Working with **messy data**

Insights

The companies with most open positions	Frequently mentioned job positions
Amazon.com: 358	Data scientist: 1261
Ball Aerospace: 187	Research analyst: 318
Microsoft: 137	Research scientist: 274
Google: 134	Data engineer: 186
NYU Langone Health: 77	Data analyst: 131
Fred Hutchinson Cancer Research Center: 70	Machine learning engineer: 122

We decided to focus on the kind of skills are most mentioned in data positions. We looked for:

- **Degrees**
- **Years of Experience**
- **Hard skills**
- **Soft skills**

How hard / easy was the data to work with?

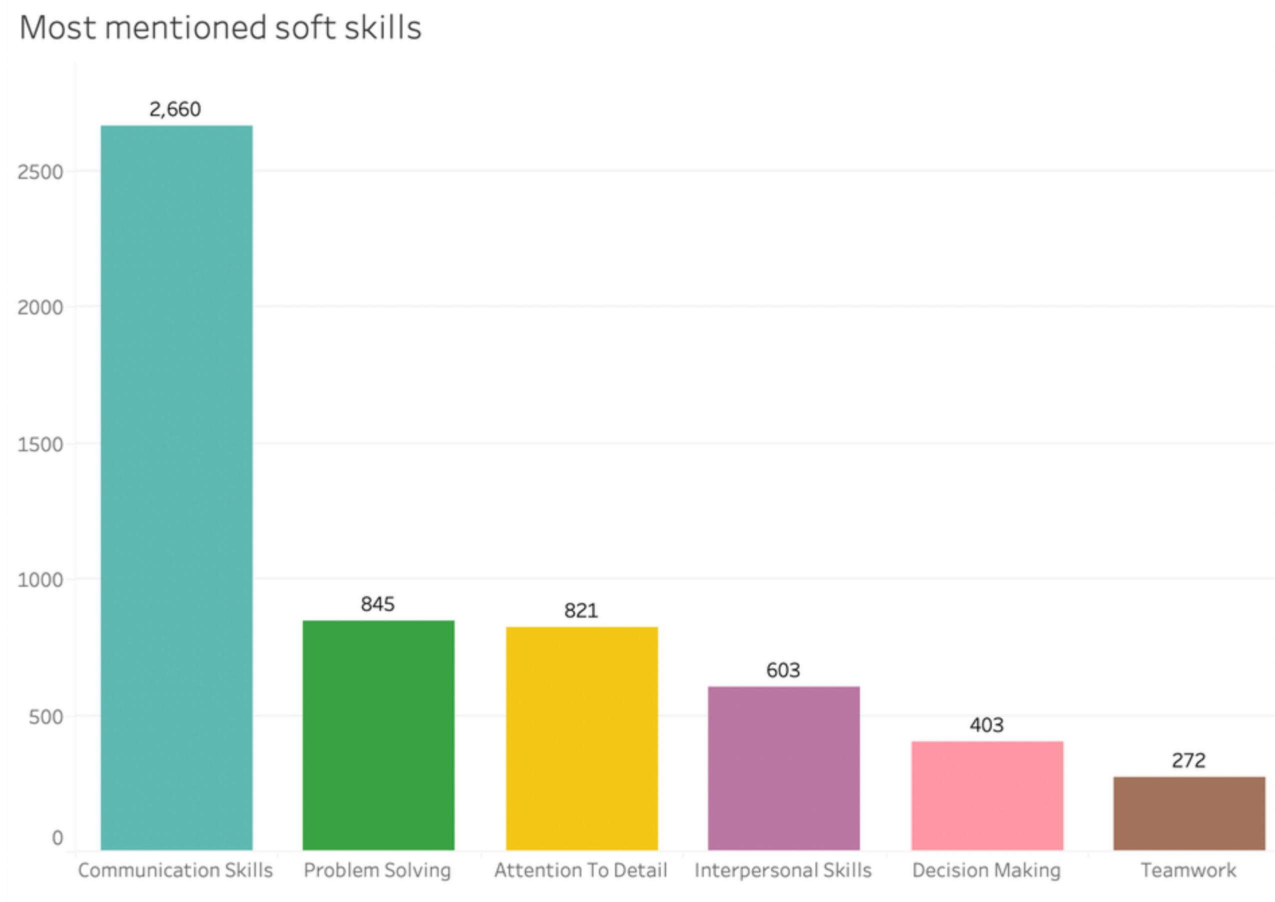
We found some limitations:

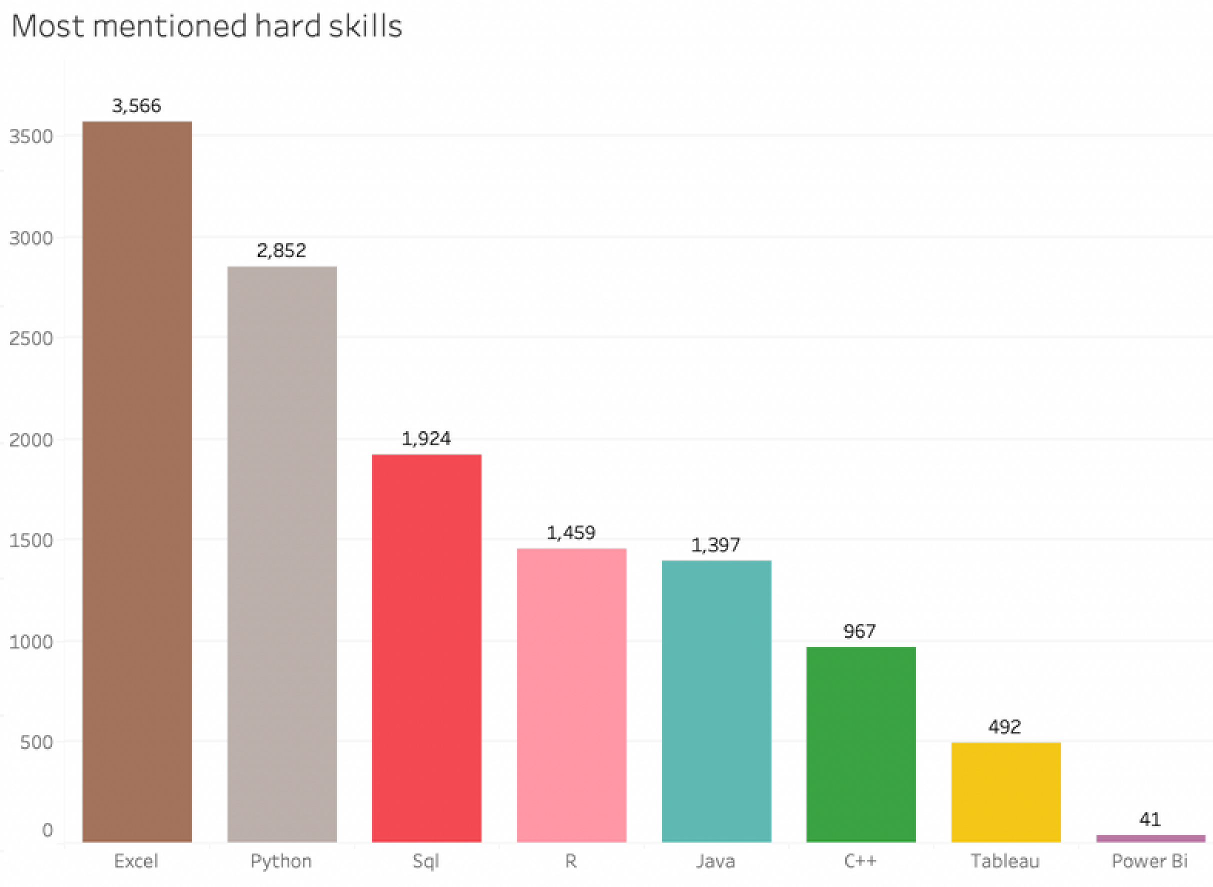
- Unclear sampling method, only one job portal utilised.... is the data representative?
- Some samples include zip codes, others not → more zip code data could allow for more specific geographic targeting down to zip code level. Also, some locations were only mentioned in the position
- The most challenging one - unstructured data in job description:

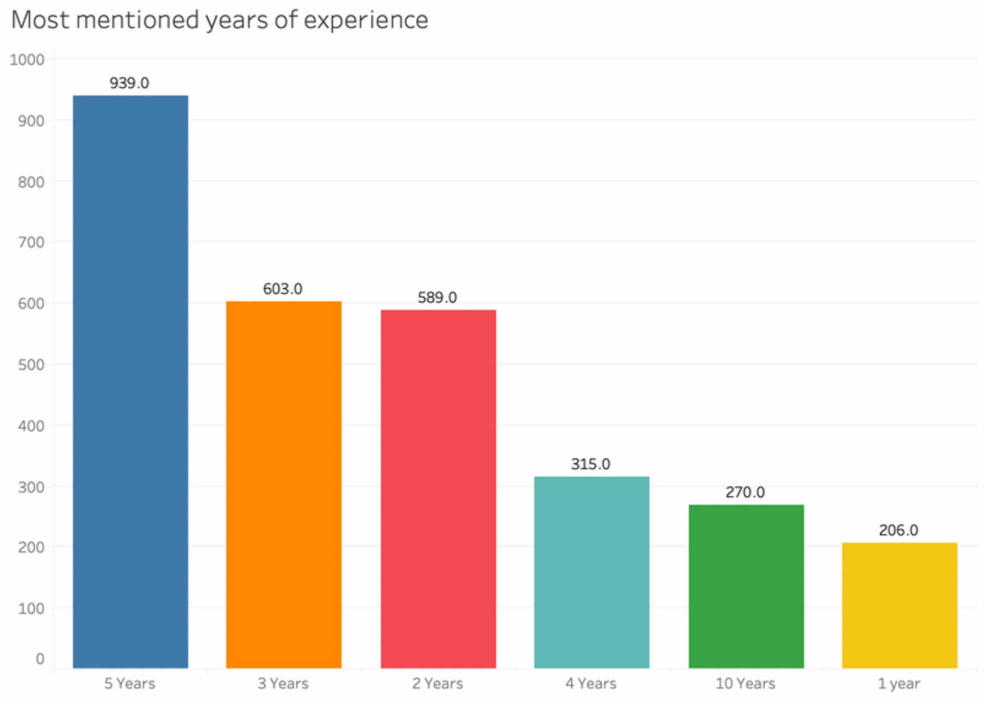
Possible to search for mentions of certain keywords such as skills but hard to determine, for instance, skill level (ex: “advanced python skills” vs. “intermediate python” vs. “some python experience”)

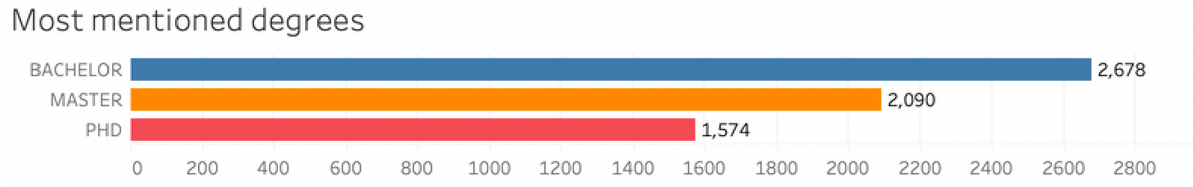
May result in some inaccuracies, for example: while searching for programming language “ r ” or “5 years” experience, may have counted “r & d” and “the company is 5 years old.”

Difficult to account for “advanced” degree


Most mentioned soft skills


Most mentioned hard skills


Most mentioned years of experience


Most mentioned degrees

Asterix group