Project 7: Conclusions

Task 1

1. What is the optimal number of store formats? How did you arrive at that number? The optimal number of store formats is 3. I determined this by first summarizing sales data for each store in 2015. Then, I determined the optimal number of clusters assuming a kmeans cluster method through the Calinski-Harabasz, and numerous other, cluster validation methods.

2. How many stores fall into each store format? Per the table below, cluster 1 contains 17 stores, cluster 2 contains 35 stores, and cluster 3 contains 33 stores.

```
## # A tibble: 3 x 2
##   cluster count
##     <int> <int>
## 1       1    17
## 2       2    35
## 3       3    33
```

3. Based on the results of the clustering model, what is one way that the clusters differ from one another? The table below summarizes the differences between the 3 clusters. One way that the clusters differ from each other, is the percentage of general merchandise sales, with Cluster 1 at the highest percentage and Cluster 2 at the lowest percentage.

```
## # A tibble: 85 x 14
## # Groups:   cluster [3]
##    store  year dry_grocery dairy frozen_food  meat produce floral  deli
##  * <fct> <int>       <dbl> <dbl>       <dbl> <dbl>   <dbl>  <dbl> <dbl>
##  1 S0001  2015        46.1 10.3         7.72  10.8    9.72  0.677  4.35
##  2 S0002  2015        45.8 10.6         7.88  11.5   10.1   0.744  3.98
##  3 S0003  2015        42.1 10.2         6.90  11.5   12.5   0.963  4.18
##  4 S0004  2015        45.5  9.71        8.03  12.8   10.0   0.612  4.18
##  5 S0005  2015        44.0 10.6         8.63  10.2   13.1   0.888  3.54
##  6 S0006  2015        46.0  9.86        7.44  10.7   10.2   0.827  4.63
##  7 S0007  2015        44.2 10.2         8.92  11.5   10.3   0.784  4.63
##  8 S0008  2015        42.8 10.9         8.03  10.7   12.4   0.786  3.26
##  9 S0009  2015        44.2 10.0         7.71  11.9   10.4   0.729  4.42
## 10 S0010  2015        43.1 11.5         8.03  10.7   12.2   1.04   3.71
## # ... with 75 more rows, and 5 more variables: bakery <dbl>,
## #   general_merchandise <dbl>, total <dbl>, cluster <int>,
## #   total_sales <dbl>
```

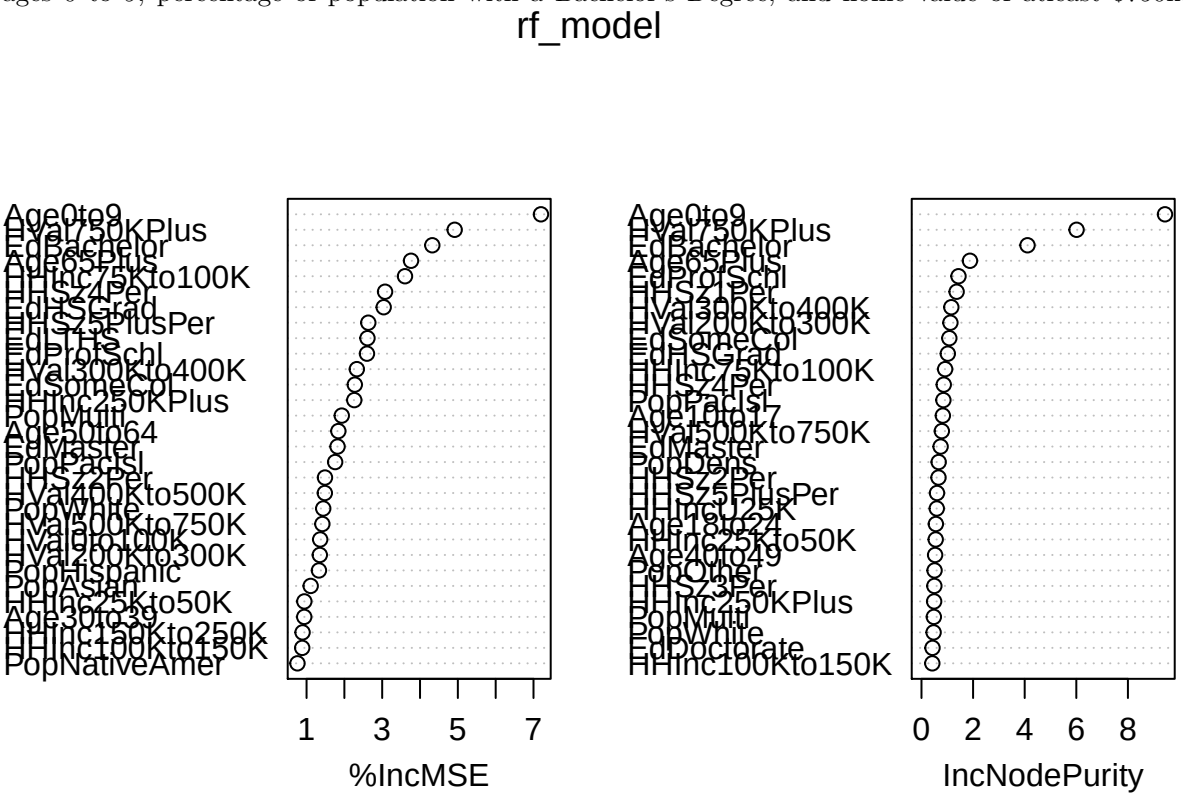
```
## # A tibble: 3 x 10
##   cluster dry_grocery dairy frozen_food  meat produce floral  deli bakery
##     <int>       <dbl> <dbl>       <dbl> <dbl>   <dbl>  <dbl> <dbl>  <dbl>
## 1       1        45.8  9.94        8.04  10.3    10.1  0.740  3.77   2.18
## 2       2        45.4 10.00        7.65  12.0    10.1  0.689  4.41   2.87
## 3       3        43.7 10.6         8.07  10.6    12.3  0.989  3.80   3.04
## # ... with 1 more variable: general_merchandise <dbl>
```

4. Please provide a map created in Tableau that shows the location of the existing stores, uses color to show cluster, and size to show total sales. Make sure to include a legend! Feel free to simply copy and paste the map into the submission template. The below Tableau map displays the store locations,

colored by cluster. https://public.tableau.com/profile/michael.gysel#!/vizhome/StoreClusters_10/Dashboard1?publish=yes

Task 2

1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? I used the Random Forest Model because this had the greatest accuracy when predicting correct clusters of the 85 existing stores.
2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization. Per the variable importance plot of the boosted model below, the three most important factors are percentage of population ages 0 to 9, percentage of population with a Bachelor's Degree, and home value of atleast $750k.



3. What format do each of the 10 new stores fall into? Please provide a data table. The data table below shows the 10 new stores and corresponding clusters.

```
##     X     Type
## 85 85 Existing
## 86 86      New
## 87 87      New
## 88 88      New
## 89 89      New
## 90 90      New
## 91 91      New
## 92 92      New
## 93 93      New
## 94 94      New
## 95 95      New
```

Task 3

1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision? An ETS(M,N,M) Model was used. After assessing the time series decomposition plot, it became clear the Error is Multiplicative, Trend is None, and Seasonality is Multiplicative. This ETS Model yielded a higher accuracy than the ARIMA Model, so it was chosen for the time series forecast.

2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts. The table below shows the produce sales forecasts for existing and new stores for 2016. The link below shows the data visualization of total produce sales. https://public.tableau.com/profile/michael.gysel#!/vizhome/P7-TotalProduceSales/Dashboard1?publish=yes

```
##        Existing.Stores New.Stores
## Jan-16        20705352    2435924
## Feb-16        20110037    2365887
## Mar-16        23667529    2784415
## Apr-16        22453553    2641594
## May-16        25493363    2999219
## Jun-16        25858644    3042193
## Jul-16        26340299    3098859
## Aug-16        23118849    2719865
## Sep-16        20624065    2426361
## Oct-16        20267316    2384390
## Nov-16        21293661    2505137
## Dec-16        21089480    2481115
```