**Assignment-based Subjective Questions**

**Question 1.** From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

Some of the categorical variables and their effect on average daily bike rentals are:

* Season: There is high variation in average daily bike rentals by season, with spring having the lowest rental volume, and fall having the highest.
* Month: Average daily bike rentals are highest in the June-September period, and lowest in the Jan-Feb period. This is corroborated by the insight derived from the season categorical variable.
* Weekday: The average daily bike rentals don’t vary a lot by day of the week. Tuesday has the lowest (~4229/day), while Sunday has the highest (~4690/day).
* Weather: Average daily bike rentals is highest in clear conditions and lowest in light snow/rain conditions. None of the days had heavy rain/ice pellets weather conditions, so there is no insight to be derived for that weather condition.
* Working day: Average daily bike rentals is similar on working and non-working days.

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

drop\_first=True removes one instance of the one-hot encoded dummy variable which is completely explained by the other dummy variables for that particular categorical variable. This is done to avoid multicollinearity.

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

According to the pair-plot, both temperature and feeling temperature have very similar correlation with daily count of bikes rented. This is also supplemented by the fact that temperature and feeling temperature themselves are very highly correlated with each other.

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

* Linearity: Checked scatter plots of residuals vs. each predictor for random scatter, indicating a linear relationship.
* Independence: Examined the Durbin-Watson statistic; values close to 2 suggested no autocorrelation in residuals.
* Homoscedasticity: Conducted the Breusch-Pagan test; a high p-value indicated constant variance of residuals.
* Normality of Residuals: Used a Q-Q plot; points following the 45-degree line indicated normally distributed residuals.
* Multicollinearity: Calculated VIFs for predictors; VIF values above 5 indicated significant multicollinearity.
* No perfect multicollinearity: Checked the correlation matrix; correlations close to 1 or -1 indicated perfect multicollinearity.

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top 3 independent variables of the model are:

* Weathersit\_3 (i.e. whether it is light snow/light rain or not
* Yr (whether year is 2018 or 2019)
* Mnth\_September (September has high predictability of average daily bike rentals)

**General Subjective Questions**

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Linear regression is a statistical model that estimates the linear relationship between a dependent variable and one or more independent variables. The steps in linear regression are:

* Data Collection: Gather data with dependent and independent variables
* Data Pre-processing: Missing values, one-hot encoding, scaling
* Splitting data: 80-20 or 70-30 train-test split
* Model building: Fit the regression model using training data, estimate coefficients using OLS or other methods
* Model evaluation: Use parameters such as R-squared, adjusted R-squared, RMSE, etc.
* Check assumptions: Linearity, independence, homoscedasticity, normality, no multicollinearity
* Prediction: Use model to predict outcome for test data

**Question 7.** Explain the Anscombe’s quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Anscombe’s quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. All four datasets share these summary statistics:

* Mean of x
* Mean of y
* Variance of x
* Variance of y
* Correlation between x and y
* Linear regression equation between x and y

But the four datasets reveal different patterns when graphed:

* Dataset 1: Shows a typical linear relationship with some scatter around the regression line
* Dataset 2: Forms a perfect quadratic relationship; the linear regression line does not fit well
* Dataset 3: Contains an outlier that strongly influences the regression line
* Dataset 4: All points are nearly identical except for one extreme outlier, which affects the regression line significantly

Anscombe’s quartet highlights several critical lessons in data analysis:

1. **Graphical Analysis**: Visualizing data is crucial before conducting statistical analysis; graphs can reveal patterns, relationships, and anomalies that summary statistics might obscure
2. **Data Interpretation**: Summary statistics alone can be misleading; it is essential to consider the context and the data distribution
3. **Influence of Outliers**: Outliers can significantly impact statistical measures and regression models; identifying and understanding outliers is an essential step in data analysis.
4. **Robust Analysis**: Relying solely on numerical summaries for decision-making can lead to incorrect conclusions

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Pearson’s r/ Pearson’s correlation coefficient is a measure of strength and direction of the linear relationship between two continuous variables. It lies between -1 and 1 (inclusive). It is symmetrical (corr(x,y)=corr(y,x)), unit-free, and measures only linear relationship (it fails to capture let’s say quadratic relationship for example).

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Scaling is the process of adjusting the range of feature values in your data to a standard range, typically to improve the performance and training stability of machine learning models. It ensures that each feature contributes equally to the model's performance by bringing them to a similar scale. Scaling is performed to:

* Improve model performance
* Prevent numerical issues
* Ensure equal contribution of features

|  |  |  |
| --- | --- | --- |
| Type of scaling | Normalised | Standardised |
| Definition | Technique that rescales feature values to a specific range, typically [0, 1] or [-1, 1] | Technique that transforms feature values to have a mean of 0 and a standard deviation of 1 |
| Range | Transforms data into a fixed range, usually [0, 1] or [-1, 1] | Transforms data to have a mean of 0 and a standard deviation of 1, effectively standardising the spread of the data |
| Application | * Useful when the distribution of the data is not Gaussian * Often applied when the algorithm does not assume any data distribution, such as in k-nearest neighbors and neural networks | * Useful when the data follows a Gaussian (normal) distribution * Often applied in algorithms that assume data is normally distributed, such as linear regression, logistic regression, and principal component analysis (PCA) |
| Effect on data | * Preserves the original relationships between data points * Compresses the spread of the data, making the features comparable on the same scale * Does not change the shape of the data distribution but rescales it | * Centers the data around the mean, making the mean 0 * Scales the data according to the variance, making the standard deviation 1 * Changes the shape of the data distribution by adjusting for variance and mean, making features comparable in terms of their spread |

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

VIF can become infinite in certain situations due to perfect multicollinearity among the predictor variables in a multiple regression model. This occurs when one predictor variable in a multiple regression model can be perfectly explained as a linear combination of one or more of the other predictor variables.

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

A Q-Q plot is primarily used to validate the assumption of normality for the residuals. A Q-Q plot helps visually check this assumption by plotting the residuals against a normal distribution. If the residuals deviate significantly from the straight line in the Q-Q plot, this could indicate that the residuals are not normally distributed, which can impact the validity of statistical tests and confidence intervals.