KING COUNTY HOUSING

By Paul Williams

GOAL

- Based off the attributes of certain houses, I wanted to produce a statistical model that will attempt to predict the prices of houses in the King County, Seattle, Washington area.

AGENDA

- Present Dataset
- Data Cleaning
- 1 of my EDA questions and the findings
- Regression Model
- Conclusion

CLEANING

THE DATASET

- The data set provided had 21 features and 21,597 homes
- It had missing values and duplicates
- And a few extreme values

```
--------------------------------------------------
Number of row's before cleaning:  21597
Number of columns before cleaning:  21
--------------------------------------------------
```

```
Number of houses duplicated:  177
```

```
Total missing values: 6281
--------------------------------------------------
id                    0
date                  0
price                 0
bedrooms              0
bathrooms             0
sqft_living           0
sqft_lot              0
floors                0
waterfront         2376
view                 63
condition             0
grade                 0
sqft_above            0
sqft_basement         0
yr_built              0
yr_renovated       3842
zipcode               0
lat                   0
long                  0
sqft_living15         0
sqft_lot15            0
dtype: int64
```

THE DATASET CLEANED

- Median values to replace extremes
- Taking the latest ID and dropping the rest
- Filling waterfront with zero's

```
------------------------------------------------
Final amount of rows:  21420
Final amount of columns:  11
------------------------------------------------
```

```
Amount of Missing Values after the clean
------------------------------------------------
id             0
price          0
bedrooms       0
bathrooms      0
sqft_living    0
sqft_lot       0
floors         0
waterfront     0
condition      0
grade          0
yr_built       0
dtype: int64
```

EDA QUESTIONS

QUESTION 1

Does having more bathrooms than bedrooms increase the price of a house?

QUESTION 1 RESULT

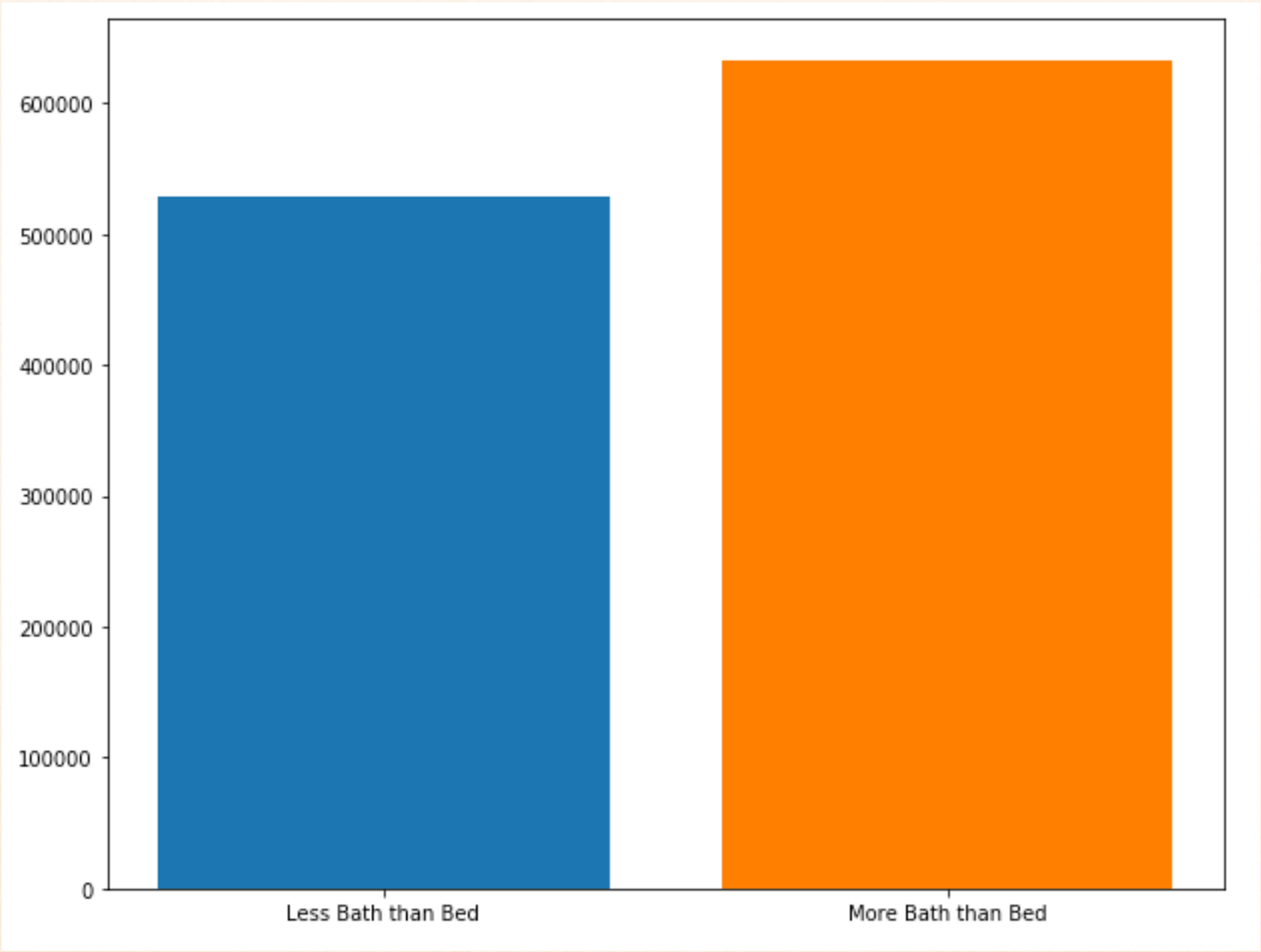
- I created 2 data frames for this and found that:

- I rounded bathrooms up to symbolise room count

- Result:

```
More Bath than Bed tend to cost: $ 633334.3 on average
Less Bath than Bed tend to cost: $ 529287.08 on average
--------------------------------------------------------
The difference in average house group price is: $ 104047.22
```

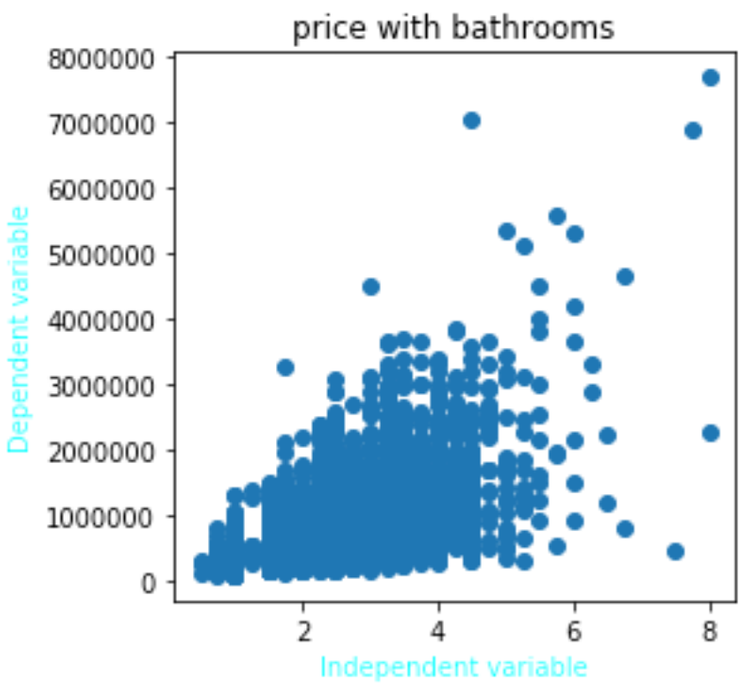
QUESTION 1 VISUAL

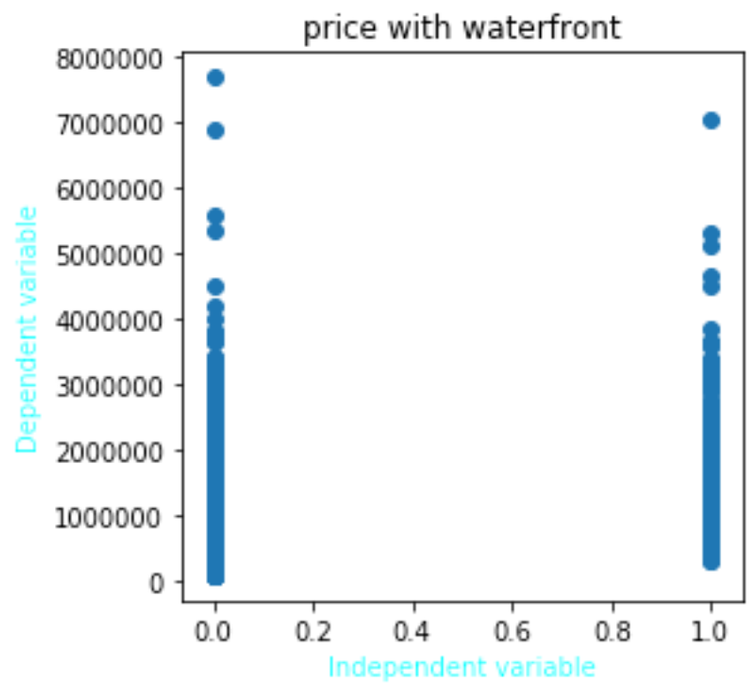


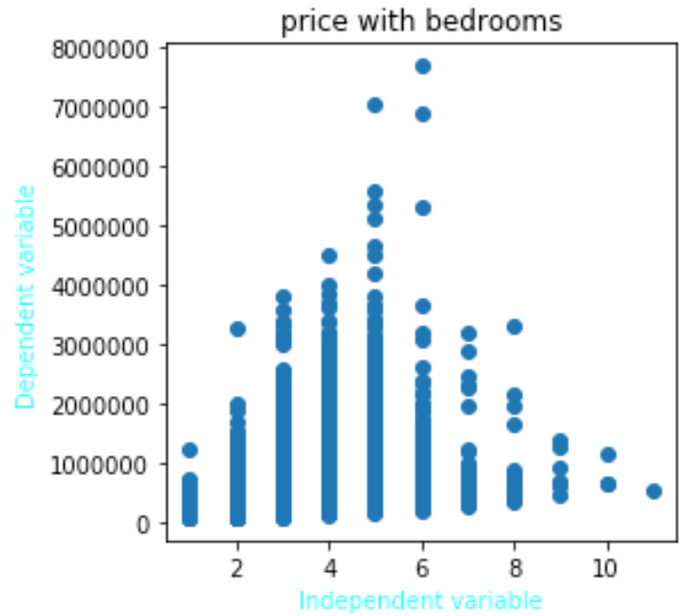
BASELINE REGRESSION MODEL

CORRELATION MATRIX

PRICE RELATIONSHIP







REGRESSION RESULTS

R-squared: 0.540

P>|t|
0.000
0.000
0.000
0.000

Baseline models equation

$$price = 305.60 * \beta_{sqft\_living} + 825,400 * \beta_{waterfront} - 55,555 * \beta_{bedrooms} + 86,270$$

Interpret Coefficients:

- const 86323.3976
  The constant in this equation says that just having a property with nothing else is worth 86,270 dollars
- bedrooms -56028.6573
  The bedrooms coeficient in this equation states that for each additional bedroom your house will lose ~56k dollars in value. This simple means it places its importance on another variable likely sqft_living in the equation
- waterfront 802913.5810
  The waterfront coefficient states that if you house has a waterfront view, it would gain an extra ~800k dollars in value
- sqft_living 306.9650
  The sqft_living coefficient states that for each square foot a house has, it will gain ~306 dollars

CONCLUSION