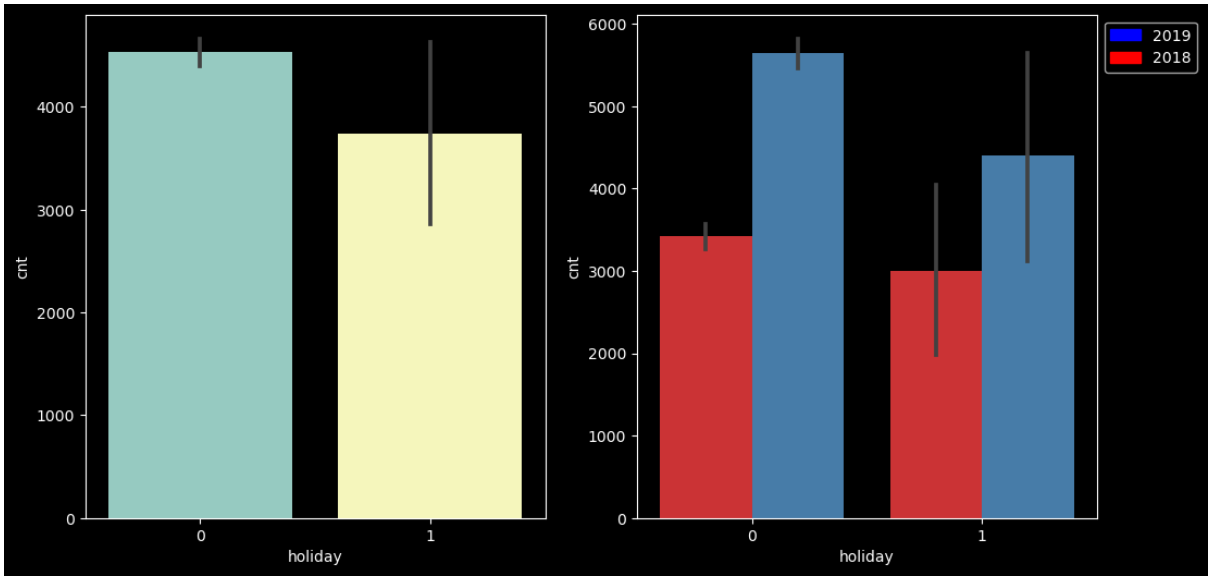
Assignment-Based Subjective Questions

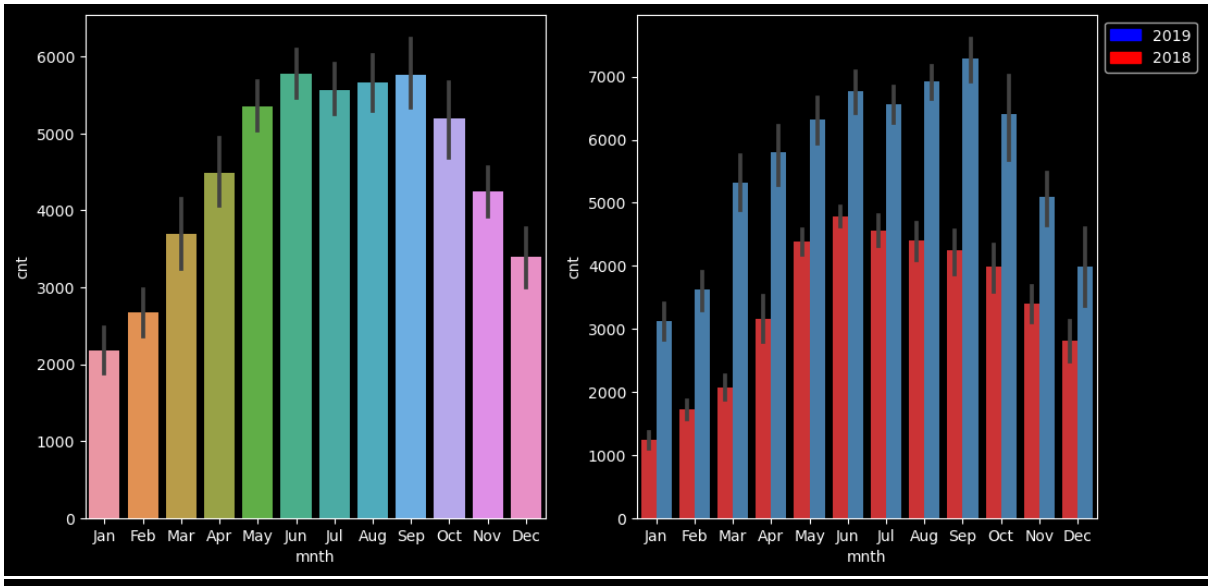
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

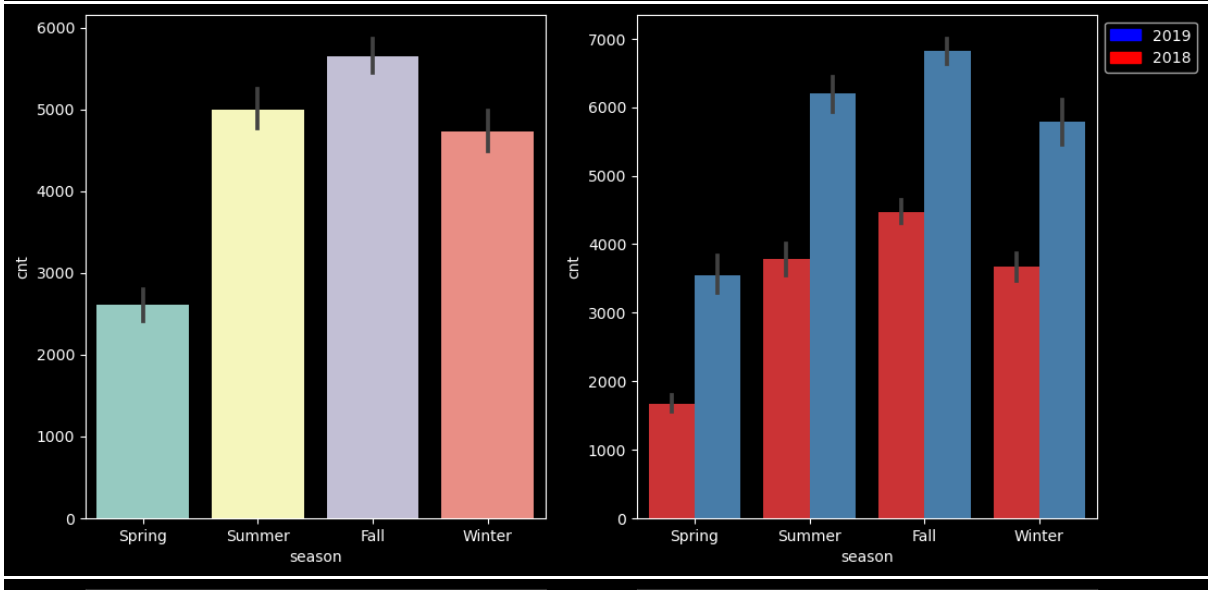
   Ans:

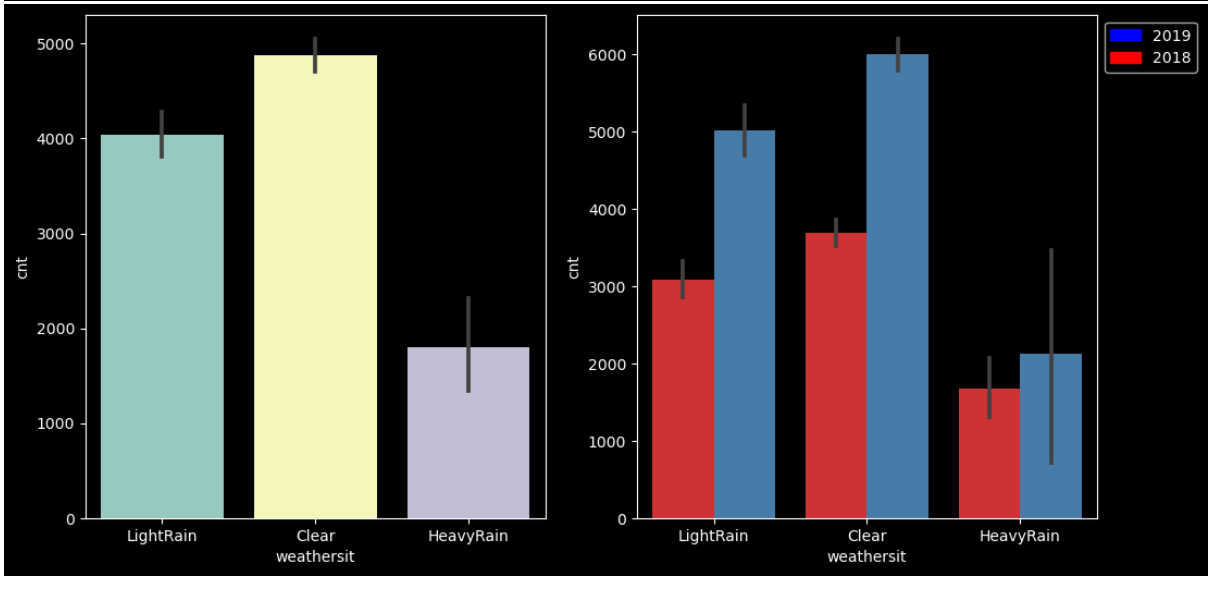
   Based on our understanding of the categorical data, the following inferences can be drawn:

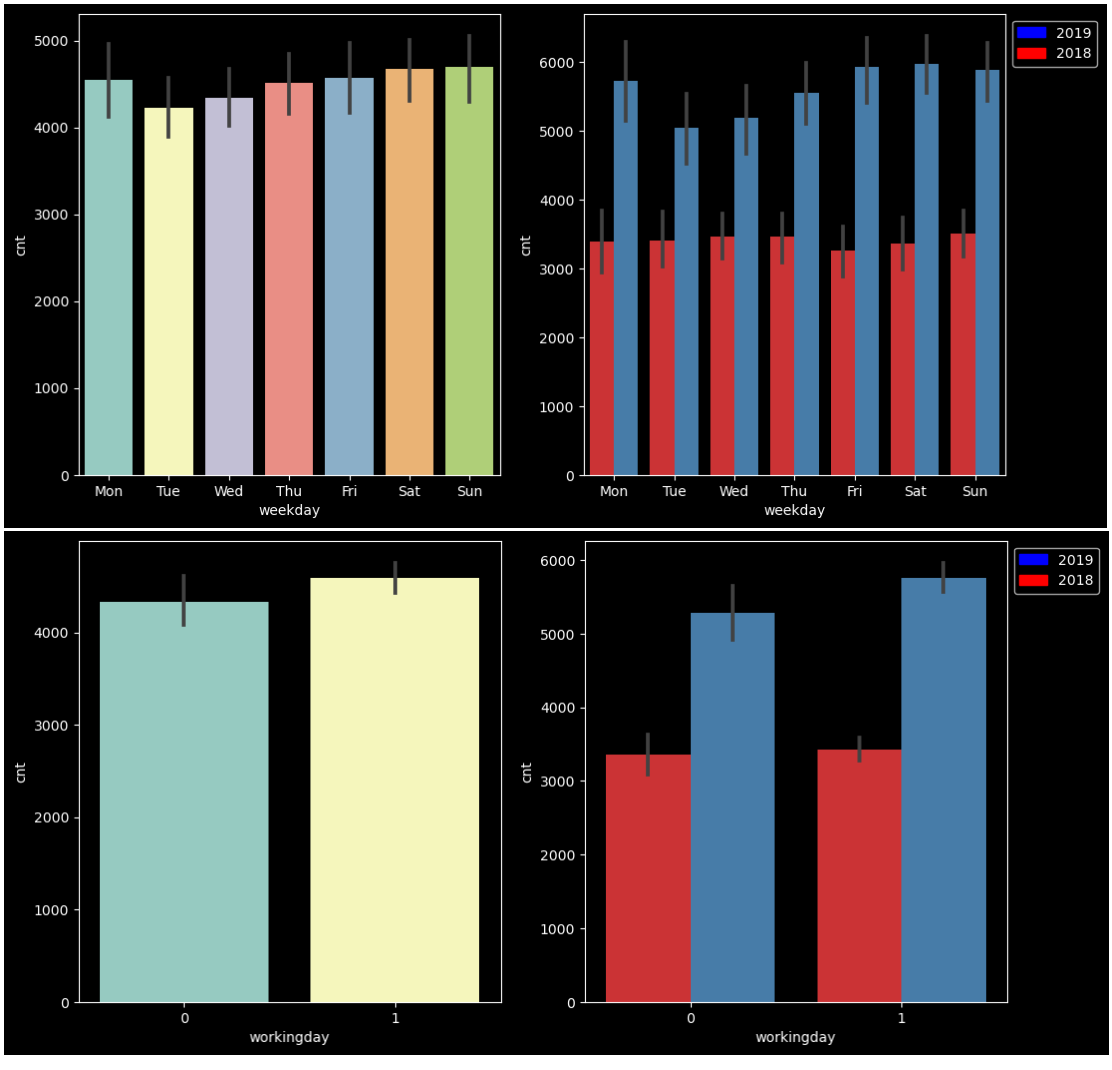
   - Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019
   - Clear weather attracted more bookings
   - Less number of bookings on holidays as people may want to spend time indoors with family or relaxing after a tiring work week
   - 2019 attracted more number of booking from the previous year, which shows good progress in terms of business




cnt
mnth
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2019
2018


cnt
season
Spring Summer Fall Winter
2019
2018


cnt
weathersit
LightRain Clear HeavyRain
2019
2018



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

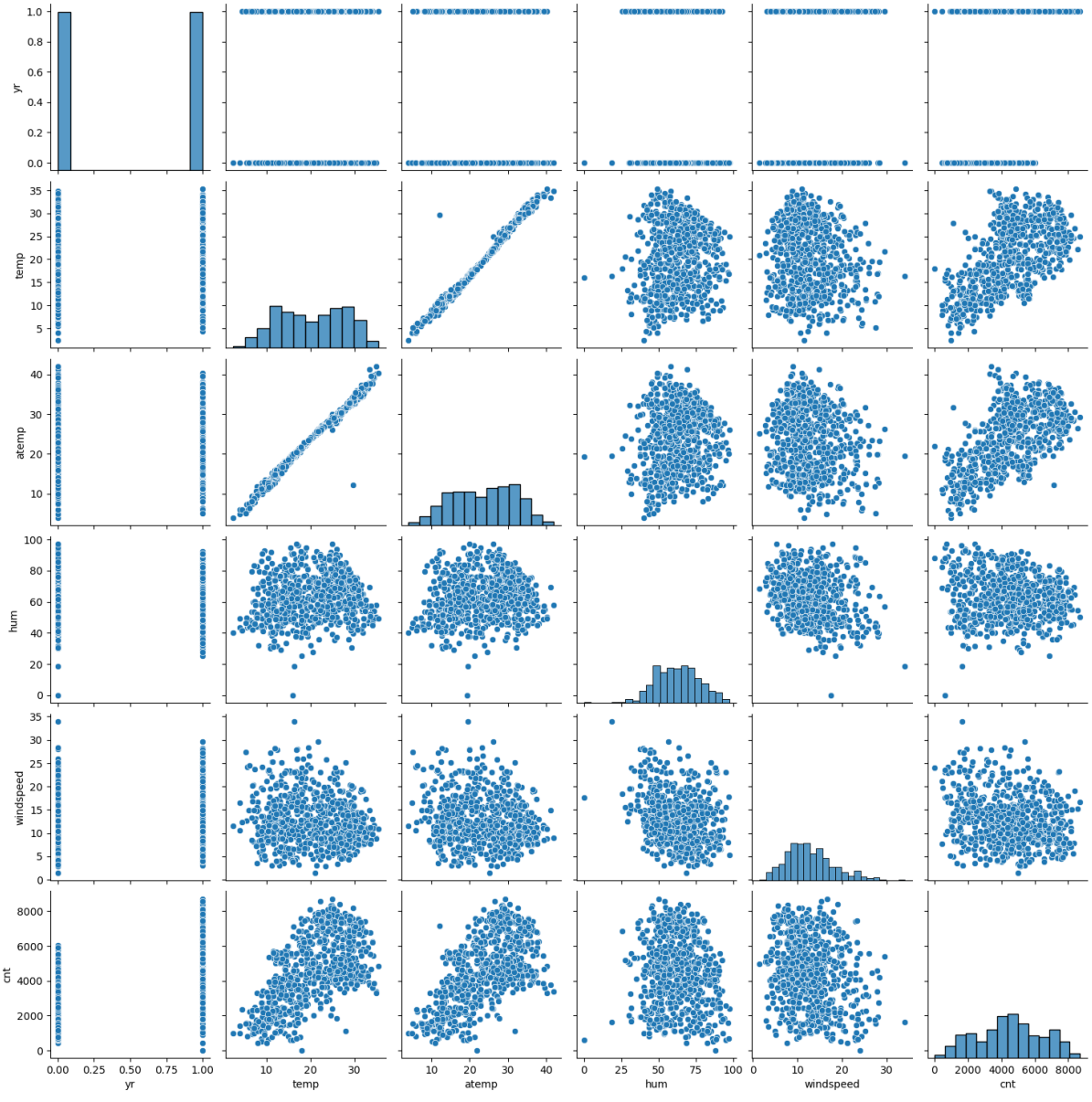
   Ans:

   While creating dummy variables, it is important to ensure that the encoding does not result in perfect multi-collinearity in the dataset. This can be avoided by using drop_first=True when creating dummy variables (one-hot encoding). Perfect multi-collinearity happens when one predictor variable in a regression model can be perfectly predicted by the other predictors. For Example: Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

   Ans:

   Based on the pair-plot generated for the numerical columns in the dataset it can be inferred that "temp" and "atemp" have the highest correlation with the target variable "cnt"



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

   Ans:

   I have validated the assumption of Linear Regression Model based on below 5 assumptions -

   - Normality of error terms: Error terms should be normally distributed
   - Multi-collinearity check: There should be insignificant multi-collinearity among variables.
   - Linear relationship validation: Linearity should be visible among variables
   - Homoscedasticity: There should be no visible pattern in residual values.
   - Independence of residuals: No auto-correlation

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

Ans:
Based on the final model: temp, holiday,months(Jun,Dec,jul,sep)

General Subjective Questions

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

Ans:
Linear regression is a machine learning algorithm which estimates how a model is following a linear relationship between one response variable (denoted by y) and one or more explanatory variables (denoted by X1, X2, X3…. Xn). The important components of Linear Regression are:

- Regression Coefficient (or β1): This component describes the change in the value of dependent variable corresponding to the unit change in the independent variable. So, for e.g. if X1 increases or decreases by one unit, then Y will increase or decrease by β1 units.
- Intercept (or β0): This component is a constant value which tells us at what point in the x-y coordinate graph, should the regression line start if it follows a linear regression. Since it is a constant value, hence it is not dependent on any change in independent variables.
- Error Terms or Residuals (ϵ): This component describes the difference between the actual and the predicted data point in the x-y coordinate graph.

The following general steps can be performed while performing Linear Analysis:

- Reading and understanding the data: Importing required libraries like pandas & numpy for data analysis and manipulation and seaborn & matplotlib for data visualization. Data Cleaning and Manipulation may also have to be done in order to standardize the data and make it fit for further exploration
- Visualizing the data (Exploratory Data Analysis): Scatter/Pairplots will mostly be used for numerical data in order to visualize and understand the data trend. Barplots/Boxplots will be used in order to interpret business/domain inferences for categorical data.
- Data Preparation: Creating dummy variables for categorical data. This step is performed for better representation of the categorical data during Model Building
- 4. Splitting the data into training and test sets: Splitting the data into two sections in order to train a subset of dataset to generate a trained (fitted) line that will very well generalize how new and unknown data (test set or new dataset) will be evaluated, and how the fitted line will be able to accurately estimate new or unknown datasets. Generally, the train-test split ratio is 70:30 or 80:20. Also, Re-Scaling of the data will have to be performed for Normalization

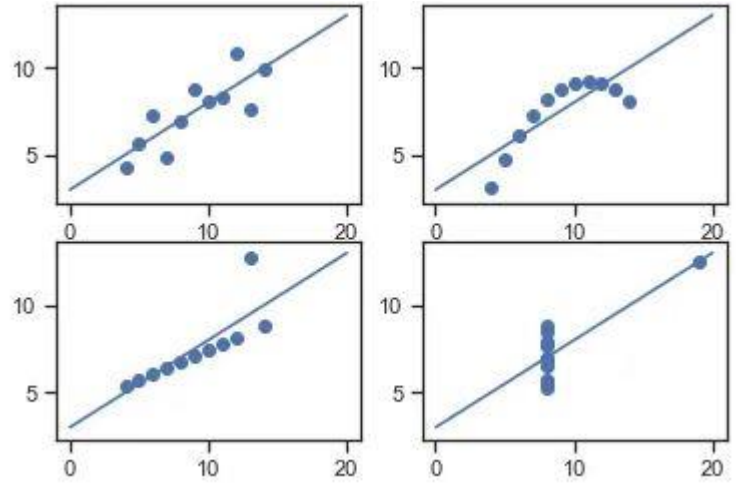
- Building a linear model:
    - Forward Selection: We start with null model and add variables one by one. These variables are selected on the basis of high correlation with the target variable.
    - b. Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p- values).
    - c. RFE or Recursive Feature Elimination is more like an automated version of feature selection technique where we select that we need “m” variables out of “n” variables and then machine provides a list of features with importance level given in terms of rankings.
- Residual analysis of the train data: This step is performed to analyze the error i.e. difference between predicted output and actual output. A good residual analysis will signify that the mean is centred around 0
- Making predictions using the final model and evaluation:
    - We will predict the test dataset by transforming it onto the trained dataset
    - Divide the test sets into X_test and y_test and calculate r2_score of test set. The train and test set should have similar r2_score. A difference of 2–3% between r2_score of train and test score is acceptable as per the standards.

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

Ans:

It is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties. It 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. Each graph plot shows the different behaviour irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
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



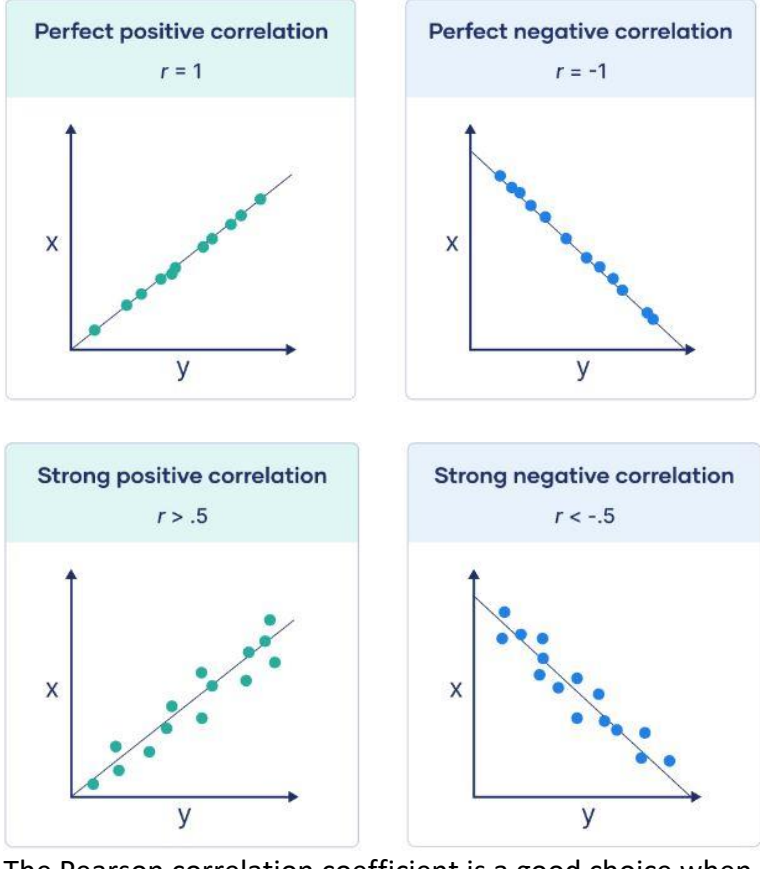
3. What is Pearson's R? (3 marks)

Ans:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 -> Data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 -> Data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 -> No linear association
0 < r < 0.5 -> Weak association
0.5 < r < 0.8 -> Moderate association
r > 0.8 -> Strong association

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses.
Specifically, we can test whether there is a significant relationship between two variables.



The Pearson correlation coefficient is a good choice when all of the following are true:

- Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
- The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

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

Ans:

Scaling or Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Some of the reasons for performing Feature Scaling are:

- Normalisation: Guarantees that all features are on a comparable scale and have comparable ranges. Larger scale features may dominate the learning process and have an excessive impact on the outcomes. One can avoid this problem and make sure that each feature contributes equally to the learning process by scaling the features.
- Algorithm performance improvement: When the features are scaled, several machine learning methods, including gradient descent-based algorithms, distance-based algorithms (such k-nearest neighbours), and support vector machines, perform better or converge more quickly.
- Preventing numerical instability: Numerical instability can be prevented by avoiding significant scale disparities between features. Examples include distance calculations or matrix operations, where having features with radically differing scales can result in numerical overflow or underflow problems.
- Scaling features makes ensuring that each characteristic is given the same consideration during the learning process. Without scaling, bigger scale features could dominate the learning, producing skewed outcomes. This bias is removed through scaling, which also guarantees that each feature contributes fairly to model predictions.

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)**

   Ans:

   A variance inflation factor (VIF) is a measure of the amount of multi-collinearity in regression analysis. Multi-collinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multi-collinearity.

   $$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

   From the above formula, it is obvious that VIF will have a value of infinity if the R^2 value is 1. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of the other variables. One approach to solve this infinite VIF issue is to review the independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model.

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

   Ans:

   The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. The first quantile is that of the variable we are testing the hypothesis for and the second one is the actual distribution that we are testing it against.

   The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

   Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, Dataset 1: the age variable has 200 records and Dataset 2: the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in the post-deployment scenarios to identify covariate shift/dataset shift/concept shift visually.

   It can be used to check the following scenarios:

   - if the two data sets come from a population with a common distribution
   - if the two data sets have a common location and scale
   - if the two data sets have similar distribution shapes
   - if the two data sets have similar tail behaviour