Final Project Interim Report №2

Sharak Sviatoslav
Kukurik Pavlo

April 2024

1 Objectives description

The goal of the project is to develop a predictive model that accurately estimates the sale price of used cars based on a comprehensive set of characteristics. Using historical sales data, the model aims to identify significant factors that affect the valuation of a car, such as make, model, year of manufacture, odometer reading, color, body style, transmission type, condition, and other relevant attributes. The ultimate goal is to increase market transparency, promote fair pricing, and simplify the buying and selling process.

2 Data Description

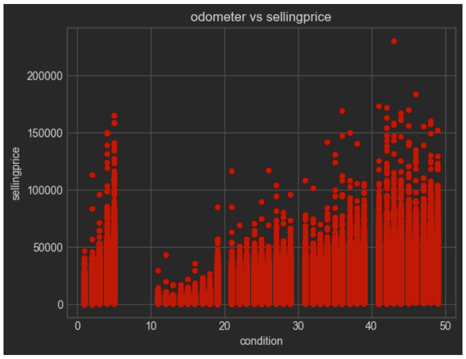
- **Size:** We have 558,837 observations (data points).
- **Features (16 total):**
  - **Quantitative Features (Features with numerical values)**

    - **year** (*int64*): The vehicle's manufacturing year.
    - **condition** (*float64*): The condition of the vehicle (likely on a numerical scale).
    - **odometer** (*float64*): The vehicle's mileage.
    - **mmr** (*float64*): The Mannheim Market Report value (estimated market value).
    - **sellingprice** (*float64*): The final price at which the vehicle was sold.
  - **Qualitative Features (Features with categorical values)**
    - **make** (*object*): The vehicle's brand/manufacturer.
    - **model** (*object*): The vehicle's specific model.
    - **trim** (*object*): Additional specification within the model.
    - **body** (*object*): The vehicle's body type (e.g., Sedan, SUV, etc.).
    - **transmission** (*object*): The type of transmission (e.g., automatic, manual).
    - **vin** (*object*): The vehicle's unique Vehicle Identification Number.
    - **state** (*object*): The state where the vehicle is registered.
    - **color** (*object*): The vehicle's exterior color.
    - **interior** (*object*): The vehicle's interior color.
    - **seller** (*object*): The entity that sold the vehicle.
    - **saledate** (*object*): The date and time of the sale.

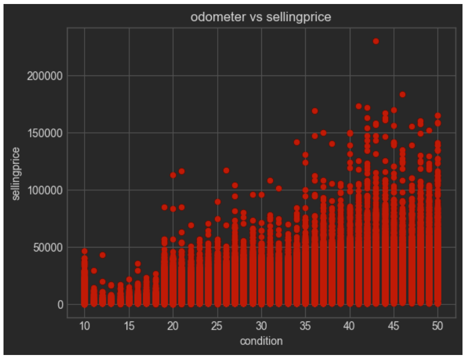
3 EDA and Preprocessing

3.1 Step №1 | Cleaning Data

First, we had to clean the data from null values, outliers, and format the data. Here are some examples:

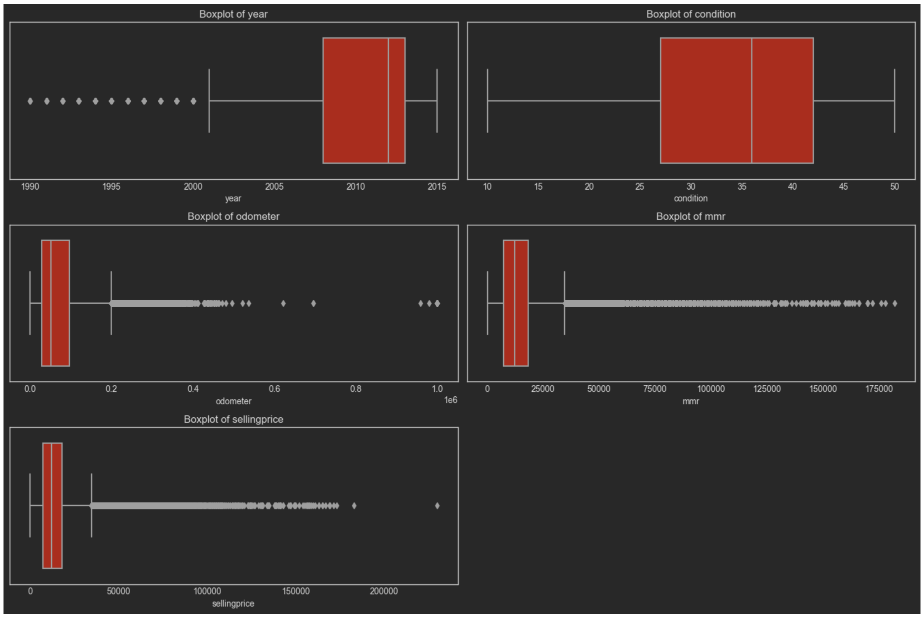


Condition before scaling

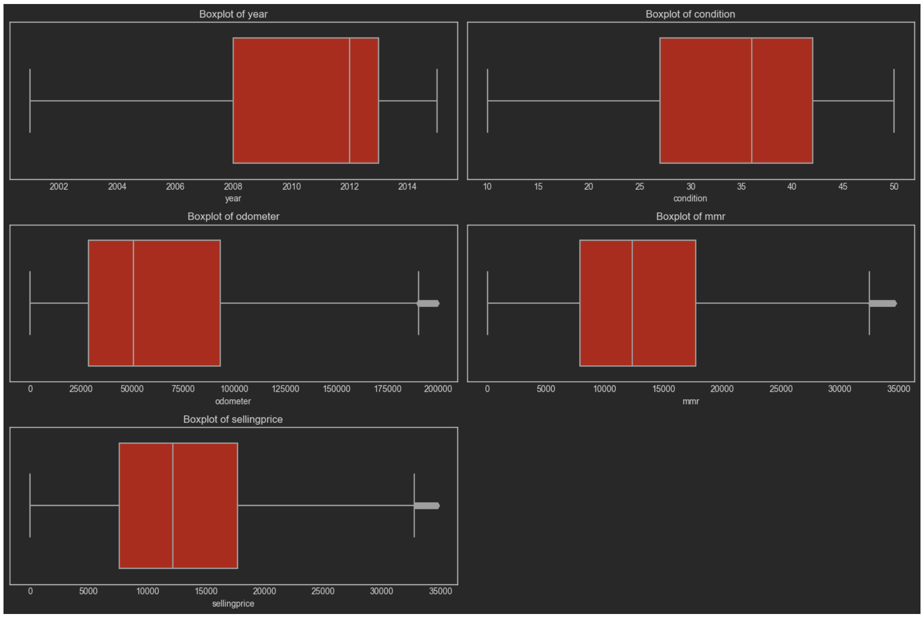


Condition after scaling

Also we cleaned data from combined values ('sedan', 'Sedan' → 'Sedan' etc.) and outliers



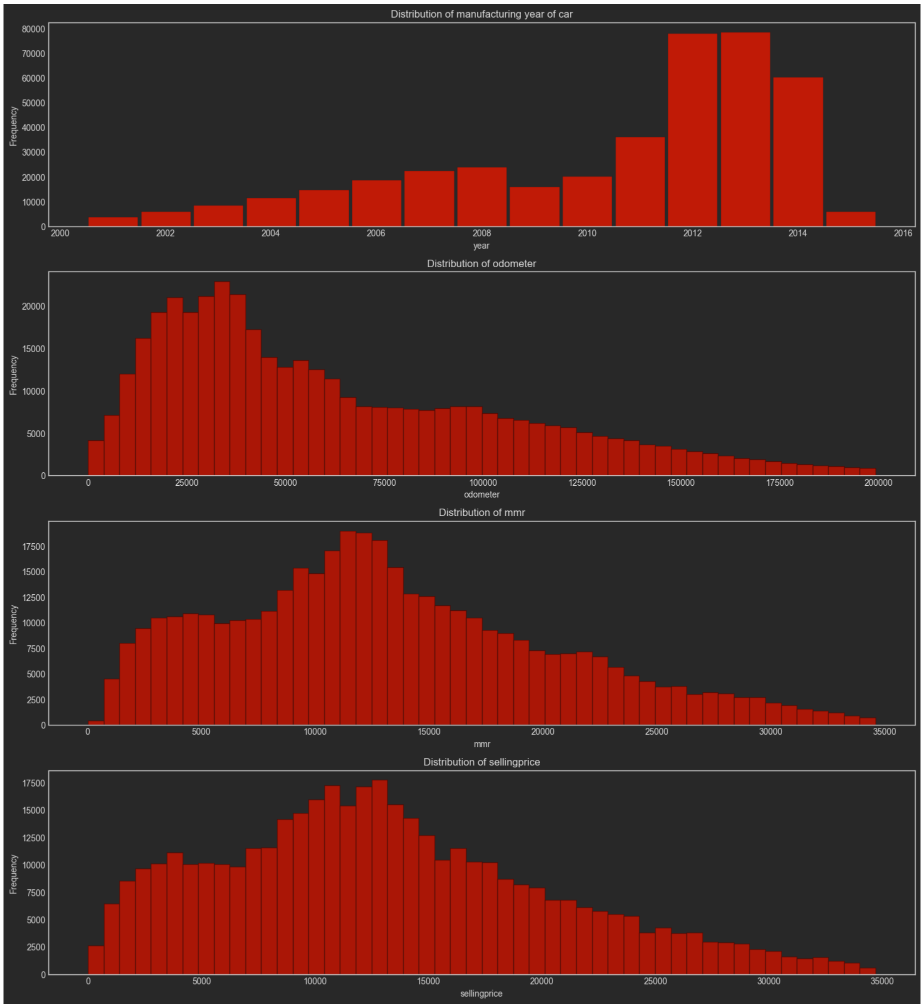
Condition before scaling



Condition after scaling

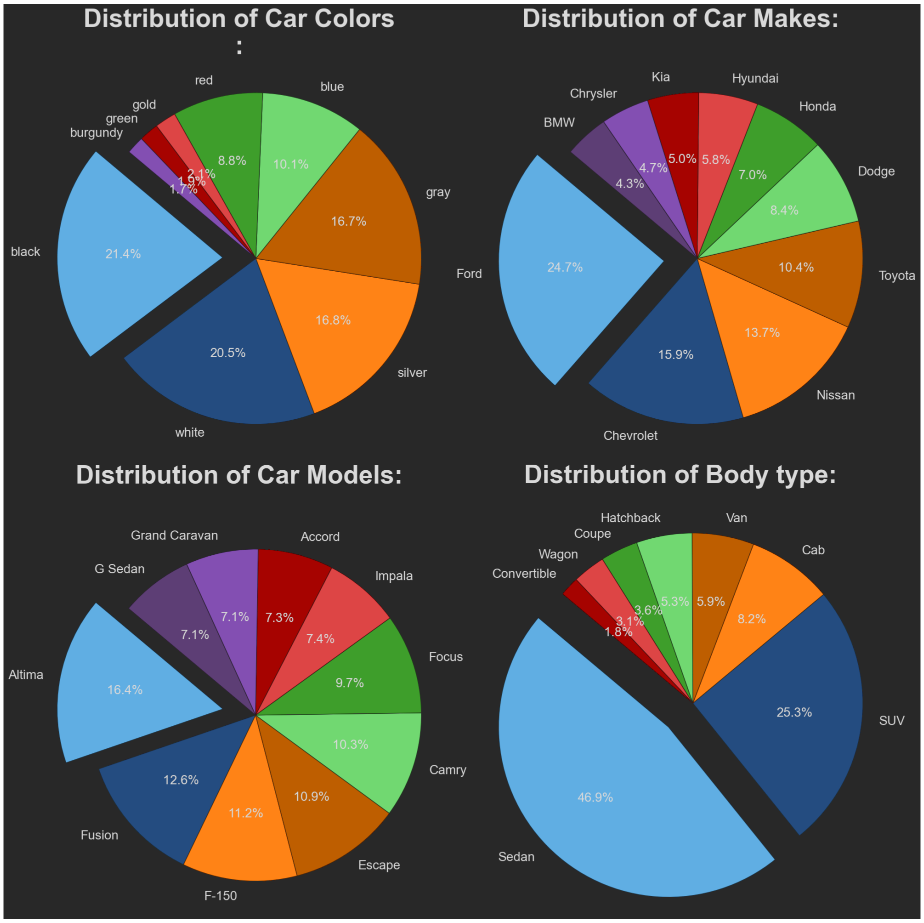
3.2 Step №2 | Visualise some Data

Also here is interesting information about our dataset like distribution of mmr and selling price etc.



Distribution of:

- Manufacturing year of car
- Miles run by car
- Conditions of car
- Selling price of car



Distribution of

Conclusions of EDA: As we can see from all information behind there the dataset includes real car sales data for 2015 of different model years. The dataset contains a lot of unclean/zero values and other things that will be harmful in the future, so we remove them as well as outliers. We also align and skew the data where it does not match the description. From the distribution in certain columns, we can see the variety of car parameters in the dataset, which allows us to make more objective assessments and conclusions when working with the data in the future.

4 Hypothesis testing

4.1 Hypothesis 1

Hypothesis 1
The transmission type (automatic or manual) significantly affects a vehicle's selling price and market demand, potentially varying by region and vehicle type.
$H_0 : \mu_{\text{automatic}} \leq \mu_{\text{manual}}$
$H_1 : \mu_{\text{automatic}} > \mu_{\text{manual}}$
Here we will use t-test, to find out is prices of car with automatic transmission is more expensive than manual

4.2 Hypothesis 2

H_0 : Vehicle preferences and market values are independent of the state of registration.
H_1 : Vehicle preferences and market values depend on the state of registration.
These hypotheses can then be tested using statistical tests for independence like the Chi-square test for categorical variables (e.g., make, model, body type vs state) and analysis of variance (ANOVA) for continuous variables

5 Methods

In this project, the Ordinary Least Squares (OLS) method is employed as the core statistical technique for constructing the regression model to predict used car selling prices. the OLS method is used to determine how well we can predict the selling price of a used car based on its attributes. After encoding the categorical variables into a format suitable for regression analysis, the OLS model helps us understand which variables are statistically significant predictors of car prices, and how changes in these variables are associated with changes in the selling price.

The OLS regression analysis's output, including the R-squared value, F-statistic, coefficients, and their p-values, helps to validate our model and informs us about the strength and nature of the relationships between the car features and their selling prices. By examining these results, we can make informed decisions about which variables to focus on in pricing strategies and can also derive insights that could lead to further questions or more in-depth analysis.

5.0.1 Pros:

1. OLS is straightforward to understand and implement, making it a popular choice for linear regression problems. It requires relatively simple calculations compared to more complex models.

2. Under certain conditions (namely, the Gauss-Markov assumptions), OLS estimators are the best linear unbiased estimators. This means that within

the class of linear and unbiased estimators, OLS provides the lowest variance.

3. The results from an OLS model are easy to interpret. Coefficients directly represent the expected change in the dependent variable for a one-unit change in an independent variable, holding all else constant.

4. OLS is a foundational tool in statistics, and its results are widely understood and trusted across various fields, making communication of findings straightforward.

5.0.2 Cons:

1. OLS requires several key assumptions (normality, homoscedasticity, independence, and no multicollinearity among predictors) for the statistical properties (like the BLUE property) to hold. If these assumptions are violated, the OLS estimates may be biased or inefficient.

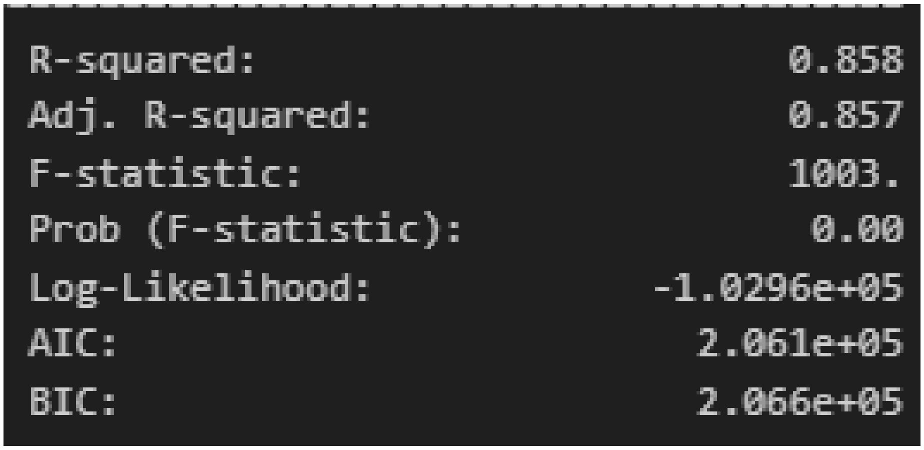
2. OLS is sensitive to outliers. Outliers can disproportionately influence the model, potentially leading to misleading results.

3. OLS can only model linear relationships unless transformations are applied to the data. This limitation can be significant if the true relationship is non-linear.

4. While multicollinearity does not bias OLS estimates, it inflates the variance of the coefficient estimates, which can make them unstable and difficult to interpret. Users must check for and address multicollinearity if present.

6 Results

For Hypothesis 1 T-Test Statistic: 34.869
P-value: approximately 0.0
Given that the t-test statistic is positive and the p-value is effectively zero. We would reject the null hypothesis in favor of the alternative hypothesis, suggesting that automatic cars are indeed more expensive than manual cars on average. During the development of our OLS model, we converted our categorical columns into dummy variables. However, we encountered an issue due to the extensive variety within certain columns, particularly 'brand' and 'model'. Given the large number of variants, we decided to simplify our approach by focusing the model on just one specific car model. We chose the Nissan Altima, as it is the most popular among our dataset. Here results of our OLS model:



```
R-squared:                     0.858
Adj. R-squared:                0.857
F-statistic:                   1003.
Prob (F-statistic):             0.00
Log-Likelihood:          -1.0296e+05
AIC:                       2.061e+05
BIC:                       2.066e+05
```

Figure 1: OLS results

7 Conclusion

In our project, we developed a predictive model designed exclusively to estimate the selling price of a specific vehicle model within a particular brand. This model has been made with a wide range of factors in mind, ensuring that it takes into account the nuances that uniquely affect the value of this particular type of vehicle. Our model allows us to estimate the market value of these vehicles, providing potential sellers and buyers with a reliable tool to determine a fair price for the vehicle in question. This model is a testament to our commitment to accuracy and specificity in car price forecasting.