AUSTRALIA DATA SCIENCE JOBS

PAULINE ARMAMENTO

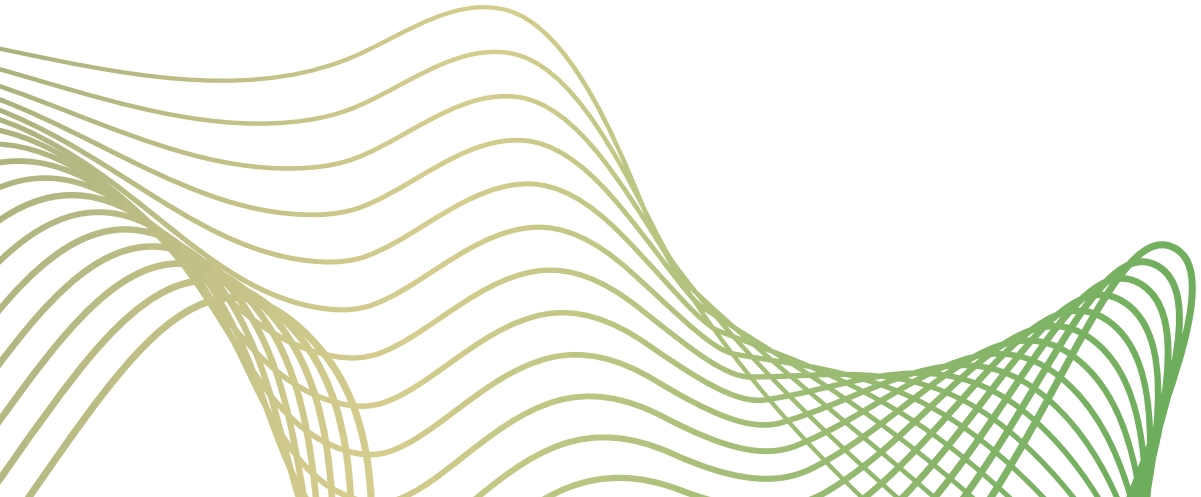
u 3 2 4 6 7 8 2

OUTLINE

1. Background
2. EDA
3. Predictive Analytics
4. Implementation and Deployment
5. References

Introduction

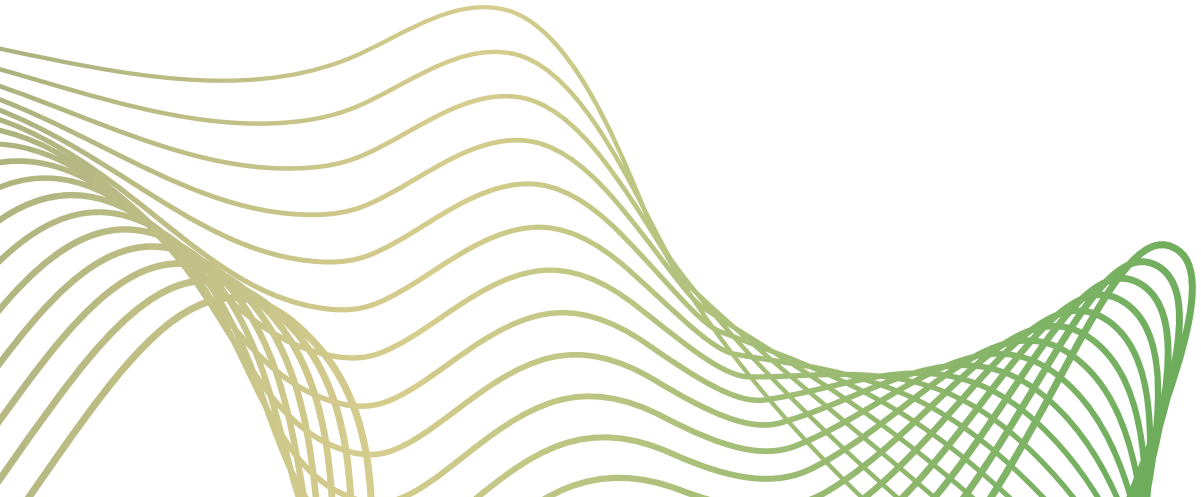
This report explores the Australia Data Science Jobs Dataset for data analytics, visualization, prediction and deployment. We performed Exploratory Data Analysis (EDA) on five identified problems in the dataset and visualized by graphs and plots.



Dataset Details

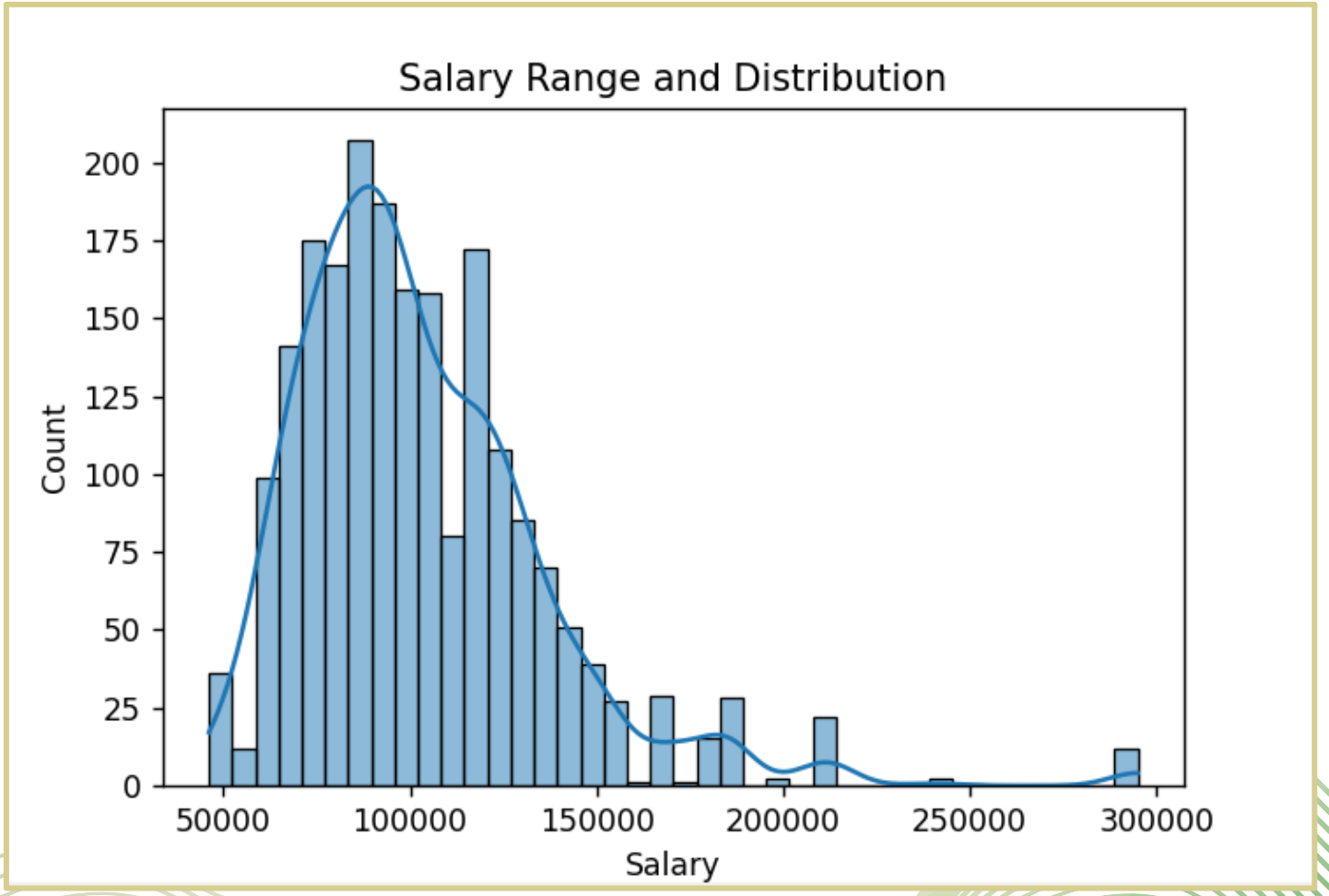
Australia Data Science Jobs Dataset

This job listings data[1] was collected from Glassdoor through web scraping algorithm Selenium (Python). It consists of 2088 observations and 53 variables. It provides valuable information regarding job titles, locations, companies, salary estimates, company ratings and required skills.

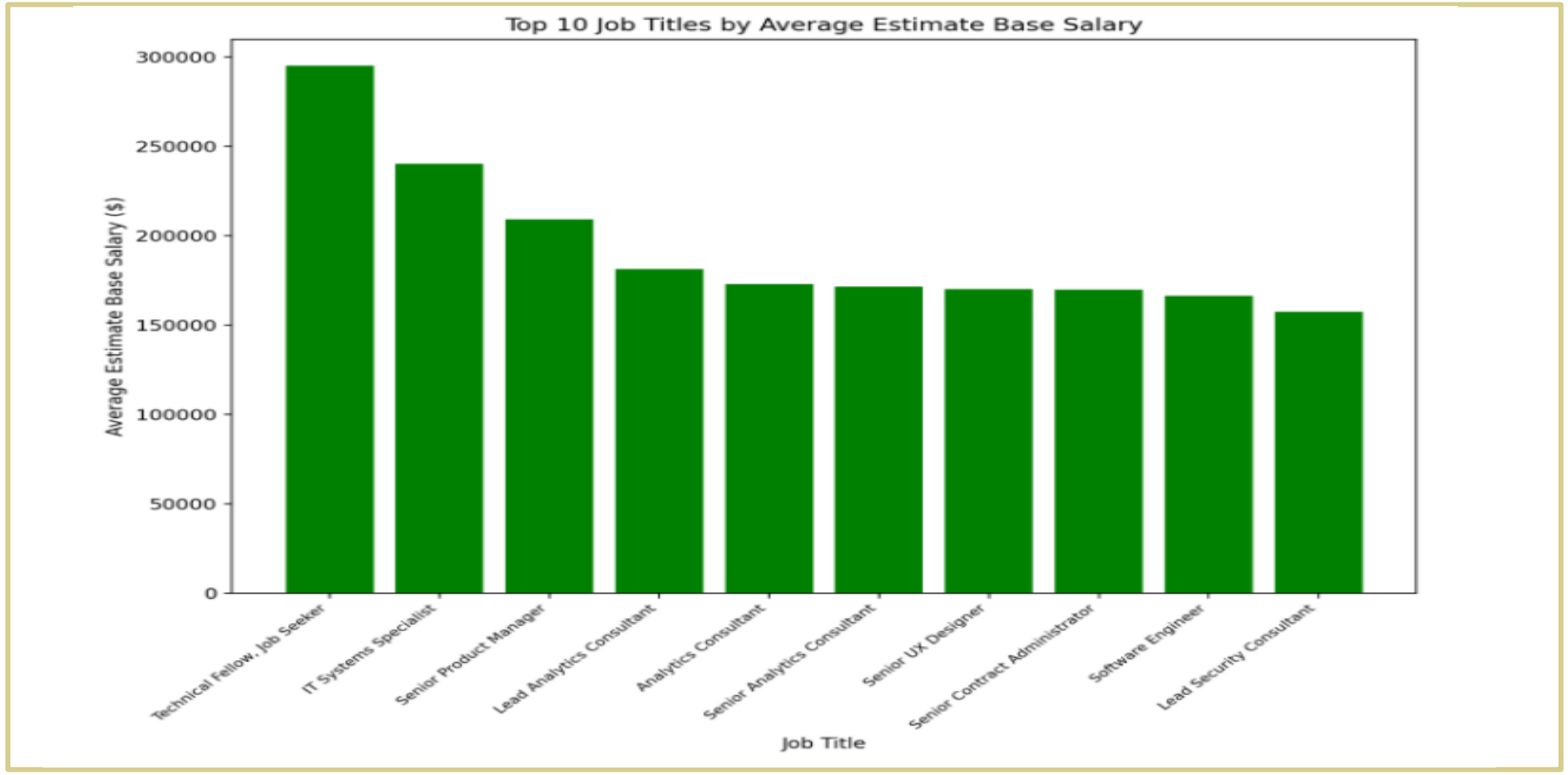


1. What is the salary distribution among jobs in the dataset?

The histogram shows that most number of jobs in the dataset have an estimate base salary around 80,000 to 90,000 with very few outliers in the range of 200,000 to 300,000. Acquiring such knowledge enables us to comprehend the process of pay scale levels and keeps us aligned with the compensation hierarchy.



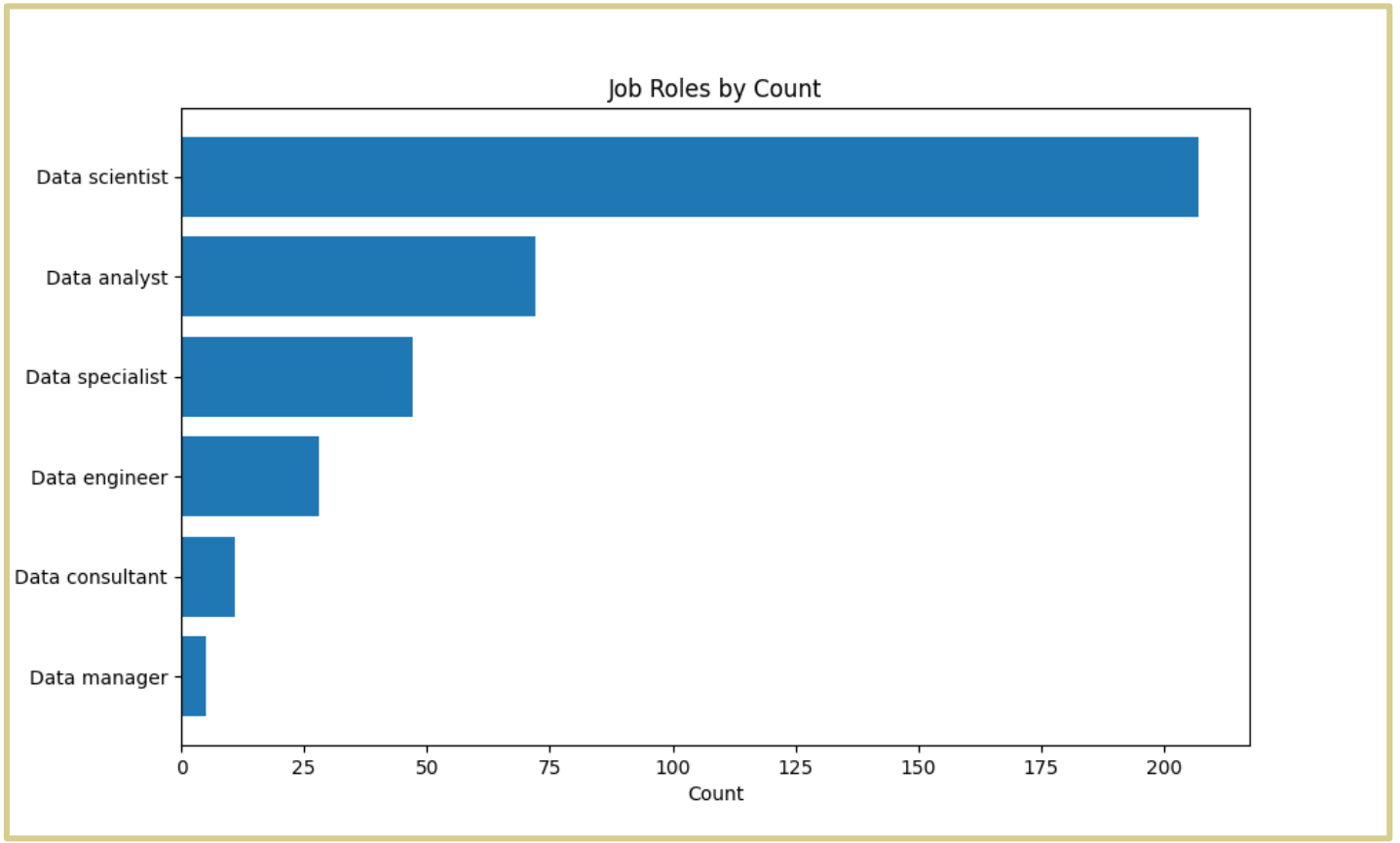
2. Which Job Titles has the highest estimated salary?



This graph displays the top occupations with the highest estimated base pay in the dataset was generated. It allows us to discern which employment positions possess substantial earning potential as previously shown in our first plot.

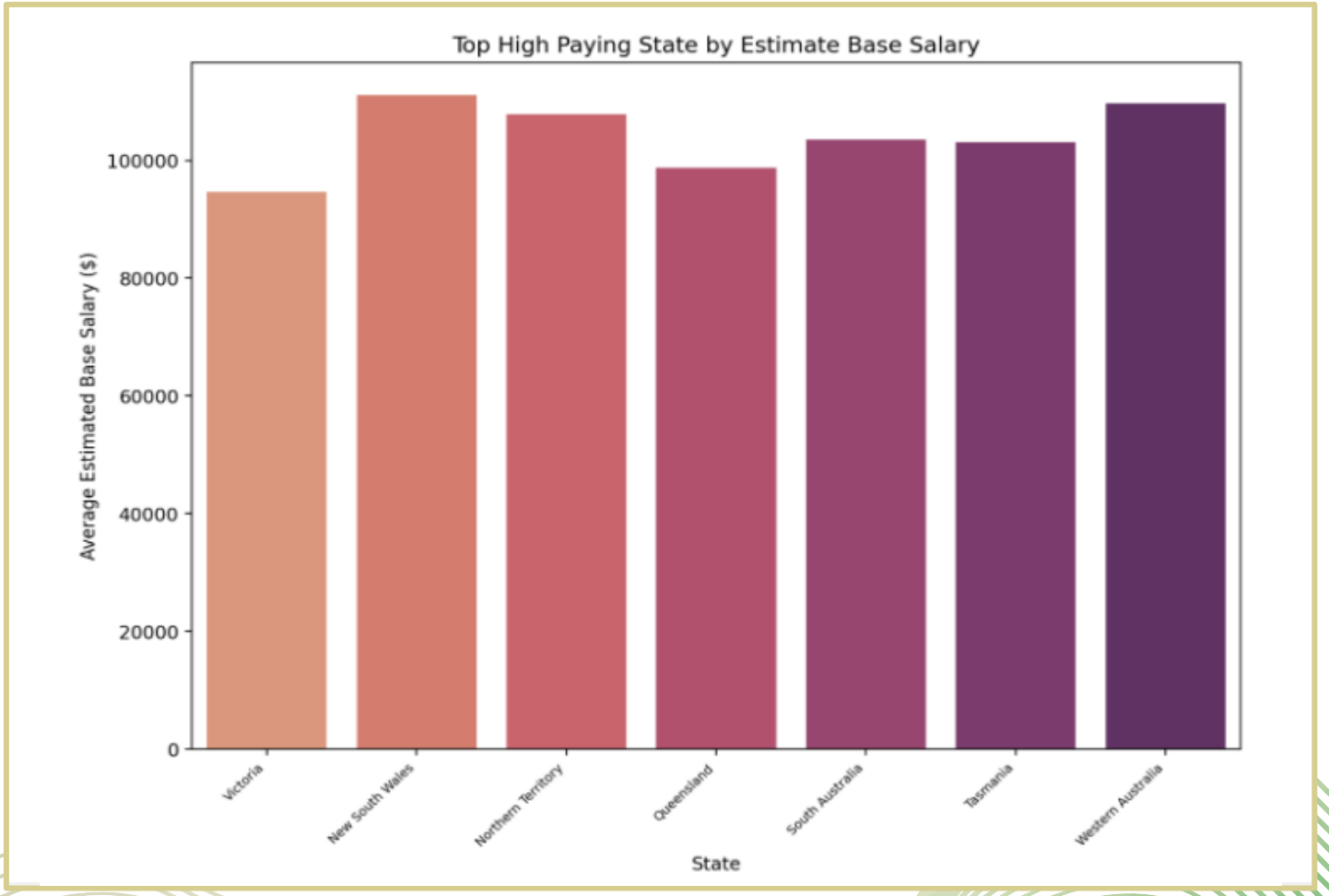
3. Determine the location that has the greatest concentration of data specialists in employment.

The research revealed that while Melbourne had the greatest number of data specialist job vacancies, Sydney held the highest volume of such postings. This information offers insight into where a significant proportion of professionals with expertise in data analysis are situated within Australia.



4. Which is the highest paying State in terms of estimated salary?

We found that New South Wales offers the highest estimated base salary among all states, followed by Western Australia and Northern Territory. This information provides valuable insights into which states provide the most significant compensation with regards to estimate base salaries.

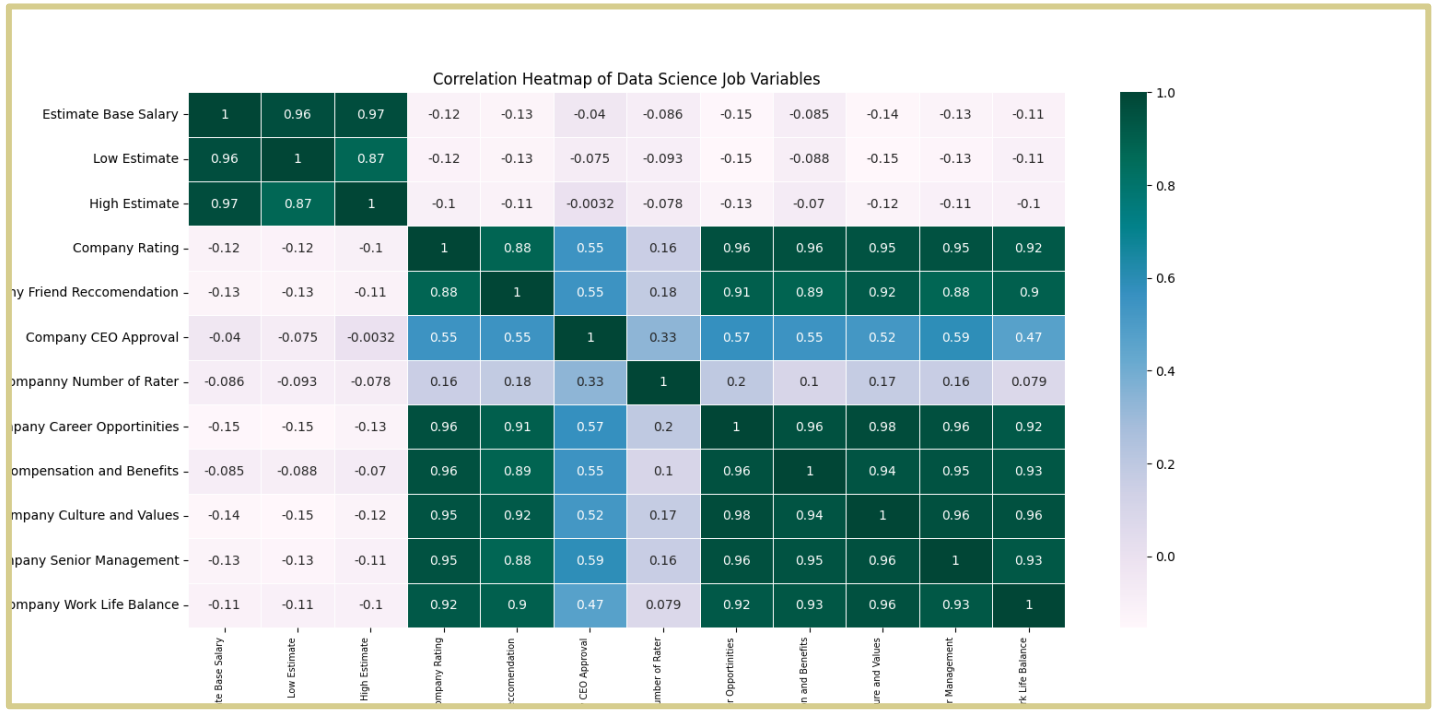


5. Identify which variables are highly correlated to Estimated Base Salary

It was observed that the variables Low Estimate and High Estimate exhibit a positive correlation with the target variable. Conversely, all other variables display a notable negative correlation with the target variable.

Random Forest Classifier

We used the popular machine learning algorithm Random Forest Classifier for this task due to its reputation for generating highly precise outcomes, it is able to handle missing values and outliers well making it a reliable choice.



Predictive Analytics

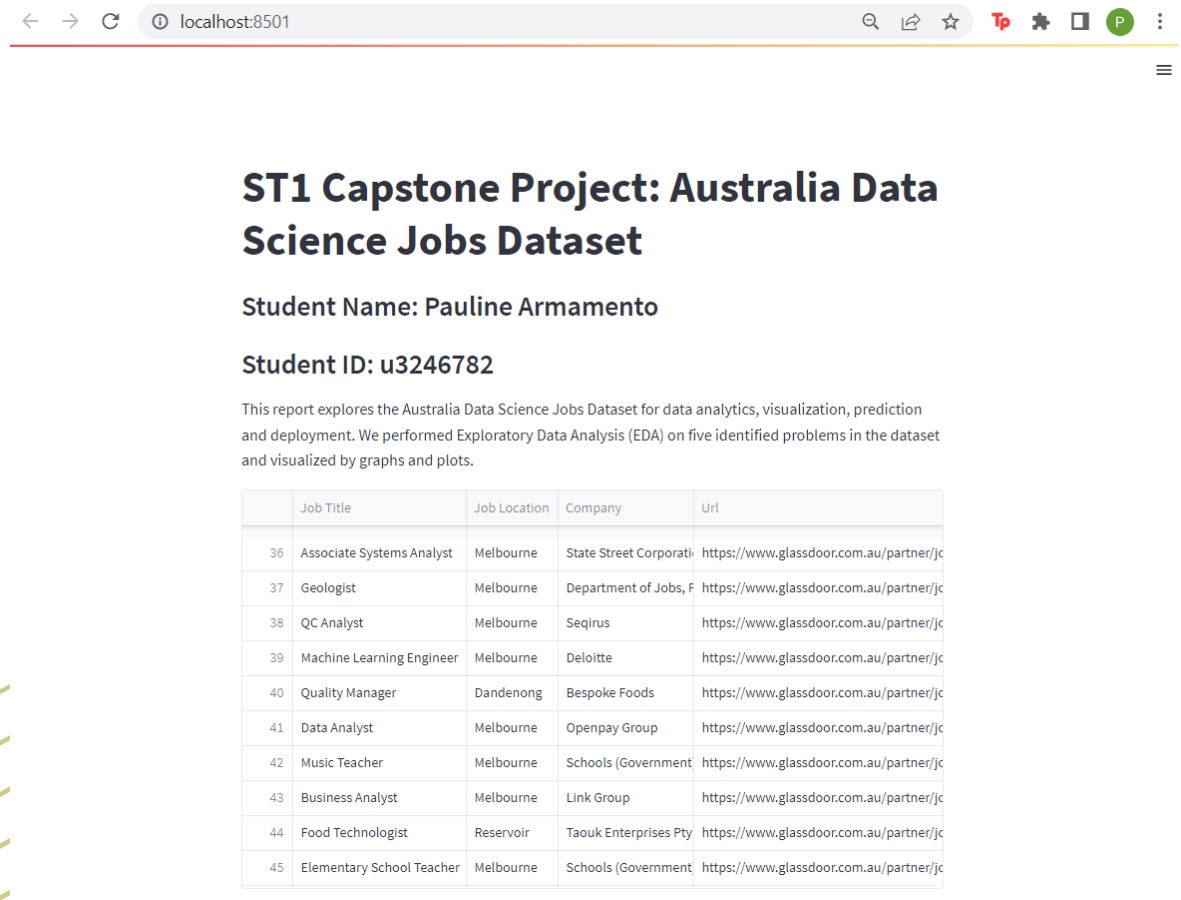
The random forest classifier's accuracy rate of 0.87 indicates that it can successfully anticipate the accurate result of the test set observations. This implies that the model displays an accuracy level of approximately 87% when predicting the correct variable label for previously unobserved data.

First 10 Predicted Values:

	0
0	125,000
1	116,190
2	101,242
3	135,000
4	134,747
5	127,500
6	90,951
7	72,675
8	102,500
9	183,000

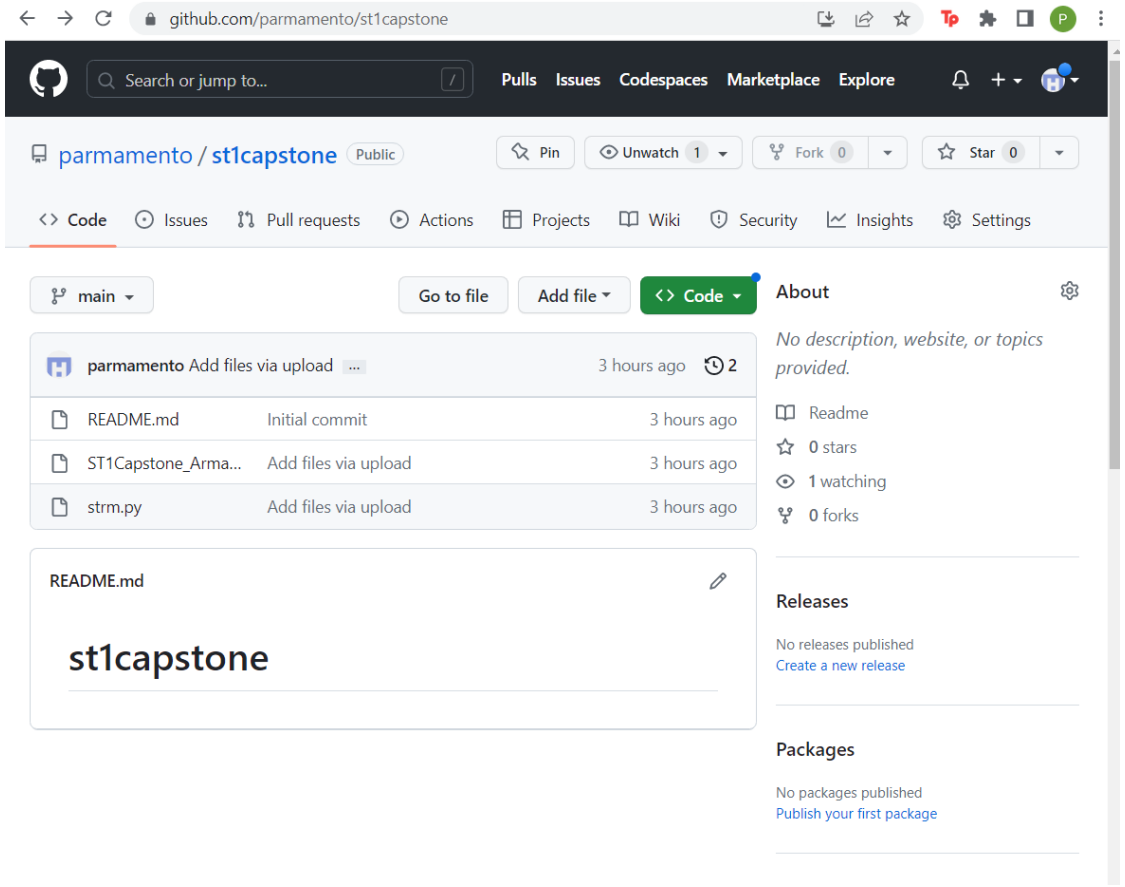
Accuracy Score on Test Set: 0.87

Streamlit Application



	Job Title	Job Location	Company	Url
36	Associate Systems Analyst	Melbourne	State Street Corporati	https://www.glassdoor.com.au/partner/jc
37	Geologist	Melbourne	Department of Jobs, F	https://www.glassdoor.com.au/partner/jc
38	QC Analyst	Melbourne	Seqirus	https://www.glassdoor.com.au/partner/jc
39	Machine Learning Engineer	Melbourne	Deloitte	https://www.glassdoor.com.au/partner/jc
40	Quality Manager	Dandenong	Bespoke Foods	https://www.glassdoor.com.au/partner/jc
41	Data Analyst	Melbourne	Openpay Group	https://www.glassdoor.com.au/partner/jc
42	Music Teacher	Melbourne	Schools (Government	https://www.glassdoor.com.au/partner/jc
43	Business Analyst	Melbourne	Link Group	https://www.glassdoor.com.au/partner/jc
44	Food Technologist	Reservoir	Taouk Enterprises Pty	https://www.glassdoor.com.au/partner/jc
45	Elementary School Teacher	Melbourne	Schools (Government	https://www.glassdoor.com.au/partner/jc

GitHub Repository



CONCLUSION

This project showcased the application of concepts learned in Data Analytics and Visualization through EDA, Prediction Analytics using machine learning algorithm Random Forest, and Deployment using Streamlit. It provided opportunities to exhibit our skills in data manipulation and visualization using python libraries such as pandas, numpy, seaborn and scikit-learn.

References:

[1] https://www.kaggle.com/datasets/nadzmiagthomas/australia-data-science-jobs?resource=download

[2] Cutler, Adele & Cutler, David & Stevens, John. (2011). Random Forests. 10.1007/978-1-4419-9326-7_5.

THANK YOU