Assignment Subjective Questions

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

From the analysis of the categorical variables from the dataset, we could see the following effects:

- **Season -** We can see that the Count of the rental bikes (which is the dependent variable), was very low in 'Spring' season, whereas in other seasons, i.e., the average count was almost similar.
- **Year ('yr') -** The count of the rental bikes significantly increased in the year 2019 as compared with the year 2018.
- **Month ('mnth') -** The count of the rental bikes were moderately increasing each month till the month of 'October' and it started decreasing in 'November' and 'December'.
- **Weather Situation ('weathersit') -** The count of the rental bikes significantly dropped when there were 'Rains'.

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

When we have a categorical variable with 'n' levels, we create dummy variables to build 'n-1' variables, indicating the levels. For example, we have a variable 'Gender' with 2 levels, namely, 'Male', and 'Female', we would create a dummy table as the following:

Gender	Male	Female
Male	1	0
Female	0	1

As we can clearly see, there is no need to define two different levels. If we drop the level 'Male', we will still be able to explain the two levels. As, the value in 0 will just tell us that the person is 'Male'. Having both levels in the data set while building the model, it would take one of them as base state and create an extra effect in the model, making it important for us to use '**drop_first=True'**, while creating the dummy variables.

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

Looking at the pair-plot among the numerical variables, in the data. The variable **'temp'** that is the Temperature is highly correlated with the **'cnt'** variable (target variable) which represents the Count of the rental bikes.

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

Performing the Residual Analysis - that is plotting a histogram to check if the error terms are normally distributed and if the error residuals have a mean value of zero, helped to validate that the assumptions of that the error terms have to be normally distributed around zero in a Linear Regression Model.

Also, performing the model evaluation where a scatter plot was used to validate the following assumption:

- The pair-plot in the beginning validated that there was Linear Relation between a few variables.
- There was no visible pattern visible which validated the assumption that Error Terms are independent of each other.
- The variance did not seem to increase or decrease which validated homoscedasticity.
- The checking of VIF values validated that there was no multicollinearity between variables.
- The p-values of the coefficients were lower than 0.05, which validated the significance of them.
- Also, checking adjusted R-squared, to check for variable significance.

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

The top 3 features contributing significantly towards explaining the demand of shared bikes are:

- **Year ('yr') -** We could see that the coefficient of the 'yr' variable was significantly high, that is, as each year goes by the total increase in the Count of rental bikes would be 0.247.
- **Spring Season ('spring') -** We could see that the coefficient of the 'spring' variable was low, statistically showing that the demand of shared bikes would decrease to -0.298 in spring season.
- **Rainy Weather Situation ('Rain') -** We could see that the coefficient of the 'Rain' variable was low, statistically showing that the count of rental bikes would decrease to -0.297 during Rains.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression Algorithm involves the following steps:

- **Reading and Understanding the Data**

  We need to first import all the required libraries, such as numpy, pandas, matplotlib, seaborn, sklearn, statsmodels, once we are done with importing. We need to import the .csv file which has the data.

  After that, we inspect the data to check the shape of the data, the number of columns and rows, to see if there are any null values or to identify the continuous and categorical variables.

- **Perform EDA and Visualise the Data**

  We need to identify the data types of the variables, to see if the categorical and continuous variables have the correct data types.

  We visualise the data by using pair-plots, scatter plots or box plots to see if there is any obvious multicollinearity and to identify if some predictors directly have a strong association with the outcome variable.

- **Data Prepaperation**

  This involves handling the categorical variables first and then performing dummy encoding. By handling the categorical variables it means that we need to convert the Categorical variables to Object data type for creating dummy variables.

  We need to convert the boolean variables i.e, the Yes and No, or True or False to binary values. We map ‘Yes’ to 1 and ‘No’ to 0.

  We create dummy indicators for other categorical variables.

- **Splitting the Data into Training and Testing Sets**

  We now split the data into Training and Testing Sets. We do this by keeping 70% of the data in the Train Set and 30% of the data in the Test Set.

- **Rescaling the Features in Train Set**

  We need to rescale the variables to have a comparable scale. If we don't rescale the features, then some of the coefficients obtained by fitting the regression model might be extremely high or extremely low as compared to the other coefficients. There are two common ways of rescaling:

  1. **Min-Max scaling -** Brings all the data in the range of 0 to 1.
  2. **Standardisation (mean-0, sigma-1) -** Brings all the data into a standard normal distribution with mean 0 and standard deviation 1.

- **Dividing the Train Data into X and Y sets**

  We need to divide the data into X and Y sets, where X set will include all the independent variables and Y will include the target variable.

- **Feature Selection**

  There are 4 approaches to perform feature selection:

  1. **Manual Feature Elimination**

     Here we build a model using all the variables then start eliminating the features one by one based on their p-values, VIF and Correlations.

  2. **Automated Approach**

     Here we would use an automated approach, i.e., Recursive Feature Selection, Forward/Backward/Stepwise Selection based on AIC or Regularisation (Lasso)

3. **Balance Approach**

   Use a combination of both, Automated (coarse tuning) + Manual (final tuning).

- **Building the Model**

  We first start with importing statsmodels.api, we can even build the model using sklearn but statsmodels gives us a Statistic Summary which is very important for considering R-squared, p-values, F-statistic, etc.

  We need to add a constant to the Train data, as statsmodels does not add a constant by default. If we do not add the constant, statsmodels fits a regression line passing through the origin, by default.

  Next step is to check the statistical summary, to see the R-Squared, coefficient p-values, F-statistic. We also need to check the Variance Inflation Factor, to check for multicollinearity.

  Now, we start eliminating the features one by one and keep rebuilding the process after removing each variable.

- **Residual Analysis**

  We perform the residual analysis to validate our assumptions, that is to check:

  1. If the error terms are normally distributed with Mean 0.
  2. If the error terms are independent of each other/
  3. If the error terms follow homoscedasticity.

- **Making Predictions on Test Data using the final model**

  Once we have fitted the model and checked the normality of error terms, we make predictions on the test data using the final model.

  For this, we need to add a constant to the test data, apply scaling to test data, divide the test set into X and Y and drop the same variables that were dropped in the Train data.

We now simply predict Y values corresponding to the X Test Set using the *predict ( )* attribute.

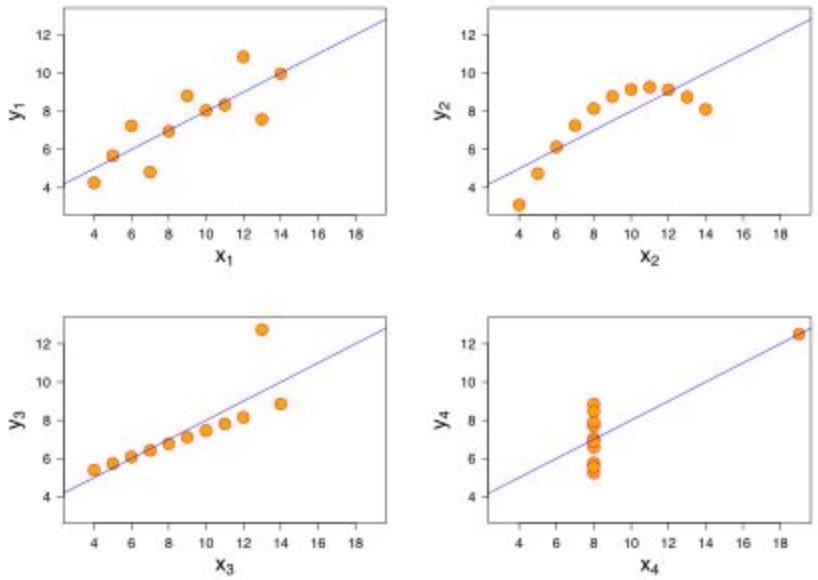
We check for R-squared and if the R-squared value is similar to the train data,

- **Model Evaluation**

  Finally, we plot graphs for actual versus predicted values and evaluate the model.

  Its also important to check the R-score in the test data, if the R-squared value is close to the R-squared of the train data, we can say that we have a best fitted model, as having similar R-squared would mean that what model has learnt on the training set, it is also able to generalise on the test set, which clearly means that is the best fitted model.

2. Explain the Anscombe's quartet in detail.



Anscombe's Quartet consists of four data sets with eleven (x, y) pairs.

The first three data sets have the same x values and the same standard output when we perform a linear regression so it is easy to assume that the data sets might also be the same.

But if we plot these values of the dataset in scatter plots, we will find that our assumption is wrong, as seen in the above image. We can clearly see in the image of scatter plots that even though it is good to use linear regression, the data is easily influenced by outliers as seen in Graph 3 and 4 and the whole model might get rendered.

This is why it is important to plot the variables before fitting a model to find and handle such outliers or any other errors in the data which will help us select the right model.

We can always plot the data again after fitting a model to validate our assumptions and decide on goodness of fit.

3. What is Pearson's R?

Pearson's R is a statistic that measures linear correlation between two variables X and Y. It has value between -1 and +1.

- 1 is positive linear correlation
- 0 is no linear correlation
- -1 is negative linear correlation

The correlation coefficient should not be calculated until and unless a Linear Relationship exists and must be calculated only on numerical variables. A scatter plot is generally used to check for correlation between the variables, as it gives a better visual idea about the correlation. If the points in the scatter plot are similar to a straight line, then it would mean that there is a higher correlation between the variables.

In Least Square Regression Analysis:

The square of the sample correlation coefficient is denoted r^2.

It estimates the fraction of the variance in Y that is explained by X in a simple linear regression. So if we have a dataset $\mathbf{Y_1,.....,Y_n}$ and the fitted dataset $\mathbf{\hat{Y}_1,....,\hat{Y}_n}$ then as a starting point the total variation in the $\mathbf{Y_i}$,

This r^2 is majorly used while building Linear Regression model.

An Adjusted r^2 is used to while performing Multiple Linear Regression.

As an adjusted r^2 penalises the model on the basis of the number of variables present in it.

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

   Scaling is a process where the values of the independent variables in a Data set are normalized, to bring them to a comparable scale while building a model. If we do not have a comparable scale while building a mode, then the parameters obtained by fitting the model might be very high or very low.

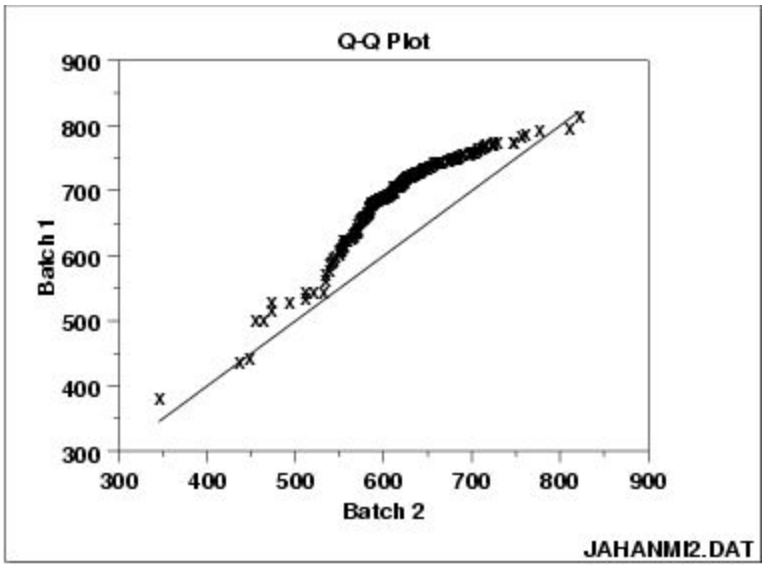
   Normalized Scaling brings all the data values in the range of 0 to 1, while **Standardized Scaling** brings the data values into a standard normal distribution with mean as 0 and Standard Deviation as 1.

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

   If the value of VIF is infinite it means that there is perfect between the independent variables. This happens when the Standard Error of the coefficient is raised by factor of 2 which implies that the confidence intervals might be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

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

   A Quantile-Quantile Plot (Q-Q Plot) is a tool which is used to assess if a data set has come from a theoretical distribution such as Normal or Exponential. This plot is basically a scatter plot which is plotted by using two sets of quantiles.



By a quantile, it means the fraction of points below the given value. If the points in the plot look similar to a straight line then it would mean the both the sets of quantiles come from the same distribution.