Car Price Case

Executive Summary

The report will indicate which factors influence the pricing of used cars by building a regression model to comprehend and test for model assumptions while discussing model selection and answering the main 3 questions in the report; building a model, predicating and lastly understanding if the price will appreciate or depreciate.

The Data

- Price The price of the car, in thousands of euros;
- Age The age of the car, in months;
- Odometer The odometer reading of the car (how many kilometers it has been driven), in thousands of kilometers;
- Inspection The scheduled time for the next inspection, in months;
- ABS An indicator variable showing if the car has anti-lock braking system (ABS) brakes (1: car has it, 0: the car does not have it);
- Sunroof An indicator variable showing if the car has a sunroof (1: car has it, 0: the car does not have it);
- Annualkms The odometer reading, divided by the age of the car, in thousands of kilometers per year. The data are presented in the .csv file usedcars.csv.

Data Model Approach

The project will require a multiple linear regression model to analyze how the continuous variable price (the dependent or response variable) is affected by the other independent (explanatory) variables. Since there is more than one independent variable, a multiple linear regression model is appropriate for answering this question. The case includes seven variables, most of which are numeric, with only two being categorical.

Explanatory Data Analysis

The variance of the variables indicates that there are four variables that need to be scaled; otherwise, the model may produce erroneous results. These four variables—Age, Inspection, Odometer, and Annualkms—are considered explanatory variables. Upon checking their distributions, Age and Inspection are skewed to the right, while Odometer and Annualkms appear more normally distributed. However, all mentioned variables require transformation to ensure accurate analysis. The response variable, Price, is skewed to the left, with a large number of values falling within the 2–4 thousand range. This strong left skew suggests that transformation is necessary. A boxplot was used after transforming the numeric variables to visualize outliers that fall outside the interquartile range (Q1 to Q3). However, the model will not omit any outliers at this stage—this step is for awareness only according to figure 1.

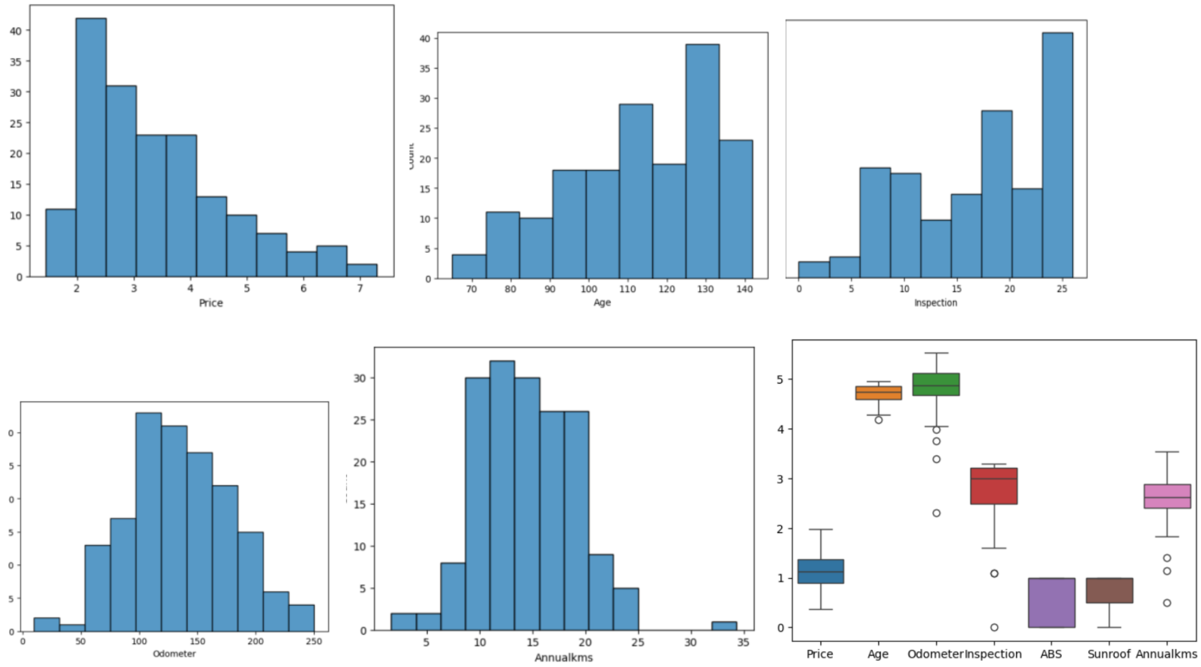
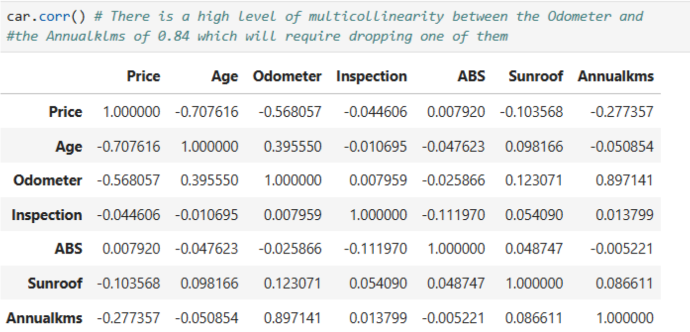


figure 1. boxplot and histplots

Using the correlation function provides insights in two key ways. First, it reveals a high degree of multicollinearity between the Odometer and Annualkms variables, with a correlation coefficient of 0.89, which is considerably high. Second, there is a strong negative correlation of -0.70 between Price and Age, which will help in explaining the effect of Age on Price later in the analysis in line with figure 2.



```
car.corr() # There is a high level of multicollinearity between the Odometer and
#the Annualklms of 0.84 which will require dropping one of them
```

	Price	Age	Odometer	Inspection	ABS	Sunroof	Annualkms
Price	1.000000	-0.707616	-0.568057	-0.044606	0.007920	-0.103568	-0.277357
Age	-0.707616	1.000000	0.395550	-0.010695	-0.047623	0.098166	-0.050854
Odometer	-0.568057	0.395550	1.000000	0.007959	-0.025866	0.123071	0.897141
Inspection	-0.044606	-0.010695	0.007959	1.000000	-0.111970	0.054090	0.013799
ABS	0.007920	-0.047623	-0.025866	-0.111970	1.000000	0.048747	-0.005221
Sunroof	-0.103568	0.098166	0.123071	0.054090	0.048747	1.000000	0.086611
Annualkms	-0.277357	-0.050854	0.897141	0.013799	-0.005221	0.086611	1.000000

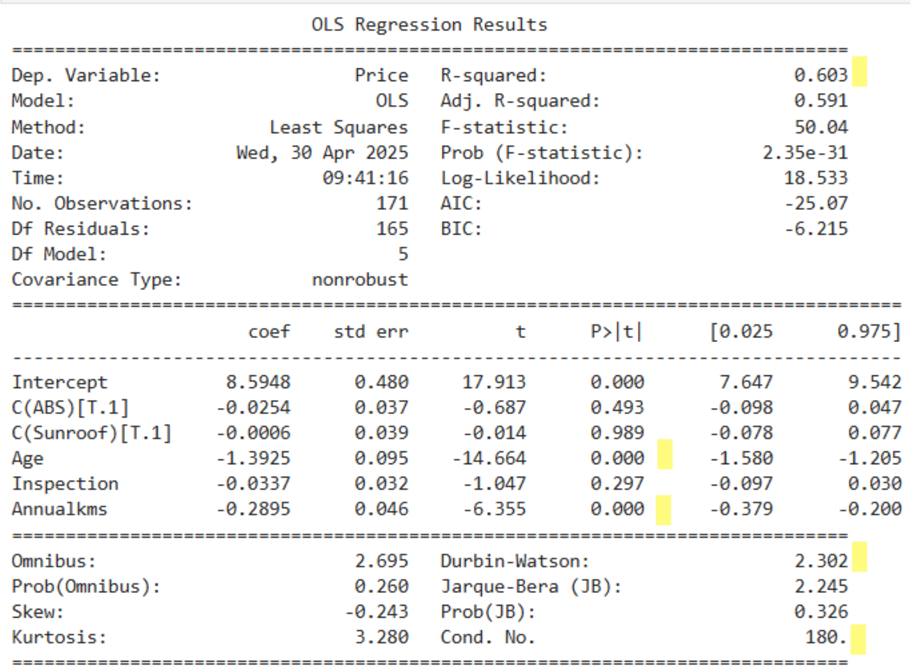
Figure 2. Correlation analysis of variables in the dataset

Model Selection

Building the model reveals something odd. There is notably high multicollinearity, which requires further investigation. A decision has been made to omit the Odometer variable due to its high correlation with Annualkms (0.90), as identified by the correlation function. This level of multicollinearity can distort the model's estimates. Additionally, Age shows a strong negative correlation with Price (-0.70), and Annualkms also has a negative correlation with Price (-0.28). These relationships will be useful in explaining the effects of these variables on Price later in the analysis. The regression model equation aligned with figure 3:

$$\log(\text{Price}_i) = \beta_0 + \beta_1 \log(\text{Age}_i) + \beta_2 \log(\text{Inspection}_i) + \beta_3 \text{ABS}_i + \beta_4 \text{Sunroof}_i + \beta_5 \log(\text{Annualkms}_i) + \varepsilon_i$$

Where the last part is the individual errors that is normally distributed



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.603
Model:                            OLS   Adj. R-squared:                  0.591
Method:                 Least Squares   F-statistic:                     50.04
Date:                Wed, 30 Apr 2025   Prob (F-statistic):           2.35e-31
Time:                        09:41:16   Log-Likelihood:                 18.533
No. Observations:                 171   AIC:                            -25.07
Df Residuals:                     165   BIC:                            -6.215
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          8.5948      0.480     17.913      0.000       7.647       9.542
C(ABS)[T.1]       -0.0254      0.037     -0.687      0.493      -0.098       0.047
C(Sunroof)[T.1]   -0.0006      0.039     -0.014      0.989      -0.078       0.077
Age               -1.3925      0.095    -14.664      0.000      -1.580      -1.205
Inspection        -0.0337      0.032     -1.047      0.297      -0.097       0.030
Annualkms         -0.2895      0.046     -6.355      0.000      -0.379      -0.200
==============================================================================
Omnibus:                        2.695   Durbin-Watson:                   2.302
Prob(Omnibus):                  0.260   Jarque-Bera (JB):                2.245
Skew:                          -0.243   Prob(JB):                        0.326
Kurtosis:                       3.280   Cond. No.                         180.
==============================================================================
```

Figure 3. Model summary after transforming the variables

Model assumptions

The linearity assumption of the residuals is met, as the majority of data points lie along the regression line in the Q-Q plot, which is a positive indication. Also, the distribution of the residuals largely follows a normal distribution, with some exceptions where residual values deviate slightly. The homoscedasticity assumption appears to be satisfied, as the residuals are randomly scattered without any clear pattern, indicating constant variance. Additionally, in terms of independence of observations, the Durbin-Watson statistic is 2.3, which is within the acceptable range and indicates no significant autocorrelation. However, multicollinearity among the independent variables is a concern. There is a very high correlation of 0.9 between *Odometer* and *Annualkms*, which suggests redundancy, but it has been omitted. The leverage plot reveals three observations 24, 62, and 78 — that exhibit high leverage and may unduly influence the model. Additionally, these points appear to be strong potential outliers, and their impact on the model should be carefully evaluated, possibly through influence diagnostics or by testing the model with and without them aligned with figure 4.

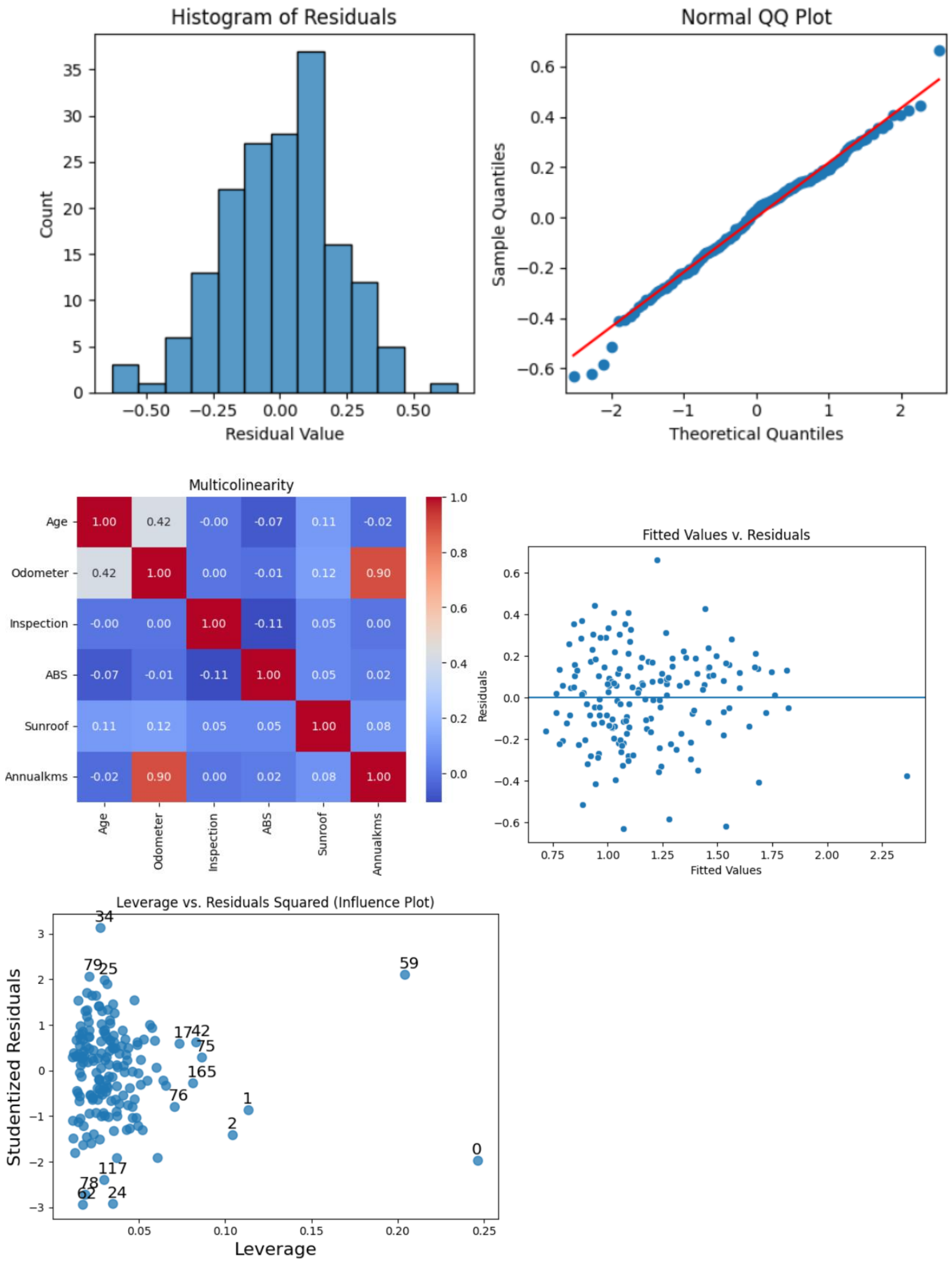


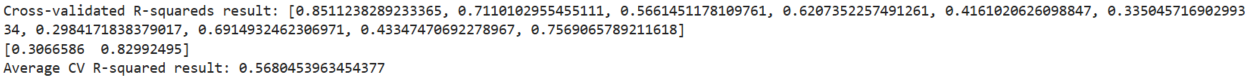
Figure 4. Model assumptions

Results

Answering the question about how the variation is explained by the model, based on the chosen model, cars with a Sunroof, cars that have ABS, and cars with Inspection are not statistically significant, as all of them have a p-value higher than 0.05, even after transformations. On the other hand, the Age variable and Annualkms are both statistically significant, with p-values of 0.00, which are excellent results. However, after omitting the Odometer variable, Annualkms went from 0.03 to 0 in the p statistical value, minimizing the multicollinearity between the variables. While a 1% increase in the age of a car will lead to a 1.39% decrease in price with confidence interval for the Age variable is −1.580 to −1.205. This aligns with the strong negative correlation between them based on the log-to-log analysis. Meanwhile, the Annualkms variable is statistically significant, and it indicates that a 1% increase in the Annualkms of a car will lead to a 0.29% decrease in price. This also reflects a negative correlation based on log-to-log analysis, with a confidence interval of −0.379 to −0.200.

Sensitivity Analysis

Managing cross-validation was challenging due to the complexity introduced by the log transformation of the target variable. This complexity led to the decision to use 10-fold cross-validation and calculate the average performance across folds. The model is not overfitting. The training R-squared was 0.67, demonstrating that the model explains 67% of the variance in the training data. K-fold cross-validation was applied to evaluate the model’s performance on unseen data, yielding a test R-squared of 0.59, which indicates the model explains 59% of the variance in previously unseen observations. To further assess the robustness of the model, 10-fold cross-validation was performed with replacement and a fixed random seed of 42. This resulted in an average R-squared of 0.57, with individual fold scores ranging from 0.31 to 0.84. These results suggest that the model is capable of predicting car prices with relatively low variance in performance, even when applied to unseen datasets with a small drop of 0.08 between the training and testing data.



```
Cross-validated R-squareds result: [0.8511238289233365, 0.7110102955455111, 0.5661451178109761, 0.6207352257491261, 0.4161020626098847, 0.335045716902993
34, 0.2984171838379017, 0.6914932462306971, 0.43347470692278967, 0.7569065789211618]
[0.3066586  0.82992495]
Average CV R-squared result: 0.5680453963454377
```

Figure 5. Cross validation k-fold method

Discussion

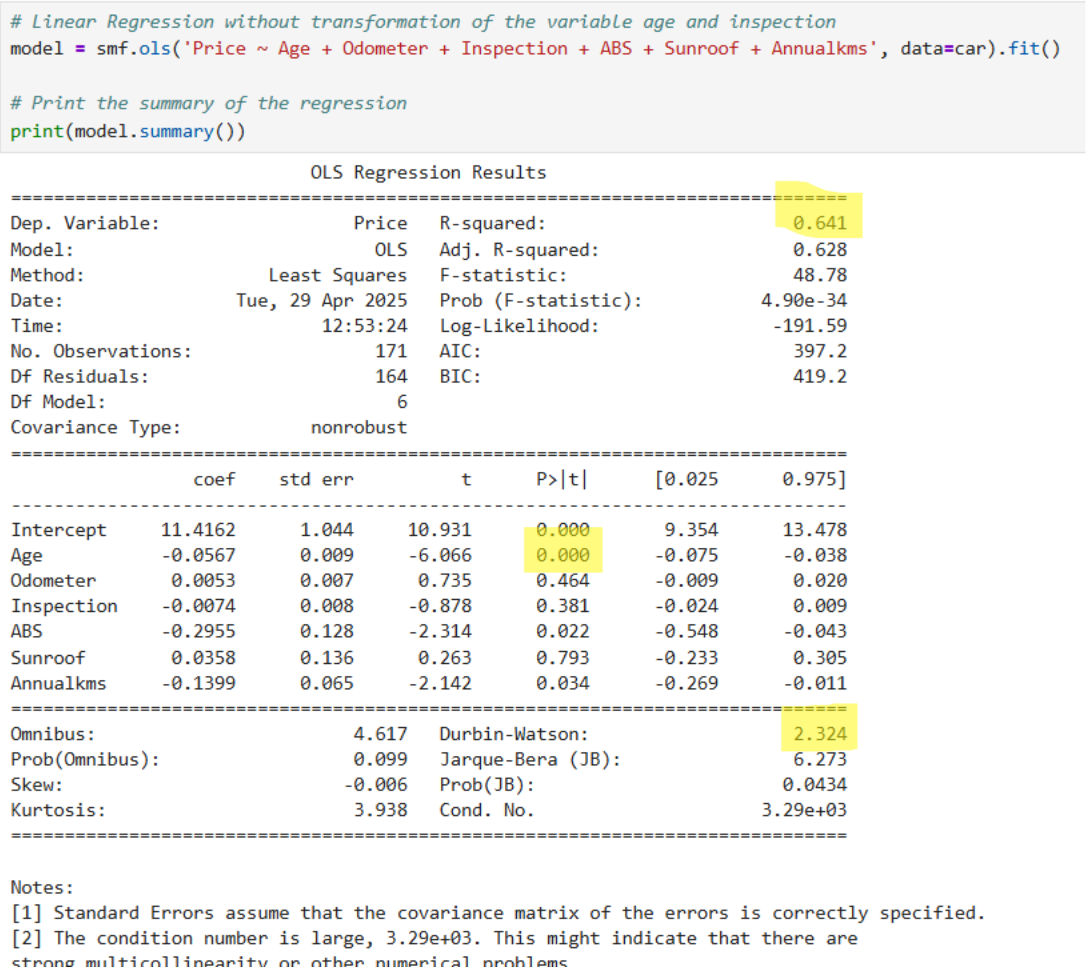
The model was trained using log transformation, so when making predictions, applying the exponential (exp) function is necessary to revert the predicted values back to the original price scale. Predicting the price without using the exp function would yield incorrect results, as the predictions are based on the transformed (log) model. The predicted value was approximately €3,012, which falls within the expected price range and the mean is €3,407 for the price variable. It's important to note that predicting the exact price of a specific car was not feasible, especially since the Odometer variable was removed due to multicollinearity. However, in my defense, I used the available information to calculate the Annualkms variable (11.5), which was included in the model even though it was not directly provided as part of the assumptions.

The final question explores a scenario where the Age of the car increases by 2 years and Annualkms rises to 40 (thousands of kilometers per year). In this case, the predicted price drops by approximately €1,376, which is considered a significant depreciation. The car would lose about 33.2% of its value, or roughly 26% per year according to Arnold Clark website "Year one: 15%-35% depreciation, leaving the car holding 65%-85% of its original value". While it is logical for the car's price to decrease over time due to age and mileage, which is reasonable in real life, the model explains only 65% of the variance. After omitting the extreme variable, the R-squared value increases by 5%. However, there are still other unexplained variables that could influence the model according to the same website; There are other variables that can influence the price of a car, popularity, brand and fuel economy. The last part of question 3 is harder to interpret, as ABS was not statistically significant in the model. However, the model still predicts an effect: switching the ABS label from 0 to 1 result in a €46 decrease in the predicted price. While this difference is relatively small, having ABS may still be considered beneficial from a safety perspective.

Extreme outliers had a significant negative impact on the model, reducing cross-validation performance by nearly 20%. While one could argue that additional outliers should be removed to further improve the model, I attempted removing some of the potential outliers, which led to only a very small improvement in performance. This suggests that the most influential outliers—identified through leverage analysis—were likely the ones hindering the model. Since linear regression is highly sensitive to outliers, which can skew both the mean and standard deviation, these extreme values had a disproportionate influence on the model's accuracy.

Appendix

Figure 1.1 The model with Odometer variable has a very high level of multicollinearity that will affect the model performance drastically.



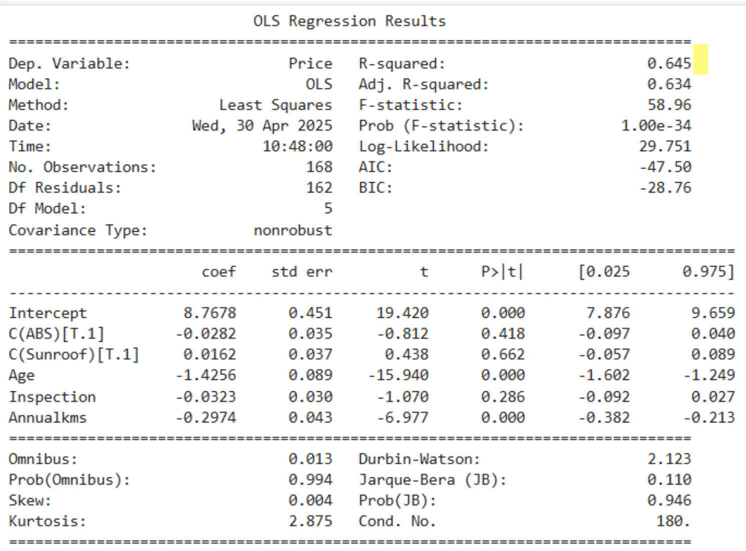
```
# Linear Regression without transformation of the variable age and inspection
model = smf.ols('Price ~ Age + Odometer + Inspection + ABS + Sunroof + Annualkms', data=car).fit()

# Print the summary of the regression
print(model.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.641
Model:                            OLS   Adj. R-squared:                  0.628
Method:                 Least Squares   F-statistic:                     48.78
Date:                Tue, 29 Apr 2025   Prob (F-statistic):           4.90e-34
Time:                        12:53:24   Log-Likelihood:                -191.59
No. Observations:                 171   AIC:                             397.2
Df Residuals:                     164   BIC:                             419.2
Df Model:                           6                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     11.4162      1.044     10.931      0.000       9.354      13.478
Age           -0.0567      0.009     -6.066      0.000      -0.075      -0.038
Odometer       0.0053      0.007      0.735      0.464      -0.009       0.020
Inspection    -0.0074      0.008     -0.878      0.381      -0.024       0.009
ABS           -0.2955      0.128     -2.314      0.022      -0.548      -0.043
Sunroof        0.0358      0.136      0.263      0.793      -0.233       0.305
Annualkms     -0.1399      0.065     -2.142      0.034      -0.269      -0.011
==============================================================================
Omnibus:                        4.617   Durbin-Watson:                   2.324
Prob(Omnibus):                  0.099   Jarque-Bera (JB):                6.273
Skew:                          -0.006   Prob(JB):                       0.0434
Kurtosis:                       3.938   Cond. No.                     3.29e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.29e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 1.2 After omitting the outliers, the model performance increase by 5%



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.645
Model:                            OLS   Adj. R-squared:                  0.634
Method:                 Least Squares   F-statistic:                     58.96
Date:                Wed, 30 Apr 2025   Prob (F-statistic):           1.00e-34
Time:                        10:48:00   Log-Likelihood:                 29.751
No. Observations:                 168   AIC:                            -47.50
Df Residuals:                     162   BIC:                            -28.76
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           8.7678      0.451     19.420      0.000       7.876       9.659
C(ABS)[T.1]        -0.0282      0.035     -0.812      0.418      -0.097       0.040
C(Sunroof)[T.1]     0.0162      0.037      0.438      0.662      -0.057       0.089
Age                -1.4256      0.089    -15.940      0.000      -1.602      -1.249
Inspection         -0.0323      0.030     -1.070      0.286      -0.092       0.027
Annualkms          -0.2974      0.043     -6.977      0.000      -0.382      -0.213
==============================================================================
Omnibus:                        0.013   Durbin-Watson:                   2.123
Prob(Omnibus):                  0.994   Jarque-Bera (JB):                0.110
Skew:                           0.004   Prob(JB):                        0.946
Kurtosis:                       2.875   Cond. No.                         180.
==============================================================================
```

Biblography

Car depreciation explained - our guide. (n.d.). https://www.arnoldclark.com/blog/selling/car-depreciation#:~:text=Average%20car%20depreciation%20per%20year%201%20Year%20one%3A,the%20car%20holding%2020%25%20of%20its%20original%20value

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, *26*(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Kotz, S., & Johnson, N. L. (1992). Breakthroughs in statistics. In Springer series in statistics. https://doi.org/10.1007/978-1-4612-0919-5

Massaron, L., & Boschetti, A. (2016). *Regression Analysis with Python*. https://openlibrary.org/books/OL26837367M/Regression_Analysis_with_Python

West RM. Best practice in statistics: The use of log transformation. Annals of Clinical Biochemistry. 2021;59(3):162-165. doi:10.1177/00045632211050531

Examples - statsmodels 0.14.4. (n.d.). https://www.statsmodels.org/stable/examples/index.html#linear-regression-models