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)**

Answer:

Based on the analysis done following points could be concluded:

- Compared to 2018 there is significant rise in demand for bikes in 2019, though the data available is only for 2 years there is still strong correlation between dependant variable and year
- Season also has good effect of dependent variable in both years there is significant rise in demand in fall season, whereas in spring season demand plummets.
- For Aug, Sep, Jun, Jul, May, Oct there is significantly higher demand for bikes, as we approach year end demand decreases and reach lowest in Jan then it starts rising again. Due to this trend, there can be impact on dependant variable (if we are predicting count in Sep and in Jan there is very high chance that predicted count for Sep will be greater than Jan)
- Clear Weather attracts more demand which is intuitive. Whereas light snow, light rain attracts less demand. For very bad weather condition there is no record present in dataset. Thus, we can conclude that there is relation between weather and count.
- In case of Wed, Thu, Fri, Sat there is high demand for year 2019, in case of 2018 there is no such pattern, which indicates that the weekday may affect the count but not in significant way.
- In case of holiday clear drop in demand can be observed which makes sense as on holidays people are less likely to commute whereas in case of non-holiday many people may opt for bikes for reaching their offices.

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

Answer:

The drop_first=True parameter in pandas' get_dummies() function is crucial when creating dummy variables from categorical features, especially when the categorical feature has a nominal scale (i.e., the categories have no inherent order).
Reasons are as follows:

1. Prevent Multicollinearity:

- When you create dummy variables for a categorical feature with n categories, you'll end up with n-1 dummy variables.
- If you don't drop the first dummy variable, you'll introduce multicollinearity. This means that one dummy variable can be perfectly predicted from the others, leading to numerical instability and potential issues in modeling.

2. Avoid Redundancy:

- The information contained in the first dummy variable is already captured by the other n-1 dummy variables.

- Dropping the first dummy variable eliminates redundancy and ensures that your model is using the most efficient representation of the categorical feature.

3. Interpretability:

- Dropping the first dummy variable often makes the interpretation of the model's coefficients more intuitive.
- The coefficients of the remaining dummy variables represent the difference in the outcome variable relative to the reference category (the category that was dropped).

Example:

Consider a categorical feature "Color" with three categories: "Red", "Green", and "Blue".

- Without drop_first=True, you'd get three dummy variables: Color_Red, Color_Green, and Color_Blue.
- With drop_first=True, you'd get only two dummy variables: Color_Green and Color_Blue.
- The coefficient of Color_Green would represent the difference in the outcome variable between "Green" and "Red" (the reference category), and the coefficient of Color_Blue would represent the difference between "Blue" and "Red".

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

Answer:

temp' variable has the highest correlation with the target variable with correlation of 0.63

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

Answer:

- Normality of error terms
- Multicollinearity check
- Linear relationship validation
- 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)**

Answer:

1. Temp
2. Year
3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (part of weather)

General Subjective Questions:

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

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find a linear equation that best fits the data points, allowing us to predict the dependent variable based on the values of the independent variables.

The Equation

The linear regression equation is typically represented as:

y = mx + b

Where:

- y: is the dependent variable (what we're trying to predict)
- x: is the independent variable (the predictor)
- m: is the slope of the line (how much y changes for every unit change in x)
- b: is the y-intercept (the value of y when x is 0)

The Learning Process

1. Data Collection: Gather a dataset with pairs of (x, y) values. These values represent the relationship between the independent and dependent variables.
2. Model Training:
    - Cost Function: Define a cost function to measure the error between the predicted values and the actual values. A common choice is the mean squared error (MSE).
    - Gradient Descent: Use an optimization algorithm like gradient descent to minimize the cost function. This involves adjusting the values of m and b iteratively to reduce the error.
3. Evaluation: Once the model is trained, evaluate its performance using metrics like R-squared, mean squared error, or root mean squared error.

Types of Linear Regression

- Simple Linear Regression: Involves a single independent variable.
- Multiple Linear Regression: Uses multiple independent variables to predict the dependent variable.

Applications of Linear Regression

Linear regression is widely used in various fields, including:

- Economics: Predicting stock prices, GDP growth, or consumer spending.
- Finance: Analyzing risk factors, predicting loan defaults, or assessing investment returns.
- Marketing: Predicting sales, customer churn, or advertising effectiveness.
- Healthcare: Analyzing patient outcomes, predicting disease progression, or identifying risk factors.
- Engineering: Modeling physical processes, predicting system performance, or optimizing designs.

Assumptions of Linear Regression

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: The observations are independent of each other.

- Homoscedasticity: The variance of the errors is constant across all values of the independent variable.
- Normality: The errors are normally distributed.

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

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties (mean, median, variance, standard deviation, correlation coefficient). However, when visualized, they reveal strikingly different patterns, highlighting the importance of visual data exploration in addition to numerical analysis.

The four datasets consist of:

1. **Dataset 1:** A perfect linear relationship between x and y.
2. **Dataset 2:** A quadratic relationship with a single outlier.
3. **Dataset 3:** A linear relationship with a constant x value for most points, creating a vertical line.
4. **Dataset 4:** A horizontal line with a few outliers.

Key takeaways from Anscombe's quartet:

- **Visual exploration is crucial:** While statistical measures can provide valuable information, they may not capture the full complexity of a dataset. Visualizing the data can reveal patterns, outliers, and relationships that might be missed through numerical analysis alone.
- **Outliers can significantly impact results:** Even a single outlier can dramatically change the appearance and interpretation of a dataset. It's essential to identify and consider the impact of outliers when analyzing data.
- **Correlation does not imply causation:** A high correlation between two variables does not necessarily mean that one causes the other. There might be other factors influencing the relationship, or the relationship could be coincidental.
- **Multiple models can fit the same data:** Different models may fit a dataset equally well, but they can lead to different interpretations and predictions. It's important to consider the context and underlying assumptions when choosing a model.

In conclusion**,** Anscombe's quartet serves as a powerful reminder that data analysis should be a combination of numerical and visual methods. By carefully examining both the statistical properties and the visual representation of a dataset, we can gain a deeper understanding of the underlying relationships and avoid potential pitfalls in our analysis.

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

Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1:

- **r = 1:** Indicates a perfect positive correlation, meaning the two variables increase or decrease together perfectly.
- **r = -1:** Indicates a perfect negative correlation, meaning one variable increases as the other decreases perfectly.
- **r = 0:** Indicates no correlation between the two variables.

Formula for Pearson's correlation coefficient:

$r = (n\Sigma xy - \Sigma x\Sigma y) / \mathrm{sqrt}((n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2))$

Where:

- n = number of data points
- Σxy = sum of the product of corresponding x and y values
- Σx = sum of x values
- Σy = sum of y values
- Σx^2 = sum of the squared x values
- Σy^2 = sum of the squared y values

Interpretation:

- **Strength:** The absolute value of r indicates the strength of the linear relationship. A value closer to 1 or -1 indicates a stronger relationship, while a value closer to 0 indicates a weaker relationship.
- **Direction:** The sign of r indicates the direction of the relationship. A positive r means the two variables increase or decrease together, while a negative r means one variable increases as the other decreases.

Assumptions:

- **Linearity:** The relationship between the two variables is linear.
- **Normality:** Both variables are normally distributed.
- **Independence:** The observations are independent of each other.

Example:

If r = 0.8 between the variables "height" and "weight" in a group of people, this suggests a strong positive linear relationship between height and weight. As height increases, weight tends to increase as well.

It's important to note that Pearson's correlation coefficient only measures linear relationships. If the relationship is non-linear, Pearson's r may not accurately reflect the association between the variables. In such cases, other correlation measures like Spearman's rank correlation coefficient might be more appropriate

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

Answer:

Scaling is a technique used in data preprocessing to transform data to a common scale, usually between 0 and 1 or -1 and 1. This is essential for many machine learning algorithms, especially those that involve distance calculations or gradient descent, as they often assume features are on a similar scale.

Why is Scaling Performed?

1. **Improved Algorithm Performance:** Many algorithms, such as K-means clustering, support vector machines (SVMs), and neural networks, rely on distance calculations. If features have vastly different scales, the algorithm might be biased towards features with larger magnitudes. Scaling ensures that all features contribute equally to the model.
2. **Faster Convergence:** Gradient descent, a common optimization algorithm used in many machine learning models, converges faster when features are on a similar scale. This is because the gradients of the loss function are more balanced, preventing the algorithm from taking excessively large steps in the direction of features with larger magnitudes.
3. **Regularization:** Some regularization techniques, like L1 and L2 regularization, assume features are on a similar scale. If features have different scales, the regularization penalty might not be applied fairly.

Normalized Scaling vs. Standardized Scaling

Normalized Scaling (Min-Max Scaling)

- **Transforms data to a specified range, typically between 0 and 1.**
- **Formula:** (x - min(x)) / (max(x) - min(x))
- **Use case:** When you want to preserve the relative differences between values but ensure they are within a specific range. For example, if you want to compare percentages or proportions.

Standardized Scaling (Z-score Standardization)

- **Transforms data to have a mean of 0 and a standard deviation of 1.**
- **Formula:** (x - mean(x)) / std(x)
- **Use case:** When you want to remove the influence of outliers and ensure that features are centered around 0. This is often useful for algorithms that assume a normal distribution of features.

Choosing the Right Scaling Method:

- **Normalized Scaling:** Suitable when you want to preserve the relative differences between values and the data is not heavily skewed.
- **Standardized Scaling:** Suitable when you want to remove the influence of outliers and the data is approximately normally distributed.

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

Answer:

VIF (Variance Inflation Factor) is a measure of how much the variance of a regression coefficient is inflated due to multicollinearity. A high VIF indicates that the variable is highly correlated with one or more other predictors in the model.

When VIF becomes infinite, it essentially means that the variable is perfectly collinear with one or more other predictors. This means that the variable can be perfectly predicted by a linear combination of the other predictors.

Here are some common reasons why VIF can become infinite:

1. **Perfect Multicollinearity:** If two or more predictors are perfectly correlated, their VIF will be infinite. This can happen due to various reasons, such as:
    - **Redundant Variables:** Including the same variable twice in the model, or including a variable that is a linear combination of other variables.
    - **Dummy Variable Trap:** When creating dummy variables for categorical variables, it's essential to avoid the dummy variable trap. This occurs when all dummy variables are included in the model, leading to perfect multicollinearity.

2. **Numerical Issues:** In some cases, numerical instability can lead to VIF values approaching infinity. This can happen due to:
    - **Small Sample Size:** With small sample sizes, the correlation matrix can become ill-conditioned, leading to numerical problems.
    - **Rounding Errors:** In computations involving floating-point numbers, rounding errors can accumulate and cause numerical instability.

Detecting and Addressing Infinite VIF:

- **Correlation Matrix:** Examine the correlation matrix to identify highly correlated variables.
- **VIF Analysis:** Calculate VIF for each predictor. A VIF of 10 or higher is often considered indicative of a significant multicollinearity problem.
- **Remove Redundant Variables:** Identify and remove redundant variables from the model.
- **Center and Scale Variables:** Centering and scaling variables can sometimes help mitigate numerical instability.
- **Consider Alternative Models:** Explore alternative model formulations or techniques that are less sensitive to multicollinearity, such as ridge regression or Lasso regression.

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

A **Q-Q plot** (Quantile-Quantile plot) is a graphical technique used to compare the distribution of a sample data set to a theoretical distribution, often the normal distribution. It's a valuable tool for assessing whether a dataset follows a particular distribution.

How a Q-Q Plot Works

1. **Rank the data:** Sort the observed data in ascending order.

2. **Calculate theoretical quantiles:** Determine the corresponding theoretical quantiles from the desired distribution (e.g., normal distribution).
3. **Plot:** Plot the observed quantiles against the theoretical quantiles.

If the data follows the theoretical distribution, the points on the Q-Q plot will fall approximately along a straight line. Deviations from this line indicate that the data does not follow the theoretical distribution.

Importance of Q-Q Plots in Linear Regression

In linear regression, the assumption of normality of the residuals (the differences between the observed and predicted values) is crucial for the validity of statistical inferences. A Q-Q plot can be used to assess this assumption:

1. **Residual Normality:** Create a Q-Q plot of the residuals against the theoretical quantiles of a normal distribution.
2. **Interpretation:**
    - **Linearity:** If the points fall approximately along a straight line, it suggests that the residuals are normally distributed.
    - **Deviations:** Deviations from the line indicate non-normality. For example, a "S" shape might suggest a skewed distribution, while a "U" shape might indicate a heavy-tailed distribution.

Why is Normality of Residuals Important?

- **Statistical Inference:** Many statistical tests and confidence intervals rely on the assumption of normally distributed residuals.
- **Model Validity:** Non-normal residuals can invalidate the model's assumptions and affect the accuracy of predictions.
- **Model Selection:** Understanding the distribution of residuals can help identify appropriate model transformations or alternative models.

In conclusion, Q-Q plots are a valuable tool for assessing the normality of residuals in linear regression.