IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Max Waterhout
April 21 2023

Outline

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

Executive Summary

- Following methods for data analyzing:
  - Data collecting via web scraping wikipedia and the SpaceX API
  - Exploratory data analysis including data wrangling, data visualization and interactive visual analytics
  - Used machine learning for predicting rocket launch
- Summary of results:
  - Data analysis allowed the best features for a good prediction
  - Machine learning showed the best models to predict the cost of a rocket launch

Introduction

- This project is based on finding ways to compete with Space X with our new company Space Y
- We want to find answers with data science on different questions:
    - The best way to estimate the total cost for launches

Section
1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from two sources
    - Space X API
    - Web scraping wikipedia
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

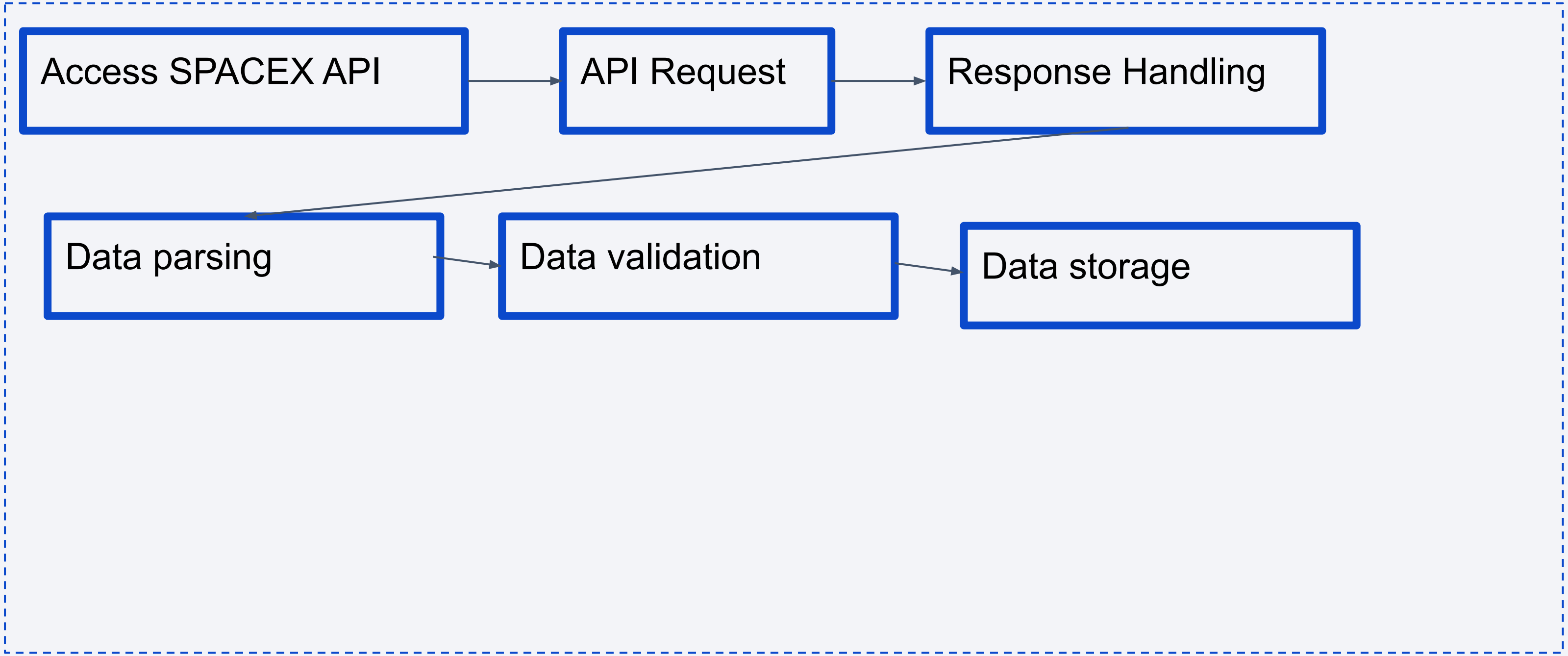
Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The collected data has undergone normalization and partitioning into training and test datasets. Subsequently, four distinct classification models were applied to the data. The accuracy of each model was assessed across various parameter combinations to determine the optimal configuration.

Data Collection - SpaceX API

- Accessing SpaceX API: Initial step involved accessing the SpaceX API endpoint to retrieve launch data.
- API Request: Sent a GET request to the SpaceX API endpoint to fetch the required data.
- Response Handling: Received JSON response containing launch data from the API.
- Data Parsing: Parsed the JSON response to extract relevant information such as launch dates, booster versions, launch sites, and landing outcomes.
- Data Validation: Validated the extracted data to ensure accuracy and completeness.
- Data Storage: Stored the extracted data in a structured format for further processing and analysis.
- Github-URL: https://github.com/maxiew123/falcon_9_project/blob/master/week1/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – SpaceX API



Access SPACEX API

API Request

Response Handling

Data parsing

Data validation

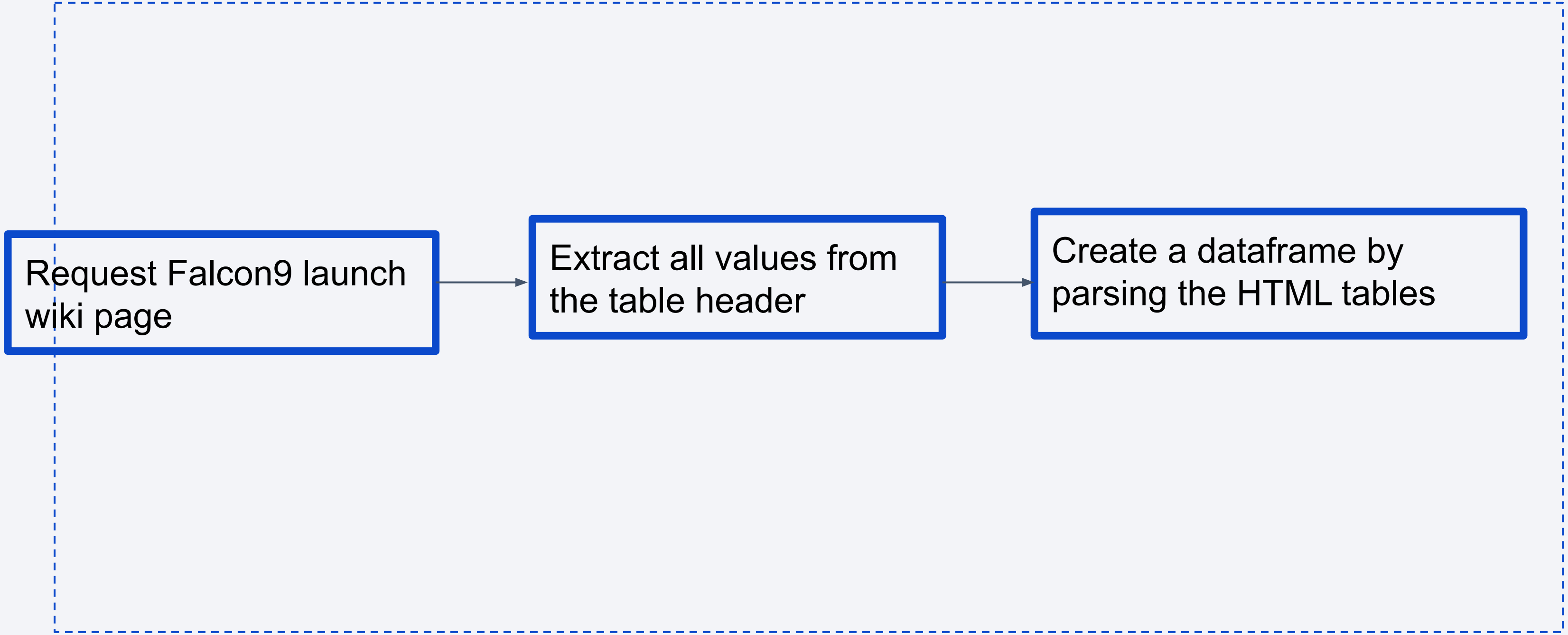
Data storage

Data Collection - Scraping

- • Data from SpaceX launches can also be obtained from Wikipedia;
- Data is downloaded from Wikipedia according to the flowchart in the next slide

- Github url: https://github.com/maxiew123/falcon_9_project/blob/master/week1/jupyter-labs-webscraping.ipynb

Data Collection - Scraping



Request Falcon9 launch wiki page

Extract all values from the table header

Create a dataframe by parsing the HTML tables

Data Wrangling

- The data exploration process commenced with preliminary Exploratory Data Analysis (EDA) to gain insights into the dataset.
- Then summaries were generated for launches per site, occurrences of each orbit, and occurrences of mission outcomes per orbit type.
- Lastly, a landing outcome label was derived from the 'Outcome' column.

EDA

Summarize

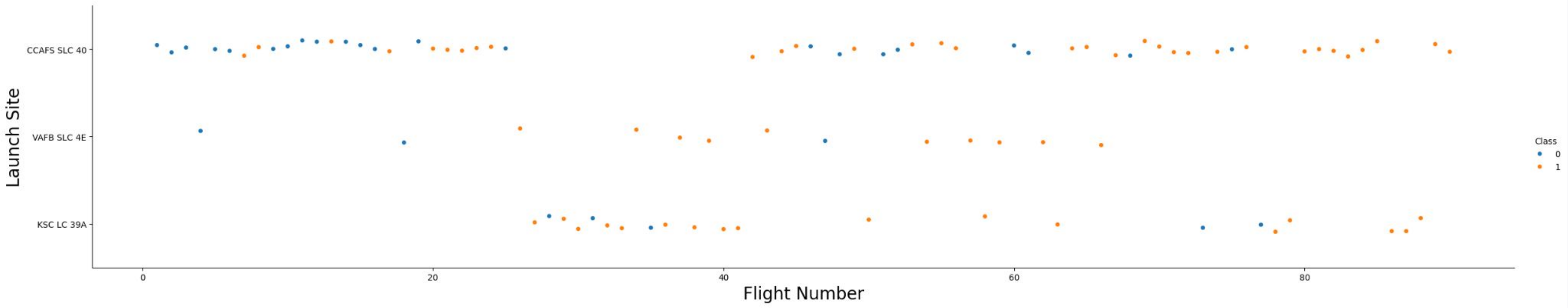
Creation of outcome label

https://github.com/maxiew123/falcon_9_project/blob/master/week1/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore the data, scatterplots and barplots were utilized to visualize the relationships between pairs of features. Specifically, the following relationships were examined:
  - Payload Mass vs. Flight Number
  - Launch Site vs. Flight Number
  - Launch Site vs. Payload Mass
  - Orbit vs. Flight Number
  - Payload vs. Orbit

Github url: https://github.com/maxiew123/falcon_9_project/blob/master/week2/jupyter-labs-eda-dataviz.ipynb



EDA with SQL

1. Display the names of the unique launch sites in the space mission:
   a. %sql select distinct Launch_Site from SPACEXTBL
2. Display 5 records where launch sites begin with the string 'CCA':
   a. %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
3. Display the total payload mass carried by boosters launched by NASA (CRS):
   a. %sql select sum(payload_mass__kg_) from SPACEXTBL WHERE customer = 'NASA (CRS)'
4. Display average payload mass carried by booster version F9 v1.1:
   a. %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE booster_version = 'F9 v1.1'
5. List the date when the first succesful landing outcome in ground pad was achieved:
   a. %sql select min(DATE) from SPACEXTBL WHERE landing_outcome = 'Success (ground pad)'
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
   a. %sql select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
7. List the total number of successful and failure mission outcomes:
   a. %sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery:
   a. %sql select booster_version, payload_mass__kg_ from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015:
   a. %sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME

Github url: https://github.com/maxiew123/falcon_9_project/blob/master/week2/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Map Objects Created and Added to Folium Map:

- Markers: Markers were added to denote specific locations on the map, such as launch sites or important landmarks.
- Circles: Circles were added to represent areas of interest, such as the range of a booster's landing or the coverage area of a ground station.
- Polyline: Polyline was added to draw lines connecting multiple points on the map, such as the flight path of a spacecraft or the trajectory of a rocket launch.
- Popup: Popup windows were added to provide additional information when a marker is clicked, such as the name of a launch site or details about a specific event.

Reasons for Adding Map Objects:

- Markers: Used to pinpoint specific locations and make them easily identifiable on the map.
- Circles: Used to visualize areas of influence or coverage, such as the landing zone for boosters or the range of ground stations.
- Polyline: Used to visualize paths or routes, such as the trajectory of a rocket launch or the flight path of a spacecraft.
- Popup: Used to provide additional information or context when interacting with map markers, enhancing the user experience and understanding of the data.

Github url: https://github.com/maxiew123/falcon_9_project/blob/master/week3/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Launch Success Pie Chart:

- This pie chart displays the distribution of launch outcomes (success, failure, etc.) for the selected launch site.
- Users can select a launch site from the dropdown menu, which dynamically updates the pie chart to show the corresponding launch success distribution.
- The purpose of this plot is to provide users with an overview of launch success rates at different launch sites.

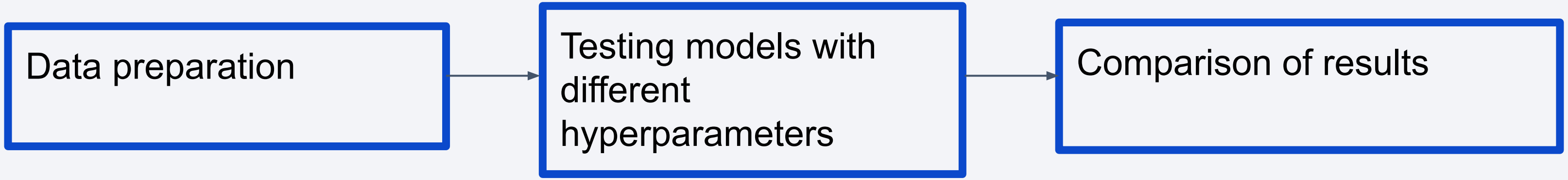
Payload Success Scatter Plot:

- This scatter plot visualizes the relationship between payload mass and launch success.
- Users can use the range slider to select a specific range of payload masses, which updates the scatter plot to display successful and failed missions within that range.
- The scatter plot helps users analyze the impact of payload mass on launch success and identify any trends or patterns.

Github url: https://github.com/maxiew123/falcon_9_project/blob/master/plotly_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.
- Github url: https://github.com/maxiew123/falcon_9_project/blob/master/week4/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Data preparation

Testing models with different hyperparameters

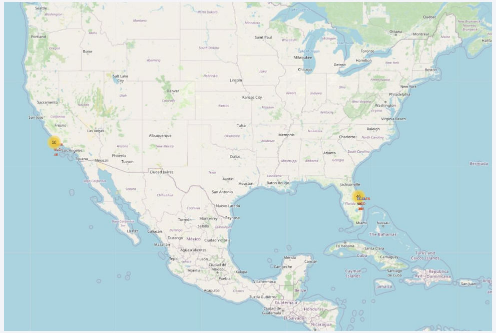
Comparison of results

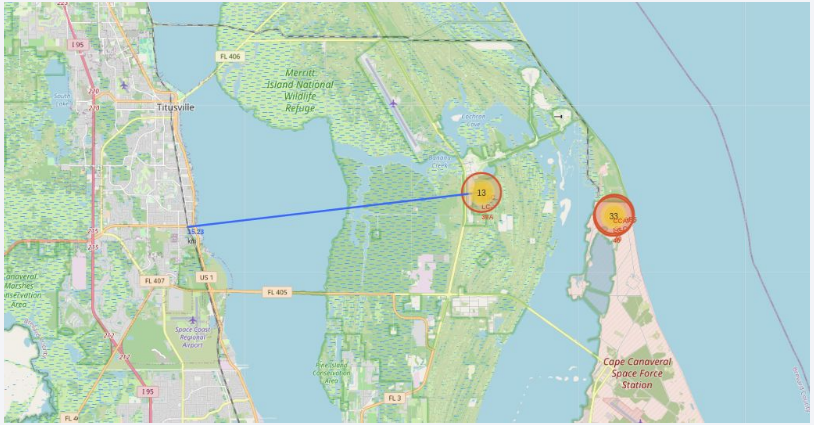
Results: EDA

- Launch Sites:
  - Identified four unique launch sites where SpaceX missions were conducted: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.
  - CCAFS LC-40 was the most frequently used launch site, indicating its significance in SpaceX's mission operations.
- Payload Mass Distribution:
  - Observed a wide range of payload masses, with some missions carrying payloads exceeding 15,000 kg.
  - Payload masses varied depending on the mission objectives, with certain missions focusing on smaller payloads for specific purposes.
- Mission Outcomes:
  - Classified mission outcomes into categories such as success, failure (in flight), failure (pre-launch), and others.
  - Most missions were successful, highlighting SpaceX's reliability in executing space missions.

Results: Interactivate analytics

- By employing interactive analytics, it became apparent that launch sites are strategically located in secure areas, often near bodies of water such as the sea. This positioning ensures enhanced safety measures and facilitates logistical operations, underscoring the importance of robust infrastructure in close proximity to the launch sites.

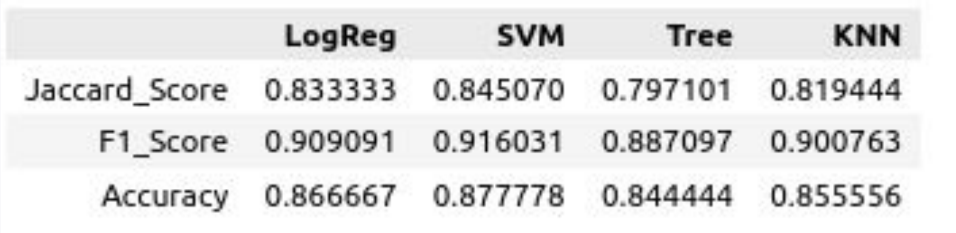




Titusville
Merritt Island National Wildlife Refuge
I 95
FL 406
FL 407
FL 405
US 1
FL 3
15.23 km
13
33
Space Coast Regional Airport
Cape Canaveral Space Force Station

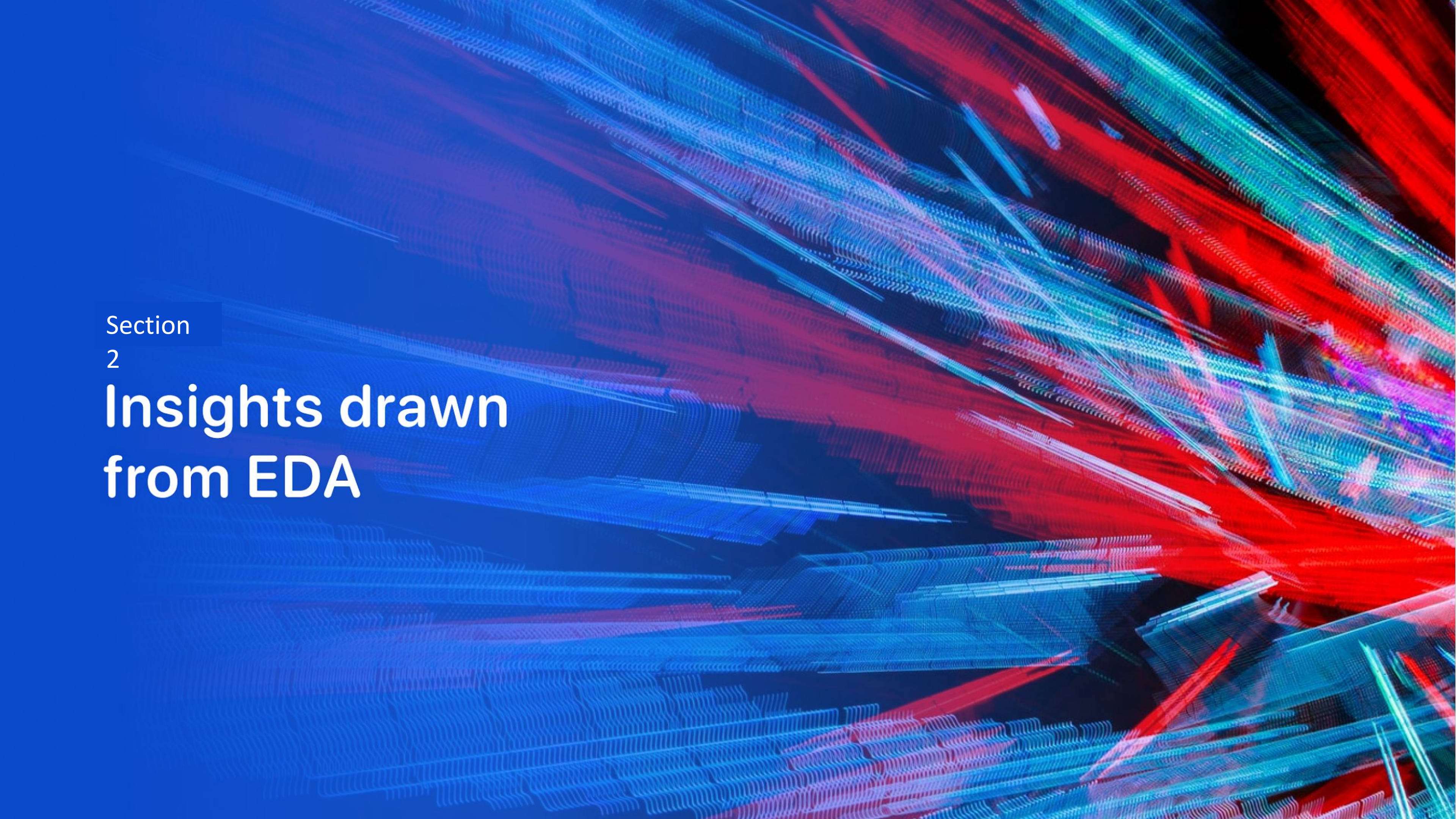
Results: Predictive analysis results

- Here are the performance metrics for the different classification models:



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.797101	0.819444
F1_Score	0.909091	0.916031	0.887097	0.900763
Accuracy	0.866667	0.877778	0.844444	0.855556

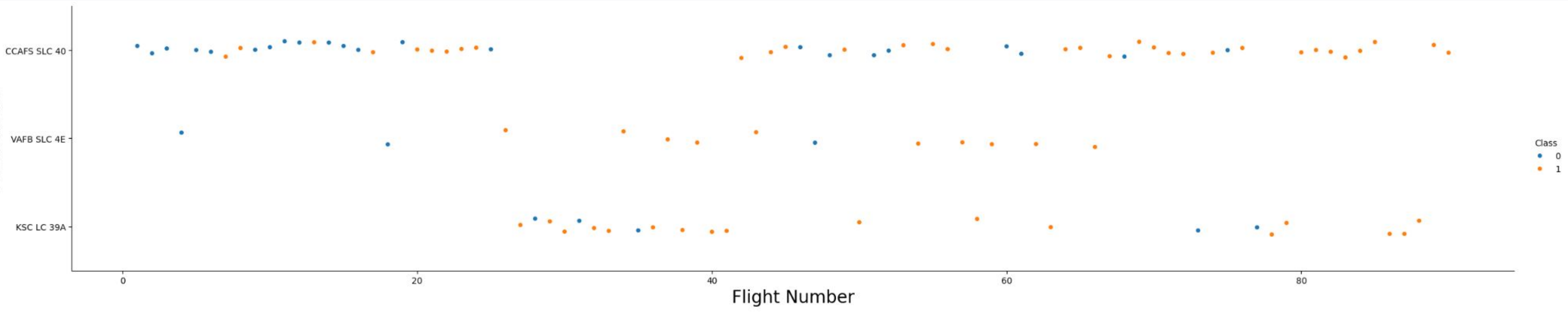
- These scores indicate the effectiveness of each model in predicting the outcome of SpaceX launches. SVM achieved the highest accuracy and F1 score, closely followed by KNN. However, LogReg and Tree models also demonstrated respectable performance across all metrics.



Section
2

Insights drawn from EDA

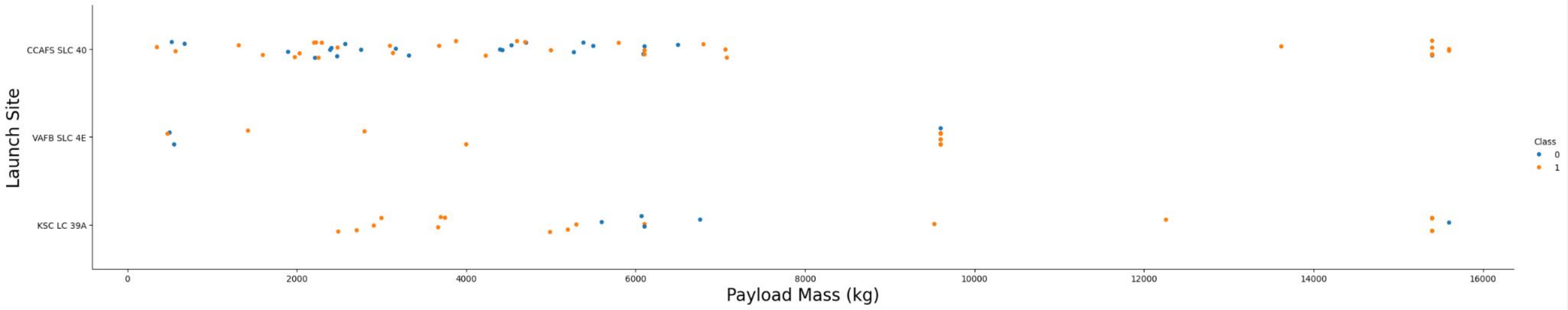
Flight Number vs. Launch Site



- Based on the plot above, it's evident that the optimal launch site currently is CCAF5 SLC 40. The majority of recent launches from this site have been successful.

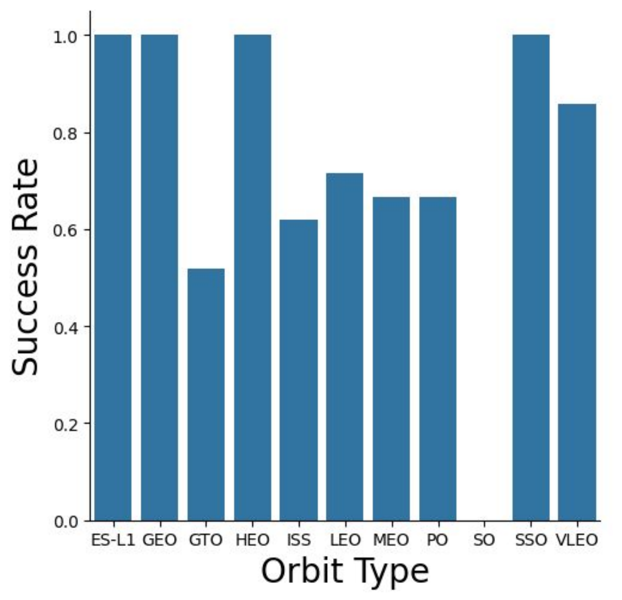
Payload vs. Launch Site

- Payloads surpassing 12,000kg appear feasible primarily at the CCAFS SLC 40 and KSC LC 39A launch sites.



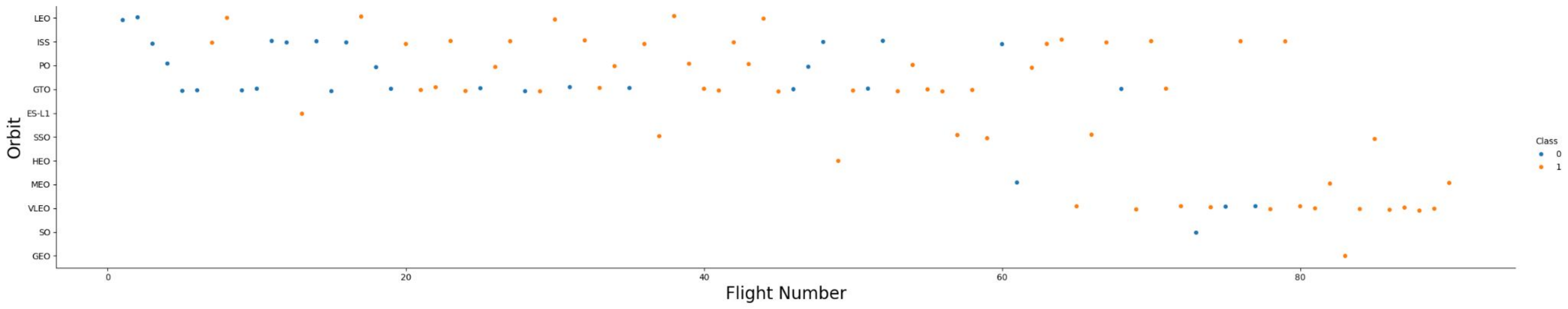
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits: ES-L1, GEO, HEO and SSO.



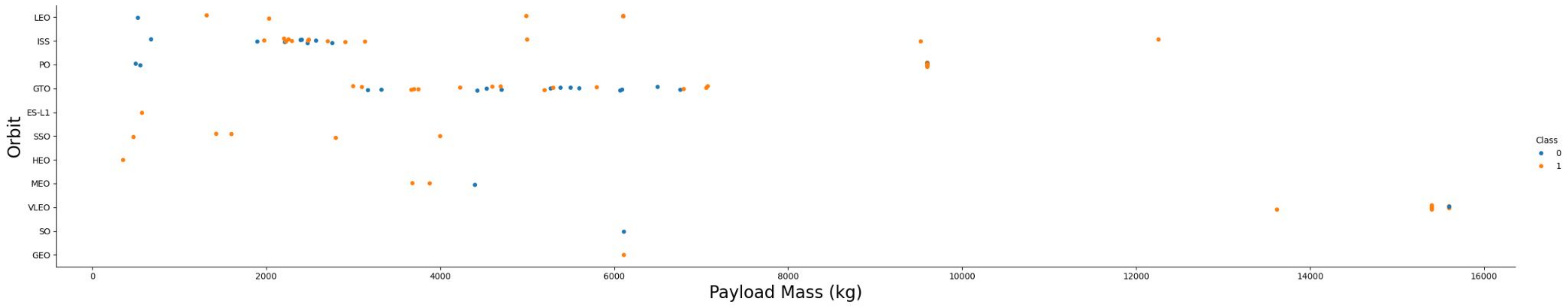
Flight Number vs. Orbit Type

- Success rate improved over time to all orbits;



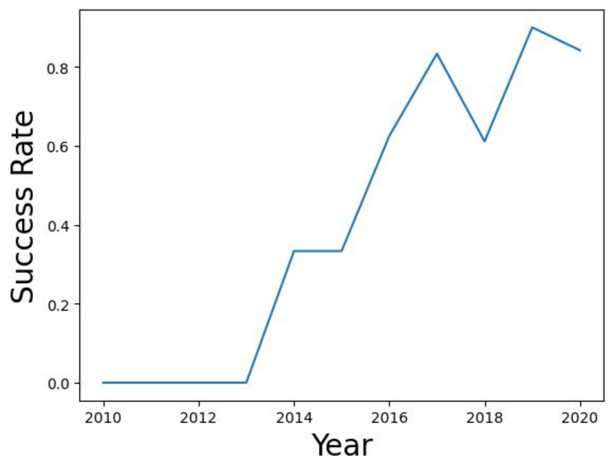
Payload vs. Orbit Type

- There have been relatively few launches to the Synchronous Orbit (SO) and Geostationary Orbit (GEO).



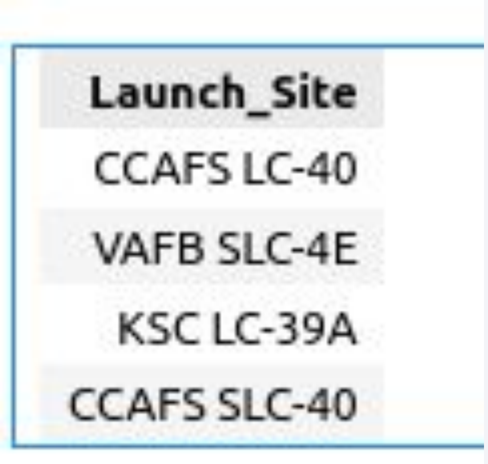
Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;



All Launch Site Names

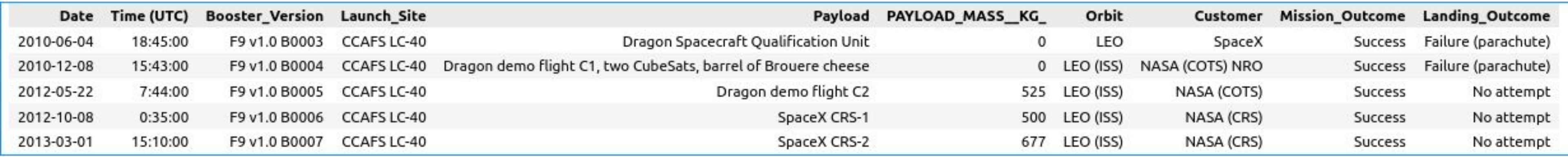
- These are the unique launch sites



Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

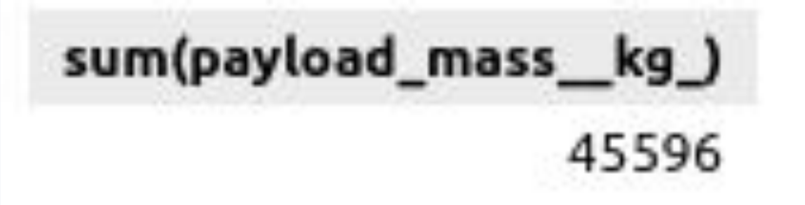
- These are the first 5 records where launch sites begin with `CCA`



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

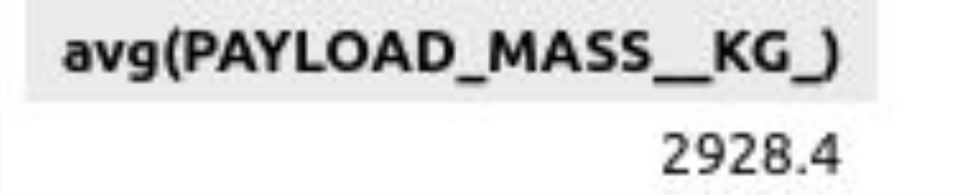
- This is the calculation for the total payload carried by boosters from NASA



sum(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1

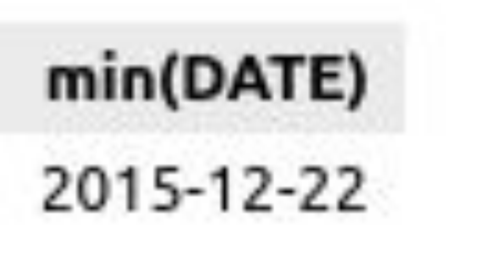
- Calculate the average payload mass carried by booster version F9 v1.1



avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

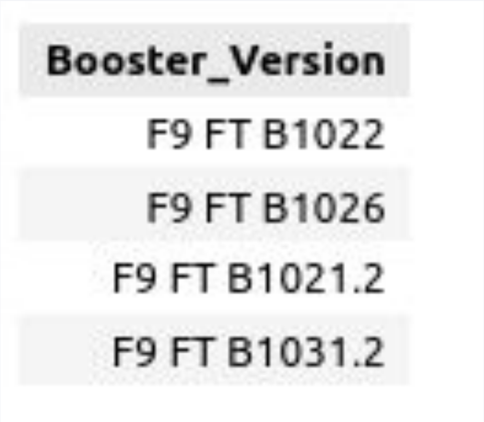
- Find the dates of the first successful landing outcome on ground pad



min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

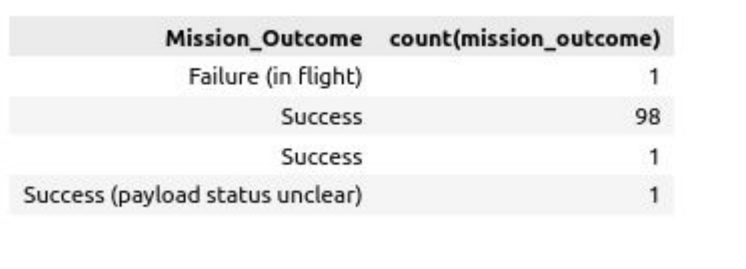
- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

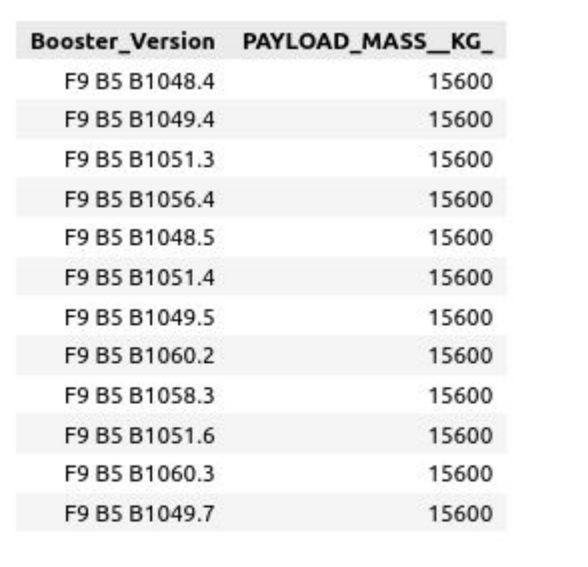
- Calculate the total number of successful and failure mission outcomes



Mission_Outcome	count(mission_outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

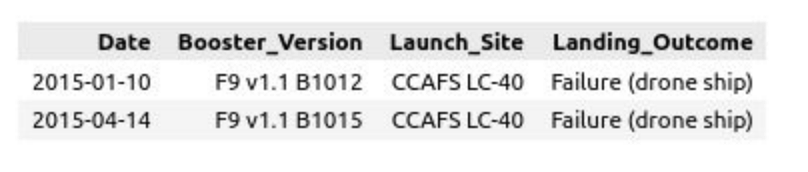
- List the names of the booster which have carried the maximum payload mass



Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

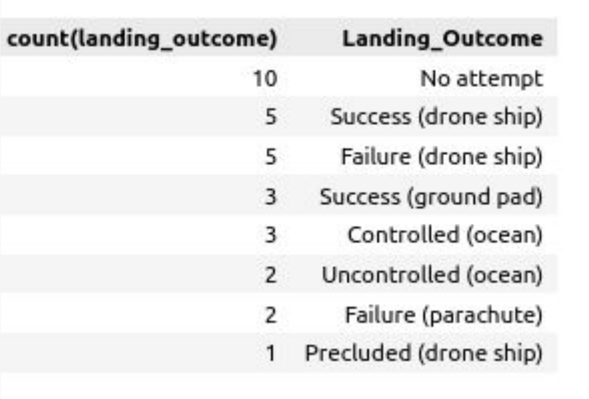
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



Date	Booster_Version	Launch_Site	Landing_Outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order



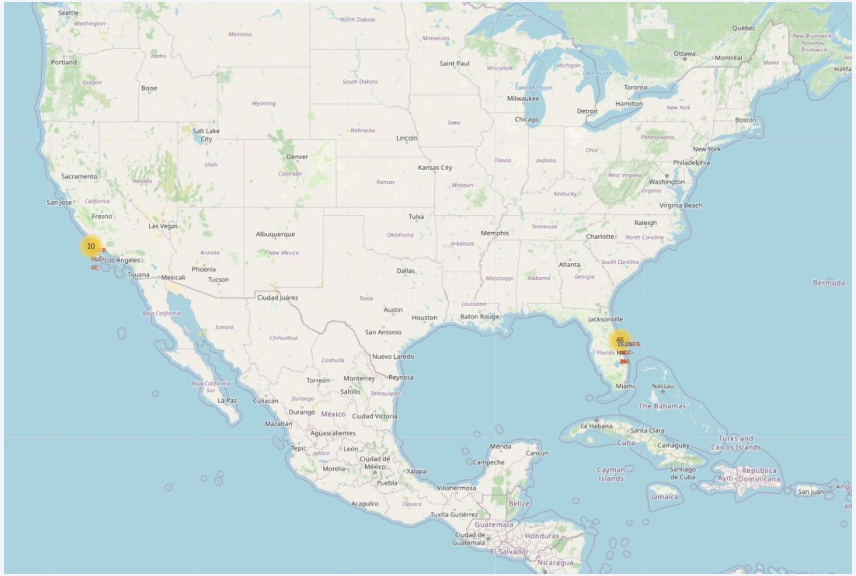
count(landing_outcome)	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

Section
3

Launch Sites Proximities Analysis

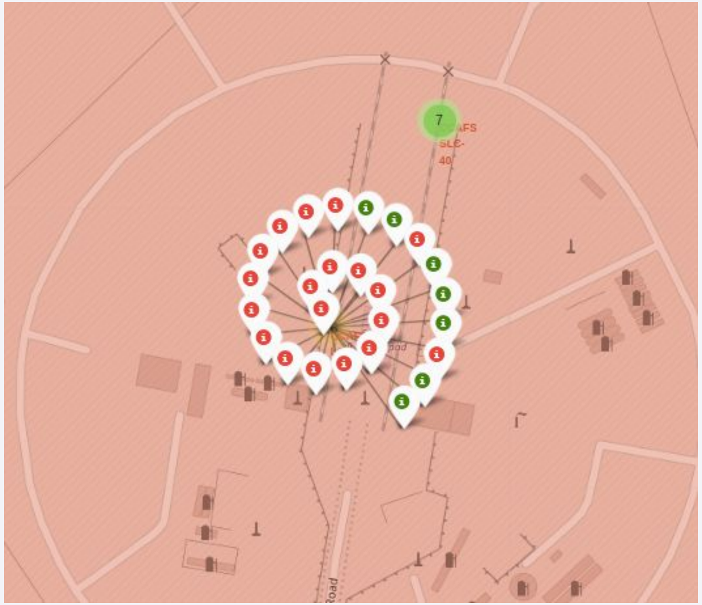
All launch sites

- Launch sites are near sea and not too far from roads and railroads.



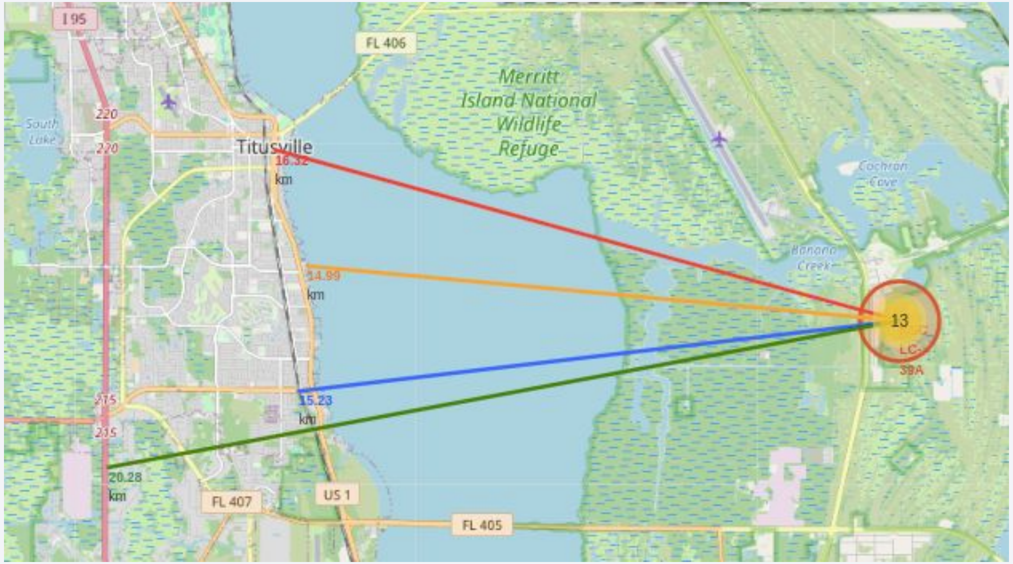
Launch Outcomes by Site

- Green markers indicate successful and red ones indicate failure.



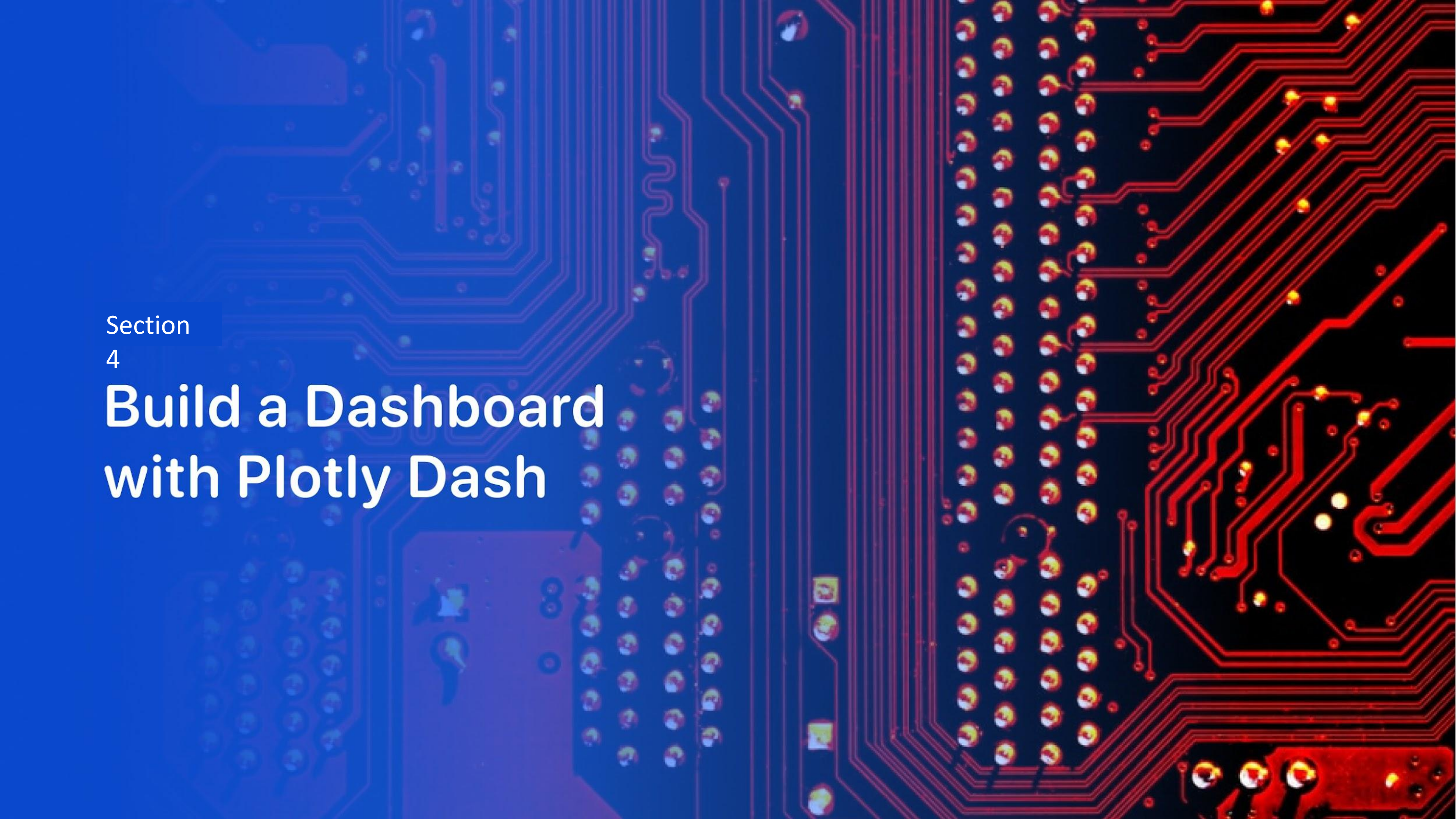
Logistics and Safety

- This launch site has good logistics aspects, being near railroad and road and far from areas.



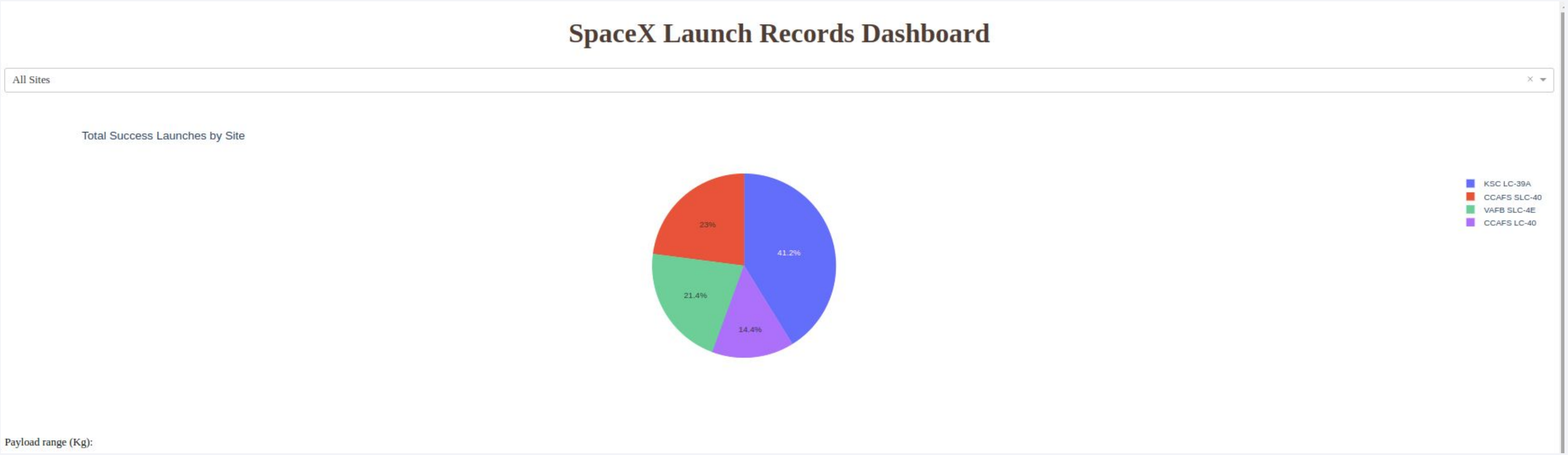
Section
4

Build a Dashboard with Plotly Dash



Total success launches

- The location from which launches are conducted appears to be a crucial factor influencing the success of missions.



SpaceX Launch Records Dashboard

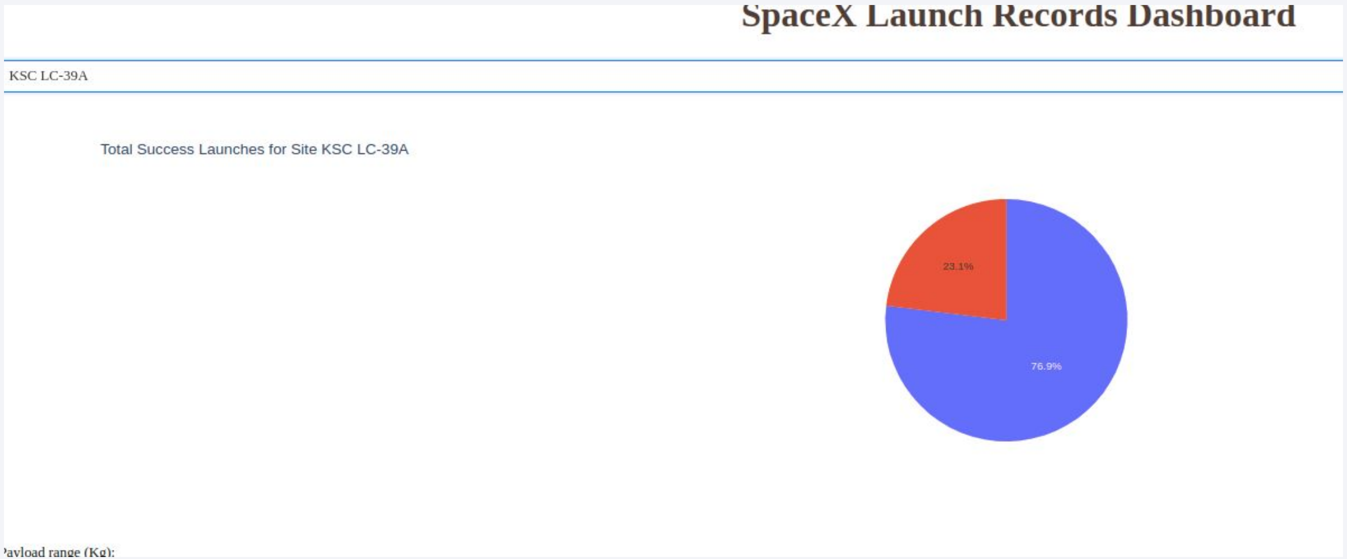
All Sites

Total Success Launches by Site

Payload range (Kg):

Highest success rate

- KSC LC-39A has a succes rate of 77%



SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for Site KSC LC-39A

'avload range (Kg):

<Dashboard Screenshot 3>

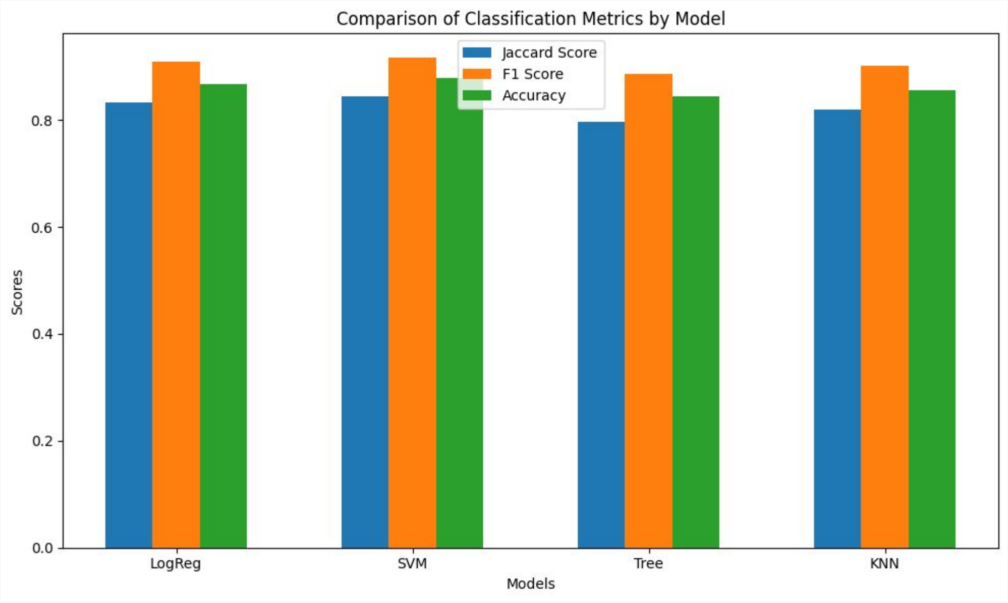
-

Section
5

Predictive Analysis (Classification)

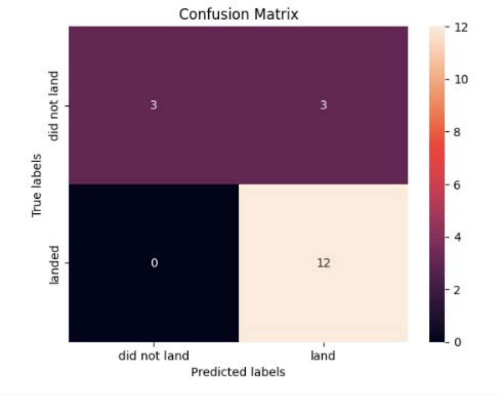
Classification Accuracy

- The model with the highest scores is SVM followed by LogReg



Confusion Matrix

- Confusion matrix of SVM proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- Various data sources underwent analysis, with conclusions evolving throughout the process.
- KSC LC-39A emerged as the optimal launch site.
- Launches surpassing 7,000kg demonstrated lower risk levels.
- While the majority of mission outcomes were successful, successful landing outcomes showcased an upward trend over time, indicating advancements in processes and rocket technology.
- SVM stands out as a viable option for predicting successful landings, potentially leading to increased profits

Thank you!