IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vassileios Kekessis
13/10/2023

Outline

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

Executive Summary

➢ Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a dashboard with Plotly Dash
- Predictive Analysis (Classification)

➢ Summary of results

- Exploratory Data Analysis Results
- Interactive Analytics demo
- Predictive Analysis results

Introduction

➢ Project background and context

SpaceX is one of the most successful companies of the commercial space age, making space travel affordable. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can predict whether the first stage will launch, we can determine the cost of the launch. Based on public information and machine learning models we are going to predict if SpaceX will reuse its first stage.

➢ Questions to be answered

- How do variables such as payload mass, launch site, number of flights and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- Which machine learning model can classify the landing outcome with highest accuracy?

Section 1

Methodology

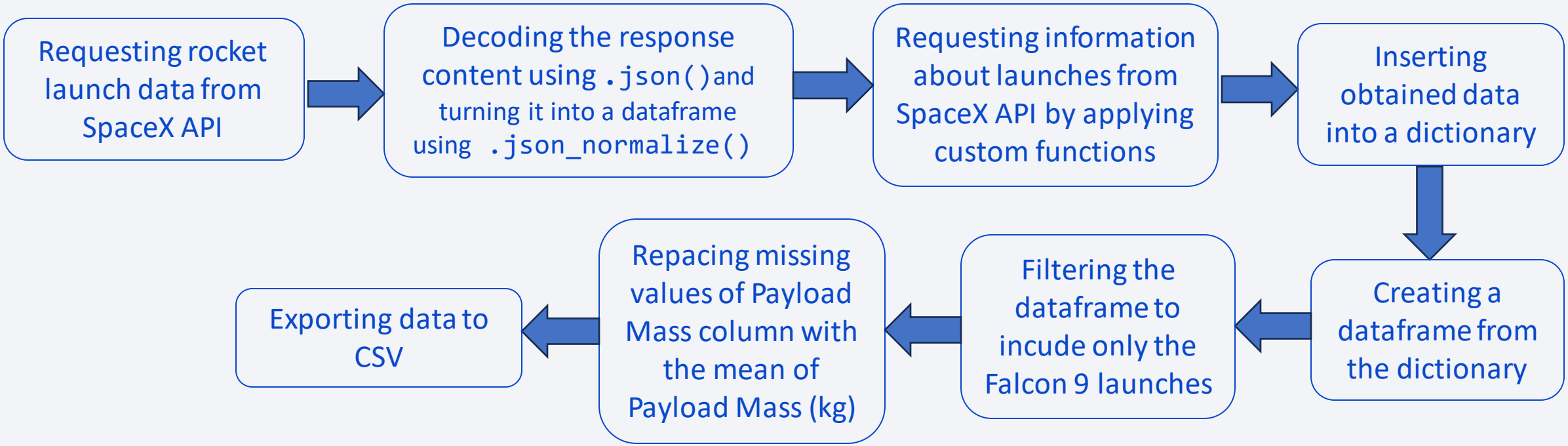
Methodology

- Data collection methodology:
  - SpaceX Rest API
  - Webscrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using one hot encoding to prepare the data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluating classification model for maximum accuracy

Data Collection

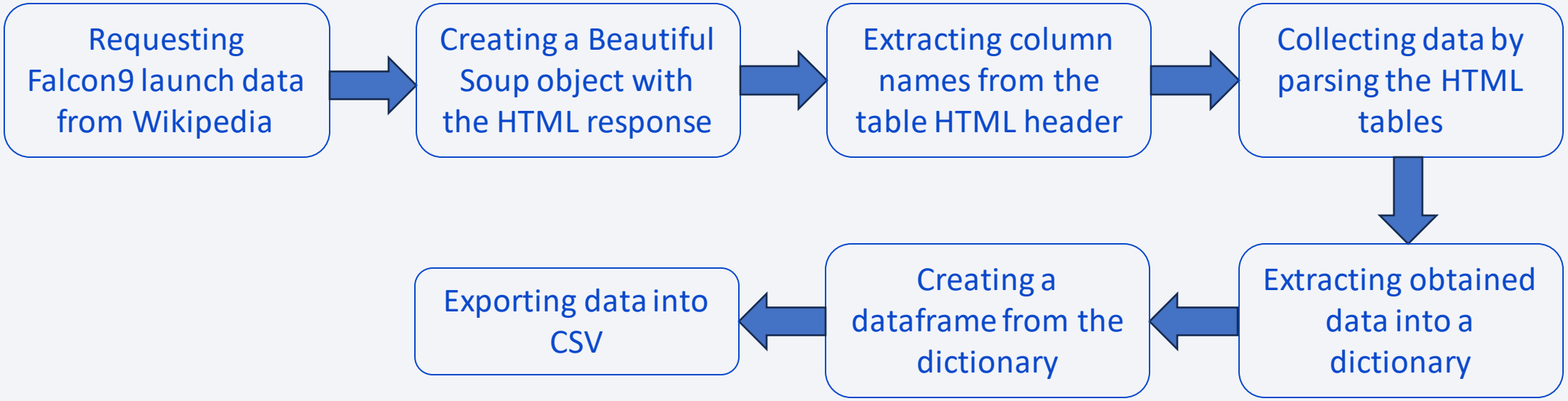
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX' s wikipedia page.
- Both methods had to be used to get complete information about the launches for an in depth analysis.
- Fields collected by using the SpaceX REST API:
  - FlightNumber
  - Date
  - BoosterVersion
  - PayloadMass
  - Orbit
  - LaunchSite
  - Outcome
  - Flights
  - GridFins
  - Reused
  - Legs
  - LandingPad
  - Block
  - ReusedCount
  - Serial
  - Longitude
  - Latitude
- Fields collected by using Wikipedia web scrapping:
  - FlightNo
  - Launch Site
  - Payload
  - PayloadMass
  - Orbit
  - Customer
  - Launch outcome
  - Version Booster
  - Booster Landing
  - Date
  - Time

Data Collection – SpaceX API



GitHub URL: Data Collection API

Data Collection - Scraping



Github URL: Data Collection with Web Scrapping

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, `True Ocean` means the mission outcome was successfully landed to a specific region of the ocean while `False Ocean` means the mission outcome was unsuccessfully landed to a specific region of the ocean. `True RTLS` means the mission outcome was successfully landed to a ground pad `False RTLS` means the mission outcome was unsuccessfully landed to a ground pad.`True ASDS` means the mission outcome was successfully landed on a drone ship `False ASDS` means the mission outcome was unsuccessfully landed on a drone ship.

Those outcomes are mainly converted into Training Labels with `1` means the booster successfully landed, `0` means it was unsuccessful.

- Calculate number of launches for each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence and mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Exporting data to CSV

Github URL: Data Wrangling

EDA with Data Visualization

- Charts plotted:
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Success rate per Orbit Type
  - Flight Number vs Orbit Type
  - Payload Mass vs Orbit Type
  - Yearly Success Rate
- *Scatter plots* are used to show to relationship between different variables. If there is a relationship, they could be used in machine learning model for classifying landing outcomes.
- *Bar chart* is used to show the contribution of categorical variables to the landing success rate.
- *Line chart* is used to show how the success rate changes over time.

Github URL: EDA with Data Visualization

EDA with SQL

- Performed SQL queries:

  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first succesful landing outcome in ground pad was acheived
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster_versions which have carried the maximum payload mass.
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names per month for the year 2015.
  - Ranking 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.

Github URL: EDA with SQL

Build an Interactive Map with Folium

- Markers of all launch sites
  - Added marker with Circle, Popup Label and Text Label for NASA Johnson Space Center using its latitude and longitude coordinates as a start location
  - Added marker with Circle, Popup Label and Text Label for all launsh sites using their latitude and longitude coordinates to show theiw geographical location and proximity to Equator and coasts.
- Coloured markers of launch outcomes for each launch site
  - Added coloured markers of successful (green) and failed (red) launch outomes per launch site to display launch sites with highest success rates.
- Distances between a launch site to its proximities
  - Added coloured lines to show the distance between launch site KSC LC-39A and its proximities including Railway, Highway, Coastline and closest city.

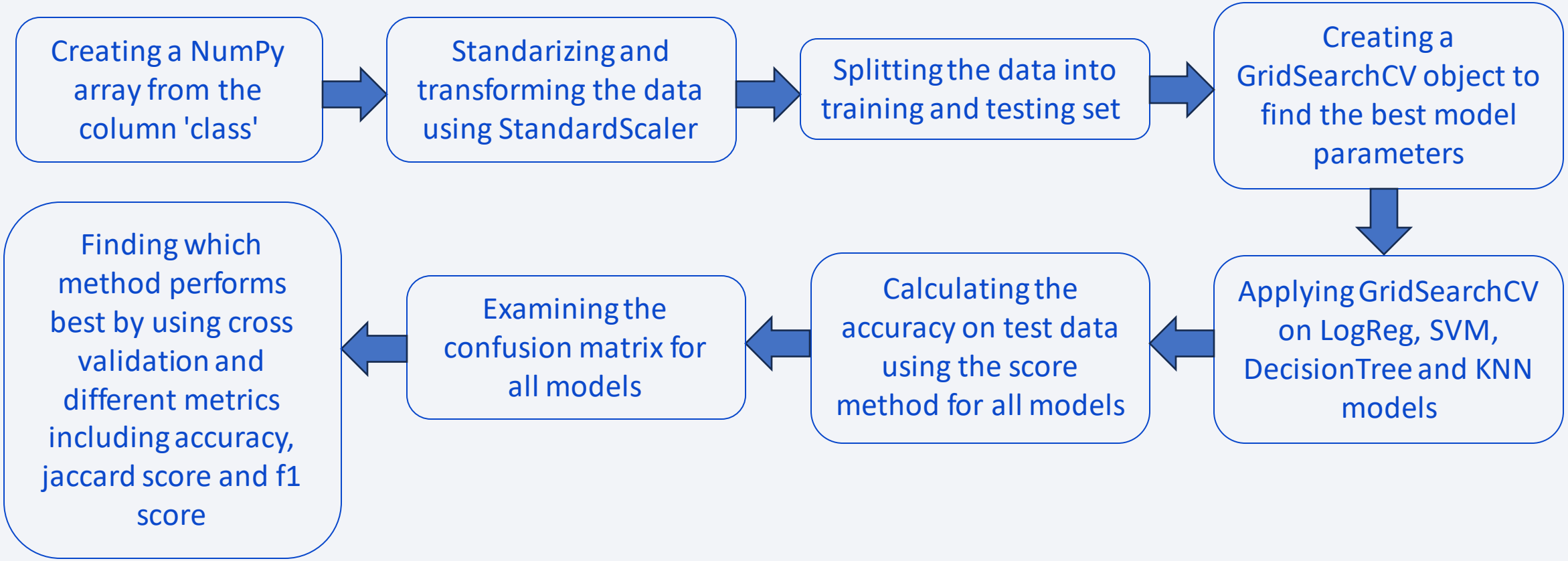
Github URL: Interactive Analytics with Folium

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List
  - Added a dropdown list to enable launch site selection.
- Pie Chart showing success launches for all sites
  - Added a pie chart to display the total count of successful launches for all sites and the Success vs Failed launches count for a site, if a specific site was selected.
- Slider of Payload Mass range
  - Added a pie chart to display the total count of successful launches for all sites and the Success vs Failed launches count for a site, if a specific site was selected.
- Scatter plot of Payload Mass vs Success Rate for different Booster Versions
  - Added a scatter chart to show the correlation between Payload and Launch Success

Github URL: SpaceX Dash App

Predictive Analysis (Classification)



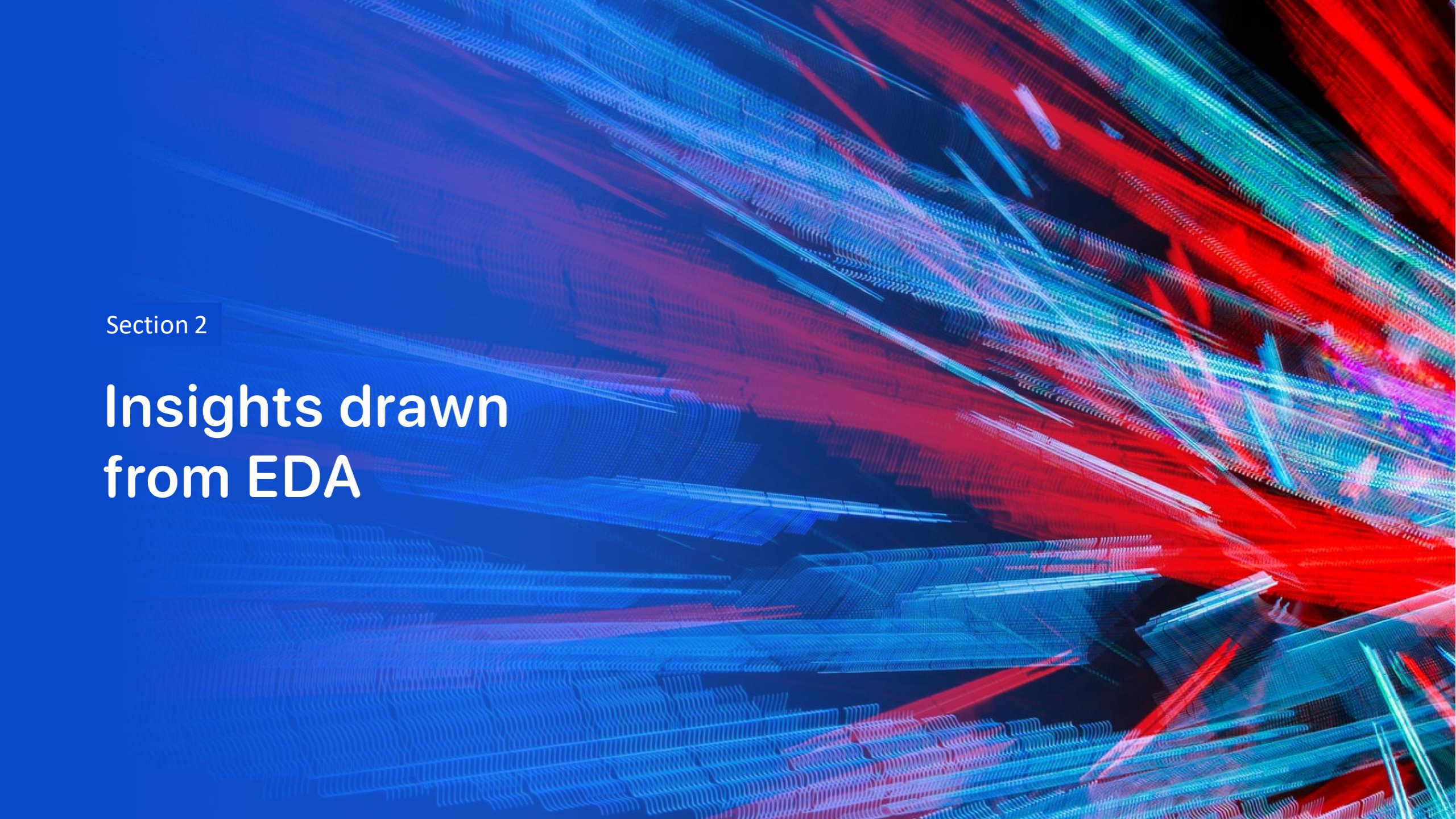
Github URL: Machine Learning Prediction

Results

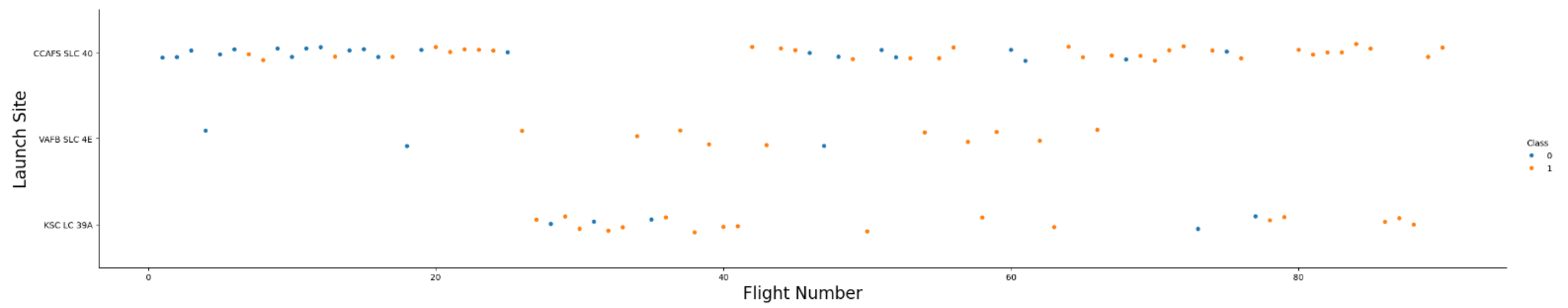
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

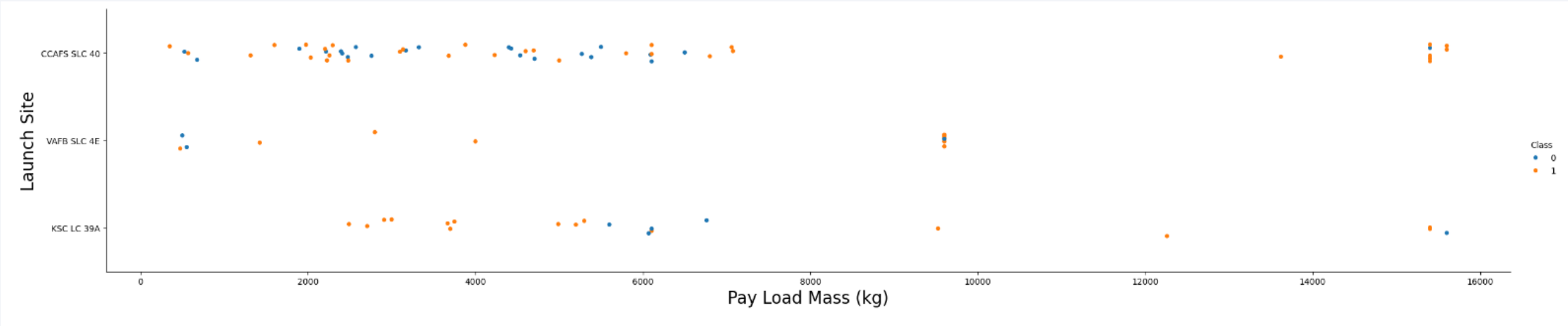


Flight Number vs. Launch Site



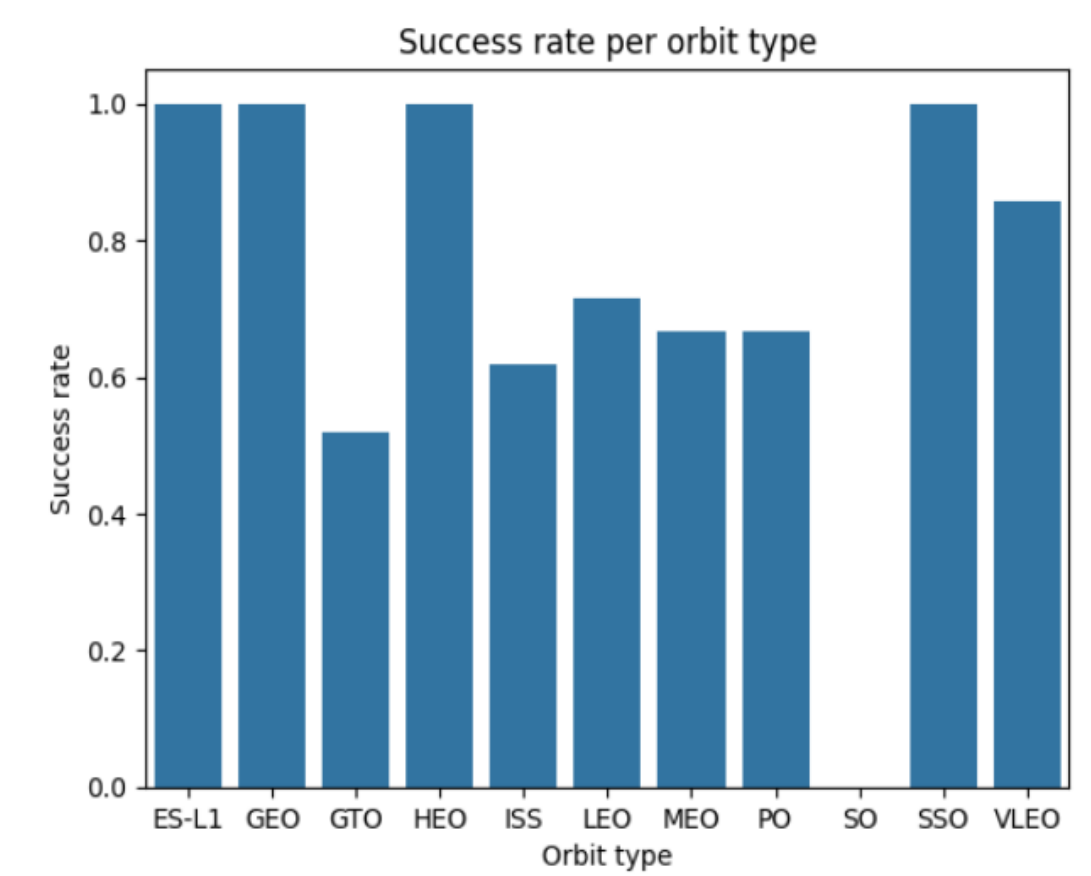
- Explanation
  - The earliest flights failed while the latest flights succeeded
  - The launch site CCAFS SCL 40 has about half of all launches
  - VAFB SLC 4E and KSC LC 39A have higher success rates

Payload vs. Launch Site



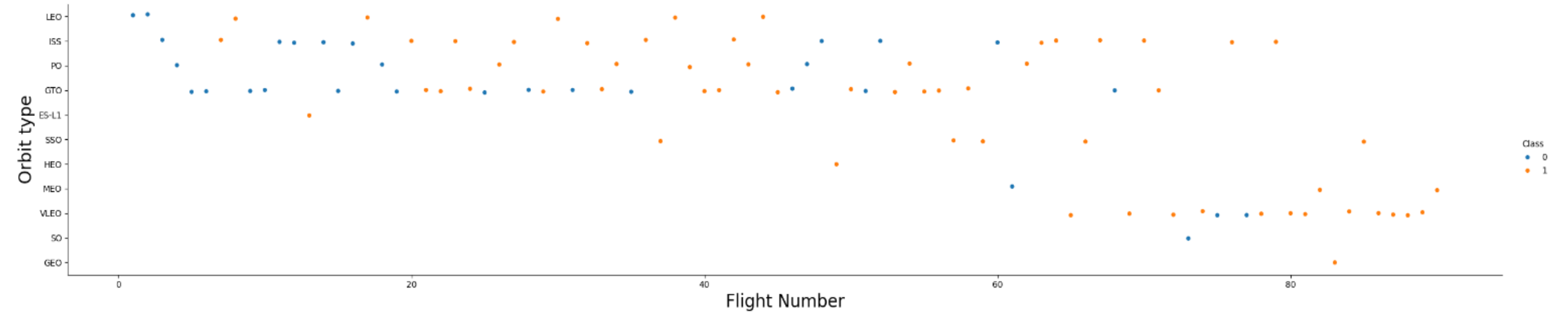
- Explanation
  - The higher the payload mass, the higher the success rate for each launch site
  - Most of the launches with payload mass over 7000 kg were successful
  - KSC LC 39A has 100% success rate for payload mass under 5500 kg

Success Rate vs. Orbit Type



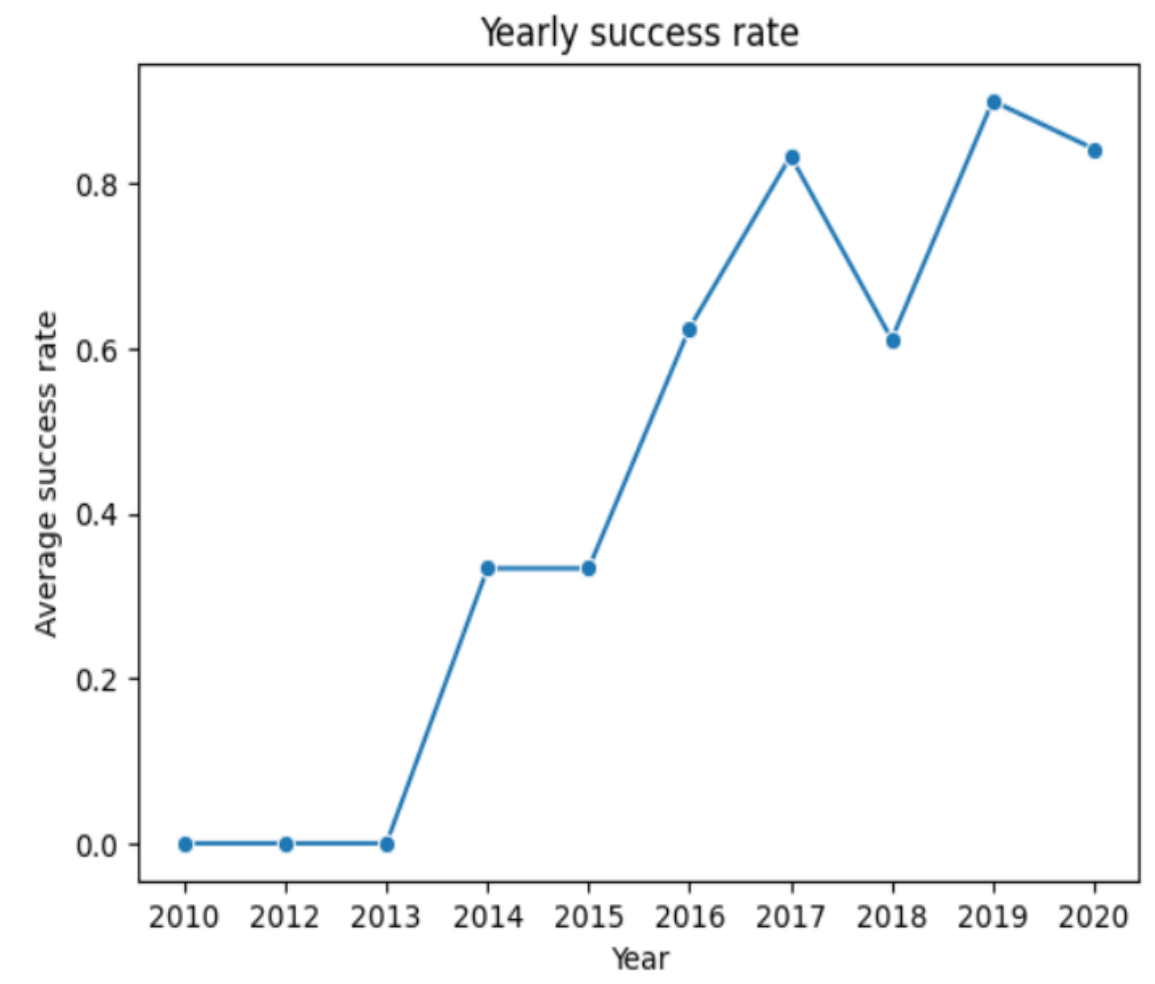
- Explanation
  - Orbits ES-L1,GEO,HEO AND SSO have 100% success rate
  - Orbits GTO,ISS,LEO,MEO and PO have success rate between 50% and 80%
  - Orbit SO has 0% success rate

Flight Number vs. Orbit Type



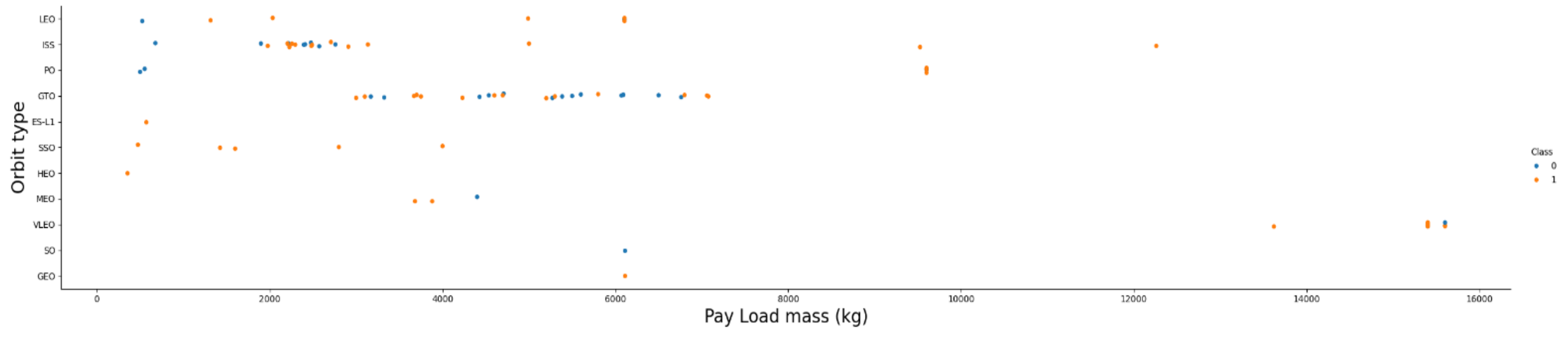
- Explanation
  - In the LEO orbit the success rate appears related to the number of flights. On the other hand, there seems to be no relationship between flight number when in GTO orbit.

Launch Success Yearly Trend



- Explanation
  - We can observe that success rate kept increasing since 2013 till 2020.

Payload vs. Orbit Type



- Explanation
  - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
  - However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are here and there.

All Launch Site Names

In [6]:
```
%%sql

select distinct "Launch_Site" from SPACEXTABLE
```

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

Out[6]:

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

- Explanation: displaying the names of unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

In [8]:
```
%%sql

select * from SPACEXTABLE
where "Launch_Site" like 'CCA%' limit 5
```

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

Out[8]:

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

- Explanation: displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

In [8]:
```
%%sql

select Customer, sum("PAYLOAD_MASS__KG_") as Total_payload_mass
from SPACEXTABLE
where Customer = 'NASA (CRS)'
```

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

Out[8]:

Customer	Total_payload_mass
NASA (CRS)	45596

- Explanation: displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

In [9]:
```
%%sql

select "Booster_version", avg("PAYLOAD_MASS__KG_") as Average_payload_mass
from SPACEXTABLE
where "Booster_version" = 'F9 v1.1'
```

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

Out[9]:

Booster_Version	Average_payload_mass
F9 v1.1	2928.4

- Explanation: displaying average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

In [11]:

```
%%sql

select "Landing_Outcome", min(Date) as First_successful_landing
from SPACEXTABLE
where "Landing_Outcome" = 'Success (ground pad)'
```

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

Out[11]:

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

- Explanation: listing the date when the first succesful landing outcome in ground pad was acheived.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [13]:

```
%%sql

select distinct("Booster_Version") as Booster_Version_successful_in_drone_ship
from SPACEXTABLE
where "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_MASS__KG_" between 4000 and 6000
```

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

Out[13]:

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

- Explanation: listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [19]:
```
%%sql

select "Mission_Outcome", count("Mission_Outcome")
from SPACEXTABLE
group by "Mission_Outcome"
```

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

Out[19]:

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

- Explanation: listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

In [22]:
```
%%sql

select "Booster_Version" as max_payload_mass_booster_versions
from SPACEXTABLE
where "PAYLOAD_MASS__KG_" in
(select max("PAYLOAD_MASS__KG_")
from SPACEXTABLE)
```

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

Out[22]:

max_payload_mass_booster_versions
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

- Explanation: listing the names of the booster_versions which have carried the maximum payload mass

2015 Launch Records

In [27]:
```
%%sql

select substr(Date,6,2) as Month, "Landing_Outcome", "Booster_Version", "Launch_Site"
from SPACEXTABLE
where "Landing_Outcome" = 'Failure (drone ship)' and substr(Date,0,5) = '2015'
```

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

Out[27]:

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

- Explanation: listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch site for the months in year 2015.

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

In [29]:

```
%%sql

select Date, "Landing_Outcome", count("Landing_Outcome") as Landing_Outcome_count
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by "Landing_Outcome"
order by Landing_Outcome_count desc
```

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

Out[29]:

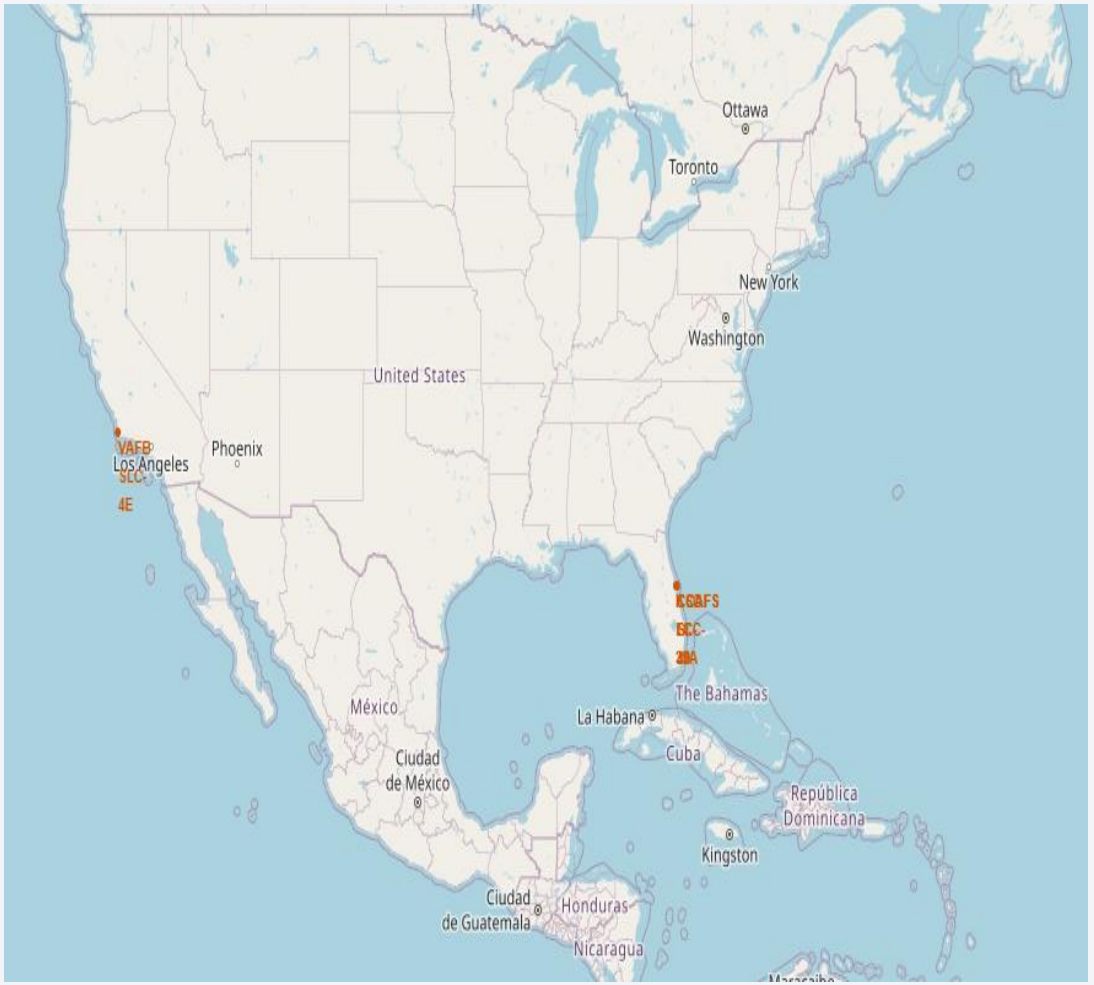
Date	Landing_Outcome	Landing_Outcome_count
2012-05-22	No attempt	10
2015-12-22	Success (ground pad)	5
2016-08-04	Success (drone ship)	5
2015-10-01	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2015-06-28	Precluded (drone ship)	1
2010-08-12	Failure (parachute)	1

- Explanation: 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.


Section 3
Launch Sites
Proximities Analysis

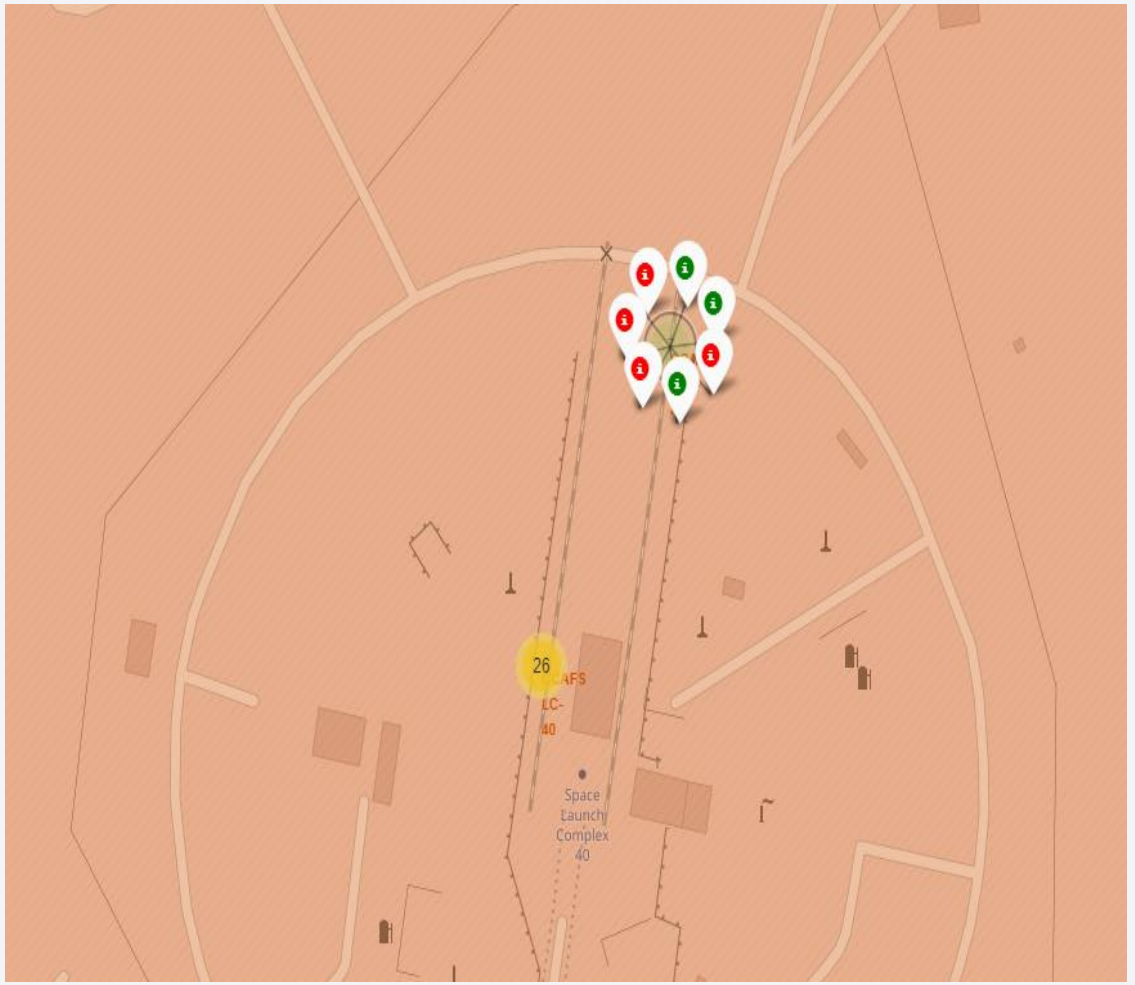
All launch sites' location markers on a global map

- Explanation
  - All launch sites are in proximity with the equator line. The land is moving faster at the equator than any other place at the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hr. If a ship is launched from the equator it goes up into the space and it is also moving around the Earth at the same time it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
  - All launch sites are in very close proximity to the coast, while launching rockets towards the ocean minimizes the risk of having any debris dropping or expoding near people.



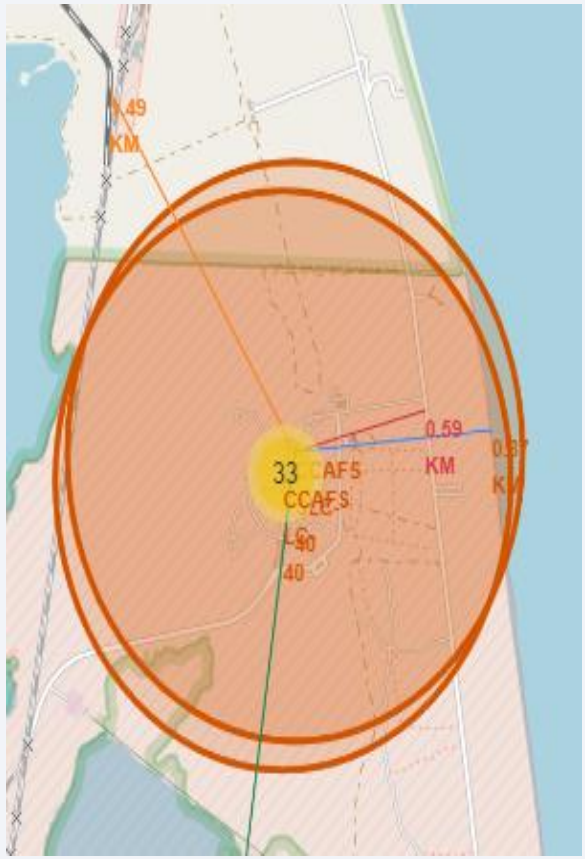
Color-labeled launch records on the map

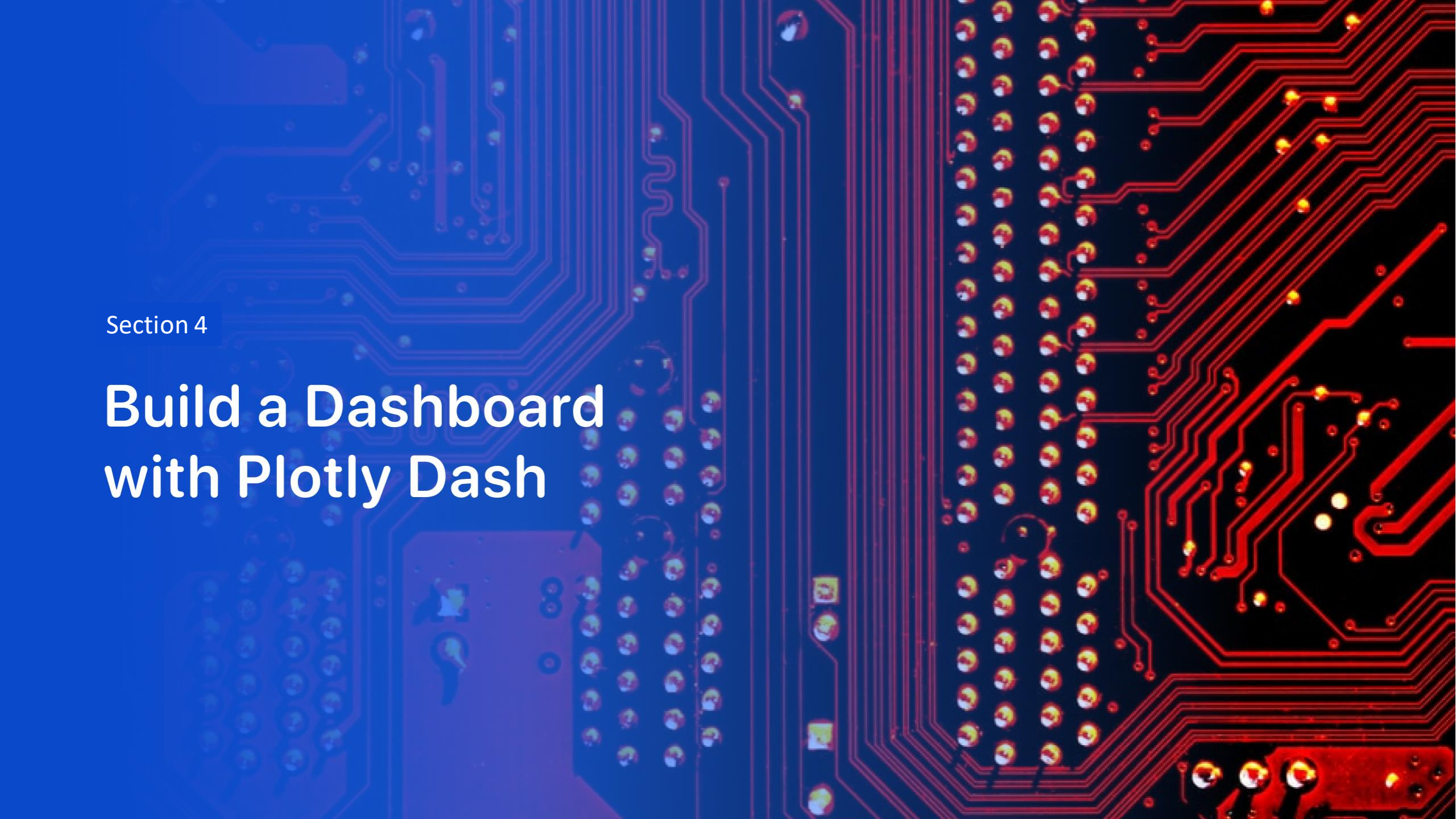
- Explanation
  - Colored markers are used to help us identify the launch sites with the highest success rate.
    - ✓ Green Marker = successful launch
    - ✓ Red Marker = failed launch
  - It seems that launch site CCAFS SLC-40 has relatively low success rate.



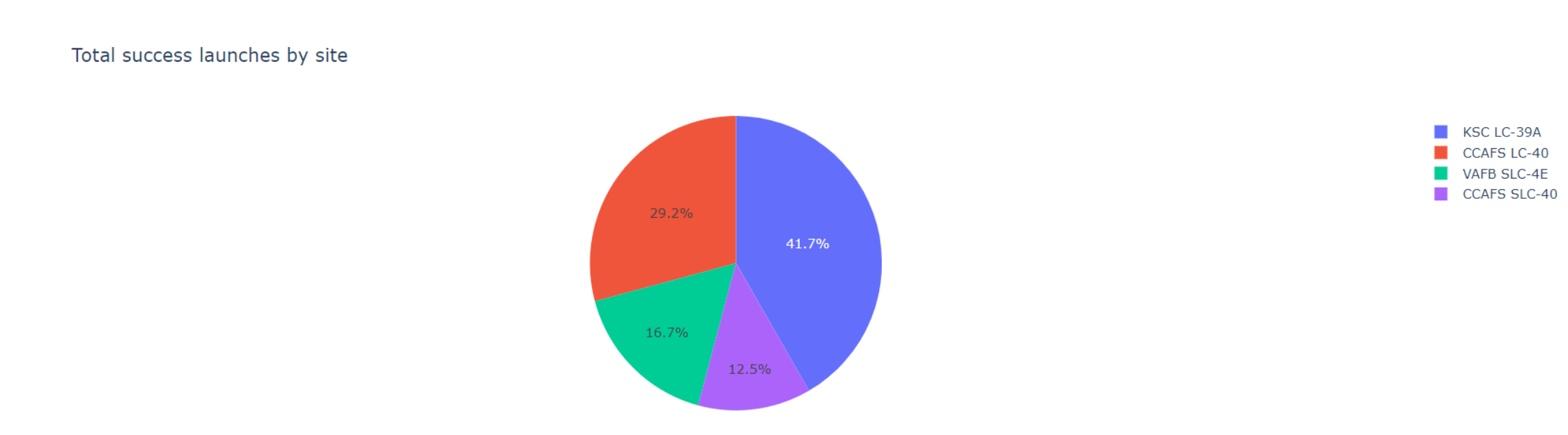
Distance from the launch site CCAFS SLC-40 to its proximities

- Explanation
  - If we draw a line between launch site CCAFS SLC-40 to its proximities and display the distance, we can see that:
    - ✓ It is very close to railway (Nasa Railway: 1.49 km)
    - ✓ It is very close to highway (Samuel C Phillips Highway: 1.49 km)
    - ✓ It is very close to coastline (0.87 km)
    - ✓ It is relatively close to its closest city (Cape Carneval: 18.16 km)



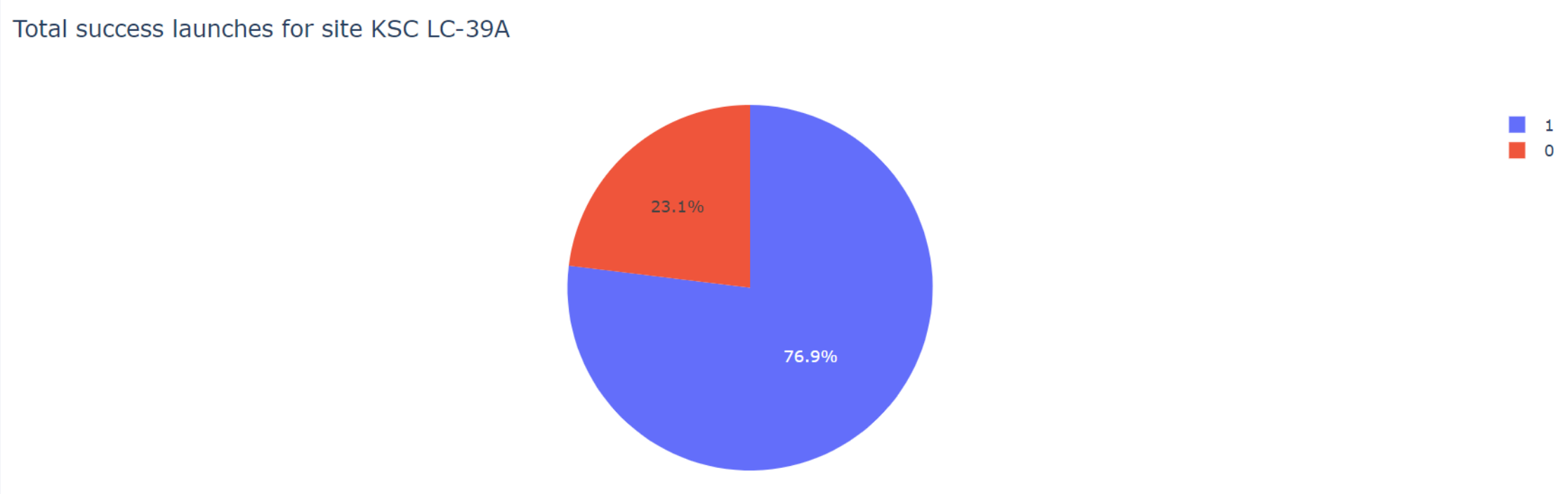

Section 4
Build a Dashboard with Plotly Dash

Launch Success count for all sites



- Explanation
  - From the pie chart we can clearly see that that launch site KSC LC-39A has the highest success rate in launches.

Total success launches for site KSC LC-39A



- Explanation
  - It seems that launch site KSC LC-39A has the highest success rate (76.9%)

Payload Mass vs Launch Outcome for all sites

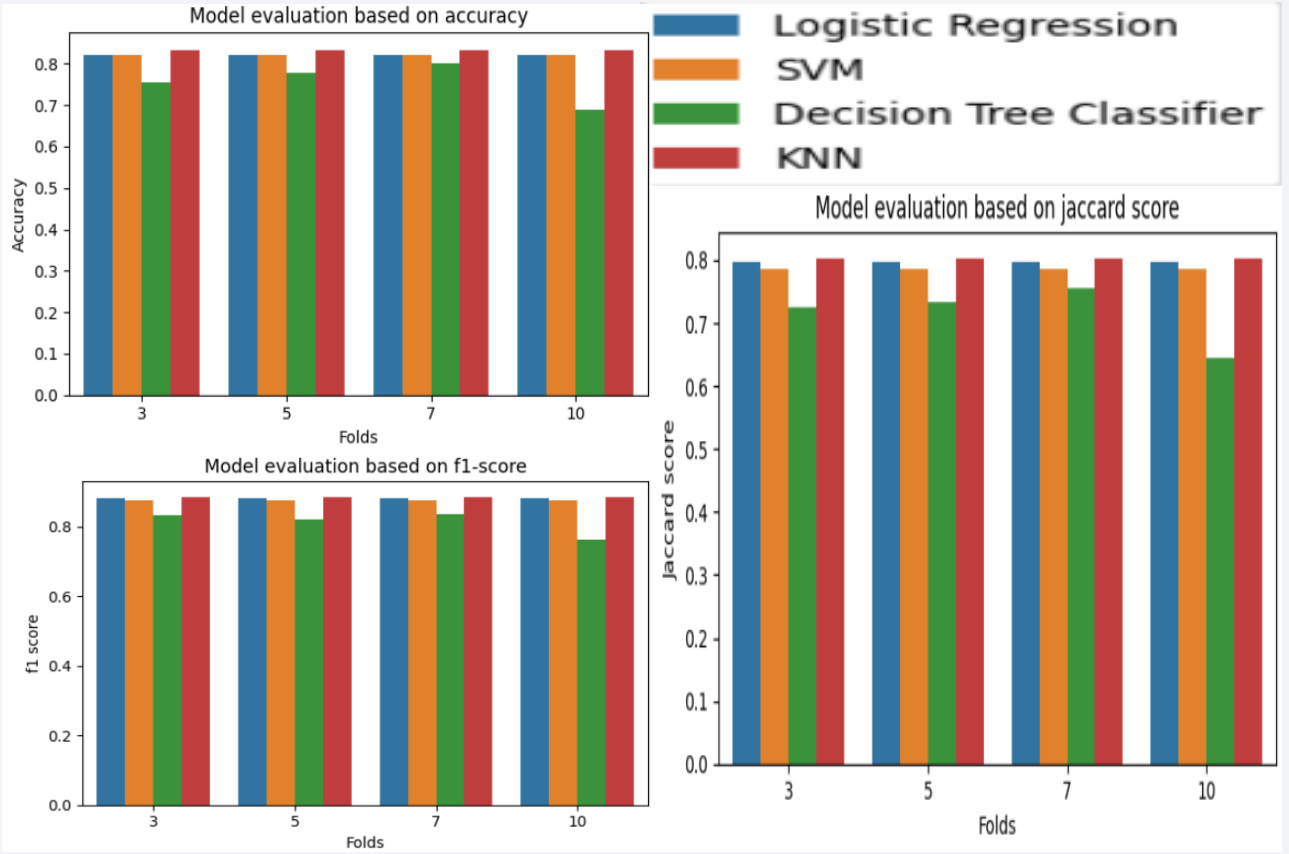
- Explanation
  - Payloads between 2000 and 5000 kg have the highest success rate.
  - Payloads between 5000 and 10000 kg have the lowest success rate.
  - It is surprising that only one launch took place with the F9 Booster Version of type B5, which was successful. Therefore, Booster Version B5 has the highest launch success rate (100%).

Section 5

Predictive Analysis (Classification)

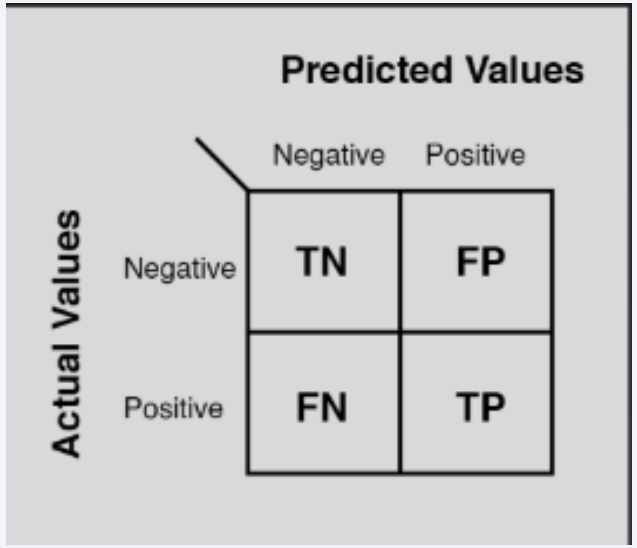
Classification Accuracy

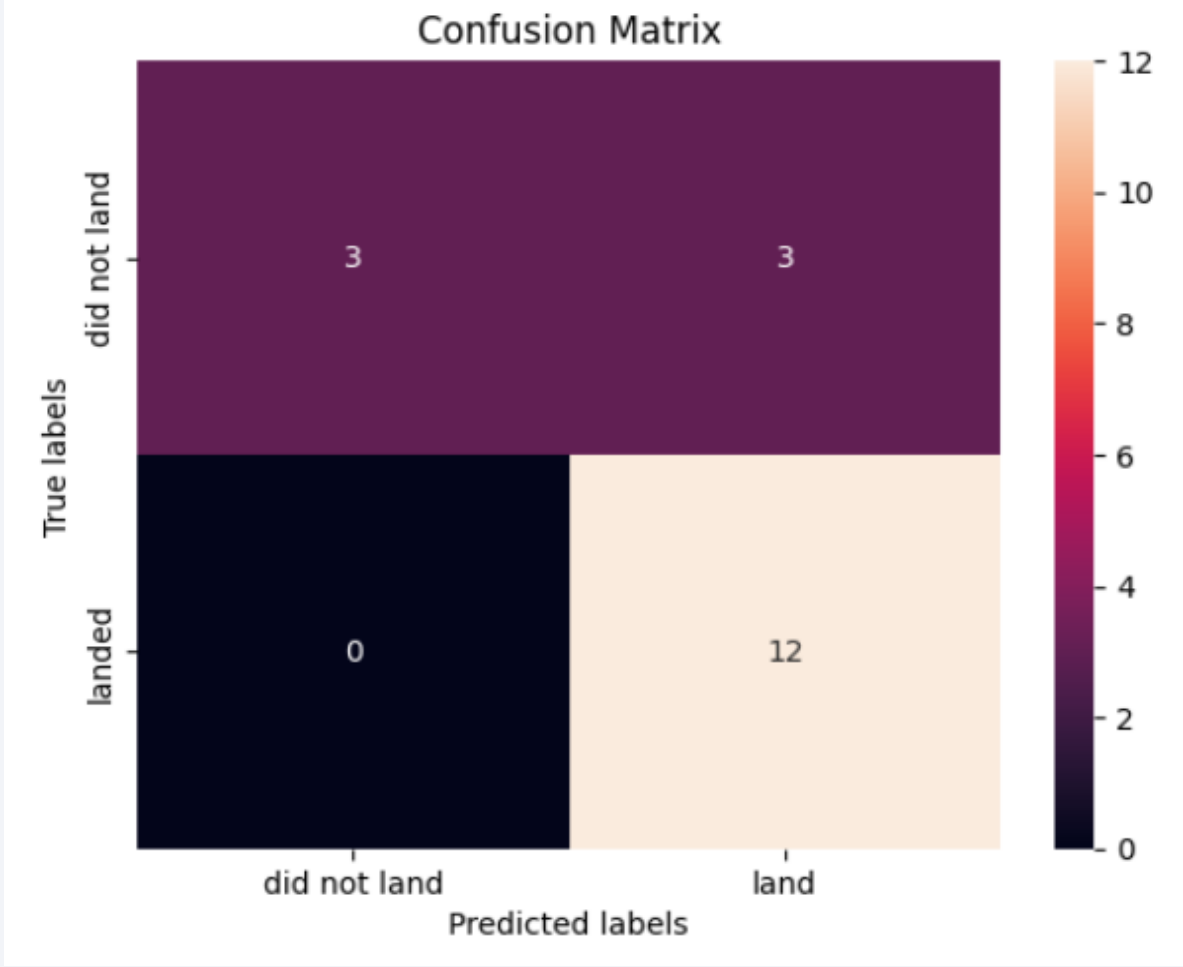
- Explanation
  - Because of the small test sample size we use K-Fold cross-validation to determine the model that performs the best. We split the dataset into different number of folds and for each number of folds we calculate the mean of different classification metrics, including accuracy, f1-score and jaccard score. From the bar plots it is obvious that KNN has the best performance across all folds and metrics.



Confusion Matrix

- Explanation
  - Examining the confusion matrix, we see that KNN can distinguish between the different classes. We see that the major problem is false positives





Conclusions

- K Nearest Neighbors is the best classification algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a higher payload mass.
- Most of launch sites are in proximity to the Equator line and all launch sites are in very close proximity to the coastline.
- The launch success rate increases over the years.
- Orbits ES-L1,GEO,HEO and SSO have 100% success rate.
- Booster Version B5 has the highest success rate.

Appendix

Special thanks to:

Instructors

Coursera

IBM

Thank you!