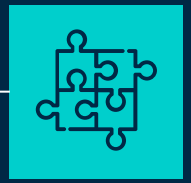
DATA SCIENCE JOB ANALYSIS

A presentation by Jessica Hoang,
Ariba Anees & Nagulan Nathan
OPMA419 W22
Group 3

Table of Contents



01

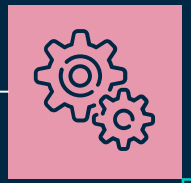
PROBLEM

What is our topic?

02

IMPORTANCE

Why did we choose this topic?



03

DATA

Data source and preparation

Table of Contents

04

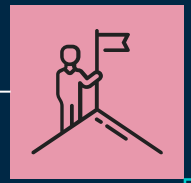
ANALYSIS

k-NN
Linear Regression
Regression Tree
Random Forest
Tableau

05

RECOMMENDATION AND VISUALS

What do we recommend to our fellow data analysts?



06

CHALLENGES

What were the challenges and limitations of the project?

Which algorithm most accurately predicts average salary for data science jobs in the US?

&

What variables are significant predictors?

Why Did We Choose This Topic?

Relevancy

The three of us are interested in Data Analytics as a career

We are in OPMA419

Related to the Audience

Usefulness

Predict salaries for future listings

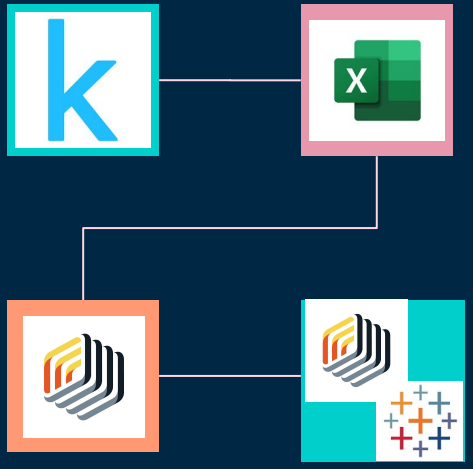
Knowing which skills are in demand

Knowing which areas have high-paying jobs

Data Workflow

Kaggle

Data downloaded as Excel csv



Data Cleaning I

Within Excel, unnecessary columns and rows with missing data removed

Data Cleaning II

In RapidMiner, we selected attributes, set role, and changed the data types as needed and made dummy variables

Analysis

Using RapidMiner to try different algorithms and Tableau for supplementary visualizations

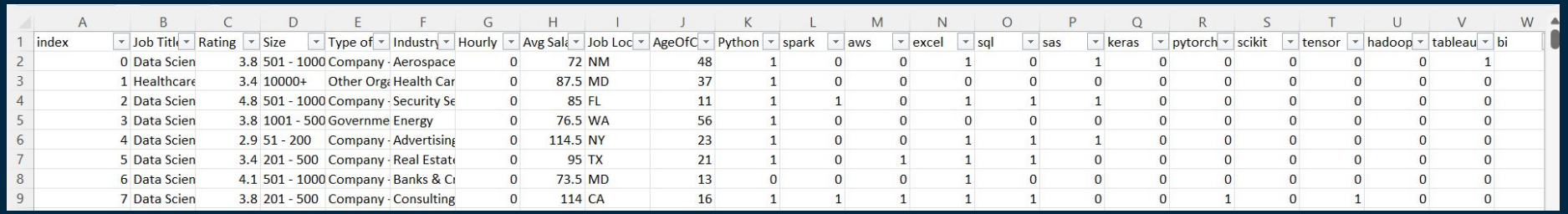
What Our Raw Data From Kaggle Looks Like

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	index	Job Title	Salary Esti	Job Descri	Rating	Company	Location	Headquart	Size	Founded	Type of ov	Industry	Sector	Revenue	Competito	Hourly	Employer	Lower Sal	Upper Sal	Avg Salary	company_	Job Locati	Age
2	0	Data Scien	$53K-$91K	Data	3.8	Tecolote	Albuquerq	Goleta, CA	501 - 1000	1973	Company -	Aerospace	Aerospace	$50 to $10	-1	0	0	53	91	72	Tecolote F	NM	48
3	1	Healthcare	$63K-$112	What You	3.4	University	Linthicum,	Baltimore,	10000+	1984	Other Orga	Health Car	Health Car	$2 to $5 bi	-1	0	0	63	112	87.5	University	MD	37
4	2	Data Scien	$80K-$90K	KnowBe4,	4.8	KnowBe4	Clearwate	Clearwate	501 - 1000	2010	Company -	Security Se	Business S	$100 to $5	-1	0	0	80	90	85	KnowBe4	FL	11
5	3	Data Scien	$56K-$97K	*Organiza	3.8	PNNL	Richland, V	Richland, V	1001 - 500	1965	Governme	Energy	Oil, Gas, Er	$500 millic	Oak Ridge	0	0	56	97	76.5	PNNL	WA	56
6	4	Data Scien	$86K-$143	Data	2.9	Affinity	New York,	New York,	51 - 200	1998	Company -	Advertising	Business S	Unknown /	Commerce	0	0	86	143	114.5	Affinity So	NY	23
7	5	Data Scien	$71K-$119	CyrusOne	3.4	CyrusOne	Dallas, TX	Dallas, TX	201 - 500	2000	Company -	Real Estate	Real Estate	$1 to $2 bi	Digital Rea	0	0	71	119	95	CyrusOne	TX	21
8	6	Data Scien	$54K-$93K	Job	4.1	ClearOne	Baltimore,	Baltimore,	501 - 1000	2008	Company -	Banks & Ci	Finance	Unknown /	-1	0	0	54	93	73.5	ClearOne /	MD	13

	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
	Python	spark	aws	excel	sql	sas	keras	pytorch	scikit	tensor	hadoop	tableau	bi	flink	mongo	google_an	job_title_s	seniority_l	Degree
48	1	0	0	1	0	1	0	0	0	0	0	1	1	0	0	0	data scien	na	M
37	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	data scien	na	M
11	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	data scien	na	M
56	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	data scien	na	na
23	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	data scien	na	na
21	1	0	1	1	1	0	0	0	0	0	0	0	1	0	1	0	data scien	na	na
13	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	data scien	na	na

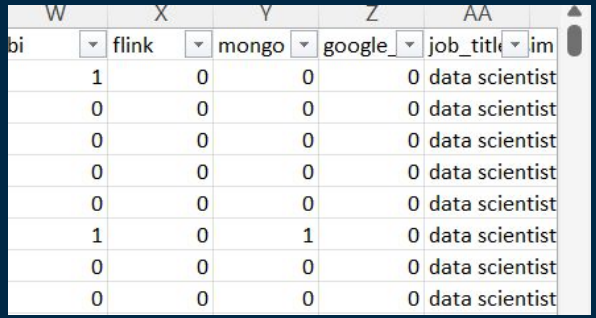
Lots of Columns!

What Our Cleaned Data Looks Like



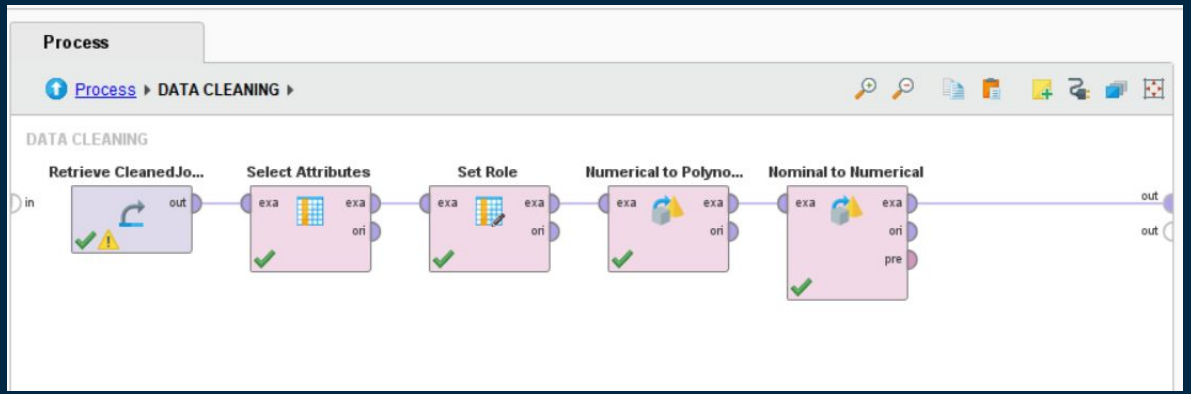
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	index	Job Titl	Rating	Size	Type of	Industr	Hourly	Avg Sal	Job Loc	AgeOfC	Python	spark	aws	excel	sql	sas	keras	pytorch	scikit	tensor	hadoop	tableau	bi
2	0	Data Scien	3.8	501 - 1000	Company -	Aerospace	0	72	NM	48	1	0	0	1	0	1	0	0	0	0	0	1	
3	1	Healthcare	3.4	10000+	Other Orga	Health Car	0	87.5	MD	37	1	0	0	0	0	0	0	0	0	0	0	0	
4	2	Data Scien	4.8	501 - 1000	Company -	Security Se	0	85	FL	11	1	1	0	1	1	1	0	0	0	0	0	0	
5	3	Data Scien	3.8	1001 - 500	Governme	Energy	0	76.5	WA	56	1	0	0	0	0	0	0	0	0	0	0	0	
6	4	Data Scien	2.9	51 - 200	Company -	Advertising	0	114.5	NY	23	1	0	0	1	1	1	0	0	0	0	0	0	
7	5	Data Scien	3.4	201 - 500	Company -	Real Estat	0	95	TX	21	1	0	1	1	1	0	0	0	0	0	0	0	
8	6	Data Scien	4.1	501 - 1000	Company -	Banks & Cr	0	73.5	MD	13	0	0	0	1	0	0	0	0	0	0	0	0	
9	7	Data Scien	3.8	201 - 500	Company -	Consulting	0	114	CA	16	1	1	1	1	1	0	0	1	0	1	0	0	

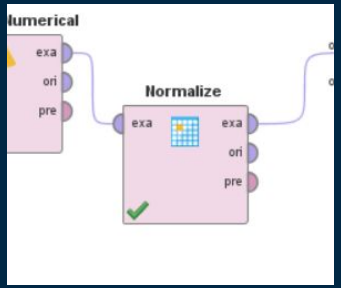
Slightly less Columns!



W	X	Y	Z	AA
bi	flink	mongo	google_	job_titl	im
1	0	0	0	data scientist	
0	0	0	0	data scientist	
0	0	0	0	data scientist	
0	0	0	0	data scientist	
0	0	0	0	data scientist	
1	0	1	0	data scientist	
0	0	0	0	data scientist	
0	0	0	0	data scientist	

The RapidMiner Data Preparation





For k-NN, to ensure scales do not skew euclidean distance

We omitted the index and Job_Title as well and made dummy variables as needed..... Now let's jump into the analysis!

Our Analysis: k-NN

	RMSE	R^2	Average Error
Training	28.707	0.454	-3.985
Validation	32.137	0.242	-4.773

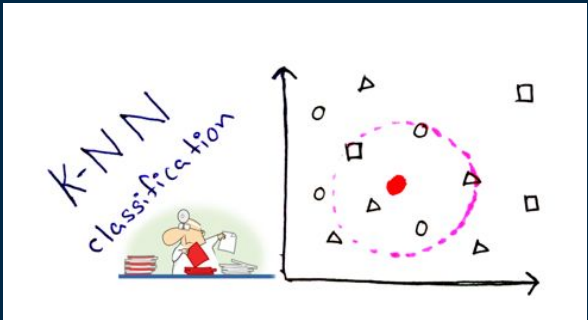


Image sourced from OPMA419 Lecture Notes on K-NN

We used a k of 10 as it yielded the lowest RMSE from k of 1-10 on the Validation set

Our Analysis: Linear Regression

MOST SIGNIFICANT PREDICTOR AT 95% (with all predictors)

0.000	0.000	0.001	0.003
Google Analytics	Python	SAS	SQL

MOST SIGNIFICANT PREDICTOR AT 95% (comparing skills to each other)

0.000	0.002	0.012	0.026
Python	SAS	SQL	Google Analytics

Our Analysis: Linear Regression

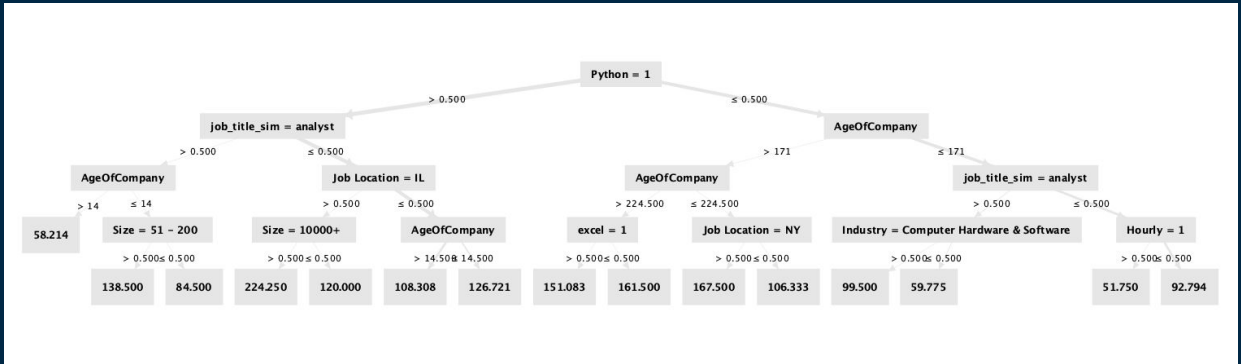
NEGATIVE COEFFICIENT ANALYSIS

SQL: -7.88
Google Analytics: -69.85

PERFORMANCE MEASURES

	RMSE	R^2	Average Error
Training	22.557	0.654	-0.000
Validation	34.801	0.278	3.032

Our Analysis: Regression Tree



	RMSE	R^2	Average Error
Training	28.795	0.437	0.000
Validation	31.435	0.267	0.213

Our Analysis: Random Forest

	RMSE	R^2	Average Error
Training	33.001	0.474	-0.071
Validation	32.757	0.289	-1.720

Based on our analysis, the Regression Tree algorithm performed better in RMSE and average error than all other algorithms.

Recommendations

ALGORITHM

Regression Tree had the lowest RMSE on the validation set

JOB ROLE

Work as a machine learning engineer

LOCATION

Work in Rhode Island

SKILLS

Be skillful in Flink, Python, SAS, and MongoDB

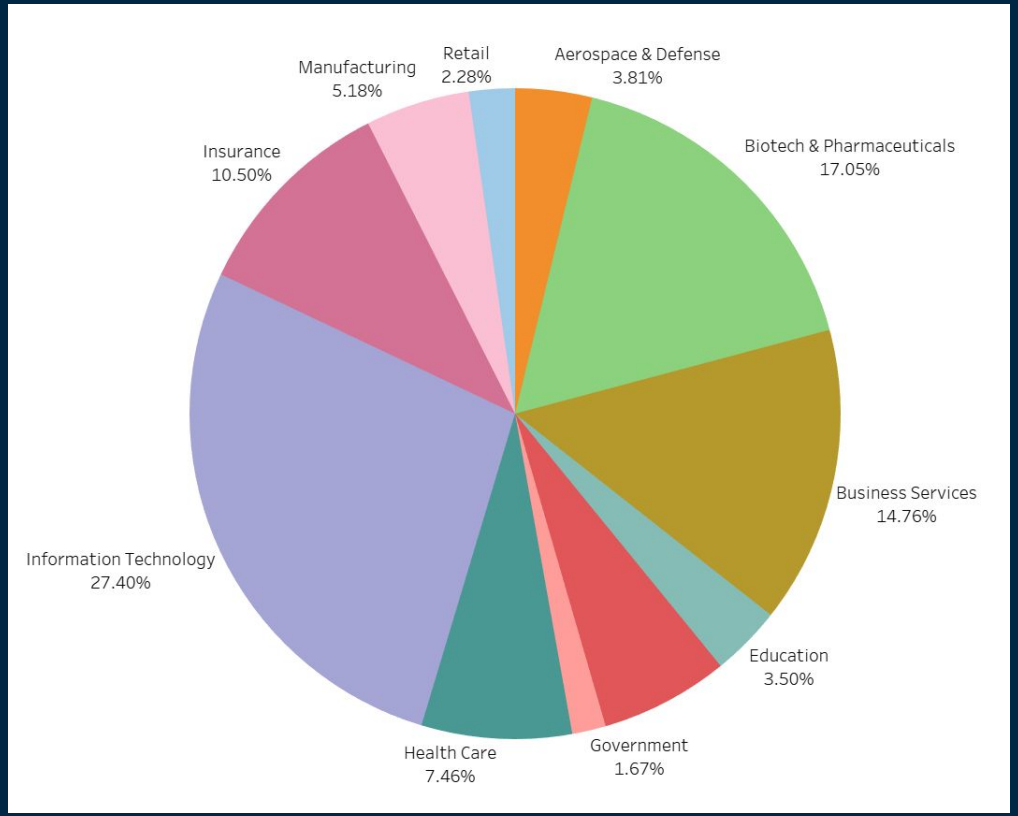
INDUSTRY

Work in the trucking industry

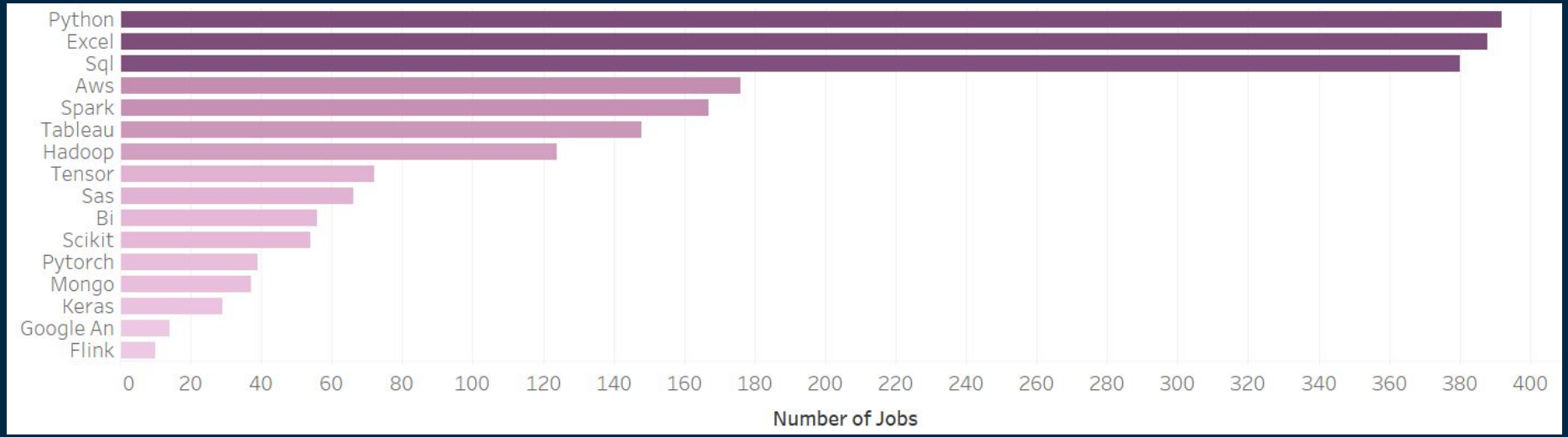
SIZE

Work for a company with 51-200 employees

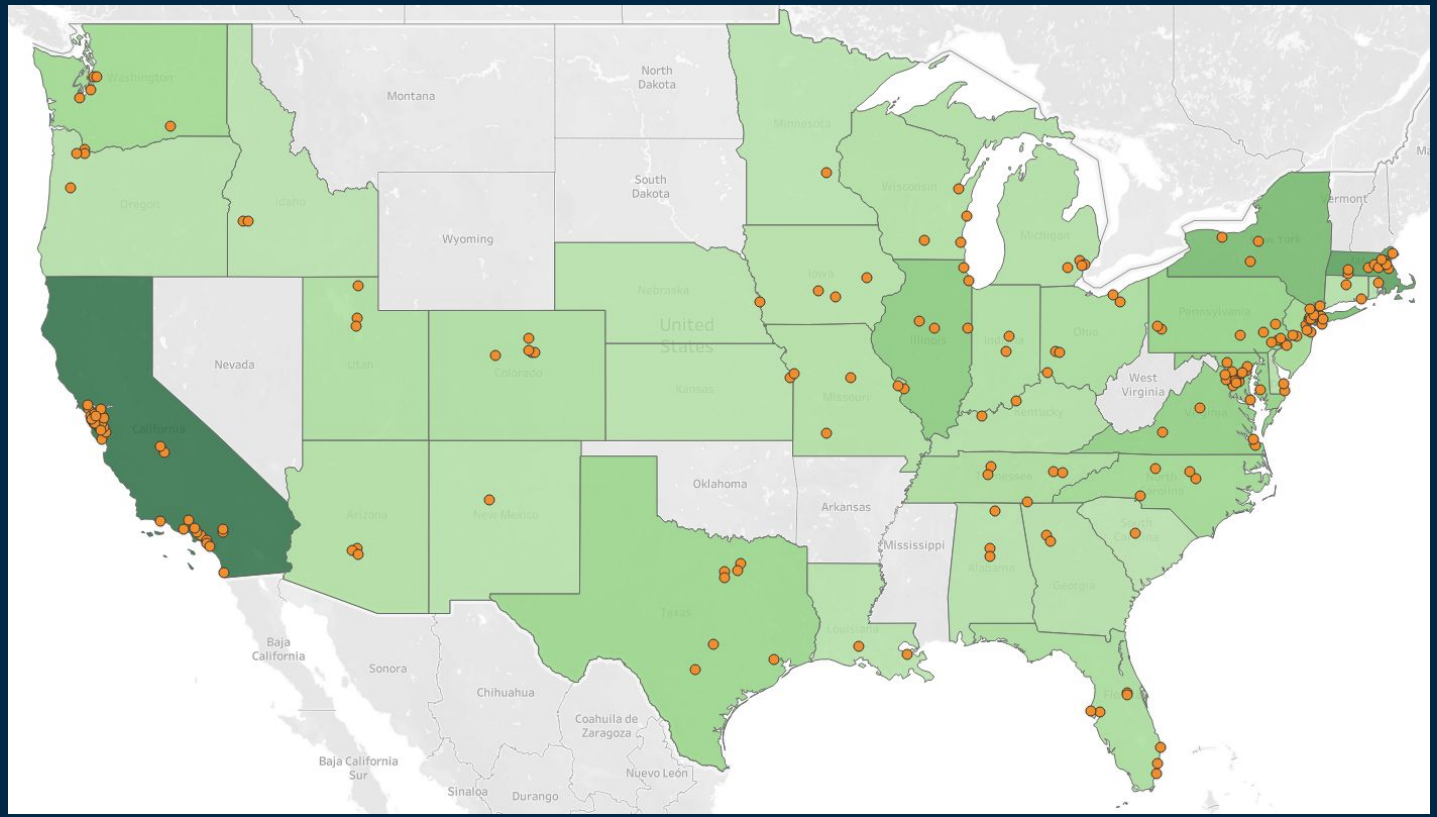
Amount of Jobs by Sector



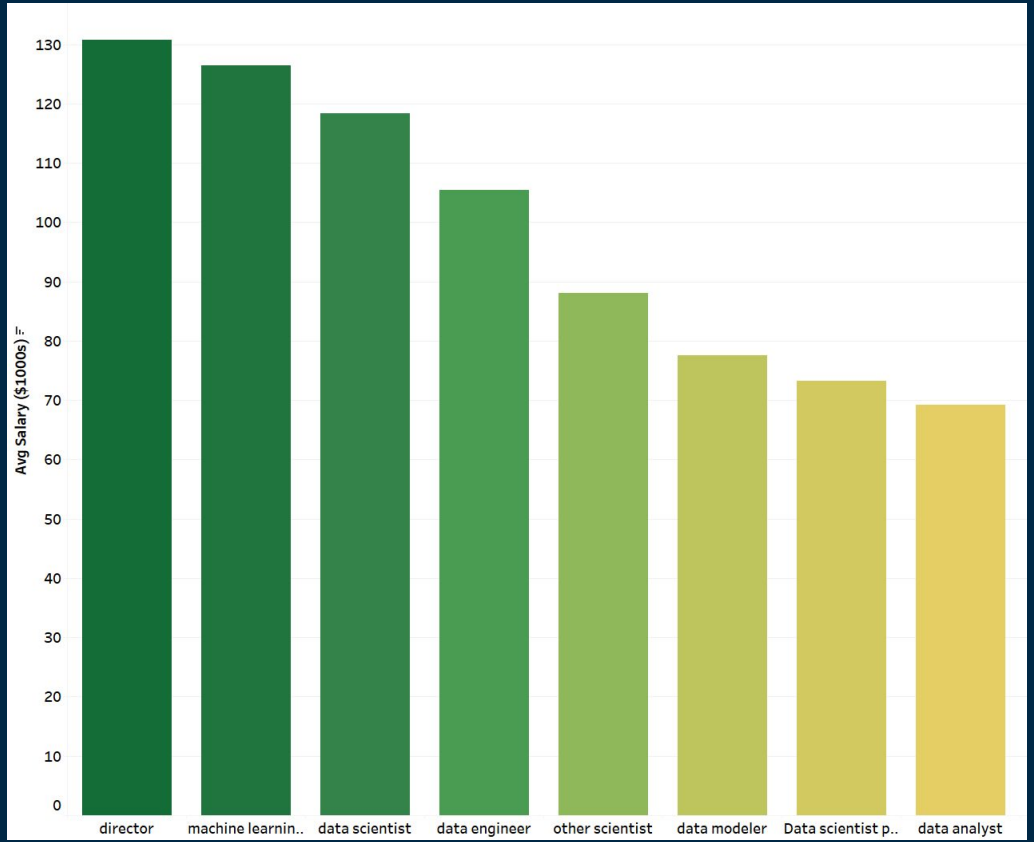
Skills in Demand



Distribution of Data Science Jobs and Average Salary by State



Average Salary by Job Title



Challenges & Limitations Faced

DIRTY DATA

There were missing values, and redundant columns that were time intensive for cleaning, or data that needed restructuring for Tableau

HIGH # OF CATEGORICAL PREDICTORS

Most data was polynominal, computationally intensive for BE when there are too many dummy variables

ROWS OF DATA

Just over 700 rows of data was not ideal especially when partitioned for algorithms like k-NN

THANKS

QUESTIONS?