# REAL ESTATE MARKET ANALYSIS

- **Dataset source:** Divar website
- **Group members:**

Mahdi Rafati
Ramin Badri

🔗 Project Repository

# Project Overview

## PROJECT PIPELINE

- Exploratory Data Analysis (EDA)
- Descriptive Statistics
- Statistical Hypothesis Testing
- Clustering (K-Means, DBSCAN)
- Predictive Modeling (Price & Total Credit)

# Dataset Overview

## DIVAR REAL ESTATE ADS

This data is extracted from **Divar**, one of the largest online classified marketplace applications in Iran, containing **1M** records with **60** features.

Key features groups:

- Price variables
- Location (lat/lon)
- Physical attributes (area, rooms, age)
- Amenities

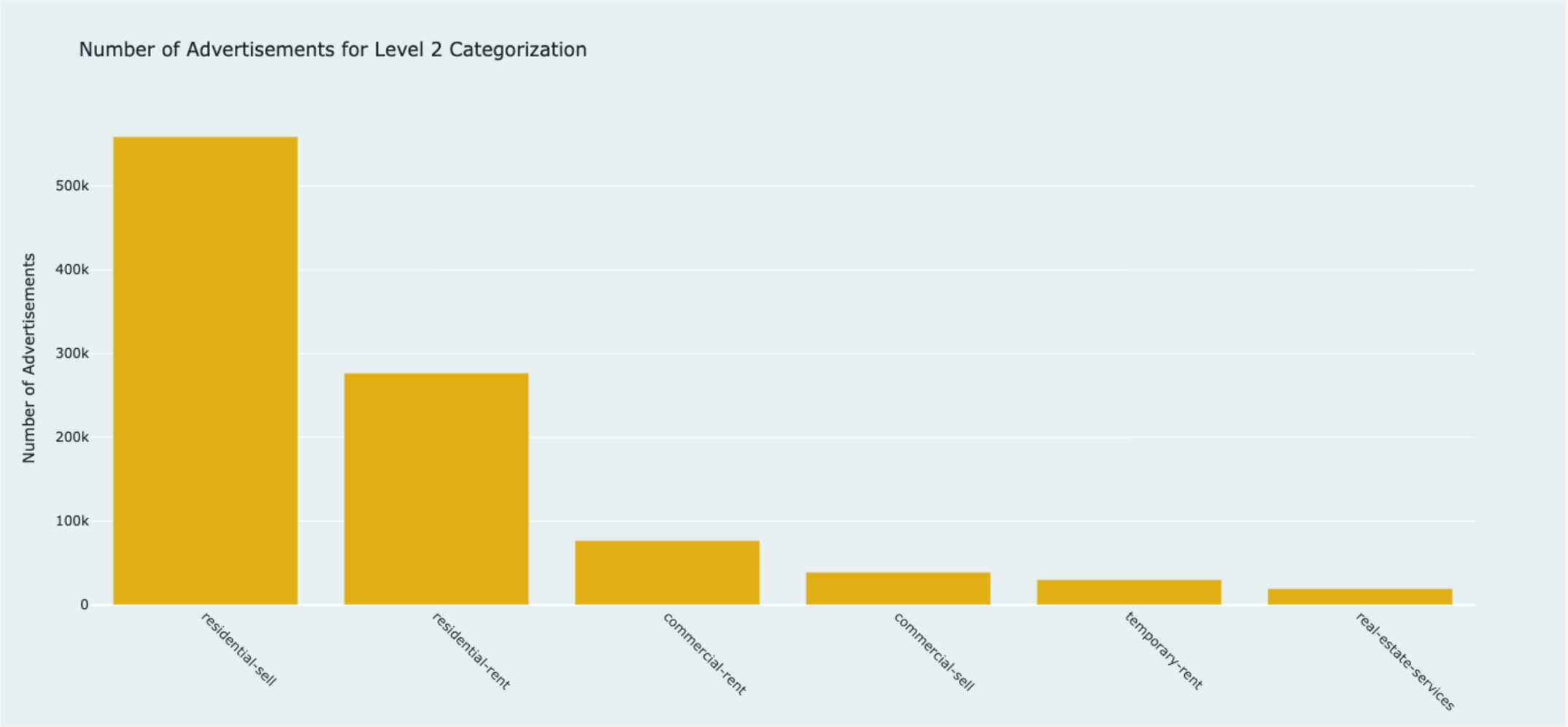
Highlight:

- This is advertisement data, not actual sales

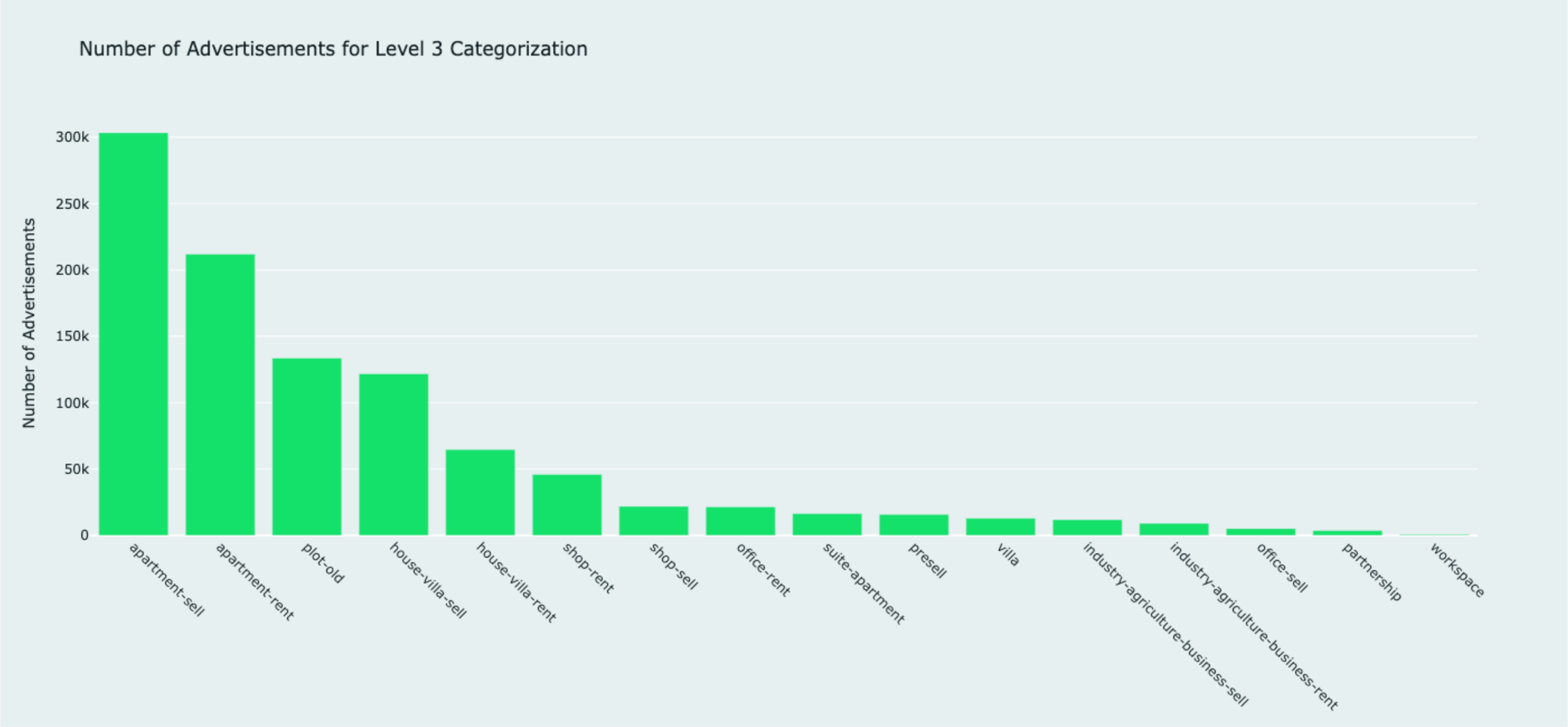
# DESCRIPTIVE STATISTICS

Answering some questions to grasp a good understanding & insights from the data

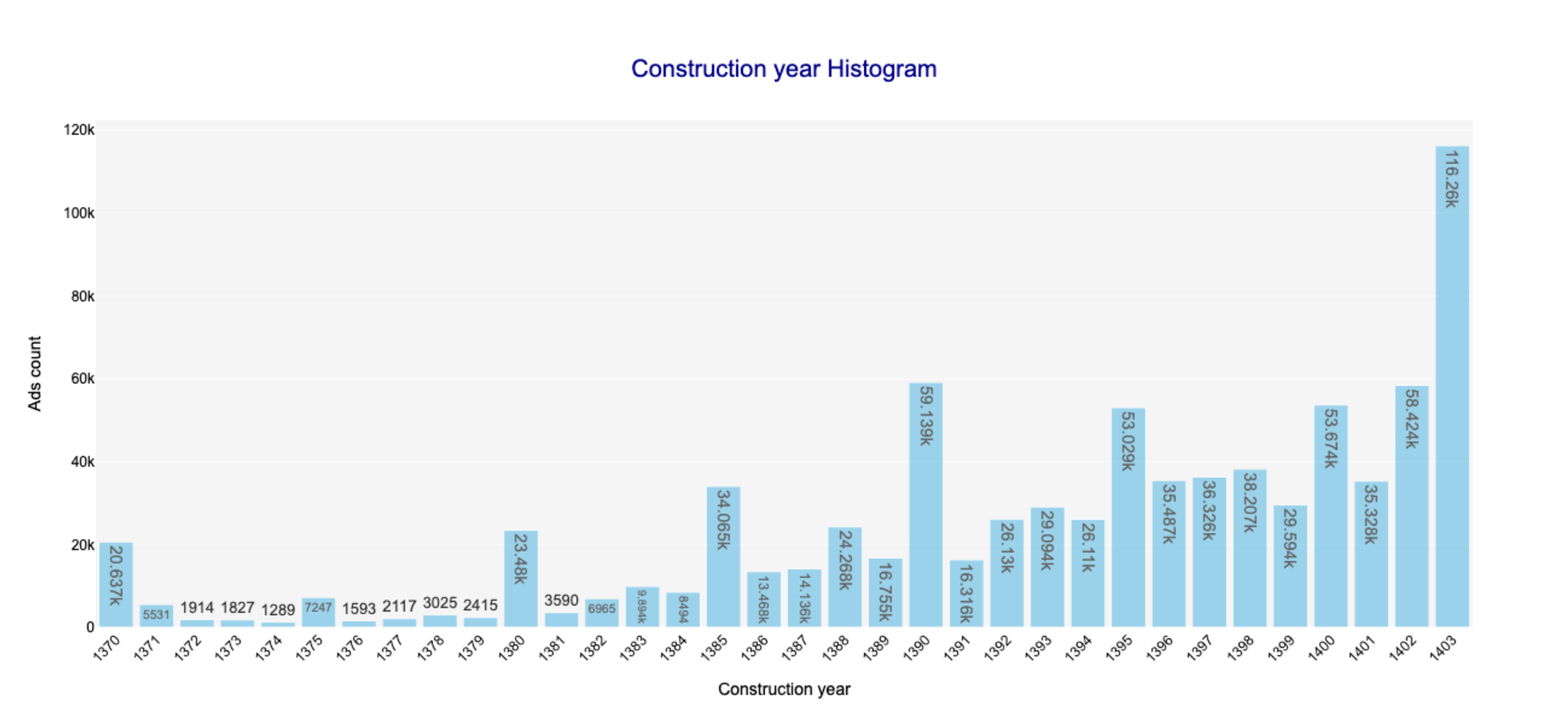
# Distribution of Ads in Different Categories for level 2 & level 3 Categorization



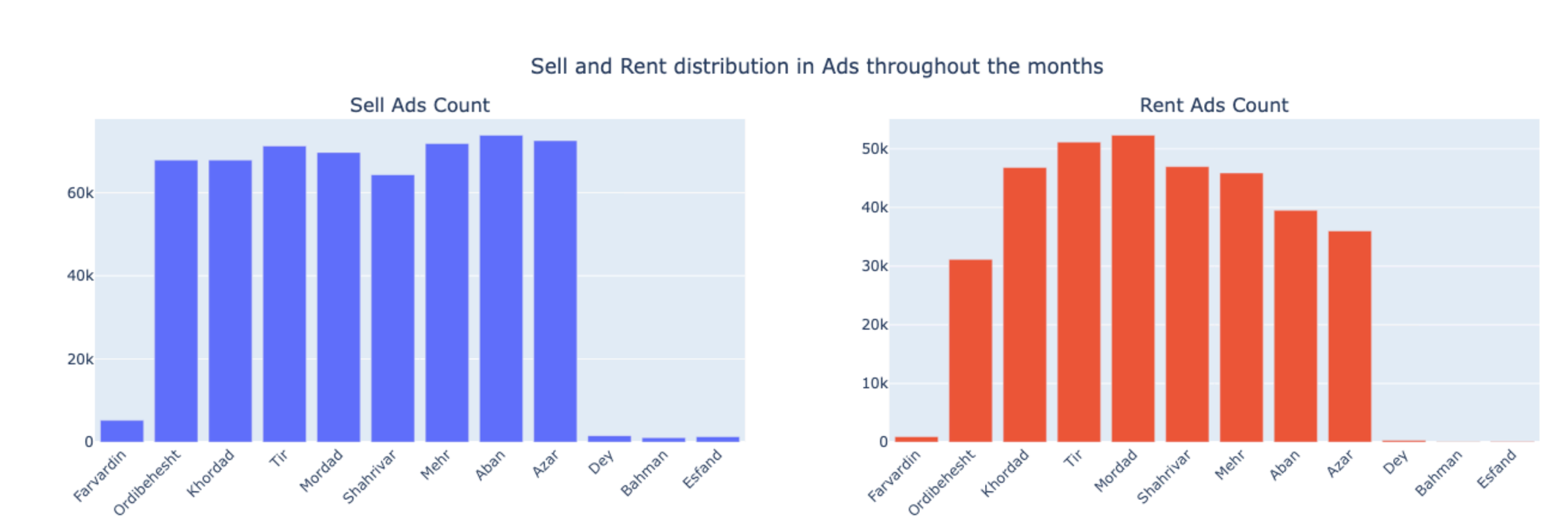
# Distribution of Ads in Different Categories for level 2 & level 3 Categorization



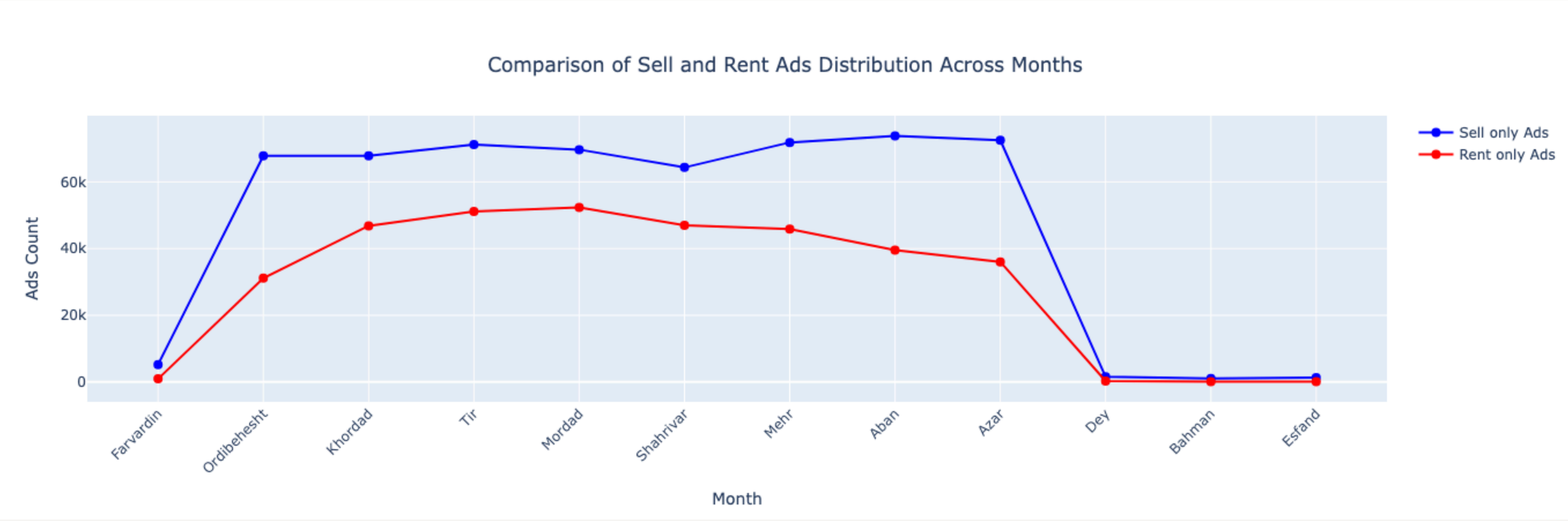
# Histogram Chart of Construction Year



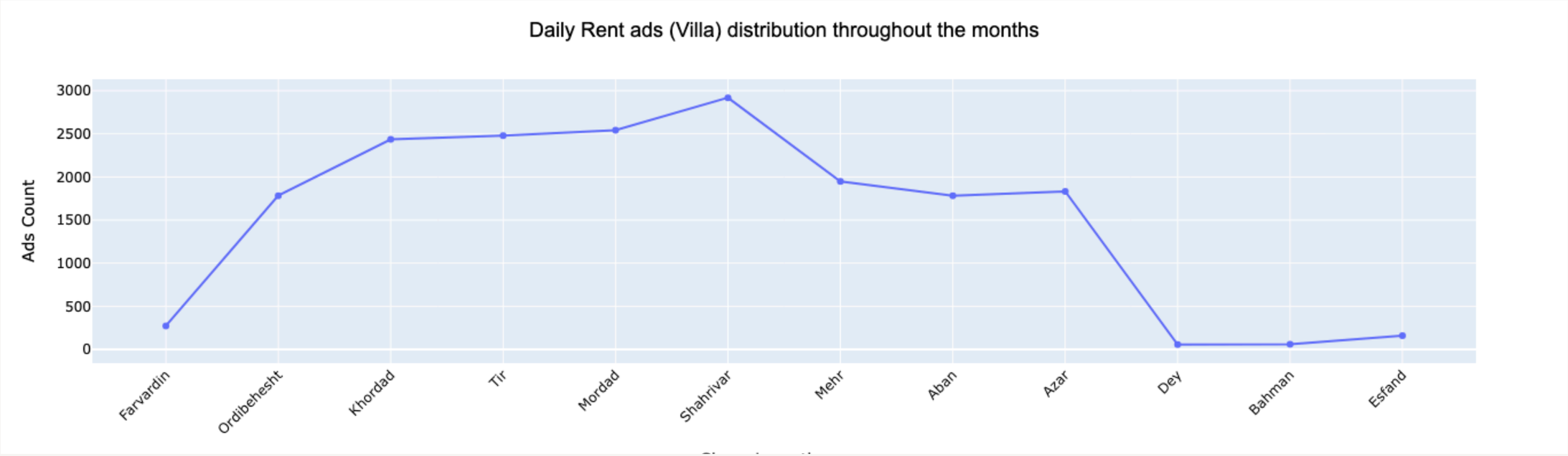
# The Number of Ads Published in Different Months for Sale and Rent



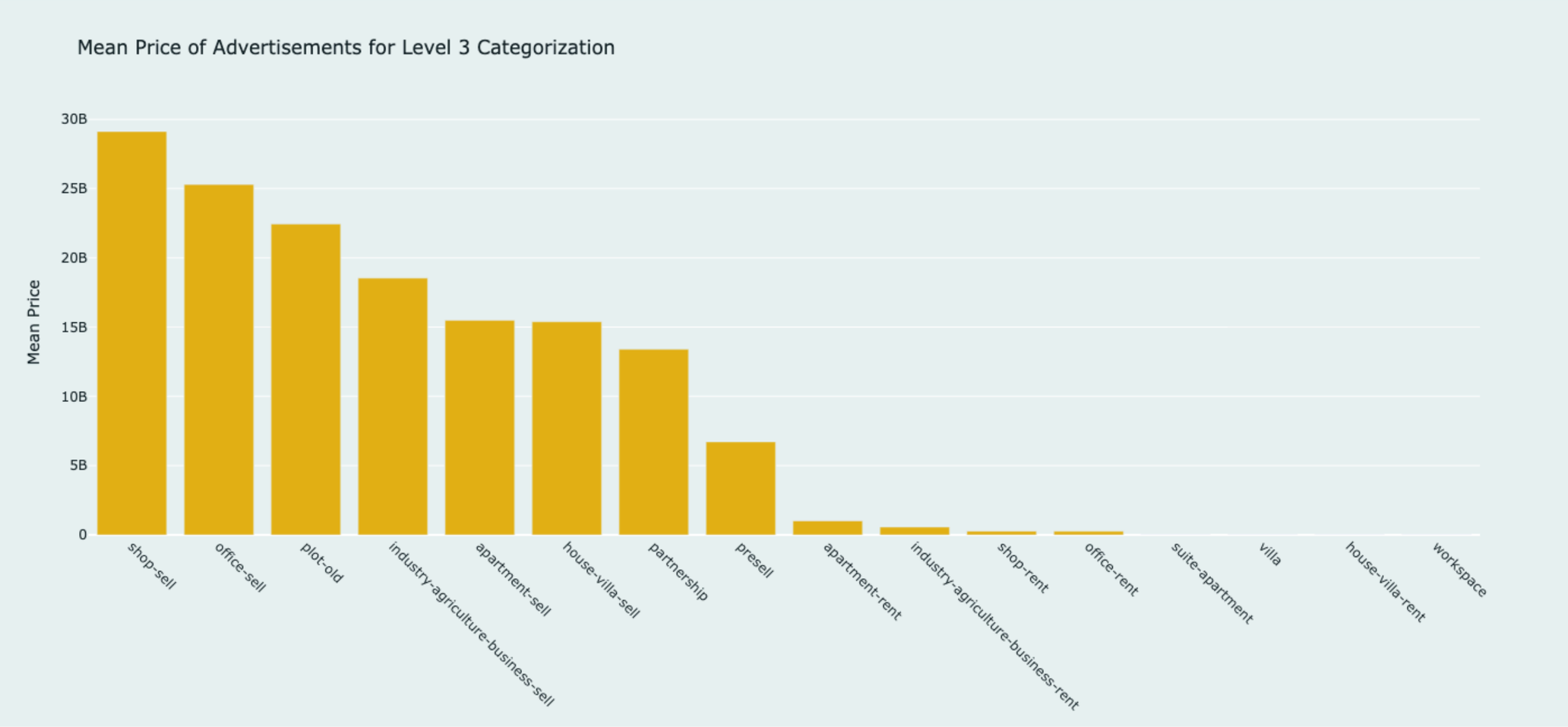
# The Number of Ads Published in Different Months for Sale and Rent



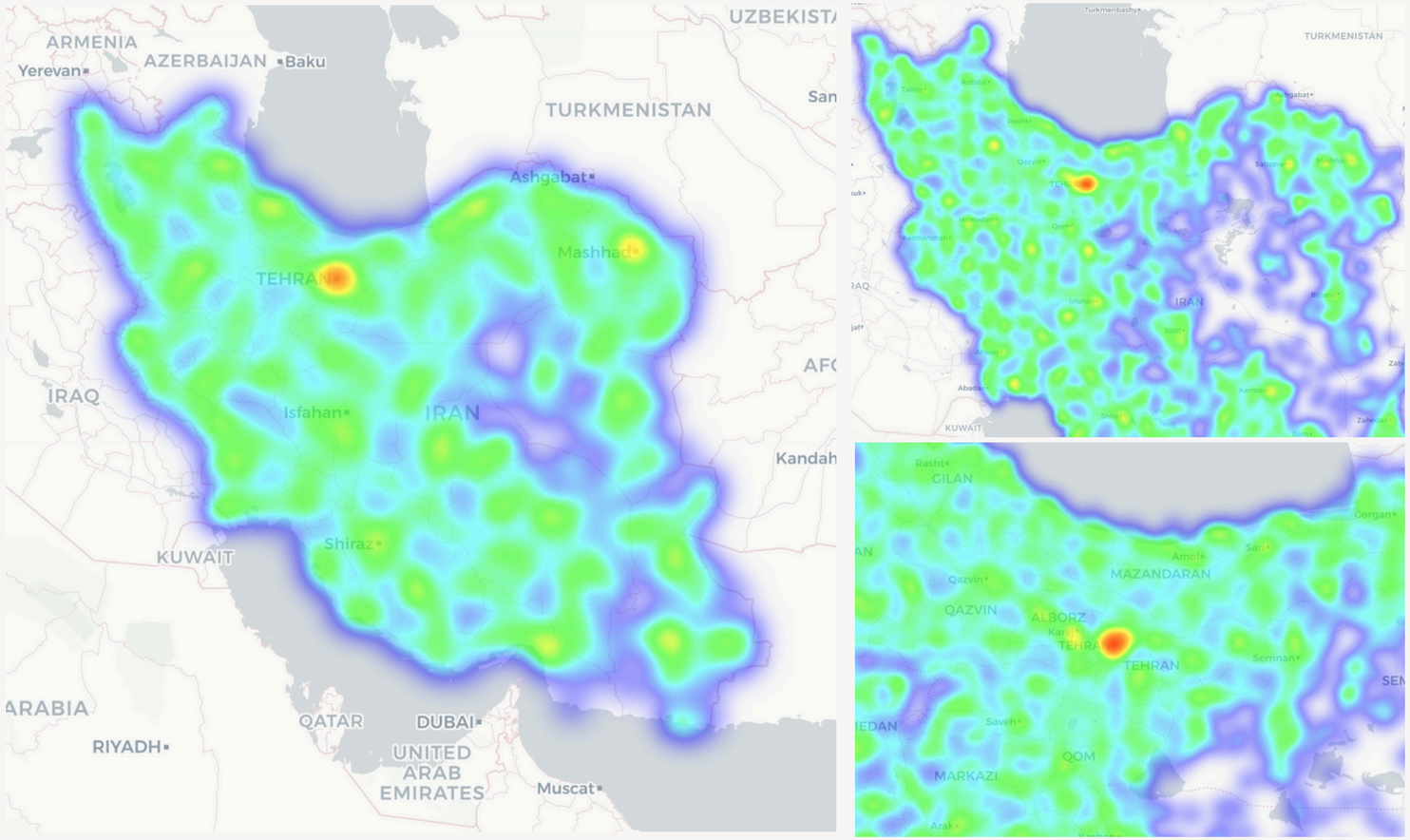
# The Number of Ads Published in Different Months for Sale and Rent



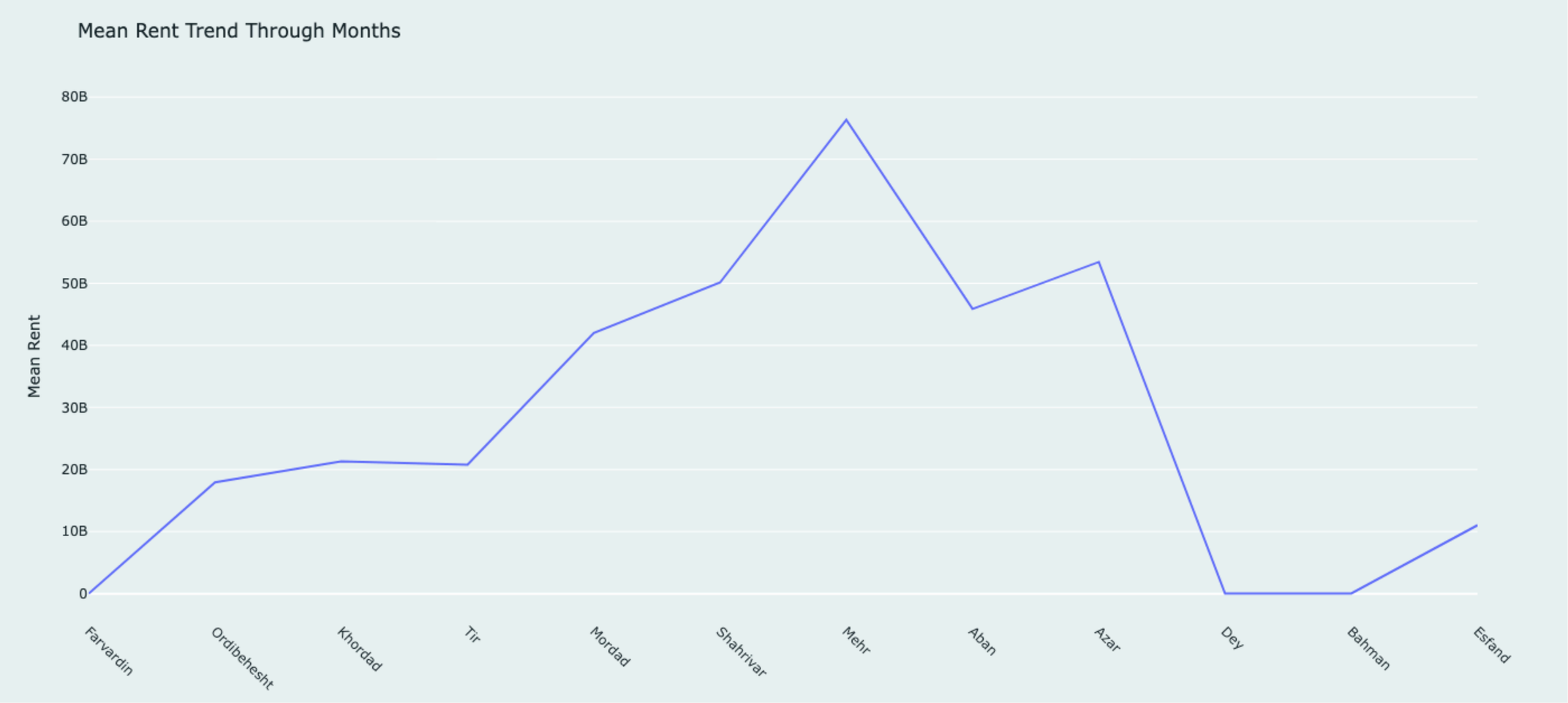
# The Distribution of Sales Price (price_value) for Level 3 Categories in a Graph



# The ads in each region on a geographic heatmap. Which region has the highest density of ads? **Tehran**

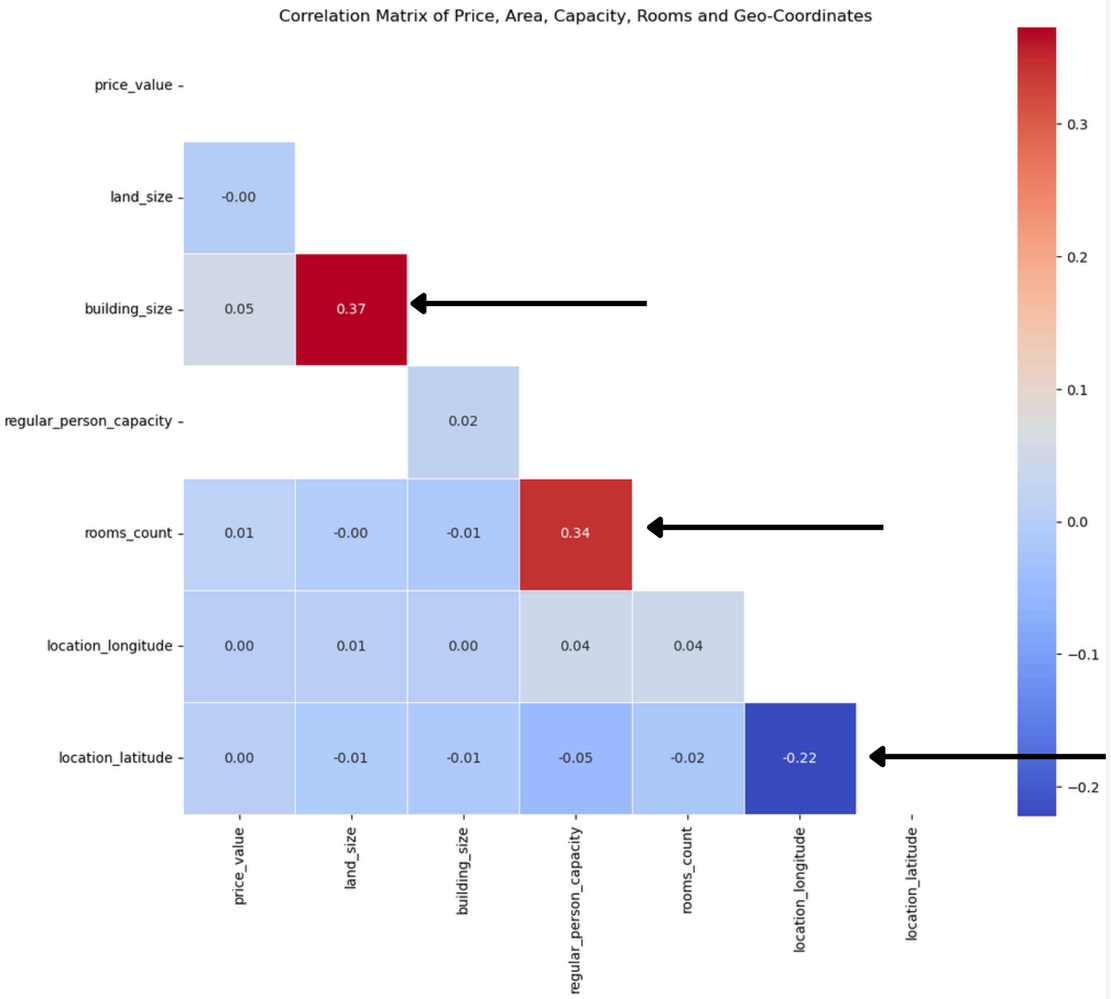


# The average rental price trend by months of ads.

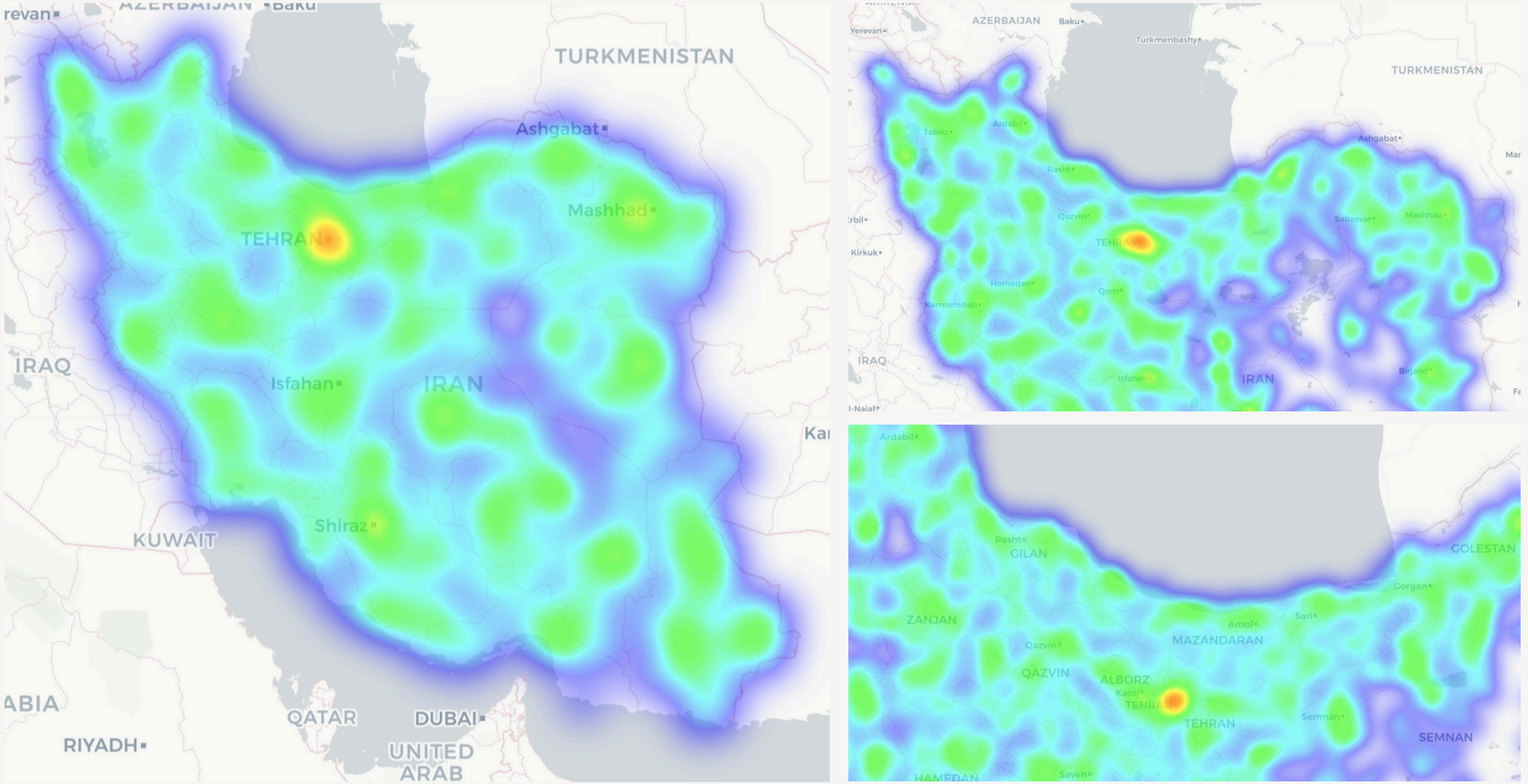


# Calculate the real price for the average price in the years 1400 to 1403 and compare the real price trend with the nominal price.

# Correlation matrix for price amount, land area, infrastructure, capacity, number of rooms, and latitude and longitude.



# In which areas are houses with balconies, elevators, security guards, barbecues, and swimming pools mainly located? **Tehran**



GROUP-3

DEC 2025

# STATISTICAL HYPOTHESIS TESTING

Answering some questions to grasp a good understanding & insights from the data

FIRST PROJECT REPORT

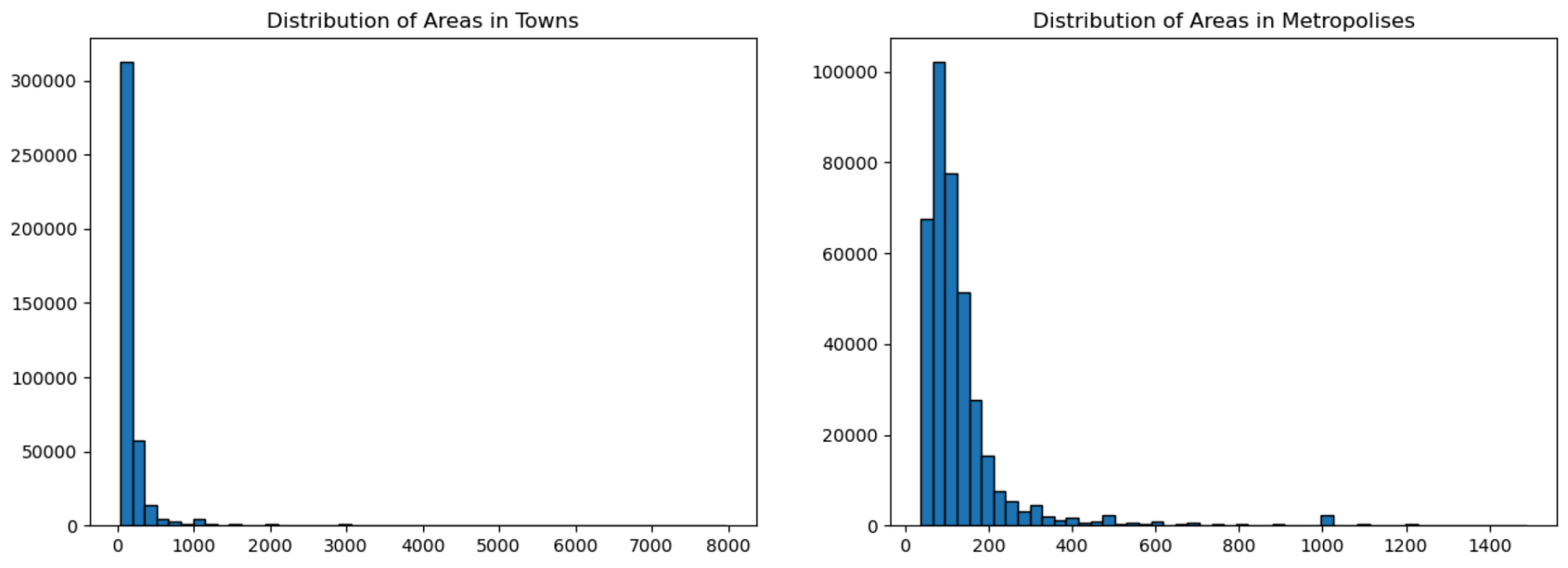
GITHUB.COM/MEHDIRT/DIVAR-PROJECT

# Is the average residential property area in metropolitan cities smaller than that in small towns and villages?

## Null Hypothesis ($H_0$):

The average residential area in metropolitan cities is greater than or equal to that in small towns and villages. (**one-tailed** Hypothesis)

## Data Sampling & Normality Check:



# 1st Approach: Use non-parametric methods

Since both distributions are extremely right-skewed, we can use **Mann–Whitney U test** to test our hypothesis.

**New $H_0$:** These two distributions are not similar.
🔁 **Result:** p-value < 0.05
✅ **$H_0$ is Rejected** → The distributions are similar, so the main hypothesis might be true.

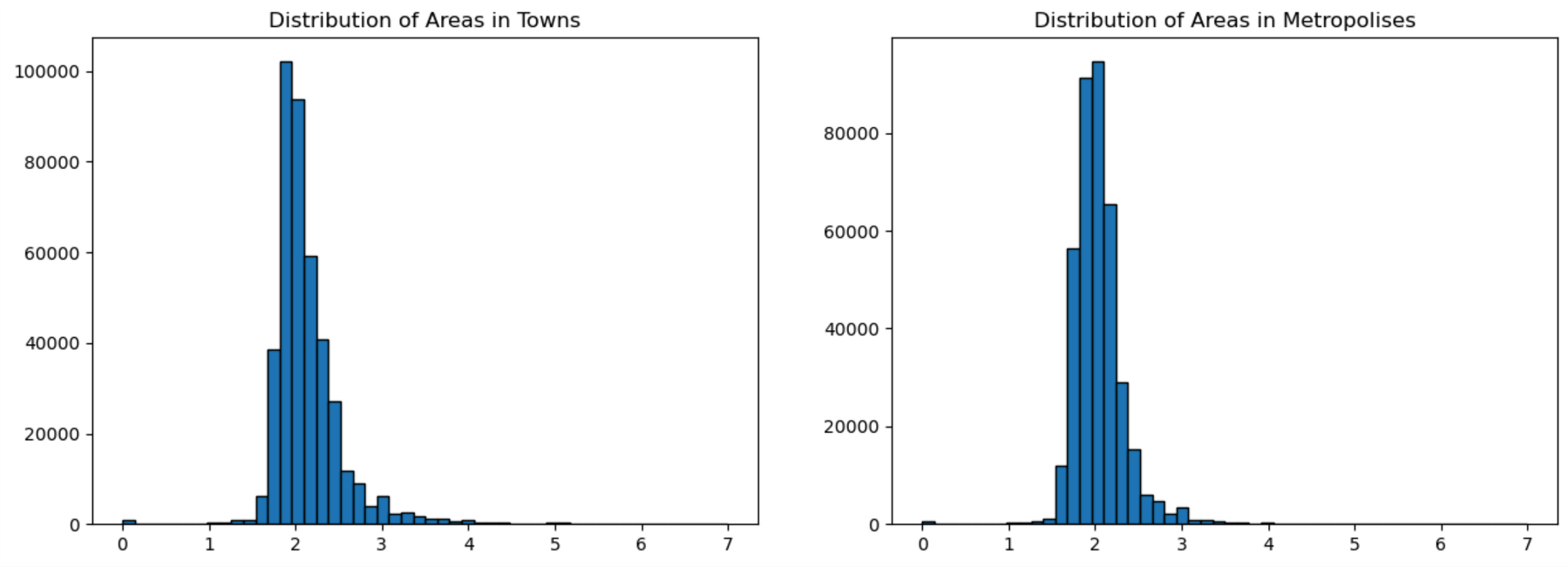
# 2nd Approach: Use parametric methods

We normalized our distributions using a log transformation, then we used a t-test to check our hypothesis.

🔁 **Result:** p-value < 0.05
✅ **$H_0$ is Rejected** → Residential Properties in Metropolises Do Have Smaller Average Area Than in Towns

# 2nd Approach: Use parametric methods

## Normalized Distributions After Log Transformation:

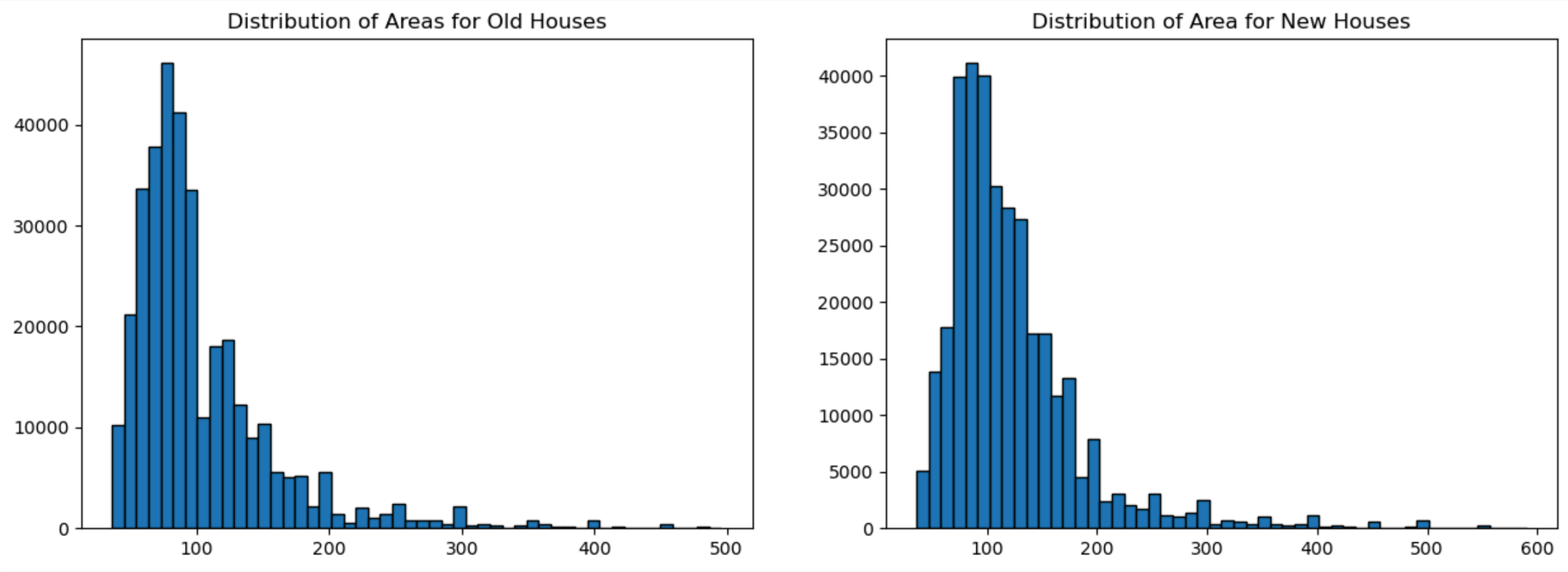


# Is the average area of older homes larger than newly built homes?

## 🔁 Null Hypothesis ($H_0$):

Older residential properties are smaller on average. (**one-tailed** Hypothesis)

## Data Sampling & Normality Check:




⚠️ **Note:** Since the distributions have almost 1M samples, we can't use the Shapiro Method to check normality.

# 1st Approach: Use non-parametric methods

Since both distributions are right-skewed, we can use **the Mann–Whitney U test** to check whether the distributions are similar.

**New $H_0$:** These two distributions are not similar.

🔄 **Result:** p-value > 0.05

❌ **Failed to Reject $H_0$ →** The distributions are not similar, so the main hypothesis might be false.

# 2nd Approach: Use parametric methods

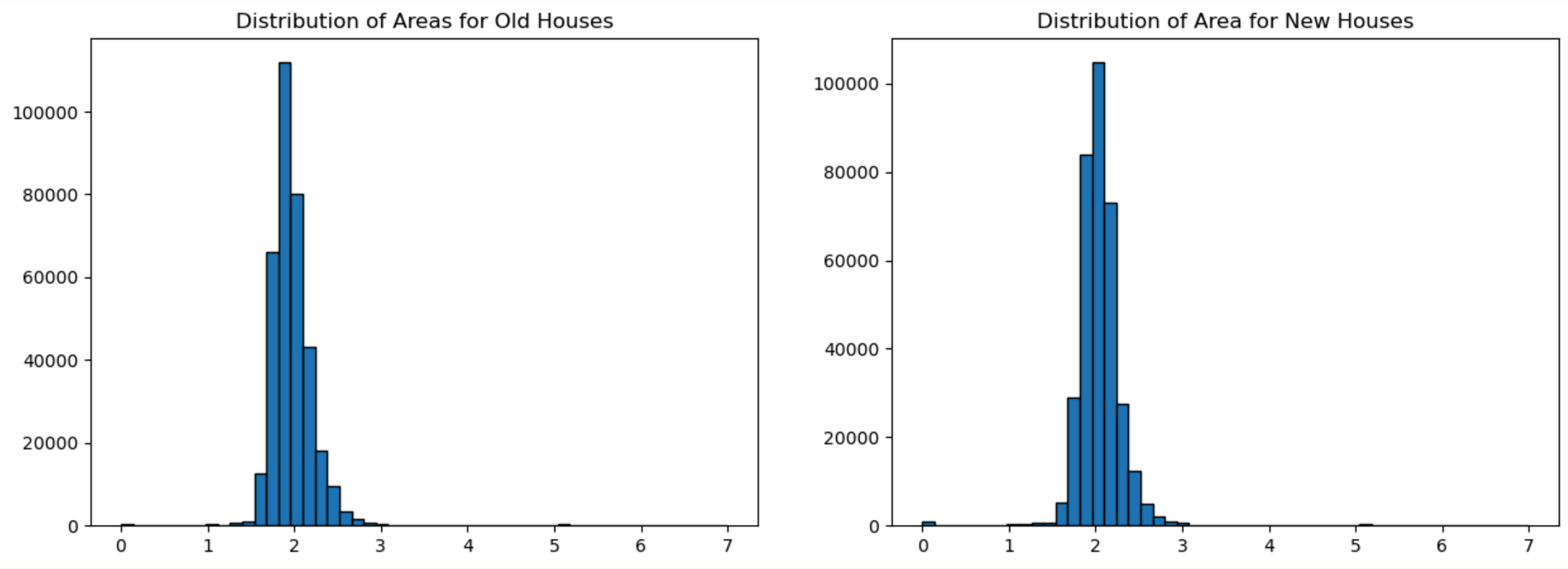
We normalized our distributions using a log transformation, then we used a t-test to check our hypothesis.

🔄 **Result:** p-value > 0.05

❌ **Failed to Reject $H_0$ →** No Sufficient evidence to Prove That Older Houses Were Bigger

# 2nd Approach: Use parametric methods

## Normalized Distributions After Log Transformation:



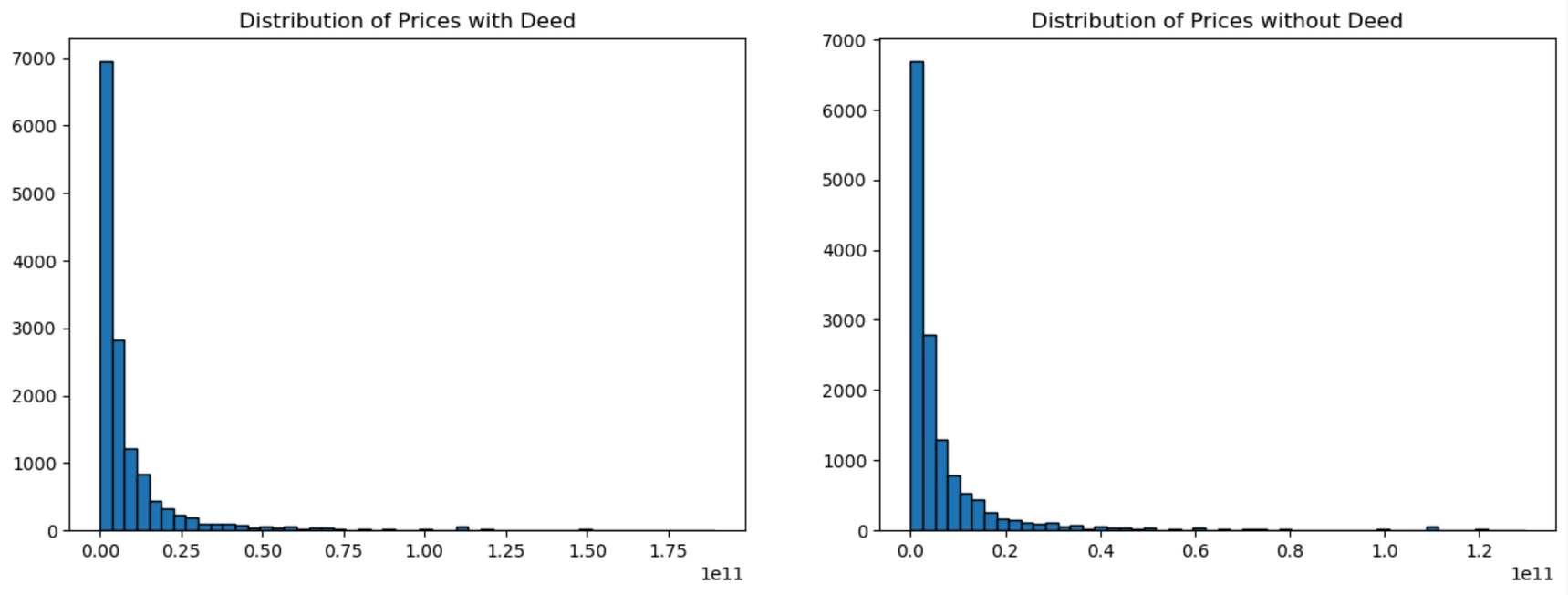
# Does having a business deed have a significant impact on the average sales price of commercial property?

## Null Hypothesis ($H_0$):

Having a business deed doesn't have a significant impact on the average sales price of commercial property.

(**two-tailed** Hypothesis)

## Data Sampling & Normality Check:



# Using non-parametric methods

Since both distributions are extremely right-skewed, we can use **Mann–Whitney U test** to test our hypothesis.

**New $H_0$:** These two distributions are not similar.

🔄 **Result:** p-value < 0.05

✅ **$H_0$ is Rejected** → The distributions are similar, so the main hypothesis might be true.

# The average price increases significantly for luxury features. But does this average also differ significantly for non-luxury features?

We have two hypotheses to check:

🔁 **First null-hypothesis:**

**$H_0$:** Having a luxury feature does not affect the average price. (two-tailed)

**New $H_0$:** These two distributions are not similar.

🔁 **Result:** p-value < 0.05

✅ **$H_0$ is Rejected** → Having Luxury Features Probably Affects the Average Price.

🔁 **Second null-hypothesis:**

**$H_0$:** Having non-luxury features does not affect the average price. (two-tailed)

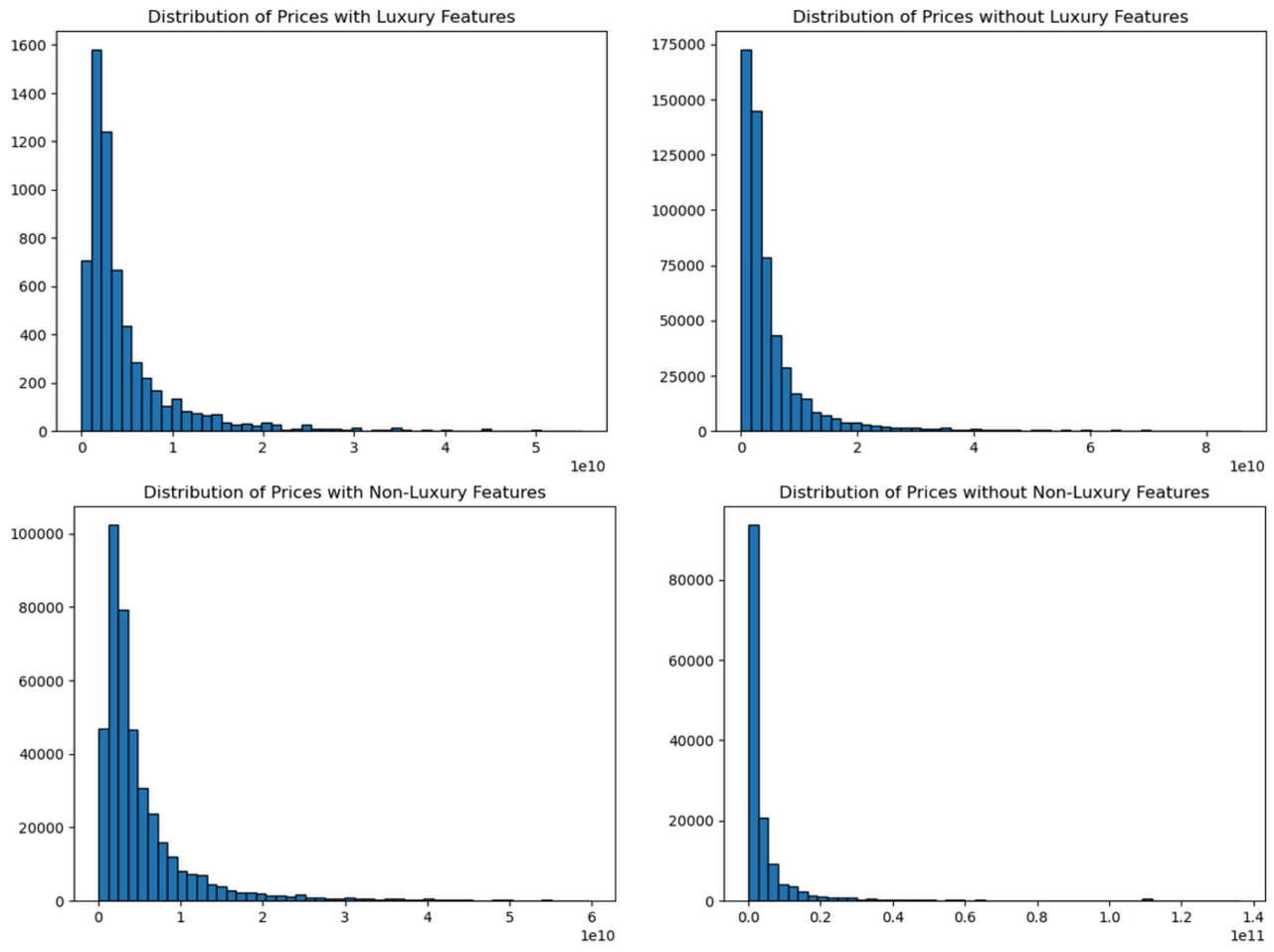
**New $H_0$:** These two distributions are not similar.

🔁 **Result:** p-value < 0.05

✅ **$H_0$ is Rejected** → Having Non-Luxury Features Probably Affects the Average Price.

# Approach for Both Hypotheseses

Both pairs of distributions are highly right-skewed, so we used the **Mann–Whitney U** test to test them. Here is the sample distribution of both problems:



GROUP-3

DEC 2025

# ML / RECOMMENDER SYSTEM

Uncovering hidden structures by clustering the data

FIRST PROJECT REPORT

GITHUB.COM/MEHDIRT/DIVAR-PROJECT

# Part 1 - K-Means Clustering

## FINDING BEST FEATURE SET

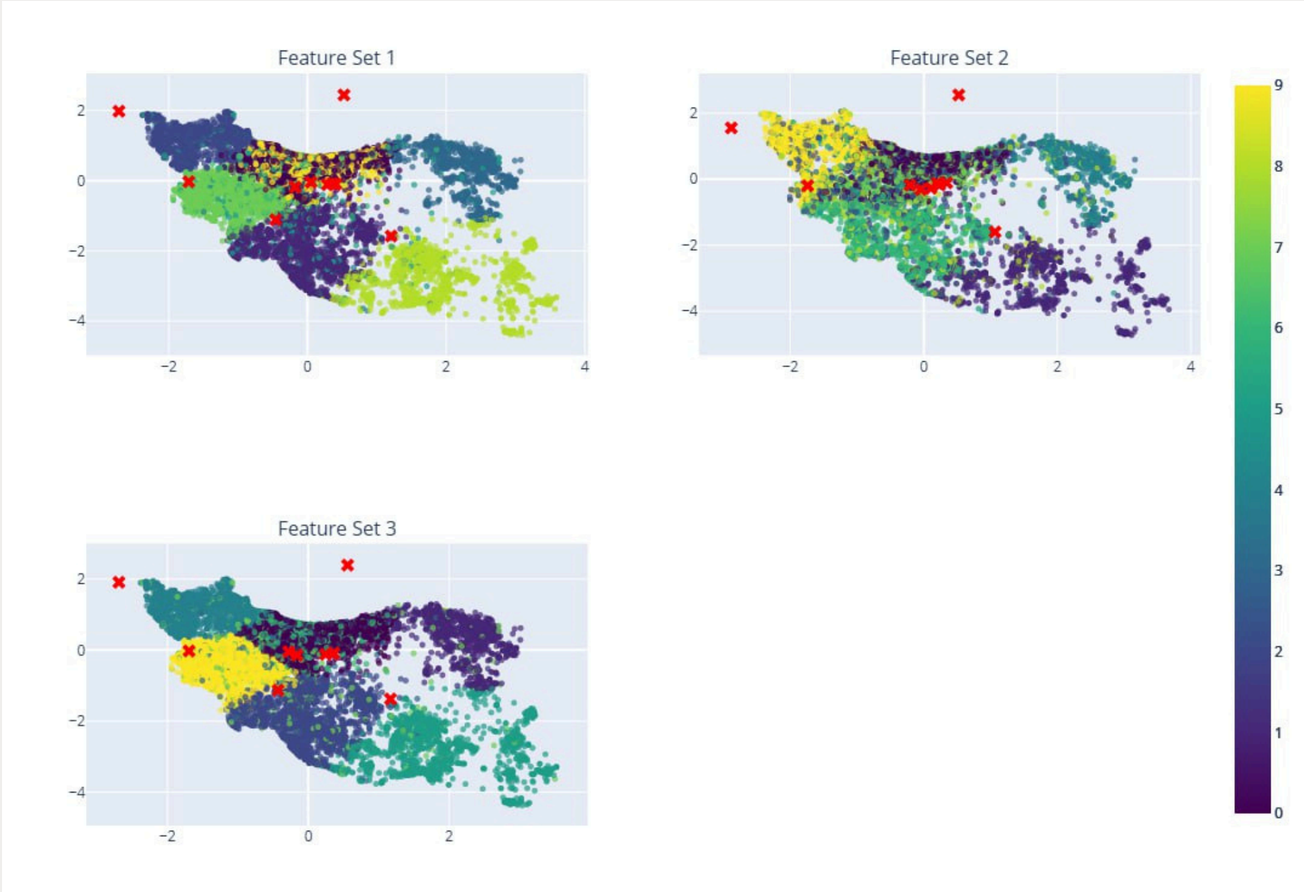
- **Feature set 1:** location_latitude, location_longitude, price_value
- **Feature set 2:** location_latitude, location_longitude, price_value, building_size, construction_year
- **Feature set 3:** location_latitude, location_longitude, price_value, rent_value, credit_value

| FEATURE SET | INERTIA | DAVIES-BOULDIN | CALINSKI-HARABASZ | SILHOUETTE |
|---|---|---|---|---|
| Feature Set 1 | **223523** | 0.878 | 165152 | 0.347 |
| Feature Set 2 | 571894 | 1.149 | 54498 | 0.251 |
| Feature Set 3 | 353549 | **0.801** | **279838** | **0.381** |

✅ So we chose feature set 3

# Part 1 - K-Means Clustering

## FINDING BEST FEATURE SET

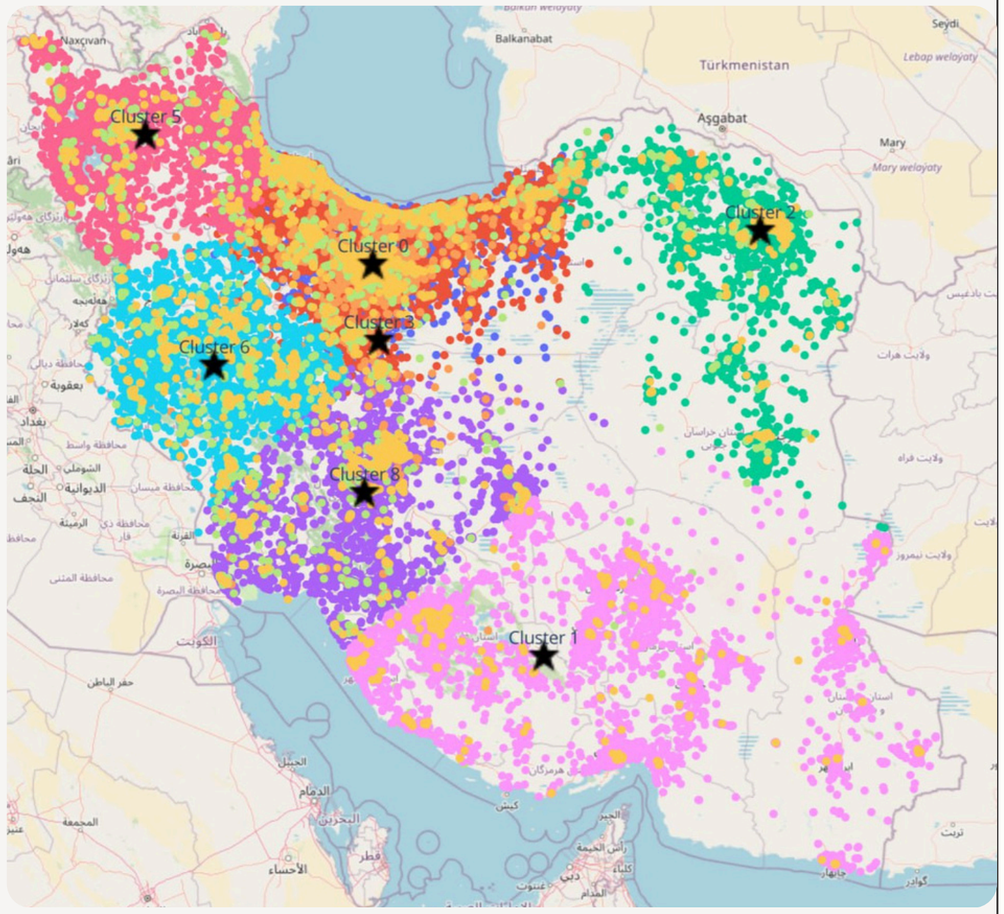


# Part 1 - K-Means Clustering by 10 Clusters

We clustered our data based on UTM coordinates and converted price value.

## Insights:

- One cluster covers most properties in the **Northern-Center.**
- One Cluster covers most properties in the **Northwest**.
- One cluster covers most properties in the **West**.
- One cluster covers most properties in the **Northeast**.
- One cluster covers most properties in the **South** and **Southeast**.
- One Cluster covers most properties in the **Southwest**.



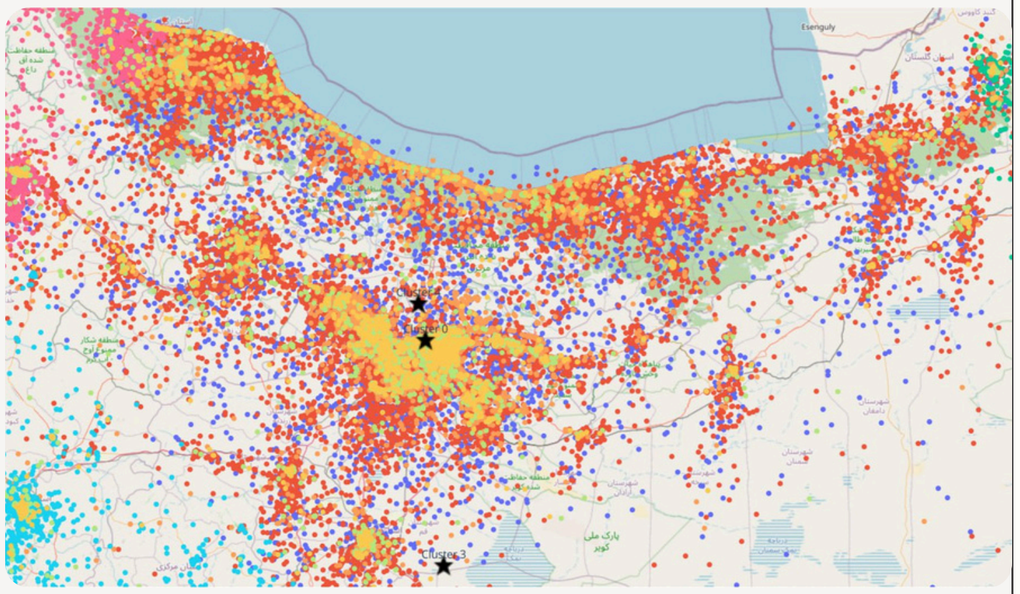

🔍 Some regions contain multiple clusters. One conclusion that can be drawn from this is the diversity of prices between different neighborhoods or **class differences** within the region.

# Part 1 - K-Means Clustering by 10 Clusters

We clustered our data based on UTM coordinates and converted price value.

**Insights:**

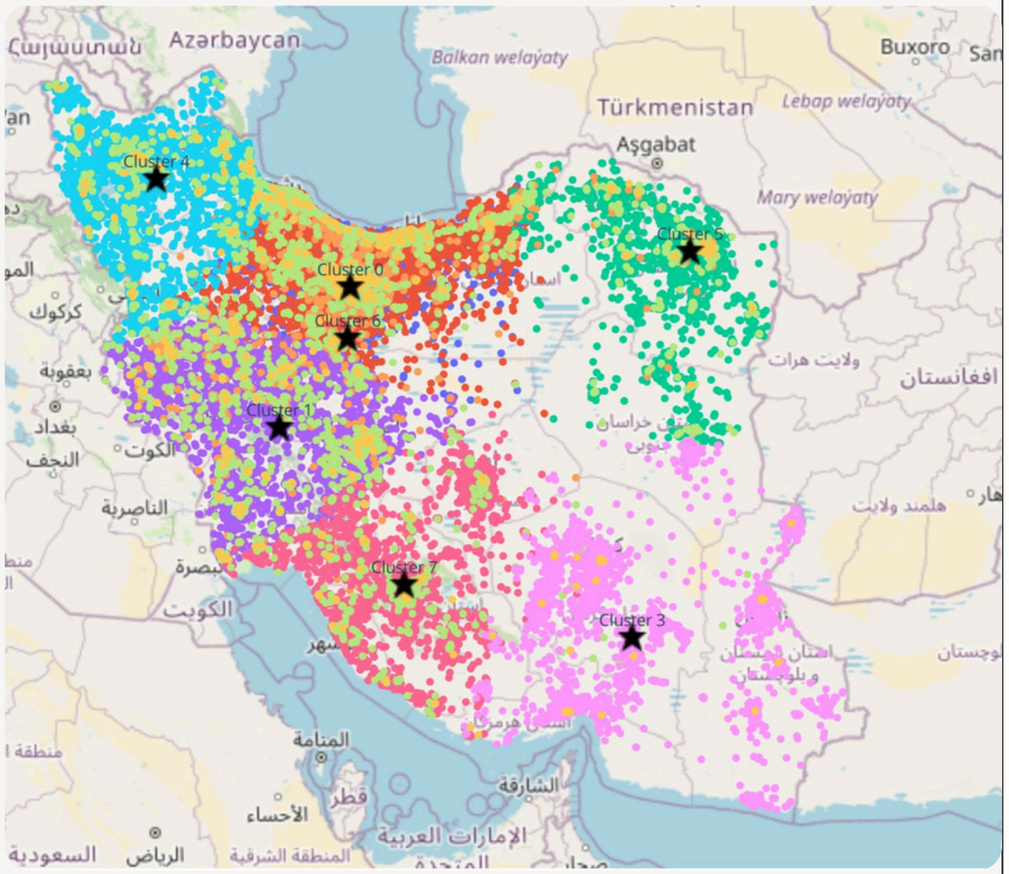
- One cluster covers most properties in the **Northern-Center.**
- One Cluster covers most properties in the **Northwest**.
- One cluster covers most properties in the **West**.
- One cluster covers most properties in the **Northeast**.
- One cluster covers most properties in the **South** and **Southeast**.
- One Cluster covers most properties in the **Southwest**.




🔍 Some regions contain multiple clusters. One conclusion that can be drawn from this is the diversity of prices between different neighborhoods or **class differences** within the region.

# Part 1 - K-Means Clustering by 10 Clusters

We also used **Yeo-Johnson's** power transformer to normalize our price data, and it ended up with a very similar clustering to when we applied the **Log transformation**.



# Part 2 - Finding Best K

Utilizing multiple methods, we aimed for finding most suitable values for **K.**

---

Here we also found optimum K values for both approaches (Log & Yeo-Johnson transformations)

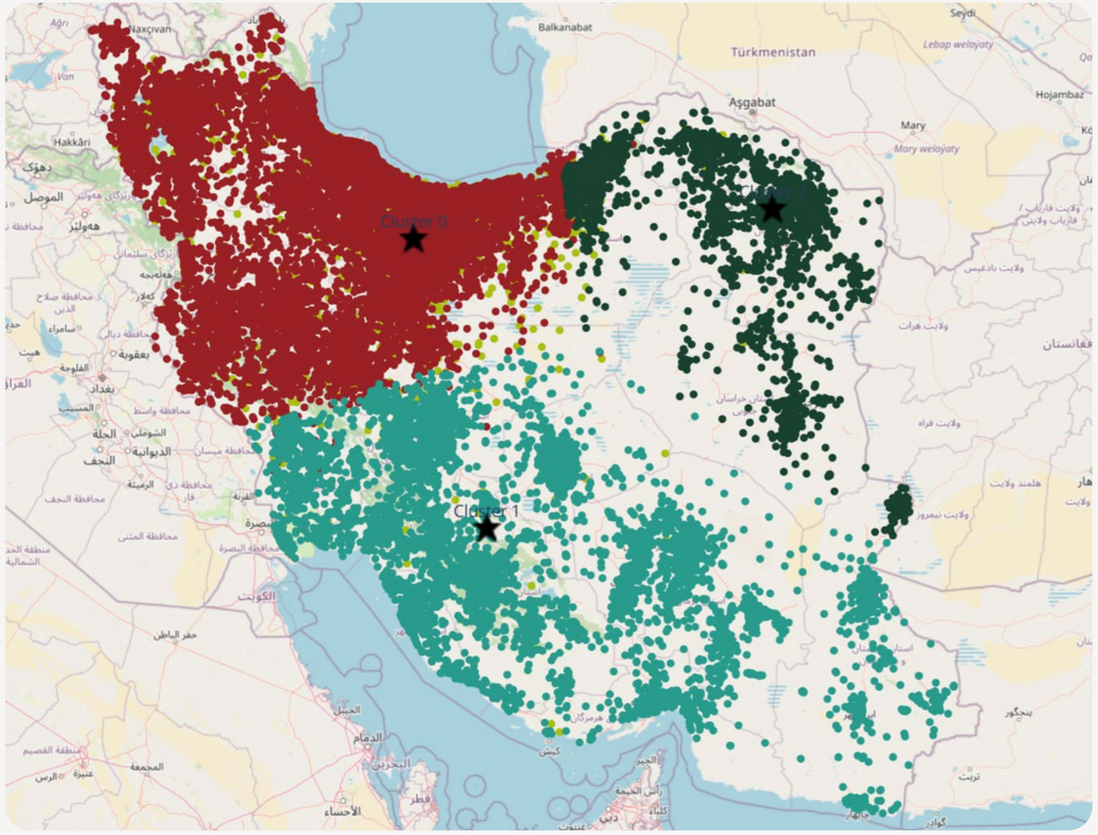
## Summerized Table

| TRANSFORMATION APPROACH | INERTIA (ELBOW) | DAVIES-BOULDIN | CALINSKI-HARABASZ | SILHOUETTE |
|---|---|---|---|---|
| Log | 10 or 7 | 4 | 7 | 4 |
| Yeo-Johnson | 10 or 7 | 7 | 10 | 3 |

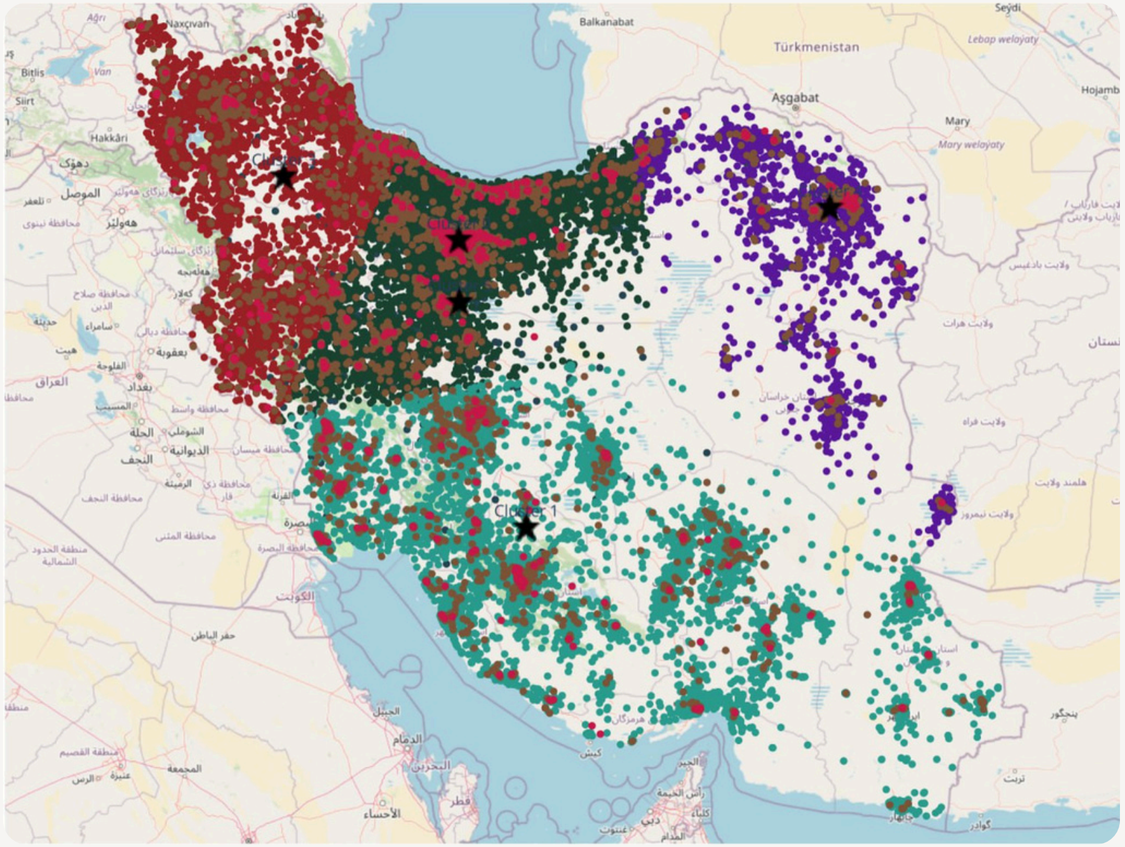
🔄 According to the table, numbers 4, 7, and 10 are the most repeated. So we plot their clusters on the map.

# Part 2 - Finding Best K

## Plot with **4** Clusters

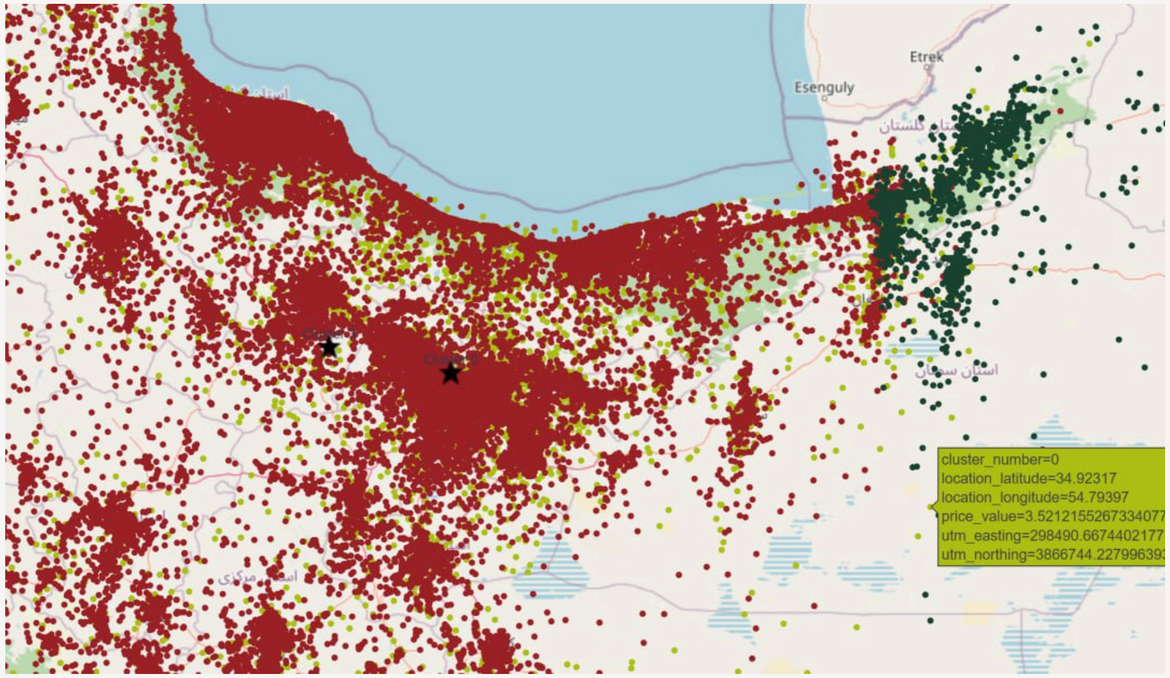



## Plot with **7** Clusters

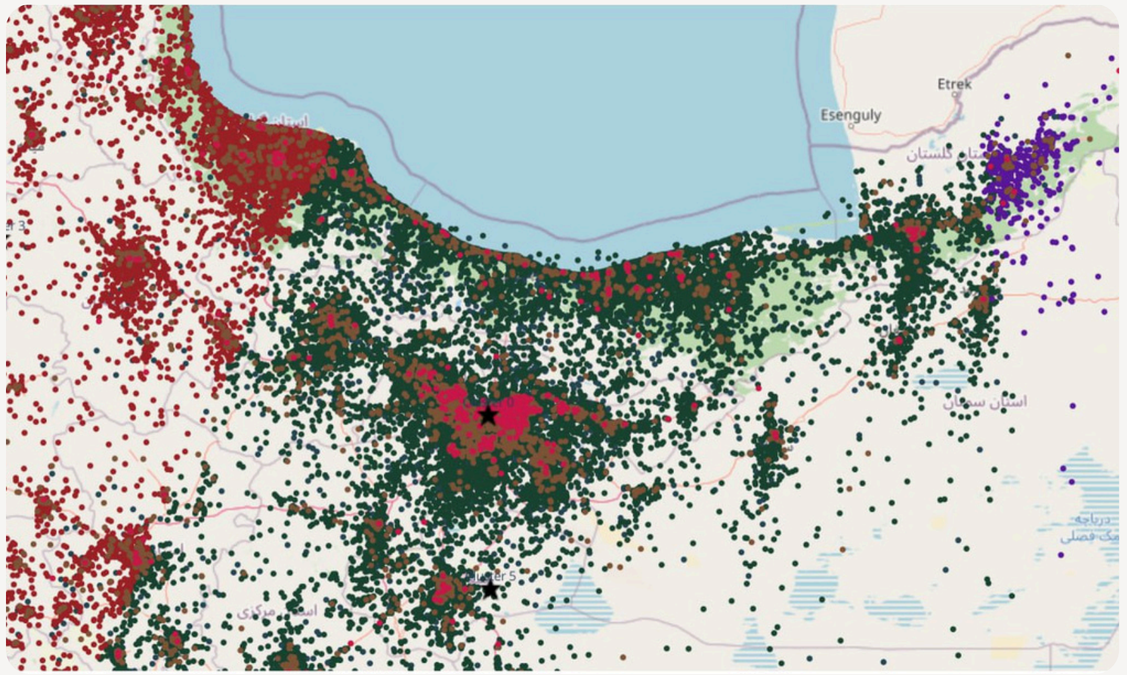


# Part 2 - Finding Best K

## Plot with **4** Clusters

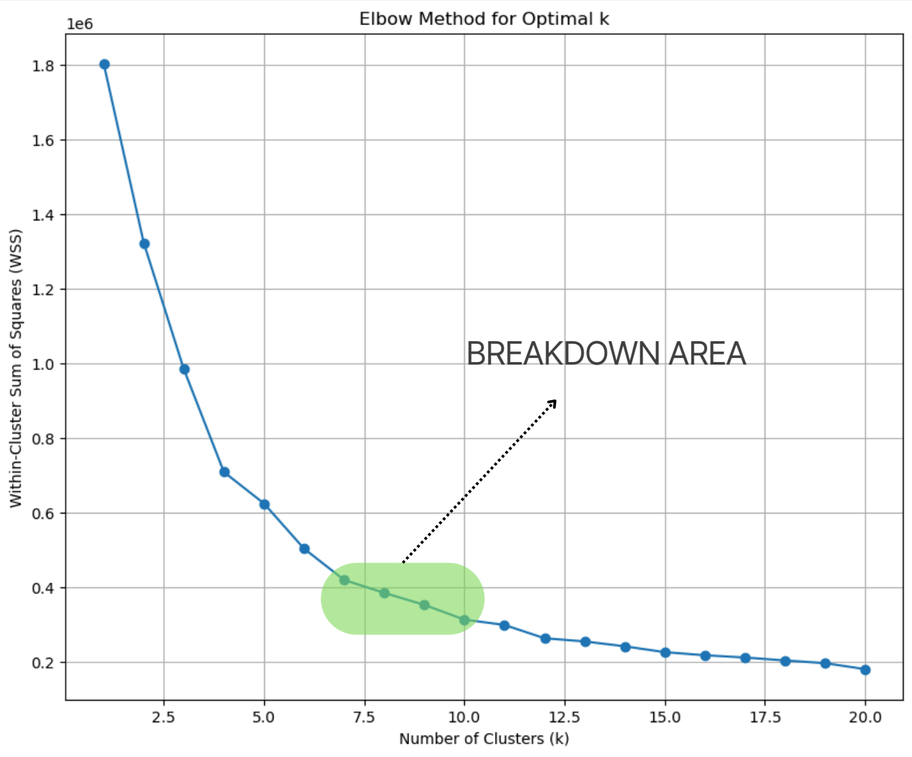



## Plot with **7** Clusters

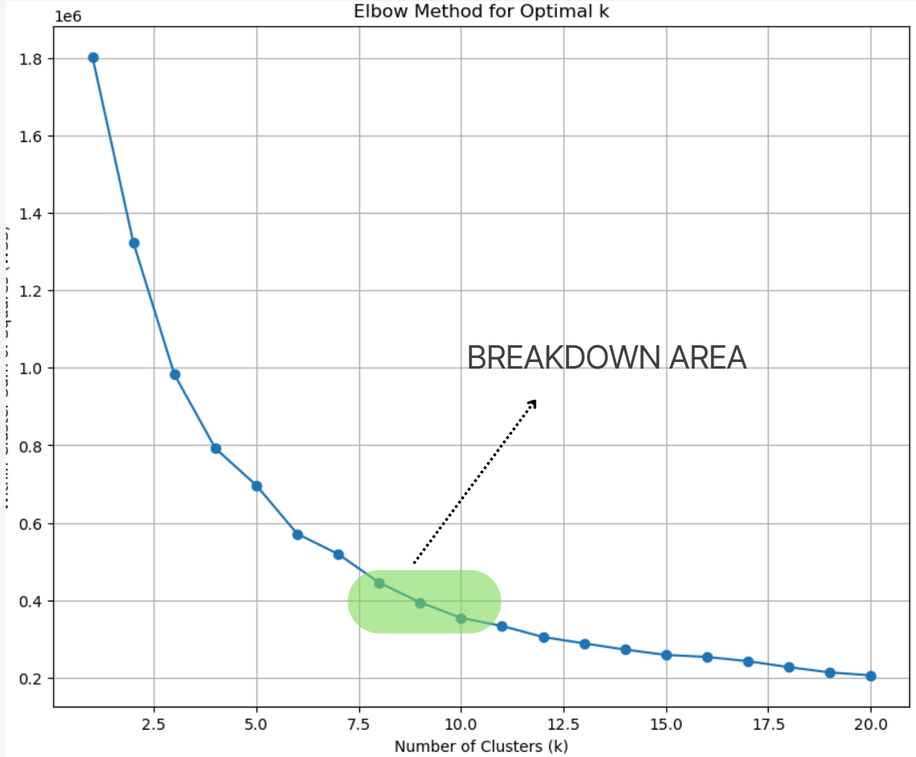


# Part 2 - Methods Plots: Elbow

## Log Transformation

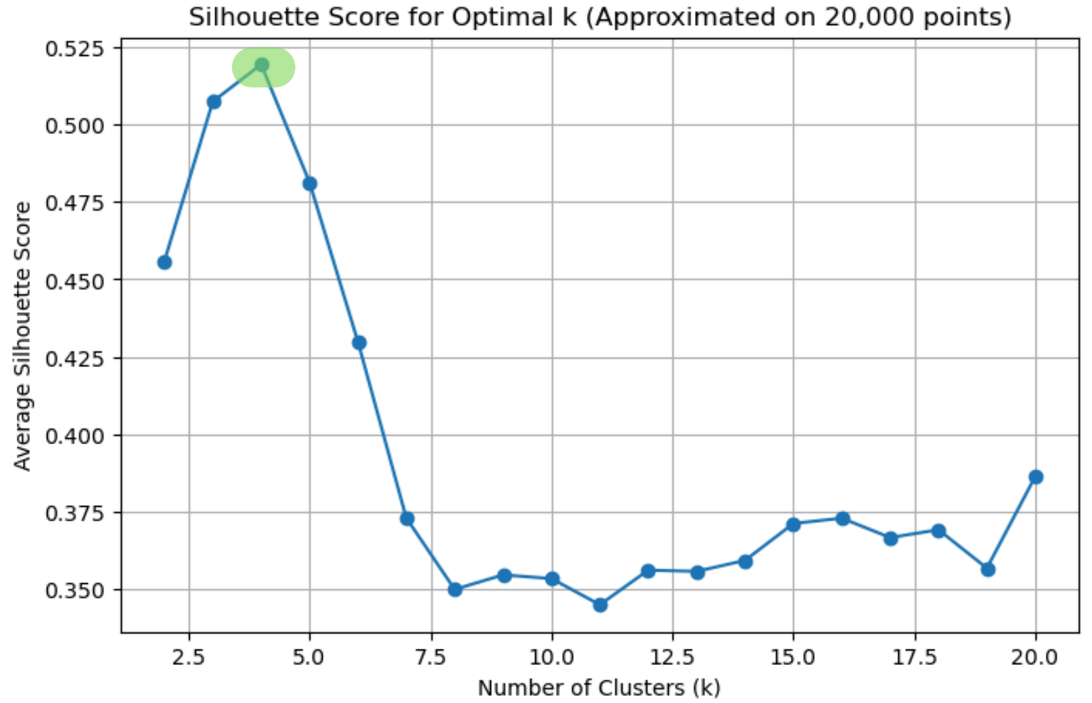



## Yeo-Johnson Transformation

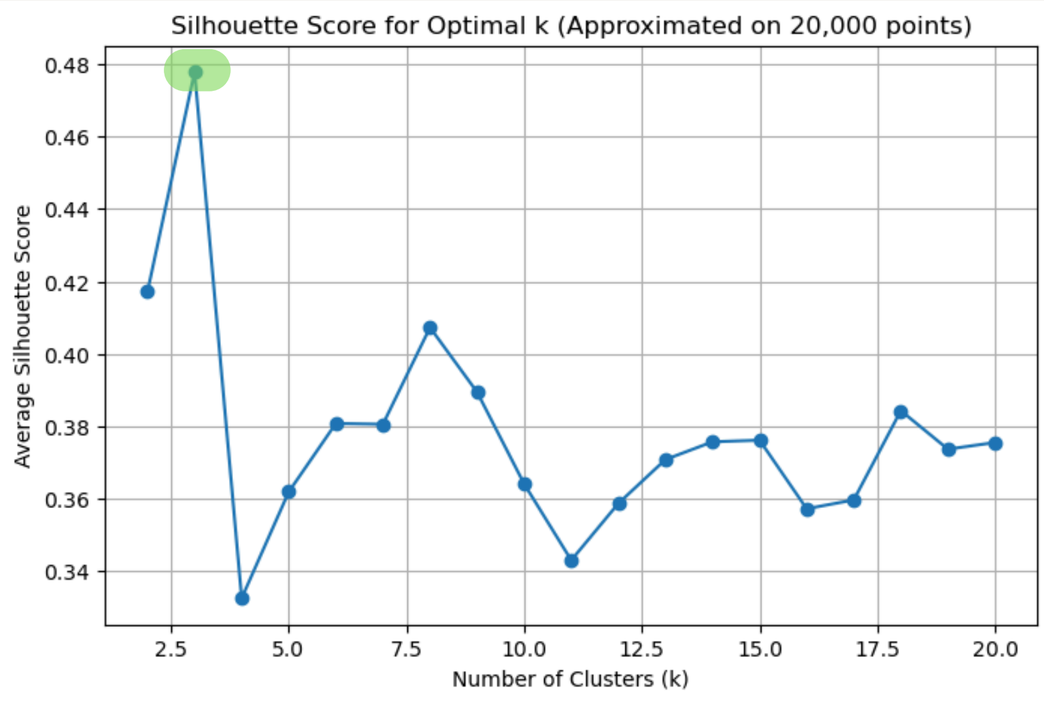


# Part 2 - Methods Plots: Silhouette

## Log Transformation

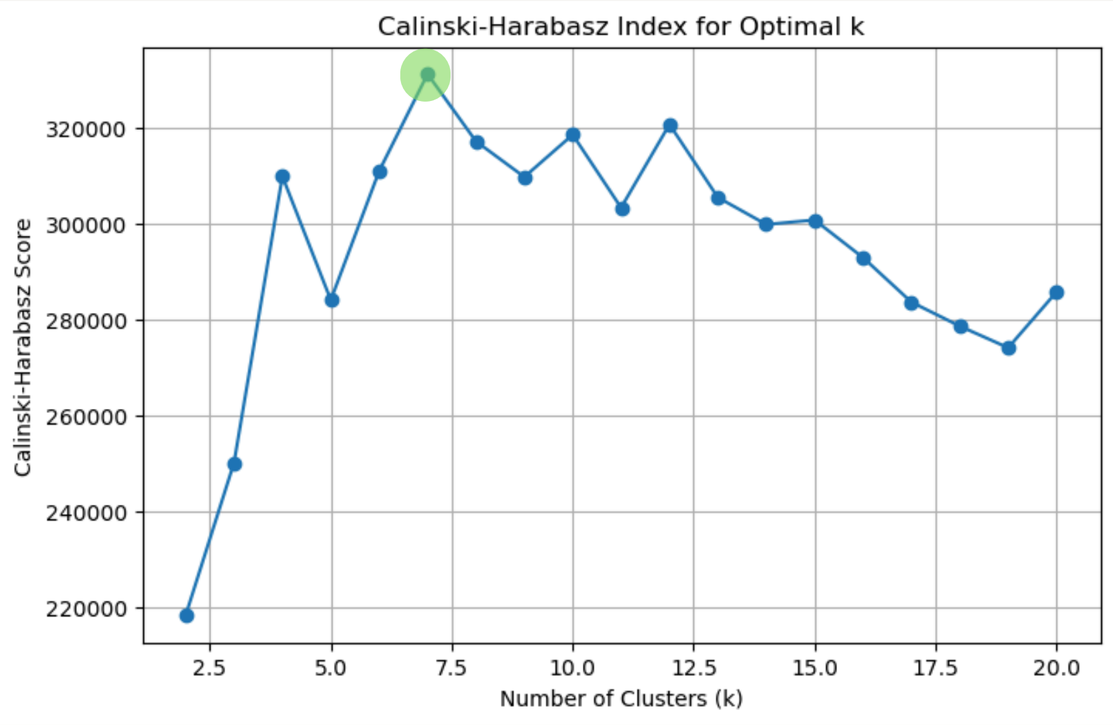



## Yeo-Johnson Transformation

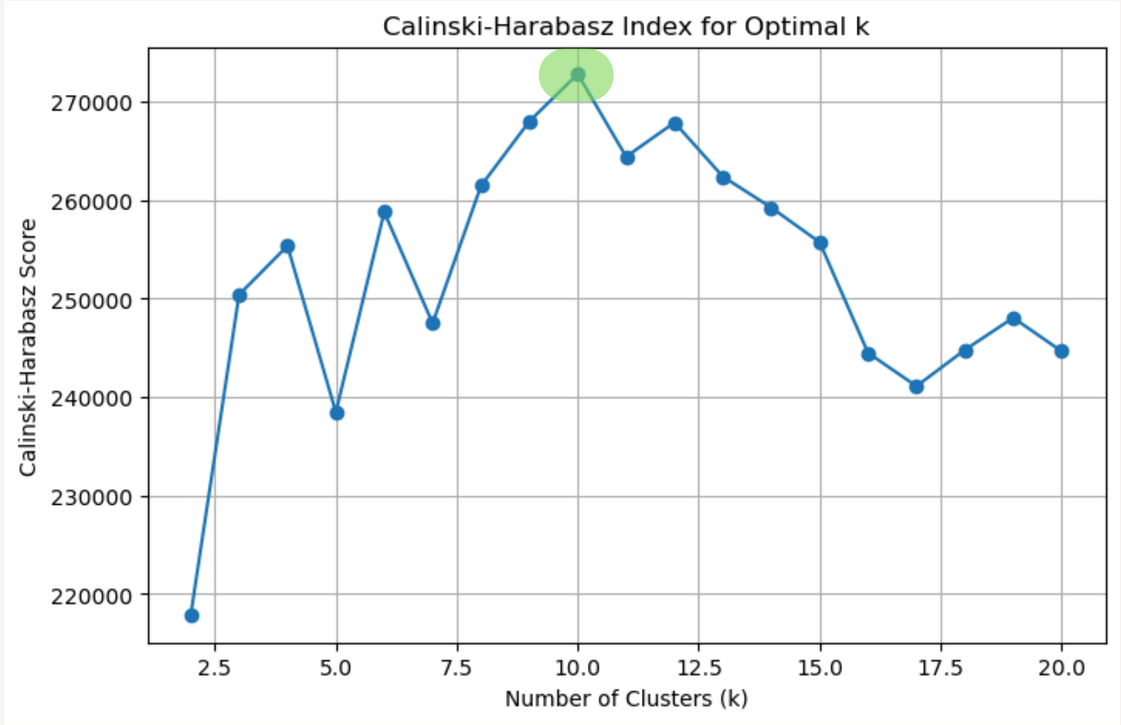


# Part 2 - Methods Plots: Calinski

## Log Transformation

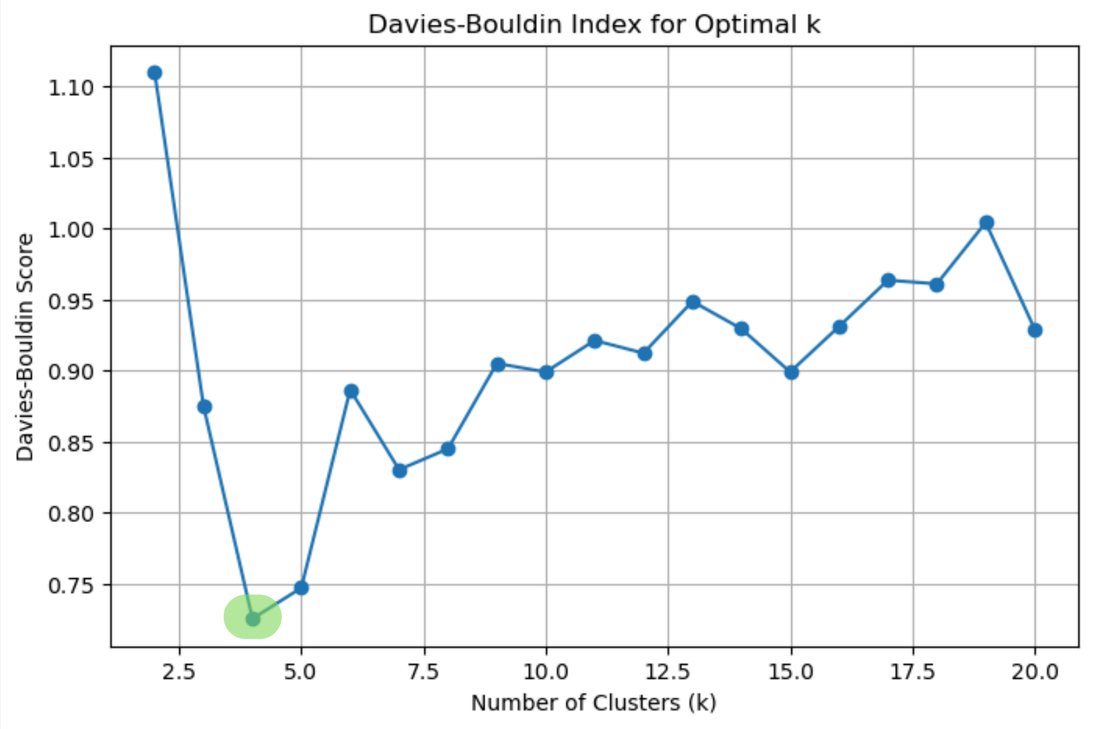



## Yeo-Johnson Transformation

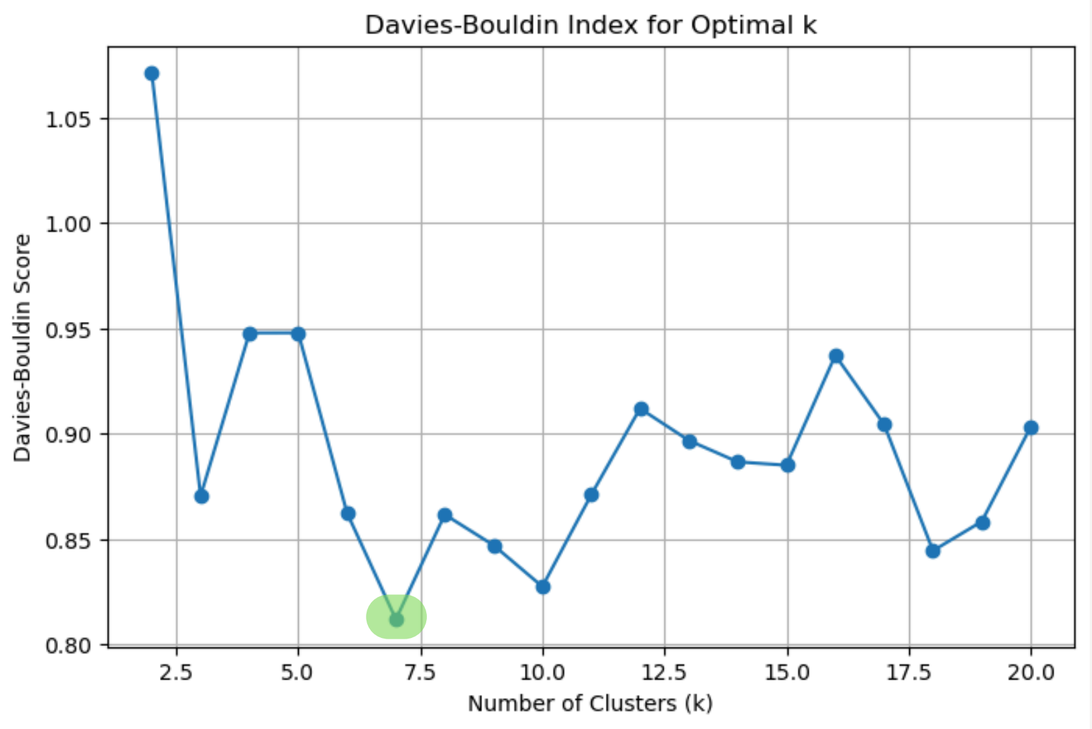


# Part 2 - Methods Plots: Davies-Bouldin

## Log Transformation




## Yeo-Johnson Transformation



# Part 3 - DBSCAN Clustering

## PURPOSE

Like the first part, we apply clustering on feature set 3, and our purpose is to reach 3 meaningful clusters by tuning the model's hyperparameters (esp, min_samples).

**eps (maximum neighbor distance):** Controls the spatial reach of density

- ⬇ **Smaller eps:** Points must be very close to be connected → results in **more clusters**, more fragmented data, and often more **noise points**.
- ⬆ **Larger eps:** Points can be farther apart and still connect → results in **fewer clusters** (merging distant groups), potentially one large cluster, and **less noise**.

**min_samples (minimum points for a core point):** Controls how dense a region must be to qualify as a cluster

- ⬇ **Smaller eps:** Easier to form core points → larger/more clusters, less noise, more tolerant of sparser regions (can create small or chained clusters).
- ⬆ **Larger eps:** Requires higher density to form core points → fewer clusters, more noise, ignores low-density areas (clusters are denser and more robust).

# Part 3 - DBSCAN Clustering (3 clusters)

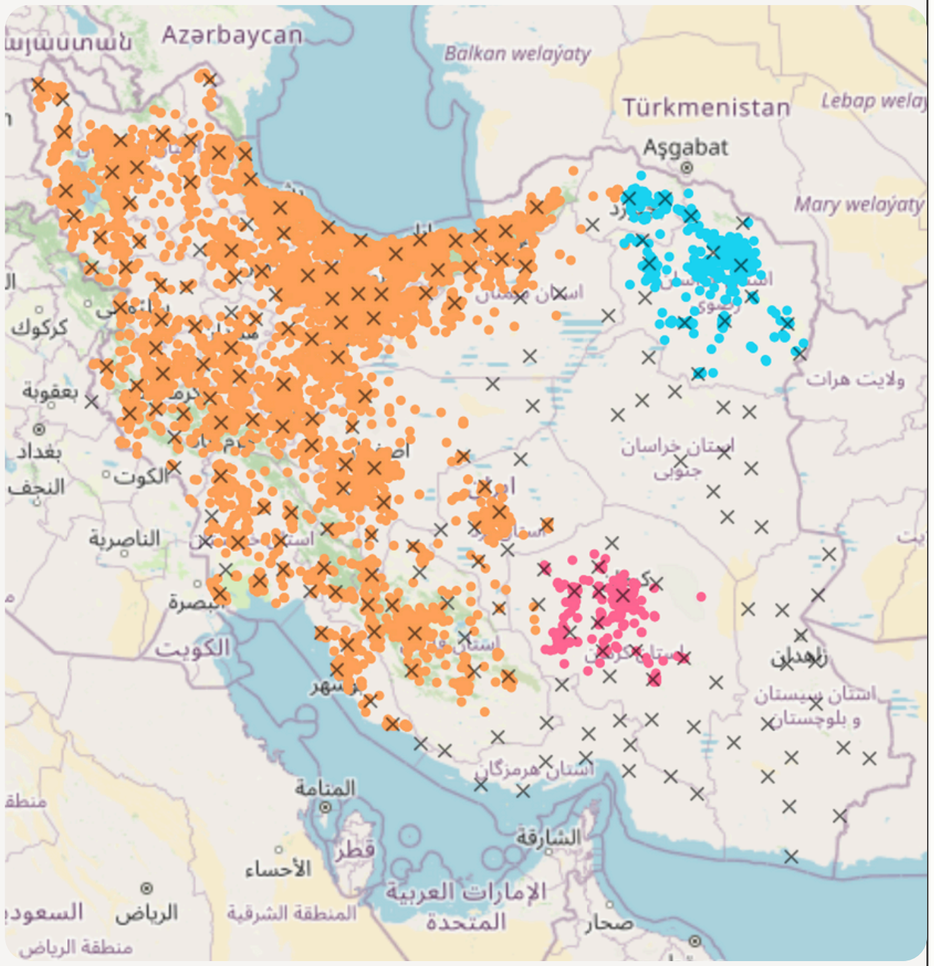
We clustered our data based on UTM coordinates and converted price value.

✔ After examining different values for **eps** and **min_samples**, we discovered that these values give us reasonable 3 clusters → eps=0.5 & min_samples=650

## Insights:

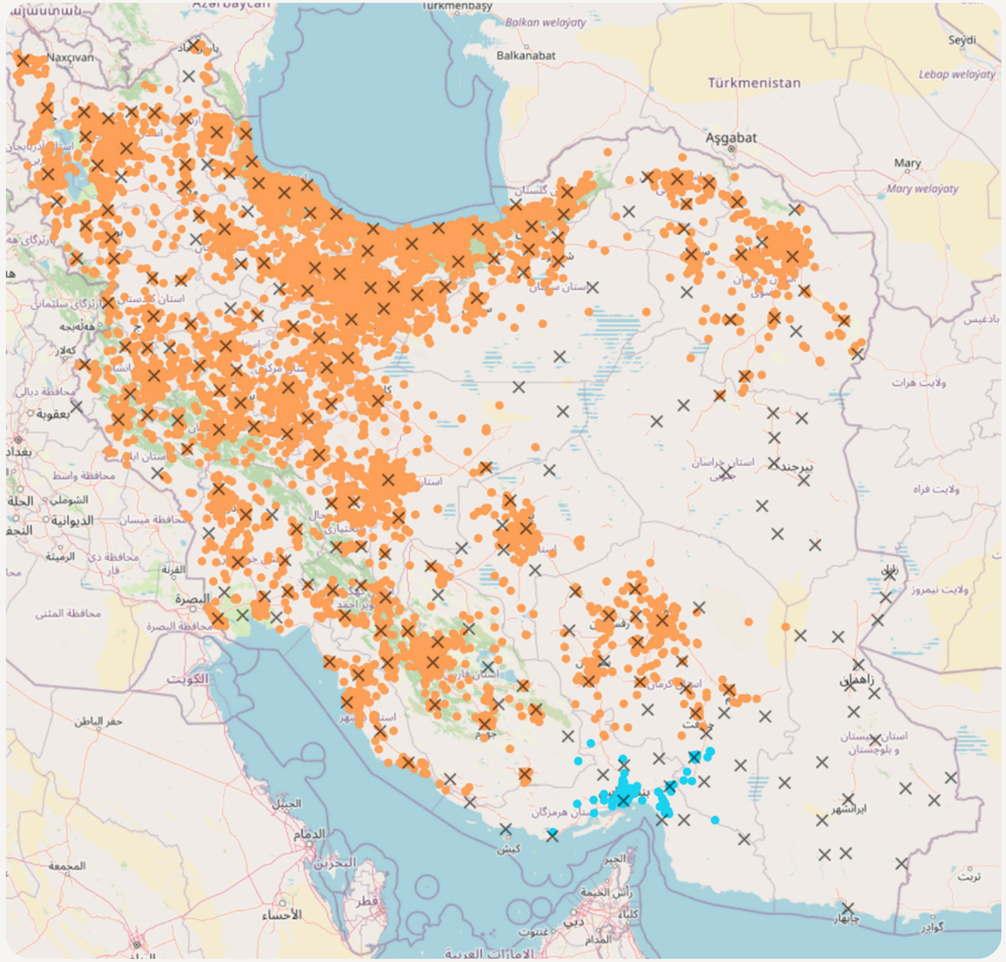
- One cluster covers most properties in the **West**.
- Properties in the **Northeast** shaped a unit cluster.
- Properties in **Southeast** shape another unit cluster.
- Crosses show noises on the map.

⚠ Since the data contains 600 thousand records, and the DBSCAN would take so long to perform, we used a smaller sample of 100 thousand rows to feed the algorithm.

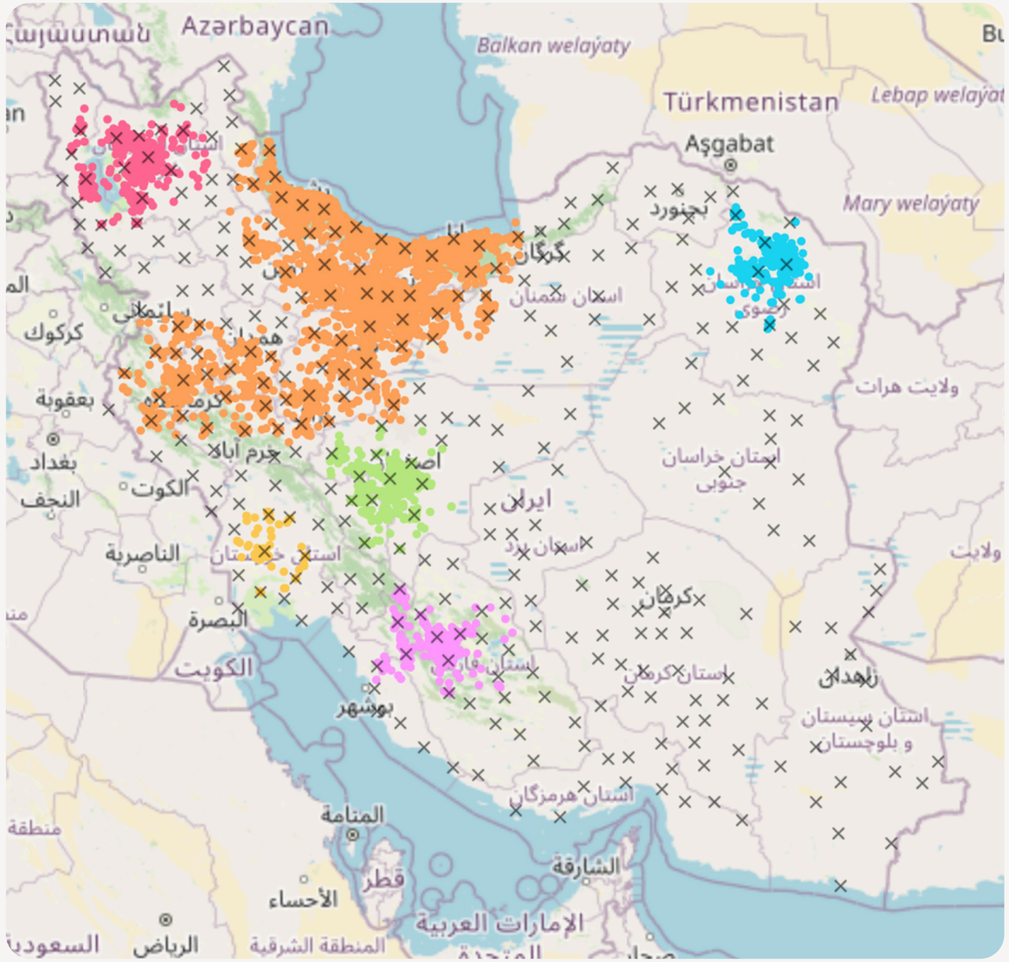


# Part 3 - Effects of eps

## esp=0.6

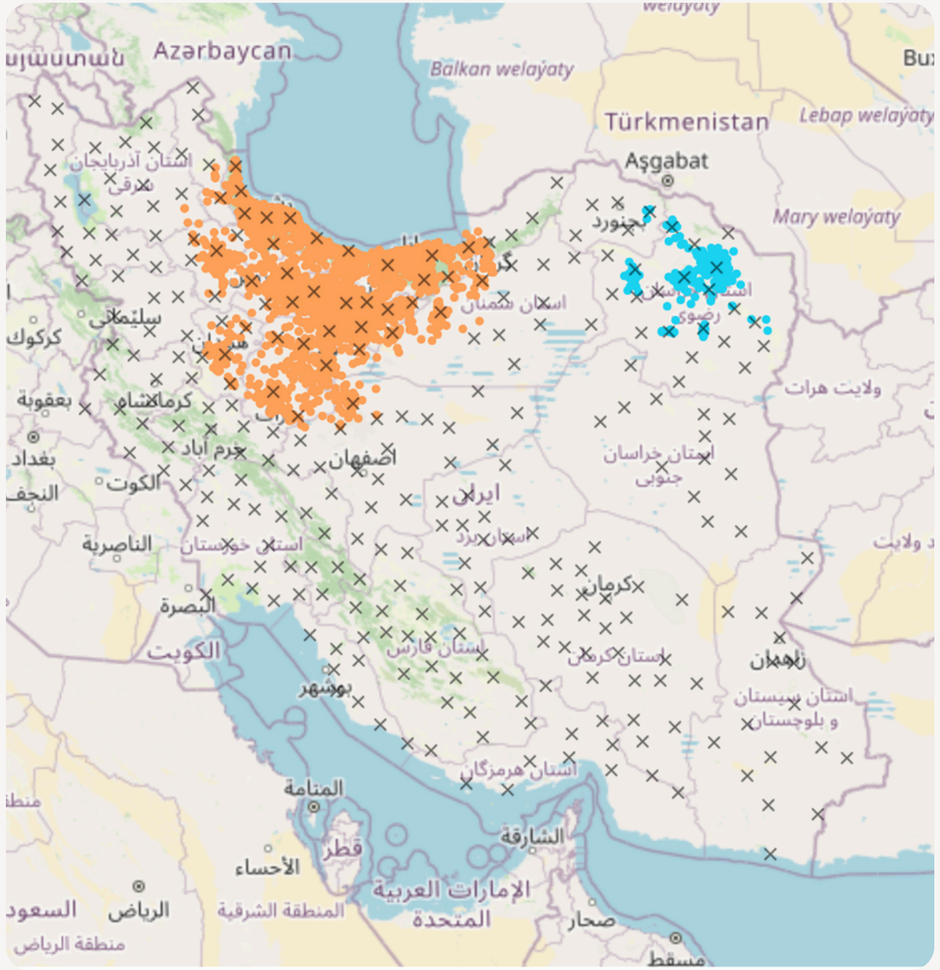



## esp=0.3

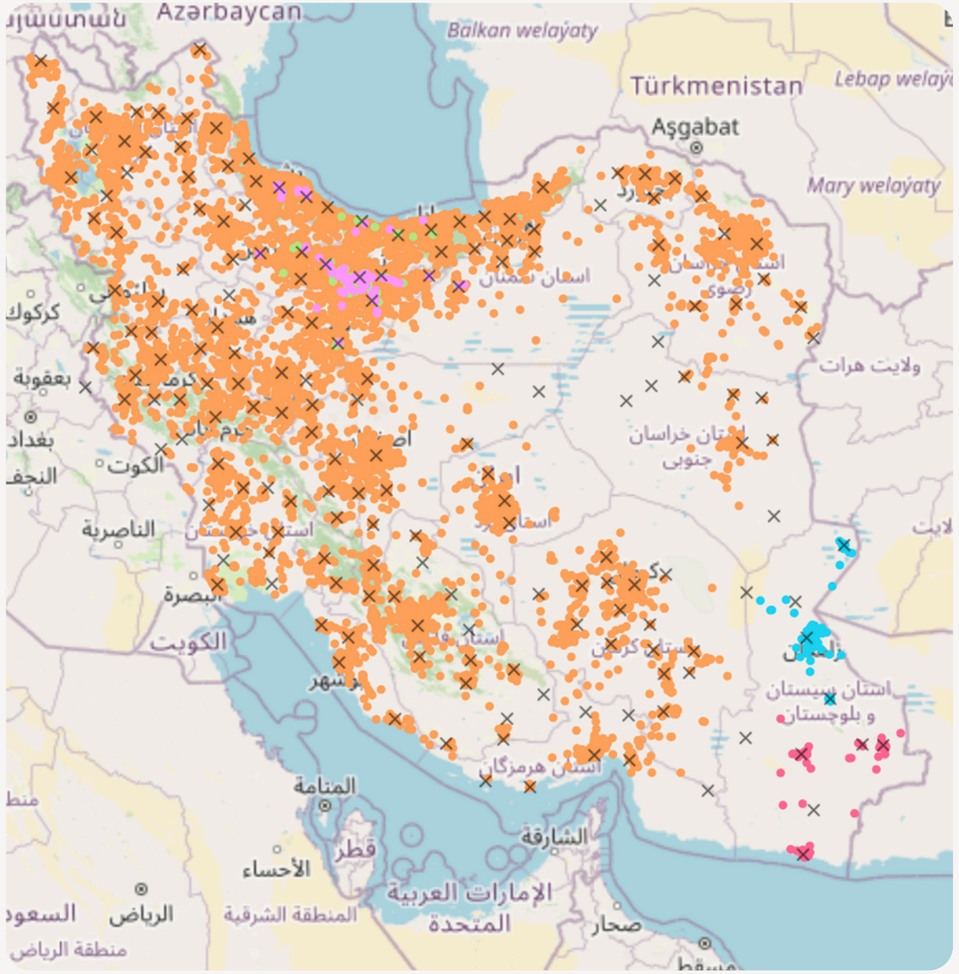


# Part 3 - Effects of min_samples

## min_samples=4000




## min_samples=50



GROUP-3

DEC 2025

# ML / PREDICTION MODELS

Predicting price & total credit values

FIRST PROJECT REPORT

45 - 55

GITHUB.COM/MEHDIRT/DIVAR-PROJECT

# Prediction Models

## PURPOSE

In this part, we made prediction models to estimate the price or total credit of a property, given its attributes.

At first, data records are separated into sales data and rent data. Then, we preprocessed the datasets, removed some features, engineered the remaining ones, encoded categoricals, and fed them to models.

Finally, we trained two tree-based regression models on each group: **Random Forrest** & **LGBoost**. Additionally, their hyperparameters were tuned with **Bayesian Search** method.

# Total Credit Prediction

## PREPROCESSING

→ At first, some useless features and the ones with too many missing values were dropped, such as:
*title, description, created_at_month, user_type, location_radius, price_mode, credit_mode, has_water, has_electricity, has_gas, land_size* ***etc***.

→Binned numeric features and made them categorical like: *floor, building_age* .

→Created new features: building_age, luxury_items, non_luxury_items

→Handle missings of categorical features: Map NaNs to unselect

→Imputed coordinate features by grouping on city_slug

# Total Credit Prediction

## OUTLIER HANDLING

There are very large and very small, unreasonable values for total credit and building size that might cause errors on model. We simply cut them from the dataset using a threshold.

**Chosen total credit range:** 50,000,000 to 80,000,000,000

**Why?**

In Iran's property market, at this time, total rents are mostly below 50 billion, and above 50 million Tomans. As we checked in Divar, one of the largest online classified marketplace applications in Iran, most of the advertisements are in this range. In addition, some prices are old and too low, counted as noise, so we try to put them aside. Plus, we are using tree-based models that act poorly when they get an input out of training range.

**Chosen building size range:** 20 to 3000

**Why?**

According to Iran's market, most of the properties published are above 20 and below 3000, as we checked at Divar application. When we go beyond this range, values become very large or low, and prices become extreme. Plus, we are using tree-based models that act poorly when they get an input out of training range.

# Total Credit Prediction

## ENCODING

→ Encoding boolean features by mapping to 1, 0, and -1.

→ Target encoding for **city_slig** and **neighborhood_slug.**

→One-Hot encoding for all categorical features

# Performance Overview

## MODELS EVALUATIONS ON VALIDATION SET

| MODEL | R2 SCORE | MAE | MSE |
|---|---|---|---|
| Random Forrest | 0.7411 | 1591485567 | 3957320608 |
| LGBoost | 0.7467 | 1545942398 | 3914446589 |

## MODELS EVALUATIONS ON TEST SET

| MODEL | R2 SCORE | MAE | MSE |
|---|---|---|---|
| Random Forrest | 0.7481 | 1567528851 | 3911627743 |
| LGBoost | 0.7549 | 1539790556 | 3858565435 |

# Sell Price Prediction

## PREPROCESSING

➔ At first, some useless features and the ones with too many missing values were dropped, such as:
*title, description, created_at_month, user_type, location_radius, price_mode, credit_mode, has_water, has_electricity, has_gas, land_size* ***etc***.

➔Binned numeric features and made them categorical like: *floor, building_age* .

➔Created new features: building_age, luxury_items, non_luxury_items

➔Handle missings of categorical features: Map NaNs to unselect

➔Imputed coordinate features by grouping on city_slug

# Sell Price Prediction

## OUTLIER HANDLING

There are very large and very small, unreasonable values for total credit and building size that might cause errors on model. We simply cut them from the dataset using a threshold.

**Chosen total credit range:** 500,000,000 to 80,000,000,000

**Why?**

In Iran's property market, at this time, total rents are mostly below 500 billion, and above 50 million Tomans. As we checked in Divar, one of the largest online classified marketplace applications in Iran, most of the advertisements are in this range. In addition, some prices are old and too low, counted as noise, so we try to put them aside. Plus, we are using tree-based models that act poorly when they get an input out of training range.

**Chosen building size range:** 20 to 3000

**Why?**

According to Iran's market, most of the properties published are above 20 and below 3000, as we checked at Divar application. When we go beyond this range, values become very large or low, and prices become extreme. Plus, we are using tree-based models that act poorly when they get an input out of training range.

# Total Credit Prediction

## ENCODING

→ Encoding boolean features by mapping to 1, 0, and -1.

→ Target encoding for **city_slig** and **neighborhood_slug.**

→One-Hot encoding for all categorical features

# Performance Overview

## MODELS EVALUATIONS ON VALIDATION SET

| MODEL | R2 SCORE | MAE | MSE |
|---|---|---|---|
| Random Forrest | 0.7000 | 1869806290 | 4404355876 |
| LGBoost | 0.7114 | 1874323969 | 4319188301 |

## MODELS EVALUATIONS ON TEST SET

| MODEL | R2 SCORE | MAE | MSE |
|---|---|---|---|
| Random Forrest | 0.6969 | 1567528851 | 1869783237 |
| LGBoost | 0.7107 | 1870714791 | 4415601567 |

# THANKYOU

GITHUB.COM/MEHDIRT/DIVAR-PROJECT