



# King County House Price Prediction

Candice Wu

DA 485 | Data Analytics Capstone Project

Spring 2023

# Agenda

1. Project Overview
2. Data Used for this Project
3. Data processing
4. Methods
5. Exploration
6. Model Comparison
7. Findings
8. Recommendations
9. Q&A
10. Sources





# Project Overview

## Objective:

As a data scientist at FlyHomes, my role is to delve into the house sales data in the King County area. The task involves building predictive models for sale prices and pinpointing the factors that significantly influence house prices.

## Business Question:

- **Location:** Which areas in King County command the highest average house prices?
- **Inner Factors:** What house features have the most substantial impact on the sale price?
- **External Factors:** How do external elements such as property tax rates, school ratings, and demographic data sway house prices in King County?

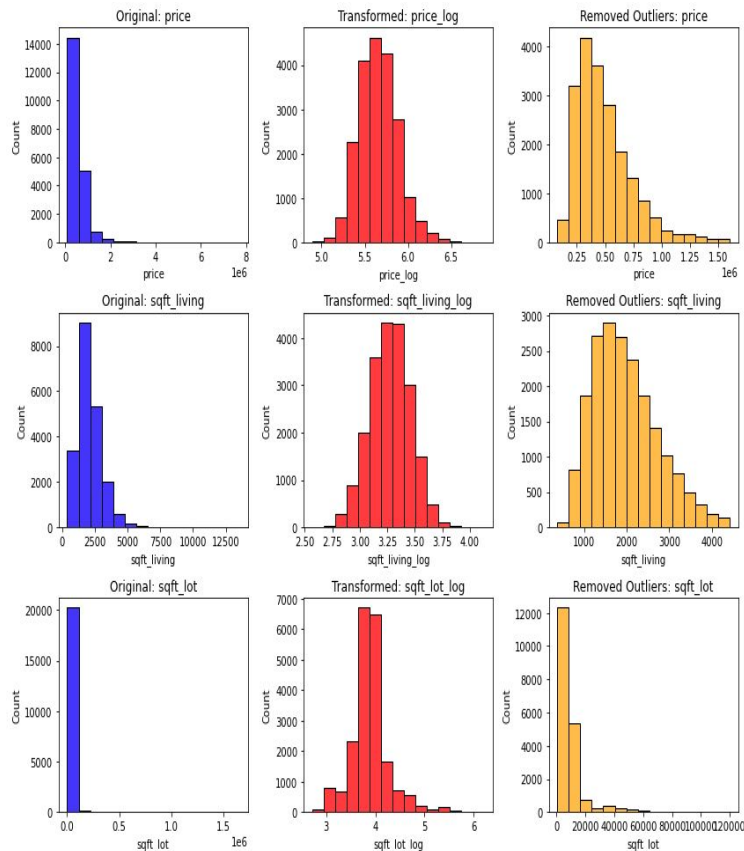
## Goal:

This project aims to guide potential investors and homebuyers in making informed decisions in the King County real estate market.

# Data Used for this Project

	Dataset Name	Data Source Link	Source
1	House Sales in King County	<a href="https://www.kaggle.com/datasets/harlfoxem/housesaleprediction">https://www.kaggle.com/datasets/harlfoxem/housesaleprediction</a>	Kaggle
2	School Dataset	<a href="https://www.schooldigger.com/go/WA/cityrank.aspx?">https://www.schooldigger.com/go/WA/cityrank.aspx?</a>	Schooldigger
3	Crime Dataset	<a href="https://ucr.fbi.gov/crime-in-the-u.s">https://ucr.fbi.gov/crime-in-the-u.s</a>	UCR
4	Property Tax Rate Dataset	<a href="https://kingcounty.gov/en/legacy/depts/finance-business-operations/treasury/property-tax.aspx">https://kingcounty.gov/en/legacy/depts/finance-business-operations/treasury/property-tax.aspx</a>	Kingcounty
5	Demographic Dataset	<a href="https://data.census.gov/">https://data.census.gov/</a> Tables S2301, DP03, S0801, S0101	Census

# Data Processing





**Data Tools:** Python, Excel

**Feature Selection:**

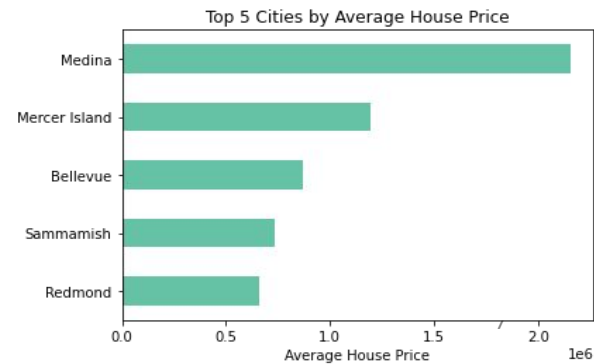
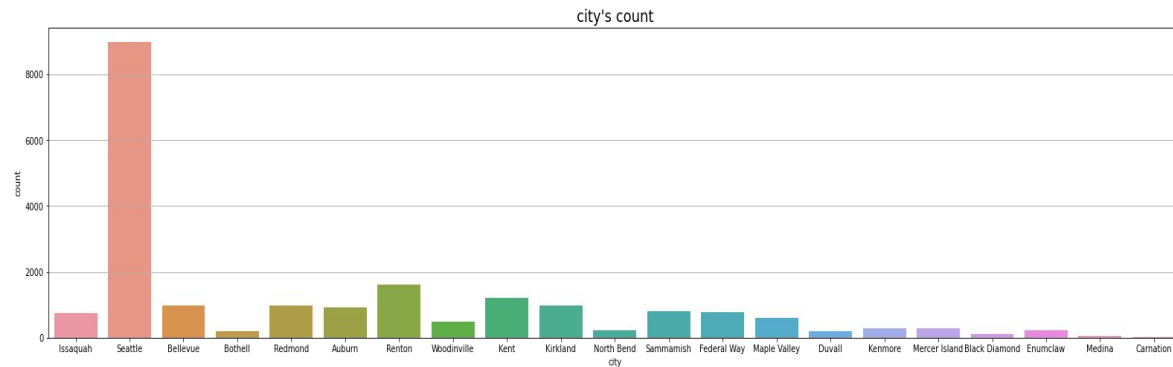
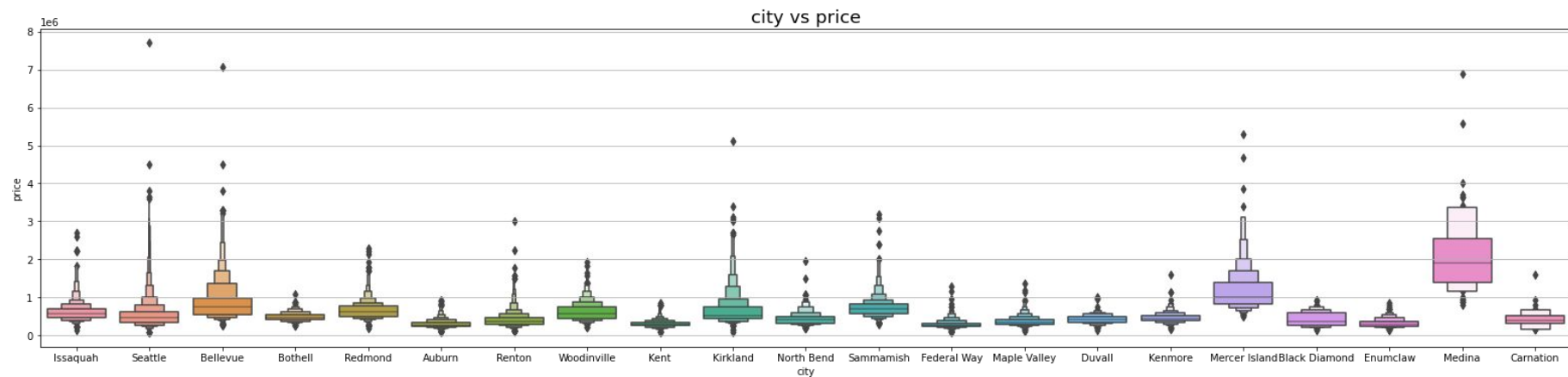
Targeted Variable	price_log
Inner Factors	condition, grade, floors, bedrooms, bathrooms, sqrt_living log, sqrt_lot_log, house_age,
External Factors	school_rate, unemployment_rate, travel_time_to_work, total_population, typical_levy_rate, median_age, median_household_income, area_crime

**Data Transformation:** applied logarithmic transformation to the 'price', 'sqft\_living' and 'sqft\_lot'

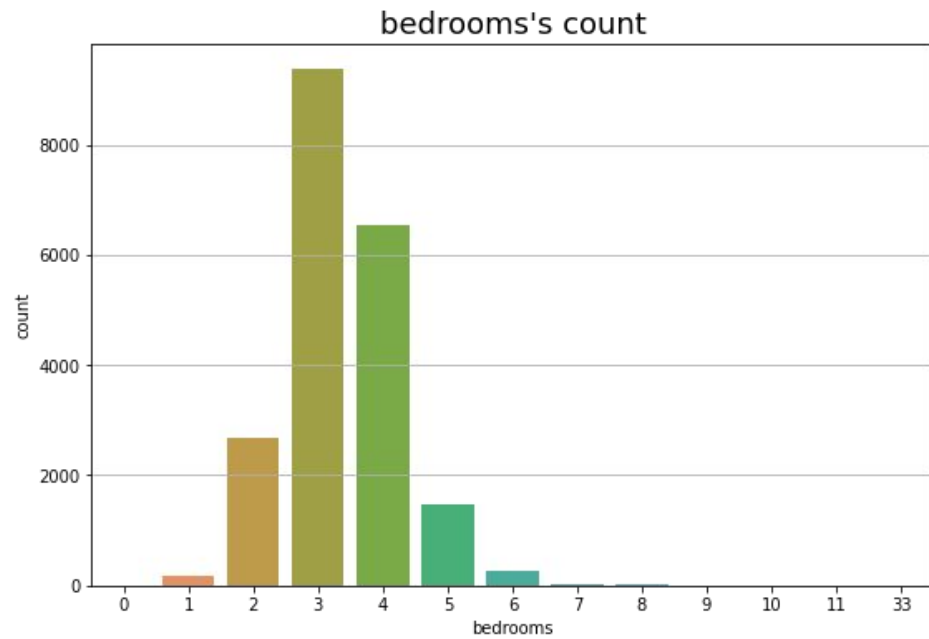
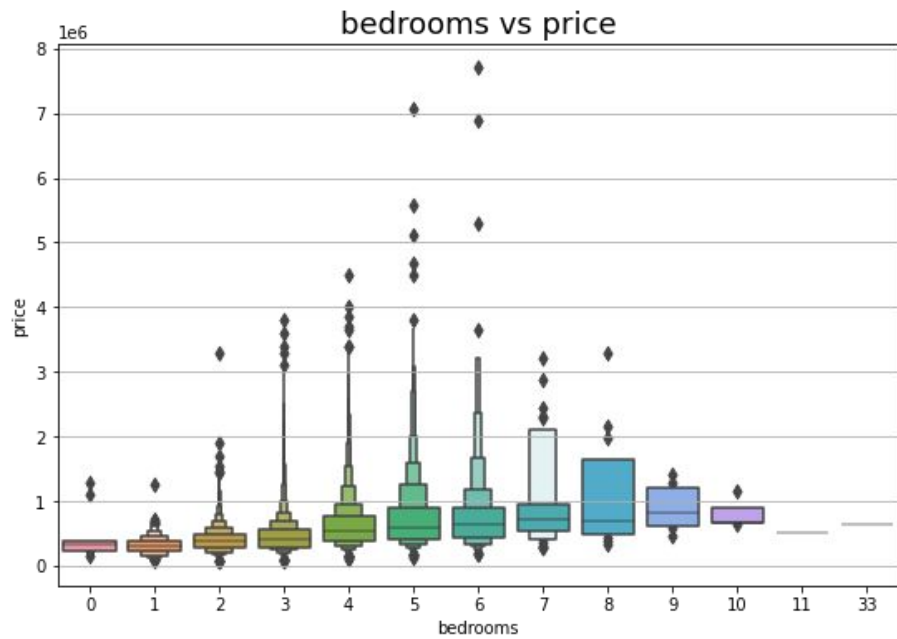
# Methodologies

Method	 Pros	 Cons
EDA	<ul style="list-style-type: none"><li>• Gain initial insights</li><li>• Identify and rectify data errors</li><li>• Select Features</li></ul>	<ul style="list-style-type: none"><li>• Time-consuming with large datasets</li><li>• Subjective insights</li></ul>
MLR	<ul style="list-style-type: none"><li>• Clear and understandable model</li><li>• Enables predictive modeling</li><li>• Identifies impactful variables</li></ul>	<ul style="list-style-type: none"><li>• Relies on strict assumptions</li><li>• Issues with correlated predictors</li><li>• Sensitive to outliers</li></ul>
XGBoost	<ul style="list-style-type: none"><li>• High predictive accuracy</li><li>• Handles missing data automatically</li><li>• Highlights important features</li></ul>	<ul style="list-style-type: none"><li>• Complex and hard to interpret</li><li>• Requires substantial computational resources</li></ul>

# Exploration

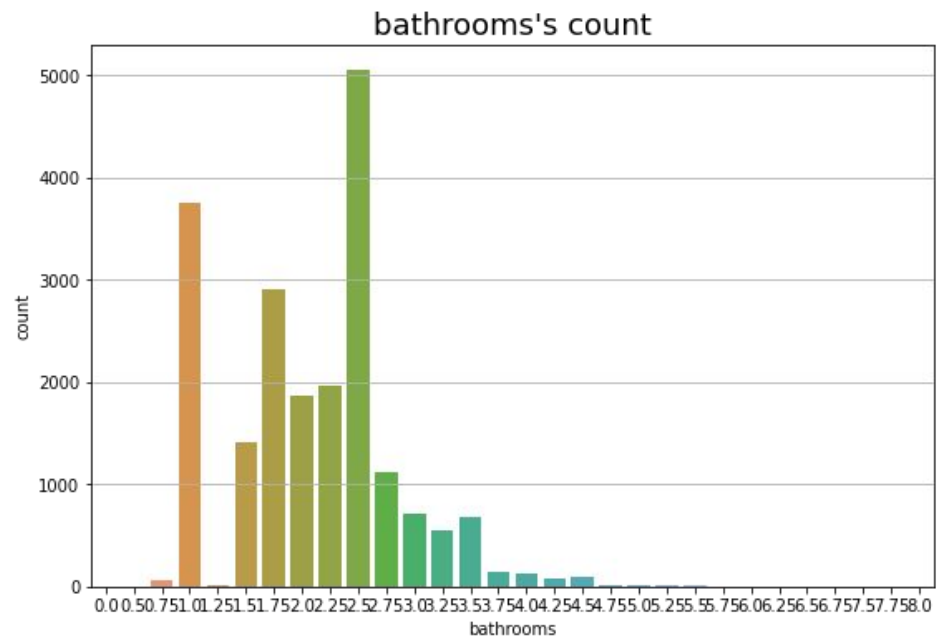
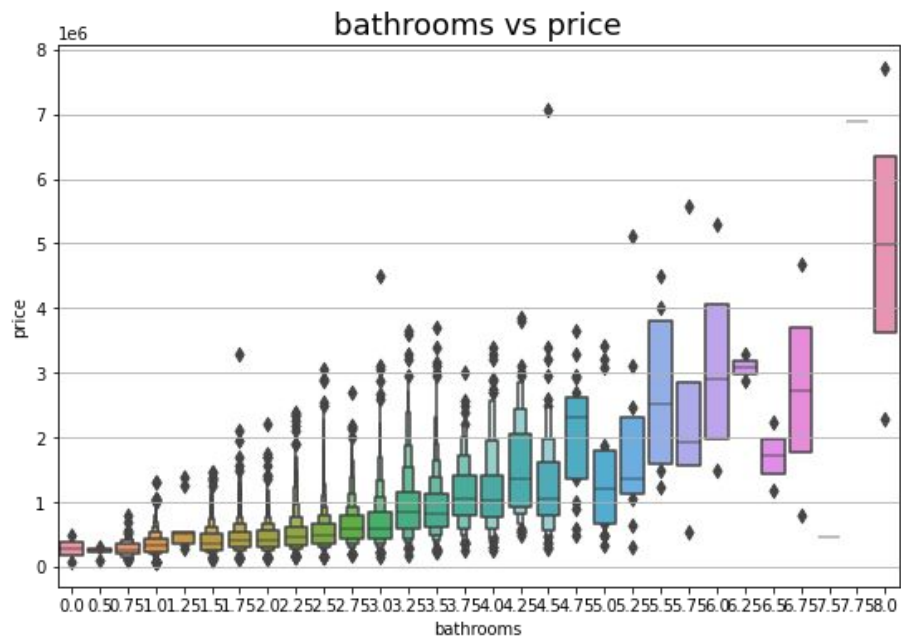


# Exploration

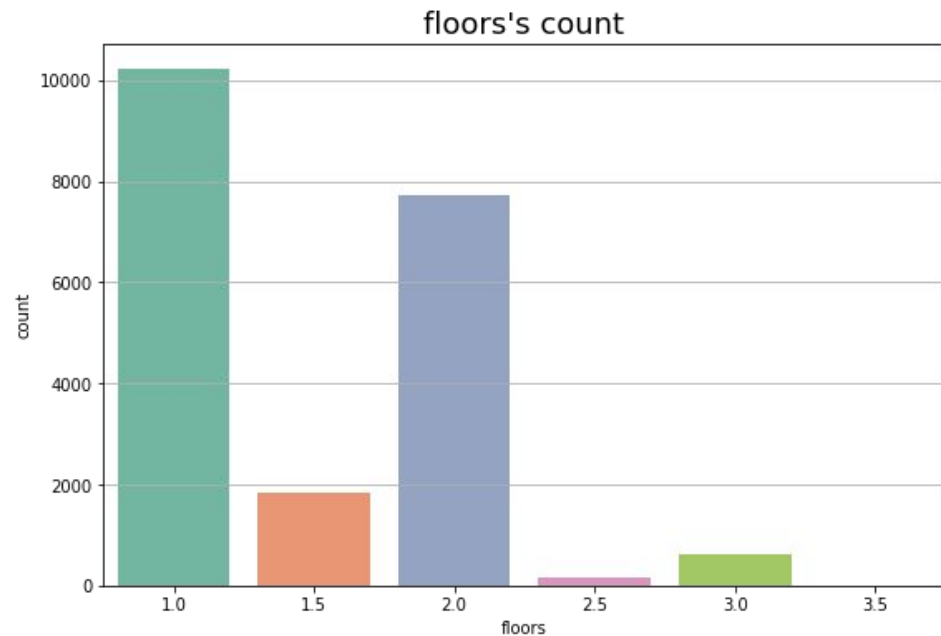
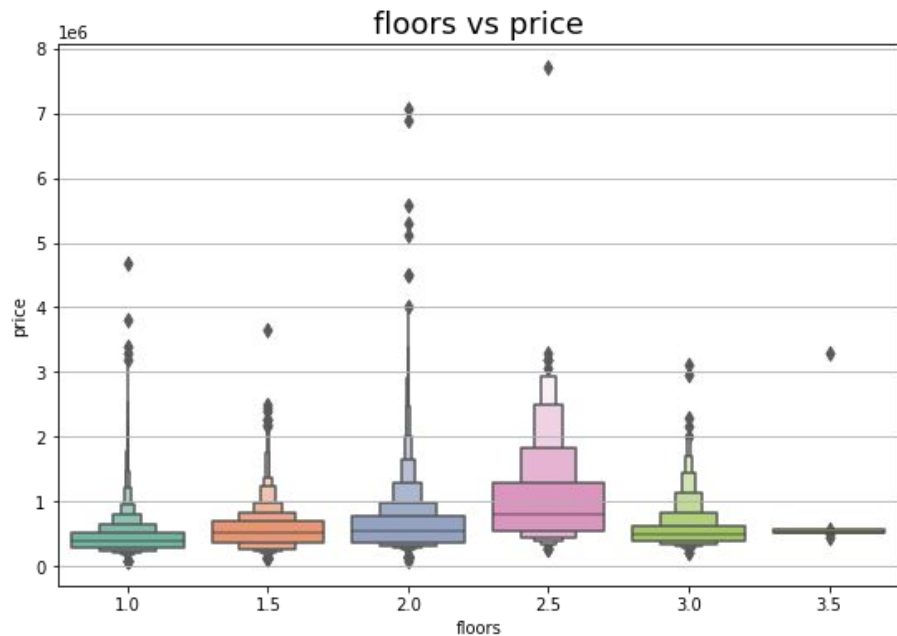




# Exploration

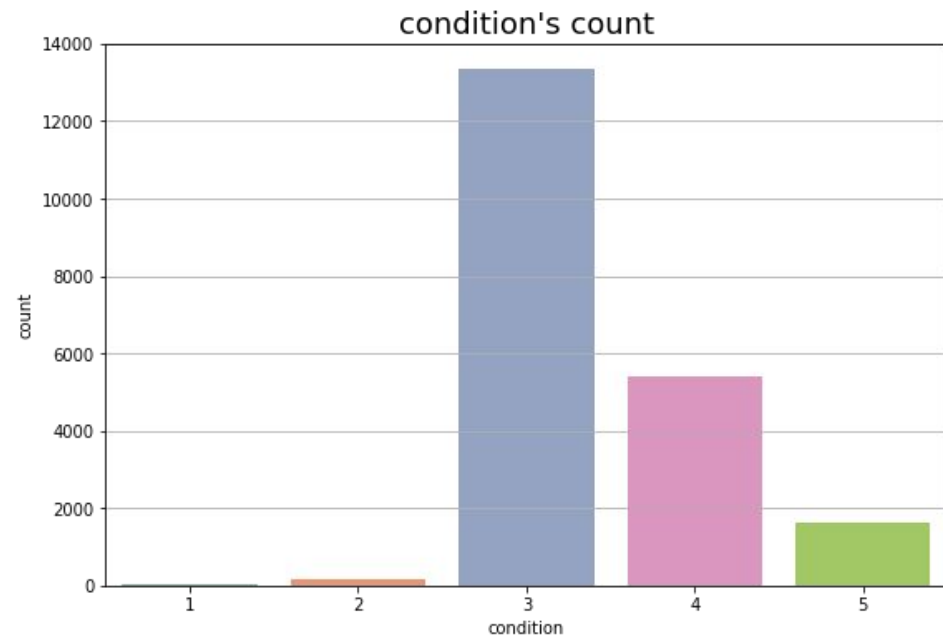
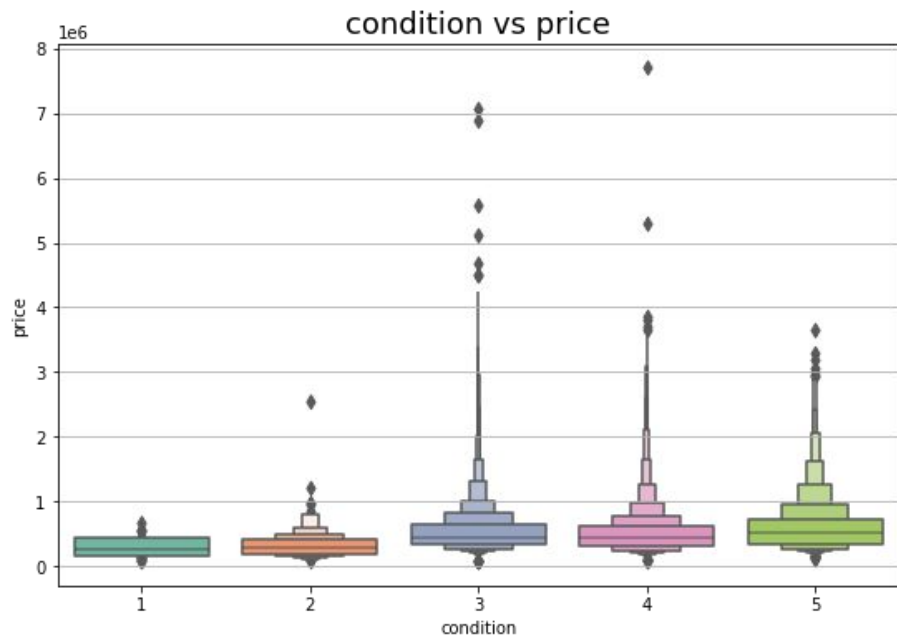


# Exploration

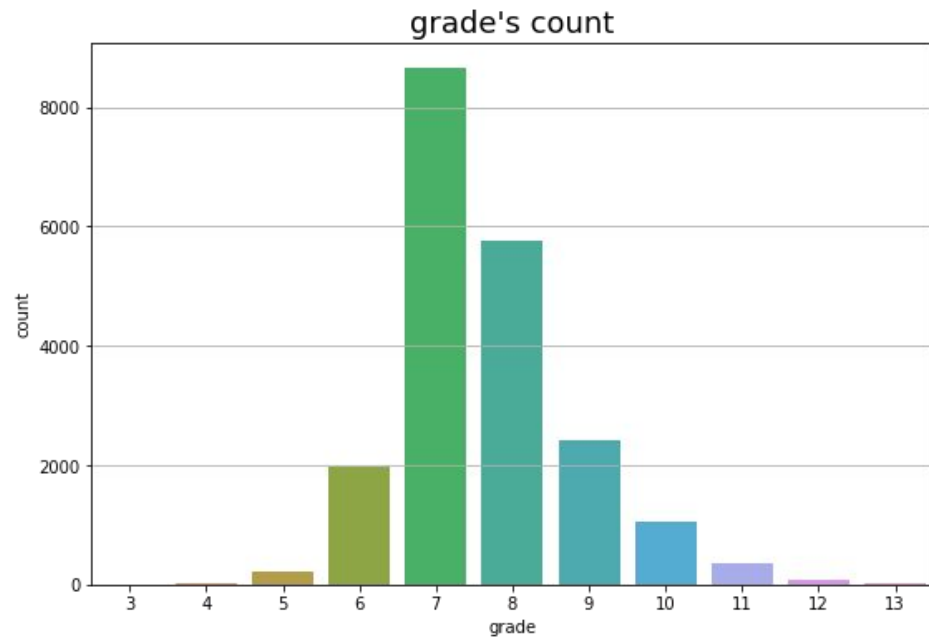
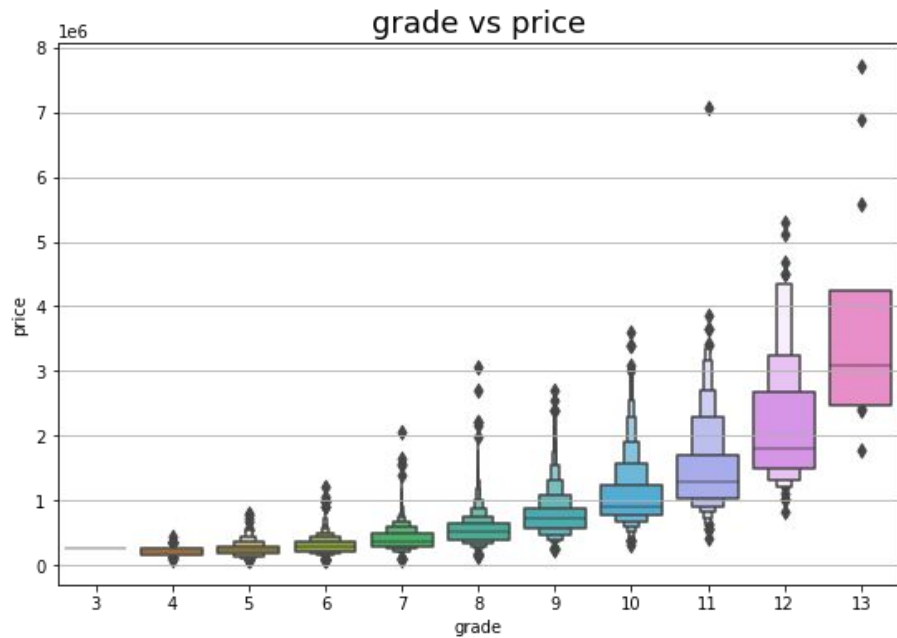


10

# Exploration



# Exploration



12

# Model Comparison

	Inner Factors		External Factors		Overall Factors	
	MRL	XGBoost	MRL	XGBoost	MRL	XGBoost
Top 5 Significant variables	living_log, gade, lot_log, bathrs, bedrs	lot_log, living_log, h_age, bathrs, grade	unemp_rate, levy_rate, Sch_rate, travel_time, med_age	med_age, med_hh_income, tot_pop, travel_time, unemp_rate	tevy_rate, unemp_rate, living_log, schol_rate, grade,	lot_log, living_log, h_age, bathrs, grade
MAE	0.097	0.100	0.110	0.115	0.063	0.056
MSE	0.014	0.016	0.019	0.023	0.006	0.006
RMSE	0.312	0.128	0.332	0.153	0.250	0.079
R-Squared	0.688	0.678	0.528	0.541	0.865	0.877



# Findings

## Location Insights

- **Top Cities for Investment:** Medina, Mercer Island, Bellevue.
- Seattle: Major hub with dense housing landscape.

## House Features

- **Optimal Features:** Homes with 3 bedrooms, 2.5 bathrooms, and 2.5 floors tend to fetch the highest prices.
- **Grade:** A higher grade generally correlates with a higher price; the condition score is less influential.

## Model

- **Inner Factors:** The `sqft_living_log`, `sqft_lot_log`, `grade`, and the number of bathrooms significantly impact house prices.
- **External Factors:** The unemployment rate, travel time to work, and median age are significant factors.
- **Overall:** The `sqft_living_log` and `grade` being pivotal factors.

## Challenges

**Multicollinearity:** Present in both inner and external factors, yet better managed by the XGBoost model compared to MLR

# Recommendation



Utilize the insights derived from the analysis to create data-driven marketing strategies. For example, highlighting the optimal house features (like the number of floors and grade score) in marketing.



Offer advisory services to clients, helping them make informed decisions based on the significant factors influencing house prices in King County.



Collaborate with local authorities to gather more comprehensive data on external factors such as crime rates and school ratings to further enhance the predictive models.



Recommend leveraging the XGBoost model for predictive analyses given its slightly superior performance compared to the MLR model, especially in handling multicollinearity issues effectively.

# Further Work

---

## More sales data:

The dataset covers the period from May 2014 to May 2015. To reach a more accurate conclusion, it is essential to obtain data that is both more recent and spans a longer period

## Temporal Analysis:

Conduct a temporal analysis to understand how house prices have evolved over time and identify any seasonal trends or patterns.

## Interactive Dashboard:

Develop an interactive dashboard that allows users to explore the data and insights visually and to generate custom reports based on their preferences.







# Audience Q&A

---

# Thank you

---

# Sources

1. "2015 Annual Reports." King County, 2015, [kingcounty.gov/depts/assessor/Reports/annual-reports/2015.aspx](https://kingcounty.gov/depts/assessor/Reports/annual-reports/2015.aspx). Accessed 1 May 2023.
2. "Crime in the U.S. 2015." FBI, 3 May 2016, [ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015](https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015). Accessed 1 May 2023.
3. Gupta, Vishesh. "Correlation in XGBoost." Medium, [medium.com, https://vishesh-gupta.medium.com/correlation-in-xgboost-8afa649bd066#:~:text=Conclusion,in%20the%20presence%20of%20multicollinearity](https://vishesh-gupta.medium.com/correlation-in-xgboost-8afa649bd066#:~:text=Conclusion,in%20the%20presence%20of%20multicollinearity). Accessed 1 May 2023.
4. "Property Tax." King County, King County Government, 2023, <https://kingcounty.gov/en/legacy/depts/finance-business-operations/treasury/property-tax.aspx>. Accessed 1 May 2023.
5. United States Census Bureau. "Explore Census Data." [data.census.gov, 2015.data.census.gov/table?t=Income%2B%28Households%2C%2BFamilies%2C%2BIndividuals%29&g=050XX00US53033%248600000&y=2015](https://data.census.gov/2015.data.census.gov/table?t=Income%2B%28Households%2C%2BFamilies%2C%2BIndividuals%29&g=050XX00US53033%248600000&y=2015). Accessed 1 May 2023.
6. "Washington City Rankings." SchoolDigger, 2015, [www.schooldigger.com/go/WA/cityrank.aspx?y=2015](https://www.schooldigger.com/go/WA/cityrank.aspx?y=2015). Accessed 1 May 2023.