Data Analyst Jobs Market

Group 8:

Mengqi(Maggie) Weng , Jinyu (Janet) Chen, Bingxin Li , Ziyi Gao, David Shin, Yamato Tadokoro

Bureau of Labor and Statistics predicts **18%** growth of job outlook for **Data Analyst** through **2029** in the US.

* More than 4 times faster than all other industries

Who Are We?

We are a consulting agency that focus on helping the best talent in the Digital Analytics market to find rewarding careers.

With over 10 years experience working solely in the Data & Analytics sector our consultants are able to offer detailed insights into the industry.

Problem Statement

Jason is trying to work in either the Accounting & Legal or the Business Services industry. He is favoring to work in either New York, Texas or California.

- Do jobs in New York, Texas and California have higher salary estimates than other states?
- Are the average salary estimates for Accounting & Legal or Business Services higher than other sectors

Ideal Experiment

Access to all data analyst/scientist jobs, and all variable information

Completely randomized sampling of jobs

Possible Outcomes

State

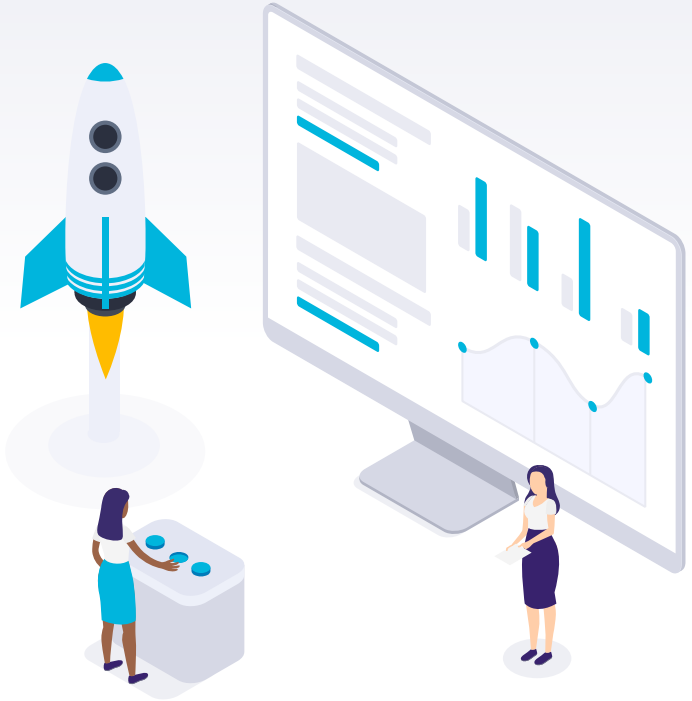
1) California, New York and Texas could each have the highest salary estimate.
2) The highest salary estimate could also not be in one of those states

Sector

1) Accounting & Legal or Business Services could have the highest salary estimate.
2) There are other sectors that have the highest salary estimate in those states

1

Data Set Description & EDA



Data Set

Source

- Kaggle

Variables of Interest

- Industry
- Sector
- Location
- Company Size
- Company Revenue
- Senior Role
- # of Hard skills

Job_Title	Salary_Estimate	Job_Description	Rating	Company_Name	Location	Headquarters	Size	Founded	Type_of_ownership
Data Analyst, Center on Immigration and Justic...	37K−66K (Glassdoor est.)	Are you eager to roll up your sleeves and harn...	3.2	Vera Institute of Justice\n3.2	New York, NY	New York, NY	201 to 500 employees	1961	Nonprofit Organization
Quality Data Analyst	37K−66K (Glassdoor est.)	Overview\n\nProvides analytical and technical ...	3.8	Visiting Nurse Service of New York\n3.8	New York, NY	New York, NY	10000+ employees	1893	Nonprofit Organization
Senior Data Analyst, Insights & Analytics Team...	37K−66K (Glassdoor est.)	We're looking for a Senior Data Analyst who ha...	3.4	Squarespace\n3.4	New York, NY	New York, NY	1001 to 5000 employees	2003	Company - Private
Data Analyst	37K−66K (Glassdoor est.)	Requisition NumberRR-0001939\nRemote:Yes\nWe	4.1	Celerity\n4.1	New York, NY	McLean, VA	201 to 500 employees	2002	Subsidiary or Business Segment

1. We took mean of "Salary_Estimate"
2. We got the location code
3. We took mean of the company "Size"
4. We took mean of company "Revenue"

	Industry	Sector	Revenue	Competitors	Easy_Apply
on	Social Assistance	Non-Profit	100 to 500 million (USD)	-1	True
on	Health Care Services & Hospitals	Health Care	2 to 5 billion (USD)	-1	-1
ate	Internet	Information Technology	Unknown / Non-Applicable	GoDaddy	-1

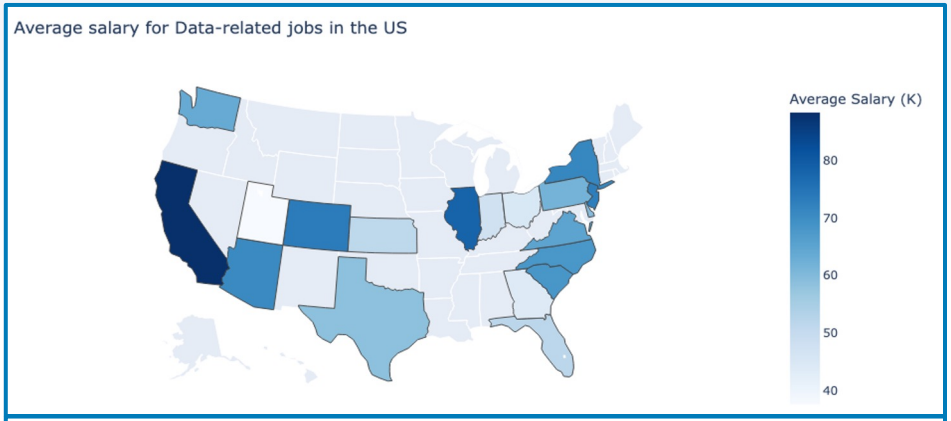
EDA(Data Cleaning)

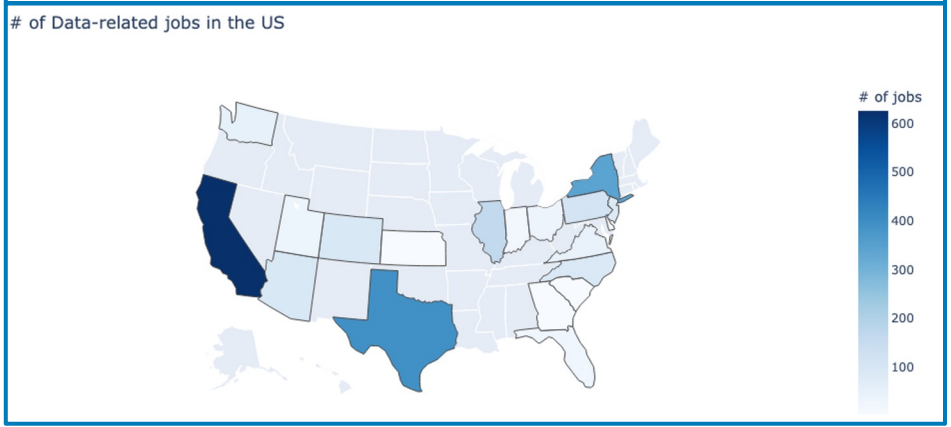
Job_Title	Salary_Estimate	Job_Description	Location	Size	Industry	Sector	Revenue
Senior Analyst, Data Science	66.5	The Role:\n\nRoku is looking for a Senior Anal...	New York, NY	1001 to 5000 employees	IT Services	Information Technology	$500 million to $1 billion (USD)
Data Science Analyst	56.0	NewGen is seeking a Data Science Analyst with ...	Norfolk, VA	1 to 50 employees	-1	-1	Less than $1 million (USD)
Data Analyst	138.5	Job Summary:\nCompany: Artech Information Syst...	San Francisco, CA	5001 to 10000 employees	Staffing & Outsourcing	Business Services	$500 million to $1 billion (USD)

Job_Title	Salary_Estimate	Industry	Sector	Senior	Location_code	Size_new	Revenue_new	SQL	Python	Machine Learning
Data Analyst	51.5	IT Services	Information Technology	0	NY	350	75.0	1	0	0
Senior SAS Data Analyst	60.5	Staffing & Outsourcing	Business Services	1	CA	3000	300.0	1	0	0
Data Analyst BHJOB11946_	36.0	Staffing & Outsourcing	Business Services	0	TX	750	-1.0	1	0	0

	Mean	Median	Min	Max
Salary_Estimate	72.09	69.0	1.0	150.0
Revenue_new	1348.51	17.5	-1.0	10000.0
Size_new	2586.77	350.0	-1.0	10000.0
Senior	12.65%	0.0	0.0	1.0
SQL	61.65%	1.0	0.0	1.0
Python	28.27%	0.0	0.0	1.0
Machine Learning	7.99%	0.0	0.0	1.0

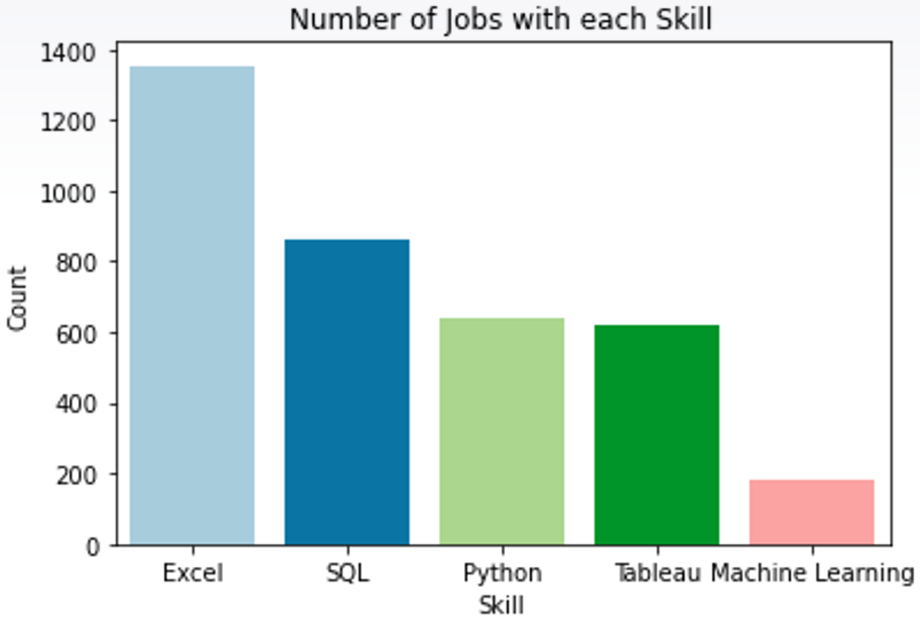
EDA(States Vs Jobs and Salary)





Top 5	Bottom 5
CA: $88.43K	UT: $37.53K
IL: $78.31K	GA: $44.00K
CO: $73.51K	OH: $45.20K
NJ: $73.00K	IN: $47.50K
NY: $71.41K	KS: $51.50K
CA: 626	SC: 3
TX: 394	KS: 3
NY: 345	GA: 4
IL: 164	DE: 11
PA: 114	IN: 23

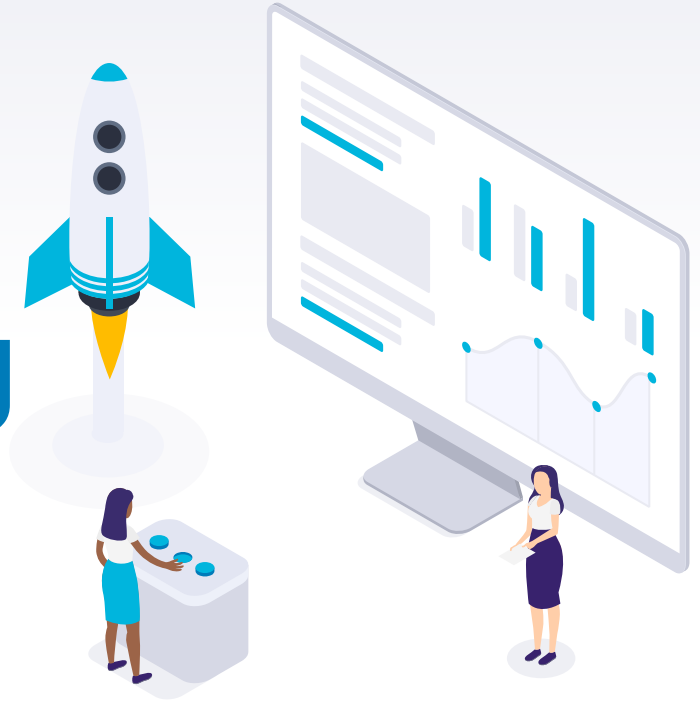
EDA(Job Vs Skills)



Skills

Hadoop
Machine Learning
MATLAB
MongoDB
Python
SAS
Spark
SPSS
SQL
Tableau

2 Model/Testing



Linear Regression Model

Dependent Variable

Salary Estimate

Independent Variables

Company Size

Company Revenue

Years of Experience

Skills

X_State

X_Sector

Variable Dictionary

Binary, 1: above 500 employees

Binary, 1: above $500M

Binary, 1: year>=0

Integer, 0-10

Binary,1: in [CA, NY, TX]

Binary,1: in [BS, Accounting]

Variables Analysis

```
result = sm.ols(formula="log_sal ~ com_size_u500+com_rev_u500m+ye0+skills+states+sec",data=df).fit()
```

OLS Regression Results

Dep. Variable:	log_sal	R-squared:	0.023
Model:	OLS	Adj. R-squared:	0.020
Method:	Least Squares	F-statistic:	8.854
Date:	Sat, 11 Dec 2021	Prob (F-statistic):	1.45e-09
Time:	12:25:19	Log-Likelihood:	-694.03
No. Observations:	2253	AIC:	1402.
Df Residuals:	2246	BIC:	1442.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.1572	0.020	203.608	0.000	4.117	4.197
com_size_u500	-0.0041	0.015	-0.266	0.790	-0.034	0.026
com_rev_u500m	0.0195	0.016	1.247	0.213	-0.011	0.050
ye0	0.0006	0.015	0.042	0.967	-0.029	0.030
skills	-0.0035	0.005	-0.691	0.489	-0.013	0.006
states	0.1014	0.014	7.093	0.000	0.073	0.129
sec	0.0158	0.017	0.954	0.340	-0.017	0.048

Company Size, State, and Sector have a positive relationship with salary estimate

Excluding Company size (x>500 employee) and skills variables

Comparing CA and TX

```
result = sm.ols(formula="log_sal ~ ny*bs+ca*bs+com_rev_u500m*ye0", data=data).fit()
```

OLS Regression Results

Dep. Variable:	log_sal	R-squared:	0.288
Model:	OLS	Adj. R-squared:	0.264
Method:	Least Squares	F-statistic:	11.76
Date:	Sat, 04 Dec 2021	Prob (F-statistic):	5.02e-14
Time:	15:11:40	Log-Likelihood:	-46.576
No. Observations:	241	AIC:	111.2
Df Residuals:	232	BIC:	142.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.9022	0.104	37.577	0.000	3.698	4.107
ny	0.4012	0.168	2.395	0.017	0.071	0.731
bs	0.1367	0.106	1.289	0.199	-0.072	0.346
ny:bs	-0.1585	0.177	-0.897	0.371	-0.506	0.190
ca	0.6582	0.135	4.883	0.000	0.393	0.924
ca:bs	-0.2728	0.143	-1.910	0.057	-0.554	0.009
com_rev_u500m	0.0043	0.069	0.062	0.950	-0.132	0.141
ye0	0.0115	0.045	0.255	0.799	-0.077	0.101
com_rev_u500m:ye0	0.0791	0.095	0.836	0.404	-0.107	0.265

Comparing Salary Estimates CA (1) vs. TX(CA=0,NY=0) for Business Services

- BS=1, Other variables controlled
- D = exp(0.658 - 0.273) = 1.4696
- 46.96% higher average salary in CA than in TX

Comparing Salary Estimates CA (1) vs. TX(CA=0,NY=0) for Accounting and Legal

- BS=0, Other variables controlled
- D = exp(0.658) = 1.9309
- 93.09% higher average salary in CA than in TX

Comparing NY and TX

```
result = sm.ols(formula="log_sal ~ ny*bs+ca*bs+com_rev_u500m*ye0", data=data).fit()
```

OLS Regression Results

Dep. Variable:	log_sal	R-squared:	0.288
Model:	OLS	Adj. R-squared:	0.264
Method:	Least Squares	F-statistic:	11.76
Date:	Sat, 04 Dec 2021	Prob (F-statistic):	5.02e-14
Time:	15:11:40	Log-Likelihood:	-46.576
No. Observations:	241	AIC:	111.2
Df Residuals:	232	BIC:	142.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.9022	0.104	37.577	0.000	3.698	4.107
ny	0.4012	0.168	2.395	0.017	0.071	0.731
bs	0.1367	0.106	1.289	0.199	-0.072	0.346
ny:bs	-0.1585	0.177	-0.897	0.371	-0.506	0.190
ca	0.6582	0.135	4.883	0.000	0.393	0.924
ca:bs	-0.2728	0.143	-1.910	0.057	-0.554	0.009
com_rev_u500m	0.0043	0.069	0.062	0.950	-0.132	0.141
ye0	0.0115	0.045	0.255	0.799	-0.077	0.101
com_rev_u500m:ye0	0.0791	0.095	0.836	0.404	-0.107	0.265

Comparing Salary Estimates NY(CA=0) vs. TX (CA=0,NY=0) for Business Services

- BS=1, Other variables controlled
- D = exp(.401 - 0.159) = 1.2738
- 27.38% higher average salary in NY than in TX

Comparing Salary Estimates NY(CA=0) vs. TX (CA=0,NY=0) for Accounting and Legal

- BS=0, Other variables controlled
- D = exp(.401) = 1.4933
- 49.33% higher average salary in NY than in TX

Comparing CA and NY

```
result = sm.ols(formula="log_sal ~ tx*bs+ca*bs+com_rev_u500m*ye0", data=data).fit()
```

OLS Regression Results

Dep. Variable:	log_sal	R-squared:	0.288
Model:	OLS	Adj. R-squared:	0.264
Method:	Least Squares	F-statistic:	11.76
Date:	Sat, 04 Dec 2021	Prob (F-statistic):	5.02e-14
Time:	15:22:04	Log-Likelihood:	-46.576
No. Observations:	241	AIC:	111.2
Df Residuals:	232	BIC:	142.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.3035	0.135	31.842	0.000	4.037	4.570
tx	-0.4012	0.168	-2.395	0.017	-0.731	-0.071
bs	-0.0218	0.141	-0.154	0.878	-0.300	0.257
tx:bs	0.1585	0.177	0.897	0.371	-0.190	0.506
ca	0.2570	0.162	1.591	0.113	-0.061	0.575
ca:bs	-0.1144	0.170	-0.673	0.502	-0.449	0.221
com_rev_u500m	0.0043	0.069	0.062	0.950	-0.132	0.141
ye0	0.0115	0.045	0.255	0.799	-0.077	0.101
com_rev_u500m:ye0	0.0791	0.095	0.836	0.404	-0.107	0.265

Comparing Salary Estimates CA (1) vs. NY(CA=0,TX=0) for Business Services

- BS=1, Other variables controlled
- D = exp(0.257 - 0.1144) =1.1533
- 15.33% higher in average salary in CA than in NY

Comparing Salary Estimates CA(1) vs. NY (CA=0,TX=0) for Accounting and Legal

- BS=0, Other variables controlled
- D = exp(0.257) =1.2930
- 29.3% higher in average salary in CA than in NY

Hypothesis Testing

Is the sector's average salary higher than the state average salary?
(Two sample, one-sided t-test | α =0.05)

California

Accounting & Legal
>
All Other Sectors in CA

New York

Business Services
>
All Other Sectors in NY

Texas

Business Services
>
All Other Sectors in TX

Model Analysis (California)

$H_o : \mu_{CA,AccountingLegal} = \mu_{CA,Average}$ $H_A : \mu_{CA,AccountingLegal} \neq \mu_{CA,Average}$

```
(                          Salary_Estimate        std  count
 Sector
 Accounting & Legal             101.708333  25.814424     12
 Information Technology          94.254054  28.139217    185
 Health Care                     93.756757  30.946105     37
 Biotech & Pharmaceuticals       89.718750  15.331850     16
 Manufacturing                   88.192308  31.665267     13
 Business Services               87.004167  27.625610    120
 Education                       83.725000  17.321629     20
 Insurance                       83.625000  31.796488     16
 Finance                         83.596774  32.219409     31
 Media                           72.475000  24.392339     20,
 Ttest_1sampResult(statistic=1.8163885099143071, pvalue=0.04831674784576253))
```

- **In CA, comparing Accounting & Legal average salary vs all the salary estimate excluding the Accounting & Legal sector**

Conclusion:

- **Since, the p-value is 0.048, CA's Accounting & Legal Sector is statistically significant**

※only including sectors that has more than 10 job descriptions

Model Analysis (New York)

$H_o : \mu_{NY,BusinessServices} = \mu_{NY,Average}$ $H_A : \mu_{NY,BusinessServices} \neq \mu_{NY,Average}$

```
(                        Salary_Estimate        std  count
 Sector
 Business Services             76.195946  18.766558     74
 Information Technology        70.653846  18.371051     78
 Health Care                   70.023256  19.892255     43
 Non-Profit                    67.125000  11.874581     12
 Finance                       67.064516  18.710844     31
 Media                         64.428571  13.660523     14,
 Ttest_1sampResult(statistic=2.1930793537369984, pvalue=0.0157439207978425))
```

- **In NY, comparing Business average salary vs all the salary estimate excluding the Business Services sector**

Conclusion:

- **Since, the p-value is 0.015, NY's Business Services Sector is statistically significant**

※only including sectors that has more than 10 job descriptions

Model Analysis (Texas)

$H_o : \mu_{TX,BusinessServices} = \mu_{TX,Average}$ $H_A : \mu_{TX,BusinessServices} \neq \mu_{TX,Average}$

Sector	Salary_Estimate	std	count
Business Services	60.516129	18.797469	93
Health Care	59.981481	22.568052	27
Information Technology	59.247475	16.395988	99
Government	57.333333	17.745657	21
Finance	56.562500	18.287908	32,

Ttest_1sampResult(statistic=0.9054247320243054, pvalue=0.1838027585280902))

Conclusion:

- **Texas' are not significant, we think it might the sample's salary estimates are not so different from each other.**
- **Though If Jason is really interested in Business Services, Texas probably won't be the best choice**

Conclusion

1) Accounting Legal sector 's estimate average salary in CA is significantly higher than the state average

2) The Business Service sector 's estimate average salary in NY is significantly higher than any other sectors.

Limitations and Next Step

1) Estimated Salaries, Revenue and Company Size were given in a range, creating a very narrow , non-robust distribution.

2) The dataset only provided a limited States selection.

THANK YOU!

Any questions?

References

- Dataset: https://www.kaggle.com/andrewmvd/data-analyst-jobs
- Problem Statistics: https://blog.galvanize.com/data-analyst-salary-map-2021-how-does-your-state-compare/