Assigment 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)

Categorical analysis is done to see the effect of independent variable on dependent variable . I have anlaysed the categorical variables using boxplot and barplot.

1. Bike riding is generally avoided during the spring season.
2. Fall and summer seasons are preferred for bike riding.
3. There was an increase in demand for bike riding in 2019 compared to 2018.
4. Demand for bike riding is slightly lower on holidays than on non-holidays.
5. Light snow conditions lead to a decrease in bike riding demand.
6. Demand for bike riding is very high when the weather is clear.
7. Maximum bike sales occur in the months of May, June, July, August, and September.

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

Drop first =True is used to drop the first category in each categorical variable. Because it is easy to predict n number of variables use n-1 variable.

If we will not drop the first category it will create the multicollinearity because that category can be easily predicted by the other variables, due which issue will create in a model.

Understanding the dependability will become more clear of the independent variable on the dependent variable.

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

'temp' variable has the highest correlation with the target variable count because people are likely to use bike during summer or clear weather condition.

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

Linear regression model is done on the basis of four assumptions:

- ➔ Mean of the distribution of error is Zero.
- ➔ All the error are independent of each other.
- ➔ There should be linear relationship among variables
- ➔ Multicollinearity among the predictor variables should be insignificant which checked using VIF where VIF should be less than 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)

Top three features explaining the demand of bike share.

→Target variable is positively correlated with 'temp' and 'season_winter

→Target variable is negatively correlated with 'windspeed'.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Introduction: Linear regression models the relationship between a dependent variable (target) and independent variables (predictors) with a linear approach.

Types:

1. **Simple Linear Regression:** equation = y = mx+c

➔ y: dependent variable
➔ x: independent variable
➔ m: slope – change in y variable because of the change in x- variable
➔ c: y-intercept when x=0

2. **Multiple Linear Regression: Equation** ➔ y = b0 + b1*x1 + b2*x2 + ... + bn*xn + e

➔ y is the dependent variable
➔ x1, x2, ..., xn are the independent variables
➔ b0, b1, b2, ..., bn are the coefficients
➔ e is the error term

Step for model building:

1. **Data Collection:** Gather data.
2. **Model Creation:** Define the model.
3. **Parameter Estimation:** Use OLS to estimate coefficients.
4. **Model Evaluation:** Assess with R-squared, MSE, RMSE, Adjusted R-squared.

Assumptions:

➔ **Linearity:** Linear relationship.
➔ **Independence:** Independent observations.
➔ **Homoscedasticity:** Constant residual variance.
➔ **Normality:** Normally distributed residuals.

Challenges:

- ➔ **Multicollinearity:** High correlation between predictors.
- ➔ **Outliers:** Affect performance.
- ➔ **Heteroscedasticity:** Variable residual variance.
- ➔ **Model Selection:** Choose right predictors.

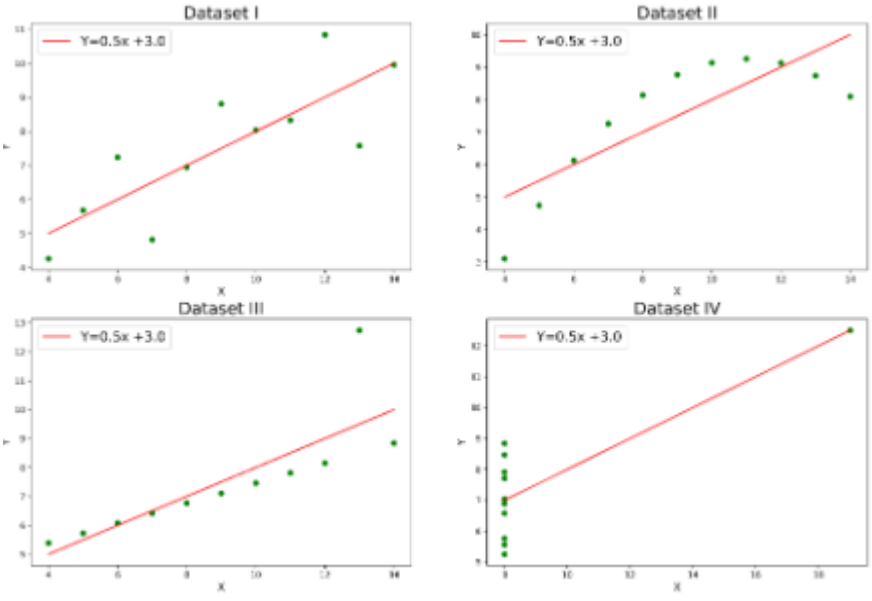
2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's Quartet Dataset

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The scatter plot and linear regression line for each datasets



Inferences:

- ➔ In the first one if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- ➔ In the second one if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- ➔ In the third one you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- ➔ Finally, the fourth one shows an example when one high-leverage point is enough to produce a high correlation coefficient.

The Quatret emphasis on the importance of data visualization just by looking at statistics can not help in interpreting the data is important to visualize through graphs.

3. What is Pearson's R?

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by r and ranges from -1 to 1.

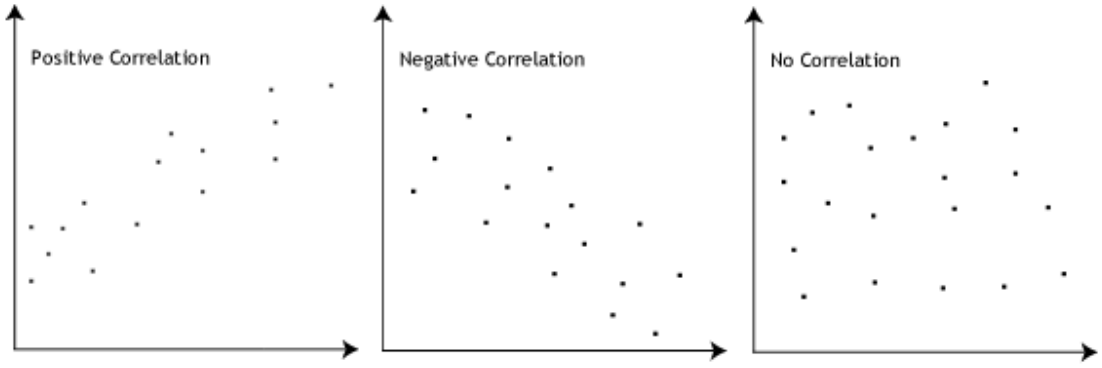
Formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Interpretation:

- ➔ A r=0 indicates that there is no association between the two variables.
- ➔ A r>0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
- ➔ A r<0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Figure for better understanding:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling : Scaling is done in multilinear regression model on dataset to make the value in numeric column under specific range. If scaling will not be done it will result into large coefficient values.

Two Types of scaling method are there:

- ➔ Standardization
- ➔ Min Max scaling

Standardization	Min Max Scaling
Standardization transforms features to have a mean of 0 and a standard deviation of 1.	Mix max scaling rescale data to fix range between 0 and 1.
Less sensitive to outliers.	Sensitive toward outliers.
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?

1. **Perfect Multicollinearity:**
    - This occurs when one predictor is explained by the other predictors.
    - In such cases, the matrix used to calculate the VIF becomes singular or non-invertible, leading to an infinite VIF value.
2. **Redundancy in Predictors:**
    - Redundancy is created when predictor is not adding any additional information.
    - This redundancy makes the variance of the coefficient infinitely large.

Implications:

Infinite VIF indicates that the predictor should be removed or combined with other predictors to solve multicollinearity issues.

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

A Q-Q plot is a graphical tool used to compare data distribution with theoretical distribution by plotting quantile of both the dataset.

Steps to Create a Q-Q Plot:

- ➔ **Calculate Quantiles:** Compute the quantiles of the dataset and the theoretical distribution.
- ➔ **Plot Points:** Plot the quantiles of the dataset against the quantiles of the theoretical distribution.
- ➔ **Draw Reference Line:** Draw a 45-degree reference line for comparison.

Importance of Q-Q plot

- ➔ It validates the normality if the residuals align closely with the straight line.
- ➔ Check outliers if point are deviating from the straights which can affect the model accuracy

It is used to check the appropriateness of the model and also for the reliability of the model