Team tidy

DS 0311-01

Instructor: Norman Lo (lokman@mail.sfsu.edu)

Participants:

- Alleyda Barraza
- Audrey De Leon
- Pyimoe Than
- Paul Hartung
- Maximilian Leitschuh

11 December 2022

DS 311 Group Project

Table of contents

Introduction

Our group selected the salary dataset from the datasets offered.

First, we looked at the data to get an idea of what it was about and to get a feel for the data distribution. This was basically our group task 1.

We decided to clean the data for further tasks, but that work is described in the next chapter.

For the second group task, we answered the given questions and added 3 more to get an overall picture of what factors affect salary the most / least. We wanted to find out facts about American salary and salary for technology in particular.

Data Cleaning

Before we started answering the questions and analyzing the data, we took a closer look at the data and found that it needed to be cleaned first. After reading through

the entire dataset once we realised that there were a lot of missing datapoints (“NaN”). We replaced those missing datapoints with “0”, “Unknown”, “Nothing” or “n”. In the following figure, you can see how we cleaned the data within our Jupyter notebook:

```
#clean the dataset, remove nas
salary['EDUCATION_LEVEL_REQUIRED'] = salary['EDUCATION_LEVEL_REQUIRED'].fillna("Nothing")
salary['COLLEGE_MAJOR_REQUIRED'] = salary['COLLEGE_MAJOR_REQUIRED'].fillna("Nothing")
salary['EXPERIENCE_REQUIRED_Y_N'] = salary['EXPERIENCE_REQUIRED_Y_N'].fillna("n")
salary['EXPERIENCE_REQUIRED_NUM_MONTHS'] = salary['EXPERIENCE_REQUIRED_NUM_MONTHS'].fillna(0)
salary['COUNTRY_OF_CITIZENSHIP'] = salary['COUNTRY_OF_CITIZENSHIP'].fillna("Unknown")
salary['WORK_POSTAL_CODE'] = salary['WORK_POSTAL_CODE'].fillna("Unknown")
salary['FULL_TIME_POSITION_Y_N'] = salary['FULL_TIME_POSITION_Y_N'].fillna("n")
salary['PREVAILING_WAGE_PER_YEAR'] = salary['PREVAILING_WAGE_PER_YEAR'].fillna(0)
salary['WORK_CITY'] = salary['WORK_CITY'].fillna("Unknown")
```

Question 1 (answered by Alleyda Barraza)

Do specific sub-types of data-related jobs have higher or lower salaries than others?

Specific sub-types of data-related jobs do differ in salaries than others. When comparing a business analyst, data analyst, and software engineer, we can see that a Software engineer makes the most money of all 3. The average paid wage per year was referenced.

1. software engineer 92505.30
2. business analyst 71300.08
3. data analyst 70030.08

Which companies have the highest salaries for those sub-types?

The companies that have the highest salaries for software engineer are

- KPI PARTNERS, INC. 1400000.00
- ABACUS TECHNICAL SERVICES, LLC 1224800.60
- INSIDE, INC. 1222788.00
- ROVI CORPORATION 1176000.00
- GOOGLE INC. 1139001.00

Changes with location of the job?

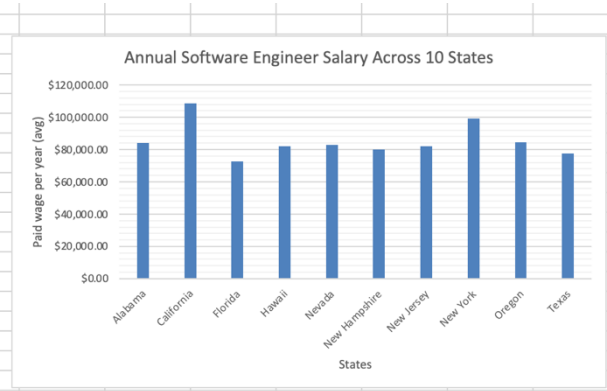
Yes, annual salary is affected by the location of the job. When analyzing the location, we can see that the top 5 highest paying states for the respective job titles are

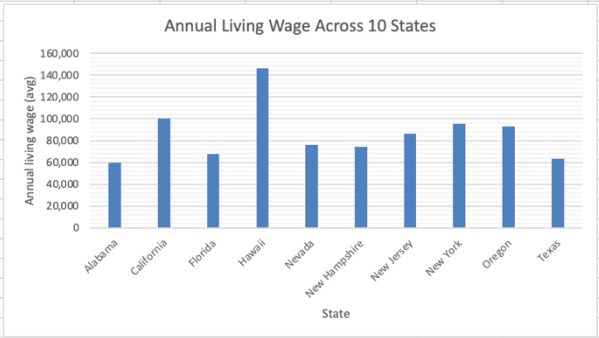
- Software engineer
  - California
  - Washington
  - Virgin Islands
  - New York

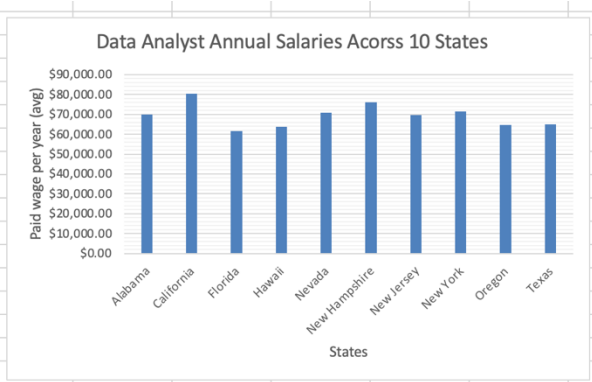
    - Massachusetts
- Business Analyst
    - Washington
    - New York
    - Wyoming
    - California
    - Puerto Rico
- Data analyst
    - Connecticut
    - California
    - Delaware
    - New Hampshire
    - Maine

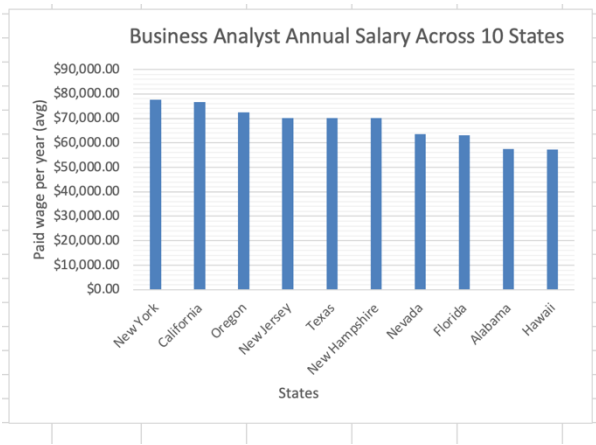
Will the answer change if I take standard of living into account?

Yes, the Standard of living affects the amount of take home annual salary. I compared 10 states and cross all job titles, the standard of living significantly affected annual salary. In Hawaii's case the cost of living per year was $146,437, but the wages for the three data subgroup jobs were significantly lower averaging at about $80,000 per year. Employees should consider location and annual living wage when interviewing or applying to a job within their respective fields.



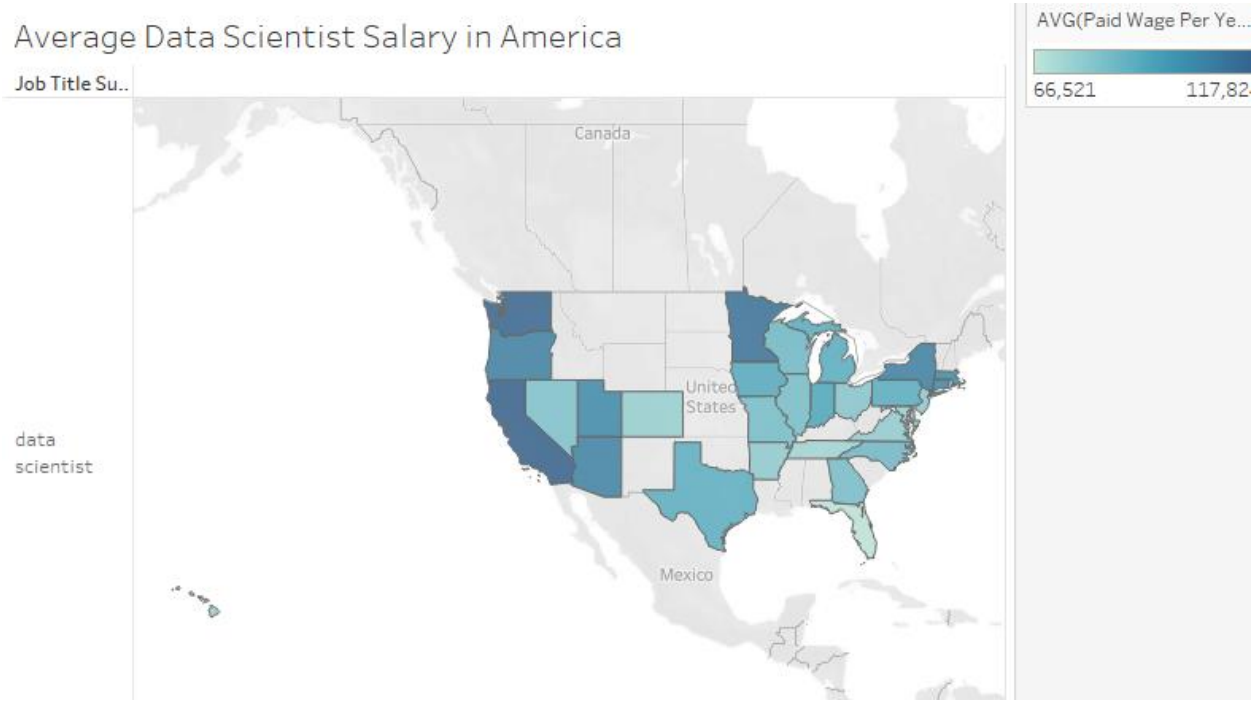




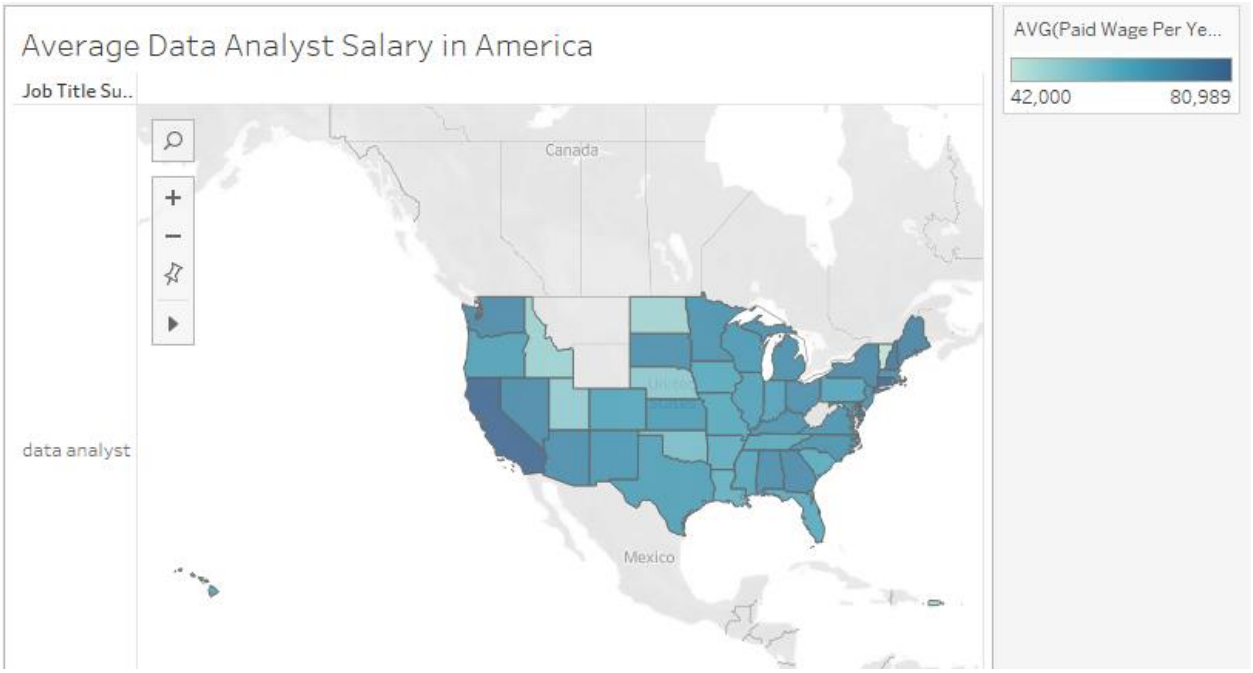


Question 2 (answered by Pyimoe Than)

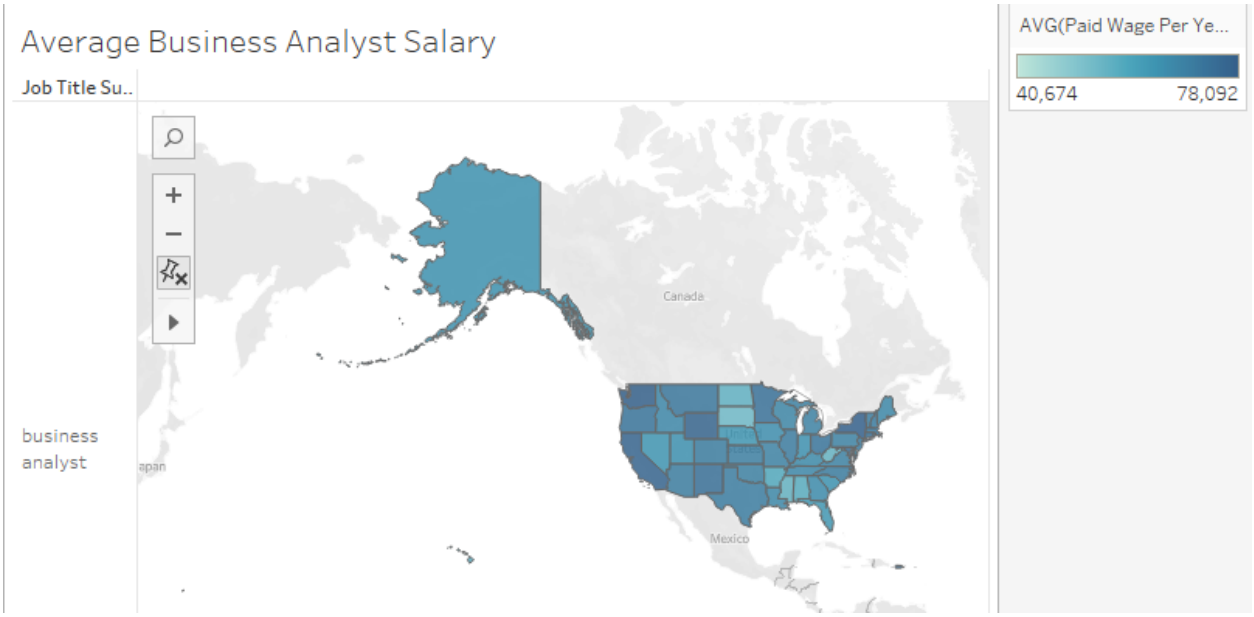
Average Data Scientist salary in America



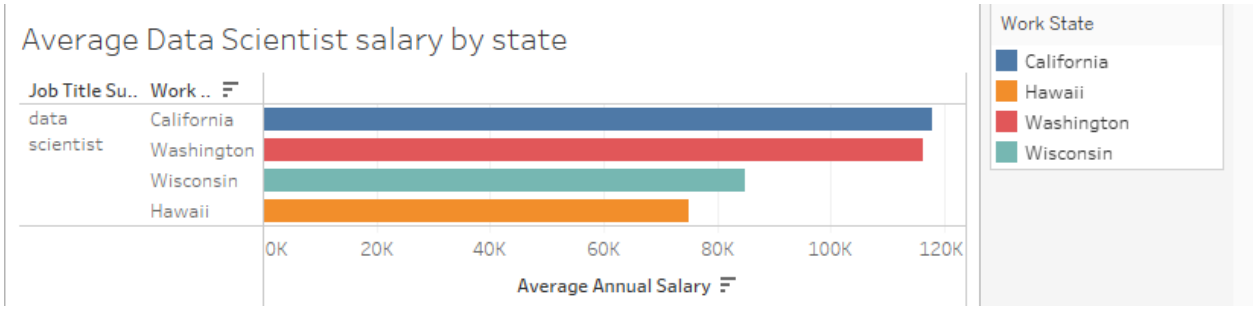
Average Data Analyst salary in America



Average Business Analyst salary in America

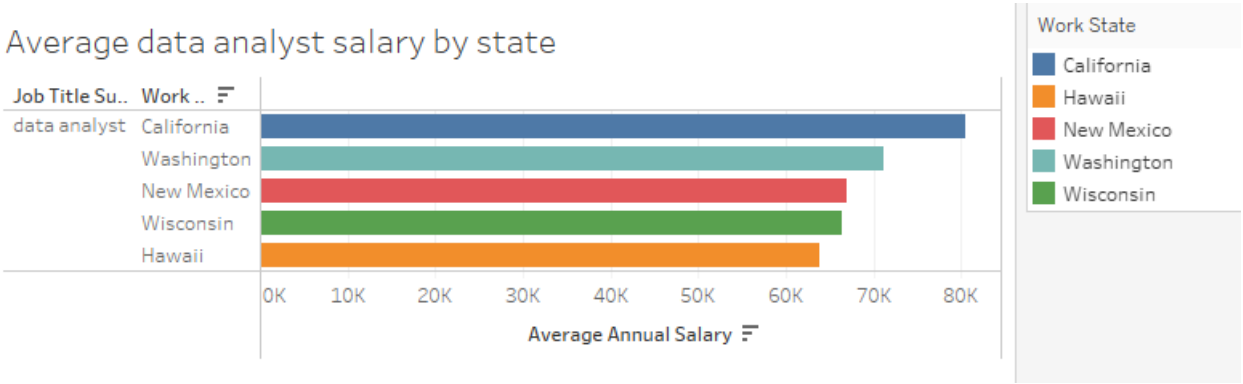


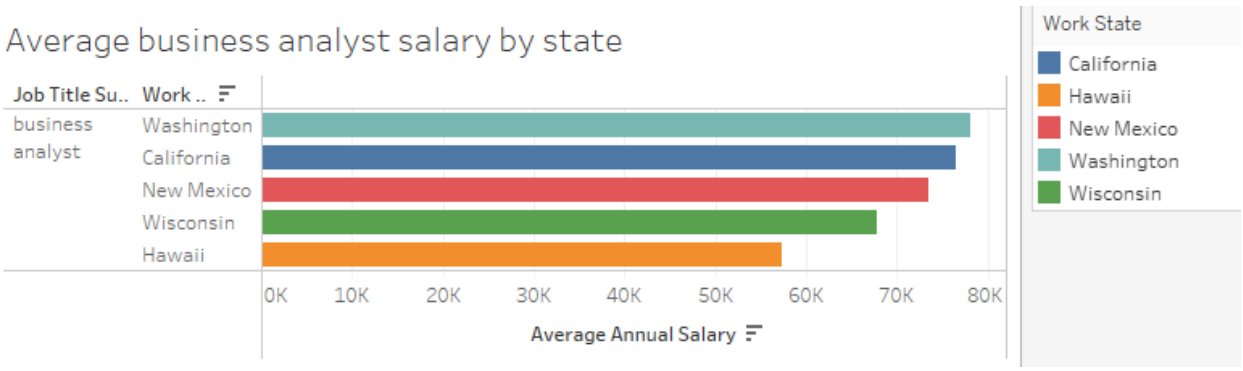
Out of all US states, I choose California, Washington, Hawaii, Wisconsin and New Mexico as my preferred state.

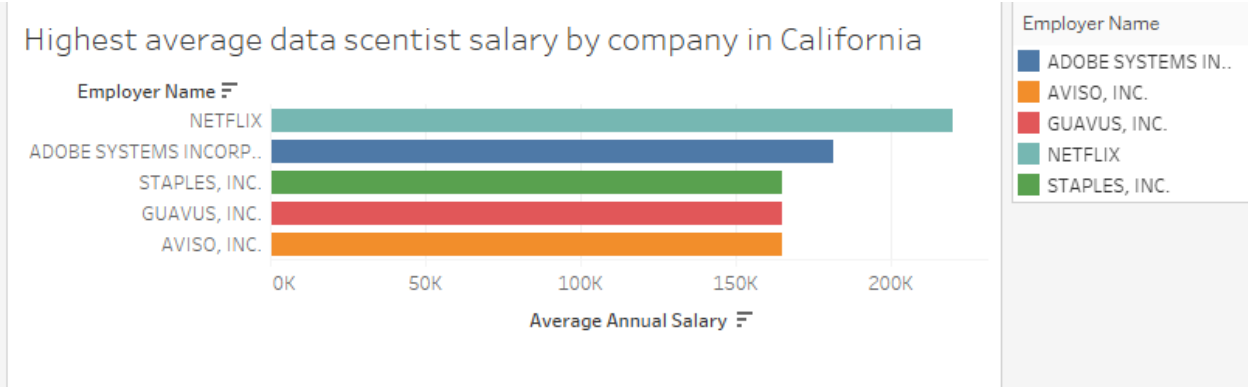


Note: Don't have data for New Mexico for data scientists.

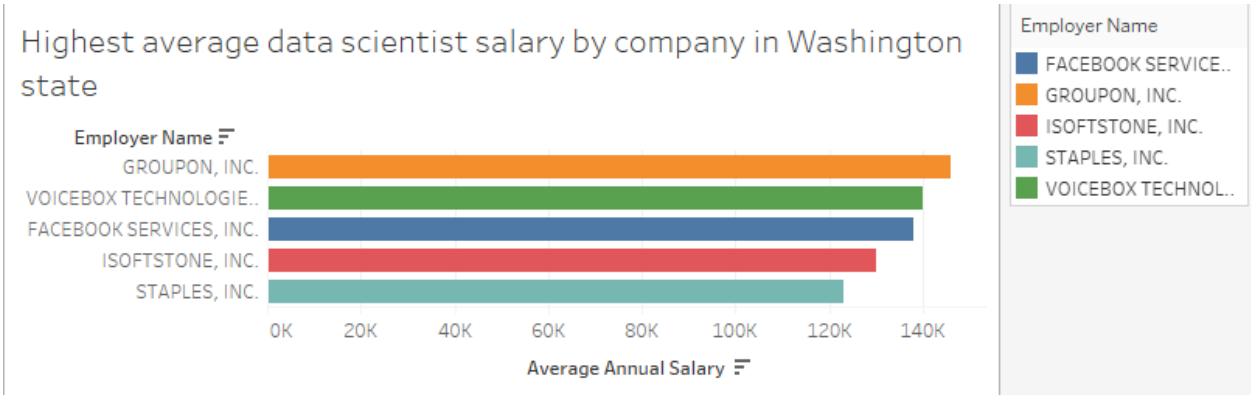
When we compare the salary of data scientists, California and Washington have the highest average salary. In contrast, Wisconsin and Hawaii have the lowest average salary when we compare them with California and Washington. Because of the six figure data scientist salary, many people want to move to live in California and Washington state. They want to become rich. However, we should not only look at the salary. We also need to determine the cost of living and taxes. According to *businessinsider.com*, the living wage in California is $99,971 and the living wage in Wisconsin is $67,667. According to *gobankingrates.com*, the real value of $100 in Wisconsin is $108.10 and the real value of $100 in California is $83.60. Therefore, the dollar value changes depending on where you live. When we add a cost of living in our calculation, average data scientists in Wisconsin get paid more than the average data scientists in California.



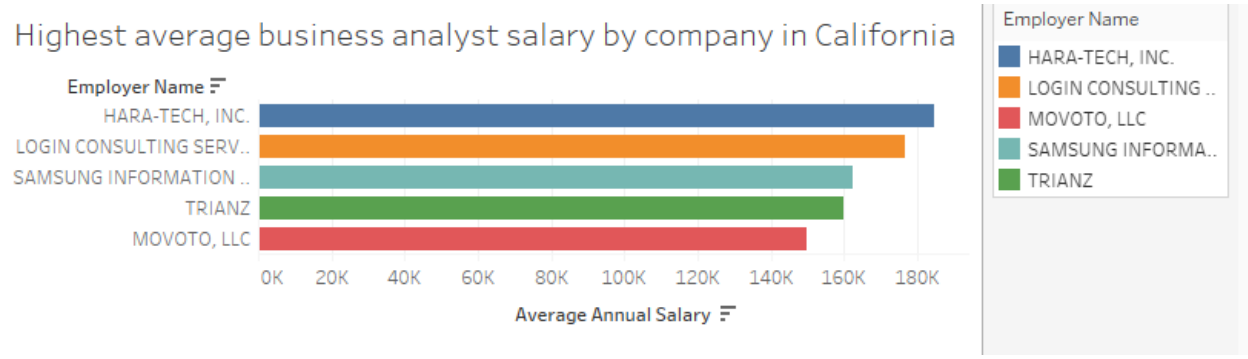




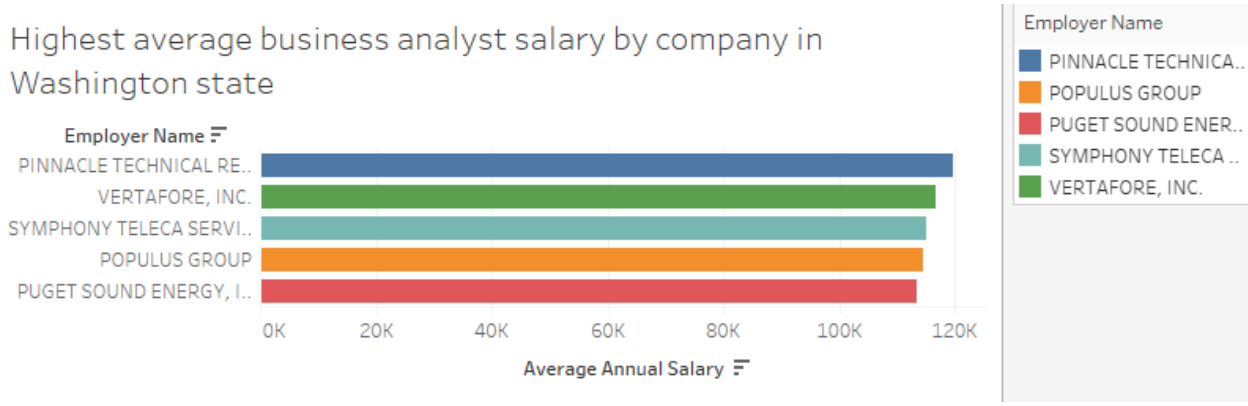
NETFLIX has the highest average data scientist salary in California.



GROUPON, INC has the highest average data scientist salary in Washington state.



Hara-tech, inc has the highest average business analyst salary in California.



Pinnacle Technical has the highest average business analyst salary in Washington state.

Question 3 (answered by Paul Hartung)

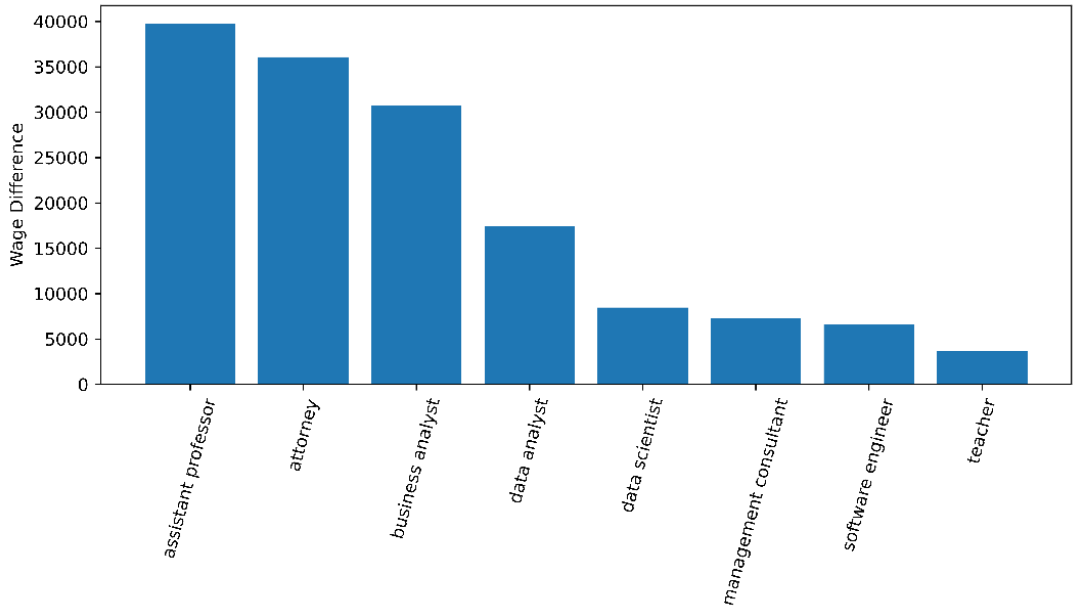
How do offered salaries compare to the prevailing wage?

```
JOB_TITLE_SUBGROUP
attorney                39765.22
assistant professor     36082.45
management consultant   30743.88
data scientist          17475.46
software engineer        8464.93
data analyst             7292.16
business analyst         6568.72
teacher                  3697.37
dtype: float64
```

Are there job sub-categories that tend to get over-paid or under-paid?

All job 8 sub-categories tend to be overpaid, when I subtract prevailed wage from real paid wage.

“Attorneys” are the most overpaid and “Teachers” are the least overpaid.

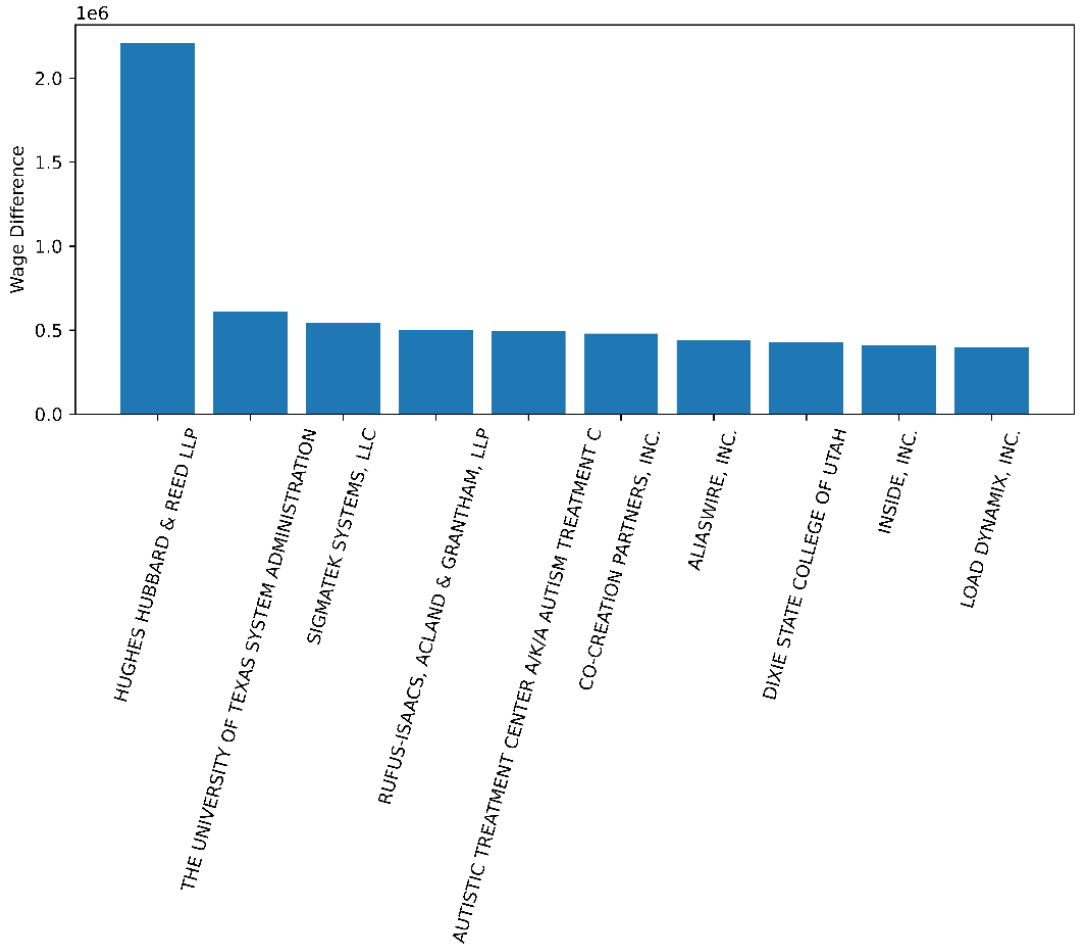


Are there companies that tend to over-pay or under-pay?

Companies such as the following tend to overpay their employees.

The top 3 overpaying companies are:

HUGHES HUBBARD & REED LLP, then “THE UNIVERSITY OF TEXAS SYSTEM ADMINISTRATION”, then “SIGMATEK SYSTEMS, LLC”



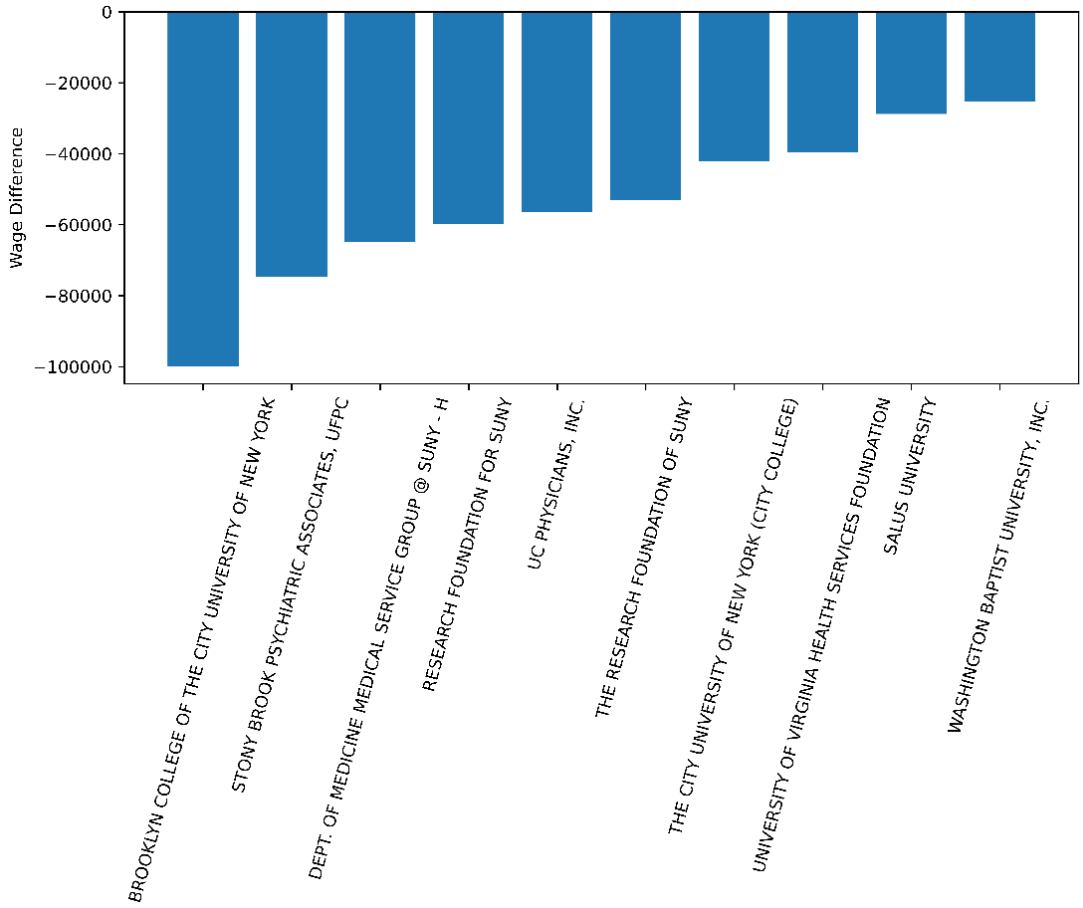
Companies such as the following tend to underpay their employees.

The top 3 underpaying companies are:

BROOKLYN COLLEGE OF THE CITY UNIVERSITY OF NEW YORK

STONY BROOK PSYCHIATRIC ASSOCIATES, UFPC

DEPT. OF MEDICINE MEDICAL SERVICE GROUP @ SUNY - H



Will the answer change if I take standard of living into account?

Where I found standard of living data: https://taxfoundation.org/real-value-100-state-2019/#

I found “worth of 100$” for all US States.

Top 5 States worth of $100 are:

Mississippi $116.69,

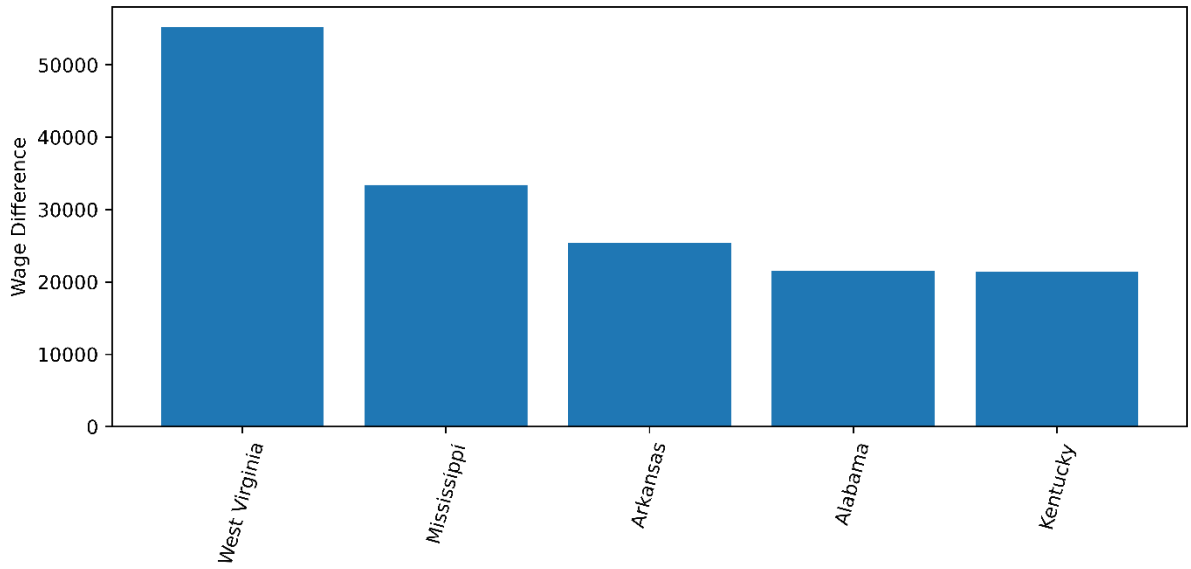
Arkansas $115.61,

Alabama $115.34,

West Virginia $114.94,

Kentucky $113.77,

Coincidentally the top 5 overpaid state are exactly the same:



When I now take into consideration the worth of 100$ in those states. The order of the top 5 does not change.

Before:

```
WORK_STATE
West Virginia   55253.00
Mississippi     33413.18
Arkansas        25347.88
Alabama         21493.09
Kentucky        21384.56
dtype: float64
```

After:

```
   West Virginia  Mississippi  Arkansas  Alabama  Kentucky
0       63507.80     38989.84  29304.68 24790.13  24329.21
```

Question 4&5 (answered by Maximilian Leitschuh)

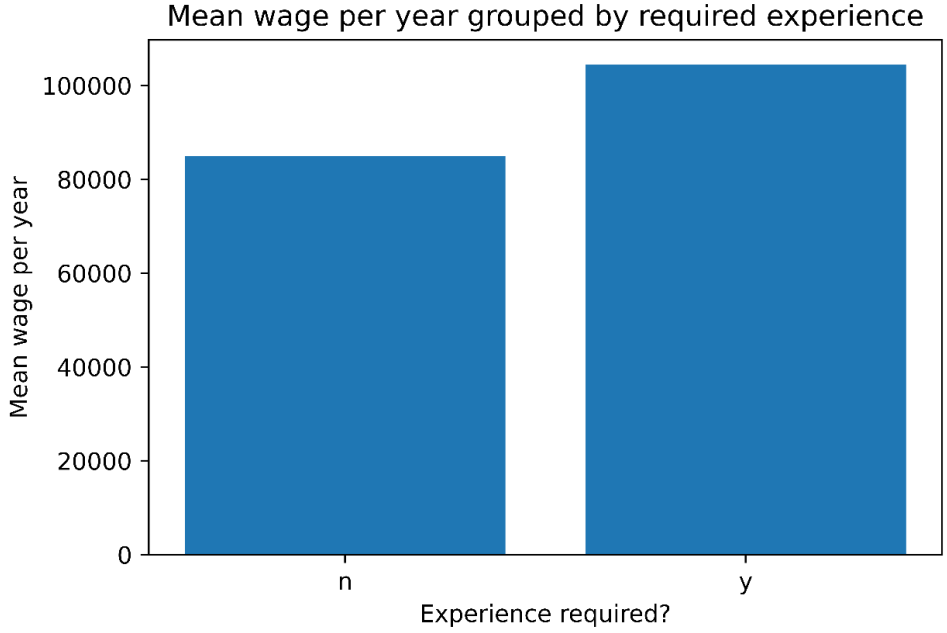
4. Does more required work experience increase the likelihood of a higher salary?

a. What are the differences between the length of required work experience in relation to average salary?

The figure shows that the higher the prior work experience in months, the higher the average annual salary. This trend breaks down at an assumed work experience of 140 months, but this is probably due to too few values within the data set.

b. Do you earn more at a company that requires work experience?

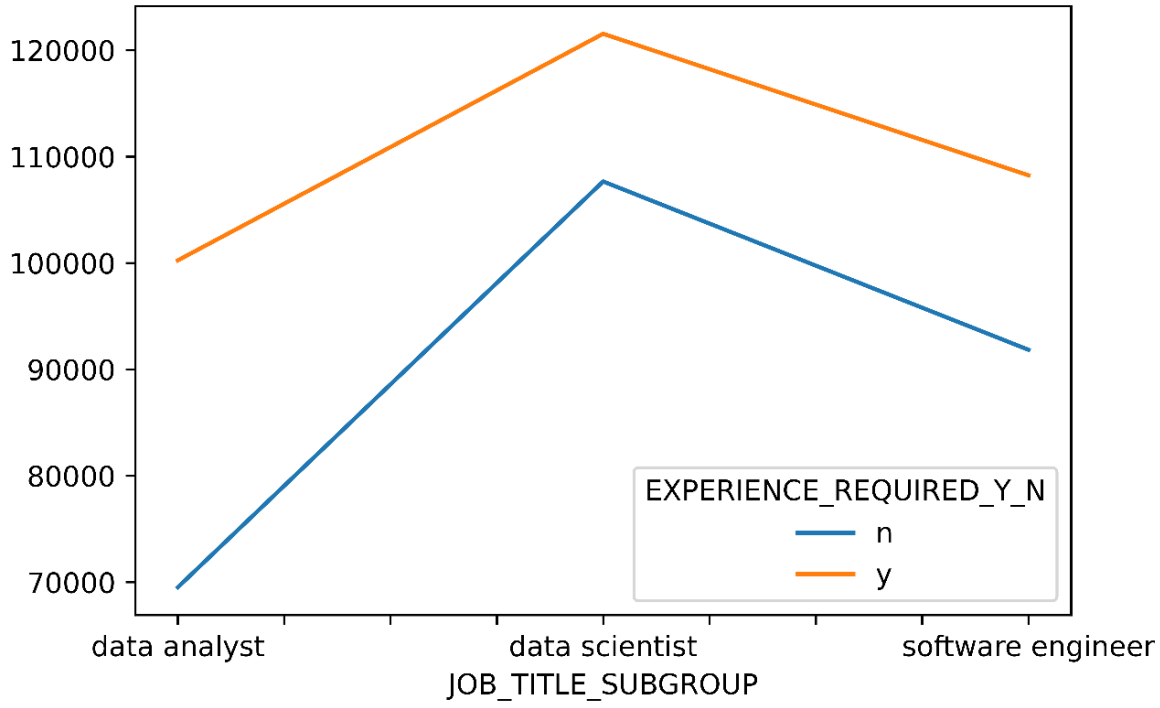
As shown in the figure, the average salary at companies that require work experience is just over $100 thousand. Thus, at companies that require work experience, one earns on average about 20 thousand dollars more per year.



c. **What is the rate of companies in the tech industry that require work experience and what is the mean salary?**

3.95% in the tech industry require work experience

3.95 percent of tech companies within this data set require work experience. This is not a lot and shows that without work experience, finding a job is not a problem. However, it can be seen the figure that tech companies that require work experience pay significantly better. For example, the difference for a data analyst is about $30 thousand per year on average.



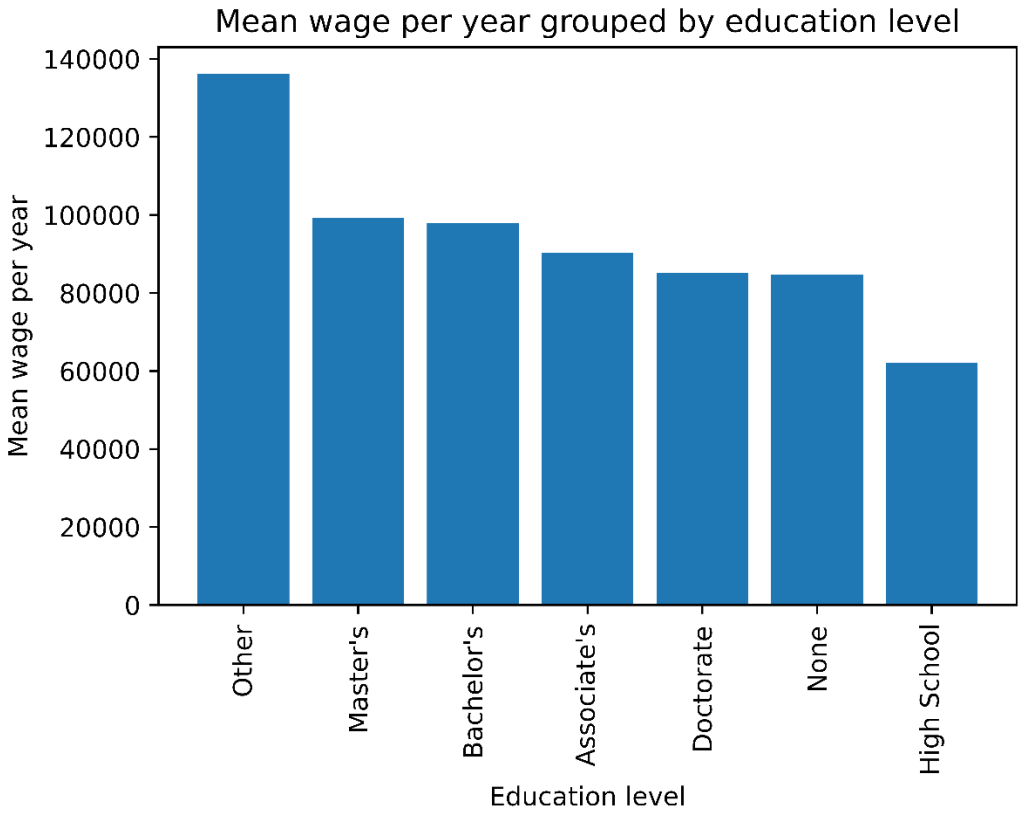
Conclusion:

All in all, required work experience increases the likelihood of a higher salary but it is not often necessary, so that, for example, during the study is not mandatory to gain work experience.

5. Does better required education increase the likelihood of a higher salary?

d. **What are the differences between different levels of education in terms of average salary?**

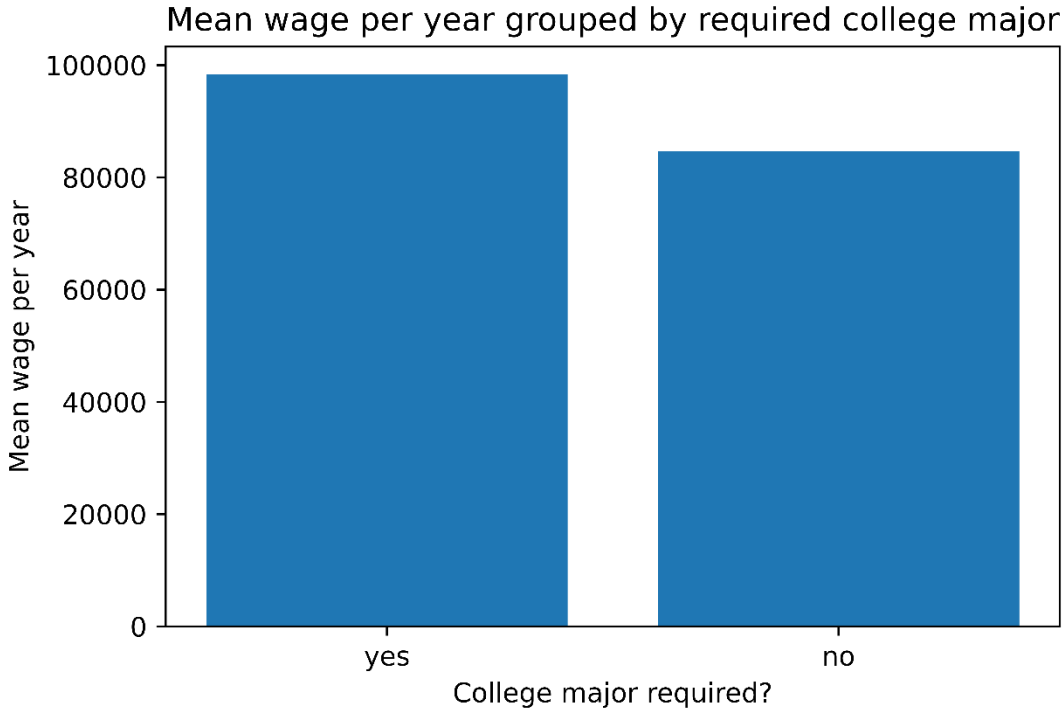
The highest average salary within this dataset is achieved by individuals with the educational level "Other". This is just under 140 thousand dollars and is not specified in more detail. Furthermore, it can be deduced from the figure that people with a bachelor's and master's degree earn the most with an average of just under 100 thousand dollars. High school graduates are somewhat behind at just under 60 thousand dollars.



e. Do you earn more at a company that requires a college degree?

6.61% require a college degree

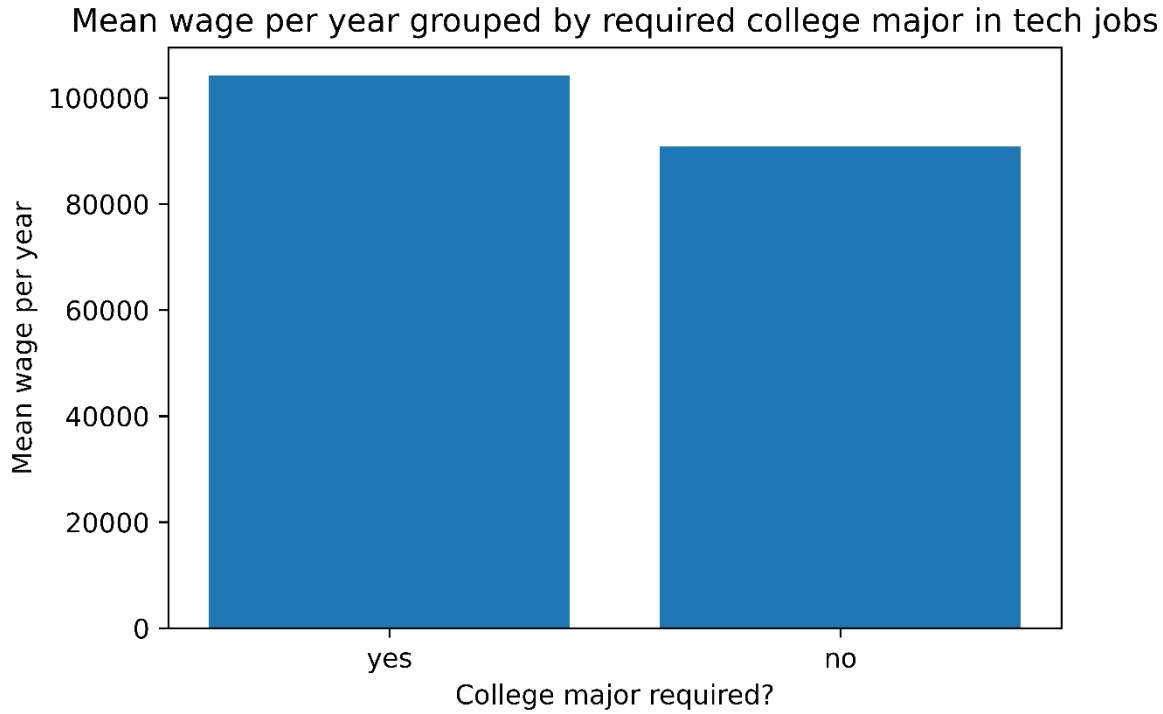
Only 6.61 percent of the companies in this data set require a college major. This is not much. In fact, the average salary with and without a college major differs by only about $15 thousand a year.



f. **What is the rate of companies in the tech industry that require a college major and what is the average wage?**

7.70% of the companies in the tech industry require a college major

Even in tech companies, only 7.7 percent require a college major. Again, the range in annual salary between one and no college is around $15 thousand.



Conclusion: With a college major or a good education, one earns more on average than with a poor or no education. However, the margins are not as high as for work experience. This suggests that work experience is more important to companies today than the highest possible degree. Of course, this is not always the case, and education is and always will be an important factor for a successful professional life.

Question 6 (answered by Audrey De Leon)

What is the distribution of Salaries Country of Citizenship and Visa Class

What was the distribution of salaries based on Country of Citizenship (Max/Min/Mean)

Top 5 Max Countries:

COUNTRY_OF_CITIZENSHIP	VISA_CLASS	
Unknown	E-3 **Australian**	2500000.0
Unknown	H-1B	2400000.0
INDIA	greencard	746323.0
CANADA	greencard	500000.0
AUSTRALIA	greencard	450000.0

Top 5 Mean Countries:

COUNTRY_OF_CITIZENSHIP	VISA_CLASS	
SYRIA	greencard	178755.625000
ESTONIA	greencard	163480.000000
LIBYA	greencard	161540.000000
KUWAIT	greencard	148803.000000
LEBANON	greencard	138500.888889

Top 5 Min Countries:

PERU	greencard	20190.0
RUSSIA	greencard	18960.0
SOUTH KOREA	greencard	18510.0
Unknown	E-3 **Australian**	12000.0
Unknown	H-1B	10500.0

Which Visa_Class has the most requests for sponsorship?

When looking at the results of the highest paying salaries based on Country of Citizenship, the visa class "E-3 Australian" had the most requests and the highest and lowest range in paid wages per year. It had a total of 1,393 --- it seems that the E-3 Australian visa class had a large disparity when compared to the other visa classes (with the exception of the greencard visa class at 11,093).

In [101]:

```
#Top Earning COUNTRY_OF_CITIZENSHIP Non Green Card Holder
salary[salary["VISA_CLASS"] == "E-3 Australian"]
```

Out[101]:

	CASE_NUMBER	CASE_STATUS	CASE_RECEIVED_DATE	DECISION_DATE	EMPLOYER_NAME	JOB_TITLE	WORK_CI
4	I-203-14259-128844	denied	9/16/2014	9/23/2014	SIGNAL SCIENCES CORPORATION	SENIOR SOFTWARE ENGINEER	PORTLA
75	I-203-13325-889631	certified	11/21/2013	11/27/2013	NEW YORK UNIVERSITY	ADJUNCT ASSISTANT PROFESSOR	NEW YO
94	I-203-14043-770403	certified	2/19/2014	2/25/2014	JONES DAY	ASSOCIATE ATTORNEY	PALO AL
121	I-203-14227-436034	certified	8/15/2014	8/24/2014	BRACEWELL & GIULIANI LLP	ASSOCIATE ATTORNEY	NEW YO
128	I-203-14307-410538	certified	11/3/2014	11/7/2014	WHITE & CASE LLP	ASSOCIATE ATTORNEY	NEW YO
...	...	...	...	...	...	...	
166971	I-203-12352-723226	certified	12/17/2012	12/21/2012	JIMBO GYMNASTICS, INC	TEACHER	SCARSDA
166973	I-203-13291-409736	denied	10/18/2013	10/29/2013	PROFESSIONAL RACQUET ORGANIZATION SPORTS CLUB	LIFESTYLE AND WEIGHT MANAGEMENT CONSULTANT	BELLEV
166995	I-203-13350-875765	certified	12/16/2013	12/20/2013	ACADEMY OF ST JOSEPH	TEACHER ASSISTANT	NEW YO
167233	I-203-13177-004887	denied	7/9/2013	7/11/2013	VUBIQUITY INTERNATIONAL HOLDINGS, INC.	ENTERPRISE BUSINESS ANALYST	BURBA
167268	I-203-15014-844277	denied	1/14/2015	1/22/2015	RESCUE RESPONSE GEAR INC.	TEACHER AND INSTRUCTOR	BE

1393 rows × 23 columns

In [99]:

```
#Top Earning COUNTRY_OF_CITIZENSHIP Non Green Card Holder
salary[salary["VISA_CLASS"] == "greencard"]
```

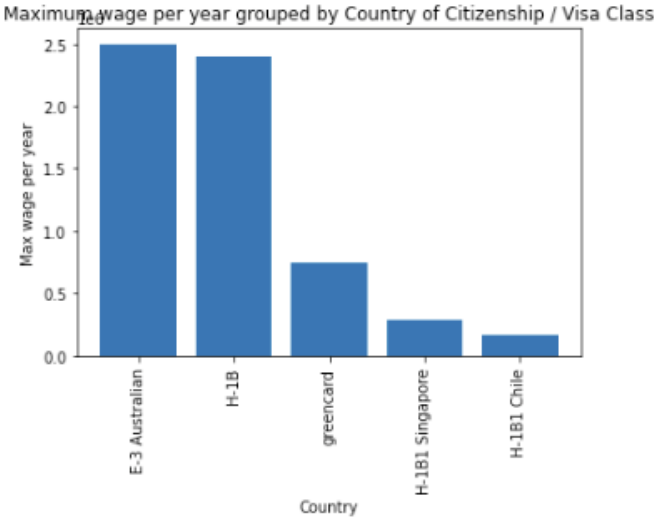
Out[99]:

	CASE_NUMBER	CASE_STATUS	CASE_RECEIVED_DATE	DECISION_DATE	EMPLOYER_NAME	JOB_TITLE	WORK_CITY
1	A-15061-55212	denied	3/19/2015	3/19/2015	SAN FRANCISCO STATE UNIVERSITY	Assistant Professor of Marketing	SAN FRANCISCO
109	A-14239-02058	certified	10/17/2014	6/25/2015	GOOGLE INC.	Software Engineering Manager	Mountain View
113	A-14115-63722	certified-expired	5/13/2014	10/16/2014	GOOGLE INC.	Software Engineer	Mountain View
114	A-14220-96670	certified	10/8/2014	3/4/2015	APPLE INC.	Software Engineer Applications Manager	Cupertino
172	A-14274-11865	certified	11/21/2014	4/30/2015	UNIVERSITY OF MICHIGAN	Assistant Professor (Clinical Track)	Ann Arbor
...	...	...	...	...	...	...	...
167162	A-14106-61006	certified-expired	5/19/2014	10/8/2014	BOWLING GREEN STATE UNIVERSITY	Assistant Professor	Bowling Green
167163	A-14241-02859	certified	9/10/2014	2/3/2015	TEXAS TECH UNIVERSITY	Assistant Professor	Lubbock
167200	A-14269-10623	certified	11/19/2014	5/7/2015	Glenville State College	Assistant Professor of Computer Science	Glenville
167223	A-14283-14734	certified	10/29/2014	3/19/2015	BOISE STATE UNIVERSITY	Assistant Professor of English	Boise
167226	A-14196-88468	certified-expired	7/29/2014	12/16/2014	UNIVERSITY OF VIRGINIA	Assistant Professor of Spanish	Wise

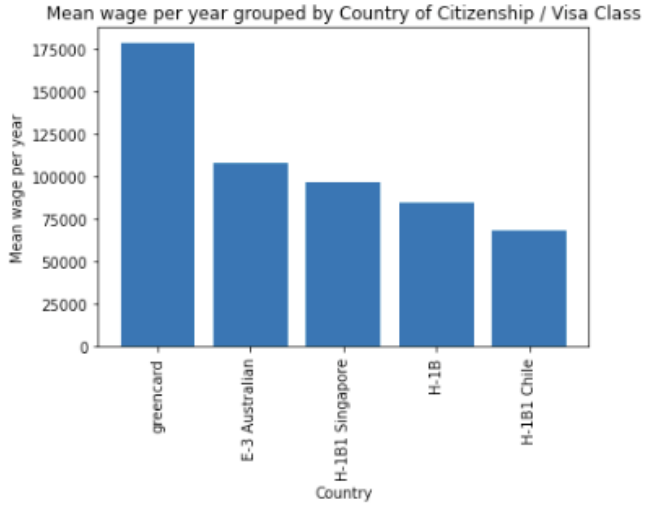
11093 rows × 23 columns

Distribution of Wage by Country of Citizenship and Visa Class (Max, Mean, Min)

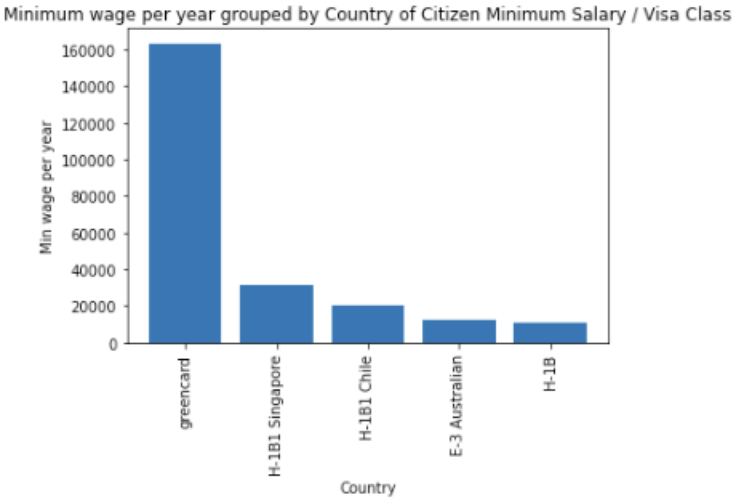
Max



Mean



Min



Conclusion

As can be seen in the previous response to each question, some interesting insights were gleaned from the salary data set. Python was predominantly used to analyze the data and one team member used Tableau for analysis. Overall, this was a very harmonious and good team project in which each team member contributed. Each team member was able to gain some valuable experience in the field of data analysis.