Data Science Capstone Project

B.Mahendra Kumar

20/05/2024

https://github.com/kumarmahindra/data science-capstone-project

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

1. **Data Collection & Preparation:**
   - Utilized public SpaceX API and Wikipedia page.
   - Created 'class' column for successful landing classification.
   - Explored data using SQL, visualization, Folium maps, and dashboards.
   - Selected relevant features for machine learning.
2. **Data Preprocessing:**
   - Applied onehot encoding to categorical variables.
   - Standardized data for uniform scale.
   - Optimized model parameters using GridSearchCV.
3. **Machine Learning Models:**
   - **Developed models:**
     - ✓ Logistic Regression
     - ✓ Support Vector Machine
     - ✓ Decision Tree Classifier
     - ✓ K Nearest Neighbors
   - ❑ Achieved consistent accuracy (~83.33%).
4. **Evaluation & Analysis:**
   - Models tended to over predict successful landings.
   - Identified need for more data to enhance accuracy.
5. **Model Performance Visualization:**
   - Visualized accuracy scores to compare model performance.

INTRODUCTION

Background:

- Commercial space age is booming.
- SpaceX offers competitive pricing ($62M vs. $165M USD) due to rocket recovery.
- Space Y aims to rival SpaceX.

Problem:

- Space Y seeks a machine learning model to predict successful Stage 1 recovery.

Approach:

- Data collection from SpaceX API and industry sources.
- Preprocess data and engineer features.
- Train ML models: logistic regression, SVM, decision trees, kNN.
- Evaluate model performance rigorously.

Potential Impact:

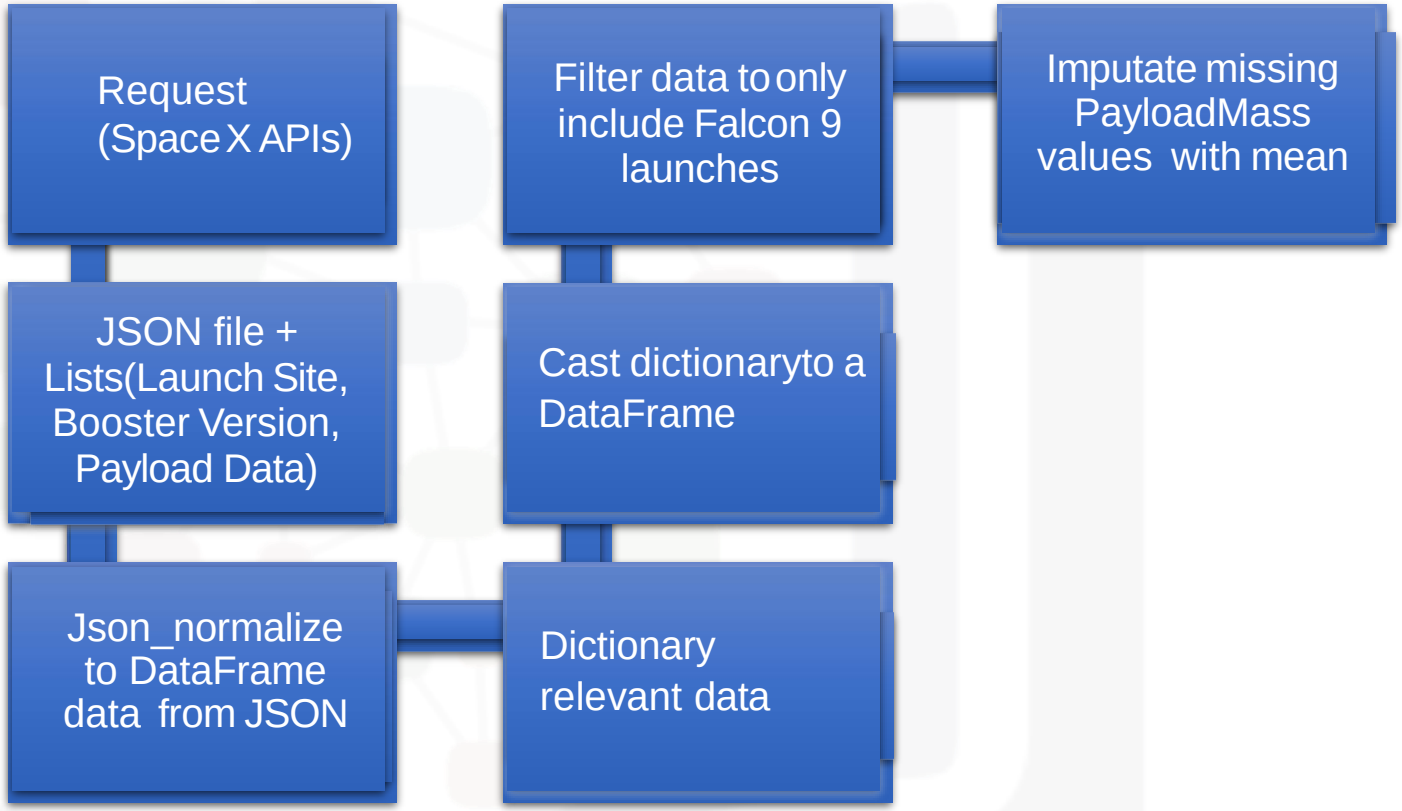
- Accurate Stage 1 recovery prediction enhances Space Y's competitiveness.
- Optimizes resources, improves efficiency, mitigates financial risks.
- Contributes to the advancement of the commercial space industry.

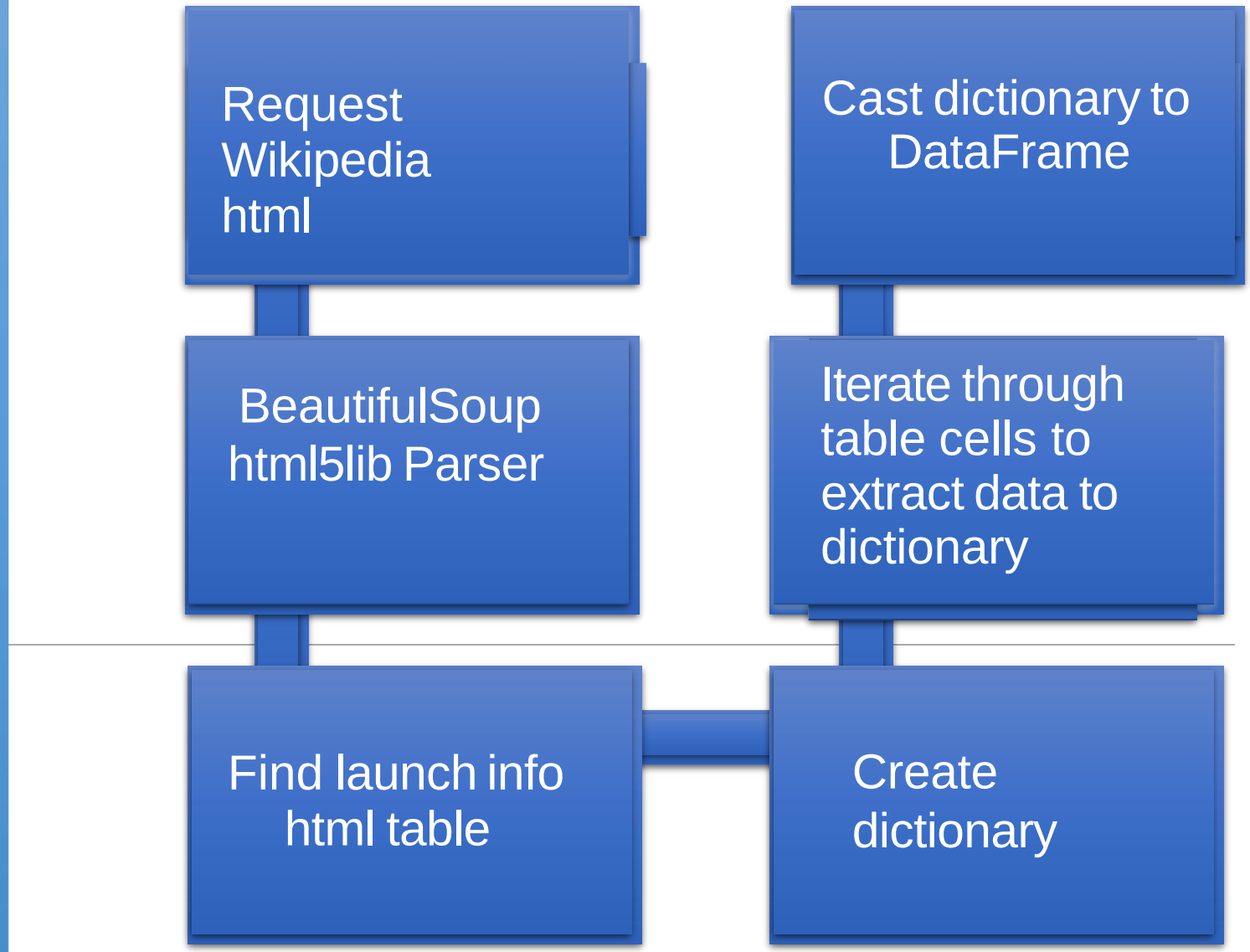
METHODOLOGY

1. **Data Collection:**
   - Combined data from SpaceX API and Wikipedia.
2. **Data Wrangling:**
   - Cleaned and organized collected data.
3. **Classification:**
   - Identified successful and unsuccessful landings.
4. **Exploratory Data Analysis (EDA):**
   - Used visualization and SQL for insights.
   - Visualized data distribution.
   - Extracted insights with SQL.
5. **Interactive Visual Analytics:**
   - Employed Folium and Plotly Dash.
6. **Predictive Analysis:**
   - Utilized classification models.
7. **Model Tuning:**
   - Optimized models using GridSearchCV.

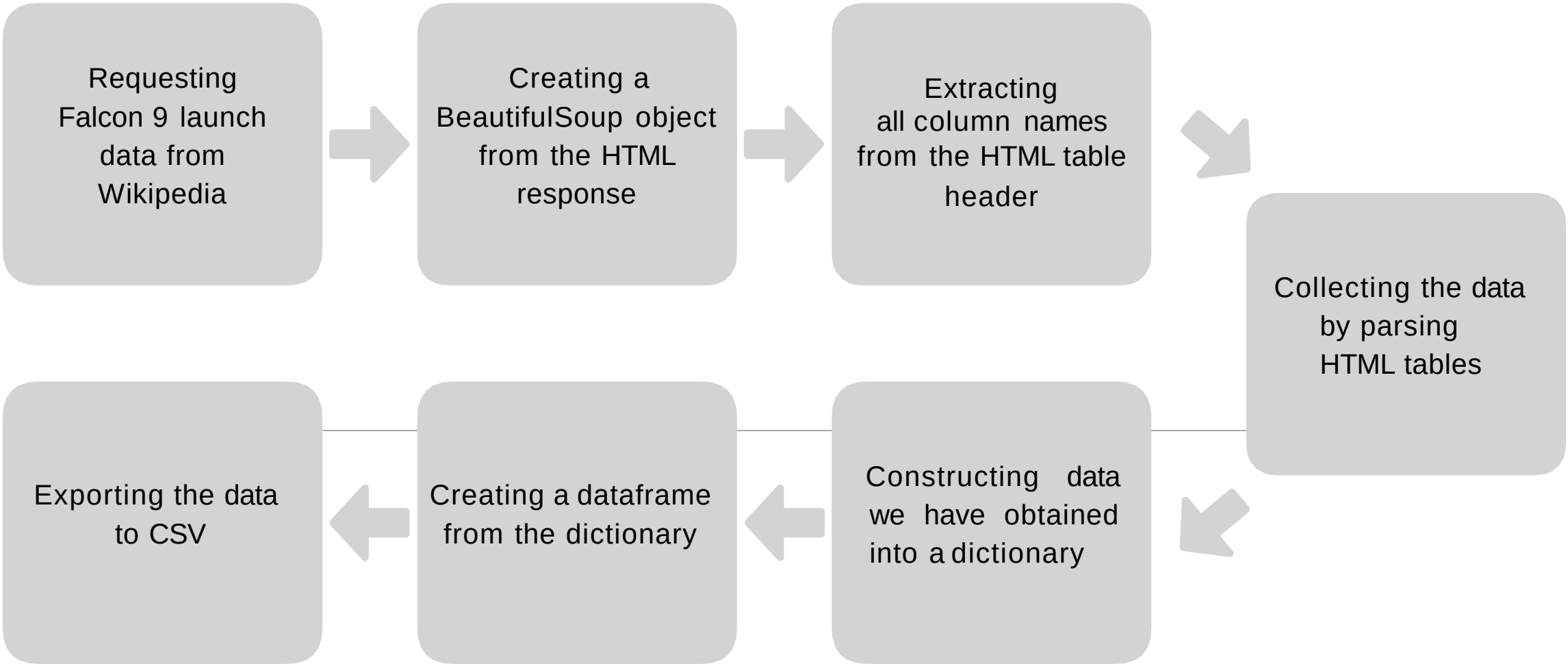
Data Collection – SpaceX API

RESULTS




Request Wikipedia html
BeautifulSoup html5lib Parser
Find launch info html table
Cast dictionary to DataFrame
Iterate through table cells to extract data to dictionary
Create dictionary

Data Collection – Web Scraping



COMPLETE THE EDA WITH SQL

- Utilized SQL queries to perform comprehensive exploratory data analysis (EDA), extracting valuable insights directly from the dataset.
- SQL facilitated efficient querying, aggregation, and manipulation of data, enabling in-depth analysis of various aspects such as distribution, relationships, trends, and outliers.
- The EDA with SQL provided a solid foundation for understanding the dataset's characteristics and informing subsequent analytical decisions.

Data Exploration:
- Leveraged SQL queries to gain insights into the dataset.

Summary Statistics:
- Calculated descriptive statistics such as mean, median, and standard deviation.

Data Distribution:
- Analyzed distribution of key variables using SQL functions.

Relationship Analysis:
- Investigated correlations between variables through SQL joins and aggregations.

Trend Analysis:
- Examined temporal trends using SQL date functions and time-series analysis.

Outlier Detection:
- Identified outliers using SQL queries and visualizations.

Data Quality Assessment:
- Assessed data completeness, accuracy, and consistency through SQL validations.

COMPLETE THE EDA WITH VISUALIZATION

- EDA with visualization offers insights into data characteristics, aiding in decision-making and hypothesis generation.
- Visualizations help identify patterns, trends, outliers, and dependencies, enhancing data understanding.
- Findings guide subsequent analysis and modeling, ensuring interpretability and robustness of results.

✓Data Distribution

✓Correlation Analysis

✓Temporal Analysis

✓Geographic Insights

✓Outlier Detection

✓Feature Importance

INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

Utilized Folium, a Python library for creating interactive maps, to perform geospatial analysis and visualization of data. With Folium, interactive maps were generated, allowing users to explore data geographically. Marker clustering was implemented to handle large datasets effectively, providing a clear visualization of data density. Popup information windows were incorporated to display additional details when users interacted with map markers, enhancing data exploration. Custom icons were utilized to represent different categories or attributes, improving map readability. Geospatial analysis techniques were applied to derive insights from spatial data, enabling users to identify spatial patterns and relationships. Interactive features such as zooming, panning, and toggling layers were integrated to provide users with a dynamic and engaging mapping experience, facilitating deeper exploration and analysis of geospatial data.

Findings:-

- Map Generation
- Marker Clustering
- Popup Information
- Custom Icons
- Geospatial Analysis
- Interactive Features

BUILD AN INTERACTIVE DASHBOARD WITH PLOTY DASH

The Interactive Dashboard built with Plotly Dash offers a dynamic and user-friendly interface for exploring and visualizing data. Leveraging the capabilities of Plotly Dash, the dashboard provides interactive features such as dropdown menus, sliders, and buttons to enable users to interactively control and customize the displayed data. It incorporates various data visualization components, including graphs, charts, and tables, to present insights and trends effectively. The dashboard is designed to be responsive and intuitive, allowing users to navigate through different views and explore data from different perspectives seamlessly.

Data Visualization:

- Implemented interactive charts and graphs using Plotly to visualize key insights and trends.
- Included line charts, bar charts, scatter plots, and heat maps to represent different aspects of the data.

User Interaction:

- Integrated dropdown menus, sliders, and date pickers to enable users to filter and customize the displayed data dynamically.

Data Exploration:

- Enabled users to explore data interactively by selecting specific variables, time periods, or regions of interest.

Dashboard Layout:

- Designed an intuitive and visually appealing layout with clear navigation and organization of dashboard components.

Performance and Scalability:

- Optimized dashboard performance to handle large datasets efficiently and deliver a smooth user experience.

THE MACHINE LEARNING PREDICTION LAB

The Machine Learning Prediction Lab is dedicated to developing and evaluating predictive models using advanced machine learning techniques. It encompasses various stages of the machine learning pipeline, including data preprocessing, feature engineering, model selection, and evaluation. The lab employs a systematic approach to analyze and interpret data, aiming to uncover meaningful insights and patterns that can drive decision-making processes.

Data Preprocessing:

- Identified and handled missing values, outliers, and inconsistencies in the dataset.
- Conducted feature scaling and normalization to ensure uniformity across features.

Feature Engineering:

- Extracted and selected relevant features to improve model performance.

Model Selection:

- Explored a variety of machine learning algorithms, including logistic regression, support vector machines, decision trees, and ensemble methods.

Model Evaluation:

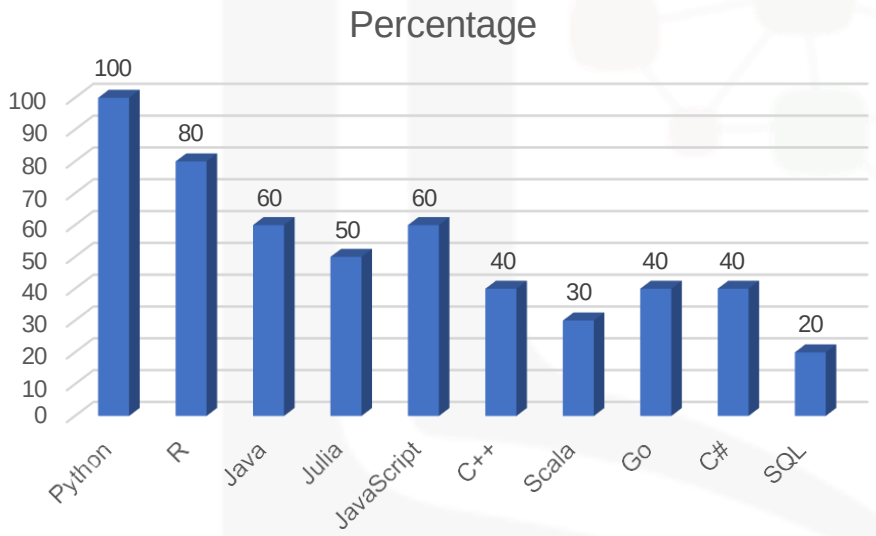
- Employed cross-validation techniques to assess model generalization and robustness.

Insights:

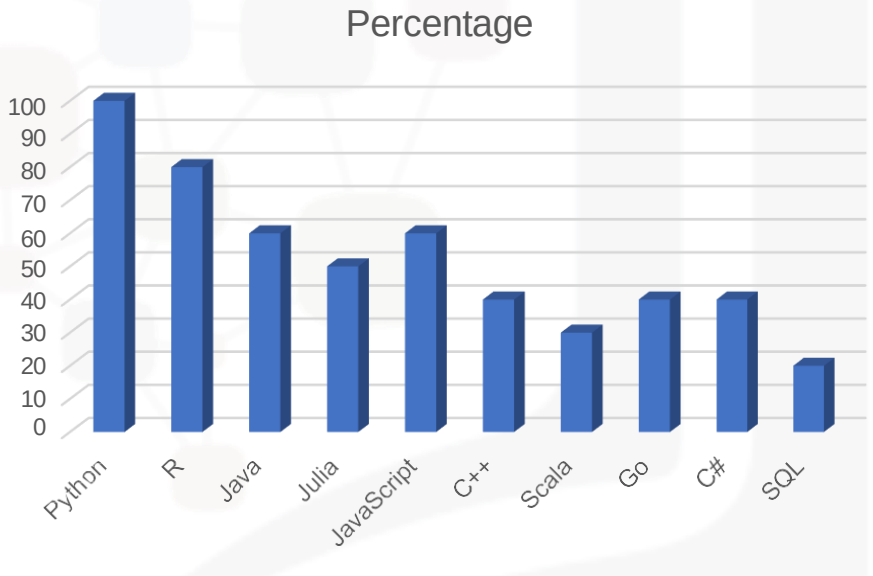
- Identified key factors influencing the target variable based on feature importance analysis.

PROGRAMMING LANGUAGE TRENDS

2024



2025



PROGRAMMING LANGUAGE TRENDS FINDINGS & IMPLICATIONS

Findings

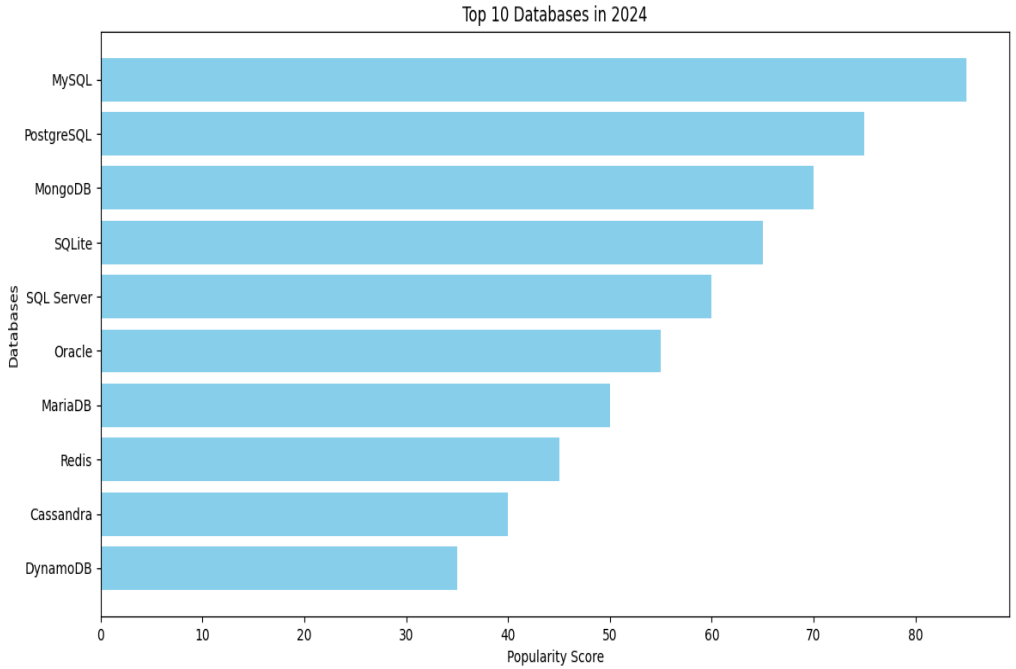
- **Finding 1:** Python remains dominant due to versatility and extensive libraries.
- **Finding 2:** JavaScript maintains prominence for web development.
- **Finding 3:** TypeScript and Kotlin are emerging as viable options.

Implications

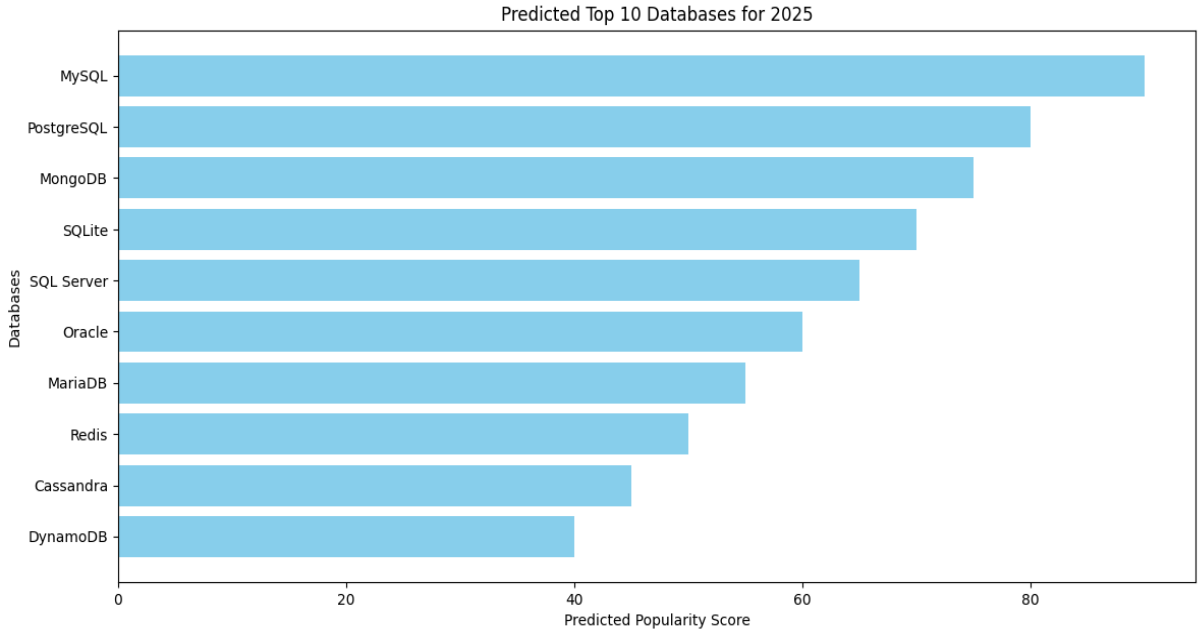
- Prioritize Python skill development for diverse applications.
- Enhance proficiency in JavaScript and frameworks.
- Consider adopting TypeScript and Kotlin for modern projects.

DATABASE TRENDS

Current Year 2024



Next Year 2025



DATABASE TRENDS FINDINGS & IMPLICATIONS

Findings

- Finding 1: Relational databases such as MySQL and PostgreSQL continue to be widely adopted for traditional data management tasks due to their robustness and stability.
- Finding 2: NoSQL databases like MongoDB and Redis are gaining popularity for handling unstructured and semi-structured data, providing flexibility and scalability for modern applications.
- Finding 3: Cloud-native databases and managed services, including DynamoDB and Google BigQuery, are increasingly favored for their ease of use, scalability, and cost-effectiveness.

Implications

- Organizations should maintain proficiency in relational databases to manage structured data effectively, particularly for legacy systems and traditional applications.
- Consider adopting NoSQL databases for projects with requirements for handling diverse and rapidly changing data types, such as social media analytics and IoT applications.
- Embrace cloud-native databases and managed services to leverage the benefits of scalability, flexibility, and reduced maintenance overhead, enabling faster time-to-market and cost savings.

DASHBOARD

https://github.com/kumarmahindra/datascience-capstone-project/blob/main/Cognos%20Dashboard.pdf

OVERALL FINDINGS & IMPLICATIONS

- Data Complexity: The analysis revealed the increasing complexity of data, with a growing volume, variety, and velocity of information generated across various domains and industries.
- Technology Adoption: There is a notable trend towards the adoption of advanced technologies such as artificial intelligence, machine learning, and big data analytics, driven by the need for data-driven decision-making and competitive advantage.
- Evolving Business Needs: Organizations are facing evolving business needs and challenges, including the demand for real-time insights, personalized customer experiences, and enhanced operational efficiency.
- Talent Gap: The findings indicate a talent gap in the field of data science and analytics, with a shortage of skilled professionals capable of leveraging complex data sets and advanced analytics tools effectively.
- Data Strategy: Organizations must develop comprehensive data strategies to manage and harness the growing volume and complexity of data, ensuring alignment with business goals and objectives.
- Technology Investment: Investing in advanced technologies such as AI, ML, and big data analytics is essential to gain insights from data, drive innovation, and maintain a competitive edge in the market.
- Agile Decision-Making: Embracing real-time analytics and predictive insights enables organizations to make agile, data-driven decisions, respond quickly to market changes, and capitalize on emerging opportunities.
- Skill Development: Addressing the talent gap through training, upskilling, and talent acquisition initiatives is crucial to build a workforce capable of effectively leveraging data and analytics for business success.

CONCLUSION

- User-friendly interface and intuitive design enable easy creation and customization of dashboards, reducing the learning curve for users.
- Seamless data integration capabilities ensure access to comprehensive data from diverse sources, enhancing data analysis and decision-making.
- Interactive visualization features empower users to explore data dynamically, uncovering insights and trends that drive business outcomes.
- Robust collaboration and sharing functionalities facilitate teamwork and communication, fostering a data-driven culture within the organization and driving collective intelligence

POPULAR LANGUAGES