IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Arijit Karmakar
20/05/2024

Table of contents

Executive Summary

- This project was done using Skills Network Labs and Jupyter notebooks. In order to fulfill all objectives multiple tasks were performed: Data collection with Web Scraping, Data Wrangling, Exploratory Analysis with SQL, Pandas and Matplotlib, created Interactive Visual Analytics and Dashboards, and Classification.
- The final goal of this project is to determinate the price of a launch for our company SpaceY, by gathering public data from SpaceX and training a **Machine Learning model.**
- The results found present us predictions on whether or not the first stage of Falcon 9 will land successfully, since this is most impactful element on the expenses.

Introduction

- The cost of launching a rocket is typically high, but this reality started to change not long ago, when the company SpaceX successfully launched rockets with the cost around 62 million dollars, while other providers expenses are upwards 165 million dollars. This cost difference is mostly due to the fact that SpaceX can reuse the first stage of a rocket, by making it land safely after a launch.
- However, the first stage does not land successfully every time. Therefore, if we were able to predict if the first stage will land, we could determinate the costs of a launch. Through detailed analysis from previous launches, this presentation contains the results on how accurate can we make such predictions.

Section 1

Methodology

Methodology

- Data collection methodology:
  - Request and parse the SpaceX launch data from Open Source REST API for SpaceX, store and filter the Dataframe with the desired data, deal with missing values
  - Web scrap Falcon 9 launch records with BeautifulSoup: Extract a Falcon 9 launch records HTML table from Wikipedia. Parse the table and convert it into a Pandas data frame
- Perform data wrangling
  - Perform exploratory Data Analysis and determine Training Labels
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of mission outcome of each orbit
  - Create a landing outcome label from Outcome column

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL
  - Understand the SpaceX DataSet
  - Load the dataset into the corresponding table in a Db2 database
  - Execute SQL queries to answer assignment questions
  - Predict if the first stage of Falcon 9 will land
- Perform interactive visual analytics using Folium and Plotly Dash
  - Launch Sites Locations Analysis with Folium

Methodology

- Perform predictive analysis using classification models
  - Determine Training labels:
    1. Create a column for the class
    2. Standardize the data
    3. Split into training data and test data
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression:
    1. Create confusion matrix with different methods and test accuracy
    2. Find the method performs best using test data

Data Collection

Collecting Data from SpaceX Open Source API

- Request to the SpaceX API, turn it into a Dataframe, select the desired global variables, create a dictionary, clean the data exclusively for Falcon 9 results, deal with missing data.


```
[
    {
        "fairings": {
            "reused": false,
            "recovery_attempt": false,
            "recovered": false,
            "ships": []
        },
        "links": {
            "patch": {
                "small": "https://images2.imgbox.com/94/f2/NN6Ph45r_o.png",
                "large": "https://images2.imgbox.com/5b/02/QcxHUb5V_o.png"
            },
            "reddit": {
                "campaign": null,
                "launch": null,
                "media": null,
                "recovery": null
            },
            "flickr": {
                "small": [],
                "original": []
            },
            "presskit": null,
            "webcast": "https://www.youtube.com/watch?v=0a_00nJ_Y88",
            "youtube_id": "0a_00nJ_Y88",
            "article": "https://www.space.com/2196-spacex-inaugural-falcon-1-rocket-lost-launch.html",
            "wikipedia": "https://en.wikipedia.org/wiki/DemoSat"
```


SpaceX API URL – How the data looked without any processing

Data Collection – SpaceX API

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details	crew	ships	capsules	payloads
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	Engine failure at 33 seconds and loss of vehicle	[]	[]	[]	[5eb0e4b5b6c3bb0006eeb1e1]
1	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 301, 'altitude': 289, 'reason': 'harmonic oscillation leading to premature engine shutdown'}]	Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage	[]	[]	[]	[5eb0e4b6b6c3bb0006eeb1e2]

Sample of the inital dataframe, created from raw data

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome
1	2010-06-04	Falcon 9	None	LEO	CCSFS SLC 40	None None
2	2012-05-22	Falcon 9	525	LEO	CCSFS SLC 40	None None
3	2013-03-01	Falcon 9	677	ISS	CCSFS SLC 40	None None
4	2013-09-29	Falcon 9	500	PO	VAFB SLC 4E	False Ocean
5	2013-12-03	Falcon 9	3170	GTO	CCSFS SLC 40	None None
...	...	...	...	...	...	...

Sample of the final dataframe, created with processed data

Data Collection – SpaceX API

```
FlightNumber      0
Date              0
BoosterVersion    0
PayloadMass       5
Orbit             0
LaunchSite        0
Outcome           0
Flights           0
GridFins          0
Reused            0
Legs              0
LandingPad       26
Block             0
ReusedCount       0
Serial            0
Longitude         0
Latitude          0
dtype: int64
```

Count of missing values on the data

- Initially, there was five missing values on the column 'PayLoadMass' and twenty six missing values on the column 'LandingPad'.
- The treatment done was to replace the missing values on 'PayLoadMass' by the mean of that column, leaving missing values only on 'LandingPad' column.
- https://github.com/Arijit1707/Applied-Data-Science-Capstone

Data Wrangling

Web scrap Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table, create a dictionary and convert it into a Pandas data frame

[hide] Flight No.	Date and time (UTC)	Version, booster[a]	Launch site	Payload[b]	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]								

Structure of the records displayed on Wikipedia

Data Wrangling

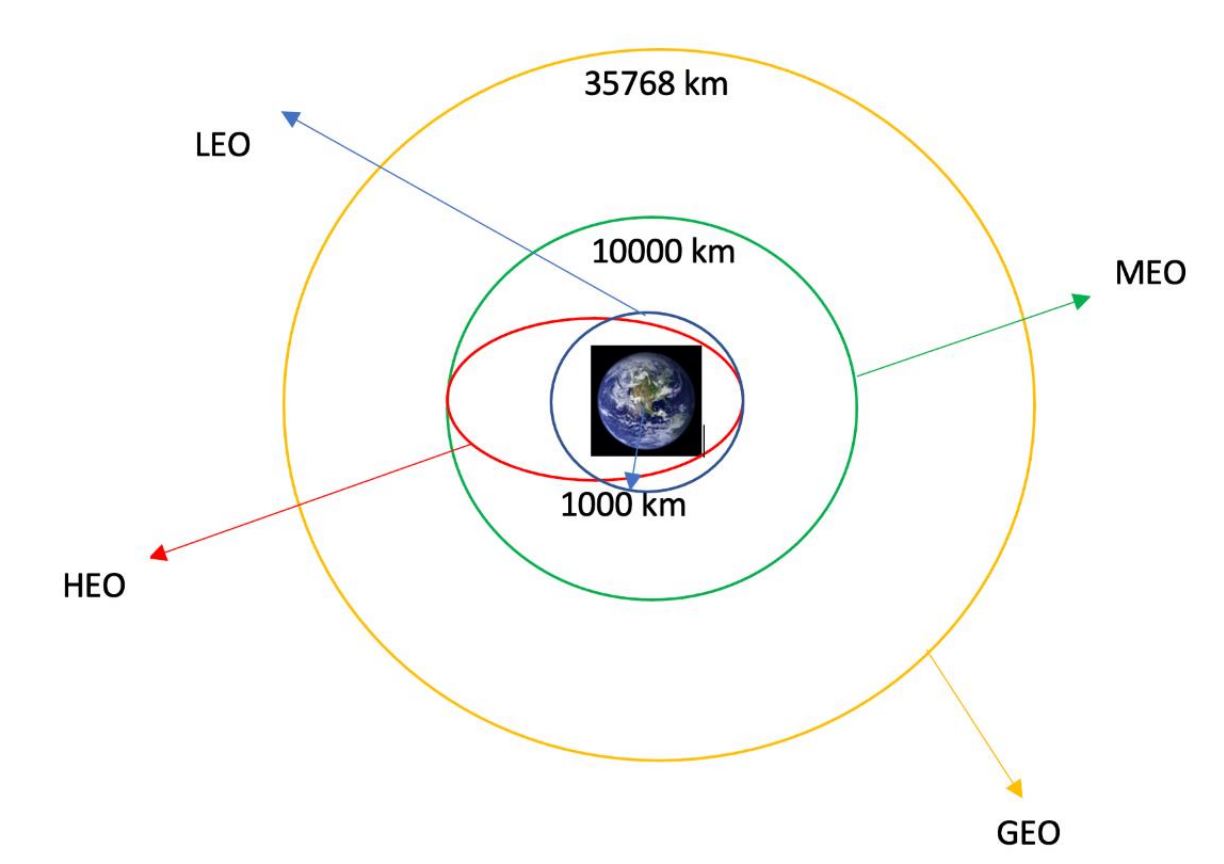
	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	1	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	4 June 2010	15:43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

Sample from the final dataframe created with the data from HTML table on Wikipedia

```
LaunchSite
CCAFS SLC 40     55
KSC LC 39A       22
VAFB SLC 4E      13
dtype: int64
```

Launches grouped by Launch Site

Data Wrangling



Different orbit types and their respective altitude

- The type of orbit a rocket is sent to was also na important information. All launches were grouped based on orbit for further analysis.

```
Orbit
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
GEO       1
HEO       1
SO        1
dtype: int64
```

All launches grouped by orbit type

Data Wrangling

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

Types of landing outcome

- There was eight different landing outcomes such as drone-ship success, drone-ship failure, ground-pad success and etc.
- For the pupose of this reasearch all results were divided in two c lasses, success or failure.
- By the ending of the Data Wrangling phase we had a new dataframe with the column 'Class' identifying all successfull landings with (1) and all the failures with (0)
- https://github.com/Arijit1707/Applied-Data-Science-Capstone

That being said, we are able to determinate the success rate of all records, calculating the mean of the column 'Class', coming to the result of approximately 66,66%

EDA with Data Visualization

- For the Exploratory Data Analysis (EDA) with Data Visualization, the majority of the charts plotted are the type scatter plot
- The scatter plot allow us to quick visualize the distribution of the data between the variables selected on the Y axis.
- There were also used a line chart and a bar chart to better visualize the success rate
- All plots are presented on the Section 2 Insights drawn from EDA
- https://github.com/Arijit1707/Applied-Data-Science-Capstone

EDA with SQL

- For the Exploratory Data Analysis (EDA) with SQL a total of ten queries were carefully selected. The objective of each query was:
- 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 achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA with SQL

- 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 records which will display the date, failure landing outcomes in drone ship ,booster versions, launch site for the 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.
- All plots are presented on the Section 2 Insights drawn from EDA
- https://github.com/Arijit1707/Applied-Data-Science-Capstone

Build an Interactive Map with Folium

- During the Interactive Visual Analytics with Folium lab, three main goals were settled:
  1. Mark all launch sites on a map
  2. Mark the success/failed launches for each site on the map
  3. Calculate the distances between a launch site to its proximities
- To mark all launch sites, Folium marker and circle were used.
- There were also used markers for the launch records, as well as a marker cluster.
- For calculating the distances, the coordinates of strategic locations were used, those being the nearest: Costline, Railway, Highway and City.
- https://github.com/Arijit1707/Applied-Data-Science-Capstone

Build a Dashboard with Plotly Dash

- The Dash Application is an interactive source to easily visualize strategic data through pie charts and scatter plots.
- The pie chart display the success rate across the launch sites selected. The scatter plot display the payload mass, the booster version and the landing outcome.
- These plots can be filtered by launch site. The scatter plot can also be filtered to show only a specific payload range.
- The plots and interactions added were carefully chosen to provide an easy and clear insight in the relation between successful landings, their location, payload mass and booster version.
- https://github.com/Arijit1707/Applied-Data-Science-Capstone

Predictive Analysis (Classification)

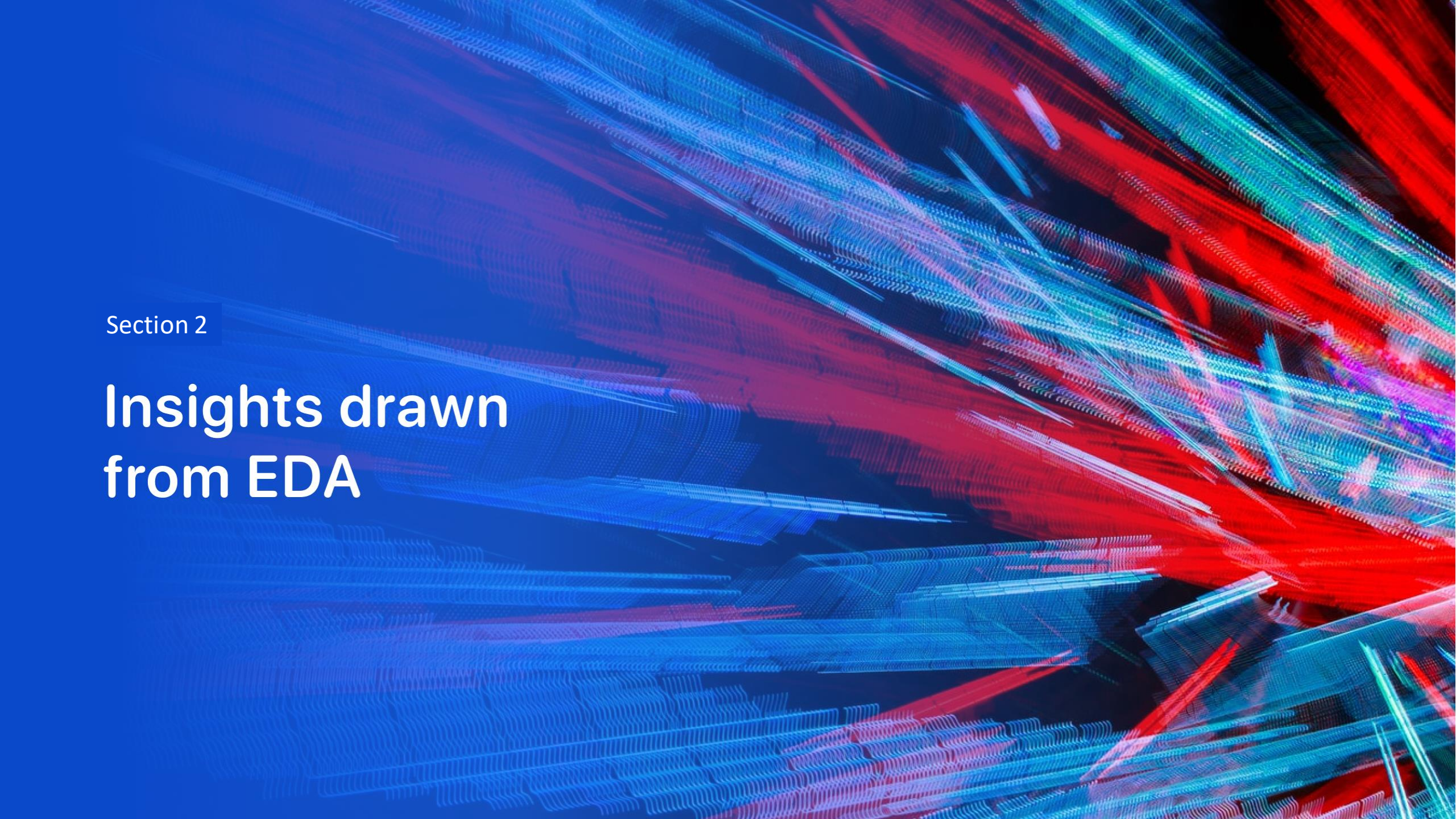
- For the predictive analysis, four different prediction methods were used, those being: Logistic Regression, Support Vector Machine (SVM), Tree Classifier and Key Nearest Neighbour (KNN).
- The first step of the process is to load the Dataframe, standardize the data and split the data in train set and test set. After, we are able to use this data in the different methods mentioned, verify their accuracy and create a confusion matrix
- https://github.com/Arijit1707/Applied-Data-Science-Capstone

Results

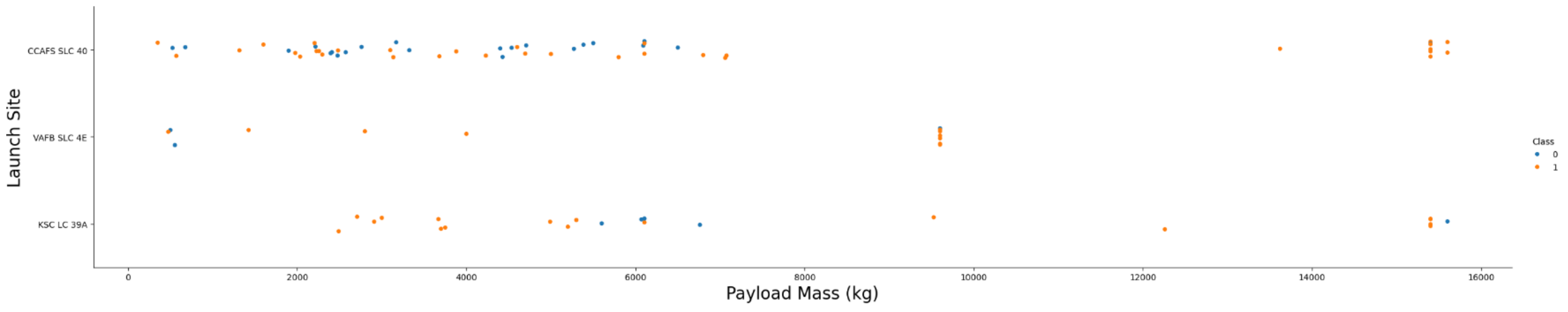
- **Some of the results obtained include:**
- The success rate on the landing of the first stage from Falcon 9 Launches
- Key insights from each launch site and type of orbit
- Analysis on the relation between data
- Interactive Visual Analytics through the Dash application and Folium Map.
- Predictive analysis results testing different classification models

Section 2

Insights drawn from EDA

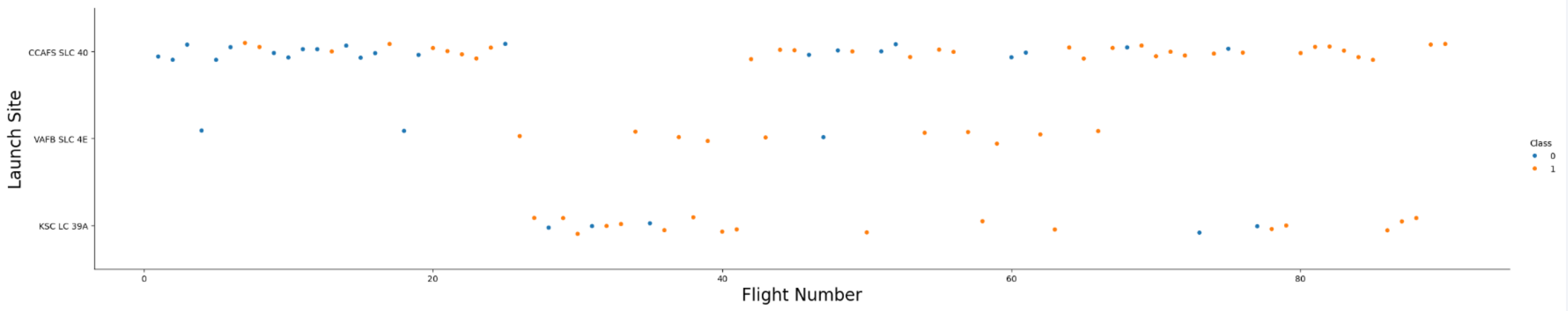


Payload vs. Launch Site



Scatter plot showing the relation between Payload Mass and Launch Site. When Payload Mass is above 8000kg the success rate is very high.

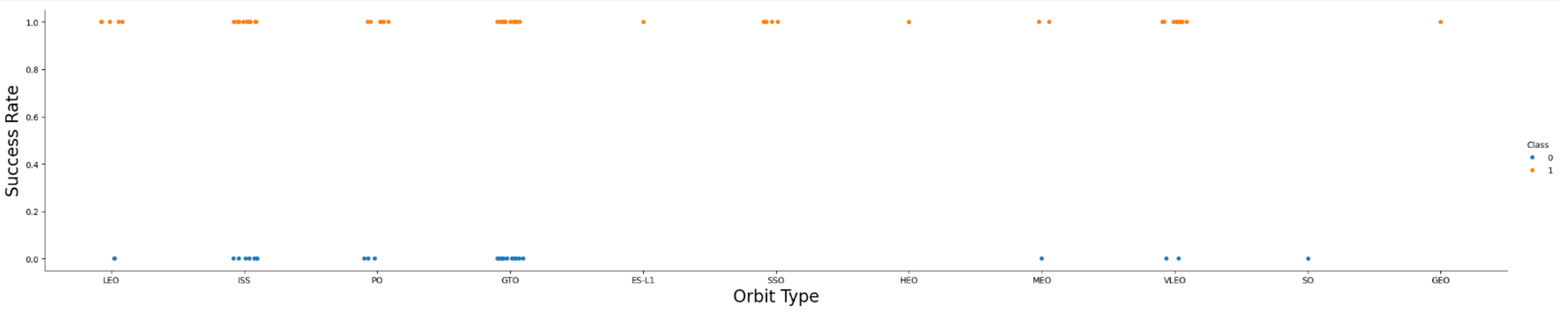
Flight Number vs. Launch Site



Scatter plot showing the relation between Flight Number and Launch Site.

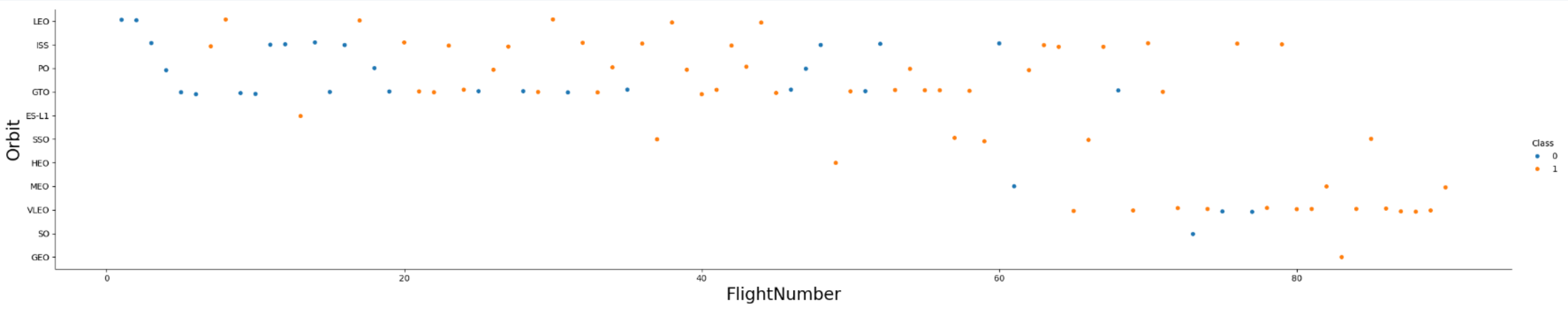
*Please note que "Class" label on the right identifying the successful results in Yellow (1) and the failure in Blue (0) about the landing of the first stage.

Orbit Type vs. Success Rate



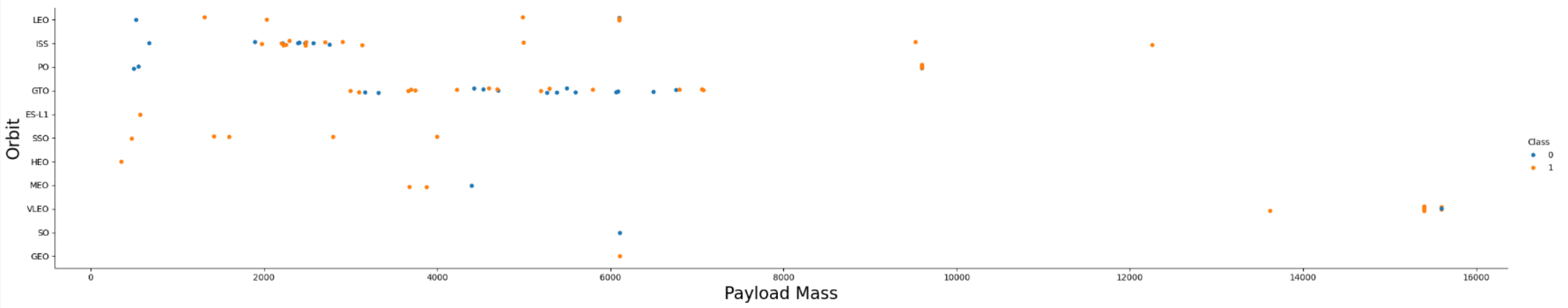
Scatter plot showing the relation between Orbit Type and Success Rate. Most of the failures are displayed on the orbits ISS and GTO.

Flight Number vs. Orbit Type



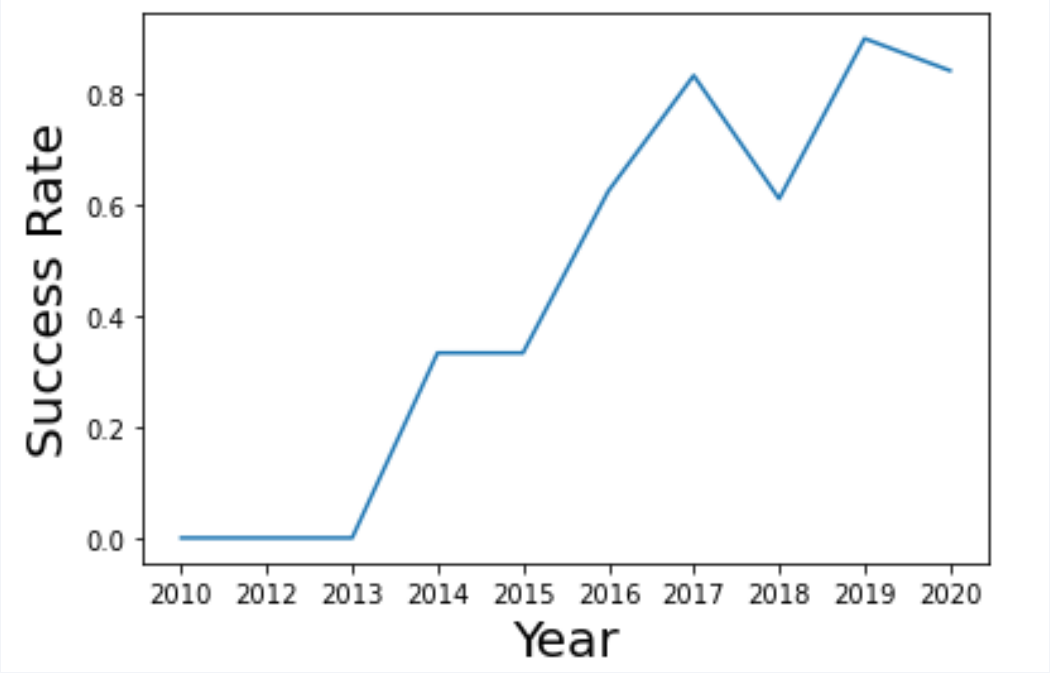
Scatter plot showing the relation between Flight Number and Orbit Type. Between the first 20 Flights occured the most failures.

Payload vs. Orbit Type

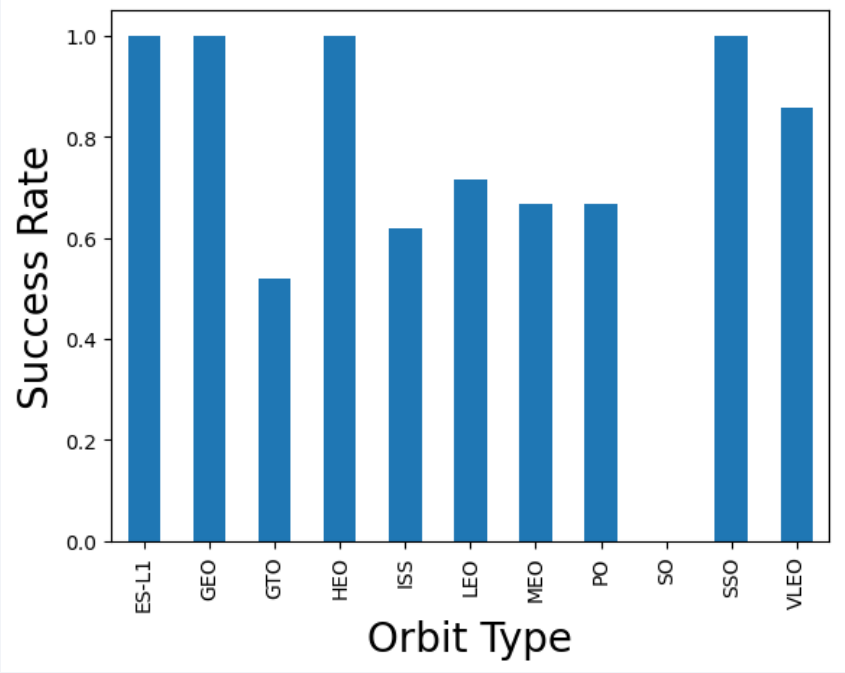


Scatter plot showing the relation between Payload Mass and Orbit Type. Most of the rockets launched have Payload between 2000kg and 6000kg, sent mostly to ISS and GTO orbits.

Launch Success Trend



Line chart showing the relation between Year and Success Rate.



Bar chart showing the relation between Orbit Type and Success Rate.

All Launch Site Names

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

All launch sites present in the data used

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Five example launch records from launch sites names begin with 'CCA'

Total Payload Mass

Customer	Total_Payload_Mass_CRS
NASA (CRS)	48213

Total Payload mass carried by boosters launched by NASA

Average Payload Mass by F9 v1.1

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1 B1003	2534.6666666666665

The average Payload mass carried by boosters F9 v1.1

First Successful Ground Landing Date

Date	Landing_Outcome
2015-12-22	Success (ground pad)

The date of the first successful landing

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	Payload	PAYLOAD_MASS__KG_	Mission_Outcome
F9 FT B1022	JCSAT-14	4696	Success
F9 FT B1026	JCSAT-16	4600	Success
F9 FT B1021.2	SES-10	5300	Success
F9 FT B1031.2	SES-11 / EchoStar 105	5200	Success

All the successfull drone ship landing with Payload mass between 4000kg and 6000kg

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	TOTAL_NUMBER
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Total Number of Successful and Failure Missions

Boosters Carried Maximum Payload

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

List of all boosters that carried the maximum payload in missions

2015 Launch Records

Landing_Outcome	Booster_Version	Launch_Site	Date
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-10-01
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

Mission records from the year 2015

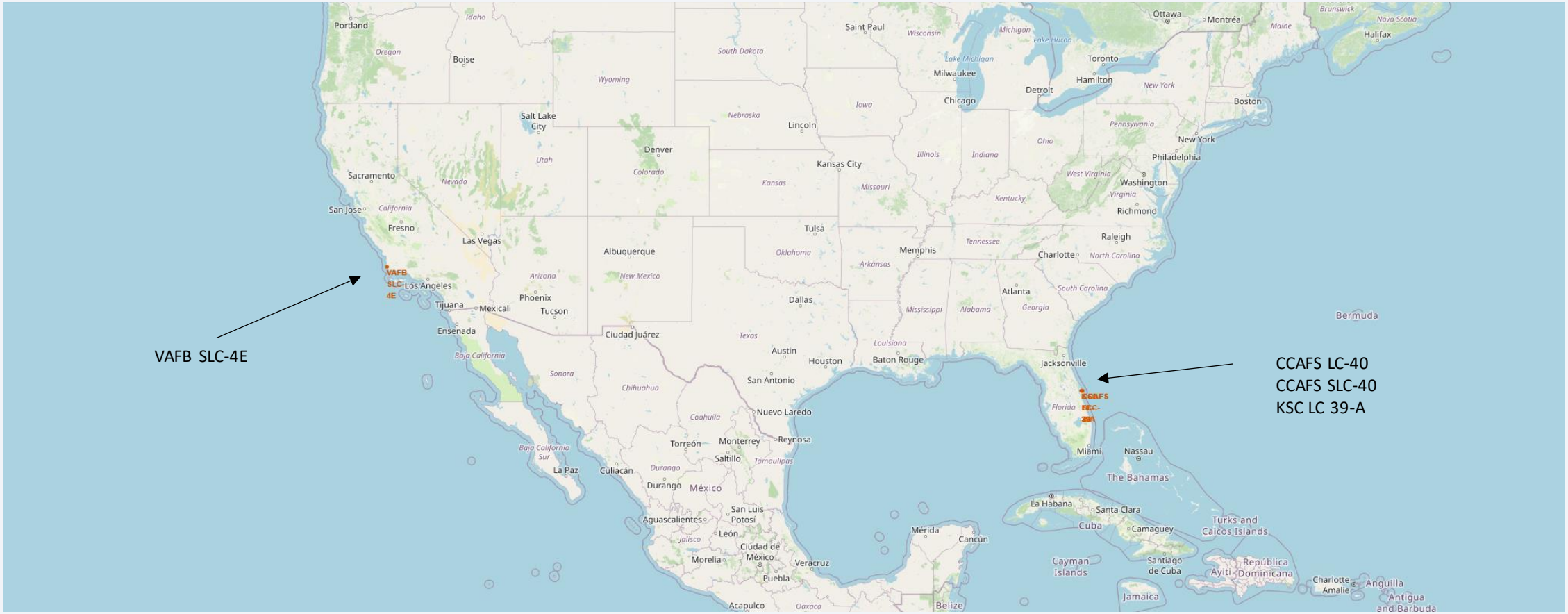
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Landing_Outcome	TOTAL_NUMBER
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

The amount of each landing outcome between June 4th 2010 and March 20th 2017

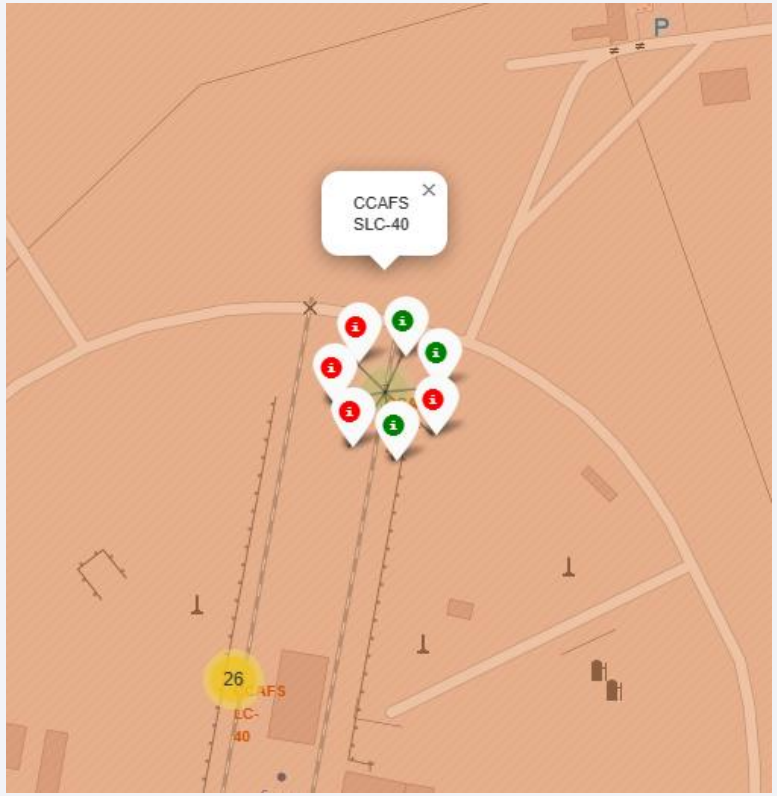

Section 3
Launch Sites
Proximities Analysis

Folium Map – All launch sites location



All launch sites location. As we can see, all of them are located near the coast line.

Folium Map – Launch records



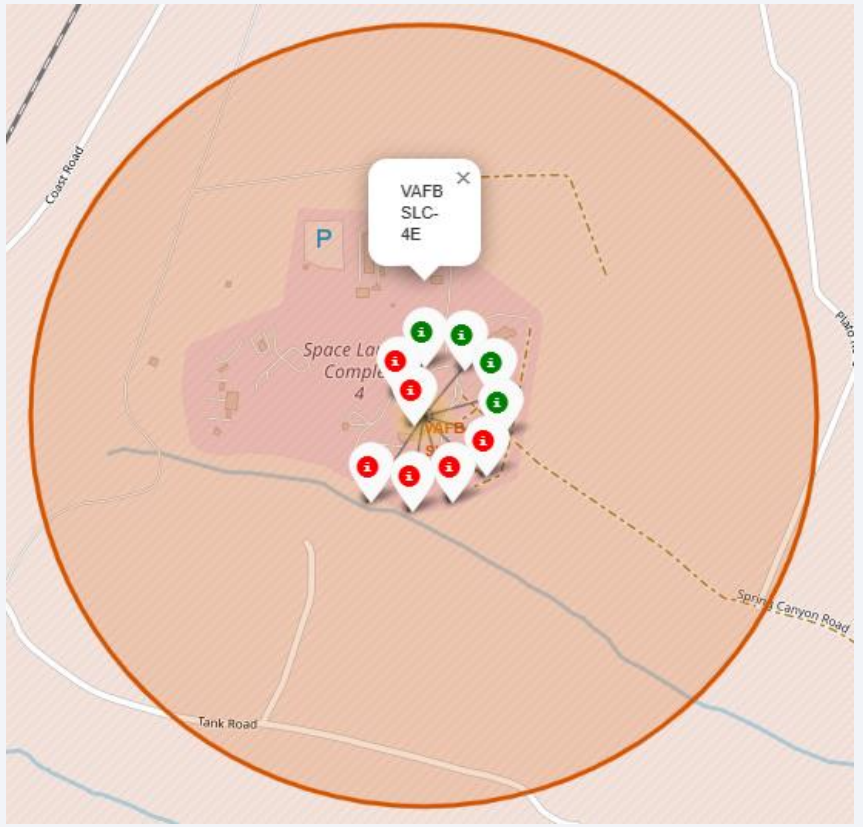
Launch records on CCAFS SLC-40

Launch records on CCAFS LC-40

*Note: The records showed in green represent success and the red ones represent failure

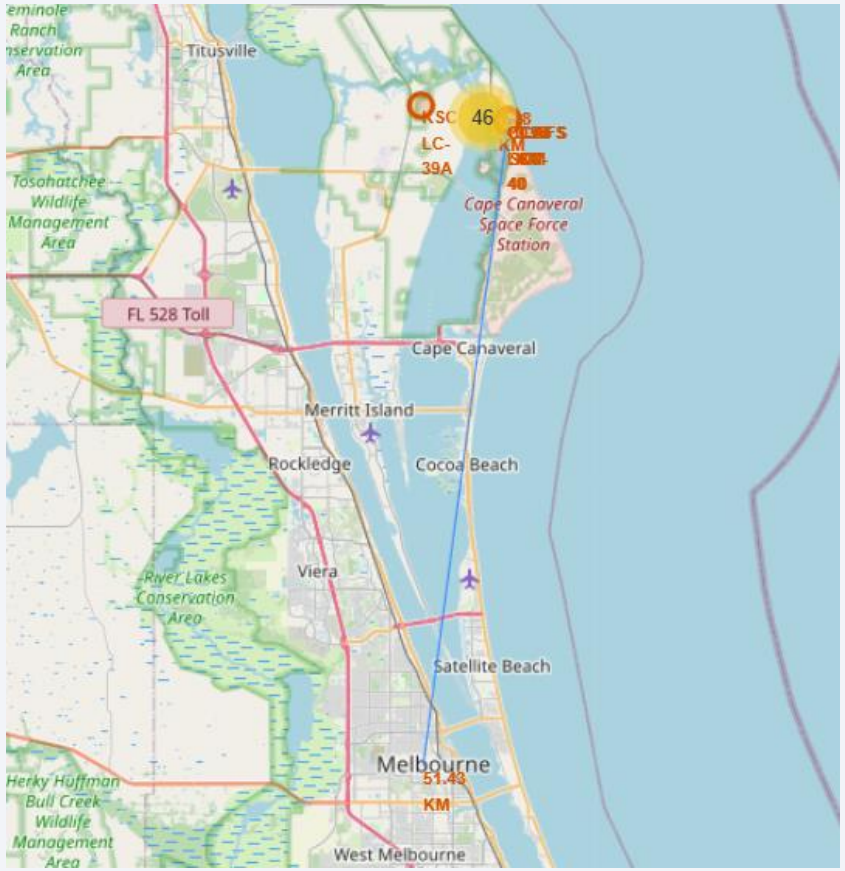
Folium Map – Launch records

Launch records on KSC LC-39A

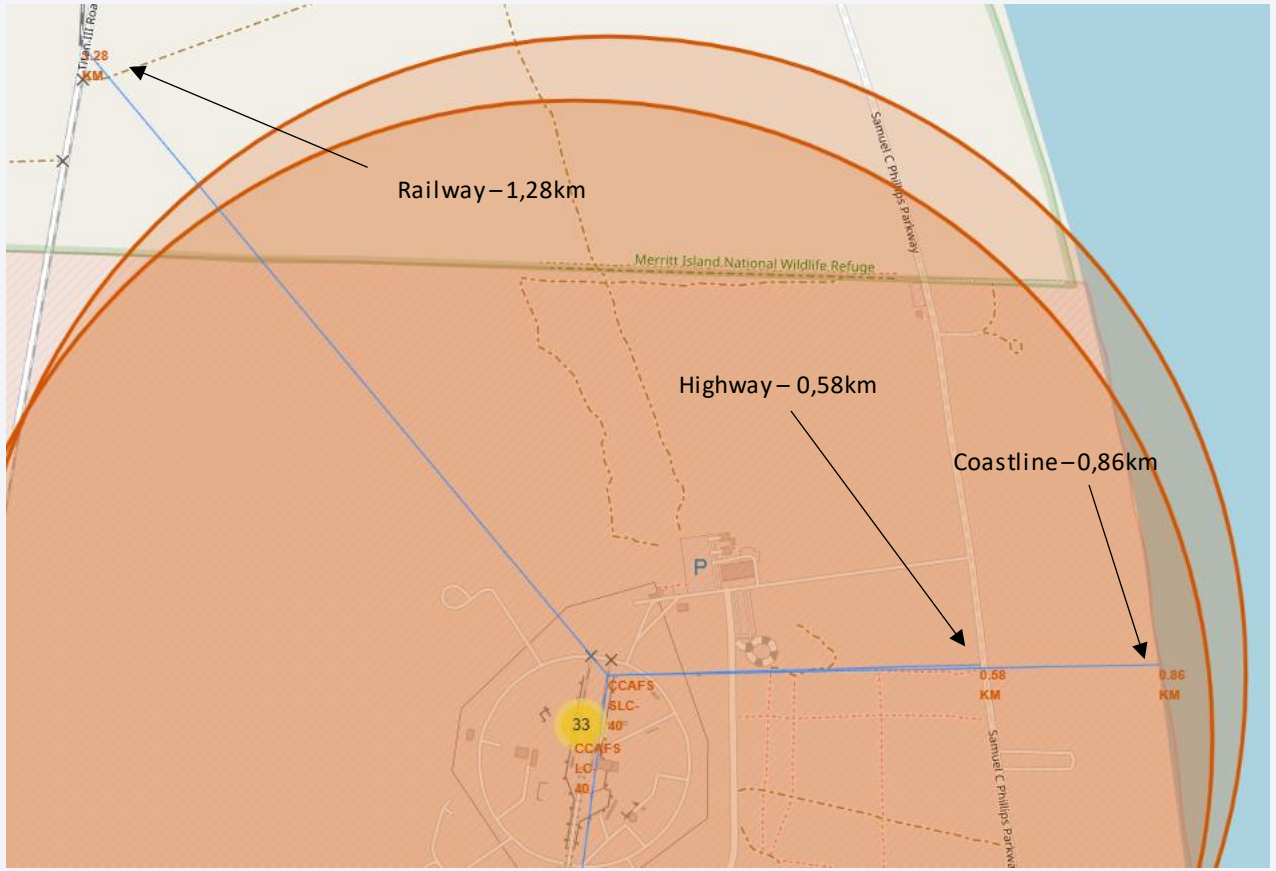


Launch records on VAFB SLC-4E

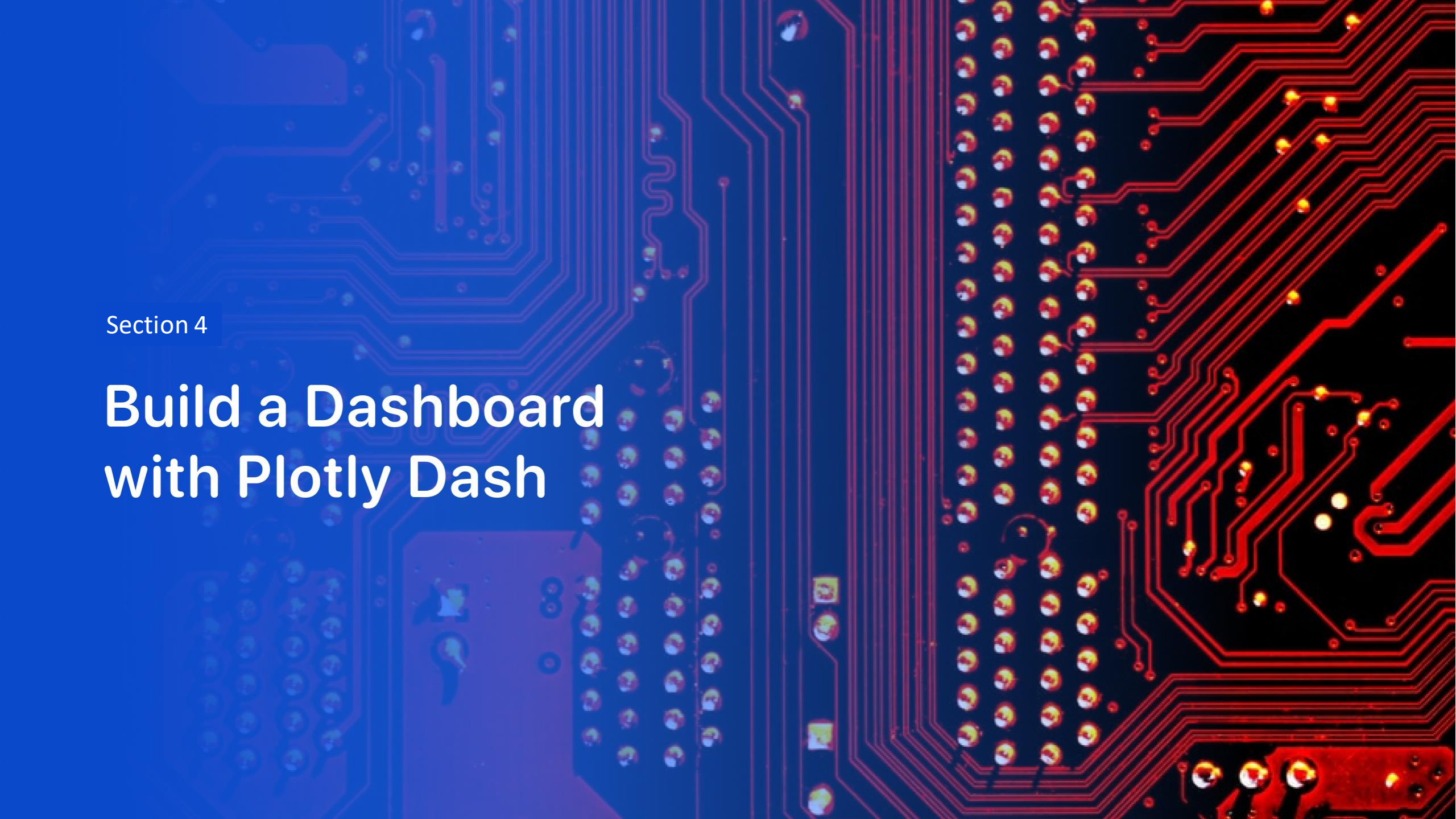
Folium Map – Distance from target locations



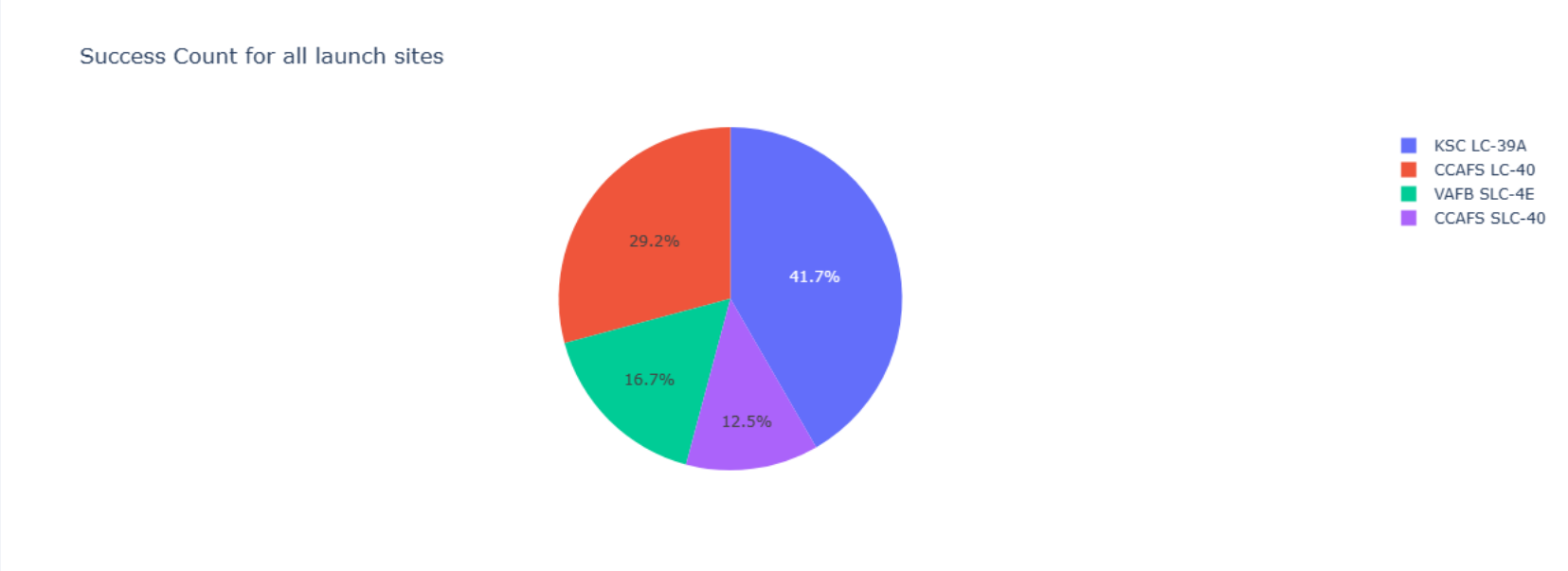
Distance of CCAFS SLC-40 to the nearest City, Melbourne. The distance in a straight line is 51,43Km



Distance of CCAFS SLC-40 to the Coastline, Highway and Railway. As we can see, this launch site is located near to transport lines and far from the city.


Section 4
Build a Dashboard with Plotly Dash

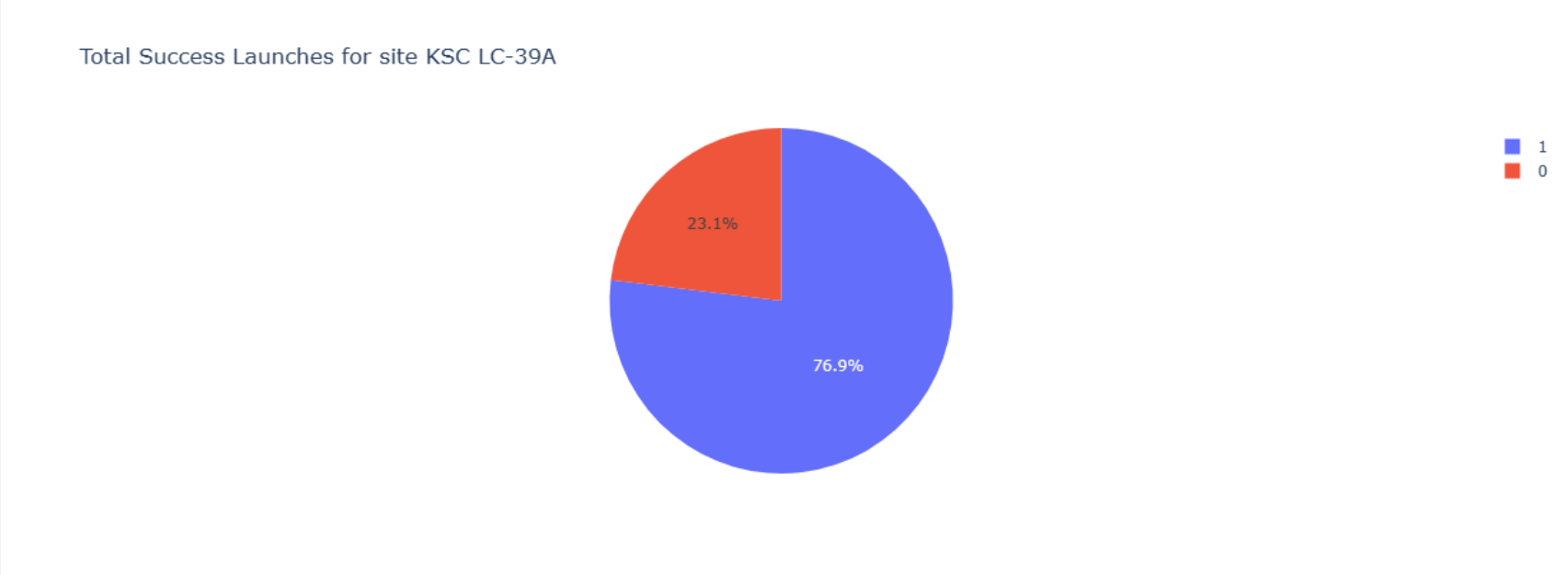
Dash Application - Piecharts



On the piechart above we are able to see the percentage each launch site represents from the total number of successful landings. The launch site with the largest success percentage is KSF LC-39A with 41,7% of the total and CCAFS SLC-40 with the lowest success percentage of only 12,5%.

*Note: The launch site **CCAFS LC-40** is just the old name of **CCAFS SLC-40**.

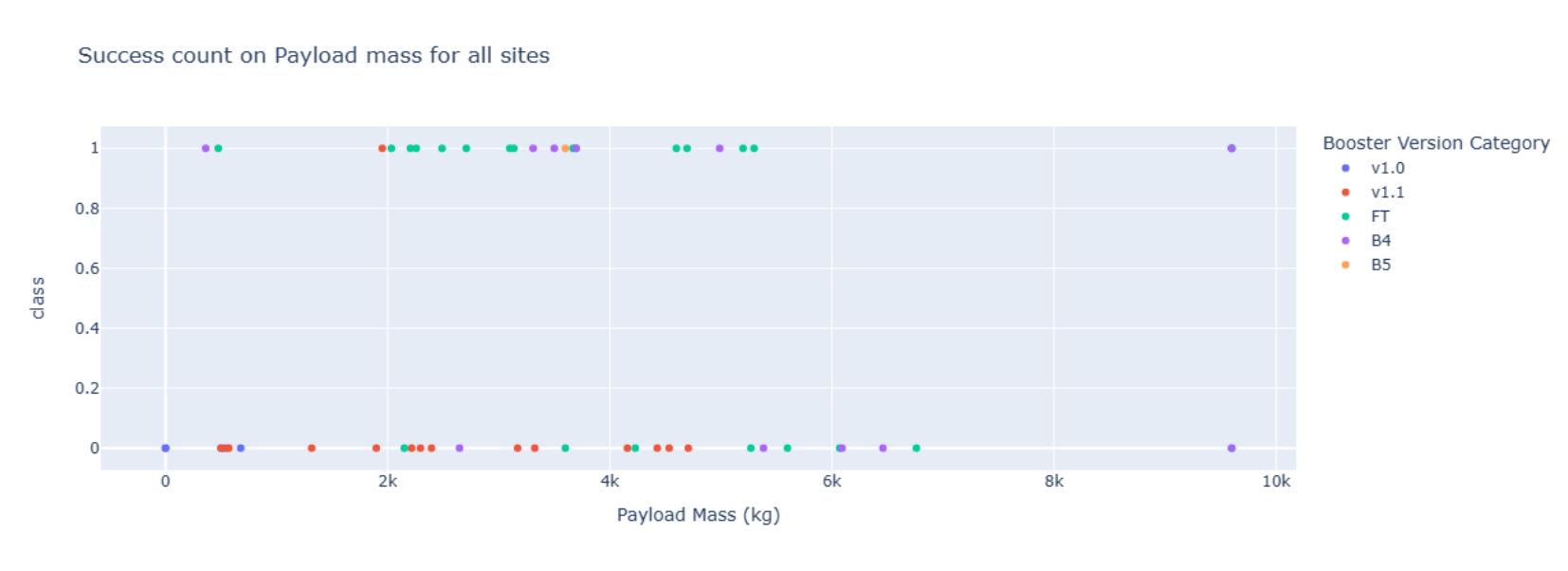
Dash Application - Piecharts



On the piechart above we have the success rate of 76,9% for the launch site KSC LC-39A. Note the subtitles on the right indicating the color blue (1) for success and orange (0) for failure.

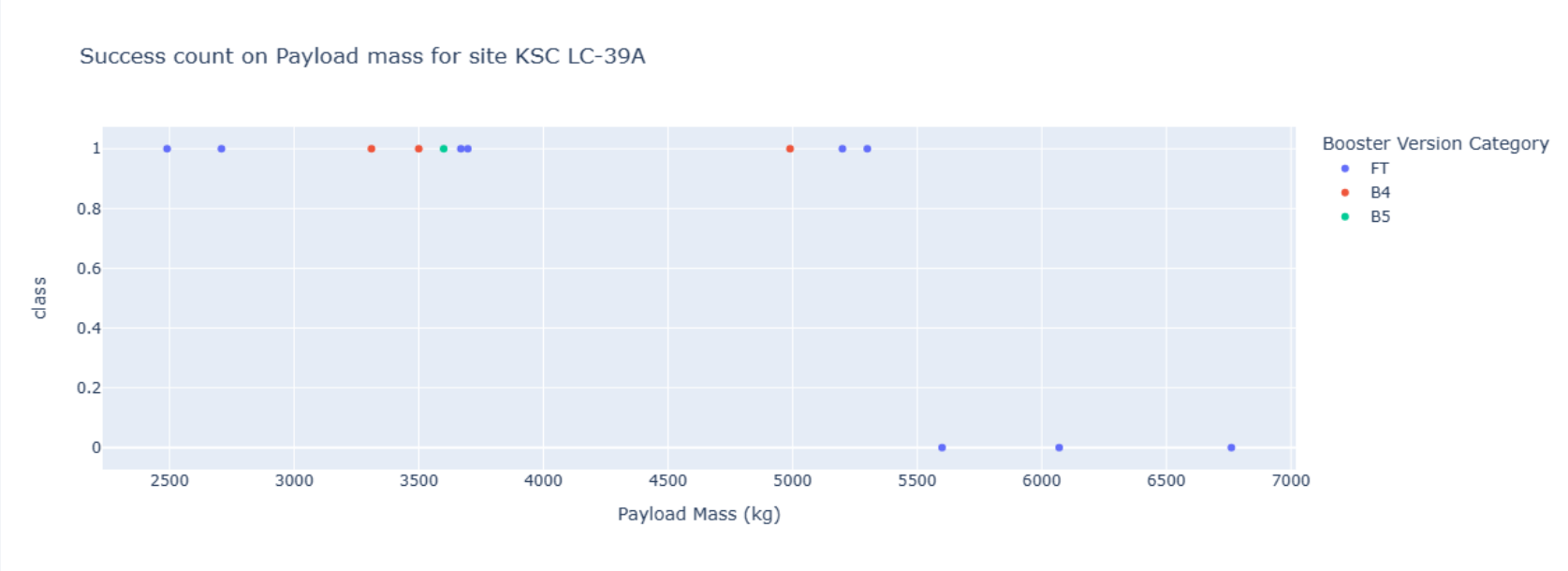
- Through the dash application built we are capable of selecting interactively any launch site to visualize its respective piechart.

Dash Applications – Scatter plot



On the scatter plot above we can see the relation between the Payload mass range (From 0kg to 10.000kg) and the landing outcome (Class 0 or 1) for all launch sites. The Subtitles on the right identify each Booster version. The most successful booster version is FT, showed in green in this plot.

Dash Applications – Scatter plot



On the scatterplot above we can see the relation between the Payload mass range (From 2.500kg to 7.000kg) and the landing outcome (Class 0 or 1) for the launch site KSC LC-39A.

- Through the dash application built we are capable of selecting interactively any launch site and any payload range to visualize its respective scatterplot.

Section 5

Predictive Analysis (Classification)

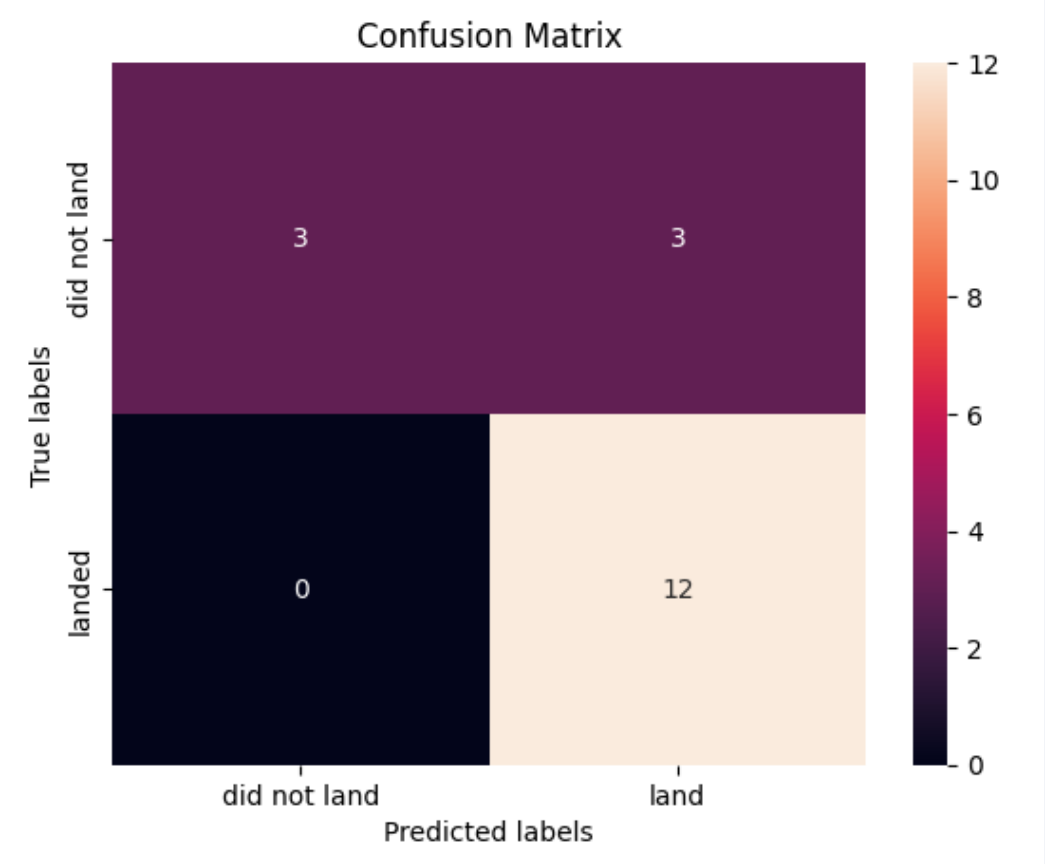
Classification Accuracy

- After setting the parameters and applying a classification model, the final step was to test the accuracy on the test data using the method score.
- All models tested: Logistic Regression, Support Vector Machine (SVM), Decision Tree and K Nearest Neighbour (KNN) scored the same accuracy of approximately **83,33%.**

	Prediction Model	Accuracy
0	Logistic Regression	0.833333
1	SVM	0.833333
2	Decision Tree	0.833333
3	KNN	0.833333

Accuracy results obtained from each model

Confusion Matrix



The Confusion Matrix for all models tested had the same result as showed in the image above

Conclusions

- Through thorough data collection, preparation, analysis, processing and visualization, it was possible to extract **key insights** that help us understand better the whole process, from the launch of a rocket until the landing of it's first stage.
- Through the use of **Machine Learning**, the company SpaceY is now capable of predicting whether the first stage of the rocket will land or not with a precision of **83,33%**.
- Being able to predict if the first stage will land or not could guarantee a higher chance of dramatically reduce the expenses, and put SpaceY in a more competitive scenario.
- The success rate on the landing of the first stage obtained from the data was 66,66%, however, it is important to notice that not all failures were by accident, since in some cases there was no try to land at all.

Appendix

- Wikipedia link used on the Webscraping phase: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Github link for all project content: htthttps://github.com/Arijit1707/Applied-Data-Science-Capstone
- Course link in coursera: https://www.coursera.org/learn/applied-data-science-capstone?myLearningTab=IN_PROGRESS
- IBM Data Science professional certificate: https://www.coursera.org/professional-certificates/ibm-data-science?myLearningTab=IN_PROGRESS#courses

Thank you!