Burlington Property Analysis

Yujia Zhang, Ya Tuo

Data Science I (STAT 287), Instructor: James Bagrow

College of Engineering and Mathematical Sciences

The University of Vermont

Abstract

Burlington is the most populous city in the U.S. state of Vermont. It is the city where the University of Vermont is located and Ben and Jerry opened their first ice cream scoop shop. Studying the property value data of Burlington leads a better understanding of local property market which is essential for investment activities and property tax, the major revenue source for local government. Six factors: the grade of the building, the type of building, the land use type of the building, the year of built, the distance to the closest park and the improvement cost of parks nearby were studied about how their impact on building values.

Keywords: Data Analysis, Regression Test, Time Series, Kruskal-Wallis test

1. Data

Two datasets were used in this study. Both of the datasets were found in the City of Burlington's Open Data Portal. One dataset, City of Burlington Property Detail, provides a detailed list of all taxable property in Burlington VT, including current valuations, ownership, location, building type, area, grade and most recent sales information. The other dataset, Penny for Parks Improvement Data, provides a list of park improvement projects from 2010 to 2014 in Burlington VT, including project name, project status, location, and cost.

For the dataset, City of Burlington Property Detail, six variables were kept: Year of built, Grade, Building Type, Land Use, Current Building Value,

and Location. For the other dataset, Penny for Parks Improvement Data, three variables were kept: Project Cost, Project Location, and Park Name. The distance between each property to its closest park was also analyzed. It was calculated based on the geographic locations of the property and the parks. Both of the datasets were filtered out missing values and values that do not make sense (negative values for building value for example) before starting any further analysis.

2. Methods

Both descriptive analysis and inferential analysis were used in this study. For the dataset City of Burlington Property Detail, the impact of Grade, Building Type, Land Use and on the Current Building Value was studied separately. Firstly, for each variable, the boxplot of current building value over different levels of that variable was generated to visualize the how the current building value was distributed among different levels. Since the Current Building Value was highly skewed to the right, instead of ANOVA, where normally distributed data is required, Kruskal-Wallis test by ranks, a non-parametric method was used to test whether samples originate from the same distribution. If the null hypothesis (all means are equal) was rejected, then the conclusion that buildings with certain levels of that variable have higher values can be drawn, which also means that variable does have impacts on the building's value.

Since both the current building value and the Shortest distance to parks were skewed, both of those two variables were boxcox transformed before applying linear regression test to determine the correlation between those two variables. Then the property was grouped by the closest park and the difference in current building value among different closest park was analyzed if properties near certain parks had higher value, and how the improvement cost will impact the building's value.

3. Results and Discussion

Six factors that may impact the building's value in Burlington were studied. The study results of Grade, Building Type, Land Use, Year of Built, Distance to the Closest Park, and the Improvement Cost on the Closest Park

were shown below with figures, calculated statistics, and interpretations.

10,222 properties were analyzed. The mean of the building value was 3,238,40 US dollars, and the median of the building value was 150,100 US dollars. The max was 367,113,200 US dollar, and the min was 100 US dollar.

- How would the Grade of the property impact the value?

There were seven levels of the Grade. The most frequent grade was "Average" with 5389 counts as frequency. In figure 1, three levels that had the local median value closest grand median value were "Average", "Fair", and "Good Grade", which also implied that the values of buildings with those three grades can present the general building values more accurately than other grades.

In figure 1, seven levels of the Grade of Buildings were sorted by the median value of the property value with respect to each level. As the median value went higher and higher from the left side of the plot to the right side, some patterns of the levels were also noticed: the grade was also getting better and better, which means that those buildings with better grade tend to have higher values. The statistics of the Kruskal-Wallis Test was 4290, the p-value was 0, which showed a very significant evidence that the building value various among different grades. In other words, Grade had a strong impact on the building value in Burlington.

- How would the Building Type of the property impact the value?

There were 118 levels of the building types. The most frequent one was "Old Style" with 2169 counts as frequency. To make the boxplots more effective and readable, only the 10 levels with the highest frequencies were plotted. In figure 2, 9 out of 10 boxplots had local medians building values that were very close to the grand median of the building values. Only one level, the "Apartment" type had significantly higher median values.

The statistics of the Kruskal-Wallis H-test was 3883. The result was significant, some certain levels of the building type did imply higher building value. But in figure 2, most the levels were closed to the average levels of building values, so the building type may not play such an important role in the determinants of the building values.

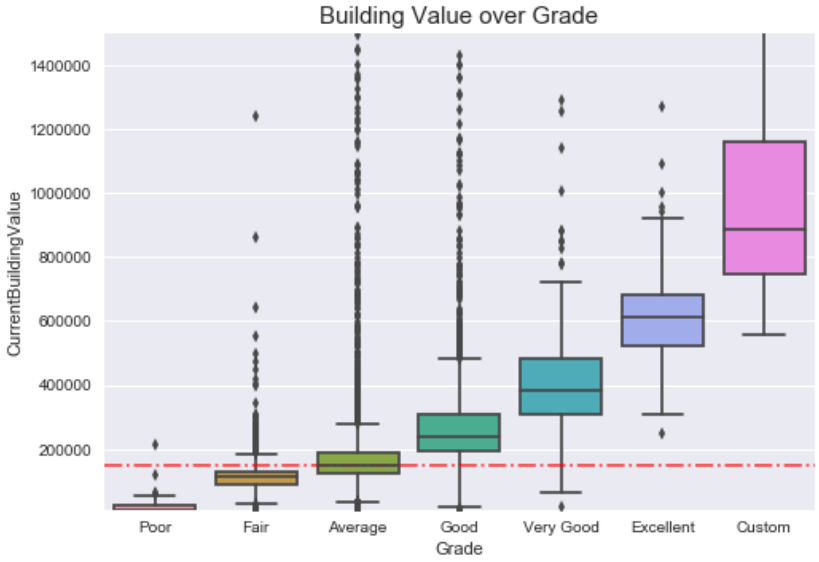

Figure 1: The building value over grades in boxplots with grand median of the building value (the horizontal red dash line)

- How would the Land Use of the property impact the value?

There were 22 levels of the building types. The most frequent one was "Single Family" with 5156 counts as frequency. To make the boxplots more effective and readable, only the 10 levels with the highest frequencies were plotted. In figure 3, 9 out of 10 boxplots had local medians building values that were very close to the grand median of the building values. Only one level, the "Exempt" land use had significantly higher median values and significantly higher range of building values.

The statistics of the Kruskal-Wallis H-test was 2272. The result was significant, some certain levels of the land use did imply higher building value. But since in figure 3, most the levels were closed to the average levels of building values, so the land use may not play such an important role in the determinants of the building values.

- How would the year of built of the property impact the value?

The earliest building on this dataset was built in the year 1798. The most recent one was built in 2015. 75 percent of the buildings were built between 1910 and 1975. The records for buildings built earlier than 1900 were few, so

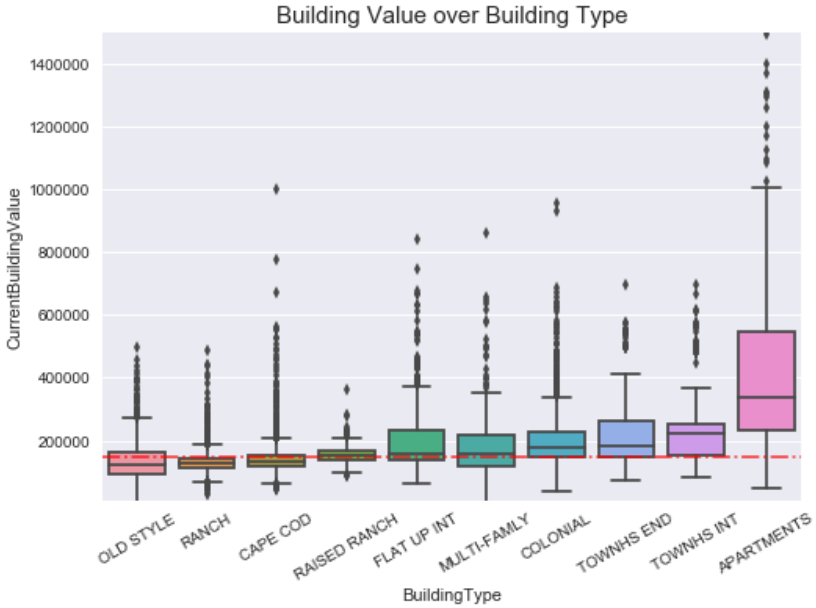

Figure 2: The building value over building types in boxplots with grand median of the building value (the horizontal red dash line)

figure 4 showed the building values and their year of built for those buildings that were built after 1900.

In figure 4, Those there were fewer records about recently built properties, the number of buildings distributed roughly evenly among all years. By the year 1980, the median value of properties that were built in that year remained steady and very close to the grand median of all buildings over all years. Start from the year 1980, there were more fluctuations in the median of the building value. One reason could be that newer buildings did tend to have higher values. Another reason could be the lack of records. Especially after the year 2000, from the density of points on figure 4, there were clearly fewer points. With smaller sample size, there tends to be more variance. the fluctuation of the median value was a reflection of that.

- How would the distance to the closest park impact the value?

The median value of the distance to the closest park was 436 meters and mean value of the distance to the closest park was 507 meters. The max

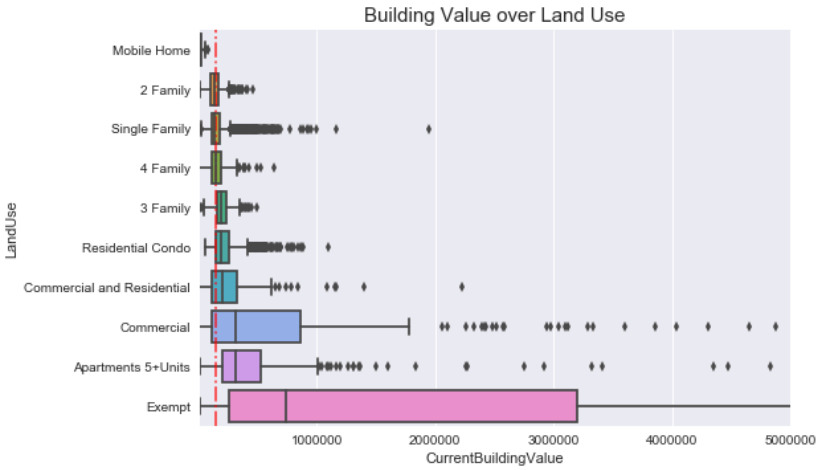

Figure 3: The building value over land use in boxplots with grand median of the building value (the horizontal red dash line)

shortest distance was 1,940 meters and the min shortest distance was 11 meters.

Since both the building value and the shortest distance data were skewed, figure 5 showed a scatter plot of the transformed building values and the transformed shortest distance with the regression line. The density of the points showed that building values were uniformly distributed among the shortest distances, which means there may not be correlations between building values and the shortest distances. The horizontal regression line also supported that conclusion.

In figure 6, most dots located closely to the median building value, which means there was no significant difference in building value across various distances to parks. Even though some dots implied significantly higher median values, the number of buildings near that park was small. So it did not contrast the conclusion that the property value was independent of the distance to the closest park.

- How would the to the closest park impact the value?

The total improvement cost on a park was calculated by summing the cost of each improvement project on a park. Figure 7 showed a bar chart

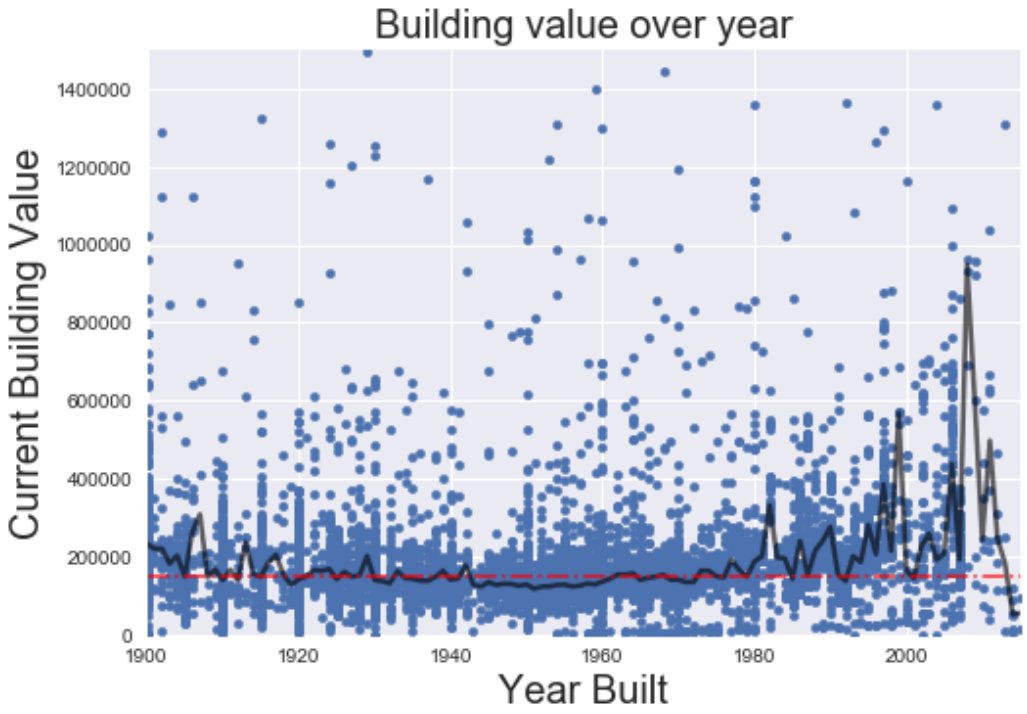

Figure 4: The building value over years with the median values for each year (the black line) and the grand median of building values in Burlington (the red dash line).

of median building value over the improvement cost by parks. Some parks with significantly higher median building values nearby did not have much improvement cost. Some parks with average median building values nearby had the highest improvement cost (Leddy Park for example). So the building values are also independent of the improvement costs on parks nearby.

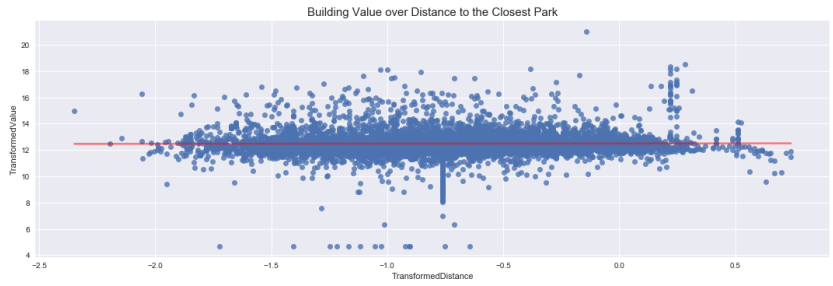

Figure 5: Boxcox transformed building value over the boxcox transformed shortest distance with a regression line

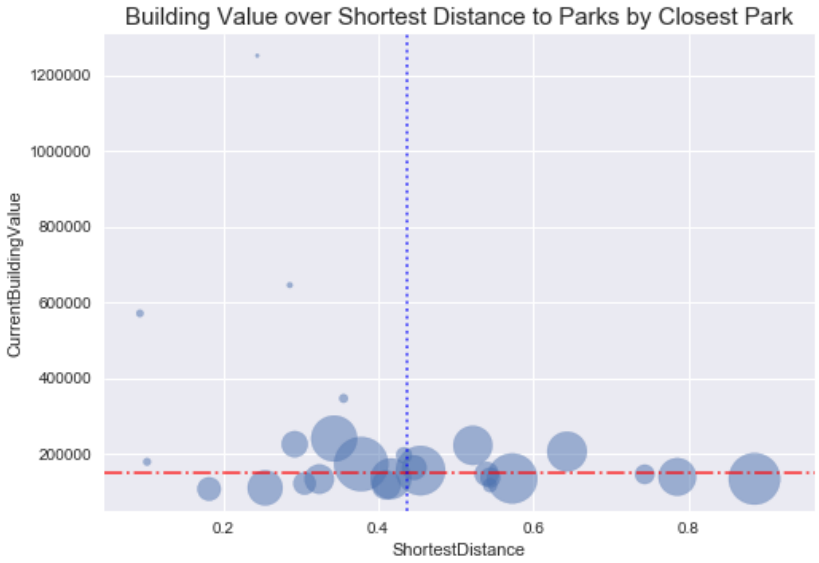

Figure 6: each dot represents a park. The coordinate of the dot represents the median shortest distance from buildings nearby to that park and median of the building values near that park. The size of the dots represents the number of buildings that have that park as their closest park. Blue dash line is the median shortest distance, and the red dash line is the median building value.

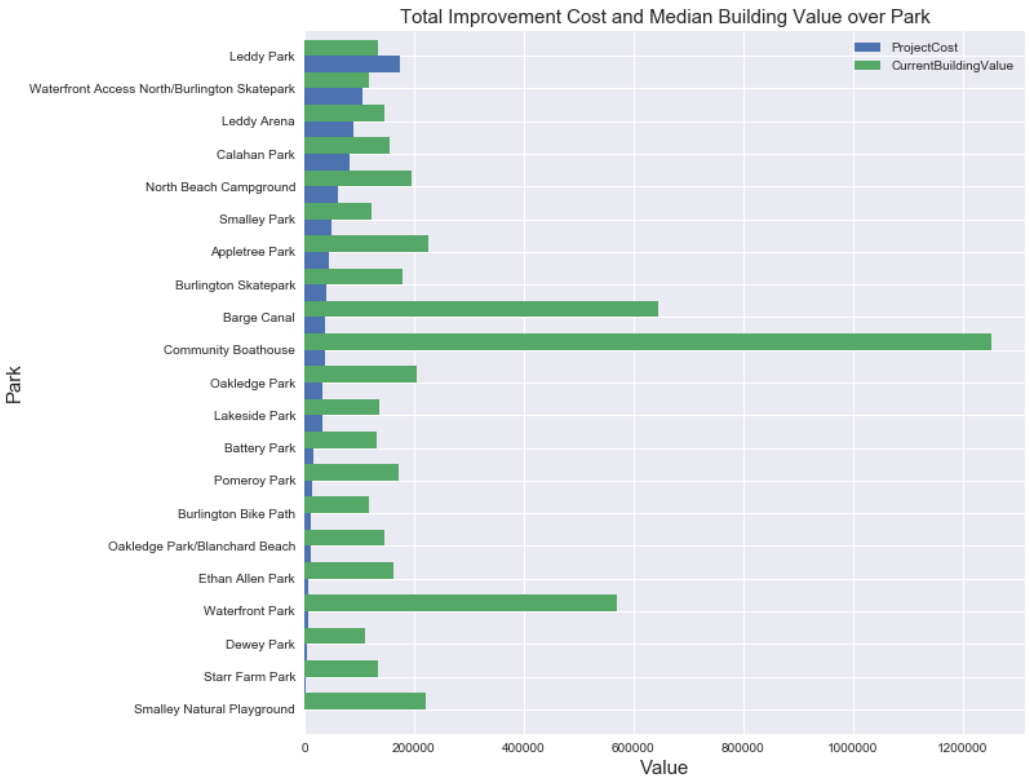

Figure 7: Building Value over the improvement cost by park