Assignment-based Subjective Questions

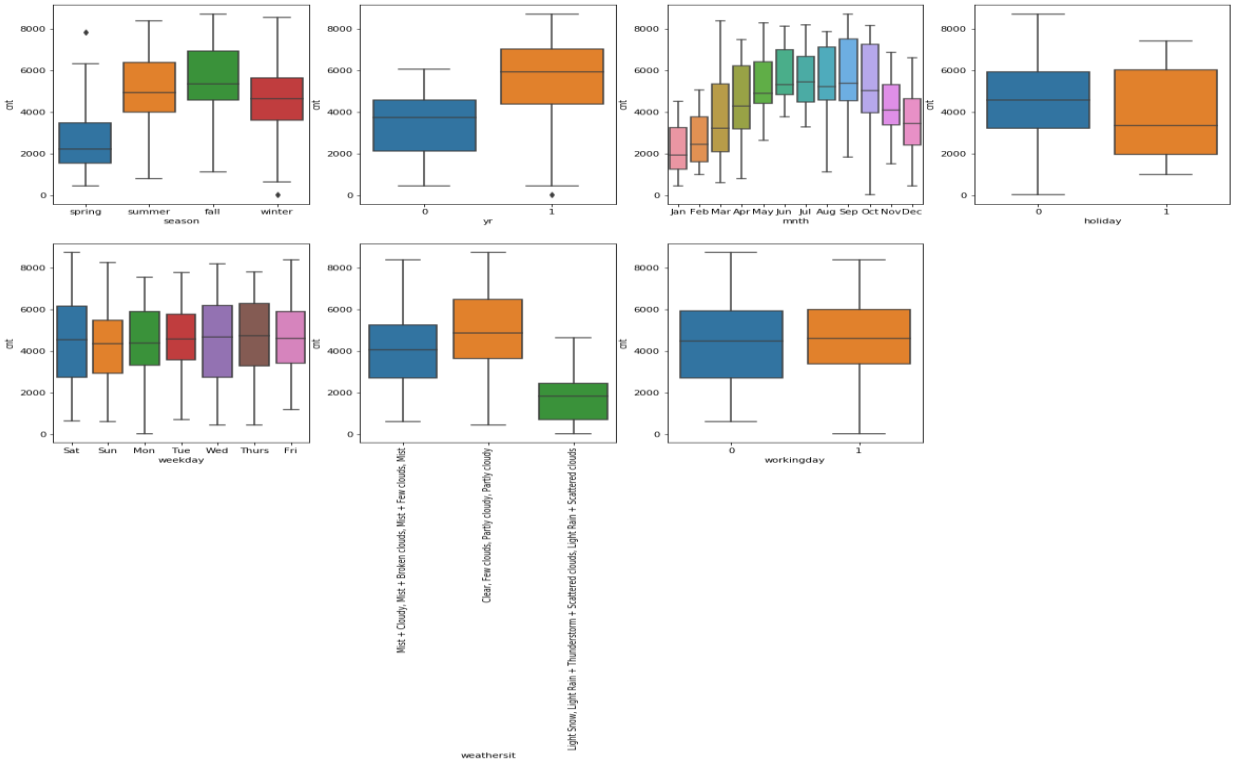
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

-> Lets discuss the relationship between the categorical variable and the dependent variable one by one:

- yr : year (0: 2018, 1:2019), In the year 2019, there is a considerable hike(from the year 2018) in the count of people that started using BoomBikes services.

- Season: The count of people using the services is also affected seasonally. Count of people using the services is higher during fall as compared to summer, winter and spring. The least usage is observed during spring season.

- Month: The count of people using the services is seen to be higher in the month of September as compared to all other months. With the least usage being observed during the first month of the year that is January.
- Holiday: Fluctuation in count of people is observed if it's a holiday v/s when it is not.
- Workingday: Fluctuation in count of people is observed if it's a working day v/s when it is not. We can ignore either workingday or holiday going further as they are highly related(We have used the RFE to finalise 15 variables to be used while building the model, which has eliminated this variable)
- Weekday: We will notice that there is only slight difference in the count of users every day.
- weathersit : We observe that there is a high drop in count of users for the weather situation Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. There is still some amount of drop noticed for weather situation Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist but not as heavy as the above mentioned weather situation.



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

-> Let us begin by taking an example:

Gender has two categories – Male and Female.

If I want to represent Gender column in a numeric sense, I will specify 1 when a person is a male or a female. So moving forward we create two columns, one for Male and other for Female specifying 1 or 0. Considering the person is a male I mark the male column as 1 and update female column as 0.

We can observe that if we eliminate one column, say the first one, we still get the same understanding of the data as, if Female is marked as 0 we will know the person is a male. In this example the person can be identified using just one column.

Hence, drop_first helps us avoid addition of an extra unnecessary column to the dataset.

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

-> Numerical variable 'atemp' has the highest correlation with the target variable.

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 state that:

1. There should be linear relationship between x and y. – Scatter plots can show the linearity between the dependent and the independent variables.

2. Error terms are normally distributed. – With the help of a histogram of error terms, I observed that the error terms are normally distributed with mean centred at 0.

3. Error terms are independent of each other. – Observed that the error terms are not dependent on one another

4. Error terms have constant variance. – With the help of a histogram of error terms, I observed that the variance of the error terms are constant and not increasing or decreasing with the change error values.

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 features contributing significantly towards explaining the demand of the shared bikes 'yr', 'temp' and 'holiday'.

General Subjective Questions

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

-> Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. There are two types of regression–

1. Simple linear regression: Model with one independent variable.
2. Multiple linear regression: Model with more than one independent variable.

Steps that we take while building a model:

1. Reading and understanding the data.
2. Visualising the data: If there is some obvious multicollinearity going on, this is the first place to catch it. This is where you'll also identify if some predictors directly have a strong association with the outcome variable.
3. Data preparation:
    - You can see if your dataset has columns with values as 'Yes' or 'No'. To fit a regression line, we would need numerical values and not string. Hence, we need to convert them to 1s and 0s, where 1 is a 'Yes' and 0 is a 'No'.
    - Also convert the categorical variables into numerical using dummy variables.
    - Treating the outliers if observed.
    - Treating the missing values if observed.
4. Splitting the data into train and test set.
5. Rescaling the data, if required, using the minmax scaling or standardisation.
6. Dividing the train data into X and y sets for model building.
7. Building a linear model: Fit a regression line through the training data using statsmodels if statistics is of importance or else sklearn can also be used.
8. Add/remove variables unless the model has all variables with p values, VIF, r-square and prob(F-statistics) in acceptable range.
9. Residual analysis of the train data: check if the error terms are also normally distributed and also other assumptions of linear regression.
10. Making Predictions using the final model.
11. Evaluating the model.

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

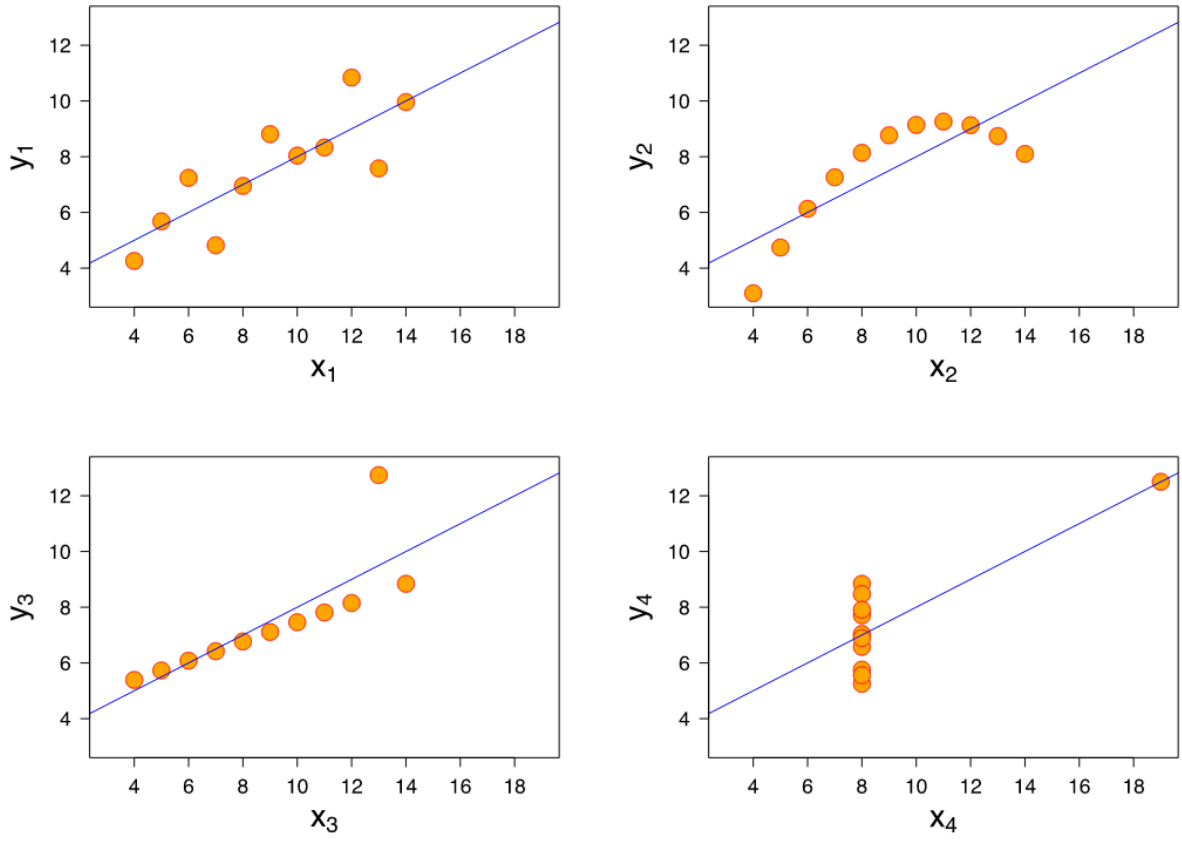
-> Anscombe's Quartet. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs as follows:

	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 :

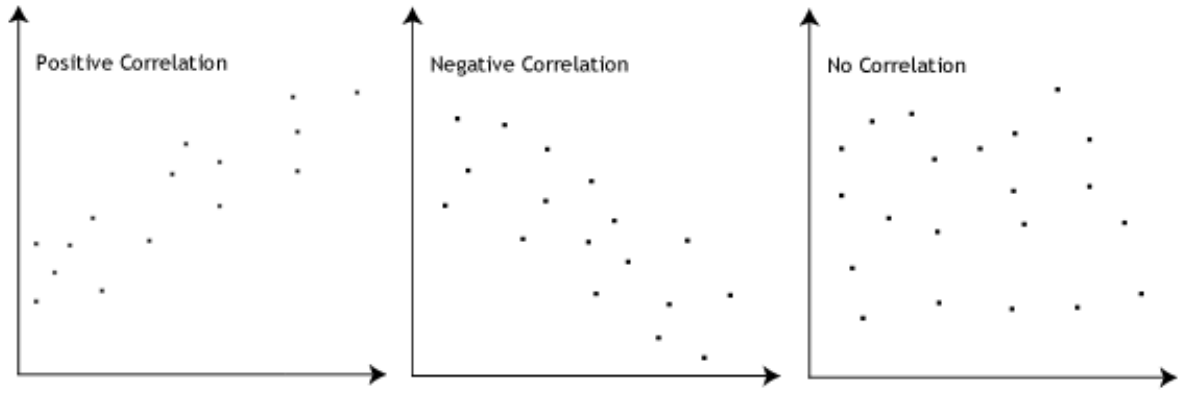


Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but does not follow a linear relationship. Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well.

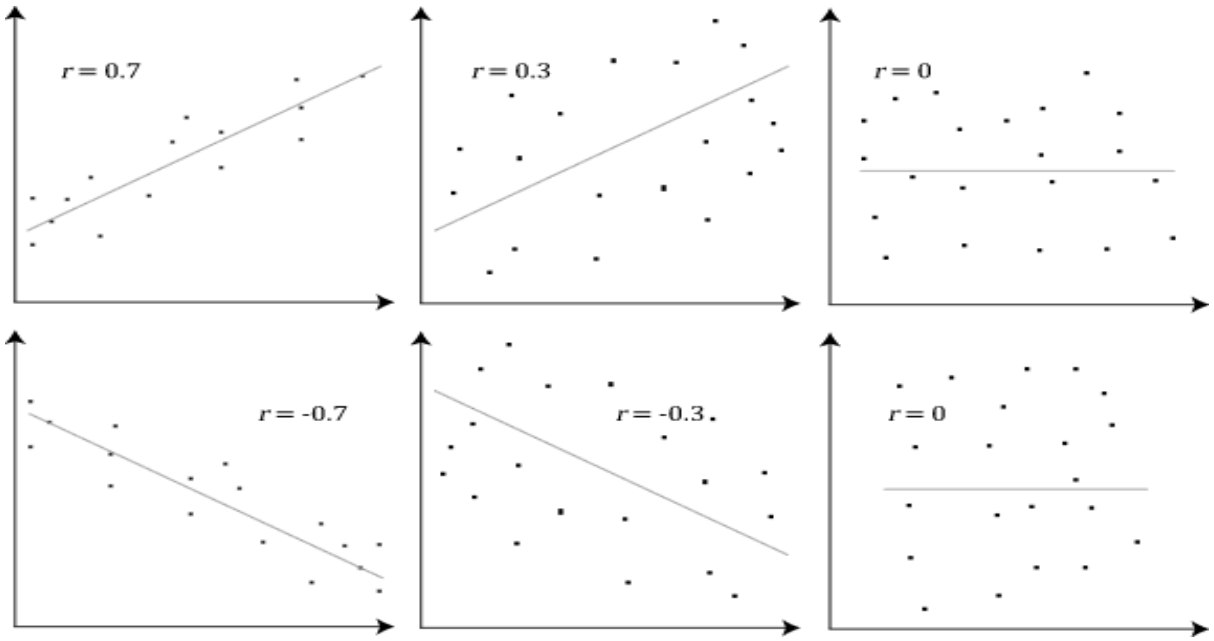
3. What is Pearson's R? (3 marks)

-> The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
-> Feature scaling is a method used to normalize the range of independent variables or features of data. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. The formula for normalization:

X' = (X-Xmin)/(Xmax – Xmin)

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. The formula for standardization:

X' = (X - μ)/σ

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

-> The variance inflation factor quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing *collinearity/multicollinearity*. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

A VIF can be computed for each predictor in a predictive model. A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.

If there is perfect correlation, then VIF = infinity.

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

-> The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. An example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.