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)

Ans: Season, Weather condition, Holiday, months, working day, and weekday were the categorical variables in the dataset. A Boxplot was used to visualise these variables influenced our dependent variable in the following ways

a) **Season**: the box plot reveal that spring season had the lowest value of count while the fall season had the highest value of count summer and winter had count values that were in the middle.

b) **Weathersit**: situations when there is heavy rain or snow there are no users indicating that weather is extremely unfavourable the highest count was observed when the weather forecast was clear partly cloudy.

c) **Holiday**: rentals were found to be lower during the holidays #4

d) **Month**: September had the most rentals while December had the fewest this observation is compatible to the one made in weathersit variable. The weather in December is typically cold and snowy.

e) **Working day**: it had little effect on the dependent variable

f) **Weekday**: weekends saw a significant increase in bike hiring compared to weekdays.

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

Ans: Dummy variables are typically correlated if we don't remove the first column which is redundant. This may have a negative impact on some models, and the effect is amplified when the cardinality is low. Iterative models, for example, may have difficulty convergent, and lists of variable importance may be distorted. Another argument is that having all dummy variables results in multi collinearity between them. We lose one column to keep everything under control.

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

Ans: “temp” and “atemp” variables were highly correlated to with the target variable (“count”).

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

Ans: the distribution of residuals should be normal and centred around zero this mean is 0. Next we take test this residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not. The residuals are scattered around mean = 0 as seen in the diagram (refer Jupyter notebook).

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

Ans: The top 3 predictor variables that influence bike booking, according to our final model, are **Temperature**: with the coefficient of 0. 0654, a unit increase in the temperature variable increases the number of bike rentals by 0. 0654 units

Weather situation: with the coefficient of 0.2332 a unit increase in the winter weather, reduces the number of bike hires by 0. 2332 units as compared to other weathersit (positive correlation). Where weathersit equal to light snow, mistycloud have negative impact on the count (target variable).

Year: with a coefficient of 0. 02993, a unit increase in the year variable increases the number of bike rentals by 0. 02993 counts.

General Subjective Questions

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

Ans:

Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables). It assumes a linear relationship between the input variables and the output variable.

Detailed explanation:

Simple Linear Regression:

Equation:

In simple terms, linear regression is like drawing a straight line to predict one thing based on another. The basic equation looks like this:

Y = B0 + B1X + e

Y is what we want to predict.

X is what we use to make the prediction.

B0 is the starting point of the line.

B1 is how steep the line is.

e is the error or what we couldn't predict.

The goal is to find the best fitting line by adjusting B0 and B1 so that our predictions are as close as possible to the actual values we find B0 and B1 using methods like least squares or gradient descent.

Multiple Linear Regression:

If we have more than one thing to predict from, the equation becomes:

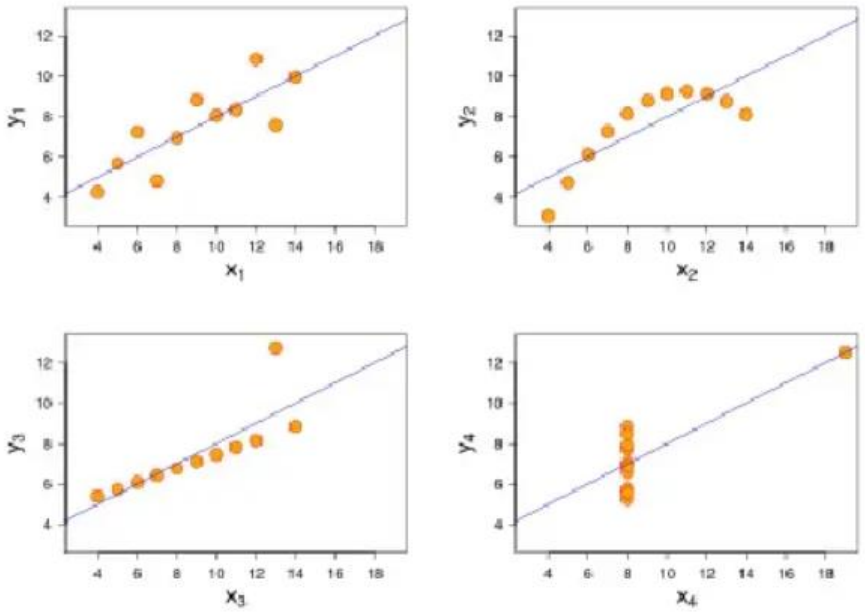
Y = B0 + B1X1 + B2X2 + … + BpXp + e

The idea is the same predict Y using multiple factors the assumptions that are made here are like the relationship with being linear errors being independent and other technical things.

Linear regression is a simple but powerful tool used for predicting outcomes based on input factors.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:



Anscombe's quartet was developed by statistician Francis Anscombe. It includes 4 datasets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasise both the importance of graphing data before analysing it and the effect of outliers and other influential observations.

Statistical properties:

- The first scatter plot appears to be a simple linear relationship.
- the second graph is not distributed normally well there is a relation between them which is not linear.

- In the third graph the distribution is linear but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- finally, the 4th graph shows an example where one high leverage point is enough to produce a high correlation coefficient even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R? (3 marks)
Ans: Pearson's R is a numerical representation of the strength of the linear relationship between the variables. Its value ranges from -1 two plus one. It depicts the linear relationship of two sets of data. In layman's terms it asks if we can draw a line graph to represent the data.
r = 1 Means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Ans: Feature scaling is a method used to normalise or standardise the range of independent variables or features of data. It is performed you during the data pre-processing stage to deal with varying value values in the data set. 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, irrespective of the units of the values.
Normalisation is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-nearest neighbours and Neural Networks.
Standardisation, on the other hand can be helpful in cases where the data follows Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalisation, standardisation does not have a bounding range. So, even if you have outliers in your data they will not be affected by standardisation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans: VIF – Variance Inflation Factor: the VIF indicates how much collinearity has increased the variance of the coefficient estimate VIF = 1 / (1-R^2). VIF = Infinity if there is perfect correlation. Where R - 1 denotes the R-squared value of the independent variable for which we want to see how well it is explained by other independent variables. If an independent variable can be completely described by other independent variables, it has perfect correlation and has an R-squared value of 1. As a result, VIF = 1 / (1-1) provides VIF = 1/0, which is "infinity".

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans: the quantiles of the first data set are plotted against the quantiles of the second data set in a Q-Q graphic it is a tool for comparing the shapes of different distributions A scatter plot generated by plotting two sets of quantiles against each other is known as Q-Q plot.
Because both sets of quantiles came from the same distribution the points should form a line.
That's a fairly a straight line.

The Q-Q plot is used to answer the following questions:
- Do two datasets come from populations with a common distribution?
- Do two datasets have common location and scale?
- Do two datasets have similar distributional shapes?
- Do two datasets have similar tail behaviour?