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?**

The categorical variables available in the assignment are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of may, jun, jul, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat, Sun and Mon have a greater number of bookings.
- When it’s holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted a greater number of bookings from the previous year, which shows good progress in terms of business.

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

When creating dummy variables for categorical features in a dataset, the `drop_first=True` parameter is used to avoid the dummy variable trap or multicollinearity issues in linear regression models. The dummy variable trap occurs when one dummy variable can be predicted from the others. The issue arises because the presence of all dummy variables in the model can create a linear dependence among them.

By setting `drop_first=True`, one level of the categorical variable is dropped, and the remaining dummy variables are sufficient to represent all possible categories.

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

‘temp’ variable has the highest correlation with the target variable. The scatter plot between temp variable and cnt is linear show that as temp increases the count of bike sharing increases.

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

I have validated the assumption of Linear Regression Model based on below 5 assumptions -

1. Linear relationship validation
   Linearity should be visible among independent variables and dependent variable

2. Normality of error terms
   Error terms should be normally distributed

3. Homoscedasticity
   There should be no visible pattern in residual values as the error terms change.

4. Multicollinearity check
   The VIF values has to be less than 10, and the p-value has to less than 0.05

5. Independence of residuals
   No auto-correlation

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:**
  Temperature is the Most Significant Feature which affects the demand of shared bikes
- **Yr:**
  The growth year on year seems organic given the geological attributes
- **Season_spring:**
  Spring season plays the crucial role in the demand of shared bikes shows negative correlation with ct.

General Subjective Questions

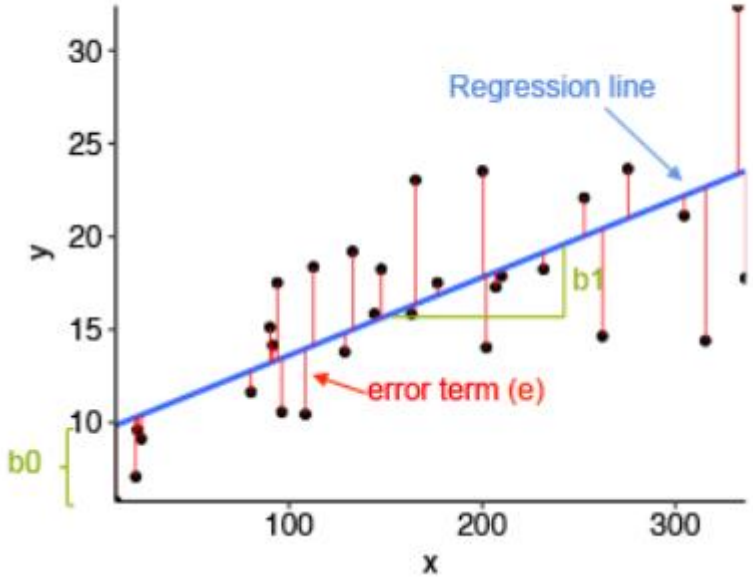
1. Explain the linear regression algorithm in detail.

Types of Linear Regression

1. **Simple Linear Regression:**
   This is like when we have only one thing that might affect another. For instance, the number of hours you study (one thing) might influence your exam score (another thing).
2. **Multiple Linear Regression:**
   Sometimes, more than one thing can affect another. For example, in addition to study hours, maybe sleep hours and exercise also matter for exam scores. Multiple linear regression helps consider all these things together.

Simple Linear Regression

A simple linear regression model attempts to explain the relationship between a dependent variable and an independent one using a straight line.



The formula for simple linear regression is

$y = b0 + b1X + e$

where,

- y is the dependent variable.
- X is the independent variable.
- b0 is the y-intercept (the value of y when X is 0).
- b1 is the slope of the line (the change in Y for a one-unit change in X).
- e is the error

Multiple Linear Regression

The main difference between simple linear regression and multiple linear regression lies in the number of independent variables they use. Multiple linear regression involves predicting a dependent variable based on multiple independent variables.

The formula for Multiple linear regression is

$\hat{y} = \theta 0 + \theta jxi + \theta jxi + \cdots + \theta nxn$

Where,

- $\hat{y}$ is the predicted value.

- n is the number of features.
- xi is the ith feature value.
- θj is the jth model parameter

If this formula is compared with a simple linear regression formula, then
θ0 is the y-intercept (which is **b0** in simple linear regression)
θ1, θ2, ⋯, θn are the slopes for each feature xi (which is **b1** in simple linear regression)

Assumptions of linear regression

- **There is a *linear relationship* between X and Y:**
  X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.

- **Error terms are *normally distributed* with mean zero:**
  The assumption of normality is made, as it has been observed that the error terms generally follow a **normal distribution with mean equal to zero** in most cases.

- **Error terms are *independent* of each other:**
  The error terms should not be dependent on one another (like in time-series data wherein the next value is dependent on the previous one).

- **Error terms have *constant variance* (homoscedasticity):**
  The variance should not increase (or decrease) as the error values change. Also, the variance should not follow any pattern as the error terms change.

2. Explain the Anscombe's quartet in detail.

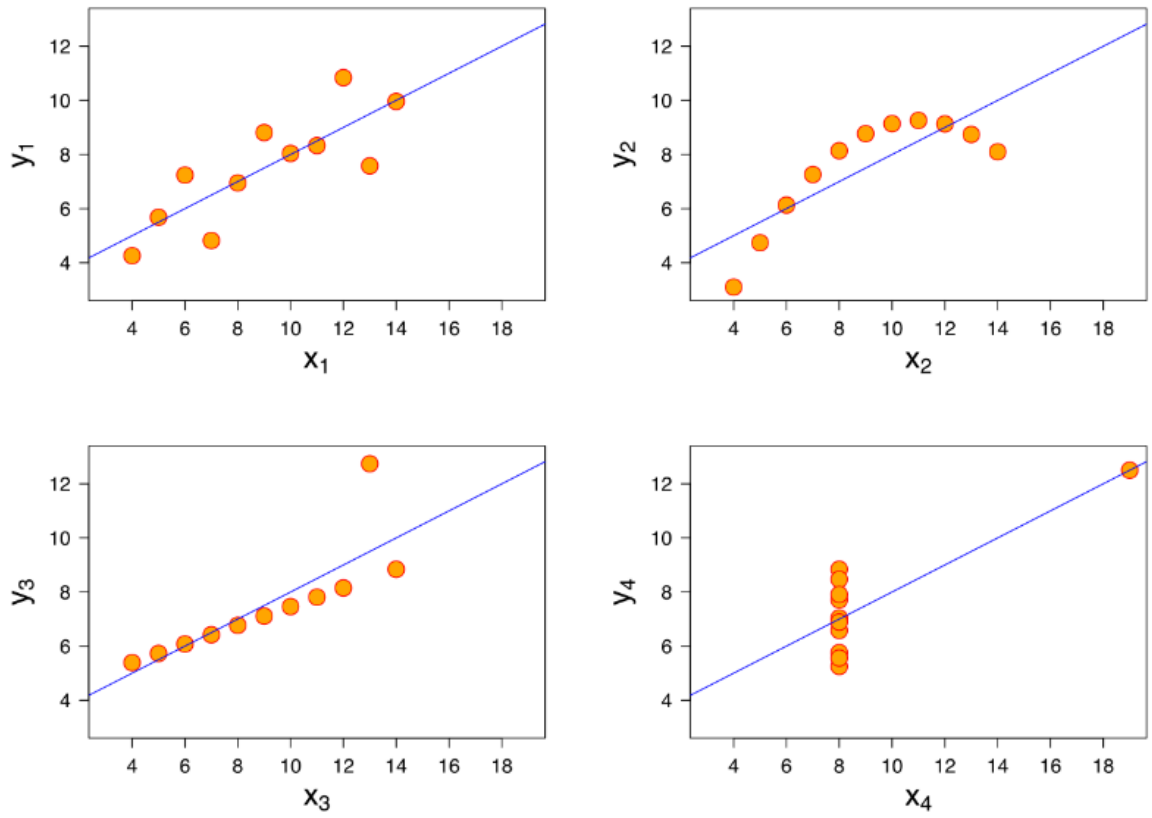
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when 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 appear 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.

3. What is Pearson's R?

Pearson's correlation coefficient, often denoted as r , is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by

Karl Pearson and is commonly used in statistics to assess the degree to which two variables are linearly related.

The formula for Pearson's correlation coefficient r between variables X and Y with sample size n is given by:

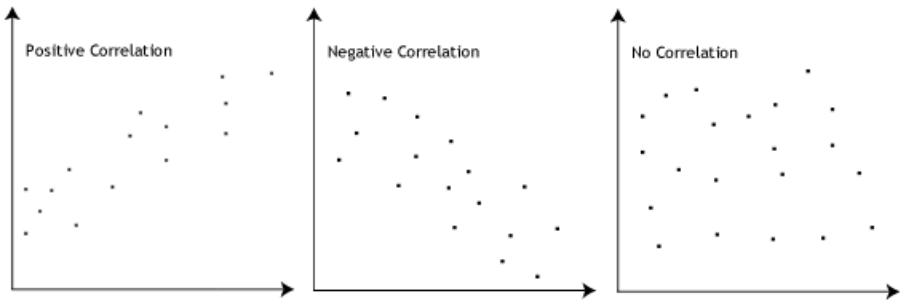
$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Here:

- Xi and Yi are individual data points.
- X_bar and Y_bar are the means of variables X and Y, respectively.

The resulting r value ranges from -1 to 1:

- r = 1: Perfect positive linear correlation.
- r = -1: Perfect negative linear correlation.
- r = 0: No linear correlation.



Interpretation:

- The magnitude of r indicates the strength of the linear relationship: closer to 1 or -1 implies a stronger relationship.
- The sign of r indicates the direction of the relationship: positive for direct, negative for inverse.

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

Scaling:

Scaling is the process of transforming numerical variables to a standard range or distribution. The goal is to bring different variables or features to a similar scale, making them comparable and preventing certain features from dominating others due to their larger magnitudes. It is a crucial step in many machine learning algorithms, particularly those that are distance-based or involve gradient descent optimization.

Why is Scaling Performed?

1. Equality of Importance: Ensures that all features contribute equally to the model training, preventing the dominance of variables with larger scales.

2. Convergence: Accelerates the convergence of gradient-based optimization algorithms by providing a more spherical or isotropic cost function.

3. Distance-Based Algorithms: Improves the performance of algorithms sensitive to distances, such as k-nearest neighbors or support vector machines.

4. Regularization: Enhances the effectiveness of regularization techniques that penalize large coefficients.

Difference between Normalized Scaling and Standardized Scaling:

Normalized scaling formula:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Standardized scaling formula:

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	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.

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line

Importance of Q-Q plots in Linear regression:

In the context of linear regression, Q-Q plots play a crucial role in checking the normality assumption of residuals. The residuals are the differences between the observed and predicted values, and their normality is vital for valid hypothesis testing, confidence interval estimation, and the overall reliability of the regression model. By examining the Q-Q plot, analysts can visually inspect whether the residuals adhere to a normal distribution. Detecting deviations from normality informs researchers about potential issues in the model, such as non-linearity or the presence of outliers. Addressing these concerns can lead to improved model validity and more accurate statistical inferences.