Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Demand is continuously growing each month till June whereas September month has highest demand and After September, demand is  decreasing. demand has decreased on holidays. During September, bike sharing is more. During the year end and beginning, it is less, could be due to extreme weather conditions.

Demand for the next year is seems to be increased. May be due to weather conditions Bike sharing has been reduced during the beginning and end of the year.

Q2. Why is it important to use drop_first=True during dummy variable creation?

It helps in reducing the extra columns created during dummy variable creation. Lets suppose if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Looking at the pair plot among the numerical variables temp and atemp columns has the highest correlations with target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

1.linear relationship: by plotting pair wise scatter plots as it indicates the linear relationship

2. homoscedasticity: I draw residual plot and verified that the variance of the error terms is constant across the values of the dependent variable.

3. checking the normality of errors: to validate this I draw the distribution of residuals against levels of dependent variables.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Below are the significant features to predict the demand.

1.holiday

2.temp

3.humidity

Q1. Explain the linear regression algorithm in detail.

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of data based on some variables. In the case of linear regression, the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

an example is let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$Y=a + b X$, where a is the y-intercept of line and b is slope. X is the independent variable and Y is the dependent variable.

Here are some Use Cases of Linear Regression:

Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.

Price Prediction – Using regression to predict the change in price of stock or product.

Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.

However, Linear Regression is a very vast algorithm and it will be difficult to include all here. You can improve the model in various ways could be by detecting collinearity and by transforming predictors to fit nonlinear relationships. This article is to get you started with simple linear regression. Let's quickly see the advantage and disadvantage of linear regression algorithm:

Linear regression provides a powerful statistical method to find the relationship between variables. It hardly needs further tuning. However, it's only limited to linear relationships.

Linear regression produces the best predictive accuracy for linear relationship whereas its little sensitive to outliers and only looks at the mean of the dependent variable.

Q2. Explain the Anscombe’s quartet in detail

Anscombe’s quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

Q3. What is Pearson’s R?

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

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.

**Min Max Scaling:**

It brings all data in the range of 0 to 1. Below is the code to use in python to implement this scaling.

**sklearn.preprocessing.MinMaxScaler**

Min Max Scaling: x = x - min(x)/ max(x)-min(x)

**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). Below is the syntax used to implement same in python.

**sklearn.preprocessing.scale**

Standardization: x = x – mean(x) / Sd(x)

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

In case of perfect corelation between two independent variables, generally we observe infinite VIF.

In such scenario R2 is equals to 1 and as per the formula 1/ 1-1 equals to 1/0 that is infinity. We can resolve this issue by dropping any one of the columns which is causing perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Q6. 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.