**Location, Location, Location ?!**

Connecting with realtors and the traditional intuition of home pricing

As machine learning is deployed more widely, across many industries, we strive to better understand the stakeholders’ points of view, and help others understand and trust new analysis. In this article I will review some of the ways to think about presenting to stakeholders, in this case realtors, using real estate data.

Realtors have a lot of information available to them, and generally use a few key “rules of thumb” for estimating home pricing:

* Location, being the top metric (which is why it’s repeated 3 times)
* Square footage of the living space
* Number of bedrooms, baths

But “location” means different things to different people: proximity to schools, places of work, friends and family, city entertainment or beautiful scenery, all apply. With the advent of machine learning, we create algorithms which clarify more specifically the important features that increase home values.

Our data set is a modified version of the Seattle competition dataset. More info on this dataset and the cleanup of the ETL of the data can be found here: *github repo*

From the point of view of a non-technical audience who is looking at a report in their area of expertise, there are three main items to include in the story of the analytics:

* Data relevance - Where the data comes from
* Data results - relative to historical perspective
* Data model – other insights

**Data relevance**

When we first show data to someone else, they are going to want to know where the data from and whether it’s relevant to their situation. They may want to know if there were enough datapoints to make a generalization.

The Seattle dataset is based on approximately 21,500 single unit house sales in King County, Washington, from the one year period from May 2014 to May 2015.

The data provided includes 20 different features for each property which we can feed into a model, including:

* **id** – house identifier
* **date** – Sales date
* **price** – Sales price
* **bedrooms** - # of Bedrooms/House
* **bathrooms** - # of bathrooms/bedrooms
* **sqft\_living** - square footage of the home
* **sqft\_lot** - square footage of the lot
* **floors** - total floors (levels) in house
* **waterfront** - House which has a view to a waterfront
* **view** – House which has been viewed
* **condition** – The overall condition of the house
* **grade** - Overall grade given to the housing unit, based on King County grading system
* **sqft\_above** - square footage of house apart from basement
* **sqft\_basement** - square footage of the basement
* **yr\_built** - Built Year
* **yr\_renovated** - Year when house was renovated, if applicable
* **zipcode** – zipcode of the property
* **lat** - Latitude coordinate of the property
* **long** - Longitude coordinate of the property
* **sqft\_living15** - The square footage of interior housing living space for the nearest 15 neighbors
* **sqft\_lot15** - The square footage of the land lots of the nearest 15 neighbors

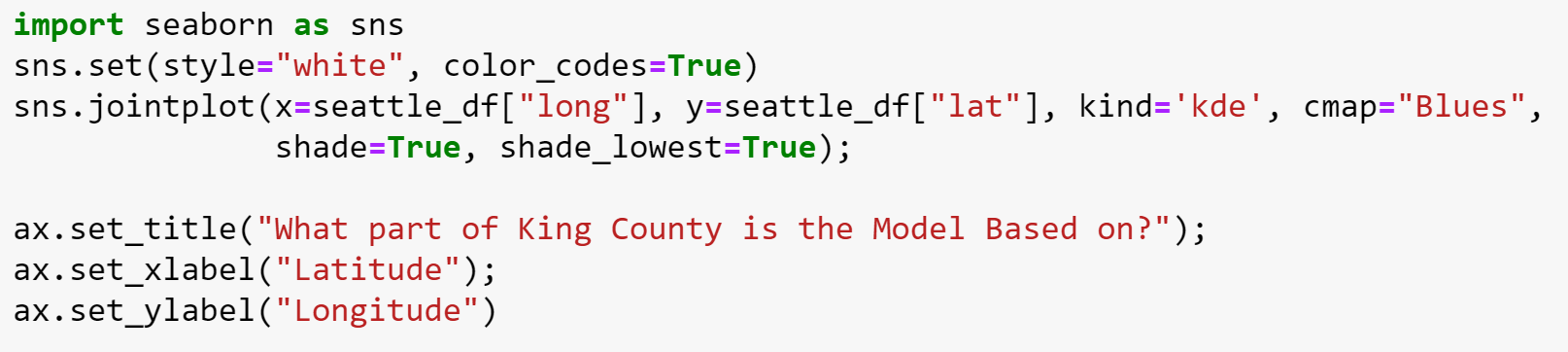
But how do we know that this data is relevant more specifically to a realtor ? As sales people, most realtors have a territory they cover. One way would be to do a 3D Geo-location plot to which shows where the data comes from.

We can use the KDE (Kernal Density Estimation) plot in Seaborn which is a non-parameterized way to show the distribution of the data as a function of two features, in this case longitude and latitude.

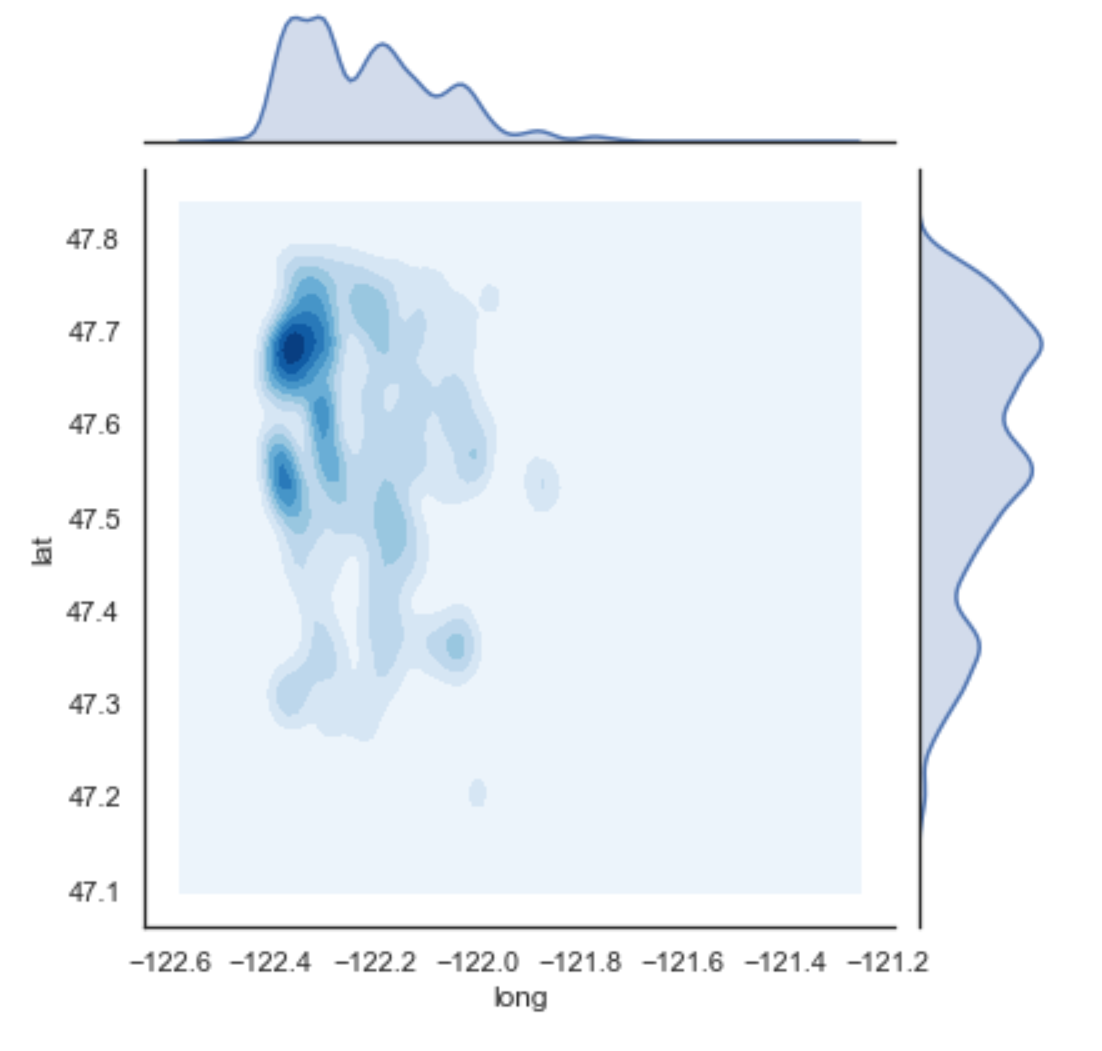
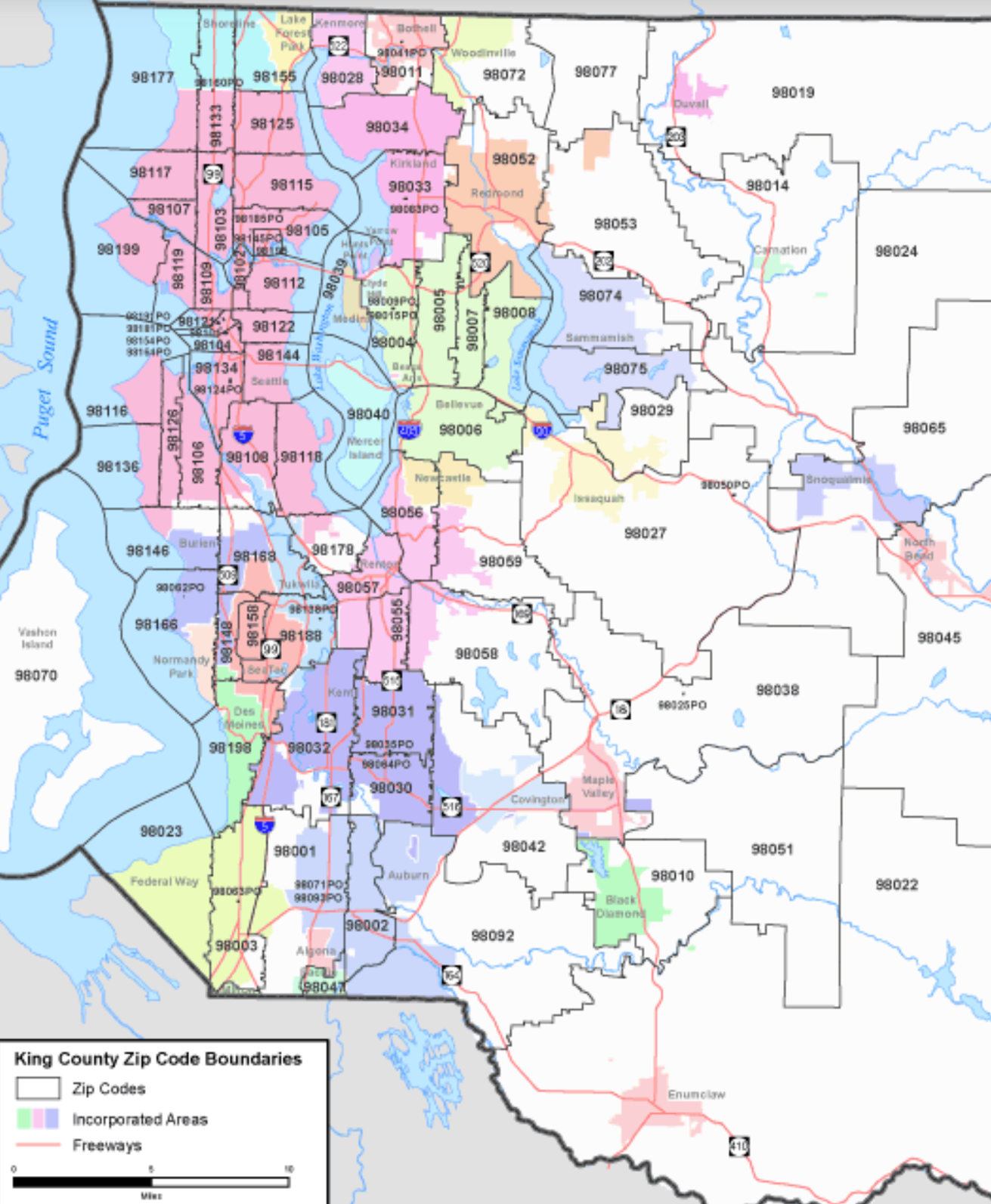


In plain English, this plot shows where the majority of the home sales data comes from - the upper Northwest portion of the county, including all areas between the Puget Sound and Lake Washington. The darker the hue, the more home sales in that region. Very little of the data comes from outlying areas of the county.





Data primarily from Northern part of King County. Darker regions indicate more sales

Puget Sound

Lake Washington

Longitude

Longitude



Latitude

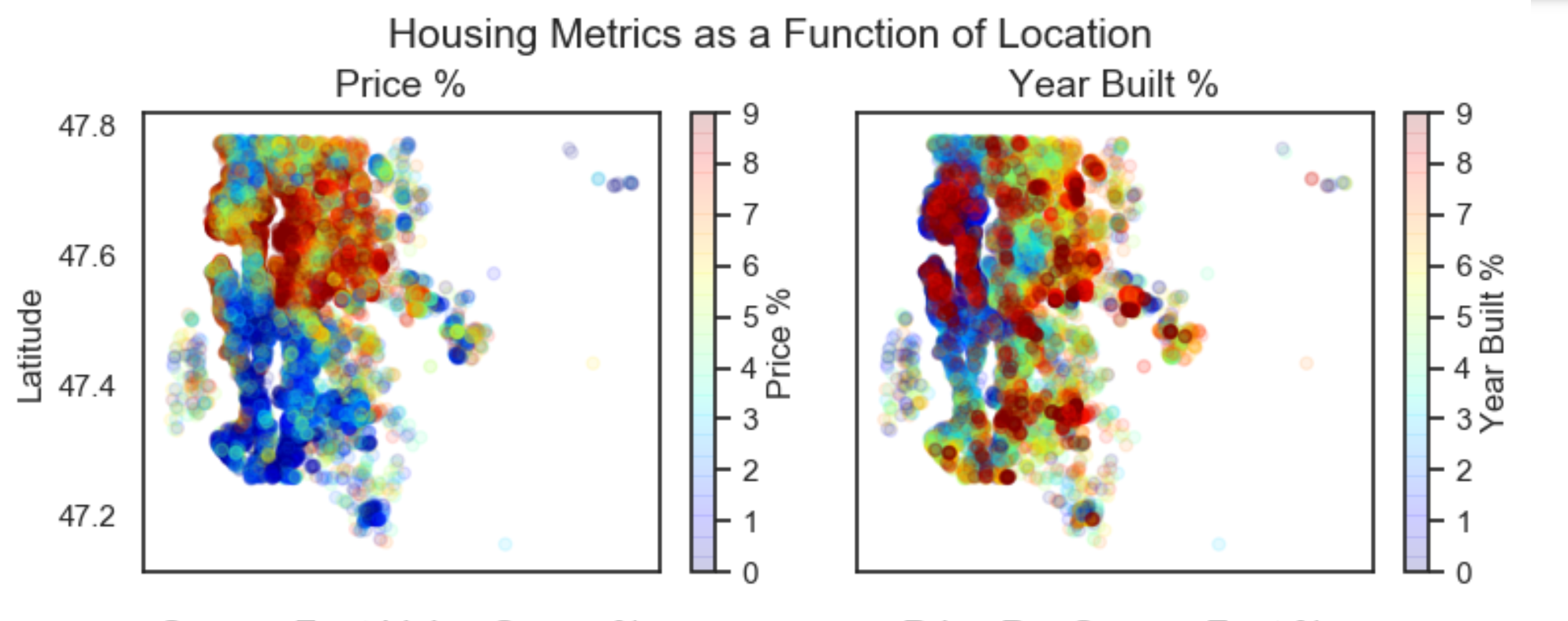
**Data Results Relative to Historical Perspective**

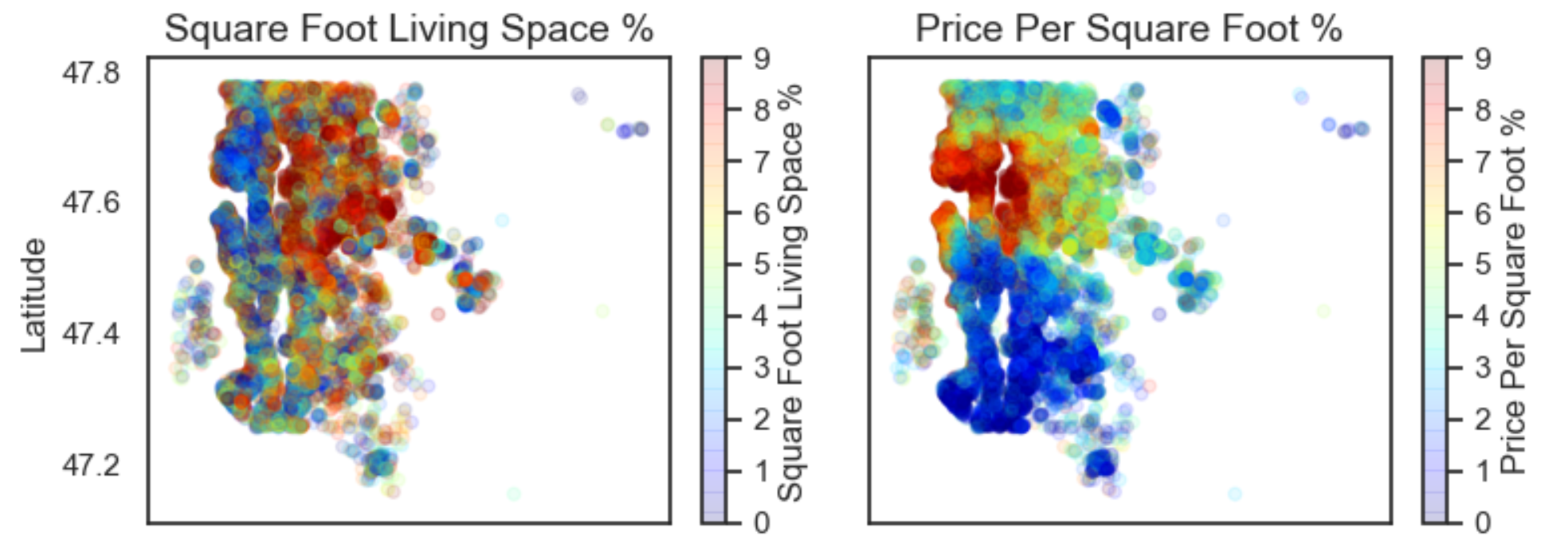
Now we zoom into some more detailed results which show realtors specifically how pricing, square footage, age of properties and price per square foot varies as a function of longitude and latitude. This connects to the terms realtors use on a day-to-day basis, with a geographical view they might not have seen before.

The color scale is based on percentiles, such as quartiles (0%, 25%,50%, 100%) , or in this case deciles (10%, 20%, 30%, etc.) In plain English, the color scale is setup so that red means at the 90th percentile, or highest values and blue is the 10th percentile, or lowest of the values.

Interestingly, for house prices the highest prices (90th percentile) are around Lake Washington and the coastline, as could be expected. But there are high priced homes also to the east of the Lake Washington area. For the distribution of houses by age, we can look at the Year Built. There we see a lot of new home construction all throughout the peninsula, mixed in with many of the oldest homes. To the east of Lake Washington we see a the houses are in the median for age.

The bottom two graphs are perhaps the most interesting. On the left, the largest homes, in terms of square foot living space, the smallest homes are on the peninsula. The largest homes are in East of Lake Washington.





**Longitude**

**Longitude**

As a result, when we look at the price per square foot, we see an amazingly clear picture of hot the price per square foot varies across the region, with a “high priced zone” all along Lake Washington and the Puget Sound coastline, the lowest prices per square foot directly to the South, and a mixture of newer, more spacious houses to the East of Lake Washington.

Most likely for realtors who are familiar with King county, this type of graphic will feel comfortable as this says our data confirms a systematic approach to pricing which is consistent with their existing intuition.

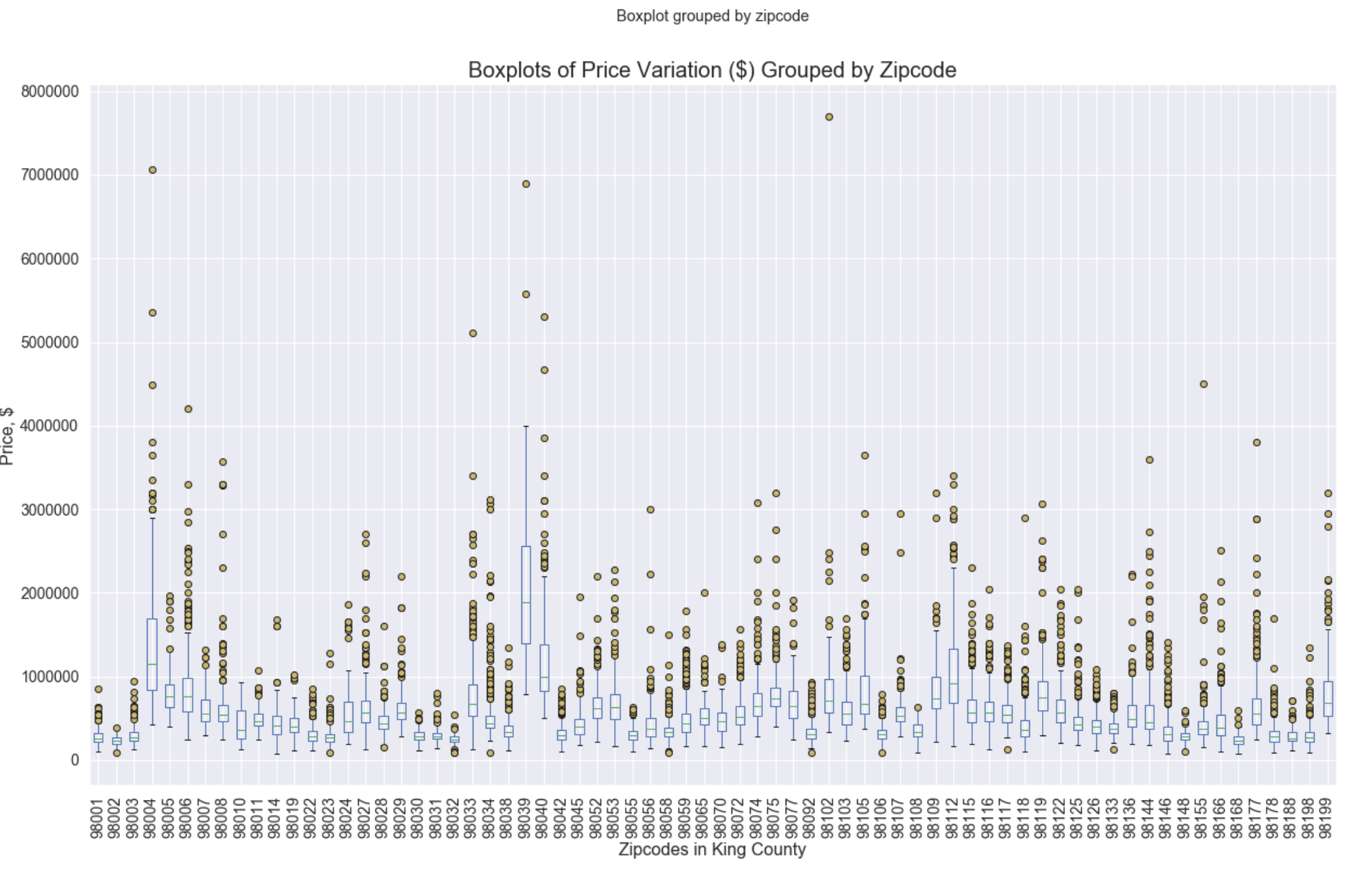
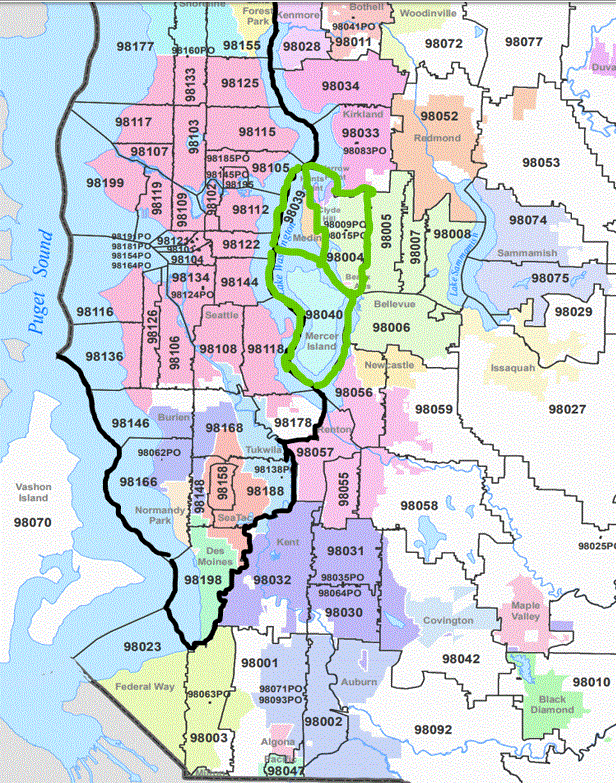
Several questions based on geo-location may come to mind:

1. Are there any significant trends between zipcode and price ?
2. What's the home price impact of having waterfront housing ?
3. Is there a significant difference between house prices on the peninsula versus away from the coastline ?
4. What other factors are key for increasing home prices?

The next graphic may surprise our realtor audience – because it might be counter-intuitive. Many people would like to know the trend of prices, across zip codes.

The boxplot figure below shows the distribution of home prices across all the zip codes in King County. The height of the box and the lines up and down (often called whiskers) show how much prices vary in a single zipcode. The longer the box and its whiskers, the more price variation. The median house price is shown as the horizontal line in the middle of each box. The yellow circles are what as known as “outliers” – they are the house prices that fall outside of the “bell curve” or normal distribution.

As one can seen, almost all of the zip codes have “outliers” relative to the other homes in their zipcode. And some zipcodes have really large outliers overall – houses greater than $4M.

981XX Zip codes

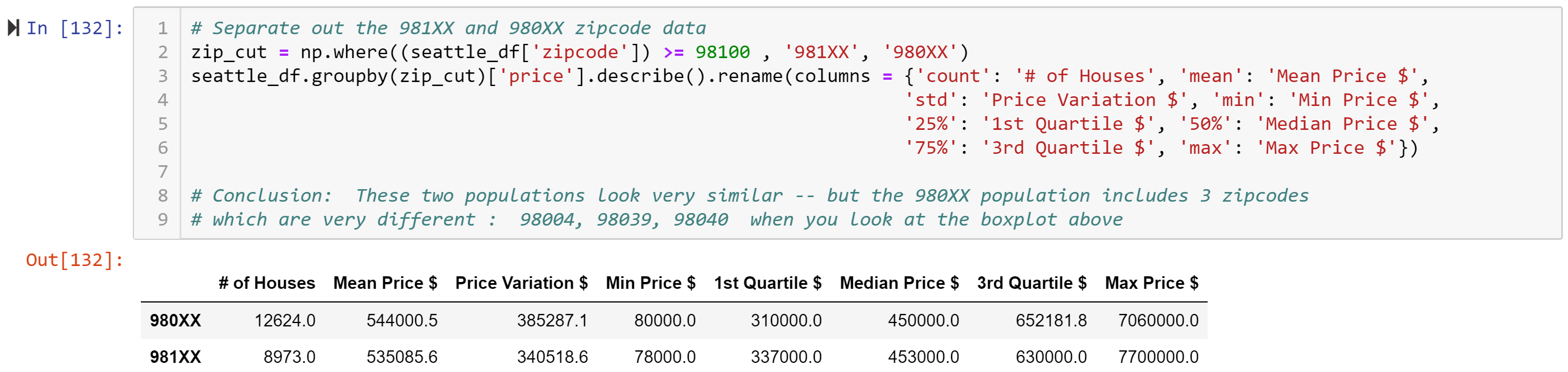
Mercer Island

Medina Island

Lake Washington

For almost all zip codes, the median house price is << $1M dollars. There are a few zip codes which generally seem to have higher prices (the median is far above $1M): 98004, 98039, and 98040 and may not be representative of the population as a whole. As you can see from the map outlined in green, these are special zipcodes with scarcity of housing supply: Mercer Island, Medina Island and the Eastern bank of Lake Washington.

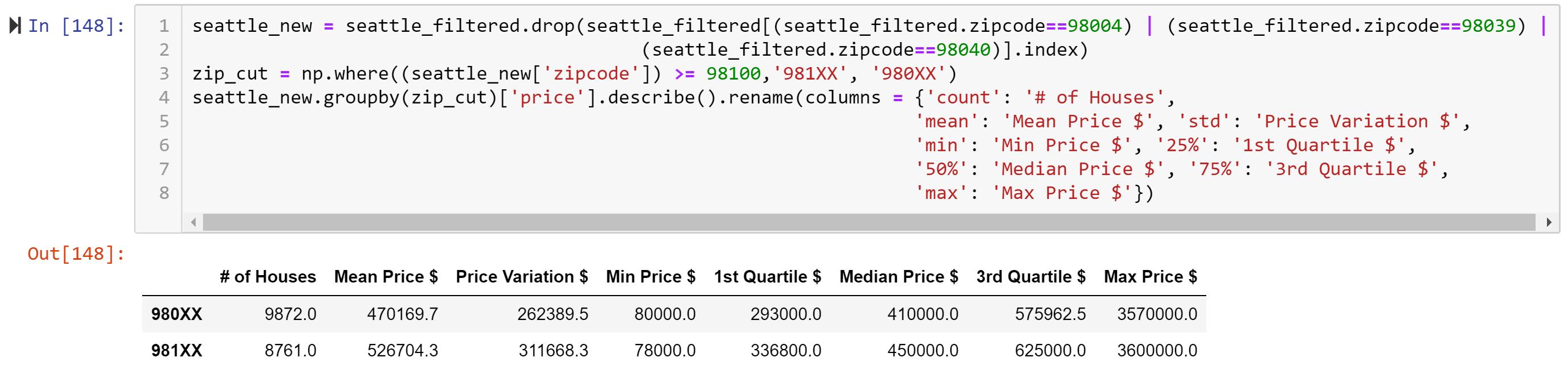
An immediate question that comes to mind is : what’s the difference in median house prices between the 980XX zip code and the 981XX zipcode ? Is there a significant difference?



Initially, it looks like there is no difference for Median home prices in the different zipcodes. But that may be skewed because the 908XX group of zipcodes contains the “key area zip codes” (98004, 98039, 98040).

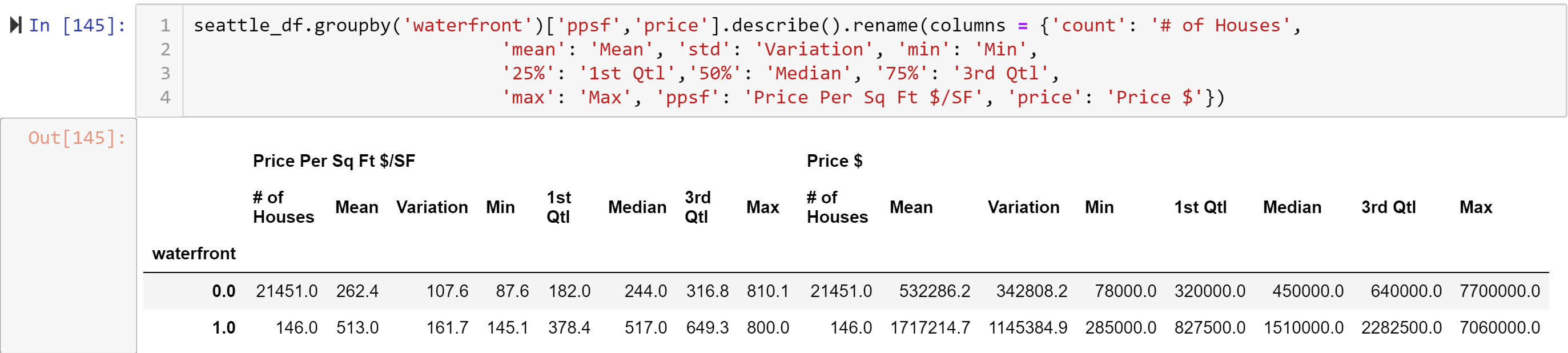
Performing some filtering on the data may lead to a more precise conclusion.

Filtering was performed on the original data, to remove outliers in lot size across all zip codes (seattle\_filtered). And the special zipcodes (98004, 98039, 98040) were also removed.



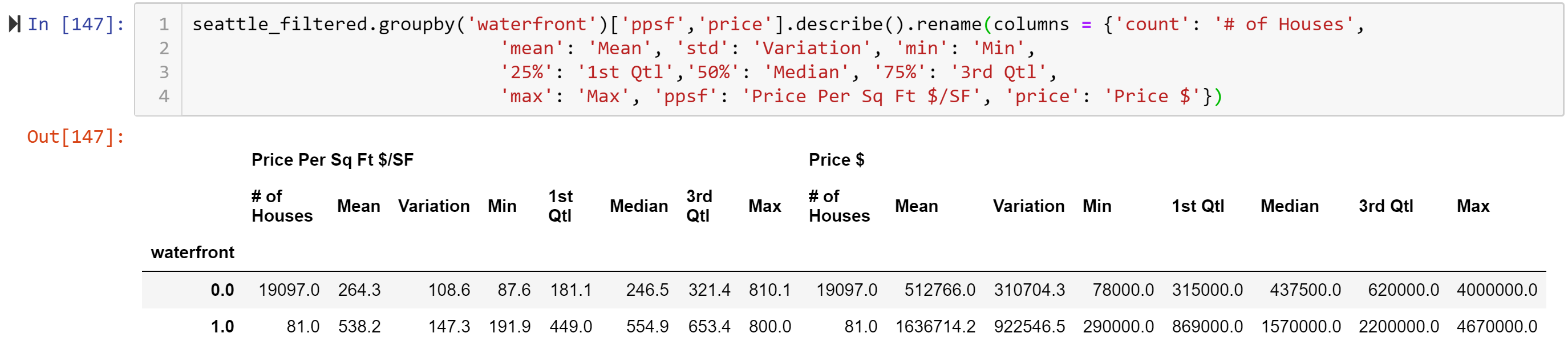
The conclusion is that there is ~ 10% difference ($40,000) in median house prices primarily based on zipcode.

Another question is whether the outliers are all properties on the waterfront or with views – and what is the average house price “add-on” for waterfront property. These questions can be answered with the analysis.



It should be noted that the maximum sales price highly varies, dependent on many other features as shown by the large number of outliers :

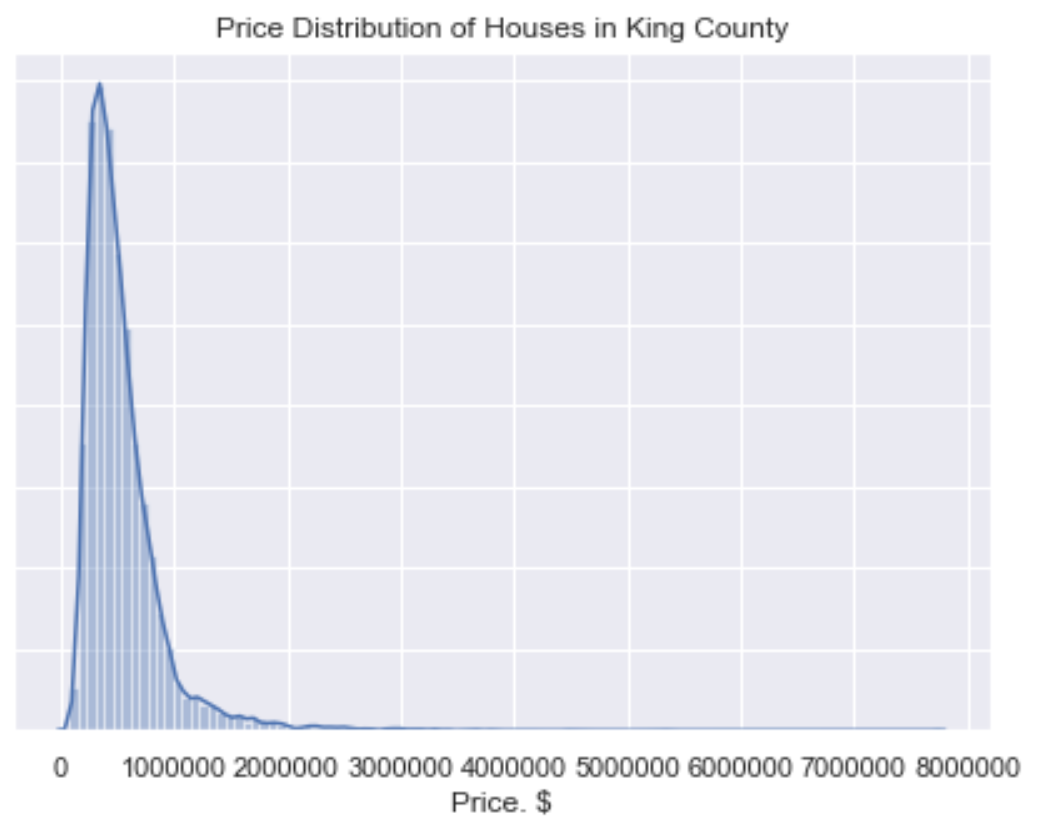




There is a clear pricing trend, irrespective of filtering: more than double (2X) the Median Price Per Sq Ft and ~3X the Median Price can be due to a waterfront feature.

**Data Model – Developing Intuition**

The next area of discussion turns to a more mathematical look at our modelling which helps build intuition and trust into the model. The first thing we do is try to look at if house pricing follows a normal “bell curve” as shown in the Figure 3 below.



**Number of Houses (A.U.)**

**0 $1M $2M $3M $4M $5M $6M $7M $8M**

**Figure 3 Distribution of Home Prices May 2014-2015**

“Tail” of the distribution

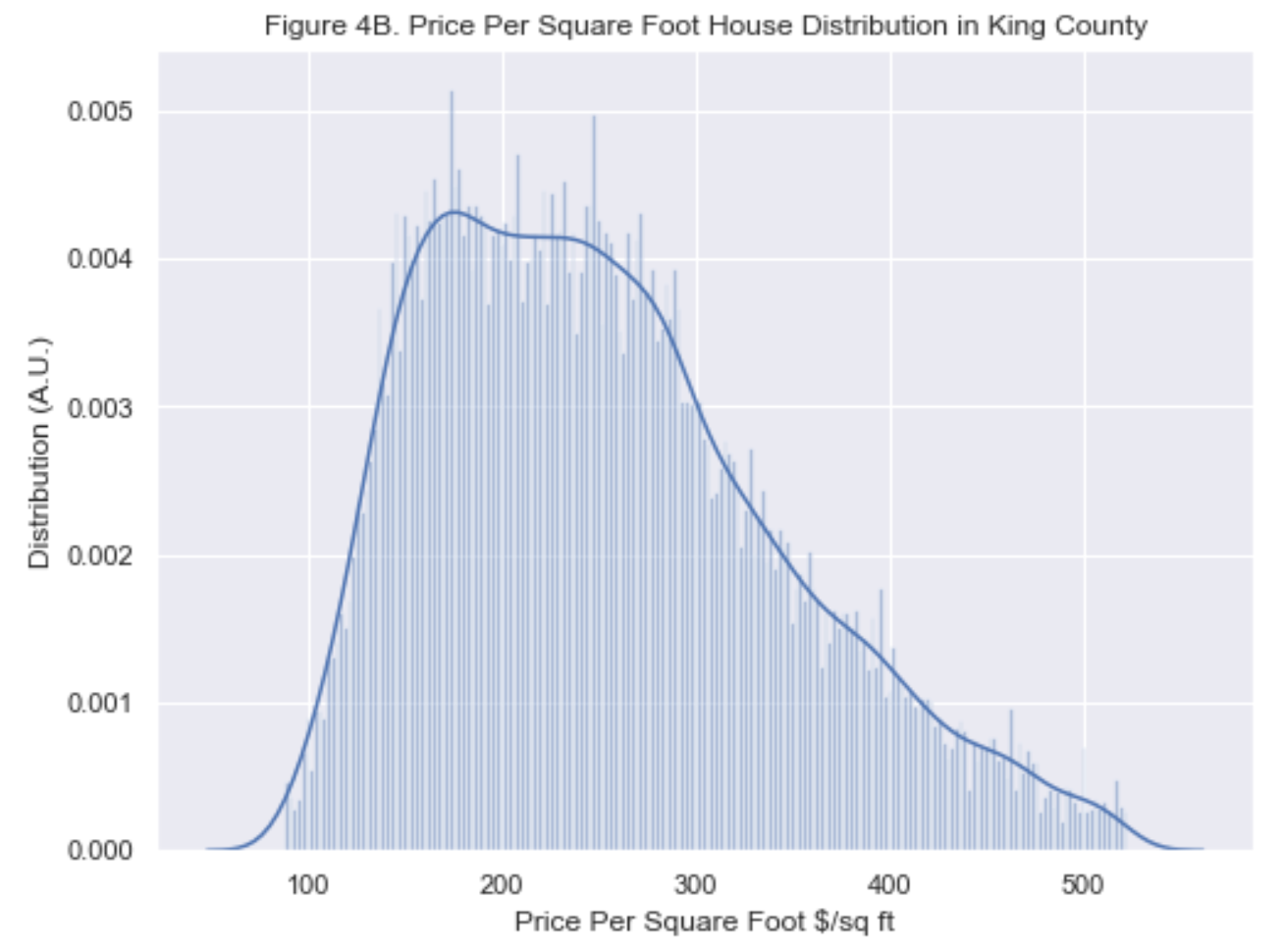
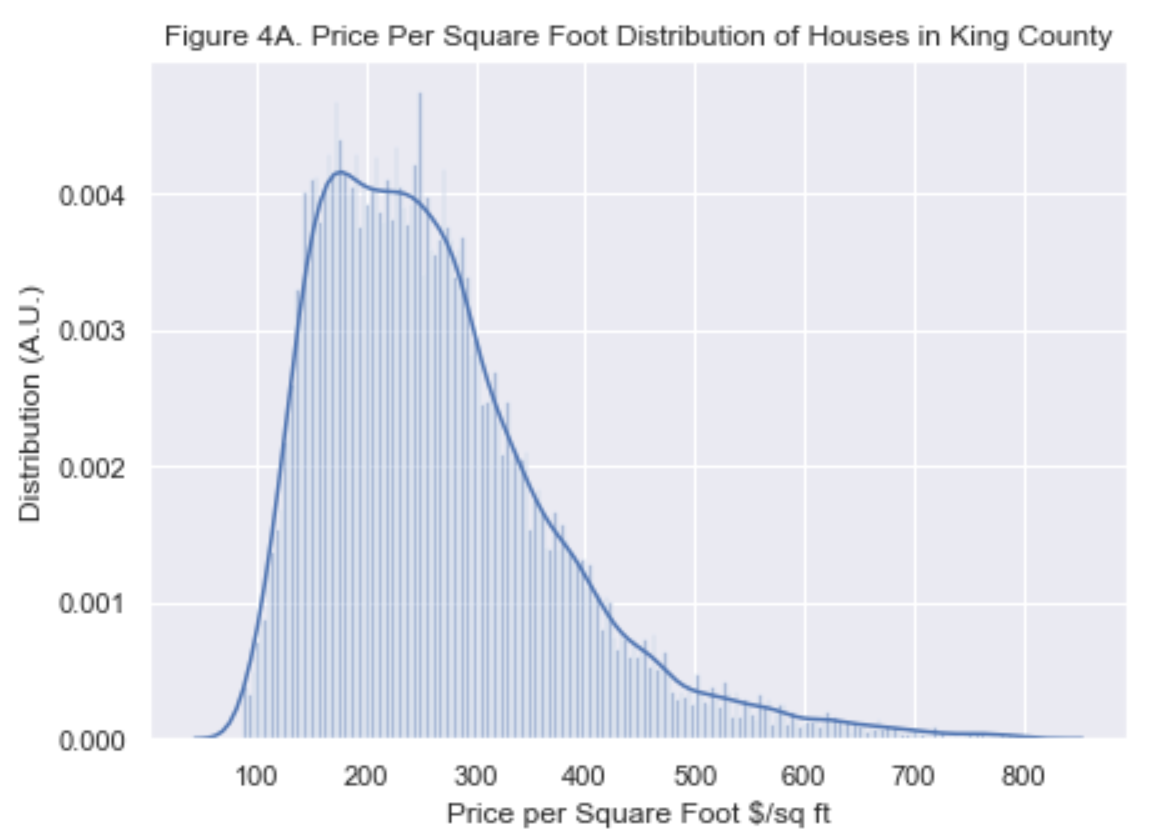
We can see that the shape of the price distribution is not normal. The majority of the house prices are below $1 million dollars, but there is a long “tail” of very high priced homes.

For the average realtor whose business is in the territory of houses below $1M dollars, a model which represents only those > $1M would not be of much interest.

For a multi-linear regression model to work well, a more normal distribution curve of the data tends to result in a higher accuracy model.

One way of trying to improve the accuracy is to filter those datapoints, which do not match to the general population of the data, by removing what we call outliers, through a statistical calculation of outliers. The details are shown in my github repository.

Another way to look at it is through the distribution of the Price per Square Foot, which as we saw was very well behaved across geographically as shown in Figure 4A and 4B.



Looking at Figure 4A, we can see that dividing the home price to the size of the living space, something of a normalizing effect, eliminating an extremely long tail of the distribution. If we apply a statistical filter that removes statistical outliers, you can see the truncation of the data on Figure 4B.

In general, it’s pretty risky to create filters so that a model fits with a higher relative accuracy, but it can be done if there is a reasonable explanation. In this case, houses which are generally much higher priced (or higher price per square foot) are not likely to be the house the average realtor would be selling.

To test out the impact this has to linear regression models, the raw and filtered data were used in a linear regression model provided by Statmodel.

The table below shows the model fit results. The use of price per square foot filtered data helped raise the adjusted R2 accuracy, but the collinearity remained very high. It was extremely difficult to model this dataset to a higher accuracy with low Condition number. Scaling and normalization were extra steps taken, but they didn’t substantially change the skew and kurtosis. The Condition Number (collinearity) of the model was improved, but the accuracy suffered, indicating that the model accuracy is heavily dependent on collinear features.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Model 1 –  Original Data | Model 2 Filtered on Square Footage of the Lot | Model 3  Model 1, Filtered on Price | Model 4,  Model 1 filtered on Price per SqFt | Model 5  Model 1 with 1 hot encoding and Normalization |
| Normalized? | No | No | No | No | yes |
| R2 Adj | 0.737 | 0.737 | 0.738 | 0.737 | 0.594 |
| Skew | 1.56 | 1.56 | 1.57 | 1.57 | 4.87 |
| Kurtosis | 13.35 | 13.36 | 13.51 | 13.45 | 71.19 |
| Cond. Number | 2.01e8 | 2.05e8 | 2.09e8 | 2.02e8 | 223 |

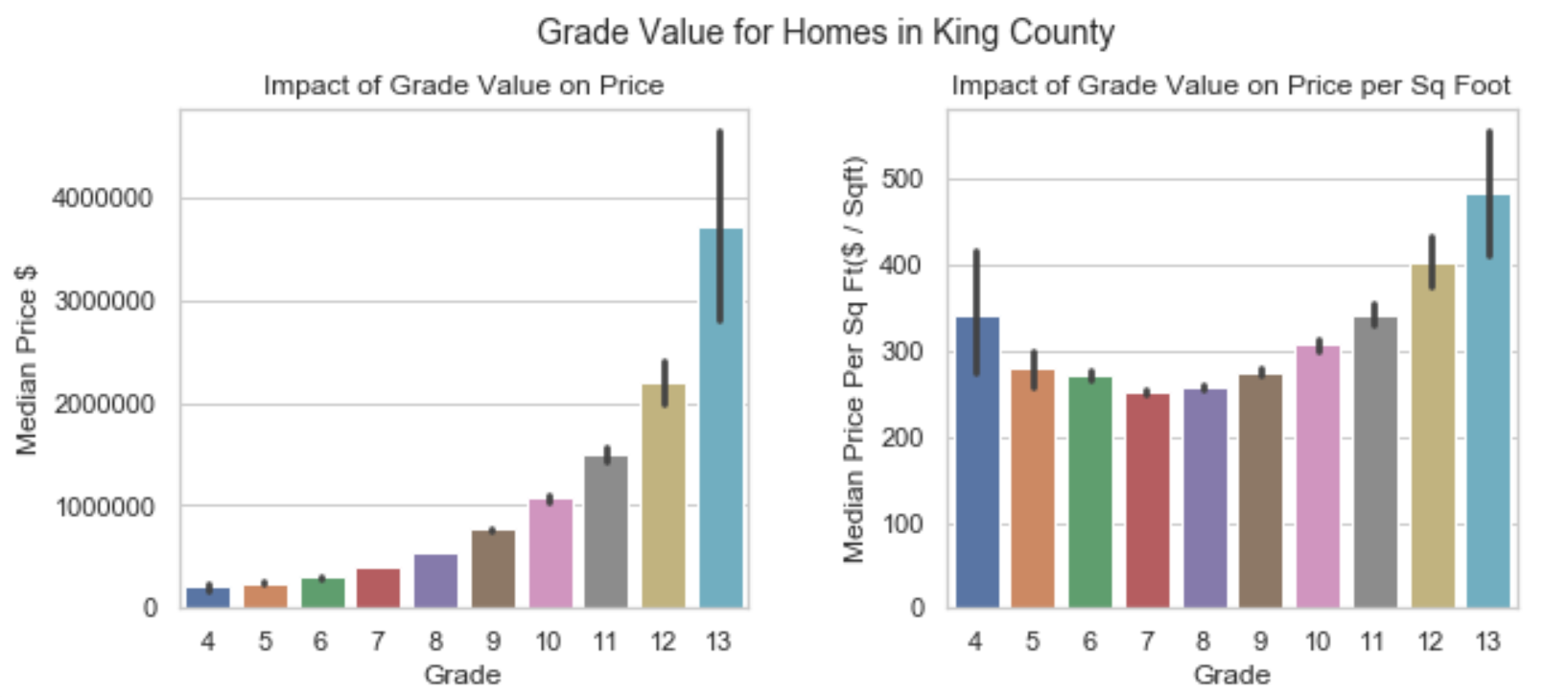
What were the most important features as determined by the model fitting ? There was a fair amount of consistency that the top features were latitude, grade, the house living area, the view and the age of the house. Most of these make intuitive sense.



But the two that should probably be investigated further are grade and square foot living. These are important to the model, and the question is – what kind of price increase could be expected, if there was improvement.

Let’s first take a look at grade. The one grade 3 looks to be a typo / error, and will be dis-regarded as there is only a single datapoint in that category, for this sample.





The King County Grade is dependent on the living area, and the Median Home Price appears to scale monotonically with the Grade. What’s a bit more interesting is that the Median Price Per Sq. Foot doesn’t scale monotonically with grade. Since the living area (sqft\_living) is part of the grade, the behavior of Median Price Per Sq. Foot suggests two things: First, that the sqft\_living tends to dominate the “Grade” and second, because that there are other influencers of price, particularly at low grade homes.

**Final Conclusions**