Winning Space Race with Data Science

Thomas Oluwasegun Olubunmi

25th Jan., 2022

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 an interactive Dashboard with plotly Dash
- Predictive Analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Background:

- Commercial Space age is here
- SpaceX has been the best and most successful company of the commercial age, making space travel affordable ($62 million vs. $162 million (USD))
- The success is due to the ability to recover part of rocket first stage
- SpaceY is trying to compete with SpaceX

- Questions that need answers
  - How do variables such as payload mass, launch site, orbits and number of flights affect the success rate of the first stage landing?
  - Does the rate of successful landings increase over the years?
  - What is best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using one Hot Encoding to prepare the data to a 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 evaluation of classification models to ensure the best results

Data Collection

- Data collection process involved a combination of API requests from Spacex REST API and web Srapping data from a table in SpaceX's Wikipedia entry.

  Data Columns are obtained by using Spacex REST API:

  FlightNumber, Date, BooterVersion, PayloadMass, Orbit, LaunhSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ResuedCount, Serial, Longitude, Latitude.

  Data columns are obtained by using Wikipedia Web Scrapping:

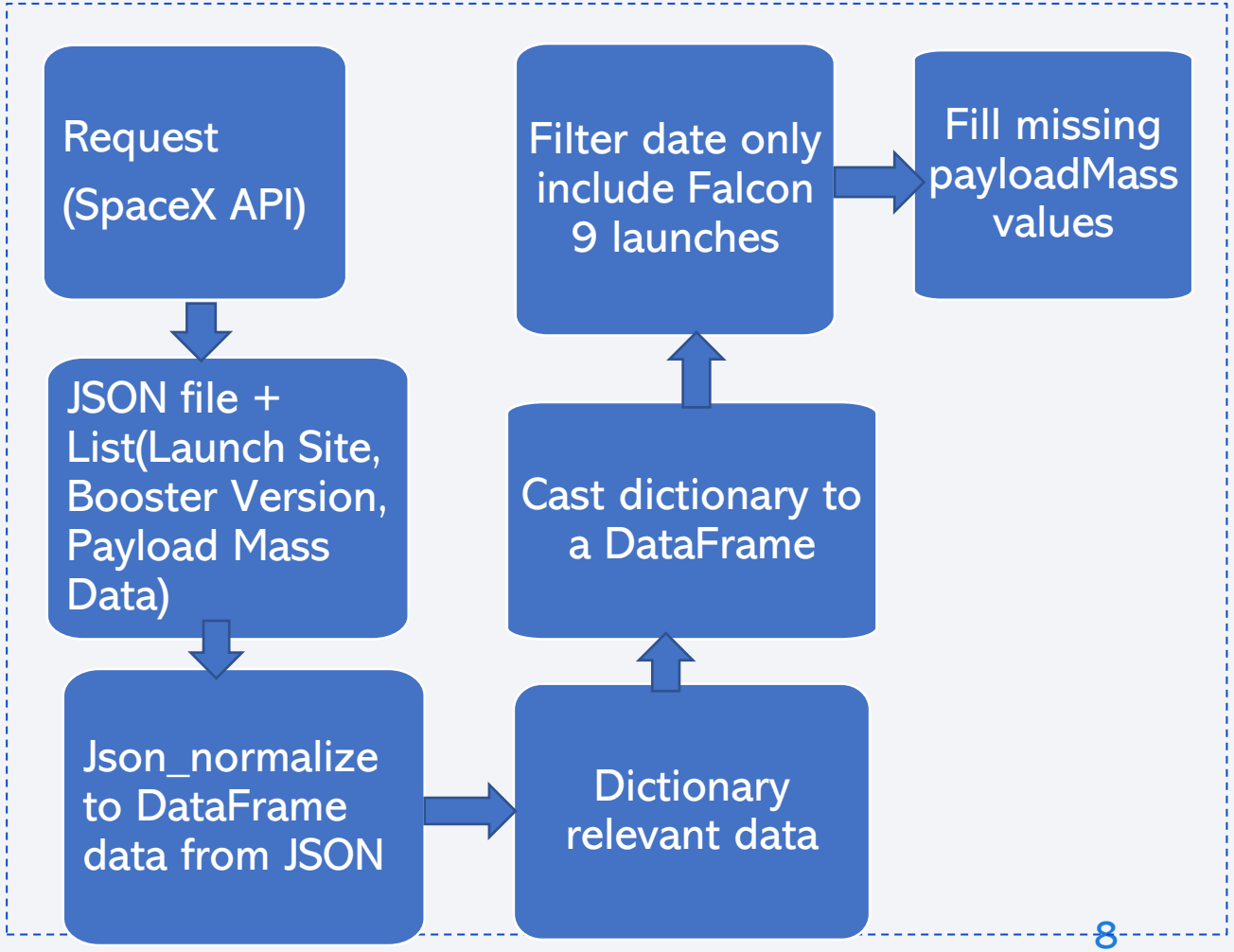
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch, Outcome, Version Boster, Booster landing, Date, Time.

Data Collection – SpaceX API

Data Collection SpaceX API

GitHub url:

https://github.com/primestute/primestute/blob/main/Data%20Collection%20API.ipynb

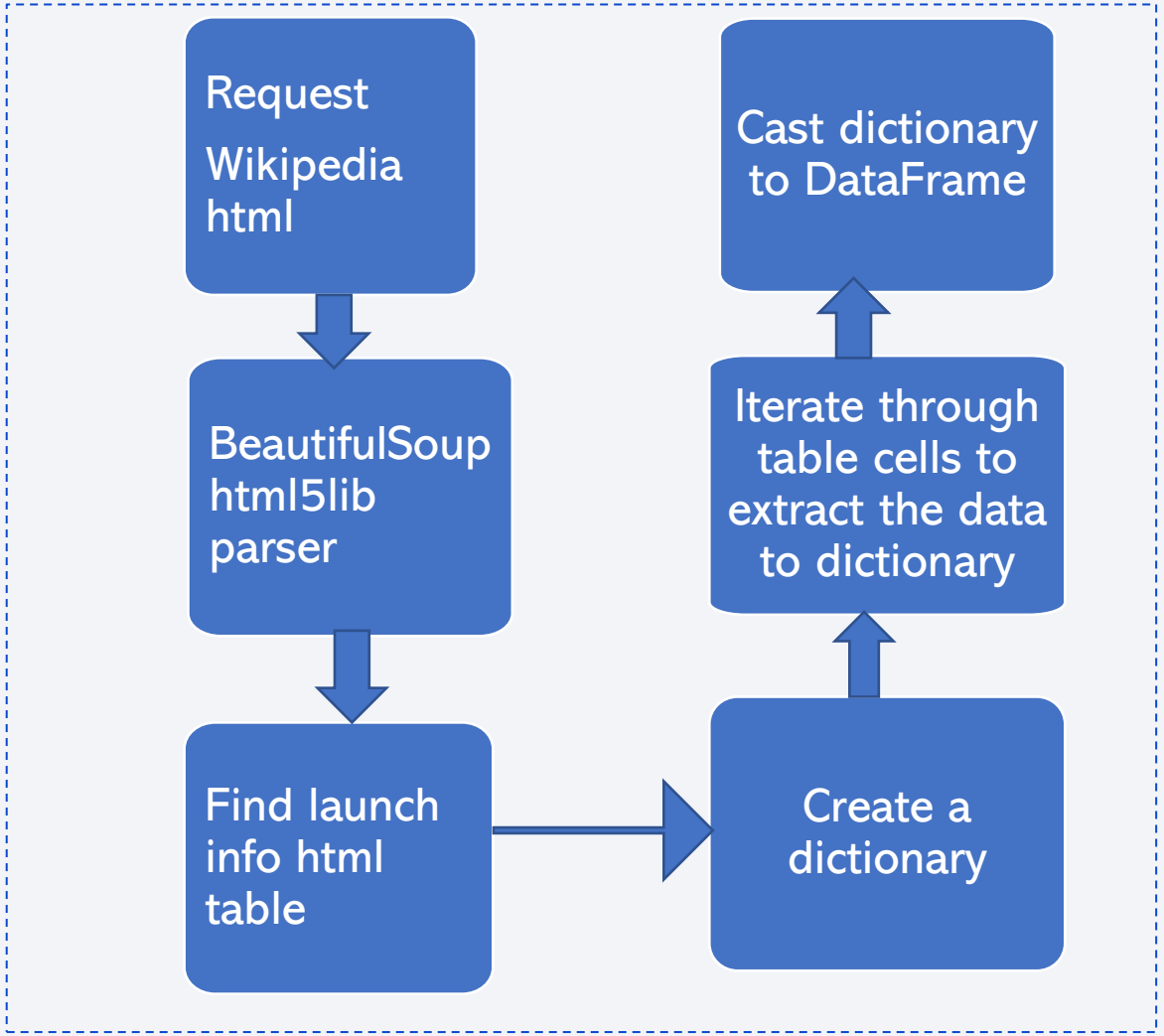


Data Collection - Scraping

Data collection Web Scrapping.

GitHub url:

https://github.com/primestute/primestute/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

The objective here is to perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
ship.
We will mainly convert those outcomes into Training Labels with ”1” means the booster successfully landed ”0” means it was unsuccessful.
There are two components in the outcome column: “Mission Outcome”, “Landing Location”

New training label column “Class” with a value of “1” if “Mission Outcome” is true and “0” otherwise.

GitHub url: https://github.com/primestute/primestute/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate Yearly Trend.

Scatter plots show the relationship between variables. If a relationship between variables. If a relationship exists, they could be used in Machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts show trends in data over time (time series)

GitHub url:

https://github.com/primestute/primestute/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Loaded dataset into IBM Database.
- Queried using sql-python integration
- Queried were better to have a deeper understanding of the data set.
- Queried information about the names of the unique launch sites in the space mission, 5 records were names begin with “CCA”, different payload mass of customers and boosters versions, and landing outcomes.

GitHub url:

https://github.com/primestute/primestute/blob/main/EDA%20with%20SQL.ipynb

Interactive Map with Folium

- Markers of all Launch Sites:
  - Added Marker with Circle, Popup Label and Text Label NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Colored Markers of the launch outcomes for each Launch Site:

  Added colored Markers of success (Green) and filled (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
  - Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) its proximities like Railway, Highway, Coastline and Closest City.

https://github.com/primestute/primestute/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch was selected.

Slider of Payload Mass Range:

- Added a slider to select payload range

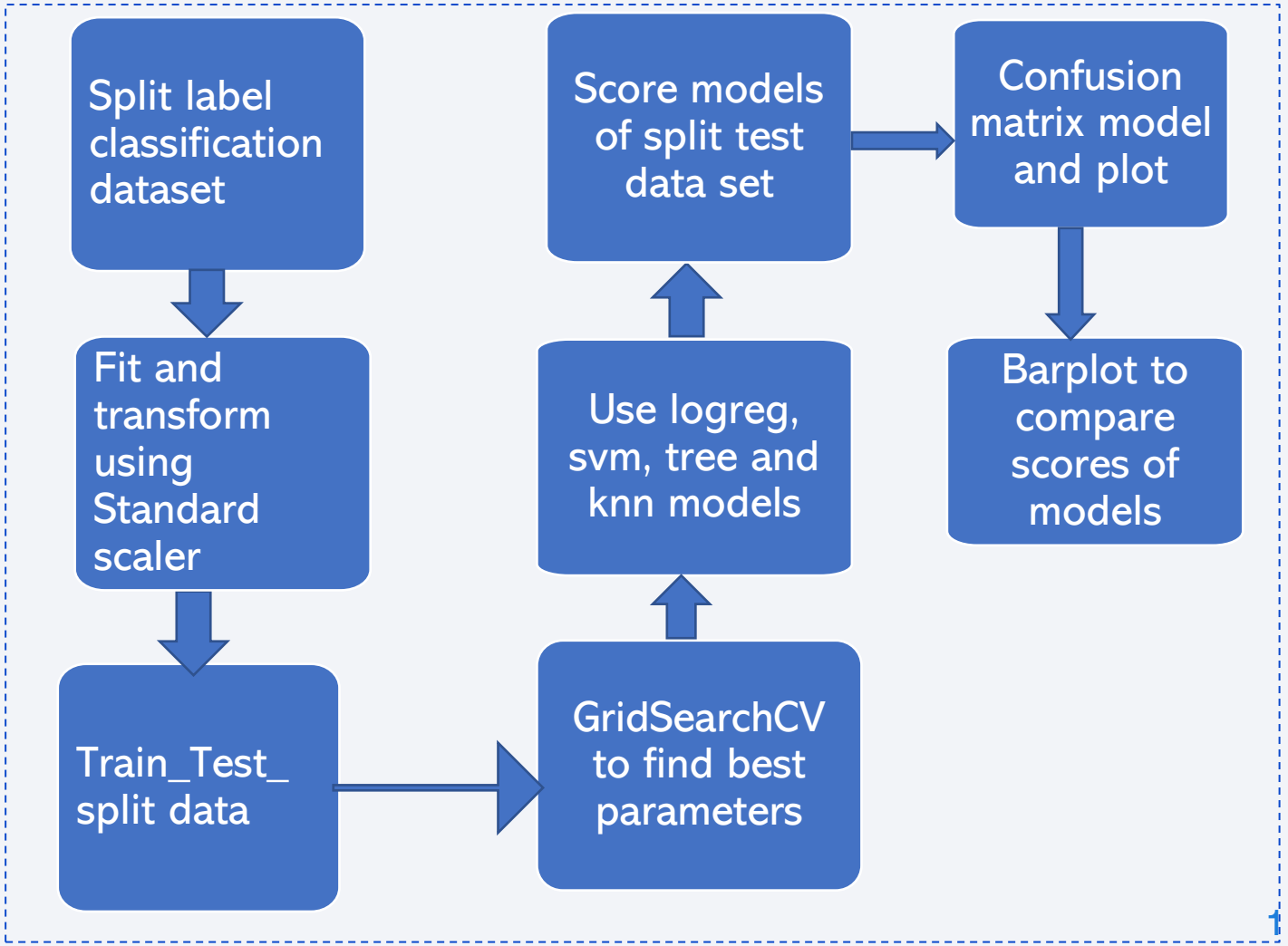
Scatter Plot of payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter plot to show the correlation between payload and Launch Success.

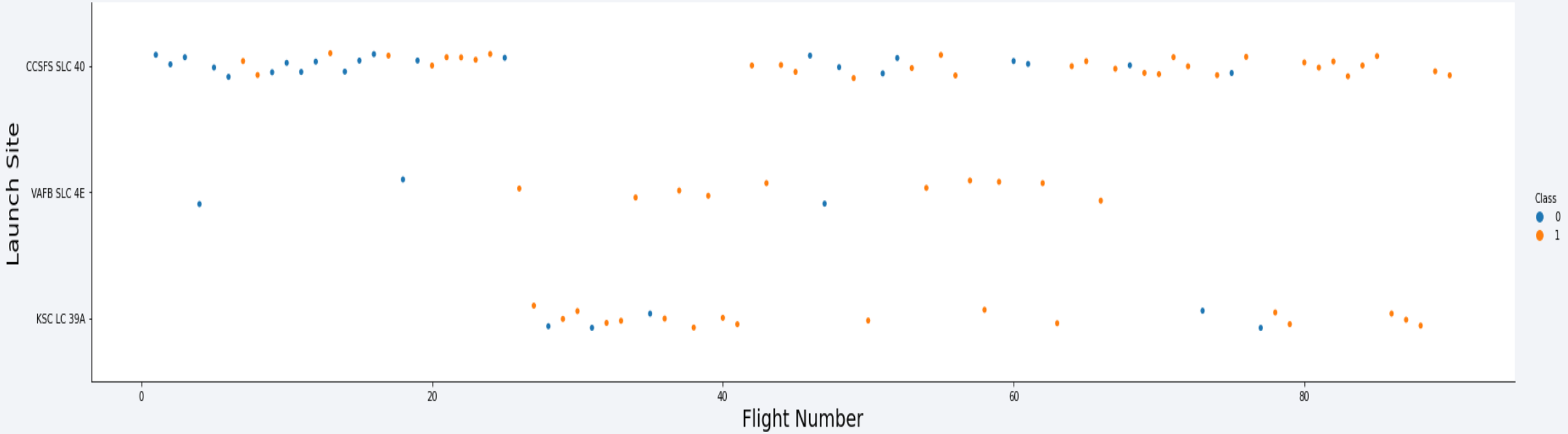
GitHub url: https://github.com/primestute/primestute/blob/main/dash.py.py

Predictive Analysis (Classification)

GitHub url:
https://github.com/primestute/primestute/blob/main/Machine%20Learning%20Prediction_Part_5.ipynb



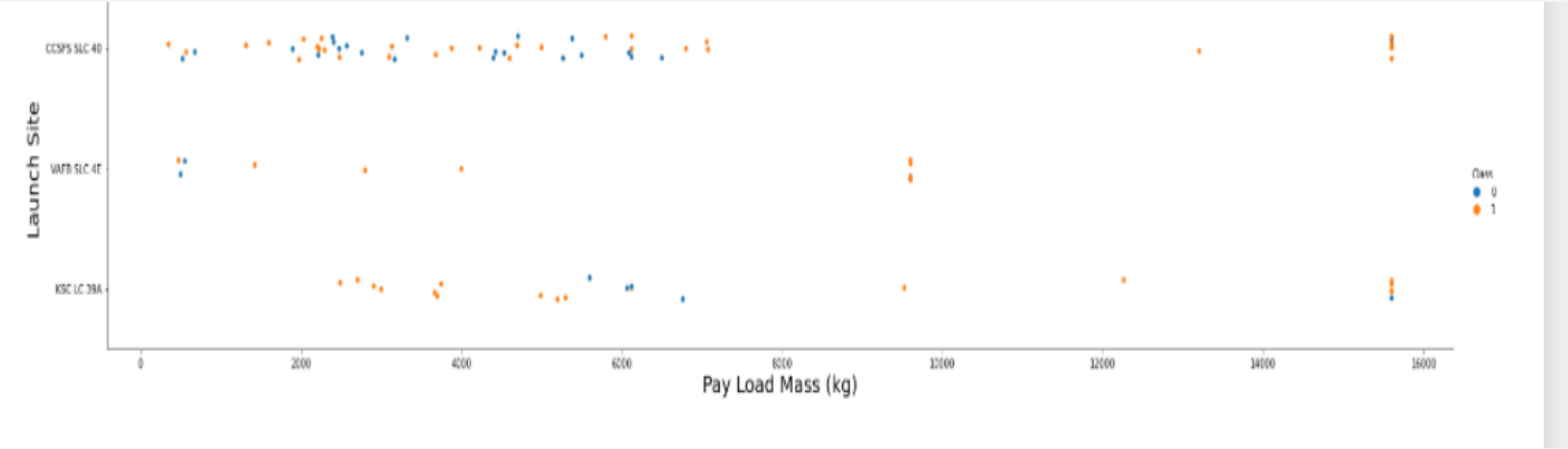
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC 39A have higher success rates
- It can be assumed that each new launch has a higher rate of success.

Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success rate
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

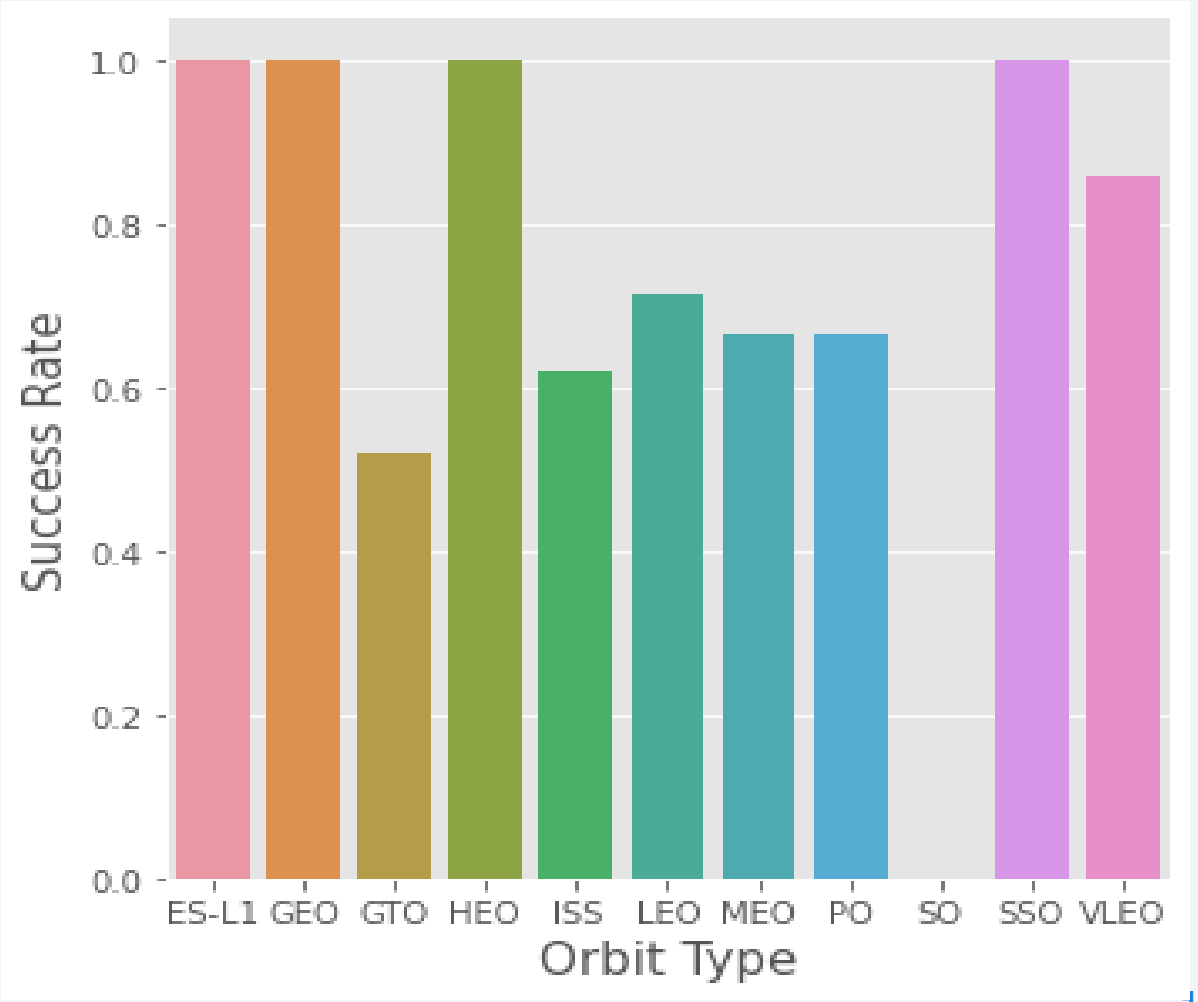
Success Rate vs. Orbit Type

Explanation:

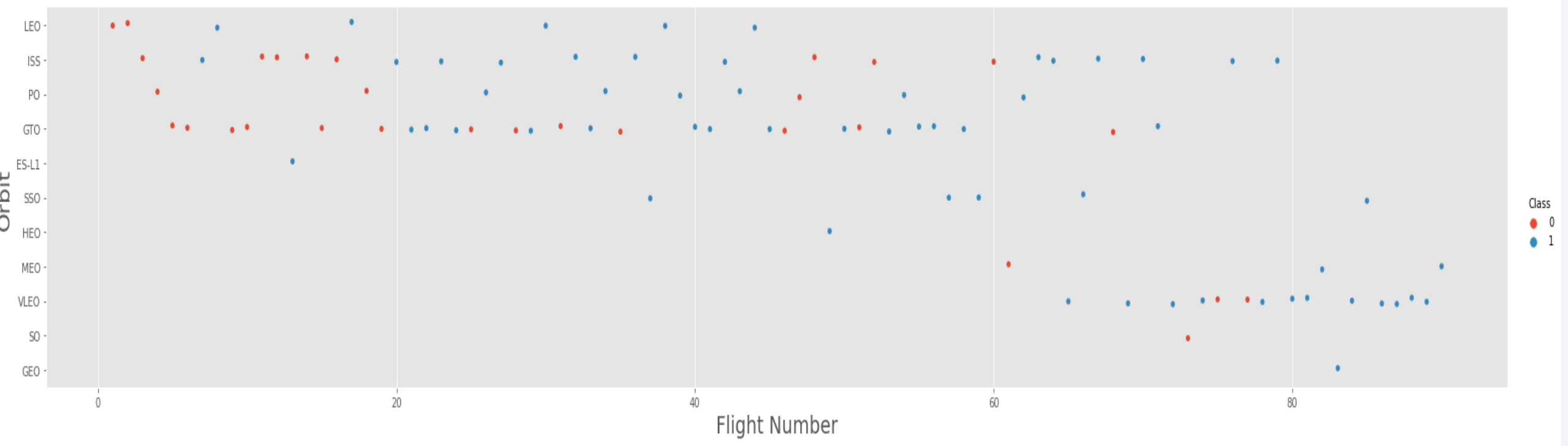
- Orbits with 100% rate:
  - ❖ ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - ❖ SO

Orbits with success rate between 50% and 85%:

- ❖ GTO, ISS, LEO, MEO and PO



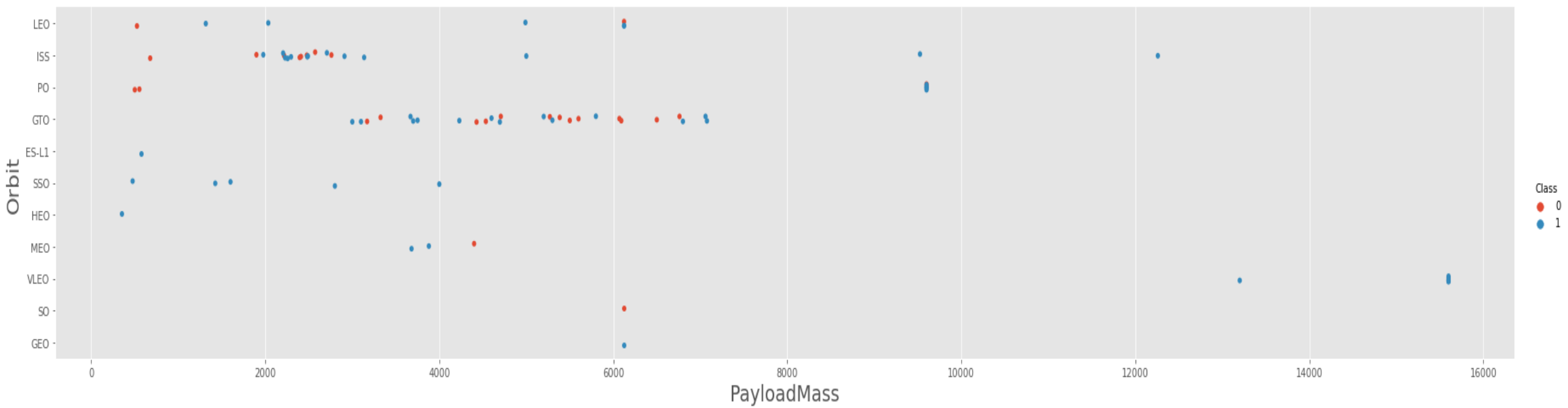
Flight Number vs. Orbit Type



Explanation:

- In the LEO orbit the success appears related to the number of flights; Looking at the plot, it appears there is no relationship between flight number when in GTO Orbit.

Payload vs. Orbit Type



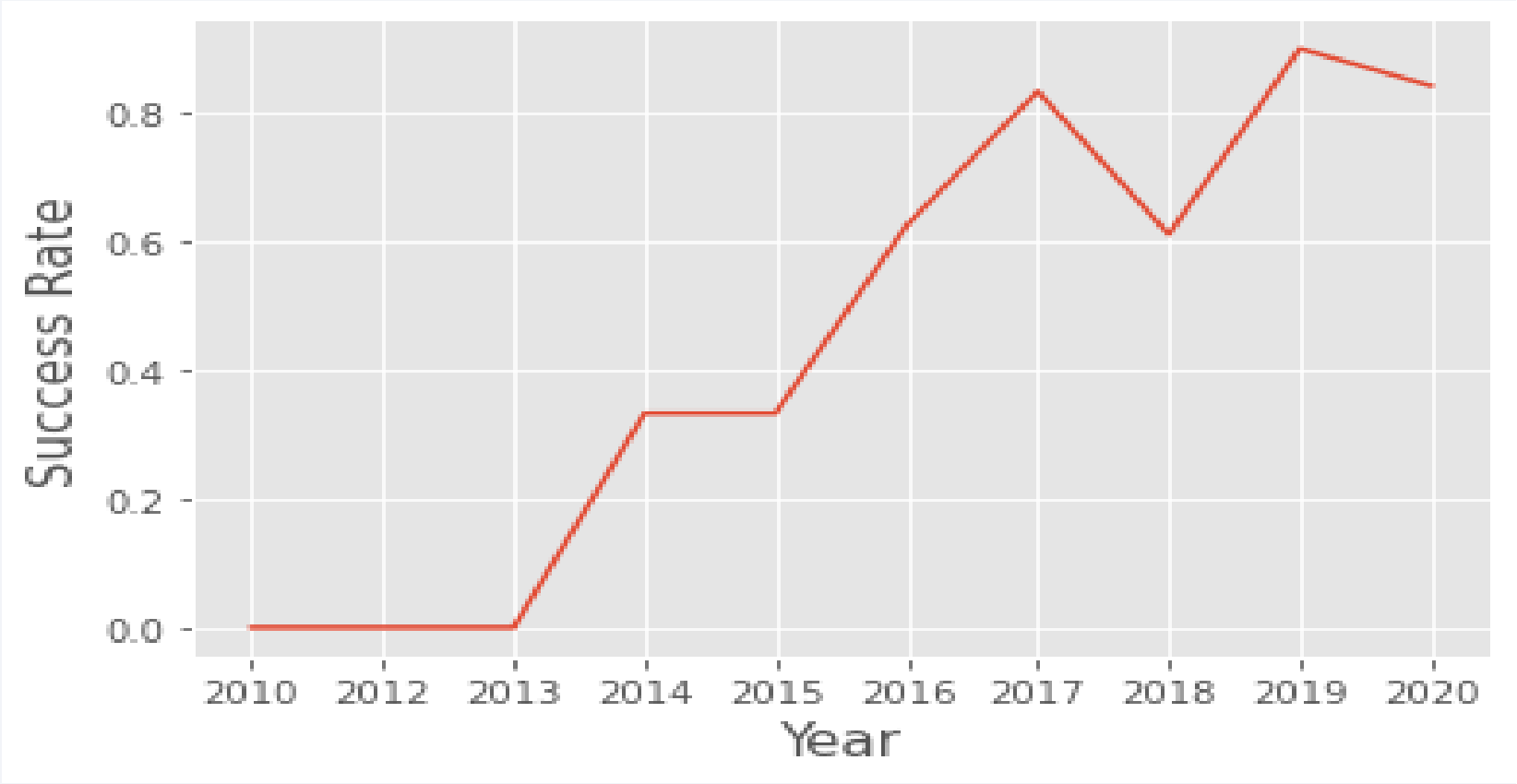
Explanation:

- Heavy payloads have a negative influence GTO orbits and positive on GTO and polar LEO (ISS0 orbits.

Launch Success Yearly Trend

Explanation:

- There is an upward trend in the success rate since 2013 till year 2019.



All Launch Site Names

```
1 %sql select distinct launch_site from SPACEXDATASET;
```

```
 ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
one.
```

launch_site
CCAFS LC-40
CAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Explanation:

Query unique launch site names from database

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Explanation:

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

Total Payload Mass

```
1 %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

total_payload_mass
45596

Explanation:

- Showing the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
1 %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/blud
Done.
```

average_payload_mass
2534

Explanation:

Displaying average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
1 %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

first_successful_landing
2015-12-22

Explanation:

Listing the date of the first successful ground landing

Successful Drone Ship Landing with Payload between 4000 and 6000

```
1 %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
one.
```

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

Explanation:

Displaying the name of the boosters drone ship landing with payload between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

```
1 %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludl
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Displaying the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
1 %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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

Explanation:

Listing the boosters carried maximum payload

2015 Launch Records

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:

Listing the failed landing outcomes in drone ship, their booster versions and launch site names in 2015

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

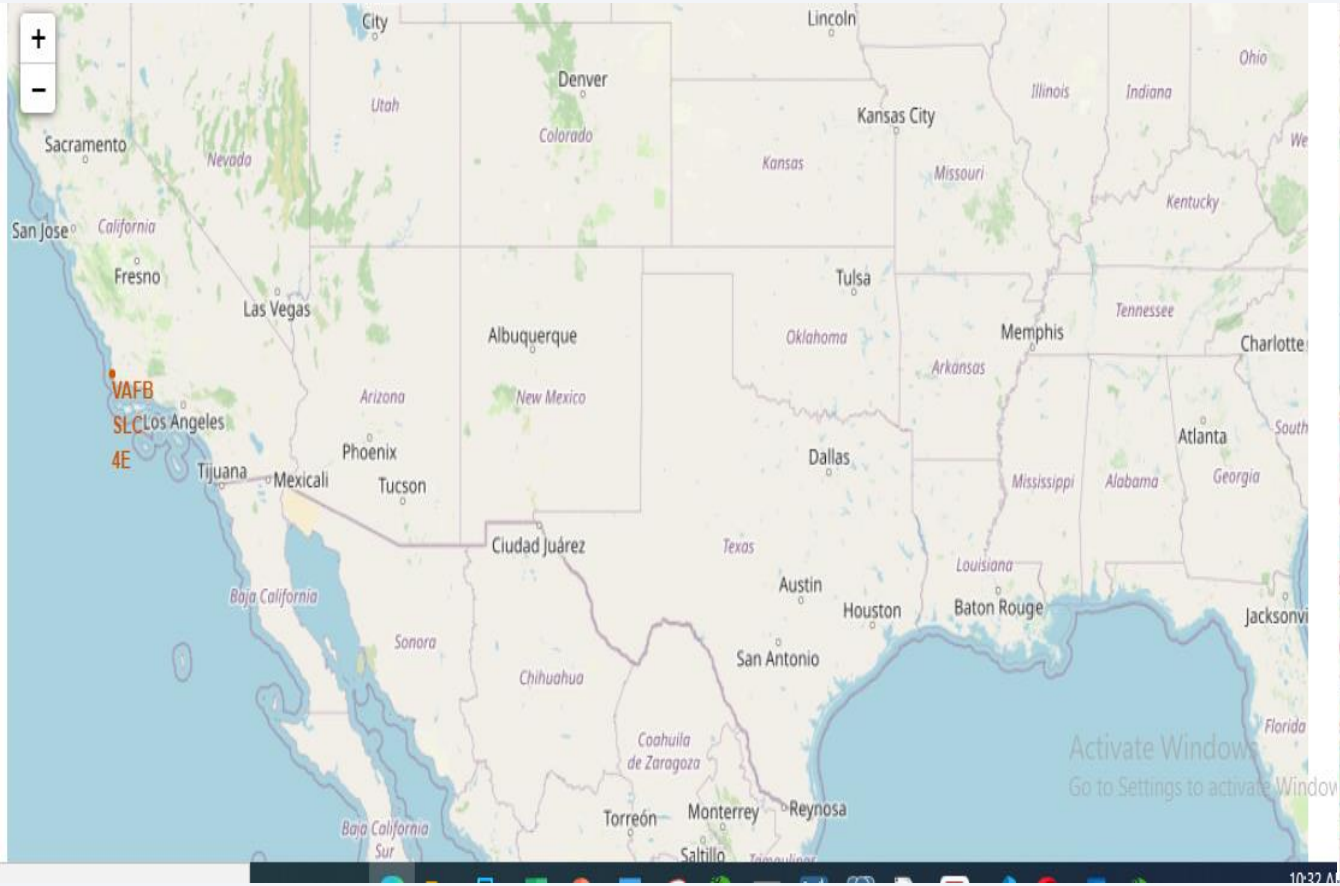
Explanation:

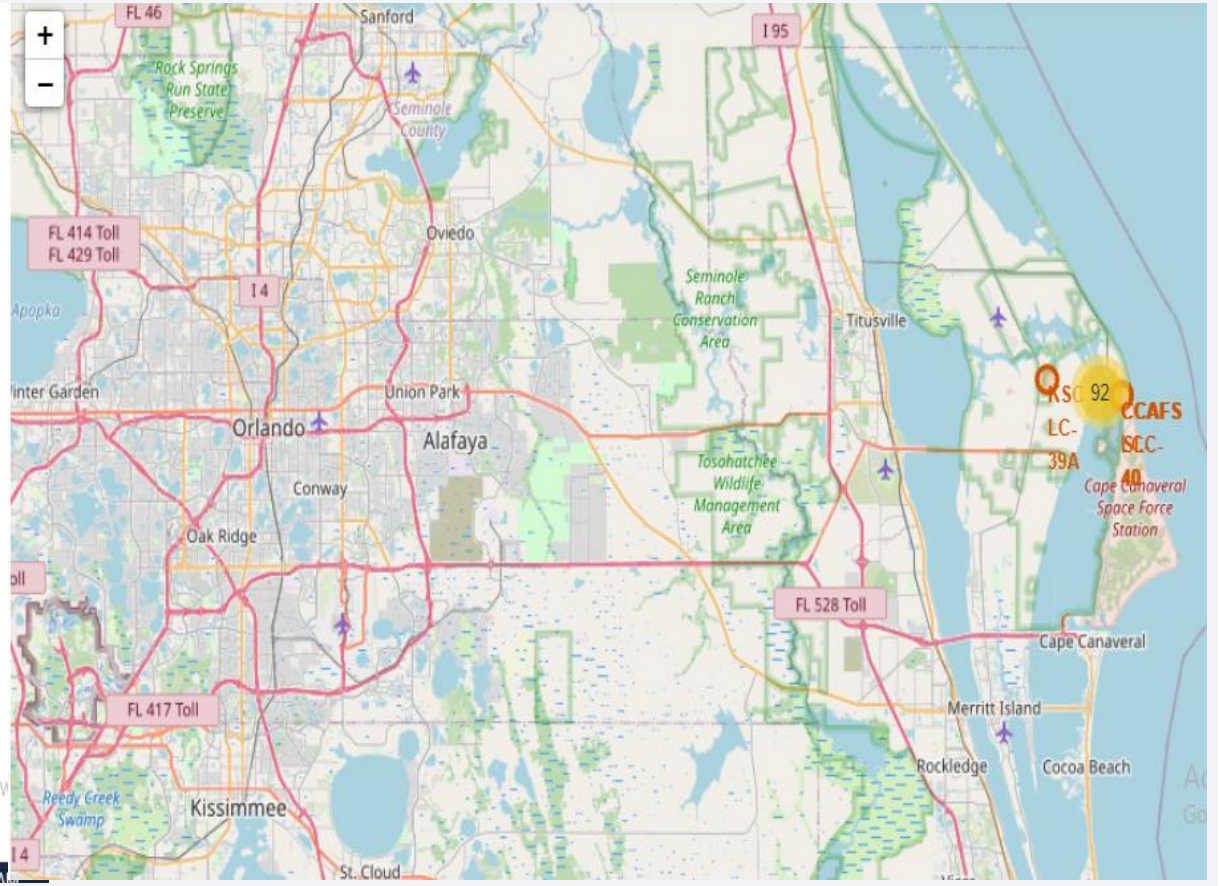
Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Section 4

Launch Sites Proximities Analysis

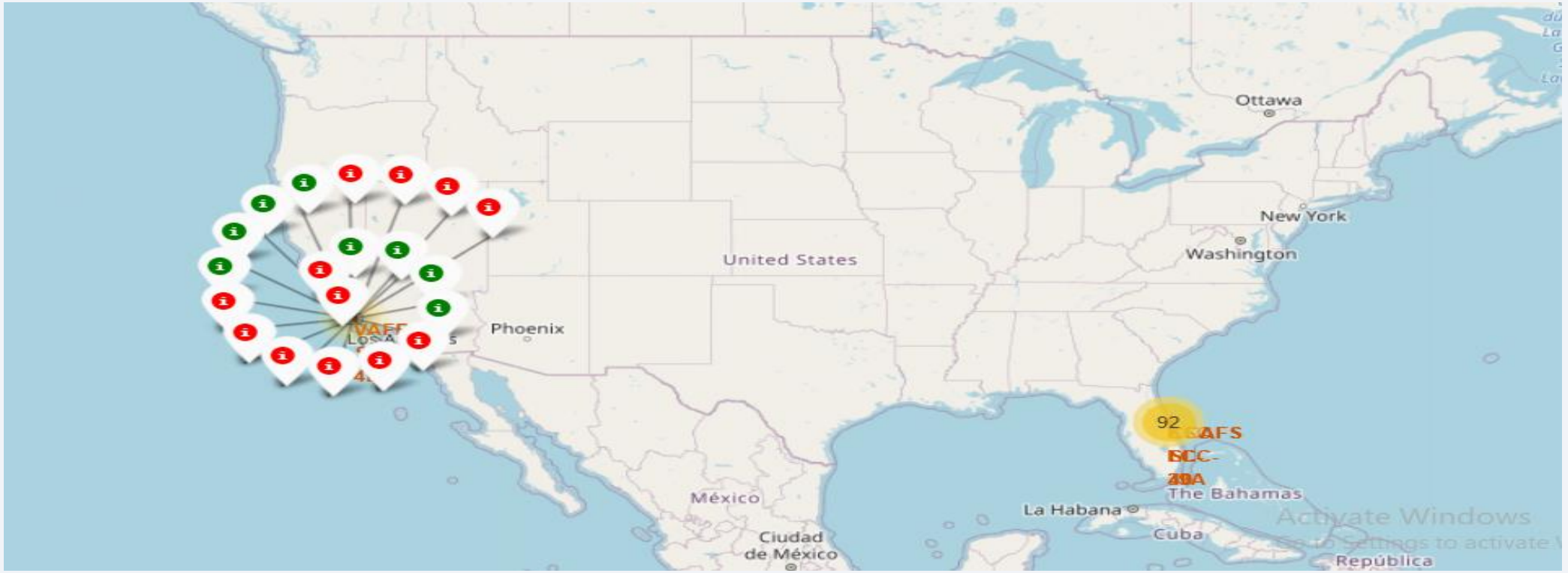
Launch Sites Locations





The left map shows all launch site relative US map. The map on the right shows the two Florida launch sites since their proximity to each other is close. All lunch sites are close to the ocean.

Color- labeled launch records on the map



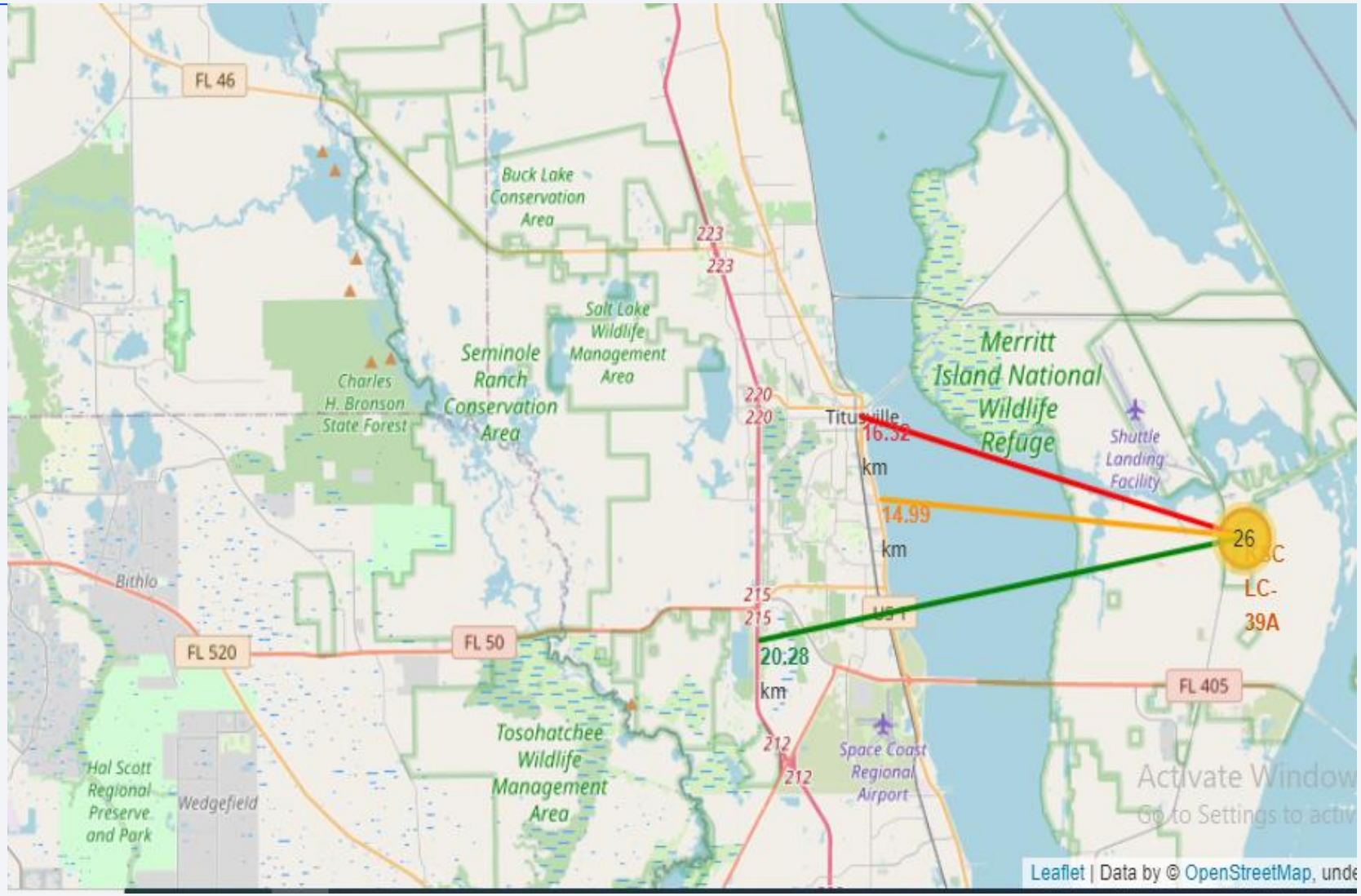
Explain:

We can click on the clusters on Folium map to show each successful landing (represented by the green icon) while failure is represented by the red icon.

Distance from the lauch site KSC LC-39A to its proximities

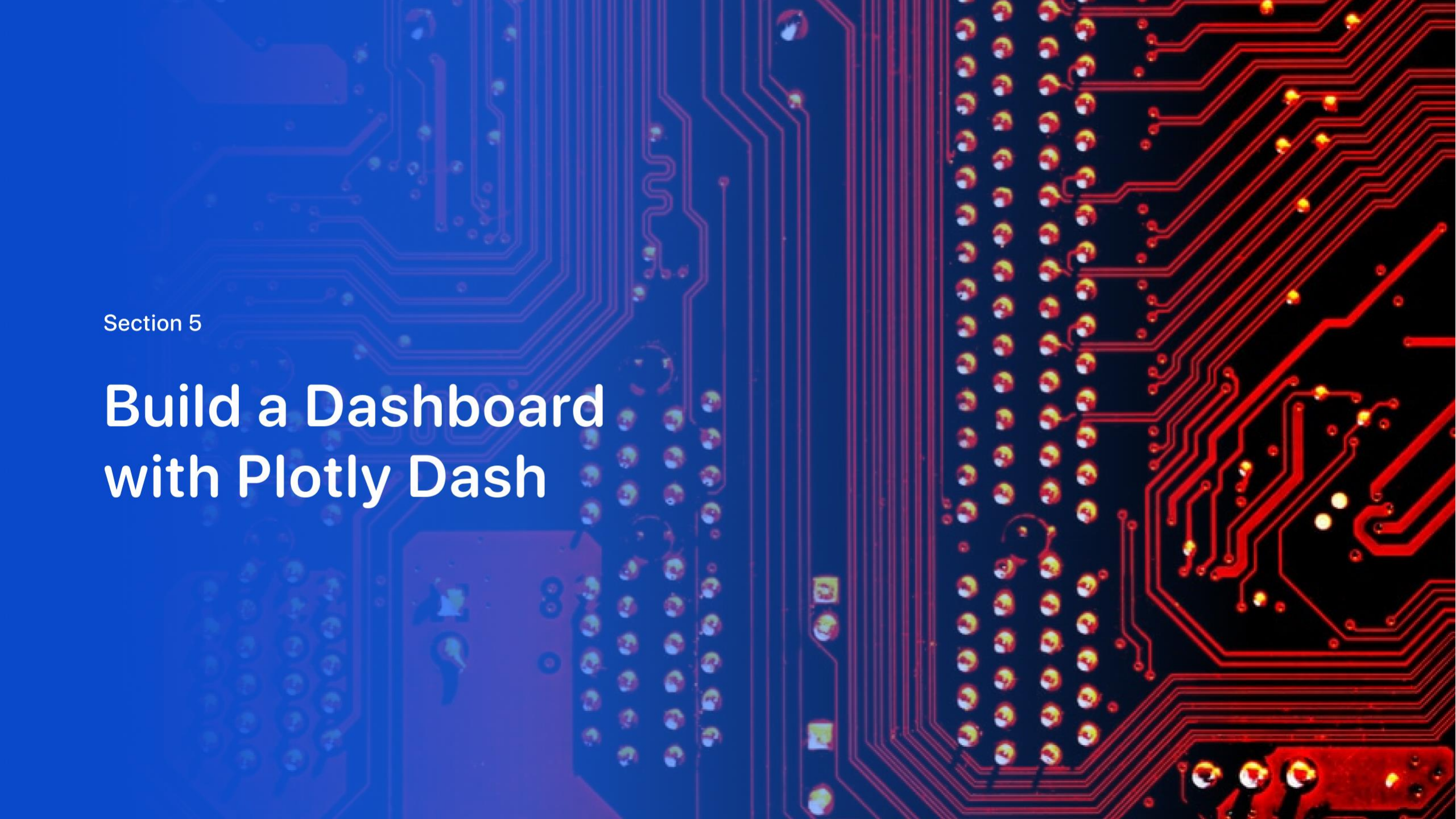
Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

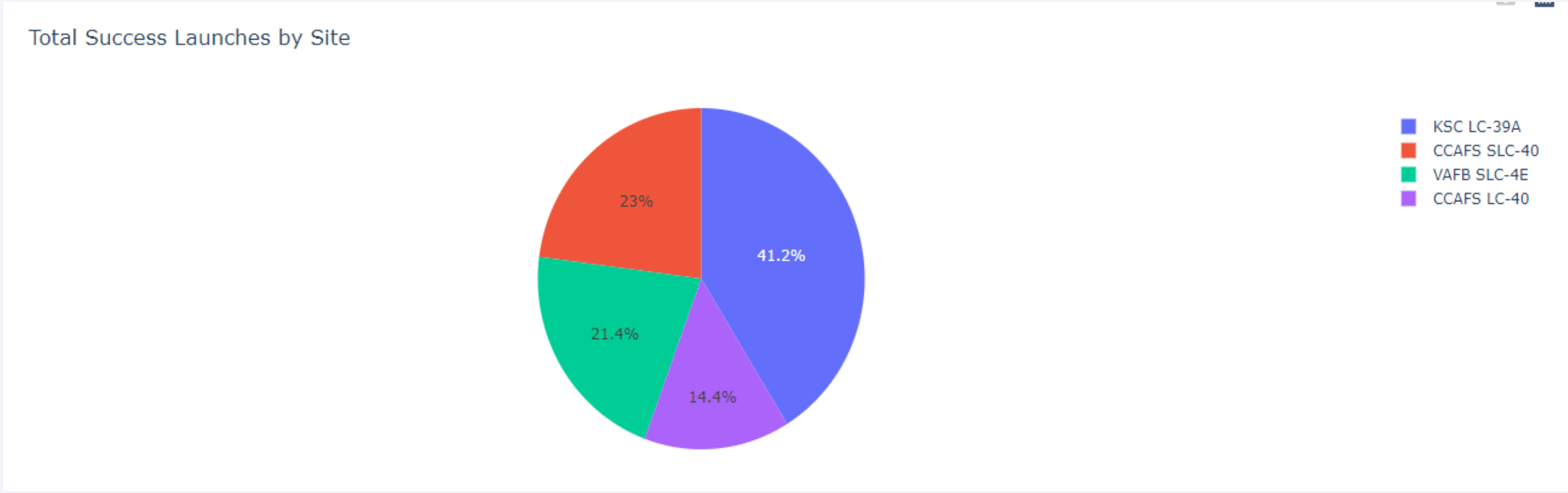


Section 5

Build a Dashboard with Plotly Dash



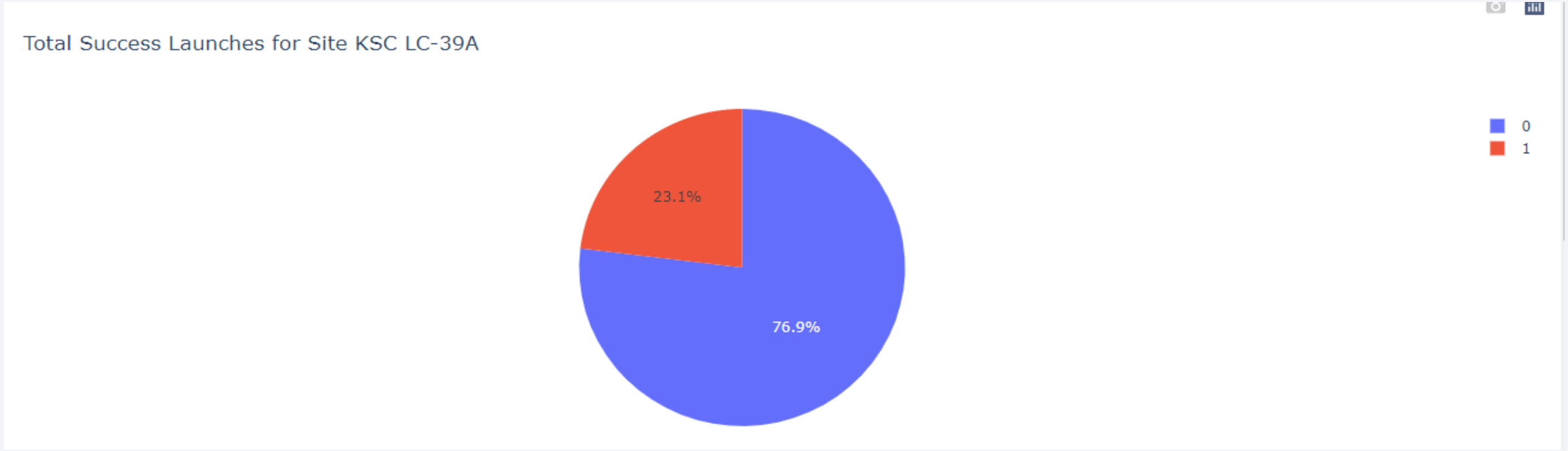
Launch success count for all sites



Explanation:

It can been seen clearly from the graph that KCS LC-39 has the highest rate of successful launches.

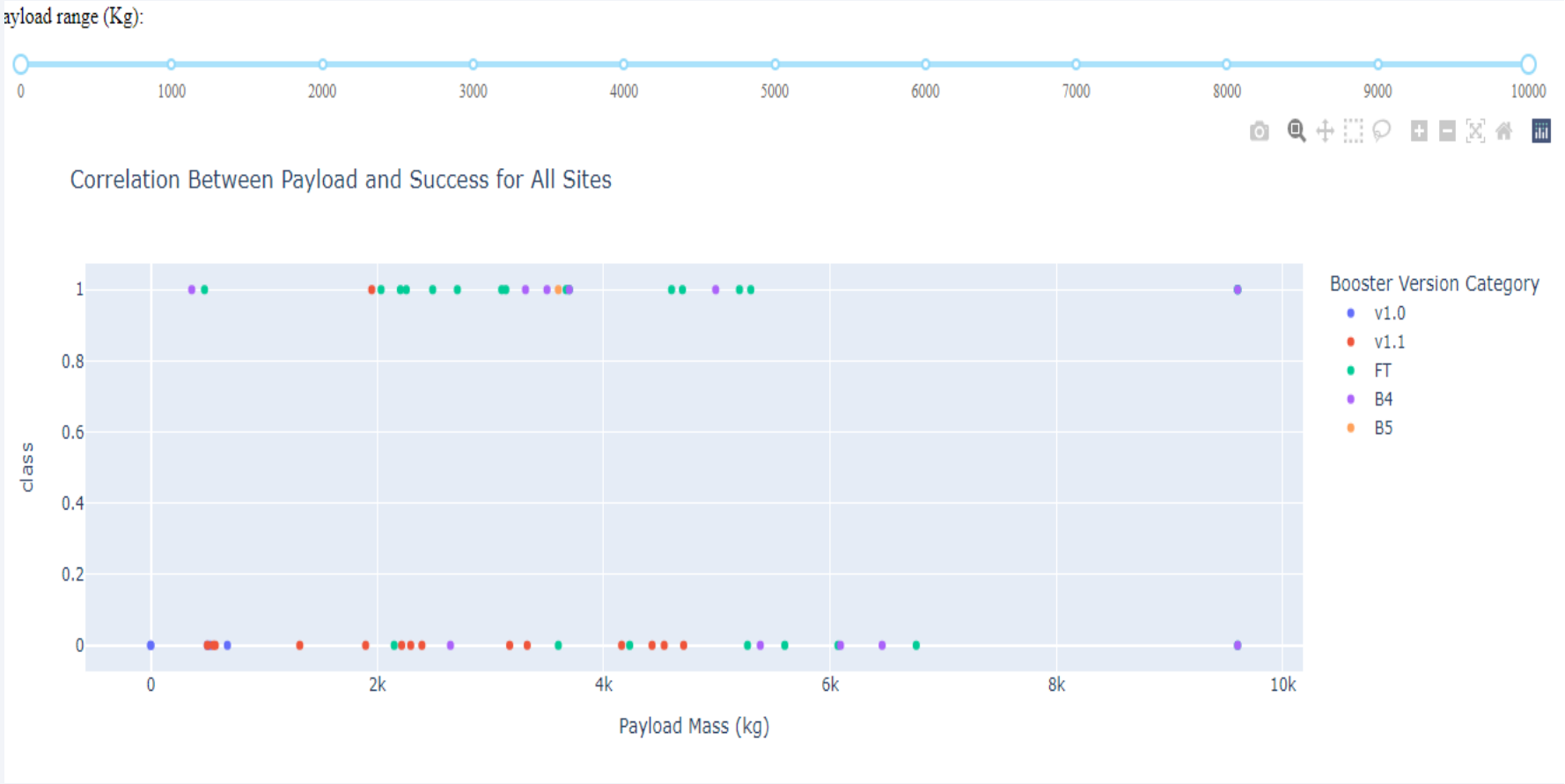
Launch site with highest launch success ratio



The success rate of KCS LC-39A is the highest (79.9%) with 10 successful and only 3 failed landings.

Payloads Mass vs. Launch Outcomes for all sites

Explanation:
The charts reveals that payload between 2000 kg and 5500 kg have the highest success rate



Section 6

Predictive Analysis (Classification)

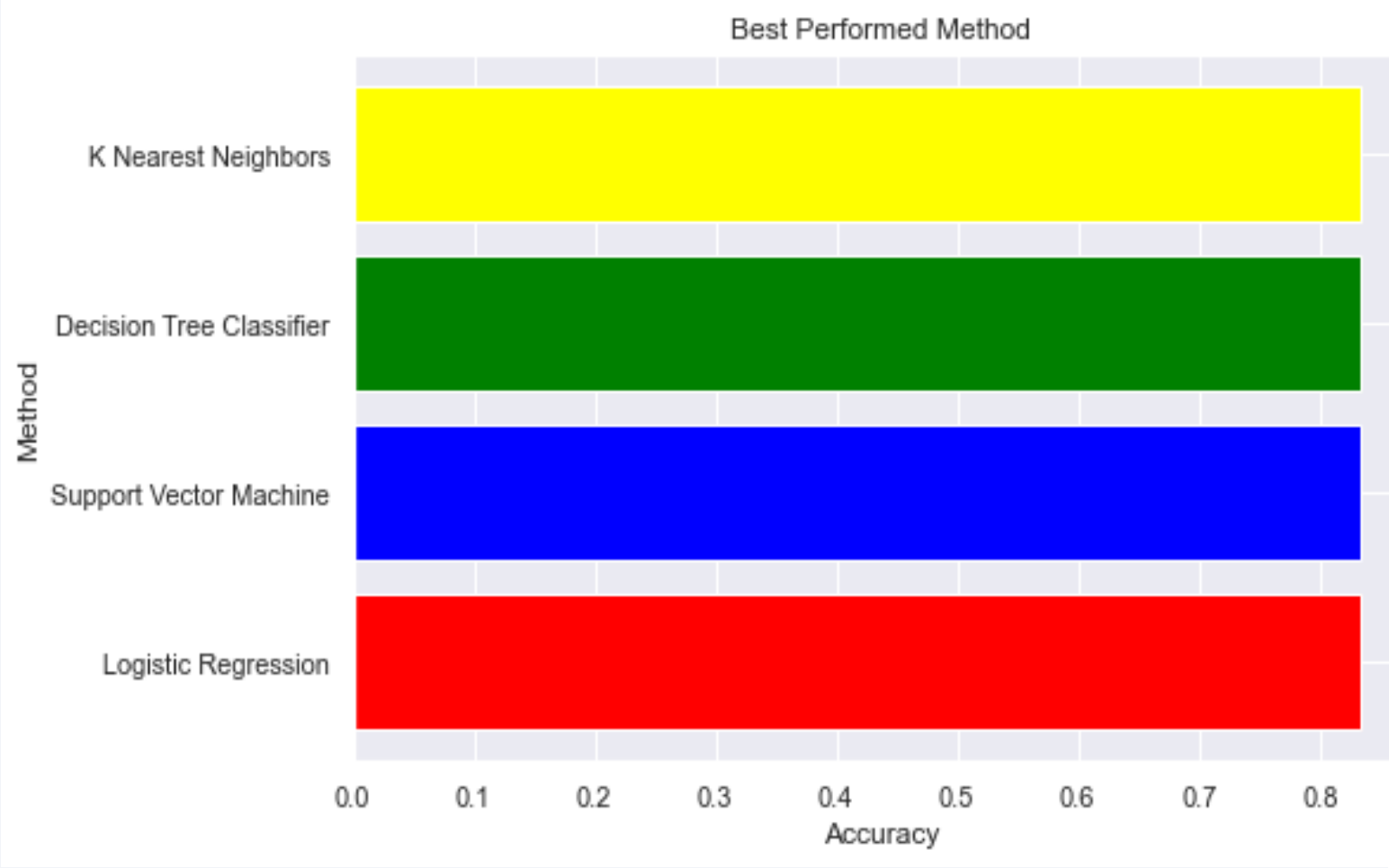
Classification Accuracy

Explanation:

The accuracy for all models is relatively equal on the test data set and it is approximately 83.33 percent. We need to note that the test size of 18 is small.

This can lead to large variance in accuracy results, such as those in Decision tree classifier model in repeated runs.

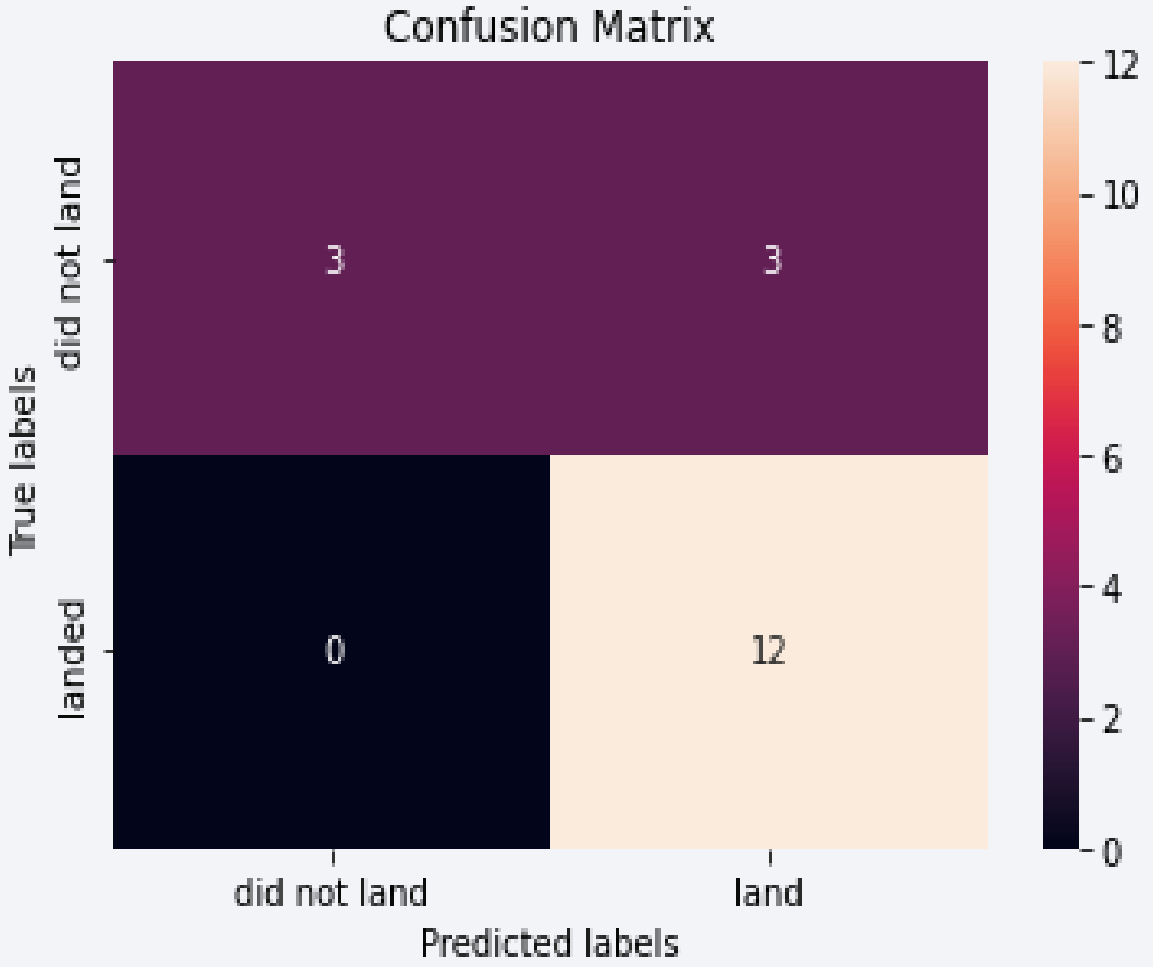
Therefore, we may need more data to determine the best model.



Confusion Matrix

Explanation:

By carefully examining the confusion matrix of logistic regression, it seems that it can differentiate between different classes.



Conclusions

- Decision tree model is the best algorithm for this dataset
- Launches with a low payload mass show better results than those with a larger payload mass
- Most of all launch sites are in proximity to the equator line and all the sites are very close to the coast.
- Over the years, the success rate has been increasing.
- KSC LC-39A has the highest success rate of the launches from all the sites
- Orbits ES-L1, GEO, HEO and SSO HAVE 100% success rate.

Appendix

Instructors:

Rav Ahuja, Alex ZKLson, Aije Egwaikhide, Svetlana Levitan, Romeo Kiezler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorogi, Yan Luo.

Special Thanks to All instructors: