1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

->

- Weather: Almost 67% of the bike booking were happening during clear weather with a median of around 4900.
- Season: Fall and Summer seasons have a median of more than 5000 which indicates that weather shows some trend in bike booking.
- Weekday: All weekday shows a similar trend which indicates that they have a minor influence on prediction
- Month: From May to October have a higher booking which indicates that it can be a good indicator for prediction.
- Year: Year on yearly basis there is an increase in demand which might be because bike rentals are getting popular.
- Working day: There is higher demand on working days than non-working which might be due to they are the using rental bikes for daily office commutes.

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

-> If we do not drop the feature, we will get one redundant feature which will create a multi-collinearity issue. But blindly dropping the first column is not a solution we can drop a feature that has a low correlation with the target variable.

Example: In the weather column we have four values and alphabetically first value will be a clear sky which will eventually get dropped in the end if we follow the drop first approach, we will not be able to draw conclusions based on that important feature.

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

-> Temp and atemp have the highest correlation with the target variable.

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

->

- Vif: VIF<=5 implies no multicollinearity, whereas VIF>=5 implies serious multicollinearity.
- Homoscedasticity: Homoscedasticity means the residuals have constant variance at every level of x. The absence of this phenomenon is known as heteroscedasticity. Heteroscedasticity generally arises in the presence of outliers and extreme values. Create a scatter plot that shows residual vs fitted value. If the data points are spread across equally without a prominent pattern, it means the residuals have constant variance (homoscedasticity).
- Normal Distribution of error terms: The distribution of the error terms is normally distributed and didn’t find any pattern in the distribution.

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

->

- Temperature: Most important factor affecting demand is temperature. With a coefficient of 0.5502, for every change in temperature of 1 degree, demand increases by a factor of 0.5502 (temperature x 0.5502).
- Light snow: Second most important factor is Light Rain or Snow with a coefficient of (-0.2848) Hence if a particular day has light rains, it is expected to reduce the demand by 28.48%.
- Year: Third most important factor is the year with a coefficient value of 0.2365.

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

-> Linear regression shows the relation between a set of independent variables to a dependent variable or in simple terms it is a method of finding the best straight line between the dependent and independent variable.

- Simple Linear Regression: when we draw a line based on a single independent variable then it is called Simple linear regression.
- Multiple Linear regression: when we draw a line based on multiple independent variables then it is called multiple linear regression.

Mathematically, we can write a linear regression equation as:

$y = mx + c$

Where m and c have given by the formulas:

$m(slope) = (n\Sigma xy - \Sigma x\, \Sigma y) / (n\Sigma x^2 - (\Sigma x)^2)$

$c(intercept) = (n\Sigma y - b(\Sigma x)) / n$

Here, x and y are two variables on the regression line.

1. m = Slope of the line
2. c = y-intercept of the line
3. x = Independent variable from dataset
4. y = Dependent variable from dataset

2. Explain the Anscombe's quartet in detail. (3 marks)

-> It can be defined as a group of four data sets that are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Because there are two points for the same x value and equidistance y values on the opposite side of the line nullifying the effect so there is no impact on the line.

It tells us the importance of visualizing the data before applying various algorithms.

The four datasets were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

- Applications: The quartet is still often 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? (3 marks)

-> It is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviation.

Mathematically, we can write Pearson's R can be represented as:

- Correlation = covariance (gen x, gen y) /(std(x) * std(y))

Covariance numbers can be any value between positive and negative infinity.

When all the data fall on a straight line then covariance and product of the std terms are the same and division gives us -1 or 1.

When data do not fall on a line then covariance and product of the std terms give a value close to zero.

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

->

- Scaling is applied to independent variables to normalize the data within the range. It helps to reduce the time required for calculation.

A. Scaling range

Normalization: Scales values between [0, 1] or [-1, 1].

Standardization: It is not bounded to a certain range.

D. Outliers.

Normalization: It is really affected by outliers.

Standardization: It is much less affected by outliers.

B. Formulae

Normalization: Minimum and maximum values of features are used for scaling.

$$X\_new = (X - X\_min)/(X\_max - X\_min)$$

Standardization: Mean and standard deviation are used for scaling.

$$X\_new = (X - mean)/std$$

D.

Normalization: It is often called Scaling Normalization

Standardization: It is often called Z-Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

-> VIF is the measure of how much the variance of an estimated regression coefficient increases due to collinearity.

- Formulae :(VIF) = 1 / (1- R2 ^ 2)

where the R2 score is the proportion of the variance in the dependent variable that is predictable from

the independent variable.

There are three cases of R2 score

1. If the R2 score is 0

   In this case, VIF will be 1 as the variables are independent.

2. If the R2 score lies between 0 and 1.

In this case, VIF will be greater than 1. The high value of VIF indicates that there is high multicollinearity.

3. If the R2 score is 1.

   If all the features are dependent then VIF is 1.

In the last case, it will be infinite and it indicates a large value of VIF indicates that there is a high correlation between the variables

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

(3 marks)

--> Q-Q plots are the plots of two quantiles against each other.

QQ plots are very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it is met using this.
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

Purpose: It is used to find whether two data sets come from the same distribution.

A 45-degree angle is plotted on a Q-Q plot and if the data set come from a common distribution

then most of the points will fall on that line.

It is used to compare properties such as location, scale, and skewness by

the shapes of distribution providing a graphical view.