General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is type of supervised machine-learning algorithm It provides valuable insights for prediction and data analysis.

It learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets. It computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation with observed data. It predicts the continuous output variables based on the independent input variable.

Types of Linear Regression

When there is only one independent feature it is known as Simple Linear Regression or Univariate Linear Regression and when there are more than one feature it is known as Multiple Linear Regression or Multivariate Regression.

Key Concepts:

1. **Dependent and Independent Variables**:
   - **Dependent Variable (Y)**: The outcome we're trying to predict.
   - **Independent Variable(s) (X)**: The input feature(s) used to make predictions.
2. **Equation of a Line**: In simple linear regression, the relationship between the dependent variable and the independent variable is described by the equation of a line:

Y=b0+b1X+ϵY

YY: Dependent variable (predicted value)

- b0b_0: Y-intercept (the value of Y when X=0)
- b1b_1: Slope of the line (the change in Y for a one-unit change in X)
- ϵ\epsilon: Error term (accounts for variability in Y not explained by X)

3. **Cost Function**: The objective of linear regression is to find the best-fitting line, achieved by minimizing the cost function. Commonly, the cost function used is the Mean Squared Error (MSE).
4. **Optimization**: Techniques like Gradient Descent or Ordinary Least Squares are used to minimize the cost function and find the optimal values.

Steps in Performing Linear Regression:

1. **Data Collection**: Gather data with known values of the dependent and independent variables.
2. **Data Preprocessing**:
    - Handle missing values.
    - Encode categorical variables, if any.
    - Split the dataset into training and testing sets.
3. **Model Training**: Use the training set to fit the linear regression model, determining the intercept and slope that minimize the cost function.
4. **Model Evaluation**: Assess the model's performance using metrics such as Mean Squared Error (MSE), R-squared, etc.
5. **Prediction**: Make predictions on new data using the trained model.

2.Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that, despite having nearly identical basic statistical properties like mean, standard deviation, and correlation coefficient, appear vastly different when visualized on a scatter plot, highlighting the crucial need to always graph data alongside relying solely on summary statistics in data analysis.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Key points about Anscombe's quartet:

- **Identical summary statistics:**

Each of the four datasets in the quartet has almost the same mean, standard deviation, and correlation coefficient for both the x and y variables, making them appear statistically similar when only looking at numerical summaries.

- **Distinct visual patterns:** When plotted on a graph, however, the four datasets reveal very different patterns: one shows a clear linear relationship, another has a strong non-linear pattern, one has a single outlier significantly impacting the data, and the last appears almost random.

3.What is Pearson's R?

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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.

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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

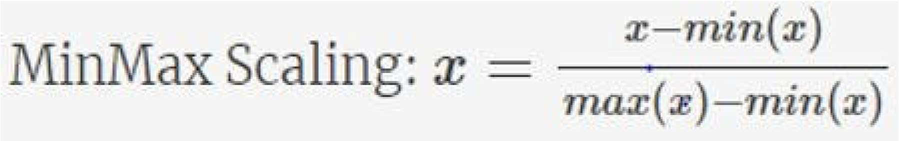
Why?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence results in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

*It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.*

Normalization/Min-Max Scaling:

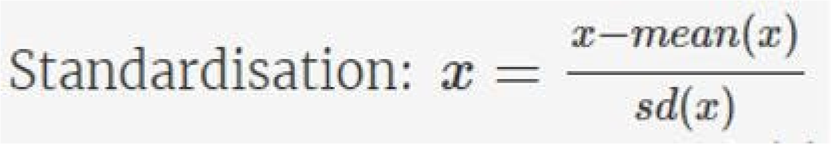
- *It brings all of the data in the range of 0 and 1.* ***sklearn.preprocessing.MinMaxScaler*** *helps to implement normalization in python.*



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean* ***(μ)*** *zero and standard deviation one (σ).*



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- ***sklearn.preprocessing.scale*** *helps to implement standardization in python.*
- *One disadvantage of normalization over standardization is that it* ***loses*** *some information in the data, especially about* ***outliers****.*

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

A VIF value is considered infinite when there is perfect multicollinearity between variables in a regression model, meaning one variable can be perfectly predicted by a linear combination of other variables, resulting in a correlation coefficient of exactly 1 and causing the VIF calculation to become undefined due to division by zero in the formula; essentially, it indicates that one variable is completely redundant with another in the model.

Key points about infinite VIF:

- **Perfect correlation:**

When two variables are perfectly correlated (R-squared = 1), the VIF becomes infinite because the denominator in the calculation approaches zero.

- **Interpretation:**

An infinite VIF signifies a serious issue in your regression model where one variable is essentially a linear combination of other variables, making it impossible to accurately estimate the individual coefficients.

Steps to deal with infinite VIF:

- **Remove redundant variables:**

Identify the variables with perfect correlation and remove one of them from the model to eliminate the redundancy.

- **Check data entry errors:**

Double-check your data for potential errors that might be causing unexpected perfect correlations.

- **Transform variables:**

Consider transforming variables to potentially reduce the level of correlation.

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

A Q-Q plot, which stands for "Quantile-Quantile plot," is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, most commonly the normal distribution, by plotting the quantiles of the data against the quantiles of the theoretical distribution; in linear regression, it is particularly important for visually assessing whether the residuals (the difference between predicted and actual values) follow a normal distribution, which is a key assumption for many statistical inferences made from the model.

Key points about Q-Q plots in linear regression:

- **Visualizing normality:**

If the points on a Q-Q plot fall roughly along a straight line, it indicates that the data is likely normally distributed. Deviations from this line suggest non-normality, which can affect the reliability of hypothesis tests based on the regression model.

- **Identifying outliers:**

Outliers can appear as points significantly far from the line on a Q-Q plot, allowing for further investigation into potential data issues.

- **Checking model assumptions:**

Since linear regression often relies on the assumption of normally distributed residuals, a Q-Q plot is a critical diagnostic tool to assess if this assumption is met.

How to interpret a Q-Q plot:

- **Straight line:**

If the points closely follow a straight line, the data is considered to be approximately normally distributed.

- **Curved pattern:**

If the points curve away from the line, especially at the tails, it suggests the data is not normally distributed and may be skewed.

- **Large deviations from the line:**

Significant outliers will appear as points far away from the line on a Q-Q plot.

Important considerations:

- **Subjectivity:**

While a Q-Q plot provides a visual assessment of normality, it is not a definitive test and should be interpreted with caution.

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?

Seasons:

- Spring has the lowest median and the narrowest range in user counts, while the other three seasons have similar medians and ranges.

Year:

- There were fewer users in 2018 compared to 2019, showing a significant increase in user count over the year.

Days of the Week:

- The user count remains consistent across different days of the week and working days, indicating no significant impact from the day of the week.

Weather Conditions:

- Extreme weather conditions lead to a notable decrease in user counts.

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

When creating dummy variables for a categorical feature, using drop_first=True is crucial to avoid the "dummy variable trap" - a situation where the dummy variables are perfectly collinear with each other, causing issues in regression model due to multicollinearity;

Essentially, by dropping the first level of the categorical variable, you ensure that one level acts as the reference category, allowing for meaningful interpretation of the coefficients.

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

Registered users has the highest positive correlation with the target variable.

Windspeed has highest negative correlation with the target variable .

Apart from that Casual users, Temperature and Atemperature variables also have a good corelation with the target variables.

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

VIF values – The VIF values of all the predictor values are less than 5 which indicate not multicollinearity.

Residual Analysis- The residual analysis of predicted vs actual values are randomly scattered around the horizontal line (y = 0).

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

-->Variables which define the target variable most are

1.Year— With each passing year, the app’s reach has expanded, attracting more customers and leading to an increase in the total count of bike users. This indicates positive growth in user adoption over time.

2.Holiday- Holidays are negatively correlated with bike usage, as many people tend to stay home. This results in a lower count of bike users on holidays compared to regular days.

3.Feeling Temperature – Pleasant weather conditions significantly impact bike usage. Higher atemp values, indicating comfortable temperatures, are associated with an increase in the number of bike users. Weather plays a critical role in influencing the choice to bike.

4.Casual Users – The number of casual users has a direct and substantial impact on the total count of users. Casual user counts can vary significantly day-to-day, contributing to fluctuations in the total bike user count.