Bentley  
6.1 - Sourcing Open Data

**Kings County Data Set**

**Data Source**

**Source:** The source of this data set came from [Kaggle](https://www.kaggle.com/). From further research, the data came directly from [House Sales in King County, USA](https://www.kaggle.com/datasets/harlfoxem/housesalesprediction). This dataset contains house sale prices for King County, which includes Seattle. While the data I specifically pulled is from [Kaggle](https://www.kaggle.com/) I think the data can be deemed reliable.

**Collection:**

This data was collected by a combination of administrative and usage data. Each time a house is sold the information is sync in a database on the county website of Kings County. From their you can gather an array of information.

**Contents:**

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. Each house is assigned a unique ID in a row of the dataset. The data also includes the date, price, bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, waterfront, view, condition, grade, sqft\_above, sqft\_basement, yr\_built, yr\_renovated, zipcode, lat, long, sqft\_living15, and sqft\_lot15. There are 21613 rows and 21 columns in the dataset.

**Limitations:**

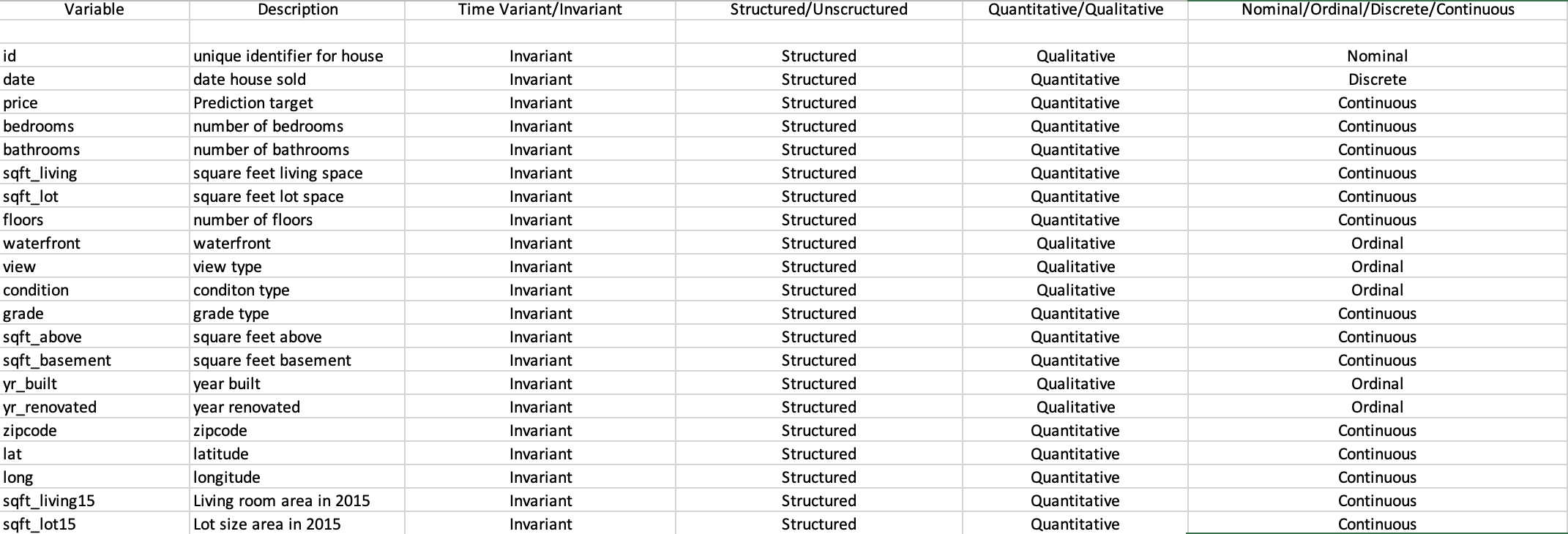
A limitation of the data is that I am unsure if the houses that are recorded are every single house sale that occurred. Also, it would be nice if we knew how many offers each listing had prior to it being sold as well as the bid/ask price possibly which would assist is being able to gauge a more precise understanding of the cost associated with each purchase.

**Ethics:**

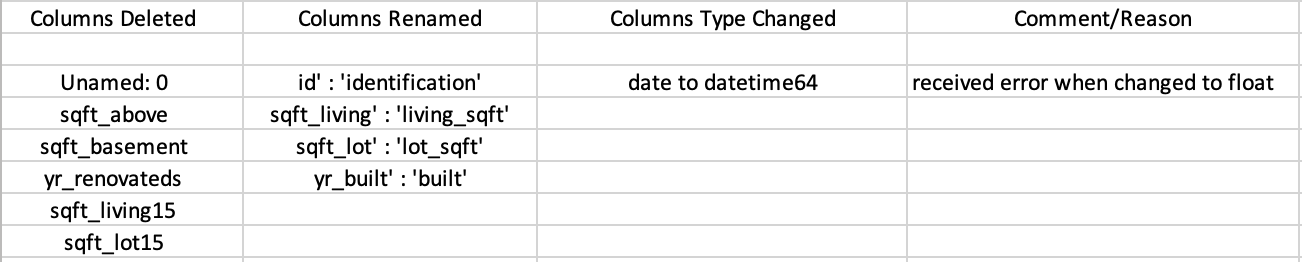
This information was gathered and displayed on their own website as being available to the public. It states that all the data is provided in accordance with NYCBS Data Use Policy. Clearly on the website Kings County Dataset Page goes over its Data License Agreement. It appears that there are no ethical concerns with this data.

**Relevance:**

I believe this data set meets the necessary requirements for this project as it is open source, includes a geospatial component, and meets the size and variable requirements. While it isn't as recent as 3 years old, this data set was provided in the CareerFoundry Project Brief. I used this dataset because I figured it would be a great use in predicting house prices using regression.

**Data Profile**

**Data Cleaning Measure**

**Wrangling**

**Consistency Checks**

****

**Questions to Explore for Analysis:**

* What are the fluctuations in daily average house prices
* What are the average house prices according to zipcode
* What are the distributions of house prices
* What are the distributions of bedrooms and bathrooms per house
* Is there a correlation of house condition and view on house prices