DATA SCIENCE CAPSTONE PROJECT

Presented By : Werisson Mendonca

TABLE OF CONTENT

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 results

INTRODUCTION

Background and context of the project the most prosperous business of the commercial space age, SpaceX lowers the cost of space travel. Due in large part to SpaceX's ability to reuse the first stage, the business offers Falcon 9 rocket launches on its website, which cost 62 million dollars, whereas other companies charge upwards of 165 million dollars per. Thus, we can calculate the cost of a launch if we can predict whether the first stage will land. We will make a prediction about whether SpaceX will reuse the first stage based on publicly available data and machine learning techniques.

Questions to be answered:
What effects do factors like orbits, cargo mass, launch location, and flight count have on the first stage landing's success?
Does the number of successful landings rise with time?
Which algorithm works best in this situation for binary classification?

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

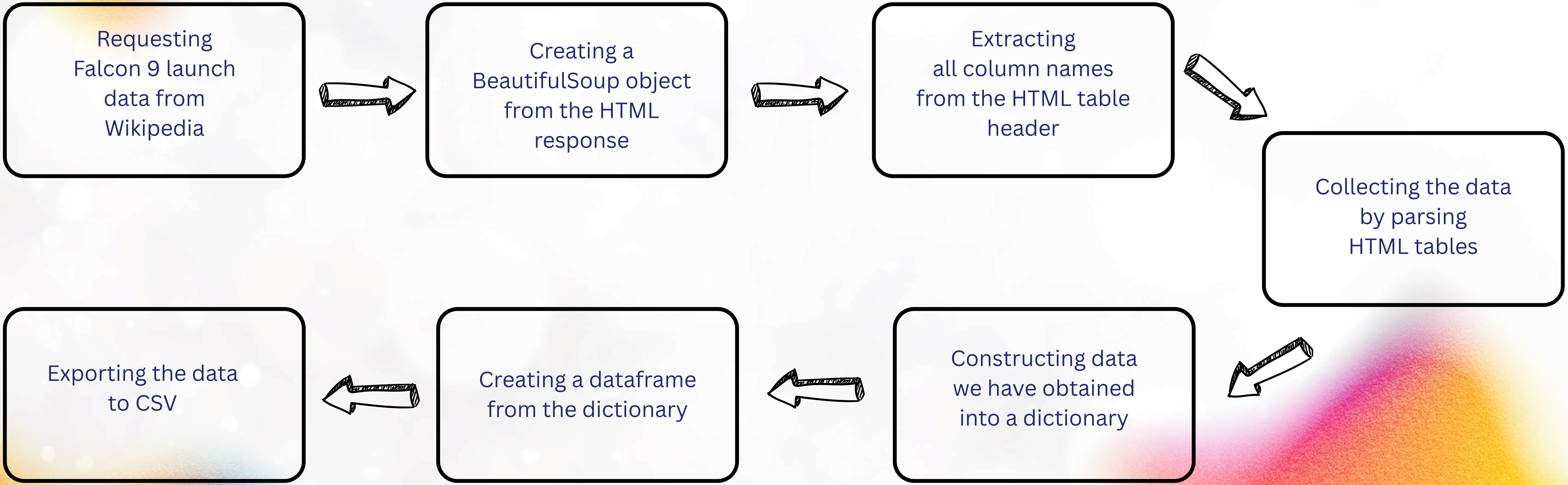
DATA COLLECTION

A combination of web scraping information from a table in SpaceX's Wikipedia entry and API queries from the SpaceX REST API were used in the data collection procedure.
To obtain comprehensive information about the launches for a more thorough analysis, we had to employ both of these data collection techniques.
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude are among the data columns that may be retrieved using the SpaceX REST API.
Wikipedia is utilized to collect the data columns. Flight number, launch site, payload, payload mass, orbit, customer, launch result, version booster, booster landing, date, and time are all scraped from the web.

DATA COLLECTION - SPACEX API

1. Requesting rocket launch data from SpaceX API
2. Decoding the response content using .json() and turning it into a dataframe using .json_normalize()
3. Requesting needed information about the launches from SpaceX API by applying custom functions
4. Constructing data we have obtained into a dictionary
5. Creating a dataframe from the dictionary
6. Filtering the dataframe to only include Falcon 9 launches
7. Replacing missing values of Payload Mass column with calculated .mean() for this column
8. Exporting the data to CSV

DATA COLLECTION - WEB SCRAPING



DATA WRANGLING

Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

In the dataset, booster landing outcomes are categorized to indicate success or failure. These outcomes are used to create training labels, where “1” represents a successful landing and “0” represents a failed attempt. A landing is considered successful if it is marked as True Ocean (successful landing in the ocean), True RTLS (successful landing on a ground pad), or True ASDS (successful landing on a drone ship). Conversely, failed landings are indicated by False Ocean (unsuccessful ocean landing), False RTLS (unsuccessful landing on a ground pad), or False ASDS (unsuccessful landing on a drone ship). These distinctions help convert mission outcomes into clear binary labels for modeling and analysis.

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 therelationship between the specific categories being compared and a measured
value.

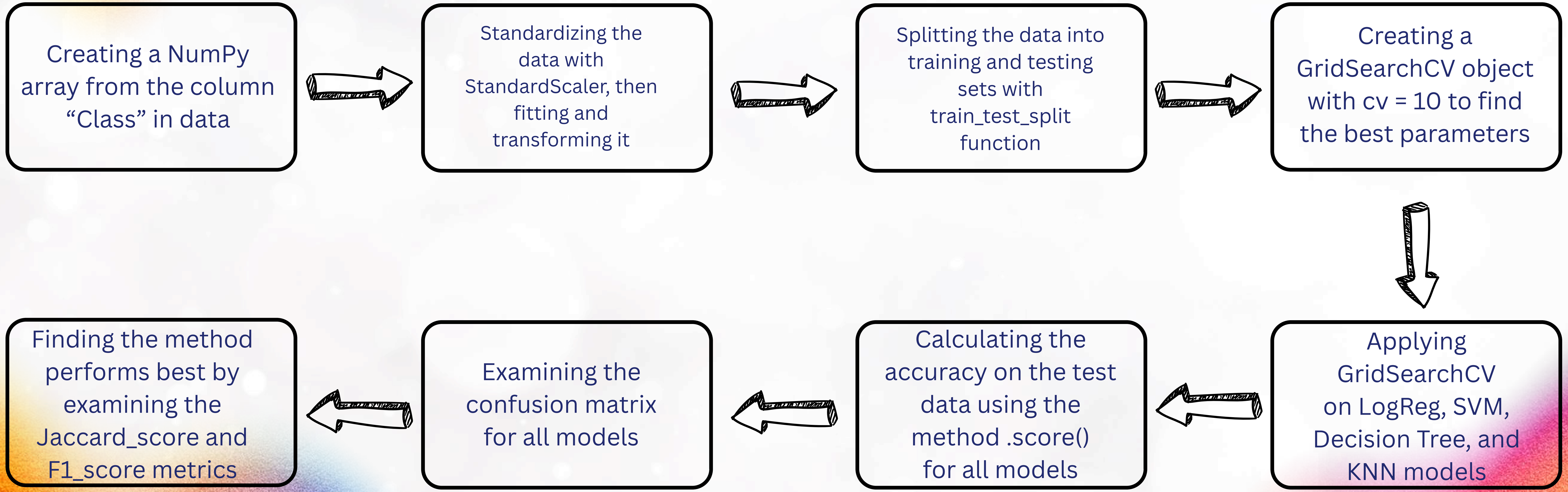
Line charts show trends in data over time (time series).

EDA WITH SQL

PERFORMED SQL QUERIES:

- 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 successful landing outcome in ground pad was achieved
- 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 for the months in 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

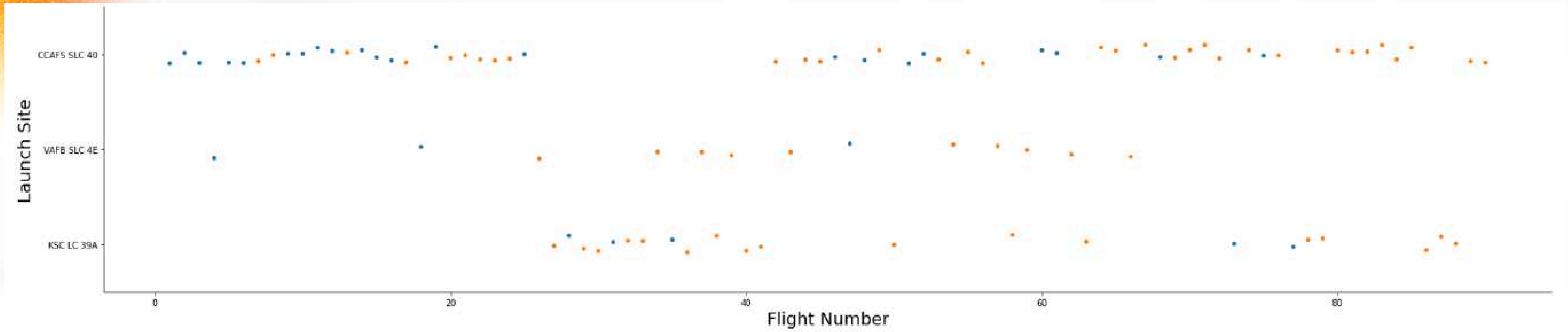
PREDICTIVE ANALYSIS (CLASSIFICATION)



RESULTS

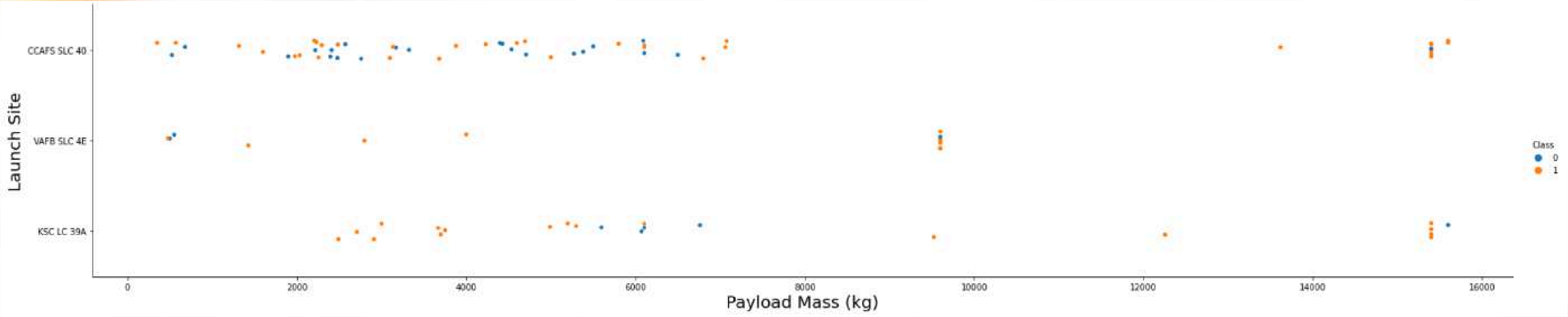
- EXPLORATORY DATA ANALYSIS
- ANALYTICS DEMO IN SCREENSHOTS
- PREDECTIVE ANALYSIS RESULTS

FLIGHT NUMBER VS LAUNCH SITE



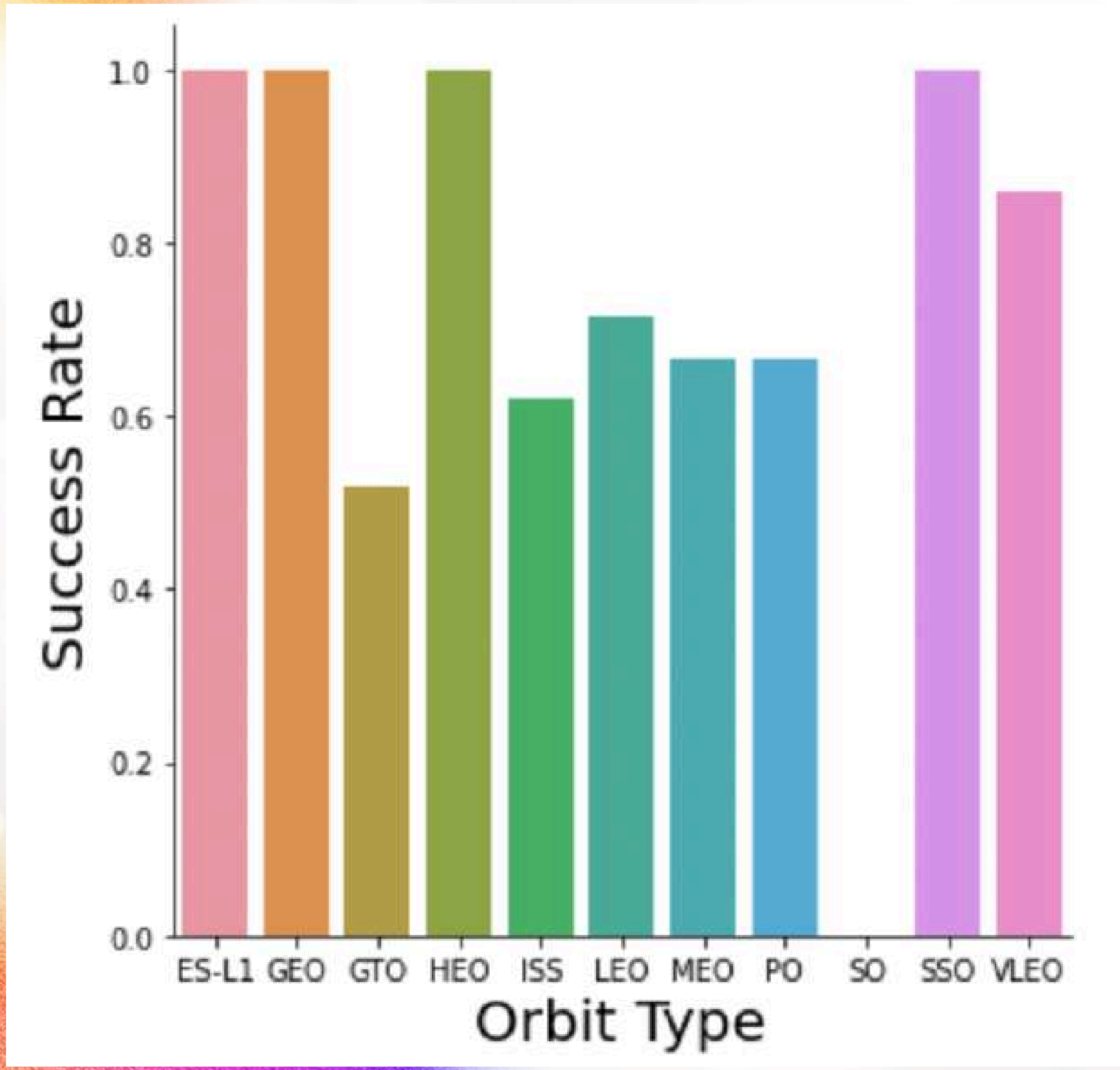
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

PAYLOAD VS LAUNCH SITE



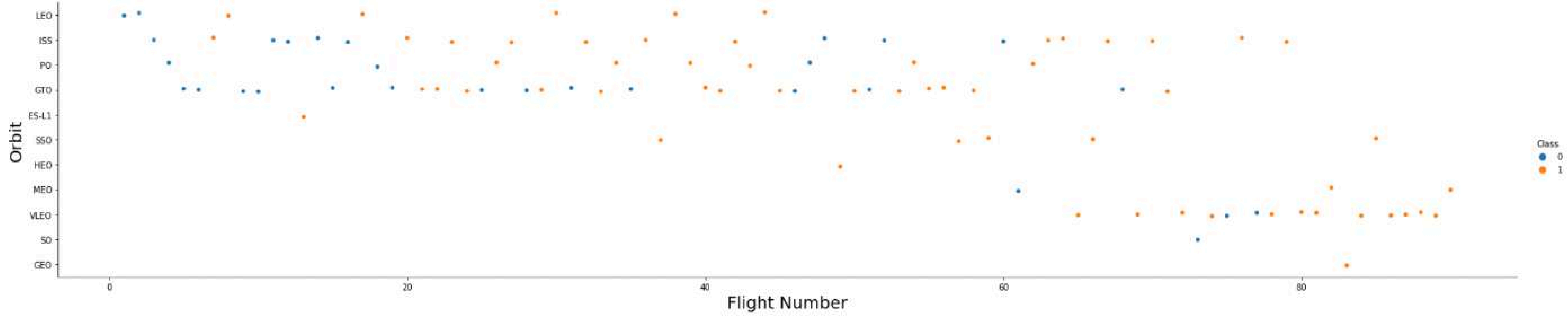
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

SUCCESS RATE VS ORBIT TYPE



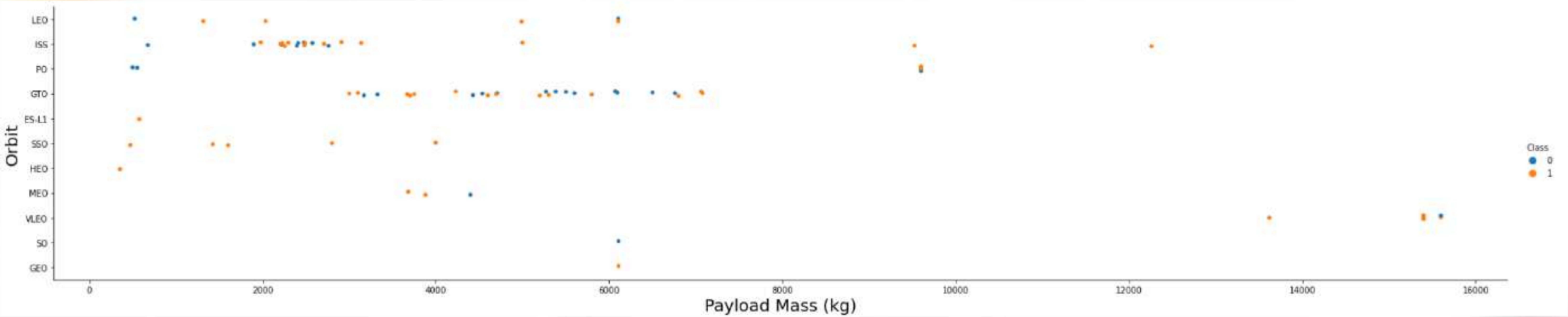
- Orbits with 100% success rate:
- ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
- SO
- Orbits with success rate between 50% and 85%:
- GTO, ISS, LEO, MEO, PO

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.

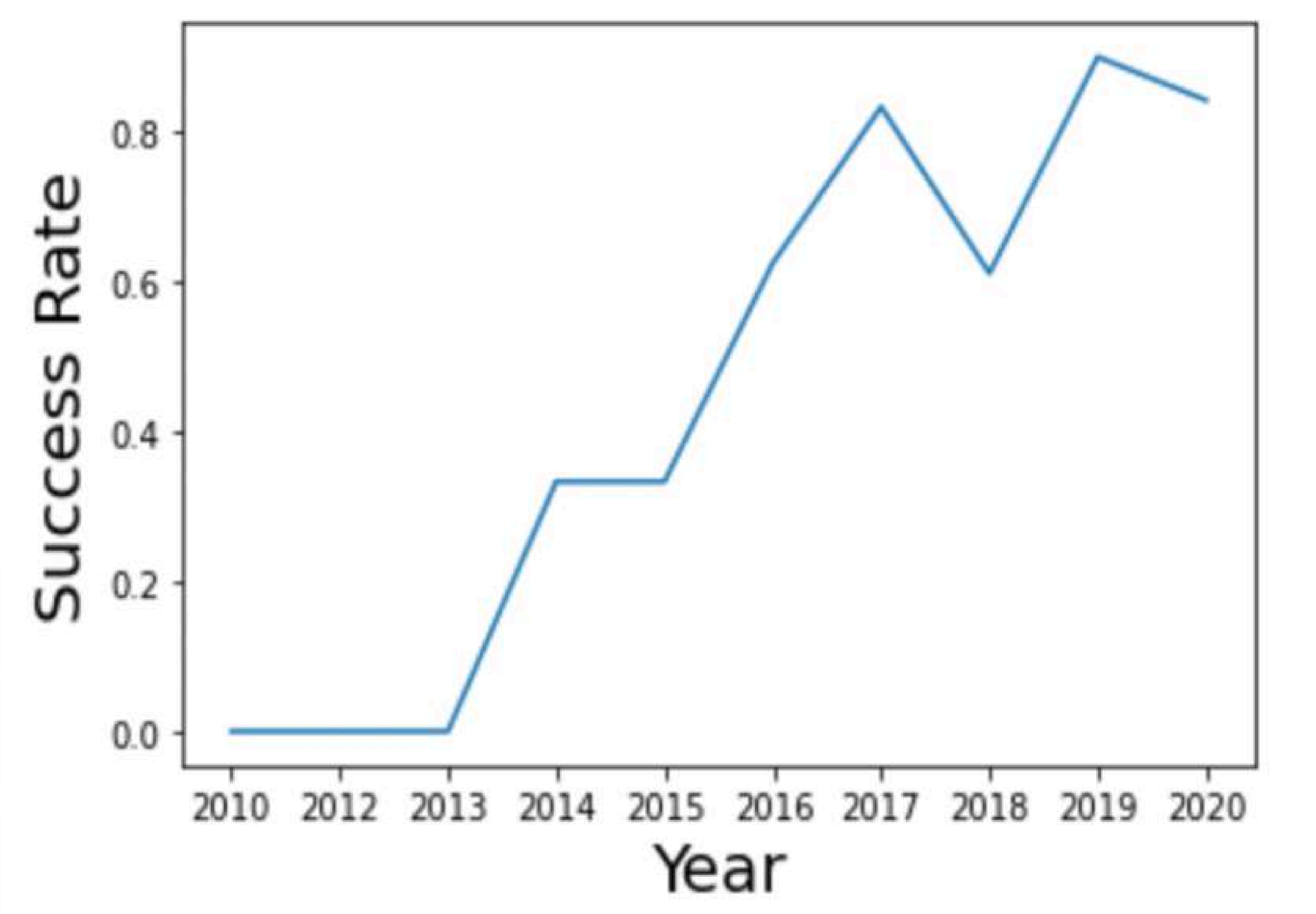
PAYLOAD MASS VS ORBIT TYPE



• 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 since 2013 kept increasing till 2020.



DATA VISUALIZATION WITH SQL

ALL LAUNCH SITE NAMES

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

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

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

TOTAL PAYLOAD MASS

```
In [6]: %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.
```

Out[6]:

total_payload_mass
45596

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

AVERAGE PAYLOAD MASS BY 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

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

FIRST SUCCESSFUL GROUND LANDING DATE

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

SUCCESSFUL DRONE SHIP LANDING

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 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

```
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 total number of successful and failure mission outcomes.

BOOSTERS CARRIED MAXIMUM PAYLOAD

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 names of the booster versions which have carried the maximum payload mass.

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

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.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

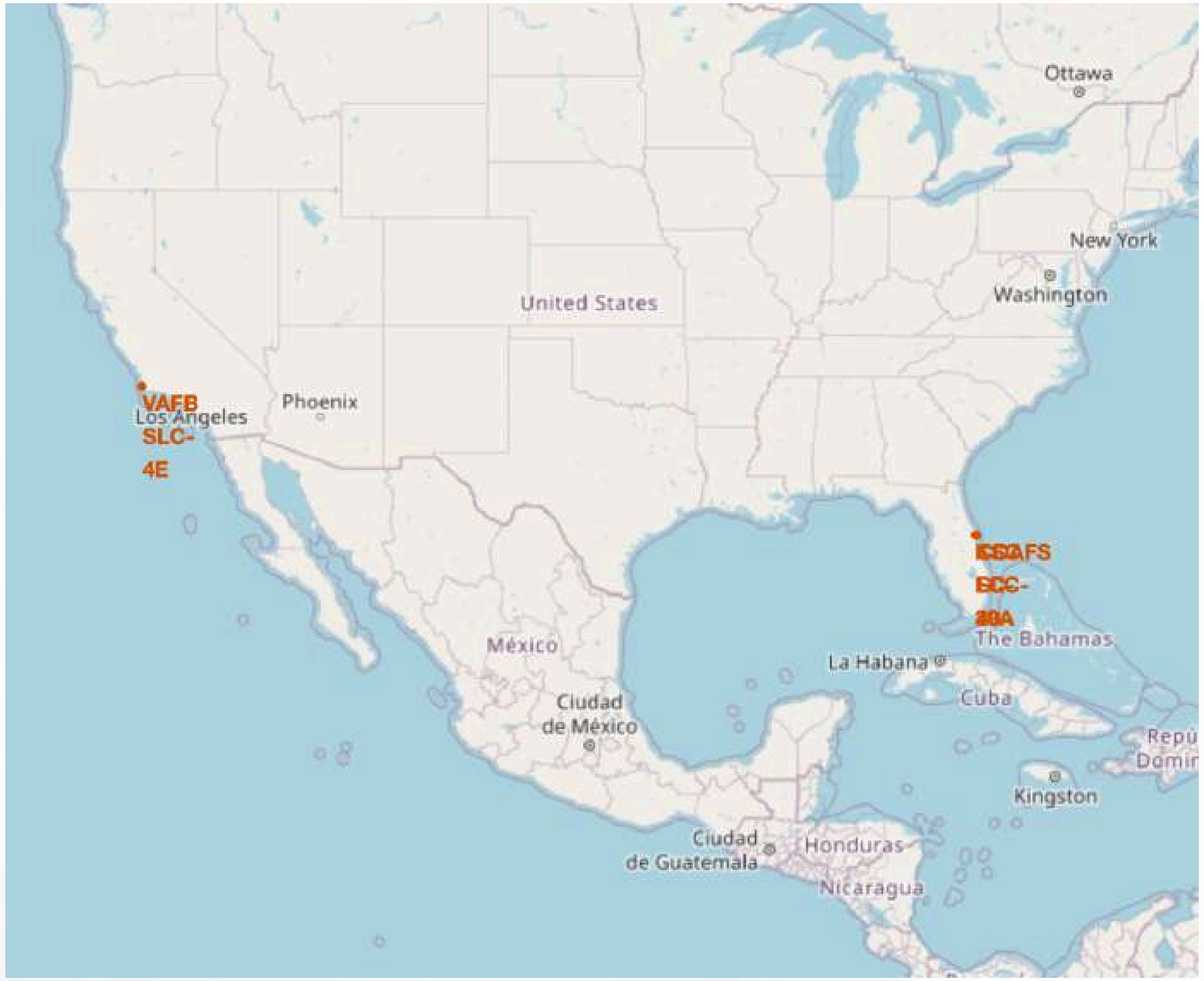
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- Ranking the count of landing outcomes (such as Failure (drone ship) or Succes (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending or

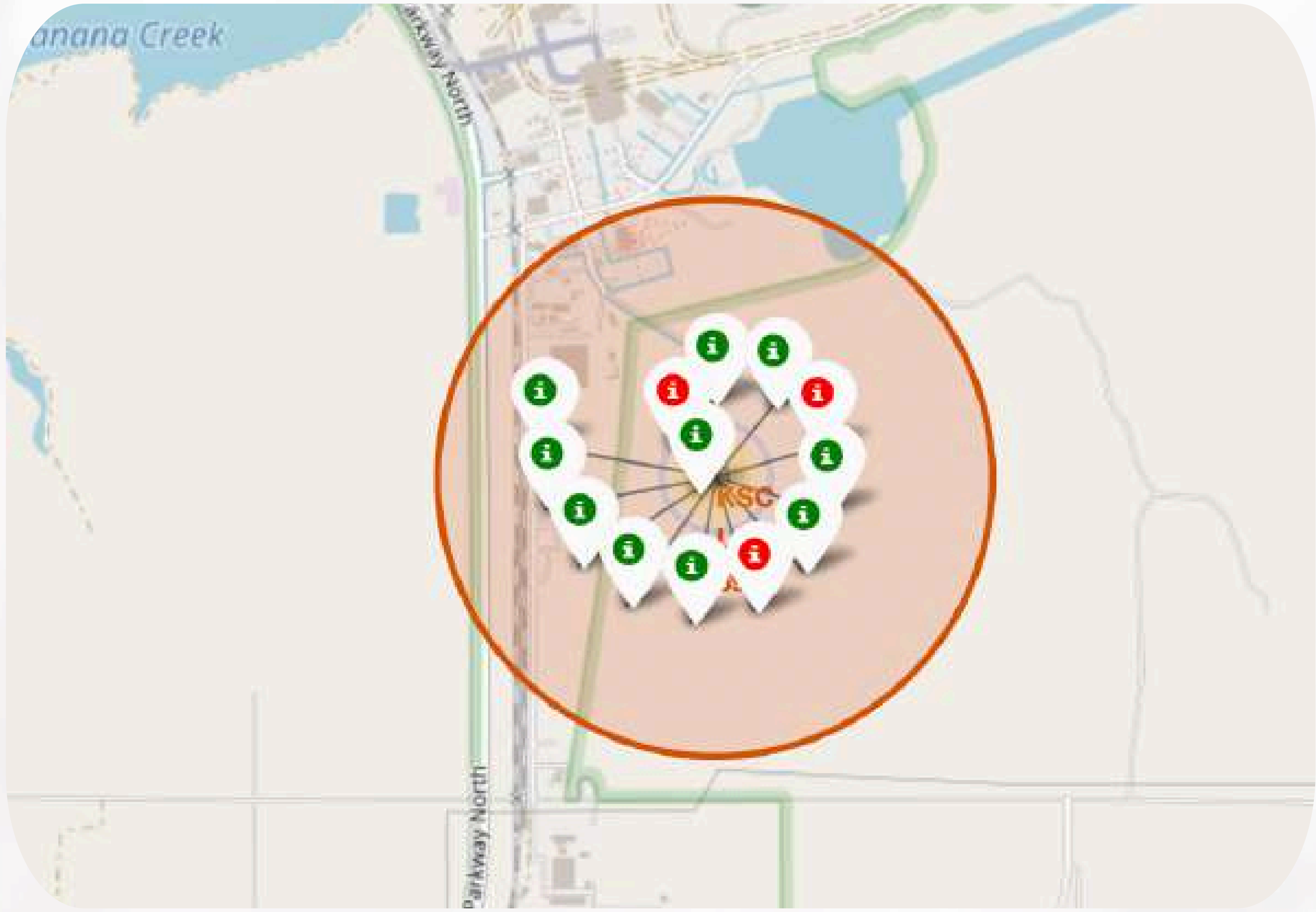
REFERENCED MAP WITH FOLIUM

ALL LAUNCH SITES' LOCATION MARKERS ON A GLOBAL MAP

There are multiple instances in the data set where the booster failed to land. A landing attempt may occasionally be unsuccessful owing to an accident; for instance, True Ocean indicates that the mission outcome was successfully landed in a certain area of the ocean, whereas False Ocean indicates that the mission outcome was unsuccessfully landed in a particular area of the ocean. If the mission outcome was successfully landed on a ground pad, it is known as true RTLS. An unsuccessful landing to a ground pad is indicated by a false RTLS. A successful mission outcome landing on a drone ship is referred to as true ASDS. An unsuccessful mission outcome landing on a drone ship is indicated by a false ASDS.



COLOUR-LABELED LAUNCH RECORDS ON THE MAP



- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- Green Marker = Successful Launch
- Red Marker = Failed Launch

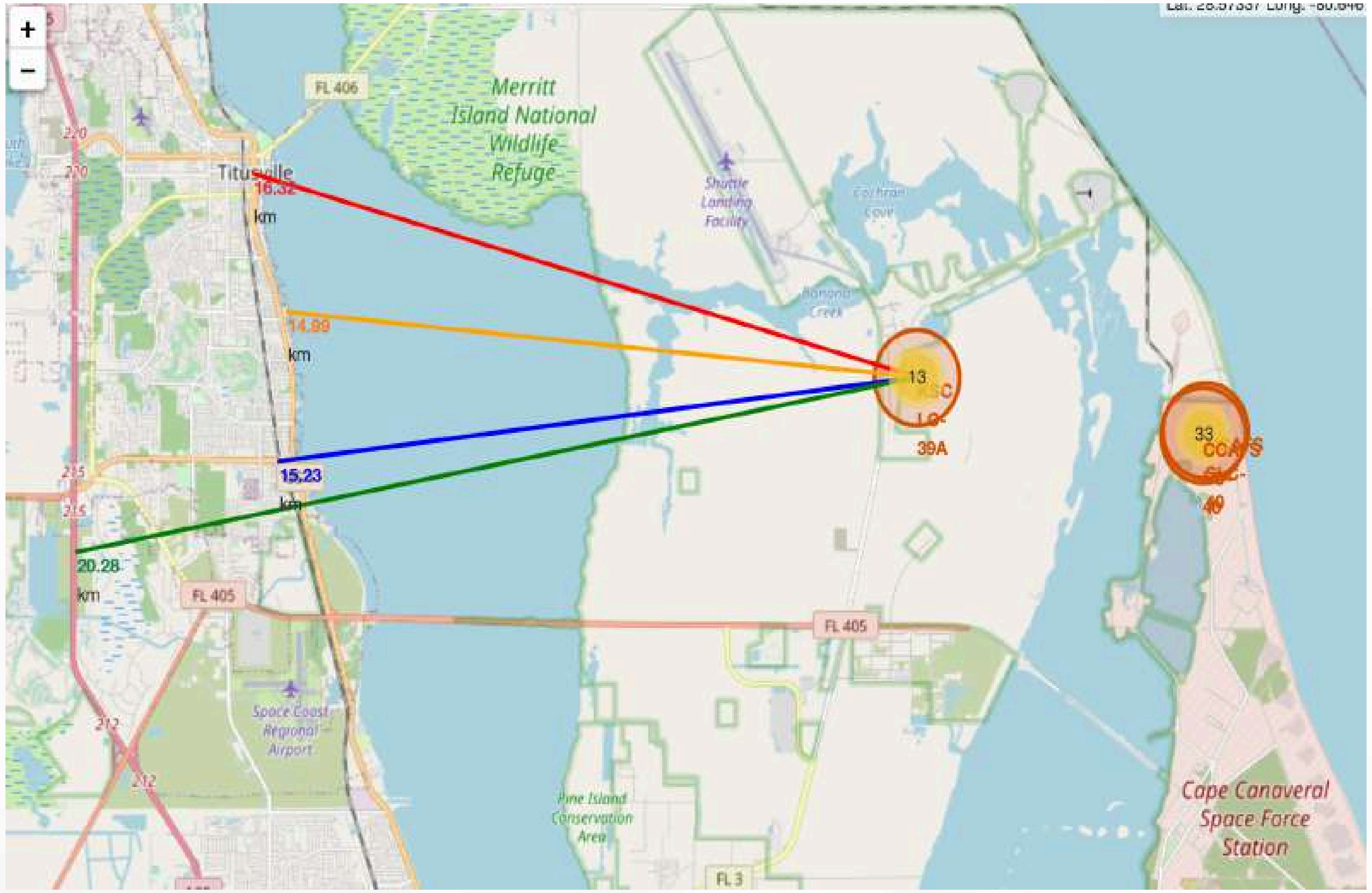
- Launch Site KSC LC-39A has a very high Success Rate.

DISTANCE FROM THE LAUNCH SITE KSC LC-39A TO ITS PROXIMITIES

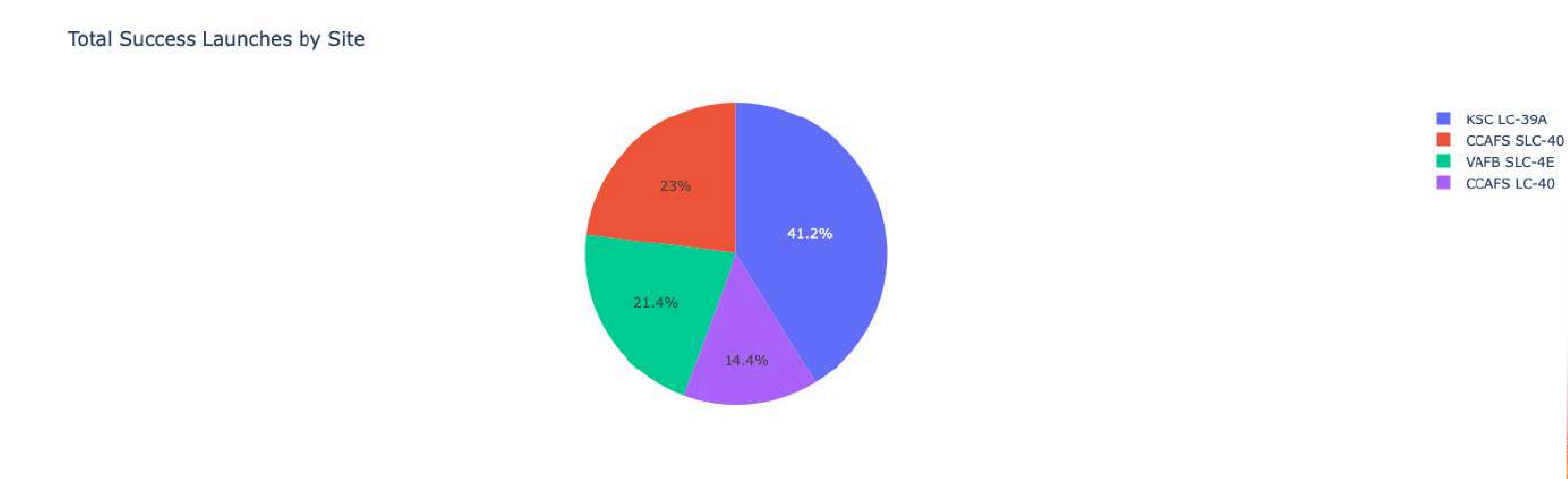
• From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)

• 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.

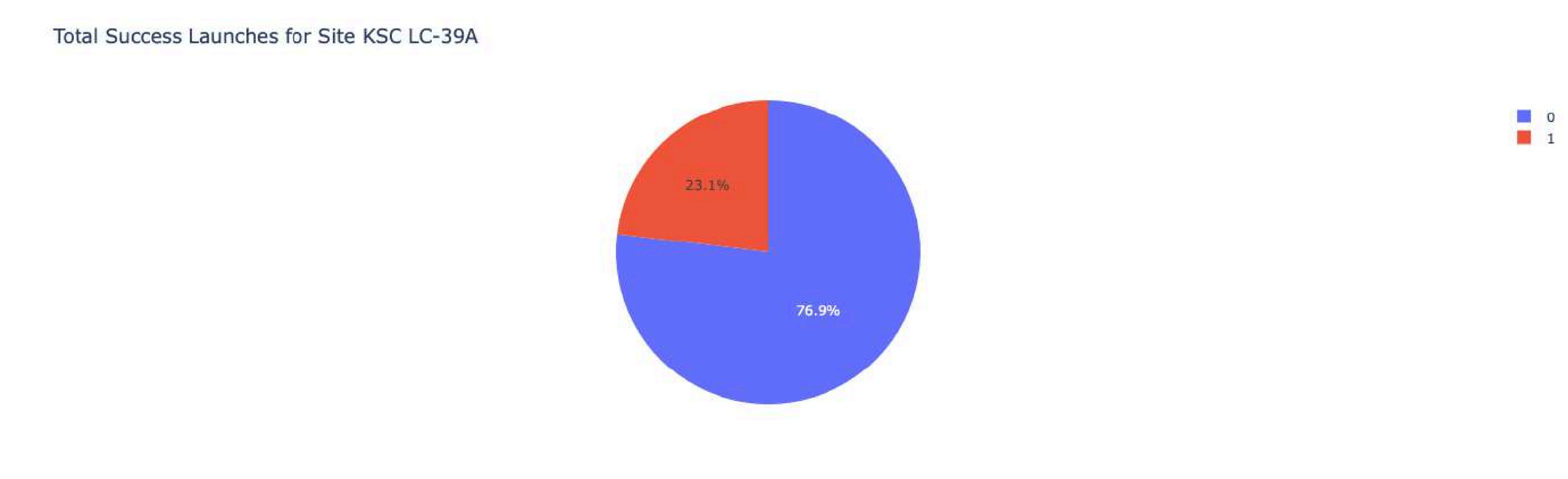


LAUNCH SUCCESS COUNT FOR ALL SITES



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

LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO



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

PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

SCORES AND ACCURACY OF THE TEST SET

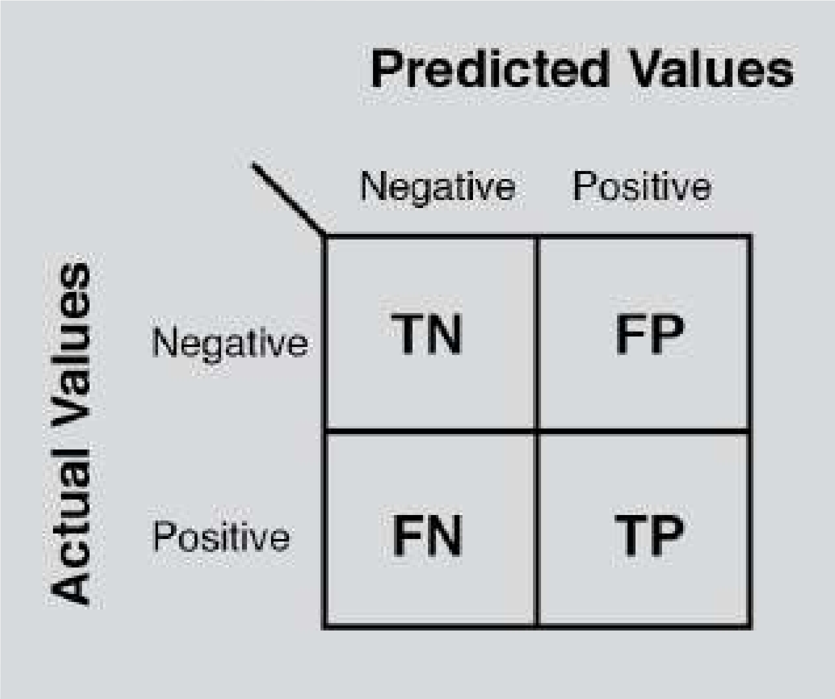
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

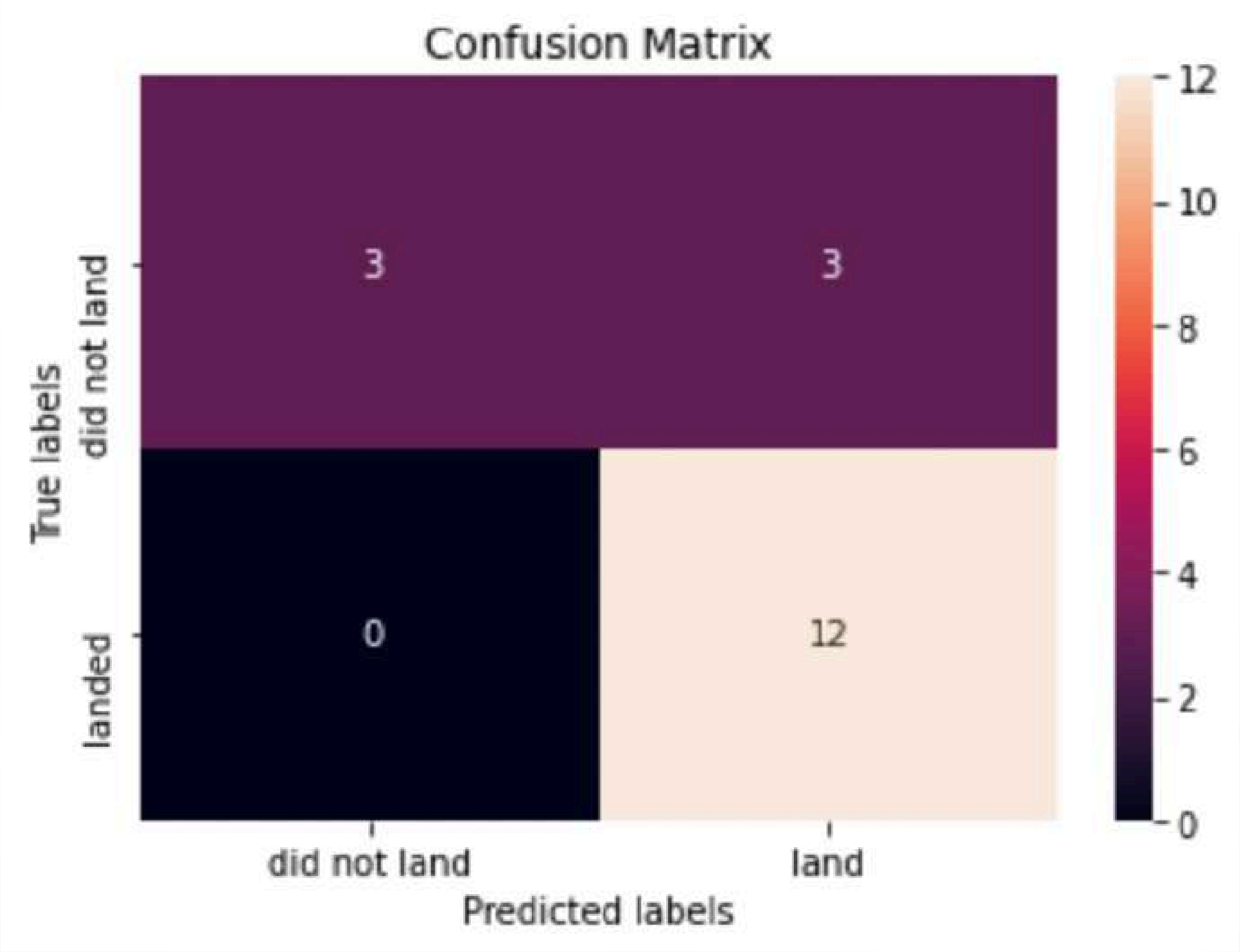
SCORES AND ACCURACY OF THE ENTIRE DATA SET

	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

CONFUSION MATRIX

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





CONCLUSION

- Decision Tree Model is the best algorithm for this dataset.
- • Launches with a low payload mass show better results than launches with a larger payload mass.
- • Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- • The success rate of launches increases over the years.
- • Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- • KSC LC-39A has the highest success rate of the launches from all the sites.

THANK YOU

If you have any questions or would like further discussion, please feel free to contact me.

https://www.linkedin.com/in/werissonm/