An Analysis of the Association Between Walkability and Housing Factors

Scott Chase

April 2016

1 Introduction

As part of my work in the Social, Spatial, and Dynamic Analysis lab and the Undergraduate Research Opportunities Program at the University of Minnesota, I have contributed to the development of two R functions to extract data from Zillow's API and Walkscore's API. [1] While the details of how these functions will not be discussed in this paper, they will be published in an upcoming R package. Zillow is a real estate information company that provides an API for publically available information about houses, such as the number of bedrooms, the date last sold, and the estimated value of the house. [7] Walkscore is a company that estimates a number $\in$ [0, 100] for a given address, that measures the ability to walk to daily destinations. [5] Walkscores closer to 0 are more dependent on a car, while a walkscore of closer to 100 indicates the ability to walk just about anywhere that may be needed. Walkscore also estimates two quantities called the Bike Score and the Transit Score, which measure the ability to bike to nearby destinations and the availability of public transit.

Originally, this research was supposed to include measures of public health. The data that would have related an original address to several measures of public health proved to be unavailable, and thus I cannot include public health as a component. As such, this research will focus on housing factors and walkability. A significant portion of the work on this project included coding API functions for both Zillow and Walkscore to even gather the data in the first place.

2 Data

I have a list of addresses for the city of Portland, Oregon. While I will not publish the addresses or data due to privacy reasons [2] and compliance with the Zillow API terms of agreement, the address, city, and zip code (as well as a Walkscore and Zillow API key) are needed for work with the API. I will suppress any output that includes a specific address, and assign each address a unique ID number for internal use in this analysis. This internal ID number will be in no way linked to anything meaningful associated with Zillow or Walkscore; it will be used as an arbitrary primary key. Data about the house such as year built, number of bedrooms, walkscore, and estimated dollar value will be used in analysis. While there were originally nearly 800,000 houses in the data set, due to the fact that Walkscore only allows for 5000 API calls per day, and Zillow only allows for 1000 API calls per day, I have randomly selected 3000 houses from this data set. This also assists with computational matters; it is easier to make computations with 3000 observations than 800,000.

[1] I want to thank Professor Zack Almquist for being a great mentor and the UROP Office at the University of Minnesota for sponsoring my research!

[2] The same reason they always put 555 in phone numbers in movies, essentially

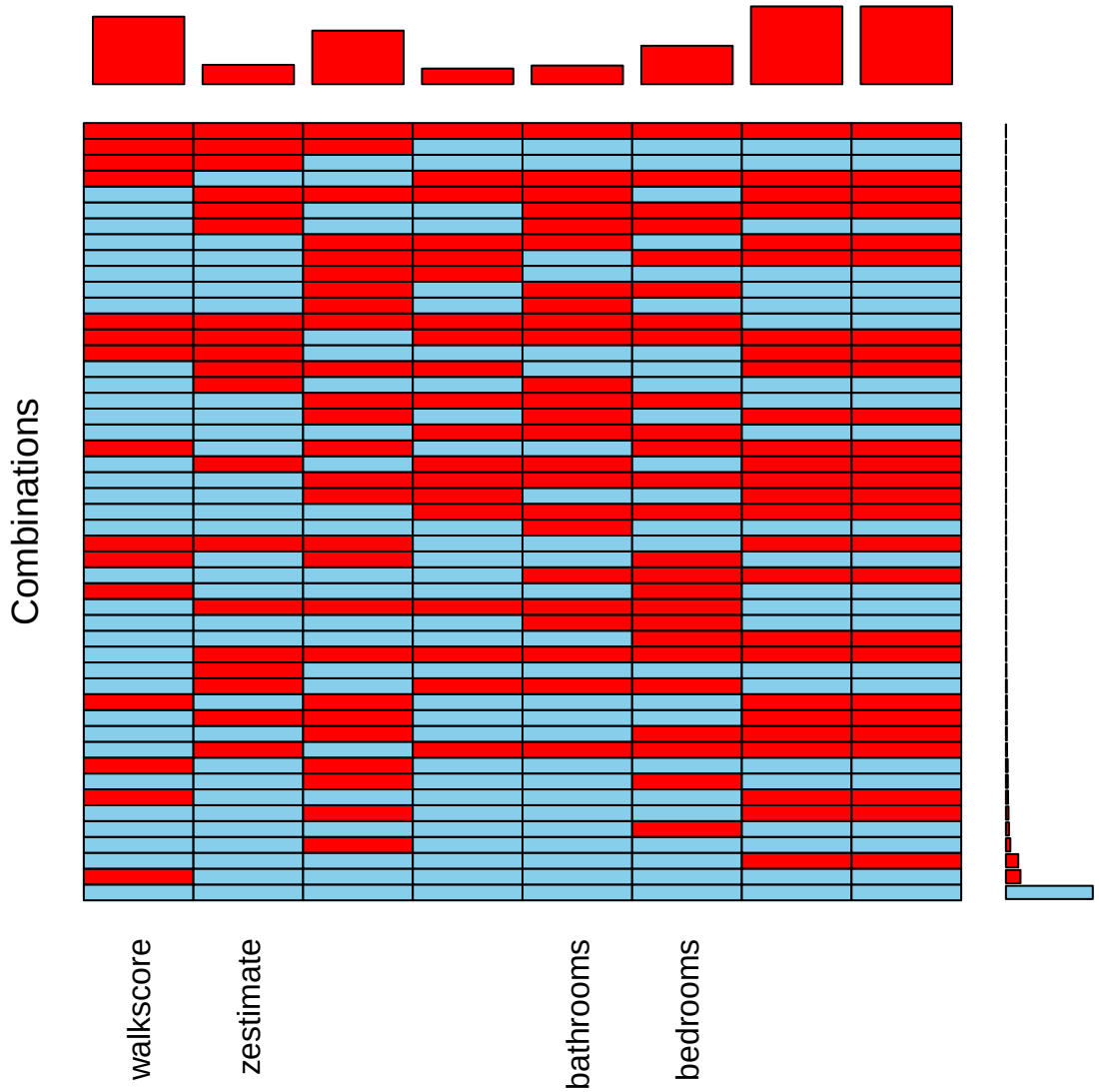
3 Portland, Oregon

The houses in the data set encompass three counties within and around the city of Portland Oregon. Fips 41005 is Clackamas County, fips 41051 is Multnomah County, and fips 41067 is Washington County. While county boundaries in the United States are arbitrary administrative lines, sometimes they can be a sort of proxy-classification for a meaningful neighborhood. For this reason, I will color the points in each scatterplot based on county throughout this document. If there appears to be a meaningful pattern, great! If not, then we might have evidence that the county lines do not define a meaningful classification here.

4 Missing Data

In some cases, there was no information about the house. k-Nearest Neighbor imputation requires that we at least have *some* information about the house to fill in potentially missing values, as it is basically attempting to fill in values based on what similar observations exhibit. The houses for which we have no information whatsoever are marked by a missing id number. I'll remove these. There are 1914 observations left. While this might be potentially concerning as to missing data, there isn't anything else we can do here.

One useful method for assessing whether data are missing at random vs not at random is a matrixplot for missing data. This is implemented through the `VIM` package using the function `aggr`. [3] The red cells indicate missing values in combinations of variables. There does not appear to be a consistent pattern in the missingness of the data, suggesting that they are missing at random (as opposed to missing not-at-random, see [2] for further discussion) and that multiple imputation is appropriate. k-Nearest Neighbor imputation is done through the `kNN` function, also in the `VIM` package.




```
## Time difference of 3.711335 secs
```

At this point, we have kNN-imputed data for the following variables:

- `walkscore`. The walkscore of the particular address, taking values $\in [0, 100]$. Walkscores closer to 0 imply heavier reliance on a car, while walkscores closer to 100 imply the ability to walk to everyday destinations.
- `zestimate`. The Zillow-estimated value (USD) of the house.
- `fips`. The county and state FIPS code the house is located in.
- `yearBuilt`. The year the house was built.
- `lotSizeSqFt`. The square footage of the lot size.

- `finishedSqFt`. The square footage of the house.
- `bathrooms`. The number of bathrooms in the house.
- `bedrooms`. The number of bedrooms in the house.
- `lastSoldPrice`. The price (USD) that the house was last sold at.

5 Analysis

5.1 Ordinary Linear Regression

The primary analysis question outlined in this report is how the walkscore is associated with different housing measures. For instance, do houses that are valued at higher rates tend to be more walkable? I will use the standard linear regression for this data for ease of interpretation. I will use backward selection (via AIC). First I will fit the full model, then use backward selection to arrive at a smaller model. I will then compare the two via ANOVA, and assess the assumptions of the chosen model.

	Estimate	Std. Error	t value	Pr(>|t|)
fips41005	483.1505	36.3024	13.31	0.0000
fips41051	511.7854	35.9657	14.23	0.0000
fips41067	495.8776	36.4424	13.61	0.0000
zestimate	0.0000	0.0000	0.40	0.6859
yearBuilt	-0.2249	0.0184	-12.22	0.0000
lotSizeSqFt	0.0000	0.0000	0.48	0.6299
finishedSqFt	0.0000	0.0001	0.27	0.7896
bathrooms	0.3933	0.3655	1.08	0.2821
bedrooms	-4.0933	0.5006	-8.18	0.0000
lastSoldPrice	-0.0000	0.0000	-0.12	0.9016

Table 1: Full Model

	Estimate	Std. Error	t value	Pr(>|t|)
fips41005	476.1912	35.1832	13.53	0.0000
fips41051	504.9138	34.8450	14.49	0.0000
fips41067	488.9154	35.3429	13.83	0.0000
yearBuilt	-0.2211	0.0178	-12.41	0.0000
bedrooms	-3.8520	0.4700	-8.20	0.0000

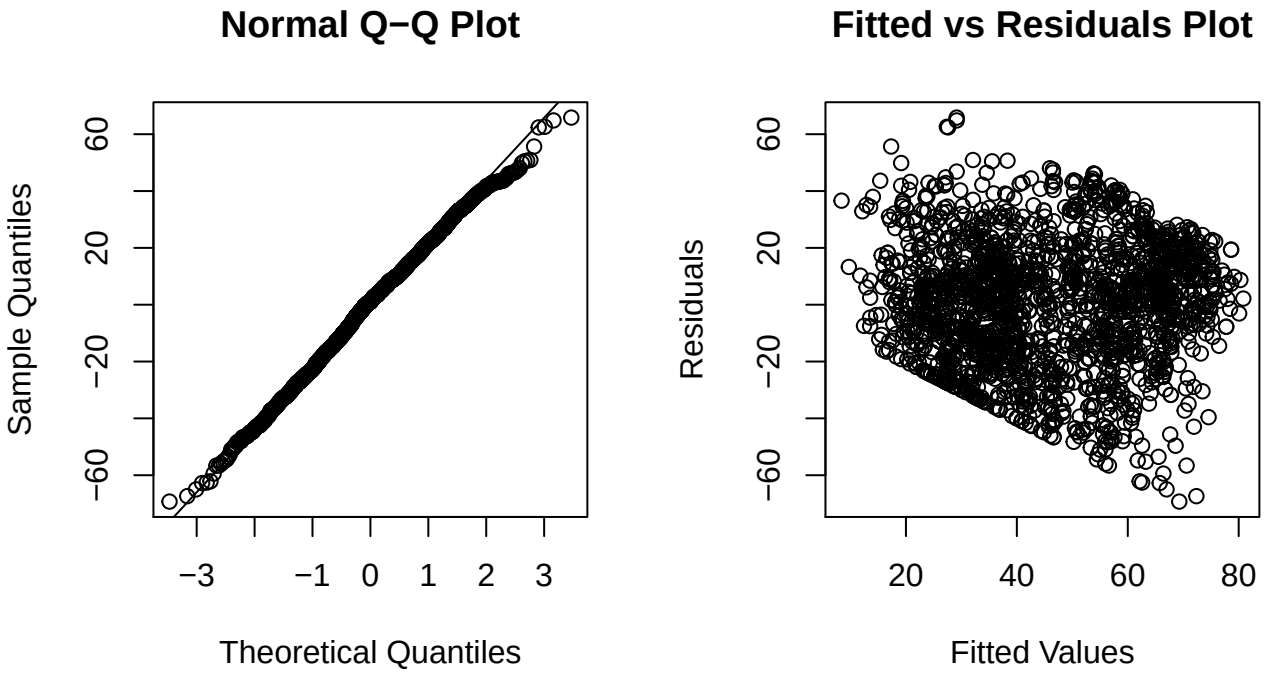
Table 2: Smaller Model chosen by Backward AIC Selection

It seems that the FIPS code, the year built, and the number of bedrooms are the most appropriate for predicting the walkscore of a particular address, as measured by statistical significance and the AIC. Each of these coefficients are additive; there is a different base walkscore estimate depending on which county the house is in, more recent houses tend to be less walkable, and houses with larger numbers of bedrooms also tend to be less walkable.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1909	876754.86				
2	1904	875747.12	5	1007.73	0.44	0.8221

The larger model does not appear to fit any better than the smaller model. For this reason, I will use the smaller model for simplicity. This also suggests that the estimated valuation of the house does not seem to significantly correlate with the walkscore. For a visualization, I will graph the Zillow estimated valuation of the house. Most visualizations throughout this document will use the R package `ggplot2`. [6]

Standard least squares regression assumes normally distributed residuals and a constant variance. As we see below, the normality assumption is justified, and at initial glance we might suspect that the variance is not constant. If the residual is defined as $y - \hat{y}$, and the fitted value is 20, then the smallest negative number the residual could take is -20, since we can't observe a walkscore smaller than 0. This is why the line in the lower left (and the upper right, for that matter) is there. The response is bounded, and the residual is too. This does not infer a systematic problem in the model. But in any case, we should try a generalized linear model. The walkscore is probably more closely associated with count data, and I'll model it with a log link. I will use the `glmbb` package to fit all hierarchical generalized linear models, and select the best one based on AIC. [1]



5.2 Poisson Generalized Model with Log Link

With an AIC of 34370, the `glmbb` suggests a model of `walkscore` regressed on `fips + yearBuilt + bathrooms + bedrooms`. This is what I will fit.

	Estimate	Std. Error	z value	Pr(>|z|)
fips41005	12.3953	0.2311	53.64	0.0000
fips41051	13.0986	0.2286	57.29	0.0000
fips41067	12.7702	0.2322	55.01	0.0000
yearBuilt	-0.0045	0.0001	-38.19	0.0000
bathrooms	0.0072	0.0021	3.48	0.0005
bedrooms	-0.0888	0.0034	-26.06	0.0000

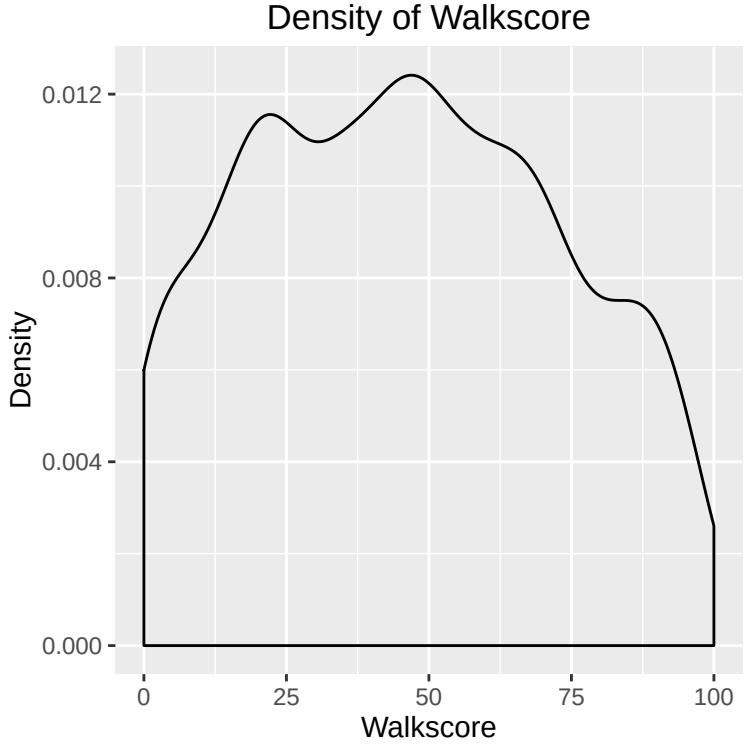
Table 3: Generalized Linear Model, Poisson Log Link

fips41005	fips41051	fips41067	yearBuilt	Bathrooms	Bedrooms
241668	488245	351591	0.9955	1.0072	0.9151

Table 4: Coefficients of GLM, raised to the power of e

The exponentiated coefficients of the generalized linear model are reported above. Note that these are all multiplicative effects: the coefficients larger than one are positively associated with the response, while the coefficients lower than one are negatively associated with the response.

Generalized linear models assume that the response is Poisson distributed, that the data are not overdispersed, and that the variance of the residuals is constant. Let's assess those. The variance of the walkscore is 726, while the mean is about 45. This suggests that we have overdispersion. I will use the negative binomial distribution to model this, implemented through the `MASS` package. [4]



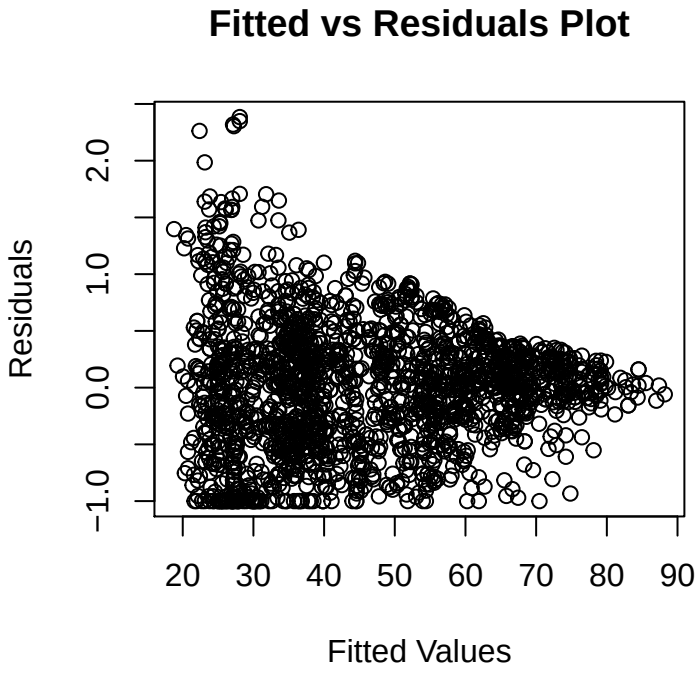
5.3 Negative Binomial Regression

The negative binomial generalized linear model introduces an additional parameter θ to deal with the overdispersion that we seem to be finding in the walkscore data. I will fit the model with the same coefficients as the poisson generalized model (just dropping the bathrooms parameter - after I initially fit the model, it was not statistically significant).

	Estimate	Std. Error	z value	Pr(>|z|)
fips41005	12.0673	1.0965	11.01	0.0000
fips41051	12.7655	1.0860	11.76	0.0000
fips41067	12.4339	1.1015	11.29	0.0000
yearBuilt	-0.0043	0.0006	-7.78	0.0000
bedrooms	-0.0746	0.0147	-5.07	0.0000

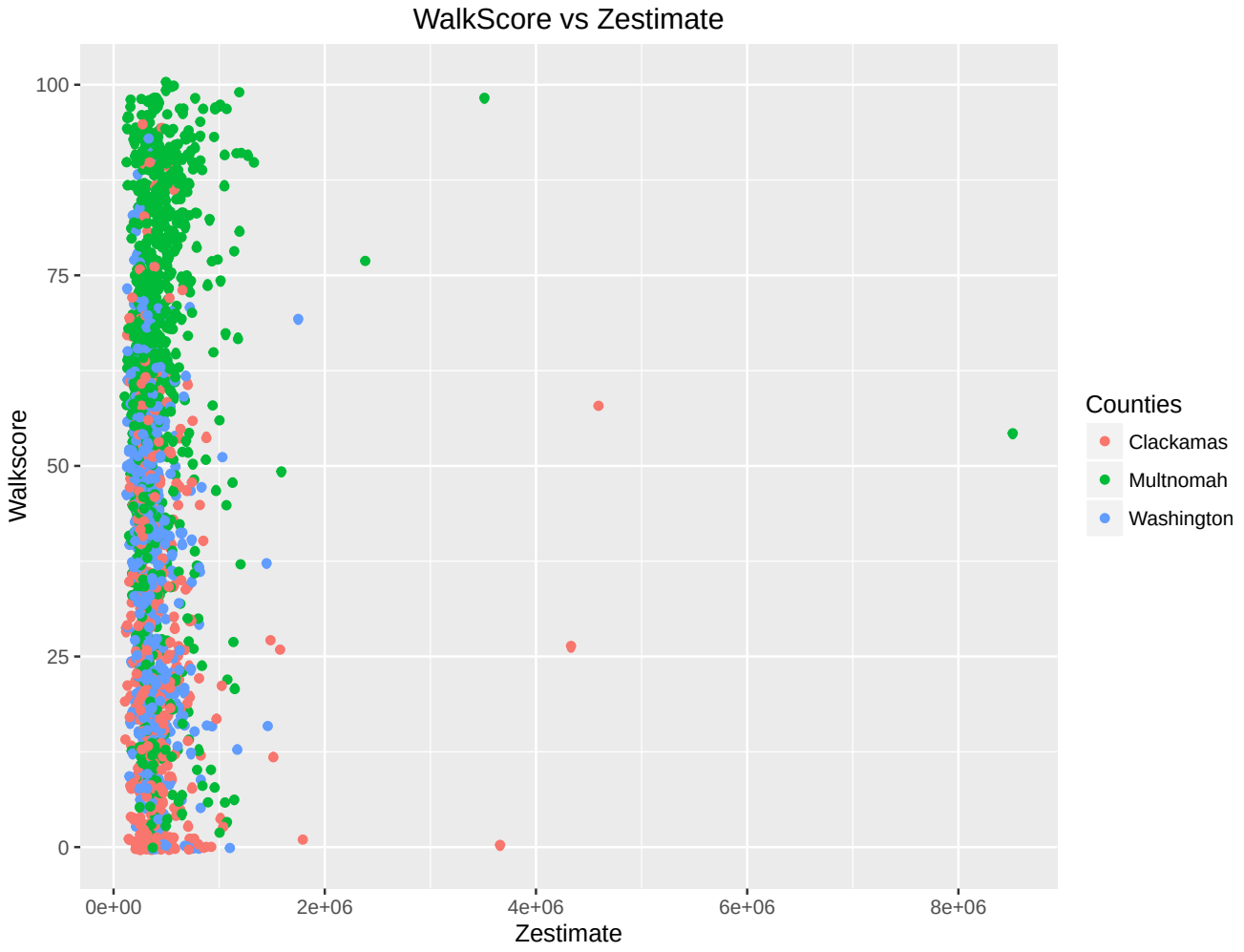
Table 5: Negative Binomial Regression

The dispersion parameter θ was reported to be 2.3444, with a standard error 0.0844. This suggests that we have overdispersion, but that this is being properly dealt with. Now, all we have to assume is that the residuals have a constant variance.



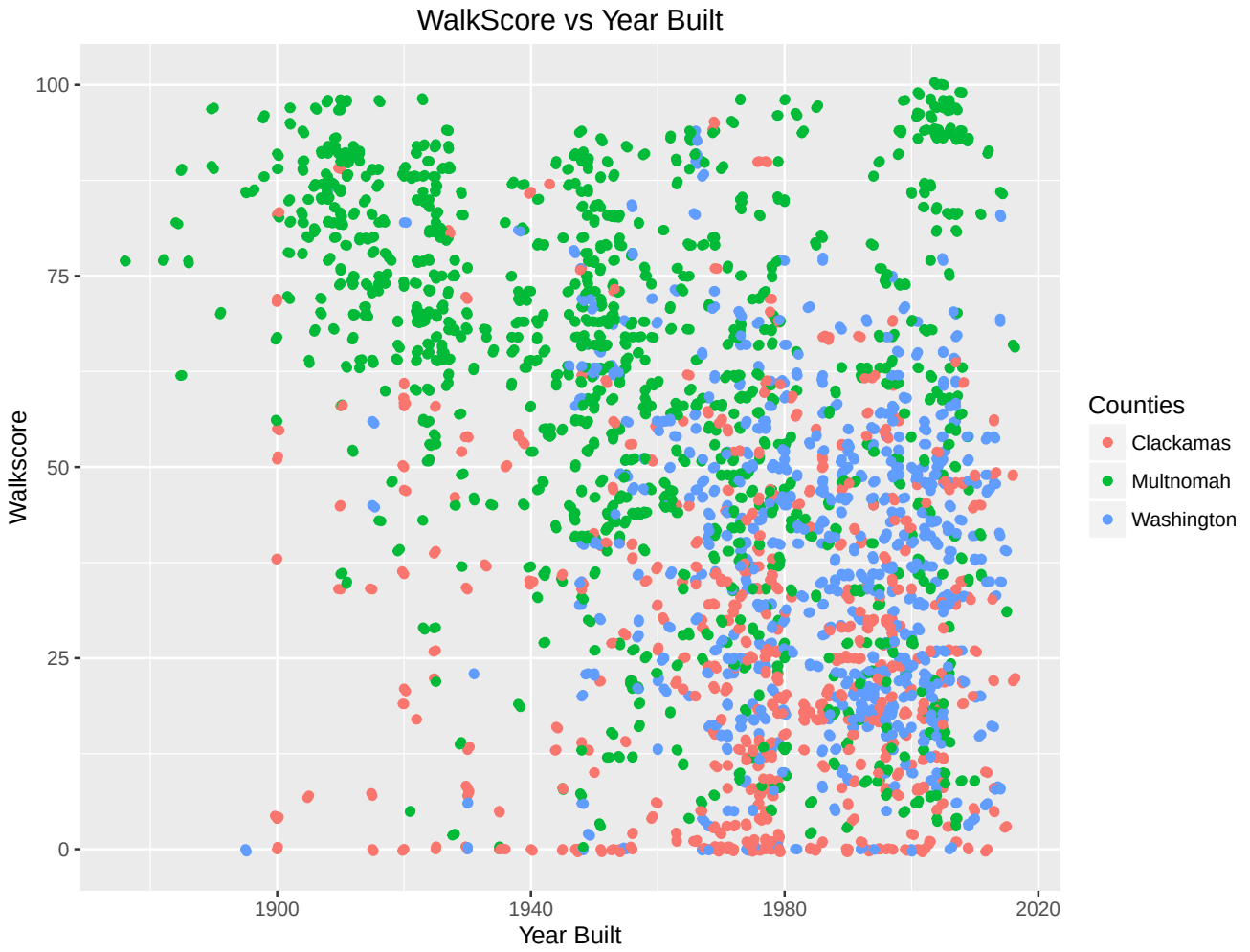
While it might look like the residuals are not constant, I suspect this is really due to the fact that the response variable is bounded $\in$ [0,100]. We see a fairly distinct decreasing line of residuals in the upper right portion of the graph. If the predicted walkscore is, say 90, the largest positive value the residual (defined by $y - \hat{y}$) could take is 10. We don't see quite as much of a distinctive line as we did in the linear model, but that is because the generalized linear model is multiplicative.

5.4 Zestimate



With the exception of a few outliers, the plot is almost a vertical line. This suggests that the estimated valuation of the house and the walkscore are almost independent. There also isn't a clear pattern between the estimated value and the county, suggesting that the houses are of similar value throughout Portland. There does, however, seem to be a pattern between walkscore and the county. Most of the green points (Multnomah) are of high walkability, while the middle blue points (Washington) are of moderate walkability. The pattern corresponding to Clackamas county is less clear.

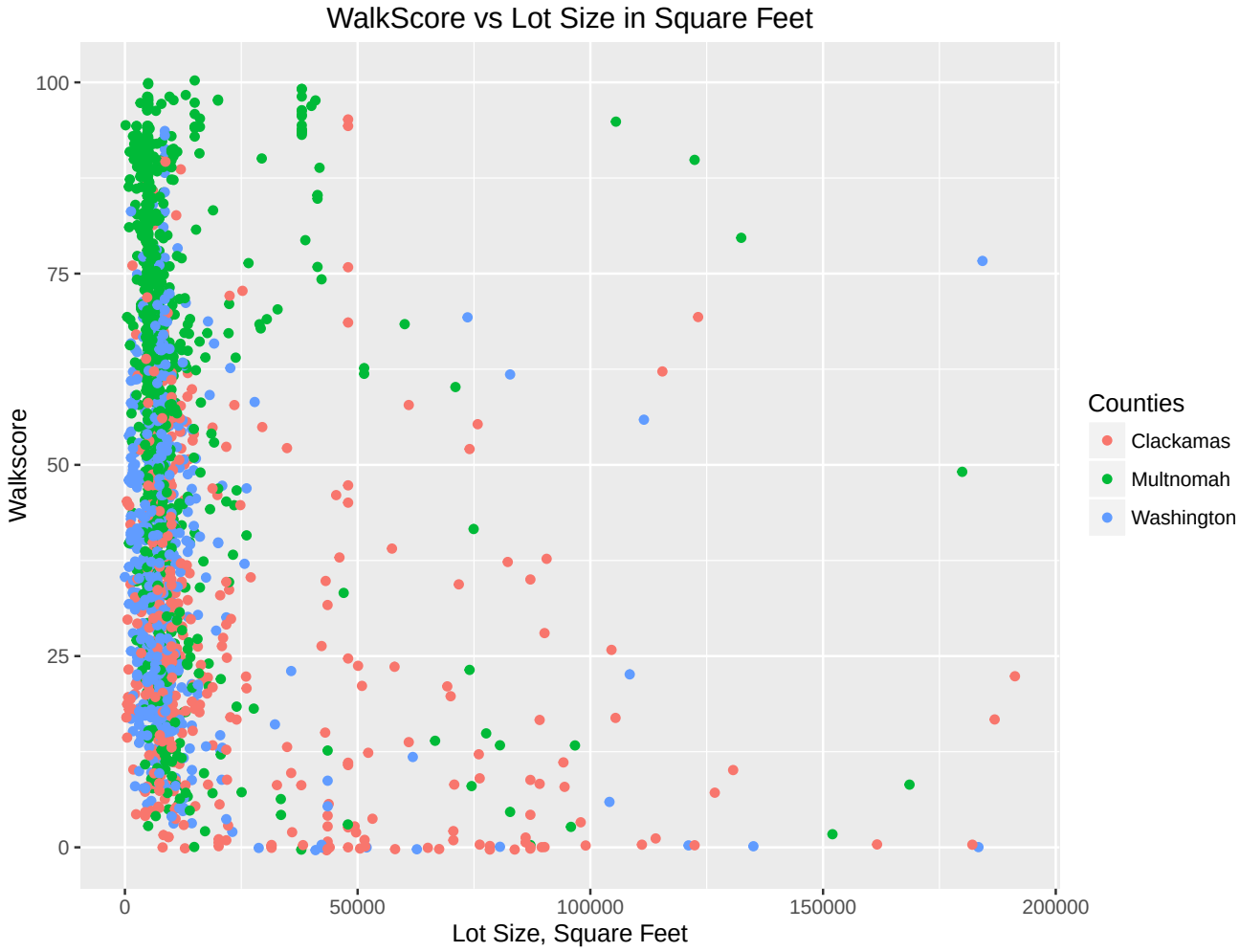
5.5 Year Built



Interestingly, there do not appear to be a lot of houses that were built before 1940 with walkscores under 40. There does, however, appear to be a larger number of houses built before 1940, but with a walkscore of over 50. This suggests that neighborhoods that originally encapsulated older houses are becoming more walkable after time. It doesn't appear that there just aren't that many houses still standing built before 1940, as there are a sizable amount of houses built before 1940 and with a higher walkscore. By examining the houses built after 1950, there appear to be an equal number of low and high walkscore houses. I think that this suggests that houses, even as far back as 1950, are being built with equal frequency in low and high walkscore areas. Perhaps World War 2 was the turning point with the development of the suburbs, with the expansion of the suburbs allowing for more houses to be less walkable.

We also see that the houses in Washington county seem to have been built after World War 2, while many of the houses in Multnomah county seem to have been built before World War 2. This suggests that Washington county has many houses that were part of the suburbian expansion in the mid 20th century in the United States.

5.6 Lot Size

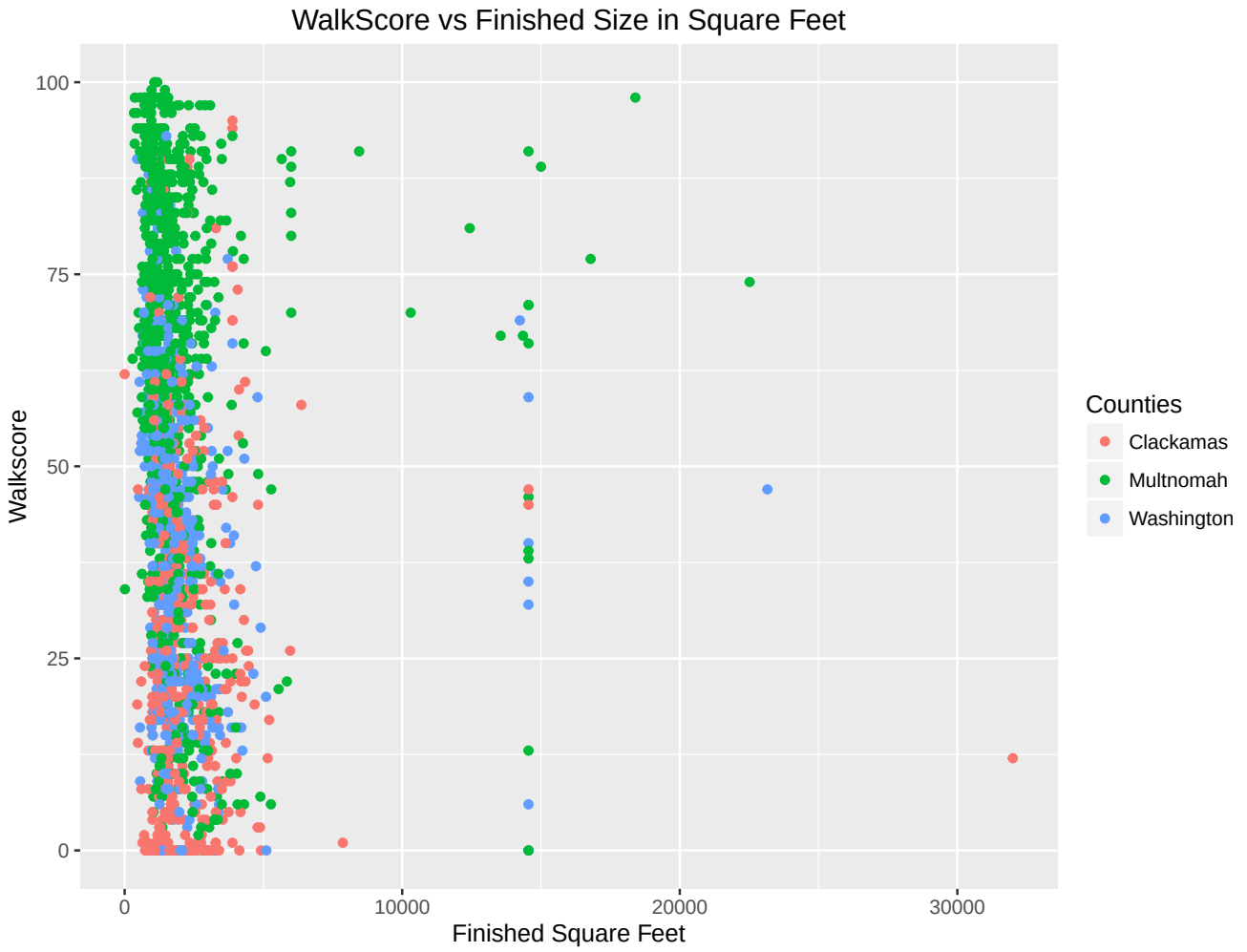


It does not appear that there is a significant pattern between the lot size in square feet and the walkscore. [3] Most of the lot sizes in square feet are under 25,000 square feet and encompass all levels of walkability. While it is a very slight pattern (due to the smaller number of data points), considering the lot sizes larger than about 25,000 square feet, there are fewer and fewer high walkability houses with significantly larger lots. My intuition is that the larger lot sizes are in more rural areas, which are less walkable.

In terms of the county lines, it appears that the blue dots corresponding to Washington county have low to medium walkscores, while the green dots corresponding to Multnomah county are highly walkable. There doesn't appear to be a large difference between counties based on the lot size. The differences in walkscore based on county is probably more a function of location, however. Multnomah county might just be more centrally located to where people's day-to-day destinations are

5.7 Finished Square Feet

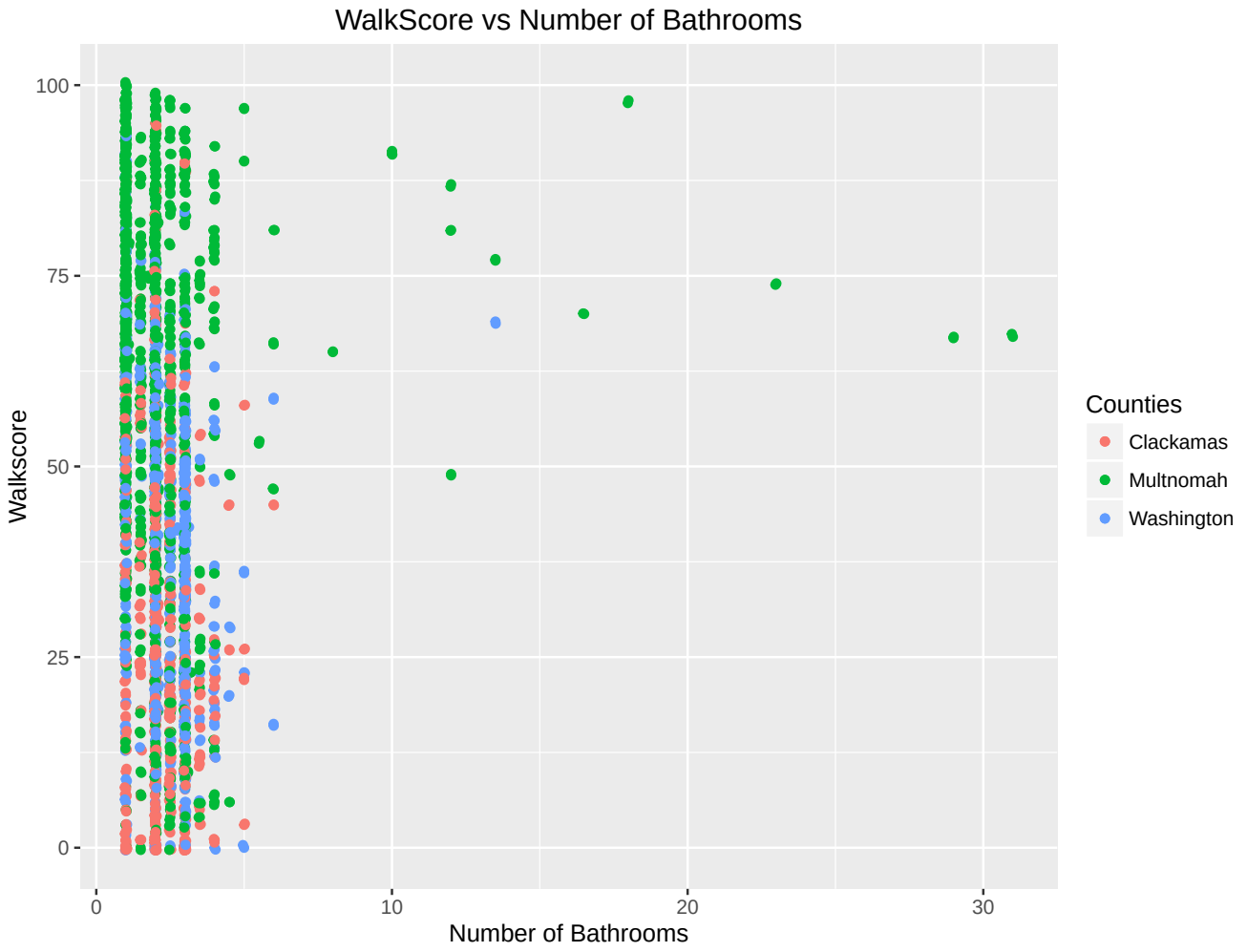
[3]Note that in creating this plot, I only included the observations with lot sizes below 200,000 square feet, which is over 96% of houses I did this as there were a few outliers that were causing everything to be clustered on the left side of the plot and it allows us to easier see the pattern. The outliers did not appear to be associated with high or low walkability.



There does not appear to be any sort of pattern between the size of the house in square feet and the walkability. [4] We do see a similar pattern with walkscore, though. The most walkable houses appear to be in Multnomah county, and the low-to-middle walkable houses appear to be in Washington county.

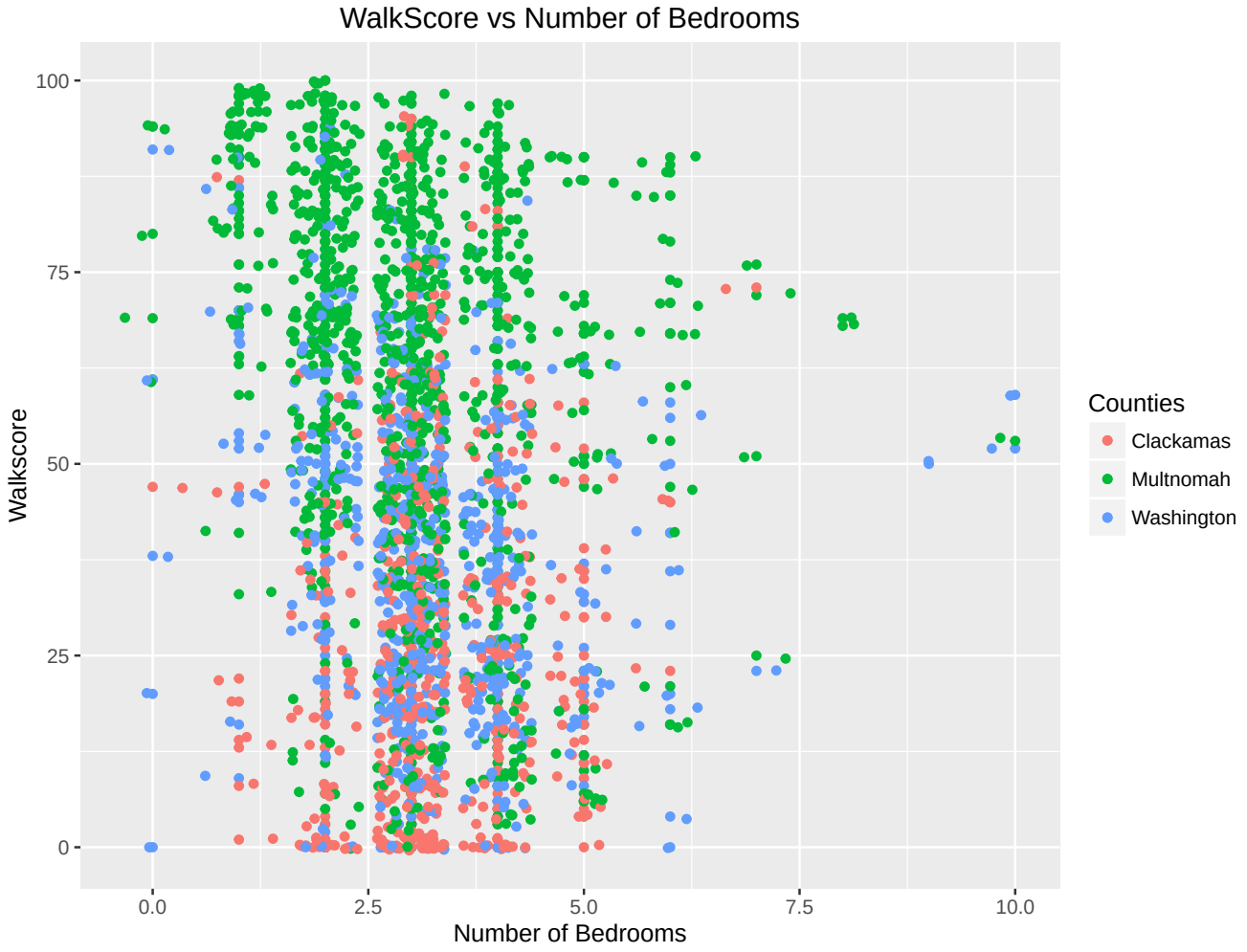
5.8 Number of Bathrooms

[4]Note that I excluded a small number of outliers from the walkscore vs finished size in square feet as well to make the plot less cluttered. The data plotted encompass over 99% of observations.



I'm guessing that the addresses with at least 10 bathrooms are apartment complexes. They can probably be ignored. Judging by how the houses with under 5 bathrooms all have every level of Walkscore, there doesn't appear to be a pattern. In hindsight, this is not surprising. You can fit plenty of bathrooms in a house, but that doesn't mean it is near anything! Again, we see the same repeated pattern with the walkscore and the county. It doesn't appear to be associated with the number of bathrooms.

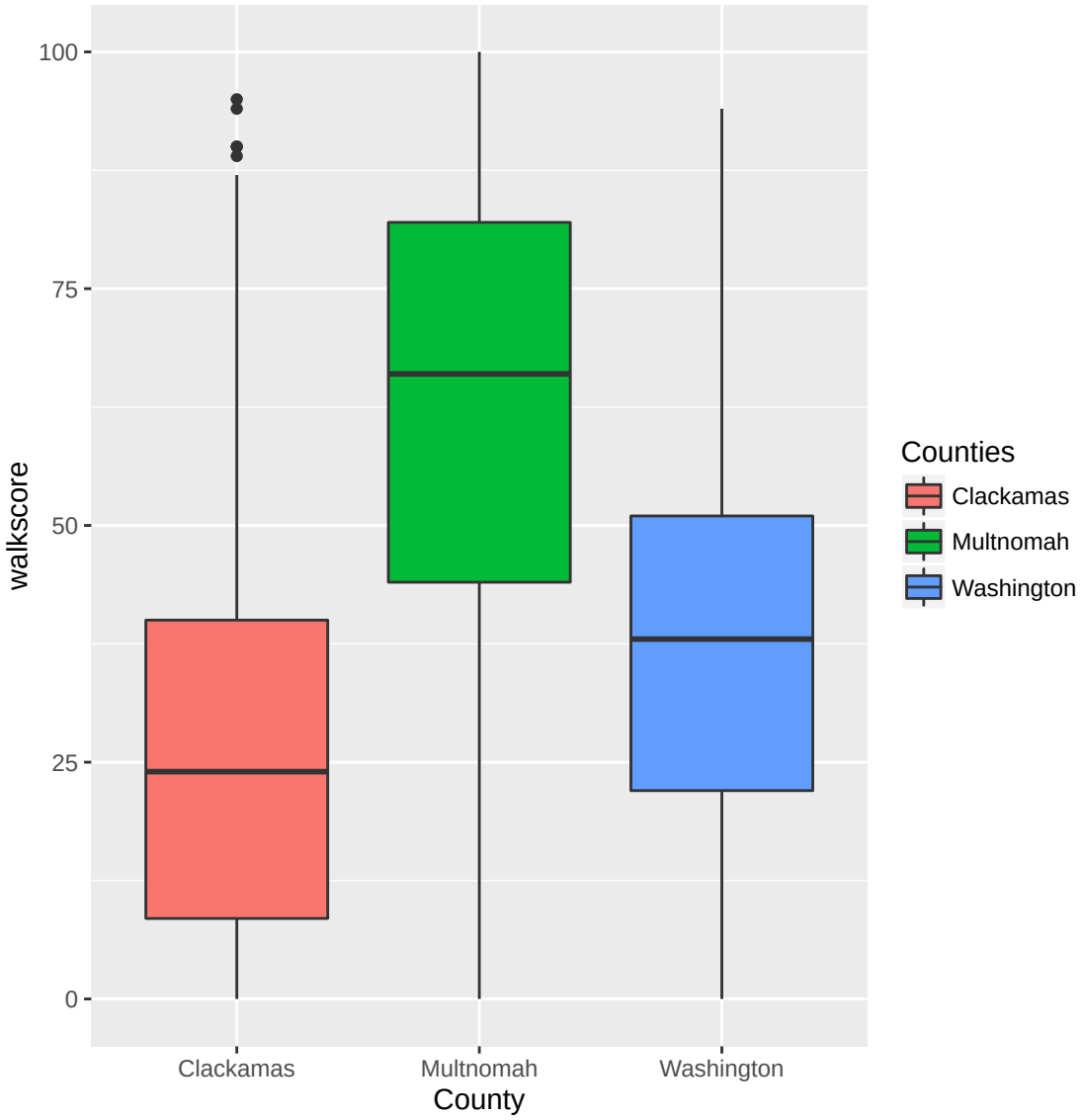
5.9 Number of Bedrooms



Similarly to bathrooms, you can have just about any number of bedrooms in a house and that doesn't appear to imply a particular walkscore value. There does not appear to be an association between the number of bedrooms and the county. While this contradicts the regression model interpretation, note that this is just a look at the pattern between the number of bedrooms and the walkscore. This does not take anything else into account.

6 Statistical Tests

Throughout this document I referenced the fact that the walkscore appeared to be different based on the county. I want to dig a little more into that.

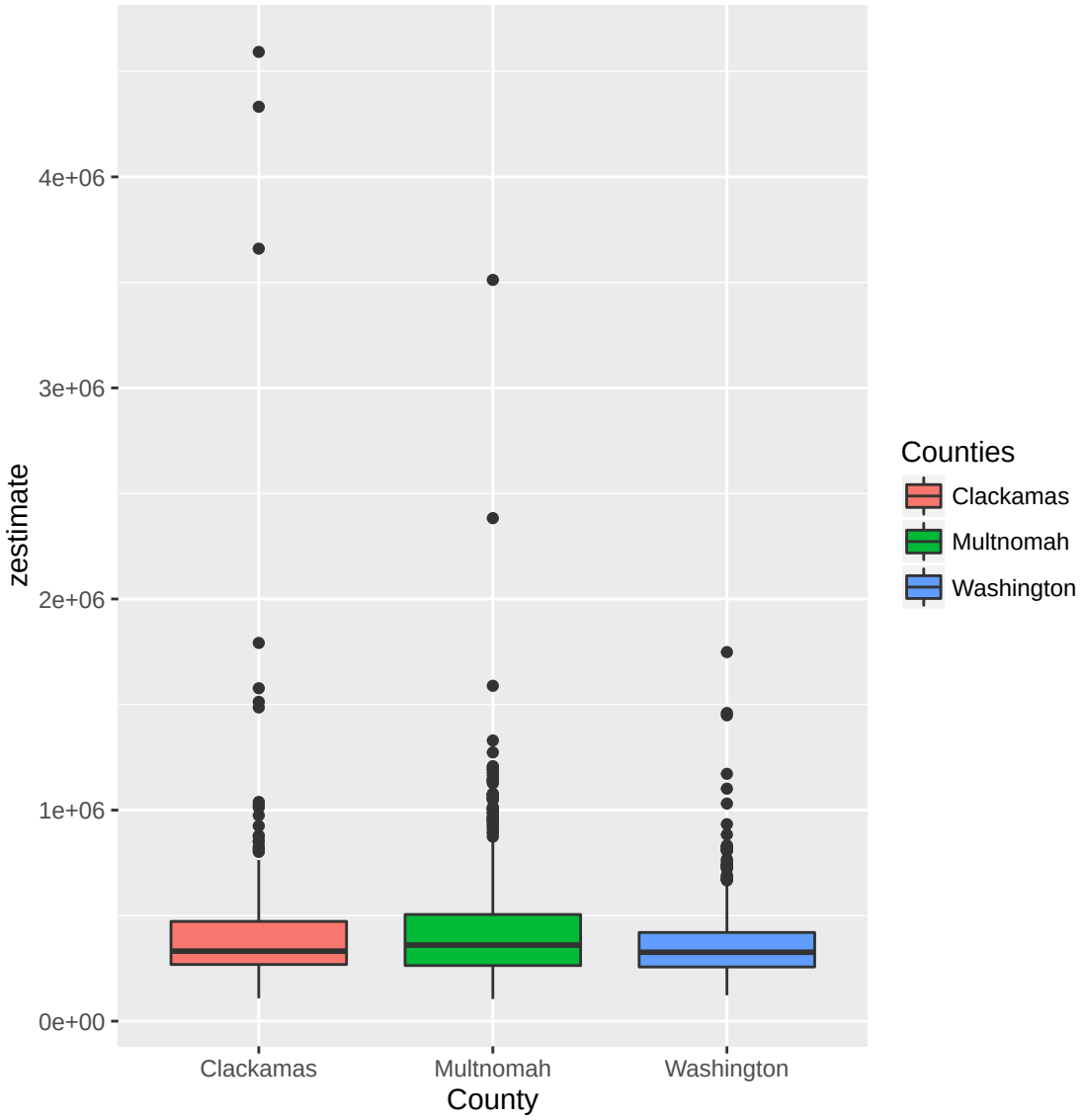


I suspect that there is a difference in Walkscore (on average) between these counties. To formally test this, I will use a one-way ANOVA test. We probably don't need normality as the sample size is *fairly* large.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
County	2	402950.52	201475.26	390.40	0.0000
Residuals	1911	986214.70	516.07		

There is evidence that the average walkscore is different across the different counties. Given the boxplot, this is not surprising. But is the county related to the Zillow-estimated value of the house at all? [5]

[5]Note that there is one drastic outlier in this data. I will remove it as I'm pretty sure Bill Gates lives in Washington; at least Microsoft is headquartered there. Who knows?



Now, let's run the ANOVA tests comparing the zillow-estimated house value across each county (with the top outlier removed).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
County	2	1209245156863.70	604622578431.85	8.82	0.0002
Residuals	1910	130944629025682.72	68557397395.65		

While the p value is small, and normally we'd reject the null hypothesis, I suspect that this is really just due to the outliers. I'm skeptical of the result from this ANOVA test.

7 Conclusion

k-Nearest Neighbors imputation was performed on data gathered from functions built in R to gather data from Walkscore and Zillow, using a subset of addresses in the city of Portland Oregon. According to the

different regression models built, the county, the year built, and the number of bedrooms seemed to be most associated with the Walkscore, and explained most of the variation in the Walkscore as measured by the multiple R^2 value. After visualizing the data, the Walkscore seemed to be unrelated to the Zillow-estimated value of the house, and the number of bedrooms & bathrooms, casting doubt onto the results of the regression model. The year the house was built seemed to reflect the expansion into the suburbs seen in the United States after World War Two. Walkscore differed based on county, but the Zillow-estimated value of the house did not have strong evidence of differing based on county, despite the ANOVA test. Walkability seems to be most related to location and neighborhood.

References

[1] Charles J. Geyer. *glmbb: All Hierarchical Models for Generalized Linear Model*, 2016. R package version 0.1.

[2] Kristin L. Sainani. Dealing with missing data. *American Academy of Physical Medicine and Rehabilitiation*, 7:990–994.

[3] Matthias Templ, Andreas Alfons, Alexander Kowarik, and Bernd Prantner. *VIM: Visualization and Imputation of Missing Values*, 2015. R package version 4.4.1.

[4] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[5] WalkScore. *Walk Score Professional API*, Accessed 2016. Available at `https://www.walkscore.com/professional/walk-score-apis.php`.

[6] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[7] Zillow. *Zillow Real Estate and Mortgate Data for Your Site API*, Accessed 2016. Available at `http://www.zillow.com/howto/api/APIOverview.htm`.