Linear Regression Subjective Questions

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)**

    - Fall season has the highest number of bike rentals followed by summer.
    - Year 2019 has higher rentals than 2018
    - When it comes to Months, the sales have a positive trajectory from Jan to July, then a downward trend from July to Dec.
    - July has the highest bike rentals.
    - A user is more likely to rent a bike on a holiday than a working day as per EDA.
    - A user is more likely to rent a bike when the weather is "clear" and less likely to rent it when it snows.

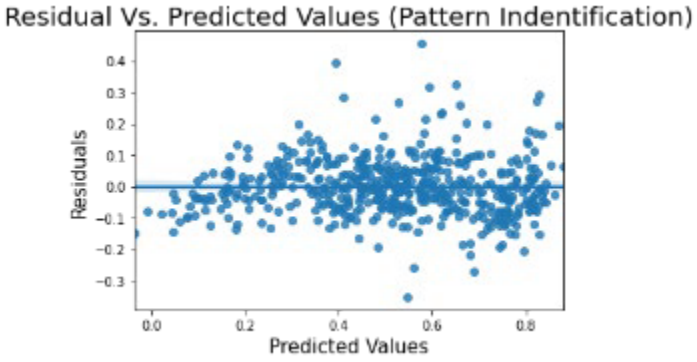
2. **Why is it important to use drop_first=True during dummy variable creation? (2 marks)**

    It is known that if a categorical variable has n levels of categorization, it can be analyzed with n-1 levels. drop_first = True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
    Syntax -
    drop_first: bool, default False, which implies to get k-1 dummies out of k categorical levels by removing the first level.
    Let's say we have 3 types of values in the Categorical column and we want to create a dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need the 3rd variable to identify the C.

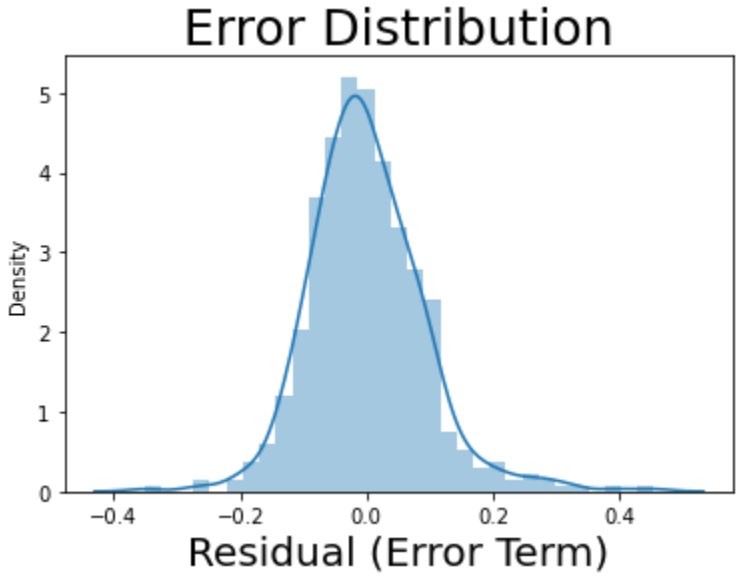
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
    atemp and temp both have the highest correlation with cnt. But since we choose only atemp for our analysis, its atemp in this case.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
    Linear relationship between features and target (tmp/atemp with cnt)
    Little or no multicollinearity between features (Removed temp since it was correlated with atemp)

Homoscedasticity was confirmed from the plot in the regression between residuals and predicted values has uniform variance.



Error Terms are normally distributed with mean zero as shown in the regression model .



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Top 3 features contributing significantly towards explaining target variable cnt are:

i. atemp (0.4117)
ii. Year (0.2357)
iii. Light Snow (-0.2912)

if the weathersit is combined it sums up to : - 0.6823

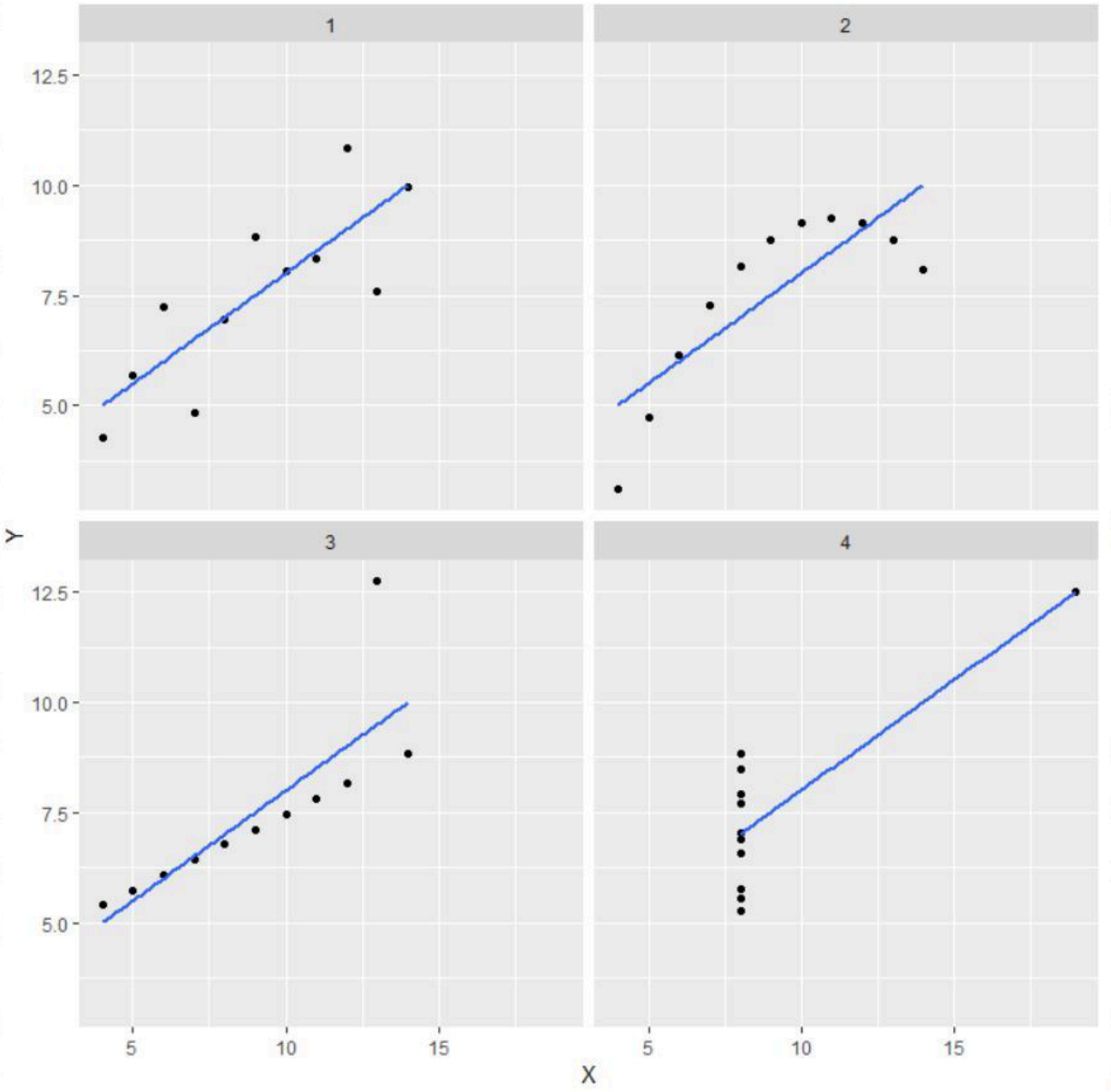
General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   It is the simplest form of regression. It is a technique in which the dependent variable is continuous in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature. There are two types of linear regressions
   a) Simple Linear Regression: it only deals with one dependent and one independent variable.
   b) Multi Linear Regression: it deals with 2 or more predictor variables.
   The independent variables are known as “predictor variables” and the dependent variables are known as “output” or “target” variables.
   Linear regression at each X finds the best estimate for Y. As the model predicts a single value, there is a distribution of error terms.
   As we are making inferences on the population using a sample, the assumption that variables are linearly dependent is not enough to generalize the results from sample to the population. So, we should have some assumptions to make inferences.
   Assumptions of Linear regression:
   1. There is a linear relationship between X and Y
   2. Error terms are normally distributed with mean zero.
   3. Error terms are independent of each other
   4. Error terms have constant variance. (homoscedasticity)
   When you fit a straight line through the data, you will obviously get the two parameters of the straight line: intercept ($\beta 0$) and the slope ($\beta 1$). You start by saying that $\beta 1$ is not significant, i.e. there is no relationship between X and y.
   So in order to perform the hypothesis test, we first propose the null hypothesis that $\beta 1$ is 0.
   And the alternative hypothesis thus becomes $\beta 1$ is not zero.
   • Null Hypothesis (H0): $\beta 1=0$
   • Alternate Hypothesis (HA): $\beta 1\neq 0$
   If you fail to reject the null hypothesis that would mean that $\beta 1$ is zero which would simply
   mean that $\beta 1$ia insignificant and of no use in the model. Similarly, if you reject the null hypothesis, it would mean that $\beta 1$ is not zero and the line fitted is a significant one.
   To perform the hypothesis test, you need to derive the p-value for the given β. If the p-value turns out to be less than 0.05, you can reject the null hypothesis and state that $\beta 1$ is indeed significant.
   The first important step before building a model is to perform the train_test_split. To split the model, you use the train_test_split function and check the summary statistics that was outputted by the model: F-statistic, r-squared, coefficients and their p-values.
   Then perform the Residual analysis by plotting histogram of the error terms to check the normality and independence. Then make predictions on the test set.

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   In simple terms Anscombe's quartet can be explained as follows. Lets assume we have 4 data sets each consisting of elevel (x,y) points, they have same statistical properties like variance, mean and linear regression. This would lead us to believe that the data sets are similar and relative. But when you plot them, they are quite different from one another. Hence Anscombe's quartet gives us the importance of plotting data on a graph before making conclusions. In 1973 Anscombe plot four such plots and its called anscombe's quartet . This demonstrates both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
   Data Points:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Corresponding Graph:



Explanation of this output:
In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

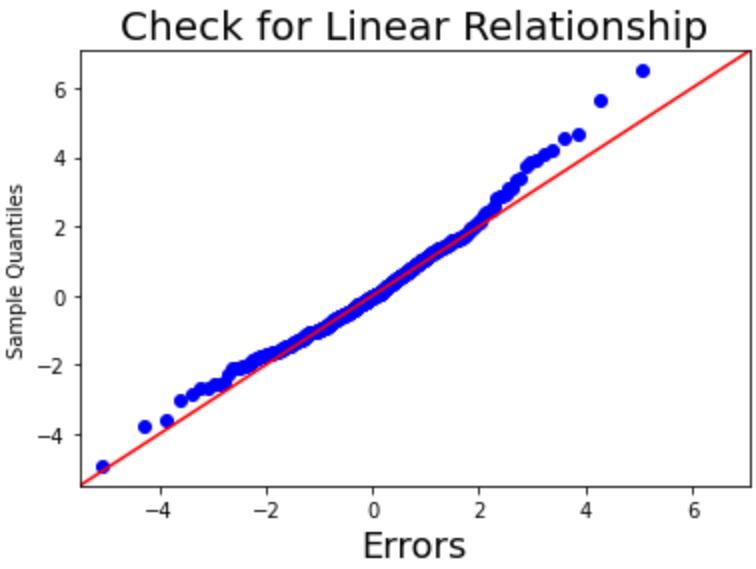
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistical properties for describing realistic datasets.

3. **What is Pearson's R? (3 marks)**
   Pearson's R a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with
   covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
   There are two types of scaling: Normalized and Standardized. Normalization means it rescales the values between 0 and 1 whereas Standardized typically rescales the values to have mean 0 and a standard deviation 1.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   Formula for VIF is 1/1-R2
   The only way the above can be infinity is when R2 (Rsquared) is 1 . We know that when Rsquared is 1 the regression predictions perfectly fit the data and model and it learns the data too well. **Hence there is perfect correlation between variables.**

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
   This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with the same distributions.
   QQ Plot from our regression model :


Check for Linear Relationship
Sample Quantiles
Errors
6
4
2
0
−2
−4
−4
−2
0
2
4
6