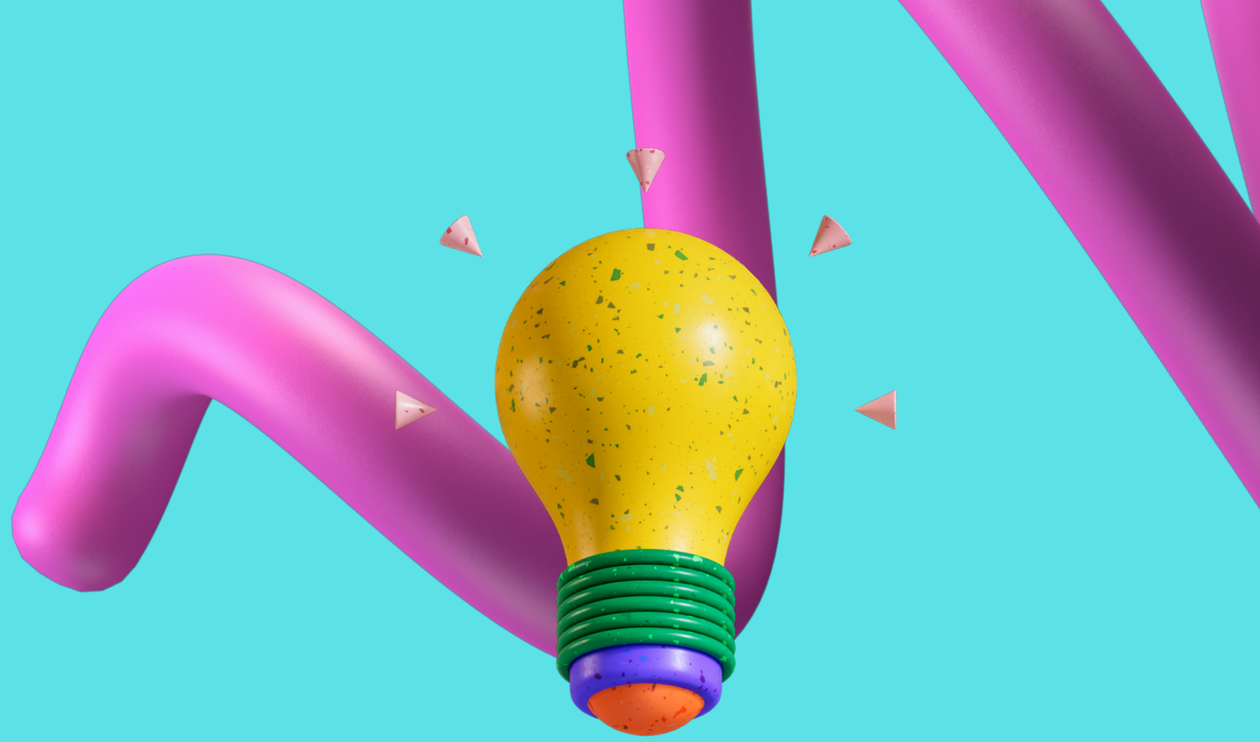
# GoDaddy Microbusiness Density Forecasting

Mohammad Solki
Statistical Data Analysis Course



# Introduction

- GoDaddy is the world's largest services platform for entrepreneurs around the globe.
- Microbusinesses are businesses with an online presence and ten or fewer employees.

- The challenge is to forecast the density of microbusinesses in US counties.
- This work will help policymakers gain visibility into microbusinesses.

# Datasets

## Train Data

| row_id | cfips | county | state | first_day_o... | microbusi... | active |
|---|---|---|---|---|---|---|
| 1001_2019-08-01 | 1001 | Autauga County | Alabama | 2019-08-01 | 3.0076818 | 1249 |
| 1001_2019-09-01 | 1001 | Autauga County | Alabama | 2019-09-01 | 2.8848701 | 1198 |
| 1001_2019-10-01 | 1001 | Autauga County | Alabama | 2019-10-01 | 3.0558431 | 1269 |
| 1001_2019-11-01 | 1001 | Autauga County | Alabama | 2019-11-01 | 2.9932332 | 1243 |
| 1001_2019-12-01 | 1001 | Autauga County | Alabama | 2019-12-01 | 2.9932332 | 1243 |

## Census Data

| pct_bb_20... | pct_bb_20... | pct_bb_20... | pct_bb_20... | pct_bb_20... | cfips | pct_colleg... | pct_colleg... | pct_colleg... | pct_colleg... |
|---|---|---|---|---|---|---|---|---|---|
| 76.6 | 78.9 | 80.6 | 82.7 | 85.5 | 1001 | 14.5 | 15.9 | 16.1 | 16.7 |
| 74.5 | 78.1 | 81.8 | 85.1 | 87.9 | 1003 | 20.4 | 20.7 | 21.0 | 20.2 |
| 57.2 | 60.4 | 60.5 | 64.6 | 64.6 | 1005 | 7.6 | 7.8 | 7.6 | 7.3 |
| 62.0 | 66.1 | 69.2 | 76.1 | 74.6 | 1007 | 8.1 | 7.6 | 6.5 | 7.4 |
| 65.8 | 68.5 | 73.0 | 79.6 | 81.0 | 1009 | 8.7 | 8.1 | 8.6 | 8.9 |

# Language & IDE


# Libraries

# Missing Data

- There are only **4** rows with missing values in census dataset:
  1. Row 93 has missing values in 10 columns
  2. Row 1817 has missing values in 2 columns.
  3. Row 2645 has missing values in 1 column.
  4. Row 2674 has missing values in 1 column.

- We used the "MICE" package with "*PMM*" method to impute missing values.

- "MICE" creates multiple replacement values for multivariate missing data.
- "PMM" method is a flexible and widely used imputation method
- This method works well with continuous variables.

# Timeframes

1. The ***train*** dataframe includes the data from **August 2019** to **October 2022.**

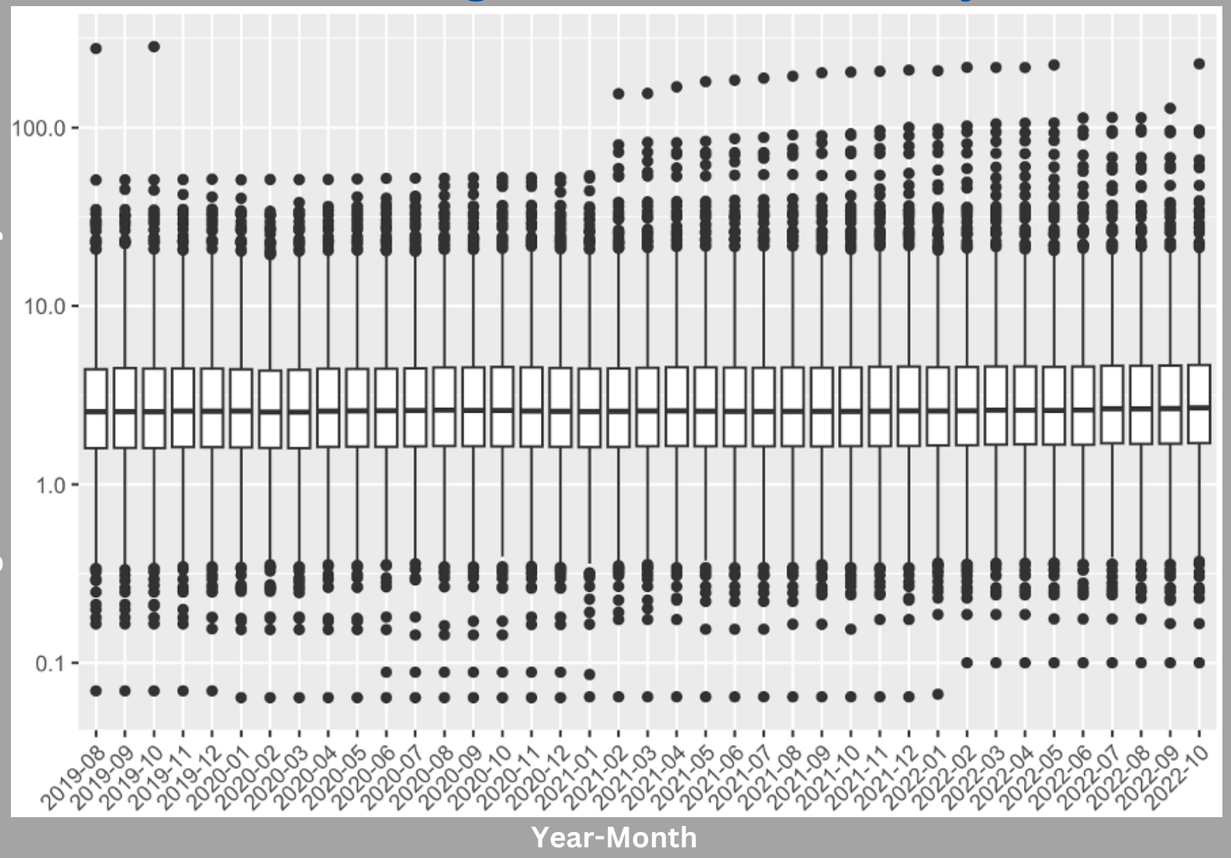
2. The target ***forecast*** timeframe is from **November 2022** to **June 2023**.

3. The ***cencus*** dataframe includes data from **2017** to **2021**.

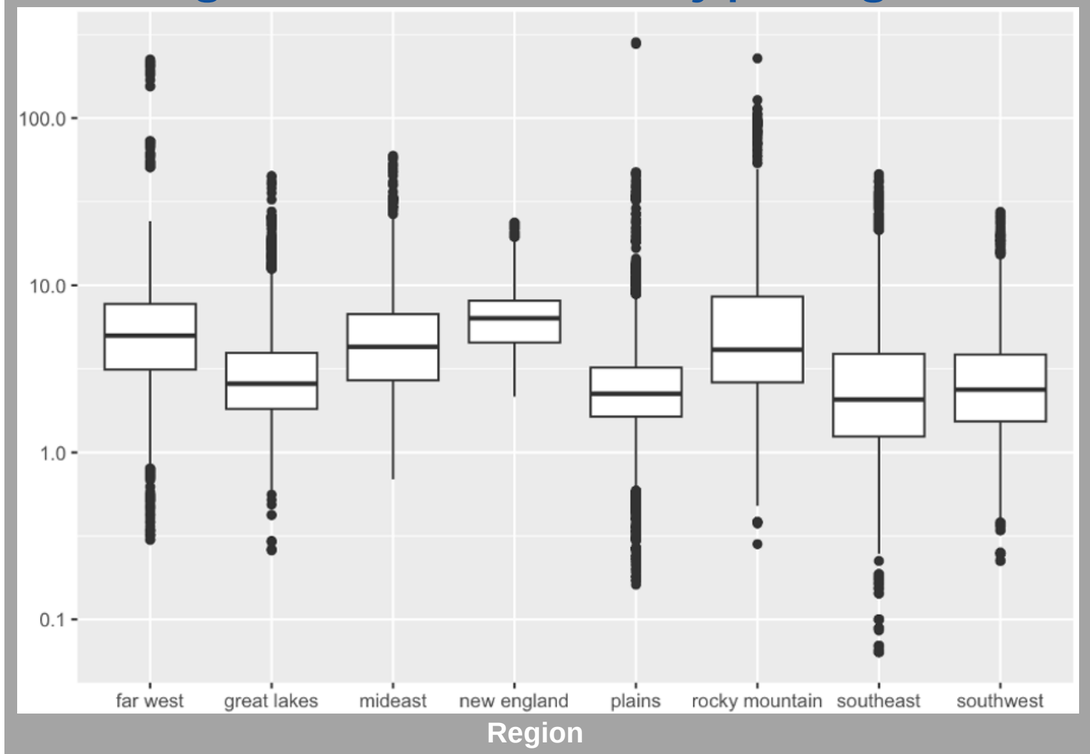
4. **Census** data have a 2-year lag to match the **train** data. We considered this gap while merging the data.

# Data Visualization

## Overall Avg. Microbusiness Density

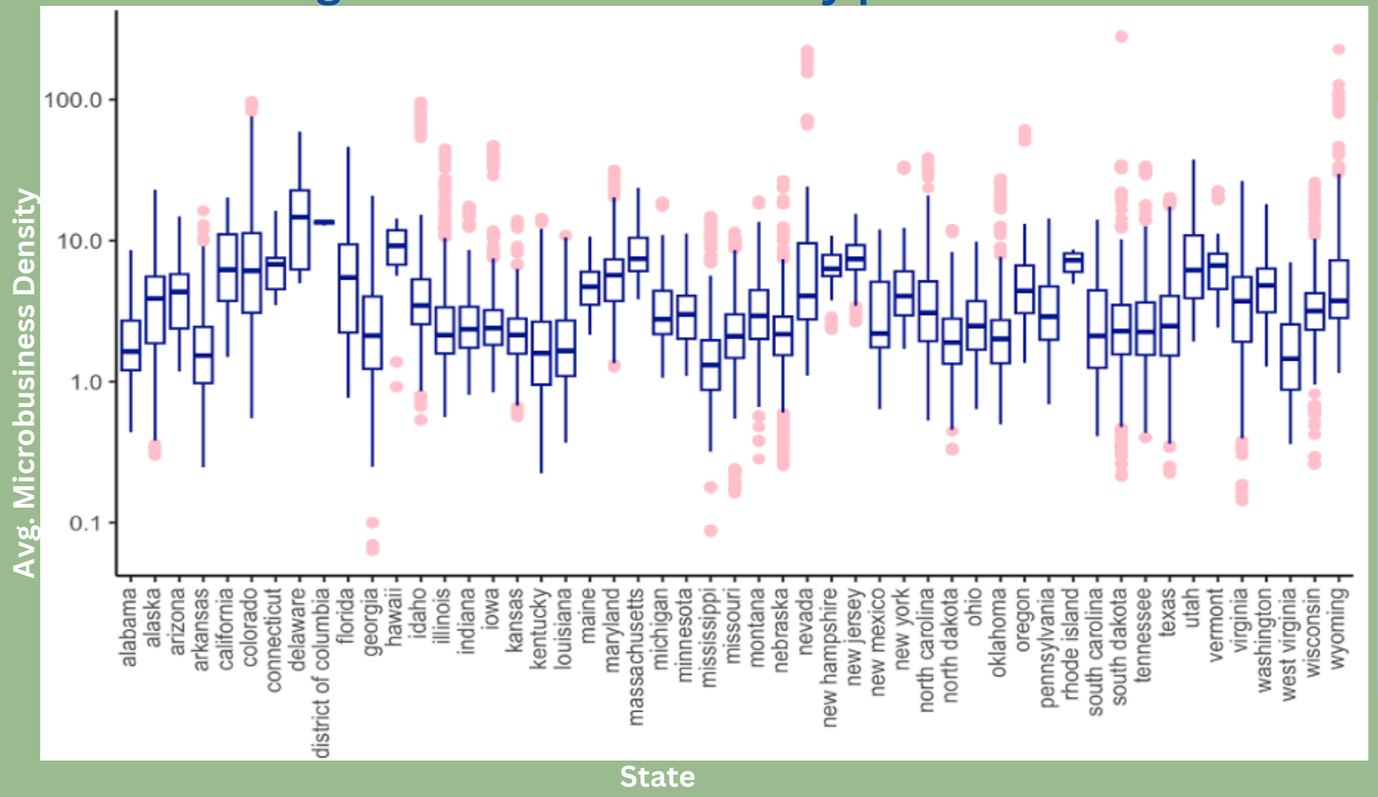



## Avg. Microbusiness Density per Region

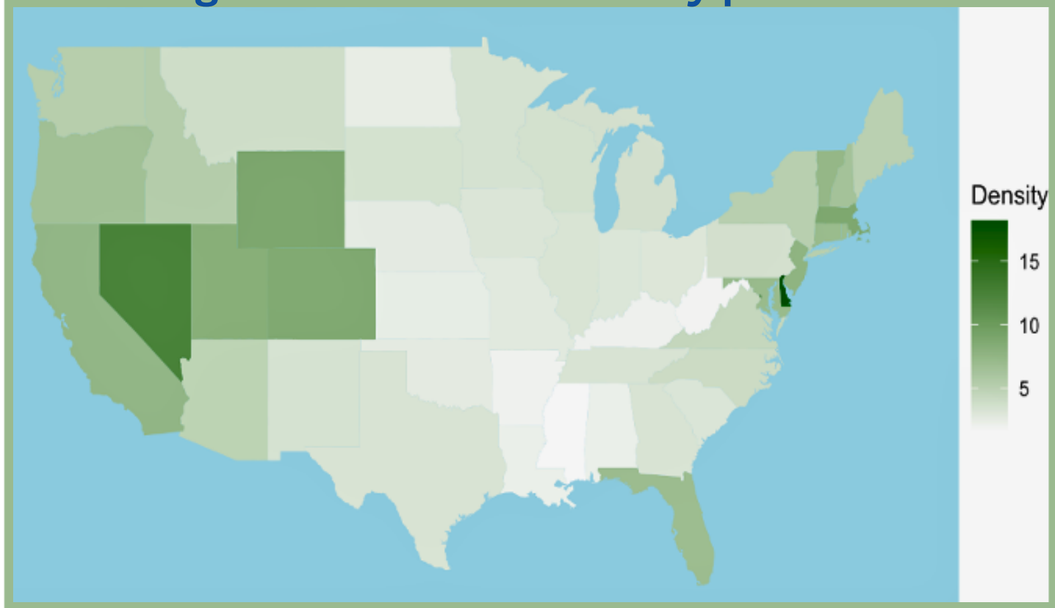


# Data Visualization

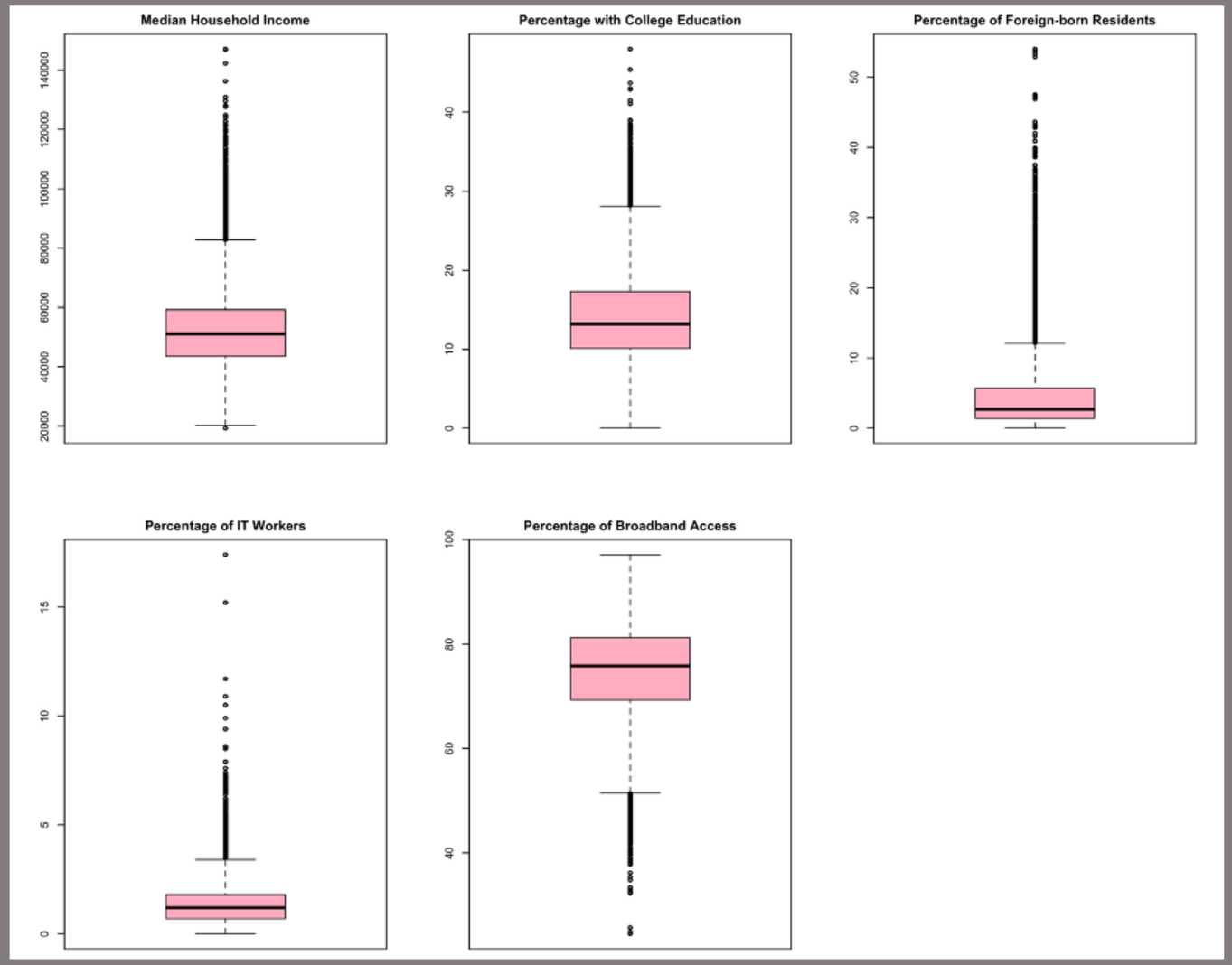
## Avg. Microbusiness Density per State




## Avg. Microbusiness Density per State

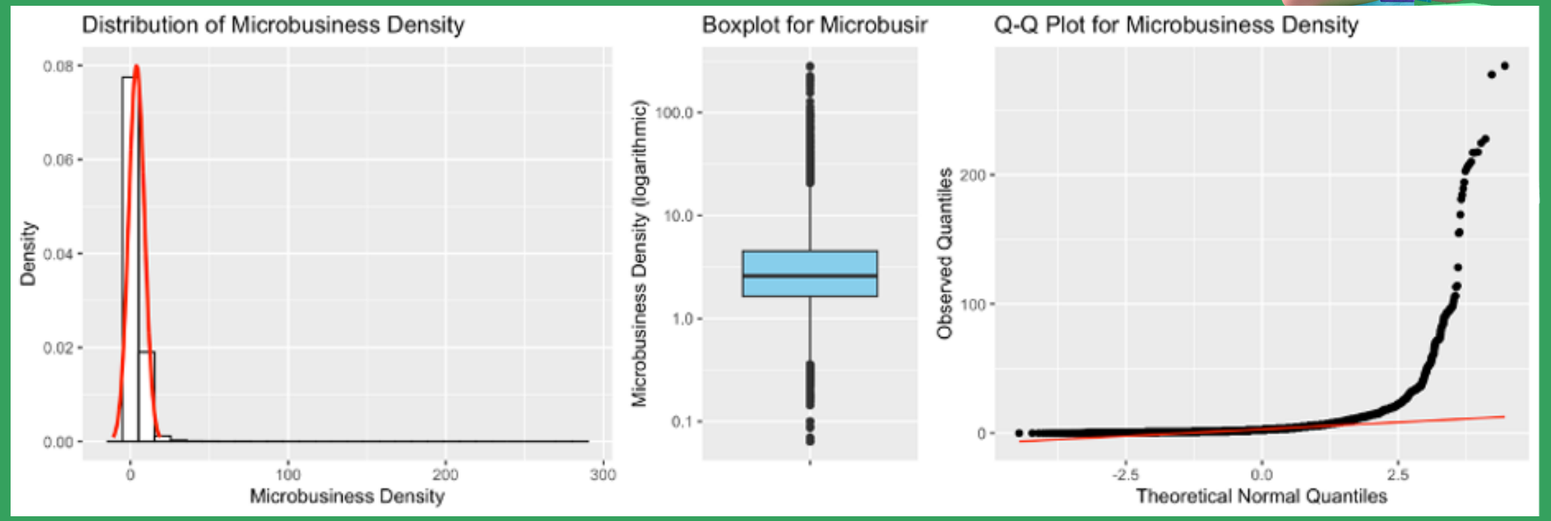


# Data Visualization

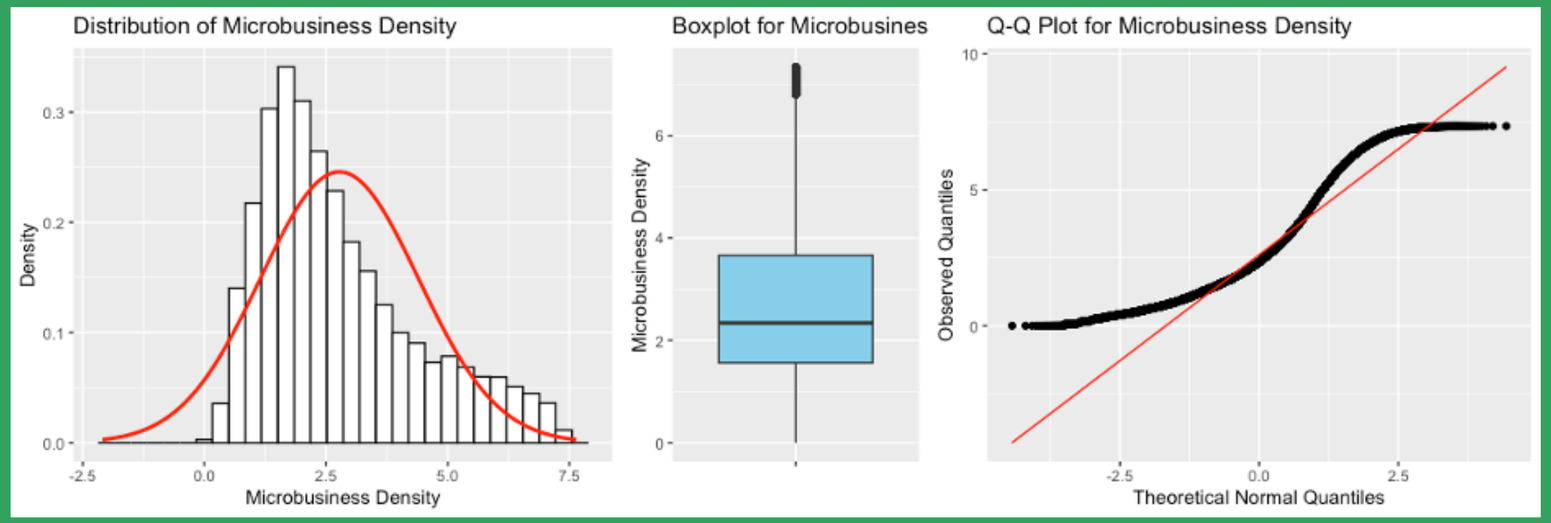



# Microbusiness Density Distribution

**Before Detecting and Removing Outliers**




**After Detecting and Removing Outliers**

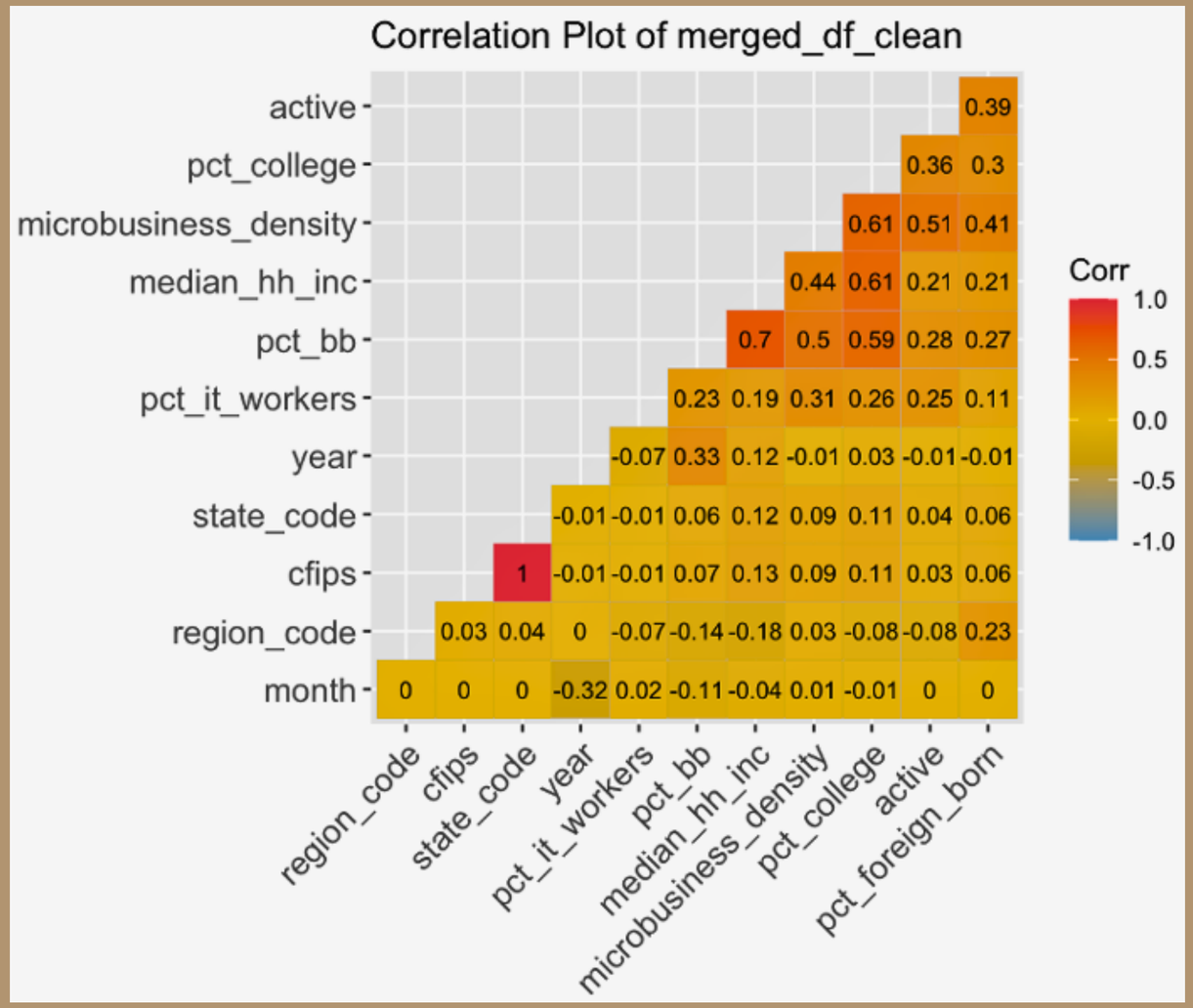


# Outlier Detection And Removal

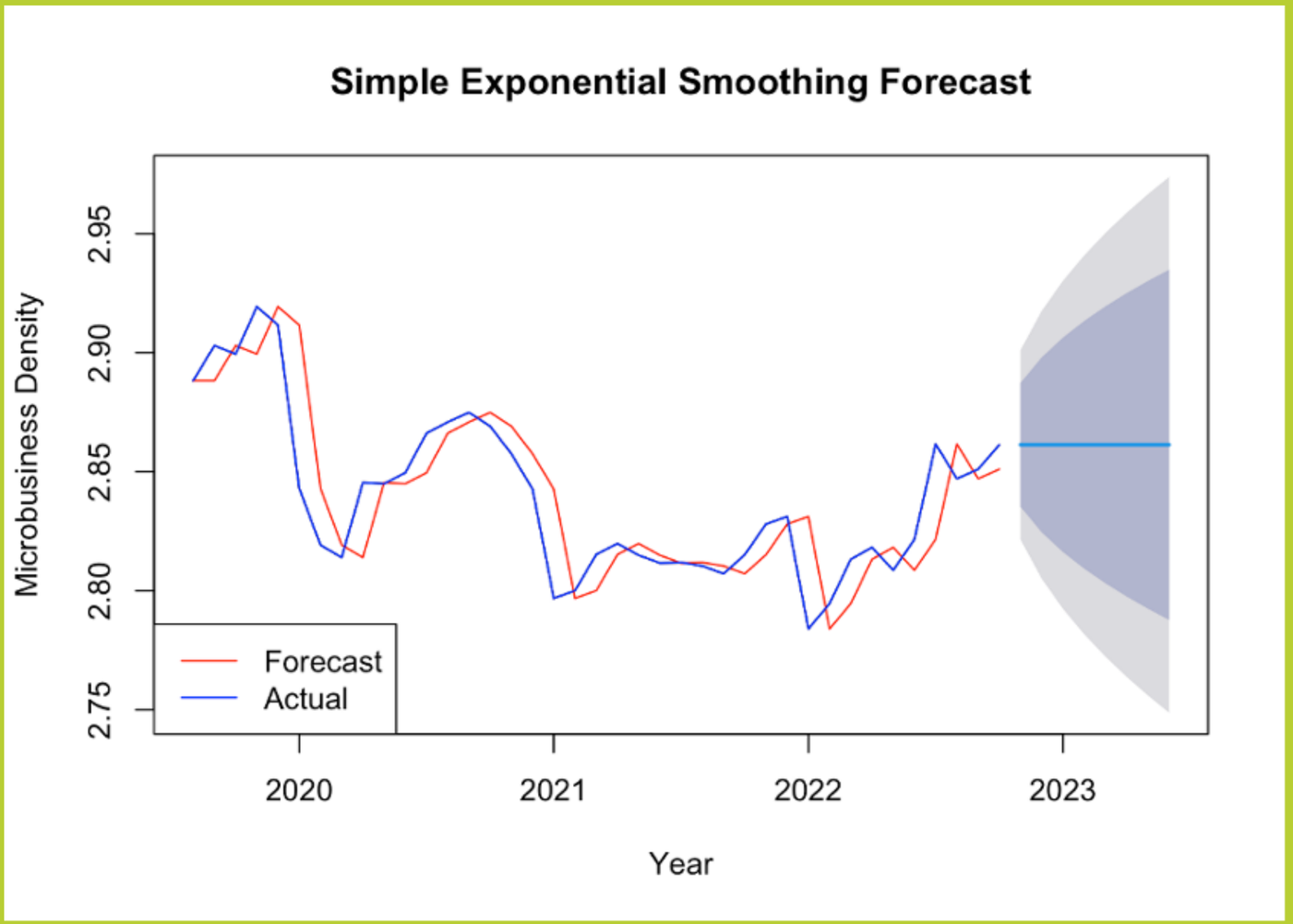
| | |
|---|---|
| Total Rows Removed | 25,167 |
| % of Rows Removed | 20.58 % |
| Total Records Count | 122,265 |

| Feature | Outliers Count | % |
|---|---|---|
| microbusiness_density | 12,692 | 10.38 % |
| pct_bb | 2,507 | 2.20 % |
| pct_college | 1,428 | 1.25 % |
| pct_it_workers | 2,556 | 2.25 % |
| pct_foreign_born | 8,329 | 7.33 % |
| median_hh_inc | 3,410 | 3 % |

# Correlation Plot



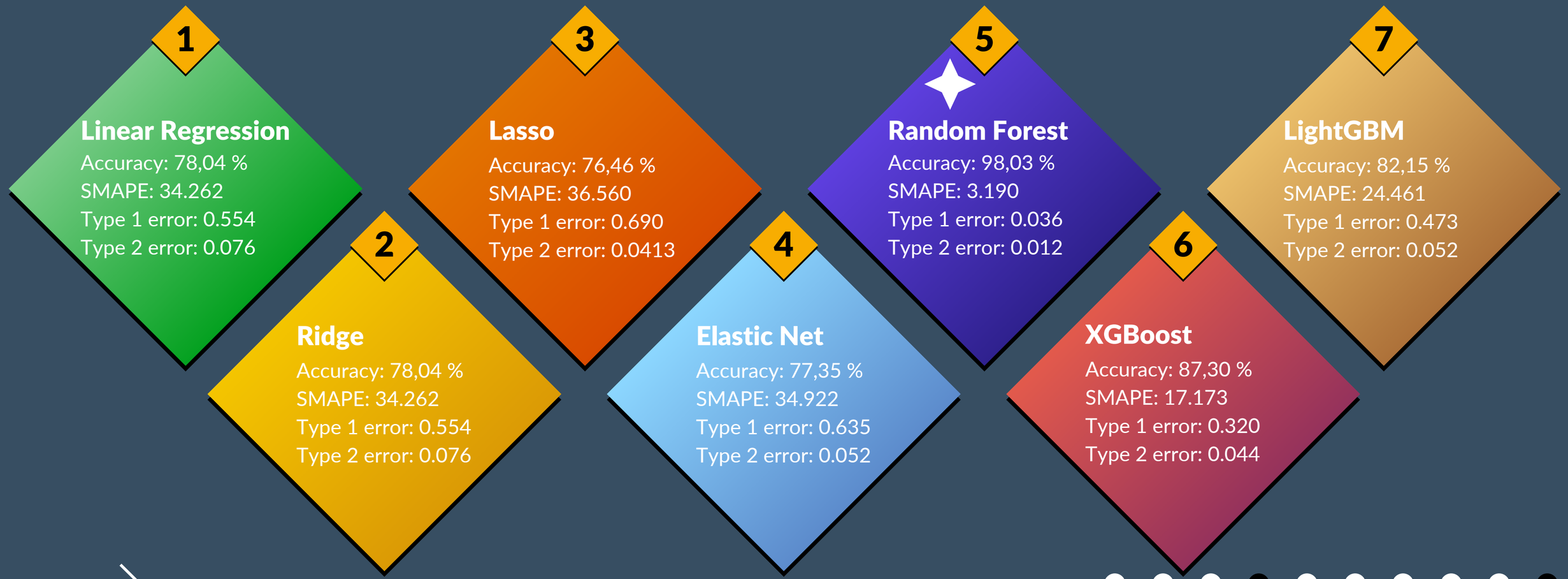
# Simple Exponential Smoothing Forecasting




# Model Training

1. Linear Models
    1. Linear Regression
    2. Ridge
    3. Lasso
    4. ElasticNet
2. Ensemble Models
    1. Random Forest
    2. Extreme Gradient Boosting (XGB)
    3. Light Gradient-Boosting Machine (LGBM)

# Model Training



# Conclusion

- Random Forest is the most reliable model to predict microbusiness density.
- The linear models achieved reasonable results but with higher errors.
- XGB and LGBM also performed well but were outperformed by Random Forest.

- When considering the implications of Type 1 and Type 2 errors, it is essential to strike a balance.
- Type 1 errors (false negatives) result in missed opportunities to identify areas with high microbusiness potential.
- Type 2 errors (false positives) lead to allocating resources to areas with low potential.

# Thank You For Your Attention!