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?**

➔ We can see that in our dataset we have following categorical variables.
(Year, month, holiday, weekday, working day, weather sit)

Year – Year has positive correlation with target variable which suggests that demand for rental bikes is increasing every incremental year.

Month – Demand of bikes is uniformly distributed across 12 months (where fall season sep-nov has seen maximum demand for bikes)

Holiday – A holiday can significantly impact the demand of bikes (negatively)

Weekday - Surprisingly Monday has seen lowest demands for bikes, otherwise increasing from Mon to Sun

Working day – has no such relation with target variable

Weather – Spring season has seen lowest demand for bikes. Fall season has highest demand for bikes.

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

➔ Dummy variables are created on categorical columns to get information regarding effect of each category on target variables.

Now let's say we have a category named Gender which has 2 values Male and Female, We are converting this column in dummy variables and not using drop_first+true.

Let's say Male will be assigned 1 and Female will be assigned 0, now in order to convey this information we created two columns.

But what if we would have used only one column which has 2 values 0 and 1 (using drop_first = True) then in that case we would need to create only one column to convey the same information.
Hence we use drop_first = True to feed less no of columns to model.

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

➔ temp and atemp columns

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

➔ *Find the residual*.
Residual is the diff between y_train and y_train_pred. Then we'll plot the distribution Plot for residuals. As per assumption residual plot should have normal distribution so if our graph will have normal dist then we can conclude that first assumption is satisfied

Graph of X_train vs res
If we don't see any pattern in the graph then we can conclude that second condition is satisfied.

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

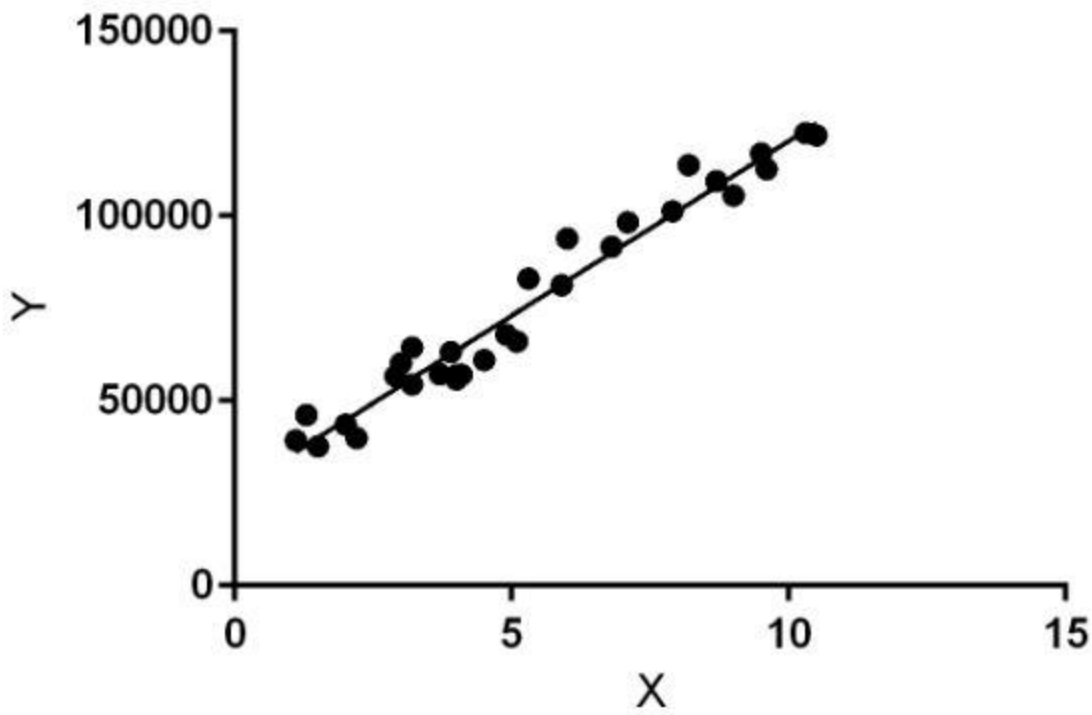
➔ *Year –has strong positive correlation*
➔ *Temp - has strong positive correlation*
➔ *Light Snow, Light Rain, Thunderstorm ,Scattered clouds, Light Rain, Scattered clouds- has strong negative correlation*

General Subjective Questions

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

➔ Linear regression is the statistics algorithm which is used to find relation between independent variable and dependent variable (something which we want to predict)

Broadly linear regression is classified into two categories:-
1) Simple linear regression
2) Multiple linear regression



Simple linear regression: - In simple linear regression we have only one independent variable and one dependent variable.

Multiple linear regression:-In multiple linear regression we have more than one independent variable and one dependent variable.
Model representation: -

Simple linear regression – y = B0 + B1*x

Multiple linear regression – y = B0 + B1*x + B2*x ……. + Bn*x

B0 = interceptor
B1 ….. Bn = coefficients of X

Example –

Imagine we want to predict weight (y) from given height(x). Our linear regression equation will look like below –

y = B0 + B1 * x1 or weight =B0 +B1 * height

Let's say we know the intercept B0 (0.1) and coefficient of height B1 = 0.5 then our equation will become –

y = 0.1 + 0.5 * x1

Now if you know the height of the person you can predict its weight so let's say if height is 160 then his/her weight should be 0.1 + 0.5*160 = 80.1

Assumptions of the linear regression

1) **Residual should have normal distribution**
2) **Graph of X_train vs residual should not show any pattern**

2. Explain the Anscombe's quartet in detail

➔ Anscombe's quartet is the group of four dataset which tells you the importance of visualization of the data and not to rely on summary statistics too much.

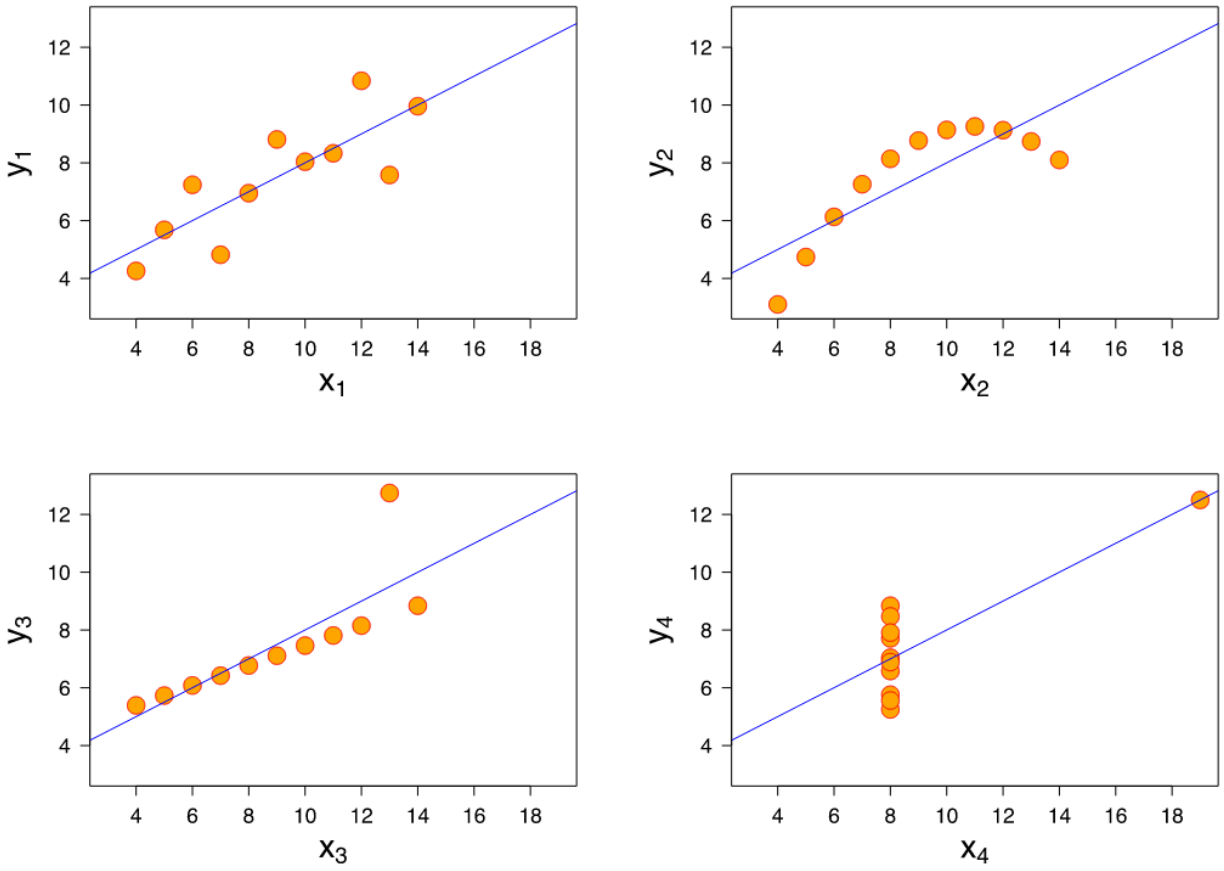
To prove this point we will look into these 4 datasets

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

All the summary statistics you'd think to compute are close to identical:

1. The average x value is 9 for each dataset
2. The average y value is 7.50 for each dataset
3. The variance for x is 11 and the variance for y is 4.12
4. The correlation between x and y is 0.816 for each dataset
5. A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So looking into above points we can say that all the datasets are same. Now let's visualize the data.



We can see that all the datasets follow different relationships with target variables. First and Third dataset have linear relationship while second has nonlinear relationship.

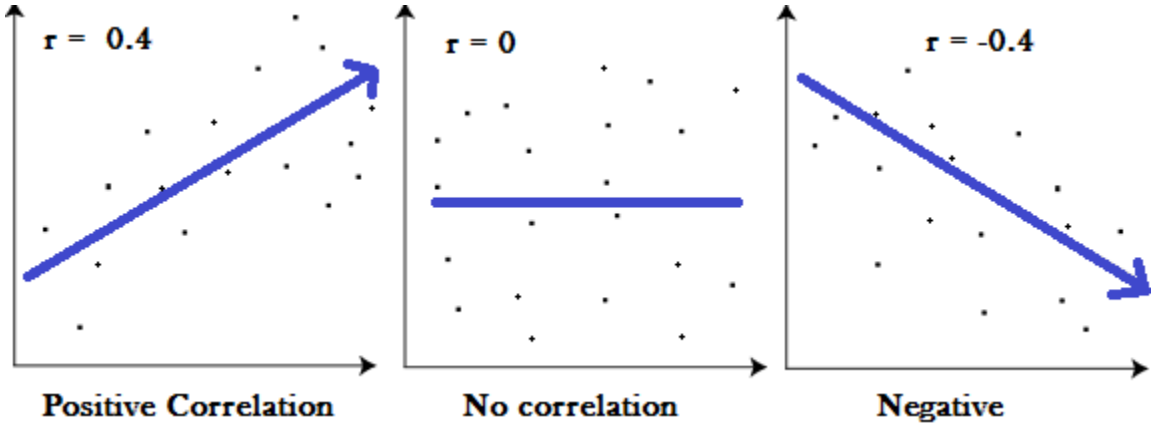
So these four graphs suggest that visualization of the data has utmost important while doing data analysis.

3. What is Pearson's R?

➔ Its correlation coefficient which tells the relationship between two variables numerically, this coefficient has value between -1 and 1.

1 means you've perfect positive correlation.

-1 means you've perfect negative correlation.



➔ A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

➔ A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

➔ Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

➔ Sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

➔ Population correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

➔ Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.

Suppose we have two features of weight and price, as in the below table. The "Weight" cannot have a meaningful comparison with the "Price." So the assumption algorithm makes that since "Weight" > "Price," thus "Weight," is more important than "Price."

So these more significant number starts playing a more decisive role while training the model. Thus feature scaling is needed to bring every feature in the same footing without any upfront importance.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

➔ Variance Inflation Factors (VIFs) provide a one-number summary description of collinearity for each model term

If feature has high collinearity means that variable is expressed by some other feature of combination of features.

Formula –

$$1/1 - r2$$

So if r2 is 1 the value of VIF will be infinity.
That means concerned feature is perfectly expressed by other feature or combination of feature.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

➔ Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

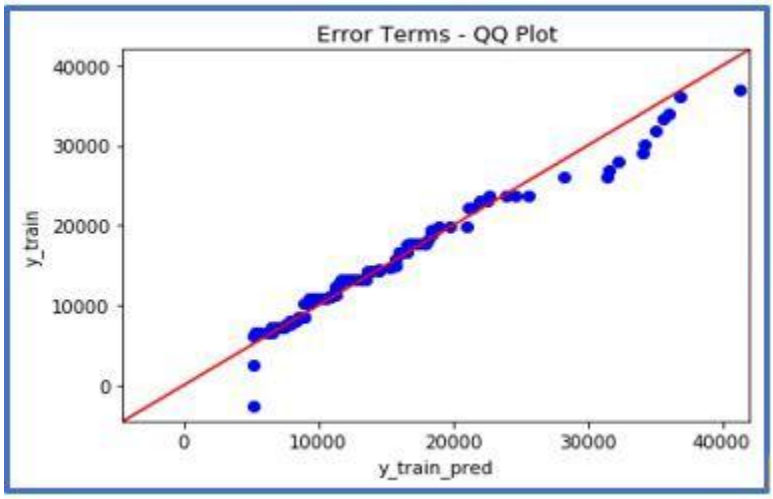
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Use and Importance –

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

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) 

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

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis