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?
   Answer:
   The analysis of categorical columns is done using the boxplot.

   - season: Most of the bike bookings were happening in season 2 and season 3 with a median of over 5000 bookings. This indicates season can be a good predictor for the dependent variable.

   - mnth: Most bike bookings were happening in the months 5,6,7,8,9 & 10 with a median of over 4000 bookings per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

   - weathersit: Most bike bookings happened during ‘weathersit1 with a median of close to 5000 bookings followed by weathersit2 with a median above 4000. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

   - holiday: Most bike bookings were happening when it was not a holiday. This indicates holidays can't be a good predictor for the dependent variable for increasing bookings.

   - weekday: Weekday variable shows a very close trend having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence on the predictor.

   - workingday: Almost 69% of the bike booking were happening on ‘working day’ with a median of close to 5000 bookings (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

   - yr: 2019 is clearly a better year for bike booking. This indicates it could be useful.

2. Why is it important to use drop_first=True during dummy variable creation?
   Answer:
   drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

   Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

   Let’s say we have 3 types of values in the Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.’

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

   Answer: ‘temp’ variable has the highest correlation with the target variable.

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

   Answer:

   The linear regression model is being evaluated with the below assumptions:

   - Error terms should be normally distributed
   - There should be insignificant multicollinearity among variables.
   - Linearity should be visible among variables
   - No auto-correlation
   - There should be no visible pattern in residual values.

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

   Answer:

   Below are the top 3 features contributing significantly towards explaining the demand for shared bikes –

   - Temperature (temp) - A coefficient value of ‘0.0936’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.0936 units.
   - Year (yr) - A coefficient value of ‘0.2266’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2266 units.
   - Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.2678’ indicated that a unit increase in weathersit_3 variable decreases the bike hire numbers by 0.2678 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions.

Mathematically the relationship can be represented with the help of following equation:

y = mx + c

Here, y is the dependent variable we are trying to predict.

x is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect x has on y

c is a constant, known as the Y-intercept. If x= 0, ywould be equal to c.

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between x and the mean of y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of x.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of x, y is normally distributed.

2. Explain the Anscombe’s quartet in detail.

Answer:

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. Each dataset consists of eleven (x,y) points.

The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and must emphasize completely, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics

3. What is Pearson’s R?

Answer:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

Answer:

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalized Scaling	Standardized Scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1]	It is not bounded to a certain range
Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.

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

Answer:

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.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests