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?

ANS:

Inference about categorical variables effect on dependent variable:

- The median of bike rentals has increased from 2018 to 2019. It also shows that the number of bike rentals has increased in the year 2019.
- Bike rentals counts increase significantly on holidays.
- Fall and summer are the 2 seasons with best count of bike rentals. Whereas, spring is the worst count.
- Bike rentals during heavy snow/rain climate are almost negligible. Moreover, clear weather rentals are quite high.
- Surprisingly, the overall medians of working day and non-working day are the same.
- Bike rentals from the month May to June are quite high.

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

ANS:

Importance of dropping the first dummy column:

- Dummy variables are created to convert each level of a categorical variable in a scaled number in order to perform Regression on it.
- In the process, we do not have to create a column for each level of a a categorical variable, since, if '01' could mean the presence of the 2^{nd} level and '10' could mean the presence of the 1^{st} level, then '00' can mean the presence of the 3^{rd} level of the categorical variable.
- ML methodologies can recognize these patterns and perform regression.
- This method reduces the number of columns which proves to be very useful when dealing with a large number of variables in the model building phase.
- This method can help reduce the memory occupied, time for execution and chances of overfitting, when large number of variables are present.

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

ANS:

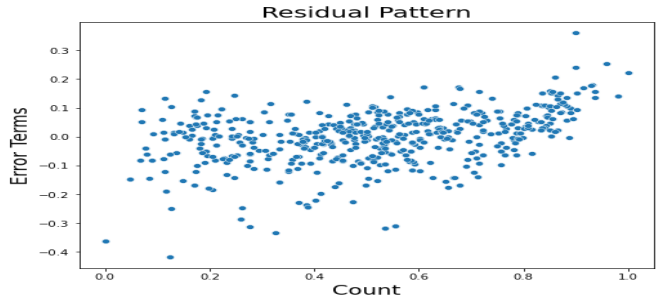
Looking at the pair-plot for numerical variables, we can say that both the variables "temp"(meaning the actual atmospheric temperature) and "atemp"(meaning the temperature we feel) have the highest and almost equal co-relation with the target variable.

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

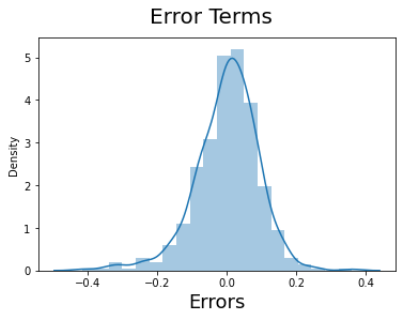
ANS:

Validation of Assumptions of Linear Regression:

- The assumptions of linear regression are:
  a) Linearity.
  b) Homoscedasticity
  c) Independence
  d) Normality
- Linearity of the Model was validated by the pair plot and co-relation heatmap between the independent variables and the target variable.
- Homoscedasticity was validated by plotting a scatter plot of residuals(errors) along with the target variable and we observed no pattern.



- Independence of independent variables was validated by observing that no variables has a VIF>5.
- Normality was validated by plotting the distribution of error terms and obtaining a near perfect normal distribution centred at 0.



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

ANS:

Top 3 features:

1. **temp**: This variable has the highest positive co-efficient value of 0.4782, meaning, this variable has a positive high influence on the dependent variable, since, the dependent variable(i.e Count of bikes) will increase significantly with the increase in this independent variable(i.e Temperature).

2. **Light_snowrain**: This variable has the highest negative co-efficient value of -0.2860, meaning, this variable has a high negative influence on the dependent variable, since, the dependent variable(i.e Count of bikes) will decrease significantly with the increase in this independent variable(i.e Light_snowrain).
3. **year:** This variable also has a high positive co-efficient value of 0.2341, meaning, this variable has a positive high influence on the dependent variable, since, the dependent variable(i.e Count of bikes) will increase significantly with the change in the year.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

ANS:

Definition:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task.

Types of Linear Regression:

Linear Regression can be broadly classified into two types of algorithms:

1. Simple Linear Regression

A simple straight-line equation involving slope (dy/dx) and intercept (an integer/continuous value) is utilized in simple Linear Regression. Here a simple form is:

$y=mx+c$

Where, y denotes the output x is the independent variable, and c is the intercept when x=0.

2. Multiple Linear Regression

When a number of independent variables more than one, the regression equation is:

$y= c+m1x1+m2x2... mnxn$

Where, the coefficient responsible for impact of different independent variables (x1, x2 ,…) are (m1,m2,…) .

3. Non-Linear Regression:

When the best fitting line is not a straight line but a curve, it is referred to as Non-Linear Regression.

Assumptions of Linear Regression:

1. **Linear and Additive:** The variables must have a linear relationship between the target variable.
2. **Independence:** The independent variables must not be dependent on other independent variables.
3. **Normal Distribution of error terms:** The distribution of error terms must follow Normal Distribution with a mean=0.
4. **Autocorrelation:** The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.
5. **Heteroskedasticity:** The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.

Uses:

- Linear regression is mainly used for Prediction i.e identifying the variables that affect a target variables.
- Ex: We can find the driving factors(independent variable) for the revenue (as the target variable) in a new business.

2) Explain the Anscombe's quartet in detail.

ANS:

- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.
- They have very different distributions and appear differently when plotted on scatter plots.

- They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.
- Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).
- Moreover, the linear regression can only be considered a fit for the data with linear relationships and for any other dataset the result will not be accurate.

3) What is Pearson's R?

ANS:

- The Pearson correlation coefficient (r) is a measure of calculating the linear relationship between 2 variables.
- Given 2 variables X and Y, the co-relation co-efficient is given by,

$$\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\,\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}.$$

- If the value of Pearson's r is closer to 1, the value of one variable increases with an increase in the other variable i.e Positive linear Co-relation.
- If the value of Pearson's r is closer to -1, the value of one variable decreases with an increase in the other variable i.e Negative linear Co-relation.
- If the value of Pearson's r is closer to 0, there is a weak linear relationship between variables.
- **It does not mean that,** if**,** Pearson's r is weak, there is no type of relationship between variables, it only means there is no linear relationship between then.

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

ANS:

Scaling:

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

Why is scaling performed:

- Many a times, different variables have data in different units,(EX: Variable Salary could be in lakhs and variable Revenue could be in Crores). Scaling helps in converting the data in a particular range.
- It also helps in speeding up the calculations in an algorithm.

- If scaling is not done then algorithm only takes magnitude in account and not units hence, it could lead to a wrong interpretation where a few co-efficients have a higher magnitude to it's original unit and not it's actual significance in the model.

The difference between normalized scaling and standardized scaling:

- Normalization means rescaling the values into a range of [0,1].
- Standardization means rescaling data to have a mean of 0 and a standard deviation of 1 .
- In Normalized scaling, each and every value of the variables will be in fixed range, that might not necessarily be the case in standardized scaling

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

ANS:

- VIF stands for Variation Inflation factor, meaning, it calculates the multicollinearity of an independent variable with other independent variables.
- This is useful in identifying which variables must be included in a linear regression model.
- VIF becomes infinite, when there is perfect co-relation between variables.
- This, basically means there is perfect co=relation between 2 independent variables.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).
- In such a case, we usually drop only one of the 2 variables and that to the one that might have the least significance with a business point of view.

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

ANS:

- Q-Q plot i.e quantile-quantile plot, describe the quantiles of a sample distribution against quantiles of a theoretical distribution.
- Doing this helps us determine if a dataset follows any particular type of probability distribution.
- Probability distributions are essential in data analysis and decision-making.
- Knowing which distribution we are working with can help us select the best model.
- Moreover, Q-Q plots are used to visually check that your data meets the homoscedasticity and normality assumptions of linear regression as well.
- This helps us in validating the assumptions of a linear model.
- Q-Q plot can also be used to detect outliers.