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?**

Answer:

There are six categorical variables including the holiday and working day along with season, month, humidity, and weathersit

Below are the observations:

- Most of the bookings are done in the months of April to October, there is consistent trend which was observed here.
- Booking had very less effect when compared to the holiday variable, it has negative raise
- Working day population initially seemed a good fit but eventually it hasn't contributed much to the final prediction
- Season fall has more bookings when compared to spring,winter,summer.
- Clear weather has more bookings
- Year has gradual growth and 2019 has more booking compared to 2018
- Weekend seems to have more bookings say"Saturday" then weekdays

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

Answer:

- It helps to reduce the extra column created during the dummy variable creation
- It is as the syntax follows for 'k'categorical variables present there will be 'k-1' dummy variables

Example:

For 3 categorical variables say 'R','P','Q', we have 2 dummy variables 'R','P',one will be dropped. This is with a notion that what is not present in P & Q will be in R.so we will ignore Q

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

Answer:

- Temp 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?

Answer:

Validated on the basis of

- Normality – Error terms are normally distributed and are centered to Zero.
- Multicollinearity- No multicollinearity was found as per VIF checks.
- Linearity- Strong linear relationships were found between input and target variable.
- Residual patterns- Error terms are independent of each other and no specific patters were found
- Constant Vraince(Homoscedasticity)- They follow constant variance indicating homoscedasticity.

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

Answer:

- **Temperature** (temp) - A coefficient value of 0.4918 increase indicates the bike hire numbers increase by 0.4918 units
- **Weather Situation 3** (weathersit_3) - A coefficient value 0.3051 indicated that a decrease in variable decreases the bike hire numbers same units
- **year(yr) -** A coefficient value 03308' indicates that an increase will increase the bike hire by same units

* weathersit_3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

General Subjective Questions with Answers & Examples:

1)Explain the linear regression algorithm in detail.

Answer:

An algorithm that provides the linear relationships between independent and dependent variables to predict future events' outcomes.

- This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals, where the residual is the difference between the predicted variable to the actual variable.

The purpose is to examine:

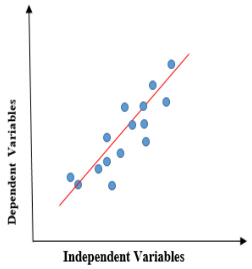
1) If a predictor variable is good for predicting an outcome variable
2) Which variables, in particular, are significant predictors of the outcome variable, and in what way do they impact the outcome variable

There are 2 types of LinearRegression Models:

- Simple Linear Regression - Uses one independent variable and 1 dependent variable

- Multiple Linear Regression- Uses multiple independent variable and 1 dependent variable.

To calculate best-fit line linear regression uses a traditional slope-intercept form



Model: $Y=mx+c$ (Equation of line)

where c- constant, m-intercept, dependent variable -‘y’ and independent variable- ‘x’

Step-by-step procedure in LR

1) Read and understand the data
2) Train and Test the data
3) Residual Analysis
4) Predict the data

Example using simple linear regression

X- Hours Studied (x)

Y- Test Score (y)

Basic understanding: number of hours effect the test score

Hours Studied (x)	Test Score (y)
1	50
2	55
3	65
4	70
5	75

We want to find the relationship between hours studied and test scores.

Steps to Perform Linear Regression

- **Plot the Data**: Visualize the data points on a scatter plot to see if there is a trend.
- **Find the Line of Best Fit**:
  - Use a method (like least squares) to find the line that best fits the data. This line minimizes the distance between the line and all data points.
  - The line we get after the calculation is line is:
    - $y=6x+44y$

Using the Regression Line

- To predict the test score for a student who studies 6 hours:

$$y=6\times6+44=36+44=80$$

So, the predicted test score for 6 hours of study is 80.

Interpretation

- **Slope (m)**: The slope of 6 means that for each additional hour studied, the test score increases by 6 points.
- **Y-intercept (c)**: The y-intercept of 44 is the predicted test score if no hours are studied.

2)Explain the Anscombe’s quartet in detail

Answer:

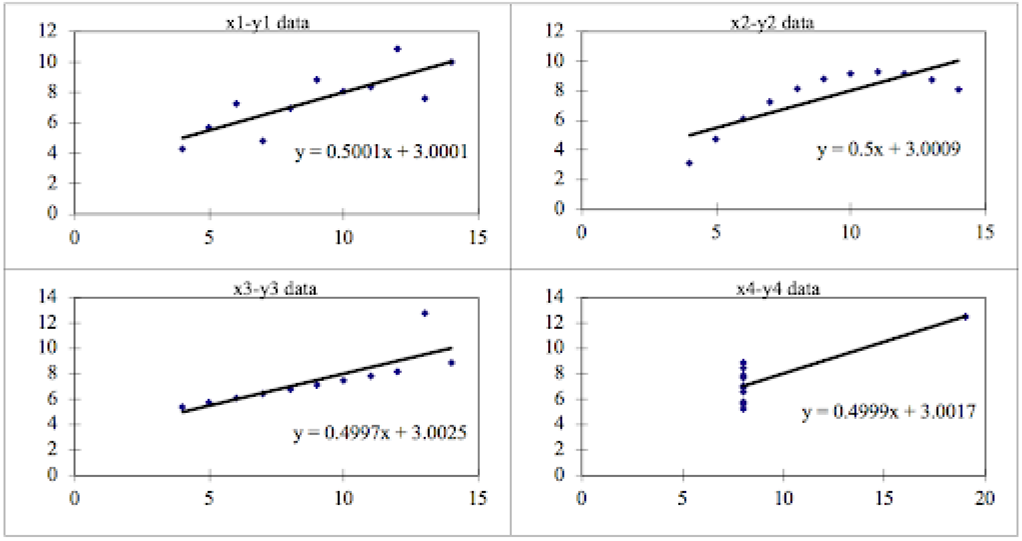
Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model.

- These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean, correlations, r-squared) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Purpose:

- Tells us about the importance of visualizing data
- This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).
- This thus plays a crucial step in exploratory data analysis and linear regression by helping us understand the linear relationships between variables.

Example



Anscombe’s Quartet Four Datasets

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3) What is Pearson's r

Answer:

Pearson correlation coefficient (*r*) is the most widely used correlation coefficient.

- It is a number between –1 and 1 that measures the strength and direction of the linear relationship between two quantitative variables.
- Is also an inferential statistic, which can be used to test whether there is a significant relationship between two variables.

□ 1 indicates a perfect positive linear relationship,

□ −1 indicates a perfect negative linear relationship, and

- 0 indicates no linear relationship.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure

Pearson correlation coefficient is used when:

- Both variables are quantitative
- The variables are normally distributed
- The data have no outliers
- The relationship is linear

Formula with example:

Here r is the Pearson correlation coefficient, xi are the individual values of one variable e.g. age, yi are the individual values of the other variable e.g. salary and x̄ and ȳ are the mean values of the two variables respectively.

x_i are the **individual values** of one **variable** e.g. age

y_i are the **individual values** of the other **variable** e.g. salary

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where r is the Pearson correlation coefficient,

$\bar{x}$ and $\bar{y}$ are respectively the **mean values** of the two variables.

In our example, we calculate the mean values of age and salary, subtract the mean values of age and salary, and then multiply. Then sum up the individual results of the multiplication.

The expression in the denominator ensures that the correlation coefficient is scaled between -1 and 1

Testing correlation coefficients for significance

- Correlation coefficient is calculated using data from a sample. In most cases, however, we want to test a hypothesis about the population to understand the correlation
- For this, we test whether the correlation coefficient in the sample is statistically significantly different from zero

Hypotheses

Null hypothesis: The correlation coefficient is not significantly different from zero (There is no linear relationship).

Alternative hypothesis: The correlation coefficient deviates significantly from zero (there is a linear correlation**)**

In above example:

- The null hypothesis is then: There is no correlation between salary and age in the German population.
- and the alternative hypothesis: There is a correlation between salary and age in the German population.

Significance and the t-test

Whether the Pearson correlation coefficient is significantly different from zero based on the sample surveyed can be checked using a t-test. Here, r is the correlation coefficient and n is the sample size.

r is the correlation coefficient

and *n* is the sample size

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

p-value can then be calculated from the test statistic, if p value is < 5% null hypothesis is rejected.

Assumptions

- To calculate r, only two metric variables must be present. Metric variables are, for example, a person's weight, a person's salary or electricity consumption.
- The Pearson correlation coefficient, then tells us how large the linear relationship is. If there is a non-linear correlation, we cannot read it from the Pearson correlation coefficient.
- Variables must also be normally distributed.

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

Answer:

Scaling is a technique used to adjust the range of features in your dataset so that they fit within a specific scale. It is commonly used in data preprocessing before applying machine learning algorithms.

Scaling is performed for several reasons:

1. **Improving Convergence Speed**: Many optimization algorithms converge faster when features are scaled properly.
2. **Ensuring Equal Contribution**: Features with larger ranges can dominate the learning process. Scaling ensures that all features contribute equally.
3. **Improving Model Performance**: Algorithms like k-nearest neighbors (KNN), support vector machines (SVM), and gradient descent-based methods can perform better with scaled data.
4. **Reducing Sensitivity to Outliers**: Scaling can reduce the impact of outliers.

Types of Scaling: Normalization vs. Standardization

Normalization (Min-Max Scaling):

- **Definition**: Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1].

  Formula: X_scaled = (X — X_min) / (X_max — X_min)

  where X is the original feature value, Xmin is the minimum value of the feature, and Xmax is the maximum value of the feature.

- **Use Cases**: Useful when you know the data has a bounded range and when the distribution is not Gaussian (normal). Often used in image processing and neural networks.

Standardization (Z-score Scaling):

- **Definition**: Standardization scales the data to have a mean of 0 and a standard deviation of 1.
- **Formula**: $X' = (X - \mu)/\sigma$

where X is the original feature value, μ is the mean of the feature, and σ is the standard deviation of the feature.

- **Use Cases**: Useful when the data follows a Gaussian distribution. Often used in algorithms that assume or benefit from normally distributed data (e.g., linear regression, logistic regression).

Key Differences

Range:

Normalization: Scales data to a specific range, usually [0, 1] or [-1, 1].

Standardization: Scales data to have a mean of 0 and standard deviation of 1.

Effect on Distribution:

Normalization: Does not change the shape of the original distribution.

Standardization: Standardizes the distribution but can be affected by outliers.

Suitability:

Normalization: Suitable when the features have different units or when no assumptions about the data distribution are made.

Standardization: Suitable when the data follows a normal distribution or when the algorithm requires features to have zero mean and unit variance

Example:

Sample	Height (cm)	Weight (kg)
1	150	50
2	160	60
3	170	70
4	180	80

Heights will be scaled to [0, 1] as follows:

- Minimum height = 150 cm, Maximum height = 180 cm.
- For height of 150 cm: X′=0
- For 180cm X′=1

Standardization (Z-score Scaling):

Heights will be standardized as follows:

- Mean height (μ) = 165 cm, Standard deviation (σ) = 12.91 cm.

- For height of 150 cm-1.16
- For height of 180 cm: X′=1.1s

Conclusion:

Scaling is a critical preprocessing step in machine learning that helps improve the performance and accuracy of models. Normalization and standardization are two common methods used for scaling, each suitable for different types of data and machine learning algorithms

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

Answer:

The Variance Inflation Factor (VIF) is a measure used to detect the presence of multicollinearity in regression models. Multicollinearity occurs when two or more predictor variables in a model are highly correlated, leading to unreliable estimates of regression coefficients

Formula:

The VIF for a predictor Xi is calculated as:

VIF(Xi)=1/(1−Ri2)

where Ri2 is the coefficient of determination obtained by regressing Xi on all the other predictors in the model.

- If VIF(Xi)=1, there is no multicollinearity.
- If VIF (Xi)>1, there is some degree of multicollinearity.
- Higher values indicate a higher degree of multicollinearity.

Why Does VIF Become Infinite?

VIF becomes infinite when Ri2=1. This means that the predictor Xi can be perfectly predicted using a linear combination of the other predictors in the model.

Causes of Infinite VIF

1. **Perfect Multicollinearity**: This occurs when one predictor is a perfect linear combination of one or more other predictors. X3=2X1+3X2 then X3 can be perfectly predicted from X1 and X2.
2. **Duplicate Variables**: If the same variable is included in the dataset more than once (or variables that are exact duplicates of each other), VIF will be infinite.
3. **Constant Variables**: If a predictor variable is constant (i.e., it has the same value for all observations), it will lead to infinite VIF because it does not add any information to the model.

Example:

Consider a dataset with three predictors X1, X2, and X3 where: X3=2X1+3X2

When we compute Ri2 for X3:

- The regression of X3 on X1 and X2 will have Ri2=1 because X3 is perfectly predicted by X1 and X2.
- Substituting Ri2=1 into the VIF formula:
    - VIF(X3) =1/(1−1) =1/0 = ∞

To handle infinite VIF values

1. **Remove Perfectly Collinear Variables**: Identify and remove one of the variables that are perfectly collinear.
2. **Principal Component Analysis (PCA)**: Use PCA to transform the predictors into a set of uncorrelated components.
3. **Regularization**: Use techniques like Ridge Regression, which can handle multicollinearity by adding a penalty to the size of the coefficients.

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

Answer:

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the points on the Q-Q plot fall approximately along a straight line, it indicates that the dataset follows the theoretical distribution.

Interpret a Q-Q Plot

- **Straight Line**: If the points lie roughly on a straight line, the data is approximately normally distributed.
- **S-Shaped Curve**: If the points form an S-shaped curve, the data may have heavier tails (leptokurtic) or lighter tails (platykurtic) than the normal distribution.
- **Bowed Shape**: If the points form a bowed shape (concave up or down), the data may be skewed.

Use and Importance of Q-Q Plots in Linear Regression

In linear regression, Q-Q plots are used to assess the normality of the residuals (errors). The assumptions of linear regression include that the residuals are normally distributed. Normality of residuals is important for the following reasons:

1. **Validating Model Assumptions**: The assumption of normally distributed residuals underpins many statistical tests for linear regression, including hypothesis tests for the coefficients, confidence intervals, and the F-test for overall significance.
2. **Improving Predictive Performance**: If the residuals are normally distributed, the model's predictions are more reliable and the confidence intervals for predictions are accurate.
3. **Diagnosing Model Fit**: Deviations from normality can indicate issues with the model, such as missing variables, incorrect functional form, or the presence of outliers.

Example of a Q-Q Plot in Linear Regression

Let's consider a simple example where we have performed linear regression and want to check the normality of the residuals.

Steps

1. **Fit a Linear Regression Model**: Suppose we have a dataset with X(independent variable) and Y (dependent variable), and we fit a linear regression model.
2. **Obtain Residuals**: Calculate the residuals from the fitted model. Residuals are the differences between the observed values and the values predicted by the model.
3. **Generate a Q-Q Plot**: Plot the quantiles of the residuals against the quantiles of a standard normal distribution.

Interpretation

- **Normal Residuals**: If the residuals are normally distributed, the points on the Q-Q plot will lie approximately on a straight line.
- **Non-normal Residuals**: If the points deviate significantly from the straight line, this suggests that the residuals are not normally distributed.

Example Visualization

Let's consider the residuals from a linear regression model:

Residuals:

[0.1, -0.2, 0.3, -0.4, 0.2, -0.1, 0.05, -0.05, 0.1, -0.2]

Generate a Q-Q plot using these residuals:

- The x-axis represents the theoretical quantiles from the standard normal distribution.
- The y-axis represents the actual quantiles from the residuals.

If the residuals are normally distributed, the points will align along the 45-degree reference line.

In Brief:

- A Q-Q plot is a vital diagnostic tool in linear regression for assessing the normality of residuals.

- Deviations from normality, as indicated by the Q-Q plot, can guide the analyst in improving the model through transformations, adding/removing variables, or addressing outliers.