IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Nguyễn Anh Đức>
<04/08/2022>

Outline

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

Executive Summary

- Summary of methodologies:
  - ➢ Data collection using web scraping and SpaceX APIs
  - ➢ Exploratory Data Analysis (EDA), including data wrangling, data visualization, and interactive visual analytics
  - ➢ Machine learning
- Summary of all results:
  - ➢ Collect valuable data from public resource
  - ➢ EDA allowed to identify which features are the best to predict success of launchings;
  - ➢ • Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

Project background and context:

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

Problems you want to find answers:

• The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;

• Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX APIs (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Creating an outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Try different models to find most suitable model

Data Collection

- Describe how data sets were collected.

➢From SpaceX APIs (https://api.spacexdata.com/v4/rockets/)

➢From Wikimedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- You need to present your data collection process use key phrases and flowcharts

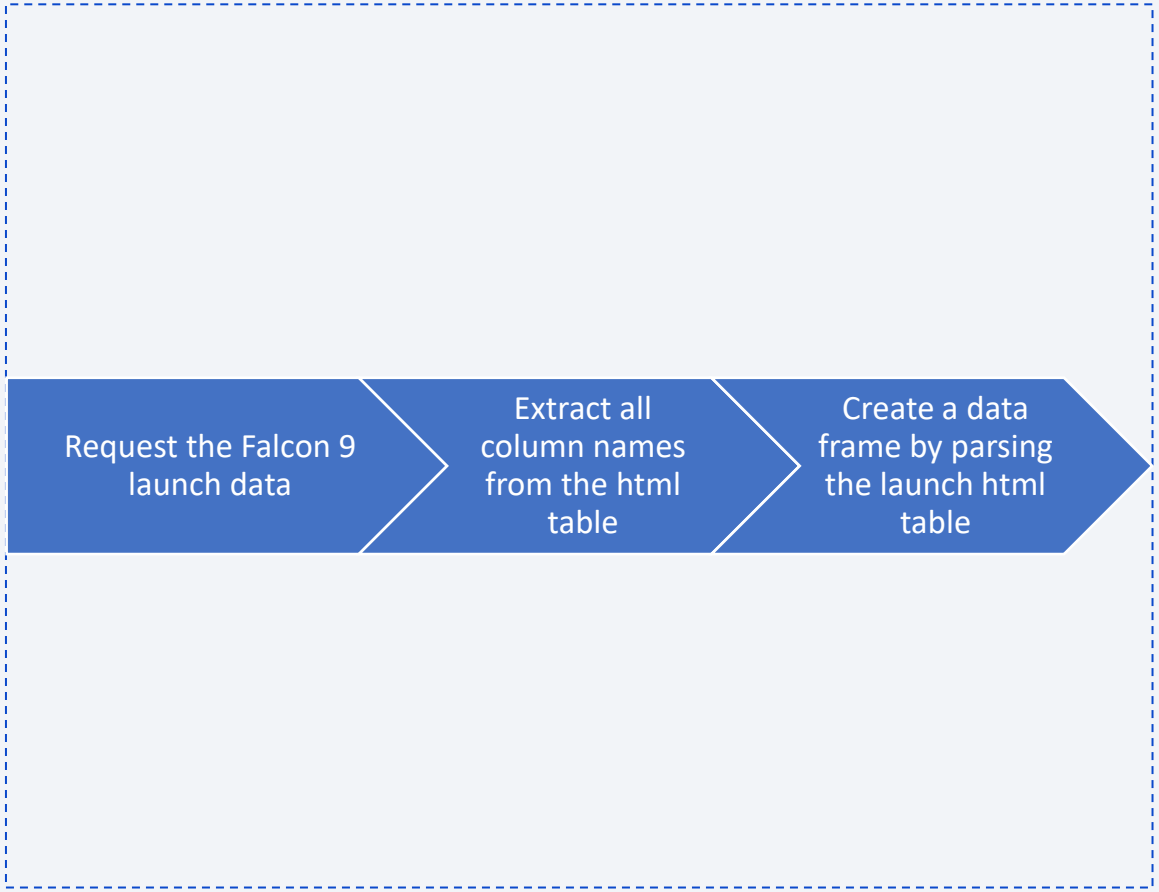
Data Collection – SpaceX API

- SpaceX APIs is a public APIs which user can obtain and use

- Github: https://github.com/nguyenanhduc6479/data-science/blob/2a2550a60ca5ff14ee9f01bd7bf849f704b63f32/15.jupyter-labs-spacex-data-collection-api.ipynb

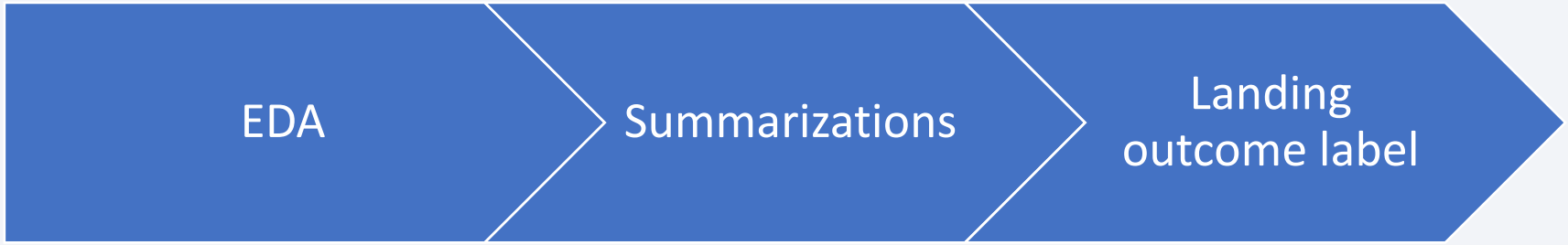
Data Collection - Scraping

- GitHub: https://github.com/nguyenanhduc6479/data-science/blob/2a2550a60ca5ff14ee9f01bd7bf849f704b63f32/16.jupyter-labs-webscraping.ipynb



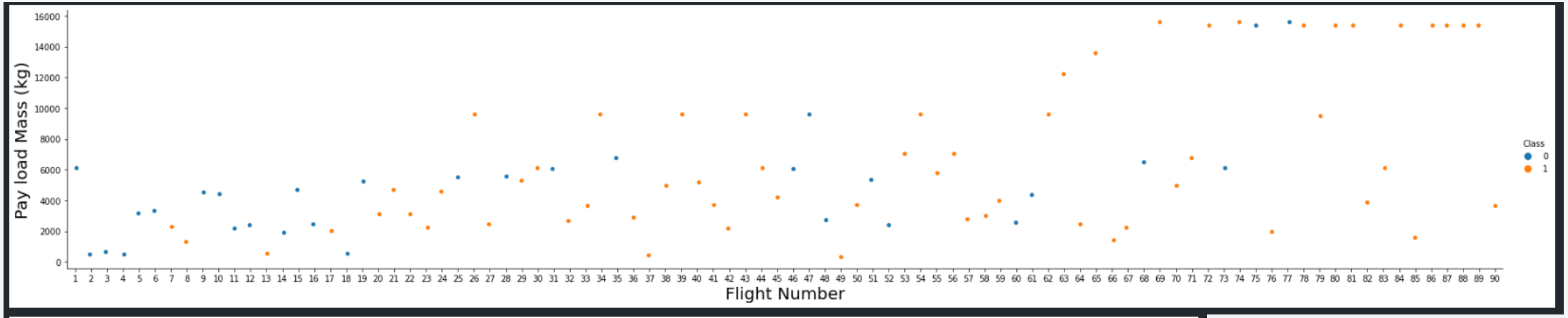
Data Wrangling

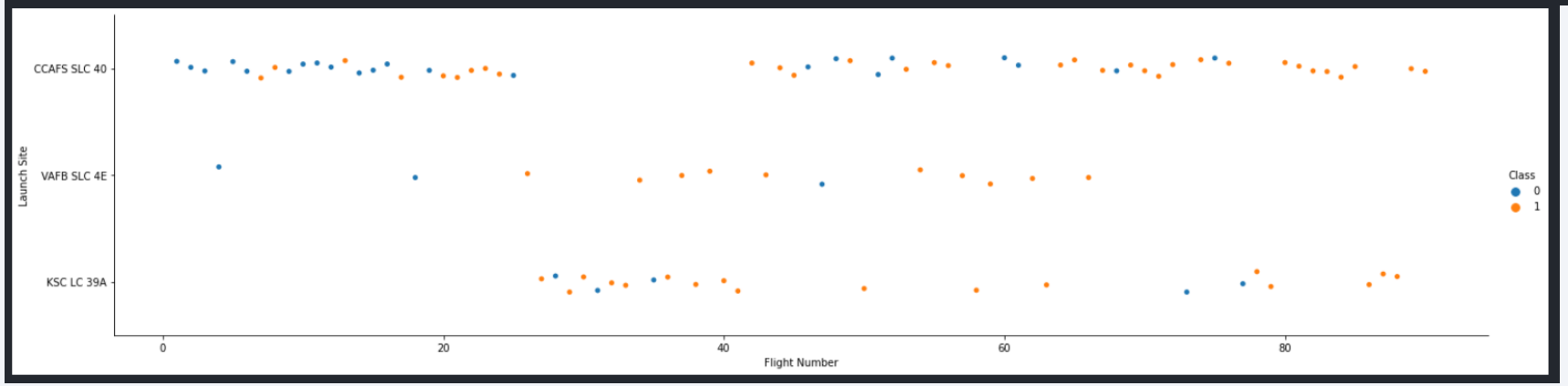
- Firstly, some EDA was performed on the dataset
- Secondly, launch per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated
- Finally, we created the landing outcome label based on outcome column
- https://github.com/nguyenanhduc6479/data-science/blob/2fb96a22a7b006d66a6268e59c1c44a1e8840b9e/17.labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- To see relationship between Payload Mass and Flight Number, Launch Site and Flight Number, Launch Site and Payload Mass, Orbit and Flight Number, Payload and Orbit, we use scatter plot and bar plot





https://github.com/nguyenanhduc6479/data-science/blob/2fb96a22a7b006d66a6268e59c1c44a1e8840b9e/19.jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Some SQL queries we 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, and launch site names for in 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.

- https://github.com/nguyenanhduc6479/data-science/blob/97526b9c0d63e861dd0650023f27ec37e1e533a1/18.jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Markers indicate points like launch site
- Circles indicate areas like NASA Johnson Space Center
- Marker cluster indicate groups of events in each coordinate
- Lines indicate distances between two coordinates
- https://github.com/nguyenanhduc6479/data-science/blob/97526b9c0d63e861dd0650023f27ec37e1e533a1/20.lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data

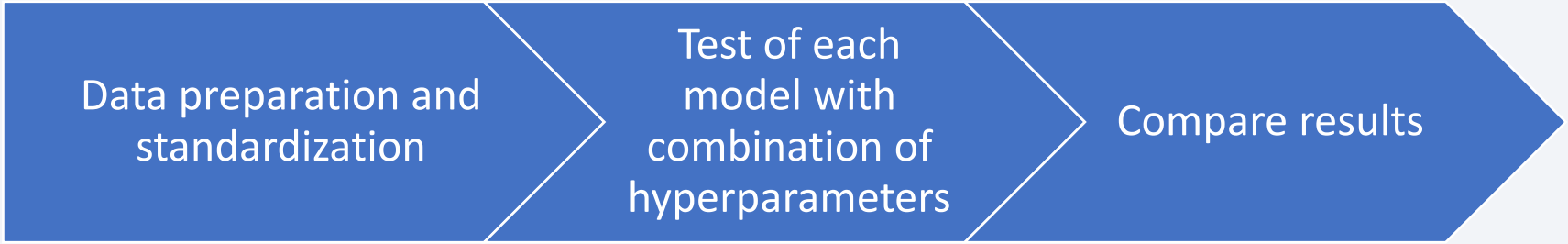
- Percentage of launches by site
- Payload range

These two plots help us analyze the relation between payload and launch site. These insights allow you to choose place to launch based on payload

- https://github.com/nguyenanhduc6479/data-science/blob/852d2de9750d4e3f976d9b88a86a2fbaedb26c68/spacex_dash.py

Predictive Analysis (Classification)

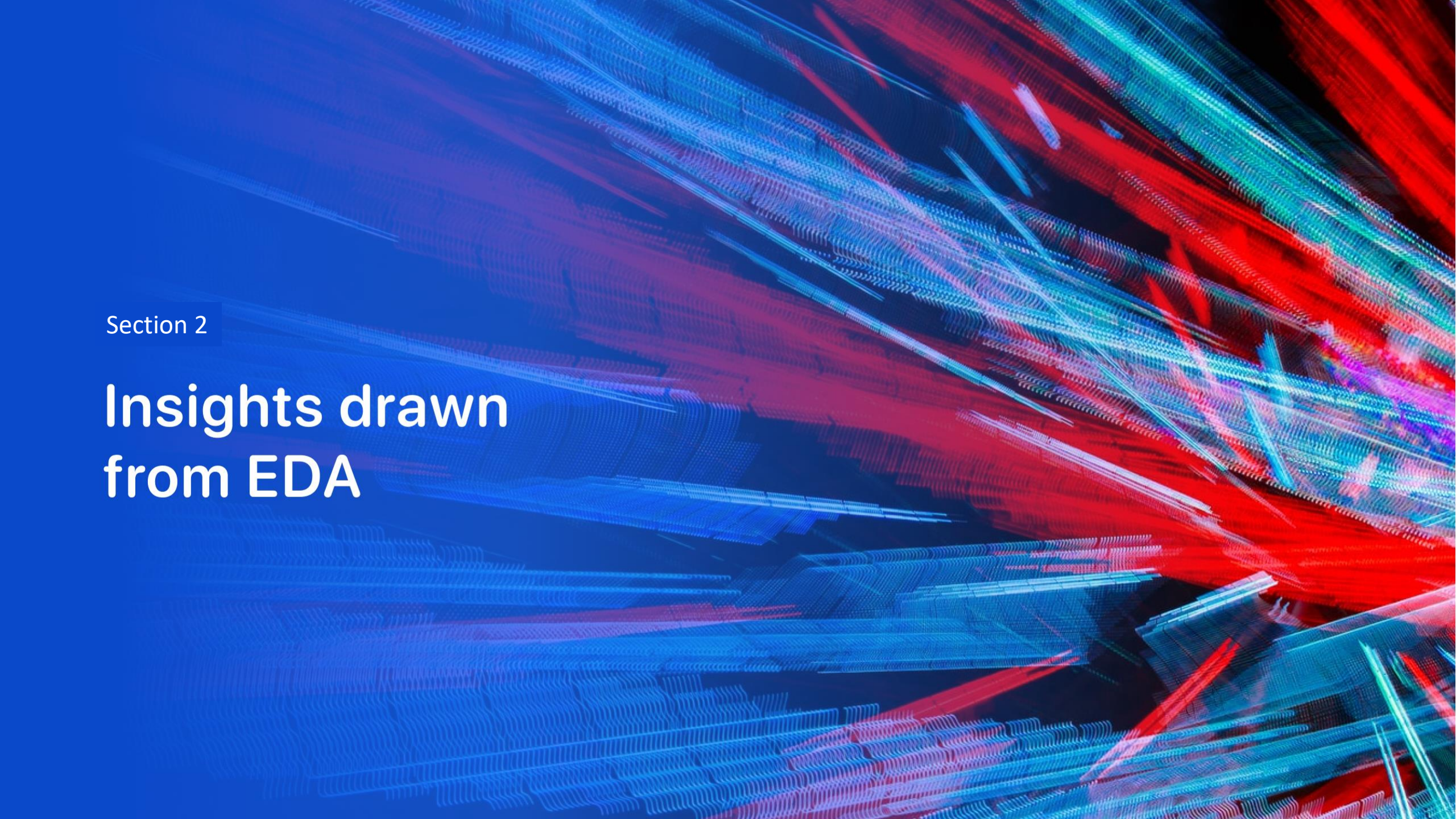
- We build 4 classification models: logistic regression, support vector machine, decision tree, k nearest neighbors



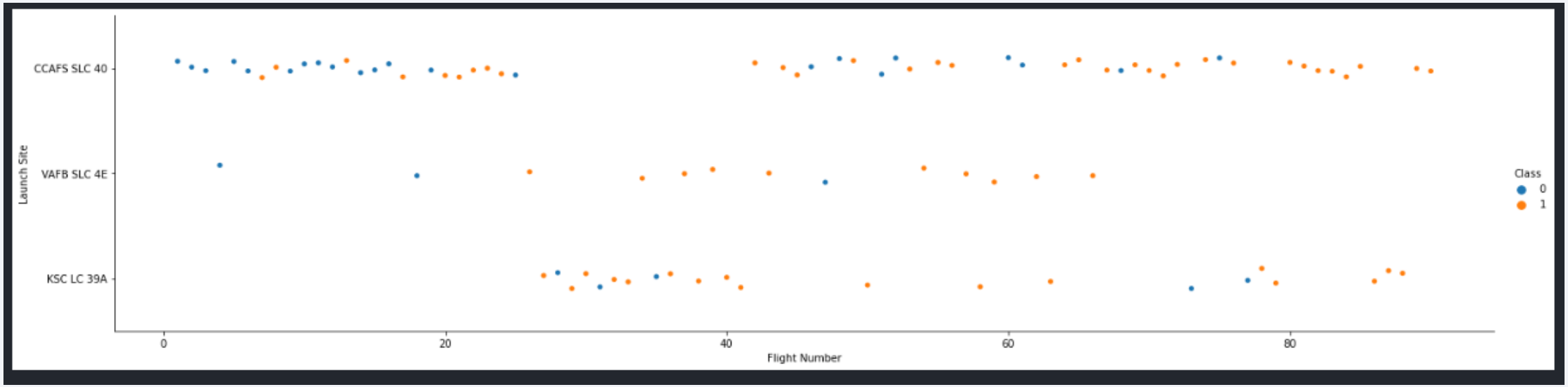
- https://github.com/nguyenanhduc6479/data-science/blob/92221f8ee82bbf6974882edfdfd8e53382830d0e/21.SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- 4 different launch sites
- The average payload of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2015
- Many Falcon 9 booster versions were successful at landing in drone ships
- Almost 100% of mission outcomes were successful
- The number of landing outcomes became as better as years passed
- Predictive Analysis showed that Logistic Regression is the best model to predict successful landings, having accuracy over 84% and accuracy for test data over 83%.

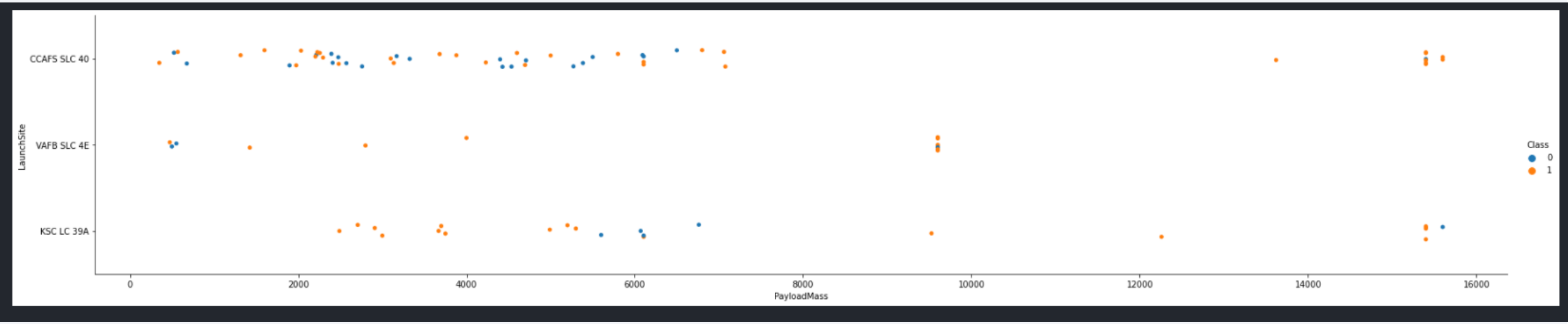

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Launch Site CCAFS SLC 40 recently gives the best performance when most of the launches were successful
- General success rate improve over time

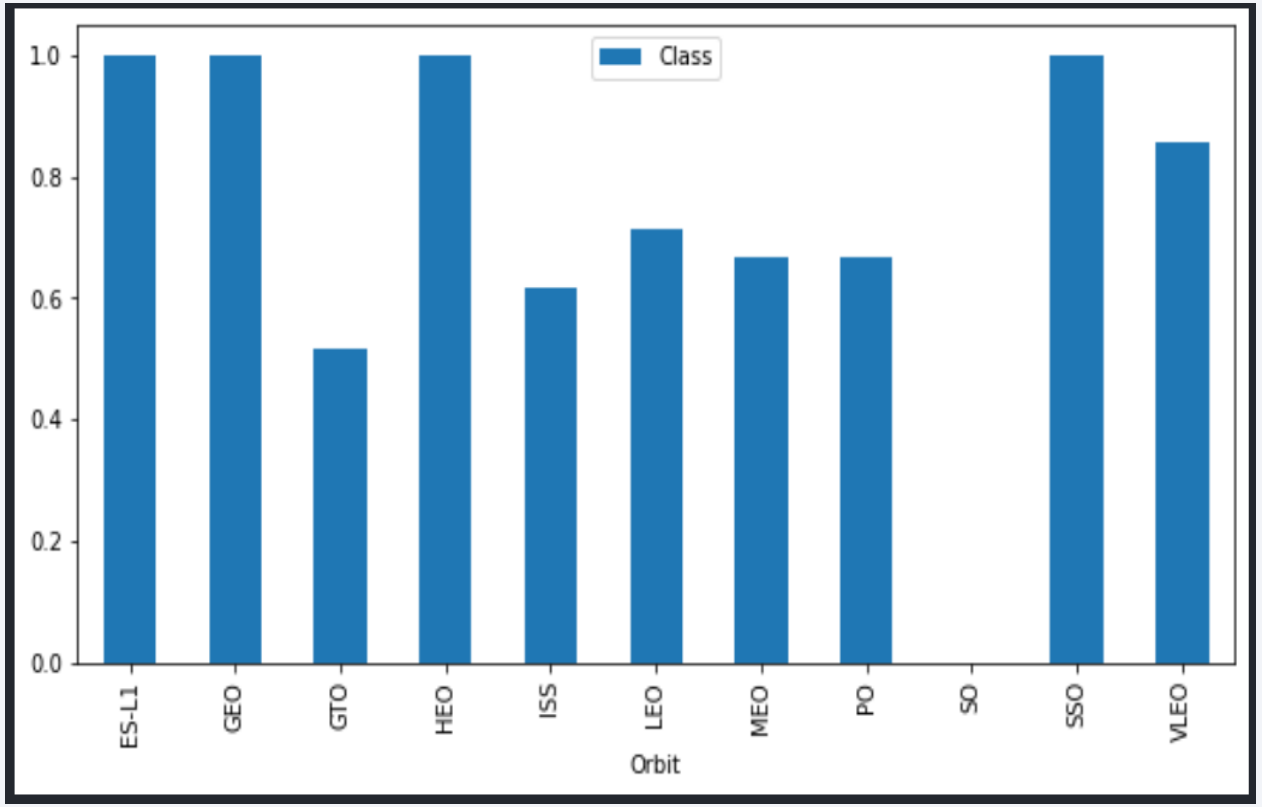
Payload vs. Launch Site



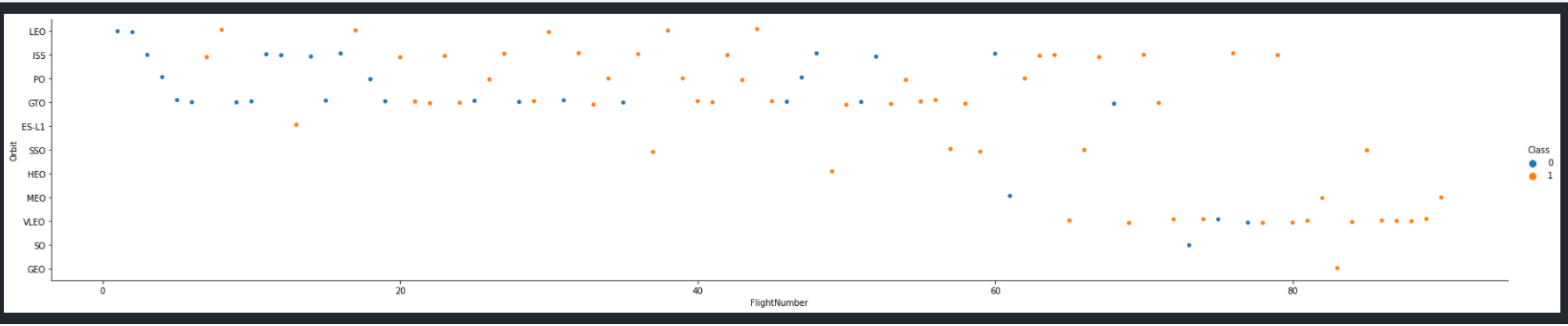
- KSC LS 39A has very good success rate with every payload mass
- Over 9000kg has good success rate
- With payload over 12000kg we should choose CCAFS SLC 40 or KSC LC 39A because they had succeed with this mass several times

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO have the highest success rate

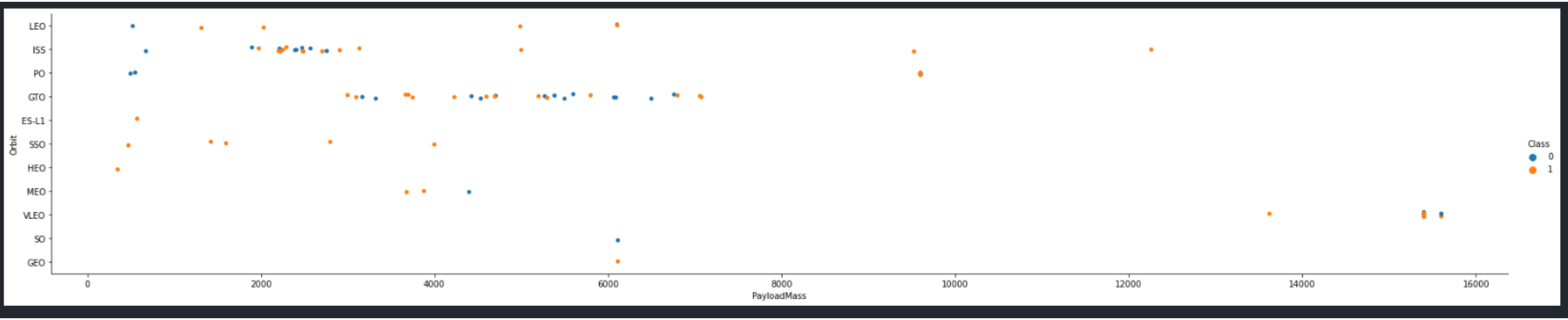


Flight Number vs. Orbit Type



- Success rate improved over time
- VLEO has been used recently and its performance was great

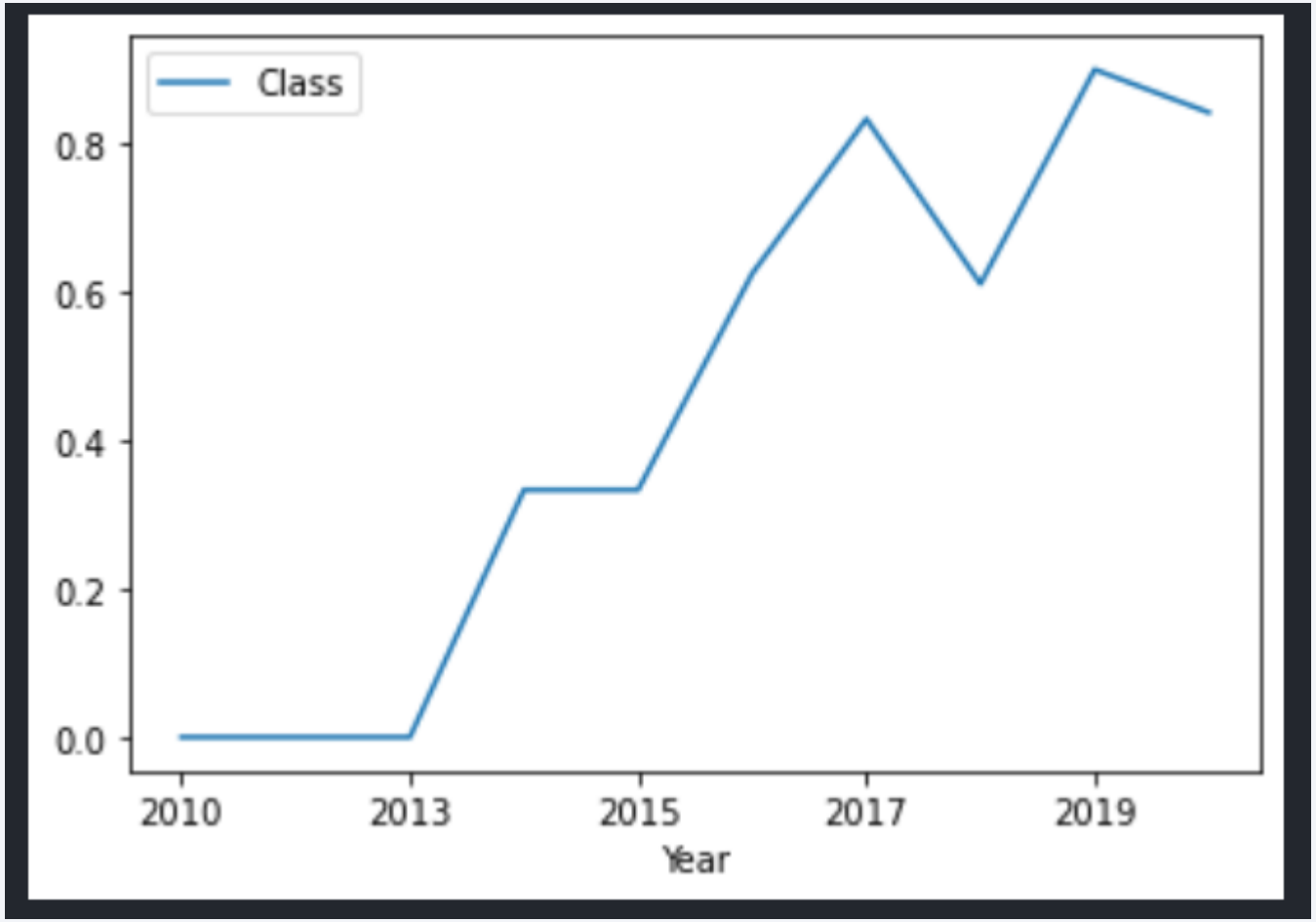
Payload vs. Orbit Type



- ISS has the widest range of payload
- With massive payload like over 14000kg they seemed to like using VLEO

Launch Success Yearly Trend

- Success rate keep improving since 2013
- Highest success rate is in 2019



All Launch Site Names

- Find unique launch site name in data frame

Launch Site Names Begin with 'CCA'

- Show first 5 result which launch site 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

Total Payload Mass

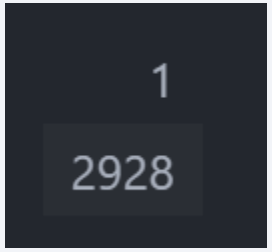
- Using SUM function with customers have ‘CRS’ in it

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER LIKE '%CRS%'
```

1
48213

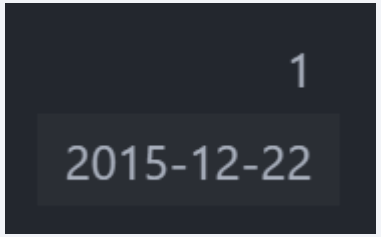
Average Payload Mass by F9 v1.1

- Choose all F9 v1.1 booster version and then add all the payload mass together



First Successful Ground Landing Date

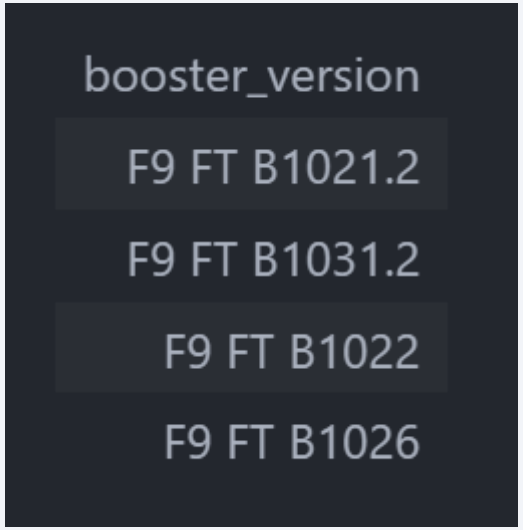
- Filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on 12/22/2015.



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

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



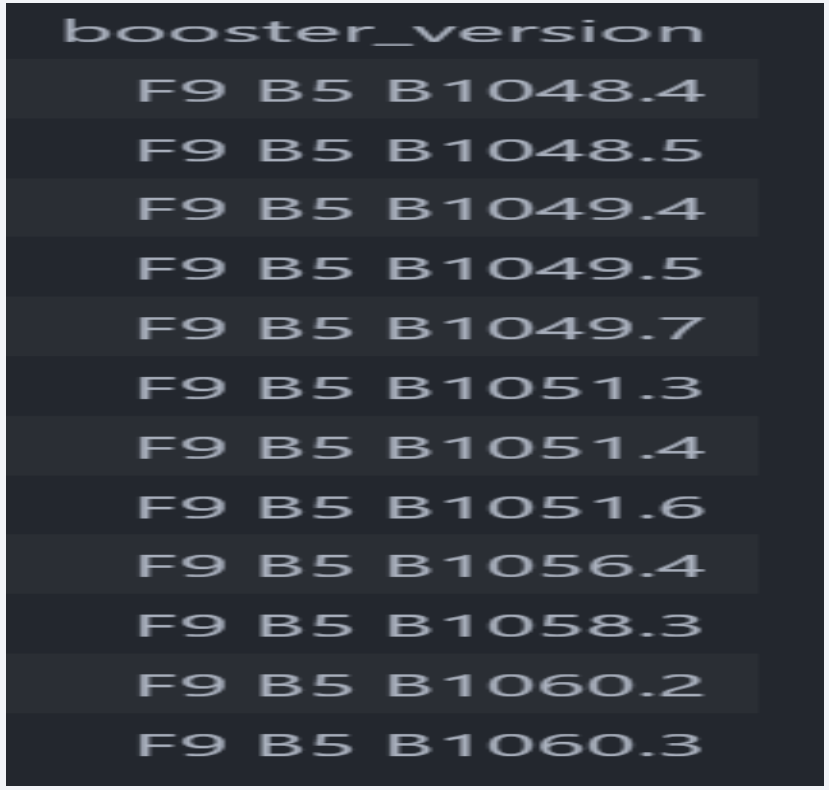
Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Grouping mission outcome and count records help us to get quantities of each outcome

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Get the maximum payload and find the booster version with that payload



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

2015 Launch Records

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

booster_version	landing_outcome
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

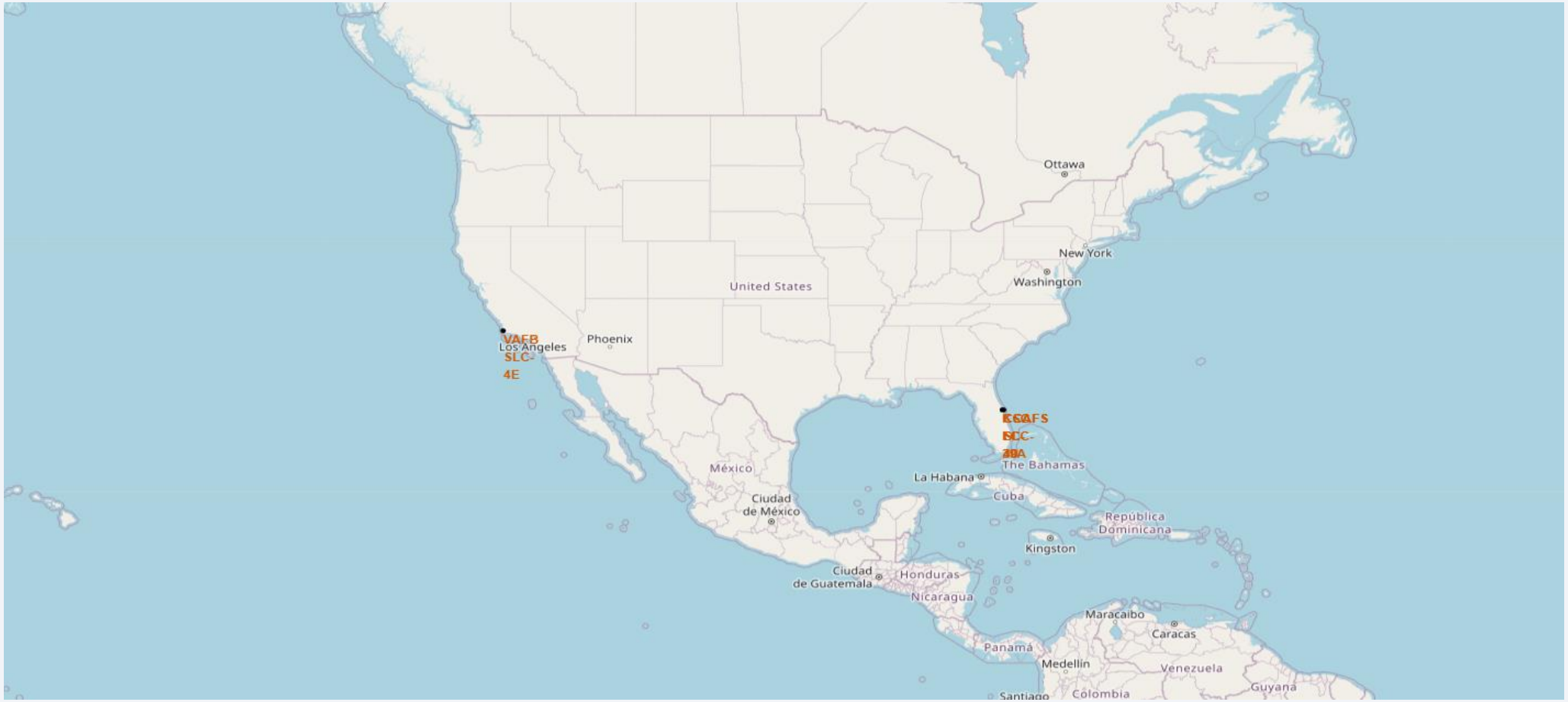
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

landing_outcome	qty
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

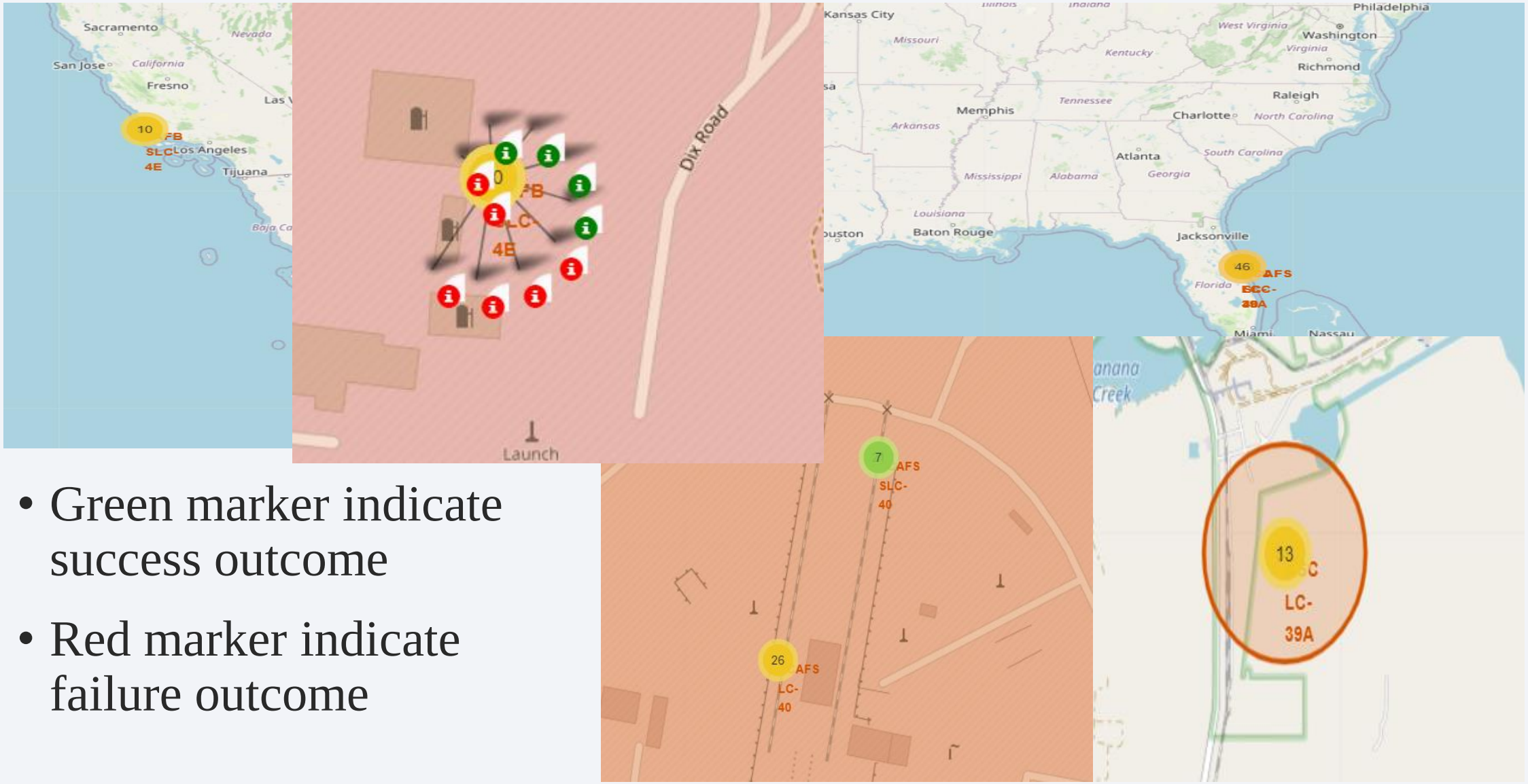
Launch Sites Proximities Analysis

ALL LAUNCH SITES



Launch site located near sea, but not too far from roads and railroads

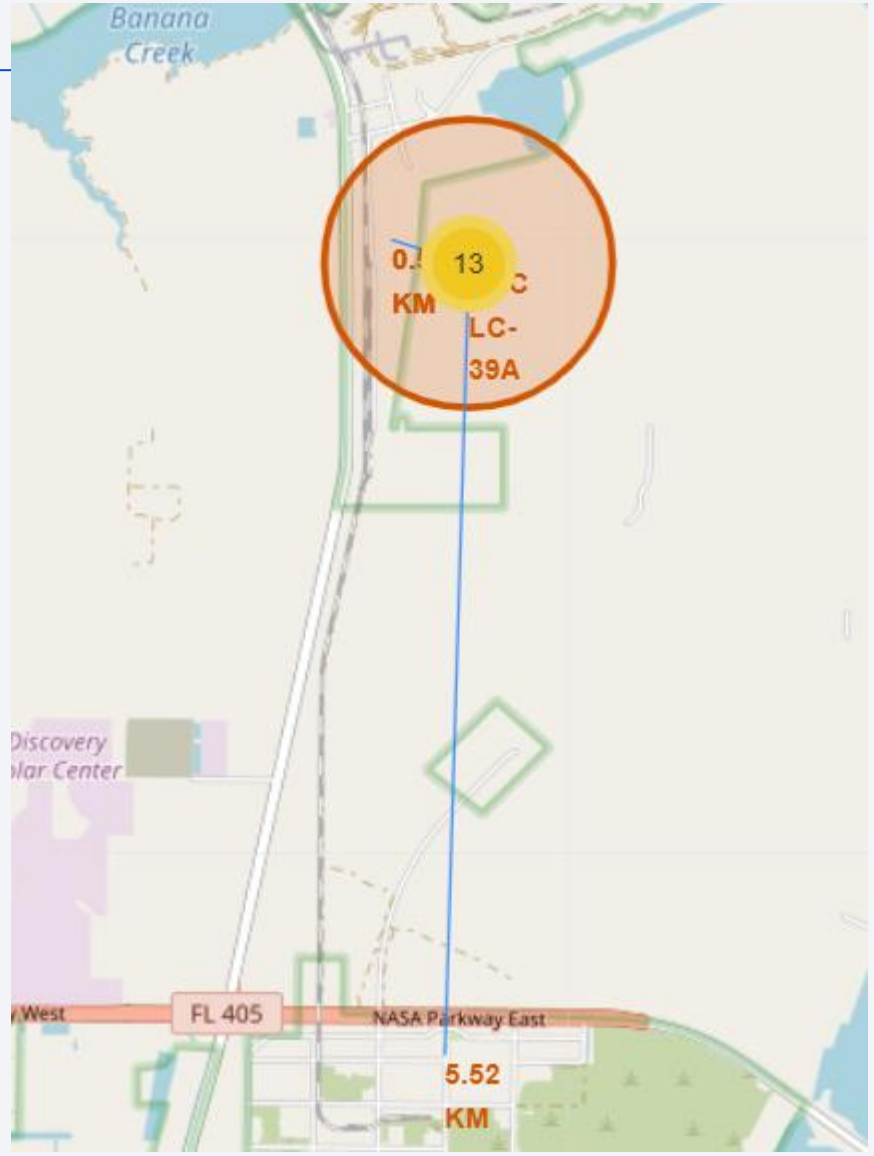
LAUNCH OUTCOME BY SITE

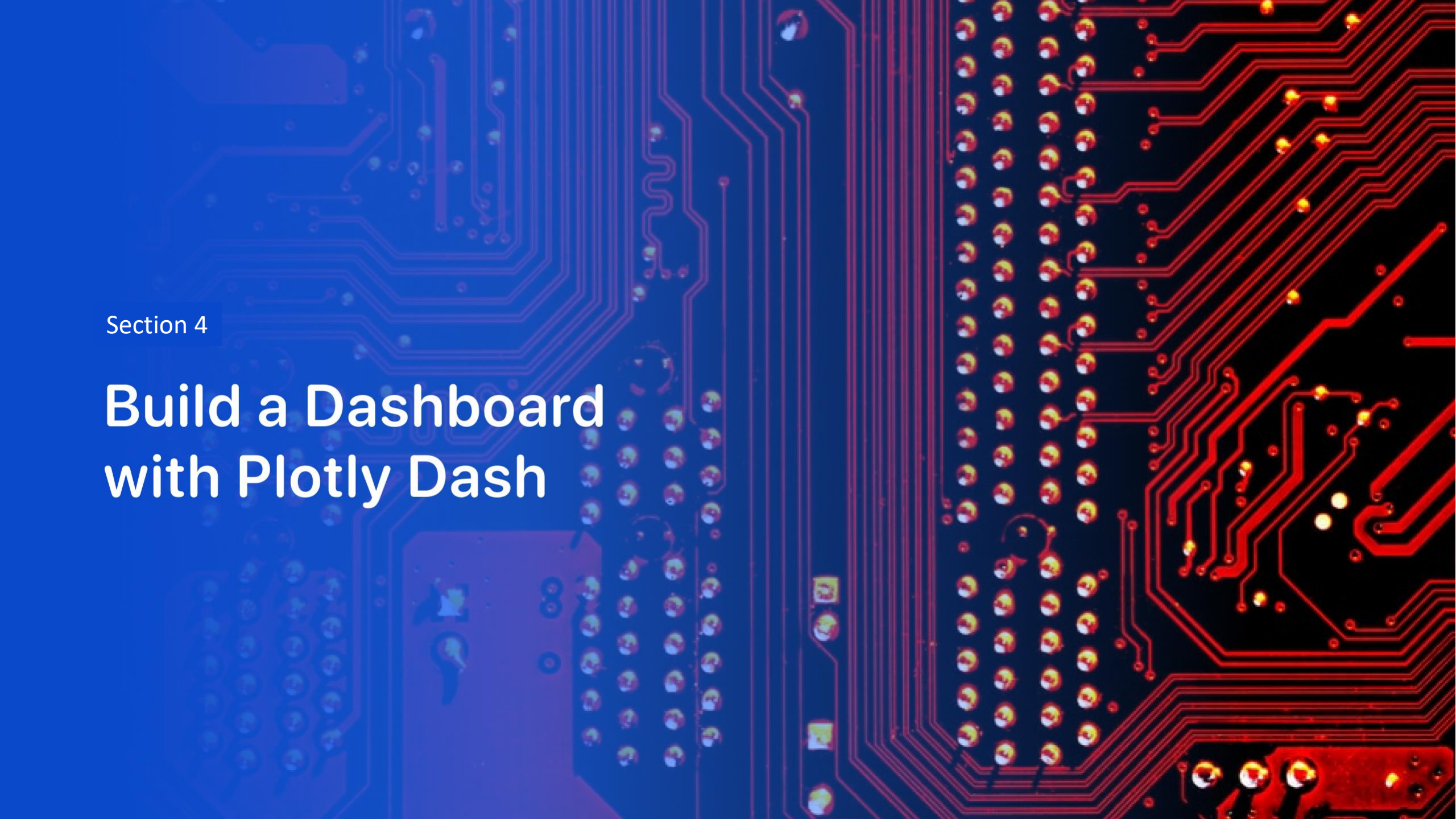


- Green marker indicate success outcome
- Red marker indicate failure outcome

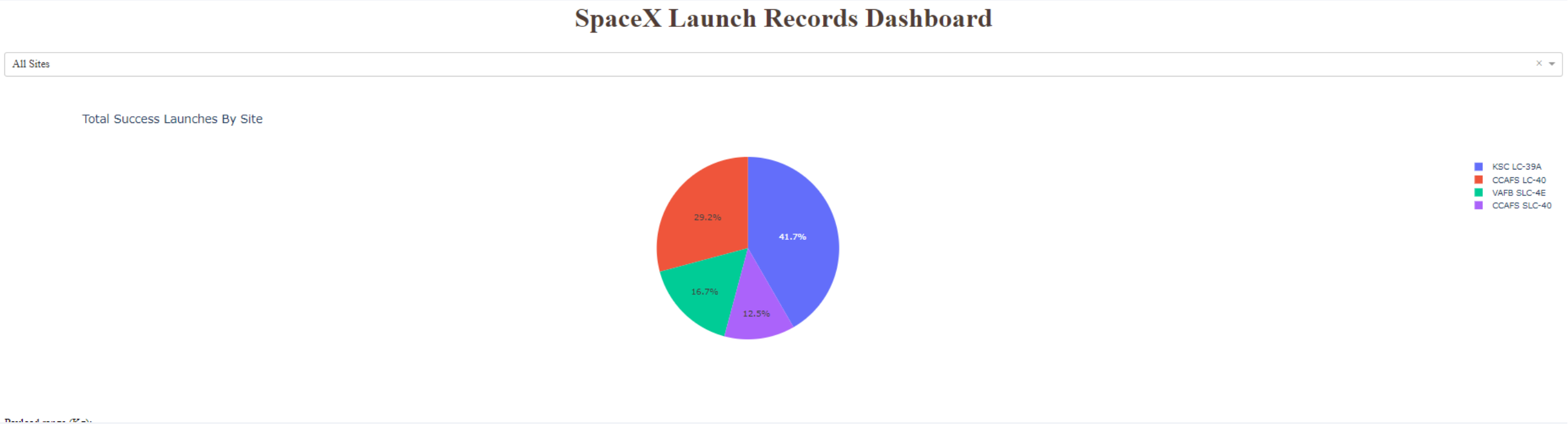
NEAREST LOCATION

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



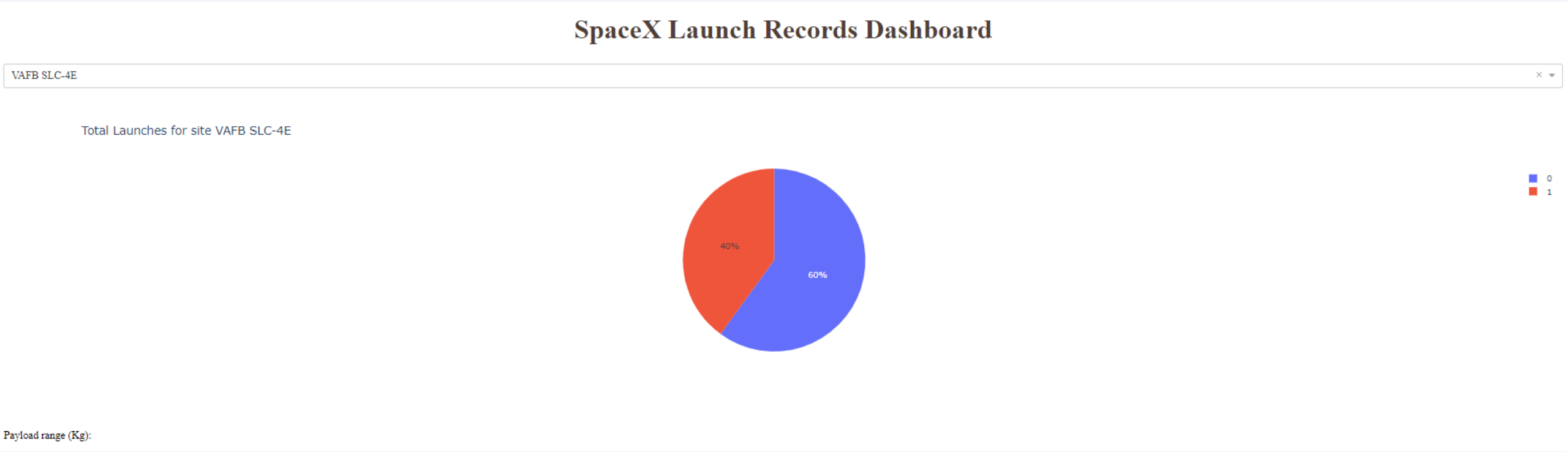

Section 4
Build a Dashboard with Plotly Dash

SUCCESSFUL LAUNCH BY SITE



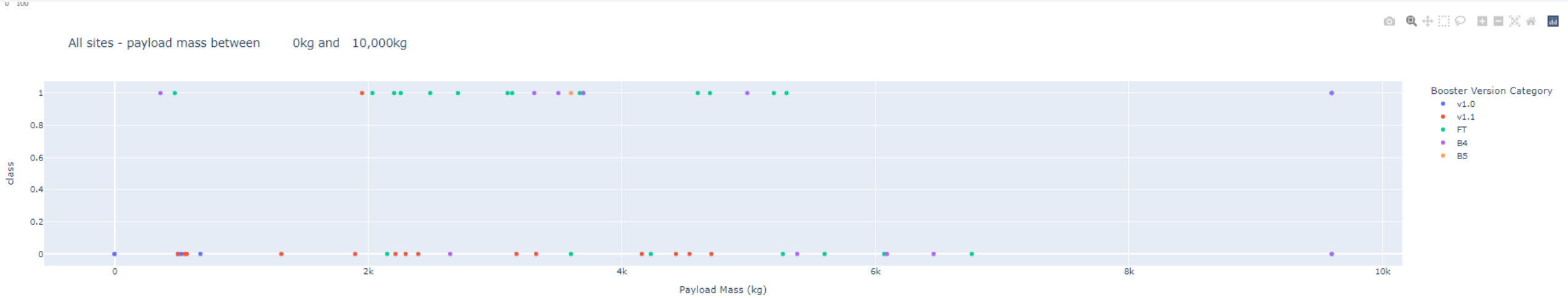
- KSC LC-39A has the most successful launch
- Next to that is CCAFS LC-40
- The other two launch site has similar number of successful launch

SUCCESS RATE FOR VAFB SLC-4E



- This site has only 40% success rate

PAYLOAD VS OUTCOME



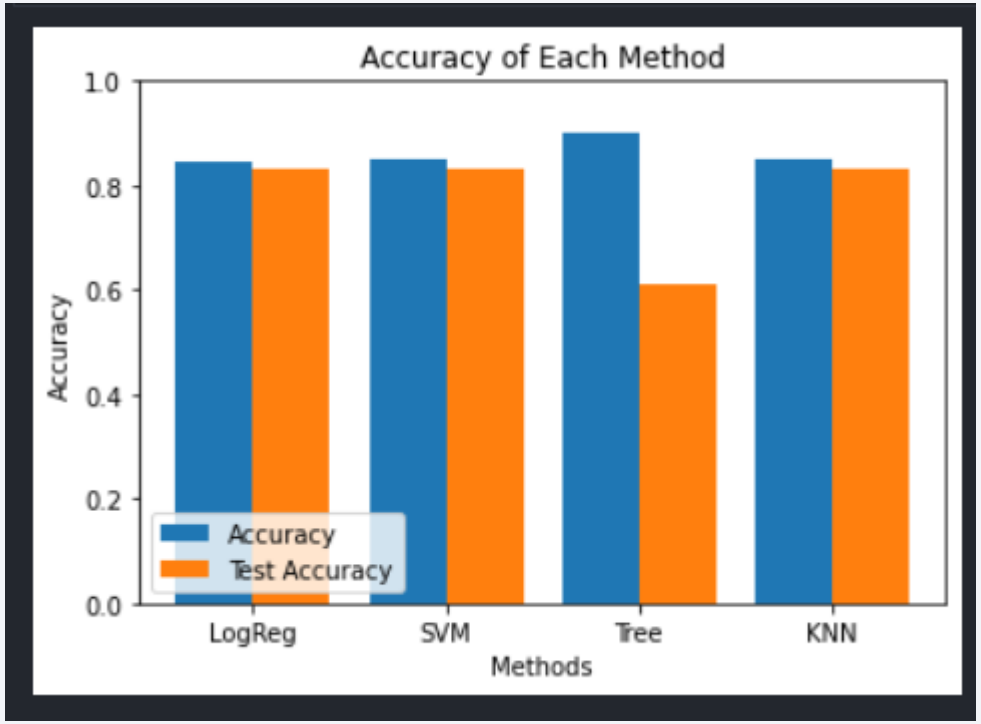
- FT has the most successful outcome

Section 5

Predictive Analysis (Classification)

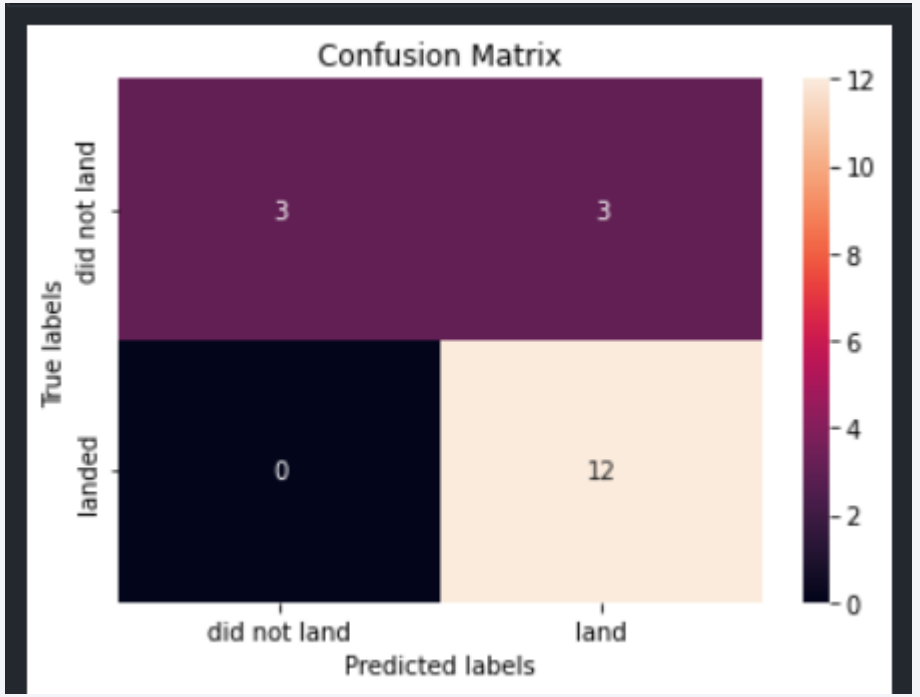
Classification Accuracy

- The tree model has the highest accuracy on train data but also the lowest accuracy on test data



Confusion Matrix

- Confusion matrix of Logistic Regression Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.
- 84% on train data and 83% on test data



Conclusions

- The best launch site is KSC LC-39A
- Success rate seem to be improved over time
- Logistic Regression can be used to predict successful outcome
- Payload about 7000kg seem to be ideal

Thank you!