Assignment-based Subjective Questions

Question 1- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer- Categorical variables which helps us in determining our target variable are –

1. Season – Summer and winter plays a crucial roles in understanding our target variables. From the final model, we can understand that more people prefer rental bikes during Winters.
2. Yr – Based on the model, more people chose rental vehicles during 2019.
3. Weathersit – Count of people chosing rental bikes decreases with Light Rain and Mist.

Question 2- Why is it important to use drop_first=True during dummy variable creation?
Answer – While creating dummy variable for a categorical column, if there are *n* unique values as part of the column, *n-1* dummy variables are sufficient to consider all n values.

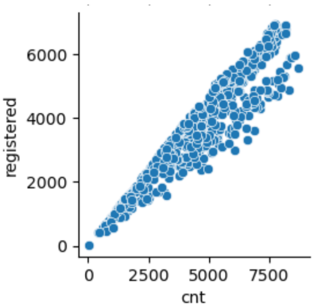
For Eg – In our assignment, we have categorical variable season with values Spring, Summer, Fall and Winter. You can see from the below table that using 3 i.e., (n-1) dummy variables, you can define all variables.

Variables	Spring	Summer	Fall
Spring	1	0	0
Summer	0	1	0
Fall	0	0	1
Winter	0	0	0

This is the reason why drop_first = True is used as you can infer from the below image, it will remove the first row creating (n-1) variables.

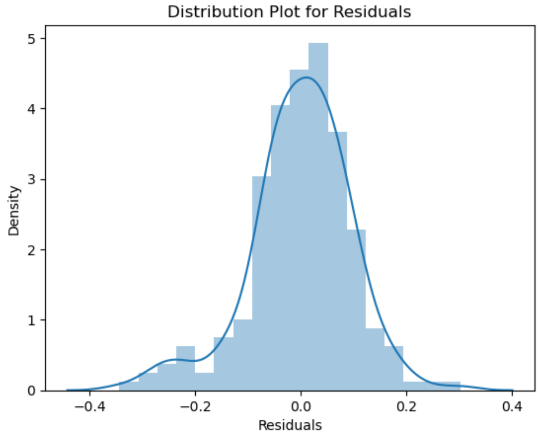
Variables	Spring	Summer	Fall	~~Winter~~
Spring	1	0	0	~~0~~
Summer	0	1	0	~~0~~
Fall	0	0	1	~~0~~
Winter	0	0	0	~~1~~

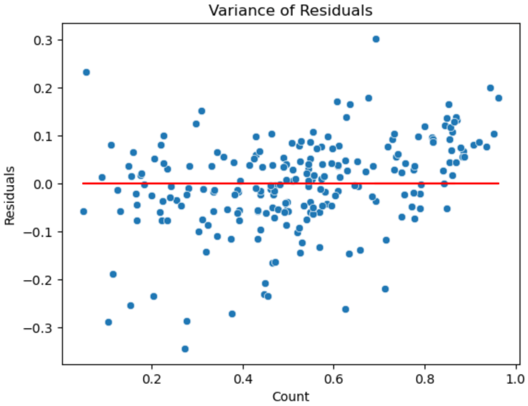
Question 3 - Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer – Column ***registered*** is highly correlated with target variable –



Question 4 - How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer – Assumptions considered for Linear Regression are –

1. There is a linear relationship between X and Y
2. Error terms are normally distributed with mean 0
3. Error terms are independent of each other
4. Error terms have constant variance





Question 5 - Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer – temperature, Light rain, yr are the 3 top features contributing significantly towards explaining the demand. The coefficients of these variables are as follows –

Temperature - 0.5946
Light Rain - -0.2400
yr - 0.2284

General Subjective Questions

Question 1 - Explain the linear regression algorithm in detail.
Answer – Linear Regression analysis is used to predict value of a variable based on another variables. It uses independent variables *(variables from which target variable needs to be predicted)* and analyses and comes up with solution which can be used to understand the behaviour of our target variables. Concept of Best-Fit line is used in linear regression where we assume that there is a line which is dependent on our independent variables and can help us in deducing our target variable.

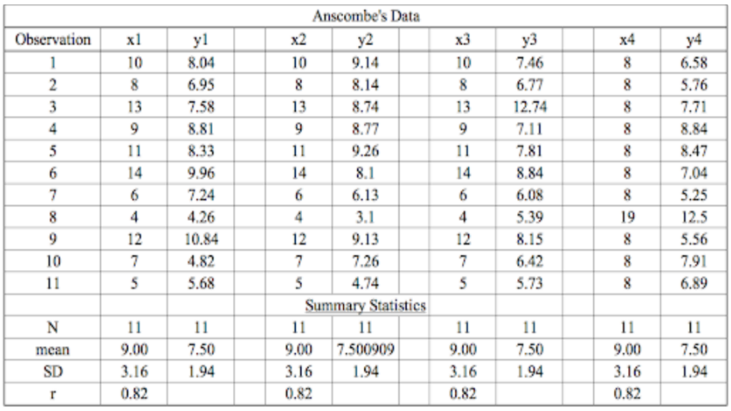
We assume that there is a straight line (or best–fit line) which considers the independent variables and use them to understand the behaviour of our target variable. With Linear regression model, we try to come up with this best-fit line that helps us predict and understand the behaviour of our target variable.

There are certain assumptions that needs to be considered for Linear regression which are –

1. There is a linear relationship between X and Y
2. Error terms are normally distributed with mean 0
3. Error terms are independent of each other
4. Error terms have constant variance

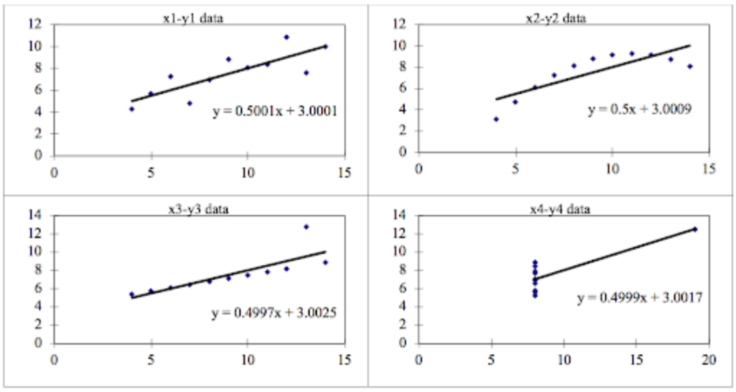
Question 2 - Explain the Anscombe's quartet in detail.
Answer – Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. It tells us about the importance of visualizing 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.).

Let us understand Anscombe's quartet with the help of an example. Here we have 4 data points with similar mean and standard deviation.



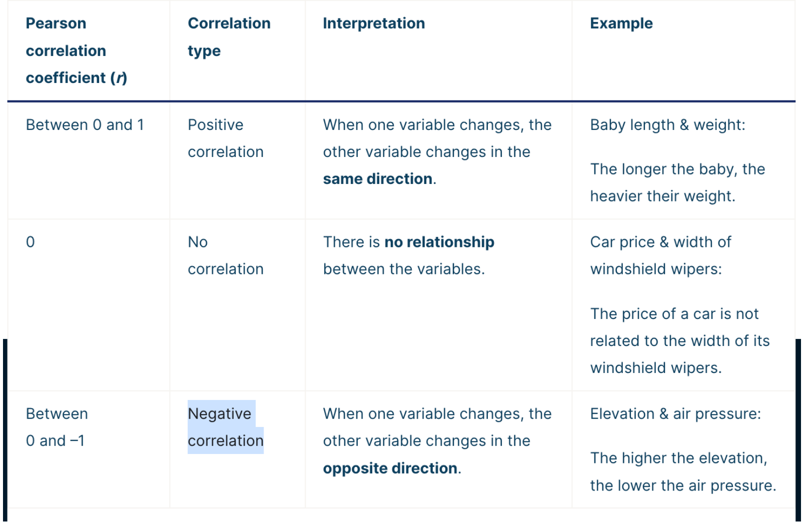
Anscombe's Data

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	



But from it's scatter plot, we can see that how different these 4 datapoints looks like when looked from the plot's perspective.

Question 3 - What is Pearson's R?
Answer – Pearson's R or **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.



Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

Question 4 - What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer – In machine learning, feature scaling refers to putting the feature values into the same range. A technique to scale data is to squeeze it into a predefined interval.

Normalization Scaling - In normalization, we map the minimum feature value to 0 and the maximum to 1. Hence, the feature values are mapped into the [0, 1] range:

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

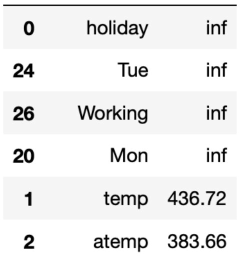
Standardization Scaling - In standardization, we don't enforce the data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1:

$$z = \frac{x - \mu}{\sigma}$$

It not only helps with scaling but also centralizes the data.

Question 5 - You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer – An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well). This occurs when you have variables which has very high correlation with each other.

For our model, infinite VIF can be seen for -



0	holiday	inf
24	Tue	inf
26	Working	inf
20	Mon	inf
1	temp	436.72
2	atemp	383.66

Question 6 - What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer - Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

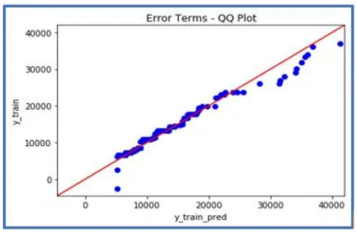
It is used to check if two datasets :

a. come from populations with a common distribution
b. have common location and scale
c. have similar distributional shapes
d. have similar tail behaviour

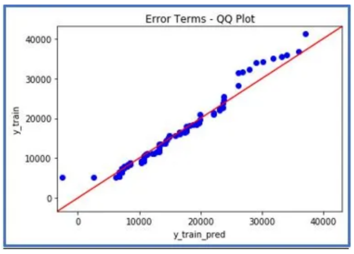
Interpretation of Q-Q plot -

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) ***Y-values < X-values:*** If y-quantiles are lower than the x-quantiles.



c) ***X-values < Y-values:*** If x-quantiles are lower than the y-quantiles.


Error Terms - QQ Plot
y_train
y_train_pred
0
10000
20000
30000
40000