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?**
    a. Year
    b. Working day
    c. Winter (season : 4)
    d. Weathersit 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
    e. Weathersit 2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist)
    f. Month (January and September)
    g. Weekday(Saturday)

2. **Why is it important to use drop_first=True during dummy variable creation?**
    a. To encode categorical data, one hot encoding is done, where a dummy variable is to be created for each discrete categorical variable for a feature. This can be done by using pandas.get_dummies() which will return dummy-coded data. Here we use parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.
    b. If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    a. Linear Relationship between the dependent variable and independent variable using pair plot
    b. Plotted the error terms using Histogram which shows mean towards zero and normally distributed
    c. Using VIF check, confirmed that there is no multicollinearity
    d. Error terms are not following any pattern. We can confirm by plotting the scatterplot on error terms

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    a. Temp
    b. Weathersit3 which is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (negative correlation)
    c. year

General Subjective Questions

1. Linear Regression:

Regression is a supervised learning technique. In Linear Regression, the output of the variable is continuous in nature.

Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). ***If there is a single input variable (x), such linear regression is called simple linear regression. And if there are more than one input variable, such linear regression is called multiple linear regression***. The linear regression model gives a sloped straight line describing the relationship within the variables.

Relationship between independent and dependent variable:

1. Positive correlation
2. Negative correlation

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \implies y = a_0+a_1x$$

Assumptions of Linear Regression Model:

- Linear relationship between X and y.
- Normal distribution of error terms.
- Independence of error terms.
- Constant variance of error terms.

Hypothesis testing in linear regression

- To determine the significance of beta coefficients.
- H0:β1=0;HA:β1≠0.
- T-test on the beta coefficient.
- t score=^βiSE(^βi).

2. Anscombe's quartet:

Anscombe's quartet has a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines. When we plot those dataset using scatter plot it has different representations. The datasets are created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data. Statistics summary alone is not sufficient and we have to visualize the data before model building.

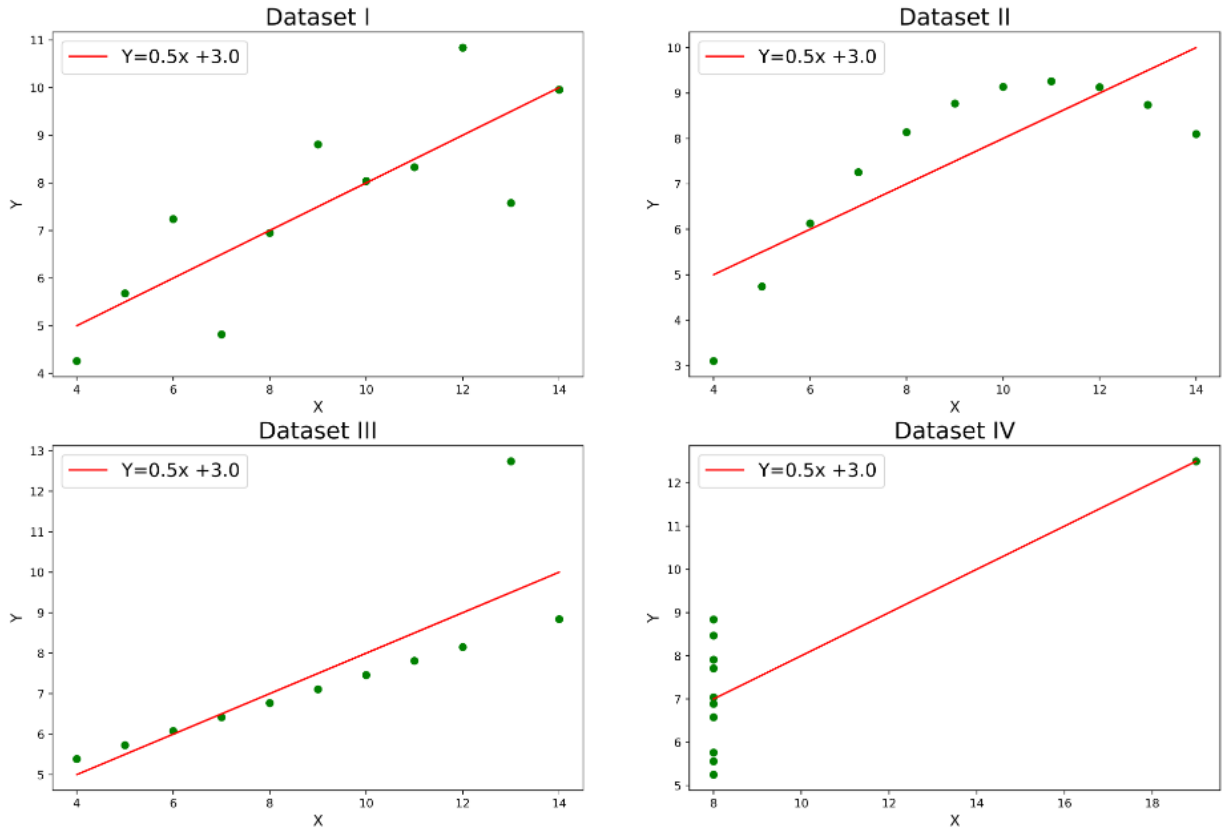
Anscombe's quartet is used to illustrate the importance of EDA and the drawbacks of summary statistics. It also emphasizes the importance of using data visualization to spot

trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Summary Statistics of four dataset:

Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

Graph representation of four dataset:



Explanation of this output:

- In the first one(top left) if you look at the scatter plot there seems to be a linear relationship between x and y.
- In the second one(top right) there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. Pearson's R

Pearson correlation coefficient, also known as Pearson R, is a statistical test that estimates the strength between the different variables and their relationships. Hence, whenever any statistical test is performed between the two variables, it is always a good idea for the person to estimate the correlation coefficient value to know the strong relationship between them.

The correlation coefficient of -1 means negative relationship. Therefore, it shows a perfect negative relationship between the variables. If the correlation coefficient is 0, it shows no relationship. If the correlation coefficient is 1, it means a strong positive relationship. Therefore, it shows a perfect positive relationship between the variables.

The Pearson correlation coefficient shows the relationship between the two variables calculated on the same interval or ratio scale. In addition, It estimates the relationship strength between the two continuous variables.

Formula:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

4. Scaling:

Scaling is the process of normalizing the range of features in a dataset. It is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Why scaling is performed:

Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling

Difference between normalized scaling and standardized scaling

Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

5. VIF Infinity:

VIF:

In regression analysis, the variance inflation factor (VIF) is a measure of the degree of multicollinearity of one regressor with the other regressors

Reason for Infinity:

VIF = 1 / 1- R2. If R2 is 0 then the VIF will be infinity for variable.

If there is perfect correlation between two independent variables, we get R2 =1, which lead to 1/(1-R2) infinity. A large value of VIF indicates that there is a correlation between the variables. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity

6. Q-Q Plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

The datasets that we are comparing are the same type of distribution type, we will get a roughly straight line

Use and Importance of Q-Q plot in Linear Regression:

This helps in linear regression when we have training and test data set comes separately and then we can confirm using Q-Q plot that both data sets are from populations with same distributions.

Residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.

Skewness of distribution

Python:

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively