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?

   Answer:

   categorical variables-season,year,month,holiday,workingday,weekday,weathersit

   a. Bike rentals are high in the fall season

   b..Bike rentals have increased almost by 2000 from 2018 to 2019

   c..Bike rentals are high in the month of September and low in the month of January

   d..Bike rentals are almost same on both holidays and not a holiday but the upper variation is high in case of not a holiday

   e.With respect to workingday,there is no significant difference

   f.weekdays doesn't show any difference except that the lower variation is high on Monday and Friday

   g.Bike rentals are high on clear weather compared to cloudy and snow/rain days

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

   Answer:

   a.To avoid multicollinearity ,specifically dummy variable trap,we make k-1 dummies for categorical variable with k categories.

   b.Dummy variable trap is if we create k dummies from a categorical variable with k categories,including all k dummies in the regression model introduces perfect multicollinearity because the sum of all the dummies will always be one .

   For example, if there is a season variable-winter,spring,summer and fall.if we can take only three variables as winter is 100,spring is 010 and summer is 001. We don't need fall as 000 indicates fall.

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

   Answer:

   1.temp and atemp looks like having positive correlation with cnt

   2.casual and registered looks highly correlated which is expected because cnt is the sum of casual and registered.

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

   Answer:

   Assumption1:

   Linear relationship-Plotted a scatterplot between target varaiable(cnt) and other fearures to see if there is linear relationship-temp and atemp looks like having linear relationship with cnt.

   Assumption2:

Normality-plotted distplot and checked for the normality-distplot of numerical variables shows almost normal distribution(gaussian curve). Also the distplot of the residuals also shows normal distribution(gaussian curve)
Assumption3:
Multicollinearity-made sure are no linear relationship between the predictor variables using Variation Inflation Factor(VIF)
Assumption4:
Homoscedasticity(variance of residuals is constant)-plotted scatterplot between residuals and predicted values . the residuals are randomly dispersed around the horizontal line (y=0) with no apparent pattern.
Assumption5:
Independence(no autocorrelation in the residuals)-used Durbin-Watson Test-the Durbin-Watson statsistic was close to 2, which confirmed that residuals were not correlated.

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

Answer:
Based on the final model , the top three features are,
1.temp (beta coefficient-0.4677)
2.month_september(beta coefficient-0.794)
3.month_october(beta coefficient-0.0668)

General Subjective Questions

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

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.
1.Simple Linear Regression(one dependent and one independent variable)
2.Multiple Linear Regression(one dependent and two or more independent variables)

The Linear Regression Model:

Assuming there is a linear relationship between dependent variable Y and independent variables ,
Y=β0+β1X1+β2X2+⋯+βnXn+ϵ
Where:
Y is the dependent variable.
X1,X2,…,XnX_1, X_2, \ldots, X_nX1,X2,…,Xn are the independent variables.
β0 is the intercept.
β1,β2,…,βn are the coefficients of the independent variables.
ϵ is the error term, representing the difference between the observed and predicted values of Y.

Objective:

The objective of linear regression is to find the values of the coefficients β0,β1,β2,…,βn that minimizes the sum of squared differences(residuals) between the observed and predicted values of Y.

Step1: Data Collection and understanding

Gather the data that includes both the dependent and independent variables and understand the data

Step2: Data Preprocessing

1.Handling Missing values
2.Encode categorical varibles by creating dummies
3.Handling outliers

Step3: Splitting the dataset

Split the data set to training data and testing data sets to evaluate the model's performance.
Normalize the train data.
Normalize the test data with respect to traindata

Step4: Model Building

Using recursive elimination technique filter the top features that are need for linear model

Fit the linear regression model to the training data. This involves finding the optimal coefficients β using methods like Ordinary Least Squares (OLS). the model should have pvalue<0.05 and VIF values of all the varaibles should be <5.

Step5: Residual Analysis of train data:

Check if the error term(residuals) is normally distributed using distplot

Step6: Model Prediction:

Use the fitted model to make predictions on the test data

Step7: Model Evaluation:

Model is evaluated using R-squared,Adj.R-squared ,MSE (Mean Square Error),RMSE(Root Mean Square Error)

Now, you can use the built model to evaluate the new data

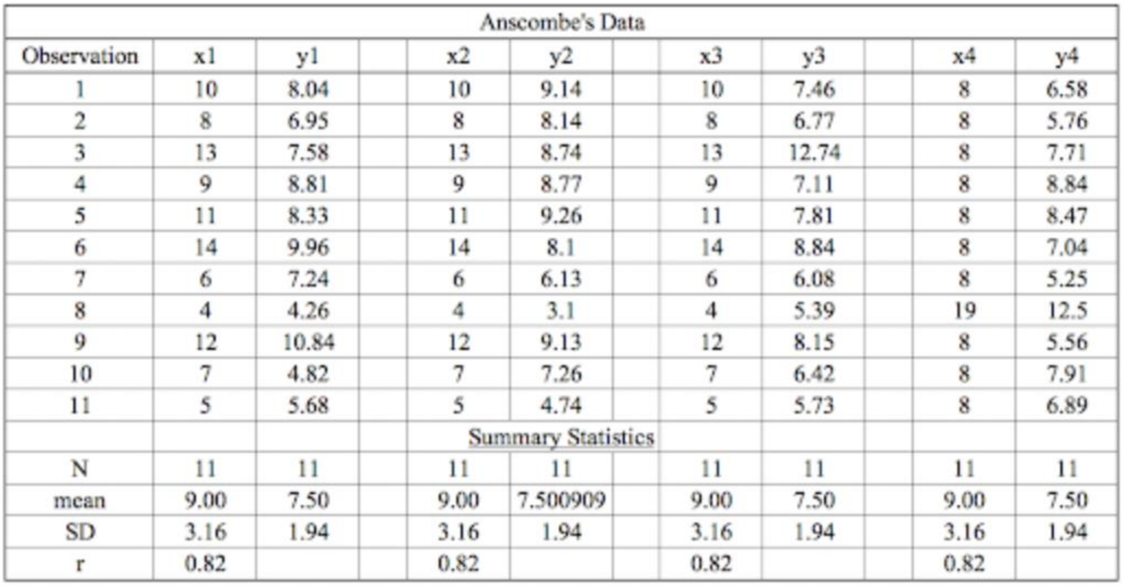
2. Explain the Anscombe's quartet in detail. (3marks)

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. The quartet demonstrates the importance of graphing data before analyzing it and the dangers of relying solely on summary statistics.

The Four Datasets

Each dataset consists of eleven (x, y) points. Despite their differences, the four datasets share the following identical statistical properties:

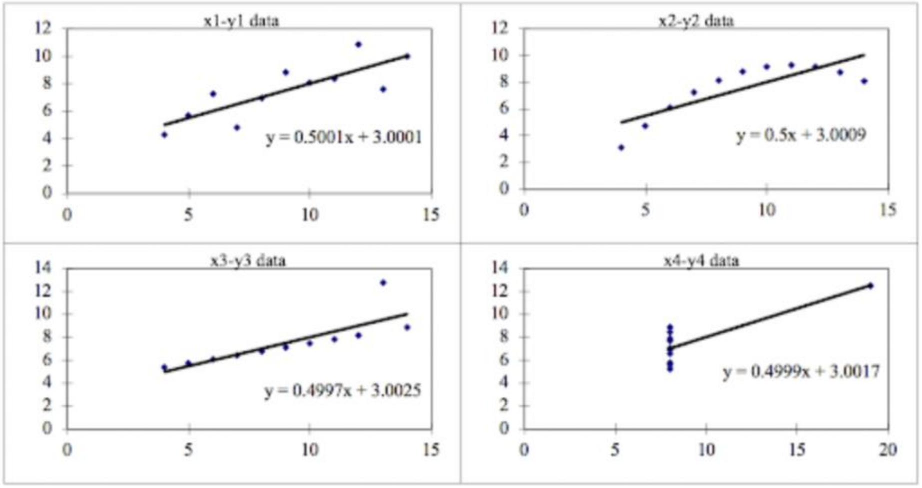
- Mean of x: 9
- Variance of x: 11
- Mean of y: 7.50
- Variance of y: 4.12
- Correlation between x and y: 0.816
- Linear regression line: y=3.00+0.500xy = 3.00 + 0.500xy=3.00+0.500x



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when plotted, the datasets reveal significant differences:



1. Dataset I: This is a typical example of a linear relationship with some noise.
2. Dataset II: Despite having a strong linear relationship, it includes an outlier that greatly influences the regression line.
3. Dataset III: The relationship appears to be curvilinear, not linear. The linear model does not fit well.
4. Dataset IV: Most of the data points are vertically aligned except for an outlier, making the linear regression line nearly horizontal.

Importance of Anscombe's Quartet

1. **Visual Inspection**: It emphasizes the importance of plotting data to understand its structure and relationships.

2. **Statistical Caution**: It warns against relying solely on summary statistics without exploring the data visually.
3. **Model Validation**: It highlights the need for model validation, ensuring that the chosen model is appropriate for the data structure.
4. **Outlier Detection**: It showcases how outliers can dramatically influence the results of statistical analyses.

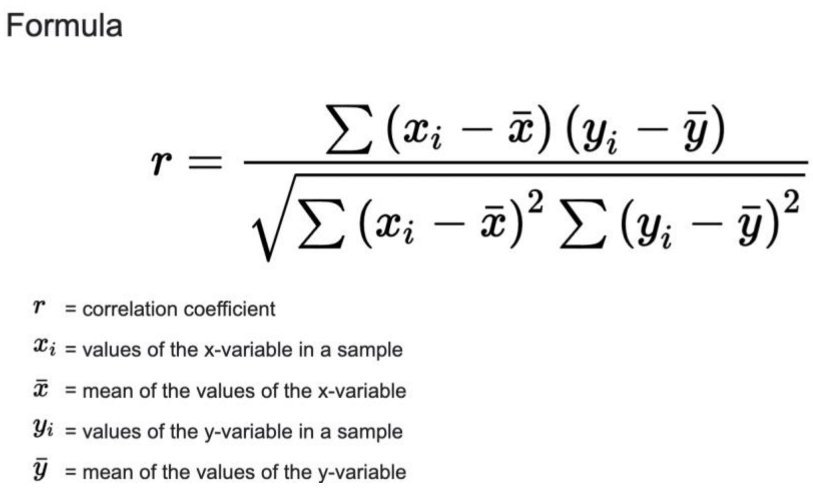
Anscombe's Quartet serves as a powerful reminder for data analysts to always visualize their data, consider the context, and validate their models. It underscores the complexity of data analysis and the potential pitfalls of over-relying on numerical summaries.

3. What is Pearson’s R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient or Pearson's correlation, is a measure of the strength and direction of the linear relationship between two variables. It is denoted by r and ranges from -1 to 1.

Formula

The formula for Pearson's R is:



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Interpretation

1. r=1: Perfect positive linear relationship. As one variable increases, the other variable increases perfectly.
2. r=−1: Perfect negative linear relationship. As one variable increases, the other variable decreases perfectly.
3. r=0: No linear relationship. The variables do not have any linear correlation.

Characteristics

- Direction: The sign of Pearson's R (positive or negative) indicates the direction of the relationship.
  - Positive R: Indicates a direct relationship. As one variable increases, the other also increases.
  - Negative R: Indicates an inverse relationship. As one variable increases, the other decreases.
- Magnitude: The absolute value of Pearson's R indicates the strength of the relationship.
  - |r| close to 1: Strong linear relationship.
  - |r| close to 0: Weak linear relationship

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

Scaling refers to the process of transforming the features of a dataset so that they are on a same scale. This process is crucial in many machine learning algorithms that use distance-based metrics (e.g., k-nearest neighbors, support vector machines) or gradient-based optimization (e.g., linear regression, neural networks).

Why Scaling is Performed

1. Improved Algorithm Performance: Many algorithms perform better or converge faster when the features are on a similar scale.
2. Reduced Bias: Ensures that all features contribute equally to the model.
3. Preventing Dominance: Prevents features with larger ranges from dominating the learning process.
4. Enhanced Model Interpretability: Makes the model easier to interpret by ensuring that all features have comparable scales.

Types of Scaling

There are two primary types of scaling: normalized scaling and standardized scaling.

Normalized Scaling

Normalization (or Min-Max Scaling) rescales the feature to a fixed range, usually 0 to 1 or -1 to 1. This is useful when you know the data distribution does not follow a Gaussian distribution and the algorithm doesn't make any assumptions about the distribution of data.

Formula

$Xnorm=X-Xmin/Xmax-XminX$

Where:

- X is the original value.
- Xmin and XmaxX are the minimum and maximum values of the feature.

Characteristics

- Range: The transformed data will be within the specified range (e.g., [0, 1]).
- Sensitive to Outliers: If there are outliers in the data, normalization can squash the majority of the data into a small range.

Use Cases

- When you need to bound values to a specific range.
- When dealing with features that have different units or scales.

Standardized Scaling

Standardization rescales the data to have a mean of 0 and a standard deviation of 1. This method is useful when the features follow a Gaussian distribution.

Formula

$Xstd=(X-\mu)/\sigma$

Where:

- X is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

Characteristics

- Mean: The transformed data will have a mean of 0.
- Standard Deviation: The transformed data will have a standard deviation of 1.
- Less Sensitive to Outliers: Compared to normalization, standardization is less sensitive to outliers.

Use Cases

- When the data follows a Gaussian distribution.
- When the algorithm assumes Gaussian distribution (e.g., Linear Regression, Logistic Regression).

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

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity among the predictor variables. In other words, it quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

Reasons for Infinite VIF Values

1. Perfect Multicollinearity**:**
   - When there is perfect multicollinearity among the predictor variables, it means that one predictor variable can be perfectly explained by a linear combination of the other predictor variables. In such cases, the denominator in the VIF formula becomes zero, causing the VIF to become infinite.

2. Linear Dependence**:**

    - Even if multicollinearity is not perfect, severe linear dependence between the predictors can cause very high VIF values, which can approach infinity.

VIF Formula

The VIF for a predictor Xi is calculated as:

VIF(Xi)=1 / (1−R ^2)

Where R^2 is the coefficient of determination of the regression of Xi on all the other predictors.

- Infinite VIF: If Ri^2=1, the denominator becomes zero, resulting in an infinite VIF. This indicates perfect multicollinearity.
- High VIF: If Ri^2 is close to 1, the VIF value will be very large, indicating severe multicollinearity.

Why It Matters

- Model Instability**:** High multicollinearity can make the coefficient estimates very sensitive to small changes in the model.
- Interpretation Difficulty: It becomes difficult to determine the individual effect of correlated predictors on the dependent variable.

Addressing Infinite or High VIF

1. Remove Variables: Identify and remove one or more of the highly collinear variables.
2. Combine Variables: Combine collinear variables into a single predictor through techniques like Principal Component Analysis (PCA).
3. Regularization**:** Use regularization techniques like Ridge Regression that can handle multicollinearity by adding a penalty to the regression.

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

A Q-Q (quantile-quantile) plot is a graphical tool to help assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution. Here's a detailed explanation of what a Q-Q plot is, how it's used, and its importance in linear regression:

6.What is a Q-Q Plot?

A Q-Q plot plots the quantiles of the sample data against the quantiles of a specified theoretical distribution. If the sample data comes from the theoretical distribution, the points should fall approximately along a straight line.

How to Create a Q-Q Plot

1. Sort the Data**:** Sort the sample data in ascending order.
2. Calculate Quantiles**:** Compute the quantiles for both the sample data and the theoretical distribution.
3. Plot Points: Plot the quantiles of the sample data on the y-axis and the quantiles of the theoretical distribution on the x-axis.

Use and Importance in Linear Regression

1. Assessing Normality of Residuals*:*

In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot can be used to check this assumption.

- Normal Distribution: If the residuals are normally distributed, the points on the Q-Q plot will lie on or near the 45-degree reference line.
- Skewness or Kurtosis: Deviations from the reference line indicate departures from normality. For example, points curving away from the line at the ends indicate heavy tails (kurtosis), and a systematic curve might indicate skewness.

2**.** Identifying Outliers**:**

Outliers can often be identified in a Q-Q plot as points that deviate significantly from the reference line.

3**.** Checking Other Distributions**:**

Although commonly used to check for normality, Q-Q plots can be used to compare data against any theoretical distribution by calculating the appropriate quantiles.

Here's how you might create and interpret a Q-Q plot in Python using statsmodels:

python
Copy code

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import statsmodels.api as sm
```

```
# Generate sample data
data = np.random.normal(0, 1, 1000)  # Sample from a normal distribution

# Q-Q plot
sm.qqplot(data, line='45')
plt.title('Q-Q Plot')
plt.show()
```

Interpreting a Q-Q Plot

- Linear Pattern: Indicates the data follows the theoretical distribution (e.g., normal distribution).
- S-Shape Curve: Suggests skewness in the data.
- Upward/Downward Curve at Ends: Suggests heavier or lighter tails than the theoretical distribution.

Importance in Linear Regression

1. Validating Model Assumptions**:** Ensuring residuals are normally distributed validates one of the key assumptions of linear regression, leading to more reliable confidence intervals and hypothesis tests.
2. Improving Model Performance: Identifying deviations from normality can prompt data transformations or model adjustments to better meet assumptions, improving model performance.
3. Detecting Issues**:** Quickly reveals issues with residual distributions that could impact the validity of the regression results.

Summary

A Q-Q plot is a crucial diagnostic tool in linear regression for checking the normality of residuals, identifying outliers, and ensuring that the assumptions of the linear regression model are met. Its ability to visualize deviations from a theoretical distribution makes it invaluable for validating the robustness of a regression model.