King County House Prices Data Analysis

Title	Introduction	Structure	Executive Summary	Geographic Distribution of Houses

BY: OLISA UZONDU
FOR: SPRINGBOARD

King County House Prices Data Analysis

PURPOSE: **To create a predictive model that estimates pricing for houses in King County**

WHERE: King County, a county in Washington state situated in **an area with a popular city**, Seattle

OPPORTUNITY CASE: With a pricing model, **real estate firms and investors** would find it easier to make housing investment decisions, thus **attracting more business opportunity** to the area

DELIVERABLES: **Descriptive analysis** of house prices, **inferential analysis, predictive price model**

King County House Prices Data Analysis

1. Geographic Price Distribution
   - High-value areas/low-value areas
   - Price ranges

2. Geographic Average House Price distribution

3. Distribution of Prices
   - Condition
   - Grade
   - Bedroom
   - Bathroom

4. Age of Houses
   - Old/New
   - Condition
   - Grade

5. Price Relationship to Variables
Correlated/Not correlated?

6. Predictive Price Model

King County House Prices Data Analysis

* Houses with **good views and/or a waterfront area** are typically **higher priced** (**$5M price range with an average of $1.65M**)

* House **prices are more clustered around the $200k - $800k range (82.5% probability)**

* **Good quality** (6 - 13 out of 13 grade levels) **and in good condition** (3 - 5 out of 5 condition levels) houses are expected to be found in houses of between **0 - 70 years old** (age range of houses between 0 and 115 years) (**83% probability**)

* Variables used for prediction models:
    a. sqft_living
    b. bedroom
    c. bathroom
    d. grade
    e. condition

* 2 **price prediction models** created:
    a. Waterfront area
    b. Non-waterfront area

* **Linear regression model not a good fit**

* **Other more suitable model tests required to accurately predict house prices**

King County House Prices Data Analysis

Structure | Executive Summary | Geographic Distribution of Houses | House Price Distribution | House Quality

Houses with **good views and/or a waterfront area** are typically **higher priced** (**$5M price range with an average of $1.65M**)

Price bin
All

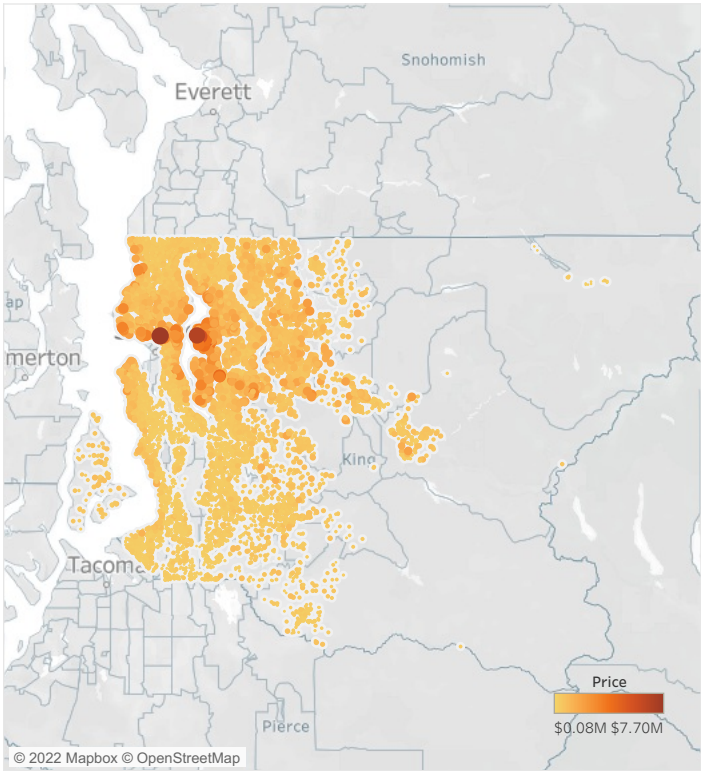
Sqft Living
All values

View
All

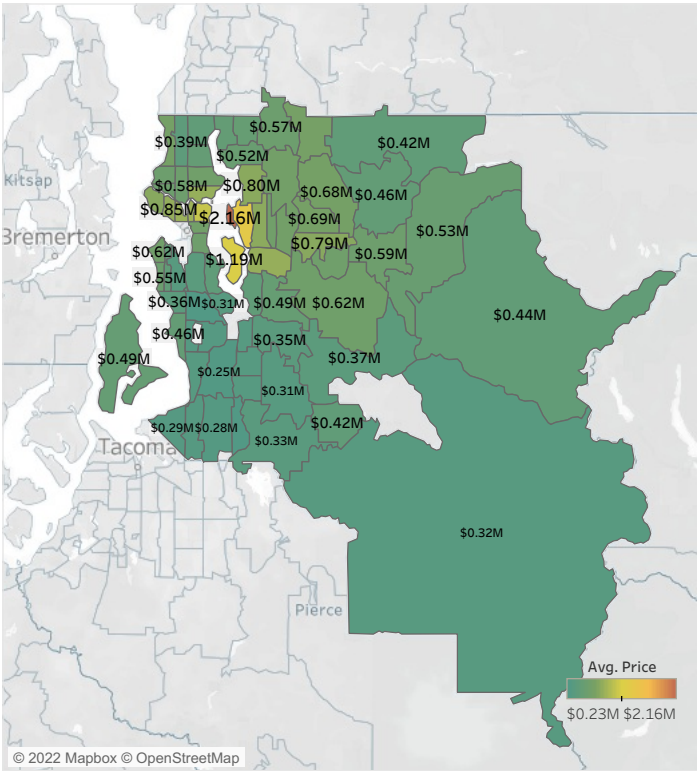
Waterfront (0 - No, 1 ..
☑ 0
☐ 1

Average Price
All values

Geographic Distribution of Houses Sold



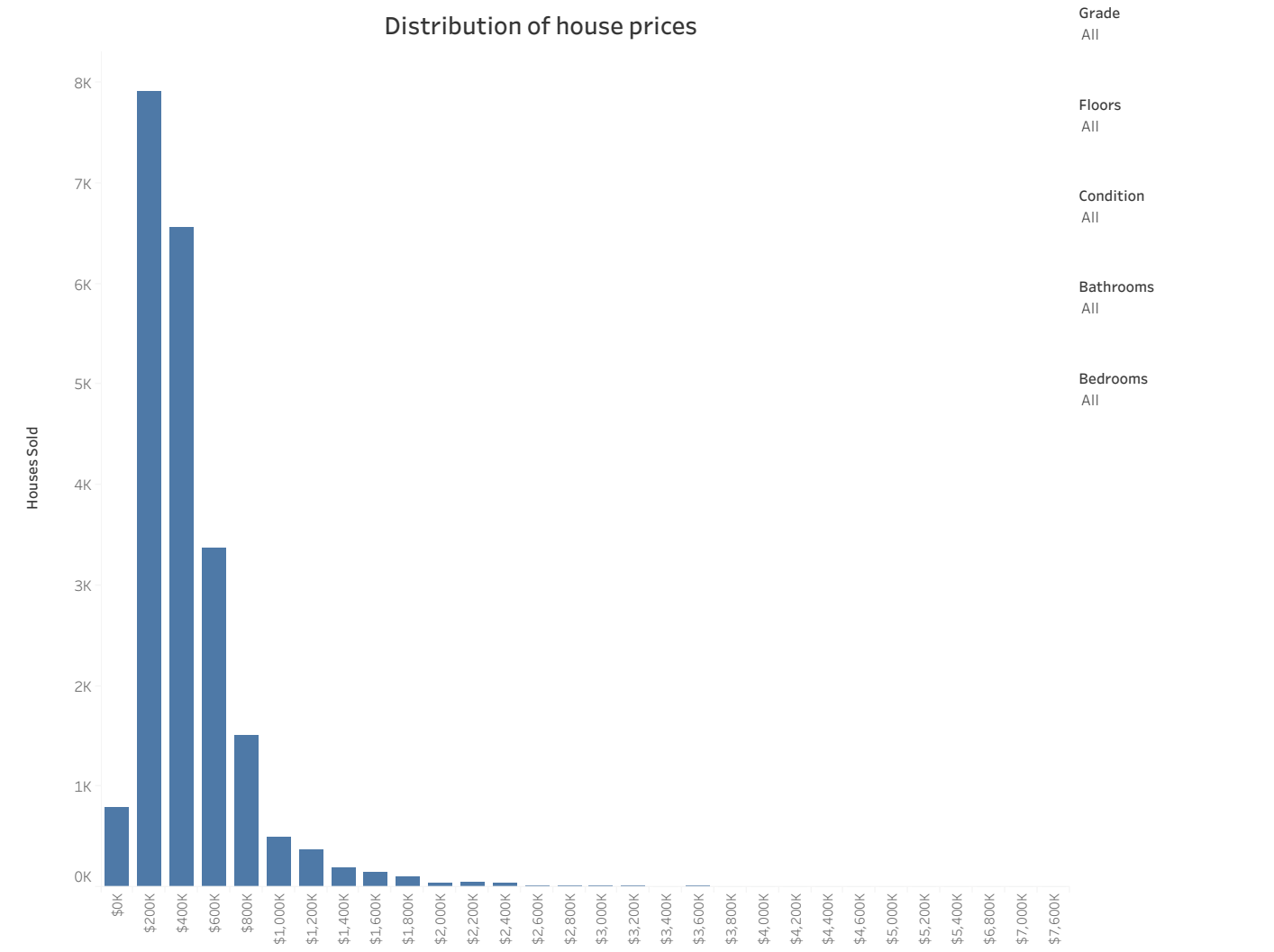
Average Prices of Houses Sold per Zipcode



Source: https://www.kaggle.com/harlfoxem/housesalesprediction

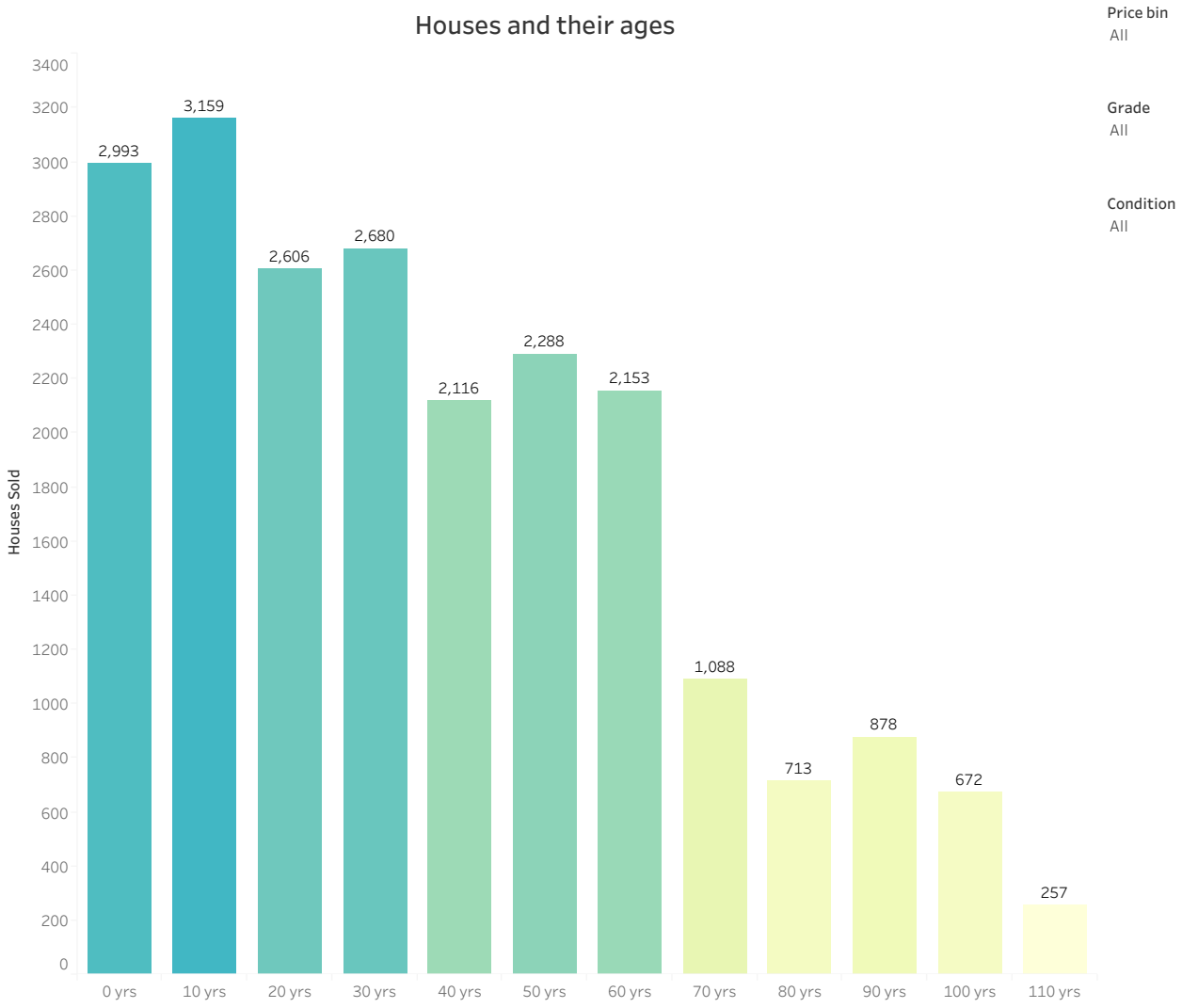
King County House Prices Data Analysis

House prices are more clustered around the $200k - $800k range (82.5% probability)



King County House Prices Data Analysis

Good quality (6 - 13 out of 13 grade levels) **and in good condition** (3 - 5 out of 5 condition levels) houses are expected to be found in houses of between **0 - 70 years old** (**83% probability**)



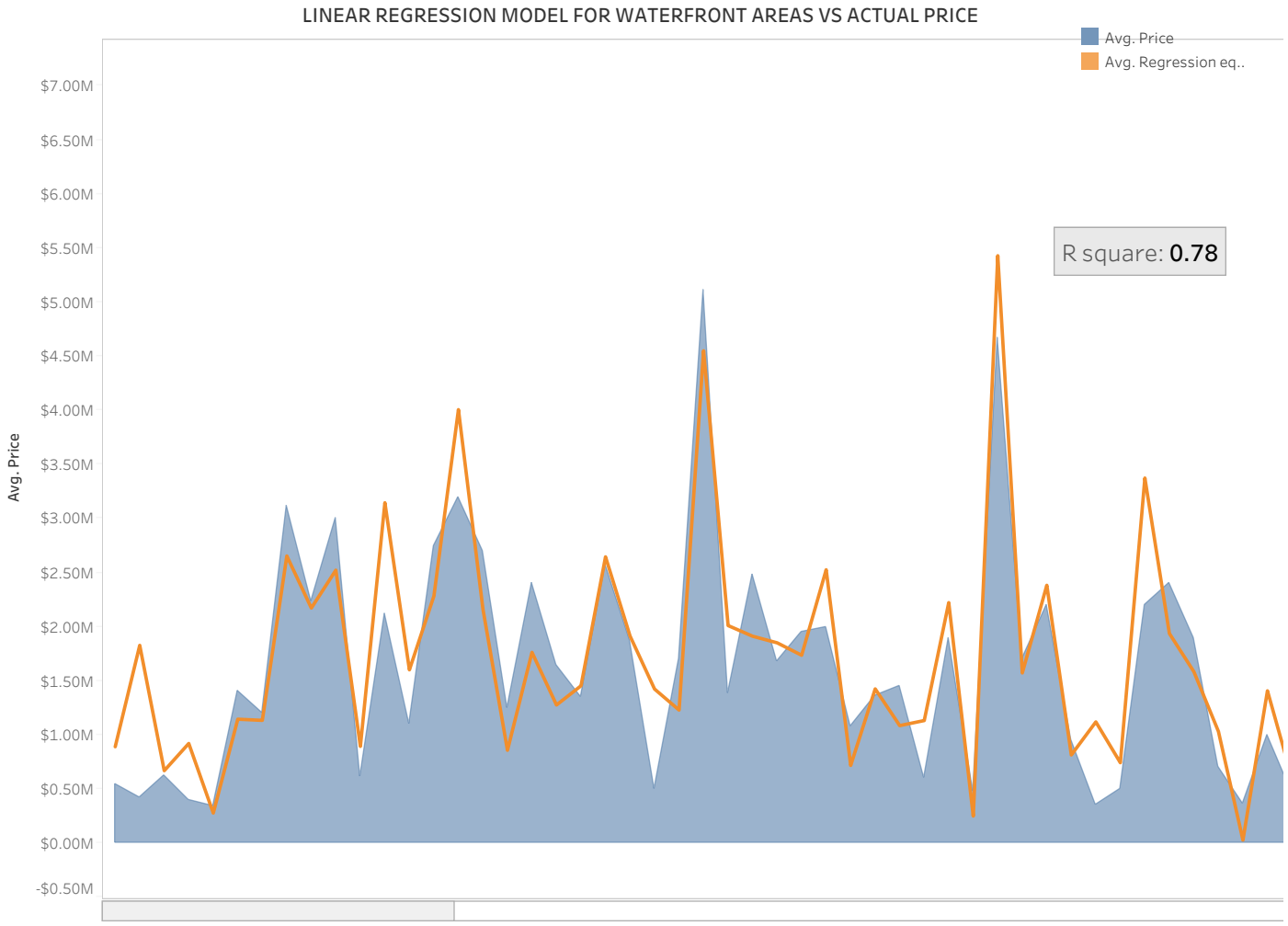
King County House Prices Data Analysis

The variables with a considerable relationship to price include: **sqft_living, bedroom, bathroom, grade, and condition** (view intentionally excluded for its binary nature)

Column1	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_baseme	Age	qft_living1	sqft_lot15
price	1.00														
bedrooms	0.32	1.00													
bathrooms	0.53	0.53	1.00												
sqft_living	0.70	0.59	0.76	1.00											
sqft_lot	0.09	0.03	0.09	0.17	1.00										
floors	0.26	0.18	0.50	0.35	-0.01	1.00									
waterfront	0.27	-0.01	0.06	0.10	0.02	0.02	1.00								
view	0.40	0.08	0.19	0.28	0.07	0.03	0.40	1.00							
condition	0.04	0.02	-0.13	-0.06	-0.01	-0.26	0.02	0.05	1.00						
grade	0.67	0.37	0.67	0.76	0.11	0.46	0.08	0.25	-0.15	1.00					
sqft_above	0.61	0.49	0.69	0.88	0.18	0.52	0.07	0.17	-0.16	0.76	1.00				
sqft_basement	0.32	0.31	0.28	0.44	0.02	-0.25	0.08	0.28	0.17	0.17	-0.05	1.00			
Age	-0.11	-0.17	-0.54	-0.34	-0.05	-0.51	0.00	0.02	0.40	-0.46	-0.44	0.10	1.00		
sqft_living15	0.59	0.40	0.57	0.76	0.14	0.28	0.09	0.28	-0.09	0.71	0.73	0.20	-0.32	1.00	
sqft_lot15	0.08	0.03	0.09	0.18	0.72	-0.01	0.03	0.07	0.00	0.12	0.19	0.02	-0.07	0.18	1

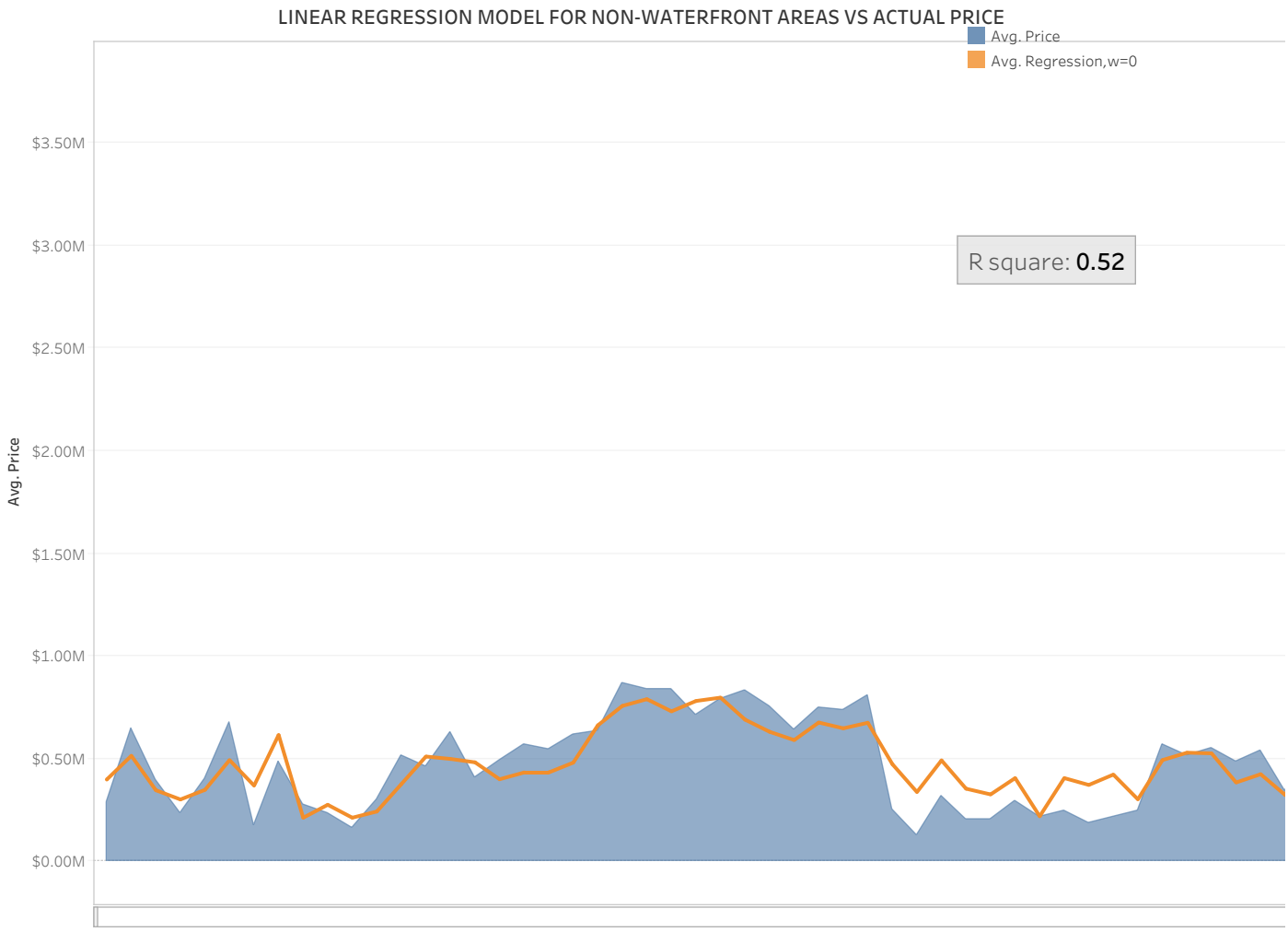
King County House Prices Data Analysis

The linear regression model for prices for **houses in the watershed areas** seems to follow the data but **exhibits large errors** in multiple instances



King County House Prices Data Analysis

The linear regression model for the **prices of houses outside a watershed area** showed **similar behavior to the houses in the watershed areas**, following the data but exhibiting **occasional large errors**.



King County House Prices Data Analysis

1. There is an **~82.5% probability** that the price of a house will be between **$200k - $800k** across the County.

2. Houses between **0-70 years old** (age range of houses between 0 and 115 yrs) are typically of **good quality and design and in good condition**

3. **Linear regression model not a good fit** for current data

2. **Lots of errors** in model estimation

3. Possible causes of poor model fit: outlier data - prices from **$3.4M upwards**

Recommendations:
1. Use a **more suitable type of analysis** to properly estimate average prices

2. **Use a larger dataset**

3. **Get rid of outliers** in data