Bike Sharing assignment by Abhishek Gupta

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?

Answer -

1. **Season** - Fall season has the highest count of people using boom bikes. Spring season has the least
2. **Mnth** - Most of the bookings have been made during the months of May, June, July, Aug, Sep, and Oct. The trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
3. **Weekday** - Sunday has the least users and the users increase from Monday to Friday.
4. **weathersit** - more people tend to use bikes when the weather is

   'clear'.
5. **Workingday** - Booking seemed to be almost equal either on working day or non-working day.
6. **Holiday** - Bikes are more used when there is no holiday.
   Yr - 2019 attracted more bookings than the previous year.

Question 2

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

Answer -

- drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

- We use drop_first=True to avoid dummy variable trap. Dummy variable trap is when we have perfect multicollinearity between the dummy variables.

Question 3

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

Answer -

- The `'temp'` variable exhibited the strongest correlation with the target variable.

Question 4

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

Answer -

Assumptions :

- **Multicollinearity:** To ensure model stability and interpretability, it is assumed that there is minimal multicollinearity among the independent variables. This means that the variables are not highly correlated with each other.
- **Normality of Error Terms:** The residuals (the differences between the actual and predicted values) are assumed to be normally distributed. This assumption is crucial for statistical inference and hypothesis testing.
- **Linear Relationship:** A linear relationship is assumed between the independent and dependent variables. This means that the relationship can be adequately represented by a straight line.
- **Homoscedasticity:** The variance of the residuals should be constant across different values of the independent variables. This assumption ensures that the model's predictions are equally reliable for all data points.
- **Independence of Residuals:** The residuals should be independent of each other, meaning that the error in predicting one observation should not be correlated with the errors in predicting other observations. This assumption is essential for valid statistical inference.

Question 5

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

Answer -

Here are the top 3 features contributing significantly towards explaining the demand of the shared bikes –
1. temp
2. yr
3. Light_snow_rain

General Subjective Questions

Question 1

Explain the linear regression algorithm in detail.

Answer -

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting line that minimizes the distance between the predicted values and the actual values.

Key Components

- **Model Equation:**
  - The linear regression model is represented by the equation:

    where:

    `y = β0 + β1x1 + β2x2 + ... + βpXp + ε`

    - `y` is the dependent variable
    - `β0` is the intercept (constant term)
    - `β1`, `β2`, ..., `βp` are the coefficients for the independent variables (`x1`, `x2`, ..., `Xp`)

        - ε is the error term, representing the random variation not explained by the model

- **Ordinary Least Squares (OLS):**
    - The most common method for estimating the coefficients in linear regression is OLS.
    - OLS minimizes the sum of squared residuals (the differences between the predicted and actual values).
- **Assumptions:**
    - **Linearity:** A linear relationship exists between the dependent and independent variables.
    - **Independence:** The observations are independent of each other.
    - **Normality:** The error terms are normally distributed.
    - **No multicollinearity:** The independent variables are not perfectly correlated with each other.

Evaluation Metrics

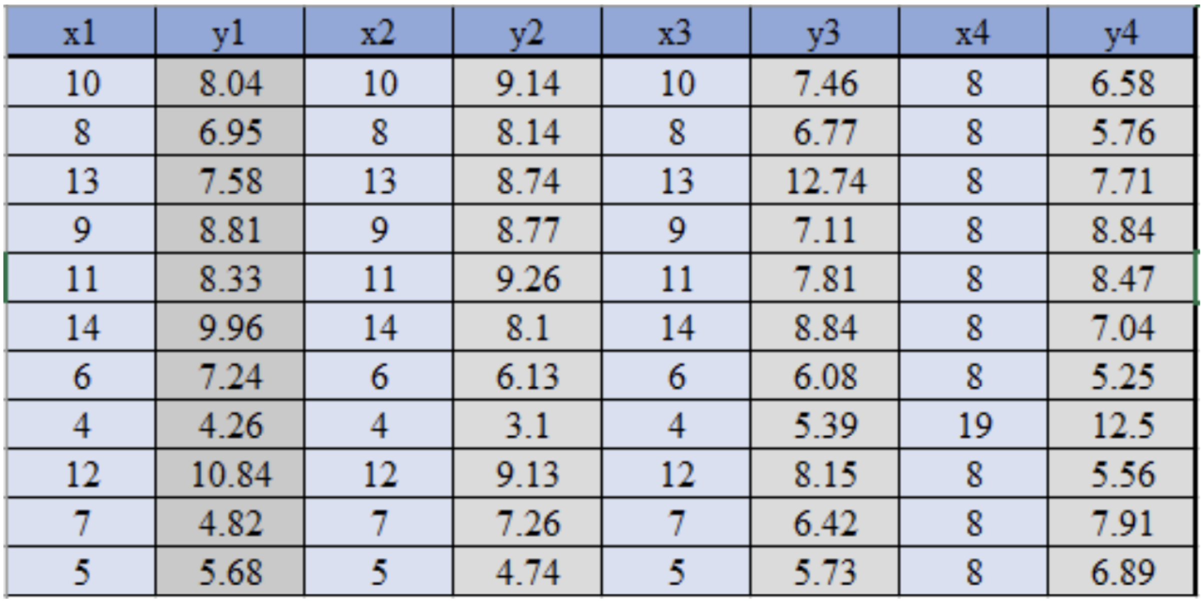
- **R-squared (R^2):** Measures the proportion of variance in the dependent variable explained by the independent variables.
- **Adjusted R-squared:** Penalizes models with excessive variables.
- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values.
- **Root Mean Squared Error (RMSE):** The square root of MSE, providing a measure in the same units as the dependent variable.

Question 2

Explain the Anscombe’s quartet in detail.

Answer -

Anscombe's quartet is a famous set of four datasets, each with identical statistical properties (mean, median, variance, correlation). Despite these identical summary statistics, the datasets exhibit dramatically different visual patterns when plotted, highlighting the importance of visualizing data beyond relying solely on numerical summaries.

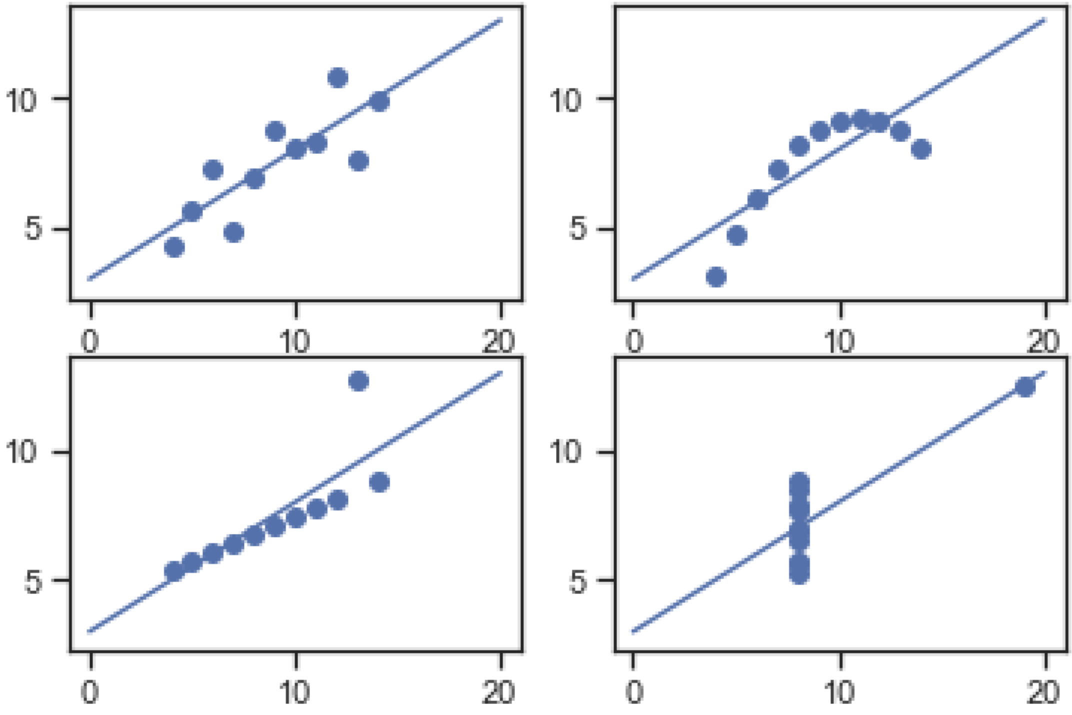


x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

The four datasets share the following statistical properties:

- **Mean of x:** 9
- **Mean of y:** 7.5
- **Variance of x:** 11
- **Variance of y:** 4.12
- **Correlation between x and y:** 0.816
- **Linear Regression Equation** : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



1. Data-set 1 — Consists of a set of (x,y) points that represent a linear relationship with some variance.
2. Data-set 2 — Shows a curve shape but doesn't show a linear relationship (might be quadratic).
3. Data-set 3 — Looks like a tight linear relationship between x and y, except for one large outlier.
4. Data-set 4 — Looks like the value of x remains constant, except for one outlier as well.

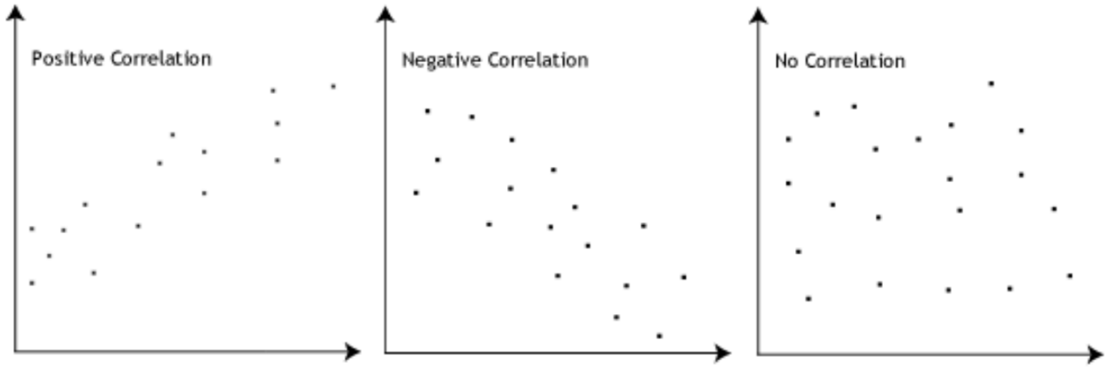
Datasets that exhibit identical statistical properties but yield distinct visual representations are often used to emphasize the significance of graphical exploration in data analysis.

Question 3

What is Pearson's R?

Answer -

- Pearson's r, also known as the Pearson correlation coefficient, is a statistical measure that describes the linear relationship between two continuous variables. It is a value between -1 and 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.
- Pearson's r measures the degree to which the variables are related by calculating the ratio of the covariance between the variables to the product of their standard deviations. In other words, it measures how much the variables vary together relative to how much they vary independently.



Question 4

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer -

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
- Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that’s actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Question 5

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

Answer -

- **VIF (Variance Inflation Factor):** A measure of how much the variance of a regression coefficient is inflated due to multicollinearity.
- **High VIF:** A large VIF value indicates a strong correlation between the variable and other independent variables, potentially leading to unstable coefficients and difficulty in interpreting results.
- **Perfect Multicollinearity:** A VIF of infinity indicates perfect multicollinearity, meaning one variable can be perfectly predicted from others.
- **Addressing Multicollinearity:** Techniques like feature engineering, feature selection, PCA, or regularization can help mitigate multicollinearity.
- **Importance:** Addressing multicollinearity is crucial for building reliable and interpretable models.

Question 6

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

Answer -

- **Purpose:** The Q-Q plot is a graphical technique used to compare the distributions of two datasets.
- **Quantiles:** A Q-Q plot plots the quantiles of one dataset against the quantiles of another.
- **Reference Line:** A 45-degree reference line is used as a benchmark. If the points fall closely along this line, it suggests similar distributions.
- **Deviations:** Deviations from the reference line indicate differences in the distributions.
- **Insights:** Q-Q plots provide valuable insights into the nature of differences between distributions, going beyond simple statistical tests like chi-square and Kolmogorov-Smirnov.