**Instructions for Use**

**Insurance Claim Frequency Modeling with Poisson Regression Meltem Çap-201504056**

**1. Establishing the Necessary Environment:**

Anaconda and PyCharm Community were used to run this project.

**2-Running the Project**

Before running the project, you need to access the dataset used. The dataset is a CSV file named **car-insurance-claim.csv** and will be used in the project.

**3-Expected Results and Outputs**

After running the code, the following outputs are expected:

* Model performance metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Pseudo R² values.
* Real vs Predicted graph.
* Residuals Distribution graph.

**4-Visualizations Used in the Project**

The following visualizations were created using Matplotlib and Seaborn:

* Real vs Predicted scatter plot.
* Residuals distribution plot.
* Correlation heatmap for numerical variables.

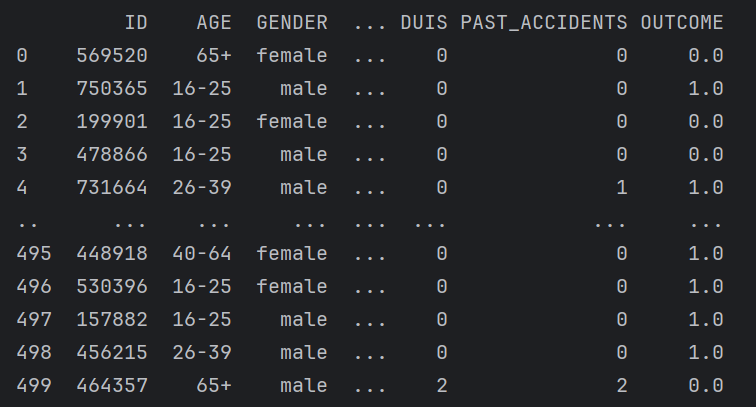
**Project Objective**The aim of the project is to develop a Poisson regression model that predicts the claim frequency of customers using vehicle insurance data. This model will help insurance companies optimize their risk assessment and pricing strategies.

**Dataset Review**  
**Dataset:** Car-Insurance\_Claim.csv

**Data Loading and Initial Exploration**

-The dataset is loaded using Pandas

-The first 500 rows are displayed for a quick review, and unique values of the RACE column are checked: data\_frame['RACE'].unique()



**Missing Data Analysis**Checking for Missing Values: The script identifies missing values using isnull() and calculates the proportion of missing data in each column.

**Filling Missing Values:**

* CREDIT\_SCORE and ANNUAL\_MILEAGE columns are filled with their respective means
* After this step, the dataset contains no missing values.

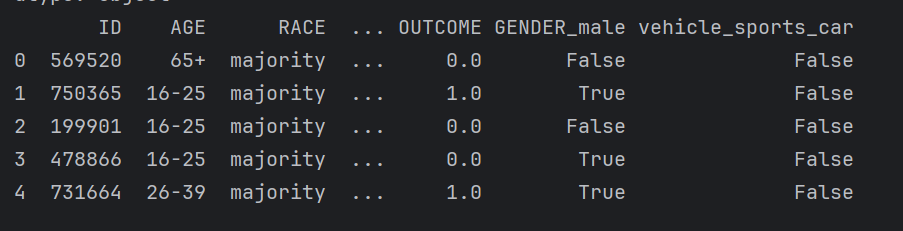
**Data Types and Transformations**

-The data types of each column are inspected to ensure compatibility with modeling steps.

Dummy Variable Encoding:

-Categorical variables (GENDER and VEHICLE\_TYPE) are converted into dummy variables using pd.get\_dummies():

-The VEHICLE\_TYPE\_sports car column is renamed to vehicle\_sports\_car for better readability.

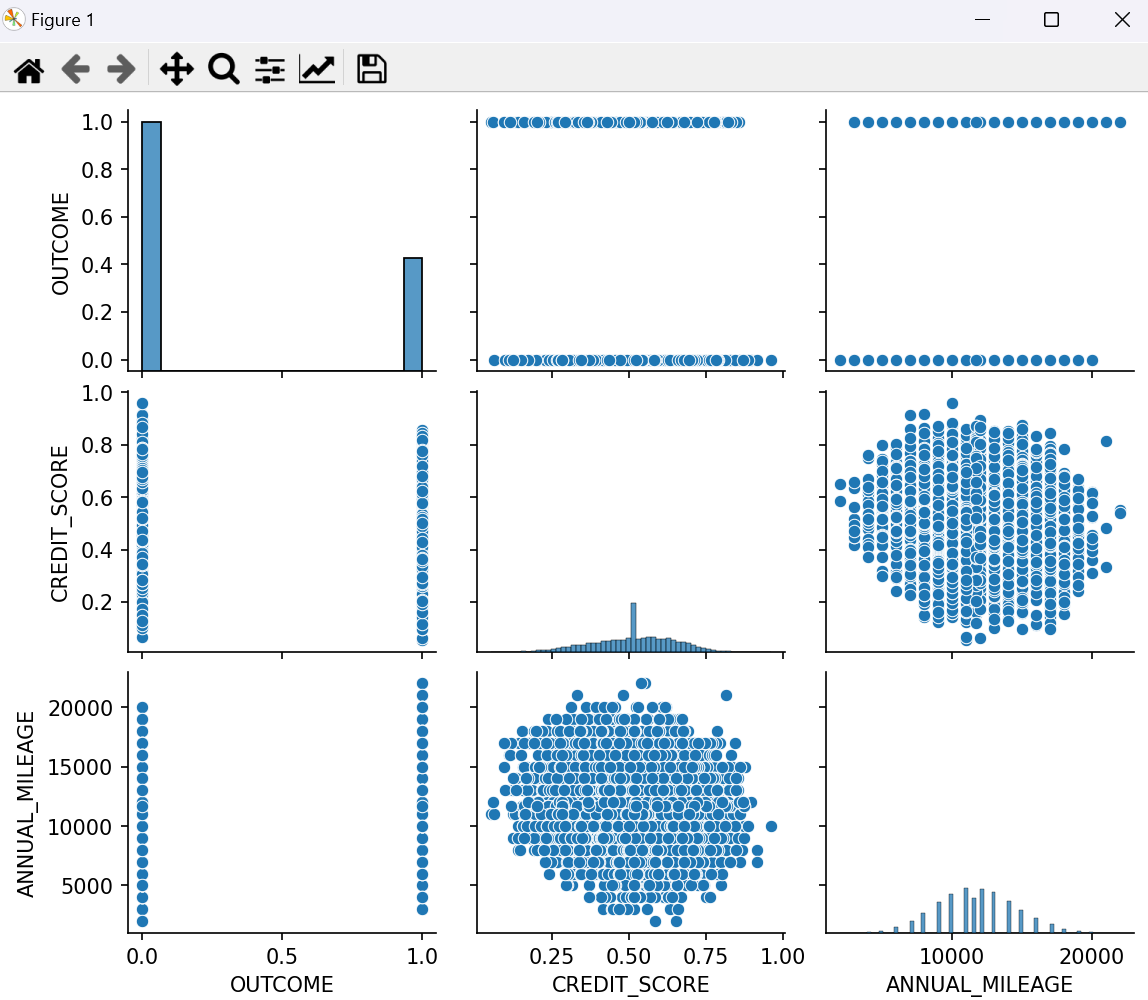


\*The dataset is now ready for analysis and modeling.

**Data Visualization**

Using seaborn and matplotlib, the script visualizes relationships and residuals:

1. Pairwise Relationships:Plots relationships between OUTCOME, CREDIT\_SCORE, and ANNUAL\_MILEAGE.
2. Residual Analysis:Distribution of residuals is analyzed with a histogram.
3. Actual vs. Predicted Values:A scatter plot compares actual OUTCOME values to model predictions.



**Train-Test Split**

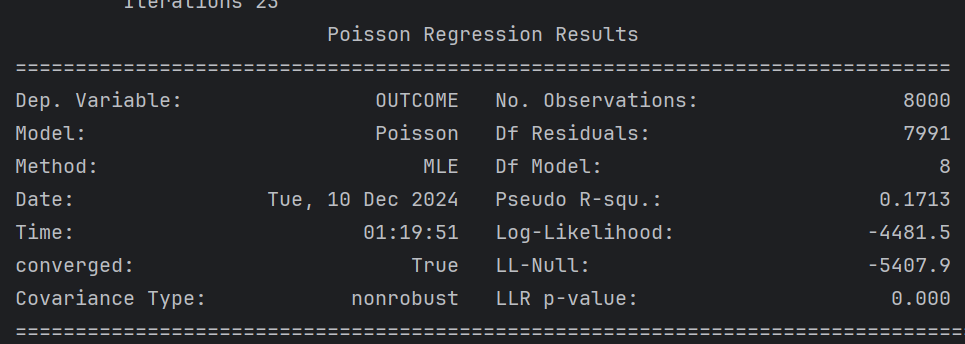
* The dataset is split into training (80%) and testing (20%) sets

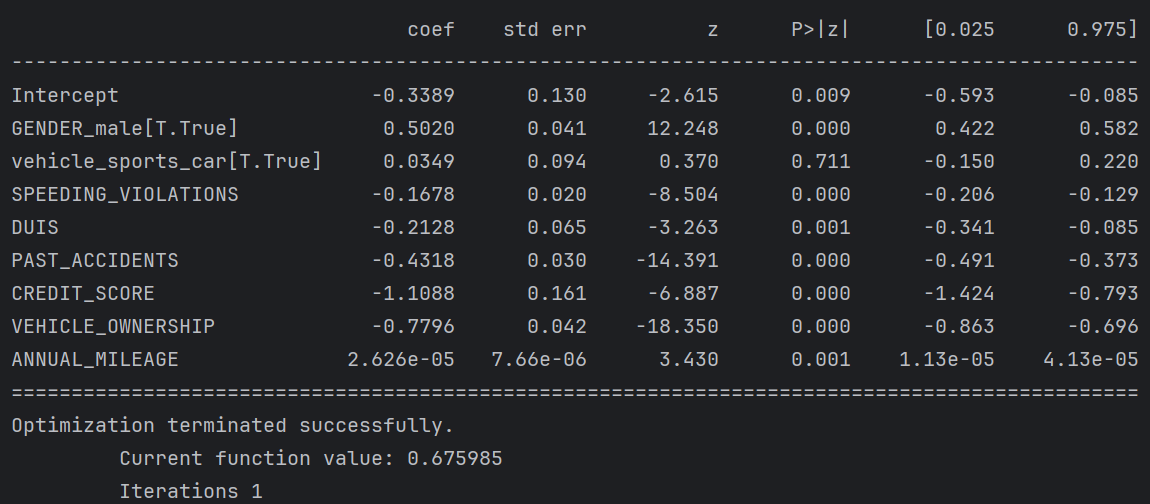
**Poisson Regression Model**

* A Poisson regression model is built to predict OUTCOME using the following formula:
* OUTCOME ~ SPEEDING\_VIOLATIONS + DUIS + PAST\_ACCIDENTS + CREDIT\_SCORE + VEHICLE\_OWNERSHIP + ANNUAL\_MILEAGE + GENDER\_male + vehicle\_sports\_car

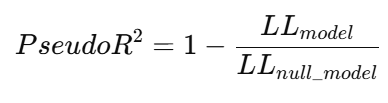
**Fitting the Model**: smf.poisson(formula=formula, data=train).fit()

**Model Summary:** The model results are displayed to analyze coefficient estimates, significance, and model fit.





**Pseudo R²:** A pseudo R² value is calculated to measure the model’s explanatory power:





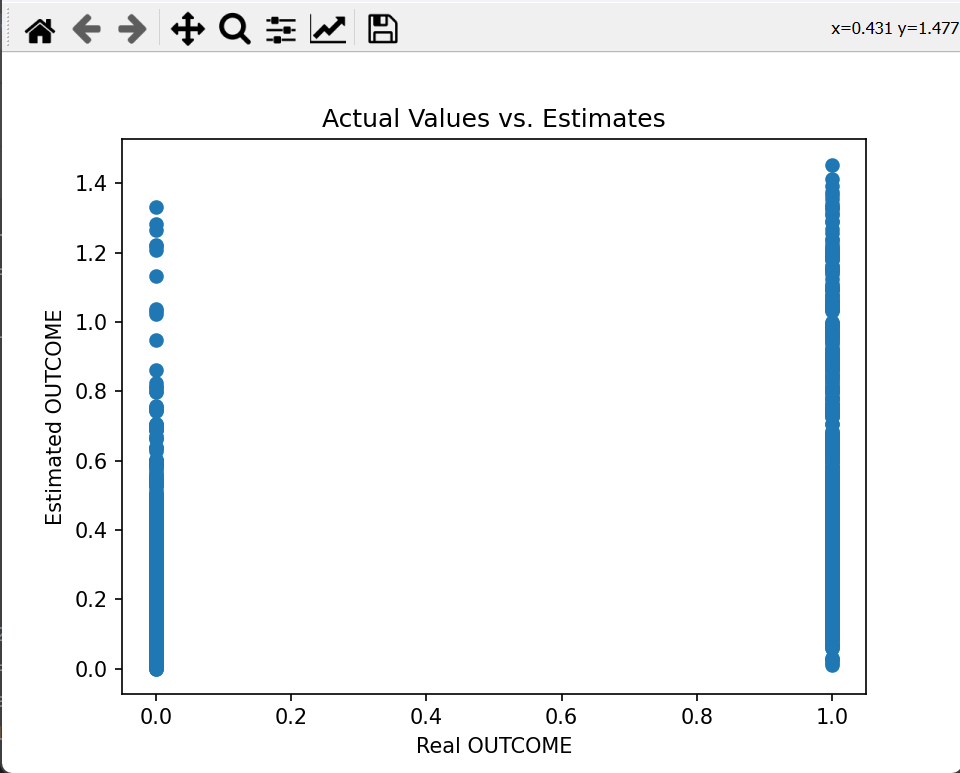
While a Pseudo R² of 0.1713 might seem low, it is not inherently "bad" in the context of Poisson regression. It indicates that the model explains some, but not all, of the variability in the outcome.

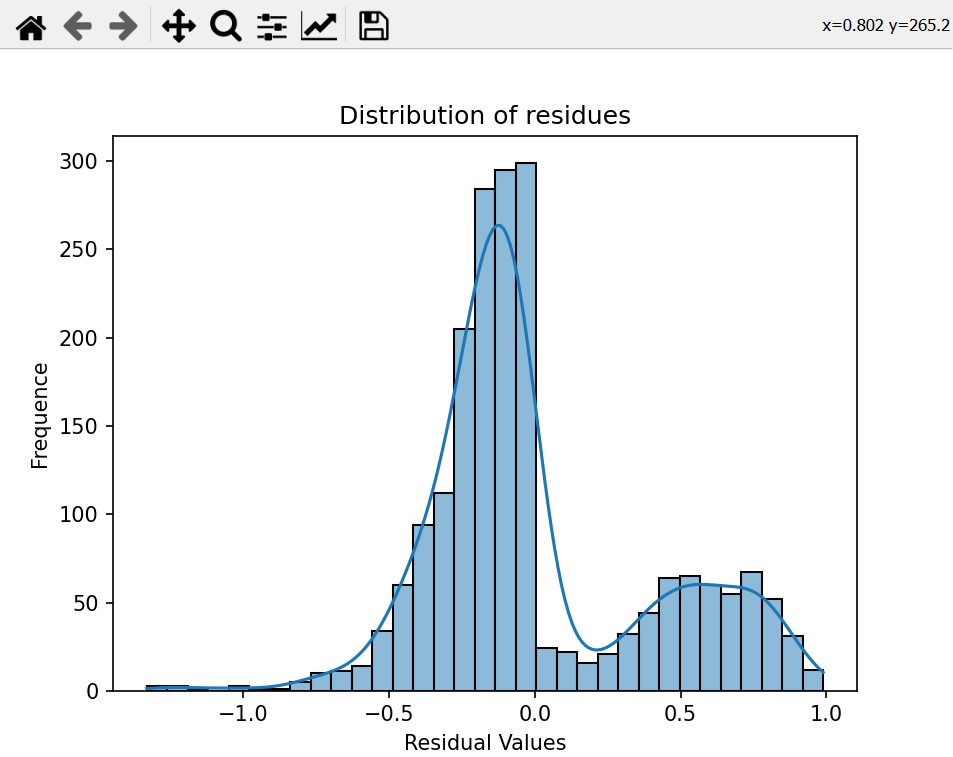
**Predictions and Model Diagnostics**

Predictions for OUTCOME are made on the test dataset: test['predicted\_OUTCOME'] = results\_train.predict(test)

Residual Analysis:Residuals (difference between actual and predicted values) are plotted.

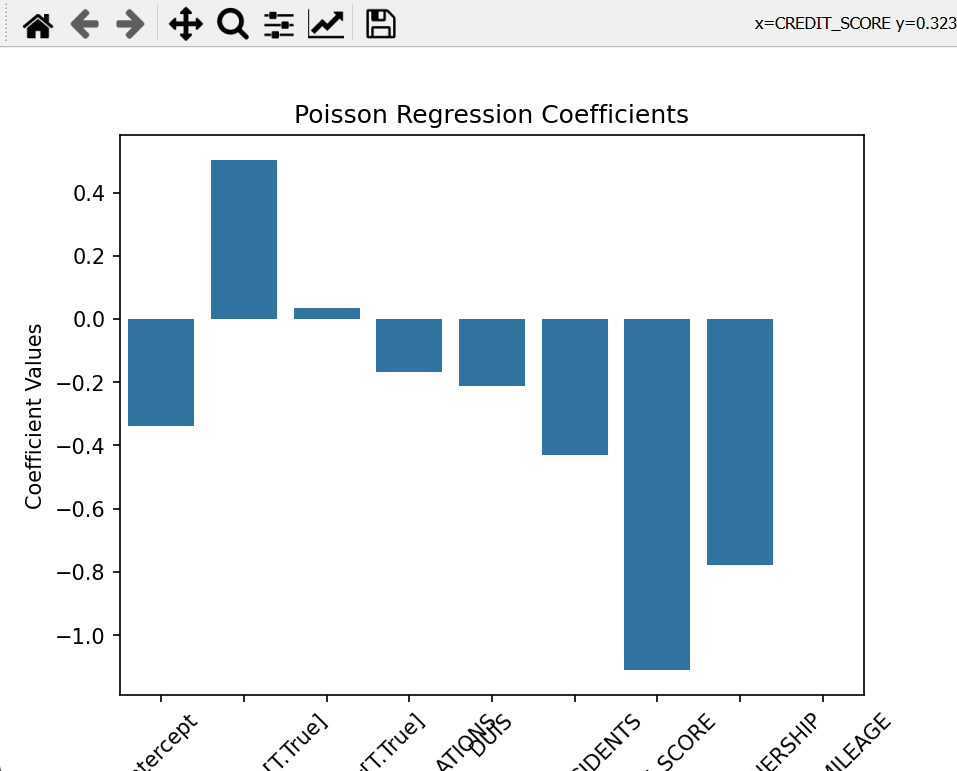
Actual vs. Predicted Values:Scatter plot visualizes the relationship between actual outcomes and predictions.





**Coefficients and Visualization**

The coefficients of the model are extracted and visualized:

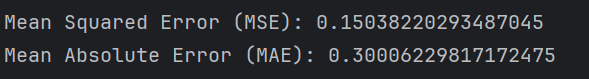
* Positive coefficients indicate a positive effect on the likelihood of claims.
* Negative coefficients indicate a negative effect.
* A bar plot is used to display coefficient values.
* 

**Model Performance Metrics**

The model's performance is evaluated using two metrics:

1. **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values.
2. **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual values.

Lower values for both metrics indicate better model performance.



The errors, MSE: 0.1504 and MAE: 0.3001, indicate reasonable model performance, assuming the scale of OUTCOME is small.

**Overdispersion Check**

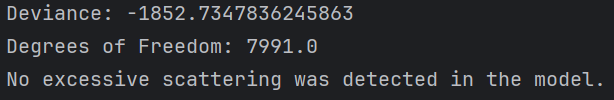
Poisson regression assumes that the mean and variance of the dependent variable are equal. To check for overdispersion:

**Deviance:**  
Deviance measures the difference between the fitted model and a saturated model:



**Degrees of Freedom:**  
The degrees of freedom are compared to the deviance.

If the deviance exceeds the degrees of freedom, overdispersion is present



The Deviance/Degrees of Freedom Ratio is approximately −0.23. A much smaller ratio (closer to 0) indicates underdispersion. In this case, the model may be considered to be making incorrect or incomplete predictions. However, it does indicate that there is no overdispersion and that the model generally fits well. This shows that the model accurately reflects the variability of the data and does not make excessive errors. That is, there is not a lot of unnecessary variability in the model.

**Results:**

There is no overdispersion in the model, but underdispersion may be observed. There may be no over-scattering or insufficient scattering in the model. However, since there is no over-scattering, my model generally appears to fit the data.

In the insurance sector, the use of models such as Poisson regression allows insurance companies to make more accurate risk analysis. The model can be a great help in determining insurance premiums and estimating damage claims. However, more accurate results can be obtained by improving the model, using additional variables and alternative models. In addition, insurance companies can identify high-risk customer groups and develop special strategies for these groups.