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: The categorical variables in the dataset are season, mnth, weekday and weathersit. Based on the data analysis of these categorical variables following are the inferences:
   a. The median count of target variable is much higher for seasons summer and fall, while it is much lower for season spring.
   b. Weathersit value "clear" has the highest median count of target variable cnt. This is followed by weathersit value "misty", and "cloudy". While there are no data points at all for weathersit value "stormy".
   c. The median count of bike sharing is seen to be higher in the months June-September.
   d. Weekday does not seem to have much impact on the count of bike sharing as there is not a big variation on the median values of the cnt target variable across the days of the week.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Ans: When a categorical variable has n number of unique values – all these values can be uniquely represented by n-1 dummy variables (with values either 0 or 1). Whereas the pandas package get_dummies used for getting dummy variables creates n dummy variables – therefore it is important to use drop_first = True so that the first dummy variable will be dropped and we just use an optimal number of necessary variables for the modelling in order to make it computationally efficient.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Ans: As per the pair plot the numerical variable "atemp" has 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? (3 marks)
   Ans: After building the model, the assumptions of linear regression are validated as follows:
   a. **Linearity** is validated with the fact that change in the values of the independent variables causes a proportional change in the value of the dependent variable – this is seen from the coefficients of the dependent variables from the model summary and by validating that the coefficients are statistically significant as seen based on a low p value.
   b. **Independence of the errors** is validated with help of plotting a scatter plot between the residuals and the predicted values of train set. As observed, the variation of residuals around zero does not show a dependency with the value of target variable predicted by the model. This indicates that the errors are independent of the predicted value of target variable.
   c. **Normality of errors** is validated with the help of a "displot" that shows that the errors are normally distributed with the center of distribution around zero.

d. **No perfect multicollinearity** is validated by calculating the variance inflation factor of the independent variables and validating that none of the variables have a VIF above 10.

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: Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:
   a. temp: temperature in Celsius – with a coefficient of 0.5499
   b. weathersit – cloudy – with a coefficient of -0.2871
   c. yr : year (0: 2018, 1:2019) – with a coefficient of 0.2331

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans: Linear regression is a machine learning technique that aims to predict the relation between set of input variables and a single output variable such that the relationship can be represented in a linear plot that can minimize the error between the predicted and real values of the output variable. When the input is a single variable, it is called as simple linear regression, while if the input is a set of variables, then it is called multiple linear regression.
   In a linear regression algorithm, this relationship is represented as:

$$y_P = w_1 f_1(x) + w_2 f_2(x) + w_3 f_3(x) \cdots w_n f_n(x)$$

Here, y_P is the predicted value of the output variable. x is the input set which consists of features $f_1, f_2, f_3 \ldots f_n$ each of these is an independent variable and w_1 to w_n are the coefficient or weights that indicate the proportion in which the input features contribute to a change in the output variable.

The predicted value of output variable in terms of matrix notation can thus be represented as:

$$y_P = w^T X$$

Here X is a matrix consisting of m datapoints with each data having n independent variables. The aim of the linear regression algorithm is to minimize the error in prediction, which is represented as $y_P - y$ where y is the actual value of output variable in training data and y_P is the predicted value by the algorithm.

However, this error can be positive or negative, while our objective is to have a loss/error function that can achieve a global minimum at certain values of the coefficients, therefore we use the mean squared error as the function to be minimized, this is represented as below:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(w^T x_i - y_i)^2$$

The sum of errors can be witted in matrix form as below:

$$\sum_{i=1}^{m}(w^T x_i - y_i) = Xw - y$$

Here X is the input matrix with dimensions m x n (m data points with n independent variables) and w is n x 1 matrix (weights for the n features), thus the dot product of these matrices gives an m x 1 matrix i.e. y_P which has same dimensions as y.

Now, to get the square of the sum of errors, we multiply the matrix with its transpose. Thus MSE can be represented as:

$$E(w) = \frac{1}{m}(xw - y)^T(xw - y)$$

Now, as per rule for a convex function, to find the minima value w that will minimize the above function, we have to find its derivative and equate it to zero

$$E'(w) = \frac{2}{m}\, x^T(xw - y) = 0$$

$$=> x^T(xw - y) = 0$$
$$=> x^T xw - x^T y = 0$$
$$=> x^T xw = x^T y$$

Now to get w, we multiply on both sides by inverse matrix of $x^T x$

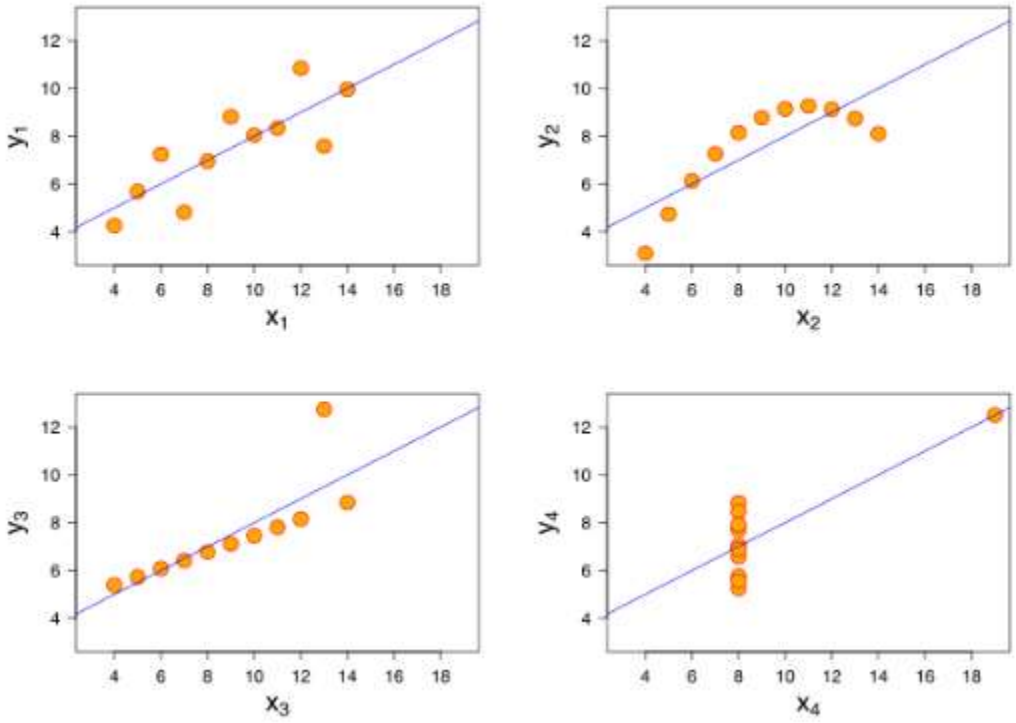
$$=> (x^T x)^{-1}(x^T x)\, w = (x^T x)^{-1} x^T y$$
$$=> w = (x^T x)^{-1} x^T y$$

Thus, the coefficient values given by above formula represent the minima that will minimize the error/loss function and help to get the best fit line as per the linear regression algorithm.
The linear regression algorithm assumes that the errors are normally distributed and are independent of the input or output variables, also the variance of errors is constant.
As part of application of the linear regression algorithm the model is trained on a set of input and output data so as to find the best fit line that minimizes the loss function and then subsequently a new set of inputs is provided to the model to predict the output variable.
If the model is able to predict the output variable with the same level of accuracy (measured by some metrics such as R-squared) as the training data, then the model is said to have generalized well.

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

Ans: Anscombe's quartet includes four data sets that have similar statistical metrics such as mean, standard deviation and best fit regression line, however have a different representation when plotted graphically. It is primarily used to emphasize that the relationship between independent variable and dependent variable must be plotted graphically as well rather than just representing with statistical measures.

Below are the scatter plots of the four data sets:



All of these have the same statistical measures such as:

Mean of x = 9

Mean of y = 7.50

Sample variance of x = 11

Sample variance of y = 4.125

Correlation between x and y = 0.816

Linear regression line represented by y = 3.00 + 0.500 x

R-Squared = 0.67

However as seen from the plots, each data set has differences as marked below:

1. The first plot represents a linear relationship between x and y with the two being correlated variables. The data points seem to be evenly distributed on both sides of the regression line.
2. In the second plot, while there is a relationship between two variables it is not linear in nature. A more general regression is applicable in this case.
3. In the third plot, the modelled relationship is linear however should have had a different regression line as there is a significant outlier and the data points are not evenly distributed on both sides of the regression line.

4. Finally, the fourth plot depicts how one high average point is enough to get a high correlation coefficient even though the other data points do not necessarily show a relation between the independent and dependent variable.

Thus, the Anscombe's quartet stresses on the importance of visualizing the relation between independent and dependent variables graphically to get more clarity on the real influence of the independent variable on the dependent variable.

3. What is Pearson's R? (3 marks)

Ans: Pearson's R or Pearson's correlation coefficient is the most common way of measuring linear correlation. It is a number between -1 and 1, with -1 representing a perfect negative correlation and 1 representing a perfect positive correlation.
Pearson's R represents the strength and direction of the correlation between the two variables.
A value of 0 represents no correlation which means that a change in value of one variable does not influence a change in the other variable in a certain direction.
A value greater than 0 and less than or equal to 1 represents a positive correlation between the two variables, which means that when one variable changes the other variable changes in same direction.
A value less than 0 and greater than or equal to 1 represents a negative correlation between the two variables, which means that when one variable changes the other variable changes in opposite direction.
As a general rule of thumb:
R value > 0.5 – represents strong positive correlation
R value > 0.3 and <= 0.5 – represents a moderate positive correlation
R value > 0 and <= 0.3 – represents a weak positive correlation
R value = 0 – represents no correlation
R value < 0 and >= -0.3 – represents a weak negative correlation
R value < 0.3 and >= -0.5 – represents a moderate negative correlation
R value < -0.5 – represents strong negative correlation

Pearson's R value is an inferential statistic as it can be used to test statistical hypothesis.
Pearson's R value can be used when:
a. Both variables are quantitative
b. Variables are normally distributed
c. The data does not have outliers
d. The relationship is linear

If any of the above rules have a significant deviation then it is better to use a measure other than Pearson's R.
The formula for Pearson's R is given by:

$$r = \frac{[n\sum xy - \Sigma x \Sigma y]}{\sqrt{[n\Sigma x^2 - \sum x^2][n\Sigma y^2 - \sum y^2]}}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Feature scaling is the process by which the independent and dependent variables are scaled appropriately such that they have values lying on a common scale such as between 0 and 1.

In case of practical problems for linear regression the data sets may have independent variables that may have values at very different scales numerically and if the modelling is done with such scales the coefficients for these variables may turn out to be weird and at varied scales – due to which it becomes very difficult to understand the relation strength between the independent variable and the dependent variable.

Therefore, feature scaling is important for two reasons:

1. Facilitating easy interpretation of the coefficients
2. Faster convergence when using gradient descent models

There are two scaling methods most commonly used:

1. Standardized scaling in which the variables are scaled in such a way that their mean is zero and standard deviation is 1. For a variable x the standardized scaled value is given by the formula as below:

   $$x = \frac{x - \bar{x}}{\sigma_x}$$

   Here $\bar{x}$ is the mean of x and σ_x is the standard deviation of x.

2. MinMax Scaling in which the variables are scaled in such a way that the scaled values lie between zero and 1, with zero representing the minimum value for the variable and 1 representing the maximum value for the variable. For a variable x the minmax scaled value is given by the formula as below:

   $$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

A disadvantage of minmax scaling or normalization is that it may lead to loss of some information, especially about outliers.

As standardization scales the data with a mean zero – whereas minmax scaling squeezes the data between 0 and 1 it may cluster the data too close to each other and this may lead to some algorithms like gradient descent to take longer to converge.

A key point to note is that scaling only affects the coefficients and does not impact any statistical metrics such as f-statistic, R-squared etc.

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

Ans: Variance Inflation Factor (VIF) is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine the VIF, a regression model may be fit between the independent variables and the VIF for a variable X_1 will then be represented as below:

VIF_1 = $\frac{1}{1-R_1^2}$

As per the above formula, R_1^2 is the R-squared value of the linear regression model fit to predict the X_1 as output with the other independent variables as input. Now, in case of a significant linear relationship i.e. collinearity between X_1 and the other independent variables the value of R_1^2 will be high and therefore the VIF_1 will be high as a result. When the value of R_1^2 will tend to 1 (perfect linear relationship between X_1 and other independent variables), then the VIF_1 will tend to infinity.
Thus, value of VIF -> infinity indicates a perfect linear relationship between the variable being considered and other independent variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans: A Q-Q plot is a scatter plot created by plotting two sets of quantiles against each other. It is a used to visualize and assess if the two datasets come from populations with a common distribution.
In linear regression, Q-Q plots provide a graphical way of assessing the normality of residuals. Normality of residuals is on of the key assumptions of linear regression and violation of this assumption may lead to incorrect inferences and biased estimates. If the residuals are identified as not normal, this indicates towards a need for further investigation on the model built so that the model fits better.
For plotting a Q-Q plot of the residuals – first the residuals should be ordered from smallest to largest, then for each residual value calculate the quantiles under a standard normal distribution. Thereafter you can plot a scatter plot of the ordered residuals against the corresponding quantiles – this is the residuals Q-Q plot. If the residuals are normally distributed then the points on the Q-Q plot will approximately fall on a straight line. Deviations of the points from straight line indicate a variation from normality of residuals.