Course Project 1 - Linear Regression

Disclaimer: Since we had constraint on the number of pages, we were unable to add code snippets which we have added them to Jupyter Notebooks in the GitHub repo.

1 Motivation

In this project, we aim to apply linear regression techniques to predict house prices based on various features extracted from real estate listings obtained from the bina.az website.

The motivation behind this project is to create a trustworthy forecasting model that will help stakeholders make well-informed decisions about transactions and investments in real estate. In particular, we wanted to investigate how well linear regression predicts outcomes when there is only one predictor and when there are several predictors.

2. Problem statement

Our objective is to develop a prediction model that, given a collection of predictor variables, including area, number of rooms, floor, and other important properties, can reliably forecast the selling price of homes. We want to comprehend the fundamental variables impacting property prices in Azerbaijan's real estate market by examining historical data from the bina.az website.

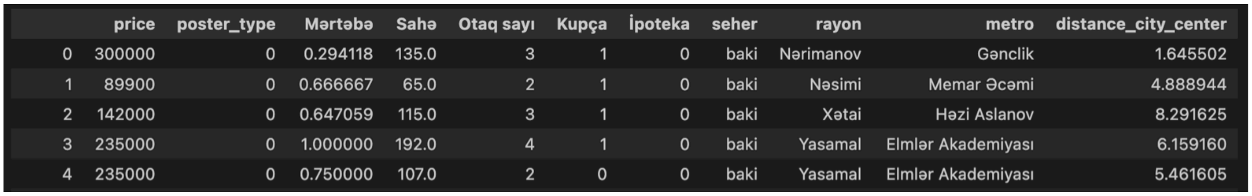
3. Methodology

We used the linear regression technique for our study in this assignment. It works well for generating the relationship between features, which are independent variables, and housing price, which is a continuous target variable. In our experiments, we have used both single predictor and multiple predictor simple linear regression approaches.

4. Experiments

4.1 Data collection and Preprocessing

The data was scraped from the bina.az website, extracting relevant features such as floor, size, number of rooms, Ipoteka, Kupca, longitude, latitude and price. The collected data were preprocessed to handle missing values, outliers, and categorical variables. Additionally, we performed feature engineering to transform coordinate columns to distance_city_center and normalized the data for model training. Below is an example snapshot from the dataset that we have preprocessed and cleaned:



	price	poster_type	Mərtəbə	Sahə	Otaq sayı	Kupça	İpoteka	seher	rayon	metro	distance_city_center
0	300000	0	0.294118	135.0	3	1	0	baki	Nərimanov	Gənclik	1.645502
1	89900	0	0.666667	65.0	2	1	0	baki	Nəsimi	Memar Əcəmi	4.888944
2	142000	0	0.647059	115.0	3	1	0	baki	Xətai	Həzi Aslanov	8.291625
3	235000	0	1.000000	192.0	4	1	0	baki	Yasamal	Elmlər Akademiyası	6.159160
4	235000	0	0.750000	107.0	2	0	0	baki	Yasamal	Elmlər Akademiyası	5.461605

Figure 1. Samples from the dataset

4.2 Single predictor linear regressio

We implemented a custom linear regression model from scratch to predict house prices based on the area ("Sahə") of the properties. As you can see from the plot there is linear relationship between the price of the house and the area of it. The model was trained using 80% of the data and tested on the remaining 20%. We evaluated the model's performance using the following metrics:

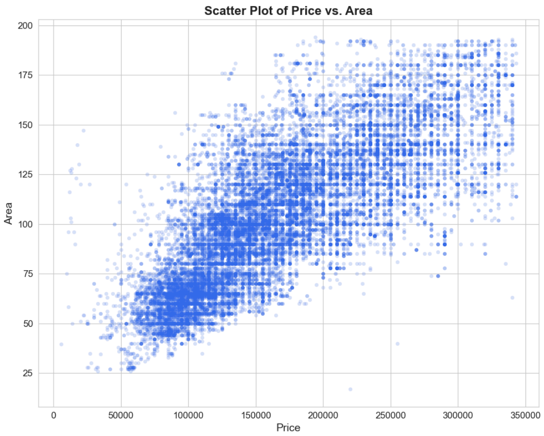


Figure 2. Relationship between area and price

- Residual Sum of Squares (RSS): Measures the discrepancy between the observed and predicted house prices.
- Root Mean Squared Error (RMSE): Provides a measure of the average prediction error.
- Standard Error (SE): Estimates the variability of coefficient estimates.
- Confidence Interval: Indicates the range of values within which the true coefficient is likely to fall with 95% probability.
- T-Statistic: Evaluates the significance of coefficient estimates.
- P-Value: Indicates the probability of observing the given t-statistic under the null hypothesis.
- R-Squared: Measures the proportion of variance in the target variable explained by the model.
- Correlation: Calculates the Pearson correlation coefficient between the predictor and target variables.

4.3 Multiple Predictor Linear Regression

In this experiment, we extended our analysis to incorporate multiple predictor variables for predicting house prices. Features such as Kupca, number of rooms, Ipoteka, Distance to city center and other relevant factors were included in the model. We applied a custom linear regression implementation to train the model on 80% of the data and evaluated its performance on the remaining 20%. The experiment involved the following steps:

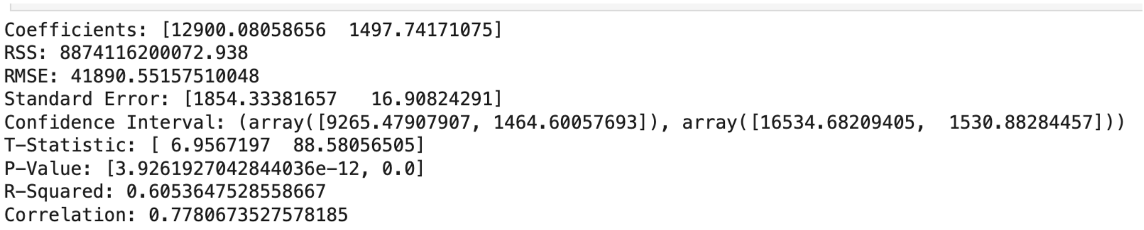
- Feature Selection: We used all available features except the target variable ("price") for training the multiple predictor linear regression model.
- Model Training and Evaluation: The model was trained using the training dataset, and predictions were made on the test dataset. We calculated the following metrics to assess the model's performance: Residual Sum of Squares (RSS), Root Mean Squared Error (RMSE), Standard Errors (SE), T-Statistics, P-Values, R-Squared.

We assessed the usefulness of predictors, model fit, and prediction accuracy. Additionally, we investigated interactions between variables and analyzed non-linear effects to enhance the model's performance.

5 Results

5.1 Single predictor regression results

- Coefficients: The intercept is approximately 12900.08, and the slope (associated with the predictor variable "Sahə" or area) is approximately 1497.74. This suggests that, on average, for every additional square meter of area, the predicted house price increases by approximately 1497.74 AZN.
- RSS: It is approximately 8.87 trillion AZN, indicating the overall discrepancy between the model's predictions and the actual prices.
- RMSE: Approximately 41890.55 AZN indicates the typical difference between the predicted and observed house prices.
- SE: The standard error estimates the variability of coefficient estimates. In this case, the standard error for the intercept is approximately 1854.33, and for the slope, it's approximately 16.91.
- Confidence Interval: For the intercept, the confidence interval ranges from approximately 9265.48 to 16534.68, and for the slope, it ranges from approximately 1464.60 to 1530.88.
- T-Statistics and P-Value: Both coefficients have low p-values (close to zero), suggesting that they are statistically significant predictors of house prices.
- R_Squared: With an R-squared of approximately 0.61, it indicates that around 61% of the variability in house prices can be explained by the predictor variable "Sahə" (area).
- Correlation: Coefficient of approximately 0.78, it indicates a strong positive correlation between the area of properties and their prices.



```
Coefficients: [12900.08058656  1497.74171075]
RSS: 8874116200072.938
RMSE: 41890.55157510048
Standard Error: [1854.33381657   16.90824291]
Confidence Interval: (array([9265.47907907, 1464.60057693]), array([16534.68209405,  1530.88284457]))
T-Statistic: [ 6.9567197  88.58056505]
P-Value: [3.9261927042844036e-12, 0.0]
R-Squared: 0.6053647528558667
Correlation: 0.7780673527578185
```

Figure 3. The results of single predictor linear regression

5.2 Multiple predictor regression results.

In this experiment the Residual Sum of Squares (RSS), Root Mean Squared Error (RMSE), Standard Errors (SE), T-Statistics, P-Values, R-Squared metrics that we got can be interpreted the same way as in the single predictor regression.

a) Is at least one of the predictors X1, X2, …, Xp useful in predicting the response?

This can be seen from the statistically significant coefficients associated with several predictor variables, as indicated by low p-values (< 0.05). A low p-value suggests that the corresponding predictor variable is statistically significant and contributes to explaining the variability in house prices. See the Figure 5.

b) Do all the predictors help to explain Y, or is only a subset of the predictors useful?

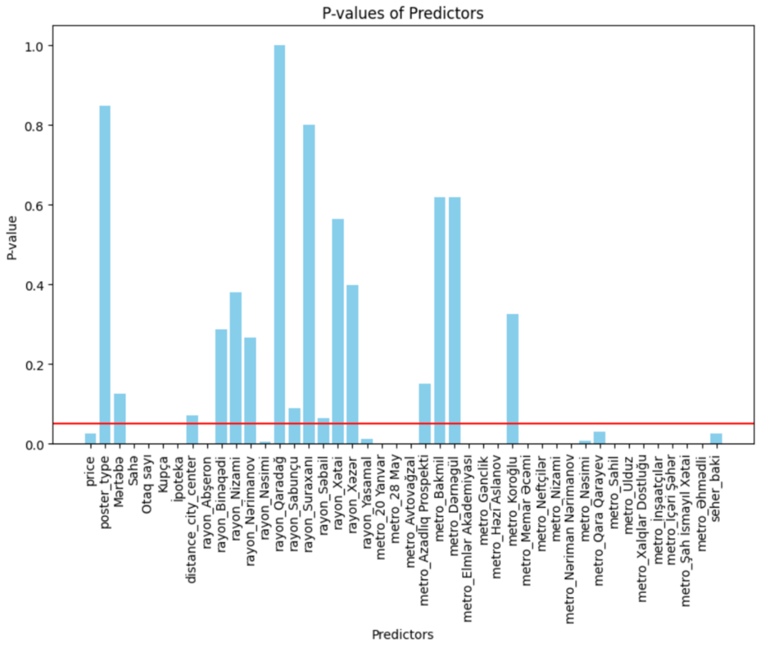


Figure 4. P-Values of predictors.

Not all predictors may help explain house prices, as indicated by the presence of statistically non-significant coefficients (with p-values > 0.05) (figure 4). We can find the subset of useful predictors by using the backward elimination. By systematically eliminating less important features, backward elimination helps to identify a subset of predictor variables that are most relevant for predicting the target variable. See figure 7.

c) How well does the model fit the data?

The model exhibits a relatively high R-squared value of approximately 0.75, indicating that around 75% of the variability in house prices is explained by the predictor variables included in the model. This suggests that the model provides a reasonably good fit to the data, capturing a substantial portion of the variability in house prices.

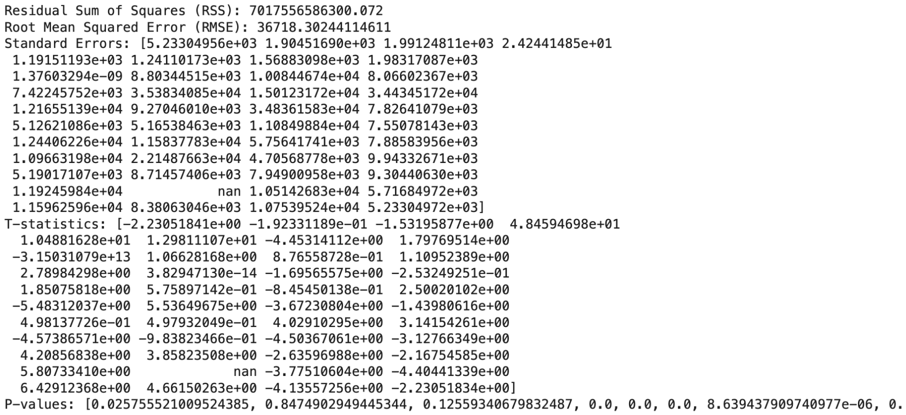
d) Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

The accuracy of our prediction can be assessed using metrics such as RMSE (Root Mean Squared Error), which provides an estimate of the average prediction error. In our case we see very high RMSE which could be due to the lack of the features or the Linear Regression model is way too simple for our data. We plan to experiment to find answer to our question.

e) Analyze Interactions between qualitative and quantitative Variables

We analyze the interactions between qualitative and quantitative variables using simple Pearson correlation. You can see some negative and positive relationship between them upon inspecting the plot below. For example there is a slightly high positive relationship between the "poster_type" and "kupça". See figure 6.

f) Analyze Non-linear Effects of Predictors



```
Residual Sum of Squares (RSS): 7017556586300.072
Root Mean Squared Error (RMSE): 36718.30244114611
Standard Errors: [5.23304956e+03 1.90451690e+03 1.99124811e+03 2.42441485e+01
 1.19151193e+03 1.24110173e+03 1.56883098e+03 1.98317087e+03
 1.37603294e-09 8.80344515e+03 1.00844674e+04 8.06602367e+03
 7.42245752e+03 3.53834085e+04 1.50123172e+04 3.44345172e+04
 1.21655139e+04 9.27046010e+03 3.48361583e+04 7.82641079e+03
 5.12621086e+03 5.16538463e+03 1.10849884e+04 7.55078143e+03
 1.24406226e+04 1.15837783e+04 5.75641741e+03 7.88583956e+03
 1.09663198e+04 2.21487663e+04 4.70568778e+03 9.94332671e+03
 5.19017107e+03 8.71457406e+03 7.94900958e+03 9.30440630e+03
 1.19245984e+04            nan 1.05142683e+04 5.71684972e+03
 1.15962596e+04 8.38063046e+03 1.07539524e+04 5.23304972e+03]
T-statistics: [-2.23051841e+00 -1.92331189e-01 -1.53195877e+00  4.84594698e+01
  1.04881628e+01  1.29811107e+01 -4.45314112e+00  1.79769514e+00
 -3.15031079e+13  1.06628168e+00  8.76558728e-01  1.10952389e+00
  2.78984298e+00  3.82947130e-14 -1.69565575e+00 -2.53249251e-01
  1.85075818e+00  5.75897142e-01 -8.45450138e-01  2.50020102e+00
 -5.48312037e+00  5.53649675e+00 -3.67230804e+00 -1.43980616e+00
  4.98137726e-01  4.97932049e-01  4.02910295e+00  3.14154261e+00
 -4.57386571e+00 -9.83823466e-01 -4.50367061e+00 -3.12766349e+00
  4.20856838e+00  3.85823508e+00 -2.63596988e+00 -2.16754585e+00
  5.80733410e+00            nan -3.77510604e+00 -4.40441339e+00
  6.42912368e+00  4.66150263e+00 -4.13557256e+00 -2.23051834e+00]
P-values: [0.025755521009524385, 0.8474902949445344, 0.12559340679832487, 0.0, 0.0, 0.0, 8.639437909740977e-06, 0.
```

Figure 5. The results of multiple predictor linear regression

We have also trained two additional models for getting polynomial features and interactions to see if we could improve the results. As you can see from the image below we do not gain very much improvement over the metrics upon using more complex models. So, we strongly conclude that the feature set is not enough to be able to predict the price of the house.

- **Polynomial features:**
  - **RSS**: 6307179433659.83
  - **RMSE**: 34810.25
  - **R-Squared**: 0.77
- **Interactions:**
  - **RSS**: 6430645388941.551
  - **RMSE**: 35149.31
  - **R-Squared**: 0.76

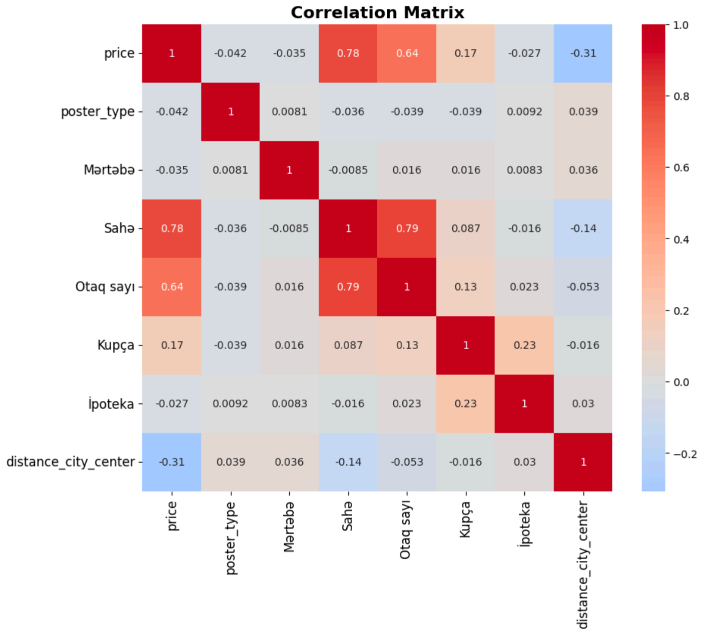


Figure 6. Pearson correlation between features

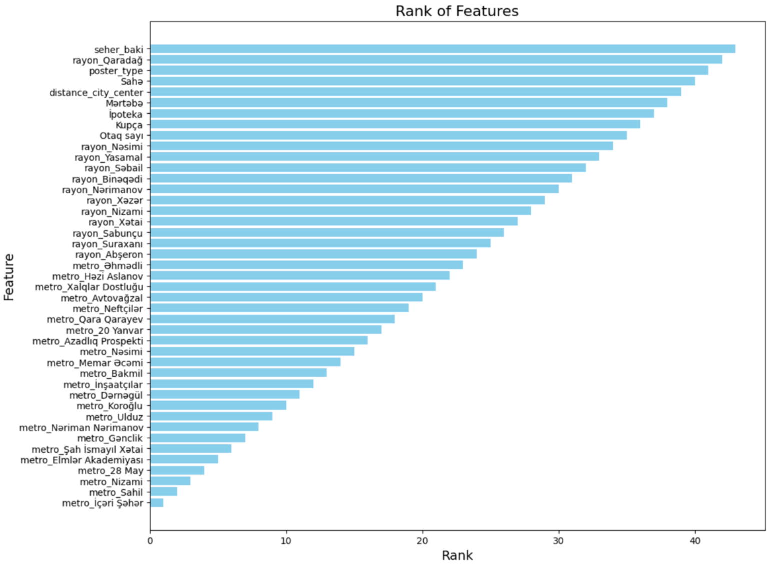


Figure 7. Feature Selection using backward elimination