IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Arnav Goel
11/26/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - EDA results
  - Interactive analytics
  - Predictive analysis

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Problems you want to find answers
  - The project task is to predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully

Section 1

Methodology

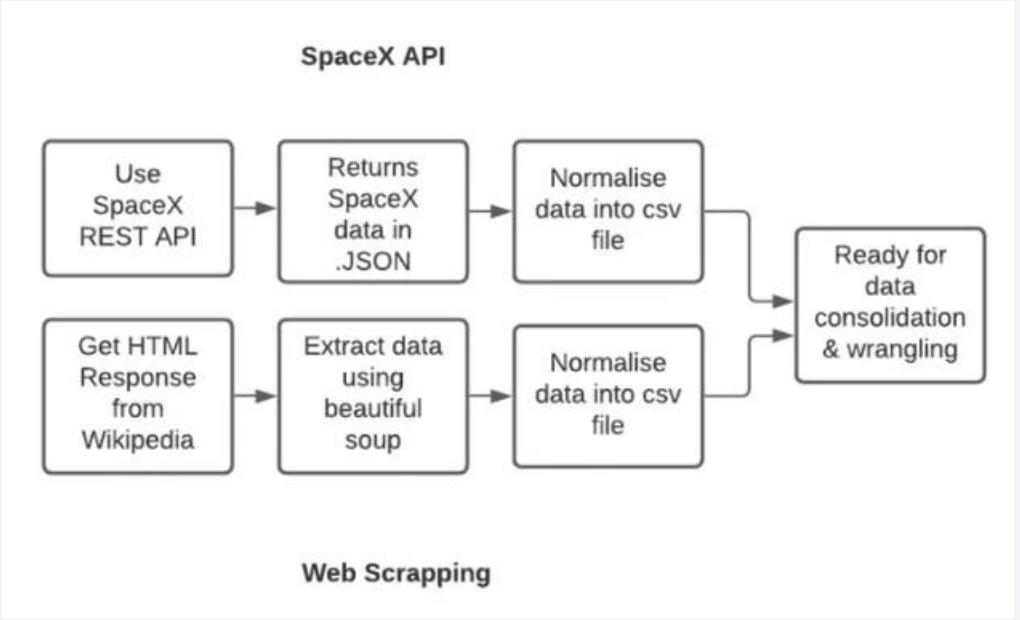
Methodology

Executive Summary

- Data collection methodology:
  - Source: Utilized publicly available SpaceX data, web scrapping Wikipedia.
  - Scope: Collected information on launch dates, launch pads, payloads, and mission outcomes.
- Perform data wrangling
  - Involved handling missing values, identifying numerical and categorical columns, and creating a binary classification variable, 'Class,' representing successful (1) or unsuccessful (0) Falcon 9 first stage landings.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built classification models LR, KNN, SVM, DT to predict launch success.

Data Collection

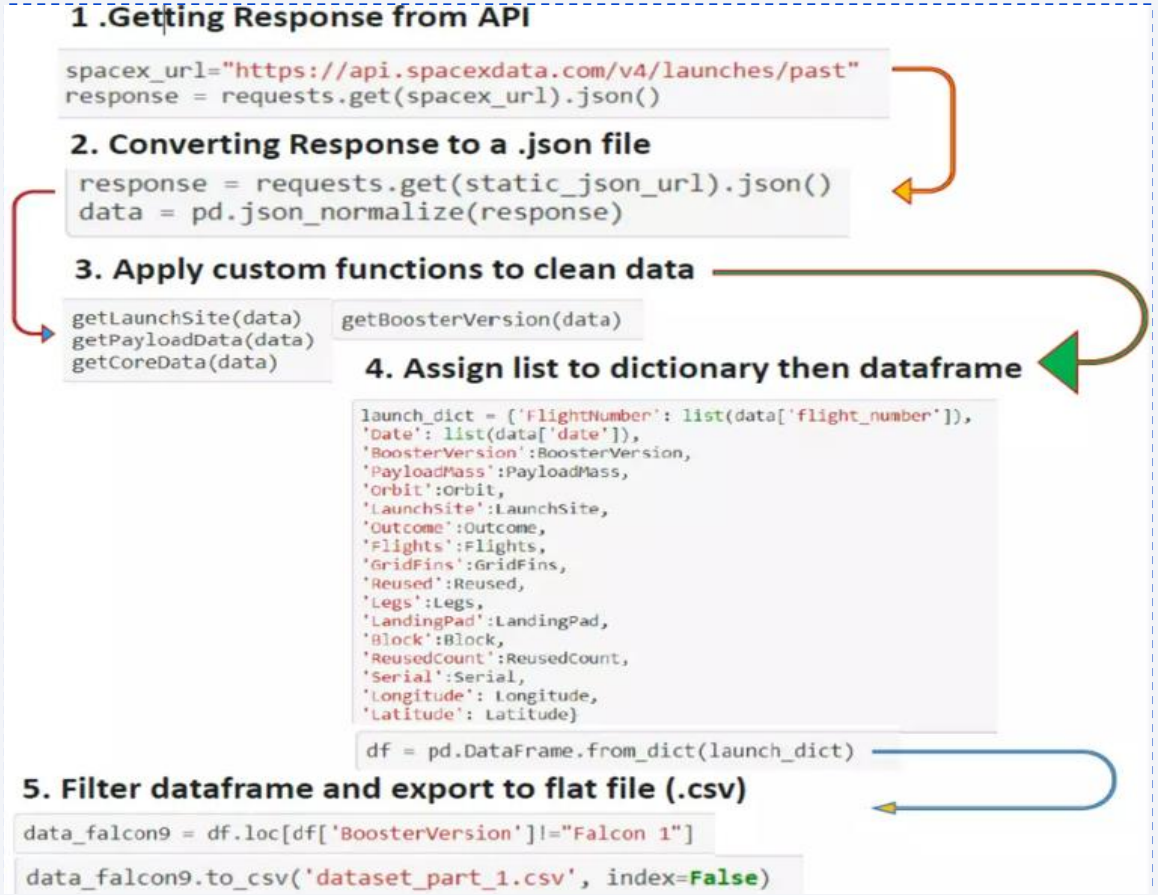
- Web Scraping Falcon 9 Launch Records:
  - Utilized BeautifulSoup for web scraping.
  - Extracted Falcon 9 launch records from Wikipedia
- Request to SpaceX API:
  - Initiated a GET request to the SpaceX API.
  - Retrieved additional launch data.
- Data Cleaning:
  - Conducted basic data wrangling on the retrieved data.
  - Addressed inconsistencies and missing values.
- Integration and Quality Assurance:
  - Combined data from web scraping and SpaceX API.
  - Ensured data consistency and accuracy.



Data Collection – SpaceX API

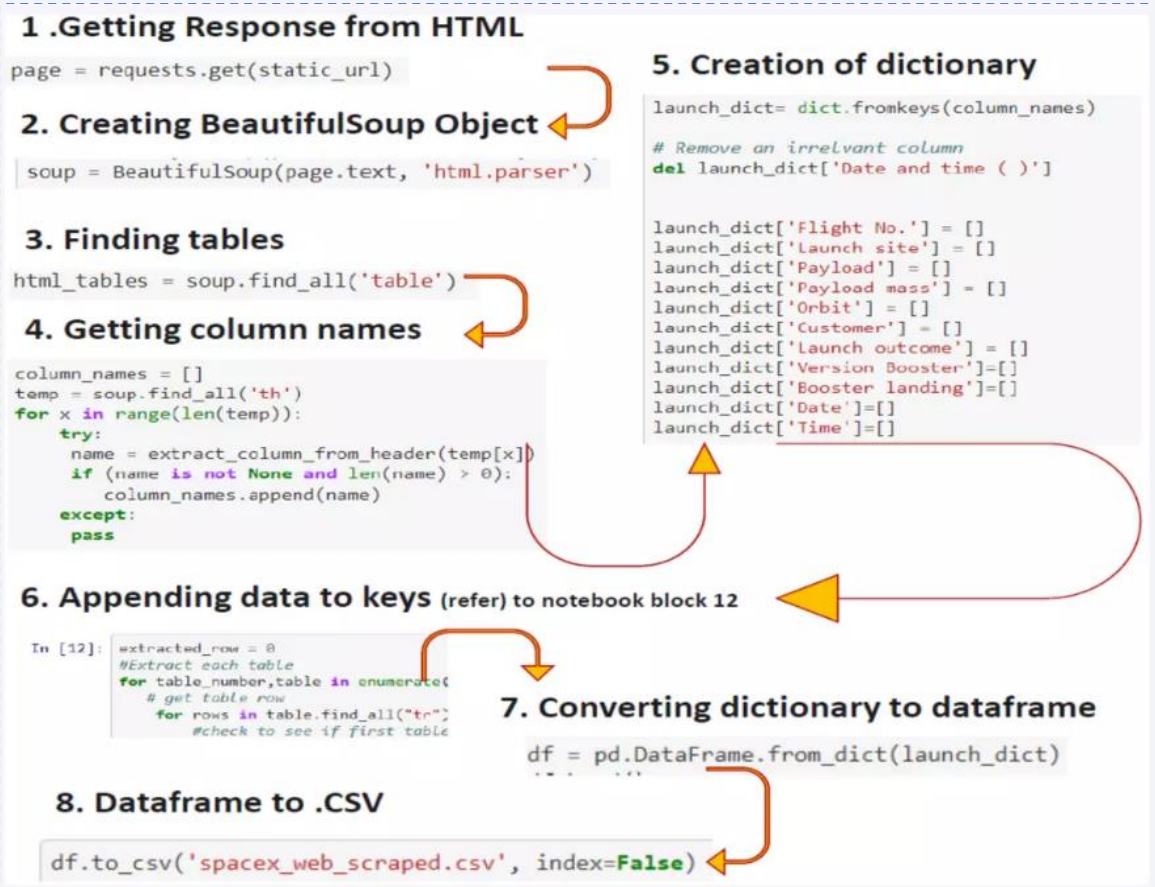
- Data collection done using SpaceX Api

- https://github.com/ArnavGoel1/SpaceX/blob/5b813b5146a8f4a9093fb21d4af63aa953bf9116/jupyter-labs-spacex-data-collection-api%20(1).ipynb



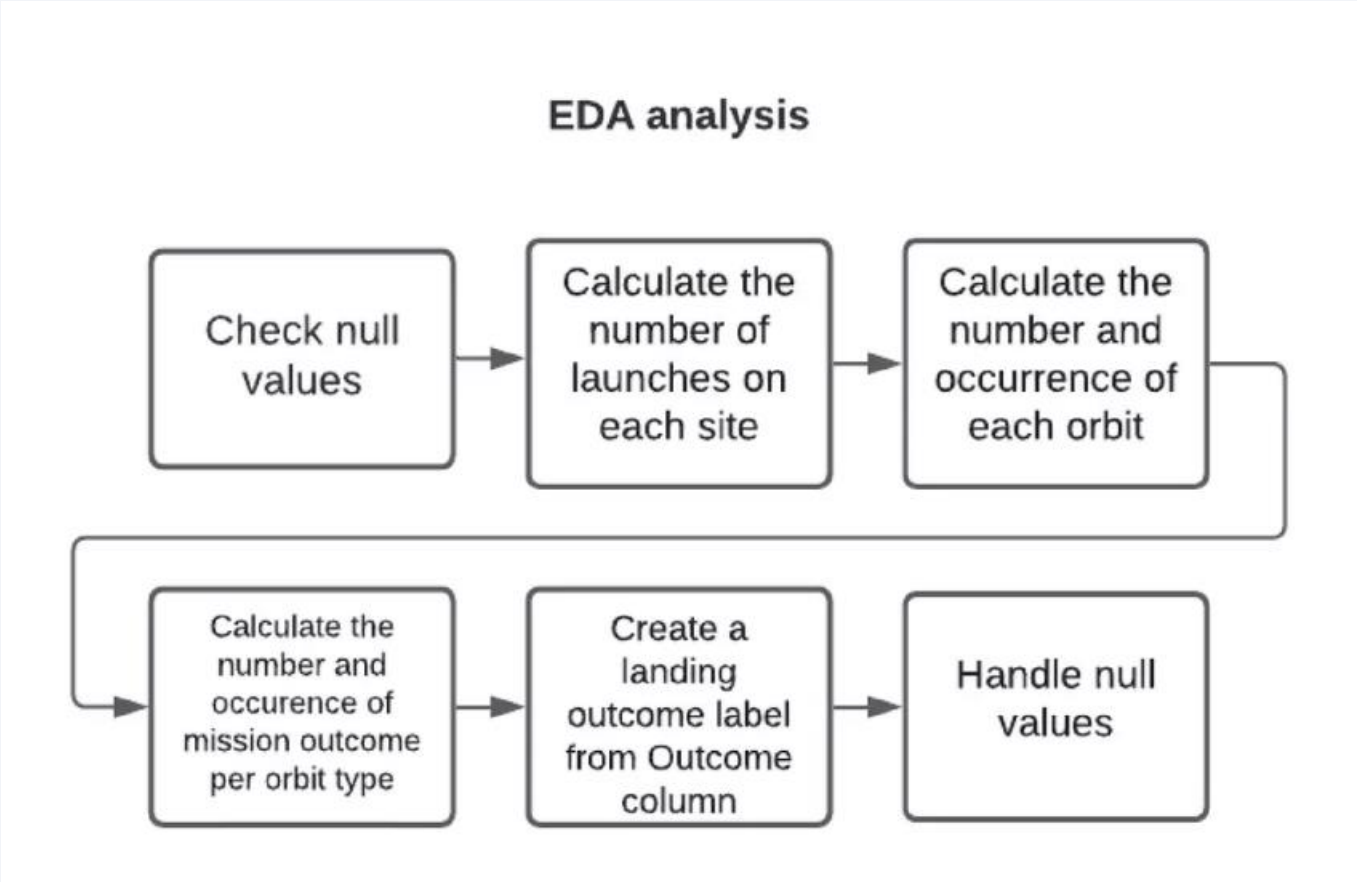
Data Collection - Scraping

- Web scraping process done from Wikipedia
- https://github.com/ArnavGoel1/SpaceX/blob/5b813b5146a8f4a9093fb21d4af63aa953bf9116/jupyter-labs-webscraping.ipynb

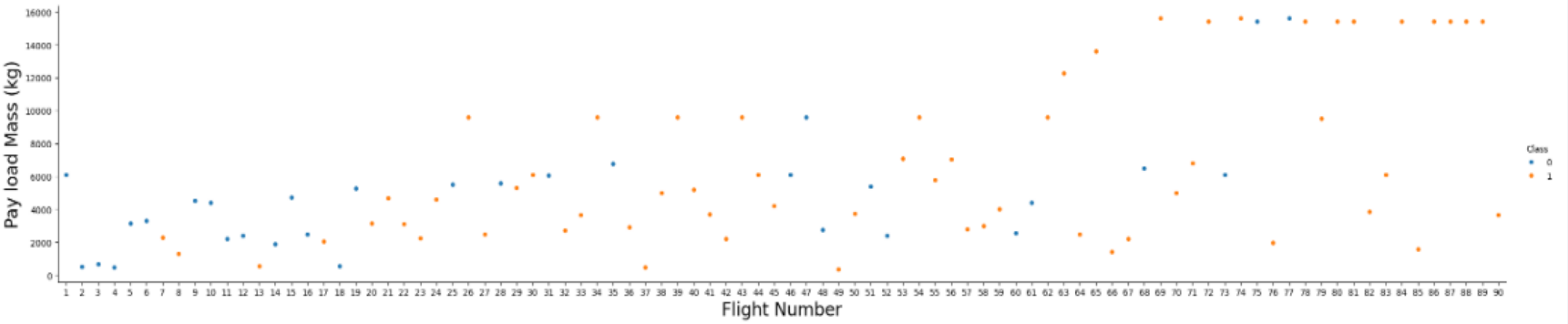


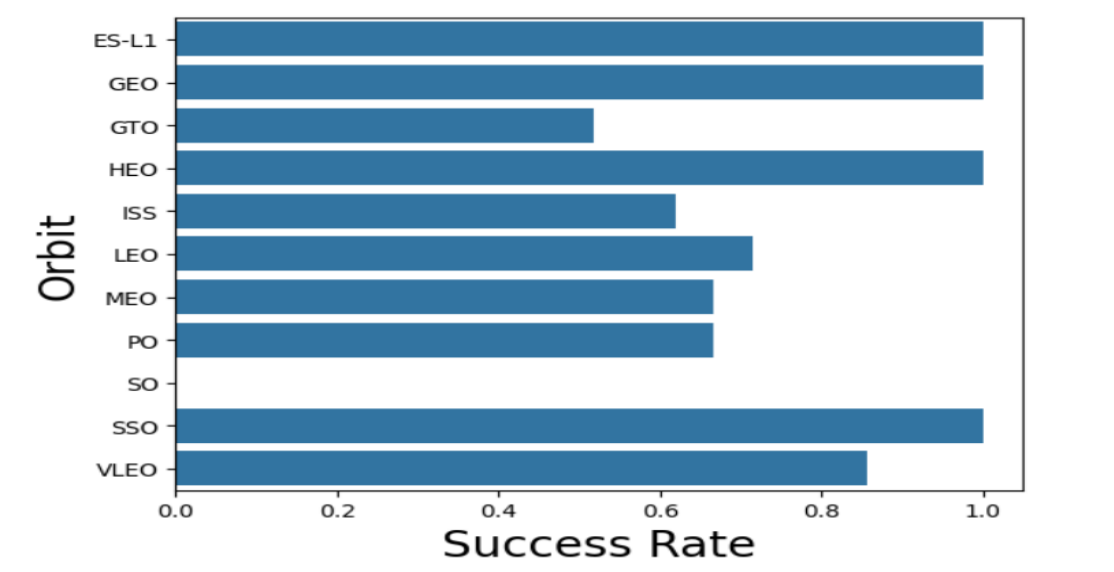
Data Wrangling

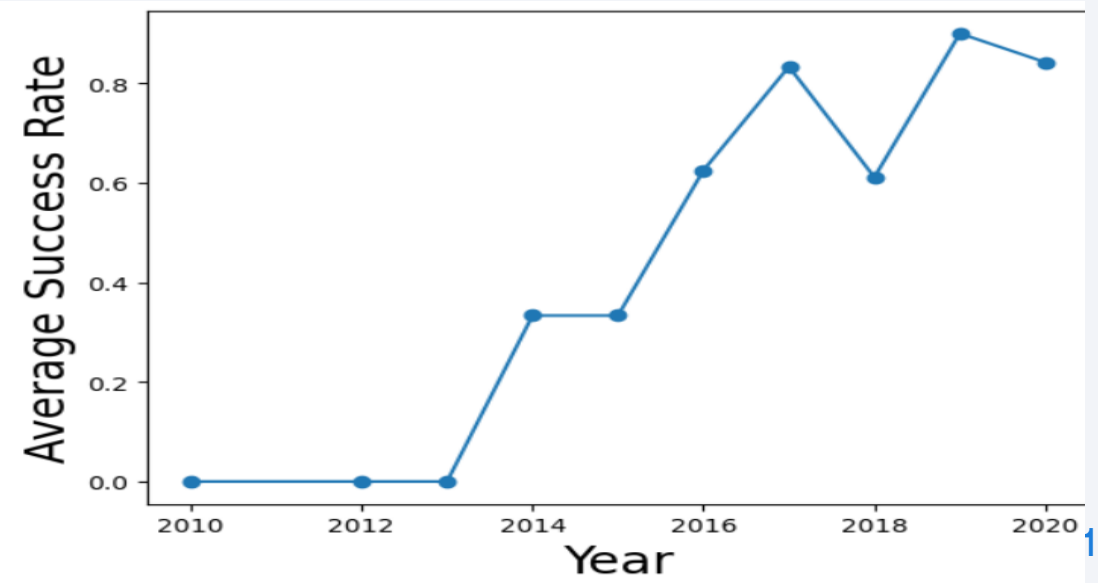
- Cleaning Process: Rigorous cleaning ensured data reliability.
- Tasks: Addressed missing data, removed duplicates, and corrected inconsistencies.
- https://github.com/ArnavGoel1/SpaceX/blob/5b813b5146a8f4a9093fb21d4af63aa953bf9116/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization



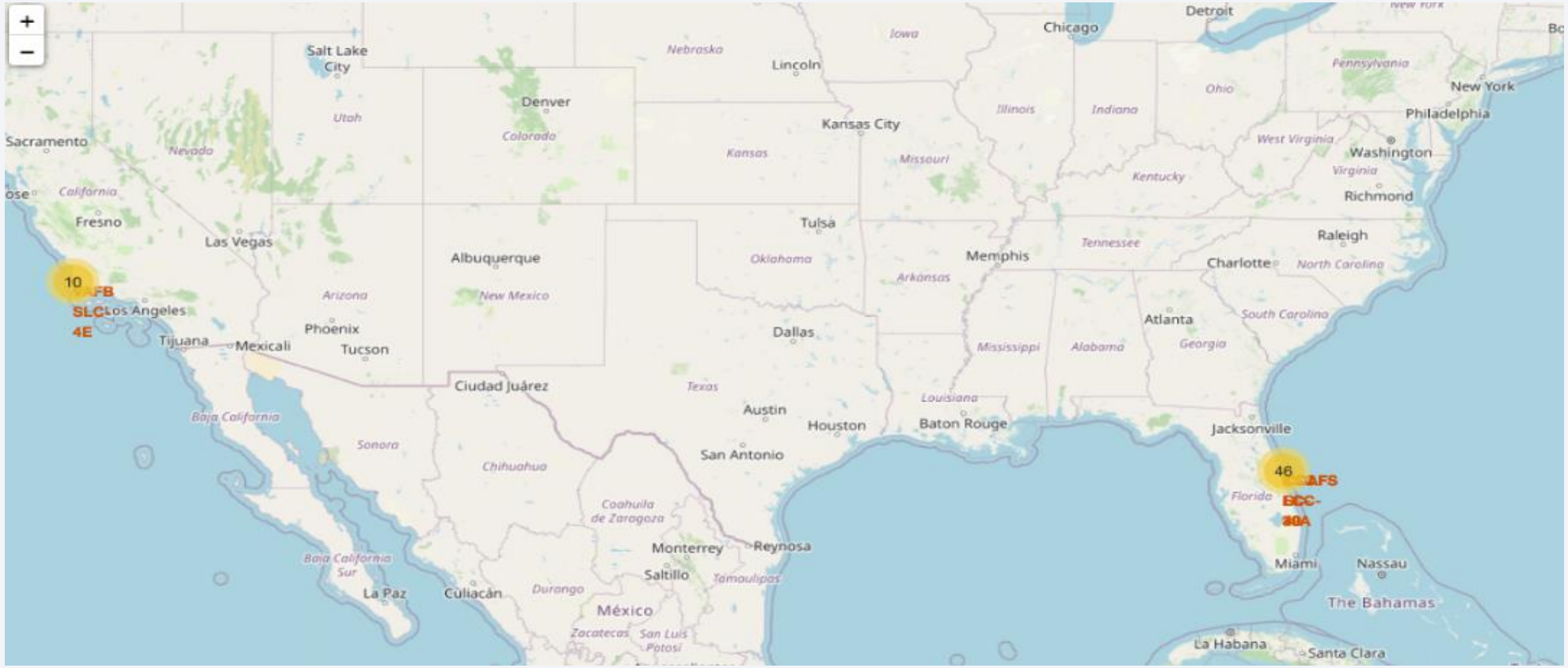




EDA with SQL

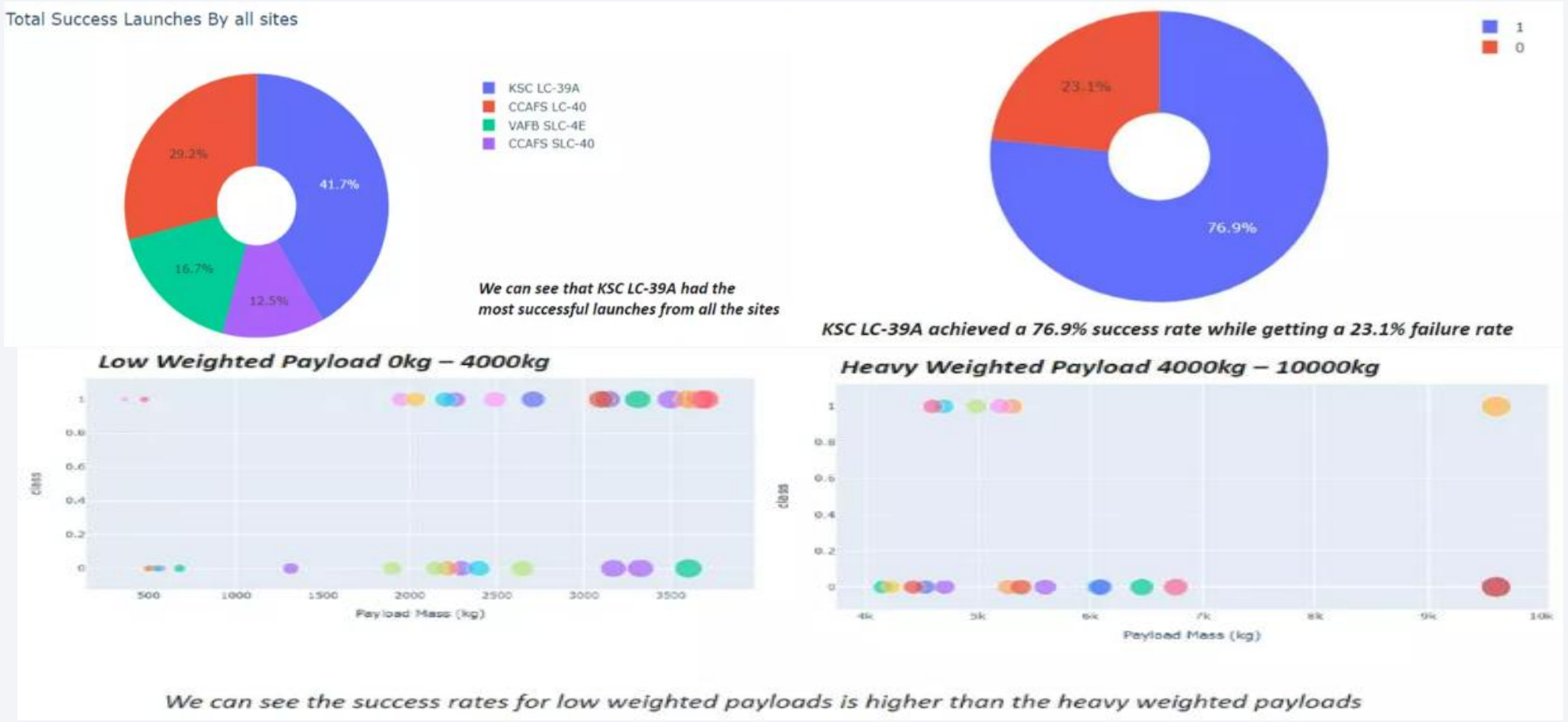
- SQL queries performed include:
  - Displaying the names of the unique launch sites in the space mission.
  - Displaying 5 records where launch sites begin with the string 'KSC'.
  - Displaying the total payload mass carried by boosters launched by NASA (CRS).
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date where the successful landing outcome in drone ship was achieved.
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000.
  - Listing the total number of successful and failure mission outcomes.
  - Listing the names of the booster versions which have carried the maximum payload mass.
  - Listing the records which will display the month names, successful landing outcomes in ground pad, booster versions, launch site for the months in year 2017.
  - Ranking the count of successful landing outcomes between the date 2010 06 04 and 2017 03 20 in descending order.
- https://github.com/ArnavGoel1/SpaceX/blob/5b813b5146a8f4a9093fb21d4af63aa953bf9116/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium



https://github.com/ArnavGoel1/SpaceX/blob/5b813b5146a8f4a9093fb21d4af63aa953bf9116/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

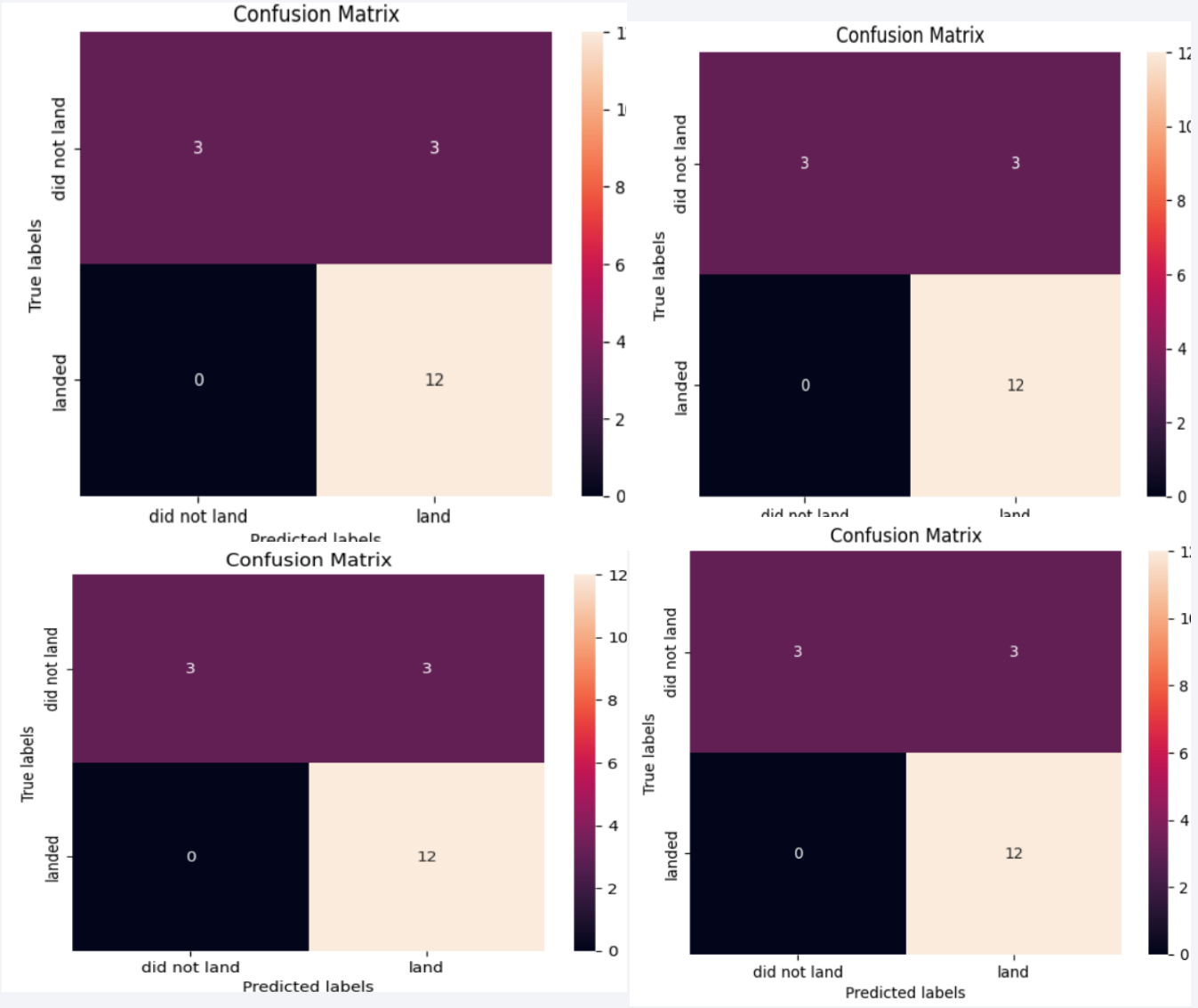
Build a Dashboard with Plotly Dash



https://github.com/ArnavGoel1/SpaceX/blob/5b813b5146a8f4a9093fb21d4af63aa953bf9116/spacex_dash_app.py

Predictive Analysis (Classification)

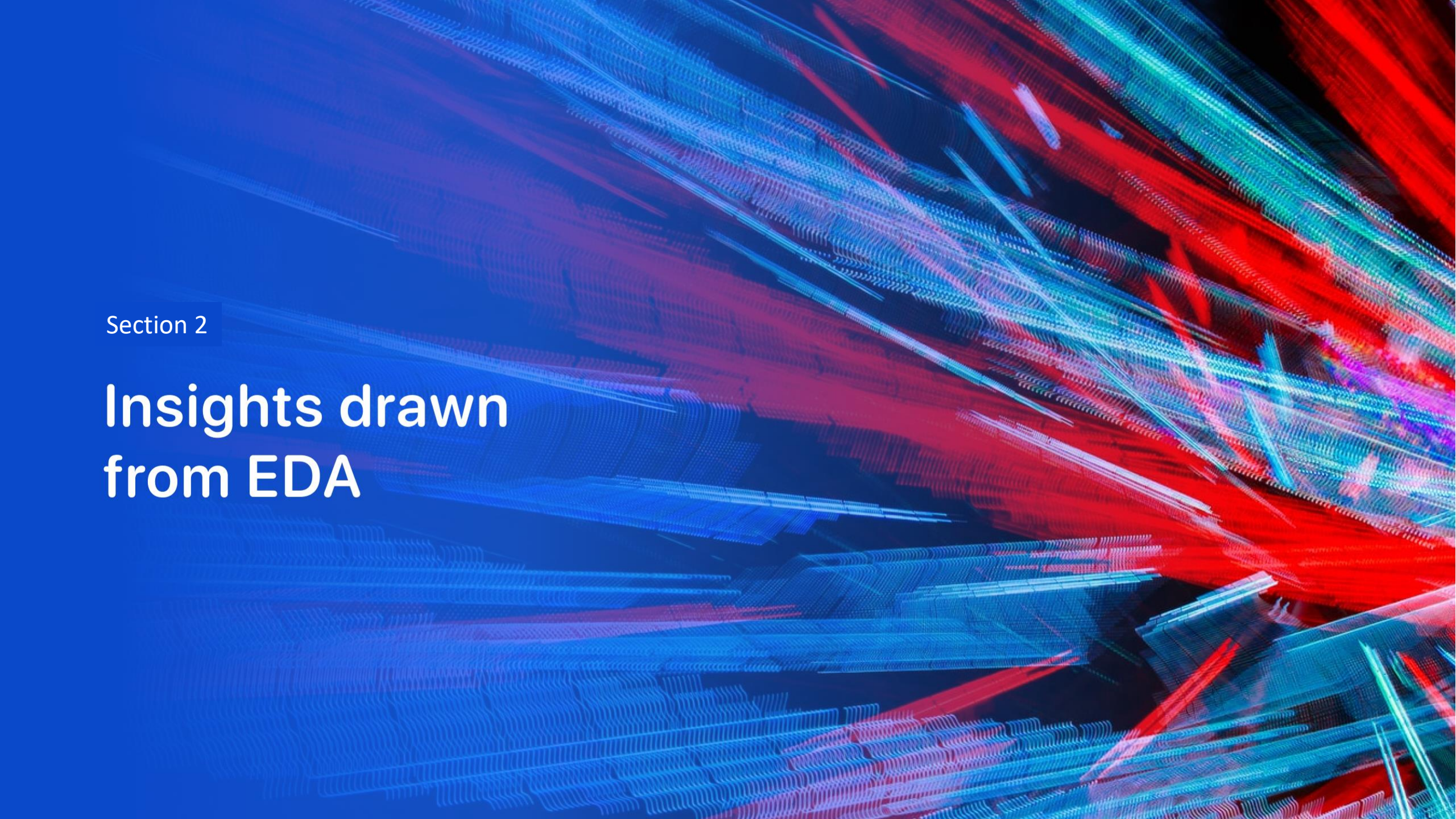
- The SVM, KNN and Logistic Regression model achieved the accuracy of 83.3%.



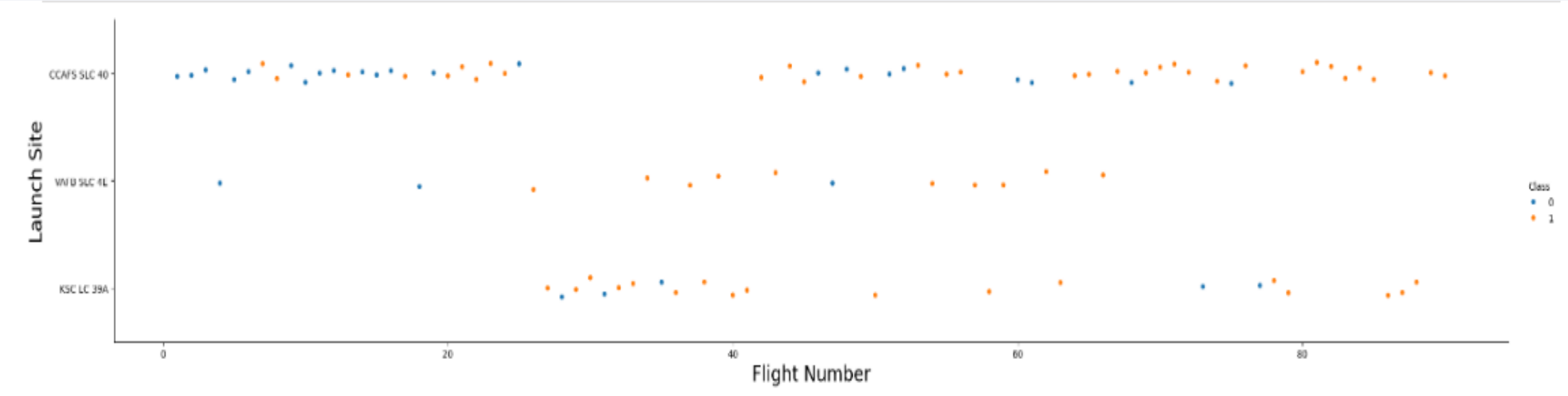
https://github.com/ArnavGoel1/SpaceX/blob/5b813b5146a8f4a9093fb21d4af63aa953bf9116/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO,HEO,SSO, ES L1 has the best Success Rate.

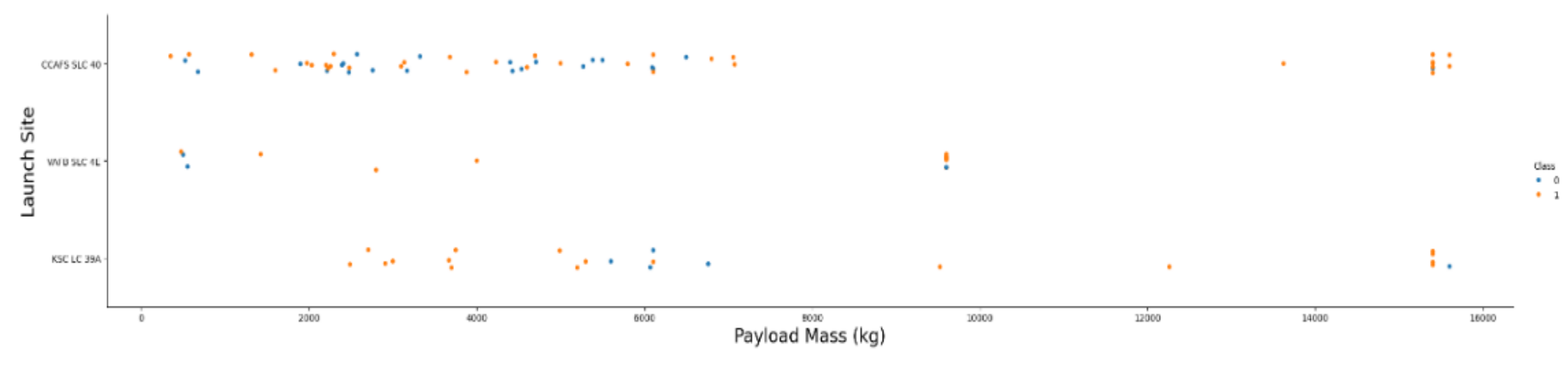

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Launches from the site of CCAFS SLC 40 are significantly higher than the launches from other sites.

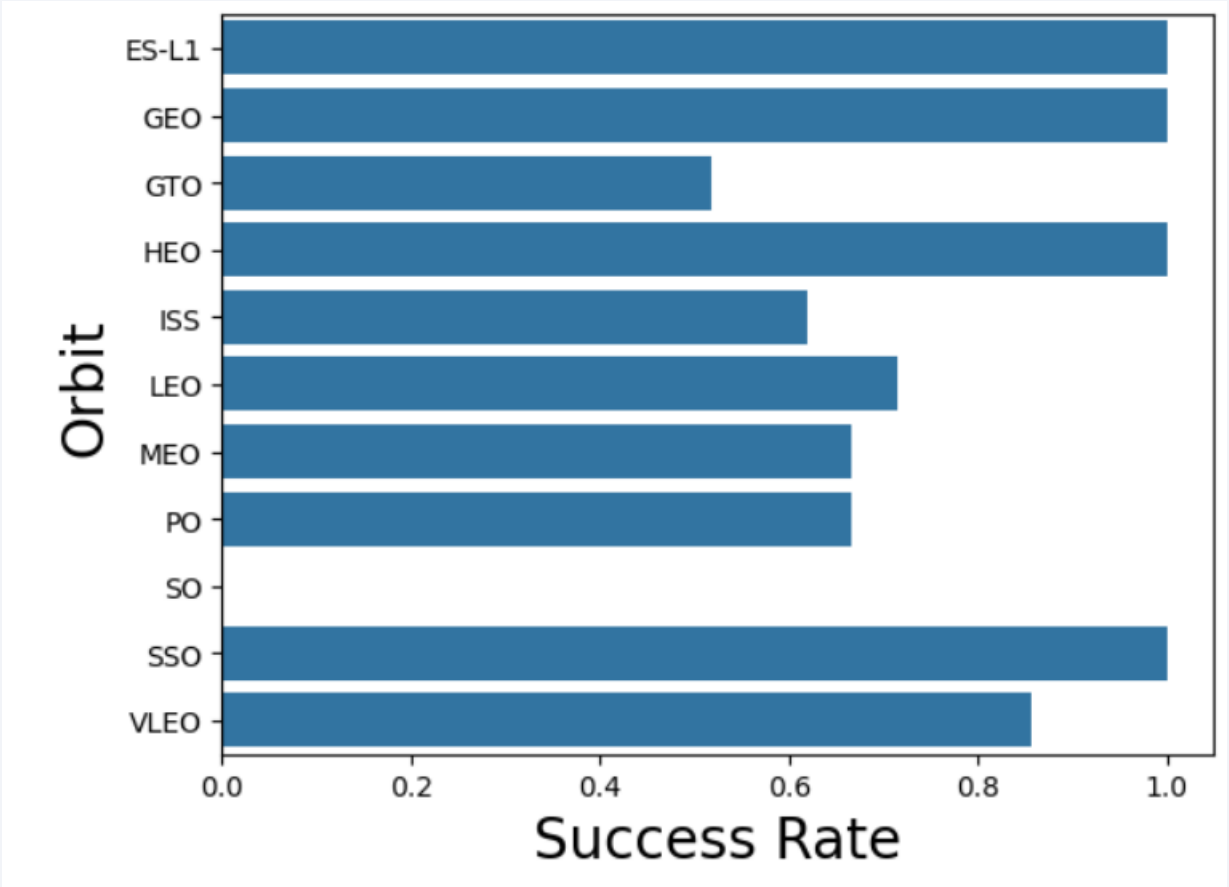
Payload vs. Launch Site



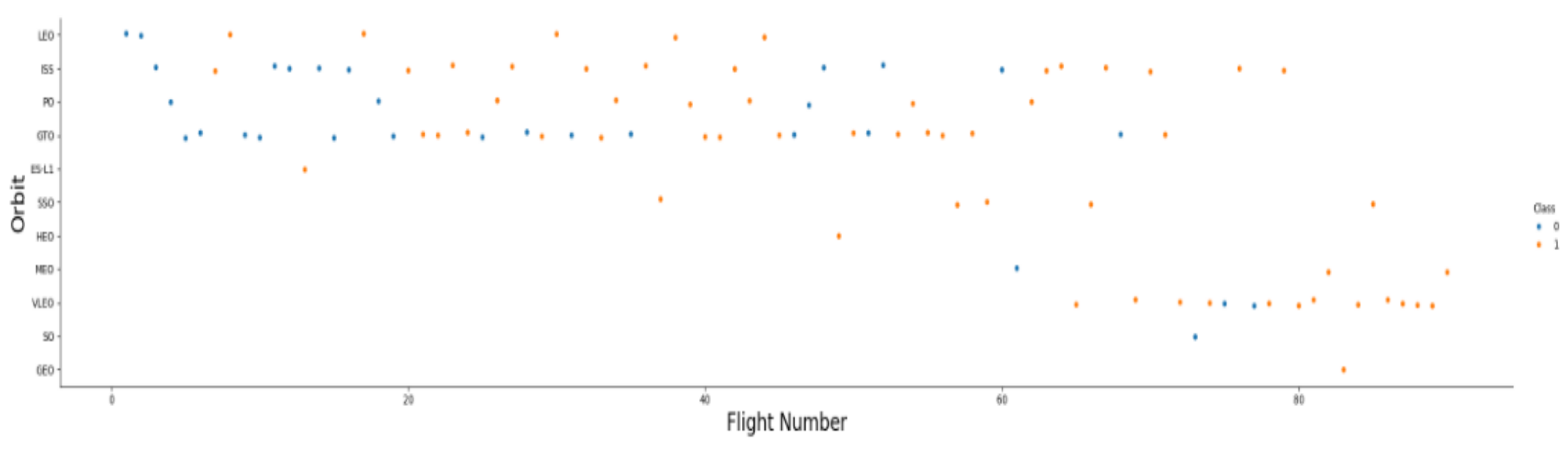
- The majority of IPay Loads with lower mass have been launched from CCAFS SLC 40.

Success Rate vs. Orbit Type

- The orbit types of ES-L1, GEO, HEO and SSO are among the highest success rate.

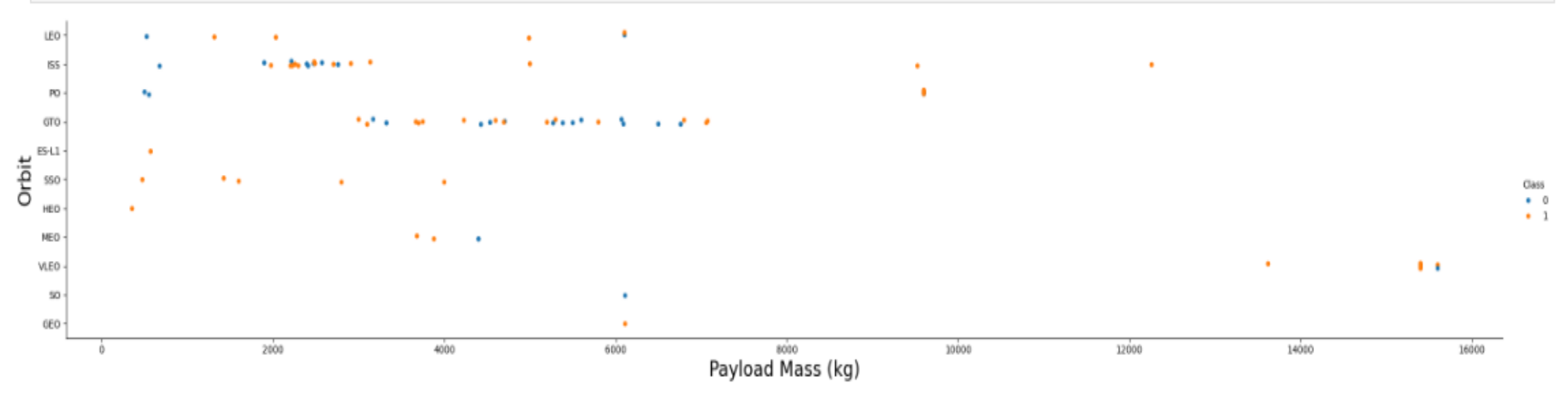


Flight Number vs. Orbit Type



- A trend can be observer of shifting to VLEO launches in recent years.

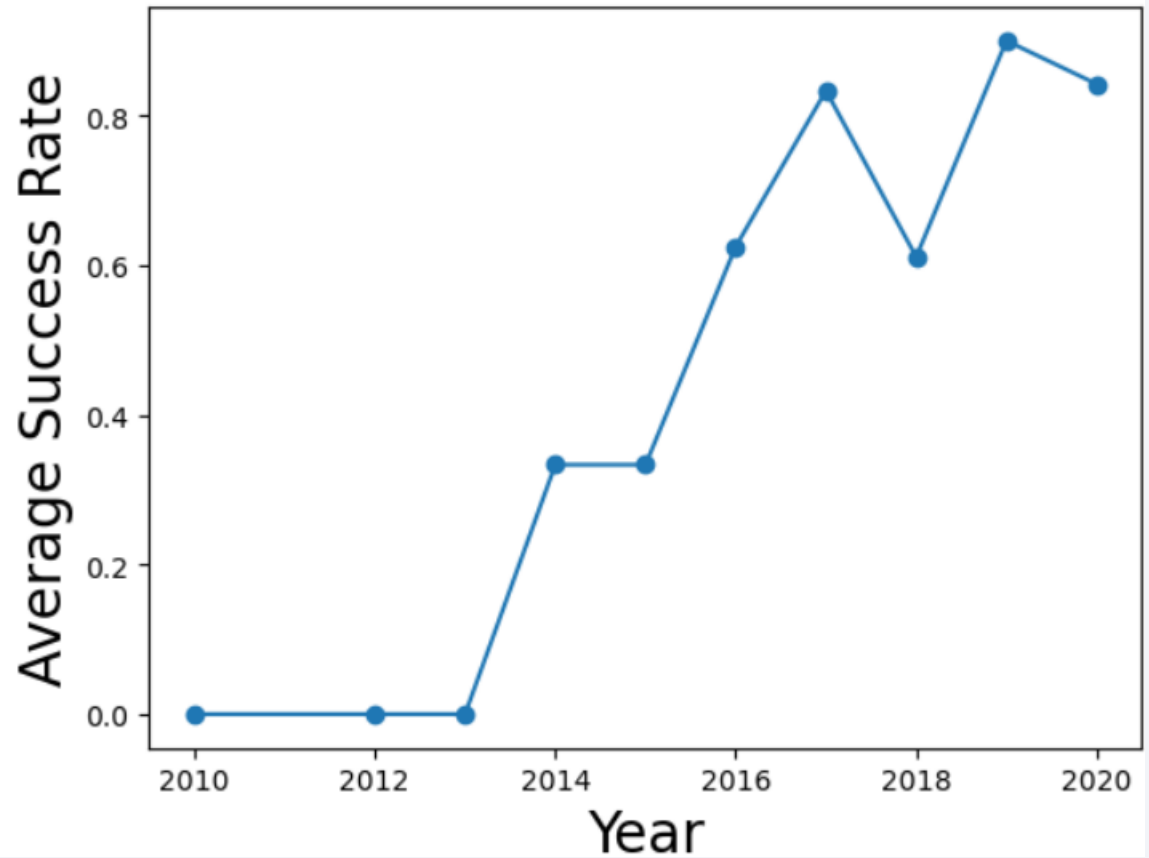
Payload vs. Orbit Type



- There are strong correlation between ISS and Payload at the range around 2000, as well as between GTO and the range 4000-8000.

Launch Success Yearly Trend

- Launch success rate has increased significantly since 2013 and has stabilized since 2019.



All Launch Site Names

- %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;

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

Launch Site Names Begin with 'CCA'

- %sql SELECT * FROM SPACEXTABLE 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	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

- %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- %sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (ground pad)%';

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)%' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 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

- %sql SELECT "Mission_Outcome", COUNT(*) FROM SPACEXTABLE GROUP BY "Mission_Outcome";

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);

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

2015 Launch Records

- %sql SELECT strftime('%Y-%m', "Date") AS "Month","Landing_Outcome","Booster_Version","Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Failure (drone ship)%' AND strftime('%Y', "Date") = '2015';

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

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

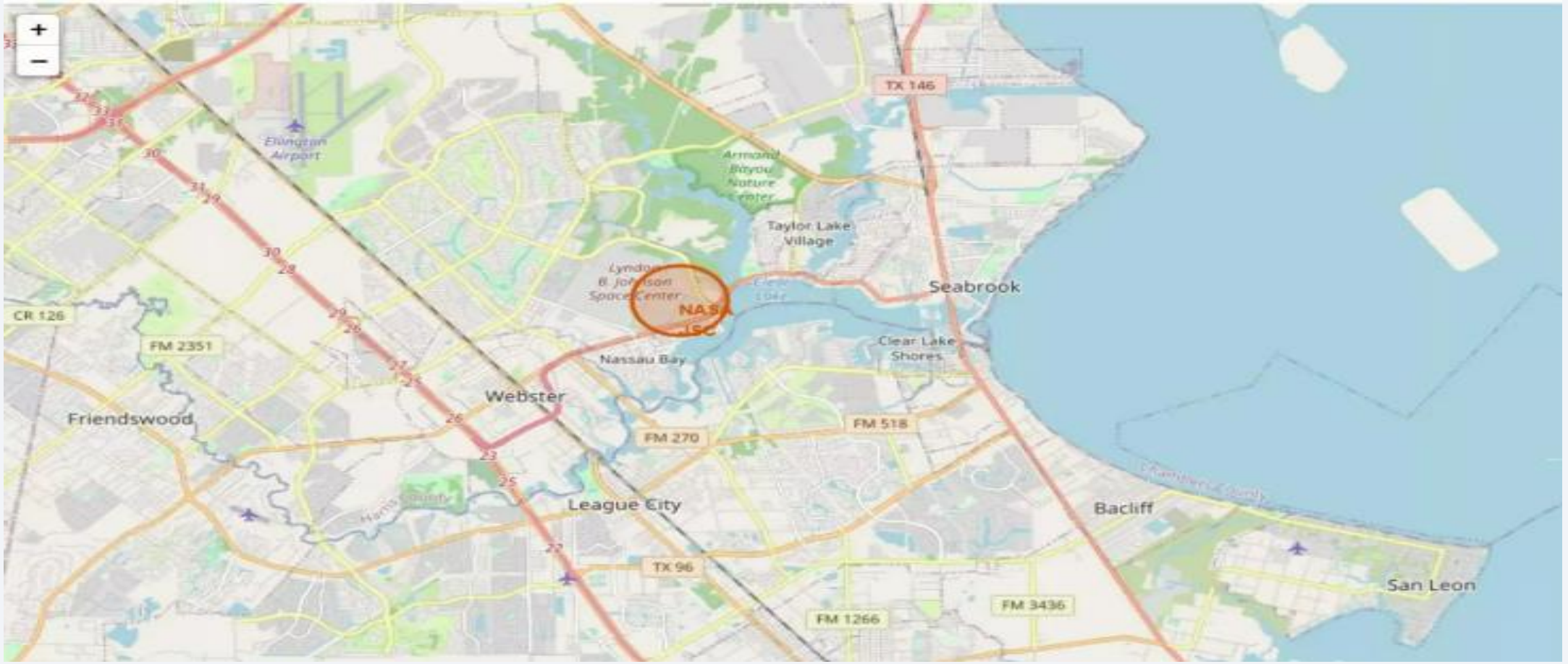
- %sql SELECT "Landing_Outcome", COUNT(*) AS "Count" FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY "Count" DESC;

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

Section 3

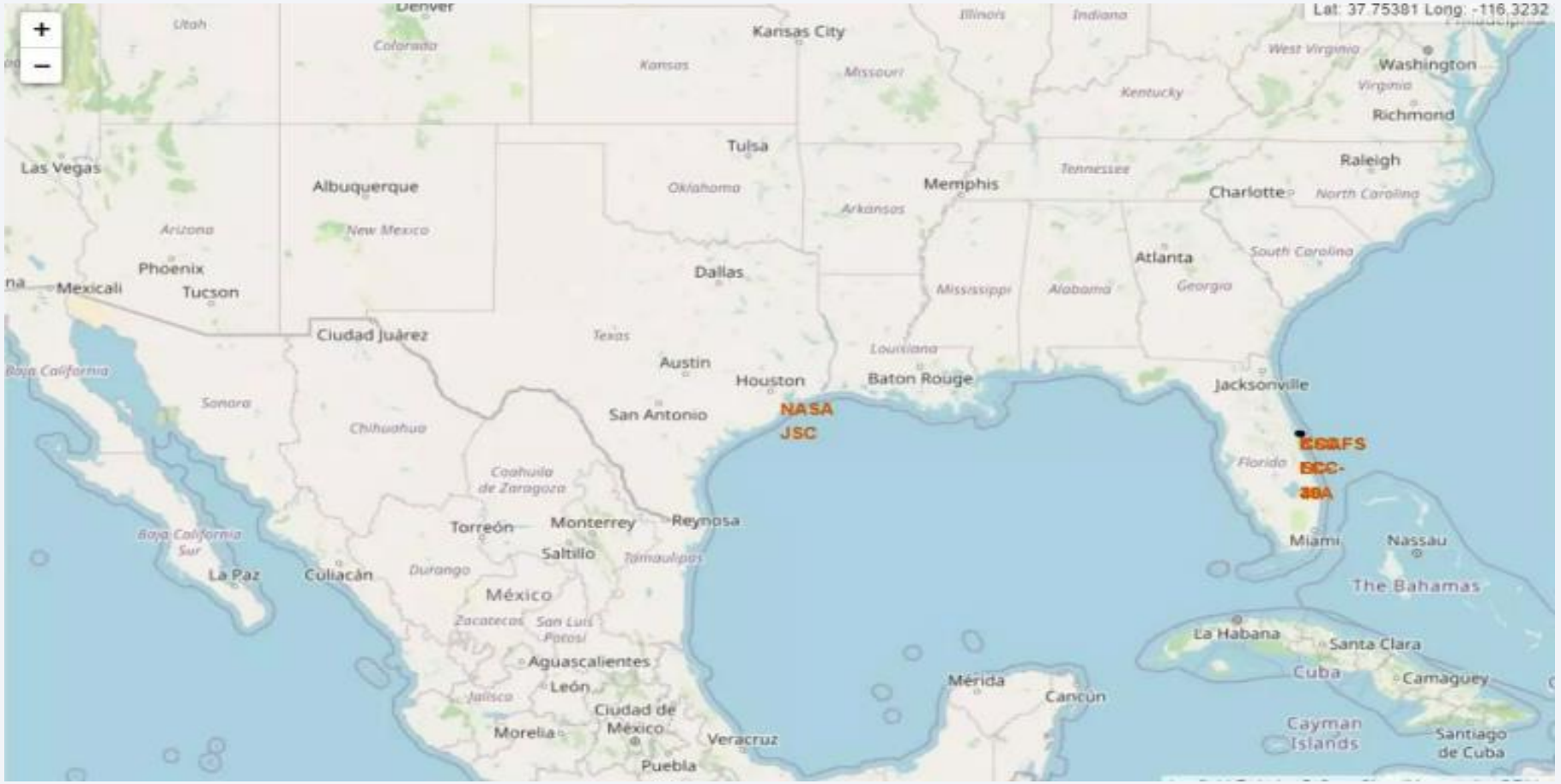
Launch Sites Proximities Analysis

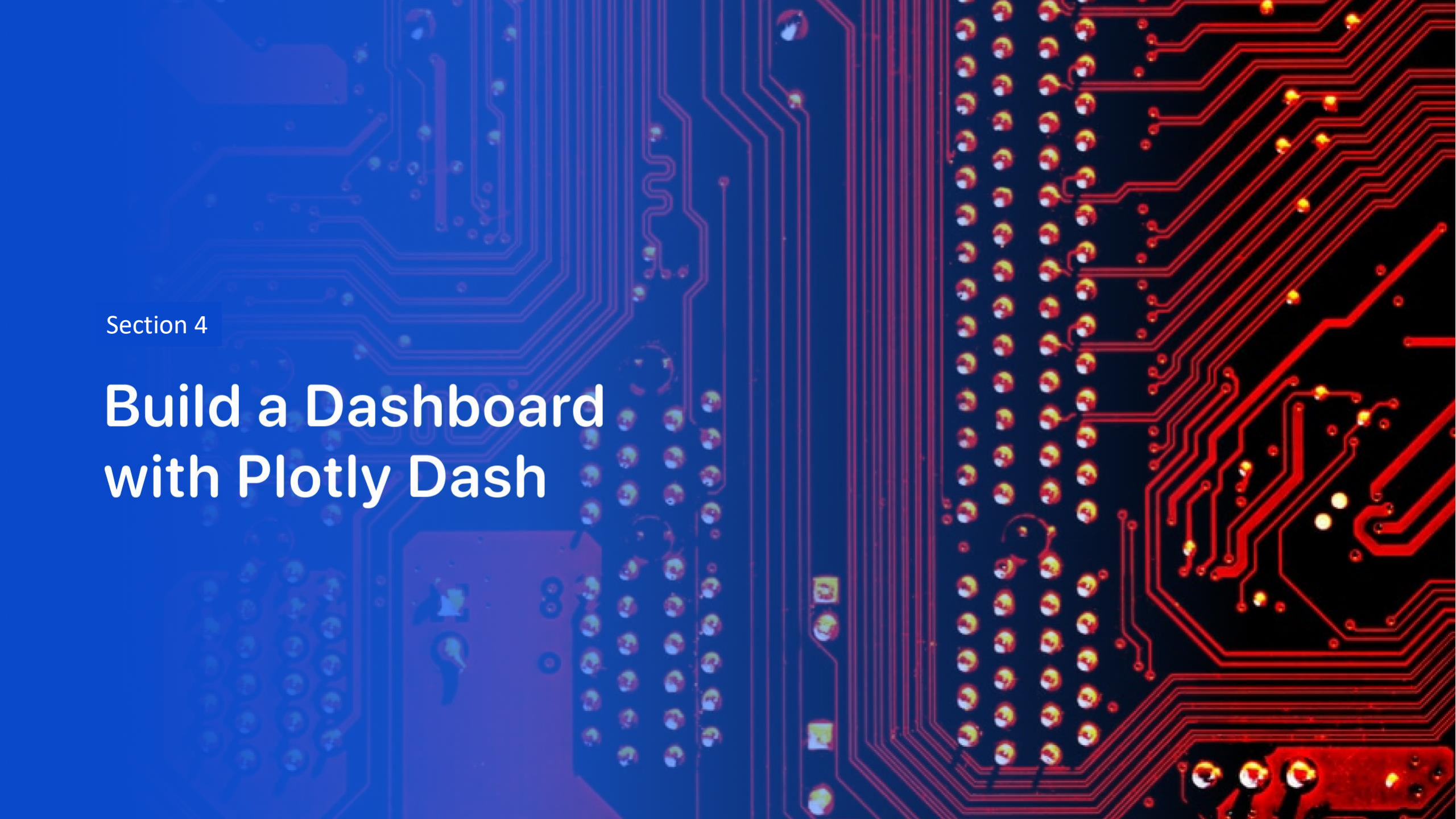
All Launch Sites



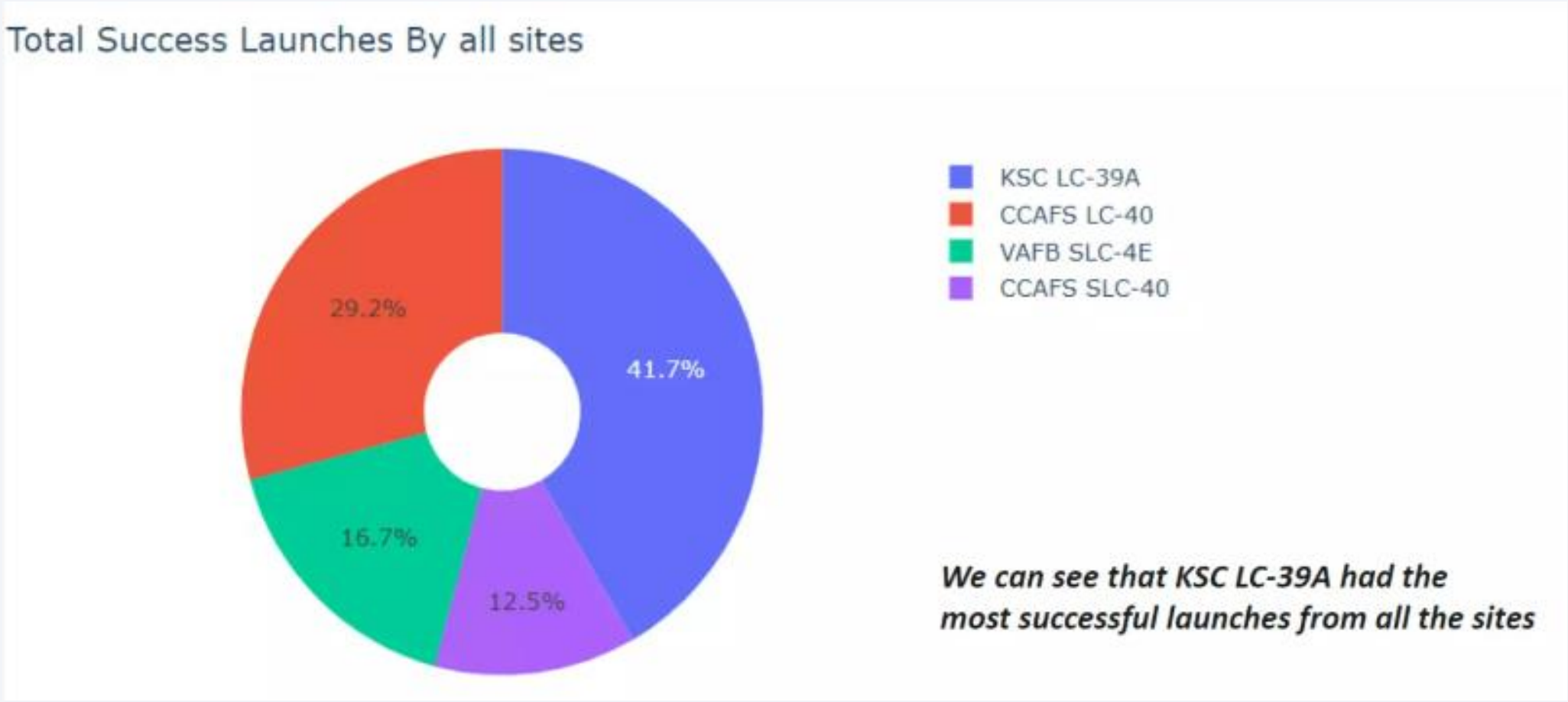
Success/Failed launches

Distance between launch sites to its proximities



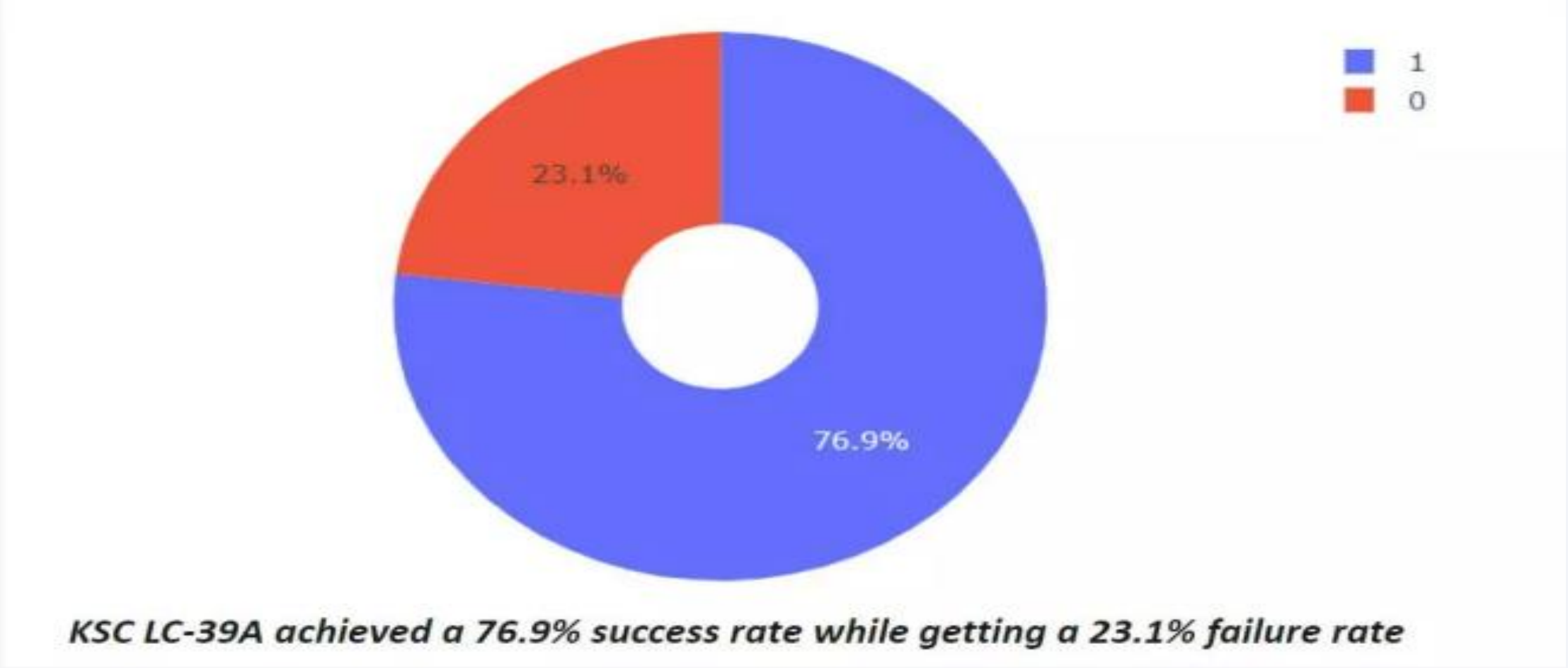

Section 4
Build a Dashboard with Plotly Dash

Total success launches by all sites



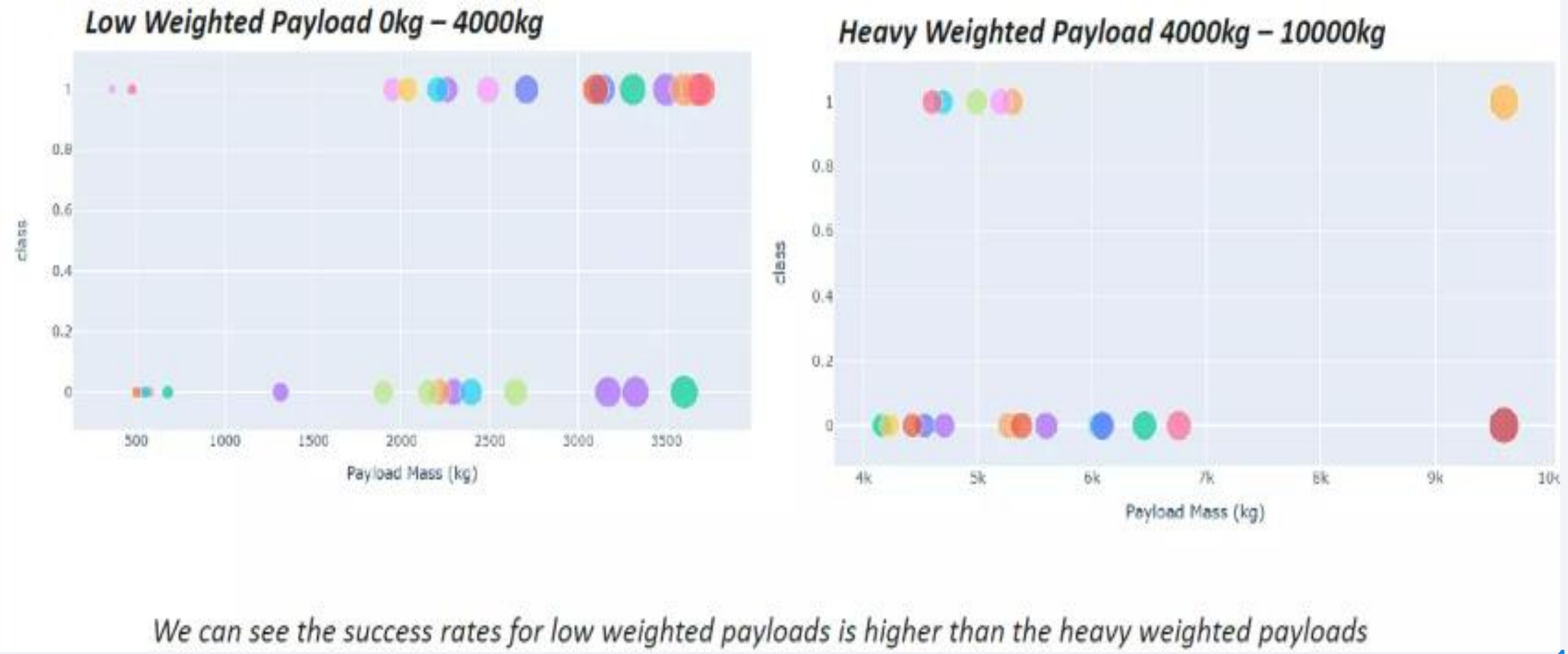
We can see that KSC LC-39A had the most successful launches from all the sites

Success rate by site



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs Launch Outcome

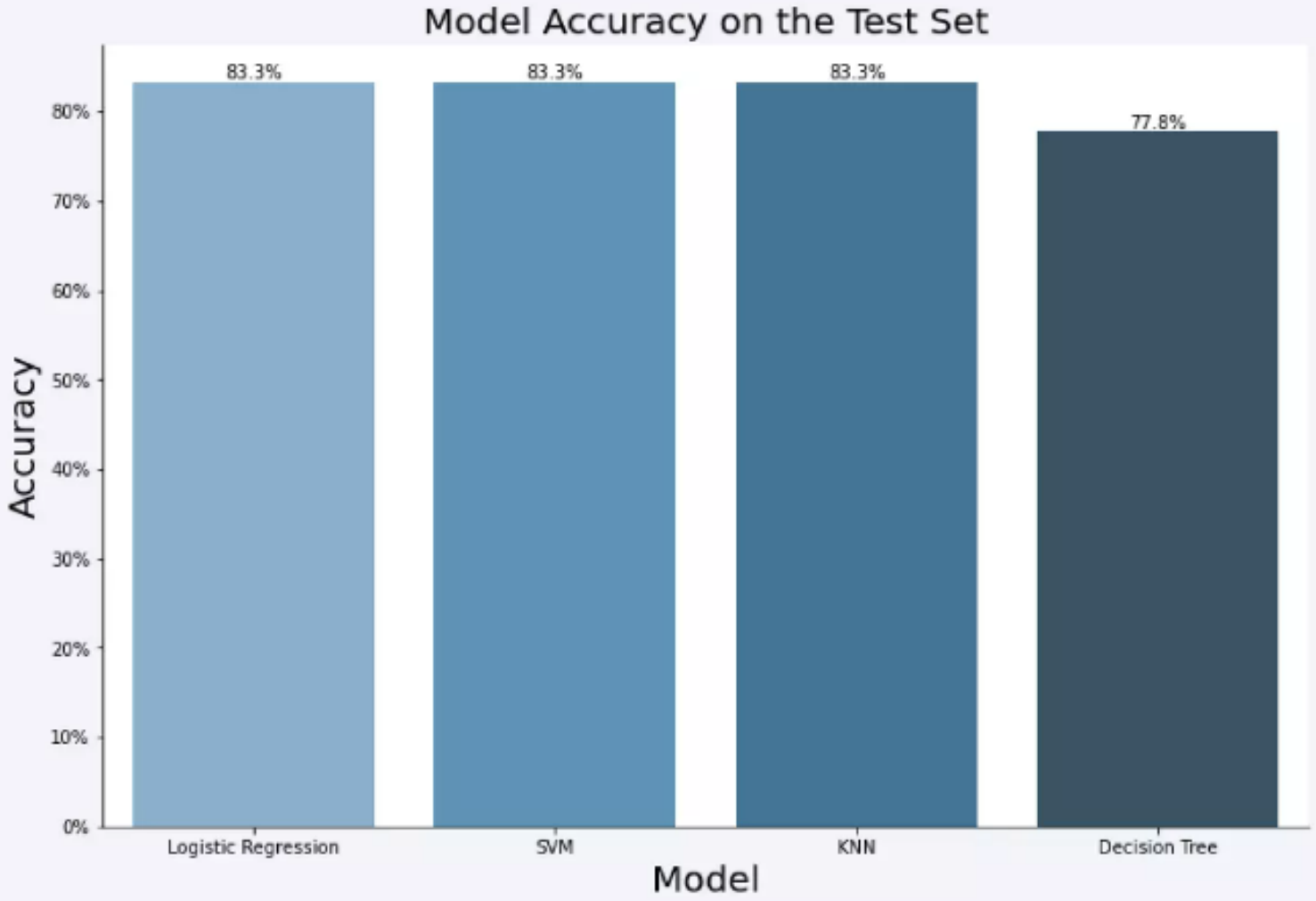


We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

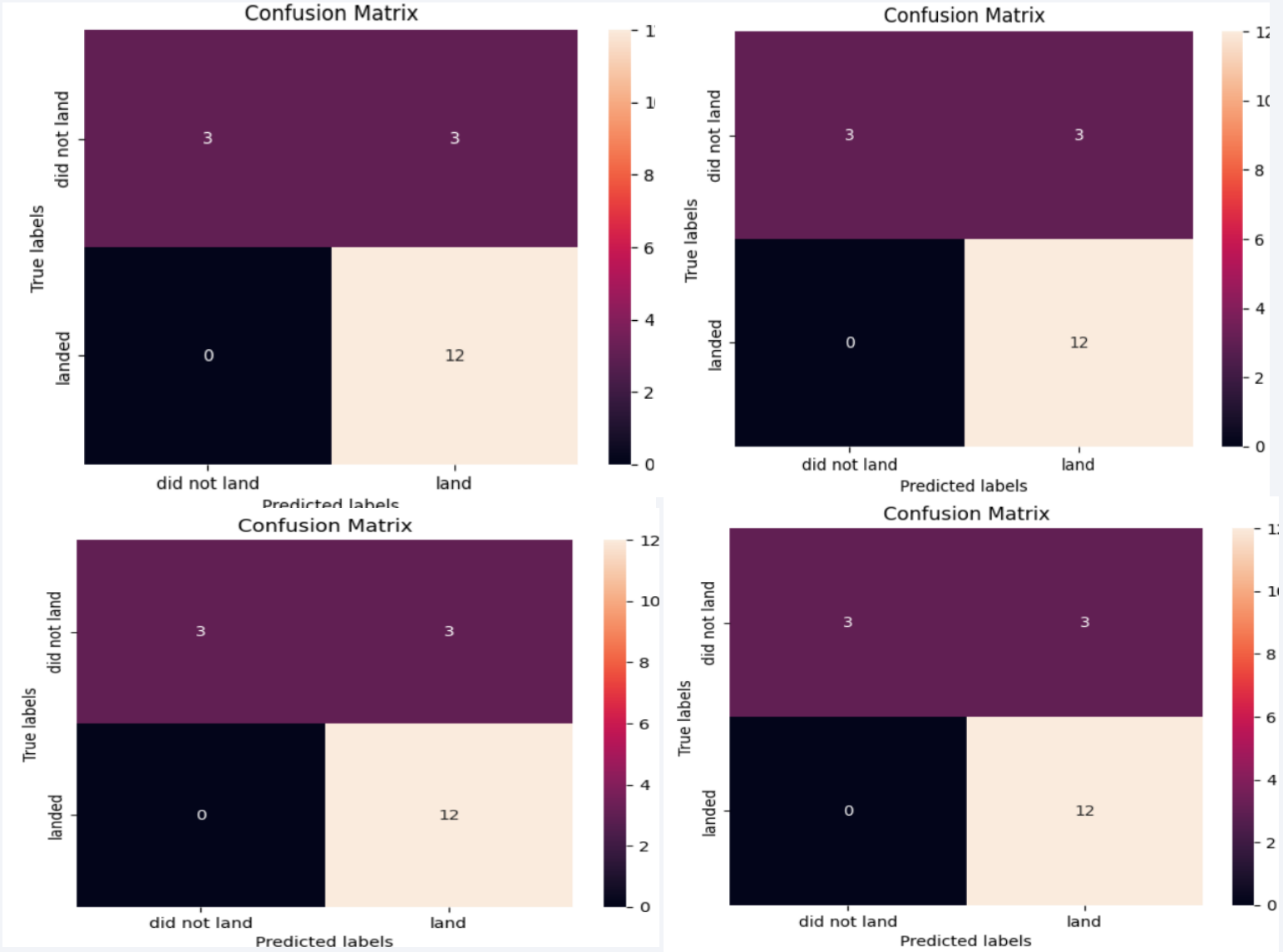
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO,HEO,SSO, ES L1 has the best Success Rate.

Appendix

https://github.com/ArnavGoel1/SpaceX/tree/main

Thank you!