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:

Based on the exploratory data analysis and data visualization done on categorical columns following points can be inferred:

- Fall season has more bookings than any other season and has increased in the subsequent year.
- Clear weather attracts more booking than any other weather.
- The bookings gradually starts rising from the beginning of the year, highest bookings are observed in the months of May, June, July, August, and September and then gradually starts to decrease.
- On holidays demand seems to drop, which might be possible because, people might want to spend some time with family.
- Year 2019 has more numbers of booking than 2018, which indicates there is progress in terms of business with increase in year.

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

Answer :

- Drop_first = True is important to use as it helps in reducing the extra column created during dummy variable creation. It is important to avoid multicollinearity issues and improve the interpretability of the model.

- Syntax -
  drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.
  Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.
- Here is an example of its usage in pandas
  ```
  import pandas as pd

  # Assuming df is your DataFrame with a categorical column 'Color'
  df = pd.get_dummies(df, columns=['Color'], drop_first=True)
  ```

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

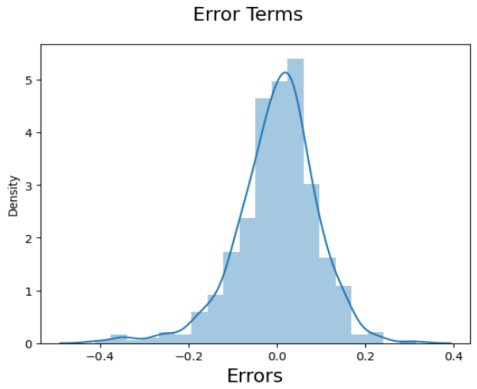
Answer :

Temp has the highest correlation with the target variable.

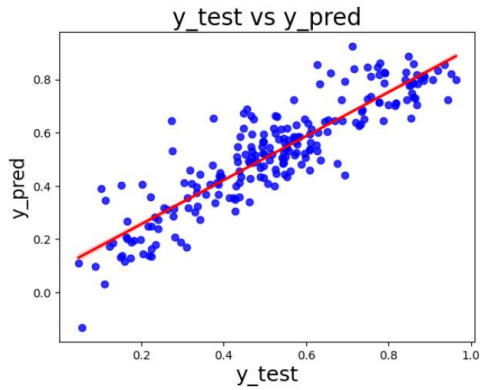
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Answer : I have validated the assumptions of linear regression as follows:

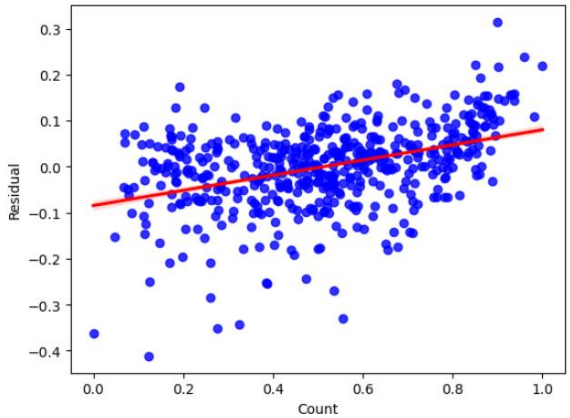
- **Normality of error terms:**
  Error terms should be normally distributed and centered around zero. To validate this I have plotted a distplot on my training data as per below screenshot.



- **Linearity**: Assumes that the relationship between the independent and dependent variables is linear. Below plot validates this assumption.



- **Homoscedasticity**: Assumes that the variance of the errors is constant across all levels of the independent variable. There should be no visible pattern in residual values. Below plot validates this assumption.



- **Independence**: Assumes that the observations are independent of each other. No auto-correlation. It was validated using correlation matrix.
- **No multicollinearity**: There should be insignificant multicollinearity among the variables. A correlation matrix was constructed to validate this assumption

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: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

1. **Temp** : Increase in temp will increase the demand. **(positive correlation)**
2. **Year :** with increase in consecutive year demand is rising. **(positive correlation)**
3. **Light snow :** decrease in light snow will increase in demand for shared bikes **(negative correlation)**

General Subjective Questions

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

Answer : Linear regression is a supervised machine learning method that is used to train a model and find a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

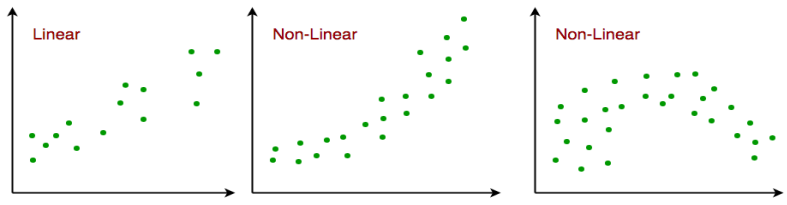
The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + ...$$

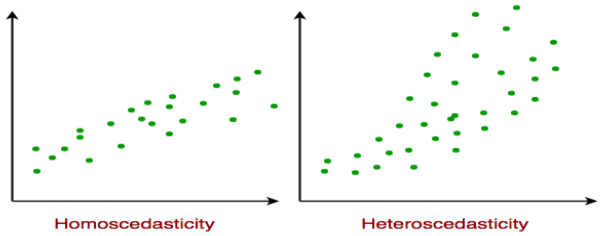
In the example above, y is the dependent variable, and x1, x2, and so on, are the explanatory variables. The coefficients (b1, b2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

Assumptions:

1. Linearity: Assumes that the relationship between the independent and dependent variables is linear.



2. Independence: Assumes that the observations are independent of each other. No auto-correlation
3. Homoscedasticity: Assumes that the variance of the errors is constant across all levels of the independent variable. There should be no visible pattern in residual values



4. Normality: Assumes that the errors are normally distributed and centered around zero.
5. No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it difficult to determine the individual effect of each variable on the dependent variable.

The two main types of Linear regression:

- Simple Linear Regression : When the number of independent variables is 1
- Multiple Linear Regression: When the number of independent variables is more than 1

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

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

Answer :

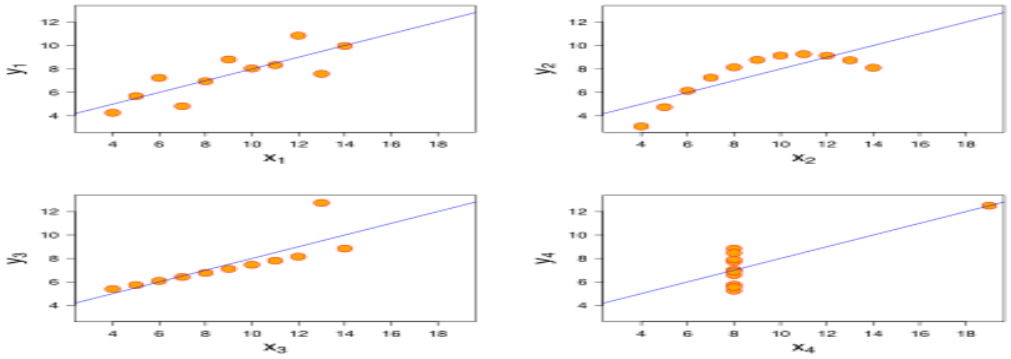
Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

	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

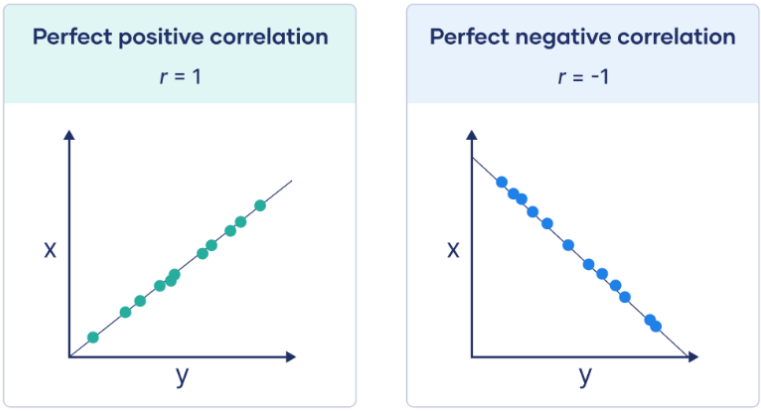
Mean for x = 9
Mean for y = 7.50
Standard Deviation for x = 3.32
Standard Deviation for y = 2.03
Even if the Statistical data is same for all data set the graph representation is different for all.



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

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.
- If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
- The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.
- 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:



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 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. Example: If an algorithm is not using feature scaling method then it can consider the value 4000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes.

The goal of scaling is to bring all features to a comparable level, ensuring that no particular feature dominates or biases the machine learning model due to its scale. Scaling is particularly important for algorithms that are sensitive to the magnitude of input features.

Difference Between Normalized Scaling and Standardized Scaling:

Normalized Scaling:

- Also known as Min-Max scaling, this method scales the features to a specific range, usually [0, 1]. The formula for normalized scaling is:

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

- If the data contains outliers and does not follow a Gaussian distribution, normalized scaling might be a better choice.
- MinMax Scaler is often used as an alternative to Standard Scaler if zero mean and unit variance want to be avoided.

Standard Scaler :

- Also known as z-score normalization, standardized scaling transforms the features to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

$$z = \frac{x - \mu}{\sigma}$$

- This method assumes that the data follows a Gaussian distribution. It is less sensitive to

outliers compared to normalized scaling. If outliers are not a major concern, standardized scaling is often recommended.

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

Answer:
The scenario where the VIF becomes infinite is often a result of perfect multicollinearity. Perfect multicollinearity occurs when one predictor variable in a regression model can be exactly predicted from the other variables with perfect accuracy.
The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess the extent of multicollinearity among predictor variables. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to issues in estimating the individual contribution of each variable to the dependent variable.
The formula for VIF is:
ViF = 1 / 1 – Ri^2

Common causes of perfect multicollinearity leading to infinite VIF include:

- Duplicated or Linearly Dependent Predictors:

Two or more predictors are identical or can be expressed as a linear combination of each other. For example, if you have two columns that are exact duplicates or are related by a constant factor, this can lead to perfect multicollinearity.

- Incorrect Specification of the Model:

Including a variable that is a sum or difference of other variables in the model can introduce perfect multicollinearity.

- Data Issues:

Extremely rare data points or data errors can sometimes lead to perfect multicollinearity.

Dealing with infinite VIF values involves identifying and addressing the source of multicollinearity in the data. This might include removing one of the perfectly correlated variables, transforming variables, or reconsidering the model specification. In practice, infinite VIF values are a sign that the model needs adjustment to handle the multicollinearity issue appropriately.

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

Answer:
A QQ 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.
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.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. 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 value. A 45-degree reference line is also plotted. If the two sets come from a population with the

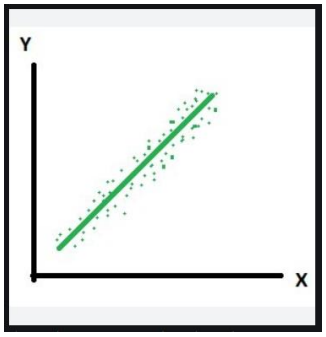
same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance :

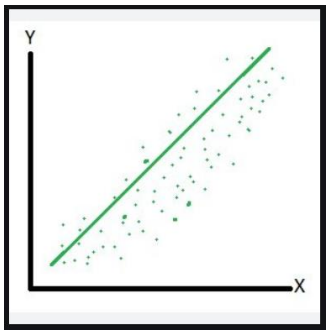
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

Interpretation :

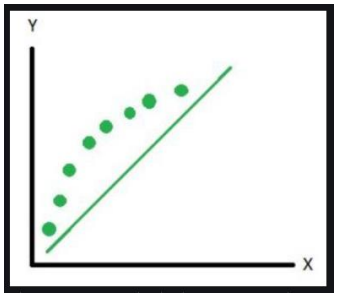
1. All point of quantiles lie on or close to straight line at an angle of 45 degree from x – axis. It indicates that two samples have similar distributions.



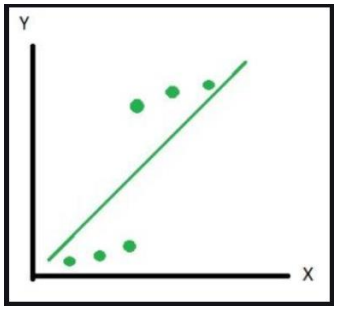
2. The y – quantiles are lower than the x – quantiles. It indicates y values have a tendency to be lower than x values.



3. The x – quantiles are lower than the y – quantiles. It indicates x values have a tendency to be lower than the y values.



4. Indicates that there is a breakpoint up to which the y – quantiles are lower than the x – quantiles and after that point the y – quantiles are higher than the x – quantiles.


Y
X