# Phase 2 Project

Flatiron Online Data Science Course

# Summary

This is the second end-of-module project in the Flatiron Online Data Science Course

The project aims at testing and reinforcing the **scientific computing and quantitative methods** taught in phase 2 of the Flatiron School curriculum

Applying the tools of **linear regression enables data driven recommendations and predictions** for a multitude of different scenarios.

# Outline

- Business Problem
- Data & Methodology
- Results
- Conclusions & Actionable Insights
- Further studies

# Business Problem

For its next marketing campaign, a real estate agency is requiring investigation into the house sales of the King County area (a northwestern county of the US).

**Preferences of potential house buyers** and support with **reliable price predictions** are the main drivers for the campaign’s success.

# Data & Methodology

- The King County housing data consists of **more than 21,000 prices and features of house sales** that occurred between **May 2014 and 2015**

- After careful exploration and cleaning of the data, an in-depth analysis was performed to reveal and select those **variables** that may provide the most **useful and interesting insights** for the marketing campaign

- Ultimately, the data was used to build models for predicting housing prices

- These models‘ qualities were compared and the **predictive abilities** of the best performing model were tested

# Results

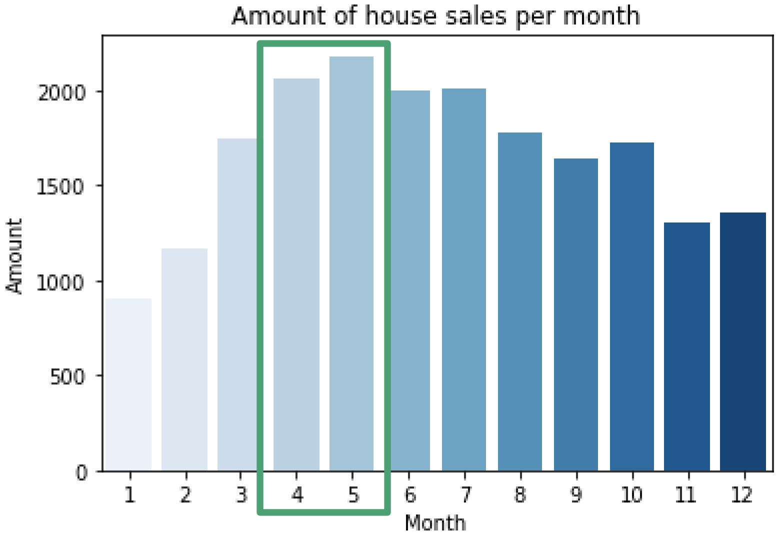
The following three key features were explored:

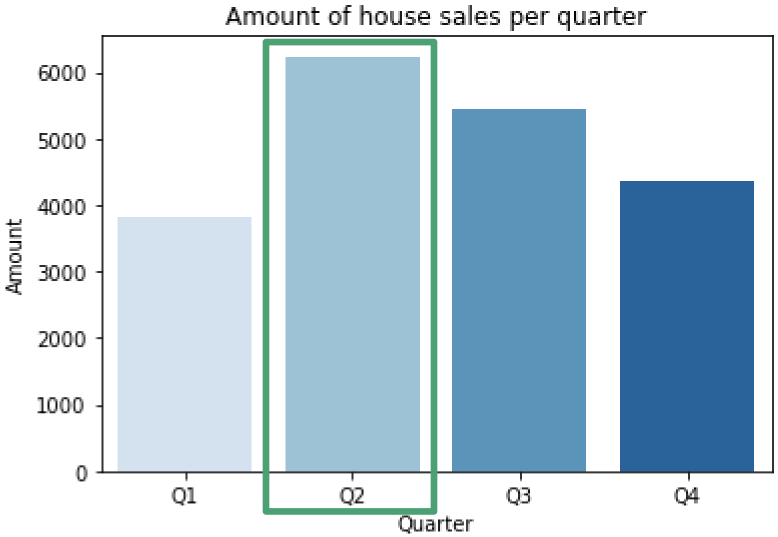
1. Time
2. Location
3. Attributes
   - Bedroom, bathroom, floor count
   - Basement and renovations
   - Grade and condition

The best performing regression model has been identified for price predictions

# 1. Time

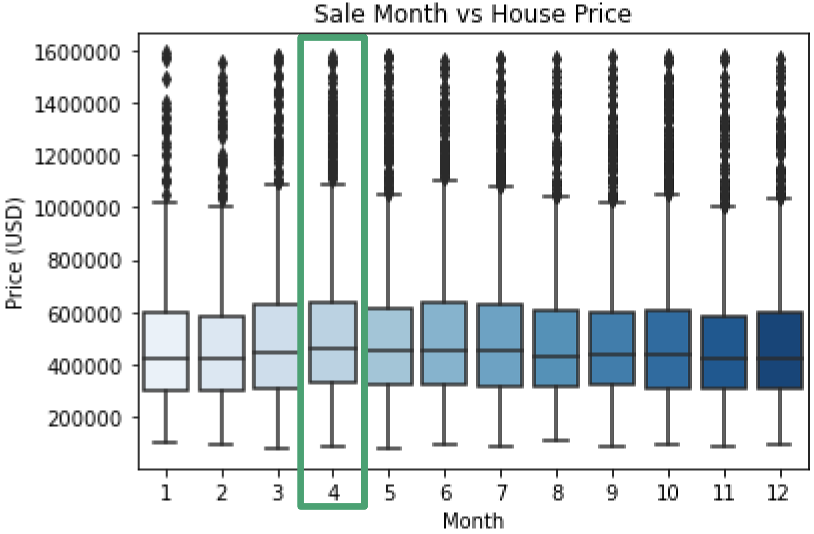
- Does the timing of a house sale affect sales price?

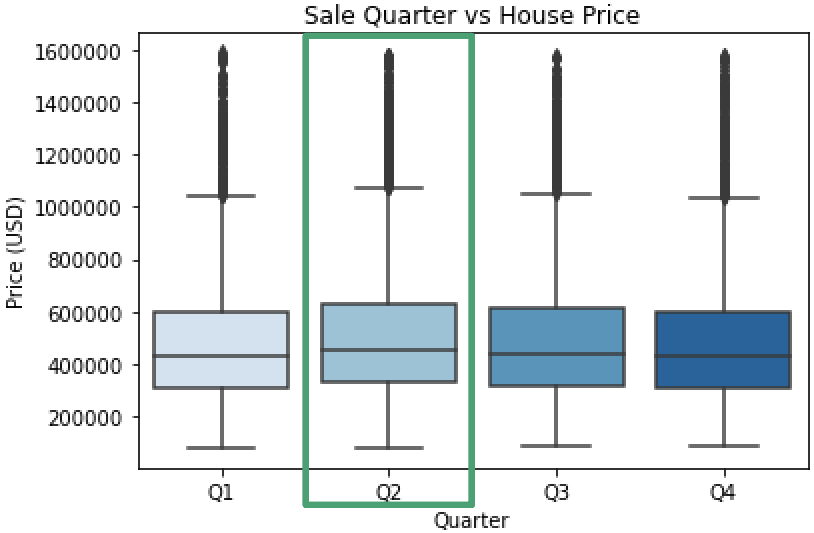




# 1. Time

- Does the timing of a house sale affect sales price?





- April: 461,000 USD median sales price

- Q2: 453,500 USD median sales price

# 2. Location

- Which areas to focus on for the highest sales prices?

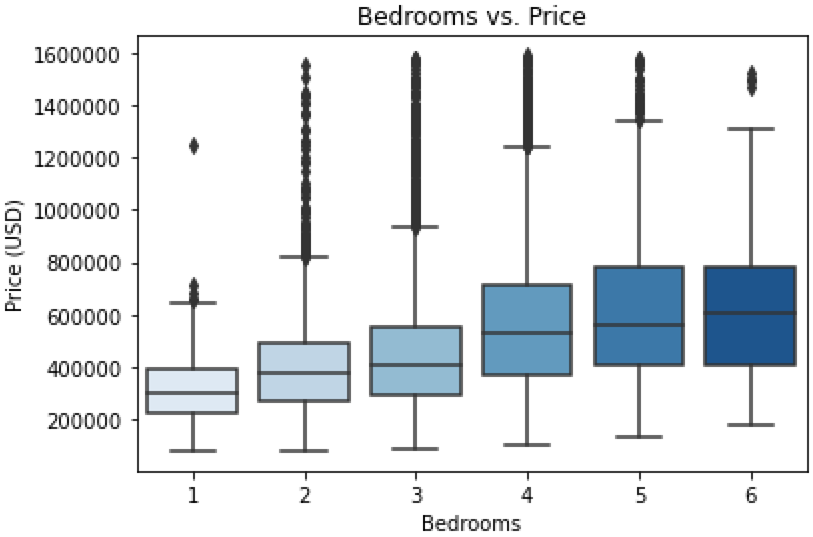
| Top 5 Areas by Sales Price | |
|---|---|
| **Zipcode** | **Top Sales Prices** |
| 98004 | 1,590,000 USD |
| 98040 | 1,580,000 USD |
| 98033 | 1,580,000 USD |
| 98105 | 1,580,000 USD |
| 98166 | 1,580,000 USD |

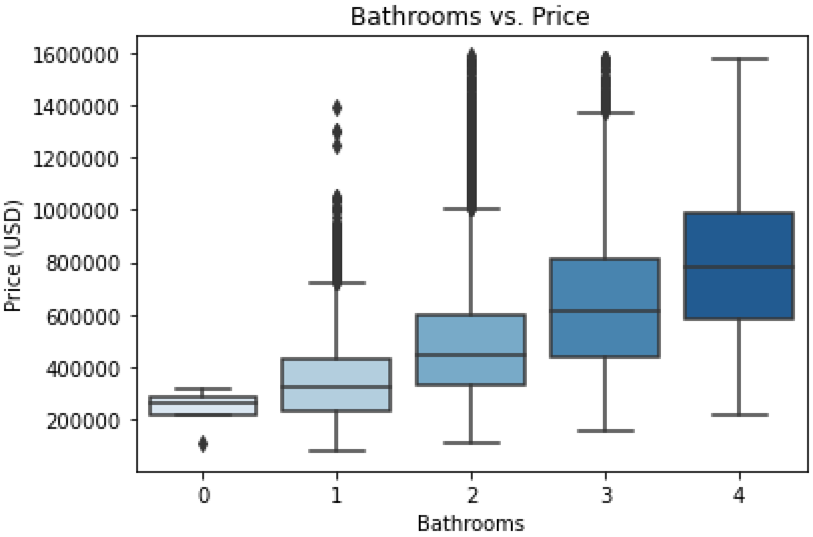
| Low 5 Areas by Sales Price | |
|---|---|
| **Zipcode** | **Top Sales Prices** |
| 98168 | 78,000 USD |
| 98014 | 80,000 USD |
| 98168 | 81,000 USD |
| 98146 | 82,000 USD |
| 98032 | 85,000 USD |

# 3. Housing Attributes

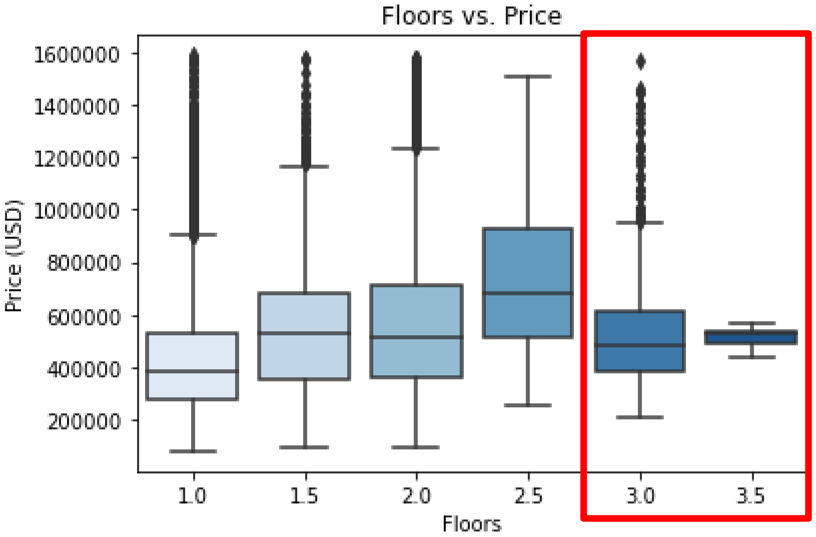
- How do the following attributes affect sales prices?
  - Bedroom, bathroom, floor count
  - Basement and renovations
  - Grade and condition

# 3. Housing Attributes – Bedroom & Bathroom count

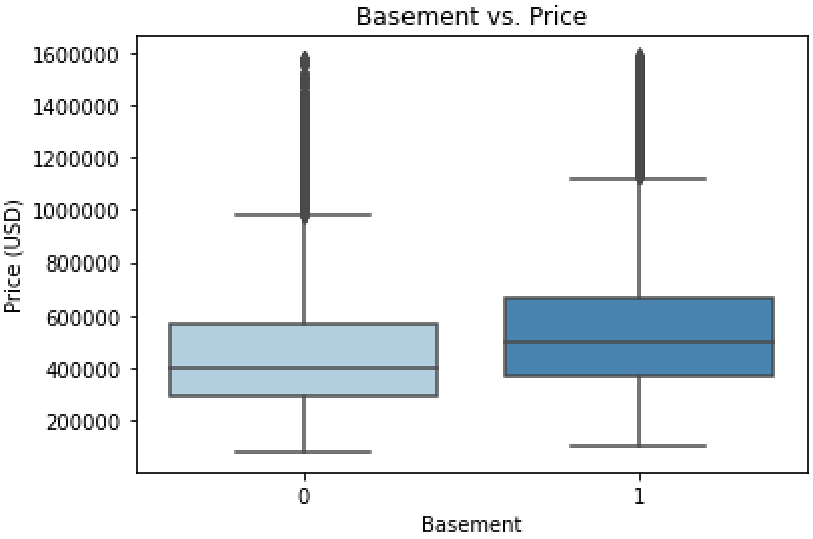


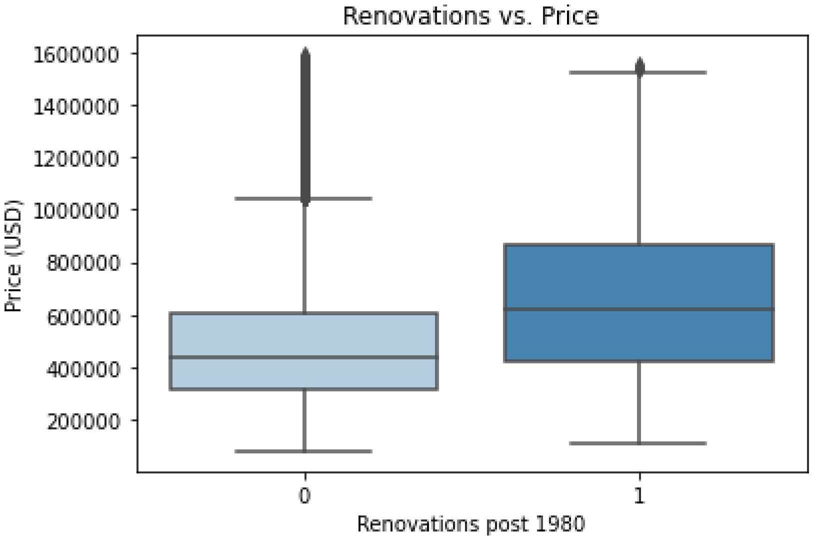


# 3. Housing Attributes – Floor count

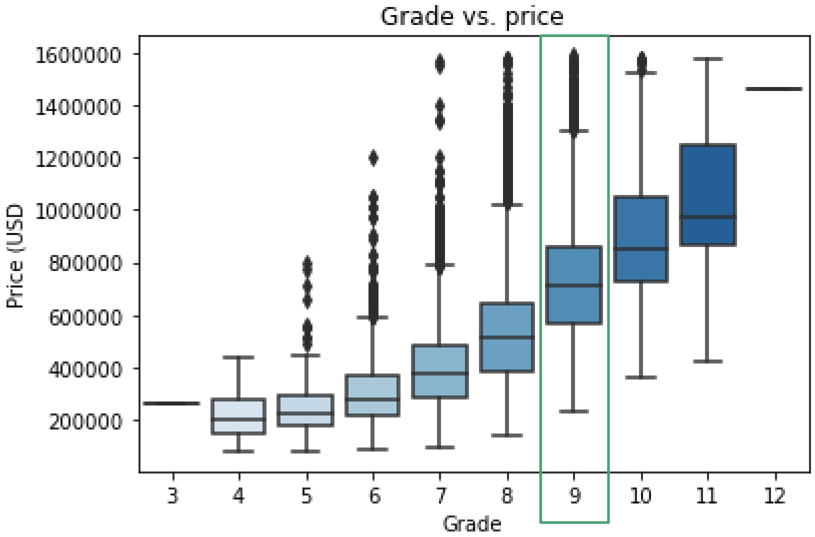


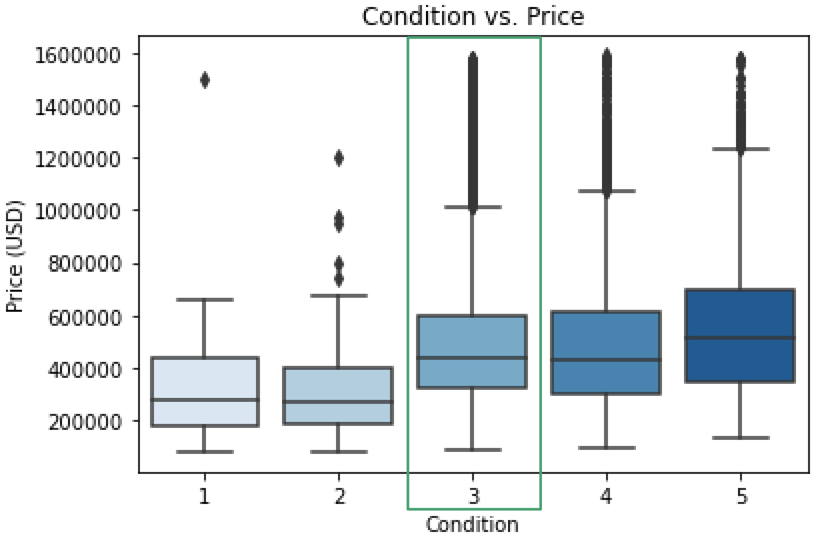
# 3. Housing Attributes – Basement & Renovations





# 3. Housing Attributes – Grade & Condition



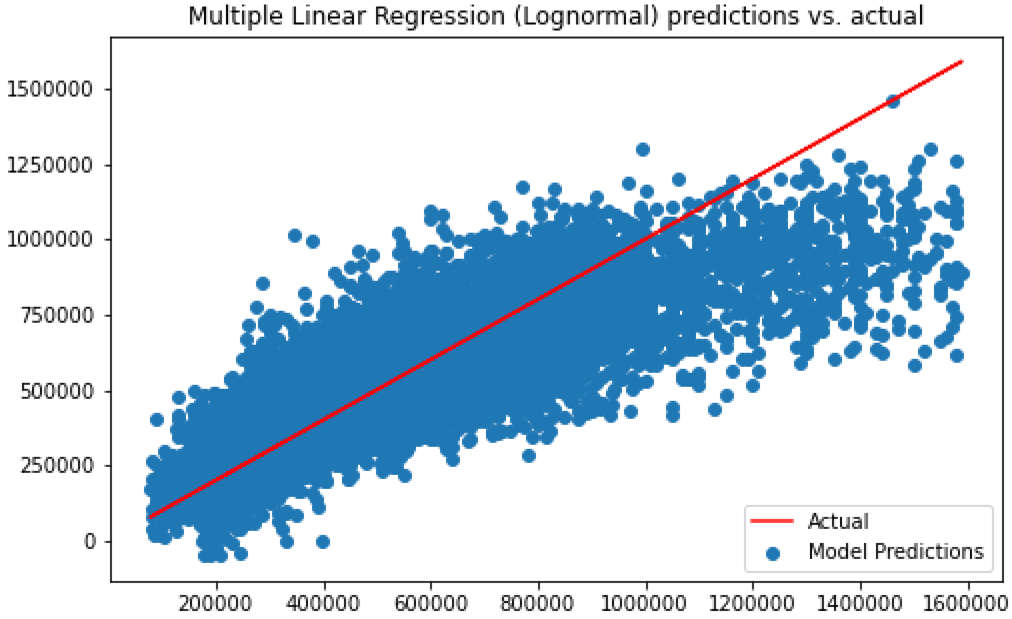


# Regression models

| Model type | Description | Number of Features | Variance in price explained ($R^2$ Adjusted) | On average deviation of predicted from actual price (RMSE) |
|---|---|---|---|---|
| Multiple Linear Regression | Lognormal transformation | 46 | 69,1% | 134,904 USD |
| Multiple Linear Regression | One-hot encoding | 43 | 68,2% | 136,640 USD |
| Multiple Linear Regression | Scaled Features | 43 | 68,2% | 136,640 USD |
| Multiple Linear Regression | Lognormal transformation (RFE) | 28 | 68,2% | 136,820 USD |
| Multiple Linear Regression | RFE | 41 | 67,7% | 137,769 USD |
| Multiple Linear Regression | Without yr_built & quarter features | 16 | 65,9% | 141,357 USD |
| Multiple Linear Regression | one-hot encoding &<br>p-value elimination | 24 | 61,8% | 149,690 USD |
| Simple Linear Regression | Sqft Living | 1 | 38,1% | 190,204 USD |

# Regression models – Best performing model

- ca. 69% of variation in price can be explained
- Average deviation between predicted and actual sales price at ca. 135,000 USD
- Tendency to underestimate higher sales prices



# Conclusions & Actionable Insights

1. **Time**
   - Launch advertising campaign in **April & June**
   - Avoid winter months and Q1 in general (esp. January)

2. **Location**
   - Focus on areas with highest sales prices in the past

3. **Housing Attributes**
   - **More bedrooms, bathrooms and floors** are favorable on prices (<u>floors only until 2.5</u>)
   - **Performed renovations and availability of basements** positively affect prices
   - **Grades from 9 upwards and conditions from 3** upwards should be focused on to achieve highest median sales prices

# Further Studies

1. **Polynomials/Advanced regression algorithms**
   - Use higher order functions and advanced algorithms for potentially better predictive power

2. **Zipcode analysis**
   - Employ a more in-depth look into location feature

3. **Model refinements**
   - Use more combinations of different analysis techniques
   - Investigate interactions between features that might distort results

# Thank You!

**Email:** kontakt@oliverzimmer.eu
**GitHub:** @senseize
**LinkedIn:** linkedin.com/in/username/oliver-zimmer-cfa-8824881ab/