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: Analysis on the categorical columns is done using the box plot

- In Fall season demand for the bikes is most followed by Summer and Winter. In spring the demand is least.
- 2019 saw a good jump in shared bike demand compared to 2018.
- Demand for bikes increases gradually as the month passes and then declines at the end of year. Bookings are quite high from the month of ‘may’ to ‘Oct’ with ‘Sep’ seeing the highest demand.
- Demand for bikes is highest on Saturdays dips on Sunday and from mon to thu demand remains more or less consistent.
- Bookings are more on Working days compared to holidays.
- During ‘Clear’ weather situations demand for bikes is high followed by Misty+Cloudy weather followed by Light snow weather, which is quite obvious people prefer bike rides in favorable weather.

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

Answer: Setting **drop_first = True**, will drop the first category. This helps in avoiding multicollinearity. Example: We have 3 options Education say High School, Diploma, Graduation When we create dummy variables for the same

High School	Diploma	Graduation
1	0	0
0	1	0
0	0	1

If we drop the first variable, say High School then we are left with 2 variables Diploma and Graduation, then with value as 0 0 we still be able to determine that the variable is High School. Ideally we don’t need the first variable.

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

Answer: Variable ‘temp’ and ‘atemp’ have the highest correlation with ‘cnt’.

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

Answer: I have validated the assumptions of Linear Regression Model based on the training set using below factors:

Residual analysis:

- On doing residual analysis we see that histogram is following normal distribution so our model has **‘Normality’**
- From scatter plot we see that the residuals shows constant variance and hence follow **‘homoscedasticity’**
- This also marks that our model is **‘linear’.**

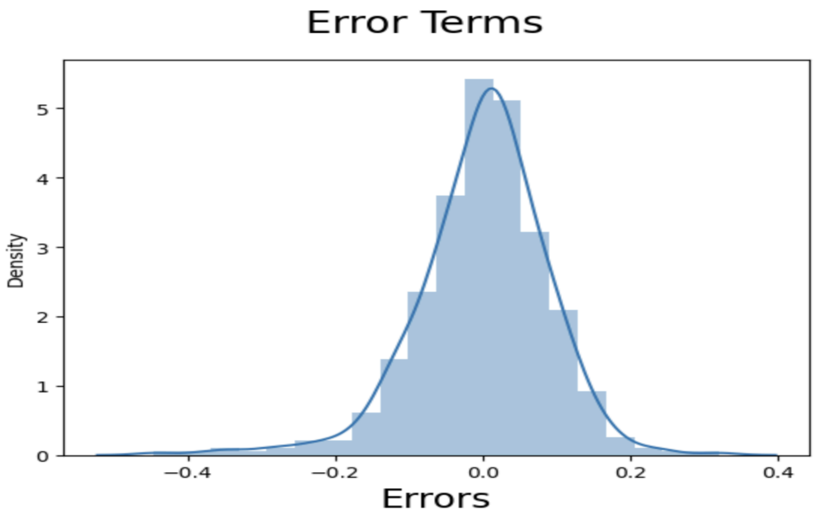
P-value analysis: For our final model none of the variables have p-value more than 0.05.

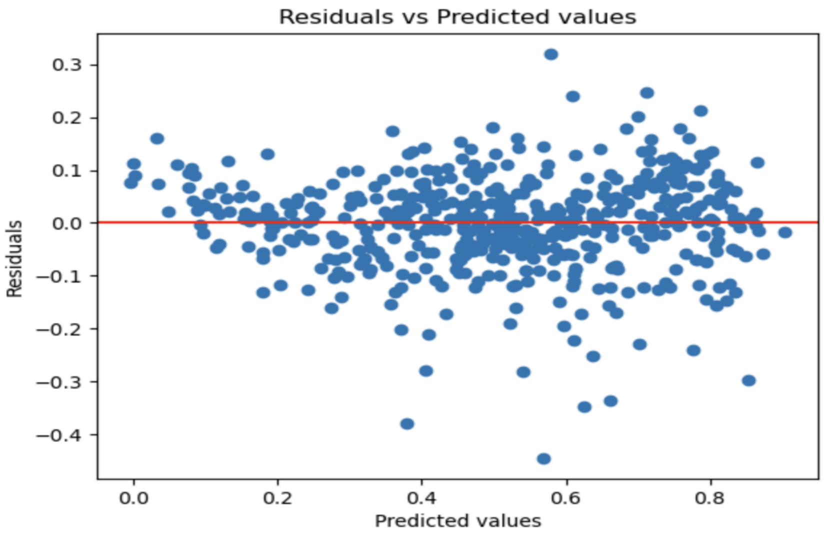
R^2 value: For the final model **R-squared:0.843** this means the model is good.

Multicollinearity: There is insignificant multicollinearity among variables based on the VIF calculated.

Testing the model on test set and verifying the R^2 value, adjusted R^2 value and

```
Text(0.5, 0, 'Errors')
```





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 : Temperature
- yr: Year
- sep: Month September

General Subjective Questions

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

Answer: Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable and one or more independent variables.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable.

Types of Linear Regression

Linear Regression can be broadly classified into two types of algorithms:

1. Simple Linear Regression

A simple straight-line equation involving slope (dy/dx) and intercept (an integer/continuous value) is utilized in simple Linear Regression. Here a simple form is:

Equation: $Y = \beta_0 + \beta_1 X + \varepsilon$

Where:

- Y is the dependent variable
- X is the independent variable
- β_0 is the y-intercept (value of Y when $X = 0$)
- β_1 is the slope (change in Y for a unit change in X)
- ε is the error term

With this equation, the algorithm trains the model of machine learning and gives the most accurate output.

Assumptions:

- **Linearity:** The relationship between X and Y is linear
- **Independence:** Observations should be independent of each other
- **Homoscedasticity**: Variance is constant
- **Normality:** Residuals are normally distributed

2. Multiple Linear Regression

When a number of independent variables more than one, the governing linear equation applicable to regression takes a different form like:

Equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{□} X_{□} + \varepsilon$

Where:

- Y is the dependent variable
- X_1, X_2, ..., X□ are k independent variables
- β_0, β_1, β_2, ..., β□ are the coefficients
- ε is the error term

Assumptions: Same as simple linear regression, plus:

- No multicollinearity: Independent variables should not be highly correlated with each other.

Estimation: OLS is commonly used.

Feature selection: Techniques to choose the most relevant predictors:

- Forward selection : Selecting one variable after another, starting point being 1 or 2 variables.
- Backward elimination: Putting all the variables at once and then removing based on p-value and VIF
- RFE

Evaluation:

- R^2 or R squared: R2 is a number which explains what portion of given model is explained by the developed model. So higher the R-squared value the better model fits our data.
- Adjusted R-squared: Accounts for the number of predictors in the model
- F-statistic: Tests the overall significance of the model
- Variance Inflation Factor (VIF): Detects multicollinearity

Industry relevance:

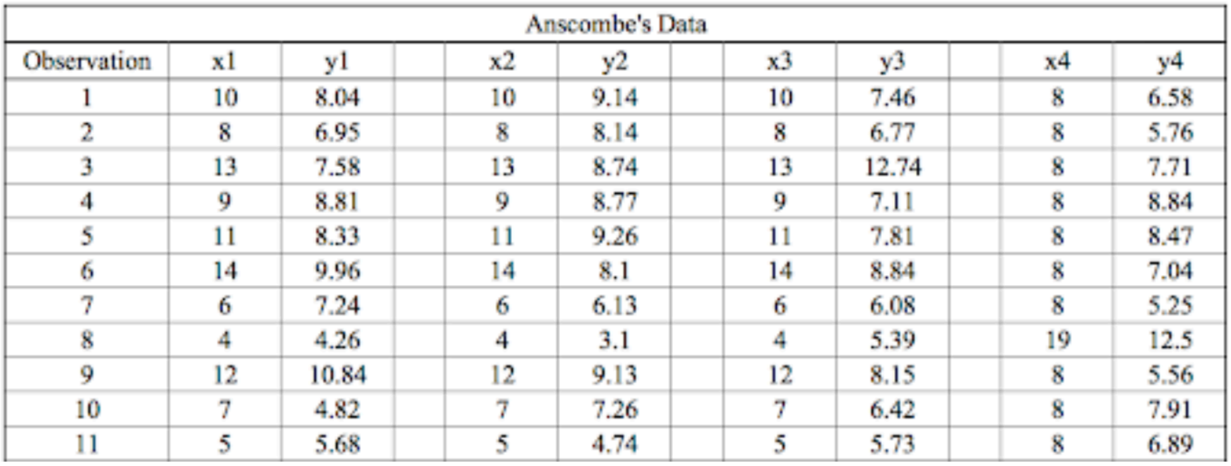
1. Predicting the sales of a company considering various factors
2. Predicting the salary of a candidate keeping in mind the years of experience, degree, past CTC, Education etc.

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

Answer: Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

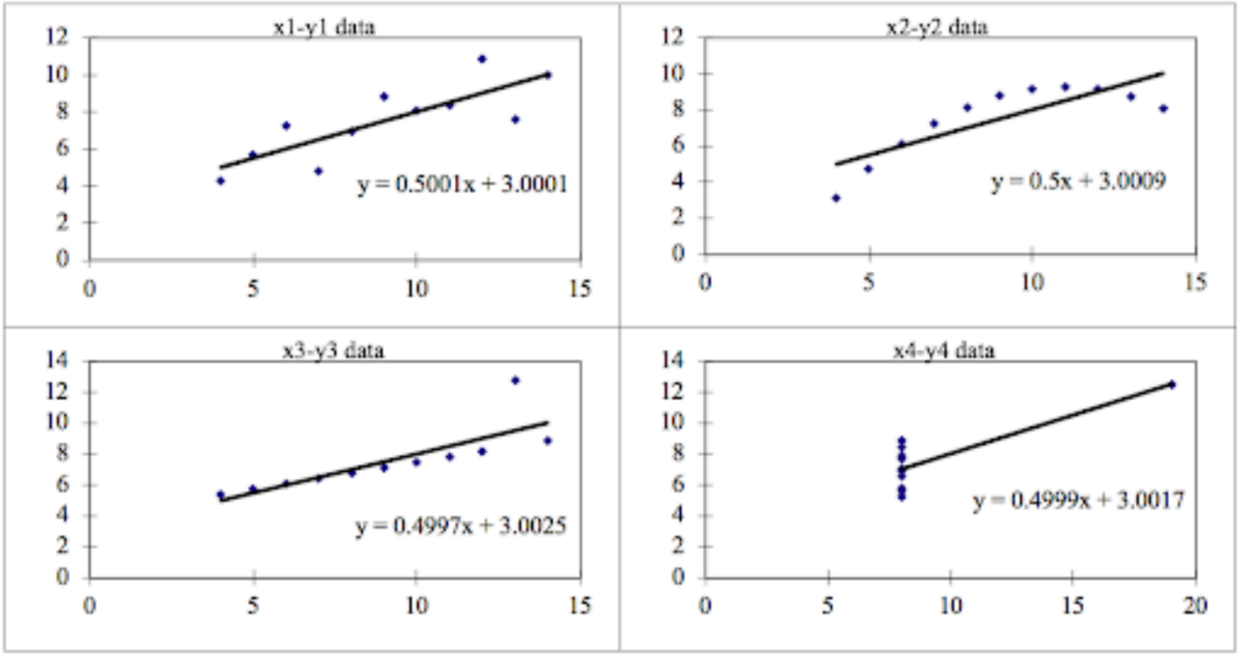
We can define these four plots as follows:

The statistical information for these four data sets is approximately similar. We can compute them as follows:



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

ANSCOMBE’S QUARTET FOUR DATASETS

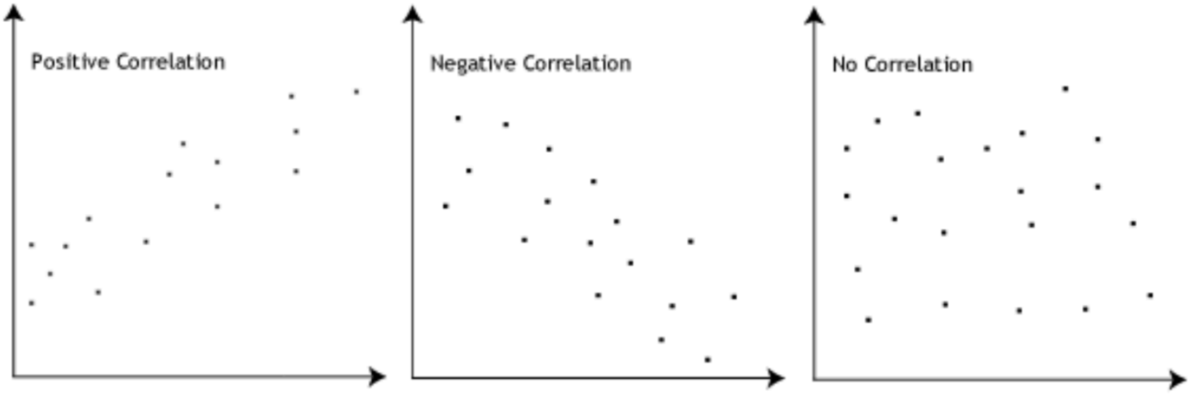
- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

  As we can see, Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set to help build a well-fit model.

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**:** Suppose we want to do Linear Regression for a set of data and we come across two variables a, b where a has values in meters e: 150 m and b has values in centimeters eg: 600 cm. If we try to develop a linear regression model without scaling so in that case variable b will be weighed more which is not correct

There are two types of scaling which we perform:

a. Normalized scaling (Min-max scaling)
b. Standardized scaling (Z-score normalization)

Sno.	Normalized scaling	Standardized scaling
1.	In normalized scaling rescaling of features happens to values between 0 and 1. X_norm = (X - X_min) / (X_max - X_min)	Rescales data to have a mean of 0 and a standard deviation of 1. Formula: X_stand = (X - μ) / σ (where μ is the mean and σ is the standard deviation)
2.	Doesn't center the data around zero	Centers the data around zero

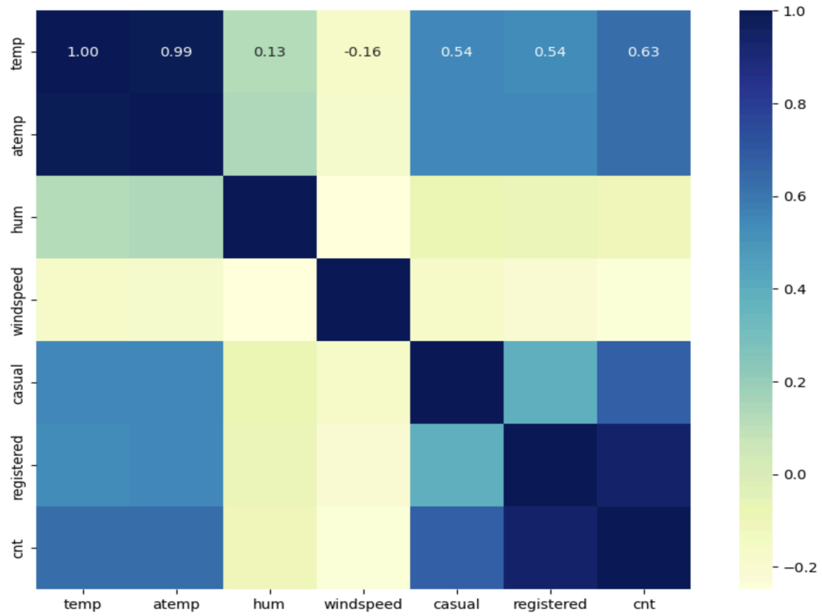
3.	Bounds values to a specific range eg: Fixed range (typically 0 to 1).	Doesn't bound values to a specific range
4.	It's usually affected by outliers	Less affected by outliers.
5	Preserves zero values	May turn zero values into non-zero values

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

Answer: VIF = 1 / (1 - R^2) Where R^2 is the coefficient of determination of a regression of that predictor on all other predictors. An infinite VIF occurs when $R^2 = 1$ for a predictor. This means that the predictor variable can be perfectly explained by a linear combination of the other predictor variables.

This situation, also called Perfect Multicollinearity, happens when there's an exact linear relationship between two or more predictors.

For eg: While working on a boombike assignment there was a situation when I got Inf VIF while doing through the manual model, reason could be understood from below heatmap. Variables 'atmep' and 'temp' have perfect multicollinearity.



Solution: We need to identify and drop the multicollinear variables.

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 (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. In the context of linear regression, Q-Q plots are particularly useful for checking the normality assumption of residuals.

Basic Concept: A Q-Q plot compares the quantiles of the observed data against the quantiles of a theoretical distribution.

Construction:

- Sort the observed data in ascending order.
- Calculate the expected quantiles from the theoretical distribution.
- Plot observed quantiles against theoretical quantiles.

Interpretation:

- If the points fall approximately along a straight line, the data follows the theoretical distribution.
- Deviations from the straight line indicate departures from the assumed distribution.

Use in Linear Regression: In linear regression, Q-Q plots are primarily used to check the normality of residuals, which is one of the key assumptions of linear regression.

Importance in Linear Regression:

- Helps verify the normality assumption of residuals.
- Important for the validity of t-tests and F-tests in regression analysis.
- Helps in determining outliers, points far from the line at the tails may indicate outliers.
- Helps in assessing skewness, an S-shaped curve suggests skewness in the data.
- Helps in assessing the overall fit of the model.
- Can suggest appropriate data transformations if normality is violated.