IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mayank Bhandari

17-11-2023

Outline

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

Executive Summary

- The following methodologies were used to analyze the data:
  - Data collection using web scraping and the SpaceX API
  - Exploratory Data Analysis(EDA), including data wrangling and interactive visual analytics
  - Landing outcome predictions using machine learning
- Summary of all results:
  - Data wrangling and visualization techniques helped identify patterns in the data.
  - EDA helped to identify which features are important for predicting the success landing outcomes.
  - Machine Learning Prediction helped identify the best model for predicting the landing outcomes.

Introduction

- The objective of this project is evaluate the viability of company Space Y to compete with Space X.

- Desirable answers:
  - The best way to estimate the total cost for launches, by predicting the landing outcomes of the first stage of rockets
  - Which is the best site to make launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - The missing values were replaced in the collected data and labels (0/1) were assigned to the landing outcome column.
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

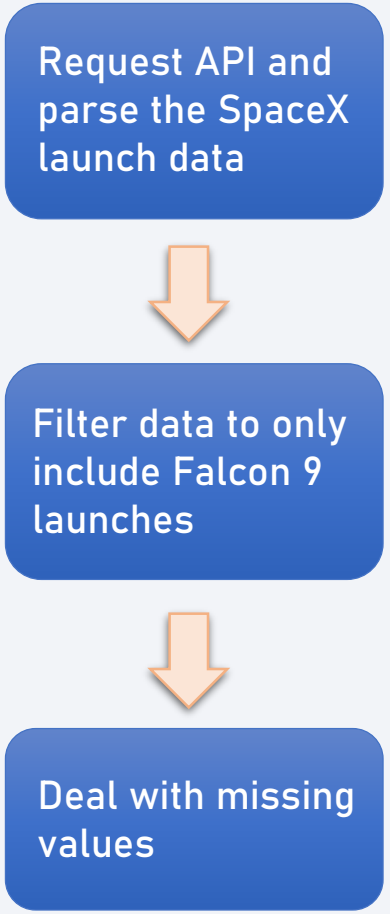
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was normalized, divided into training and testing data sets and evaluated using four different ML models.
  - The accuracy each model was measured using different combination of parameters.

Data Collection

- The datasets were collected from two sources:
  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - Wikipedia using Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

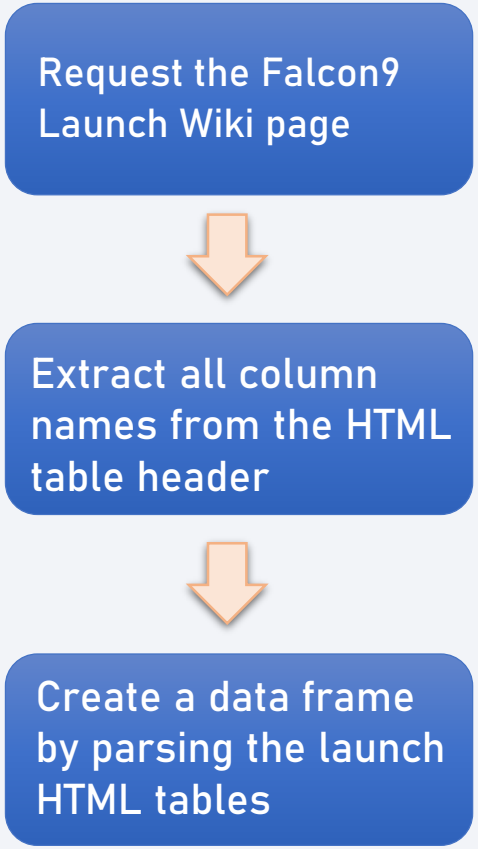
- SpaceX offers a public API from where data can be obtained about the launch of rockets
- This API was used according to the flowchart shown and then data is processed.
- Source Code: https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/42838ce5cbd70151840a1d827ec8b709146e93c0/Data%20Collection%20API.ipynb



Data Collection - Scraping

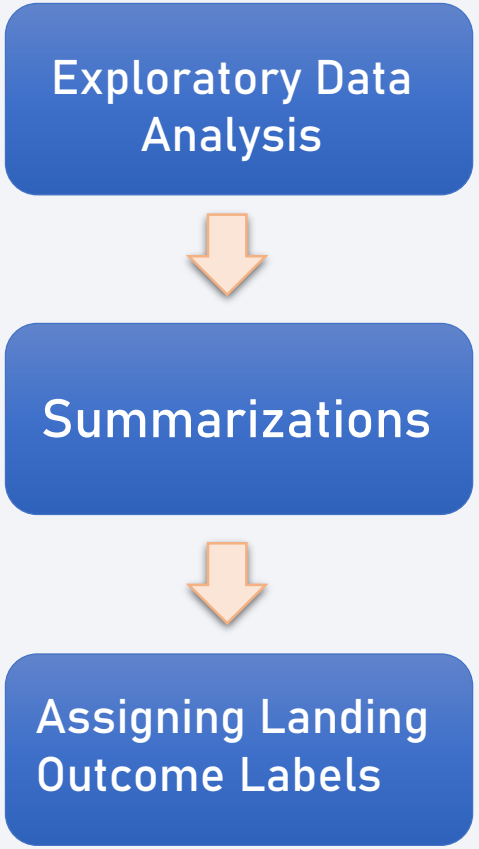
- Data about SpaceX launches is also obtained from Wikipedia.
- Data are downloaded from Wikipedia according to the flowchart and then persisted.

- Source Code: https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/42838ce5cbd70151840a1d827ec8b709146e93c0/Data%20Collection%20with%20Web%20Scraping.ipynb



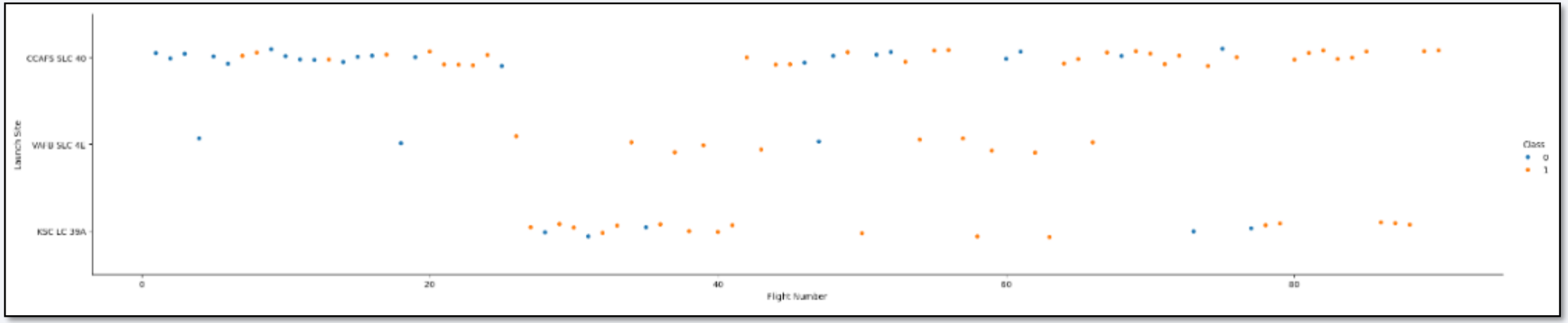
Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.
- Source Code: https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/42838ce5cbd70151840a1d827ec8b709146e93c0/Data%20Wrangling.ipynb



EDA with Data Visualization

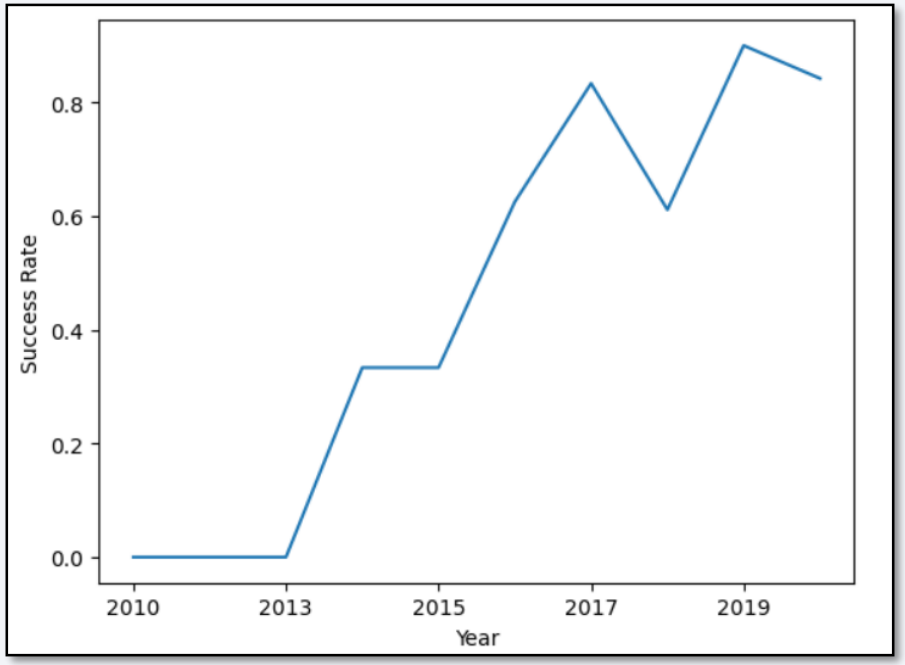
- To explore the data, scatter plots and bar charts were used to visualize the relation between different pairs of features:
  - Launch Site X Flight Number, Payload Mass X Flight Number, Flight Number X Orbit Type, Payload Mass X Launch Site, Payload Mass X Orbit, Orbit X Success Rate



- Source Code: https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/42838ce5cbd70151840a1d827ec8b709146e93c0/EDA%20with%20Visualization.ipynb

EDA with Data Visualization

- To visualize the yearly trend of Success Rate of landing of the first stage Line chart is used.



Source Code: https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/42838ce5cbd70151840a1d827ec8b709146e93c0/EDA%20with%20Visualization.ipynb

EDA with SQL

- The following SQL queries were 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 boosters launched by NASA (CRS)
  - Average 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, launch site names and month for 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.
- Source code: https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/3b3933b8b64c039472ff160ea6ce8ca8f1e692c0/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - 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.

- **Source Code:** https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/42838ce5cbd70151840a1d827ec8b709146e93c0/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

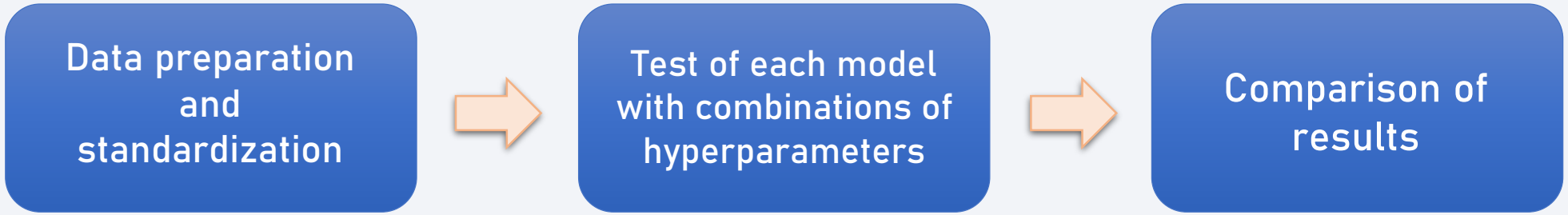
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data:
  - Percentage of launches by site
  - Payload range
- 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.

- **Source code**: https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/42838ce5cbd70151840a1d827ec8b709146e93c0/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: Logistic Regression, Support Vector Machine, Decision Tree and K - Nearest Neighbors



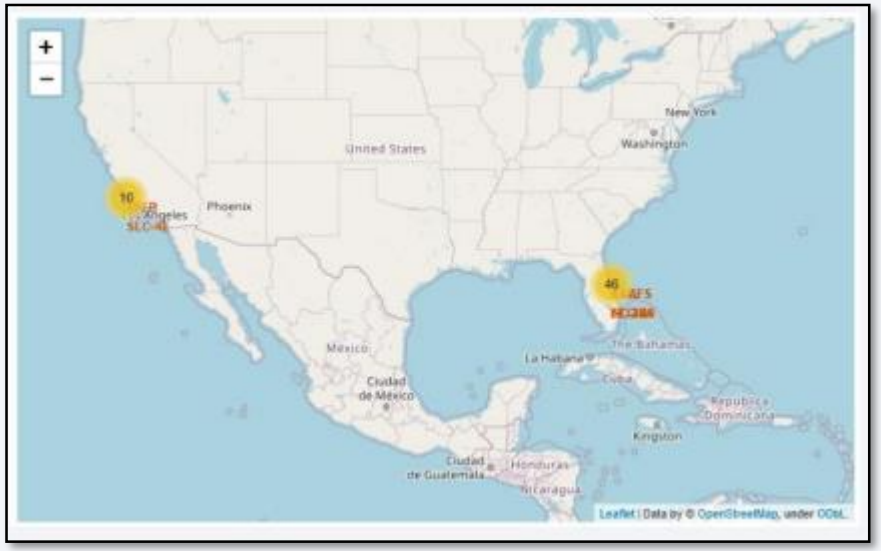
- Source Code: https://github.com/xJoey01x/Data-Science-Capstone-Repo/blob/42838ce5cbd70151840a1d827ec8b709146e93c0/Machine%20Learning%20Prediction.ipynb

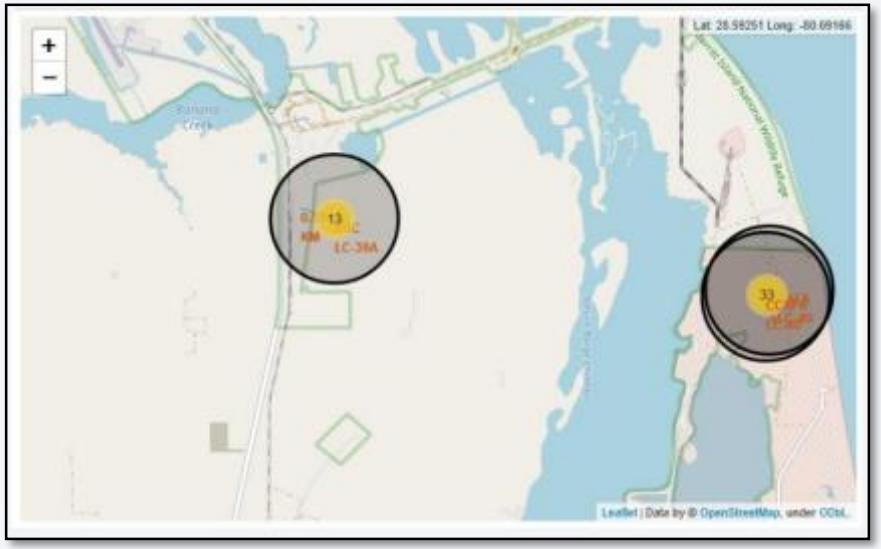
Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites
  - The first launches were done to Space X itself and NASA
  - The average payload of F9 v1.1 booster is 2,928 kg
  - The first success landing outcome happened in 2015 fiver year after the first launch
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
  - Almost 80% of mission outcomes were successful
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
  - The success of landing outcomes became as better as years passed

Results

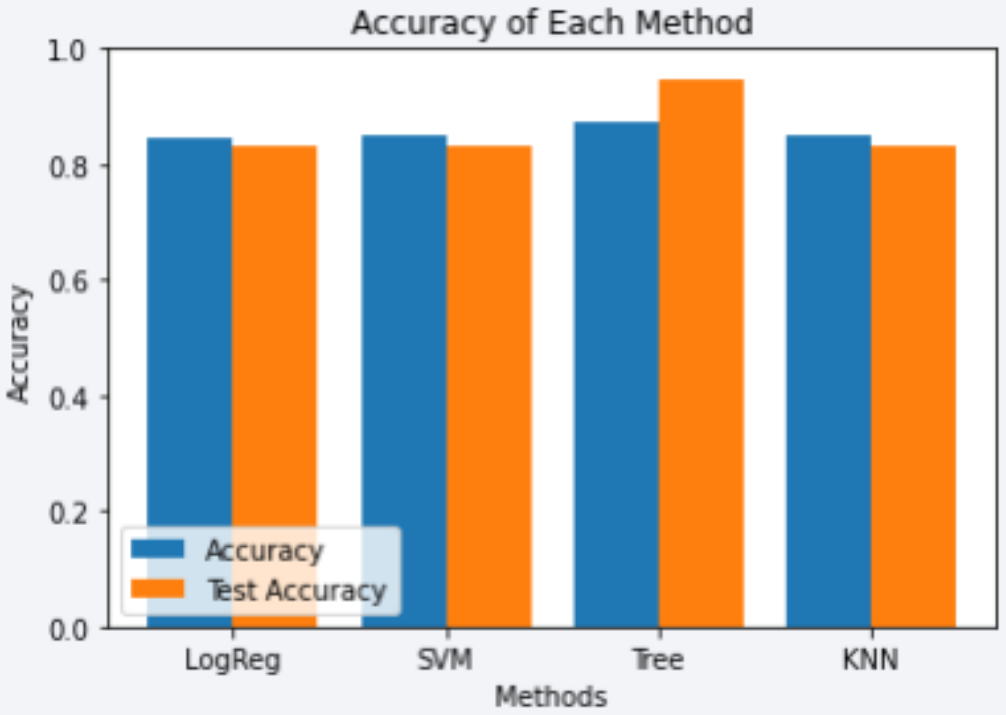
- Using interactive analytics, it was possible to identify that launch sites are built in safety places, for example near the sea, and have a good logistic infrastructure around.
- Most launches happens at east coast launch sites.

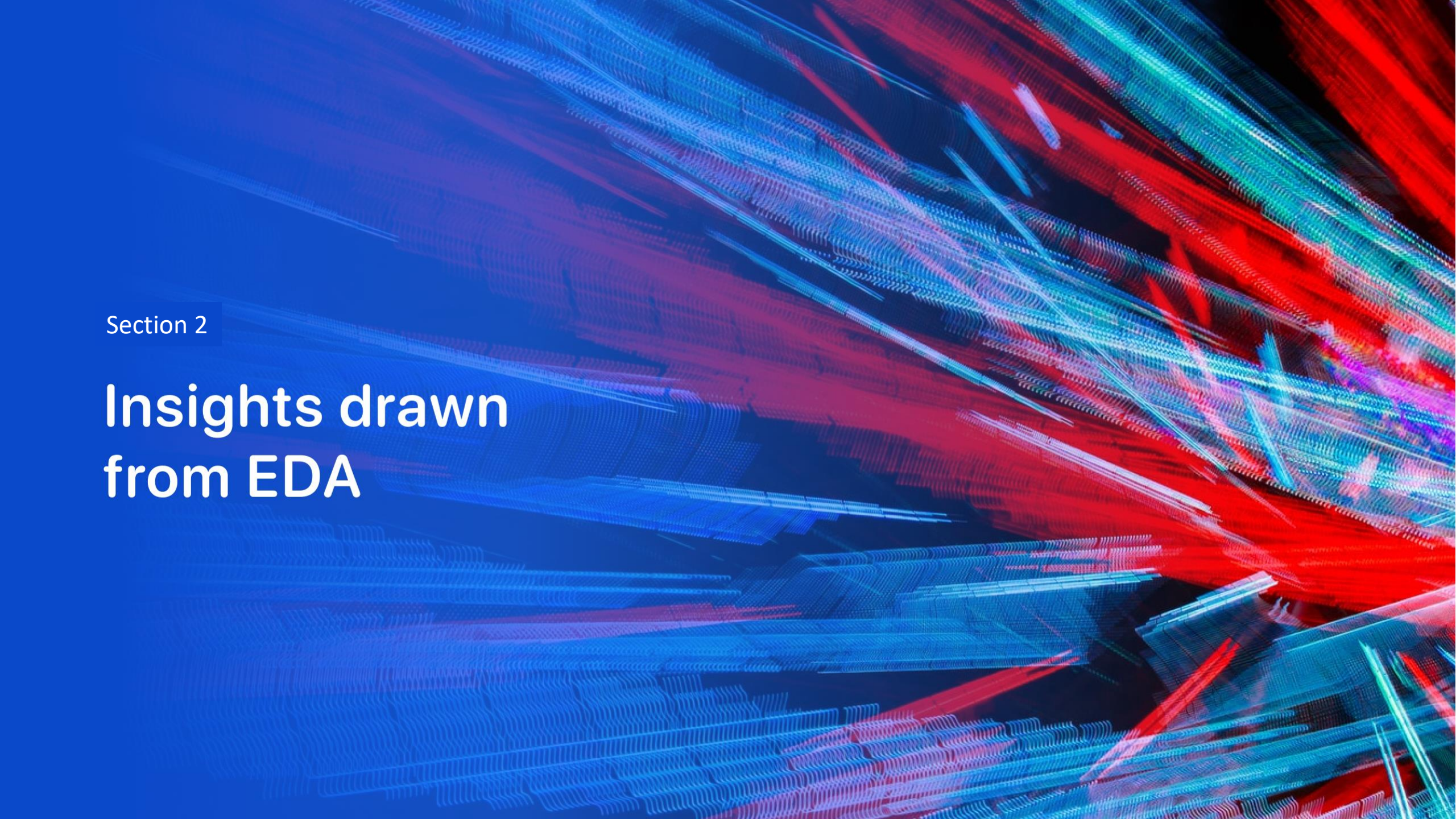




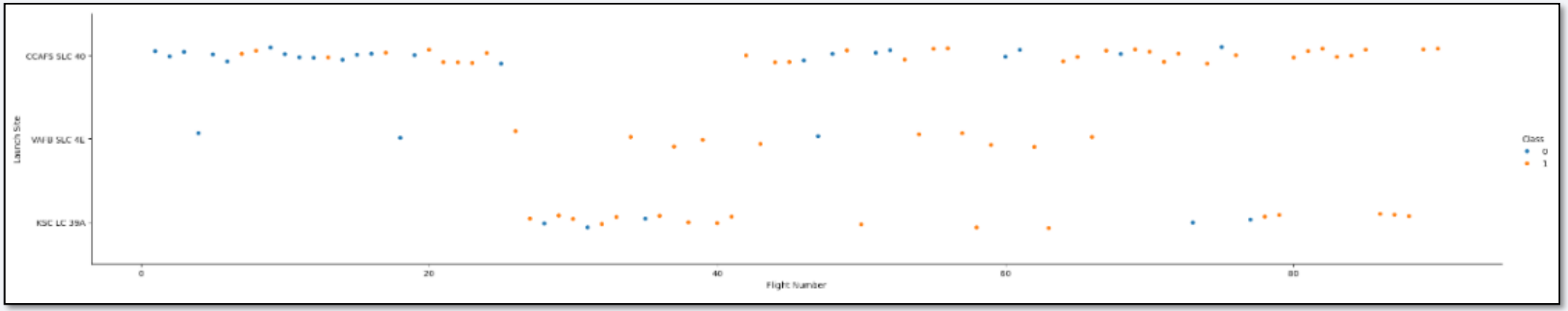
Results

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.



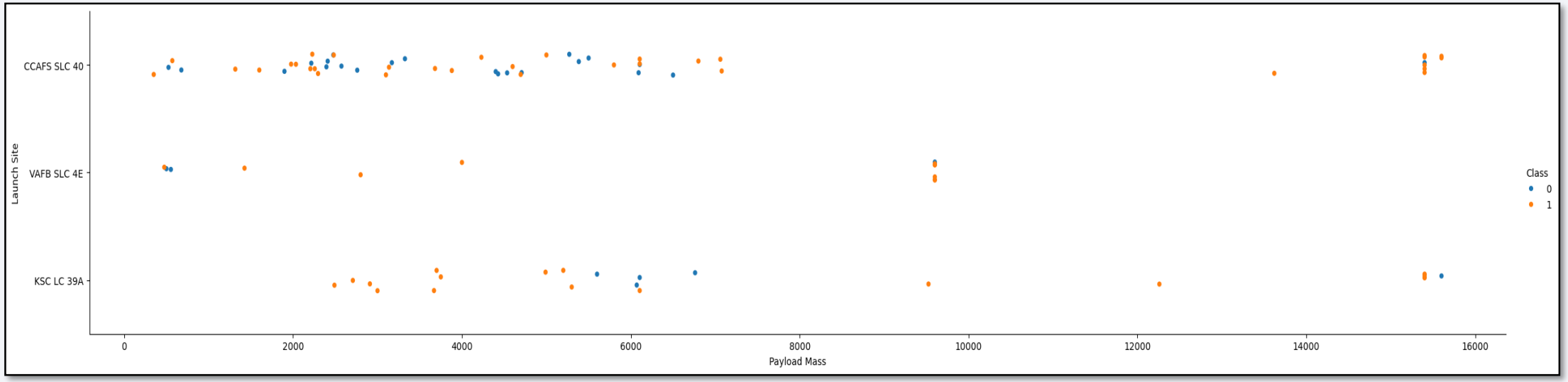

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site is CCAF5 SLC 40, where most of recent launches were successful
- In second place VAFB SLC 4E and third place KSC LC 39A
- It is also evident that the overall success rate improved over time

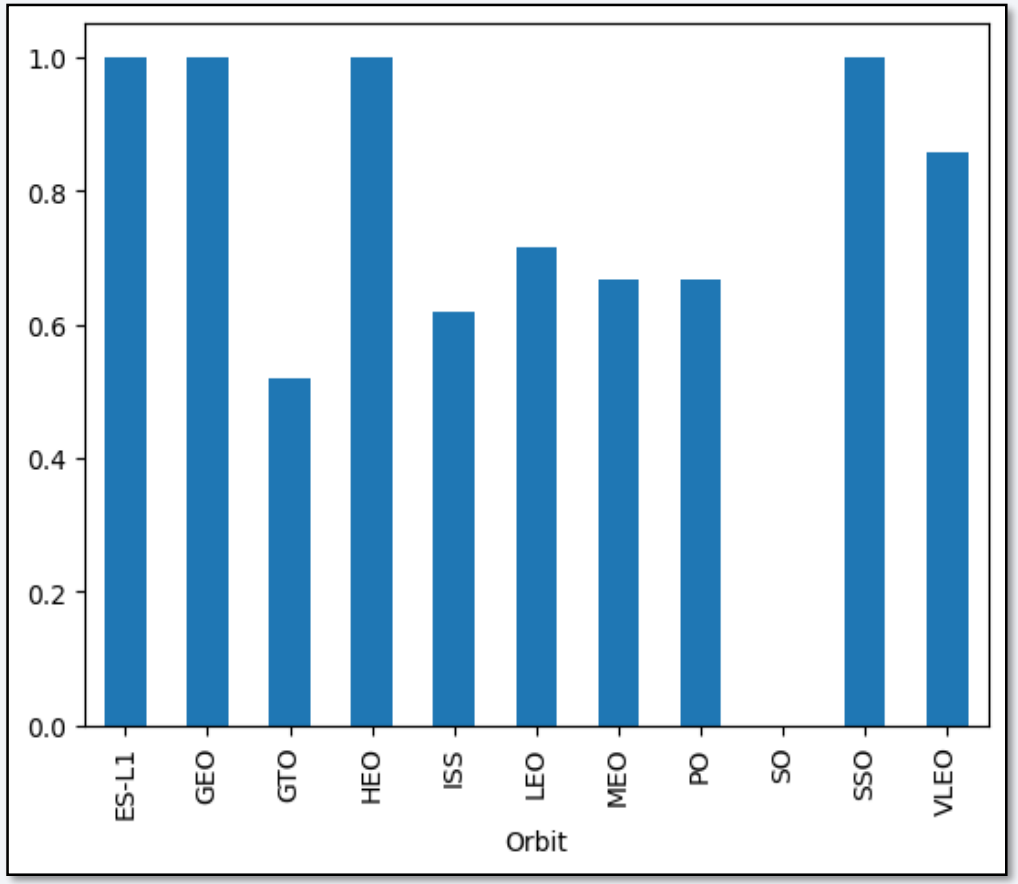
Payload vs. Launch Site



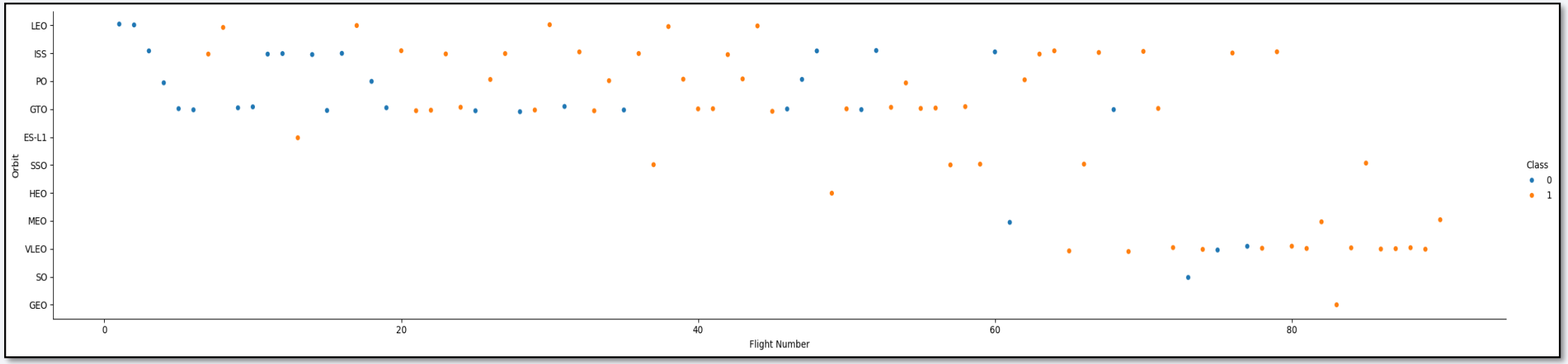
- Payloads over 9,000kg have excellent success rate.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The highest success rates happens to be in orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Followed by:
  - VLEO (around 80%)
  - LFO (around 70%)

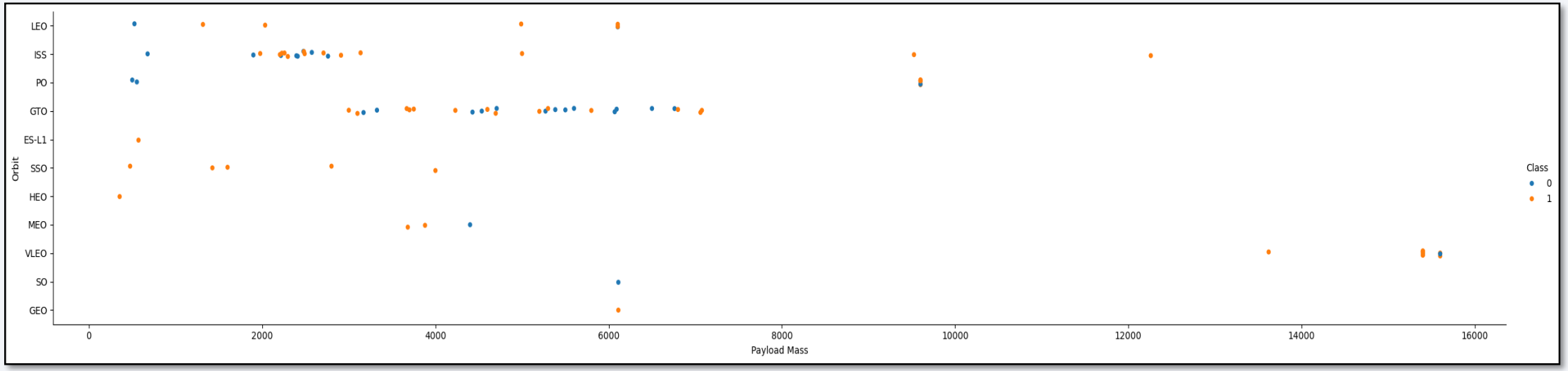


Flight Number vs. Orbit Type



- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.
- The success rate has improved over time for all orbits.

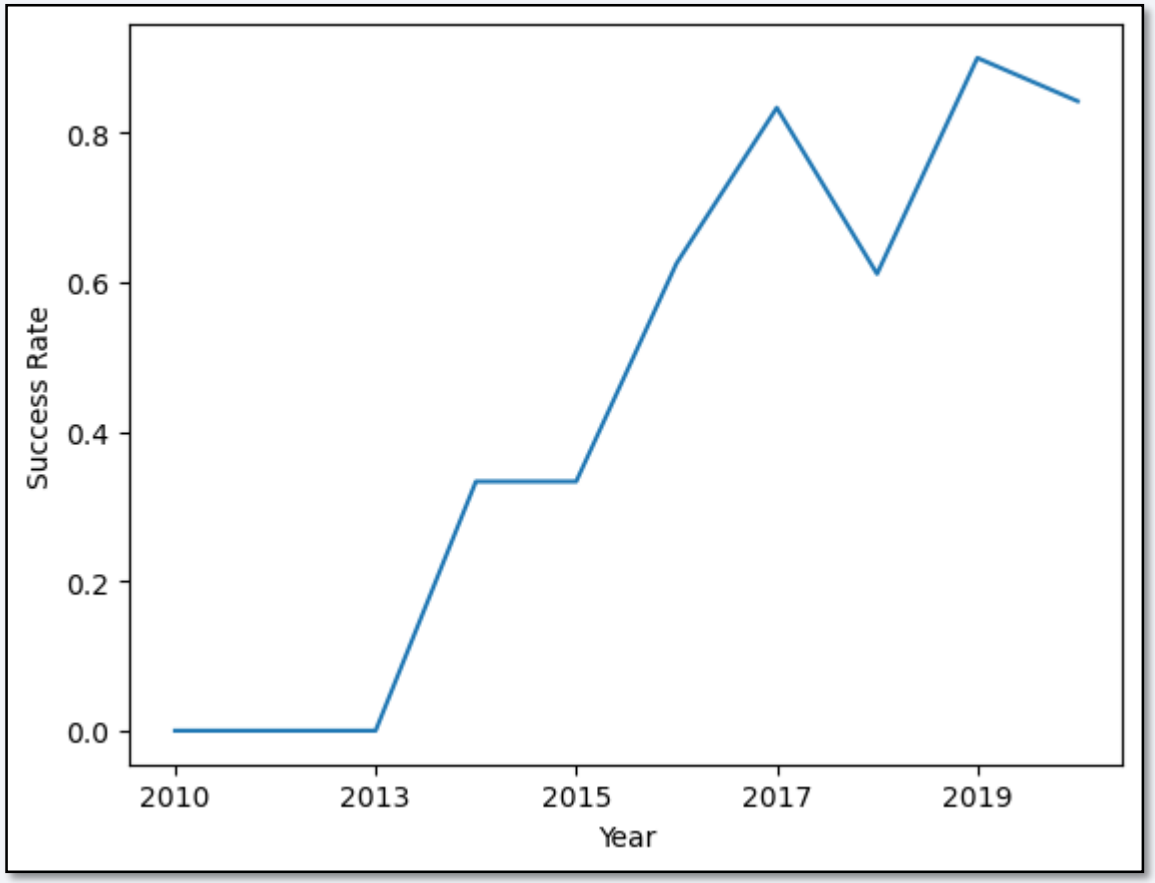
Payload vs. Orbit Type



- Apparently, there is no relation between payload and success rate for the GTO orbit
- ISS orbit has the widest range of payload and a good rate of success
- The VLEO orbit has launches only for higher payload mass.

Launch Success Yearly Trend

- Success rate started increasing exponentially in the year 2013
- There seems to be an abnormal dip in the success rate in the year 2018.
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are four launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- The result is obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

- 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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- The result is obtained using the LIKE operator in SQL.

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
45596

- Total payload calculated above is achieved by summing all payloads whose customer is ‘NASA (CRS)’.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Avg Payload (kg)
2928.4

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928.4 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
22/12/2015

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it is possible to identify the first occurrence, that happened on 22/12/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences.

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

- Ranking of all landing outcomes between the date 04/06/2010 and 20/03/2017:

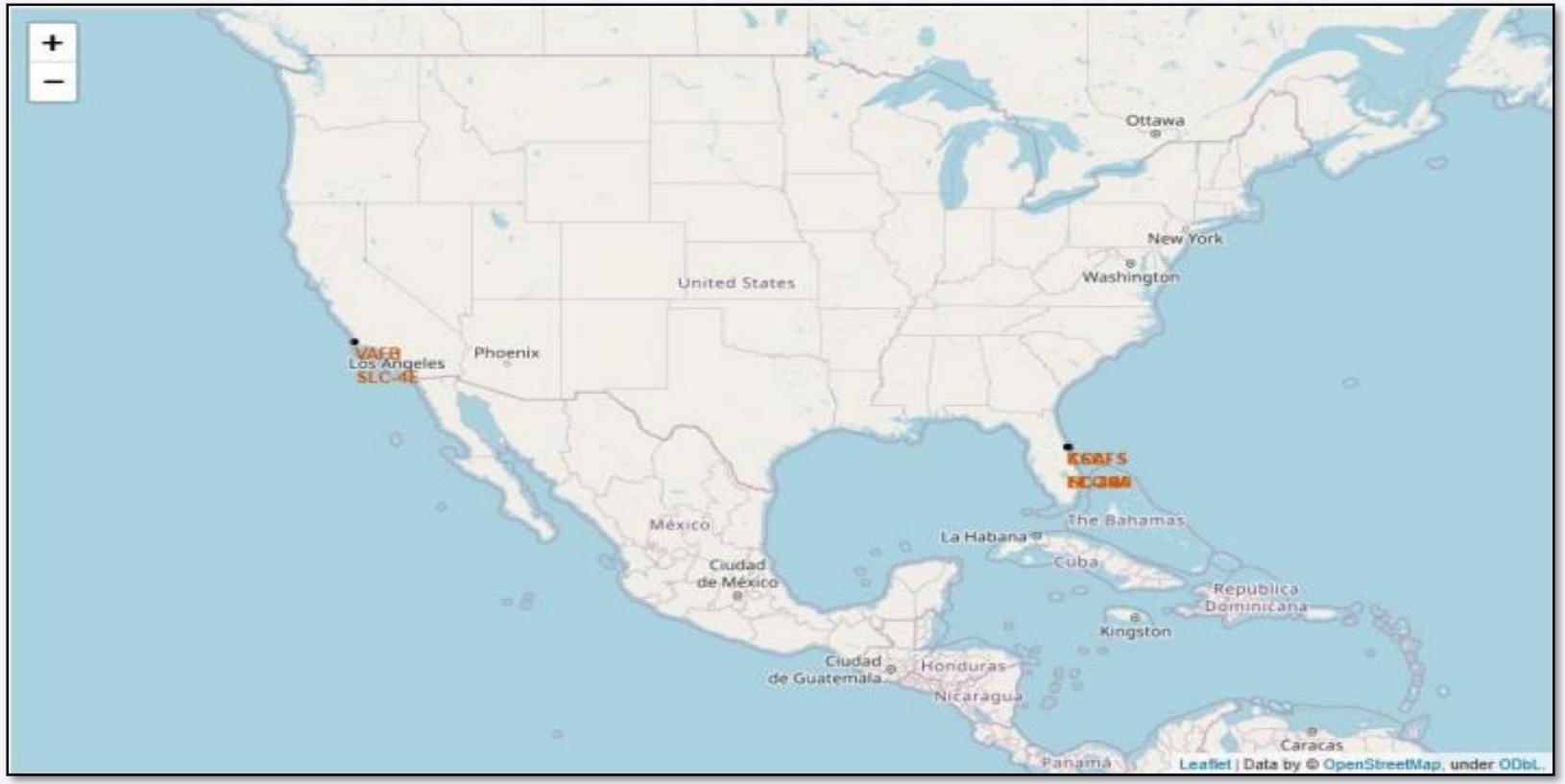
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- This view of data alerts us that “No attempt” must be taken in account.

Section 3

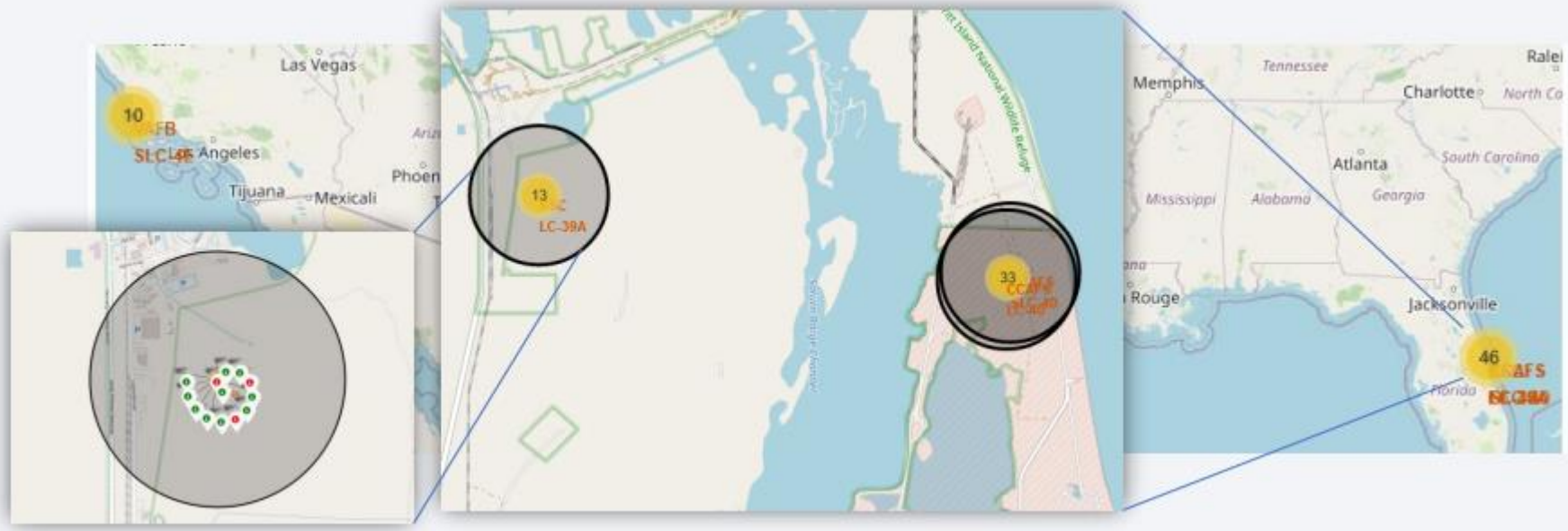
Launch Sites Proximities Analysis

Launch Sites



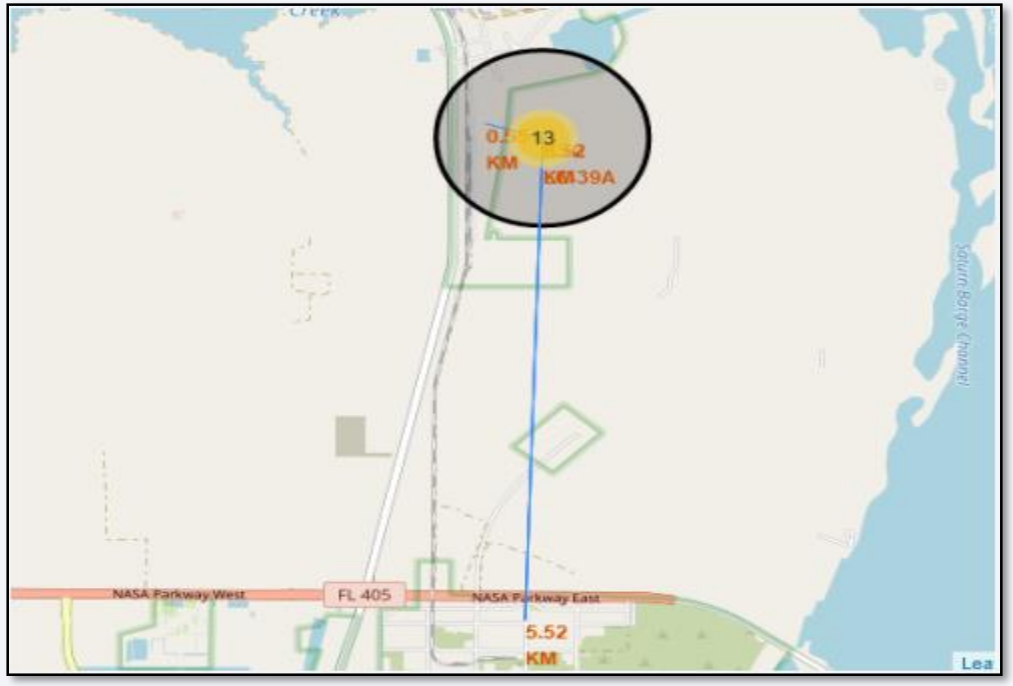
- The launch sites are located near the sea, probably for safety reasons, but they are also well connected by roads and railroads.

Launch Outcomes

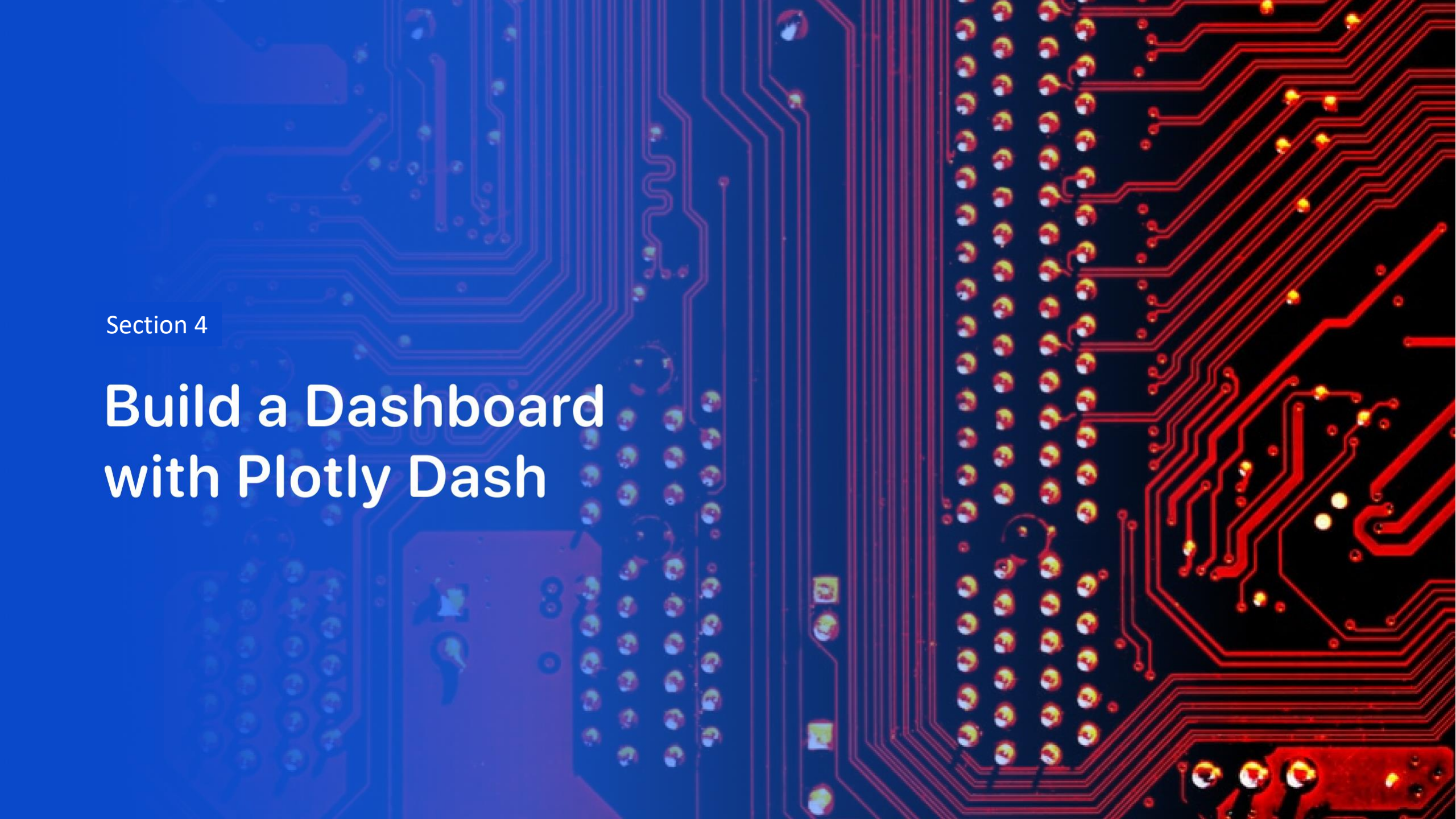


- The above image shows the KSC LC-39A launch site launch outcomes.
- The green markers indicate success and the red markers indicate failure.

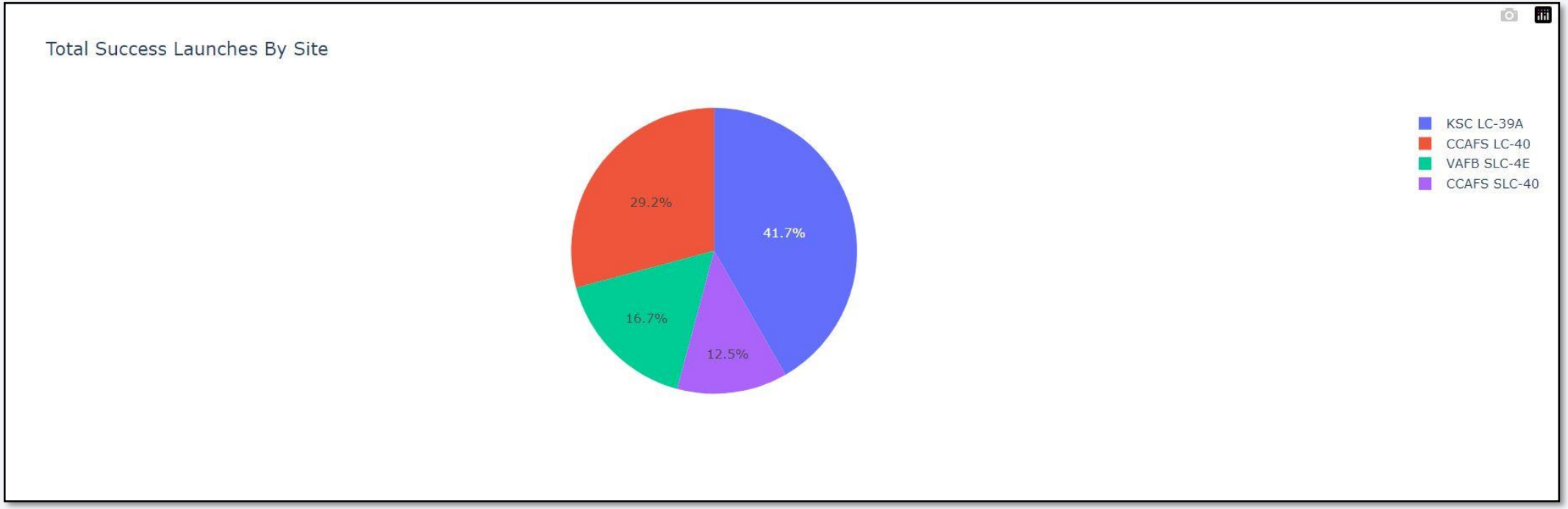
<Folium Map Screenshot 3>



- Launch site KSC LC-39A has good logistic aspects, being near railroad and road and relatively far from inhabited areas.

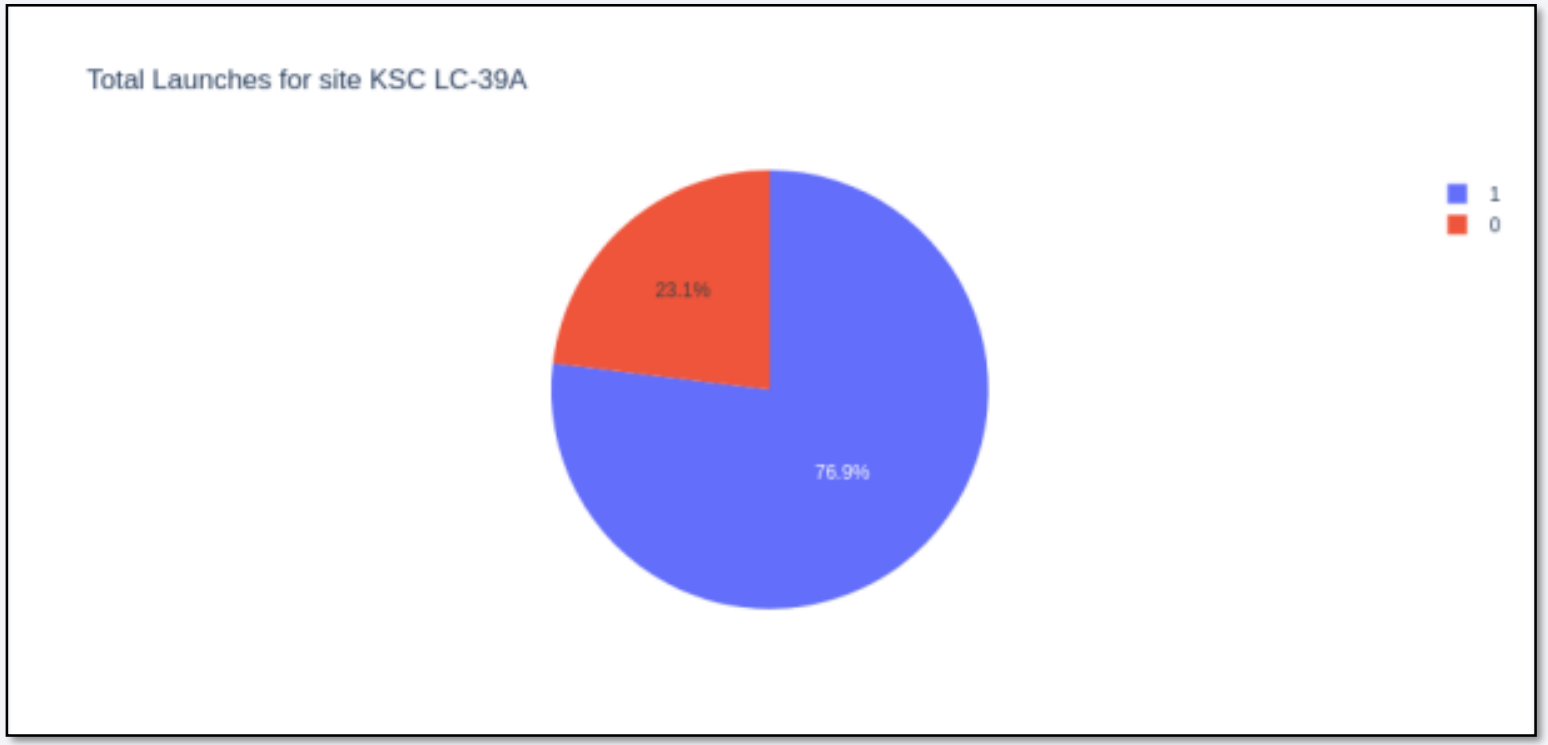

Section 4
Build a Dashboard with Plotly Dash

Success Rate per Site



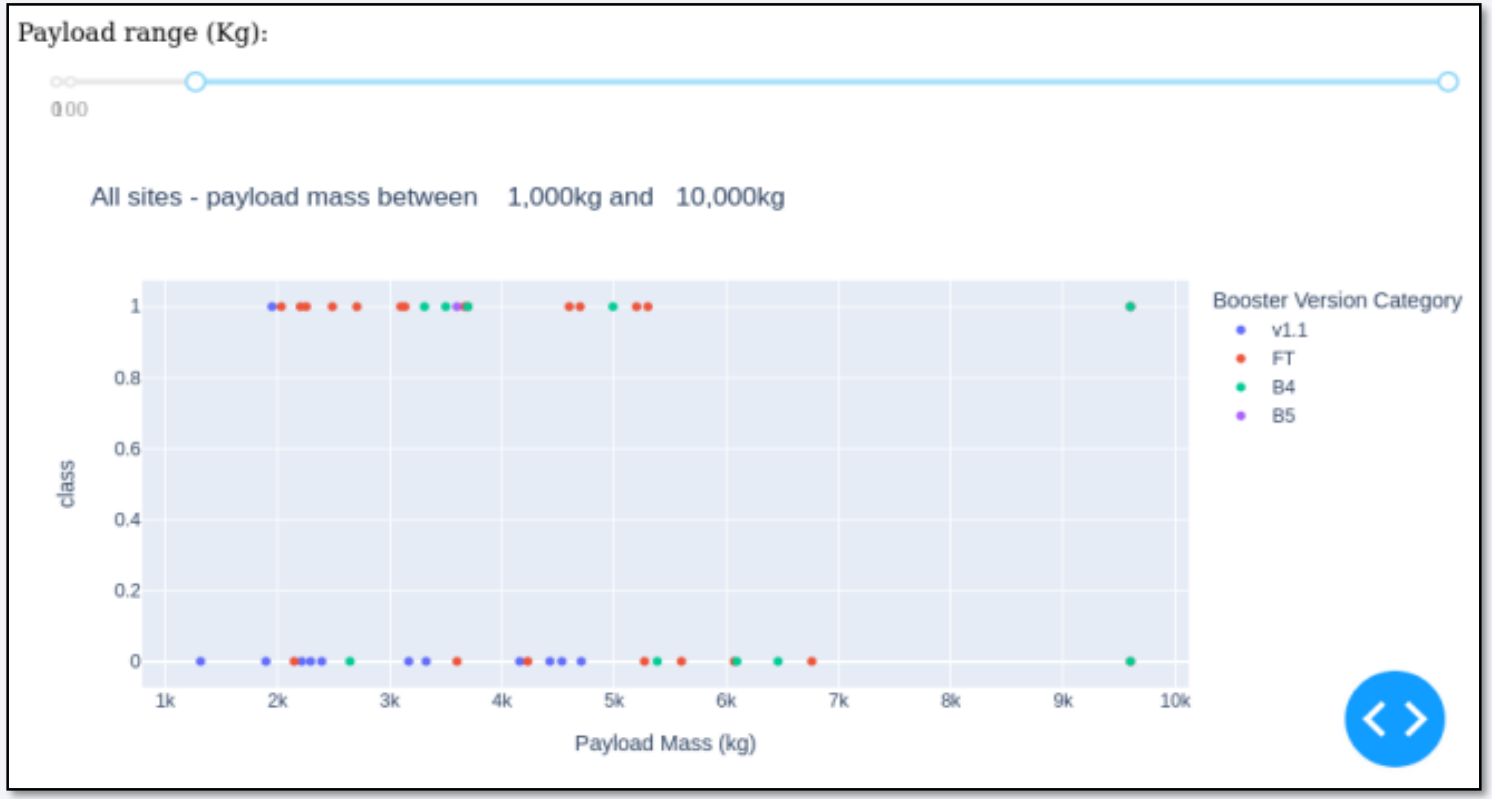
- The KSC LC-39A has the highest success rate among all the launch sites and the CCAF5 SLC-40 has the lowest.

Success Ratio for KSC LC-39A



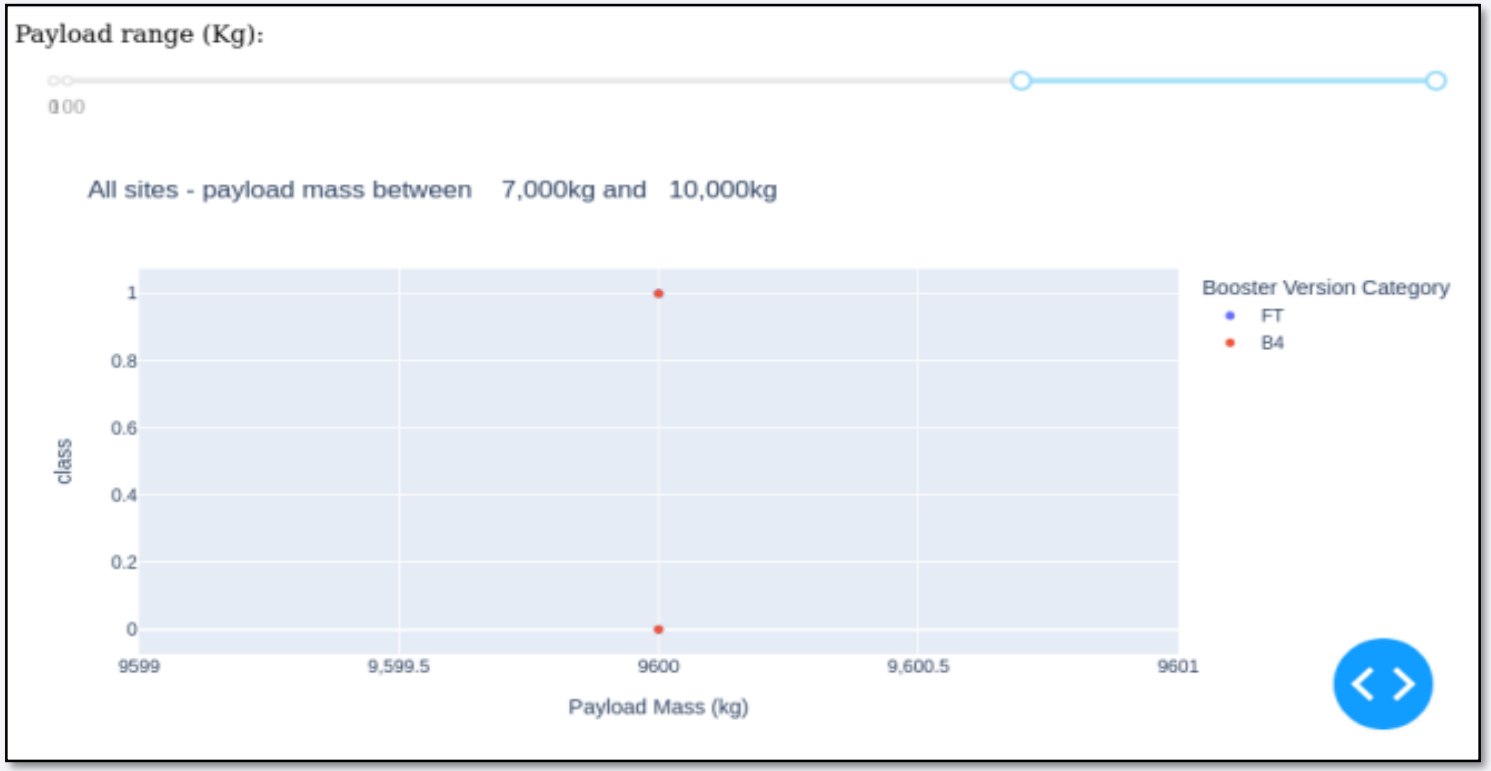
- The KSC LC-39A site has a 76.9% success rate.

Payload vs Launch Outcome



- Payloads under 6,000kg and FT boosters are the most successful combination.

Payload vs Launch Outcome



- There's not enough data to estimate risk of launches over 7,000kg

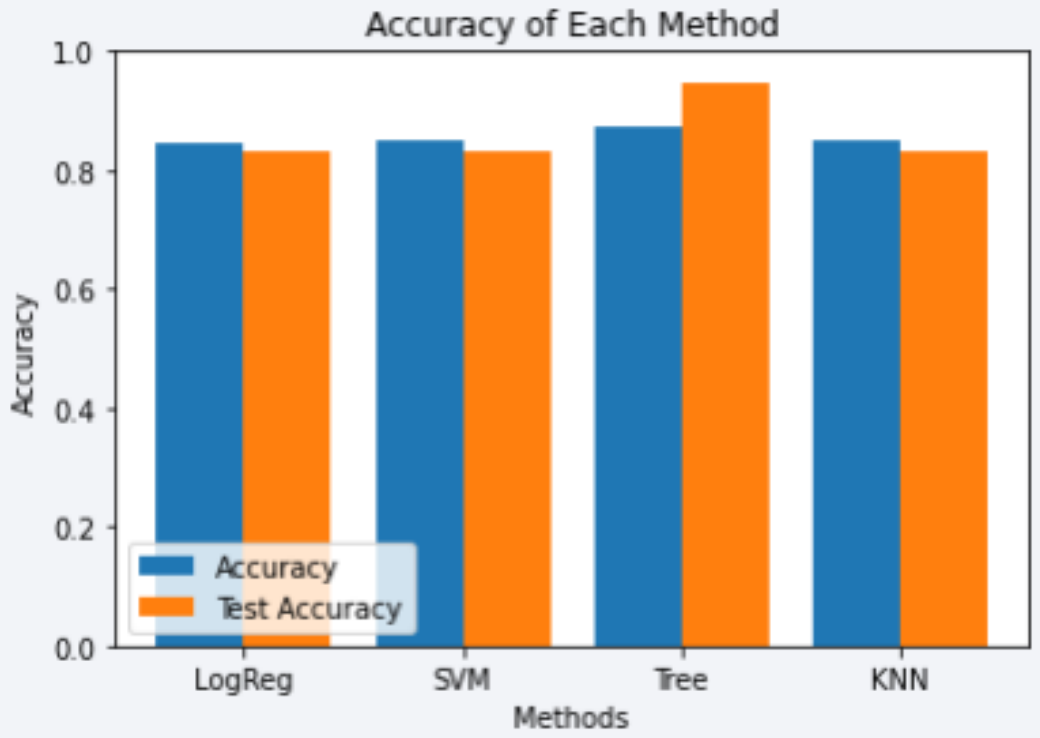
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside

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



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing large numbers of true positive and true negative compared to the false ones.

Conclusions

- Different data sources were analyzed, refining conclusions along the process
- The best launch site is KSC LC-39A
- Launches above 7,000kg are less risky
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets
- Decision Tree Classifier can be used to predict successful landings and increase profits

Appendix

- As an improvement for model tests, it's important to set a value to np.random.seed **variable**
- To generate labels for the landing class, I used a different approach, i.e., landing_class = df['Outcome'].map(lambda x: 0 if x in bad_outcomes else 1**). This uses the list comprehension and lambda function to assign labels (0/1).**

Thank you!