Data Science 2 Group project:
Kien Pham - 2027844
Elisa Martinez Fuentes - 1953368
Sevilay Ozturk -1799704

Data Scientist Salary Analysis

I. Data description

Dataset link: https://drive.google.com/file/d/18Mjg2RwTVr_rd41TLhSQxN6kQ69nAp7V/view

Overview:

The objective of the project is to select a dataset of our interest and apply machine learning models to the dataset to predict helpful outcomes. The chosen dataset is about data scientist job postings since we wanted to gain insights into the data science job market. The dataset was scraped from over 12000 data scientist job postings across the US from Glassdoor.com. The size of the dataset is 12079 rows x 20 columns. Each row in the dataset contains information about a job listing such as job title, company name, location, skill requirement, etc.

In order to develop an ML model, a response variable needs to be determined. In this case, it is the average salary, which is included in each job posting. Kindly note that the average salaries are estimated from Glassdoor's proprietary machine learning model because in reality, not all job listings include this information. According to Glassdoor.com, the salary estimates provide job seekers a likely salary range so they can make more informed job decisions, and help employers recruit informed and quality candidates. With the response variable in mind, we can proceed to work on data preprocessing and exploratory data analysis before building the predictive models.

Column and data type - brief description:

title:
Original title of the job posting - Text

title_normalized:
Normalized job title extracted from the title - Text

company:
Name of the company - Text

location:
Location of the company - Text

rating:
Company rating - Numerical

skill:
Describe the skills required for the job poster - Text

years_of_experience:

Describes the years of experience asked - Categorical

education:

Describes the education required - Text

headquarters:

Location of the company's headquarters - Text

size:

Describes the size of the company - Categorical:

"1 to 50 Employees" "51 to 200 Employees"
"201 to 500 Employees" "501 to 1000 Employees"
"1001 to 5000 Employees" "5001 to 10000 Employees"
"10000+ Employees" "Unknown"

sector:

Describes the sector the company posting the job belongs to - Categorical
There are 24 categories. Examples: Manufacturing, Aerospace & Defense, Business Services, Information Technology, Finance, etc.

industry:

Describes the industry the company posting the job belongs to - Categorical.
There are 103 categories. Examples: Auto Repair & Maintenance, Grocery Stores & Supermarket, Cruise Ships, Wholesale, etc.

company_type:

Describes the company type - Categorical:

"Company - Private" "Company - Public"
"Subsidiary or Business Segment" "Government"
"Nonprofit Organization" "Unknown"
"College / University" "Self-employed"
"School / School District" "Hospital"
"Private Practice / Firm" "Contract"

revenue:

Describes the yearly revenue - Categorical:

"$25 to $50 million (USD)" "$10+ billion (USD)"
"Unknown / Non-Applicable" "$100 to $500 million (USD)"
"$5 to $10 billion (USD)" "$1 to $5 million (USD)"
"$1 to $2 billion (USD)" "$500 million to $1 billion (USD)"
"$10 to $25 million (USD)" "$5 to $10 million (USD)"
"$50 to $100 million (USD)" "$2 to $5 billion (USD)"
"Less than $1 million (USD)"

yearFounded:

Describes the year the company was founded - Numerical

job_link:

Contains the URL to the job posting on Glassdoor - Text

date_posted:

Describes the date the job was posted - Date

Note: Some jobs are set to post in the future

pay_percentile90:

The 90th percentile pay for data scientists - Numerical

pay_percentile50:

The 50th percentile pay for data scientists - Numerical - *Response variable*

pay_percentile10:

The 10th percentile pay for data scientists - Numerical

II. Data preprocessing / Feature Engineering

Since the dataset was in raw format, there were many preprocessing steps that needed to be done to make the data usable. First of all, we dropped the *title* column and kept the *title_normalized* column because *title* contains the subject of the job listing with many redundant words while *title_normalized* has the standardized job title that we need for our analysis. We also dropped the *pay_percentile10* and *pay_percentile90* and kept only the *pay_percentile50* as our response variable since we're only interested in predicting the average salary. *Date_posted, job_link, yearFounded* columns were also dropped due to unneeded information.

title	title_normalized
Data Scientist, GFCCP Model Analytics	Data Scientist
Sr. Data & ML Engineer	Senior Machine Learning Engineer
Customer Transformation Data Scientist - Senio...	Senior Data Scientist

Next, we checked if there were any duplicates or missing values in the dataset. Unfortunately, the dataset has a significant amount of duplicated rows since Glassdoor.com intentionally displays repeated job listings on their platform and due to the way our web scraper works, it collected all the duplicated job postings. Consequently, the size of the dataset got reduced from 12079 to 4167 rows after we performed pandas' *drop_duplicates* method on the dataset. Then, we inspected missing values and found that there were a lot of them as well:

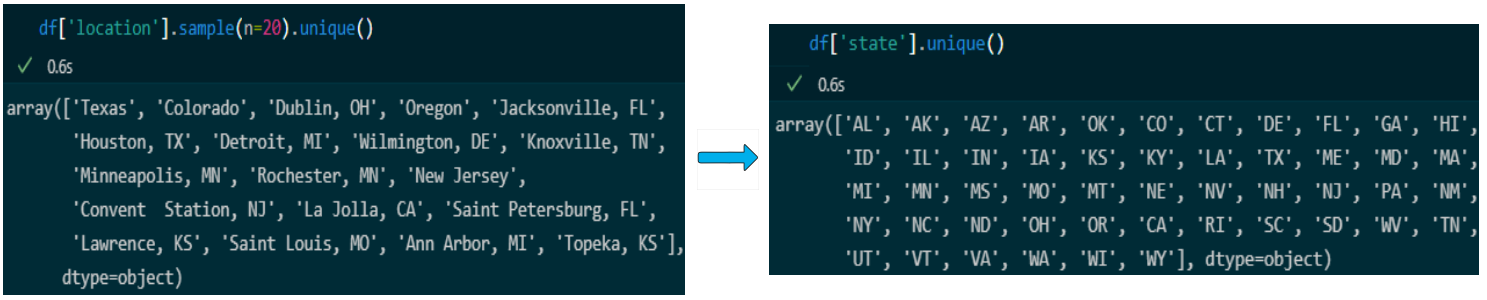
```
title_normalized        0.767939
company                 3.647708
location                0.000000
rating                  3.623710
pay_percentile50       21.310295
skill                   0.191985
years_of_experience     0.191985
education               0.191985
headquarters            3.839693
size                    3.647708
sector                 10.775138
industry               10.799136
company_type            3.647708
revenue                 3.647708
dtype: float64
```

- The response variable *pay_percentile50* had the most missing values that account for 21% of the dataset's size.

- *sector* and *industry* had about 10% of missing values.
- Other columns such as company, headquarters, etc. had less than 4% of missing values. Inspecting further *rating* column, we also found that there were several rows with a value of -0.1, which stands for not available.

Since these missing values came with the data itself and there was no good imputation method available, we had no choice but to drop these rows. The final size of the dataset after dropping duplicates and missing values is 2873 rows.

Looking at the *location* column, we see that the locations are not normalized since some locations have the city name while others do not. Thus, we wrote a function to extract the state from the location then convert it to US state abbreviations and store it into a new variable called *state.* Moreover, we created a new binary variable *at_headquarters* that takes value 1 if the job location is the same as the company headquarters, otherwise 0. This variable could be used to determine whether working at headquarters affects the pay or not. Subsequently, the *location* and *headquarters* columns were dropped since they had become obsolete.

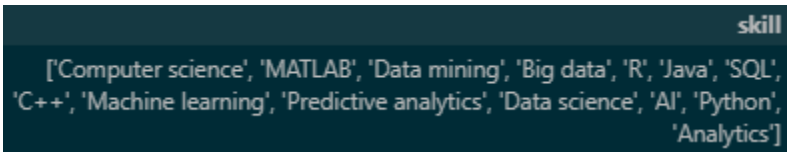


We proceeded to clean the *years_of_experience* column. The values of this column were stored as strings of a list in the data frame such as "['2']", ['Under 1'], etc. So we created a function to extract the years of experience from the value and convert it into numbers. If the value is an empty list, we assume that the job does not require any experience, and assign 0 to the value. If the years of experience is under 1 year, we assign 0.5 to the value. And any jobs that require more than 11 years of experience will get assigned a value of 11.

Moving on to the *education* column, the values were also stored as strings with five education levels: high school or GED, associate, bachelor's, master's, and doctoral. For each degree level, we accordingly

created a binary column that takes 1 if the job requires such level, otherwise 0. The *education* column was dropped afterward.

Finally, we looked into the *skill* column and saw that the values were stored neatly as lists of skills separated by a comma. Then we ran Python's Counter function to find the most sought-after skills in the dataset and use this information for the exploratory data analysis part. However, we haven't performed any transformation on this column yet since there are too many skills existing in the column and we are not sure what other insights we can extract for the model. We will come back to this column when we start building the model later on. Moreover, all other columns such as companies' *size*, *industry*, *sector*, *type*, etc. were left intact because they are categorical columns. Below is the summary of the cleaned dataset that we will use for exploratory data analysis.




```
df.info(verbose=False)
df.columns
```

✓ 0.7s Python

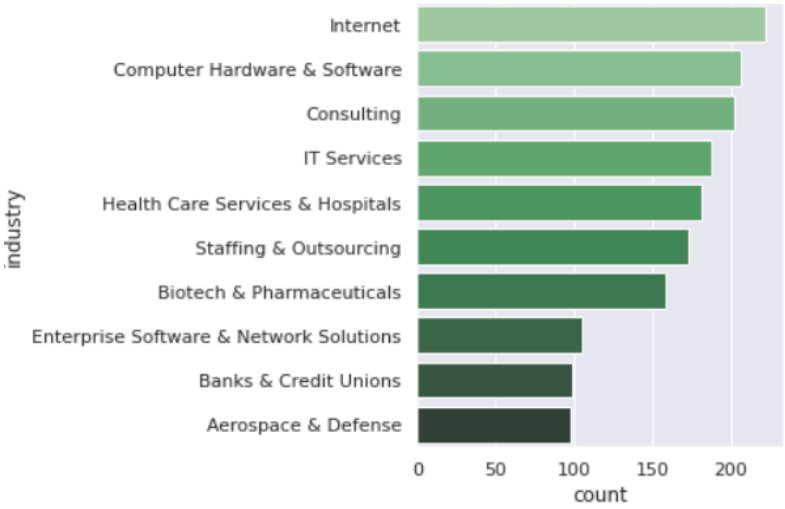
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2873 entries, 0 to 4166
Columns: 18 entries, title_normalized to doctoral
dtypes: float64(3), int64(6), object(9)
memory usage: 426.5+ KB
```

```
Index(['title_normalized', 'company', 'rating', 'pay_percentile50', 'skill', 'years_of_experience', 'size',
'sector', 'industry', 'company_type', 'revenue', 'state', 'at_headquarters', 'highschool/GED', 'associate',
'bachelor', 'master', 'doctoral'], dtype='object')
```

III. **Exploratory Data Analysis**

- Which industries are in demand?

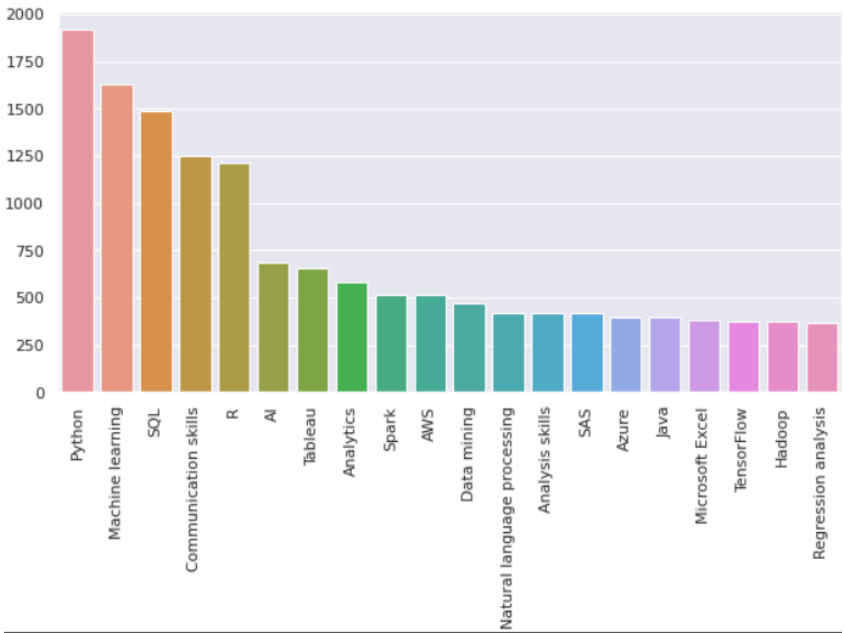
➢ As future data scientists, we wanted to know which industries are in demand for data science. We discovered that the internet industry was in demand the most followed by computer hardware & software with a little distinction from consulting.



- What is the average data scientist's salary?

➢ The average data scientist salary is around 95k although the range is very high with 160k. This tells us that there are still low-paying data science jobs but the median tells that on average the salary is not too bad, especially the high outliers that are above the maximum salary suggest data scientists can be important assets to the companies.

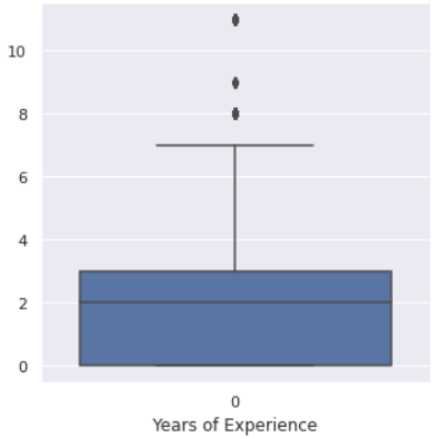
- Which skills are in demand?



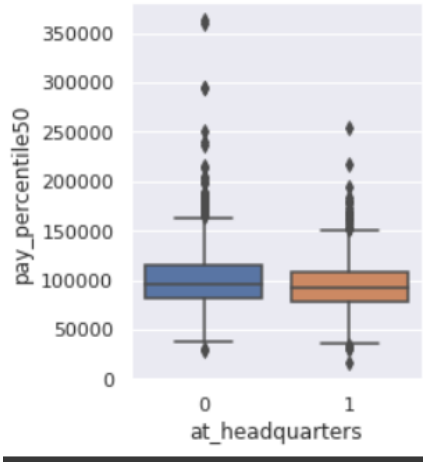
➢ Python and Machine learning turned out to be essential skills for data scientists, followed by SQL and communication skills. This tells us that in the data science industry, an important skill like R could be less demanding than communication skills, and good data scientists should be able to explain their work.

- What is the average year of experience required for data science positions?

➢ We found out that the average years of experience required is 2 years, although the data suggest that this range could be up to 10 years which tells a lot about the expected qualities of a data scientist, some companies seem to think it takes people to have 4-10 years of experience to perform and excel the skills required of the job.



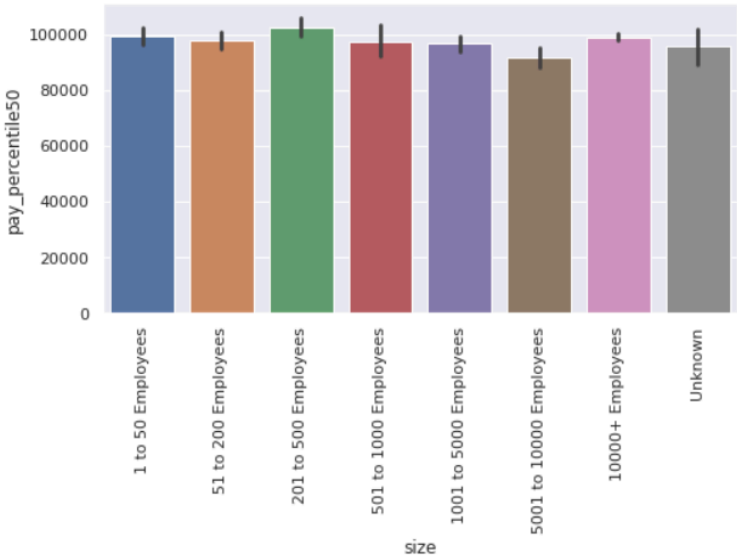
- Does working at the headquarters affect salary?



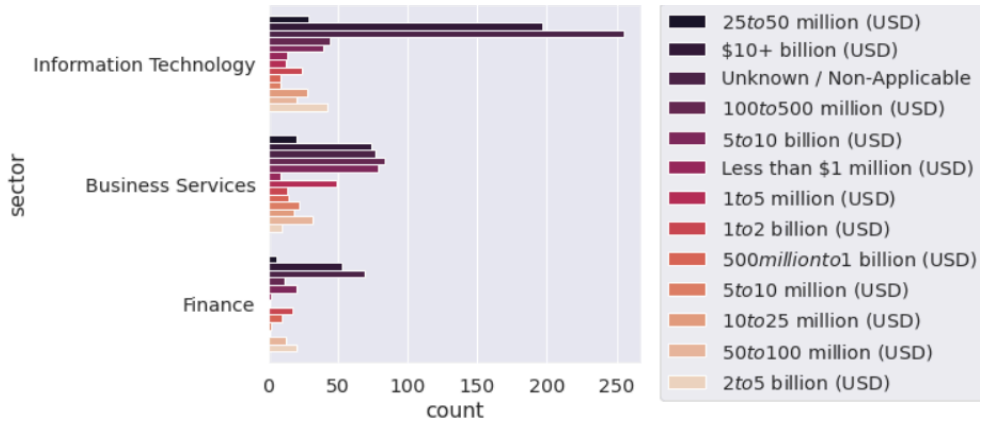
➢ With 1 indicating yes and 0 indicating no, we see that the average pay for people who work at the headquarters is almost the same as the ones who do not. Contrary to our assumption, outliers for people who work at the headquarters seem to be lower than that of people who do not. This could be due to the fact that work has become more flexible since the pandemic with lots of online jobs that pay very well.

- Do bigger companies pay better?

➢ We initially assumed that bigger companies would offer better compensations but we were mistaken as the data shows that there is no correlation between size and pay. We can attribute this to startups as they receive capital from investors they can offer higher salaries to developers and data scientists and compete with the bigger companies by paying market-rate salaries. This, of course, is not the norm.

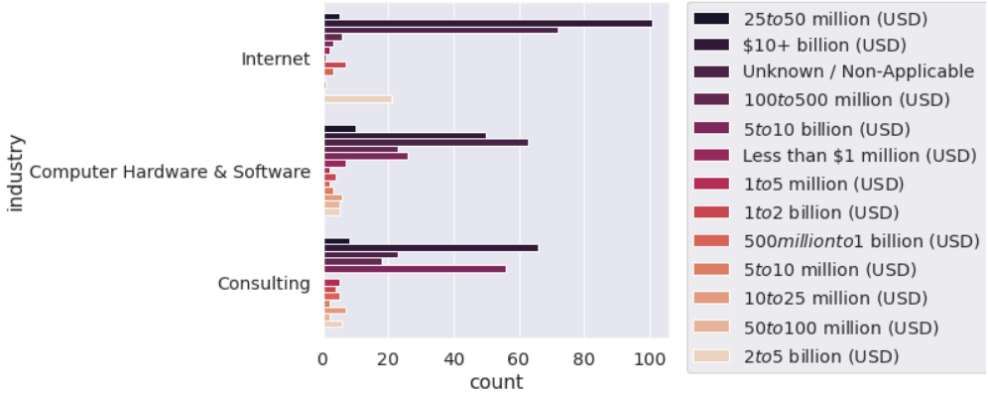


- Which sector makes the most money?



➢ The top 3 sectors that make the most money are IT, Business Services, and Finance. IT on the other hand passes its opponents by billions of dollars, although it is also the industry with the most unknown revenue.

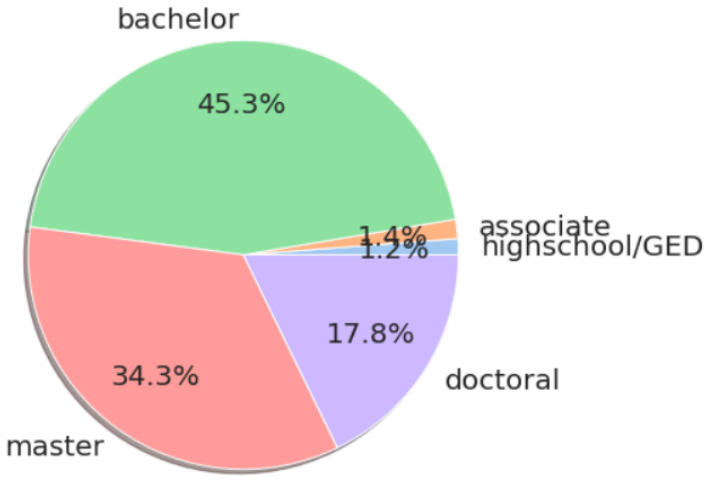
- Which industry makes the most money?



➢ In terms of industry, the top 3 that make the most money are Internet, Computer Hardware & Software, and Consulting industries in order.

- What's the required education on average and education distribution?

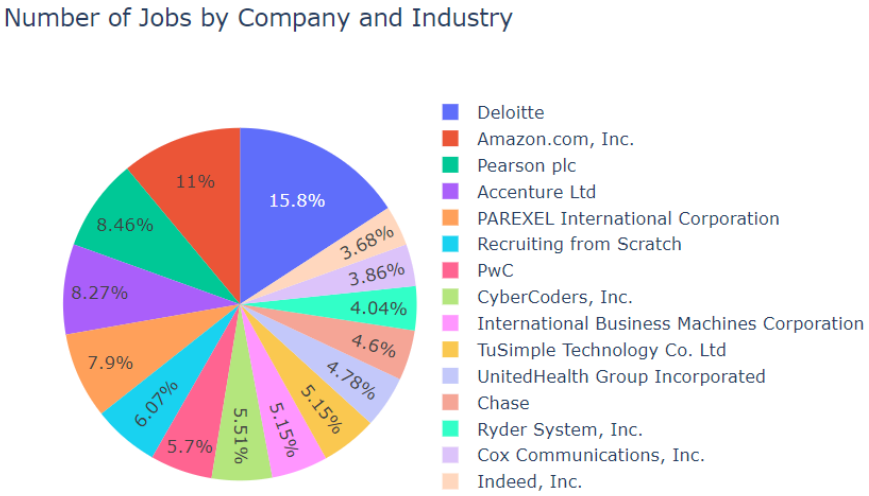
➢ Education distribution tells us that the majority of the data scientists hold bachelor's degrees; therefore, a master's or a doctoral degree is not required, but a master's can be helpful since it follows bachelor's by very close.



- Does years of experience affect rating?

➤ It seems like with less than a year and 3 years of experience they have a wide range of ratings, on the other hand after 4 years of experience the rating range seems to go lower and more consistent. This concludes that the years of experience affect the rating.

- Top 15 companies with a higher number of jobs

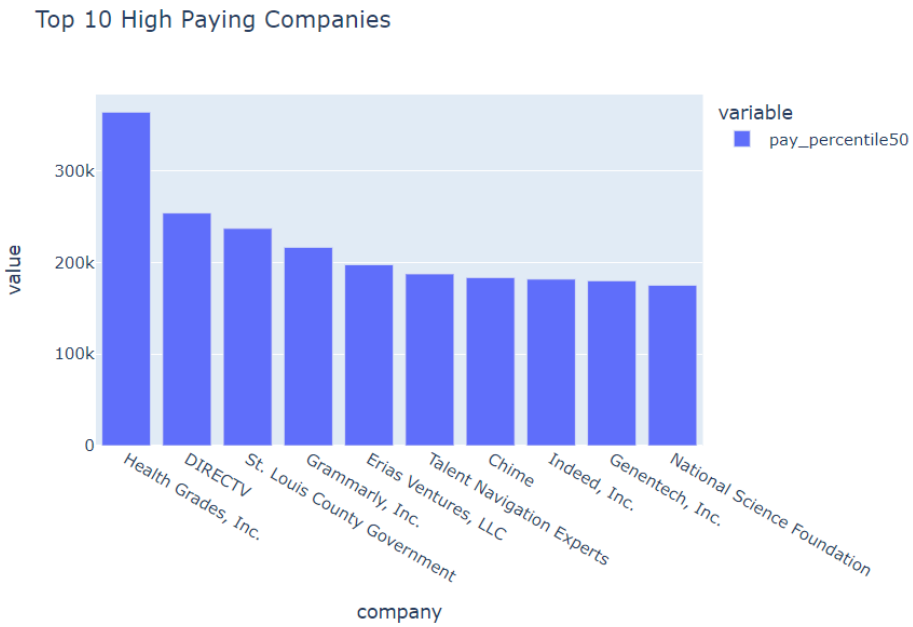


➤ We wanted to know what companies post most jobs in total and found out that Deloitte, a UK private company that offers consulting, advisory, and tax services has the most jobs in the market at the moment followed by Amazon.com, Inc, and Pearson, also a UK based company which is one of the largest publisher companies in the world.

- What are the worst-paying companies?

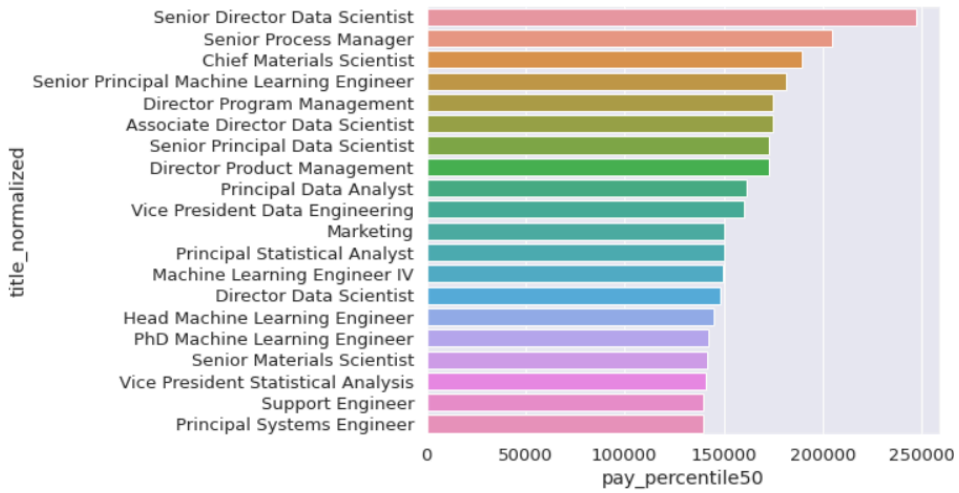
➤ We were able to see that universities tend to pay less and also jobs that are located in Tennessee.

- What are the best-paying companies?



➤ Interestingly, we were able to see that the best paying company was Health Grades, Inc which specializes in providing information about physicians, hospitals, and healthcare providers, which made sense to us. As they rely heavily on data scientists and statistics as their main service. The next company is DIRECTV, a digital satellite service. The third company is Grammarly, Inc which provides natural language processing algorithms to aid writers online.

- What job titles make the most money?



➤ To answer this question we looked at several bar plots and found that the best-paying jobs are “Senior Director Data Scientists”, “Senior Process Manager”, “Chief Materials Scientist”, and “Senior Principal Machine Learning Engineer.” All of these titles have two things in common, they are more senior positions, and they are more management and leadership positions. At the same time, we can see that Research Analyst III and Research Fellow are the least paying jobs, which is unfortunate because it provides a high barrier to entry for several aspiring scientists, as the positions are very competitive and the pay is low compared to entry-level junior level software engineers, for example.