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?

We can make following deductions

- there are more riders on clear day and fall season
- Working-day/holiday doesnt seem to have an impact on number of riders
- september has highest number of riders
- Number of riders have increased wrt previous year

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

Drop_first=True avoids creation of extra column during dummy variable creation.
This can be understood using an example

Lets say home furnishing has 3 values furnished,semi-furninshed and unfurnished.

If for a given record we know that furnished is false and unfurnished is false we can deduce that the given house is semi-furnished.

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

Temperature has highest correlation with the target variable

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

1. Homoscedasticity : Plot the residuals against the fitted values. The spread of the residuals should be roughly constant across all levels of the fitted values.
2. Linearity : Plot the residuals (the differences between the observed and predicted values) against the predicted values. If the plot shows no clear pattern, the linearity assumption is likely met.
3. No Multicollinearity : Calculate the VIF for each predictor. A VIF value greater than 10 indicates high multicollinearity.

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

Temperature, Light snow rain and year have significant impact

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable (response variable) and one or more independent variables (predictors). The goal is to find the best-fitting linear equation that describes how the dependent variable changes as the independent variables change. Here's a detailed explanation of the linear regression algorithm:

1. The Model

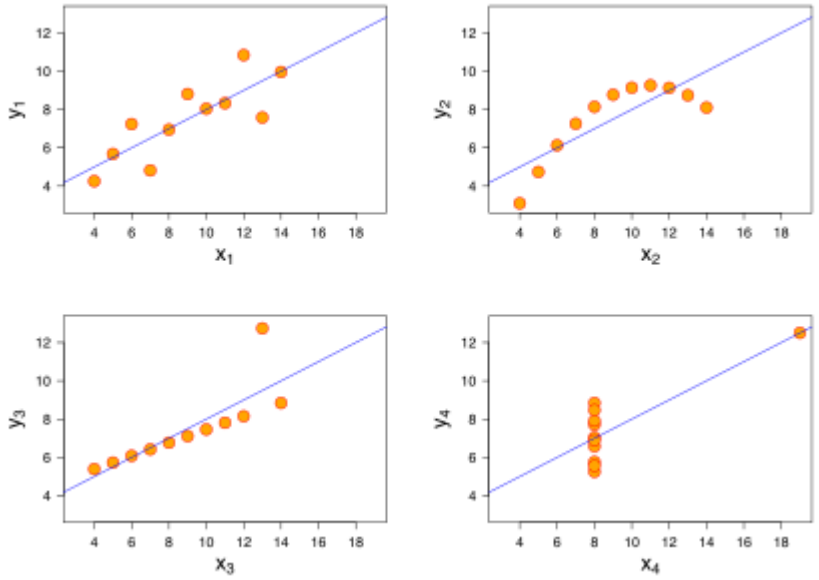
In simple linear regression, we model the relationship between two variables y (dependent variable) and x (independent variable) using a linear equation: $y=\beta0+\beta1x+\epsilon y$ where:

- y is the dependent variable.
- x is the independent variable.
- $\beta0$ is the intercept (the value of y when x=0).
- $\beta1$ is the slope (the change in y for a one-unit change in x).
- ϵ is the error term (the difference between the observed and predicted values of y).

In multiple linear regression, the model extends to include multiple independent variables: $y=\beta0+\beta1x1+\beta2x2+\ldots+\beta pxp+\epsilon y$ where p is the number of independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.
It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.



3. What is Pearson's R?

The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.
0	No correlation	There is no relationship between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling in Machine Learning

Scaling adjusts the range and distribution of features in a dataset to improve the performance and convergence speed of many machine learning algorithms.

Why Perform Scaling?

1. **Improves Algorithm Performance**: Algorithms like gradient descent converge faster.
2. **Prevents Feature Dominance**: Ensures no single feature dominates the model.
3. **Reduces Numerical Instability**: Mitigates issues in mathematical computations.

Types of Scaling

1. Normalization (Min-Max Scaling)

- **Rescales** the feature to a fixed range, typically [0, 1].
- **Formula**: $\mathbf{X}norm = (\mathbf{X} - \mathbf{X}_{min})/(\mathbf{X}max - \mathbf{X}min)$
- **Pros**: Maintains relationships between data points.
- **Cons**: Sensitive to outliers.

2. Standardization (Z-score Scaling)

- **Transforms** the feature to have a mean of 0 and a standard deviation of 1.
- **Formula**:
- **Pros**: Less sensitive to outliers.
- **Cons**: Does not bound values within a fixed range.

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

The Variance Inflation Factor (VIF) is a measure used to detect the presence and severity of multicollinearity in regression models. A high VIF indicates that a predictor variable has a strong linear relationship with one or more other predictor variables.

Why VIF Can Be Infinite

VIF can become infinite when perfect multicollinearity is present in the dataset. Perfect multicollinearity occurs when one predictor variable is a perfect linear combination of one or more other predictor variables. This situation makes it impossible to isolate the individual contribution of the perfectly collinear variable because its effect is indistinguishable from the combination of the other variables.

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

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. The plot is used to assess whether the residuals of a linear regression model follow a normal distribution, which is one of the key assumptions in linear regression.

In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. The Q-Q plot is crucial for validating this assumption. Here's why it's important:

1. **Model Validation**: Checking the normality of residuals helps validate the linear regression model. Non-normal residuals can indicate issues with the model, such as omitted variables, incorrect functional form, or heteroscedasticity.
2. **Inference Accuracy**: Many inferential statistics (e.g., confidence intervals, hypothesis tests) rely on the normality assumption. If residuals are not normal, these statistics may be inaccurate.
3. **Identifying Outliers**: The Q-Q plot can help identify outliers or unusual observations that may unduly influence the model.