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:

The dataset contains several categorical features, including Season, Year, Month, Weekday, Weathersit, Workingday, and Holiday. To analyze their impact on the dependent variable (cnt), I used bar plots and box plots.

- **Season**: The highest bike rental counts were observed in the Fall, surpassing other seasons such as Spring, Summer, and Winter, indicating higher demand during this time.
- **Year**: There was significant growth in bike rentals from 2018 to 2019, with the total count increasing from 1.2 million to 2 million—a 60% increase, reflecting substantial business growth.
- **Month**: Bike demand was strong from January to September, but began to decline starting in October, continuing through December.
- **Weekday**: The demand for bikes remained fairly consistent across all days of the week.
- **Weathersit**: The highest usage occurred when the weather was clear, while usage was moderate during misty or cloudy conditions, and lowest in light snow or rain.
- **Working Day**: Bike rentals were higher on working days, with similar counts observed in both 2018 and 2019.
- **Holiday**: There was noticeably lower bike usage on holidays, with fewer people opting for shared bikes.

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:

Using drop_first=True is important because it helps avoid creating extra columns when generating dummy variables, thereby reducing multicollinearity among them. For example, if a categorical column has three categories (e.g., furnished, semi-furnished, and unfurnished), creating dummy variables for all three is unnecessary. If a property is not furnished or semi-furnished, it must be unfurnished. Thus, only two dummy variables are needed to represent the three categories.

In general, if a categorical variable has *n* levels, using *n-1* columns is sufficient to represent the dummy variables.

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: <Your answer for Question 3 goes below this line> (Do not edit)

The variables 'temp' and 'atemp' show the strongest correlation with the target variable (cnt).

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:

- **Linear Relationship**: The relationship between the independent and dependent variables should be linear.
- **Independence of Residuals**: The residuals should be independent, with no autocorrelation.
- **Homoscedasticity**: Residuals should display constant variance with no discernible pattern, as seen in the predicted vs. residuals plot.
- **Normality of Residuals**: The residuals should follow a normal distribution, as illustrated in the attached image.
- **No Multicollinearity**: Independent variables should not be excessively correlated. The Variance Inflation Factor (VIF) was used to detect multicollinearity, and variables with high VIF values were removed.
- **No Endogeneity**: There should be no correlation between the error terms and the independent variables.

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:
The following three features significantly contribute to the demand for shared bikes:

- temp
- atemp
- year

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)

Answer:

Linear Regression is the supervised Machine Learning model used for predicting a continuous output
variable based on one or more predictor variables.
There are two types of Regression.

- Simple:- Only one independent variable and one dependant feature.
- Multiple:- More than one independent variable and one dependant feature.

Simple Linear Regression:
This involves only one independent variable and one dependent variable. The equation for simple
linear regression is: $y= \beta 0 + \beta 1 \, X$
where:

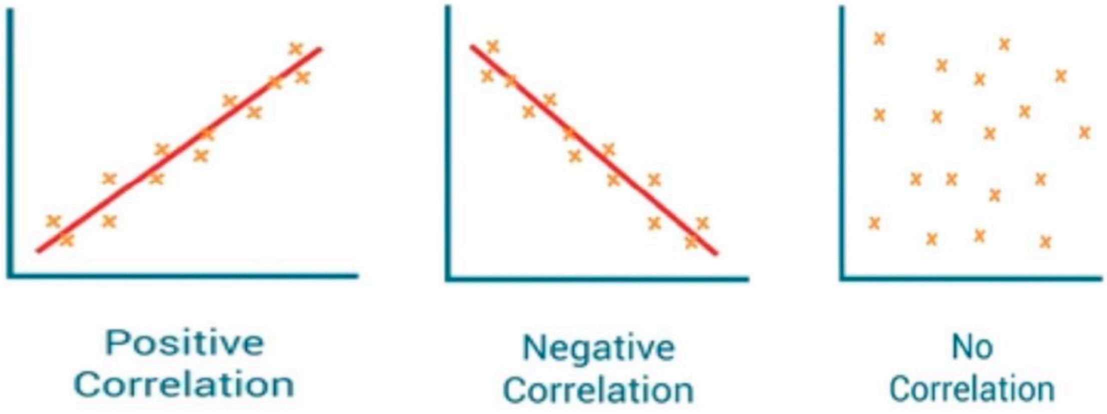
- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression:
This involves more than one independent variable and one dependent variable. The equation for
multiple linear regression is: $y= \beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots\ldots\ldots \beta n \, Xn$

- Y is the dependent variable
- X1, X2, …, Xn are the independent variables
- B0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values
based on the independent variables.

There are 3 types of relationships they are described as below.➢ Positive Correlation: If one variables increases tends to increase the value of other variable.
➢ Negative Correlation: If one variables increases tends to decrease the value of other variable.
➢ No Correlation: If one variable increases, the other variable may or may not increase. It can



Positive Correlation

Negative Correlation

No Correlation

either increase or decrease.

Assumptions of linear regression include:
1. Linearity: The relationship between the dependent and independent variables is linear.
2. Independence: The observations are independent of each other.
3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
4. Normality: The errors follow a normal distribution.
5. No multicollinearity: The independent variables are not highly correlated with each other.
6. No endogeneity: There is no relationship between the errors and the independent variables.
7. Autocorrelation: There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Auto correlation

R-Squared:

R-square(R^2) is also known as the coefficient of determination, It is the proportion of variation in Y explained by the independent variables X. It is the measure of goodness of fit of the model.Higher the R^2, the more variation is explained by your input variable and hence better is your model

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 → coefficient of determination

RSS → sum of squares of residuals

TSS → Total sum of Squares.

Cost Function For Linear Regression

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

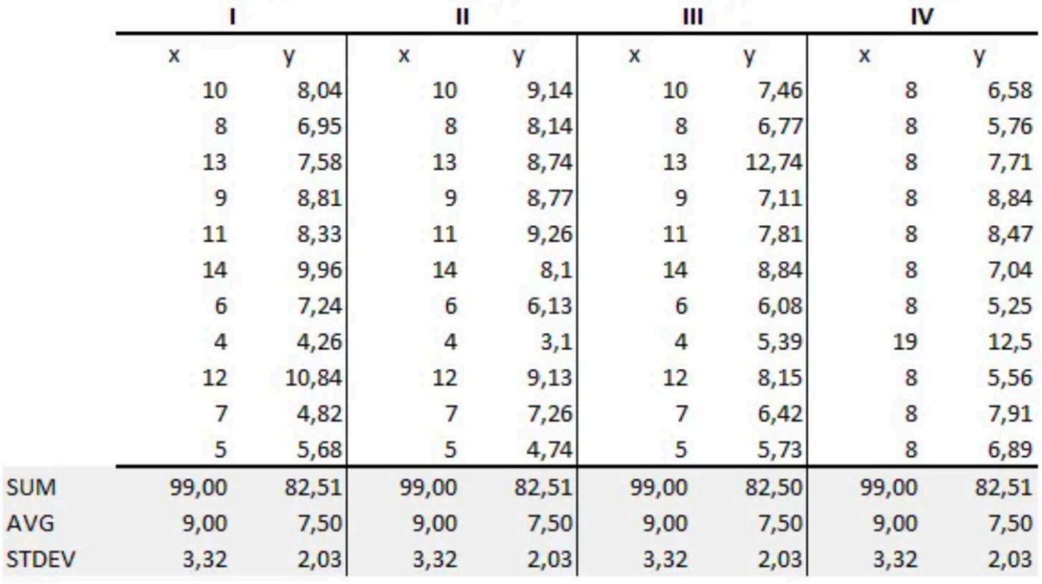
n → number of data points

yi → actual value

ŷ → predicted value.

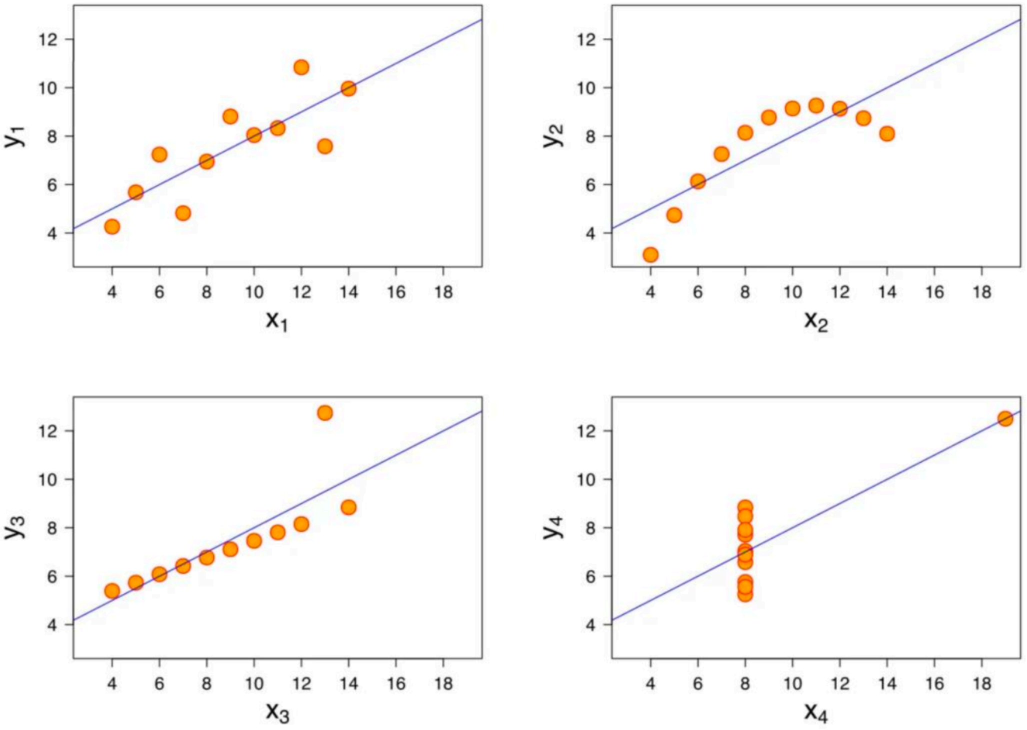
Question 7. Explain the Anscombe’s quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

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

Quartet’s Summary Stats



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 appears 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.

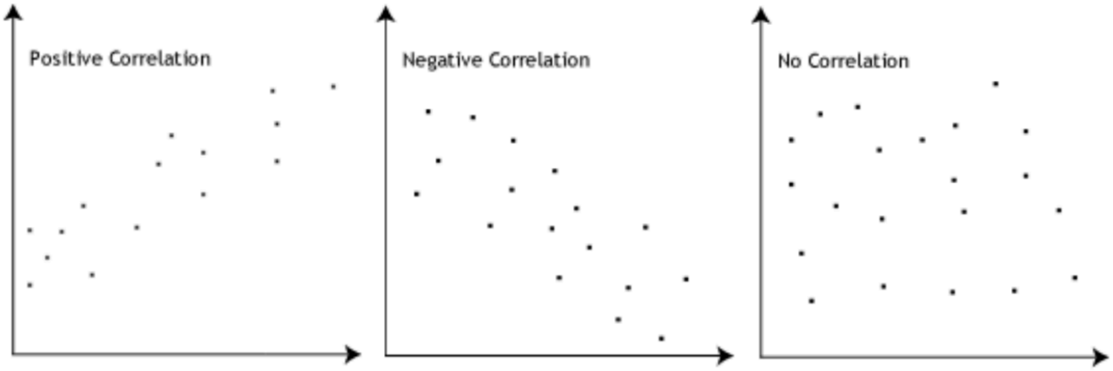
Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

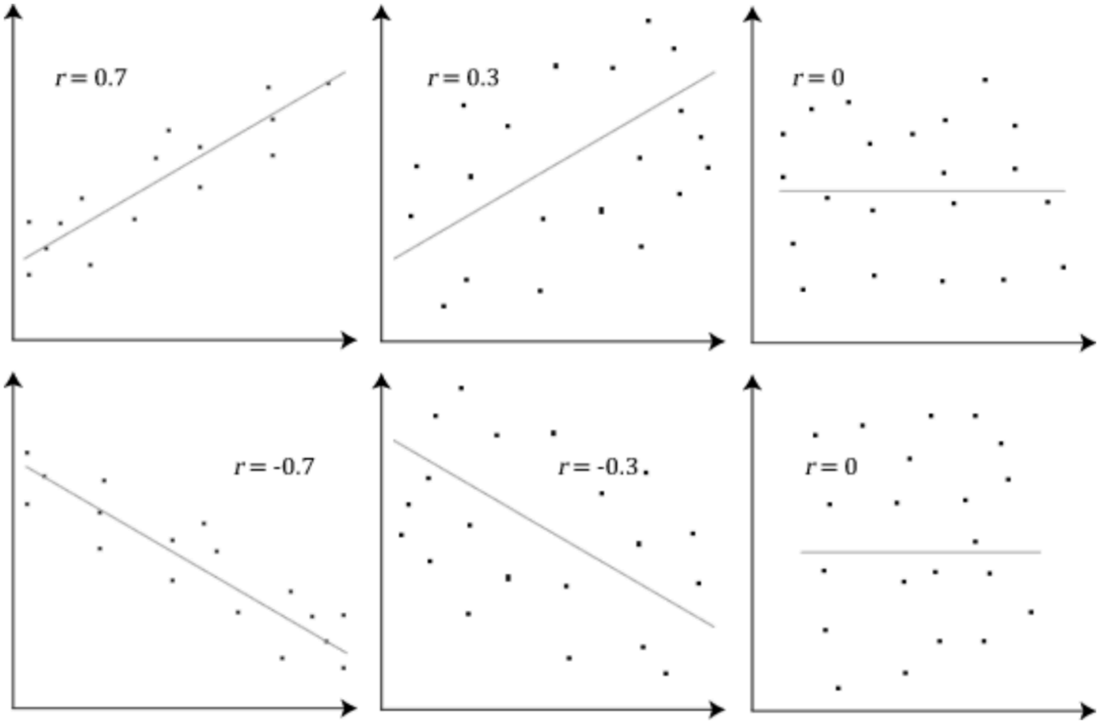
Pearson correlation coefficient, is a measure of the strength of a linear association between two variables and is denoted by r

The mathematical representation is below.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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:





- R=1 → perfect positive corelation
- R is between 0 to 1 → positive corelation
- R is 0 → No corelation
- R is -1 to 0 → negative corelation
- R is -1 → perfect negative corelation

heir values lies between -1 to 1 they are described above.

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:

Data scaling is the process of transforming the values of the features of a dataset till they are within a specific range, e.g. 0 to 1 or -1 to 1. This is to ensure that no single feature dominates the distance calculations in an algorithm, and can help to improve the performance of the algorithm 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.

	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	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
7	MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$	Standardisation: $x = \frac{x - mean(x)}{sd(x)}$

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 value of VIF is calculated by the below formula:

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

When we calculate the VIF for one independent variable using all the other independent variables, if the R^2 value we get equal to 1 then VIF will become infinite. This is quite possible when one of the independent variables is strongly correlated with many of the other independent variables. It denotes perfect correlation in variables

A rule of thumb for interpreting the Variance Inflation Factor:

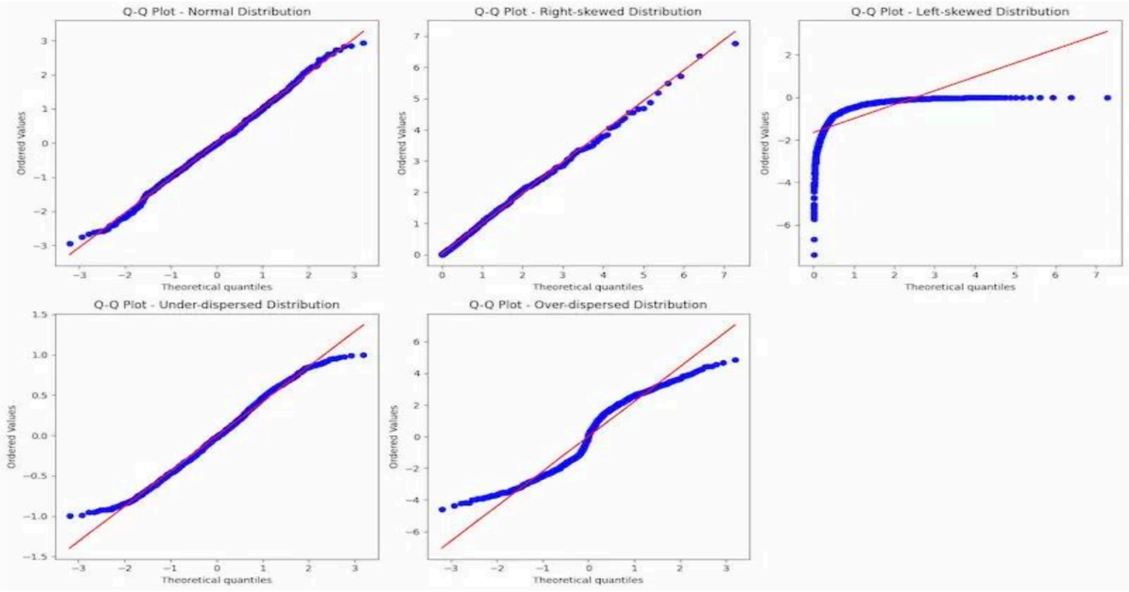
- VIF is 1 = not correlated.
- VIF Between 1 and 5 = moderately correlated.
- VIF Greater than 5 = highly correlated.

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:

The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption.

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

Below is the different distribution using QQ plot


Q-Q Plot - Normal Distribution
Ordered Values
Theoretical quantiles
Q-Q Plot - Right-skewed Distribution
Ordered Values
Theoretical quantiles
Q-Q Plot - Left-skewed Distribution
Ordered Values
Theoretical quantiles
Q-Q Plot - Under-dispersed Distribution
Ordered Values
Theoretical quantiles
Q-Q Plot - Over-dispersed Distribution
Ordered Values
Theoretical quantiles