Answers to the Subjective Questions:

1. Categorical variables influence the target variable positive or negatively. When it sign is positive, it means that the occurring of that variable will increase a point in the target variable, whereas a negative sign indicates that the occurrence of the categorical variable will lower the value of target variable. Weather Low in my module has a positive influence, thus when the weather severity is low, the count of cycles will increase. July variable has a negative influence on my target variable and this is justified as July is basically a month of rainy season and therefore not many people would go out.
2. For a categorical variable of n levels, the number of dummy variable created is n-1. Thus if we have gender as a categorical variable with Male and Female as levels, the get_dummies function of Pandas will create two variables as Male and Female, 0 indicating False and 1 indicating True. Thus Female 0 will actually indicate that the user is not a female, but in fact is male. Thus one column here is sufficient to explain the gender. We would therefore use **dropfirst** to drop the first column of gender, making the count of dummy variables as n-1
3. I got **atemp** as the highest correlation with the target variable followed by **temp**.
4. I plotted a **histogram** for the residuals and checked if the **curve** was **normal**. I also plotted a **scatterplot** that showed the **residuals** in a **linear pattern**.
5. According to my model, **atemp**, **weather_Low** and **July** contributed significantly

General Questions

- Answer to Q1: Linear Regression is a Machine Learning algorithm that predicts the value of Target or the dependant variable with help of the independent variable. The algorithm follows Least squared method and fits a straight line through the data-points. Thus it follows the y=mX+c formula where *y is the dependent variable, x is the independent variable* and *c is the constant.* Before we could implement Linear Regression onto a data, we need to check if the variable that we are using is having a positive correlation with the target variable.
- Answer to Q2: Anscombe's quartet comprises of 4 datasets that have similar descriptive statistics but when they're plotted on a graph, each one of them is distinct from the rest. This concept emphasizes that the data needs to analysed graphically first, before we move to draw inferences. It also portrays the effect of Outliers on the statistics of the dataset. Elaborating more on the datasets, one of them showed a simple linear relationship, the second one was distributed normally but did not show a linear relationship, third dataset was linear but it should've had a different line for an improved model whereas the last one did not show a linear relationship but had one data-point that was able to produce high correlation coefficient
- Pearson's coefficient is a measure of how to variables are correlated to each other. The value for this coefficient ranges from -1 to 1, as is interpreted as Linear relationship if the coefficient is 1, no linear relationship if coefficient is equals to 0 and an inverse linear relationship if it is -1. Numerically, it is equal to the covariance of the two variables divided by the product of standard deviations of the two variables respectively.
- Answer to Q4: Each variable has its own data type, thus to make it uniform, we scale it and bring all of the variables to a single level.
- Answer to Q5: VIF would come as INF if a variable has been repeated. Let's say, you were to erringly add **temp** column twice, then the VIF for **temp** would come as INF.
- Answer to Q6: A probability plot or Q–Q (quantile-quantile) is a graphical method by which you compare two probability distributions by plotting their quantiles against each other. The

concept involves determining the intervals for the quantiles. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line.