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: I have plotted the categorical variables with the target variables on boxplot and has inferred following effect on target:

Season: 3: fall has highest demand for rental bikes

I see that demand for next year has grown

Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing

When there is a holiday, demand has decreased.

Weekday is not giving clear picture about demand.

The clear weathers hit has highest demand

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.

If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable?

Answer: There is a Linear relation between 'temp','atemp' and 'cnt'. As temp,atemp value increases cnt value also increase.

4. How did you validate the assumptions of Linear Regression after building the model on the

training set?

Answer: Error terms are normally distributed with mean 0.

Error Terms do not follow any pattern.

Multicollinearity check using VIF(s).

Linearity Check.

Ensured the overfitting by looking the R2 value and Adjusted R2.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Answer:

Features “holiday”, “temp” and season “hum” are highly related with target column, so these are top contributing features in model building.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: The TA linear regression algorithm uses a straight line to attempt to explain the relationship between the independent and dependent variables. Only numerical variables are applicable.

When performing linear regression, the following actions are taken:

- Test and training data are separated from the dataset.
- Target (dependent) and features (independent) datasets are separated from the train data.
- The training dataset is used to fit a linear model. The gradient descent algorithm is used internally by the Python APIs to determine the coefficients of the best fit line. The cost function is minimised by the gradient descent process. Residual sum of squares is a common instance of a cost function.
- When there are numerous characteristics, a hyperplane is projected as the variable rather than a line. The anticipated variable looks like this:

$Y = \beta 0+\beta 1x1+\beta 2x2+\beta 3x3+\cdots+ \beta nx$

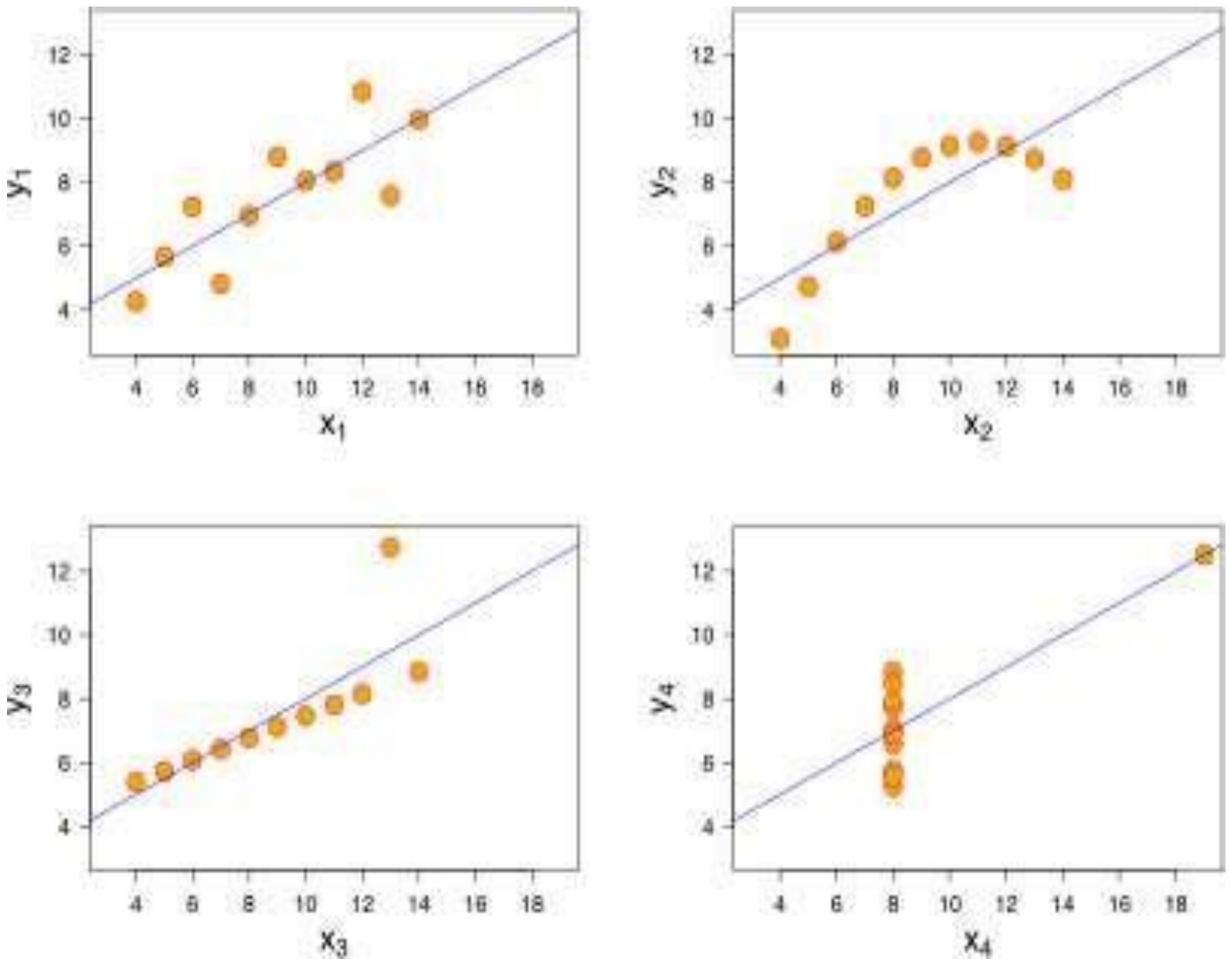
- The anticipated variableThe predicted variable is than compared with test data and assumptions arechecked.

2. Explain the Anscombe’s quartet in detail.

Answer: Anscombe's quartet consists of four data sets with virtually similar simple descriptive statistics, but when represented graphically, the distributions are very different.

The mean, sample variation of x and y, correlation coefficient, linear regression line, and R-Square value make up the simple statistics.

Anscombe's Quartet demonstrates how graphing can nevertheless reveal significant differences between numerous data sets with many comparable statistical features. The charts are displayed below:



1. The first plot (top left) seems to represent a straightforward linear relationship.

2. The correlation coefficient is useless because the second figure (top right) depicts a nonlinear relationship and is not normally distributed.

3. The third plot is linear but uses a different regression line (bottom left). This is taking place as a result of the outliers in the data.

4. The fourth plot (bottom right) does not demonstrate a linear relationship, but the data were changed because of outliers.

To put it simply, it is better to visualise data and eliminate outliers before examining it.

3. What is Pearson's R?

Answer: The strength of a relationship between two variables is measured by Pearson's R.

It is calculated by dividing the covariance of two variables by the sum of their standard deviations. Its range of values is +1 to -1.

•A value of 1 denotes a complete linear positive correlation. It implies that if one variable rises, the others will follow suit.

•Zero indicates there is no association.

•A score of -1 indicates a completely negative association. It implies that if one variable rises, another will fall.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling?

Answer:

Scaling is a technique used in data preprocessing to standardize the range of features in a dataset. It involves transforming the values of each feature so that they fall within a specific range or distribution. Scaling is performed to make sure that no feature dominates the model's prediction, which could lead to a biased or inaccurate model. 1. Normalization scaling (also called min-max scaling) transforms each feature so that its values are between 0 and 1. It is done by subtracting the minimum value of the feature and dividing by the range of the feature. The formula for normalization is: x_normalized = (x - min(x)) / (max(x) - min(x))

where x is the original feature, x_normalized is the normalized feature, min(x) is the minimum value of the feature, and max(x) is the maximum value of the feature. Normalization scaling is useful when the distribution of the data is not known, and it does not change the shape of the distribution.

2.Standardization scaling (also called z-score scaling) transforms each feature so that its values have zero mean and unit variance. It is done by subtracting the mean of the feature and dividing by the standard deviation of the feature. The formula for standardization is: x_standardized = (x - mean(x)) / std(x) where x is the original feature, x_standardized is the standardized feature, mean(x) is the mean value of the feature, and std(x) is the standard deviation of the feature. Standardization scaling is useful when the distribution of the data is known or approximately known, and it transforms the data into a standard normal distribution with a mean of zero and standard deviation of one.

Normalised Scaling	**Standardized scaling**
Called min max scaling, scales the variable such that the range is 0-1	Values are centred around mean with aunit standard deviation
Good for non- gaussian distribution	Good for gaussian distribution
Value id bounded between 0 and 1	Value is not bounded
Outliers are also scaled	Does not affect outliers

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

Answer: The formula for VIF is

VIFi = 1

1− $Ri2$

Basically, VIF becomes limitless if R square is 1.

It indicates that the characteristics perfectly align with one another.

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

Answer: A Q-Q plot is a scatter plot that compares two sets of quantiles.

To determine whether the two sets of data came from the same distribution is its goal.

Data is being visually checked. If all of the data are from the same source, the plot will seem like a line.