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:

   From the analysis of season, yr, mnth, holiday, weekday, weathersit are categorical variables in the dataset. We can say that as below:

   - In the Fall season, there will maximum booking received and a result maximum active customers
   - In the month of September, maximum booking received
   - The year 2019 have registered more bookings in comparison of year 2018
   - Holidays affect the active count which drops
   - No bookings when bad weather like: heavy rain
   - More bookings when partly cloudy/clear sky

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

Answer:

it important to use **drop_first=True.** It helps us make the dummy variables correlated to each other. So, we can remove redundant, which helps to reduce the extra column creation while doing 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:

atemp and **temp** 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? (3 marks)

Answer:

One of the fundamental assumptions of a Linear **Regression Model** is that the **error terms** should correspond to a **normal curve, when plotted on histogram**.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

The **top 3 features directly influencing the count** are the features with **highest coefficients**. These are given as below:

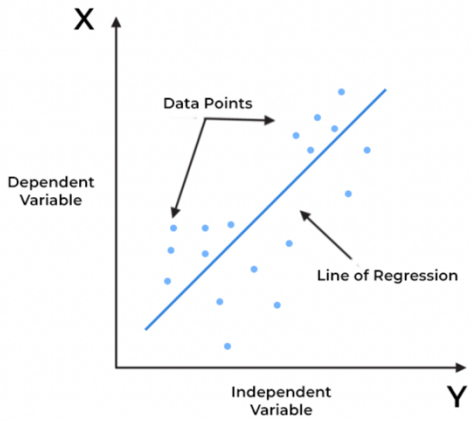
- Temp,
- Year (positively influencing)
- snowy and rainy weather (negatively influencing)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear regression is an algorithm that **provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events**. It is a statistical method used in data science and machine learning for predictive analysis



An interpolation technique used to predict correlation between variables and how an independent variable is influenced by the dependent variable(s).

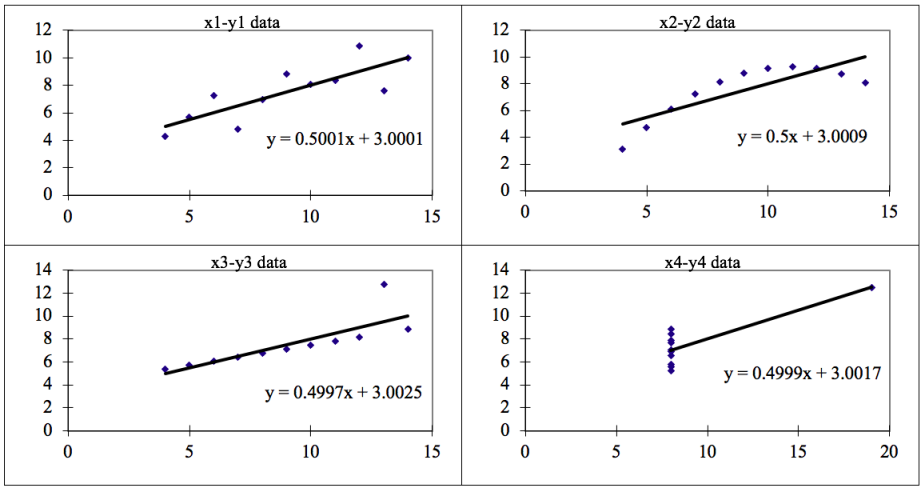
As we have looked into the bike sharing data set and cleaned it with exploratory data analysis, we split the dataset into training set (which would be used to train a model) and the testing set (which would be used to check how close is our model to the actual output). After checking the collinearity of variables and using the requisite variables to train the model and checking the R-value of the model and the p-values of dependent variables, after dealing/dropping the necessary columns and reiterating the steps (feature elimination), we come to a final model. According to the conditions of linear regression which states that the error curve must be a normal one, we proceed to testing the model with the test dataset. The conclusion, we have drawn on the model would be used to provide valuable insights/predictions on datapoints in the range of the model.

Note: Please see next page

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots



The four datasets can be described as:

- **Dataset 1:** this **fits** the linear regression model pretty well.
- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
- **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

Note: Please see next page

3. What is Pearson's R? (3 marks)

Answer:

Pearson's R is a numerical summary of the strength of the **linear association between the variables**. If the variables tend to go up and **down together, the correlation coefficient** will be positive.

The value of Pearson's R always lie between -1 and +1, the latter indicating a perfectly positive and linear correlation and the former indicating a perfectly linear negative regression. The values in between denotes the relative collinearity of two variables

Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

Answer:

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.

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 typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Note: Please check next page

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. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

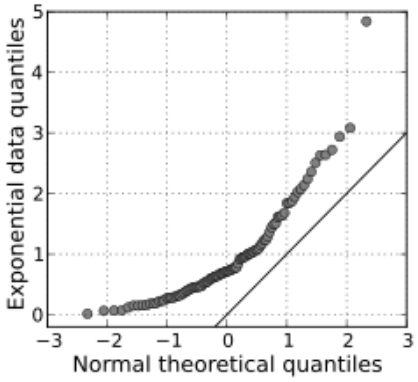
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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:



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.