I. Summary of raw data

a) In order to answer the question of what are the most important data science skills, our team looked at the 2017 Kaggle Machine Learning and Data Science Survey.

c) The survey was conducted in the span of two weeks in August 2017 and contained 16716 total responses internationally that was communicated through Kaggle by email, discussion boards and social media channels.

b) The data included 291 of both multiple and freeform responses along with a schema to decode responses and questions and a csv containing conversion rates for currency.

II. Loading and Cleaning Data

a) The data came as a .csv file that we loaded into a sql Database but as the data was available publicly, we loaded the dataset into github and called the data from there into a dataframe in R.

b) There were two main sets of data that we worked with namely the Free Form questions and Multiple Choice that were separated to preserve anonymity of the responders. The supplementary schema and conversion tables were used for reference.

c) We double checked that all of the questions from the mcq and freeform were contained in the schema and checked whether or not the free form questions were paired with the multiple choice questions which were not the case.

d) Since there were over 200 questions asked, some cleaning was required to extract info useful to our study.

e) Cleaning of Free Form Data

i) Excluding the questions to non data science professionals or learners excluded 3 columns.

ii) The resulting dataframe was then filtered to contain only questions that contained data science skill related words which included but not limited to “skill, ability, competency, experience, aptitude etc.

iii) The resulting vector of the previous filter was used to subset our free form questions further so that it contained questions with that criteria.

iv) Further cleaning included removing columns that had all values missing in every row and double checking that remaining columns had no missing values.

v) A final schema of 35 columns were left out of the original 62 was created for further reference.

f) Cleaning of Multiple Choice Data

i) Filtered necessary columns by google sheet.

ii) Found the mode of a specified column and replaced all NA values with that value.

iii) For values that contained numeric values, the missing data was replaced by the median of that specific column.

III. Analysis

1. Filter was provided further to limit the responders answers to only those of Data Analysts and Data Scientists.
2. For our analysis, we looked at the following five criteria for analysis namely

Language Recommendation

Most Important DS Tech Skills

Most Important DS Tools

DS tools that are monetarily valued

DS Algorithms that are monetarily valued

1. Language Recommendation

i) Here the question “What would you recommend a new data scientist learning f first?” is analyzed. The counts of which are represented in this table and more easily viewed by this bar graph.

ii) We can see clearly here that Python is the language is that most scientists and analysts recommend and is chosen twice more than R with SQL coming in at a far third.

1. Most Important DS Tech Skills

i) Here the question “What are the most important DS Tech skills are?” is asked. Like above, we have a table and graph to better visualize our findings.

ii) Almost every skill was highly rated though Enterprise Tools had the largest deviance from the rest of the group. More digging must be done to find the relevant technical chops they most value.

1. Most Important DS Tools

i) To further add detail to the above analysis, we looked to answer the following questions.

“For work, which data science/analytics tools, technologies, and languages have you used in the past year?” and “At work, which algorithms/analytic methods do you typically use?”

ii) The data used to answer this was contained in “WorkToolsSelect” and “WorkAlgorithmSelect” fields. The information was parsed and extracted and placed in the following bar graphs.

1. DS tools and algorithms that are monetarily valued

i ) Some issues arose when trying to analyze by defining a quantitative measure of value being compensation. One being that location affecting salaries as they might differ by country and also experience will skew the correlation as more experienced professionals will tend to earn more.

ii) To help with our first issue, we disregarded respondents who lived outside of

the US and focused nationally. This unexpectedly did not remove respondents

Who were not paid in US dollars so we used the table given to us to convert

Currencies.

iii) Just to double check that Experience does increase salary, we have provided a graph below with linear regression to illustrate. From the graph, we can see that it is indeed true.

iv) To remove the skewness, we used the line of best fit in the above graph to estimate the compensation of every participant if they had an experience level of 1.

v) After removing a possibly complicating skew received from location and experience level, we evaluated then which skills correlate to the highest levels of compensation. The results are in the graphs shown below.

vi) The barcharts above reveal some interesting results: a good portion of the tools/algorithms that we deemed as non-important in the earlier section actually seem to have higher total compensation (for example, GANs for algorithms, and Angoss for tools). We can evaluate this a little bit further by making some scatter plots:

vii) When looking at the plot of the algorithms, there are some not so commonly used ones in which the average compensation of those who use are either very high or very large. The majority of the more commonly used algorithms seem to have average compensation values that fall closer to the middle of the plot. For the tools plot, we see that there might be a slight negative correlation between the two, indicating that those who use more obscure DS tools might have higher average compensations. We can analyze this a little further by taking a look at the correlation coefficients between average compensation and usage for both tools and algorithms:

VI. Conclusion

1. The above results provide us with some interesting conclusions: while there is a clear group of tools and methods that are most used by data science professionals (i.e. Python, SQL, R, etc.), there are some more highly specific skills (i.e. using Angoss, being able to implement GANs) that are less used but might provide a data scientist with a higher compensation. This makes sense, seeing as the skills that might land someone their first DS job are likely to be those that are more generalized. Once spending some time in the industry, more specialized tools will likely provide a more experienced data science professional with higher compensation.

While these results are interesting, the data used is nowhere near from being fully analyzed. The whole data set comprises the responses to about 230 data science related questions meaning that there are likely many more insights that could be drawn. In terms of this analysis, a next step might be to further inspect this negative correlation between compensation and usage via more advanced modelling techniques.