**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 : There were 11 independent variables identified which has impact on the dependent variable cnt. Below are the variables.

windspeed, year, season spring,Weather2(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) and weather3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds), months (march,may,jun,august,september) and holiday.

All the variables have a P value of less than 0.05 which means these are highly significant.

VIF also has the values less than 5, which tells us that the final variables selected in the model are independent and don't have any correlation with other variables.

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

Ans : For any variable for which we need to create dummies , always N-1 dummies are sufficient. Since the presence of N-1 dummies itself signifies the value of the Nth dummy which has been dropped. This is done to avoid creating a multicollinearity issue between variable which might lead to errors and OLS not being correct. When there is a large number of dummies which can be created, dropping even 1 would cause considerable impact on the model. We can drop any of the dummy variables, it's not mandatory to drop First one .

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

Temp and atemp seems to have high linearity , but when using heat map, season and year also shows high correlation.

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

We can validate the assumptions by checking the p values of the variables.

Though temp and atemp has high correlation with target, but these variables also had high correlation with other variables like spring, which was finally selected in the model with p value of 0.00

Spring and Cnt also had a high correlation of 0.53.

Cnt and yr had correlation of 0.57. Yr was also selected in the final model with p value of 0.00.

Windspeed has correlation of 0.25, which was selected in the final model with p value of 0.00

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

Ans - Year, spring and weather 3 have high coeff of 0.24 , 0.23 and 0.28 respectively

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Ans : Linear regression algorithm is Machine Learning algorithm that computes the linear relationship between dependent or target variable with one or more independent variables.

This is a type of Supervised Learning algorithm which can be used when the target variable is a continuous variable.

The purpose of the algorithm is to find the best fit line or best fit equation that can predict the values of a target variable based on the one or more independent variables.

The slope of the line indicates the change in target variable based on the unit change in independent variables.

Linear regression can be of two types:

1. Simple linear regression - This has 1 target variable and one independent variable.
2. Multiple linear regression - This has 1 target variable and 1 or more than 1 independent variable.

This helps to identify the response or change in the target variables based on the variation of the feature variables.

Linear regression is performed on Train and test datasets having Y as target variable and X as one more independent variable.

First the model is set to learn or predict the Y based on the training dataset, recursively adding or deleting the features to get the best fit equation or line.

Once we get the best fit equation on the training data set, the model is used to predict Y on the test dataset.

Best fit line is identified by below equation:

$Y = \beta_0 + \beta_1 X$

Where Y is target variable

$B_0$ = constant/intercept

$\beta_1$= Slope or gradient

X is dependent variable

For multiple linear regression:
$Y = \beta_0 + \beta_1 X1 + \beta 2 X2 \ldots . \beta n Xn$

The strength of the regression model is assessed using below 3 parameters:

1. R2 or coefficient of determination - This measures the goodness of fit of a model. It is the statistical measure of how well the regression predictions approximate the read data points.
   R2 is calculated based on least Sum of squares method.r2 increases as the number of variables increases in the model
   It always takes a value between 0 and 1. Overall the higher the R2 better is the model.
   Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$
   R2 has some problems:
   Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently, a model with more terms may appear to have a better fit simply because it has more terms.
   If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as overfitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions. To overcome this we can use Adjusted R2
2. Adjusted R2 : The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors.
   The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

3. Variance inflation factor VIF :
   Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.
   VIF measures the number of inflated variances caused by multicollinearity.
   VIF basically helps explain the relationship of one independent variable with all the other independent variables. The formula of VIF is given below:

$$VIF_i = \frac{1}{1 - R_i^2} = \frac{1}{\text{Tolerance}}$$

Steps to perform Linear Regression:
Linear regression can be done manually or combination of automated and manual approach.

When the number of variables are high, then manual approach is very difficult and time taking.
So, we use combination of automated and manual approach.
Below are the steps:

1. Analyze the data and do cleanup
2. Create dummies for the categorical variables to convert them to numeric, since linear regression models can not be made on categorical values.
3. Perform scaling on numeric variables to bring all the variables in the same scale. We can either use MinMax Scaling or transformation
4. Use RFE - recursive feature elimination to get some list of features based on which we can go with the calculation of model summary
5. Create a linear model and check the summary.
6. Check Adjusted R2 and P values of all the predictors
7. If any of the predictors has the high P value more than 0.05, then drop that predictor and create the model again.
8. Repeat this step unless we have all the predictors having p value less than 0.05.
9. Finally check the VIF . if any predictor has VIF more than 5, it indicates the predictor has high correlation with other predictors. Remove this predictor and recreate the linear model, until all the predictors have p values less than 0.05 and VIF less than 5.
10. Use this final model on the test set and calculate predicted values of target variable using the trained model.
11. Also, calculate r2 for test data using a trained linear model. If R2 of test and train is close, then the model seems to be a good fit.
12. Finally we can plot graph between predicted values vs actual values to check if the residual is a linear fitted line or not.

2. Explain the Anscombe's quartet in detail.

Ans : Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. When we plot the four data sets the plot looks different.
All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.
Each dataset consists of eleven x,y points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other observations on statistical properties.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)

3. What is Pearson's R?

Pearson's correlation coefficient, also called correlation coefficient, a measurement of the strength of the association between two variables. Pearson's correlation coefficient r takes on the values of −1 through +1. Values of −1 or +1 indicate a perfect linear relationship between the two variables, whereas a value of 0 indicates no linear relationship.

The Pearson's correlation coefficient formula is

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

In this formula, x is the independent variable, y is the dependent variable, n is the sample size, and Σ represents a summation of all values.

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

Ans : Feature scaling is a way of transforming your data into a common range of values.

Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead to a model with very weird coefficients that might be difficult to interpret.

So we need to scale features because of two reasons:

1. Ease of interpretation 2. Faster convergence for gradient descent method

There are two common scalings:

1.Standardizing :Standardization takes our data and makes it follow a Normal distribution, usually of mean 0 and standard deviation 1. It is best to use it when we know our data follows a standard distribution or if we know there are many outliers

2.Normalizing or Min Max scaling:

This scales our features to a predefined range (normally the 0–1 range), independently of the statistical distribution they follow. It does this using the minimum and maximum values of each feature in our data set.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans : An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.
Infinite VIFs tells there is perfect collinearity.
Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. 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.

If we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.
A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions.

As we are plot the graph between the actual and observed residuals, and if the linear model is good enough then the distribution of the actual and observed residuals statistic variable will be very much similar, so by using QQ plot we can determine this similarity between the distribution and thus verifying our linear model

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 the same distributions.

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 behavior