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:

Following are some of the analyses made after detailed dependence check on the variables:

a) There has been drastic increase in the number of bookings made in the year 2019 as compared to the year 2018. Especially during Fall of the year, bike bookings have been shown maximum counts
b) The mid of the year stating from month May until October, bookings have shot up compared to the rest of the year. Eventually decreased by the end of the year
c) It became quite evident that 'Clear' weather showed good amount of traction to ride a bike, while 'Misty' weather also had fairly good enough bike bookings.
d) As compared to the start of the week, Thu to Sun have seen more number of bookings
e) People tend to spend more time with bike rides on holidays
f) Be it working or non-working day, bookings seemed to be almost equal

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

Answer:

Using drop_first= True, during dummy variable creation is important to avoid multicollinearity issues, particularly the dummy variable trap, and to simplify the model`s interpretation by establishing a clear reference category. It reduces redundancy in the representation of categorical variables in linear regression models.

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

Answer:

The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

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

Answer:

Following are the assumptions made after building Linear Regression model with the training dataset:

i. Normality of Error Terms - should have normal distribution
ii. Multicollinearity - should have zero insignificant multicollinear variables
iii. Linearity - amongst the chosen variables, linearity should be evident
iv. Homoscedastity - necessary absence of patterns in residual values
v. Independence of residuals - no presence of auto-correlation

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

Answer:

'temp', 'sep' and 'winter' are the top 3 features found contributing significantly towards demand of shared bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning that, changes in the independent variables are associated with a constant change in the dependent variable.

The objective of Linear regression is to find the values of intercepts or the constant terms and coefficients of independent variables that can minimize the sum of squared differences between the predicted and actual values. This involves model training, after selection of a model to serve the requirement. Iterative training of the model has to be performed in order to arrive at convergence. Based, on these iterative modeling, predictions and evaluations are made with test data, while relying on certain assumptions such as, linearity, independence, homoscedasticity, normality.

Linear regression is widely used in applications such as, finances, stock marketing, social sciences etc.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe`s quarter is a set of four datasets that have nearly identical simple descriptive statistics but vary widely when graphed. This dataset was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics. The quartet is deigned to emphasize the need for exploring data graphically before drawing conclusions. Characteristics of Anscombe`s Quartet:

a. Four datasets: Anscombe`s quarter consists of four sets of 11 (x, y) points, labeled I, II, III, and IV.
b. Similar Descriptive Statistics: Despite having different patterns, each dataset shares nearly identical summary statistics, such as means, variances, and correlation coefficients. This highlights the limitation of relying solely on summary statistics for understanding the data.

3. What is Pearson's R?

Answer:

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

Its usecases are as follows - 1. Pearson`s correlation is commonly used in statistics, machine learning and data analysis to assess the strength and direction of relationships between variables. 2. It is often used to measure associations between variables in bivariate data.

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

Answer:

Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:
1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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.