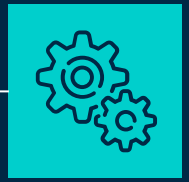
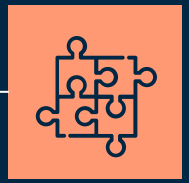
# How to have the best DATA SCIENTIST PROFILE

# TABLE OF CONTENTS



01

PROBLEM & DATA



02

OUR DISCOVERY

03

HOW TO BE A CRACK

# Problem & data

How we managed the data

01

# Problem and data - 1

1 - Exploring the **data**
2 - Dropping rows with all **nulls** values
3 - Cleaning **location** column
4 - Creating **state and city** columns
5 - Cleaning **Description** Column
6 - Creating **category** Column

**Creating category Column**

In [350]:
```
df['position'] = [str(item) for item in df['description']]
```

In [351]:
```
category = []

for position in df['position']:
    if 'data science' in position.lower() or 'machine learning' in position.lower() or 'data scientist' in position.low
        category.append('Data Scientist')

    elif 'analyst' in position.lower() or 'analytics' in position.lower():
        category.append('Data Analyst')

    elif 'artificial intelligence' in position.lower() or 'engineer' in position.lower() or 'ai' in position.lower():
        category.append('Data Engineer')

    else:
        category.append('Other')
```

In [352]:
```
df['category'] = category
```

In [353]:
```
df.head()
```

Out[353]:

| | position | company | description | reviews | location | city | state | category |
|---|---|---|---|---|---|---|---|---|
| 0 | Development Director ALS Therapy Development I... | ALS TDI | Development Director ALS Therapy Development I... | NaN | Atlanta, GA | Atlanta | GA | Data Engineer |
| 1 | Job Description "The road that leads to accomp... | The Hexagon Lavish | Job Description "The road that leads to accomp... | NaN | Atlanta, GA | Atlanta | GA | Data Engineer |
| 2 | Growing company located in the Atlanta GA are... | Xpert Staffing | Growing company located in the Atlanta GA are... | NaN | Atlanta, GA | Atlanta | GA | Data Scientist |
| 3 | DEPARTMENT: Program OperationsPOSITION LOCATIO... | Operation HOPE | DEPARTMENT: Program OperationsPOSITION LOCATIO... | 44.0 | Atlanta, GA | Atlanta | GA | Data Scientist |
| 4 | DESCRIPTION The Emory University Department of... | Emory University | DESCRIPTION The Emory University Department of... | 550.0 | Atlanta, GA | Atlanta | GA | Data Scientist |

# Problem and data - 2

7 - Including **Tools** (skills_list) and **Skills** (soft_skills)
8 - **Making column** for it
9 - Same process for **education** and other data we needed

**skills to includes**

In [354]:
```
# we will do r separately

skills_list = ['sql', 'tableau', 'oracle', 'qliksense',
 'python', 'hadoop', 'perl', 'java', 'c++', 'spark', 'kafka',
 'azure', 'spss', 'sas', 'excel', 'aws', 'big data', 'power bi']

soft_skills = ['statistics', 'mathematics', 'probability', 'machine learning',
'data analysis', 'data visualization', 'modeling', 'deep learning',
'a b testing', 'web scraping', 'neural network', 'interpersonal skills',
'communication skills', 'artificial intelligence']

# crm, salesforce, ux, cloud
```

**Columns for Skills**

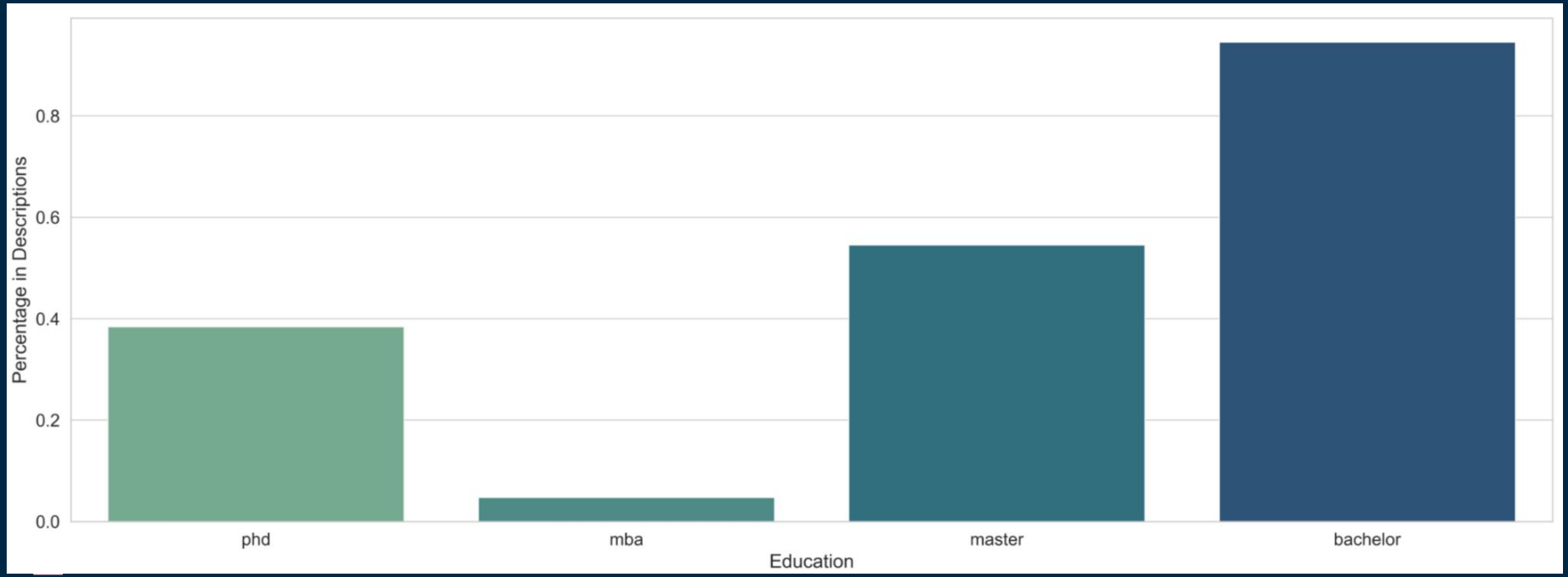
In [355]:
```
for skill in skills_list:

    i = []

    for description in df['description']:
        if skill in description.lower():
            i.append(1)
        else:
            i.append(0)

    df[skill] = i
```

# The perfect profile

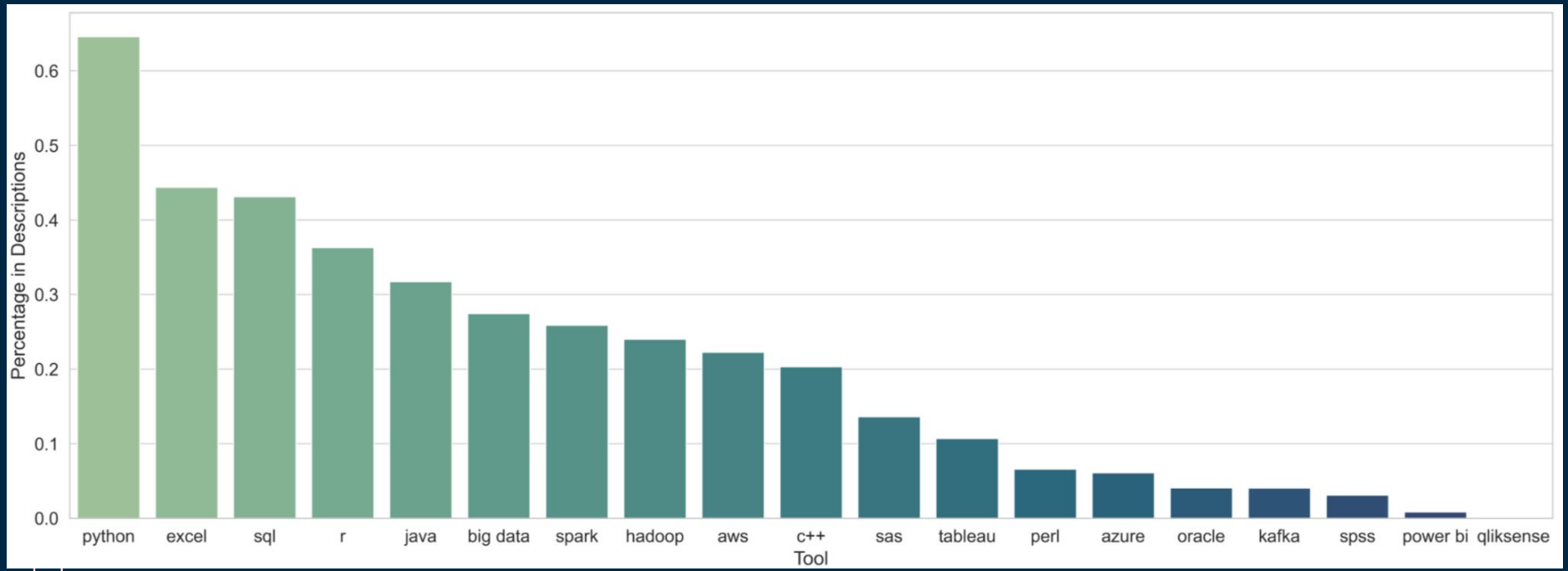
02

How to get the perfect profile ?

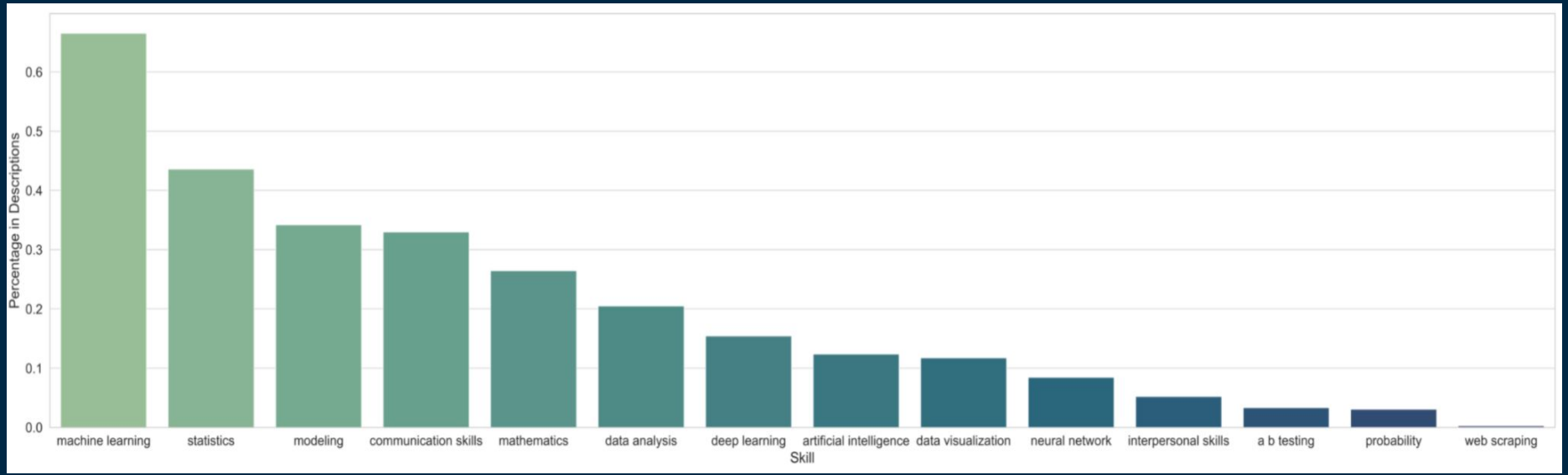
# The perfect profile – Education



# The perfect profile – Software



# The perfect profile – Skills



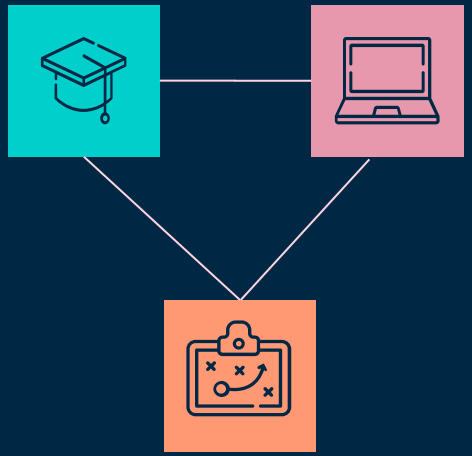
# How to be a crack !

03

How to get the perfect profile ?

# Top profile of a data analyst

**Education**
Bachelor - Master



**Software**
Python - SQL - R

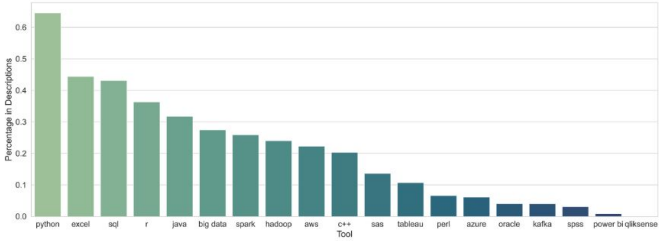
**Soft Skills**
Machine Learning -
Statistics - Modeling

# HOW TO BE A 'CRACK' DATA SCIENTIST

**The Process**: This data needed a lot of cleaning since the description column was particularly messy. We used a variety of string functions to achieve this, removing most of the punctuation to be able to perform searches.

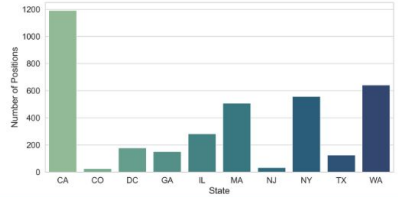
## TOOLS

Python is by far the most requested tool for data science positions. Excel, SQL and R are also useful tools to know, with a variety of other tools often asked for.




## LOCATION

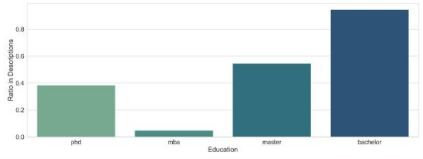
California has the most positions available and may be the best place to search for a job in the US. Washington, New York and Massachusetts are also good locations.




**Limitations**: Not having access to date-time information for this data inhibits our ability to make predictions and leaves us unable to be sure it is representative over any length of time.

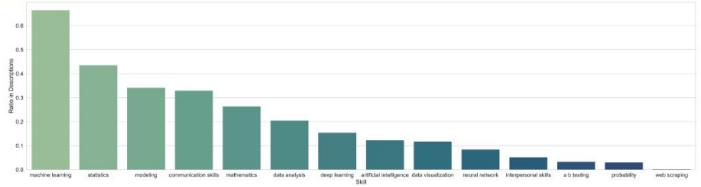
## EDUCATION

Almost 90% of positions ask for a bachelors degree, with over 50% asking for a masters. A PHD is also very useful to have, while an MBA is not commonly needed.




## SKILLS

Machine learning is a highly sought after skill in data science. While, statistics, modelling, communications skills and mathematics are also valuable. Occasionally, more specialised skills such as deep learning and artificial intelligence are requested.



Do you have any questions?

# THANKS