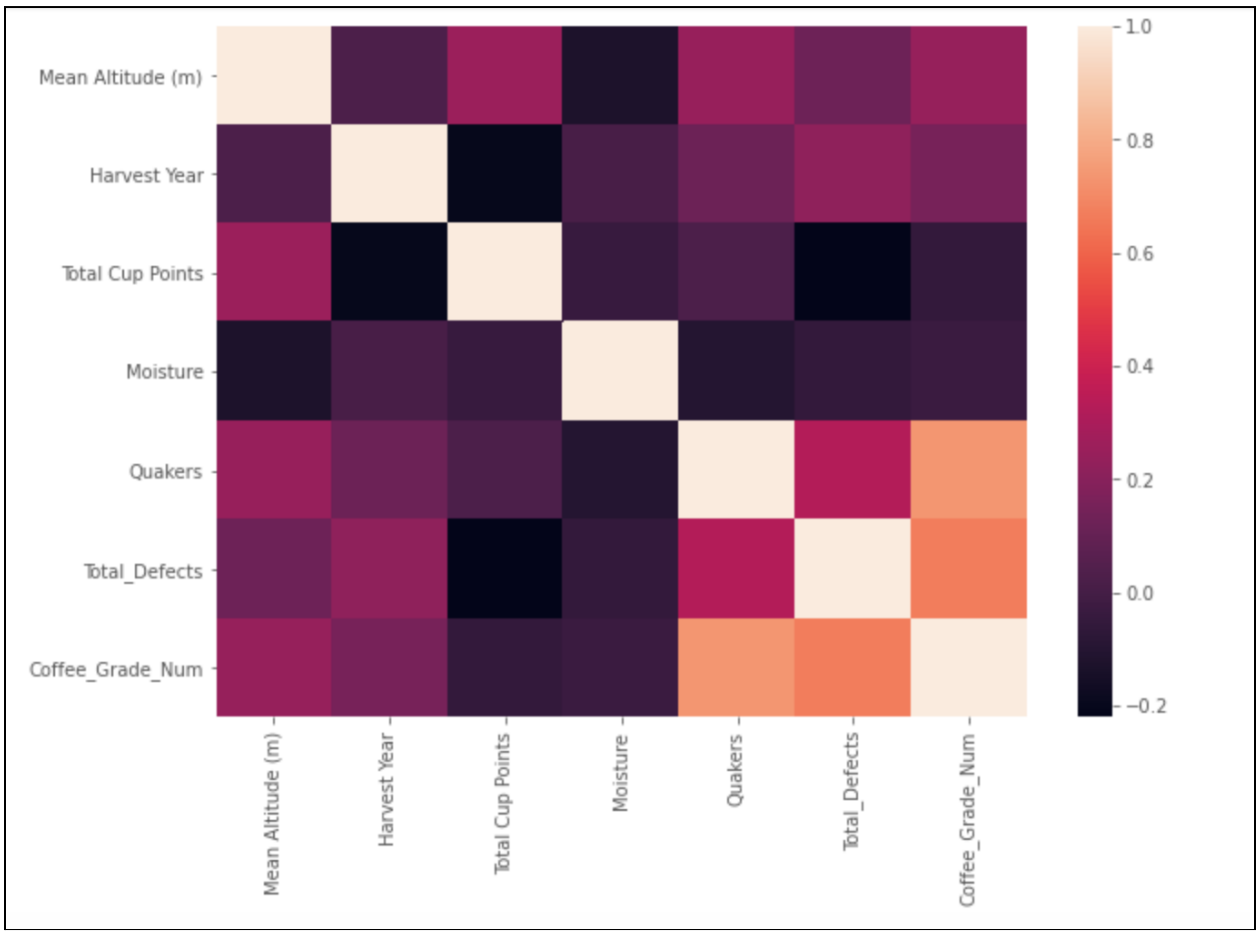
In our data, the Catimor, Java, Mundo Novo, and SHG varieties are all Specialty grade.

Coffee Grade and Other Categorical Features

Our other categorical features looked into were Harvest Year, Processing Method, and Color.
Based on our analysis, it is hard to say whether there is a clear advantage to any of the categorical features above. The Harvest Year of 2019 seems to be out performing the other years consistently but this could be due to random error.



Heat & Correlation Matrix



	Mean Altitude (m)	Harvest Year	Total Cup Points	Moisture	Quakers	Total_Defects	Coffee_Grade_Num
Mean Altitude (m)	1.00	0.03	0.25	-0.13	0.24	0.13	0.24
Harvest Year	0.03	1.00	-0.21	0.01	0.12	0.22	0.15
Total Cup Points	0.25	-0.21	1.00	-0.04	0.02	-0.22	-0.06
Moisture	-0.13	0.01	-0.04	1.00	-0.10	-0.06	-0.03
Quakers	0.24	0.12	0.02	-0.10	1.00	0.33	0.74
Total_Defects	0.13	0.22	-0.22	-0.06	0.33	1.00	0.67
Coffee_Grade_Num	0.24	0.15	-0.06	-0.03	0.74	0.67	1.00

Based on the heatmap and correlation matrix, it's easy to see that **there aren't any highly correlated features** except for coffee grade and total number of defects, which makes sense considering the fact that coffee grade is a function of total defects. For our model, we will not include Total Defects as an independent variable because of this relationship. The idea is that we want to predict the coffee grade before it is evaluated for defects.

EDA Conclusion

Based on our EDA, we have seen how lower elevations tend to increase the coffee grade and that there is statistical significance between the altitudes of the Specialty coffee and the Premium Coffee. There also seems to be a correlation between Total Cup points (which comes from people grading the coffee's more subjective features) and Coffee Grade (which comes from the number of defects and quakers and is therefore more objective).

We have seen how the country from where the coffee was produced, as well as the specific variety of coffee affects the Coffee Grade with some of **our best coffee (in general) coming from Taiwan, Burundi, Ecuador, and Vietnam** and **our best varieties being Gesha, SHG, SL28, and Catimor**.

Lastly, we looked at our other categorical features such as Color, Harvest Year, and Processing Method to gain insights on how each affects the Coffee Grade.

IV. Feature Engineering

In this section, the features will be engineered and modified in preparation for modeling. We will combine small values into lumped categories, select which features to use for our models, and one-hot-encode categorical data to label them numerically.

Combining Variables

Because of our small dataset, our categories with the fewest values were combined into one lumped category for our Countries, Varieties, and Coffee Grade features. This will improve the modeling process and help prevent over/under fitting of the test data. The resulting value counts are the following:

Country	Value Count
Other	54
Guatemala	36
Brazil	21
Colombia	17
Ethiopia	12
Costa Rica	10

Variety	Value Count
Other	55
Caturra	41
Bourbon	32
Catuai	12
SL28	10

Coffee Grade	Value Count
Specialty	58
Premium	79
Grade 3 or Lower	13

Even with this lumping, the dataset is highly unbalanced and that will be taken into consideration while modeling and analyzing results.

Feature Selection

Because our problem requires us to predict coffee grade without having already graded it for defects, quakers, and all other graded criteria, we will create a new data frame without these features. The useful features we will include are Country of Origin, Variety, Mean Altitude (m), Harvest Year, Processing Method, Moisture, Color, and our target variable, Coffee Grade.

One-Hot-Encoding

Lastly our data frame was finalized by one-hot-encoding our categorical features. This will allow us to run the models on the data.

V. Modeling

This section describes the models we tested and the tuning that went into each test. The results are analyzed and a final model is chosen. Because this is a **multi-class classification problem**, the models tested in this project were:

- **Logistic Regression,**
- **Random Forest Classifier**
- **Gradient Boosting**
- **Catboost**

Defining Model Success

Because multiple models were tested, there needed a way to compare them to choose the 'best' model. For this particular context -- that is, coffee grade predictions -- it is important that a coffee is graded accurately. If coffee is graded higher than it should be, the customer will feel like they are being lied to or ripped off. They expect a certain quality and are receiving a worse quality product. On the other hand, coffee that is graded lower than it should be means that the seller is not receiving just compensation for the quality of the product they are selling. Though both errors are important to consider, I think that most businesses would rather make the mistake of selling *really* good low-grade coffee than *really* bad "specialty" coffee. So, though **Accuracy and F1 score will be important metrics** to consider while conducting our tests, we want to also look at the model closely as to what is being labeled what. In other words, given a similar Accuracy and F1 score, does one model tend to label coffees *above* or *below* the actual grade compared to others. In these instances, the model that tends to low-ball the coffee grade, so-to-speak, will be prefered.

Model Results

	model	name	train_accuracy	test_accuracy	train_f1	test_f1
9	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	Random Forest Balanced Subsample	0.9911	0.8421	0.9911	0.8421
7	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	Random Forest Out of the Box	0.9911	0.8421	0.9911	0.8421
8	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	Random Forest Balanced	0.9911	0.8421	0.9911	0.8421
22	RandomizedSearchCV(cv=3,\n estimator=<catboost.core.CatBoostClassifier object ...	CatBoost w/ Random Search	0.6696	0.7895	0.6696	0.7895
21	<catboost.core.CatBoostClassifier object at 0x7f85d855b5e0>	Catboost Out of the Box	0.6696	0.7895	0.6696	0.7895
20	<catboost.core.CatBoostClassifier object at 0x7f85d828be80>	Catboost Out of the Box	0.6696	0.7895	0.6696	0.7895
19	RandomizedSearchCV(cv=3, estimator=GradientBoostingClassifier(), n_iter=100,\n ...	Gradient Boosting RandomSearch	0.6875	0.7632	0.6875	0.7632
18	([DecisionTreeRegressor(criterion='friedman_mse', max_depth=3,\n random_sta...	Gradient Boosting Out of the Box	0.9732	0.7632	0.9732	0.7632
12	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	RandomForestClassifier(random_state=10) with top 2 features	0.8304	0.7105	0.8266	0.7155
0	LogisticRegression()	Log Reg Out of the Box	0.7589	0.7105	0.7589	0.7105
16	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	RandomForestClassifier(random_state=10) with top 10 features	0.9911	0.6842	0.9911	0.7044
17	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	RandomForestClassifier(class_weight='balanced', random_state=10) with top 10 features	0.9911	0.6842	0.9911	0.7044
6	GridSearchCV(cv=3, estimator=LogisticRegression(),\n param_grid={'C': [100, 10, 1.0,...	Log Reg w/ GridSearch	0.7500	0.6579	0.7500	0.6579
5	LogisticRegression(max_iter=10000)	Log reg model with 17 PCA features	0.7500	0.6579	0.7464	0.6174
3	LogisticRegression(max_iter=10000)	Log reg model with 7 PCA features	0.6786	0.6579	0.6455	0.6431
2	LogisticRegression(max_iter=10000)	Log reg model with 4 PCA features	0.6429	0.6579	0.6084	0.6270
1	LogisticRegression(max_iter=10000)	Log reg model with 2 PCA features	0.5089	0.6316	0.4724	0.5834
13	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	RandomForestClassifier(class_weight='balanced', random_state=10) with top 2 features	0.8214	0.6316	0.8257	0.6523
4	LogisticRegression(max_iter=10000)	Log reg model with 12 PCA features	0.7232	0.6316	0.7206	0.6222
10	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	Balanced Random Forest Classifier	0.7679	0.6053	0.7679	0.6053
14	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	RandomForestClassifier(random_state=10) with top 4 features	0.9554	0.6053	0.9555	0.5776
15	(DecisionTreeClassifier(max_features='auto', random_state=1165313289), DecisionTreeClassifier(ma...	RandomForestClassifier(class_weight='balanced', random_state=10) with top 4 features	0.9554	0.6053	0.9554	0.5776
11	RandomizedSearchCV(cv=3,\n estimator=RandomForestClassifier(class_weight='bala...	Random Forest Balanced RandomSearch	0.9911	0.6053	0.9911	0.6053

Comparing Top Models

As you can see by the table, the best performing models are the Random Forest Classifiers. The top three models have the same Accuracy and F1 scores so to determine which model is the best, the confusion matrices must be analyzed.

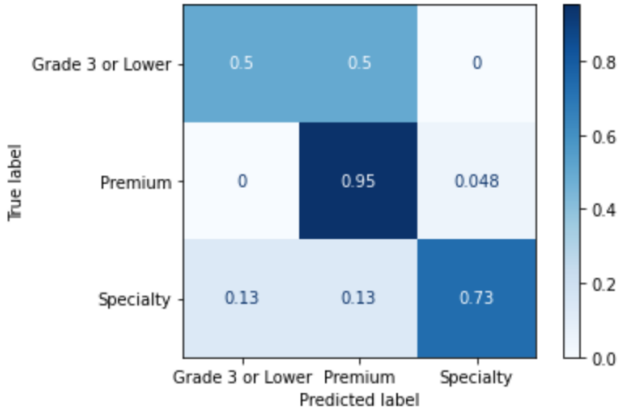
Random Forest Out of the Box

```
Validation - Classification report
                   precision    recall  f1-score   support

Grade 3 or Lower       0.33      0.50      0.40         2
         Premium       0.87      0.95      0.91        21
       Specialty       0.92      0.73      0.81        15

        accuracy                           0.84        38
       macro avg       0.71      0.73      0.71        38
    weighted avg       0.86      0.84      0.85        38
```

```
<Figure size 648x648 with 0 Axes>
```



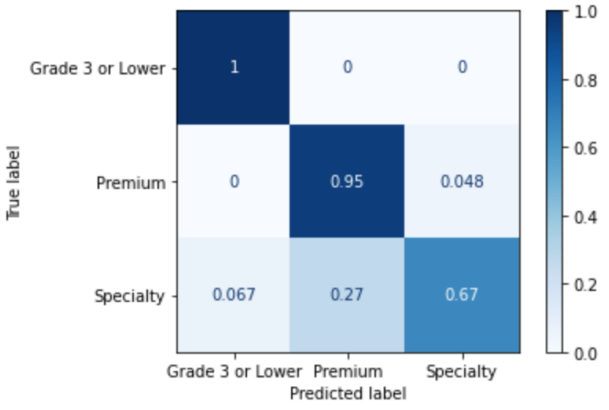
Random Forest Balanced

```
Validation - Classification report
                   precision    recall  f1-score   support

Grade 3 or Lower       0.67      1.00      0.80         2
         Premium       0.83      0.95      0.89        21
       Specialty       0.91      0.67      0.77        15

        accuracy                           0.84        38
       macro avg       0.80      0.87      0.82        38
    weighted avg       0.85      0.84      0.84        38
```

```
<Figure size 648x648 with 0 Axes>
```



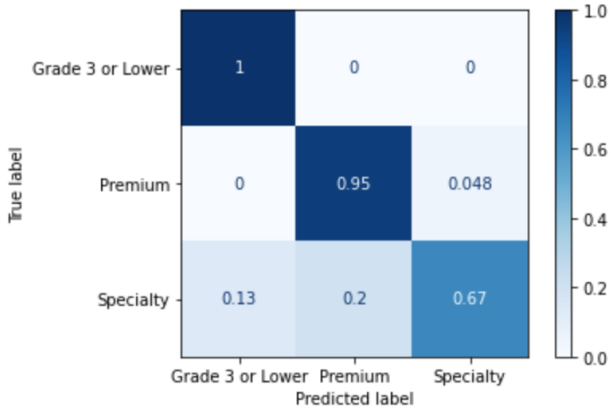
Random Forest Balanced Subsample

```
Validation - Classification report
                  precision    recall  f1-score   support

Grade 3 or Lower       0.50      1.00      0.67         2
         Premium       0.87      0.95      0.91        21
       Specialty       0.91      0.67      0.77        15

        accuracy                           0.84        38
       macro avg       0.76      0.87      0.78        38
    weighted avg       0.87      0.84      0.84        38
```

```
<Figure size 648x648 with 0 Axes>
```



VI. Conclusion and Recommendations

Based on the results from these three models, our **Random Forest Balanced slightly out-performed the others**. Though they had similar accuracy and overall F1 scores, the precision and recall scores for the individual classes were better. In other words, the Random Forest Balanced model did a better job at not giving the coffees a *higher* grade than what they actually were. Based on the context of our problem, this is the preferred model.

When it comes to purchasing high grade coffee, the recommendations would be to prioritize coffees with the following features:

Grown in altitude ranging from 1250-1600m
Countries: Taiwan, Burundi, Ecuador, and Vietnam
Varieties: Gesha, SHG, SL28, and Catimor

VII. Suggestions for Further Analysis

- Collect more data for better modeling
- More intensive hyperparameter tuning of models
- Incorporate weather/precipitation data along with the country and altitude
- Include robusta coffees