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 summary of the regression model is as follows: -

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.822
Model:                            OLS   Adj. R-squared:                  0.819
Method:                 Least Squares   F-statistic:                     256.4
Date:                Sat, 21 Dec 2024   Prob (F-statistic):           4.93e-181
Time:                        18:17:09   Log-Likelihood:                 478.53
No. Observations:                 510   AIC:                            -937.1
Df Residuals:                     500   BIC:                            -894.7
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0804      0.024      3.393      0.001       0.034       0.127
yr             0.2345      0.009     27.427      0.000       0.218       0.251
workingday     0.0567      0.012      4.884      0.000       0.034       0.080
temp           0.4305      0.029     14.668      0.000       0.373       0.488
Spring        -0.1164      0.016     -7.433      0.000      -0.147      -0.086
Winter         0.0571      0.013      4.490      0.000       0.032       0.082
Showers       -0.2273      0.026     -8.766      0.000      -0.278      -0.176
Sunny          0.0788      0.009      8.659      0.000       0.061       0.097
month_9        0.0732      0.016      4.579      0.000       0.042       0.105
Monday         0.0642      0.015      4.292      0.000       0.035       0.094
==============================================================================
Omnibus:                       76.518   Durbin-Watson:                   2.044
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              216.563
Skew:                          -0.724   Prob(JB):                     9.42e-48
Kurtosis:                       5.845   Cond. No.                         15.0
==============================================================================
```

VIF values of predictors are as follows: -

```
2        temp   5.77
1   workingday   4.22
6        Sunny   2.79
0           yr   2.05
8       Monday   1.72
3       Spring   1.60
4       Winter   1.40
7      month_9   1.15
5      Showers   1.10
```

Overall Model Fit:
1) R-squared (0.822): Indicates that 82.2% of the variance in the dependent variable (cnt) is explained by the model.
2) Adjusted R-squared (0.819): Slightly lower than R-squared, accounting for the number of predictors in the model, indicating a good fit

Categorical Variables and Interpretation

Categorical Variables	Coefficient, p-value	Interpretation
Spring	-0.1164, 0.000	Spring has a negative impact on bike demand (cnt). When compared to the reference season, bike demand decreases in Spring.
Winter	0.0571, 0.000	Winter positively impacts bike demand. Compared to the reference season, there is an increase in bike demand during Winter
Showers	-0.2273, 0.000	Showers significantly reduce bike demand. On days with showers, bike demand decreases.
Sunny	0.0788, 0.000	Sunny weather positively influences bike demand, leading to an increase in bike rentals.
month_9	0.0732, 0.000	September (month_9) sees an increase in bike demand compared to the reference month.
Monday	0.0642, 0.000	Mondays see a significant increase in bike demand compared to the reference day of the week.

Inference:
The categorical variables in the dataset significantly impact bike demand in various ways:

1. **Seasonality**: Spring reduces, while Winter increases bike demand.
2. **Weather**: Showers decrease, while sunny weather increases bike demand.
3. **Month**: September shows increased demand.
4. **Day of the Week**: Mondays see higher 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)

Encoding refers to the process of converting categorical data into a numerical format that can be easily processed and analyzed by machine learning models. There are various types of encoding such as Binary, label or One-Hot encoding,

Every encoding technique has pros and cons. In this linear regression model building, One-Hot encoding is used. The python notebook documents the reason behind adopting this technique. On this technique, while *creating dummy variables 'drop_first=True' is used to avoid the dummy variable trap*.

To summarize - The dummy variable trap occurs when there is perfect multicollinearity in the regression model, meaning one predictor can be perfectly predicted from the others. This situation leads to redundancy in the variables and can cause issues in model estimation

Example - For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', a dummy variable would be created as follows –

Relationship Status	Single	In a relationship	Married
Single	1	0	0

In a relationship	0	1	0
Married	0	0	1

With 'drop_first=True', the same encoding is presented as

| Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

The reason it is important is because
1) **Avoid Multicollinearity**: Reduces the risk of perfect multicollinearity by removing redundancy.
2) **Model Stability**: Makes the regression model more stable and reliable.
3) **Interpretability**: Simplifies interpretation by comparing each category against the reference category (the dropped category).

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)

Variable 'atemp' has the highest correlation with the target variable ('cnt') with correlation of 0.65
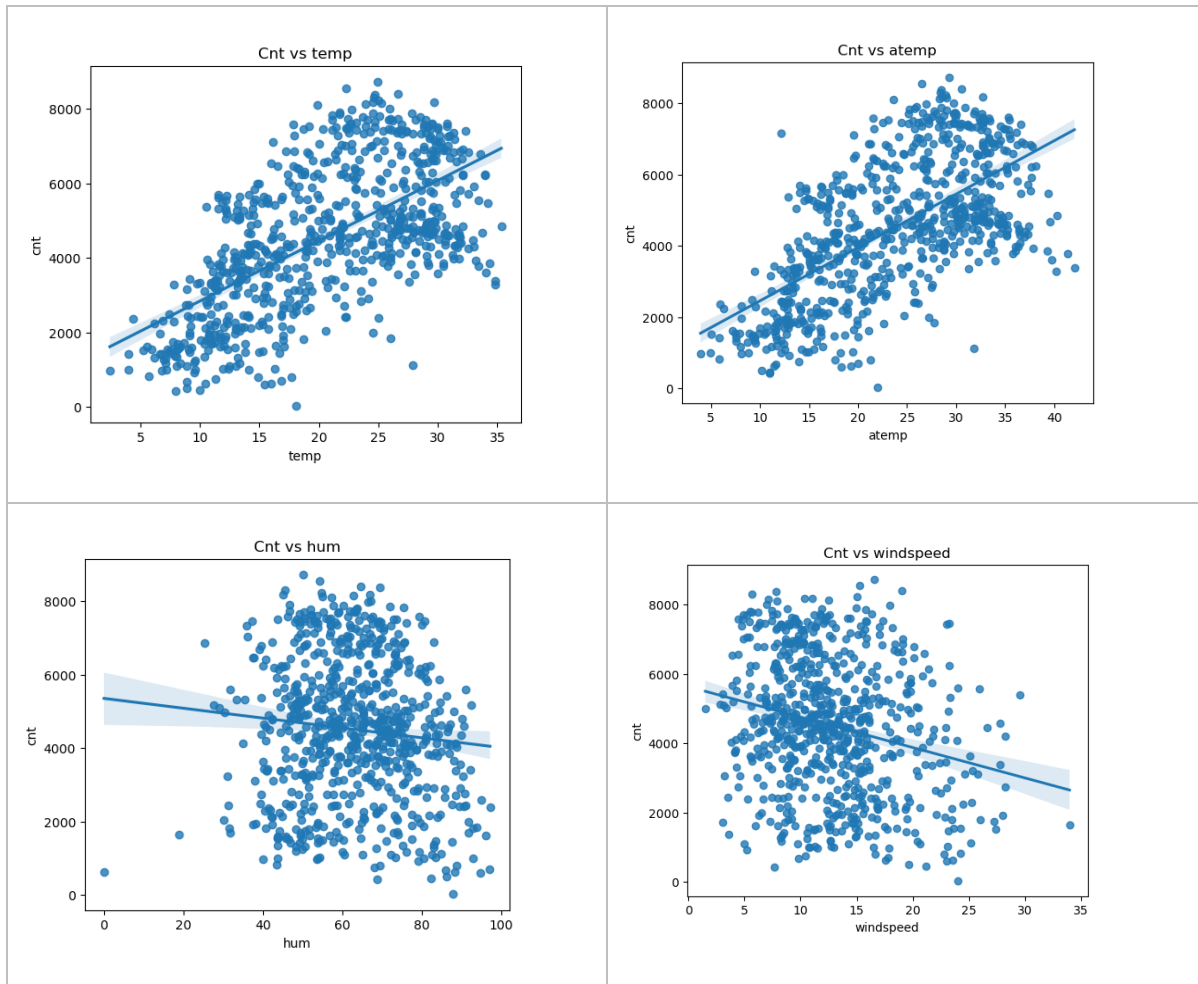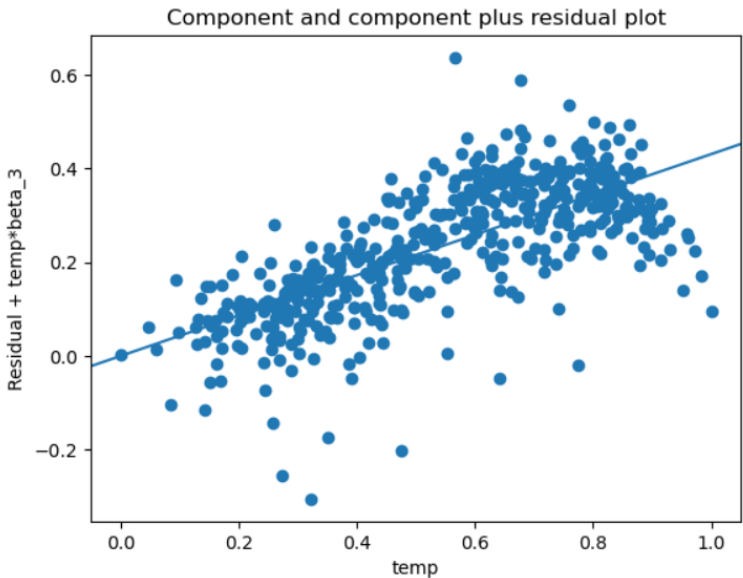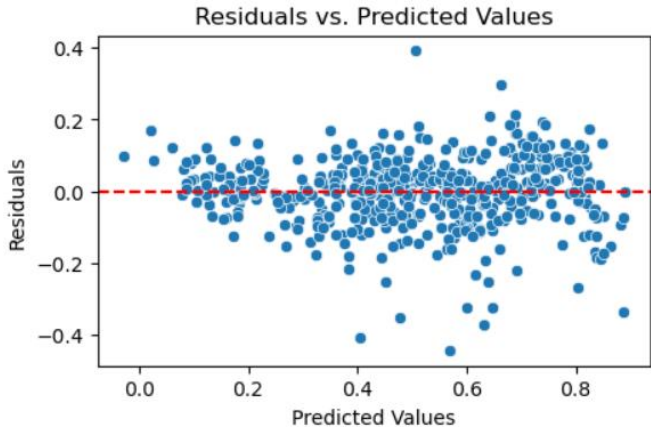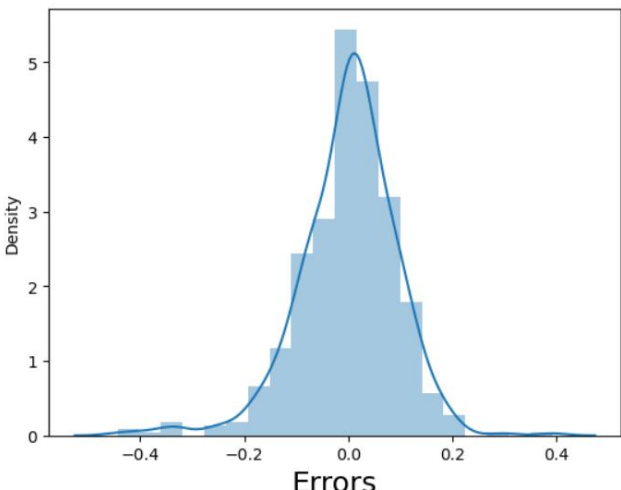
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Following validations are conducted for assumption of linear regression

Assumption	Validation steps
Linearity - The relationship between the independent variables and the dependent variable is linear	Conditional Component-plus-Residual (CCPR) plots helps determine if the relationship between each predictor and the response variable is approximately linear. One of the predictors is 'temp' Below is CCPR plot for the same. Since the relationship is linear, the residuals are randomly scattered around zero.
Homoscedasticity - The residuals have constant variance (homoscedasticity)	

Normality of Residuals	Error Terms

Multicollinearity	VIF values below 5 indicate low multicollinearity.

2	temp	5.77
1	workingday	4.22
6	Sunny	2.79
0	yr	2.05
8	Monday	1.72
3	Spring	1.60
4	Winter	1.40
7	month_9	1.15
5	Showers	1.10

Except for temp, VIF values are blow 5. Temp variable is not removed because there is a high correlation between temp and bike travelling

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 3 features were chosen based on the magnitude of their coefficients and the absolute value of their t-values, which reflect their statistical significance and impact on the dependent variable (cnt). Below are the top 3 features
  1) Year (yr):
  2) Temperature (temp):
  3) Showers (Showers):

Feature	Coefficient	t-value	p-value	Interpretation
Year (yr)	0.2345	27.427	0.000	Indicating a strong positive relationship with bike demand
Temperature (temp)	0.4305	14.668	0.000	Higher temperatures are associated with increased bike demand, as more people prefer to ride bikes in warmer weather
Showers	-0.2273	-8.766	0.000	Negative impact on bike demand. Rainy weather conditions lead to a decrease in bike rentals

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)

Linear regression falls under predictive analytics. It is one of the most widely used model. Primarily it is used for linear relations.

Linear regression attempts to find the best-fitting linear relationship between the dependent variable 'y' and one or more independent variables 'X' with the help of supervised learning.

Linear regression is broadly categorized into 2 categories

| Simple Linear Regression | Involves one independent variable. |
| Multiple Linear Regression | Involves two or more independent variables. |

Linear regression model can be represented as

credit - link

Below are the key high-level steps to be carried out for building a model:

Sr.	Steps
1	Importing and understanding data
2	Visualizing the Data - numerical and categorical variables
3	Data preparation - includes choice between Binary Encoding and One-Hot encoding, inferences
4	Splitting the data into training and testing set
5	Rescaling the features
6	Building a linear model - includes comparison between various models
7	Residual Analysis of the train data
8	Making predictions using final model
9	Model Evaluation

Visualizing the Data:
Various graphs are used to represent the data information.

Data preparation:
This is an important step. It aims to increase the accuracy of the data by removing null values or imputing values where ever possible. Removing outliers based on domain context is also an important step.
As a part of answer to question (2), encoding is already explained.

Splitting the data into training and testing set:
In supervised learning, it is important to split the data into training and test. The model building happens in training data and then validated on test data

Rescaling Features:
Rescaling features, also known as normalization or standardization, is a crucial preprocessing step in machine learning. It ensures that all features contribute equally to the model by transforming the data to a common scale. Here are the key methods

Methods	Description
Min-Max Scaling	Transforms the data to a range between 0 and 1.
Standardization (Z-score Normalization):	Centers the data around 0 with a standard deviation of 1.

Rescaling features helps improving convergence and reduces bias

Building a linear model:
This step involves feature selection and estimating the Coefficients
1. Manual feature selection - A very tedious task in order to select the correct set of features.
2. Automated feature selection - The three-step process is involved.
    a. Select top 'n' features (Recursive feature elimination)
    b. Forward/backward/Stepwise selection based on AIC
    c. Regularization

Finding a balance between the two - A balance of both manual and automatic feature selection is required to attain the features.

Ordinary Least Squares (OLS): The most common method to estimate the coefficients is by minimizing the sum of the squared differences between the observed and predicted values. This method is known as Ordinary Least Squares (OLS).

Recursive feature elimination: Based on the idea of repeatedly constructing a model (for example, an SVM or a regression model) and choosing either the best or worst performing feature (for example, based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.

Typically, model building starts with RFE and followed by evaluating / removing irrelevant features based on model evaluating parameters, fine tuning in the context of domain.

Model prediction and evaluating:
It is an important step in building the confidence and stability of the model. The various metrics used are:
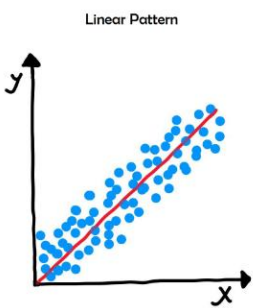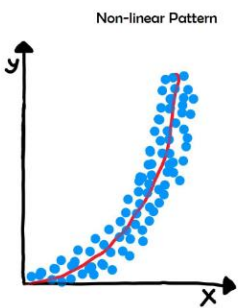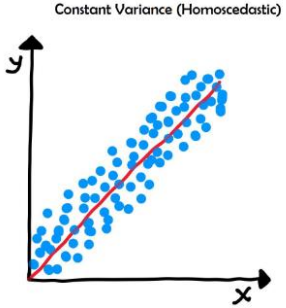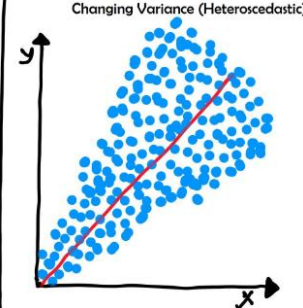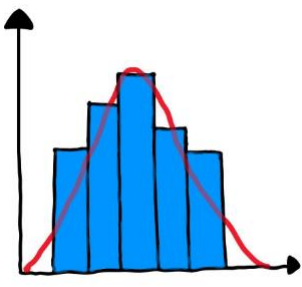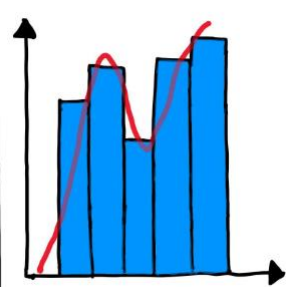
Metric	Interpretation

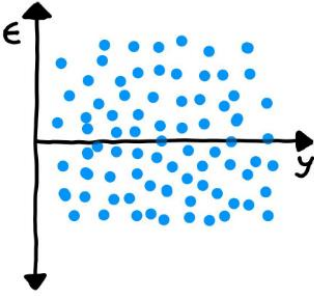| R-squared	Indicates the proportion of the variance in the dependent variable explained by the independent variables
Adj. R-squared	value for the number of predictors, providing a more accurate measure in multiple regression
P-values	Indicate the significance of individual predictors
F-statistic and Prob (F-statistic)	The F-statistic assesses the overall significance of the regression model, indicating whether there is a meaningful relationship between the dependent and independent variables, while the Prob(F-statistic) provides the p-value, showing the probability that the observed F-statistic would occur under the null hypothesis of no relationship. A low p-value (typically < 0.05) suggests that the model is statistically significant

Once the final model is selected based on the metrics mentioned above, run the model on test data. Consider rescaling the feature that was performed on the training data.

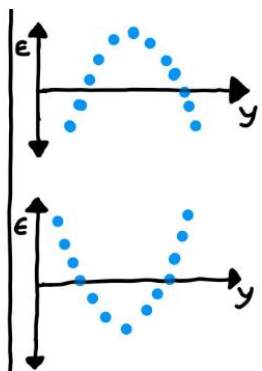**Assumptions of Linear Regression that should be verified on the final model**:
Residual Analysis of the train data is a part of this step. For the model to be valid, certain assumptions must be met:

Linearity	The relationship between the independent and dependent variables is linear
Homoscedasticity	The variance of the residuals (errors) is constant across all levels of the independent variables
Normality of Residuals	The residuals should be approximately normally distributed

Multicollinearity	The independent variables should not be highly correlated with each other. Defined by variance inflation factor (vif)
Independence	The residuals should be independent of each other (no autocorrelation). No visible pattern – Error terms independent Visible pattern – Error terms dependent

---

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)

In the field of Data science, the first step is essentially to understand the data. While statics results (a.k.a numbers) help understanding the data, it is important to plot the data before analyzing it. Anscombe's Quartet is the modal example to demonstrate the importance of data visualization.

Anscombe's quartet emphasizes that relying solely on summary statistics can be misleading. Visualizing data is crucial for detecting patterns, relationships, and anomalies that summary statistics might obscure. Always graph your data to understand its true nature.

Anscombe's quartet is a set of four datasets. These datasets are specifically designed to demonstrate the importance of visualizing data before performing statistical analyses. While these four datasets have nearly identical statistical properties, they reveal very different distributions and patterns when graphed.
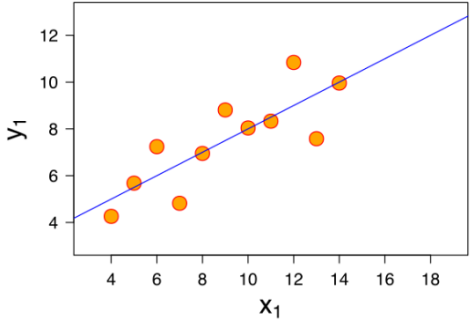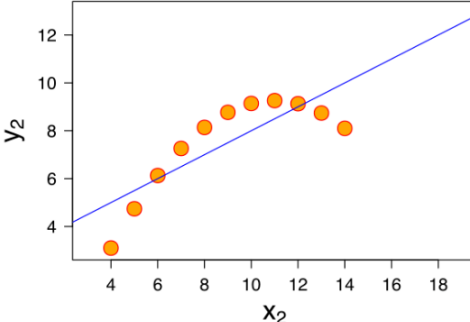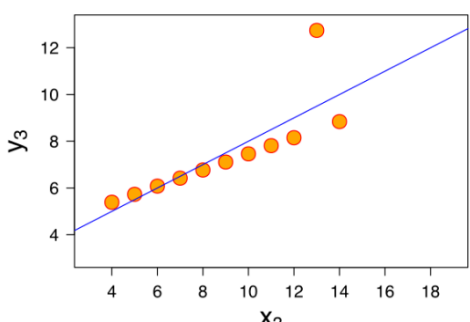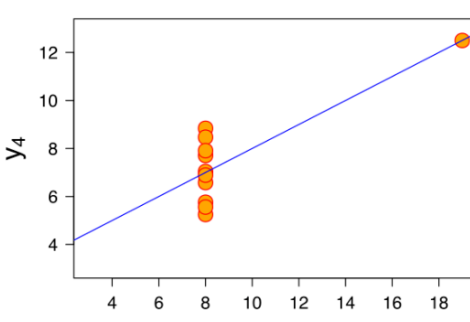
Here's a detailed explanation:

Statistical Properties: All four datasets in Anscombe's quartet share the following characteristics:
1) Mean of x: The average of the x-values is the same across all datasets.
2) Mean of y: The average of the y-values is the same across all datasets.
3) Variance of x and y: The variances of the x and y values are equal.
4) Correlation: The correlation between x and y is identical for all datasets.
5) Linear regression line: All datasets produce the same linear regression equation.
6) R-squared: The coefficient of determination (R-squared) is the same for all datasets, indicating a similar goodness-of-fit

Despite these similarities, the visual patterns of the datasets are strikingly different. Here are the four datasets in Anscombe's quartet.

Dataset	Characteristics	Graph (credit – wiki)

Dataset I	Appears as a typical linear relationship when plotted	
Dataset II	Contains a clear nonlinear relationship.	
Dataset III	Exhibits a linear relationship with an outlier	
Dataset IV	Consists of vertical clustering and a single influential point.	

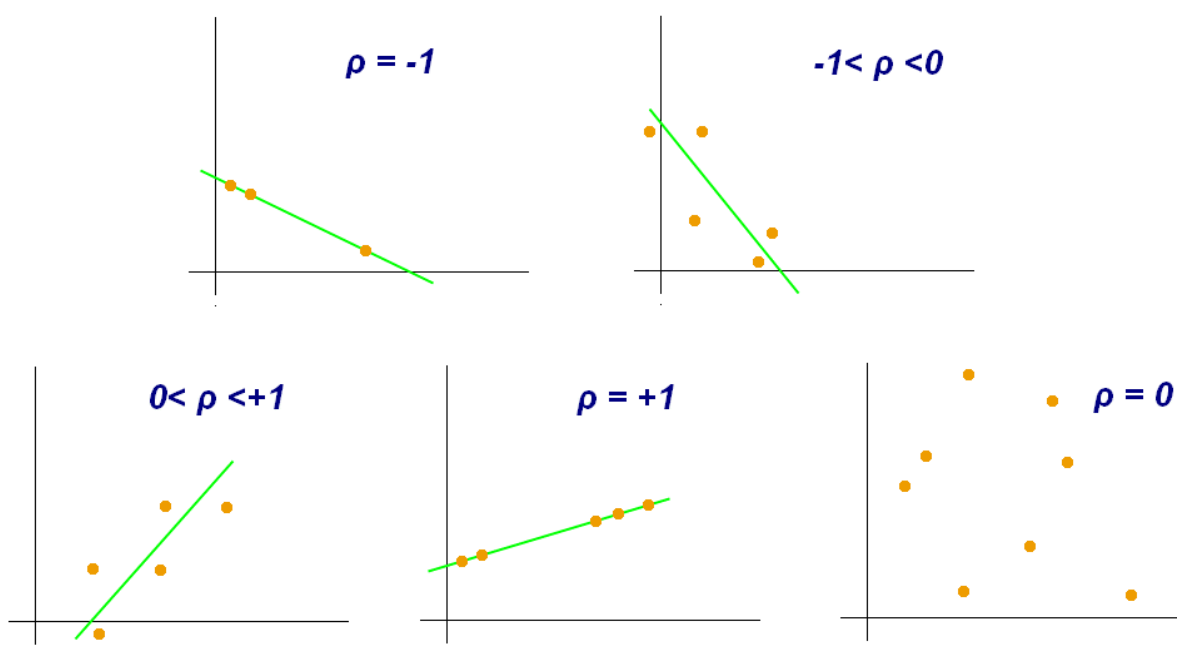**Question 8.** What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, or the Pearson correlation coefficient (r), quantifies linear relationship between two continuous variables.
- It is the most common way of measuring a linear correlation.
- It measures the strength and direction of the relationship between two variables.
- It ranges from -1 to 1.

Pearson correlation coefficient (r)	Correlation type	Interpretation
$0 < r \leq 1$	Positive correlation	When one variable changes, the other variable changes in the same direction.
0	No correlation	no relationship between the variables
$0 > r \geq -1$	Negative correlation	When one variable changes, the other variable changes in the opposite direction.

credit - wiki

The closer the value of Pearson's R is to -1 or +1, the stronger the linear relationship

---

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)

Scaling is the process of transforming features to a common scale to ensure that no single feature dominates the model due to its magnitude.

The range of values of raw data varies widely. The model might think the weight is way more

important just because the numbers are bigger.

Example: Two features: one measuring height in centimeters (like 150 cm) and another measuring weight in kilograms (like 70 kg)

Scaling helps make sure every feature contributes equally to the model. Also, gradient descent converges much faster with feature scaling than without it. In this way, it enhances the performance of the model.

There are two types of scaling:
1) **Normalization**
   Also called as min-max scaling or min-max normalization. This technique transforms data to a range between 0 and 1. The formula representation is

   $$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

   max(x), min(x) - maximum and the minimum values of the feature respectively.
   Mostly useful for algorithms sensitive to feature magnitude.

2) **Standardization**
   This technique centers data around 0 with a standard deviation of 1. The formula representation is

   $$x' = \frac{x - \bar{x}}{\sigma}$$

   σ is the standard deviation and x̄ is the mean
   Ideal for algorithms assuming normally distributed data.

The difference between normalization and standardization is that while normalization brings all the data points in a range between O and 1, standardization replaces the values with their Z scores.

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)

Addressing multicollinearity is one of the important steps in machine learning. To effectively detect multicollinearity, the Variance Inflation Factor (VIF) is widely used. A high VIF means feature variable A is being well explained by the other feature variables.
The formula representation of VIF is

$$VIF = \frac{1}{1 - R^2}$$

where R^2 is coefficient of determination of the regression of one predictor on the others. Typically,
$R^2 = \frac{RSS}{TSS}$

RSS = Residual sum of squares and TSS = Total sum of squares
R^2 of 1 means all points are on the best fit line

If $R^2 = 1$ (indicating perfect multicollinearity), the denominator becomes zero, causing VIF to be infinite. An infinite VIF means that the regression coefficient for that predictor cannot be estimated uniquely because of the exact linear relationship with other predictors.

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)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, like the normal distribution. It is a plot of two quantiles against each other. E.g., median is a quantile where 50% of the data fall below and above it.

It identifies deviations from normality, such as skewness or heavy tails, which might affect the model's performance

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

statsmodels.api provide qqplot to plot Q-Q graph for single and two different data sets

```
sm.qqplot((y_train - y_train_pred), fit=True, line='45')
plt.show()
```
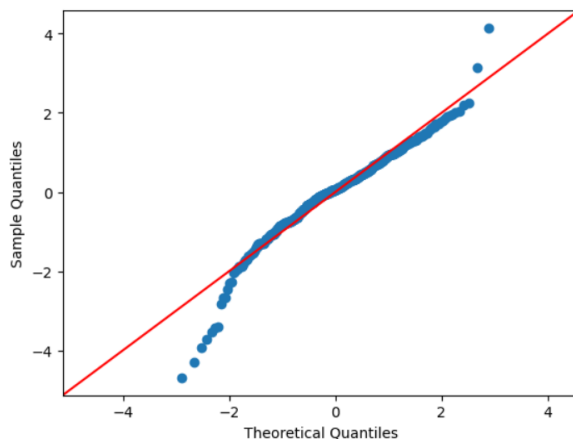

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions
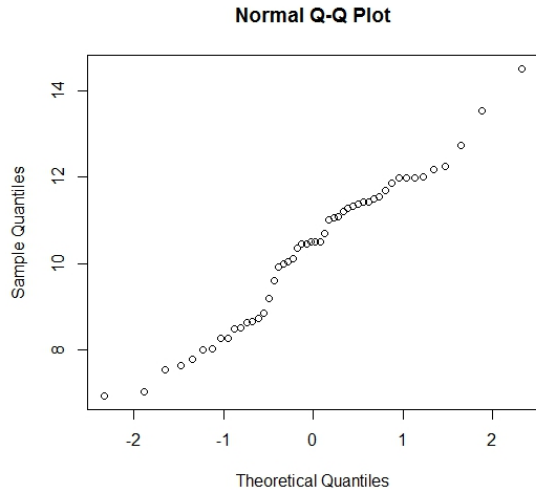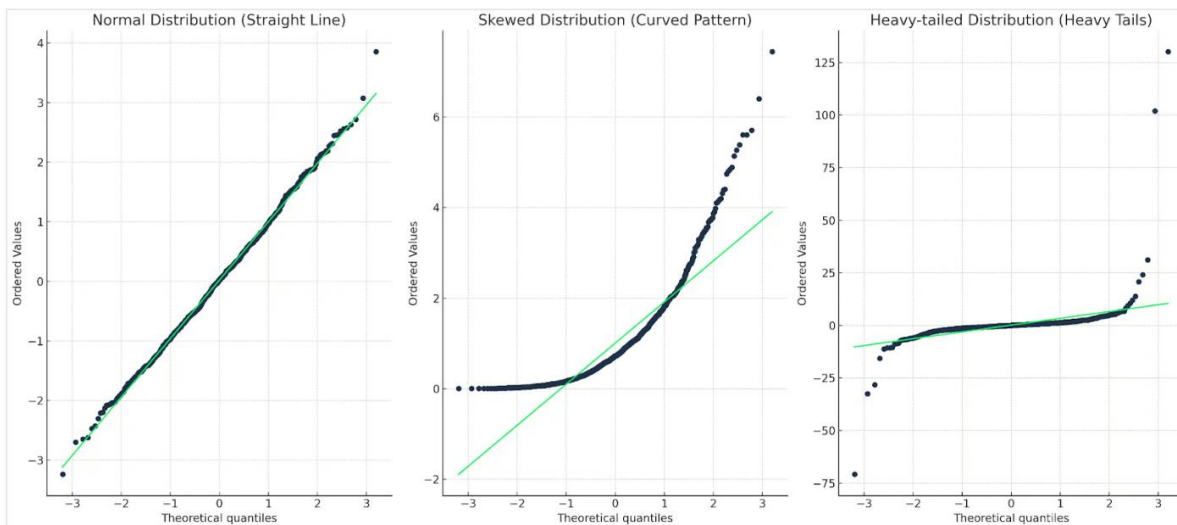
Normal Q-Q Plot

In linear regression, one key assumption is that the residuals (errors) are normally distributed. The Q-Q plot helps check this. If the points lie on a straight line, it means the residuals are normally distributed.

There are about three things need to check:
- Straight Line: Data aligns well with the theoretical distribution.
- Curved Patterns: Indicate skewed data or non-normal distributions.
- Heavy or Light Tails: If points deviate at the ends, the data might have heavier or lighter tails than expected.

Let's show with an example for each:

Credit - link