This notebook is based my original submission for the last year's competition. However, I cleaned up my coded and changed many other things.

Everyone lived in peace and profit until digitalization and with it came the far north there came an evil creature, the Data Monster. It ravaged through the data land, his most dreadful weapson was a messy and big dump of data that overflowing it with a massive stream of data . The Council of Data decided to convene and find a solution. After hours long and arduous discussions, one of most wisdom of council said lets consult the sacred Kaggle scripts.

I assumed that the number of people in each role who poses a skill (such as programming with python) would be a good indication of how important and commonly used this skill for this particular role is. For example, if 50% of the data analysts in the survey said that they use Python but 75% of the data scientists in the survey said they work with Python then it would indicate two things:

1.Naturally Python is a more important skill for the data scientists than the data analysts

2. Data scientists are more skilled in Python than data analysts.

Also, I compare the percentage of respondents that use a tool and compare it with the average percentage of all Kaggle survey respondents (all roles) that use this tool. Say if the percentage of Python users among data scientists is higher than the average percentage then I'd assume that Python is a distinct skill for data scientists.

# Software engineers

While they possess great programming skills and can work with cloud services they are not particularly skilled in wielding most data science related tools and libraries. However, surprisingly they seem to be skilled in computer vision

Their visualization skills lags behind most champions

They would not be your first choice for training machine learning models

They have no main quest in the data land.

## How did I create the skill bars

The idea behind my skill bars is very simple

## How I computed the median salary and XP

To compare different roles, I thought it’d be interesting to calculate the median value of their salary/compensation and coding experiences. However, the salary/compensation and coding experiences columns all contain categorical values. This would make calculating the median salary very difficult. So, I used a different method to calculate median scores for categorical values.

## How did I come up with Quests

In most of the RPG games, characters have several main quests that they must complete to finish the game. There also side quests, which means they are optional, but a character can get rewards if he/she completes them.

I used the ‘Select any activities that make up an important part of your role at work’ question from the survey data to decide the main and side quests for each character. It’s a select all that applies question which means that respondents can select several options. It’s specifically designed for data-related problems so it doesn’t cover all the activities that a respondent might do. For instance, there is no option for teaching activities but obviously for some roles such as teachers/professors this could be their primary activity.

1.First I compute the number of times an activity was selected for the whole survey population and divide it by the total number of respondents. This would be the average ratio of this activity for a typical kaggler (For instance, 45% of all Kaggle respondents said “” makes up an important part of your role at work’. This would be the average ratio of this activity

2.Then I compute the number of times an activity was selected per group and divide it by the total number of respondents per group

3. I compare 1 and 2. If 2 is larger than 1 (average kaggler) then I’d consider this activity as one of their main quests otherwise I’d consider it as one their side quests.

Mistakes I made:

I did most of my analysis in the RStudio IDE and RMarkdown files on my local machine. RMarkdown files are more stable tools . However,

Coming up with quotes was extremely challenging

Alternatively one could make the assumption that for example salary/compensation is uniformly distributed random distribution I used a uniform random distribution generator number to generate compensation estimates for each salary bracket.

7198000057740969

It would be great to look at them and see what are their strength and weaknesses. However, sometimes there are overlaps between these roles. For example, both data scientists and statisticians might wield Python but we want to know who is more proficient or a more frequent user in using Python. To answer these types of questions, I look at the percentage of data scientists that use Python and compare it with the average percentage of all Kaggle survey repsondents (all roles). If the percentage of Python users among data scientists is higher than average then I'd assume that Python is a distinct skill of data scientists. This is clear from *the video-game like* plot below which I show the difference between these two values between programming languages in colored numbers.

unfold\_moreShow hidden code

Diagram, table

Description automatically generated with medium confidence

For example here we can see that the percentage of Python user among data scientists is more than average by 10 points.

Having explained this important detail, let's go and look at every role ordered by median salary (reward) in more depth!