



┌ Epsilon

Data Science Project

Glass Door Salaries Prediction

Project Overview

- Making a model that estimates data science salaries to help data scientists negotiate their income when they get a job.
- Do some features engineering from the text of each job description to quantify the value companies put on python, SQL, R, and Power BI.
- Perform ML Models like SVR , Decision trees , and Random Forest Regressors to reach the best model performance .
- Deployment the model with flask , HTML , CSS with using Heroku .



Features of the data

Job title
Salary Estimate
Rating
Location
Company Size
Type of Ownership
Sector

Competitors
Job Description
Company
Company Headquarters
Company Founded Date
Industry
Revenue

Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector
Data Scientist	53K–91K (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 to 1000 employees	1973	Company - Private	Aerospace & Defense	Aerospace & Defense m
Healthcare Data Scientist	63K–112K (Glassdoor est.)	What You Will Do:\n\nI. General Summary\n\nThe...	3.4	University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+ employees	1984	Other Organization	Health Care Services & Hospitals	Health Care
Data Scientist	80K–90K (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 to 1000 employees	2010	Company - Private	Security Services	Business Services m
Data Scientist	56K–97K (Glassdoor est.)	*Organization and Job ID**\n\nJob ID: 310709\n\n...	3.8	PNNL\n3.8	Richland, WA	Richland, WA	1001 to 5000 employees	1965	Government	Energy	Oil, Gas, Energy & Utilities 50
Data Scientist	86K–143K (Glassdoor est.)	Data Scientist\nAffinity Solutions / Marketing...	2.9	Affinity Solutions\n2.9	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Advertising & Marketing	Business Services

Package I used :

Pandas, Numpy, Sklearn, Matplotlib, Seaborn, ggplot , Datasist , joblib ,Flask, , pickle



Data Frame



pandas



Machine Learning



Vectors



NumPy



Visualization

Data preprocessing

1- Doing some feature engineering & extraction

- Extract jobs skills like :

python

SQL

R

Power_BI

Excel

Computer Vision



Data preprocessing

- Simplify job title to :

Data scientist
Data engineer
Analyst
Manager
ML
Director



Data preprocessing

- Simplify Seniority to :

Senior
Junior

- Extract numeric data out of salary :

Minimum ,Maximum & Average
Between.



Data preprocessing

- Create a column Calculate jobs that offered in per hour and convert it to annually paid.
- Removed rows without Salary Estimate, Size , Sector & Location
- Outliers Handling in Salary



Visualization

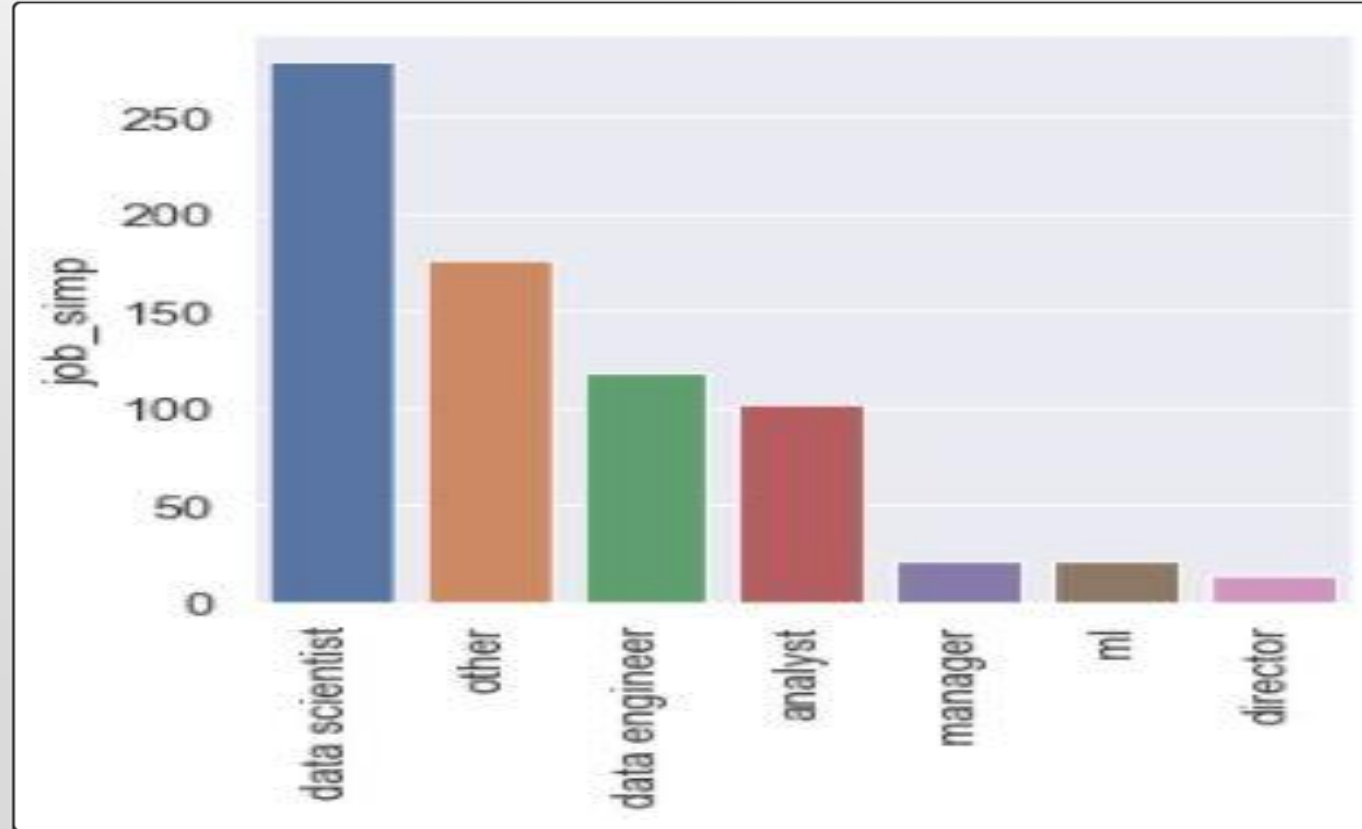
A presentation of data with insight information in a graph, chart, or other visual format. It communicates relationships of the data with images. And find patterns to be more easily seen.



Visualization

It is noted that

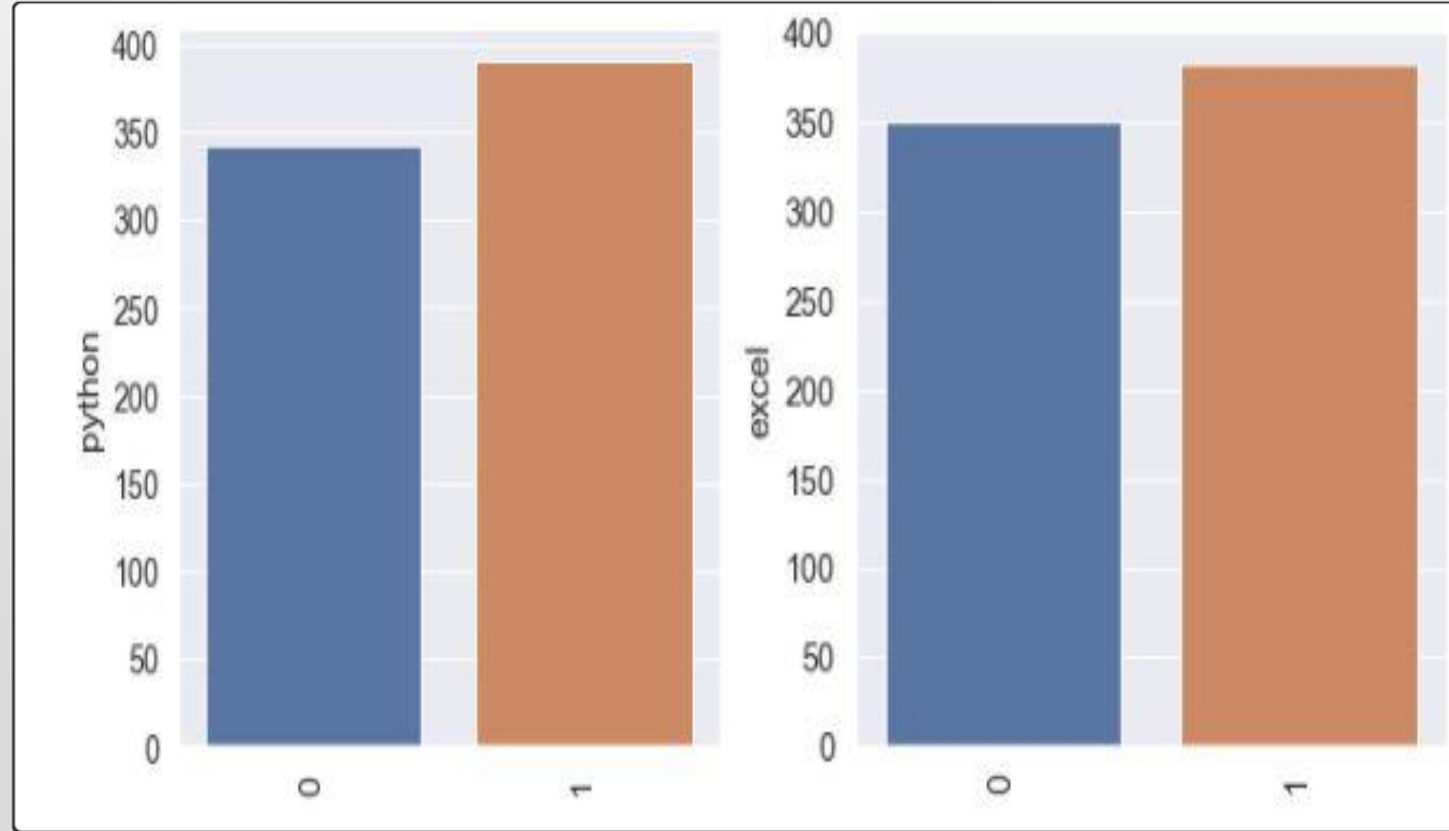
Data scientist have the highest job applications with over 250+ job vacancy
ml & analyst have the lowest in jobs



Visualization

It is noted that

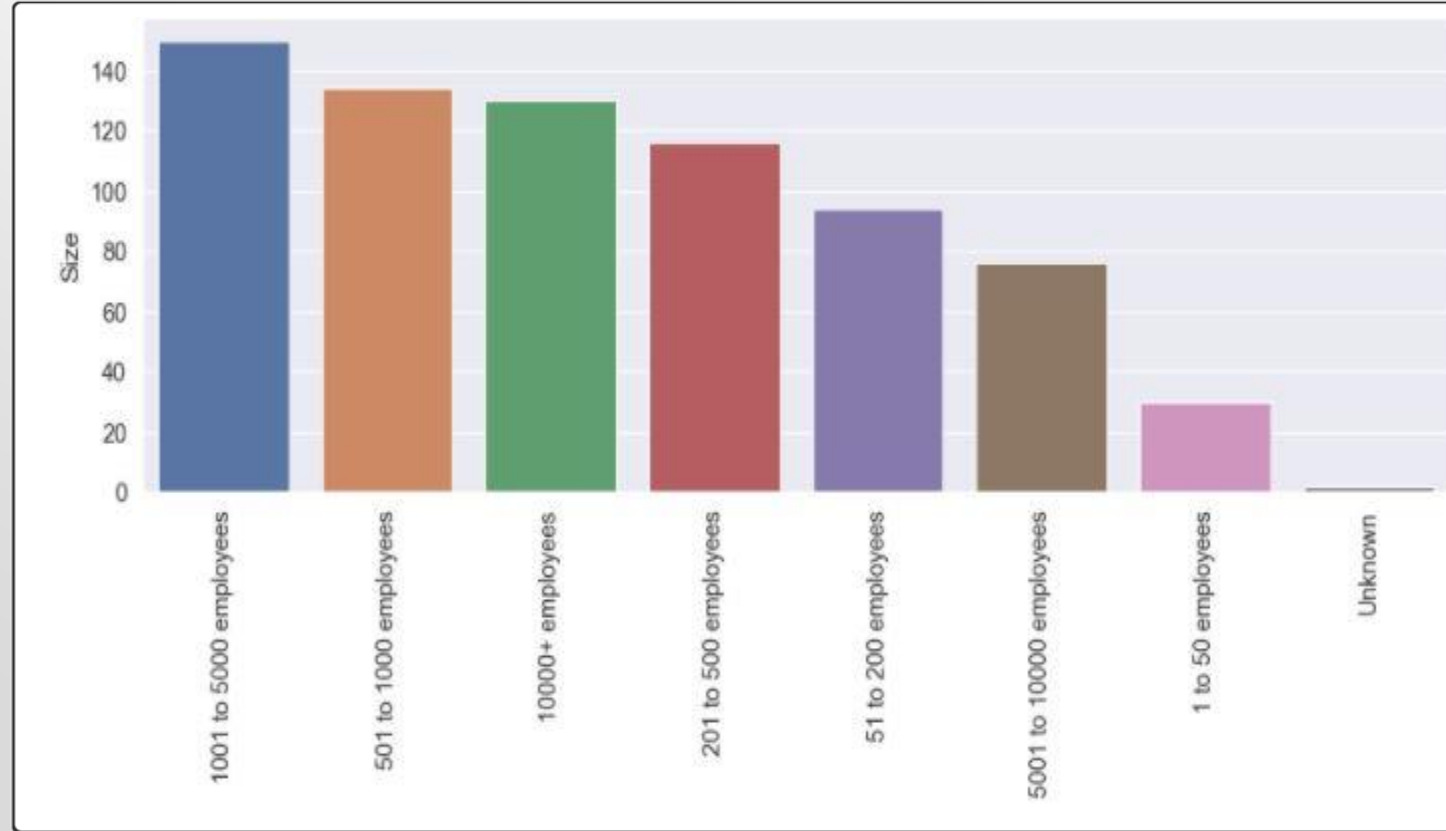
Python is more required in
50% of job opportunity &
Office suite Excel is the
same



Visualization

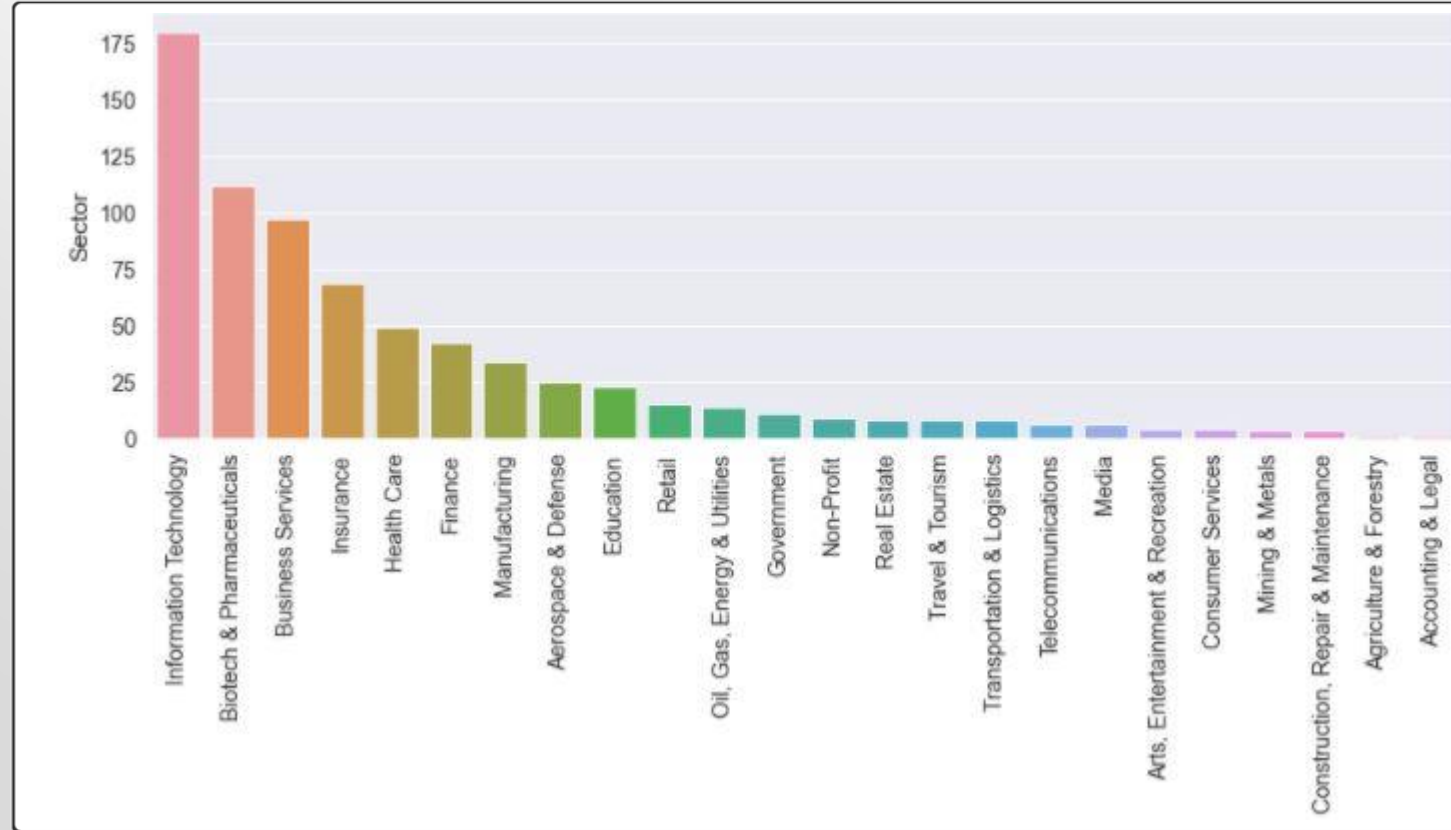
It is noted that

Average size companies have the highest demand for jobs



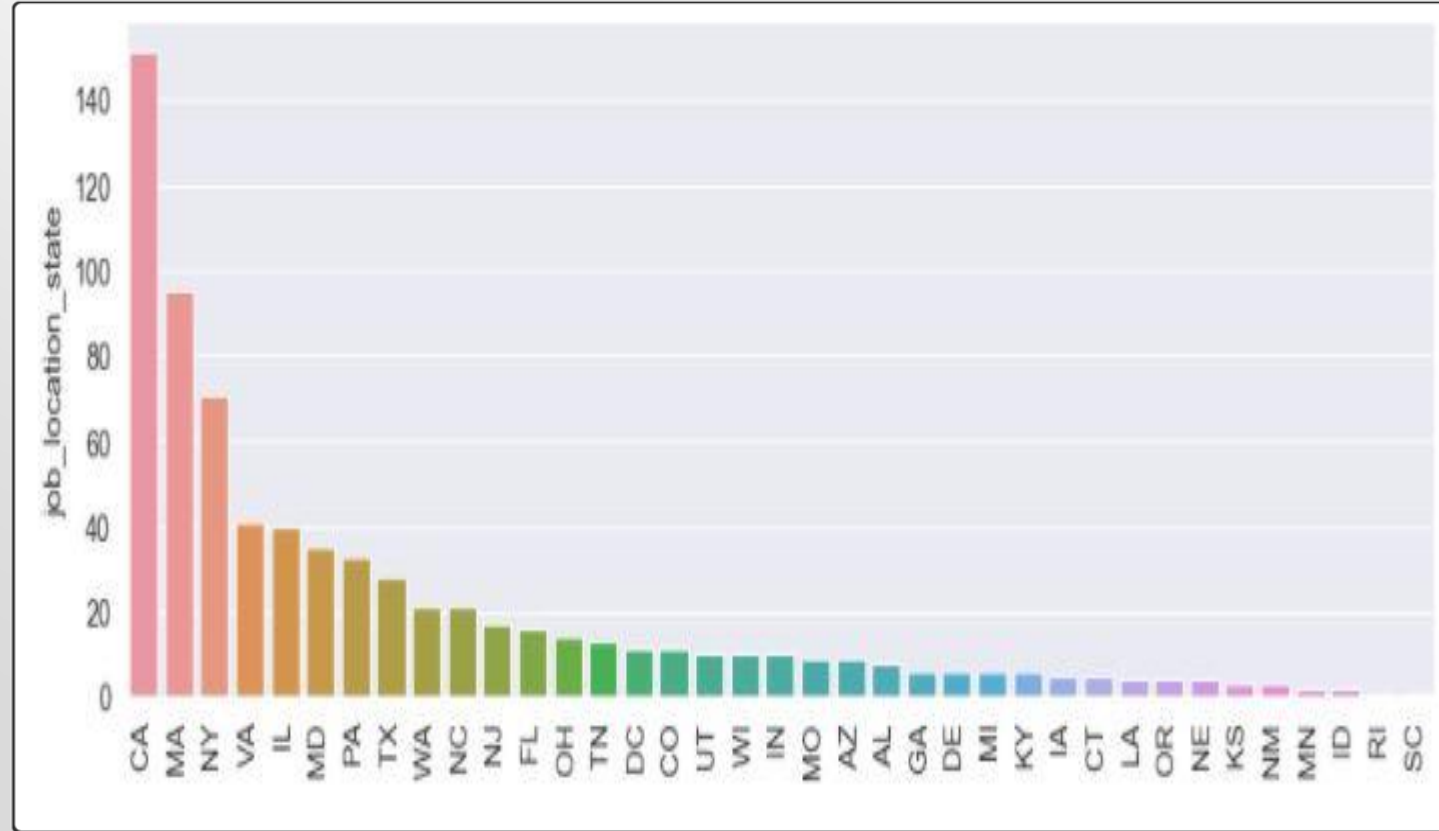
Visualization

It is noted that
IT companies have The
highest demand for jobs
with 150+ job
applications



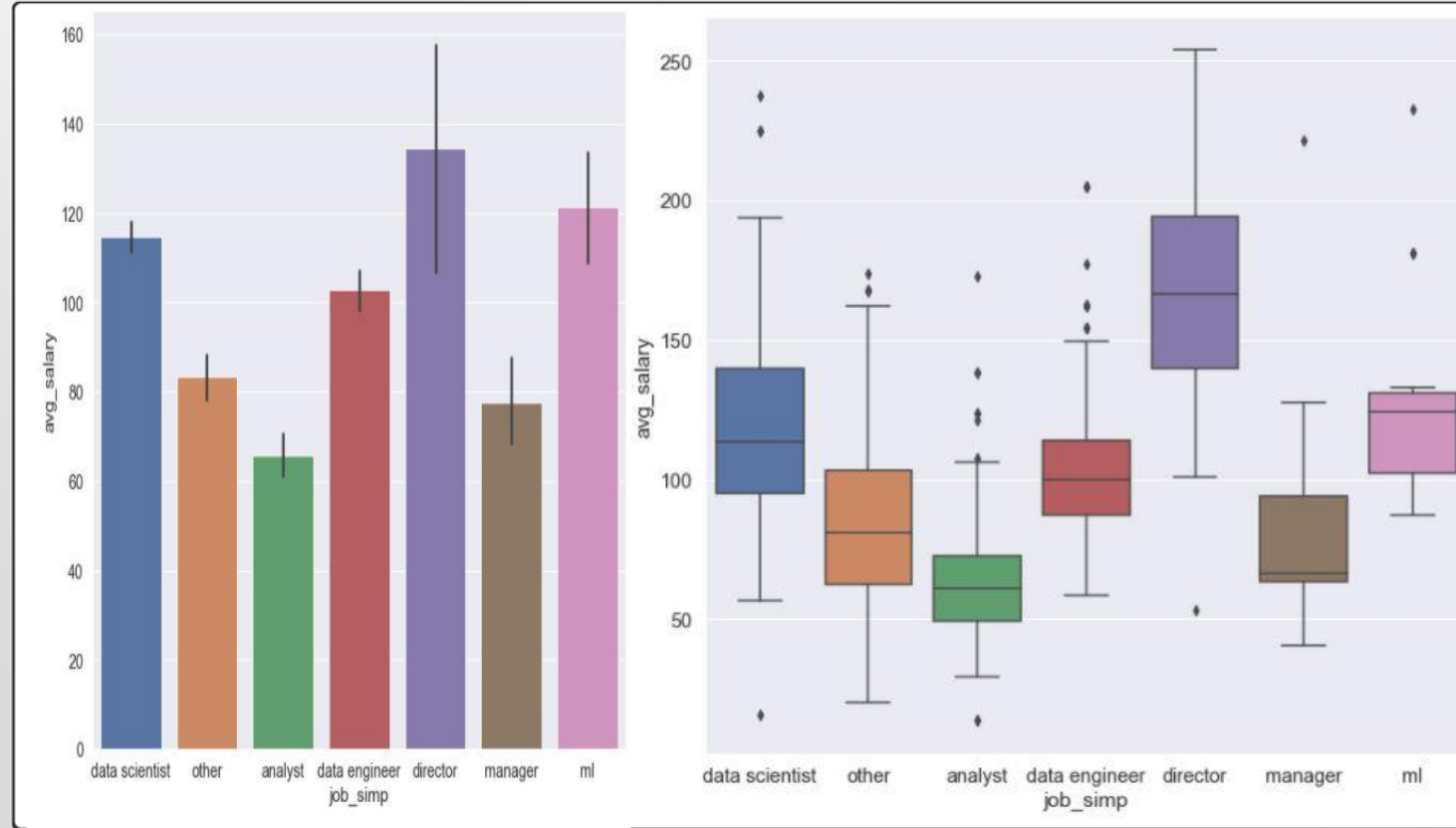
Visualization

It is noted that
California have The
highest demand for jobs
with 140+ job
applications &
Minnesota has the lowest



Visualization

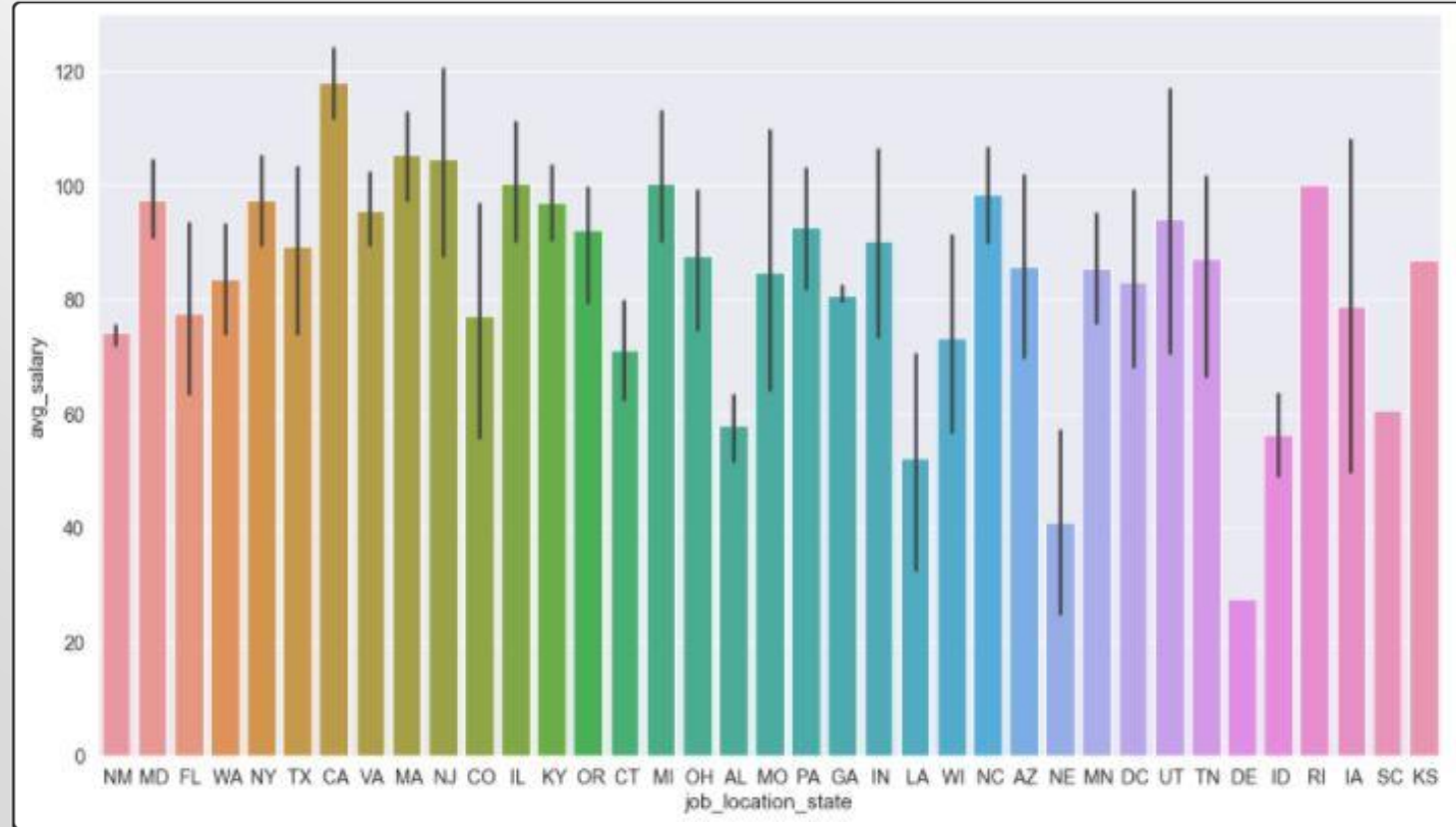
It is noted that
50% of job applications
have average salary
between 98000 , 14000 in
DS Jobs



Visualization

It is noted that

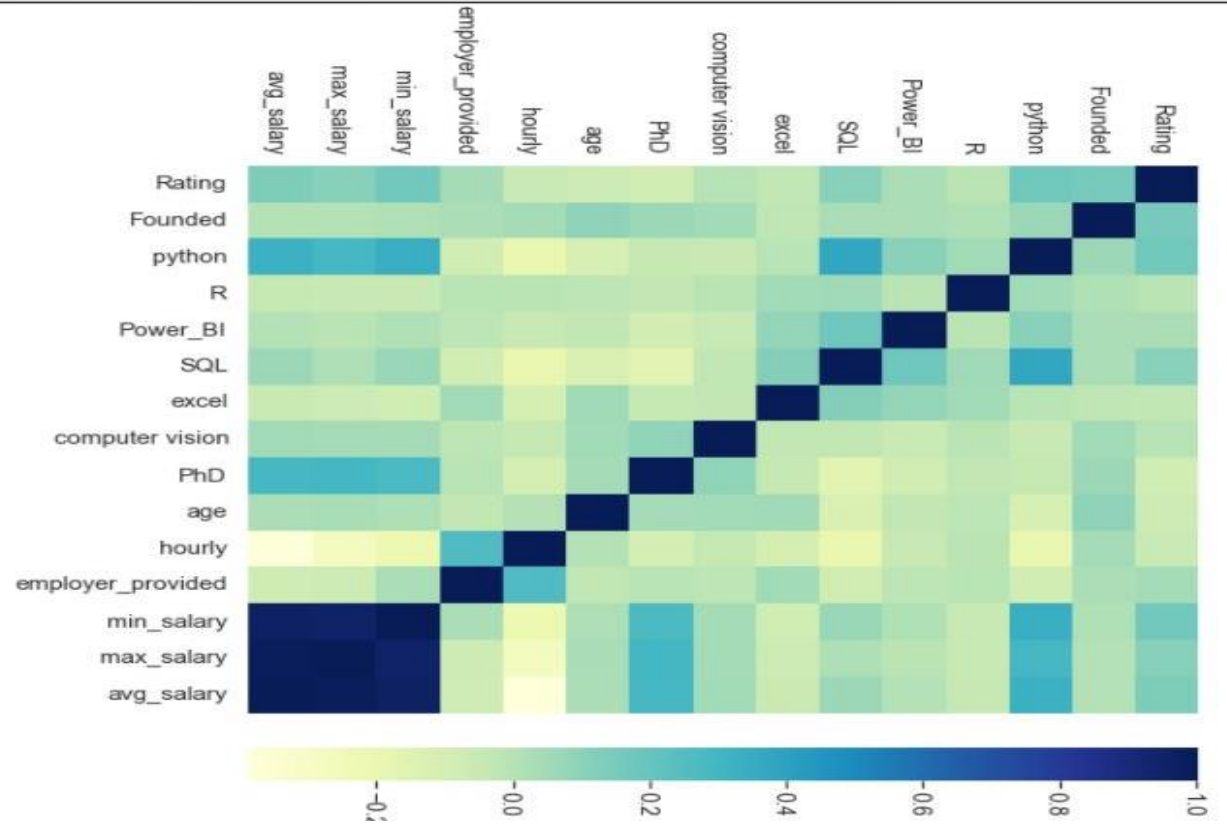
CA have the highest paid salary & Delaware the lowest one with 25000\$ /year



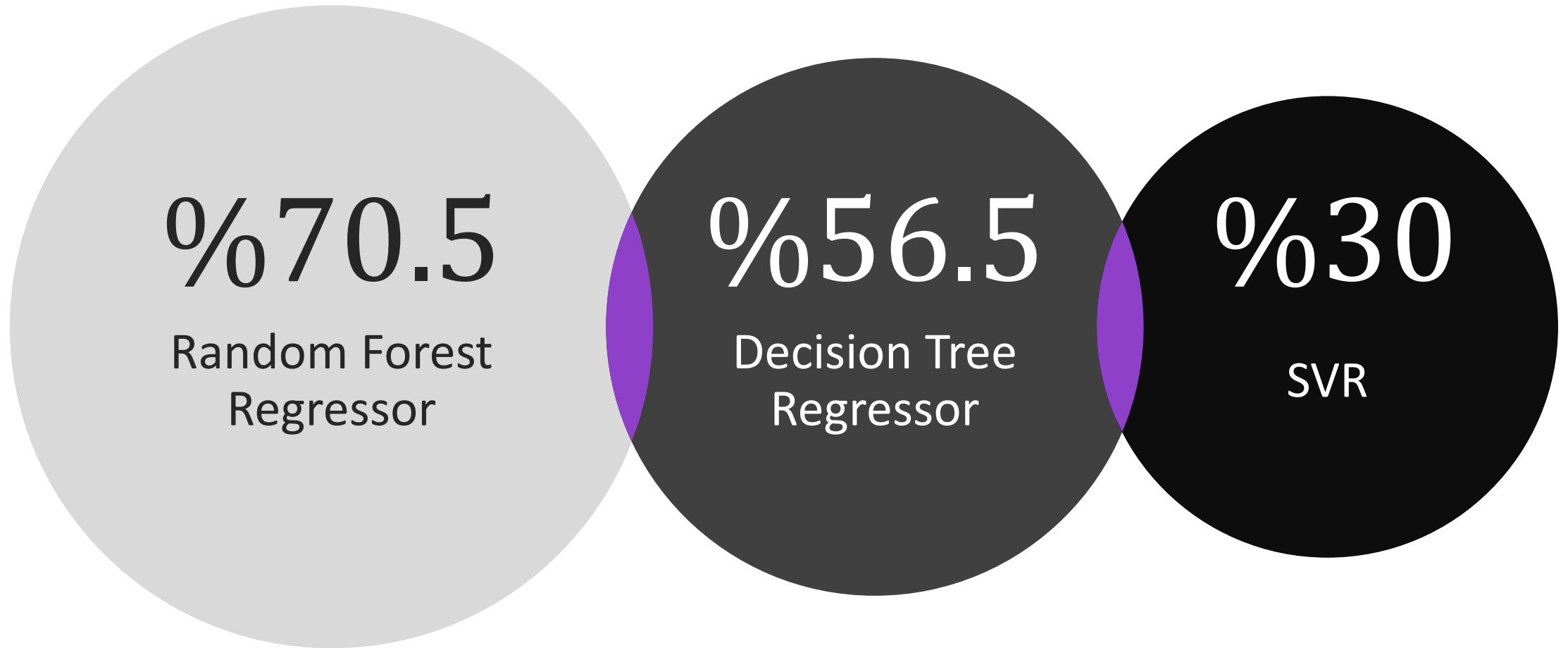
Visualization

It is noted that

There is no correlation between features & it's a good sign to build our machine learning model



ML Model Accuracy





Deployment

Create a web app with modern html interface to help data scientists job seeker to figure out the estimated salary that applies for it with Heroku platform.

Link:

<https://ds-jobs-salary-pred-with-ml.herokuapp.com>

Biography

|

A Highly competent Data Scientist with Engineering Degree & Quality Control experience background.

Have a certificate in Data Analysis and Machine Learning.

Deploying complex machine learning and statistical modeling algorithms/techniques for identifying patterns and extracting valuable insights.



Maged
Elsabagh

Cons. Engineer | Data Scientist

┌ Epsilon.

Thank You

👤 Maged ElSabagh

📞 +2 0122-293-2663

✉ Maged_magdy9000@gmail.com

🔗 [linkedin.com/in/maged-magdy-00210](https://www.linkedin.com/in/maged-magdy-00210)

