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)

➔ To analyse the categorical variables from the dataset, bivariate (continuous-categorical) analysis is done using boxplots from which following are the inferences made:

1. season vs cnt:

- From the plot there is only one outlier after higher and lower fence during the season, 1(spring) and 4(winter) respectively.

- Number of bikes rented is more during the season 2(summer)and 3(fall). Also their median values falls around 5000.

2. mnth vs cnt:

- Bikes rented is less in the month of 1, 2, 5, 6, 7, 11 and 12.

3. weekday vs cnt:

- In this plot the median values are almost around the same value ie., 4300 for all days.

- Only on days 2 and 5 the IQR is minimum and for 3 and 6 it is maximum.

4. weathersit vs cnt:

- It looks like no one has rented bikes when weaathersit = 4 (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog).

- Most of the bikes are rented when weather was clear or mist or with few clouds indicated by the weathersit = 1 and weathersit = 2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist). Though their median values are slightly different.

- Renting bikes during rain or when the weather situation is little bad some customers have avoided it and hence the number is less as indicated by the weathersit = 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds).

➔ Summarizing the above inferences it can be said that the number of bikes rented is less during spring(1) and winter(4) as the weather condition will not be good for riding in those seasons thereby affecting the target variable 'cnt'.

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

- ➔ Given the categorical variables with 'n' levels, the dummy variable creation is to build 'n-1' variables indicating the levels.
- ➔ For a variable 'X', with three levels, say, 'a', 'b', 'c', the dummy table will be as follows:

	a	**b**	**c**
a	1	0	0
b	0	1	0
c	0	0	1

After using drop_first = True ➔

	b	**c**
a	0	0
b	1	0
c	0	1

- ➔ As it can be seen that defining 3 levels is not required (i.e., one of the column will be a redundant), so dropping a variable say 'a' using drop_first =True, it is possible to explain 3 levels as follows:
  - o Value 00 is 'a'
  - o Value 10 is 'b'
  - o Value 01 is 'c'

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

- ➔ The highest correlation is between **'registered'** and the target variable **'cnt'**.

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

- ➔ **Assumption of Linear Regression Model:**
  1. Error terms are normally distributed: First find the difference between the y_train and y_predicted and then plot histogram on the obtained result
  2. Linear Relationship between X and y.
  3. Error terms are independent to each other.
  4. Error terms have constant variance i.e., Homoscedasticity.

     The remaining 3 assumptions were validated by plotting a scatter plots.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

➔ Following are the top 3 features and corresponding absolute coefficients that contributing significantly towards the demand of the shared bikes:

   1. atemp is 3789.69
   2. Light_Snow is 2321.24
   3. yr is 2048.41

General Subjective Questions

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

➔ It is a Machine Learning algorithm based on Supervised Learning where the output variable to be predicted is continuous variable. For example, marks scored by the student in a subject.

➔ It focuses on the relationship between the dependent and the independent variables.

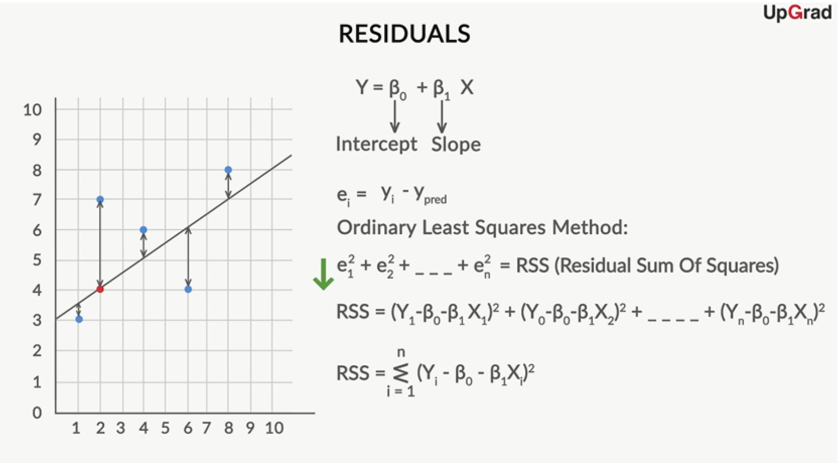
➔ **Types:**

   1. ***Simple Linear Regression (SLR):*** Changing only one variable at a time.
      General Equation: $Y = \beta_0 + \beta_1.X$ where β_0 is intercept and β_1 is slope.
   2. ***Multiple Linear Regression (MLR):*** Changing multiple variables at a time.
      General Equation: $Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 +\ldots. + \beta_n.X_n$.

➔ **Assumptions on Linear Regression:**

   1. There is a Linear Relationship between X and y.
   2. Error terms are normally distributed
   3. Error terms are independent to each other.
   4. Error terms have constant variance (Homoscedasticity).
   5. Multicollinearity (in case of MLR) – Determines whether an independent variable is dependent on a combinations of other independent variables.

➔ The best fit line is found by minimising the expression of **Residual Sum of Squares(RSS)** which is equal to the sum of squares of the residual for each data point on the plot. Residuals for any data point is the difference between the predicted value and actual value of the dependent variable.



RESIDUALS

$Y = \beta_0 + \beta_1 X$

Intercept Slope

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\,\_\,\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$RSS = (Y_1-\beta_0-\beta_1X_1)^2 + (Y_0-\beta_0-\beta_1X_2)^2 + \_\,\_\,\_\,\_ + (Y_n-\beta_0-\beta_1X_n)^2$

$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1X_i)^2$

- ➔ **Strength of Linear Regression:**

1. ***R-square:***

Formula: R2 = 1 – (RSS/TSS)

Where, RSS is residual sum of squares and TSS is Total Sum of Squares which are given by

$$RSS = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$$

$$\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Ranges between 0 indicating bad fit and 1 indicating the good fit.

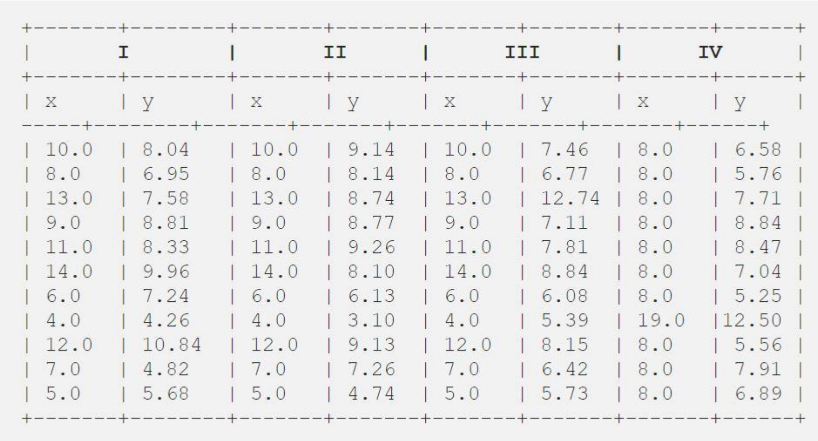
2. ***Residual Squared Error(RSE):***

Formula: RSE = square root(RSS/df)

Where df is degrees of freedom = n-1 (n is the number of data points.)

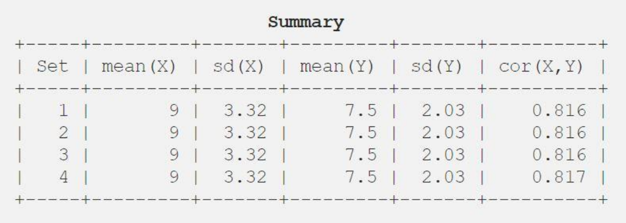
2. Explain the Anscombe’s quartet in detail. (3 marks)

- ➔ Anscombe’s quartet was constructed in 1973 by the statistician Francis Anscombe. It comprises of four data sets having nearly identical statiscal properties i.e., mean, standard deviation and correlation, but appear differently when plotted. Each dataset consists of 11 (x,y) poinits as shown below.



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       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 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

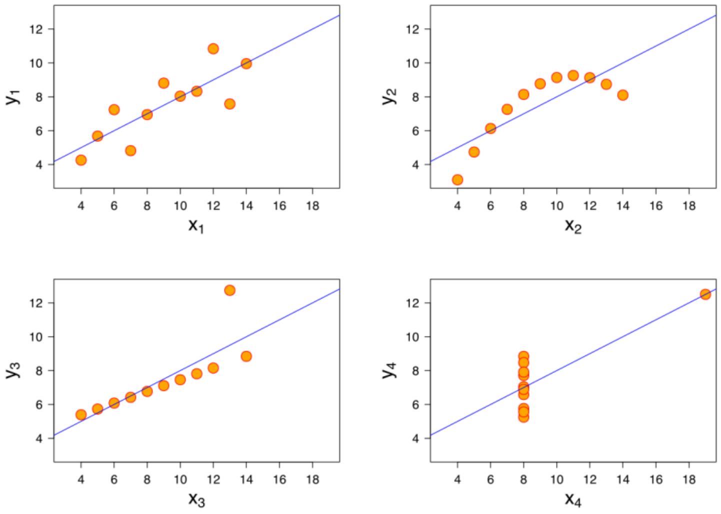
➔ The following is the corresponding summary statistics for all 4 data sets.



```
                         Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

➔ When the same data points are plotted, each plot tell a different story.

1. The first plot (x1, y1) shows that there seems to be a linear relationship between x and y.
2. The second plot (x2, y2) shows that there is non-linear relationship.
3. The third plot (x3,y3) has a linear relationship but the calculated regression is changed because of an outlier.
4. Finally, the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though other points do not show any relationship between the variables.

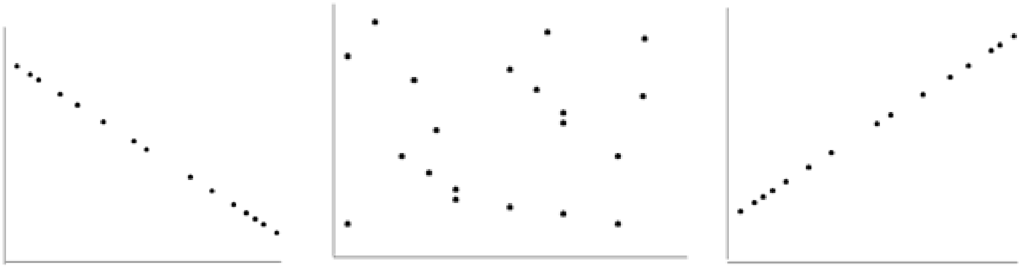


➔ Therefore, Anscombe quartet demonstrates the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson’s R? (3 marks):

➔ Pearson’s R or Pearson Correlation Coefficient is a measure of the strength of a linear association between 2 variables. The coefficient value ranges between +1 and -1.

➔ A value of +1 is total positive linear relationship, indicating both variables increase and decrease together. A value 0 is no linear correlation indicating no relationships exist. A value of -1 is total negative linear correlation, indicating that as one variable increases other variable decreases and vice versa. Following plots depicts the Pearson Correlation for all three values.



r = -1 **r = 0** **r = 1**

➔ **Formulae for finding Pearson’s Coefficient Correlation:**

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Where, r_{xy} → Pearson Correlation Coefficient

n → no. of observations

x_i → value of x

y_i → value of y

➔ **Degree of Correlation:**

- ***Perfect:*** Value is near to ±1 or ±1
- ***High Degree:*** Coefficient value lies between ±0.5 to ±1
- ***Moderate Degree:*** Coefficient value lies between ±0.30 to ±0.49
- ***Low Degree:*** Coefficient value is less than ±0.29
- ***No Correlation:*** Value is 0.

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

- ➔ Scaling means transformation of given data so that it can fit to a particular scale like., 0-100 or 0-1.
- ➔ Scaling is necessary for following reasons:
  1. Ease of interpretation
  2. Faster convergence for gradient descent methods.
- ➔ Scaling affects the coefficients and none of the other parameters such as t-statistic, F-statistic, p-value and R-squared.
- ➔ **Difference between normalized and standardized scaling:**

Normalized Scaling	**Standardized Scaling**
Brings all data in the range of 0-1.	Brings all the data into a standard normal distribution with mean = 0 and standard deviation(sd) = 1
Formulae: X = (x – min(x)) / (max(x) – min(x))	***Formulae:*** X = (x –mean(x))/sd(x)
It does not handle outliers.	It handles outliers and hence it is robust.

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

- ➔ Variance Inflation Factor (VIF) determines whether a independent variable is dependent on a single or a combination of independent variables i.e., multicollinearity among independent variables.
- ➔ Common condition followed for VIF values are
  1. VIF >10 is very high and the variable needs to be eliminated.
  2. VIF >5 is fine but needs to be inspected.
  3. VIF < 5 is good and no need to eliminate.
- ➔ VIF is given by 1/(1-R2), so if VIF reaches infinity only R2 reaches 1. This means that there is perfect correlation with each other which happens when the varaibles are orthogonal to each other.

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

➔ Q-Q plots or Quantile-Quantile plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quartile.

➔ Purpose: To find out if two sets of data come from the same distribution.

➔ In linear regression it is generally used to fit a model as it is simple. It checks if the points lie on the fitted lie if it does not fit then the errors are not normal.

➔ Steps to make a QQ plot:

   1. Given a set of values, sort the values in ascending order.
   2. Draw a normal distribution curve
   3. Find the Z-scores.
   4. Plot the data set values against the Z-scores values.