Winning Space Race with Data Science

Adithya S
25-01-2023

Outline

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

Executive Summary

- Data is collected from SpaceX API and by web scraping (BeautifulSoup library).
- Exploratory Data Analysis (EDA) is done using data wrangling, visualization (interactive dashboards) – Folium and Poltly libraries.
- Four Machine Learning models with comparison of their accuracies.

Introduction

- Objective is to evaluate viability of first stage of rocket to land.
- Expected outcomes:
  - Cost of rocket launches, success of first stage of rocket landing.
  - Best location to land first stage of rockets.

Section 1

Methodology

Methodology

Executive Summary

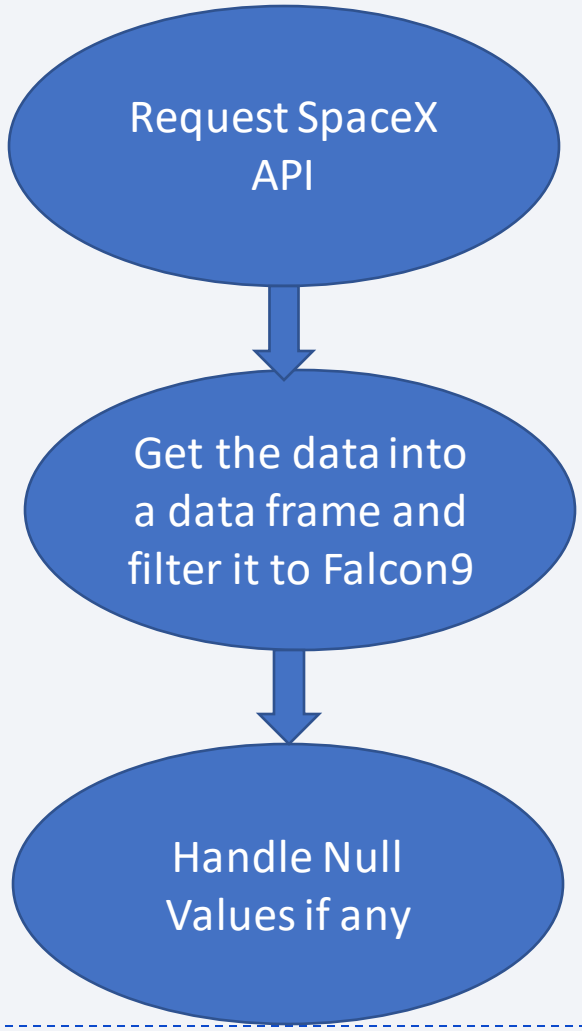
- Data collection methodology:
  - SpaceX launches and landings data is collected from two sources (SpaceX API, Wikipedia)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based data after analyzing the data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using four classification models

Data Collection

- Data is collected from two sources
  1. SpaceX API Link: https://api.spacexdata.com/v4/rockets/
  2. Web Scrapping – BeautifulSoup library. Link: 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
- Source Code: https://github.com/AdithyaSwarna/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection.ipynb

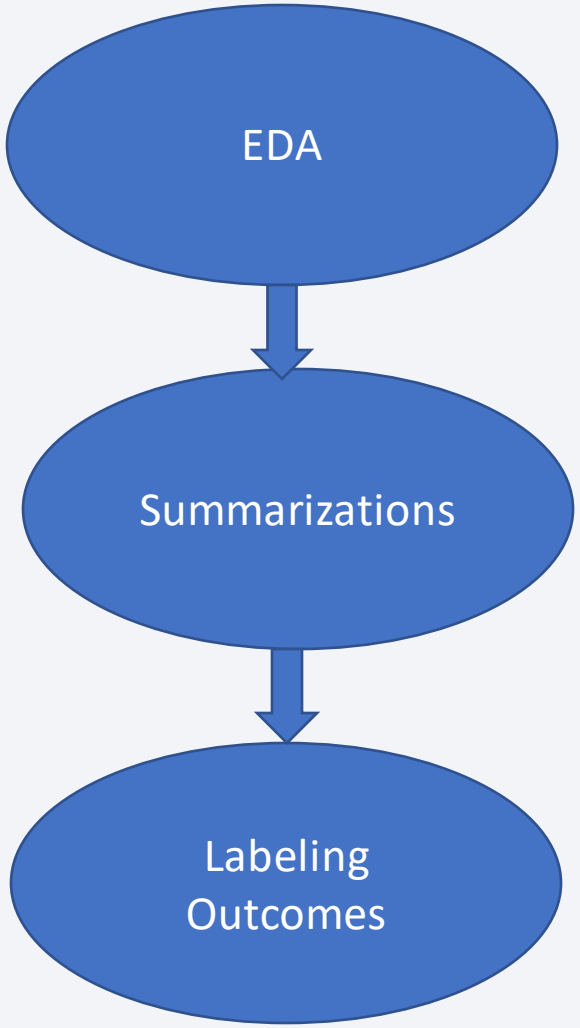


Data Collection - Scraping

- Data collected from wikipedia.
- Source Code: https://github.com/AdithyaSwarna/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- The flowchart presented here showcases the data wrangling steps followed.
- Source Code: https://github.com/AdithyaSwarna/IBM-Applied-Data-Science-Capstone/blob/main/EDA-Data%20Wrangling.ipynb



EDA with Data Visualization

- Bar Charts and Scatter Plots are created as process of EDA.
- Relation between Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Payload Mass, Orbit vs Flight Number, Payload vs Orbit
- Source Code: https://github.com/AdithyaSwarna/IBM-Applied-Data-Science-Capstone/blob/main/EDA-DV.ipynb

EDA with SQL

- The following SQL queries were performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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
- Source Code: https://github.com/AdithyaSwarna/IBM-Applied-Data-Science-Capstone/blob/main/EDA-SQL.ipynb

Build an Interactive Map with Folium

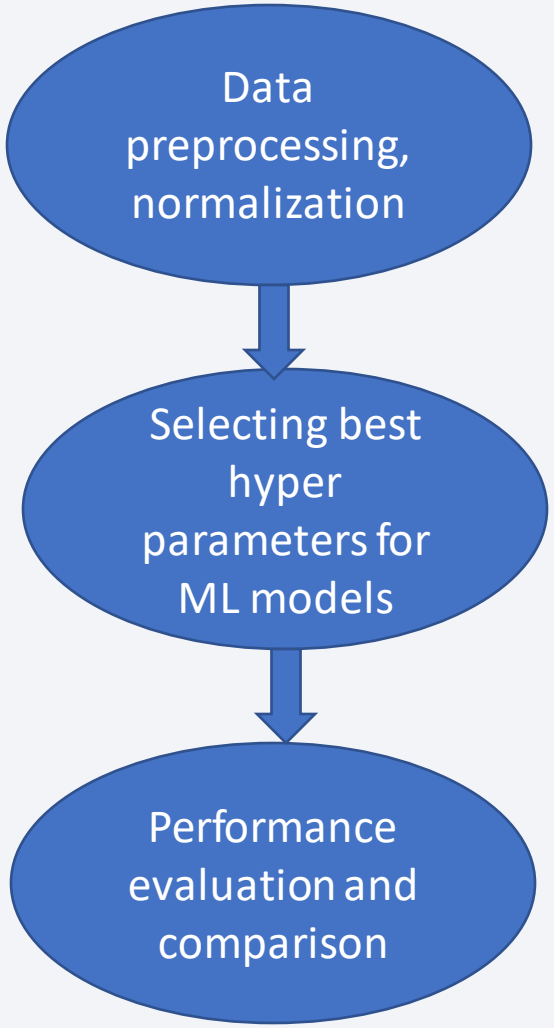
- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates.
- Source Code: https://github.com/AdithyaSwarna/IBM-Applied-Data-Science-Capstone/blob/main/DV-FOLIUM.ipynb
- Note: Github isn't capturing Follium maps.

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/AdithyaSwarna/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four models were trained and accuracies compared: Logistic Regression, Support Vector Machine, Decision Trees and K-Nearest Neighbors.
- Source Code: https://github.com/AdithyaSwarna/IBM-Applied-Data-Science-Capstone/blob/main/ML%20Prediction_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

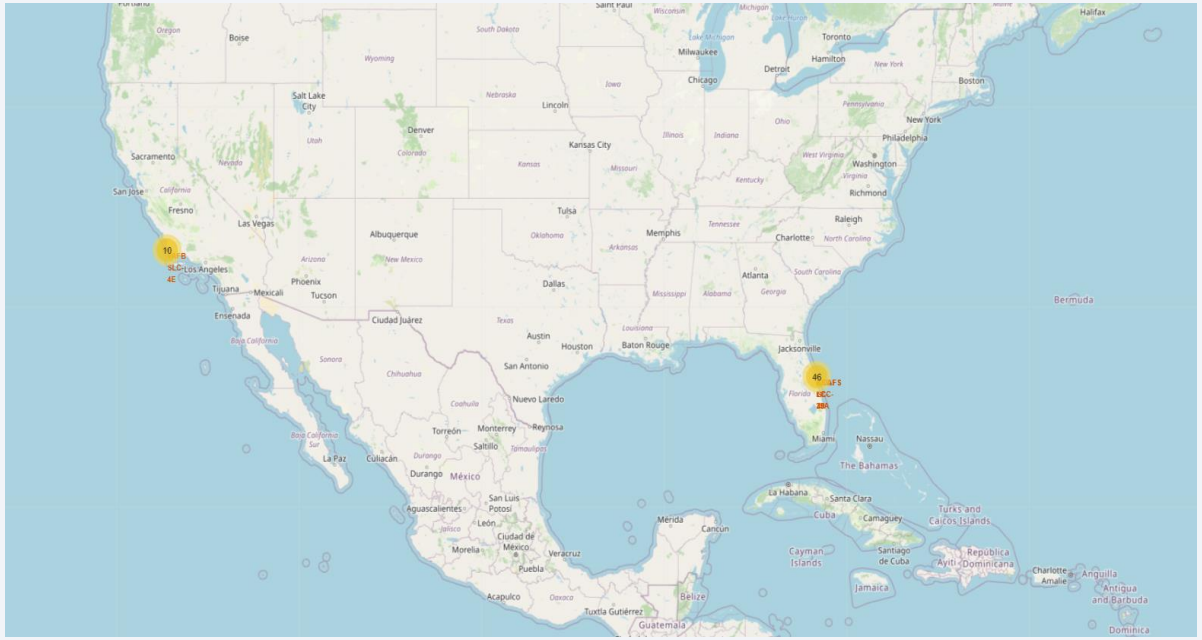


Results

- 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 100% 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 number of landing outcomes became as better as years passed.

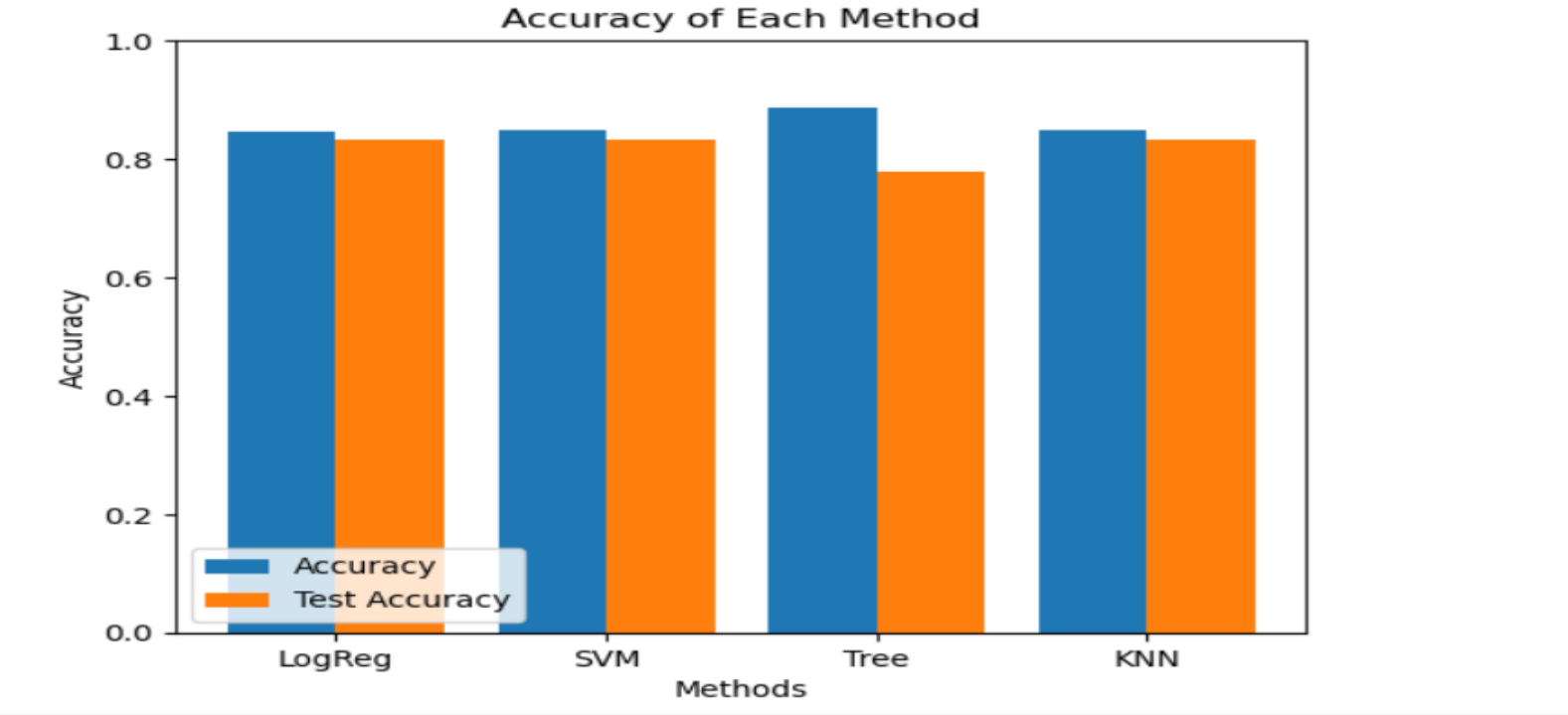
Results

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



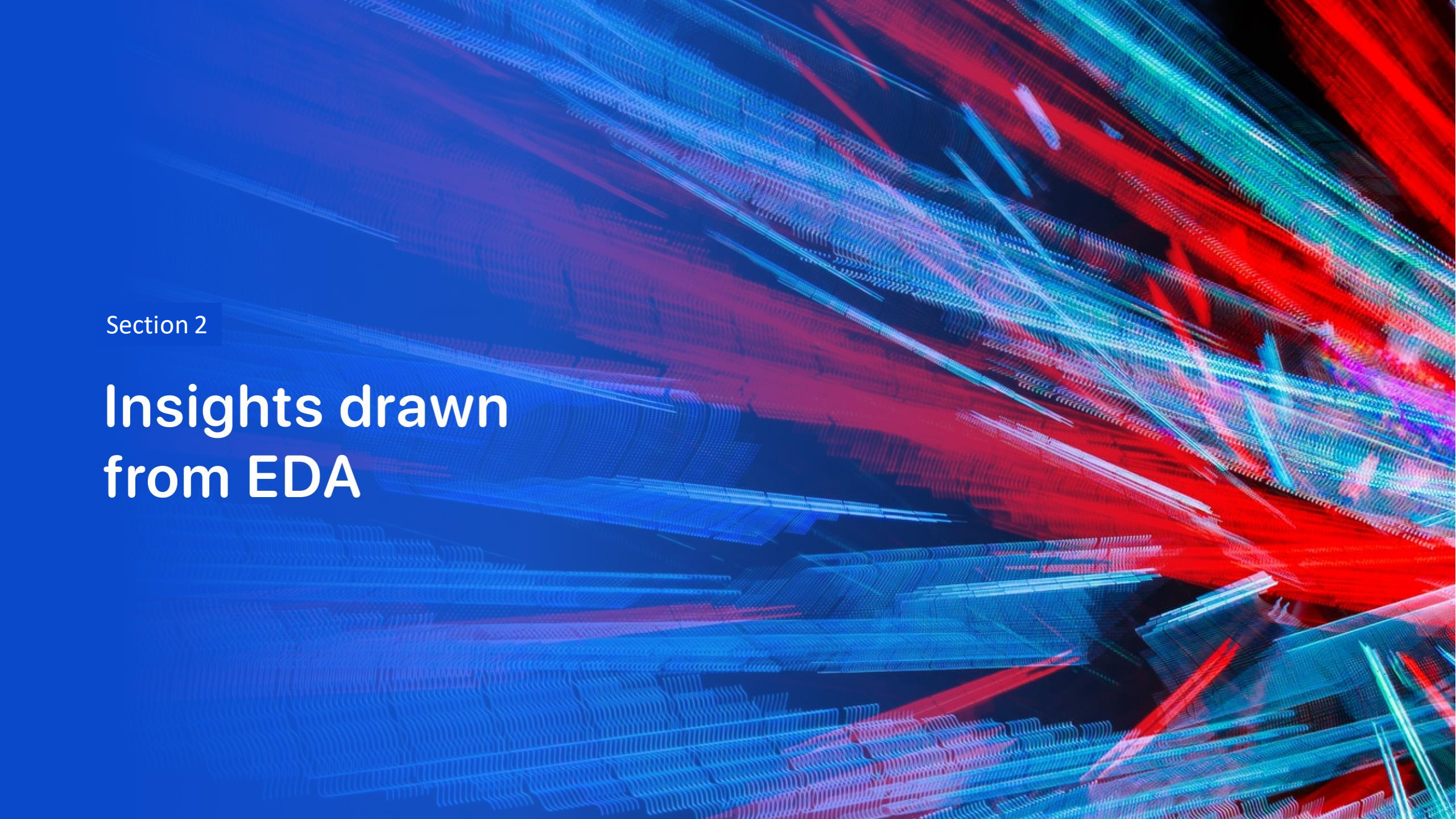
Results

- 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%.

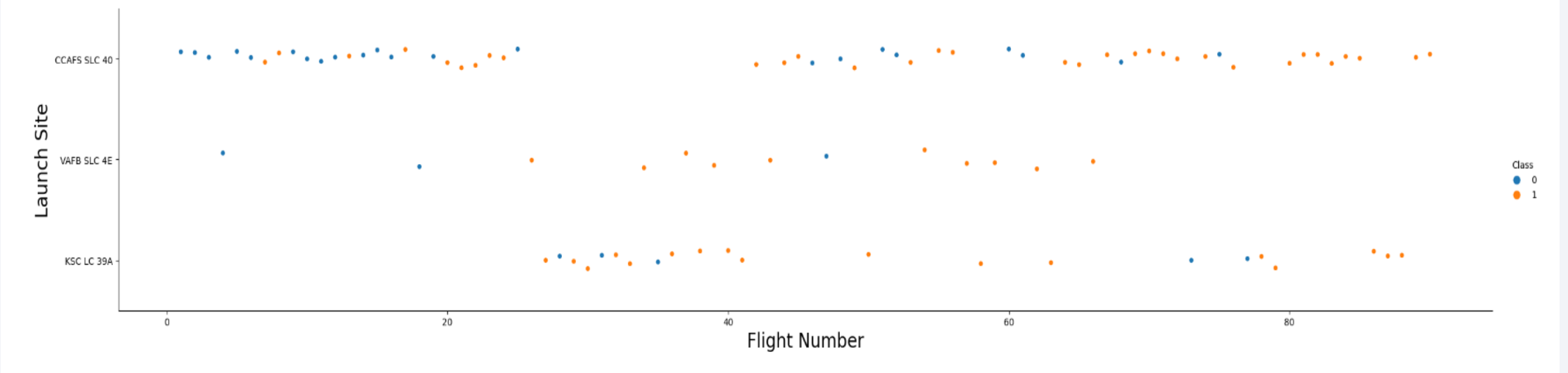


Section 2

Insights drawn from EDA

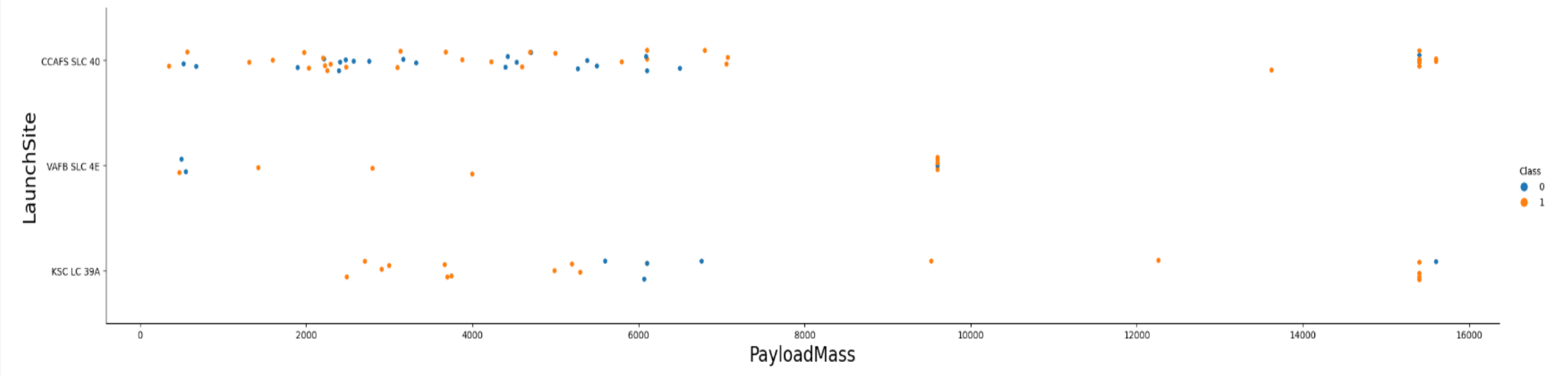


Flight Number vs. Launch Site



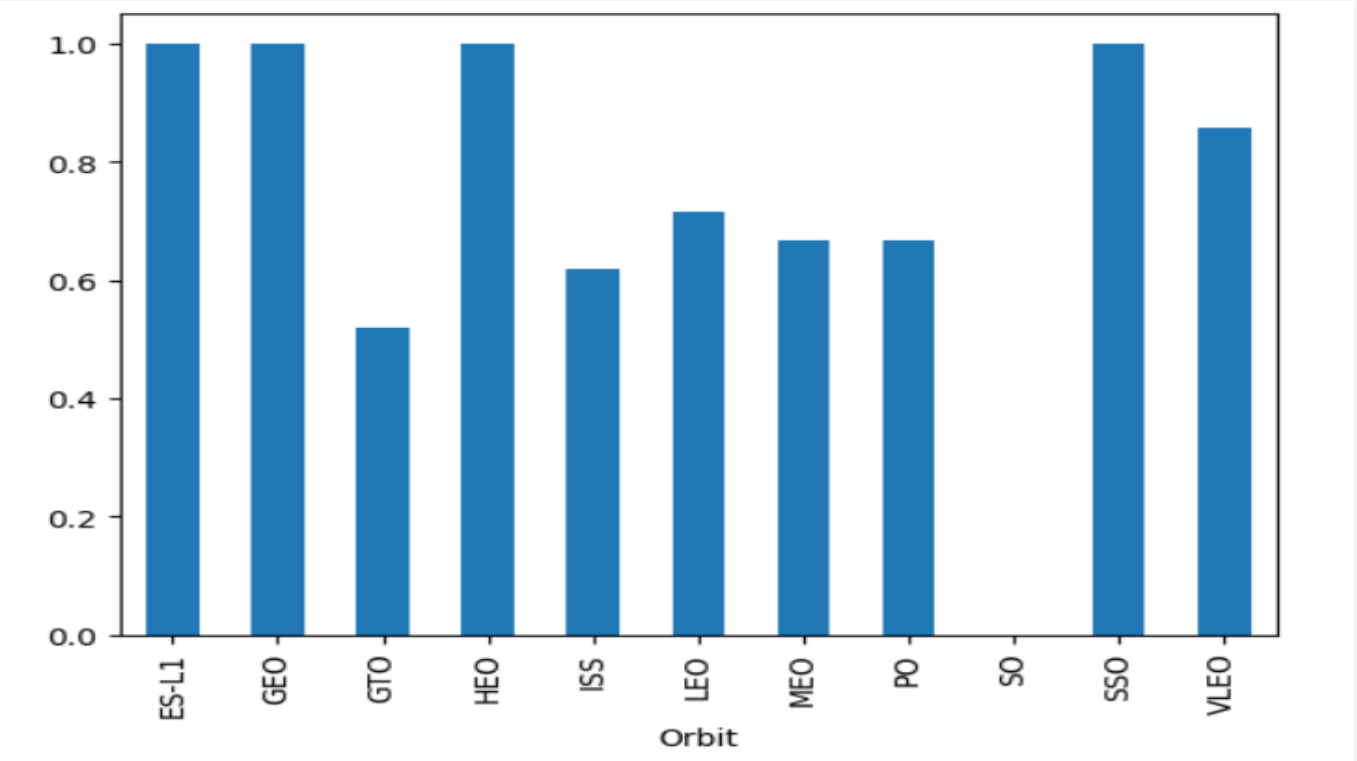
- As flight number increase success rate also improved.
- Based on above plot it can be said that best launch site is CCAFS SLC 40 followed by VAFB SLC 4E and KSC LC 39A.

Payload vs. Launch Site



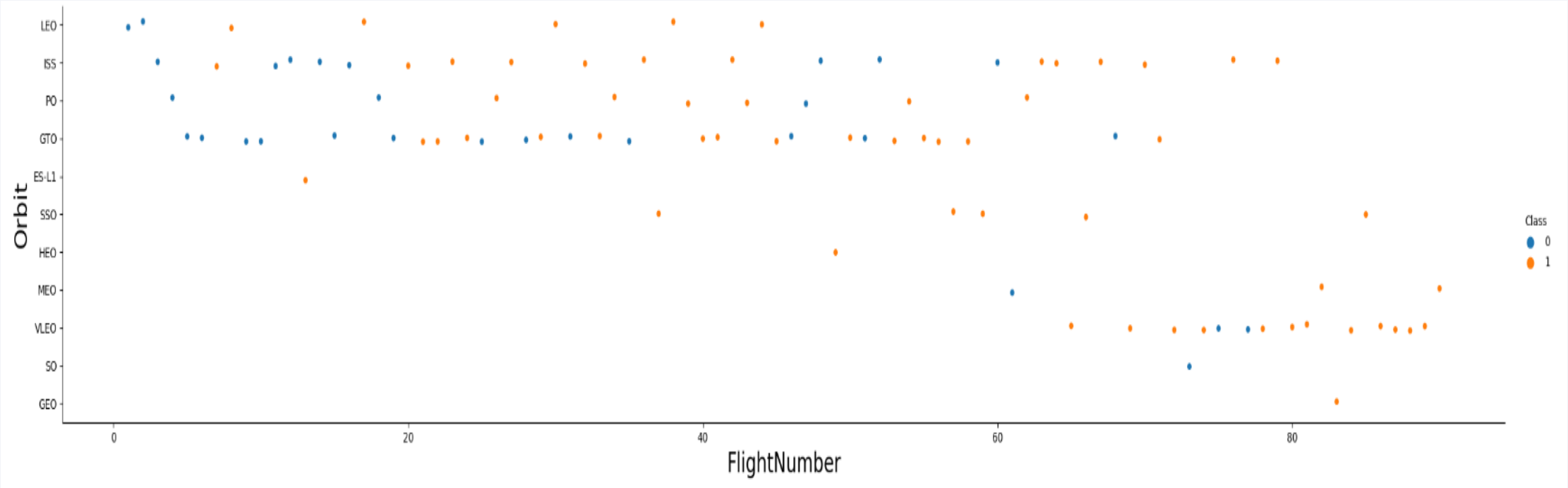
- Payloads over 9,000kg have excellent success rate.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type



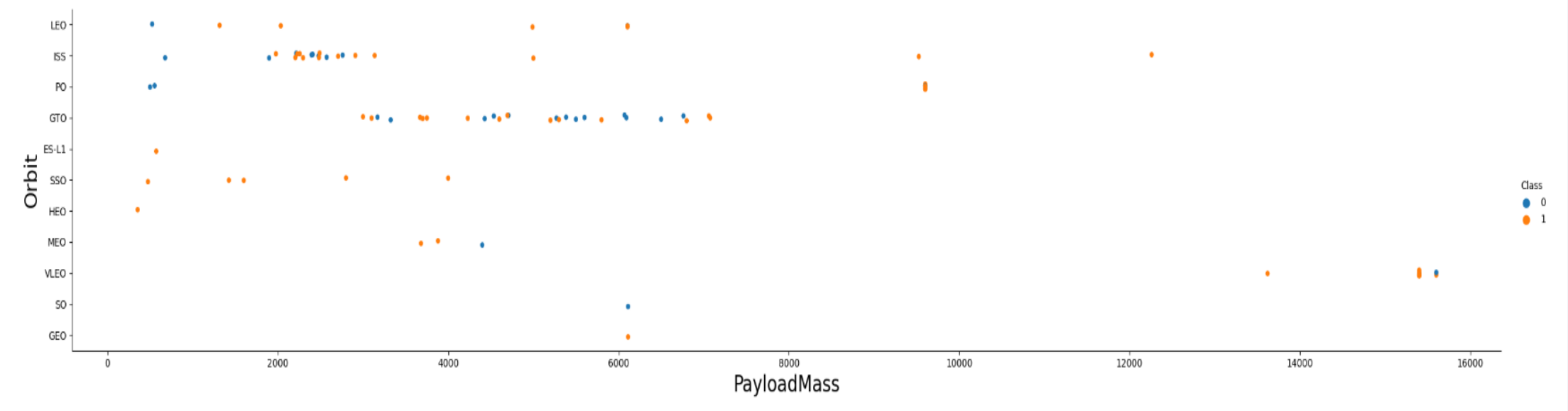
- ES-L1, GEO, HEO, SSO have highest success rates.

Flight Number vs. Orbit Type



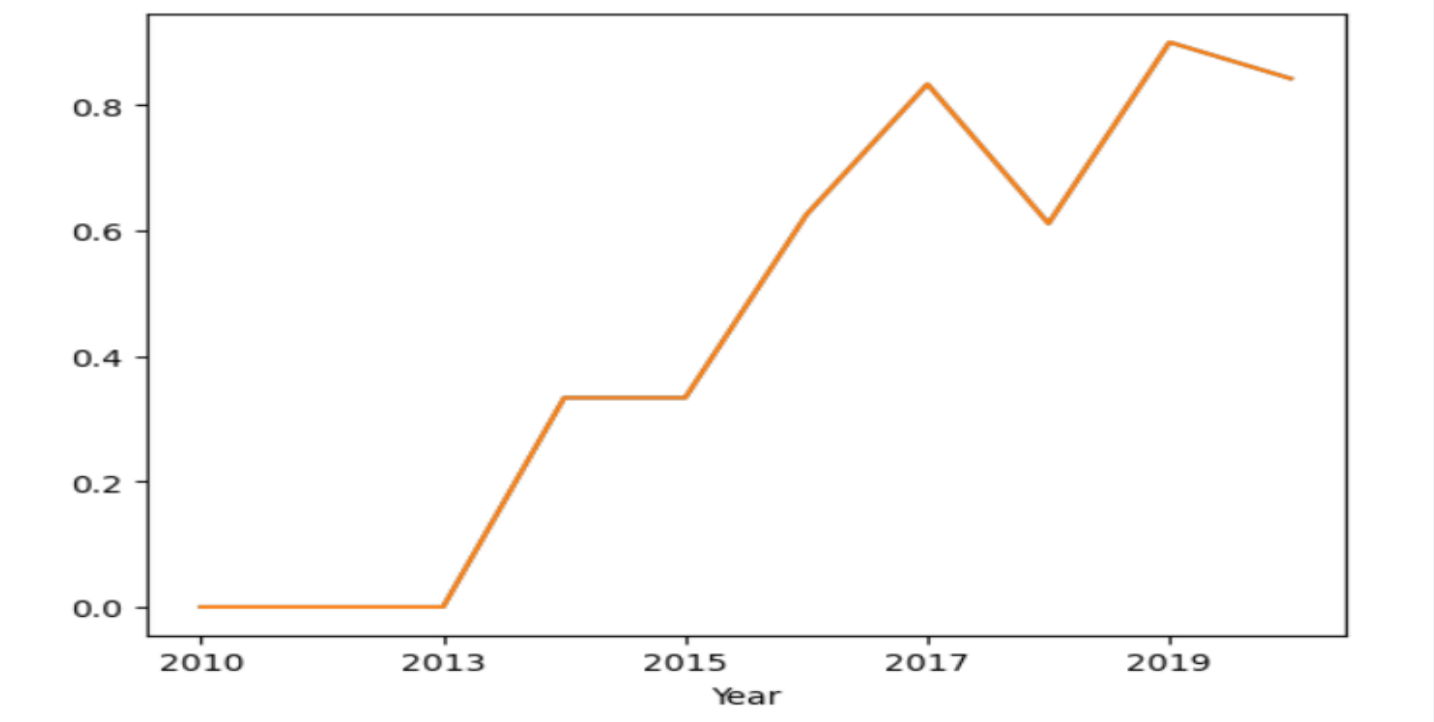
- Over the time all orbits are showing improvement in success rate.

Payload vs. Orbit Type



- Few launches to orbits SO and GEO.
- With descent success rate and wide range of payloads, rocket launch is towards International Space Station.

Launch Success Yearly Trend



- From year 2013 there is a rise in success rate. From around 0% to around 80%.

All Launch Site Names

- Names of the unique launch sites is given in the table.
- There are four launching sites in data, obtained writing a SQL query.

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

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`
- Achieved by using LIMIT.

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

- Total payload carried by boosters from NASA is 45596.
- Calculated by summing all payloads whose codes contain 'NASA (CRS)'.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2928.
- Calculated using avg function by filtering data with BOOSTER_VERSION = 'F9 v1.1'

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad is 2015-12-22.
- Query is `select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'`

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

- Query used is

```
select BOOSTER_VERSION

from SPACEXTBL

where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_
> 4000 and PAYLOAD_MASS__KG_ < 6000
```

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

Mission Outcome	Occurences
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

- Names of the booster which have carried the maximum payload mass are shown in beside table.
- Query used: `select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)`

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

- 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

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

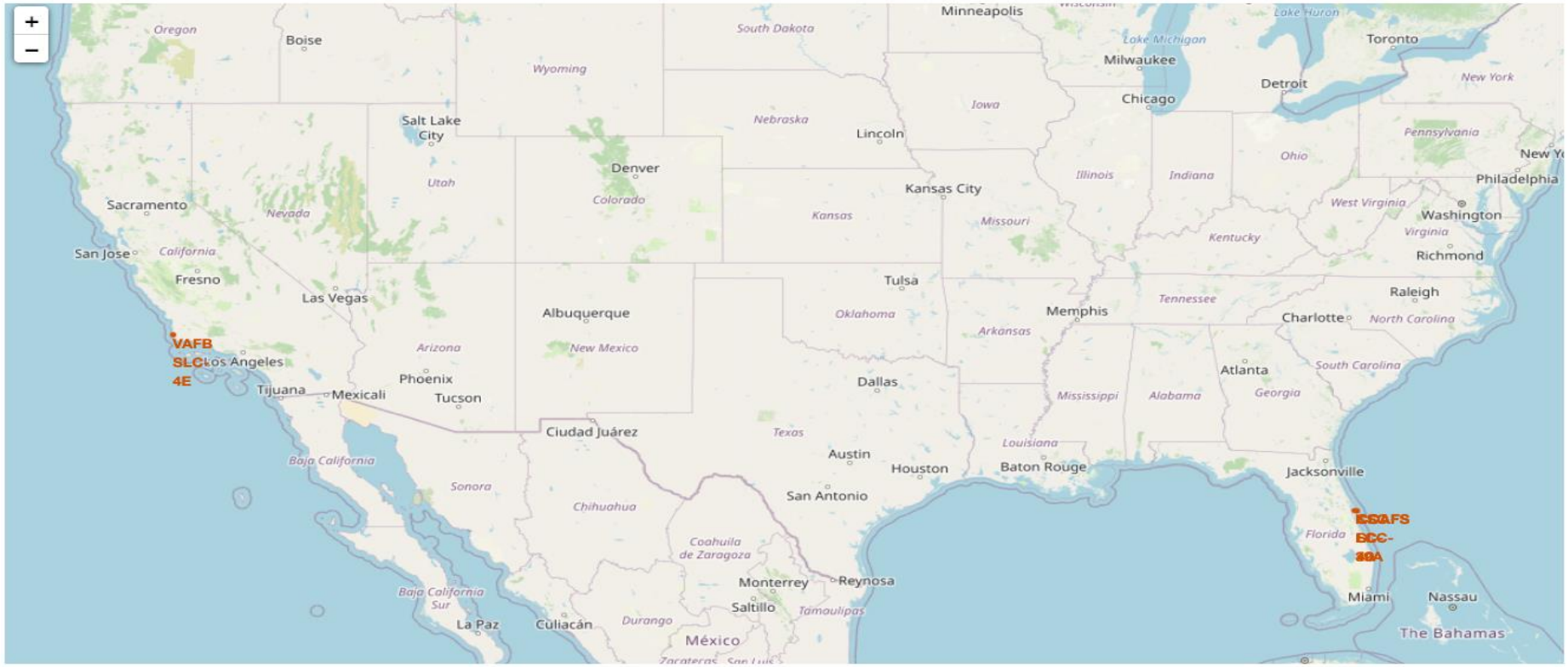
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing Outcome	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

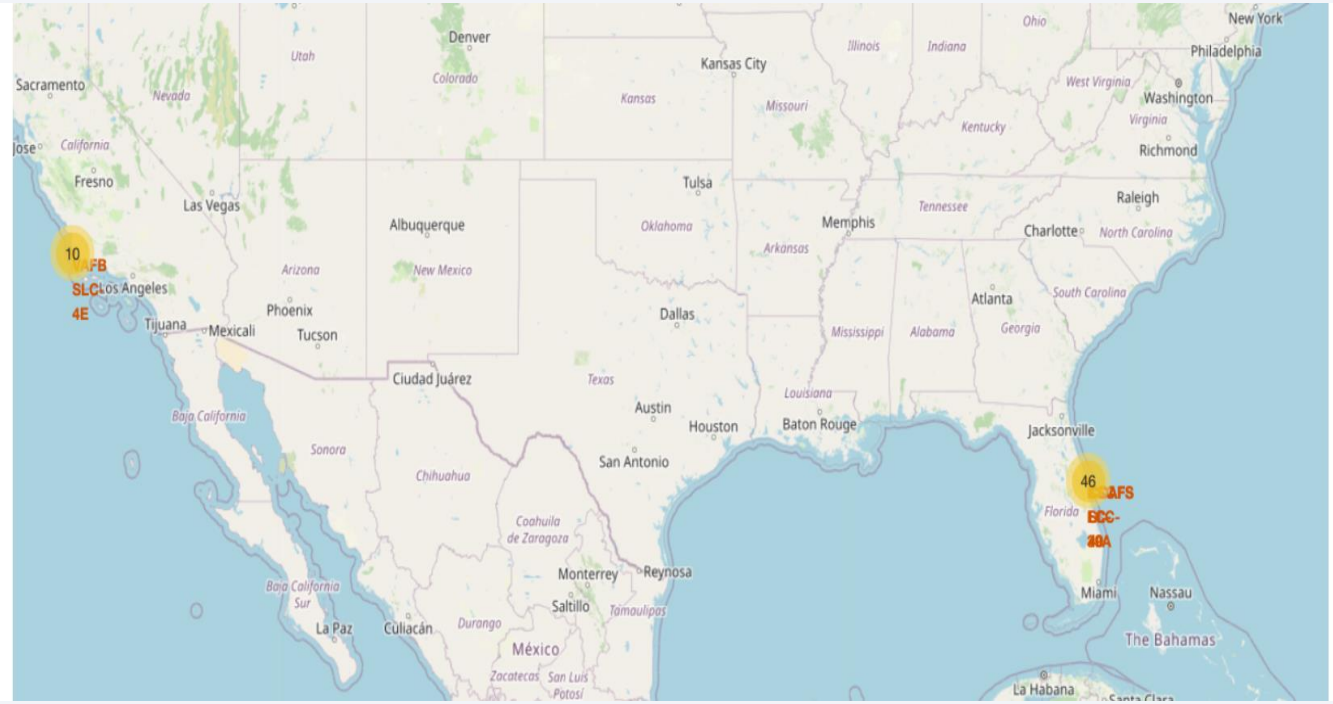

Section 3
Launch Sites
Proximities Analysis

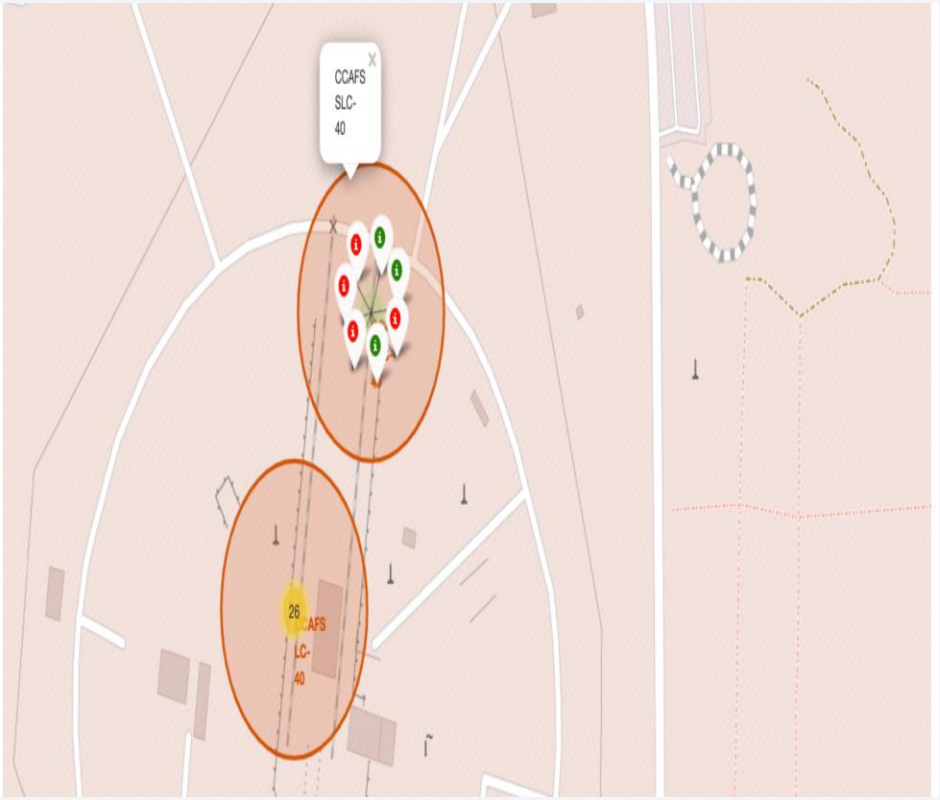
All Launch Sites

- All launch sites are near ocean, this might be for two reasons, safety and logistics.



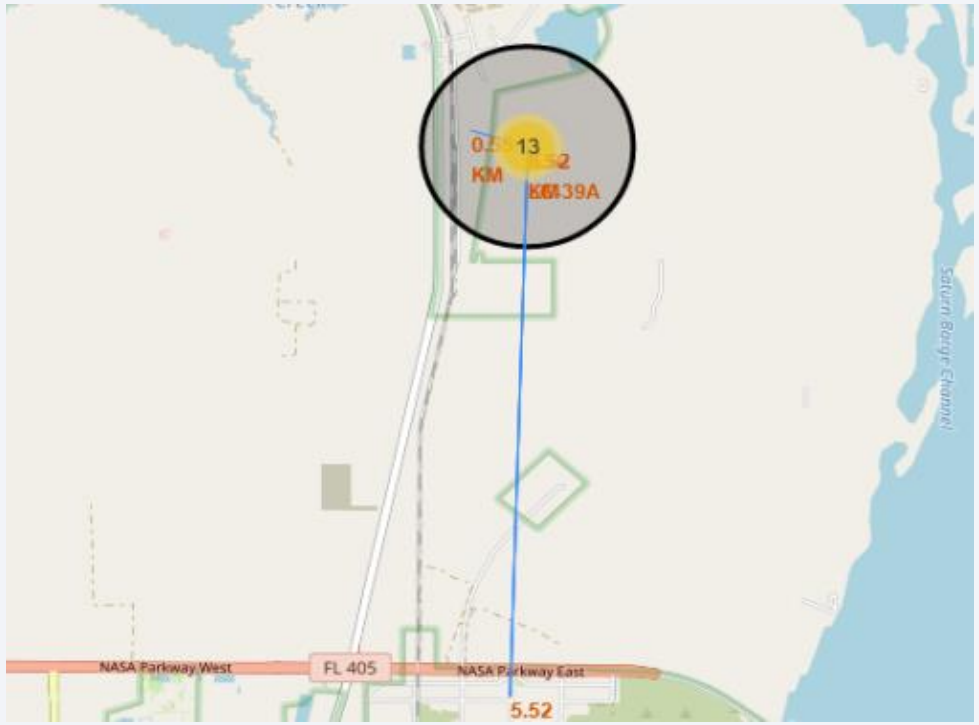
Launch outcomes by site



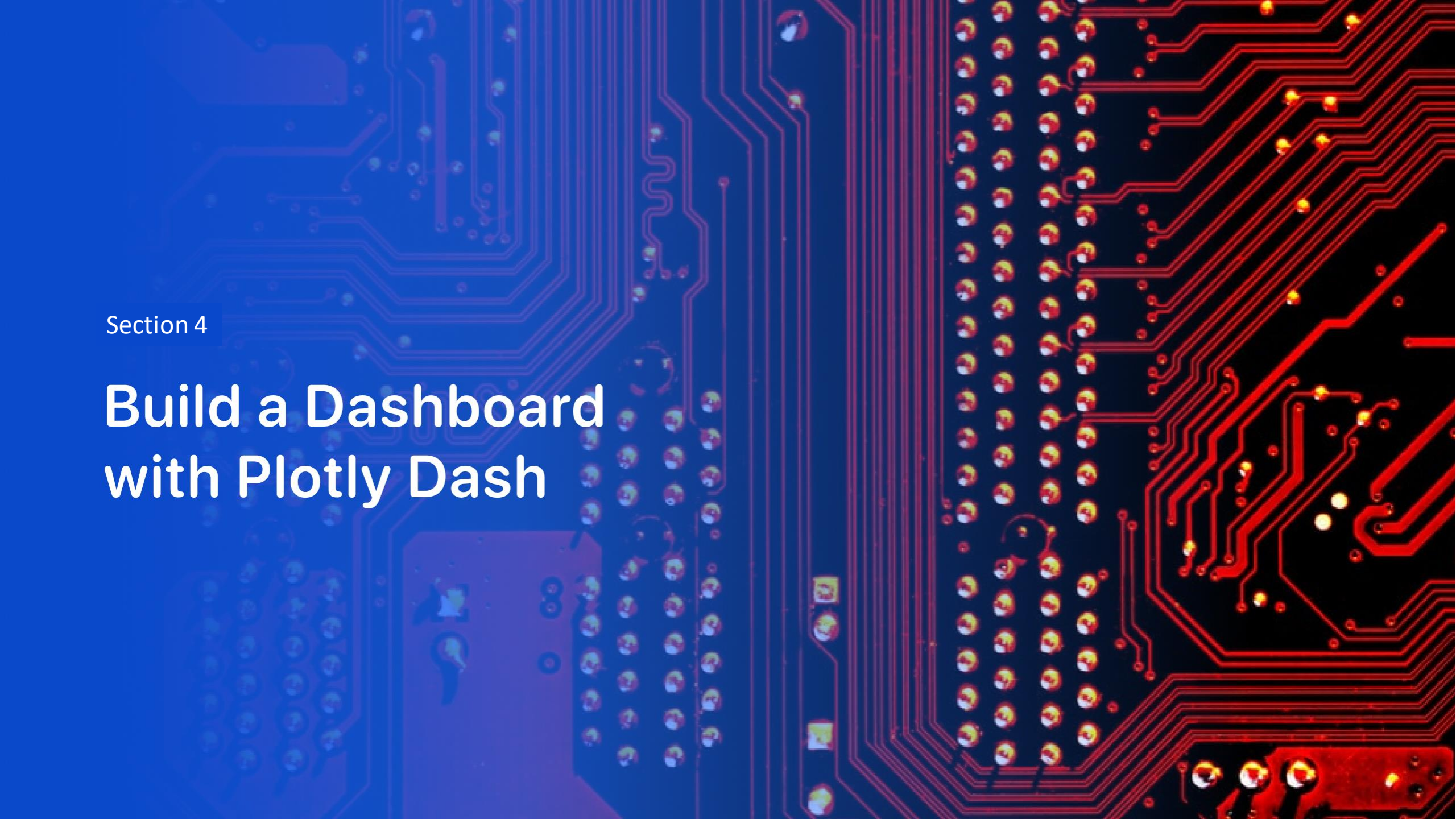


- Right side image is zoom in of right most marker in first map.

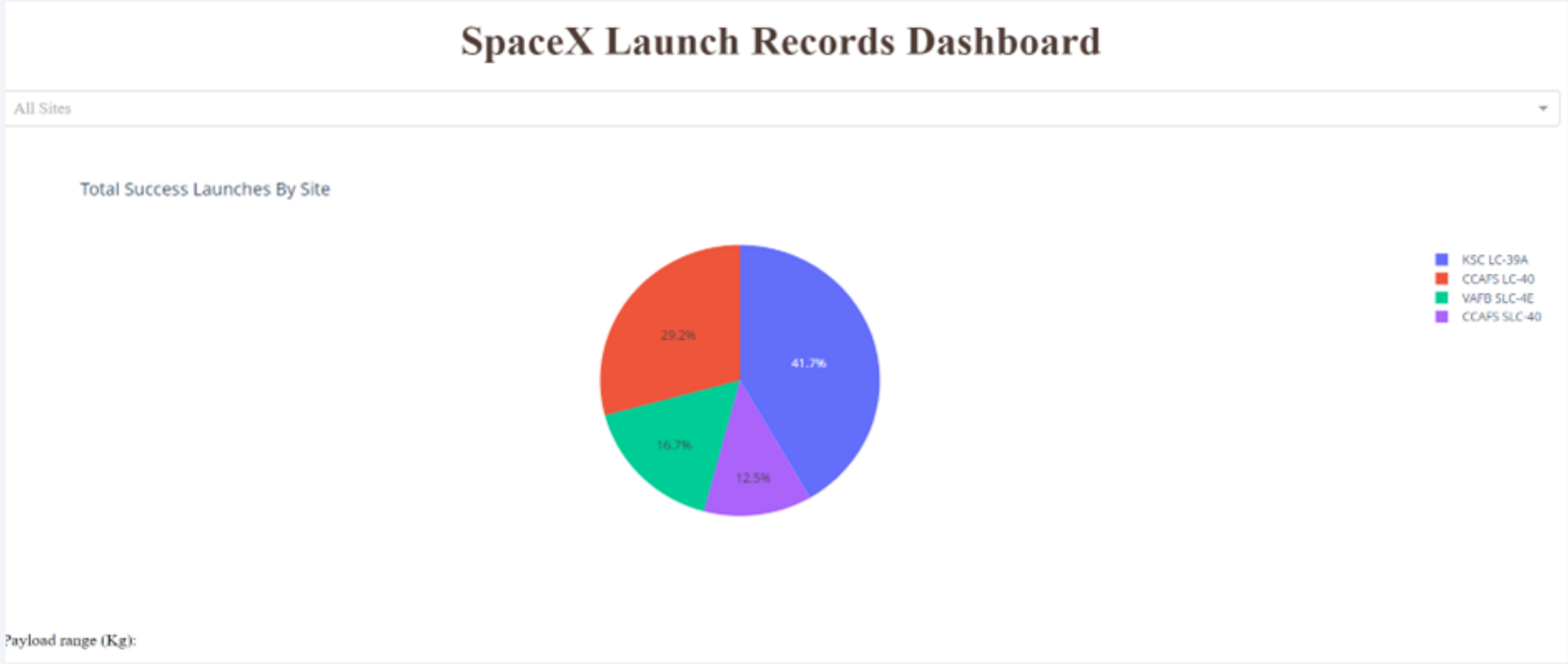
Logistics and Safety



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

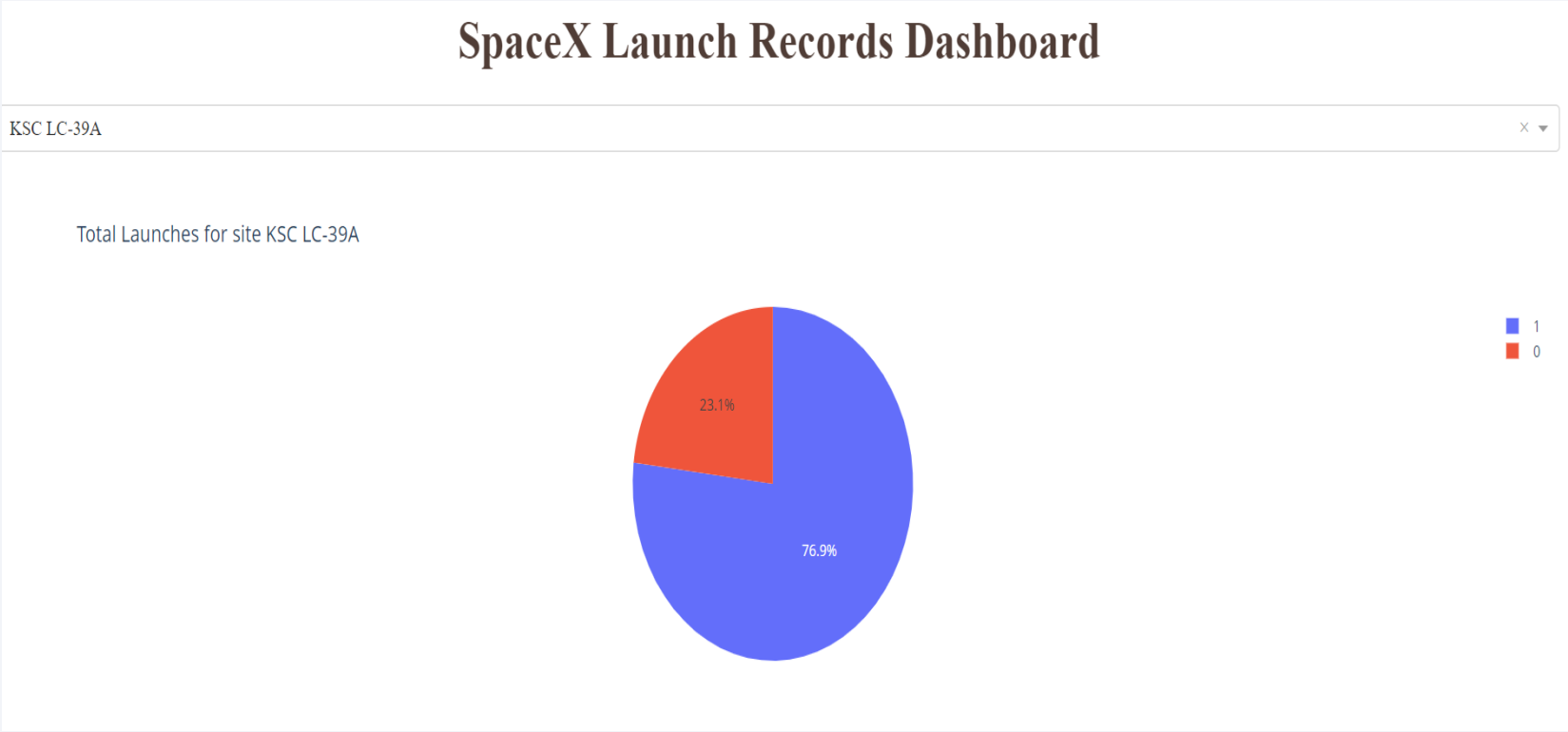

Section 4
Build a Dashboard with Plotly Dash

Successful Launches All Sites



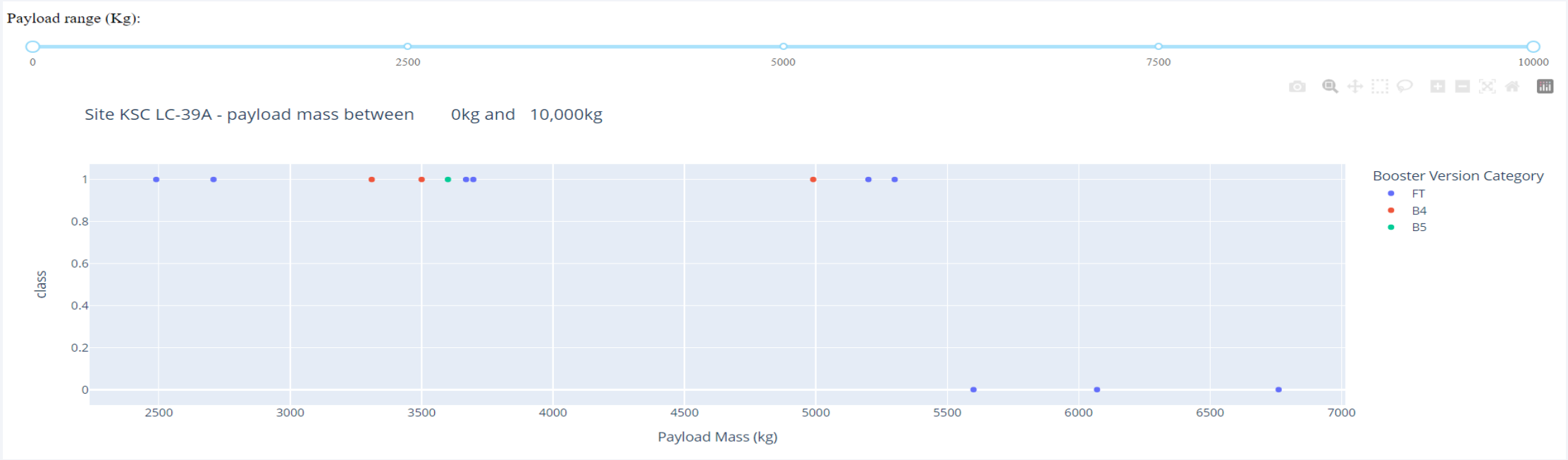
- The place from where launches are done seems to be a very important factor of success of missions.

Highest Launch Success Ratio



- KSC LC-39A has 76.9% success rate.

Payload vs. Launch Outcome



- Payloads under 6,000kg and FT boosters are the most successful combination.
- There's not enough data to estimate risk of launches over 7,000kg.

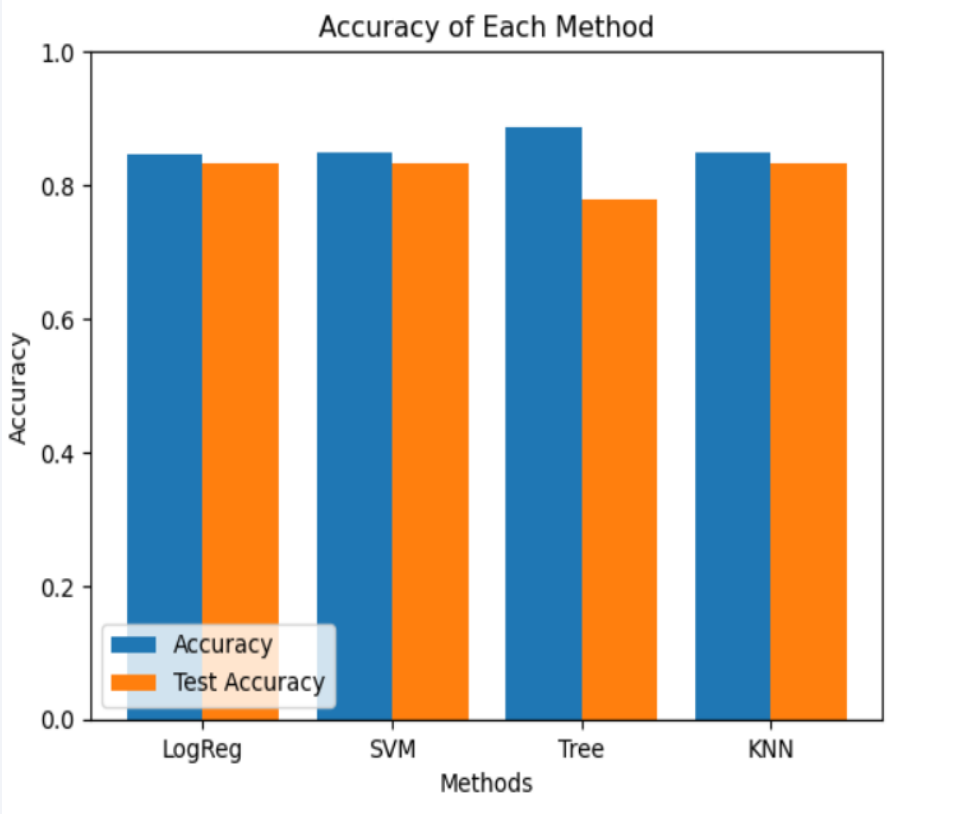
Section 5

Predictive Analysis (Classification)

Classification Accuracy

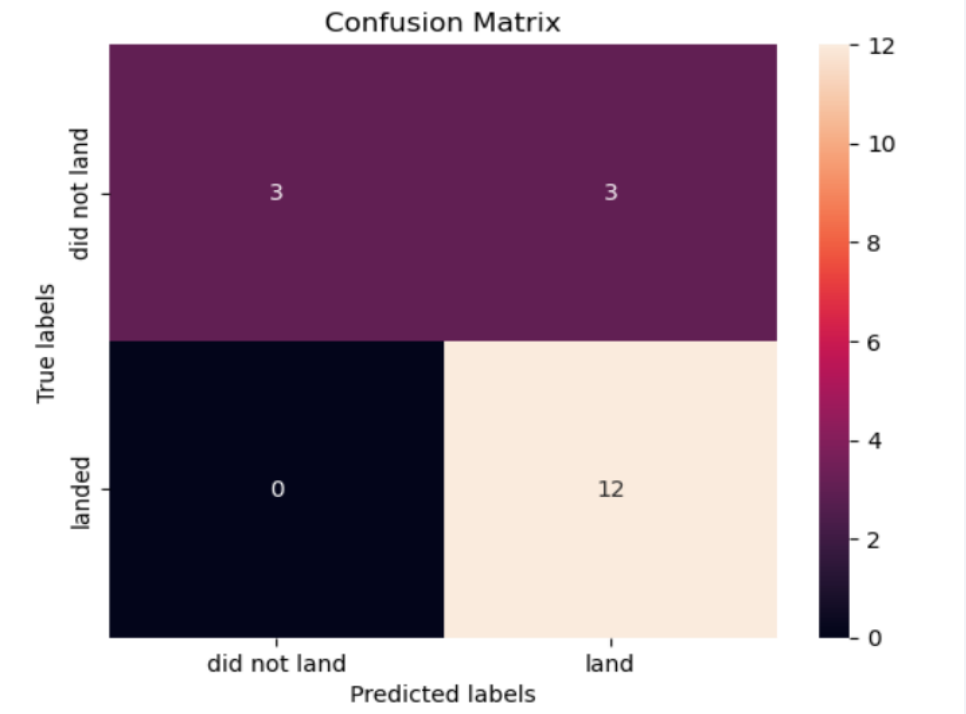
- Four classification models were tested, and their accuracies are plotted beside.

- Logistic Regression and SVM have test accuracy around 80%. Decision Tree has around 85% accuracy with training data.



Confusion Matrix

- Confusion matrix of Logistic regression proves its accuracy by showing big true positive and true negative compared to 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 to the evolution of processes and rockets.
- Logistic regression or SVM Classifier can be used to predict successful landings and increase profits.

Appendix

- Link to all Jupyter notebooks and python code: https://github.com/AdithyaSwarna/IBM-Applied-Data-Science-Capstone

Thank you!