IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abhishek Tripathi
20-11-2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data Wrangling
  - EDA with data visualization
  - EDA with SQL
  - Build an interactive map with Folium
  - Building a Dashboard with Plotly Dash
- Summary of all results
  - EDA results
  - Interactive analysis
  - Predictive analysis results

Introduction

- In this analysis we will focus on the study of the first stage of the SpaceX Falcon 9 rocket in order to obtain conclusions that allow us to make cost projections as well as obtain insight on the implications in the area beyond the economic ones
- We will work on finding the following solutions:
  - How to predict if the rocket will land successfully?
  - What parameters can determine the success rate of a landing successful?
  - Project and determine the costs of future launches.

Section 1

Methodology

Methodology

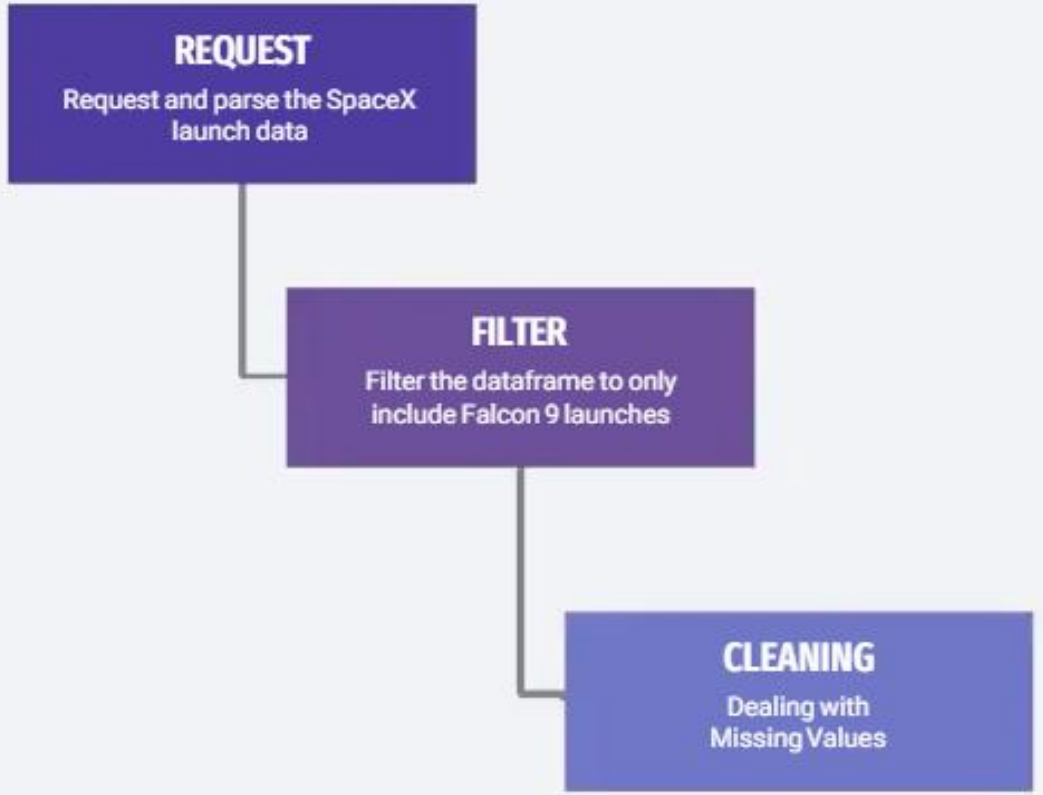
- Data collection methodology:
  - Data was collected from the past SpaceX missions.
  - Web scraping from Wikipedia
- Perform data wrangling
  - Calculated the number of launches on each site.
  - Calculated the number and occurrences of each orbit.
  - Calculated the number and occurrences of mission outcome per orbit type.
  - Created a landing outcome label from Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Datasets were collected from previous SpaceX mission and Wikipedia pages.
- Below processes were obtained to filter, clean and transform the data to prepare for modelling

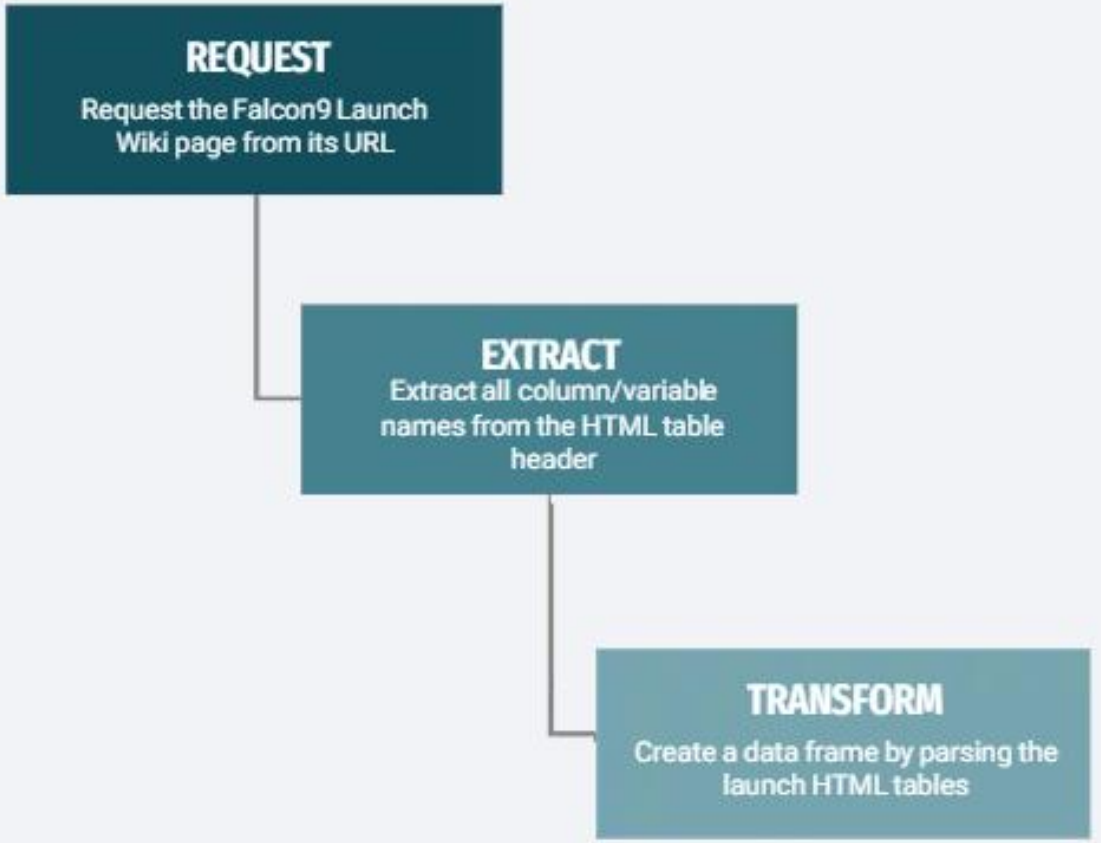
Data Collection – SpaceX API

- We make a get request to the SpaceX API. We also perform some basic data wrangling and formatting.
- The procedure is summarized in the flowchart.
- It can be seen in detail in the following GitHub link.



Data Collection - Scraping

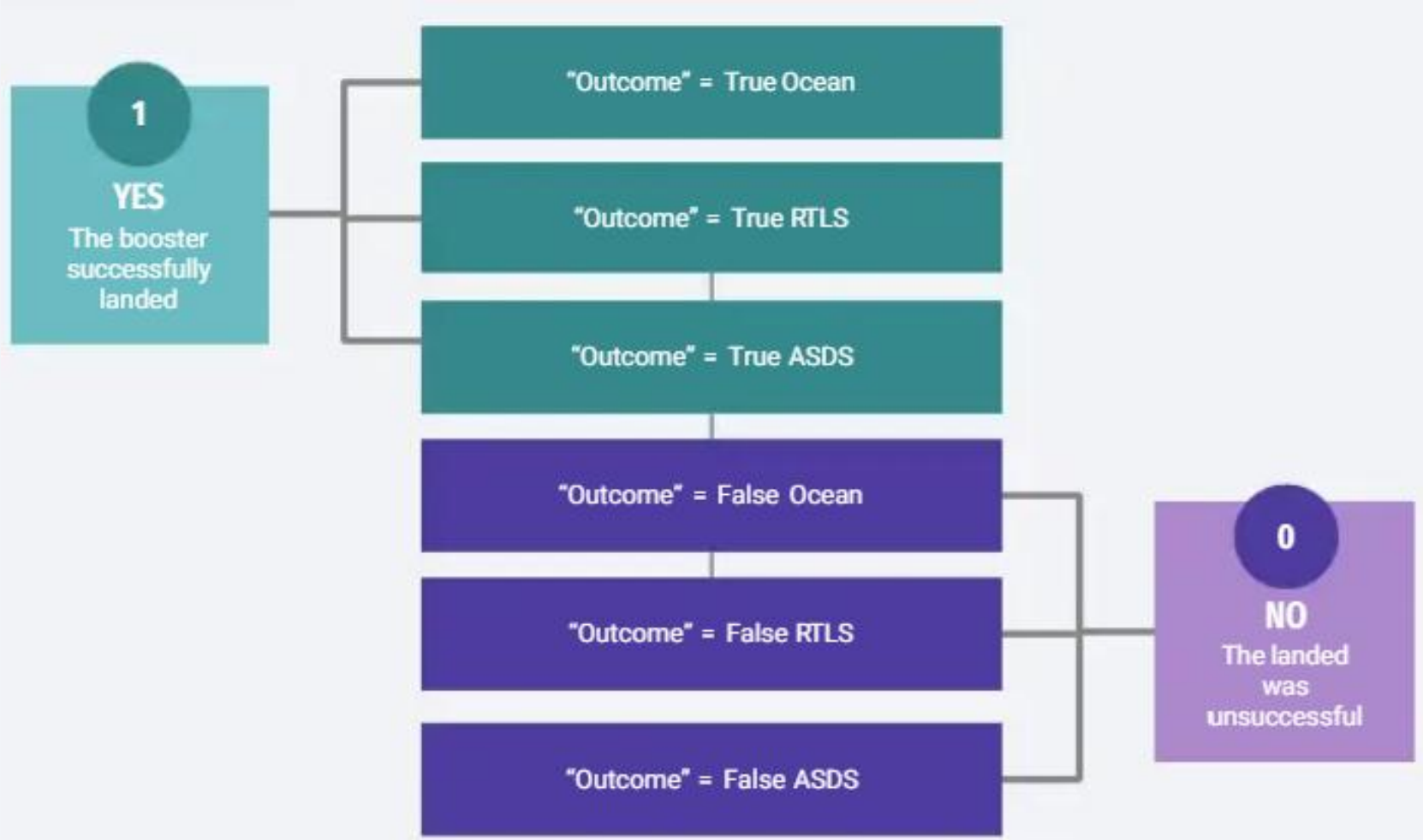
- We performed web scraping to collect historical Falcon 9 launch records from a Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches”.
- The procedure is summarized in the flowchart.
- It can be seen in detail in the following GitHub link.



Data Wrangling

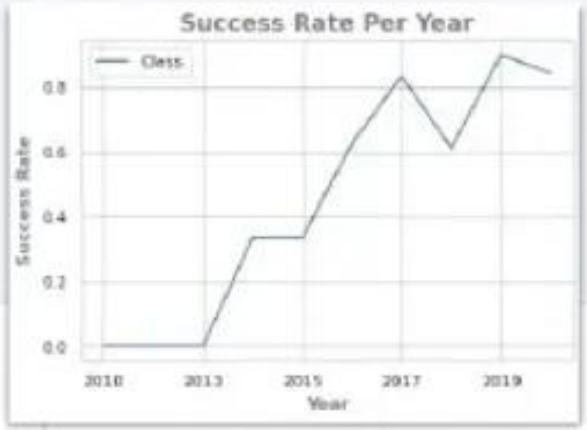
- Through a preliminary exploratory analysis identifying the transformations that are required in the data set to prepare them.
- We will process the landing data into valid tags for training the predictive models later:
  - Training tags with “1” will mean the rocket landed successfully and “0” will mean it was unsuccessful.
- The procedure is summarized in the flowchart in the next slide.
- It can be seen in detail in the following GitHub link.

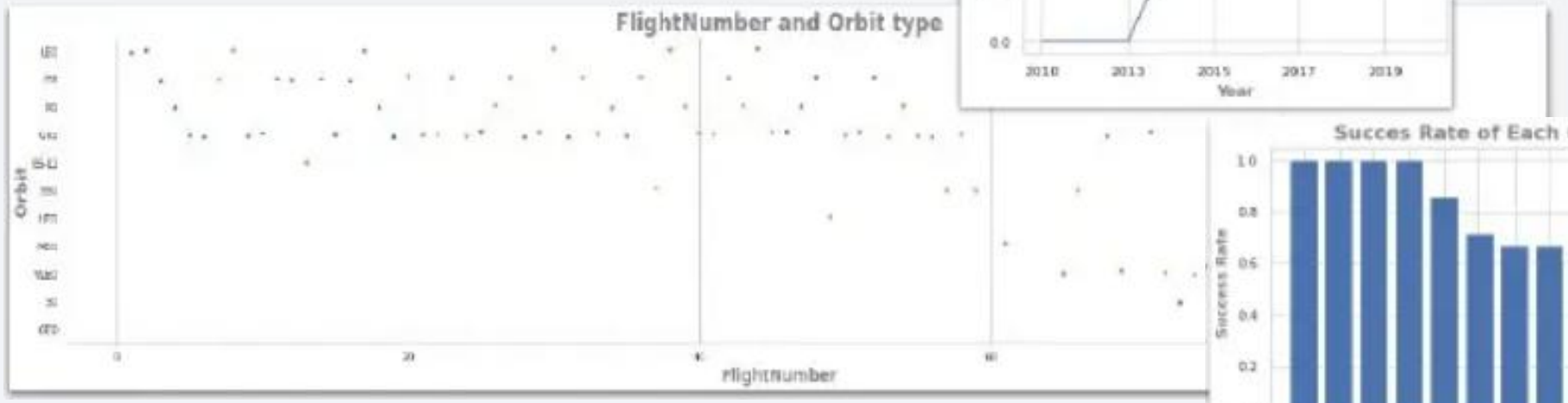
Data Wrangling – *Convert outcomes into training labels*



EDA with Data Visualization

- Exploratory Data Analysis to visualize the relationship between:
  - Flight Number and Launch Site.
  - Payload and Launch Site.
  - Success rate of each orbit type.
  - Flight Number and Orbit type.
  - Payload and Orbit type.
  - Visualize the launch success yearly trend.





- Scatterplot, Bar chart and line graph were used for simplifying the comparative analysis.
- It can be seen in detail in the following GitHub link.

EDA with SQL

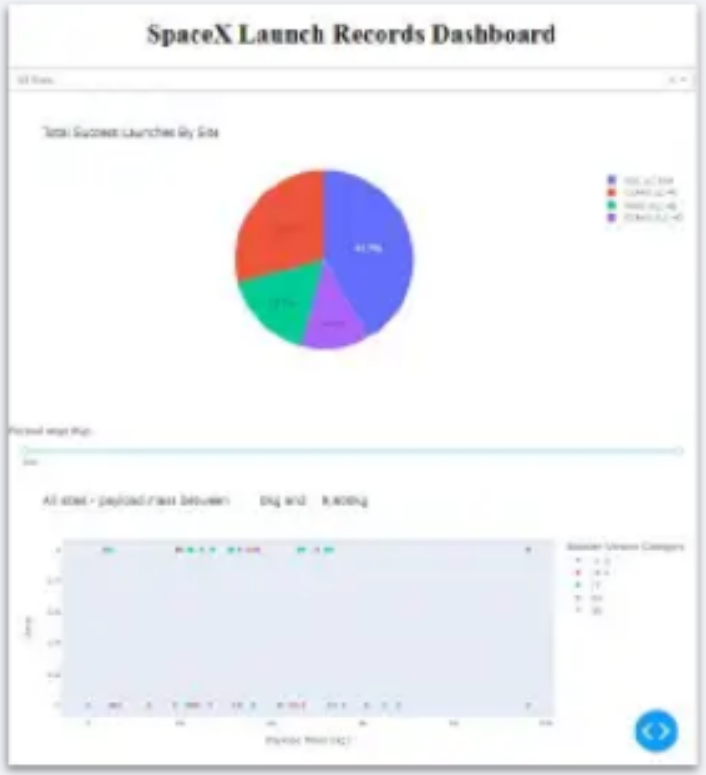
- SQL queries performed:
  - Names of the unique launch sites in the space mission.
  - Top 5 launch sites whose name begin with the string 'CCA'.
  - Total payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved.
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg.
  - Total number of successful and failure mission outcomes.
  - Names of the booster versions which have carried the maximum payload mass.
  - Failed landing outcomes in drone ship, their booster versions and launch site names in the year 2015.
  - Rank of the count of landing outcomes (such as failure(drone ship) or success(ground pad)) between the date 2010-06-04 and 2017-03-20.
- It can be seen in detail in the following GitHub link.

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps.
- Indications of each element:
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates.

- It can be seen in detail in the following GitHub link.

Build a Dashboard with Plotly Dash



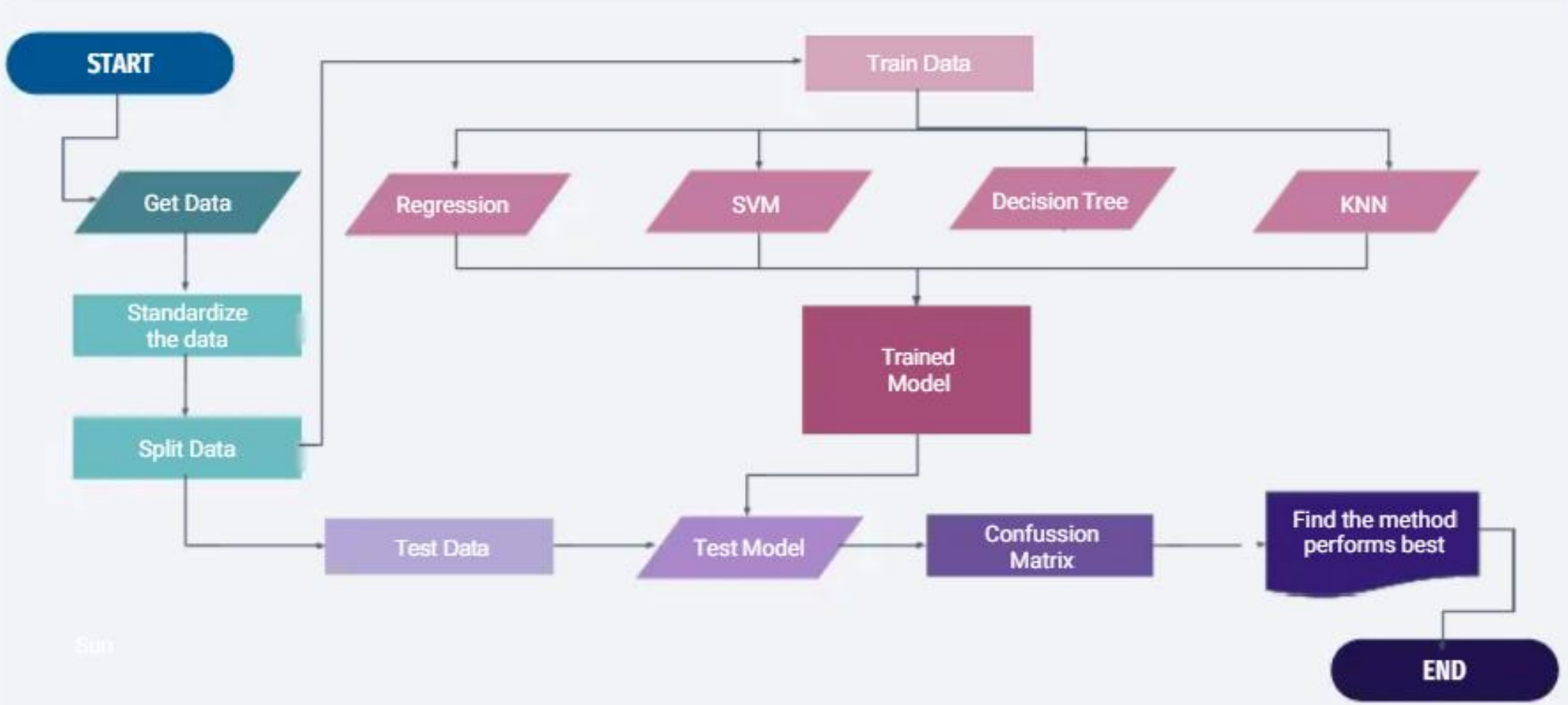
It can be seen in detail in the following GitHub link.

- Elements
  - Dropdown list for the launch site.
  - RangeSlider for selecting the payload mass.
  - PieChart: for showing the success rate of each launch site, or showing the number of successful landing outcomess.
  - Scatterplot: Show success/failure by payload and booster version.
- Findings:
  - Which site has the largest successful launches? KSC LC-39A.
  - Which site has the highest launch success rate? KSC LC-39A (success rate 76.9%).
  - Which payload range(s) has the highest launch success rate? 2000-4000.
  - Which payload range(s) has the lowest launch success rate? 6000-8000.
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest.
  - Launch success rate? B5(only one successful start), apart from that FT(15 successes and 8 failures).

Predictive Analysis (Classification)

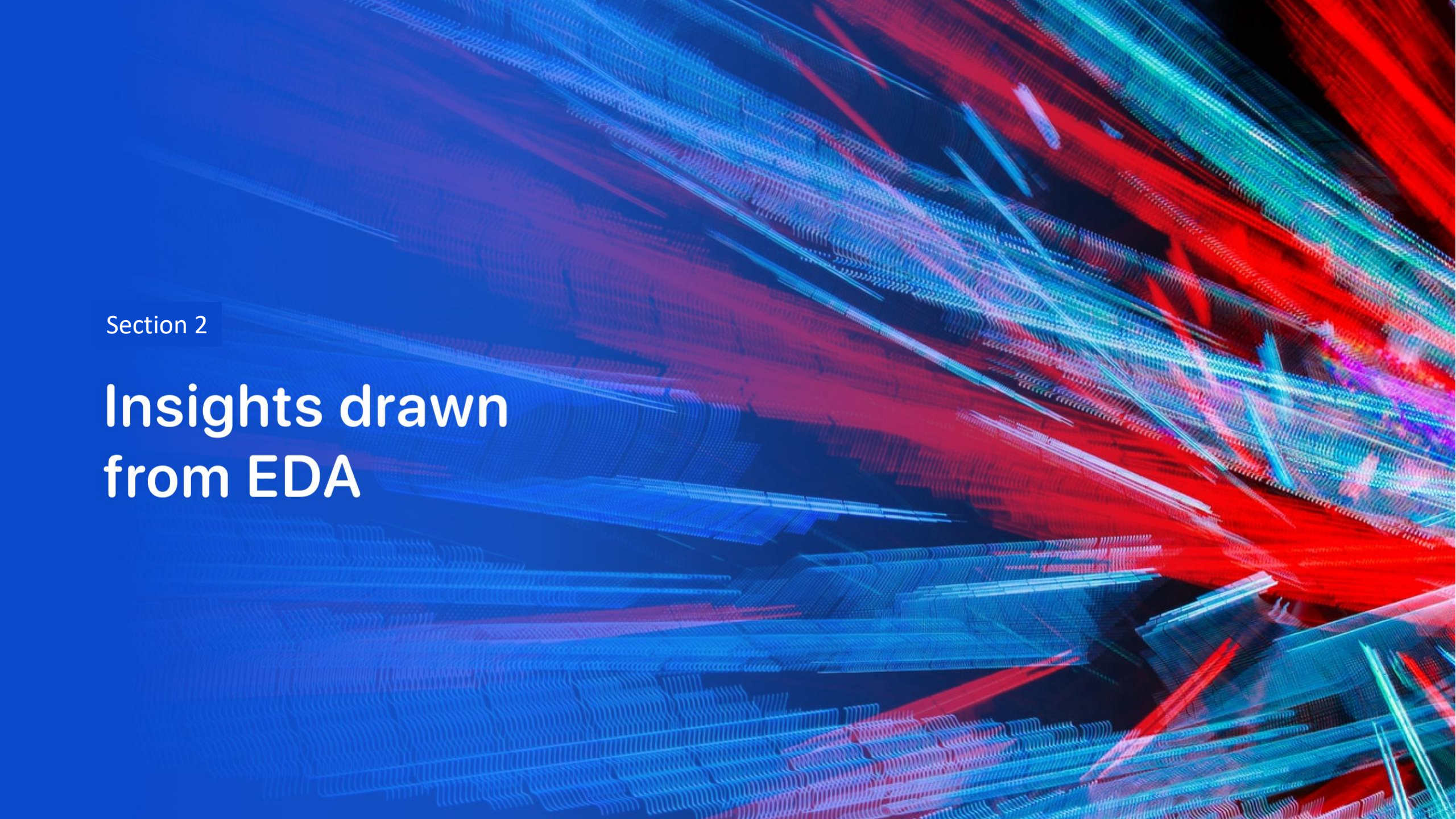
- We create a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.
- Performed EDA and determined Training labels:
  - Create a column for the class.
  - Standardize the data.
  - Split the data into training and testing data.
- Find the best hyperparameter for SVM, Classification Trees and Logistic Regression.
  - Find the method that performs the best using test data.
- The procedure is summarized in the flowchart in the next slide.
- It can be seen in detail in the following GitHub link.

Predictive Analysis (Model flowchart)

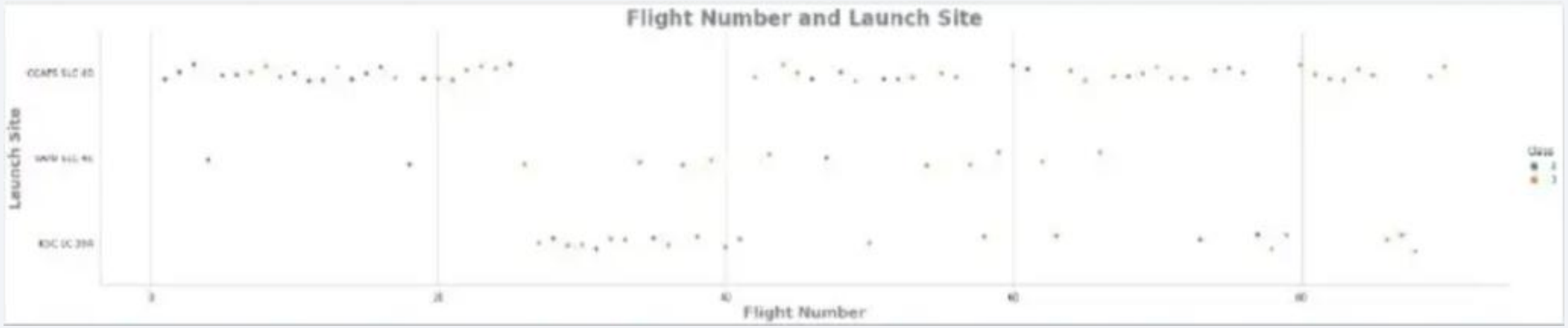


Results

- Exploratory data analysis results
  - Launch success rate increases over time.
  - Higher success rate for higher orbits.
- Interactive analytics demo in screenshots
  - Higher success rate for higher payload mass.
  - Low success rate for booster versions v1.0, v1.1 and high success rate for FT, B4, B5.
  - Higher success rate for Kennedy Space Center and recent starts at Cape Canaveral.
- Predictive analysis results
  - Best prediction results with Logistic regression and Support Vector Machine.


Section 2
Insights drawn from EDA

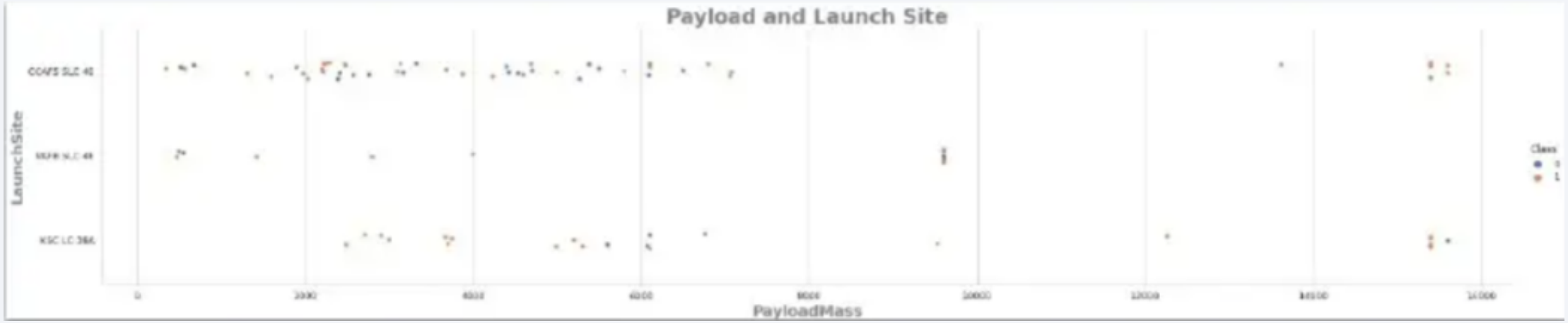
Flight Number vs. Launch Site



Explanation:

- We can see that the CCAFS LC-40 launch site has more attempts than KSC LC-39A and VAFB SLC-4E.

Payload vs. Launch Site



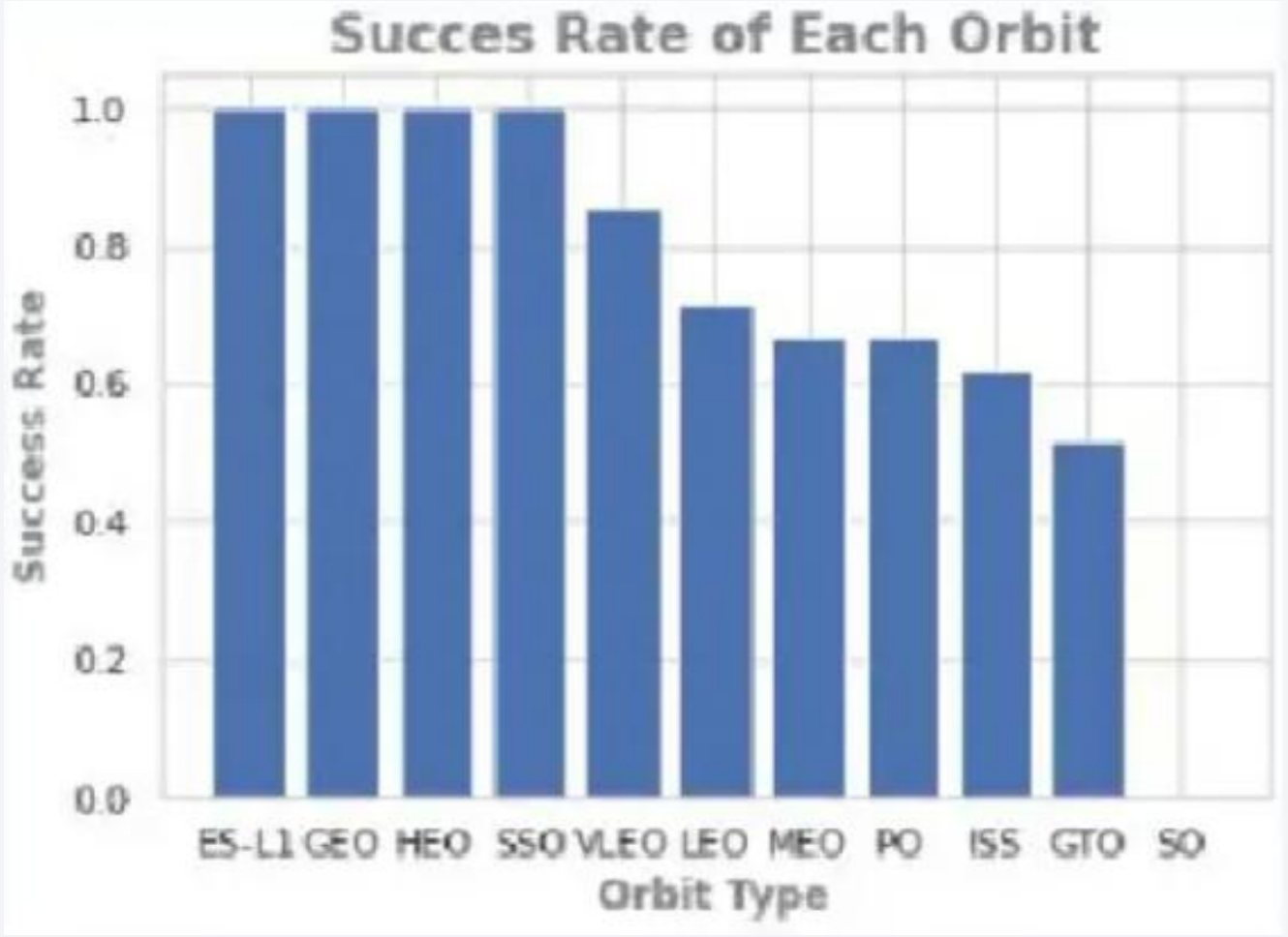
Explanation:

- We can see that for the VAFB SLC launch site, there are no rockets launched for heavy payload mass(greater than 10000).

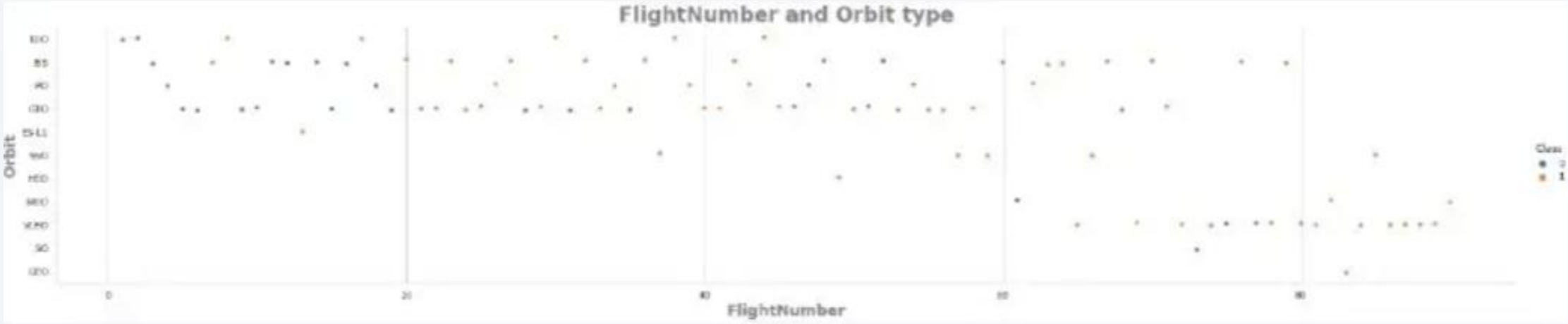
Success Rate vs. Orbit Type

Explanation:

- Low Earth Orbits
  - GTO
  - ISS
  - LEO
  - MEO
  - PO
  - VLEO
- High Earth Orbits
  - ES-L1
  - GEO
  - HEO
  - SSO



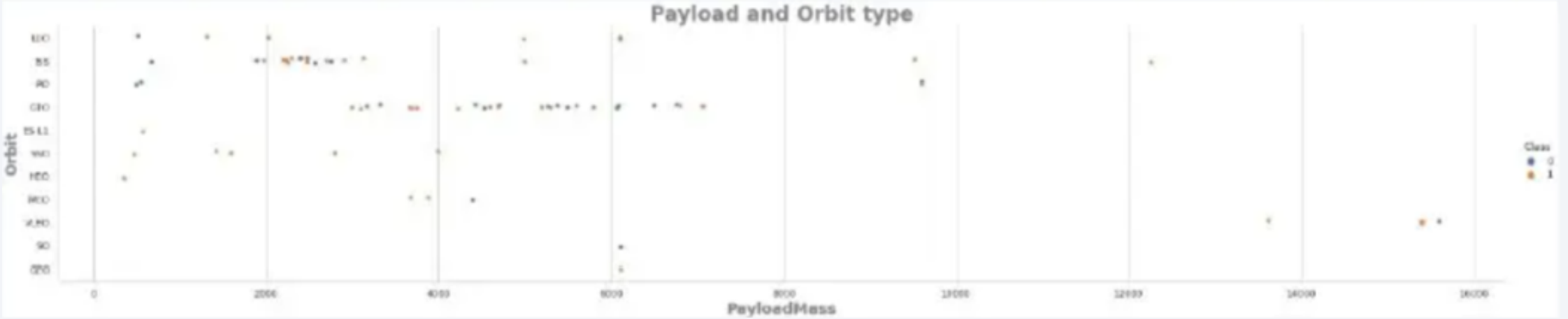
Flight Number vs. Orbit Type



Explanation:

- We can see that in the LEO orbit, the Success appears related to the number of flights.
- On the other hand, there seems to be no relationship between flight number in the GTO orbit.

Payload vs. Orbit Type



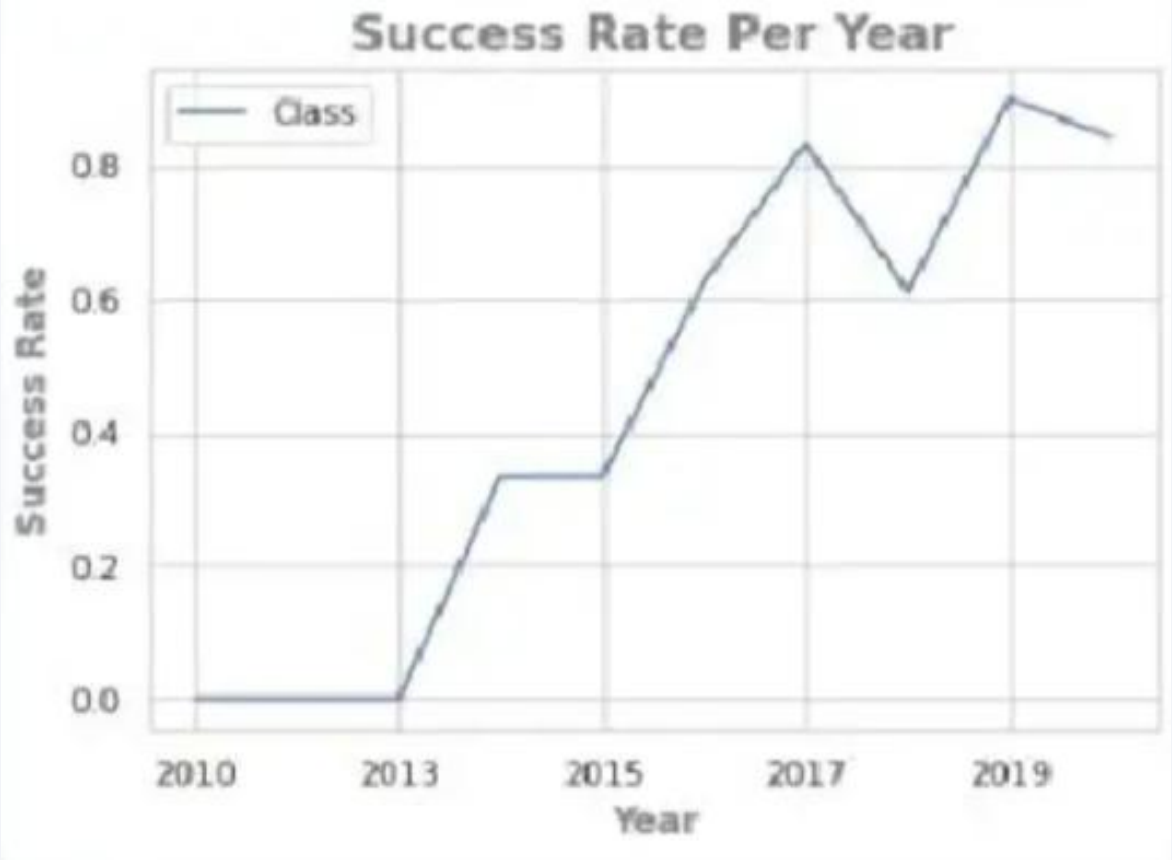
Explanation:

- With heavy payloads, the successful landing rates are more for Polar, LEO and ISS.
- However, for GTO, we cannot distinguish this well as both positive and negative landing rates are present.

Launch Success Yearly Trend

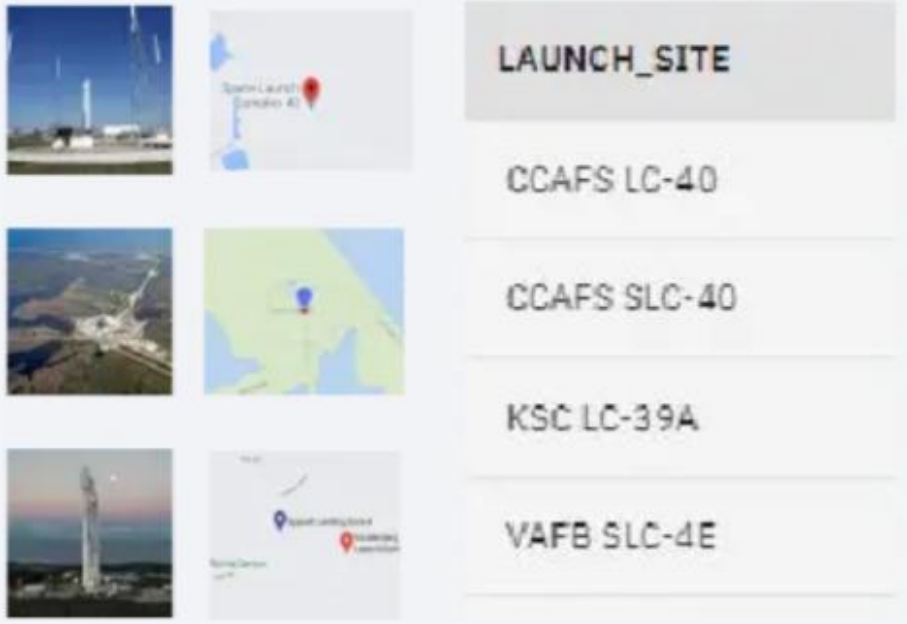
Explanation:

- We can observe that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

- Query:
  - %sql SELECT DISTINCT launch_site FROM SPACEXDATASET;
- The query result presented with the 4 launch sites in the entire dataset:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E



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

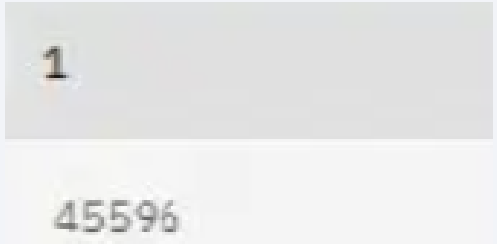
Launch Site Names Begin with 'CCA'

- Query:
  - %sql SELECT * FROM SPACEXDATASET WHERE launch_site like ‘CCA%’ LIMIT 5;

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

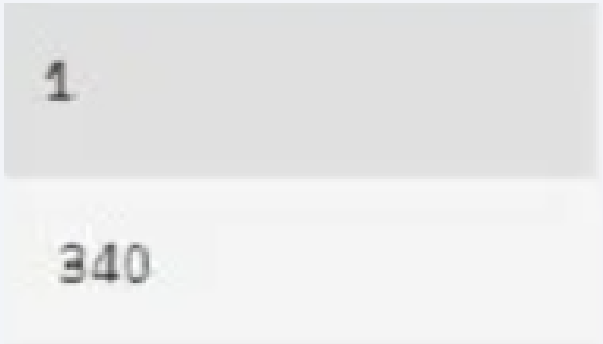
Total Payload Mass – by boosters from NASA

- Query:
  - %sql SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET WHERE customer='NASA (CRS)';



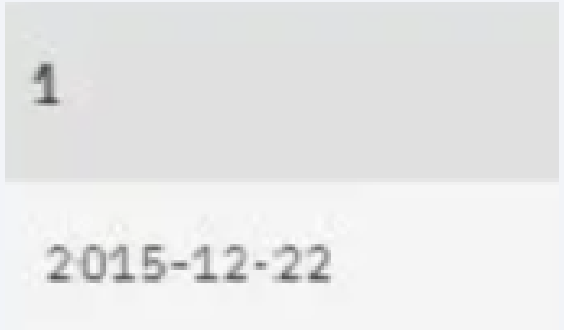
Average Payload Mass by F9 v1.1

- Query:
  - %sql SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET WHERE booster_version like 'F9 v1.0%';



First Successful Ground Landing Date

- Query:
  - %sql SELECT MIN(date) FROM SPACEXDATASET WHERE landing_outcome='Success (ground pad)';



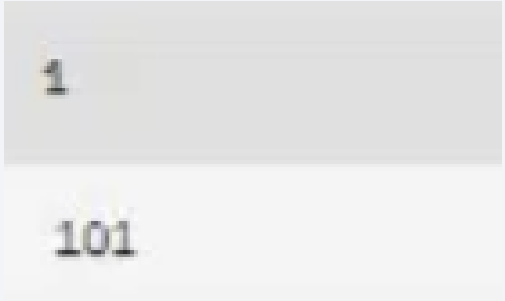
Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:
  - %sql SELECT DISTINCT booster_version FROM SPACEXDATASET WHERE landing_outcome='Success (drone ship)' AND 4000<payload_mass__kg_<6000;

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

Total Number of Successful and Failure Mission Outcomes

- Query:
  - %sql SELECT COUNT(mission_outcome) FROM SPACEXDATASET;



Boosters Carried Maximum Payload

- Query:
  - %sql SELECT DISTINCT booster_version FROM SPACEXDATASET WHERE payload_mass__kg_=(SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET);

BOOSTER_VERSION
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7

2015 Launch Records

- Query:
  - %sql SELECT landing_outcome, booster_version, launch_site FROM SPACEXDATASET WHERE landing_outcome='Failure (drone ship)' AND YEAR(date)=2015;

LANDING_OUTCOME	BOOSTER_VERSION	LAUNCH_SITE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

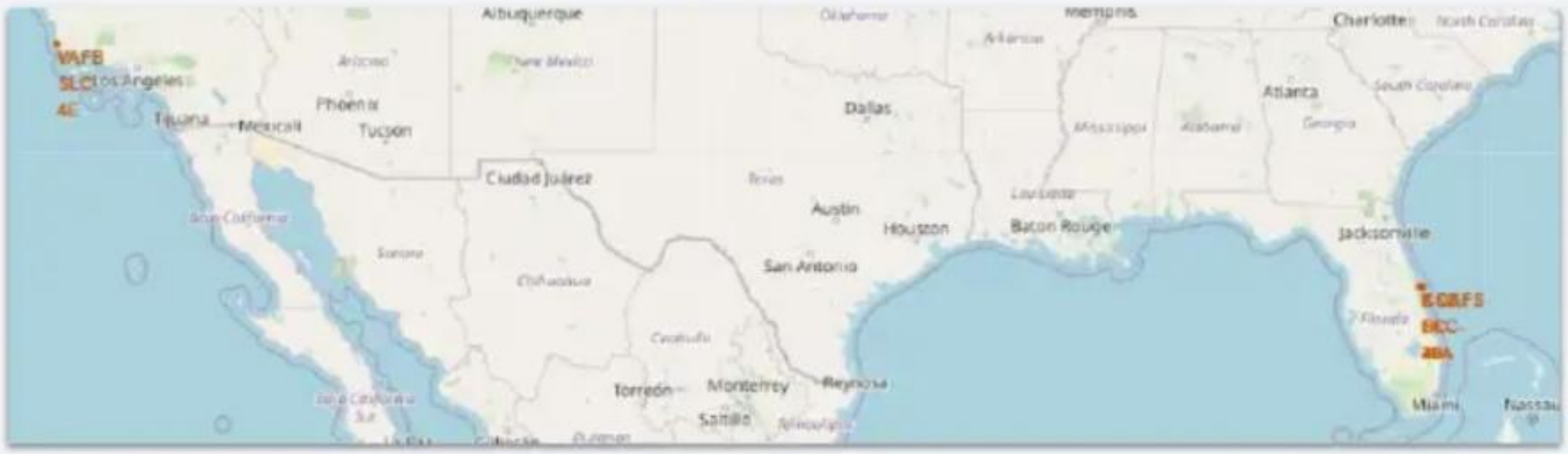
- Query:
  - %sql SELECT landing_outcome, COUNT(landing_outcome) AS total_number FROM SPACEXDATASET WHERE date BETWEEN '2010-06-04' and '2017-03-20' GROUP BY landing_outcome ORDER BY total_number DESC;

LANDING__OUTCOME	TOTAL_NUMBER
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

Folium Map: Launch Sites



- Launch sites are at the East and West coast, near the southernmost U.S.mainland area, ;which is Florida and; California

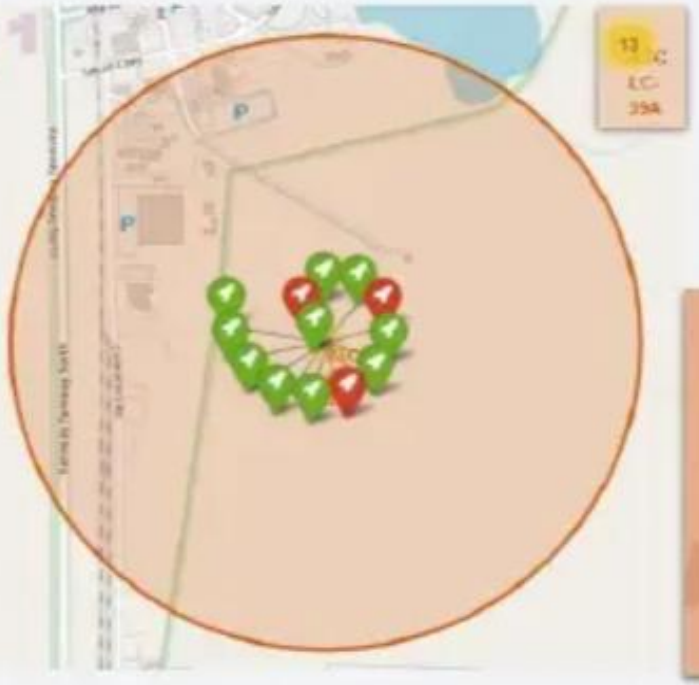
CCAFS Cape Canaveral Space Launch Complex
KSC Kennedy Space Center Launch Complex
VAFB Vandenberg Space Launch Complex

Folium Map: Stage-1 Landing Success by Launch Site

Vandenberg Space Launch Complex

VAFB SLC-4E
40.00% Success

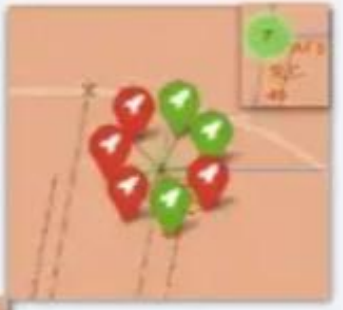
Kennedy Space Center Launch Complex



KSC LC-39A
76.92% Success

Cape Canaveral Space Launch Complex

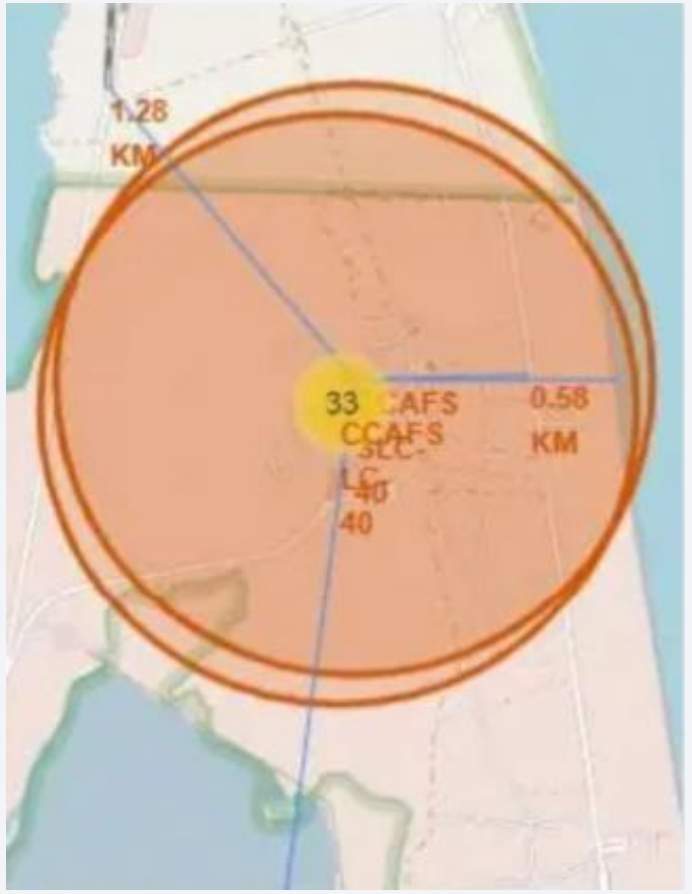
CCAFS SLC-40
42.85% Success

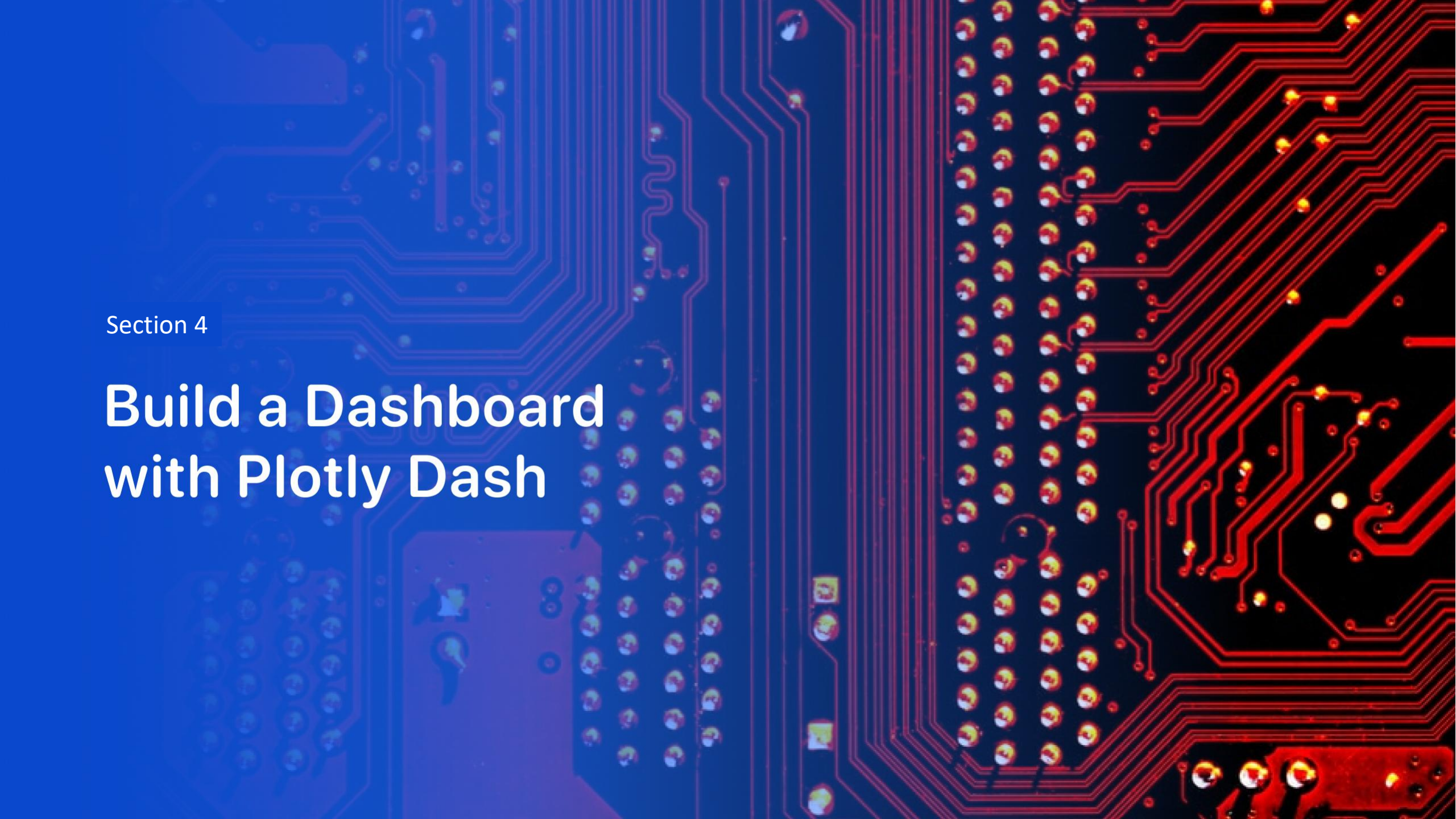


CCAFS LC-40
26.92% Success

Logistics and Safety

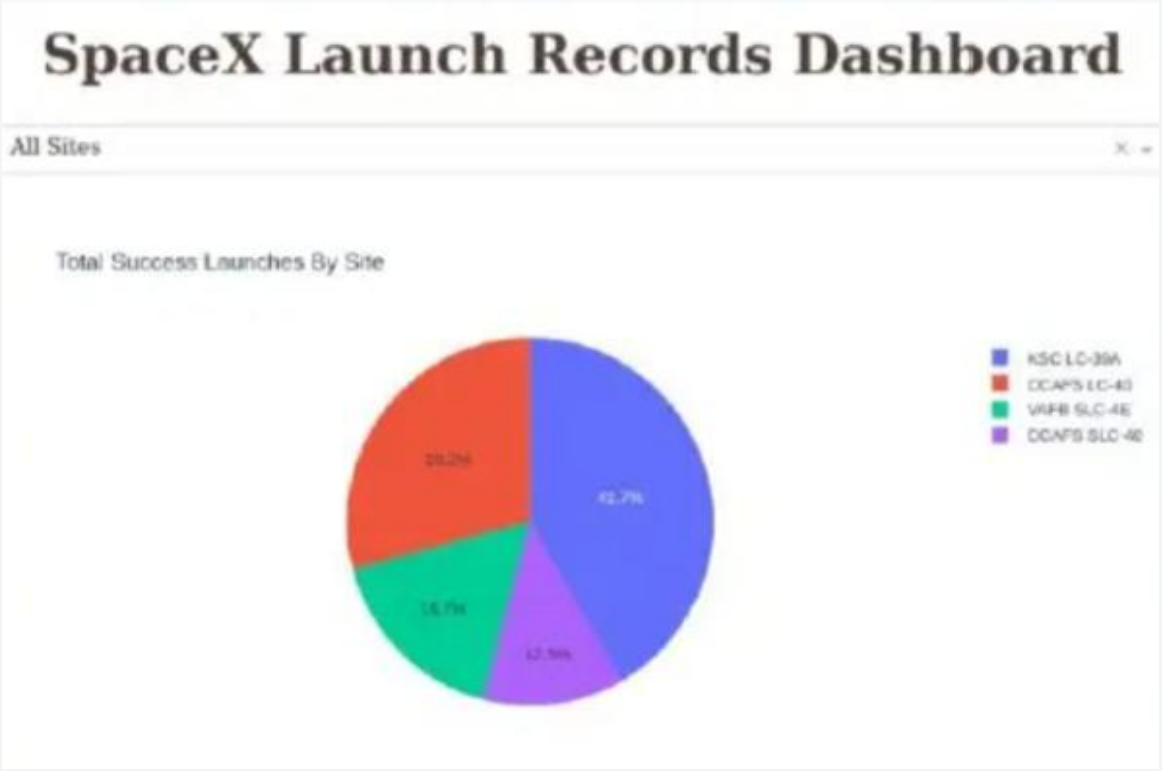
- Launch Site KSC LC-39A has good logistics aspects, being near railroad and relatively far away from inhabited areas.




Section 4
Build a Dashboard with Plotly Dash

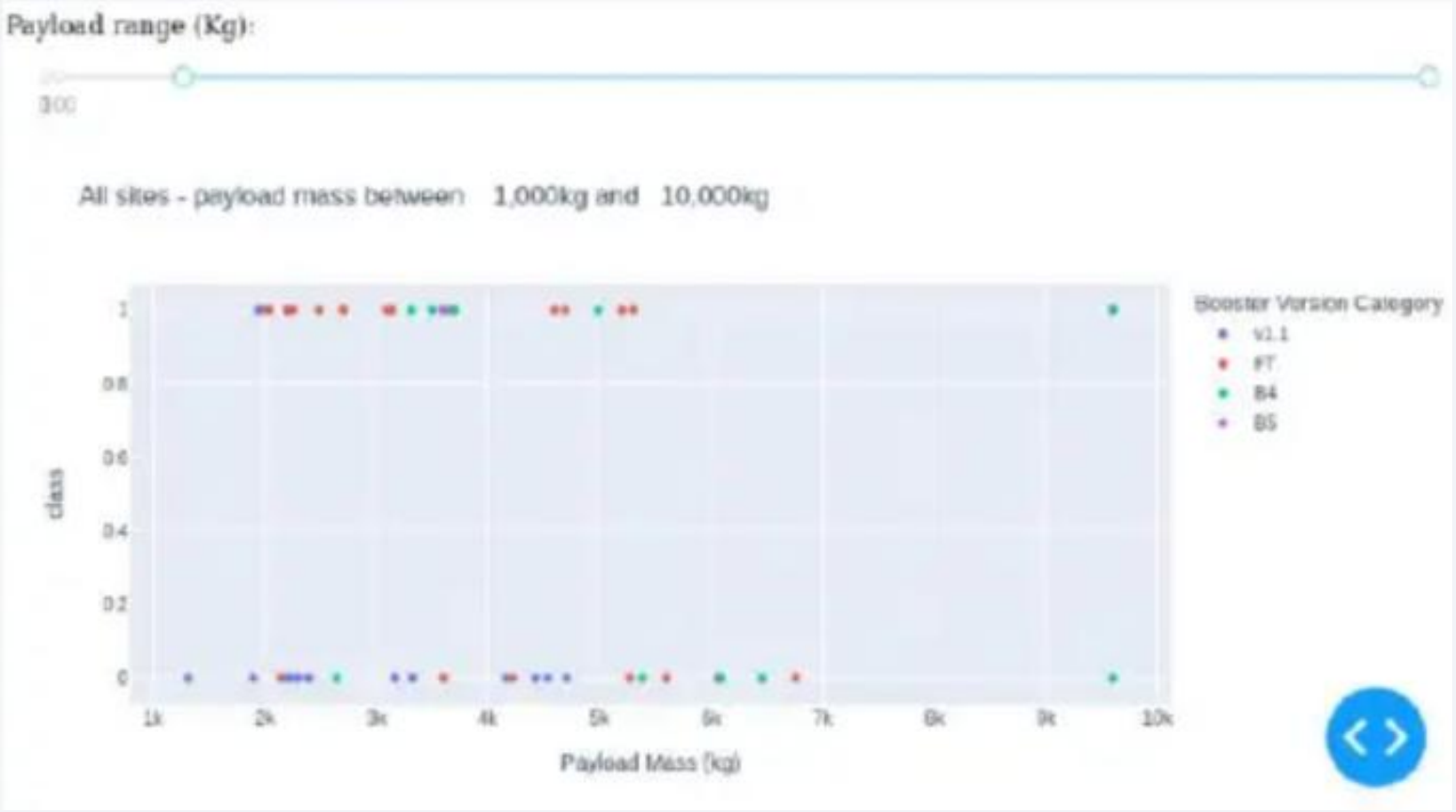
Dashboard: Launch Success Count for all Sites

- KSC LC-39A has the most successful stage-1 landings.
- VAFB SLC-4E has the least number of successful stage-1 landings.



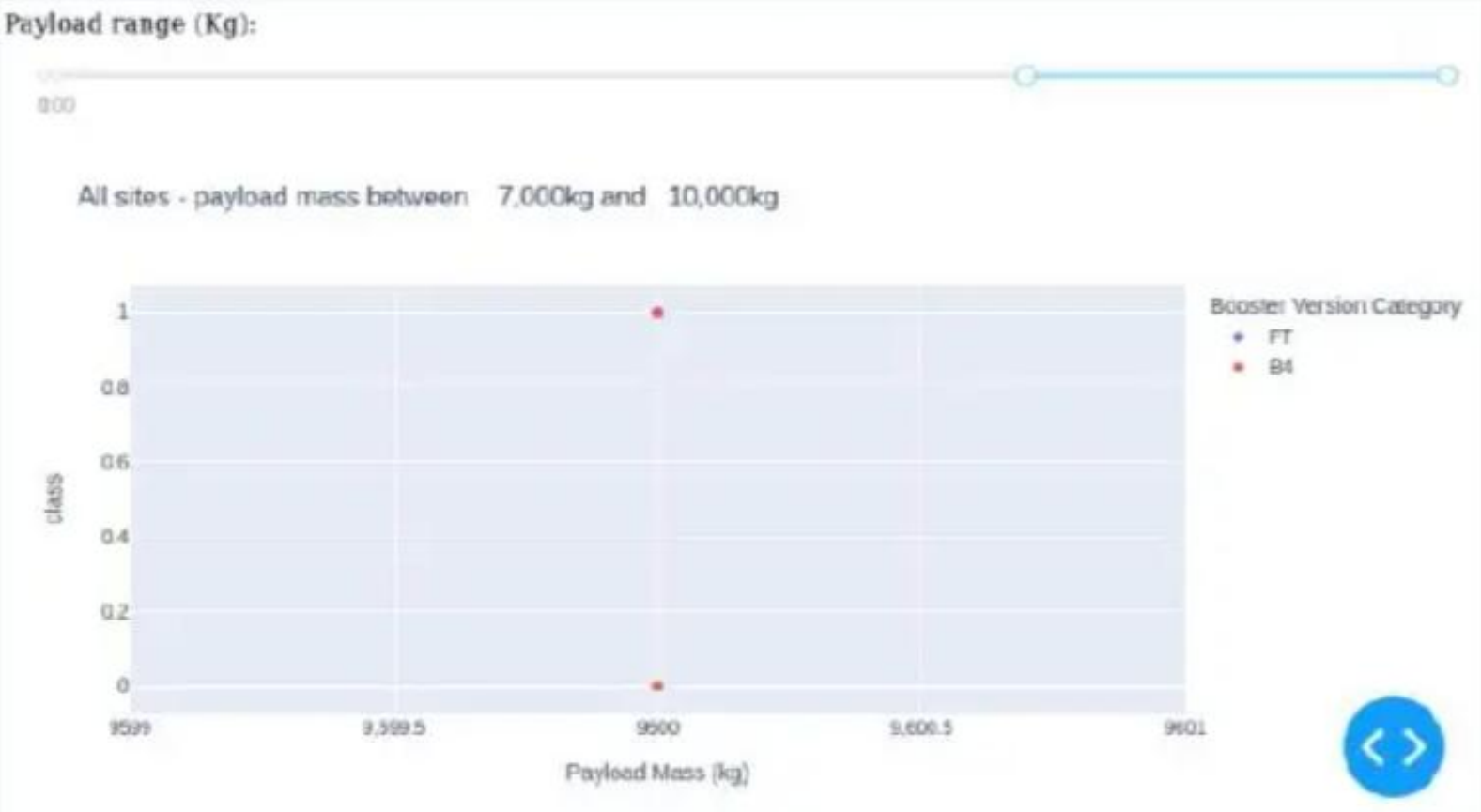
Payload vs. Launch Outcome (under 6000kg)

- Payloads under 6000kg and FT boosters are the most successful combination.



Payload vs. Launch Outcome (7000 to 10000kg)

- There is not enough data to estimate risk of launches over 7000kg.



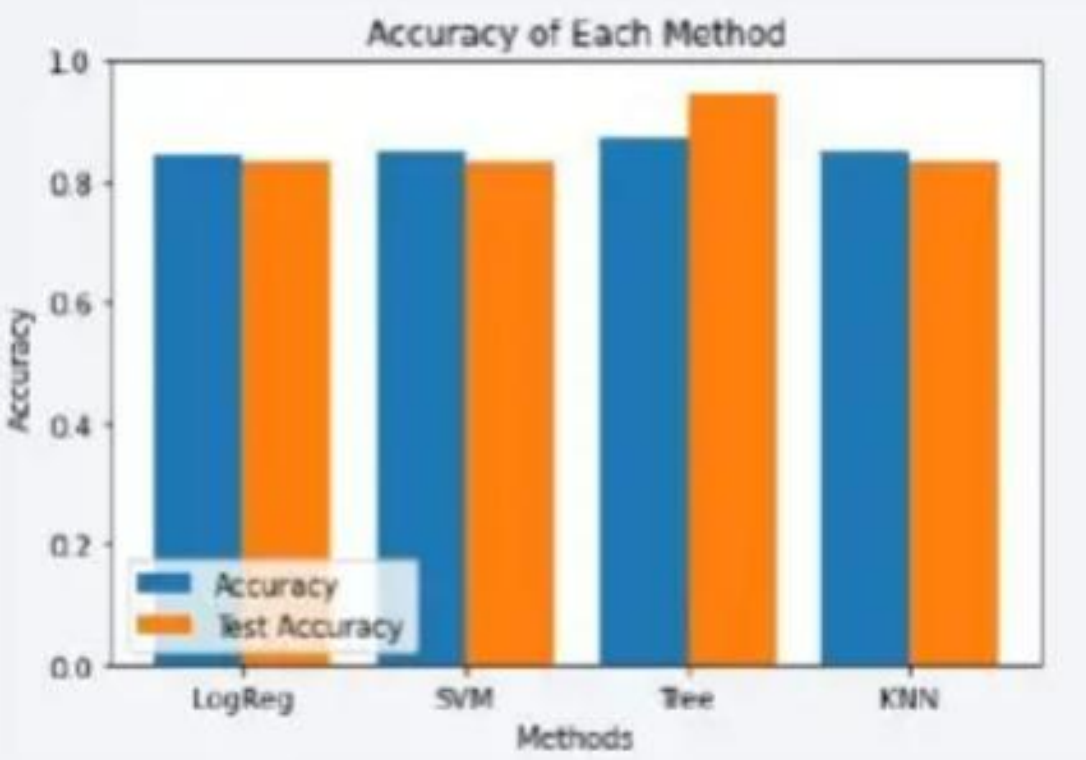
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested and their accuracies are shown in the bar chart beside.

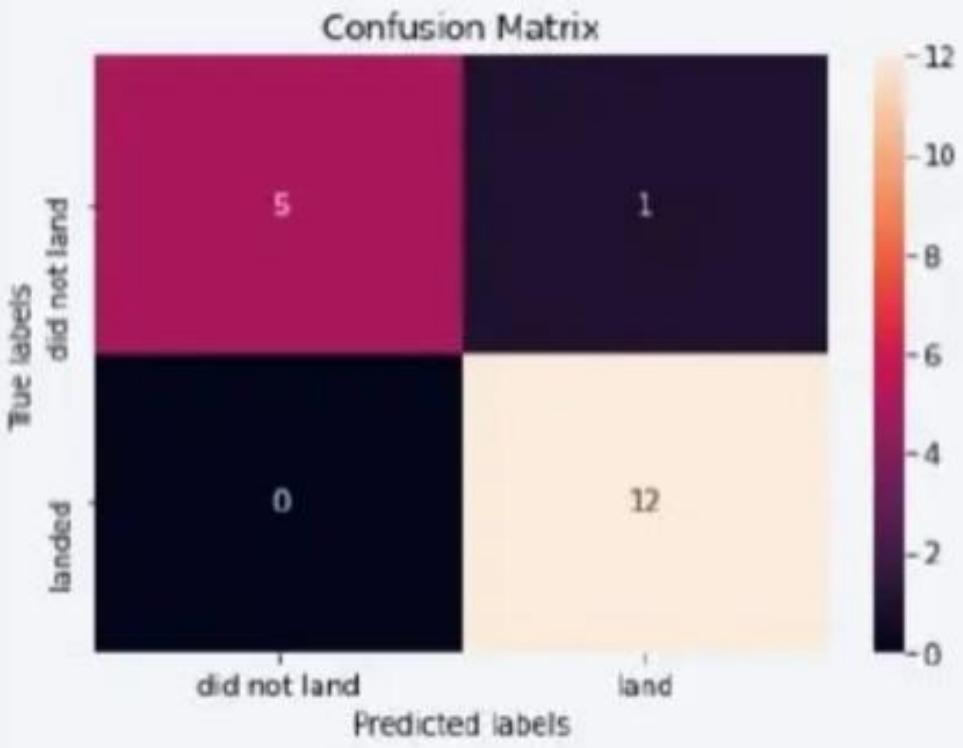
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracy over 87%.



Confusion Matrix of Decision Tree Classifier

- This confusion matrix proves its accuracy by showing the big numbers of true positives and true negatives compared to the false ones.

True Positives	12
True Negatives	5
False Positives	1
False Negatives	0



Conclusions

- The best launch site is KSC LC-39A.
- Launches above 7000kg are less risky.
- Although most of the mission outcomes are successful, successful landing outcomes seem to improve over time, according to the evolution of processes and rockets.
- None of the models had false negatives.
- All models had atleast one false positive.
- Prediction with Logistic Regression is quite accurate.
- Support Vector Machine also provides a good result for predicting the landing outcome.

Appendix

- You can see all the references and details of the project at this link.

Thank you!