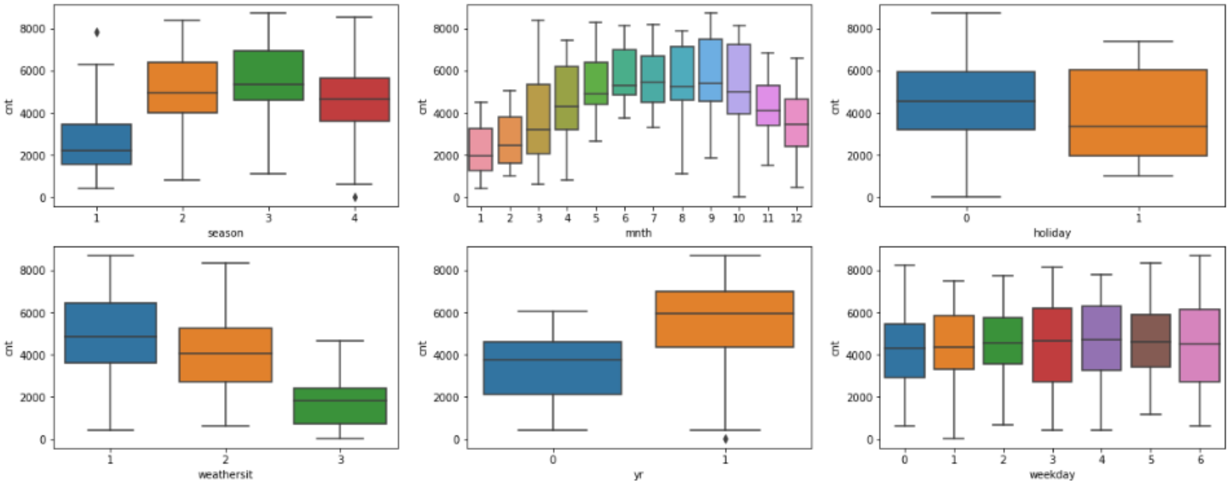
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer:



Categorical variables mnth, yr, weekday, holiday, season and weathersit were observed. From the boxplots plotted, we can conclude that;

- **mnth**: demand increases from January to September and then gradually decreases towards end of the year (winter months)
- **yr**: demand is increasing year-on-year
- **season**: demand is highest in fall and lowest in spring
- **holiday**: demand is more on non-holidays as compared to holidays
- **weathersit**: demand is highest when weather is clear and lowest during light snow.
- **weekday**: demand is higher in non-working day

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

Answer:

drop_first=True is important to use, as **it helps in reducing the extra column created during dummy variable creation**. Hence it reduces the correlation created among dummy variables.

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

Answer:

As per the pair plot, '**temp**' and '**atemp**' are the two numerical variables which are highly correlated with the target variable '**cnt**'.

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

Answer:

According to this assumption, there is a linear relationship between features and target. Linear regression captures only linear relationship:

- This was validated by plotting a scatter plot between the features and target
- Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent and feature variables.
- The error distribution was observed and it was a normal distribution.
- **Linear relationship:** Strong linear relationship observed in the pair plots between 'temp' and 'cnt'. The same was confirmed in the final model summary. The coefficient of 'temp' was **>0.5**
- **No or little multicollinearity:** Variables with **p-values>0 and VIF>0.05** were dropped from the final model
- **No auto-correlation:** Variables 'temp' and 'atemp' had strong correlation. The later was dropped from the final model.

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

Answer:

As per the final model, the top 3 features are:

temp – coefficient: 0.491508

yr – coefficient: 0.233482

weathersit_Light Snow & Rain – coefficient: -0.285155

General Subjective Questions

1. **Explain the Linear Regression algorithm in detail.**

Answer:

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values.Linear Regression is the most basic form of regression analysis.Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation “y = mx + c”. It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

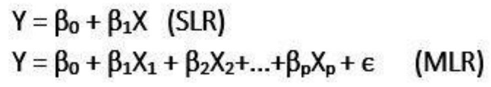
In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.

Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

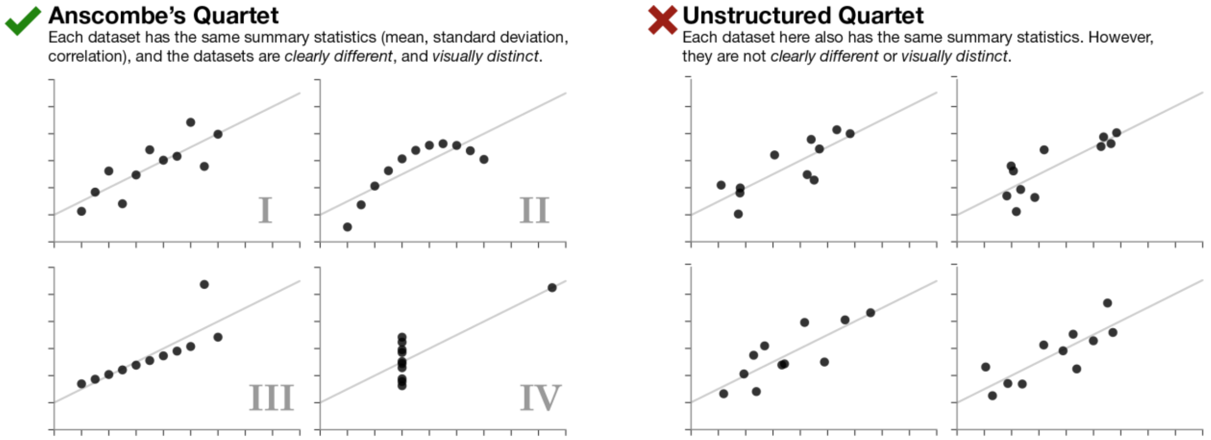


$$Y = \beta_0 + \beta_1 X \quad \text{(SLR)}$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \quad \text{(MLR)}$$

β1 = coefficient for X1 variable
β2 = coefficient for X2 variable
β3 = coefficient for X3 variable and so on…
β0 is the intercept (constant term).

2. Explain the Anscombe’s quartet in detail.

Answer:
Anscombe’s Quartet demonstrates the importance of data visualization. Some people are of the impression that charts are simply "pretty pictures", while all of the important information can be divined through statistical analysis. An effective (and often used) tool used to demonstrate that visualizing your data is in fact important is Anscombe's Quartet. Developed by F.J. Anscombe in 1973, Anscombe's Quartet is a set of four datasets, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quite similar. However, after visualizing (plotting) the data, it becomes clear that the datasets are markedly different. The effectiveness of Anscombe's Quartet is not due to simply having four different datasets which generate the same statistical properties, it is that four ***clearly different*** and ***visually distinct*** datasets are producing the same statistical properties. In contrast the "Unstructured Quartet" on the right in below Figure also shares the same statistical properties as Anscombe's Quartet, however without any obvious underlying structure to the individual datasets, this quartet is not nearly as effective at demonstrating the importance of visualizing your data.



✓ Anscombe’s Quartet
Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.

✗ Unstructured Quartet
Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.

- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.

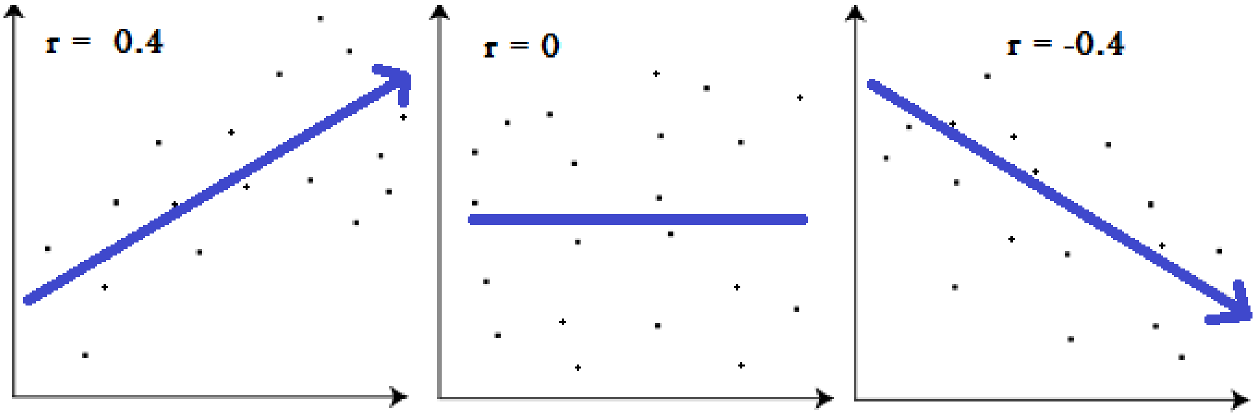
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example 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.

3. What is Pearson’s R?

Answer:
Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson’s. Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression. The formula below returns a value of ‘r’ between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$



Positive Correlation No correlation Negative

As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation.)

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

Answer:

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

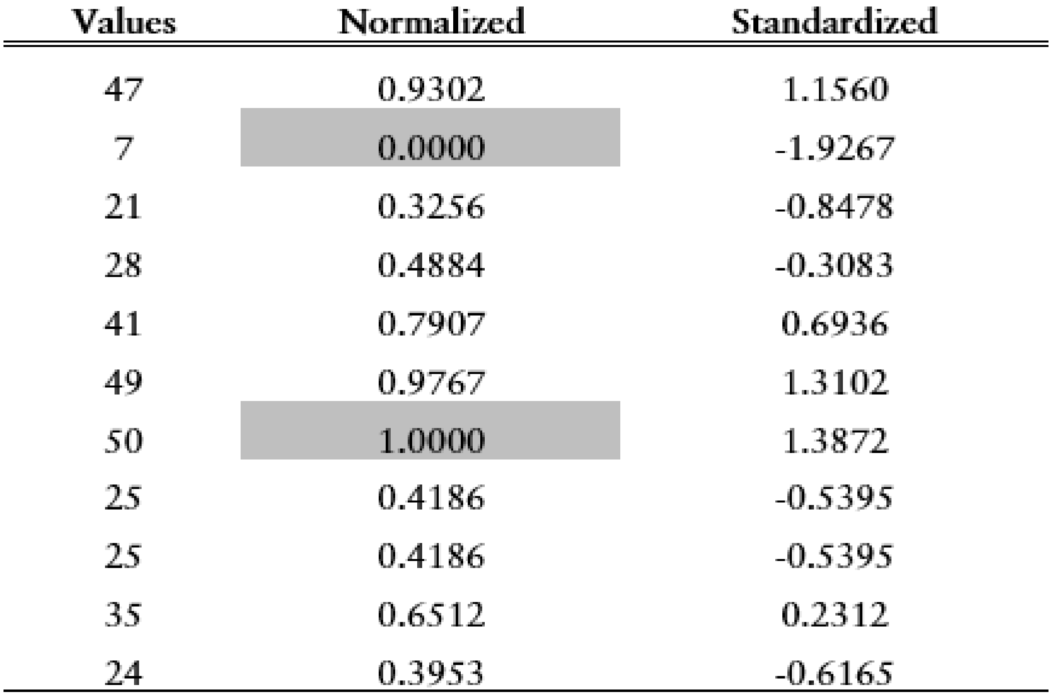
Feature scaling or simply scaling means adjusting data that has different scales so as to avoid biases from big outliers. The most common techniques of feature scaling are Normalization and Standardization.

Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
It is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.
While Standardization transforms the data so as to have zero mean and a variance of 1. The table below compares raw data with its two transformations, the second column is processed through normalization and the third column is calculated using the standardization function:

$$x_{new} = \frac{x - \mu}{\sigma}$$



Values	Normalized	Standardized
47	0.9302	1.1560
7	0.0000	-1.9267
21	0.3256	-0.8478
28	0.4884	-0.3083
41	0.7907	0.6936
49	0.9767	1.3102
50	1.0000	1.3872
25	0.4186	-0.5395
25	0.4186	-0.5395
35	0.6512	0.2312
24	0.3953	-0.6165

Feature scaling does not alter how data is shown. If you plot the three columns above, you will get exactly the same figure. This step is vital for the success of any machine learning model with the exception of the Random Forest algorithm who can be run without the need to scale data although it’s always best practice to do so. Scaling (precisely Normalization) is also helpful to detect historical highs and lows.

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

Answer:

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient
estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity.Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it’s R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in “infinity”

The common heuristic we follow for the VIF values is:

- **10**: Definitely high VIF value and the variable should be eliminated.
- **5**: Can be okay, but it is worth inspecting.
- **<5**: Good VIF value. No need to eliminate this variable.

$$VIF_1 = \frac{1}{1 - R^2}$$

R-square in this formula is the coefficient of determination from the linear regression model which has:

- X1 as dependent variable
- X2 and X3 as independent variables

R-square =1 means perfect correlation. If this is the case, then VIF will be **infinity**.

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

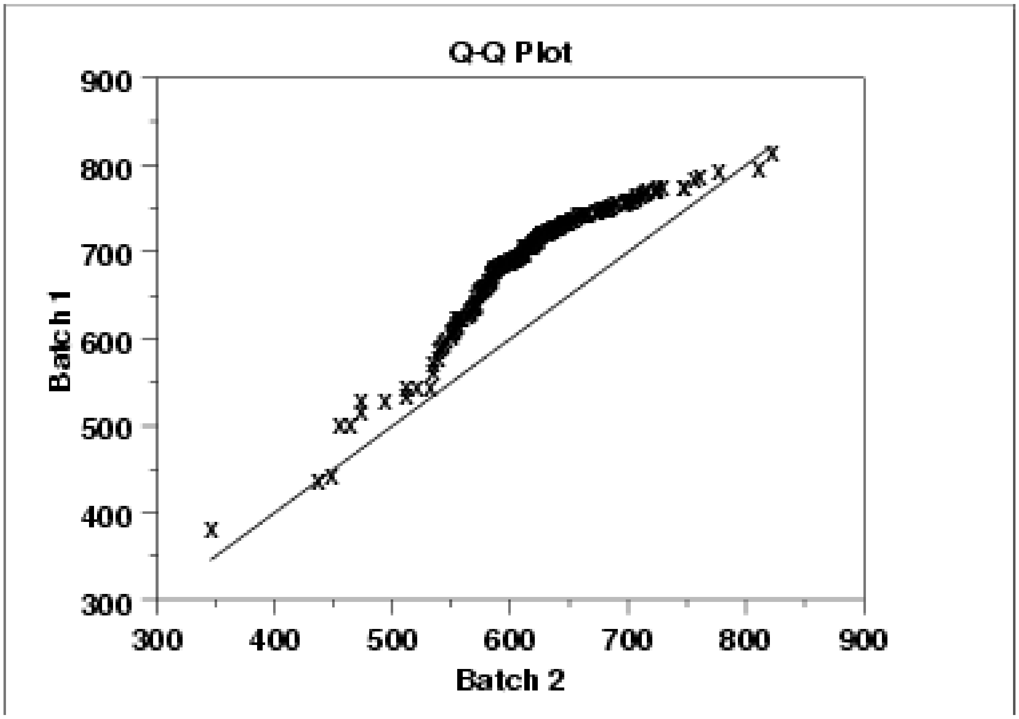
Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight.

By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.



Q-Q plot 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.

The Q-Q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?