

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?

Clearly the categorical variable are influencing the dependent variable i.e cnt

1. **For Season variable:**

During fall i.e 3 count of rental bikes is more

2. **For yr variable:**

From this variable we can see a significant progress as count of rental bike is increased in the year 2019

3. **For mnth variable:**

From this variable we can understand that the peak season is from 5th to 9th month

4. **For holiday variable**

Holiday slightly effect the rental as comparatively during the rentals are less

5. **For variable weekday, workingday there is no effect in rental of bikes**

6. **For variable weathersit:**

During the rainy season the rentals are less

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

By using the drop_first=True eliminate the creation of the extra variable while creation of the dummy variable

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

From the pair plot temp and atemp has the highest correlation

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

The residual i.e $y_{\text{actual}} - y_{\text{pred}}$ was normal distributed and there was no relation between the error term

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

Temp, Season, mnth are top 3 feature

General Subjective Questions

1. Explain the linear regression algorithm in detail

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 getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

2. Explain the Anscombe's quartet in detail

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 analyzing it and the effect of outliers on statistical properties.

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r , the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by 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 .

The Pearson's correlation coefficient varies between -1 and $+1$ where:

- $r = 1$ means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- $r = -1$ means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- $r = 0$ means there is no linear association
- $r > 0 < 0.5$ means there is a weak association
- $r > 0.5 < 0.8$ means there is a moderate association
- $r > 0.8$ means there is a strong association

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

Scaling 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.

Scaling is performed because *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.*

Normalized Scaling brings all of the data in the range of 0 and 1.

Standard Scaling 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).

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

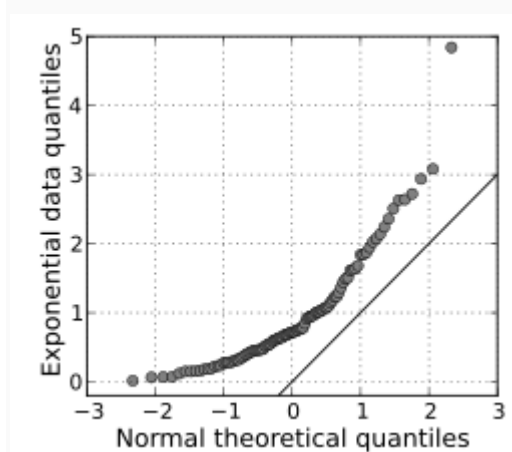
If there is perfect correlation, then $VIF = \infty$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which lead to $1/(1-R^2)$ infinity. To solve this problem we need to drop one of the variables from the dataset which is causing 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).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to 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.