Project 4 PRESENTATION

Group 2

Greg Thomas

Hilari Waters

Jon Haas

Housing Prices Machine Learning

University of Kansas – Data Analytics Boot Camp

Project Description & Motivation:

Analysis of USA Real Estate Dataset (2.2M+ listings from Kaggle.com)

- **For businesses, employees and people looking to move their permanent residence location, it is important to know and understand the cost of purchasing a home in the new location.**
    - **Often times, wages/salaries may be similar, between locations.**
        - **However the cost of housing maybe considerably different.**

Specific Needs and Why?

- HME, Inc.’s headquarters are located in Topeka, Kansas where all major operations and fabrications are completed.
  - However, HME has also established satellite offices in larger cities which include Kansas City, Denver, and Dallas.
  - The satellite offices include key positions for Project Management, Estimating, Sales, Accounting, Steel Detailing (drafting), and Engineering.
  - HME also recently added Data Analytics to list!

Specific Needs and Why? (continued)

- As HME has grown to over 500 employees, the management team has realized there is a shortage of talented, experienced, and qualified individuals in the Topeka market.
  - HME must recruit outside of Topeka to fullfill key positions.
    - College Grads want to live in big cities.
    - If HME is going to be competitive in the job marketplace at major universities,
      - HME must have positions available in attractive areas of large cities.

Specific Needs and Why? (continued)

- In order to bring in young engineers, accountants, construction managers, etc.
  - HME must also relocate experienced managers capable of training and managing the college grads.
    - Thus, existing Topeka managers moving to other larger cities, may need compensated for cost of living adjustments.
    - Housing can be a major portion of adjustment.
    - For the convenience and attractiveness of it's satellite offices, HME needs to consider attractive locations close to the housing needs of it's current and future employees.

The Dataset Being Considered:

We choose to Kaggle.com's "USA Real Estate Dataset.

- **It is a large dataset with over 2.2 M home listings in the U.S.**

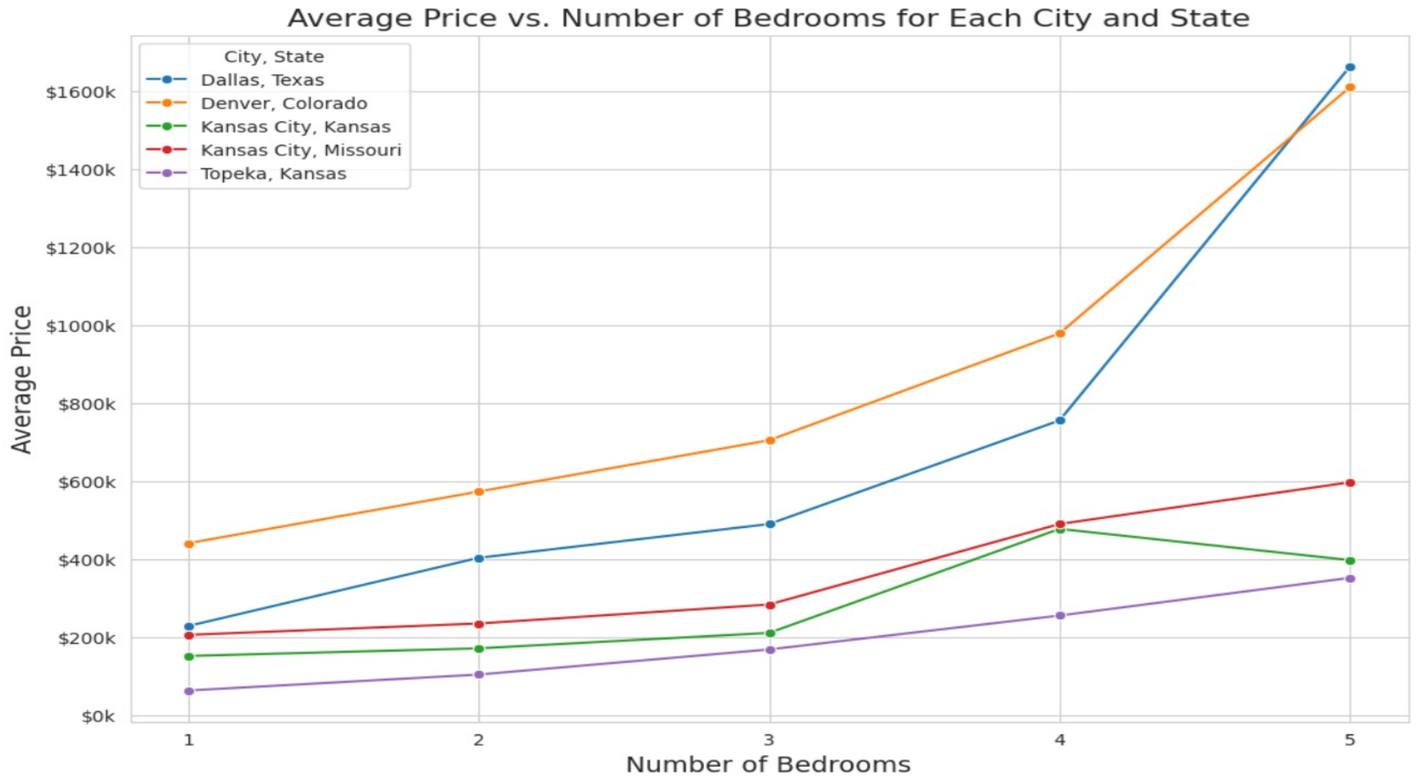
- **The dataset included 10 columns of information:**
  - **Price – Price/sqft, and Price for number of bedrooms were considered**
  - **Bed – We felt very important, but limited our analysis to 2, 3, and 4 bedrooms**
  - **City – We looked at HME's Satellite office cities (KC, Denver, Dallas, and Topeka)**
  - **State – We found important. 3 of the cities in analysis were in multiple states**
  - **zip_code – We used to identify key areas of cities**
  - **House-size – We used house size as square feet to calculate $/sqft**

The Resources:

- Kaggle.com’s “USA Real Estate Dataset.
- Google Colab “Notebook”
- Google Docs
- Matplotlib
- Numpy
- Pandas
- Seaborn
- Sklearn
- Tableau
- Github
- One Hot Encoding
- Random Forest
- Gradient Boosting Regression

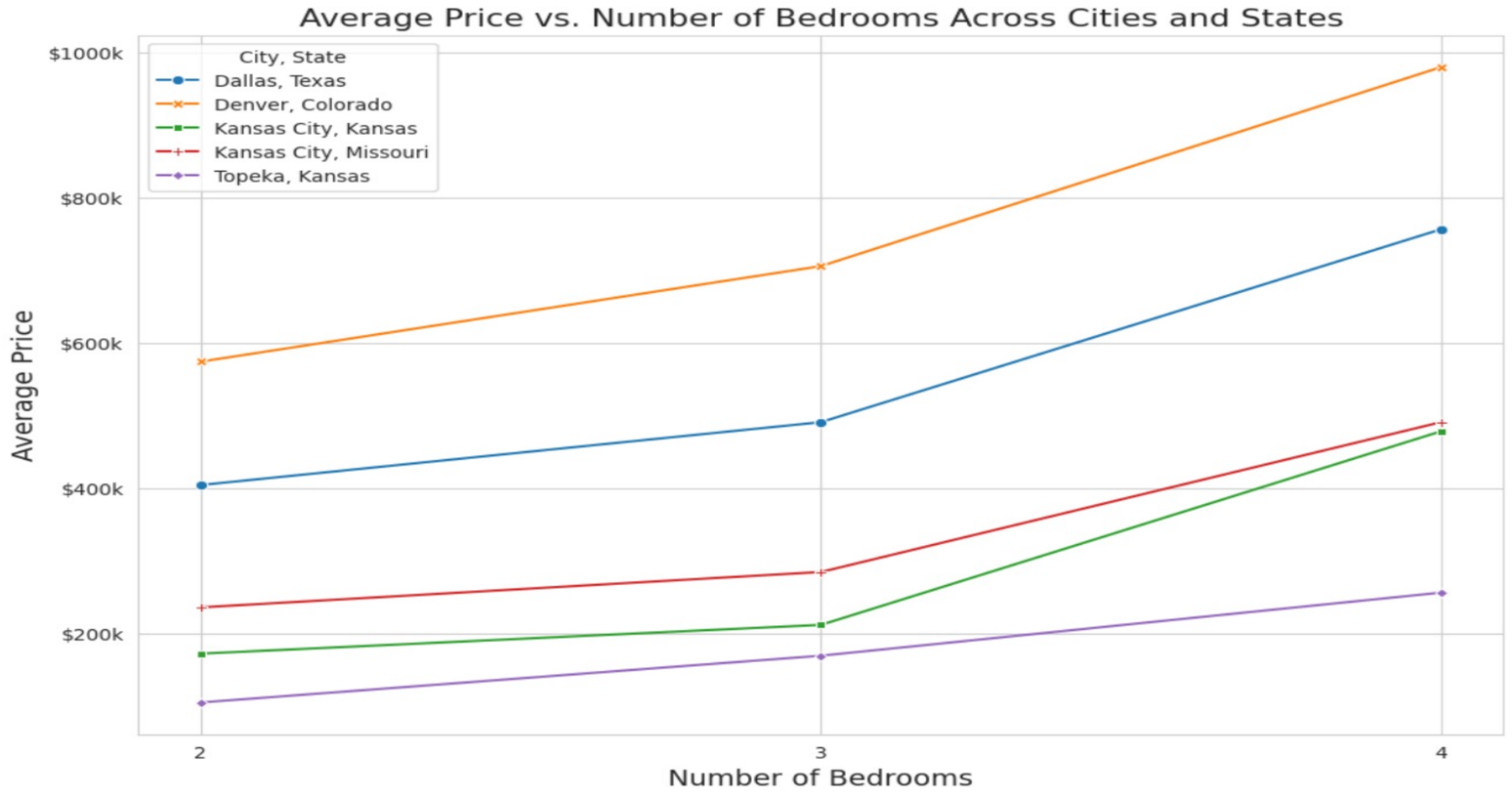
The Data:

- **Data included from 1 to 22 bedrooms. First we choose to get a feel for the data for the for each city, but limited from 1 to 5 bedrooms.**



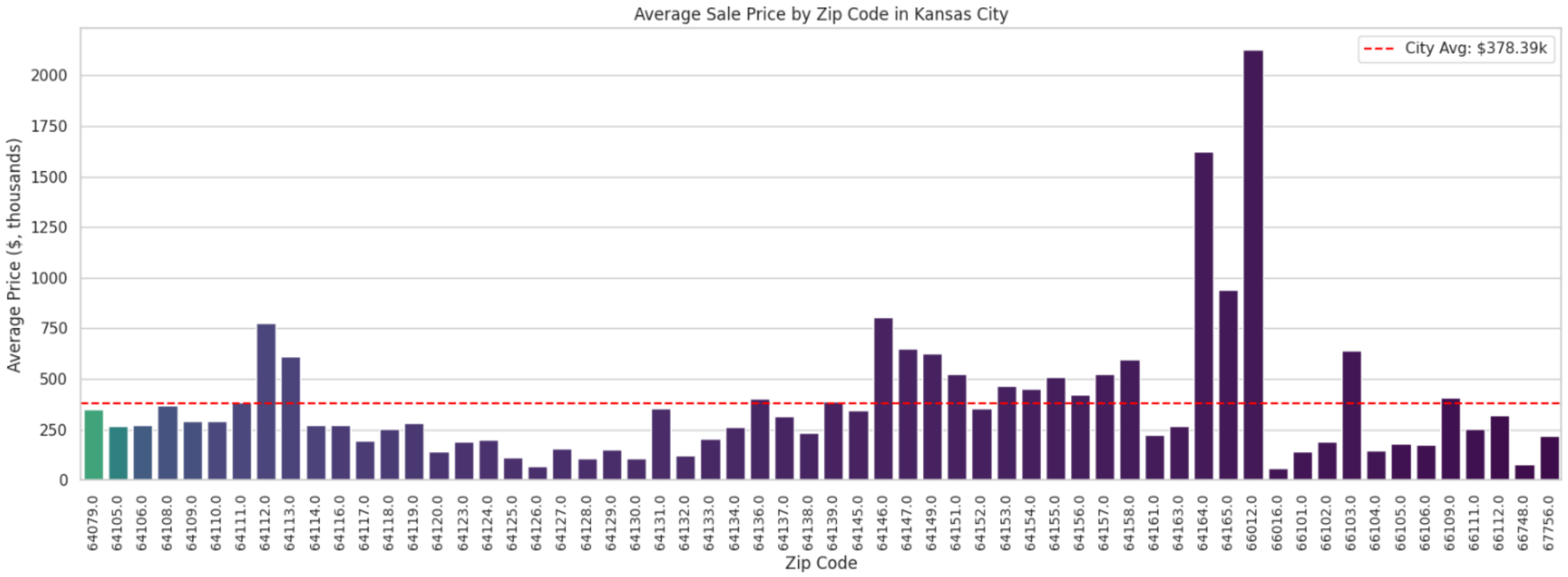
The Data:

- **We then limited to min. of 2 bedrooms and max. of 4 bedrooms.**



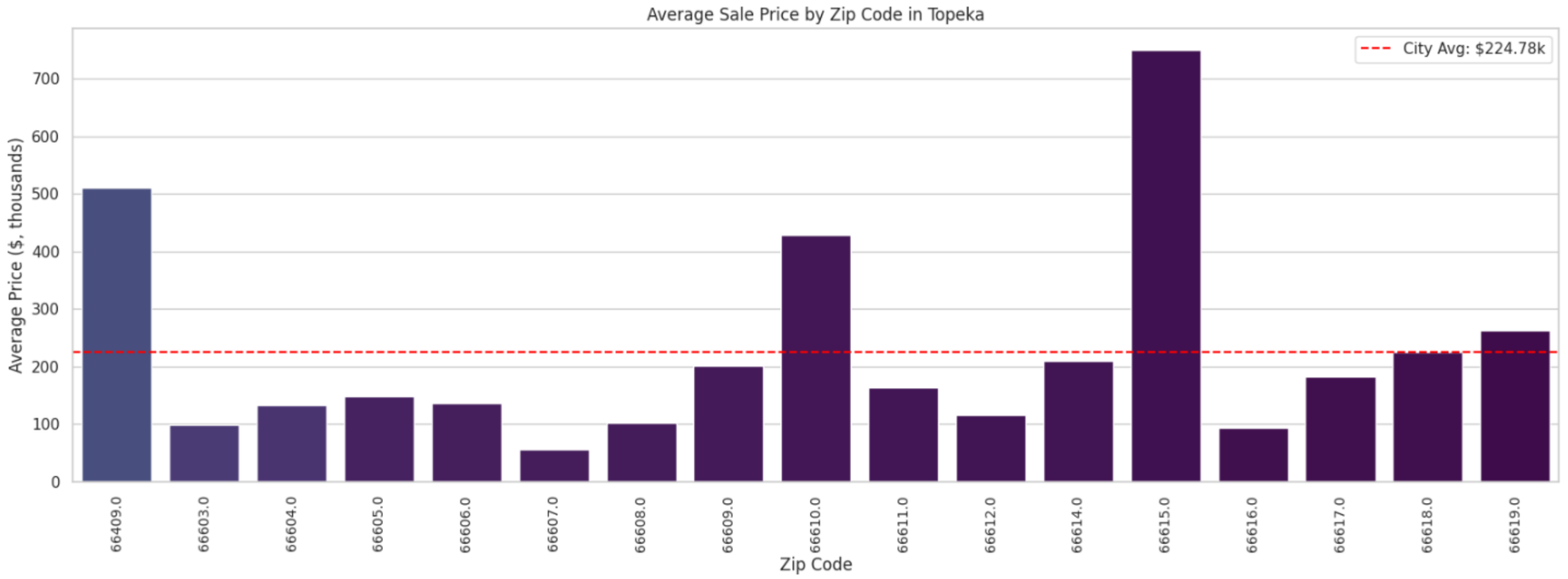
The Data:

- **We then wanted to get a feel for the data by City and prices by zip code.**



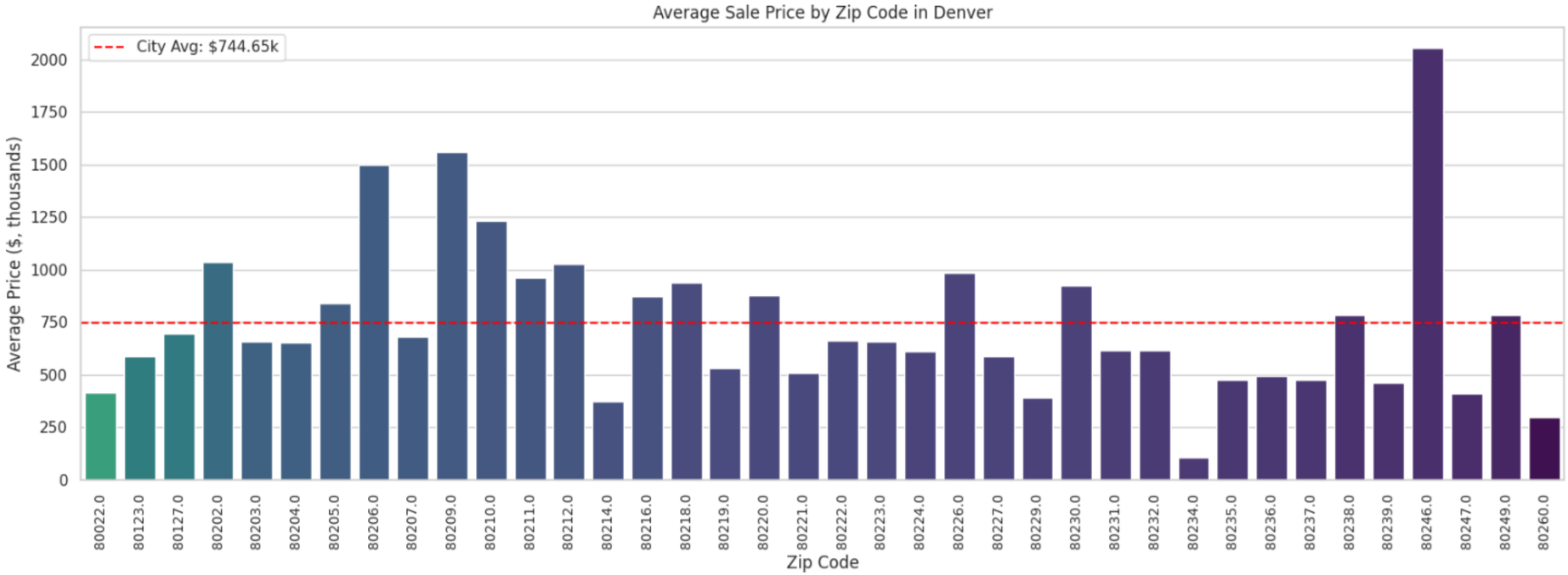
The Data:

- **We then wanted to get a feel for the data by City and prices by zip code.**



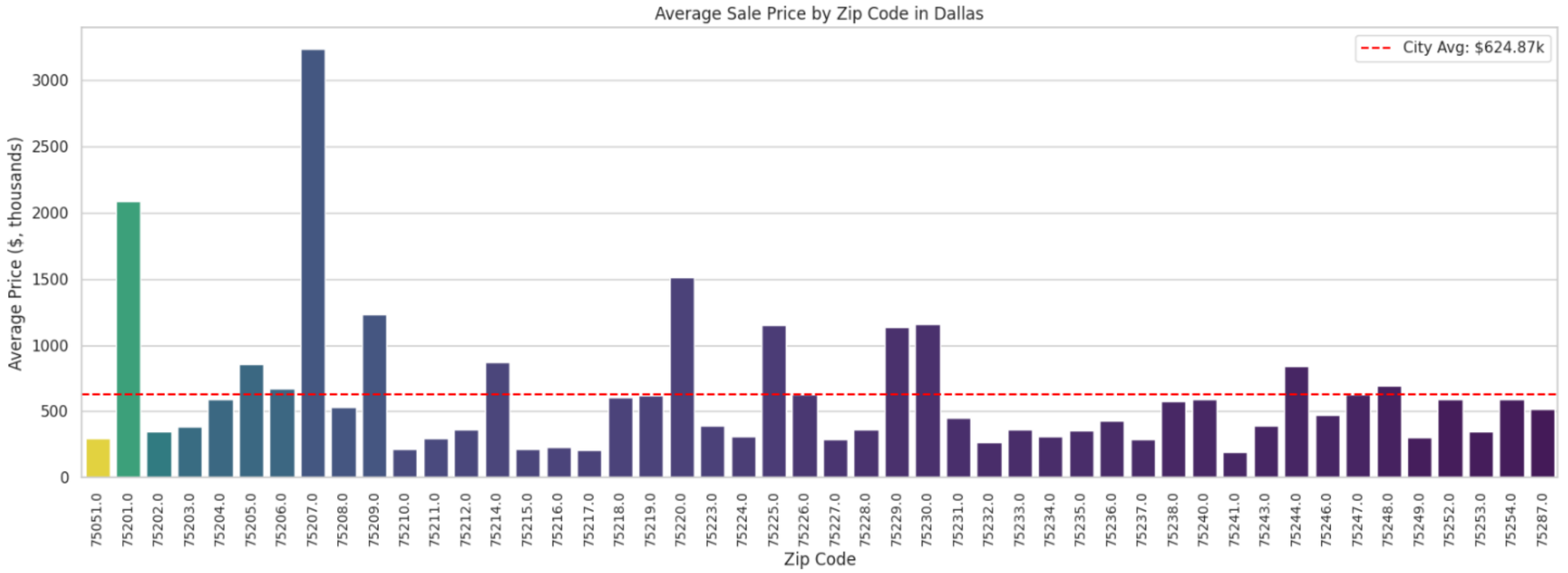
The Data:

- **We then wanted to get a feel for the data by City and prices by zip code.**



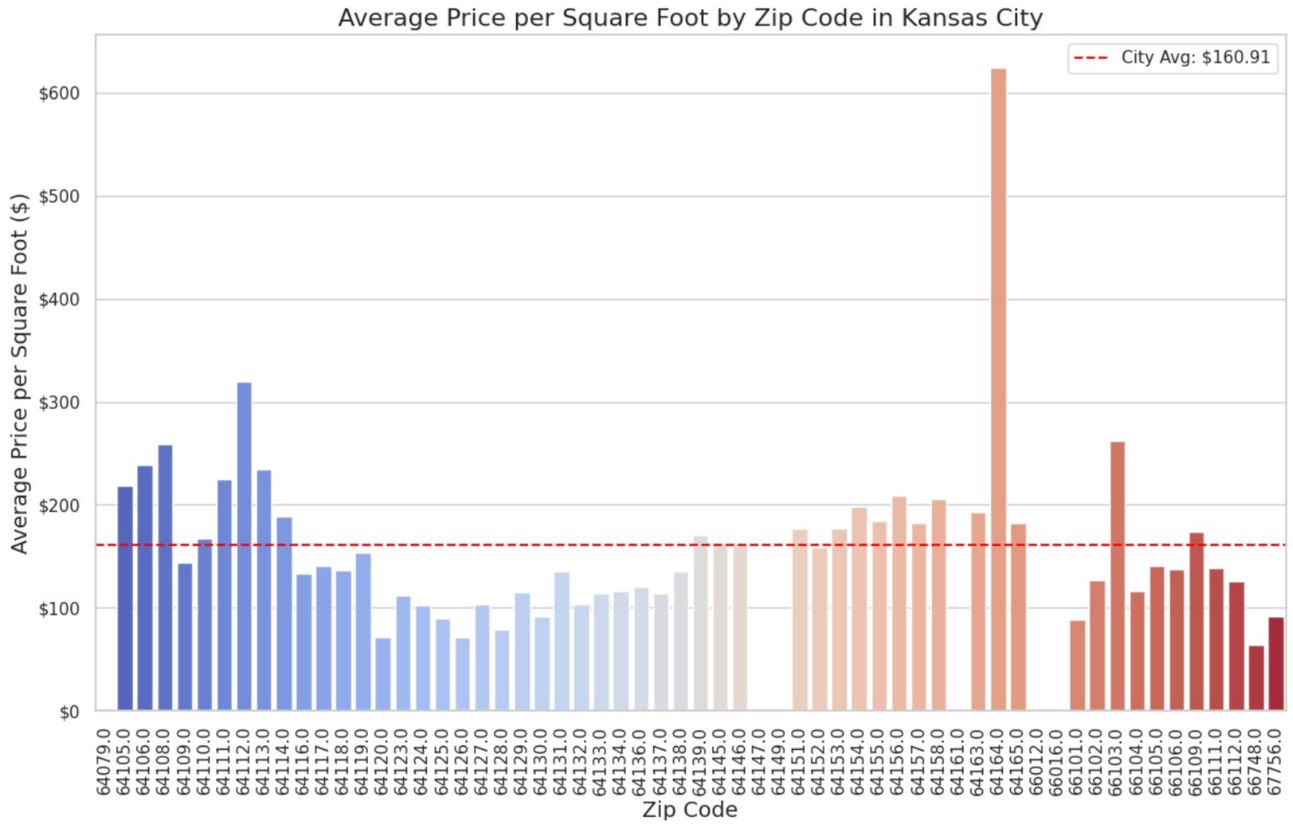
The Data:

- **We then wanted to get a feel for the data by City and prices by zip code.**



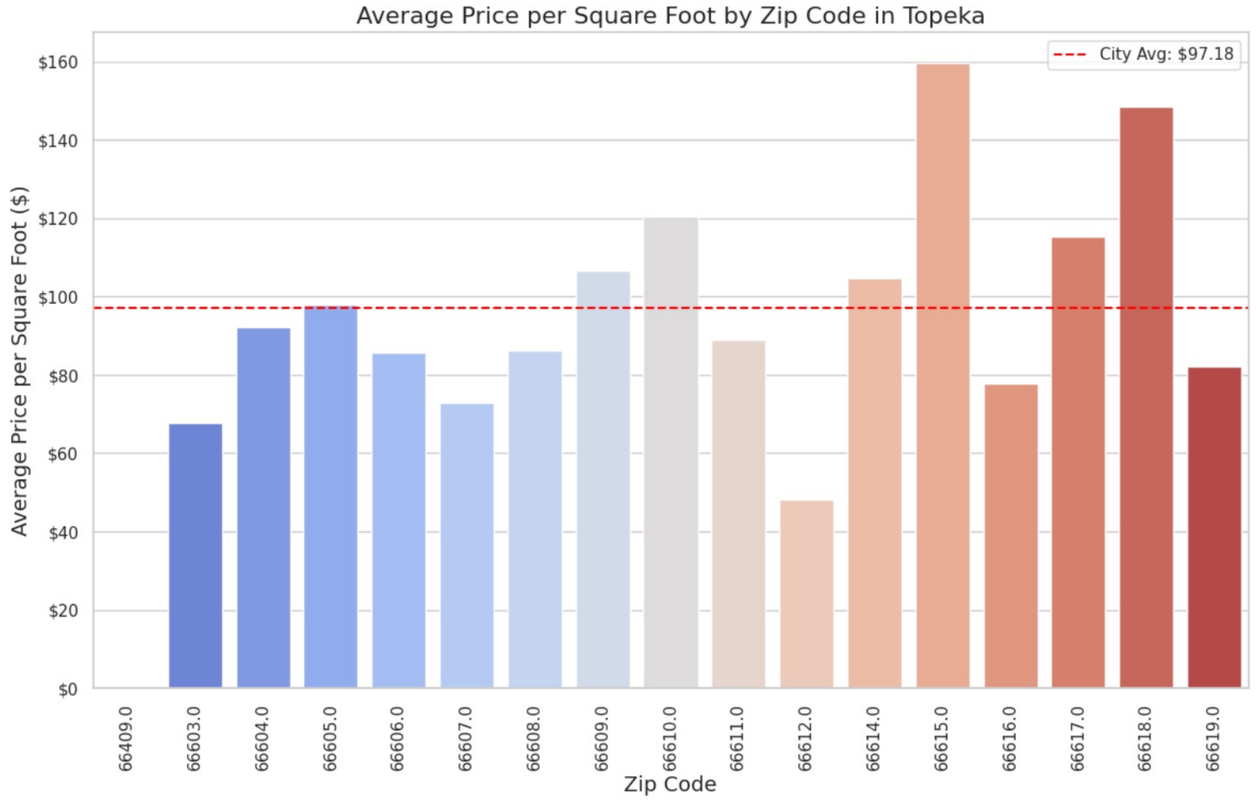
The Data:

- **When then wanted to get a feel for the data by City and prices/sqft by zip code.**



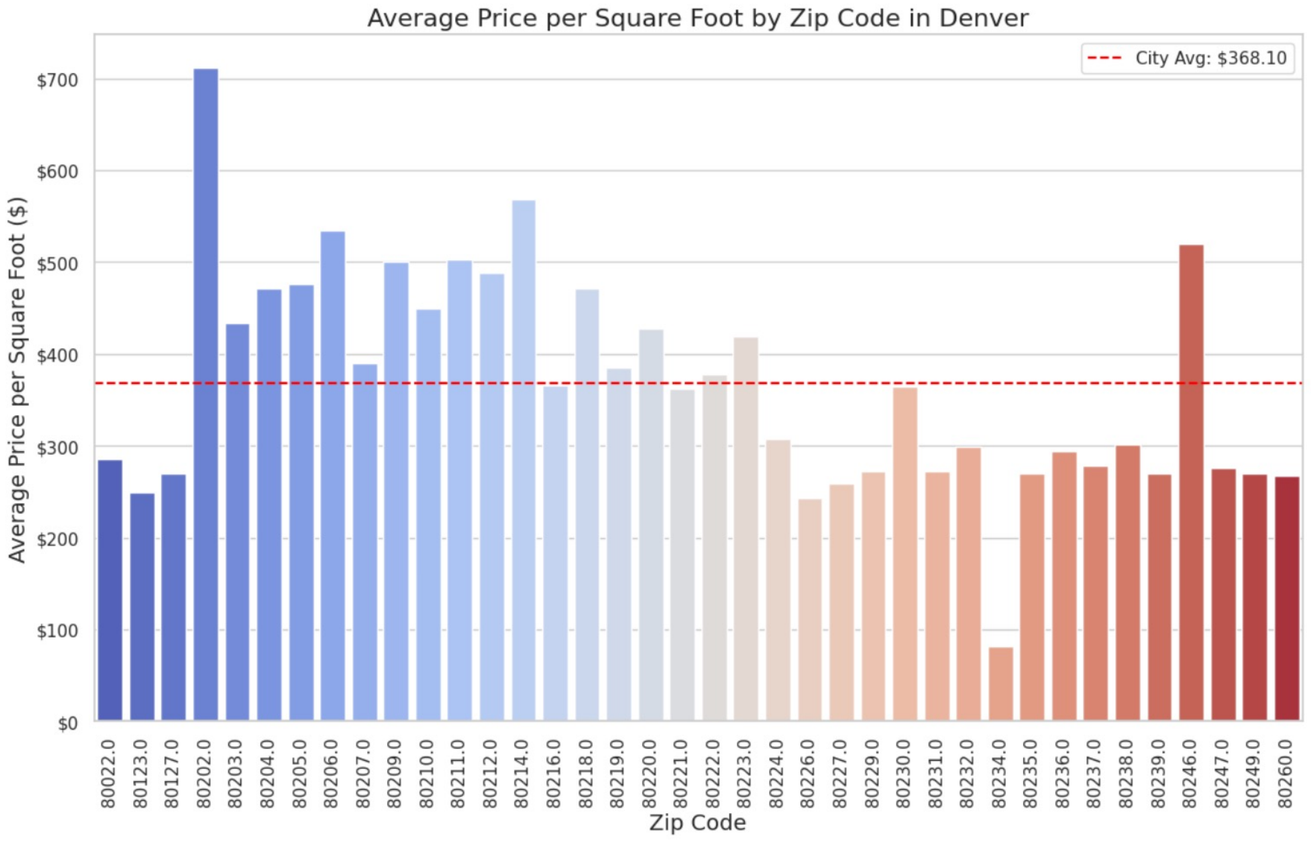
The Data:

- **We then wanted to get a feel for the data by City and prices/sqft by zip code.**



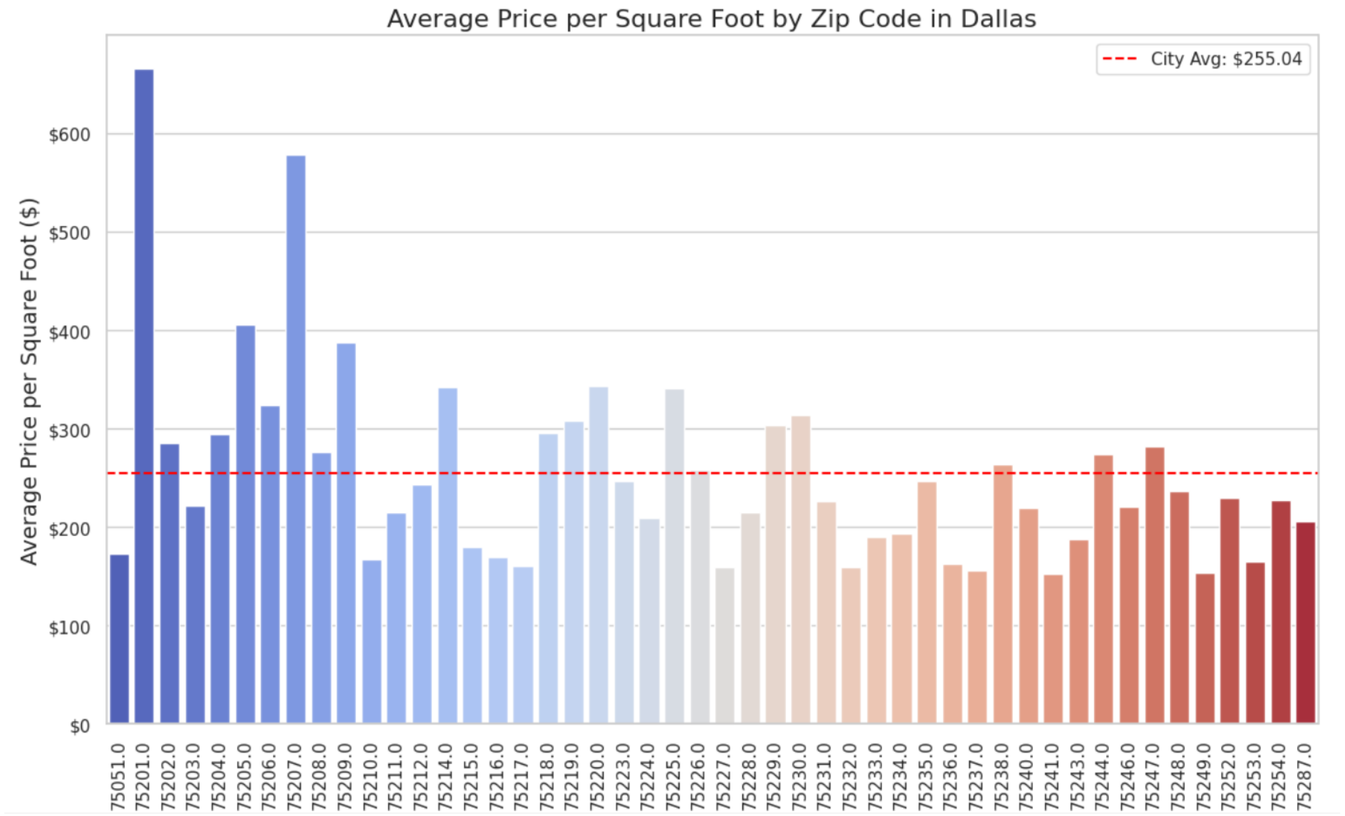
The Data:

- **We then wanted to get a feel for the data by City and prices/sqft by zip code.**



The Data:

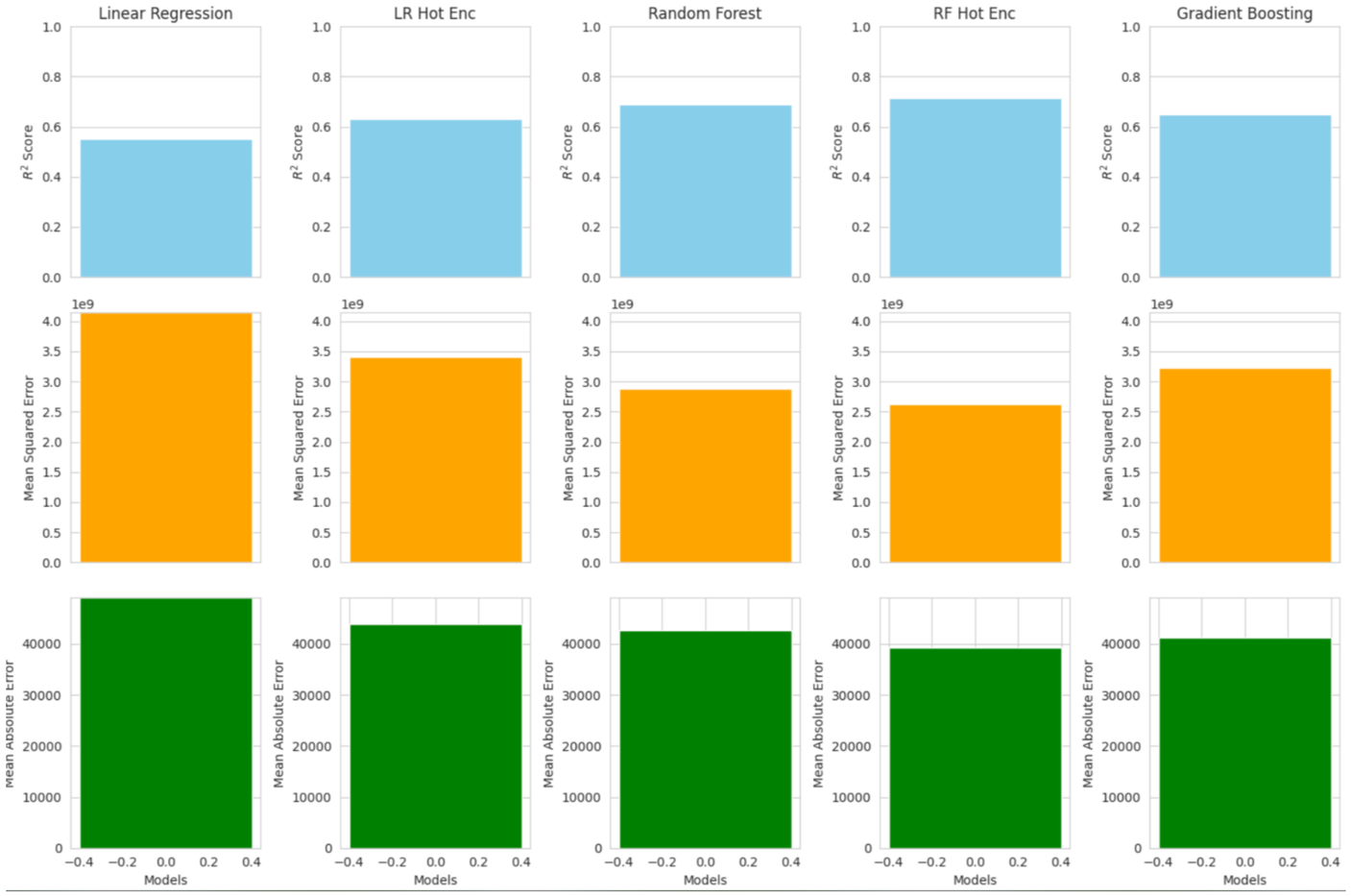
- **We then wanted to get a feel for the data by City and prices/sqft by zip code.**



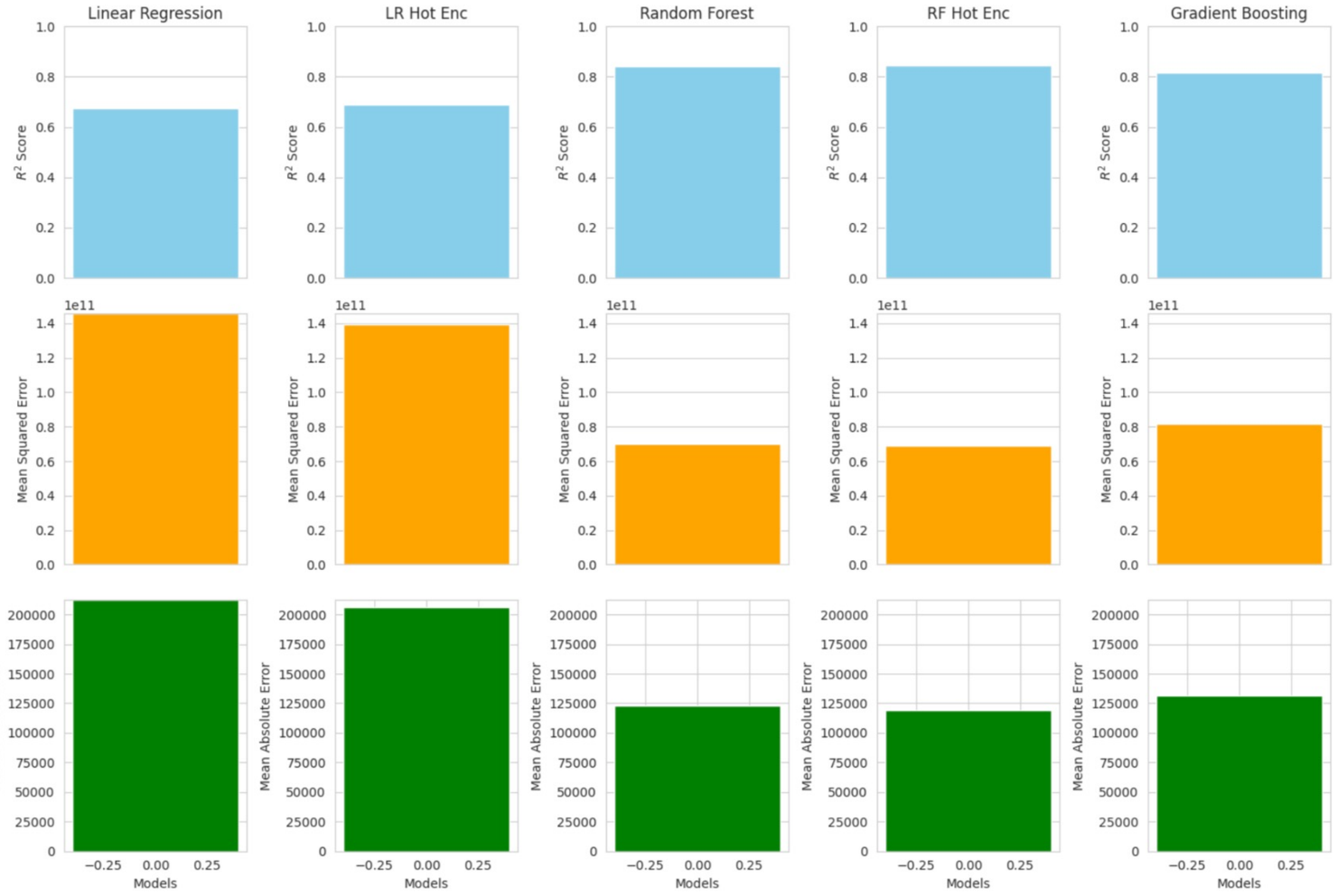
The Models

- Linear Regression
  - One Hot Encoding
- Random Forest
  - One Hot Encoding
- Gradient Boosting Regression

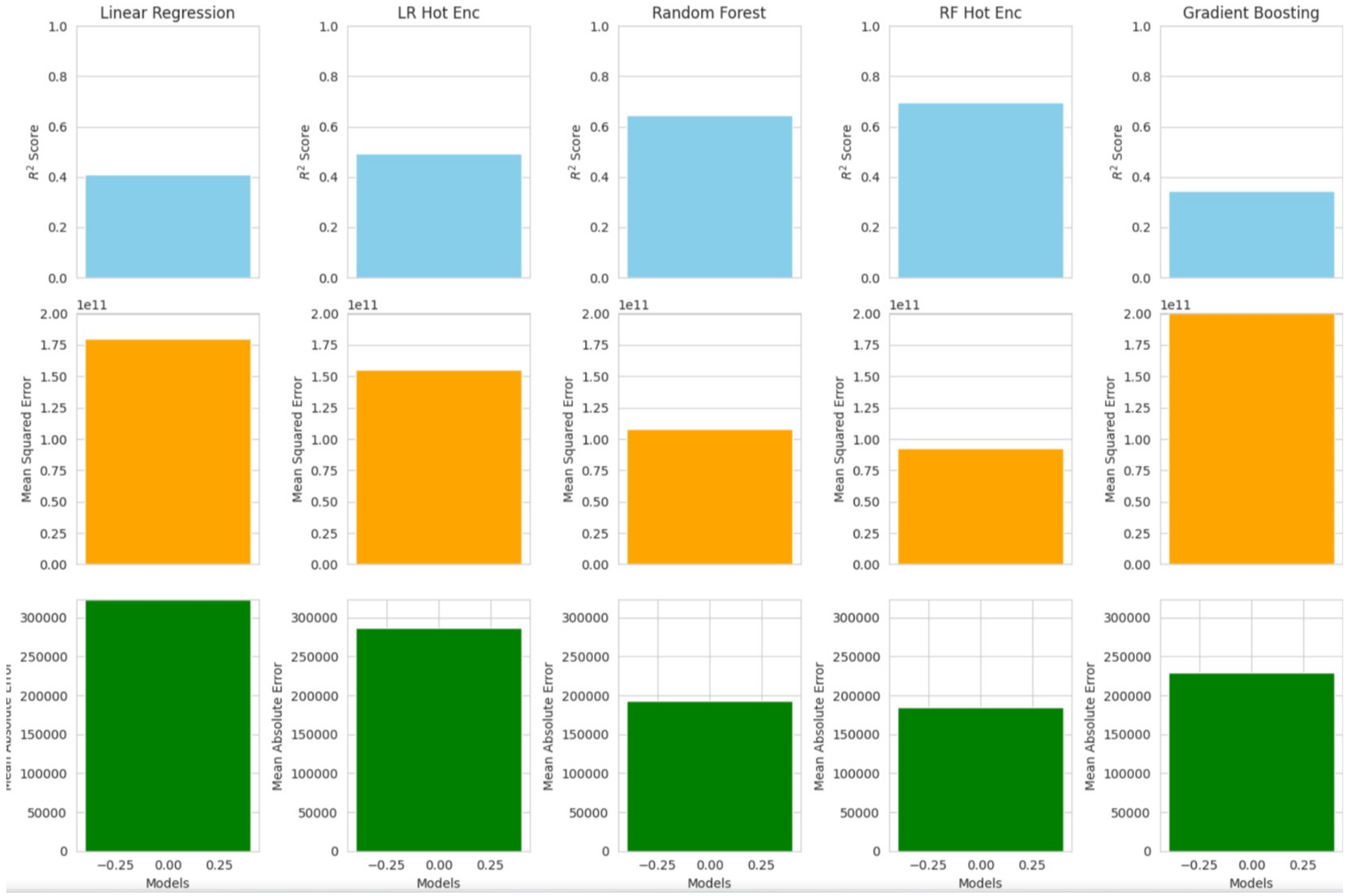
The Models: Topeka – 66609, 4 bed, 2000 sqft



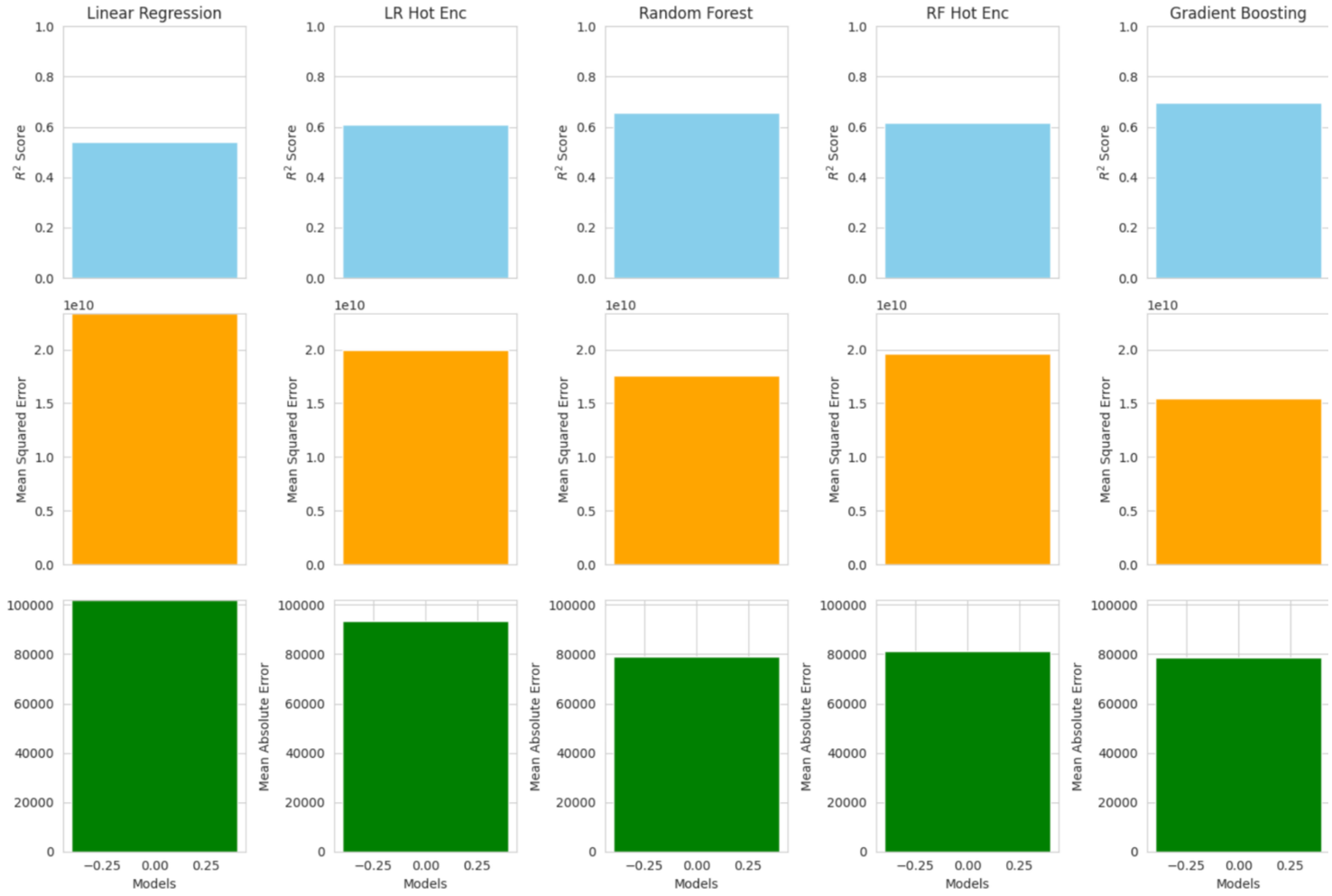
The Models: Dallas – 75218, 4 bed, 2000 sqft



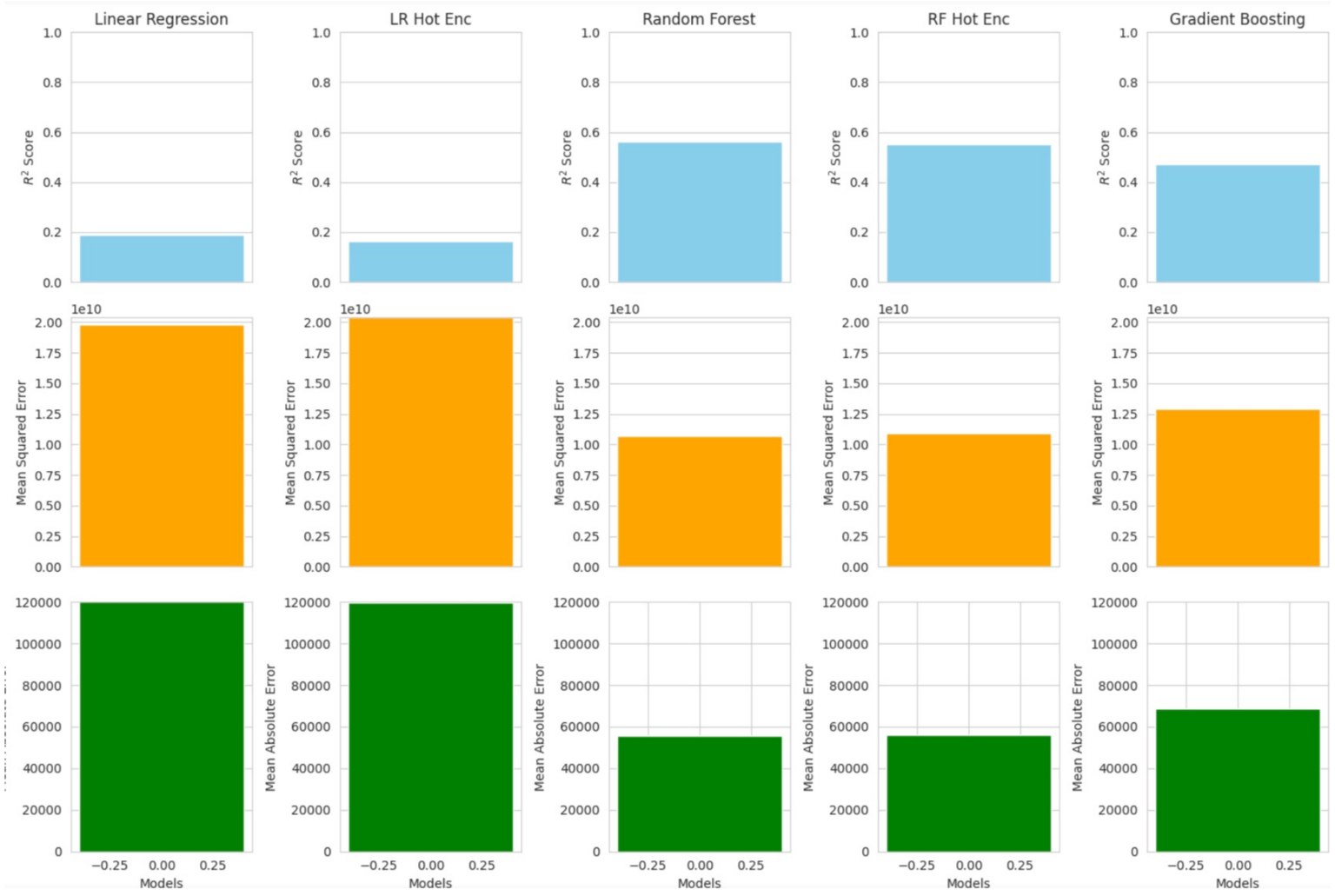
The Models: Denver – 80204, 4 bed, 2000 sqft



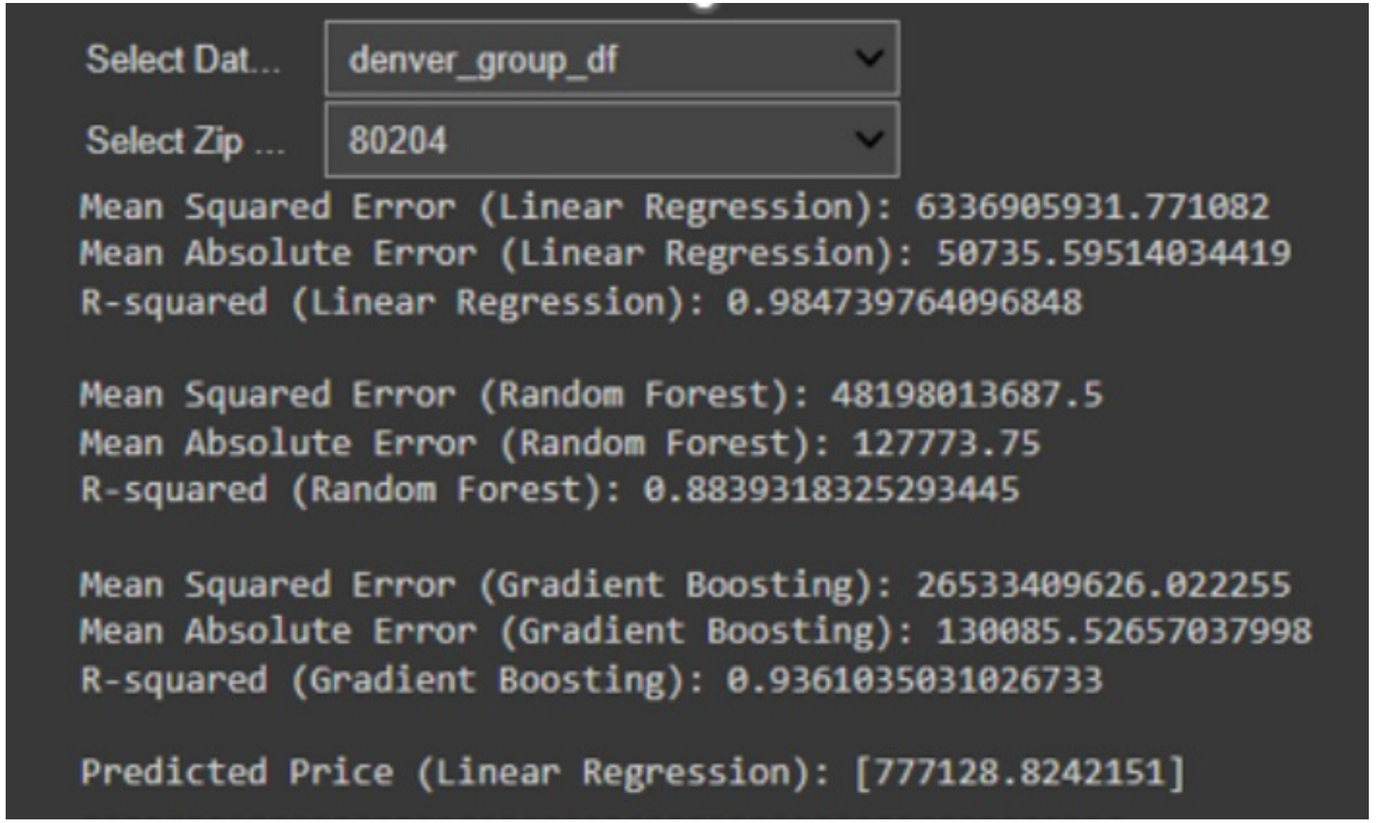
The Models: Kansas City, MO – 64128, 4 bed, 2000 sqft



The Models: Kansas City, KS – 66102, 4 bed, 2000 sqft



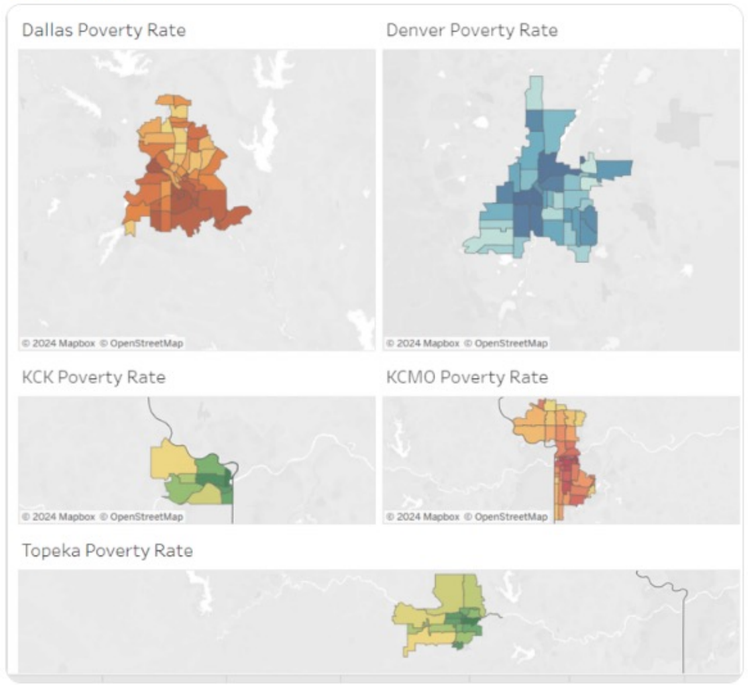
The Analysis: User Interaction!



The Analysis:

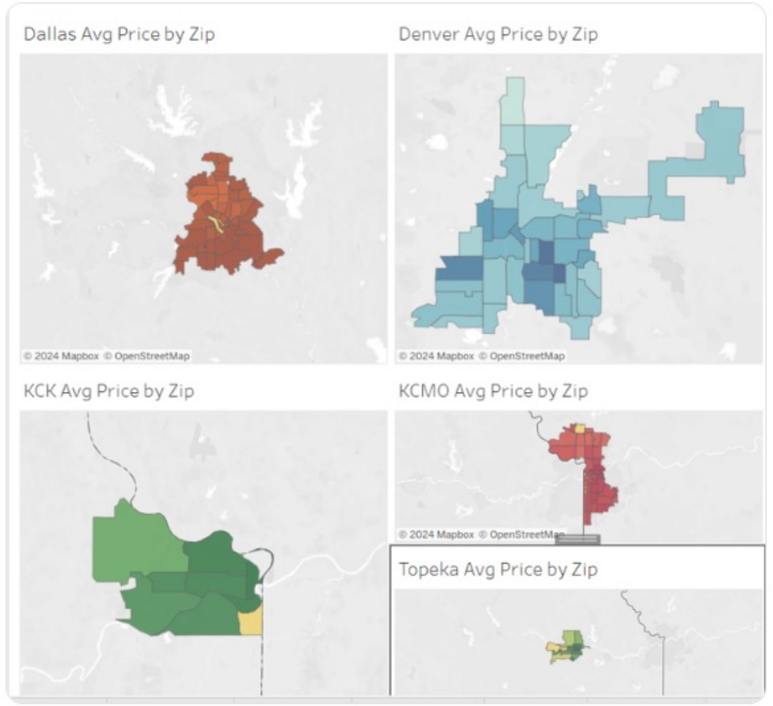
Avg Price by Zip - Interactive

image.png



Poverty Rate by Zip - Interactive

image.png



Summary:

- **The better performing models tended to be random Forest and Gradient Boosting**
- **The better technique for preprocessing is to add “One Hot Encoding” on the zip codes**

Next Steps:

- We limited the data to our particular needs to solve our problem.
  - Next time we think not to limit the data would give us a more dynamic model.
    - This would allow any zip, in any city, for any house size or bedroom quantity.

- We searched for other data we thought might influence home prices, however the free data was hard to find. Examples:
  - Actual Addresses
  - Crime Rates
  - Market Supply and Demand
  - Possibly adding indexes such as consumer pricing