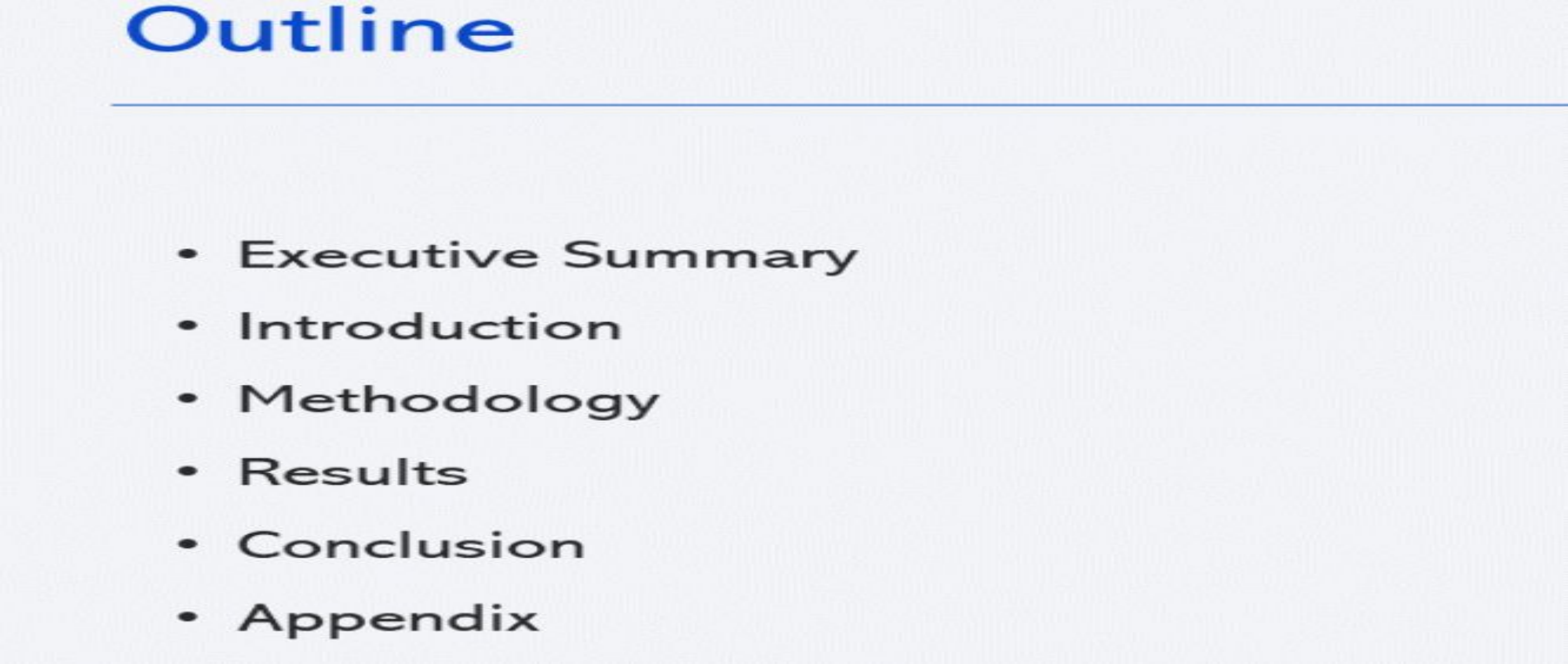
Q1
Q2
Q3
Q2
Q3
Q4

DATA SCIENCE CAPSTONE FINAL PROJECT

Octavia Wellsi 2024/01/24

IBM Developer

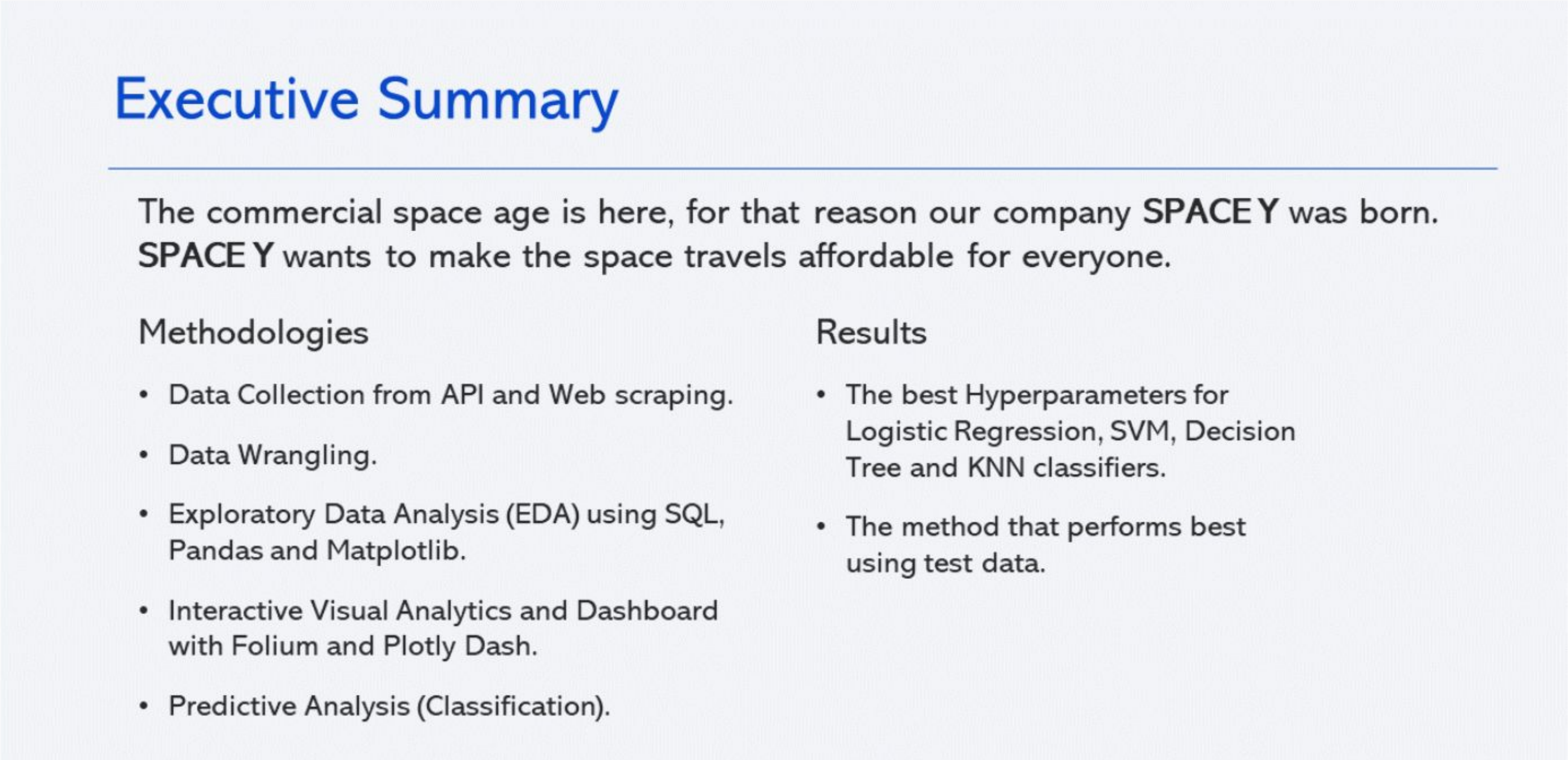
SKILLS NETWORK



Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

EXECUTIVE SUMMARY



Executive Summary

The commercial space age is here, for that reason our company **SPACE Y** was born. **SPACE Y** wants to make the space travels affordable for everyone.

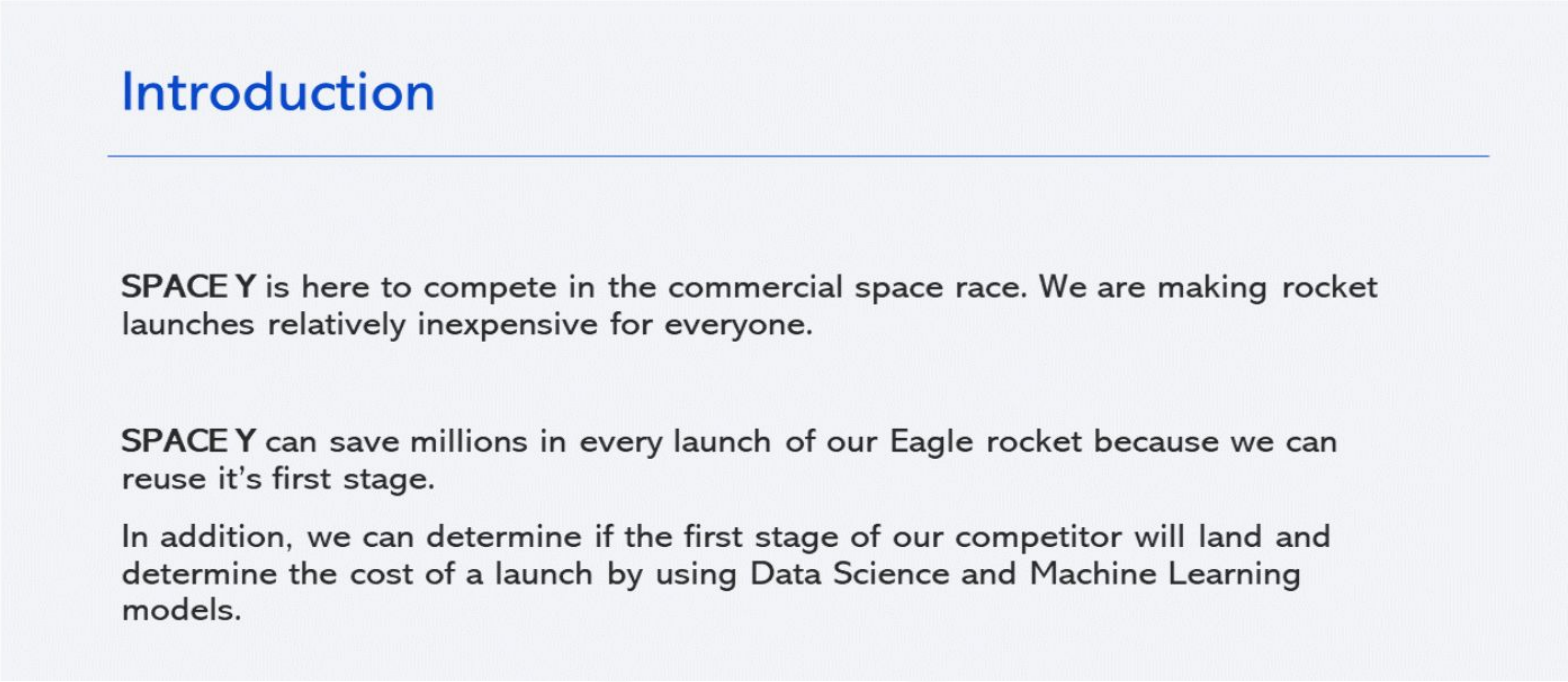
Methodologies

- Data Collection from API and Web scraping.
- Data Wrangling.
- Exploratory Data Analysis (EDA) using SQL, Pandas and Matplotlib.
- Interactive Visual Analytics and Dashboard with Folium and Plotly Dash.
- Predictive Analysis (Classification).

Results

- The best Hyperparameters for Logistic Regression, SVM, Decision Tree and KNN classifiers.
- The method that performs best using test data.

INTRODUCTION



Introduction

SPACE Y is here to compete in the commercial space race. We are making rocket launches relatively inexpensive for everyone.

SPACE Y can save millions in every launch of our Eagle rocket because we can reuse it's first stage.

In addition, we can determine if the first stage of our competitor will land and determine the cost of a launch by using Data Science and Machine Learning models.

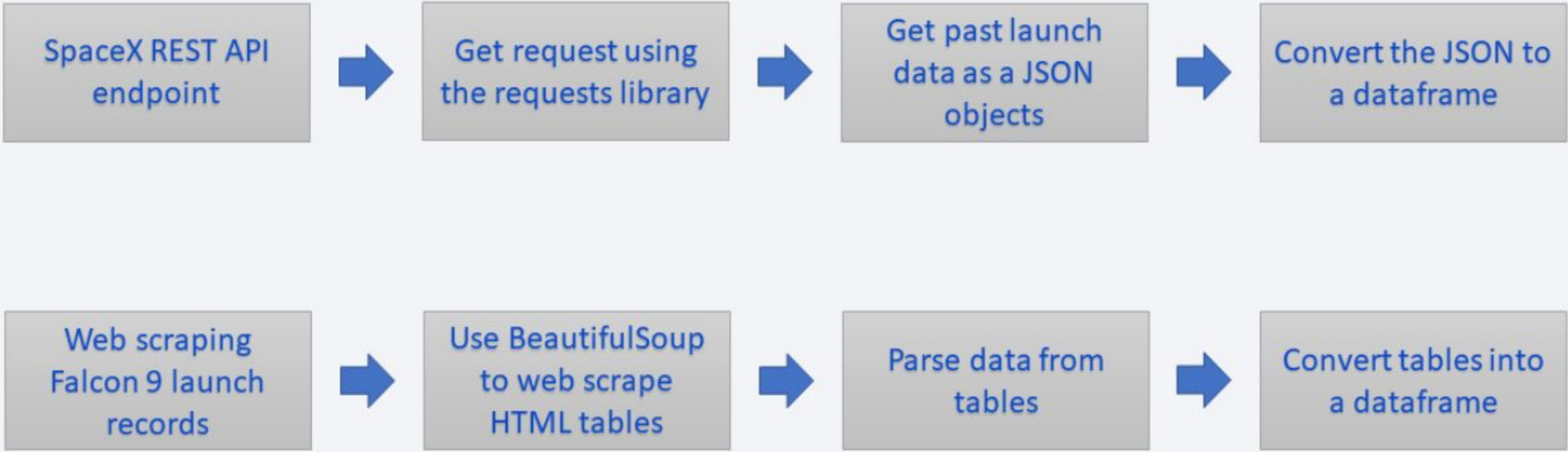
Methodology

Executive Summary

- Data collection methodology:
  - The data was gathered from the SpaceX REST API and web scraping from wiki pages.
- Perform data wrangling
  - The data collected is in form of a JSON object and HTML tables, after that the data is converted into a Pandas dataframe for visualization and analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use of machine learning to determine if the first stage of Falcon 9 will land successfully.

Data Collection

The data was gathered from the SpaceX REST API and web scraped from wiki pages

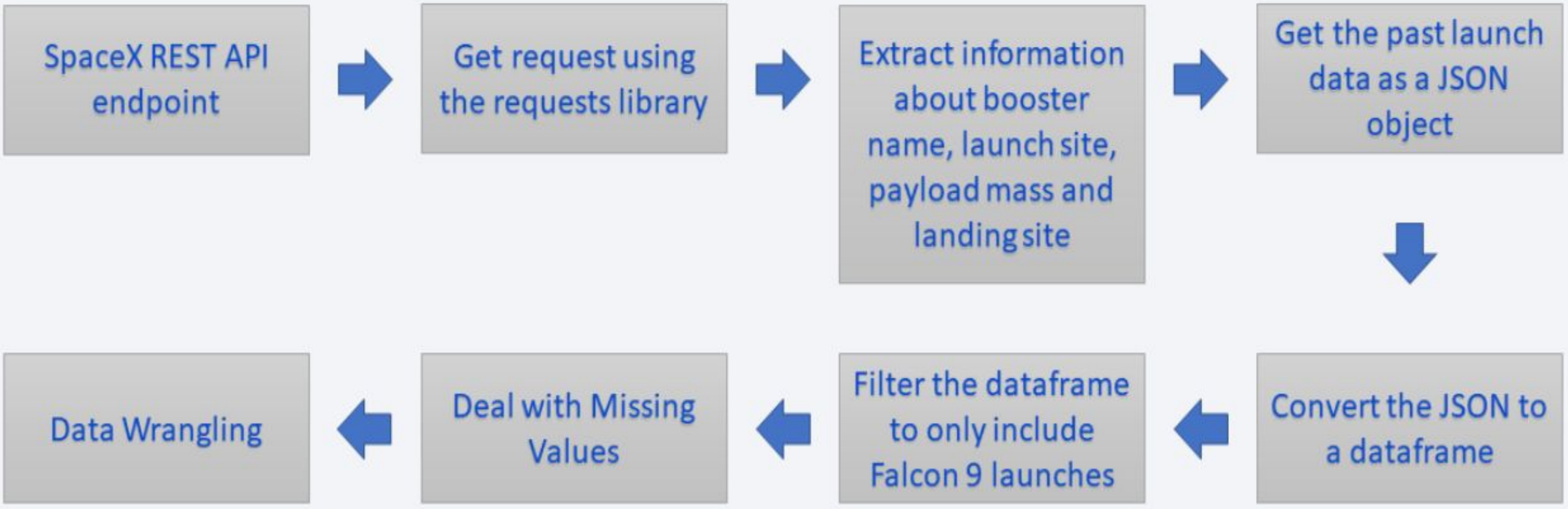


SpaceX REST API endpoint → Get request using the requests library → Get past launch data as a JSON objects → Convert the JSON to a dataframe

Web scraping Falcon 9 launch records → Use BeautifulSoup to web scrape HTML tables → Parse data from tables → Convert tables into a dataframe

Data Collection – SpaceX API

Collect and make sure the data is in the correct format from an API



SpaceX REST API endpoint

Get request using the requests library

Extract information about booster name, launch site, payload mass and landing site

Get the past launch data as a JSON object

Convert the JSON to a dataframe

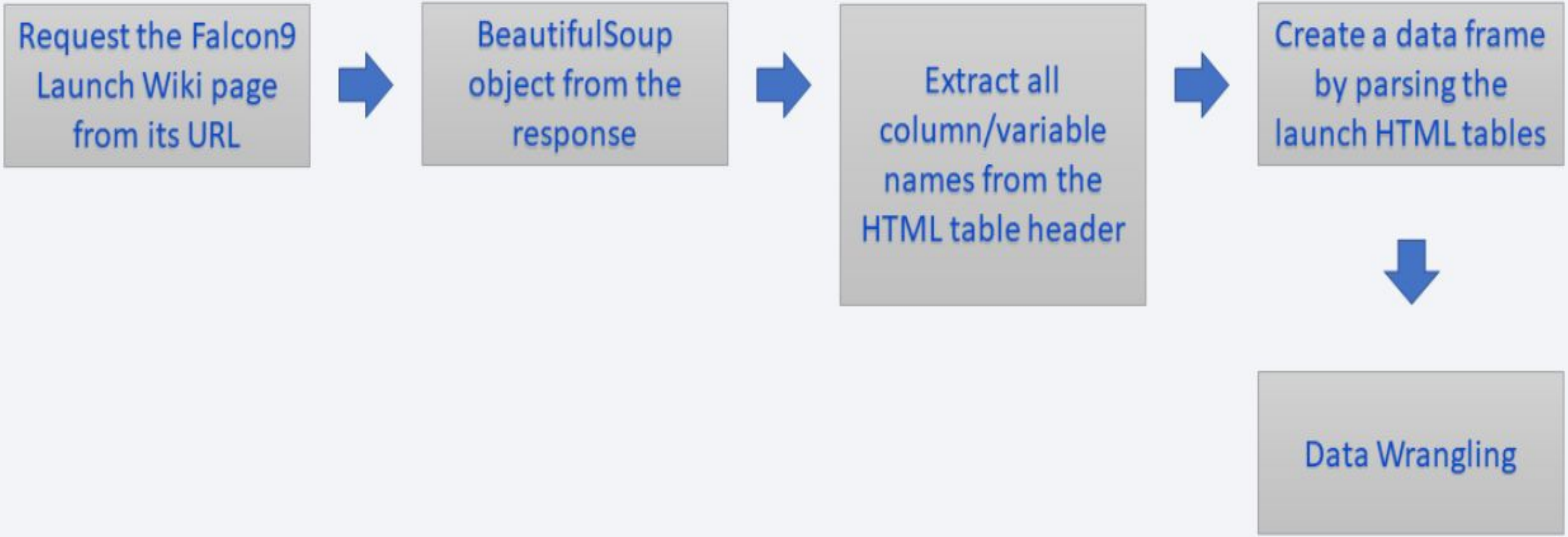
Filter the dataframe to only include Falcon 9 launches

Deal with Missing Values

Data Wrangling

Data Collection - Scraping

Perform web scraping to collect Falcon 9 historical launch records from Wikipedia page



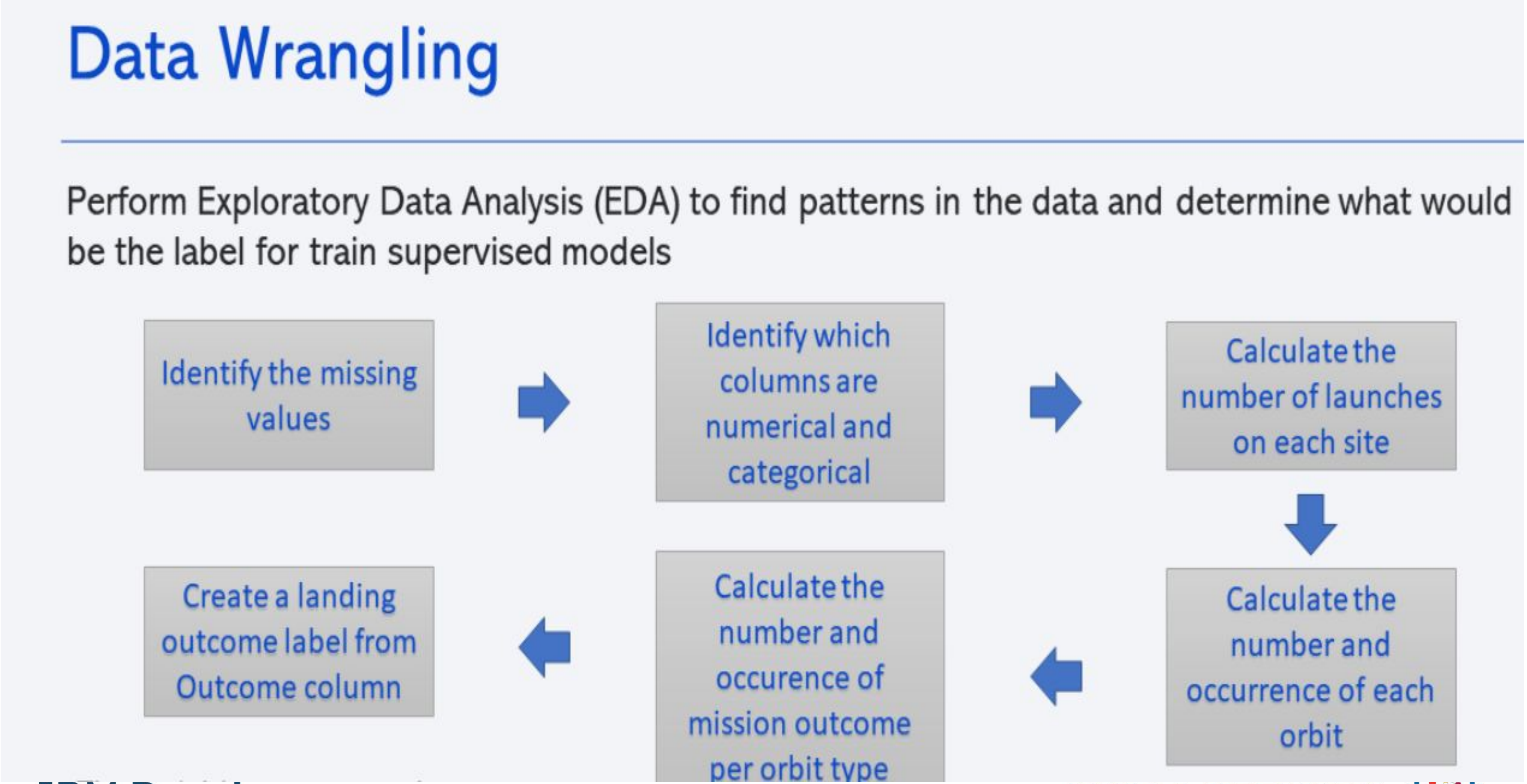
Request the Falcon9 Launch Wiki page from its URL

BeautifulSoup object from the response

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling



Data Wrangling

Perform Exploratory Data Analysis (EDA) to find patterns in the data and determine what would be the label for train supervised models

Identify the missing values

Identify which columns are numerical and categorical

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome per orbit type

Create a landing outcome label from Outcome column

EDA with Data Visualization

Summary of charts that were plotted:

- Catplot to visualize the relationship between Flight Number and Payload.
- Catplot to visualize the relationship between Flight Number and Launch Site.
- Catplot to visualize the relationship between Payload and Launch Site.
- Bar chart to visualize the relationship between success rate of each Orbit type.
- Catplot to visualize the relationship between Flight Number and Orbit type.
- Catplot to visualize the relationship between Payload and Orbit type.
- Line chart to visualize the launch success yearly trend.

EDA with SQL

SQL queries performed:

- Display the names of the unique launch sites in the space mission:
  ```
  SELECT DISTINCT(launch_site) FROM SPACEXTBL;
  ```
- Display 5 records where launch sites begin with the string 'CCA’:
  ```
  SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;
  ```
- Display the total payload mass carried by boosters launched by NASA (CRS):
  ```
  SELECT SUM(payload_mass__kg_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE
  customer='NASA (CRS)’;
  ```
- Display average payload mass carried by booster version F9 v1.1:
  ```
  SELECT AVG(payload_mass__kg_) AS AVG_PAYLOAD_MASS FROM SPACEXTBL WHERE
  booster_version='F9 v1.1’;
  ```
- List the date when the first successful landing outcome in ground pad was achieved:
  ```
  SELECT MIN(DATE) AS first_successful_landing FROM SPACEXTBL WHERE (landing_outcome)='Success
  (ground pad)’;
  ```

Build an Interactive Map with Folium

Summary of map objects that were created and added to the Folium map

- *folium.Circle* and *folium.Marker* to add a highlighted circle area with a text label on a specifi coordinate for each launch site on the site map.
- *MarkerCluster* object for simplify a map containing many markers having the same coordinate.
- *MousePosition* on the map to get coordinate for a mouse over a point on the map.
- *folium.PolyLine* object to draw a line between a launch site to its closest city, railway and highway.

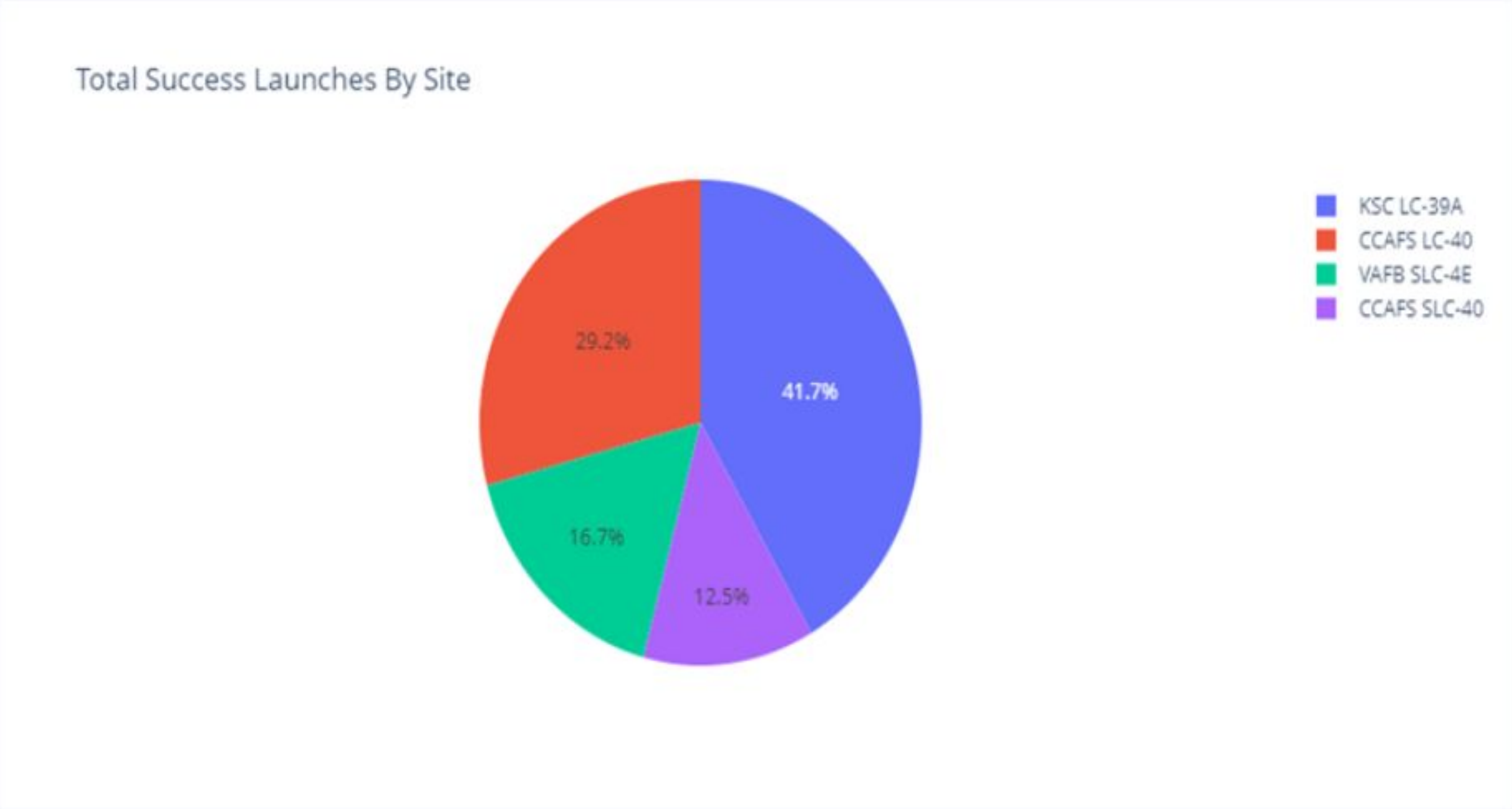
Build a Dashboard with Plotly Dash

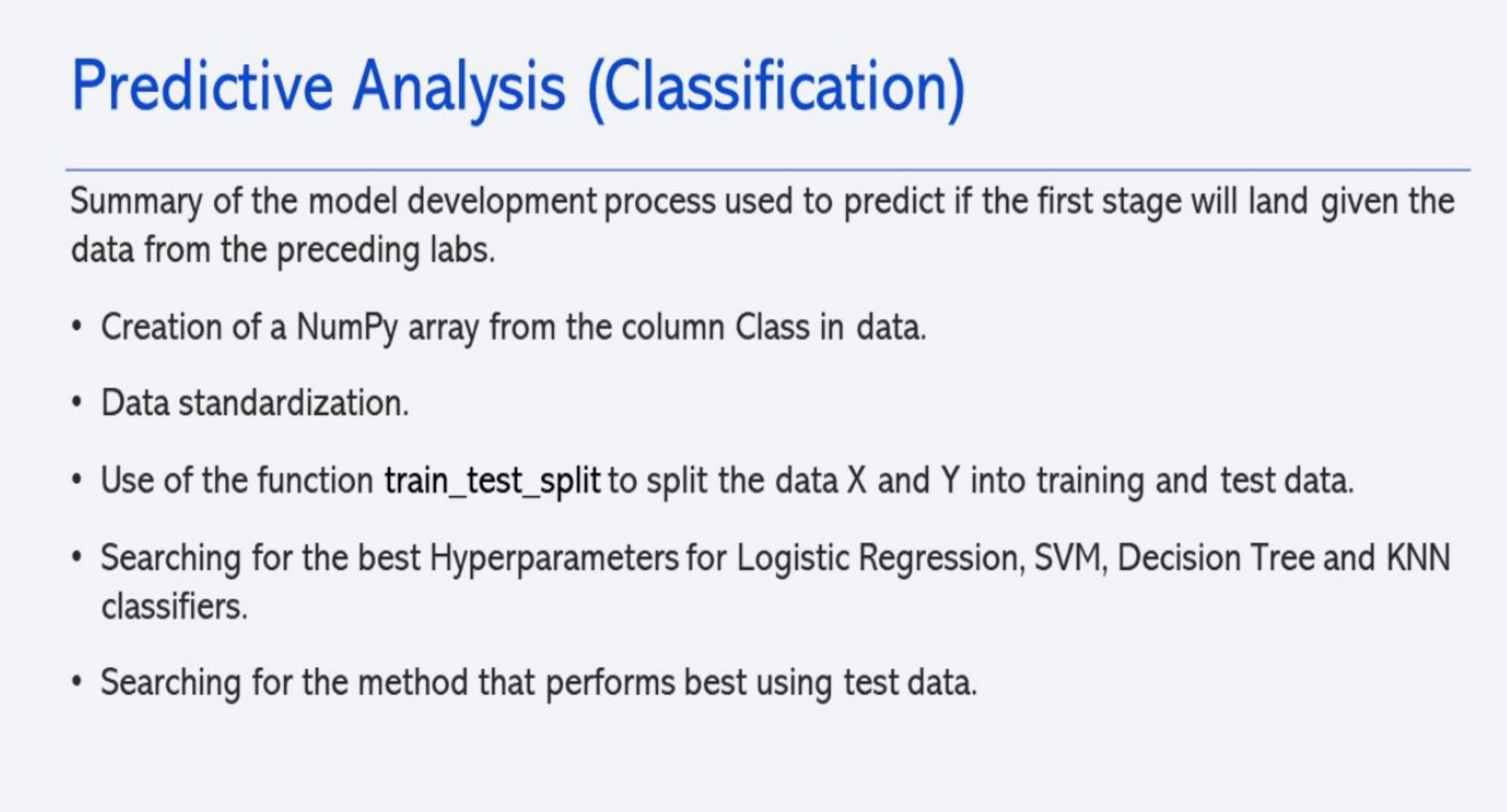
Summary of plots/graphs and interactions that were added to the dashboard to perform interactive visual analytics on SpaceX launch data in real-time.

This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.

- A launch Site Drop-down Input Component.
  There are four different launch sites and a dropdown menu let us select different launch sites.
- A callback function to render *success-pie-chart* based on selected site dropdown.
  The general idea of this callback function is to get the selected launch site from site-dropdown and render a pie chart visualizing launch success counts.
- A range Slider to Select Payload.
  The Slider is to be able to easily select different payload range and see if we can identify some visual patterns.

KSC LC-39A is the site with the higher success launches followed by CCAFS LC-40.



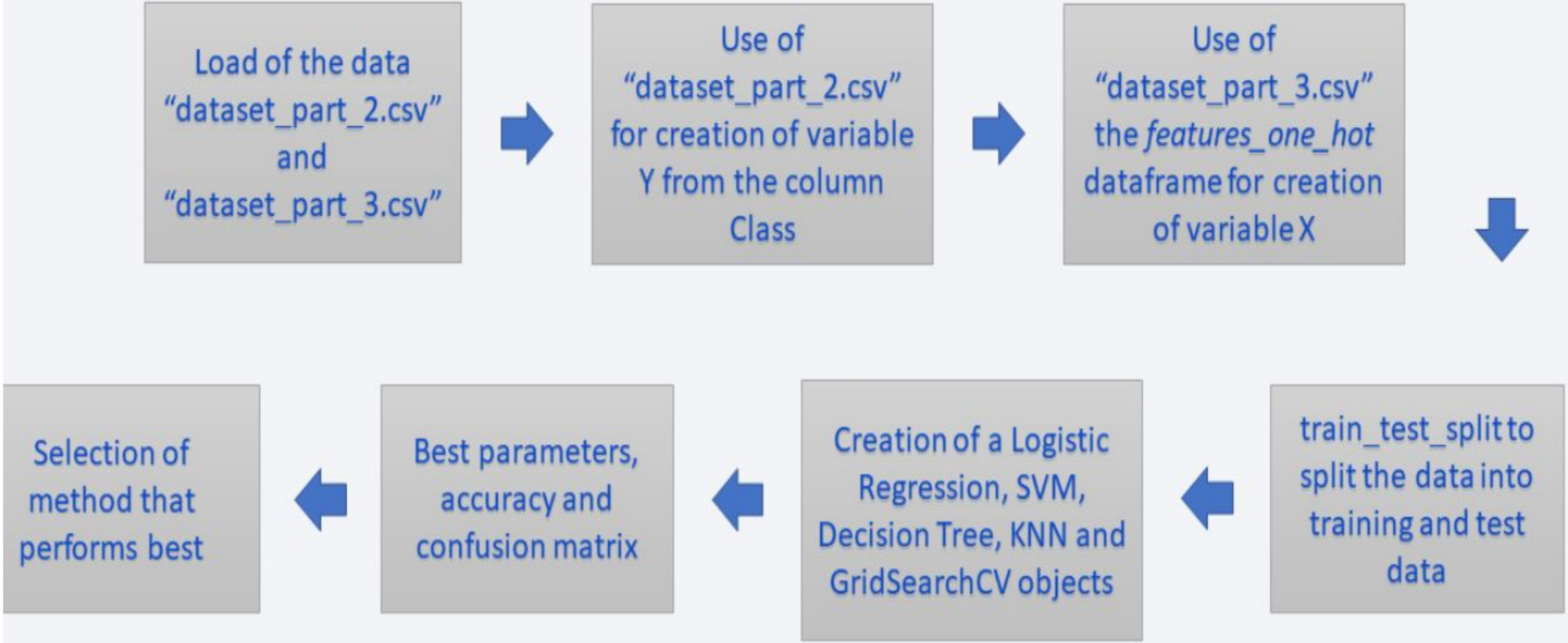


Predictive Analysis (Classification)

Summary of the model development process used to predict if the first stage will land given the data from the preceding labs.

- Creation of a NumPy array from the column Class in data.
- Data standardization.
- Use of the function **train_test_split** to split the data X and Y into training and test data.
- Searching for the best Hyperparameters for Logistic Regression, SVM, Decision Tree and KNN classifiers.
- Searching for the method that performs best using test data.

Predictive Analysis (Classification)



Load of the data "dataset_part_2.csv" and "dataset_part_3.csv"

Use of "dataset_part_2.csv" for creation of variable Y from the column Class

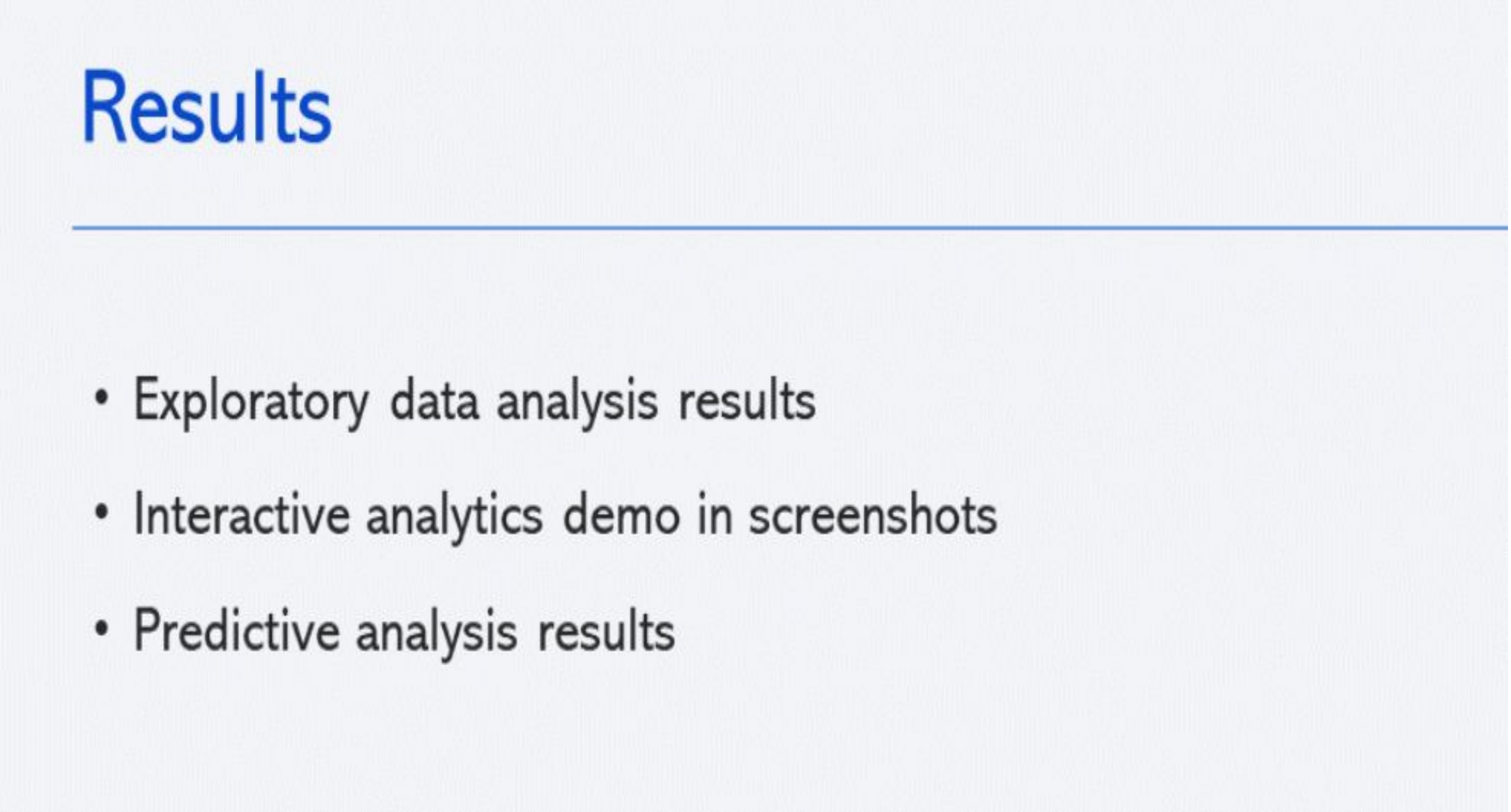
Use of "dataset_part_3.csv" the *features_one_hot* dataframe for creation of variable X

train_test_split to split the data into training and test data

Creation of a Logistic Regression, SVM, Decision Tree, KNN and GridSearchCV objects

Best parameters, accuracy and confusion matrix

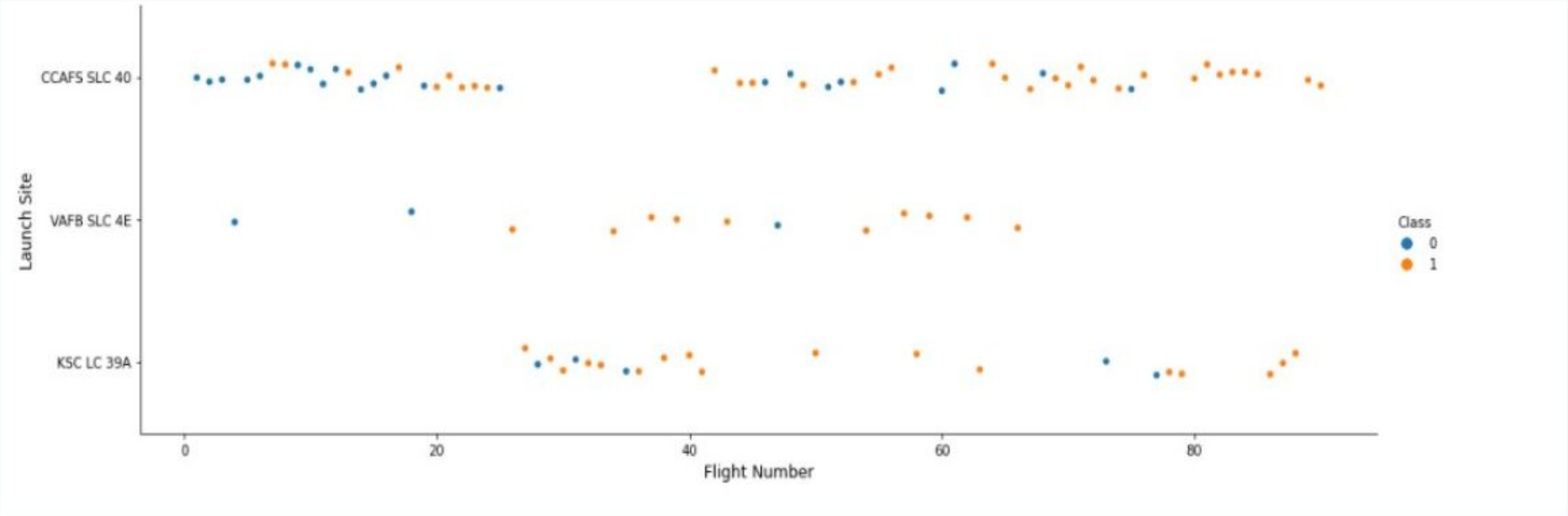
Selection of method that performs best



Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

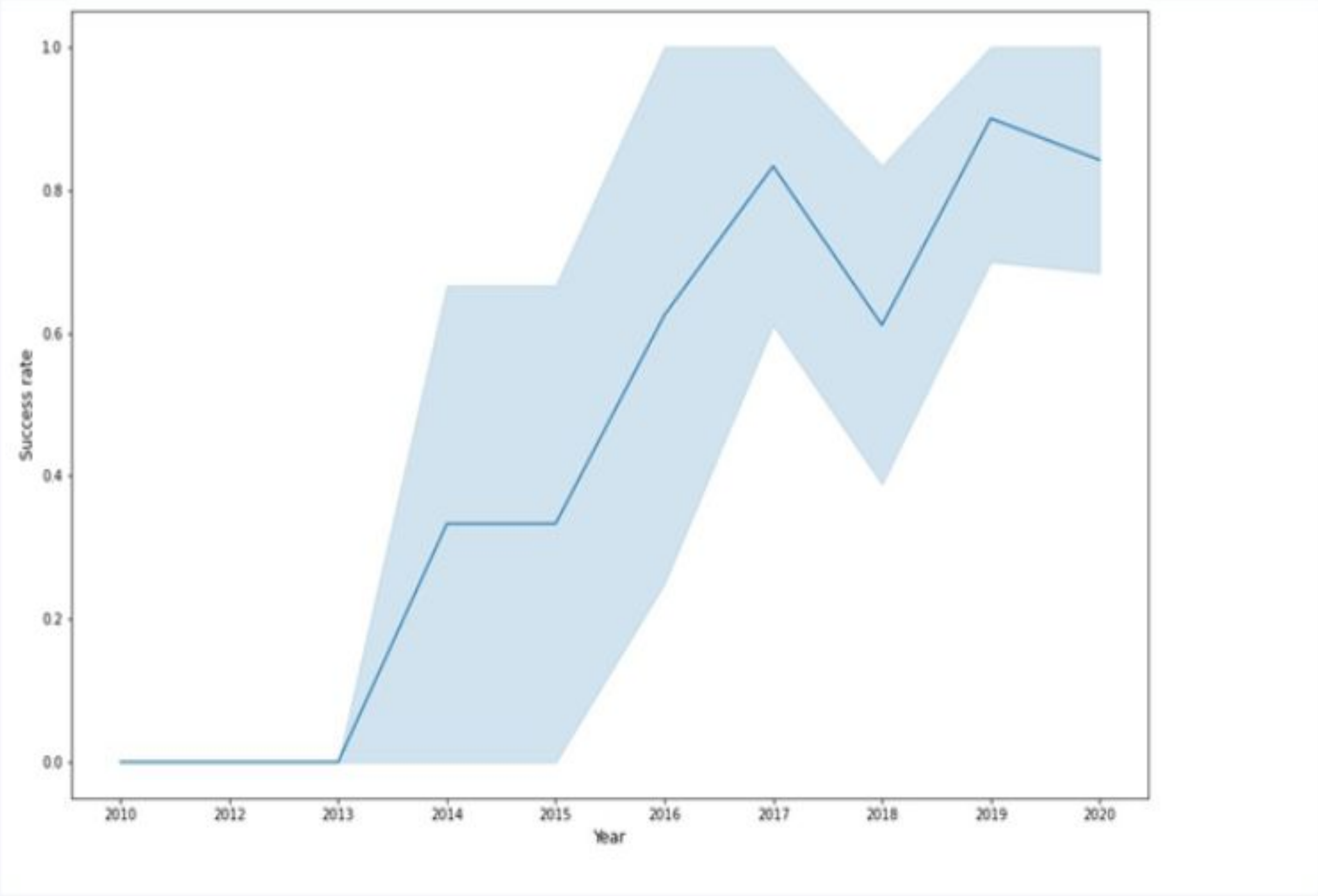
Flight Number vs. Launch Site



- With time the successful rate has increased for every Launch Site, especially for CCAFS SLC 40, where are concentrated the majority of the launches.
- VAFB SLC 4E and KSC LC 39A has a higher successful rate but represents one third of the total launches.

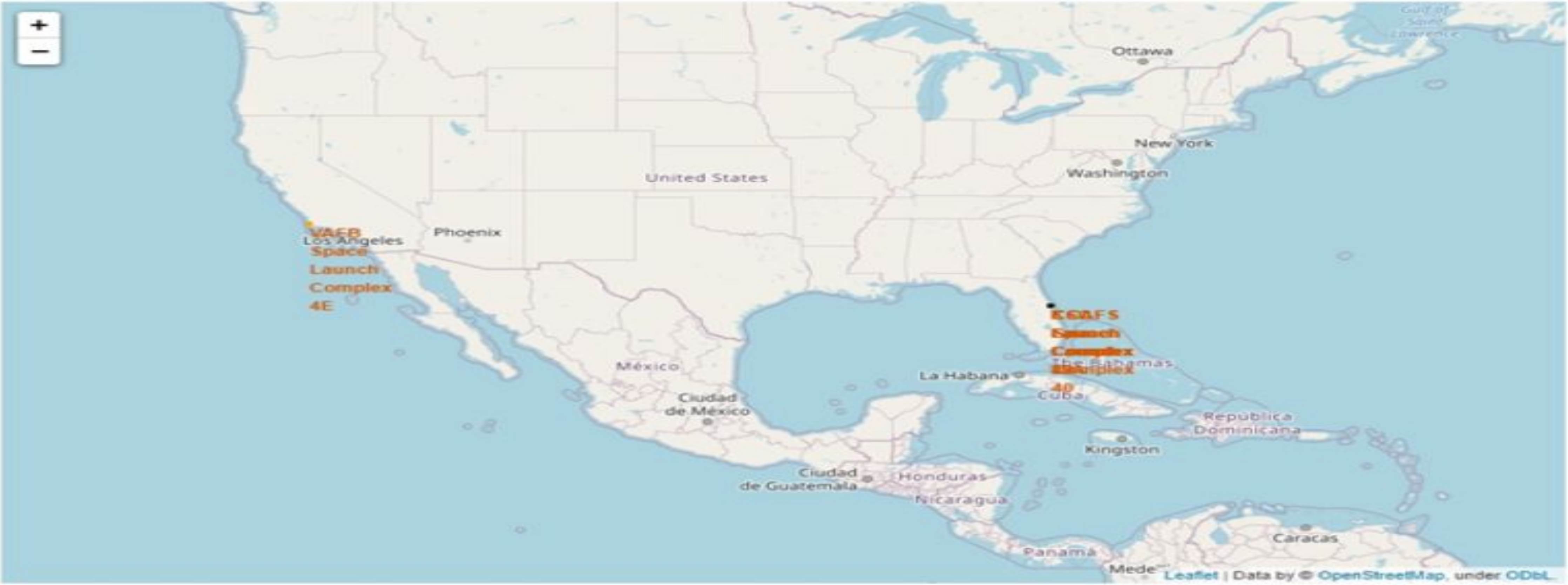
Launch Success Yearly Trend

- The success rate since 2013 kept increasing until 2020.



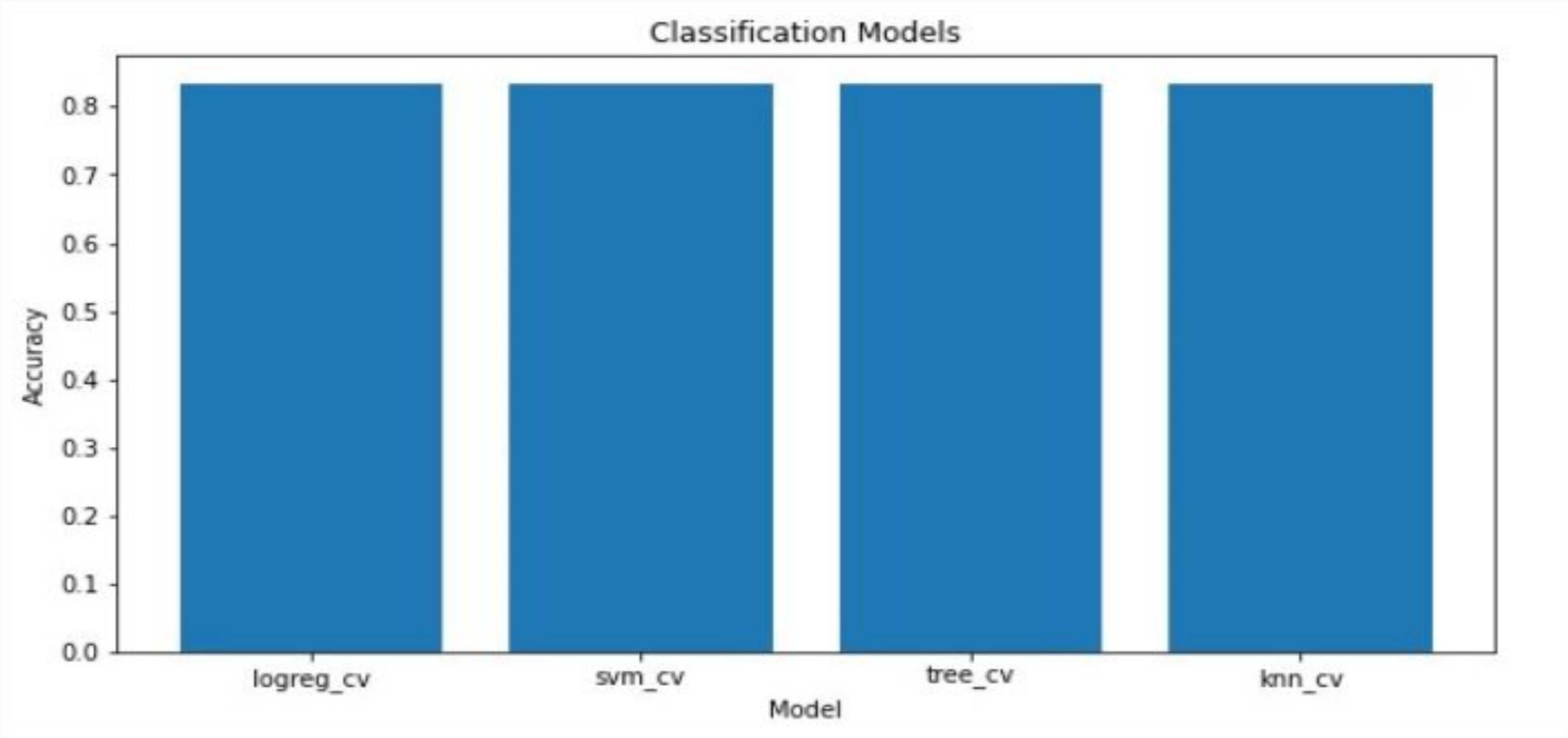
All Launch Sites

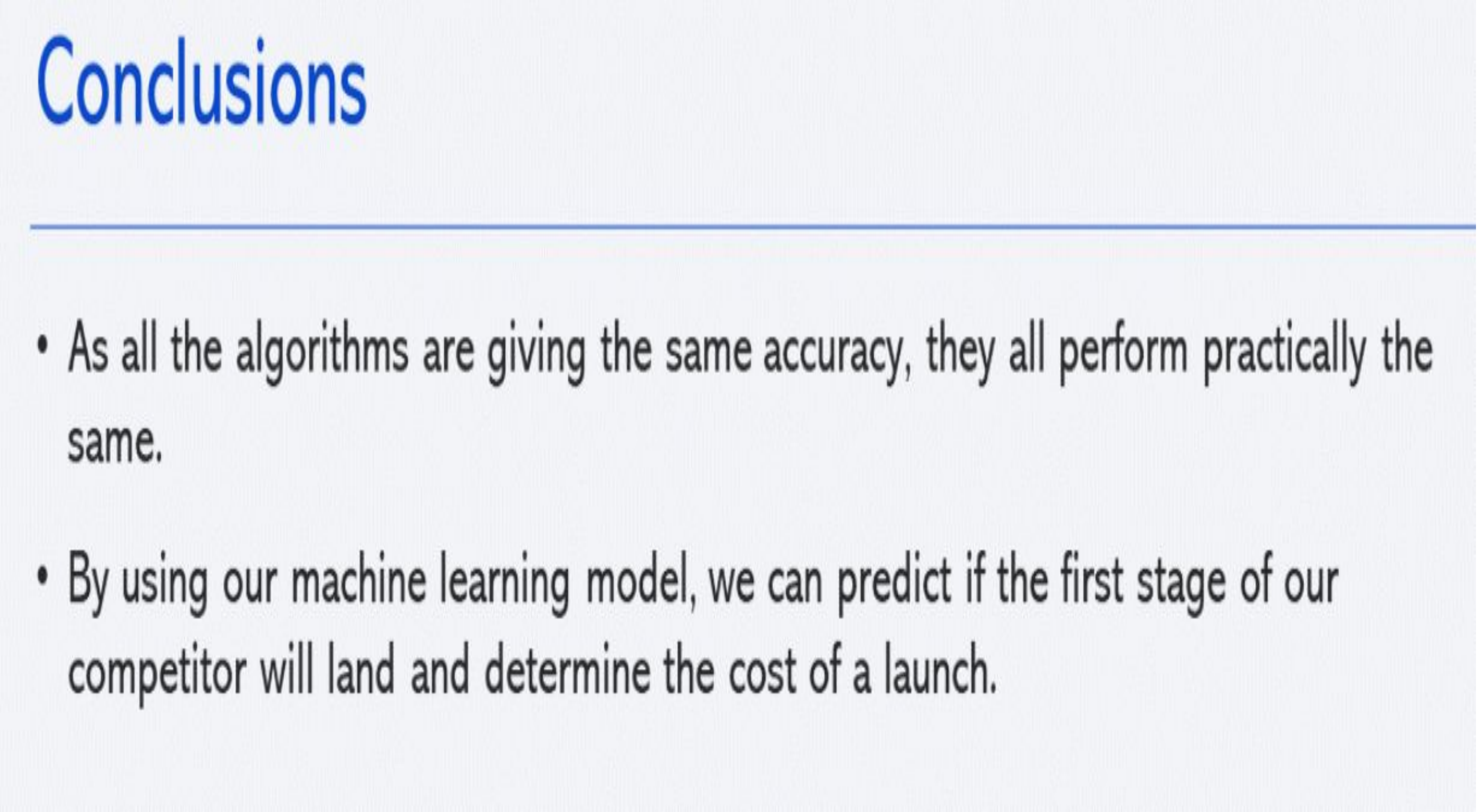
All launch sites are in very close proximity to the coast



Classification Accuracy

The accuracy is the same for all models.





Conclusions

- As all the algorithms are giving the same accuracy, they all perform practically the same.
- By using our machine learning model, we can predict if the first stage of our competitor will land and determine the cost of a launch.