IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aryan Dani
19/01/2024

Outline

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

Executive Summary

- Collected data from public SpaceX API and SpaceX Wikipedia page. Created labels column 'class' which classifies successful landings. Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features. Changed all categorical variables to binary using one hot encoding. Standardized data and used Grid Search CV to find best parameters for machine learning models. Visualize accuracy score of all models.
- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

Background:

- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Introduction

Problem:

Space Y tasks us to train a machine learning model to predict successful Stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping

  (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Classifying true landings as successful (1) and unsuccessful (0) otherwise

Methodology

Executive Summary

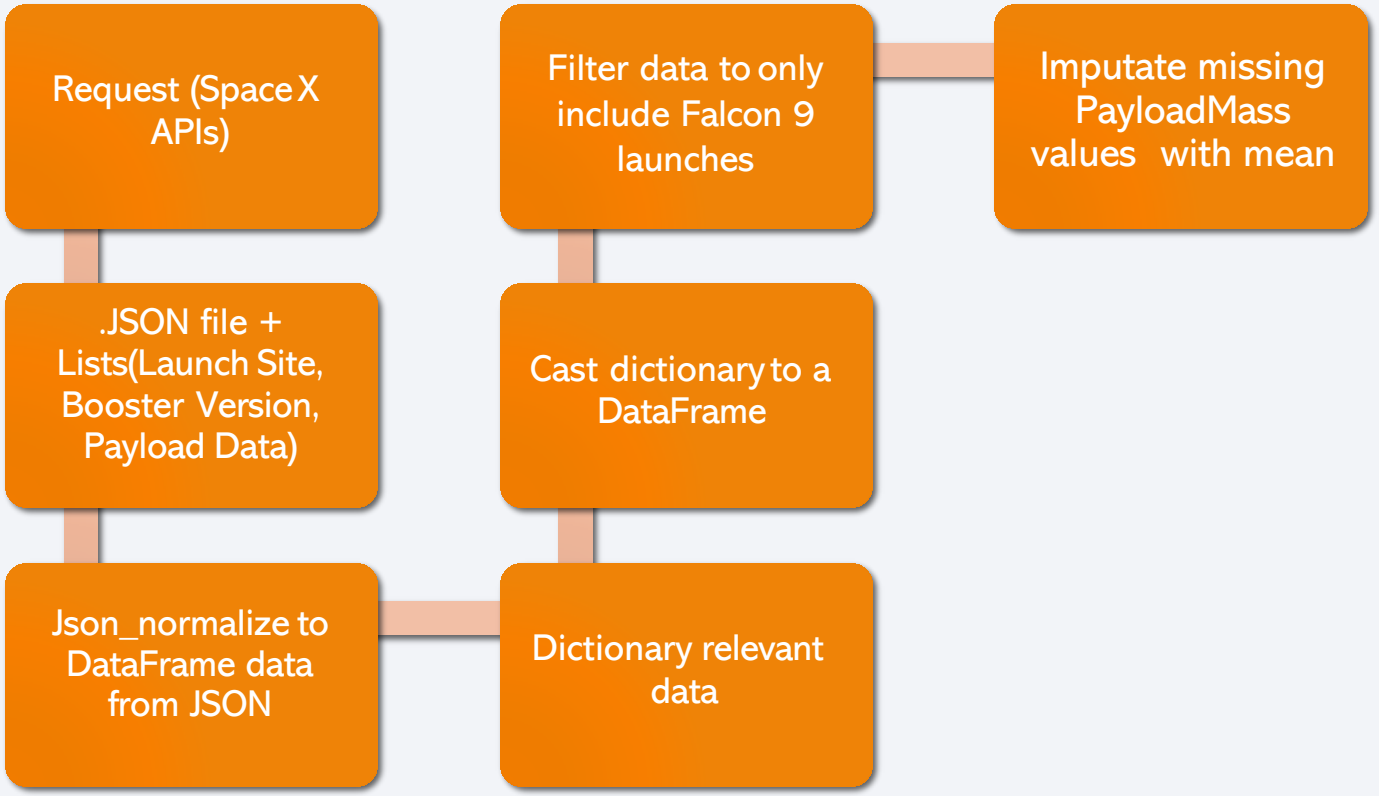
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training
  - and test data sets and evaluated by four different classification models, being
  - the accuracy of each model evaluated using different combinations of
  - parameters.
  - Tuned models using Grid Search CV

Data Collection

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.
- The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from web scraping.
- Space X API Data Columns:
- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,
- Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Webscrape Data Columns:
- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

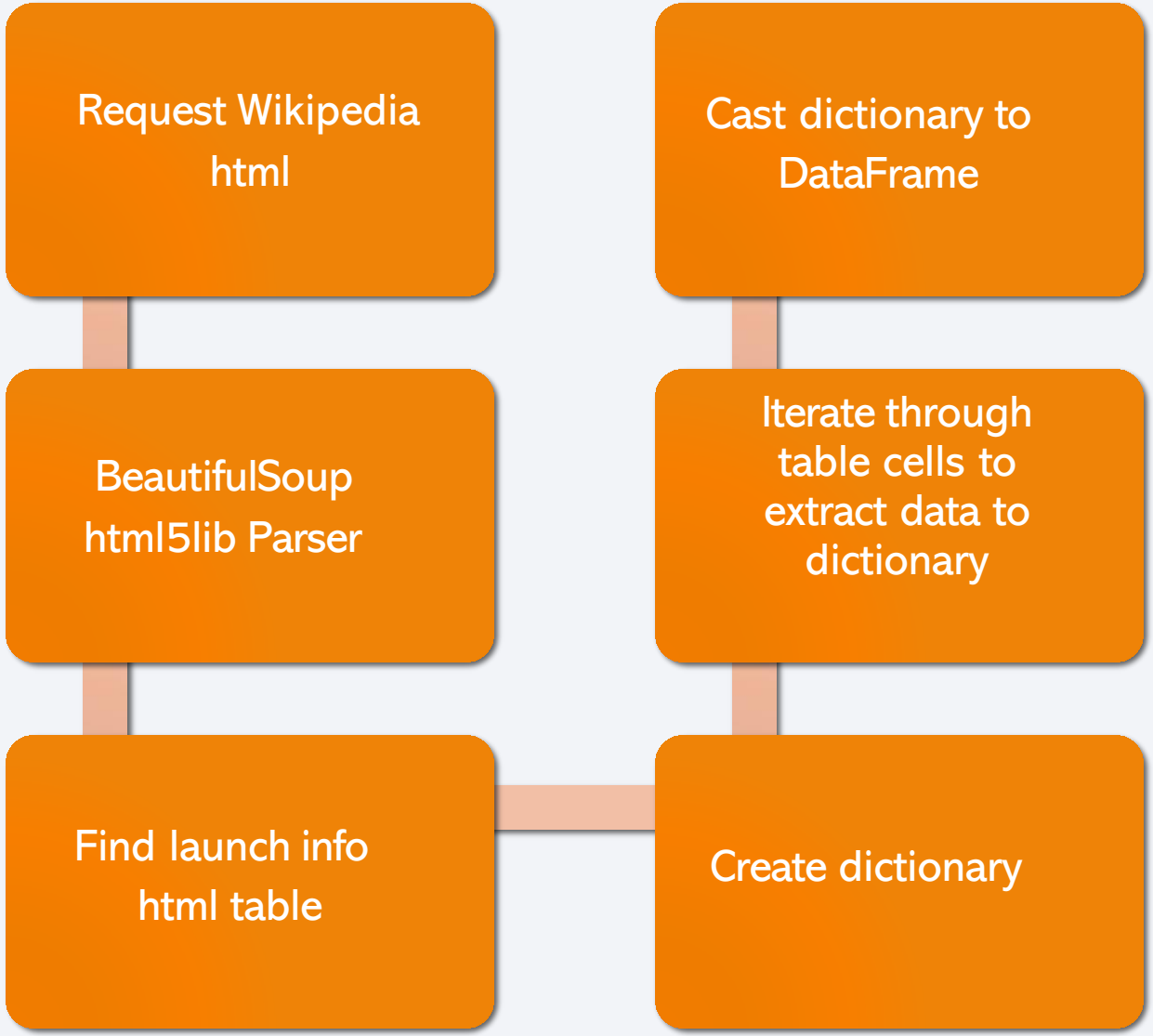
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used.
- This API was used according to the flowchart beside and then data is persisted
- The GitHub URL of the completed SpaceX API calls notebook (https://github.com/duneshime/Capstone_ds/blob/main/Data%20Collection%20API.ipynb)



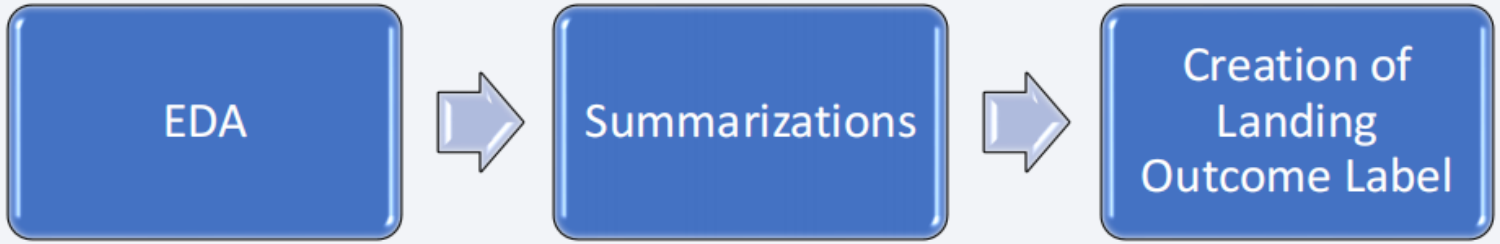
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data is downloaded from Wikipedia according to the flowchart and then persisted.
- The GitHub URL of the completed web scraping notebook: https://github.com/duneshime/Capstone_ds/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

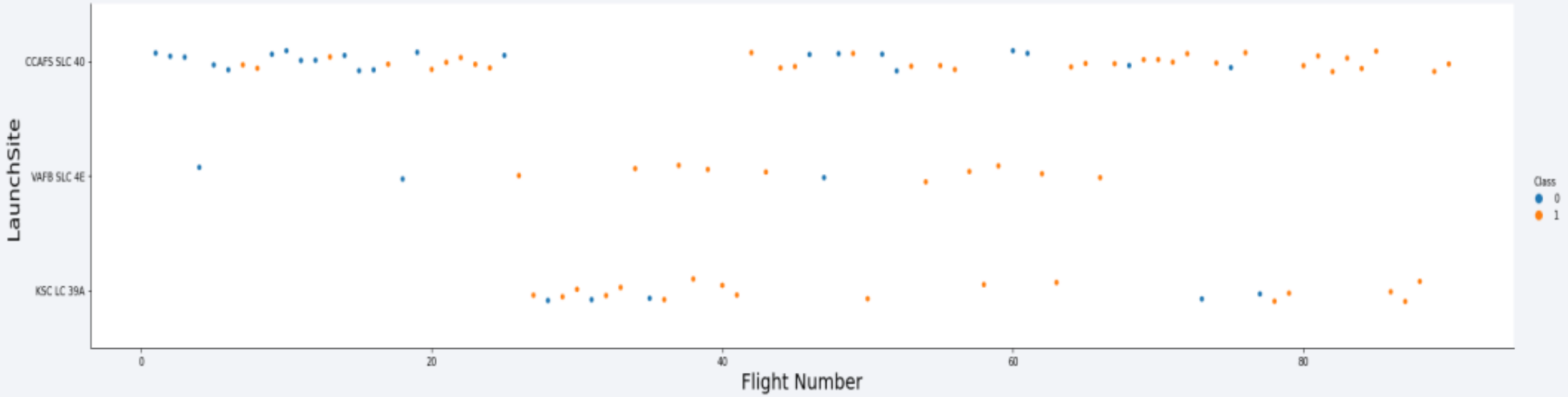
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



- The GitHub URL of the completed data wrangling related notebooks, as an external reference and peer-review purpose is https://github.com/duneshime/Capstone_ds/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- To explore data, 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



- The GitHub URL of the completed EDA with data visualization notebook [Capstone_ds/EDA with Data Visualization.ipynb at main · duneshime/Capstone_ds]

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, and launch site names for in year 2015; and
- 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/duneshime/Capstone_ds/blob/main/EDA.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 coordinate

Source code:
https://github.com/duneshime/Capstone_ds/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data

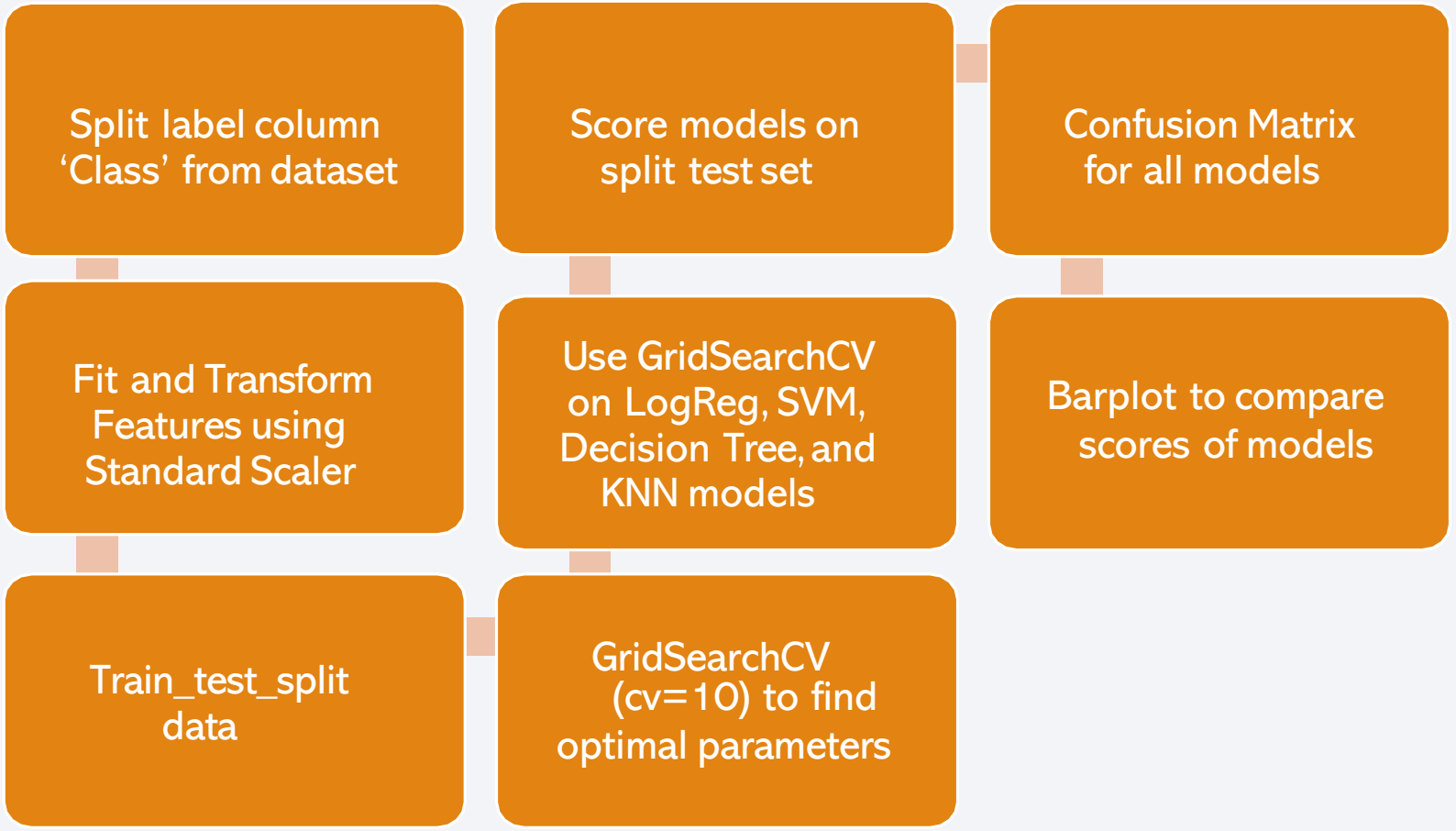
- Percentage of launches by site
- Payload range

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/duneshime/Capstone_ds/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

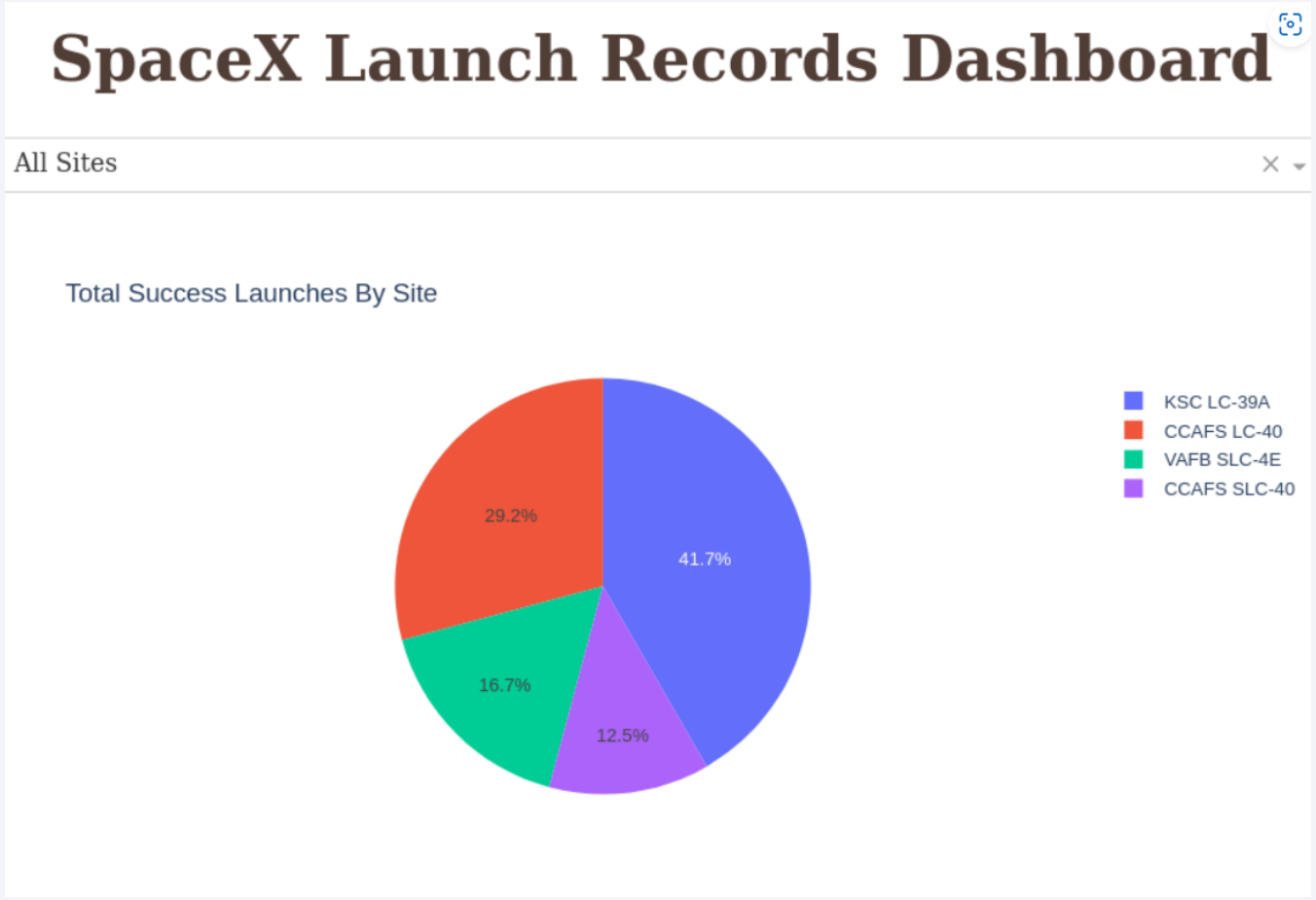


Results

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.

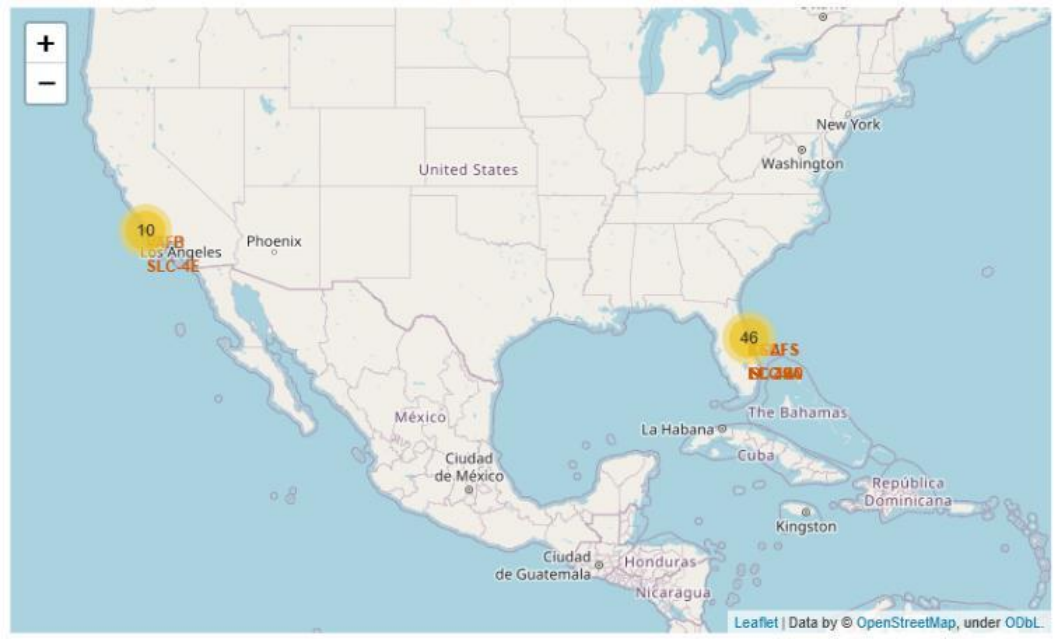
Results



This is a preview of the Plotly dashboard.

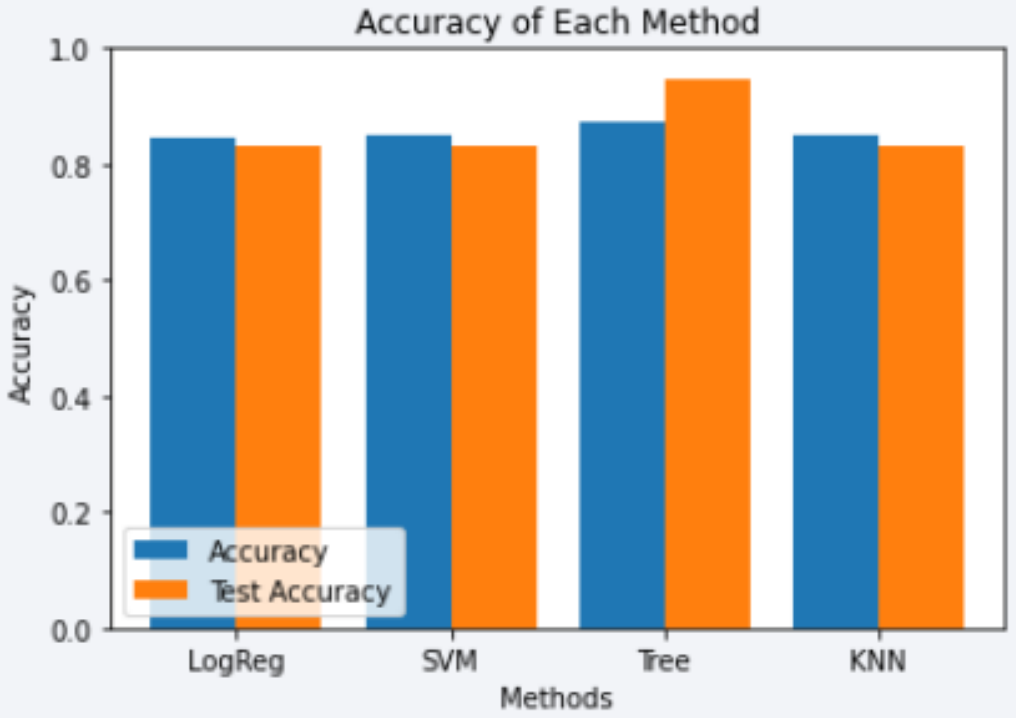
Results

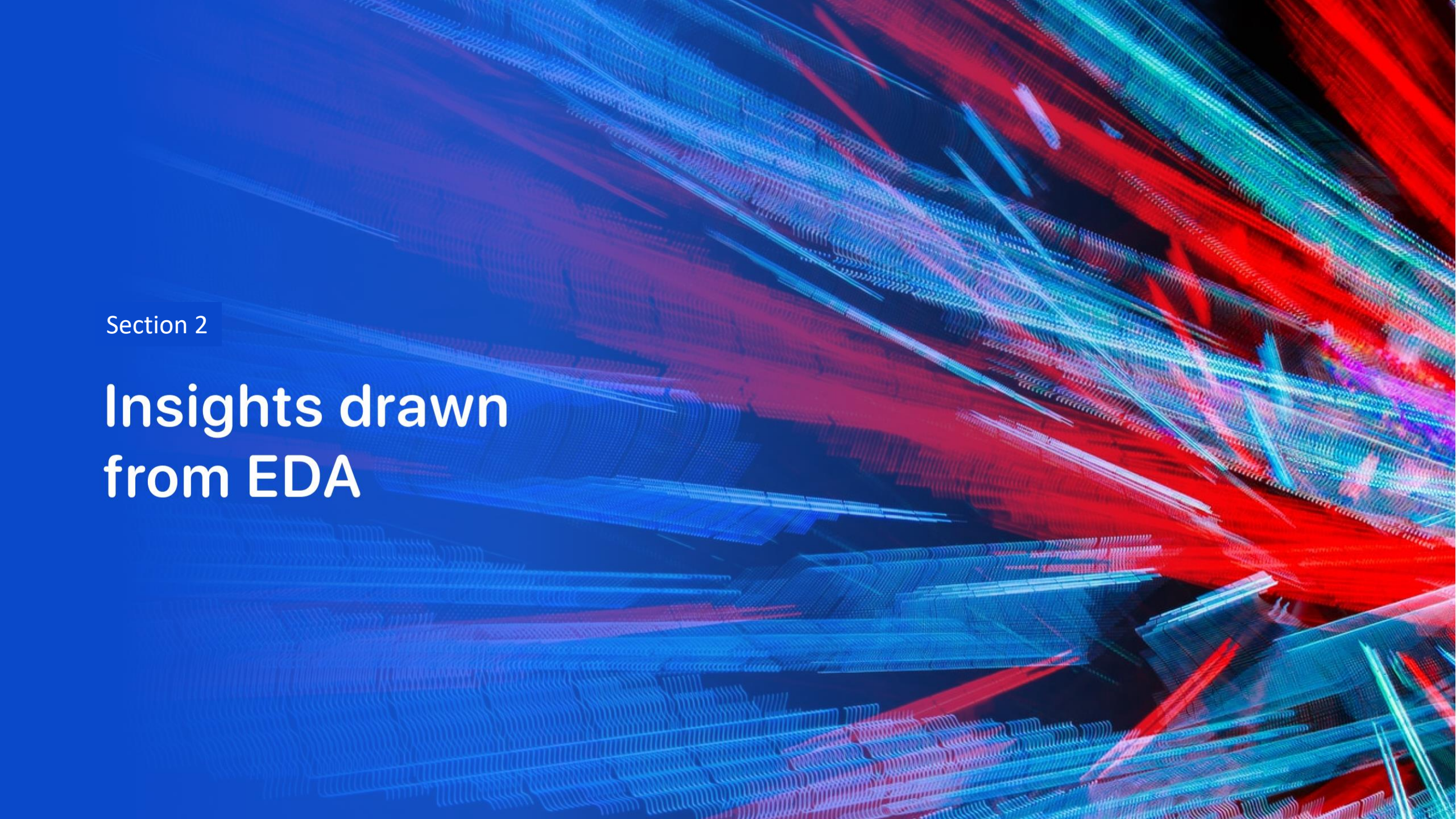
- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.



Results

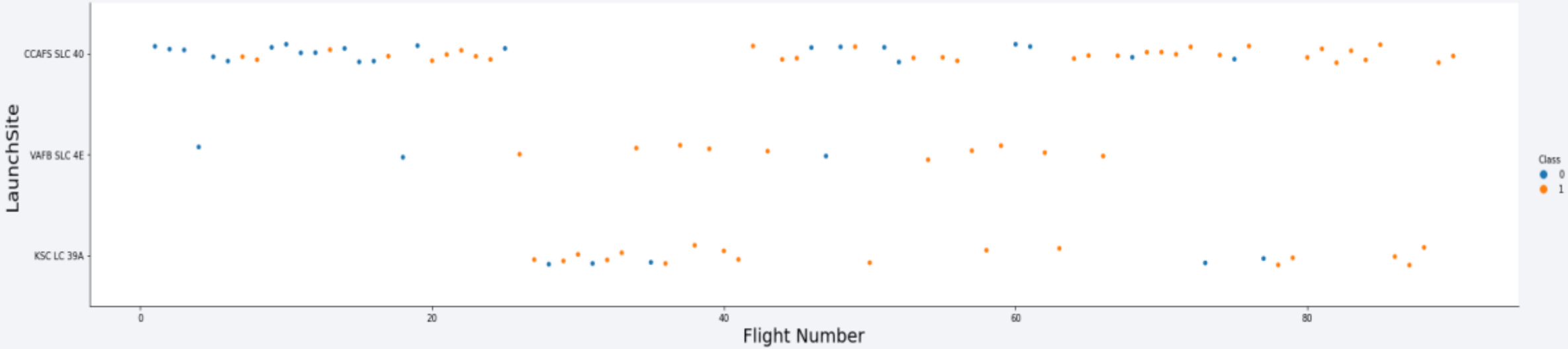
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.




Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Green indicates successful launch; Purple indicates unsuccessful launch.

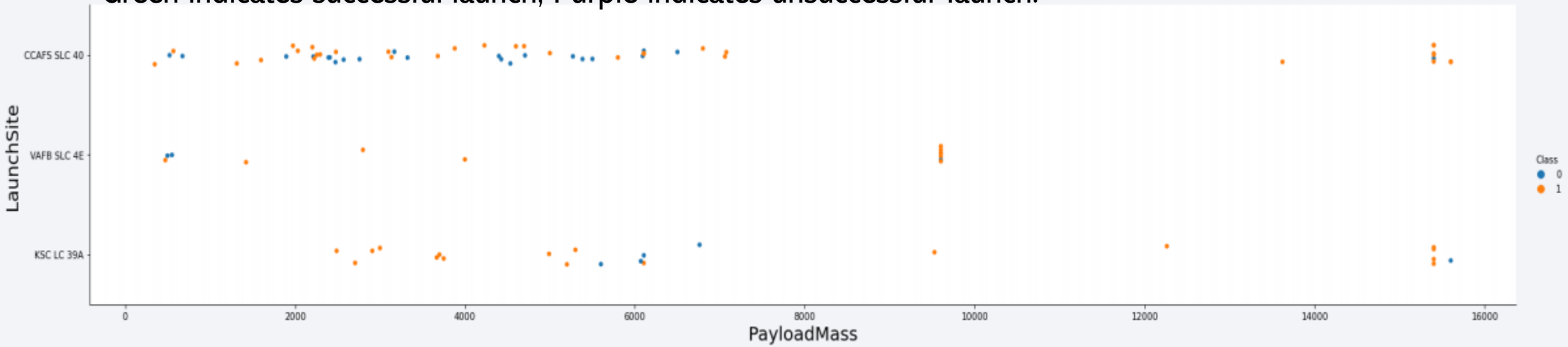


- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

Payload vs. Launch Site

Green indicates successful launch; Purple indicates unsuccessful launch.



- 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.

Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.

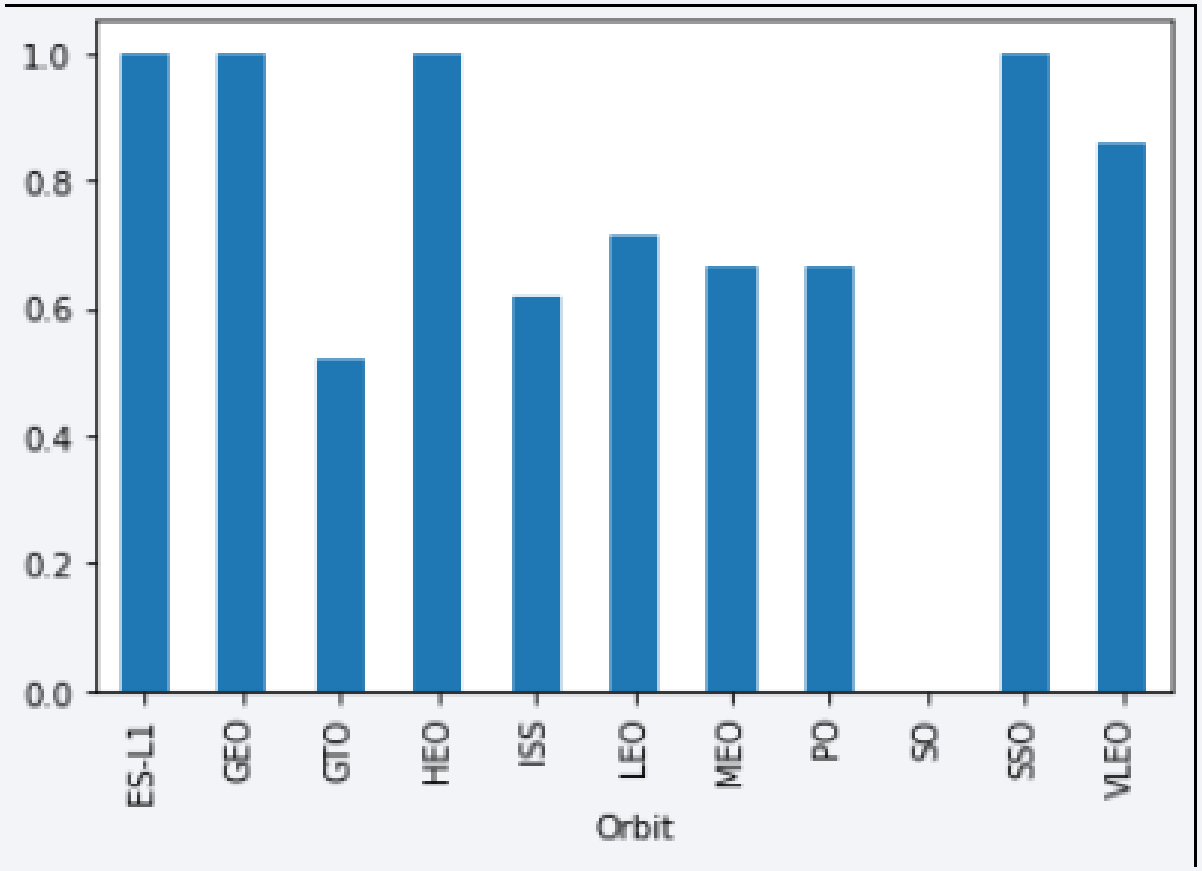
Success Rate vs. Orbit Type

The biggest success rates happens to orbits:

- ES-L1;
- GEO;
- HEO; and
- SSO.

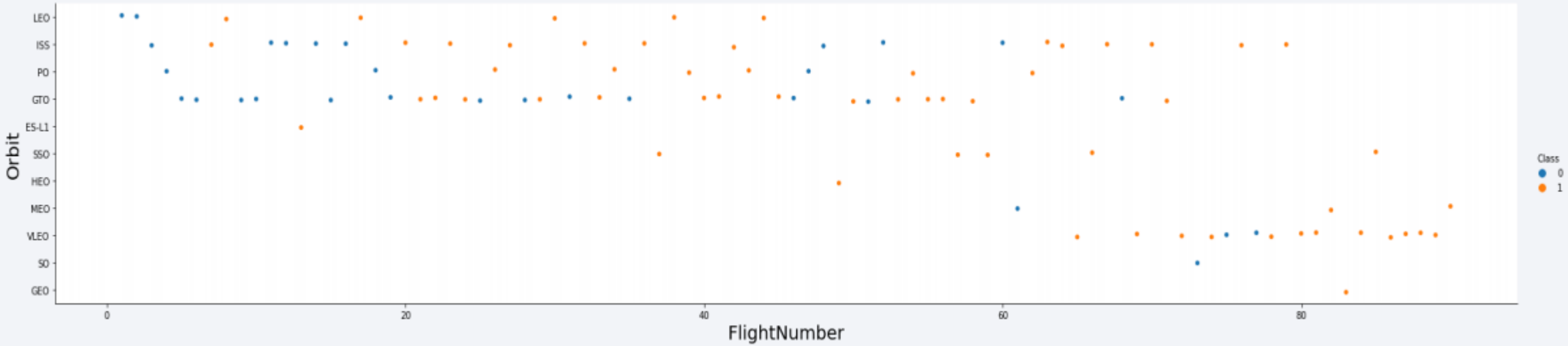
Followed by:

- VLEO (above 80%); and
- LFO (above 70%)



Flight Number vs. Orbit Type

Green indicates successful launch; Purple indicates unsuccessful launch.

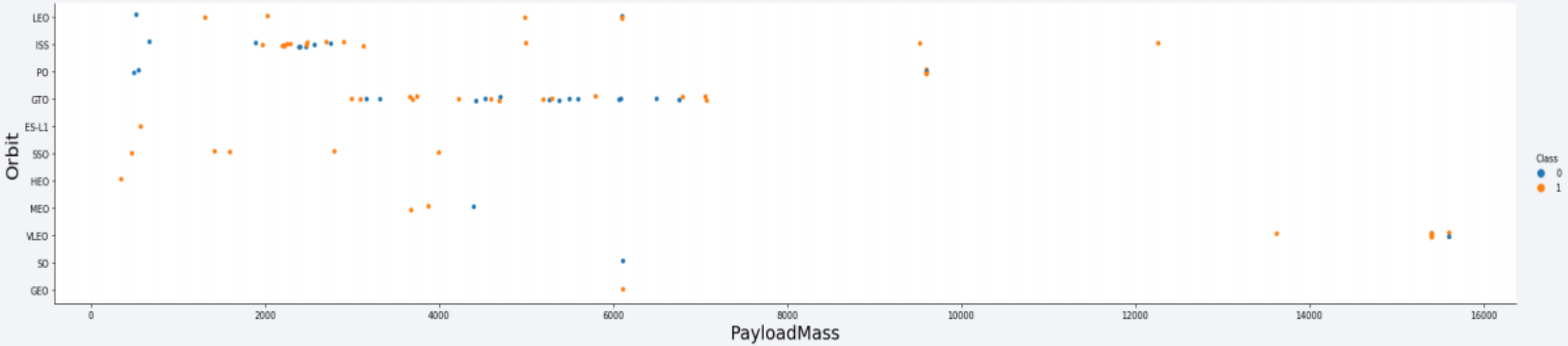


- Apparently, success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type

Green indicates successful launch; Purple indicates unsuccessful launch.

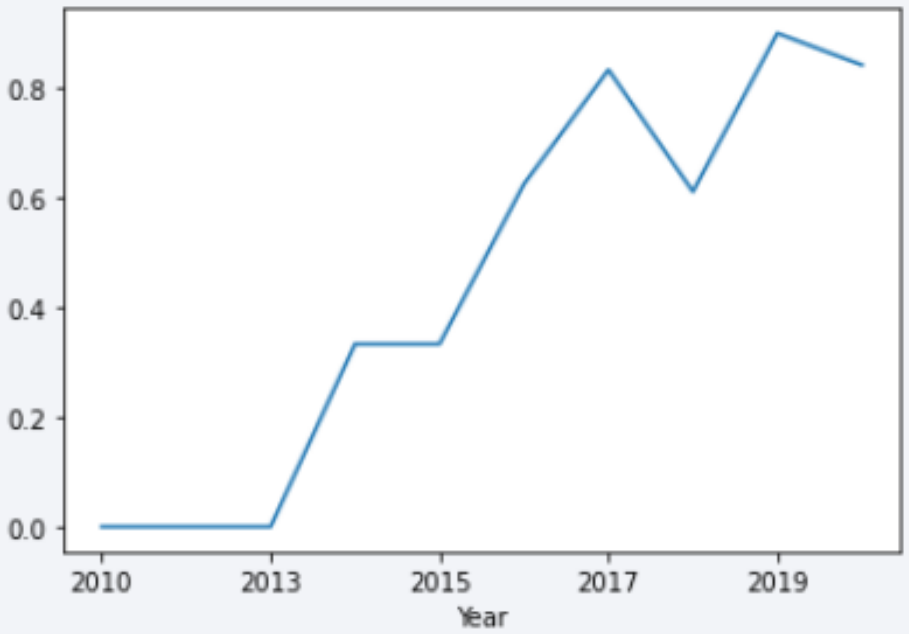


- Apparently, there is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Payload mass seems to correlate with orbit
LEO and SSO seem to have relatively low payload mass
The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.



Success generally increases over time since 2013 with a slight dip in 2018.
Success in recent years at around 80%

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

They are 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`:

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[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	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:

Total Payload (kg)
111.268

Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

Avg Payload (kg)
2.928

Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg

First Successful Ground Landing Date

First successful landing outcome on ground pad:

Min Date
2015-12-22

By 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

- 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

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- The list above has the only two occurrences.

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

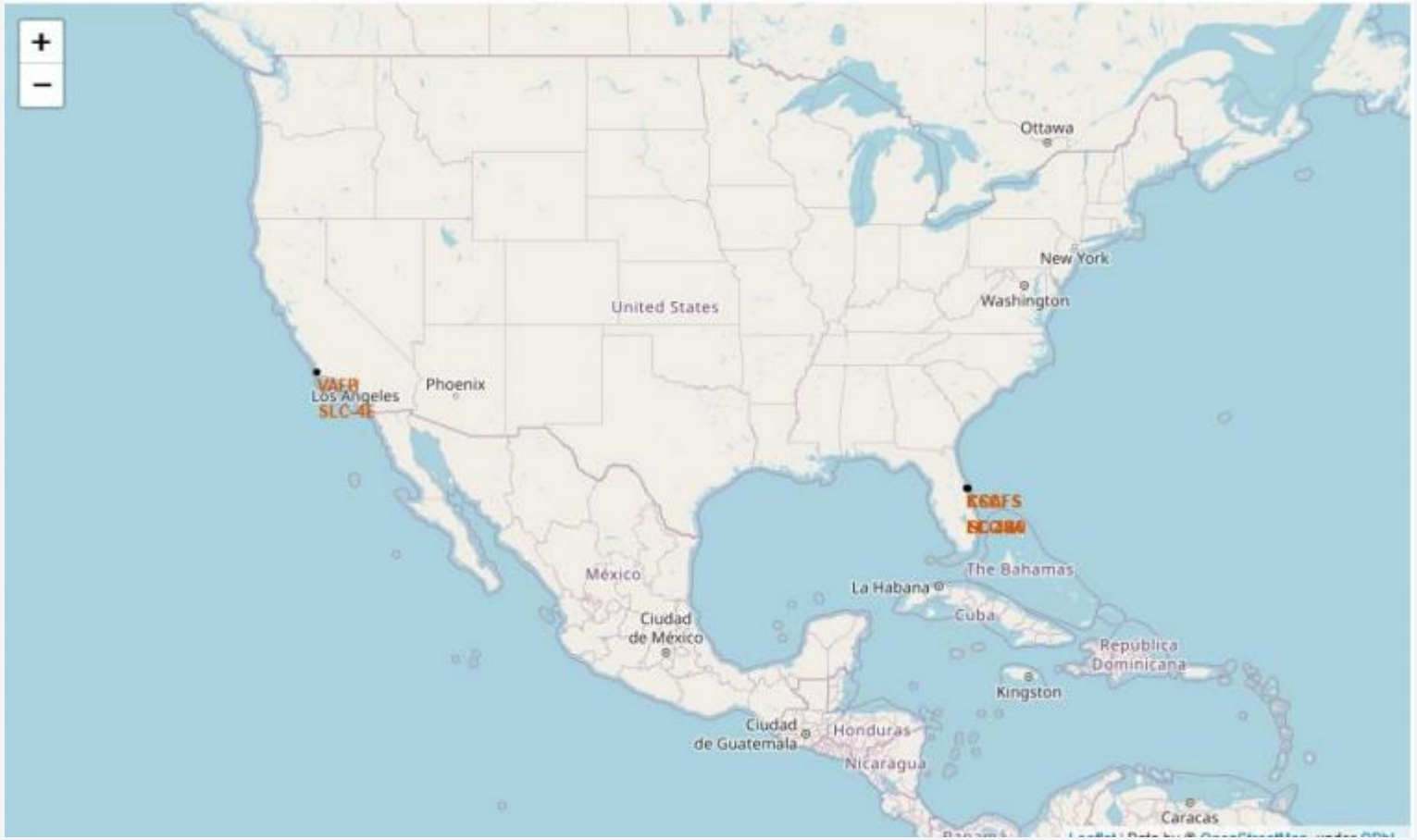
- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

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

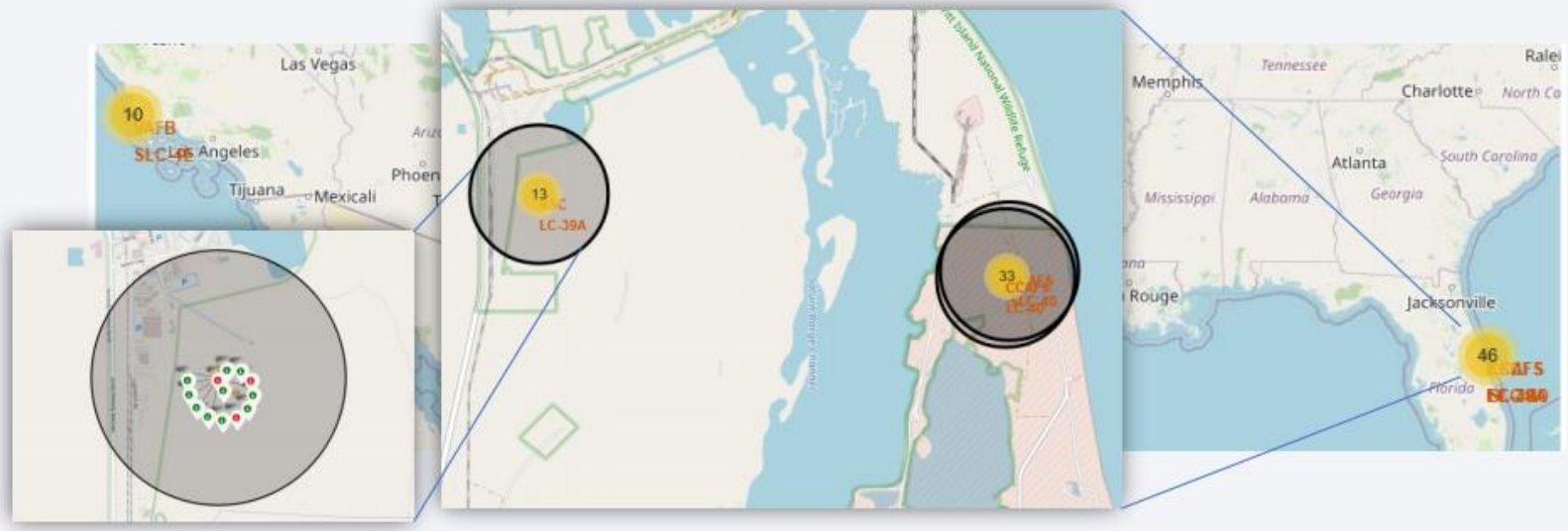
All launch sites



Launch sites are near sea, probably by safety, but not too far from roads and railroads.

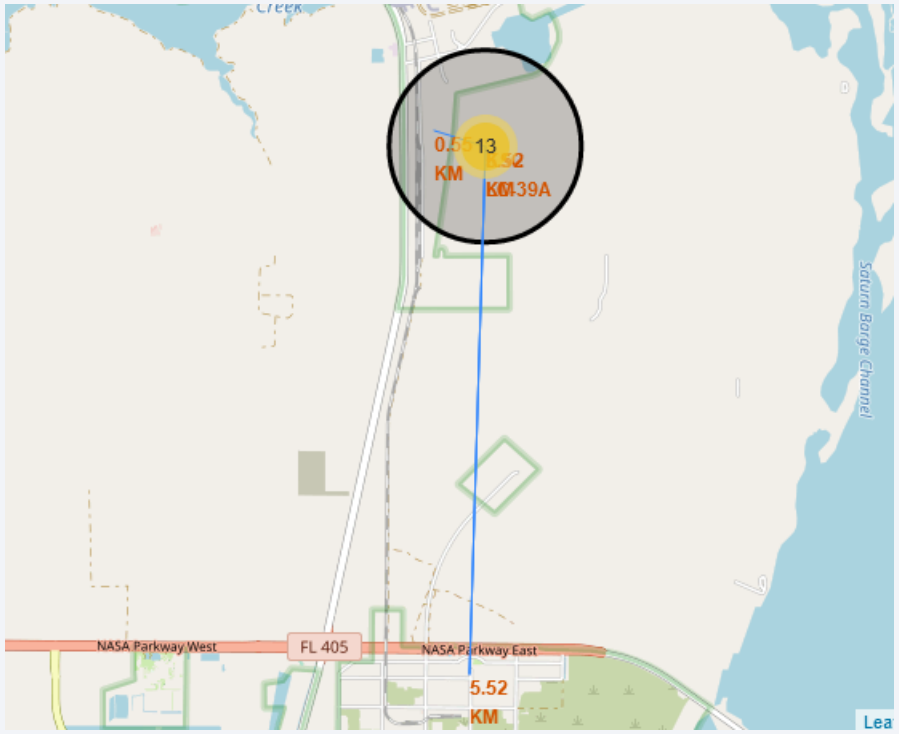
Launch Outcomes by Site

Example of KSC LC-39A launch site launch outcomes.

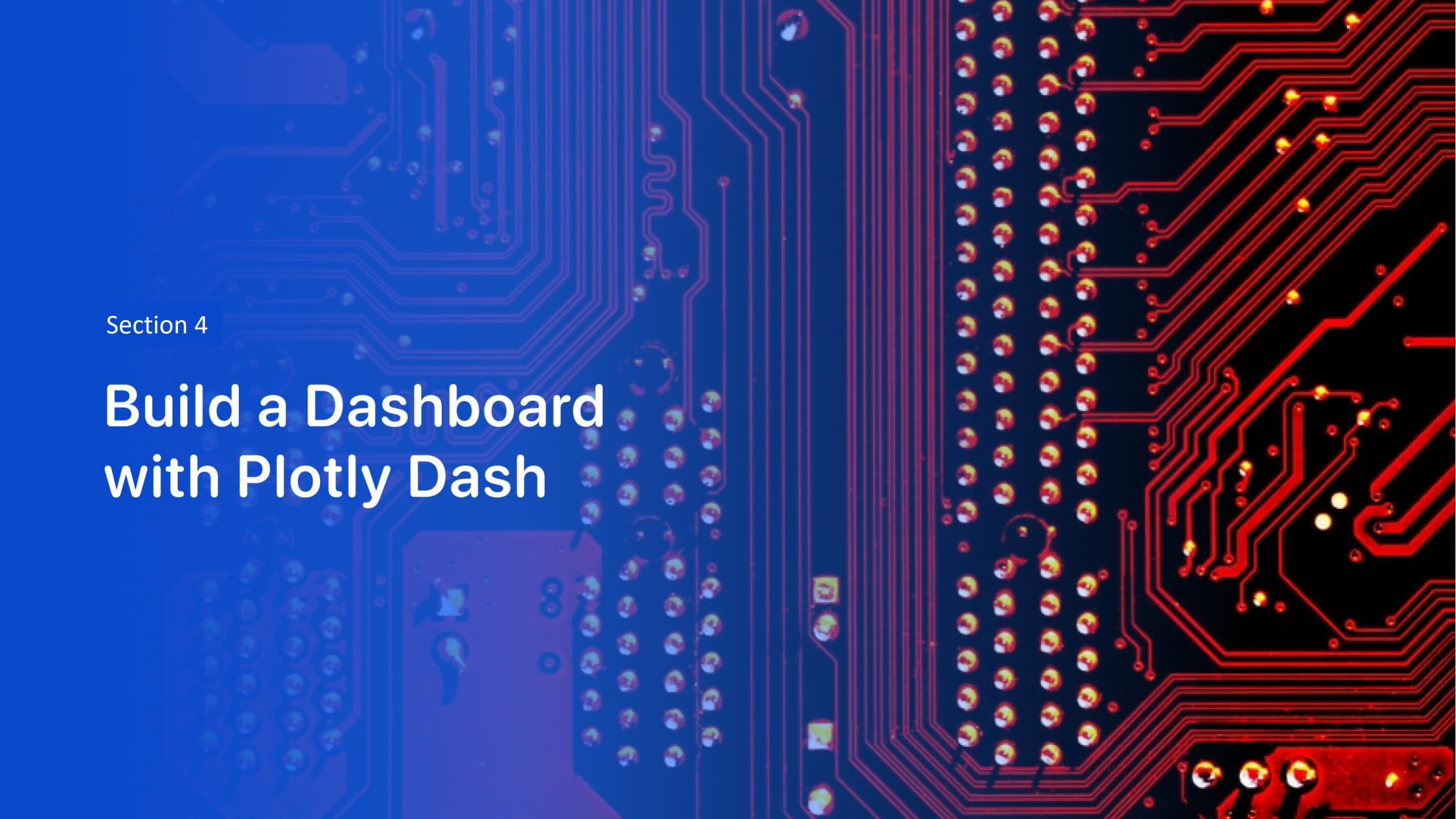


Green markers indicate successful and red ones indicate failure.

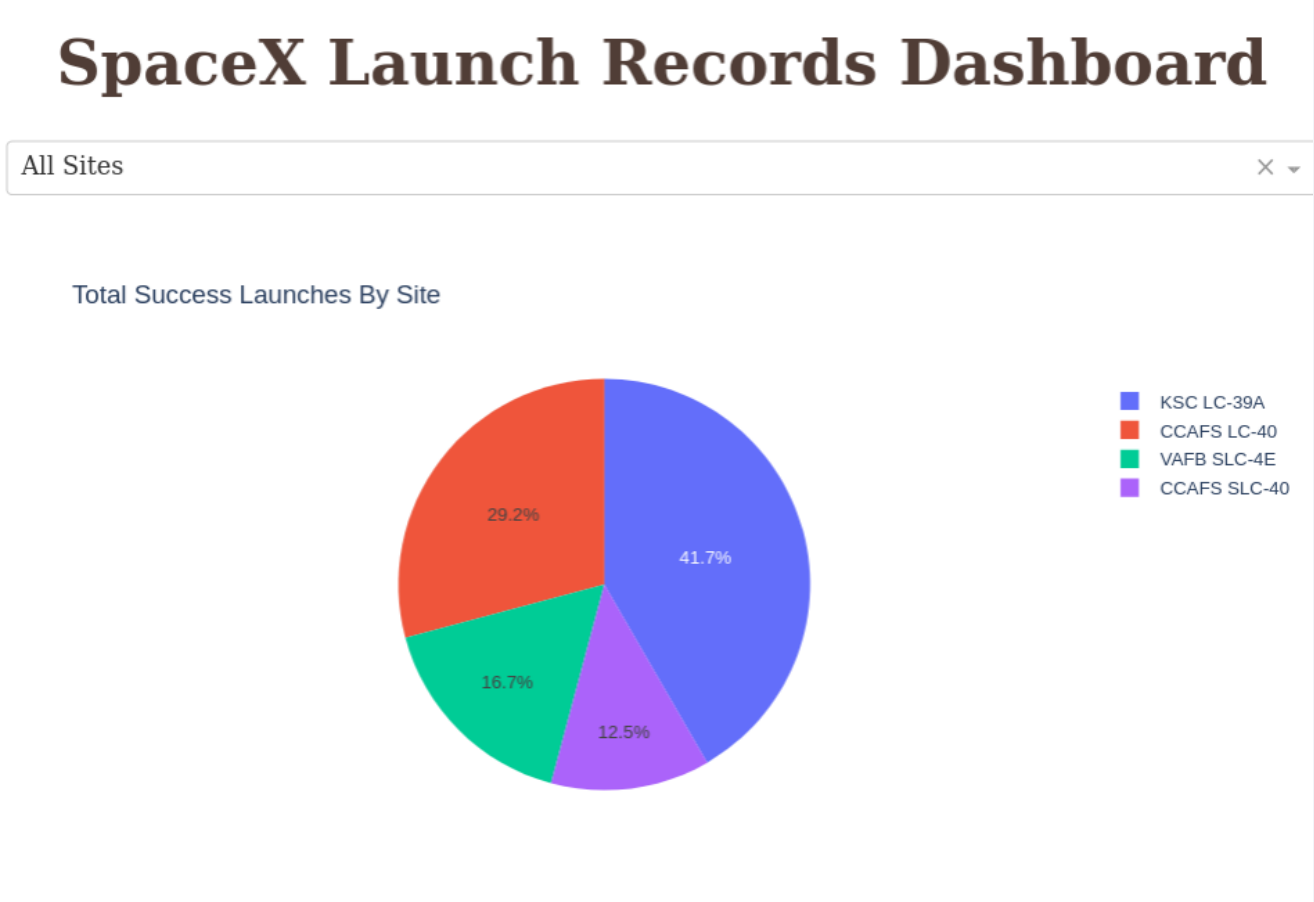
Logistics and Safety



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

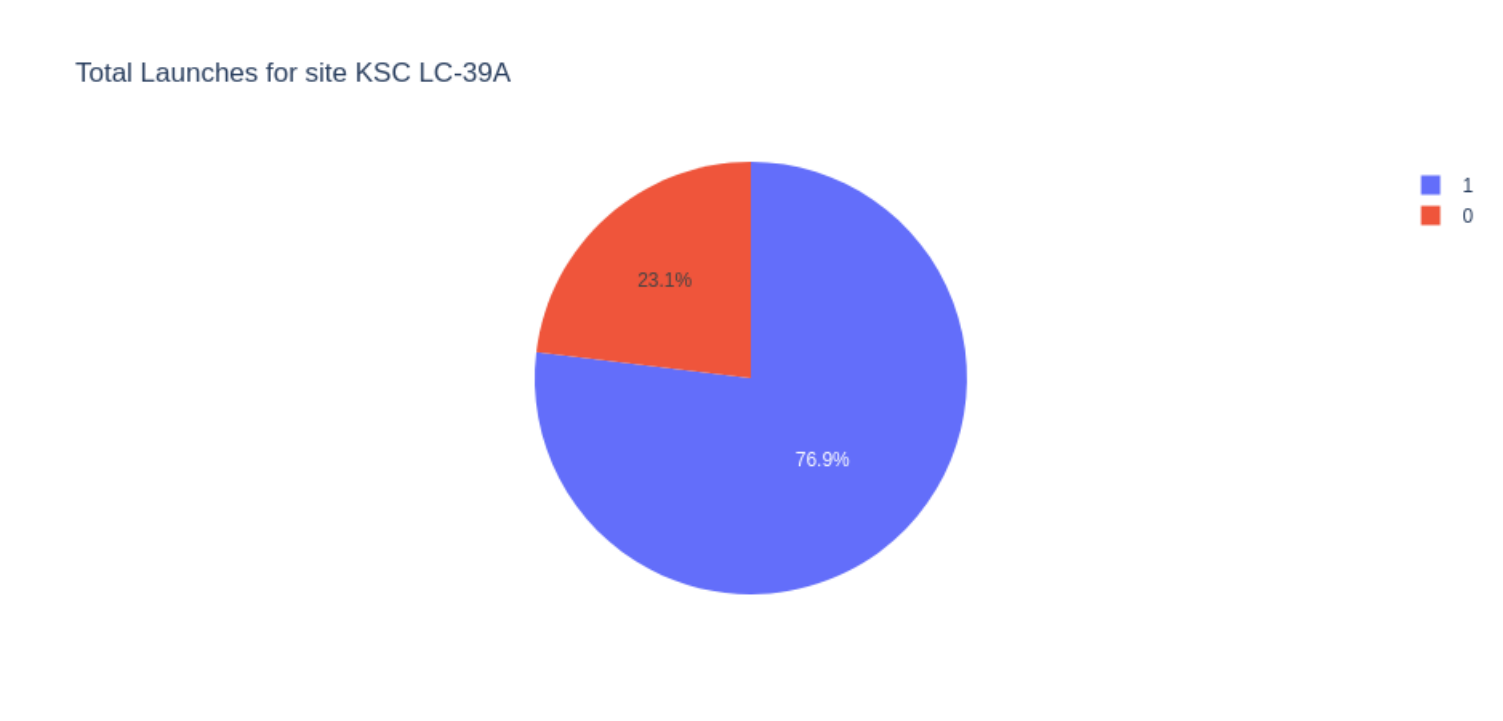

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



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

Launch Success Ratio for KSC LC-39A



76.9% of launches are successful in this site

Payload vs. Launch Outcome

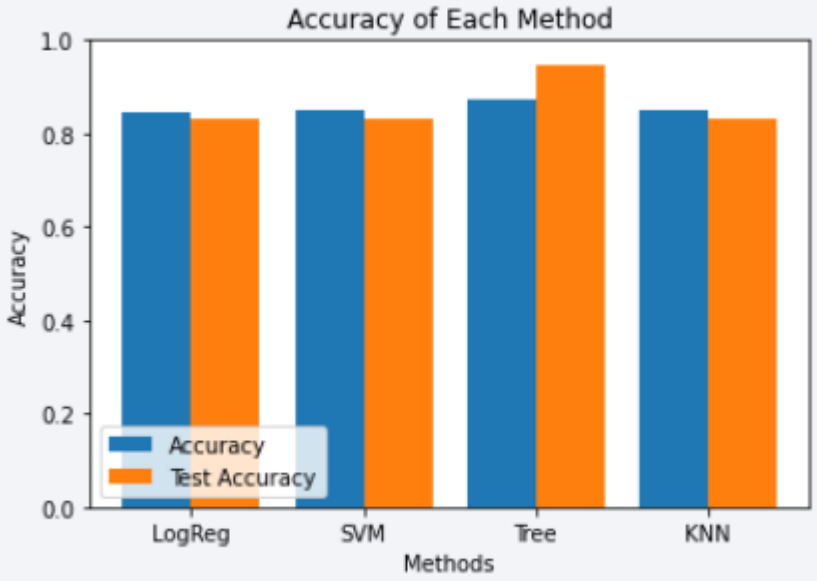
Payloads under 6,000kg and FT boosters are the most successful combination.

Section 5

Predictive Analysis (Classification)

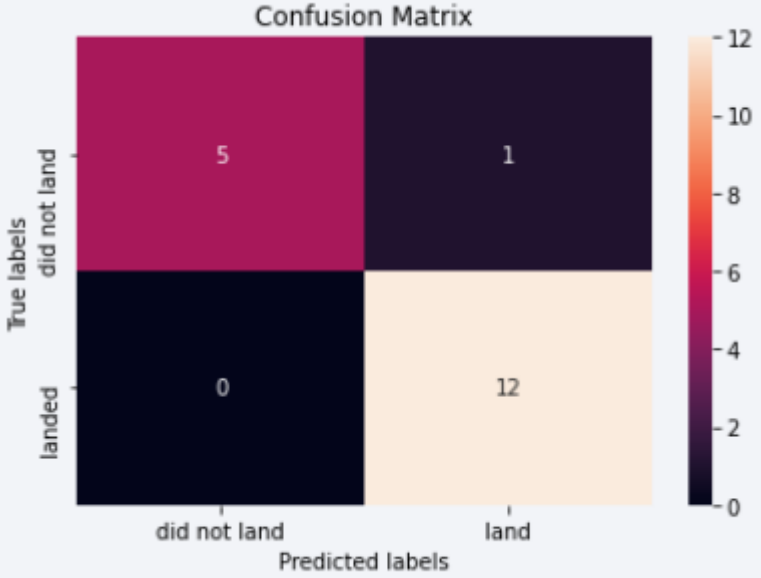
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big 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.
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

Appendix

GitHub repository url:

https://github.com/duneshime/Capstone_ds

Instructors:

- Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Special Thanks to All Instructors:

https://www.coursera.org/professional-certificates/ibm-data-science?#instructors

- As an improvement for model tests, it's important to set a value to np.random.seed variable;
- Folium didn't show maps on Github, so I took screenshots.

Thank you!