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:

   Effect of categorical variables on the dependent variable:

   - Fall season attracted more bookings, and booking counts increased in each season from 2018 to 2019.
   - Most bookings occurred from May to October.
   - Clear weather attracted more bookings.
   - Bookings were higher on Thu, Fri, Sat, and Sun.
   - Fewer bookings were made on holidays.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Answer:
   It reduces multicollinearity by avoiding redundancy in dummy variables, thus reducing correlations created among them.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Answer:
   The '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:
   The assumptions validated include:
   - Normality of error terms
   - Multicollinearity check
   - Linear relationship
   - Homoscedasticity
   - 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:
   'temp', 'winter', and 'sep' are the top 3 features contributing significantly to the demand for shared bikes.

General Subjective Questions

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

Answer:

Linear regression is a statistical model that examines the linear relationship between a dependent variable and one or more independent variables. A linear relationship means that as the independent variable(s) change, the dependent variable changes in a corresponding manner. The relationship is expressed by the equation

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

The relationship can be positive (both variables increase) or negative (one increases while the other decreases).

Types of Linear Regression:

- Simple Linear Regression: One independent variable.

- Multiple Linear Regression: More than one independent variable.

Assumptions:

1. Minimal multicollinearity (independent variables are not highly correlated).

2. No autocorrelation (errors are independent).

3. Linear relationship between variables.

4. Error terms are normally distributed.

5. Homoscedasticity (consistent variance in residuals).

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

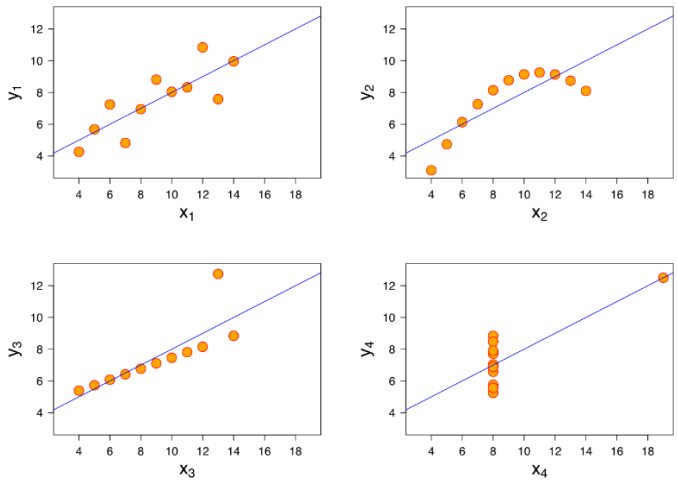
Answer:

Anscombe's Quartet, created by Francis Anscombe, consists of four datasets, each with 11 (x, y) pairs. Despite sharing identical descriptive statistics, their graphs tell completely different stories.

	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

Each dataset has:

- A mean of 9 for x and 7.5 for y,
- A variance of 11 for x and 4.13 for y,
- A correlation coefficient of 0.816 between x and y.



However, when plotted:

- Dataset I shows a clean, linear relationship.
- Dataset II is not normally distributed.
- Dataset III is linear but skewed by an outlier.
- Dataset IV is dominated by a single outlier.

This highlights the importance of data visualization in revealing deeper insights that summary statistics alone cannot show.

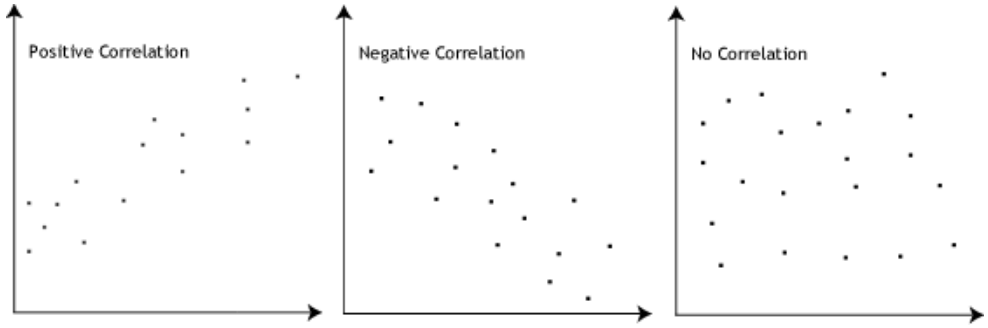
3. **What is Pearson's R?** **(3 marks)**

Answer:

Pearson's r measures the strength of the linear relationship between two variables. A positive correlation means both variables increase or decrease together, while a negative correlation indicates that as one variable increases, the other decreases.

The Pearson correlation coefficient, r, ranges from +1 to -1:

- r = 0 means no correlation.
- r > 0 indicates a positive relationship.
- r < 0 indicates a negative relationship.



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

Answer:

Feature scaling is a technique used to standardize the range of independent variables in the data. It helps handle differences in magnitude, units, or values during preprocessing. Without scaling, machine learning algorithms may assign more weight to larger values and overlook smaller ones, leading to inaccurate predictions.

For example, without scaling, an algorithm might interpret 3000 meters as greater than 5 kilometres, which is incorrect. Scaling ensures that all features are on the same scale, improving the model's accuracy.

Criteria	Normalized Scaling	Standardized Scaling
Definition	Uses minimum and maximum values to scale data.	Uses mean and standard deviation to scale data.
Purpose	Useful when features have different scales.	Ensures data has zero mean and unit variance.
Range	Scales values between [0, 1] or [-1, 1].	Not limited to a specific range.
Sensitivity to Outliers	Highly affected by outliers.	Less affected by outliers.
Scikit-learn Transformer	MinMaxScaler	StandardScaler

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

Answer:

If the VIF (Variance Inflation Factor) is infinite, it indicates perfect correlation between two independent variables. A high VIF means multicollinearity is present, inflating the variance of the model coefficients. For instance, a VIF of 4 means the variance is inflated four times due to multicollinearity.

When VIF is infinite, the R-squared (R^2) value is 1, leading to an undefined result in the calculation $1/(1 - R^2)$. To fix this, you need to remove one of the perfectly correlated variables from the dataset to resolve the multicollinearity issue.

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

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 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.

Importance of Q-Q plot:

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.