King County, Washington Real Estate Pricing Analysis

Tara Rosen

Goals:

- to create a real estate pricing model that sets our firm above all others
- create a pricing model that accurately prices our seller's properties so they sell quickly and at their listed price
- create a pricing model that reassures our buyer's that the price of the home they just fell in love with is priced fairly

Data Included In the King County, Washington Property Records

- Over 20,000 records in data set
- Home Sales from May 2014 through May 2015
- Data Includes:

  - Unique identifier for a house
  - Date House Was Sold
  - Price (Prediction Target)
  - Number of Bedrooms
  - Number of Bathrooms
  - Square Footage of the Home
  - Square Footage of the Lot
  - Floors (levels) in House
  - Waterfront view
  - sqft_living15
  - sqft_lot15
  - Has Been Viewed
  - Overall Condition of Property
  - Overall Property Grade
  - Above Grade Square Footage
  - Below Grade Square Footage
  - Year Built
  - Year Renovated
  - Zip Code
  - Latitude Coordinate
  - Longitude Coordinate

Methodology:

- Visually Inspect The Data
  - Look for Missing Values
- Clean The Data
  - Handle Missing Values
  - Drop Irrelevant Data
- Visualize the Data
- Transform the Data
- Test Pricing Model
- Use Pricing Model!

Visually Inspecting the Data

Missing Values

```
price            False
bedrooms         False
bathrooms        False
sqft_living      False
sqft_lot         False
floors           False
waterfront        True
view              True
condition        False
grade            False
sqft_above       False
sqft_basement    False
yr_built         False
yr_renovated      True
zipcode          False
lat              False
long             False
sqft_living15    False
sqft_lot15       False
dtype: bool
```

```
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
```

Data Features

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21597 entries, 0 to 21596
Data columns (total 21 columns):
id               21597 non-null int64
date             21597 non-null object
price            21597 non-null float64
bedrooms         21597 non-null int64
bathrooms        21597 non-null float64
sqft_living      21597 non-null int64
sqft_lot         21597 non-null int64
floors           21597 non-null float64
waterfront       19221 non-null float64
view             21534 non-null float64
condition        21597 non-null int64
grade            21597 non-null int64
sqft_above       21597 non-null int64
sqft_basement    21597 non-null object
yr_built         21597 non-null int64
yr_renovated     17755 non-null float64
zipcode          21597 non-null int64
lat              21597 non-null float64
long             21597 non-null float64
sqft_living15    21597 non-null int64
sqft_lot15       21597 non-null int64
dtypes: float64(8), int64(11), object(2)
memory usage: 3.5+ MB
```

Handling The Missing/Questionable Data:

- Removed the rows with missing values in the view column
- Replaced the missing values in the waterfront column with zeros
- Removed the yr_renovated, id and date column
- Replaced '?' in sqft_basement column with zeros

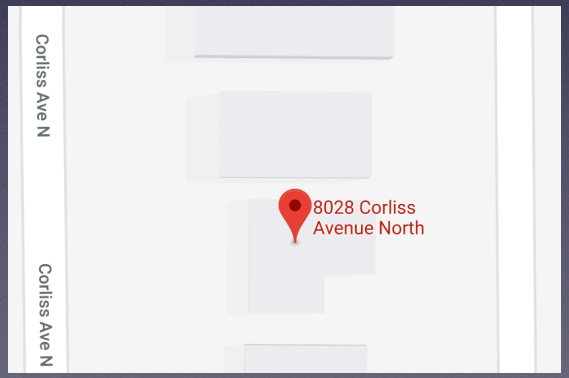
Confirmation of Missing Data Corrected

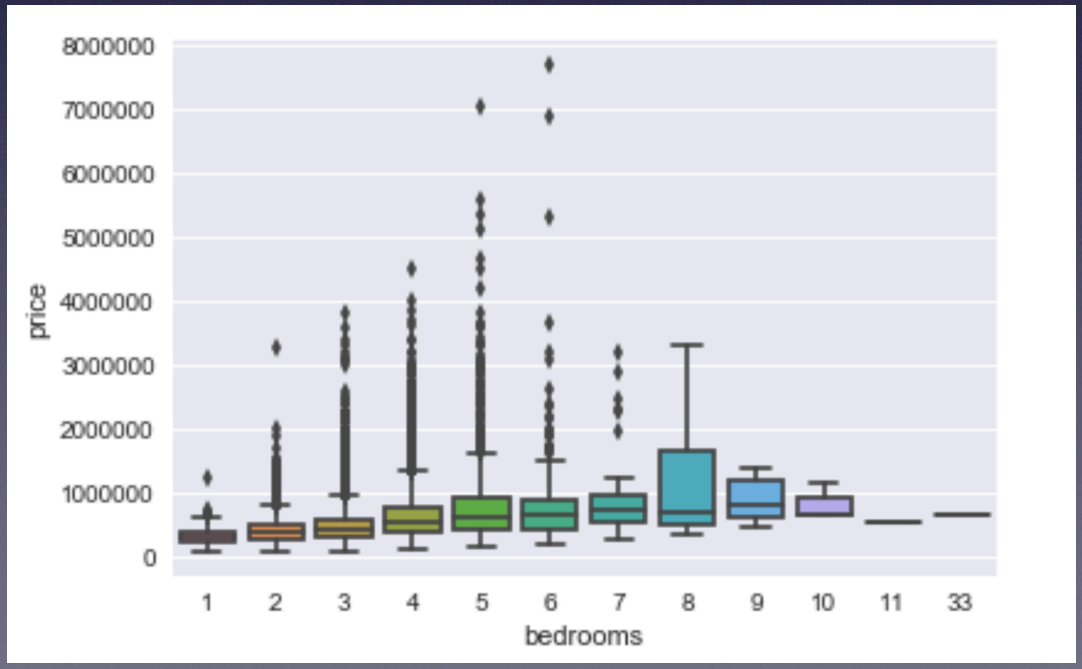
```
price            False
bedrooms         False
bathrooms        False
sqft_living      False
sqft_lot         False
floors           False
waterfront       False
view             False
condition        False
grade            False
sqft_above       False
sqft_basement    False
yr_built         False
zipcode          False
lat              False
long             False
sqft_living15    False
sqft_lot15       False
dtype: bool
```

```
price            0
bedrooms         0
bathrooms        0
sqft_living      0
sqft_lot         0
floors           0
waterfront       0
view             0
condition        0
grade            0
sqft_above       0
sqft_basement    0
yr_built         0
zipcode          0
lat              0
long             0
sqft_living15    0
sqft_lot15       0
dtype: int64
```

Deeper Dive Into Data

33 Bedrooms?


Corliss Ave N
8028 Corliss Avenue North
Corliss Ave N


8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
price
1
2
3
4
5
6
7
8
9
10
11
33
bedrooms

```
housing_df.corr()   #.308787
```

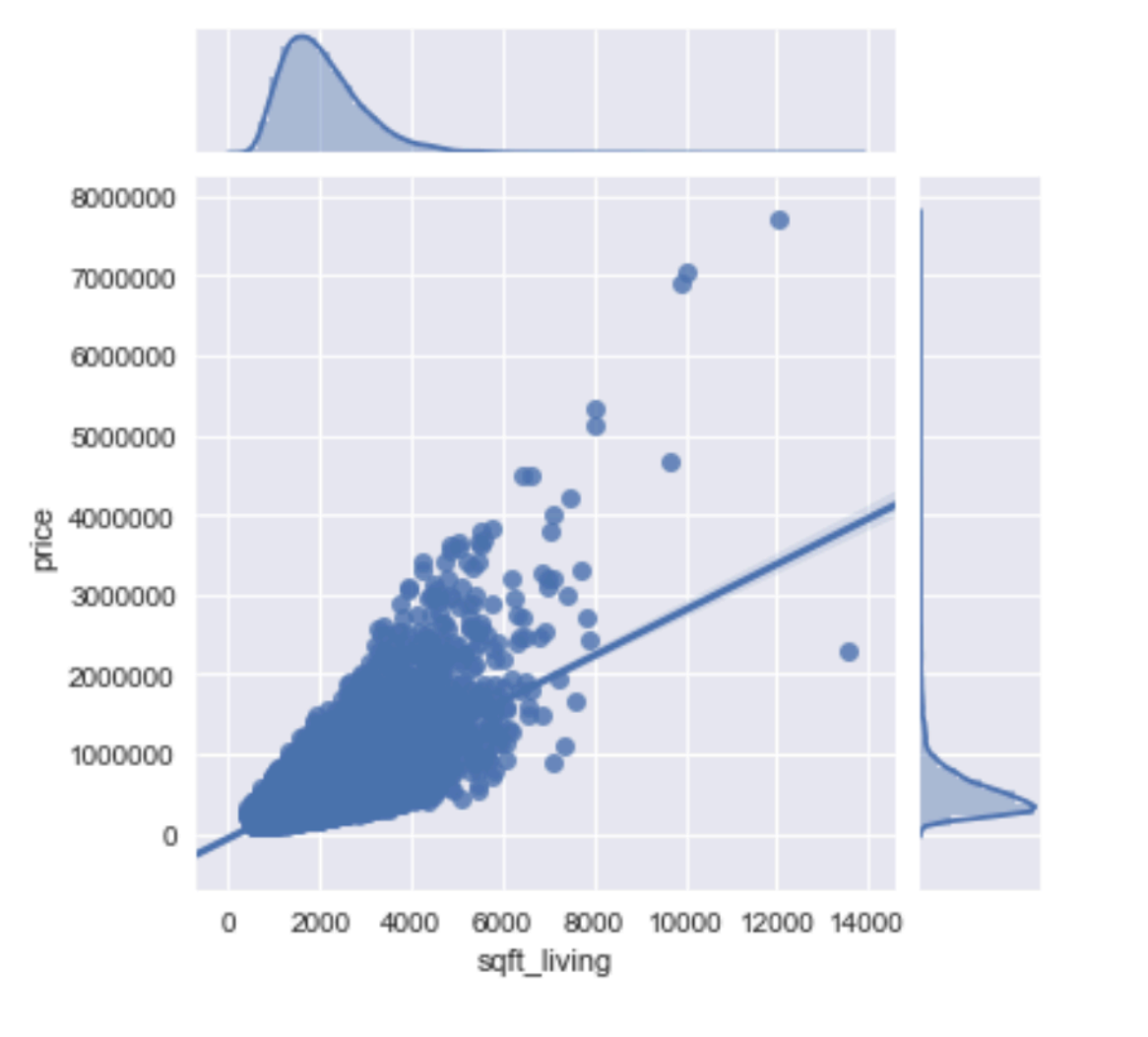
	price	bedrooms	ba
price	1.000000	0.308787	
bedrooms	0.308787	1.000000	
bathrooms	0.525906	0.514508	
sqft_living	0.701917	0.578212	

```
housing_df.corr()      #0.315954
```

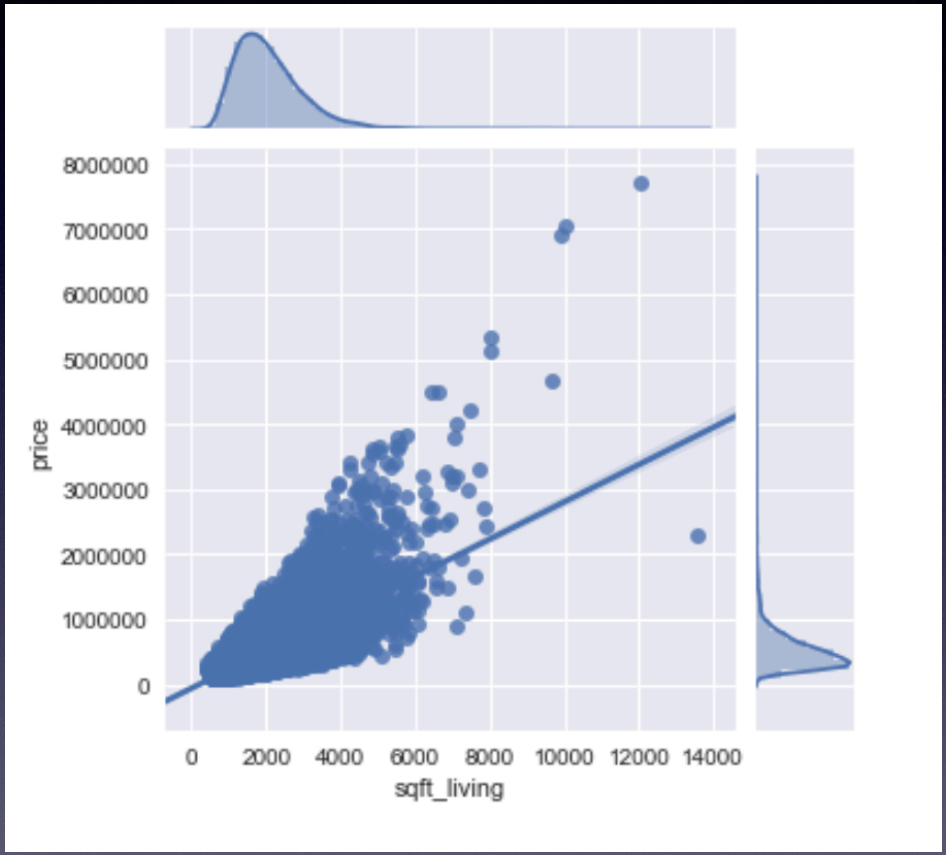
	price	bedrooms	bathro
price	1.000000	0.315954	0.525
bedrooms	0.315954	1.000000	0.527
bathrooms	0.525906	0.527874	1.000

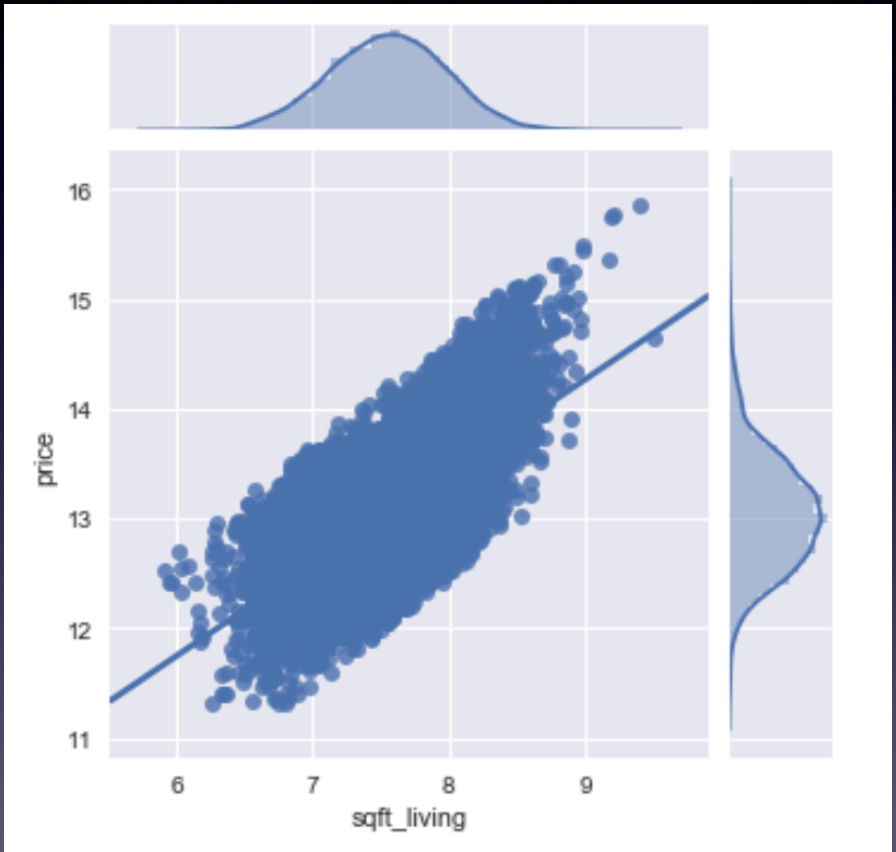
Improvement in Correlation Between Price and Bedrooms

Another Example of Data Inspection



Transforming the Data


8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
price
0
2000
4000
6000
8000
10000
12000
14000
sqft_living


16
15
14
13
12
11
price
6
7
8
9
sqft_living

Rinse
Lather
Repeat
(break)
with
Data

pricing model will satisfy our customers and improve our sales