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?**

   The categorical variables are : 'mnth', 'weekday', 'season' , 'weathersit' ,'yr' and 'holiday'. Based on the boxplot analysis we can infer following:
   • Yr: Bike bookings are higher in 2019 as compared to 2018, it might be due to the fact bike rentals are getting popular and people are becoming more aware about environment.
   • season: Highest booking happening in season3(fall) with a median of over 5000 booking. This was followed by season2(summer) & season4(winter) of total booking.
   • mnth: Bike booking is quite high in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
   • weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking. This was followed by weathersit2. Clear weather is most optimal for bike renting.
   • holiday: The bike booking were happening mostly when it is not a holiday.
   • weekday: weekday variable shows very close trend. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
   • workingday: Median is quite close, does not have much impact

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

   Reasons for dropping first dummy variable :
   1. To reduce multicollinearity : If we do not drop the first column during dummy variable creation , it will lead to high multicollinearity between the dummy variables and will adversely affect the model.
   2. To reduce redundancy. For eg for a variable gender , both male and female dummy are not required. Male=0 will be a female. However , sometimes it depends on the number of values in a categorical variables. For a categorical variable with large number of values, the drop first could be avoided to see the effect of all the values of the variable.

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

Answer: Temp[temperature] 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 set?

Error terms are normally distributed with mean zero(not X,Y)

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

Solution:

i. Temperature[0.6076] Increase in temp implies increase in cnt

ii. light snow weather condition (-0.2173) has Negative Impact

iii. Year[0.2324] 2019 has more sales, increase in year, increase in cnt

General Subjective Questions:

1) Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

The best fit regression line $Y = \beta_o + \beta_1 X$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

β_o = intercept of the line (Gives an additional degree of freedom)

β_1 = Linear regression coefficient (scale factor to each input value).

Finding the best fit line:

Answer: When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

Assumptions of Linear Regression

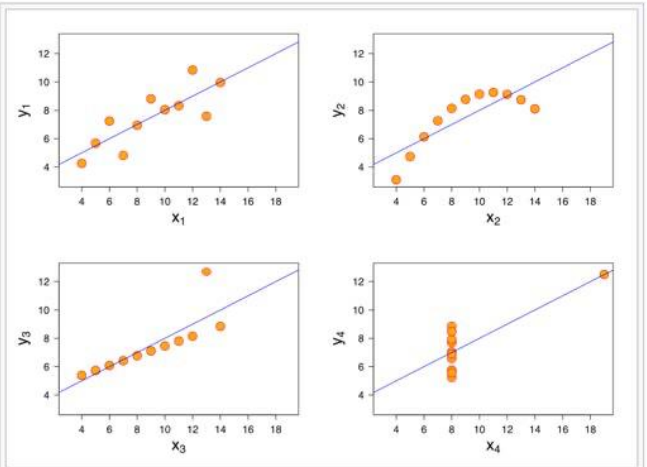
1. Linear relationship between the features and target
2. Small or no multicollinearity between the features:
3. Homoscedasticity Assumption: there should be no clear pattern distribution of data in the scatter plot.
4. Normal distribution of error terms: It can be checked using the q-q plot.
5. No autocorrelations: no dependency between residual errors.

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.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it, and the effect of outliers and other influential observations on statistical properties.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

• The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
• The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
• In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
• Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Importance

The quartet is used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3) **What is Pearson's R?**

Answer: It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

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

Answer:

Scaling :

It is a data pre-processing technique 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 to perform Scaling

The variables in the data set under analysis may have highly varying magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence results in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. So we need to scale features because of two reasons:

• Ease of interpretation

• Faster convergence for gradient descent methods

Normalization/Min-Max Scaling:

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

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization 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).

$$x = \frac{x - mean(x)}{sd(x)}$$

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

Answer:

Using this formula, VIF=infinity implies R2=1, which implies a perfect correlation between two Independent variables. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

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

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution. It helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

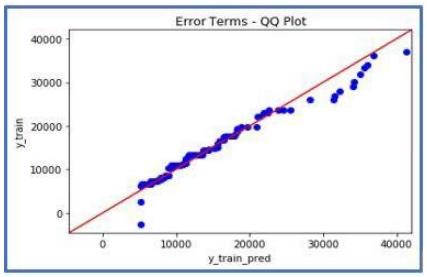
It compare the shapes of distributions, providing a graphical view of how properties such as location,scale, and skewness are similar or different in the two distributions.

Interpretation:

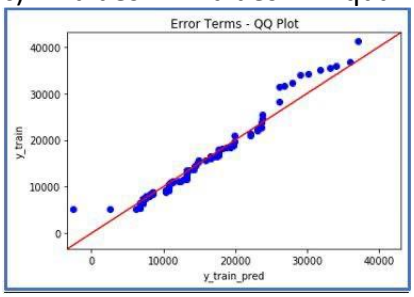
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree
from x - axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45
degree from x -axis