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 box plots generated from the exploratory data analysis (EDA), we can infer the effect of categorical variables on the target variable (count of bike rentals):

Year: The box plot for the year indicates an increase in bike rentals from one year to another. This suggests a positive trend in bike rentals over time.

Season: The box plot shows variations in bike rentals across different seasons. Typically, we observe higher bike rentals during seasons like fall, winter and summer compared to spring.

Month: Bike rentals vary across different months. Generally, we see higher rentals during the month of May to October.

Holiday: The box plot does not show much differences in bike rentals on holidays compared to non-holidays. It appears that bike rentals are slightly lower on holidays.

Weekday: Bike rentals vary depending on the day of the week. There might be higher rentals on certain weekdays compared to others, indicating potential differences in commuting patterns or recreational activities.

Weather Situation: The box plot for weather situation suggests that bike rentals tend to be lower during adverse weather condition like 2 and 3

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

These observations provide insights into how categorical variables such as year, season, month, holiday, weekday, and weather conditions can influence the count of bike rentals. They can help in understanding the factors affecting bike rental patterns and in making decisions related to bike rental management and marketing strategies.

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

Answer: Using drop_first=True during dummy variable creation is important to avoid multicollinearity issues in regression analysis, specifically when dealing with categorical variables with more than two levels.

For example, consider a categorical variable "season" with three levels: spring, summer, and fall. If we create dummy variables without dropping the first one, we'll have:

Dummy variable 1: 1 if spring, 0 otherwise

Dummy variable 2: 1 if summer, 0 otherwise

Dummy variable 3: 1 if fall, 0 otherwise

In the case of the "season" variable, dropping the first dummy variable means we only include two dummy variables:

Dummy variable 1: 1 if summer, 0 otherwise

Dummy variable 2: 1 if fall, 0 otherwise

Now, each dummy variable provides unique information about a specific category.

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

Answer: ‘**temp** and **atemp’** feature are highly corelated with the target variable ‘**cnt’**.

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

Answer: We have done the following assumptions:

1. There is a linear relationship between actual and predicted values of ‘cnt’
2. The error terms are evenly distributed

Question 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:

Year (yr): The coefficient for the "yr" variable is 0.2354 with a very low p-value ($p < 0.001$), indicating a highly significant positive effect on bike demand. This suggests that there has been a significant increase in bike rentals over time.

Temperature (temp): The coefficient for the "temp" variable is 0.4078 with a very low p-value ($p < 0.001$), indicating a highly significant positive effect on bike demand. This suggests that higher temperatures lead to increased bike rentals, which is intuitive as people are more likely to ride bikes in warmer weather.

Weather Situation (weathersit): Both "weathersit_2" (partly cloudy) and "weathersit_3" (rain/snow/fog) variables have significant coefficients with very low p-values ($p < 0.001$). However, their coefficients have negative values, indicating a negative effect on bike demand. This suggests that adverse weather conditions (partly cloudy, rain, snow, fog) lead to decreased bike rentals, which is reasonable as people may be less inclined to ride bikes in such weather conditions.

General Subjective

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

Answer: Detailed explanation of the linear regression algorithm:

1. **Problem Statement**: Linear regression is used to model the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the independent variables and the dependent variable.

2. **Model Representation**: The linear regression model is represented by the equation:

   $y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n+\epsilon$

   where:

   - y is the dependent variable (target),
   - $x1,x2,...,xn$ are the independent variables (features),
   - $\beta 0$ is the intercept,
   - $\beta 1,\beta 2,...,\beta n$ are the coefficients (slopes) corresponding to each independent variable,
   - ϵ is the error term representing the deviation of the observed values from the true linear relationship.

3. **Assumptions**: Linear regression relies on several assumptions:

   - Linearity: Assumes that the relationship between the independent variables and the dependent variable is linear.
   - Independence: Assumes that the observations are independent of each other.
   - Homoscedasticity: Assumes that the variance of the residuals (the differences between observed and predicted values) is constant across all levels of the independent variables.
   - Normality: Assumes that the residuals follow a normal distribution.
   - No multicollinearity: Assumes that the independent variables are not highly correlated with each other.

4. **Fitting the Model**:

   - The goal of linear regression is to estimate the coefficients $\beta 0,\beta 1,...,\beta n$ that minimize the sum of squared errors (SSE) between the observed and predicted values.
   - This is typically done using the method of least squares, where the coefficients are estimated to minimize the sum of the squared differences between the observed and predicted values.

5. **Model Evaluation**:

   - Once the model is fitted, it needs to be evaluated to assess its performance and validity.
   - Common evaluation metrics include R-squared (the proportion of variance explained by the model), adjusted R-squared, mean squared error (MSE), and root mean squared error (RMSE).

6. **Interpretation**:

- After fitting the model, the coefficients can be interpreted to understand the relationship between the independent variables and the dependent variable.
- The intercept $\beta 0$ represents the value of the dependent variable when all independent variables are zero.
- The coefficients $\beta 1, \beta 2, \ldots, \beta n$ represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.

7. **Predictions**:

    - Once the model is validated, it can be used to make predictions on new data by plugging in the values of the independent variables into the regression equation.

Overall, linear regression is a powerful and interpretable tool for modelling the relationship between variables, but it's essential to check the assumptions and validate the model to ensure its reliability and accuracy

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

Answer: Anscombe's quartet is a set of four small datasets that have nearly identical statistical properties, yet they appear very different when graphed.

Here's a detailed explanation of Anscombe's quartet:

Description of the Quartet:

Anscombe's quartet consists of four small datasets, each containing 11 data points. Despite having identical summary statistics (e.g., mean, variance, correlation coefficient), the datasets have different patterns and relationships when graphed.

Properties:

Dataset I: Linear relationship, no outliers.

Dataset II: Non-linear relationship, one outlier.

Dataset III: Linear relationship, one outlier with high leverage.

Dataset IV: Linear relationship, one outlier with high influence.

Statistical Properties:

Each dataset has the same mean, variance, correlation coefficient, and regression line parameters.

For example, all datasets have a mean of approximately 9 for both the x and y variables, a variance of approximately 11 for the x variable and 4 for the y variable, and a correlation coefficient of approximately 0.816 for the x and y variables.

Visualization:

When plotted, each dataset exhibits a different pattern, highlighting the importance of visualizing data.

Dataset I shows a clear linear relationship between x and y variables.

Dataset II shows a non-linear relationship with a quadratic curve.

Dataset III appears to have a linear relationship, but it is heavily influenced by an outlier with high leverage.

Dataset IV also appears to have a linear relationship, but it is heavily influenced by an outlier with high influence.

Implications:

Anscombe's quartet demonstrates that summary statistics alone may not be sufficient to fully understand the relationship between variables. It emphasizes the importance of graphical exploration and visualization in data analysis to uncover patterns, outliers, and relationships that may not be apparent from summary statistics alone. The quartet serves as a cautionary example against relying solely on summary statistics and highlights the value of exploratory data analysis (EDA) in understanding datasets.

Question 3. What is Pearson’s R? (3 marks)

Answer: Pearson's correlation coefficient, often denoted as r, is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It quantifies the degree to which the variables move together in a linear fashion.

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

Answer: Scaling is a preprocessing technique used in machine learning and data analysis to standardize the range of features or variables. It involves transforming the data such that it falls within a specific range or distribution. The goal of scaling is to ensure that all features contribute equally to the analysis and to improve the performance of machine learning algorithms.

Scaling is used for:

1. ***Equal Weights***: Many machine learning algorithms use distance-based calculations (e.g., Euclidean distance) to measure the similarity between data points. If features have different scales, those with larger scales may dominate the calculation, leading to biased results. Scaling ensures that all features contribute equally to the analysis.
2. ***Faster Convergence***: Some optimization algorithms (e.g., gradient descent) converge faster when features are on similar scales. Scaling can help speed up the convergence process and improve the efficiency of training.

3. ***Model Interpretability***: Scaling can make the coefficients or weights of the features more interpretable. It ensures that the coefficients represent the impact of the features on the outcome in a meaningful and comparable way.

Differences between Normalized Scaling and Standard Scaling:

1. Normalized scaling scales the features to a fixed range (e.g., 0 to 1), while standardized scaling centers the features around the mean and scales them based on the standard deviations.
2. Normalized scaling maintains the relative relationships between data points but does not center the data around zero.
3. Standardized scaling centers the data around zero and maintains the shape of the original distribution, making it more suitable for algorithms that assume Gaussian-distributed data.

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

Answer: This typically occurs when there is perfect multicollinearity among the independent variables in the regression model.

Perfect multicollinearity occurs when one or more independent variables in the regression model can be exactly predicted by a linear combination of other independent variables.

In other words, one independent variable is a perfect linear function of the others, leading to a situation where the coefficients cannot be uniquely estimated.

Questions 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 (Quantile-Quantile) plot is a graphical tool used to assess whether a given set of data follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the observed data to the quantiles of a theoretical distribution, typically a normal distribution. Q-Q plots are commonly used in statistics to visually inspect the distribution of data and to assess assumptions made in statistical analyses, such as linear regression.

It can be used as follows:

Assumption Checking:

In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) follow a normal distribution.

Q-Q plots are used to visually inspect whether the residuals follow a normal distribution. If the residuals are normally distributed, the points on the Q-Q plot will lie approximately along the straight line.

Departures from the straight line indicate violations of the normality assumption, which can impact the validity of the regression analysis.

Residual Analysis:

Q-Q plots help identify patterns or trends in the residuals that may indicate violations of other regression assumptions, such as homoscedasticity (constant variance) or linearity.

For example, if the residuals exhibit non-linear patterns in the Q-Q plot, it may suggest non-linear relationships between the independent and dependent variables.

Model Diagnostics:

Q-Q plots are an essential diagnostic tool for assessing the adequacy of the regression model and identifying areas for improvement.

They provide insights into the distributional properties of the residuals and guide decisions regarding model refinement or transformation of variables.