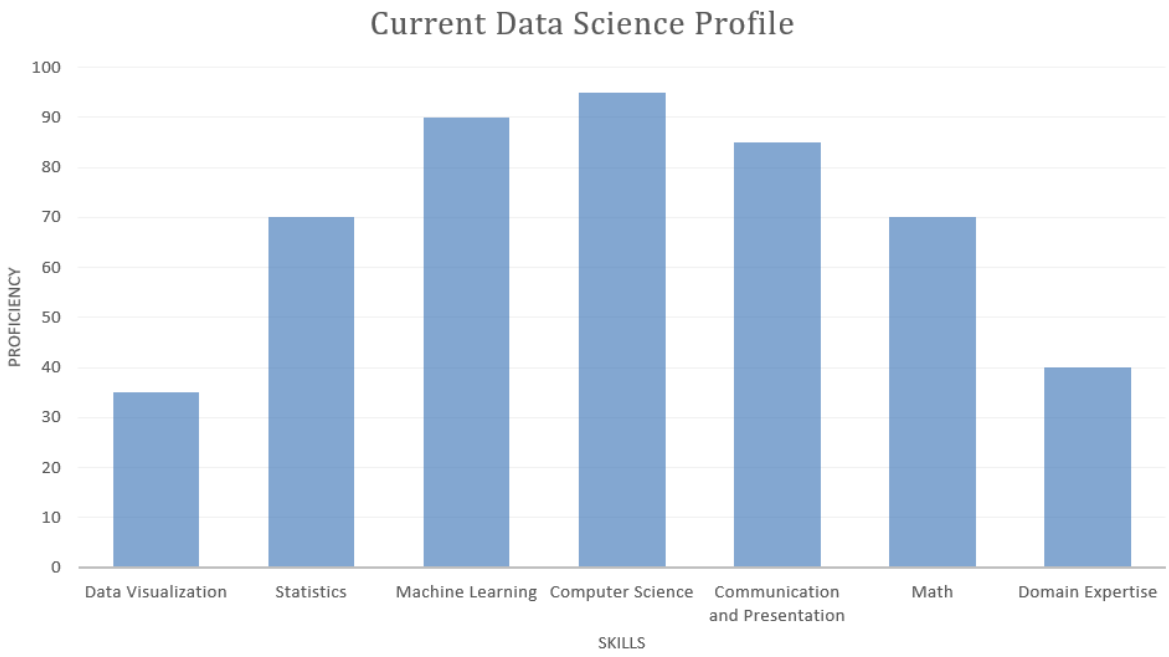
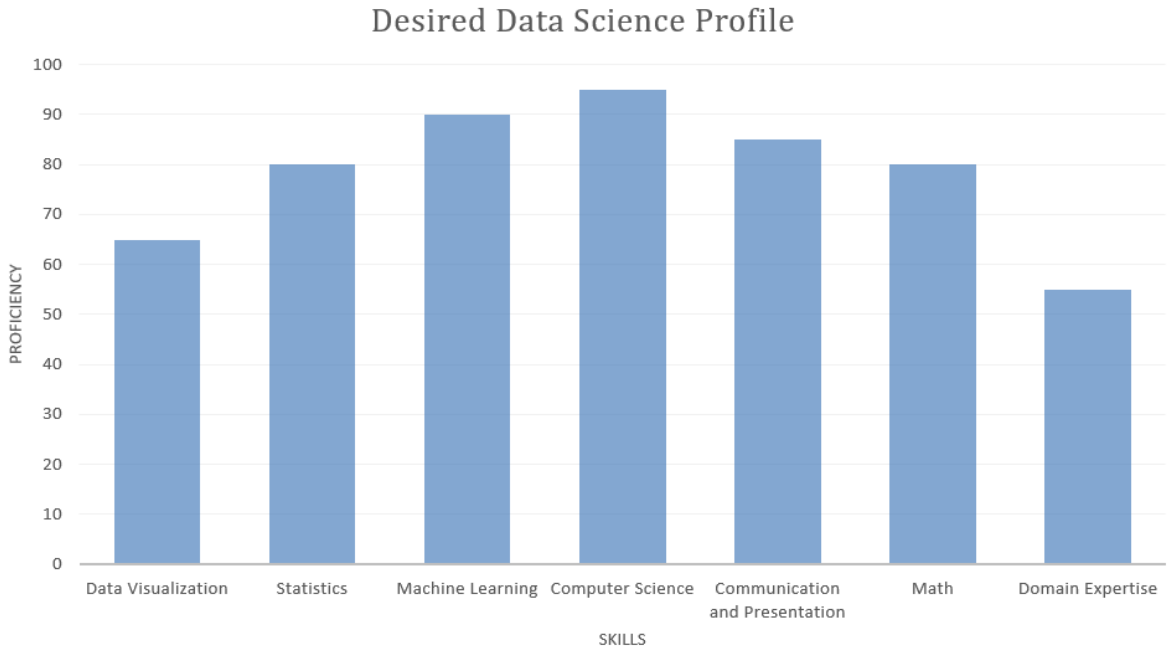
CptS 575 Data Science
Assignment 1

Task 1:





1. A) Symmetry is always aesthetically pleasing. Hence, I placed the skills I am the most proficient in (Computer Science, Machine Learning) in the centre and the ones with lower values on either side. If I had ordered the skills in a decreasing or increasing order of proficiency, it might have unintentionally implied a trend of some sorts at a cursory glance which I didn't want. [Note: I have been working on a project involving data analysis of climate and population data for over 3 months now, and hence I have entered a non-zero value for Domain Knowledge].
1. B) No, I think all these skills are the essential ones in a data scientist's repertoire. The list of skills provided is a good one.

Task 2:

2. A) The data that is used in the field of statistics is structured by nature. This is not so in the field of data science. The data gathered in the latter is gathered from various fields and hence has a heterogeneous and unstructured nature and can be in any format – text, images, video, etc. These lead to complex relationships between data entities and consequently, analysis requires integration, interpretation and identifying actionable intelligence.
   The use of markup languages in data science lets computers interpret data automatically and make intelligent decisions. Markup languages are not a part of the field of statistics.
   Statistics can be viewed as an essential part of data science (a subset of data science). But data science also involves systematic study of organization and analysis of data, its role in inference and confidence in that inference. These factors are not present in statistics.
   Data Science, unlike statistics doesn't end at displaying information/data for human consumption. It also involves the automation of the data generation of the process to be consumed by machines and make decisions (this sub-field is where Big Data comes in).

2. B) For firmer grounds in terms of theory development, big data provides large amount of data which can be used for accurate predictive models. There might be less causal insights, but as long as prediction errors are small, the ground becomes more firm in terms of theory development.
   If an observed pattern can be used to make predictions, it might be worthy of publication and deep inquiry. Simple predictive models can be used as potential components instead of a causal model tested by the data.
   Big data allows data scientists to reduce the types of errors that come up in prediction. They are misspecification of a model, samples used for estimating parameters and the third is randomness. Big data allows us to consider models that make fewer assumptions as there is large amount of data to test models and compute error bounds. The second error can also be significantly reduced by big data as sample estimates become reasonable proxies for the population.
   The main theoretical limitation of observational data is that data is generally passive. It represents what actually happened in contrast to the things that could have happened if the circumstances were different. This can be dealt with as in the internet era, inexpensive large scale randomized experiments can be conducted.

Big data also allows social scientists to observe human behaviour at a high level of granularity and variability with the increasing interactions and activities mediated by the internet.
I think along with Big Data, an infusion of domain knowledge will result in the perfect cocktail. If the model that is produced by Big Data can be guided by already established domain knowledge, then that might be useful in bringing the number of false positives down (By false positives, I am referring to predictions that fit the pattern uncovered by Big Data but goes against irrefutably established domain knowledge).
I once did a project on time series prediction of stock market data and tried out different models. The ones that could come up with the best predictions were the ones which were the least supervised ones. That is, they were left to their own devices to identify patterns as opposed to those which had some inherent bias (example: a linear model used to fit a non-linear dataset. A univariate model which forces you to choose what you think is the best predictor to describe a multivariate dataset).

2. C) Headline: Data Science demystified.
Summary: With an increasing glut in the world of data, the identification of actionable intelligence has never been easier. Machine Learning, and Big Data comes together to change the ever shifting landscape of predictive modelling and analysis.