CptS 475/575:
Data-Science
Fall 2024
SHARATH KUMAR KARNATI
011852253
Assignment 1

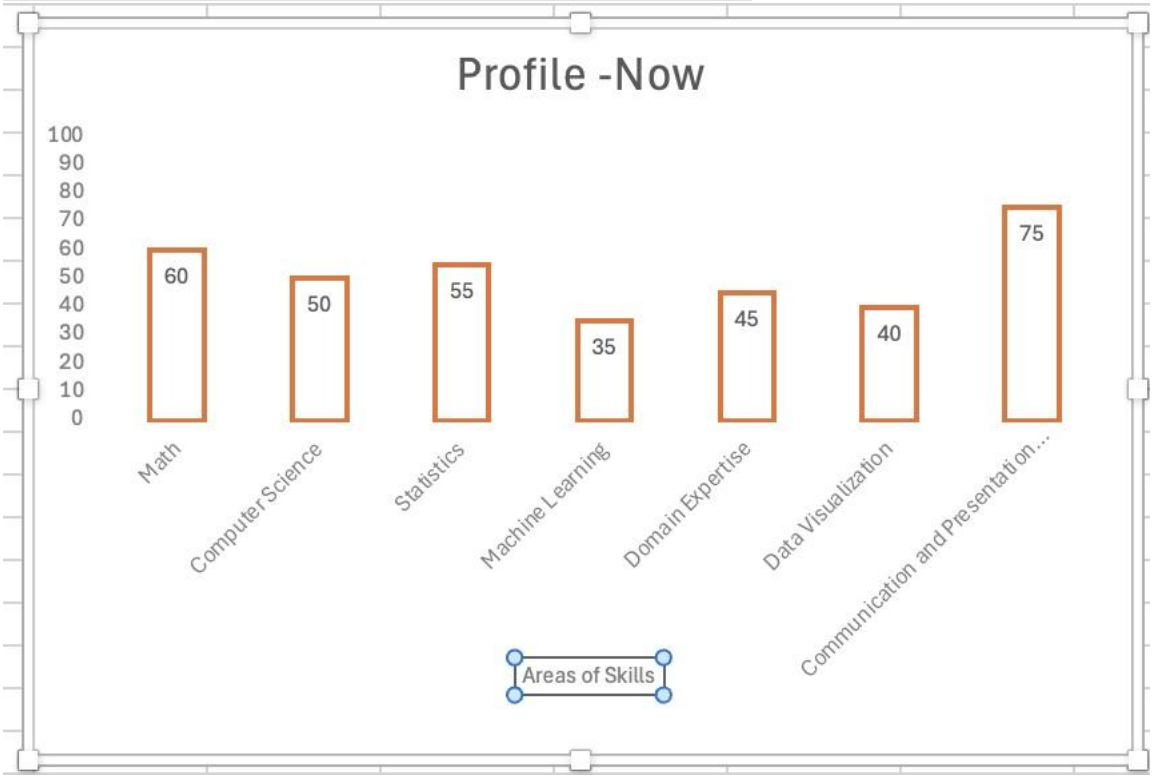
Create Data Science Profile of Yourself and Reflect on an Article on Data Science.

TASK-1:

1.a. The areas in the horizontal axis could be ordered in a number of different ways. What ordering in your opinion would be most effective(and aestetically pleasing) and why? Create your profile in the order you chose.

ANSWER:

CURRENT PROFILE :



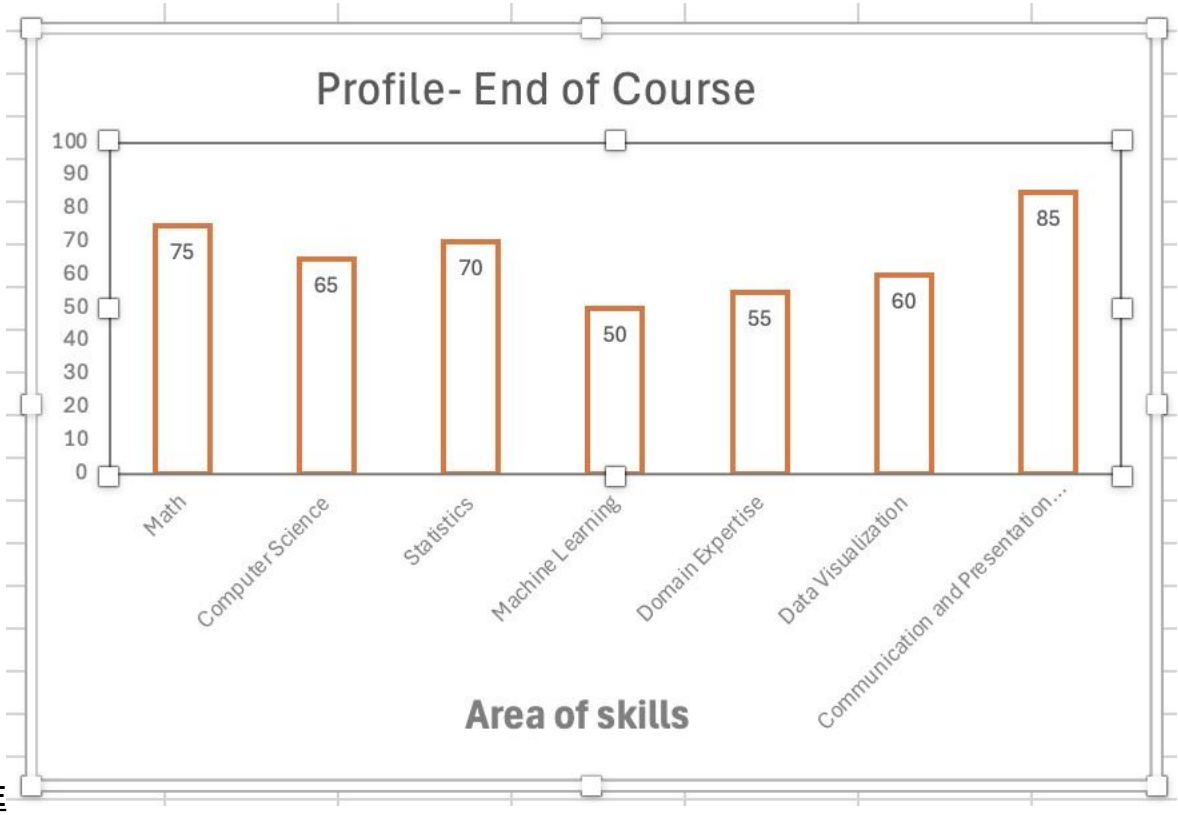
I believe that the perfect order to visualize these skills will be :
Math --- Computer Science ---Statistics --- Machine Learning --- Domain Expertise --- Data Visualization --- Communication and presentation skills.

It is because, Math, Computer Science and Statistics are all the basics that we need to have a solid foundation on to understand how machine learning, data visualization and other things that are related to computer science works. We need to have a clear understanding of these three so that we can evolve easily into any branch of computer science.

Then, we can try learning and understand advanced things like Machine Learning and Domain Expertise etc. You need to know the basics of Machine Learning since different models are key to making predictions. Being an expert in the domain helps you make sense of the results and figure out how to use the data effectively to solve problems.
Later, Data Visualization comes into play to project our results perfectly so that everyone will easily understand about our work or applications.

Finally, Communication and Presentation skills which are very important but we should first have a good understanding of the anything to present. So, I feel like knowledge is more important than any other thing. If we have knowledge we can easily grasp other essential things later.

AFTER COURSE COMPLETION, PROFILE I LIKE TO HAVE



1.b. (10 points) Is there a skill (bucket) you think should be added to this data science profile? A skill you think should be removed? Specify and justify briefly.

ANSWER:

I think that the all the above given skills are crucial and important in the view of a computer science student. They all are essential for an individual to succeed in this field. Hence, there is no need to remove a skill from above.

The skill that I think to add will be Artificial Intelligence because when we have a good understanding of AI, ML and DS it will be like a super power, we can create wonders with them and also AI can speed up the data analysis process.And also, we can add python to the list, because python is so crucial when we have to implement any of these in real life problems or projects. So, it would a good idea to add it to the above list.

TASK-2:

2.a. (15 points) The author identifies a few ways in which data science differs from statistics. What are those ways?

ANSWER:

In summary, Vasant Dhar explains that data science differs from traditional statistics in several ways. Data science deals with large, unstructured, and diverse data types like text, images, and videos, unlike statistics, which usually works with structured and homogeneous data. To handle this complexity, data science integrates tools and methods from various fields such as computer science, linguistics, and sociology. It also focuses on predictive modeling using machine learning, which allows for forecasting future outcomes, whereas statistics primarily explains existing data without emphasizing prediction. Additionally, data science often involves automated processes and advanced techniques to discover patterns in data, beyond the capabilities of traditional databases and statistical methods.

2.b. (25 points) In the section of the article headed "Knowledge Discovery"(pages 70 to 72 of the article), the author makes a distinction between domains in terms of the predictive power of their theories (models). Secif-ically, the author points out that models in the physical sciences are generally expected to be "complete", whereas in the social sciences they are generally "incomplete". The author discusses ways in which "big data"could potentially put domains on both ends of this spectrum on firmer grounds in terms of theory development. Give a brief summary of the ways the author identifies. Do you see any additional ways than what the author sees? (If the discussion in this section of the article reasonated in some ways with your own research or work you do, feel free to incorporate that in your answer.)

ANSWER:

In the article, Vasant Dhar highlights the differences in how predictive models are viewed in physical and social sciences. In physical sciences, models are often considered "complete," meaning they can predict outcomes with high accuracy based on established theories. For example, the behavior of a space shuttle can

be predicted reliably using physical laws. In contrast, social sciences deal with "incomplete" models that simplify reality and are based on assumptions about human behavior, which are less precise.

The author argues that big data has the potential to strengthen theory development in both domains. In physical sciences, big data can enhance the precision of predictions by reducing errors related to model specification and sample size limitations. In social sciences, big data enables the discovery of patterns and connections that might not be evident with smaller datasets. This can lead to the refinement of existing models and the development of new theories. Additionally, big data allows for large-scale experiments and real-time monitoring, which can improve the accuracy of social science models.

Beyond the author's points, big data can also help address complex real-world problems by integrating knowledge from various fields, leading to more accurate and adaptable models. This interdisciplinary approach can provide a more holistic view of the problems being studied, enhancing both prediction and theory development in multiple domains.

2.c. (10 points) Imagine you were asked to write a "head-line" (as you see in newspapers) for this article, followed by two or three very telling summary sentences. What would your headline and the summary sentences be?

ANSWER:

Headline:
"Data Science: The Future's Predictive Powerhouse or Premature Hype?"

Summary:
As industries increasingly rely on data science and machine learning, the potential to predict future outcomes is more promising than ever. However, questions remain about the reliability of these predictions, particularly across different fields like physical and social sciences. The debate centers on whether data science's current capabilities can fully be trusted to shape critical decisions for the future.