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? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The categorical variables as independent variables have shown before and after the analysis that these have some effect to the dependent (target variable). Below are the list of observations made during the analysis

- Summer and Fall seasons are the TOP two seasons for high demand for bike rental.
- Clear weather is the TOP most weather for high demand for the bike rental.
- May, Jun, Jul, Aug, Sep, Oct are the TOP 6 months to focus for a higher demand for bike rental.
- Sat, Wed , and Thu – these 3 days have higher bookings, however other days are not far behind with demand. Demand for bike rental on Sunday is lower.
- There is low demand on holidays.
- There is almost equal demand for bike rental be it a working day or a non-working day. (exception is holidays where there is low demand)

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

- A dummy variable is a binary variable that takes on the value of 1 if an observation falls into a particular category and 0 otherwise.
- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. In such a case it will create k-1 which equates to 3-1 = 2 dummy variables.
- Drop_first=True is important to reduce redundancy and to increase performance.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The ‘temp’ variable has the highest correlation with the target variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

After building the model on the training set the below validation can be done on the assumptions.

- R-Squared values of the training data set and test data set should be close to each other.
- The test data set should be linear on actual test data and predicted data. (homoscedasticity)
- Further validations include
    - No high VIF
    - Very low p-values
    - Residual errors are normally distributed.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top three features significantly contributing towards the demand of the shared bikes are:

- 'temp' meaning temperature – this can be interpreted as higher the temperature higher is the booking.
- 'yr' meaning year – this can be interpreted as the second year has more booking than first year. Thus indicating the demand in the bike sharing rental business in the first two years of operations. Hence further business can be expected in the upcoming year.
- 'winter' meaning winter season – this can be interpreted as even if booking count point of view winter has lower bookings than summer and fall but still winter has significant bookings while the temperature is low. Thus proves the demand for the service.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear Regression:
Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

- When there is only one independent feature, it is known as Simple Linear Regression
- When there are more than one feature, it is known as Multiple Linear Regression.

Similarly,

- when there is only one dependent variable, it is considered Univariate Linear Regression,
- when there are more than one dependent variables, it is known as Multivariate Regression.

Why Linear Regression is Important?

The interpretability of linear regression is a notable strength. The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics. Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.

Linear regression is not merely a predictive tool; it forms the basis for various advanced models. Techniques like regularization and support vector machines draw inspiration from linear regression, expanding its utility. Additionally, linear regression is a cornerstone in assumption testing, enabling researchers to validate key assumptions about the data.

Types of Linear Regression

There are two main types of linear regression:

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$y = \beta_0 + \beta_1 X$

where:

- Y is the dependent variable
- X is the independent variable
- β_0 is the intercept
- β1 is the slope

Multiple Linear Regression

This involves more than one independent variable and one dependent variable.

The equation for multiple linear regression is:

$y=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\ldots\beta n Xn$

where:

- Y is the dependent variable
- X_1, X_2, …, Xn are the independent variables
- β_0 is the intercept
- β_1, β_2, …, βn are the slopes

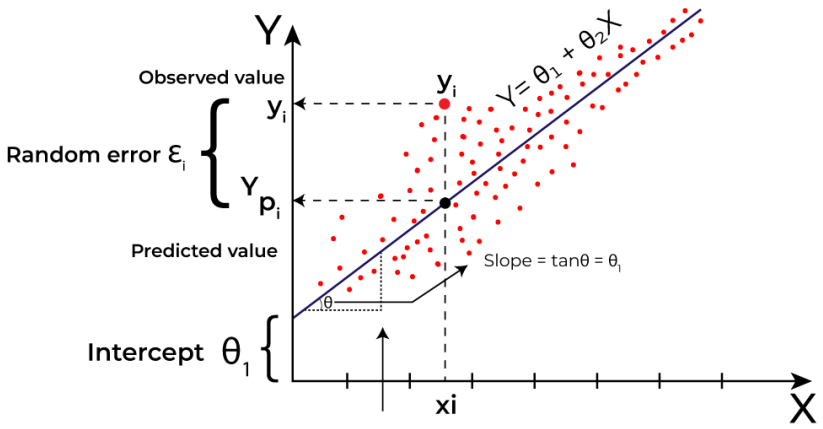
The goal of the algorithm is to find the **best Fit Line** equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

What is the best Fit Line?

Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).



Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

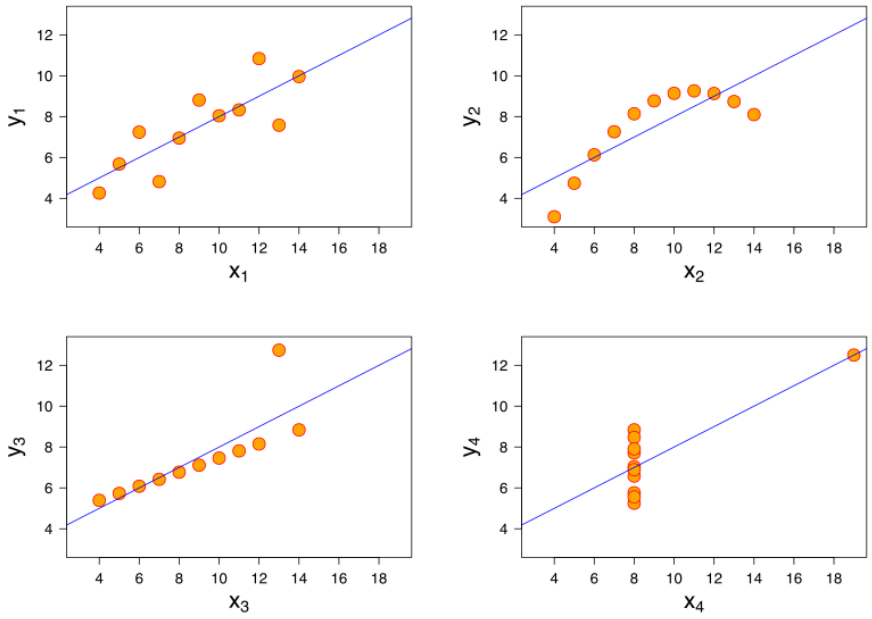
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best-fit line for our model.

We utilize the cost function to compute the best values in order to get the best fit line since different values for weights or the coefficient of lines result in different regression lines.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

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, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



Anscombe's quartet							
Dataset I		**Dataset II**		**Dataset III**		**Dataset 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

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:

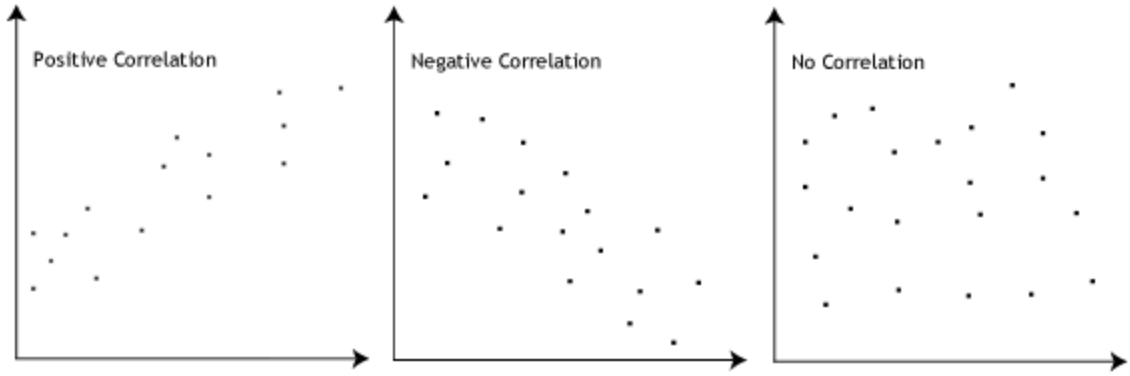
- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

- 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.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
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:



Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Feature scaling is one of the most important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.

There are some feature scaling techniques such as Normalization and Standardization that are the most popular and at the same time, the most confusing ones.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

X_new = (X - X_min)/(X_max - X_min)

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

Difference between Normalization and Standardization

S.NO.	Normalization	Standardization
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.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is often called as Z-Score Normalization.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

VIF explains the relationship between correlations among independent variables. 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.

This happens because there are multiple variables affecting the target variable, thus impacting the target variable being affected by similar relations affecting the predictability of target variable. Hence we must remove variables with high VIF.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same **population** or not. Q-Q plots are particularly useful for assessing whether a dataset is **normally distributed** or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Quantiles And Percentiles

Quantiles are points in a dataset that divide the data into intervals containing equal probabilities or proportions of the total distribution. They are often used to describe the spread or distribution of a dataset. The most common quantiles are:

1. **Median(50th percentile)**: The median is the middle value of a dataset when it is ordered from smallest to largest. It divides the dataset into two equal halves.

2. **Quartiles(25th, 50th, and 75th percentiles)**: Quartiles divide the dataset into four equal parts. The first quartile (Q1) is the value below which 25% of the data falls, the second quartile (Q2) is the median, and the third quartile (Q3) is the value below which 75% of the data falls.

3. **Percentiles**: Percentiles are similar to quartiles but divide the dataset into 100 equal parts. For example, the 90th percentile is the value below which 90% of the data falls.

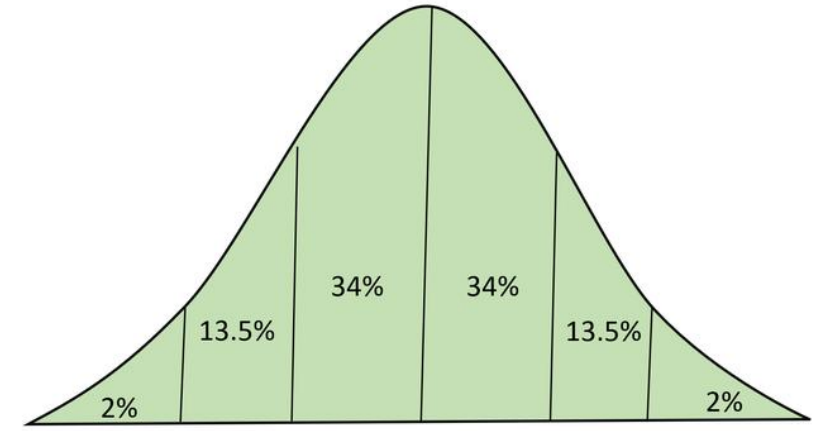
Note:

- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

- For reference purposes, a 45% line is also plotted; **For** if the samples are from the same population then the points are along this line.

Normal Distribution:

The normal distribution (aka Gaussian distribution Bell curve) is a continuous probability distribution representing distribution obtained from the randomly generated real values.



Normal Distribution with Area Under CUrve

Interpretation of Q-Q plot

- If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.
- Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.
