In this part of our project, we want to evaluate course materials with perspective of salary from analyzing Kaggle survey from 2018 and 2019. We will focus on the salary buckets for both 2018 and 2019 surveys as our target variable. We will investigate which skills have a more positive correlation with higher annual compensation. In MIE 1624, introduction to data science and analytics, we are focusing on Python as the programming language, however, we need to explore thoroughly results of the surveys to be sure our choice is good enough. First, we need to clear our data since both have a significant amount of noise associated with them. We are going to only focus on Canada and the United States for this report. Owing to the fact that we know each country has its average salary based on the many different factors. Take figure 1 as an example, as we can see USA has almost a normal distribution, on the other hand, India’s salary is more skewed toward lower salary buckets.

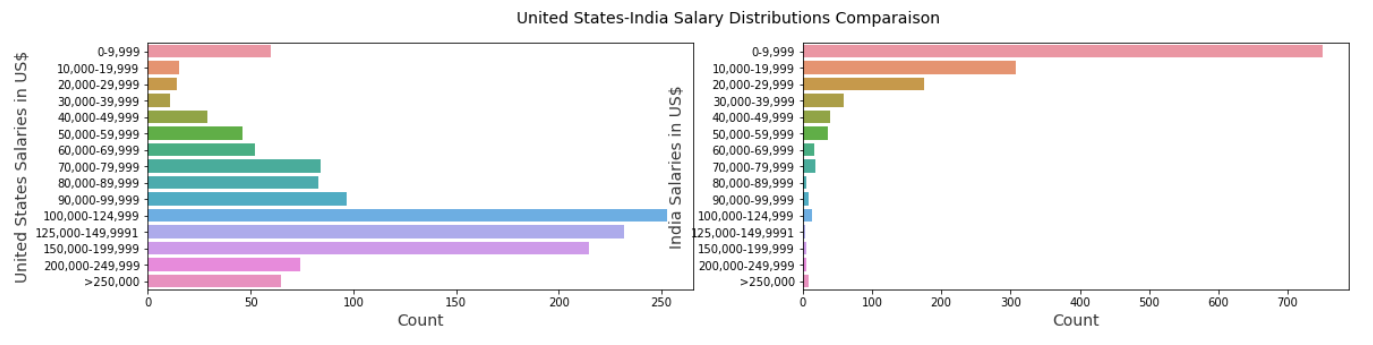


Figure 1- USA vs India salary buckets

We want to be very precise with the choices that we want to make to redesign the curriculum for MIE 1624 with a salary-wise perspective. Therefore, we are going to focus on the two countries that have similar characteristics. Also, many factors would contribute to a higher salary other than skills. Factors such as years of experience, age, gender, etc. However, in the scope of this project, since we are only looking for skills, we are not going to consider other factors. To start, we first explore the 2018 data set and compared the result to what we had in the case of 2019 surveys. We want to see whether passing one year from the survey would change things or not.

We are going to start our analysis by looking at the difference we would have in the case of programming language. As we mentioned before, in MIE1624 we are focusing on Python as the primary programming language. Even though the result of the 2018 Kaggle survey indicates that Python is used for almost all the salary buckets, SQL and R seem to be useful tools to achieve higher annual compensation for a data scientist.

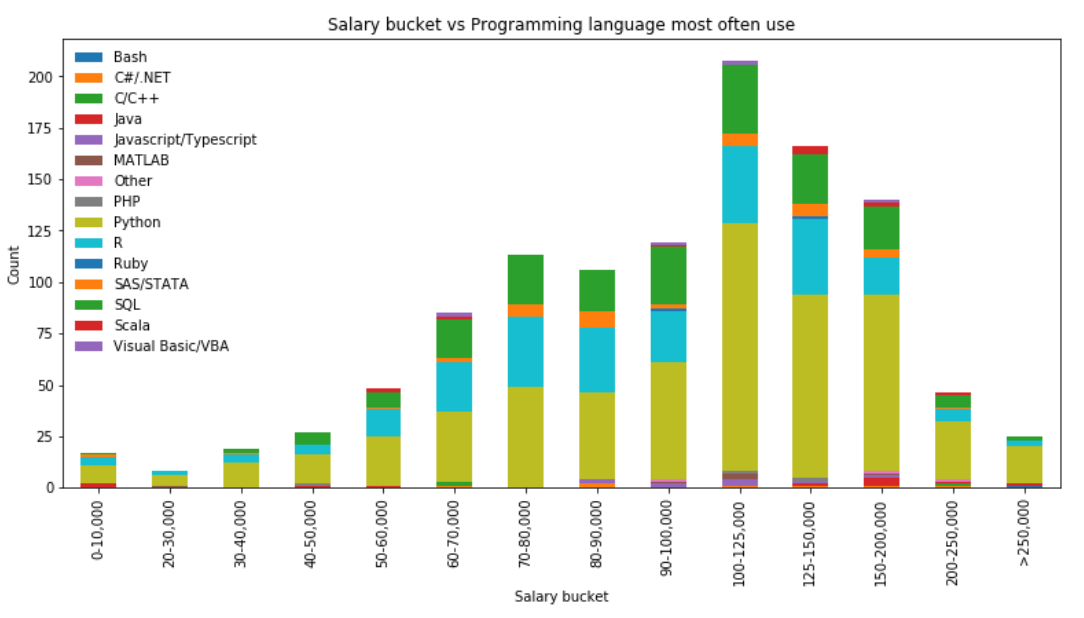


Figure 2- Programming language based on salary buckets 2018

In the case of the 2019 survey, since we did not have a specific question that asks about the primary programming language which a data scientist would use, we decided to use LASSO regularization to see which of the programming languages had a more positive coefficient which would indicate the importance of that language. Interestingly we have seen similar results in the case of the 2019 survey. Result states that SQL and R are quite useful programming languages if a data scientist plans to achieve a higher salary. In the case of skills that would contribute to a higher salary for a data scientist, we look at the question which asked the primary tool that respondents used in case of a higher salary. This question was repeated for both years. In the 2018 Kaggle survey, we can interpret that in general Cloud-based data software and APIs are mainly shown in higher salary buckets. In addition, Local or hosted development environments such as Rstudio or Jupyter notebook are presented in almost all of the salary buckets and they have the highest frequency.

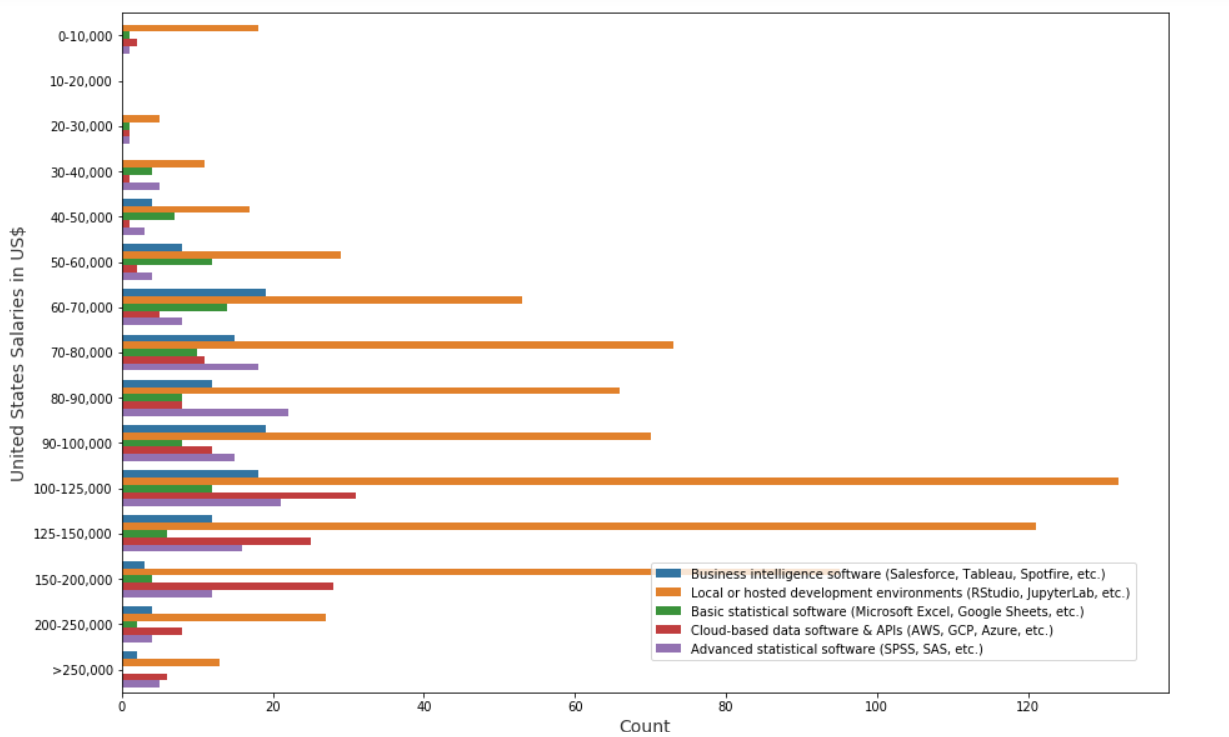


Figure 3- Salary bucket VS primary tool used by respondents 2018

Now, if we look at the data set which is provided by Kaggle for 2019, we can see that Local or hosted development environments, are used in most of the salary buckets; however, Cloud-based software and APIs are mainly shown in higher salaries.

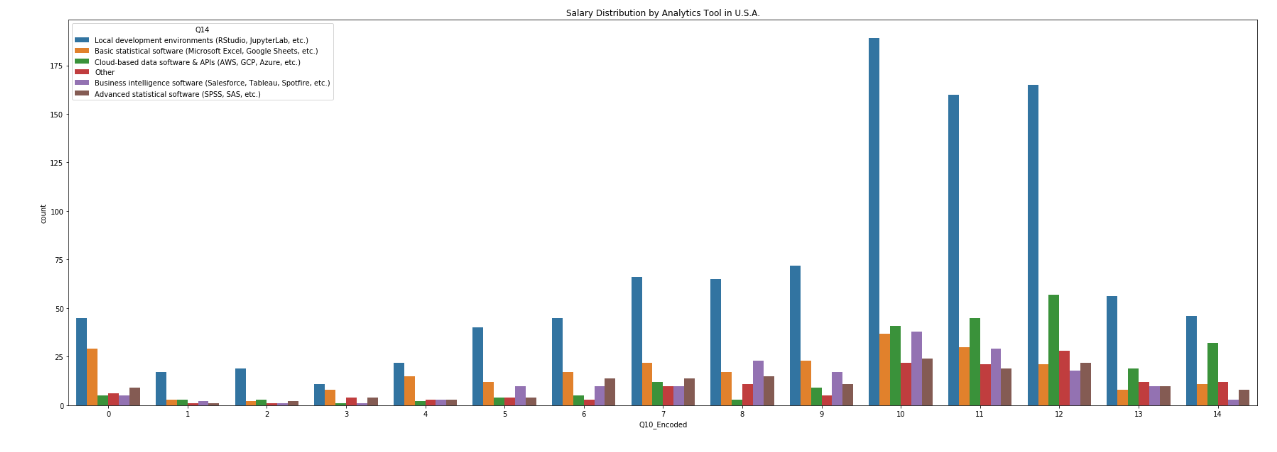


Figure 4- Salary bucket VS primary tool used by respondents 2019

In the last step of our analysis on Kaggle surveys, we realized that Building prototypes to explore applying machine learning to new areas is a very important skill that would increase the salary bucket for a data scientist. In addition, if a data scientist knows how to use IBM Cloud Analytics Engine it can increase her or his salary. As we saw from the results of Lasso, most big data analytics products are useful tools for future data scientists. Finally, from the conclusion that we have of our analysis, a data scientist must know how to work with Tabular.

In conclusion, salary-wise analysis of Kaggle surveys from 2018 and 2019, would indicate that additionally to Python, a data scientist should be familiar with SQL and R. Moreover, we need to start focusing more on cloud-based software and APIs since they would contribute to a higher salary than other tools.

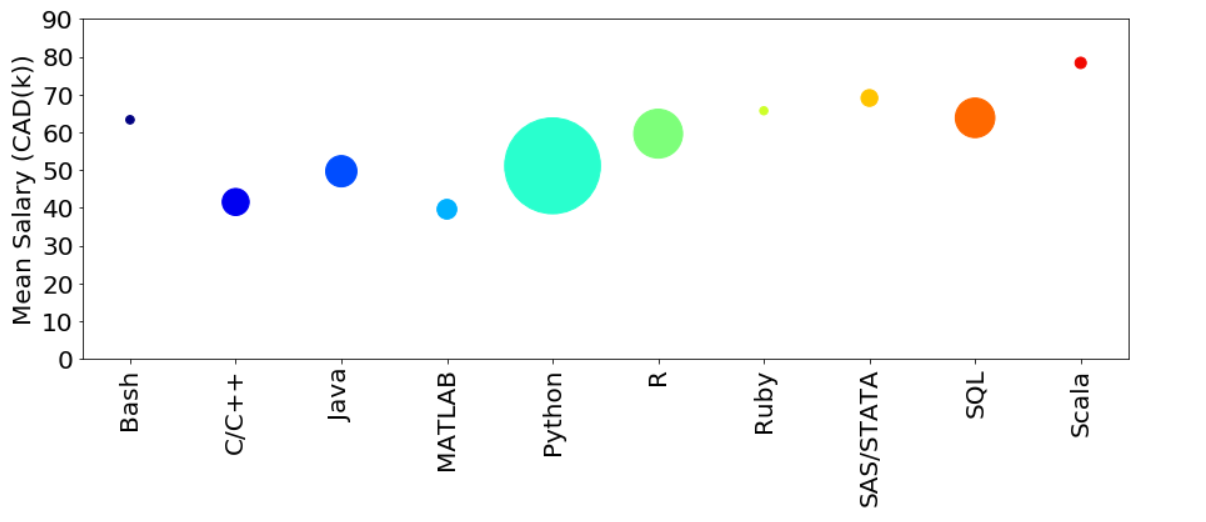


Figure 5- Programming language based on average salary

The size of each circle shows the number of occurace of each specific programming language in the average salary. We can see that again Python, R and SQl are the most important programming languages in the kaggle 2018 survey.

In the second part, we worked on the data that we scraped from galssdoor. For this project, by web scraping we were able to obtain roughly 3000 job postings in the field of data science and analytics from glassdoor website. In the last part of this project, we evaluated the skills required to redesign curriculum for MIE 1624 from salary perspective by analyzing Kaggle's surveys from 2018 and 2019. In this section, we would like to evaluate skills from the perspective of job opportunities. After this course, students should be able to at least apply for internship positions in data science field. Thus, these postings would explain to use what are the main skill that firms are looking for from their applicants.

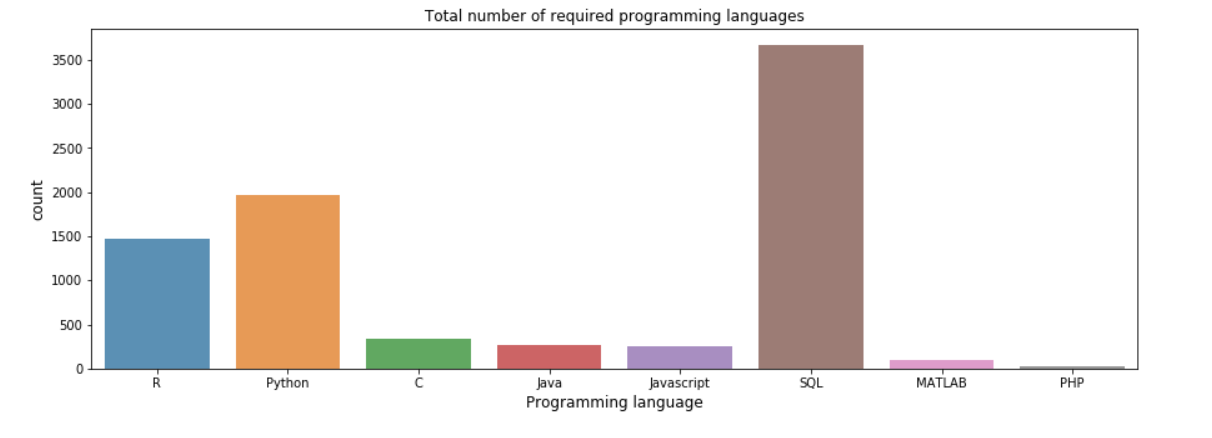


Figure 6- Required Programming language for job postings

As we can, similar to what we had in case of Kaggle's surveys most of the job postings required SQL as the programming language, in the second position we have about 1000 job postings that require Python, and at the third place we have about 750 job postings published R as their required programming language. However, many of these postings state multiple programming languages. In case of MIE 1624, since the course name is stated as introduction to data science and analytics, it is better to use Python as primary programming language, similar to the current case. Due to the fact that many of the students are more familiar with this programming language because it is one of the most user-friendly environment for coding. However, we suggest that as an extra work, class would have several tutorials about SQL and an assignment that need to be done with SQL as the programming language. This would engage and familiarize students with the programming language that seems to have a high demand in the field of data science and analytics.

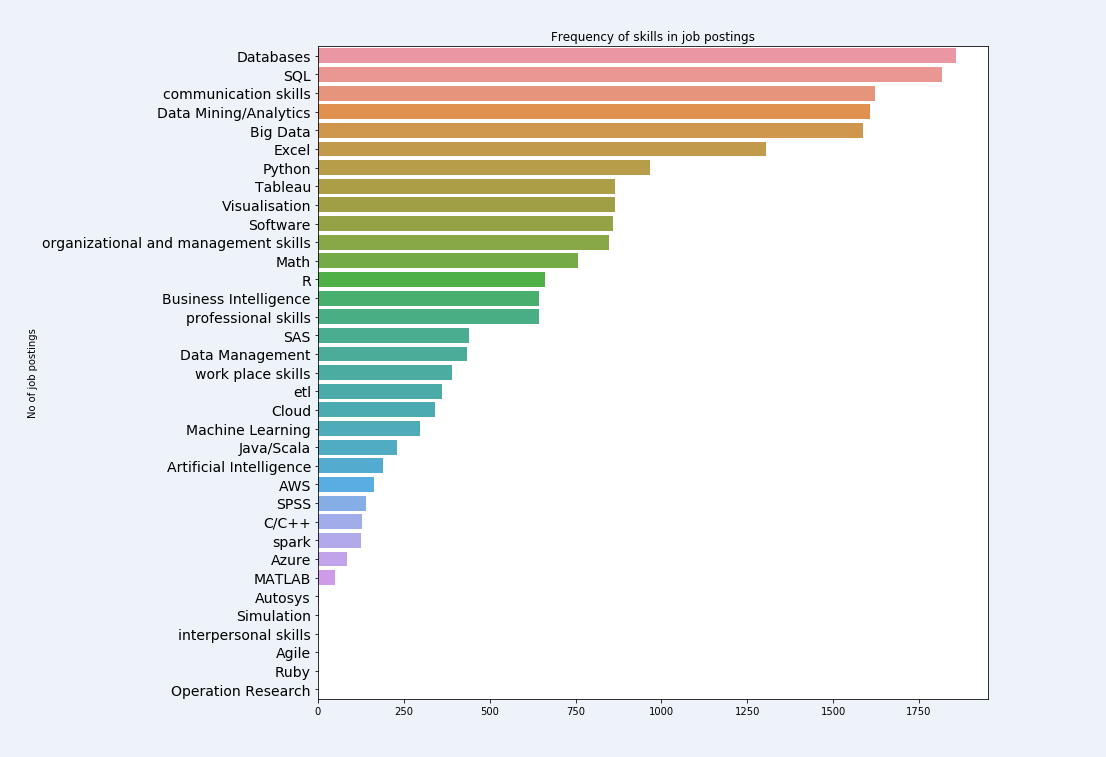


Figure 7- Required Skill frequency based on glassdoor job postings

AS we discussed before, considering programming languages, Python and SQL are the most important ones that the focus of this course should be on that. In addition, from the graph we can interpret, that database is one of the most popular skills that job postings are seeking for. However, in MIE 1624, introduction to data science and analytics, we have not focused on this skill. Thus, from the analysis, we are intrigued to add concept of database in this course. Communication skill is a critical skill that a data scientist should have to be able to progress in his or her career. Nevertheless, it is not considered a skill that have in mind when designing our course. Data mining seems to be quite important in the field of data science. We are exposed to data analytics in some degree, however we have not work either in big data or data mining. Tableu, which is visualization tool, is a required skill in most of job postings. Cloud and machine learning are among the skills that were asked in more than 250 job postings.

The result of this analysis gave us some perspective about what needs to be considered during redesigning the course. However, we are going now to use another data set that we were able to find in Kaggle website. It includes more than 10000 job postings including internship, full time, contractor and part time jobs. We are going to focus only on full time jobs. The data frame is not fully cleaned so we need to also clean it in this part. From the results that we had from this data frame and the previous works that were done, we then going to decided how to change the structure of the course.

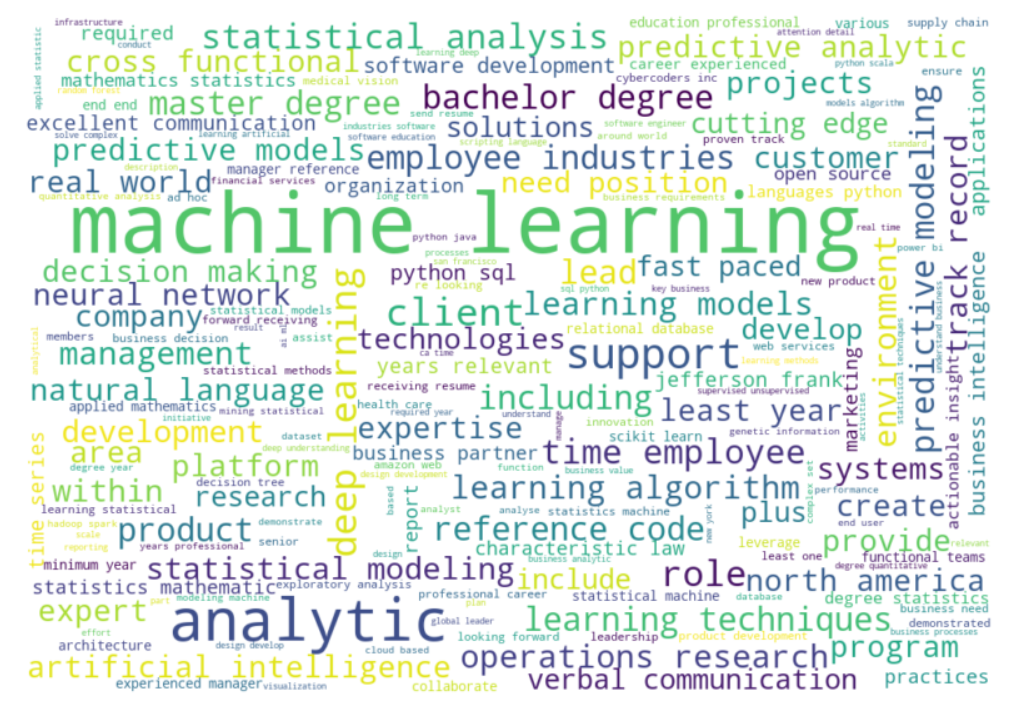


Figure - Kaggle's job posting word cloud

Machine learning, analytics, statistical analysis, artificial intelligence are the most important features that are represented in most of the job postings.

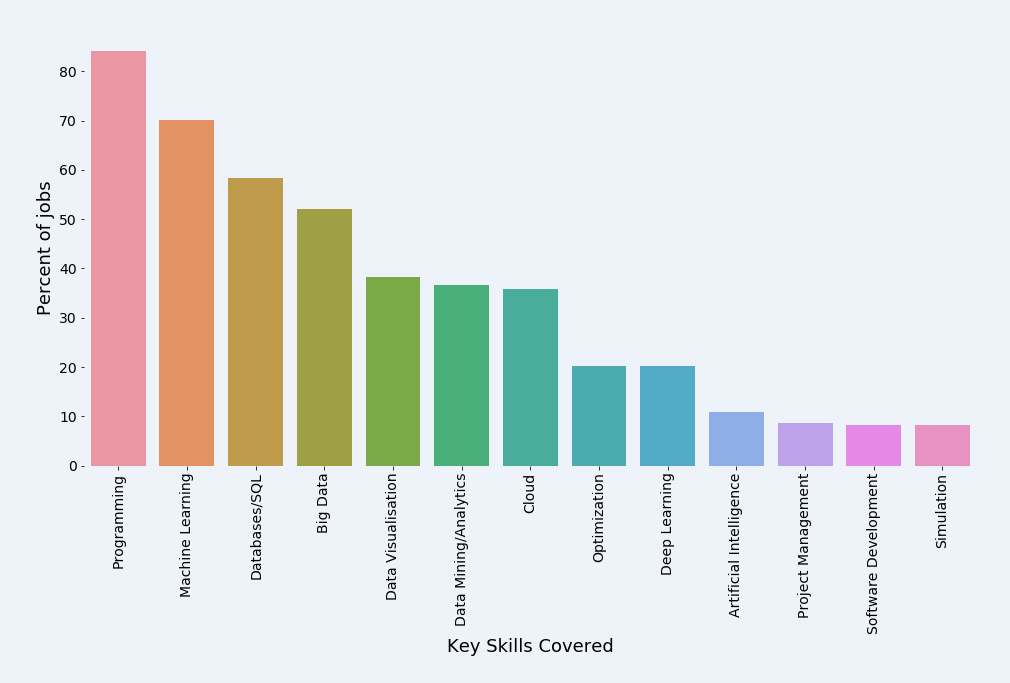


Figure 9- Kaggle's job psoting general required skills

Based on the analysis on the kaggle's job posting data frame, Programming is should be included in our course, which is considered. Machine learning algorithm should be taught in the course. However, we suggest to go with the presentations about different ML algorithm and teach only two or three important ones. Database and SQL is not covered in this course, which is in the opposite of our analysis. We need to devote a lecture or two to big data. Data visualization should be implemented in the course project and assignments as we have in case of current curriculum. Data mining is another skill that we are not covering in this course. Nevertheless, giving a introduction to data mining will be useful for students. Cloud computing also is a topic that can be covered in the form of an assignment or small project. Deep learning can be allocated to another course, however since optimization is a useful tool for most of

engneers we are proposing to have it included in this course. The following is our course structure:

