Salary Analysis and Prediction Case Study

- *Analyzing trends, predicting salaries, and uncovering insights through data.*

M. Rahul 142201022
G. Sai Rohith 142201019
B.Sai Punith 142201021

Problem Statement:

- Salaries vary widely across roles, industries, and experience levels.
- job seekers need insights to benchmark salaries and identify key drivers.

Key Implementation Steps :

- Perform a detailed case study to understand salary trends.
- Build machine learning models to predict salaries.
- Extract actionable patterns through association rules.

Dataset Overview :

- We have Prepared dataset for dynamic salary analysis.
- The datasets we used are Fully preprocessed and cleaned data, ready for analysis.

Key Features:

- Categorical Features: Include roles, industries, locations, etc., for grouping and comparison.
- Numerical Features: Include salary, experience, and other measurable metrics.
- Skill-Specific Columns: Highlighting technical skills like Python, SQL, Excel, etc., for skill-based analysis.

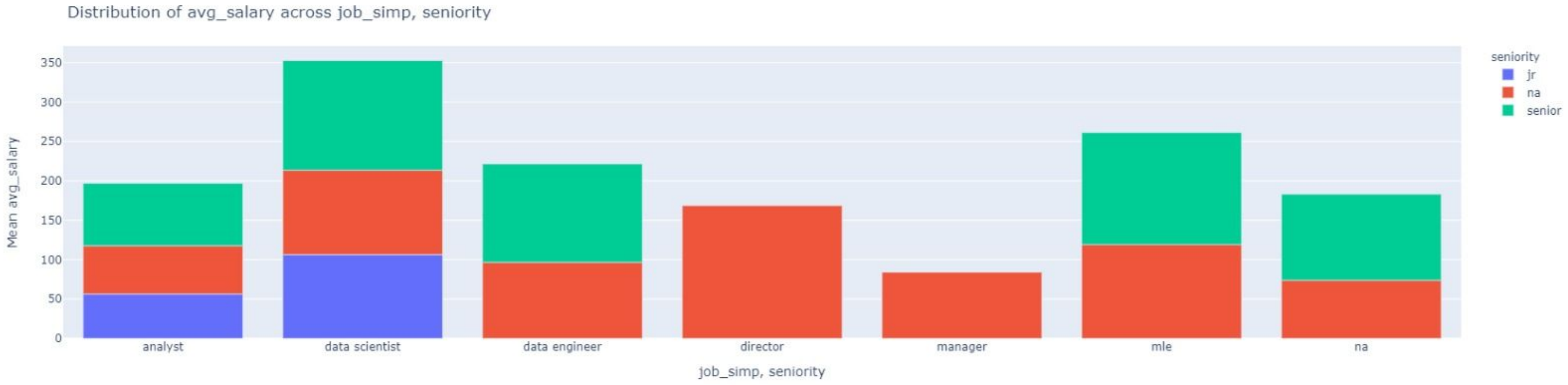
Preprocessing Steps:

- We have converted categorical columns to numerical using label encoders

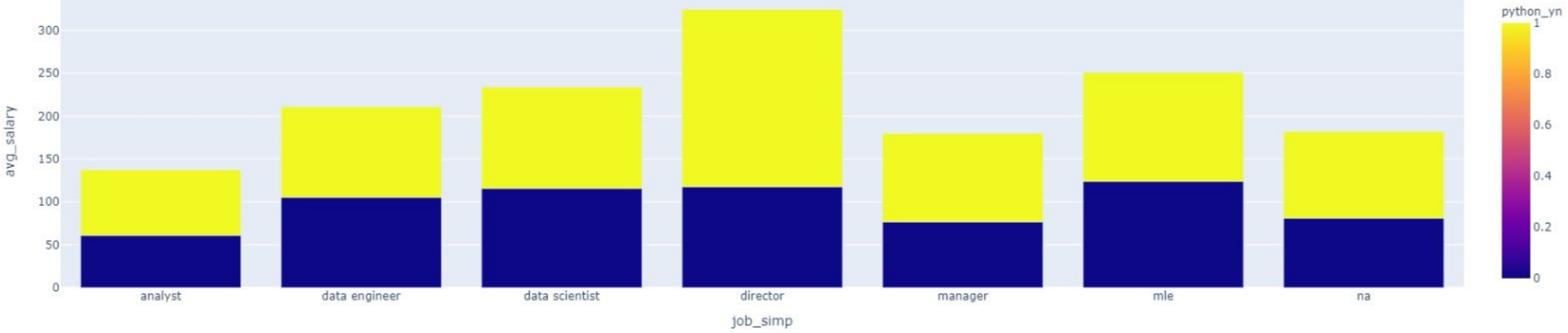
Case Study :

- User has the option to select some combination of features out of listed categorical , numerical and skill based features

- Then user can select type of graph for that combination , if that graph requirements are matched then graph is generated.
- Some examples are :

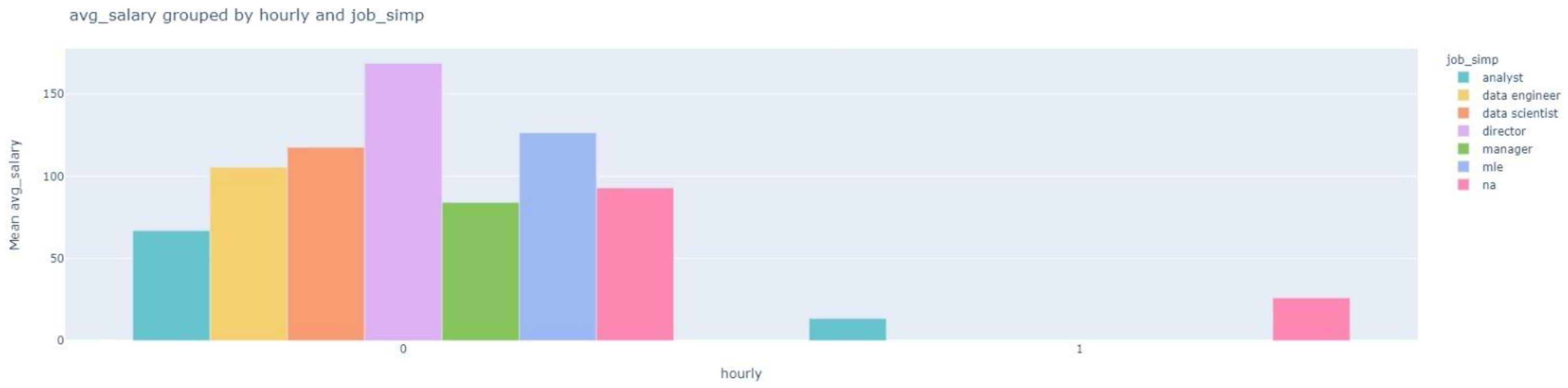


Distribution of avg_salary across job_simp and skill-python_yn



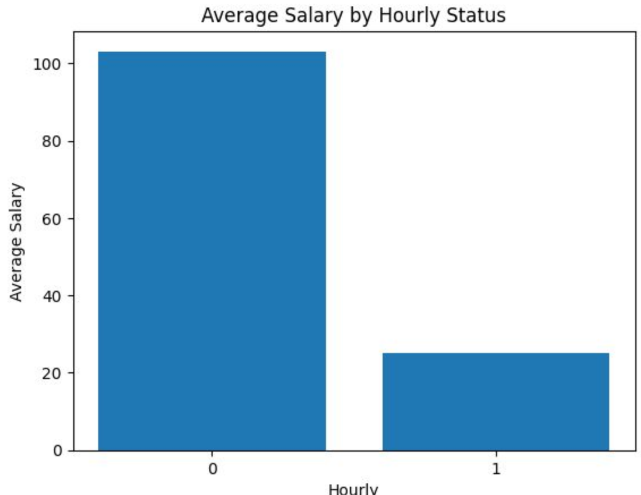
Distribution of avg_salary across job_simp

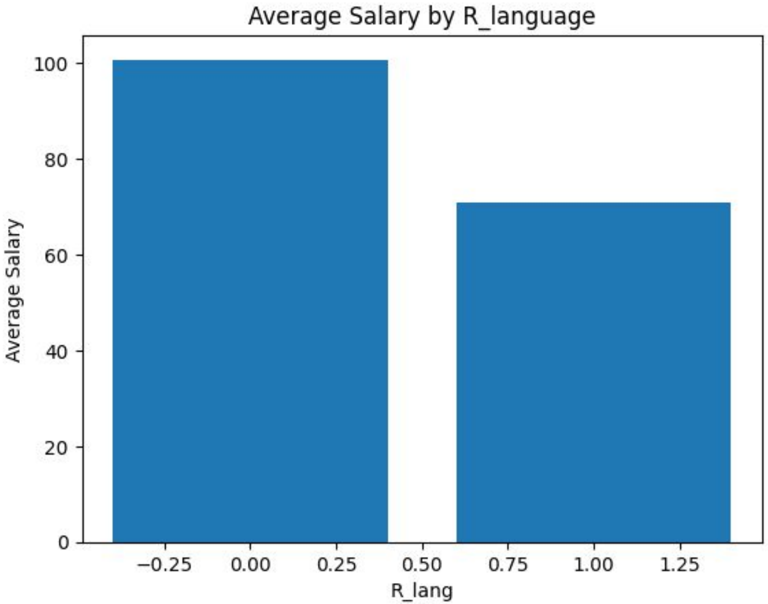
Some insights on static dataset:

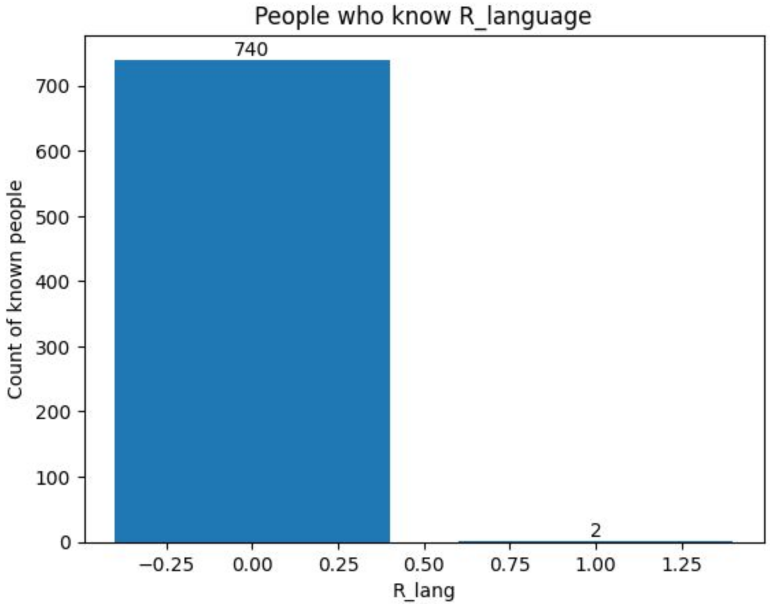


- On an average free lancers are earning less than the regular employees

- Data engineer,data scientist,manager,mle are not working on an hourly basis





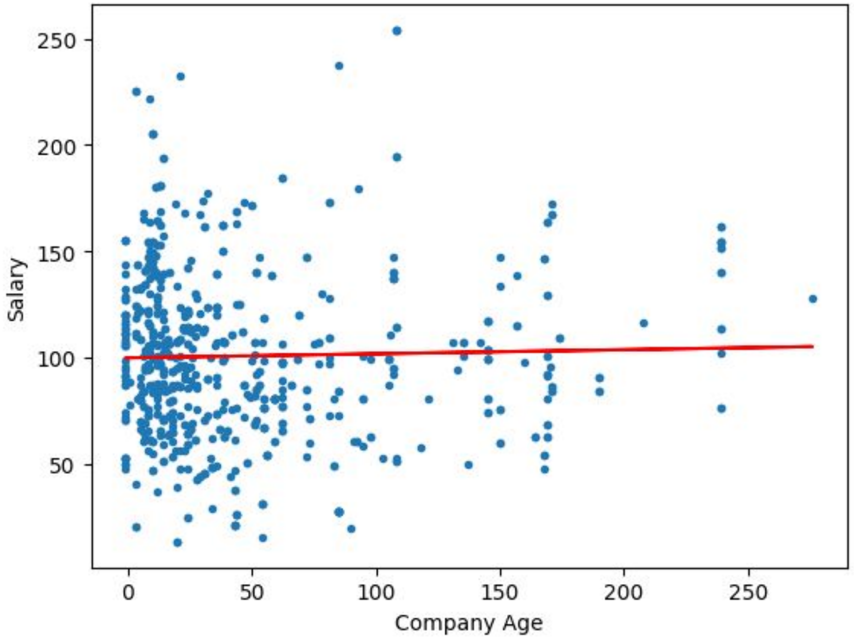


- R_language is not mandatory in many of the job applications

- The people who know R are earning ¾ of the avg_salary of people who don't know R

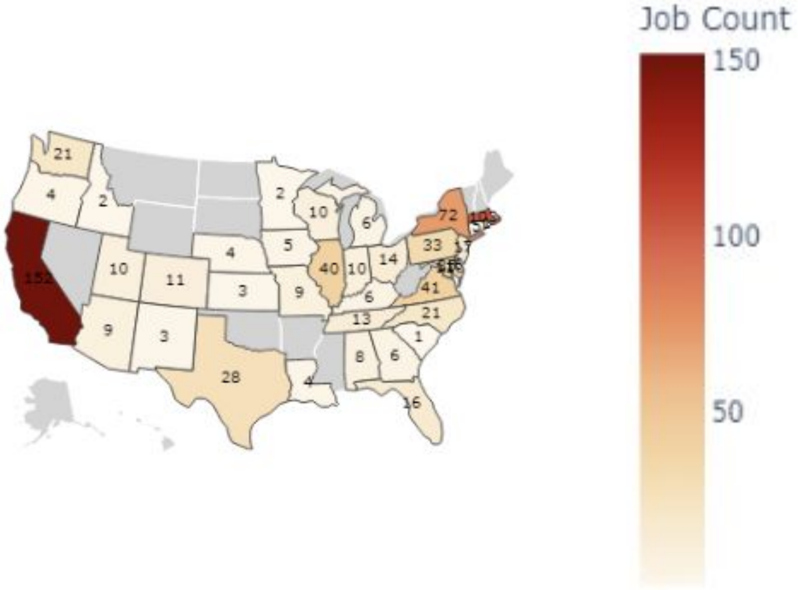
Company age vs Avg_salary:

- The flat regression line suggests a **weak or no correlation** between company age and salary.
- Company age has minimal influence on salary levels based on this dataset.
- Salaries appear to be more dependent on other factors like job role, location, or company size.



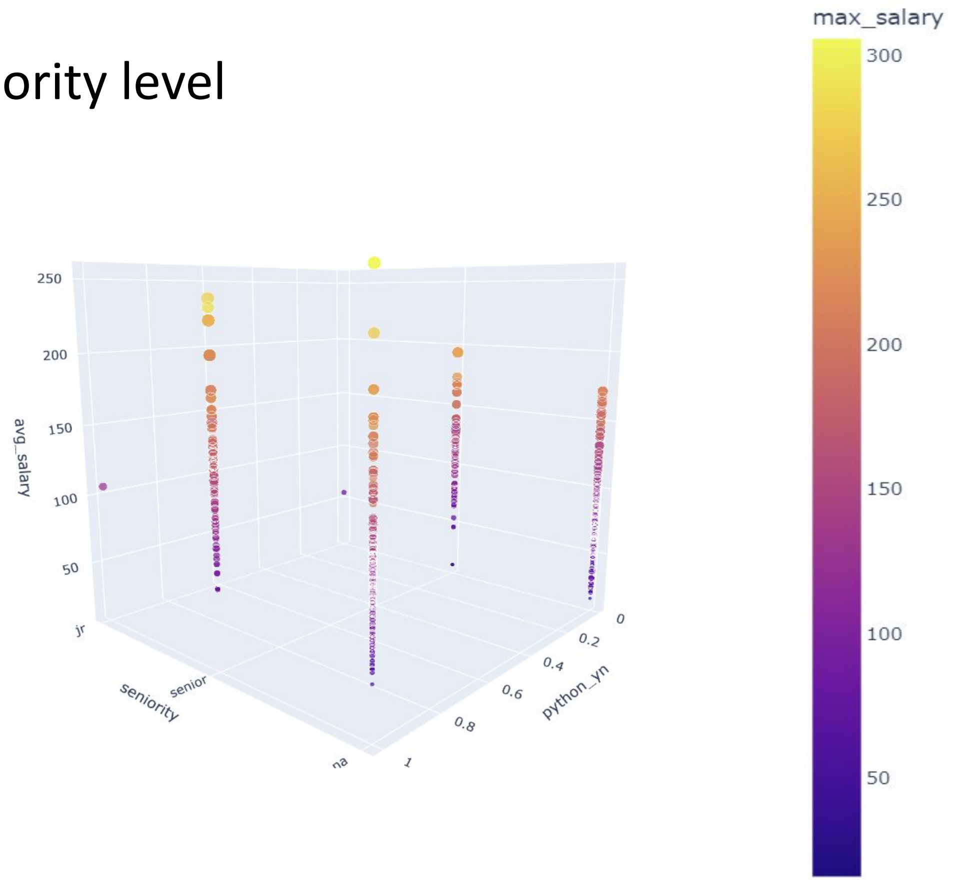
Number of Jobs by state :

California created 14,700 new jobs in September, averaging 16,500 new jobs per month this year, as the state's economy has grown faster than the nation's over the past 25 years. If California were a nation it would rank in terms of nominal GDP as the world's fourth largest economy.



Python has more weight than seniority level

seniority	python_yn	avg_salary
jr	0	56.5$
jr	1	106.5$
na	0	79.72$
na	1	103.32$
sr	0	107.06$
sr	1	132.7$



SALARY PREDICTION :

- Models used: Random Forest, Linear Regression, Ridge, SVR.
- Feature selection: users can select features for training based on correlation values
- Key features include job title, remote ratio, and experience level.
- Evaluation metric: Mean Squared Error (MSE).R2 score
- A plot of actual value vs predicted value will be shown

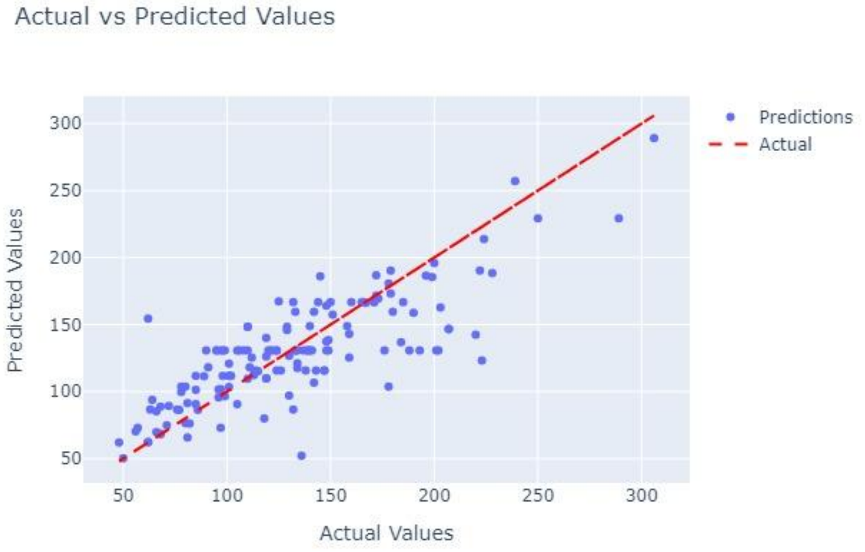
Prediction for the dataset

Target Variable: avg_salary

Correlation Table

Select	Feature	Correlation with avg_salary
☐	avg_salary	1.0
☐	max_salary	0.99
☐	min_salary	0.97
☐	seniority	0.35
☐	python_yn	0.33
☐	Job Title	0.22
☐	spark	0.18
☐	aws	0.17
☐	num_comp	0.09
☐	Competitors	0.08
☐	employer_provided	0.08
☐	Headquarters	0.02
☐	age	0.02
☐	Rating	0.01
☐	Revenue	-0.01
☐	Location	-0.01
☐	Founded	-0.02
☐	Size	-0.02
☐	same_state	-0.03
☐	R_yn	-0.04
☐	job_simp	-0.05
☐	excel	-0.06
☐	Sector	-0.07
☐	Industry	-0.1
☐	Type of ownership	-0.13
☐	job_state	-0.19
☐	hourly	-0.36

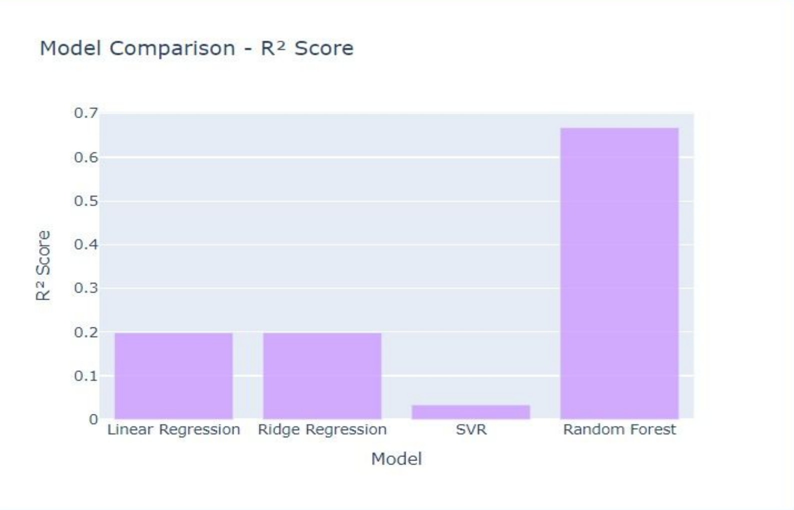
Train Model



Model Training Results

Mean Squared Error (MSE)

R^2 Score



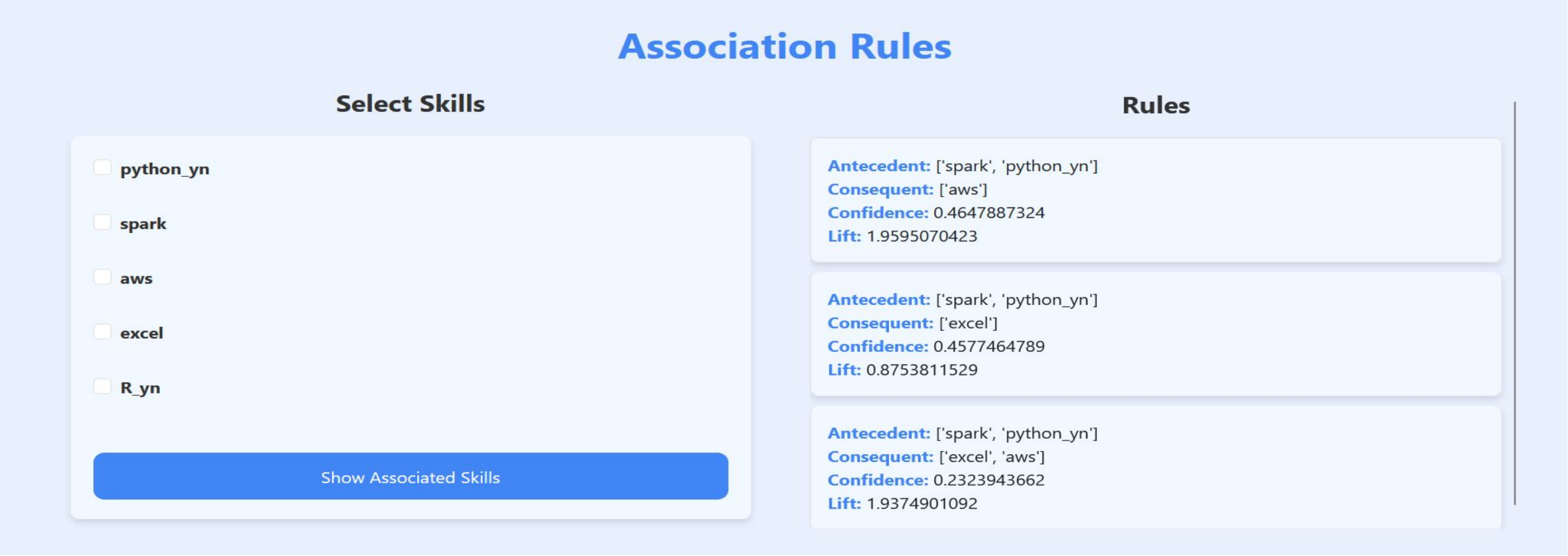
Select a Model to Use for Prediction:

Linear Regression

Go to Prediction

Association Rules :

- Apriori algorithm and association_rules for generating rules.
- Metrics like support, confidence, and lift.



Contributions:

- Sai Rohith: html,python and styling related to prediction,some insights in static dataset

- Rahul : html,python and styling related to case study,home page

- Sai Punith:html,python and styling related to association rules,slides

THANK YOU