

# Module 1 Data Science Project: Housing

David Goldstein

[dgoldstein24@gmail.com](mailto:dgoldstein24@gmail.com)

[dg2996@columbia.edu](mailto:dg2996@columbia.edu)

# Obtain: KC Housing Data

Target  
Variable

Predictor  
Variables

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	gr
0	7129300520	10/13/2014	221900.0	3	1.00	1180	5650	1.0	NaN	0.0	...	
1	6414100192	12/9/2014	538000.0	3	2.25	2570	7242	2.0	0.0	0.0	...	
2	5631500400	2/25/2015	180000.0	2	1.00	770	10000	1.0	0.0	0.0	...	
3	2487200875	12/9/2014	604000.0	4	3.00	1960	5000	1.0	0.0	0.0	...	
4	1954400510	2/18/2015	510000.0	3	2.00	1680	8080	1.0	0.0	0.0	...	

5 rows × 21 columns

# Scrubbing: Changing columns with string values

```
<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)
```

Not relevant

Change "?" to median value

## Scrubbing: Dealing with null values

id	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
----------	---

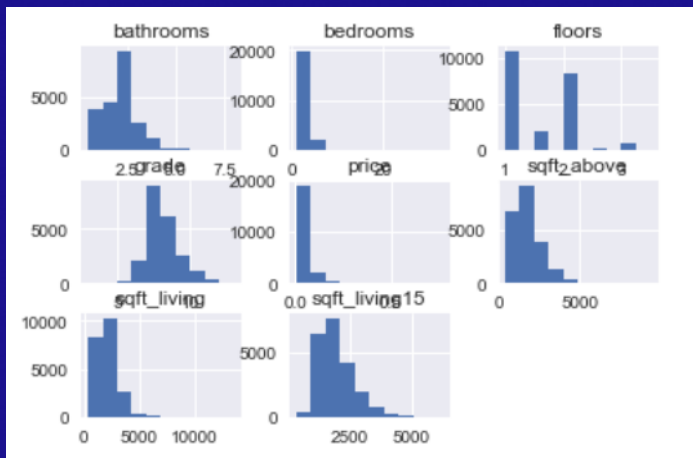
yr_renovated	3842
--------------	------

Make null its own bin

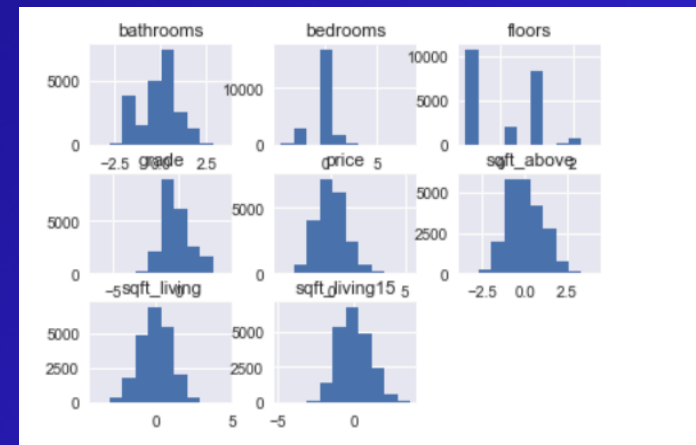
Change to median

Too many 0 values

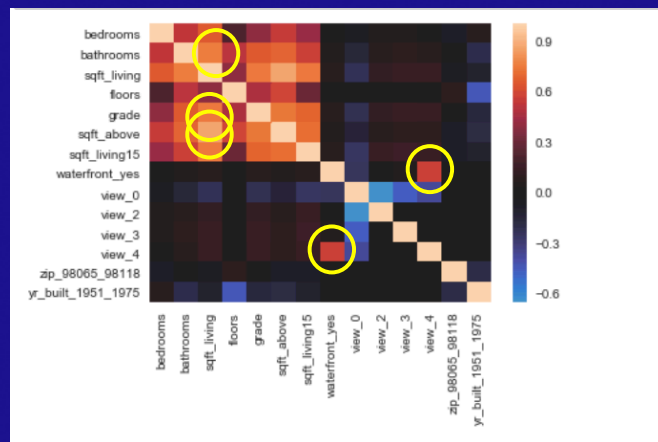
# Scrubbing: Transforming numerical categories



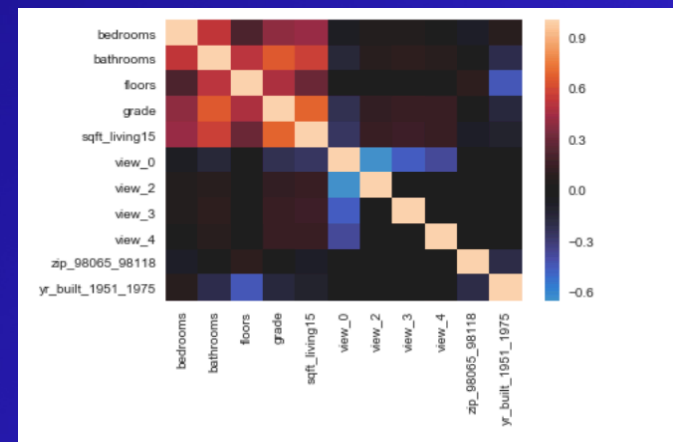
1. Log
2. Standardize



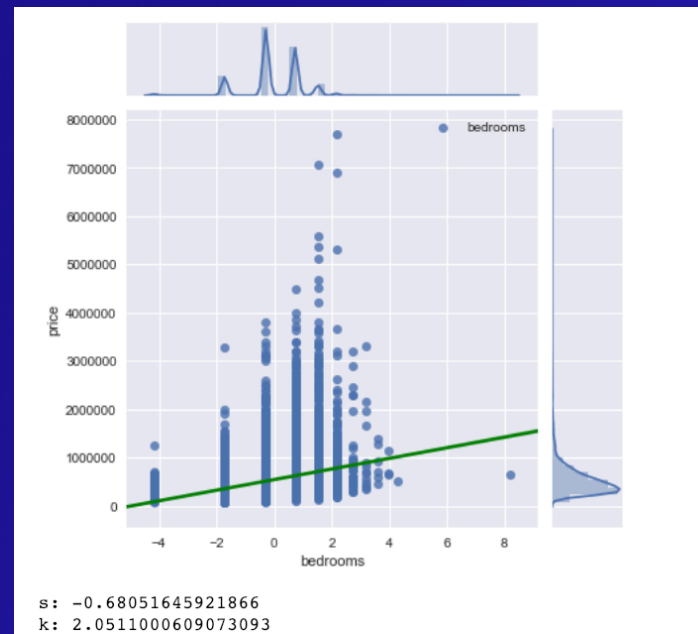
# Scrubbing: Reducing collinearity



Collinearity < .75



# Explore: Correlations, Skewness, and Kurtosis






# Model: Least Squares Regression

OLS Regression Results

<b>Dep. Variable:</b>	price	<b>R-squared:</b>	0.505
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.505
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	2000.
<b>Date:</b>	Tue, 01 Jan 2019	<b>Prob (F-statistic):</b>	0.00
<b>Time:</b>	14:28:28	<b>Log-Likelihood:</b>	-2.9980e+05
<b>No. Observations:</b>	21597	<b>AIC:</b>	5.996e+05
<b>Df Residuals:</b>	21585	<b>BIC:</b>	5.997e+05
<b>Df Model:</b>	11		
<b>Covariance Type:</b>	nonrobust		



## Validation: Alternative Models

1. Regression Model without View columns 
2. Regression Model without columns with (-) kurtosis 
3. Best model: original model – r-squared = 50.5% 

## Interpret: Recommendations

1. Rec 1: to increase housing price by about \$15,780,  
build another bedroom
2. Rec 2: to increase housing price by about \$13,200,  
build another bathroom



## Interpret: Project Takeaways

1. OSEMiN is a helpful data science process
2. Scrubbing makes up bulk of work
3. Multicollinearity table and graph – very useful



Thank You

David Goldstein