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?**

- ➔ Analysis of categorical variables - “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth” has been done through boxplot and bar graph. Below the are point infer from the visualization are as follows:
    - Demand for Bike Rental is higher in Fall season. And booking count is drastically increased from 2018 to 2019
    - Most of the bookings has been done during the month of May, June, July, August and September. The number of booking increased from 2018 to 2019.
    - More people prefer to rent bike on working days and during holiday people prefer to stay home.
    - Wednesday, Thursday, Friday and Saturday have more number of users as compared to the start of the week. The trend is increased from 2018 to 2019.
    - Clear weather or few clouds weather attracted more booking and booking increased from 2018 to 2019.
    - Overall, the year 2019 have done more business as compared to year 2018.

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

➔ During dummy variables the attribute drop_first = True is very important to use because as it helps in reducing the extra column and make the model less complex. And hence it reduces the correlations created among dummy variables.

For the attribute drop_first: bool, the default value is False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

For example:

Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If value A is 1 then value of B & C is 0, if value B is 1 then value of A & C is 0. Therefore if the value of A & B is 0 then definitely it would be C . So we don’t need three variables.

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

➔ Temp and atemp variable are more corelated with target variable cnt.

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

➔ I validated the assumption of Linear Regression based on the below parameter are as follows:

- Normality of error terms
    - ♦ Error terms should be normally distributed
- Multicollinearity check
    - ♦ There should be insignificant multicollinearity among variables.

- Homoscedasticity
  - ♦ There should be no visible pattern in residual values.
- Linear relationship validation
  - ♦ Linearity should be visible between actual value and predicted value

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

➔ Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

- ➢ Temp
- ➢ Year
- ➢ Winter

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

- ➔ Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.
- The algorithm uses the best fitting line to map the association between independent variables with dependent variable.
- There are 2 types of linear regression algorithms
  - Simple Linear Regression – Single independent variable is used.
    - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
  - Multiple Linear Regression – Multiple independent variables are used.
    - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
  - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
  - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$
- Cost functions – The cost functions helps to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained**.
  - Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
    - The straight-line equation is $Y = \beta_0 + \beta_1 X$
    - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as Y_i.
    - $Now\ the\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum (y_i - \beta_1 x_i - \beta_0)^2$
  - The unconstrained minimization are solved using 2 methods
    - Closed form
    - Gradient descent
- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
  - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
  - OLS is used to minimize the total e^2 which is called as Residual sum of squares.

- $\text{RSS} = = \sum_{i=1}^{n}(y_i - y_{pred})^2$
- Ordinary Least Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

Assumptions
The following are some assumptions about dataset that is made by Linear Regression model

1. Multi-collinearity
   - ✓ Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
2. Auto-correlation
   - ✓ Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
3. Relationship between variables
   - ✓ Linear regression model assumes that the relationship between response and feature variables must be linear.
4. Normality of error terms
   - ✓ Error terms should be normally distributed
5. Homoscedasticity
   - ✓ There should be no visible pattern in residual values.

2. **Explain the Anscombe's quartet in detail.**

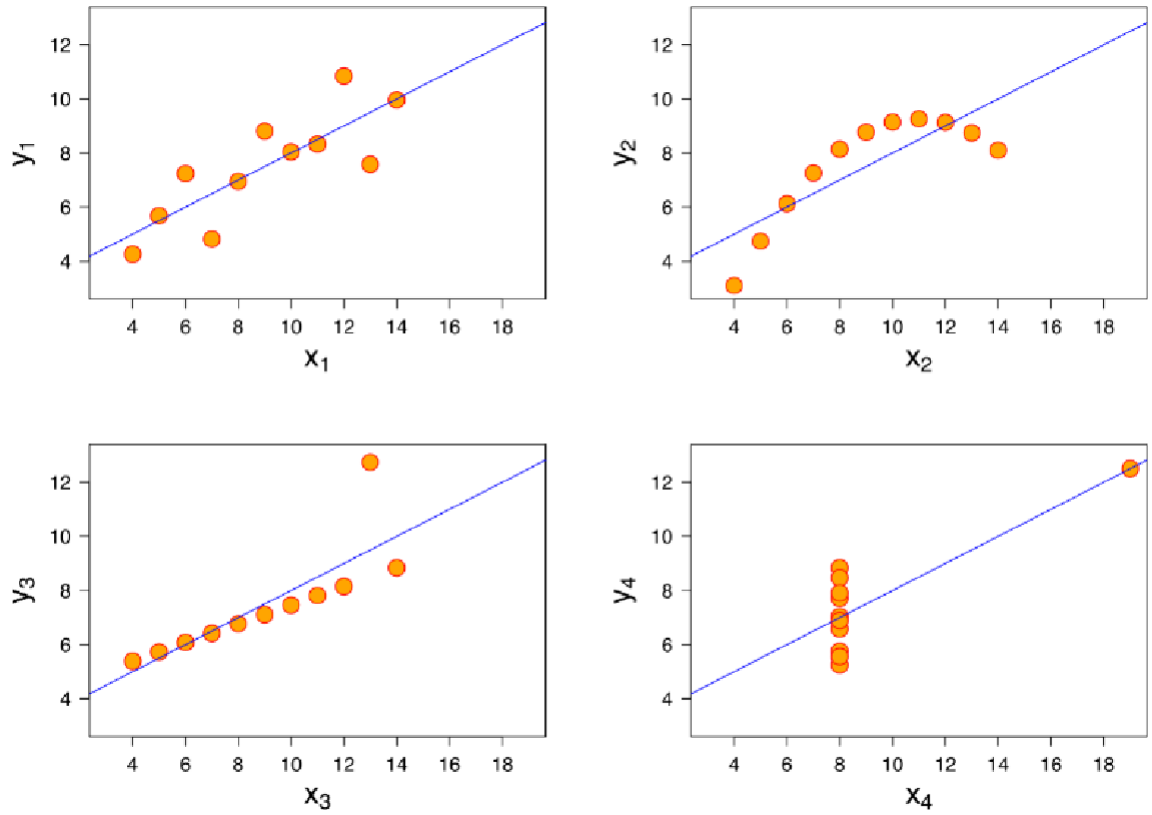
➔ Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	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:



• Dataset I appears to have clean and well-fitting linear models.
• Dataset II is not distributed normally.
• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

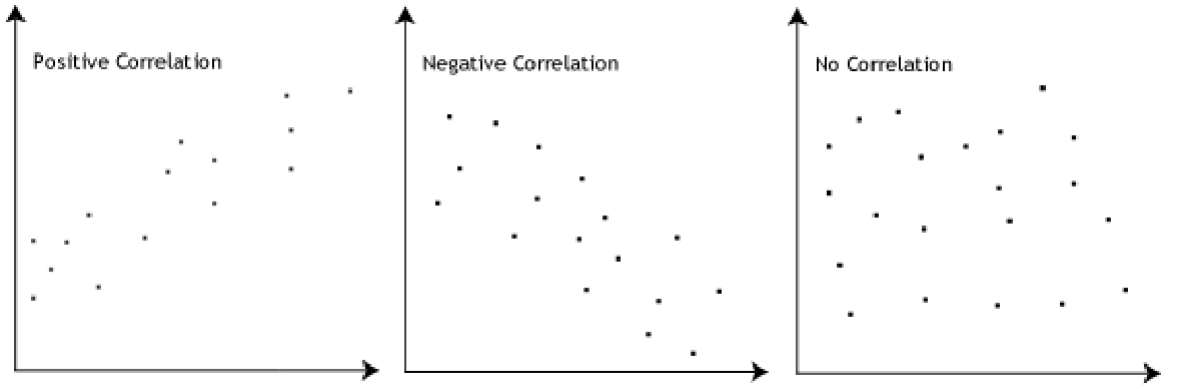
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset

3. **What is Pearson's R?**

➔ Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be

positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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:



---

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

➔ Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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.

Example**:** If an algorithm is not using feature scaling method then it can consider the value 1000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalized scaling	**Standardized scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization

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

➔ If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

➔ The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.