Assignment-based Subjective Questions

Que 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Effect of categorical variables was as follows:

Target variable: Count ('cnt' column in dataset)

1. Weekday: People prefer renting bikes on weekdays and very few rent on weekend
2. Weathersit: Count of renting is more when the weather is clear and lowest when it rains and when there is snowfall
3. Year: Demand for bikes is more in 2019 as compared to 2018
4. Holiday: Very few people seem to rent bikes on holidays
5. Month: Demand for bikes is high in June - September and low in January - February

Que 2: Why is it important to use drop_first=True during dummy variable creation?

Answer: We use 'n-1' levels to describe n variables so that we can indicate the variable with minimum number of dummy variables and convey same information as conveyed by the main variable (for which dummy variables are created)

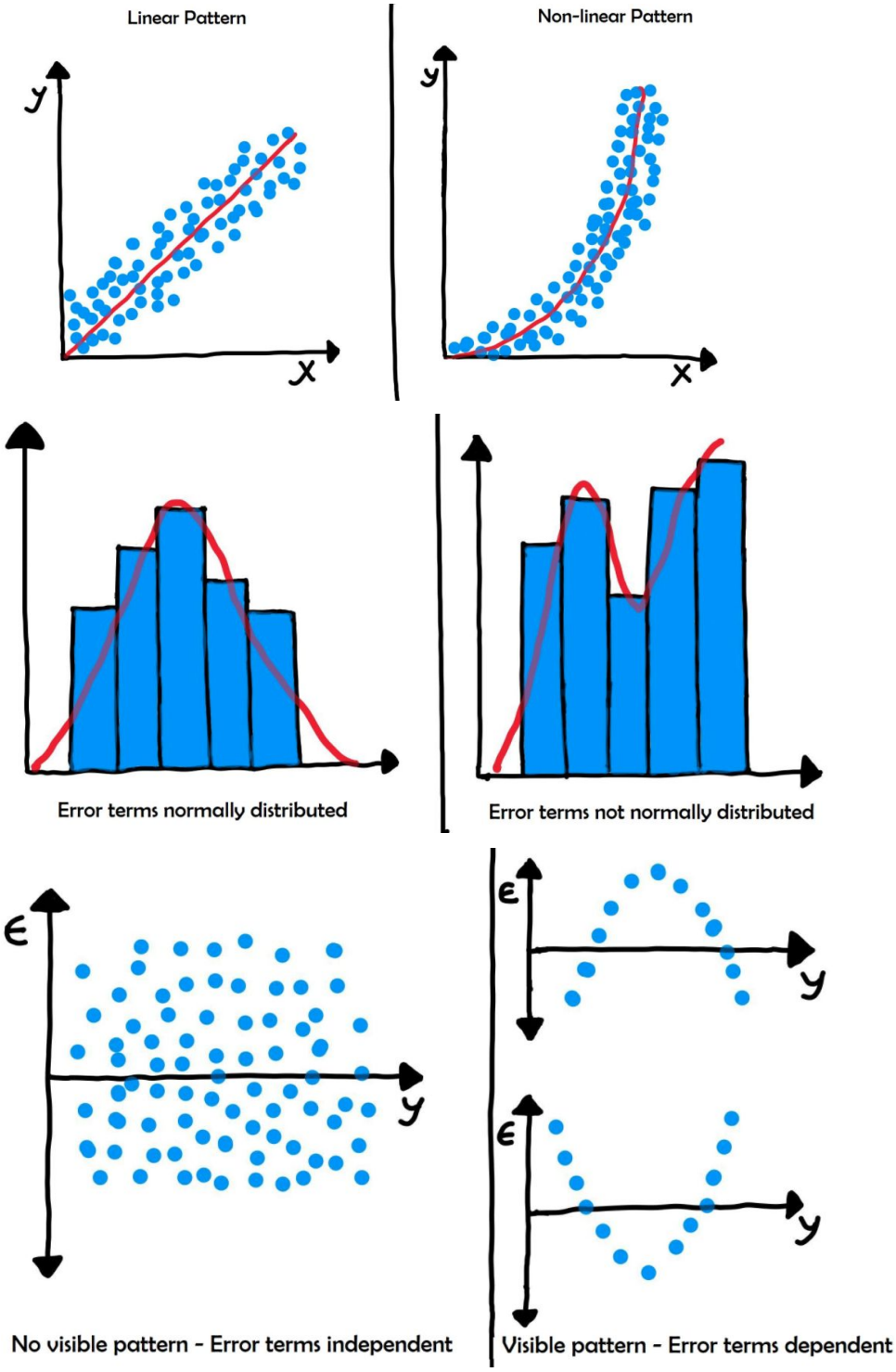
Que 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

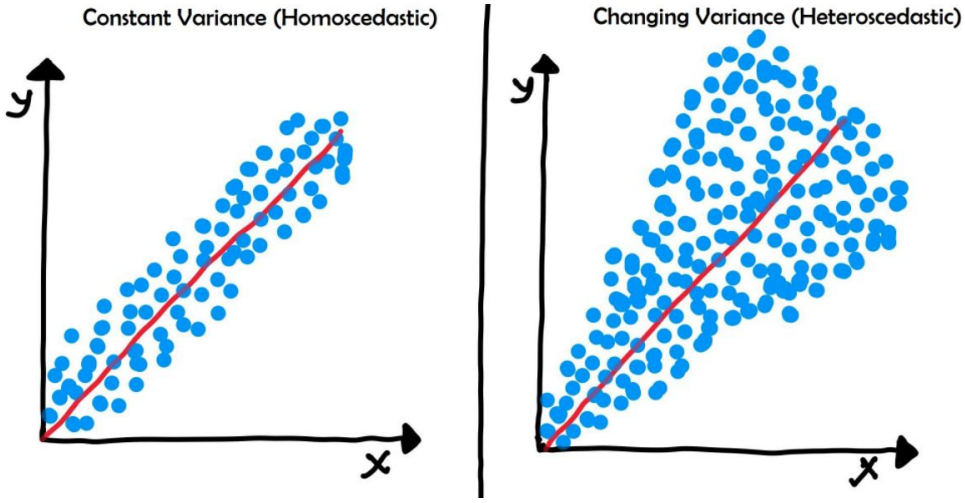
Answer: From the pair plot we can infer that the variables namely 'temp', 'atemp' and 'yr' have the highest correlation with the target variable

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

Answer: Assumptions of Linear Regression can be validated on the basis of:

1. By plotting residuals and checking if error terms are normally distributed and mean is centered at zero
2. Plotting spread of error terms and checking if error terms are independent (no pattern should be observed in the graph)
3. Error terms should have constant variance
4. There should be linear relation between dependent and independent variables


Linear Pattern
y
x
Non-linear Pattern
y
x
Error terms normally distributed
Error terms not normally distributed
ϵ
y
No visible pattern - Error terms independent
ϵ
y
ϵ
y
Visible pattern - Error terms dependent



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

Answer: Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Year 2019: yr_2019 column
2. Temperature: temp_diff column
3. Weather: weathersit_2 and weathersit_3 columns

General Subjective Questions

Que 1: Explain the linear regression algorithm in detail.

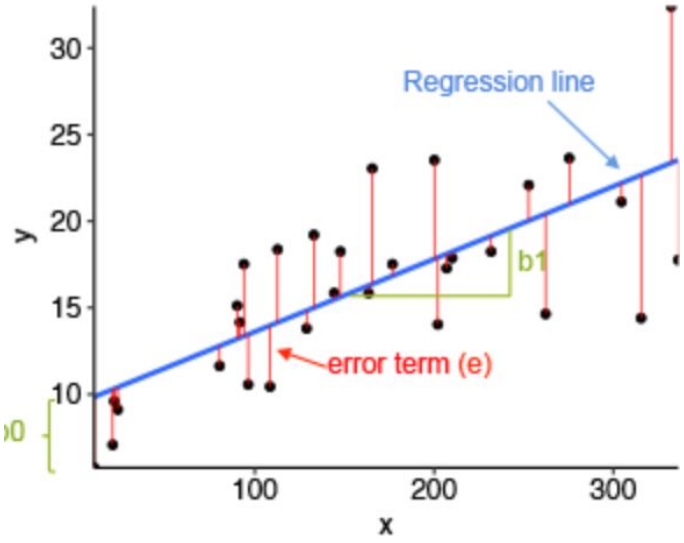
Answer: In Regression analysis, we find the relationship between one dependent and one or more independent variables.
Linear Regression is a type of modelling where we explore the dataset and find the 'Linear' relationship between the dependent (Target) variable and independent (Predictor) variables.

Linear Regression can be used for:

1. To estimate upcoming trends
2. Predict the change in output/Target variable upon change in predictors/Input variables
3. Determining the strength of Predictors

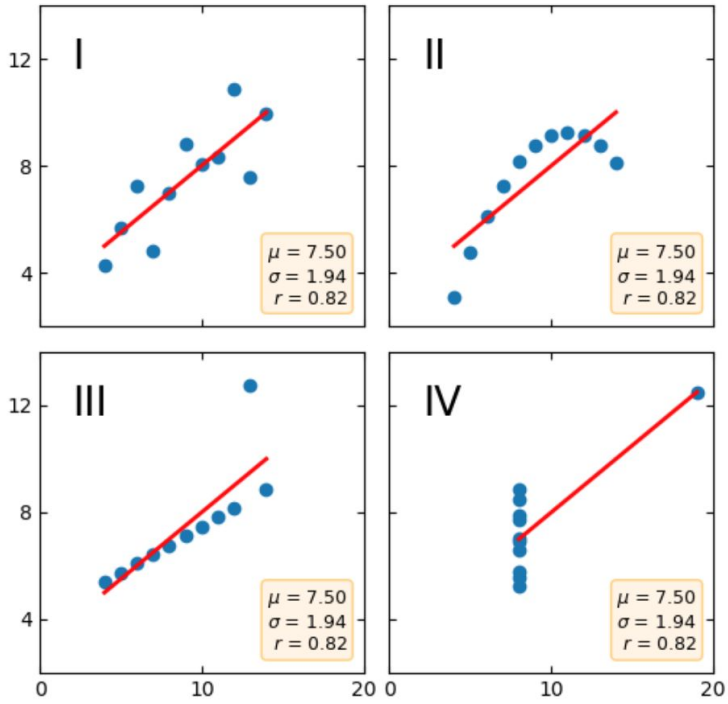
Linear Regression Algorithm:

1. Importing required Libraries
2. Reading, understanding and cleaning the dataset
3. Choosing the Target variable based on problem definition
4. Visualising the Dataset to find the relations between dependent /target variable and independent variables/predictor variables
5. Data preparation: converting categorical columns to dummy variables
6. Splitting the data into Train set and Test set
7. Scaling the columns in Train dataset so that all columns comply to same scale and provide better interpretation making backend processing much faster
8. Training the model using statsmodel or scikit learn library
9. Analysing the model parameter such as p value, F stat, prob (F stat), R square, adjusted R square and coefficients
10. If the model parameters as satisfactory then we can move to Residual analysis otherwise repeat the above steps with different features in dataset till you get the acceptable model parameters
11. Residual Analysis to check if error terms follow normal distribution or not
12. Check for plot of error terms: error terms should be independent and there should be constant variance in error terms
13. Make prediction on Test dataset
14. Evaluate model: Plot the graph between predicted values of target variable and actual values of target variable
15. Difference in R square of Train set and Test set should be less than 5% for stable model



Que 2: Explain the Anscombe's quartet in detail.

Answer:
Anscombe's quartet has four data sets.
These datasets have nearly identical simple descriptive statistics but have very different distributions.
They appear very different when graphed and each dataset comprises eleven (x,y) points.



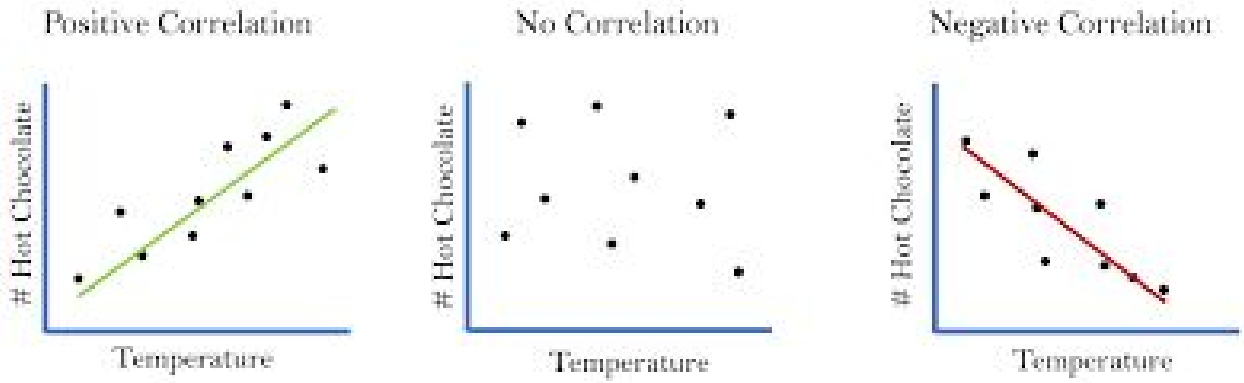
Plot 1 (top left): Date set seems to have clean and very well fitted linear model
Plot 2 (top right): Dataset is not distributed normally
Plot 3 (bottom left): Dataset shows linear distribution but calculated regression goes wrong due to outlier
Plot 4 (bottom right): This graph shows that one outlier is enough to create high correlation coefficient

This proves that even if stats of datasets are similar, the visualisation can completely reveal different hidden patterns and thus it is very important to go through Exploratory Data Analysis before concluding the problem definitions.

Que 3: What is Pearson's R?

Answer: Pearson's R measures the linear relationship between two variables X and Y.

Value of Pearson's R lies between -1 to 1:
-1: implies strong negative linear relationship between X and Y
0: implies no linear relationship between X and Y
1: implies strong positive linear relationship between X and Y



Pearson's R is given by:

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

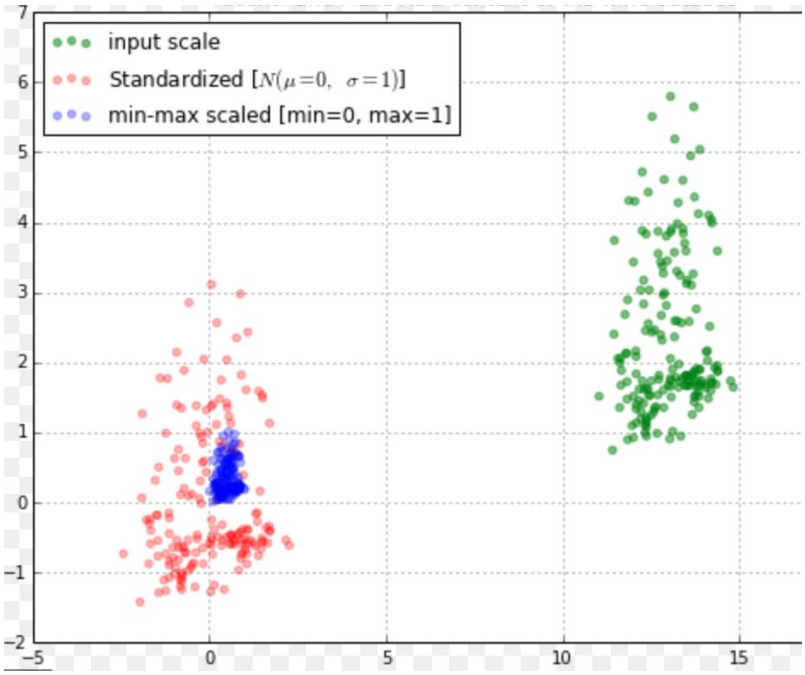
Que 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:
Scaling:
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/ values/ units.

Need for Scaling:
If we do not perform feature scaling, then our chosen machine learning algorithm will weigh greater values higher and weigh smaller values as lower values, regardless of the unit of the values.

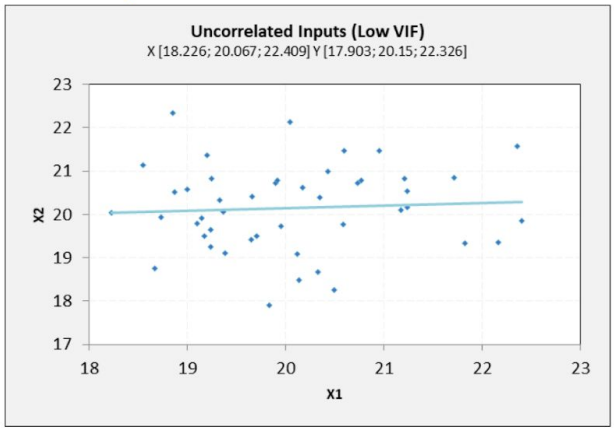
Types of scaling:
1. Normalization or MinMaxScaling**:** This technique re-scales a feature/ observation value with distribution value between 0 and 1 strictly.

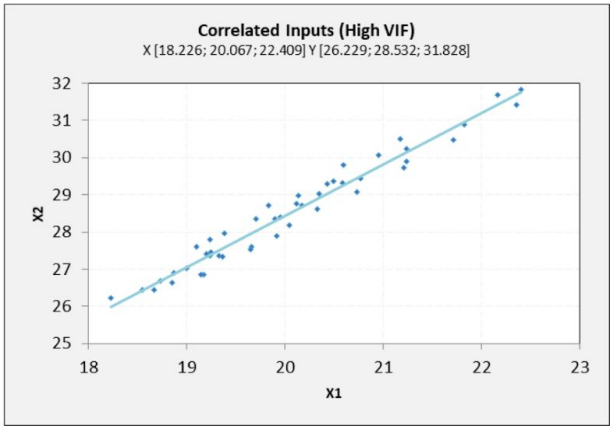
2. Standardization: It re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.



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

Answer: We start by observing VIF values and drop the features with VIF values more than 10. We might sometimes observe that VIF value for some feature is infinite.
This infinite value of VIF indicates that corresponding variable can be exactly expressed by the combinations of other present variables.





VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

We should definitely drop the features having VIF more than 10 as they can be easily explained by other existing features in Dataset. Keep the features with VIF less than 5 and do check the features once with VIF value in between 5 - 10.

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

Answer:

Quantile-Quantile (Q-Q) plot:
It is a graphical tool to assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

Use of Q-Q plot:
It helps to determine if two data sets came from populations with a common distribution or not. This basically helps in linear regression where we have a training set and test data set received separately and then we can confirm using Q-Q plot that both the data sets (train and test) are from populations with same distributions.

Impotence of Q-Q plot:

1. It can be used with sample sizes
2. We can detect the outlier present in dataset, change in symmetry, location or scale with Q-Q plot

Sample: