Todd Moore

Class: CS 539
Instructor: Dr. Ricky Sethi

Classification, Prediction and Analysis of Tungsten-based Alloy Densities from Trace Element Properties, Powder Characteristics, and Processing Parameters Data Using EDA, Random Forest, SVM and Decision Tree Models

Full Analysis

EDA Analysis

During EDA it became evident that the data was at best semi-chaotic in the results I obtained. All of the variables chosen for this project seemed to balance out equally during visual evaluation within the box plots, distribution plots, histograms and scatter plots. The number of outliers seemed to be overwhelming as well, potentially pointing to errors in the dataset itself.

Some key features to the dataset stood out however, some being C, K, Zn and BulkDensity, and FSSS as these showed some form of slight regression, either positive or negative. These are expected as the alloy is a K-based alloy that relies on doping of other elements to alloy it. We know already some of the embrittling agents, such as C, within the material that stops grain boundaries from expanding and thus lowering sinter density, so the observations here make sense which is a good validation of the work itself.

Correlation Plot

The correlation plot, as shown below in Figure 1, showed that some elements correlated with one another which may represent a problem in our analysis equipment, which uses refracted waves of light in order to identify elements, and if the wavelength chosen to observe is slightly off of the element it may point to another element that may not actually exist. Notedly the Ni and Cr and Ni and Fe are both correlated moderately as well. The Ni and Cr doesn't make much sense but we traditionally use Ni and Fe together so it makes sense they both would show up. Also, Ca and Zr and Ca and Co correlate which is both surprising as there should be no reason for these too. However, we observe such limited amount of these in our materials anyway its perceived not to be a problem

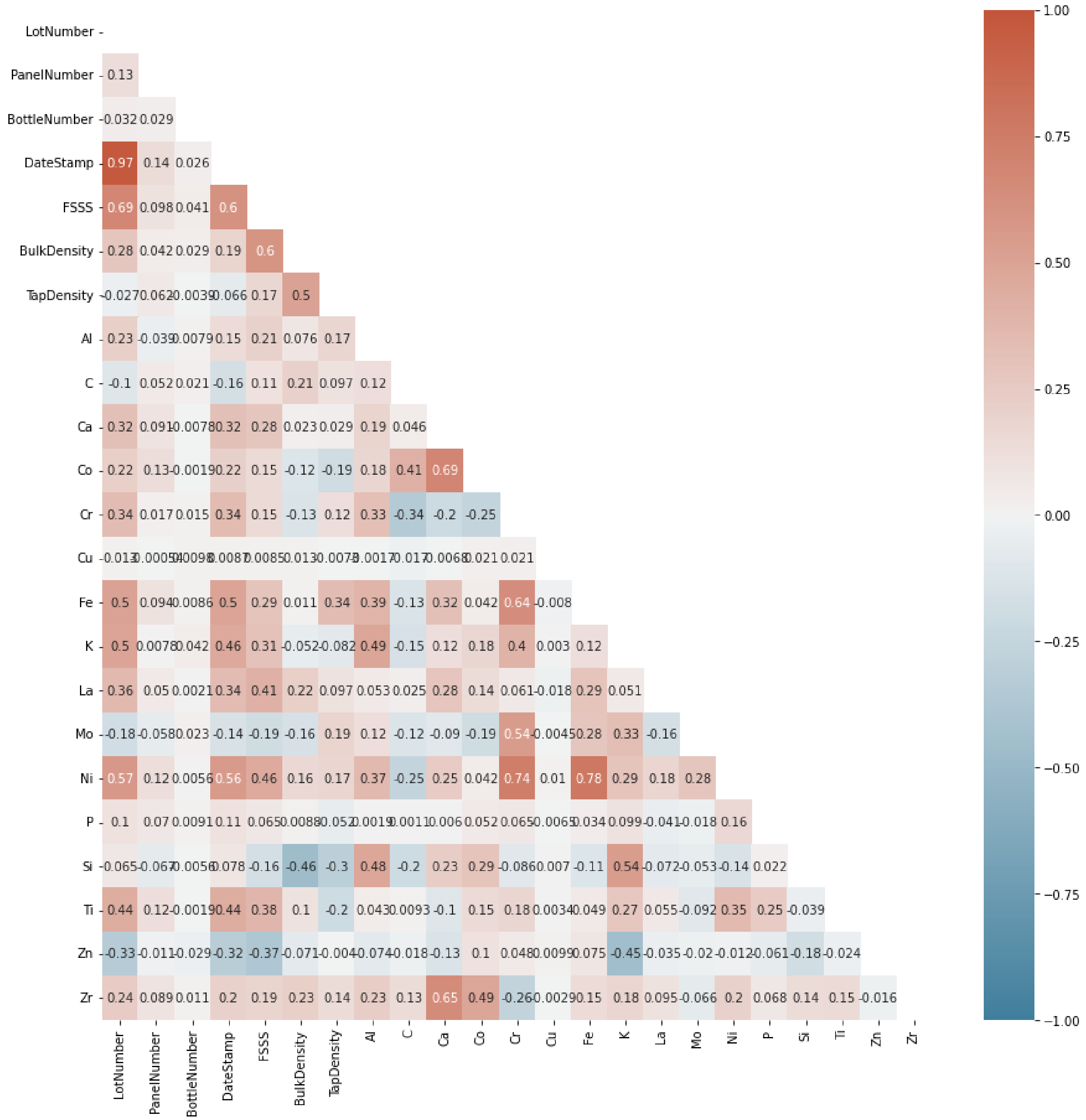


Figure 1: Correlation Plot of all Features

Principal Component Analysis

Principal Component Analysis (PCA) showed what features were most important in the dataset as far as variance influence was concerned.The results themselves were confusing as it did not include what I was expecting it too, being FSSS, BulkDensity and TapDensity. Instead it showed that Date, Zn and some other elements were of most influence over the variance. This was unexpected but not surprising, as these factors can in fact greatly change density of the alloys during sintering. The breakdown of these components is shown in Appendix 1.1. This lead to some reductions in the dataset itself which elimited some of the features that did not seem to be of as much importance. It was surprising that the top 2 variables only accounted for ~22% of the variance in the dataset. The plot of the top 2 variables is shown in Figure 2 below.

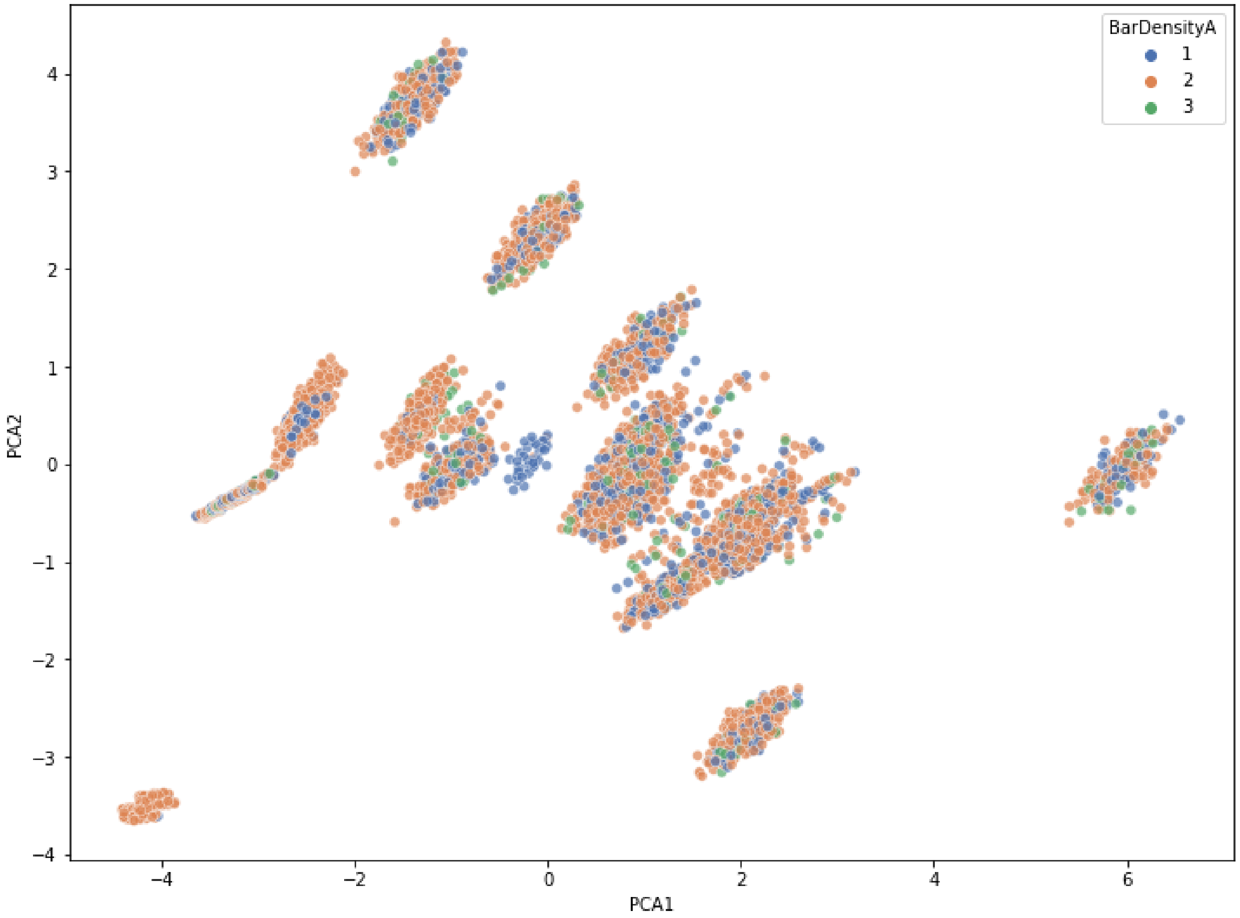


Figure 2: PCA scatterplot of the first 2 components

Box Plots

Multiple box plots were visually plotted for an understanding of the general distribution of the data. The LotNumber column shows that earlier lots in this dataset seemed to show higher densities, possibly pointing to some type of date-related phenomena. Interestingly enough, the DateStamp follows a similar pattern.

The instruments used (PanelNumber and BottleNumber) didn't seem to vary much according to this.

The average partical size (FSSS) shows that there is some potential benefit to smaller partical size < 3.5, however, it is negated with the large group of 1 based densities around the same partical size.

Most of the other chemistry based datapoints have too many outliers to accurately represent their profiles. The figures were redrawn without outliers. These are seen in Figure 3 below.

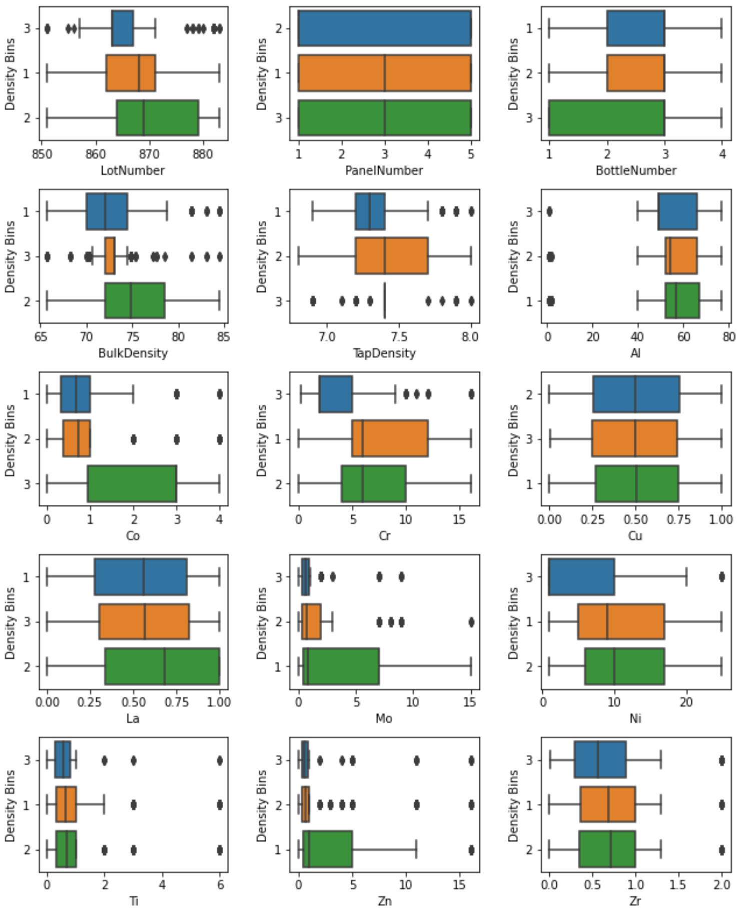


Figure 3: Box plot spread of features compared to target.

Scatter Plots

Scatter plots were used to represent the general trend of the data. For the most part, it is underwhelming that we see a bunch of low slope regression lines. A slightly interesting observation is that there is a slightly positive slope in the K element, which is good because this is a K based alloy.

There is a negative slope with the Mo vs density, which could show that it lowers density as it gets higher, which is unfortunate as we deal with a lot of Mo powder near our process. This observation goes for Zn as well.

An interesting takeaway is that C actually seems to increase the density as it goes up, and in fact, may have the largest slope out of all of them which is surprising, as this is usually classified as an embrittling agent in this material, the same observation can be made about the FSSS (average particle size) which has somewhat been studied, but not greatly. It is generally assumed that lower particle size provides greater densities, but, this shows otherwise. All plots are shown in Figure 4 below.

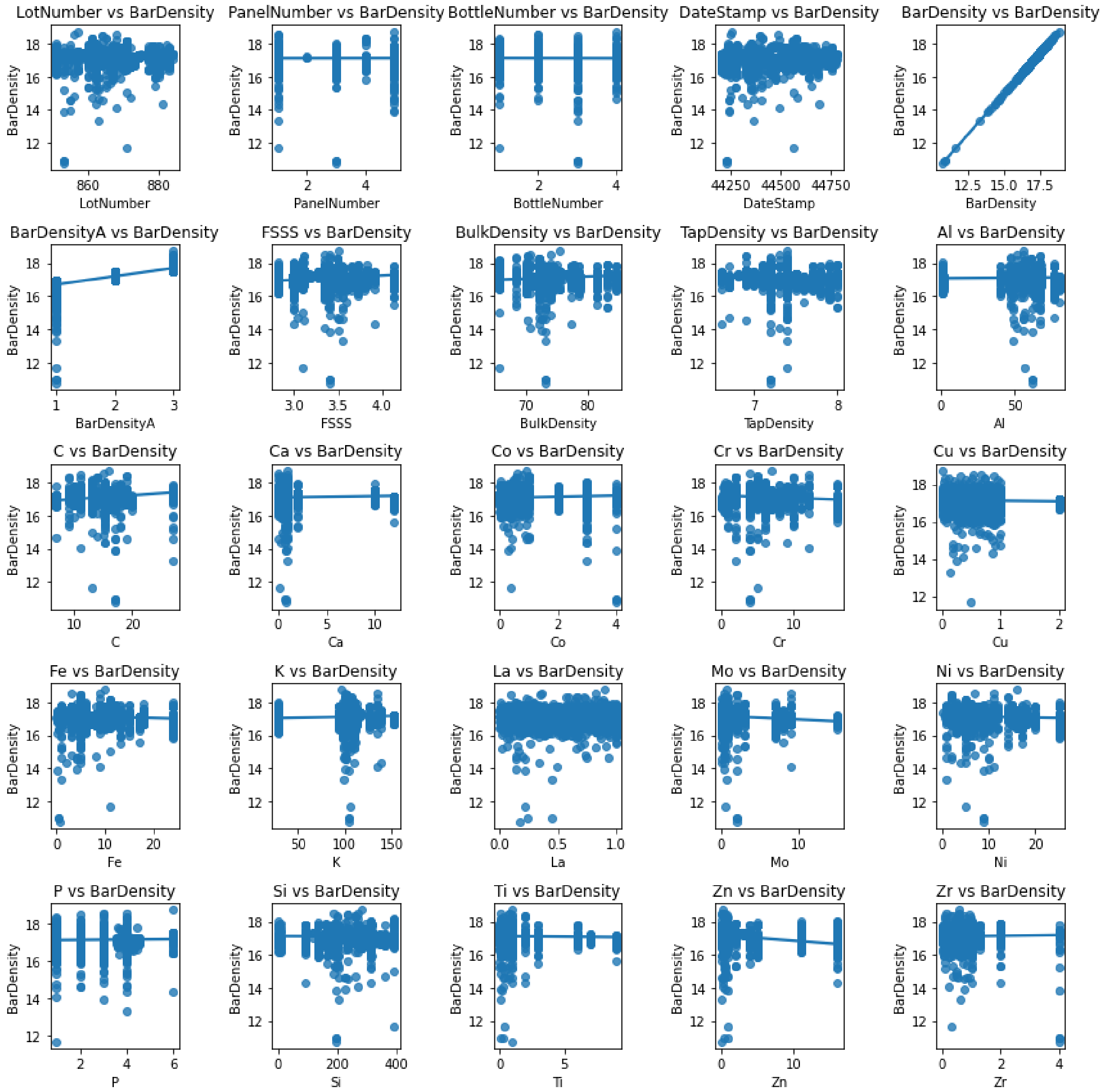


Figure 4: Scatter plot array of features vs target with regression lines

Next, I tried to expand and plot some columns of interest to get more an insight into the data provided here. The new plots are much more interesting. We can see how truely random this dataset seems. Due to the skew of the collected data points count per bin of density it is tough to tell what is true. However, it looks like we can see some patterns regardless.

For FSSS it seems that majority of the 3 bin show up at 3.5 or less, although the regression line is ever slightly positive, the same exact pattern can be seen with Bulk Density and C.

Lower Mo and Zn seems to show the opposite of this relationship. The lower the values is where the data shows higher density. K shows a neutral relationship, which is somewhat expected, up to a certain point, and shows that 100+ in the observed range possibly more beneficial. The upclose plots for this can be seen in Appendix 1.2.

Strip Plots

Strip plots are another unique way to visualize densities of data for specific columns. It's similar to a scatter plot except it bins and visually displays the spread of each X variable. Below is a strip plot of an interested column, Bulk Density, where it shows some data point density of higher ingot density near slightly lower values of Bulk Density. This is very insightful for process development.

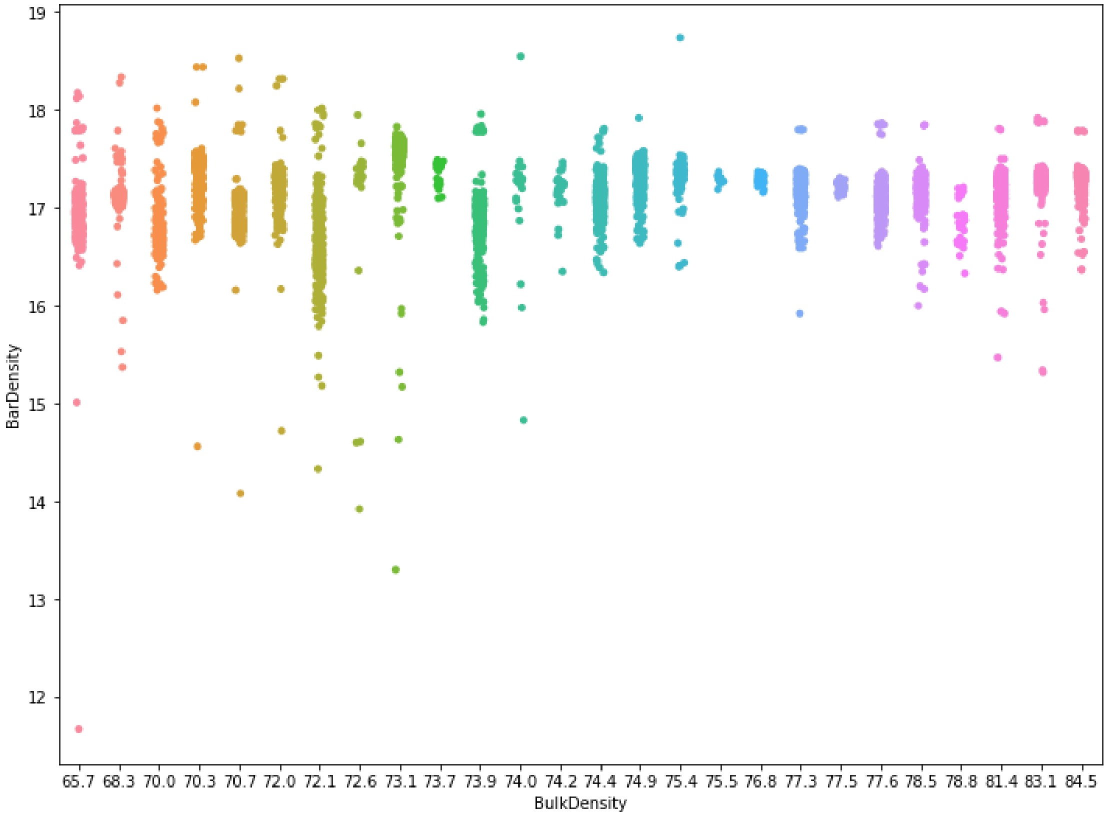


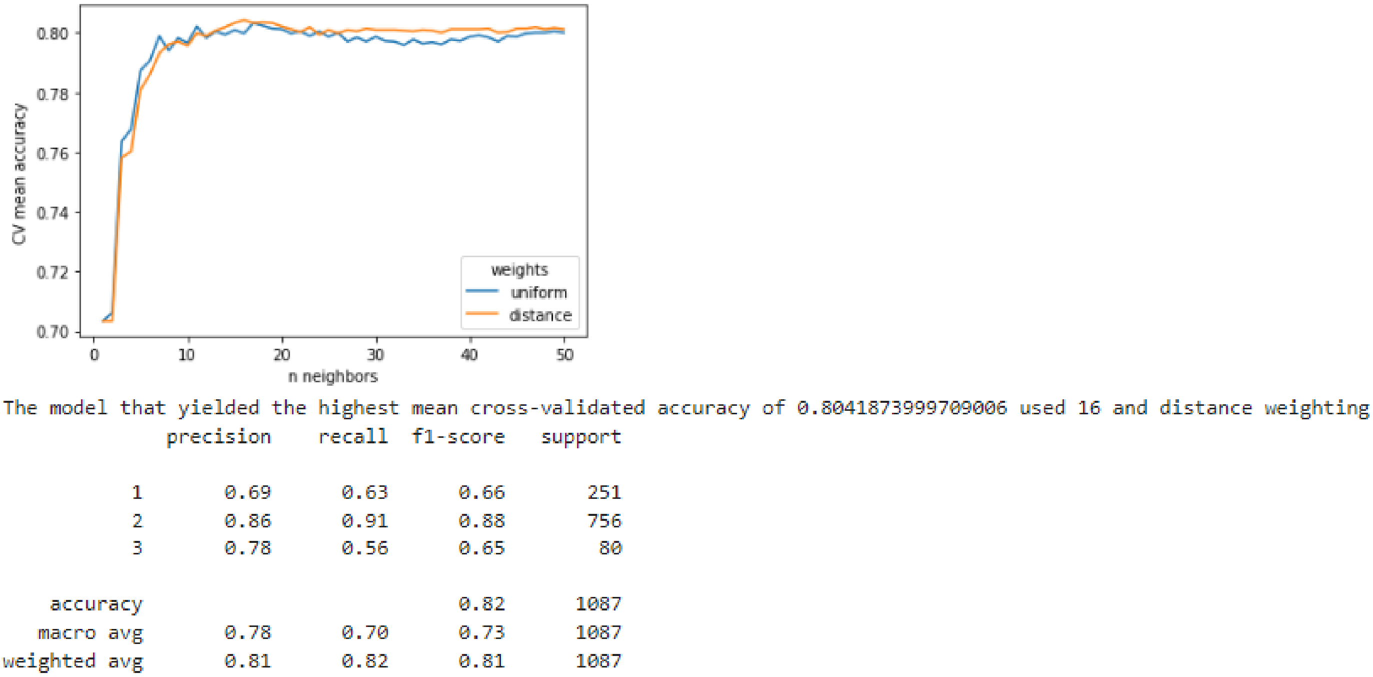
Figure 5: Strip plot of interested column, Bulk Density, vs target Bar Density.

Models Analysis

The models chosen for the analysis were KNN, SVM, Random Forest, and Decision tree. They were chosen primarily to combat the randomness of the dataset and the spread of the data, as well as the limited categorical data, i.e. bias toward the class 2 and 1 densities with limited data for class 3.

KNN

KNN proved to be a very useful model that reached a model accuracy of just over 80% with little mistakes in the confusion matrix. The model was hypertuned in order to accomodate for the spread of the data and looked for nearest neighbors up to 50 away. This is shown in Figure 6 below. The best fit resulted in 16 neighbors away which was surpising and good to hear.



```
The model that yielded the highest mean cross-validated accuracy of 0.8041873999709006 used 16 and distance weighting
              precision    recall  f1-score   support

           1       0.69      0.63      0.66       251
           2       0.86      0.91      0.88       756
           3       0.78      0.56      0.65        80

    accuracy                           0.82      1087
   macro avg       0.78      0.70      0.73      1087
weighted avg       0.81      0.82      0.81      1087
```

Figure 6: Gridsearch analysis for best-fit K value in KNN Modeling

The confusion matrix for KNN provided some insight into the limitation of the dataset, namely the bias toward the lower class of the target, classes 1 and 2, having many more data points than class 3. Thai lead to confusion of over 50% of class 3 density compared to nearly 30% or less in the other 2 classes. This could also be due to random data.

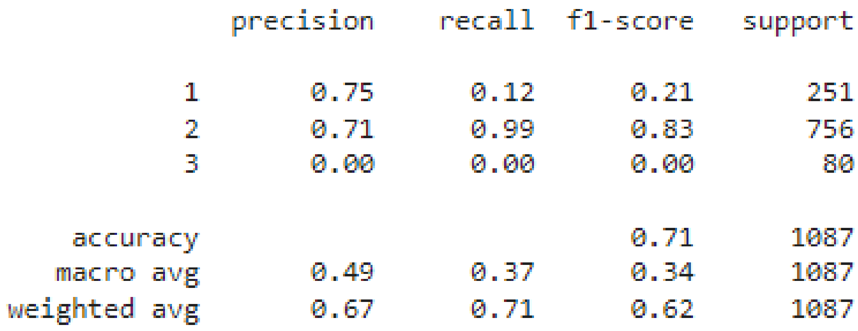
Table 1: Confusion Matrix of KNN Model

	1	2	3
1	157	88	6
2	60	689	7
3	10	25	45

SVM

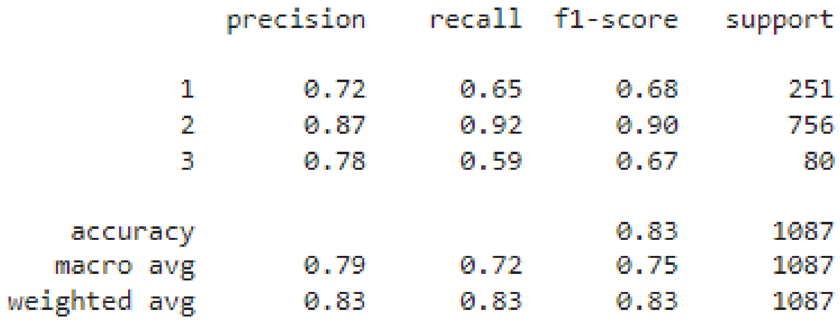
SVM was chosen primarily because it was able to form different shapes for analysis as the shape of the trend of the data was uncertain. Four different kernels were tested being rbf, poly, sigmoid, and linear. The default settings of the models all had similar results showing that there is no immediate trend to the data itself. After optimization of the parameters tested over the 4 different models it was shown that rbf with a C value of 1, and a gamma of 0.1 had the best accuracy of just under 83%. The classification report showed that there was some confusion in the model predicting primarily toward the most common feature the class 2 density, which is somewhat expected with the distribution of the data being the way it is. This was corrected using optimization techniques used in SVM, the results of that change is shown below in Tables 2 and 3.

Table 2: Classification report of SVM Model before optimization.



```
              precision    recall  f1-score   support

           1       0.75      0.12      0.21       251
           2       0.71      0.99      0.83       756
           3       0.00      0.00      0.00        80

    accuracy                           0.71      1087
   macro avg       0.49      0.37      0.34      1087
weighted avg       0.67      0.71      0.62      1087
```

Table 3: Classification report of SVM Model after optimization



```
              precision    recall  f1-score   support

           1       0.72      0.65      0.68       251
           2       0.87      0.92      0.90       756
           3       0.78      0.59      0.67        80

    accuracy                           0.83      1087
   macro avg       0.79      0.72      0.75      1087
weighted avg       0.83      0.83      0.83      1087
```

There is a major improvement over the original classification report. The scores alone are not the greatest, however, they aren’t terrible either, with the recal of the class 3 density scores being the worst, but again, this is expected due to the major bias in data density. Overall, for what the model was handed, the results aren’t bad.

Random Forest and Decision Tree

Random Forest and Decision Tree were both chosen due to their statistical prowess allowing for more randomization which somewhat fit the dataset the best.The base model had ab accuracy just above 84% which is the best out of the previous models. The feature scores of the model showed that DateStamp and Zn and Ni were of most useful in the model. A new model was tested by removing the last 5 of the feature scores, but this did not result in a significant increase in accuracy.

A useful feature of the Random Forest is that it can produce features that had the greatest weight in the decisions made in the model. Below is a table of these results. Itw as surprising to see that the date in which the data was collected had a major role. This is useful information because it was always speculated that the change in humidity and temperature, an inherent obstacle of Maine’s weather, was an important factor in material quality.

Also surprisingly, is the elements that are repeatedly topping the analysis charts, including Ni and Zn. Lastly, it is interesting to see that TapDensity has very little influence on the density predicted according to this model, which may mean the range of our TapDensity is good and doesn't differ enough to create variance.

Table 4: Feature Scores used in the Random Tree Model

Feature:	Weight:
DateStamp	0.183327
Zn	0.109681
Ni	0.082128
Cu	0.076788
La	0.068842
Co	0.068332
Ti	0.066008
BulkDensity	0.064476
Mo	0.061358
Ca	0.049551
PanelNumber	0.038351
P	0.032236
LotNumber	0.026099
K	0.023302
Cr	0.022559
Al	0.015453
TapDensity	0.011510

Results Summary and Future Prospects

Result Summary and Future Prospect

Overall, I am happy with where I've started with this project and ended with this project. Primarily the knowledge I've gained in both ML and the dataset itself. I believe I have successfully answered my question and have a few models that will provide a bases for physical testing in the future.

The major take-aways of this project so far has been that there are some chemistry links that I need to look into. The general trends between Zn and Cr and other trace elements seem too coincidental, and are likely standard errors in the testing. Reforming this could help give better and more accurate chemistry results. Furthermore, it was evident that Bulk Density and Tap Density were not as important in this project as I was suspecting they would be, although they did provide some minor regression-type behavior, which means perhaps our current standard is good for the application.

The Datestamp feature came up as an important feature which was surprising but not unexpected. In Maine the weather hits some major extremes, dry in the winter, humid in the summer. It is known that these types of swings may produce an impact on the properties of our materials but it was never speculated how big of an impact it has. This will be explored further as this could have major upsides to production during the change in the weather.

My next steps would be to implement attribute testing in the models in order to form theoretical predictions of parameters entered. I also plan to expand on the project and encorperate other variables such as pressing pressure, ingot geometry, mold type, and other processing specifications in the hopes that I can bolster the models and find some certainty in the influence of density. Further beyond that I hope to expand into the territory of the paper mentioned above and target mechanical properties of our materials, including the several other alloys that we make. The alloy chosen is a difficult one and has very little information on it, so a paper on this would likely be successful.

Some questions that arose during the project are:

1. If I were to clear out outliers and low-weighted features (based on model representation) would my EDA Analysis show different results?
2. If I were to observe my data more, on a statistical level, could I re-engineer some of the data to be less, random, and more normalized.

Appendix

Appendix 1: Additional Tables

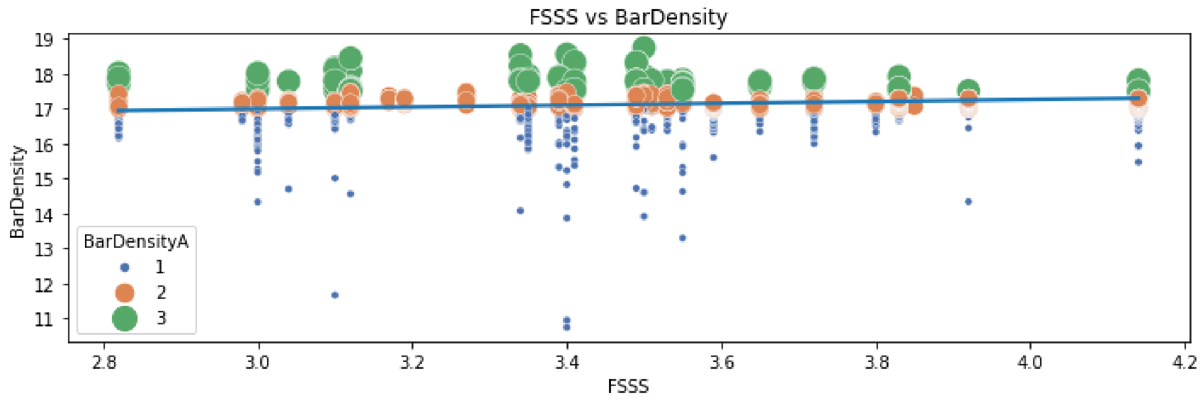
Appendix Table 1.1: PCA Breakdown

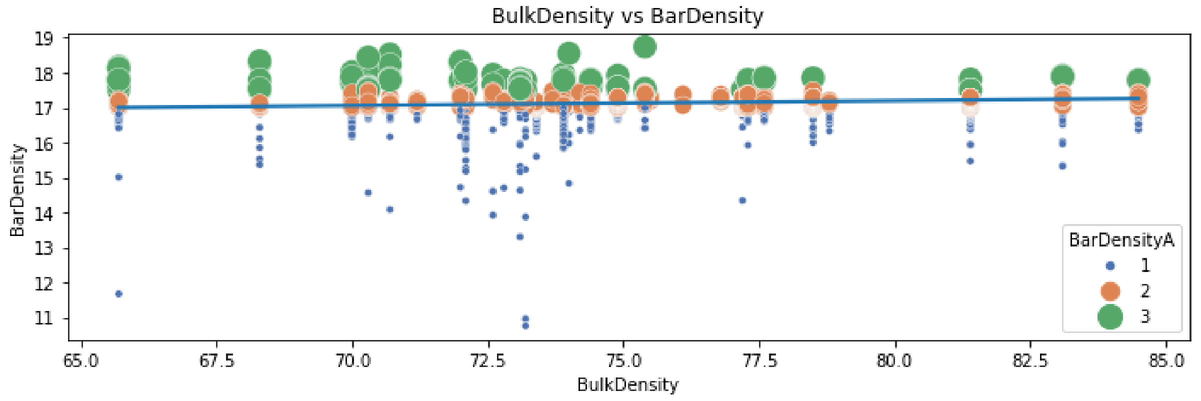
```
   LotNumber  PanelNumber  BottleNumber  DateStamp      FSSS  BulkDensity  \
0   -0.397612    -0.074879     -0.016696  -0.381717 -0.326004    -0.128134
1   -0.055555    -0.059524      0.000965  -0.037382 -0.111802    -0.091070
2    0.028695     0.076390      0.007169   0.000328  0.242111     0.475095
3   -0.221165    -0.007777     -0.021544  -0.237382 -0.171674    -0.010283
4    0.038859     0.199729     -0.070646   0.100296 -0.179672    -0.307125
5   -0.122537     0.200059      0.081300  -0.149844 -0.015044     0.106182
6    0.009604    -0.475039     -0.653427  -0.034624  0.066829     0.078383
7    0.022913    -0.115573      0.342806  -0.002149  0.040070     0.023439
8    0.022378    -0.060185      0.531472  -0.002895  0.047785    -0.082104
9    0.013155     0.702796     -0.385641   0.016810  0.011659    -0.134330
10   0.010682    -0.364349     -0.092273   0.047196  0.046485    -0.145866
11  -0.048861    -0.066829     -0.016380  -0.048379  0.007098    -0.170213
12  -0.168490     0.106641     -0.015093  -0.175025  0.014137     0.125966
13   0.254723    -0.109770      0.012329   0.337527 -0.422495    -0.185129
14   0.179667     0.014436     -0.031771   0.182094  0.079157     0.327758
15   0.202404     0.074531     -0.036889   0.170843 -0.306105     0.169363
16   0.064784     0.038620     -0.004417   0.189945  0.064326     0.213438
17  -0.002447    -0.009568      0.018681   0.135132 -0.642354     0.319390
18  -0.054557    -0.018195      0.002380   0.062055 -0.080403     0.185381
19  -0.046112    -0.009974     -0.003285  -0.257311 -0.106280     0.118413
20   0.041648     0.001173     -0.003307   0.074632  0.124199     0.006473
21  -0.232347    -0.006941      0.002750   0.469182  0.145658    -0.364723
22  -0.732951    -0.003153     -0.002069   0.448291  0.009774     0.216498

    TapDensity        Al         C        Ca  ...        La        Mo  \
0    -0.064943 -0.208095  0.042450 -0.203790  ... -0.182814 -0.029906
1     0.115579  0.029266 -0.277621 -0.336594  ... -0.103880  0.363079
2     0.336691 -0.214684  0.164799 -0.091250  ...  0.173716 -0.133874
3     0.356744  0.285152  0.143206  0.324897  ...  0.013316  0.268004
4    -0.309628 -0.289528 -0.145923  0.115461  ...  0.068760 -0.133538
5     0.045799  0.148901  0.385363 -0.205461  ... -0.394482  0.194401
6    -0.068820  0.301100  0.185421 -0.182514  ...  0.087240 -0.172618
7    -0.077670  0.127688  0.127026 -0.101200  ...  0.040167 -0.067503
8    -0.150624  0.192163  0.345756 -0.165604  ...  0.207079 -0.062077
9     0.008854  0.044986  0.289655 -0.101065  ...  0.064624  0.051333
10   -0.203492 -0.398906  0.308910  0.159582  ... -0.021569  0.481121
```

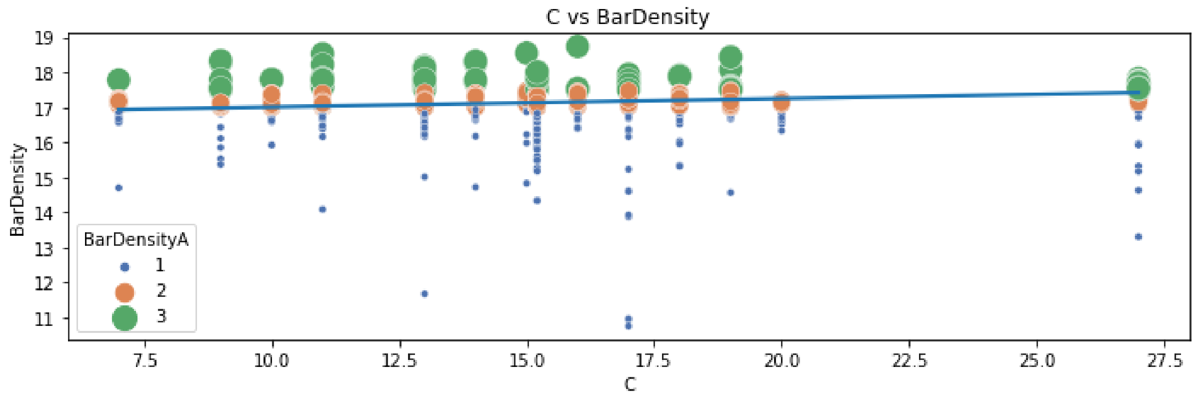
```
11    0.021974  0.040760  0.243708  0.047460  ...  0.417493 -0.040955
12    0.012311 -0.109354 -0.198625 -0.050393  ...  0.672048  0.334196
13    0.531050 -0.224547  0.143426 -0.225032  ...  0.051224 -0.019333
14    0.006902 -0.165231 -0.093555  0.024668  ... -0.217227  0.088620
15   -0.473721  0.127938  0.220151 -0.201765  ...  0.100689  0.065996
16   -0.149372  0.106820  0.003244  0.184499  ... -0.075264  0.481023
17   -0.008529  0.266089 -0.118757  0.196785  ...  0.099553 -0.010805
18    0.030274 -0.351915  0.239828 -0.206180  ...  0.059555  0.061240
19   -0.111797 -0.094153 -0.131037 -0.258873  ...  0.005735 -0.090175
20   -0.024508  0.170900 -0.297269 -0.531626  ...  0.044687  0.210935
21    0.123514  0.235579 -0.017464 -0.085512  ...  0.030028  0.108468
22   -0.100608 -0.100310  0.012590 -0.010444  ... -0.019728 -0.154918

         Ni         P        Si        Ti        Zn        Zr
explained_var  \
0  -0.351254 -0.070696 -0.045142 -0.196494  0.132722 -0.171246
0.216986
1   0.236642 -0.004621 -0.132029 -0.025142  0.066246 -0.318994
0.123929
2   0.046675 -0.034968 -0.540891  0.007211  0.087488 -0.022309
0.106980
3   0.122750 -0.143294  0.013157 -0.343170  0.267049  0.290233
0.082950
4   0.195258  0.137858 -0.160024  0.214128  0.493851  0.017789
0.066235
5   0.076012  0.474270 -0.096347  0.415879  0.100621  0.143271
0.055511
6   0.013650 -0.080454  0.064884  0.189407  0.266228 -0.102927
0.044541
7  -0.012152 -0.213920  0.020216  0.040765  0.164923 -0.097970
0.043951
8  -0.048066 -0.259696  0.006334  0.057603  0.175040 -0.251128
0.043006
9  -0.081255 -0.332498  0.001362 -0.055813 -0.137210 -0.260303
0.040167
10  0.012451 -0.172109 -0.344773  0.032889 -0.263923  0.003975
0.036827
11 -0.064903  0.665987  0.031586 -0.287511 -0.189821 -0.328276
0.035400
12 -0.116387  0.014750  0.066726  0.233399  0.228079  0.149378
0.028606
13 -0.313314 -0.002539  0.097869  0.158346  0.074235  0.096174
0.018629
14 -0.139918  0.129675  0.011279 -0.436822  0.419390 -0.290547
0.017824
15 -0.055851  0.058244 -0.146916 -0.343826  0.069022  0.429834
0.014578
```

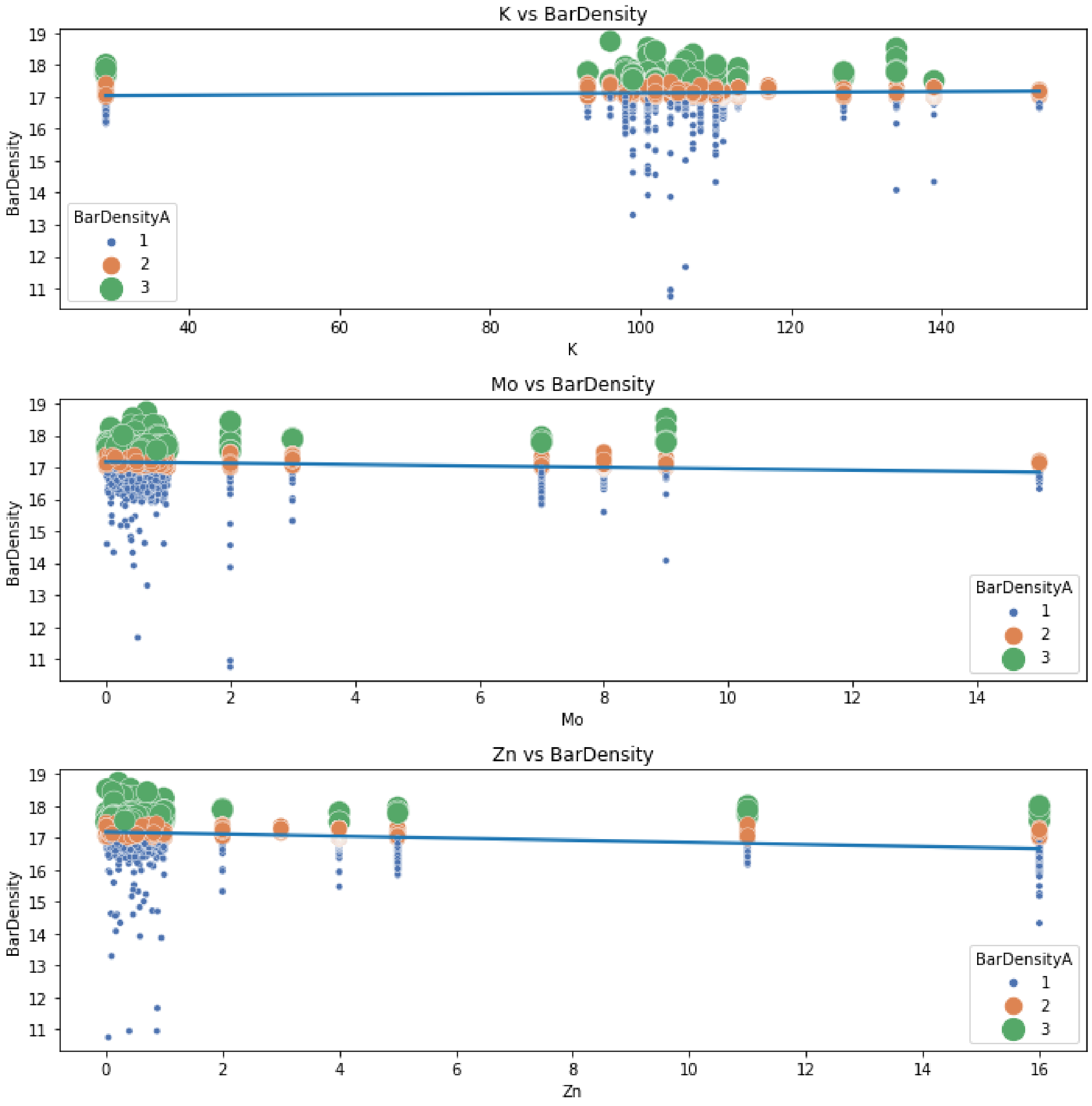
```
16 -0.239865  0.028376  0.260697  0.200654  0.117910 -0.265151
0.008050
17  0.181079 -0.029095 -0.197323  0.149164 -0.239427 -0.275139
0.006319
18  0.591104 -0.048216  0.548107 -0.110947 -0.032467  0.005970
0.004064
19  0.195833 -0.020941 -0.140972 -0.015131 -0.047191 -0.124210
0.002203
20 -0.135095  0.024428 -0.004070 -0.112476 -0.219618  0.189491
0.001934
21  0.291496  0.016972 -0.248013 -0.135922  0.158656 -0.044109
0.000791
22 -0.173633 -0.012835  0.062572  0.040811 -0.075416  0.066972
0.000517

    explained_var_cumsum
0               0.216986
1               0.340915
2               0.447895
3               0.530845
4               0.597079
5               0.652591
6               0.697132
7               0.741083
8               0.784088
9               0.824256
10              0.861083
11              0.896483
12              0.925089
13              0.943718
14              0.961542
15              0.976120
16              0.984171
17              0.990490
18              0.994554
19              0.996758
20              0.998692
21              0.999483
22              1.000000
```

Appendix 2: Additional Figures









Appendix Figure 2.1: Scatter plot of interested columns with density hue and regression line.