

Introduction	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Tha nk...
--------------	------	---------	-------------	----------------	------	------	------------	----------	------------	--------------




The housing story

Introduction	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Tha nk...
--------------	------	---------	-------------	----------------	------	------	------------	----------	------------	--------------



MERCEDES SANTANA
RODRIGUEZ




NIRANJAN KONDO

The People
Behind the
Project


Introduction	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Tha nk...
--------------	------	---------	-------------	----------------	------	------	------------	----------	------------	--------------

Problem



What factors impact the house prices in Seattle?

Let's use a dataset of 21597 records



Introduction	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Tha nk...
--------------	------	---------	-------------	----------------	------	------	------------	----------	------------	--------------

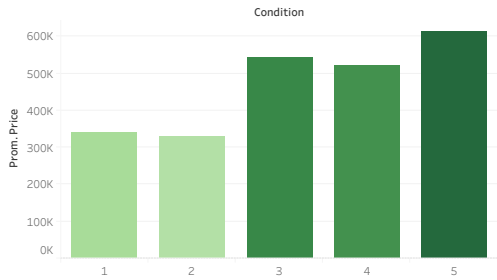
Initial Assumptions

Positive	Unsure	Negative
Area in square feet	Waterfront	Year built
Basement	View	
Zip code	Bathrooms	
Distance from city center	Sold date	
Floor	House id	

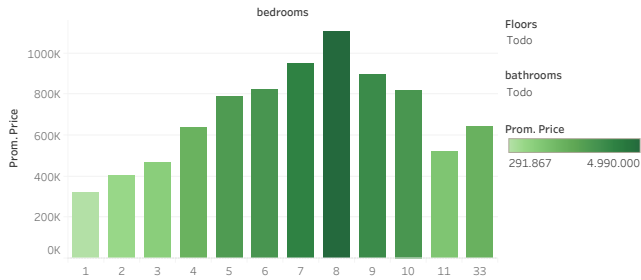
The housing story

Introduction	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Tha nk...
--------------	------	---------	-------------	----------------	------	------	------------	----------	------------	--------------

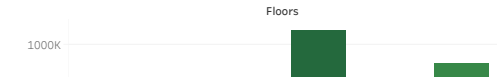
Price - Condition



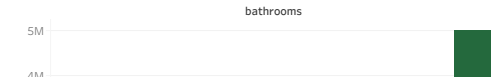
Price - Bedroom



Price - Floor



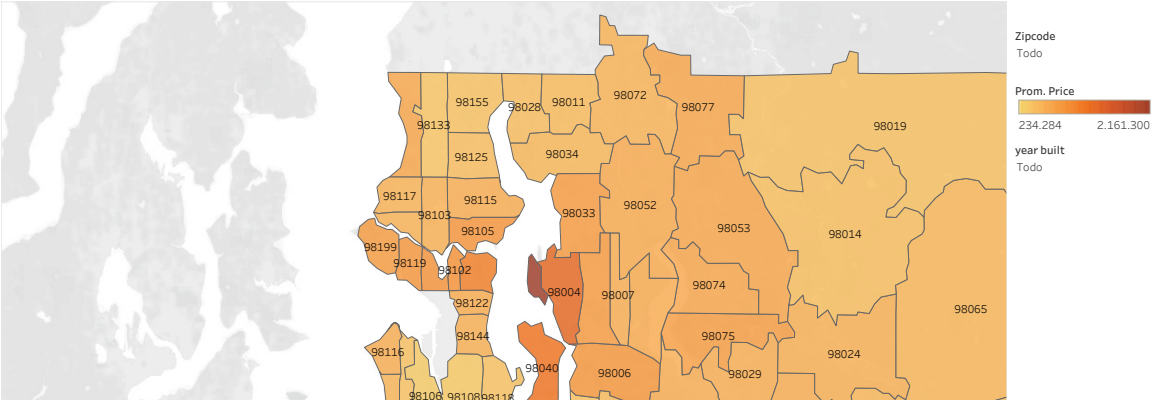
Price - Bathroom



The housing story

Introduction	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Thank You
--------------	------	---------	-------------	----------------	------	------	------------	----------	------------	-----------

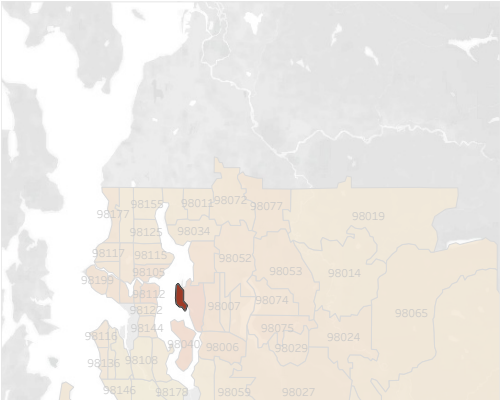
Mapping ZipCodes



The housing story

Intr od...	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Thank You
---------------	------	---------	-------------	----------------	------	------	------------	----------	------------	-----------

Mapping ZipCodes

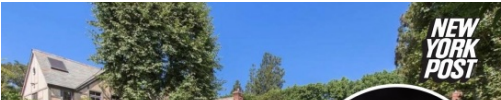


Prom. Price
234.284,035175879 a 2...

Zipcode
Todo

Prom. Price
234.284 2.161.300

year built
Todo



Intr od..	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Thank You
-----------	------	---------	-------------	----------------	------	------	------------	----------	------------	-----------

0.82

R2 Ridge

0.82

R2 Lasso

0.821

Regression Model

OLS Regression Results

Dep. Variable:	price	R-squared:	0.821
Model:	OLS	Adj. R-squared:	0.819
Method:	Least Squares	F-statistic:	720.4
Date:	Thu, 18 Nov 2021	Prob (F-statistic):	0.00
Time:	11:09:31	Log-Likelihood:	-2.2725e+05
No. Observations:	16964	AIC:	4.547e+05
Df Residuals:	16856	BIC:	4.556e+05
Df Model:	107		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	8.99e+05	3.9e+04	23.074	0.000	8.23e+05	9.75e+05
distance_from_seattle	-1.085e+04	230.541	-43.612	0.000	-1.05e+04	-9602.483
med_sqft_lot15_trans	4.034e+04	1981.795	20.354	0.000	3.65e+04	4.42e+04
med_sqft_living15_trans	3.564e+04	2077.493	17.156	0.000	3.16e+04	3.97e+04
bedrooms_2	3.246e+04	1.23e+04	2.648	0.008	8435.397	5.65e+04
bedrooms_3	6.138e+04	1.21e+04	5.075	0.000	3.77e+04	8.51e+04

Intr od..	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Thank You
--------------	------	---------	-------------	----------------	------	------	------------	----------	------------	-----------

Outcomes

Positive
Area in square feet
Basement
Zip code
Distance from city
center
Curb

Unsure
Year renovated
Year built
Floor
Bedroom

Negative



Intr od..	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Thank You
--------------	------	---------	-------------	----------------	------	------	------------	----------	------------	-----------

Conclusions

- City center neighborhoods have high price per sqft. E.g. code: 98039.
- Geographic information was the most important feature in our model.
- Houses with 2.5 floor has the highest selling price.
- Houses with 8 bedrooms have the highest selling price.
- Condition 3 & 5 have the highest selling price.




Intr od..	Team	Problem	Assumptions	Price analysis	Map1	Map2	Regression	outcomes	Conclusion	Thank You
-----------	------	---------	-------------	----------------	------	------	------------	----------	------------	-----------

Contents

- 1 Midterm-Project: Housing Market Study 2014-15
 - 1.1 Loading Data
 - 1.2 Final data set
 - 1.3 Important Functions
 - 1.4 Data type
 - 1.5 Removing duplicates
 - 1.6 Creating data groups
 - 1.6.1 Convert Integer into categorical
 - 1.7 Correlation analysis
 - 1.8 Removing outliers
 - 1.8.1 Numerical variables
 - 1.8.2 Outplots
 - 1.8.2.1 Removing outliers from "log1_jct15"
 - 1.8.2.2 Removing outliers from "log1_living15"
 - 1.8.2.3 Removing outliers from "distance_from_seattl"
 - 1.8.3 Normality (numerical variables)
 - 1.9 Power transformation
 - 1.9.1 Power transformation in "log1_jct15"
 - 1.9.2 Power transformation in "log1_living15"
 - 1.9.3 Data frame with transformed variables
 - 1.10 Predictive modeling
 - 1.10.1 XY split
 - 1.10.2 Hot coding categorical variables
 - 1.11 Linear regression model
 - 1.11.1 Unscaled
 - 1.11.1.1 Lineal regression model without irrelevant var
 - 1.11.1.2 Most important coefficients in our model
 - 1.11.2 Standardize
 - 1.11.2.1 Scaled Lineal regression model without irrele
 - 1.11.3 Model comparison
 - 1.11.3.1 Lineal regression with Scikit Learn
 - 1.11.3.2 Lasso
 - 1.11.3.3 Ridge

1 Midterm-Project: Housing Market Study 2014-15

```
In [2]: # we need to import the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.collections import LineCollection
%matplotlib inline
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import OneHotEncoder
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import PowerTransformer
from sklearn.linear_model import LinearRegression
from scipy.stats import boxcox
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error as mse
from sklearn.metrics import mean_absolute_error as mae
pd.options.display.max_rows = 50
from IPywidgets import interactive
import statsmodels.api as sm
from sklearn.preprocessing import MinMaxScaler
import math
from sklearn import linear_model
from sklearn.linear_model import Ridge
from haversine import haversine
import joblib
```


GitHub

1.1 Loading Data