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)

Solution:

1. **weathersit** value **clear** is having high median than other values Mist + Cloudy (2), Light Snow (3)
2. **season** values fall and summer are having high median
3. **mnth** the trend is in increasing until July and then following the decreasing trend
4. Working day registrations are high

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

Solution:

If we don't use this setting, it creates 'n' dummy variables for a categorical column with 'n' unique values but the same can be managed with less number of variables 'n-1' using this setting. Unnecessary extra columns can be reduced using this option and improve model performance.

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

Solution:

temp : temperature in Celsius is having highest correlation

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

Solution:

The residuals are centered around mean value 0(Zero). Tested with histogram based on y-predicted and y-actual.

The scattered plot is following a linear regression model.

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

Solution:

temp : temperature in Celsius

yr : year (0: 2018, 1:2019)

Jun : June Month

General Subjective Questions

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

Solution:

Linear regression is a basic yet powerful predictive modeling technique. In simple terms, linear regression uses a straight line to describe the relationship between a predictor variable (x) and a response variable (y). The linear regression equation takes the form of:

$$y=b_0+b_1*x$$

Where b0 is the intercept and b1 is the slope of the line. The goal of linear regression is to find the line that best fits the data. This allows us to make predictions about y based on values of x.

linear regression is a statistical technique for modeling the linear relationship between a dependent variable y and one or more independent variables x. The dependent variable is also called the outcome or response variable. The independent variables are also called explanatory or predictor variables.

For example, we could use linear regression to understand the relationship between advertising spending (predictor variable) and sales revenue (response variable). Or we could predict home prices (response) based on features like square footage, number of bedrooms, etc (predictors).

The linear regression model assumes a linear functional form:

$$y=b_0+b_1*x_1+..+b_n*x_n$$

Where y is the response, x1 through xn are the predictors, b0 is the intercept, and b1 through bn are the coefficients.

The coefficients describe the size and direction of the relationship between each predictor and the response. Once we've trained a linear regression model on our data, we can use it to make predictions for new data points.

Types of Linear Regression Models

There are several types of linear regression models, each used for different purposes:

Simple Linear Regression: Used when there is only one predictor variable, simple linear regression takes the form:

$$y=b_0+b_1*x$$

For example, predicting home price based only on square footage.

Multiple Linear Regression: Used when there are two or more predictor variables. The general form is:

$$y=b_0+b_1*x_1+..+b_n*x_n$$

For example, predicting home price based on square footage, number of bedrooms, location, etc.

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

Solution:

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician ***Francis Anscombe*** in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

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	7.71
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	7.81	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	9.13	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

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

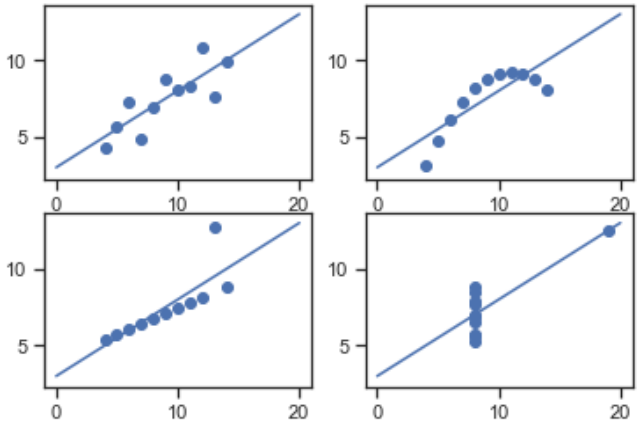
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



Graphical Representation of Anscombe's Quartet

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

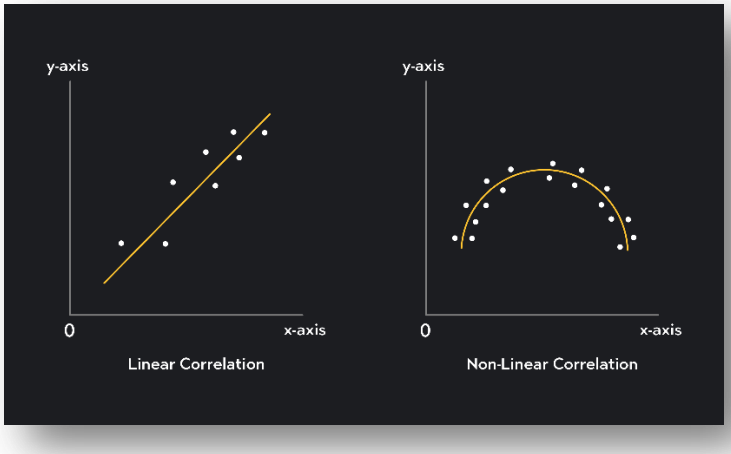
3. What is Pearson's R? (3 marks)

Solution:

The Pearson correlation coefficient is one of the most common methods for measuring correlation. You might also hear this term being called Pearson's r, a bivariate correlation, the Pearson product-moment correlation coefficient (PPMCC), or just, the correlation coefficient for short.

A Pearson correlation coefficient measures a linear correlation's direction and magnitude. A linear association—as opposed to a non-linear one—is a correlation approximated by a straight line, where the change in one variable is approximately proportional to the observed change in the second variable.

A linear correlation is strictly positive or negative, whereas a non-linear correlation can change with the values of x and y.



How To Determine the Strength of Association

Pearson's r ranges from -1 to 1, where -1 represents a perfect negative correlation, and 1 represents a perfect positive correlation. The closer the absolute value of r is to 1, the stronger the correlation, and the closer the absolute value is to 0, the weaker the correlation.

A strong correlation means a stronger association between the two variables. If X and Y are strongly correlated, knowing the value of X gives you more information about Y—and vice versa—compared to when the variables are weakly correlated.

1. Perfect Negative Correlation (r=-1)

A perfect negative correlation is an association between two variables where an increase in one is always associated with a perfectly proportional decrease in the other. In other words, the two variables have a perfectly proportional inverse relationship. The correlation coefficient for a perfectly negative correlation is -1.

2. Negative Correlation ($-1 \leq r < 0$)

A negative correlation is any inverse correlation where an increase in the value of X is associated with a decrease in the value of Y. For a negative correlation, Pearson's r is less than 0 and greater than or equal to -1.

3. Zero Correlation (r=0)

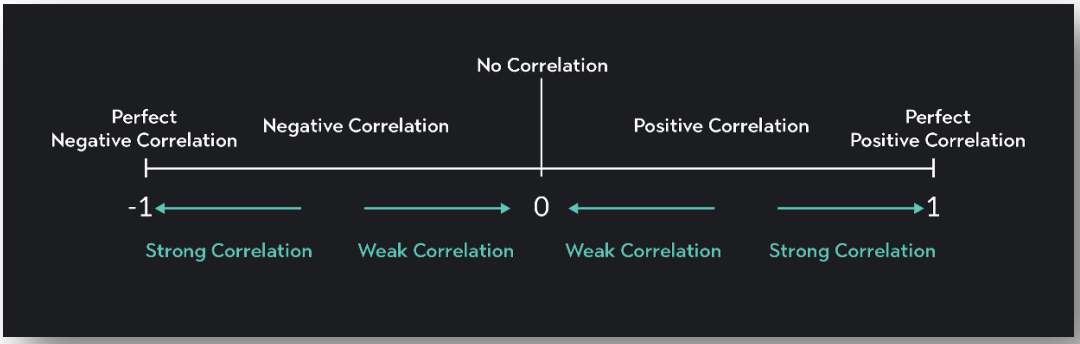
A zero correlation indicates there is no observable linear relationship between your two variables. Knowing the values of X will not tell you anything about the value of Y.

4. Positive Correlation ($0 < r \leq 1$)

A positive correlation is any correlation where an increase in the value of X is associated with an increase in the value of Y, and a decrease in the value of X is associated with a decrease in the value of Y. For a positive correlation, Pearson's r will be greater than 0 or less than or equal to 1.

5. Perfect Positive Correlation (r=1)

A perfect positive correlation is an association between two variables where an increase in one is always associated with a perfectly proportional increase in the other. The correlation coefficient for a perfectly positive correlation is 1.



Correlations and Scatter Plots

Scatter plots are a useful way of visualizing correlations. A scatter plot is a graph that maps the values of one variable—measured along the x-axis—to the values of the second variable—measured along the y-axis.

If there is a linear correlation between your two variables, you can draw an upward or downward-sloping straight trend line through your data to approximate the association.

By looking at a scatterplot, you should be able to determine both the direction and magnitude of a linear correlation.

1. Is the correlation positive or negative?

   Looking at a scatter plot, you can tell whether a correlation is positive or negative by the slope of the trend line.

   - A negative downward-sloping line indicates a negative correlation
   - A positive upward-sloping line indicates a positive correlation
   - If no linear trend line can be drawn through the data, there is no linear correlation.

2. Is the correlation strong or weak?

   You can approximate the strength of a correlation by looking at how close the data points are to the trend line.

   - The more closely clustered the data points are around the trend line, the stronger the correlation.
   - The further away the data points are from the trend line, the weaker the correlation.

When To Use the Pearson Correlation Coefficient

You can use the Pearson coefficient under the following circumstances:

- The variables you are comparing are both quantitative variables. If you are working with ordinal variables, you can use the Spearman rank correlation or Kendall's tau, and If your variables are nominal, you can use a correlation measure called Cramér's V.
- Each variable is normally distributed.
- You have no outliers in your data. Correlations are extremely sensitive to outliers. If you include even one outlier in your calculations, you will get misleading results.
- The relationship between the two variables appears to be linear rather than non-linear

How To Find the Pearson Correlation Coefficient

How To Calculate Pearson's r by Hand

Pearson's correlation coefficient is equal to the covariance of your two variables divided by the product of their standard deviations.

What is Pearson's R? What is Pearson's R?

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

Where:

- r is the Pearson correlation coefficient
- x and y are the two variables
- cov(x,y) is the covariance of x and y
- sx is the standard deviation of x
- sy is the standard deviation of y
- xi is each individual observation of x
- yi is each individual observation of y
- $\bar{x}$ is the mean of x
- $\bar{y}$ is the mean of y

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

Solution:

Scaling refers to transforming the features in your dataset so that they have similar scales or distributions. This process is essential because features with different scales can adversely affect the performance of certain algorithms, such as gradient descent-based methods or distance-based algorithms.

Why is Scaling Performed?

Scaling is performed for several reasons:

1. **Improved Algorithm Performance:**
    - **Gradient Descent:** Algorithms like linear regression, logistic regression, and neural networks use gradient descent, which can converge faster if features are scaled. Features with vastly different scales can cause slow convergence or poor results.
    - **Distance-Based Algorithms:** Algorithms such as k-Nearest Neighbors (k-NN) and clustering algorithms (e.g., k-means) rely on distance calculations. If features are on different scales, the distance metric can be dominated by features with larger scales, leading to biased results.
2. **Equal Weighting of Features:**
    - Scaling ensures that all features contribute equally to the model, preventing features with larger ranges from disproportionately influencing the model's behavior.
3. **Improved Numerical Stability:**

- Algorithms that involve matrix operations can suffer from numerical stability issues if features have very large or very small values. Scaling helps mitigate these issues.

Types of Scaling

There are two common types of scaling: **normalization** and **standardization**.

1. Normalization (Min-Max Scaling)

Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1]. The formula for normalization is:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where:

- x is the original value.
- $\min(x)$ and $\max(x)$ are the minimum and maximum values of the feature, respectively.
- x_{norm} is the normalized value.

Pros:

- Useful when you want to bound your data within a specific range.
- Ensures that all features are on the same scale, which can be particularly helpful for algorithms that require bounded input features.

Cons:

- Sensitive to outliers. Outliers can significantly affect the min and max values, which can lead to skewed results.

2. Standardization (Z-score Scaling)

Standardization transforms data to have a mean of 0 and a standard deviation of 1. The formula for standardization is:

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

Where:

- x is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.
- x_{std} is the standardized value.

Pros:

- Not sensitive to outliers as much as normalization because it doesn't bound the data to a specific range.
- Standardized features often lead to better performance in algorithms that assume normally distributed data.

Cons:

- Transformed data does not have a fixed range. If the data needs to be bounded for some reason, standardization alone might not be sufficient.

Comparison Between Normalized and Standardized Scaling

- **Range:**
    - **Normalization** bounds the data to a fixed range [0, 1] (or [-1, 1]).
    - **Standardization** does not bound the data; it scales data to have a mean of 0 and standard deviation of 1.
- **Impact of Outliers:**
    - **Normalization** is sensitive to outliers because it relies on the minimum and maximum values.
    - **Standardization** is less affected by outliers because it uses the mean and standard deviation, which are less sensitive to extreme values.
- **Use Cases:**
    - **Normalization** is often used when features need to be on the same scale, especially in algorithms that require bounded data (e.g., neural networks).
    - **Standardization** is used when the data needs to be normalized to a distribution with a mean of 0 and a standard deviation of 1, particularly for algorithms that assume normally distributed data.

Choosing Between Normalization and Standardization

The choice between normalization and standardization depends on the specific requirements of the algorithm and the nature of the data. In general:

- Use **normalization** if your algorithm is sensitive to the range of data values or if you need to bound your data.
- Use **standardization** if you need to handle features with different variances and if your algorithm assumes normally distributed data.

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

(3 marks)

Solution:

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are

highly correlated, which can lead to issues in estimating the coefficients of the model accurately. VIF quantifies how much the variance of an estimated regression coefficient increases due to multicollinearity.

What is VIF?

The VIF for a predictor variable X_i is calculated as:

$$\text{VIF}(X_i) = \frac{1}{1-R_i^2}$$

where R_i^2 is the coefficient of determination obtained by regressing X_i on all the other predictor variables.

Why Might VIF Be Infinite?

A VIF value becomes infinite when the denominator in the VIF formula is zero, i.e., when:

$$1 - R_i^2 = 0$$

or equivalently,

$$R_i^2 = 1$$

This scenario implies perfect multicollinearity. Perfect multicollinearity occurs when:

1. **Perfect Linear Relationship:**
    - One of the predictor variables is a perfect linear combination of one or more other predictor variables. For example, if X_1 can be exactly expressed as a linear combination of X_2 and X_3, then R_i^2 for X_1 will be 1 when regressed on X_2 and X_3, leading to an infinite VIF.
2. **Redundant Predictors:**
    - The presence of a redundant predictor in the dataset can cause this issue. For instance, if you include both a variable and its linear transformation in the model (e.g., including both X and $2X$), you will have perfect multicollinearity.

Implications of Infinite VIF

- **Model Interpretation:**
    - When VIF is infinite, it indicates that the variable in question is redundant and does not contribute additional unique information to the model. This makes it difficult or impossible to estimate the coefficient for that variable accurately.
- **Parameter Estimates:**
    - Infinite VIF implies that the variance of the estimated regression coefficients for the involved variables is extremely high, leading to unreliable parameter estimates.

How to Address Infinite VIF

1. **Remove Redundant Variables:**

- Identify and remove variables that are perfectly collinear or highly redundant. This can often be done through domain knowledge or exploratory data analysis.

2. **Combine Variables:**
    - If appropriate, combine collinear variables into a single predictor to reduce redundancy.
3. **Principal Component Analysis (PCA):**
    - Use PCA or other dimensionality reduction techniques to transform the predictors into a set of uncorrelated components.
4. **Regularization:**
    - Apply regularization techniques such as Ridge Regression or Lasso, which can help in dealing with multicollinearity by adding a penalty to the size of the coefficients.
5. **Check Data:**
    - Recheck data entry to ensure that there are no mistakes leading to perfect multicollinearity.

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

(3 marks)

Solution:

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution, such as the normal distribution. In the context of linear regression, Q-Q plots are primarily used to evaluate the normality of residuals, which is an important assumption for many regression models.

What is a Q-Q Plot?

A Q-Q plot compares the quantiles of a dataset with the quantiles of a theoretical distribution. Here's how it works:

1. **Compute Quantiles:**
    - **Empirical Quantiles:** These are the quantiles of your actual data. For example, you might rank your data points and compute the quantiles at 10%, 20%, 30%, etc.
    - **Theoretical Quantiles:** These are the quantiles from a theoretical distribution. For example, if you assume your data should follow a normal distribution, you compute the quantiles from the normal distribution.
2. **Plot Quantiles:**
    - Plot the empirical quantiles on the y-axis and the theoretical quantiles on the x-axis.
3. **Interpretation:**

- If the points on the Q-Q plot fall approximately along a straight line (typically the 45-degree line, or line y = x), this suggests that the empirical data distribution is close to the theoretical distribution.
- Deviations from the line indicate departures from the theoretical distribution.

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

In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. This assumption is important for hypothesis testing and for making reliable inferences about the regression coefficients. Here's how a Q-Q plot is used and why it's important:

1. Checking Normality of Residuals

- **Purpose:** The primary use of a Q-Q plot in linear regression is to assess whether the residuals of the model are approximately normally distributed.
- **Process:**
  - After fitting a linear regression model, you calculate the residuals.
  - Generate a Q-Q plot of these residuals against the quantiles of the normal distribution.

2. Validating Model Assumptions

- **Assumption of Normality:** Many statistical tests and confidence intervals in linear regression rely on the assumption that residuals are normally distributed. A Q-Q plot helps validate this assumption.
- **Model Diagnostics:** By checking the normality of residuals, you can identify if the model is a good fit for the data or if there are deviations that might suggest model improvements.

3. Detecting Outliers and Influential Points

- **Outliers:** Deviations from the Q-Q line may indicate outliers or influential points in the data. For example, points that are far from the line in the tails of the plot might be outliers.
- **Influence:** Extreme deviations from the line might suggest that the data has influential points that could affect the regression results.

4. Assessing Model Fit

- **Good Fit:** If the Q-Q plot shows the residuals closely following the theoretical distribution line, it suggests a good fit of the model.
- **Poor Fit:** Significant deviations from the line indicate problems with the model fit, suggesting that the model may need improvement or that the normality assumption might be violated.

Interpreting a Q-Q Plot

- **Points on the Line:** If the points lie approximately on the line, the residuals are normally distributed.

- **Systematic Deviations:** If the points curve away from the line (e.g., S-shaped curves), it indicates non-normality.
- **Heavy Tails:** If points diverge at the ends, it may suggest heavy tails in the residuals compared to the theoretical normal distribution.

Example in Linear Regression

Assume you have fitted a linear regression model and obtained the residuals. You generate a Q-Q plot of these residuals against a normal distribution:

- **Straight Line:** If the residuals follow the straight line, it suggests that the normality assumption is satisfied.
- **S-Curve or Other Patterns:** If you see deviations from the line, it might indicate issues such as skewness, kurtosis, or other deviations from normality.