Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Season , Weather Situation ,holiday ,month, Working day, and weekday were the categorical variables in the dataset. A boxplot was used to visualise these. These variables influenced our dependent variable in the following ways:

1 Season: The boxplot revealed that the spring season had the lowest value of cnt, while the fall season had the highest value of cnt. Summer and winter had cnt values that were in the middle.
2. Weather Situation: When there is heavy rain/snow, there are no users, indicating that the weather is extremely unfavourable. The highest count was observed when the weather forecast was 'Clear, partly cloudy.

3. Holiday: Rentals were found to be lower during the holidays.

4. Month: September had the most rentals, while December had the fewest. This observation is comparable to the one made in weathers it. The weather in December is typically cold and snowy.

5. Weekday: Weekends saw a significant increase in book hiring compared to weekdays.

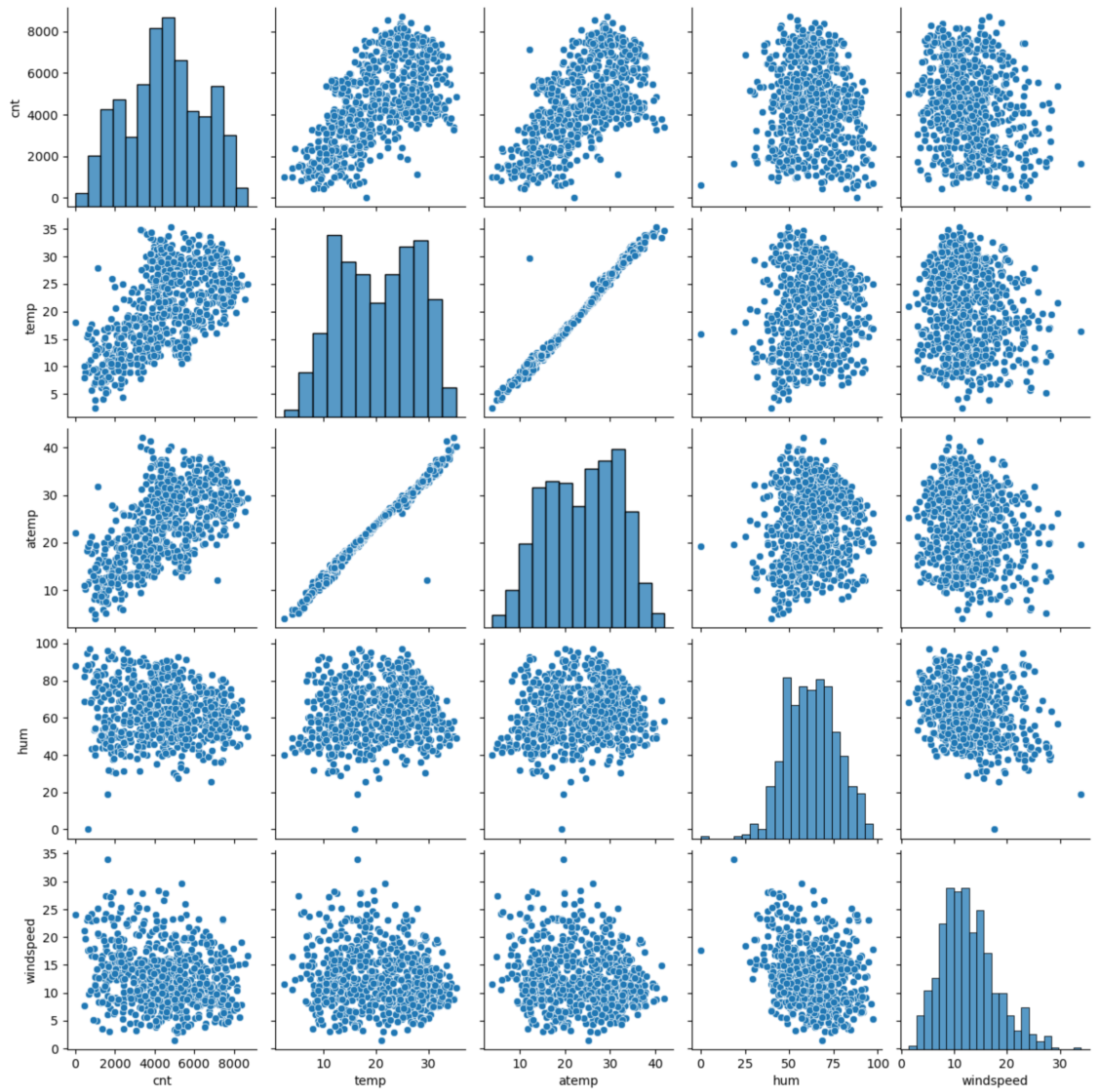
6. Working day: It had little effect on the dependent variable.

Q2. Why is it important to use drop _first=True during dummy variable creation?

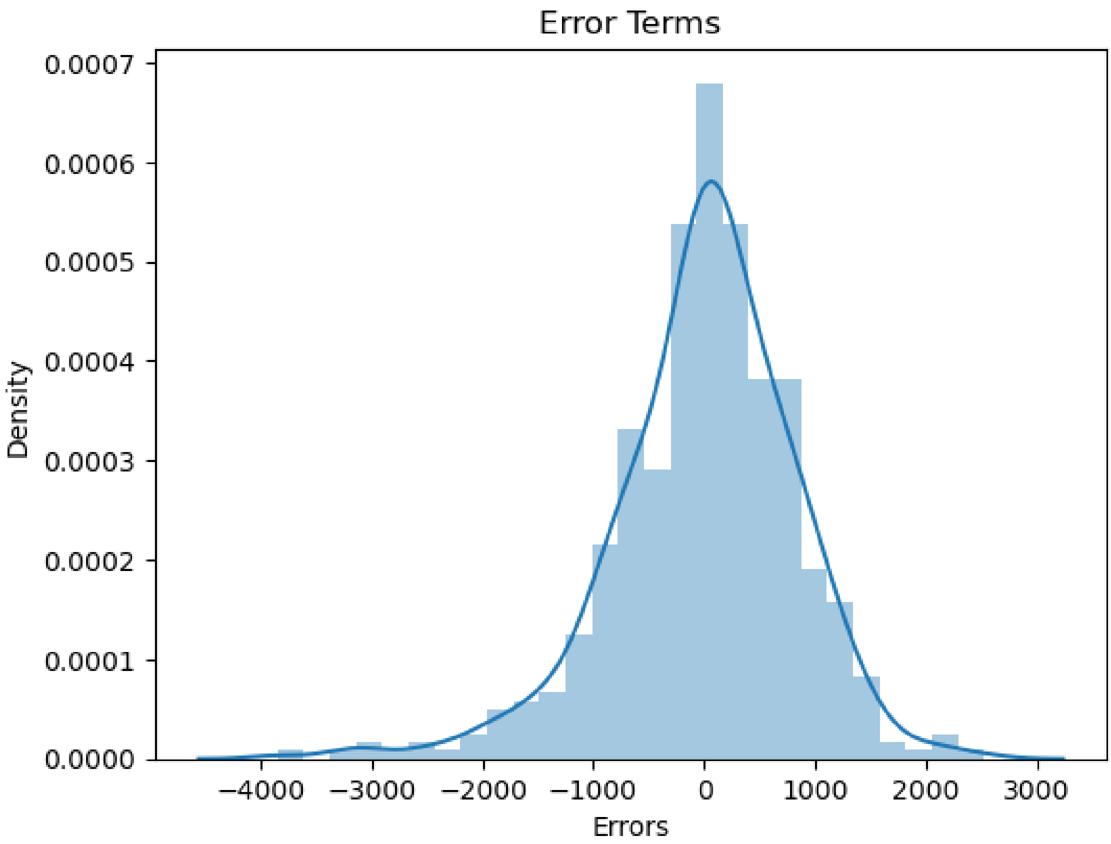
Answer:- Your dummy variables will be correlated if you don't remove the first column (redundant). This may have a negative impact on some models, and the effect is amplified when the cardinality is low.
Iterative models, for example, may have difficulty convergent, and lists of variable importance may be distorted. Another argument is that having all dummy variables results in multicollinearity between them. We lose one column to keep everything under control.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:-



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: - The top three predictor variables that influence bike booking, according to our final Model, are:

A. Temperature(temp) : With a coefficient of 0.5173, a unit increase in the temp variable increases the number of bike rentals by 0.5173 units.

B. Weather Situation 3 (weathers it 3): With a coefficient of '-0.2828.' a unit increase in the Weathersit3 variable reduces the number of bike hires by 0.2828 units as compared to Weathers it_ 1.

Where weathersit_3 = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain

C. Scattered clouds

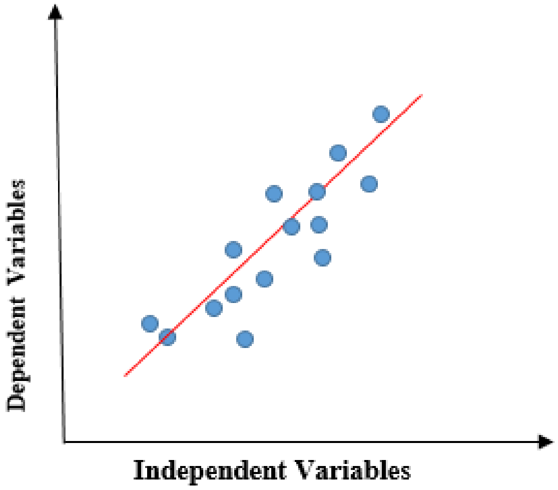
Year(yr) : With a coefficient of 0.2324, a unit increase in the yr variable increases the number of bike rentals by 0.2324.

General Subjective questions:-

Q1. Explain Linear regression algorithms in details

Answer: -

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. *If there is a single input variable (x), such linear regression is called* ***simple linear regression****. And if there is more than one input variable, such linear regression is called* ***multiple linear regression*****.** The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

Need of a Linear regression

As mentioned above, Linear regression estimates the relationship between a dependent variable and an independent variable. Let's understand this with an easy example:

Let's say we want to estimate the salary of an employee based on year of experience. You have the recent company data, which indicates that the relationship between experience and salary. Here year of experience is an independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.

1. Q2. Explain the Anscombe’s quartet in detail.

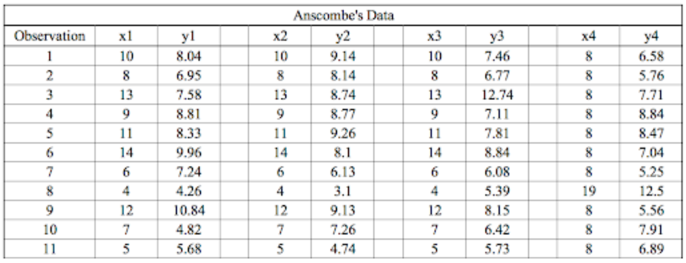
Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe’s quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

What Is the Purpose of Anscombe’s Quartet in Data Visualization?

Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

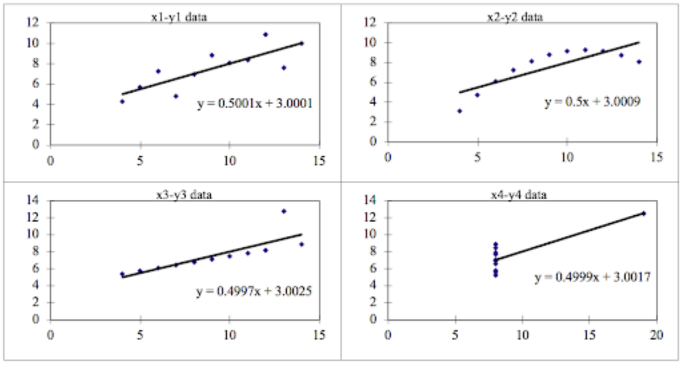
We can define these four plots as follows:



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn’t interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

ANSCOMBE’S QUARTET FOUR DATASETS

2 **Data Set 1:** fits the linear regression model pretty well.

3 **Data Set 2:** cannot fit the linear regression model because the data is non-linear.

4 **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
5 **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

What is Pearson's r?

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.

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables	
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction	

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson’s r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)

- The correlation coefficient

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.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (r) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

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.

When to use the Pearson correlation coefficient

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

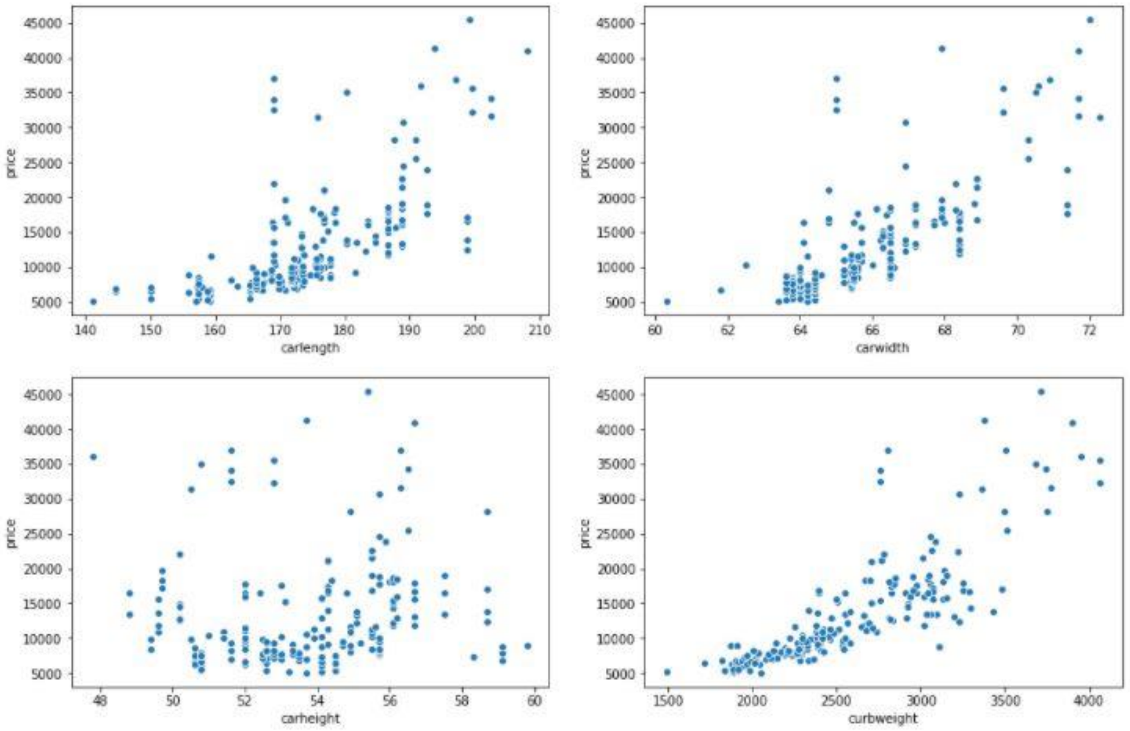
- Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
- The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables 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.

1. Q3. What is Pearson's R?

Answer: -

When we try to infer something from what we have heard or read, the first step we do is relate a few of the parameters or scenes, etc. with each other and then proceed. Correlation means to find out the association between the two variables and Correlation coefficients are used to find out how strong the is relationship between the two variables. The most popular correlation coefficient is Pearson’s Correlation Coefficient. It is very commonly used in linear regression.

Consider the example of car price detection where we have to detect the price considering all the variables that affect the price of the car such as carlength, curbweight, carheight, carwidth, fueltype, carbody, horsepower, etc.



We can see in the above scatterplot, as the carlength, curbweight, carwidth increases price of the car also increases. So, we can say that there is a positive correlation between the above three variables with car price. Here, we also see that there is no correlation between the carheight and car price.

Assumptions for a Pearson Correlation:

1. Data should be derived from random or least representative samples, draw a meaningful statistical inference.

2. Both variables should be continuous and normally distributed.

3. There should be Homoscedasticity, which means the variance around the line of best fit should be similar.

4. Extreme outliers influence the Pearson Correlation Coefficient. You need to consider outliers that are unusual only on one variable, called as 'univariate variable' or for both of the variables known as 'multivariate outliers'. 2 variables are measured independently from each other pairs. e.g. If we plot age vs amount then, we can certainly, see that there is a correlation between the age of a person and loan the amount is given to that person, as age increases the loan amount given to the person decreases and vice versa. But if we plot the loan amount vs age, it is not possible to draw any conclusion from it. It would violate the assumption.

1. Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: -

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.

Why to use scaling:-

Why use Feature Scaling?
In machine learning, feature scaling is employed for a number of purposes:
6 Scaling guarantees that all features are on a comparable scale and have comparable ranges. This process is known as feature normalisation. This is significant because the magnitude of the features has an impact on many machine learning techniques. Larger scale features may dominate the learning process and have an excessive impact on the outcomes. You can avoid this problem and make sure that each feature contributes equally to the learning process by scaling the features.
7 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. The algorithm's performance can be enhanced by scaling the features, which can hasten the convergence of the algorithm to the ideal outcome.

8 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. Stable computations are ensured and these issues are mitigated by scaling the features.

9 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.

Data processing includes Normalization and standardization as essential components. We frequently come across several variables with varied original scales while processing data. Using these scales, variables with wide data ranges can be given more weight.

The two feature scaling techniques—Normalization vs. Standardization—will be covered in this article. Both phrases are occasionally used synonymously. But they relate to different things.

What is Data Normalization?

One of the most popular methods for preparing data is Normalization, which enables us to alter the values of numerical columns in the dataset to a standard scale.

Normalization is the method used to arrange the data in a database. It is a scaling method that reduces duplication in which the numbers are scaled and moved between 0 and 1.

When there are no outliers since it can't handle them, normalization is employed to remove the undesirable characteristics from the dataset.

One technique to process data to produce easily comparable findings within and across several data sets is the normalization procedure. Anyone reading data can benefit from it, but those using machine learning and significant amounts of data may find it most regularly helpful. Understanding the normalization formula will help you decide if it is the best way to handle your data set.

What is Data Standardization?

Standardization, often referred to as z-score Normalization, occasionally is a method for rescaling the values that meet the characteristics of the standard normal distribution while being similar to normalizing.

Standardization is crucial because it enables reliable data transmission across various systems. It would be easier for computers to exchange data and communicate with one another with standardization. Additionally, standardization makes it simpler to process, analyze, and store data in a database. Businesses can use their data to make better judgments with this method. Companies can more readily compare and evaluate data when standardized, allowing them to gain insights into how to run their businesses better.

When the data is distributed Gaussianly, standardization can be helpful. But it's okay for this to be the case. Standardization also lacks a bounding range, in contrast to normalizing. Therefore, normalization will have no effect on any outliers you may have in your data.

Normalization vs Standardization

Normalization	Standardization

This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
When features are on various scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful

Normalization vs Standardization Key Differences

Normalization is a suitable choice when your data's distribution does not match a Gaussian distribution. A practical transformation approach that helps your model perform and be more accurate is normalization. Normalization of a machine learning model is helpful when you are unsure about the precise feature distribution. To put it another way, the data's feature distribution does not have a Gaussian distribution. Outliers in your data will be impacted by normalization because it needs a wide range to function correctly.

When you are entirely aware of the feature distribution of your data, or, to put it another way, when your data has a Gaussian distribution, standardization in the machine learning model is useful. This need not necessarily be the case, though. In contrast to Normalization, Standardization does not always have a bounding range; therefore, any outliers in your data won't be impacted by it.

Scales for normalization fall between [0,1] and [-1,1]. Standardization has no range restrictions. When the algorithms don't make any assumptions about the distribution of the data, Normalization is taken into account. When algorithms create predictions about the data distribution, standardization is applied.

1. Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

〖VIF〗_1=1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$$X_2=C+\alpha_1 X_1+\alpha_3 X_3+\cdots$$

$$\text{VIF}_2=1/(1-R_2^2)$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

What to do if VIF is large?

If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:

10 One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.

11 A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.

12 The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.

13 The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.

14 Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

In conclusion, when you are building a multiple regression model, always check your VIF values for your independent variables and determine if you need to take any corrective action before building the model.

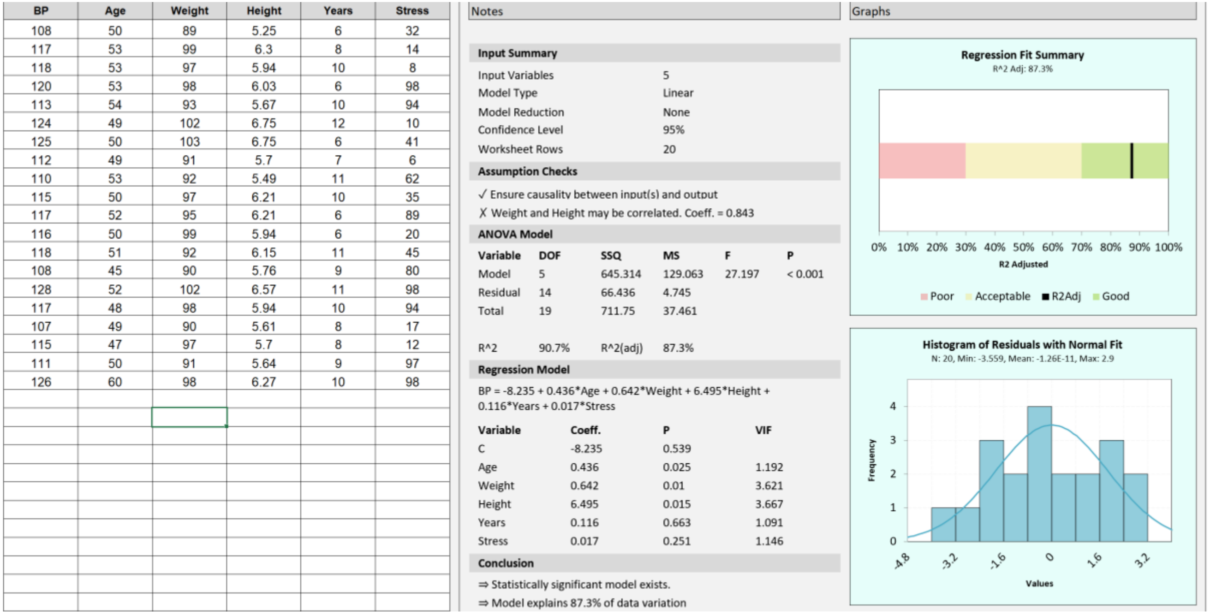
Example

The Blood Pressure (BP) measurements for several individuals was collected along with a few independent variables as shown in the table below. The explanatory variables are age of the person (years), weight of the person (kg), height of the person (feet), duration the person is suffering from hypertension (years), and stress level (score on a scale of 0-100). Develop a multiple regression model between the input(s) and output.

BP	Age	Weight	Height	Years	Stress
108	50	89	5.25	6	32
117	53	99	6.3	8	14
118	53	97	5.94	10	8
120	53	98	6.03	6	98
113	54	93	5.67	10	94
124	49	102	6.75	12	10
125	50	103	6.75	6	41
112	49	91	5.7	7	6
110	53	92	5.49	11	62
115	50	97	6.21	10	35
117	52	95	6.21	6	89
116	50	99	5.94	6	20
118	51	92	6.15	11	45
108	45	90	5.76	9	80
128	52	102	6.57	11	98
117	48	98	5.94	10	94
107	49	90	5.61	8	17
115	47	97	5.7	8	12
111	50	91	5.64	9	97
126	60	98	6.27	10	98

Solution:-

Model 1 Let's build a model with all the factors in the model. The regression model between the output and the inputs is shown below.



BP	Age	Weight	Height	Years	Stress
108	50	89	5.25	6	32
117	53	99	6.3	8	14
118	53	97	5.94	10	8
120	53	98	6.03	6	98
113	54	93	5.67	10	94
124	49	102	6.75	12	10
125	50	103	6.75	6	41
112	49	91	5.7	7	6
110	53	92	5.49	11	62
115	50	97	6.21	10	35
117	52	95	6.21	6	89
116	50	99	5.94	6	20
118	51	92	6.15	11	45
108	45	90	5.76	9	80
128	52	102	6.57	11	98
117	48	98	5.94	10	94
107	49	90	5.61	8	17
115	47	97	5.7	8	12
111	50	91	5.64	9	97
126	60	98	6.27	10	98

Notes

Input Summary

Input Variables	5
Model Type	Linear
Model Reduction	None
Confidence Level	95%
Worksheet Rows	20

Assumption Checks

✓ Ensure causality between input(s) and output

✗ Weight and Height may be correlated. Coeff. = 0.843

ANOVA Model

Variable	DOF	SSQ	MS	F	P
Model	5	645.314	129.063	27.197	< 0.001
Residual	14	66.436	4.745		
Total	19	711.75	37.461		

R^2 90.7% R^2(adj) 87.3%

Regression Model

BP = -8.235 + 0.436*Age + 0.642*Weight + 6.495*Height + 0.116*Years + 0.017*Stress

Variable	Coeff.	P	VIF
C	-8.235	0.539	
Age	0.436	0.025	1.192
Weight	0.642	0.01	3.621
Height	6.495	0.015	3.667
Years	0.116	0.663	1.091
Stress	0.017	0.251	1.146

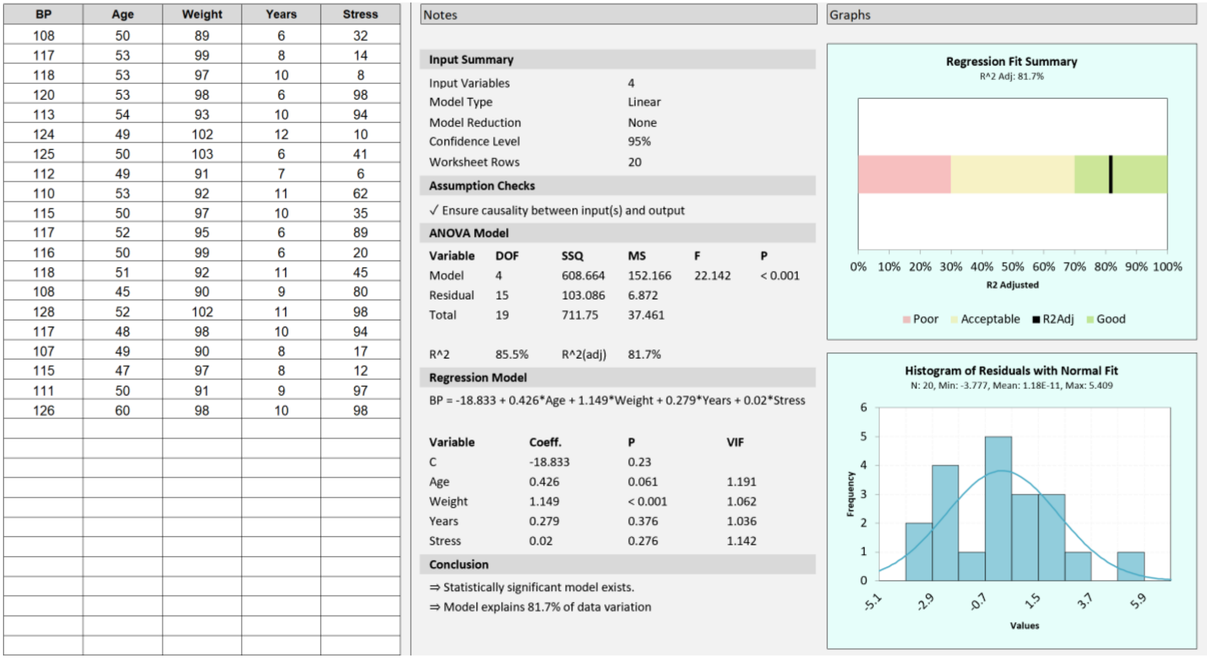
Conclusion

⇒ Statistically significant model exists.

⇒ Model explains 87.3% of data variation

Looking at the P values, we would also conclude that the number of years of hypertension and stress are not correlated to BP. Let’s look at the correlation plots between the input variables – it looks like the height and weight are correlated but the other factors are not. From the analysis results, we can see that the VIF for height and weight are higher than the other factors. It can also be seen that the two factors were correlated with a correlation coefficient of 0.843. Since, some of the VIF factors are large, we are not sure at this point if we can trust the model coefficients. Hence, we need to build a better model where the VIF factors are not that large. Since, height and weight seem to be correlated, we need to pick only one of these terms in our model. If we feel that it is easier to include weight in our analysis, we would drop the height factor and rebuild the model.

Model 2 If we were to drop the height from our model, we would get the following model.



BP	Age	Weight	Years	Stress
108	50	89	6	32
117	53	99	8	14
118	53	97	10	8
120	53	98	6	98
113	54	93	10	94
124	49	102	12	10
125	50	103	6	41
112	49	91	7	6
110	53	92	11	62
115	50	97	10	35
117	52	95	6	89
116	50	99	6	20
118	51	92	11	45
108	45	90	9	80
128	52	102	11	98
117	48	98	10	94
107	49	90	8	17
115	47	97	8	12
111	50	91	9	97
126	60	98	10	98

Notes

Input Summary

Input Variables	4
Model Type	Linear
Model Reduction	None
Confidence Level	95%
Worksheet Rows	20

Assumption Checks

✓ Ensure causality between input(s) and output

ANOVA Model

Variable	DOF	SSQ	MS	F	P
Model	4	608.664	152.166	22.142	< 0.001
Residual	15	103.086	6.872		
Total	19	711.75	37.461		

R^2 85.5% R^2(adj) 81.7%

Regression Model

BP = -18.833 + 0.426*Age + 1.149*Weight + 0.279*Years + 0.02*Stress

Variable	Coeff.	P	VIF
C	-18.833	0.23	
Age	0.426	0.061	1.191
Weight	1.149	< 0.001	1.062
Years	0.279	0.376	1.036
Stress	0.02	0.276	1.142

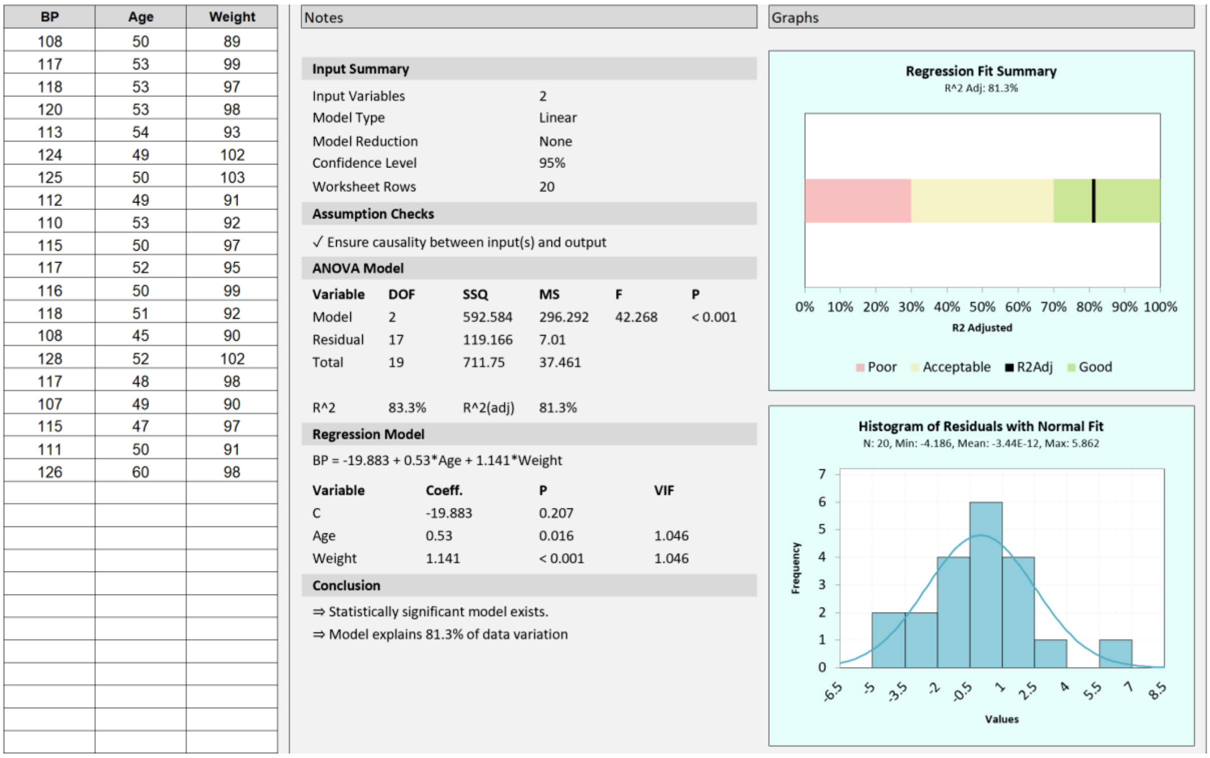
Conclusion

⇒ Statistically significant model exists.

⇒ Model explains 81.7% of data variation

Looking at the VIF values, our model does not exhibit any multicollinearity. Now, we can trust the model coefficients. Since the P values for number of years and stress are not-significant, let’s drop these terms to build the final model.

Model 3 The final model with only the significant factors is:



BP	Age	Weight
108	50	89
117	53	99
118	53	97
120	53	98
113	54	93
124	49	102
125	50	103
112	49	91
110	53	92
115	50	97
117	52	95
116	50	99
118	51	92
108	45	90
128	52	102
117	48	98
107	49	90
115	47	97
111	50	91
126	60	98

Notes

Input Summary

Input Variables	2
Model Type	Linear
Model Reduction	None
Confidence Level	95%
Worksheet Rows	20

Assumption Checks

✓ Ensure causality between input(s) and output

ANOVA Model

Variable	DOF	SSQ	MS	F	P
Model	2	592.584	296.292	42.268	< 0.001
Residual	17	119.166	7.01		
Total	19	711.75	37.461		

R^2 83.3% R^2(adj) 81.3%

Regression Model

BP = -19.883 + 0.53*Age + 1.141*Weight

Variable	Coeff.	P	VIF
C	-19.883	0.207	
Age	0.53	0.016	1.046
Weight	1.141	< 0.001	1.046

Conclusion

⇒ Statistically significant model exists.

⇒ Model explains 81.3% of data variation

This model shows that the BP increases 0.53 units with every year and 1.141 units for every kg increase in weight. From model 1 to model 3, the R^2 adjusted value drops from 87% to 81%. SInce all the terms are now statistically significant, we can use this model to make predictions and/or optimization.