Econometrics Assignment

Moatez Abbassi

Abstract:

* **Data Preprocessing and Feature Selection**
* **Identification of key variables: Visibility, Humidity, and Month.**
* **Individual Feature Analysis**
  + Sinusoidal transformation of Month to improve model accuracy.
  + Individual Analysis of other features
* Best model equation and validation of assumptions using the **Exhaustive Search method.**
* **Final model equation:** 
  + **Y = 26 - 0.0568(Visibility) - 19.55(Humidity) - 5.21sin(Month) - 7.72cos(Month)**
* **Model Assumptions Testing**
* **Conclusion on the Temperature Prediction Model.**

1. **Introduction**

During this assignment, I will be analyzing a Temperature dataset with the objective of making a significant and accurate model that predicts the Temperature using relevant features. Also, I will be using different regression techniques and tests that I have learned during this course.

Dataset URL: <https://www.kaggle.com/datasets/budincsevity/szeged-weather>

Notebook:<https://colab.research.google.com/drive/1VwYL4zi8TpSDIyVRp9dof2CRaedpwCyC?usp=sharing>

1. **Testing Features Individually:**

**2.1) Data Cleaning**

This is the first step in building our model, we need to ensure that our dataset is clear for easy analysis.

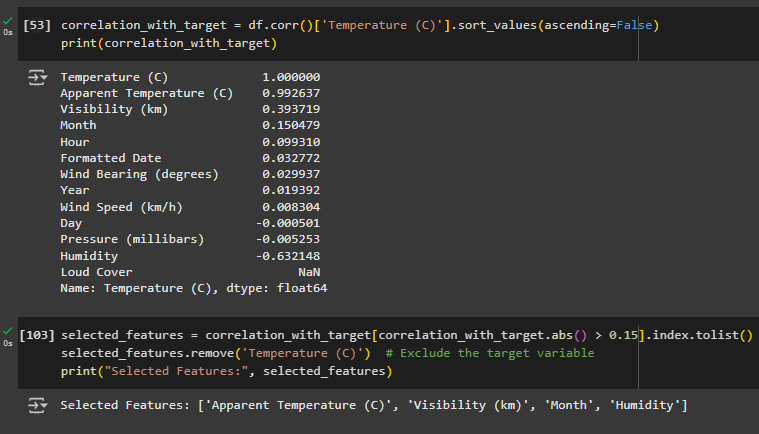
This step includes:

- Dropping Irrelevant Columns.

- Drop Null Rows.

- Fix the Formatted Date column.

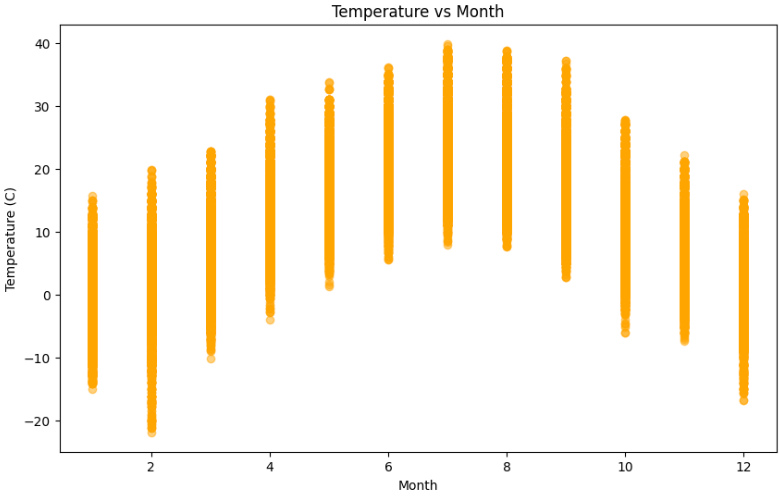
**2.2) Identifying Main Features:**



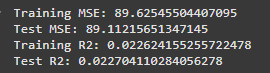
* Dependent Variable: Temperature.
* To identify possible Features for our model, we will exclude every variable that contributes by less than 0.15 to the dependent variable.
* We will be left with: **Visibility, Month, Humidity.**
* I will exclude **Apparent Temperature** because it is already a similar column to the Actual Temperature column.

**2.3) Individual Analysis of the Main Features:**

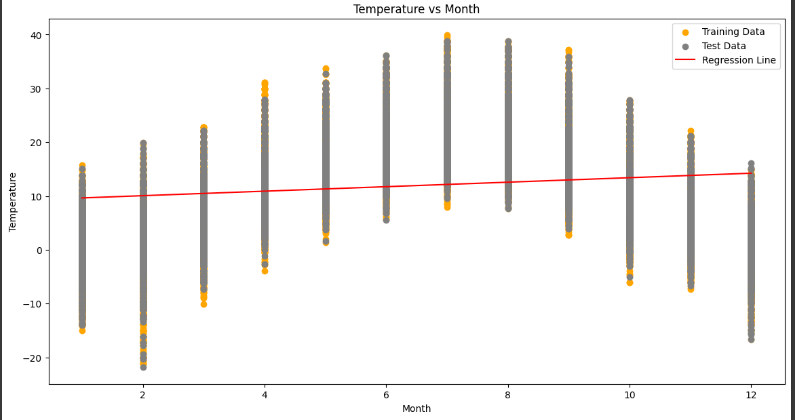
* **Month:**
  + Upon visualization, we can notice that the Month to Temperature plot shows a non-linear relationship.



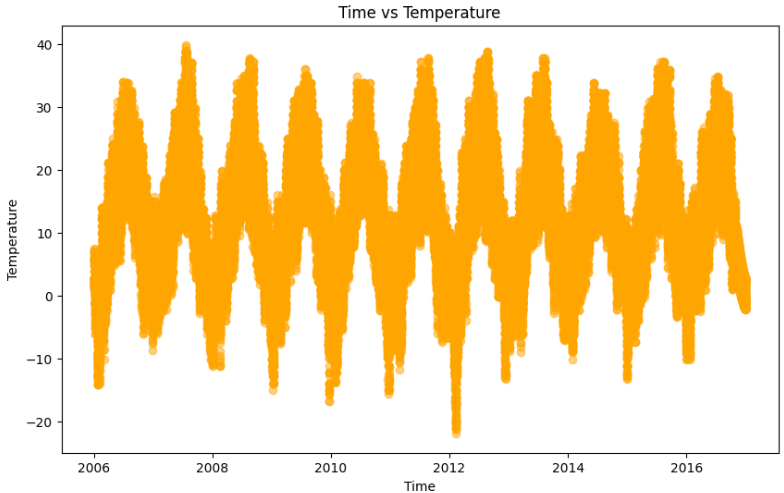
* + Before performing any tests, it looks like the Month is actually a significant independent variable, but not with a linear pattern, I believe it is either quadratic or sinusoid relationship. To identify the actual relationship, we should look at a bigger scale of time. I will first perform regular month variable analysis so I can compare it after I transform it to a sinusoid variable.
  + After making the simple linear model and training it, we obtain the following:



* + R²=0.02 in both training and testing, which is **extremely low, also the regression line looks terrible:**



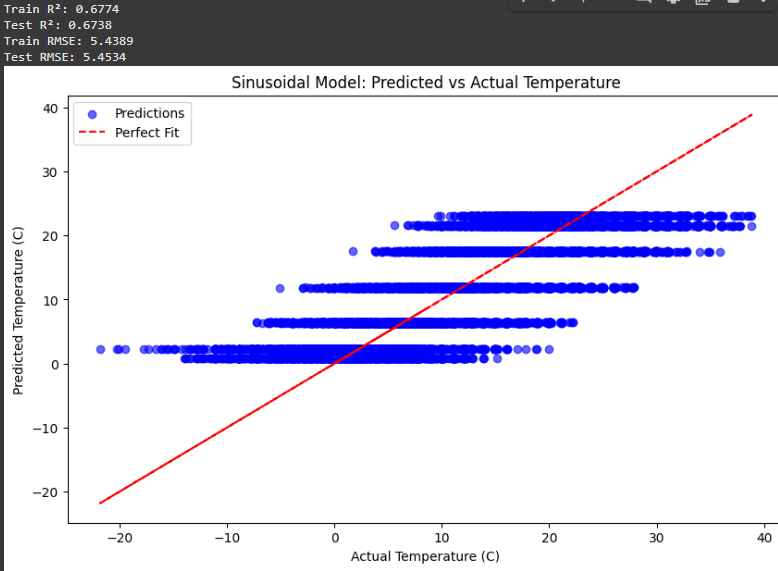
* + The prediction line is not fitting, so I will be looking for a proper transformation. In order to do that, I will look at a bigger time scale to search for a monthly pattern.



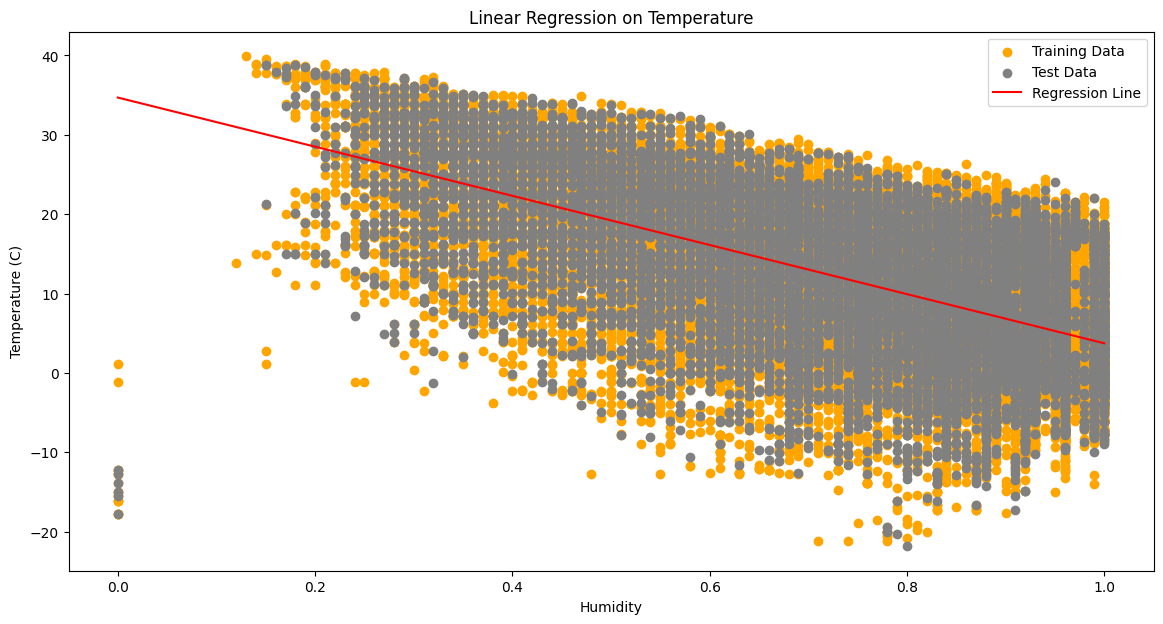
* + **It looks like a sinusoid pattern between Time and Temperature.**
  + I will **Transform the Model into Sinusoid model for a better fit** by making these 2 new columns:



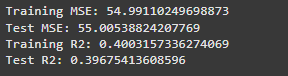
* + The new Prediction model for month:



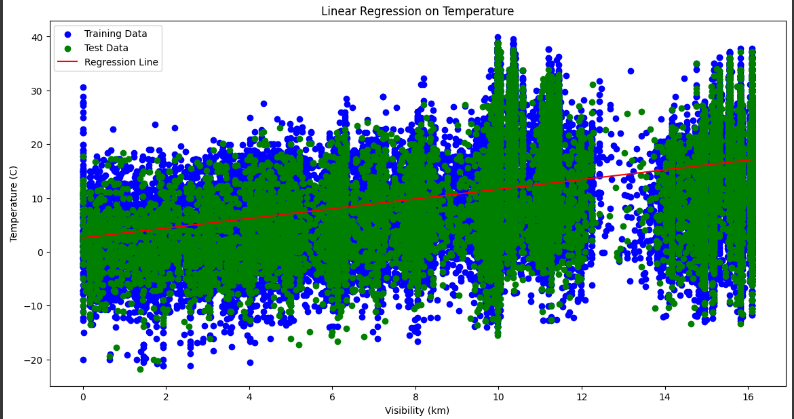
* + R² moved from 0.02 to 0.67, which **confirms that this transformation is valid.**
* **Humidity Variable:**
  + Prediction model plot:



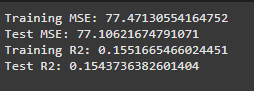
* + Model Testing:



* Humidity Variable explains approximately 40% of variance in the dependent variable in both Training and Testing, which suggests that Humidity has a good predictive power.
* **Visibility Variable:**
  + Prediction model plot:

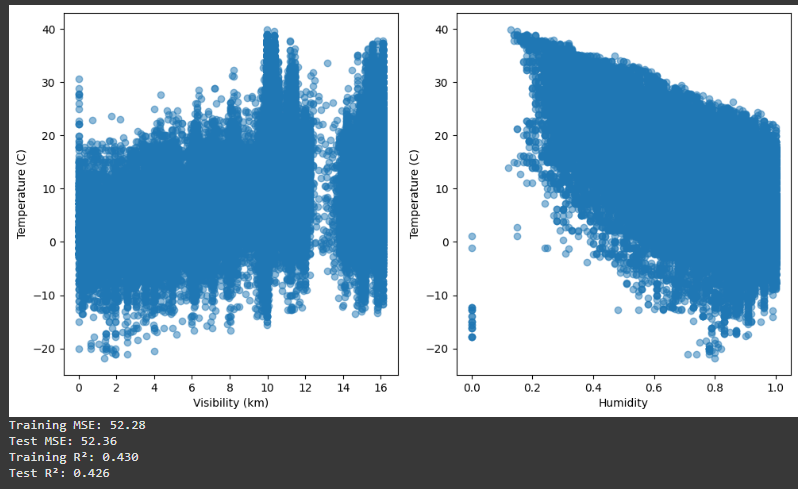


* + Model Testing:



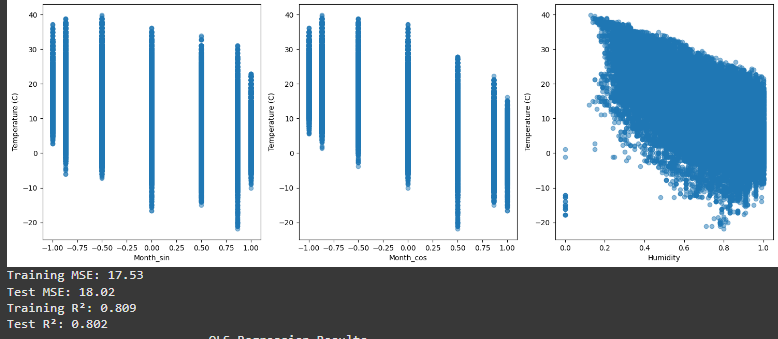
* + Visibility explain approximately 15% of variance in Temperature in both Training and Testing, which that Humidity has a weak, but reliable predictive power.

1. Choosing a Multiple Regression Model:
   * In this phase: We will use **Exhaustive Search** in this assignment to find the best multiple regression model. We test all combinations until we find the best Multiple Regression Model.
   * We ended up with 3 Independent Features: Visibility, Humidity and Month.
   * **Trial #1: Humidity + Visibility**



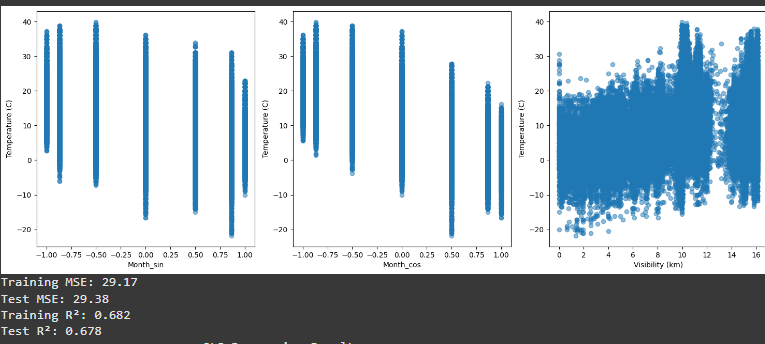
Both Variables Humidity and Visibility explain approximately 43% of variance in Temperature. It is not that significant but we may take it if it is the best model among others.

* + **Trial #2: Humidity + Month Variables:**



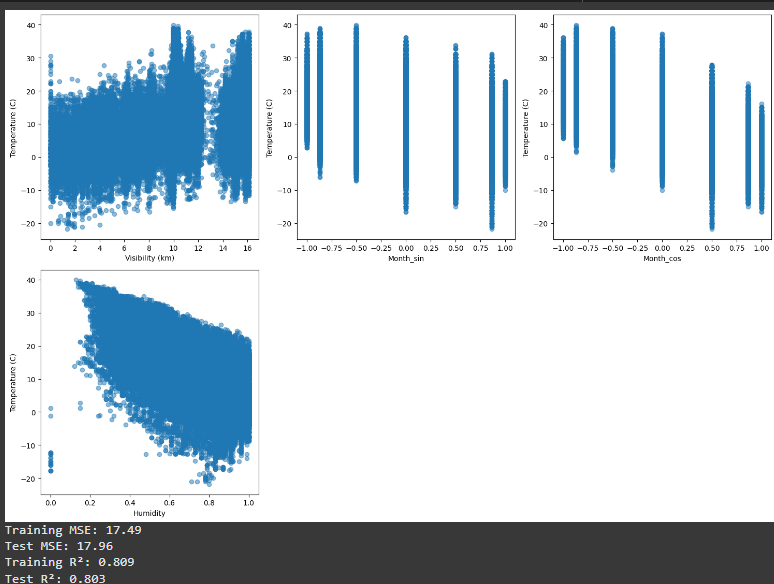
Humidity and Month variables explain approximately 80% of variance in Temperature, which is much larger than Trial #1’s model. It is also significant and reliable.

* + **Trial #3: Visibility + Month Variables:**



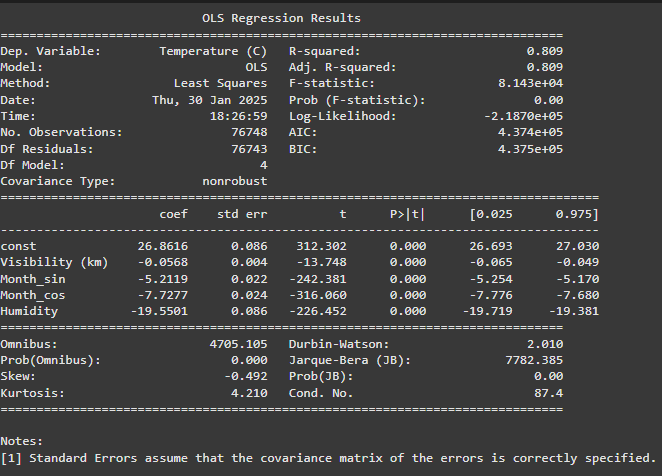
Visibility and Month Variables explain approximately 68% of variance in Temperature. It is actually significant and reliable but it is worse than the model in Trial #2.

* + **Trial #4: Visibility + Humidity + Month Variables:**



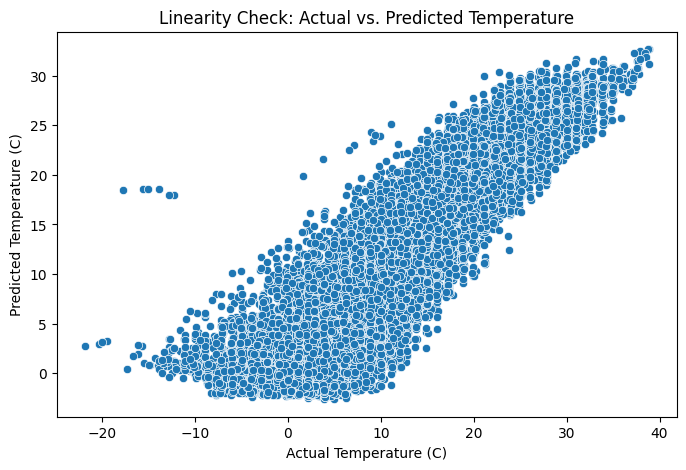
This model explains 80% of variance in Temperature, with a larger Test R² than Test#2 and better MSE results.

* + **Best Model: Trial #4(Visibility +Humidity +Month)**
  + **Model Summary:**

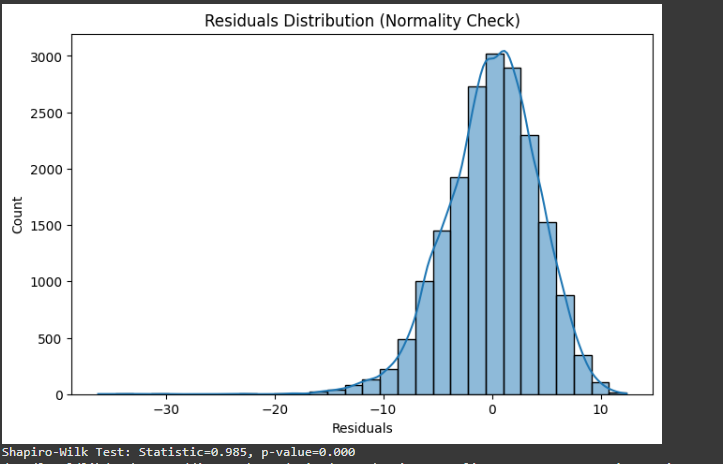


**Model Equation:**

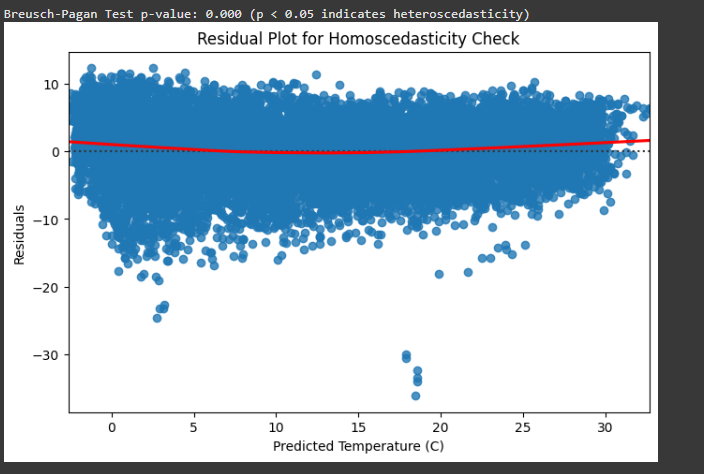
1. **Model Analysis for assumptions:**
   * **Linearity:**



* The plot shows a linear pattern between actual and predicted Temperature. Therefore, we can conclude that the **Linear assumption is valid.**
* **Normality Assumption:**



* The plot shows a bell plot for the residuals.
* Shapiro test> p-value. Therefore, we won’t reject H0.
* **Therefore, Normality assumption is valid**
* **Homoscedasticity Assumption:**



* Breusch-Pagan Test <p-value. Reject H0. Therefore, **The Homoscedasticity assumption is violated.**
* **Autocorrelation Assumption:**

Durbin-Watson Statistic: 1.982 (Values near 2 indicate no autocorrelation)

* Don’t Reject H0.
* Autocorrelation Assumption is Valid.

**Conclusion**

The study successfully identified an optimal multiple regression model for predicting temperature using Visibility, Humidity, and a sinusoidal transformation of Month. The model explains 80% of the variance, making it a strong predictive tool. While most econometric assumptions were satisfied, the model exhibited heteroscedasticity, suggesting potential improvements through weighted regression or transformation techniques. Overall, the analysis demonstrates the effectiveness of econometric principles in predictive modeling and highlights the importance of feature transformation in improving model performance.