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?**

Ans: Following are some conclusions as per my analysis of the categorical variables from the dataset:

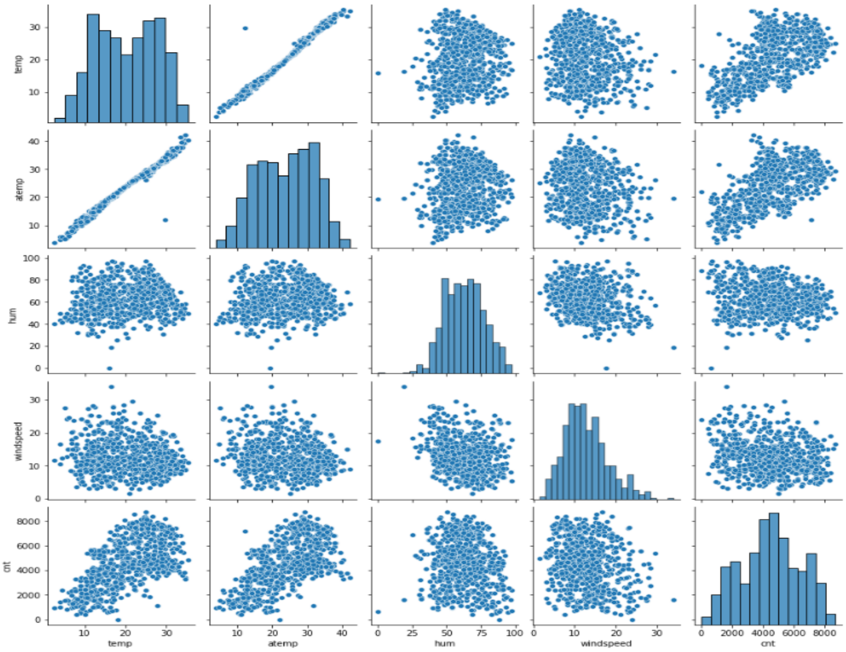
- Yearly the demand of bike sharing is increasing
- Demand of bike sharing is increasing in non-holydays as compare to holyday.
- Demand of bike sharing is more when the weathersit is clear and few clouds.
- Demand of bikes are increasing till month sep then from month oct it starts reducing till dec.
- Demand is almost same in all the weekdays.
- Demand is same in working days and non-working days.
- Demand of bikes was lowest during the spring season.

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

Ans: The drop_first=True command is used to avoid multicollinearity if we don't drop the dummy variables, they will correlate and affects the model adversely.
A variable with **n** levels can be represented by **(n-1)** dummy variables as we drop the first column, we can also represent the data.
We have a few categorical columns in our dataset where drop first might be utilised.
Example: shift =
pd.get_dummies(data[['season','weekday','mnth','weathersit']],drop_first=True)

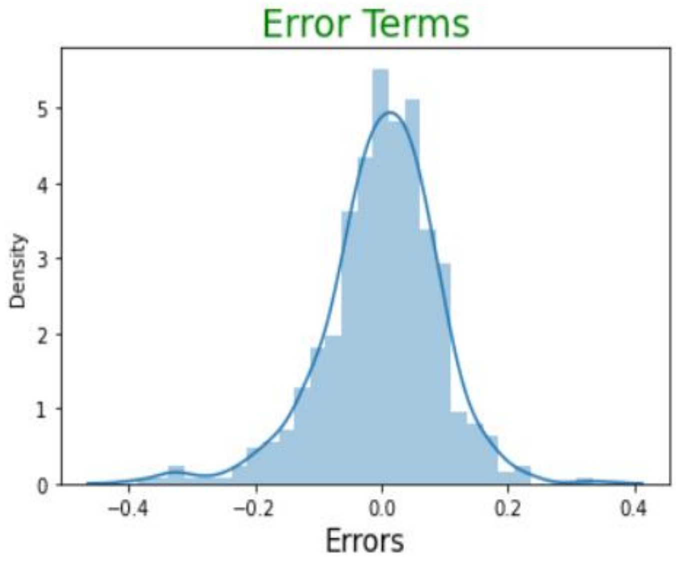
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

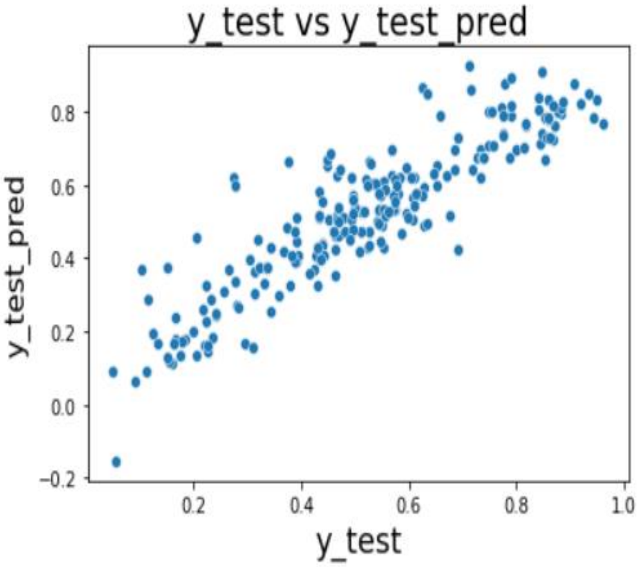
Ans: The variable Temp (Temperature) is highest correlation with target variable (cnt).



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

Ans: After developing the model on the training set, I do Residual Analysis between predictions and actual values to check the errors follow normal distribution. And maintains linear relation between dependant variable (test and predicted)





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

Ans: The top 3 features contributing significantly towards explaining the demand of the shared bikes are followed:

1. Temp (Temperature) with positive relation
2. Yr (Year) with positive relation
3. Weathersit - Light Snow & Rain with negative relation

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans: Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent variable (X) and dependent variables (y). Increases or decreases in the values of one variable have the same effect on the other variable in a linear relationship. It is mostly used for forecasting.

Linear regression is of the 2 types:

- **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line.

  Mathematical Equation: $\mathrm{Y} = \beta 0 + \beta 1X$

- **Multiple Linear Regression:** It explains the relationship between one dependent variable and more than two independent variables.

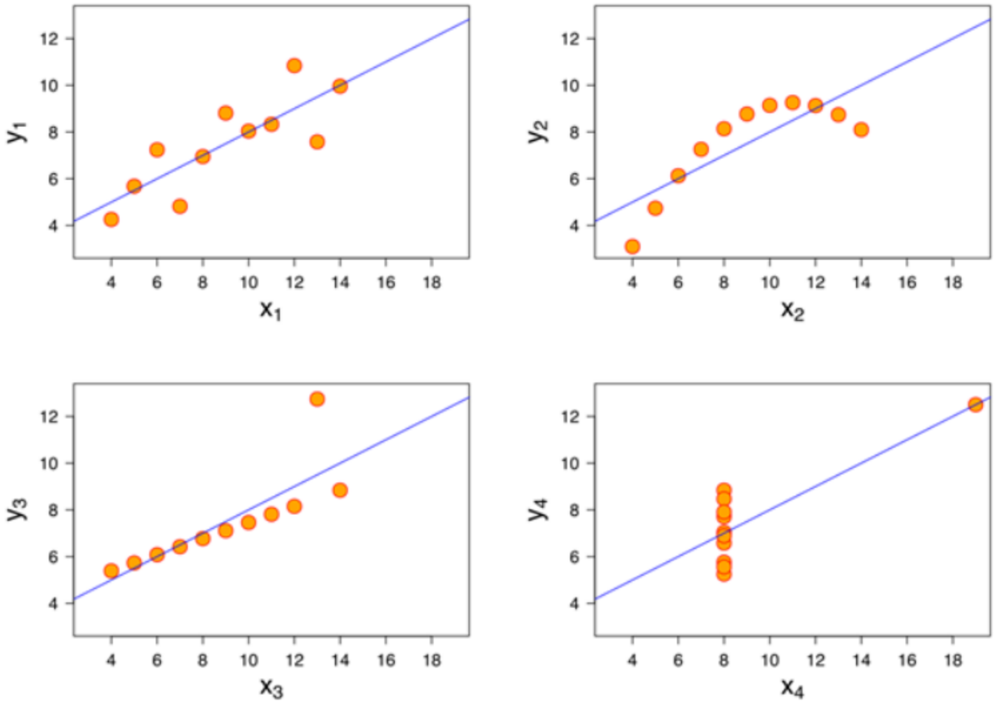
  Mathematical Equation: $\mathrm{Y} = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta \mathrm{p}X\mathrm{p}$

  Where,

  - Y is target/dependent variable
  - X1/X2 …. /Xp are independent variables
  - β1/ β2/ …. / βp is the coefficient of X
  - β0 is the intercept

2. Explain the Anscombe's quartet in detail.

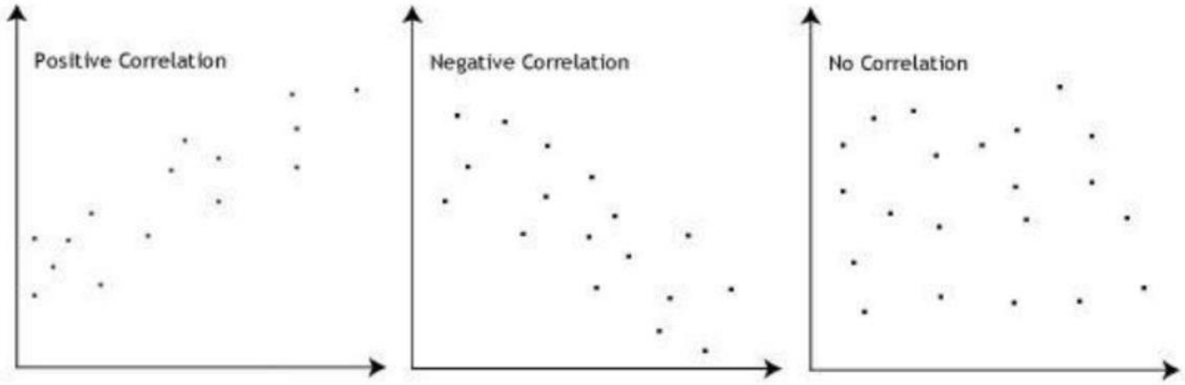
Ans: Anscombe's quartet are of four datasets that have almost same statistical features but seem substantially different when displayed. Each dataset has eleven (x, y) points. It was built by statistician Francis Anscombe to highlight the significance of charting data before analysing it and its impact of outliers on statistical features.



- 1st dataset I appears to have clean and well-fitting linear models.
- 2nd dataset II is not distributed normally.
- 3rd dataset is linear distributed but it should have a different regression line (a robust regression would have been called for). The calculated regression is offset by one outlier that appears to be far from the line.
- 4th dataset shows that one outlier is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Ans: Pearson's R is a correlation coefficient which is use to measure the strength of a linear association between two variables and it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is denoted by 'r'.



Mathematical Equation:

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

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

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

Ans: Scaling is use to normalising the range of independent variables. It is done to bring all of the independent variables on the same scale, as continuous variables have high scale as compare to categorical variable. So, it is a very important step in data pre-processing. Scaling has no effect on the p-value, r-squared value, or other parameters.

Normalization scaling or Min-Max scaling brings all the data in the range of 0 and 1.
Min-Max scaling: X = $(\mathbf{x}-\mathbf{min}(\mathbf{x}))$ / $(\mathbf{max}(\mathbf{x}) - \mathbf{min}(\mathbf{x}))$

Standardization scaling replaces the values by z-scores. It brings all the data into standard normal distribution which as mean(µ) as 0 and standard deviation(σ) as 1.
Standardization scaling: $\mathbf{x}$ = $(\mathbf{x}-\mathbf{mean}(\mathbf{x}))$ / $\mathbf{sd}(\mathbf{x})$

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

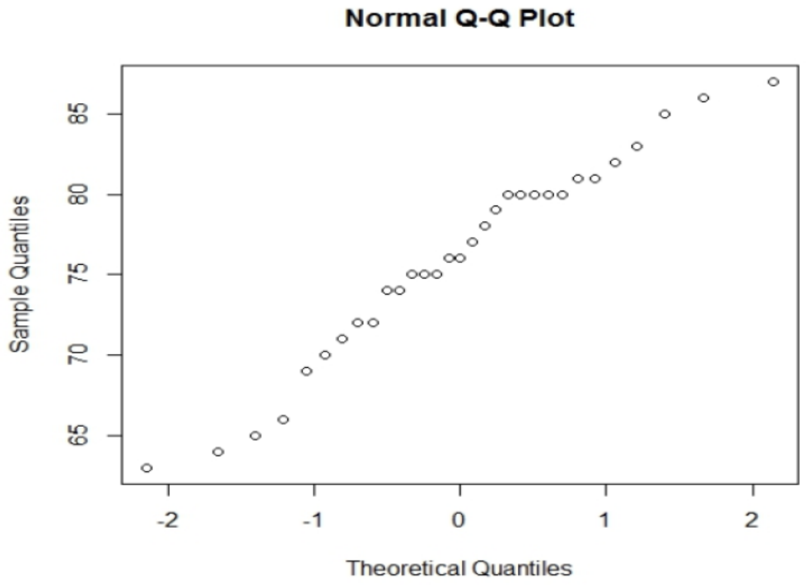
Ans: The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the R-squared statistic of the regression where the predictor of interest is predicted by all the other predictor variables. The variance inflation for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

When R-squared reaches 1, VIF reaches infinity. When R-squared reaches 1 then it means multicollinearity exists. Different variables are highly correlated with each other.

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

Ans: The Q-Q plot, also known as the quantile-quantile plot. It is a graphical tool use for detecting if two data sets are from populations with a similar distribution or not.
A Q-Q plot is a scatterplot formed by charting two quantile sets against each other. If both sets of quantiles originate from the same distribution, the points should form a relatively straight line. The example of a Normal Q-Q plot is shown below, where both sets of quantiles are drawn from Normal distributions.



The Application of the Q-Q Plot in Linear Regression: The Q-Q plot is used to determine if the points are roughly on the line. If they don't, it suggests our residuals aren't Gaussian (Normal), and our errors aren't either.

The Importance of the Q-Q Plot:

- Sample sizes do not have to be equal.
- Many distributional features may be investigated at the same time. For instance, variations in position, scale, symmetry, and the existence of outliers.
- The q-q plot, rather than analytical approaches, can give more information about the nature of the difference.