**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)**

The demand of bike is less in the month of spring when compared with other seasons

Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.

Bike demand is less in holidays in comparison to not being holiday.

The demand of bike is almost similar throughout the weekdays.

There is no significant change in bike demand with working day and non working day.

1. **Why is it important to use drop\_first=True during dummy variable creation? (2 mark)**

To encode categorical data, one hot encoding is done, where a dummy variable is to be created for each discrete categorical variable for a feature. This can be done by using pandas.get\_dummies() which will return dummy-coded data. Here we use parameter drop\_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.

If we do not use drop\_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

Mathematically it can be explained by considering a regression model which is used to find

population rise in 3 different states as below where X3 represents the 3 different states name.

Y = β0 + β1 (X1) +β2 (X2) +β3(X3) + ε ---(i)

As X3 is a categorical variable that contains 3 different state names, we can assign 3 dummy variables D1 as [100], D2 as [010], and D3 as [001] for each state in our equation.

D1 + D2 + D3 = 1

D3 = 1- (D1 + D2) ---(ii)

The last equation indicates D3 is perfectly explained by the other two dummy variables D1 and D2.

Assigning dummy variables to equation (i), we get

Y = β0 + β1 (X1) +β2 (X2) +β4D1 + β5D2 + β6D3 + ε

Y = β0 + β1 (X1) +β2 (X2) +β4D1 + β5D2 + β6 (1- (D1 + D2)) + ε [using equation (ii)]

Y = β0 + β1 (X1) +β2 (X2) +β4D1 + β5D2 + β6- β6D1+β6 D2+ ε

Y = (β0 + β6) + β1 (X1) +β2 (X2) +(β4 - β6) D1+ (β5- β6) D2+ ε

Y = β0! + β1 (X1) +β2 (X2) +β4! D1+β5! D2+ ε ---(iii)

So from equation (iii) coefficient β0, β4 and β5 are impacted by D3 coefficient β6. So,

one dummy variable should be dropped to avoid being trapped by the dummy variable.

In a broader sense, we can conclude that if there are n dummy variables, n-1 dummy  
variables will be able to predict the value of the nth dummy variable, so one dummy variable should be dropped to avoid multicollinearity.

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

The above Pair-Plot tells us that there is a LINEAR RELATION between 'temp'and'atemp'

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

We can validate the assumptions of Linear Regression after building the model on the following training set by below method:

1)Fitted regression line is linear.

2)Error terms came out normally distributed with mean as 0.

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

Based on final model top three features contributing significantly towards explaining the demand are:

* 1. Temperature (0.552)
  2. weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)
  3. year (0.256)

General Subjective Questions

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

Answer: Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

Y= b0+b1X1+b2X2+….

In the example above, y is the dependent variable, and x1, x2, and so on, are the explanatory variables. The coefficients (b1, b2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

1. **Explain the Anscombe’s quartet in detail. (3 marks)**

Anscombe’s quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

1. **What is Pearson’s R? (3 marks)**

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

* Pearson’s r
* Bivariate correlation
* Pearson product-moment correlation coefficient (PPMCC)

The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

| Pearson correlation coefficient (r) value | Strength | Direction |
| --- | --- | --- |
| Greater than .5 | Strong | Positive |
| Between .3 and .5 | Moderate | Positive |
| Between 0 and .3 | Weak | Positive |
| 0 | None | None |
| Between 0 and –.3 | Weak | Negative |
| Between –.3 and –.5 | Moderate | Negative |
| Less than –.5 | Strong | Negative |

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

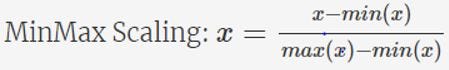
1. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

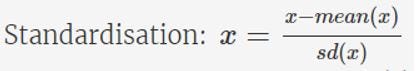
Normalization/Min-Max Scaling:

* It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



Standardization Scaling:

* Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).



* sklearn.preprocessing.scale helps to implement standardization in python.
* One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

(3 marks)

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 VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X\_1=C+ α\_2 X\_2+α\_3 X\_3+⋯

〖VIF〗\_1=1/(1-R\_1^2 )

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

X\_2=C+ α\_1 X\_1+α\_3 X\_3+⋯

〖VIF〗\_2=1/(1-R\_2^2 )

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behaviour

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis