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 categorical data visualization
   - During the fall season the bookings are increased
   - Bookings are good between the months of May and October
   - Bookings showed increasing trend until 3rd quarter of the year and dropped in the last quarter
   - During clear weather we see more bookings
   - The weekday did not impact the average bookings
   - Bookings increased drastically in 2019 compared to 2018
   - Bookings are low during holidays
   - Working day and non-working day has no impact on the bookings

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

   Answer:
   drop_first=True is used when creating dummy variables, and it removes the extra column that is created.
   We use a technique which converts categorical data into a ml understandable mode.
   Assuming we have a categorical column with **K** categories/values, using pd.getdummies() we convert them into a binary, that is one column for each unique value of the original column and wherever this value is true for a row it is indicated as 1 else 0.
   In this case we end up with **K** columns, with **drop_first=True** we are left **K+1** columns.
   Here if K+1 columns are 0 which indicates the first column or dropped column is True/1.

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

   Answer:
   Temperature (temp) variable 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:
   I have validated the assumption of Linear Regression Model based on assumptions
   - Error terms follow normal distribution
   - Multicorrelation
   - Homoscedasticity
   - Mean of residuals
   - Independence of residuals

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:

Top 3 features which influence the bike bookings are

1. Temperature (temp)
2. Weather (Weathersit) and
3. Year

General Subjective Questions

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

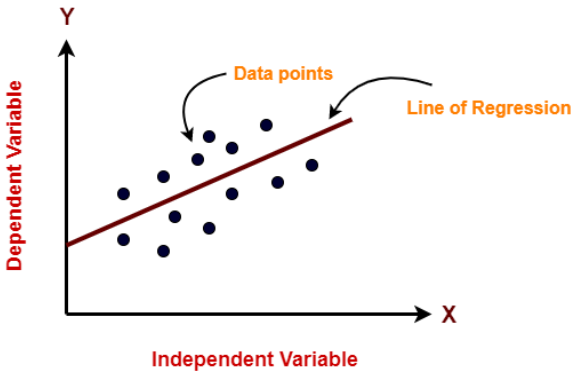
Answer:

Linear regression is a modelling technique or a way of calculating the relationship between dependent (target) variables and independent (predictor) variables.

The standard equation for linear regression is $\mathbf{y = \beta_o + \beta_1.X}$ or $\mathbf{y = mX + c}$

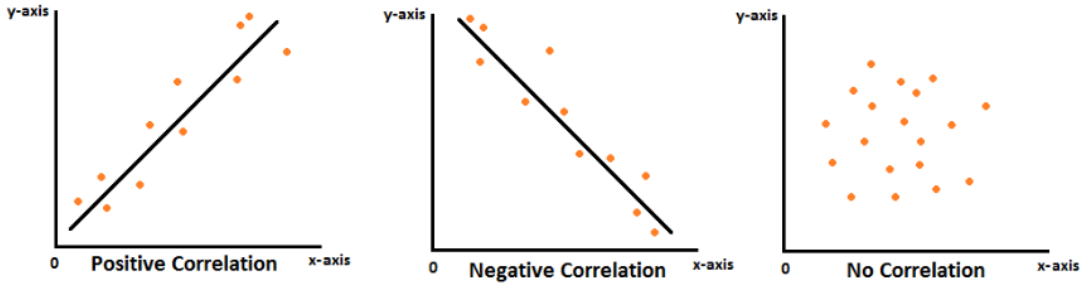
where **y** is the dependent variable, **X** is the independent variable. $\mathbf{\beta_o}$, $\mathbf{\beta 1}$ are coefficients determining intercept and slope

Linear regression algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s).



In the figure above, on X-axis is the independent variable and on Y-axis is the output. The regression line is the best fit line for a model. And our main objective in this algorithm is to find this best fit line.

Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y. A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables x and y.

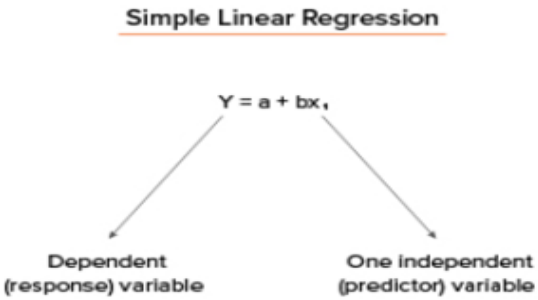


Positive linear relationship: dependent variable increases as dependent variable increases

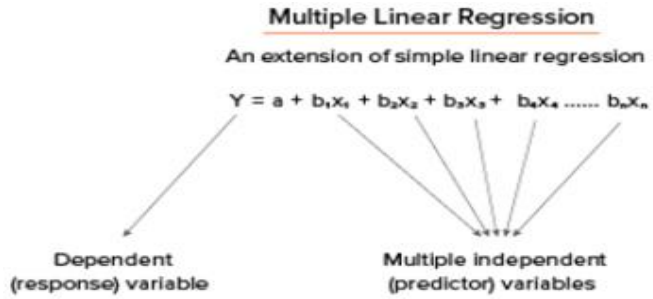
Negative linear relationship: dependent variable increases as dependent variable decreases

Linear regression models can be classified into two types depending upon the number of independent variables.

Simple linear regression explains the relationship between a dependent variable and one independent variable using a straight line.



Multiple linear regression explains the relationship between one dependent variable and several independent variables (explanatory variables).



Assumptions:

1. **Linearity:** The relationship between X and the mean of Y is linear
2. **Homoscedasticity:** The variance of residual is the same for any value of X
3. **Independence:** Observations are independent of each other
4. **Normality of residuals:** For any fixed value of X, Y is normally distributed

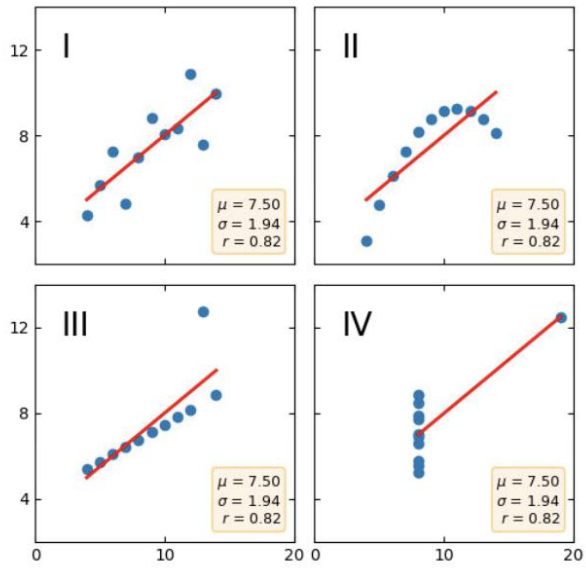
The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

Where RSS= Residual Sum of Squares and TSS= Total Sum of Squares

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

Answer:

Anscombe's quartet highlights the importance of plotting data and not just rely on statistical measures when analyzing data.



To illustrate this statistician Francis Anscombe created 4 data sets which would produce nearly identical statistical measures.

Statistical measures

1. Mean of x values in each data set = 9.00
2. Standard deviation of x values in each data set = 3.32
3. Mean of y values in each data set = 7.50
4. Standard deviation of x values in each data set = 2.03
5. Pearson's Correlation coefficient for each paired data set = 0.82
6. Linear regression line for each paired data set: y = 0.500x + 3.00

When looking at this data we would be forgiven for concluding that these data sets must be very similar – but really, they are quite different.

When datasets are scatter plotted as shown in the above figure reveals 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.

So Anscombe's quartet emphasizes the importance of visualization in Data Analysis.

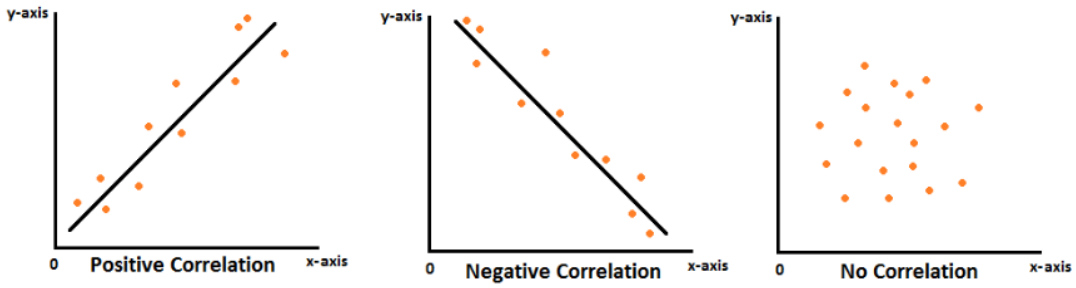
3. What is Pearson's R? (3 marks)

Answer:

Correlation coefficients are used 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 Pearson correlation coefficient, r, can take a range of values from +1 to -1.

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



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
- Zero means that for every increase, there is not a positive or negative increase.

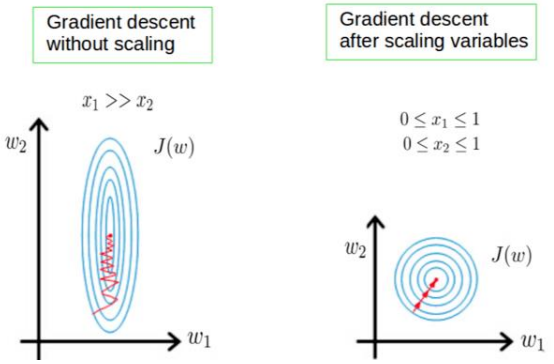
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is a technique to normalize or standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.

Machine learning algorithm just sees number, if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So, these more significant number starts playing a more decisive role while training the model.

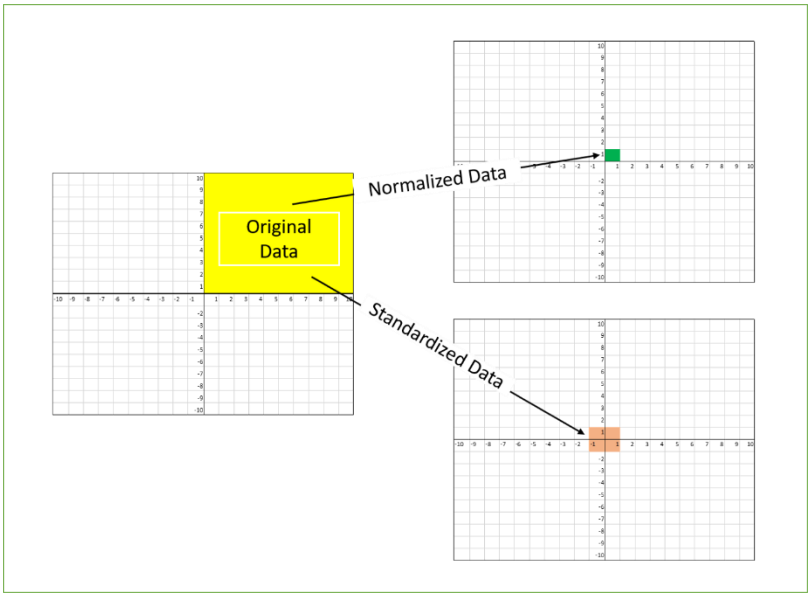
Another reason why feature scaling is applied is that few algorithms like Neural network gradient descent converge much faster with feature scaling than without it.



The most common techniques of feature scaling are Normalization and Standardization both make the data unitless

Normalization is used when we want to bound our values between two numbers, typically, between [0, 1] or [-1, 1]

Standardization transforms the data to have zero mean and a variance of 1



#	Normalization	Standardization
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 an often called as Scaling Normalization	It is an often called as Z-Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

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).

VIF	Conclusion
1	No multicollinearity
4 – 5	Moderate
10 or greater	Severe

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

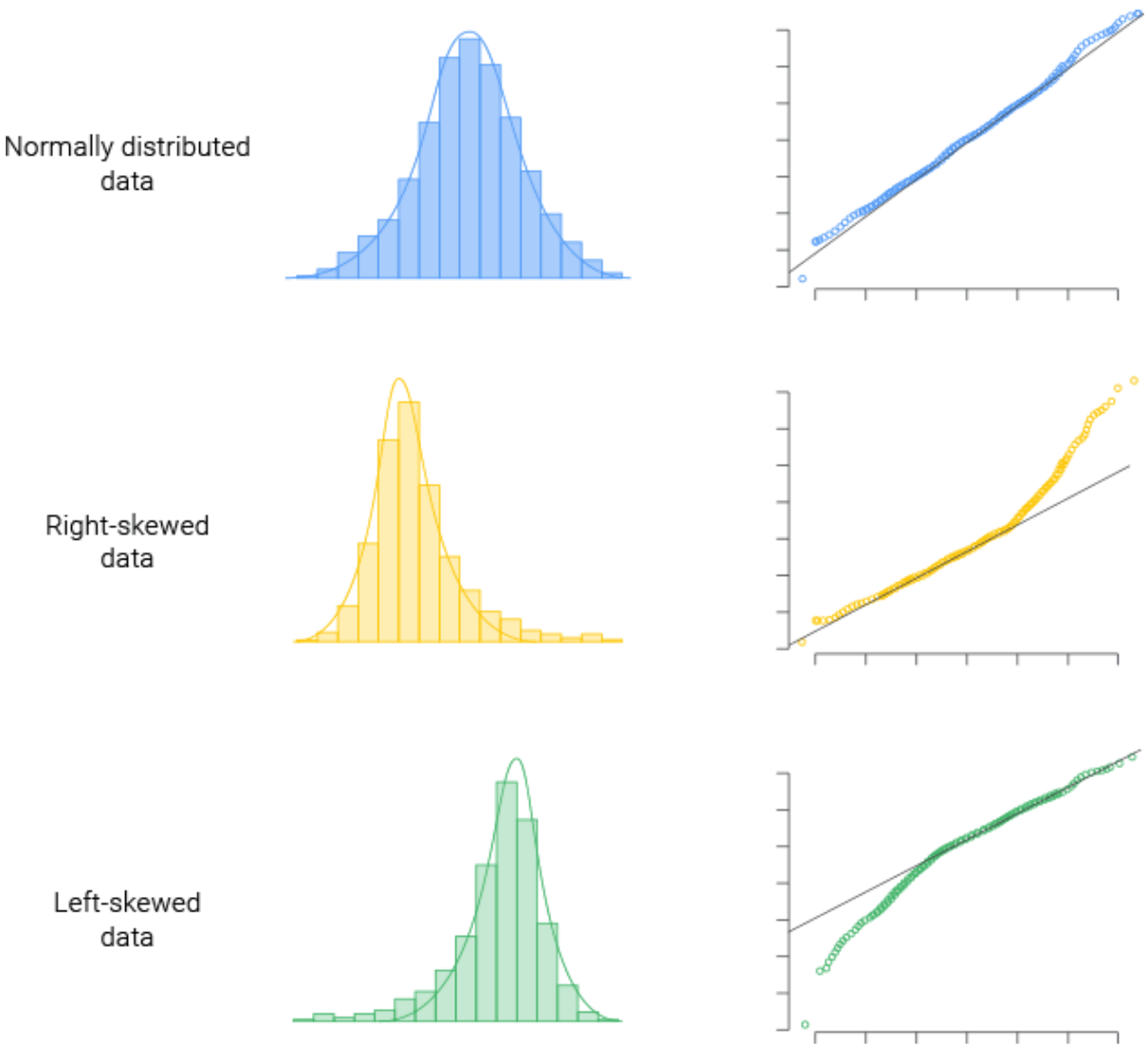
The quantile-quantile or q-q plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set.

For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption.

It 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 is roughly straight.

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.

If the data points were normally distributed, most of the points would be on the 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.