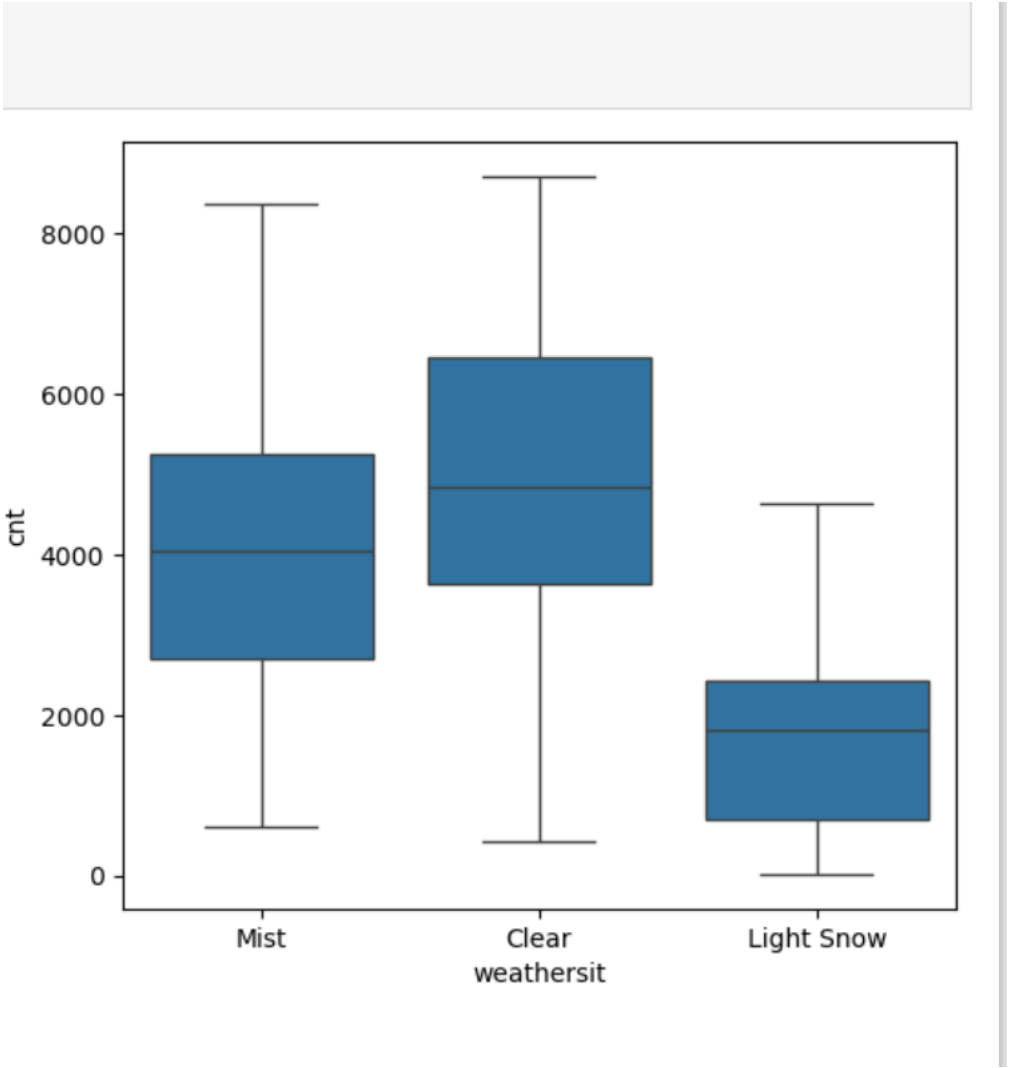
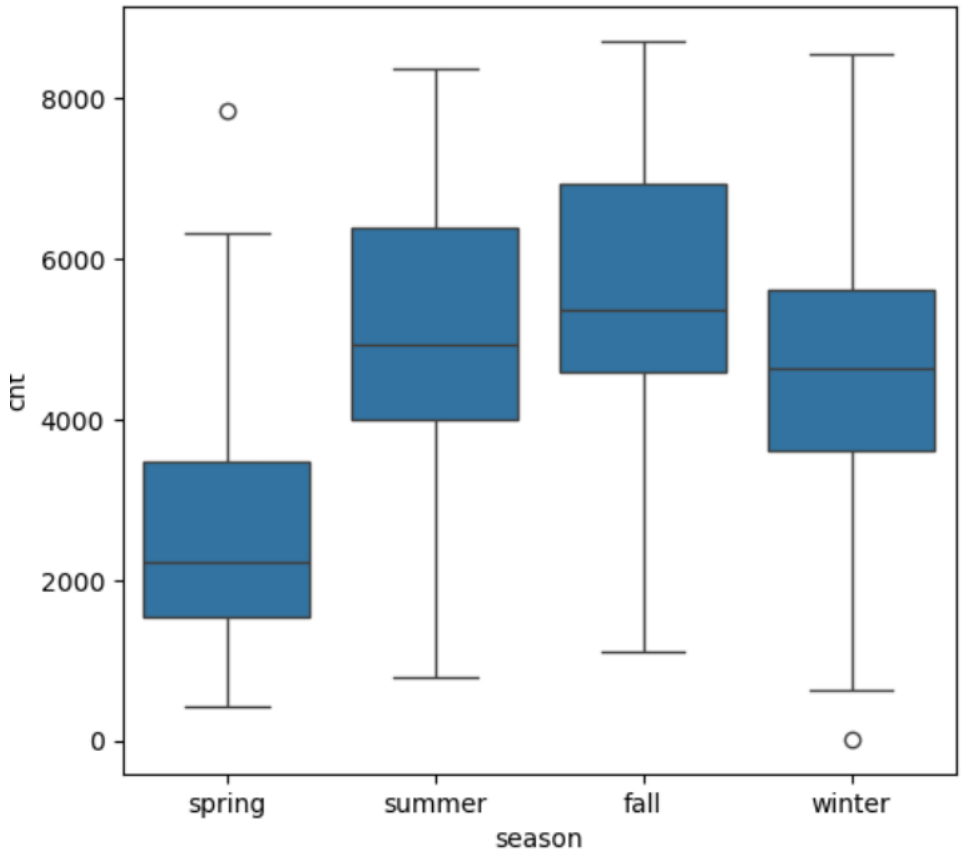
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    a. There are seven categorical variables
        i. Season, Year, Month, Weekday, Holiday, Working Day, Weather
    b. WeatherSit has a clear impact on our target variable i.e Count of Bookings received.
        i. There is raise in Bookings when the weather is Clear
        ii. Bookings Dip drastically when its Snowing/ Raining
        iii. Moderate decrease in bookings when there is Mist / its Cloudy.

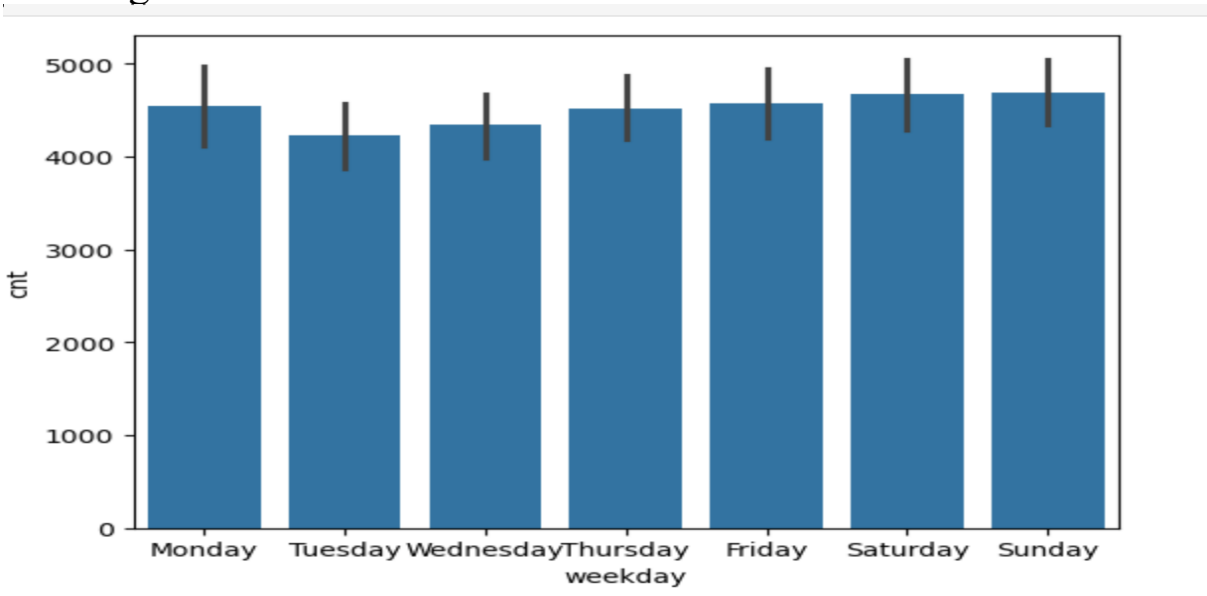


    c. Season : Bookings count varies based on season
        i. Spring : Bookings are dropped by significant number, Mean being around 2000 Bookings

ii. Summer & Fall : This season is more favourable for business, we see increase in orders in this season.
iii. Winter : There is a small dip in bookings but not as significant as Spring.

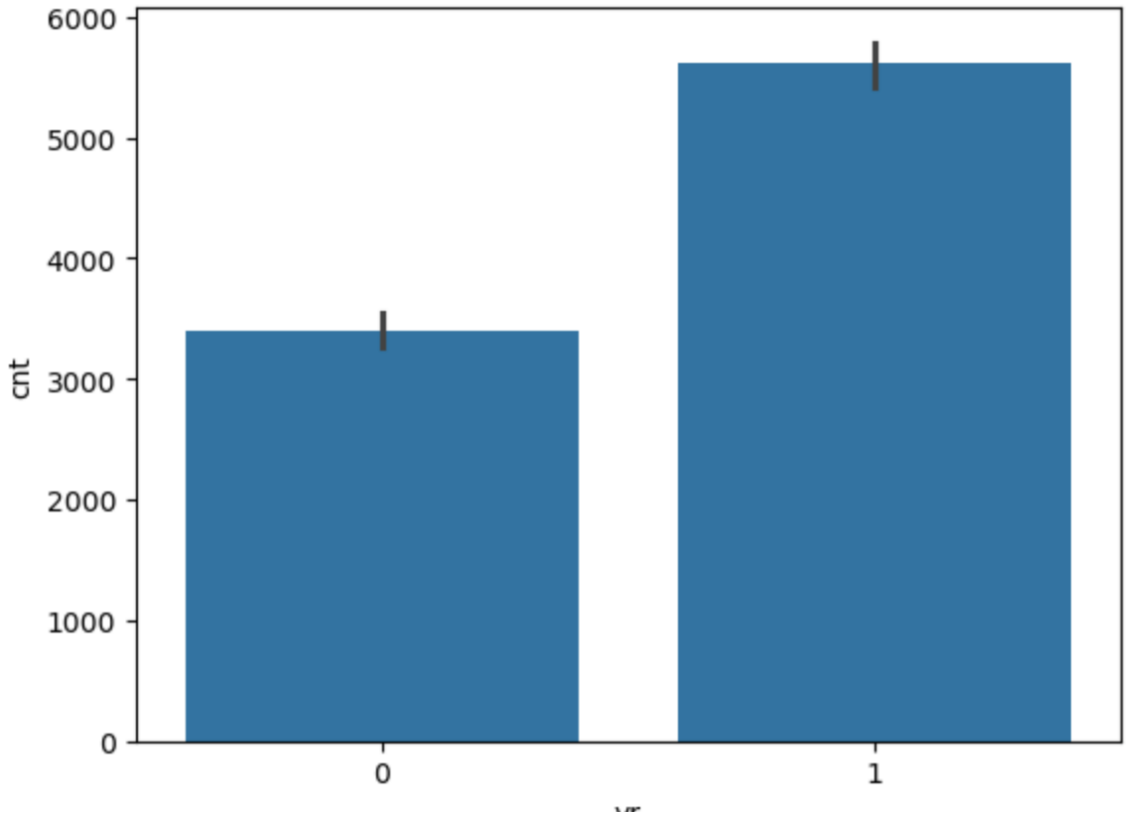


d. There is no much impact of Weekend and working day on Bookings.

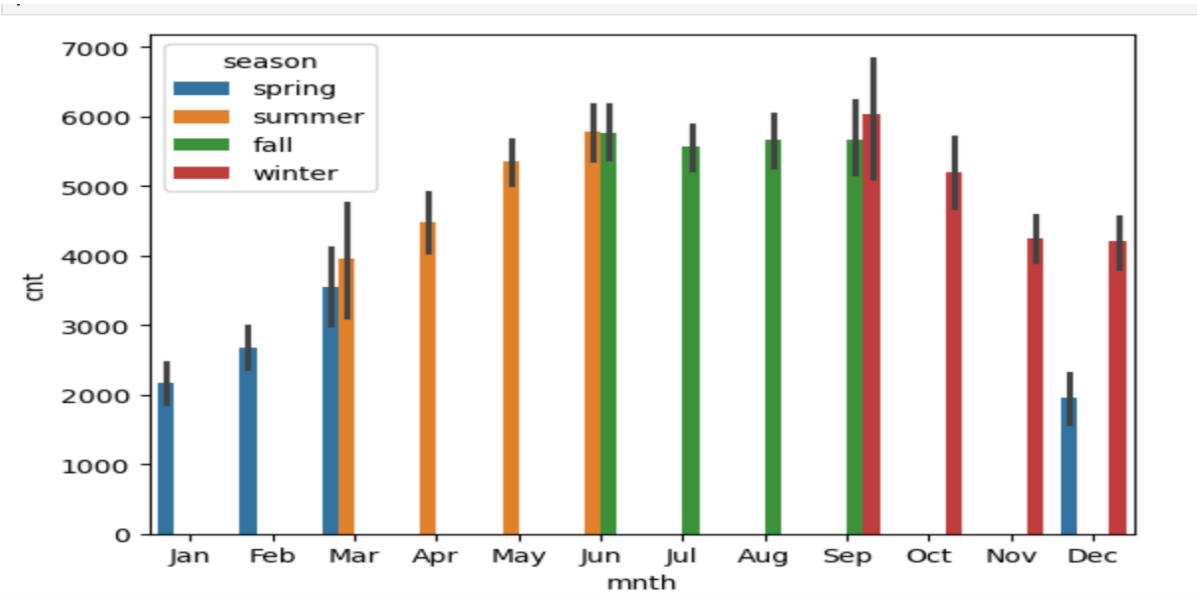


e. Year : There is significant raise in Bookings in the year 2019 compared to previous year

```
plt.show()
```

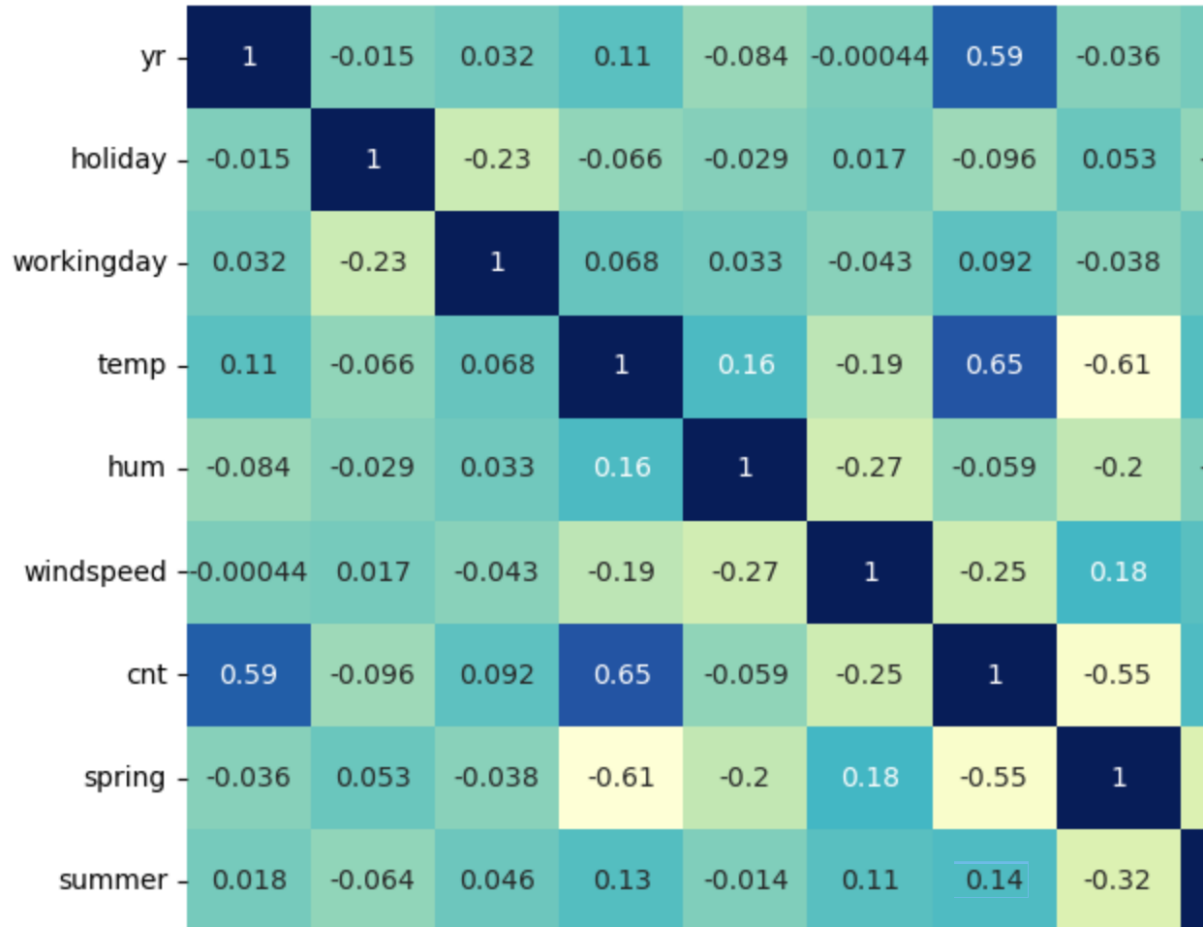


f. Month : Bookings are high in the month of March to September which is also Summer / Fall season as predicted earlier in season analysis.



g. HeatMap also shows that co-relation is high for Summer season, Year

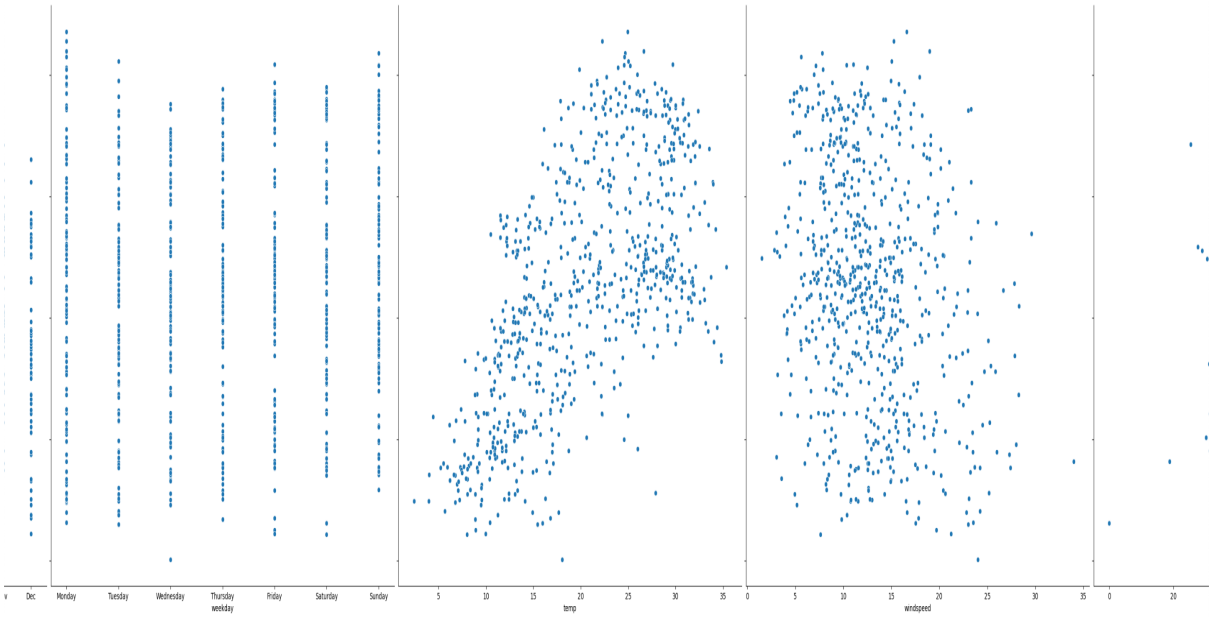
```
plt.show()
```



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

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
- If there are 3 Categorical variables we can infer additional value based on the 2 dummy variables created.

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



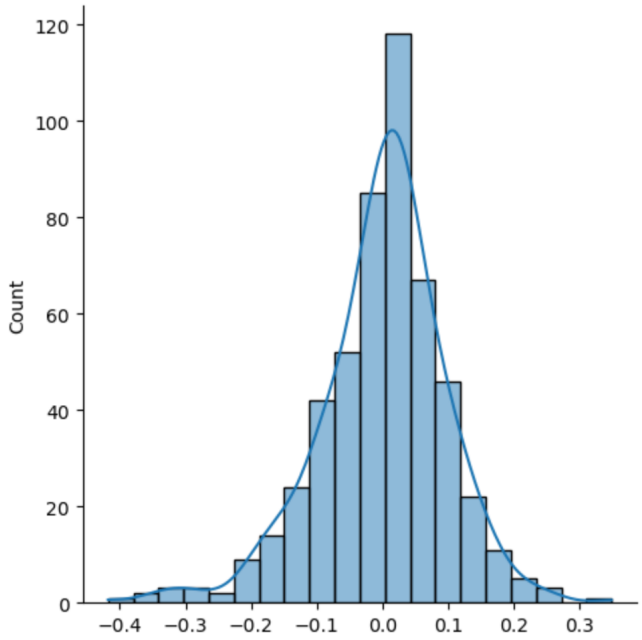
Looking at pair plot Temp has highest corelation with cnt(target variable)

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

one of the assumptions that you studied was that the error terms should be normally distributed with mean equal to 0. After the model is fit, calculated the Y predicted using the model and plotted the histogram for residuals (Y Actual – Y Pred) Residuals mean is Zero and residuals are normally distributed around mean.

```
plt.figure(figsize=(20,160))
sns.displot(y_train-y_train_pred, bins = 20, kde = True)
plt.show()
```

```
<Figure size 2000x16000 with 0 Axes>
```



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

Based on stats model summary,

1. Temp
2. Season
3. Year

These are the top features impacting the demand

Df Residuals:	500	**BIC:**	-866.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3293	0.024	13.474	0.000	0.281	0.377
yr	0.2318	0.009	26.286	0.000	0.214	0.249
holiday	0.0310	0.023	1.361	0.174	-0.014	0.076
workingday	0.1091	0.010	11.087	0.000	0.090	0.128
temp	0.3908	0.025	15.851	0.000	0.342	0.439
hum	-0.2145	0.033	-6.533	0.000	-0.279	-0.150
windspeed	-0.1978	0.027	-7.233	0.000	-0.251	-0.144
spring	-0.1551	0.013	-11.996	0.000	-0.181	-0.130
Light Snow	-0.2007	0.027	-7.414	0.000	-0.254	-0.148
Monday	0.1187	0.012	9.750	0.000	0.095	0.143
Tuesday	0.0705	0.013	5.593	0.000	0.046	0.095

Omnibus:	49.165	**Durbin-Watson:**	2.020
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	92.947
Skew:	-0.587	**Prob(JB):**	6.56e-21
Kurtosis:	4.731	**Cond. No.**	1.36e+15

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    - Linear regression is supervised machine learning algorithm.
    - Model is built from historical data , maps data points to optimised linear function which can be used for prediction of new data points.

- Linear regression computes the linear relationship between dependent feature to one or more independent features, when only one independent feature is present it is called Simple Linear Regression, if multiple independent features are present then its called Multiple Linear Regression.
- Simple Linear Regression : $y = \beta 0 + \beta 1 X$
- Multiple Linear Regression :
  $y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots\ldots\ldots \beta n Xn$

$\beta 1, \beta 2, \ldots. Bn$ are n independent features.

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

- Primary objective of Linear Regression is to locate best fit line with least residuals.
- A linear regression model can be trained using the optimization algorithm gradient descent by iteratively modifying the model's parameters to reduce the mean squared error (MSE) of the model on a training dataset.

Assumptions of Simple Linear Regression Algorithm

Linearity: The independent and dependent variables have a linear relationship with one another

Independence: The observations in the dataset are independent of each other.

Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant

Normality: The residuals should be normally distributed.

Assumptions of Multiple Linear Regression:

No multicollinearity: There is no high correlation between the independent variables.

Additivity: The model assumes that the effect of changes in a predictor variable on the response variable is consistent regardless of the values of the other variables.

Overfitting: Overfitting occurs when the model fits the training data too closely,

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

3. What is Pearson's R?

Pearson's correlation coefficient is a statistical measure that evaluates the strength and direction of the relationship between two continuous variables. It is considered the most effective method for assessing associations due to its reliance on covariance. This coefficient not only reveals the magnitude of the correlation but also its direction.

Key Assumptions:

Independence: Each case should be independent of others.

Linearity: There must be a linear relationship between the variables, which can be verified through a scatterplot. If the plot forms a straight line, the criterion is met.

Homoscedasticity: The scatterplot of residuals should approximate a rectangular shape.

Degrees of Correlation:

Perfect: Values near ±1 indicate a perfect correlation, where one variable's increase (or decrease) is mirrored by the other.
High Degree: Values between ±0.50 and ±1 suggest a strong correlation.
Moderate Degree: Values between ±0.30 and ±0.49 indicate a moderate correlation.
Low Degree: Values below +0.29 are considered a weak correlation.
No Correlation: A value of zero implies no relationship.

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

Scaling : 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.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and
1 sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
Formula = X – Xmin(x) / max(x)-min(x)

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

X = x-mea(x) / sd(x)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   VIF(i) = 1 / (1- (Ri)2)
   When R2 is 1 which indicates highest co-relation then VIF becomes infinity as 1-1 becomes 0.

   What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.