Derrick Priebe

Pierre Augustamar

Matthew Peters

INFX 575 Project Follow Up

**Part 1:**

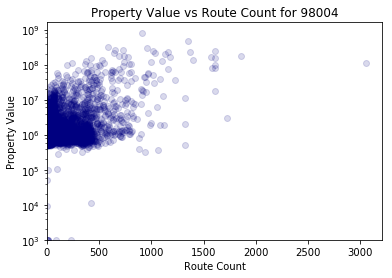
***“I'd like to see the effect for two different regions / tracts. One expensive, one less expensive. Say, medina and renton. You can just show the plot of route count vs. log(price) for Medina and route count vs. log(price) for Renton and then discuss the qualitative results.”***

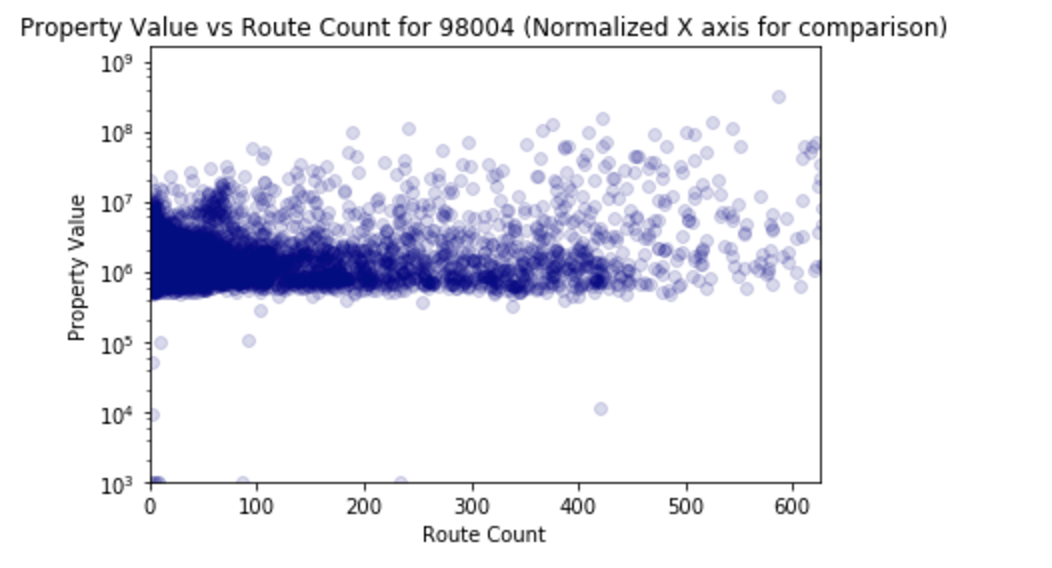
For region, we decided to use zip code. In order to select one expensive and less expensive, we consolidated the data into zip code (grouped), count and median price. We then applied a sort and limited the selection to zip codes with at least 5,000 properties. This led us to 98004 (West Bellevue area) as the representative for high price and 98168 (South Seattle / SeaTac / Tukwila area) as representative for the low price.

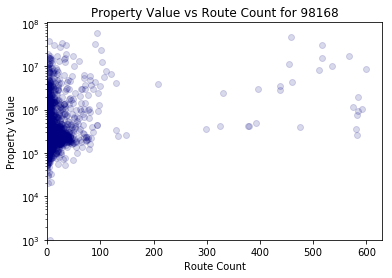
|  |  |  |  |
| --- | --- | --- | --- |
|  | **zip5** | **count** | **median** |
| **3** | 98004.0 | 6763 | 1345000.0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **zip5** | **count** | **median** |
| **90** | 98168.0 | 8946 | 239000.0 |

The route count and property value information was consolidated into a plot for both zip codes. A qualitative analysis shows some pretty interesting results. First, it does not appear that for a majority of the points that property values increase significantly as route count increases. In fact, in each zip code the lower ranges of housing prices appear to have the greatest preponderance of routes. Second, it does appear that the higher property value zip code has a greater distribution of properties overall that intersect routes than the lower value zip code.



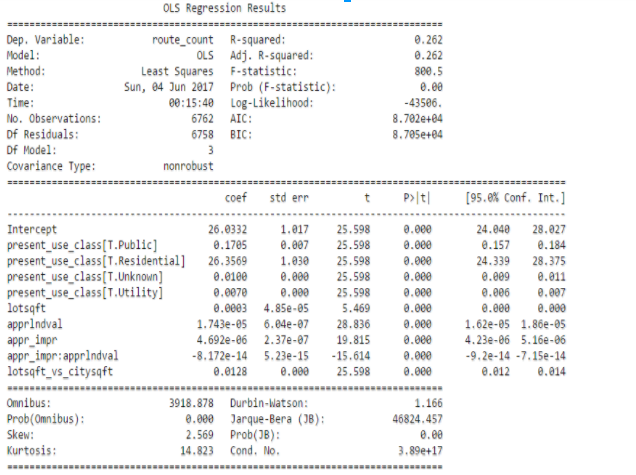




**Part 2:**

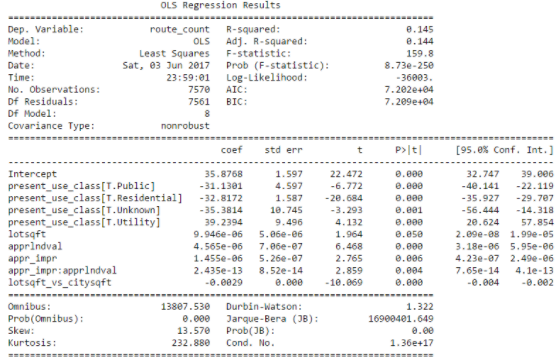
***“Re-run the regression for these two neighborhoods to discuss the error, etc. These items will help deal with the bias in data acquisition (since you can hopefully select neighborhoods for which you have good data) as well as the effects of location (since you are selecting representative neighborhoods deliberately.)”***

**Analysis for 98004 (West Bellevue area) about the route\_count.**



Key analysis:

* R-Squared: With the entire location the fit covered 5.8% of the total variation while with zip code 98004 the fit covers 26.2% of the total variation.
* P-Value for this single location matches the entire location as all coefficients are under
* Coefficients: When using all the locations, three coefficients negatively impacted the response while with this zip code, 98004, only one coefficient caused the dependent variable to decrease.
* Intercept: The intercept has increased from 11.2978 for all locations to 26.0332 for locations that are associated with zip code: 98004.

**Analysis for 98168 (Tukwila area) in relation to the route\_count** 

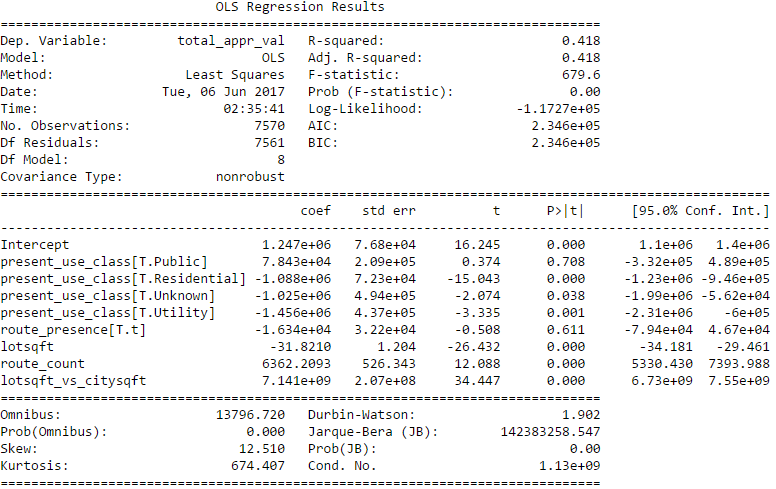
* R-Squared: With the entire location the fit covered 5.8% of the total variation while with zip code 98168 the fit covers 14.5% of the total variation.
* P-Value for this single location matches the entire location as all coefficients are under 0.05.
* Coefficients: When using all the locations, three coefficients negatively impacted the response while with this zip code, 98168, four coefficient caused the dependent variable to decrease.
* Intercept: The intercept has increased from 11.2978 for all locations to 35.8768 for locations that are associated with zip code: 98168.

**Analysis for 98168 (Tukwila area) VS 98004 (West Bellevue area) for route\_count.**

* Both linear regressions show that there is a statistically significant relationship between the terms and the response based on the p-value.
* Most Coefficients for locations in 98004 seem to have a positive impact on the response except for one coefficient that has a very minimal negative impact. On the other hand, three coefficients for locations in 98168 have significant negative impact on the response.
* The model for locations in 98004 have a better sample for analysis at 26.2% compare to locations in 98168 which have 14.5% fit of the total variation.
* The intercept shows that it takes on average 35.8768 unit for running routes in locations in 98168 to see any significant changes while it takes on average 26.0332 unit for running routes in locations in 98004. Thus, changes in running routes will occur at a faster rate for locations in 98004 as opposed to 98168.
* There is stronger relationship or correlation for running routes that are located in 98004 than for locations in 98168 for the same coefficients when comparing the number of coefficients that have a positive impact on the response.

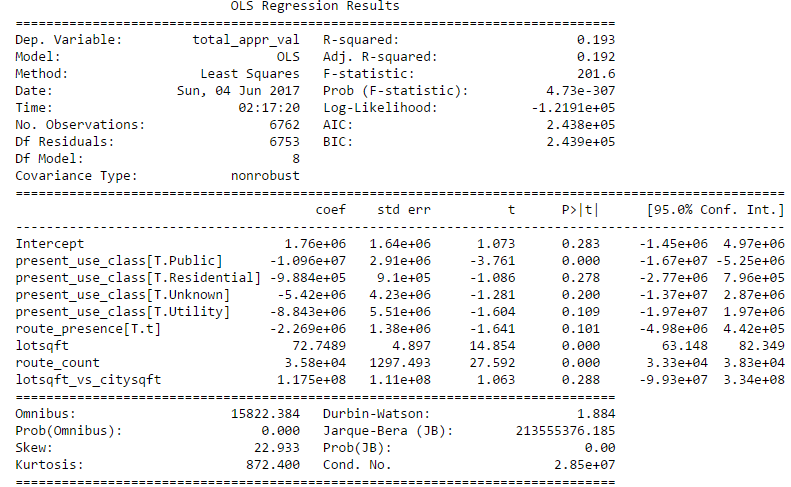
There is stronger relationship or correlation for running routes that are located in 98004 than for locations in 98168 for the same coefficients when comparing the number of coefficients that have a positive impact on the response.

**Analysis for 98168 (Tukwila area) about total\_appr\_val**



* R-Squared: With the entire location the fit covered at .04 of the total variation of the data while with zip code 98168 the fit is now at 0.418 of the total variation.
* P-value: With all the locations the p-values for each of the coefficients were under 0.05. However, when using locations associated with zip code 98168 two components are over the 0.05 threshold.
* Slope: The intercept increased from 585,275 for all locations to 124,7000 for locations in 98168.

**Analysis for 98004 (West Bellevue area) about total\_appr\_val**



* R-Squared: With the entire population set the fit covered at .04 of the total variation of the data while with zip code 98004, the fit is now at 0.193 of the total variation.
* P-value: With all the locations the p-values for each of the coefficients were under 0.05. However, when using locations associated with zip code 98004, six components including the intercept are over the threshold.
* Slope: the intercept increased from 585,275 for all locations to 176,0000 for locations in 98004

**Analysis: 98168 (Tukwila area) VS 98004 (West Bellevue area for total\_appr\_val.**

* Locations in 98004 only have three coefficients out of nine that are statistically significant while locations in 98168 have seven coefficients that are statistically significant.
* Both locations in 98004 and 98168 only have three coefficients out of nine that have a positive impact on the total\_appr\_val.
* Locations in 98004 have an R-squared of 0.193 while locations in 98168 have a larger variation with an R-squared of 0.418.

**Overall analysis of selected individual locations against the entire locations**

In the case of total\_appr\_val, it seems that using all locations show a better relationship or correlation between the intercept and the coefficients as opposed to using the locations for both zip code in 98004 and 98168. We came to this conclusion mainly when comparing the number of coefficients that are statistically significant for individual locations and those that are associated with the entire locations.

In the case of route\_count, it seems that using locations for 98004 appear to show a better relationship or correlation between the intercept and the coefficients as opposed to using the entire locations and locations in 98168. We arrive at this conclusion based on the fact that there was only one coefficient that has a minimal negative impact on the response for 98004.

Finally, for all four individual analysis, we can say that the relationship or correlation between the response and the terms is a lot stronger for the locations associated with zip code 98004 than when using all locations or locations related to zip code 98168