

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? (3 marks)**

Answer:

yr: In 2019, the cnt is increased more than 2018.

holiday: In holiday, the median count is less.

season: In fall season, the cnt is high. And in the spring season it's very less.

weathersit: In clear weather, it's more affecting the cnt var. The registration is very high in the clear weather.

mnth: June-September, the cnt affecting more.

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

Answer: It helps in reducing the extra column created during dummy variable creation.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer: The temp and atemp variables have the highest correlation with cnt.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Answer: By finding the r-squared for the test set and comparing it with the r-squared of the train set. It should be very similar.

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

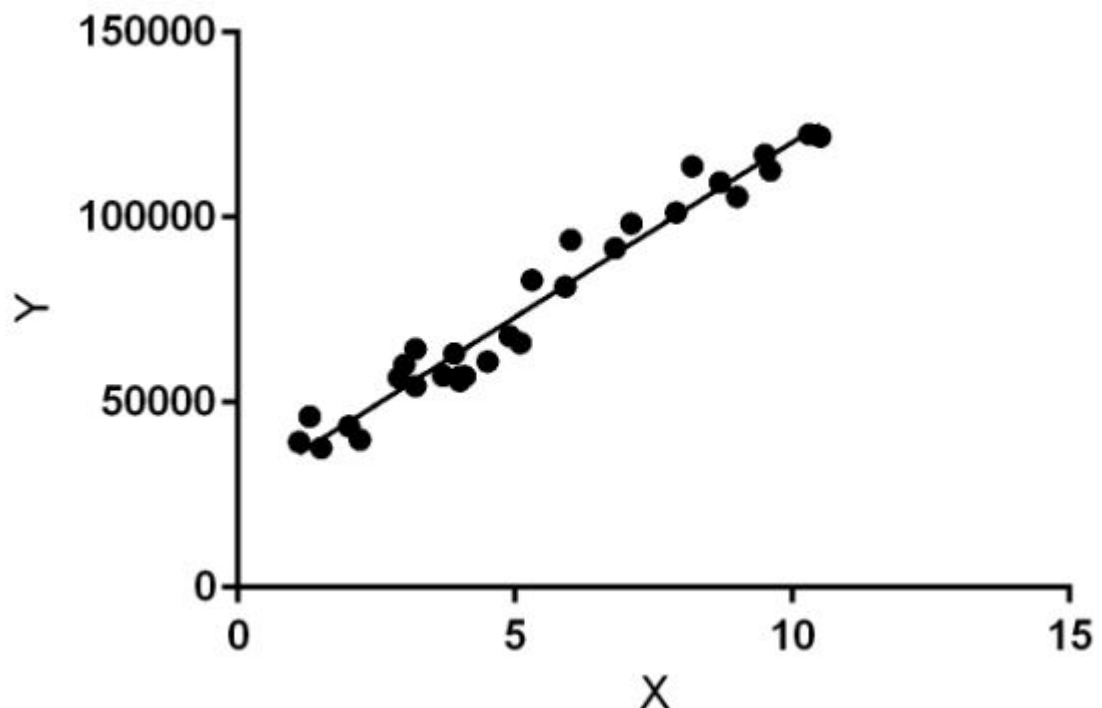
Answer: yr, temp, weathersit

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression Algorithm: 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.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Steps to create the linear regression model:

1. Reading understanding and visualising the data
2. Preparing the data for modelling (split training and testing data, rescaling etc.)
3. Train the model

4. Residual Analysis
5. Predictions and evaluation on test set

2. Explain the Anscombe's quartet in detail.

Answer:

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?

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. When one variable changes, the other variable changes in the same direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

What?

It 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.

Why?

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.

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.

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).

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

Answer:

If there is perfect correlation, then $VIF = \text{infinity}$. 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.

Answer:

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:

