IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Visualization –Charts

•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. 2. Data Preprocessing: Applied onehot encoding to categorical variables. Standardized data for uniform scale. Optimized model parameters using GridSearchCV.
3. 3. Machine Learning Models: Developed models: Logistic Regression Support Vector Machine Decision Tree Classifier K Nearest Neighbors Achieved consistent accuracy (~83.33%).
4. 4. Evaluation & Analysis: Models tended to over predict successful landings. Identified need for more data to enhance accuracy.
5. 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: Stage 1 recovery. Approach:

⍰ Space Y seeks a machine learning model to predict successful

⍰ Data collection from SpaceX API and industry sources.

⍰ Preprocess data and engineer features.

⍰ Train ML models: logistic regression, SVM, decision trees

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.

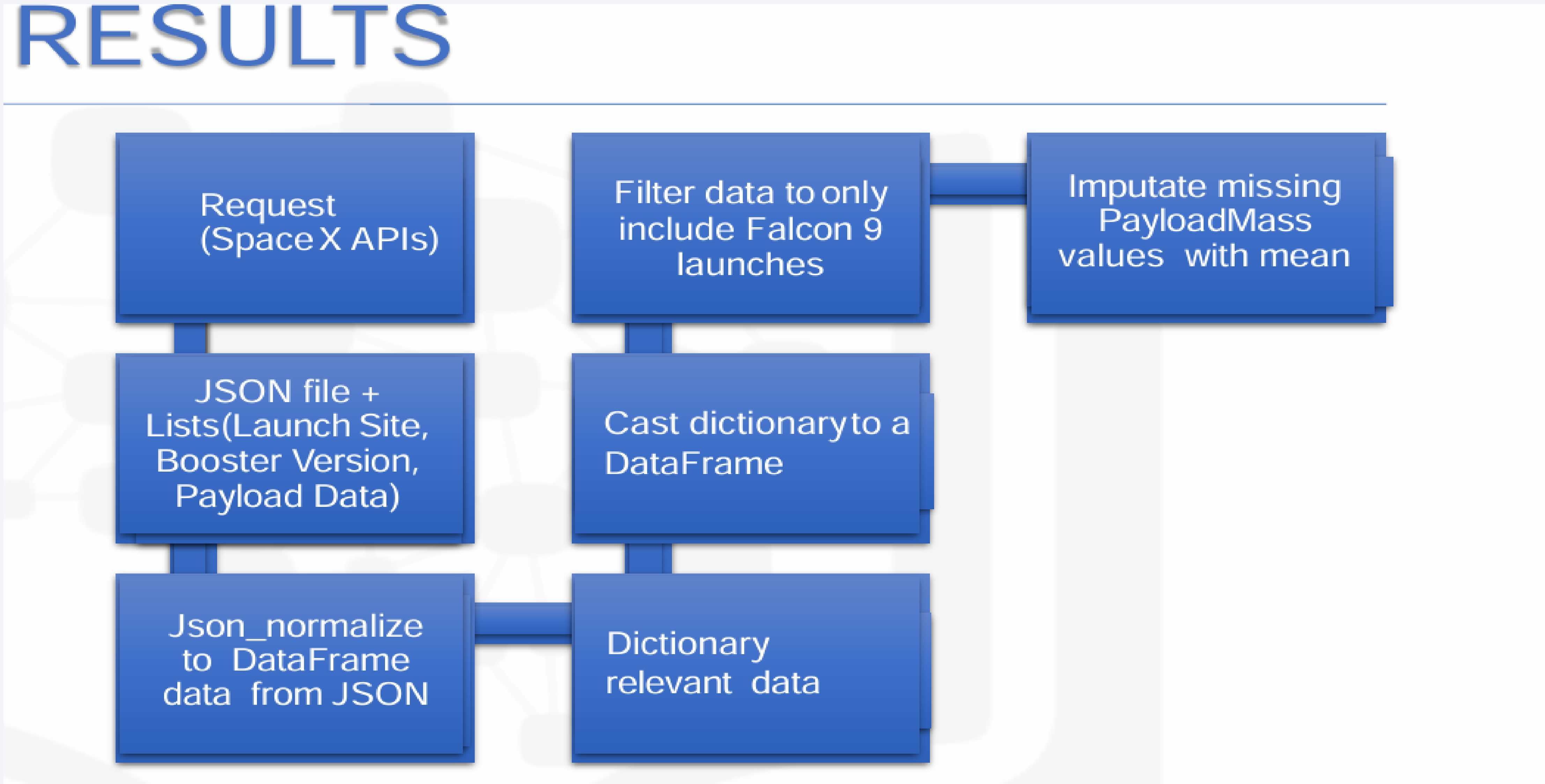
⍰ Visualized data distribution.

⍰ Extracted insights with SQL.

5. Interactive Visual Analytics:

⍰ Employed Folium and Plotly Dash.

Data Collection



RESULTS

Request (SpaceX APIs)

JSON file + Lists(Launch Site, Booster Version, Payload Data)

Json_normalize to DataFrame data from JSON

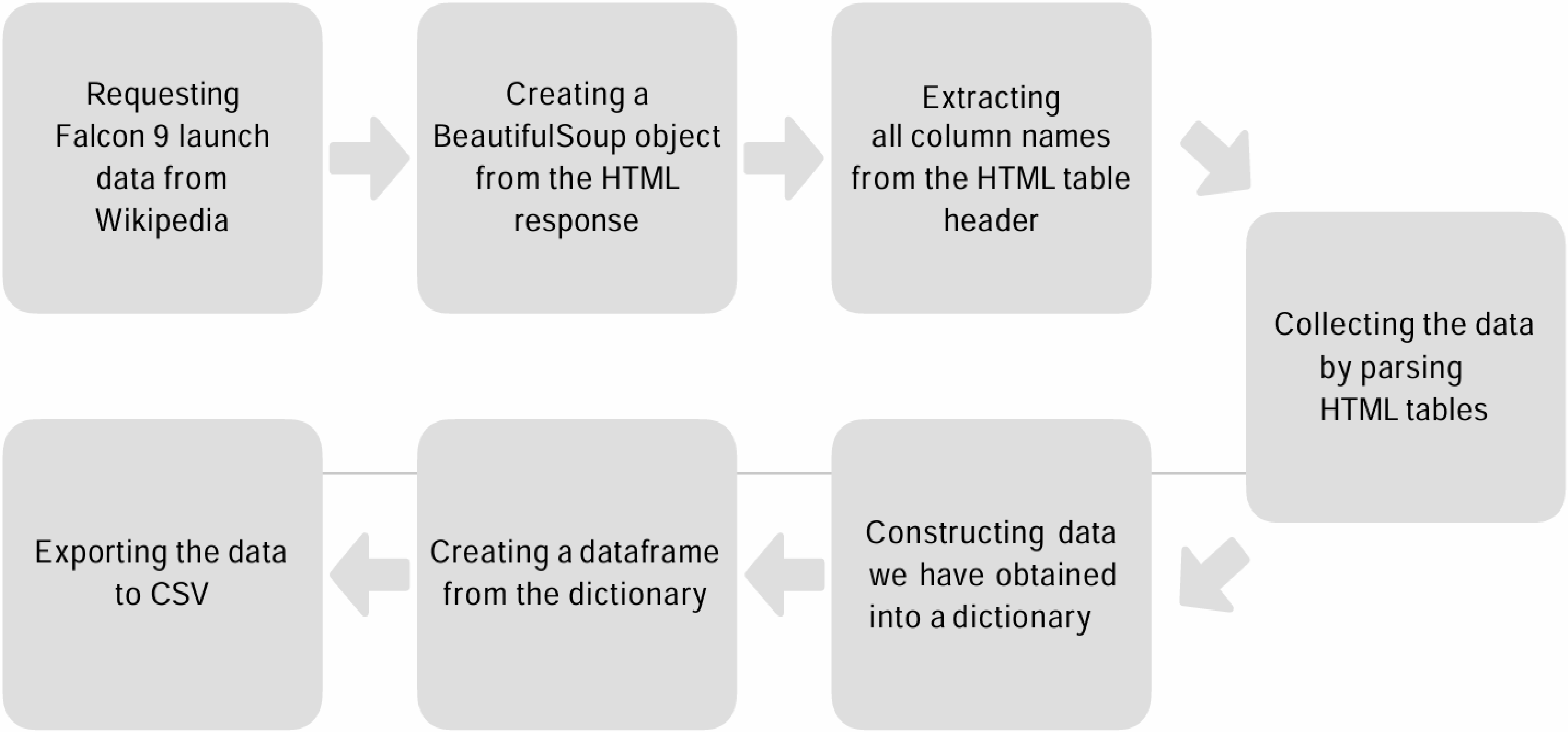
Dictionary relevant data

Cast dictionary to a DataFrame

Filter data to only include Falcon 9 launches

Imputate missing PayloadMass values with mean

Data Collection – SpaceX API



Requesting Falcon 9 launch data from Wikipedia

Creating a BeautifulSoup object from the HTML response

Extracting all column names from the HTML table header

Collecting the data by parsing HTML tables

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

Exporting the data to CSV

EDA with Data 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, interpretability and robustness of results.

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 EDAwith SQL provided a solid foundation for understanding the dataset's characteristics and informing subsequent analytical decisions.

Build an Interactive Map with Folium

- Utilized Folium, a Python library for creating interactive maps, to perform geospatial analysis and visualization of data. Popup information windows were incorporated to display additional details when users interacted with map markers, enhancing data exploration. Interactive features such as zooming, panning, and toggling layers were integrated to provide users with a dynamic and
- GitHub Findings:
- Map Generation
- Marker Clustering
- Popup Information

Build a Dashboard with Plotly Dash

- The Interactive Dashboard built with Plotly Dash offers a dynamic and user-friendly interface for exploring and visualizing data.
- Data Visualization:
- Implemented interactive charts and graphs using Plotly to visualize key insights and trends. User Interaction:
- Included line charts, bar charts, scatter plots, and heat maps to represent different aspects of the data.
- Integrated dropdown menus, sliders, and date pickers to enable users to filter and customize the displayed data dynamically.

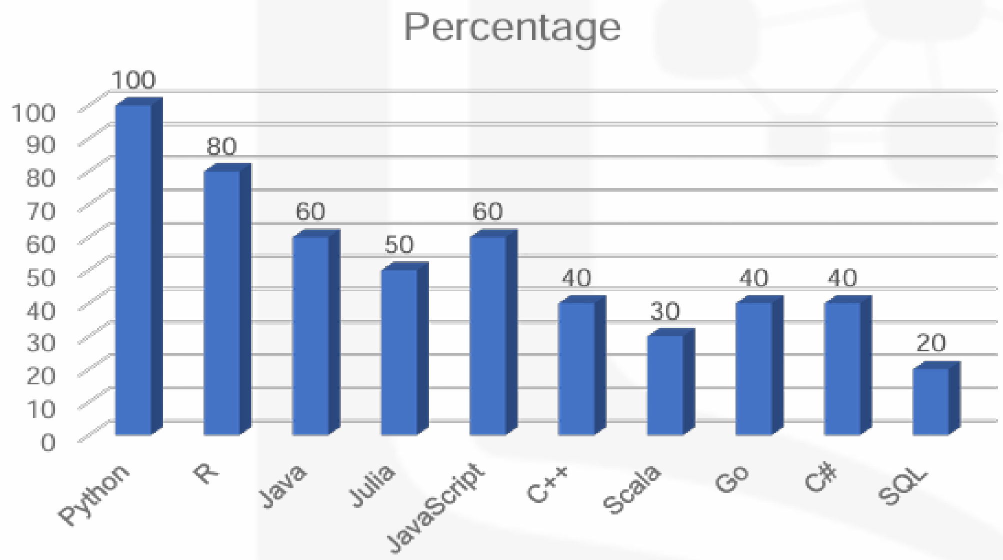
Predictive Analysis (Classification)

- The Machine Learning Prediction Lab is dedicated to developing and evaluating predictive models using advanced machine learning techniques.
- Model Evaluation:
- Employed cross-validation techniques to assess model generalization and robustness.
- Identified key factors influencing the target variable based on feature importance analysis.

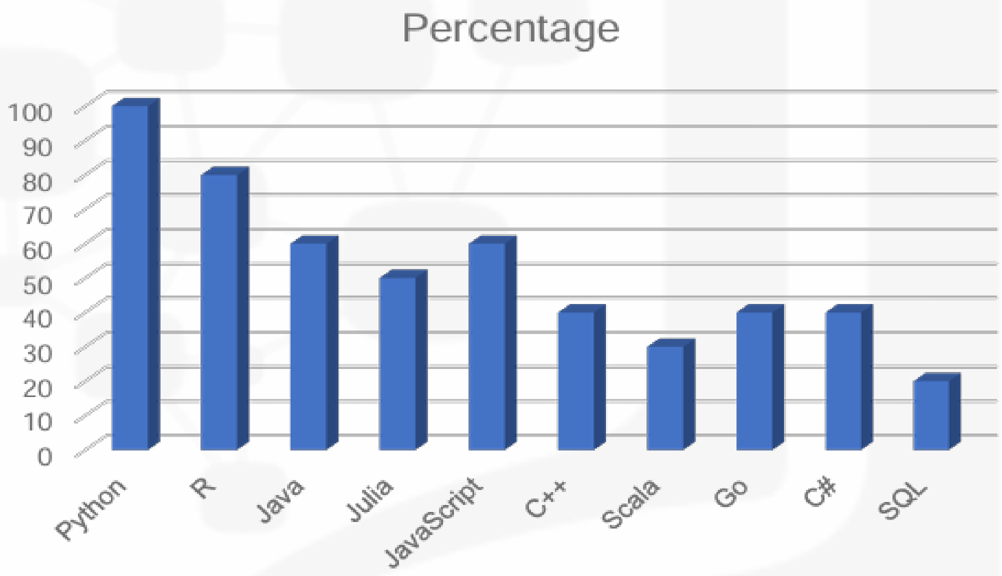
Results

PROGRAMMING LANGUAGE TRENDS

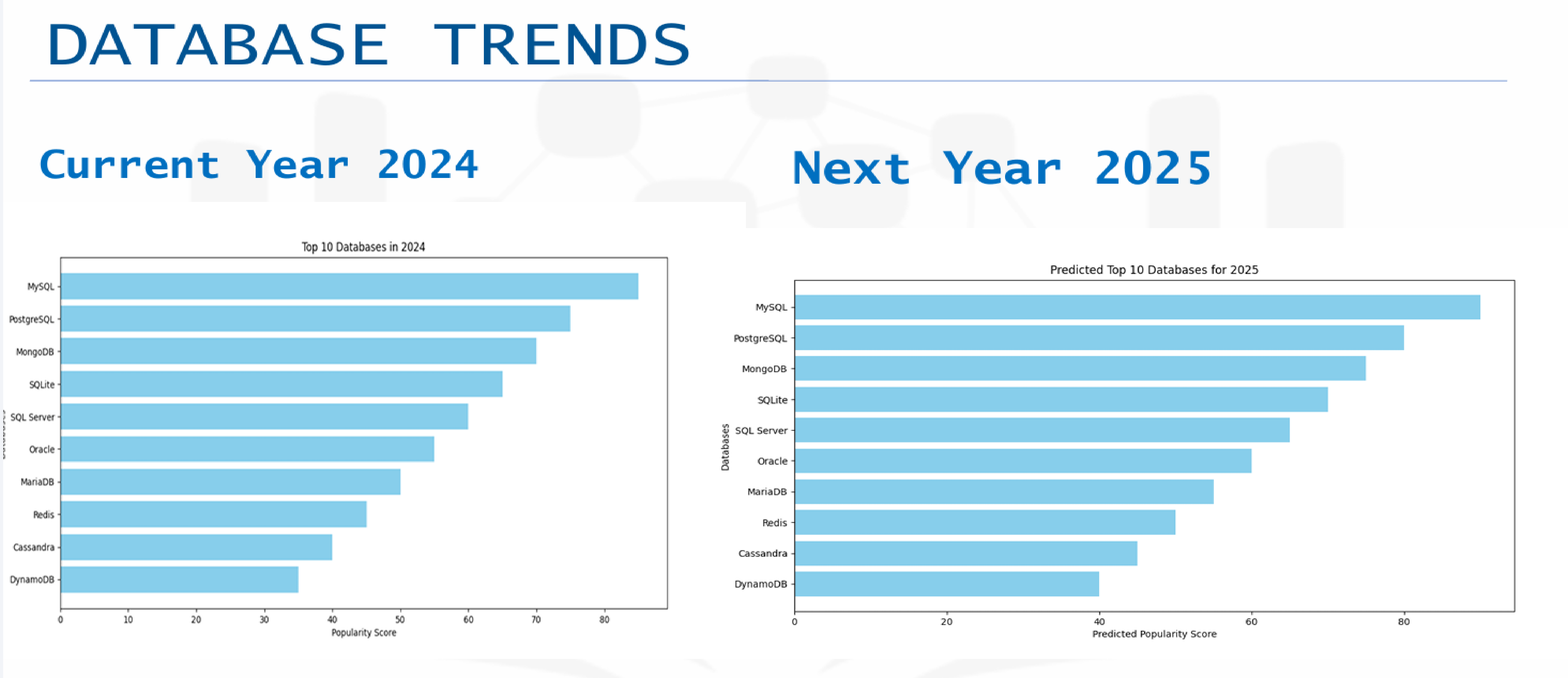
2024



2025



Flight Number vs. Launch Site



DATABASE TRENDS

Current Year 2024

Next Year 2025

All Launch Site Names

- Findings
- Finding1:Relational databasessuchas MySQLandPostgreSQLcontinue tobe widely adopted for traditional data management tasks due to their robustnessandstability.
- Finding 2: NoSQL databases like MongoDB and Redis are gaining popularity for handling unstructured and semi-structured data, 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.
- 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..

Conclusions

- 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,enhancingdataanalysisanddecision-making.
- Interactivevisualizationfeaturesempowerusers toexploredatadynamically,uncoveringinsights andtrendsthatdrivebusinessoutcomes.
- Robust collaborationandsharing functionalities facilitate teamwork and communication, fostering a data-driven culture within the organizationanddrivingcollectiveintelligenc

Thank you!