ignment-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?

answer-> •Bike demand in the fall is the highest.
•Bike demand takes a dip in spring.
•Bike demand in year 2019 is higher as compared to 2018.
•Bike demand is high in the months from May to October.
•Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.
•The demand of bike is almost similar throughout the weekdays.
•Bike demand doesn’t change whether day is working day or not

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

answer-> •It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

•For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don’t need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

•It is also used to reduce the collinearity between dummy variables.

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

answer-> •atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

answer-> assumption is validated

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

answer->Temp, Year, and snowy and rainy weather

General Subjective Questions

1). Explain the linear regression algorithm in detail.

answer-> Linear regression is one of the very basic forms of machine learning in the field of data science where we train a model to predict the behaviour of your data based on some variables.In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Regression analysis is a statistical technique for investigating and modeling the relationship between variables . Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique.

As an example of a problem in which regression analysis may be helpful, suppose that an industrial engineer employed by a soft drink beverage bottler is analyzing the product delivery and service operations for vending machines. He suspects that the time required by a route deliveryman to load and service a machine is related to the number of cases of product delivered. The engineer visits 25 randomly chosen retail outlets having vending machines, and the in - outlet delivery time (in minutes) and the volume of product delivered (in cases) are observed for each.

2. Explain the Anscombe's quartet in detail.

answer-> 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

Arguing for Graphics in 1973->

In 1973, Francis J. Anscombe published a paper titled, Graphs in Statistical Analysis. The idea of using graphical methods had been established relatively recently by John Tukey, but there was evidently still a lot of skepticism. Anscombe first lists some notions that textbooks were "indoctrinating" people with, like the idea that "numerical calculations are exact, but graphs are rough.

3. What is Pearson's R?

answer-> Correlation coefficients are used to measure how strong a relationship is between two variables.
There are several types of correlation coefficient,but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression.
If you're starting out in statistics, you'll probably learn about Pearson's R first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson's

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

answer-> What?
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.

Why?

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.

*differnce between scaling and standardized->

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

standardized 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).

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

answer-> The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the
fraction (or percent) of points below the given value.

For the reference purpose, a 45% line is also plotted, if the samples are from the same population then the points are along this line.

*Importance of Q-Q Plot in Linear Regression:

1:- Assumption Checking:- Q-Q plots are crucial for assessing the normality assumption of residuals. If the residuals are not normally distributed, it can affect the accuracy of hypothesis tests and the reliability of confidence intervals.

2:- Identifying Outliers:- Outliers in the residuals can also be identified through the Q-Q plot. Outliers may cause deviations from the expected straight line pattern in the plot.

3:- Model Validity:- The validity of the linear regression model relies on the normality assumption for making valid statistical inferences. Q-Q plots provide a visual tool to assess this assumption.

4:- Model Improvements:- If the Q-Q plot indicates non-normality, it may suggest potential improvements to the model, such as transforming the response variable or including additional predictor variables.