PRESENTATION

START UP

Success Prediction

TABLE OF CONTENTS

01 About dataset

02 mission

03 vision

About Dataset

The dataset provides insights into the characteristics and outcomes of startups, such as their funding amounts, business categories, locations, and the number of funding rounds they have undergone. It highlights the current status of each startup, such as whether it is still operating, has been acquired, or has closed down. This data helps understand the factors contributing to a startup's success or failure, offering valuable information for investors, entrepreneurs, and analysts.

MISSION

To analyze and predict the success or failure of startups by leveraging data on funding, categories, and geographical factors. This analysis aims to provide actionable insights for investors, entrepreneurs, and policymakers, helping them make informed decisions to foster innovation and economic growth.

VISION

To empower decision-makers in the startup ecosystem with data-driven insights that enhance their ability to identify promising opportunities, mitigate risks, and drive sustainable growth. Through advanced analytics, we aim to contribute to the success of startups worldwide, fostering innovation and economic progress across industries and regions.

Problem Statement:

Startups face high risks of failure, often due to factors such as insufficient funding, poor market fit, or unfavorable locations. Investors and entrepreneurs struggle to predict the likelihood of a startup's success or failure. The challenge is to develop a predictive model that can accurately determine the status of a startup (operating, closed, or acquired) based on key features such as funding, category, and location. Solving this problem will enable more informed decisions, reducing risks and fostering the growth of successful ventures.

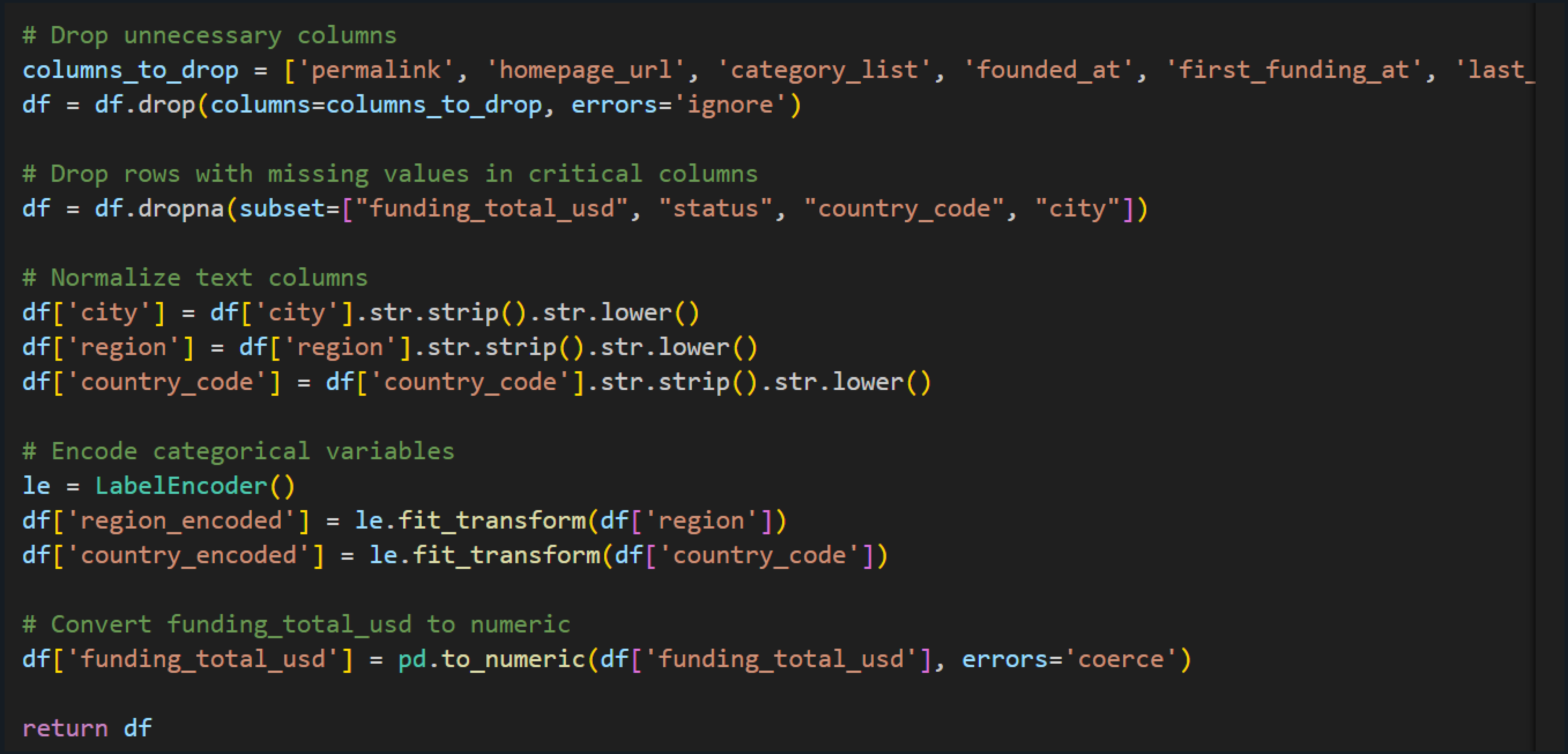
Overview of Files

1. app.py : Core Streamlit application.
2. main.py: Main script for data processing and modeling.
3. Supporting modules:
   - DataLoader : Handles data loading.
   -DataPreprocessor: Prepares and cleans datasets.
   -Plotter: Generates visualizations.
   - ModelTrainer: Manages machine learning model training and evaluation.

Steps :

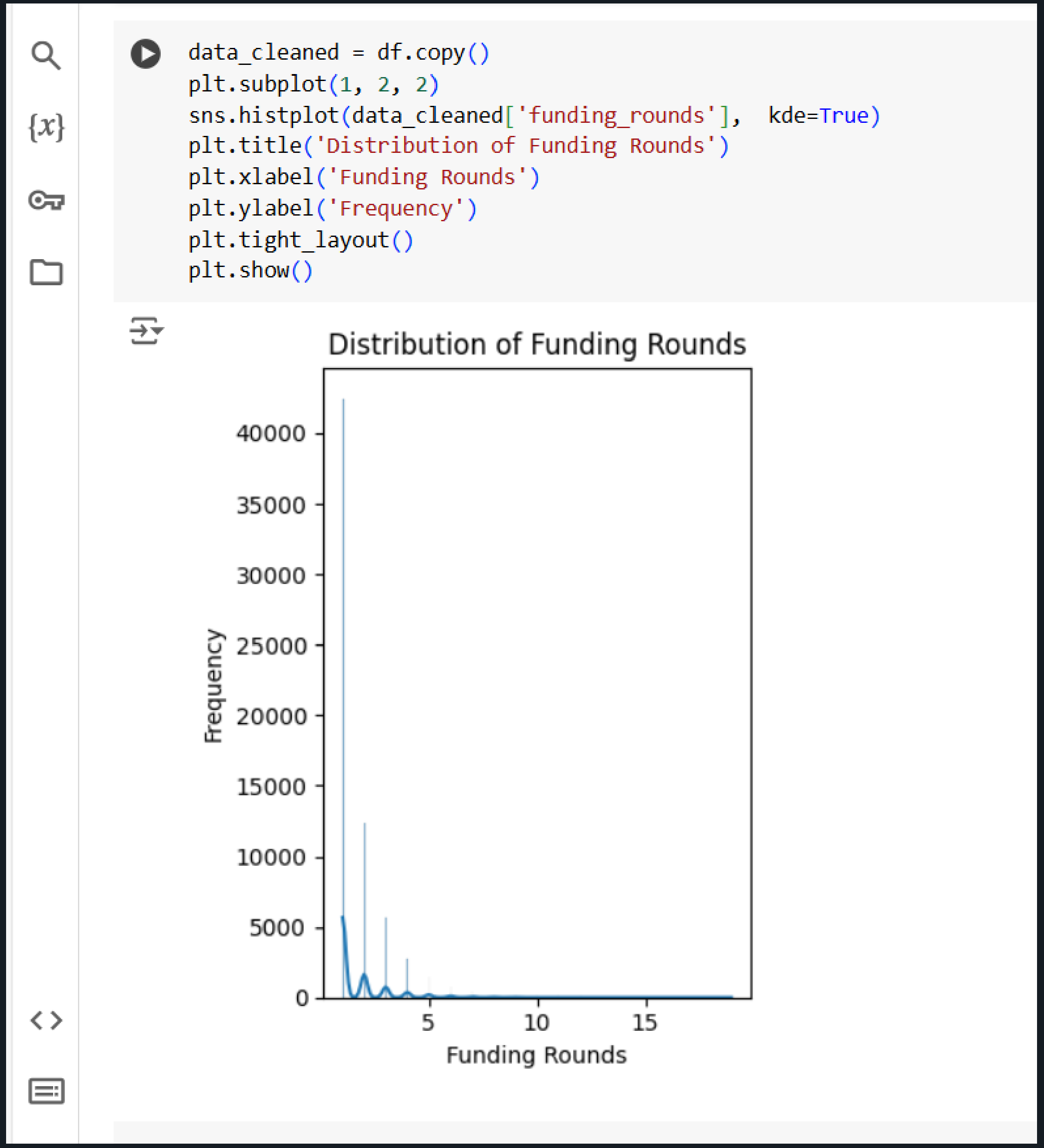
1. **Load and Preprocess Data:**
   - Loads raw data using DataLoader.
   - Cleans and saves the processed dataset

2. Data Cleaning –
Handling Missing Values and Conversion of Data Types

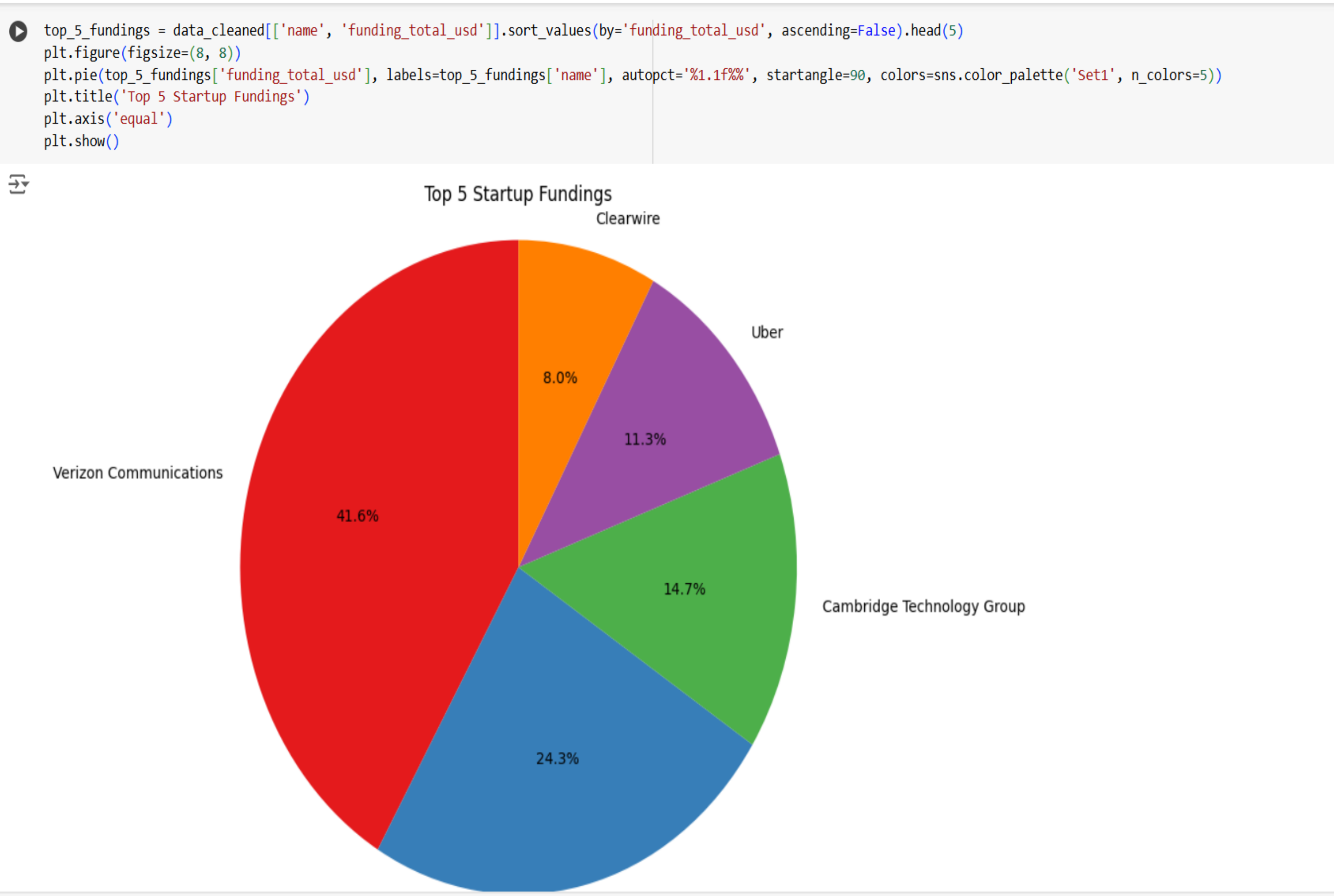


```
# Drop unnecessary columns
columns_to_drop = ['permalink', 'homepage_url', 'category_list', 'founded_at', 'first_funding_at', 'last_
df = df.drop(columns=columns_to_drop, errors='ignore')

# Drop rows with missing values in critical columns
df = df.dropna(subset=["funding_total_usd", "status", "country_code", "city"])

# Normalize text columns
df['city'] = df['city'].str.strip().str.lower()
df['region'] = df['region'].str.strip().str.lower()
df['country_code'] = df['country_code'].str.strip().str.lower()

# Encode categorical variables
le = LabelEncoder()
df['region_encoded'] = le.fit_transform(df['region'])
df['country_encoded'] = le.fit_transform(df['country_code'])

# Convert funding_total_usd to numeric
df['funding_total_usd'] = pd.to_numeric(df['funding_total_usd'], errors='coerce')

return df
```

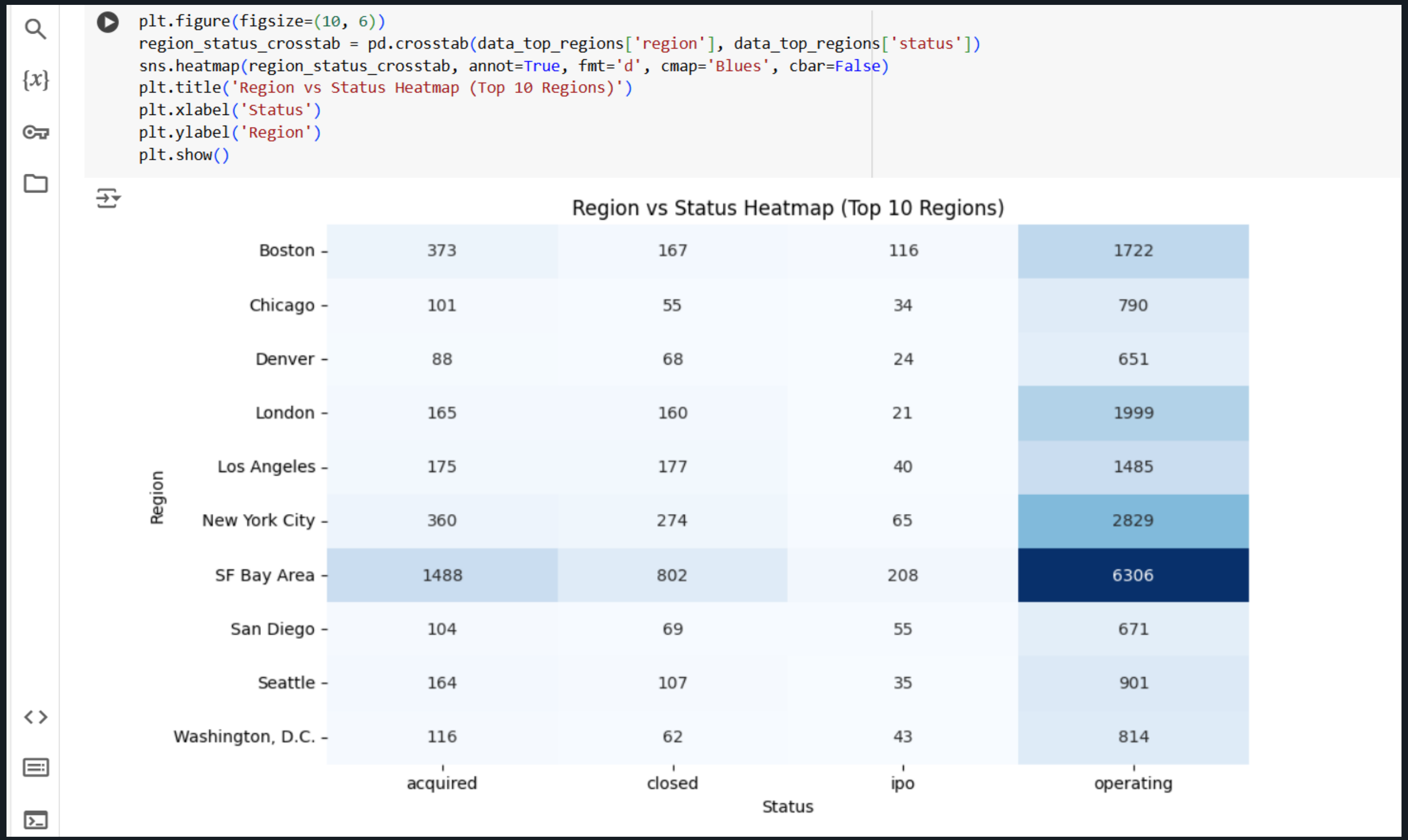
3. Exploratory Data Analysis (EDA) –



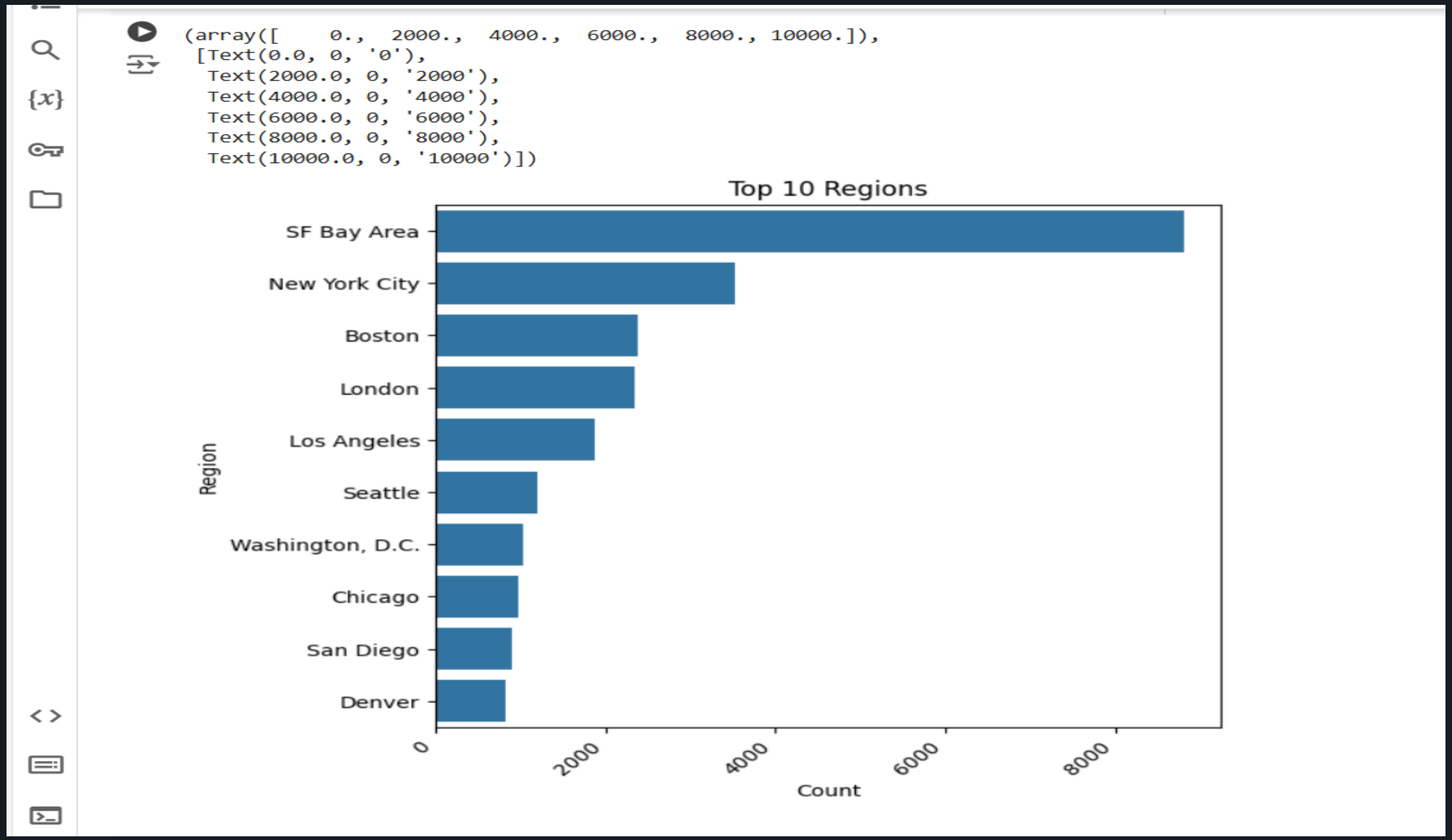
```
data_cleaned = df.copy()
plt.subplot(1, 2, 2)
sns.histplot(data_cleaned['funding_rounds'],  kde=True)
plt.title('Distribution of Funding Rounds')
plt.xlabel('Funding Rounds')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```



```
top_5_fundings = data_cleaned[['name', 'funding_total_usd']].sort_values(by='funding_total_usd', ascending=False).head(5)
plt.figure(figsize=(8, 8))
plt.pie(top_5_fundings['funding_total_usd'], labels=top_5_fundings['name'], autopct='%1.1f%%', startangle=90, colors=sns.color_palette('Set1', n_colors=5))
plt.title('Top 5 Startup Fundings')
plt.axis('equal')
plt.show()
```



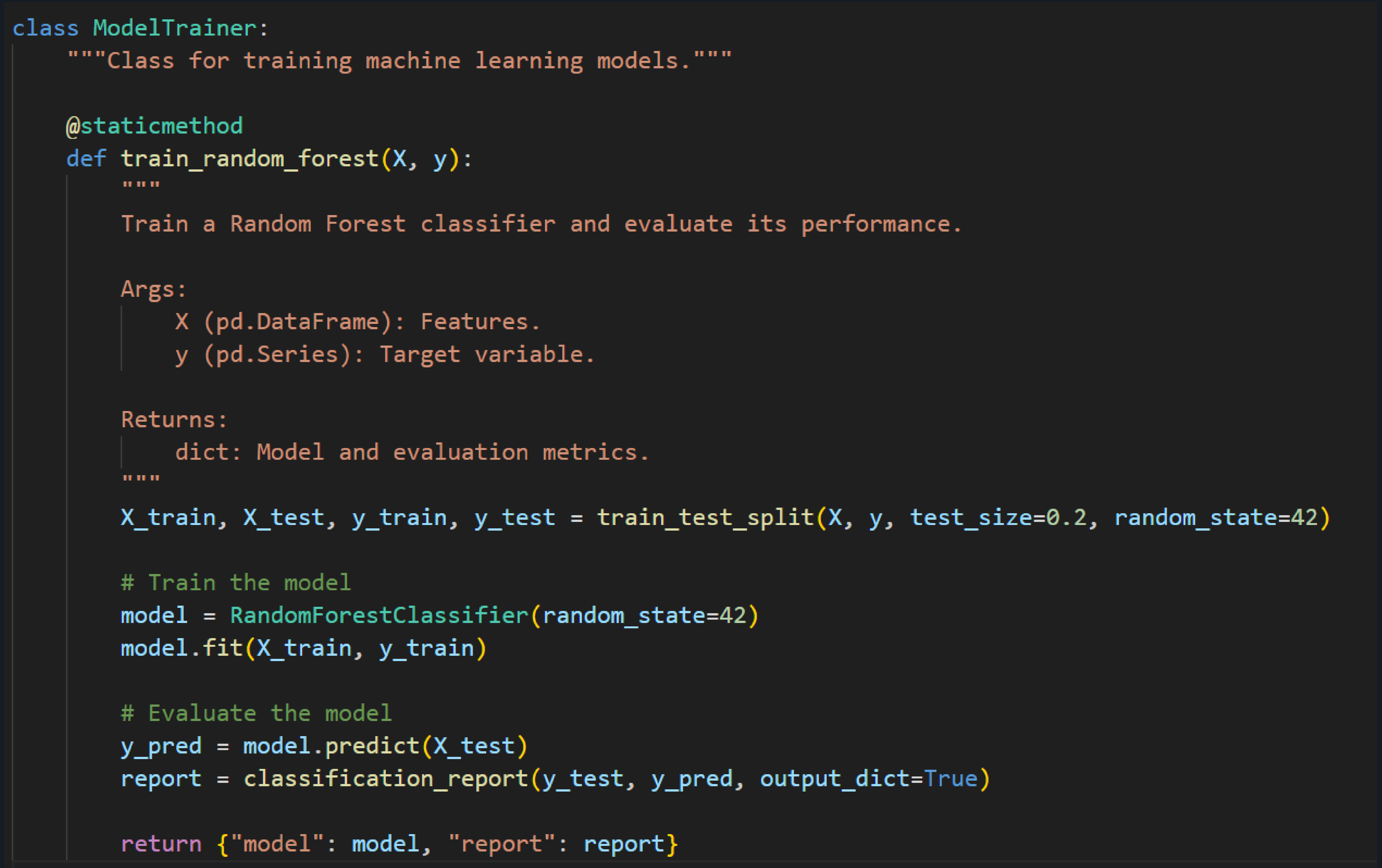
```
plt.figure(figsize=(10, 6))
region_status_crosstab = pd.crosstab(data_top_regions['region'], data_top_regions['status'])
sns.heatmap(region_status_crosstab, annot=True, fmt='d', cmap='Blues', cbar=False)
plt.title('Region vs Status Heatmap (Top 10 Regions)')
plt.xlabel('Status')
plt.ylabel('Region')
plt.show()
```



```
(array([    0.,  2000.,  4000.,  6000.,  8000., 10000.]),
 [Text(0.0, 0, '0'),
  Text(2000.0, 0, '2000'),
  Text(4000.0, 0, '4000'),
  Text(6000.0, 0, '6000'),
  Text(8000.0, 0, '8000'),
  Text(10000.0, 0, '10000')])
```

4. **Model Training:**

- Prepares data for training using Helpers.
- Trains a Random Forest model via ModelTrainer.
- Evaluates the model and prints performance reports.



```python
class ModelTrainer:
    """Class for training machine learning models."""

    @staticmethod
    def train_random_forest(X, y):
        """
        Train a Random Forest classifier and evaluate its performance.

        Args:
            X (pd.DataFrame): Features.
            y (pd.Series): Target variable.

        Returns:
            dict: Model and evaluation metrics.
        """
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

        # Train the model
        model = RandomForestClassifier(random_state=42)
        model.fit(X_train, y_train)

        # Evaluate the model
        y_pred = model.predict(X_test)
        report = classification_report(y_test, y_pred, output_dict=True)

        return {"model": model, "report": report}
```

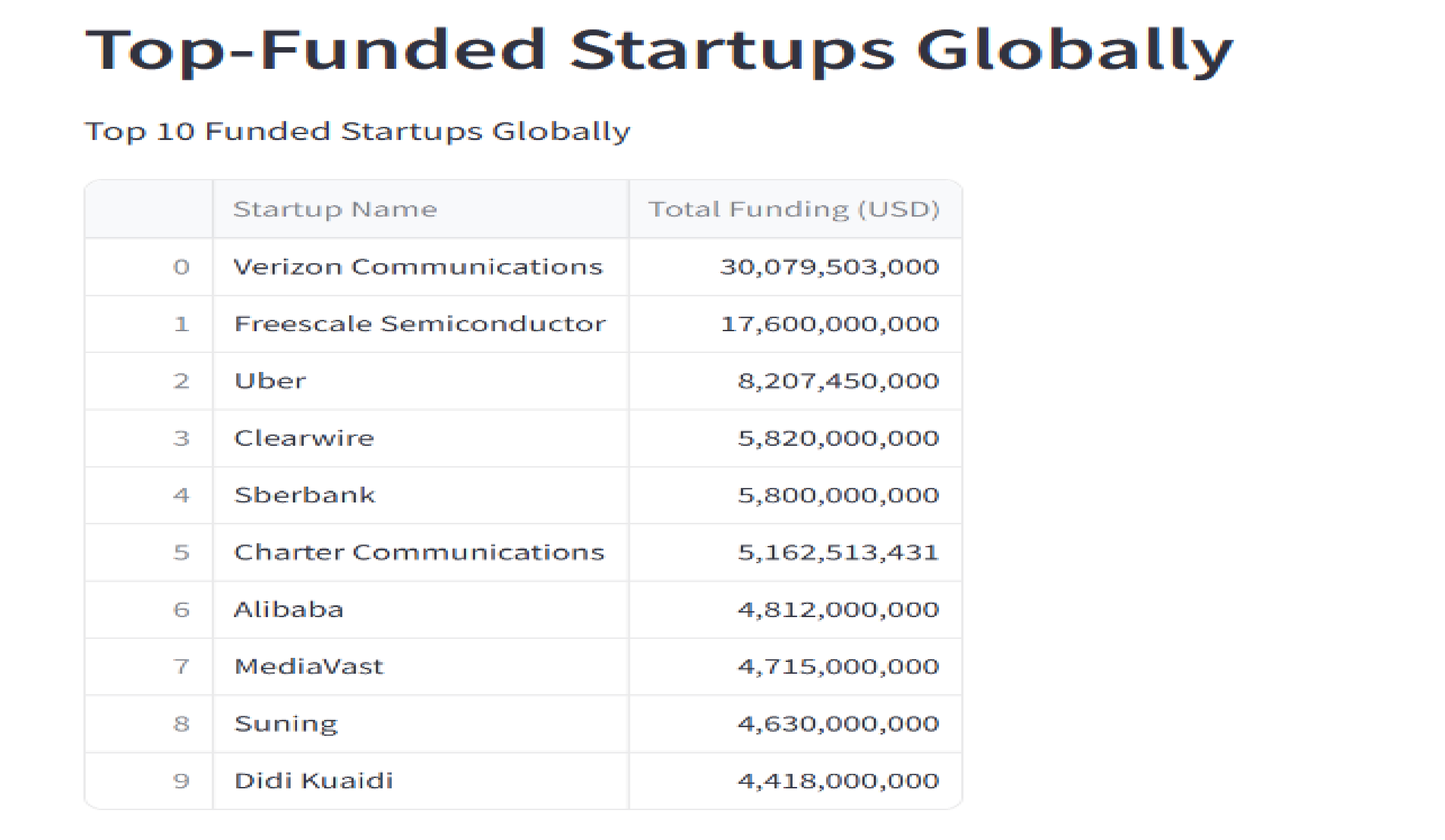
5. **app.py**

- **Purpose:** Streamlit-based interactive application for exploring startup funding trends
- **What we did is:**

1. **Load Data:**
   - Uses DataLoader and DataPreprocessor to load and clean datasets.
   - Caching implemented for performance.

2. **Global Analysis:**

- Displays top 10 funded startups globally using a dataframe .



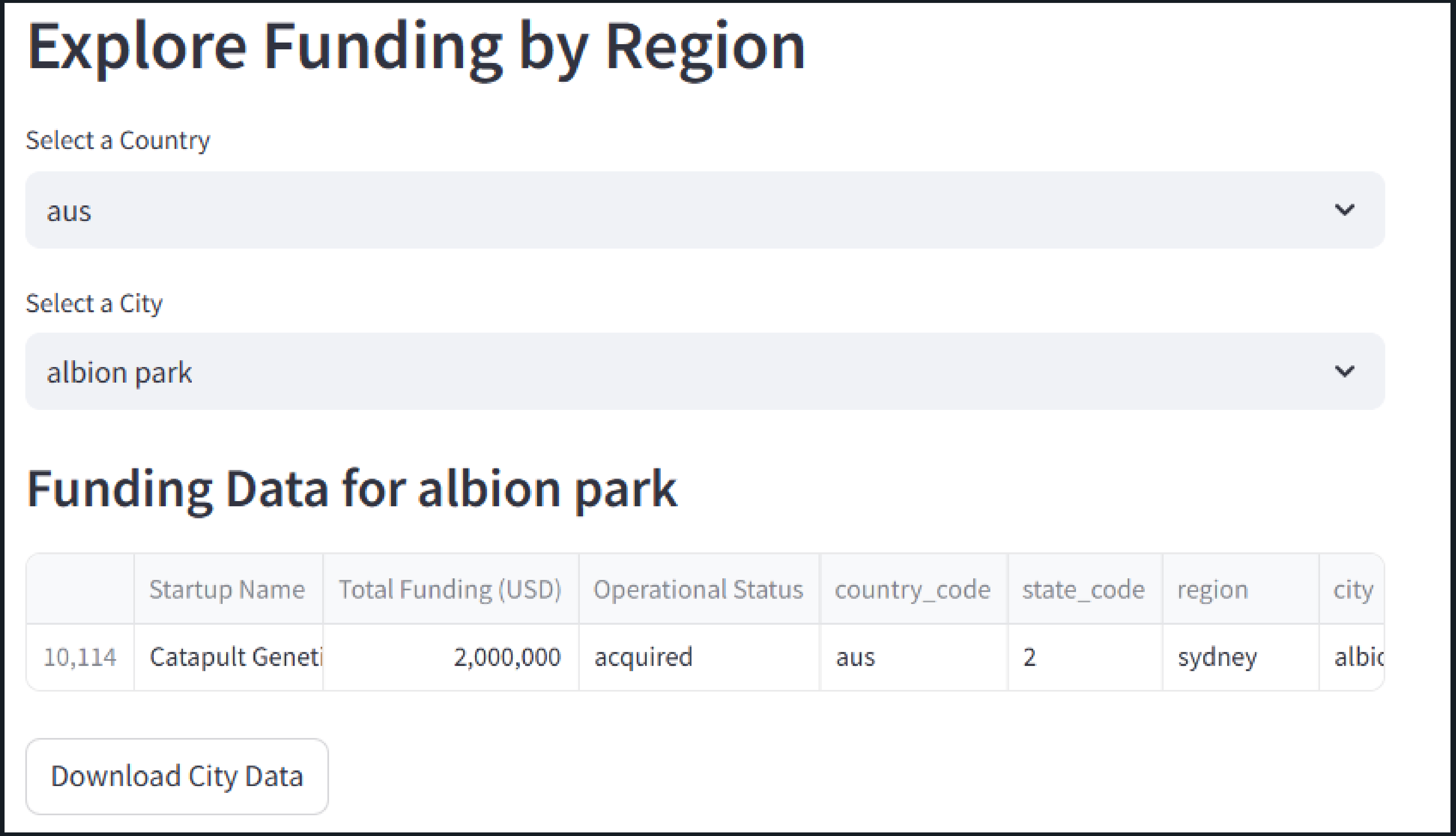
Top-Funded Startups Globally

Top 10 Funded Startups Globally

	Startup Name	Total Funding (USD)
0	Verizon Communications	30,079,503,000
1	Freescale Semiconductor	17,600,000,000
2	Uber	8,207,450,000
3	Clearwire	5,820,000,000
4	Sberbank	5,800,000,000
5	Charter Communications	5,162,513,431
6	Alibaba	4,812,000,000
7	MediaVast	4,715,000,000
8	Suning	4,630,000,000
9	Didi Kuaidi	4,418,000,000

3. **Regional Analysis:**

- Allows user to select a country and city.
- Provides detailed funding insights.
- Offers data download option



Explore Funding by Region

Select a Country

aus

Select a City

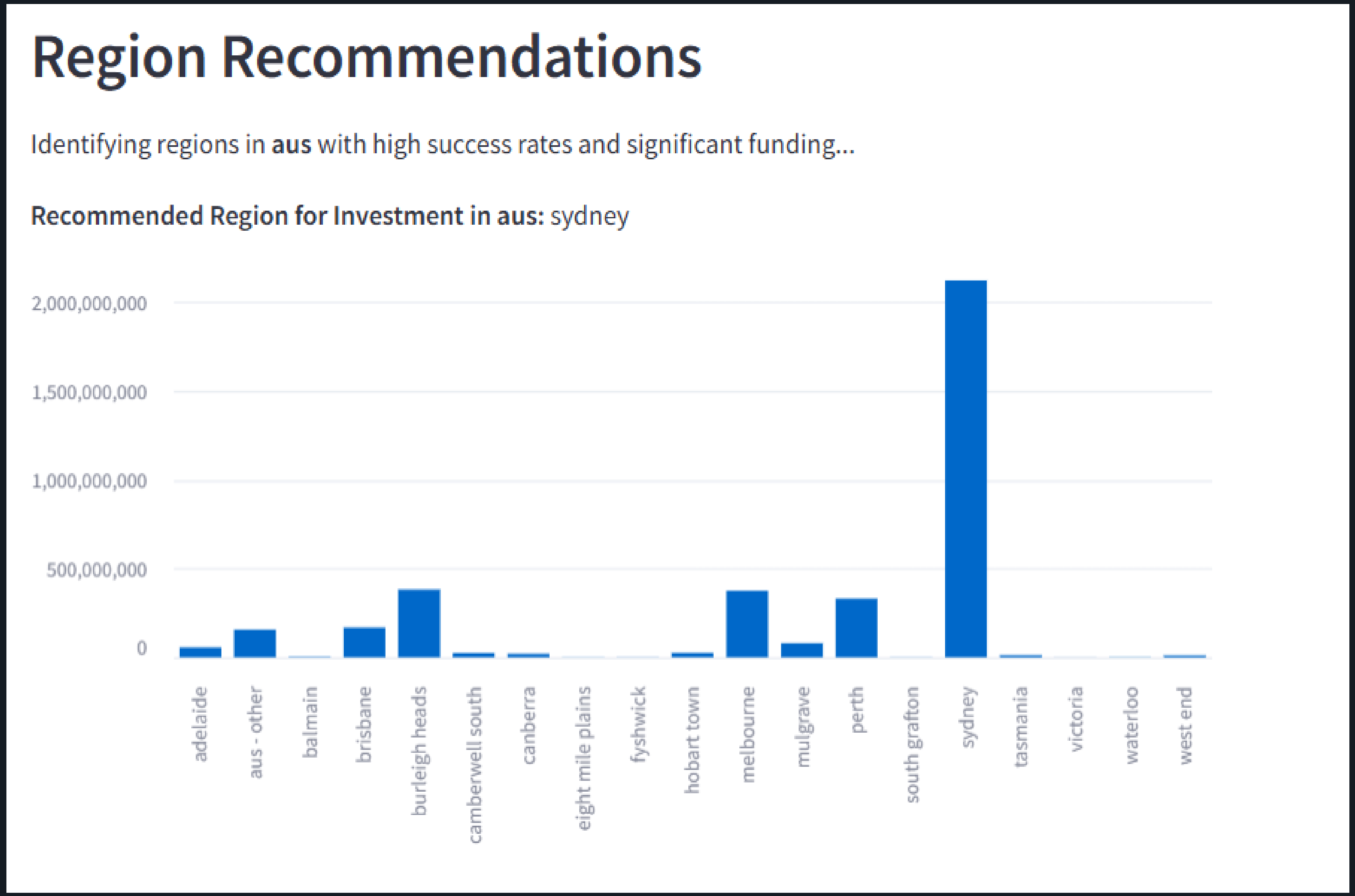
albion park

Funding Data for albion park

	Startup Name	Total Funding (USD)	Operational Status	country_code	state_code	region	city
10,114	Catapult Geneti	2,000,000	acquired	aus	2	sydney	albi

4.Recommendations:

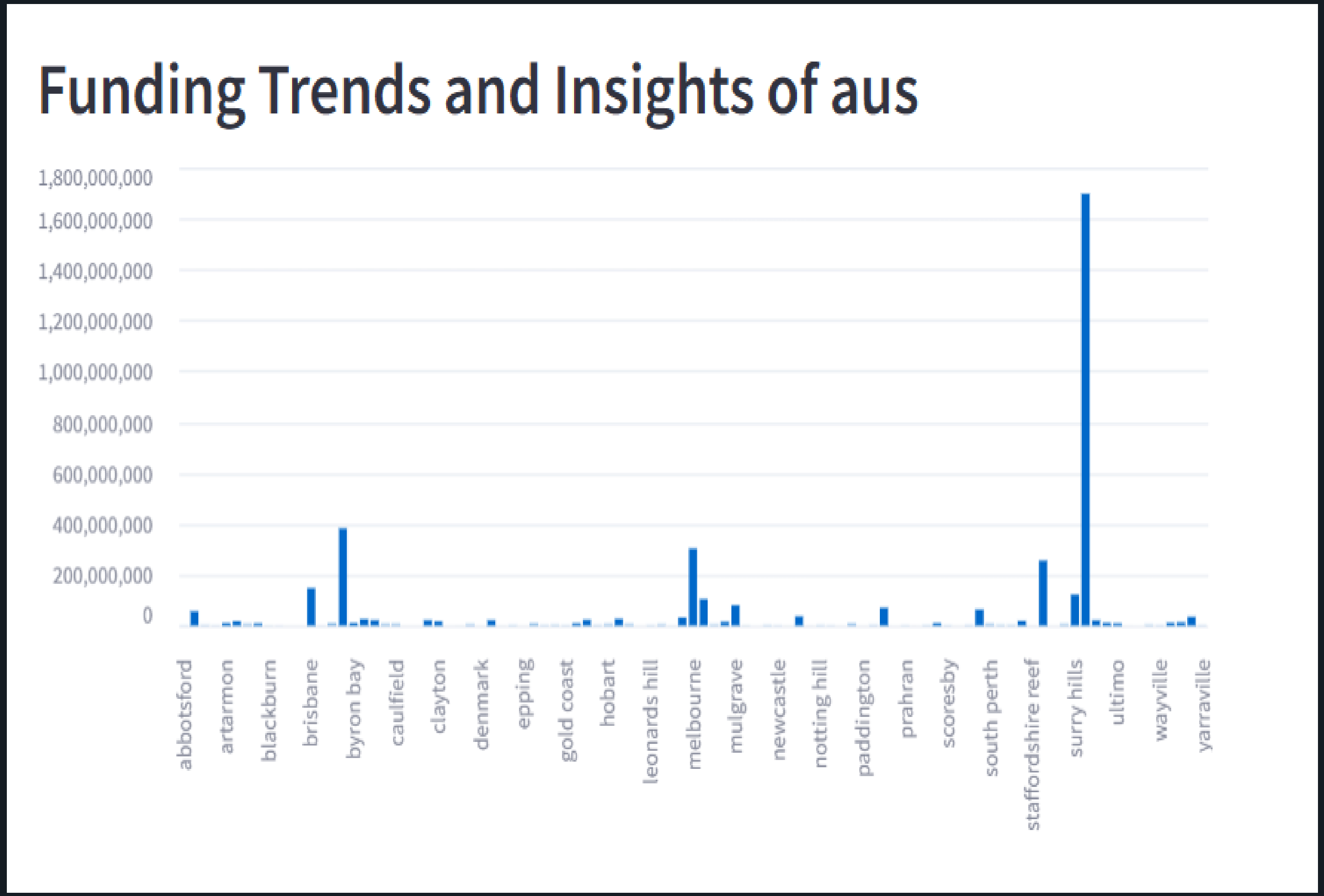
- Highlights top-performing regions in selected countries.
- Visualizes regional funding trends using bar charts.
- Displays funding patterns by city within the selected country.



Region Recommendations

Identifying regions in **aus** with high success rates and significant funding...

Recommended Region for Investment in aus: sydney



Funding Trends and Insights of aus

Conclusion

This project integrates data preprocessing, visualization, and machine learning to deliver actionable insights on startup success. Exploratory Data Analysis (EDA) identifies key trends and influential variables. Machine learning models like Random Forest provide accurate and explainable predictions to guide decision-making.

Streamlit enables a user-friendly interface for customized, interactive exploration.

The modular design ensures scalability and adaptability for future enhancements.