Assignment-based Subjective Questions & Answers

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

   Some of the inferences made from the analysis of categorical variables from the dataset on the dependent variable(cnt) are:

   - Season 3 i.e., Fall has the highest median, which is expected as the most optimal weather condition for bike ride.
   - Median for bike rents has increased in year 2019, it might be due the fact that bike rentals are getting popular and people are becoming more aware about the environment.
   - People rent more on non-holidays compared to holidays, this might be because they prefer to spend time with family and use personal vehicle instead of bike rentals.
   - Clear weather is most optimal for bike renting, as temperature is optimal, humidity is less and temperature is less.

2. Why is it important to use drop_first=True during dummy variable creation?
   - drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
   - Example

Furnishing status	Furnished	Semi-Furnished
Furnished	1	0
Semi-Furnished	0	1
unfurnished	0	0

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - After data preparation, when we drop registered due to multicollinearity the numerical variable ‘atemp’ and ‘temp’ have the highest correlation with the target variable ‘cnt’.

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

   After building the model on the training set, I have performed

   - Linear relationship between x and y
   - Error terms are normally distributed
   - Error terms are independent of each other
   - Error terms have constant variance.

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 that significantly explain the demand of the shared bikes are: -

   - 'atemp'- feeling temperature in Celsius
   - 'yr'- year (0: 2018, 1: 2019)
   - 'winter'- A subcategory of 'season' (4: winter)

General Subjective Questions

1. Explain the linear regression algorithm in detail

   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

   Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

   Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

   The linear regression model can be represented by the following equation:

   $$Y = \theta_o + x_1 \theta_1 + x_2 \theta_1 \ldots. \theta n \, xn$$

   where, Y is the predicted value θ_o is the constant term.

   $\theta_1,\ldots,\theta n$ are the model parameters

   $x_1, x_2,\ldots,xn$ are the feature values.

   The goal of regression analysis is to create a trend line based on the data you have gathered.

2. Explain the Anscombe's quartet in detail.

   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.

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

   All the summary statistics for each dataset are identical

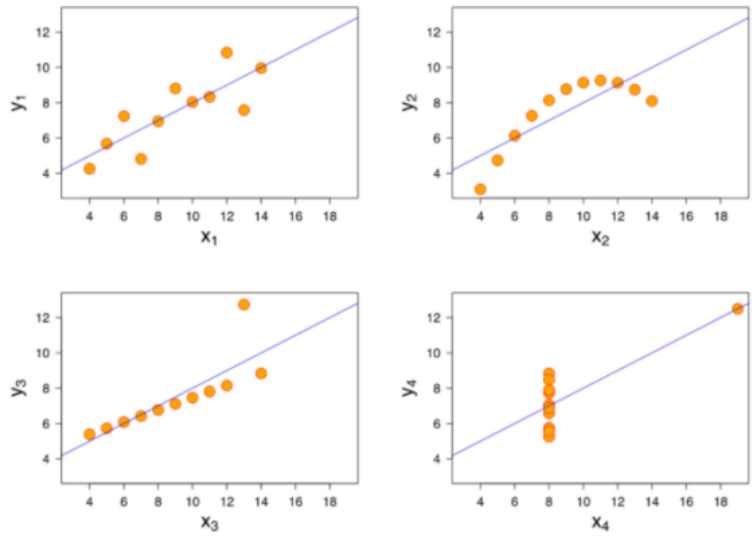
   1. The average value of x is 9.

   2.The average value of y is 7.5.

   3. The variance for x is 11 and y is 4.12

4. The correlation between x and y is 0.816
5. The line of best for is y = 0.5x + 3.

But the plots tell a different and unique story for each dataset.



3. What is Pearson's R?

Pearson's R is a numerical summary of the strength of the linear association between the variables. It varies between -1 and +1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association r > 8 means there is a strong association

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where, N = the number of pairs of scores
Σxy = the sum of the products of paired scores, **Σx =** the sum of x scores
Σy = the sum of y scores, **Σx2 =** the sum of squared x scores
Σy2 = the sum of squared y scores

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

   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.

   It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. Normalized scaling means to scale a variable to have values between 0 and 1, while standardized scaling refers to transform data to have a mean of zero and a standard deviation of 1.

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

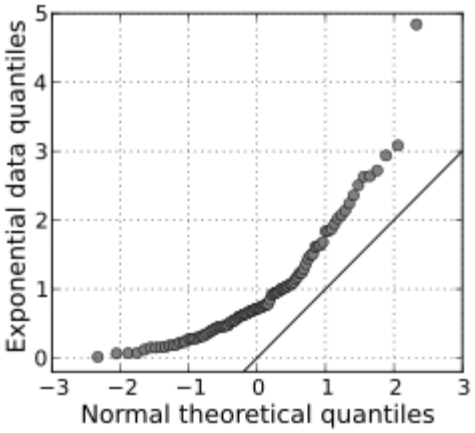
   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   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).

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

   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

   A Q Q plot showing the 45-degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.