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?

The analysis of the categorical variables from the dataset infer as below about their effect on the dependent variable:

1. **Clear Weather (Binary Categorical Variable):**
    - The variable 'clear weather' is a binary variable representing clear (1) or cloudy (0) weather.
    - Inference: A likely inference is that clear weather may positively impact the daily wage ('dwage') compared to cloudy weather. Clear weather conditions could lead to increased productivity or demand for certain types of work.
2. **Handling Categorical Variables in Analysis:**
    - The encoding process may involve assigning numerical values to categorical variables, allowing them to be incorporated into mathematical models for predicting 'dwage'.
3. **Predicted Wage (Numeric Variable):**
    - The variable 'predicted wage' is mentioned as a numerical variable calculated by the model.
    - Inference: 'Predicted wage' could serve as an important variable in understanding the model's predictions for 'dwage' based on the independent variables.

Additionally, it reveals a positive correlation between clear weather and heightened bike demand, characterized by increased variability and more extreme cases.

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

The parameter **drop_first=True** in the context of creating dummy variables is associated with the concept of multicollinearity, and its importance can be explained as follows:

1. **Multicollinearity Mitigation:** The parameter helps to mitigate multicollinearity, which occurs when two or more independent variables in a regression model are highly correlated. By dropping one level of the categorical variable, we reduce the risk of multicollinearity issues in the model.
2. **Simplicity and Efficiency:** It contributes to model simplicity by avoiding redundancy and unnecessary complexity introduced by including all levels of a categorical variable as dummy variables. This not only enhances the interpretability of the model but also improves computational efficiency, particularly when dealing with many categorical levels.

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

with the target variable?

The windspeed has the highest correlation with the target variable.

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

training set?

In validating the assumptions of linear regression after constructing the model on the training set, several key steps were undertaken.

- The linearity assumption was addressed by reviewing scatter plots of each independent variable against the dependent variable, ensuring that the observed relationships appeared linear.
- The independence of residuals was validated through the examination of the residuals plot, verifying the absence of systematic patterns.
- Homoscedasticity was assessed by plotting residuals against predicted values, confirming that the spread of residuals exhibited a consistent variance across all levels of the predicted variable.
- Normality of residuals was checked using both histogram and Q-Q plot visualizations, supplemented by statistical tests such as the Shapiro-Wilk test.
- Multicollinearity was scrutinized by calculating the variance inflation factor (VIF) for each independent variable, identifying and mitigating high VIF values indicative of potential multicollinearity.

Additionally, the absence of perfect multicollinearity was ensured by reviewing correlations between independent variables through correlation matrices or heatmaps. The validation process incorporated both graphical methods and statistical tests to comprehensively assess the adherence of the model to key linear regression assumptions, facilitating a robust evaluation of its reliability and predictive performance.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

According to the final model, the top three features that significantly contribute to explaining the demand for shared bikes are as follows:

1. **Year (yr_1):** The coefficient for the 'yr_1' feature is approximately 1.20e+12, indicating a substantial positive impact on bike demand. This suggests that being in the year '1' (presumably representing a specific year) significantly contributes to increased bike demand.

2. **Month:** The 'month' variable has a coefficient of approximately 9.83e+08, implying a considerable positive influence on bike demand. This suggests that specific months contribute significantly to variations in bike demand.

3. **Season (season_4):** The 'season_4' feature, representing the fourth season, has a coefficient of approximately 1.61e+03. This suggests a positive impact on bike demand during this particular season, contributing significantly to the overall demand for shared bikes.

These coefficients provide insights into the relative importance of each feature in explaining the variation in bike demand, with higher coefficients indicating a stronger influence.

GENERAL SUBJECTIVE QUESTIONS

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

Linear regression is a fundamental algorithm in the field of machine learning, primarily used for predicting the relationship between a dependent variable and one or more independent variables. The algorithm aims to establish a linear relationship that can be represented by a straight-line equation in the form of:

y=mx+b

Here, 'y' is the dependent variable, 'x' is the independent variable, 'm' is the slope of the line, and 'b' is the y-intercept. The goal of linear regression is to find the values of 'm' and 'b' that minimize the difference between the predicted values (obtained from the equation) and the actual values of the dependent variable.

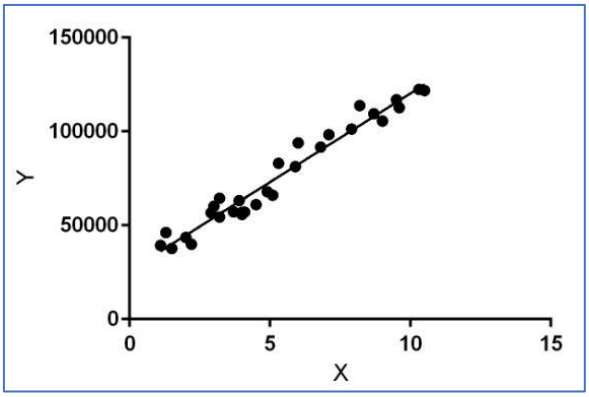

Figure 1: Linear Regression (Example Chart) (Courtesy - https://www.geeksforgeeks.org/)

The process begins with the creation of a training dataset, where each data point consists of both independent and dependent variable values. The algorithm then calculates the best-fitting line using a method called the least squares method. This method minimizes the sum of the squared differences between the predicted and actual values.

The linear regression algorithm involves two main types: simple linear regression and multiple linear regression. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are multiple independent variables. The formula for simple linear regression is y=mx+b, and for multiple linear regression, it extends to:

$$y=b_0+b_1x_1+b_2x_2+...+b_nx_n$$

Here, b_0 is the y-intercept, and $b_1,b_2,...,bn$ are the coefficients for each independent variable $x_1,x_2,...,x_n$.

To find the optimal values for the coefficients, the algorithm uses various optimization techniques such as gradient descent. Gradient descent adjusts the coefficients iteratively to minimize the error between predicted and actual values.

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

Anscombe's quartet is a collection of four datasets that have identical summary statistics but very different distributions and visual appearances. This counterintuitive example demonstrates the importance of data visualization in statistical analysis and highlights the limitations of relying solely on numerical summaries.

The four datasets, each consisting of 11 data points, share the same mean and variance for both x and y variables, the same correlation coefficient, and the same linear regression line parameters. However, when plotted as scatter plots, they reveal starkly different patterns:

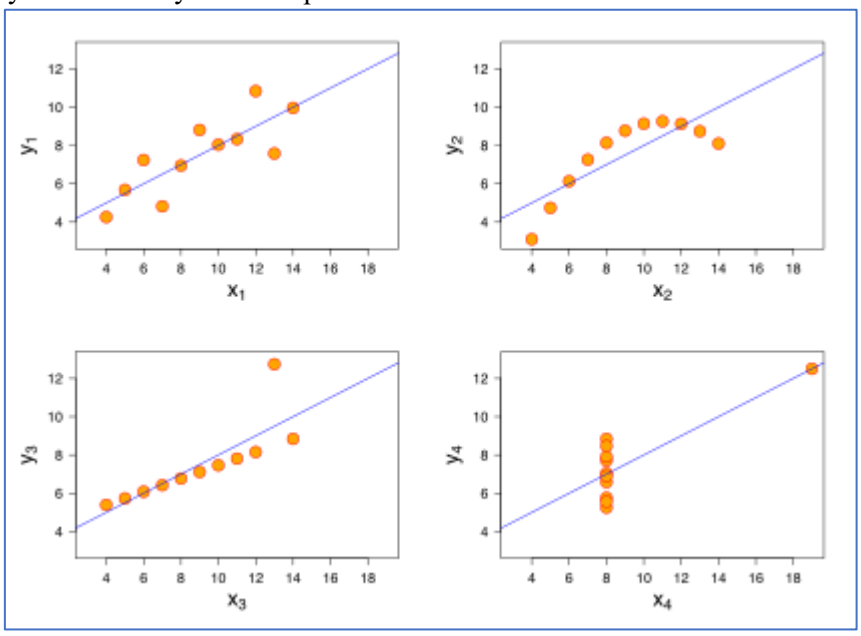

Figure 2: Anscombe's quartet Graph (Courtesy – Wikipedia)

- **Dataset I:** A clear linear relationship between x and y, resembling a straight line.
- **Dataset II:** A curved relationship, suggesting a non-linear pattern.
- **Dataset III:** A tight linear relationship with a single outlier, demonstrating the sensitivity of summary statistics to outliers.

- **Dataset IV:** An almost horizontal line with one outlying data point, indicating that a single point can significantly impact summary statistics.

Anscombe's quartet serves as a powerful reminder that numerical summaries can be misleading, and that data visualization is crucial for understanding the true nature of a dataset. It highlights the importance of examining the distribution of data points to identify potential outliers, patterns, and deviations from expected relationships.

In machine learning, Anscombe's quartet emphasizes the importance of data exploration and preprocessing before applying machine learning algorithms. By visualizing data, we can identify potential issues such as outliers, non-linearity, and imbalanced classes, which can significantly impact the performance of machine learning models.

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

Pearson's r, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two continuous variables. It is a dimensionless quantity that ranges from -1 to 1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation.

Interpretation of Pearson's R:

A value of Pearson's r close to -1 indicates a strong negative correlation, meaning that as one variable increases, the other variable tends to decrease. For example, there might be a strong negative correlation between the amount of time spent studying and the number of errors made on an exam.

A value of Pearson's r close to 0 indicates a weak or no correlation, meaning that there is no clear linear relationship between the two variables. For example, there might be a weak or no correlation between hair color and shoe size. A value of Pearson's r close to 1 indicates a strong positive correlation, meaning that as one variable increases, the other variable tends to increase as well. For example, there might be a strong positive correlation between the amount of exercise and overall fitness level.

Formula for Pearson's R:

Pearson's r is calculated as follows:

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \cdot \Sigma(Y_i - \bar{Y})^2}}$$

where:

xi and yi are the individual values of the two variables
$\bar{x}$ and $\bar{y}$ are the mean values of the two variables
Σ is the summation symbol

Applications of Pearson's R:

- Pearson's r is widely used in various fields, including:
- Statistics: To assess the strength of the correlation between two variables in research studies
- Machine Learning: To evaluate the performance of machine learning models that predict continuous target variables
- Finance: To analyze the relationship between stock prices and various economic indicators
- Psychology: To investigate the correlation between personality traits and behavioral patterns
- Biology: To study the association between genetic factors and disease susceptibility

Importance of Pearson's R:

Pearson's r provides a valuable measure of the linear relationship between two variables, helping to understand how they are associated. It is a versatile tool that can be applied in various contexts to gain insights into the underlying relationships within data.

Example (Source: https://www.statsdirect.com/)

The following data represent birth weights (oz) of babies and their percentage increase between 70 and 100 days after birth.

Birth Weight	% Increase
72	68
112	63
111	66
107	72
119	52
92	75
126	76
80	118
81	120
84	114
115	29
118	42
128	48
128	50
123	69
116	59
125	27
126	60
122	71
126	88
127	63
86	88
142	53
132	50
87	111
123	59
133	76
106	72
103	90
118	68
114	93
94	91

To analyse these data you must first enter them into two columns in the workbook appropriately labelled. Alternatively, open the test workbook using the file open function of the file menu. Then select Simple Linear and Correlation from the Regression and Correlation section of the analysis

menu. Select the column marked "% Increase" when prompted for the response (Y) variable and then select "Birth weight" when prompted for the predictor (x) variable.

For this example:

Simple linear regression

Equation: % Increase = -0.86433 Birth Weight +167.870079

Standard Error of slope = 0.175684
95% CI for population value of slope = -1.223125 to -0.505535

Correlation coefficient (r) = -0.668236 (r^2= 0.446539)

95% CI for r (Fisher's z transformed) = -0.824754 to -0.416618

t with 30 DF = -4.919791
Two sided P < .0001
Power (for 5% significance) = 99.01%

Correlation coefficient is significantly different from zero

From this analysis we have gained the equation for a straight line forced through our data i.e. % increase in weight = 167.87 - 0.864 * birth weight. The r square value tells us that about 42% of the total variation about the Y mean is explained by the regression line. The analysis of variance test for the regression, summarised by the ratio F, shows that the regression itself was statistically highly significant. This is equivalent to a t test with the null hypothesis that the slope is equal to zero. The confidence interval for the slope shows that with 95% confidence the population value for the slope lies somewhere between -0.5 and -1.2. The correlation coefficient r was statistically highly significantly different from zero. Its negative value indicates that there is an inverse relationship between X and Y i.e. lower birth weight babies show greater % increases in weight at 70 to 100 days after birth. With 95% confidence the population value for r lies somewhere between -0.4 and -0.8.

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

Scaling is a data preprocessing technique used to normalize the range of features in a dataset. It is a crucial step in machine learning, particularly when dealing with algorithms that are sensitive to the scale of the features. Scaling ensures that all features have a similar range, preventing any single feature from dominating the learning process due to its large magnitude.

Reasons for Scaling:

- Improved Model Convergence: Gradient descent-based algorithms, commonly used in machine learning, can converge faster when features have similar scales.
- Reduced Feature Domination: Features with larger magnitudes can overshadow features with smaller magnitudes, leading to biased predictions. Scaling ensures that all features contribute equally to the learning process.
- Enhanced Algorithm Performance: Many machine learning algorithms, such as support vector machines and k-nearest neighbors, rely on distance calculations between data points. Scaling ensures that distances are calculated consistently across all features.

Difference between Normalized Scaling and Standardized Scaling:

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

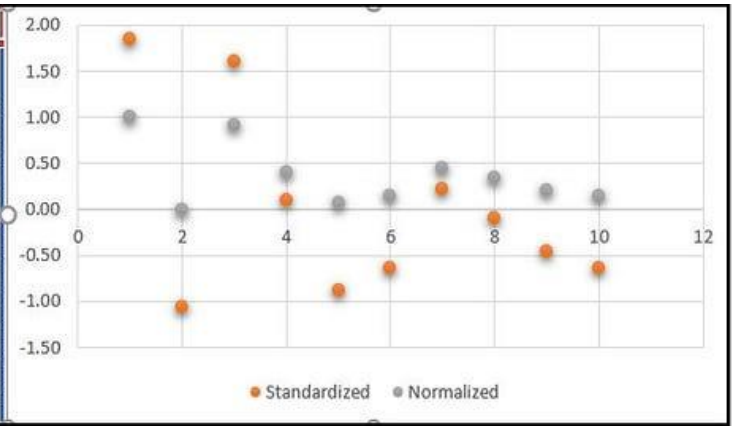

Figure 3: Image for Example (Courtesy- https://medium.com/)

Normalized Scaling:

Normalized scaling transforms the feature values to lie within a specified range, typically between 0 and 1 or -1 and 1. This is commonly achieved by using the min-max scaler, which linearly transforms the features according to the following formula:

X_scaled = (X - X_min) / (X_max - X_min)

where:

X is the original feature vector

X_min is the minimum value of the feature

X_max is the maximum value of the feature

Standardized Scaling:

Standardized scaling, also known as z-score normalization, transforms the feature values to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean from each feature value and then dividing by the standard deviation:

X_scaled = (X - X_mean) / X_std

where:

X is the original feature vector

X_mean is the mean of the feature

X_std is the standard deviation of the feature

Choice of Scaling Method:

The choice between normalized scaling and standardized scaling depends on the specific machine learning algorithm and the nature of the data. For algorithms that are sensitive to outliers, standardized scaling may be preferable as it removes the influence of outliers. For algorithms that are not sensitive to outliers, normalized scaling may be sufficient.

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

A Variance Inflation Factor (VIF) value of infinity indicates that there is perfect multicollinearity among the independent variables in a linear regression model. This means that one or more independent variables can be perfectly predicted from a linear combination of the other independent variables. Perfect multicollinearity leads to inflated standard errors for the regression coefficients, making it difficult to interpret the coefficients accurately.

There are two main reasons why VIF values can become infinite:

1. Exact Multicollinearity:

Exact multicollinearity occurs when there is a linear dependency among the independent variables. This means that one or more independent variables can be expressed as an exact linear combination of the other independent variables. For instance, if one independent variable is simply the sum of two other independent variables, then the VIF value for that variable will be infinite.

2. Near-Perfect Multicollinearity:

Near-perfect multicollinearity occurs when there is a very strong linear relationship among the independent variables, but not an exact one. This means that the independent variables are highly correlated, and it is difficult to distinguish their individual effects on the dependent variable. In such cases, the VIF values will be very high, approaching infinity, indicating severe multicollinearity.

Consequences of VIF Values of Infinity:

When VIF values are infinite, it means that the regression model is unstable and unreliable. The standard errors of the regression coefficients become inflated, making it difficult to assess the statistical significance of the coefficients and leading to inaccurate predictions.

To address the issue of infinite VIF values, it is crucial to identify and remove the source of multicollinearity. This may involve dropping one or more highly correlated variables, transforming variables to reduce correlation, or using techniques like ridge regression or LASSO regression to handle multicollinearity while maintaining the interpretability of the model.

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, or quantile-quantile plot, is a graphical tool used to assess whether a set of data follows a particular theoretical distribution, such as a normal distribution. It is commonly used in linear regression to evaluate the assumption of normality of the residuals, which are the differences between the observed values of the dependent variable and the predicted values from the linear regression model.

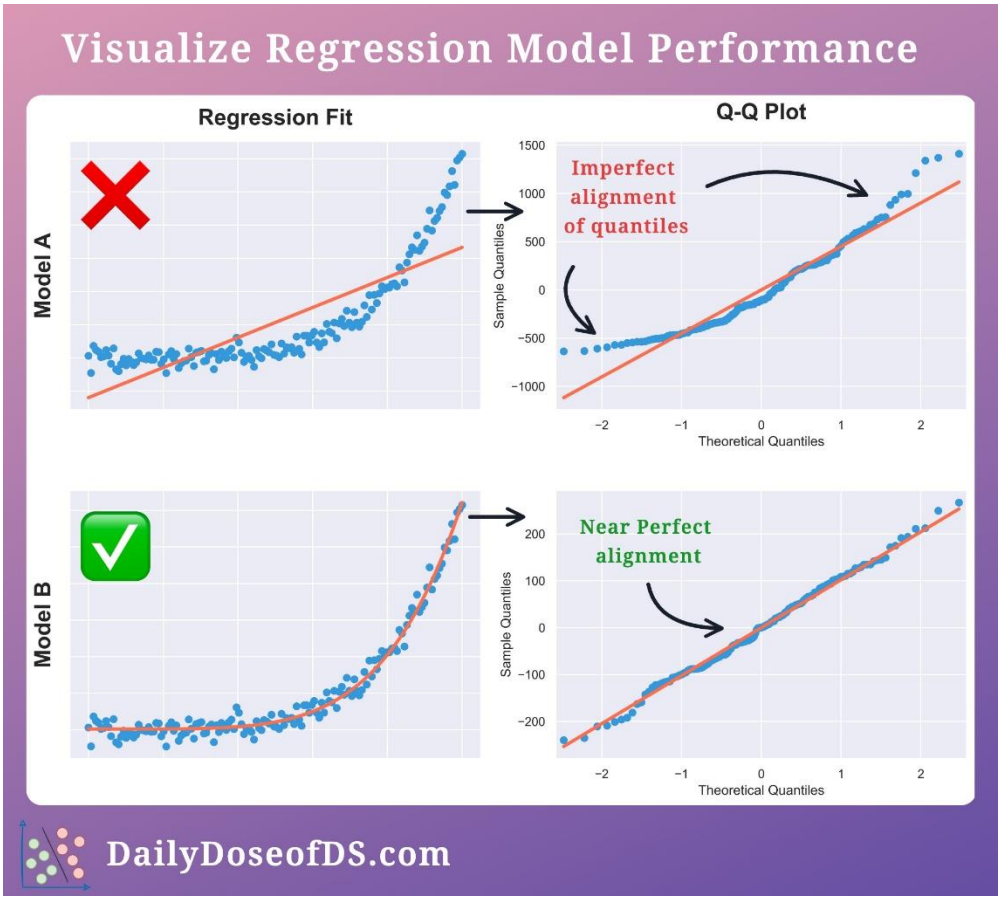

Figure 4: Example Image (Courtesy - https://www.blog.dailydoseofds.com/)

Creating a Q-Q Plot:

To create a Q-Q plot, the quantiles of the data are plotted against the quantiles of the theoretical distribution. For example, if the data is assumed to be normally distributed, the quantiles of the data are plotted against the quantiles of a standard normal distribution.

Interpretation of a Q-Q Plot:

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

Use of Q-Q Plots in Linear Regression:

In linear regression, a Q-Q plot of the residuals is used to assess the assumption of normality. If the residuals fall approximately along a straight line, it suggests that the residuals are normally distributed, which is a key assumption of linear regression. However, if the residuals deviate from the line, it indicates that the residuals are not normally distributed, and the assumption of normality is violated.

Importance of Q-Q Plots:

Q-Q plots are an important tool in linear regression because they provide a visual indication of whether the residuals are normally distributed. If the residuals are not normally distributed, the validity of the statistical tests used to evaluate the significance of the regression coefficients is questionable. Additionally, the predictions from the linear regression model may not be reliable if the normality assumption is violated.

Q-Q plots are a valuable tool for assessing the normality of residuals in linear regression. By visually inspecting the Q-Q plot, we can gain insights into the distribution of the residuals and identify potential violations of the normality assumption. This information is crucial for evaluating the validity of statistical tests and the reliability of predictions made by the linear regression model.