DS 0311-01 / Instructor: Norman Lo (lokman@mail.sfsu.edu) / 07 October 2022

DS 311 Group Assignment 1

Tasks: Write a page of summary of the dataset includes 5 fun facts that your team found.
The 1-page write up is not including your code or coding file.
Push the coding file and summary report to the group project repository.

Fun facts about the dataset, option 2: “salary_data_states.xlsx”

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”. *(Figure 5)*

Fact number 1:

Sorting the mean wage per year by job subgroups, we can conclude that “attorney” related jobs make the most money, they make 140,000$ on average per year. The subgroup “teacher” is in last place and makes less than half of an attorney.

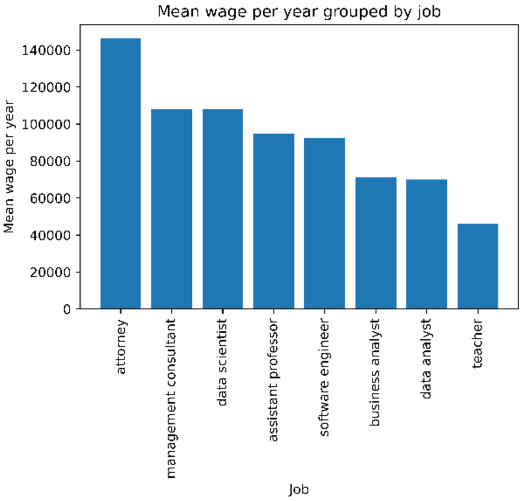
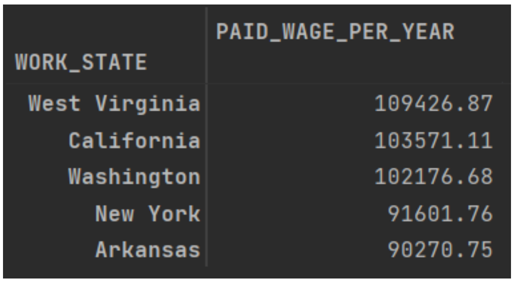


Figure 1 Mean wage per year, grouped by job category

Fact number 2:

Sorting the mean wage per year by states, we can conclude that employees in “West Virginia” make the most money, they make around 109,000$ per year. We displayed the top 5 states in *(Figure 2)*



	PAID_WAGE_PER_YEAR
WORK_STATE	
West Virginia	109426.87
California	103571.11
Washington	102176.68
New York	91601.76
Arkansas	90270.75

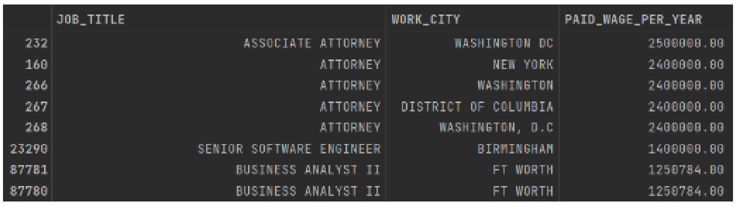
Figure 2 Mean wage per year, grouped by state.

Fact number 3:

The mean salary per year in San Francisco is around 123,850$

Fact number 4:

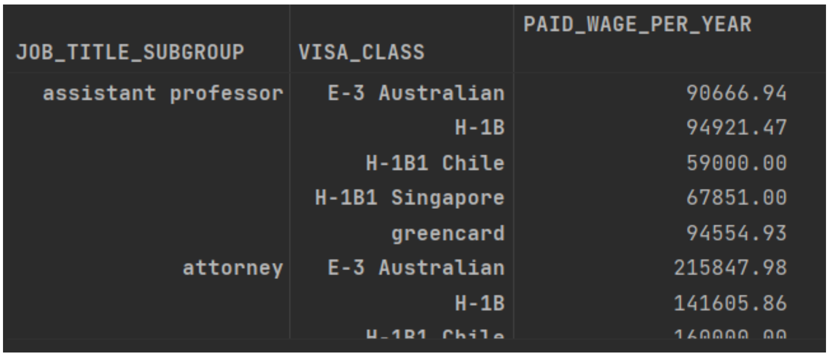
Looking at who makes the most money per year as a single person, we can see that there are a few attorneys that make substantially more than the rest in the dataset. Number 1 to 5 are attorneys that each make over a million more than number 6.



	JOB_TITLE	WORK_CITY	PAID_WAGE_PER_YEAR
232	ASSOCIATE ATTORNEY	WASHINGTON DC	2500000.00
160	ATTORNEY	NEW YORK	2400000.00
266	ATTORNEY	WASHINGTON	2400000.00
267	ATTORNEY	DISTRICT OF COLUMBIA	2400000.00
268	ATTORNEY	WASHINGTON, D.C	2400000.00
23290	SENIOR SOFTWARE ENGINEER	BIRMINGHAM	1400000.00
87781	BUSINESS ANALYST II	FT WORTH	1250784.00
87780	BUSINESS ANALYST II	FT WORTH	1250784.00

Figure 3 Best paid people with job and city.

Fact number 5:



JOB_TITLE_SUBGROUP	VISA_CLASS	PAID_WAGE_PER_YEAR
assistant professor	E-3 Australian	90666.94
	H-1B	94921.47
	H-1B1 Chile	59000.00
	H-1B1 Singapore	67851.00
	greencard	94554.93
attorney	E-3 Australian	215847.98
	H-1B	141605.86
	H-1B1 Chile	160000.00

Figure 4 Visa class vs paid wage

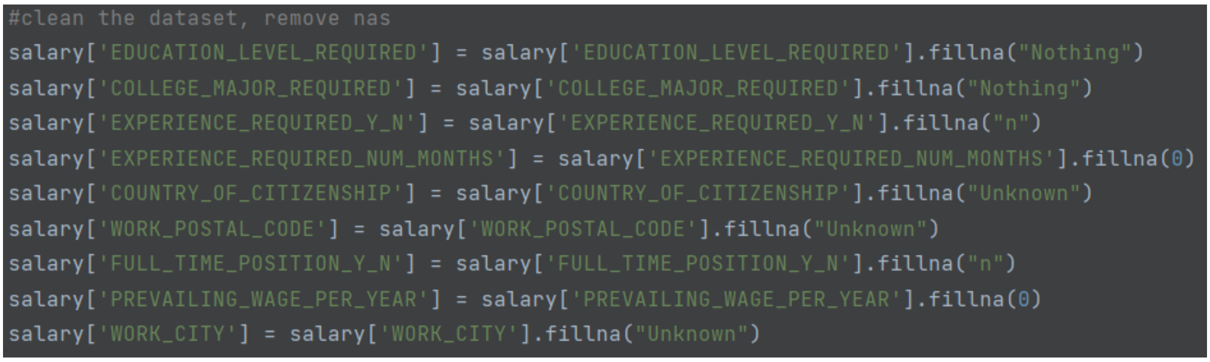
Considering the diversity within companies, we wanted to look at how salaries of those on work visas compare to green card holders. The data was sorted by job title subgroup, visa class, and lastly by organizing the mean paid wage per year. Across the board we note that those not holding green cards are paid significantly less than green card holders. The most notable jump is a data scientist from Chile gets paid a mean annual wage of 90000.00 compared to a greencard holder who makes 116507.76, the amount of experience would have to also be analyzed to gather more conclusive results. When isolating wage by visa status we can notably see that greencard holders on average make more than those on work sponsored visas.

Summary:

The data we have shown is of course only representative of the given dataset and should not be understood in general terms.

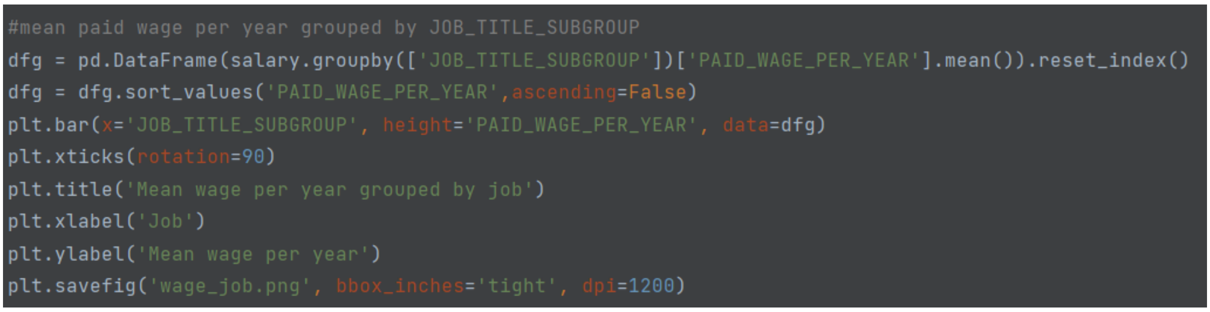
To summarize our first look at the dataset, we can say that the data provided contains quite a bit of intel about deployment data, but is also very broad and provides a lot of information about its relationship/relevance to visa status, the state and city where one works, etc....

Code screenshots attachment:



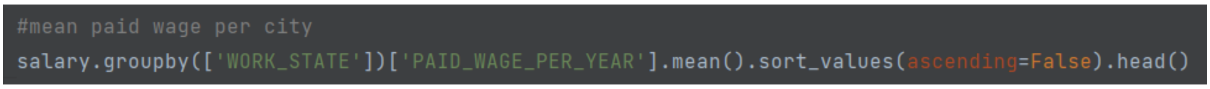
```
#clean the dataset, remove nas
salary['EDUCATION_LEVEL_REQUIRED'] = salary['EDUCATION_LEVEL_REQUIRED'].fillna("Nothing")
salary['COLLEGE_MAJOR_REQUIRED'] = salary['COLLEGE_MAJOR_REQUIRED'].fillna("Nothing")
salary['EXPERIENCE_REQUIRED_Y_N'] = salary['EXPERIENCE_REQUIRED_Y_N'].fillna("n")
salary['EXPERIENCE_REQUIRED_NUM_MONTHS'] = 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")
```

Figure 5 replacing missing data



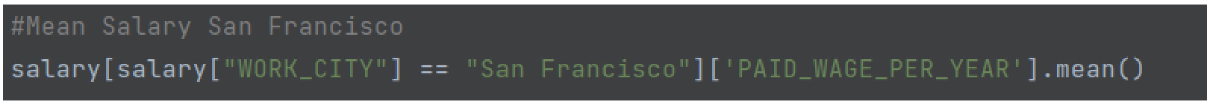
```
#mean paid wage per year grouped by JOB_TITLE_SUBGROUP
dfg = pd.DataFrame(salary.groupby(['JOB_TITLE_SUBGROUP'])['PAID_WAGE_PER_YEAR'].mean()).reset_index()
dfg = dfg.sort_values('PAID_WAGE_PER_YEAR',ascending=False)
plt.bar(x='JOB_TITLE_SUBGROUP', height='PAID_WAGE_PER_YEAR', data=dfg)
plt.xticks(rotation=90)
plt.title('Mean wage per year grouped by job')
plt.xlabel('Job')
plt.ylabel('Mean wage per year')
plt.savefig('wage_job.png', bbox_inches='tight', dpi=1200)
```

Figure 6 Code for fact 1



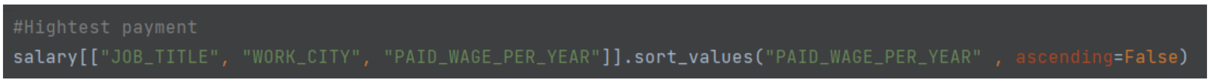
```
#mean paid wage per city
salary.groupby(['WORK_STATE'])['PAID_WAGE_PER_YEAR'].mean().sort_values(ascending=False).head()
```

Figure 7 Code for fact 2



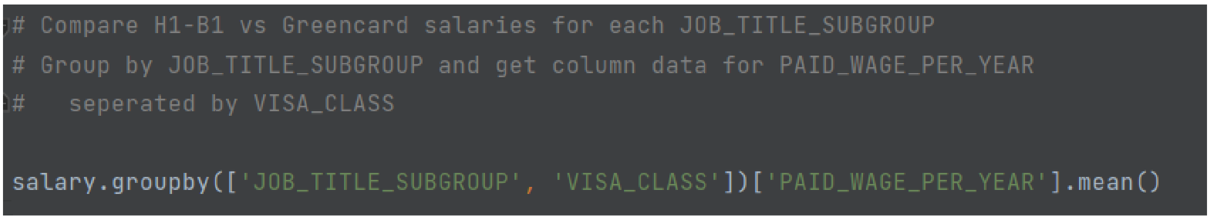
```
#Mean Salary San Francisco
salary[salary["WORK_CITY"] == "San Francisco"]['PAID_WAGE_PER_YEAR'].mean()
```

Figure 8 Code for fact 3



```
#Hightest payment
salary[["JOB_TITLE", "WORK_CITY", "PAID_WAGE_PER_YEAR"]].sort_values("PAID_WAGE_PER_YEAR" , ascending=False)
```

Figure 9 Code for fact 4



```
# Compare H1-B1 vs Greencard salaries for each JOB_TITLE_SUBGROUP
# Group by JOB_TITLE_SUBGROUP and get column data for PAID_WAGE_PER_YEAR
#   seperated by VISA_CLASS

salary.groupby(['JOB_TITLE_SUBGROUP', 'VISA_CLASS'])['PAID_WAGE_PER_YEAR'].mean()
```

Figure 10 Code for fact 5