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?

Ans – From the analysis of the categorical variables, the following inferences could be made on the effect of these features on the dependent variables:

a) The total bikes rented increased as we went from 2018 to 2019. This could be because of more advertisement, more popularity about the service and, the service being accessible in a greater number of areas and, an increase in the inventory.

b) The demand of the bike rental is way more on a clear or a partly cloudy day than a rainy day, or when there is snowfall.

c) The demand for the shared bikes has a positive correlation with the feature atemp.

d) The median of the demand for rental bikes is lower on holidays than other days.

e) The demand was the highest during the summer and the fall seasons and the lowest during spring.

f) The demand was much higher during the summer and fall months (April-Oct) than the other months i.e., during the summer and fall seasons as observed earlier.

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

Ans – When creating dummy variables for 'n' levels, the number of dummy variables required to explain all the information in all the categories/levels is 'n-1'. So, having 'n' dummy variables becomes redundant and hence, the drop_first=True parameter is used to drop the first dummy variable. This ensures that there are 'n-1' dummy variables preventing redundancy. This also ensures that there are fewer input features thus, reducing computational time and model complexity and increasing efficiency.

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

Ans – The pairwise scatterplots and the heatmap both revealed the feature '***atemp***' ***(Feeling temperature)*** to have the highest correlation with the target variable and this correlation is positive.

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

Ans – The assumptions of Linear Regression are:

a) Linear relationship between X and y – This was validated through scatterplots and regression plots which revealed certain features to have a linear relationship with the target variable.

b) The residuals (error terms) have a normal distribution with approximately zero mean – This was validated by calculating the error terms (y_true – y_pred) and then plotting a distribution plot of the error terms/residuals.

c) Error terms should be random and independent of each other – This was validated by plotting scatter plot of error terms/residuals vs the y_true values.

d) Homoscedasticity (error terms have constant variance) – This was validated by plotting a scatterplot of error terms/residuals vs the predicted y values (y_pred).

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

Ans – Apart from the constant, the final model reveals the top three features contributing to explaining the demand of shared bikes to be:

a) ***atemp*** – The feeling temperature has the most significant contribution in explaining the demand and has a positive relation with the output.

b) ***weathersit_Light_Rain_Snow*** – The weather condition of light rain, light snowfall or thunderstorms is the second most significant feature in explaining the demand and it has a negative relation with the output.

c) ***yr*** – The year feature is the third most significant feature in explaining the demand for shared bikes and has a positive relation with the output.

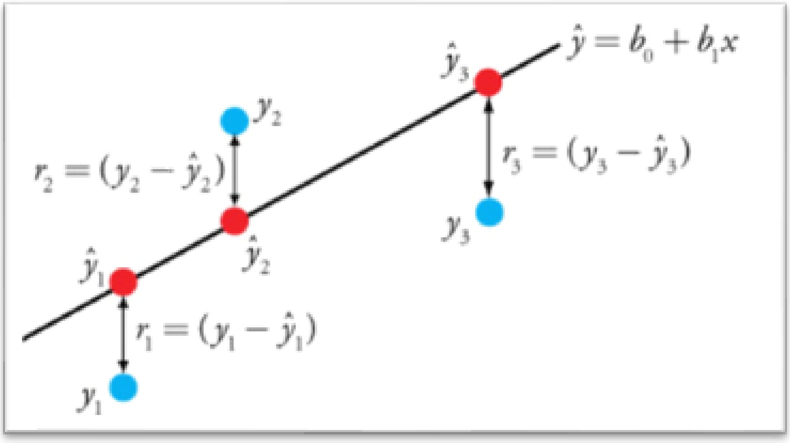
General Subjective Questions:

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

Ans – Linear Regression is a predictive machine learning algorithm that models/learns the linear relationship between the input/independent variables and the target/dependent variable. It is a form of supervised machine learning. If there is only one independent variable it is called *Simple Linear Regression,* and if there are multiple independent variables, it is called *Multiple Linear Regression.*

A linear relationship implies a constant straight-line relationship. In real world data, a perfect straight-line relationship between the inputs will not exist. However, there can exist some linear relationship between the inputs and the output, and this relationship can be explained approximately by a “best-fit” line. This is what Linear Regression tries to determine i.e., finding the equation of the best fit line to describe the relationship between the inputs and the output as accurately as possible. Once, the equation and thus, the coefficients of the inputs are determined, these coefficients can then be used to predict the output variable for different values of input variables.

The best-fit line is determined (i.e., calculation of coefficients of the input variables) by reducing the normal distance of the data points from the straight line passing through those data points. This distance/error is also known as the residual error. The figure below shows the residual errors r_1, r_2 and $r_{3.}$


$\hat{y} = b_0 + b_1 x$
$\hat{y}_3$
y_2
$r_2 = (y_2 - \hat{y}_2)$
$r_3 = (y_3 - \hat{y}_3)$
y_3
$\hat{y}_2$
$\hat{y}_1$
$r_1 = (y_1 - \hat{y}_1)$
y_1

The general equation for Linear Regression is:

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

where,

$\hat{y}$ = Output/Dependent variable

x_i = Input/Independent variables

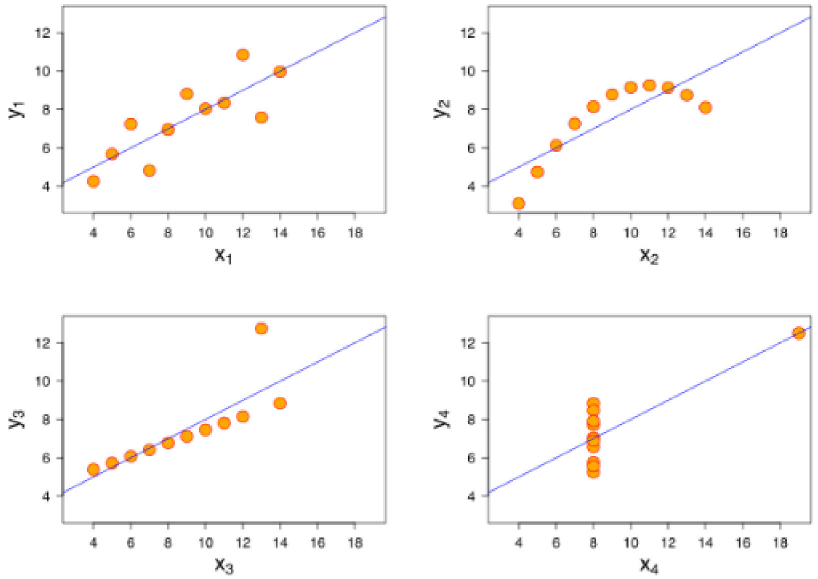
β_0 = y-intercept/constant term

β_i = Coefficients of x

The method used to minimize this error is known as the Ordinary Least Squares. Ordinary Least Squares works by minimizing the sum of the squares of the residual errors i.e., the differences between the observed dependent variable in the given dataset and those predicted by the linear equation/function. Another method used to minimize this error is known as the Gradient Descent algorithm in which a cost function is iteratively minimized until a global minima is reached. There are various cost functions which can be used, some which are: Mean Squared Error (MSE). Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

2) Explain the Anscombe's quartet in detail.

Ans – Anscombe's quartet is a set of four different sets of data points each of which have the same summary/descriptive statistics – For each of these datasets, the means and variances of x and y, and the correlation between x and y are the same. Also, from the plots shown below, the best regression lines are also the same, but the data points differ vastly.



From the figure, we see that only the first dataset appears to be a simple linear relationship. The other plots reveal that they are not suitable for linear regression as they don't have a linear relationship.

The quartet is 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 statistic properties for describing realistic datasets.

3) What is Pearson's R?

Ans – Pearson's R or Pearson's Correlation Coefficient is a measure of the linear correlation between two sets of data. By formula, it is the covariance of two variables, divided by the product of their standard deviations. Covariance is the joint variability of two random variables, i.e., a measure of the relationship between two random variables and to what extent, they change together. The formula for Pearson's R is given below:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt[2]{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt[2]{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where,

x_i = sample values of x

$\bar{x}$ = mean of x

y_i = sample values of y

$\bar{y}$ = mean of y

A positive value of Pearson's R indicates positive correlation between the features, whereas a negative value indicates a negative correlation between the features. Also, a value of 0.3-0.5 indicates moderate correlation and anything higher than that indicates a strong correlation.

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

Ans – Scaling or Feature Scaling is the process of normalizing the range of the features in a dataset. In other words, it involves transforming the scale (range of values) of each feature to the same scale or range.

Input features can have different units and hence, different scales and distributions. As such, it may make it difficult to model the relationship between the features. An example of this is that large input values (e.g., a spread of hundreds or thousands of units) can result in a model that learns large weight values. A model with large weight values is often unstable, meaning that it may suffer from poor performance during learning and sensitivity to input values resulting in higher generalization error. Thus, scaling is an important pre-processing step performed for the following reasons:

a) Scaling helps the model converge faster on the minima.
b) Having the features in the same scale helps in interpreting the weights or coefficients properly
c) Helps in preventing poor model performance due to large spread of values of input features.

Normalized scaling refers to scaling the input features to the range between 0 and 1. The min-max scaling technique is a normalized scaling technique. The formula for min-max scaling is given below:

$$x_{norm} = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$$

where,

x_i = i^{th} sample of x

x_{min} = min value of x

x_{max} = max value of x

Standardized scaling is a scaling technique that refers to centering the distribution of the data around the value 0 and the standard deviation to the value 1 i.e., transform the data in such a way that the mean becomes 0 and standard deviation becomes 1. The formula for standardized scaling is given below :

$$z = \frac{(x - \mu)}{\sigma}$$

where,

x = values of feature x

μ = mean of x

σ = standard deviation of x

Standardization is a scaling technique that assumes that the data conforms to a normal distribution, whereas Normalization is a scaling technique that does not assume any specific distribution. If a given data attribute is normal or close to normal, this is probably the scaling method to use. If the data is not normally distributed, consider normalizing it prior to applying your machine learning algorithm.

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

Ans – VIF stands for Variance Inflation Factor which is measure of multicollinearity in the input features i.e., how much variation in an input feature can be explained by the other input features. Multicollinearity exists when there is a correlation between the input/independent features in a multiple linear regression model. The formula for VIF is:

$$VIF = \frac{1}{1 - R^2}$$

where,

R^2 = R-squared value of coefficient of determination

The formula for R-squared is:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where,

RSS = Residual sum of squares and is given as $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$

TSS = Total sum of squares and is given as $\sum_{i=1}^{n}(y_i - \bar{y})^2$

For input features which have high a high multicollinearity with other input features, the RSS value will be close to 0. Hence, The R-squared value will be close to 1. As a result, the denominator of VIF will tend to 0, causing the VIF to have an infinitesimally large value or essentially a value of infinity. Thus, an infinite VIF value implies that the given input feature is highly correlated to the other input features. In other words, almost all the variation in this input variable can be explained by the other input variables.

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

Ans – A quantile-quantile plot, also known as a Q-Q plot is a probability plot that is used to compare two probability distributions by plotting the quantiles of the first data set against the quantiles of the second data set. It helps us assess if a set of data came from some theoretical distribution like Normal, Uniform, etc., and can also help us determine if two datasets come from populations with a common distribution, have a common location and scale, or have similar tail behavior.

In case of linear regression, the Q-Q plot helps us in the scenario when we have training and test data sets received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

In a Q-Q plot, in addition to the quantiles, a 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points in the Q-Q plot should fall approximately along this reference y=x line. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line but not necessarily on the y=x line. If the two data sets have come from populations with different distributions, then the data points will be far from the reference line.