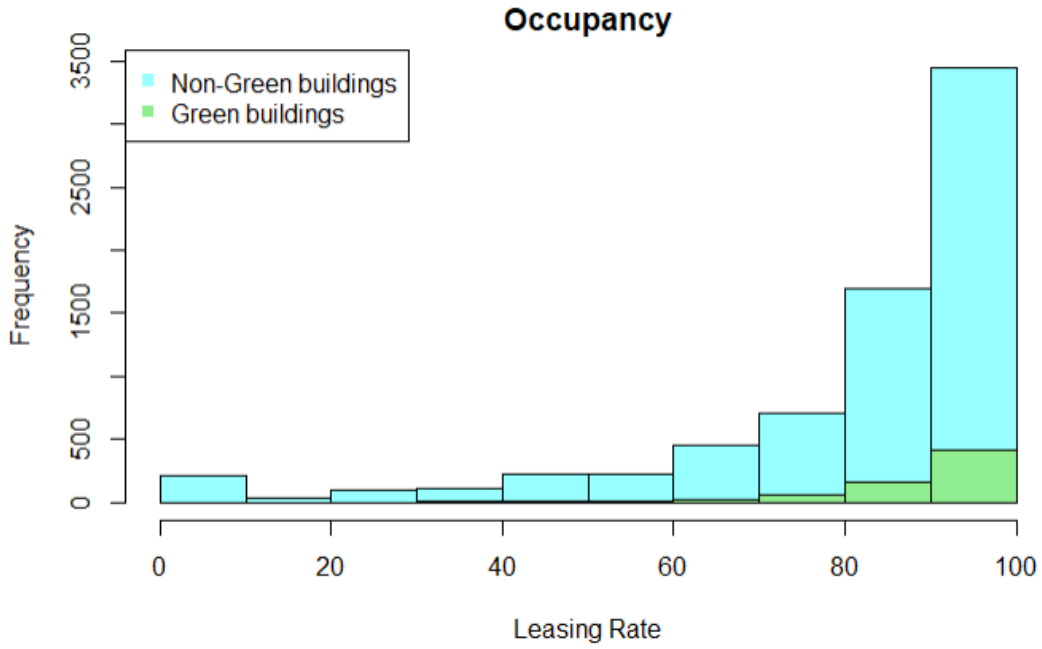
STA 380, Part 2: Exercises

Kevin Cheng

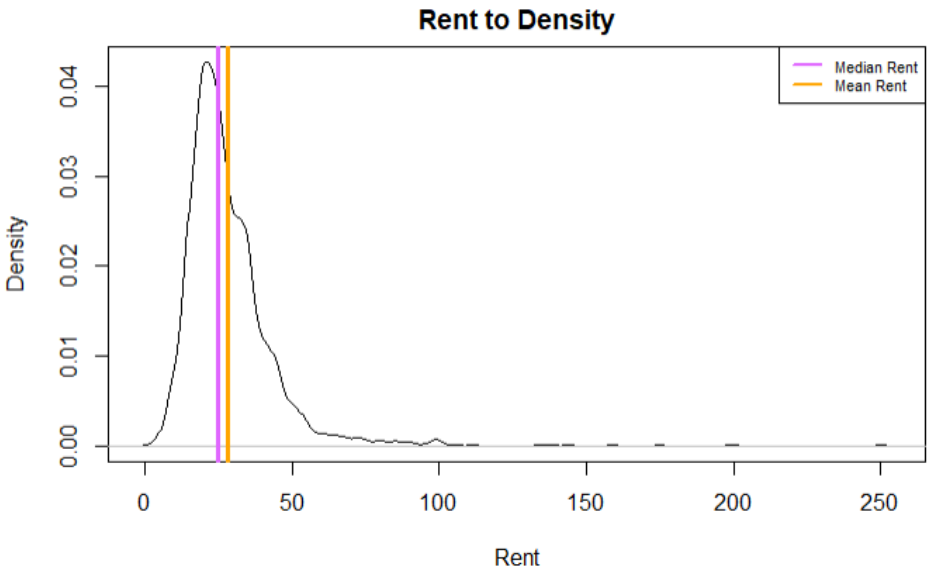
8/17/2020

1. Visual Story Telling Part 1: Green Buildings

First, I look into the distribution of leasing rate of green buildings and non-green buildings. The distribution of non-green buildings' leasing rate has a shoot up in the range below 10%. Therefore, I believe these buildings are "weird" and should be removed from our analysis.



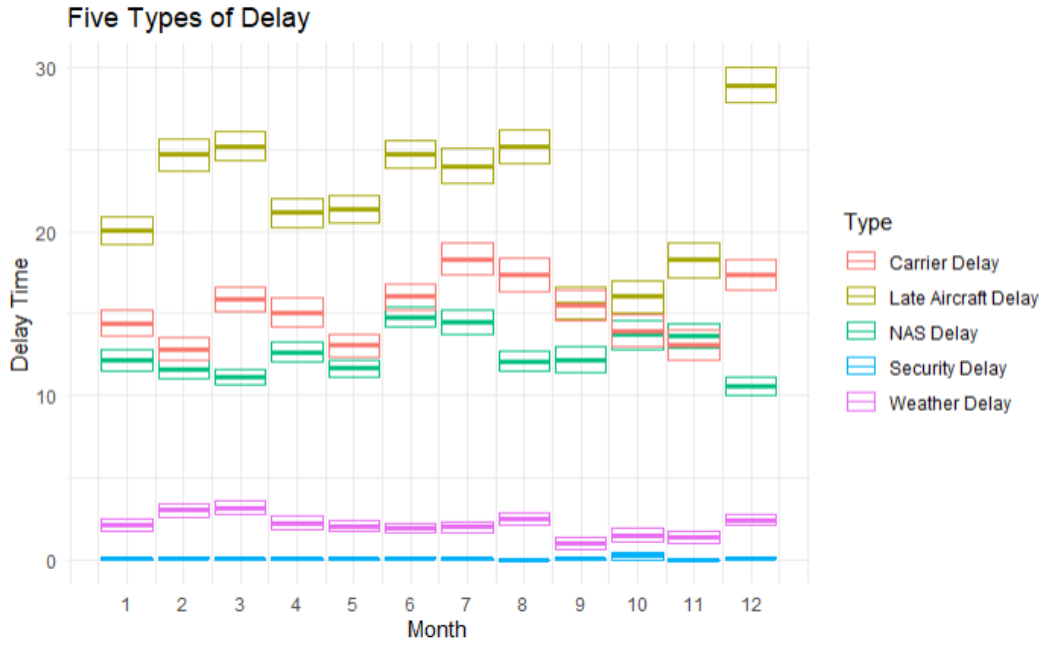
Second, the leasing rate for green buildings is highly left-skewed, so the median is a better estimation for building, which is 92.9%.



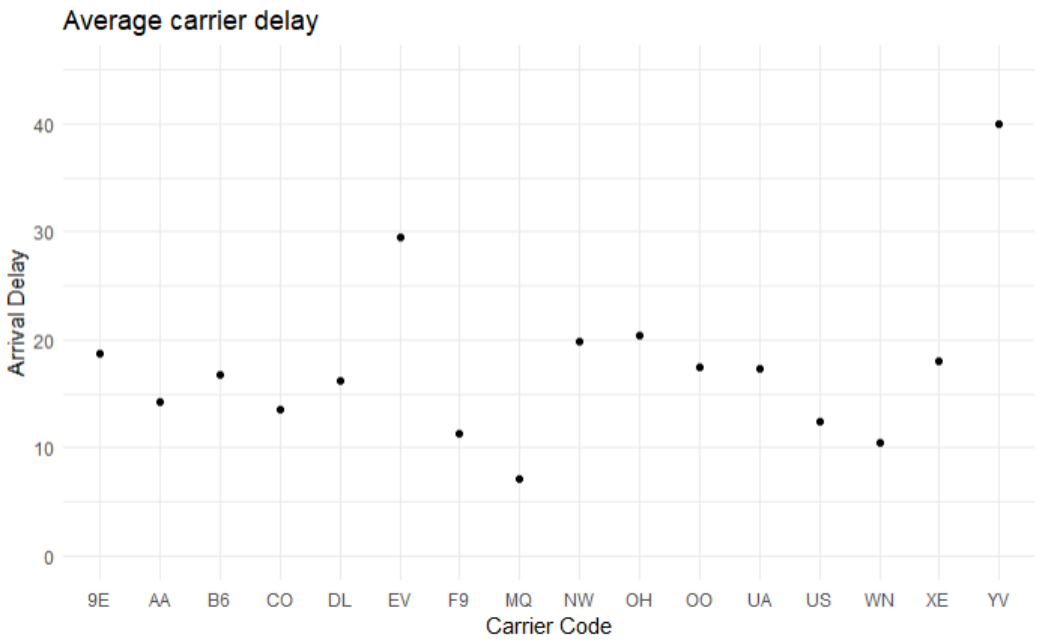
Third, the way the author calculates the premium rent for green buildings is too generic as there are confounding variables. With these confounding variables, we are not sure how the green rating directly influences the rent. Age is one of the confounding variables. As shown in the plot, green buildings are highly concentrated in the lower range of age, which means they are relatively new, thus having higher rent.

We have seen that the analysis by stats guru is flawed since he fails to account for all the factors that affect the rent. Though green building might be a great idea in terms of rent revenue, the stats guru ignored too many factors into his analysis.

2. Visual Story Telling Part 2: Flights at ABIA



According to this plot, I can see late aircraft delay is much higher than any other delay types except for September and October. Overall, the most delays occur in the winter months.



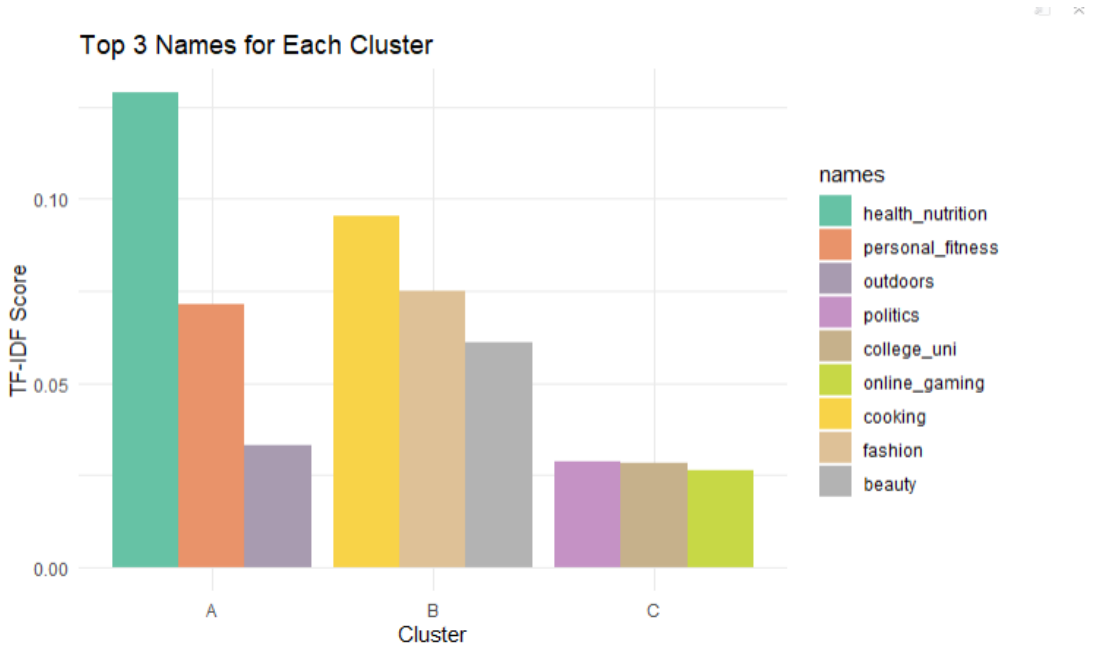
From these plots I can see that the higher average arrival delay for each carrier may not be due to their own fault. Airline YV and EV have both higher average arrival delay and average carrier delay, so it may be unwise to choose from these two carriers. Aircraft Delay has the largest delay time among all types of reasons. STL also ranks the first among this group. Therefore, based on the average total delay time for different airports, we can determine that STL is the worst airport to fly in.

4. Market Segmentation

I use TF-IDF to recalculate the weight of each term for every follower. TF stands for term-frequency, measuring how frequent a term occurs in a follower's tweets: the more frequent a term occurs, the more important it is to the follower; IDF stands for inverse-document-frequency, measuring how frequent the term occurs in the whole dataset: the more frequent a term occurs, the less important it is to every follower.

I use ‘cosine’ as a measurement for the similarity. It calculates the cosine of the angle between two vectors. It measures difference in orientation instead of magnitude. For example, I have 3 following A,B,C with features like A={‘food’:8,‘shopping’:7}, B={‘online_gaming’:13,‘outdoors’:4}, C={‘sports_playing’:6,‘shopping’:6}, I would consider A more similar with C than B.

By looking at different outputs of different Ks, I chose k=4 as our final parameter since its output makes more sense to us.



From the topics of high TFIDF-socres in the clusters, I can infer that first cluster represents people who care a lot about health and fitness; the second cluster represents college/high school students; the third cluster represents people who care about current events, most likely working people.

5. Author Attribution

There are 32,589 words in document-term matrix for the test data, however there are only 3325 words which are also common in the train data set. So, I will drop the remaining words for the classification problem.

Create the TF-IDF matrix for test and train data:

3325 words are still high to conduct classification. Thus I will reduce the dimensions using Principal Component Analysis. I will run the PCAs on the train data set and take the top words which explain 75% of the variability in data.

Run PCA on TF-IDF to reduce the number of words:

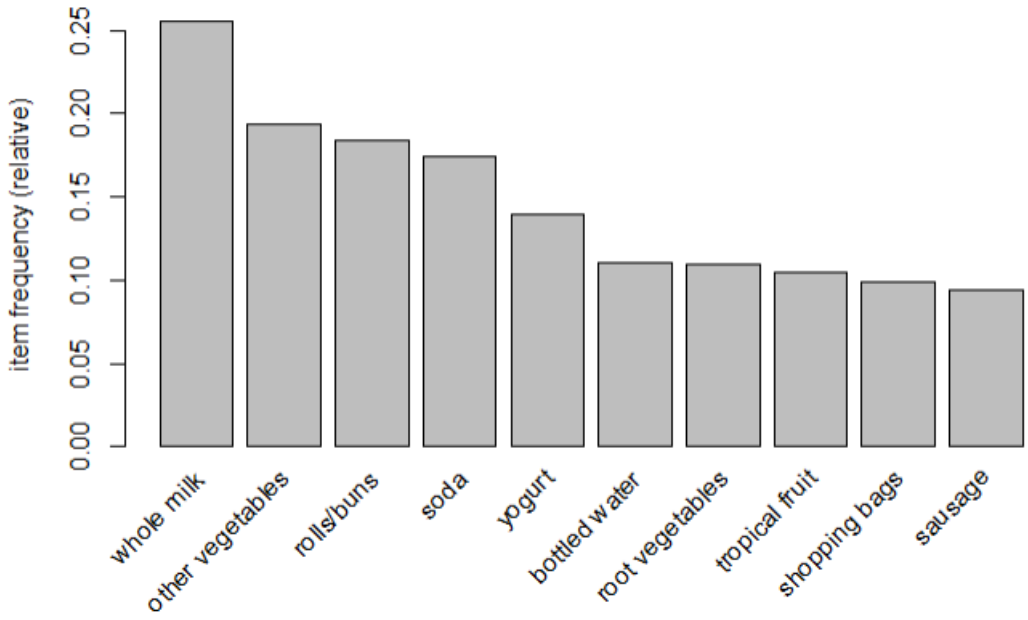
Based on the PCA on train data, I can say that 330 words define 75% of the variability. Using the PCAs on the train data set, I predicted the PCAs on test data set. I will use these data sets for our classification of the authors.

Classification - Random Forest

After running Random Forest, overall, the output of the model gives an accuracy of ~59%. Overall accuracy seems to be low as there are some authors who are not predicted that well. One potential reason behind this could be that I have dropped quite a few words form the train and test data sets. However, there are many more words in the test data (which if incorporated could improve accuracies).

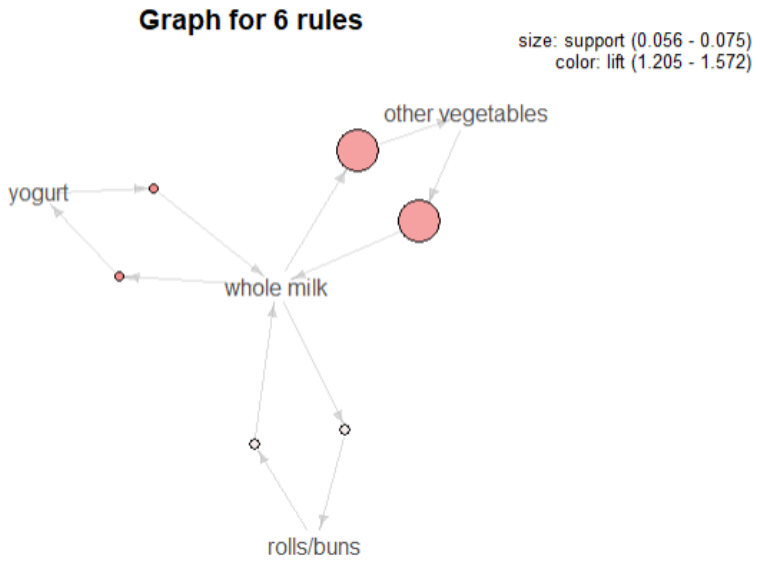
6. Association Rule Mining

I transform the data into a “transactions” class before applying the apriori algorithm. The summary of the dataset reveals the following: 1. There are total of 9835 transactions in our dataset 2. Whole milk is the present in 2513 baskets and is the most frequently bought item



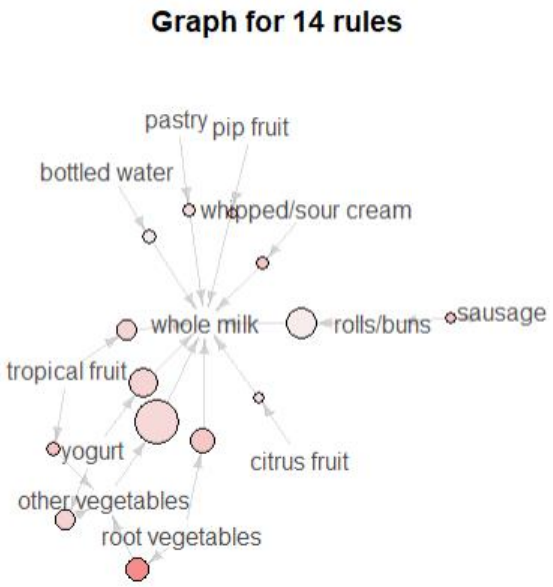
Rules with support > 0.05, confidence > 0.2 and length <= 2

I also notice that most relationships in this item set include whole milk, yogurt and rolls/buns which is in accordance with the transaction frequency plot we saw earlier. These are some of the most frequently bought items..



Rules with support > 0.03, confidence > 0.3 and length <= 2

This item set contains 14 rules and includes a lot more items. However, whole milk still seems to be a common item.



In conclusion, whole milk is the most common item purchased by customers. Whole milk, yogurt, rolls/buns are all often purchased together, which could be due to the close proximity to one another in a grocery store.