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, and, in each season the booking count has increased from 2018 to 2019.
   - Year 2019 has higher rentals than 2018.
   - When it comes to months, the sales have a positive trajectory from Jan to Jun, then a downward trend from July to December.
   - A user is more likely to rent a bike when the weather is "clear" and less likely to rent it when it snows.
   - Thursday, Friday, Saturday have more bookings when compared to the start of the week.
   - When it's a holiday, bookings seem to be less in number.

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 levels of values (a, b, c) for a categorical column X and we want to create a dummy variable for that column. If one value is not a and b, then it is obviously c, so we do not need the 3rd variable to identify c.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   temp and *atemp* both have the highest correlation with *cnt* (the target variable). However, as part of the RFE *atemp* is ignored (as *rfe.support_ value* for *atemp* is False) so temp is considered for analysis.

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 (*temp/atemp* with *cnt*)
   - Little or no multicollinearity between features (Removed *atemp* as part of RFE)
   - 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. temp (0.5473)
   ii. Year (0.2332)
   iii. Light Snow (-0.2890)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

   Linear Regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with a given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

   There are two types of linear regressions
   Simple Linear Regression: it only deals with one dependent and one independent variable.
   Multi Linear Regression: it deals with 2 or more predictor variables.

   Linear Regression is the simplest form of regression. 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 (β0)** and the **slope (β1)**. You start by saying that β1 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 β1 is 0.
   The alternative hypothesis thus becomes β1 is not zero.

- Null Hypothesis (H0): $\beta 1=0$
- Alternate Hypothesis (HA): $\beta 1 \neq 0$

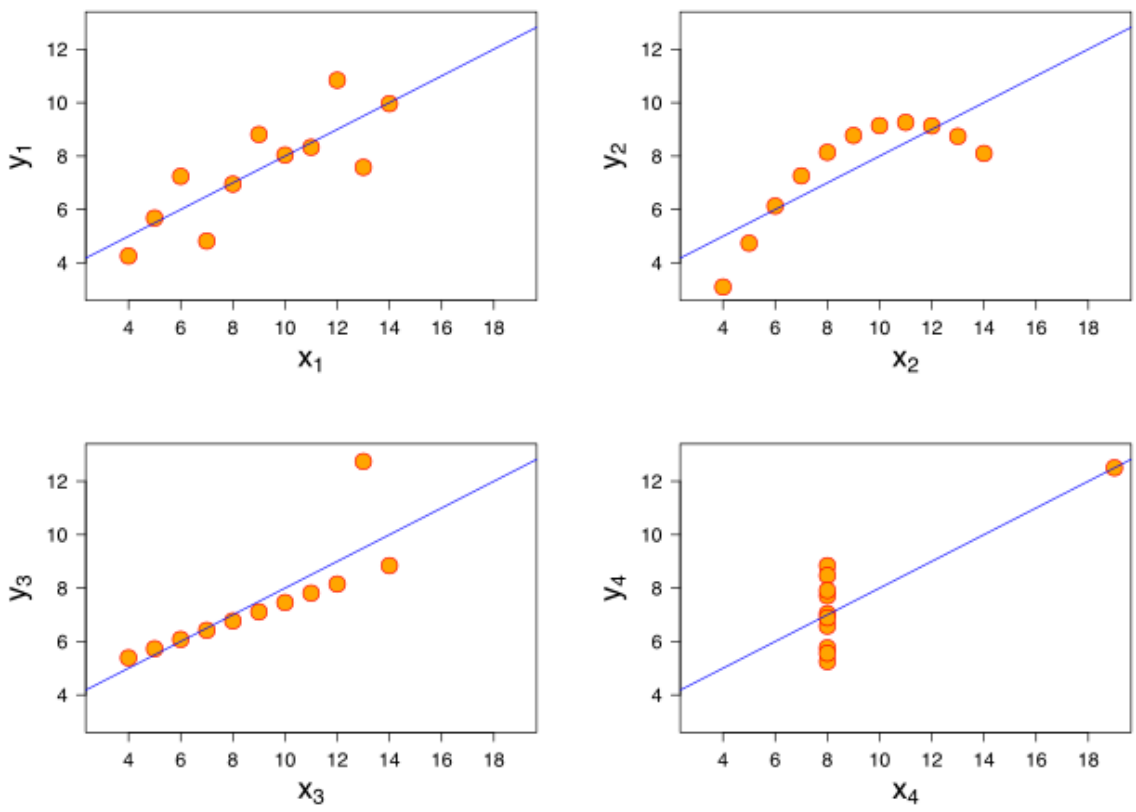
If you fail to reject the null hypothesis that would mean that $\beta 1$ is zero which would simply mean that $\beta 1$ is insignificant and of no use in the model. Similarly, if you reject the null hypothesis, it would mean that $\beta 1$ is not zero and the line fitted is a significant one.
To perform 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 a 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. Let's assume we have 4 data sets each consisting of eleven (x,y) points, they have the 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 **Francis Anscombe** plotted four such plots and it's 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 the scatter plot 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 to 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, also known as the Pearson correlation coefficient, is a statistical measure of the strength and direction of the linear relationship between two variables. It is commonly used to determine how strongly two variables are related to one another.

Interpretation:

1. r = 1: Perfect positive linear correlation (as one variable increases, the other does too).
2. r = -1: Perfect negative linear correlation (as one variable increases, the other decreases).
3. r = 0: No linear correlation.

Properties:

- The value of r ranges from -1 to 1.
- It assumes that the relationship between the variables is linear.
- Pearson's r is sensitive to outliers, which can skew the result.

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**.
Normalized scaling rescales the values between 0 and 1 whereas Standardized scaling typically rescales the values to have mean 0 and a standard deviation 1.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum values are used for scaling	Mean and standard deviation are used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bound to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

   When the **Variance Inflation Factor (VIF)** value is infinite, it indicates perfect multicollinearity between one or more features in a dataset. This occurs when one feature (or variable) can be perfectly predicted by one or more other features. This inflates the variance of regression coefficients and indicates a serious issue with the independence of features in your dataset.

   VIF becomes infinite when there is a Perfect Multicollinearity. In such cases, the R^2 value becomes 1.
   VIF calculation with R^2 = 1:
   Formula for VIF is 1/(1-R^2). If R^2 = 1, the denominator becomes zero, leading to VIF becoming infinite.
   To solve infinite VIF
   Remove one of the highly correlated features: Drop one of the variables from the dataset which is causing this perfect multicollinearity.
   Combine features: If multicollinearity arises due to similar variables, consider combining them into a single feature using dimensionality reduction techniques like Principal Component Analysis (PCA).

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a data set with a theoretical distribution (such as a normal distribution) or to compare two data sets. It helps assess whether the data follows a specific distribution by plotting the quantiles of one distribution against the quantiles of the other.

Use of Q-Q Plots in Linear Regression:

In linear regression, one key assumption is that the residuals (the differences between the observed and predicted values) are normally distributed. A Q-Q plot is often used to check this assumption by comparing the quantiles of the residuals with the quantiles of a normal distribution.

Steps:

- Calculate the residuals from the linear regression model.
- Plot the quantiles of the residuals against the quantiles of the normal distribution.
- Include a 45-degree reference line (representing perfect normality).

If the residuals are normally distributed, the points in the Q-Q plot will fall along the reference line. Deviations from this line indicate departures from normality.

Importance of Q-Q Plots in Linear Regression:

Assessing Normality of Residuals: One assumption of linear regression is that the residuals are normally distributed. Violations of this assumption can lead to incorrect conclusions about the model. A Q-Q plot helps check whether this assumption holds.

Identifying Outliers and Skewness: A Q-Q plot can reveal outliers (points far from the reference line) or skewness (a systematic pattern of points curving away from the line), indicating that the data may not fit the linear regression model well.

Model Diagnostics: By visually assessing the distribution of residuals, Q-Q plots provide insights into whether transformations (e.g., log transformation) or alternative models (e.g., robust regression) might be needed to improve model fit.

In summary, Q-Q plots are an essential diagnostic tool in linear regression, used to check the normality of residuals, identify outliers, and ensure the validity of model assumptions.