Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   a) Year – In the year 2019 more number of bikes are rented compared to 2018. So YOY grow is there before pandemic.
   b) Season – More bikes are rented in Summer and Fall seasons compared to Winter and Spring
   c) Weather Situation - During the Clear and Mist weather situations most of the people are renting bikes. The usage is very less during snow. No bikes are rented during heavy rain.
   d) Working day – Most of the people using the bikes on working days. In the Weekend during Saturday most of the people are using. Otherwise on Sunday usage is very less.
   e) Holiday – In case of Holidays the usage of bikes is less compared to non-holidays.
   f) Month - In the month of September most of the people are using the bikes. In the month of June there is decline in the usage due to holidays other than winter/fall months.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   When you use get dummies, you should use drop first = True so that you don't have dummy variable trap. You need to drop those columns which have redundant data. This could be the reason for the difference you see between one hot encoder and get dummies. If you drop first in get dummies, there will be no difference in model accuracy.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   According to pair plots temp has highest correlation with cnt target variable.

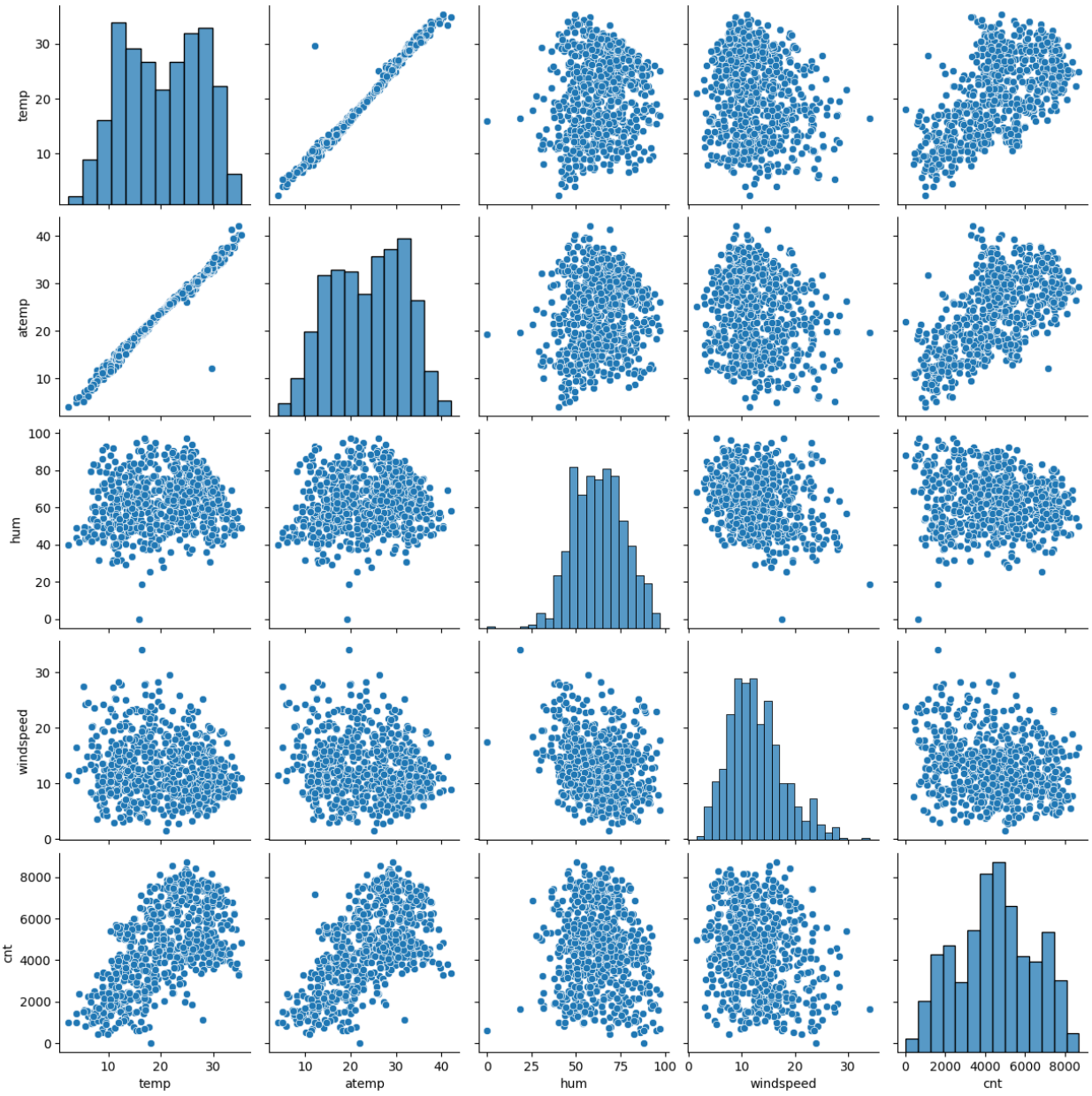
5 **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

a) Temp – Positive correlation.
b) 2019 - Positive correlation
c) Humidity – Negative correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Linearity :

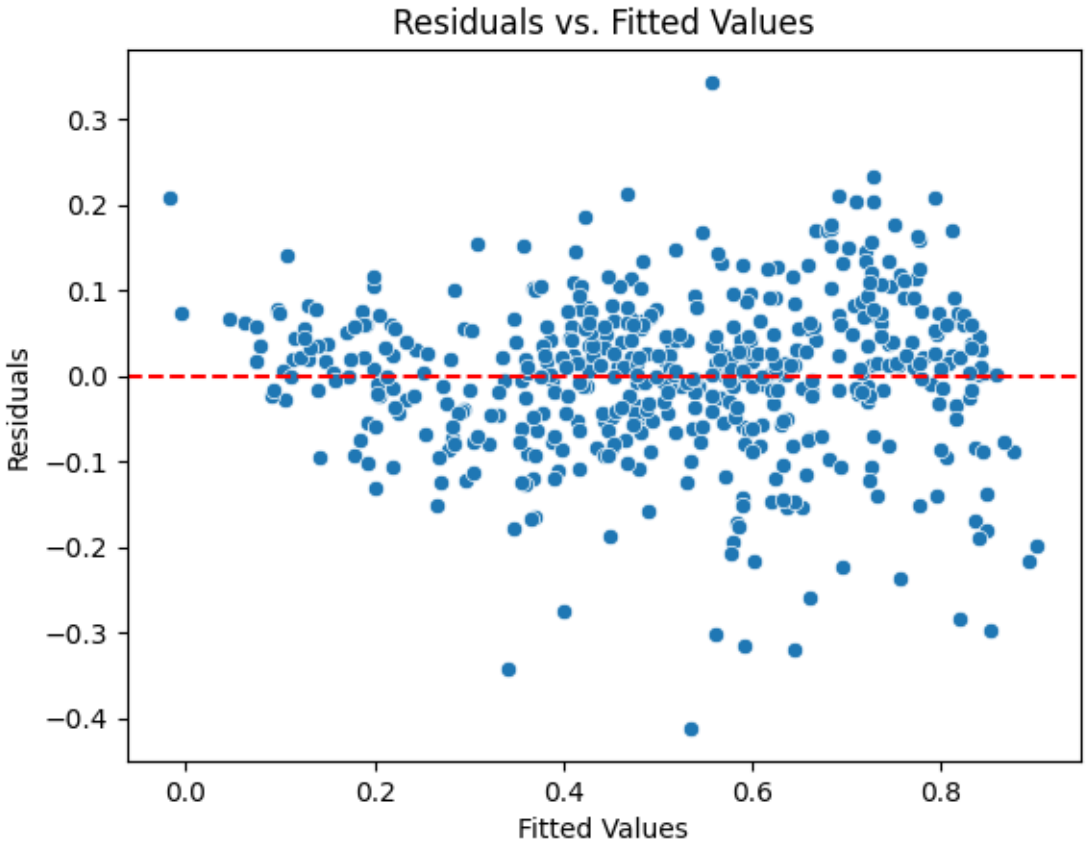
Added below scatter plots between numerical variable and dependent cnt variable to check the linearity



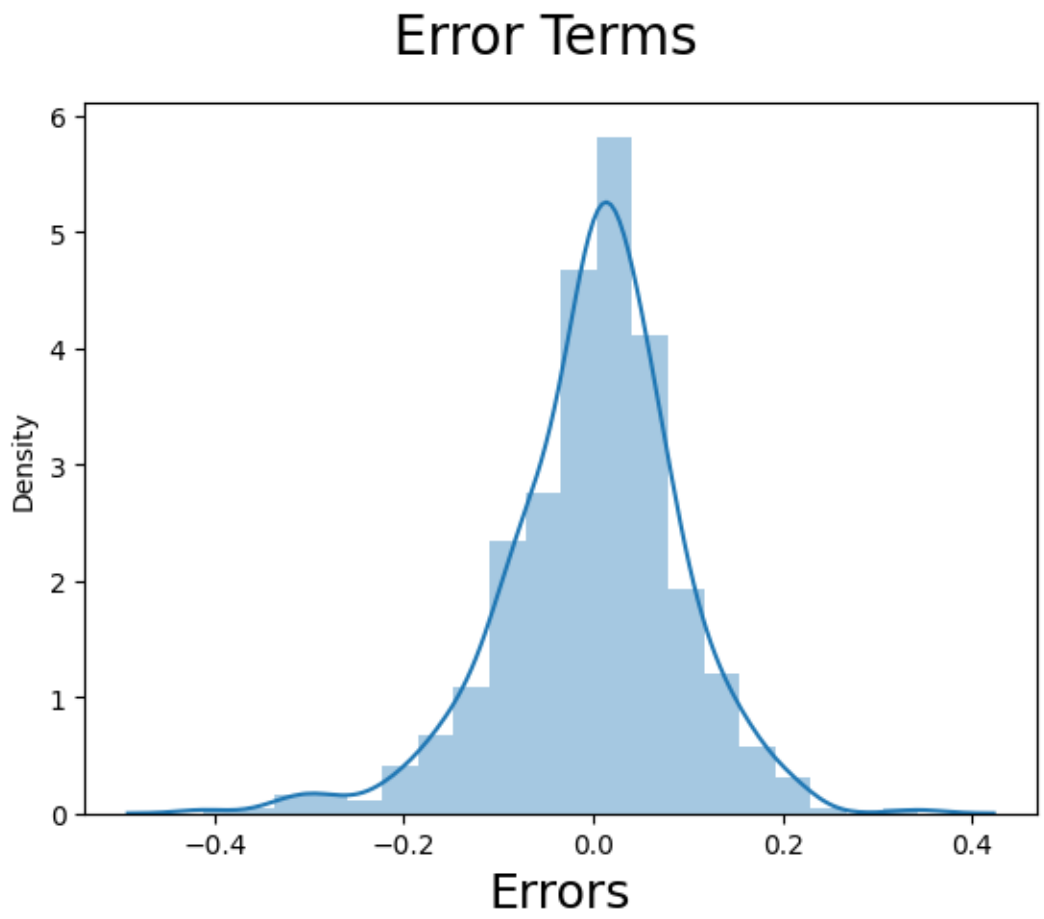
Homoscedasticity:

This plot helps to check the Homoscedasticity feature on final regression model

\



Normal Distribution of Error terms



Independence/No Multicollinearity:

Added the VIF factor value final model features. None of the feature are above 5 except constant.

0	const	56.13
2	temp	2.87
5	spring	2.53
6	winter	1.75
9	Jul	1.29
3	hum	1.26

4	windspeed	1.15
7	snow	1.11
10	Sep	1.11
8	2019	1.03
1	holiday	1.01

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is the supervised machine learning model in which model finds the best fit linear line between independent (x) and dependent variables(y)

Linear regression is of two types Simple and Multiple Linear regression.

In case simple Linear one independent variable will be there to predict the dependent. But in case of multiple linear regression multiple independent variables are there to predict the dependent variable.

The basic formula for Multiple Linear regression is:

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots . + b_nx_n$$

here b0 – Intercept and b1,b2…bn are coefficients for dependent variables

A linear regression model aim is to find best fit linear line with optimal value of intercept and coefficient where errors are minimized.

Assumptions of Linear Regression –

1. **Linearity** : The dependent variable Y should be linearly related to independent variables. We can check this by adding scatter plots.
2. **Homoscedasticity**: The variance of error term should be constant for all all values of independent variable. We can check this by adding residual plots
3. **Independence/No Multicollinearity:** There should be no correlation between independent variables. We can check this by using the correlation matrix or VIF score

4. The error terms should be normally distributed. We can check this by histograms

Evaluation of Linear regression:

1. R squared or Coefficient of Determination: The most used metric for model evaluation in regression analysis is R squared

2. Adjusted R squared: It is the improvement to R squared. The problem/drawback with R2 is that as the features increase, the value of R2 also increases which gives the illusion of a good model. So, the Adjusted R2 solves the drawback of R2. It only considers the features which are important for the model and shows the real improvement of the model.
3. Mean Squared Error (MSE): Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.
4. Root Mean Squared Error (RMSE): It is the root of MSE i.e Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE doesn't.

2 Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics.

3 What is Pearson's R?

The Pearson correlation coefficient (named for Karl Pearson) can be used to summarize the strength of the linear relationship between two data samples.

The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation

of each data sample. It is the normalization of the covariance between the two variables to give an interpretable score.

The result of the calculation, the correlation coefficient can be interpreted to understand the relationship.

The coefficient returns a value between -1 and 1 that represents the limits of correlation from a full negative correlation to a full positive correlation. A value of 0 means no correlation. The value must be interpreted, where often a value below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation.

The *pearsonr()* SciPy function can be used to calculate the Pearson's correlation coefficient between two data samples with the same length**.**

4 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

QQ — Plots, otherwise called as Quantile-Quantile plots are used to assess if a set of data come from theoretical distribution like Normal distributions. It takes theoretical quantiles on the x-axis and data on y-axis.

In Linear regression we used to assess the one the assumption like the residuals of data should for a normal distribution with mean 0 and constant variance. So to check this assumption we will use QQ plots

This the process we will follow to measure the normal distribution.
Take the residual values and compute the percentile ranks values for residual values. Generate the Z-scores from these percentile rank values.
Then plot the QQ plot between Residuals and Z-scores to check the normal distribution.

5 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units, *and range. If scaling is not* done, *then algorithm only takes magnitude in account and not units hence incorrect*

modelling. To solve this issue, we must *do scaling to bring all the variables to the same level of magnitude.*

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outlier

6 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If there is perfect correlation between two independent variables in the dataset then the VIF will become infinity. According to this you have redundant features in your dataset.