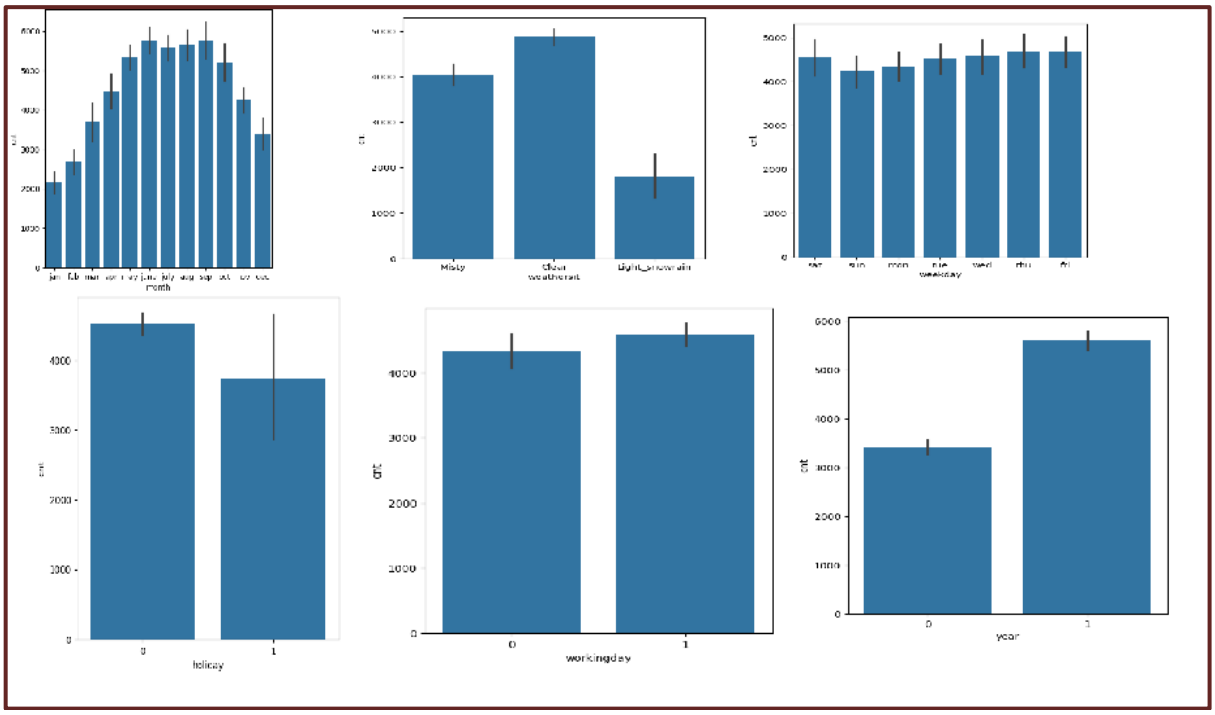
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

- The fall season has seen an increase in bookings. Additionally, the booking count for each season has significantly risen from 2018 to 2019.
- Most of the bookings were made during May, June, July, August, September, and October. The trend increased from mid-year and then decreased as the year approached its end.
- The clear weather resulted in an increase in bookings, which is to be expected.
- There are more bookings on Thursdays, Fridays, and Saturdays compared to the beginning of the week.
- When it is not a holiday, the number of bookings tends to be lower, which is understandable considering the increased demand during holiday periods.
- 2019 saw an increase in bookings compared to the previous year, indicating positive business progress.



Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)

Answer:

If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

The purpose of drop_first=True is usually to avoid multicollinearity.
In one-hot encoding, you set 1 to the position associated to a discrete value among n possible options. When you one-hot encode a value, there is redundant information because you can figure out the value of any of the positions by computing 1 minus the sum of all other values. This means that any position of the one-hot encoded variable is a linear combination of the other positions.
This linear correlation, however, can be a problem in some cases. One example is when you want to know the effect the input features have on the prediction of a logistic or linear regression model

One solution to the multicollinearity of one-hot encoding is simply to, remove one of the values. With that, you don't loose information and, at the same time, you remove the multicollinearity. drop_first=True is precisely for that.
In cases where multicollinearity is not a problem, however, it is desirable to have all the dummies and therefore use drop_first=False. Nevertheless, it should be decided case by case. For instance, in decision trees, it's easier to learn the association between the dummy having value 1 than to learn that all the others have value 0. In the case of k nearest neighbours, having n - 1 dummies make the distances between samples with the missing value different than distances between samples with the other values, so it would be best to have all the dummies (at least with the Euclidean distance).
Other cases where it's specifically better to have all dummies is the case where there are missing values, and you are encoding them as all zeroes.

Finally, in models with regularization, it is also needed to keep all dummies because otherwise, the predictions may depend on which column you leave out

---

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer:

'temp' variable has the highest correlation with the target variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

- Scatter Plot
  - Independent variable vs dependent variable to check for linear relationship
  - Residuals vs fitted value – to check if any assumptions violated.
- Outliers
  - Check for outlier as linear regression sensitive to outlier
- Normality of error terms
  - Error terms should be normally distributed
- Multicollinearity check
  - There should be insignificant multicollinearity among variables.
- Linear relationship validation
  - Linearity should be visible among variables
- Homoscedasticity
  - There should be no visible pattern in residual values.
- Independence of residuals
  - No autocorrelation

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes
Temp, winter, year

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

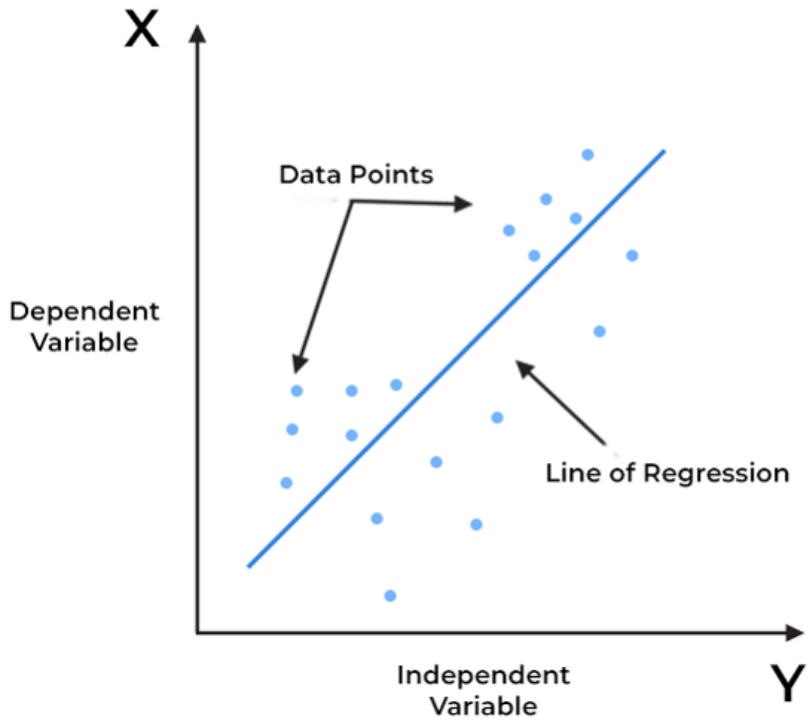
Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.

A sloped straight line represents the linear regression model.



Best Fit Line for a Linear Regression Model
In the above figure,
X-axis = Independent variable
Y-axis = Output / dependent variable
Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the 'best fit line.' The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

Key benefits of linear regression

Linear regression is a popular statistical tool used in data science, thanks to the several benefits it offers, such as:

1. Easy implementation

The linear regression model is computationally simple to implement as it does not demand a lot of engineering overheads, neither before the model launch nor during its maintenance.

2. Interpretability

Unlike other deep learning models (neural networks), linear regression is relatively straightforward. As a result, this algorithm stands ahead of black-box models that fall short in justifying which input variable causes the output variable to change.

3. Scalability

Linear regression is not computationally heavy and, therefore, fits well in cases where scaling is essential. For example, the model can scale well regarding increased data volume (big data).

4. Optimal for online settings

The ease of computation of these algorithms allows them to be used in online settings. The model can be trained and retrained with each new example to generate predictions in real-time, unlike the neural networks or support vector machines that are computationally heavy and require plenty of computing resources and substantial waiting time to retrain on a new dataset. All these factors make such compute-intensive models expensive and unsuitable for real-time applications.

The above features highlight why linear regression is a popular model to solve real-life machine learning problems.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

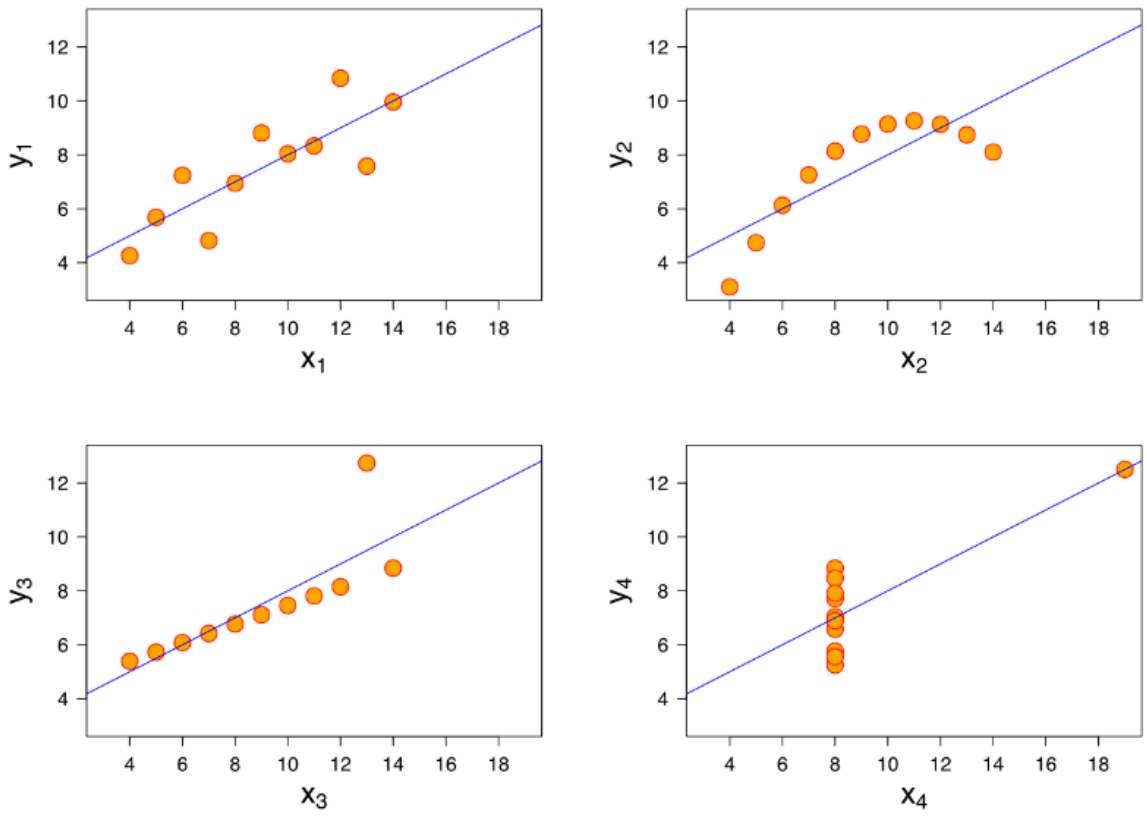
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's correlation coefficient is a statistical measure that evaluates the strength and direction of the relationship between two continuous variables. It is considered the most effective method for assessing associations due to its reliance on covariance. This coefficient not only reveals the magnitude of the correlation but also its direction.

Key Assumptions:
Independence: Each case should be independent of others.
Linearity: There must be a linear relationship between the variables, which can be verified through a scatterplot. If the plot forms a straight line, the criterion is met.

Homoscedasticity: The scatterplot of residuals should approximate a rectangular shape.

Characteristics:

Range: The coefficient's value ranges from +1 (perfect positive correlation) to -1 (perfect negative correlation), with 0 indicating no correlation.

Unit Independence: The coefficient is unaffected by the units of measurement, ensuring comparability across different scales.

Symmetry: The correlation between two variables remains consistent, regardless of the variable order (X with Y or Y with X).

Degrees of Correlation:
Perfect: Values near ±1 indicate a perfect correlation, where one variable's increase (or decrease) is mirrored by the other.

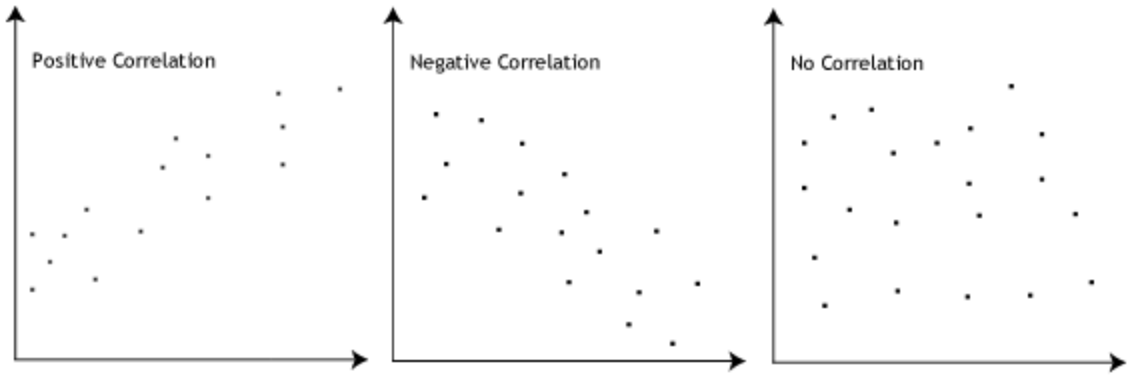
High Degree: Values between ±0.50 and ±1 suggest a strong correlation.

Moderate Degree: Values between ±0.30 and ±0.49 indicate a moderate correlation.

Low Degree: Values below +0.29 are considered a weak correlation.

No Correlation: A value of zero implies no relationship.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

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.

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.

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.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

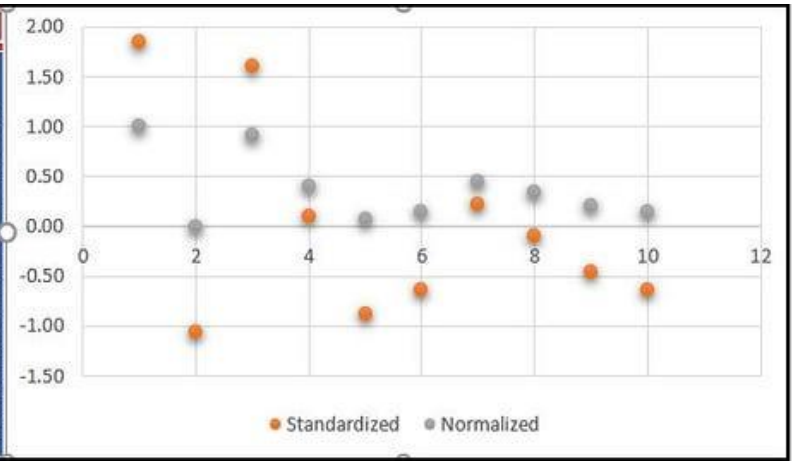
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Example:
Below shows example of Standardized and Normalized scaling on original values.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

The Variance Inflation Factor (VIF) can be infinite when there is perfect correlation between independent variables in a multiple regression equation. This happens when one independent variable is equal to a linear combination of the other independent variables

The value of VIF is calculated by the below formula:

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

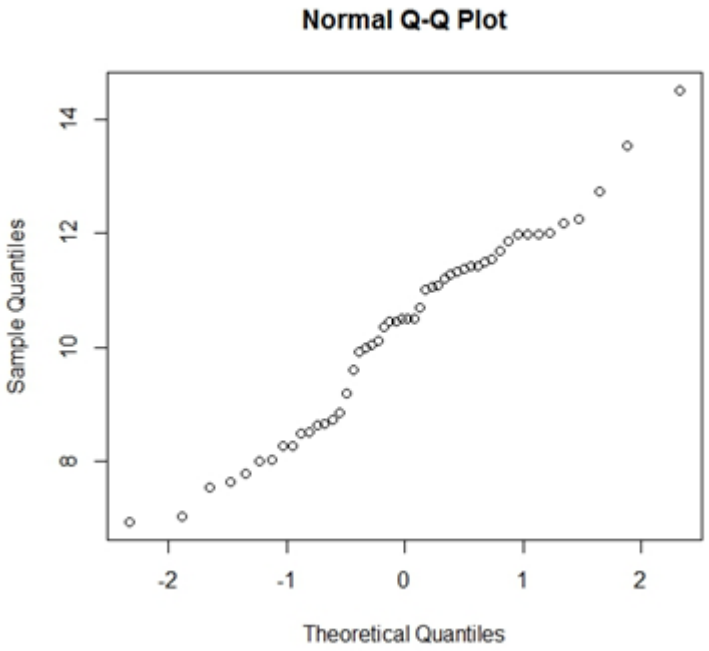
Where, 'i' refers to the ith variable.
If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables

---

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

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

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. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.
