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:

1. Majority of bikes hired on working day / non holiday
2. less bikes hired in spring & max in summer /fall & their corresponding months
3. Light snow/rain has lower bike rentals
4. 2019 has more rentals showing growth in bike uptakes

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

Answer:

drop_first= true is redundancy reduction technique used in dummy variable creation. If there are 3 stages, the 3 dummies will be created. Now if a value is not in first 2 its obviously in the 3rd one. So the 3rd variable is reductant which is dropped.

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

Answer: temp and atemp [these 2 have very high correlation as well]

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

I checked for the following:

1. Error term normally distributed
2. Heteroscedasticity
3. Multi Collinearity / VIF
4. Linearity
5. No auto correlation among residuals

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

Year, temperature and windspeed are the top 3 contributors in my model

General Subjective Questions

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

Linear regression aims to find a linear relationship between a dependent variable YYY (also called the target or outcome variable) and one or more independent variables XXX (also called predictors or features). In simple linear regression, there is only one independent variable, while multiple linear regression involves two or more independent variables.

The equation of a simple linear regression model can be expressed as:

Y=β0+β1X+ϵY

where:

- YYY is the dependent variable we are trying to predict,
- XXX is the independent variable used for prediction,
- β0\beta_0β0 is the y-intercept of the line (the value of YYY when X=0X = 0X=0),
- β1\beta_1β1 is the slope of the line (indicating the rate of change in YYY for a one-unit change in XXX),
- ϵ\epsilonϵ represents the error term or residual, which accounts for the variability in YYY that XXX does not explain.

In multiple linear regression, with more than one independent variable, the model becomes:

Y=β0+β1X1+β2X2+⋯+βnXn+ϵY

Assumptions of Linear Regression

Linear regression relies on several key assumptions:

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: The variance of errors (residuals) is constant across all levels of the independent variables.
4. **Normality of Errors**: The residuals (differences between observed and predicted values) are normally distributed.
5. **No Multicollinearity (for Multiple Regression)**: Independent variables should not be too highly correlated with each other.

How Linear Regression Works: Least Squares Method

The **least squares method** is used to estimate the coefficients β0\beta_0β0 and β1\beta_1β1 by minimizing the sum of the squares of the residuals:

$$\text{Residual} = Y_{\text{observed}} - Y_{\text{predicted}}$$

The objective is to minimize the **sum of squared residuals (SSR)**:

$$\text{SSR} = \sum (Y_{\text{observed}} - Y_{\text{predicted}})^2$$

Minimizing SSR helps us find the best-fitting line that captures the relationship between YYY and XXX with minimal error.

The estimates for β0 and β1 in simple linear regression can be calculated using:

$$\beta_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

where $\bar{X}$ and $\bar{Y}$ are the mean values of X and Y respectively.

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

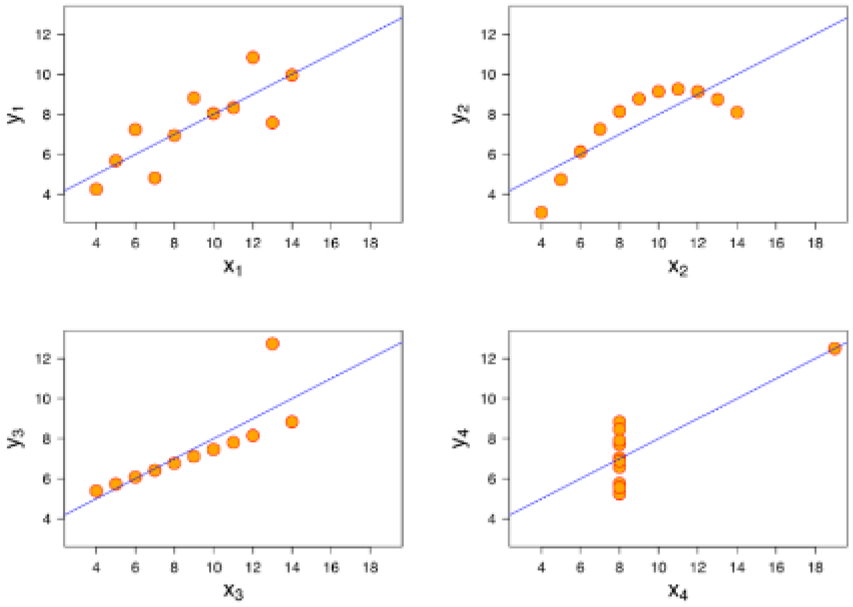
Anscombe's quartet is a set of four different datasets that have nearly identical descriptive statistics, such as mean, variance, correlation, and linear regression line. Created by statistician Francis Anscombe in 1973, the quartet demonstrates the importance of visualizing data in addition to relying on statistical measures. Each dataset in Anscombe's quartet produces the same statistical results, yet their scatter plots reveal dramatically different distributions and relationships between the data points. This example highlights how relying solely on summary statistics can be misleading and stresses the value of data visualization in uncovering underlying patterns, trends, and outliers.

Structure of Anscombe's Quartet

Each of the four datasets in Anscombe's quartet consists of 11 pairs of (x,y)(x, y)(x,y) values. The datasets share the following approximate statistical properties:

- **Mean of x**: Approximately 9.0
- **Mean of y**: Approximately 7.5
- **Variance of x**: Approximately 10
- **Variance of y**: Approximately 3.75
- **Correlation between x and y**: Approximately 0.82
- **Linear regression line**: y=3+0.5xy = 3 + 0.5xy=3+0.5x

Despite these identical statistics, each dataset represents a different relationship between x and y.



Dataset I

This dataset resembles a simple linear relationship between xxx and yyy. It has a roughly linear pattern, with data points closely following the linear regression line y=3+0.5xy = 3 + 0.5xy=3+0.5x. This dataset aligns with what we typically expect when we see these summary statistics.

Dataset II

Dataset II is different from Dataset I because it has a non-linear pattern. The values in Dataset II follow a parabolic, or U-shaped, relationship rather than a linear one. Despite having the same mean, variance, and regression line as Dataset I, a plot reveals that this data set curves upwards in a way that the linear regression line doesn't capture well. This shows that summary statistics can fail to capture non-linear relationships.

Dataset III

In Dataset III, most of the data points line up perfectly along a linear trend except for one outlier that skews the data. This outlier is far from the others and significantly affects the statistical results, particularly the correlation and regression line. When plotted, the outlier is clearly visible and demonstrates how a single extreme value can distort summary statistics and create a misleading representation of the overall data trend.

Dataset IV

Dataset IV is unique because it contains a vertical line of data points with one outlier at the end. While all the data points share the same xxx value except for one, the summary statistics are identical to the other datasets. This dataset serves as a strong example of how the same regression and correlation results can be meaningless if the data distribution is atypical or clustered around specific values.

Anscombe’s quartet provides critical insights into the limitations of relying on summary statistics without visualization:

1. **Visualization Complements Statistics**: The quartet demonstrates that two-dimensional scatter plots can reveal patterns, trends, or outliers that summary statistics miss.
2. **Outliers Can Skew Results**: Dataset III shows that even one outlier can significantly distort the regression line and correlation coefficient, leading to a misinterpretation of the data.
3. **Non-Linear Relationships**: Dataset II reveals that linear regression is not suitable for all datasets. While summary statistics may suggest a linear relationship, a plot can reveal non-linear patterns.
4. **Importance of Context**: Data should be analyzed within the context of the problem it addresses. Context can inform the appropriate methods for exploring relationships, detecting patterns, and making predictions.

The principles highlighted by Anscombe’s quartet are especially relevant today, as data analysis has expanded with more sophisticated algorithms and larger datasets. In fields like machine learning and artificial intelligence, where models are often complex, visualizing data remains essential to ensure the results are meaningful and interpretable. Understanding the underlying data relationships can improve the accuracy of models, prevent the influence of outliers, and ensure appropriate model selection.

3. What is Pearson’s R? (3 marks)

Pearson's r, also known as the **Pearson correlation coefficient**, is a statistical measure that quantifies the linear relationship between two variables. It indicates both the **strength** and **direction** of a linear association, making it widely used in fields like psychology, social sciences, economics, and any domain where relationships between two variables are analyzed. Pearson's r can range from -1 to 1, where:

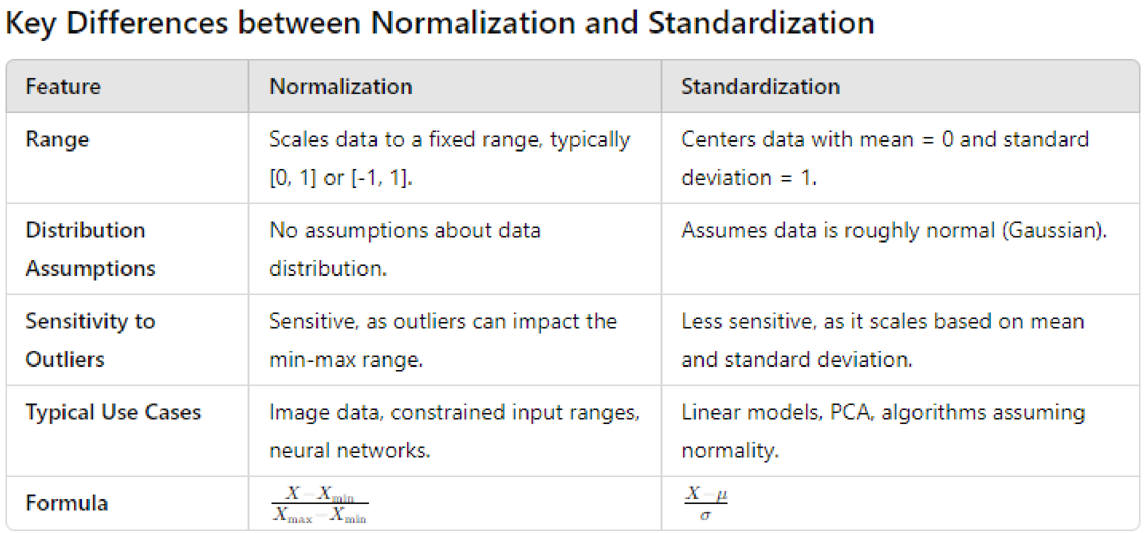
- **+1** indicates a perfect positive linear relationship.
- **-1** indicates a perfect negative linear relationship.
- **0** indicates no linear relationship.

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 adjust the range and distribution of values in a dataset. In data analysis and machine learning, scaling is particularly important because many algorithms perform better or converge faster when features are on a similar scale.

Scaling is necessary for several reasons:

1. **Improving Model Performance**: Algorithms that calculate distances (like k-nearest neighbors and clustering algorithms) and gradient-based models (like linear regression and neural networks) work better when features are scaled.
2. **Ensuring Faster Convergence**: For optimization algorithms (e.g., gradient descent), scaling helps achieve faster convergence, as large differences in feature scales can lead to oscillations or slow down convergence.
3. **Preventing Bias from Larger Scales**: Features with larger ranges may dominate features with smaller ranges, which can bias the model toward certain attributes.
4. **Interpretable Coefficients**: In linear models, scaling ensures that the coefficients are comparable across features, making interpretation more meaningful.



Key Differences between Normalization and Standardization

Feature	Normalization	Standardization
Range	Scales data to a fixed range, typically [0, 1] or [-1, 1].	Centers data with mean = 0 and standard deviation = 1.
Distribution Assumptions	No assumptions about data distribution.	Assumes data is roughly normal (Gaussian).
Sensitivity to Outliers	Sensitive, as outliers can impact the min-max range.	Less sensitive, as it scales based on mean and standard deviation.
Typical Use Cases	Image data, constrained input ranges, neural networks.	Linear models, PCA, algorithms assuming normality.
Formula	$\frac{X - X_{\min}}{X_{\max} - X_{\min}}$	$\frac{X - \mu}{\sigma}$

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

VIF equals infinite is a case of perfect correlation where R-square = 1 and consequently (1/(1-R-square) is infinite. We have to drop a variable in that case.

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

Quantile Quantile Plot a graphical technique used to determine of 2 datasets come from population with the same distribution.