**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)
2. To infer the effect of categorical variables on the dependent variable in a dataset, we would typically perform statistical analysis such as
3. **Frequency Distribution**
4. **Bar Charts**
5. **Chi-Square Test**
6. **ANOVA (Analysis of Variance)**
7. **Regression Analysis**
8. **Covariates**

the specific analysis and interpretation will depend on the nature of our data, research question and the statistical methods we choose to use. It's crucial to consider both statistical significance and practical significance when drawing conclusions about the effect of categorical variables on the dependent variable.

1. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)
2. Using “**drop\_first=True”** during dummy variable creation is important for two main reasons:
3. **Avoiding Multicollinearity.**
4. **Interpretability**
5. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

A. looking at the test data set we find that the highest co relation is between spring and temperature(-0.61) but in negative.

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

A. common ways to validate these assumptions:

1. Residual Analysis.

2. Linearity Assumption.

3. **Homoscedasticity (Constant Variance).**

4. **Multicollinearity.**

**5. VIF(Variance inflation factor).**

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

A. Temperature , Light snow/rain and year

**General Subjective Questions**

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

A. Linear regression is a widely used statistical method for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The fundamental idea behind linear regression is to find the best-fitting straight line (or hyperplane, in the case of multiple independent variables) that minimizes the sum of squared differences between the observed and predicted values. Here's a detailed explanation of the linear regression algorithm:

**1. Data Preparation:**

* Gather your dataset, which should include one or more independent variables (predictors) and a dependent variable (target) that you want to predict.
* Preprocess the data by handling missing values, scaling or standardizing variables if necessary, and splitting it into training and testing sets for model evaluation.

**2. Linear Equation:**

* In simple linear regression (with one independent variable), the linear equation takes the form:

makefileCopy code

y = b0 + b1 \* x + ε

* + **y** is the dependent variable (target).
  + **x** is the independent variable (predictor).
  + **b0** is the intercept, representing the predicted value of **y** when **x** is 0.
  + **b1** is the slope, representing the change in **y** for a one-unit change in **x**.
  + **ε** represents the error term, accounting for the noise or unexplained variance in the data.

**3. Model Training:**

* The goal of training is to find the values of **b0** and **b1** that minimize the sum of squared residuals (errors). Residuals are the differences between the observed **y** values and the predicted values (given by the linear equation).
* This minimization is often achieved using the method of least squares, which calculates **b0** and **b1** as follows:

scssCopy code

b1 = Σ((xi - x̄) \* (yi - ȳ)) / Σ((xi - x̄)^2) b0 = ȳ - b1 \* x̄

* + **xi** and **yi** are the individual data points.
  + **x̄** and **ȳ** are the means of the independent and dependent variables, respectively.

**4. Model Evaluation:**

* After training, you need to assess the model's performance. Common evaluation metrics for linear regression include:
  + Mean Squared Error (MSE): Measures the average squared difference between observed and predicted values.
  + Root Mean Squared Error (RMSE): The square root of MSE, providing error in the original units.
  + R-squared (R²): Measures the proportion of variance in the dependent variable explained by the model.

**5. Model Prediction:**

* Once the model is trained and evaluated, you can use it to make predictions on new, unseen data. Plug new values of **x** into the linear equation to predict **y**.

**6. Assumptions:**

* Linear regression assumes that there is a linear relationship between the independent variables and the dependent variable.
* It assumes that the errors (residuals) are normally distributed and have constant variance (homoscedasticity).
* Independence of errors assumption implies that the residuals are not correlated.

**7. Extensions:**

* Linear regression can be extended to multiple independent variables, leading to multiple linear regression.
* Regularization techniques like Ridge and Lasso regression are used to prevent overfitting by adding regularization terms to the least squares objective function.

1. Explain the Anscombe’s quartet in detail. (3 marks)
2. Anscombe's quartet is a famous example in statistics that demonstrates the importance of not relying solely on summary statistics or visualizations when analyzing data. It was created by the statistician Francis Anscombe in 1973 to emphasize the concept that datasets with very different characteristics can produce identical or nearly identical summary statistics. The quartet consists of four small datasets, each containing 11 data points, but they have remarkably different properties when graphed or analyzed.
3. What is Pearson’s R? (3 marks)
4. Pearson's correlation coefficient, often denoted as "r" or "Pearson's R," is a statistic that quantifies the strength and direction of the linear relationship between two continuous variables. It measures how well the data points of two variables fit a straight line. Pearson's correlation coefficient can take values between -1 and 1,

where: *r*=(∑*i*=1*n*​(*Xi*​−*X*ˉ)2∑*i*=1*n*​(*Yi*​−*Y*ˉ) / sqrt(​∑*i*=1*n*​(*Xi*​−*X*ˉ)(*Yi*​−*Y*ˉ))^2

1. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
2. Scaling is a data preprocessing technique used in machine learning and statistics to transform the values of numerical variables to a common scale without changing the underlying relationships between the variables. Scaling is performed to ensure that all variables have a similar influence on model training and to prevent issues that can arise when variables are on different scales. The two common methods of scaling are normalized scaling and standardized scaling, and they serve slightly different purposes:

**Normalized Scaling (Min-Max Scaling):**

* In normalized scaling, also known as min-max scaling, the values of the variable are transformed to a specific range, typically between 0 and 1.
* The formula for min-max scaling is as follows:

X\_normalised = (X -Xmin) / (Xmax – Xmin)

**Standardized Scaling (Z-Score Standardization):**

* In standardized scaling, the values of the variable are transformed to have a mean (average) of 0 and a standard deviation of 1.
* The formula for standardization is as follows:

X\_standardized = (X – Xmean) / Xstd

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

A. The phenomenon of the Variance Inflation Factor (VIF) becoming infinite can occur in certain situations when calculating VIF values for predictor variables in a multiple linear regression model. This happens when there is perfect multicollinearity among a set of predictor variables.

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

A. A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles (ordered values) of the observed data against the quantiles of the expected theoretical distribution. Q-Q plots are widely used in statistics and data analysis to check the assumption of normality and to identify departures from the assumed distribution.