Housing Market
Austin, TX

By
Ammanuel F. WoldeAregay

AGENDA

Introduction

Assess the housing market data from Zillow

Build a price predicting model

Share insights

Final tips & takeaways

Problem Statement

What opportunities exist for real estate developer who aims to make informed investment decisions in and around Austin, TX, considering the potential for housing price increases in 2022 by analyzing the historical housing price data from 2018 to 2021?

Raw Data Sample

	Unnamed: 0.1	Unnamed: 0	address_city	address_neighborhood	address_state	address_streetAddress	address_subdivision	address_zipcode
0	0	15	Austin	NaN	TX	12801 Wooded Lake Ct	NaN	78732
1	1	16	Austin	NaN	TX	904 Lakewood Hills Ter	NaN	78732
2	2	17	Austin	NaN	TX	13701 Montview Dr	NaN	78732
3	3	18	Austin	NaN	TX	700 Lakewood Hills Ter	NaN	78732
4	4	19	Austin	Steiner Ranch-Lakewood Hills	TX	1008 Lakewood Hills Ter	NaN	78732

5 rows × 736 columns

16482 Rows
736 Columns

Data Wrangling

Original dataset had 16482 rows and Rows 736 columns

- Only consider data from data from 2018 to 2021
- Drop redundant columns
- Apply correlation between independent variables and drop features that has less than 0.005 correlation value.
- Drop rows with more than 80% of data missing values
- Create a “price/square feet” feature
- Target Variable: House Price

Data Wrangling

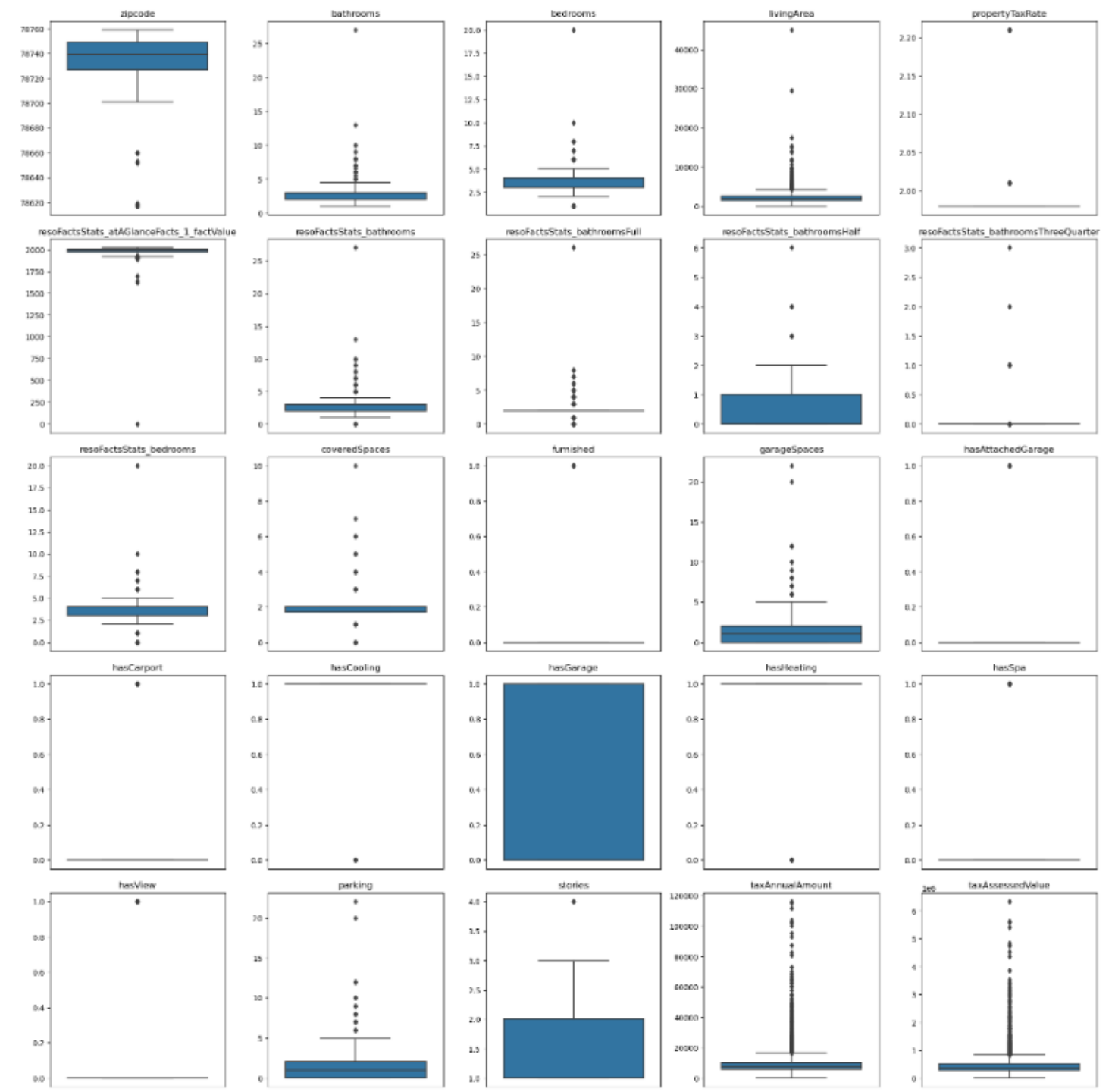
Additional Data

- Austin, TX income by zipcode
- Rows: 11
- Columns: 111

- ❖ Load US Income data to access Austin's dara
- ❖ Merge it with Austin House Listings semi-cleaned data

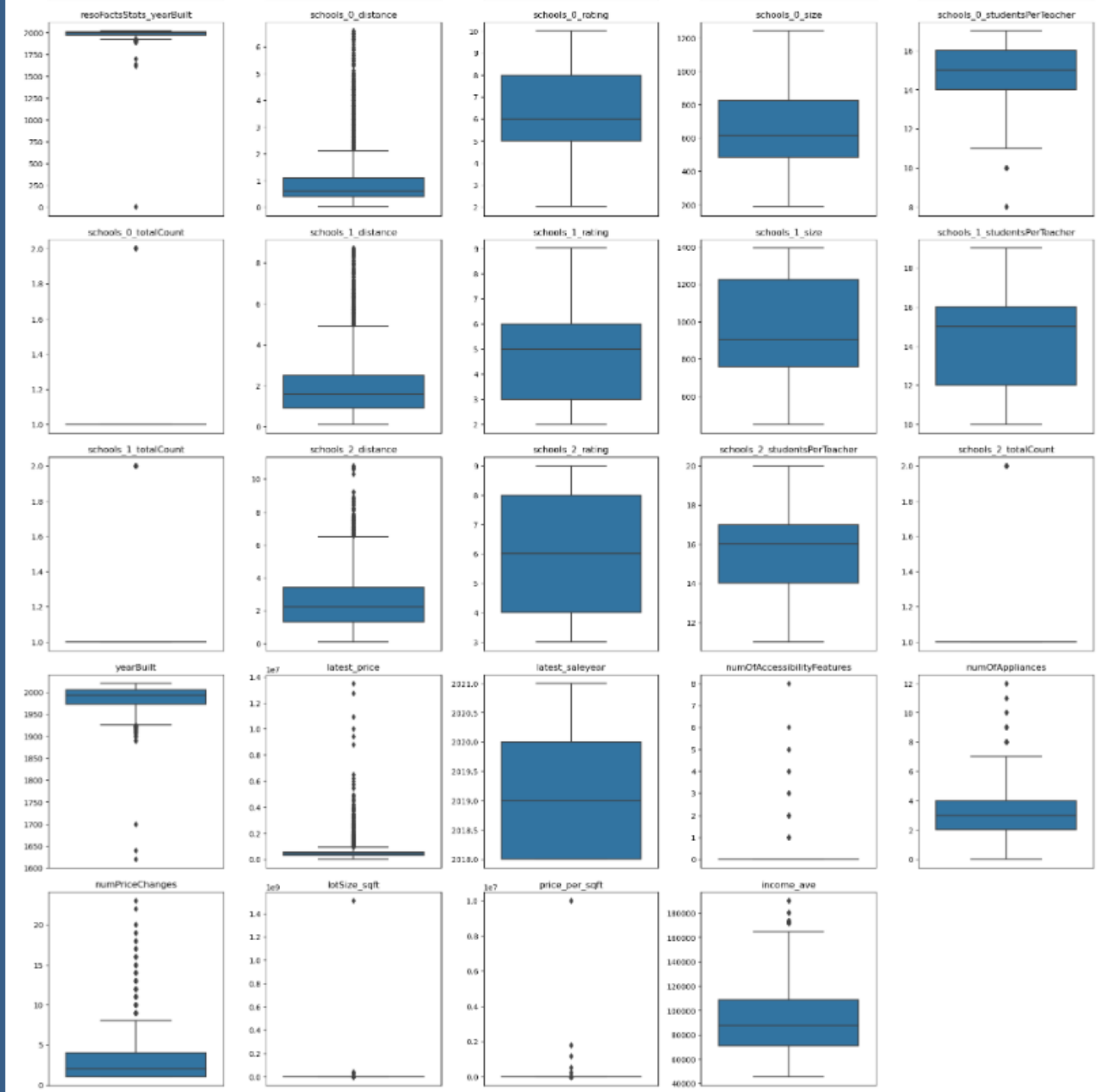
Exploratory Data Analysis

- Analyzing Feature (Independent Variables) Distributions



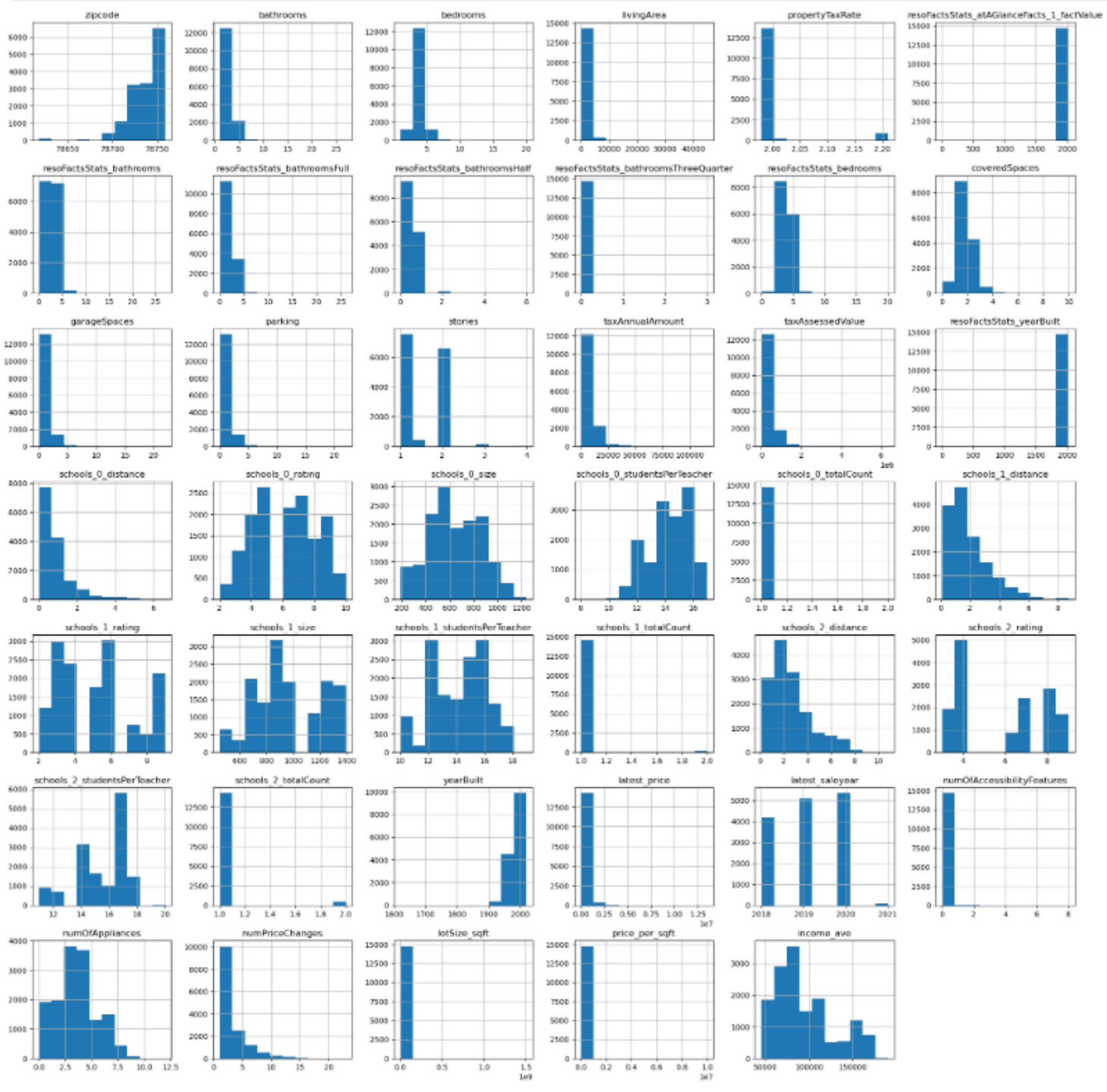
Exploratory Data Analysis

- Analyzing Feature (Independent Variables) Distributions



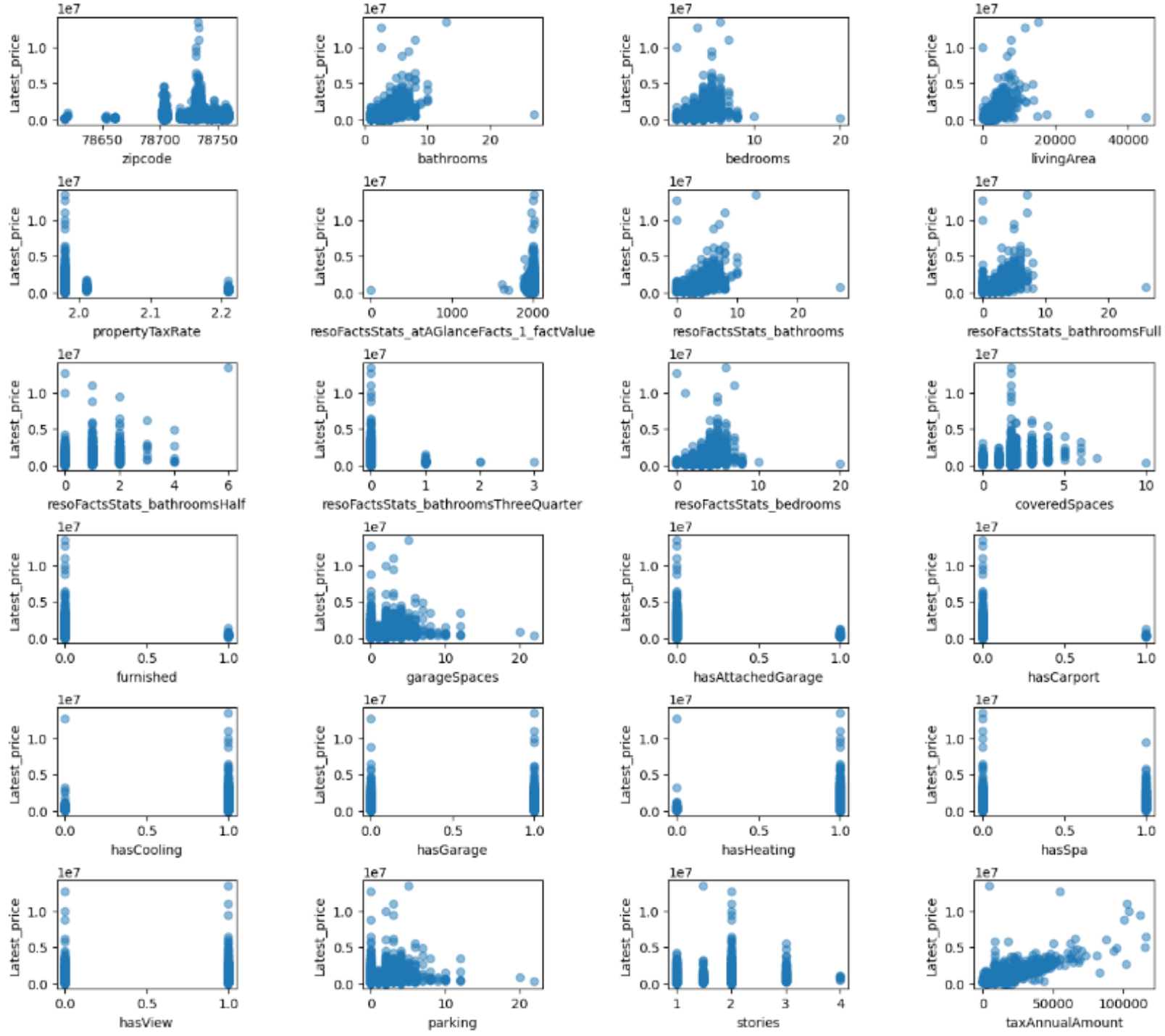
Exploratory Data Analysis

- Comparing features that have numerical values



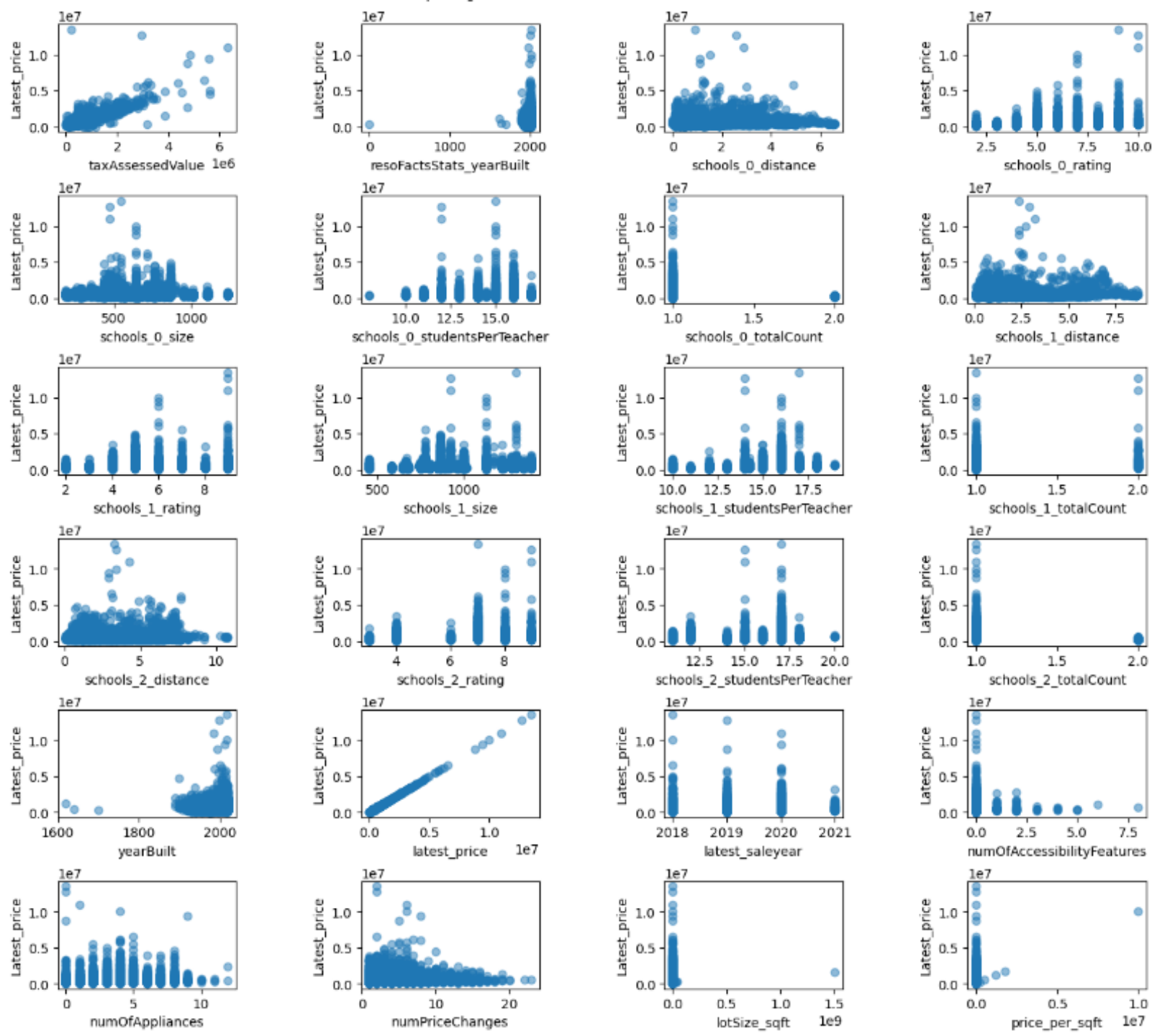
Exploratory Data Analysis

- Comparing individual features (independent variables) with the target (dependent variable)



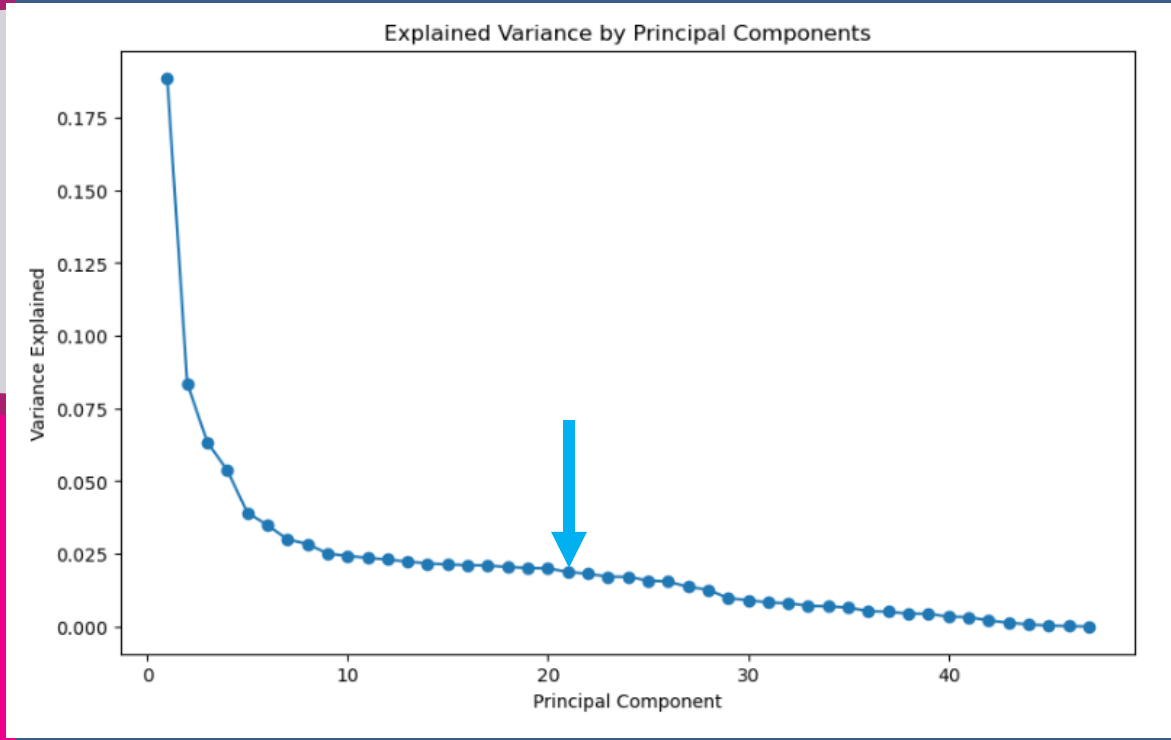
Exploratory Data Analysis

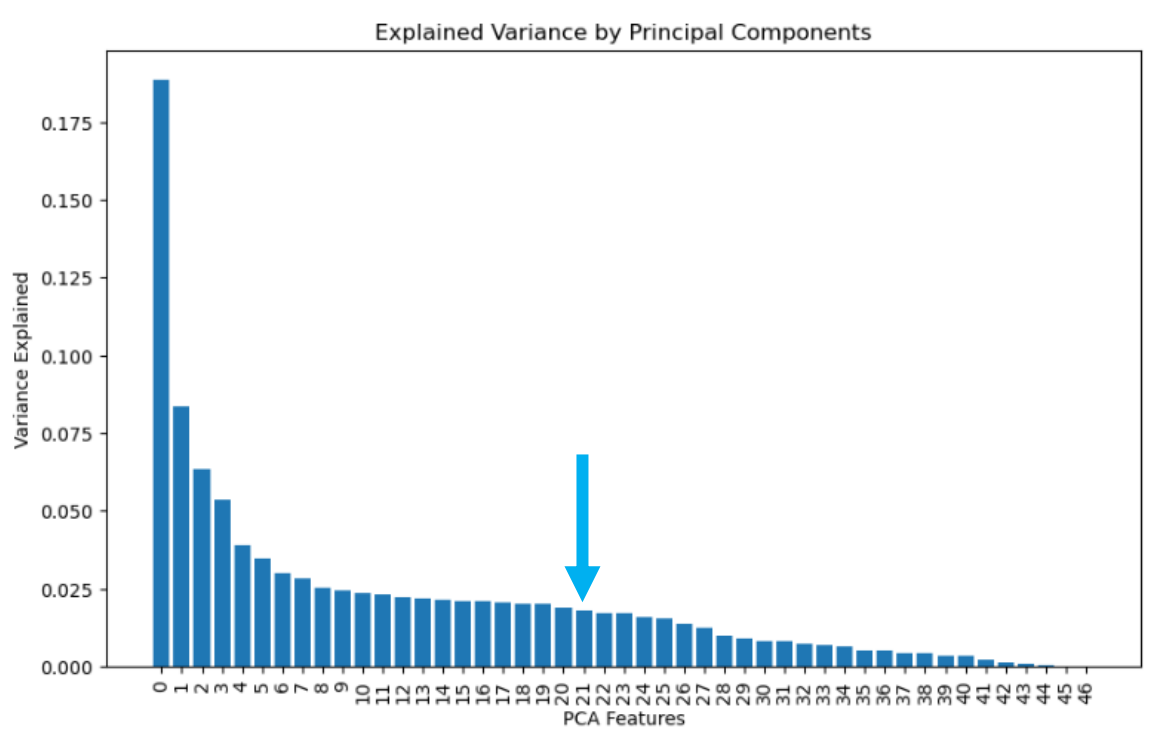
- Comparing individual features (independent variables) with the target (dependent variable)



PCA Components Selection

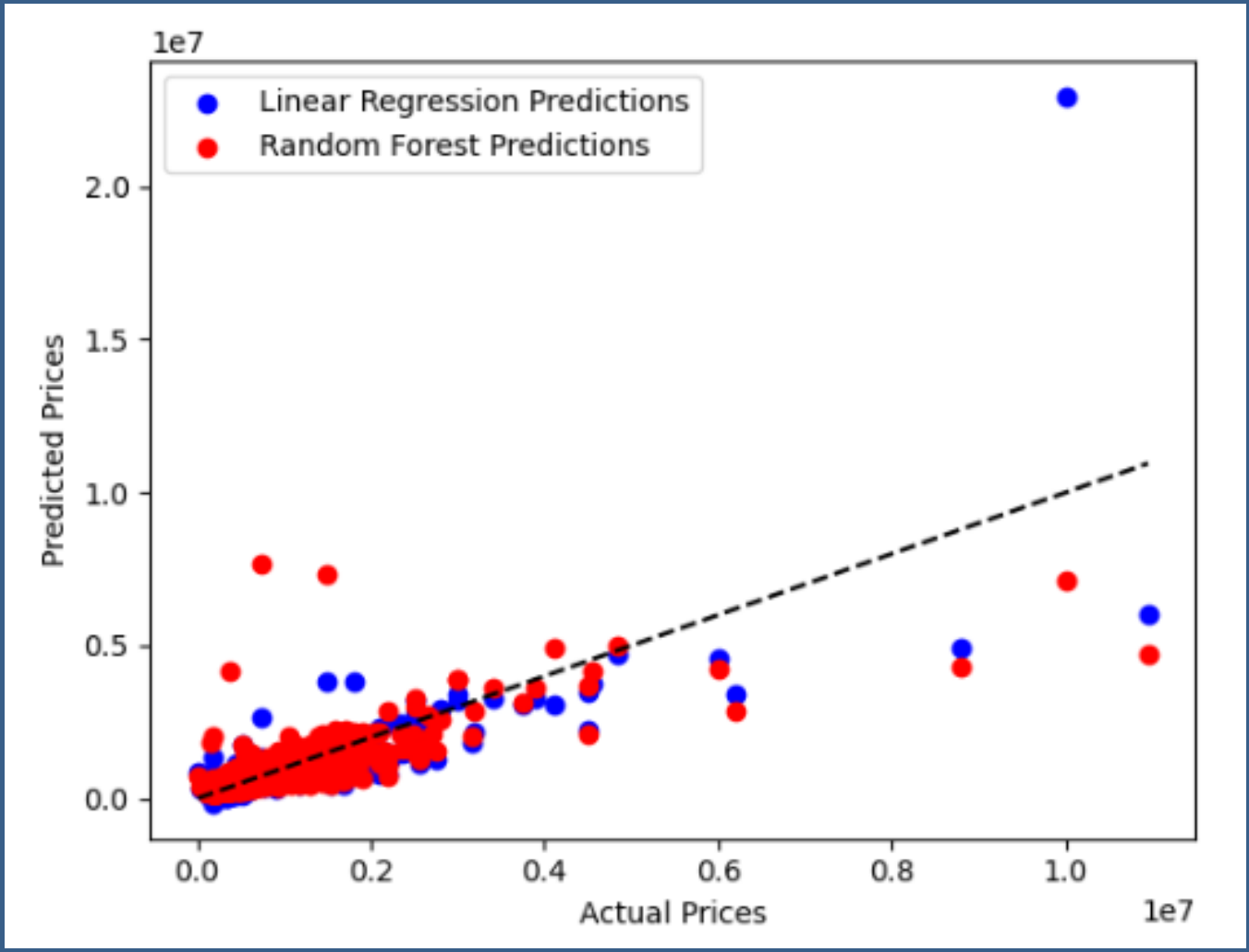
Number of PCA components selected: 21





Model Selection

- **Compared models:**
  - Logistic Regression
  - Random Forest Regressor



Takeaways & Final Tips

- Random Forest Regressor is the preferred model
- Test more parameters

THANK YOU

Ammanuel F. WoldeAregay