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? (3 marks)

The categorical feature variables in the study include 'season,' 'workingday,' 'weathersit,' 'weekday,' 'yr,' 'mnth,' and 'holiday.' Their impact on the dependent variable "cnt," which represents the count of bike rentals, is as follows:

Season

The dataset is well-balanced across the four seasons: Fall, Summer, Spring, and Winter. Fall records the highest number of bike rentals, indicating its popularity among users for biking. Conversely, Spring sees the lowest rental numbers, suggesting a potential area for improvement. Summer ranks as the second most popular season for bike rentals, closely followed by Winter. The data implies that Fall and Summer are peak seasons for bike rentals, presenting opportunities to further enhance rental numbers.

Working Day

This variable differentiates between weekdays and weekends/holidays. Interestingly, bike rentals are slightly higher on working days, suggesting a propensity for using bikes for commuting. Rentals on non-working days show more variability, pointing towards leisure or recreational use on these days.

Weather Situation

Clear and partly cloudy conditions see the highest bike rental activity, which is expected given the favorable conditions for outdoor activities. A notable, albeit smaller, number of rentals also occur on days with light rain/snow, primarily from registered users, indicating a committed user base. Heavy rain/snow conditions see no rentals, as would be expected due to safety and comfort concerns.

Weekday

Analysis of rentals across different days of the week reveals a uniform median count, indicating no significant pattern based on the day of the week. However, a positive correlation with the rental count suggests a consistent use of rentals for daily commuting.

Year

Data from two consecutive years shows an increase in bike rentals from 2018 to 2019, highlighting a growing trend in bike usage over time.

Holiday

Non-holiday days experience higher bike rentals compared to holidays. This observation could correlate with the use of bikes for commuting to work on regular days.

Month

Bike rentals peak during the middle months of the year, from June to October, where monthly rental counts surpass 5,000. Additionally, the expanding gap between the median and the 75th percentile in these months suggests a growing demand for bike rentals, potentially due to favorable weather conditions and increased leisure activities.

These insights provide a nuanced understanding of the factors influencing bike rentals and can guide strategies to boost rental numbers, particularly by leveraging seasonal trends and understanding user behavior in relation to the workweek and weather conditions.

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

While creating dummy variables we use drop_first=True as we can have a based or reference category instead of creating a new category. The reason for this also is to avoid the multicollinearity getting added into the model if all dummy variables are included. The reference category can be deduced from a row where all the dummy variables have zero values.

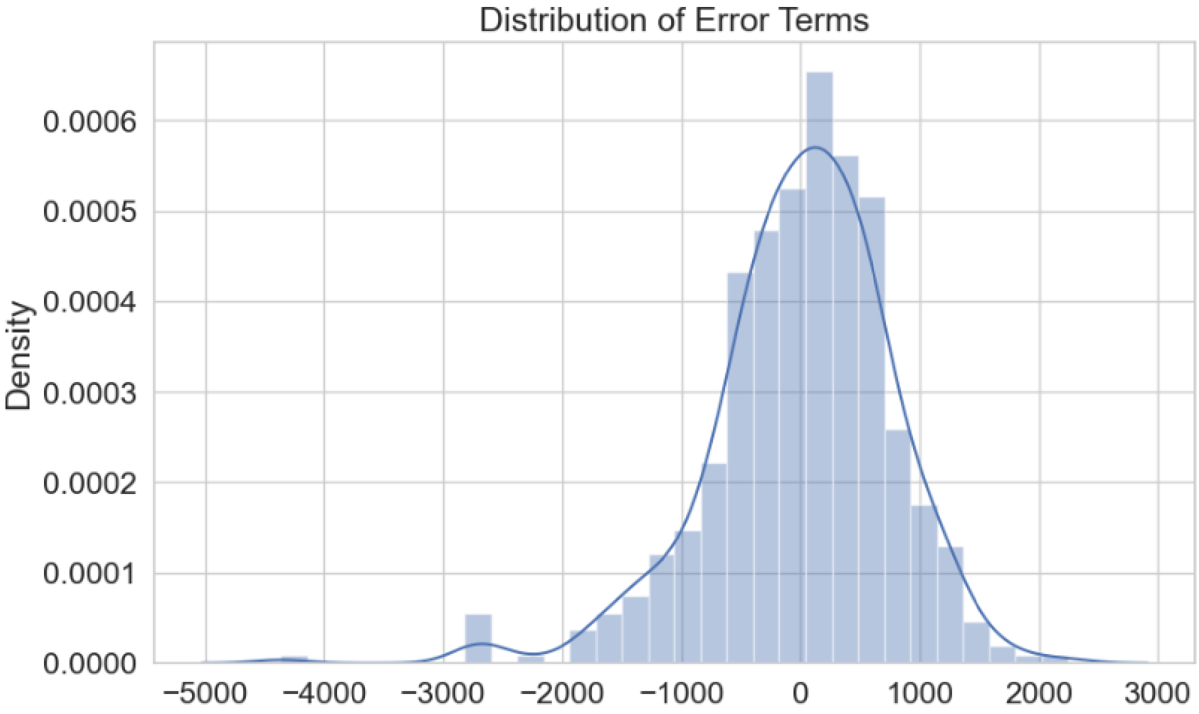
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

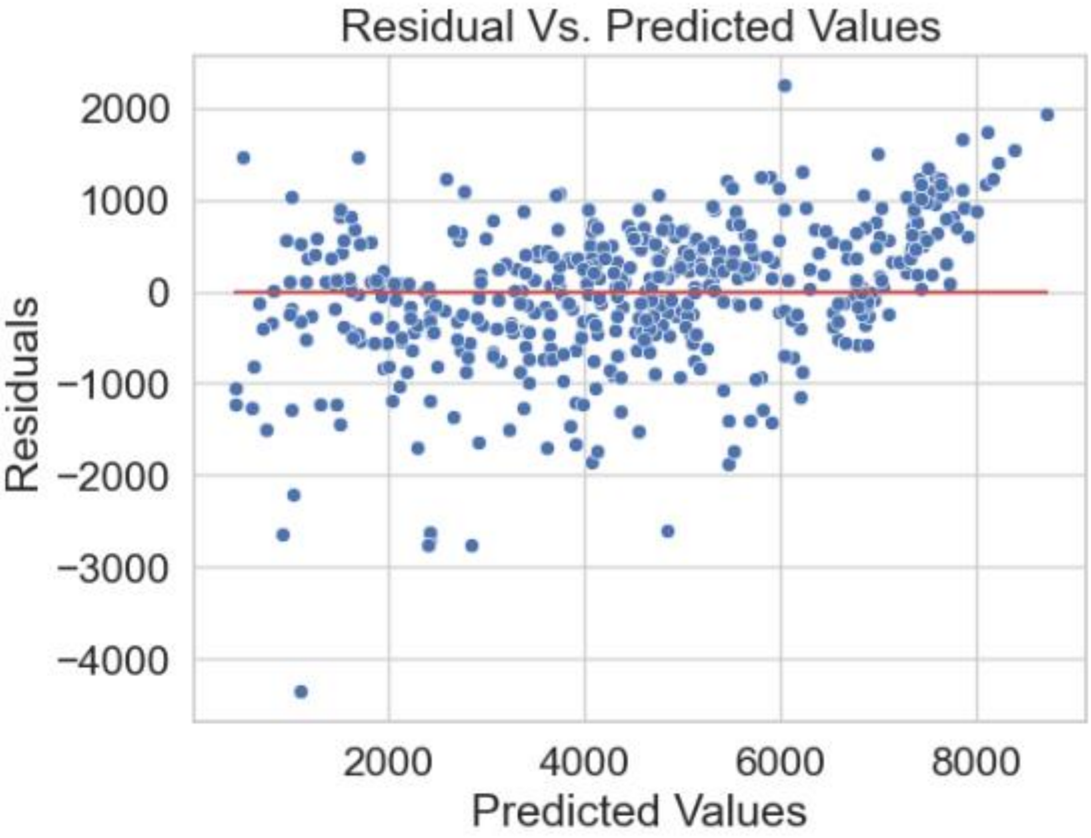
- temp” variable has the highest correlation with target variable i.e. 0.63.
- atemp, casual and registered have been dropped as they are represented by other variables in the dataset in the EDA Steps.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

1. **Residual Analysis – Residuals should be normally distributed**

a. Histogram and distribution plot and Regplot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.

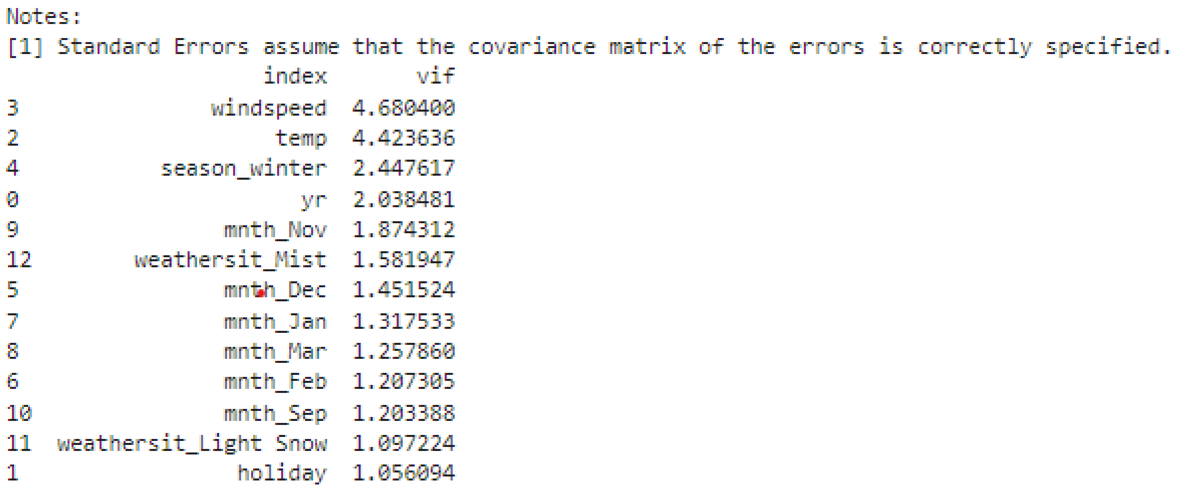




Residual errors follow a normal distribution with mean=0
Variance of Errors doesnot follow any trends
Residual errors are independent of each other since the Predicted values vs Residuals plot doesn't show any trend.

2. **No Multicollinearity** –

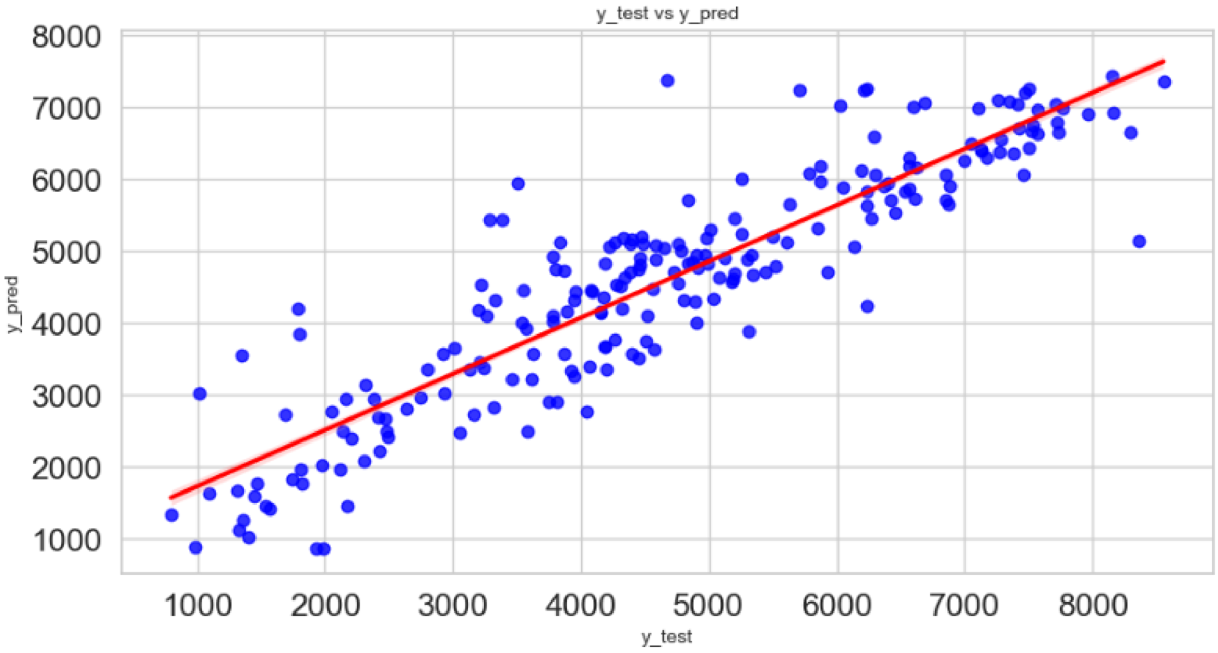
As we can see **VIFs of all feature variables below 5**, so there is no multicollinearity.



```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                   index       vif
3              windspeed  4.680400
2                   temp  4.423636
4          season_winter  2.447617
0                     yr  2.038481
9               mnth_Nov  1.874312
12       weathersit_Mist  1.581947
5               mnth_Dec  1.451524
7               mnth_Jan  1.317533
8               mnth_Mar  1.257860
6               mnth_Feb  1.207305
10              mnth_Sep  1.203388
11  weathersit_Light Snow  1.097224
1                holiday  1.056094
```

3. Linear relationship between target and feature variables

b. Here we can see that for numerical feature cnt increases with increase in value



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top predictor variables:

- temperature(temp): coefficient of 2907.59, indicates a unit increase in temperature variable will increase the bike hire number by 2907.59.
- year: coefficient of 2017.83, indicates a unit increase in year will increase the bike hire number by 2017.83.
- weather Situation(weathersit): coefficient of -2193.78, indicates if the weather situation gets bad by one level up then bike hire number goes down by -2193.78.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a statistical method used for predictive modeling, which establishes a linear relationship between a dependent variable (Y) and one or more independent variables (X). The core objective of linear regression is to find a "line of best fit" that predicts the value of the dependent variable based on the values of the independent variable(s). This relationship is represented by a linear equation, where the predicted values of Y for different values of X fall on this line.

Linear regression models can be categorized into two types based on the number of independent variables involved:

- **Simple Linear Regression (SLR):** Utilized when there is only one independent variable. The equation for the line of best fit in SLR is \(Y = \beta_0 + \beta_1X\), where \(\beta_0\) is the y-intercept, and \(\beta_1\) is the slope or coefficient of the independent variable \(X\).

- **Multiple Linear Regression (MLR):** Applied when there are multiple independent variables. The line equation for MLR is \(Y = \beta_0 + \beta_1X_1 + \ldots + \beta_pX_p +

\epsilon\), where \(\beta_0\) is the y-intercept, \(\beta_1, \beta_2, \ldots, \beta_p\) are the coefficients for each independent variable, and \(\epsilon\) is the error term.

When building a multiple linear regression model, it is essential to consider:

- **Overfitting:** Ensuring the model is not overly complex, which could make it perform well on training data but poorly on unseen data.

- **Multicollinearity:** Avoiding highly correlated independent variables to prevent redundancy and instability in the coefficient estimates.

- **Categorical Variables:** Utilizing dummy variables through one-hot encoding to include categorical data in the model.

- **Feature Scaling:** Implementing standardization or MinMax scaling to normalize the range of independent variables, ensuring no single feature dominates the model due to its scale.

The accuracy of the fitted line is assessed based on the residuals, which are the differences between the actual and predicted values (\(e_i = y_i - \hat{y}_i\)). The goal is to minimize the sum of squared residuals, known as the Residual Sum of Squares (RSS), through Ordinary Least Squares (OLS). The cost function, representing the RSS, guides the optimization of coefficient values to minimize prediction errors.

Model performance is primarily evaluated using:

- **R-squared (\(R^2\))**: Indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s), with 1 being perfect prediction.

- **Adjusted R-squared**: Adjusts \(R^2\) to account for the number of predictors in the model, improving reliability in the context of multiple independent variables.

- **AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion):** Criteria for model selection that balance model complexity and fit.

Optimizing the model involves minimizing the cost function through methods such as differentiation or gradient descent. The ultimate aim is to create a robust, predictive model that accurately forecasts the dependent variable using the independent variables.

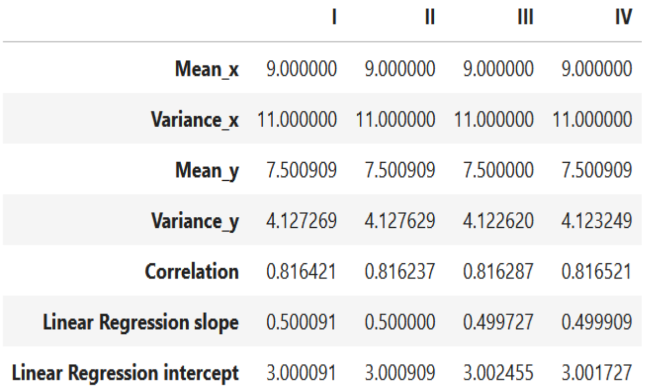
2. Explain the Anscombe's quartet in detail. (3 marks)

Statistical methods like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great to for describing the general trends and aspects of the data.

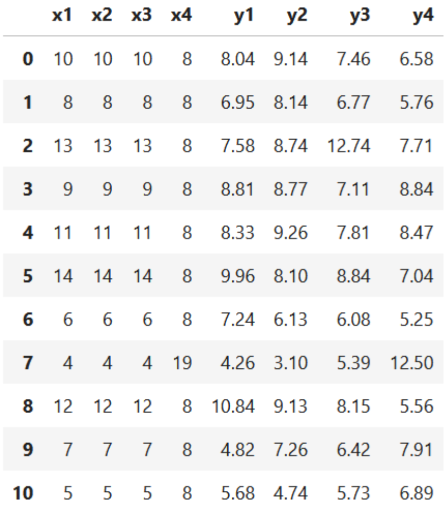
Anscombe's Quartet is a classic statistical demonstration devised by the statistician Francis Anscombe in 1973. It consists of four distinct datasets, each containing eleven (x, y) data points. What makes Anscombe's Quartet remarkable is that despite having very similar statistical properties, the datasets exhibit vastly different relationships when graphed, showcasing the importance of visualizing data alongside numerical analysis.

Lets understand this with an e.g. dataset –

The quartet consists of four different datasets, each containing 11 points, with two variables: x and y; such as x1 & y1, x2 & y2, x3 & y3, x4 & y4.

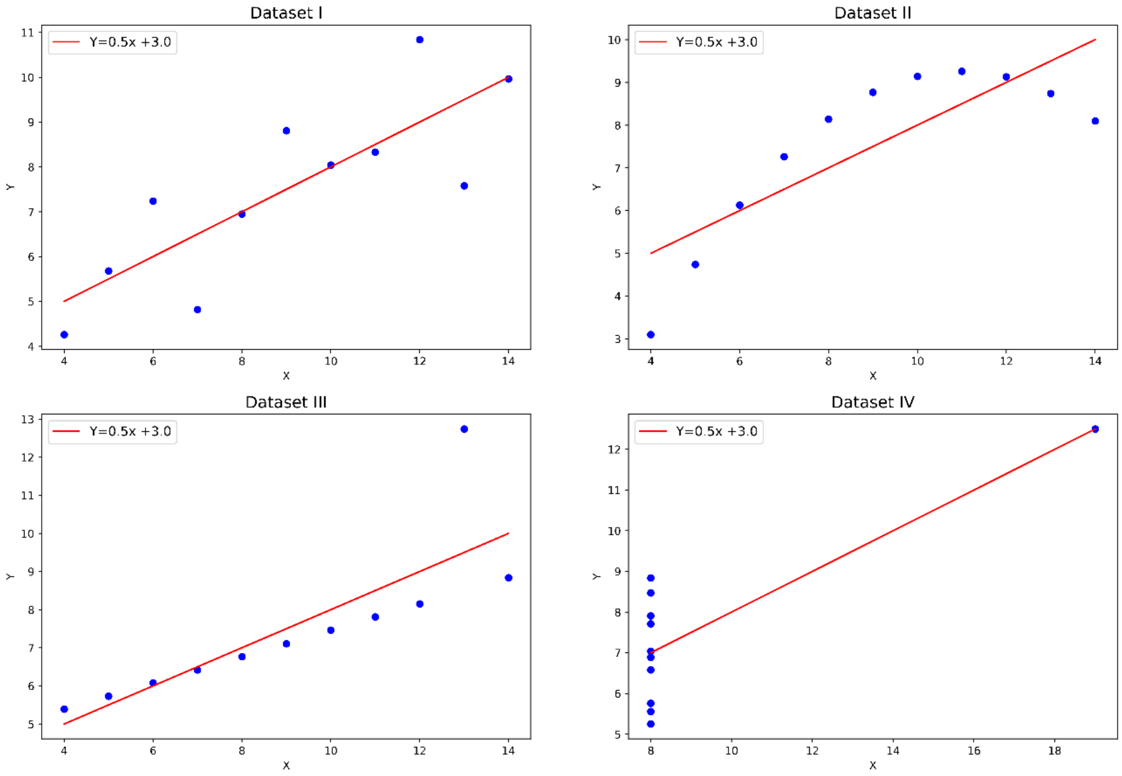


	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727



	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

- **Data-set I** — the first dataset appears to be a simple linear relationship, where y increases as x increases.
- **Data-set II** — this follows a perfectly quadratic relationship, with a clear curve. This highlights the fact that data can exhibit nonlinear patterns, and relying solely on linear regression can lead to incorrect conclusions.
- **Data-set III** — the dataset, shows a linear trend, a single outlier affects the regression line, creating a misleading representation of the data.
- **Data-set IV** — the fourth dataset adds a new layer of complexity to the situation. There is one data point that is outlier from the others and entirely contradicts the pattern, which causes the linear regression line to shift in a significant way.



Anscombe's Quartet Significance:

- **Diverse Relationships**: Despite having identical means, variances, correlation coefficients, and linear regression lines, the datasets portray drastically different relationships when graphed. This illustrates the importance of visualizing data to understand its underlying structure fully.
- **Cautionary Tale**: Anscombe's Quartet serves as a cautionary tale in statistical analysis, reminding researchers not to rely solely on summary statistics. Even when statistical properties seem similar, the actual data may be fundamentally different.
- **Exploratory Data Analysis (EDA):** Anscombe's Quartet supports the necessity of exploratory data analysis (EDA) before drawing conclusions from statistical analyses. Visualization techniques can reveal nuances and patterns that summary statistics alone might miss.

Outliers need to be treated: Based on the last 2 graphs we can see how outliers impact the overall model.

3. What is Pearson’s R? (3 marks)

The Pearson’s R (aka Pearson’s correlation coefficients) measures the strength between the different variables and the relation with each other.

The Pearon’s R returns values between -1 and 1. The interpretation of the coefficients are:

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship. r= n(Σx∗y)−(Σx)∗(Σy)√[nΣx2−(Σx)2]∗[nΣy2−(Σy)2] Where: N = the number of pairs of scores Σxy = the sum of the products of paired scores Σx = the sum of x scores Σy = the sum of y scores Σx2 = the sum of squared x scores Σy2 = the sum of squared y scores

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

From formula we can say, if the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables, i.e the independent variables are orthogonal to each other

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Working of Q-Q plots –

Quantiles: Quantiles divide a dataset into intervals of equal probability. For example, the median is the 50th percentile, dividing the dataset into two equal parts.

Comparison: In a Q-Q plot, the quantiles of the sample dataset are plotted on the horizontal axis, while the quantiles of the theoretical distribution (e.g., normal distribution) are plotted on the vertical axis.

Linearity: If the sample dataset follows the theoretical distribution closely, the points on the Q-Q plot will fall close to a diagonal line (the line of equality). Deviations from the diagonal line indicate departures from the theoretical distribution.

Types of Q-Q plots

4. **Normal Distribution**: A symmetric distribution where the Q-Q plot would show points approximately along a diagonal line if the data adheres to a normal distribution.
5. **Right-skewed Distribution**: A distribution where the Q-Q plot would display a pattern where the observed quantiles deviate from the straight line towards the upper end, indicating a longer tail on the right side.
6. **Left-skewed Distribution**: A distribution where the Q-Q plot would exhibit a pattern where the observed quantiles deviate from the straight line towards the lower end, indicating a longer tail on the left side.
7. **Under-dispersed Distribution**: A distribution where the Q-Q plot would show observed quantiles clustered more tightly around the diagonal line compared to the theoretical quantiles, suggesting lower variance.
8. **Over-dispersed Distribution**: A distribution where the Q-Q plot would display observed quantiles more spread out or deviating from the diagonal line, indicating higher variance or dispersion compared to the theoretical distribution.

Advantages of Q-Q plot -

- Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
- Easily detects departures from assumed distributions, aiding in identifying data discrepancies.
- The plot has a provision to mention the sample size as well can compare datasets of different sizes without requiring equal sample sizes.

The use and importance of a Q-Q plot in linear regression include:

9. **Normality Assumption**: In linear regression, it is often assumed that the residuals (the differences between the observed and predicted values) follow a normal distribution. A Q-Q plot of the residuals helps assess whether this assumption holds. If the points on the Q-Q plot form a roughly straight line, it suggests that the residuals are normally distributed. Deviations from the line indicate departures from normality.
10. **Model Assessment**: Q-Q plots can be used to assess the goodness-of-fit of the regression model. If the residuals follow a normal distribution, it suggests that the model adequately captures the variation in the data. However, if the residuals deviate from normality, it indicates that the model may not be appropriate for the data.
11. **Identifying Outliers**: Q-Q plots can help identify outliers or data points that do not conform to the expected distribution. Outliers may appear as points that deviate significantly from the

diagonal line on the Q-Q plot, suggesting that they may have a different distribution than the rest of the data.

12. **Model Validation**: Q-Q plots are a useful tool for validating the assumptions of linear regression models. By visually inspecting the Q-Q plot, analysts can assess whether the normality assumption holds and whether the model adequately fits the data.

Overall, Q-Q plots provide valuable insights into the distributional properties of the data and help ensure the validity and reliability of linear regression models. They are an essential tool in the diagnostic process of linear regression analysis.