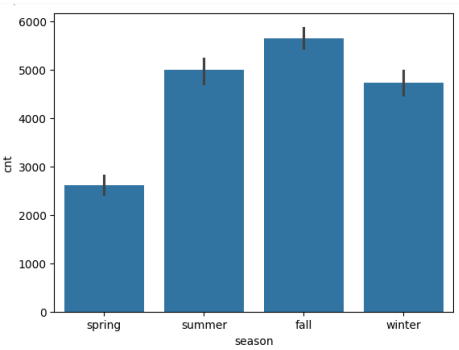
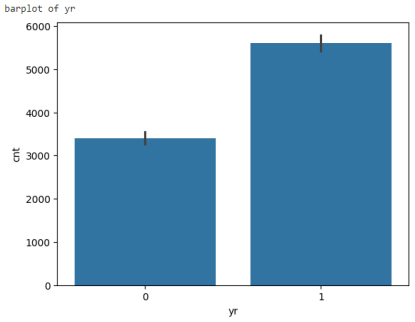
Assignment-based Subjective Questions

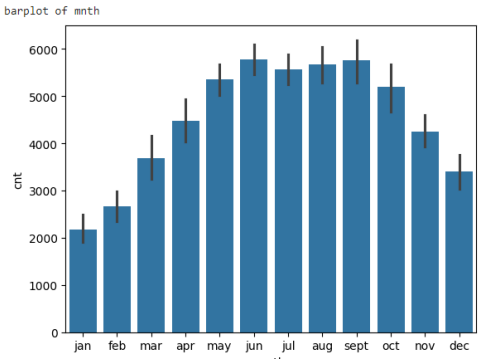
1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable?**

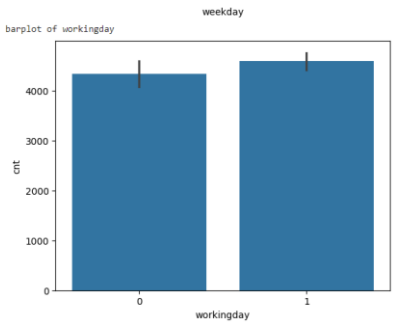
I did an analysis on categorical columns using barplots. Below are a few points what I observed from the plots:

- ❖ Most of the bookings have been done during the months of May, June, July, August, September, and October.
- ❖ Fall season seems to have attracted more bookings. In each season, the booking count has increased drastically from 2018 to 2019.
- ❖ Clear weather attracted more bookings, which seems obvious.
- ❖ Thursday, Friday, Saturday, and Sunday have more bookings compared to the start of the week.
- ❖ When it's not a holiday, bookings seem to be fewer, which is reasonable as on holidays, people may want to spend time at home and enjoy with family.
- ❖ Bookings seem to be almost equal on working days and non-working days.
- ❖ 2019 attracted more bookings compared to the previous year, which shows good progress in terms of business.









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

Using **drop_first = True** is crucial because it helps minimize the additional column generated during the creation of dummy variables. This reduces the correlations among the dummy variables.

For example, if you have a column for **seasons** with values like Spring, Summer, Fall, and Winter, you'll get four dummy columns: one for each season. However, this can cause issues with multicollinearity, which can affect your model's performance.

Syntax:

drop_first: bool, default False

Specifies whether to produce k-1 dummy variables from k categorical levels by excluding the first level.

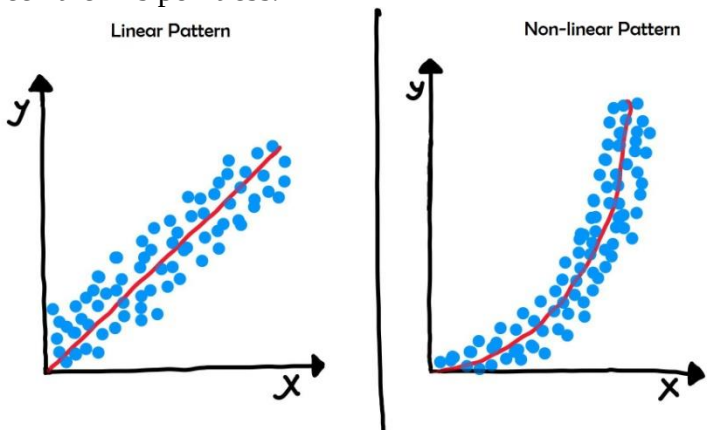
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable?** (1 mark)

The 'temp' variable has the strongest correlation with the target variable.

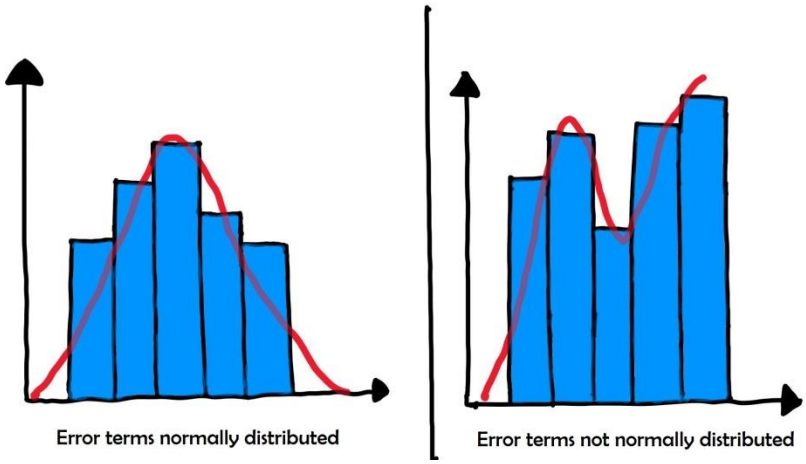
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?(3 marks)**

I have validated the assumptions of the Linear Regression Model based on the following five assumptions:

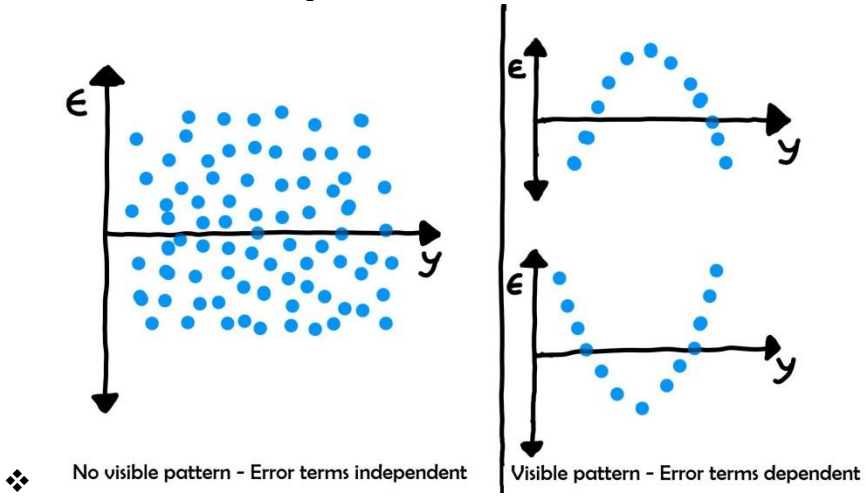
- X and Y should exhibit some form of linear relationship; otherwise, fitting a linear model between them is pointless.



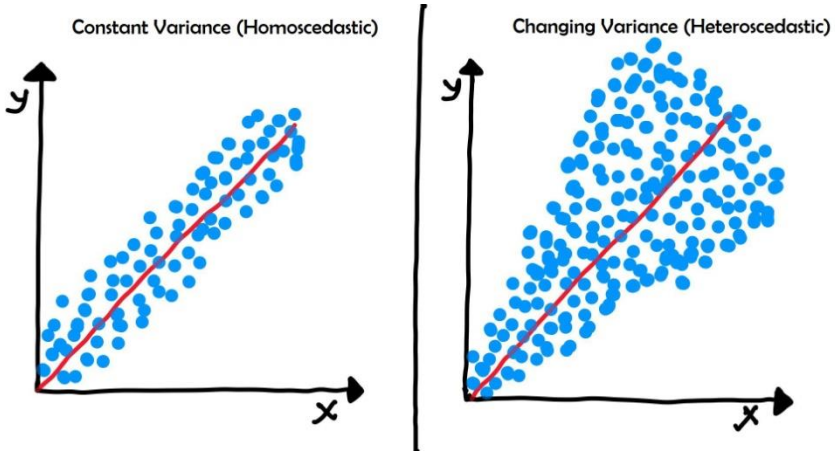
- The error terms should be normally distributed with a mean of zero (not X or Y).



- The error terms should be independent of each other.



- The error terms should have constant variance (homoscedasticity).



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

Below are the top three features that significantly contribute to explaining the demand for shared bikes:

- Temp
- Weathersit_moderate
- Workingday

General Subjective Questions

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

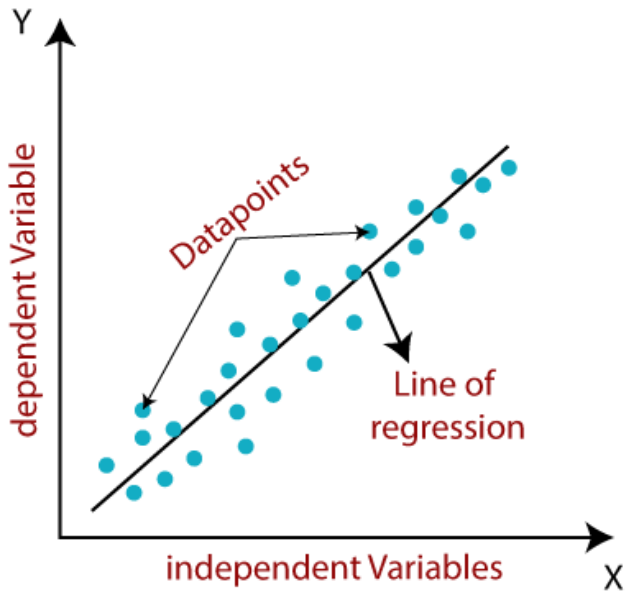
Linear regression is defined as a statistical model that analyzes the linear relationship between a dependent variable and a given set of independent variables. A linear relationship between variables means that when the value of one or more independent variables changes (increases or decreases), the value of the dependent variable changes accordingly (increases or decreases).

Mathematically, this relationship can be represented by the following equation:

$Y=mX+c$

Here:

- Y is the dependent variable we are trying to predict.
- X is the independent variable we are using to make predictions.
- m is the slope of the regression line, representing the effect X has on Y.
- c is a constant, known as the Y-intercept. If X=0, Y would be equal to c



Furthermore, the linear relationship can be positive or negative in nature, as explained below:

- **Positive Linear Relationship:** A linear relationship is positive if both the independent and dependent variables increase together.
- **Negative Linear Relationship**: A linear relationship is negative if the independent variable increases and the dependent variable decreases.

Linear regression is of the following two types:

- Simple Linear Regression
- Multiple Linear Regression

Assumptions:

The following are some assumptions about the dataset made by the Linear Regression model:

- **Multi-collinearity:**

  Linear regression assumes that there is very little or no multi-collinearity in the data. Multi-collinearity occurs when the independent variables or features have dependencies among them.

- **Auto-correlation**:

  Another assumption of the Linear Regression model is that there is very little or no auto-correlation in the data. Auto-correlation occurs when there is a dependency between residual errors.

- **Relationship between variables**:

  Linear regression assumes that the relationship between the response and feature variables must be linear.

- **Normality of error terms**:

  Error terms should be normally distributed.

- **Homoscedasticity**:

  There should be no visible pattern in residual values.

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

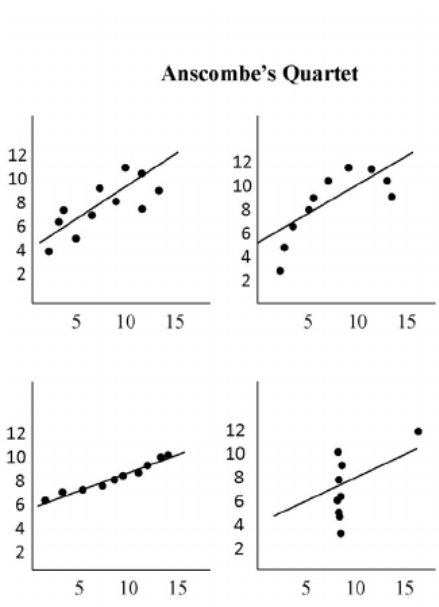
Anscombe's quartet is a famous dataset in statistics was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and to demonstrate the effect of outliers on statistical properties.

It consists of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear distinctively different when plotted.

Description of the Datasets

Let's describe each dataset in Anscombe's quartet:

1. **Dataset I:**
   - **x**: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
   - **y**: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
   - **Description**: This dataset has a linear relationship between x and y, and fits well with a linear regression model.
2. **Dataset II:**
   - **x**: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
   - **y**: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]
   - **Description**: This dataset also has a linear relationship between x and y, but with a different intercept and slight variations in the points compared to Dataset I.
3. **Dataset III:**
   - **x**: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
   - **y**: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
   - **Description**: This dataset has a non-linear relationship between x and y, and shows an outlier that influences the regression line and statistical properties.
4. **Dataset IV:**
   - **x**: [8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8]
   - **y**: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 5.56, 7.91, 6.89, 6.89]
   - **Description**: This dataset consists of a perfect relationship between x and y except for an outlier, which strongly affects the regression line and statistical properties.



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

I		II		III		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	5.76
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	8.81	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	7.26	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

Key Observations

- **Similar Summary Statistics**: All four datasets have nearly identical means, variances, correlations, and regression lines.
- **Different Distributions**: Despite having similar summary statistics, the datasets are visibly different when plotted.
- **Effect of Outliers**: Dataset III and Dataset IV show the impact of outliers on statistical properties and the regression line.

Importance

Anscombe's quartet highlights the limitations of summary statistics and the importance of visualizing data. It emphasizes that:

- Graphical representation of data can reveal patterns and relationships that summary statistics might miss.
- Outliers can heavily influence summary statistics and should be identified and analyzed separately.

3. **What is Pearson's R?** (3 marks)

Pearson's r, commonly known as Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between paired data.

Key Points about Pearson's r:

Definition: Pearson's correlation coefficient r is a statistical measure that ranges from -1 to 1, inclusive. It measures the degree of linear relationship between two variables X and Y.

Range:

- r=1: Perfect positive linear relationship.
- r=−1: Perfect negative linear relationship.
- r=0: No linear relationship (though note that it does not necessarily mean no relationship at all).

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

- r: Correlation coefficient
- xix_{i}xi: Values of the x-variable in a sample
- x¯\bar{x}x¯: Mean of the values of the x-variable
- yiy_{i}yi: Values of the y-variable in a sample

- $\bar{y}$: Mean of the values of the y-variable

Properties:

- **Direction**: The sign of r indicates the direction of the relationship. Positive r indicates a positive relationship (as one variable increases, the other tends to increase); negative r indicates a negative relationship (as one variable increases, the other tends to decrease).
- **Magnitude**: The magnitude of r indicates the strength of the relationship. Closer to 1 (either positive or negative), stronger the linear relationship. Closer to 0, weaker the linear relationship.

Assumptions:

- Pearson's r assumes that the relationship between the variables is linear.
- It assumes that the variables are approximately normally distributed.
- It assumes homoscedasticity, meaning that the variance of Y is the same for all levels of X.

Usage:

- Pearson's r is widely used in statistics, especially in fields like psychology, sociology, and economics, to measure relationships between variables.
- It is useful for determining whether and how strongly two continuous variables are related.

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

Feature scaling is a technique used to standardize the independent features in data to a fixed range. It is performed during data preprocessing to handle highly varying magnitudes, values, or units. If feature scaling is not applied, machine learning algorithms tend to give higher weights to features with greater values and consider smaller values as lower, regardless of their unit.

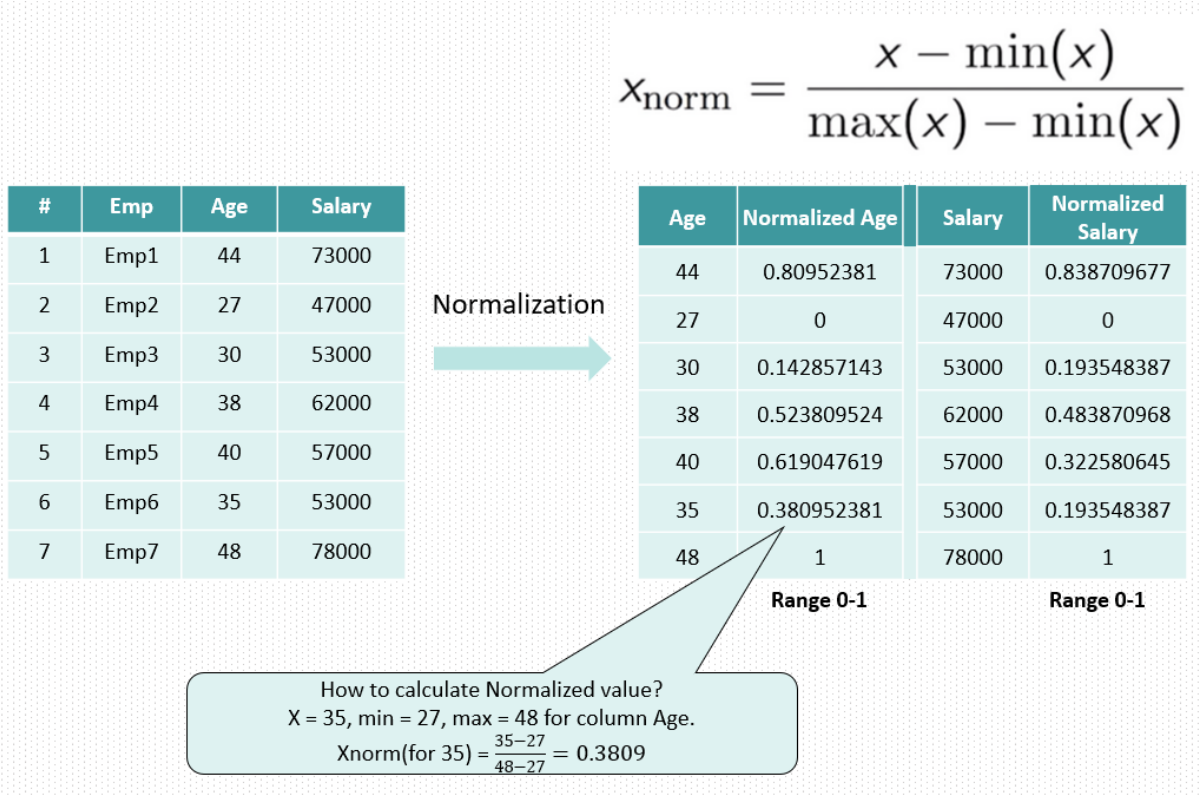
For example, without feature scaling, an algorithm might consider the value 3000 meters to be greater than 5 kilometers, which is not true. In such cases, the algorithm may make incorrect predictions. Feature scaling brings all values to the same magnitude, addressing this issue.

Feature Scaling Methods:

Normalized Scaling:

1. Uses the minimum and maximum values of features for scaling.
2. Scales values between [0, 1] or [-1, 1].
3. It is used when features are of different scales.
4. It is really affected by outliers.
5. Scikit-Learn provides a transformer called MinMaxScaler for normalization.

   Example:



#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Age	Normalized Age	Salary	Normalized Salary
44	0.80952381	73000	0.838709677
27	0	47000	0
30	0.142857143	53000	0.193548387
38	0.523809524	62000	0.483870968
40	0.619047619	57000	0.322580645
35	0.380952381	53000	0.193548387
48	1	78000	1
	Range 0-1		**Range 0-1**

Standardized Scaling:

1. Uses mean and standard deviation for scaling.
2. Ensures zero mean and unit standard deviation.
3. It is not bounded to a certain range.
4. It is much less affected by outliers.
5. Scikit-Learn provides a transformer called StandardScaler for standardization.

Example:

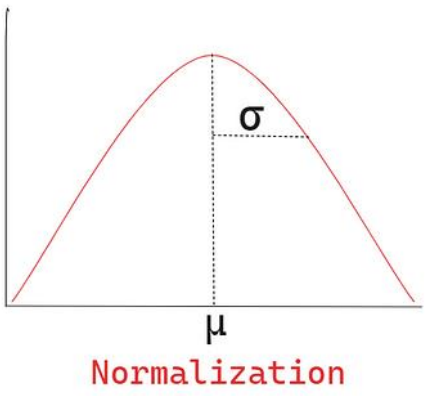
$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

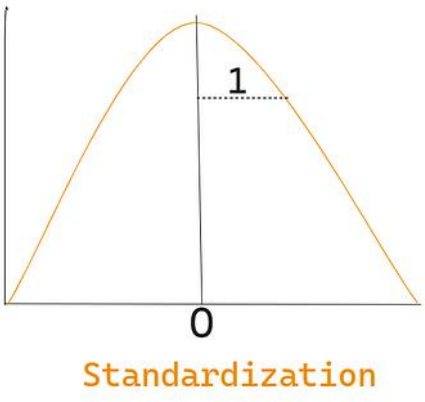
#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000
		Mean = 37.42857 Std. Dev. = 6.883876	Mean = 60428.5714 Std. Dev = 10499.7570

Standardization

How to calculate Standardized value?
X = 35, mean = 37.42, Std. Dev. = 6.88
for column Age.
Xstd(for 35) = $\frac{35-37.42}{6.88}$ = -0.3527

Age	Standardized Age	Salary	Standardized Salary
44	0.954611636	73000	1.197306616
27	-1.514927162	47000	-1.278941158
30	-1.079126198	53000	-0.707499364
38	0.083009708	62000	0.149663327
40	0.373543684	57000	-0.326538168
35	-0.352791257	53000	-0.707499364
48	1.535679589	78000	1.673508111
	Mean = 0 Std. dev. = 1		Mean = 0 Std. dev. = 1





These techniques help ensure that all features contribute equally to the analysis and prevent the model from being biased towards certain features due to their scale or range.

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

(3 marks)

The occurrence of an infinite value of VIF (Variance Inflation Factor) typically happens due to perfect multicollinearity in the dataset. Here's why this happens:

Understanding VIF:

Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A VIF value greater than 10 is often considered problematic and indicates high multicollinearity.

Formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Reason for Infinite VIF:

When the VIF for a particular predictor is infinite, it means that predictor can be perfectly explained by other predictors in the model. This situation occurs because:

1. **Perfect Multicollinearity**: One or more of the independent variables in the model are perfectly correlated with (linearly dependent on) one or more of the other independent variables. This leads to a situation where the regression model cannot distinguish between the effects of these variables.
2. **Linear Combination**: If one variable is a perfect linear combination of other variables in the model, it can cause an issue in the matrix inversion process used in calculating VIF. Specifically, the determinant of the correlation matrix used in VIF calculation is zero, leading to an infinite VIF.

Implications:

- **Model Fitting**: Infinite VIF values can cause issues in model fitting, particularly in linear regression, where they can lead to unstable coefficient estimates.
- **Interpretation**: It becomes impossible to interpret the impact of the variable with infinite VIF, as it is indistinguishable from the other correlated variables.

Resolution:

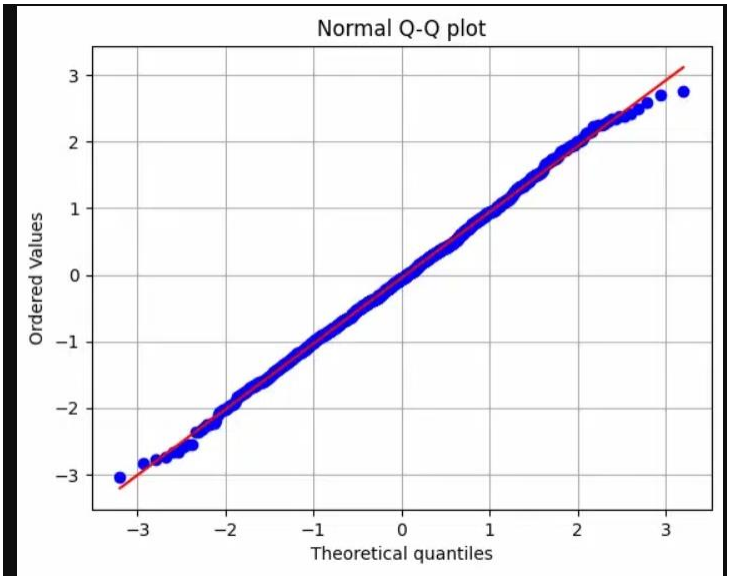
To resolve issues with infinite VIF:

- **Identify and Remove Variables**: Identify which variables are perfectly correlated and consider removing one of them from the model.
- **Data Transformation**: Sometimes, transforming variables (e.g., through differencing or combining them in a different way) can resolve multicollinearity issues.
- **Regularization**: Techniques like ridge regression can help in handling multicollinearity by penalizing large coefficients.

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

(3 marks)

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether or not a dataset follows a particular distribution, such as the normal distribution. It is particularly useful in linear regression and other statistical analyses to check the assumption of normality of residuals or to compare the distribution of a sample with a theoretical distribution.



Understanding Q-Q Plot:

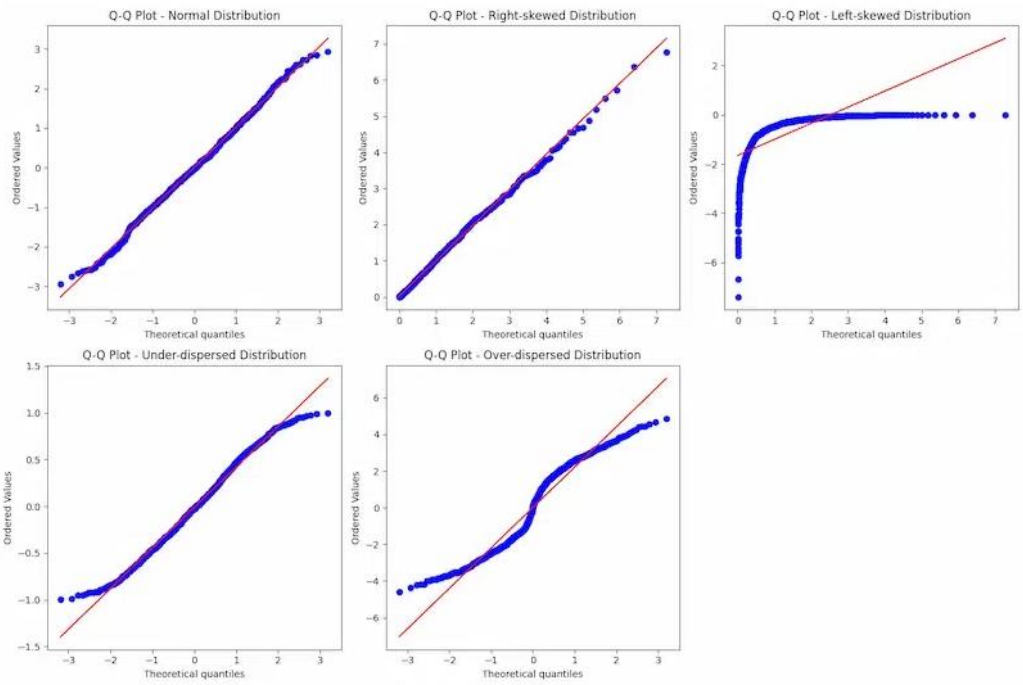
1. **Definition**:
    - A Q-Q plot compares the quantiles of the dataset against the quantiles of a specified theoretical distribution.
    - If the dataset and the theoretical distribution being compared have similar shapes, the points in the Q-Q plot will fall approximately on a straight line.
2. **Construction**:
    - The Q-Q plot is constructed by plotting the quantiles of the dataset on the x-axis and the quantiles of the theoretical distribution on the y-axis.
    - If the dataset is normally distributed, the points will fall along the diagonal line.
    - Deviations from the diagonal line indicate departures from normality.
3. **Interpretation**:
    - **Straight Line**: If the points lie approximately on a straight line, the dataset can be considered to be normally distributed.
    - **Curved Line**: If the points deviate from the straight line, it suggests that the dataset does not follow a normal distribution.

Use and Importance of Q-Q Plot in Linear Regression:

1. **Normality of Residuals**:
    - In linear regression, one of the key assumptions is that the residuals (the difference between observed and predicted values) should follow a normal distribution.
    - Q-Q plots are used to visually inspect whether the residuals are normally distributed. If the residuals are normally distributed, the points in the Q-Q plot will fall along a straight line.
2. **Detecting Departures from Normality**:

- Q-Q plots can help detect departures from normality that may affect the validity of statistical tests and confidence intervals.
- If the Q-Q plot shows deviations from the diagonal line (e.g., points curving upwards or downwards), it indicates that the residuals may not be normally distributed.

3. **Assessment of Model Assumptions**:
   - By examining the Q-Q plot, analysts can assess whether the assumptions of the linear regression model (such as normality of residuals) are met.
   - If the assumptions are not met, adjustments may need to be made, such as using transformations or considering a different model.
4. **Comparison with Theoretical Distributions**:
   - Q-Q plots can also be used to compare the distribution of a sample with other theoretical distributions, not just the normal distribution.
   - This allows for a broader assessment of whether the dataset conforms to various statistical assumptions.



Q-Q plot for different distributions

Example:

Consider a Q-Q plot used to check the normality of residuals in a linear regression model:

- If the residuals of the model are normally distributed, the points in the Q-Q plot will fall along a straight line.
- Deviations from this line suggest that the residuals are not normally distributed, indicating that the model assumptions may be violated.