Milestone 4 - Gerardo Palacios

Summary

To reiterate the group discussions, we decided on two main approaches, PCA/CFA and regularized regression analysis in order to predict cupper points and the probability of the type of processing the coffee beans received. Part of the main issues on interpreting the predictors is that of dimensionality, differences in predictor values, and high multicollinearity. In order to extrapolate latent factors and to produce a sparser model, my portion of the analysis is to conduct PCA and ElasticNet regression. These two, in addition to creating associated plots for every model iteration, would assist into a holistic analysis when the group brings together all the avenues of analysis.

In earlier discussions, I explored the possibility of combining data from two different datasets, that of a publicly available weather API database in order to link for geographical variables into the dataset based on the associated regions. However, the problem with that approach was that of accuracy and consistency. Many of the regions with the dataset were using charecters that were unreadable. This is due to it not having a latin based name (i.e Asian regions with foreign charecters). Much of the data clean up revolved around matching these two sets, without much success. As a result, we decided to excluded all of the weather data. In this milestone, I chose a simpler data clean up, which only involved removing rows that did not have complete data, as well as rows that held data that did not make logical sense for producing beans (i.e Coffee beans produced at altitudes over 3,000 or bags of coffee under 100 lbs). Officially, subsetting the coffee dataset from an original 1,343 observations and 36 elements to 875 observations with complete data.

PCA Analysis

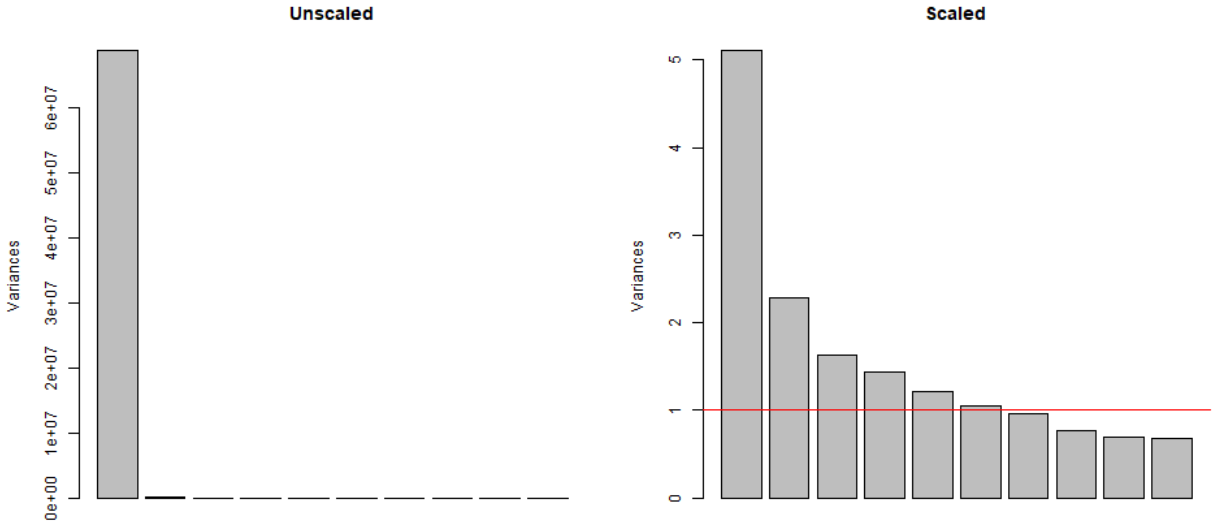
For the first portion of the analaysis I ran 2 different iterations of PCA. Each iteration involved comparing the results with its scaled|unscaled and rotated | unrotated counterparts.

1. No variables excluded

As it is apparant. Unscaled PCA held the entire variability in PC1, This would create uninterpretable results. Fortunately, using scaled PCA allowed the variability to become more evenly spread out. More importantly, I was able to use the variance value of 1 as a breaking point for determining the number of components.

```
PCA_Results[[1]]$Plots$`Scree plot Scaled vs Unscaled`
```

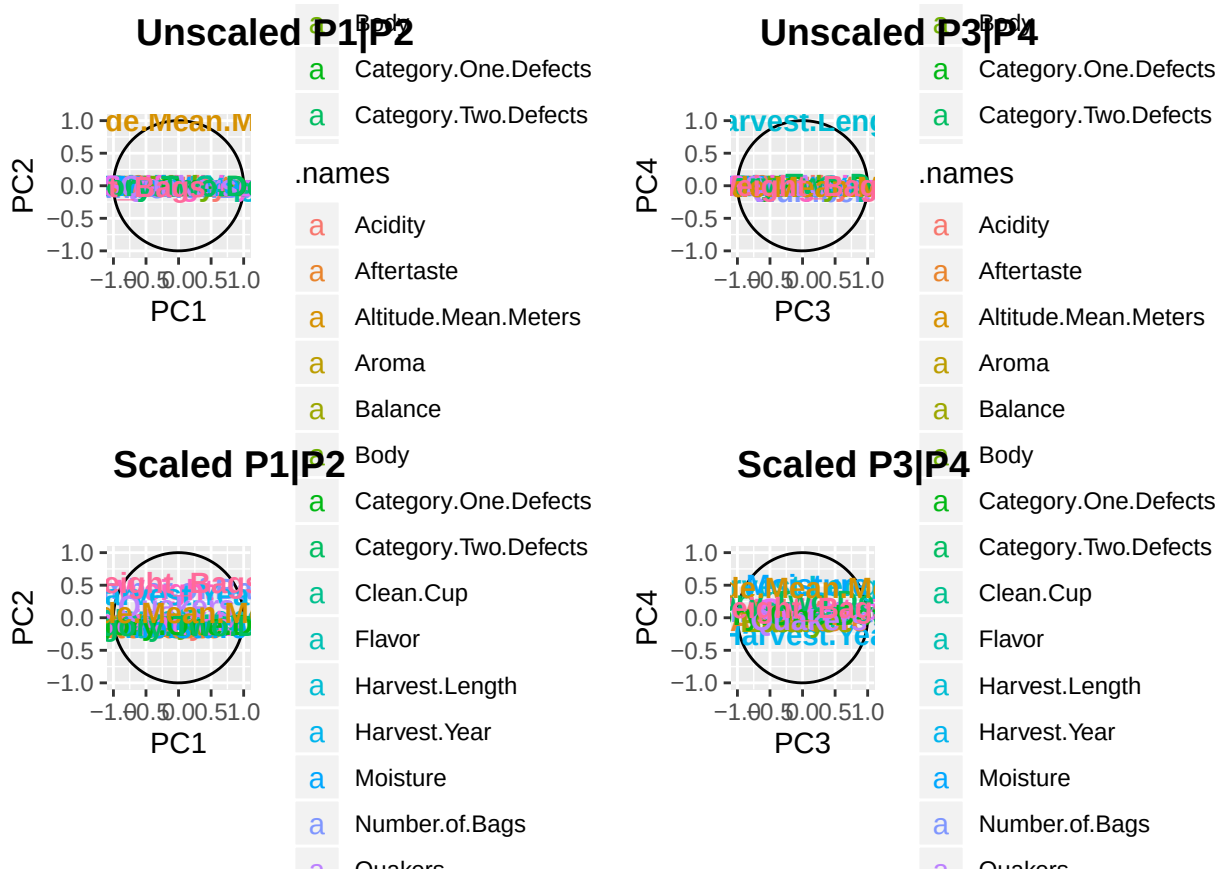
Scree Plots − Unscaled vs Scaled



As expected, using all the variables resulted in difficult interpretations. The output of the PCA results are not included due to the size. But, below are the PCA charts reflecting unscaled vs scaled PCAs. The scaled results, surprisingly are harder to interpret than that of the unscaled. Looking closlely at the Unscaled PCA plots, it becomes apparant that Altitude, Weight, number of bags, and harvest length are strongly correlated. However, since these are unscaled, it could also be suggesting that these variable values are adding artificial weights to those variables.

```
plot_grid(
PCA_Results[[1]]$Plots$`PCA Plot Unscaled - P1 and P2`,
PCA_Results[[1]]$Plots$`PCA Plot Unscaled - P3 and P4`,

PCA_Results[[1]]$Plots$`PCA Plot Scaled - P1 and P2`,
PCA_Results[[1]]$Plots$`PCA Plot Scaled - P3 and P4`,
ncol = 2,nrow = 2,align = "h", labels = c("Unscaled P1|P2","Unscaled P3|P4","Scaled P1|P2","Sca
)
```



2. Excluding at CF = 0.90

Using a correlation test with a confidence level of 90%, I was able to eliminate very high and very low correlated variables. From the original 29 numerical variables reduced to only 4 variables. The results below represent scaled vs unscaled. The scree plots are drastically different from the first iteration. Using 4 variables adjust the variability among all the components. However, it is difficult to determine an actual elbow point since so many of the components have variances close to 1. The coefficients of the PCA results suggest that number of bags and altitude seem to move in tandem with one another.

```
PCA_Results$`Excluding at CF = 0.90`$`PCA results`$`PrComp-Scaled`
```

```
## Standard deviations (1, .., p=4):
## [1] 1.0554004 1.0206754 0.9977604 0.9213174
##
## Rotation (n x k) = (4 x 4):
##                              PC1        PC2         PC3        PC4
## Number.of.Bags       -0.7487490 -0.1147776  0.01688234 -0.6526224
## Sweetness             0.3151214  0.8042087 -0.02065124 -0.5035081
## Quakers              -0.3848254  0.3557048 -0.77234983  0.3589697
## Altitude.Mean.Meters -0.4381584  0.4621132  0.63463710  0.4378406
```

```
summary(PCA_Results$`Excluding at CF = 0.90`$`PCA results`$`PrComp-Scaled`)
```

```
## Importance of components:
```

```
##                            PC1    PC2    PC3    PC4
## Standard deviation     1.0554 1.0207 0.9978 0.9213
## Proportion of Variance 0.2785 0.2604 0.2489 0.2122
## Cumulative Proportion  0.2785 0.5389 0.7878 1.0000
```

```
PCA_Results$`Excluding at CF = 0.90`$`PCA results`$`PrComp-notScaled`
```

```
## Standard deviations (1, .., p=4):
## [1] 409.6826927 126.1206534   0.7451999   0.4326634
##
## Rotation (n x k) = (4 x 4):
##                                  PC1           PC2           PC3           PC4
## Number.of.Bags       -2.671659e-02  0.9996429296 -3.852427e-04 -0.0002974405
## Sweetness            -3.449076e-05 -0.0002878968  2.714737e-02 -0.9996314002
## Quakers              -9.197895e-06  0.0003930700  9.996314e-01  0.0271472552
## Altitude.Mean.Meters -9.996430e-01 -0.0267165848  1.615939e-07  0.0000421900
```

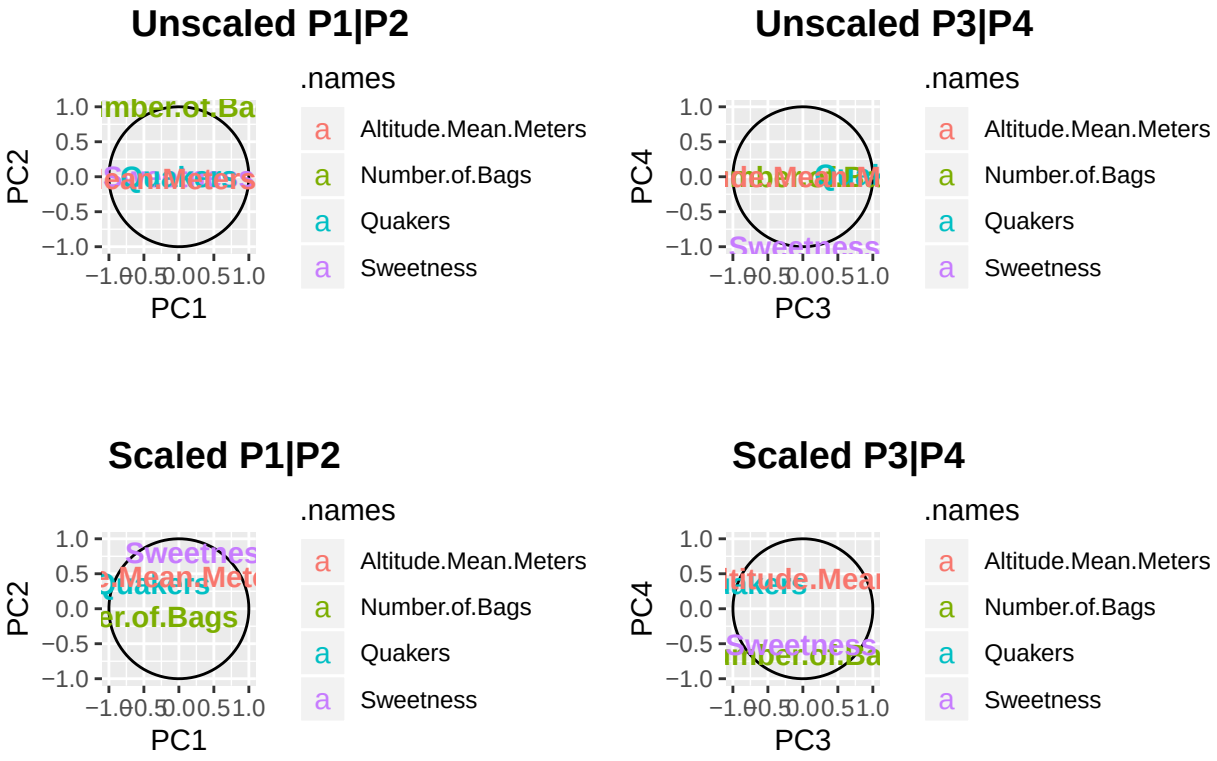
```
summary(PCA_Results$`Excluding at CF = 0.90`$`PCA results`$`PrComp-notScaled`)
```

```
## Importance of components:
##                             PC1       PC2    PC3    PC4
## Standard deviation     409.6827 126.12065 0.7452 0.4327
## Proportion of Variance   0.9134   0.08657 0.0000 0.0000
## Cumulative Proportion    0.9134   1.00000 1.0000 1.0000
```

The results of the PCA plots reveal are more promising. THe PCA plots reveal that mean altitude, quakers and sweetness are all positivly correlated with each other. It seems to suggest that quakers in the coffee bean and altitude have a strong relationship with each other.

```
plot_grid(
PCA_Results[[2]]$Plots$`PCA Plot Unscaled - P1 and P2`,
PCA_Results[[2]]$Plots$`PCA Plot Unscaled - P3 and P4`,

PCA_Results[[2]]$Plots$`PCA Plot Scaled - P1 and P2`,
PCA_Results[[2]]$Plots$`PCA Plot Scaled - P3 and P4`,
ncol = 2,nrow = 2,align = "h", labels = c("Unscaled P1|P2","Unscaled P3|P4","Scaled P1|P2","Sca
)
```



Regularization and Linear Models

When building the regularization models I decided to still run OLS in tandem to compare the results. In this case, I made two iterations of models. The difficulty to determine which model performed better was due to making comparable results. Especially when it comes to regularization. Thus, I decided to do each iteration by creating appropriate ridge,lasso, net as well as its different blend levels. In order to accomplish this, I seperated the data into training and test sets (80/20 split), then I built the model using the training data and tested with the test data. Comparing the RMSE results. This was done for both lambda.min and lambda.1se. Subsequently, OLS models were created as well including forward and backward automatic variable selection for comparisons.

1. All Predictors

Overall, The best model was produced with elasticNet using lambda at 1 SE with an alpha of 0.05. This produced the smalled RMSE error when compared to the test set. The difference between the training(0.2347) and test(0.2862) RMSE was small, nearly (.05), but it does suggest that the model was overfitting to the variables. The Training set has a calculated R2 that accounts for 63.66% of the variability in Cupper Points. However, when compared to the test set the trained model was only able to account for 46.63% of the variability in Cupper Points. This, in addition to the increased RMSE between the training and test sets, suggests that despite Regularization, there is still some overfitting. Since elasticNet was used, it also served as a way for variable selection. It produced a sparser model, driving nearly half the beta weights to zero. That being said, those that are not zero are still REALLY small that could be interpreted as zero.",

Lasso produced the most parsimonious model with only 3 predictors. Just as Net at 0.50, the beta weights included are not trivially zero. Aftertaste, Flavor, and Balance appear to be some of the most important predictors. Ridge does not give very interpretable results because all the variables have beta weights that are trivially zero. ElasticNet at 0.50 also produced a sparse model. Compared to that of alpha 0.05, the beta weights are not trivially zero. 5 distinct predictors standout - Aroma, Flavor, Aftertaste, Acidity, Body, and Balance"

Residual analysis did not yield good results. Very few elements displayed low degrees of heteroscedasticity, such as some of the date variables. There is a very obvious outlier that is apparant in all the charts with a standardized residual of -8. This is a very drastic deviation and may be strongly influencing the effects of the model. Residuals against the response = "Cupper Points Against Residuals, as expected, has a strong positive correlation with a few observations tailing off to the upper right corner suggesting some large overestimations. This may be an indication of some variables that are giving the model too much weight. Residuals against the predictors as expressed before, nearly all the elements were heteroscedastic, some of the predictors displayed positive correlations with the standardized residuals. This suggests that some of the predictions may be more or less accurate than one another. Making it very difficult to interpret relationships.

Linear models overall all three linear models produced terrible, overfit results. This is mainly due to the many levels in the categorical variables that can make the interpretations very difficult. The linear model with all the variables expectedly produced the worst model. Due to the many factor levels it is very difficult to differentiate between important and unimportant predictors". The backwards model produced a majority of the betas to be statistically significant, however, it did so by removing some of the categorical country and regions. It has an F statistic less than alpha suggesting that at least one of the betas is not equal to zero. The adjusted R^2 is 74.65%. Respectable, but it does not help interpreting the relationships. All the residuals appear to have a multiplicative process, that is, as actual values increase, so do the residuals. Additionally, all the plots display a high degree of heteroscadisticity. There may be a need to transform the variables or standardize the variables in order to correct the residual patterns. The QQplot suggests that the residuals are not normal. The majority of the points lie above and below the QQ line. The deviations are most apparant at the tails. The histogram looks normal but with extreme tails on both ends. It depicts outliers exceeeding 5 SDs. Residuals against the response appear to have a curved relationship. Suggesting the need to transform the response variable. All the residuals have a high degree of homoscedasticity with a multiplicative residual pattern.

Next Steps

The next steps in this analysis would be to take the variables that were deemed important from the regularization models and create another iteration using those variables. Subsequently, I would then build multiple models using the predictors found important from the PCA analysis. Finally, I will attempt to use the PCA scores with the regularization models built.