Task 1 - Data Modeling

In this notebook, I perform an exploratory analysis on the data tables provided in the `\data` folder within this repo, and create a sqlite database in which to launch queries to answer specific business questions regarding employers in the jobs dataset.

Data relationships

In the following I find that the .csv files are best represented using a SNOWFLAKE model where the `postings.csv` acts as the fact table, and all others as dimenstion tables. While as a whole, a SNOWFLAKE schema is best used to represent the data (because not all tables have 1 degree of separation in relation to the fact table, some have 2) a simpler STAR schema is the most practical to create a database in which I can answer basic business questions defined by the client.

Final Schema

A simple STAR schema is the most practical to create a database in which I can answer basic business questions defined by the client. Specifically, `postings.csv` acts as the fact_table `fact_job_postings` and `company_industries.csv` acts as the dimension table `dim_company`. These tables are related via the "company_id" column that exists within both tables. More like a SHARD schema.

Insights gained

The following business questions were asked about the dataset, here they are summarized, please view the respective sections for SQL queries for more detailed answers

1. How many companies have more than one job posting?: `601`
2. How many job postings are there for each job industry?: `The range from 1010 in Hospitals and Health Care, to 1 in Government Relations Services`
3. What is the average normalized salary by company industry?: `They range from 250,000 in Information Services to, NONE in sectors where there was insufficient data to state an average.`
4. Name the top 5 companies with the highest average normalized salary for their job postings.:

Company Name	Salary
Woodside Staffing Solutions & Consulting	337,500.00
Calm	337,500.00
Health eCareers	337,246.41
Buck Institute for Research on Aging	300,000.00
Spire Orthopedic Partners	284,124.00

Exorbidantly high, but this is because often the number jobs posted by that company is just 1, so the average is the single datapoint, perhaps the CEO?

Insustries with insufficient salary data

For quick reference here is the list of industries where there isnt enough information to give an average norm. salary

Industry Category	Value
Writing and Editing	None
Recreational Facilities	None
Public Safety	None
Printing Services	None
Performing Arts	None
Outsourcing and Offshoring Consulting	None
Machinery Manufacturing	None
Libraries	None
Government Relations Services	None
Civic and Social Organizations	None
Armed Forces	None
Appliances, Electrical, and Electronics Manufacturing	None
Animation and Post-production	None

In [1]:
```
# Importing standard data analysis packages
import pandas as pd
import sqlite3
import prettytable
from matplotlib import pyplot as plt
import seaborn as sns
import dash
import plotly
from IPython.display import display, HTML
```

1.1 Explore the source data

The available data has the following folder structure and is shown for convenience below. Lets try and see what variables the tables have in common, so I can identify the fact and dimension tables

In [2]:
```
'''
├── companies
│   ├── companies.csv
│   ├── company_industries.csv
│   ├── company_specialities.csv
│   └── employee_counts.csv
├── jobs
│   ├── benefits.csv
│   ├── job_industries.csv
│   ├── job_skills.csv
│   └── salaries.csv
├── mappings
│   ├── industries.csv
│   └── skills.csv
└── postings.csv
'''
```

Out[2]:
```
'\n├── companies\n│\xa0\xa0 ├── companies.csv\n│\xa0\xa0 ├── company_indust
ries.csv\n│\xa0\xa0 ├── company_specialities.csv\n│\xa0\xa0 └── employee_co
unts.csv\n├── jobs\n│\xa0\xa0 ├── benefits.csv\n│\xa0\xa0 ├── job_industrie
s.csv\n│\xa0\xa0 ├── job_skills.csv\n│\xa0\xa0 └── salaries.csv\n├── mappin
gs\n│\xa0\xa0 ├── industries.csv\n│\xa0\xa0 └── skills.csv\n└── postings.cs
v\n'
```

Postings data (Fact table)

Particularly interesting here is the job_id and the company_id, since these are identifiers that could exist in other lookup tables (dimension tables)

In [3]:
```python
# Exploring Postings data
postings_df = pd.read_csv("../data/postings.csv")
cols_posting = sorted(list(postings_df.columns))
print('n columns: ',len(cols_posting))
print(cols_posting)
```

```
n columns:  31
['application_type', 'application_url', 'applies', 'closed_time', 'company_i
d', 'company_name', 'compensation_type', 'currency', 'description', 'expir
y', 'fips', 'formatted_experience_level', 'formatted_work_type', 'job_id',
'job_posting_url', 'listed_time', 'location', 'max_salary', 'med_salary', 'm
in_salary', 'normalized_salary', 'original_listed_time', 'pay_period', 'post
ing_domain', 'remote_allowed', 'skills_desc', 'sponsored', 'title', 'views',
'work_type', 'zip_code']
```

In [4]:
```python
postings_df.head()
```

Out[4]:

	job_id	company_name	title	description	max_salary	pay_period
0	91700727	Downtown Raleigh Alliance	Economic Development and Planning Intern	Job summary:The Economic Development & Plannin...	20.0	HOURLY
1	2264355	Bay West Church	Worship Leader	It is an exciting time to be a part of our chu...	NaN	MONTHLY
2	229924287	REquipment Durable Medical Equipment and Assis...	Administrative Assistant	The Administrative Assistant will organize and...	NaN	HOURLY
3	358267047	ADEPT HRM Solutions	Production Planner (Food Technologist)	Job Summary: We are seeking a skilled Producti...	NaN	NaN
4	445337908	Food Bank of Alaska	Chief Operating Officer	The Chief Operations Officer (COO) position is...	110000.0	YEARLY

5 rows × 31 columns

Companies data

The company_id column seems to be particulary interesting here, since it is shared with the postings data

In [5]:
```python
# Exploring Companies data
companies_df = pd.read_csv("../data/companies/companies.csv")
industries_df = pd.read_csv("../data/companies/company_industries.csv")
specialties_df = pd.read_csv("../data/companies/company_specialities.csv")
employee_df = pd.read_csv("../data/companies/employee_counts.csv")

# Creating a list to show all available columns
companies = list(companies_df.columns)
industries = list(industries_df.columns)
specialties = list(specialties_df.columns)
employees = list(employee_df.columns)
print("companies:  ", companies)
print("industries: ", industries)
print("specialties:", specialties)
print("employees:  ", employees)
```

```
companies:   ['Unnamed: 0', 'company_id', 'name', 'description', 'company_si
ze', 'state', 'country', 'city', 'zip_code', 'address', 'url']
industries:  ['Unnamed: 0', 'company_id', 'industry']
specialties: ['Unnamed: 0', 'company_id', 'speciality']
employees:   ['Unnamed: 0', 'company_id', 'employee_count', 'follower_coun
t', 'time_recorded']
```

In [6]:
```python
companies_df.head()
```

Out[6]:

	Unnamed: 0	company_id	name	description	company_size	state	co
0	18	1088	NXP Semiconductors	NXP Semiconductors N.V. (NASDAQ: NXPI) enables...	7.0	Noord-Brabant	
1	27	1207	Johnson & Johnson	At Johnson & Johnson, we believe health is eve...	7.0	NJ	
2	29	1224	US Army Corps of Engineers	U.S. Army Corps of Engineers Mission: \nProvid...	7.0	DC	
3	44	1292	The Walt Disney Company	From classic animated features and exhilaratin...	7.0	CA	
4	52	1360	National Computer Systems	WHY CHOOSE NCS ?\nTop 5 reasons why clients ch...	3.0	0	

In [7]:
```python
print(companies_df["Unnamed: 0"].min(),companies_df["Unnamed: 0"].max())
```

```
18 24471
```

In [8]:
```python
industries_df.head()
```

Out[8]:

	Unnamed: 0	company_id	industry
0	18	33218	Staffing and Recruiting
1	36	7790573	Business Consulting and Services
2	49	24803	Staffing and Recruiting
3	50	13345578	IT Services and IT Consulting
4	57	54077952	Motor Vehicle Manufacturing

In [9]:
```python
print(industries_df["Unnamed: 0"].min(), industries_df["Unnamed: 0"].max())
```

```
18  24266
```

In [10]:
```
industries_df.describe()
```

Out[10]:

	Unnamed: 0	company_id
count	1432.000000	1.432000e+03
mean	12275.868017	2.064689e+07
std	7000.207157	3.178757e+07
min	18.000000	1.088000e+03
25%	6200.000000	1.661878e+05
50%	12396.000000	2.860462e+06
75%	18530.250000	2.702445e+07
max	24266.000000	1.034689e+08

In [11]:
```
specialties_df.head()
```

Out[11]:

	Unnamed: 0	company_id	speciality
0	149	33218	CSS Tec
1	150	33218	CSS ProSearch
2	151	33218	CSS Professional Staffing
3	152	33218	CSS Accounting & Finance
4	153	33218	Peergenics

In [12]:
```
employee_df.head()
```

Out[12]:

	Unnamed: 0	company_id	employee_count	follower_count	time_recorded
0	18	33218	191	36335	1712346173
1	36	7790573	16	233	1712346248
2	49	24803	130	60572	1712346323
3	50	13345578	279	85916	1712346323
4	57	54077952	74	686	1712346397

I am not sure what the Unnamed: 0 columns are, some have values in a common range, others dont..

Jobs data

The job_id column here is shared with the postings.csv data

In [13]:
```
# Exploring Jobs data
benefits_df       = pd.read_csv("../data/jobs/benefits.csv")
job_industries_df = pd.read_csv("../data/jobs/job_industries.csv")
job_skills_df     = pd.read_csv("../data/jobs/job_skills.csv")
salaries_df       = pd.read_csv("../data/jobs/salaries.csv")

# Creating a list to show all available columns
salaries = list(salaries_df.columns)
benefits = list(benefits_df.columns)
industries = list(job_industries_df.columns)
skills = list(job_skills_df.columns)
print("salaries:  ", salaries)
print("benefits:  ", benefits)
print("industries:", industries)
print("skills:    ", skills)
```

```
salaries:   ['salary_id', 'job_id', 'max_salary', 'med_salary', 'min_salar
y', 'pay_period', 'currency', 'compensation_type']
benefits:   ['job_id', 'inferred', 'type']
industries: ['job_id', 'industry_id']
skills:     ['job_id', 'skill_abr']
```

In [14]:
```
benefits_df.head()
```

Out[14]:

	job_id	inferred	type
0	3887474156	0	Medical insurance
1	3887474156	0	Vision insurance
2	3887474156	0	Dental insurance
3	3884436043	0	Medical insurance
4	3884436043	0	Vision insurance

In [15]:
```
job_industries_df.head()
```

Out[15]:

	job_id	industry_id
0	3887466990	10
1	3887473087	11
2	3887467990	96
3	3887467990	14
4	3884435035	84

In [16]:
```
print(job_industries_df["industry_id"].min(), job_industries_df["industry_ic
```

```
1 3252
```

In [17]:
```
job_skills_df.head()
```

Out[17]:

	job_id	skill_abr
0	3887466990	LGL
1	3887466990	ADM
2	3887473087	MRKT
3	3887473087	SALE
4	3887467990	CNSL

In [18]:
```
salaries_df.head()
```

Out[18]:

	salary_id	job_id	max_salary	med_salary	min_salary	pay_period	currency
0	13	3887473087	80000.0	NaN	75000.0	YEARLY	USD
1	18	3887467990	80.0	NaN	60.0	HOURLY	USD
2	65	3884433143	NaN	53000.0	NaN	YEARLY	USD
3	70	3884428699	300000.0	NaN	90000.0	YEARLY	USD
4	96	3887474156	80000.0	NaN	70000.0	YEARLY	USD

In [19]:
```
salaries_df.describe()
```

Out[19]:

	salary_id	job_id	max_salary	med_salary	min_salary
count	2088.000000	2.088000e+03	1.662000e+03	426.000000	1662.000000
mean	20072.255268	3.889088e+09	9.627357e+04	36351.924624	66036.549212
std	11559.985785	1.786400e+08	9.232996e+04	71459.274156	59313.422769
min	13.000000	2.264355e+06	1.000000e+00	0.000000	1.000000
25%	10139.750000	3.894573e+09	6.500000e+01	19.812500	50.000000
50%	19874.500000	3.901800e+09	9.000000e+04	30.000000	66300.000000
75%	29606.250000	3.904398e+09	1.500000e+05	53810.000000	100000.000000
max	40780.000000	3.906266e+09	1.000001e+06	500000.000000	400000.000000

Mapping data

The industries.csv dataset looks like it has the "industry_id" column in common with job_industries.csv

And the skills.csv dataset looks like it has the "skill_abr" column in common with the job_skills.csv

In [20]:
```
# Exploring Mappings data
industries_df = pd.read_csv("../data/mappings/industries.csv")
skills_df = pd.read_csv("../data/mappings/skills.csv")

# Creating a list to show all available columns
print("Industries: ", list(industries_df.columns))
print("Skills:     ", list(skills_df.columns))
```

```
Industries:  ['industry_id', 'industry_name']
Skills:      ['skill_abr', 'skill_name']
```

In [21]:
```
industries_df.head()
```

Out[21]:

	industry_id	industry_name
0	1	Defense and Space Manufacturing
1	3	Computer Hardware Manufacturing
2	4	Software Development
3	5	Computer Networking Products
4	6	Technology, Information and Internet

In [22]:
```
skills_df.head()
```

Out[22]:

	skill_abr	skill_name
0	ART	Art/Creative
1	DSGN	Design
2	ADVR	Advertising
3	PRDM	Product Management
4	DIST	Distribution

1.2 Design a database schema

Based on the column mappings that I have shown in the diagram below, it looks as though `postings.csv` is definitly the fact_table with links to the other dimension tables via the variables 'company_id' and 'job_id'. The data tables look to be arranged best in a SNOWFLAKE schema, with `postings.csv` at the center as a fact table. The reason this is a SNOWFLAKE schema is because the job `industries.csv` and `job_skills.csv` are linked to other tables, extending the graph relationship to `postings.csv` by more than 1 degree.

```
postings.csv is related to companies.csv,
company_industries.csv, company_specialities.csv and
employee_counts.csv via variable 'company_id'

postings.csv is related to benefits.csv job_industries.csv
job_skills.csv salaries.csv via variable 'job_id'

job_industries.csv is related to industries.csv via variable
'industry_id'

job_skills.csv is related to skills.csv via variable
'skill_abr'
```

The most practical data base scheme is the STAR schema between the `postings.csv` which will act as the fact_table and the `company_industries.csv` that will act as the dim_table

In [23]:
```
# Folder strucutre and columns in each data table is shown below
'''
├── companies
│   ├── companies.csv
│   │     companies:   ['Unnamed: 0', 'company_id', 'name', 'description', '
│   ├── company_industries.csv
│   │     industries:  ['Unnamed: 0', 'company_id', 'industry']
│   ├── company_specialities.csv
│   │     specialties: ['Unnamed: 0', 'company_id', 'speciality']
│   └── employee_counts.csv
│         employees:   ['Unnamed: 0', 'company_id', 'employee_count', 'follo
│
│
├── jobs
│   ├── benefits.csv
│   │     benefits:   ['job_id', 'inferred', 'type']
│   ├── job_industries.csv
│   │     industries: ['job_id', 'industry_id']
│   ├── job_skills.csv
│   │     skills:     ['job_id', 'skill_abr']
│   └── salaries.csv
│         salaries:   ['salary_id', 'job_id', 'max_salary', 'med_salary', 'm
│
│
│
├── mappings
│   ├── industries.csv
│   │     Industries:  ['industry_id', 'industry_name']
│   └── skills.csv
│         Skills:      ['skill_abr', 'skill_name']
│
│
└── postings.csv
        Common variables with other tables: 'company_id', 'job_id'
        postings: ['application_type', 'application_url', 'applies', 'closed
                   'currency', 'description', 'expiry', 'fips', 'formatted_e
                   'listed_time', 'location', 'max_salary', 'med_salary', 'm
```

```
                   'posting_domain', 'remote_allowed', 'skills_desc', 'spons
'''
```

Out[23]:

```
"\n├── companies\n│\xa0\xa0 ├── companies.csv\n│   |     companies:   ['Unnamed: 0', 'company_id', 'name', 'description', 'company_size', 'state', 'country', 'city', 'zip_code', 'address', 'url']\n│\xa0\xa0 ├── company_industries.csv\n│   |     industries:  ['Unnamed: 0', 'company_id', 'industry']\n│\xa0\xa0 ├── company_specialities.csv\n│   |     specialties: ['Unnamed: 0', 'company_id', 'speciality']\n│\xa0\xa0 └── employee_counts.csv\n│         employees:   ['Unnamed: 0', 'company_id', 'employee_count', 'follower_count', 'time_recorded']\n│\n│\n├── jobs\n│\xa0\xa0 ├── benefits.csv\n│   |     benefits:   ['job_id', 'inferred', 'type']\n│\xa0\xa0 ├── job_industries.csv\n│   |     industries: ['job_id', 'industry_id']\n│\xa0\xa0 ├── job_skills.csv\n│   |     skills:     ['job_id', 'skill_abr']\n│\xa0\xa0 └── salaries.csv\n│         salaries:   ['salary_id', 'job_id', 'max_salary', 'med_salary', 'min_salary', 'pay_period', 'currency', 'compensation_type']\n│\n│\n│\n├── mappings\n│\xa0\xa0 ├── industries.csv\n│   |     Industries:  ['industry_id', 'industry_name']\n│\xa0\xa0 └── skills.csv\n│         Skills:      ['skill_abr', 'skill_name']\n│\n│\n└── postings.csv\n        Common variables with other tables: 'company_id', 'job_id'\n        postings: ['application_type', 'application_url', 'applies', 'closed_time', 'company_id', 'company_name', 'compensation_type', \n                   'currency', 'description', 'expiry', 'fips', 'formatted_experience_level', 'formatted_work_type', 'job_id', 'job_posting_url', \n                   'listed_time', 'location', 'max_salary', 'med_salary', 'min_salary', 'normalized_salary', 'original_listed_time', 'pay_period', \n                   'posting_domain', 'remote_allowed', 'skills_desc', 'sponsored', 'title', 'views', 'work_type', 'zip_code']\n"
```

1.3 Create and load a local database

Two tables are loaded into a sqlite database called `job_postings.db`

`postings.csv` as `fact_job_postings` and `company_industries.csv` as `dim_company`

In [24]:

```
# Allows for displaying the sql queries
prettytable.DEFAULT = 'DEFAULT'
```

In [25]:

```
# Connecting to an existing database, or creating it if it does not exist ye
conn = sqlite3.connect("job_postings.db")

# Allows for querying using sql
cursor = conn.cursor()

# Allows for using magic statements within sql
%load_ext sql

# Creating/loading a database called job_postings.sb
%sql sqlite:///job_postings.db
```

In [26]:

```
# Reading the fact and dim table into memory using pandas
fact_job_postings_df = pd.read_csv("../data/postings.csv")
```

```
dim_company_df = pd.read_csv("../data/companies/company_industries.csv")
```

In [27]:

```
# Converting the dataframes to sql tables, linking them to job_postings.db
fact_job_postings_df.to_sql("fact_job_postings", conn, if_exists='replace',
dim_company_df.to_sql("dim_company", conn, if_exists='replace', index=False,
```

Out[27]: 1432

In [28]:

```
# What info is in the fact table again?
%sql PRAGMA table_info("fact_job_postings")
```

```
 * sqlite:///job_postings.db
Done.
```

Out[28]:

cid	name	type	notnull	dflt_value	pk
0	job_id	INTEGER	0	None	0
1	company_name	TEXT	0	None	0
2	title	TEXT	0	None	0
3	description	TEXT	0	None	0
4	max_salary	REAL	0	None	0
5	pay_period	TEXT	0	None	0
6	location	TEXT	0	None	0
7	company_id	REAL	0	None	0
8	views	REAL	0	None	0
9	med_salary	REAL	0	None	0
10	min_salary	REAL	0	None	0
11	formatted_work_type	TEXT	0	None	0
12	applies	REAL	0	None	0
13	original_listed_time	REAL	0	None	0
14	remote_allowed	REAL	0	None	0
15	job_posting_url	TEXT	0	None	0
16	application_url	TEXT	0	None	0
17	application_type	TEXT	0	None	0
18	expiry	REAL	0	None	0
19	closed_time	REAL	0	None	0
20	formatted_experience_level	TEXT	0	None	0
21	skills_desc	TEXT	0	None	0
22	listed_time	REAL	0	None	0
23	posting_domain	TEXT	0	None	0
24	sponsored	INTEGER	0	None	0
25	work_type	TEXT	0	None	0
26	currency	TEXT	0	None	0
27	compensation_type	TEXT	0	None	0
28	normalized_salary	REAL	0	None	0
29	zip_code	REAL	0	None	0
30	fips	REAL	0	None	0

In [29]:
```
%sql PRAGMA table_info("dim_company")
```

```
 * sqlite:///job_postings.db
Done.
```

Out[29]:

cid	name	type	notnull	dflt_value	pk
0	Unnamed: 0	INTEGER	0	None	0
1	company_id	INTEGER	0	None	0
2	industry	TEXT	0	None	0

1.4 Use your database to answer some questions

How many companies have more than 1 job posting?

In [30]:
```
%%sql
SELECT COUNT(count) as `Companies with > 1 job postings` FROM (SELECT compar
WHERE count > 1 ;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[30]:

Companies with > 1 job postings
601

In [31]:
```
%%sql
SELECT comp AS Company, count AS `Num Job Postings` FROM (SELECT company_nam
WHERE count > 1
ORDER BY count DESC
LIMIT 10
;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[31]:

Company	Num Job Postings
Family Dollar	288
Talentify.io	276
Rent-A-Center	136
National Staffing Solutions	134
AutoZone	131
Claire's	130
Sutter Health	120
Johnson & Johnson	108
Revature	103
LanceSoft, Inc.	95

How many job postings are there for each job industry?

This question requires me to join tables so I can use the industry type, the dim table

In [32]:

```
%%sql
SELECT industry AS Industry, COUNT(job_id) AS `Num Postings` FROM (fact_job_
GROUP BY industry
ORDER BY COUNT(job_id) DESC;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[32]:

Industry	Num Postings
Hospitals and Health Care	1010
Retail	913
Staffing and Recruiting	803
IT Services and IT Consulting	762
Software Development	489
Entertainment Providers	211
Insurance	156
Higher Education	143
Construction	126
Hospitality	106
Defense and Space Manufacturing	106
Financial Services	102
Business Consulting and Services	86
Food and Beverage Manufacturing	83
Pharmaceutical Manufacturing	80
Non-profit Organizations	77
Advertising Services	77
Food and Beverage Services	76
Real Estate	74
Telecommunications	72
Design Services	71
Manufacturing	67
Environmental Services	66
Government Administration	64
Motor Vehicle Manufacturing	62
Wellness and Fitness Services	58
Biotechnology Research	55
Truck Transportation	49
Oil and Gas	47
Law Practice	47
Individual and Family Services	42
Medical Equipment Manufacturing	39

Industry	Num Postings
Mental Health Care	37
Mining	35
Aviation and Aerospace Component Manufacturing	33
Retail Apparel and Fashion	31
Personal Care Product Manufacturing	31
Airlines and Aviation	30
Wholesale Building Materials	29
Industrial Machinery Manufacturing	29
Furniture and Home Furnishings Manufacturing	28
Packaging and Containers Manufacturing	26
Human Resources Services	24
Primary and Secondary Education	23
Restaurants	21
Retail Groceries	20
Civil Engineering	20
Banking	20
Outsourcing and Offshoring Consulting	19
Book and Periodical Publishing	19
Utilities	18
Armed Forces	18
Semiconductor Manufacturing	16
Security and Investigations	16
Plastics Manufacturing	14
Wholesale	13
Research Services	13
Education Administration Programs	13
Renewable Energy Semiconductor Manufacturing	12
Glass, Ceramics and Concrete Manufacturing	12
Textile Manufacturing	11
Architecture and Planning	11
Accounting	10
Information Services	9

Industry	Num Postings
Chemical Manufacturing	9
Appliances, Electrical, and Electronics Manufacturing	9
Transportation, Logistics, Supply Chain and Storage	8
Public Relations and Communications Services	8
Machinery Manufacturing	8
Gambling Facilities and Casinos	8
Farming	8
Facilities Services	8
Computer and Network Security	8
Venture Capital and Private Equity Principals	7
Travel Arrangements	7
Professional Training and Coaching	7
Fundraising	7
Retail Office Equipment	6
Medical Practices	6
Consumer Services	6
Automation Machinery Manufacturing	6
Computer Hardware Manufacturing	5
Broadcast Media Production and Distribution	5
Technology, Information and Internet	4
Retail Luxury Goods and Jewelry	4
Public Safety	4
Legal Services	4
Civic and Social Organizations	4
Beverage Manufacturing	4
Spectator Sports	3
Religious Institutions	3
Political Organizations	3
Paper and Forest Product Manufacturing	3
Investment Management	3
Computers and Electronics Manufacturing	3
Translation and Localization	2

Industry	Num Postings
Shipbuilding	2
Recreational Facilities	2
Public Policy Offices	2
Printing Services	2
Photography	2
Online Audio and Video Media	2
Museums, Historical Sites, and Zoos	2
Media Production	2
E-Learning Providers	2
Animation and Post-production	2
Writing and Editing	1
Tobacco Manufacturing	1
Sporting Goods Manufacturing	1
Performing Arts	1
Nanotechnology Research	1
Libraries	1
Graphic Design	1
Government Relations Services	1

What is the average normalized salary by company industry?

In [33]:
```
%%sql
SELECT industry AS Industry, AVG(normalized_salary) AS `Avg. Norm. Salary` F
GROUP BY industry
ORDER BY `Avg. Norm. Salary` DESC;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[33]:

Industry	Avg. Norm. Salary
Information Services	250000.0
Investment Management	225000.0
Automation Machinery Manufacturing	195900.0
Semiconductor Manufacturing	180000.0
Biotechnology Research	164804.125
Online Audio and Video Media	159500.0
Entertainment Providers	153425.15569620254
Venture Capital and Private Equity Principals	149366.66666666666
Personal Care Product Manufacturing	138401.95789473684
Defense and Space Manufacturing	136776.82222222222
Beverage Manufacturing	130000.0
Computer and Network Security	126875.0
Wholesale	125000.0
Technology, Information and Internet	120625.0
Computers and Electronics Manufacturing	115000.0
Staffing and Recruiting	114820.28340298506
Motor Vehicle Manufacturing	113653.75
IT Services and IT Consulting	112549.90168539326
Financial Services	111910.00733333334
Hospitals and Health Care	110733.30331560284
Farming	110500.0
Transportation, Logistics, Supply Chain and Storage	108234.5
Medical Equipment Manufacturing	107730.28571428571
Design Services	107644.05
Renewable Energy Semiconductor Manufacturing	107500.0
Software Development	107205.14064705883
Legal Services	106150.0
Civil Engineering	105416.66666666667
Business Consulting and Services	105215.33333333333
Architecture and Planning	105166.66666666667
Translation and Localization	105000.0
Utilities	104752.0

Industry	Avg. Norm. Salary
Research Services	102316.66666666667
Construction	100457.14285714286
Wellness and Fitness Services	98956.19724137931
Broadcast Media Production and Distribution	98800.0
Advertising Services	96501.85882352942
Telecommunications	96320.10806451613
Nanotechnology Research	95679.5
Pharmaceutical Manufacturing	93761.76470588235
Accounting	92795.0
Oil and Gas	92000.0
Retail Luxury Goods and Jewelry	91400.0
Aviation and Aerospace Component Manufacturing	90513.33333333333
Professional Training and Coaching	89637.59999999999
Law Practice	89623.23529411765
Public Policy Offices	88975.0
Airlines and Aviation	86800.85
Insurance	85807.7392857143
Medical Practices	85000.0
Manufacturing	84114.27857142857
Primary and Secondary Education	83948.625
Public Relations and Communications Services	83750.0
Banking	81250.0
Human Resources Services	79988.0
Retail Apparel and Fashion	78409.3090909091
Education Administration Programs	77738.3
Museums, Historical Sites, and Zoos	76250.0
Real Estate	75829.52380952382
Gambling Facilities and Casinos	75000.0
Travel Arrangements	74992.5
Environmental Services	74843.84615384616
Mining	74173.0
Government Administration	73318.85727272728

Industry	Avg. Norm. Salary
Food and Beverage Manufacturing	70589.8303030303
Chemical Manufacturing	70300.0
Food and Beverage Services	70274.94736842105
Photography	70000.0
Media Production	70000.0
Packaging and Containers Manufacturing	69684.5
Retail	69041.95952380952
Mental Health Care	67299.80725
Hospitality	66920.53333333334
Wholesale Building Materials	65030.56
E-Learning Providers	62400.0
Higher Education	60690.291445783136
Truck Transportation	60056.23684210526
Individual and Family Services	58858.89421052631
Non-profit Organizations	57630.561285714284
Tobacco Manufacturing	57500.0
Paper and Forest Product Manufacturing	55806.399999999994
Book and Periodical Publishing	55218.0
Shipbuilding	50928.8
Security and Investigations	48633.06285714286
Textile Manufacturing	48477.5
Spectator Sports	46800.0
Industrial Machinery Manufacturing	45612.8
Facilities Services	45009.333333333336
Retail Office Equipment	44990.4
Restaurants	44373.333333333336
Plastics Manufacturing	43680.0
Sporting Goods Manufacturing	43500.0
Furniture and Home Furnishings Manufacturing	43153.333333333336
Consumer Services	42293.333333333336
Glass, Ceramics and Concrete Manufacturing	39520.0
Fundraising	39520.0

Industry	Avg. Norm. Salary
Computer Hardware Manufacturing	32586.666666666668
Retail Groceries	29120.0
Graphic Design	29120.0
Political Organizations	5250.0
Religious Institutions	4200.0
Writing and Editing	None
Recreational Facilities	None
Public Safety	None
Printing Services	None
Performing Arts	None
Outsourcing and Offshoring Consulting	None
Machinery Manufacturing	None
Libraries	None
Government Relations Services	None
Civic and Social Organizations	None
Armed Forces	None
Appliances, Electrical, and Electronics Manufacturing	None
Animation and Post-production	None

Name the top 5 companies with the highest average normalized salary for their job postings

In [34]:

```
%%sql
SELECT company_name AS Company, AVG(normalized_salary) AS `Avg. Norm Salary`
GROUP BY company_name
ORDER BY `Avg. Norm Salary` DESC
LIMIT 5;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[34]:

Company	Avg. Norm Salary
Woodside Staffing Solutions & Consulting	337500.0
Calm	337500.0
Health eCareers	337246.4090909091
Buck Institute for Research on Aging	300000.0
Spire Orthopedic Partners	284124.0

In [ ]:

Verifying the averages, they seem extremely high

seems like there is only 1 postings a lot of the time, so the average is the posted value, seems reasonable

In [35]:

```
%%sql
SELECT company_name AS Company, normalized_salary FROM fact_job_postings
WHERE company_name='Woodside Staffing Solutions & Consulting'
ORDER BY company_name;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[35]:

Company	normalized_salary
Woodside Staffing Solutions & Consulting	337500.0

In [36]:

```
%%sql
SELECT company_name AS Company, normalized_salary FROM fact_job_postings
WHERE company_name='Calm'
ORDER BY company_name;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[36]:

Company	normalized_salary
Calm	337500.0

In [37]:

```
%%sql
SELECT company_name AS Company, AVG(normalized_salary) FROM fact_job_posting
WHERE company_name='Health eCareers'
ORDER BY company_name;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[37]:

Company	AVG(normalized_salary)
Health eCareers	337246.4090909091

In [38]:

```
%%sql
SELECT company_name AS Company, normalized_salary FROM fact_job_postings
WHERE company_name='Buck Institute for Research on Aging'
ORDER BY company_name;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[38]:

Company	normalized_salary
Buck Institute for Research on Aging	300000.0

In [39]:

```
%%sql
SELECT company_name AS Company, normalized_salary FROM fact_job_postings
WHERE company_name='Spire Orthopedic Partners'
ORDER BY company_name;
```

```
 * sqlite:///job_postings.db
Done.
```

Out[39]:

Company	normalized_salary
Spire Orthopedic Partners	450000.0
Spire Orthopedic Partners	118248.0

In [ ]: