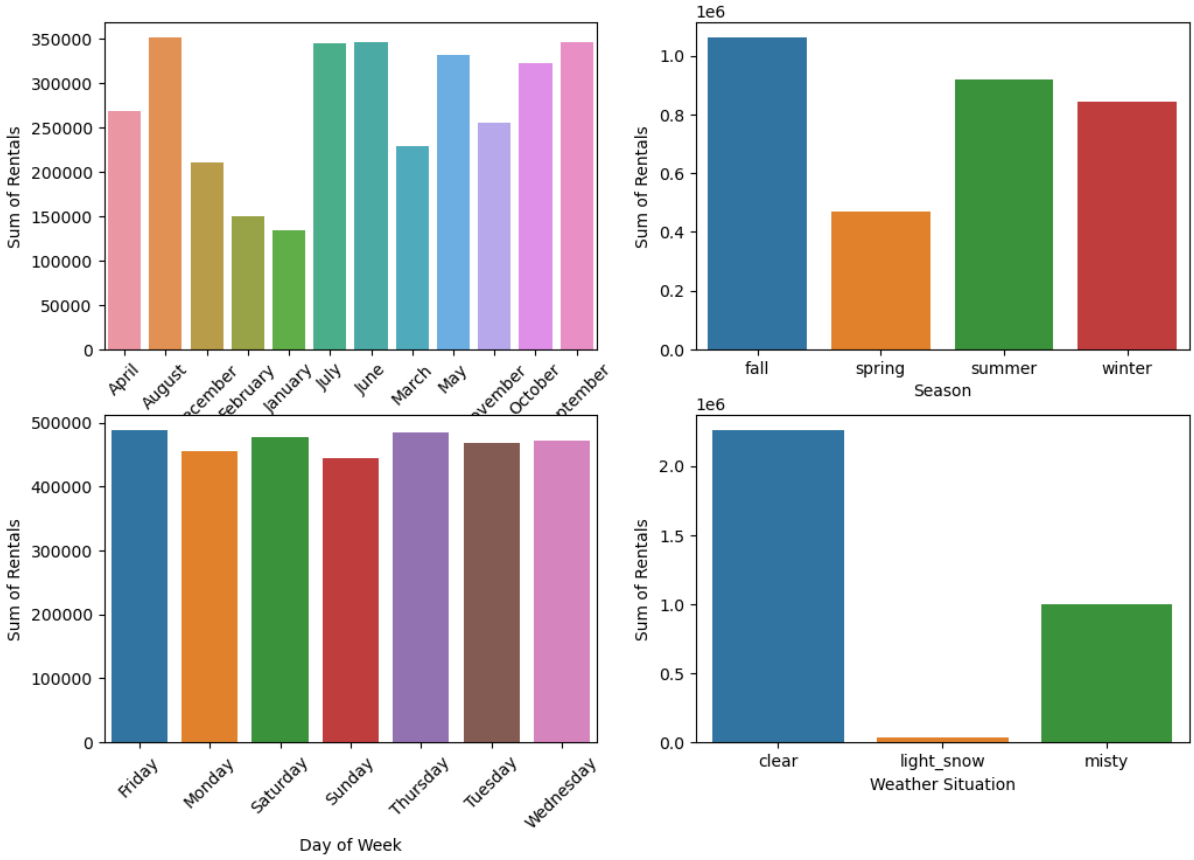
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   Answer:



   - The months June, July, August, and September have highest number of rentals whereas January, February and March have lowest Rentals
   - Fall season has the highest number of rentals whereas Spring season has lowest rentals.
   - There is not much to differentiate between days
   - When the weather is clear, the rentals are high whereas rentals are at their lowest when the weather is Snowing.

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

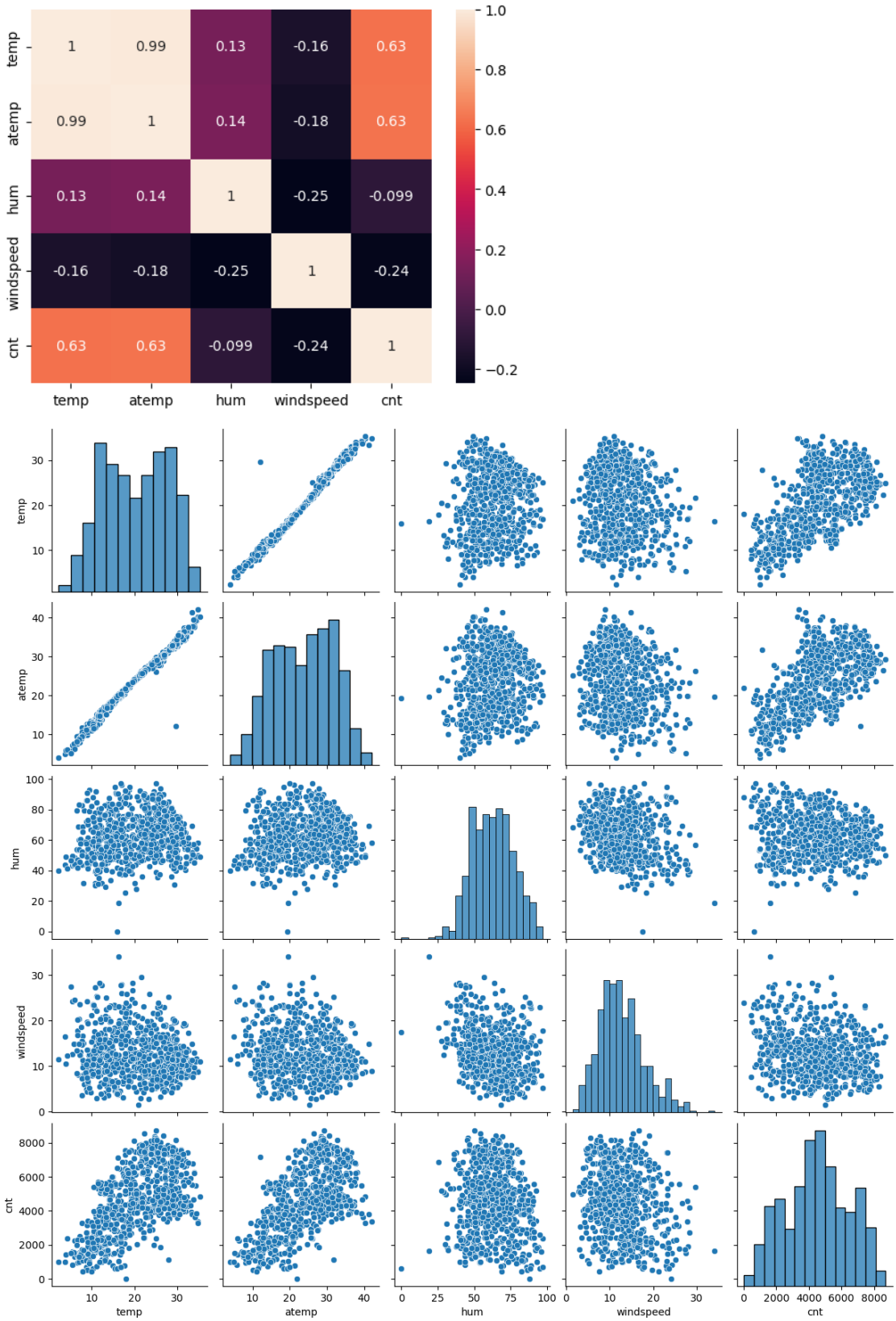
   Answer: If a categorical variable has K distinct values, then K columns will be created. drop_first=True drops one of the columns creating which in turn creates K-1 columns.

   Firstly, dropping a column results in one less column. The dropped column value can be derived by rest of the columns.

   Secondly, it reduces multicollinearity by reducing the correlation between variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:



From the heatmap and pairplot above, temp (and atemp) are the numerical variables which have highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

1. **Linearity**: Linearity is proved by p-value analysis from the model. As per the screenshot below, the p-value for all predictor variables is less than threshold (0.05).

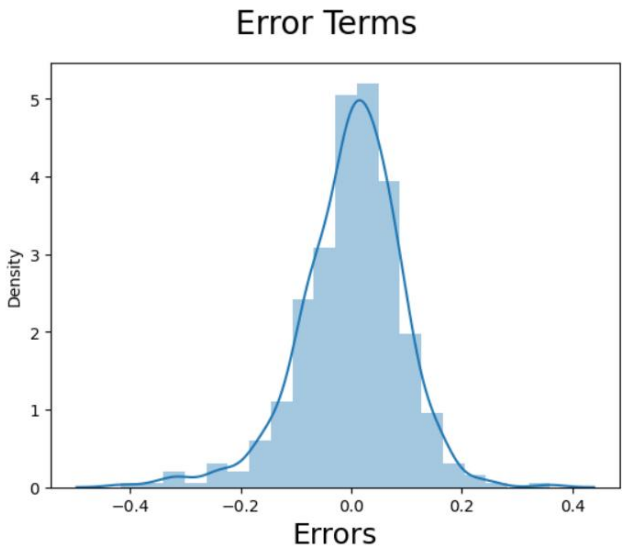
OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.833
Model:	OLS	**Adj. R-squared:**	0.830
Method:	Least Squares	**F-statistic:**	249.2
Date:	Sun, 07 Apr 2024	**Prob (F-statistic):**	7.36e-187
Time:	16:02:54	**Log-Likelihood:**	495.16
No. Observations:	510	**AIC:**	-968.3
Df Residuals:	499	**BIC:**	-921.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1910	0.030	6.456	0.000	0.133	0.249
yr	0.2341	0.008	28.246	0.000	0.218	0.250
holiday	-0.0969	0.026	-3.691	0.000	-0.148	-0.045
temp	0.4782	0.033	14.446	0.000	0.413	0.543
windspeed	-0.1482	0.025	-5.860	0.000	-0.198	-0.098
mnth_September	0.0909	0.016	5.565	0.000	0.059	0.123
season_spring	-0.0551	0.021	-2.641	0.009	-0.096	-0.014
season_summer	0.0610	0.014	4.271	0.000	0.033	0.089
season_winter	0.0959	0.017	5.730	0.000	0.063	0.129
weathersit_light_snow	-0.2860	0.025	-11.492	0.000	-0.335	-0.237
weathersit_misty	-0.0801	0.009	-9.090	0.000	-0.097	-0.063

2. **Normal Distribution of Errors**:

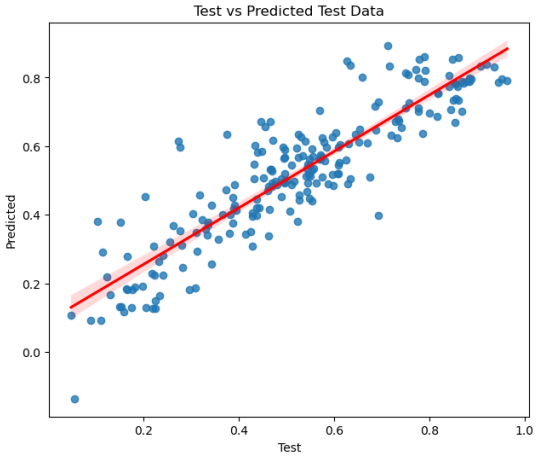
As per screenshot below, errors are normally distributed with mean close to 0




```
The mean of errors is -4.745659203299689e-16
```

3. **Error terms are independent of each other**
   Clearly, there is no pattern among the error terms.

4. **Constant Variance of Error Terms (Homoscedasticity)**
   There is no clear pattern in variance of error terms.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer: After the model building, 'temp', 'yr' and 'season_spring' are variables having most correlation with the 'cnt' variable.

```
In [17]: corr_df = data.corr()
         corr_df[['cnt']].sort_values(by='cnt', key=lambda x:abs(x), ascending=False)
```

Out[17]:

	cnt
cnt	1.000000
temp	0.627044
yr	0.569728
season_spring	-0.561702

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Answer: Linear Regression is a Statistical and Machine Learning algorithm. It is used to mathematically / statistically model the relationship between and target variable and predictor variables.

   It is mathematically represented as:

   $\mathbf{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon}$

   where:

   y is the target variable

   β_0 is the constant / intercept

   x_1, x_2 … x_n are the predictor variables

   ϵ is the error – the difference between actual and predicted values.

   In machine learning there are a few steps on how the model is built.

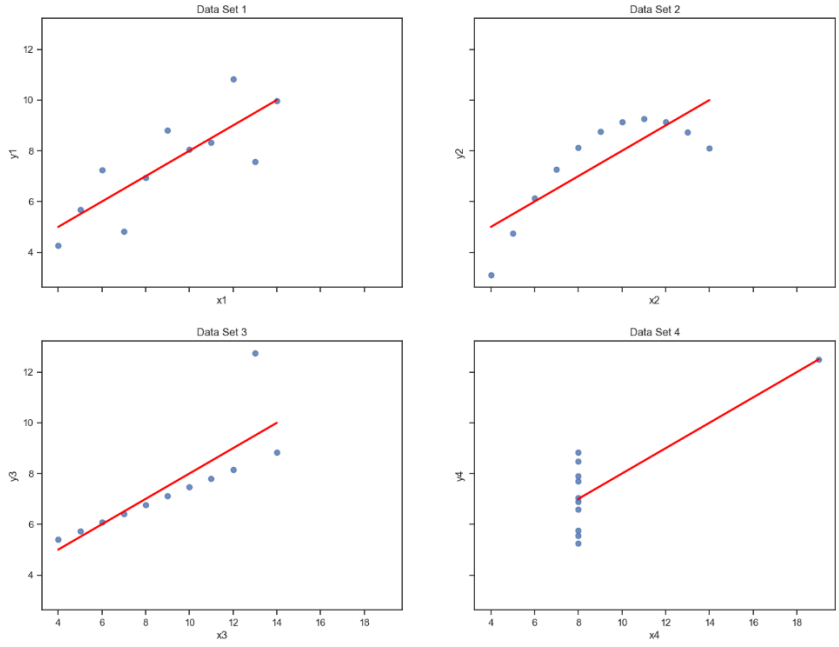
   i. The available data is cleaned up. Exploratory Data Analysis (EDA) is performed on the data. The data is made available for training. The data is analysed for relation between variables in the data set. The data is prepared to have most relevant columns.
   ii. The data is split into training data and test data.
   iii. Scaling is applied on training data which consists of continuous numerical variables.
   iv. Then the process of column selection begins.
   v. There are various ways in which column selection can happen. Recursive Feature Elimination (RFE) is one of them.
   vi. RFE is tweaked till required R-squared is achieved and all predictive variables show p-value within threshold (typically 0.05).
   vii. Then the model is examined for multi-collinearity using Variance Inflation Factor (VIF). In case some variables show higher VIF (generally the threshold is 5), the predictor variables are dropped one by one till the desired result is achieved.
   viii. Once the model is finalized, the model will predict the test data.
   ix. Then the model evaluation begins. Generally, the model is evaluated for Assumptions of Linear Regression and we make sure the assumptions are true.
      a. Linearity: Here it should be proved that all the predictor variables are linearly related to the target variable.
      b. Distribution of Errors: Here it should be proved that the distribution of Errors follows normal distribution and the mean of error terms is very close to zero.
      c. Independence of Error terms: Here it should be proved that the error terms when compared to target variables are constant and there is no pattern among the error terms.
      d. Homoscedasticity: Here it should be proved that there is no pattern in variance of error terms.
   x. When all the criteria are met, the final r2-score is calculated to make sure that the model is not over-fitting or under-fitting.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet was constructed by statistician Francis Anscombe in 1973. He built 4 data sets whose statistical parameters are the same. Each dataset has 11 co-ordinates. The parameters: mean, variance, standard deviation, correlation, Linear Regression Slope, Linear Regression Intercept, all have very similar values

Data Set:

Data Set 1		Data Set 2		Data Set 3		Data Set 4	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



Data Set 1 – represents a Linear relationship

Data Set 2 – shows a curve

Data Set 3 – Mostly Linear except for an outlier

Data Set 4 – Does not represent a Linear relationship.

This data shows the importance of visualizing the data before drawing statistical metrics and applying algorithms like Linear regression.

3. What is Pearson's R? (3 marks)

Answer: Pearson's R or Pearson's correlation co-efficient is a metric used to determine the strength and direction of a relationship between two continuous variables.

The value of the metric ranges from -1 to +1.

The formula for r is denoted by:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Where x_i and y_i are the two variables in question

$\bar{x}$ and $\bar{y}$ are means of the 2 variables

Explanation of r values:

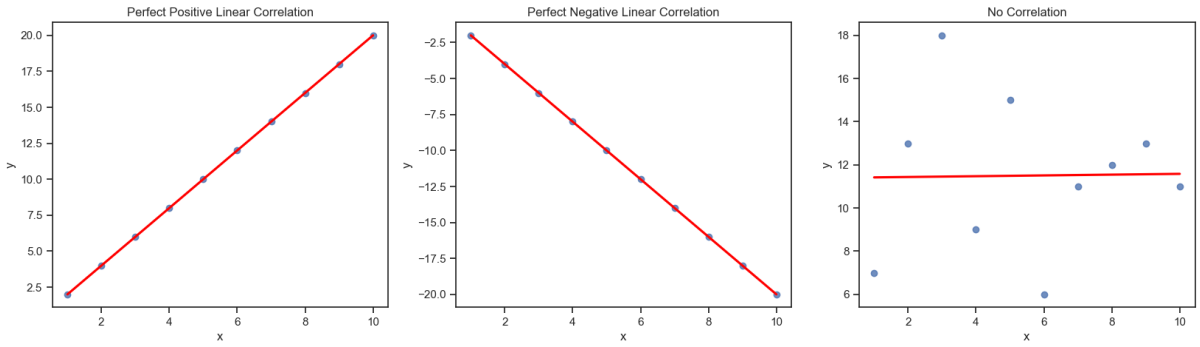
If r = 1 then it indicates perfect positive linear relationship.

If r = 0 then it indicates no linear relationship

If r = -1 then it indicates perfect negative linear relationship.

Generally, $|x| > 0.5$ suggests a strong correlation and $|x| < 0.3$ suggests a weak correlation.

Below graph illustrates a Perfect Positive Correlation, Perfect Negative Correlation and No Correlation.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling refers to the process of transforming numerical variables between 0 and 1.

Scaling is done to ensure that all the variables in data used for regression have similar scale. This is typically done to speed up some of the Machine Learning algorithms.

The most popular types of scaling are:

<u>Normalized Scaling</u>: This is also called Min-Max Scaling. When transformed, the values range from 0 to 1.

The formula for min-max scaling is

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

<u>Standardized Scaling</u>: This is also called Z-score normalization. When transformed, the values will have properties of standard normal distribution with mean=0 and standard deviation=1.

The formula for Standardized Scaling is:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Answer: When VIF has the value of infinity, it means that there is perfect multicollinearity between the predictor variables of a regression model. It means that the variable for which the VIF is infinity, one or more variables can perfectly explain and predict the target variable. To address this problem, generally the variable(s) having infinite VIF are dropped one by one and the VIF is recalculated on every iteration till we get the set of variables having VIF under threshold.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Answer: A Q-Q plot or quantile-quantile plot is a plot used to determine if a set of data follows a certain distribution like Normal Distribution.
   How the Q-Q Plot is constructed:
   - Data Set is sorted in ascending order.
   - Quantiles of data are calculated.
   - Quantiles of a given distribution (e.g. Normal Distribution) are calculated based on number of data points.
   - Now the observed quantiles are plotted against the distribution quantiles.

   The above plot is expected to be very close to a straight line if the data set is normally distributed.

   In Linear Regression, QQ plots are mainly used to validate the assumption of normality of error terms. This also means that we validate if the model is a good fit. QQ plot can also figure out if there are outliers in training data set which will skew the model training.