IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Guido Mascia
30th June 2023

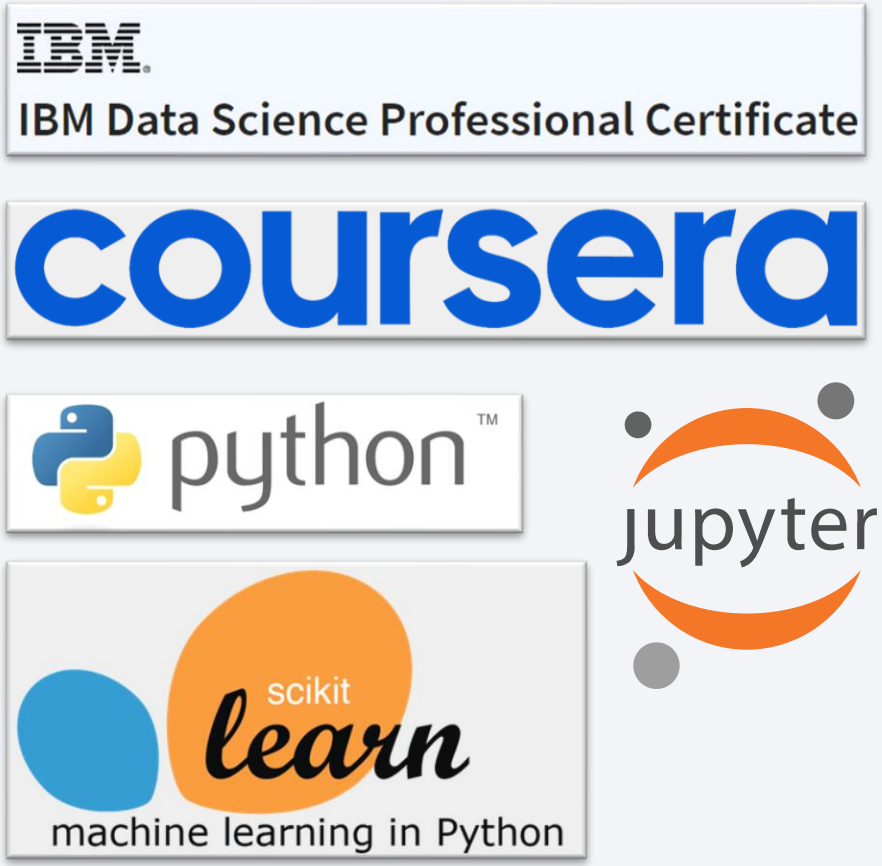
Table of Contents

Executive Summary

- **SpaceY** is a newly established rocket launch company which wants to compete against the already established SpaceX.
- To do so, SpaceY should be able to:
  - Reuse the 1st stage rocket booster.
  - Be more cost-competitive than its competitor.
- SpaceX states that their launch services – with 1st stage recovery – cost **62 million USD**, whereas **15 million USD** are required to build a 1st stage Falcon 9 booster when excluding R&D and profit margin.
- Considering the parameters in our predictive models, a Decision Tree was capable to predict the successfulness of 1st stage booster landing with an **accuracy of 89%**.
- It comes that SpaceY will be able to predict the cost of a launch exploiting the Decision Tree model as a proxy. Thus, SpaceY will be capable of making more informed bids against SpaceX for a rocket launch.

Introduction – Project Background

- This project is part of the **IBM Data Science Professional Certificate** delivered by Coursera.
- The idea of the project is to fully deploy the skills I acquired and that now I should master for taking part in a Data Science project based on real data.
- The most important aspects would be:
  - **Hard skills**:
    - Coding using Python in a Jupyter Notebook environment, using the many Data Science libraries.
    - Computational thinking, i.e., solving real world data issues by means of coding instructions.
  - **Soft skills**:
    - Understanding the patterns in the data I gathered.
    - Presenting the data in a way that stakeholders can be advised.



Introduction – Business Problem

- **SpaceY** is a newly established rocket launch company which wants to compete against the already established SpaceX.
- To do so, SpaceY should be able to:
  - Reuse the 1st stage rocket booster.
  - Be more cost-competitive than its competitor.
- SpaceX states that their launch services – with 1st stage recovery – cost **62 million USD**, whereas **15 million USD** are required to build a 1st stage Falcon 9 booster when excluding R&D and profit margin.

AIM ⇩

Use SpaceX data to build a predictive model capable of classifying successfull launches ⇨ Use the model as a proxy for predicting SpaceY Launch costs

Section 1

Methodology

Methodology

The Data Science methodology was applied to the project through:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA) – Visualization and SQL
- Interactive visual analytics – Folium and Plotly Dash
- Predictive analysis using Classification Models

Data Collection

- Two methods were used:

  1. Open-Source REST API for SpaceX for collecting historical launch data, focusing the attention on the **Falcon 9** launches

     Main Libraries: `requests`, `pandas`, `numpy`, `datetime`

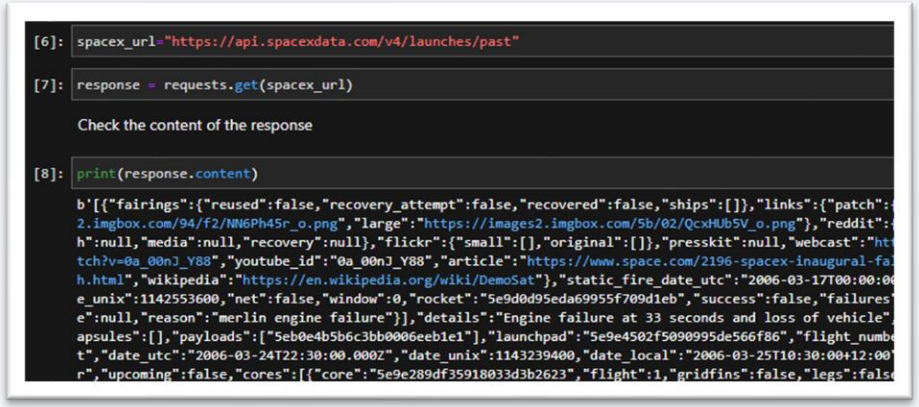
  2. Web Scraping from the Falcon 9 launch history Wikipedia page* for extracting further tabular data

     Main Libraries: `requests`, `bs4`, `re`, `unicodedata`, `pandas`, `sys`

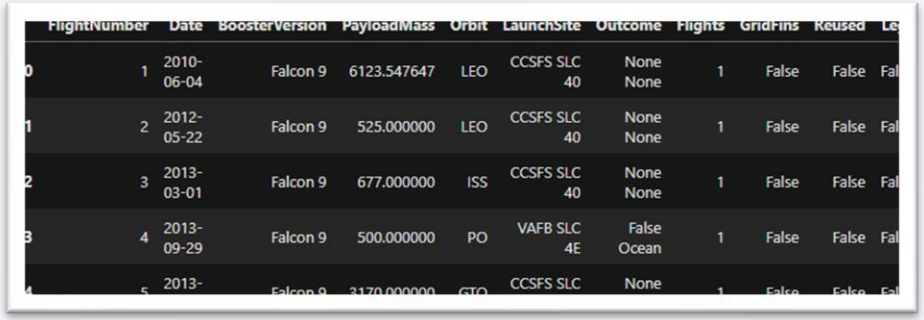
*Records limited to year 2020, as requested

Data Collection – SpaceX API

- **API** used for collecting historical launch data, focusing the attention on the **Falcon 9** launches:
  1. Request and parse the SpaceX launch data using the GET request
  2. Filter the dataframe to only include **Falcon 9** launches
  3. Replace missing Payload Mass values with the mean of the available ones



«Raw» response after GET request



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Le
0	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	Fal
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	Fal
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	Fal
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	Fal
4	5	2013-	Falcon 9	3170.000000	GTO	CCSFS SLC	None	1	False	False	Fal

Final DataFrame after **json normalization** and missing value replacement

[GitHub Jupyter Notebook link](#) for SpaceX API

Data Collection - Scraping

- **Web Scraping** from the Falcon 9 launch history Wikipedia page* for extracting further tabular data
    1. Request the Falcon9 Launch Wiki page from its URL
    2. Extract all column/variable names from the HTML table header
    3. Create a data frame by parsing the launch HTML tables

GitHub Jupyter Notebook Link for Scarping

1

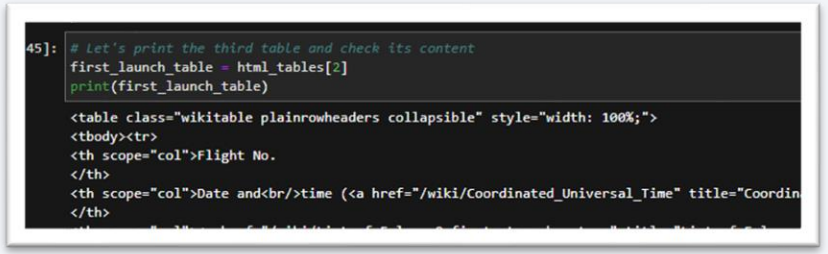
2020 [edit]

n late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[11] in addition to 14 or 15 non-Starlink aunches. At 26 launches, 14 of which were for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket amily of 2020, only behind China's Long March rocket family.[12]

[hide] Flight No.	Date and time (UTC)	Version, booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020 02:19:21[13]	F9 B5 △ B1049.4	CCSFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[14]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[15]								
	19 January 2020 15:30[16]	F9 B5 △ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[17] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[18]	NASA (CTS)[19]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi),								

Page

2

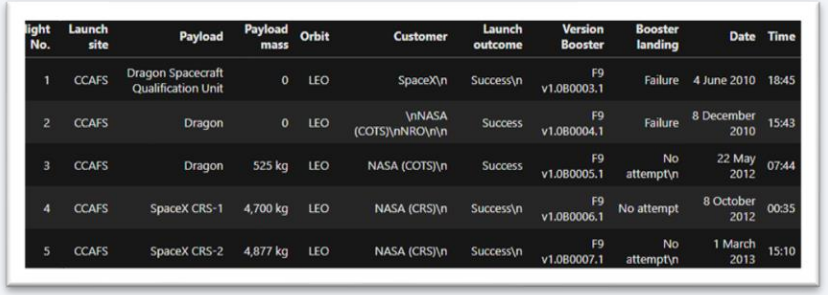


```
45]: # Let's print the third table and check its content
     first_launch_table = html_tables[2]
     print(first_launch_table)

     <table class="wikitable plainrowheaders collapsible" style="width: 100%;">
     <tbody><tr>
     <th scope="col">Flight No.
     </th>
     <th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordin
     </th>
```

HTML

3



ight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX\n	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
2	CCAFS	Dragon	0	LEO	\nNASA (COTS)\nNRO\n\n	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)\n	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)\n	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)\n	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

DataFrame

*Records limited to year 2020, as requested

Data Wrangling

- **Data Wrangling** used for creating appropriate labels referred to mission success or failure:
  1. Calculate the number of launches on each site
  2. Calculate the number and occurrence of each orbit
  3. Calculate the number and occurrence of mission outcome of the orbits
  4. Create a landing outcome label from Outcome column
     - Success ⇔ 1
     - Failure ⇔ 0

Success or Failure	Landing Location	Occurrence	Class
TRUE	ASDS	41	1
None	None	19	0
TRUE	RTLS	14	1
FALSE	ASDS	6	0
TRUE	Ocean	5	1
FALSE	Ocean	2	0
None	ASDS	2	0
FALSE	RTLS	1	0

Table 1. The possible outcome combination along with their occurrence in the dataset and the corresponding label.

GitHub Jupyter Notebook Link for Data Wrangling

EDA with Data Visualization

- To find visual insights about the relationships between the variables in the dataset and the launch success / failure
  - Main libraries: `pandas`, `matplotlib`, `seaborn`
  - Main Plot resources: Bar chart, Scatter Point Chart, Line Chart
- Relationship between:
  1. Flight Number and Payload Mass*
  2. Flight Number and Launch Site*
  3. Payload and Launch Site*
  4. Success Rate for each Orbit Type
  5. Flight Number and Orbit Type*
  6. Payload and Orbit Type*
- Trend:
  6. Success Rate yearly trend

* ⇔ Grouped by **`class`** in seaborn

GitHub Jupyter Notebook Link for EDA with Data Visualization

EDA with SQL

- Used to acquire further insights about the faced dataset
  - Main Libraries: `sqlite3`, **`pandas`**, **`csv`**
- The queries were used for acquiring information about:
  - Launch Sites
  - Payload Mass
  - Successful and Unsuccessful launches
  - Mission Outcomes in specific date ranges
  - Booster versions

Create the db query

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

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

Analyze the outcome

[GitHub Jupyter Notebook Link](#) for EDA with SQL

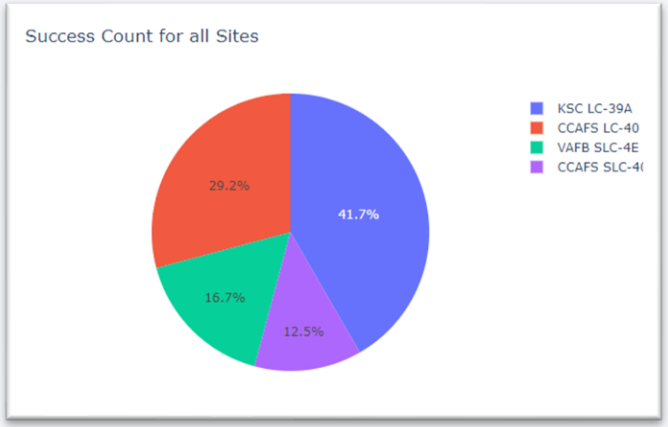
Build an Interactive Map with Folium

- Creation of a map of Launch Sites and some key info.
  - Main Libraries: `folium`, **`pandas`**
- **Why** ➔ The objects marked on the Folium map were added to navigate among Launch Sites easily and to understand common patterns related to their location
  - All the Launch Sites were marked on the map
  - Both successful and failed launches were indicated for each Launch Site
  - The distance between the **CCAFS SLC-40** Launch Site and the closest coastline, highway, and city was computed

GitHub Jupyter Notebook Link to Map with Folium

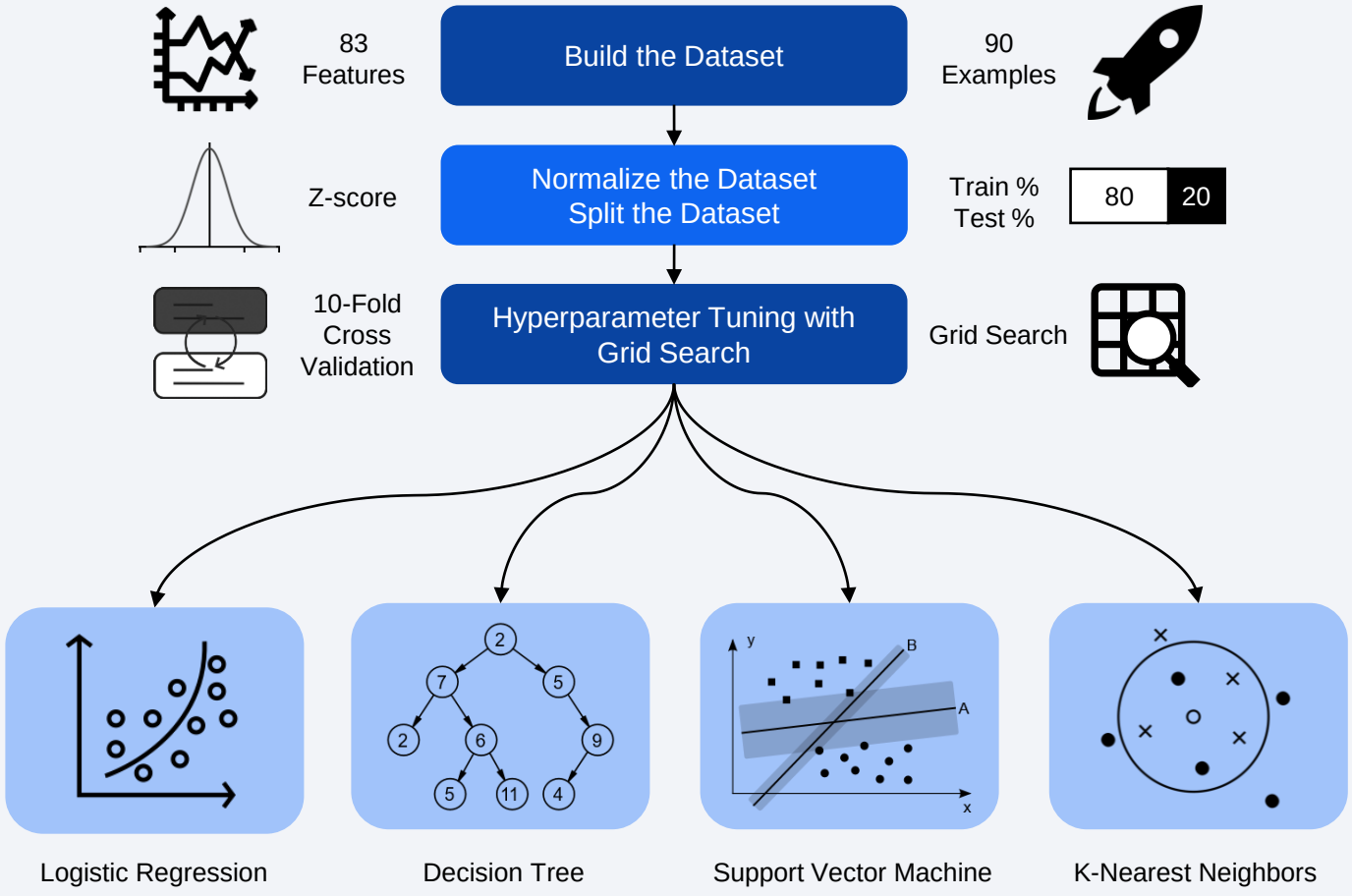
Build a Dashboard with Plotly Dash

- An interactive dashboard for data exploration was built.
  - Main libraries: `dash`, `plotly`, **`pandas`**
- **Why** ➔ for user friendly data exploration
- The dashboard includes:
  - A drop-down menu allowing to show either individual or all Launch Sites
  - A Pie Chart showing the Launch Success rate according to the option selected in the drop-down menu
  - A Scatter Point Chart allowing to visualize the relationship between payload and landing outcome
  - A slider that allows to modify the range within the latter chart is showed.



GitHub Link to Dashboard

Predictive Analysis (Classification)



- **Classification Problem** to predict whether a launch is successful or not.
  - Main Libraries: `scikit-learn`, `pandas`, `numpy`, `seaborn`
- Target labels:
  - 0 (failure)
  - 1 (success)
- Accuracy evaluated on the Test Set as the % of correctly predicted outcomes.

GitHub Jupyter Notebook Link for Predictive Analysis

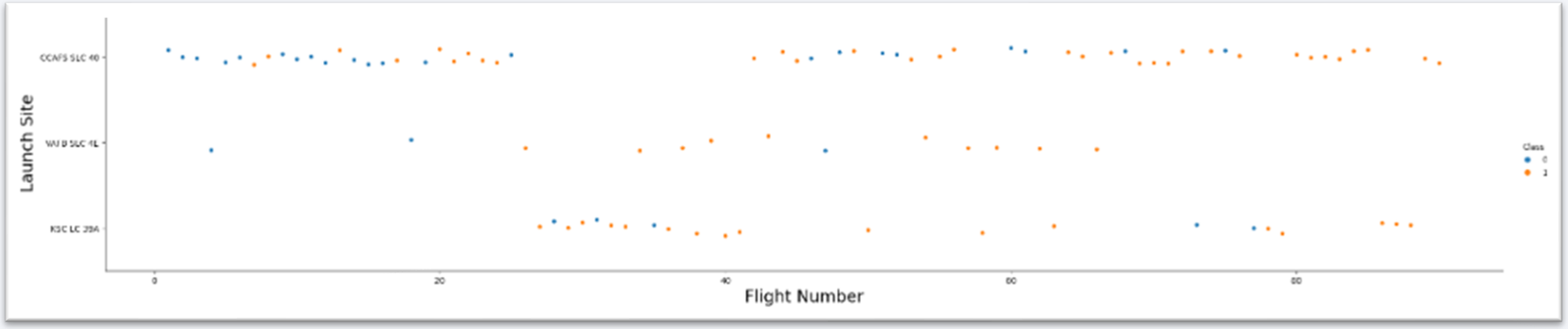
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

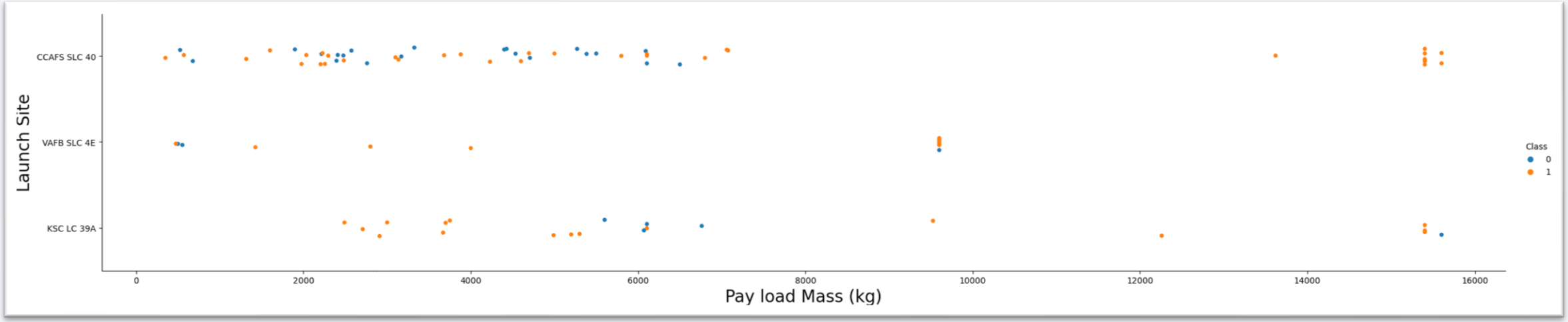
Insights drawn from EDA

Flight Number vs. Launch Site



- It is possible to see that the Launch Site CCAFS SLC-40 is the one with the most failed Stage1 outcomes.
- **However**, its Success Rate improved as the Flight Number increased.

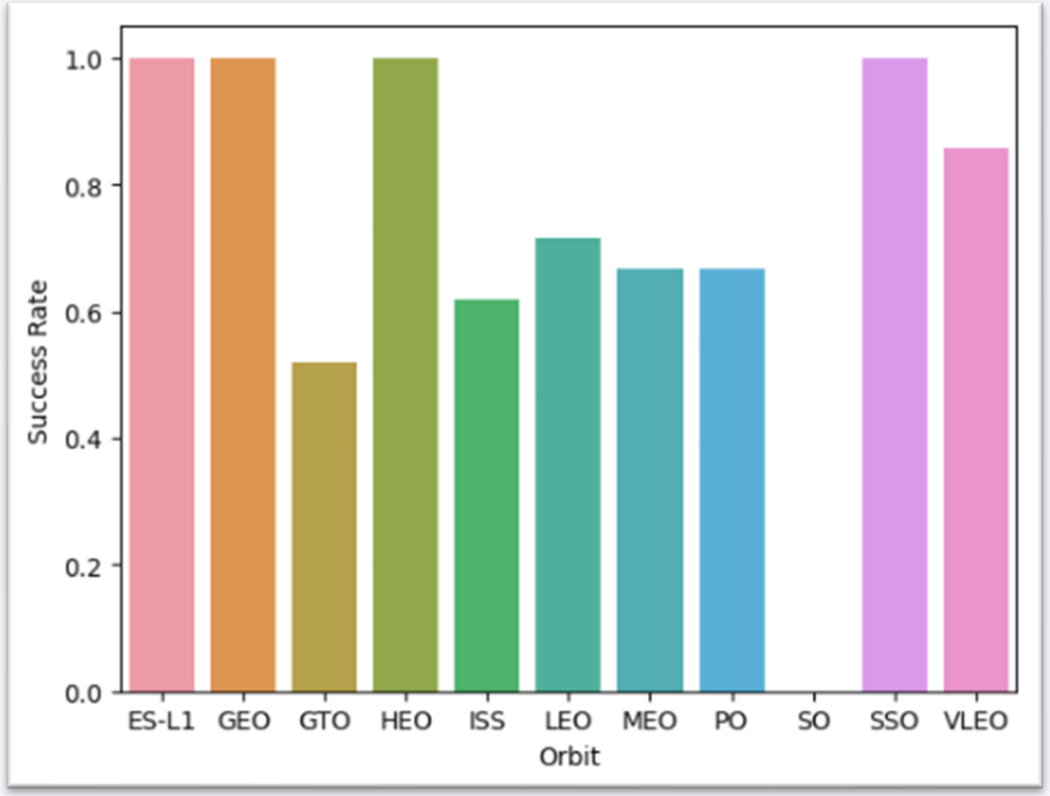
Payload vs. Launch Site



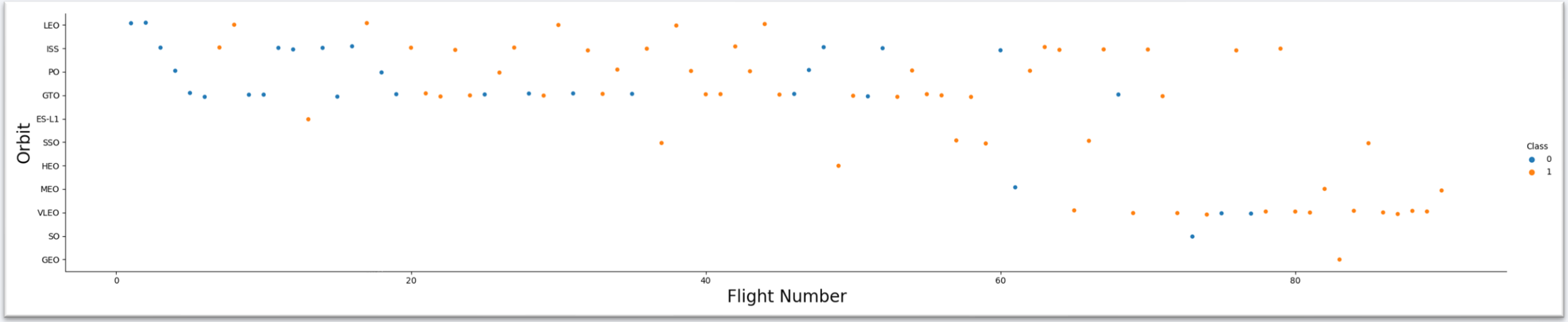
- Most of the launches for all the three Launch Site occurred with a Payload mass smaller than 8000 kg.
- From this plot it seems that there is no correlation between CCAFS SLC-40 Success and Payload.

Success Rate vs. Orbit Type

- There are some orbits with no failure ➔ **ES-L1**, **GEO**, **HEO**, **SSO**
- There is one orbits with a good Success Rate ➔ **VLEO**
- The other orbits show a poor Success Rate, except for **SO**, which exhibit null Success Rate.

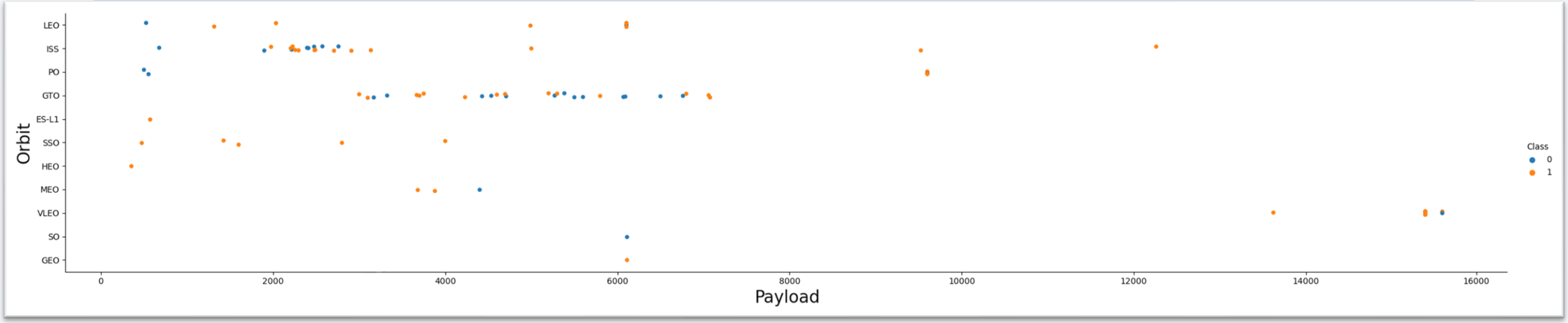


Flight Number vs. Orbit Type



- From this plot is possible to appreciate that the four Orbits that exhibit perfect Success Rate are the ones with fewer attempts.
- The same is true for the Orbit **SO**, with only one launch attempt ending into failure.
- Hence, this type of chart gives a ***weighted*** view of the Success Rate for each Orbit, although qualitatively.
- In the **LEO** Orbit, the Success is linked to the flight number.
- No correlation appears for the **GEO** orbit, instead.

Payload vs. Orbit Type

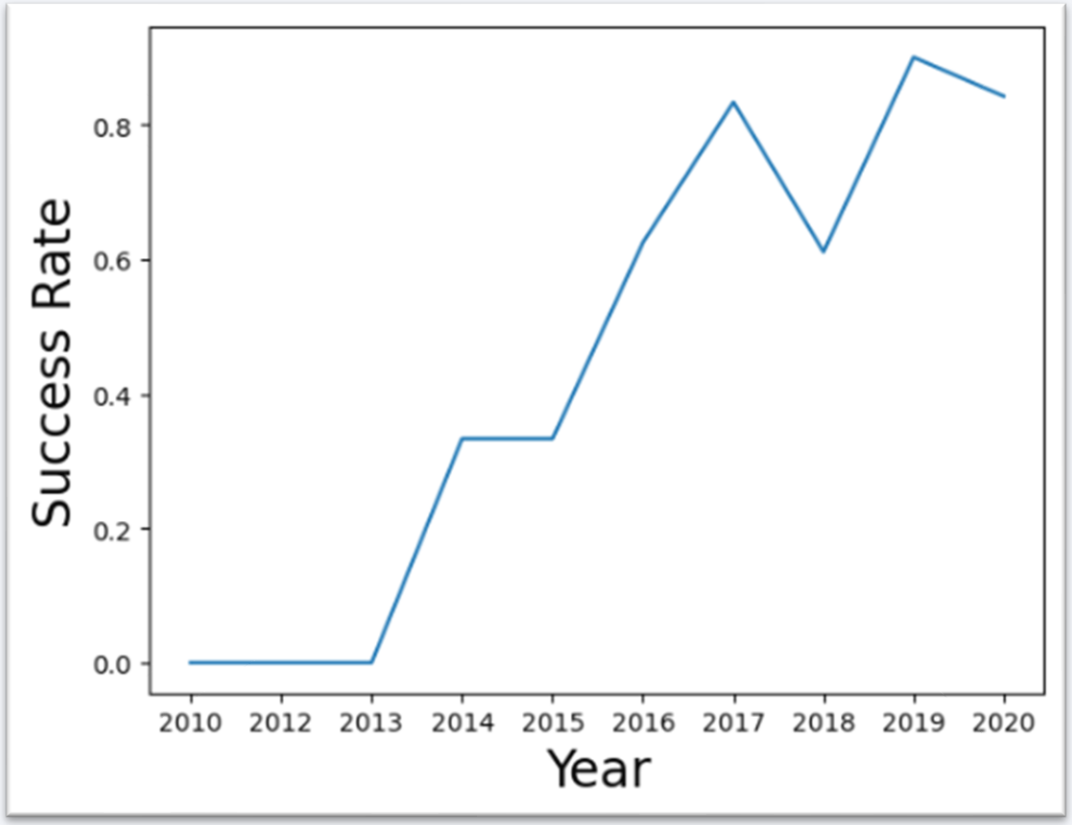


- Polar, LEO and ISS show high Success Rate with heavy Payloads.
- For GTO this cannot be distinguished, as Success and Failure classes overlap for mid-high Payloads.

Launch Success Yearly Trend

- The Success Rate continuously increases until 2017.
- It has a slight decrease in 2018.
- Increases again to its peak in 2019.

- Despite some minor oscillations in the past few years, there is a general Success Rate improvement as the year increases.



All Launch Site Names

- Cape Canaveral (Florida)
  - CCAFS LC-40
  - CCAFS SLC-40
- Vanderberg (California)
  - VAFB SLC-4E
- Merrit Island (Florida)
  - KSC LC-39A

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

- From these info, we can infer that the Launch Sites are located:
  - On the coast or close to the ocean
  - In a location where the weather is both dry and stable.

Launch Site Names Begin with 'CCA'

- It was necessary to use a LIKE in the SQL query as the Cape Canaveral Launch Sites had different codes.
- Indeed, the Launch Sites were the same, but the name changed in 2017 (source).

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcom
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachut
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachut
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attemp
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attemp
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attemp

Some Descriptive Data

Total Number of Successful And Failure Mission Outcomes

Mission_Outcome	Number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Almost all Mission Outcomes were successful, except for one

2015 Launch Records

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

All the failures regarded autonomous drone ships

First Successful Ground Landing Date

MAX(DATE)
22/12/2015

Before the first Successful Ground Landing outcome it was necessary to wait about 5 and a half years from the first Falcon 9 launch ever performed (04/06/2010)

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Successful Drone Ship Landing with Payload between 4000 and 6000

- All the Successful Drone Ship Landing with a Payload between 4000 and 6000 kg.

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535.0
F9 v1.1 B1011	4428.0
F9 v1.1 B1014	4159.0
F9 v1.1 B1016	4707.0
F9 FT B1020	5271.0
F9 FT B1022	4696.0
F9 FT B1026	4600.0
F9 FT B1030	5600.0
F9 FT B1021.2	5300.0
F9 FT B1032.1	5300.0
F9 B4 B1040.1	4990.0
F9 FT B1031.2	5200.0
F9 FT B1032.2	4230.0
F9 B4 B1040.2	5384.0
F9 B5 B1046.2	5800.0
F9 B5 B1047.2	5300.0
F9 B5 B1046.3	4000.0
F9 B5 B1048.3	4850.0
F9 B5 B1051.2	4200.0
F9 B5B1060.1	4311.0
F9 B5 B1058.2	5500.0
F9 B5B1062.1	4311.0

Boosters Carried Maximum Payload

- The list of the Boosters which carried the maximum Payload Mass contains only Falcon **F9 B5 B10XX.Y** versions.
- If considering the Launch Date too, it is possible to infer that launches with maximum payload were executed rather recently, i.e., starting from the late 2019 only.

Booster_Version	PAYLOAD_MASS__KG_	Date
F9 B5 B1048.4	15600.0	11/11/2019
F9 B5 B1049.4	15600.0	01/07/2020
F9 B5 B1051.3	15600.0	29/01/2020
F9 B5 B1056.4	15600.0	17/02/2020
F9 B5 B1048.5	15600.0	18/03/2020
F9 B5 B1051.4	15600.0	22/04/2020
F9 B5 B1049.5	15600.0	06/04/2020
F9 B5 B1060.2	15600.0	09/03/2020
F9 B5 B1058.3	15600.0	10/06/2020
F9 B5 B1051.6	15600.0	18/10/2020
F9 B5 B1060.3	15600.0	24/10/2020
F9 B5 B1049.7	15600.0	25/11/2020

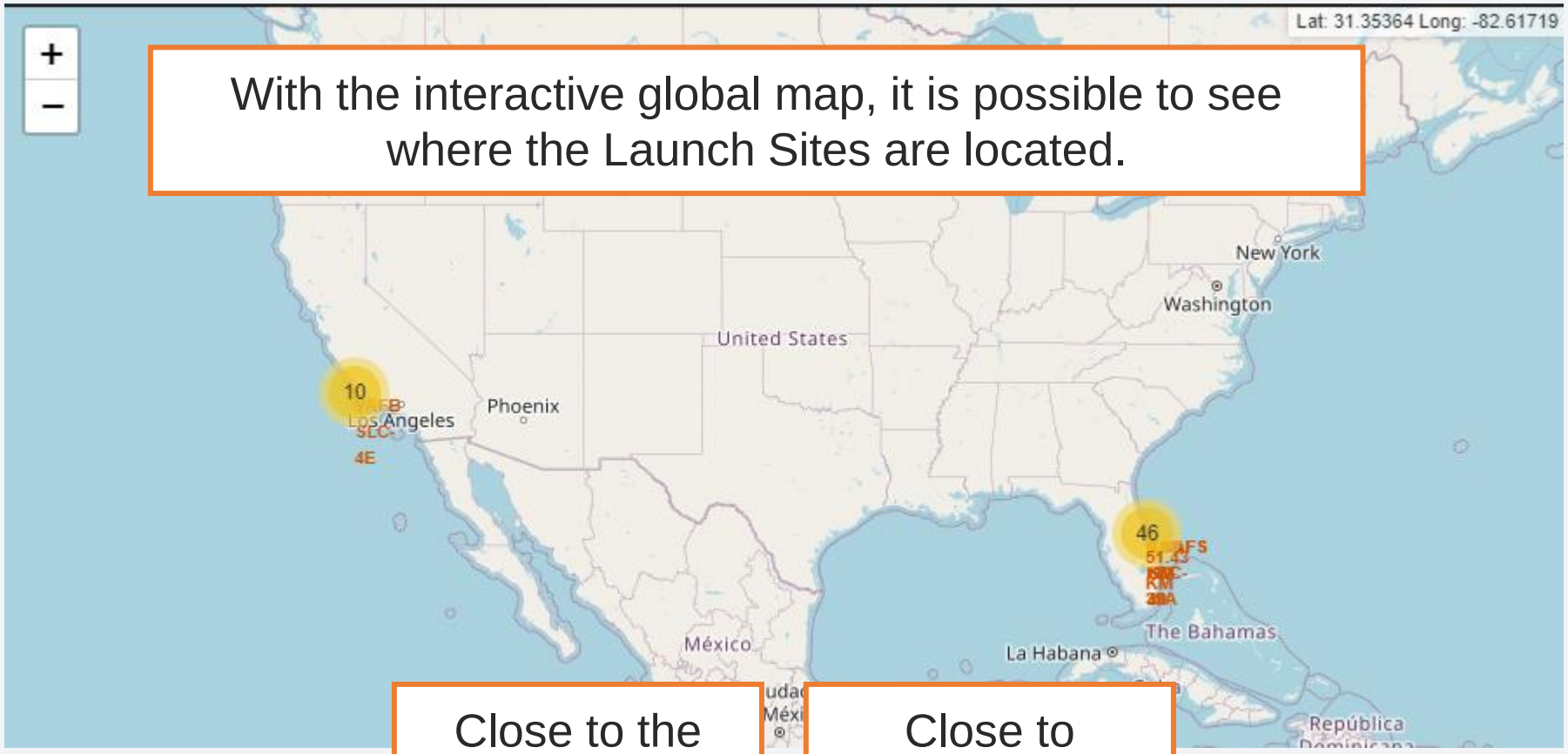
Landing Outcomes Between 2010-06-04 and 2017-03-20

Landing_Outcome	Number
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

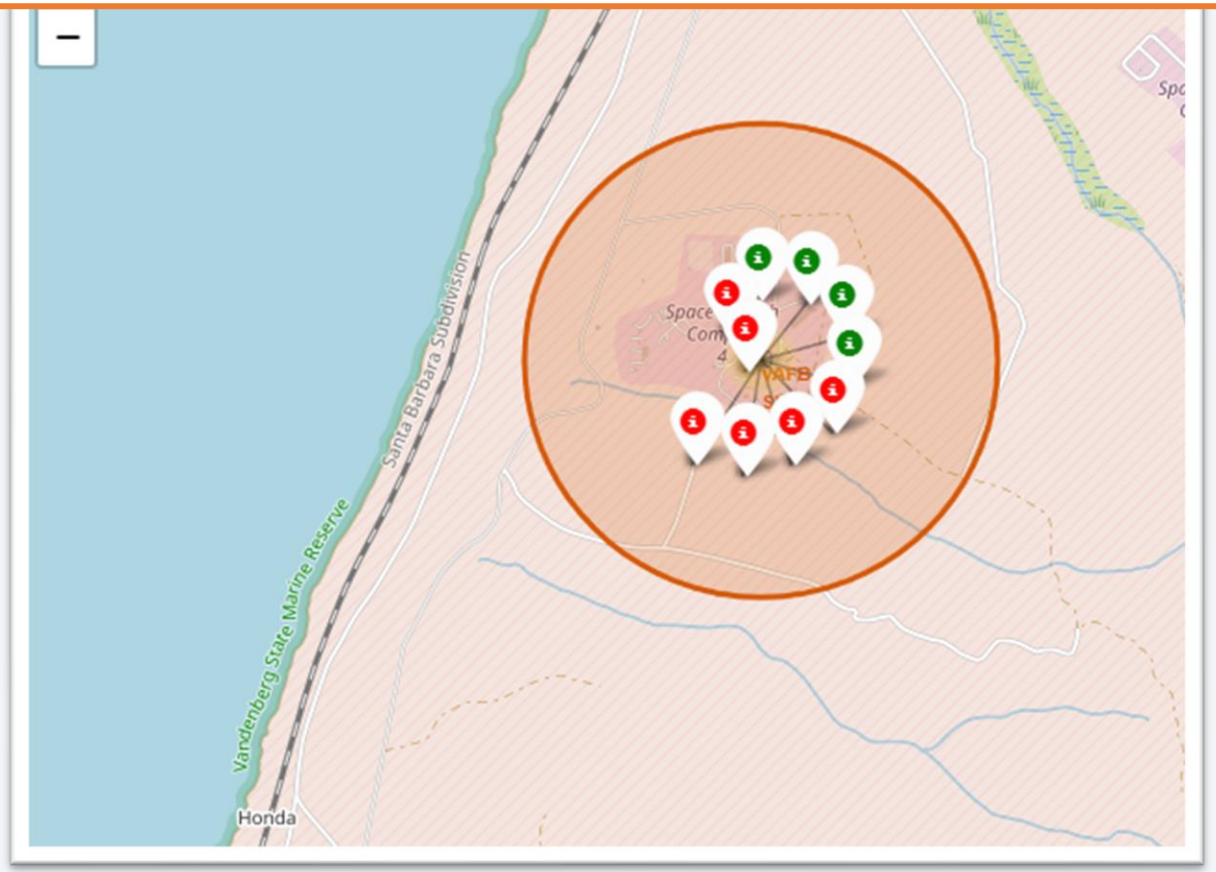
Launch Sites Proximities Analysis

Global “Marked” Map of Launch Sites

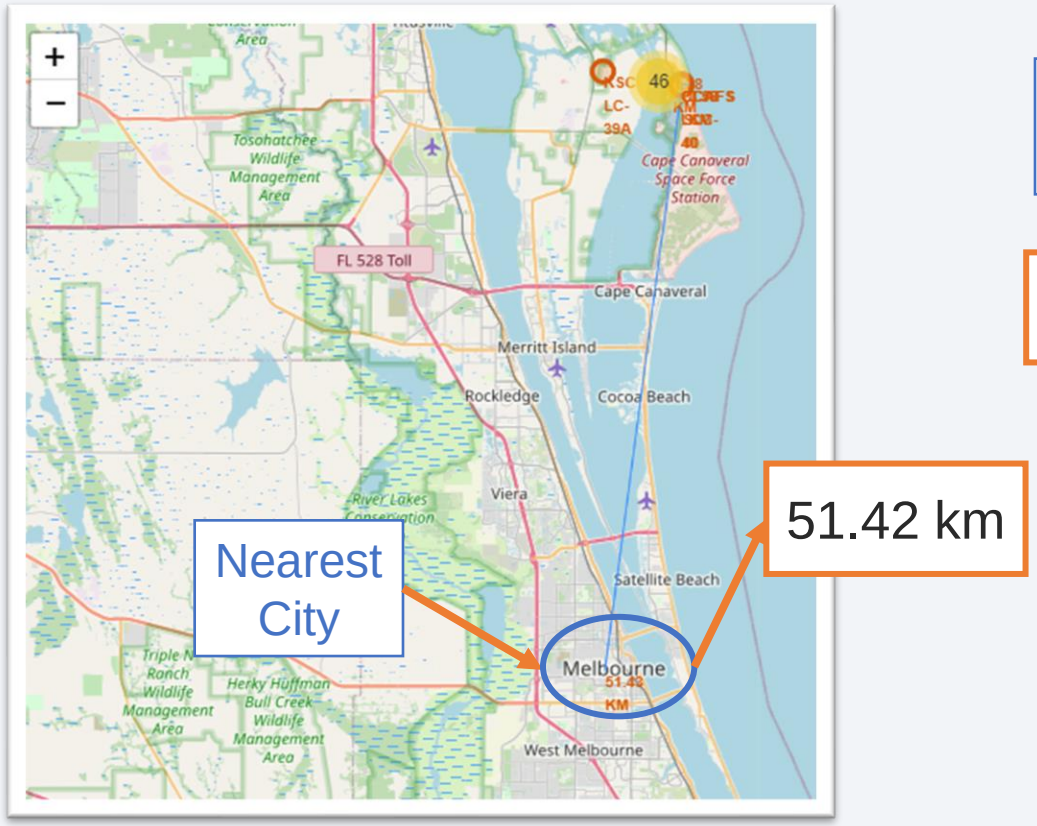


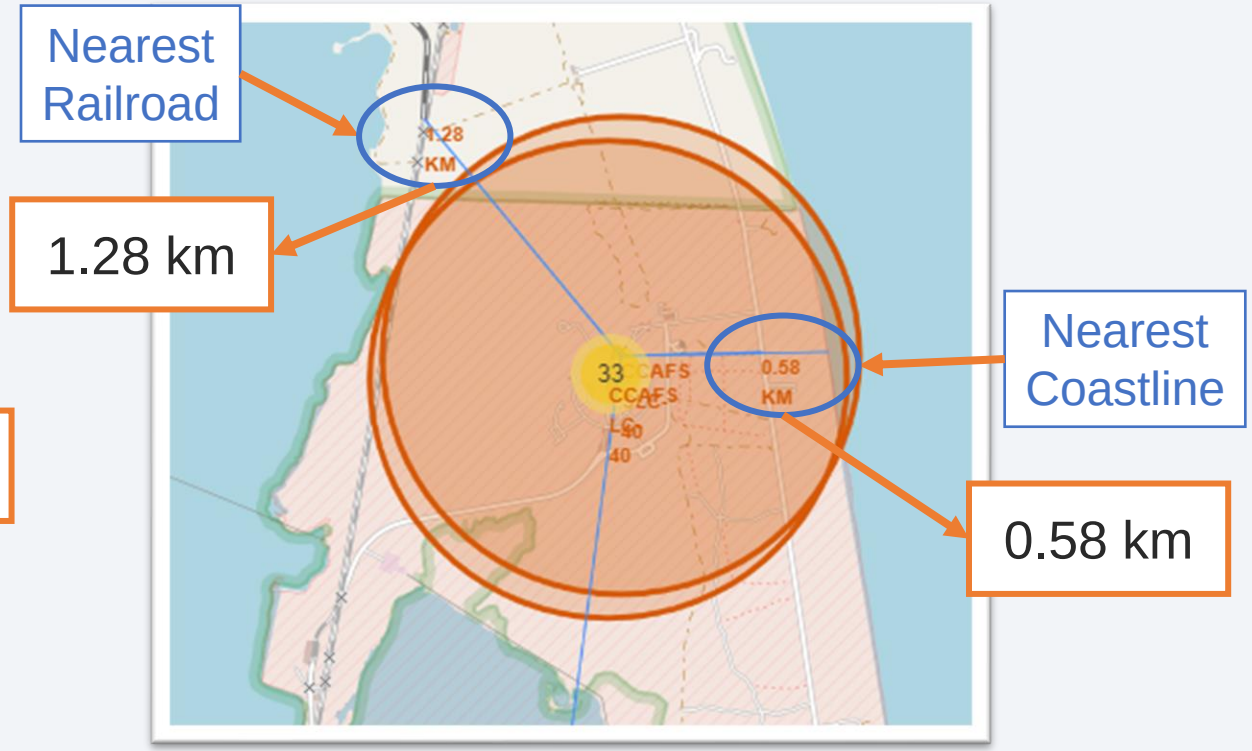
Booster Landing Outcome Visualization

Zooming and Clicking onto a Launch Site it is possible to visually inspect the launch successes and failures

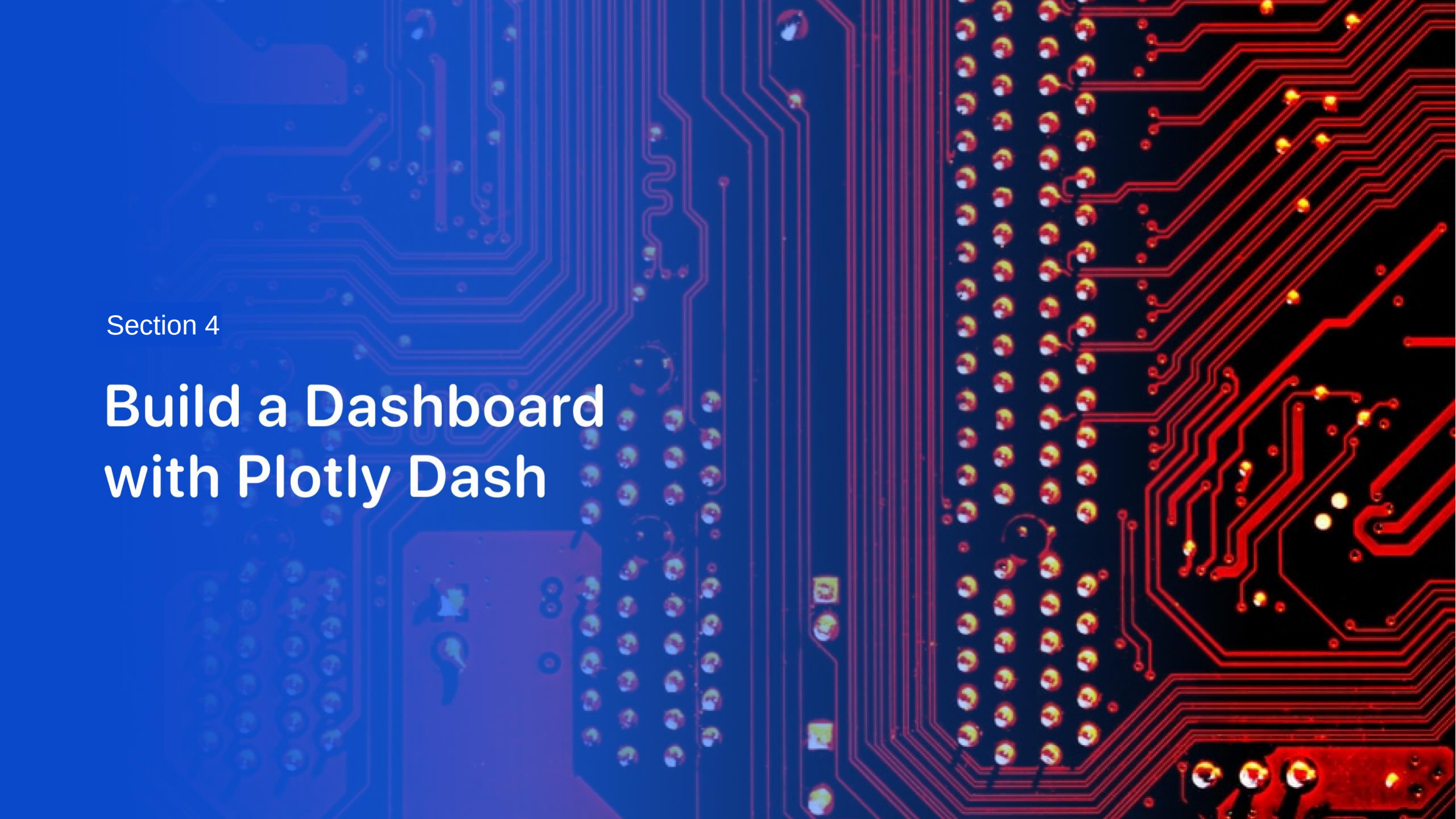


Distance from Relevant Places

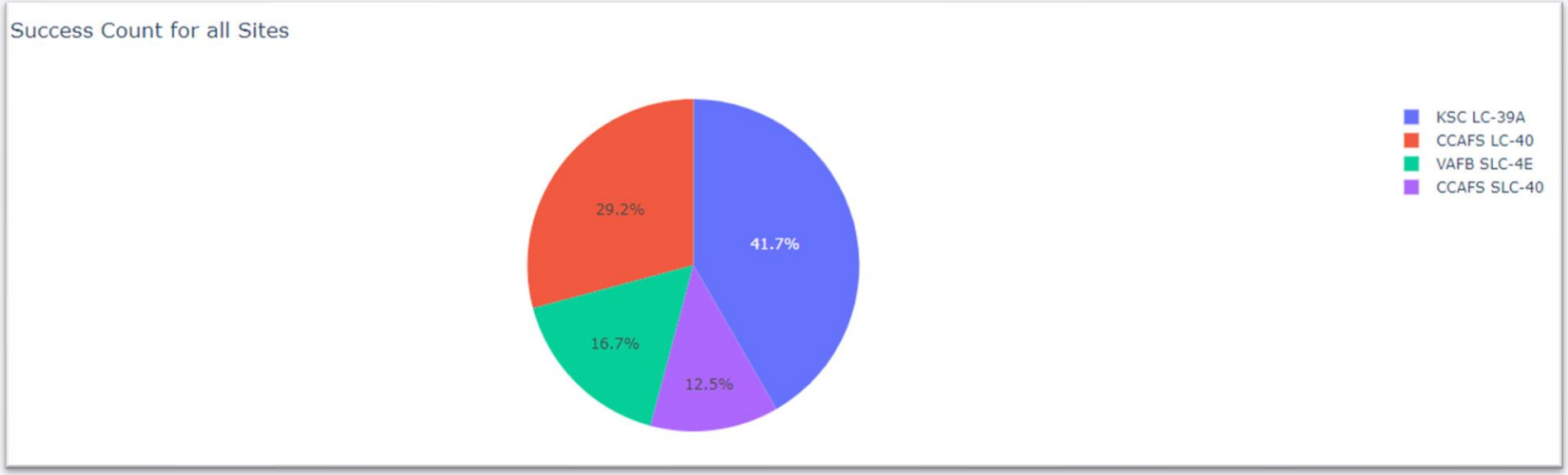




It is possible to check other main features of the Launch Sites, here in the form of “Distances”

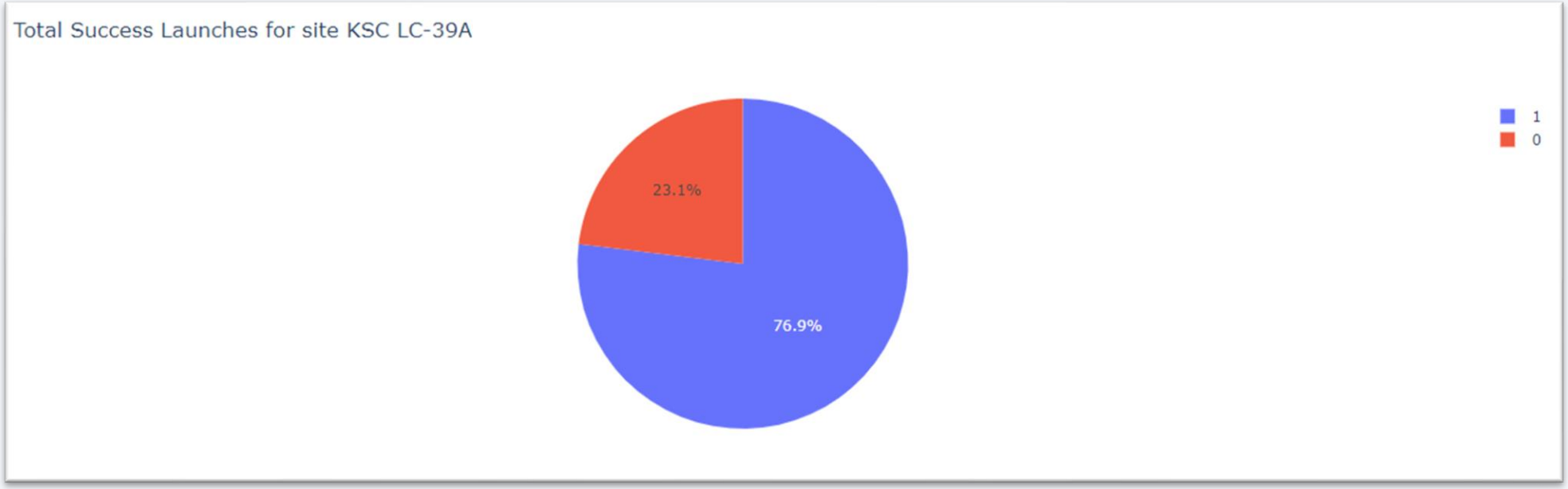

Section 4
Build a Dashboard with Plotly Dash

Launch Success for all Launch Sites



- In the pie chart it is possible to appreciate the Launch Sites that contributed to successful launches the most. The KSC LC-39A (about four over ten) contributed the most, followed by CCAFS LC-40 (about three over ten), then by VAFB SLC-4E and CCAFS SLC40 (about one over ten).

Success Rate for KSC LC-39A



- The pie chart allows us to visually understand the proportion of successful and unsuccessful launches occurred at a specific Launch Site (here KSC LC-39A is depicted).
- In this case, three over four launches performed at the selected location were successful.

Payload vs. Launch Outcome

- This scatter point plot allows us to inspect the Launch Outcome (1 = success, 0 = failure) for each Booster Version Category in a self-selected range of Payload Mass.

Payload vs. Launch Outcome (with slider)

- It is possible to focus on some relevant Payload Mass value by using the slider above.
- Here it is possible to see that the FT Booster Version Category appears to be the one with the most Success (Payload ~ 1000-7000 kg).
- On the other hand, the most unsuccesful appears to be the v1.1.

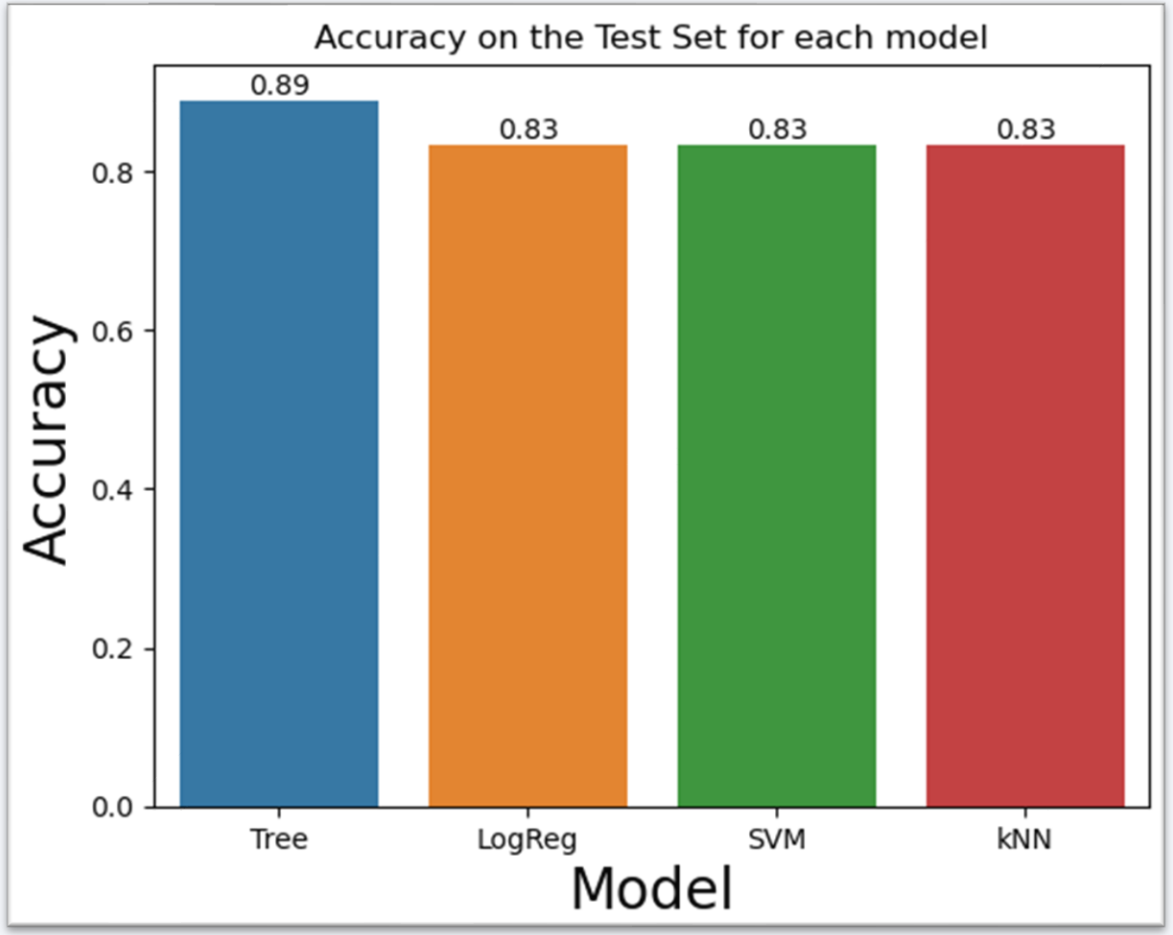
Section 5

Predictive Analysis (Classification)

Classification Accuracy

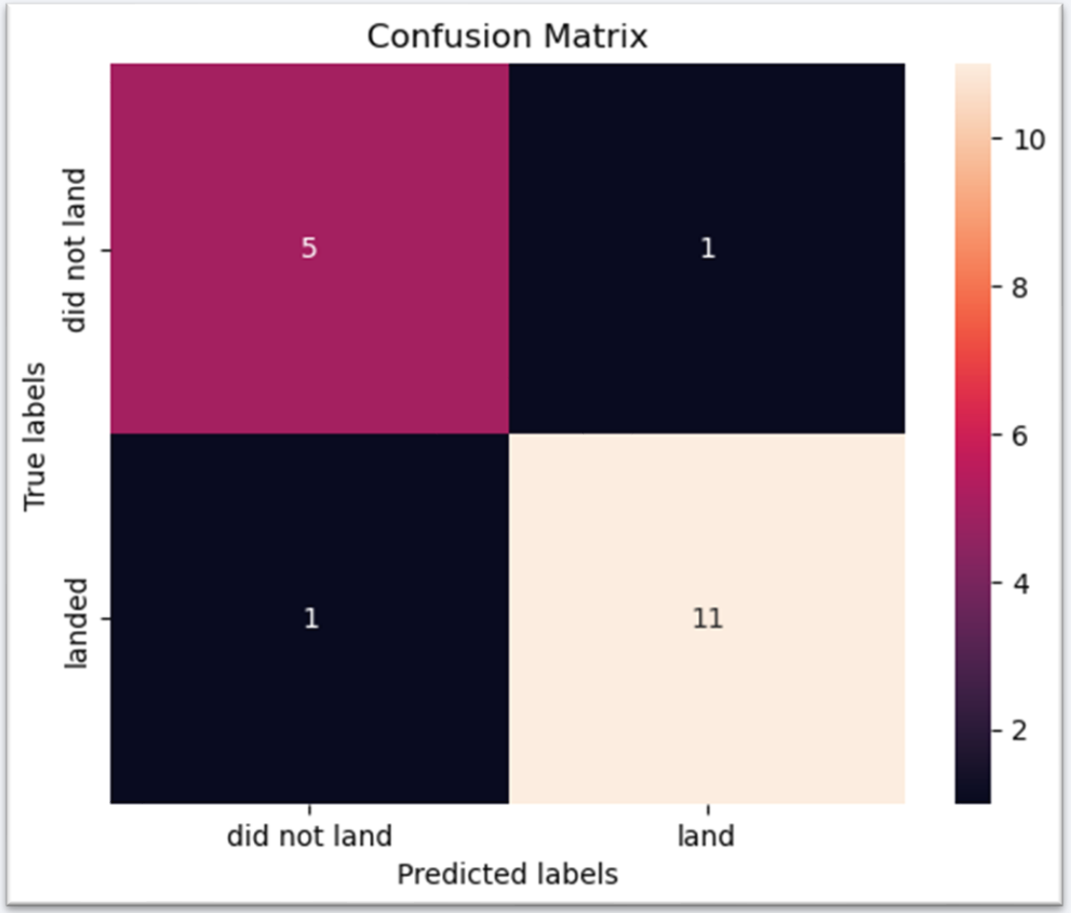
- The model with the best Test Set accuracy was the **Decision Tree**, with almost 90% of correctly predicted launch outcomes.
- The other analyzed models performed slightly worse, with about 80% of accuracy.

Under the proposed assumptions, it is recommended to make further predictions using a Decision Tree



Confusion Matrix

- The best performing model was the **Decision Tree**.
- In the confusion matrix it is possible to appreciate it correctly predicted:
  - 11 successful landings (True Positive)
  - 5 unsuccessful landings (True Negative)
- The model was erroneous in predicting 1 successful landing (False Negative – bottom left) and one unsuccessful one (False Positive – top right)



Conclusions

- API Requests and Web Scraping appeared to be a good tool for retrieving data to be further organized into a readable structure (e.g., dataframe).
- The proposed dashboard allows us to understand in advance the relationship between the payload mass and the successfulness of a launch, as well as the correlation between the latter and the booster version family.
- The interactive map gave us insights about the Launch Sites main features, such as their distance from strategic sites (cities, railroads, ocean) and the rate of success of each Launch Site.
  - It is now possible to select our own Launch Site according to the info we gathered.
- The best model for classifying the landing outcome appeared to be the **decision tree**.
  - The accuracy showed by the model (~90%) allows us to predict with a **high accuracy** if a launch would be successful or not, under the proposed assumptions.

Appendix

- All the material through which this final report was developed are available at the following GitHub Repository: github.com/Maskul93/IBM-Applied-DataScience-Captsone

Thank you!