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:

a) cnt vs season : for Spring, 50% of the user count is between 2000 to 4000 which is the lowest and has an outlier near 8000 due to which the overall count has increased. Fall has the highest median.
b) cnt vs mnth : count increases in first few months starting from January and becomes stable then gradually decreases back towards december
c) cnt vs weekday : all days seem to have median values between 4000 to 5000
d) cnt vs weathersit : among the three, high count can be observed for weathersit_1(Clear, Few clouds, Partly cloudy, Partly cloudy) whereas lowest count for weathersit_3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds). Also, weathersit_4(Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog) is not present in the dataset itself

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

Answer:

This means the first column from the dummy variables is dropped. Say, a column contains n = a, b, c different types of values, then the dummy variables created would be n (column ‘a’, column ‘b’, column ‘c’) having values 0 or 1 for each row depending if the row has that value. Eg. if for a particular row, a=1, b=0, c=0 - 100 combination means value is c. Now, if ‘a’ column is dropped, then b=0 and c=0 - this implies that the value is a=1 by default as others are 0. Therefore, the idea of dummy variable creation is to build ‘n-1’ variables for ‘n’ levels.

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

Answer: ‘temp’ and ‘atemp’

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

Answer:

By Residual Analysis. First, predicting the y train output and then plotting a distribution plot for error that is, (y_train - y_predicted values). The plot is close to mean =0 and looks like a normal distribution bell curve.

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

Answer:

a) Year (increased from 2018 to 2019)
b) Temperature (+ve correlation)
c) Spring (-ve correlation)

General Subjective Questions :

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a machine learning algorithm used to determine the best fit line that determines the linear relationship of a dependent variable (predictors) with other independent variables (target) by the process known as supervised learning. Supervised learning regression uses past data and learns, then predicts the possible outcomes from what it has learned previously. Uses of this algorithm - prediction and forecasting. Each model differs based on the relationship between dependent and independent variables and the number of independent variables.

a) Simple Linear Regression: Relationship of a target variable with one independent variable.
   Equation is, ypred = mx+c.
   Where, m is slope and c is constant
b) Multiple Linear Regression: Relationship of a target variable with 'i' independent variable.
   Equation is, ypred = c + m1x1 + m2x2 + ……+ mixi
   Where, m1, m2, …,mi are slopes and c is constant

1. Avoid overfitting - When more and more variables/features are added, say, if we keep increasing the degree of polynomial, the model will end up memorizing the data points in the training dataset, resulting in losing the generalization. Thus, reducing the accuracy of the model.
2. Error - The error terms when plotted against the training set, should be random. Indicating that there is no pattern and the errors are not due to any specific reason.
3. Feature scaling - Scaling is performed on the dataset as it makes the data points closer to each other. Thereby, increasing the performance and faster convergence.
4. Dummy variables - Linear regression modeling works on numerical data but some of the categorical data might also be helpful. So, it is essential to handle these. The idea is that, for a categorical feature with 'n' level (that is 'n' of categories), there would be 'n-1' dummy variables created and each of these levels can be assigned 0 or 1 depending on whether the particular row has that value. For example, a column contains 3 features for a column x1, say, A, B, C. Dummy variables created will be only two, say, A, and B and encoded with 0 or 1. If x1 has A as the value for a particular row, then, A=1, B=0. Here, it is understood that the value of category C is 0.
5. The difference in the actual y and ypred is calculated and square is taken. Then, the sum of all is calculated. The lesser the value, the better. Less value would mean, the predicted values are very close to actual values.
6. R squared: tells about the correlation between the variables. The higher the value, the better the correlation.
7. VIF basically helps explain the relationship of one independent variable with all the other independent variables. VIF = 1/(1-R2). The higher the VIF, the higher is the correlation. The acceptable value of VIF is < 5. During the modeling if it is found that the VIF is high, multicollinearity is detected. To deal with this either drop the variables one at a time or

create new variables. The correlation between independent variables should be close to zero because multicollinearity can result in an inaccurate model.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four datasets (x and y), each with 11 data points. All four sets have nearly identical statistical properties (mean and standard deviation of x and y, and correlation between x and y) but when plotted using scatter plot, were found to appear very different from each other.

First, had a linear relationship between x and y.

Second, had a non-linear relationship between x and y.

Third, had a perfect linear relationship but had an outlier.

Fourth, just one point on the higher side among all other points was good enough to produce a high correlation coefficient.

By this, we understood that to explain a dataset, both statistical properties and graphical analysis is necessary for interpretation.

3. What is Pearson's R?

Answer:

It is also known as the Pearson correlation coefficient (PCC), the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or simply as the correlation coefficient. It is a measure of linear relationship between two sets of variables assuming that the two sets follow normal distribution and homoscedasticity (same variance - error terms are same across all the values of independent variable).

The value of R can lie in the range [-1,1]

a) R = 1, implies strong positive linear correlation
b) R = 0, implies no linear correlation
c) R = -1, implies strong negative linear correlation

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

Answer:

Scaling is a technique used to standardize the features of the dataset points to make them closer to each other. It is performed during pre-processing. Scaling makes sure that the algorithm is not biased towards certain feature values and also helps in faster computation. Hence, giving better performance and efficiency.

Normalized Scaling(Min Max scaling) - It is called Scaling Normalization. It is the transformation of the feature by subtracting from minimum dividing by difference of max and min.

$$X\_new = (x - xmin)/(xmax - xmin)$$

Standardized Scaling - It is called Z score Normalization. It is the transformation of features by subtracting from mean and dividing by standard deviation.

$$X\_new = (x - xmean)/(std\ of\ x)$$

Normalized Scaling (Min Max)	*Standardized Scaling*
Minimum and maximum values are used	Mean and standard deviation are used
Scale values range : [0,1]	No specific range
It is affected by outliers	It is not affected by outliers

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

Answer:

VIF indicates the strength of correlation between the variables. The higher the VIF, the higher is the correlation.

$$VIF = 1/(1 - R^2)$$

VIF will be infinity when, R^2 = 1 and this is possible in case of perfect correlation.

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

Answer:

Q–Q plot is a probability plot which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. To plot, consider a point (x,y) one of the quantiles of the second distribution (y coordinate) is plotted against the same quantile of the first distribution (x coordinate). It helps in determining if the two datasets come from the population with the same distribution or not. In linear regression, there could be scenarios when the train and test data are provided separately, then by Q-Q plot we can confirm if they come from the population with the same distribution.