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:

Following are the insights :

1. Fall season seems to have attracted more booking.
2. Most of the bookings have been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
3. Clear weather attracted more booking which seems obvious
4. Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
5. When its not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
6. Booking seemed to be almost equal either on working day or non-working day

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 removes the repetition of a variable. This helps in removing the duplicate variable which will cause more issues while modelling the data. Like VIF will go to inf as both variables will present the exact same data

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

Answer:

'temp' and atemp variables has the highest correlation with the target variable. At 0.63 followed by year flag.

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

Answer:

Validated the assumption of Linear Regression Model based on below 5 assumptions -

a) Normality of error terms
   i) Error terms should be normally distributed

b) Multicollinearity check
    i) There should be insignificant multicollinearity among variables.
c) Linear relationship validation
    i) Linearity should be visible among variables
d) Homoscedasticity
    i) There should be no visible pattern in residual values.
e) Independence of residuals
    i) No auto-correlation

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 –

- temp
- lightsnow_rain
- year

General Subjective Questions

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

Answer:

Linear regression is a fundamental and widely used statistical and machine learning technique. The goal of linear regression is to model the relationship between a dependent variable (target or response) and one or more independent variables (predictors or features) by fitting a linear equation to the observed data.

Components of Linear Regression

1. **Dependent Variable (Y)**:
    - This is the variable we are trying to predict or explain.
2. **Independent Variables (X)**:
    - These are the variables that we use to make predictions.
3. **Linear Relationship**:
    - The relationship between the dependent and independent variables is assumed to be linear. This can be represented by the equation:

$$Y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon$$

    - Where:
        - Y is the dependent variable.
        - X1,X2,…, Xn are the independent variables.
        - β0 is the intercept.
        - β1,β2,…,βn are the coefficients (slopes) for each independent variable.
        - ϵ is the error term, representing the difference between the observed and predicted values.

Steps in Linear Regression

1. **Data Collection**:
   - Gather the data that contains the dependent and independent variables.
2. **Data Preprocessing**:
   - Handle missing values, normalize or standardize the data, and convert categorical variables to numerical ones using techniques like one-hot encoding.
3. **Model Fitting**:
   - Fit the linear regression model to the data using methods like Ordinary Least Squares (OLS). The goal is to find the coefficients β\betaβ that minimize the sum of the squared residuals (differences between the observed and predicted values).
   - The OLS method minimizes the cost function:

$$\text{Cost} = \sum_{i=1}^{m} (Y_i - \hat{Y}_i)^2$$

     Where m is the number of observations, Yi is the actual value, and Y^i is the predicted value.

4. **Model Evaluation**:
   - Evaluate the performance of the model using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).
5. **Model Interpretation**:
   - Interpret the coefficients to understand the relationship between the dependent and independent variables. Each coefficient represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.

Assumptions of Linear Regression

1. **Linearity**:
   - The relationship between the dependent and independent variables is linear.
2. **Independence**:
   - The observations are independent of each other.
3. **Homoscedasticity**:
   - The residuals (errors) have constant variance at every level of the independent variables.
4. **Normality**:
   - The residuals are normally distributed.
5. **No Multicollinearity**:
   - The independent variables are not highly correlated with each other.

Visualization and Interpretation

- The scatter plot you provided seems to show a simple linear regression with a single independent variable and a dependent variable.
- The linear regression line is shown, indicating the best fit line through the data points.

Conclusion

Linear regression is a powerful tool for understanding and predicting relationships between variables. However, it's essential to ensure that the assumptions of linear regression are met to produce valid and reliable results. When applied correctly, linear regression can provide valuable insights and predictions for various types of data.

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

Answer:

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and the effect of outliers and the structure of the data on statistical properties.

The Four Datasets

Each of the four datasets consists of eleven (x, y) points. Despite having nearly identical statistical properties, including:

- Mean of x
- Mean of y
- Variance of x
- Variance of y
- Correlation between x and y
- Linear regression line y=mx+by = mx + by=mx+b
- Coefficient of determination (R2R^2R2)

The datasets are structured very differently, showing that relying solely on these statistical metrics can be misleading without visualizing the data.

	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

Detailed Explanation of Each Dataset

1. **First Dataset (I)**:
    - This is a simple linear relationship with a small amount of random noise. When plotted, it looks like a standard scatter plot with points following a linear trend.
2. **Second Dataset (II)**:

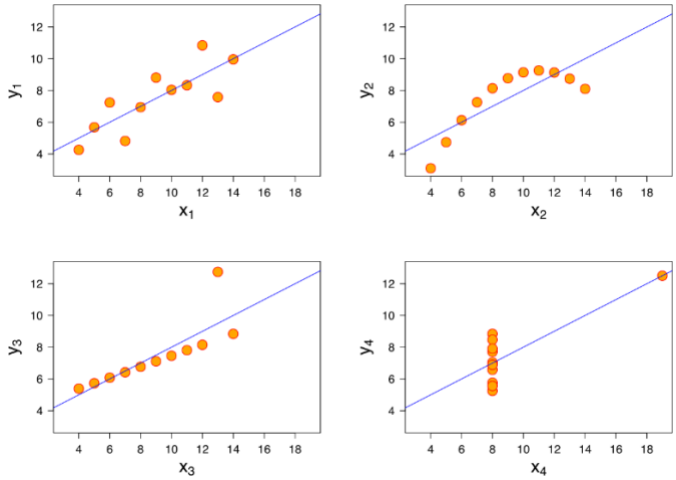
- Here, the x-values are nearly constant except for one point. This leads to a situation where most of the points are in a vertical line, with one point influencing the regression line heavily.

3. **Third Dataset (III)**:
   - This dataset has a perfect linear relationship except for one outlier. The outlier significantly affects the mean, variance, and correlation, but the linear trend is apparent when the outlier is removed.
4. **Fourth Dataset (IV)**:
   - This dataset forms a horizontal line except for one point. The single outlier affects the regression line, leading to misleading statistics if not visualized.

Visual Representation

Graphical representation of these datasets reveals their differences:

- **Dataset I**: Shows a clear linear relationship with slight deviations.
- **Dataset II**: Reveals a vertical alignment with a single influential point.
- **Dataset III**: Exhibits a linear trend disrupted by an outlier.
- **Dataset IV**: Demonstrates a horizontal line influenced by an outlier.



Importance of Anscombe's Quartet

Anscombe's quartet highlights several important lessons for data analysis:

1. **Visualization is Crucial**:
   - Visualizing data can reveal patterns, trends, and outliers that are not apparent from summary statistics alone.
2. **Influence of Outliers**:
   - Outliers can significantly affect statistical measures and regression lines, leading to potentially misleading conclusions.
3. **Context Matters**:
   - Understanding the context and structure of the data is essential for accurate interpretation and analysis.
4. **Limitations of Summary Statistics**:
   - Simple descriptive statistics can be insufficient to capture the true nature of the data distribution and relationships.

Conclusion

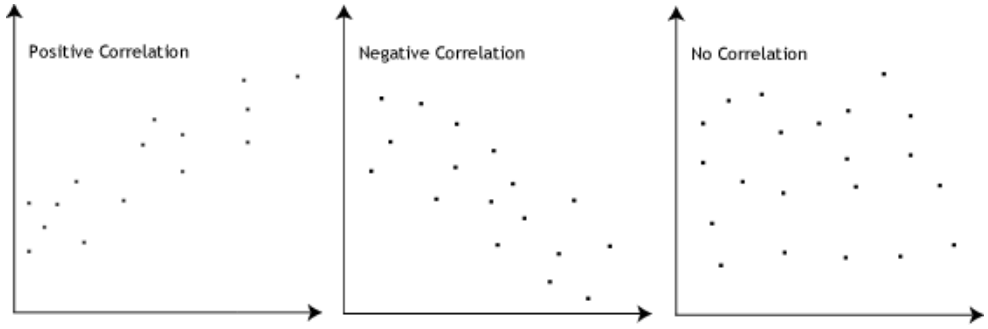
It emphasizes the importance of considering the underlying data distribution and the potential impact of outliers on statistical analysis

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

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the 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, 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:

Feature 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 3000 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 and thus, tackle this issue.

S.NO.	Normalized scaling	Standardized scaling
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.

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

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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.