CSCI 1070 Final Report

Banana Analysis

David Joseph

One thing that a lot of people never think about is where the source of our everyday objects come from. A computer may be made with parts that come from everywhere around the world, a datasheet with references from Antartica, or even something as simple as a banana. For my project, I was interested in how the process of my favorite fruit, a banana, is sourced. More specifically, I was interested in finding out what qualities of a banana makes it the best. In a more real world setting, the scenario I crafted was that I was a data scientist tasked with figuring out how to find the best types of bananas so that my company can figure out where to source their bananas from. This leads us to the research question, what variables make the best types of bananas? What varieties as well?

To answer this question, I had to hypothesize different ways to go about finding a correlation between variables and the best "quality score" of a banana. After finding my data set, I was able to craft a hypothesis, then alternative and null hypotheses as well. The variables of interest I found during my research from the dataset were quality_score, as it tells me which bananas have the best quality, and firmness, sugar content, variety and ripeness index. My hypothesis was the bananas with the highest quality score will have a positive relationship with bananas with the highest ripeness index. Thus, the null hypothesis is that there is no relationship between quality score and ripeness index. Finally, the alternative hypothesis is that there is a relationship between quality score and ripeness index. With which varieties of bananas were the best, my hypothesis was that Cavendish would be the best, as they are a very popular banana type. A null hypothesis for this would be that Cavendish does not have any correlation with quality score and an alternative would be that Cavendish has a positive correlation with quality score.

Using the data set, multiple data science strategies we learned in class can be applied to solve this question. I plan to use said strategies to make up models that I can present to help understand the relationship between the variables. First, I will check the dataset to see if it's clean, then do proper cleaning. For variables that need to be encoded with one-hot encoding, I will do that. Then I will use multiple and simple linear regression models, using MSE to evaluate how good they are, and using the coefficients they produce to draw conclusions. Afterwards, I will make correlation matrices between the three variables to see correlation between the independent variables, with our dependent variable being quality score. Then, I will one-hot encode a variety to change from categorical variables to discrete/numerical values we can work with. I will then use this to make a table that shows the correlation between all the banana varieties and quality score. After that, I will make basic models showcasing the three variables modeled against quality score, using a box plot to distinguish between the varieties and doing a scatter plot for the other variables.

Before I get into how I coded all this and my results, I will need to do some background explanation on the variables. First, let's define our dependent variable of quality score, which can be described as a numerical score, likely on a scale of 1-4 that rates the overall quality of the sample. Quality scores can tell us which banana samples are the best ones. Using quality score as a dependent variable, we can use Linear Regression models to compare other variables to it and make conclusions about the data set.

The variety column is defined as the breed of banana, such as Cavendish, Red Dacca, or Lady Finger. Variety can give us context to how the sample's physical characteristics affect other variables and how growing conditions can affect the yield. The full list of the varieties are: Manzano, Plantain, Burro, Red Dacca, Fehi, Lady Finger, Blue Java, and Cavendish. Variety was interesting to me because I wanted to know what type of banana was the best and this was the best way to do so.

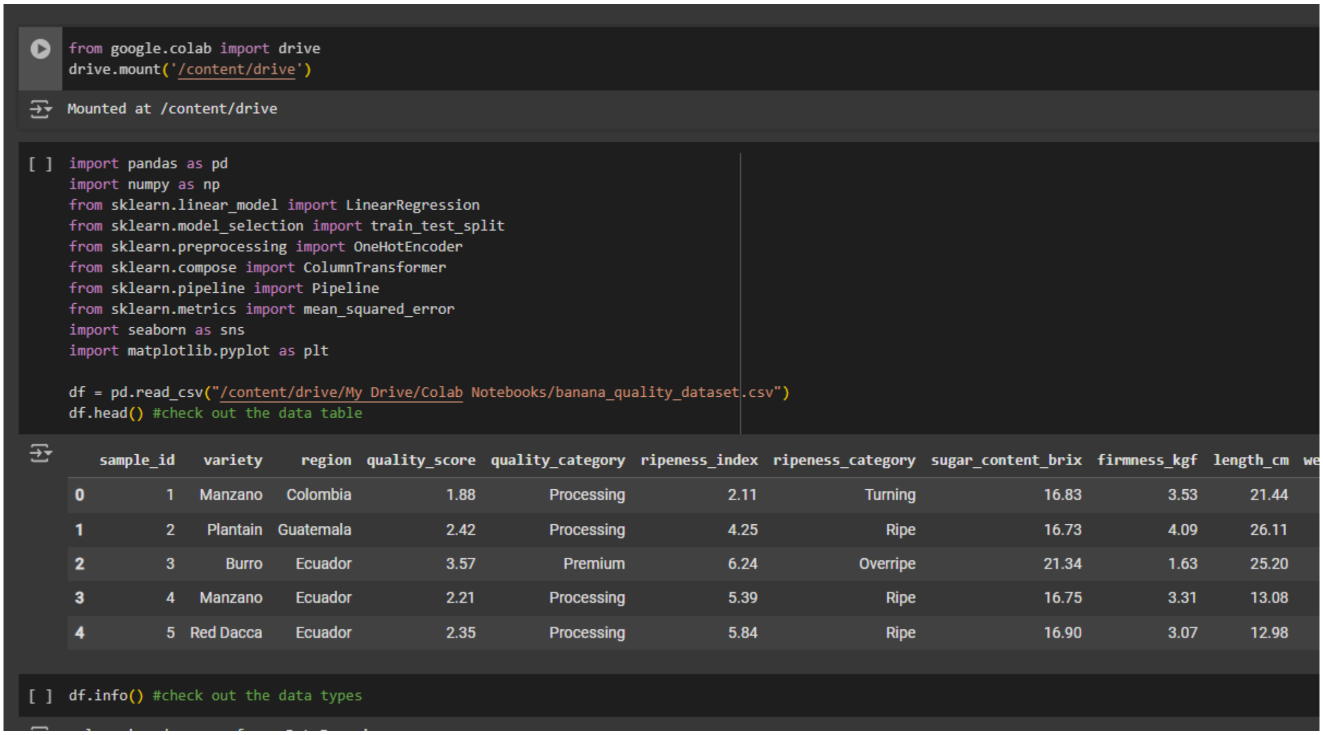
Next up is firmness. Firmness is the firmness of the banana, which can be measured in kilograms-force. What's interesting about firmness is that it can be a useful way to understand the quality and ripeness of a banana. Firmness can indicate the texture and maturity of the sample we are working with, which can provide good insights into how ripe a banana is and the overall quality of it. There's a lot of different ways to measure firmness, but this dataset measures it in kilograms-force. I actually found a research paper that measured firmness with Near-infrared (NIR) spectroscopy, a technique that utilizes the near-infrared region of the electromagnetic spectrum to identify and evaluate the physical and chemical properties of a sample, in this case firmness of the banana.

Next, we have to explain ripeness_index. Ripeness index is a numerical value that ranges from one to ten that can tell us how ripe a banana is. A ripeness index of 1 would tell us that the sample is green/unripe. A ripeness index of 10 would indicate that the sample is overripe, which is not ideal for our company's mission of finding the best banana. Ripeness index is important because we can see how riper bananas affect other variables such as quality score and sugar content.

Finally, we have sugar content. Sugar content of the banana is pretty intuitive, it is how sugary a sample of the banana is. We can measure this using Brix, which is a popular way to define sugar content in many fruits. When using the Brix method, we compare how much pure sucrose is in a sample with the equation or notation: 1 degree Brix (°Bx) = 1g of sucrose / 100g of solution. You can measure Brix with multiple different methods, Hydrometer, Pycnometer, Optical refractometers, and more. Brix also has applications in more fields than just the food industry, for example the chemical industry.

For my data and methods section, I was able to locate my data source from the website kaggle, which provides many data sets, mainly for competitions and open use. The data set is a comprehensible sheet or csv file containing important information about where a banana is sourced from and what types those bananas are. The dataset has a column named sample id, which allows anyone viewing the data to track different samples across the world and reference different samples uniquely. One upside about this dataset was that it was pre-cleaned already. The only "cleaning" I would really have to do is one-hot encode the data.

Other than that, the data was pretty straight forward to work with. All the things we did with homework assignments #4, #5, and #6 would come into action here. First I would import all my data via google drive then convert the data frame to a df. Using commands like df.info, df.head and df.describe would become handy in figuring out how clean the data was.



```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

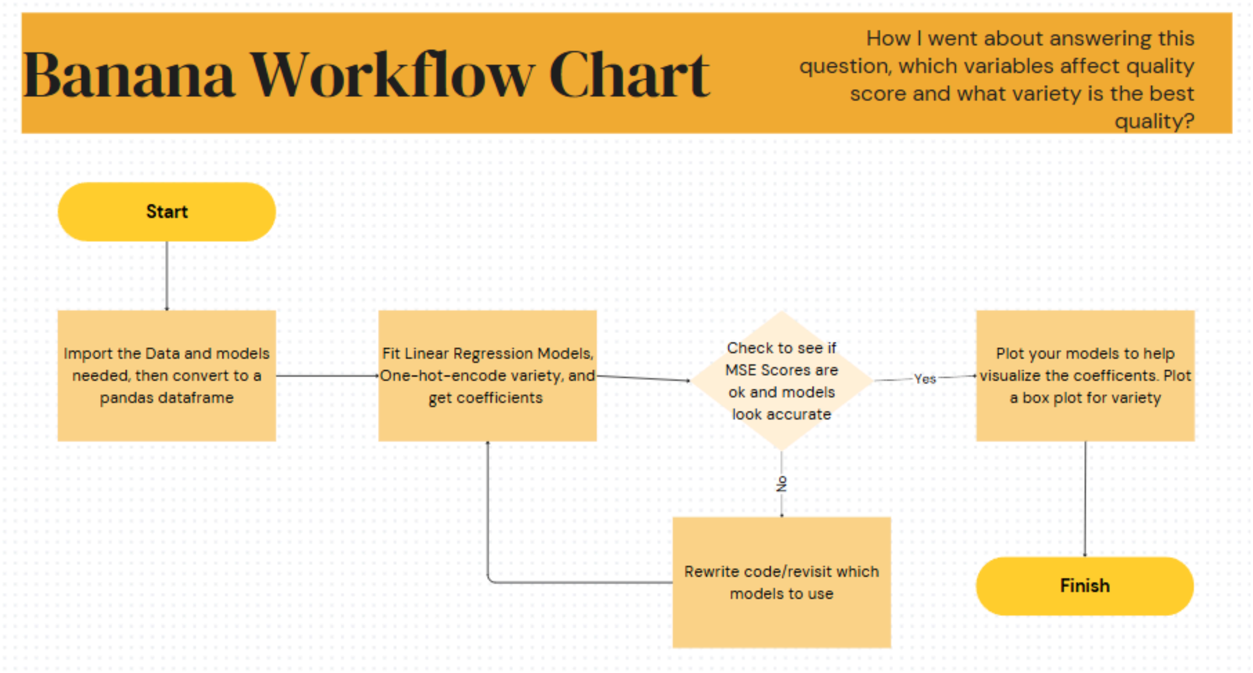
```
[ ] import pandas as pd
    import numpy as np
    from sklearn.linear_model import LinearRegression
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import OneHotEncoder
    from sklearn.compose import ColumnTransformer
    from sklearn.pipeline import Pipeline
    from sklearn.metrics import mean_squared_error
    import seaborn as sns
    import matplotlib.pyplot as plt

    df = pd.read_csv("/content/drive/My Drive/Colab Notebooks/banana_quality_dataset.csv")
    df.head() #check out the data table
```

```
   sample_id   variety     region  quality_score  quality_category  ripeness_index  ripeness_category  sugar_content_brix  firmness_kgf  length_cm  we
0          1   Manzano   Colombia           1.88        Processing            2.11            Turning               16.83          3.53      21.44
1          2  Plantain  Guatemala           2.42        Processing            4.25               Ripe               16.73          4.09      26.11
2          3     Burro    Ecuador           3.57           Premium            6.24           Overripe               21.34          1.63      25.20
3          4   Manzano    Ecuador           2.21        Processing            5.39               Ripe               16.75          3.31      13.08
4          5 Red Dacca    Ecuador           2.35        Processing            5.84               Ripe               16.90          3.07      12.98
```

```
[ ] df.info() #check out the data types
```

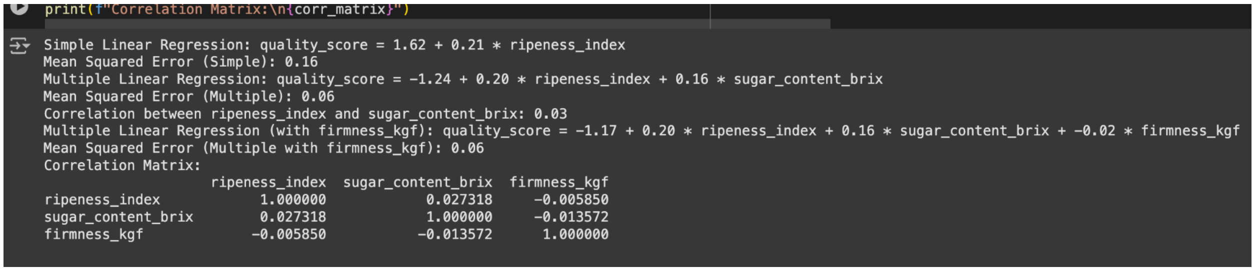
Eventually after figuring out how the data was structured, I used df.variety.unique() to figure out how many different types of banana were in the data set. Afterwards, I felt comfortable to begin making my Linear Regression models. I started with simple linear regression, using the ripeness score as my independent against the quality score as my dependent variable. I used MSE to evaluate the model and printed my coefficient for the model. Then I repeated these steps for Multiple Linear Regression. I did this for all my variables up to variety, which I would end up using pd.get_dummies to one hot encode. I then still did multiple linear regression, but using variety as my alpha. The equation can be modeled like this: "quality_score = ONE_HOT_VARIETY + beta*ripeness_score + gamma*sugar + delta*firmness".


Banana Workflow Chart
How I went about answering this question, which variables affect quality score and what variety is the best quality?
Start
Import the Data and models needed, then convert to a pandas dataframe
Fit Linear Regression Models, One-hot-encode variety, and get coefficients
Check to see if MSE Scores are ok and models look accurate
Yes
No
Plot your models to help visualize the coefficents. Plot a box plot for variety
Rewrite code/revisit which models to use
Finish

Check out this basic workflow infographic. This is the basic skeleton for my approach. The finish button implies that I would use the models in the last step to make a conclusion based on the results.

The models I will be using are also pretty straightforward. I will be using a box plot graph to showcase how different banana varieties correlate with ripeness score to see specifically how Cavendish fares with the others. I will also plot the net coefficients of all four variables against quality score in a single graph, including all the banana varieties to showcase how much the quality score really varies. Finally, I will graph scatter plots of firmness, sugar content, and ripeness index against quality score to showcase relationships between those variables. Finally, I printed a correlation matrix for firmness, sugar content, and ripeness index to see if they are correlated together in any way.

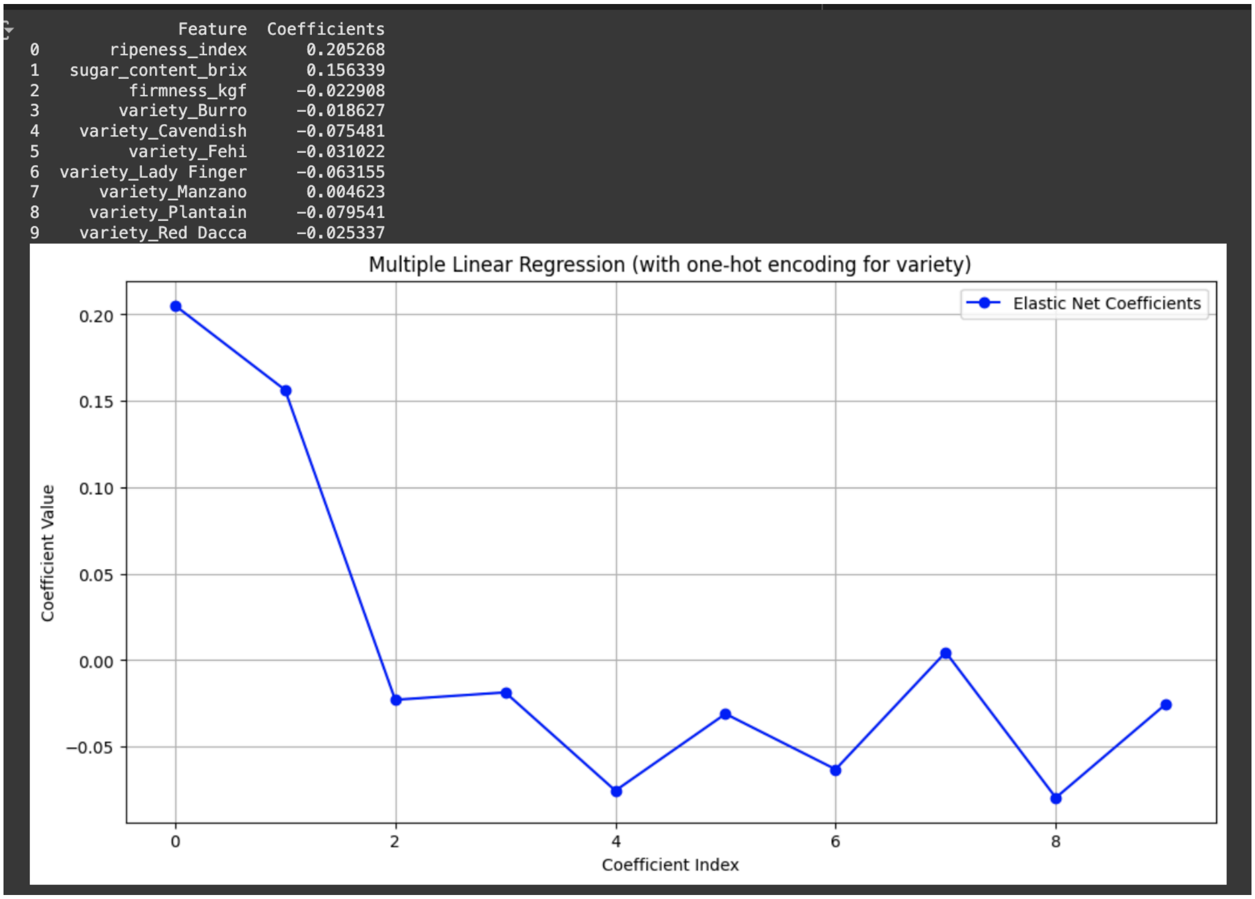
In order to decide if the results from our final models are substantial and if we can make conclusions from them, we must first evaluate if the linear regression models we made are satisfactory. To do this, we can take the MSE or mean squared error value for each of them.



```
print(f"Correlation Matrix:\n{corr_matrix}")
```

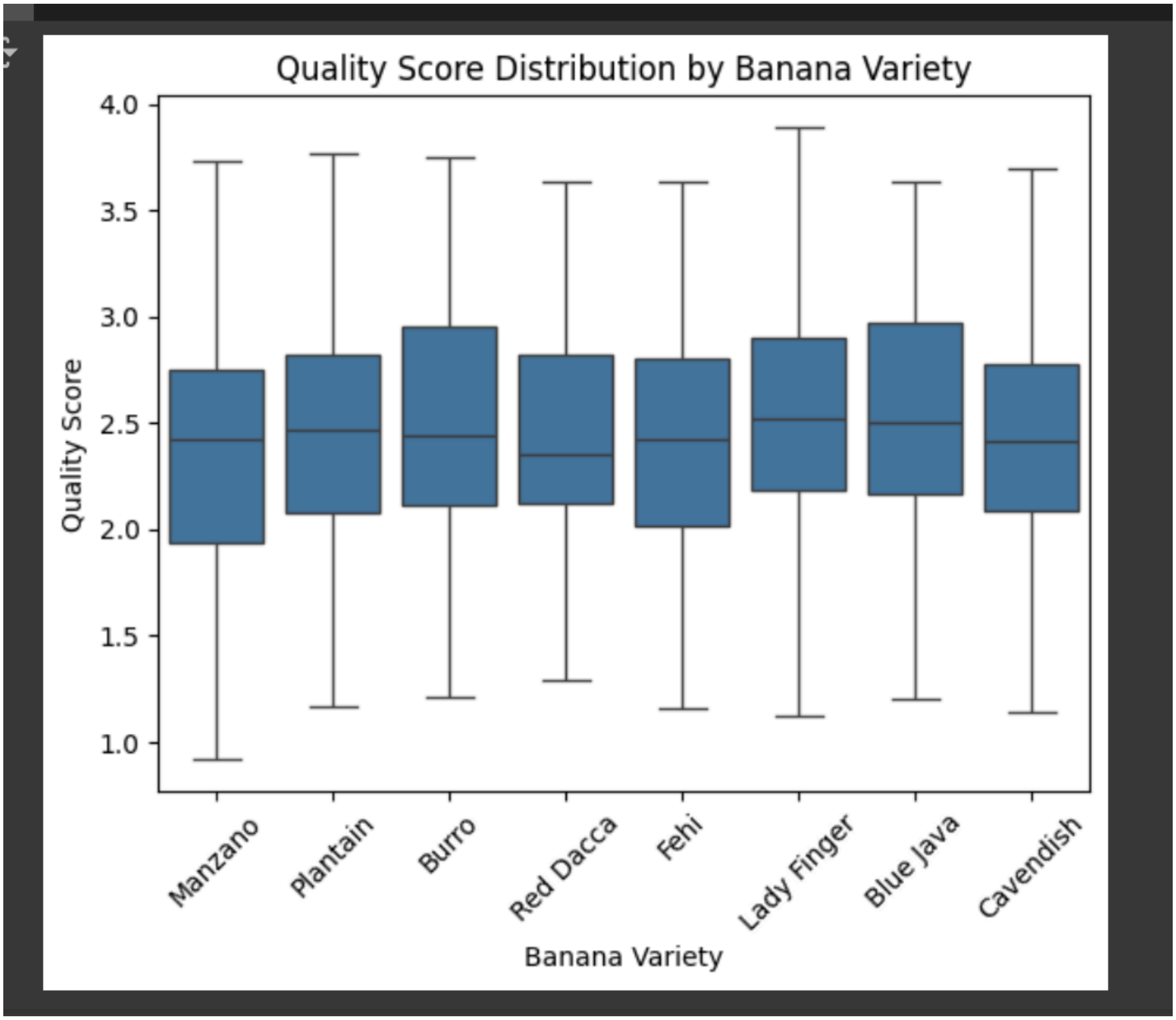
```
Simple Linear Regression: quality_score = 1.62 + 0.21 * ripeness_index
Mean Squared Error (Simple): 0.16
Multiple Linear Regression: quality_score = -1.24 + 0.20 * ripeness_index + 0.16 * sugar_content_brix
Mean Squared Error (Multiple): 0.06
Correlation between ripeness_index and sugar_content_brix: 0.03
Multiple Linear Regression (with firmness_kgf): quality_score = -1.17 + 0.20 * ripeness_index + 0.16 * sugar_content_brix + -0.02 * firmness_kgf
Mean Squared Error (Multiple with firmness_kgf): 0.06
Correlation Matrix:
                    ripeness_index  sugar_content_brix  firmness_kgf
ripeness_index            1.000000            0.027318     -0.005850
sugar_content_brix        0.027318            1.000000     -0.013572
firmness_kgf             -0.005850           -0.013572      1.000000
```

In statistics, MSE is the average of the squared residuals, which is the distance between the predicted and the actual values in the dataset. A lower MSE is better, because that means our model’s predicted values are more accurate. As we can see, our MSE of .16 for simple linear regression is good as well as our MSE of .06 for multiple linear regression. This means the analysis we use from these models will be accurate! I also printed the correlation matrix between the three variables I was using on the right side of our equations or our X = side to see if they have any correlation with each other. There is very little correlation between these variables, which may tell us that these variables do not affect each other much.

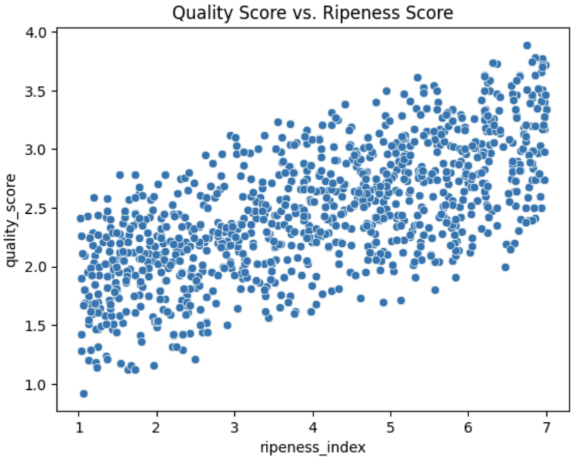


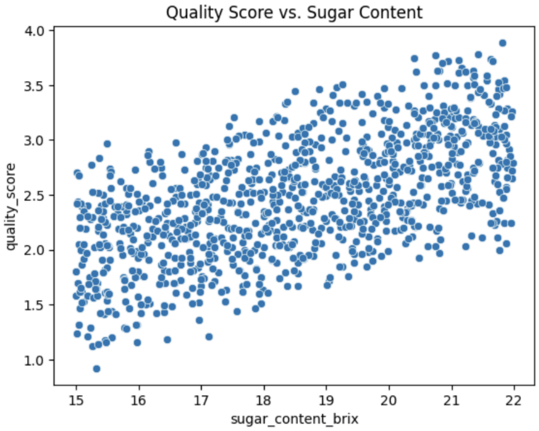
```
              Feature  Coefficients
0       ripeness_index      0.205268
1   sugar_content_brix      0.156339
2          firmness_kgf     -0.022908
3        variety_Burro     -0.018627
4    variety_Cavendish     -0.075481
5         variety_Fehi     -0.031022
6  variety_Lady Finger     -0.063155
7      variety_Manzano      0.004623
8      variety_Plantain     -0.079541
9     variety_Red Dacca     -0.025337
```

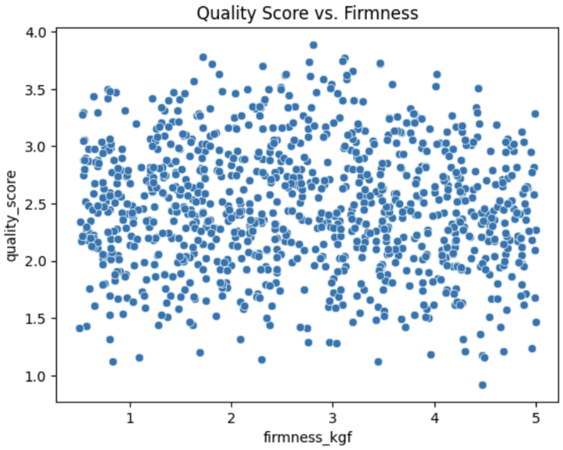
Here, I found the coefficients of the multiple linear regression, with the varieties of banana being one-hot encoded. We can see the correlation kind of drop off after you leave ripeness index, sugar content, and firmness. All the coefficients aren't super massive, which may not show a major correlation between these variables. These coefficients mean a one-unit change in the corresponding independent variable will result in a larger change in the dependent variable. So really small numbers may not point towards a high correlation.



Here, I graphed the banana variety against the quality score. This showed some interesting results. All the varieties have very similar box plots. We can tell that Cavendish has at least a lower average quality score than other varieties, like Blue Java and Burro having higher quality scores. It might be hard to make conclusions about varieties as a whole from how close these box plots are to each other. Below are the scatterplots of correlations between ripeness score, sugar content, firmness against quality score.


Quality Score vs. Ripeness Score


Quality Score vs. Sugar Content


Quality Score vs. Firmness

You can see from the models that as the ripeness index increases, quality score increases as well. The same thing happens with sugar content. This can showcase a positive correlation between sugar content with quality score and ripeness index with quality score. This tells us that the samples that are riper and the ones that have a higher sugar content. Firmness seems to have no correlation with quality score. This suggests that firmness doesn't really affect the quality score of a sample.

Now, using the information we have collected over this report, we can revisit our hypotheses and draw a conclusion as to what the best bananas are. The variety of the banana doesn't seem to have any specific banana that is better than the other. The best varieties, or the ones with technically the highest quality score are Burro and Blue Java, so these would be the bananas that I recommend. After that, I found that the best variables that affect quality score were sugar content and ripeness. It seemed that as ripeness and sugar content increased, so did the quality score. So I would advise the company to pick Burro and Blue Java bananas that have a good ripeness score and a high sugar content. My hypothesis that Cavendish bananas were a good choice was not correct. I also found out that my hypothesis that the quality score would have the highest correlation with ripeness score. This was true, as it had the highest coefficient out of all the other independent variables I used. All in all, I was able to conclude that firmness and variety did not have major impacts on quality score, while variables like sugar content and ripeness did have a more significant role. If I had to do this project again, I would try to use deeper models to have more models to work with to make better conclusions with it. I also would use a data sheet where the variables had a lot more correlation than mine did, so I didn't have to work harder to draw conclusions.

Works Cited

Lipidomic Changes in Banana (Musa Cavendish) during Ripening and Comparison of Extraction by Folch and Blighdyer Methods, https://doi.org/10.1021/acs.jafc.0c04236.s001.

Mars_1010. “Banana Quality Dataset.” *Kaggle*, 7 Nov. 2024, www.kaggle.com/datasets/mrmars1010/banana-quality-dataset/data.

“Mean Squared Error: Definition, Formula, Interpretation and Examples.” *GeeksforGeeks*, GeeksforGeeks, 13 Aug. 2024, www.geeksforgeeks.org/mean-squared-error/.

Mettler-Toledo International Inc. all rights reserved. “Brix Measurement.” *Brix: The Essential Knowledge*, 7 Feb. 2023, www.mt.com/us/en/home/perm-lp/product-organizations/ana/brix-meters.html.