**CKME 136**

**Capstone Project – Andrew Howard**

**Fall 2015**

**Youthful Cities**

**https://github.com/ajhoward/Capstone**

**Introduction:** Youthful Cities has created a list for 2014 and 2015 of cities all around the world ranked by youth friendliness. This ranking has never had any formal data analysis completed so the data is largely unexplored. The research problem I proposed was to use feature selection to determine the number of features needed to efficiently predict the final rankings. I will be using R to complete stepwise linear regression and determine the Root Mean Squared Error to evaluate the different feature selection models.

**Literature Review:** As this data set has never had any formal data analysis completed there were no papers to reference. The majority of the literature I used to assist me in my analysis was the lecture slides and lab exercise from CKMT 105, Data Analytics Advanced Methods, Week 5 Multivariate Linear Regression.

**Dataset:** This data was started in 2014 by Youthful Cities as a way for young people to determine how youth friendly a city is by comparison to other cities all around the world. The data is also useful for city governments and management as a tool to see how their services compare to other cities. The data includes data from government services and various sources on the internet to surveys of actual young persons in various cities. The index of the final ranking is what I will be using for the exploration of the data. This part of the data set is where the cities are given values for 21 different features. The 21 features are determined by normalizing and combining the raw data of 100 features down to the 21 used for the final ranking. The totals are then combined to come up with a final ranking of the 55 cities.

**Approach:**

## Step 1: Data Prep

The data was provided to each of us students as a multi sheet excel document. The sheet I used was called Attribute Scores. This was the sheet that included all of the 21 features after being normalized and totaled for each feature set. It also included the total of all the features together which was used as the value for the final ranking.

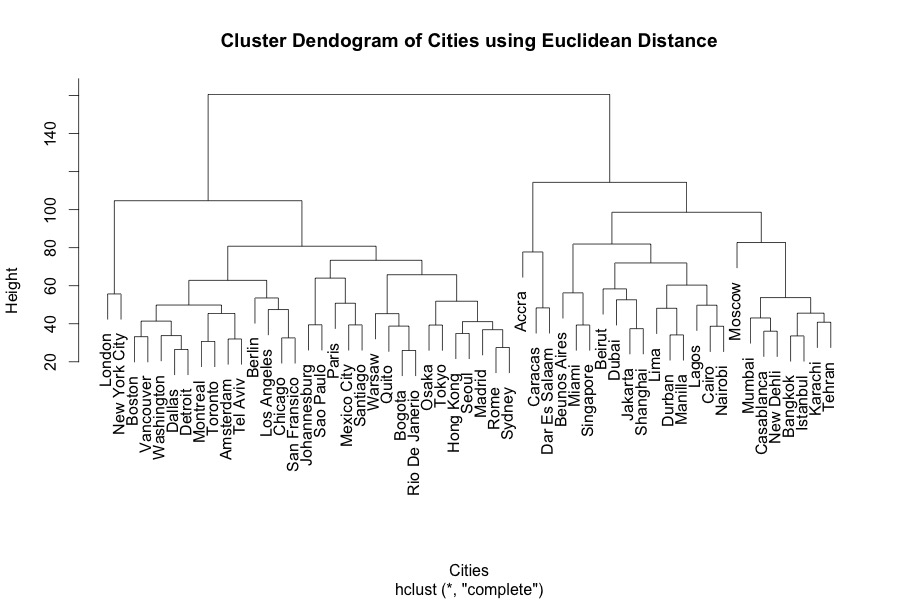
I also needed to edit the Master data sheet as a couple of the features were missing values since the Hyperlink in the Excel sheet was not working correctly. We were given the fix by Youthful Cities so that the data was still accurate.

## Step 2.: Baseline RMSE

Once the data was loaded into R the first analysis I completed was to run a linear regression model on all 21 features and I calculated the baseline Root Mean Squared Error to be included with the visualizations I created.

## Step 3: Hierarchical Clustering

The next part of my analysis I created 2 Cluster Dendograms to visualize the Features and Cities. I started with creating a distance matrix using Euclidean distance. I then completed the clustering and plotted the clustering data in a Dendogram. I also added rectangles around the Features Dendogram at the 7 different cluster cut off.



## 

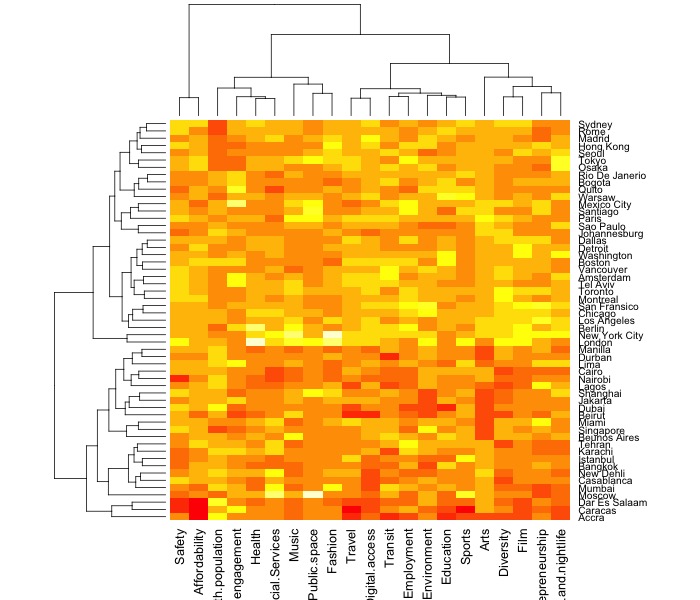
## Conclusions: I found that the Cluster Dendogram for the cities organized the cities into the left side being more of the North American cities and right being more of the Asian and European cities. This does make sense that many of the cities being geographically close would yield similar results for the different features.

## Step 4: Heat Map

With having created the Hierarchical Clustering I also decided to create a heatmap using the clusters created with the Feature Dendogram as a way to organize the Heat Map.

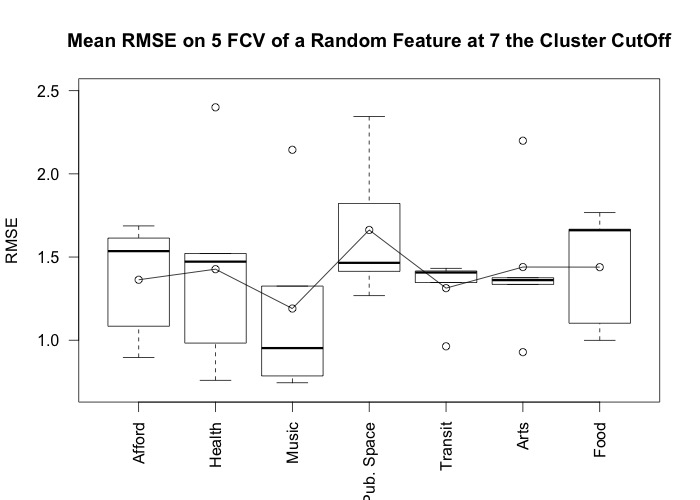
## Conclusions: I would conclude that the data is actually quite similar and that the differences between the cities and the difference between the features isn’t all that high.

There are some differences in the colour, however the difference is not all that great. As you will see later in the report the root mean squared error only generally fluctuates between 1.5 and 0.5 which is not a large variance.



## Step 5: 7 Feature Cluster Linear Model

For the next part I used the 7 Clusters from the Feature Dendogram to select a random feature from each of the clusters to get a generalized look at the 5 Fold Cross Validation of the RMSE’s for the 7 different clusters. The 5 folds were created by taking sets of 11 in a row using the original alphabetical order.



## Conclusions: I would conclude that the 7 clusters are fairly varied, considering that the data doesn’t have a large variance between the RMSE’s. One quality I do notice is that although the boxplots are quite varied when you take into account the outliers there actually isn’t a large amount of variance between the clusters.

## Step 6: 3 Stepwise Linear Regression Models using Different Starting Points

I started each of the 3 models by completing the linear regression using forward stepwise regression. I started the first model with Safety, the second with Music and the third with Education. The feature name shown on the visualizations of the 3 models is the feature added at that point of the model. I calculated the root mean squared error for all the feature selection models and put them together. I then separated the models into the 5 folds and used the same 5 folds as the previous model just selecting sets of 11 in a row using the alphabetical order. On all of these models I also included a line of the baseline RMSE to give a comparison for the models.

## Conclusions: Across the 3 experiments the general direction of the plots is in a downward trajectory. This supports the idea that the RMSE would get less as the number of features increases.

## It is interesting that although the baseline RMSE is almost at 0, when only one feature is removed the RMSE jumps to approx. 0.25. This conclusion was found across all three experiments.

## The RMSE also generally stays consistent over the next 5 or so feature removals across the 3 experiments. Depending on what would be considered an acceptable amount of error a reasonable conclusion could be made from a feature set including 15 features.

## Final Conclusions/Future Plans: Across the 3 experiments it is hard to find any concrete solutions for which features would be best for predicting final results. I would like to run further models with the stepwise linear regression starting on each of the different features. I could then better cross check the results and figure out a subset of best features. I would also be interested in further examining the clusters found in the dendogram for better selecting cities for the 5 folds to perform cross validation. I believe this would yield a more accurate report. Some of these issues are also caused by the fact that this dataset is relatively small. Therefore, I would be interested in trying to use the full dataset with 100 features and complete the normalizing manually to see if that would yield more significant results.

## 

