IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vincentius Samudro
June 28, 2022

Outline

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

Executive Summary

- The data is obtained by using SpaceX REST API and web scraping and then cleaned and made consistent
- With the data, some EDA, interactive analytics and predictive analysis are performed
- From EDA, we can observe some correlations between flight number, payload mass, orbit type and even the location of the launch site with the outcome
- Interactive analytics gives us detailed information about the success rate from different launch sites and its correlation with booster version
- Predictive analysis using Decision Tree yields the best accuracy

Introduction

- With its reusable rockets, SpaceX is leading in the aerospace industry. The innovation makes SpaceX to be able to save significant amount of cost
- Predicting whether the rockets would land successfully becomes an important matter here, since the cost then can be determined from there
- The information in this report can be used as a way to indirectly estimate the cost of a launch
- Beside the main goal to predict whether a flight will be successful, we also get to look for other important aspects of the past SpaceX flights

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data is obtained directly from SpaceX using its REST API
  - Web scraping is also done for collecting data from Wikipedia
- Perform data wrangling
  - Since the data is in JSON formatting, the data is then converted to Pandas Dataframe to make things simpler
  - Furthermore, the data is filtered so that only Falcon 9 boosters are shown
  - The data is also cleaned from null and inconsistent formatting
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Hyperparameters are exhaustively searched using GridSearchCV
  - These parameters are then applied to training data
  - Finally, the parameters are validated using the score() method to the test data
  - Confusion matrix is also utilized to put emphasis on the validation

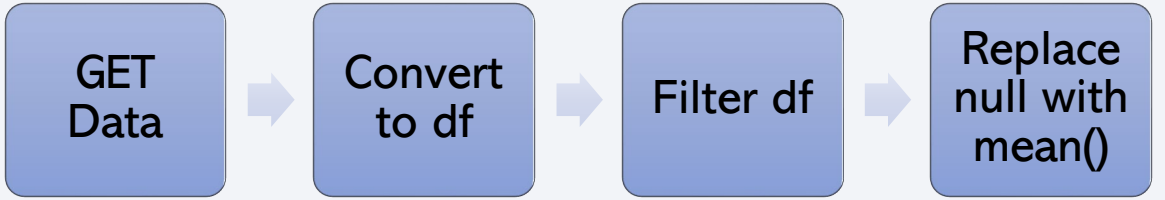
Data Collection

- The data is obtained using **SpaceX REST API**
  - Since every column/feature needed is located in different endpoints, we have to access every endpoints
  - To automate this process, the endpoint is iterated using self-made function in Python to gather data for each necessary column
- Some data is also collected from **SpaceX Wikipedia table** using **BeautifulSoup web scraper**
  - The HTML page will be treated as BeautifulSoup object so that elements contained there can be scraped easily
  - The table data in the page is iterated and every iteration is put to Pandas Dataframe

Data Collection – SpaceX API

- Flowchart of the data collection via SpaceX REST API

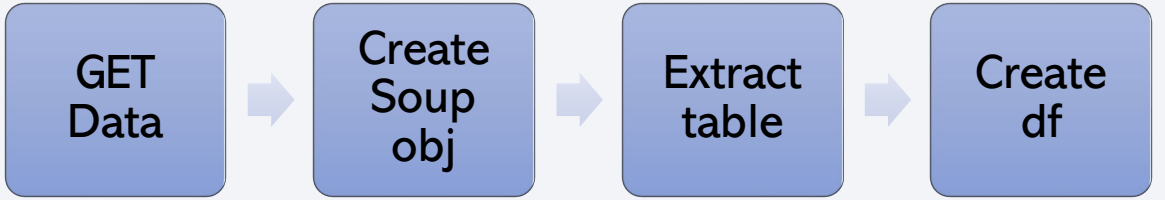
https://github.com/vincadrn/datasci/blob/main/Data%20Collection%20by%20REST%20API.ipynb



Data Collection - Scraping

- Flowchart of the data collection via web scraping using BeautifulSoup

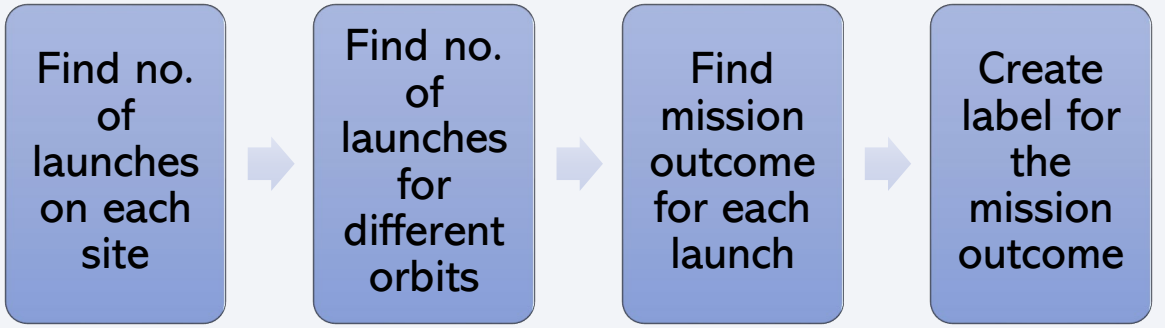
https://github.com/vincadrn/datasci/blob/main/Data%20Collection%20by%20Web%20Scraping.ipynb



Data Wrangling

- Here, data wrangling is done primarily for labeling each row with “success” or “failure”

https://github.com/vincadrn/datasci/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- These following charts are used to see the relationship between the available data and the mission outcome (“success” or “failure”):
  - Payload mass vs flight number, with color label on the outcome
  - Launch site vs flight number, with color label on the outcome
  - Launch site vs payload mass, with color label on the outcome
  - Outcome vs orbit

EDA with Data Visualization

- Orbit vs flight number, with color label on the outcome
- Orbit vs payload mass, with color label on the outcome
- Outcome vs year

https://github.com/vincadrn/datasci/blob/main/Visualization%20EDA.ipynb

EDA with SQL

- These SQL queries are performed to gain some insight about the data:
  - Display the names of the four launch sites
  - Display 5 records where the launch sites begin with ‘CCA’
  - Display the total payload mass carried by NASA (CRS)
  - Display the average payload mass carried by booster version of F9 v1.1
  - Show the first date of successful landing outcome in ground pad

EDA with SQL

- List the names of the boosters which have successful landing in drone ship AND payload mass between 4000 and 6000 kg
- List the total number of successful and failure outcomes
- List the booster versions which have carried the maximum payload mass
- List every booster versions and launch site names which have failed landing in drone ship
- Rank the count of landing outcomes between June 4, 2010 and March 20, 2017 in descending order

- https://github.com/vincadrn/datasci/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- These map objects are created and/or added to the site map:
  - Map: as the base object
  - Circle: to locate and show the given coordinate a radius
  - Marker: to make text label such as the site name and distance
  - MarkerCluster: to be able to expand and collapse the smaller marker showing the outcomes
  - MousePosition: to be able to have a mouse pointer and the pointed coordinate
  - PolyLine: to make line to show distance from the launch site to some near landmarks

https://github.com/vincadrn/datasci/blob/main/Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- These visualizations are used in Plotly Dash:
  - Pie chart to visualize the successful outcomes for all sites
  - Pie chart to visualize both outcomes for each sites
  - Scatter chart to visualize outcomes vs payload mass

https://github.com/vincadrn/datasci/blob/main/Dashboard%20with%20Plotly.py

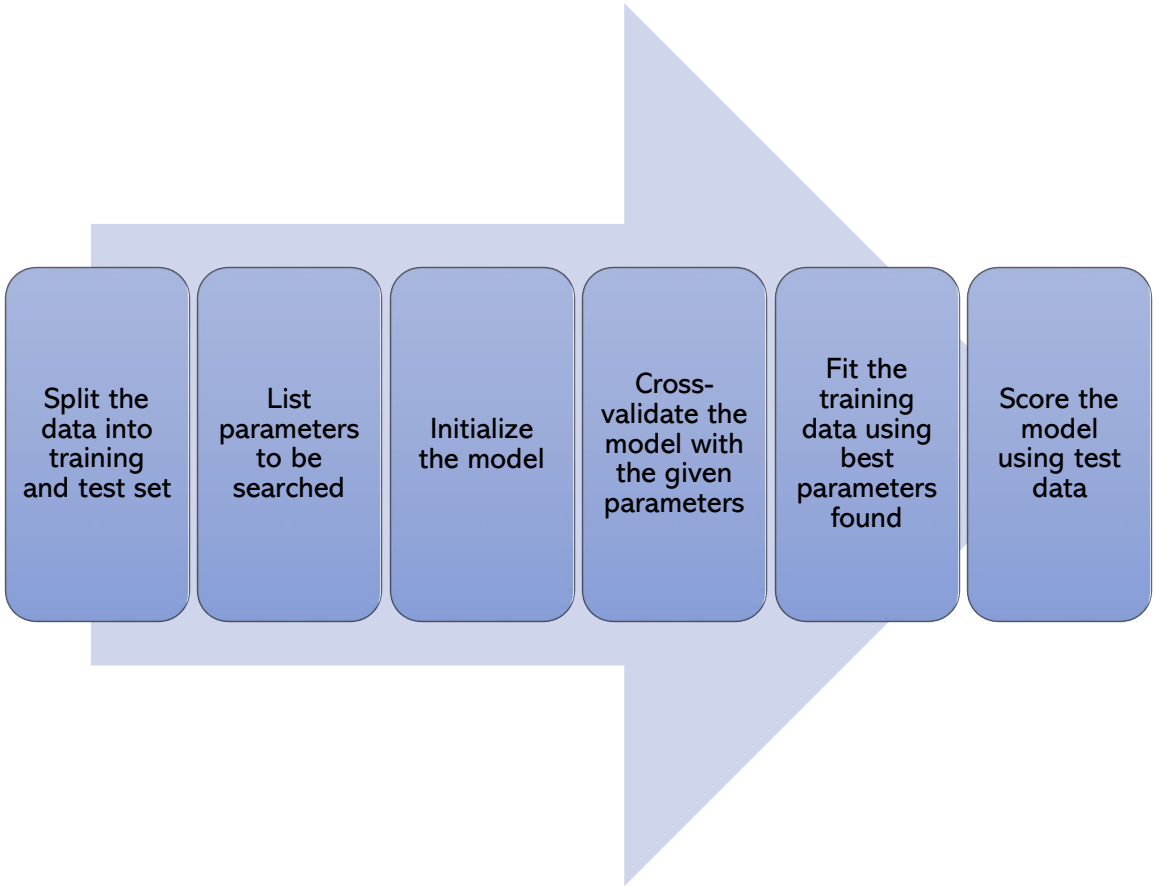
Predictive Analysis (Classification)

- Classification model are implemented using these algorithm:
  - Logistic regression
  - SVM
  - Decision tree
  - KNN
- Parameters for these algorithms are exhaustively searched using GridSearchCV

Predictive Analysis (Classification)

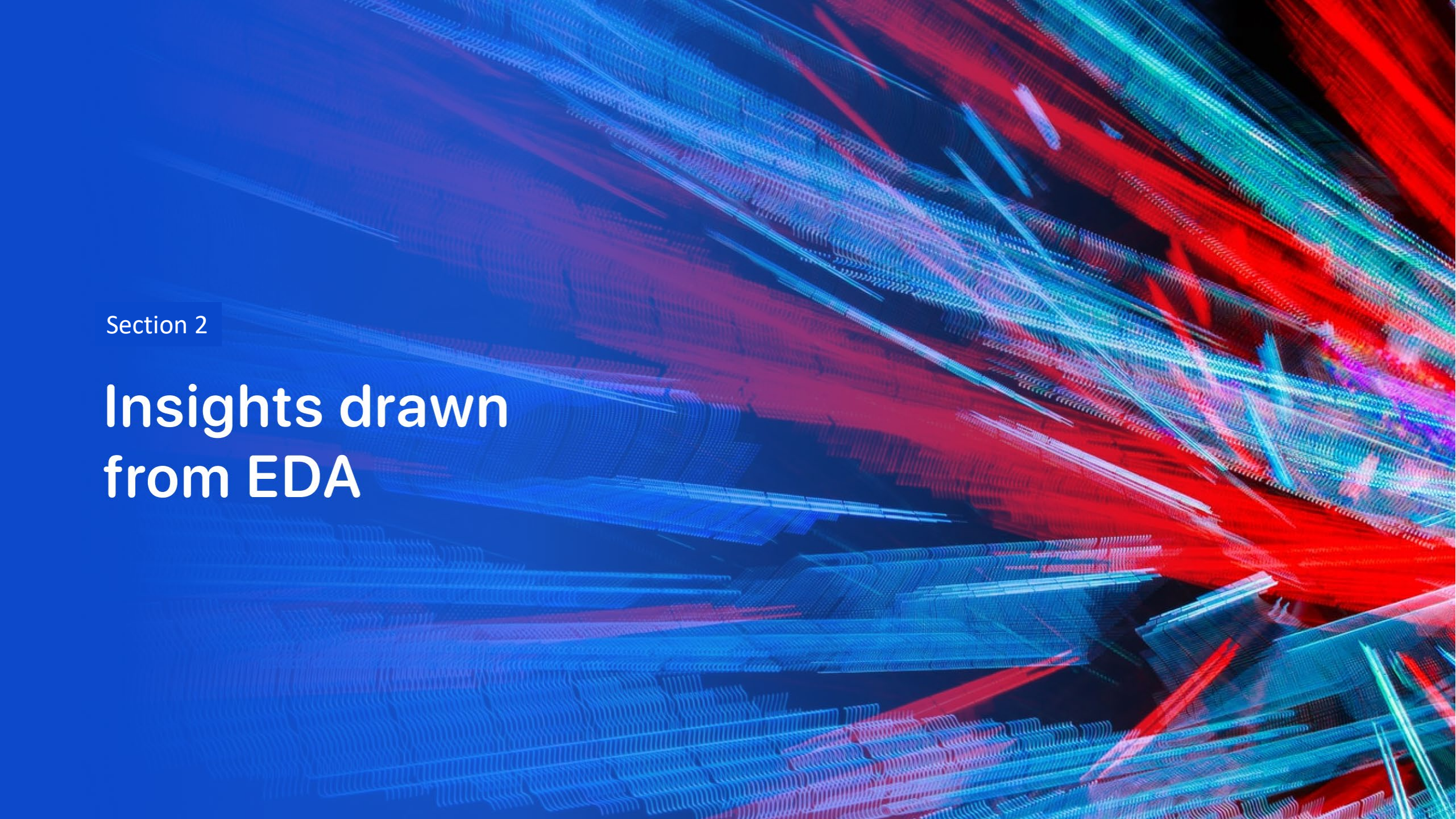
- All models are developed in similar manner as depicted in the flowchart

https://github.com/vincadrn/datasci/blob/main/Predictive%20Analysis.ipynb



Results

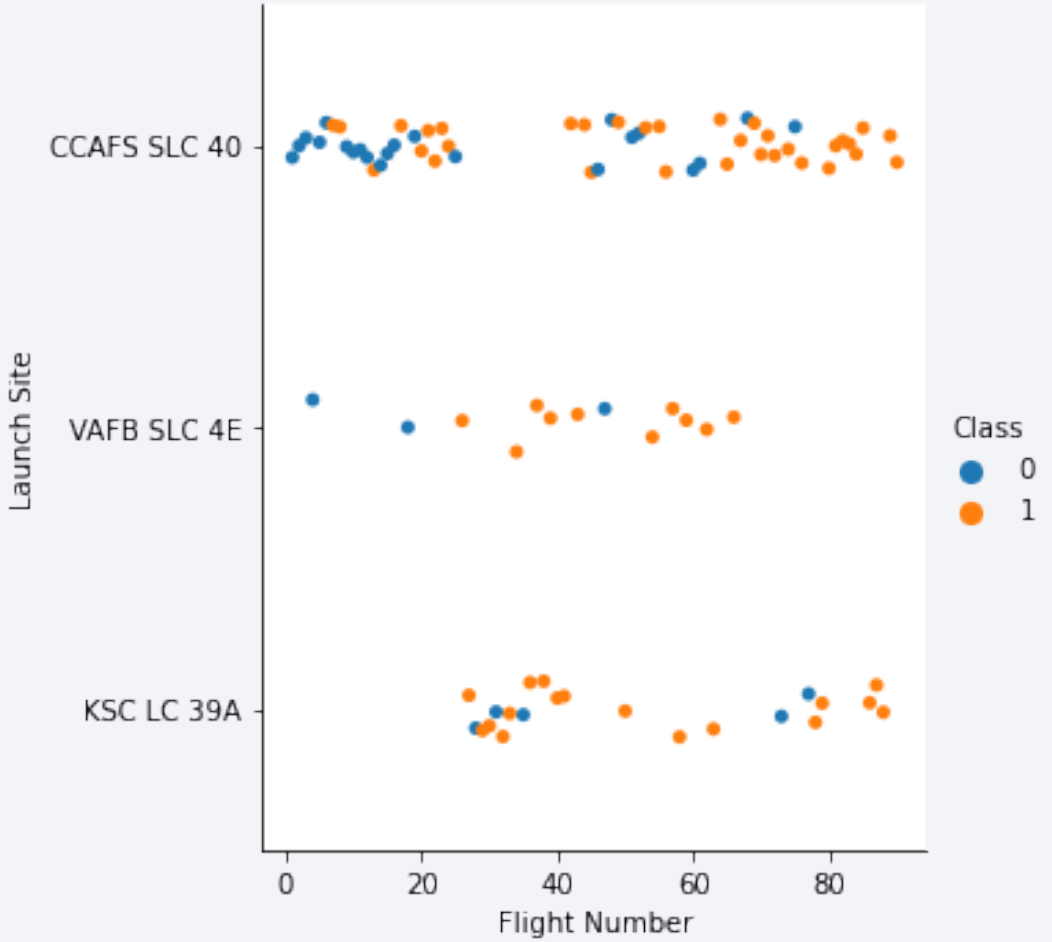
- From EDA and interactive analytics, insights are gained and will be explained further in the next section(s)
- From predictive analysis, best classifier(s) to model the data are obtained


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Significant correlation of successful outcome is found in VAFB SLC 4E and KSC LC 39A, observed from the proportion of yellow dots compared to the blue ones.

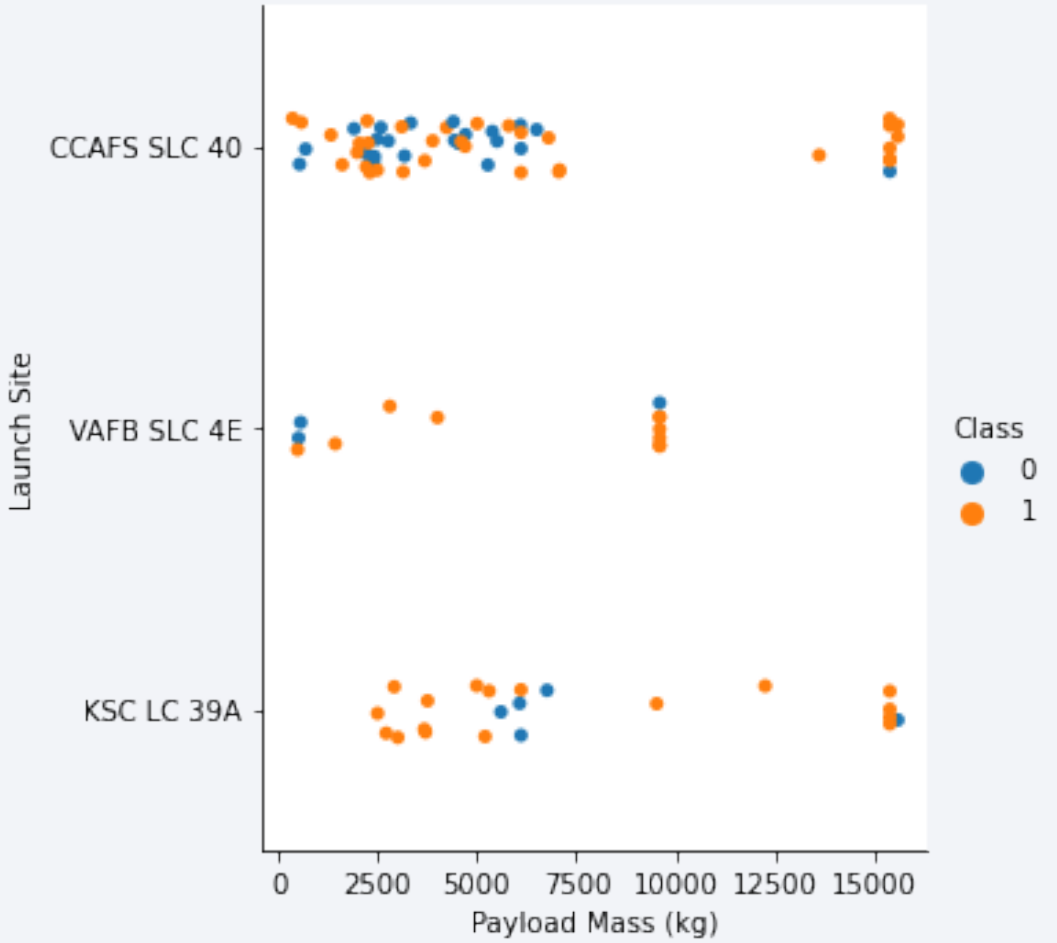
Also, as the flight number increases, the success rate also increases.



Payload vs. Launch Site

In CCAFS SLC 40, the success rate is high in the extreme end of the payload mass.

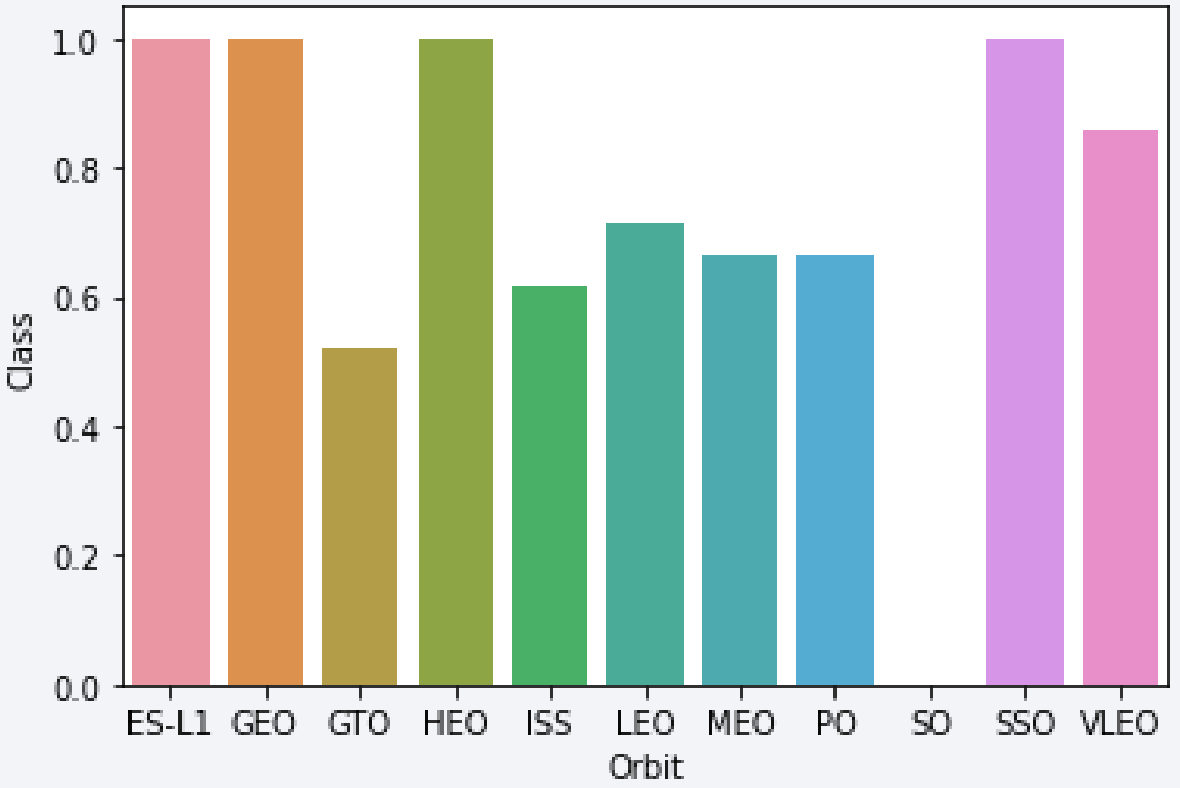
In other launch sites, the payload mass is observed to be less relevant



Success Rate vs. Orbit Type

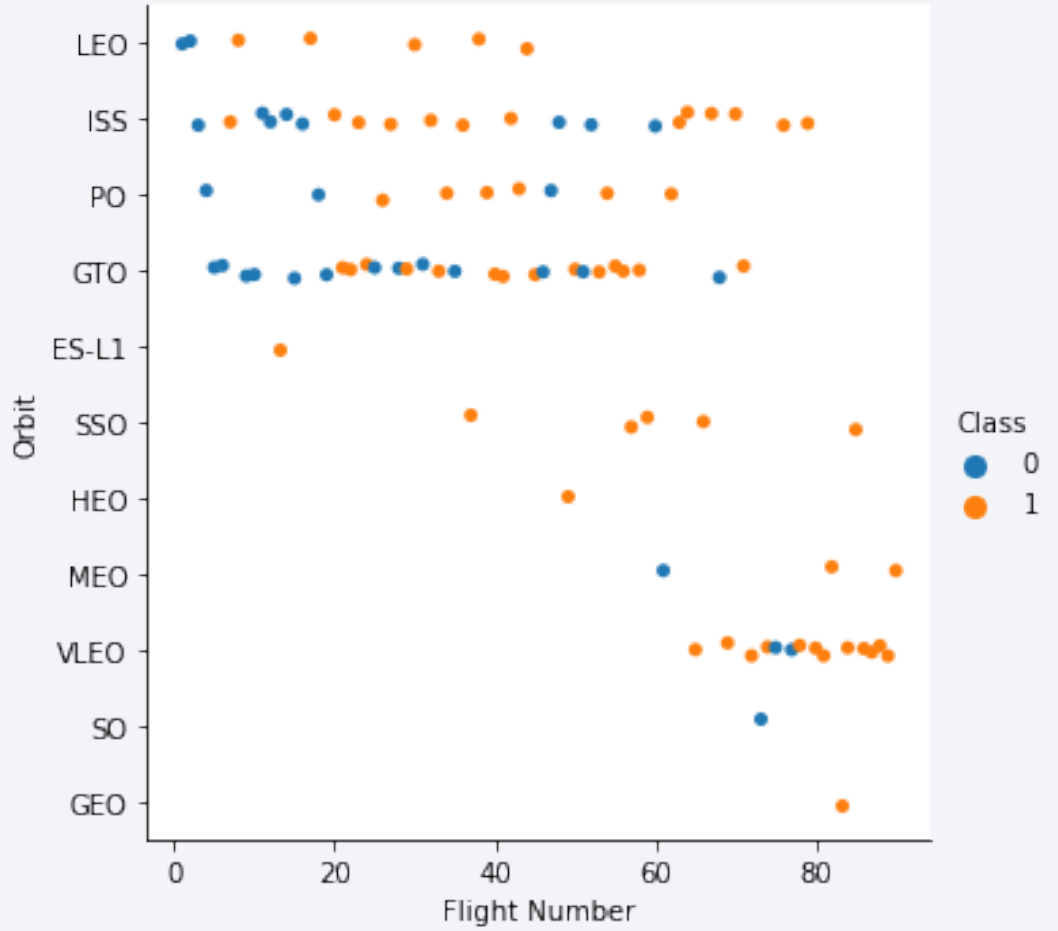
Success rate is very high at ES-L1, GEO, HEO, and SSO.

No correlation between successful outcome and distance of the orbit observed.



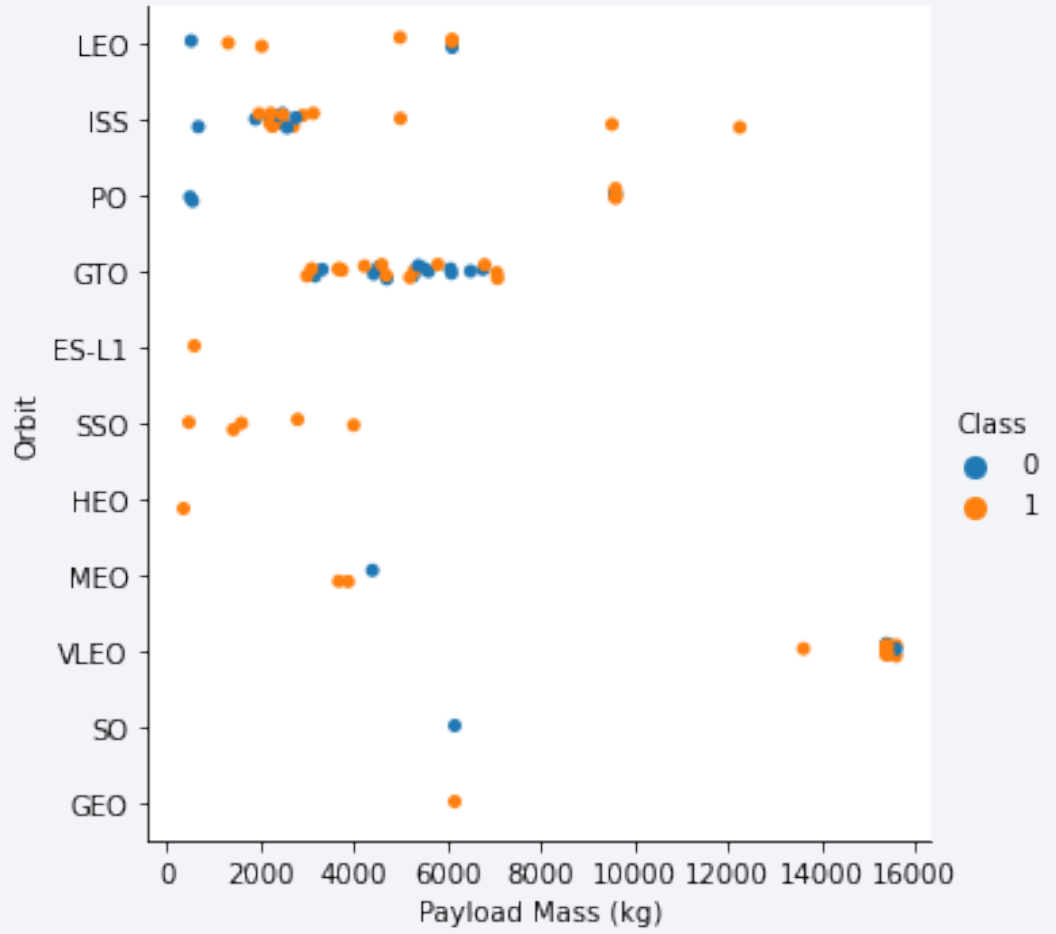
Flight Number vs. Orbit Type

Strong correlation between flight number and the success rate is found in LEO.



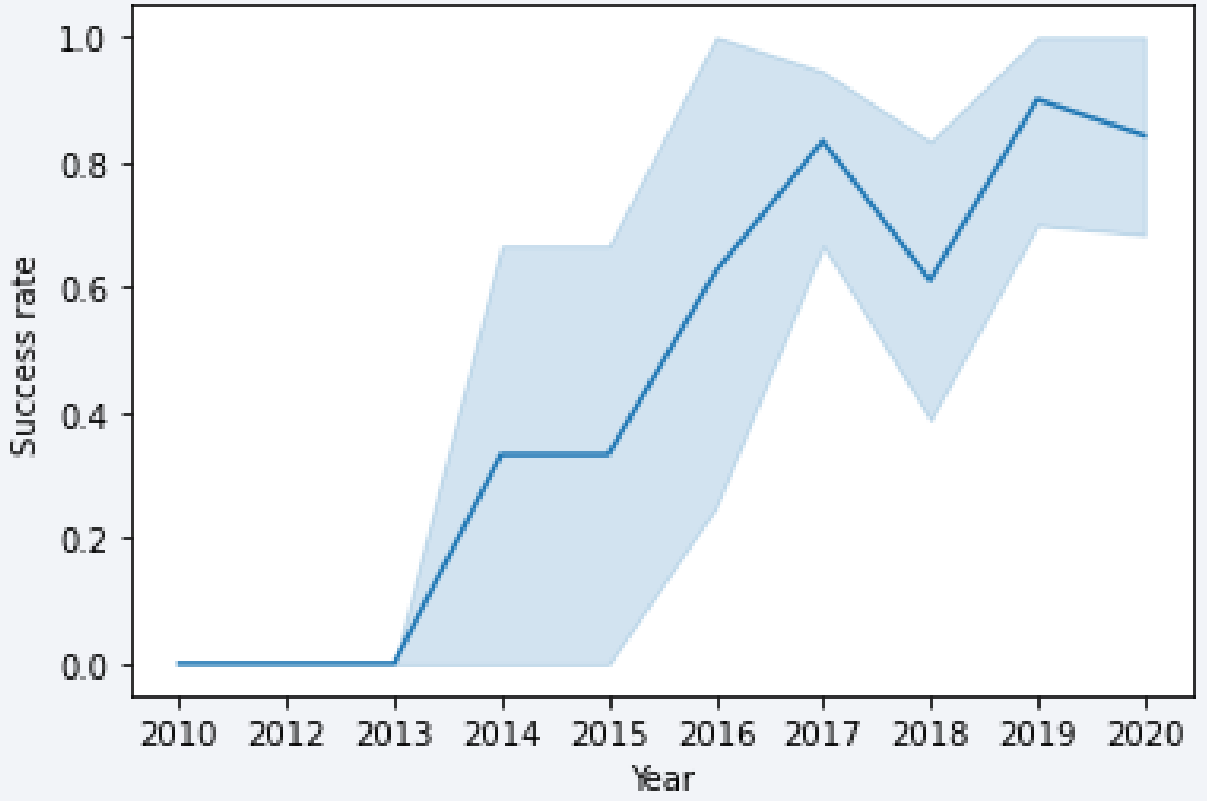
Payload vs. Orbit Type

Success rate is higher with heavier payload mass in VLEO, LEO, ISS, and PO.



Launch Success Yearly Trend

Success rate is increasing as years go by.



All Launch Site Names

The data only contains four launch sites in three location:

- Cape Canaveral Space Force Station (CCAFS LC-40 and CCAFS SLC-40),
- Kennedy Space Center (KSC LC-39A) and
- Vandenberg Air Force Base (VAFB SLC-4E)

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

Launch Site Names Begin with 'CCA'

The records shows the launches which were done in Cape Canaveral.

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 Launched by NASA (CRS)

Total payload mass carried that launched by NASA (CRS) was 45,596 kg.

1
45596

Average Payload Mass by F9 v1.1

Average payload mass carried by F9 v1.1 booster was 2,534 kg.

1
2534

First Successful Ground Landing Date

First successful landing in ground pad was achieved in December 22, 2015.

	1
	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster F9 FT B1022, FT B1026, FT B1021.2 and FT B1031.2 successfully landed on drone ship with payload varying between 4,000 and 6,000 kg.

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

Majority of the flights succeeded to do the assigned missions.

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

Boosters Carried Maximum Payload

These are the list of boosters that had carried the maximum 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

In 2015, booster F9 v1.1 B1012 and B1015 which are launched in Cape Canaveral, failed to land successfully on drone ship.

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

These are the ranking of landing outcomes between June 4, 2010 and March 20, 2017.

Ten flights had no landing attempt.

landing_outcome	number
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

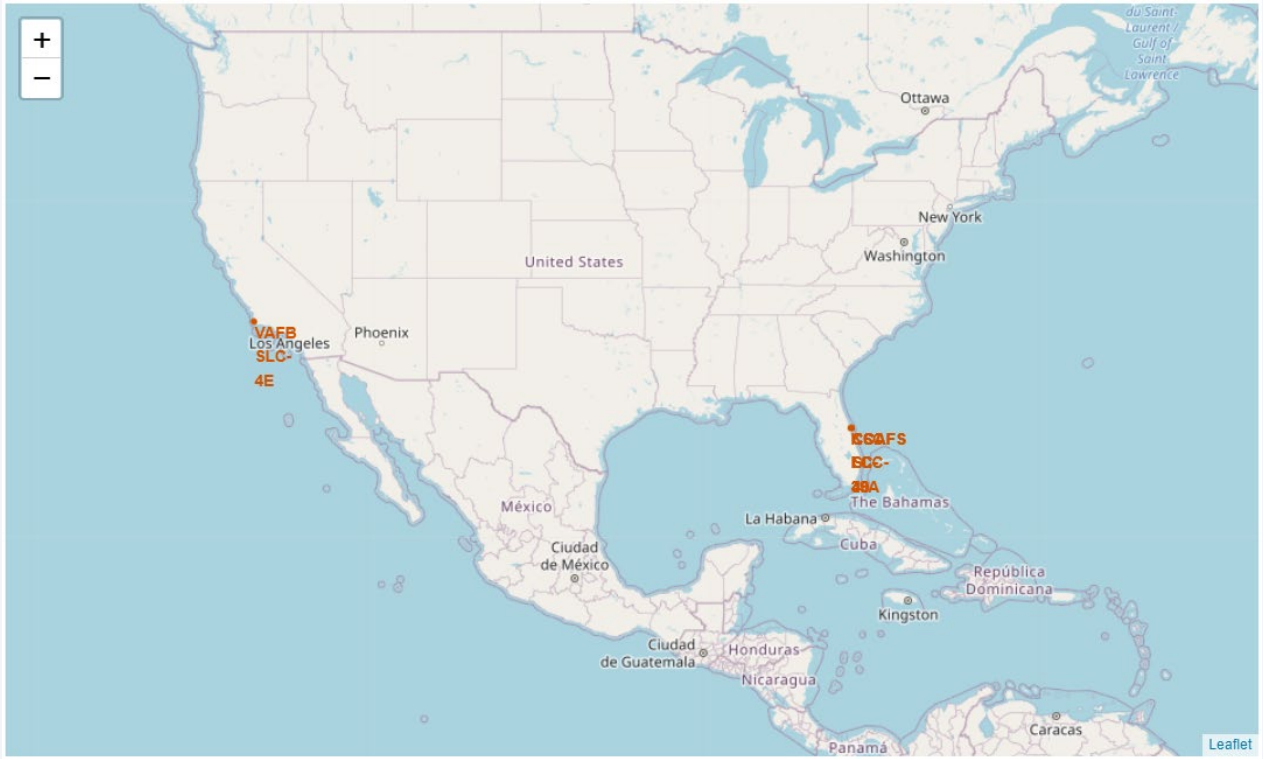
Section 3

Launch Sites Proximities Analysis

Global View

From the map, we can see that Vandenberg is located in Los Angeles, the west of the USA.

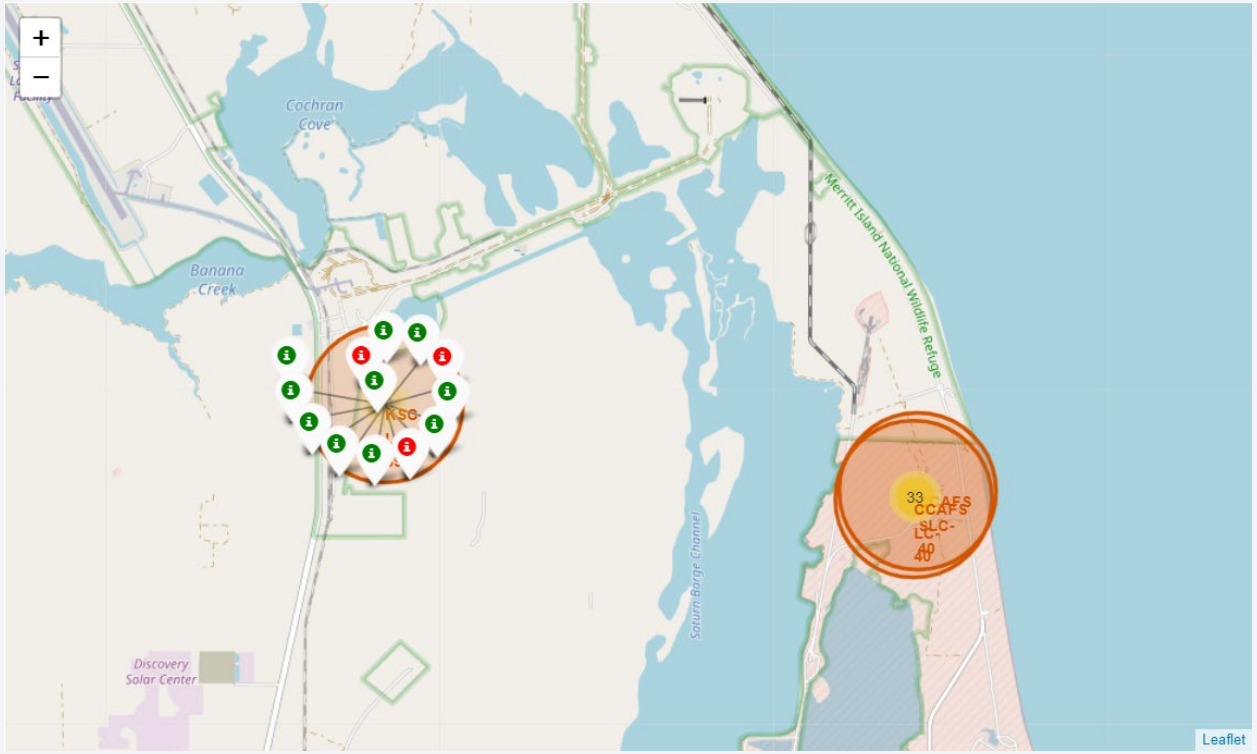
Also, the rest of the launch sites are located in Florida, the east of the USA.



Launch Outcomes

From the map, we can also see the launch outcomes for each launch site.

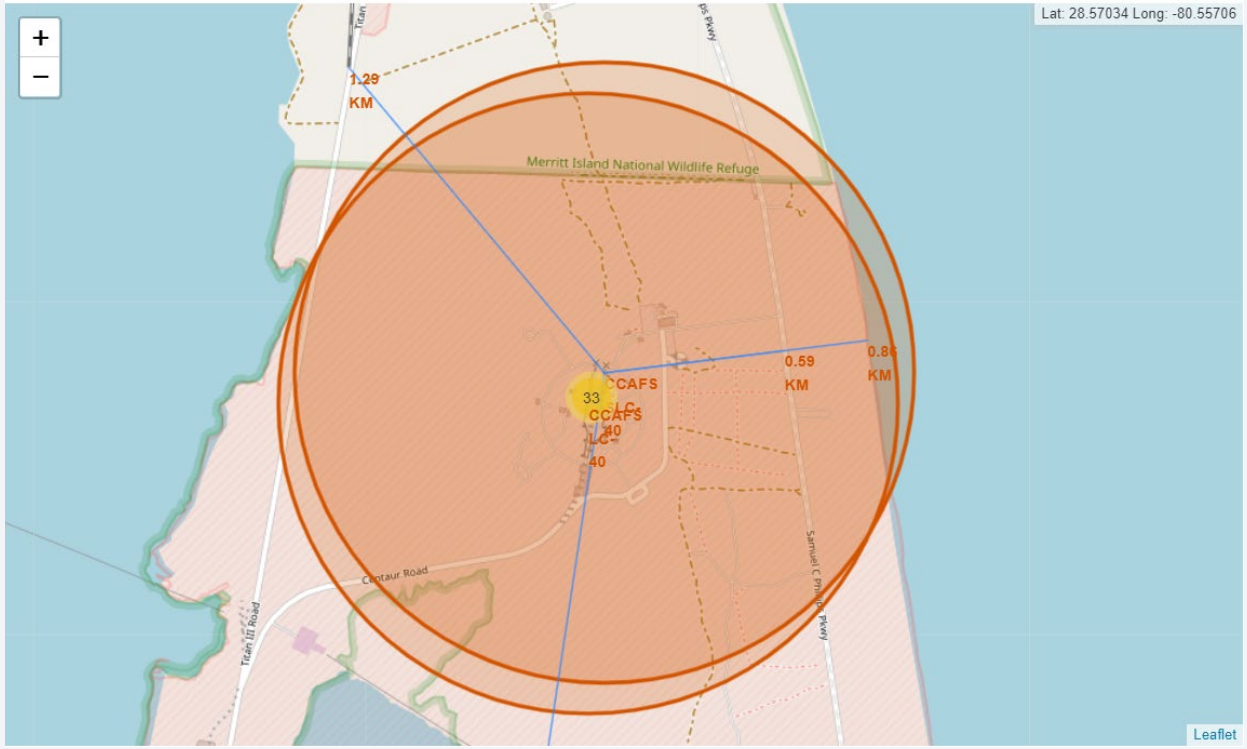
In the map, KSC is expanded and we can observe the launch outcomes visually.

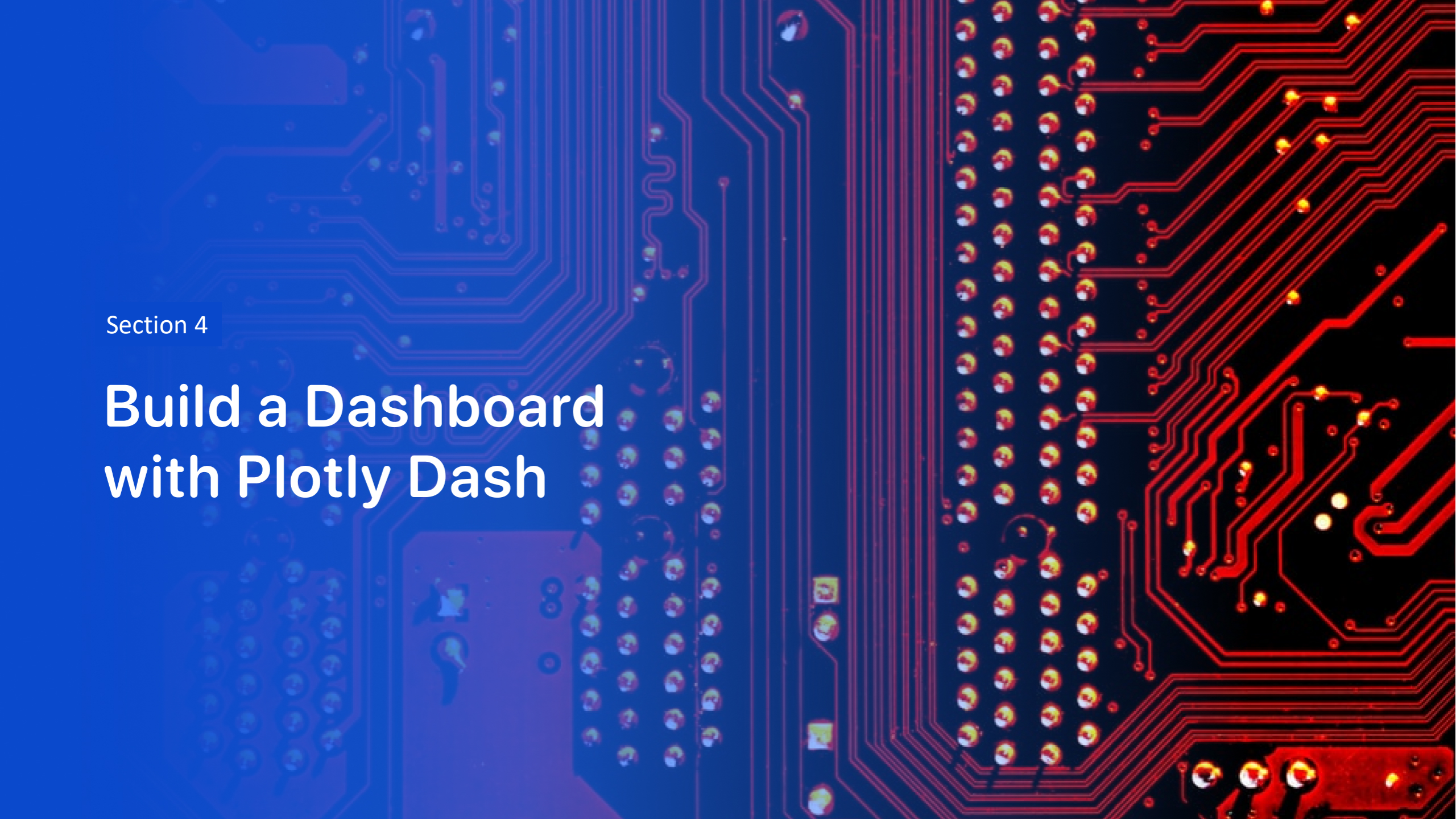


Launch Site Proximity

The CCAFS launch site is near the coastline since boosters (not SpaceX's) are typically dumped in the ocean after launch.

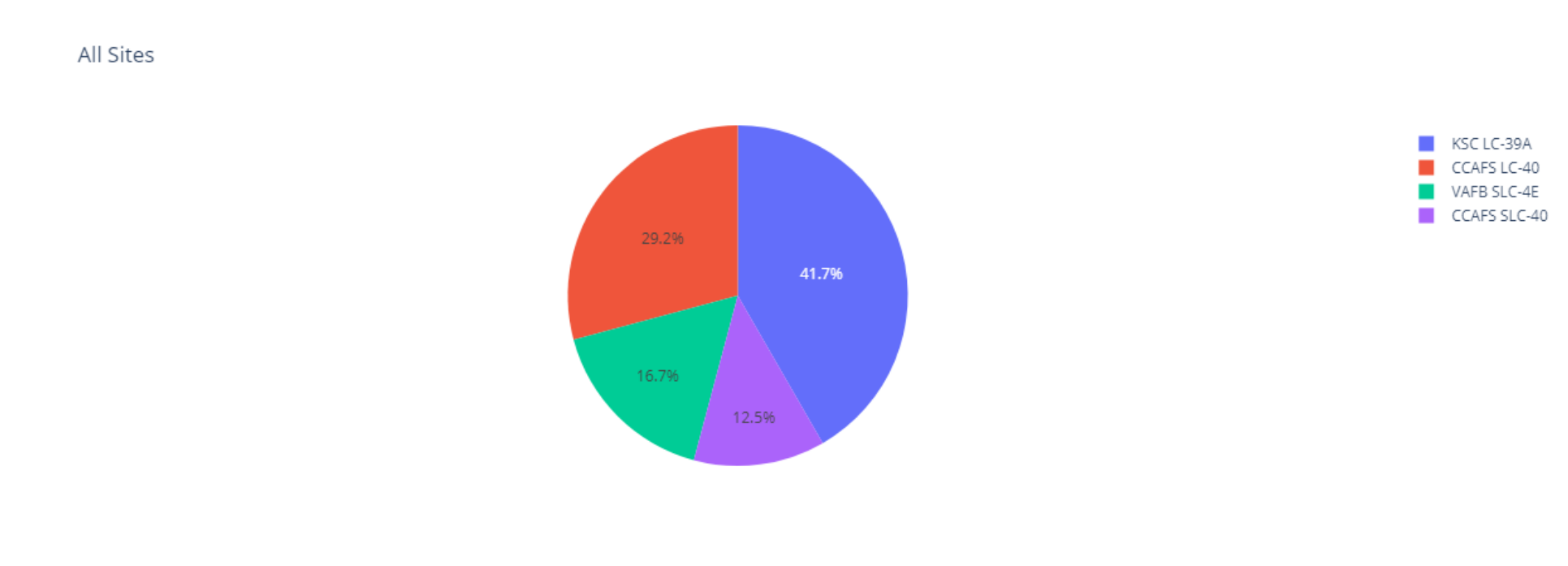
The launch site is also far from city center to avoid risk to bystanders.




Section 4
Build a Dashboard with Plotly Dash

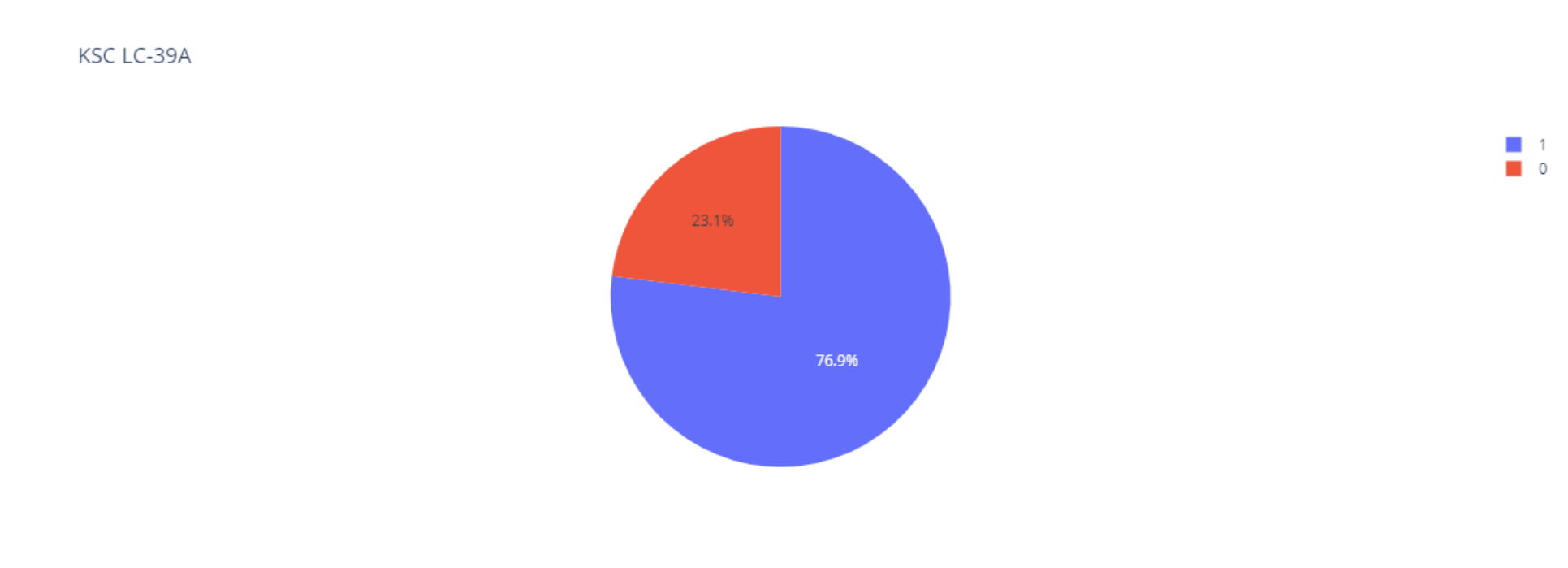
Success Rate for All Launch Sites

Launch from Kennedy Space Center has the highest success rate.



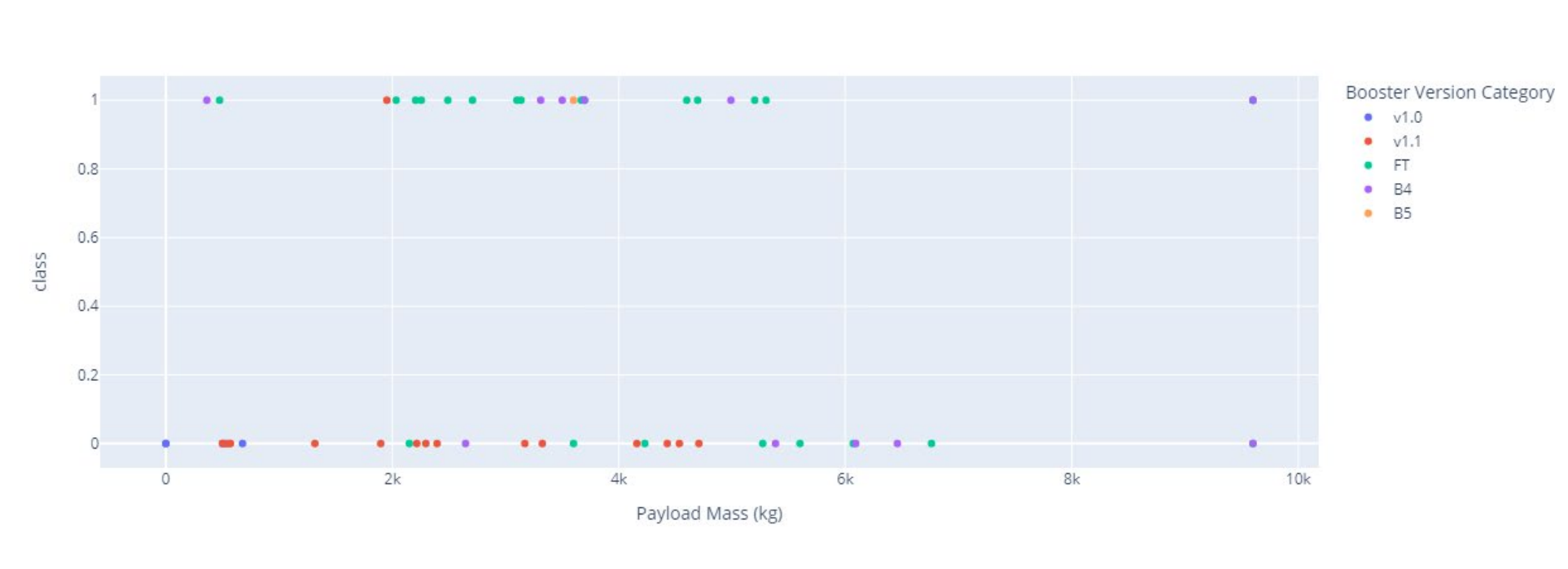
Launch Outcome in KSC

Launch from KSC yielded 10 (76.9%) successful outcome and 3 (23.1%) failed outcome.



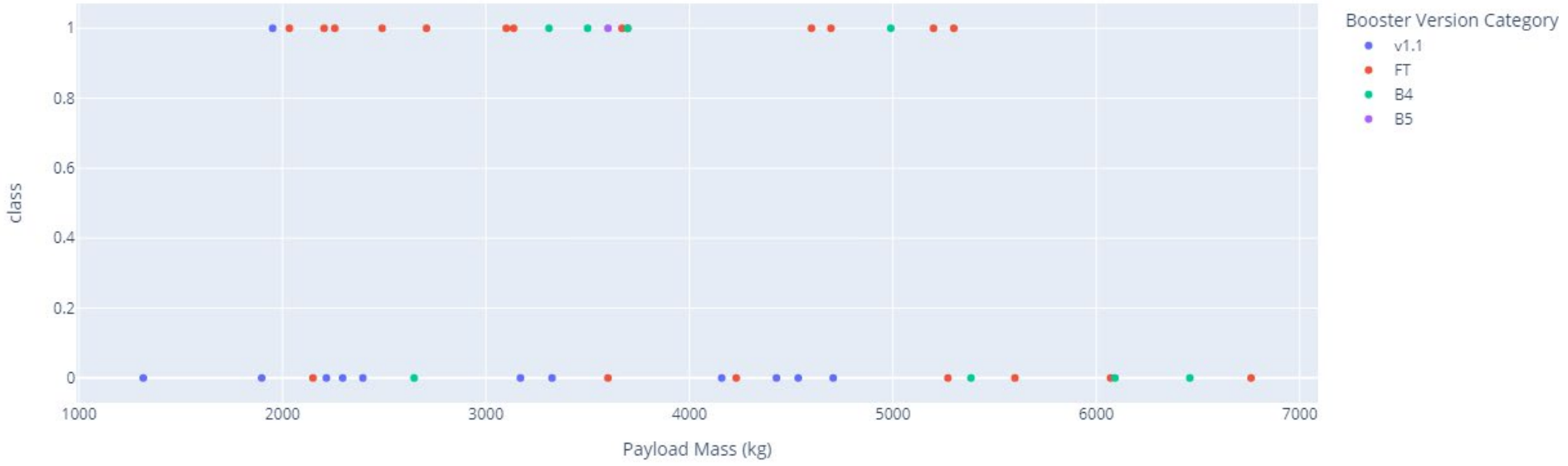
Payload vs. Launch Outcome for All Launch Sites

FT boosters have the highest success rate, followed by B4 boosters which thrive for higher payload mass.



Payload vs. Launch Outcome for All Launch Sites

Launches with v1.1 boosters only yielded one successful outcome. The next boosters improved significantly.



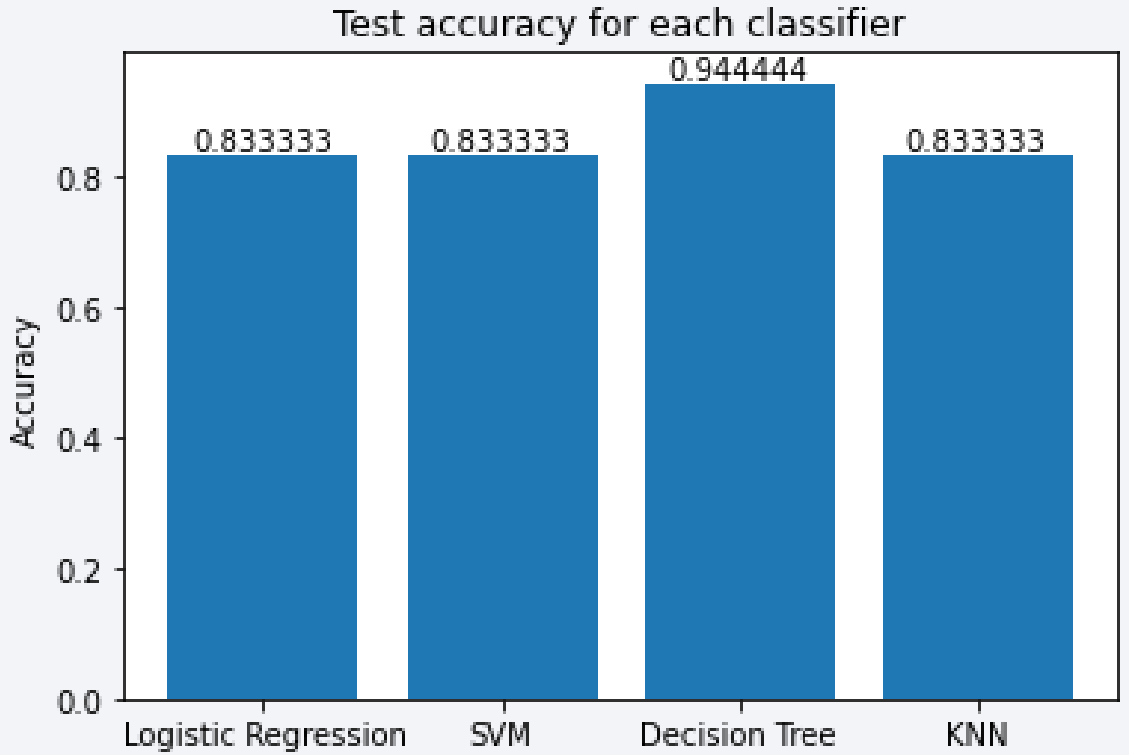
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Decision tree yields the best result with 94.44% accuracy on the test data.

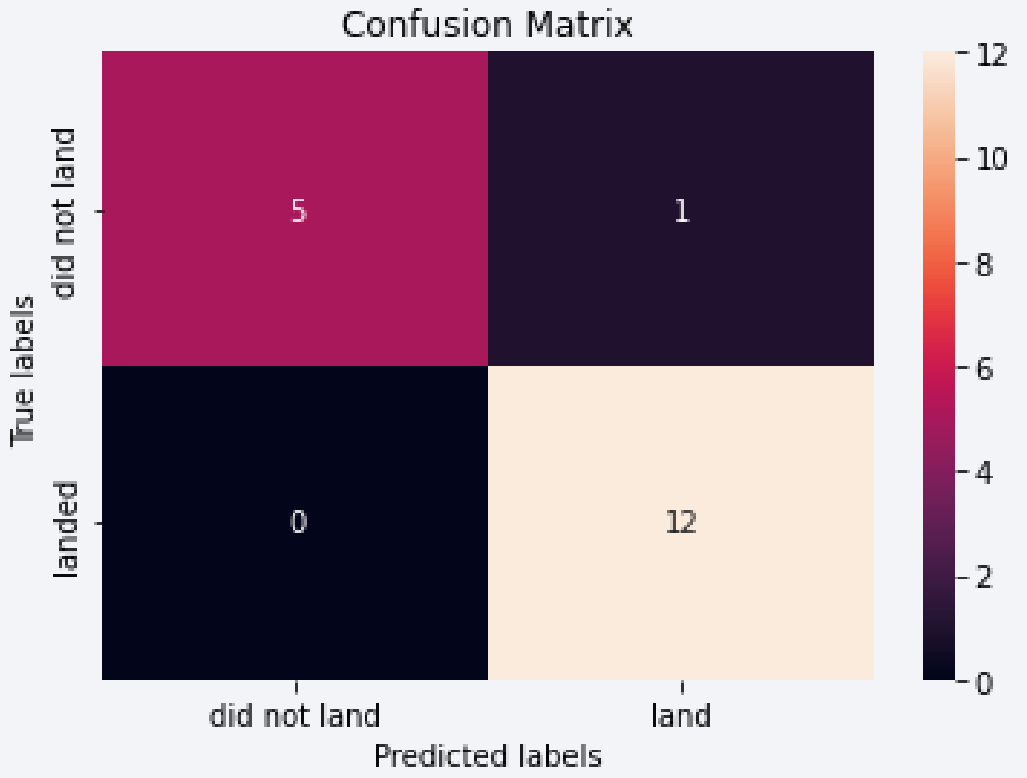
Three other models yields the same result of 83.33% accuracy.



Confusion Matrix of DT Model

As observed in the confusion matrix, there is no false negative and only one false positive out of 18 data samples.

This in turns makes the DT model very accurate.



Conclusions

- Launch site and flight number correlates strongly with the landing outcome
- Payload mass and orbit type correlates weakly with the landing outcome
- Improvement of the boosters affects the landing outcome positively
- Launch site is located strategically: near a coastline and far from city center
- Decision Tree is the best model to predict the landing outcome, compared to Logistic Regression, SVM and KNN

Appendix

- SQL queries used to yield the results in this section, in order:
  - `%sql select distinct Launch_Site from SPACEXTBL`
  - `%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5`
  - `%sql select SUM(payload_mass__kg_) from SPACEXTBL where customer like 'NASA (CRS)'`
  - `%sql select AVG(payload_mass__kg_) from SPACEXTBL where booster_version like 'F9 v1.1%'`
  - `%sql select MIN(date) from SPACEXTBL where landing_outcome like 'Success (ground pad)'`
  - `%sql select * from SPACEXTBL where landing_outcome like 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000`

Appendix

- `%sql select mission_outcome, COUNT(*) as number from SPACEXTBL group by mission_outcome`
- `%sql select distinct booster_version from SPACEXTBL where payload_mass__kg_ = (select MAX(payload_mass__kg_) from SPACEXTBL)`
- `%sql select booster_version, launch_site from SPACEXTBL where landing_outcome like 'Failure (drone ship)' and YEAR(Date) = 2015`
- `%sql select landing_outcome, COUNT(*) as number from SPACEXTBL where date between '2010-06-04' and '2017-03-20' group by landing_outcome order by number desc`

Thank you!