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)**

   We have the categorical variables ‘season’, ‘yr’, ‘mnth’, ‘holiday’, ‘weekday’, ‘workingday’ and weathersit’ in the given dataset. Below variables are having a major effect on the dependant variable ‘cnt’.

   'Season' - Bike rental counts are increasing in the season summer and the counts are high in the season fall. Then started to decrease in the season winter. Bike rental counts are least in the season spring.

   ‘yr’ – We have increased in the bike rental counts when increasing in the year.

   ‘mnth’ – Bike rental counts are started to increase in the month of March and the rental counts are at the peak in the month of September. The months January and February are having the least rental counts.

   ‘weathersit’ – When the weather is clear we are having a high number of bike rentals. When we have the weather light rain, we have the least number of rentals. Also, there is no record of rentals in the dataset for the weather heavy rain. This shows that people do not take rental bikes when there is heavy rain which is unfavourable for the user.

   ‘holiday’ – There is a drop in rental counts when there is a holiday.

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

   While creating the dummy variables for the categorical labels using pandas “get_dummies” method, we need to use drop_first = True to obtain the n-1 variables from the n category of the categorical column. This will reduce dependant variables count and we can represent the same categorical separation using n-1 columns.

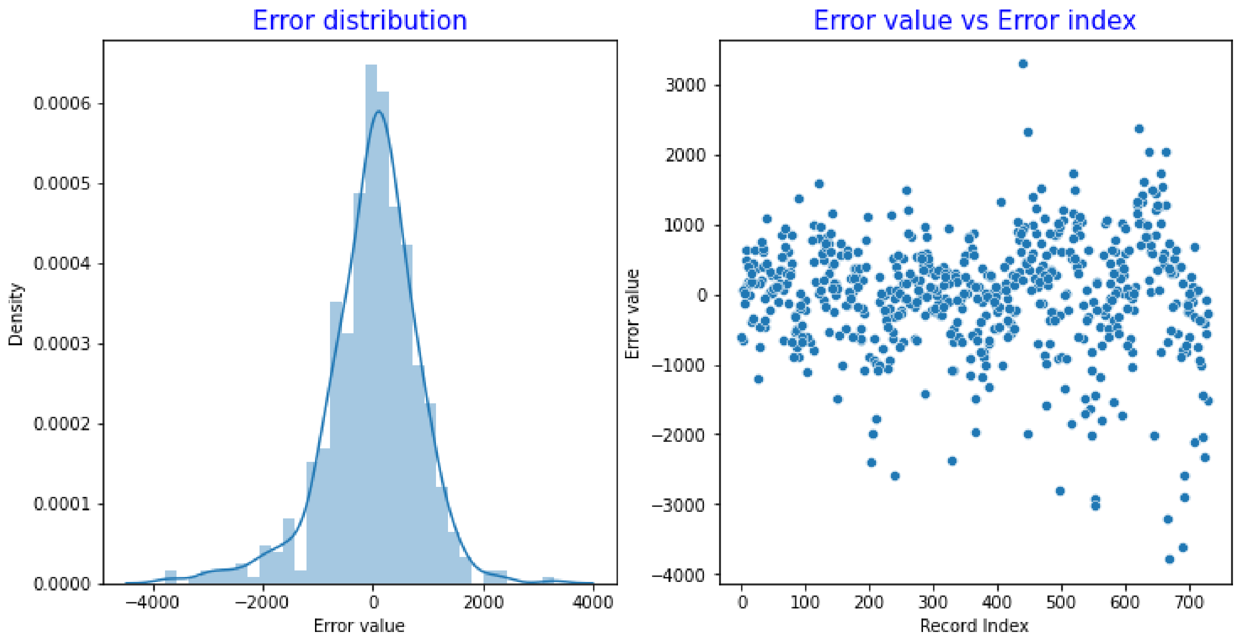
   If we have 4 categories (Season - 'spring', 'summer', 'fall' and 'winter') in a column, we can represent the same categories with 3 (4-1) dummy variables ('season_spring', ' season_summer' and ' season_winter').

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

   As per the pair-plot, the temperature (Column: “temp”) is having a high correlation with the target variable (Column: “cnt”).

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

   The assumptions of Linear Regression are the error terms are normally distributed, error terms are having the mean as zero and error terms are independent.



The above distribution plot is created using the errors/residuals (difference between the actual to predicted values) of the model. We could see that the distribution is normal and the mean value is nearly zero. Also in the second plot, we could see that there is no visible pattern in the error terms. We can say that our model satisfies the assumption of Linear Regression.

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 three features explaining the demand for shared bikes are,

i. 'temp' – coefficient value: 3809.949918
ii. 'weathersit_light_rain' - coefficient value: -2472.878584
iii. 'yr' - coefficient value: 2012.609278

General Subjective Questions

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

Linear regression algorithm is a supervised machine learning method to predict the continuous output/dependant variable as per the input/independent variables. Linear regression algorithm explains the linear relationship between the output/dependant variable and the input/independent variables. We can explain the relationship with the mathematical notation **y = mx + c**.

In linear regression, we need to find the best fit line to predict the values as per the input variables. To achieve that we need to reduce the error values between the actual values to predicted values.

We have two types are linear regression:

i. Simple linear regression (SLR)
ii. Multiple linear regression (MLR)

1. Simple linear regression (SLR):

In this method, the output/dependant variables are predicted using only one input/independent variable.

$Y = \beta0 + \beta1x$

$\beta0$ -> Y-intercept or constant term

$\beta1$ -> Coefficient of the dependant variable

2. Multiple linear regression (MLT):

More than one input/independent variables are used to predict the output/dependant variable.

$$Y = \beta 0 + \beta 1\ X1 + \beta 2\ X2 + \ldots. + \beta n\ Xn$$

$\beta 0$ –> Y intercept or constant term
$\beta 1$ –> Coefficient of X1 variable
$\beta 2$ –> Coefficient of X2 variable
βn –> Coefficient of Xn variable

For every linear regression method, we need to have the below assumptions.

i. There is a linear relationship between the y(output) and x(input) variables.
ii. Error terms are normally distributed with the mean value zero and they are independent.
iii. No multicollinearity between the input variables when we use multiple linear regression.

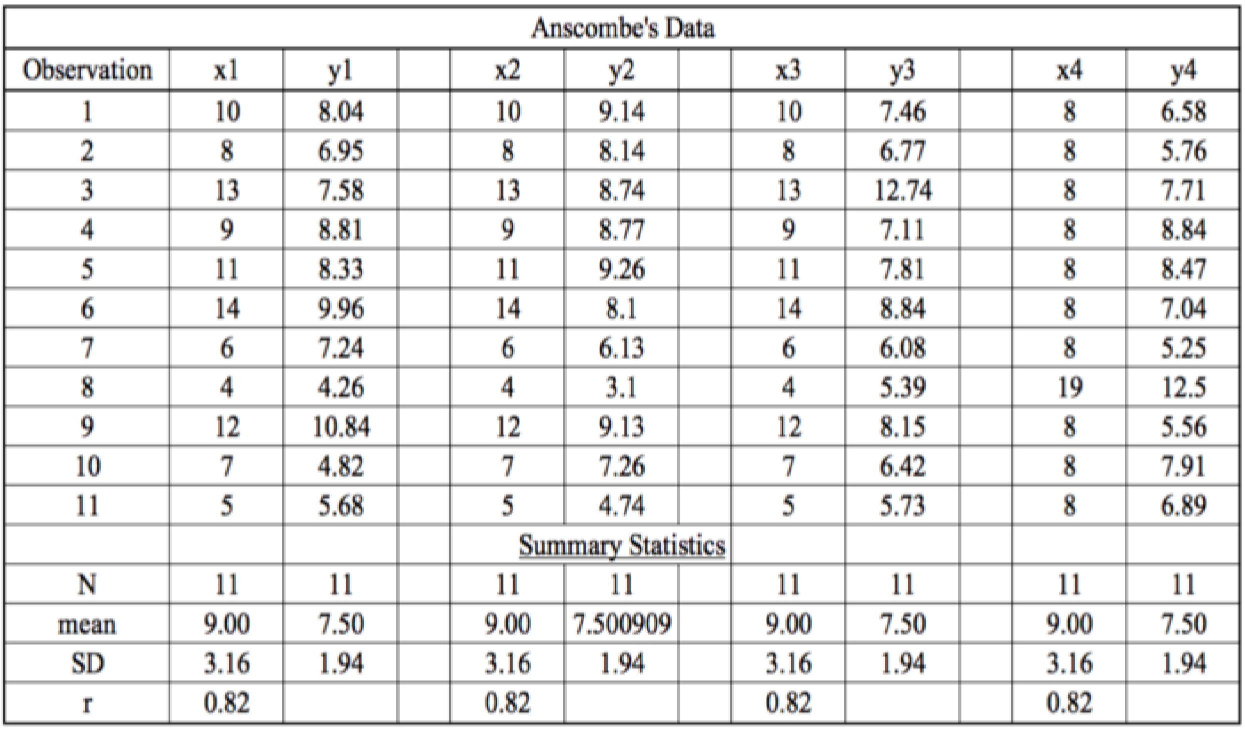
Positive relationship: When there is an increase in output variable for the increase of input variables are called positive linear relationship.

Negative relationship: When there is a decrease in the output variable for the increase of input variables are called a negative linear relationship.

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

Anscombe's quartet was developed by the statistician Francis Anscombe. Anscombe's quartet contains four datasets which are having identical simple descriptive statistics but, appear different while graphing.

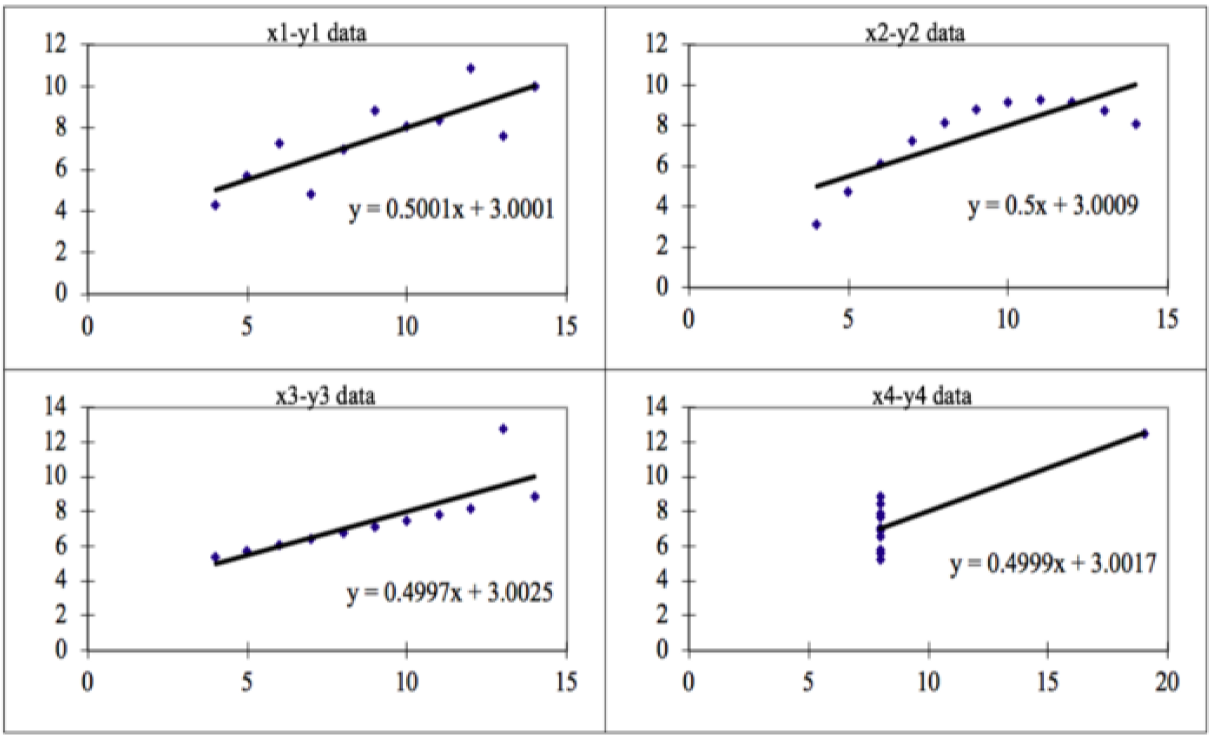
The below table has Anscombe's data which are having identical simple descriptive statistical values.



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
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Number of records - 11
Mean: input- 9.00 & output- 7.50
Std. Deviation: input - 3.16 & output – 1.94
Correlation (x and y): 0.82

Below are the plots using Anscombe's dataset and we can see that the linear regression lines are nearly the same for all four datasets.



Visualization will be useful to find the data distribution, presence of outliers/anomalies etc.

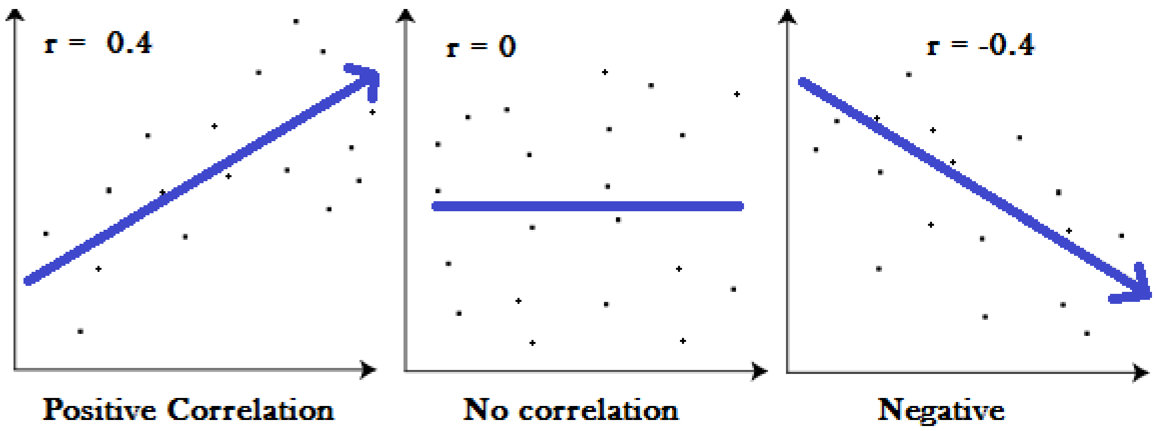
So, Anscombe's quartet illustrates the importance of visualizing the dataset before proceeding to analysing and building a machine learning model with the dataset. This will help to make a good fit machine learning model.

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

Pearson's R (r) is the numerical representation of the linear relationship between the two datasets (x and y). The Pearson's R formula will return the value between -1 to 1.

- -1 indicates a strong negative relationship between x and y.
- 1 indicates a strong positive relationship between x and y
- 0 indicates a no relationship between x and y

r = Covariance (x, y) / S.D (x) * S.D (y)



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

Feature scaling is the method to normalize or standardize the independent variables to a fixed range. This will be performed during the data pre-processing to handle the varying values. Scaling also helps in speeding up the calculation of the algorithm.

If feature scaling is not done, the machine learning algorithm tends to assign more weightage to the large numbers and low weightage to the smaller number irrespective of the unit representation.

Types of scaling and the difference between them:

i. Normalization:
- Minimum and maximum values are used for scaling.
- Scale values between 0 to 1 or -1 to 1.
- Scaling values are affected by the outliers.
- X = x - min(x) / max(x) - min(x)

ii. Standardization:
- Mean and standard deviation are used for scaling
- Boundaries are not pre-defined.
- Scaling values are not affected by the outliers.
- X = x-mean(x) / S.D(x)

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

VIF (Variance inflation factor) is the measure of the amount of multicollinearity between the predictor variables in the multiple regression model.

VIF value of each variable was calculated using the below mathematical formula.

$$\text{VIF} = 1 / (1 - R_i^2)$$

Here, (1 -R_i^2)→ Tolerance of the variable (This a measure of collinearity).

R^2 value of the independent variable is how well that particular variable is explained by the other variables. If the independent variable is explained perfectly by the other independent variables, then this will be the perfect correlation. So, the R^2 value becomes 1.

If we replace this R^2 value in the VIF formula, we will get the result as infinite. We should drop that particular variable from the dataset to build the best model.

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

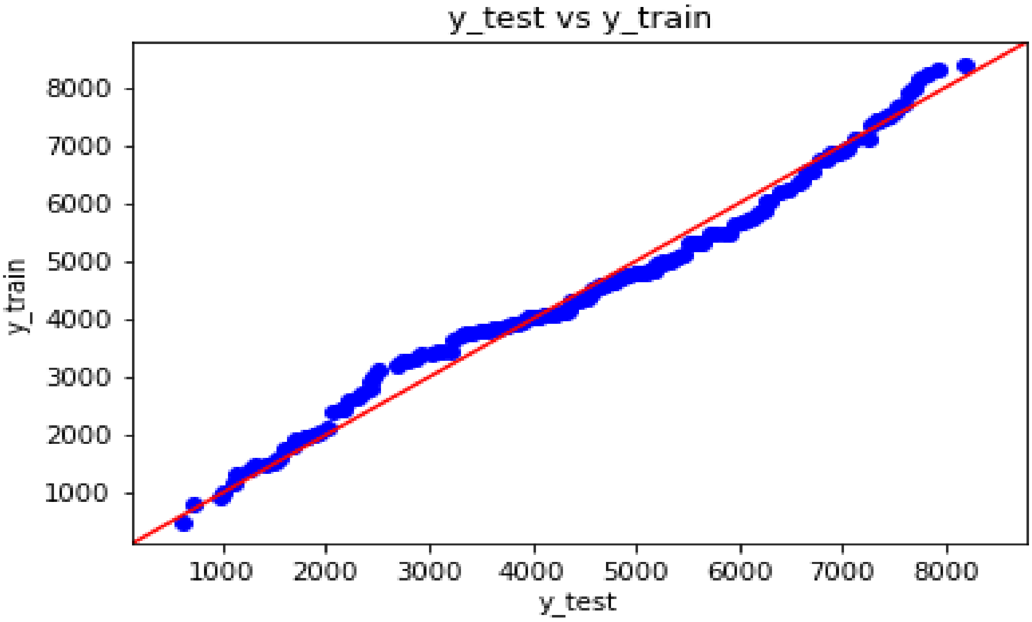
Quantile-Quantile (Q-Q) plot is a graphical representation to help to validate a dataset whether the data distribution is normal, uniform or exponential.

Also, this will help to find whether the two datasets come from a common distribution.

With the two sets, we can find the below details using the Q-Q plot.

- Whether they come from populations with the same distribution.
- Whether they have a common scale and location
- Whether they have a similar shape of the distribution.
- Whether they have similar tail behaviour.

Below is the Q-Q plot using the python stats model function “qqplot_2samples”. Here, we used the output variable ‘cnt’ of training and test data which are split from the input dataset.



We could see that all the points are near the 45-degree line. This datasets test and train are split from the same dataset and the above Q-Q plot explains the same.