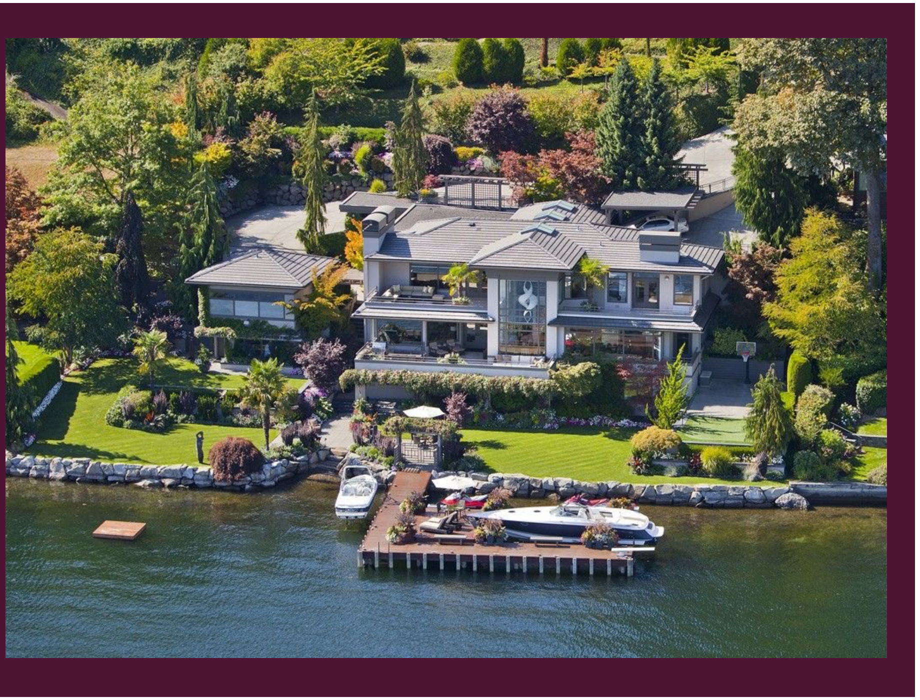
REAL ESTATE PRICES IN KING COUNTY 2014 – 2015

IRONHACK D.A. BOOTCAMP JAN.2022, MIDTERM PROJECT

ALEJANDRA PARRA, ADRIANA CUPPULERI & ODELIA AHDOUT



RESEARCH QUESTIONES

- Which factors are the ones responsible for house prices in King Counry, WA?
- The goal of the project is to design a model which would predict **selling house-price** from **a set of features used to evaluate the property**.

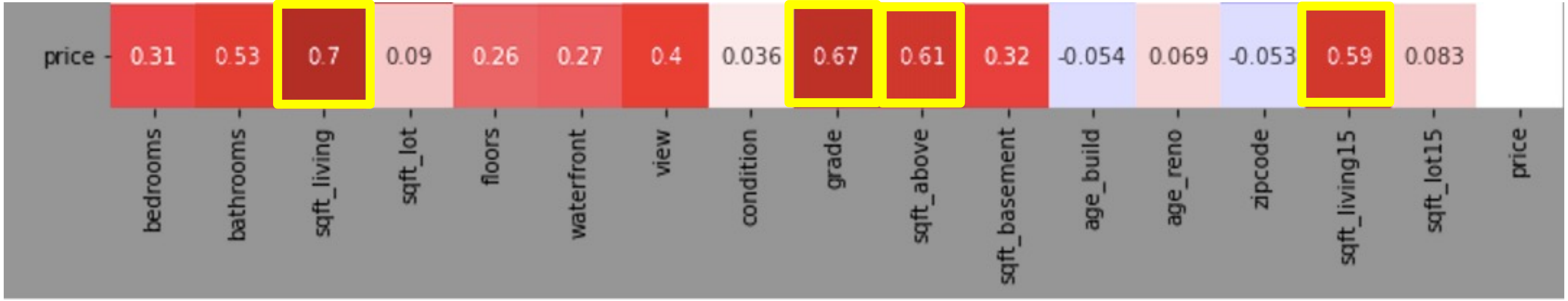
THE DATA BASE

- Consists of information on roughly **22,000 properties** in King County, WA, sold between May 2014 and May 2015.
  - ➢ No missing values
  - ➢ No duplicates
- Dropped ID, date, longitude & latitude columns

Feature	Classification	Scale/Range	Type (self determined)
ID	General		CAT
Date	General		CAT
Bedrooms	Distribution of living space	1 - 11 {33}	CAT
Bathrooms	Distribution of living space	0.5 - 8	CAT
Sqft_living	Size		#
sqft_lot	Size		#
Floors	Distribution of living space	1 - 3.5	CAT
Waterfront	Surroundings	1/0 [Yes/No]	CAT
View	Surroundings	0-4	CAT
Condition	Quality Rating	1-5	CAT
Grade	Quality Rating	1-13	CAT
sqft_above	Size		#
sqft_basement	Size		#
yr_built	Age	1900 - 2015	#
yr_renovated	Renovated?	0 [No] / 1943 - 2014	#
zipcode	Location		🌐
lat	Location		🌐
long	Location		🌐
sqft_living15	Size		#
sqft_lot15	Size		#
Price	Dep. Variable		#

INITIAL ASSUMPTIONS

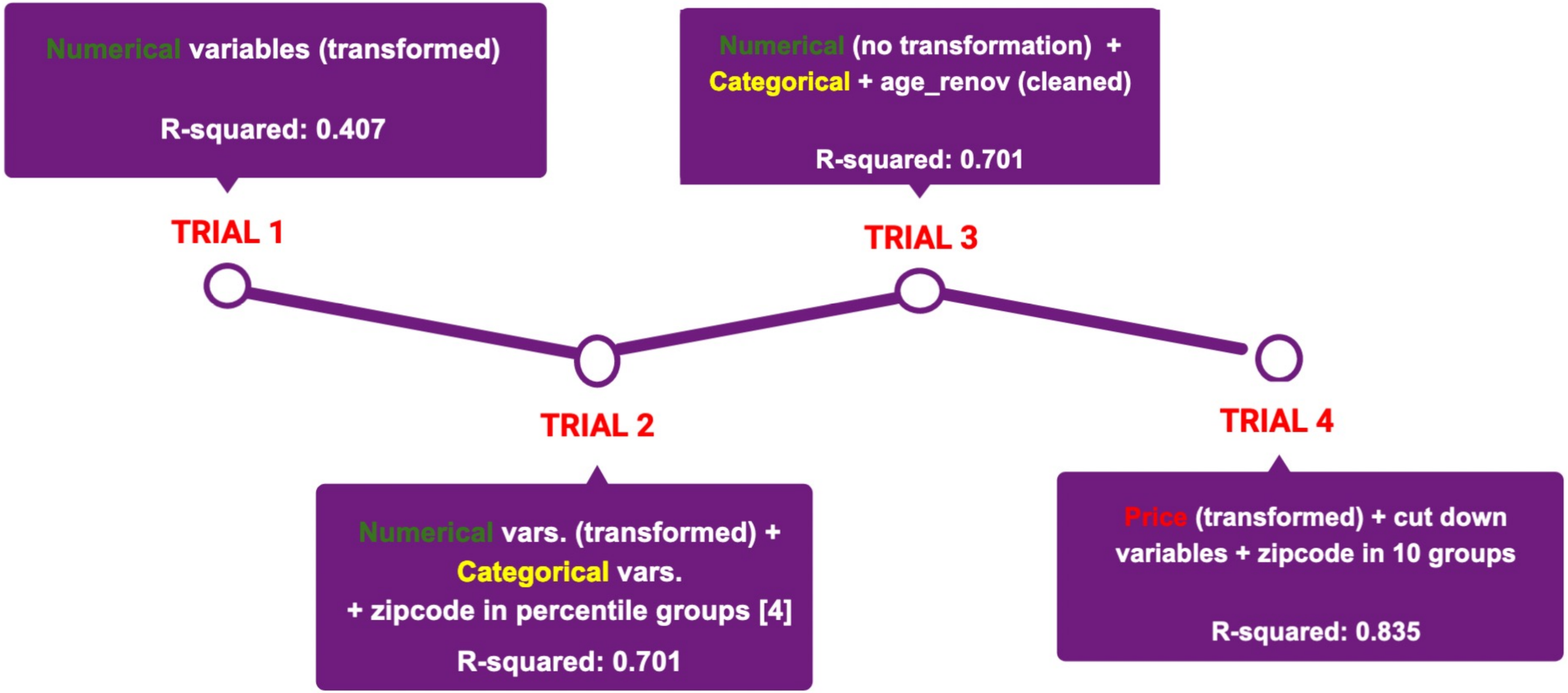
- Preliminary correlation matrix:



	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	age_build	age_reno	zipcode	sqft_living15	sqft_lot15	price
price	0.31	0.53	0.7	0.09	0.26	0.27	0.4	0.036	0.67	0.61	0.32	-0.054	0.069	-0.053	0.59	0.083	

- **Basic assumptions:**
  - Based on correlations: most influential features are <u>size</u> and <u>grade</u> <=> positively correlated with price
  - Based on general knowledge: location is important! <=> <u>zipcode</u> should be correlated with price

DATA PROCESSING PIPELINE



Numerical variables (transformed)

R-squared: 0.407

TRIAL 1

TRIAL 2

Numerical vars. (transformed) + Categorical vars.
+ zipcode in percentile groups [4]

R-squared: 0.701

Numerical (no transformation) + Categorical + age_renov (cleaned)

R-squared: 0.701

TRIAL 3

TRIAL 4

Price (transformed) + cut down variables + zipcode in 10 groups

R-squared: 0.835

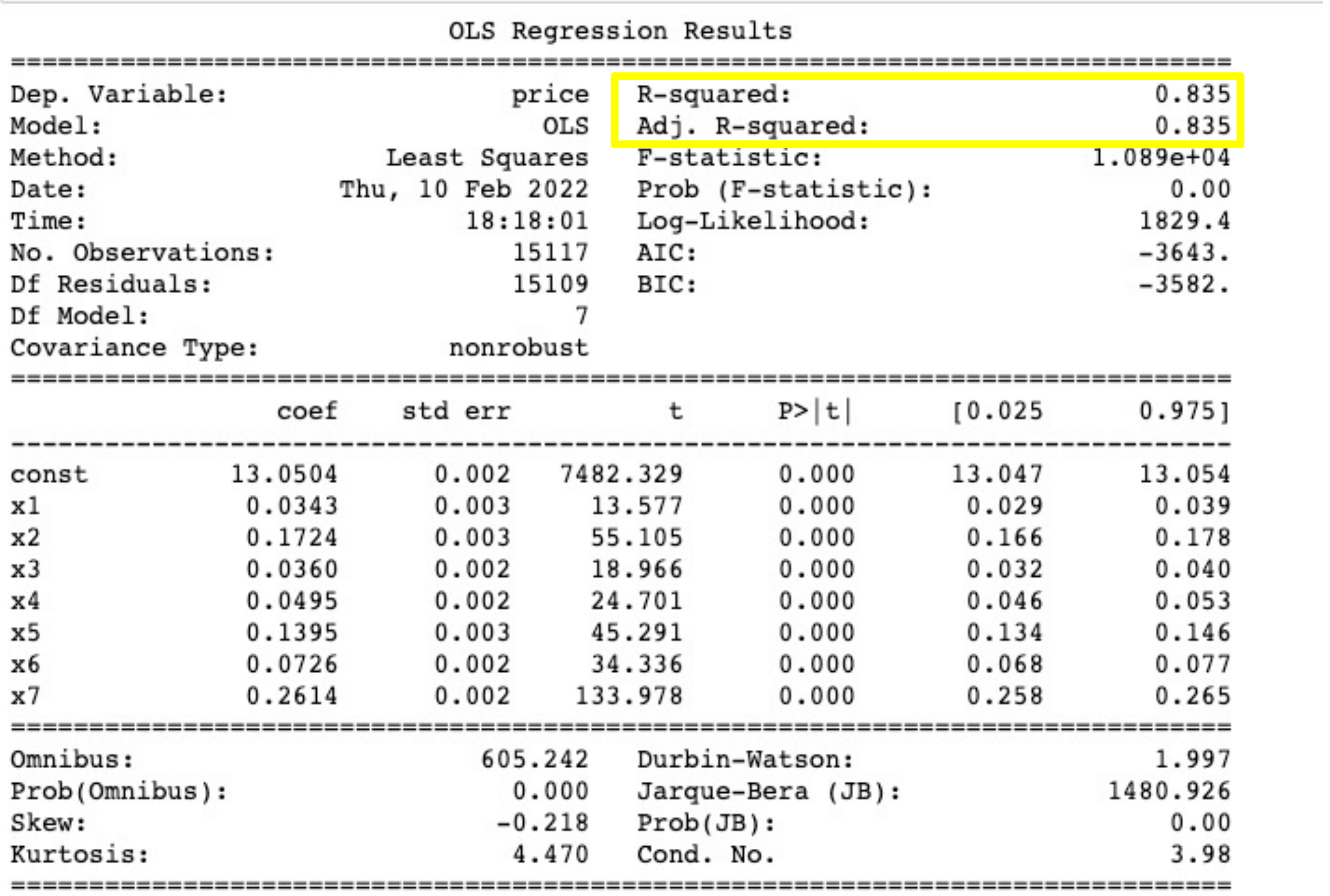
SELECTED FEATURES (FINAL MODEL)

Feature	Classification	Scale/Range	Type (self determined)
ID	General		CAT
Date	General		CAT
Bedrooms	Distribution of living space	1 - 11 {33}	CAT
Bathrooms	Distribution of living space	0.5 - 8	CAT
Sqft_living	Size		#
sqft_lot	Size		#
Floors	Distribution of living space	1 - 3.5	CAT
Waterfront	Surroundings	1/0 [Yes/No]	CAT
View	Surroundings	0-4	CAT
Condition	Quality Rating	1-5	CAT
Grade	Quality Rating	1-13	CAT
sqft_above	Size		#
sqft_basement	Size		#
Age_build	Age	1900 - 2015	#
yr_renovated	Renovated?	0 [No] / 1943 - 2014	#
Percentile_zip	Location	1-10	CAT
lat	Location		🌐
long	Location		🌐
sqft_living15	Size		#
sqft_lot15	Size		#
Price	Dep. Variable		#

FINDINGS

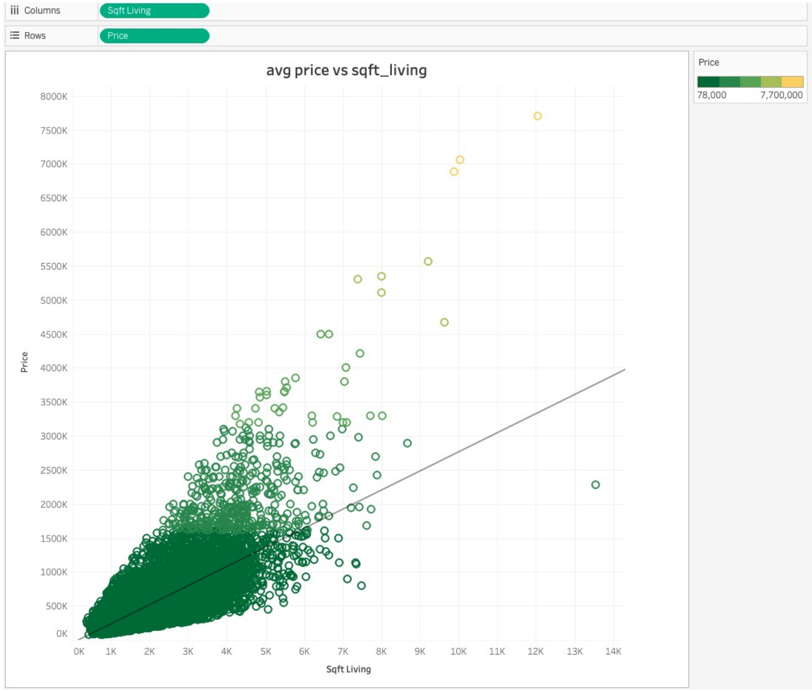
- Trial 4 (final model) – linear regression

Rank	Feature	Classification	Type
#1	**Percentile_zip**	Location	CAT
#2	**Sqft_living**	Size	#
#3	**Grade**	Quality Rating	CAT
#4	**Age_build**	Age	CAT
#5	**View**	Surroundings	CAT
#6	**Waterfront**	Surroundings	CAT
#7	**Bathrooms**	Distribution of living space	CAT
	Price	Dep. Variable	#

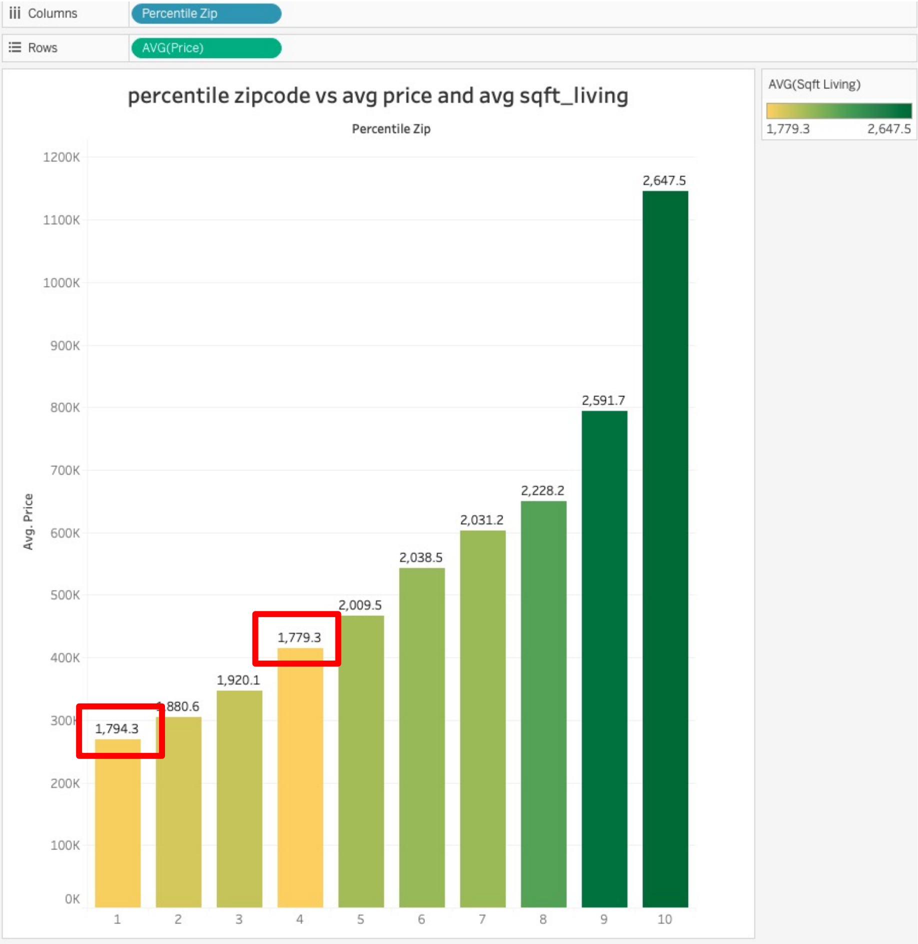


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.835
Model:                            OLS   Adj. R-squared:                  0.835
Method:                 Least Squares   F-statistic:                 1.089e+04
Date:                Thu, 10 Feb 2022   Prob (F-statistic):               0.00
Time:                        18:18:01   Log-Likelihood:                 1829.4
No. Observations:               15117   AIC:                            -3643.
Df Residuals:                   15109   BIC:                            -3582.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.0504      0.002   7482.329      0.000      13.047      13.054
x1             0.0343      0.003     13.577      0.000       0.029       0.039
x2             0.1724      0.003     55.105      0.000       0.166       0.178
x3             0.0360      0.002     18.966      0.000       0.032       0.040
x4             0.0495      0.002     24.701      0.000       0.046       0.053
x5             0.1395      0.003     45.291      0.000       0.134       0.146
x6             0.0726      0.002     34.336      0.000       0.068       0.077
x7             0.2614      0.002    133.978      0.000       0.258       0.265
==============================================================================
Omnibus:                      605.242   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1480.926
Skew:                          -0.218   Prob(JB):                         0.00
Kurtosis:                       4.470   Cond. No.                         3.98
==============================================================================
```

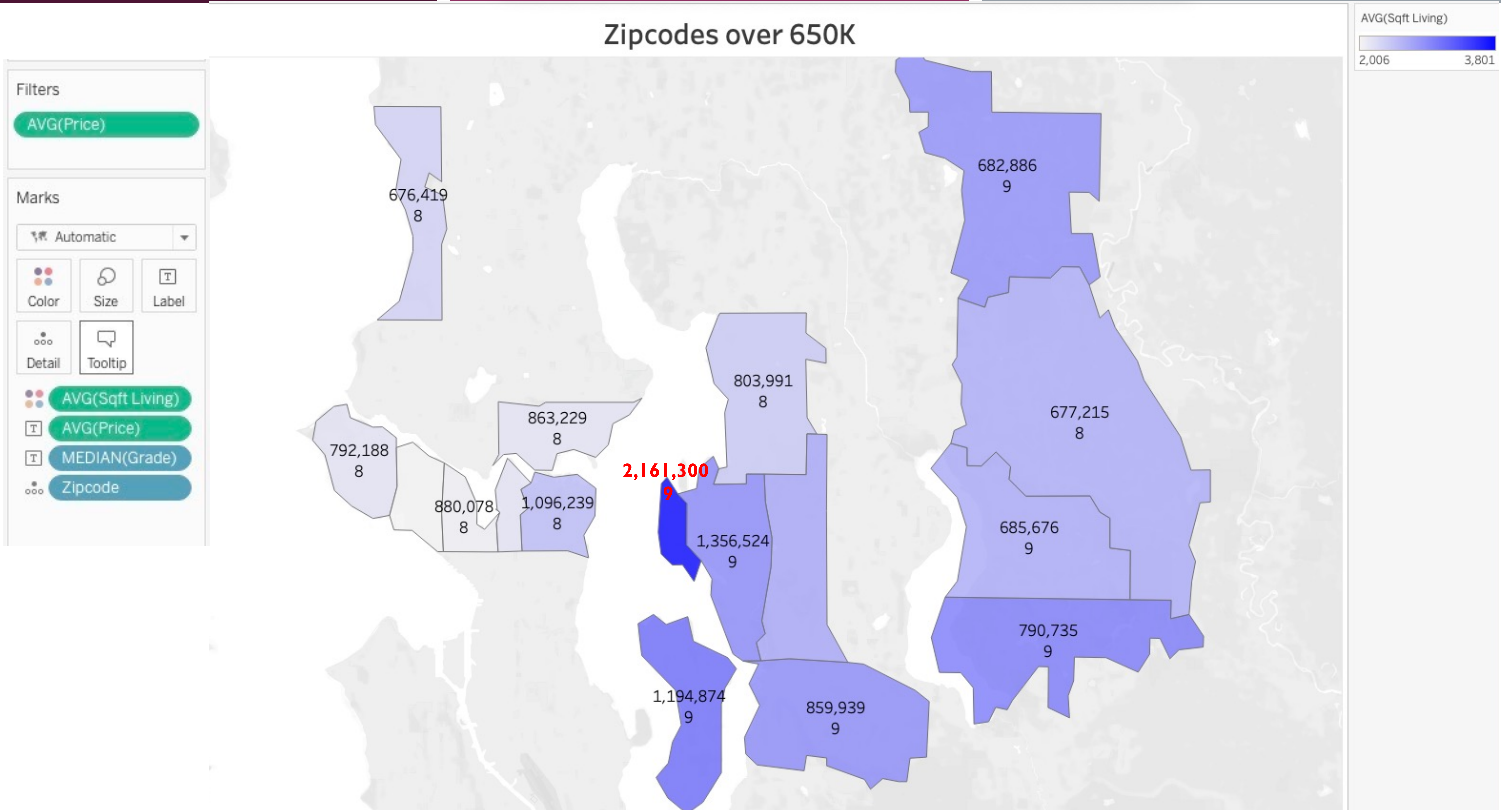
REVISITING INITIAL ASSUMPTIONS



- ➢ <u>Location</u> proved to be the most important feature for predicting the selling house price
  - ○ Interaction between price and location?



- ➢ <u>Size</u> in and of itself is not the best predictor of the selling price



Zipcodes over 650K

Filters

AVG(Price)

Marks

Automatic

AVG(Sqft Living)

AVG(Price)

MEDIAN(Grade)

Zipcode

POSSIBLE IMPROVEMENTS

- Interpreting error metrics is tricky when the dependent variable is transformed
- Feature importance – what should one do when the coefficient scales are different?

THANKS!

- ✓ Rafa
- ✓ Nelson and Kike
- ✓ Everyone who gave advice, shared insights, and helped deal with Tableau