Problem 1 and 2

Matthew Leong

8/11/2020

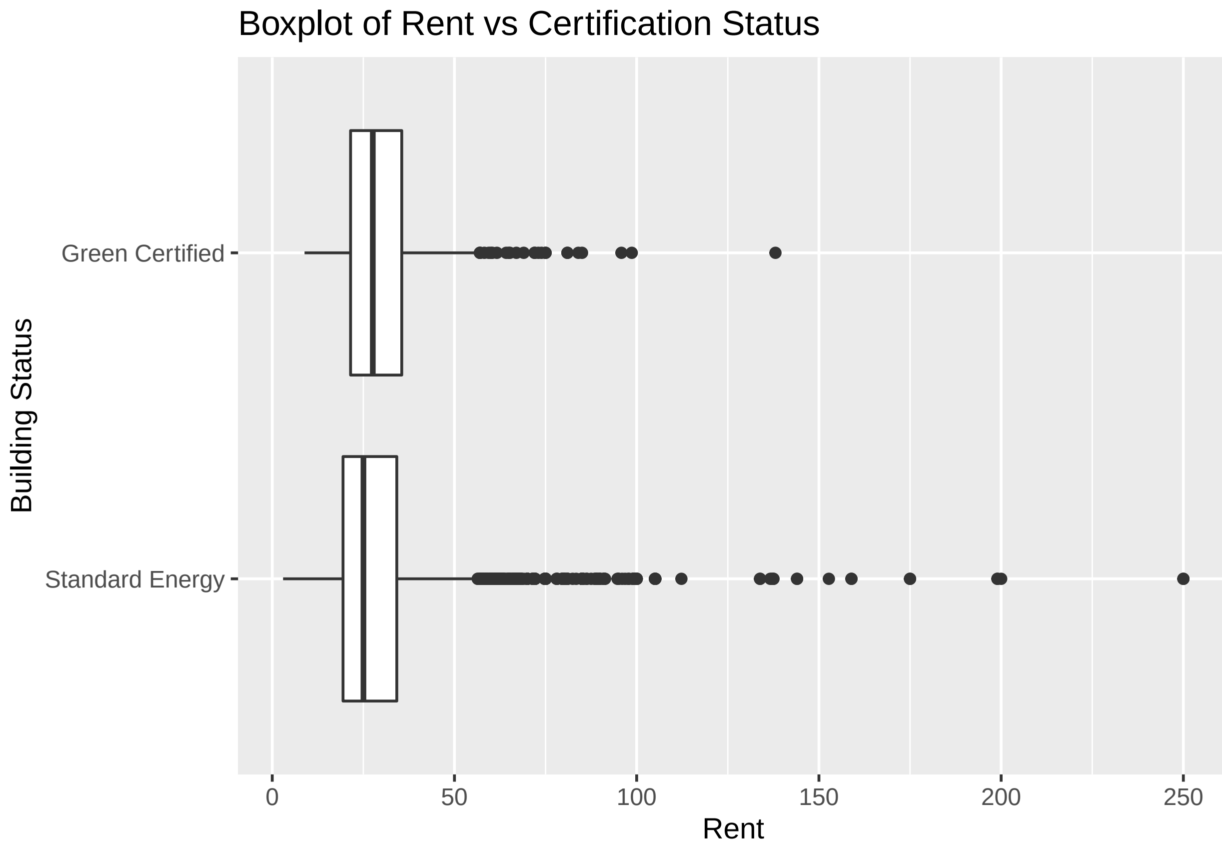
The 'Excel guru" took out the buildings with very low occupancy rates (so leasing rate) (less than 10%). He seperated green and non green buildings (based on green rating). He looked at market rent (Rent) variable. Notably we see that the guru did not look into cluster rent. He also did not take into account net: indicator for whether the rent is quoted on a "net contract" basis. Net rental meaning pay their own utility costs. These costs are not included in rent. The real estate developer wants to build her project on East Cesar Chavez, across the I-35. $100 million baseline cost plus 5% premium for green certification.

Problem 1:

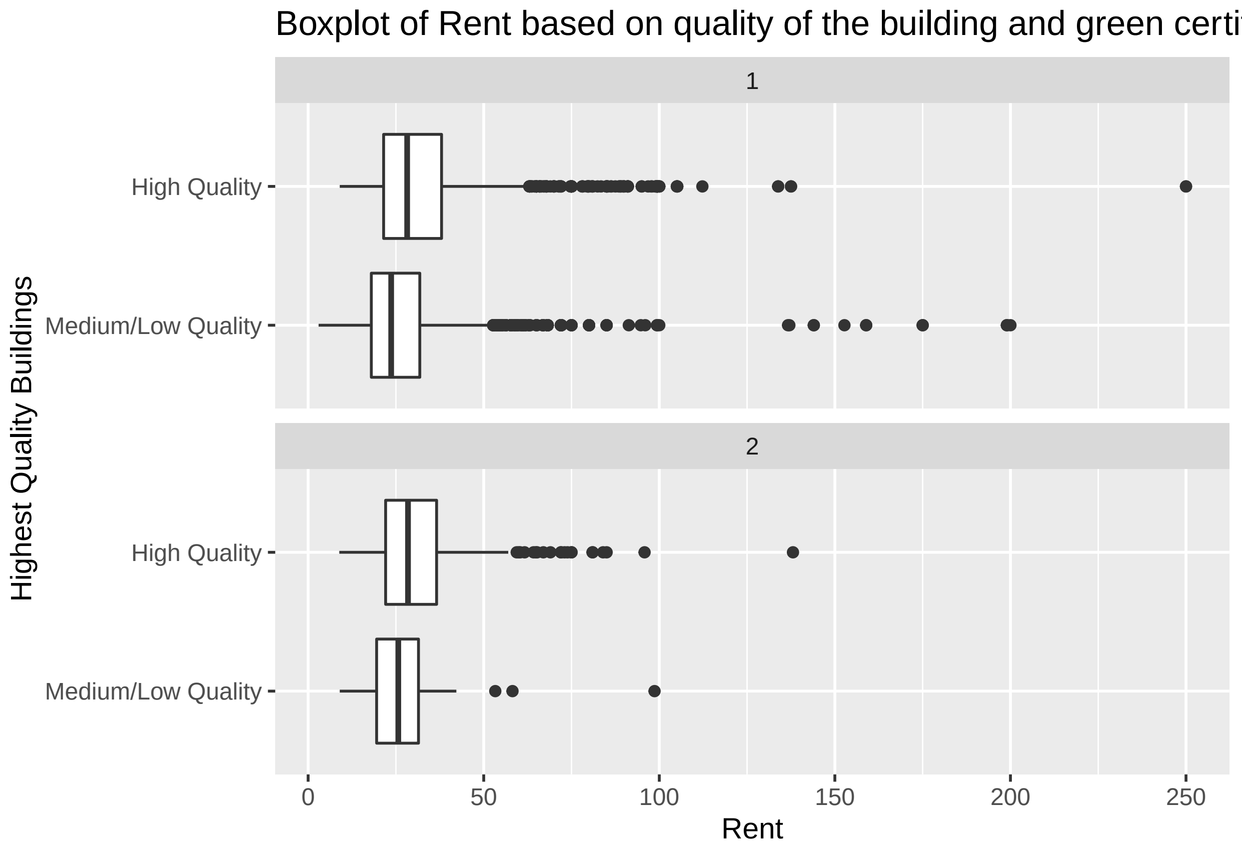
There were a number of things that sounded off about the "Excel guru"'s analysis but first, we wanted to check if there was an unacceptable amount of data lost.

```
## 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.
```

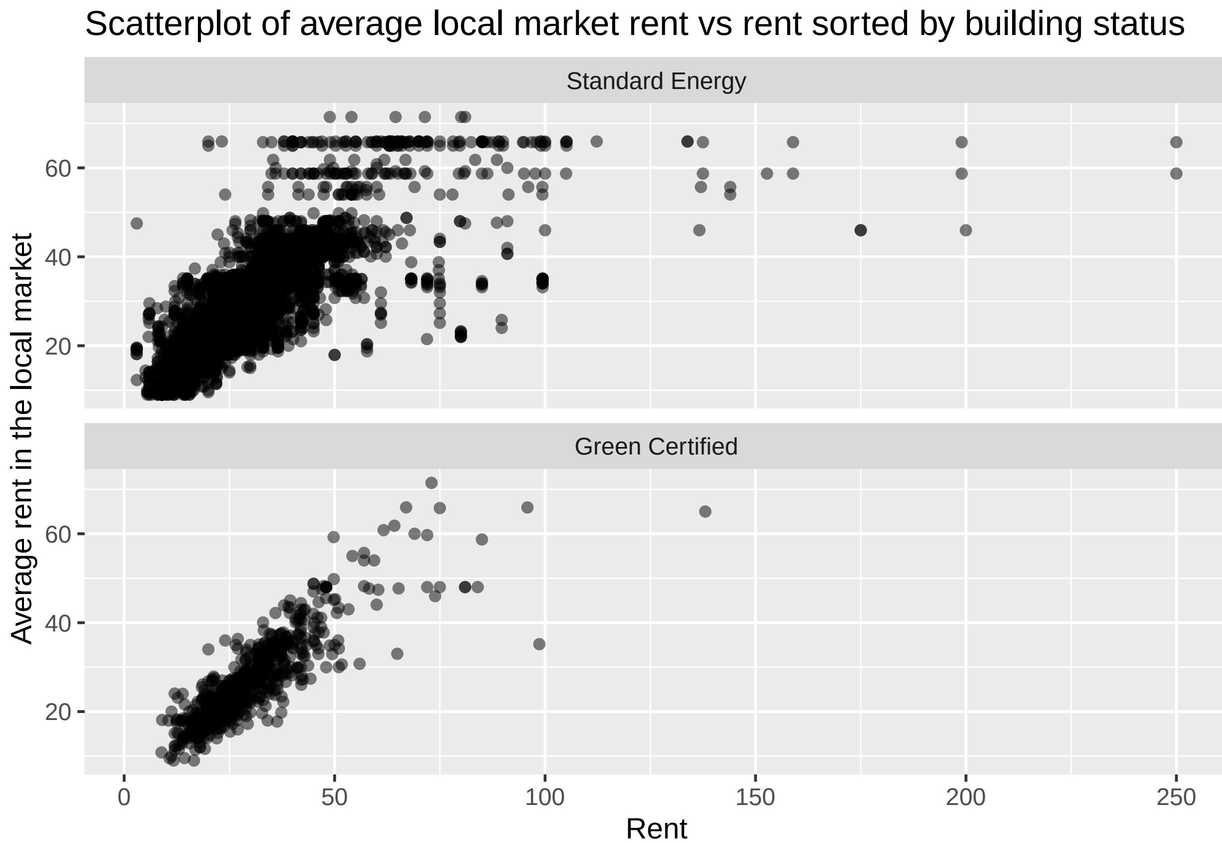
After running some code really quickly, we can see that the guru lost about 2.72% of the data which is honestly an acceptable amount. However, there is a key issue with this analysis in general. He assumes a causal relationship or strong positive correlation between rent and green rating and assumes that the other variables are unrelated. To demonstrate, this issue let's look at the boxplot 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. The guru ignores the overall distribution though and bases his analysis 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.



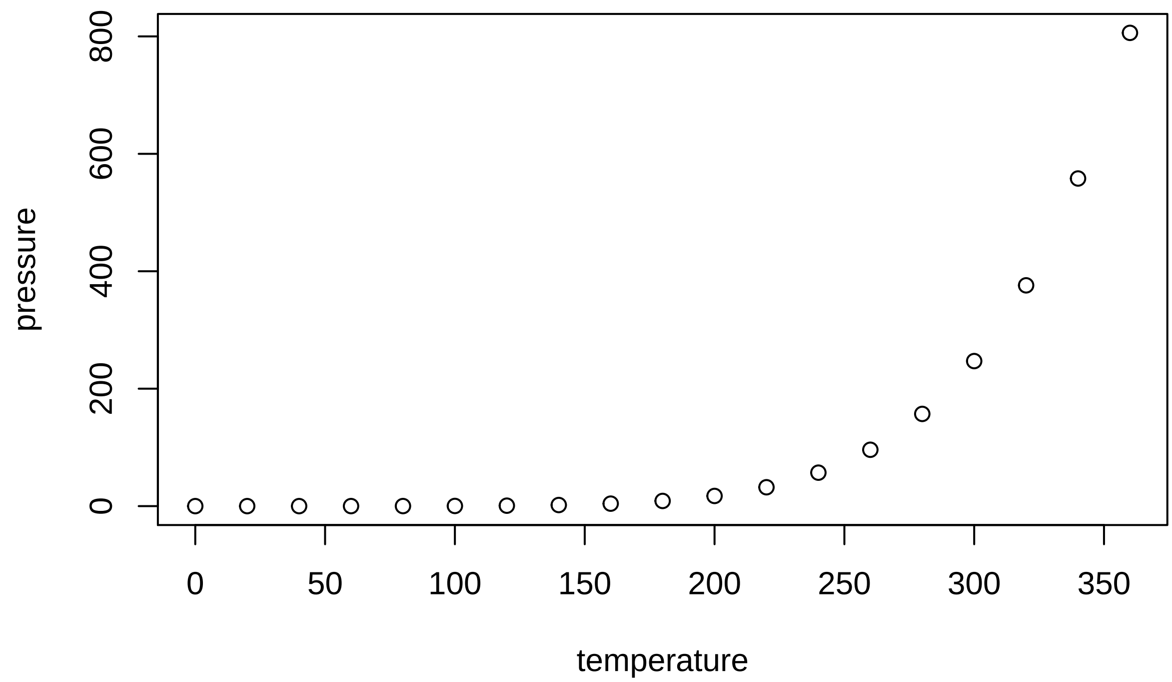
Naturally, as one would expect, high class buildings have a higher rent associated with them. Basing our analysis off of median, we can see in the box plot that while green certified medium/low quality buildings have a higher rent, standard energy (non certified green buildings) and green certified high class buildings have about the same median rent. This is essentially a confounding variable problem which entails how another different variable could affect rent and green certification status.



If we look at the scatterplot for rent and average local market rent.

Problem 2:

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.