IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dipti Theng
06/01/2025

Outline

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

Executive Summary

- **Key Methodologies**:
  - Data Collection: SpaceX API and Web Scraping from Wikipedia
  - Data Wrangling: Cleaning, One-Hot Encoding
  - EDA: Visualization & SQL
  - Interactive Visual Analytics: Folium & Plotly Dash
  - Predictive Analysis: Classification Models
- **Summary of Results**:
  - Identified key features for launch success
  - Demonstrated interactive visual analytics
  - Found Decision Tree model most effective

Introduction

- **Objective:** Evaluate SpaceY's potential to compete with SpaceX
- **Background:** SpaceX's cost-effective reusable rockets
- **Questions:**
  - Impact of various factors on launch success
  - Trend of success over years
  - Optimal algorithm for predicting success

Section 1

Methodology

Methodology

Executive Summary

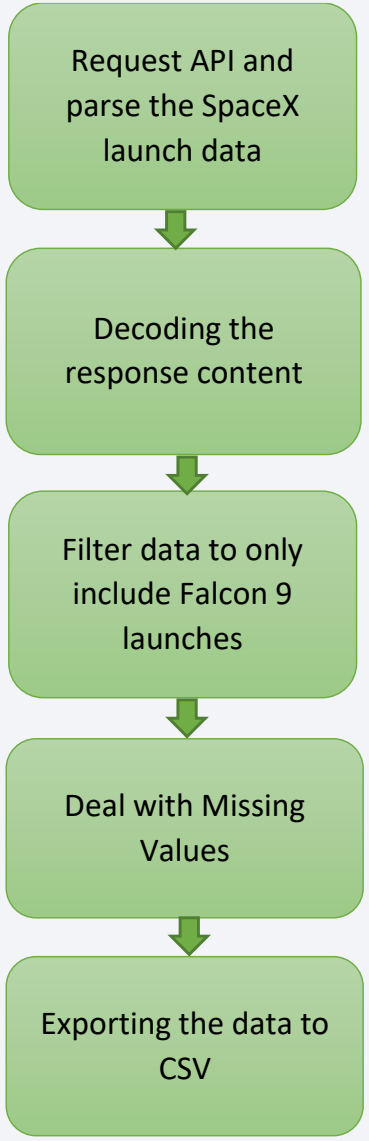
- Data collection methodology:
  - SpaceX API: Flight data, rocket specs
  - Wikipedia Scraping: Historical launch details
- Data wrangling
  - Handling missing values, filtering, encoding
  - Labeling outcomes for machine learning
- Exploratory Data Analysis (EDA):
  - SQL queries for insights
  - Visualizations: Payload, Launch Site, Success Rates

Data Collection

- Data sets were collected using web scraping technics from
  - ✓ **Space X API:** https://api.spacexdata.com/v4/rockets/
  - ✓ **Wikipedia**: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- Data Columns are obtained by using SpaceX REST API:
  - ✓ FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
  - ✓ Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

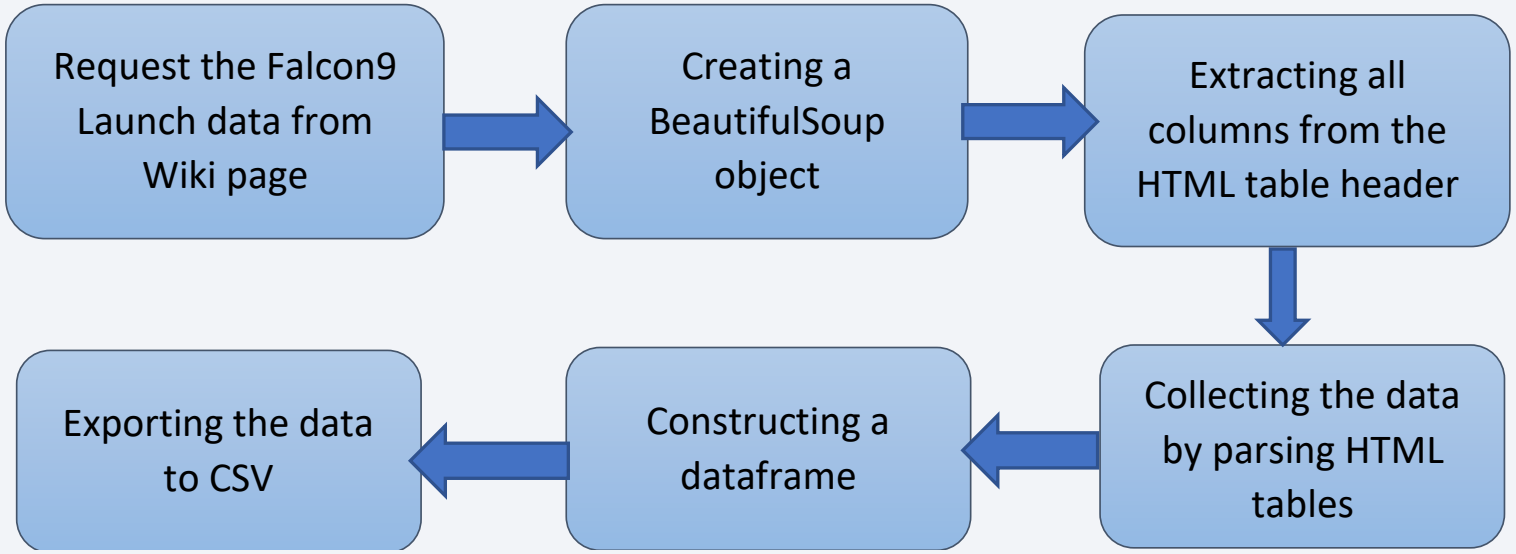
- SpaceX API: Flight data, rocket specs
- Wikipedia Scraping: Historical launch details
- GitHub URL of the completed SpaceX API calls notebook:



Data Collection - Scraping

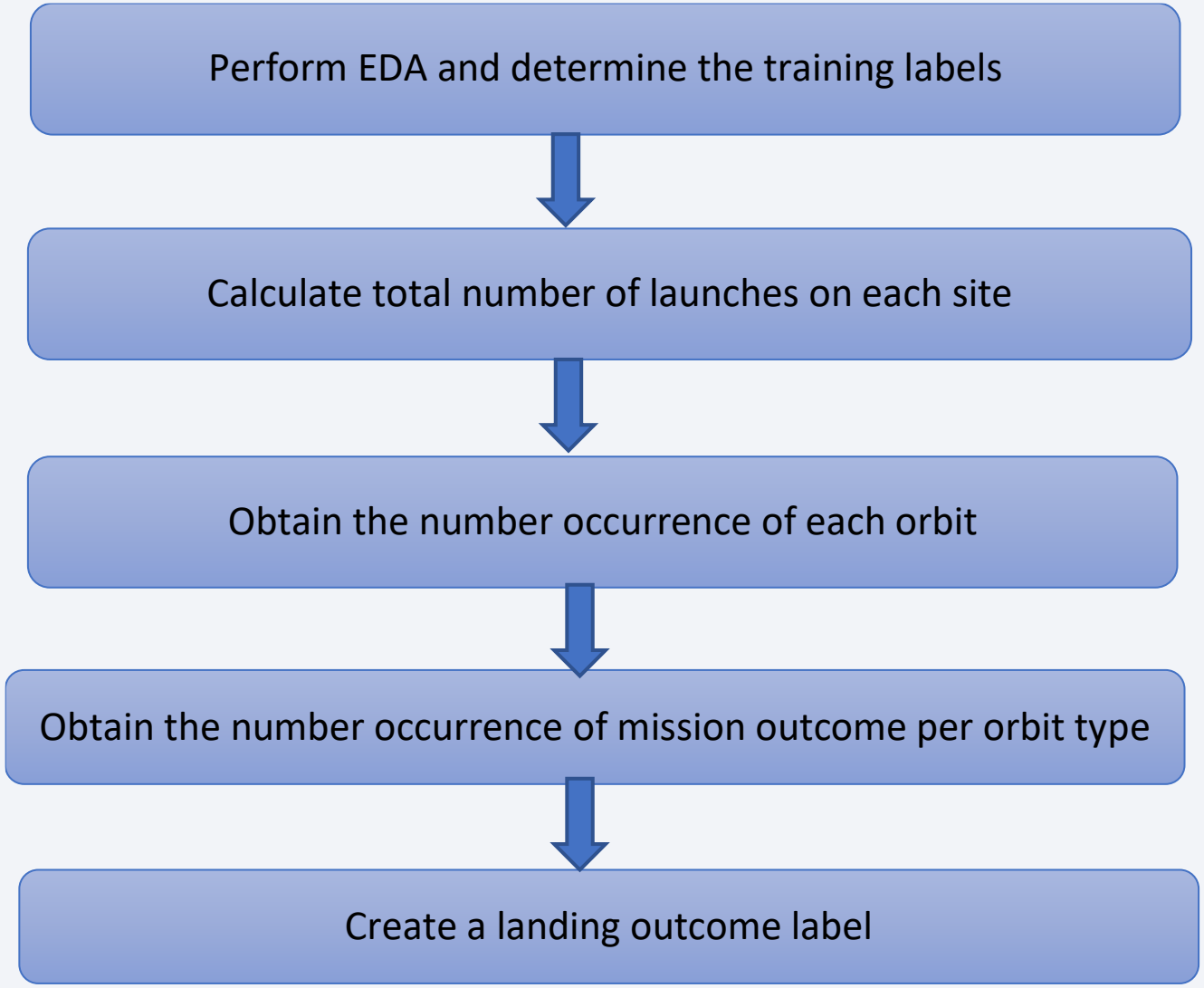
- Web scraping process is depicted in the figure here

- GitHub URL of the completed web scraping notebook:



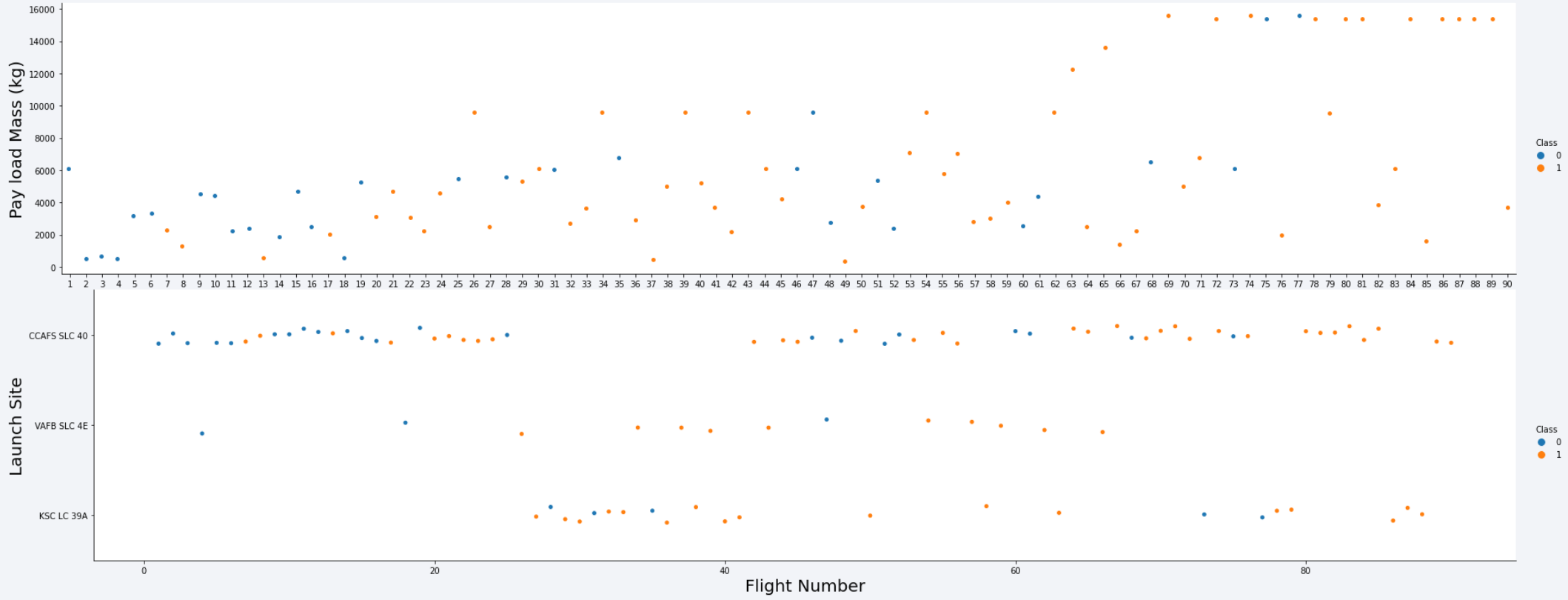
Data Wrangling

- GitHub URL of your completed data wrangling related notebooks:



EDA with Data Visualization

- Scatterplots and bar plots were used to visualize the relationship between pair of features: Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit
- GitHub URL of your completed EDA with data visualization notebook:



EDA with SQL

- SQL queries are performed for below tasks:
  - To Display the names of the unique launch sites in the space mission
  - To Display Top 5 launch sites whose name begin with the string 'CCA'
  - To Display the total payload mass carried by boosters launched by NASA (CRS)
  - To Display average payload mass carried by booster version F9 v1.1
  - To Display the date when the first successful landing outcome in ground pad was achieved
  - To Display the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - To Display the total number of successful and failure mission outcomes
  - To Display the names of the booster versions which have carried the maximum payload mass
  - To Display the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - For ranking 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
- GitHub URL:

Build an Interactive Map with Folium

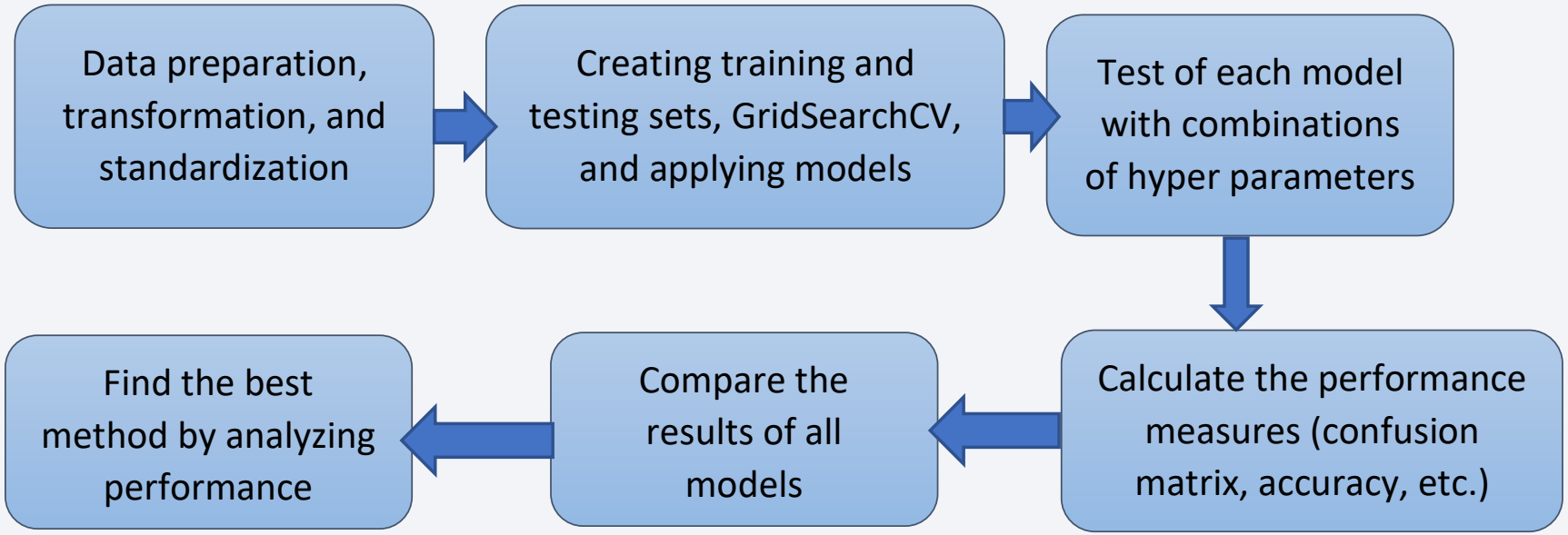
- Launch sites location markers, success/failure color-coded markers, proximity analysis of launch sites:
  - ✓ Markers indicate points like launch sites.
  - ✓ Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center.
  - ✓ Marker clusters indicates groups of events in each coordinate, like launches in a launch site.
  - ✓ Lines are used to indicate distances between two coordinates.
- GitHub URL:

Build a Dashboard with Plotly Dash

- Features: Interactive launch success visualizations, payload impact on launch outcome, user-friendly interface for analysis.
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.
- GitHub URL

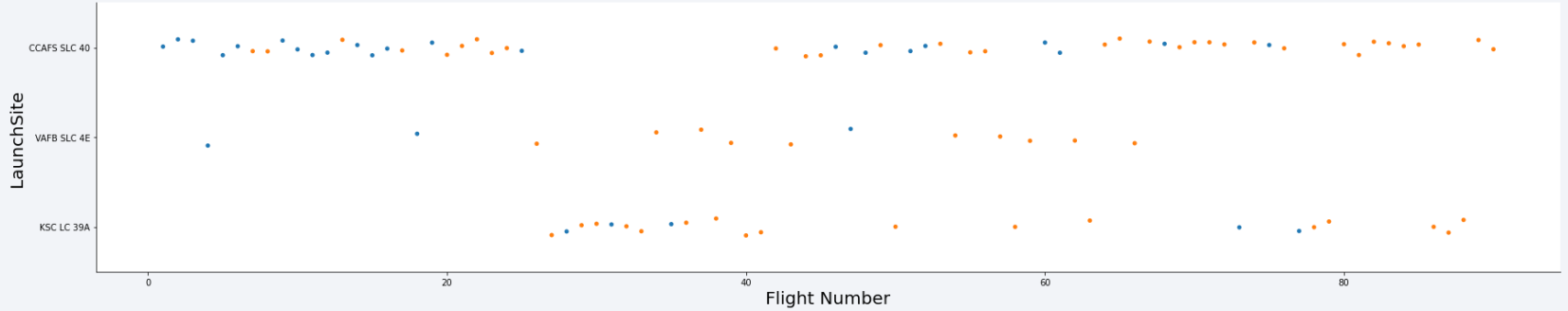
Predictive Analysis (Classification)

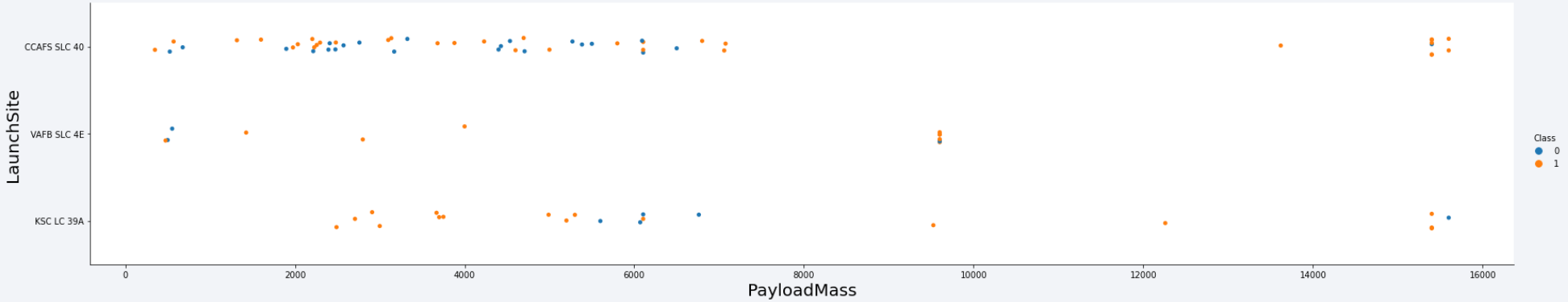
- **Model Performance:**
  - Confusion matrix, accuracy scores
  - Decision Tree model identified as best fit
- **Confusion Matrix Insights:**
  - High accuracy in distinguishing success/failure
- GitHub URL

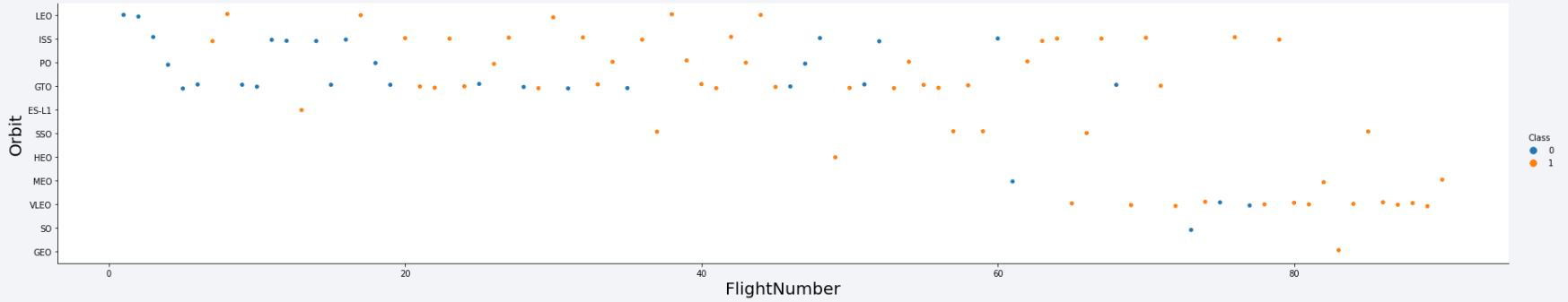


Results

- Exploratory data analysis results: Visualize the relationship between different variables

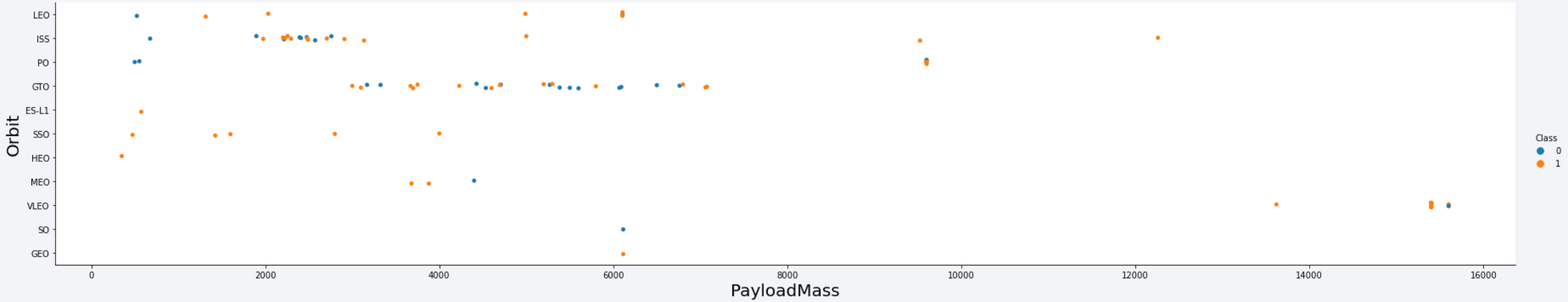


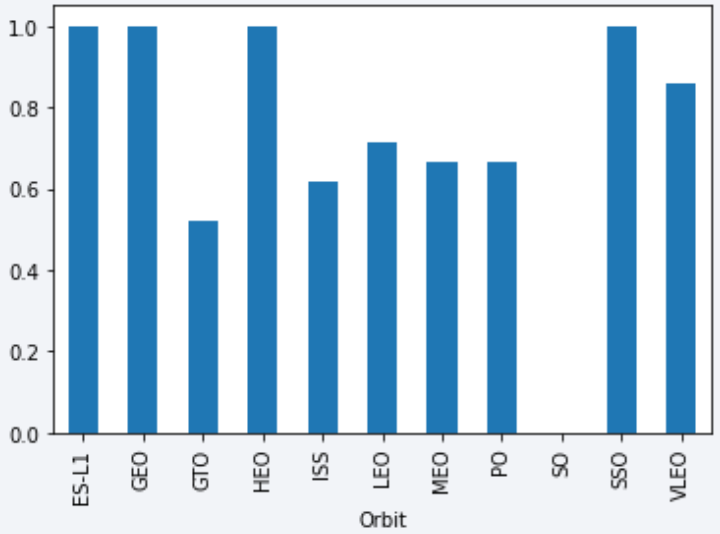




Results

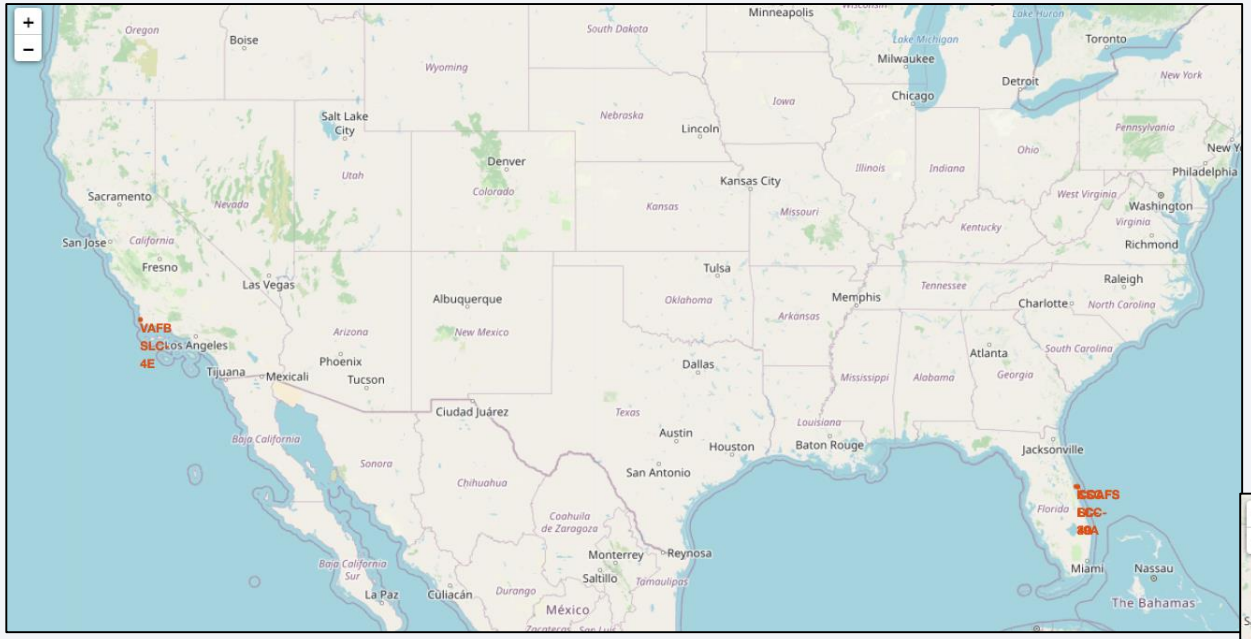
- Exploratory data analysis results: Visualize the relationship between different variables

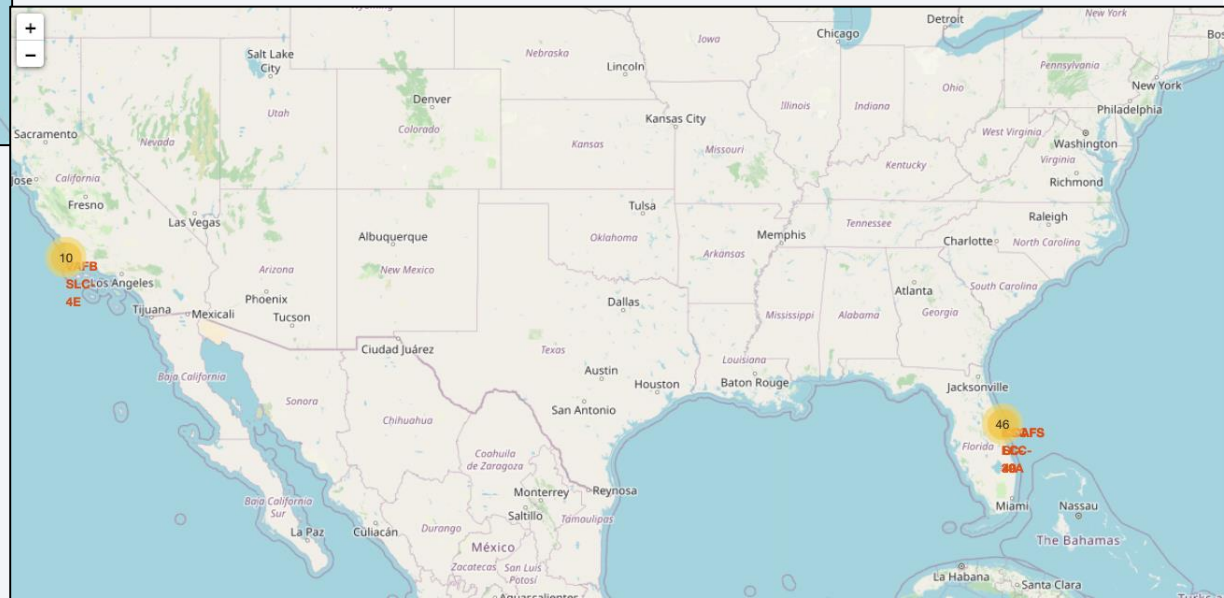




Results

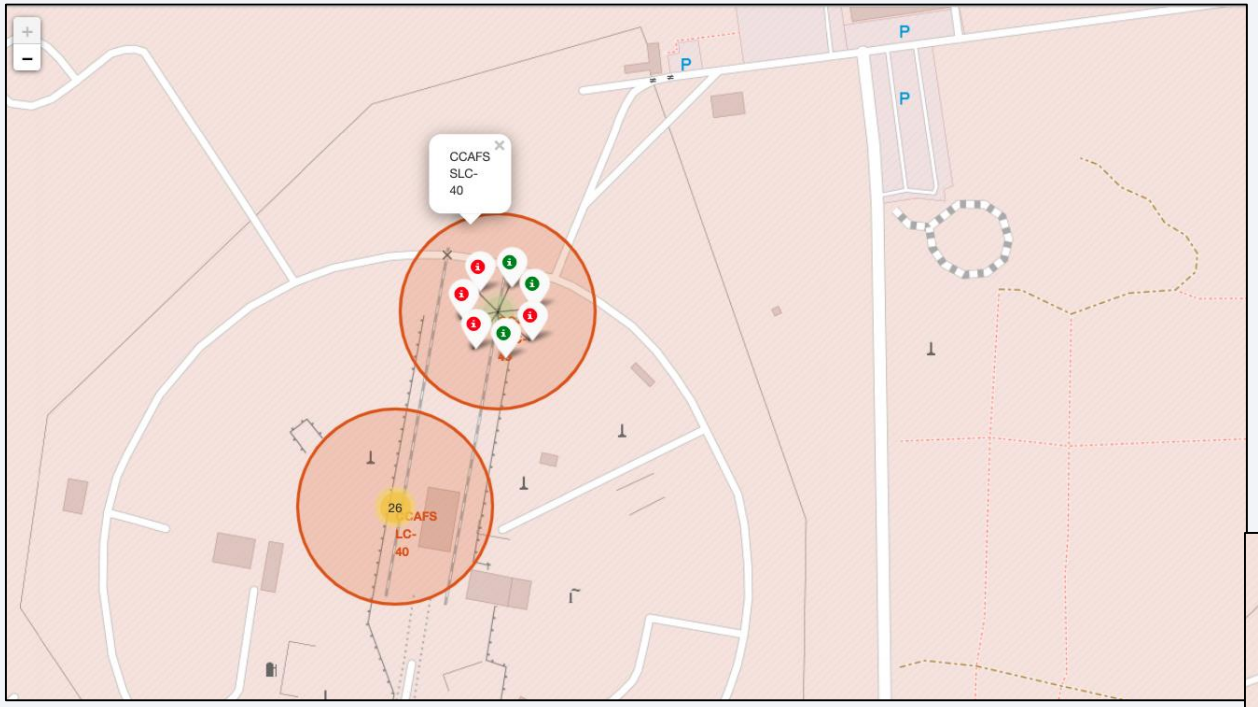
- Interactive analytics demo in screenshots

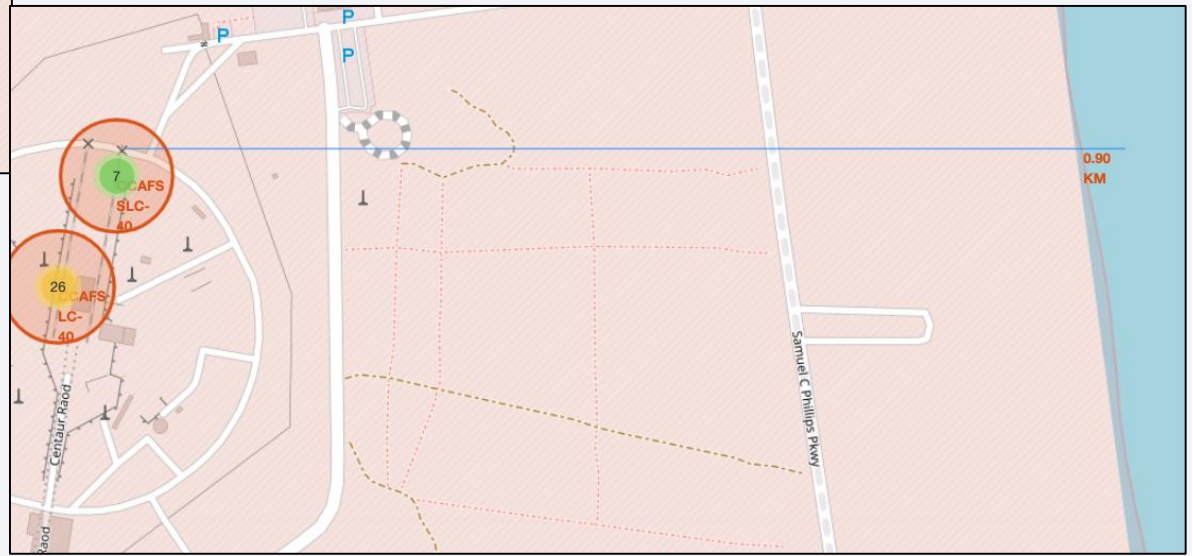




Results

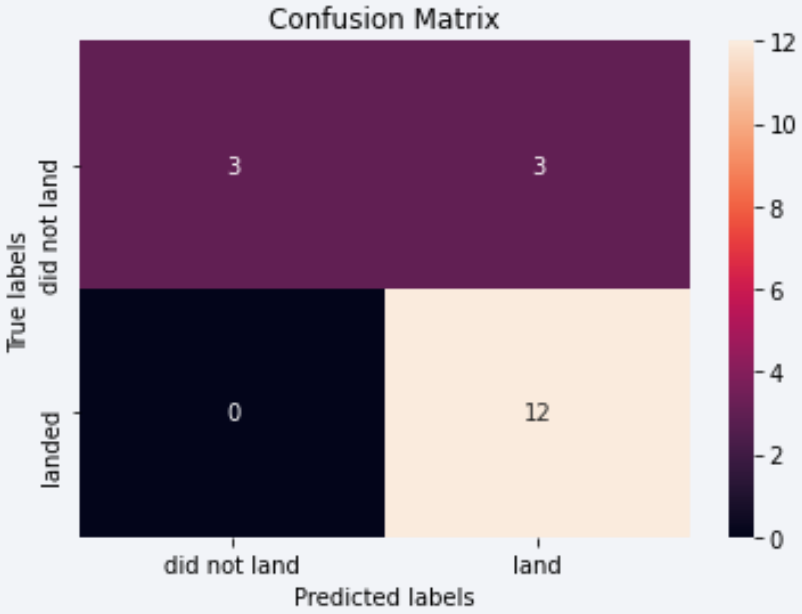
- Interactive analytics demo in screenshots



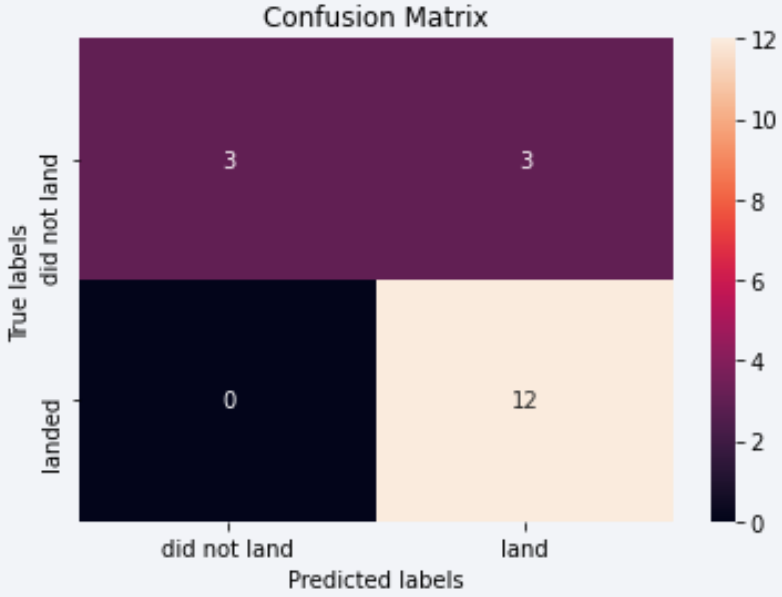


Results

- Predictive analysis results



a) LogReg



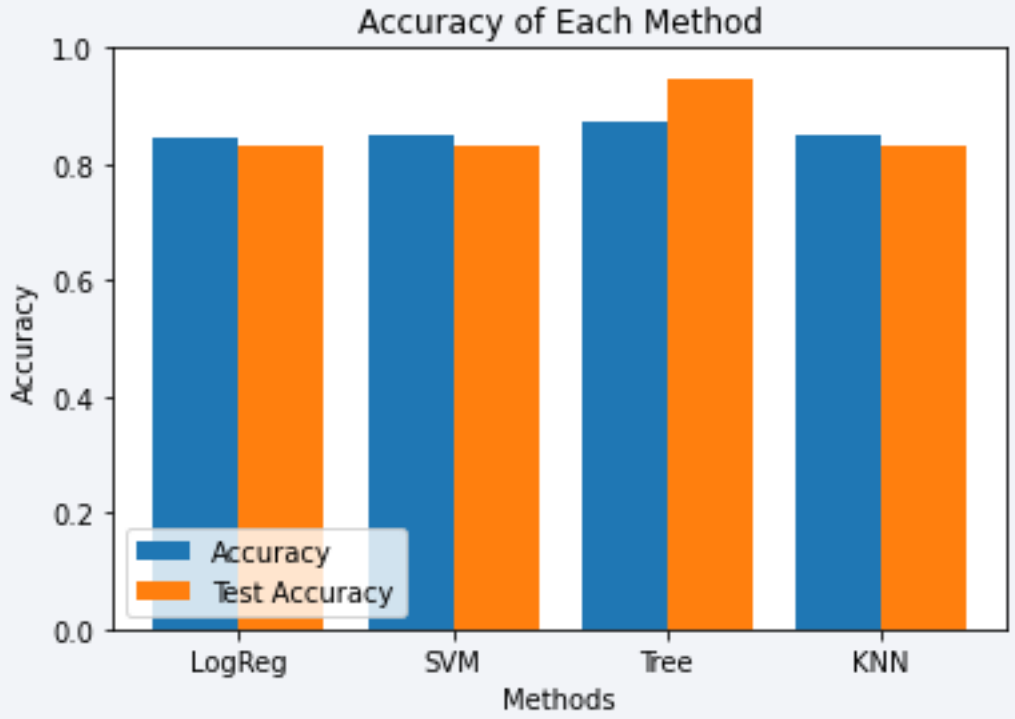
b) SVM

c) KNN

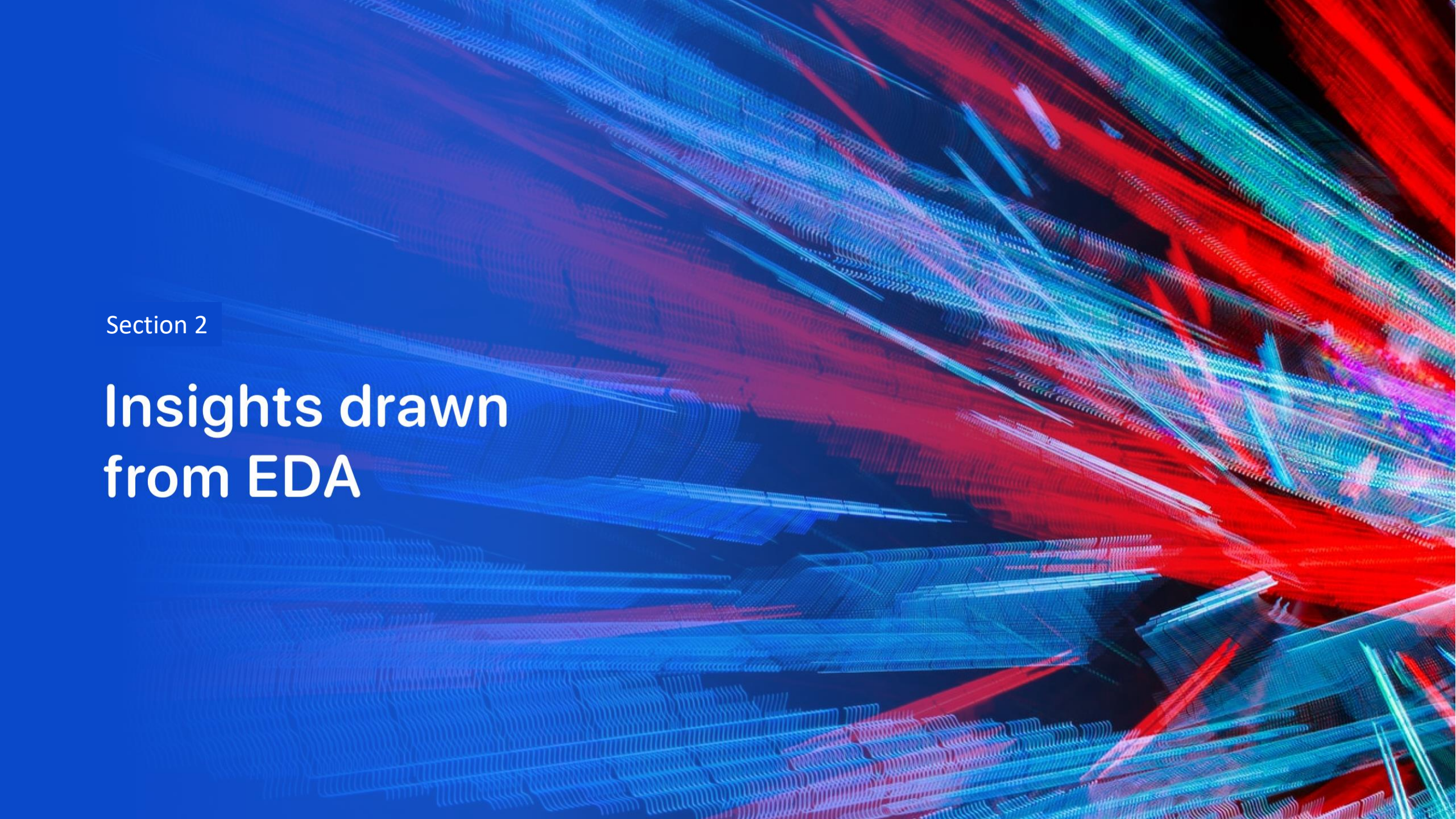
Results

- Predictive analysis results

c) Tree

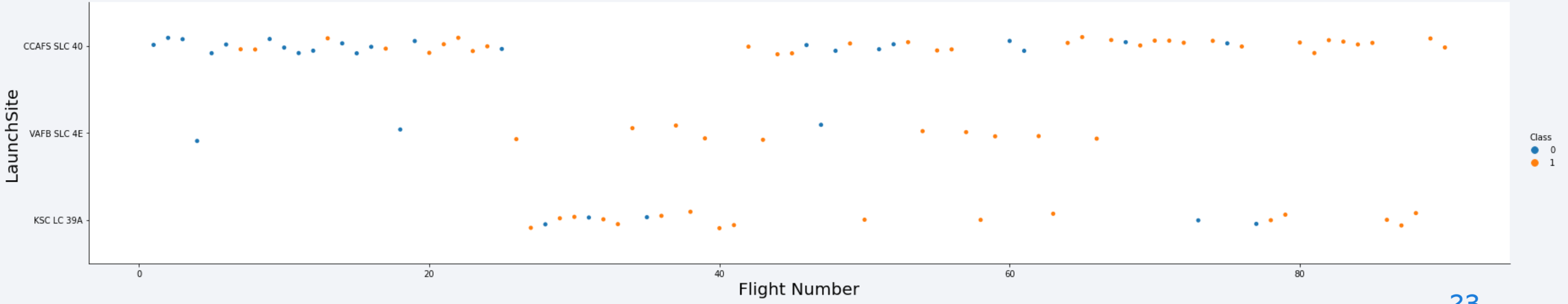


Performance Comparison


Section 2
Insights drawn from EDA

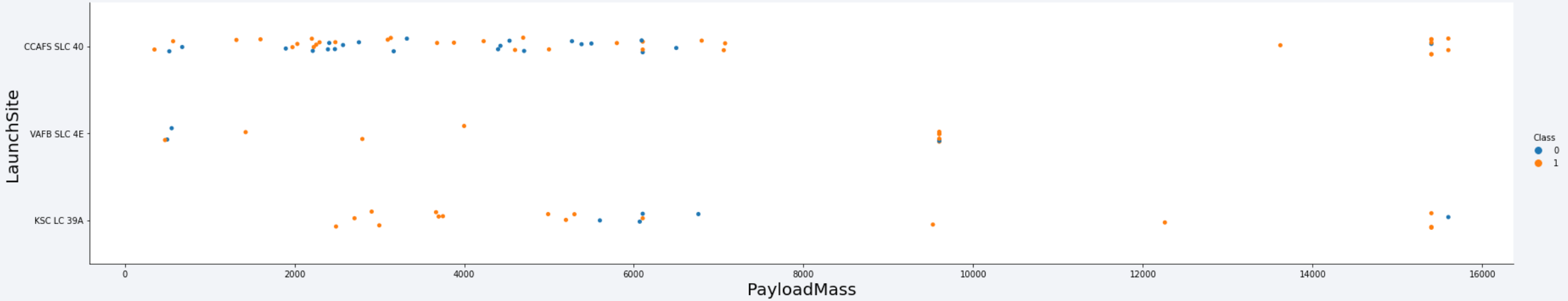
Flight Number vs. Launch Site

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.



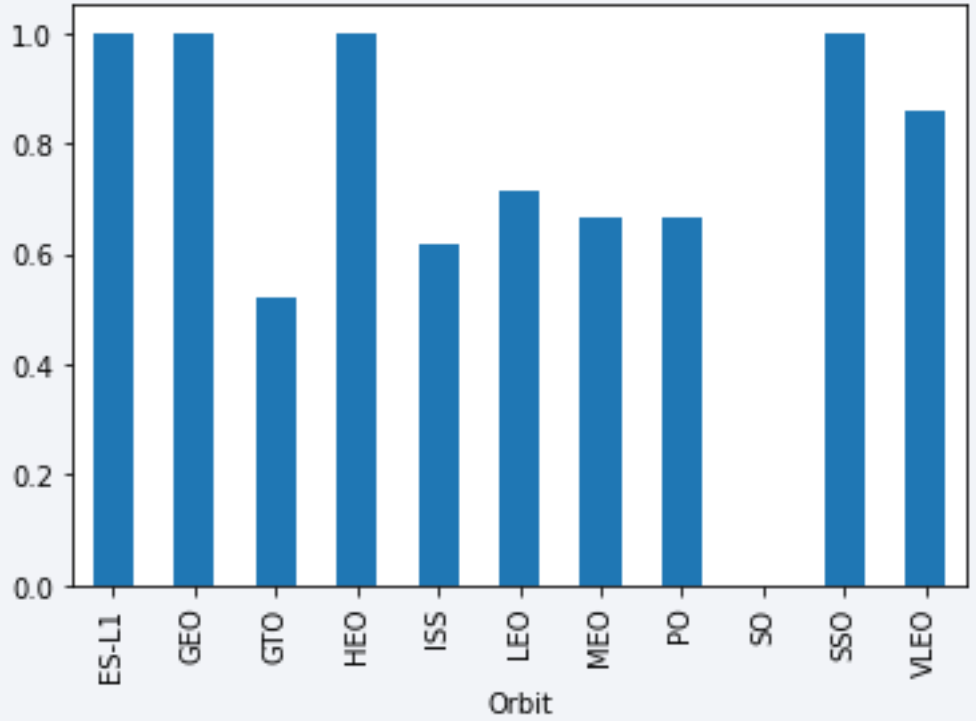
Payload vs. Launch Site

- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



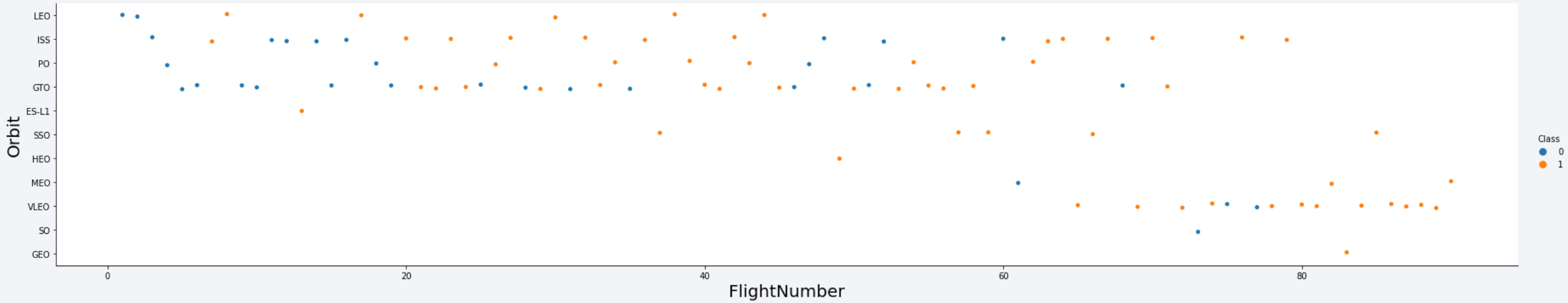
Success Rate vs. Orbit Type

- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: SO
- Orbits with success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO



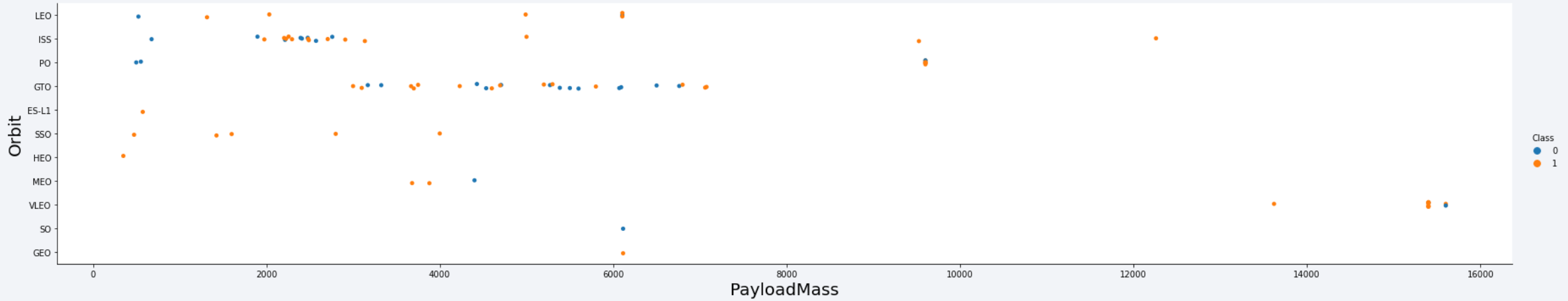
Flight Number vs. Orbit Type

- Success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency



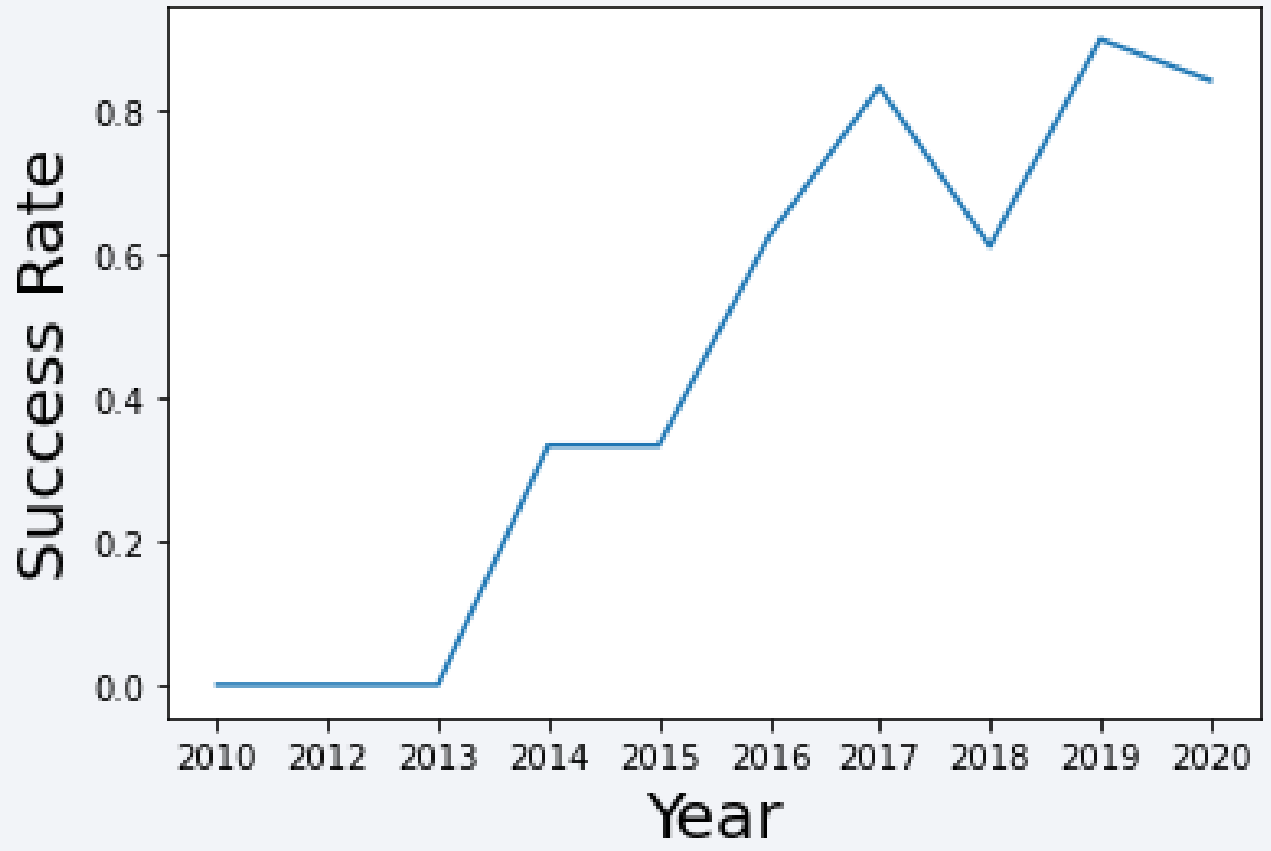
Payload vs. Orbit Type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

- The names of the unique launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- The information is obtained by selecting unique occurrences of “launch_site” values from the dataset.
- SQL Query: %sql select distinct launch_site from SPACEXDATASET;

Launch Site Names Begin with 'CCA'

- The 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	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

- SQL Query: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

Total Payload Mass

- The total payload carried by boosters from NASA

total_payload_mass
45596

- SQL Query: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

Average Payload Mass by F9 v1.1

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

average_payload_mass
2534

- SQL Query: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

First Successful Ground Landing Date

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

- SQL Query: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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

- SQL Query: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;

Total Number of Successful and Failure Mission Outcomes

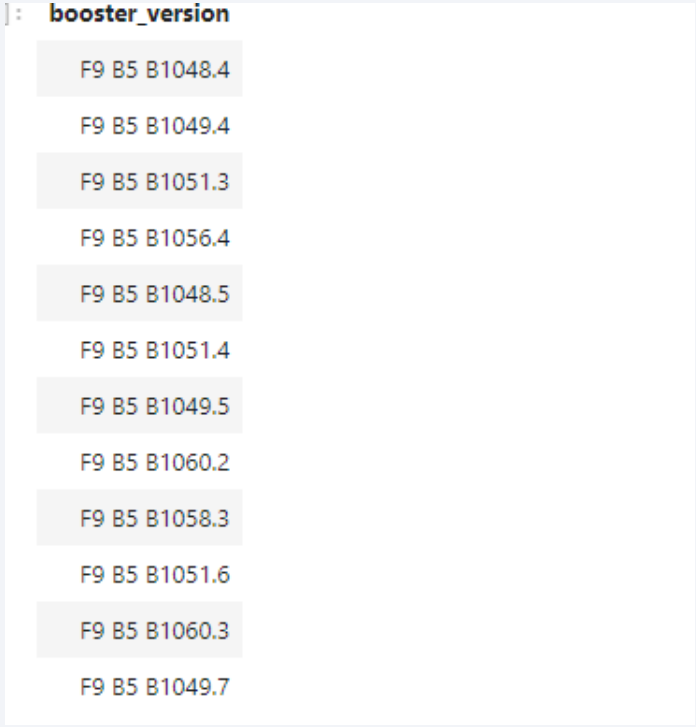
- The total number of successful and failure mission outcomes

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- SQL Query: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

Boosters Carried Maximum Payload

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



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

- SQL Query: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

2015 Launch Records

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

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- SQL Query: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

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

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

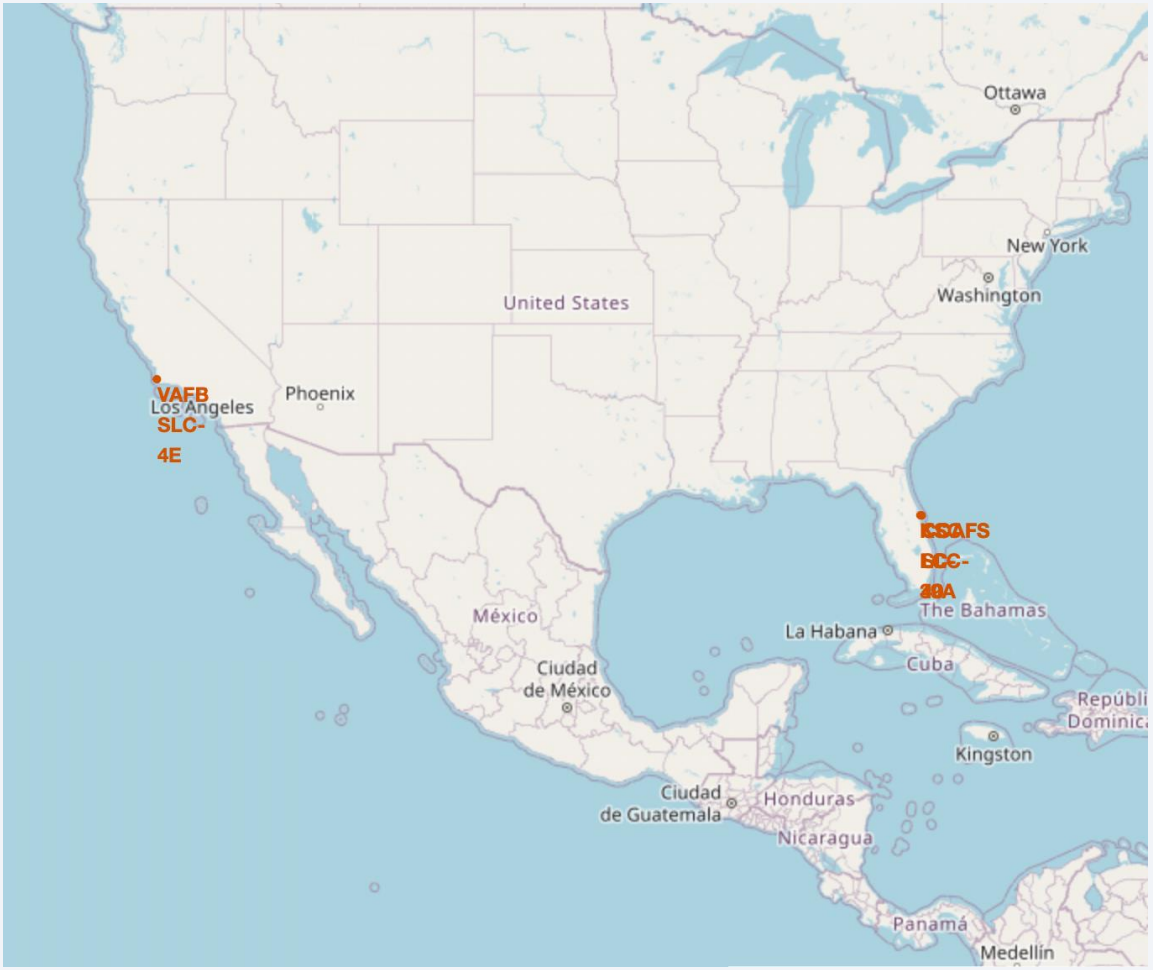
- SQL Query: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET where between '2010-06-04' and '2017-03-20' group by landing__outcome order by count_outcomes desc;

Section 3

Launch Sites Proximities Analysis

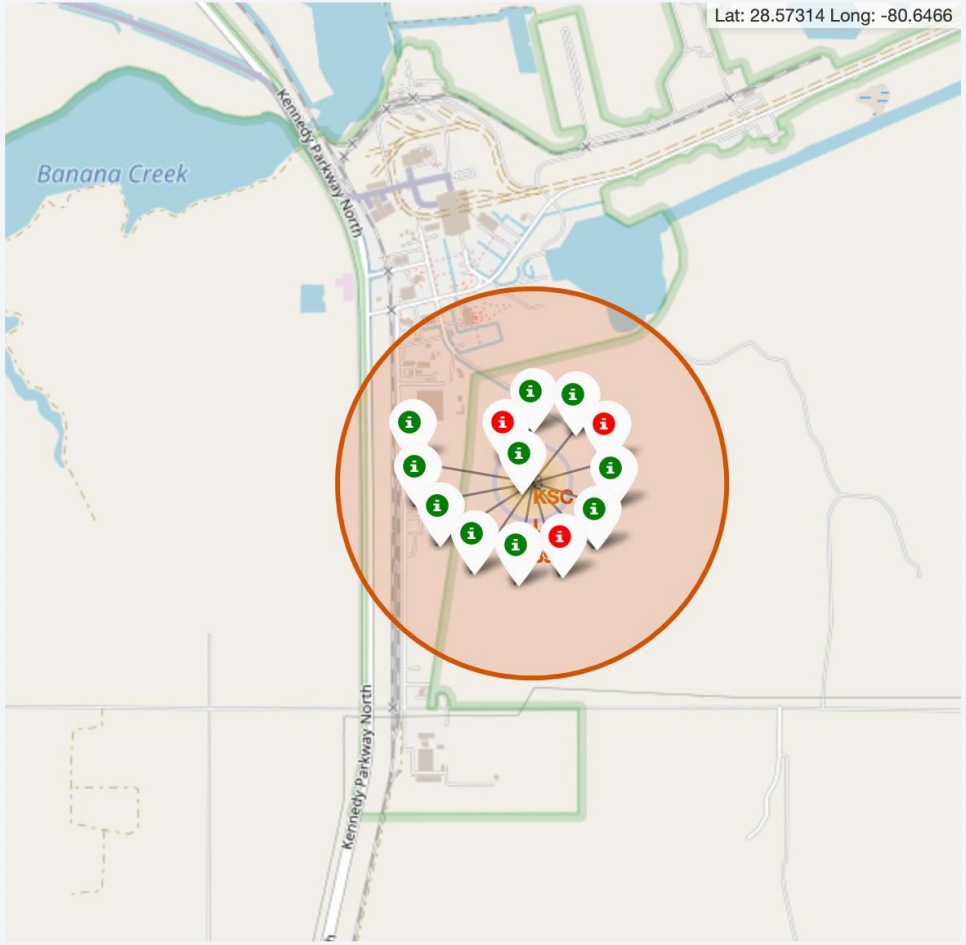
All Launch Sites' Global Map

- Launch sites are near the sea for safety but close to roads and railways for accessibility.
- Most launch sites are near the equator, where Earth's rotation speed is highest (1670 km/h), aiding spacecraft to reach orbit due to inertia.
- Coastal locations reduce the risk of debris impacting populated areas during launches.



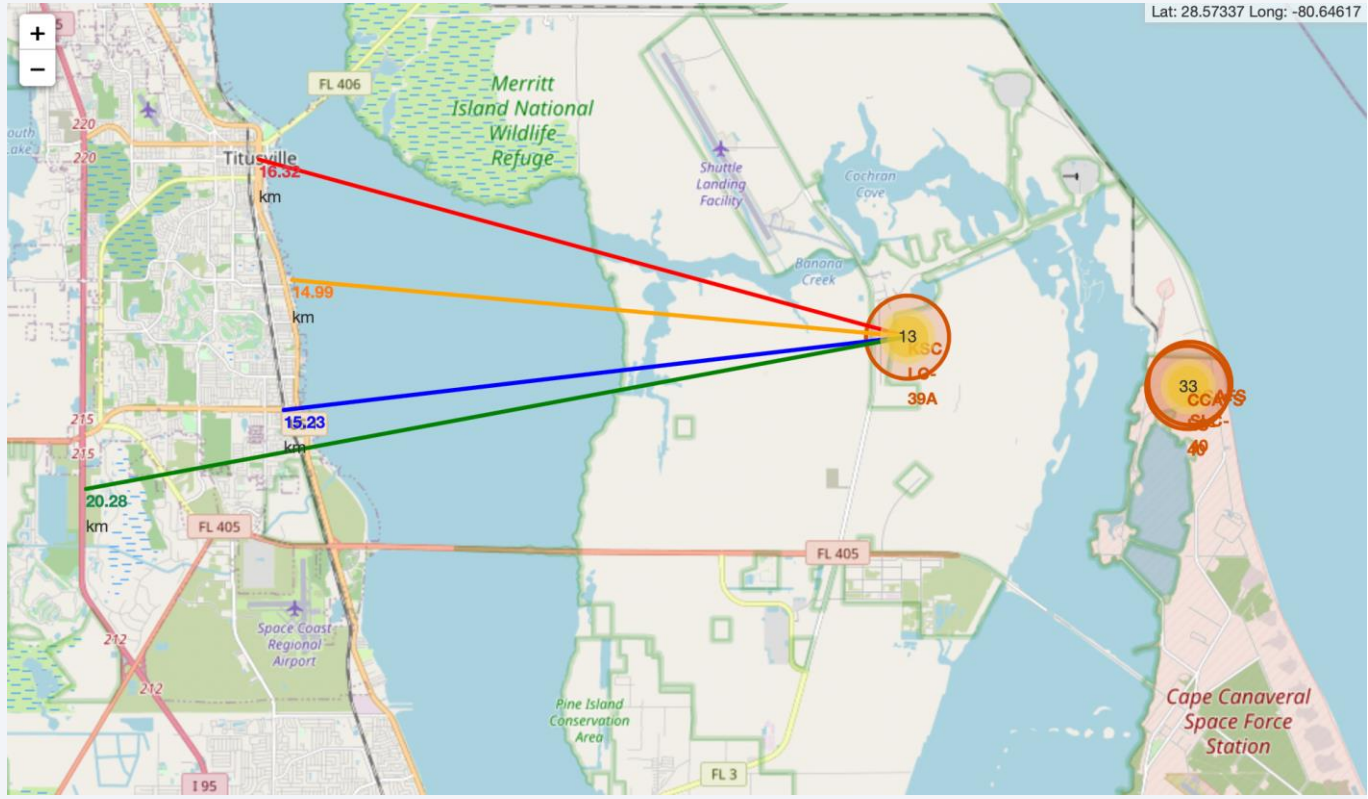
Launch Outcome Records

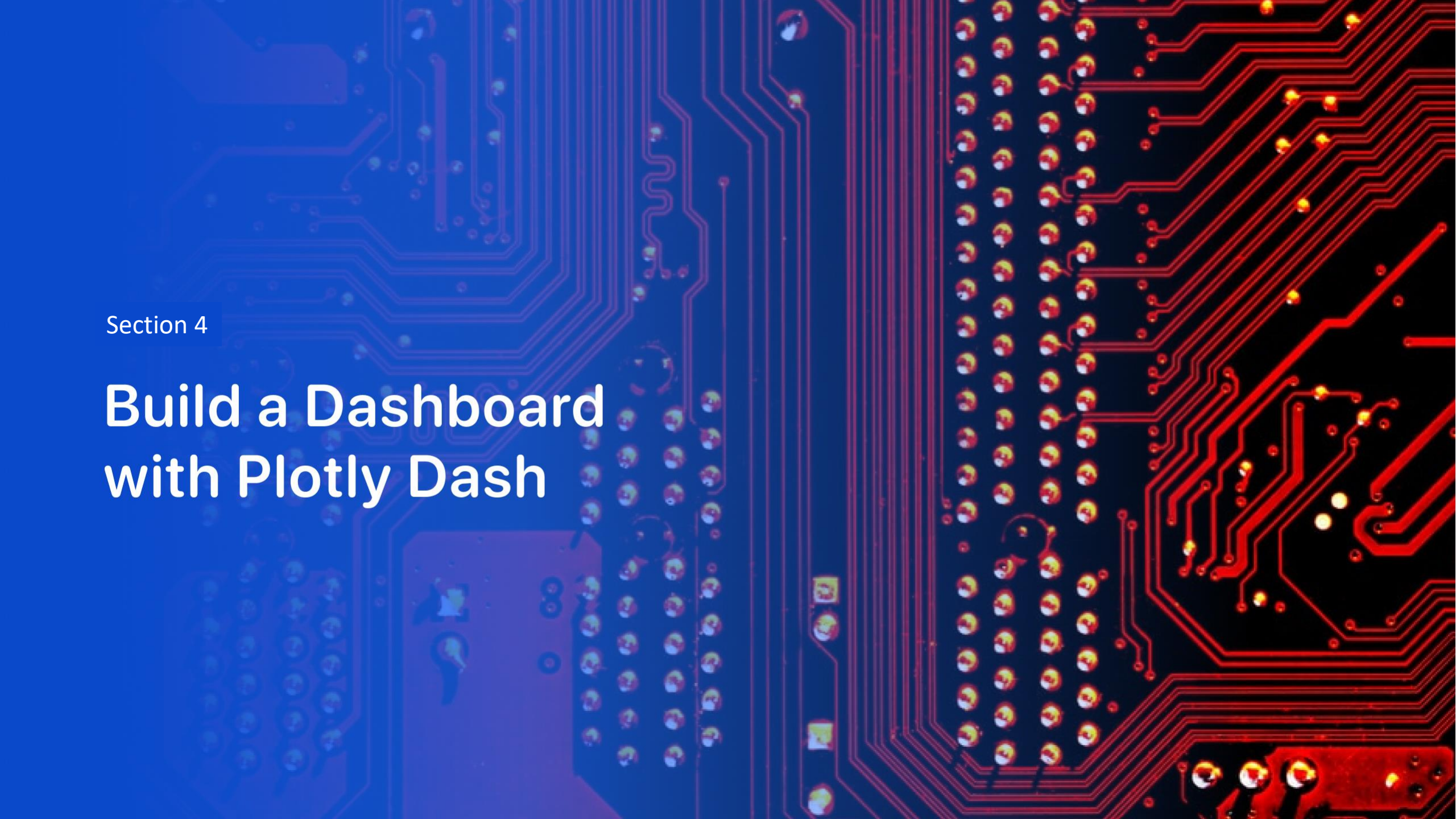
- Color-coded markers help identify launch sites' success rates: green for success, red for failure.
- Launch Site KSC LC-39A has a notably high success rate.



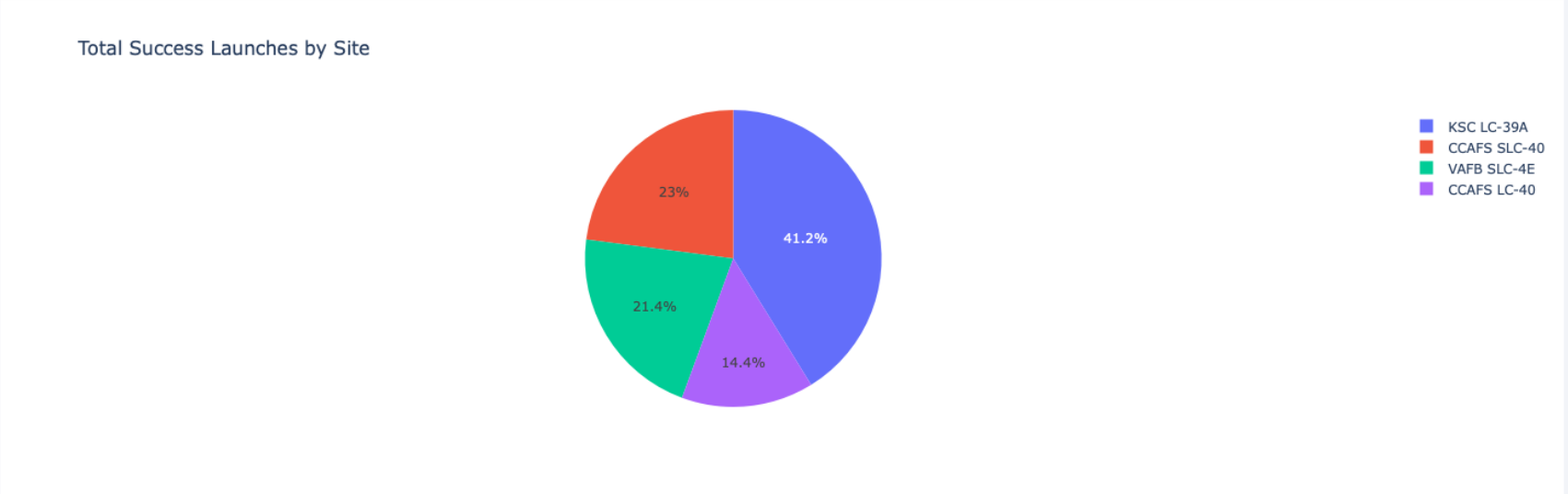
Proximity and Safety Analysis of Launch Site KSC LC-39A

- Depicts Logistics and Accessibility.
- Safety Considerations:
  - ➢ High-speed failed rockets can cover 15–20 km in seconds, posing risks to populated areas.
  - ➢ The location balances logistical convenience with safety by being near infrastructure yet away from densely populated zones.



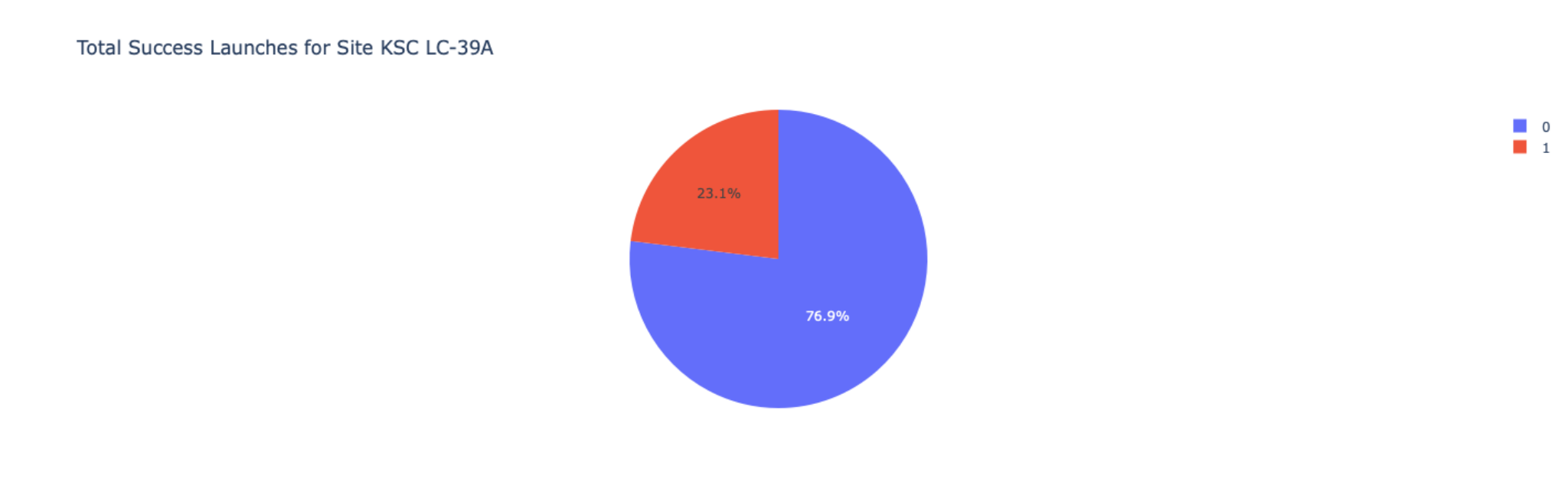

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count by Site



- The chart highlights that KSC LC-39A has the highest number of successful launches.
- The choice of launch site significantly impacts the success of missions.

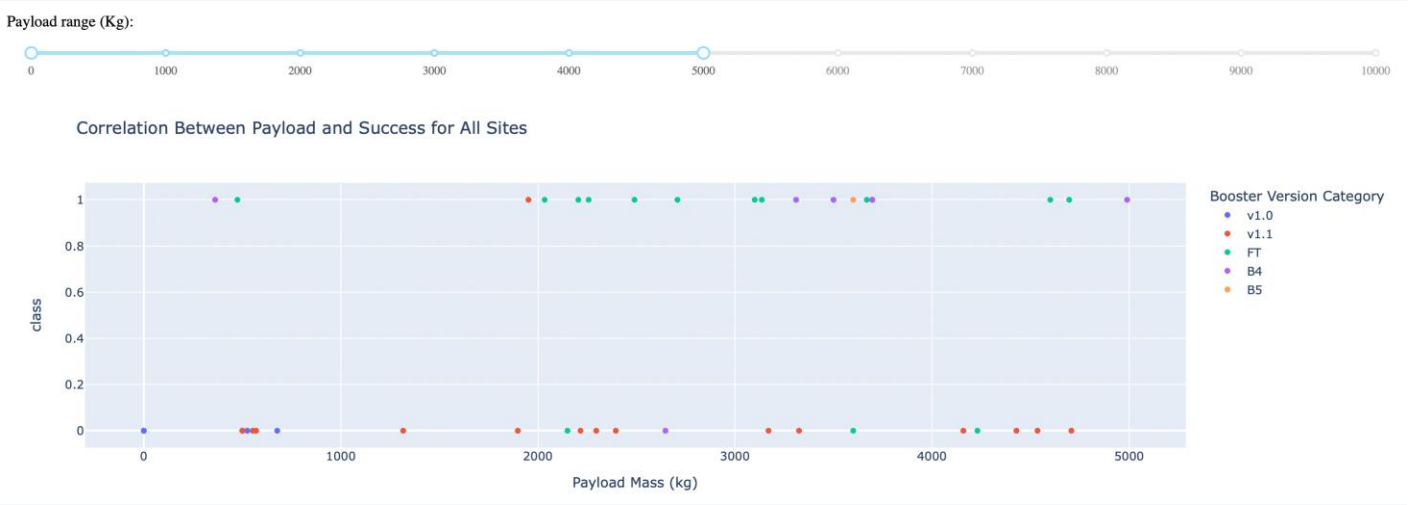
Launch Success Ratio for KSC LC-39A

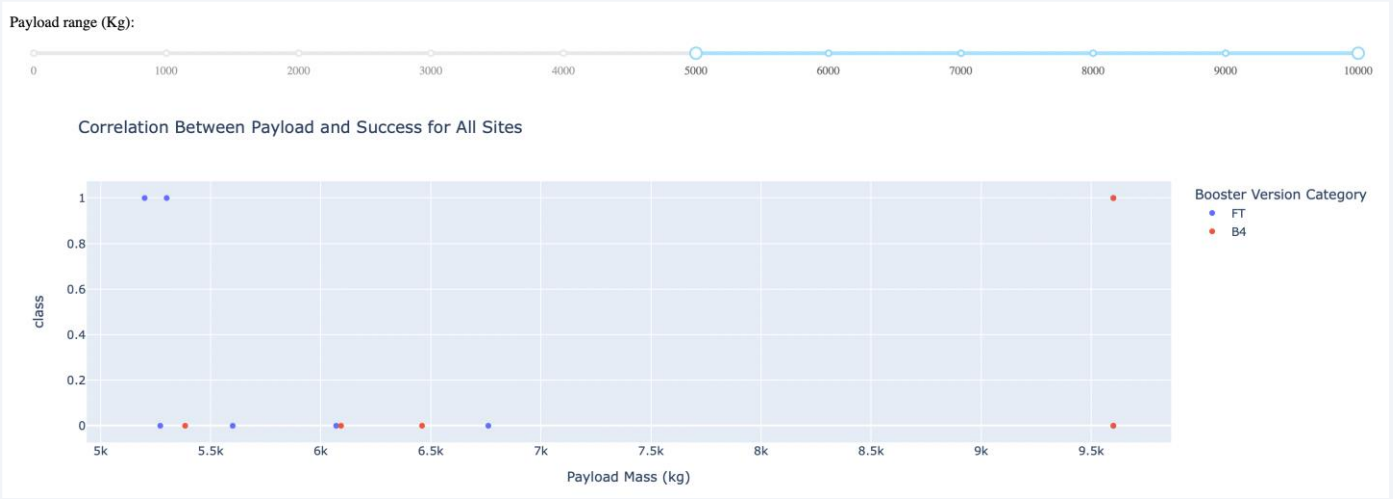


- KSC LC-39A has the highest launch success rate of 76.9%.
- Out of 13 launches, 10 were successful, and 3 resulted in failure.

Payload vs. Launch Outcome

- Payloads under 6,000 kg with FT boosters demonstrate the highest success rates.
- Payloads in the range of 2,000 to 5,500 kg show the highest overall success rates across all sites.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

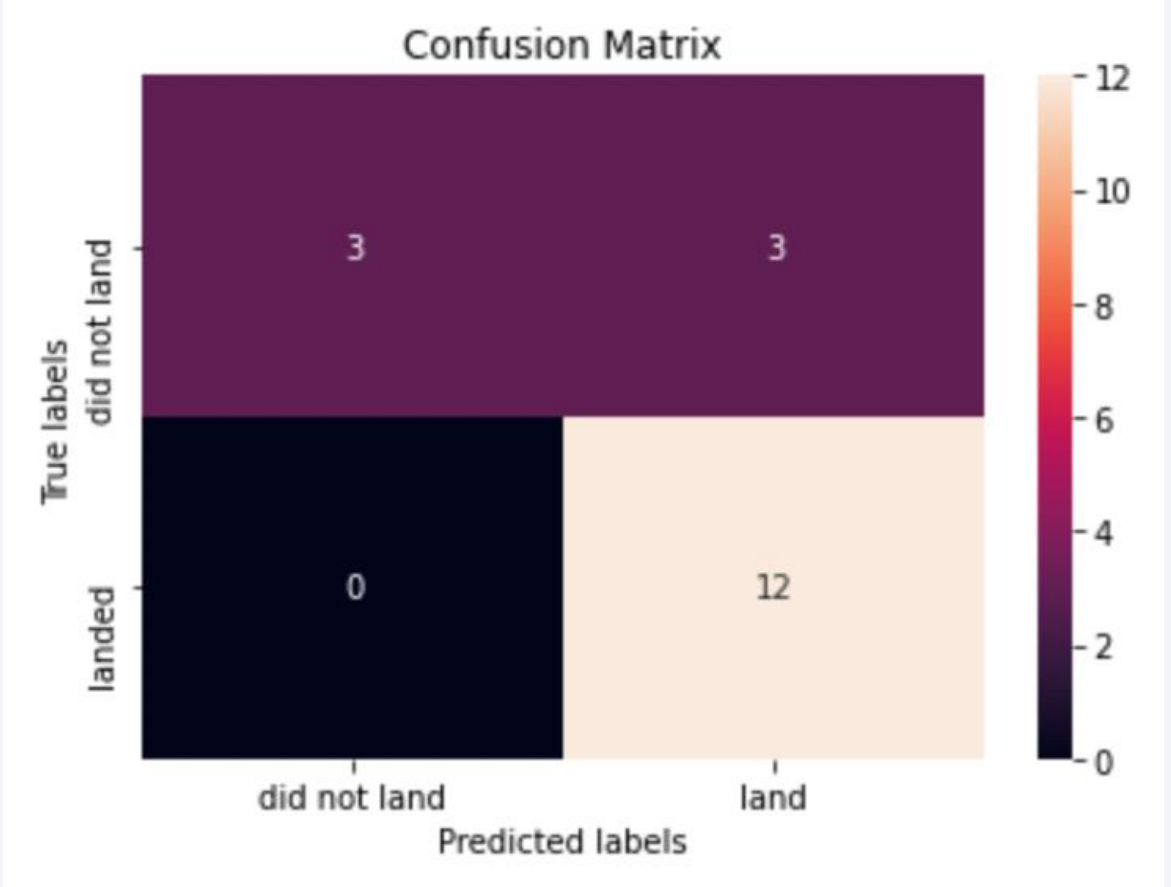
- Four classification models were tested; their accuracies are plotted.
- The Decision Tree Classifier achieved the highest accuracy, exceeding 87%.
- Test Set scores alone cannot confirm the best-performing model due to a small sample size (18 samples).
- Testing on the entire dataset confirmed that the Decision Tree Model is the best, with the highest scores and accuracy.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

- The confusion matrix shows that logistic regression effectively distinguishes between classes but struggles with false positives.

- The Decision Tree Classifier demonstrates its accuracy with a high count of true positives and true negatives compared to false results.



Conclusions

- The Decision Tree Model is the most effective for this dataset.
- Launches with lower payload mass yield better results.
- Most launch sites are near the Equator and close to coasts.
- Launch success rates have improved over the years.
- KSC LC-39A has the highest launch success rate.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.
- KSC LC-39A is the best launch site.
- Launches over 7,000kg are less risky.
- Successful landings have improved over time with advancements in processes and rockets.
- Decision Tree Classifier can predict successful landings and increase profits.

Appendix

- **Acknowledgements:**
  - Instructors for their course guidance.
  - Coursera for providing the platform and resources.
  - IBM for their contribution to the learning experience.

Thank you!