DATA ANALYSIS PORTFOLIO

An aspiring ‘Data Scientist’ with two years of successful experience in Infosys as ‘Systems Engineer’. I am specialized in Machine Learning technologies and regularly learn and implement new algorithms to my different projects. A python coder, who loves to find optimal solutions for different scenarios and implements the same. I have done analysis of 15+ datasets and predicted results with good accuracy. Last but not the least, I am a Badminton player and a big foodie!

Professional Background

Education –

Bachelors of Technology (Information Technology) [Aug 2015 - Jul 2019]

Netaji Subhas University of Technology New Delhi, DELHI,India . (**GPA: 8.23/10)**

Work Experience –

Systems Engineer (Infosys) [Dec 2019 – Present]

Responsibilities -

- Automated end to end testing in SAP PT1 environment using Java, saving engineers over 10 hours each week in manual testing.
- PDF Extraction Tool using OpenCV and PyTesseract from text and scanned pdf.
- Content mining using OCR and built regex for extracting specific details from documents which saves customer's time by 30%.
- Spell corrector tool to detect wrong spellings in scanned images and correct it for better understanding of context.
- Experienced in working with Agile methodologies.

DATA ANALYSIS PORTFOLIO

Table of Contents

DATA ANALYSIS PORTFOLIO

Udemy Project Description

This is the business problem which required step by step analysis of data and find the optimal solution of the problem.

● **Business problem**: Udemy wants to increase the revenue for their company. For that I have to identify all possible ways by which revenue can be increased. I have data of all online courses given by Udemy. I have to find most popular courses on Udemy so that they can increase the next quarterly earnings.

● **Purpose of this report**: I am trying to find most popular courses on Udemy because, if we will focus on these courses, then we can increase the price or can launch similar courses which will give us more revenue.

● **Actions**: Now I have clear business problem which needs to be solved. I started with data cleaning. I removed all duplicates and null values. I also replaced incorrect values with correct ones. Now I have correct and accurate data, next thing I did was data visualization. For that I used both excel and Tableau. I tried to derive relationships among data for find insights and patterns in data. I drew pie charts to see percentage of each course type. Then I drew bar graph to see what types of courses are enrolled by majority of students.

● **Result**: I got multiple outcomes from analysis:-

- There are more paid courses than free courses but majority of student are opting for free courses.
- Web development is most popular among all subjects. It covers 67.75% of all subjects.
- Out of 20 most popular courses, 17 are web development courses.
- Free courses are more popular than paid courses.
- Graphic design expert courses got best average rating among all courses.
- Almost all courses in top 20 are beginner friendly courses.

DATA ANALYSIS PORTFOLIO

The Problem

What is the business problem?

Udemy wants to increase the revenue for their company. For that I have to identify all possible ways by which revenue can be increased. I have data of all online courses given by Udemy. I have to find most popular courses on Udemy so that they can increase the next quarterly earnings.

How long do you have to work on this project?

I have to work for at least two weeks to identify this complete scenario and to find patterns in the data for deriving final conclusions.

What data should be collected to understand this problem?

For this problem, I will need data about enrolled learners like which course they enrolled and duration of enrolled courses. I need categories of courses, difficulty levels of each course, it's free or paid.

How should it be presented?

It should be represented in pictorial form. All increasing and decreasing patterns should be represented by graphs. Amount of share each course category has should be depicted by pie chart.

What questions would you ask to better understand the business problem?

I would ask following questions for better understanding of problems:-

- Do we have to consider age groups of learners, like which category is more famous among which age groups?
- How old is the data. It is mixture of past few years or it is just recent year data?
- What are the features given to me for analysis?

DATA ANALYSIS PORTFOLIO

Design

Initial condition of data –

- Contained null values.
- Data had duplicate records
- Some cell data was incorrect.
- Headers were incorrect
- In some numerical columns, I found some non numeric details.

Steps I took to clean the data –

- I removed all duplicates using 'Google sheet' command.
- I dropped all rows which contained null values.
- I replaced all incorrect values with corrected ones.
- I renamed few headers.
- I corrected numerical values

Visualization tools used –

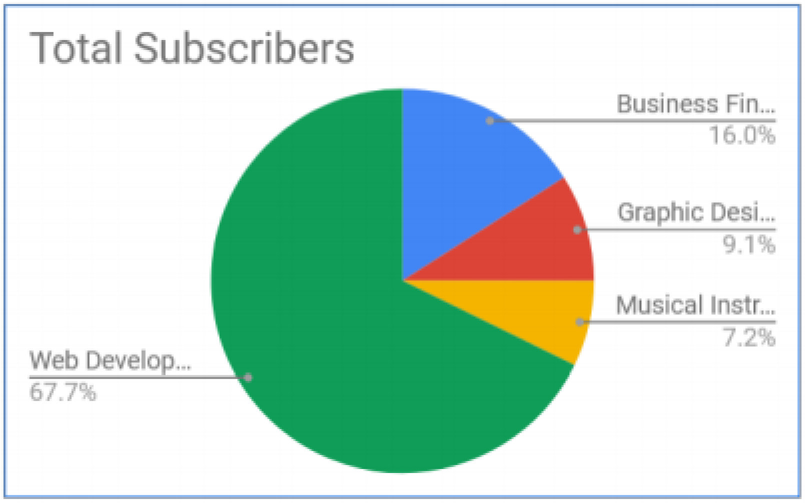
- Used excel sheet to create pivot tables to see the grouped data.
- Used 'Tableau' for plotting graphs and charts for getting the clarity and patterns in data.
- Pivot table can easily show that what range of data is there for particular category.
- 'Tablaeu' made my work much easier and it gives very clear representation for analysis.

DATA ANALYSIS PORTFOLIO

Findings

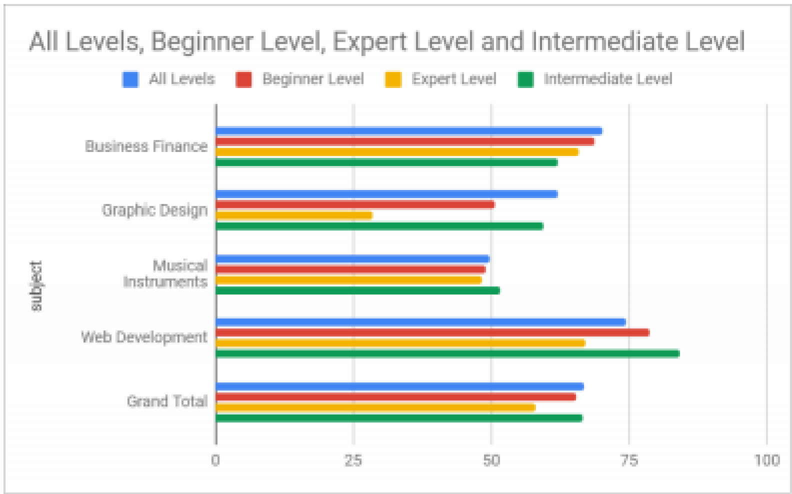
Here we have 4 categories in the data. So we are visualizing the number of subscribers of each category to check the inclination of subscribers.

Figure 1: Total Subscriptions by Category



Now we are checking category wise, the requirements of majority of users. It means that whether they are more interested in beginner or intermediate or advanced courses.

Figure 2: Users by Skill Level



DATA ANALYSIS PORTFOLIO

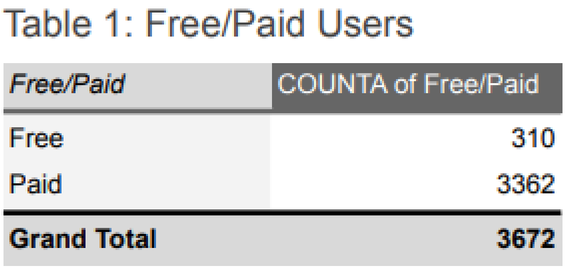


Table 1: Free/Paid Users

Free/Paid	COUNTA of Free/Paid
Free	310
Paid	3362
Grand Total	**3672**

Here we can see free users are more than paid users.

Sum and average subscribers for each subject category.

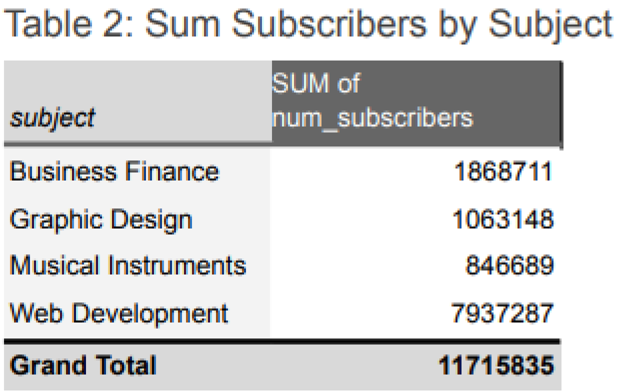


Table 2: Sum Subscribers by Subject

subject	SUM of num_subscribers
Business Finance	1868711
Graphic Design	1063148
Musical Instruments	846689
Web Development	7937287
Grand Total	**11715835**

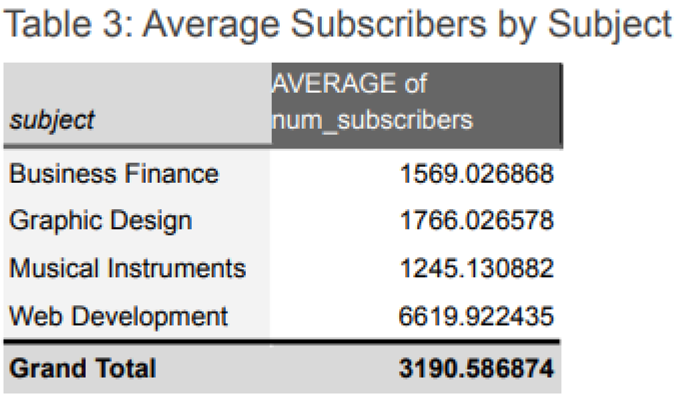
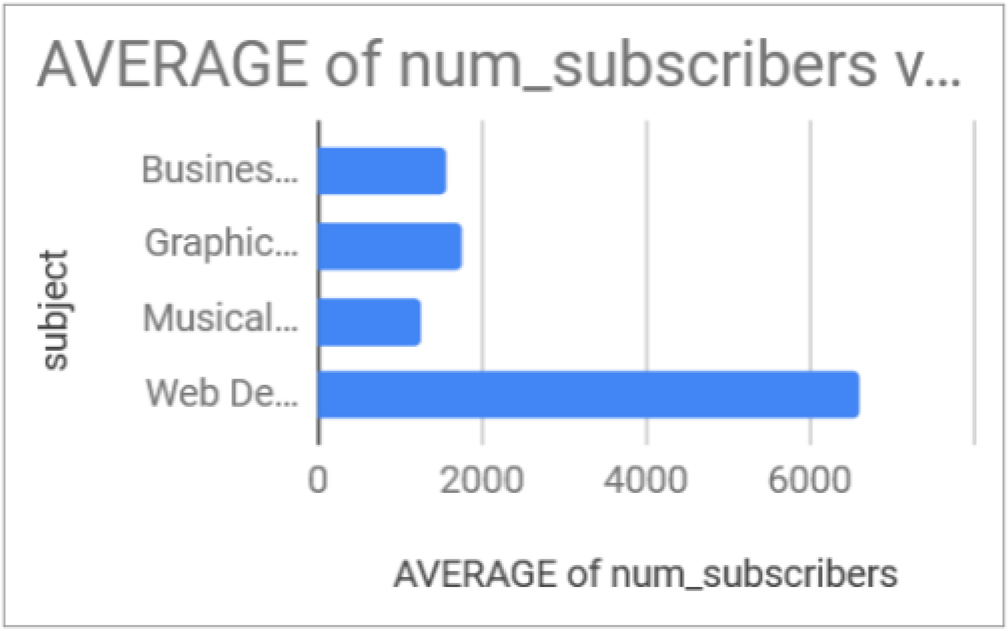


Table 3: Average Subscribers by Subject

subject	AVERAGE of num_subscribers
Business Finance	1569.026868
Graphic Design	1766.026578
Musical Instruments	1245.130882
Web Development	6619.922435
Grand Total	**3190.586874**

DATA ANALYSIS PORTFOLIO

Figure 3: Average Subscribers by Course



DATA ANALYSIS PORTFOLIO

Analysis

➢ There are more paid courses than free courses but majority of student are opting for free courses.

➢ Web development is most popular among all subjects. It covers 67.75% of all subjects.

➢ Out of 20 most popular courses, 17 are web development courses. Free courses are more popular than paid courses.

➢ Graphic design expert courses got best average rating among all courses.

➢ Almost all courses in top 20 are beginner friendly courses

Conclusion

➢ Since majority popular courses are of web development. So we need to charge fee for free courses. We will not make it very expensive since no student will opt for it then, but we can change it from free to affordable courses.

➢ We can make other courses like ‘business development’ and ‘musical instruments’ free in initial days to attract students. When enrollment will increase, we can slowly make it paid with nominal amount of charges.

➢ From the rating of students, we figured out that, some courses are less rated. We can give feedback form to students in which they can write their requirements or improvements which they need in different courses. By their feedbacks we can easily make courses for popular.

➢ By analyzing time duration of different subjects, we found that web development curriculums are more extensive than others. So we need to make other courses more extended.

DATA ANALYSIS PORTFOLIO

Iris Project Description

This is the business problem which required step by step analysis of data and find the optimal solution of the problem.

● **Business problem**: We have details of three categories of flower. We have to find category of new flower, if it is there.

● **Purpose of this report**: The purpose of this report is to find the details of petals of different species of flowers and find out that new flower will fall in which category by past analysis.

● **Actions**: Now I have clear business problem which needs to be solved. I started with data cleaning. I removed all duplicates and null values. I also replaced incorrect values with correct ones. Now I have correct and accurate data, next thing I did was data visualization. For that I used both excel and Tableau. I tried to derive relationships among data for find insights and patterns in data. I drew pie charts to see percentage of each flower type. Then I drew bar graph to see what types of species are there in majority.

● **Result**: I got multiple outcomes from analysis:-

- In data all species are equal.
- Iris-virginica has highest sepal length, petal length and petal width.
- Iris-setosa has minimum sepal length, petal length and petal width.
- All species which has more petal length has more petal width. So graph will be linear increasing.

DATA ANALYSIS PORTFOLIO

The Problem

What is the business problem?

We have details of three categories of flower. We have to find category of new flower, if it is there.

How long do you have to work on this project?

I have to work for at least two weeks to identify this complete scenario and to find patterns in the data for deriving final conclusions.

What data should be collected to understand this problem?

For this problem, I will need data about flowers with their sepal and petal length, I need categories of flowers, species of each flower.

How should it be presented?

It should be represented in pictorial form. All increasing and decreasing patterns should be represented by graphs. Amount of share each flower category has should be depicted by pie chart.

What questions would you ask to better understand the business problem?

I would ask following questions for better understanding of problems:-

- Do we have to consider color of petals, like which category has which color?
- How old is the data. It is mixture of past few years or it is just recent year data?
- What are the features given to me for analysis?

DATA ANALYSIS PORTFOLIO

Design

Initial condition of data –

- Contained null values.
- Data had duplicate records
- Some cell data was incorrect.
- Headers were incorrect
- In some numerical columns, I found some non numeric details.

Steps I took to clean the data –

- I removed all duplicates using ‘Google sheet’ command.
- I dropped all rows which contained null values.
- I replaced all incorrect values with corrected ones.
- I renamed few headers.
- I corrected numerical values

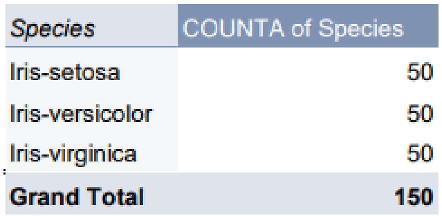
Visualization tools used –

- Used excel sheet to create pivot tables to see the grouped data.
- Used ‘Tableau’ for plotting graphs and charts for getting the clarity and patterns in data.
- Pivot table can easily show that what range of data is there for particular category.
- ‘Tablaeu’ made my work much easier and it gives very clear representation for analysis.

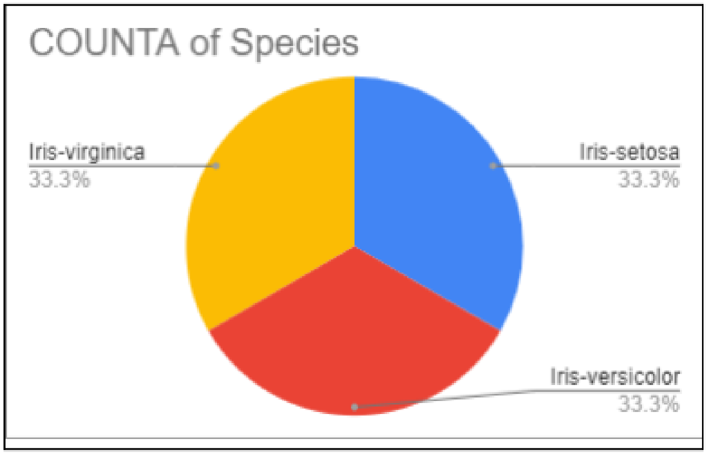
DATA ANALYSIS PORTFOLIO

Findings

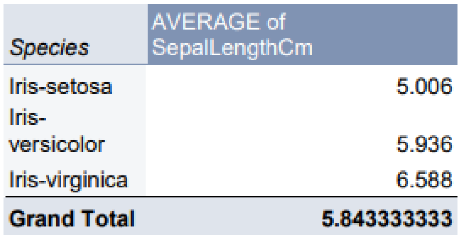
Count of each species.



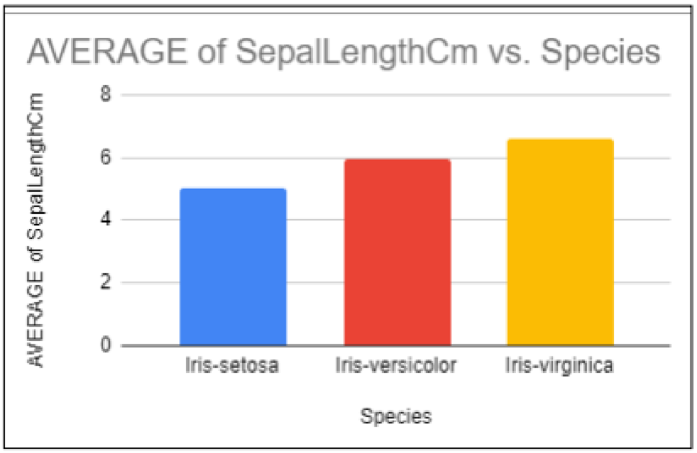
Species	COUNTA of Species
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50
Grand Total	**150**



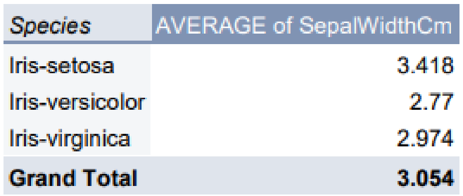
Average length and breadth of sepal in cm



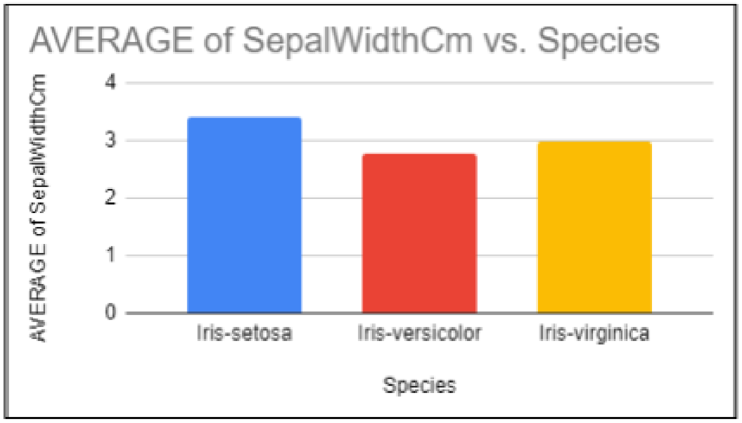
Species	AVERAGE of SepalLengthCm
Iris-setosa	5.006
Iris-versicolor	5.936
Iris-virginica	6.588
Grand Total	**5.843333333**



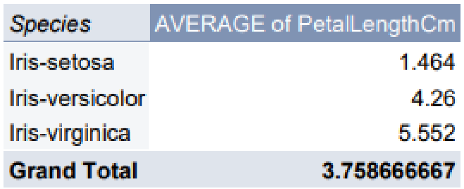
DATA ANALYSIS PORTFOLIO



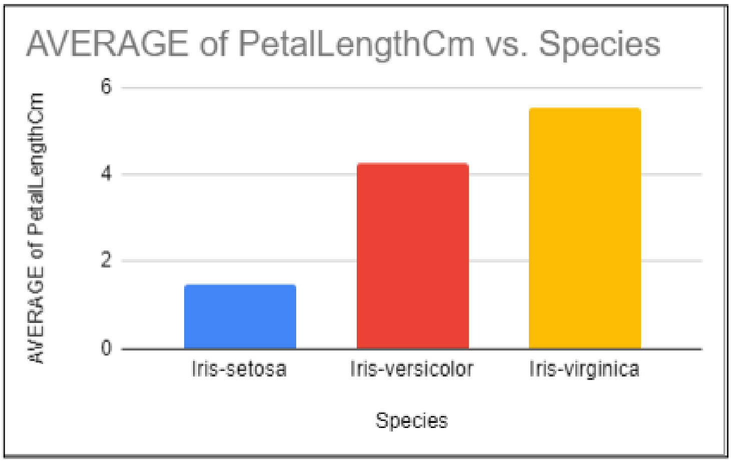
Species	AVERAGE of SepalWidthCm
Iris-setosa	3.418
Iris-versicolor	2.77
Iris-virginica	2.974
Grand Total	**3.054**



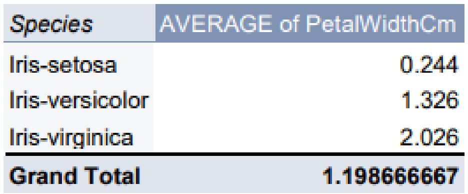
Average length and breadth of petal in cm



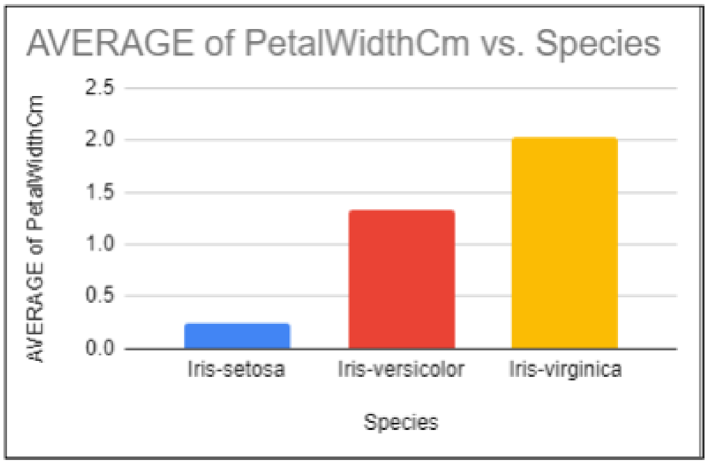
Species	AVERAGE of PetalLengthCm
Iris-setosa	1.464
Iris-versicolor	4.26
Iris-virginica	5.552
Grand Total	**3.758666667**



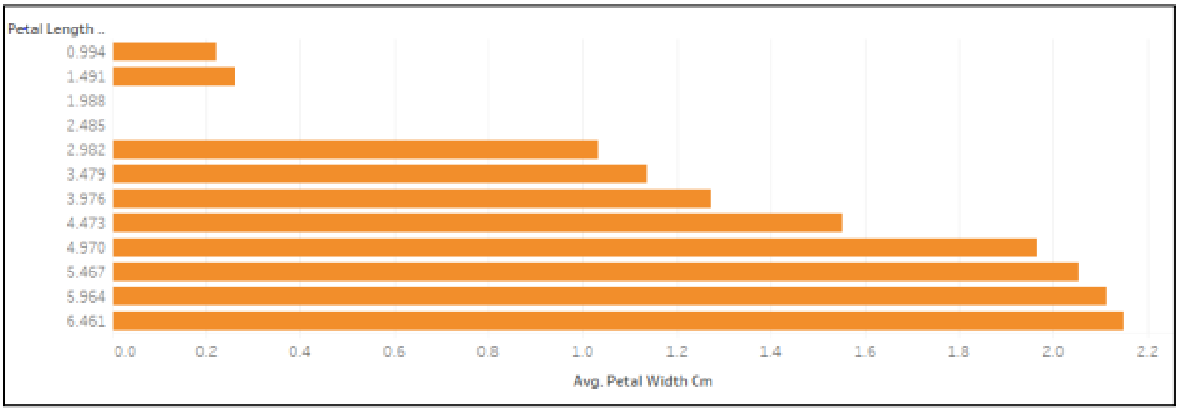
DATA ANALYSIS PORTFOLIO



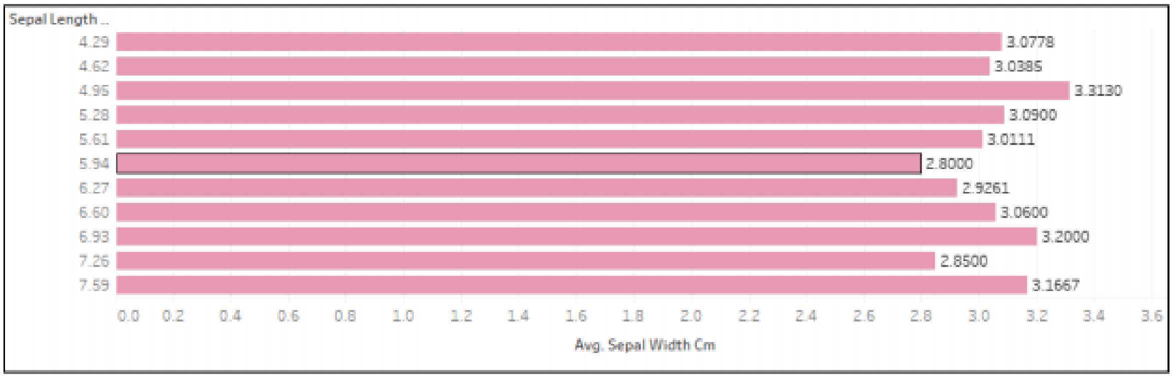
Species	AVERAGE of PetalWidthCm
Iris-setosa	0.244
Iris-versicolor	1.326
Iris-virginica	2.026
Grand Total	**1.198666667**



- **Average petal width for different petal length**



- **Average sepal width for different sepal length**



DATA ANALYSIS PORTFOLIO

Analysis

- In data all species are equal.
- Iris-virginica has highest sepal length, petal length and petal width.
- Iris-setosa has minimum sepal length, petal length and petal width.
- All species which has more petal length has more petal width. So graph will be linear increasing.

Conclusion

- Since by analysis, we can easily identify that flowers which has high petal length and width can be ‘Iris-virginica’.
- Since by analysis, we can easily identify that flowers which has low petal length and width can be ‘Iris-setosa’