1) The best fit line as derived from multiple linear regression model is as follows:
cnt=0.1945+(0.2292*yr)+(-0.5558*holiday)+(0.0444*workingday)+(0.5301*temp)+(-0.1692*hum)+(-0.1857*windspeed)+(0.1039*season_2)+(0.1348*season*4)+(0.0529*weekday_6)+ (0.560*mnth_8)+(0.1255*mnth_9)+(0.0411*mnth_10+(-0.582*weathersit_2) +(-0.24868*weathersit_3).
This can be interpreted as follows:
    - yr (coeff 0.2292): Indicates that a unit increase in the variables 'yr' causes increase in demand by 0.2292.
    - holiday (coeff -0.5558): Indicates that a unit increase in the variables 'holiday' causes decrease in demand by 0.5558
    - workingday (coeff 0.0444): Indicates that a unit increase in the variables 'workingday' causes increase in demand by 0.0444.
    - temp (coeff 0.5301): Indicates that a unit increase in the variables 'temp' causes increase in demand by 0.5301.
    - hum (coeff -0.16928): Indicates that a unit increase in the variables 'hum' causes decrease in demand by 0.1692
    - windspeed (coeff -0.1857): Indicates that a unit increase in the variables 'windspeed' causes decrease in demand by 0.1857
    - season_2 (coeff 0.1039): Indicates that a unit increase in the variables 'season_2' causes increase in demand by 0.1039.
    - season_4 (coeff 0.1348): Indicates that a unit increase in the variables 'season_4' causes increase in demand by 0.1348.
    - weekday_6 (coeff 0.0529): Indicates that a unit increase in the variables 'weekday_6' causes increase in demand by 0.0529.
    - mnth_8 (coeff 0.0529): Indicates that a unit increase in the variables 'mnth_8' causes increase in demand by 0.560.
    - mnth_9 (coeff 0.1255): Indicates that a unit increase in the variables 'mnth_9' causes increase in demand by 0.1255.
    - mnth_10 (coeff 0.0411): Indicates that a unit increase in the variables 'mnth_10' causes increase in demand by 0.0411.
    - weathersit_2 (coeff 0.582): Indicates that a unit increase in the variables 'weathersit_2' causes decrease in demand by 0.582.
    - weathersit_3 (coeff -0.2486): Indicates that a unit increase in the variables 'weathersit_3' causes decrease in demand by 0.2486.

2) When we use encoding to convert categorical variables to dummy variables, we use pd.get_dummies in Python. Suppose there are 'n' levels for the variable. Using pd.get_dummies converts the single categorical to 'n' dummy variables. If we use 'n' dummy variables, it can lead to multicollinearity. However, one of the assumptions of multiple linear regression is that the independent variables should not be correlated. Hence, using drop_first=True drops the first dummy variable so as to create 'n-1' dummy variables only. This helps in eliminating multicollinearity while creating dummy variables.

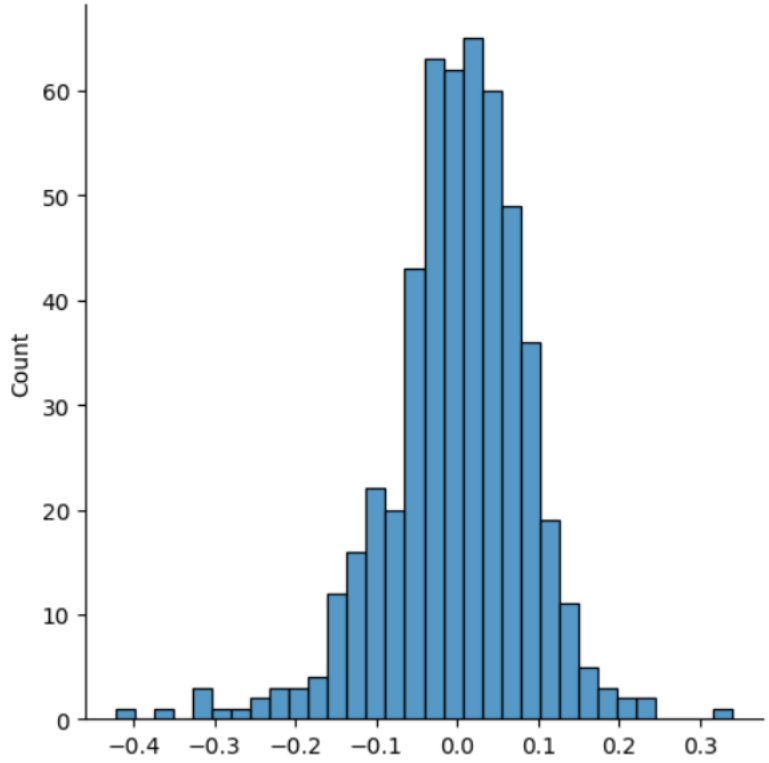
3) The variable ‘registered’ has the highest correlation with the target variable, followed by the variable ‘casual’ and ‘temp’.

4) The assumptions are validated as follows:
   - The error terms are normally distributed as shown in the plot below:

```
# Residual analysis of train data

er=df_train_y-y_train_price
sns.displot(er)
```

```
<seaborn.axisgrid.FacetGrid at 0x1ed19a12a90>
```



From this, it is clear that the mean of the error is 0.

- The VIF of all the independent variables are below 2, which validates the independence of the independent variables.
- The pairplot of the dependent variable ‘cnt’ and independent variables confirms the linear dependency between dependent and independent variables ‘registered’, ‘casual’ and ‘temp’.

5) The top 3 features contributing significantly towards explaining the demand of shared bikes are as follows:
   - temp (coeff 0.5301): Indicates that a unit increase in the variables ‘temp’ causes increase in demand by 0.5301.

- yr (coeff 0.2292): Indicates that a unit increase in the variables 'temp' causes increase in demand by 0.2292.
- weathersit_3 (coeff -0.2486): Indicates that a unit increase in the variables 'weathersit_3' causes decrease in demand by 0.2486.

General subjective Questions

1) Linear regression algorithm is used to predict a dependent variable using independent data variables. The relation is usually a straight line fit that best fits the data points as close as possible. Linear regression algorithm is used when the dependent variable to be predicted is a numeric value. Examples include sales value, revenue value, total count value etc.

   Linear regression can be simple linear regression which involves single independent variable or multiple linear regression which includes multiple independent variables.

   The equation for simple linear regression is y=mx+c; where y is the dependent variable to be predicted, 'm' is the coefficient and 'c' is the intercept.

   The equation for multiple linear regression is : y=c+m1x1+m2x2+….+mnxn
   Where m1,m2,m3.. are the coefficients and x1,x2… are the independent variables.

   Cost function:
   The algorithm used to identify the optimum values of the coefficients and intercept uses a cost function. This cost function is the summation of the difference between the actual and the predicted 'y' values and this cost function has to be minimized.

   Gradient descent:
   The goal of gradient descent is to minimize the convex function by parameter iteration. Arbitrary values are taken and it is iterated each time and the function is recalculated. The next iteration value is calculated using a learning rate 'n' and finally the value is converged.

   Assumptions of linear regression:
   - Linearity: There should be a linear relationship between the input and output variable.
   - Homoscedasticity: This means that the standard deviation and variance of the error (which is the difference between actual and predicted y values) must be the same of any value of 'y'.
   - The data should not have multicollinearity. This means that there should be any correlation between the independent variables.
   - The error terms are normally distributed with mean=0 and SD=1.
   - The error terms are independent of each other.

   Parameters to evaluate linear regression model:
   - R2: This gives the percentage of variance of the y variable that is explained by the model.

- P-value: This gives the level of confidence about the coefficients of the variable.
- Prob(F-statistic): This gives the significance of the overall model.

Process of modelling linear regression:
After exploratory data analysis is done, the categorical variables are converted to encoded to dummy variables as required. Then the data set is split to train and test data sets. The scaling fit and transform is done using a scaling technique like min-max scaling or standardization scaling for the train data set. Then, after adding constant, the model is built to calculate coefficients. This model is evaluated with parameters and optimized as per requirement. Then the r2 value of train data is calculated. Further, prediction is done on the test data using the built model. Then r2 value and the assumptions are validated on the test data.

2) Anscombe's Quartet can be defined as a group of four data sets which are nearly identical when calculating the statistical values like mean, variance, standard deviation. However, the datasets when plotted appear differently. This was constructed by the statistician Francis Anscombe to illustrate the importance of plotting the variables before building a model. It is necessary to plot the variables in order to identify whether the data is linear related and to identify the outliers present in the data.

Anscombe's quartet is as follows:

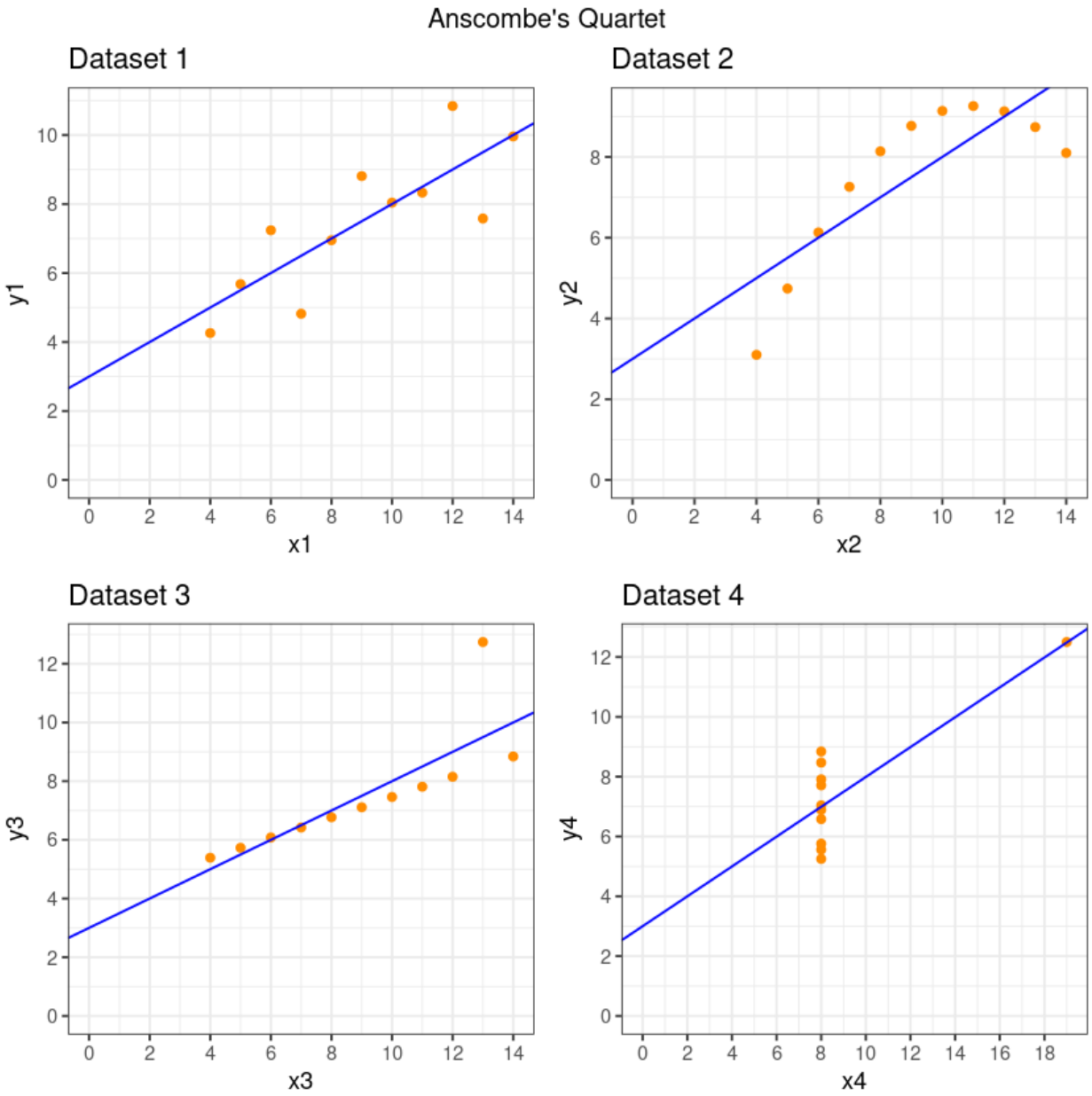
x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

The statistical description of the datasets are as follows:

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

On plotting the datasets:



From the above plots, it is evident that the datasets are extremely different from each other. This confirms the importance of plotting the datasets before building a model.

3) Pearson's correlation coefficient, denoted as R, measures the strength and direction of the relationship between two continuous variables. It can be any value between -1 and +1.

Formula for Pearson coefficient is as follows:

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

Where x and y are the variables.

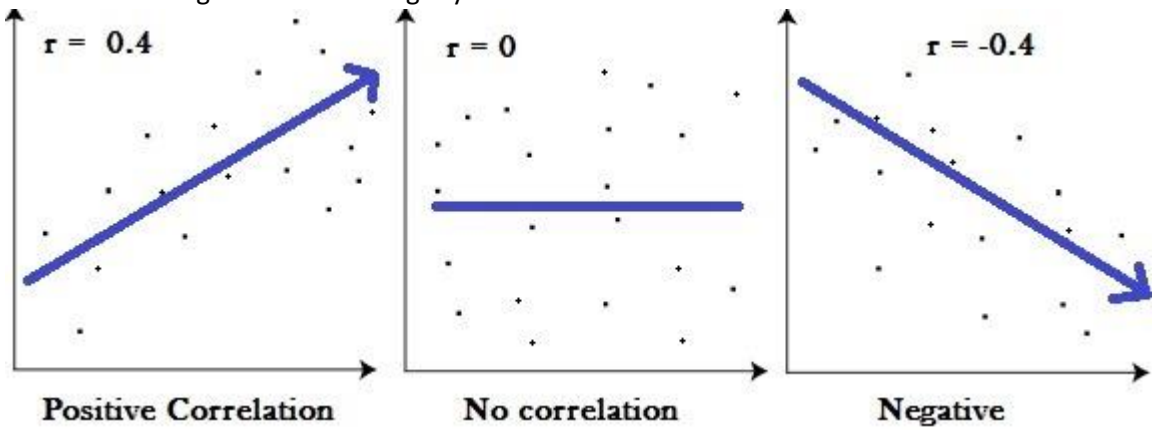
Value of 0: Specifies that there is no relation between the two variables.

Value >0: Indicates positive relation between the two variables. This means that when one variable increases, the other also increases.

Value >0: Indicates negative relation between the two variables. This means that when one variable increases, the other also decreases and vice versa.

Assumptions for a Pearson's correlation:

- Data must be from random or representative samples for meaningful statistical inferences.
- Both variables should be continuous and follow a normal distribution.
- Homoscedasticity is crucial, ensuring similar variance around the line of best fit.
- Extreme outliers, whether univariate or multivariate, impact the Pearson Correlation Coefficient. For instance, plotting age vs. loan amount reveals a correlation, but reversing the variables might yield inconclusive results.



3) Scaling is a preprocessing technique which is used to transform the values of a variable in a dataset to a similar scale. This will ensure that all features contribute equally to the model.

Scaling becomes necessary when the dataset has variables that have different ranges, units of measurement. If scaling is not done, the model building takes to account only the magnitude of the value of the variables and can generate incorrect model.
Scaling can also help in speeding up the convergence and improve the performance of the model since features would be on a similar scale.
The two different types of scaling are:

- Normalized scaling: Normalizing adjusts the values in a variable to a common scale. This helps to reduce the impact of different scales on the accuracy of the model. Normalizing shifts the values to a range between 0 and 1. It is also called Min-Max scaling.

  $$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

  Here, Xmax and Xmin are the maximum and minimum values of the variable. This ensures that all the values are between 0 and 1.

- Standardized scaling: In standardized scaling, the values are scaled such that the values are centred around mean=0 and standard deviation =1.

  X'=(X-u)/sigma

  Here, X'= Scaled value of the variable, X=Original value of the variable, u= Mean of the values and sigma is the standard deviation of the values.

- Standardized scaling is less sensitive to outliers, whereas normalized scaling is more sensitive to outliers.
- Standardized scaling may not preserve the shape of the dataset, but normalized scaling preserves the shape of the dataset.
- Standardized scaling preserves the relationship between the datapoints, but normalized scaling may not preserve the relationship between the datapoints.
- Standardization can be followed when data follows Gaussian distribution and normalization can be followed when the data does not follow Gaussian distribution.

4) VIF is Variable Inflation Factor. This is a measure of the multicollinearity in regression analysis. Multicollinearity exists when there is dependency between the independent variables in a multiple regression analysis. Multicollinearity does not affect the prediction of the model, but it will not be able to explain clearly the significance of particular variables on the model.

   VIFi = (1/1-Ri2); Here Ri is the model built with all variables except ith variable.

Hence, when VIF is infinte, it means that there is perfect correlation between the independent variables. If VIF of an independent variable is infinte, this indicates that the variable can be perfectly predicted by other variables in the model. This indicates very high multicollinearity.

The model can be made more efficient by removing multicollinearity. One method to reduce multicollinearity is to remove the variable and rebuild the model.

6)Q-Q plot is quantile-quantile plot which is a graphical technique which determines whether two data sets come from populations with a common distribution. The Q-Q plot is a plot of quantiles of the first data set against the quantiles of the second data set. A 45 degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line of 45 degree, the greater the evidence for the conclusion that the two data sets are from populations with different distributions.

The Q-Q plot can be used even when the sample size of the two data sets are not equal.
The advantage of Q-Q plot is that many distributional aspects of the data sets can be simultaneously checked with Q-Q plot. This includes shift in locations, shift in scales, changes in symmetry, presence of outliers in the data sets. For example, if the two data sets come from populations whose distributions doffer only by a shift in location, the points will lie along a straight line that is displaced either up or down from the 45 degree reference line.