Real Estate Rental Price Analysis in the US

CIS 5450 Big Data Analytics

Prekshi Vyas, Shriya Ramakrishnan & Zed Liu

University of Pennsylvania

CONTENTS

/01

Project Introduction

Aim and motivation of our project

Database we used

Aims, Motivations & Stakeholders

01. Understanding Market Dynamics

By examining variables like location, property characteristics, crime rates, economic status, and climate, the project seeks to unveil the complex interplay of factors affecting real estate rental pricing.

02. Informing Investment Decisions

For investors and property owners, insights into what drives rental prices can guide investment strategies and property management decisions to maximise returns.

03. Policy Making and Urban Planning

For policymakers and urban planners, understanding these dynamics can aid in formulating policies that promote balanced development and address housing affordability issues.

04.Our Stakeholders

Real Estate Investors and Property Owners, Renters and Housing Seekers, Policy Makers and Urban Planners, Real Estate Professionals, Academic Researchers and Economists.

Data We Used

Rental Price of Real Estate

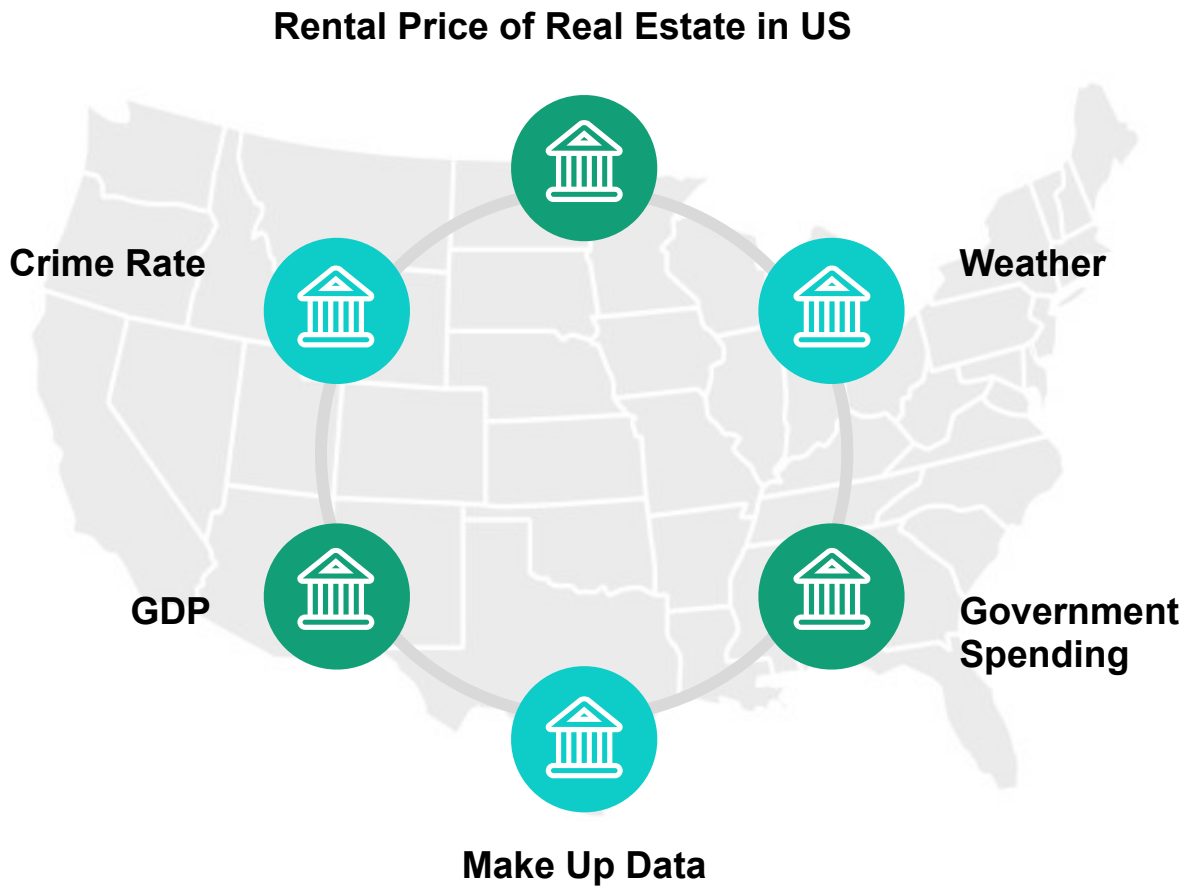
Rental pricing, Geographic location, no of Bedrooms & bathrooms ad other attributes.

Crime Rate

Crime rate statistics across all U.S. states and territories, serving as a proxy for the regional safety.

State GDP

Encapsulates the Gross Domestic Product (GDP) per capita for each state.



Weather

Covers average temperature for all states for last 10 year to evaluate the climate condition and liveability.

Spending

Conveys the government expenditures across various sectors within each state for the year 2022

Make up data

Make up data set for missing data in weather.

/02

Data loading and Preprocessing

Data from Kaggle, Statista, Tax Policy Center, Government database and Wikipedia

Data Loading & Preprocessing Pipeline

The final dataframe retains a considerable scale i.e. **125687 rows × 26 columns**, providing ample data for a thorough exploration of the underlying relationships.

Data Acquisition

Loaded data from online database, local csv files and webpage.

01

Data Cleaning

Dropped unnecessary columns and meaningless instance. Deal with missing data using additional databases.

02

Data merging

Integrated and merged different dataframes using a standard foreign key: the abbreviation of state name.

03

Highlights in Data Loading and Preprocessing

List of U.S. states and territories by GDP

10 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

"List of states by GDP" redirects here. For the list of countries worldwide by GDP, see Lists of countries by GDP.

This is a list of U.S. states and territories by gross domestic product (GDP). This article presents the 50 U.S. states and the District of Columbia and their nominal GDP at current prices.

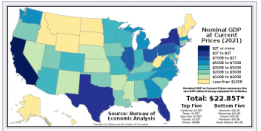
The data source for the list is the Bureau of Economic Analysis (BEA) in 2022. The BEA defined GDP by state as "the sum of value added from all industries in the state."[1]

Nominal GDP does not take into account differences in the cost of living in different countries, and the results can vary greatly from one year to another based on fluctuations in the exchange rates of the country's currency. Such fluctuations may change a country's ranking from one year to the next, even though they often make little or no difference in the standard of living of its population.[2]

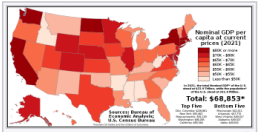
Overall, in the calendar year 2022, the United States' Nominal GDP at Current Prices totaled at $25.463 trillion, as compared to $23.315 trillion in 2021.

The three U.S. states with the highest GDPs were California ($3.6 trillion), Texas ($2.356 trillion), and New York ($2.053 trillion). The three U.S. states with the lowest GDPs were Vermont ($40.6 billion), Wyoming ($47.4 billion), and Alaska ($63.6 billion).

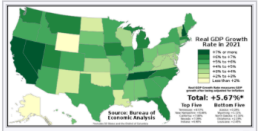
GDP per capita also varied widely throughout the United States in 2022, with New York ($105,226), Massachusetts ($99,274), and North Dakota ($96,461) recording the three highest GDP per capita figures in the U.S., while Mississippi ($47,572), Arkansas ($54,644), and West Virginia ($54,870) recorded the three lowest GDP per capita figures in the U.S. The District of Columbia, though, recorded a GDP per capita figure far higher than any U.S. state in 2022 at $242,853.



The GDP of each U.S. state and the District of Columbia in 2021 according to the U.S. Bureau of Economic Analysis



The GDP per capita of each U.S. state and the District of Columbia in 2021 according to the U.S. Bureau of Economic Analysis and the U.S. Census Bureau



Real GDP Growth Rate by U.S. state according to the U.S. Bureau of Economic Analysis in 2021

50 states and Washington, D.C. [edit]

The following list includes the annual nominal gross domestic product for each of the 50 U.S. states and the national capital of Washington, D.C. and the GDP change and GDP per capita as of 2022.[1][3]

The total for the United States in this table excludes U.S. territories. The raw GDP data below is measured in millions of U.S. Dollars. * indicates "GDP of STATE or FEDERAL DISTRICT" or "Economy of STATE or FEDERAL DISTRICT" links.

	State or federal district	Nominal GDP at current prices 2022 (millions of U.S. dollars)[1]		Annual GDP change at current prices 2022 (2021–2022)[1]		Real GDP growth rate (2021–2022)[1]	Nominal GDP per capita 2022[1][3]		% of national[1]	
		2022	2023				2022	2023	2022	2021
1	California *	3,598,103	3,755,487	▲ 224,862	▲ 11.6%	▲ 1.0%	$92,190	$96,222	14.69%	14.49%
2	Texas *	2,355,960	2,436,346	▲ 304,191	▲ 12.9%	▲ 5.6%	$78,456	$81,130	8.69%	8.55%

Difficulties when dealing with raw data:

- **Lack of a comprehensive dataset on Kaggle for our problem statement**
- **Incomplete databases**

Scrape GDP data from webpage

- First, we built up a helper function to visualize XML structure of the webpage.
- We used XPath to specifically target and extract the GDP data we needed from the web table on the Wikipedia page.

Complete the missing data

- Upon inspecting the gathered data, we noticed that the GDP data for three regions - the District of Columbia, Alaska, and Hawaii - were missing.
- To address the gaps in our data, we found an alternative or supplementary database that provided the missing GDP data

/03

Exploratory Data Analysis (EDA)

Identifying and dropping outliers.

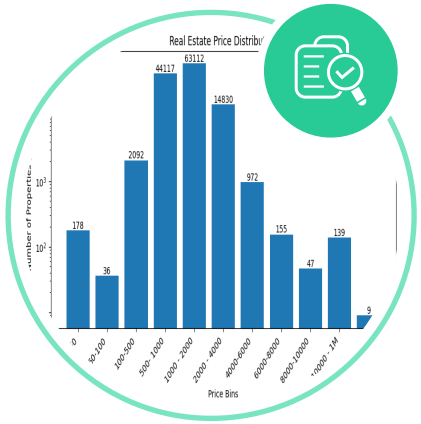
Visualisation to understand our data.

Recognize and Drop Outliers

Necessity of Dropping Outliers

First, we examined the correlation matrix and observed minimal association with our target variable – **rental price**.

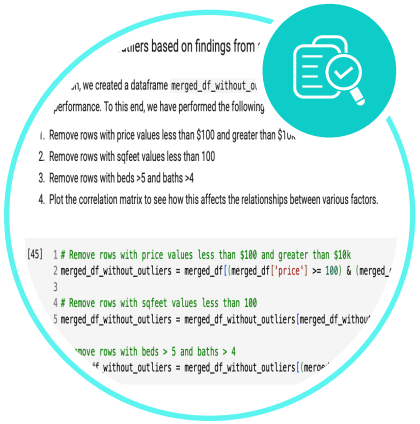
This lack of correlation is likely attributable to the **significant influence of outliers in our dataset**.



Identify Outliers

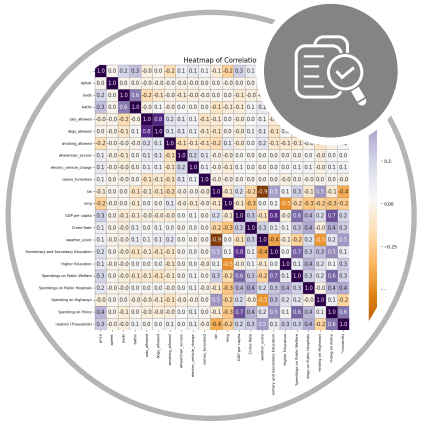
Through the bar charts of **rental prices, square feet, and the number of bedrooms and bathrooms**, we gained insights into the distribution of our dataset.

We segmented our data into several categories and excluded those with **exceedingly low frequencies**.



Dropping Criterion

1. Remove rows with price values less than $100 and greater than $10k
2. Remove rows with sq feet values less than 100
3. Remove rows with beds >5 and baths >4

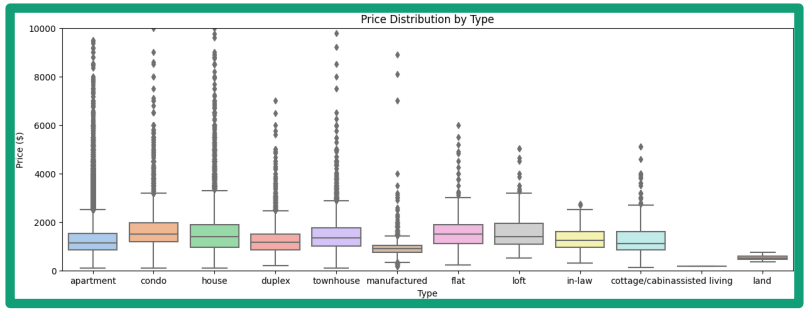


Outcomes

The correlation map reveals that price is associated with socio-economic factors, including **population, police spending, public welfare, and education**.

Additionally, house features such as **bathrooms, bedrooms, laundry, and parking** options also show correlation with price.

Visualisation - Property Level

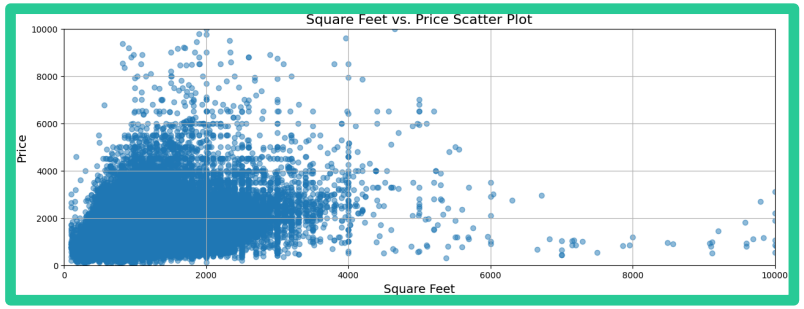


Types vs. Prices

The lowest prices for apartments, condos, houses, duplexes, townhouses, lofts, and cottages/cabins are surprisingly similar and low. Land has the highest minimum price.
Manufactured homes and land show concentrated price structures, reflecting more stable prices across the US due to lower variability in costs.

Parking vs. Prices

Most parking options have similarly low minimum prices, except for valet parking (around 600 USD), which commands higher prices due to enhanced services.
Properties with attached garages and off-street parking have more high-priced outliers, as these features are considered premium amenities.



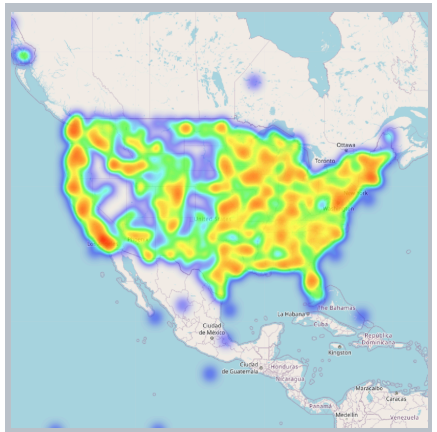
Square Feet vs Prices

There is a visible trend that as the square feet increase, the price also tends to increase. The correlation may be positive but not very tight, indicating other factors may also significantly affect the price.

Visualisation - State Level

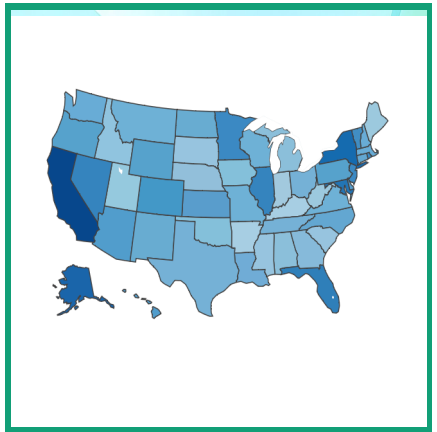
Location of Data

From the map we can infer that the distribution of data is uniform across all the states of United States.



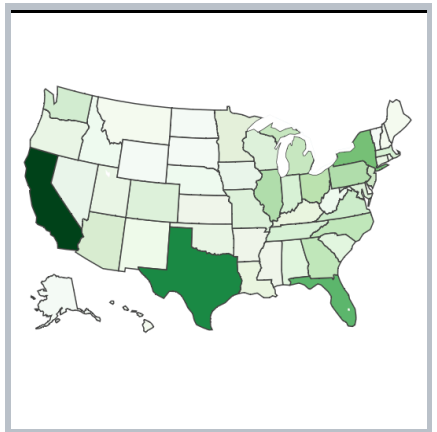
Price Heatmap

Heatmap indicates high rental prices on the West and East Coasts, moderate prices in cities like Portland, and lower prices in the Midwest and southern states.



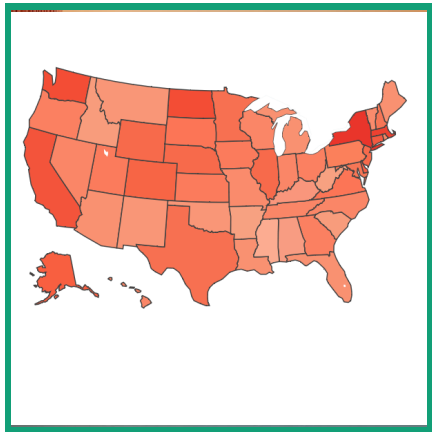
Spending - Police

States with greater spending on police have higher prices implying that they maybe safer places to live in and hence driving up housing prices.



Population

States with greater population have higher prices implying that there is a larger demand of housing.



GDP

State with a higher GDP has higher housing prices and this can be explained by the income and spending ability of the people residing in those states that drive up the housing prices.

/04

Feature Engineering

Data Encoding and Evaluation

Preparation for Modeling

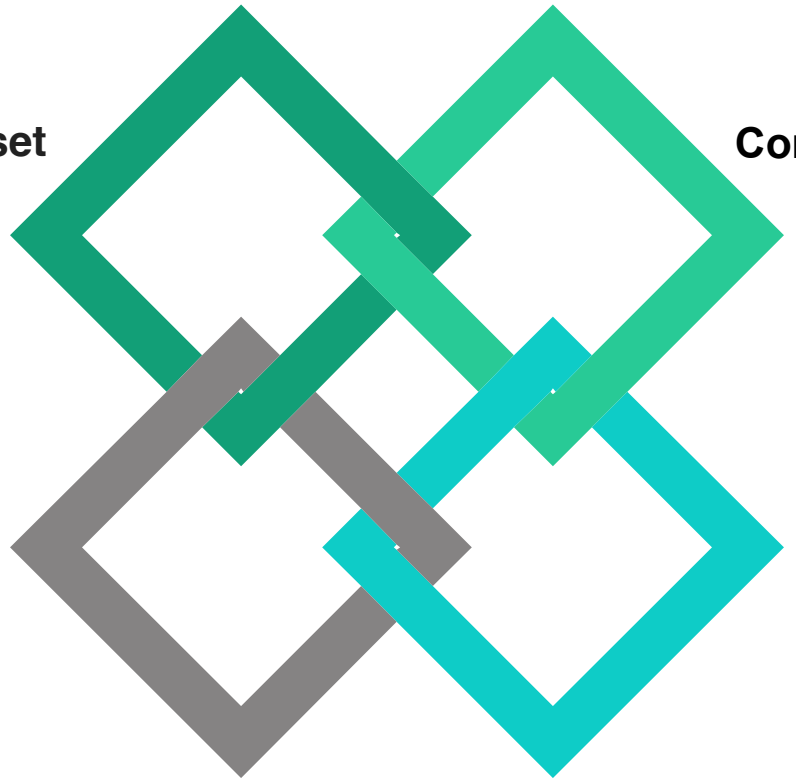
Evaluation and Encoding

Encoding for Real Estate Dataset

For the `parking_options` and `laundry_option` columns, we rate different options with different points according to their general preference in real life.

Evaluation for Weather Dataset

Assessing an area's liveability through monthly weather patterns is more accurate than using annual temperature averages, which can be skewed by extreme seasonal changes.



Combining Highly correlated Features

Based on the visualisation output of the correlation matrix, we can infer that `dogs_allowed` and `cats_allowed` are highly correlated, thus these features can be combined into one feature `pets_allowed`

One Hot Encoding

We applied one hot encoding on `type` column, because this is a categorical column, and we do not perform any evaluation on it.

Preparation for Modeling

Create Training & Test Data

We split our datasets in 80-20 ratio. This is a highly conventional split ratio, so we felt it to be an apt starting point.

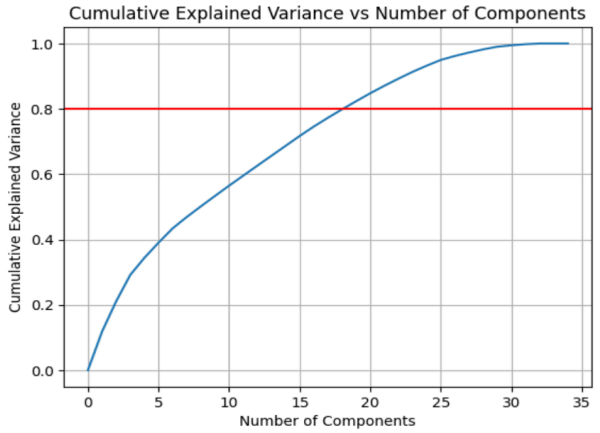
In the interest of producing reproducible results, we set our seed = 42.

PCA

With our data now split into training and testing data, we also want see whether the application of PCA can help improve the performance of our model.

We need to standardise both datasets before applying PCA. The reason for standardising is because PCA is not scale-invariant.

We also used visualisation to find the best component number.



/05

Modeling

Linear Regression, XGBoost,

Random Forest & Feedforward Neural Network

Base Line Model: Linear Regression

R^2: 0.38

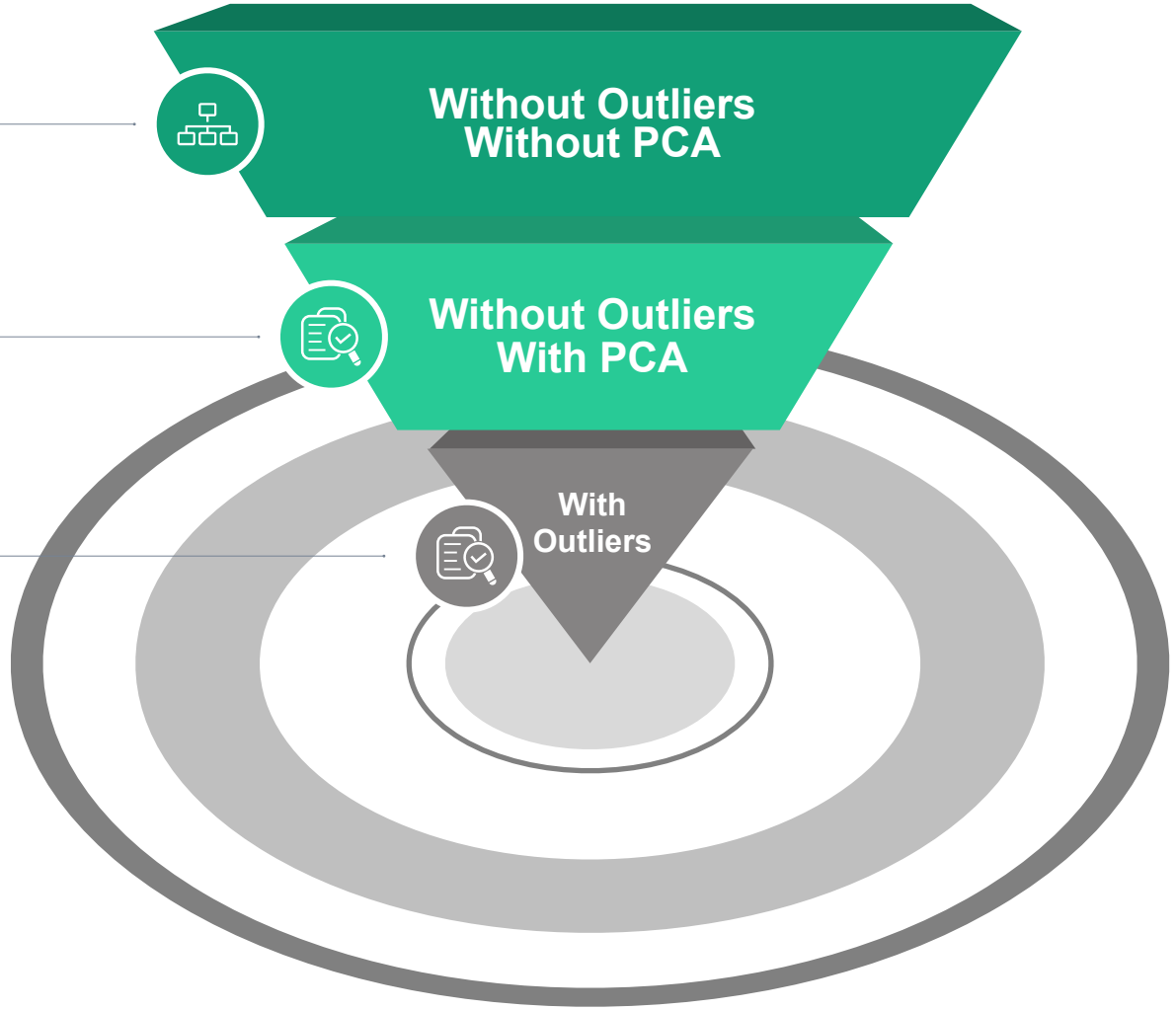
The best linear regression model we have.
Further prove PCA is unnecessary for this model.

R^2: 0.33

Dropping outliers greatly improve the predicting ability of linear regression model.
However, PCA does not benefit our model since most of our features are categorical.

R^2: -26.1

A strong indication that the model is performing worse than a simple mean model.
This suggests that the outliers are having a significantly adverse effect on the model's ability to predict the target variable.



Advanced Model

XGBoost Regressor With Two-Fold Cross Validation

Hyperparameter tuning using GridSearchCV

The best parameter setting is:
learning_rate: 0.06; max_depth: 7; n_estimators: 500; objective: reg: squarederror.

Testing result

Testing Set R^2 value: **0.8045401066375166**
Training Set R^2 value: **0.7805637121112315**
Mean Squared Error: **95920.24859932926**

Conclusion

The R-squared value is a relatively high value, indicating a **good fit** to the data.

Most influential factors seen from XGBoost model's output :

- **Spending on Police**
- **GDP**
- **bath**
- **laundry_options**

Advanced Model

Random Forest Regressor
With Two-Fold Cross Validation

Hyper-parameter tuning using GridSearchCV

The best parameter setting is:
max_depth: Default
n_estimators: 300

Testing Result of Model

Testing Set R^2: 0.814;
Training Set R^2: 0.774;
MSE: 91507

Conclusion:

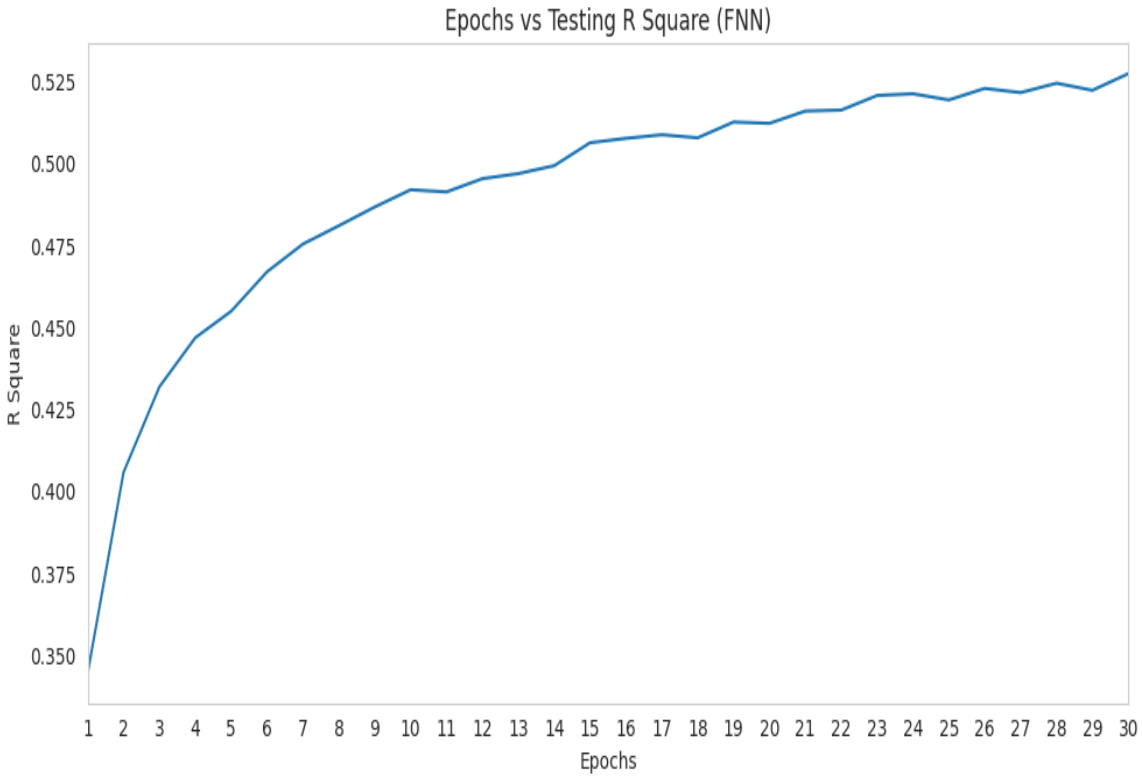
R-squared value is a relatively high value, indicating a good fit to the data.
It is also higher than the value for XGBoost thus, this is a better model.

Most influential factors

Longitude: Most influential, 21.89% importance, correlates with weather conditions.
Property Size (sqfeet): Second most crucial, 21.06% importance, significantly impacts prices.
Other Factors: Police spending, latitude, and laundry options also notably affect pricing.

Feedforward Neural Network

An architecture with 3 layers and ReLU as activation function



01 Hyperparameter Tuning

Using ParameterGrid and iterate through all the parameter combinations to find the best settings.
The best parameters are: **betas: (0.9, 0.9999); dropout_rate: 0.3; hidden1_size: 128; hidden2_size: 64; learning_rate: 0.0001.**

02 Testing Result

After 30 epochs of training, the R^2 **value of testing set is 0.52**.
According to the visualisation of the learning curve, with a greater epoch number, the R^2 **value** will keep increasing.

03 Conclusion

The model showed a positive learning curve, evidenced by declining loss and MSE values and an increasing R-squared value.
However, the model requires a better architecture or tuning to achieve a better prediction performance.

Evaluating the Performance of Models

Linear Regression

Linear Regression's assumption of linearity limits its ability to capture these complex interactions, leading to suboptimal performance.

$R^2: 0.38$

FNN

The model's ability to handle complex, non-linear relationships in datasets depends greatly on the right network architecture and hyper-parameter tuning.

$R^2: 0.53$

XGBoost

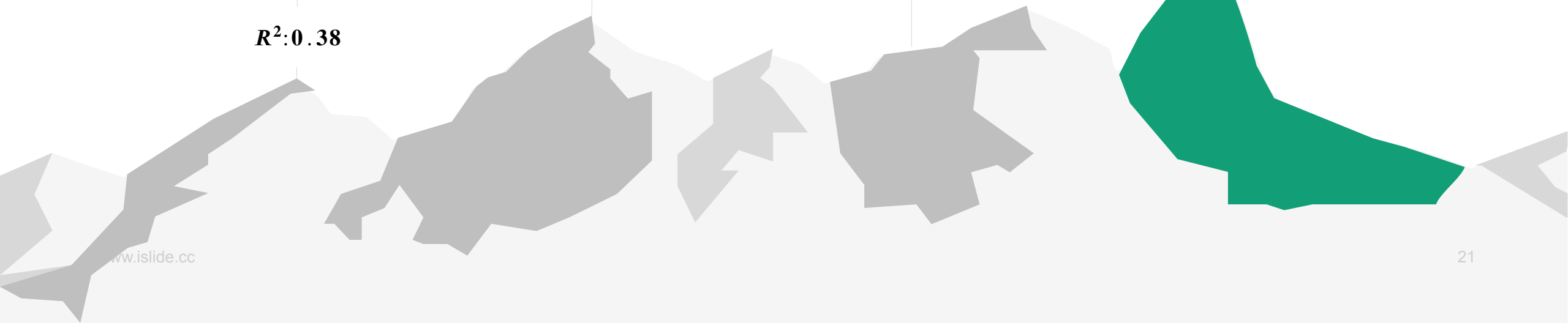
XGBoost excels in handling datasets with mixed variable types and complex patterns, effectively managing features from real estate, economic, and environmental datasets.

$R^2: 0.80$

Random Forest

The model effectively handles complex, non-linear real estate datasets, including economic, geographic, and socio-demographic factors.

$R^2: 0.81$



/06

Conclusion

Message to the stakeholders

Future Work

Message to the Stakeholders

Main Takeaways From Our Rental Price Analysis

Social Economics Factors

From the top model output we can conclude that among the social economic factors, **Spending on Police** (Safety Level) and **GDP** (Economics Level) of the state are the most influential features.

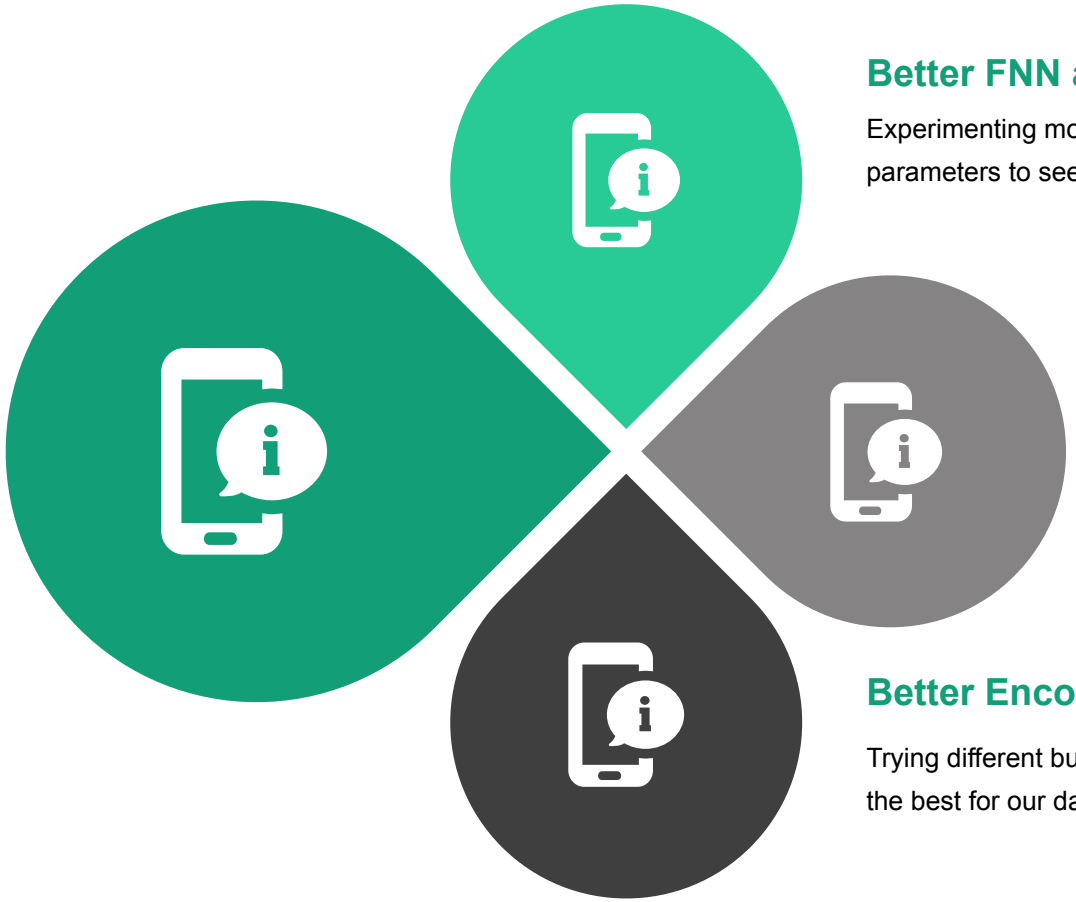
House Properties

Among the house properties, **laundry options, parking_options and bath** seem to be the most influential factors in deciding the rental prices.

Challenges and Future Work

Challenges

1. Difficulties in obtaining data for **specific areas** like counties limited the ability to develop targeted forecasts.
2. Time constraints also prevented the creation of a more advanced **Neural Networks architecture**.



Better FNN architecture

Experimenting more with Neural Network architecture and hyper-parameters to see if it can improve model performance.

More Detailed Data

Considering a smaller geographic location like county or neighbourhood to better account for region specific social economic factors like Crime Rate.

Better Encoding

Trying different business rules for attributes to see which one works the best for our data and consulting domain experts.

Thanks!

Prekshi Vyas, Shriya Ramakrishnan & Zed Liu
University of Pennsylvania