

Data Scientist Salary Study

Group Project by Yuheng Fan, Yizhang Huang, Siqi Wang, Sophia Zhang

Project Overview

In this project, we are interested to find out the requirements and skills needed and salary of data scientists job of different industry on the job market. Our project overview is as follows:

Introduction and Motivation: Why we are interested in this particular topic? What can we learn from this project

Part 1: Correlation Among Programming languages and Job Titles

Part 2: Correlation Among Educational Degree, Job Type, and Salary

Part 3: Correlation Among Job Amount, Salary, and Industry

Part 4: Correlation Among Company Rating, Salary, and Location.

Conclusion and Considerations: Does our analysis provide useful insights for our motivation? which of the factors have impacts on the salary? What advices can we give for statistics students who want to work in similar fields?

Introduction and Motivation

In this 21 century, people are all familiar with the term "big data", which refers to large amount of data that exceeds the processing capacity of traditional database systems. With the rapid development of technology, more and more daily life information transforming into digital data, while increasing demand for data analysis in various of industries makes it particularly important. Therefore, data science is developed.

Data science is an interdisciplinary field - combining subjects such as computer science, mathematics, statistics, and many other fields. It encompasses machine learning (algorithms that use statistics to find patterns in large amounts of data), data analysis (including examining the data, cleaning/validating it, and transforming it to ensure it is modeled in an efficient manner, thereby help solve business problems) and data engineering (focusing on acquiring data, preparing data, and processing data). Research in this field has been increasing exponentially for the recent decays, as well as the application of data science. It is worth noting that data science is currently associated with almost every modern industries.

Therefore, as students majoring in statistics, we would like to plan ahead to our career path, so we determine to conduct a analysis project that allows us to further explore the data related job. We are interested in not only the requirements listed by the company and the skills/qualifications that other more sophisticated data scientists use, so that we can plan ahead to acquire the skills that we need for the job of interest. But also we would like to learn about the salary level of different industry so that we can have a clear goal of the industry we want to work in. In this project, we will analyze the data using various plots.

Dataset: Data Scientist Salary (Kaggle)

In order to fully explore the factors that correlate to salary for data science related jobs, we chose to explore the [Data Scientist Salary](#) dataset from Kaggle. The dataset is scraped from the Glassdoor website, with 42 variables (such as average salary, job description, rating of the company, etc.).

```
In [ ]: # import salary dataset
import pandas as pd

data = pd.read_csv("content/data_salary_cleaned_2021.csv")
data.head(6)
```

	index	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	...	tensor	h
0	0	Data Scientist (Glassdoor est.)	53K - 91K	Data Scientistn.location: Albuquerque, NM;Edu...	3.8	Tecolote Research	Albuquerque, NM	Goleta, CA	501 - 1000	1973	...	0	
1	1	Healthcare Data Scientist (Glassdoor est.)	63K - 90K	What You'll Do:You'll General Summary:You'll...	3.4	University of Maryland Medical System	Linthicum, MD	Baltimore, MD	1000+	1984	...	0	
2	2	Data Scientist (Glassdoor est.)	80K - 99K	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4	Cleanwater, FL	Cleanwater, FL	501 - 1000	2010	...	0	
3	3	Data Scientist (Glassdoor est.)	56K - 97K	*Organization and Job ID**Job ID: 310709nrv...	3.8	PNNL	Richland, WA	Richland, WA	1001 - 5000	1965	...	0	
4	4	Data Scientist (Glassdoor est.)	86K - 143K	Data ScientistAffinity Solutions/Marketing...	2.9	Affinity Solutions	New York, NY	New York, NY	51 - 200	1998	...	0	
5	5	Data Scientist (Glassdoor est.)	71K - 119K	CyruOne is seeking a talented Data Scientist ...	3.4	CyruOne	Dallas, TX	Dallas, TX	201 - 500	2000	...	0	

6 rows × 42 columns

For our project, we will mainly focus on the the relationship among salary and the following areas:

- Company ratings
- Programming languages: Python, Java, SQL ...
- Job types: scientist, analyst, engineer, and other ...
- Industries: Media, Healthcare, Technology ...
- Geographic locations

Part 1: Correlation Among Programming languages and Job Titles

Context

Proficiency in different programming languages is a requirement for data scientist. There are so many programming languages that are helpful purpose of analyzing. People need to determine which programming languages are most used in the workplace. The following code provides some insights about different programming languages.

Methodology

```
In [ ]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# select the columns with skills
skills = data[['python', 'spark', 'aws', 'excel', 'sql', 'sas', 'keras', 'pytorch', 'scikits', 'tensorflow', 'hadoop', 'tableau', 'bi', 'flink', 'mongo', 'google_an']]

# Create a histogram to show the counts of programming languages
lst1=[]
for i in skills:
    lst1.append(skills[i].value_counts())
skills_sum=pd.DataFrame(lst1)
new_skills=skills_sum[1]
ax=pd.DataFrame(new_skills)
sns.set(font_scale=2)
plt.gcf().set_size_inches(25, 8)
ax=sns.barplot(x = a.index, y = a[1])
ax.set(title = "Bar plot of Programming language", xlabel = "Programming language", ylabel = "Counts")
ax
```

```
In [ ]: # Create a dataframe of counts of 16 programming languages
a
```

The question that we try to determine is that which programming language are most used and least used. This question can provide us some insights about which type of programming languages has relationship with salary. We create a dataframe to shows the counts of each programming language in our dataset. The dataset collect the information from 742 data scientist. There are total of 16 programming languages that we try to explore. They are the following: Python, spark,aws,excel,etc. It is clear to see that Python, excel, sql are the top 3 programming languages among data scientist and data related jobs. Having a dataframe is not that obvious to indicate the result. Thus, we need to make the data more visualize. A barplot was made to visually present the data. In the barplot, python,excel, and sql have a similar counts which the bars have same length.Next, we check back at the dataframe. We can clearly identify that how many people used python language. 392 data scientist report that they used python language Python programming languages is the most used programming language. From the bar plot, we can see that flink programming languages are the least used in the workplace according to our dataset

```
In [ ]: # Create a graph to show the counts of programming languages by job titles
lst1=[]
for i in skills:
    lst1.append((i, skills_by_title["data.groupby('job_title_sim').sum()"][lst1]
skills_by_title=skills_by_title.drop("na")
skills_by_title1.style.background_gradient(cmap='Reds', axis='columns')
```

To dig deeper about the question, we create a dataframe to shows the programming languages by job titles. This dataframe help us better understand what programming languages are being used for each job title. In our dataset, our job title include data scientist, analyst, data modeler, and so on. In addition, we use red color to show the relationship. The deeper the red color means the programming being used most often. From the dataframe, it is clear to see that data scientist favor python programming language the most. This implicate that python is the most important skill for data scientist. It looks like Python and SQL are the top skills in demand for Data Scientists, Data Engineers, and Machine Learning Engineers because they have a relative deep red color compared to other languages One interesting finding is that Excel is also a good skill to have since many jobs need that skill. In conclusion, it is imperative to process the python skills if you want to become a data scientist in the future.

Part 2: Correlation Among Educational Degree, Job Type, and Salary

Context

Besides programming languages, educational degree is also a big part in job requirements. When mentioning data science, people usually have the stereotype that for these computer science related jobs, PhD is a necessary requirement to apply. As undergraduate students that are about to finish our Bachelor's Degree, it is an important time point that we should decide whether pursuing further degree (such as Master or PhD) or not. Therefore, we would like to explore the distribution of degree requirement, as well as the relationship with salaries, to have a better sense of the real-world situation.

After finishing our educational degree, we also wonder the salary distribution between different job types. Therefore, an analysis between average salary and job types is also conducted.

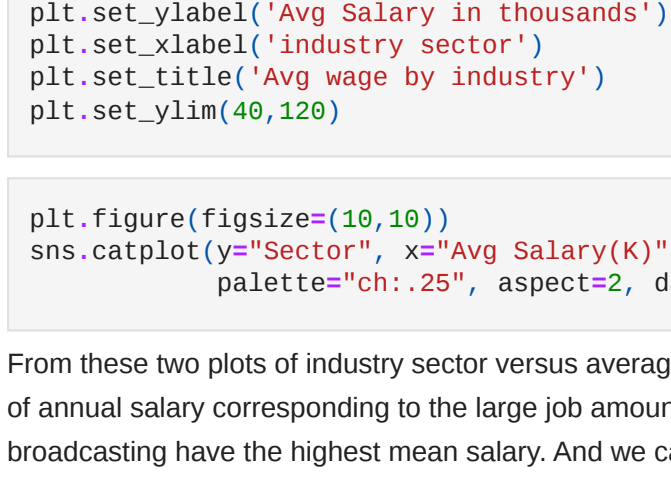
Methodology

In this dataset, we noticed that many job descriptions do not have specific degree requirements. In order to have a better data integrity, we replaced those 'na' with 'B', which stands for Bachelor's Degree or equivalent experience. We categorized the data into three groups, B (Bachelor's Degree or equivalent), M (Master's Degree or equivalent), and P (PhD or equivalent), and visualized the frequencies in a bar chart.

```
In [ ]: import seaborn as sns
import matplotlib.pyplot as plt

# copy the data for degree v.s. job type analysis
degree = data.copy()[['Degree', 'Avg Salary(K)']]
degree['degree'] = degree['degree'].str.replace('na', 'B')
# visualize the data using bar plot
sns.catplot(x="Degree", kind="count",
            order=['B', 'M', 'P'], palette="ch:25", data=degree)
plt.title("Degree Requirement Count", size=16)
```

Text(0.5, 1.0, 'Degree Requirement Count')

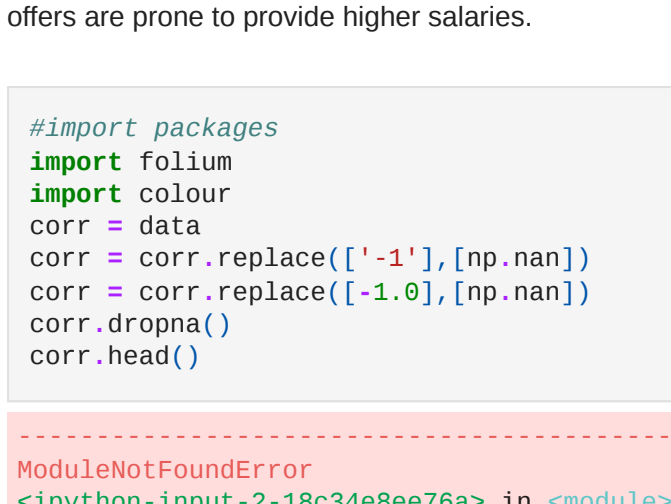


The result showed that more than 50% of the jobs only require Bachelor's Degree. There are about double the amount of jobs require for Master than for PhD. Overall, we can see that there is a decreasing trend of jobs required for higher educational degree. Therefore, it seems that a Bachelor's Degree is sufficient for finding a job in data science area.

Moreover, we conduct further analysis of the salary distribution among different degree groups. We visualize the result using a violin plot, which is more informative and shows the full distribution comparing to a plain box plot.

```
In [ ]: # visualize the data using violin plot
sns.catplot(x="Avg Salary(K)", y="Degree", kind="violin",
            order=['B', 'M', 'P'], palette="ch:25", data=degree)
plt.title("Avg Salary by Degree", size=16)
```

Text(0.5, 1.0, 'Avg Salary by Degree')



We can see that the median for average salaries demonstrate a positive correlation with the required degree. The plot indicates that PhD has the highest median at about 125K. However, there is an interesting finding that for PhD salary distribution, we have a U shape near the median average salary. This suggests that there are more jobs have either higher or lower salaries, instead of being in the middle. Our guessing is that for PhD students, they either choose to continue devoting themselves in research, or they might get hire in a company for higher positions (such as manager level). This might cause the gap in the salary distribution.

Based on these two results, we can conclude that most jobs indeed only require for a Bachelor's Degree. However, if you have a high expectation on the salary, maybe gaining a higher degree is a better choice.

Next, for career-oriented purpose, we also want to explore the relationship between different job types and the average salaries. Here we apply a boxen plot to visualize the distribution.

```
In [ ]: # copy the dataset for salary v.s. job type analysis
salary = data.copy()[['job_title_sim', 'Avg Salary(K)']]
# visualize the data using boxen plot
plt.figure(figsize=(16,18))
sns.catplot(y="job_title_sim", x="Avg Salary(K)", kind="boxen",
            palette="ch:25", aspect=2, data=salary)
plt.ylabel("Job Type")
plt.title("Avg Salary by Job Type", size=16)
```

Text(0.5, 1.0, 'Avg Salary by Job Type')

<Figure size 720x720 with 0 Axes>



From the plot above, we can see that most of the job types have its second and third quartile fall in the range between 50K and 100K. Data scientist, machine learning engineer and director are having higher salaries, with second and third quartile fall in the range between 100K and 150K, almost double the amount. Data scientist are having the best performance among all the job types, with more jobs distribute in the range of 150K and above, and also the highest salary. However, it is also having the widest range, with job salary falls below 25K. Overall, data scientist is a good choice.

Part 3: Correlation Among Job Amount, Salary, and Industry

Context


In this section, we will be examine how the data scientist salary differs base on the different industries. First we will be looking at the number of jobs provided of each industries. Since the amount of jobs of different industry sector recorded is different, we will be comparing the mean salary. And the following questions will be answered: Which industry sector offers the most job? Which industry sector has the highest mean salary? What's an example company that represents this industry sector?

Methodology

We will use bar plot and box plot to visually examine the data.

```
In [ ]: df=data.loc[:,['Sector', 'Avg Salary(K)', 'Job Location', 'job_title_sim', 'Degree']]

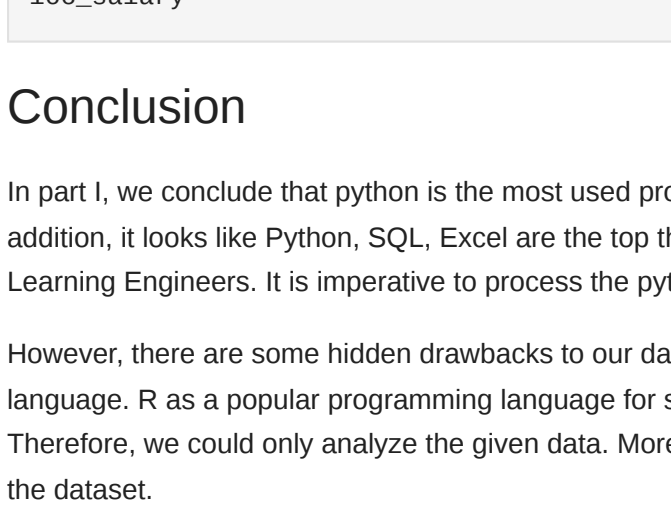
plt=df[['Sector']].value_counts().plot(kind='bar')
plt.figure()
plt.rcParams.update({'font.size': 22})
plt.set_ylabel('Avg Salary in thousands')
plt.set_xlabel('Industry sector')
plt.set_title('Avg wage by industry')
plt.set_ylim(40,120)
```



Moreover, as we can see from this histogram of job amounts vs industry sector, the top three industries that offer the most jobs are information technology firms, companies that operate in the internet or computer software business. Biotech firms such as pfizer, Business Service firms that mainly operate in marketing and advertising. These three industry sectors offer the majority of statistic related jobs on the market.

```
In [ ]: plt=df.groupby(['Sector']).mean().sort_values(by=['Avg Salary(K)'],ascending=False).plot(kind='bar')
plt.rcParams.update({'font.size': 22})
plt.set_ylabel('Avg Salary in thousands')
plt.set_xlabel('Industry sector')
plt.set_title('Avg wage by industry')
plt.set_ylim(40,120)
```

Text(0.5, 1.0, 'Avg Salary by Job Type')



From these two plots of industry sector versus average salary, we can see that information technology firms do have a widest range of annual salary corresponding to the large job amount. However, media companies that operate in video games and tv broadcasting have the highest mean salary. And we can also see that data scientists working in art industry earns the least.

Part 4: Correlation Among Company Rating, Salary, and Location.

Context:

Data scientists, analysts, engineers are known for its versatility, opening doors from healthcare to finances. Nevertheless, similar to other specializing careers, data science's job positions occurs in clusters. Regions such as Silicon Valley has more job offerings than others. In this portion, we are inspecting the correlation among rating, salary, and location of the companies. Is there a correlation between company rating and average salary? With cautious premisses, is money an important factor for job satisfactory? Does location of the job offering influence the salary? Does location influence the company rating?

Starting with the question if there's a correlation between the average salary and company. Then, further investigate the correlation between location and salary, and location and company rating.

Methodology:

1. To understand wether or not there's an correlation between salary and rating, we will perform an spearsman correlation calculation and correlation plot.
2. For the second and third questions, we will first acquire the location from location variable, then use an geocoding API(MapQuest) to obtain the latitude and longitude of the location.
3. Create separate dataframes for location vs. rating, and location vs. salary, attaching the longitude and latitude in separate dataframes. Then uses the package folium to generate interactive plots.
4. Change the color of the circles on map to demonstrate the different wage and rating intervals, using if and elif statements. A darker green circle indicate a higher salary or rating.

Observation and comments

From the correlation plot we observe no linear or other correlation between the company rating and salary. The spearsman correlation coefficient is only 0.13699152484765198. This is result is not unexpected. The data source suggest rating to be a more comprehensive evaluation of the company. Salary, despite its crucial role, cannot be the sole determinate of the company rating.

The interactive map for location vs. rating is not consistent with my initial predications. The rating map suggests a conglomeration of lower rating companies around New York, while other metropolises such as Los Angeles, Chicago, Silicon Valley, Seattles does not experience this inverse correlation.

On the other hand, the interactive map for location vs. salary is accordant with most people's expectation. In the urban regions, job offers are prone to provide higher salaries.

```
In [ ]: #import packages
import folium
import colour

corr = data
corr = corr.replace(['-1'],[np.nan])
corr = corr.replace(['-1.0'],[np.nan])
corr.dropna()
corr.head()
```

```
ModuleNotFoundError: Traceback (most recent call last)
<ipython-input-2-18c348ee76a> in <module>()
      1 import packages
      2 import folium
----> 3 import colour
      4 corr = data
      5 corr = corr.replace(['-1'],[np.nan])

ModuleNotFoundError: No module named 'colour'
```

NOTE: If your import is failing due to a missing package, you can manually install dependencies using either !pip or !apt.

To view examples of installing some common dependencies, click the "Open Examples" button below.

```
In [ ]: def select_marker_color(row):
    if float(row['Avg Salary(K)']) < 2.0:
        return '#E6F5E9'
    elif float(row['Rating']) < 2.5:
        return '#C8E6C9'
    elif float(row['Rating']) < 3.0:
        return '#A5D6A7'
    elif float(row['Rating']) < 3.5:
        return '#B1C784'
    elif float(row['Rating']) < 3.8:
        return '#68BB6A'
    elif float(row['Rating']) < 4.0:
        return '#4CAF50'
    elif float(row['Rating']) < 4.3:
        return '#43A047'
    elif float(row['Rating']) < 4.5:
        return '#38A83C'
    elif float(row['Rating']) < 4.8:
        return '#2E7D32'
    elif float(row['Rating']) < 5.0:
        return '#1B5E20'
```

```
In [ ]: locR['color'] = locR.apply(select_marker_color, axis = 1)
loc_rating = folium.Map(
    location = [39, -100],
    zoom_start = 4.4
)
loc_rating
for _, loc in locR.iterrows():
    folium.CircleMarker(
        location = [loc['lat'], loc['lng']],
        popup = loc['Rating'],
        tooltip = loc['Rating'],
        radius = 8,
        color = loc['color'],
        fill_color = loc['color']
    ).add_to(loc_rating)

loc_rating
```

```
In [ ]: #location vs. salary dataframe from API
locS = corr.loc[:,['Avg Salary(K)', 'Location']]

#API call and get the latitude & longitude values
for i, row in locS.iterrows():
    address = locS.at[i, 'Location']
    parameters = {
        "key": "f1xh0A8XusBgzhqJai01W4RuVfYVZrN",
        "location": address
    }
    response = requests.get("http://www.mapquestapi.com/geocoding/v1/address", params = parameters)
    cd = json.loads(response.text)['results']

    coord = coord[0]['locations'][0]['latLng']['lat']
    lng = cd[0]['locations'][0]['latLng']['lng']

    locS.at[i, 'lat'] = lat
    locS.at[i, 'lng'] = lng
```

```
In [ ]: def select_marker_color_s(row):
    if float(row['Avg Salary(K)']) < 50.0:
        return '#E6F5E9'
    elif float(row['Avg Salary(K)']) < 70.0:
        return '#C8E6C9'
    elif float(row['Avg Salary(K)']) < 90.0:
        return '#A5D6A7'
    elif float(row['Avg Salary(K)']) < 105.0:
        return '#B1C784'
    elif float(row['Avg Salary(K)']) < 120.0:
        return '#68BB6A'
    elif float(row['Avg Salary(K)']) < 135.0:
        return '#4CAF50'
    elif float(row['Avg Salary(K)']) < 150.0:
        return '#43A047'
    elif float(row['Avg Salary(K)']) < 180.0:
        return '#38A83C'
    elif float(row['Avg Salary(K)']) < 200.0:
        return '#2E7D32'
    elif float(row['Avg Salary(K)']) < 270.0:
        return '#1B5E20'
```

```
In [ ]: locS['color'] = locS.apply(select_marker_color_s, axis = 1)
loc_salary = folium.Map(
    location = [39, -100],
    zoom_start = 4.4
)
loc_salary
for _, loc in locS.iterrows():
    folium.CircleMarker(
        location = [loc['lat'], loc['lng']],
        popup = loc['Avg Salary(K)'],
        tooltip = loc['Avg Salary(K)'],
        radius = 3,
        color = loc['color'],
        fill_color = loc['color']
    ).add_to(loc_salary)

loc_salary
```

Interactive Map: Location vs. Salary

```
In [ ]: #location vs. salary dataframe from API
locS = corr.loc[:,['Avg Salary(K)', 'Location']]

#API call and get the latitude & longitude values
for i, row in locS.iterrows():
    address = locS.at[i, 'Location']
    parameters = {
        "key": "f1xh0A8XusBgzhqJai01W4RuVfYVZrN",
        "location": address
    }
    response = requests.get("http://www.mapquestapi.com/geocoding/v1/address", params = parameters)
    cd = json.loads(response.text)['results']

    lat = cd[0]['locations'][0]['latLng']['lat']
    lng = cd[0]['locations'][0]['latLng']['lng']

    locS.at[i, 'lat'] = lat
    locS.at[i, 'lng'] = lng
```

```
In [ ]: def select_marker_color_s(row):
    if float(row['Avg Salary(K)']) < 50.0:
        return '#E6F5E9'
    elif float(row['Avg Salary(K)']) < 70.0:
        return '#C8E6C9'
    elif float(row['Avg Salary(K)']) < 90.0:
        return '#A5D6A7'
    elif float(row['Avg Salary(K)']) < 105.0:
        return '#B1C784'
    elif float(row['Avg Salary(K)']) < 120.0:
        return '#68BB6A'
    elif float(row['Avg Salary(K)']) < 135.0:
        return '#4CAF50'
    elif float(row['Avg Salary(K)']) < 150.0:
        return '#43A047'
    elif float(row['Avg Salary(K)']) < 180.0:
        return '#38A83C'
    elif float(row['Avg Salary(K)']) < 200.0:
        return '#2E7D32'
    elif float(row['Avg Salary(K)']) < 270.0:
        return '#1B5E20'
```

```
In [ ]: locS['color'] = locS.apply(select_marker_color_s, axis = 1)
loc_salary = folium.Map(
    location = [39, -100],
    zoom_start = 4.4
)
loc_salary
for _, loc in locS.iterrows():
    folium.CircleMarker(
        location = [loc['lat'], loc['lng']],
        popup = loc['Avg Salary(K)'],
        tooltip = loc['Avg Salary(K)'],
        radius = 3,
        color = loc['color'],
        fill_color = loc['color']
    ).add_to(loc_salary)

loc_salary
```

Conclusion

In part I, we conclude that python is the most used programming language and flink is the least used programming language. In addition, it looks like Python, SQL, Excel are the top three skills in demand for Data Scientists, Data Engineers, and Machine Learning Engineers. It is imperative to process the python skills if you want to become a data scientist in the future.

However, there are some hidden drawbacks to our dataset. For example, the dataset does not include the R programming language. R as a popular programming language for statistical computing and graphic should be considered in this dataset. Therefore, we could only analyze the given data. More information about programming languages needs to be collected to improve the dataset.

In part II, we try to determine the correlation among Educational Degree, Job Types, and salaries. Among our observations, pursuing a degree in Bachelor is sufficient for many data-related jobs. Having a bachelor's degree can earn an average of 100k per year. Higher education can increase the average salary. The median average salary has a positive correlation with a higher degree.

In part three, we captures the job offers by industries to further prove the versatility of data science. Industries that offer data related jobs range from art to government. It is apparent that tech companies offers the majority of job positions, while fields such as accounting and agriculture extend the least offers. We also calculate the average salary based on industries. Media industries has the highest average salary, and art is on the other spectrum. Due to the different number of job offers across different industries, the average salary

In part four, we examine the relationship among salary, location, and rating. We conclude that companies that have data scientist job offerings are clustered around metropolises. Those regions tend to have higher salaries. However, some regions demonstrate an inverse correlation with company ratings, such as the East Coast region around New York. It is worth noticing that, from an economic standpoint, a higher salary is not equivalent to higher purchasing power. An interesting further study may calculate purchasing power based on the price index by state. Moreover, since job offerings require different seniority, the map may introduce biases by mapping all job offers while ignoring the required experience in the field.

Sources

source data: [data scientist salary](#)

MapQuest API for geo coding