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)

Below are the categorical variable are there in dataset
season, mnth, weathersit, holiday, weekday, workingday, yr

- season:- Based on data visualization summer, fall and winter season having high consumption ration while spring has low consumption ratio.
- mnth:- Bike renting ratio is increasing month by month till September and then later decrease till jan.
- weathersit:- there is not enough data or very less bike renting ratio for Heavy Rain and Fog conditions.
  Clean/few clouds days are most favourable condition for bike renting.
- holiday:- user prefer to do bike rending on working days compare to holidays.
- weekday:- bike usage is higher in weekdays , Saturday contribute most in determining renting count.
- workingday:- the effect is nullify for regular and registered user for working days because both are contradictory.
- Demands of Bike rent is more in 2019 in compare to 2018, signifies that Boom or need for this service.
- In sept month demands is highest. After that demand is decreasing till jan.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
The drop_first=True is used while creating dummy variables to drop the base/reference category. The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included. The reference category easily be calculated where all the category is 0.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
By looking at the pair-plot among the numerical variables, temp(0.64) and atemp(0.65) have the highest correlation with the target variable (cnt).

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

- There is linear relationship between independent and dependent variables. The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot.
- Actual and predicted results follow pattern. prediction of test set is really close with actual target value.
- Error terms are normally distributed, histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0.
- Error terms are independent of each others.
- Error terms have constant variance. We can see Error Terms have approximately a Constant Variance, hence it follows the Assumption of Homoscedasticity.
- R2 value for predictions on test data (0.802) is almost same as R2 value of train data(0.837). This is a good R-squared value, hence we can see our model is performing good even on test data or unseen data.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

The top 3 variables are:

1. temp: Temperature is the Most Significant Feature which affects the Business positively Demand of Bike rent signify positive corelation.
2. Yr: Demand of Bike rent is more in 2019 in compare to 2018, signifies that Boom or need for this service there is positive corelation.
3. weathersit: Whereas the other Environmental condition such as Raining, Humidity, Windspeed and Cloudy affects the Business negatively, user do not prefer to rent a bike, its signify negative corelation.

General Subjective Questions

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

Linear regression is a supervised machine-learning algorithm and maps the data points to the linear functions which can be used for prediction on new datasets. 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 two main types of linear regression:

1. Simple Linear Regression

   This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

   y= β0 + β1 X

   where:

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

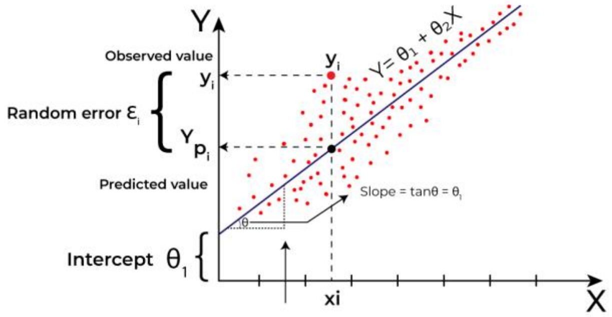
2. Multiple Linear Regression

   This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

   y= β0+ β1X1+ β2X2+......... βnX p

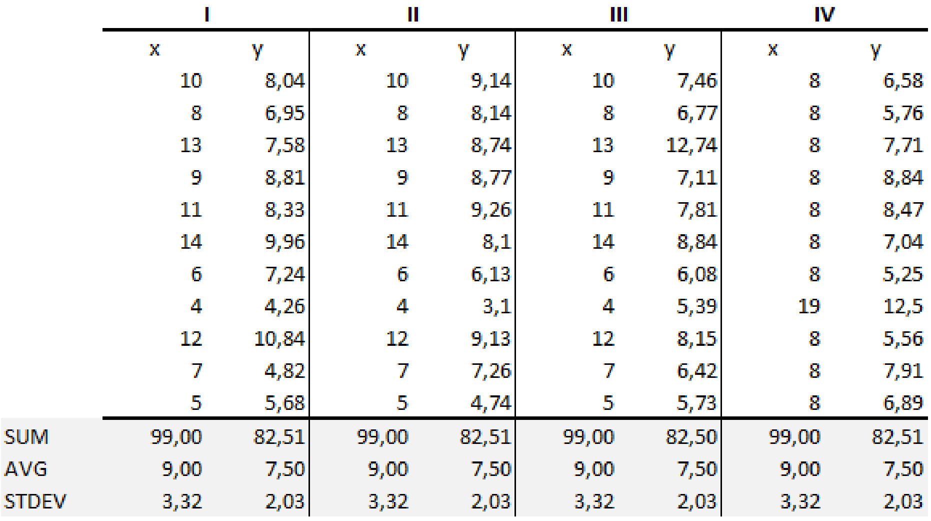
   where:

   Y is the dependent variable
   X1, X2, …, Xp are the independent variables
   β0 is the intercept
   β1, β2, …, βn are the slopes



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

Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. This is a method which keeps four datasets, each containing eleven (x, y) pairs.

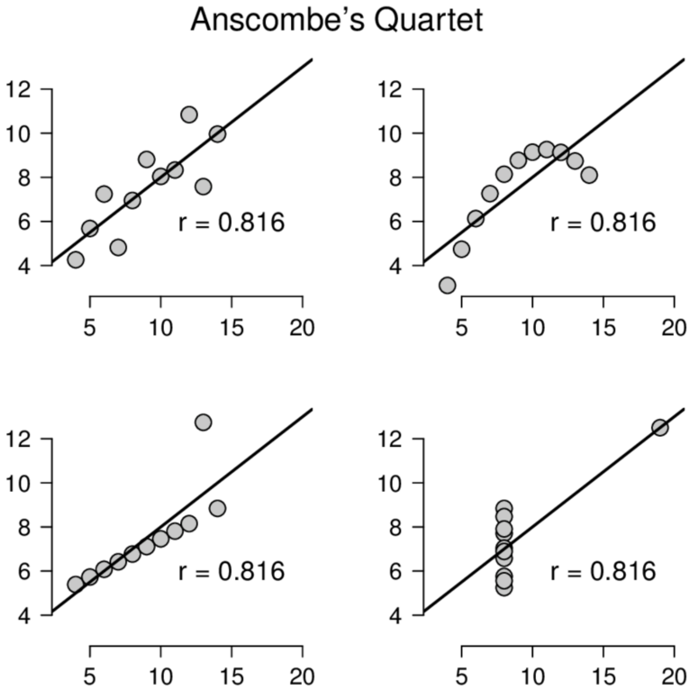


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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

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:



Anscombe's Quartet's descriptive statistics might seem homogeneous, but the accompanying visualizations demonstrate clear trends, demonstrating the need of combining statistical analysis with graphical exploration for reliable data interpretation. Anscombe's Quartet warns of the dangers of outliers in data sets.

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

Pearson's R was developed by Karl Pearson around the turn of the 20th century. Pearson correlation coefficient, is a statistical calculation of the strength of two variables' relationships. The formula for Pearson's correlation.

$$R= n(\sum xy) – (\sum x)(\sum y) / \surd[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2$$

The full name for Pearson's correlation coefficient formula is Pearson's Product Moment correlation (PPMC). It helps in displaying the Linear relationship between the two sets of the data.

The correlation coefficient formula returns a value between 1 and -1.

- -1 indicates a strong negative relationship
- 1 indicates strong positive relationships
- And a result of zero indicates no relationship at all

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

Scaling is data Pre-Processing step which is applied to independent variables to normalize the data within a particular range, also it helps in speeding up the calculations in an algorithm. Many times, in the dataset we see that multiple variables are in different ranges. Data set may contains features highly varying in magnitudes, units and range. In this case scaling becomes necessary otherwise algorithm only takes magnitude in account and not units hence this leads to incorrect modelling. So, scaling is required to bring them all in a single range. The two most discussed scaling methods are Normalization and Standardization.

1. Normalization typically scales the values into a range of [0,1].

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

$$x = \frac{x - \text{mean}(x)}{sd(x)}$$

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

An infinite Variance Inflation Factor (VIF) occurs when one variable can be perfectly predicted by other variables in the model. Imagine a situation where a variable X_1 is entirely redundant because its information is already captured by other predictors X_2 and X_3. When the coefficient of determination (R^2) approaches 1 for the relationship between X_1 and the other predictors, its VIF becomes infinite.

1. Perfect Multicollinearity:

- The occurrence of an infinite VIF is closely tied to perfect multicollinearity among predictor variables in a regression model.
- Imagine a scenario where one independent variable X_1 can be precisely predicted by other variables say X_2 and X_3.
- In this case, the VIF for X_1 becomes undefined because it provides no additional information beyond what's already captured by X_2 and X_3.

2. Interpretation:

- VIF values convey information about correlation between predictors.
- When R^2 (coefficient of determination) approaches 1 for the relationship between X_1 and other predictors, its VIF becomes infinite.
- Essentially, X_1 becomes redundant, as its contribution is already accounted for by the other variables.

3. Dataset Size Matters:

- The issue of infinite VIF can also arise due to the size of your dataset.
- If you have more variables than observations, all regressions may end up with $R^2 = 1$, leading to infinite VIF.
- Techniques like forward stepwise regression can help select a smaller subset of relevant predictors.

In summary, an infinite VIF indicates perfect multicollinearity, where one variable is akin to a silent note in a symphony of predictors.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)
Q-Q Plot (Quantile-Quantile Plot) is a graphical tool used to assess whether a dataset follows a specific theoretical distribution (such as Normal, exponential, or Uniform). It helps us compare the quantiles of our data to the quantiles of a standard normal distribution (a theoretical distribution with mean 0 and standard deviation 1). The plot provides insights into the distributional aspects of our data. Q-Q plots help us assess normality, identify outliers, and validate our regression model assumptions.

Use and Importance in Linear Regression:
1. Normality of Residuals:

- In linear regression, one of the assumptions is that the residuals (errors) follow a normal distribution.
- A Q-Q plot helps us evaluate this assumption by comparing the quantiles of the residuals to those of a standard normal distribution.
- If the points on the plot fall approximately along a 45-degree line, it suggests that the residuals are normally distributed.

2. Detection of Distributional Deviations:

- Q-Q plots reveal deviations from normality:
- If the points deviate significantly from the 45-degree line, it indicates non-normality of residuals.
- Outliers or skewness become evident as departures from the expected straight line.
- Shifts in location, scale, symmetry, and tail behavior can be detected.

3. Validation of Model Assumptions:

- Ensuring that residuals are normally distributed is crucial for reliable statistical inference.
- If the Q-Q plot shows deviations, we might need to consider transformations or address outliers.
- A well-behaved Q-Q plot validates the assumptions underlying linear regression.