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: Categorical variables are analysed using box plot. Inferences from the analysis are,

   - Bikes are rented more in fall and summer compared to other seasons (winter and spring)
   - Bikes are rented more between May and September compared to other months
   - Bikes are rented more when weather condition is clear or mist
   - Bikes are rented more on Saturday followed by sunday
   - Bikes are rented more in 2019 compared to 2018. Looks like the service usage is increasing YOY
   - Bikes are rented less when it's a holiday

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

   Answer: drop_first is used to get k-1 dummies out of k categorical levels (removing the first level)

   This is important because of two reasons:

   - Redundant info is not good for modelling. It increases time.
   - It reduces/prevents multicollinearity. Kth level can be inferred / formed using k-1 levels in categorical column.

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

   Answer:

   'temp' variable has the highest correlation with the target variable 'total_count'

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

   Answer:

   - Calculated residuals and verified error distribution
     - Error distribution is centred around 0
     - Error distribution follows normal distribution
   - All the predictor variables have linear relationship (positive or negative) with target variable 'total_count'
   - Error terms are independent of each other
   - Homoscedasticity - Error terms have constant variance
   - There's no multicollinearity between predictor variables – All of them has VIF < 5

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
2. year
3. winter

General Subjective Questions

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

Answer:

Linear regression - Supervised machine learning model

- Computes linear relationship between target/dependent variable and one or more independent/predictor variables.
- Linear relationship can be positive or negative.

For example:

If independent variable increases, dependent variable increases - positive relationship

If independent variable increases, dependent variable decreases - negative relationship

Linear regression can be mathematically represented as:

$Y = mX + c$

Here, Y = dependent variable

X = independent variable

m = slope of the line

c = intercept

Goal of linear regression model - Find the best fit line which describes the data well with least residual sum of squares

residual - difference between actual and predicted value

Linear regression can be classified into two types:

1. Simple linear regression - One independent variable and one target variable
2. Multiple linear regression - More than one independent variable and one target variable

Assumptions/Considerations of linear regression:

Simple linear regression:

1. Linear relationship should exists between X and Y
2. Error terms are normally distributed with mean centered around 0
3. Error terms are independent of each other
4. Homoscedasticity - Error terms have constant variance / standard deviation

Multiple linear regression:

1. Assumptions made in simple linear regression holds true.
2. Prevent overfitting
   - Overfit - training accuracy is high while test accuracy is low.
   - Model should not be too complex
3. Multicollinearity
4. Associations between predictor variables should be low. Use Variance Inflation Factor (VIF) to calculate this
5. Feature selection
   - Select the optimal set of features
   - Best approach is to use combined strategy - Recursive Feature Elimination (RFE) + Manual feature selection

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

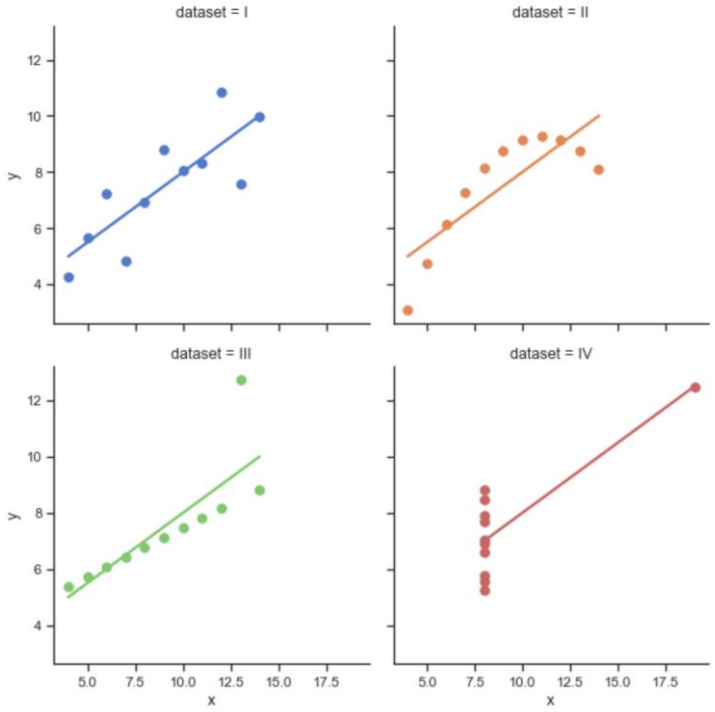
Answer:

- Anscombe's quartet (constructed by statistician Francis Anscombe) can be defined as a group of four datasets which are nearly identical in descriptive statistics (mean, variance, standard deviation) but have very different distributions when plotted on scatter plots.
- This quartet tells us the importance of plotting the graphs (visualizing the data) before building the model that can help us identify the various anomalies present in the data like outliers.

Example:

	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

For the above dataset, sum, mean, standard deviation are all same for x and y across groups. But plotting the data in graphs (as below), tells a different story about the data.



- Dataset = I – shows a linear relationship between x and y with some variance
- Dataset = II – shows a curve shape but doesn't show a linear relationship
- Dataset = III – looks like a tight linear relationship between x and y, except for one large outlier
- Dataset = IV – looks like value of x remains constant for all y except one outlier

This quartet illustrates the importance of data visualization for better understanding of data.

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

Answer:

Pearson's R (also known as Pearson's correlation coefficient) is a statistical measure of the relationship between two variables.

It also shows the strength and direction of the relationship (represented numerically)

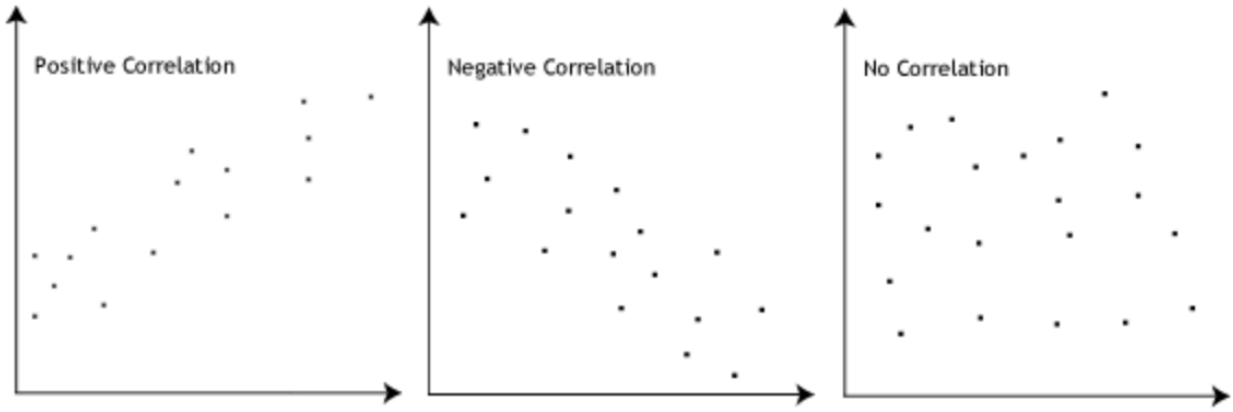
Correlation coefficient lies between -1.0 and +1.0

If correlation coefficient is positive, both the variables change in a proportion and in the same direction.

If correlation coefficient is negative, both the variables change in a proportion and in the opposite direction.

If correlation coefficient is 0, no linear relationship between those two variables.

Note: If correlation coefficient is +1.0, it's a perfect positive correlation



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

Answer:

Feature scaling – It's a process of normalizing the features in a dataset. In other words, standardizing the feature values to compare it on a same scale. It's done during data pre-processing step (before training the model on training set).

Scaling is performed because,

Real-world datasets will have features in different ranges / units / magnitude. In order to interpret these features on the same scale (easy interpretation) and for faster convergence, need to perform scaling.

S.No.	Normalized Scaling	Standardized Scaling
1.	Minimum and Maximum value of features are used for scaling	Mean and standard deviation are used for scaling. Centers data around the mean and scales to a standard deviation of 1
2.	Scales value between between [0, 1] or [-1, 1]	It's not bounded to a certain range
3.	Formula: (X-min) / (max-min)	Formula: (X-mean) / standard deviation
4.	Sensitive to outliers	Less sensitive to 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's a perfect correlation between variables, then value of VIF becomes infinite.

Formula for VIF is 1 / 1- R^2

If R-squared (R^2) value is 1 (perfect correlation), then the denominator of above formula will become 0 and overall value become infinite.

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

Answer:

A Q-Q plot is a scatterplot created by plotting two set 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. If they don't, it means, residuals are not normal (Gaussian) and thus, errors are also not normal.

Use of Q-Q plot in linear regression:

- To know whether two samples of data came from the same population or not.
- To know whether two samples have the same tail.
- To know whether two samples have the same distribution shape.

Importance of Q-Q plot in linear regression:

- Comparing two or more datasets would be helpful in machine learning (example: train and test) to see the distribution is same.

- If the compared datasets differ, then, it’s also useful to understand the nature of differences.