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?**

Equation for best fitted line we got is:

cnt= 0.235*yr - 0.0862*holiday + 0.4758*temp - 0.1325*windspeed - 0.1032*season_Spring + 0.0504*season_Winter- 0.0616*mnth_July + 0.0498*mnth_September - 0.2562*weathersit_Light Snow and Light Rain

Inference we got from our model regarding categorical variables:

1. Year (‘yr’) : Positive impact of Year on the dependent variable ‘cnt’ (bike rental count) with a coefficient of 0.235 (With per unit increase in variable ‘yr’, dependent variable ‘cnt’ will increase by 0.235 keeping all others predictors constant)
2. Holiday (‘holiday’) : Negative impact of holiday on bike rental count(cnt) with a coefficient of -0.0862
3. Season:
   a. Spring (‘season_Spring’): Negative impact of spring season on bike rental count with value of coefficient as -0.1032
   b. Winter (‘season_Winter’): Positive impact on bike rental count due to winter season with coefficient value as 0.0504
4. Month (‘mnth’) :
   a. July (‘mnth_July’): Month July is negatively associated with bike rental count(‘cnt’) with coefficient value as -0.0616
   b. September (‘mnth_September’): September month have positive impact on bike rental count with coefficient of 0.0498
5. Weathersit: Light Snow and Light Rain weather have negative impact on dependent variable ‘cnt ‘with coefficient value of - 0.2562

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

It is important to use **drop_first=True** while performing dummy encoding (transforming categorical values to vectors for use in ML algorithms), as it helps in reducing the extra redundant column created during dummy variable creation. As a guiding principle, number of dummy coded variables needed is one less (k-1) than number of possible values k.

For ex:

Categorical variable “**furnishing status**” has 3 levels/possible values (furnished, semi_furnished and unfurnished). If we wish to create a dummy variable for the same, we could do with just 2 instead of 3. See below for detailed explanation

Unfurnished	**Semi-furnished**	
0	0	➔ Furnished
0	1	➔ Semi-Furnished
1	0	➔ Unfurnished

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

Amongst all the numerical variables, temperature (temp) has the highest correlation with the target variable (cnt - bike rental counts) with value of 0.64.

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

To validate the assumptions of Linear Regression:
Residual Analysis of the train data (error terms are normally distributed around the mean value of 0) we can plot histogram using the actual values of the dependent variable of the train dataset and the predicted values obtained from the model.
To evaluate the model, use a scatter plot with the actual values of the test data and predicted values from the model.

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

As per the final model built, top 3 features that are contributing significantly towards the demand of shared bikes:

1. Temperature (‘temp’): Coefficient value of 0.4758
2. Weathersit (‘weathersit_Light Snow and Light Rain’) : Coefficient value of - 0.2562
3. Year (‘yr’): coefficient value of 0.235

cnt = **0.235*yr** - 0.0862*holiday + **0.4758*temp** - 0.1325*windspeed - 0.1032*season_Spring + 0.0504*season_Winter- 0.0616*mnth_July + 0.0498*mnth_September **- 0.2562*weathersit_Light Snow and Light Rain**

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression Algorithm is a machine learning algorithm based on supervised learning.
Linear regression i is used for the prediction\analysis of the target variable with respect to the independent variables available in the data set and there is a linear relationship between the dependent and independent variables. In linear regression on given data set we plot the best fit line for the data points. It is mostly done by the Sum of Squared Residuals Method.

Linear regression model can be represented by the following equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots \beta_n X_n$$

y is the predicted value
β_0 is the intercept
$\beta_1, \beta_2, \dots \beta_n$ are the model parameters
$X_1, X_2, \dots X_n$ are the feature values (predictor values).

When we have one predictor variable it is called Simple Linear Regression.
When more than one predictor, we call it Multiple Linear Regression.
βi are selected by choosing the line that minimizes the squared distance between each y value and the line of best fit. β's are chose such that they minimize the expression.

$$\sum(y_i - (\beta_0 + \beta_i X_i))^2$$

Assumptions in Linear Regression:

1. Relation between dependent and independent variables should be almost linear.
2. Homoscedastic data: error term have same variance (constant variance assumption).
3. Error terms are independent of each other.
4. Error terms are *normally distributed* with mean zero (not X, Y).

In Multiple Linear Regression, we need to handle some issues like Multicollinearity (refers to phenomenon of having related predictor variables in the input dataset). To resolve this, we can use VIF (variance Inflation factor).
Scaling - technique to standardize the independent features present in the data in a fixed range should be performed during the data pre-processing.

Strength of simple linear regression model is explained by R^2 :

$$R^2 = 1 - (RSS / TSS)$$

RSS: Residual Sum of Squares

TSS: Total Sum of Squares

Strength of multiple Linear regression model is measured using Adjusted R^2 (keeping in mind multicollinearity)

$$\text{Adjusted } R^2 = 1-((1-R^2)(N-1))/(N-p-1)$$

Adjusted R^2 adjusts the value of R^2 such that a model with a larger number of variables is penalized.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet - To describe the significance of visualization of data). It is group of four data sets which are nearly identical in simple descriptive statistics (**variance**, **mean** of all x, y points in all four datasets, correlation coefficients), but when plotted on scatter plot they appear different and have different distributions.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs (visualizing the data) before analyzing and model building.

This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson's R?

Correlation is a measure of how well two quantitative, continuous variables are related. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables. It shows the linear relationship between two sets of data. In simple terms, it answers the question, *Can I draw a line graph to represent the data?*

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables.

It can have values from -1 to 1

r = -1 (negative correlation) indicates that data lie on a perfect straight line with a negative slope, which means as one variable increases other decreases and vice versa.

r = 0 indicates no linear relationship between the variables

r = 1 indicates a positive correlation which means that both variables increase and decrease together (data for this lies on a straight line with positive slope).

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

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. If not scale, the feature with a higher value range starts dominating i.e. ML algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using scaling method (if the data is not scaled) then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes or range and thus, tackle this issue.

Another reason why feature scaling is applied is that few algorithms like neural network gradient descent converge much faster with feature scaling than without it.

Two most important techniques to perform Feature Scaling:

Normalization (Min-Max Normalization): This technique rescales a feature values in such a way that they are between 0 and 1.

$$X_{new} = (X_i - \min(X)) / (\max(X) - \min(X))$$

This technique responds well if the standard deviation is small and when the distribution is not Gaussian .This technique is sensitive to outliers.

Standardization: Effective technique which rescales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = (Xi- X_{mean}) / \text{Standard Deviation}$$

This technique assumes data is normally distributed within each feature. If data is not normally distributed, this is not the best scaler to use.

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

Variance Inflation Factor (**VIF**) detects multicollinearity (when there is correlation between predictors in a model); its presence can affect our regression results. VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to multicollinearity. It is calculated using the below formula:

$$VIF = 1 / (1 - R^2)$$

VIF calculates how well one independent variable is explained by all the other independent variables combined.Now if there is **perfect correlation** between the variables i.e. R square will have the value as 1, the value for VIF will be 1 / (1-1) i.e. infinity.

When faced to multicollinearity, the concerned variable should be removed, since the presence of multicollinearity implies that the information provided by this variable about the response is redundant in the presence of other predictor variables

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

QQ Plots (Quantile - Quantile plots) are plots of two quantiles against each other. (A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set) A quantile is a fraction where certain values lie below that quantile. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. For example, the median is a quantile where 50% of the data lie below that point and 50% lie above it. The purpose of Q Q plots is to find out if two data sets come from populations with the same distribution.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

References:

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

https://boostedml.com/2019/03/linear-regression-plots-how-to-read-a-qq-plot.html

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm#:~:text=By%20a%20quantile%2C%20we%20mean,reference%20line%20is%20also%20plotted.