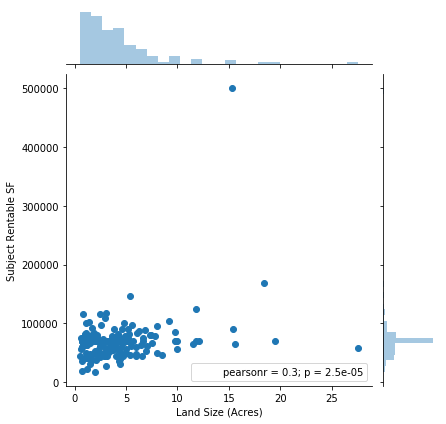
Project 1 Statistics

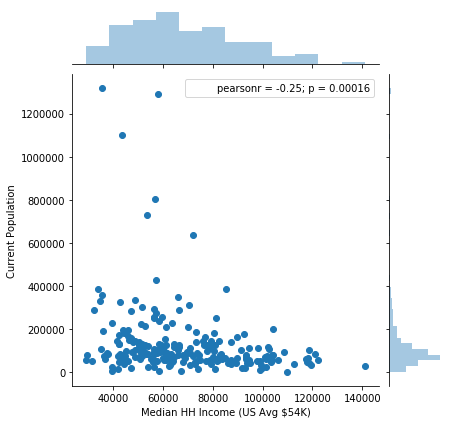
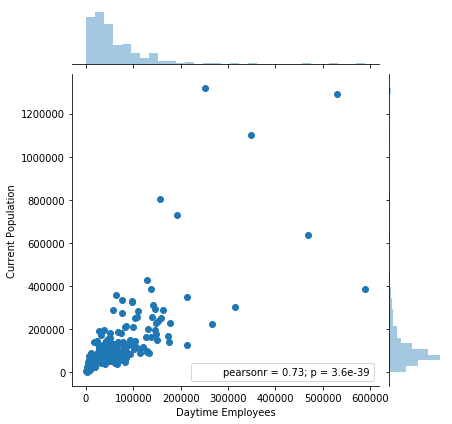
After cleaning and wrangling my data, I did some basic EDA to see if I could discover anything. First, I created violin plots for each column in the data set. I found most of the columns to have distributions with a long right tail. This made sense because for most of the data there is a lower limit,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

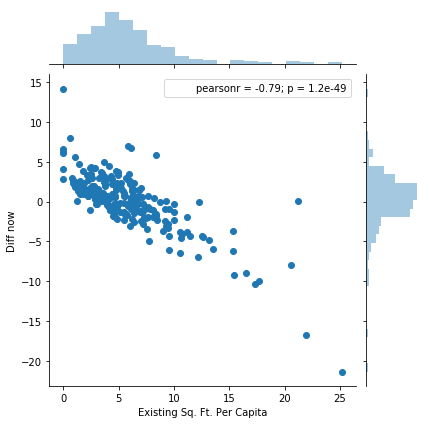
I followed this by looking at a correlation matrix of my data. Since the matrix was large (23 features), I broke down the correlations into smaller groups.



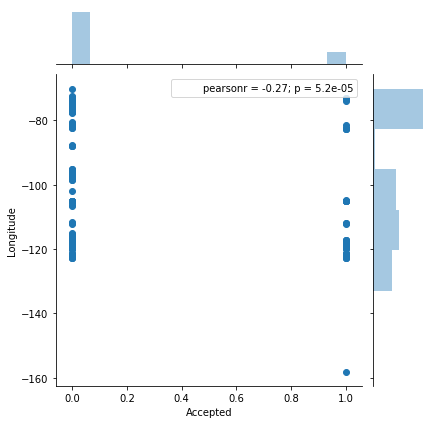
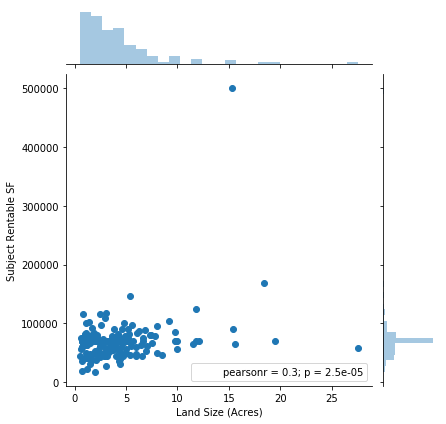
The first group was land size. Since there were two land size variables, with one in acres and one in square feet, it seemed likely that they would be nearly the same. This proved to be the case and as a result the land size square feet column was removed because it represented the same thing as the 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.

