IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Presented by Jordana Bauch

Date of publication: 30.01.2025

Outline

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

Executive Summary

- Methodology
  - Data collection, wrangling and formatting
  - Exploratory Data Analysis
  - Data Visualization
  - Machine Learning Prediction (Classification)
- Results
  - We found a correlation between some of the features of rocket launches and the success of the launch.

Introduction

- SpaceX offers relatively inexpensive rocket launches: Falcon 9 launch ~ 62 Mio. USD
- Why? – They can reuse the (quite expensive) first stage.
- To minimize the cost of a launch, it is the goal to have a successful landing of the first stage so that it can be reused
- That leads to the following questions:
  - What features of the rocket launch of the Falcon 9 have an impact on the outcome of the landing?
  - Do some of the features have a positive impact on the landing and can therefore lead to a successful outcome?
  - Can we predict if a landing will be successful or not so that we can calculate the cost of the launch accordingly?

Section 1

Methodology

Methodology

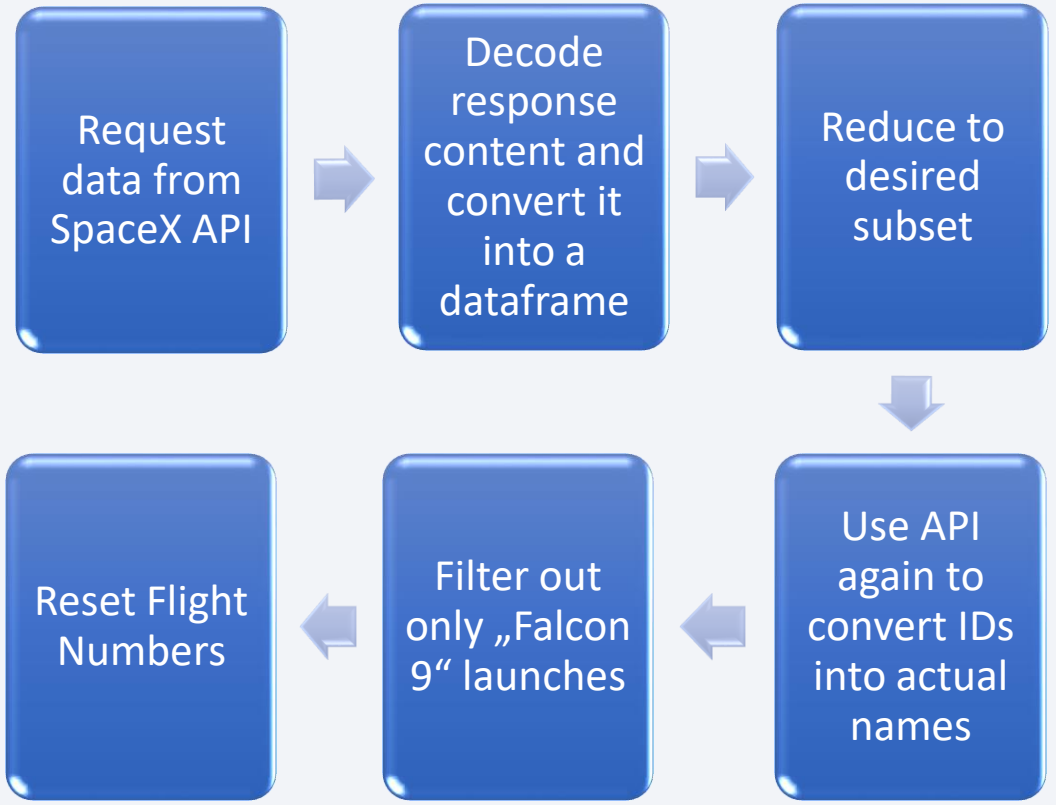
- Data collection:
  - SpaceX REST API
  - Web Scraping from Wikipedia
- Data wrangling:
  - Cleaning data, removing unwanted entries, providing a binary outcome label “Class”
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Training and testing different ML models and evaluating their accuracy

Data Collection

- To obtain relevant Data, we used:
  - API request from SpaceX REST API using Get Request
  - Webscraping from SpaceX's Wikipedia page using BeautifulSoup

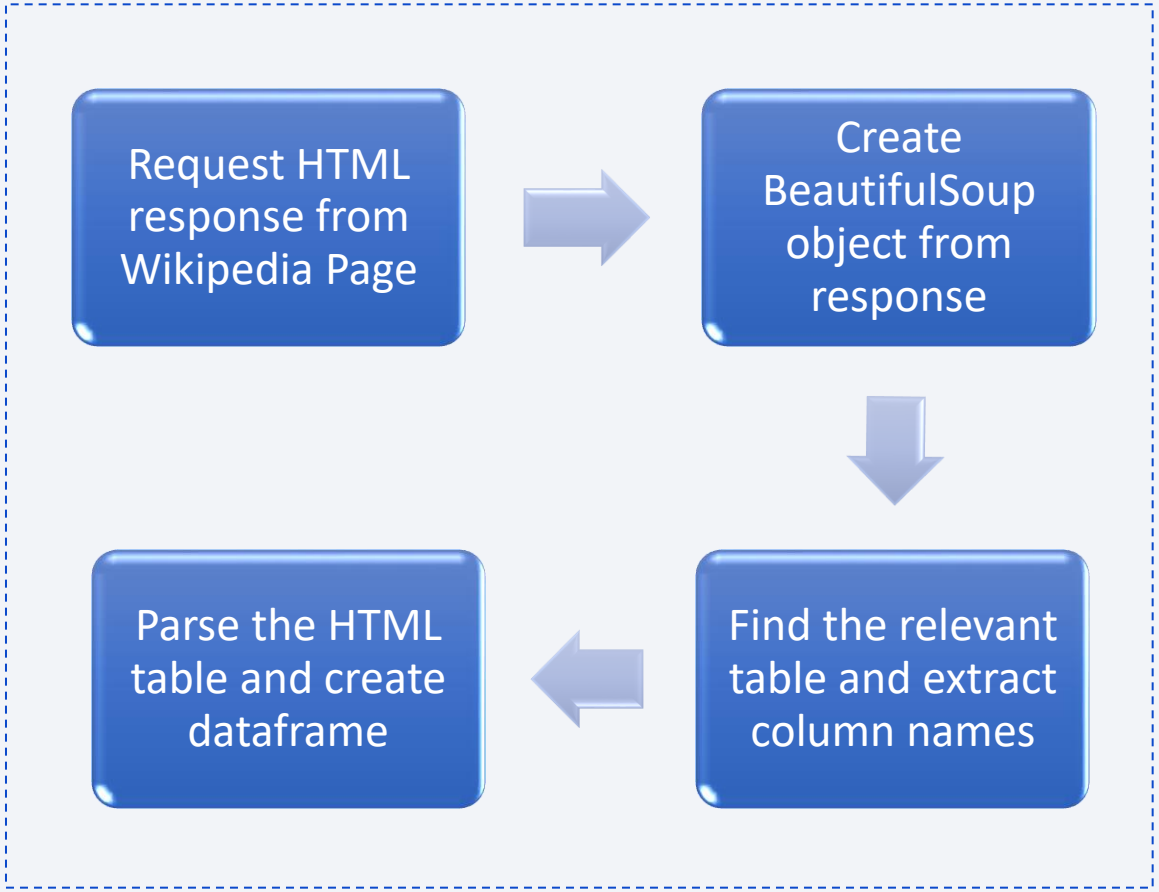
Data Collection – SpaceX API

- A get request from SpaceX REST API provides a JSON-Format result, that can be transformed into a Pandas dataframe
- Obtained values: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, and more
- https://github.com/jordi-bee/datasciencestudy/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

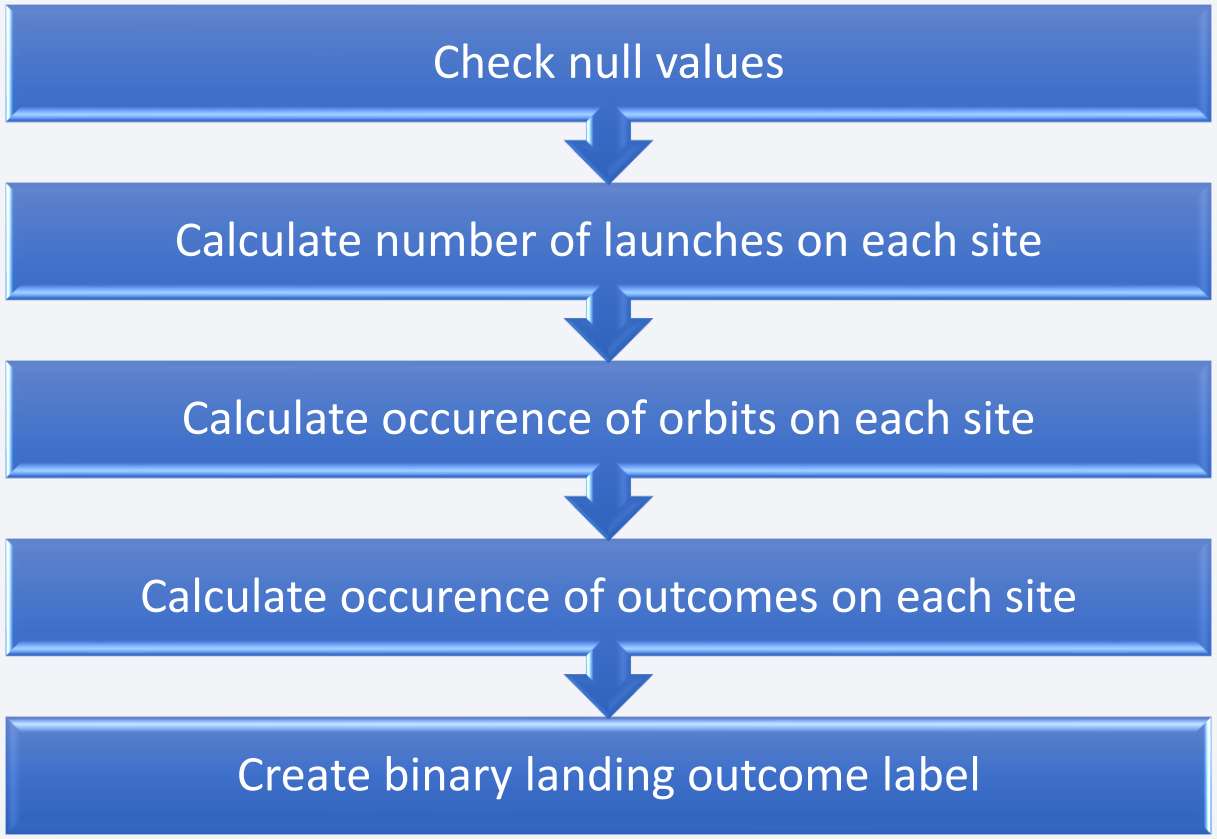
- A HTTP get request is performed to obtain an HTML response
- The HTML response is then parsed and the relevant data is stored in a Pandas dataframe
- Obtained values: Launch Site, Payload mass, Orbit, Customer, Launch Outcome, and more
- https://github.com/jordi-bee/datasciencestudy/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data was filtered to only “Falcon 9” launches
- Only relevant columns were included
- Missing values were handled
- Outcome label was changed to be binary

- https://github.com/jordi-bee/datasciencestudy/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Charts plotted to visualize a possible correlation between the variables and the outcome (Class):
  - Scatterplot: Flight No. and Payload vs. Class
  - Scatterplot: Flight No. and Launch Site vs. Class
  - Scatterplot: Payload Mass and Launch Site vs. Class
  - Bar Chart: Success Rate on each Orbit
  - Scatterplot: Flight No. and Orbit vs. Class
  - Scatterplot: Payload Mass and Orbit vs. Class
  - Line Plot: Launch Success Yearly Trend
- https://github.com/jordi-bee/datasciencestudy/blob/main/edadataviz.ipynb

EDA with SQL

- Names of unique launch sites
- Five records where launch site starts with “CCA”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing in ground pad
- Boosters with success in drone ship and payload mass between 4.000 and 6.000 kg
- total number of successful and failure mission outcomes
- Booster Versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, booster versions and launch sites for the months in 2015
- https://github.com/jordi-bee/datasciencestudy/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Launch Sites on a US-centered Folium-Map with Markers identifying each site
- Clustered markers for successful and failed landings on each site for a visual overview
- Distance between Launch site VAFB SLC4E and the coastline to show proximity of the ocean

- https://github.com/jordi-bee/datasciencestudy/blob/main/lab_jupyter_launch_site_location.ipynb

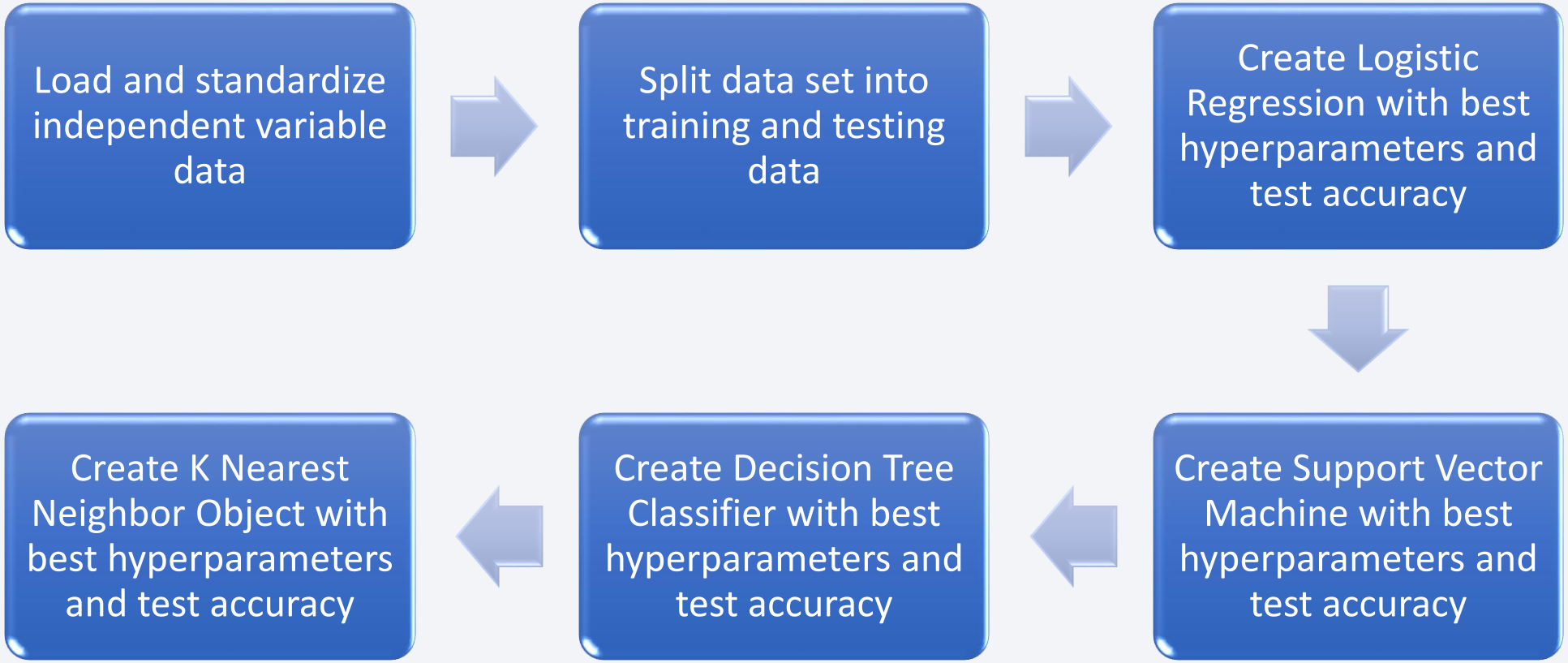
Build a Dashboard with Plotly Dash

- Interactive Pie chart to show the total success rate of each launch site or all launch sited, selected from a dropdown list
- Interactive Scatterplot of Payload Mass (selected with Slider) and the resulting Success Rate of the different Booster Versions
- https://github.com/jordi-bee/datasciencestudy/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

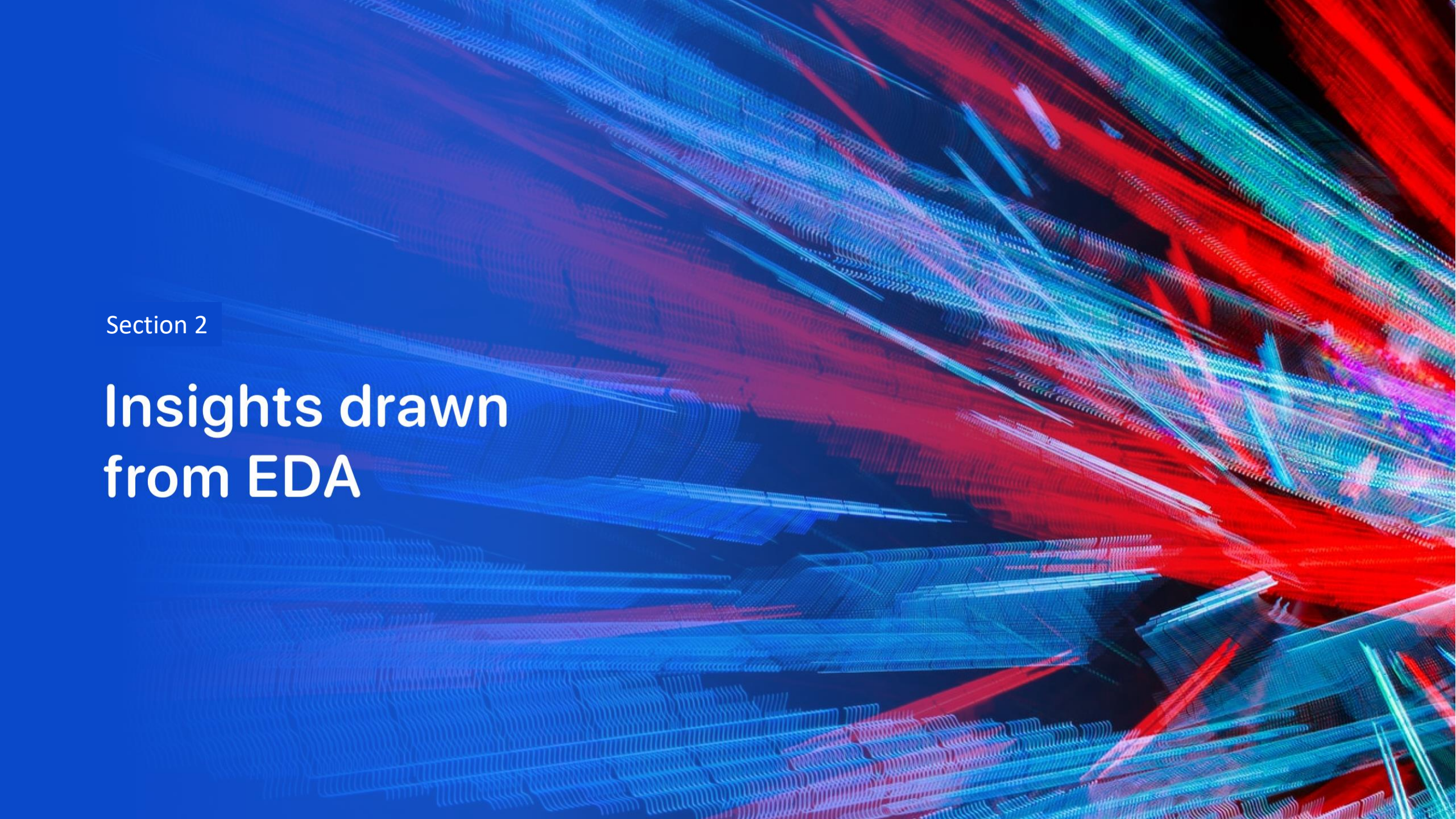
The goal: predict the success of the landing (“Class”)

- https://github.com/jordi-bee/datasciencestudy/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

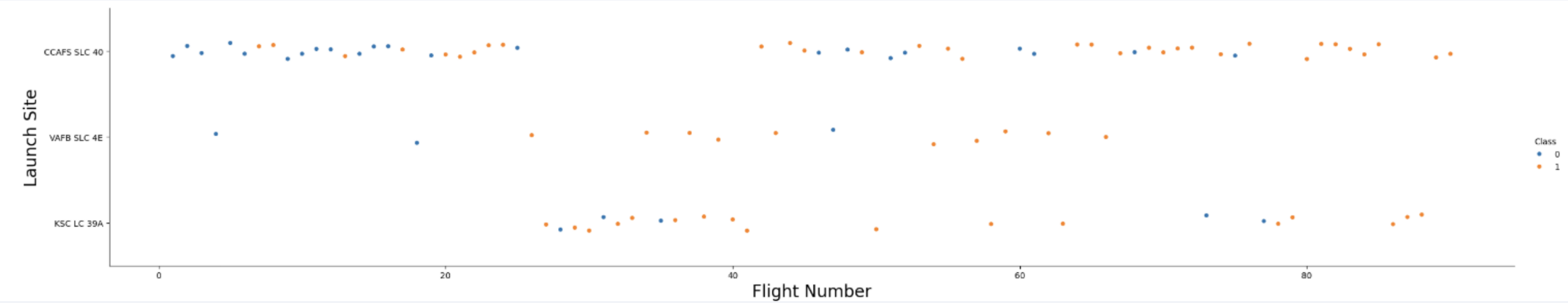


Results

- Exploratory data analysis shows:
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate amount the launch sites
  - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- Visual analytics show:
  - All launch sites are close to the coast line
- Predictive analysis results
  - Prediction is fairly accurate BUT the show a significant number of false positives


Section 2
Insights drawn from EDA

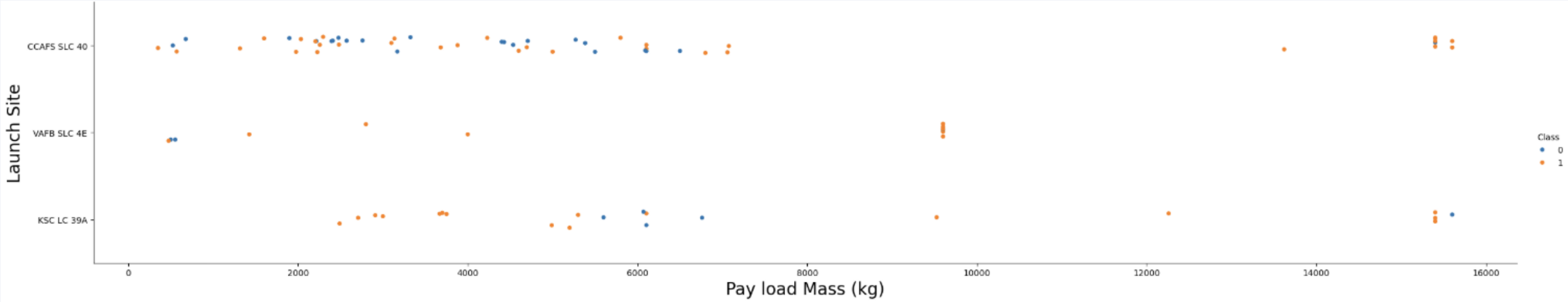
Flight Number vs. Launch Site



At first, the main launch site used was CCAFS SLC 40. The performance was mediocre so after ~ 25 flights they stopped flights and KSC LC 39A was tested. It had a better success rate and after around 15 launches, CCAFS SLC 40 was used again, resulting in more successful landings than before.

VAFB SLC 4E has only launched a few rockets, but with increasing success.

Payload vs. Launch Site



Most of the launches have been conducted with a rather low payload. However, it is visible, that a higher payload results in more successful landings.

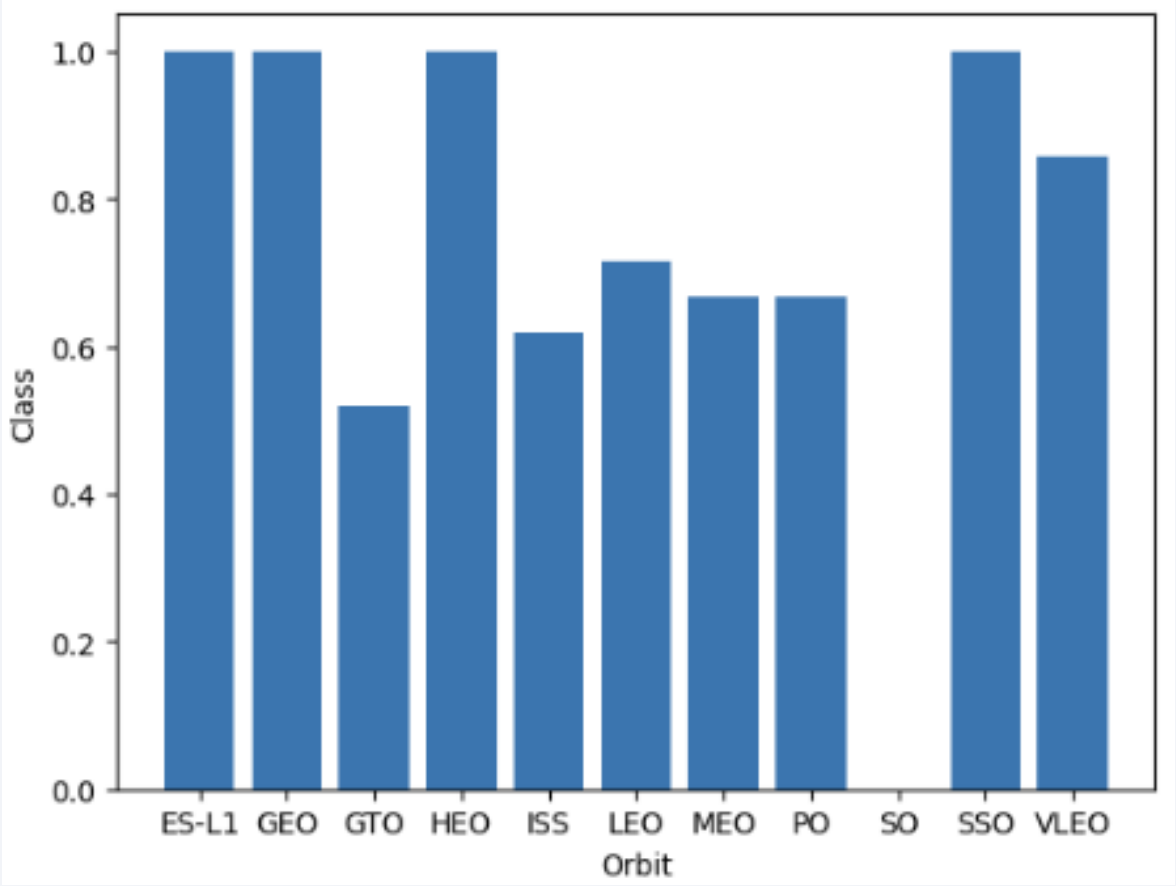
While VAFB SLC 4E seems to be unsuccessful with low payloads, KSC LC 39A performs better with low payload (> 2000 kg and < 5500 kg) and payload above 9000 kg.

CCAFS SLC 40 shows a mediocre performance with low and medium payload and a better performance with high payload.

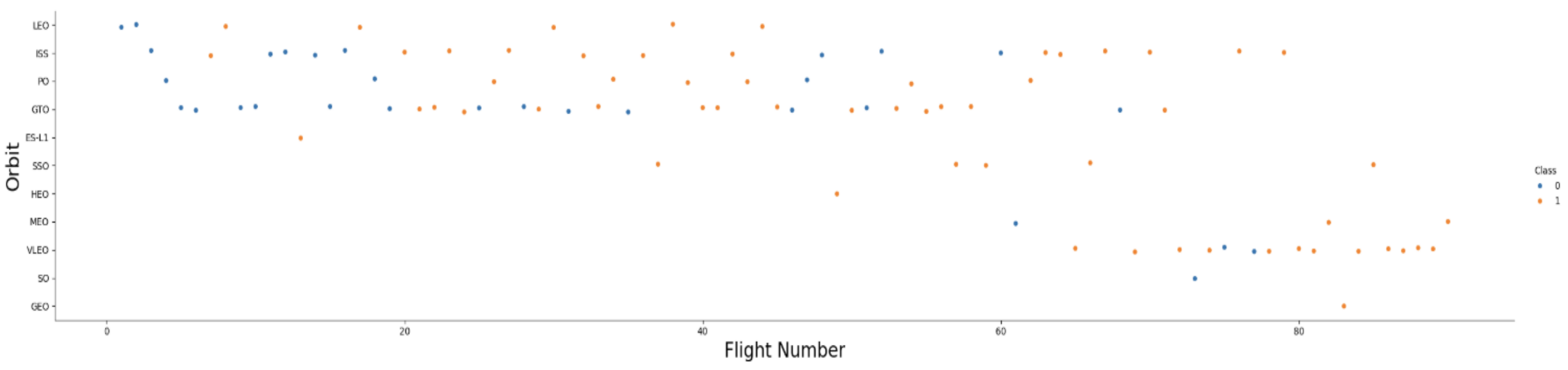
Success Rate vs. Orbit Type

As we can see, rockets that went into the orbits ES-L1, GEO, HEO, and SSO, the success rate is 100 %

Launches to other orbits were not as successful and only show a mediocre performance.



Flight Number vs. Orbit Type

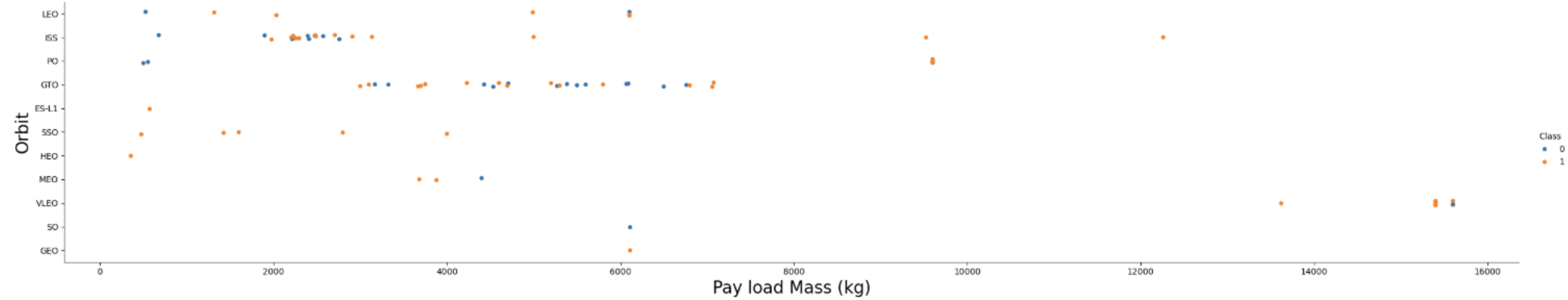


In the LEO orbit, success seems to be related to the number of flights.

Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

As you can see, some of the orbits were only reached in a later stage of test flights, but they show a good success rate.

Payload vs. Orbit Type

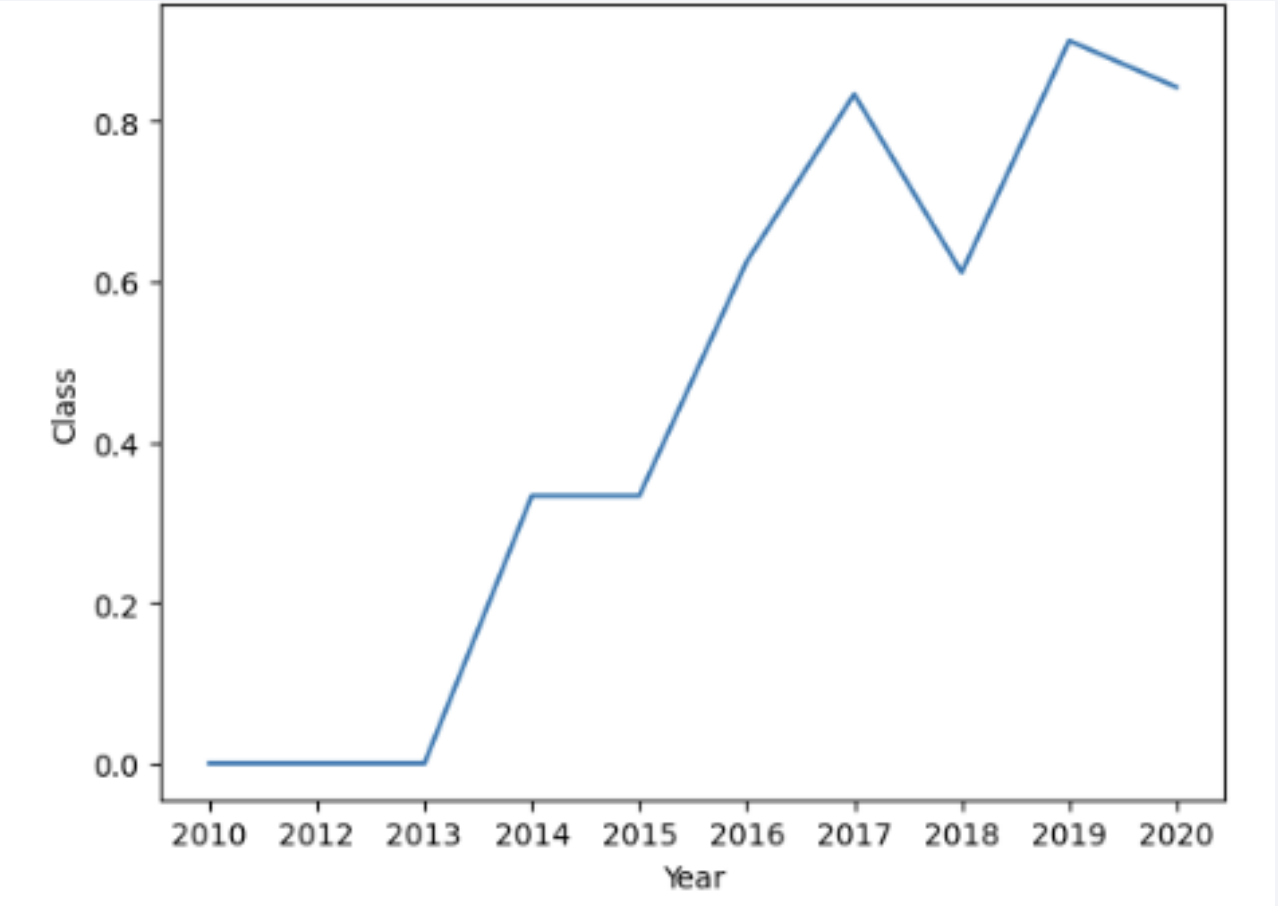


With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

Starting in 2013, the success rate kept increasing, with a little dip in 2018.



All Launch Site Names

```
%sql select DISTINCT "Launch_Site" from SPACEXTABLE
```

We have two main locations for launch sites:

Florida: CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A and

California: VAFB SLC-4E

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where "Launch_Site" like
"CCA%" LIMIT 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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Five records where the launch sites begin with the string “CCA”

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass
from SPACEXTABLE where Customer like "%NASA%"
```

- The total payload mass carried by boosters launched by Nasa (CRS) is: 107,010 kg

total_payload_mass
107010

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as avg_payload from
SPACEXTABLE Where Booster_Version like "%F9 v1.1%"
```

- The average payload mass carried by booster version F9 v1.1 is: 2534.67 kg

avg_payload
2534.6666666666665

First Successful Ground Landing Date

```
%sql select MIN(Date) as first_landing from SPACEXTABLE
WHERE Landing_Outcome like "%ground pad%"
```

- The first successful landing outcome in ground pad was achieved on Dec 22, 2015

first_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version from SPACEXTABLE WHERE
Landing_Outcome="Success (drone ship)" and
PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```

Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(*) from SPACEXTABLE WHERE
Landing_Outcome like "%success%"
```

count(*)
10

- Total number of successful mission outcomes: 10

```
%sql select count(*) from SPACEXTABLE WHERE
Landing_Outcome like "%failure%"
```

count(*)
10

- Total number of failure mission outcomes: 10

Boosters Carried Maximum Payload

```
%sql select Booster_Version from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (select
max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

- Names of the booster which have carried the maximum payload mass:

- F9 B5 B1048.4, F9 B5 B1049.4, F9 B5 B1051.3, F9 B5 B1056.4, F9 B5 B1048.5, F9 B5 B1051.4, F9 B5 B1049.5, F9 B5 B1060.2, F9 B5 B1058.3, F9 B5 B1051.6, F9 B5 B1060.3, F9 B5 B1049.7

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

2015 Launch Records

```
%sql select substr(Date, 6,2) as month, Booster_Version,
Landing_Outcome, Launch_Site from SPACEXTABLE WHERE
Landing_Outcome=="Failure (drone ship)" and
substr(Date,0,5)=='2015’
```

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

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

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

```
%sql select Landing_Outcome, count(Landing_Outcome) as Landing_Count
from SPACEXTABLE WHERE date BETWEEN "2010-06-04" and "2017-03-20"
GROUP BY Landing_Outcome ORDER BY Landing_Count DESC
```

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

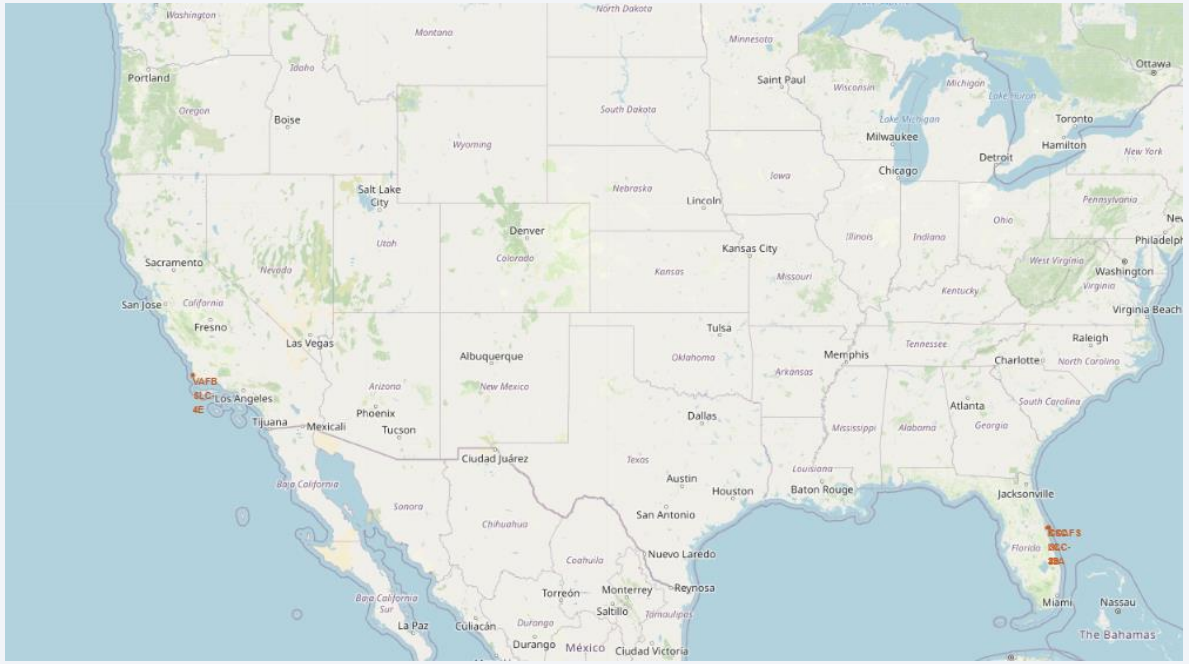
Landing_Outcome	Landing_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Rocket Launch Sites (USA)

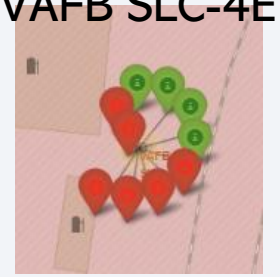
- There are two main locations for launch sites:
  California and Florida
- Both locations are close to the equator and the coast line
- In case a landing should fail, the rocket could crash (or emergency land) in the water



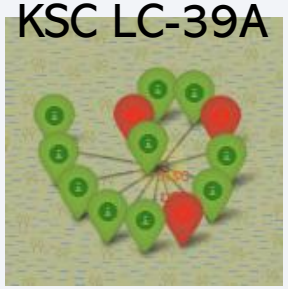
Successful and failed launches

- 10 launches at launch site VAFB SLC-4E:
  - 4 successful (green)
  - 6 failed (red)
- 13 launches at launch site KSC LC-39A:
  - 10 successful (green)
  - 3 failed (red)
- 26 launches at launch site CCAFS LC-40:
  - 7 successful (green)
  - 19 failed (red)
- 7 launches at launch site CCAFS SLC-40:
  - 3 successful (green)
  - 4 failed (red)

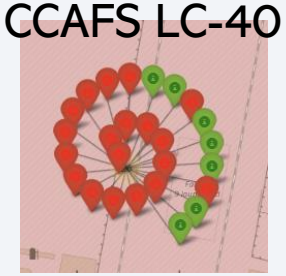
VAFB SLC-4E



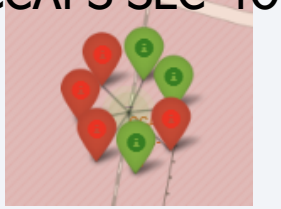
KSC LC-39A



CCAFS LC-40

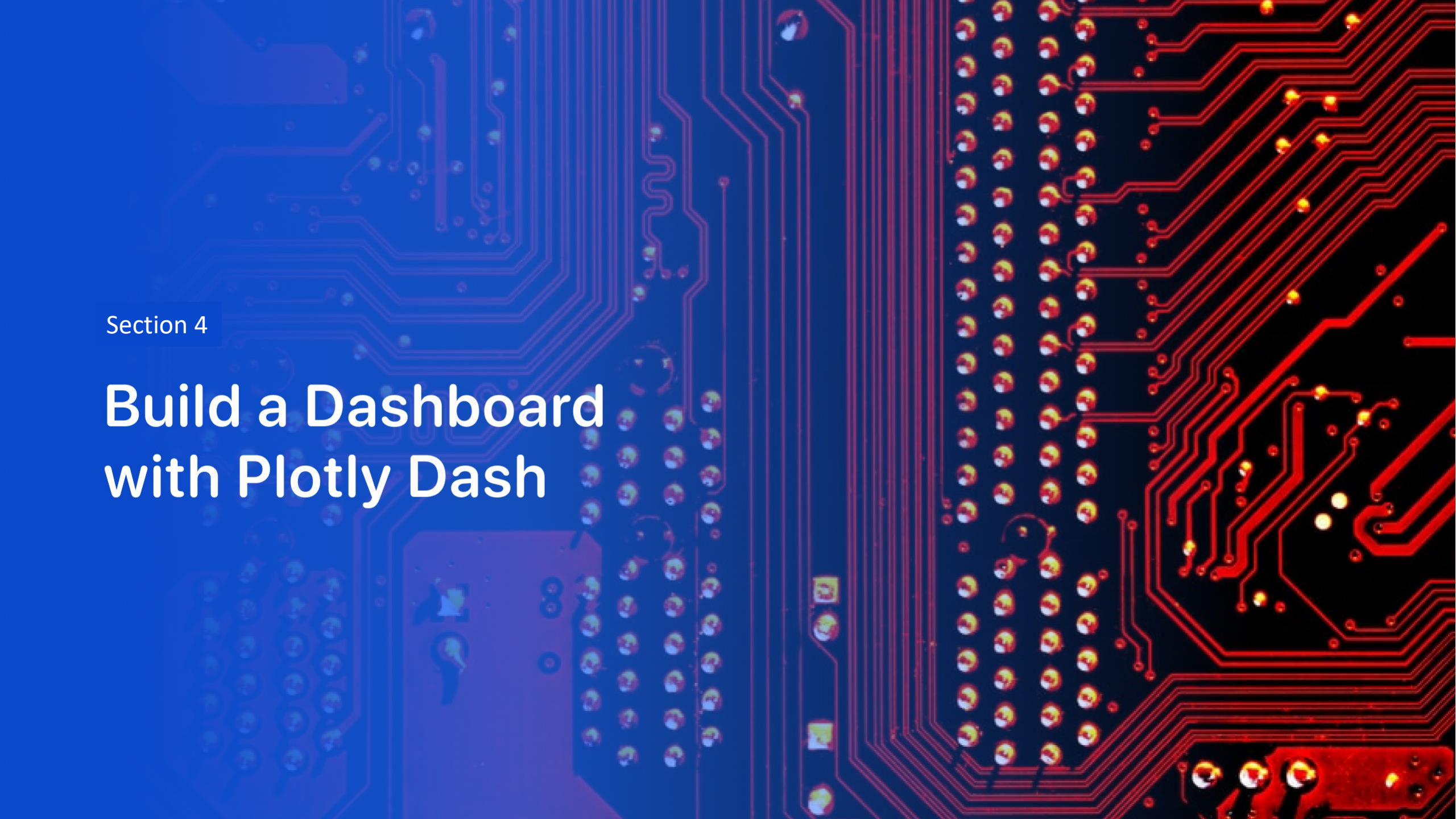


CCAFS SLC-40

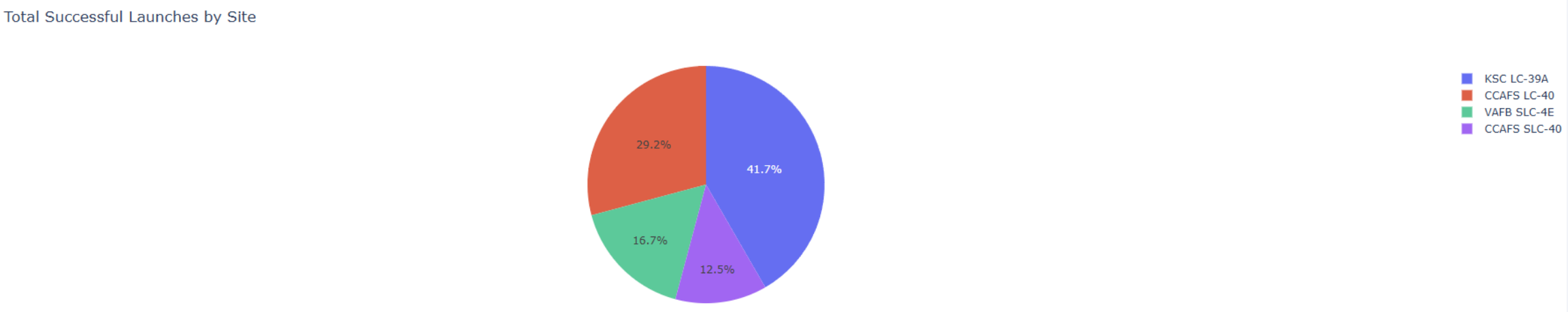


Launch site distance to coastline

Launch site VAFB SLC-4E is very close to the coastline,
with a distance of only 1.36 km from launch site to the coast

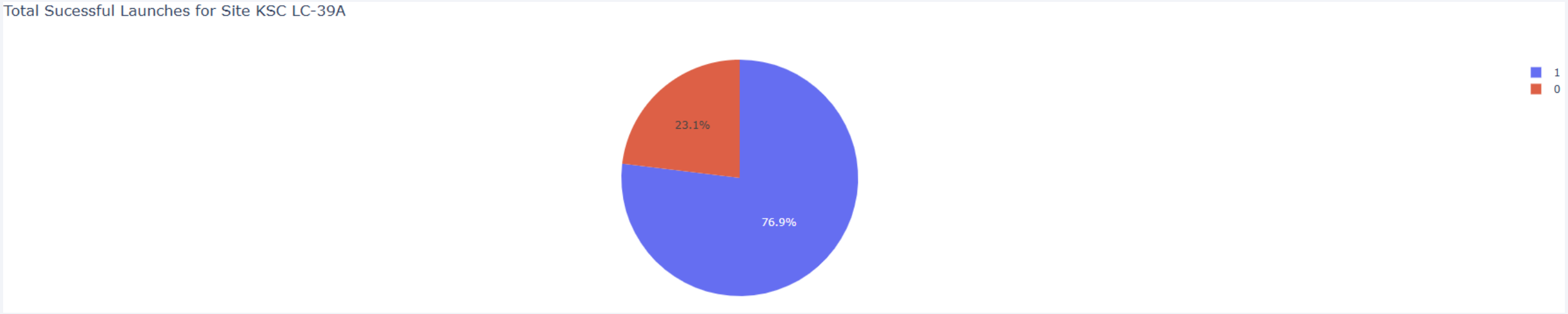

Section 4
Build a Dashboard with Plotly Dash

Total Launch Success for all sites



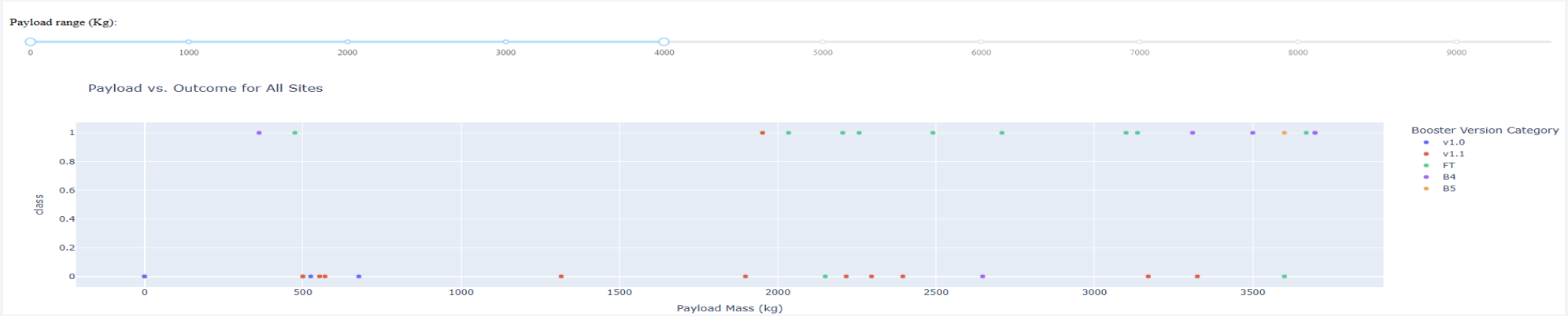
- KSC LC-39A is the launch site with the highest success rate, followed by CCAFS LC-40.
- VAFB SLC-4E and CCAFS SLC-40 only had very little success with only 4, respectively 3 successful launches

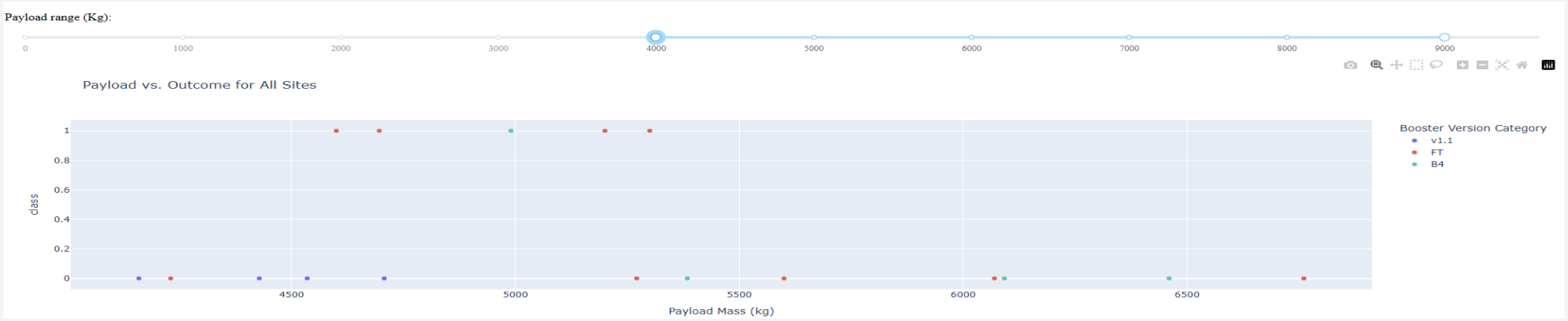
Launch Success Rate at KSC LC-39A



- KCS LC-39A has a very high success rate:
  - 76.9% of all launches at the site have been successful
  - 23.1% of the launches have failed

Payload and the effect on the Outcome



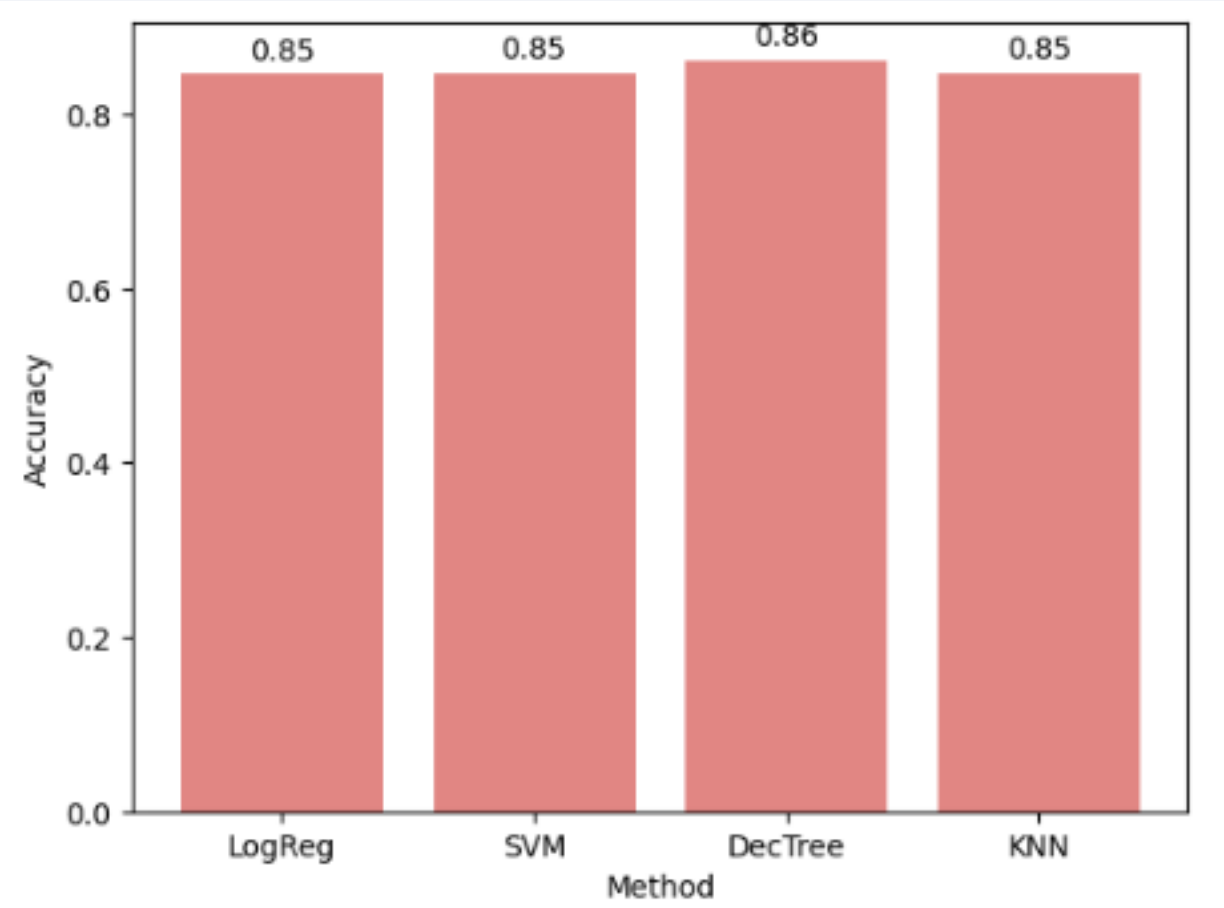


- Launches with lower payload have a higher success rate than those with high payload

Section 5

Predictive Analysis (Classification)

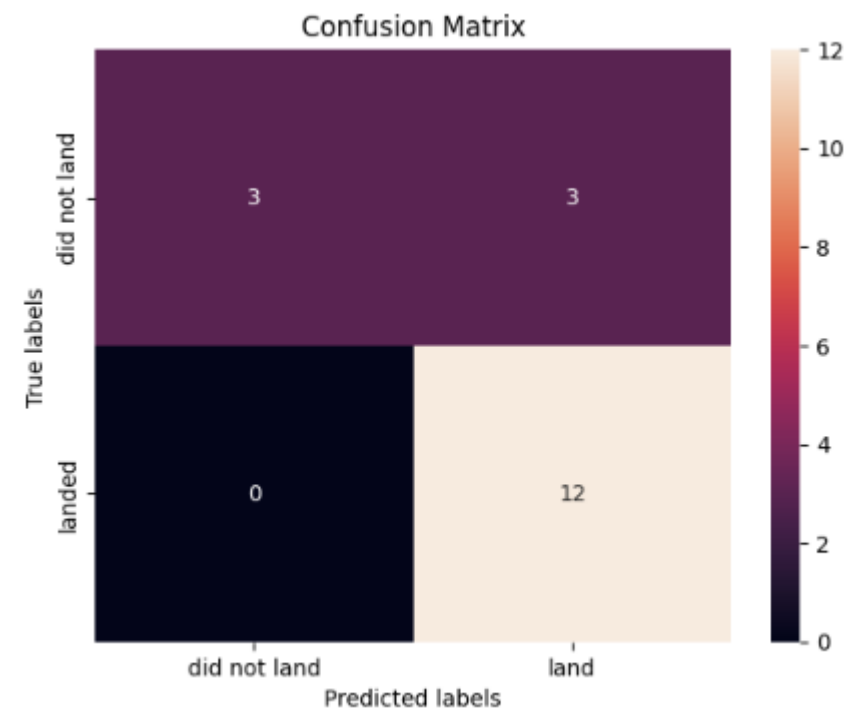
Classification Accuracy



With the right hyperparameters, Decision Tree Classification is the method with the highest accuracy.

However, all methods have the same accuracy score on the test data. (Which is probably due to a very small set of test data.)

Confusion Matrix



As previously stated, the accuracy score on the test data is the same with all methods.

Therefore the confusion matric is also the same for all methods.

The methods are very reliable for prediction negative outcomes correctly. However, there is a problem with the false positives, that could lead to severe issues.

Conclusions

- Over the years, there has been a significant improve in the launches.
- Features that show to have been correlated to a successful outcome are:
  - lower payloads
  - launch site KSC LC 39A
  - Orbits GEO, HEO, SSO, ES L1
- The accuracy of the models is very similar. Even though Decision Tree Classification seems to be the most accurate of the methods, the difference is so small that I would recommend to evaluate the models even further. I would suggest to use cross validation and/or a larger test set. As soon as there is more data from new launches, I would recommend re-evaluation all models with the bigger data set.

Appendix

- You can see all my work in my Github repository:

  https://github.com/jordi-bee/datasciencestudy

Thank you!