IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rafael Jiménez Vicente
April 23

Outline

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data wrangling
  - EDA
  - Maps with Folium
  - Dashboard with plotly dash
  - Machine Learning
- Summary of all results
- The four machine learning models used generated similar results with an accuracy rate of 83,33%. These models tend to overpredict successful landings. So if the models would be feed with more training data could improve the accuracy.

Introduction

- Project background and context
  - The competition to win the commercial space race is getting tougher.
  - In this context SpaceX advertsise Falcon 9 rocket launches with a cost of 62 millions dolars much cheaper tan other competitors.
  - This is due to its ability to reuse the first stage.
  - A new Company called SpaceY wants to compete with SpaceX.
  - So determine if the first stage will land can determinate the cost of the launch and this information its crucial for SpaceY in order to compete with SpaceX.
- Problems you want to find answers
  - The first stage landing success likelihood.
  - Which factors influence the most in determining this likelihood.
  - Can these factors be isolated and replicated in optimal conditions to increase the landing success in SpaceY?.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - A combination of two data sources is used within this process. On the one hand through the SpaceX API and on the other hand using web scraping methodology from “List of Falcon 9 and Falcon Heavy launches” Wikipedia page.
- Perform data wrangling
  - Trough cleaning data, the use of the method value_counts() to determine occurrences and expressing landing success in a binary result.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models.

Data Collection

- This section includes:

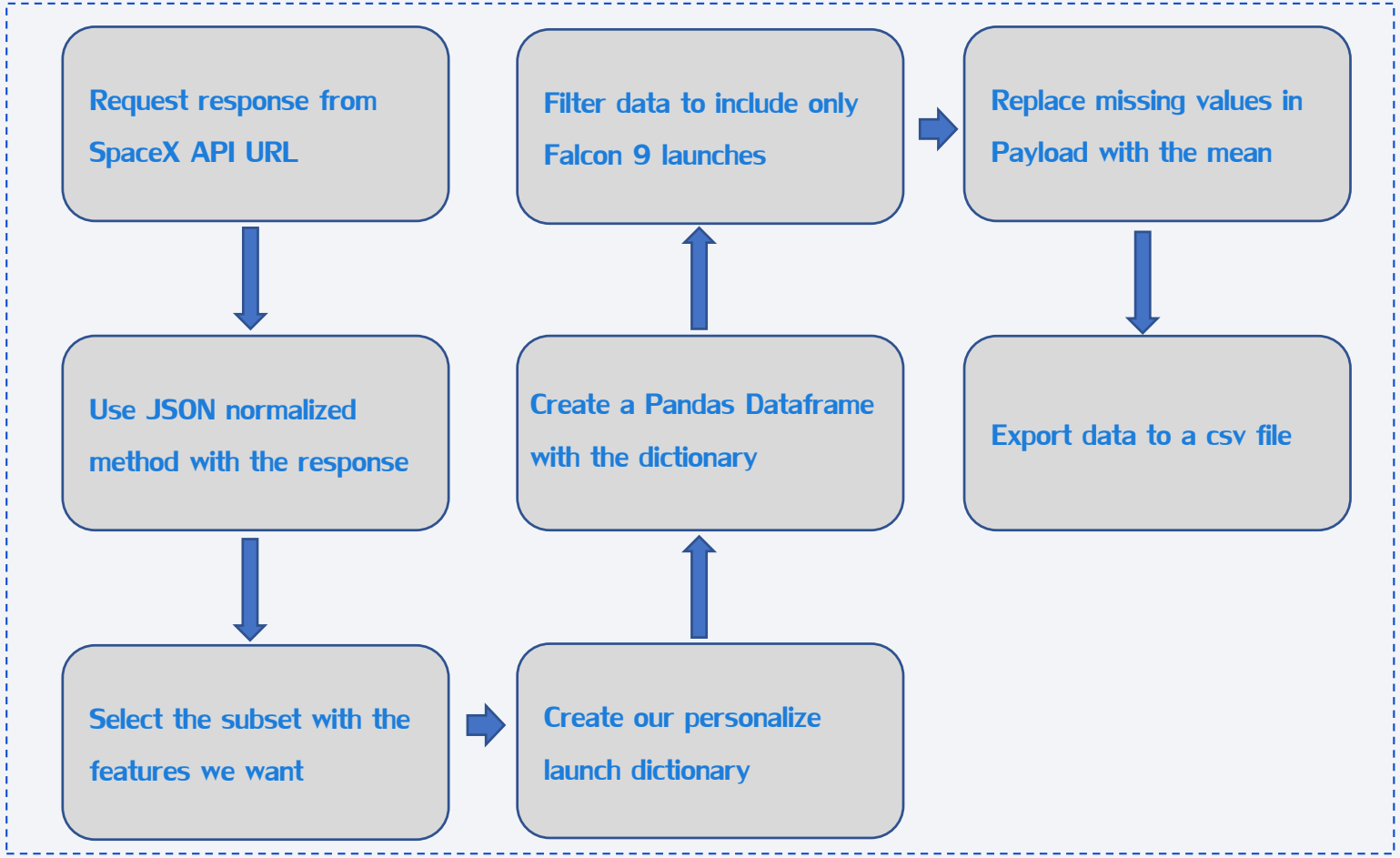
  - Data Collection- SpaceX API. Flowchart (Data Source- SpaceX Rest API).

  - Data Collection- Scraping. Flowchart (Data Source- “List of Falcon 9 and Falcon Heavy launches” Wikipedia).

Data Collection – SpaceX API

- GitHub URL:

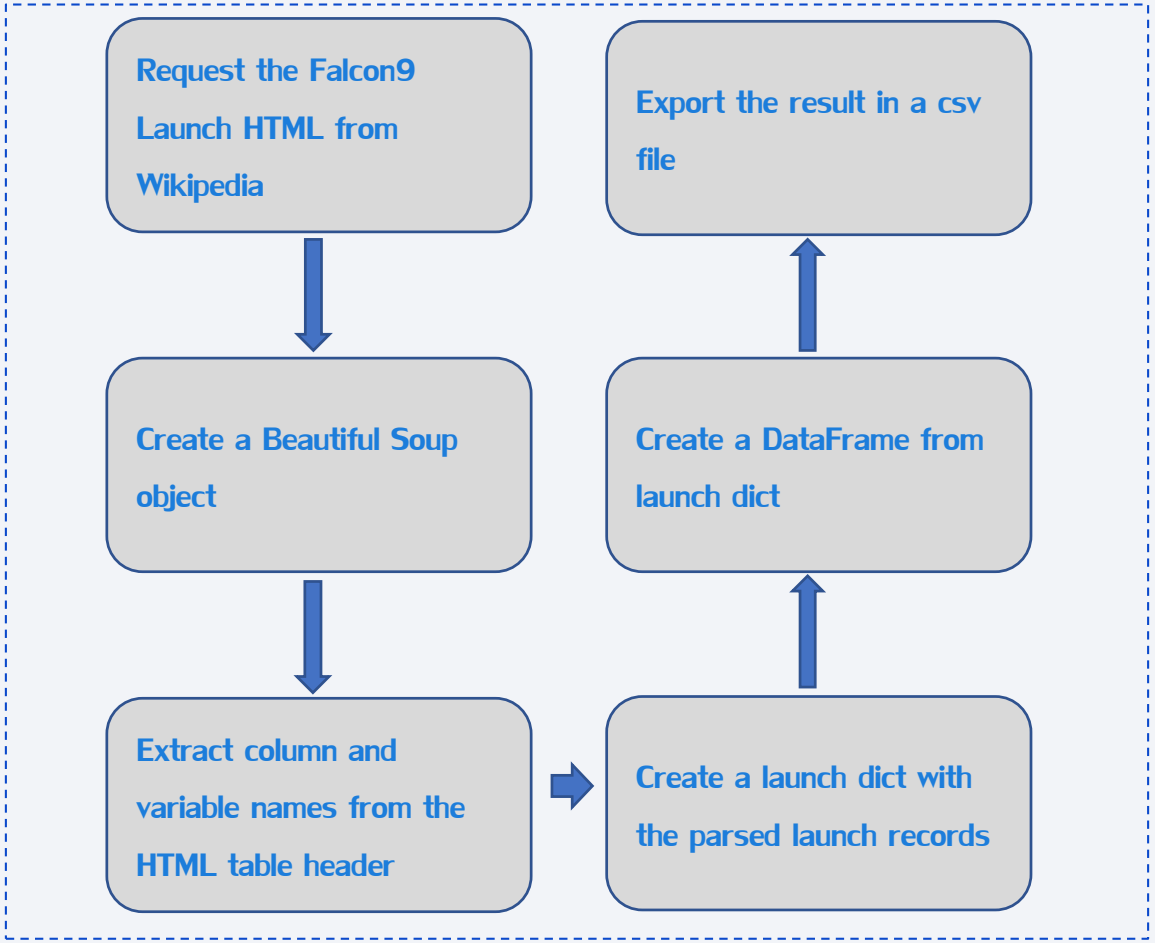
https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/1-jupyter-labs-spacex-data-collection.ipynb



Data Collection - Scraping

- GitHub URL:

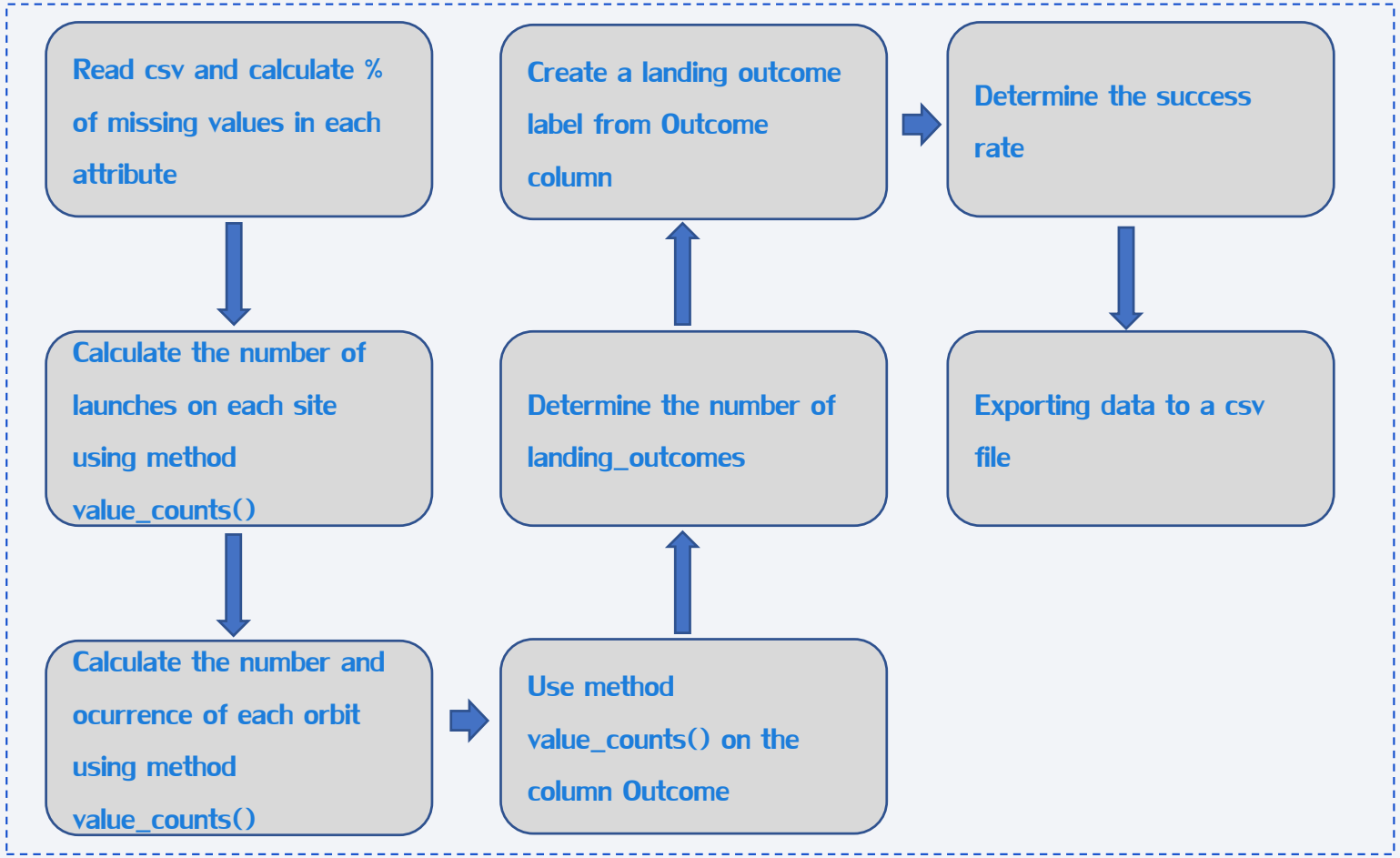
https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/2-jupyter-labs-webscraping.ipynb



Data Wrangling

- GitHub URL:

https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/3-jupyter-labs-spacex-data_wrangling.ipynb



EDA with Data Visualization

- Charts Plotted:
  - **Scatter Charts:** Flight number vs Payload Mass, Flight number vs Launch Site, Payload vs Launch Site, Orbit vs Flight Number, Payload vs Orbit Type, Orbit vs Payload Mass.
  - **Bar Charts:** Mean vs Orbit.
  - **Line Charts:** Success Rate vs Year.
- These charts were used to compare relationships between variables to determine if a relationship exists or not and if they could be used in ML Model.
- GitHub URL:

https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/5-jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Queries Performed:

- Displaying the names of unique launch sites in space mission
- Displaying 5 records where launch sites' names start with `KSC`
- Displaying the total payload mass carried by boosters launched NASA (CRS)
- Displaying the average payload mass carried by booster version F9 v1.1
- Displaying the date of the first successful landing outcome on drone ship
- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Displaying the total number of successful and failure mission outcomes
- Listing the names of the booster which have carried the maximum payload mass
- Listing the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

- https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/4-jupyter-labs-eda-sql.ipynb

Build an Interactive Map with Folium

- Folium maps visualize the launch data onto an interactive map. Using latitude and longitude coordinates of each launch site, we added labelled circle markers in each one of them. Using MarkerCluster(), we indicate successful outcomes with green markers, and red markers in unsuccessful. We also calculate the distance to key locations on the map and mark a line to visualize them (launch sites proximities);
- Distance to nearest railway
- Distance to nearest highway
- Distance to the nearest point of the coast
- Distance to nearest city
- GitHub URL:

https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/6-jupyter-labs_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot is used to visualize how success varies dependent on payload mass and booster version category.

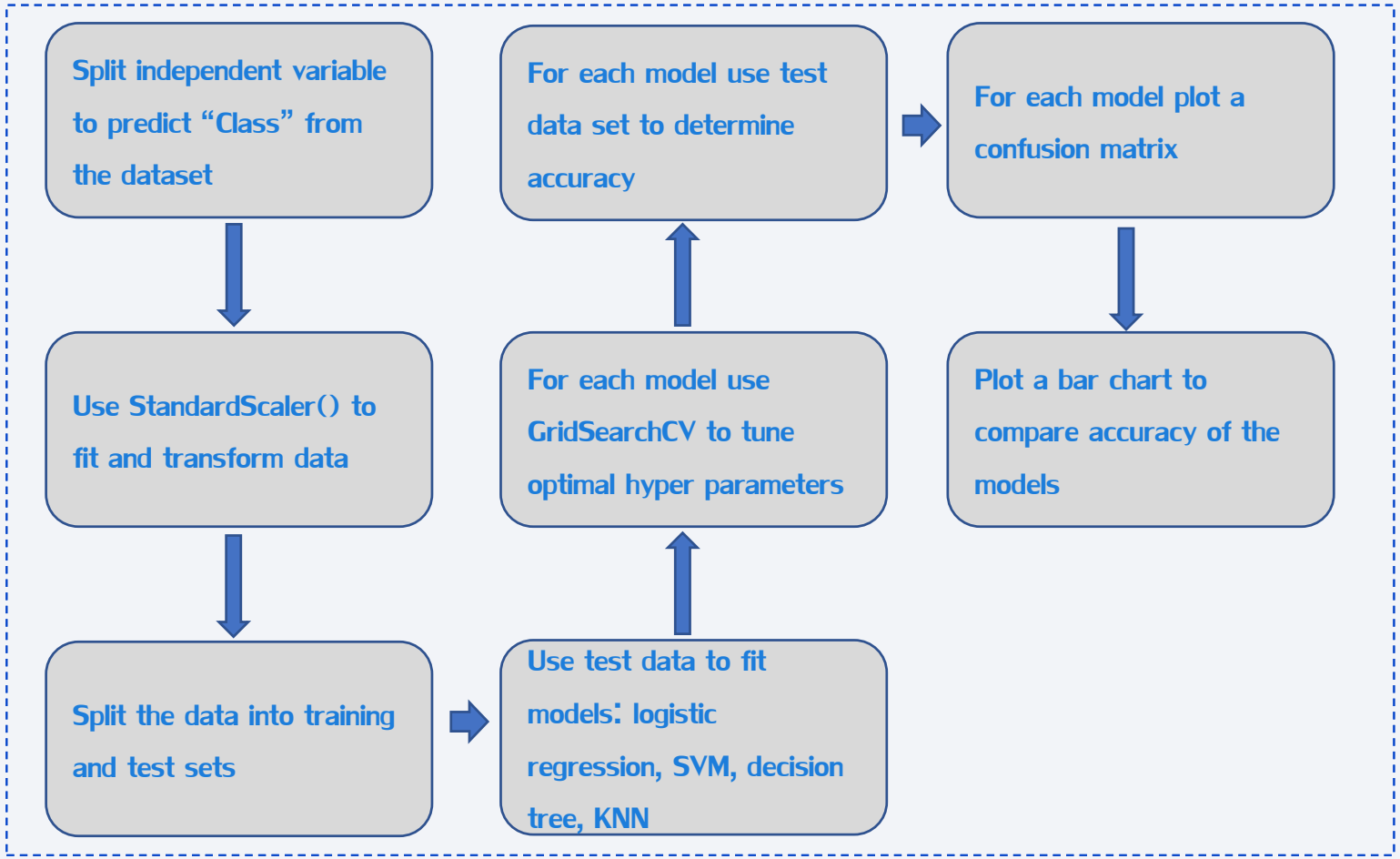
- GitHub URL:

https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/6-plotly_dash_app.py

Predictive Analysis (Classification)

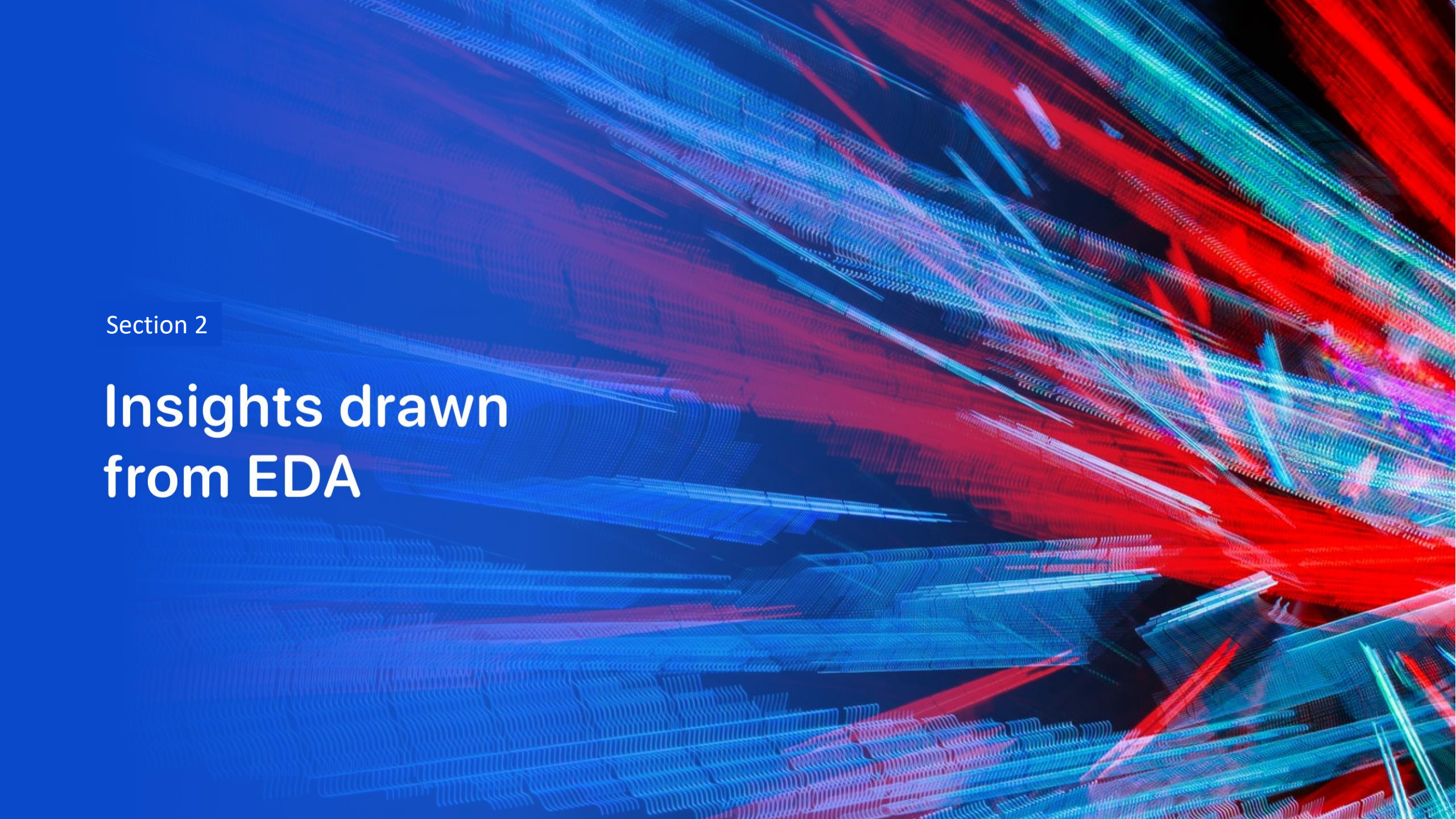
- GitHub URL:

https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/7-jupyter-labs_SpaceX_Machine_Learning_Prediction.ipynb

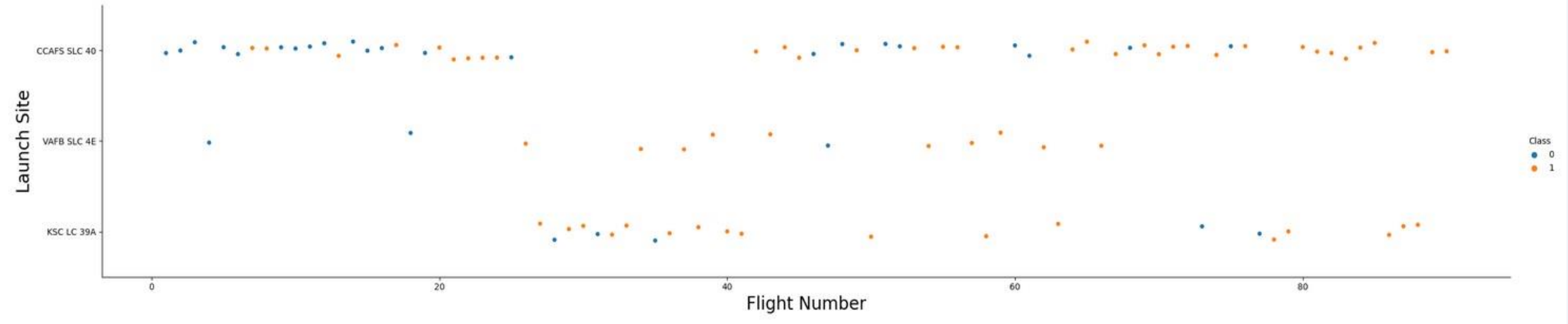


Results

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

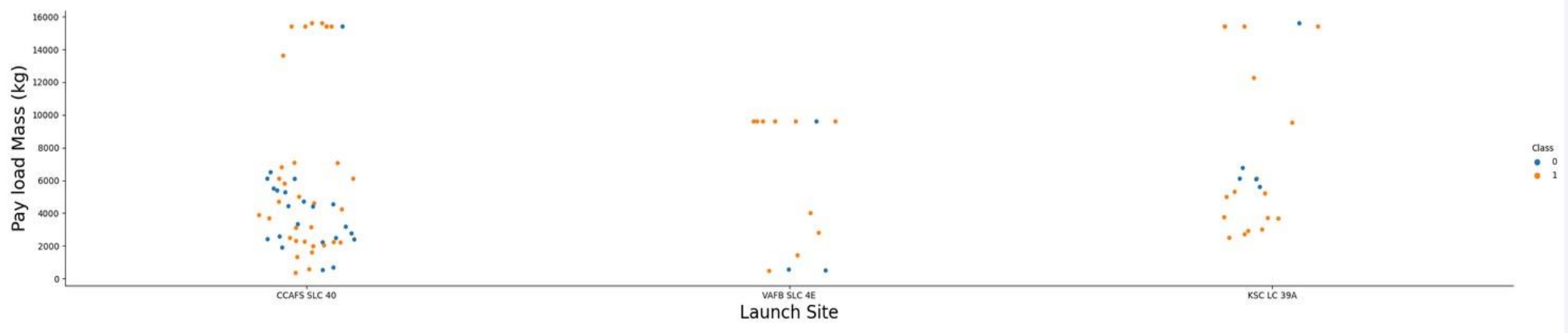

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



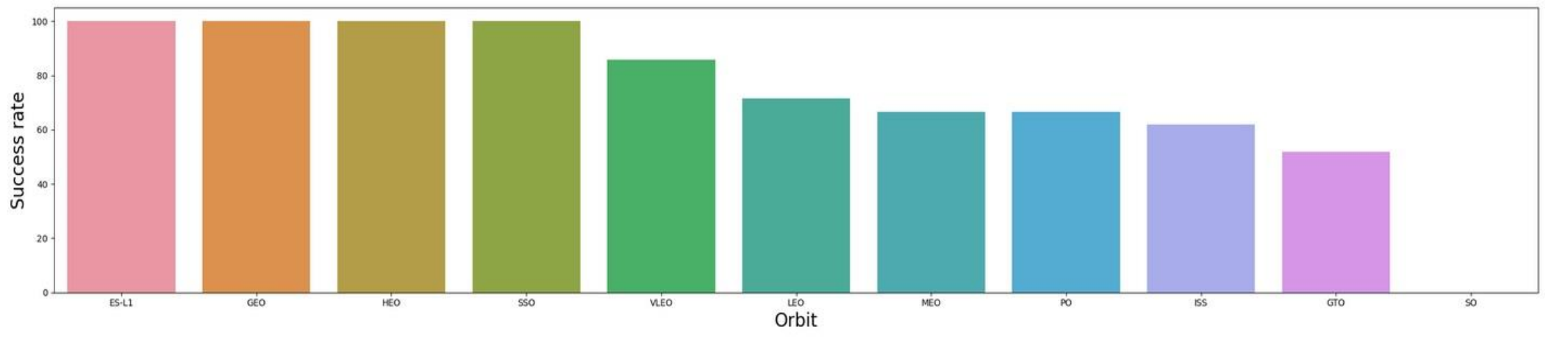
- Orange circles indicate successful launch and blue circles indicate unsuccessful ones. As the number of flights has been increasing, the result has been more satisfactory, increasing the number of successful launches. We also can see that the launch site labeled as “CCAFS SLC 40” has been where the greatest number of launches have been made.

Payload vs. Launch Site



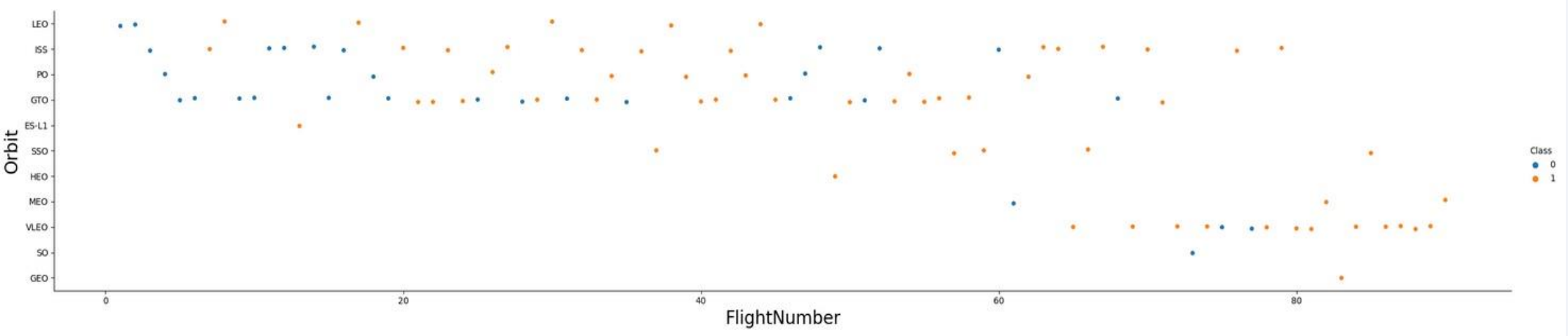
- Orange circles indicate successful launch and blue circles indicate unsuccessful ones. when the payload is at 10000 kg or above it seems that the number of successful launches becomes more evident than with lower masses. We have the exception of the launch site labeled as where with masses between 2000 and 6000 kg the launches were successful.

Success Rate vs. Orbit Type



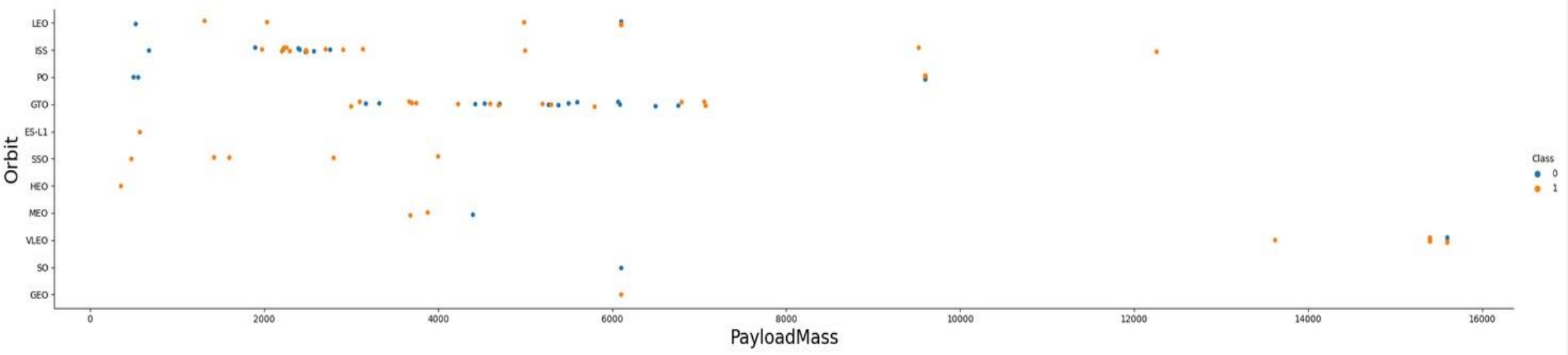
- ES-L1, GEO, HEO and SSO orbits have 100% of successful launch rate. On the other hand we can see that SO orbit has a 0% of successful rate. Also note that the largest sample of launches, and therefore its significance is concentrated in the orbits GTO, ISS and VLEO.

Flight Number vs. Orbit Type



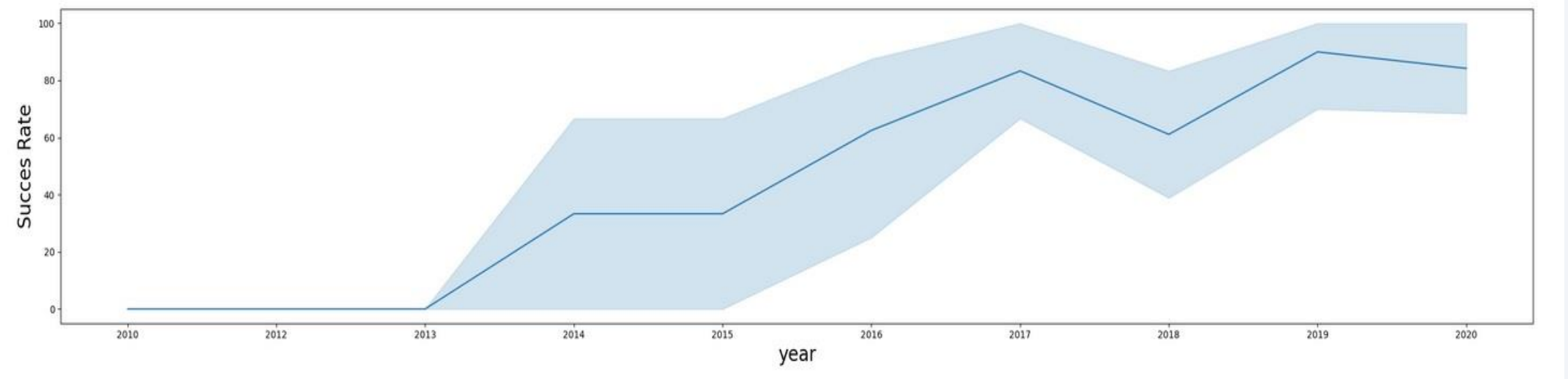
- Orange circles indicate successful launch and blue circles indicate unsuccessful ones. As the number of flights has increased, the preference of the orbits has changed. Going on to focus mainly on VLEO type orbits which seems to yield successful results more consistently.

Payload vs. Orbit Type



- Orange circles indicate successful launch and blue circles indicate unsuccessful ones. It seems that in SSO and LEO type orbits when payload mass below 6000 kg have been used launches have been mostly successful. Similarly, in orbits ISS and VLEO, when masses of over 12,000 kg have been used successful launches have also been obtained. While GTO orbit there seems to be no clear correlation between these variables.

Launch Success Yearly Trend



- We can see that success rate has generally increased from 2013 to 2020, with slight decrease in 2018, and the highest success rate was observed in 2019. **Success rate in recent years is located around 80%.**

All Launch Site Names

```
%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

- This query returns unique launch sites from database. We have really 3 unique launch sites (CCAFS SLC-40, VAFB SLC-4E, and KSC LC-39A), because CCAFS LC-40 and CCAFS SLC-40 are the same site.

Launch Site Names Begin with 'KSC'

```
%sql select * from SPACEXTBL where launch_site like 'KSC%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

`

- Return the first 5 records in database where launch site name begins with "KSC".

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

sum
45596

- This query returns the sum of payload mass (KG) where NASA (CRS) was the customer.

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

Average
2534.6666666666665

- Returns the average payload mass (kg) of launches where the booster version name begins with “F9 v1.1”

First Successful Drone Ship Date

```
%sql select min((substr(Date, 7, 4) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2))) as first_succes from SPACEXTBL where "Landing _Outcome"= 'Success (drone ship)'
```

```
 * sqlite:///my_data1.db
Done.
```

first_succes
2016-04-08

Returns the first successful drone pad landing date. Since the SQlite database manager does not have min (date) the date format must be adapted to achieve the desired result.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTBL where (payload_mass__kg_ between 4000 and 6000) and ("landing _outcome" = 'Success (drone ship)')
```

```
 * sqlite:///my_data1.db
Done.
```

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

- This query returns the four Booster versions that had successful drone ship landings and payload mass between 4000 and 6000 (kg).

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL group by mission_outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

- This query returns the count of the mission outcomes. We can see that most of the missions are successful.

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

- This query returns the names of booster versions which have carried the maximum payload mass.

2017 Launch Records

```
%sql select substr(Date, 4, 2) as Month, "Landing _Outcome", booster_version, launch_site from SPACEXTBL where substr(Date,7,4)='2017' and "landing _outcome" = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Landing _Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- This query returns the list of records (month, landing outcome, booster version, launch site) with successful landing outcomes in ground pad for 2017. The data format must be adapted again.

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

```
%sql select "Landing _Outcome", count(*) as count from SPACEXTBL where (substr(Date, 7, 4) || '-' || substr(Date, 4, 2) || '-' || substr(Date, 1, 2))
between '2010-06-04' and '2017-03-20' and ("Landing _Outcome" like "Success%") group by "Landing _Outcome" order by count Desc
```

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	count
Success (drone ship)	5
Success (ground pad)	3

- This query returns the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order. The data format must be adapted again.

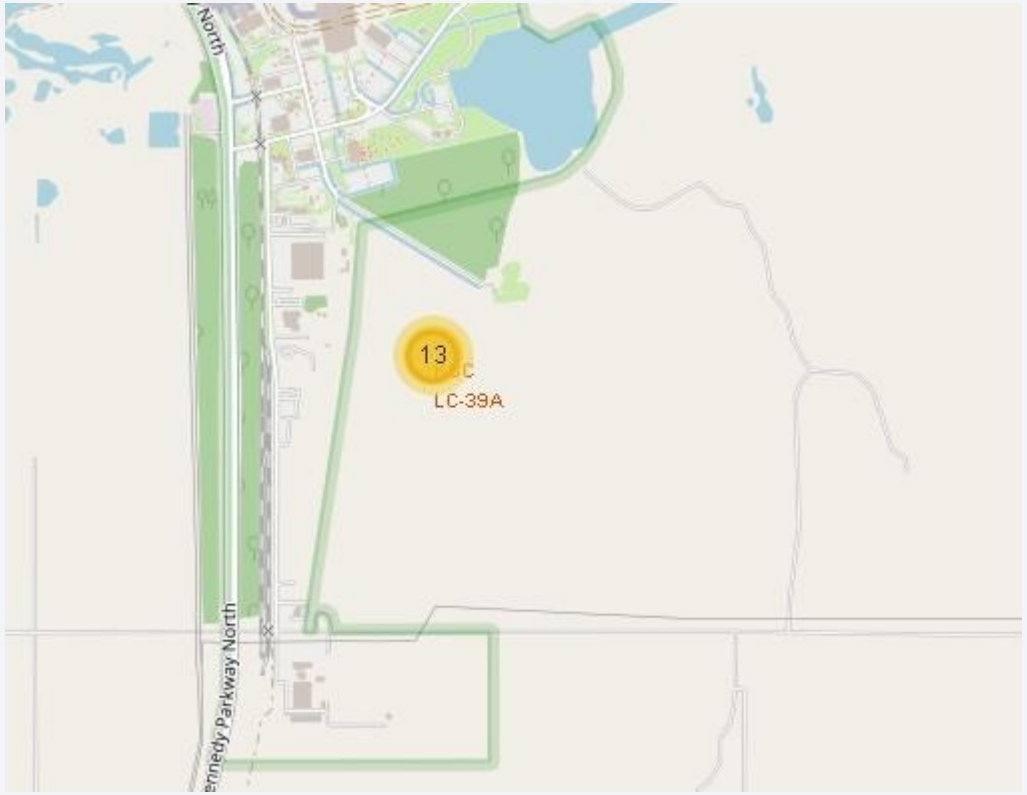

Section 3
Launch Sites
Proximities Analysis

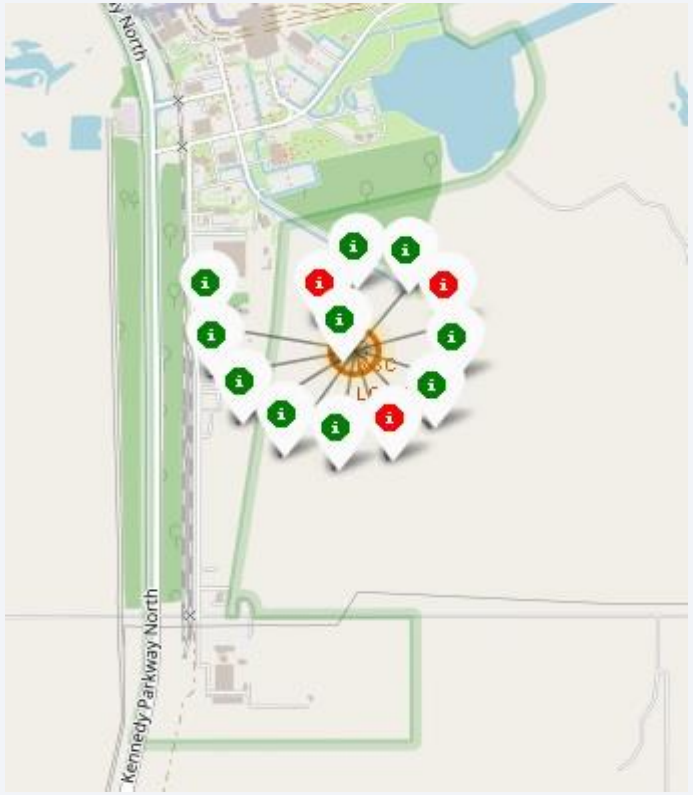
Launch Site Locations

- We can see that the launches are located both

on the east coast and on the west coast of the United States, specifically on the coasts of California and Florida.

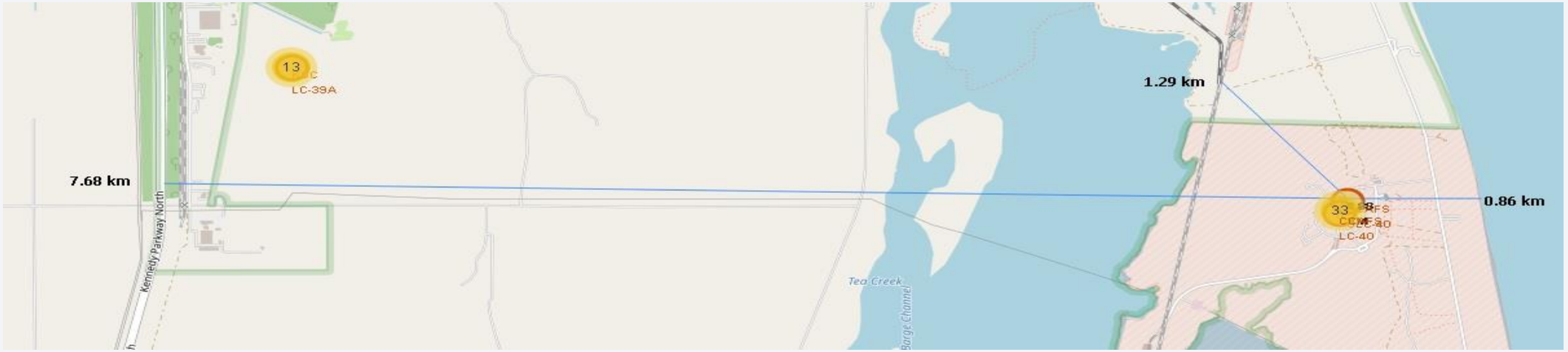
Launch Outcomes

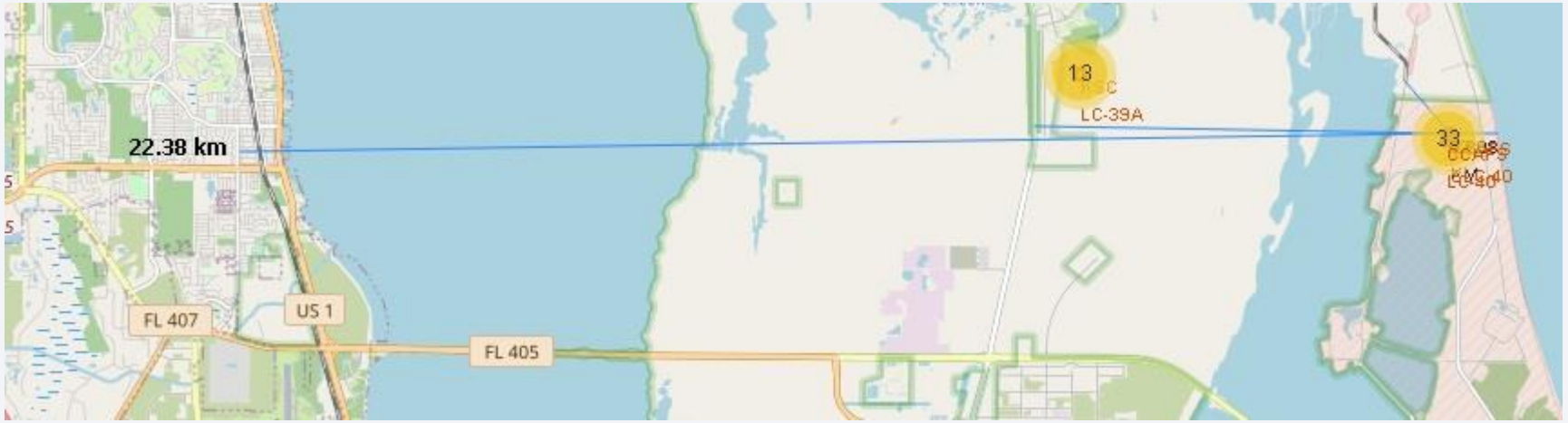




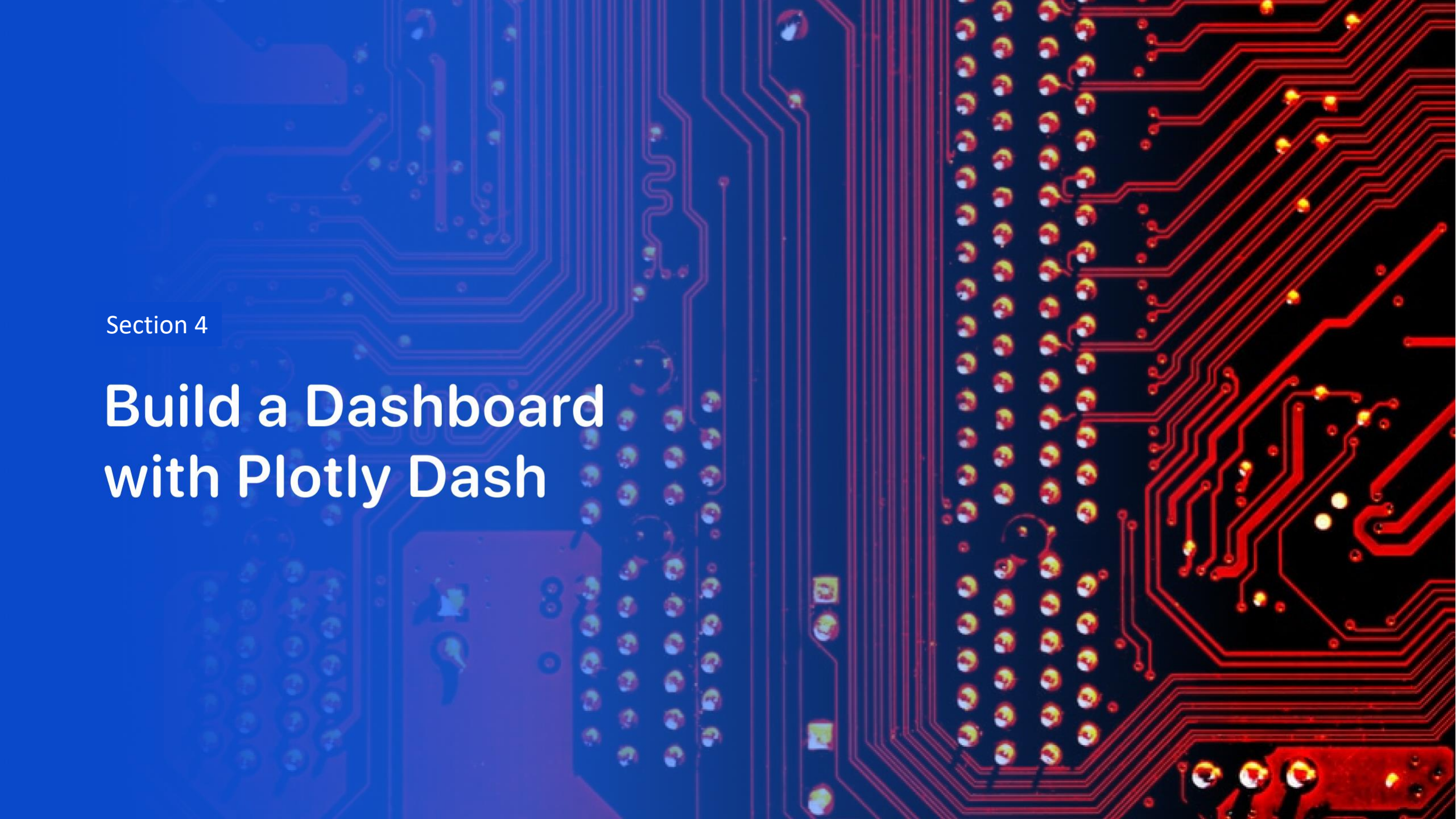
- We can see how there are 13 launches clustered in the location KSC LC 39A. Subsequently, ungrouping this location, we observe the launches in more detail, seeing in this case that 3 landings were unsuccessful while the rest were successful.

Launch Site Proximities

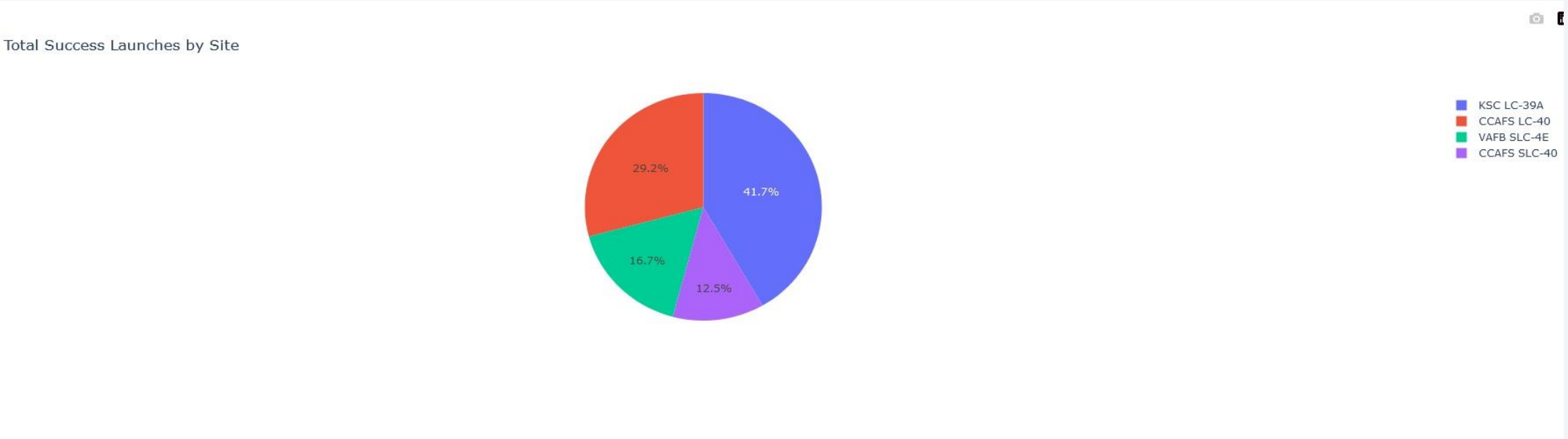




- In these images we analyze the proximity of the launch site CCAF SLC-40 to key locations. In this way we observe that there is a distance of 0.86 km with the coastline, 1.29 km with the railway line, 7.68 with the highway and 22 .38 km to the nearest city.

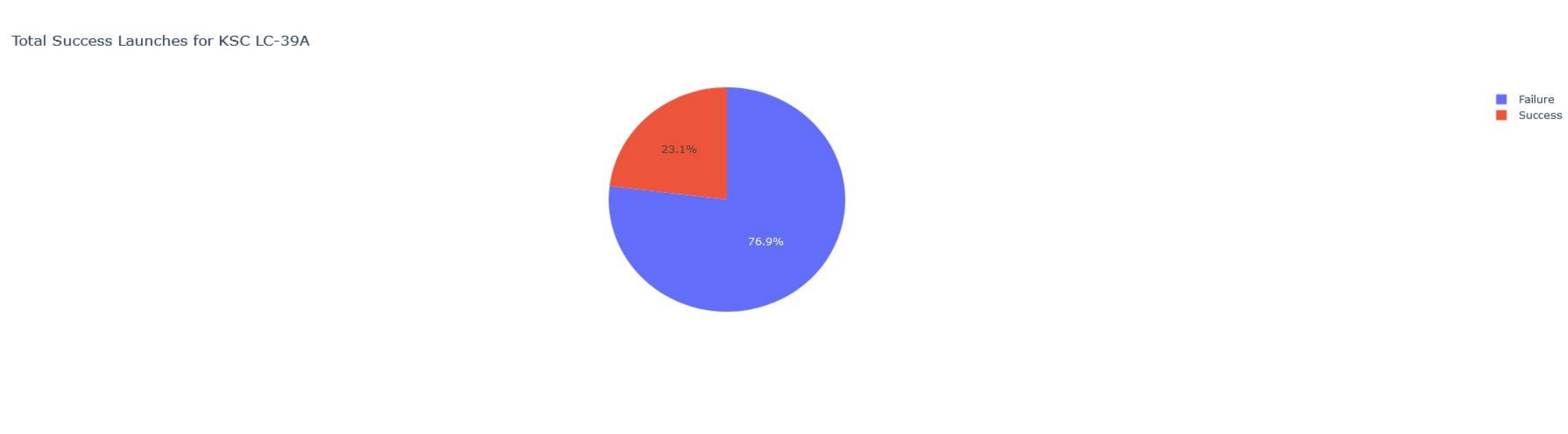

Section 4
Build a Dashboard with Plotly Dash

Percentage of successful launches by location



- In this image we can see how most of the successful launches are located at the KSC LC-39A location. Positioning the Florida coast as the best place for launches.

Percentage of successful landings on KSC LC-39A



- Carrying out a more in-depth analysis of the KSC LC-39A launch site, we can see how the success rate is relatively small, representing only 23.1% of successful landings.

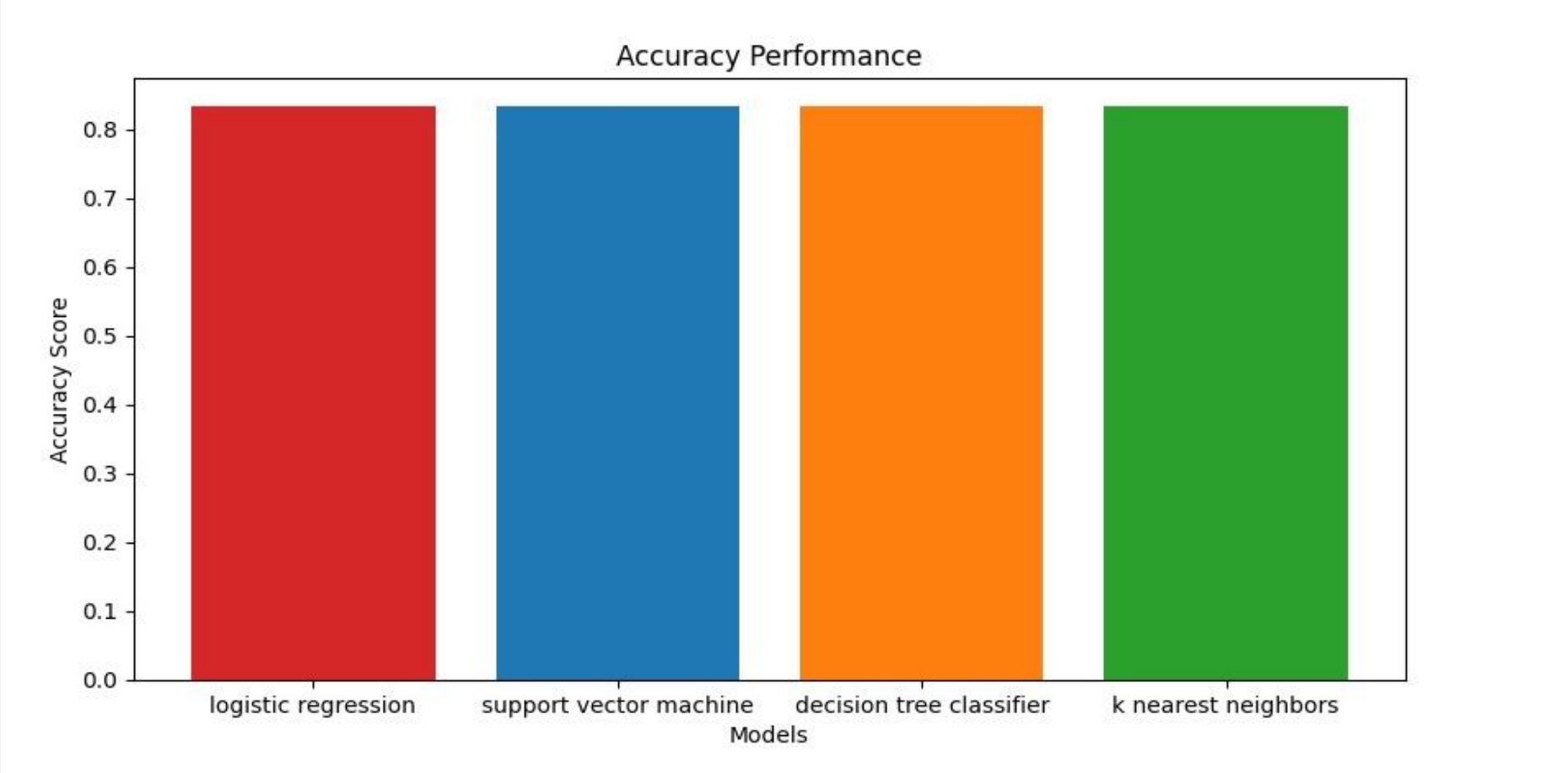
Payload vs. Launch Outcome scatter plot

- We can see how the highest success rates are concentrated in booster version FT. While the most successful payload range is between 2500 kg and 5000 kg.

Section 5

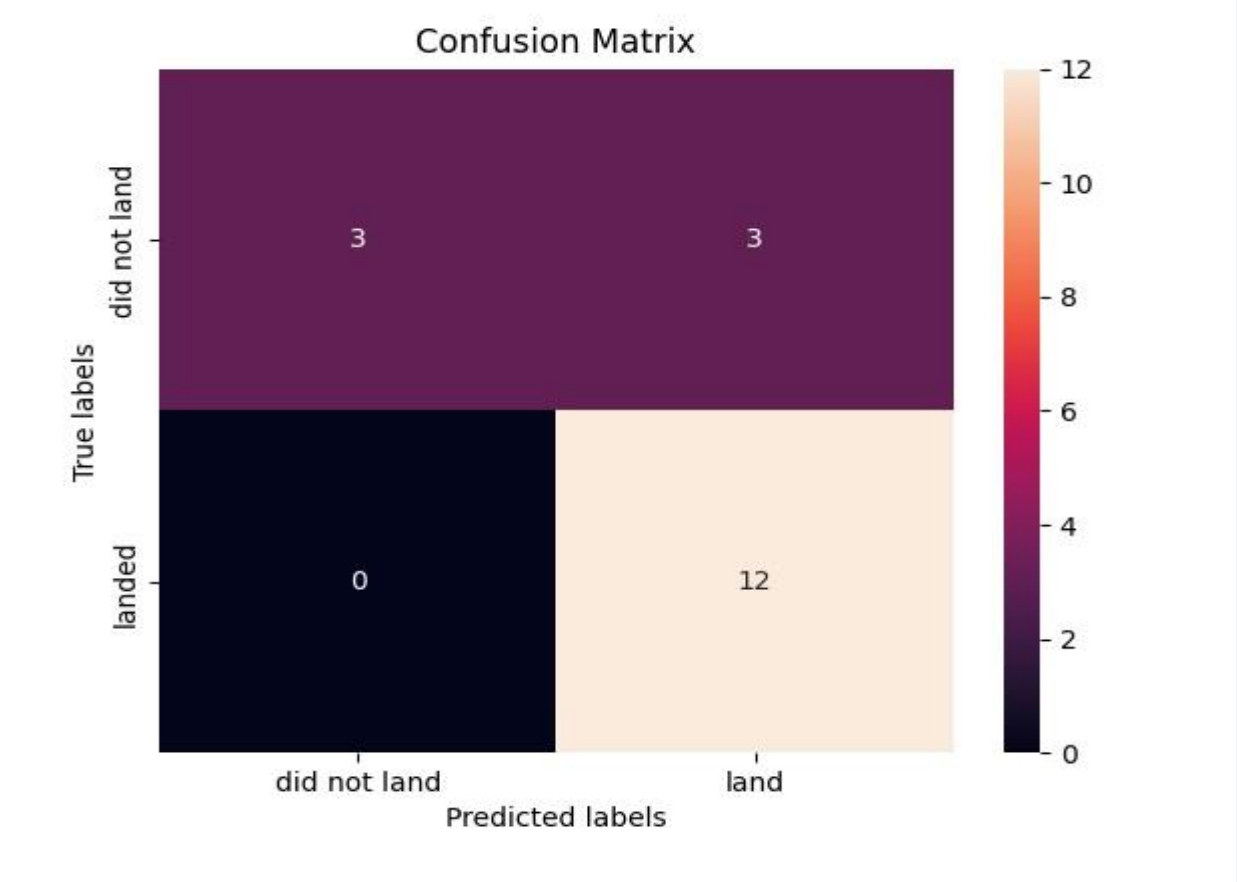
Predictive Analysis (Classification)

Classification Accuracy



- All models had the same accuracy on the test set 83.33%. This may be due to the small size of the sample, for more significant results the sampling should be larger

Confusion Matrix



- All models generated the same confusion matrix. The models predicted 12 successful landings when the true label was successful landing. The models predicted 3 unsuccessful landings when the true label was unsuccessful landing. The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). So models over predict successful landings.

Conclusions

- We have to develop a machine learning model for Space Y to bid against SpaceX, being our goal to predict when Stage 1 will successfully land.
- We have got data from a public SpaceX API and web scraping SpaceX Wikipedia page, obtaining valuables insights through the preparation of a dashboard (data visualization) and and the development of 4 machine learning models which give an accuracy of 83.3%.
- Allon Mask can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch, to determine whether the launch should be made or not.
- A larger number of data samples to feed machine learning models could improve them and obtain more reliable and representative results.

Appendix

- GitHub URL:

https://github.com/rafaj77/Data-Science-and-Machine-Learning-Capstone-Project

- Special thanks to all instructors of the IBM Data Science Professional Certificate.

Thank you!