King’s County Data Set

Module01 – Final Project

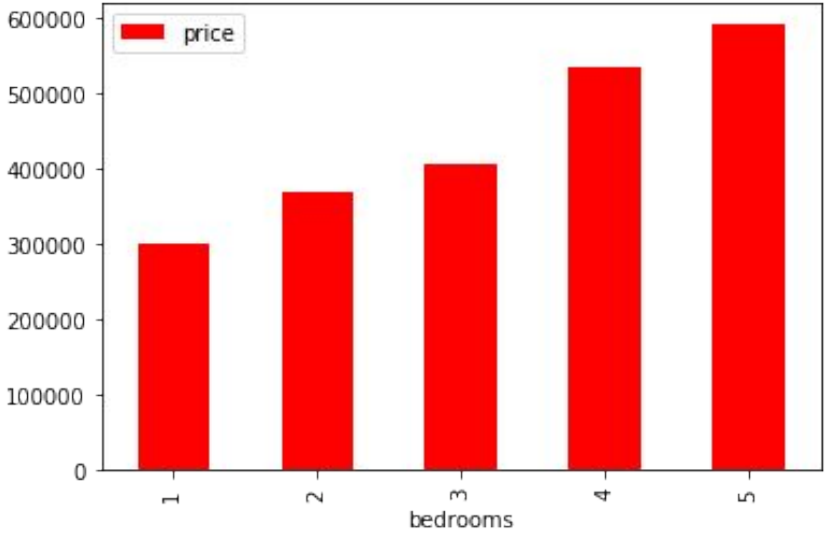
The Problem:

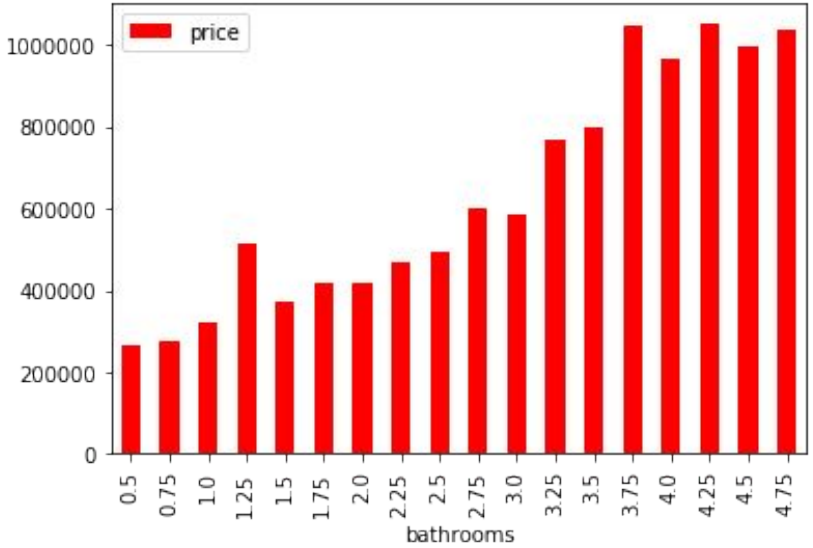
Predict the prices of house as accurately as possible.

Steps used:

- Cleaned data by first finding null values and replacing them with appropriate values
- Checked the data types and made sure to fix the one that was supposed to be a ‘float’ but was an ‘object’ -- sqft_basement
- Checked for outliers and multicollinearity and took steps to continue cleaning the data
- Created three new features to help with the data:
  - price/sqft
  - 3beds/2baths
  - Quality of homes
- Lastely, separated the data into continuous and categorical data in order to run an OLS
  - Ended with an r-squared value of 0.972 which shows how closely the model fits to the data

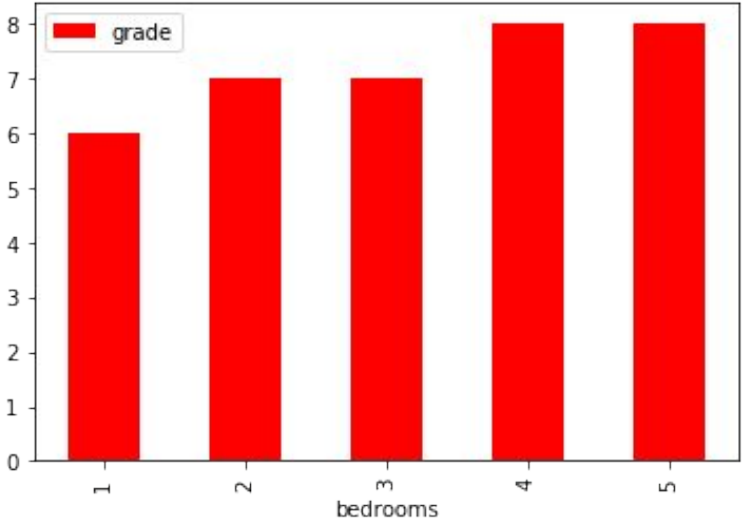
Bedrooms and Bathrooms - PRICE

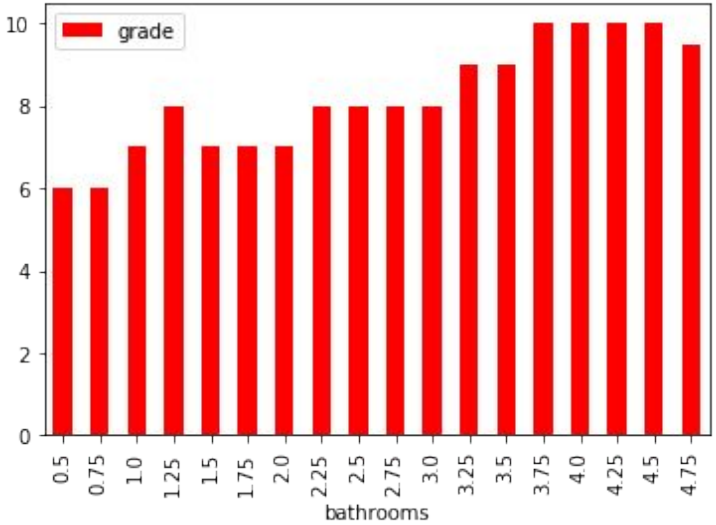




Houses that have 3 bedrooms and a range of 1 to 2 bathrooms sell for a reasonable price -- affordable

Bedrooms and Bathrooms - GRADE



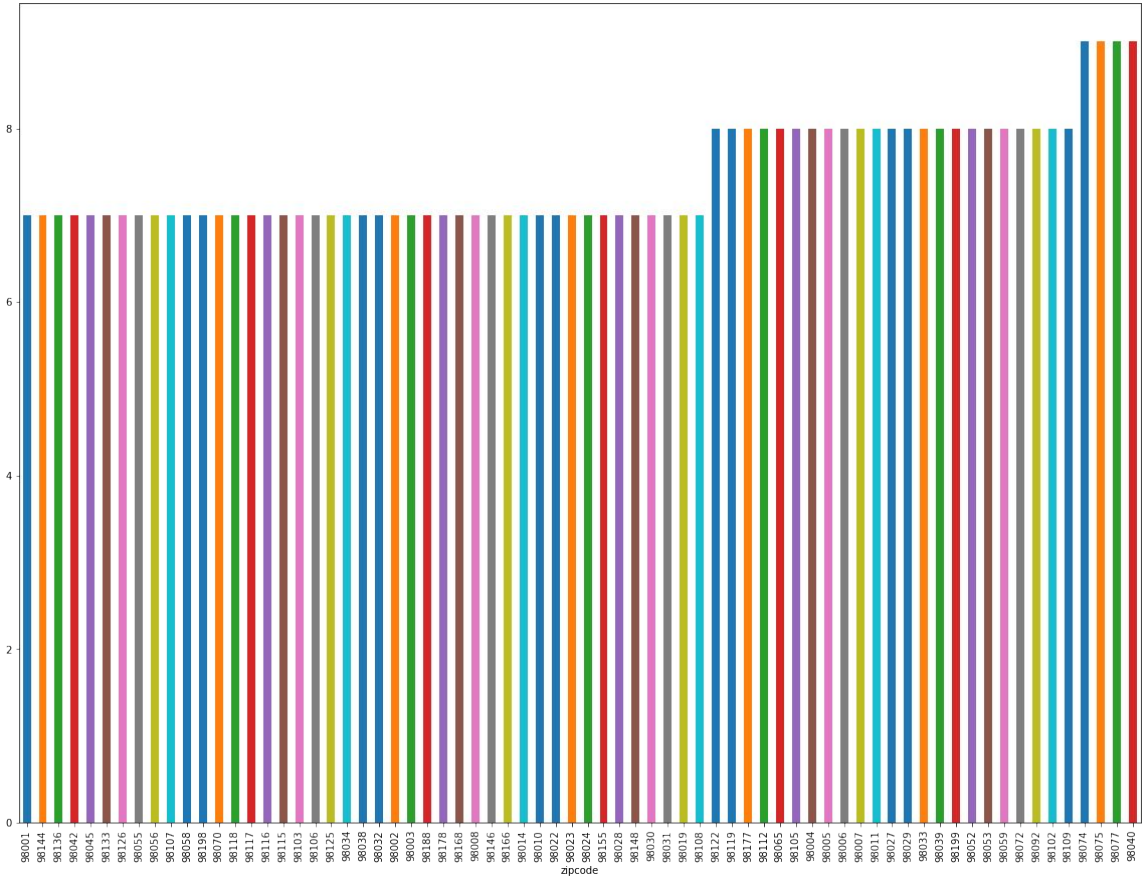


Houses that have 3 bedrooms and a range of 1 to 2 bathrooms have an average grade of 8 out of 11.

Where to buy and sell?

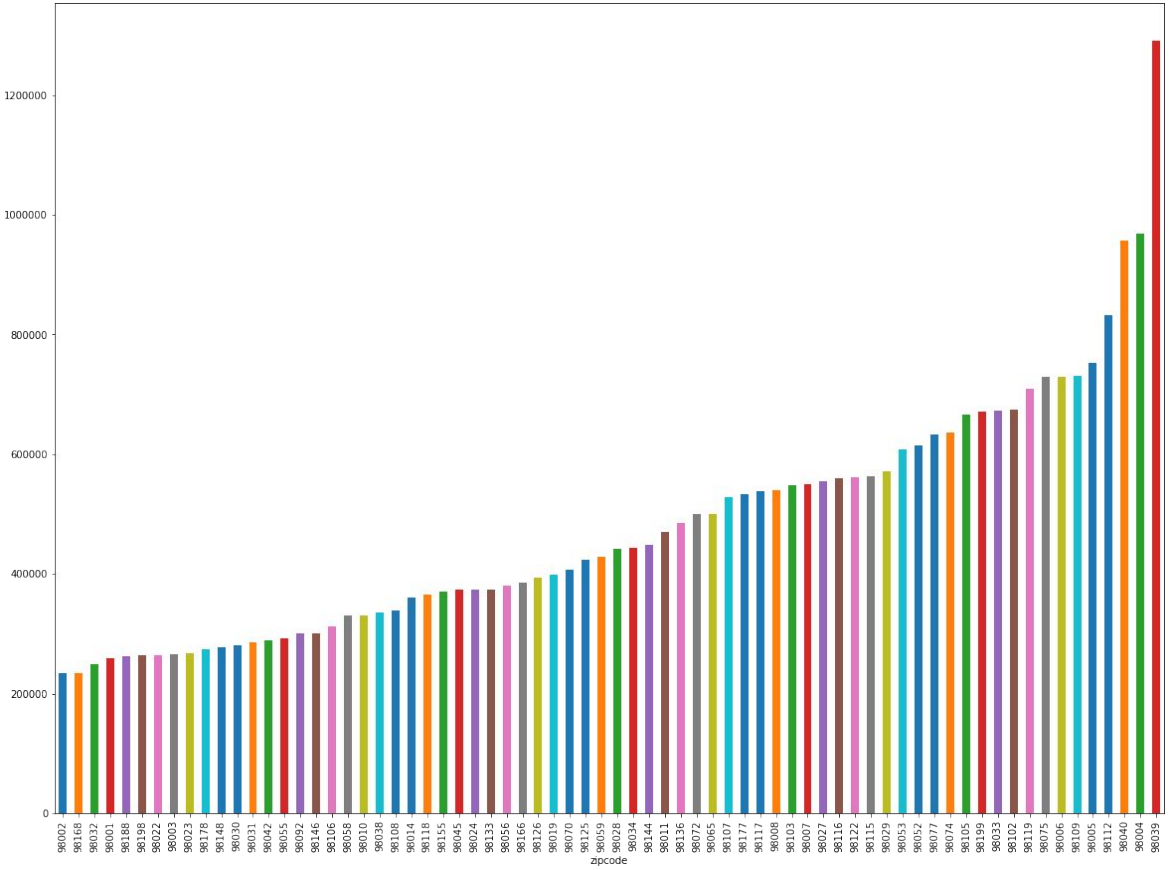
The zipcodes with the best quality houses fall into - grade 8:

- 98122
- 98119
- 98177
- 98112
- 98065
- 98105
- 98004
- 98005
- 98006
- 98007
- 98011
- 98027
- 98029
- 98033
- 98039
- 98199
- 98052
- 98053
- 98059
- 98072
- 98092
- 98102
- 98109



Average price of homes:

The homes in the zipcodes with the best quality homes range from $400,000 to the most expensive of $1M.



Variables that matter!

- Continuous variables
  - price_per_sqft
  - sqft_living
  - sqft_lot15
- Categorical variables
  - quality
  - bathrooms
  - bedrooms
  - view
  - condition
  - floors
  - waterfront

Conclusion:

With a R-Squared value of 0.972 this means that the model shows how close it fits the data -- therefore we can predict future home prices