Group Project on Modelling using Regression

Instructor: Mr. Chandan Verma

Aryan Tandon (Team Leader) Harshal Niwas Rathod
Manish Kumar Prajapati Danish Ahmad

Model for Calculating Weight of an Individual using Regression and Data Analysis

Contents

S.No.	Title	Page no.
1.	Model for Calculating Weight of an Individual using Regression and Data Analysis – Abstract (Software Used)	4
2.	Objective (Introduction, Goals of Multiple Regression)	5
3.	Methodology (Data Flow Diagram, Plan of Research)	6
4.	Data Analysis (Graphical Representation of Data, Accuracy)	8
5.	Code	10
6.	Conclusion	11
7.	Future Work	12
8.	References	13

List of Figures

Number	Caption
1.1	A regression graph
3.1	Data Flow Diagram
3.2	Data table showing the head and tail values
4.1	Importing data
4.2	Applying Regression
4.3	The line represents the regression equation

Model for Calculating Weight of an Individual using Regression and Data Analysis

1. ABSTRACT

In this paper, we reduce the dimension by principal component analysis and choose the best regression equation using multiple regression analysis. Finally, compared with the real value, we analyse the fitting accuracy of the regression equation we proposed. Weight prediction has been a popular problem in ergonomics study. There are a lot of weight-height algorithms that help us find the weight of a person based on his/her height, but they aren't very accurate. We have taken into consideration a lot more variables.

Software used

Jupyter Notebook, MS Excel

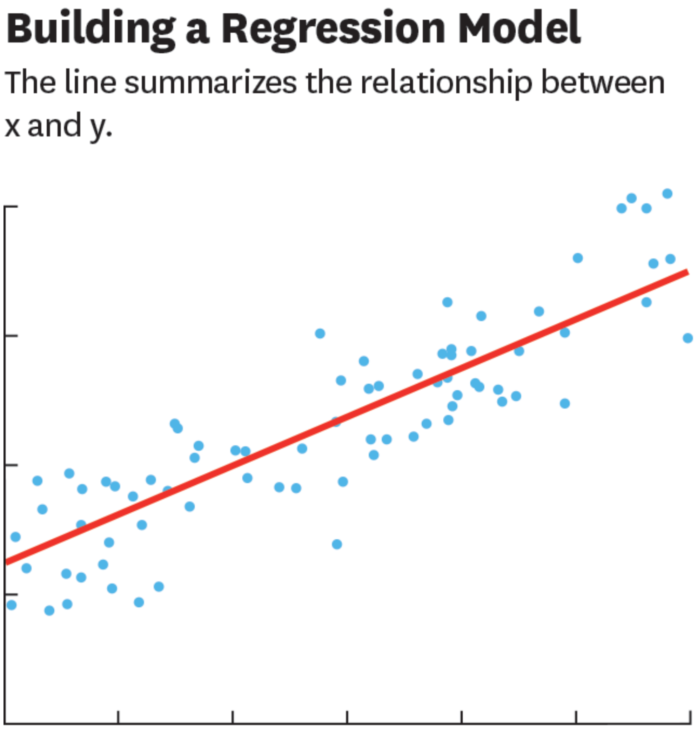


Fig 1.1: A regression graph

2. OBJECTIVE

The objective of this project is to create a machine learning model based on regression analysis which help us to determine the weight of a person. There exists a model which is based upon finding the weight of a person based on their height alone. But that model isn't very accurate, and we need to include more variables into a model so as to correctly predict the weight of an individual. We will be using multiple regression analysis in this machine learning model, and check how accurate it is.

Introduction

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

Regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Thus, it provides a good basis for estimating the cost and duration. If y is a dependent variable and x_1, ..., x_k are independent variables then the multiple regression model provides a prediction of y from the xi of the form:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$,
where $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ is the deterministic portion of the model and ε is the random error. We further assume that for any given values of the xi the random error ε is normally and independently distributed.

Regression analysis is widely used in social, economic, scientific and technological fields of data analysis, the establishment of empirical formula for regular forecasting, such as weather forecasting, earthquake prediction, stock market analysis and so on.

Goals of Multiple Regression

Often, examples in statistics courses describe iterative techniques to find the model that best describes relationships or best predicts a response variable. This data set can also demonstrate how multivariate regression models can be used to confirm theories. The most common goals of multiple regression are to:

1) Describe: Develop a model to describe the relationship between the explanatory variables and the response variable.

2) Predict: Use a set of sample data to make predictions. A regression model can be used to predict response values from explanatory variables within the range of our sample data.

3) Confirm: Theories are often developed about individual variables, such as confirming which variables, or combination of variables, need to be included in the model. Regression can then be used to determine if the contribution of each explanatory variable in a model captures much of the variability in the response variable.

3. METHODOLOGY

The data of 100 people was taken into account, the attributes being their height (cm), weight (kg), arm reach from the back (mm), hand length (mm), foot length (mm), lower extremity length (mm), neck circumference (mm), bust (mm), waist circumference (mm), head width (mm).

We have applied multiple regression analysis through the use of Python programming language on Jupyter. We're building a regression analysis model which will help us to predict the weight of an individual through the given data.

Data Flow Diagram

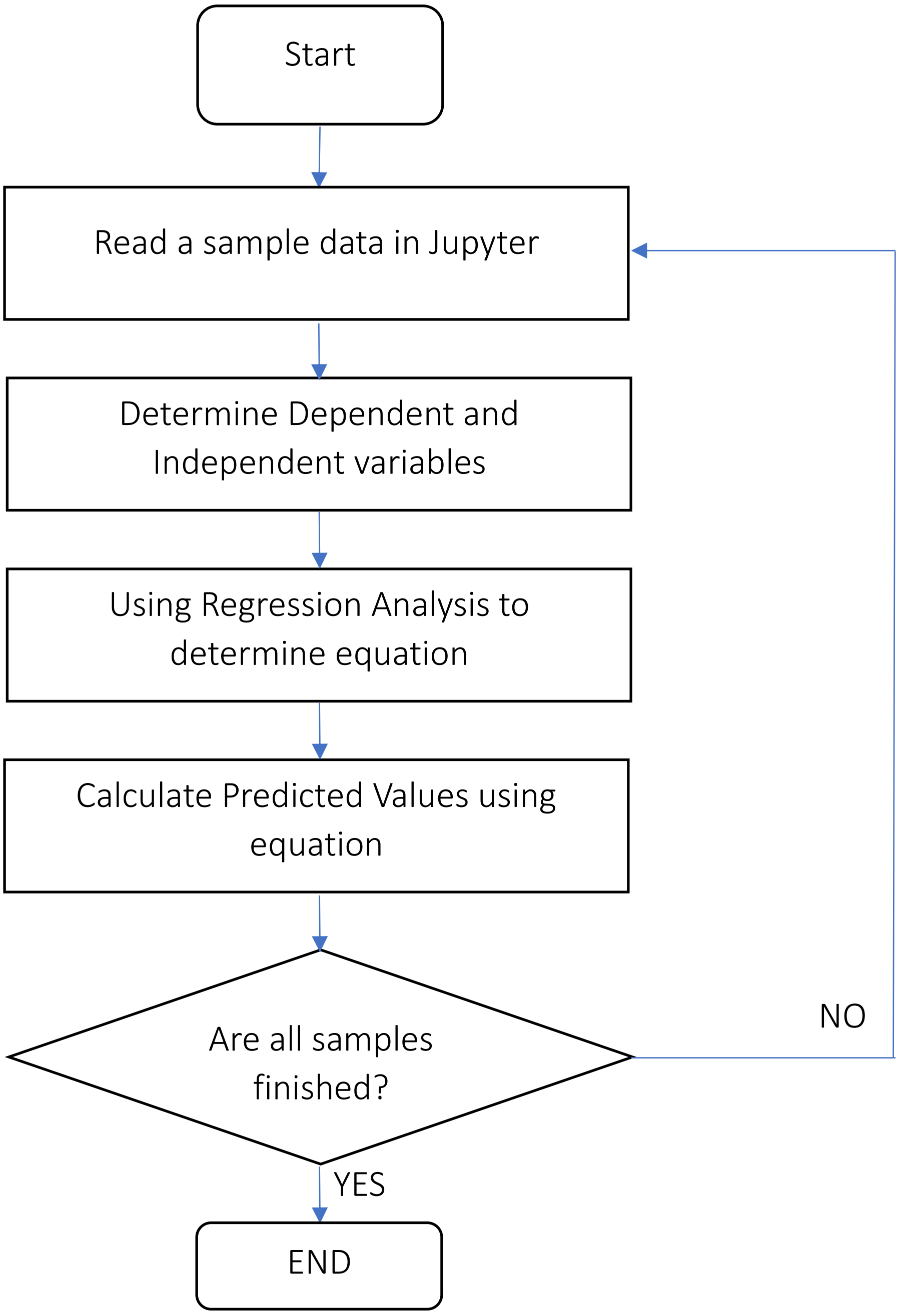
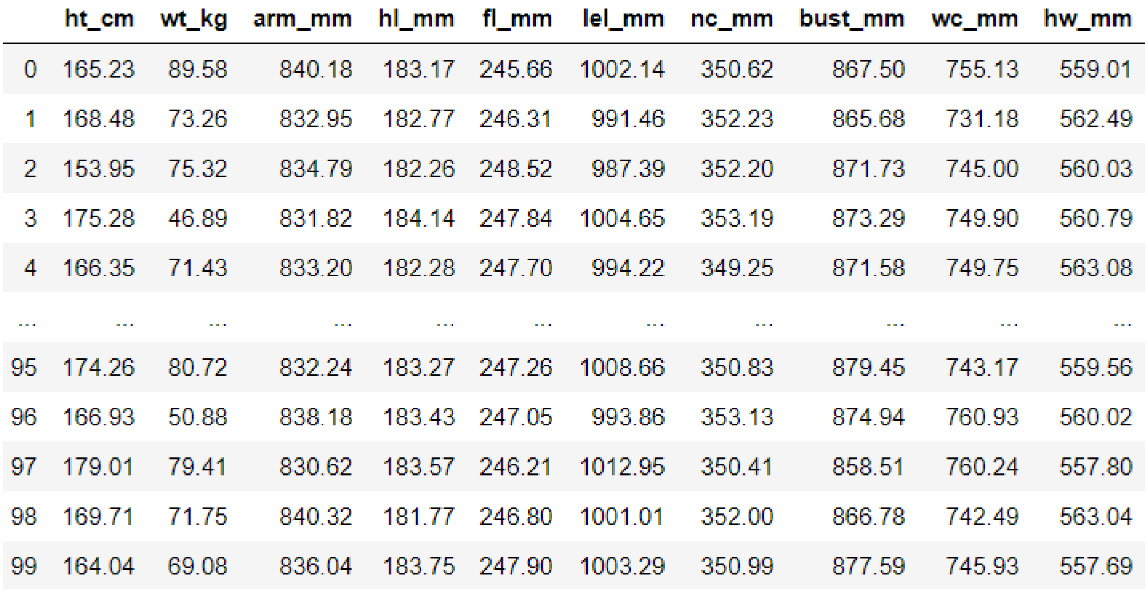

Start
Read a sample data in Jupyter
Determine Dependent and Independent variables
Using Regression Analysis to determine equation
Calculate Predicted Values using equation
Are all samples finished?
NO
YES
END

Fig 3.1: Data Flow Diagram

Plan of Research

We take weight of all the participants as the dependent variable, and the rest of the variables as the independent variables, to try and find the regression equation of weight. This will help us to predict the weight of an individual on the characteristics mentioned above.

Some of the data is given below.



	ht_cm	wt_kg	arm_mm	hl_mm	fl_mm	lel_mm	nc_mm	bust_mm	wc_mm	hw_mm
0	165.23	89.58	840.18	183.17	245.66	1002.14	350.62	867.50	755.13	559.01
1	168.48	73.26	832.95	182.77	246.31	991.46	352.23	865.68	731.18	562.49
2	153.95	75.32	834.79	182.26	248.52	987.39	352.20	871.73	745.00	560.03
3	175.28	46.89	831.82	184.14	247.84	1004.65	353.19	873.29	749.90	560.79
4	166.35	71.43	833.20	182.28	247.70	994.22	349.25	871.58	749.75	563.08
...	...	...	...	...	...	...	...	...	...	...
95	174.26	80.72	832.24	183.27	247.26	1008.66	350.83	879.45	743.17	559.56
96	166.93	50.88	838.18	183.43	247.05	993.86	353.13	874.94	760.93	560.02
97	179.01	79.41	830.62	183.57	246.21	1012.95	350.41	858.51	760.24	557.80
98	169.71	71.75	840.32	181.77	246.80	1001.01	352.00	866.78	742.49	563.04
99	164.04	69.08	836.04	183.75	247.90	1003.29	350.99	877.59	745.93	557.69

Fig 3.2: Data table showing the head and tail values

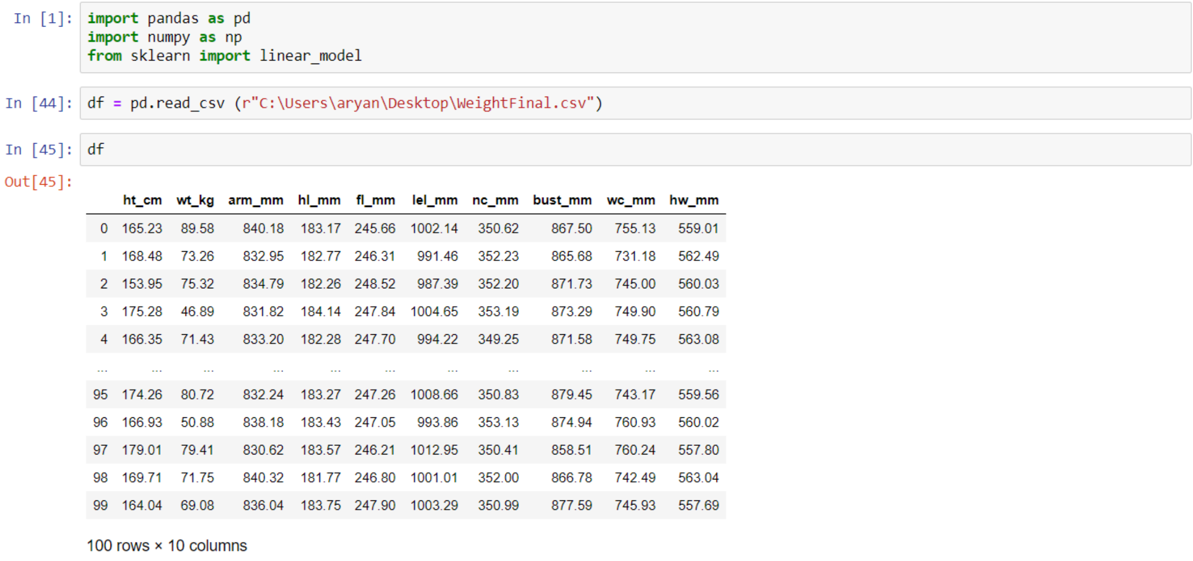
Input Dataset: https://drive.google.com/file/d/1NYWa6e-yZJ20Fd7TPe3ffYSLAXWi19NR/view?usp=sharing

4. DATA ANALYSIS

Import libraries: The advantage of working with Python is that we have access to many libraries that allow us to rapidly read data, plot the data, and perform a linear regression.

I like to import all the necessary libraries on top of the notebook to keep everything organized.

The .csv data file is imported through the functionality offered to us by Jupyter. After reading the data, we know now that it is clean.



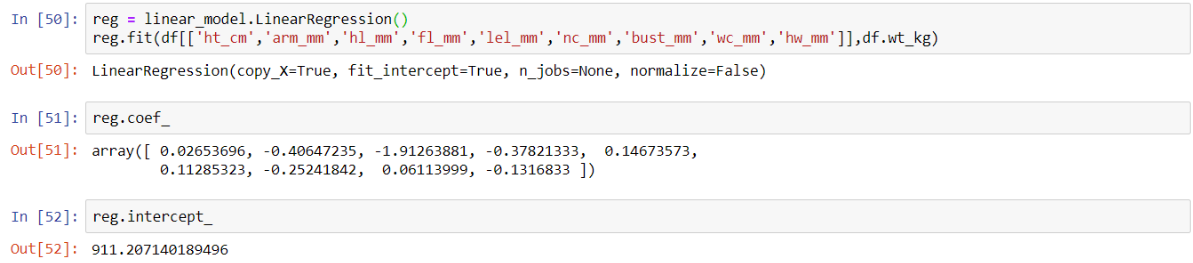
```
In [1]: import pandas as pd
        import numpy as np
        from sklearn import linear_model

In [44]: df = pd.read_csv (r"C:\Users\aryan\Desktop\WeightFinal.csv")

In [45]: df

Out[45]:
```

	ht_cm	wt_kg	arm_mm	hl_mm	fl_mm	lel_mm	nc_mm	bust_mm	wc_mm	hw_mm
0	165.23	89.58	840.18	183.17	245.66	1002.14	350.62	867.50	755.13	559.01
1	168.48	73.26	832.95	182.77	246.31	991.46	352.23	865.68	731.18	562.49
2	153.95	75.32	834.79	182.26	248.52	987.39	352.20	871.73	745.00	560.03
3	175.28	46.89	831.82	184.14	247.84	1004.65	353.19	873.29	749.90	560.79
4	166.35	71.43	833.20	182.28	247.70	994.22	349.25	871.58	749.75	563.08
...	...	...	...	...	...	...	...	...	...	...
95	174.26	80.72	832.24	183.27	247.26	1008.66	350.83	879.45	743.17	559.56
96	166.93	50.88	838.18	183.43	247.05	993.86	353.13	874.94	760.93	560.02
97	179.01	79.41	830.62	183.57	246.21	1012.95	350.41	858.51	760.24	557.80
98	169.71	71.75	840.32	181.77	246.80	1001.01	352.00	866.78	742.49	563.04
99	164.04	69.08	836.04	183.75	247.90	1003.29	350.99	877.59	745.93	557.69

100 rows × 10 columns

Fig. 4.1: Importing data

After this, we apply regression analysis to this data, considering weight as the dependent variable.



```
In [50]: reg = linear_model.LinearRegression()
         reg.fit(df[['ht_cm','arm_mm','hl_mm','fl_mm','lel_mm','nc_mm','bust_mm','wc_mm','hw_mm']],df.wt_kg)

Out[50]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

In [51]: reg.coef_

Out[51]: array([ 0.02653696, -0.40647235, -1.91263881, -0.37821333,  0.14673573,
                 0.11285323, -0.25241842,  0.06113999, -0.1316833 ])

In [52]: reg.intercept_

Out[52]: 911.207140189496
```

Fig. 4.2: Applying Regression

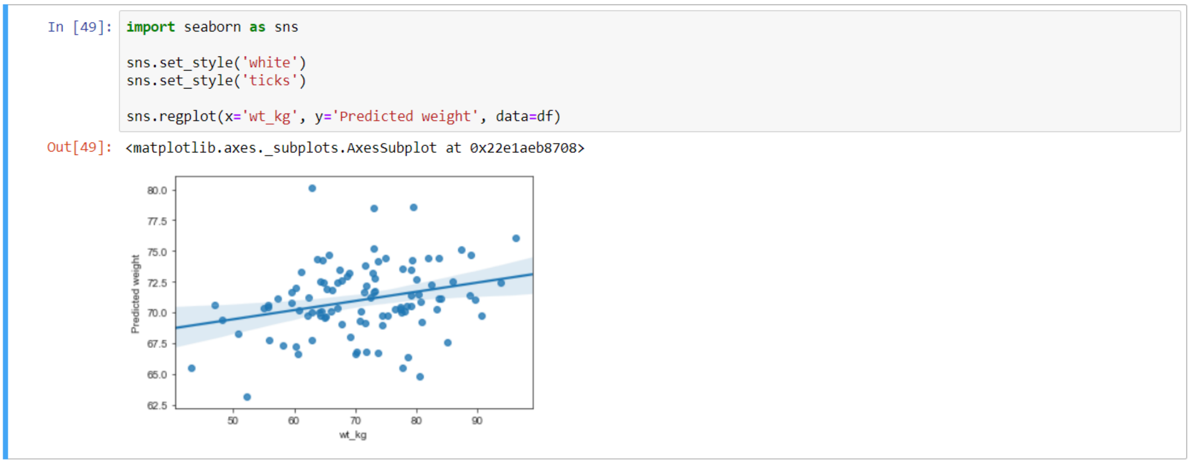
The intercept here is the error associated with the equation. The coefficients off the independent variables are as given above in their respective order.

Therefore, the regression equation is equal to,
y = 0.02653696 *(ht_mm) - 0.40647235 *(arm_mm) - 1.91263881 *(hl_mm) - 0.37821333 *(fl_mm) + 0.14673573 *(lel_mm) + 0.11285323 *(nc_mm) - 0.25241842 *(bust_mm) + 0.06113999 *(wc_mm) - 0.1316833 *(hw_mm) + 911.207140189496

Hence, we can make predictions about the weight of an individual with the help of this equation.

Graphical Representation of Data

Now we have the predicted values through the formula, we will find the scattering plot between the predicted values and the actual values of the weights of the individuals through the functionality given to us in the seaborn module.



```
In [49]: import seaborn as sns

         sns.set_style('white')
         sns.set_style('ticks')

         sns.regplot(x='wt_kg', y='Predicted weight', data=df)

Out[49]: <matplotlib.axes._subplots.AxesSubplot at 0x22e1aeb8708>
```

Fig. 4.3: The line represents the regression equation

Accuracy

We predicted the values of the weights of all individuals by the equation. Then we found out the % of error (%e) by dividing the difference (diff) between predicted value (PV) and actual value (AV) of the weights and multiplying the value by 100. Thus, the % of accuracy (%a) was calculated by subtracting 100 by % of error of all individuals.

%e = (diff/PV)*100
%a = 100 - %e

Thus, we then take the average of all the individual percentages of accuracies. That value is considered as the accuracy. The accuracy of our model is 87.65%. This model is more accurate than any other linear regression model that uses one variable, be it height or hand length.

5. CODE

```
import pandas as pd
import numpy as np
from sklearn import linear_model

df = pd.read_csv (r"C:\Users\aryan\Desktop\WeightFinal.csv")

df

reg = linear_model.LinearRegression()
reg.fit(df[['ht_cm','arm_mm','hl_mm','fl_mm','lel_mm','nc_mm','bust_mm','wc_mm','hw_mm']],df.wt_kg)

reg.coef_

reg.intercept_

reg.predict([[179.01,830.62,183.57,246.21,1012.95,350.41,858.51,760.24,557.80]])

import seaborn as sns
sns.set_style('white')
sns.set_style('ticks')
sns.regplot(x='wt_kg', y='Predicted weight', data=df)
```

6. CONCLUSION

From our analysis utilizing multiple regression technique, we have determined that an acceptable model is applicable to the data. This model has taken into consideration a number of factors, such as height, weight, arm reach from the back, hand length, foot length, lower extremity length, neck circumference, bust, waist circumference and head width. It is said that the height – weight model of calculating weight that has been in use for some time has an accuracy of about 51%. We have applied multiple regression analysis on Jupyter notebook and have also found out the accuracy of the model which is 87.65%. This makes it a much better model to predict weight.

We’ve also concluded with the help of the coefficients of regression of all variables, that hand length has a more significant effect on the weight of an individual than any other variable.

7. FUTURE WORK

This project involves implementation of data analytics and machine learning. This project work can be used as reference to learn implementation machine learning from very basic.

In future the idea can be extended by making more advanced graphical user interface with the help of newer libraries like shiny in R. An interactive page can be made, i.e. if the value of a attribute is changed on the scale the values corresponding to its graph (ggplot or histogram) will also change. We can also draw much focused conclusions by combining results we obtained.

8. REFERENCES

Draper, N.R. and Smith, H., (1981), "Applied Regression Analysis", Wiley. [Google Scholar]

Neter, J., Wasserman, W., and Kutner, M. (1985), Applied Linear Statistical Models, Homewood, Illinois: Irwin. [Google Scholar]

The R Project for Statistical Computing Retrieved from https://www.r-project.org/

Multiple Linear Regression Analysis Retrieved from http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis.

Data obtained by Project on Multiple Regression Analysis by Yin Jiang, Chen Yan Wu, Kelly Yarusso, Daobin Ye, Kan Zhang, Tianchi Zhang, Yoel Zuman

Mr. Chandan Verma
Head of Training, Computer Science Domain
Eisystems Services