

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light green color. They are positioned diagonally, with the blue one partially covering the green one.

# King County House Sales

Exploratory Data Analysis and Linear  
Regression Modeling

# The Data: House sales in King County

- 20 Features, e.g. Size Measures, Quality and Condition and Neighborhood
- 1 Target: The Price the house was sold at
- 21597 Observations over one Year, from May 2014 to May of 2015

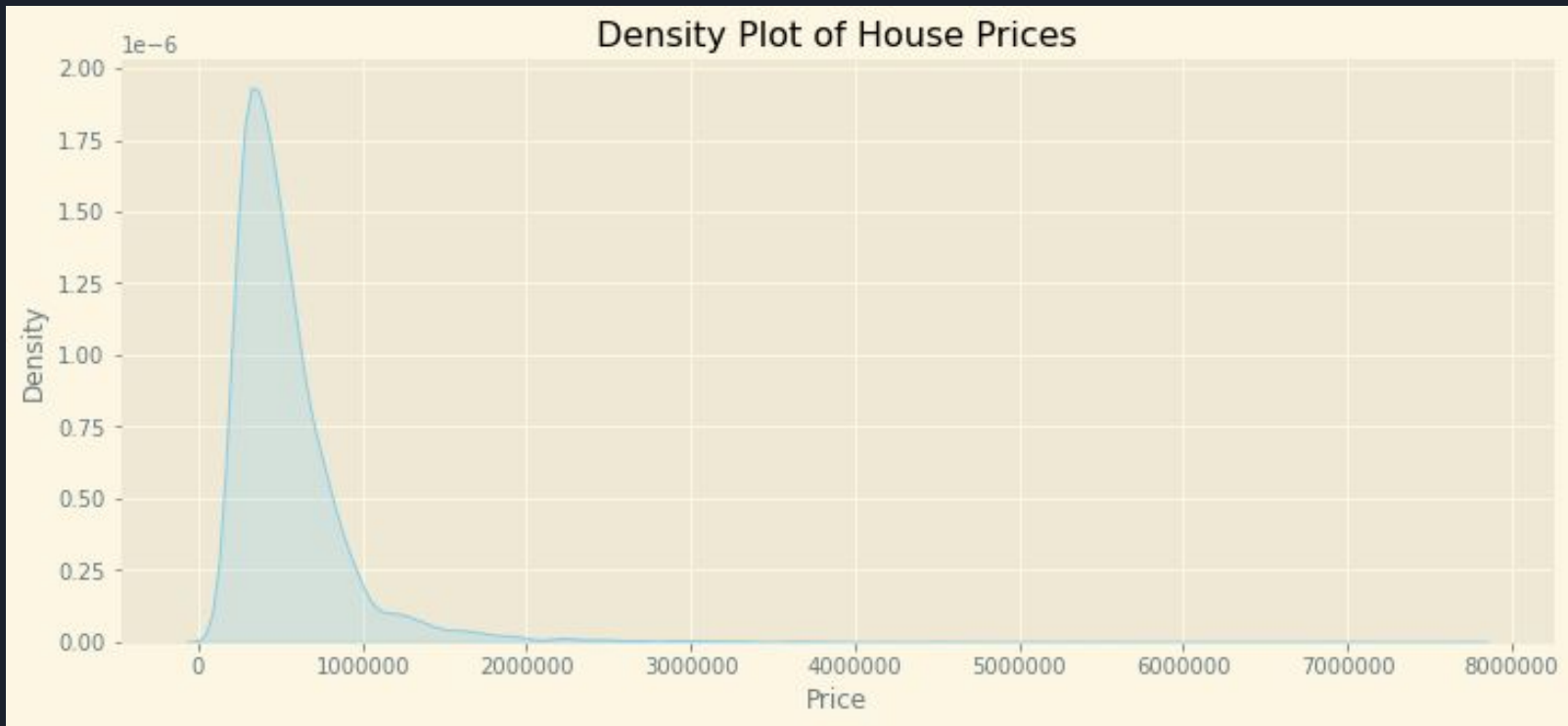




# Cleaning and Exploration of Data

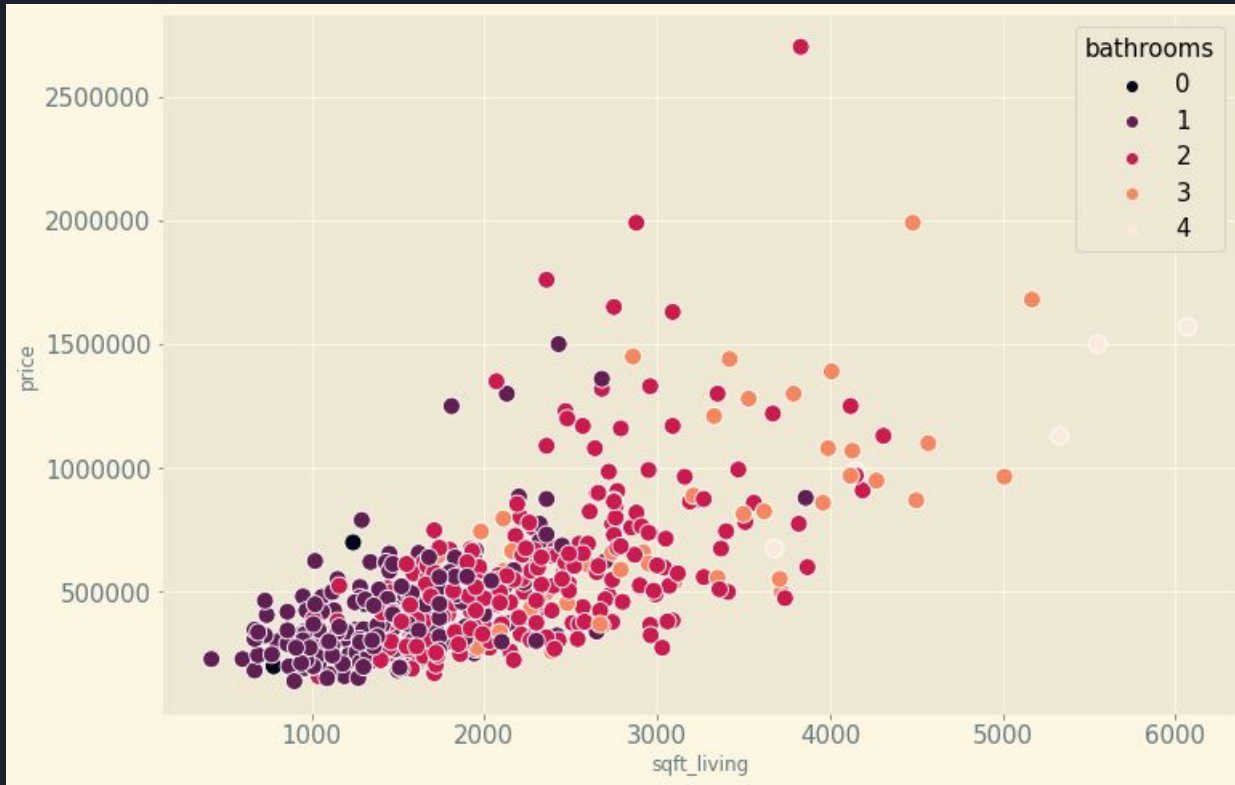
- Check Dataset, fix missing values and get formatting right
- Descriptive Stats and Visualizations
- Find the strongest Correlations of Predictors with target to form a preliminary hypothesis

# The Prices



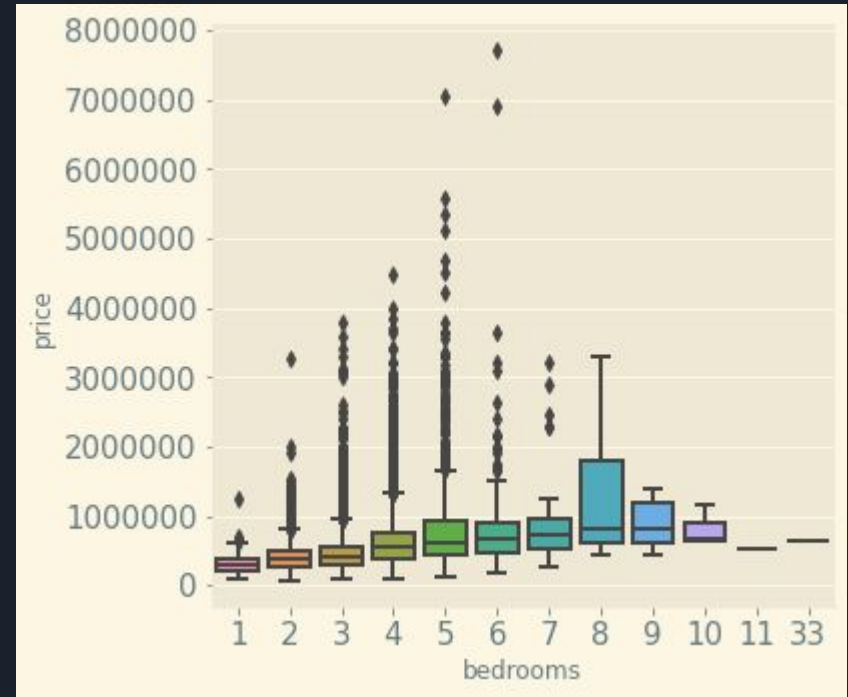
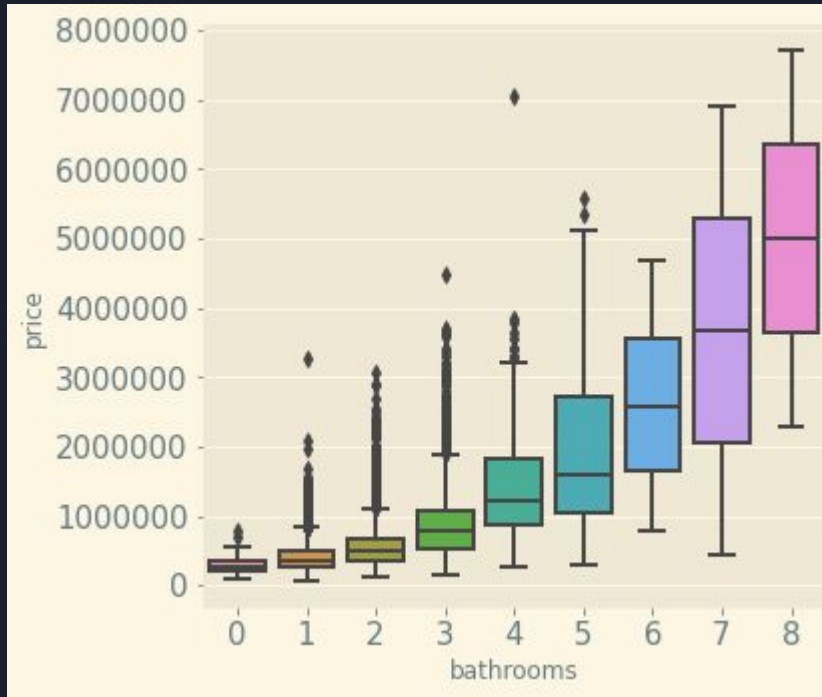
# The most promising Predictors

## Area of Living and Number of Bathrooms

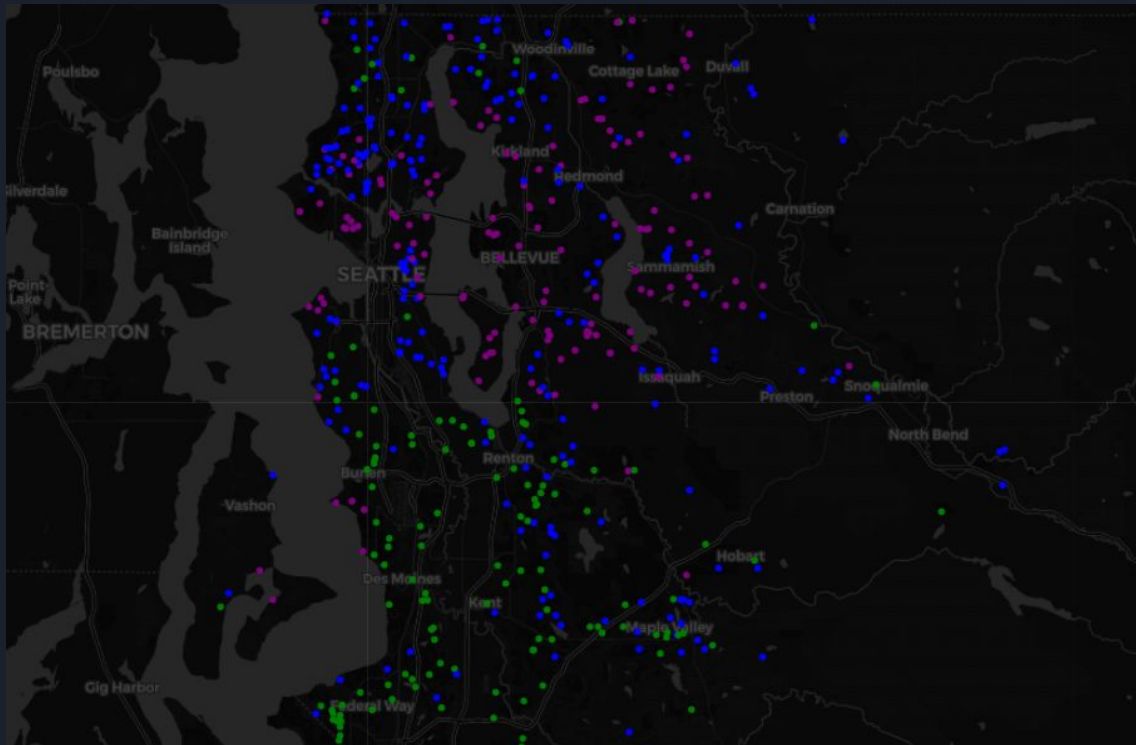


# The most promising Predictors

## Number of Bathrooms and Bedrooms



# The most promising Predictors Neighborhood



# The most promising Predictors

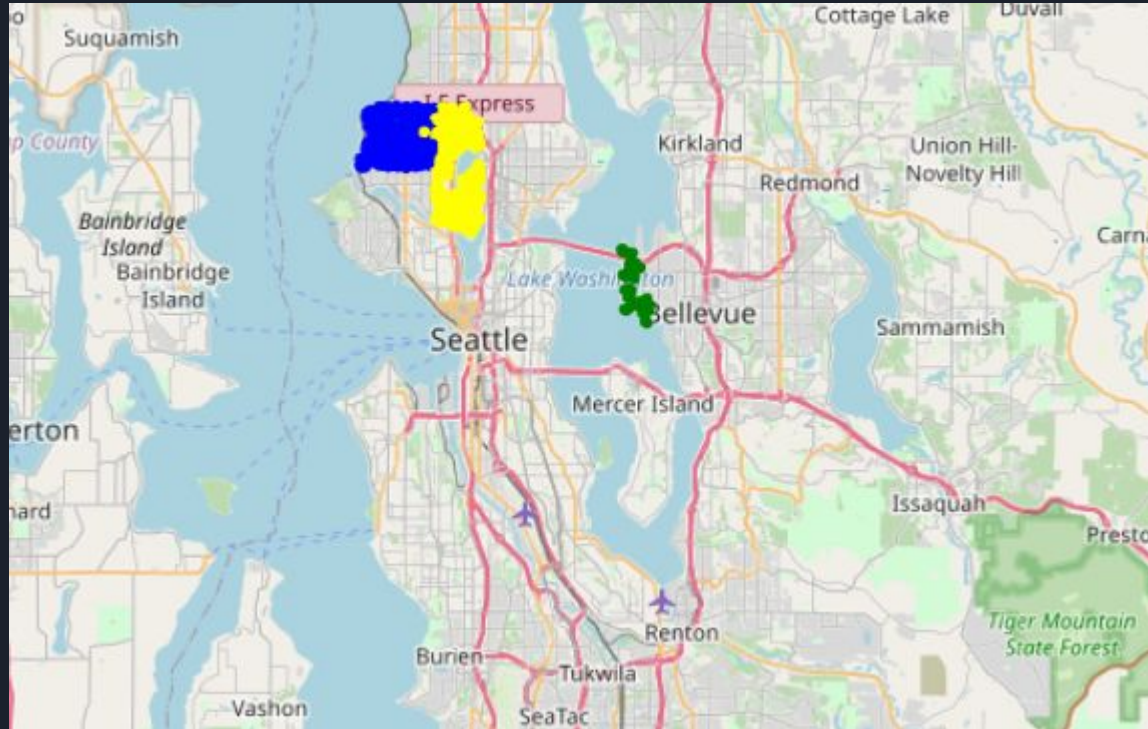
## Zip Codes





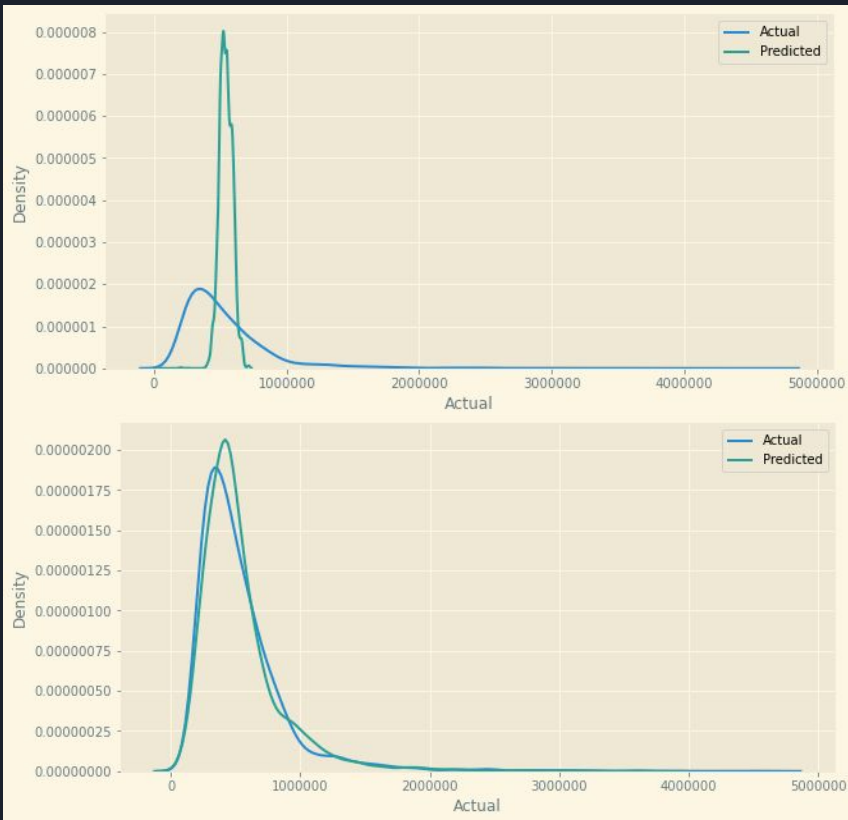
# The most promising Predictors

## Zip Codes with highest and lowest turnover



# Predicting the Price

## Results with weakest vs strongest Predictors





# Recap and Findings

A dataset of house sales in Seattle was cleaned, explored and used as a basis for modeling a predictive Model with a Linear Regression Algorithm

- The best Model included the 30 highest correlated variables with the target and had a  $R^2$  of 0.77 (RMSE 173,361). It was better than the model including all of the variables ( $R^2$  0.56)
- The sales of all months in the Interval from 2014 to 2015 were roughly the same, this might point to an issue with the dataset

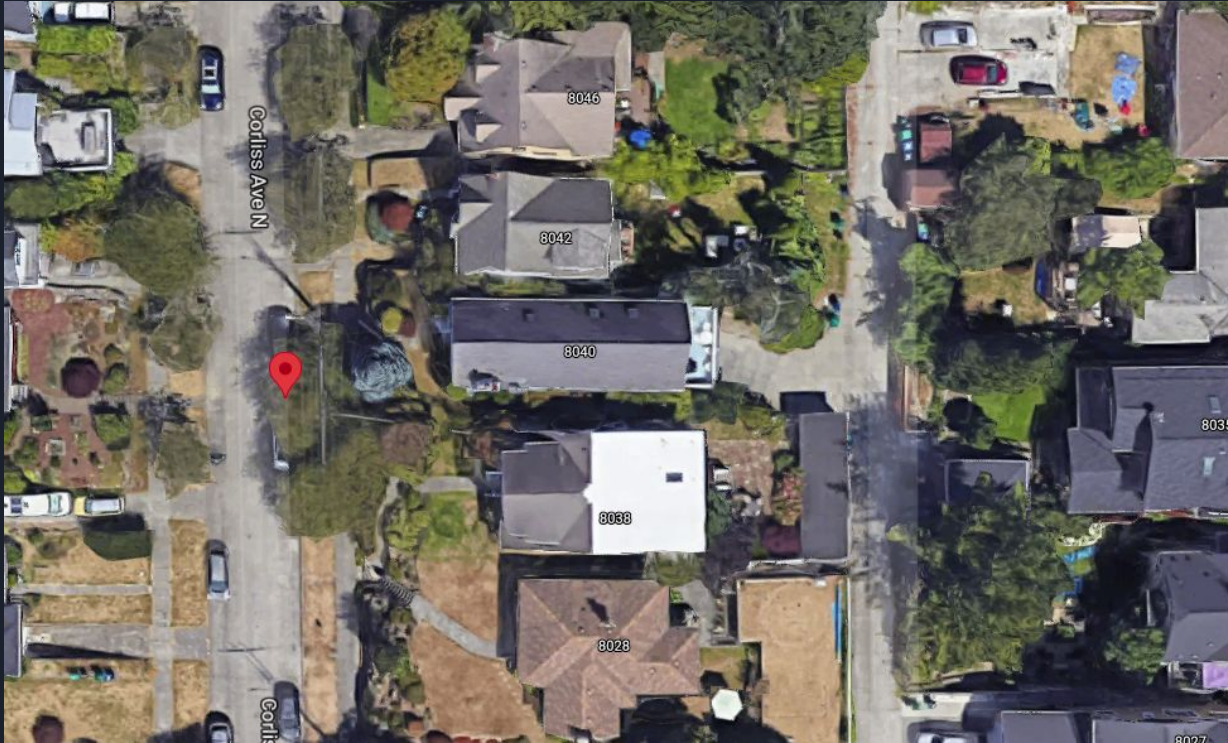


# Ideas to improve Prediction by the Linear Regression Algorithm

- Other methods for replacing missing values could be tested or more data derived from other sources. Also, geographical Data given could give hints
- Possible irregularities and typos found in the dataset should be amended (Sales and number of bedrooms, duplicates)
- A Feature pulling together Information about the Age and Condition of a House and if it was renovated and when could prove to be useful
- The Date of the Sale could be important due to the high volatility of the housing market if Market Data were included
- Demographics and Crime Data could be included
- Polynomial Features could be included to account for possible nonlinear Relationships like year built
- Transformation of the price data could be helpful to improve the distortion of house prices due to the luxury segment of the market
- Multicollinearity in the data should be reduced

# Addendum

## Who has 33 bedrooms...not this guy





Addendum: 7700000 \$ House  
looks reasonable, no typo here



# Weird Sales Data?

