Assignment-based Subjective Questions

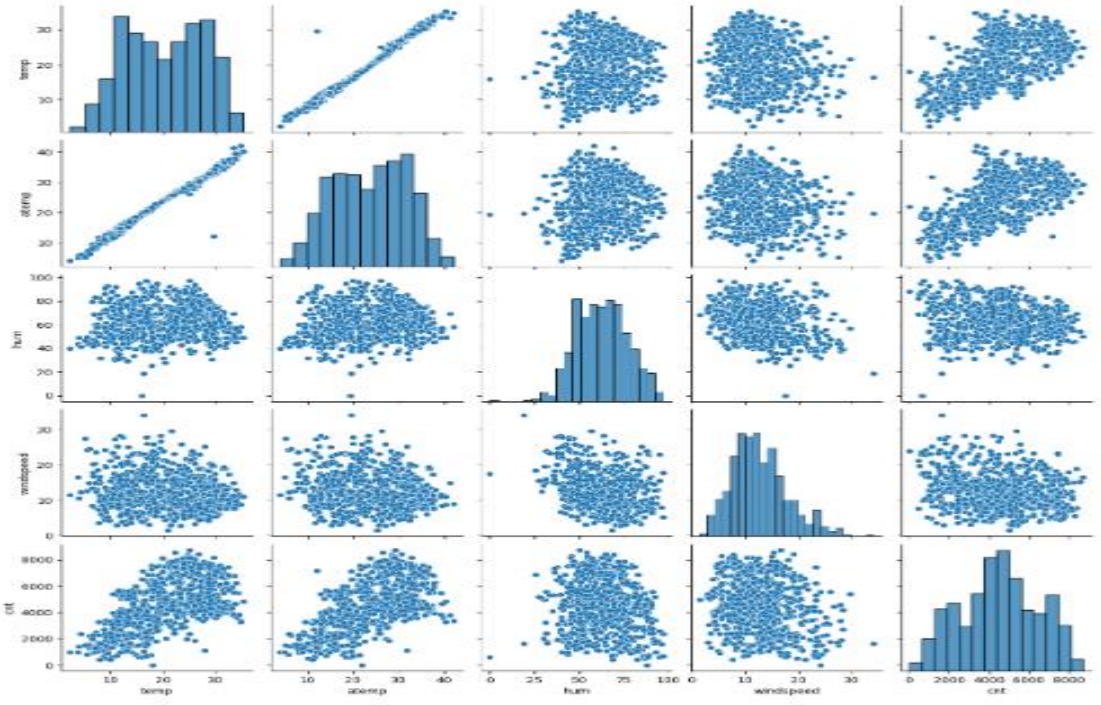
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

From our dataset, we can conclude that demand for bike rental in 2019 is high compared to 2018, Somehow people like to travel more on weekends compared to weekdays and on holiday demand is pretty low. We have observed in September and October month people like to take more bike on rents which come under summer and fall season. Clear to partly cloudy weather is more suitable for bike riding, so people are concentrating more on weather as well.

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

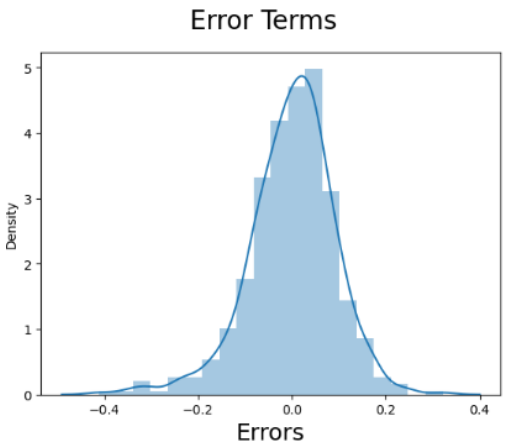
We need to use drop_first = True for reducing extra column which created during dummy variable creation. It also reduces multicollinearity which causes issue in estimation of regression coefficient and reduces model interpretability.

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



Temp and atemp are highly correlated with each other. On other way, temp and atemp are highly correlated with the target variable cnt.

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



We have validated the assumption of linear regression by checking VIF, p value, residual analysis and by checking the relationship between dependent and independent variables. Hence, model is linearly valid and normally distributed.

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

There are three features contributing towards demand for shared bikes are temperature, year and season.

Temperature: A coefficient value of 0.518784 means one unit increase in temperature variable increases bike rentals by 0.518784 unit.

Season: A coefficient value of 0.079796 means one unit increase in summer variable increases bike rentals by 0.079796 unit and A coefficient value of 0.119175 means one unit increase in winter variable increases bike rentals by 0.119175 unit.

Year: A coefficient value of 0.233748 means one unit increase in year variable increases bike rentals by 0.233748 unit.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm which used to predict the relationship between variables. There is an equation of linear regression

$y=mx+c$

where, y = dependent variable

x = independent variable

c = intercept

m = slope

$y=mx+c+e$

e = error

Where it depicts the as straight line when model is a good fit and shows a linear relationship between dependent and independent variable. Dependent variable is the target variable or numerical variable where independent variable can be numerical or categorical variable. Good model occurs when there is less difference between actual value and predicted value that means less error and as per assumption error terms should be normally distributed.

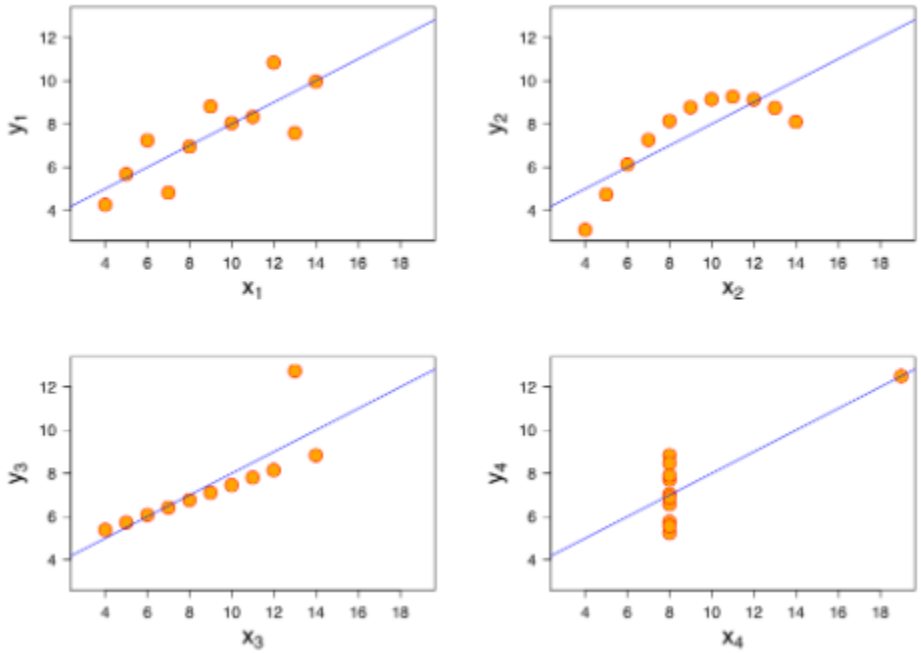
There are two types of regression model –

Simple Linear Regression (SLR) – when there is only one independent variable.

Multiple Linear Regression (MLR)-When there is more than one independent variable.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a representation of scatterplot which contains four datasets with identical simple descriptive statistics that means same mean and variance for each x and y in four datasets. When we visualize the data, it looks completely different from each other. It helps to identify outliers, linear separability of data, diversity of data before visualizing it.



Dataset 1: represents good fit of linear model

Dataset 2: represents non-linear model

Dataset 3: its shows outliers which can't be handled by linear data

Dataset 4: depicting outlier and not a linear model

3. What is Pearson's R?

Pearson's R is a linear relationship between two quantitative variables and r lies between -1 and 1.

We can interpret it in three different ways.

Positive Correlation: When both the variable increase or decrease in same direction and lies between 0 to 1.
Negative Correlation: When one variable increase, other decrease and they are inversely related and r lies between -1 to 0
No Correlation: When there is no correlation between two variables then r is 0.

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

Feature Scaling is a data pre-processing technique used to normalize or standardize independent variables within particular range. It is done to ensure that all features contribute equally to the model when features have different ranges. It helps to improve performance and efficiency of the model.

Importance of scaling:

1. It improves model convergence.
2. Scaling ensures equal distribution of features.
3. It helps to solve numerical issues.

There are two types of scaling.

Normalization/Min-Max Scaling: It brings all the data between 0 to 1. It depends on minimum and maximum value and needs to check outliers as it will impact the data. It is used for KNN and nueral network.

X = (x – min(x)) / (max(x) – min(x))

Standardization Scaling: It brings all the data into standard normal distribution with mean is 0 and standard deviation is 1. It is used for linear regression and logistic regression.

X = (x – mean(x)) / standard deviation(x)

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

VIF is variance inflation factor which is used to detect the presence of multicollinearity. Multi collinearity occurs when there are multiple independent variables are highly correlated. VIF measures how much the variance of a regression coefficient is inflated due to collinearity with another predictor.

VIF(X) = 1 / (1 – R^2)

R^2 = coefficient of determination
If R^2 is 1 then perfect multicollinearity
If denominator (1-R^2) is 0 then VIF is infinite or perfectly collinear. One predictor variable is linearly related with one or more predictor variable.

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

Q-Q plot is quantile-quantile plot. Its kind of scatterplot with the two-quantile variable. It determines whether two sample of data coming from same population or not which is nothing but probability distribution or data is normally distributed or not. It is also applicable for different distribution. If two distributions are exactly equal that means plot is perfectly lie on a straight line.