Data Science Capstone Project

Rijeesh Keloth

January 08, 2024

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 all results

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

INTRODUCTION

Context and Background of the Project:

SpaceX, a leader in commercial space exploration, has revolutionized affordability with Falcon 9 rocket launches priced at $62 million, a fraction of competitors' costs reaching $165 million. The key to this success is SpaceX's pioneering strategy of reusing the first stage of its rockets. Anticipating the successful landing of this stage is crucial in estimating launch costs. Our goal is to use public information and machine learning models to predict whether SpaceX will reuse the first stage, considering factors like payload mass, launch site, flight numbers, and orbital characteristics.

Key Inquiries to Address:

- What impact do variables such as payload mass, launch site, number of flights, and orbital characteristics have on the likelihood of a successful first stage landing?
- Is there a discernible trend indicating an increase in the rate of successful landings over the years?
- Which algorithm proves most effective for binary classification in predicting the reuse potential of the first stage in this particular context?

METHODOLOGY

Data collection methodology:

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

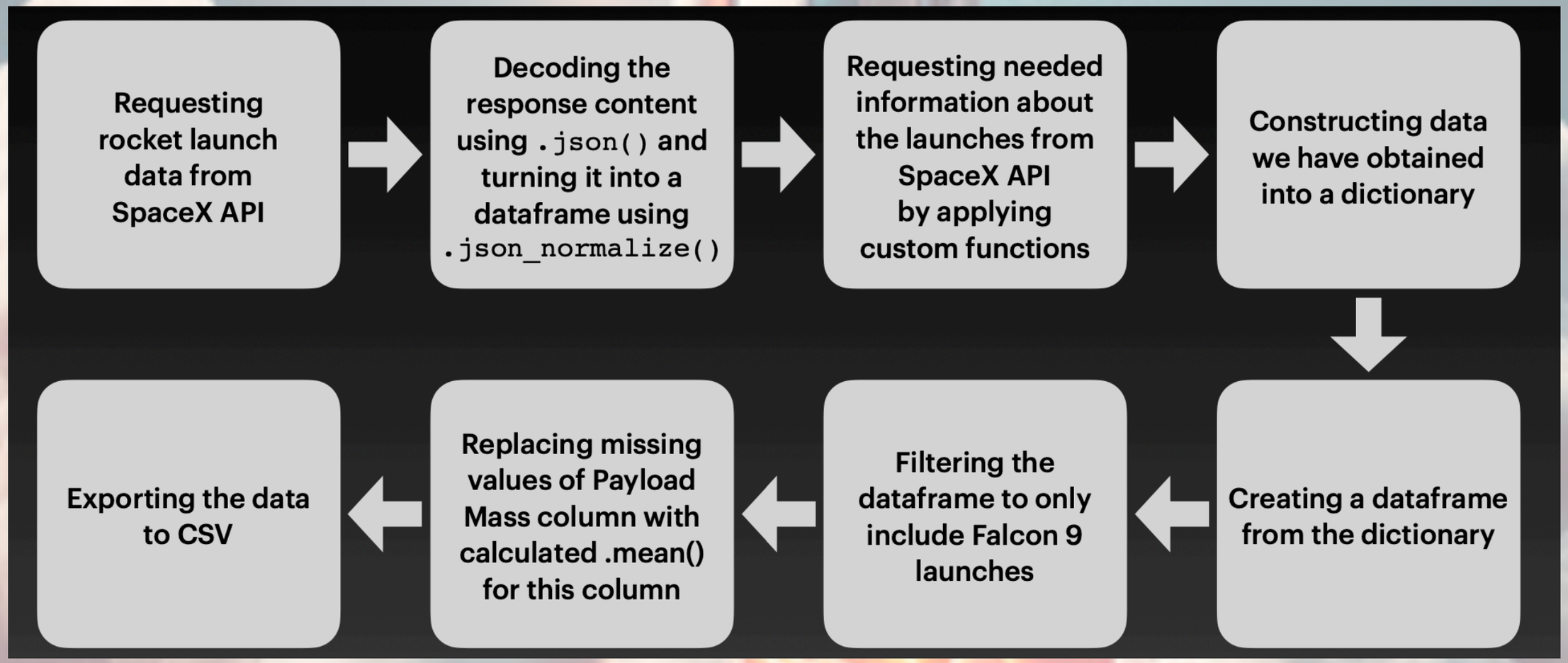
- Building, tuning and evaluation of classification models to ensure the best results

METHODOLOGY

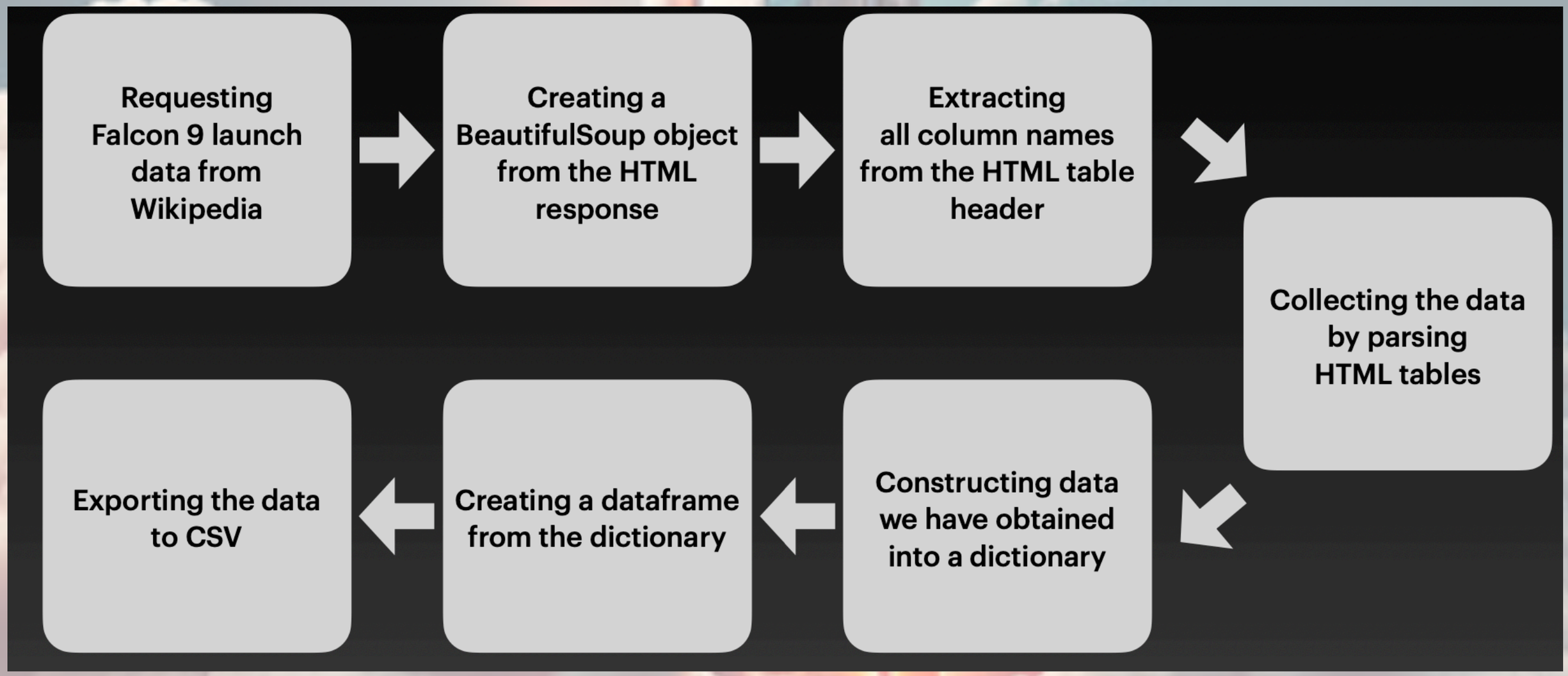
DATA COLLECTION

- Data collected through a dual approach: SpaceX REST API and Wikipedia web scraping.
- Both methods employed for a comprehensive dataset, ensuring detailed launch analysis.
- SpaceX REST API yielded columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- Wikipedia web scraping contributed columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.
- Combined dataset provides a holistic view of SpaceX launches.

DATA COLLECTION- SPACEX API

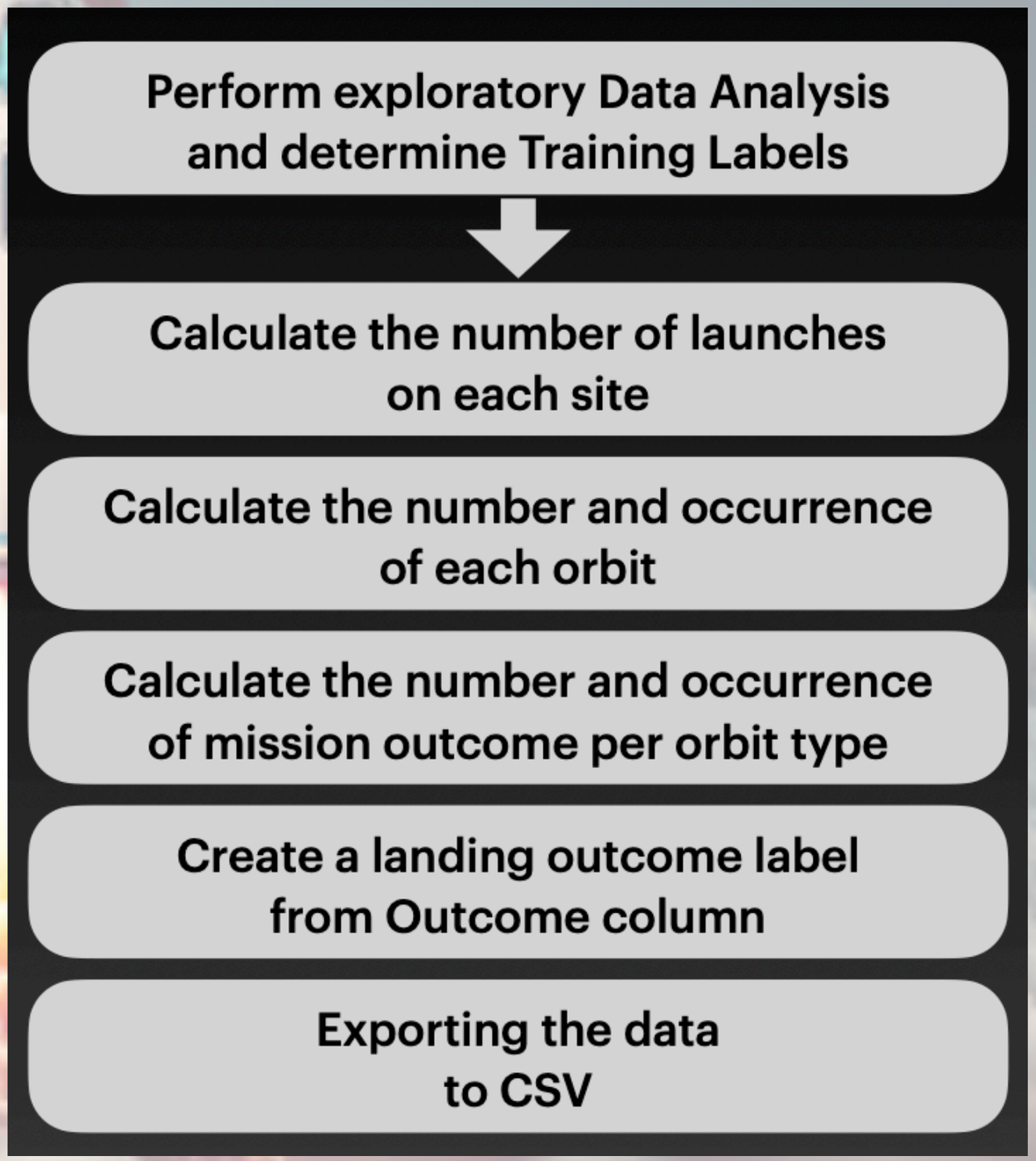


WEBSCRAPING



DATA WRANGLING

- The dataset contains information about SpaceX booster landings, where landing outcomes are categorized based on different conditions, such as landing in the ocean, on a ground pad (RTLS), or on a drone ship (ASDS).
- The outcomes are converted into binary training labels: "1" for successful landings and "0" for unsuccessful landings. This simplification facilitates the analysis of booster landing success across various scenarios.



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, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend Scatter plots show the 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).

EDA WITH SQL

Performed SQL queries include:

- Displaying unique launch site names in the space mission.
- Showing 5 records with launch sites starting with 'CCA.'
- Revealing the total payload mass carried by NASA (CRS) boosters.
- Displaying the average payload mass carried by booster version F9 v1.1.
- Listing the date of the first successful landing on a ground pad.
- Listing booster names with success on a drone ship, payload mass between 4000 and 6000.
- Displaying total counts of successful and failed mission outcomes.
- Listing booster versions carrying the maximum payload mass.
- Detailing failed landing outcomes on a drone ship, including booster versions and launch site names for 2015.
- Ranking landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

INTERACTIVE MAP WITH FOLIUM

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of 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.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

DASH BOARD WITH PLOTLY

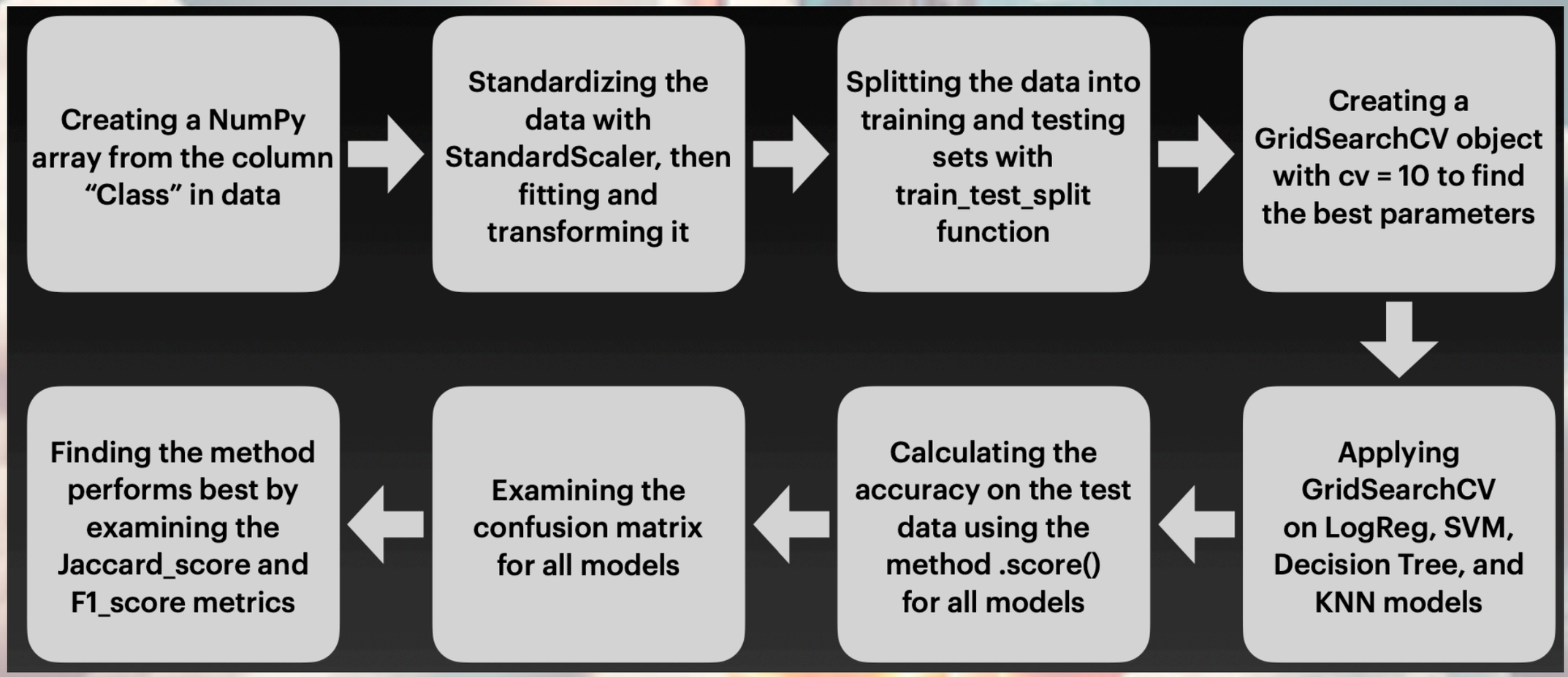
Launch Sites Dropdown List:
- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):
- 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 Site was selected.

Slider of Payload Mass Range:
- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
- Added a scatter chart to show the correlation between Payload and Launch Success.

PREDICTIVE ANALYSIS - CLASSIFICATION



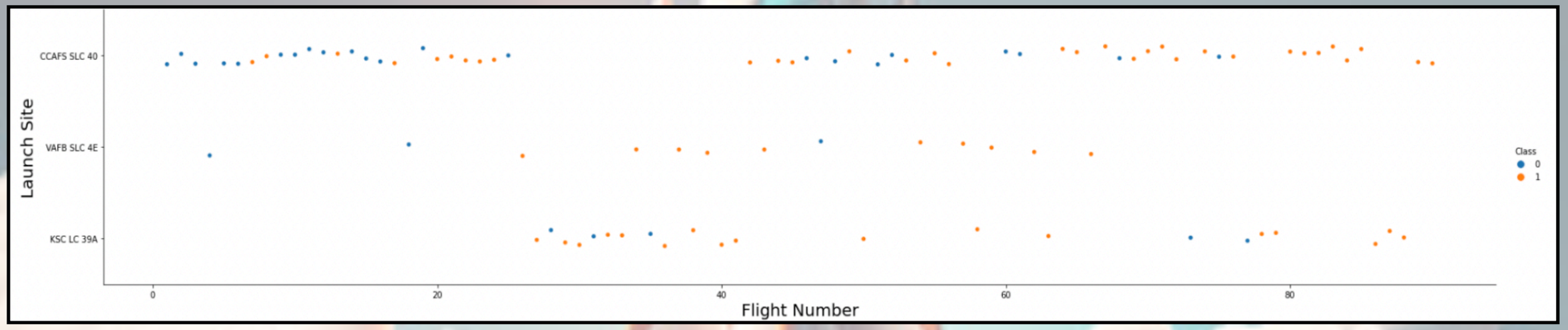
RESULTS

Exploratory data analysis results

- Interactive analytics demo in screenshots
- Predictive analysis results

EDA WITH VISUALISATION

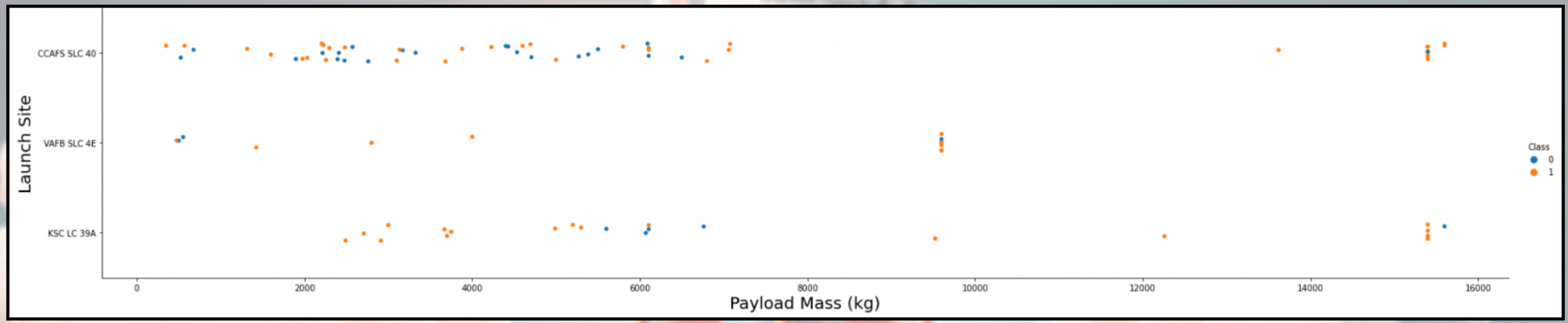
FLIGHT NO. VS LAUNCH SITE



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 LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

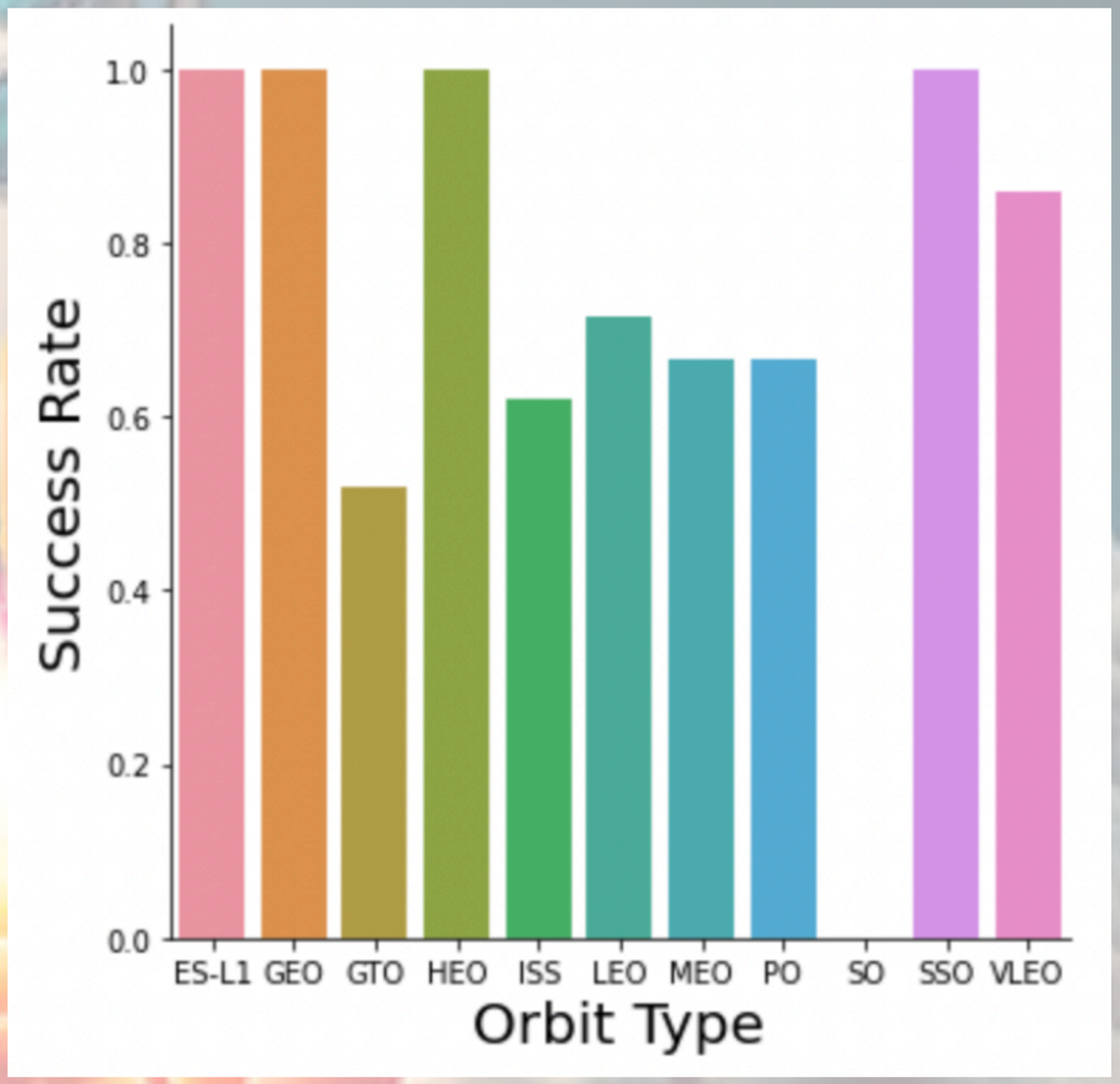
PAYLOAD VS LAUNCH SITE



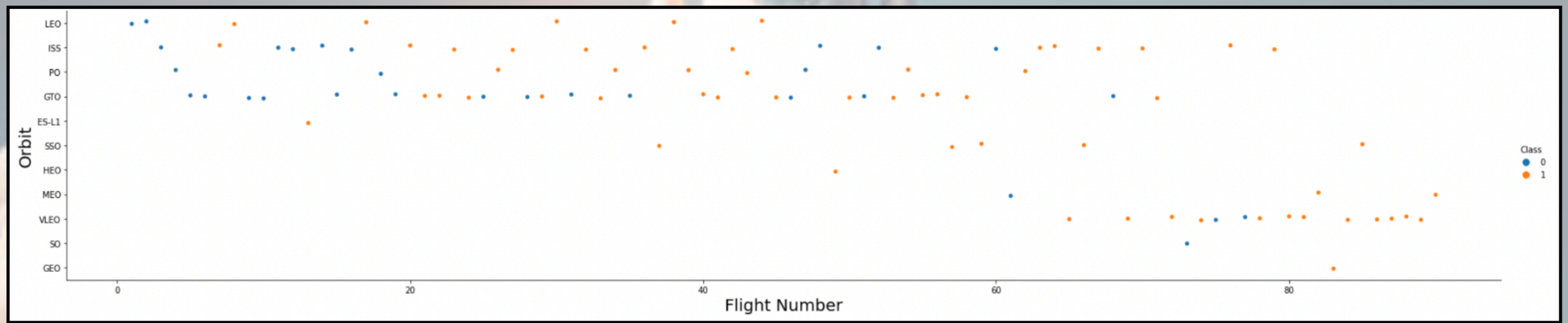
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

SUCCES RATE VS ORBIT TYPE

Orbits with 100% success rate:
- ES-L1, GEO, HEO, SSO
• Orbits with 0% success rate:
- SO
• Orbits with success rate between 50% and 85%:
- GTO, ISS, LEO, MEO, PO

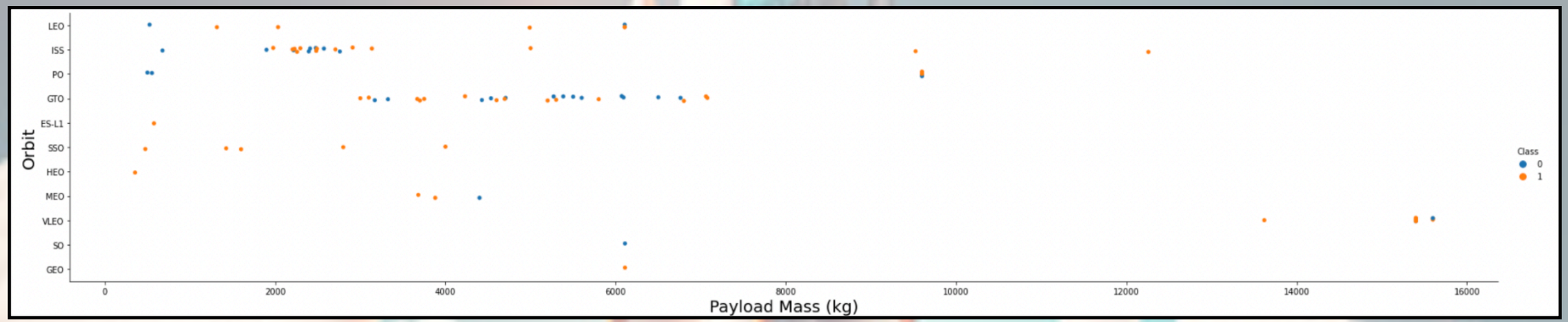


FLIGHT NUMBER VS ORBIT TYPE



In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

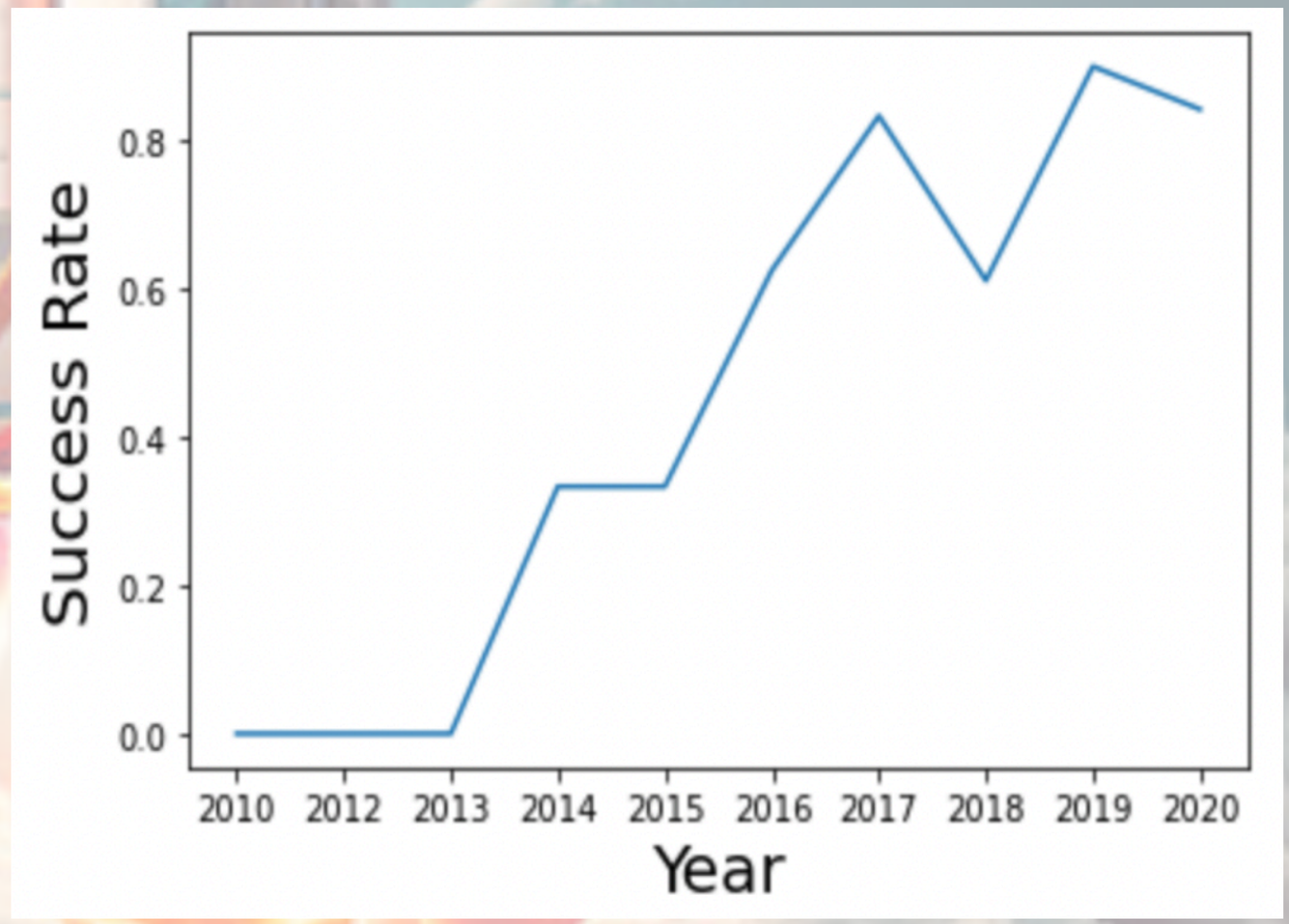
PAYLOAD MASS VS ORBIT TYPE



Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

LAUNCH SUCCESS YEARLY TREND

The success rate Increased from 2013 and kept increasing till 2020.



EDA WITH SQL

Displaying the names of the unique launch sites in the space mission.

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
```

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

Out[4]:

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

Displaying 5 records where launch sites begin with the string 'CCA'.

```
In [5]: %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.
```

Out[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	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

Displaying the total payload mass carried by boosters launched by NASA (CRS).

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
```

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

Out[4]:

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

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

```
In [7]: %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/bludb
Done.
```

Out[7]:

average_payload_mass
2534

Listing the date when the first successful landing outcome in ground pad was achieved.

```
In [8]: %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.
```

Out[8]:

first_successful_landing
2015-12-22

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

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
```

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

Out[9]:

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

Listing the total number of successful and failure mission outcomes.

```
In [10]: %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/bludb
Done.
```

Out[10]:

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

Listing the booster versions carried maximum payload mass

```
In [11]: %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.
```

Out[11]:

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

Listing the failed landing outcomes in drone ship

```
In [12]: %%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.
```

Out[12]:

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)

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

```
In [13]: %%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.
```

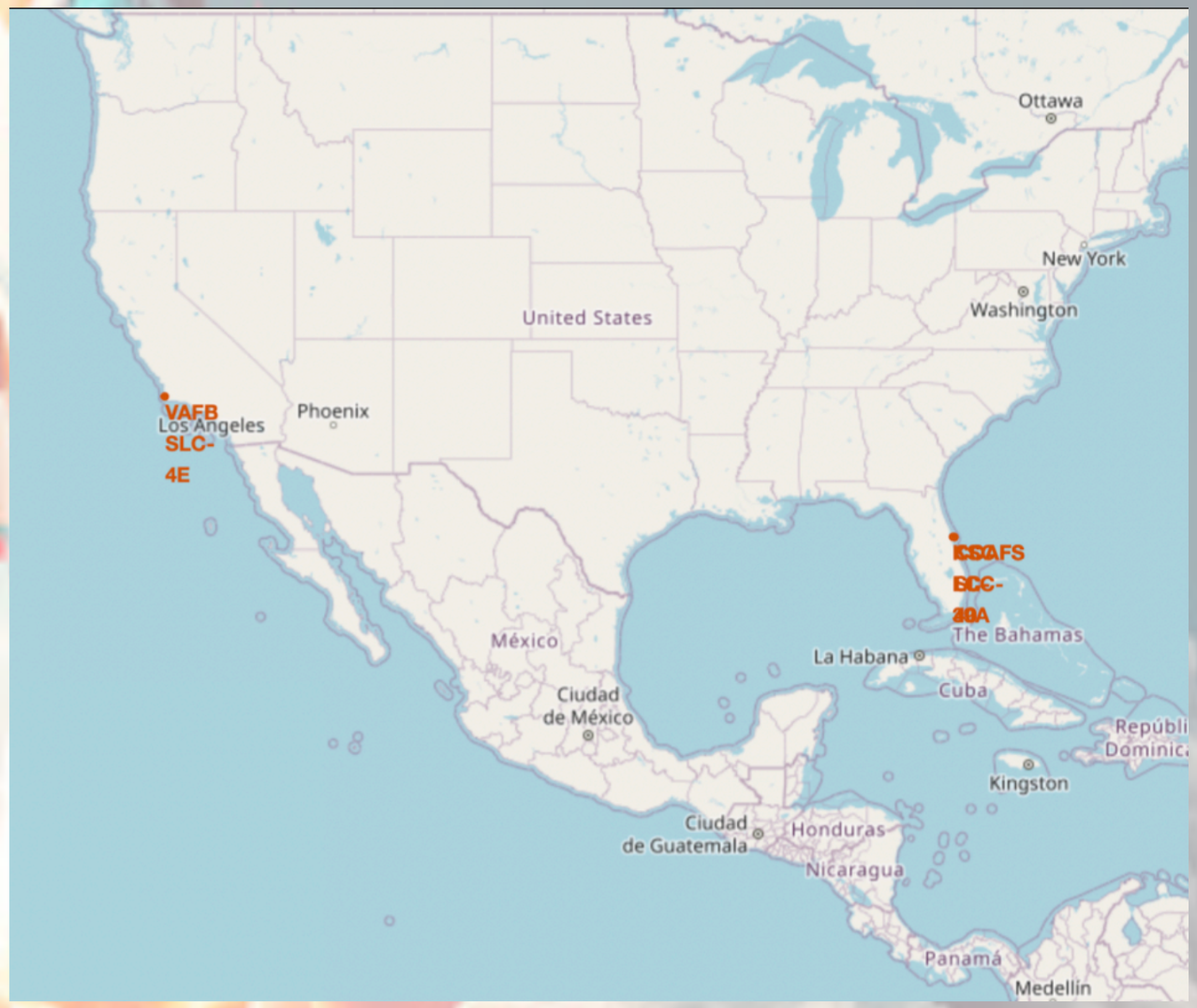
Out[13]:

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

INTERACTIVE MAP WITH FOLIUM

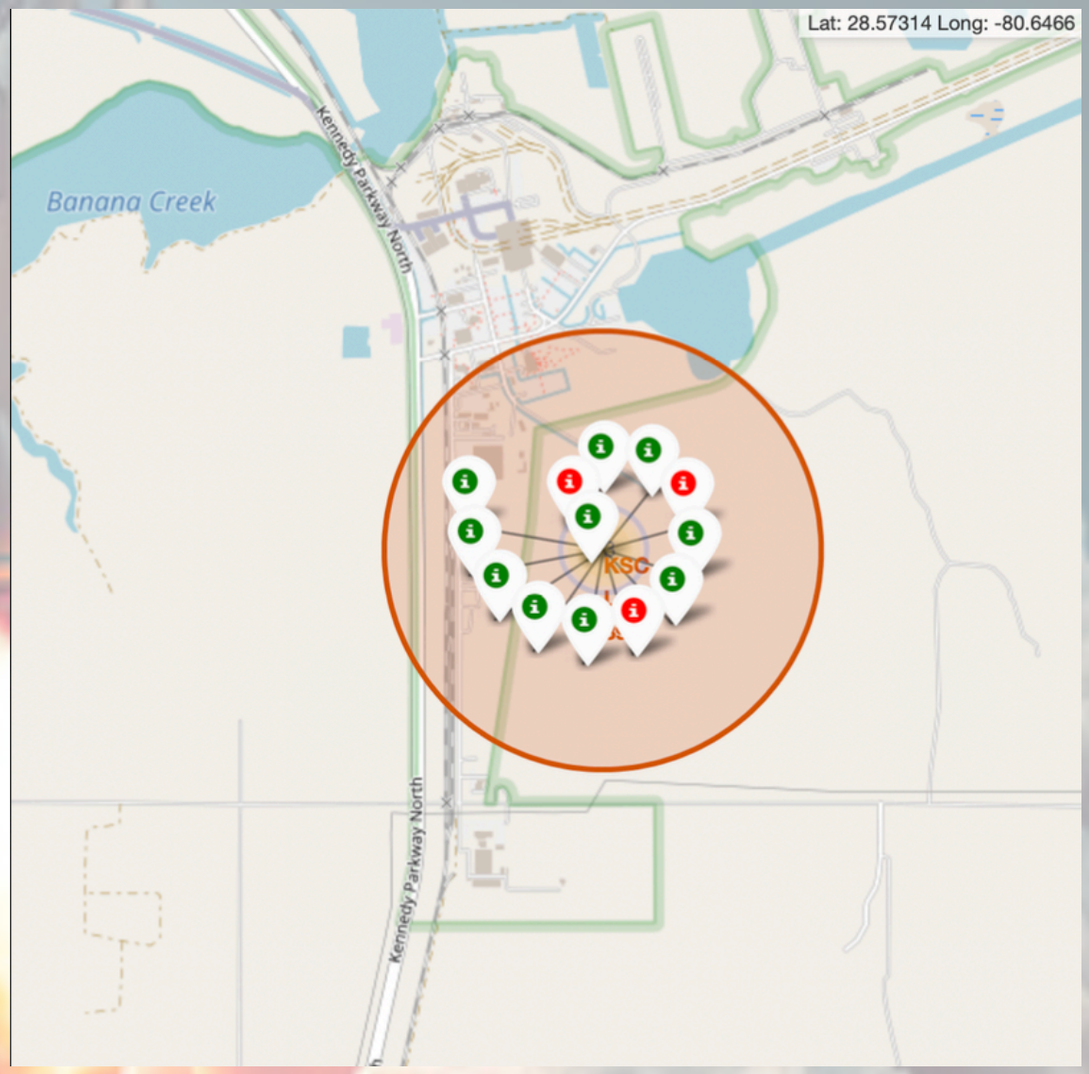
LAUNCH SITE LOCATIONS MARKED

- Launch sites near the Equator capitalize on Earth's rotational speed.
- Objects at the equator move at 1670 km/hour, benefiting spacecraft speed.
- Launching from the equator helps maintain orbital velocity.
- Proximity to coastlines reduces risk of debris near populated areas during launches.



COLOR LABELLED LAUNCH RECORDS ON THE MAP

Green Marker = SuccessfuL Launch
- Red Marker = Failed Launch

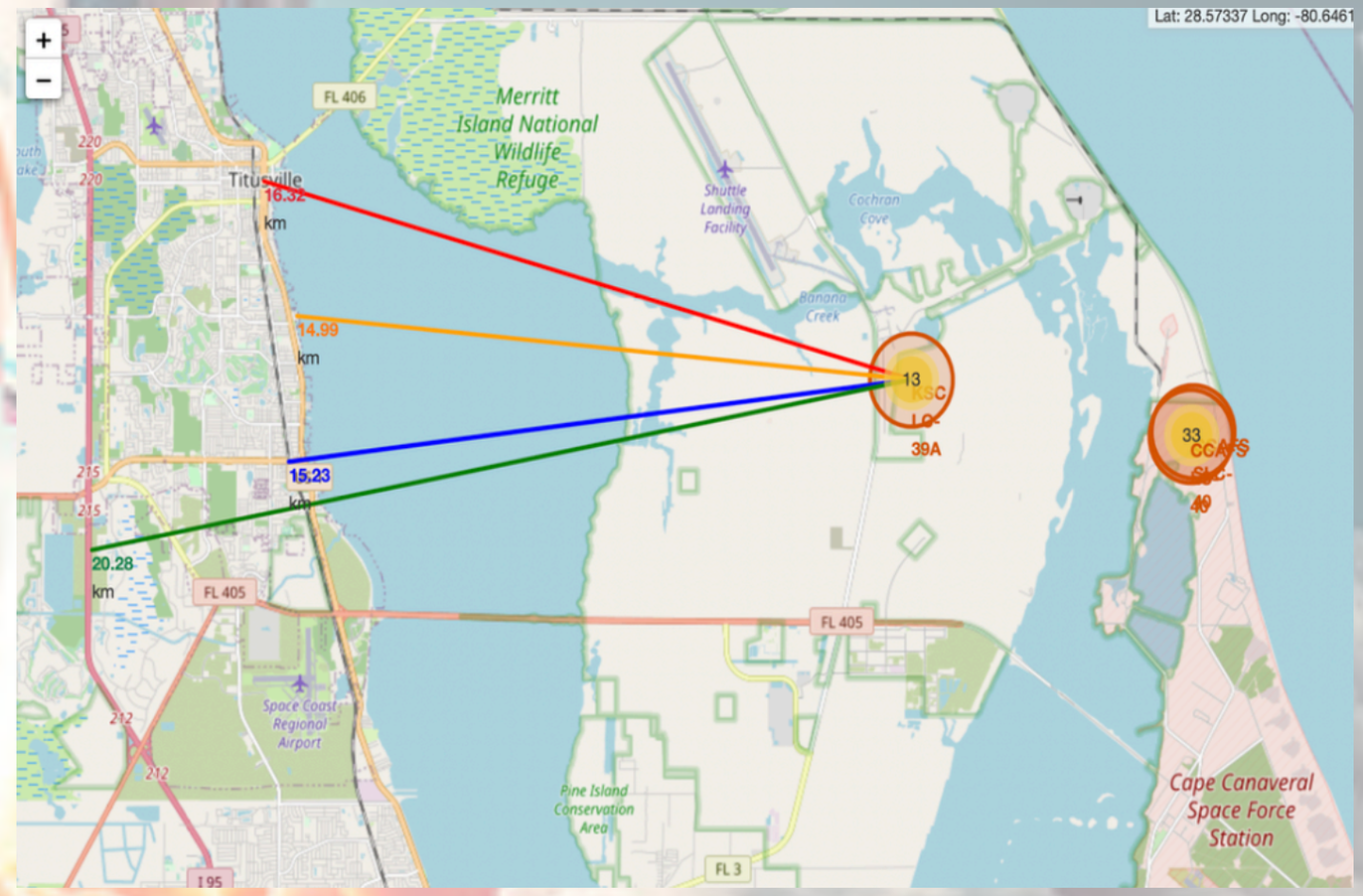


Distance from the launch site KSC LC-39A to its proximities

Upon visually examining the KSC LC-39A launch site, it is evident that:

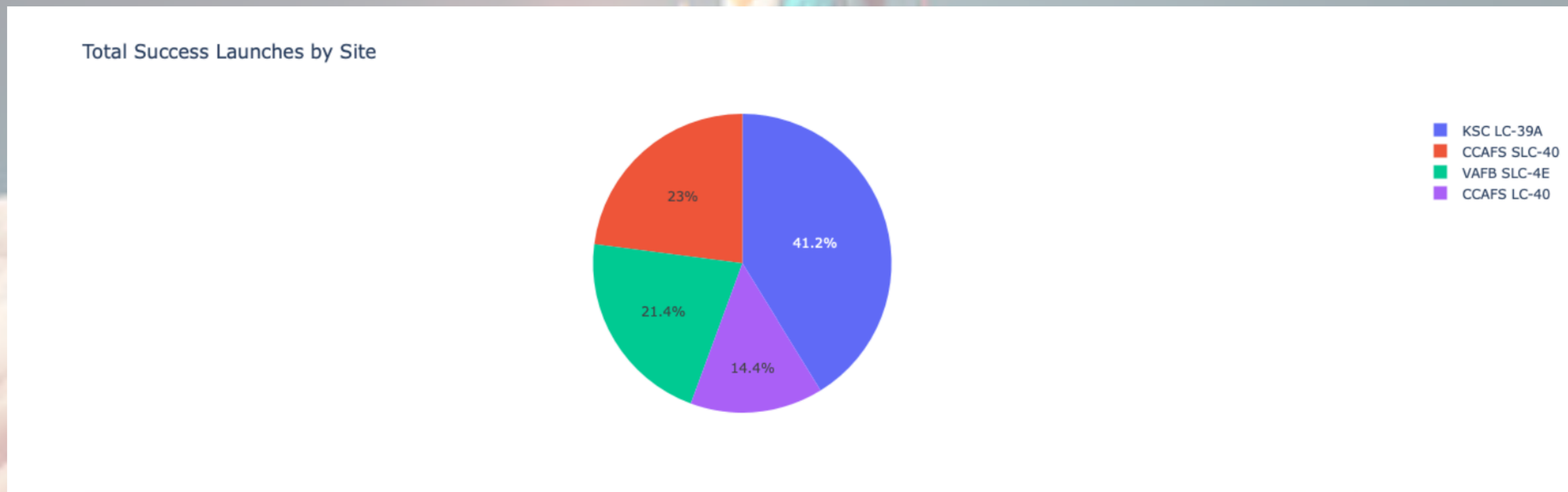
- It is in close proximity to a railway (15.23 km) and a highway (20.28 km).
- It is situated near the coastline (14.99 km).
- The launch site is relatively close to its nearest city, Titusville (16.32 km).

- Given the high speed of a failed rocket, covering distances of 15-20 km in mere seconds, there is a potential safety concern, especially in proximity to populated areas.



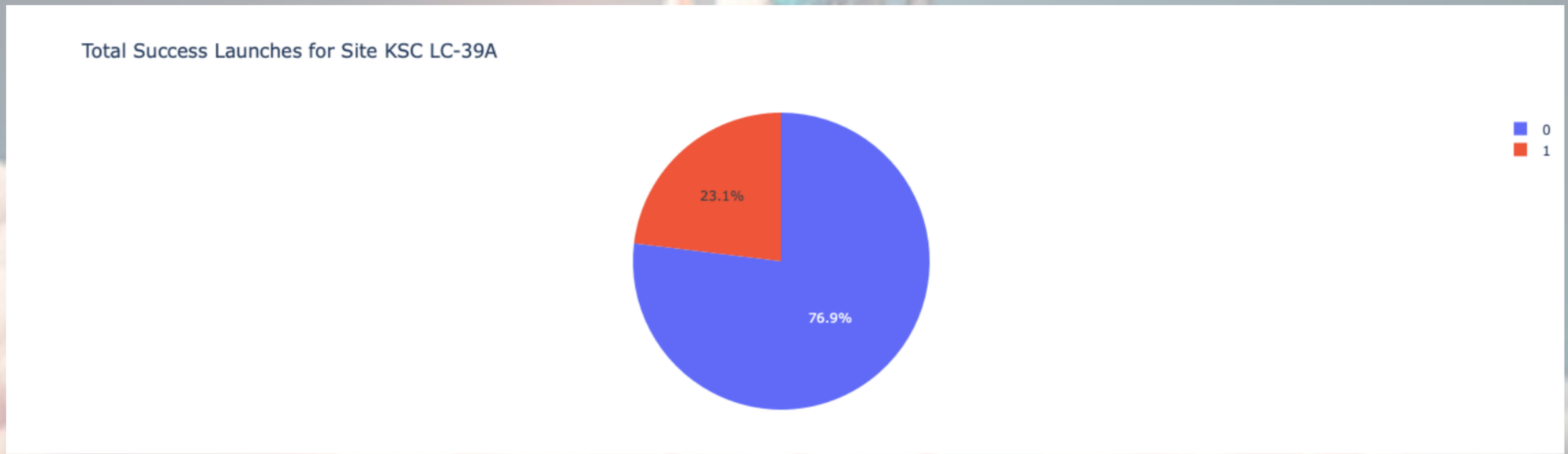
DASH BOARD WITH PLOTLY DASH

SUCCESS COUNT FOR ALL SITES



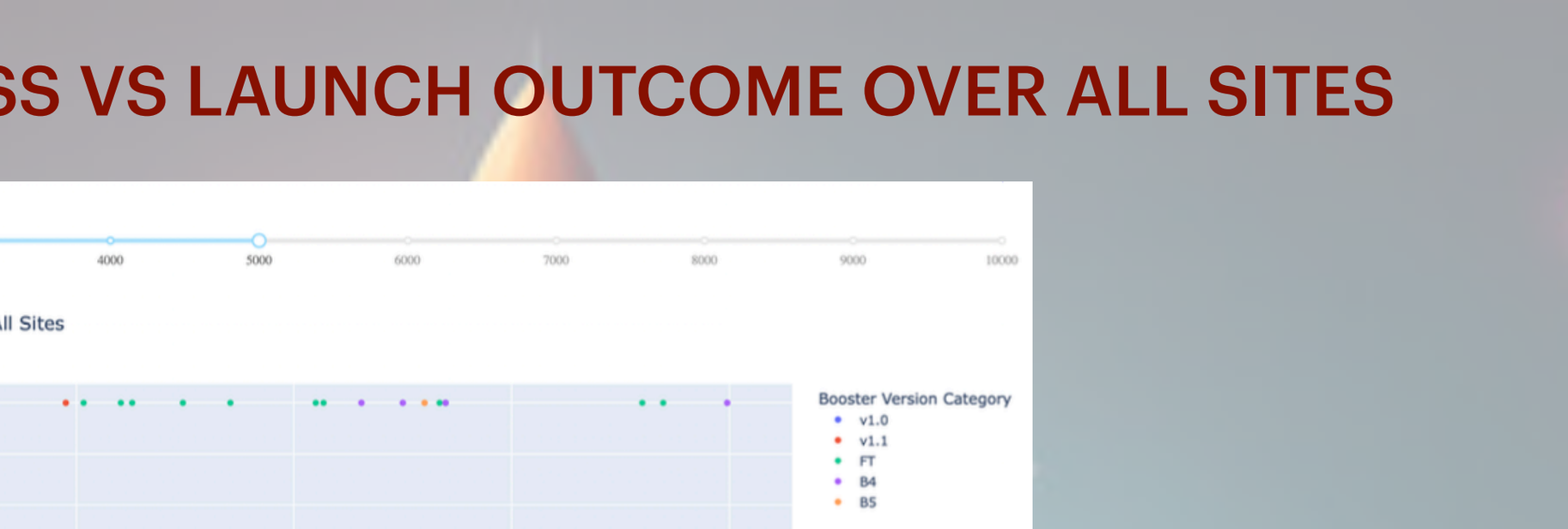
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

LAUNCH SITE WITH HIGHEST SUCCESS RATIO



KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

PAYLOAD MASS VS LAUNCH OUTCOME OVER ALL SITES



- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

PREDICTIVE ANALYSIS

CLASSIFICATION

CLASSIFICATION ACCURACY

Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

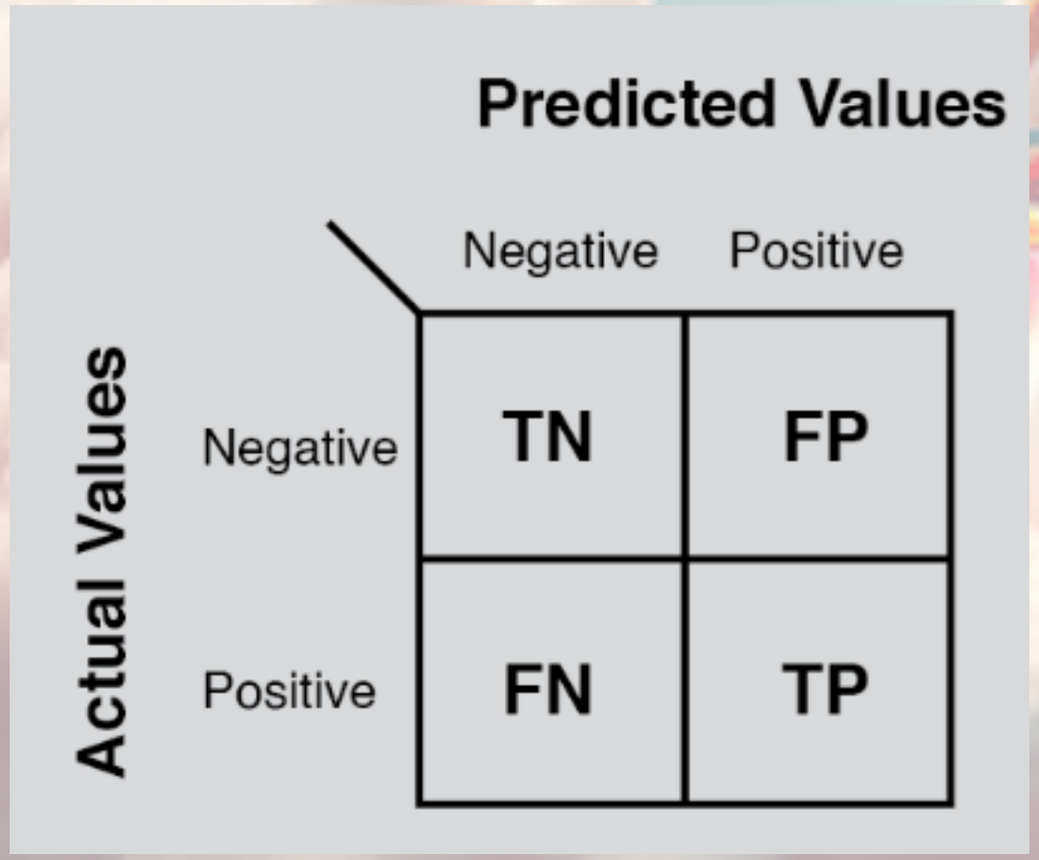
Scores and Accuracy of the ENTIRE Set

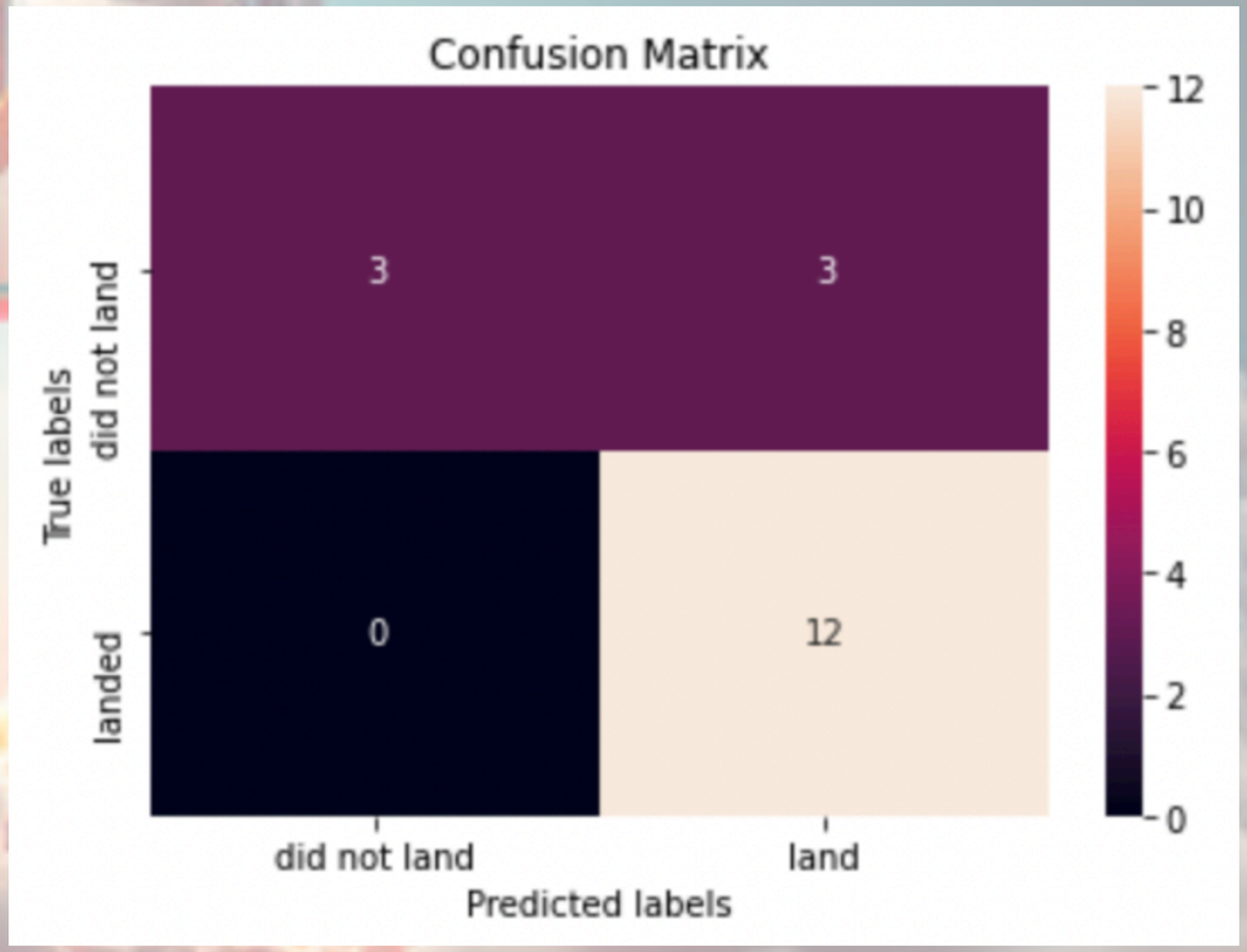
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

The scores of the whole Dataset confirm that the best model is the Decision Tree Model.

CONFUSION MATRIX

Upon analyzing the confusion matrix, it becomes evident that logistic regression effectively differentiates between distinct classes. Notably, the primary challenge observed is the occurrence of false positives.





CONCLUSION

- The Decision Tree Model stands out as the optimal algorithm for this dataset due to the following observations:
- Launches with lower payload mass exhibit superior results compared to those with larger payload mass.
- The majority of launch sites are situated near the Equator line, all in close proximity to the coast.
- The success rate of launches demonstrates an upward trend over the years.
- Among all launch sites, KSC LC-39A boasts the highest success rate.
- Orbits ES-L1, GEO, HEO, and SSO consistently achieve a 100% success rate.

APPENDIX

THANKS TO

COURSERA, IBM AND INSTRUCTORS