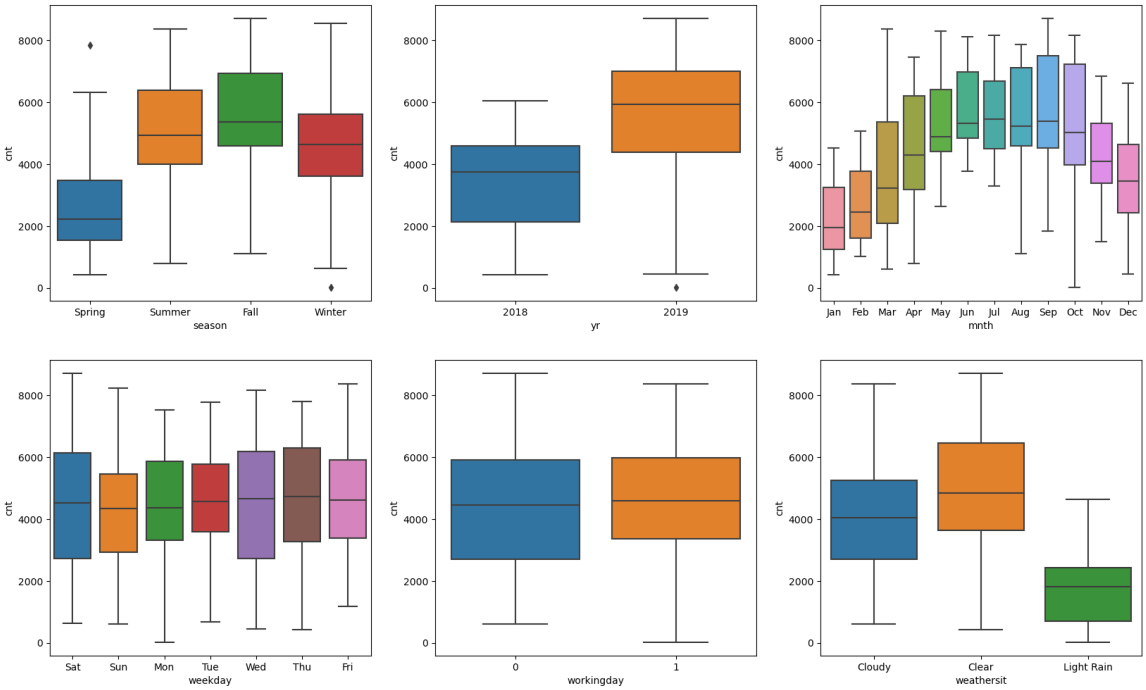
Assignment-based Subjective Questions

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

→ The following categorical variables were analysed:



Season:

Highest bike rental - Fall

Lowest bike rental - Spring

Year:

The demand of bike rentals has increased significantly in the year 2019 when compared to 2018

Month:

Maximum demand for bikes is observed in the months of September & October followed by August & June respectively.

Weekday:

The demand for bikes is highest on weekend, Saturday, followed by Sunday.
On weekdays, the demand is highest on Friday, followed by Wednesday, Thursday, Tuesday & Monday respectively.

Working day:

The demand for bikes is highest on working days.

Weather situation:

The demand for bikes is high when the weather is clear, followed by cloudy and light rain conditions.

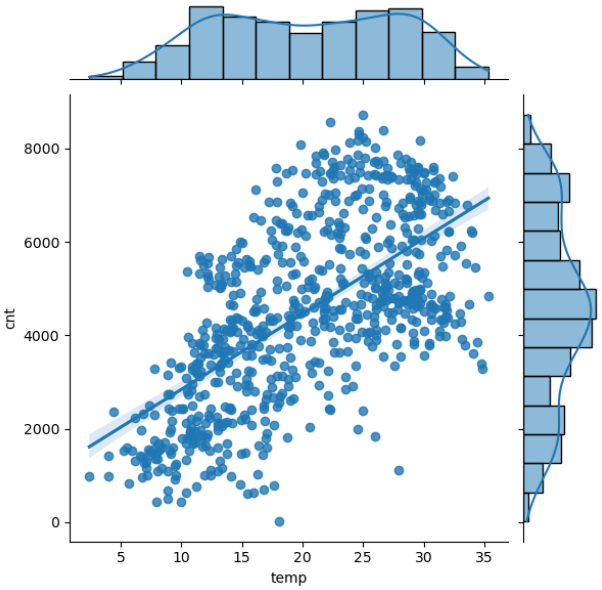
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

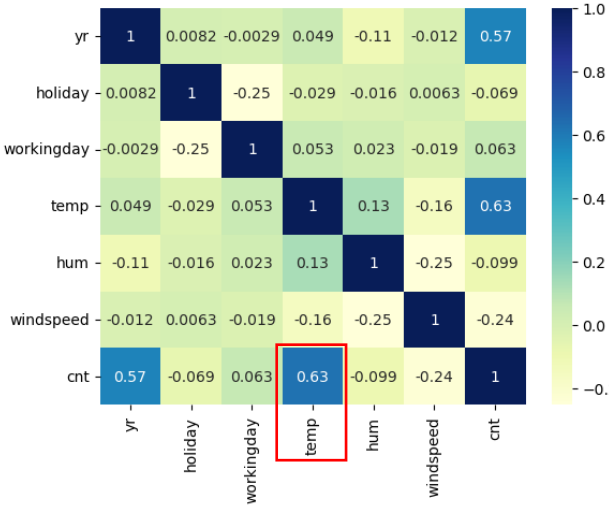
→ This has something to do with Multicollinearity in case of Multiple Linear Regression.

- Because, Keeping k dummies for k levels of a categorical variable is good idea, but there is a redundancy of one level, which is here in separate column. This is not needed since one of the combination will be uniquely representing this redundant column.
- Hence, it's better to drop one of the column and just have k-1 dummies(columns) to represent k levels.
- This Overall approach reduces Multicollinearity in the dataset, which is one of the prime Assumption of Multiple Linear Regression.

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

→ Temperature(temp) has the highest correlation with cnt.

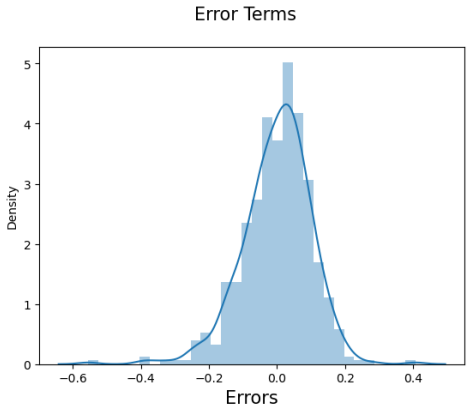




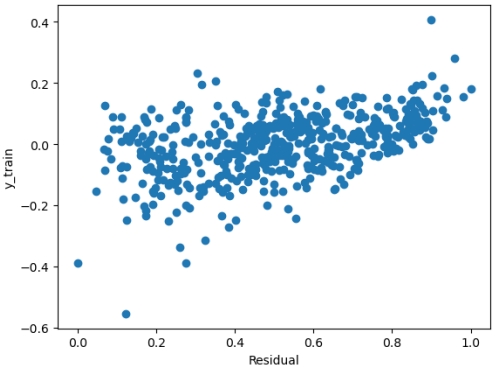
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

→ Assumptions are made based on the following properties:

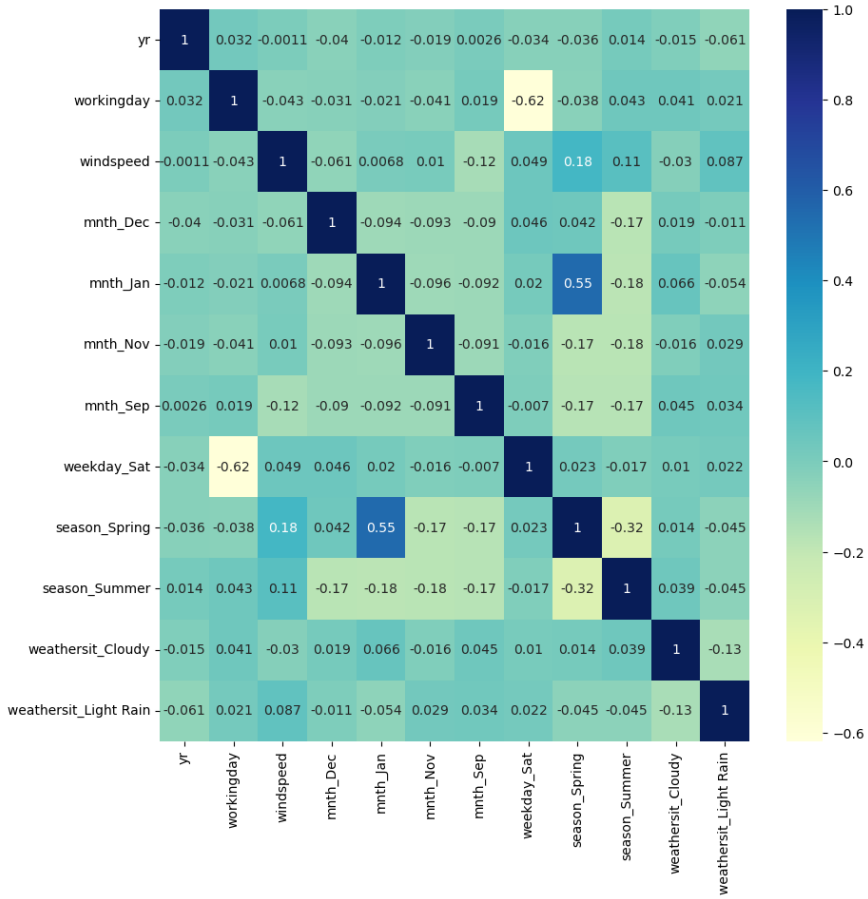
- **Normality of errors:** The error terms are normally distributed.



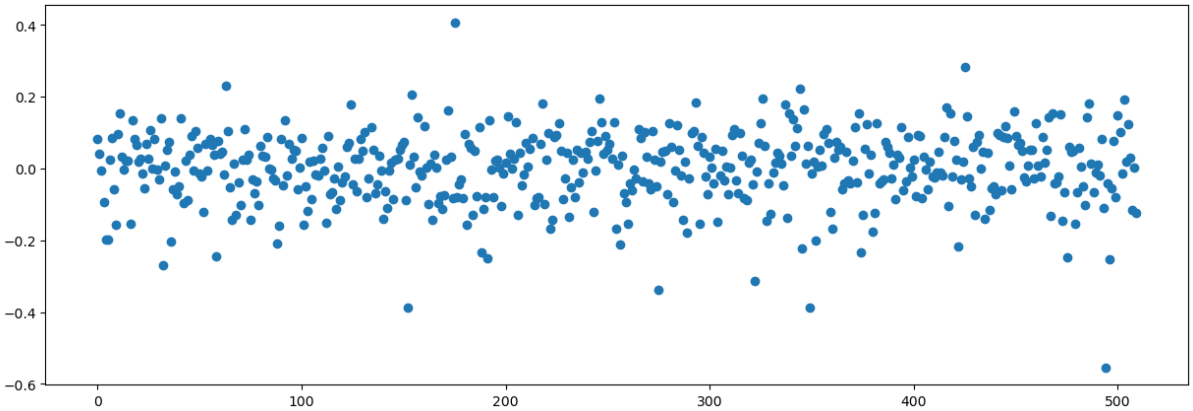
- **Linear Relationship:** The predictor and the target variables must have a linear relationship.



- **No Auto-correlation of error terms:** Error terms should be independent.
- **Absence of Multicollinearity:** There should not be high correlation between the predictor variables.



- **Homoscedasticity:** The error should be constant along the values of the predictor variables i.e., error terms are randomly distributed.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

→ Top 3 features obtained in the model:

1. Year (yr)
2. Workingday
3. mnth_Sep (Month of September)

General Subjective Questions

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

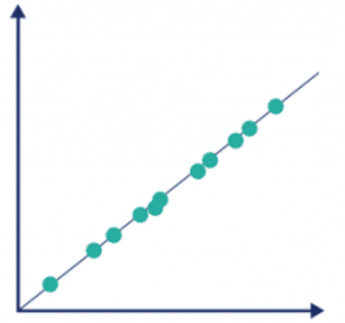
→ Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.
Linear relationship between variables means that when the value of one or more independent variables change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship is given by the equation: ***Y = mX + c***

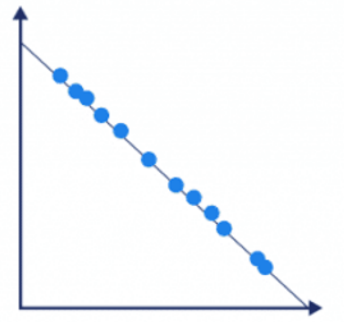
Here,
Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Also, the linear relationship could be positive or negative in nature as explained below:

- *Positive Linear Relationship*: A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph:



- *Negative Linear relationship*: A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph:



- Linear regression is of the following two types:
  - Simple Linear Regression
  - Multiple Linear Regression

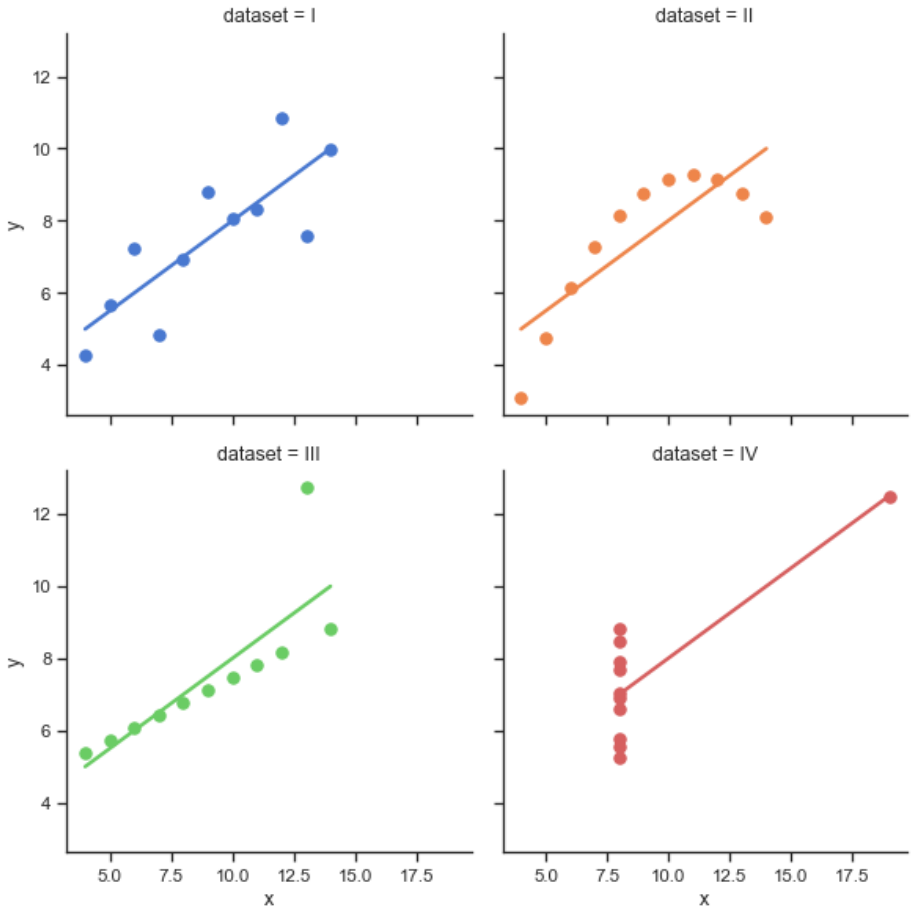
The following are some assumptions about dataset that is made by Linear Regression model

- Multi-collinearity: Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation: Linear regression model assumes that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables: Linear regression model assumes that the relationship between target and predictor variables are linear.
- Normality of error terms: Error terms should be normally distributed.
- Homoscedasticity: There should be no visible pattern in residual values.

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

→ Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built.

- They have very different distributions and appear differently when plotted on scatter plots.
- This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.
- Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:



The four datasets can be described as:

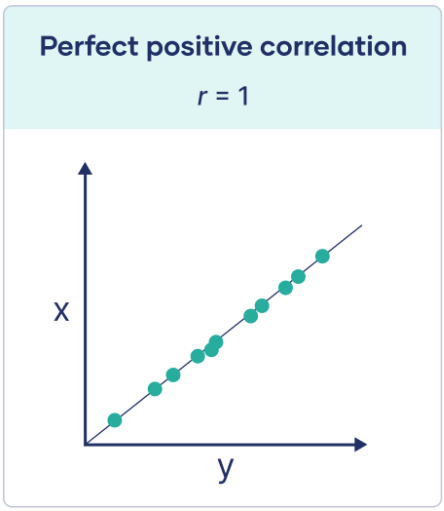
- **Dataset 1:** this **fits** the linear regression model pretty well.
- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
- **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

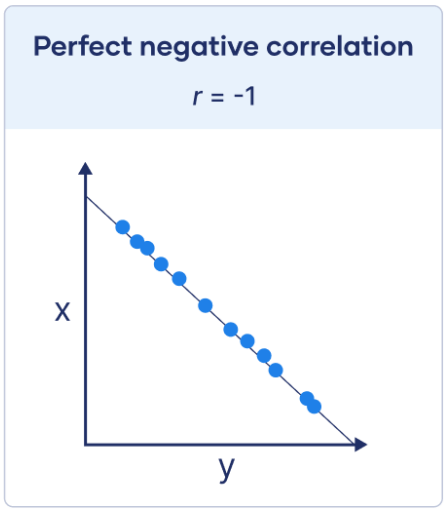
3. What is Pearson's R? (3 marks)

→ Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

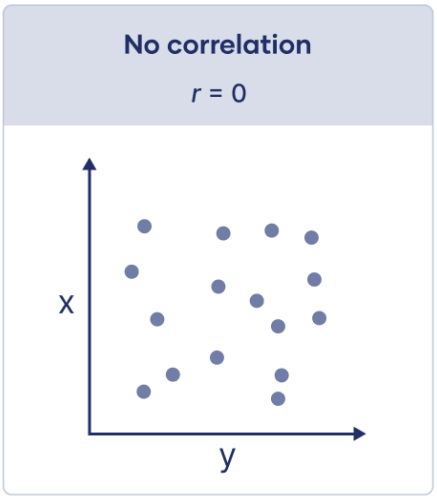
The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)





- r = 0 means there is no linear association



- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

→ It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why ?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization	Standardization
Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is often called as Scaling Normalization	It is often called as Z-Score Normalization

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

→ When the value of VIF is infinite it shows a perfect correlation between two independent variables.

Formula of **VIF = 1 / (1-R2)**

- So, if there is perfect correlation, then VIF = infinity.
- In the case of perfect correlation, we get R-squared (R2) = 1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this multicollinearity and re-calculate VIF.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

→ The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset.
- By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
- 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, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

- When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified.
- If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale.
- If two samples do differ, it is also useful to gain some understanding of the differences.
- The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.