Project 1 - Group 2

Job Market Analysis for Data Jobs

Members:

Colin

Heran

Kevin

Paola

Motivation & Summary

- Love for data
- Analytics skills are needed now
- Predict our future - post bootcamp
- Most valuable skills and if we are learning them
- Where the jobs are located

- Hypothesis :
  - Higher number of skills correlated with higher pay

    (Were we able to answer it?)

Main Questions & Data used

Data used: Indeed Dataset

- What are the top skills required for Data jobs?
- Is higher number of skills correlated with higher pay?
- What are the most valuable skills?
- Is having more skills better?

API: Google Maps

- What are the top locations?
- What locations pay the most?

Data used: US Bureau of Labor Statistics

- What is the pay premium for having a Data job?

Data used: Cost of Living Index

- Is there a pay premium that results from having these data skills?

Job Market Status for Data Jobs - The Process

Cleanup

1. Library & packages used
   - Pandas, Matplotlib
2. Import CSV
3. Reformat and clean:
   - Dropna, dropped blank skills, dropped remote/USA locations
   - Filters
   - Replace
   - Rename
   - Merge

Analysis

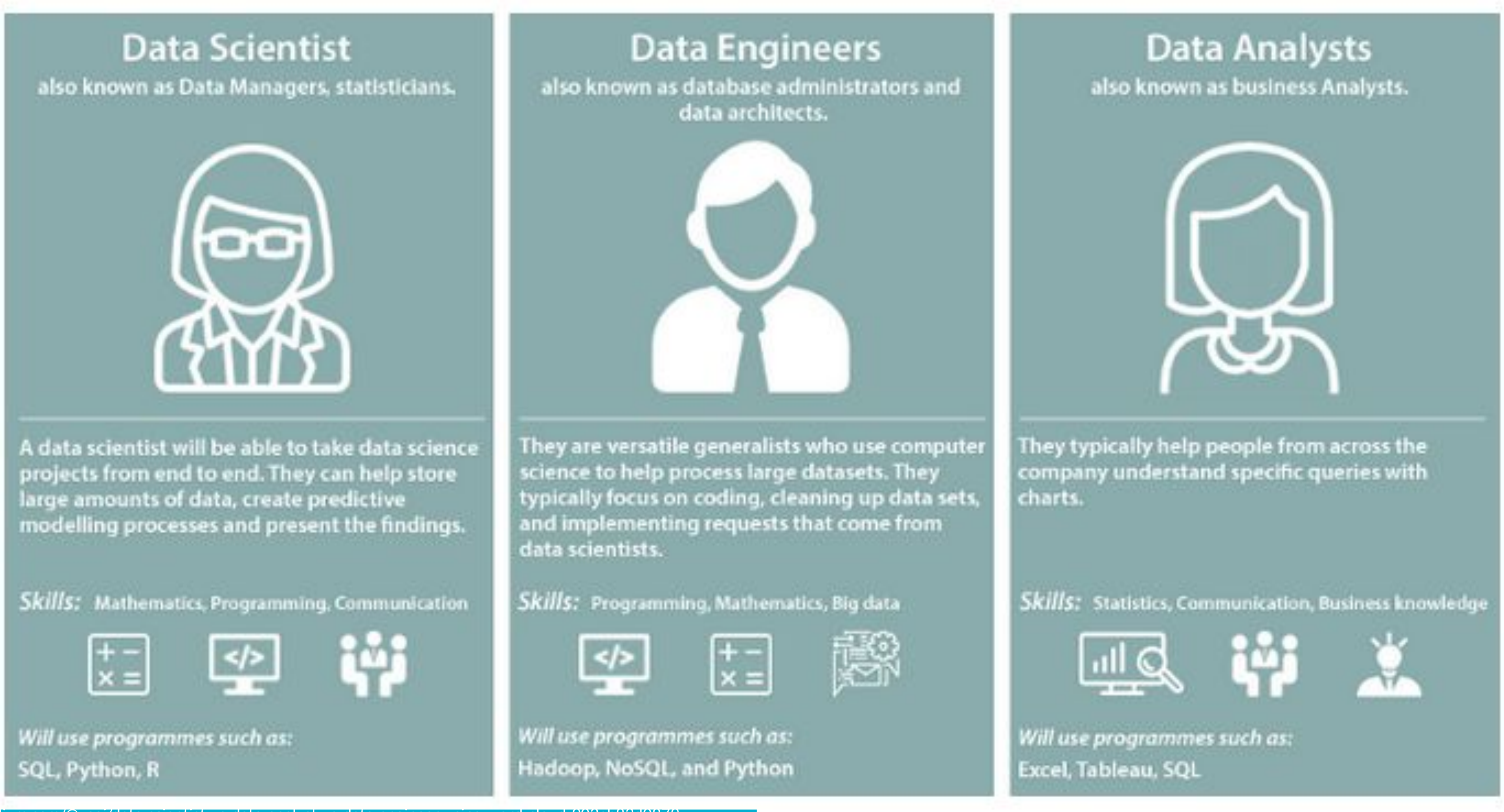
1. Present overview of data in new dataframe
   - Summary statistics - Mean, std, min and max
   - Functions - count, value_counts, dtypes, sum
   - Select data - iloc, loc
   - Chi-squared test
   - Cramer’s V
   - Multivariate Regression

Visualization

1. Make visualizations for key data points
   - Bar plots
   - Pie charts
   - Maps

Data Analysis

What is the difference between ta scientist, engineer, and yst?



Data Scientist

also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:

SQL, Python, R

Data Engineers

also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:

Hadoop, NoSQL, and Python

Data Analysts

also known as business Analysts.

They typically help people from across the company understand specific queries with charts.

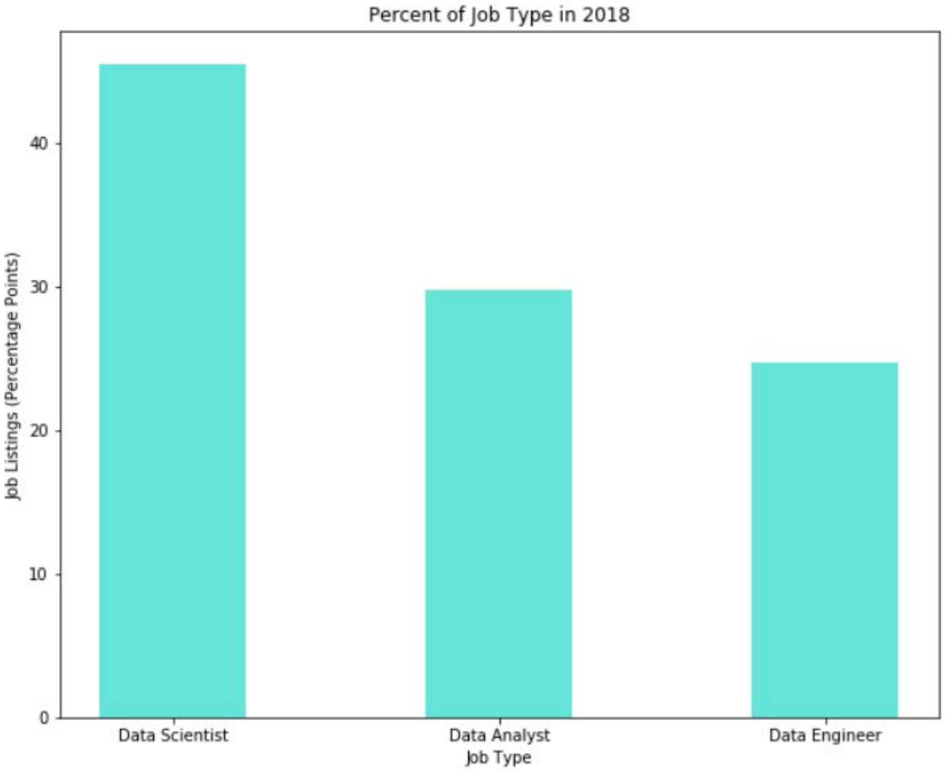
Skills: Statistics, Communication, Business knowledge

Will use programmes such as:

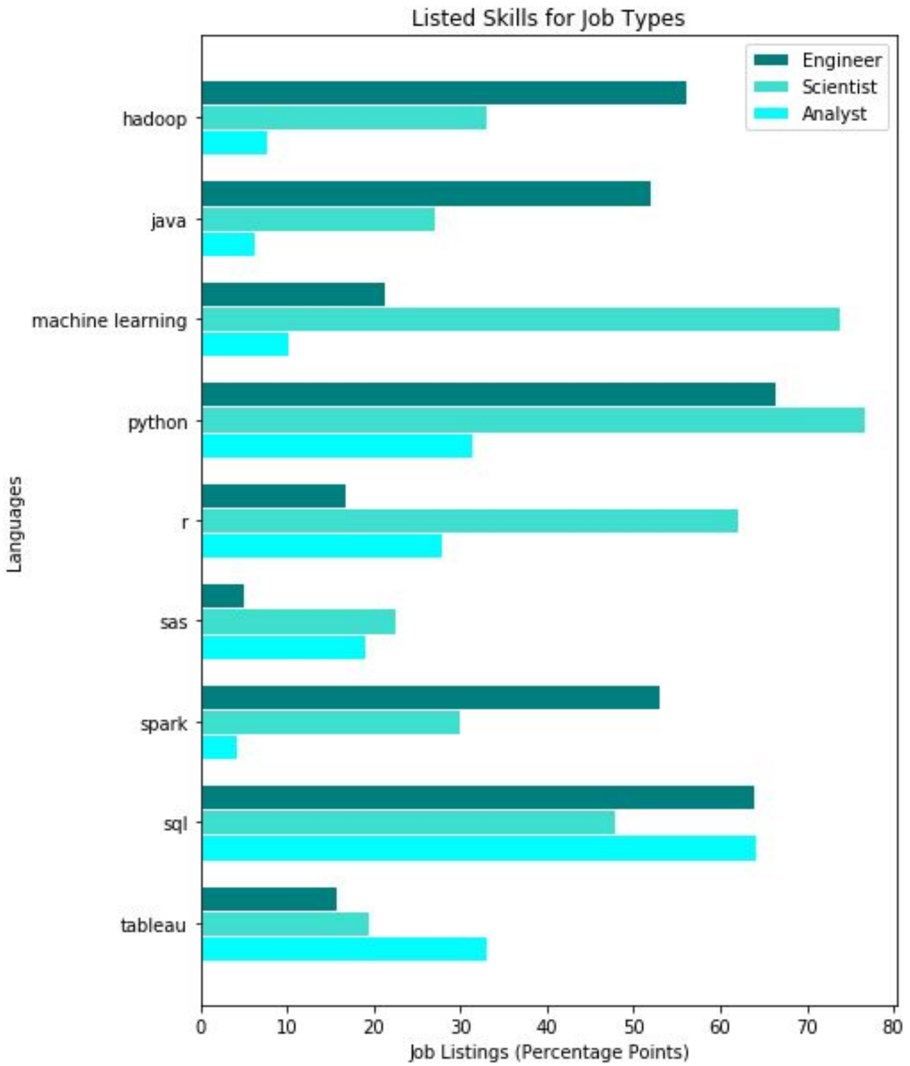
Excel, Tableau, SQL

Image source: https://medium.com/@vegi/data-scientist-vs-data-analyst-vs-data-engineer-using-word-cloud-902ab83d0879

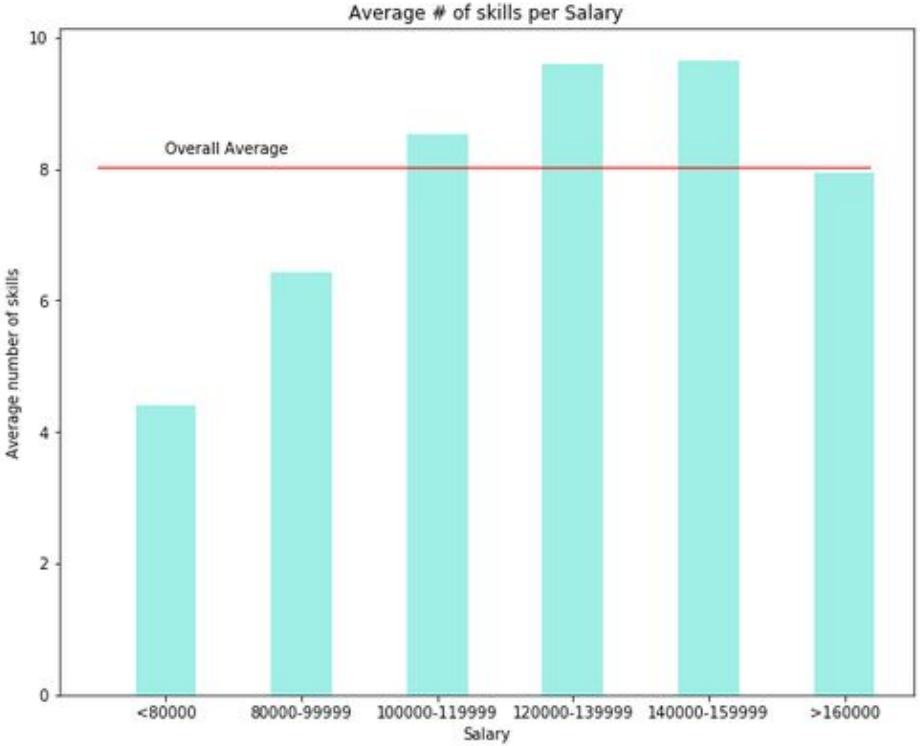
Which job types are the most listed in 2018?



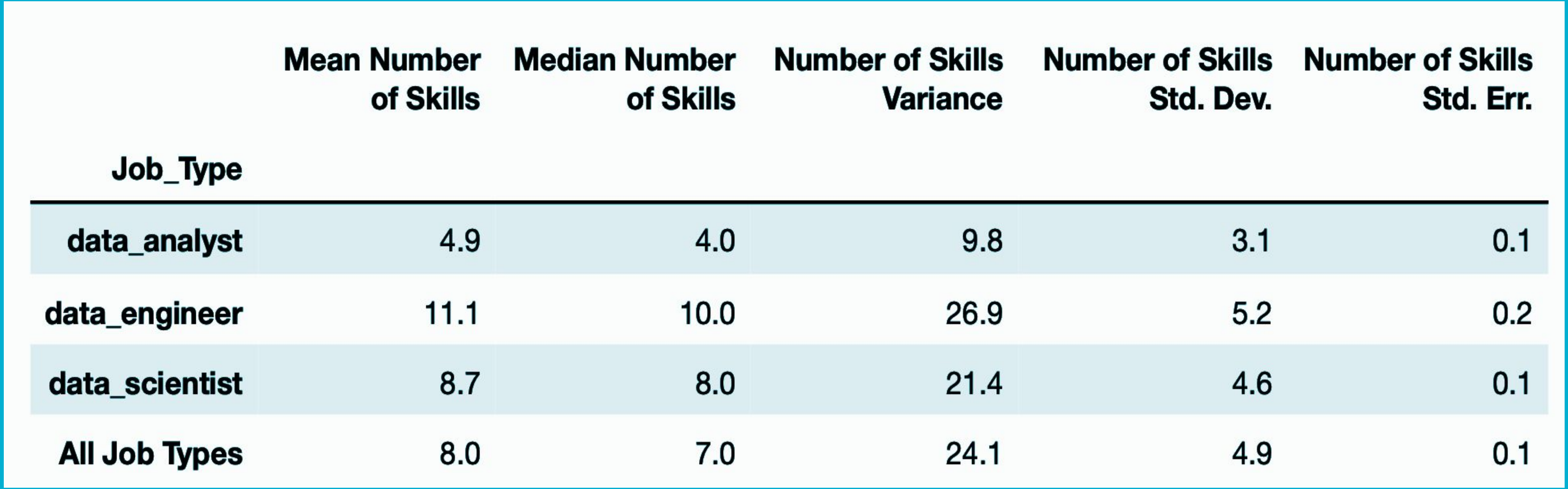
What are the top skills requested for each job type?



Does knowing more languages increase your salary?

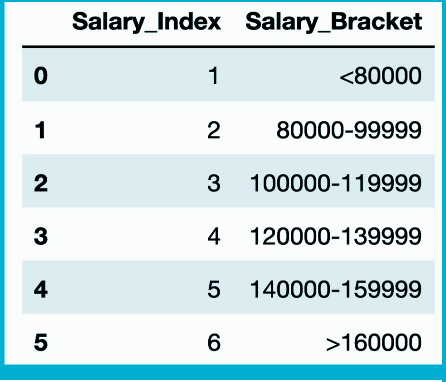


Inferential Statistics
Number of Skills vs Job Types

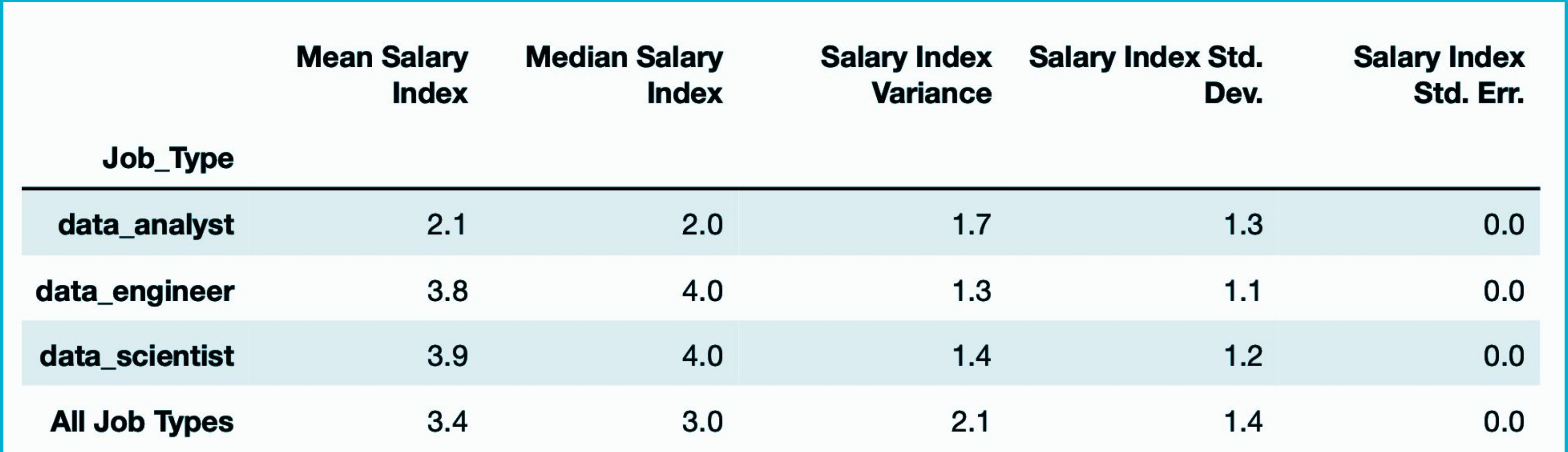


	Mean Number of Skills	Median Number of Skills	Number of Skills Variance	Number of Skills Std. Dev.	Number of Skills Std. Err.
Job_Type					
data_analyst	4.9	4.0	9.8	3.1	0.1
data_engineer	11.1	10.0	26.9	5.2	0.2
data_scientist	8.7	8.0	21.4	4.6	0.1
All Job Types	8.0	7.0	24.1	4.9	0.1

Inferential Statistics
Salary vs Job Types

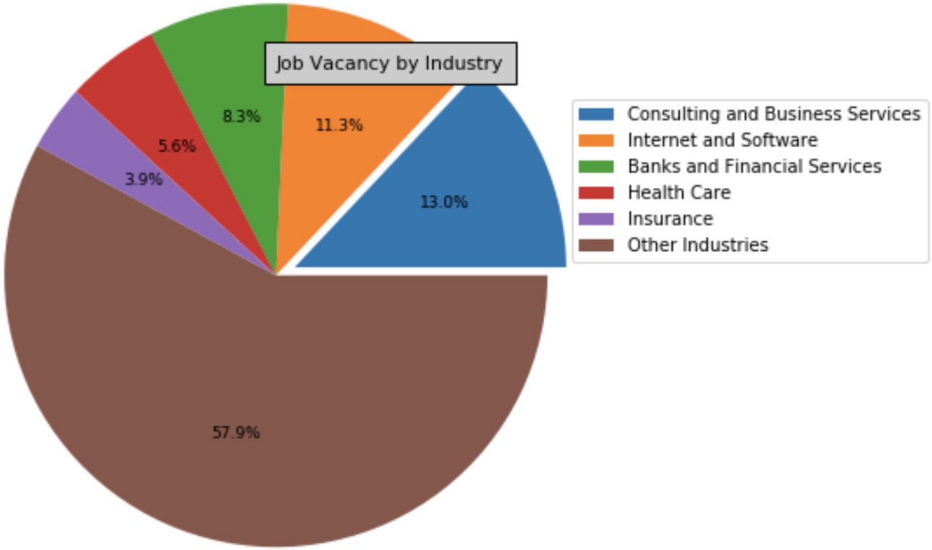


	Salary_Index	Salary_Bracket
0	1	<80000
1	2	80000-99999
2	3	100000-119999
3	4	120000-139999
4	5	140000-159999
5	6	>160000

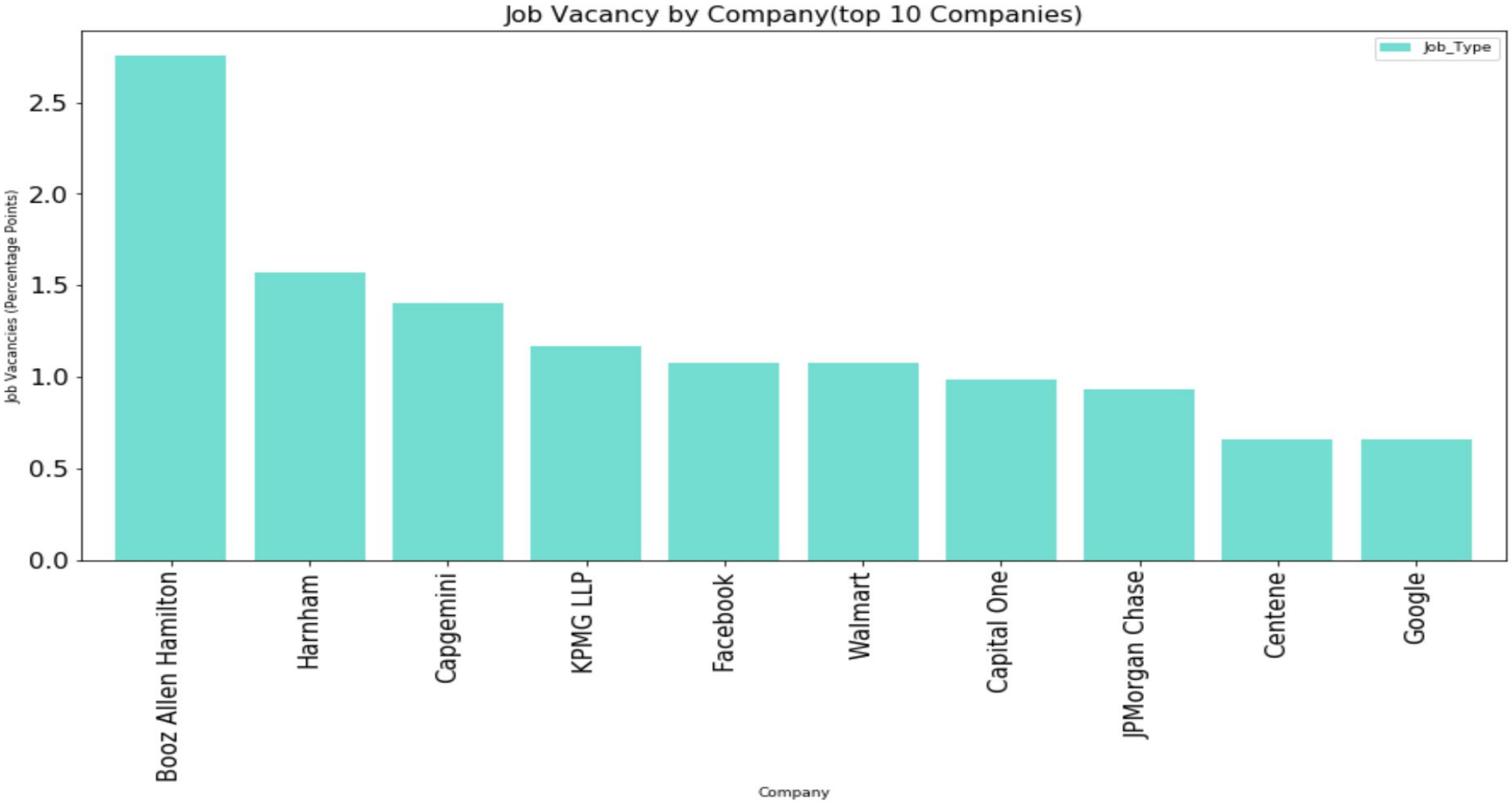


	Mean Salary Index	Median Salary Index	Salary Index Variance	Salary Index Std. Dev.	Salary Index Std. Err.
Job_Type					
data_analyst	2.1	2.0	1.7	1.3	0.0
data_engineer	3.8	4.0	1.3	1.1	0.0
data_scientist	3.9	4.0	1.4	1.2	0.0
All Job Types	3.4	3.0	2.1	1.4	0.0

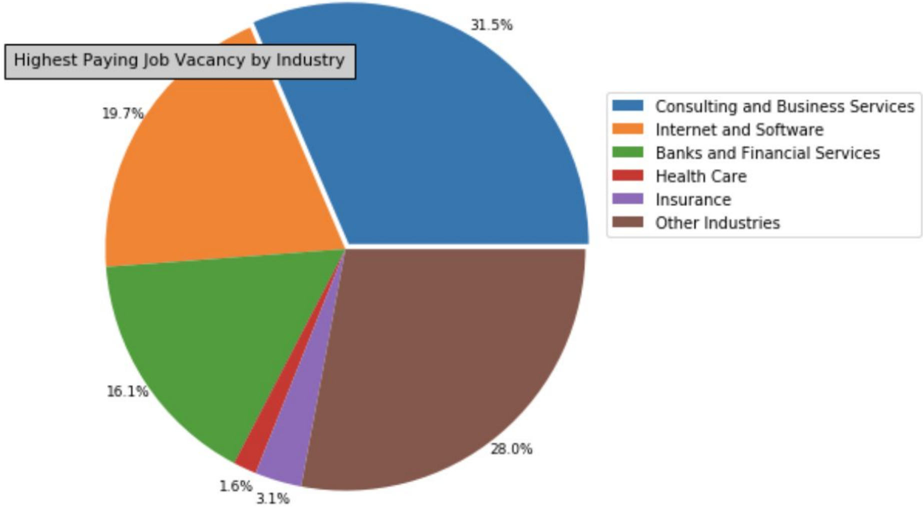
Which industries have the most job vacancies?



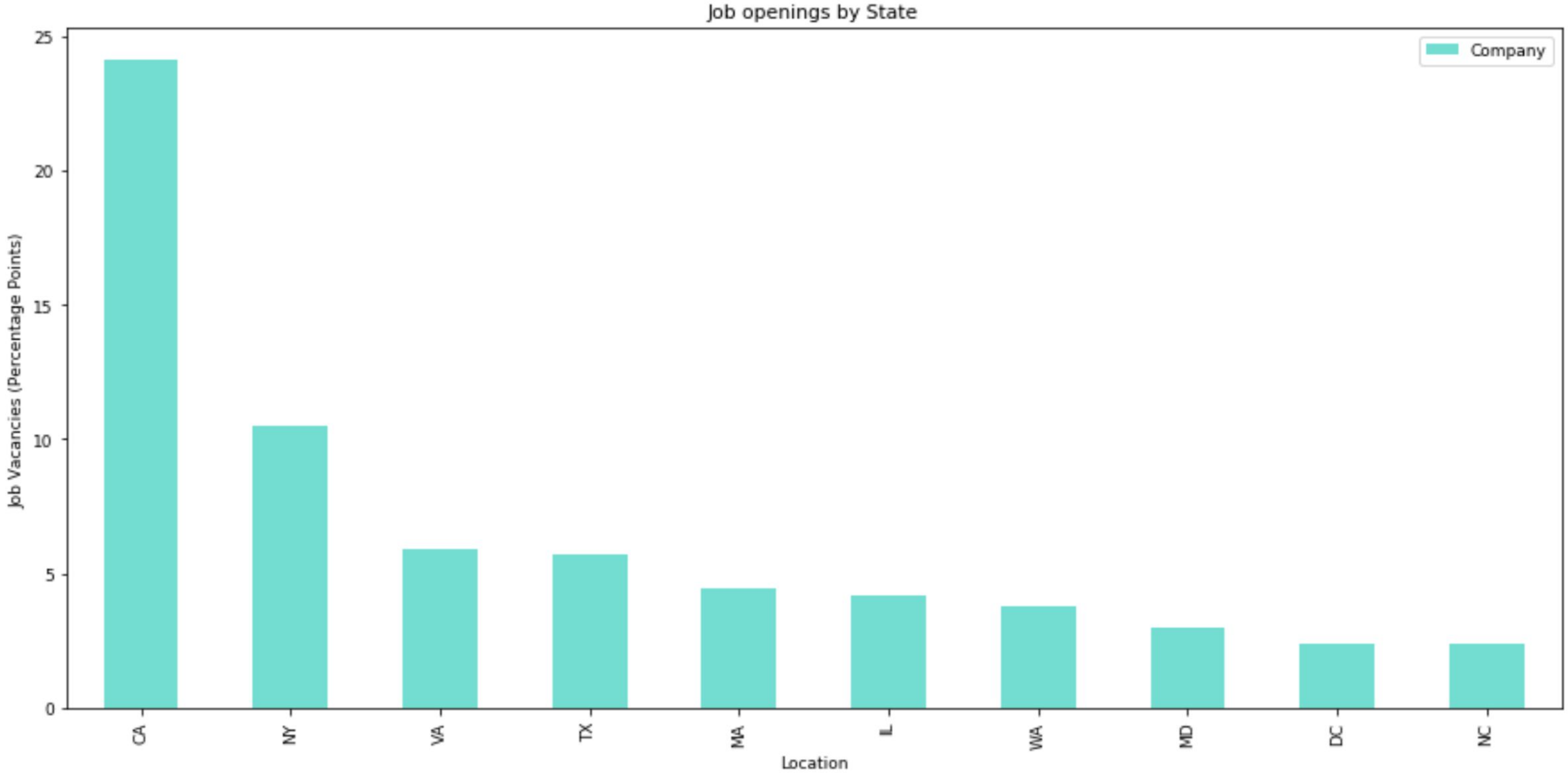
Which companies have the most job listings?



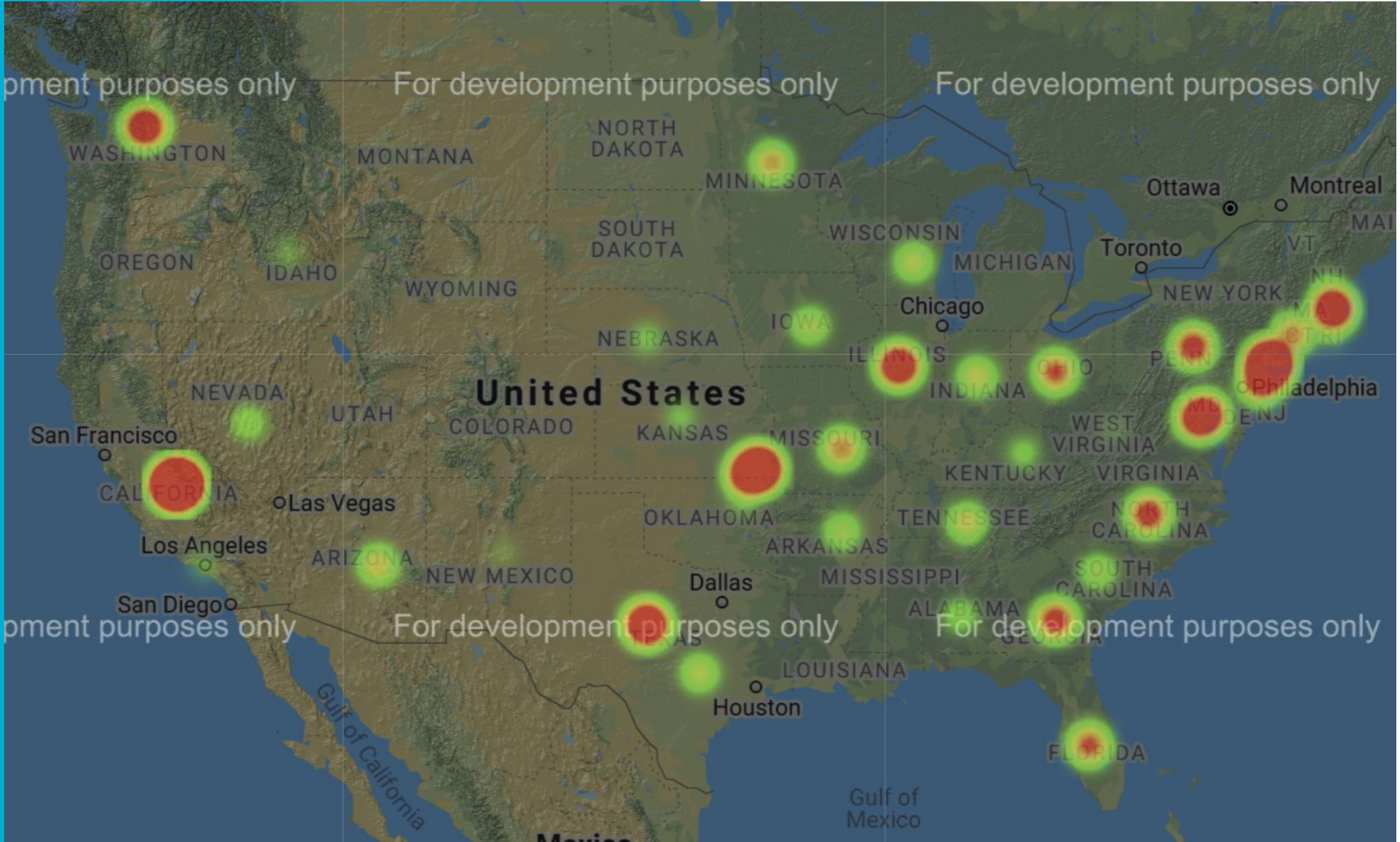
Highest Paying Industries



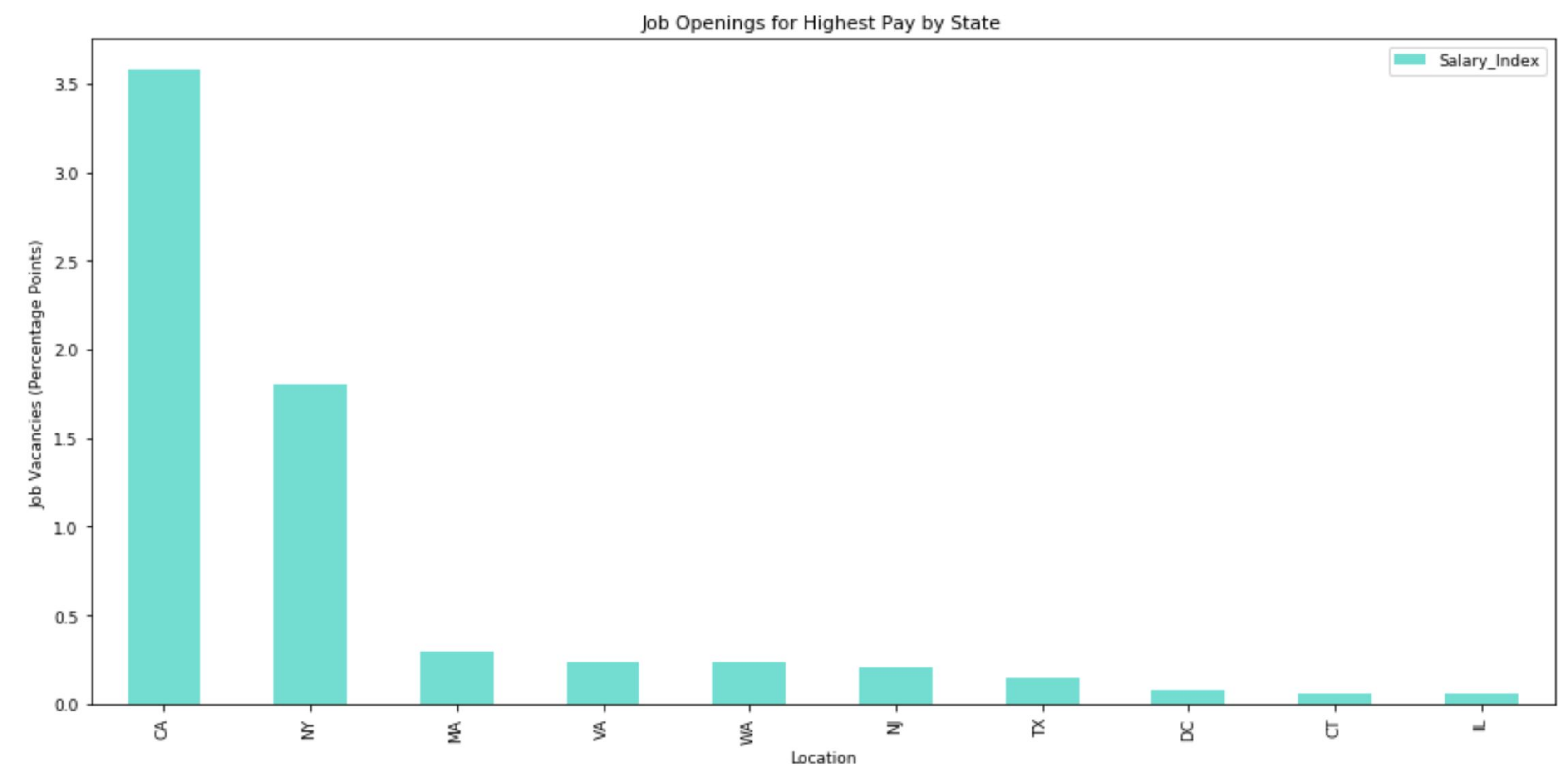
Where are the jobs located?



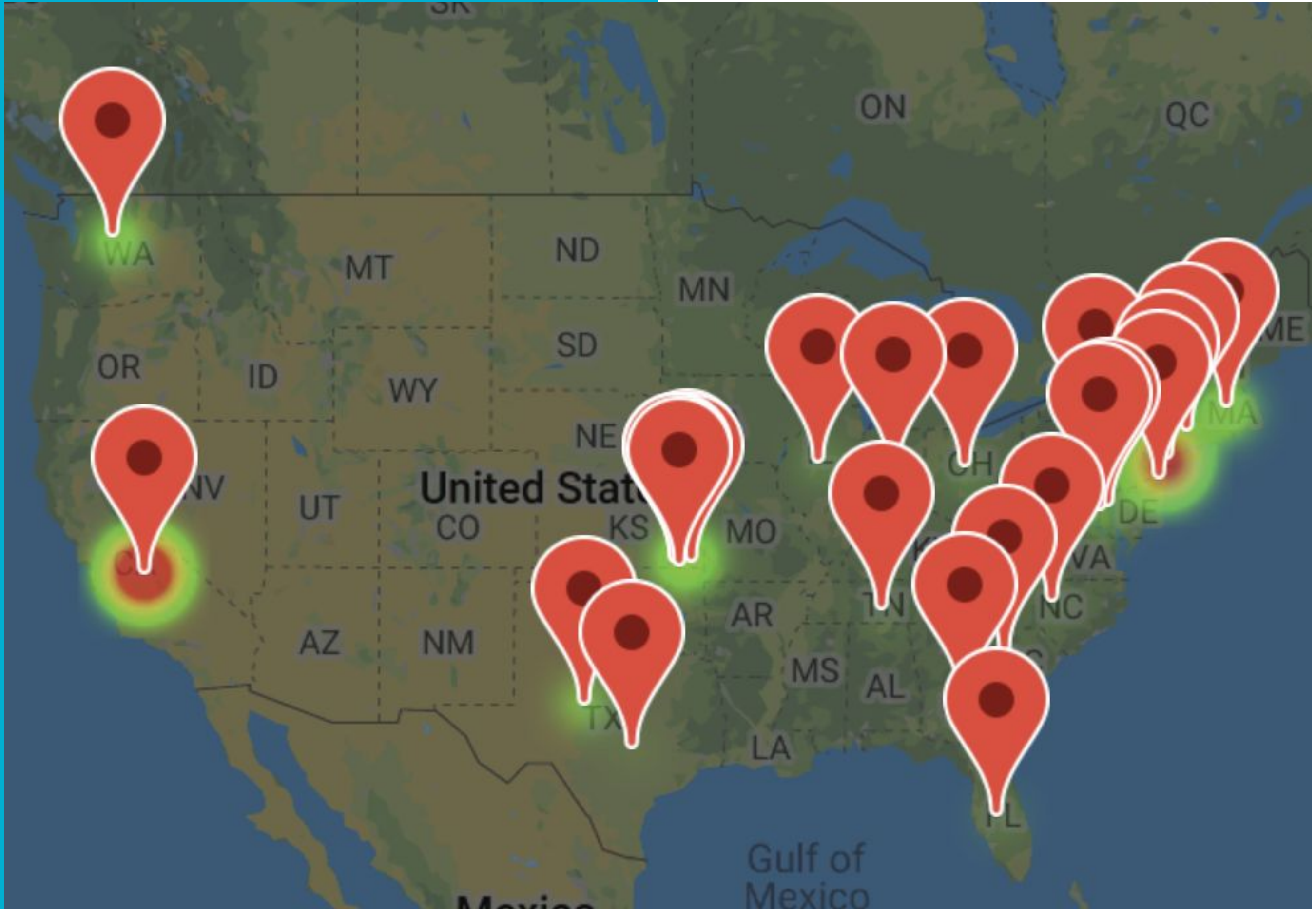
Where are the jobs?



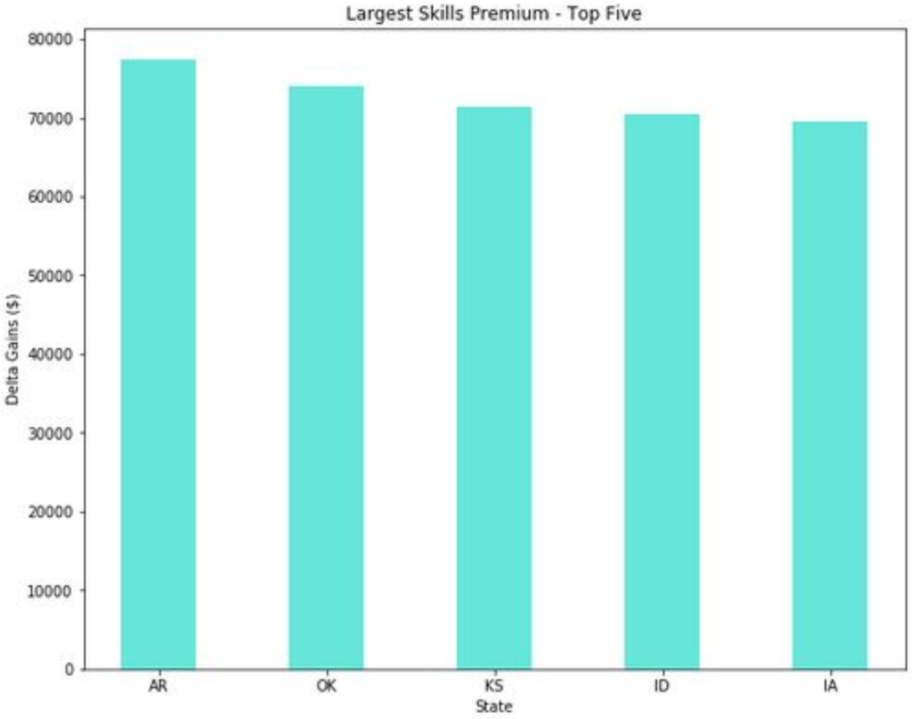
Where are the highest paying jobs?



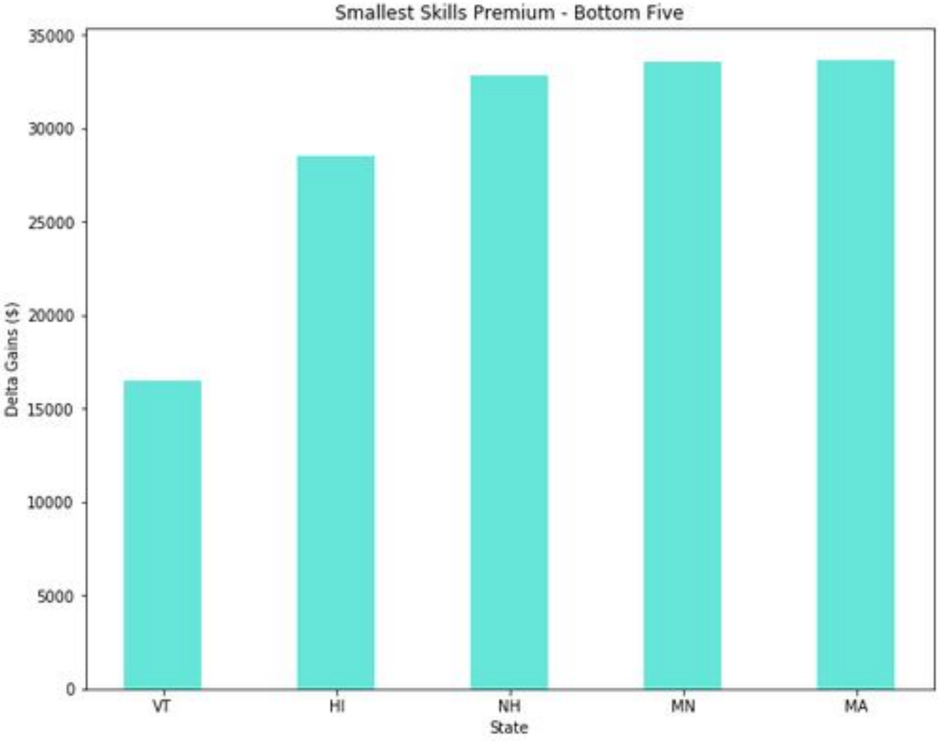
Where are the highest paying jobs?



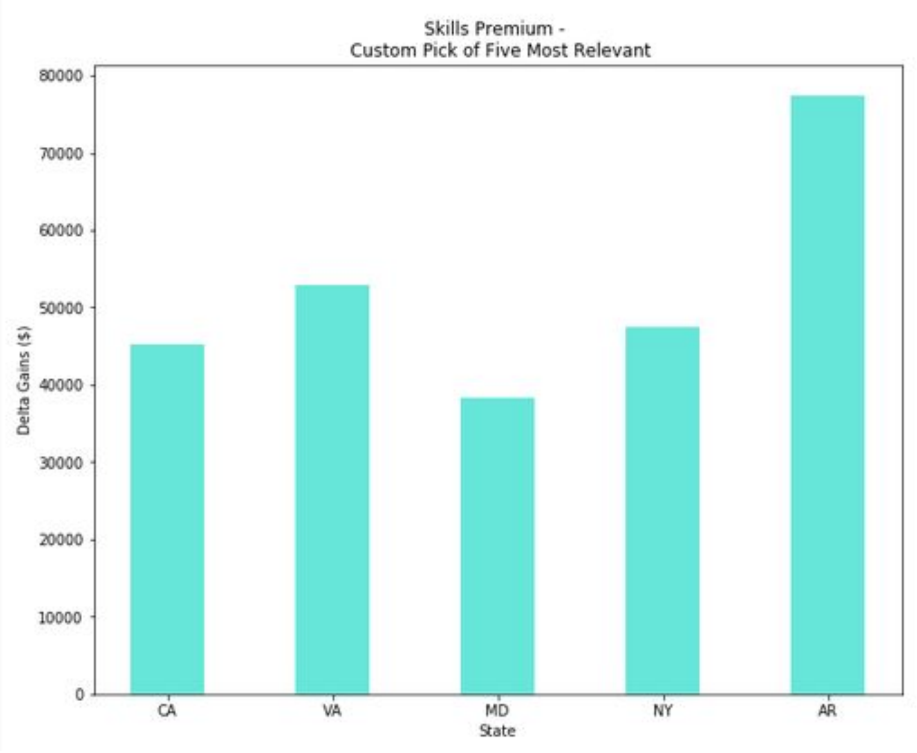
Premium 1: Largest Premium - Top Five States



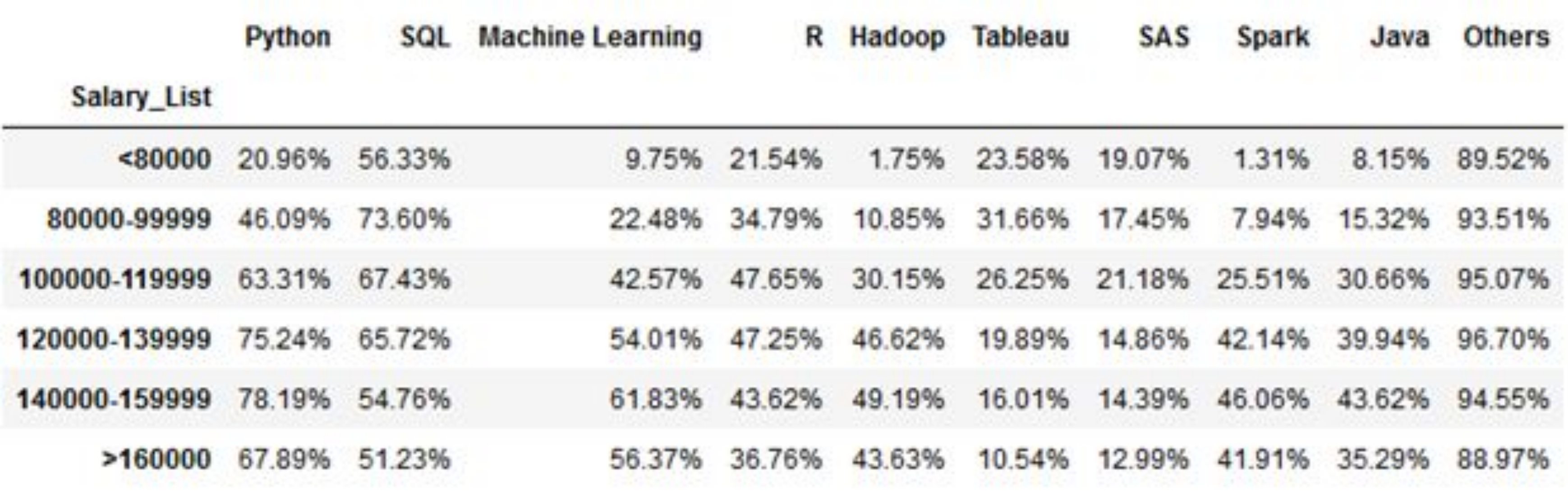
Premium 2: Smallest Premium - Bottom Five States



Premium 3: Custom Pick of Relevant States

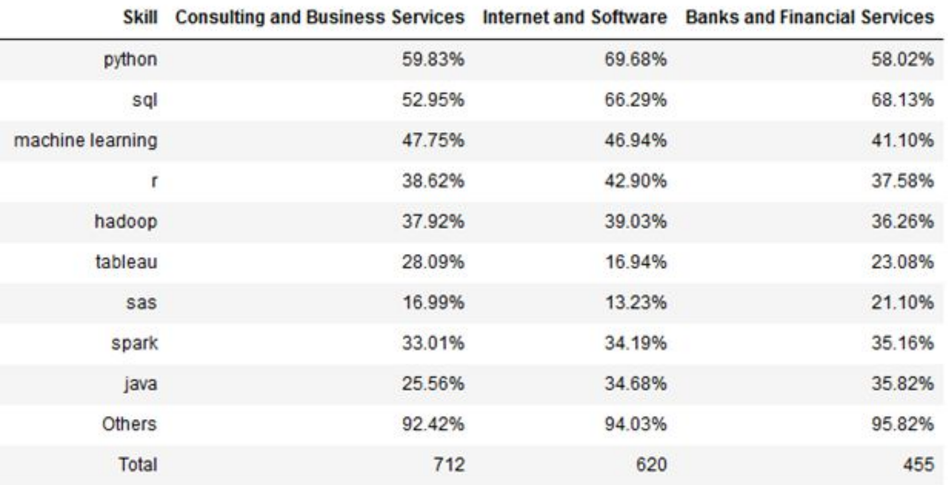


What skills are most in-demand for each income bracket?

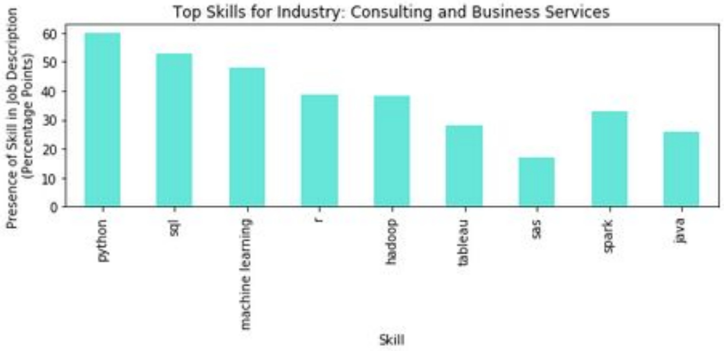


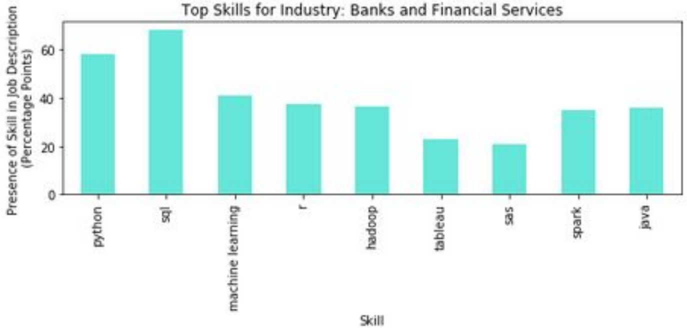
	Python	SQL	Machine Learning	R	Hadoop	Tableau	SAS	Spark	Java	Others
Salary_List										
<80000	20.96%	56.33%	9.75%	21.54%	1.75%	23.58%	19.07%	1.31%	8.15%	89.52%
80000-99999	46.09%	73.60%	22.48%	34.79%	10.85%	31.66%	17.45%	7.94%	15.32%	93.51%
100000-119999	63.31%	67.43%	42.57%	47.65%	30.15%	26.25%	21.18%	25.51%	30.66%	95.07%
120000-139999	75.24%	65.72%	54.01%	47.25%	46.62%	19.89%	14.86%	42.14%	39.94%	96.70%
140000-159999	78.19%	54.76%	61.83%	43.62%	49.19%	16.01%	14.39%	46.06%	43.62%	94.55%
>160000	67.89%	51.23%	56.37%	36.76%	43.63%	10.54%	12.99%	41.91%	35.29%	88.97%

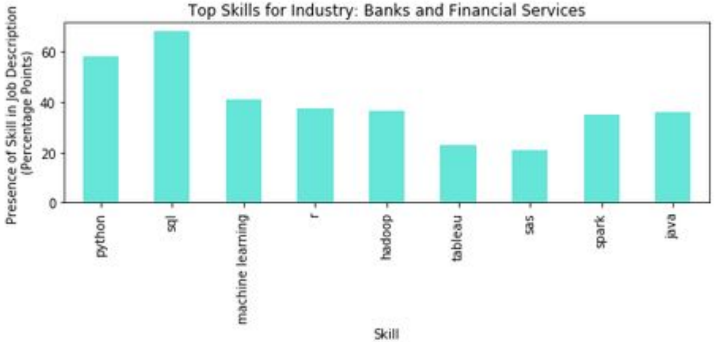
Skills for Top 3 Industries



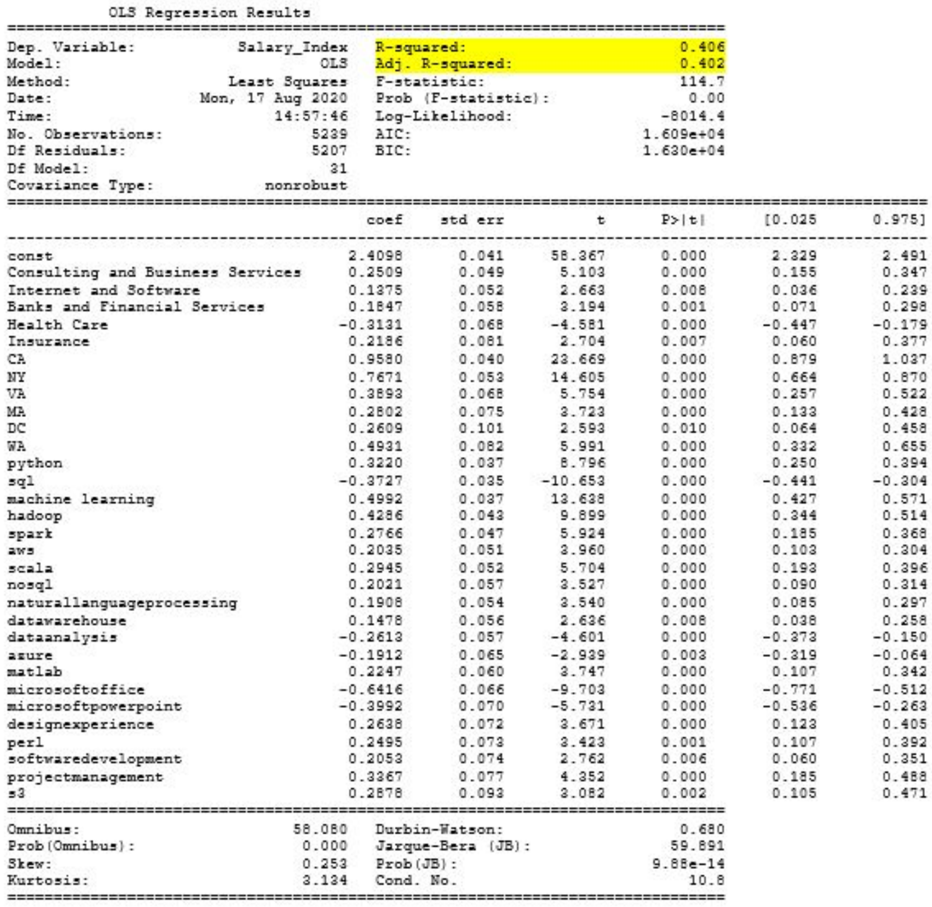
Skill	Consulting and Business Services	Internet and Software	Banks and Financial Services
python	59.83%	69.68%	58.02%
sql	52.95%	66.29%	68.13%
machine learning	47.75%	46.94%	41.10%
r	38.62%	42.90%	37.58%
hadoop	37.92%	39.03%	36.26%
tableau	28.09%	16.94%	23.08%
sas	16.99%	13.23%	21.10%
spark	33.01%	34.19%	35.16%
java	25.56%	34.68%	35.82%
Others	92.42%	94.03%	95.82%
Total	712	620	455







Preliminary Multivariate Regression Results Support Key Takeaways



```
                OLS Regression Results
===========================================================================
Dep. Variable:            Salary_Index   R-squared:                   0.406
Model:                             OLS   Adj. R-squared:              0.402
Method:                  Least Squares   F-statistic:                 114.7
Date:                 Mon, 17 Aug 2020   Prob (F-statistic):           0.00
Time:                         14:57:46   Log-Likelihood:            -8014.4
No. Observations:                 5239   AIC:                     1.609e+04
Df Residuals:                     5207   BIC:                     1.630e+04
Df Model:                           31
Covariance Type:             nonrobust
=============================================================================================
                                        coef    std err          t      P>|t|     [0.025     0.975]
---------------------------------------------------------------------------------------------
const                                 2.4098      0.041     58.367      0.000      2.329      2.491
Consulting and Business Services      0.2509      0.049      5.103      0.000      0.155      0.347
Internet and Software                 0.1375      0.052      2.663      0.008      0.036      0.239
Banks and Financial Services          0.1847      0.058      3.194      0.001      0.071      0.298
Health Care                          -0.3131      0.068     -4.581      0.000     -0.447     -0.179
Insurance                             0.2186      0.081      2.704      0.007      0.060      0.377
CA                                    0.9580      0.040     23.669      0.000      0.879      1.037
NY                                    0.7671      0.053     14.605      0.000      0.664      0.870
VA                                    0.3893      0.068      5.754      0.000      0.257      0.522
MA                                    0.2802      0.075      3.723      0.000      0.133      0.428
DC                                    0.2609      0.101      2.593      0.010      0.064      0.458
WA                                    0.4931      0.082      5.991      0.000      0.332      0.655
python                                0.3220      0.037      8.796      0.000      0.250      0.394
sql                                  -0.3727      0.035    -10.653      0.000     -0.441     -0.304
machine learning                      0.4992      0.037     13.638      0.000      0.427      0.571
hadoop                                0.4286      0.043      9.899      0.000      0.344      0.514
spark                                 0.2766      0.047      5.924      0.000      0.185      0.368
aws                                   0.2035      0.051      3.960      0.000      0.103      0.304
scala                                 0.2945      0.052      5.704      0.000      0.193      0.396
nosql                                 0.2021      0.057      3.527      0.000      0.090      0.314
naturallanguageprocessing             0.1908      0.054      3.540      0.000      0.085      0.297
datawarehouse                         0.1478      0.056      2.636      0.008      0.038      0.258
dataanalysis                         -0.2613      0.057     -4.601      0.000     -0.373     -0.150
azure                                -0.1912      0.065     -2.939      0.003     -0.319     -0.064
matlab                                0.2247      0.060      3.747      0.000      0.107      0.342
microsoftoffice                      -0.6416      0.066     -9.703      0.000     -0.771     -0.512
microsoftpowerpoint                  -0.3992      0.070     -5.731      0.000     -0.536     -0.263
designexperience                      0.2638      0.072      3.671      0.000      0.123      0.405
perl                                  0.2495      0.073      3.423      0.001      0.107      0.392
softwaredevelopment                   0.2053      0.074      2.762      0.006      0.060      0.351
projectmanagement                     0.3367      0.077      4.352      0.000      0.185      0.488
s3                                    0.2878      0.093      3.082      0.002      0.105      0.471
===========================================================================
Omnibus:                        58.080   Durbin-Watson:               0.680
Prob(Omnibus):                   0.000   Jarque-Bera (JB):           59.891
Skew:                            0.253   Prob(JB):                 9.88e-14
Kurtosis:                        3.134   Cond. No.                     10.8
===========================================================================
```

Findings and Conclusions

Post Mortem:

- Limitations
- How we deal with that
- Additional research questions (if we had two more weeks)

Did we find what we expected to find?

If not, why not?

What inferences or general conclusions can we draw from our analysis?

General conclusion:

- The analytics and technology skills vary widely (ML, Python, R, SQL as the most valuable)
- Problem-solving in the workplace, including soft skills such as communication, creativity and teamwork are also important skills
- Consulting, Internet & Software, Financial Services as the top industries

Q&A