**Introducton**

This project focuses on analyzing and modeling a dataset to understand the key factors that influence the number of hours an individual spends on coding per day. The objective is to develop a reliable linear regression model to predict hours\_coding based on various behavioral, productivity, and system usage metrics.

**Dataset Description**

The dataset consists of **daily observations** of individuals, with each row representing a single day’s record. The **dependent variable** is:

* hours\_coding: Total number of hours spent coding per day.

The **independent (predictor) variables** include:

* coffee\_intake\_mg : Caffeine intake in milligrams.
* distractions: Number of distractions or interruptions experienced.
* sleep\_hours: Hours of sleep the previous night.
* commits: Number of code commits made.
* bugs\_reported: Number of bugs reported.
* ai\_usage\_hours: Time spent using AI tools (e.g., Copilot, ChatGPT).
* cognitive\_load: Self-rated mental workload level.
* task\_success: Binary or score-based success metric of tasks completed.

This dataset is particularly relevant in understanding productivity patterns, software engineering behaviors, and how certain habits or tools (like coffee or AI usage) might influence coding performance.

To derive insights and build a predictive model, the following steps were undertaken:

1. **Data Splitting**:
   * The dataset was split into **training** (80%) and **testing** (20%) subsets using random sampling.
2. **Model Building**:
   * A **multiple linear regression model** was built using the training data with hours\_coding as the response variable.
   * All other variables were included as predictors.
3. **Statistical Analysis**:
   * The model's **coefficients**, **standard errors**, **t-values**, and **p-values** were analyzed to assess the **significance** and **impact** of each variable.
   * Key predictors like coffee\_intake\_mg, commits, and ai\_usage\_hours were found to be **statistically significant**.
4. **Assumption Checks**:
   * **Durbin-Watson Test** was performed to check for **autocorrelation** in residuals. Results indicated that residuals are independent.
   * Residual plots were reviewed to assess **homoscedasticity** (constant variance) and **model fit**.
5. **Model Evaluation**:
   * Model fit was strong with **R-squared = 0.85**, indicating that 85% of the variability in hours\_coding is explained by the predictors.

**CODE OUTCOME DOCUMENT**

Out-Come Image:

A screenshot of a computer

AI-generated content may be incorrect.

Explanation: Residual

* Residuals is the difference between the predicted Value and the actual value
* Here Residual value lies between -2.0980 and 4.2800
* It is clear that all quartiles , Median all are close to 0 so that the difference between actual value and the predicted value is less thereby model is good and there is a good prediction in the model

Explanation: Coefficient

* Coefficient is the relationship between dependent variable and the independent
* Here it shows all the relationships between Coding\_Hours and other variables
* Here intercept shows us the value of Coding\_Hours when other variables are 0
* Standard error defines how well the coefficient estimates is if it is low the model is precise
* Thereby here the model is precise because standard error is low
* T-Value test is conducted to test where there is a significant difference in coefficients.
* If T- Value is larger and P value is smaller there is a more significant relationship between predictors and Hours\_code
* So here Coffee\_intake\_mg , ai\_usage\_hours , Commits takes a larger value so that the coefficients is not 0 they affect the coding\_hours significantly
* The main idea is that every intake of coffee the Coding hours increase by 0.009 . ( the same idea applies to other variables as well)

Explanation: Residual Standard error

* Residual standard error defines as the size of the predicted errors
* If it is larger the model is not good
* There-by Here 0.769 is smaller so the model is better

Explanation: Adjusted R Squared

* It corrects the number of predictors
* This is close to 1 so indicating Strong model

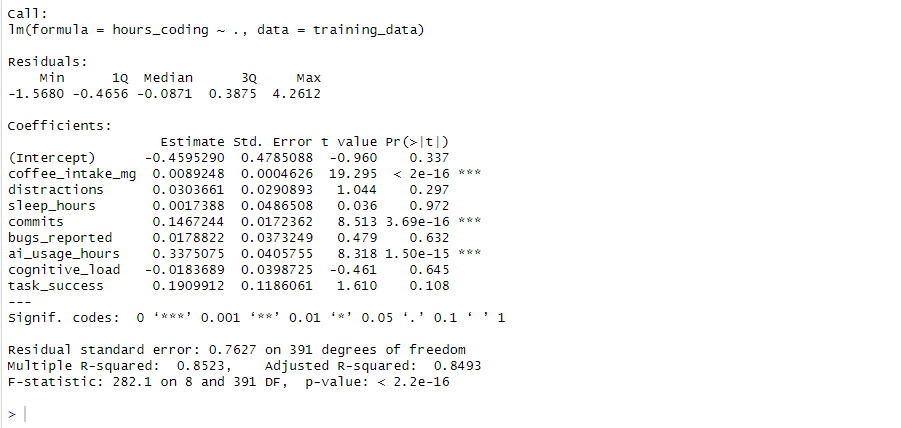
Explanation: Multiple R Squared

* This is the measure that used to find out whether the model fits best
* If R squared is closer to 1 the model best fits
* If R squared is closer to 0 the model doesn’t fit well
* There by here 0.84 which is closer to 1 that means the model fits well

Explanation: F-Statistics

* This measure ensures that the model is overall useful
* If the F statistic value is high and p value is lower it says that the model is overall good
* There-by in this data\_set the model is overall good

Analyzing after Splitting data-set to Training and Testing



Explanation: Residual

* Here after the separating the dataset for training and testing the residual lies between -1.5680 to 4.2612
* Here the median and quartiles are close to 0 there by there is no significant error there by prediction is good

Explanation: Coefficient

* Here intercept shows us the value of Coding\_Hours when other variables are 0 there by when coding\_hours is -0.45 when other variables are 0
* Here ai\_usage\_hours , commits , coffee\_intake\_mg effects significantly
* This model is precise because the standered error is low
* And when we compare the t value and p value if t value is high and p value is low it shows that there is a significant relationship between hours coding and other variables
* Generally the estimates explain us that every intake of variables the coding-hours increase or decrease accordingly

Explanation: Multiple and Adujested R square

* It is good that both adjusted and multiple R squaredare all most equal
* The R squared explain wether the model fits well or not
* There by since the r squared arecloser to 1 the model fits well

Explanation: Multiple and Adujested R square

* It is good that both adjusted and multiple R squaredare all most equal
* The R squared explain wether the model fits well or not
* There by since the r squared arecloser to 1 the model fits well

Explanation: Residual Standered error

* Here it shows less residual error there by the error size is less
* So this model is fit

Explanation: F-Statistics

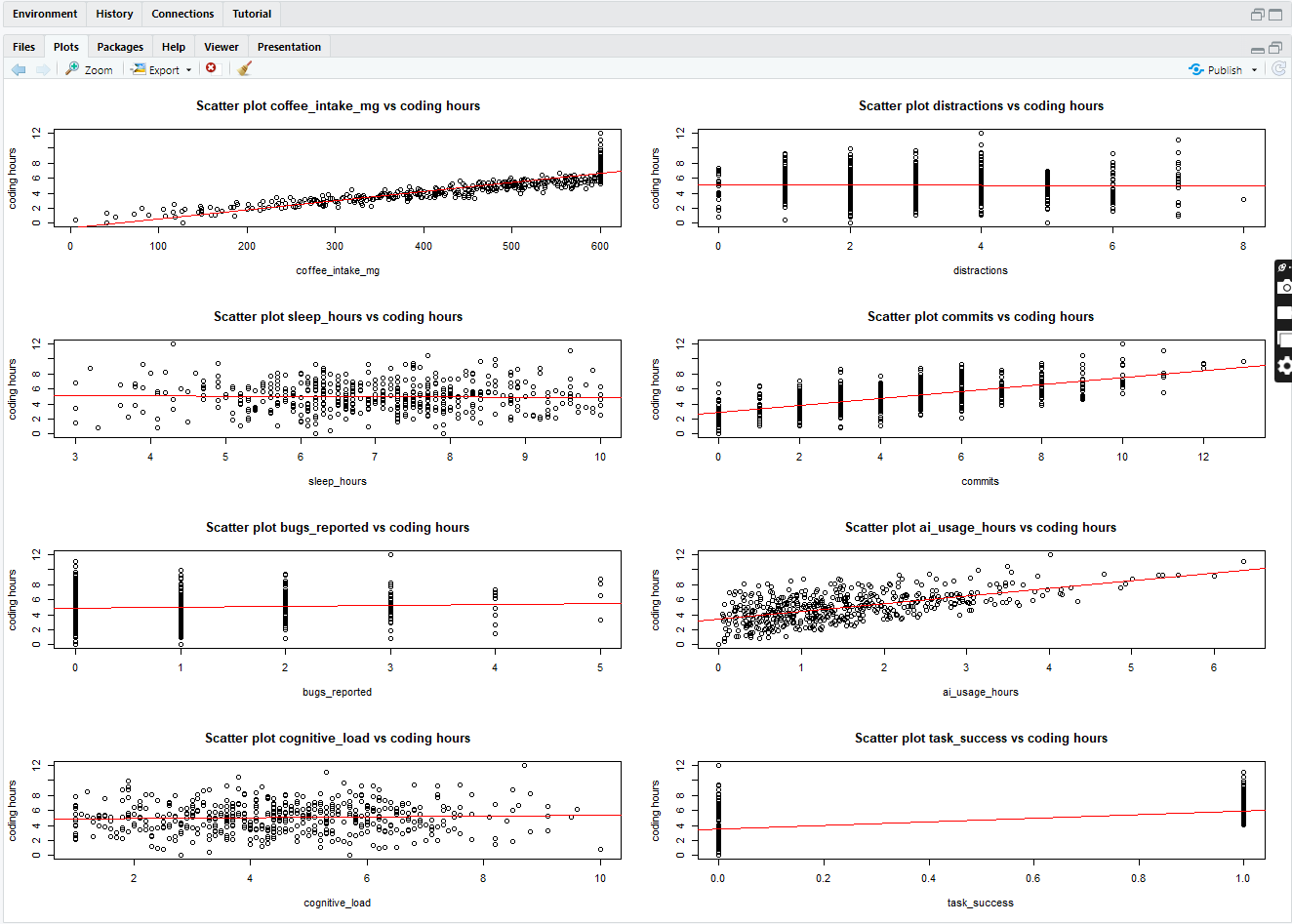
If F statistics is larger and the p value is smaller it says that the model is overall good

There by here this model is overall good

**CHECKING OF THE ASSUMPTIONS**

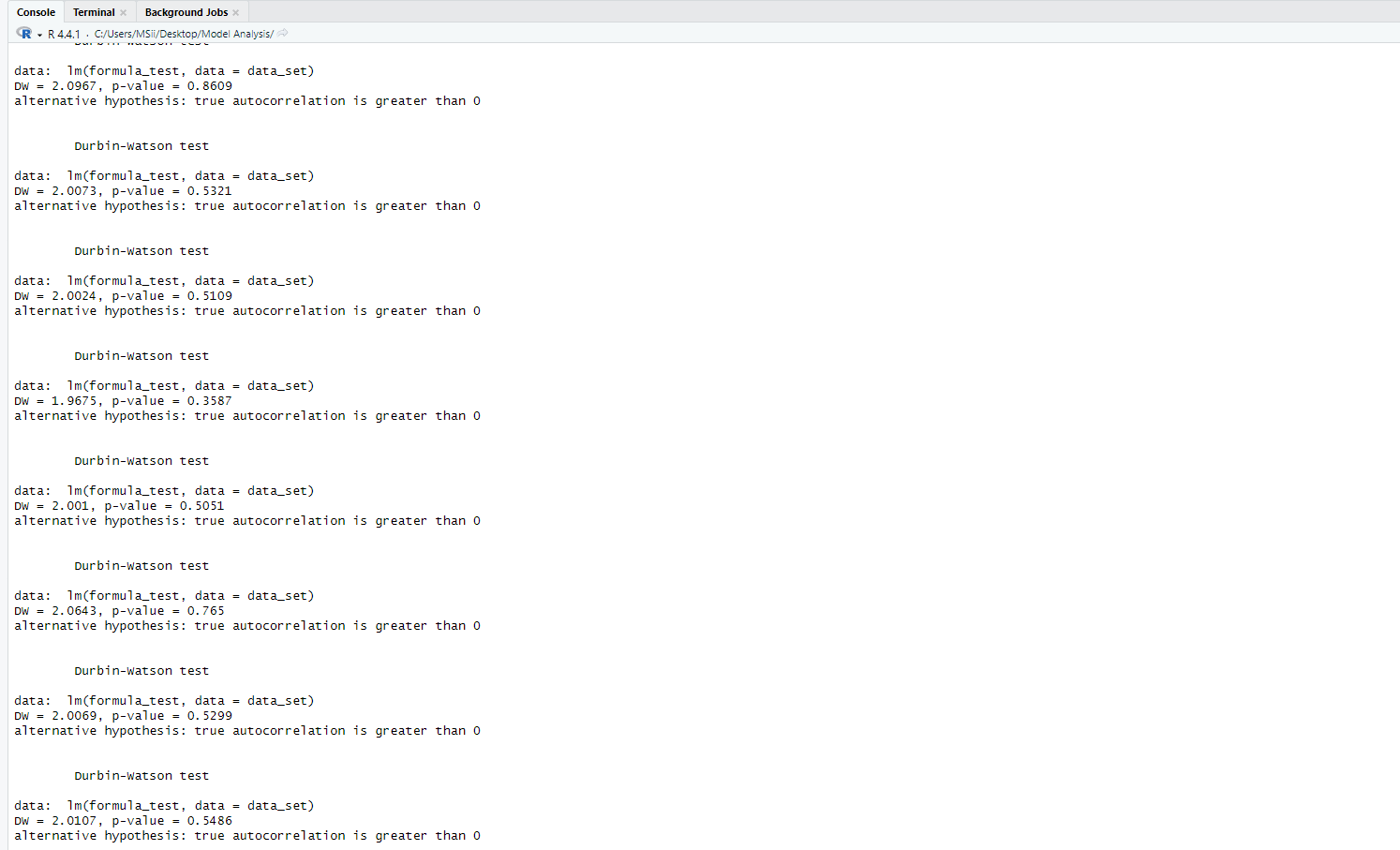
When you build a linear regression model, you're using a statistical method that relies on several key assumptions. If these assumptions are violated, your model:

* May produce biased results
* Might give incorrect p-values or confidence intervals
* Could make bad predictions

 **Scatter plots of The Linearity: This plot shows that the all relationship follows a linear model**

The Concept of the Linearity explain that the relationship between the independent variable and the dependent variable follows a linear model

Thereby Here all follows a linear model or a linear relationship

**Durbin Watson test for independence: Here There is No autocolleration since all value are close to 2** 

This is the test we do to check weather we have autocorrelation means that the errors in the data-set are correlated to each other .Thereby this violates the least square of error rule

Thereby there should not be auto correlated and In the above There is no autocorrelation there by this satisfies the assumption of independence

Code :

A computer code on a white background

AI-generated content may be incorrect.

Since we are checking the entire variable set with the independent variable we needed to loop through each and every line and get the result

**Homodecaity: All The points are scattered around 0**A screenshot of a computer screen

AI-generated content may be incorrect.

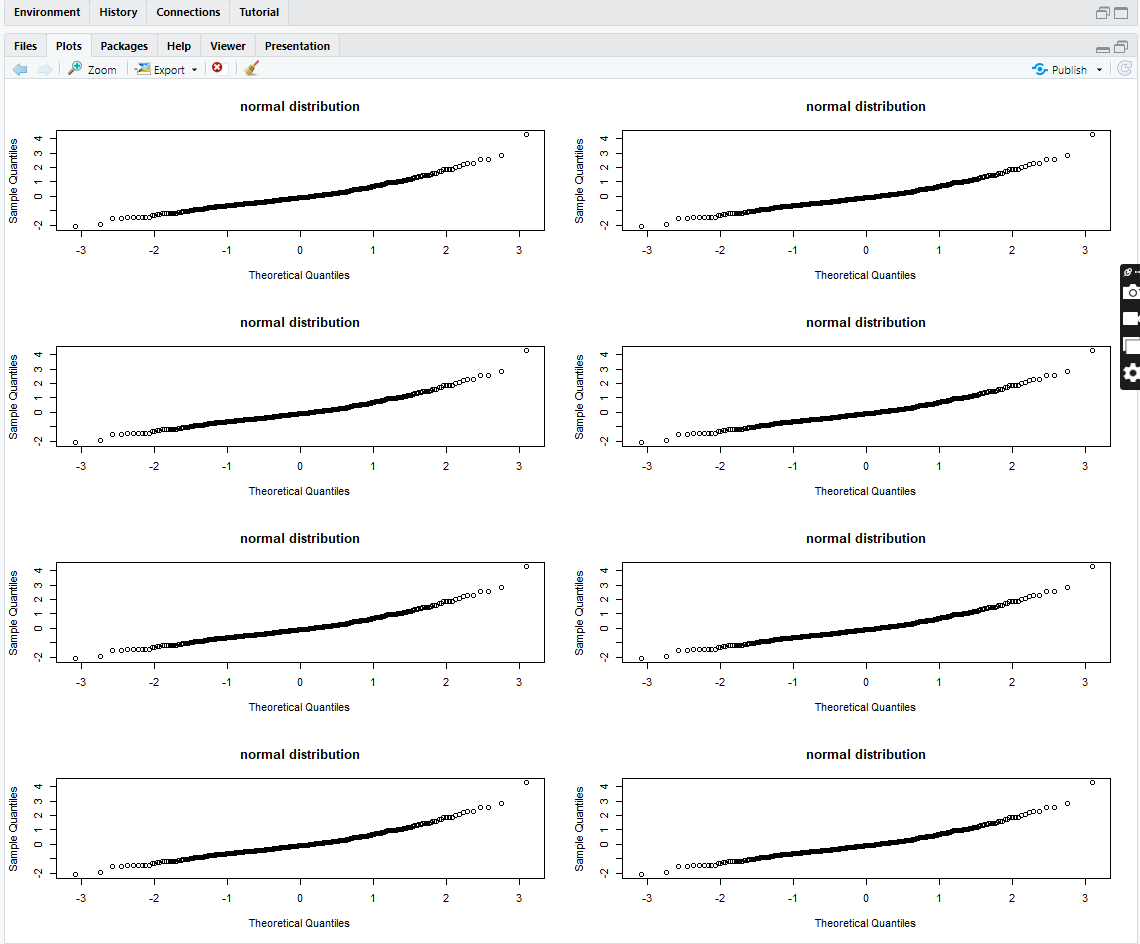
Code :

A computer code with text

AI-generated content may be incorrect.

Since we need all the graphs to show in 1 window, we need to divide the graph into the number of variables and the going through a loop to check the assumptions

**Normality : The errors of the model should be distributed normally ( Bell shaped curve )**

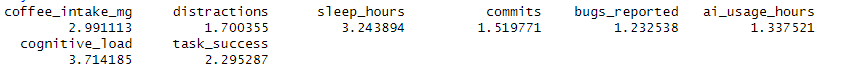


Code :

A computer code with text

AI-generated content may be incorrect.

**No Multicollinearity : Since all are less than 5 There is no significant Multicollinearity**



Code :

A close-up of a computer code

AI-generated content may be incorrect.

CAUTION: SINCE THIS IS DESCRIBING OF THE SET THERE CAN BE ERROS THERE BY BE CAUSTIOUS