Capstone Project 1 Milestone Report

I’m going to attempt to help solve the problem of selecting whether to undergo a real estate development or acquisition project. Using the rough information given to me by a self-storage company, I will analyze past project proposals and whether they are undertaken or not, to determine whether a future project should be deemed worthwhile.

By solving this problem, I will hopefully be able to significantly speed up the process of selecting which projects to take on. Many projects are proposed every week at an increasing rate, with the majority of them being turned down. Although this is the case, almost every project proposal has to be looked at extensively before it can be determined whether to turn them down or not. If by using this data I am able to determine more quickly which projects to take on then much time will be saved and more profit-generating projects can be accepted.

In the short-term, what I have listed is the goal for this problem, but in the long-term this solution can potentially offer even more. In the real estate self-storage business the ultimate profitability of a project cannot be determined for several years, so this analysis is only determining whether based on the history of the company whether a project should be considered or not. In several years, the profitability of projects will become available and using a very similar data set and model an estimate of the profitability of a project could be determined, offering a greater solution than what the current data can give.

To solve this problem, I initially will be given individual project proposal workbooks from the past four years. From this data I will use python to aggregate the relevant information in the project workbooks into a single workbook that will have the pertinent information for a given proposal in each row of the workbook. After compacting this information, I will clean and analyze the data.

To clean and wrangle the dataset, I first used pd.read\_excel() function to read in data from approximately 250 project proposals. Once they had been read, I selected the pertinent information from each proposal and assembled that information into one spreadsheet. Additionally, I set up code to add the relevant information from any future project proposals to the workbook.

Now that I had all my data in one sheet, I analyzed the data to see what needed to be cleaned. I reset the index, while dropping the existing one to more easily analyze the data in the future.

Next, I made the ‘Vehicles Per Day in Front of Site’ column machine readable. I started by making the column data astype(str). Using regular expressions, I found numerical values in the data, which were the amounts of vehicles, and not part of street names. Using that information, I summed up these numbers using lambdas and apply(sum) to have a single numerical value for the number of vehicles that passed by a given site per day.

Next, I dropped all the non-numerical data for ease of reading for the machine learning model. This involved dropping the first eight columns and an additional ‘Best Comp’ column.

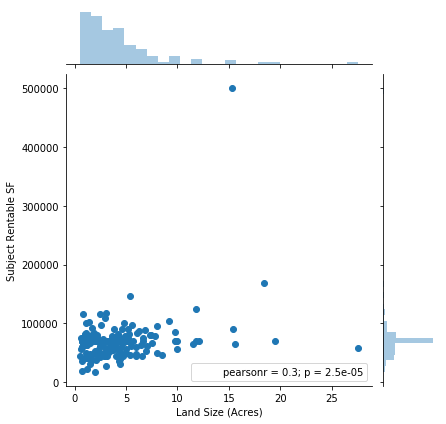
Following this, any row where the ‘Accepted’ column equaled -1 was dropped. The -1 in this case indicated that the project had neither been accepted nor rejected, these projects are still under review.

Knowing the self-storage industry, I created two new columns from the existing data that were likely to be of use to the model. These are the ‘Diff now’ and ‘Diff later’ columns. They represent the difference between the MSA SF per Capita and the corresponding existing SF per capita (now) and SF per capita including planned (later).

After this, I determined the amount of missing data in each column using the .info() function and saw that there was too much missing information on the asking price for each project. Thus the pricing data had to be dropped. The next column with a significant amount of missing data was the ‘Subject Rentable SF’ column. This column’s data was fairly consistent with few outliers, so it was determined that filling in the missing values could be acceptable. In the creation of the workbooks if the subject rentable SF is unknown it is filled in with 70000. Thus filling it in with the mean value of the ‘Subject Rentable SF’ column, which was very close to 70000 is viewed as acceptable.

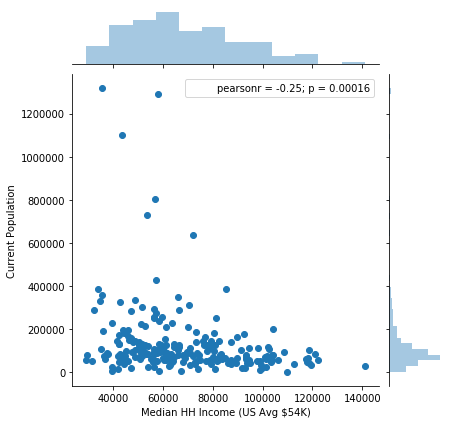
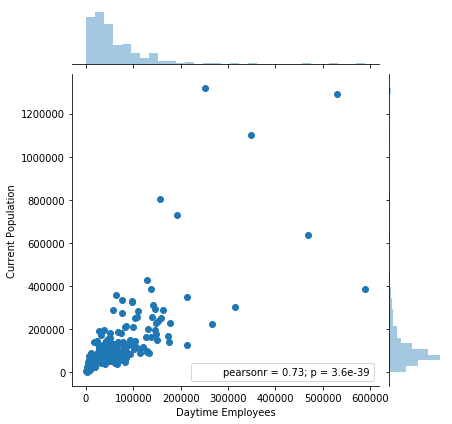
After cleaning and wrangling my data, I did some basic EDA to see what I could discover. First, I created violin plots for each column in the data set. I found most of the columns to have distributions with a long right tail. This made sense because for most of the data there is a lower limit, but no upper limit.

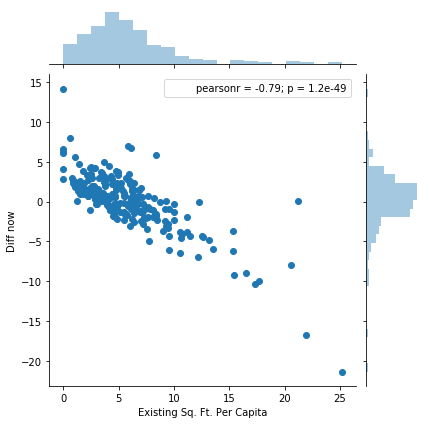
I followed this by looking at a correlation matrix of my data. Since the matrix was large (23 features), I broke down the correlations into smaller groups.



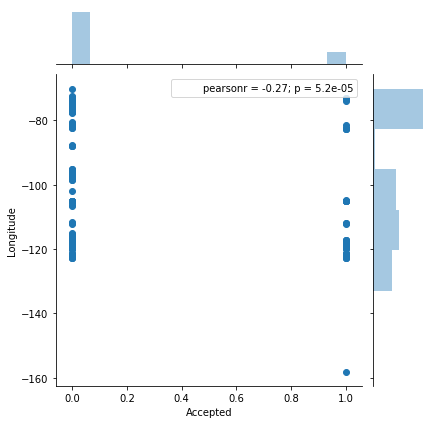
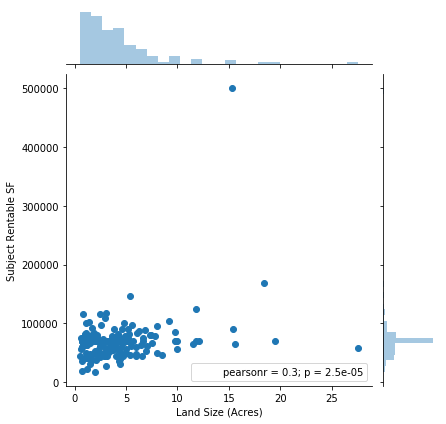
The first group was land size. Since there were two land size variables, with one in acres and one in square feet, it seemed likely that they would be nearly the same. This proved to be the case and as a result the land size square feet column was removed because it represented the same thing as the land size acres column. We should expect a strong positive correlation between land size and subject rentable SF, but the correlation is weak. This could be because of multi-story buildings where the rentable SF could be larger than the actual land it is on, skewing the relationship.

Next, I compared the demographic data along with the latitude and longitude of proposed stores. Current population, population in 5 years, and current households all had very strong correlations as expected and additionally some positive correlation with 5 year population growth and 5 year SF demand. Due to this high correlation (above .99) of the first three variables, population in 5 years and current households were dropped. The results also showed that 5 year population and 5 year SF demand are perfectly correlated, which is unsurprising because it is a formulaic calculation. As a result, the 5 yr population growth column will be dropped to avoid having essentially the same information twice. Interestingly median HH income had a slight negative correlation to all of the population variables. More population tended to mean a lower median income as well. The daytime employees column had a fairly strong positive correlation to population as expected. More people should mean more jobs. There was no strong relationship between vehicles, latitude, longitude and anything.



After the demographic data, I compared the self storage market data. I found there was a strong positive correlation between existing comparable facilities and existing SF in comparable facilities. Existing SF per capita and SF per capita including planned had a strong positive correlation. Existing SF per capita had a strong negative correlation with diff now and diff later. Subject rentable SF, MSA SF/ capita, and accepted columns did not have a strong relationship with any other column. Additional planned/ proposed facilities and additional facility SF were close as expected. SF per capita including planned had a strong negative correlation to diff later, which makes sense since the diff later column is derived from SF per capita including planned. Diff now and diff later also have a strong positive correlation as expected.

Seeing the relationships in the subgroups I then looked at the entire data set for correlations again, but with a greater understanding of what I was looking for. Looking at all of it, more correlations were discovered. The current population, population in 5 years, current HH, and daytime employees had a significant positive correlation to existing facilities and their SF. Median HH Income and existing comp facilities also had a somewhat significant negative correlation. Ultimately, longitude had the greatest correlation with our target variable, accepted, but it was still relatively weak at -0.27.

After exploring the correlations between the columns for all the data, I plotted histograms of the accepted project proposals data vs the rejected project proposals data. Comparing the histograms for each I didn’t find any obvious differences in the distribution for any given column though.