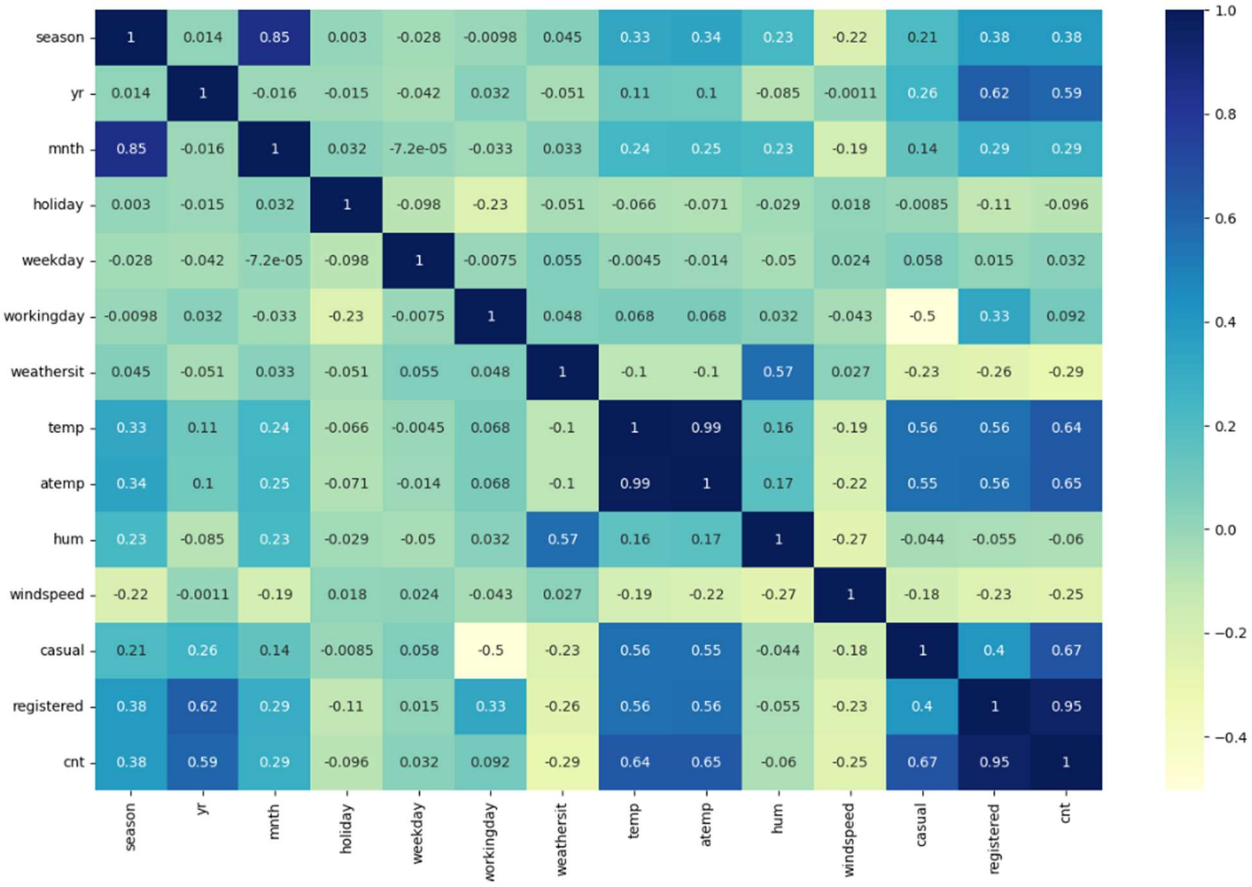
Assignment-based Subjective Questions

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



	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
season	1	0.014	0.85	0.003	-0.028	-0.0098	0.045	0.33	0.34	0.23	-0.22	0.21	0.38	0.38
yr	0.014	1	-0.016	-0.015	-0.042	0.032	-0.051	0.11	0.1	-0.085	-0.0011	0.26	0.62	0.59
mnth	0.85	-0.016	1	0.032	-7.2e-05	-0.033	0.033	0.24	0.25	0.23	-0.19	0.14	0.29	0.29
holiday	0.003	-0.015	0.032	1	-0.098	-0.23	-0.051	-0.066	-0.071	-0.029	0.018	-0.0085	-0.11	-0.096
weekday	-0.028	-0.042	-7.2e-05	-0.098	1	-0.0075	0.055	-0.0045	-0.014	-0.05	0.024	0.058	0.015	0.032
workingday	-0.0098	0.032	-0.033	-0.23	-0.0075	1	0.048	0.068	0.068	0.032	-0.043	-0.5	0.33	0.092
weathersit	0.045	-0.051	0.033	-0.051	0.055	0.048	1	-0.1	-0.1	0.57	0.027	-0.23	-0.26	-0.29
temp	0.33	0.11	0.24	-0.066	-0.0045	0.068	-0.1	1	0.99	0.16	-0.19	0.56	0.56	0.64
atemp	0.34	0.1	0.25	-0.071	-0.014	0.068	-0.1	0.99	1	0.17	-0.22	0.55	0.56	0.65
hum	0.23	-0.085	0.23	-0.029	-0.05	0.032	0.57	0.16	0.17	1	-0.27	-0.044	-0.055	-0.06
windspeed	-0.22	-0.0011	-0.19	0.018	0.024	-0.043	0.027	-0.19	-0.22	-0.27	1	-0.18	-0.23	-0.25
casual	0.21	0.26	0.14	-0.0085	0.058	-0.5	-0.23	0.56	0.55	-0.044	-0.18	1	0.4	0.67
registered	0.38	0.62	0.29	-0.11	0.015	0.33	-0.26	0.56	0.56	-0.055	-0.23	0.4	1	0.95
cnt	0.38	0.59	0.29	-0.096	0.032	0.092	-0.29	0.64	0.65	-0.06	-0.25	0.67	0.95	1

Based on the image above of the correlation coefficients, we can see as follows:

- Categorial variables “season” has a high correlation coefficient of 0.38 with target variable “cnt”. It means that the count of total rental bikes is increased according the season from spring, summer, autumn, and winter.
- Categorial variables “yr” has a high correlation coefficient of 0.58 with target variable “cnt”. It means that the count of total rental bikes in 2019 increased compared to 2018.
- Categorial variables “month” has a high correlation coefficient of 0.29 with target variable “cnt”.
- It means that the count of total rental bikes in 2019 increased compared to 2018.
- Categorial variables “weekday” has a low correlation coefficient of 0.032 with target variable “cnt”.
- Categorial variables “weathersit” has a negative correlation coefficient of -0.29 with target variable “cnt”. It means that the count of total rental bikes is decreased in the worst weather conditions such as “heavy rain + ice pallets+thunderstorm+mist” and is increased in the good weather conditions such as “clear, few clouds, partly cloudy, partly cloudy”.
- Categorial variables “holiday” has a negative correlation coefficient of -0.096 with target variable “cnt”.

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

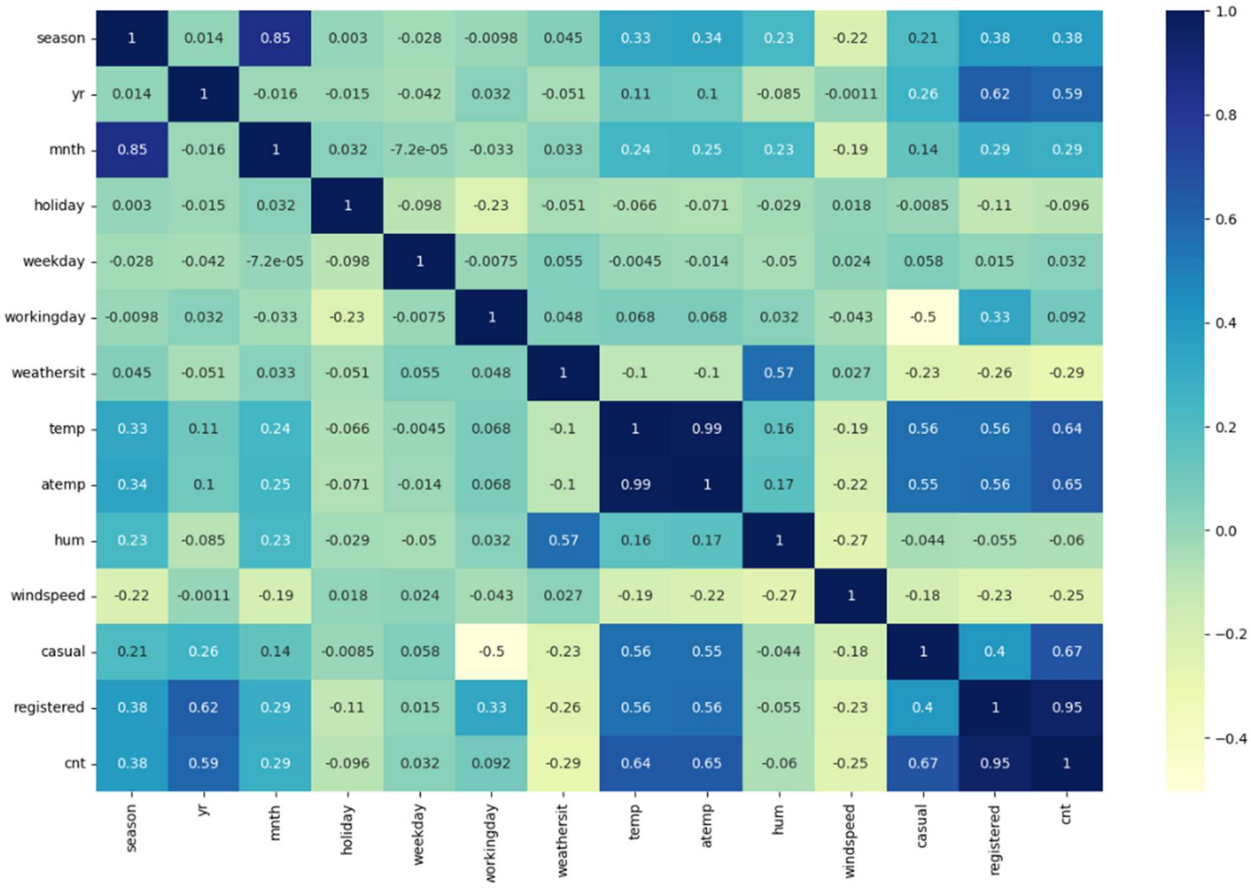
When creating dummy variables from categorical variables, it is often necessary to drop one of the categories to avoid the problem of multicollinearity in the dataset. Multicollinearity occurs when one or more predictor variables are highly correlated with each other. In the case of categorical variables, each category can be represented by a binary variable, and if we include all the binary variables, it will create perfect multicollinearity, as the sum of all the binary variables will always equal 1.

For example, let's consider a categorical variable "workingday" with two categories: "neither weeken nor holiday" is 1, "othewise" is 0. The sum of the two columns will always equal 1, as every observation will belong to exactly one of the two categories. This means that the model will not be able to distinguish between the two categories, leading to multicollinearity issues.

To avoid this issue, we can drop one of the categories when creating the dummy variables, which will allow the model to compare each category to the dropped category. By convention, the dropped category is usually the reference category, and the other categories are compared to it.

By setting **drop_first=True** when creating dummy variables using a function like **pandas.get_dummies()**, we can drop the first category, which will create k-1 binary columns for k categories, avoiding the multicollinearity issue. So it is important to use **drop_first=True** to ensure that the resulting dummy variables are properly constructed and the model can provide reliable predictions.

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



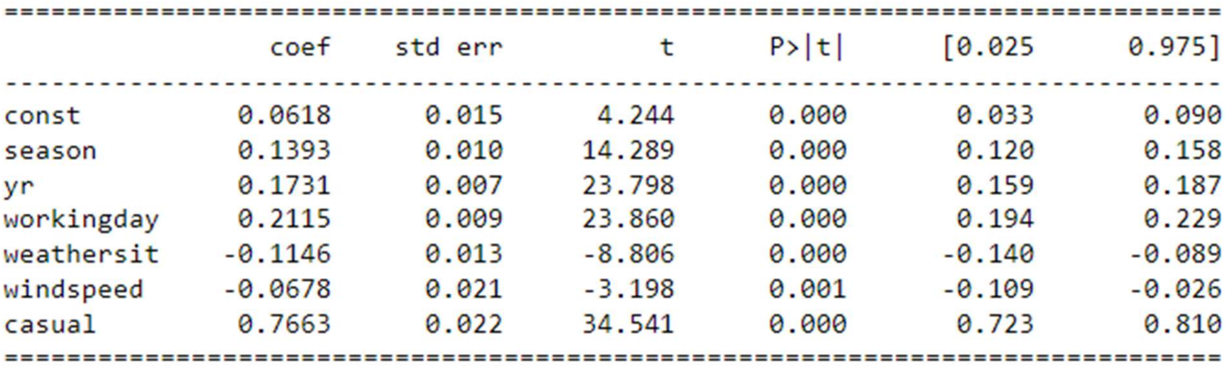
	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
season	1	0.014	0.85	0.003	-0.028	-0.0098	0.045	0.33	0.34	0.23	-0.22	0.21	0.38	0.38
yr	0.014	1	-0.016	-0.015	-0.042	0.032	-0.051	0.11	0.1	-0.085	-0.0011	0.26	0.62	0.59
mnth	0.85	-0.016	1	0.032	-7.2e-05	-0.033	0.033	0.24	0.25	0.23	-0.19	0.14	0.29	0.29
holiday	0.003	-0.015	0.032	1	-0.098	-0.23	-0.051	-0.066	-0.071	-0.029	0.018	-0.0085	-0.11	-0.096
weekday	-0.028	-0.042	-7.2e-05	-0.098	1	-0.0075	0.055	-0.0045	-0.014	-0.05	0.024	0.058	0.015	0.032
workingday	-0.0098	0.032	-0.033	-0.23	-0.0075	1	0.048	0.068	0.068	0.032	-0.043	-0.5	0.33	0.092
weathersit	0.045	-0.051	0.033	-0.051	0.055	0.048	1	-0.1	-0.1	0.57	0.027	-0.23	-0.26	-0.29
temp	0.33	0.11	0.24	-0.066	-0.0045	0.068	-0.1	1	0.99	0.16	-0.19	0.56	0.56	0.64
atemp	0.34	0.1	0.25	-0.071	-0.014	0.068	-0.1	0.99	1	0.17	-0.22	0.55	0.56	0.65
hum	0.23	-0.085	0.23	-0.029	-0.05	0.032	0.57	0.16	0.17	1	-0.27	-0.044	-0.055	-0.06
windspeed	-0.22	-0.0011	-0.19	0.018	0.024	-0.043	0.027	-0.19	-0.22	-0.27	1	-0.18	-0.23	-0.25
casual	0.21	0.26	0.14	-0.0085	0.058	-0.5	-0.23	0.56	0.55	-0.044	-0.18	1	0.4	0.67
registered	0.38	0.62	0.29	-0.11	0.015	0.33	-0.26	0.56	0.56	-0.055	-0.23	0.4	1	0.95
cnt	0.38	0.59	0.29	-0.096	0.032	0.092	-0.29	0.64	0.65	-0.06	-0.25	0.67	0.95	1

- Categorial variables “registered” has the highest correlation coefficient of 0.95 with target variable “cnt”.

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

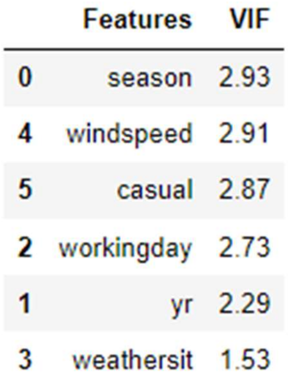
Techniques for validating the assumptions of Linear Regression:

- P-values: P-values can be used to assess the significance of the coefficients of the independent variables. A low p-value (usually less than 0.05) suggests that the independent variable is statistically significant and has a significant effect on the dependent variable. Example, the p-values is approx. zero as below:



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0618      0.015      4.244      0.000       0.033       0.090
season         0.1393      0.010     14.289      0.000       0.120       0.158
yr             0.1731      0.007     23.798      0.000       0.159       0.187
workingday     0.2115      0.009     23.860      0.000       0.194       0.229
weathersit    -0.1146      0.013     -8.806      0.000      -0.140      -0.089
windspeed     -0.0678      0.021     -3.198      0.001      -0.109      -0.026
casual         0.7663      0.022     34.541      0.000       0.723       0.810
==============================================================================
```

- Variance Inflation Factor (VIF): VIF can be used to check for multicollinearity among the independent variables. A VIF value of greater than 5 or 10 suggests a high degree of multicollinearity, which can affect the stability of the coefficients and the overall accuracy of the model. Example: our model has VIF values as follows:



	Features	VIF
0	season	2.93
4	windspeed	2.91
5	casual	2.87
2	workingday	2.73
1	yr	2.29
3	weathersit	1.53

All VIF values are less than 5

- F-statistic: F-statistic can be used to test the overall significance of the model. A high F-statistic (with a low p-value) suggests that the model is a good fit for the data. Example: our model has F-statistic = 630.3, it is high value with low p-values (approx. Zero).

- Adjusted R-squared: Adjusted R-squared is a measure of how well the model fits the data. It takes into account the number of variables in the model and penalizes for overfitting. A high adjusted R-squared value suggests that the model is a good fit for the data. Example: our model has Adjusted R-squared = 88.1%, it is high value that our model is good fit for the data.

- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): AIC and BIC are measures of the relative quality of a statistical model for a given set of data. They provide a trade-off between the goodness of fit of the model and the complexity of the model. A model with a lower AIC or BIC value is preferred over a model with a higher AIC or BIC value. Example: our model has AIC = -1156 and BIC = -1126.

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

Based on the coefficients of the final model:
cnt=0.1393×season + 0.1731×yr + 0.2115×workingday - 0.1146×weathersit-0.0678×windspeed + + 0.7663×casual + 0.0618

The top 3 features that contribute significantly towards explaining the demand of shared bikes are:

casual: This is the variable that has the highest positive coefficient in the model. This suggests that the higher the number of casual riders using the bikes, the higher the demand for shared bikes.

workingday: This variable has a positive coefficient, indicating that the demand for shared bikes is higher on working days compared to non-working days.

yr: This variable has a positive coefficient, indicating that the demand for shared bikes has increased over the years.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear regression is a machine learning algorithm used to model the relationship between a dependent variable and one or more independent variables. It is a statistical method used to determine the linear relationship between two or more variables. In this algorithm, a linear equation is constructed that best fits the data in order to predict the value of the dependent variable.

   The general equation for linear regression is:

   $y = c + mx$

   where y is the dependent variable, x is the independent variable, m is the slope of the line, and c is the intercept. The goal of linear regression is to find the values of m and c that minimize the sum of the squared residuals (the difference between the predicted value and the actual value).

   There are two types of linear regression: simple linear regression and multiple linear regression.

   Simple linear regression involves only one independent variable, while multiple linear regression involves two or more independent variables. The steps for both types of linear regression are similar.

   Here are the steps for performing linear regression:

   - **Step 1. Data collection:** Collect the data for the dependent and independent variables.

   - **Step 2. Reading and Understanding the Data**
     This involves cleaning the data by removing any missing or duplicate values, and read the dataset.

   - **Step 3. Visualising the Data:**
     - If there is some obvious multicollinearity going on, this is the first place to catch it.
     - Here's where we'll also identify if some predictors directly have a strong association with the outcome variable.

   - **Step 4. Data Preparation**
     - We can see that our dataset has many columns with values as 'Yes' or 'No'. But in order to fit a regression line, we would need numerical values and not string. Hence, we need to convert them to 1 and 0, where 1 is a 'Yes' and 0 is a 'No'.
     - Some variables have several levels. We need to convert these levels into integer as well. For this, we will use something called dummy variables.

   - **Step 5. Splitting the data into Training and Testing**

The data is split into two sets: training and testing data. The training data is used to train the model, while the testing data is used to evaluate the performance of the model.

It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation.
So, it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.
There are two common ways of rescaling:

- **Min-Max scaling**
- **Standardization (mean-0, sigma-1)**

- **Step 6. Building a linear model:** Choose the appropriate type of linear regression (simple or multiple), and select the features (independent variables) to include in the model. The model is trained using the training data by minimizing the sum of the squared residuals.

- **Step 7. Model evaluation:** Evaluate the performance of the model using the testing data. The most common evaluation metric is the mean squared error (MSE), which measures the average squared difference between the predicted value and the actual value.

When p-values is less than 5, the variables aren't really significant (in the presence of other variables). We could simply drop the variable with the highest, non-significant p value.

A better way would be to supplement this with the VIF information.
Checking VIF: Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. The formula for calculating VIF is:

$$VIFi=1/(1-Ri2).$$

Dropping the variable and updating the model
As we can see from the summary and the VIF dataframe, variables which have VIF >5 and p-values of their coefficients >0.05, such variables are insignificant and should be dropped.

- **Step 8. Residual Analysis of the train data**

So, now to check if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression), we plot the histogram of the error terms and see what it looks like.

- **Step 9: Making Predictions Using the Final Model**

Now that we have fitted the model and checked the normality of error terms, it's time to go ahead and make predictions using the final.

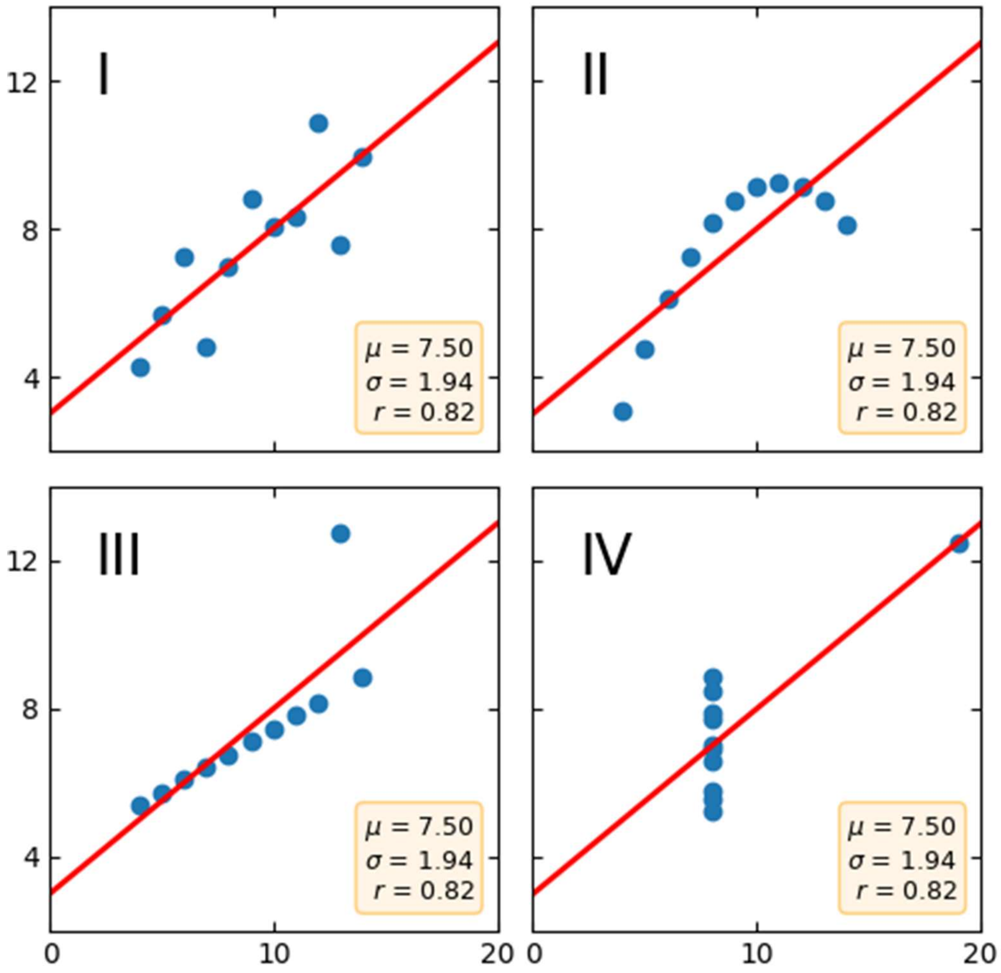
- **Step 10: Model Evaluation**

Let's now plot the graph for actual versus predicted values.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties. Example:



3. What is Pearson's R?

Pearson's R is a statistical measure that describes the strength and direction of the linear relationship between two variables. It is also known as the Pearson correlation coefficient or Pearson's product-moment correlation coefficient.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It can take on values between -1 and 1, where -1 indicates a perfect negative correlation (when one variable increases, the other decreases), 0 indicates no correlation, and 1 indicates a perfect positive correlation (when one variable increases, the other increases as well).

Pearson's R assumes that the relationship between the two variables is linear, meaning that the change in one variable is directly proportional to the change in the other variable. It also assumes that the variables are normally distributed.

Pearson's R is widely used in many fields, including psychology, social sciences, biology, and economics. It can be used to analyze data, identify trends, and make predictions based on the relationship between two variables.
Below is a formula for calculating the Pearson correlation coefficient (r):

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

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

   Scaling is a process of transforming data to a common scale or range to facilitate better analysis and comparison of variables. In machine learning and data analysis, scaling is performed to bring all the variables to the same level of magnitude.

   If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation.

   There are two common ways of scaling data:

   Normalization: Normalization rescales the data to be between 0 and 1. The formula for normalization is:

   x_normalized = (x - min(x)) / (max(x) - min(x))

   where x is the original value, and x_normalized is the scaled value.

   Standardization: Standardization rescales the data to have a mean of 0 and standard deviation of 1. The formula for standardization is:

   x_standardized = (x - mean(x)) / std(x)

   where x is the original value, mean(x) is the mean of the variable, std(x) is the standard deviation of the variable, and x_standardized is the scaled value.

   The difference between normalized scaling and standardized scaling is in the scale and range of the transformed data. Normalization scales the data to a range between 0 and 1, while standardization scales the data to have a mean of 0 and standard deviation of 1.
   Normalization is suitable when the range of the data is known, and we want to scale it to a common range, while standardization is suitable when the range of the data is unknown, and we want to scale it to have a standard deviation of 1. Both normalization and standardization are used depending on the nature of the data and the requirements of the analysis.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   The VIF (Variance Inflation Factor) is a measure of multicollinearity in linear regression models. It calculates the ratio of the variance of the estimated coefficient for a particular independent variable in the presence of other independent variables to the variance of the estimated coefficient for that same variable in the absence of the other independent variables.

When the value of VIF is infinite, it indicates that there is perfect multicollinearity among the independent variables. This means that one or more independent variables in the model can be perfectly predicted by a linear combination of the other independent variables. In other words, there is a linear relationship between two or more independent variables that can be expressed as an exact mathematical equation.

Perfect multicollinearity can happen for various reasons, such as including the same variable multiple times in the model or using a variable that is a linear combination of other variables already included in the model. When perfect multicollinearity occurs, the regression coefficients and their standard errors cannot be estimated reliably, and the regression model becomes unstable.

To avoid this, it is important to check for multicollinearity in the initial stages of building a regression model, and to remove or combine any highly correlated independent variables before fitting the model.

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

A Q-Q plot (quantile-quantile plot) is a graphical method used to compare the distribution of a sample data set with a theoretical distribution. In other words, it is a plot of the quantiles of two distributions against each other.

The Q-Q plot is useful in linear regression because it allows us to assess whether the residuals (the differences between the predicted and actual values) follow a normal distribution. If the residuals are normally distributed, it suggests that the linear regression model is appropriate for the data and that the assumptions of the model are met.

To create a Q-Q plot for the residuals of a linear regression model, we plot the quantiles of the residuals against the quantiles of a normal distribution. If the residuals follow a normal distribution, the points on the plot will form a straight line. If the residuals do not follow a normal distribution, the points on the plot will deviate from a straight line.

In addition to assessing the normality of the residuals, the Q-Q plot can also be used to check for other distributions, such as the t-distribution, depending on the assumptions of the regression model.

the Q-Q plot is a valuable tool for assessing the validity of a linear regression model and ensuring that its assumptions are met.