Problem 1 rough draft.

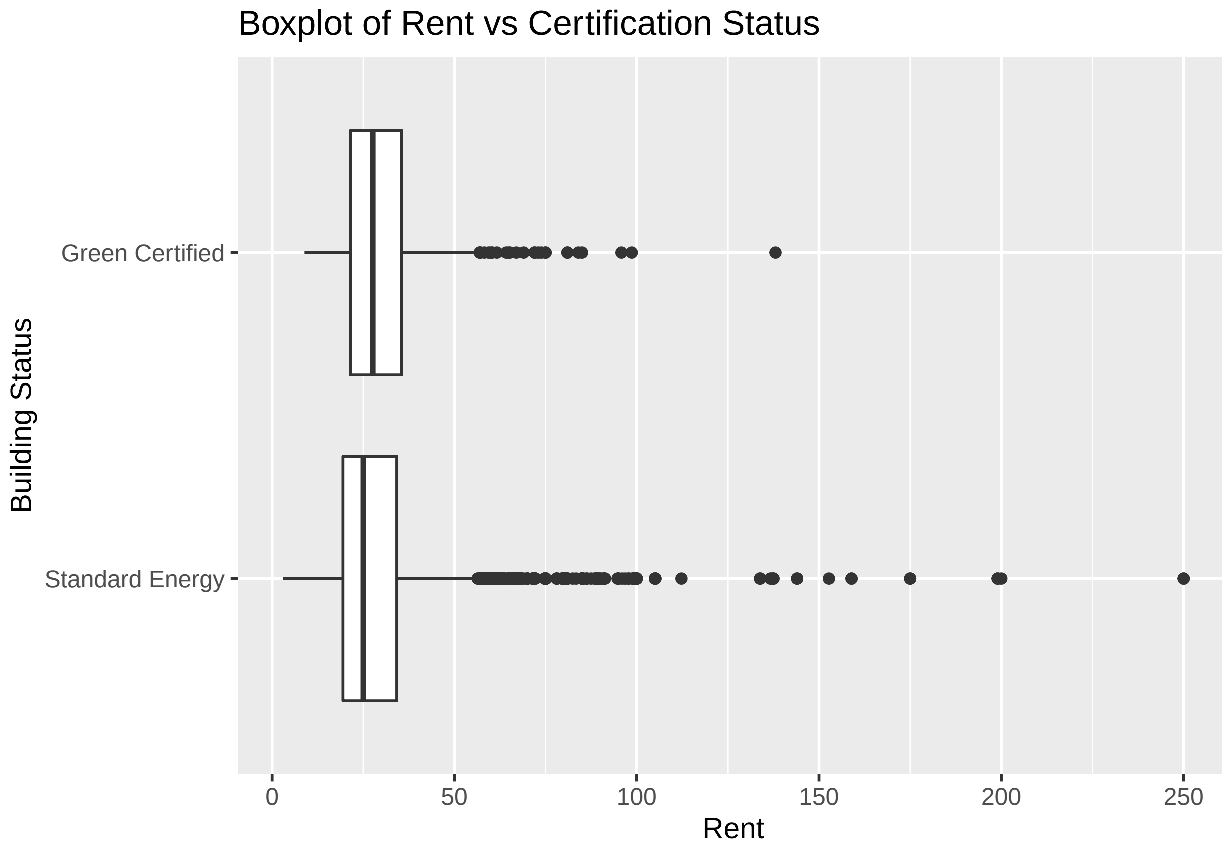
Matthew Leong

Problem 1:

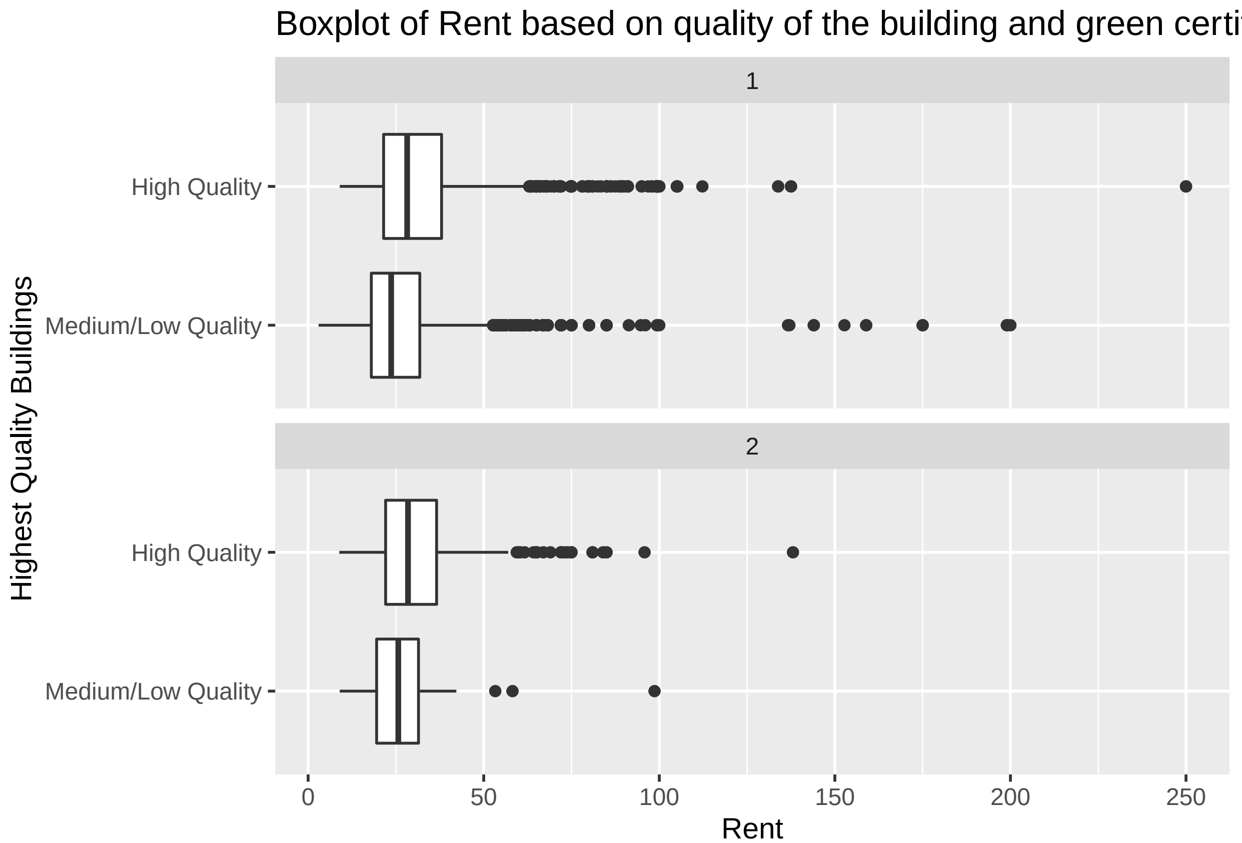
There were a number of things that sounded off about the "Excel guru"'s analysis but first, we wanted to check if there was an unacceptable amount of data lost.

```
## The original dataset originally had 7894 observations.

## The new cleaned dataset has 7679 observations.

## They differ by 215 observations

## which is a percentage difference of 2.723588 percent.

## There are 684 green certified buildings in the data set.

## There are 6995 non green certified buildings in the data set.
```

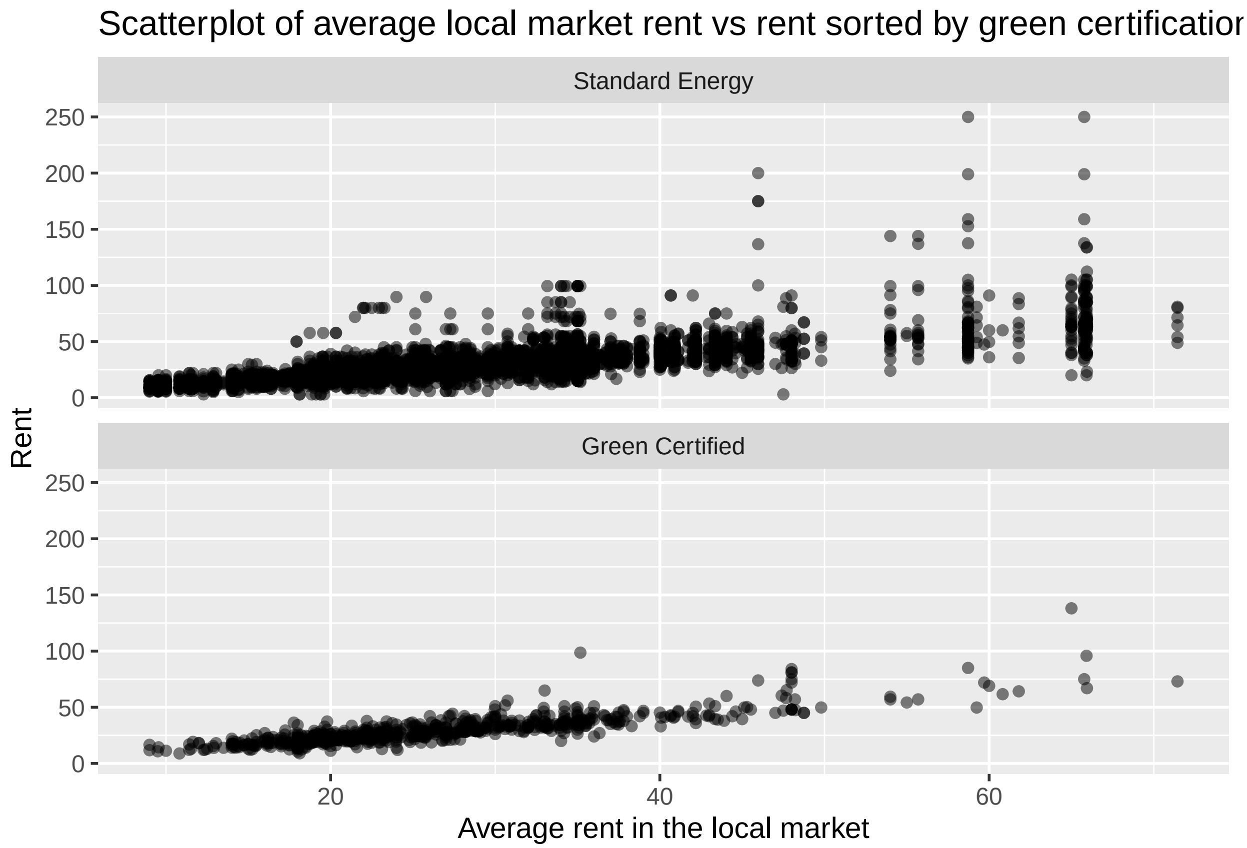
After running some code really quickly, we can see that the guru lost about 2.72% of the data which is honestly an acceptable amount. As such, we will also be using this subset of the data for our plots. However, the guru's analysis falls apart because he forgot to consider other variables. To demonstrate this issue, let's look at the box plot obtained when the other variables are not considered.



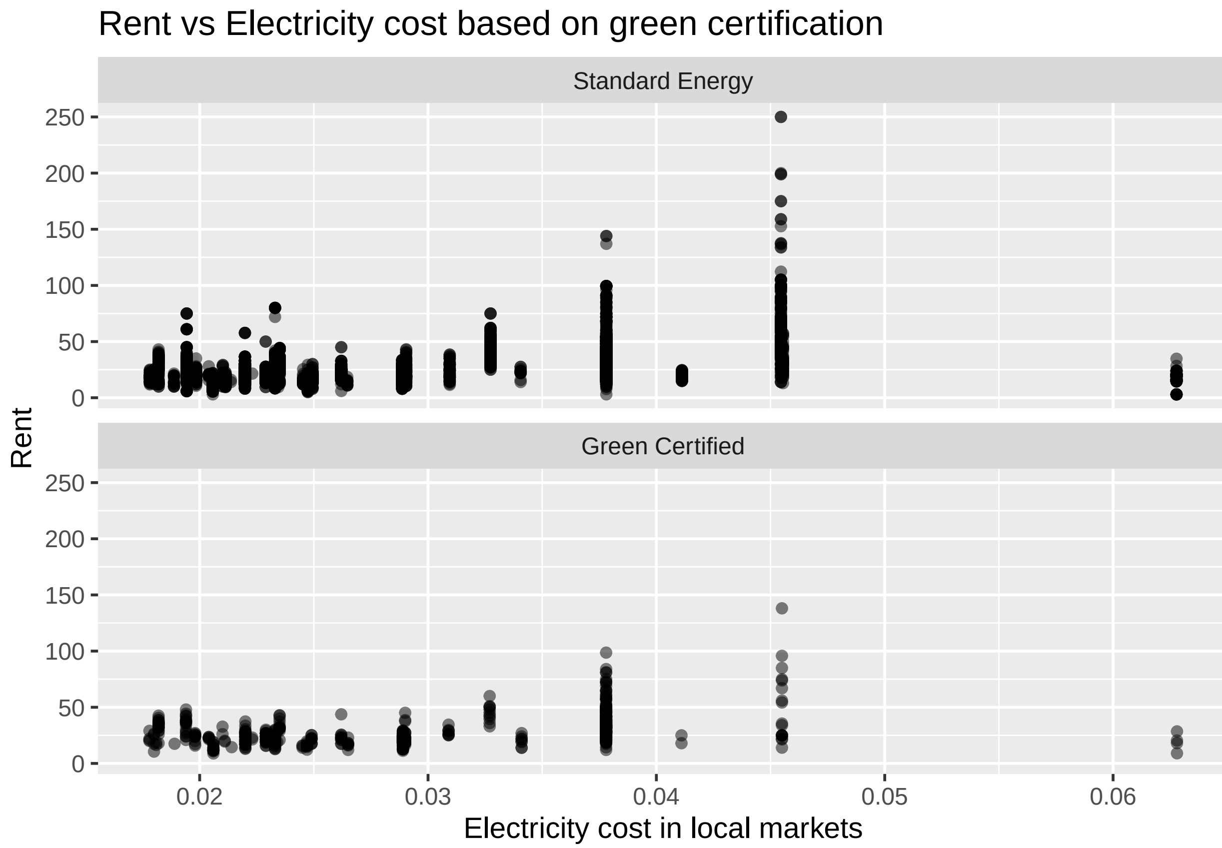
The box plot demonstrates the distribution of the split that the guru did. As we can see from the box plot, rent does not seem to differ that much between green certified and standard energy. However, the guru ignores this overall distribution in favor of focusing on the median. The median here is misleading as while it is higher on this distribution, it will not necessarily hold true when we add in additional factors. For instance, let us see what would happen when we add in whether or not a building is high quality.



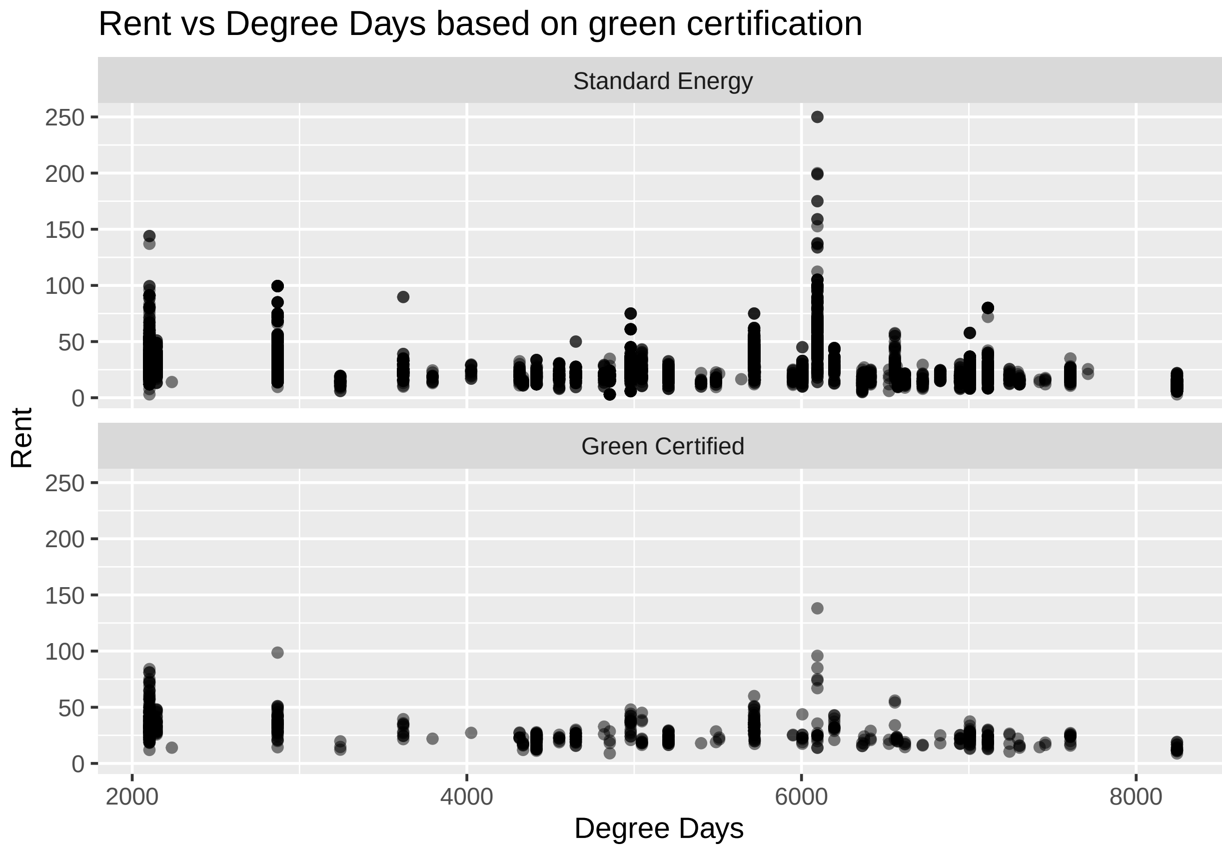
Naturally, as one would expect, high class buildings have a higher rent associated with them. If we go with the median standard, we can see that although green certified medium/low quality buildings have a higher rent, standard energy (non certified green buildings) and green certified high class buildings have about the same median rent. So this begs the question of why rent is suddenly the same. The answer is that this is a confounding variable problem which is basically how another variable like high class status affects rent and green certification status. This is not the only confounding variable.



Another such confounding variable left out by the guru is the average rent in the local market. This variable has a key relationship with rent charged in that they essentially deal with the same thing. If green certified buildings netted a higher rent, this would be reflected in the graph when controlling for average rent in the local market. However according to the graph, it seems that green certified buildings perform worse but are more consistent overall which can be explained by the fact that there are a lot fewer green certified buildings than non green certified buildings. Additionally, the distributions are fairly similar implying that green certification does not actually affect the rent price. To further support this case, let's look at a couple more potential confounding variables and see if the distributions differ.



Again, when controlling for electricity costs, we can see that the distributions do not differ significantly further implying that rent and green certification are unrelated. In fact when looking at the distribution, the rent appears to be lower in the green certified buildings. One possible explanation for this phenomenon is that green buildings have lower fixed costs than standard energy buildings which allow them to charge less rent based on energy usage. This odd phenomenon is also reflected in another energy measurement variable.



Degree days in this data set are essentially measures of demand for energy where higher values mean greater demand. As this is another energy variable, we would expect to see and do actually see a similar distribution in that green certified buildings have lower rent than standard energy buildings.

In conclusion, the confounding variable problem invalidates the guru's analysis. When controlling for such variables, we can see evidence that points towards green certified buildings have equal or even lower rent than non green certified ones. We can conclude that going green does not necessarily mean higher rent. TO further verify this conclusion, we recommend a linear regression or some other similar predictive modeling algorithm to control for all the variables and see which ones would actually lead to higher rent. In fact having run the model ourselves, we found that green rating has no significant relationship with rent. Additionally, we would recommend running such a model on the green rating to determine other potential factors that are actually related to green buildings. Regardless, we do not recommend investing in a green building as it does not yield higher rent and has an additional premium attached to it.

Problem 2: