WINNING SPACE RACE WITH DATA SCIENCE

IBM Developer
SKILLS NETWORK

Pankati Bhatt

02-May-2023

OUTLINE

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

Project background and context

The most prosperous business of the commercial space age, SpaceX has reduced the cost of space travel. On its website, the corporation promotes Falcon 9 rocket flights for 62 million dollars; alternative providers charge upwards of 165 million dollars each launch. Because SpaceX can reuse the first stage, there will be significant savings. So, if we can figure out whether the first stage will land, we can figure out how much a launch will cost. We are going to make a prediction about whether SpaceX will reuse the first stage based on available data and machine learning models.

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?

What is the best algorithm that can be used for binary classification in this case?

METHODOLOGY

SECTION 1

METHODOLOGY

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 Scraping data from a table in SpaceX's Wikipedia entry.

First the data collection was done using get request to the SpaceX API.

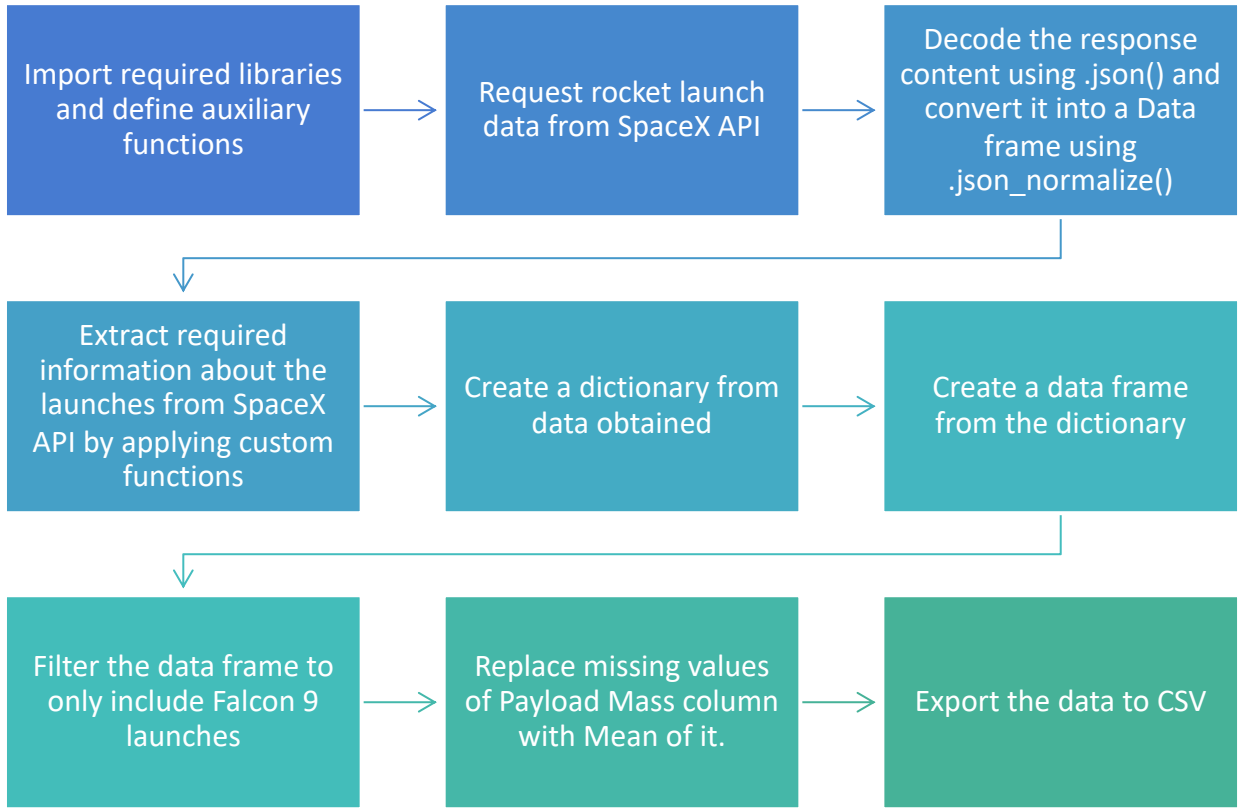
Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().

We then cleaned the data, checked for missing values and fill in missing values where necessary.

In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.

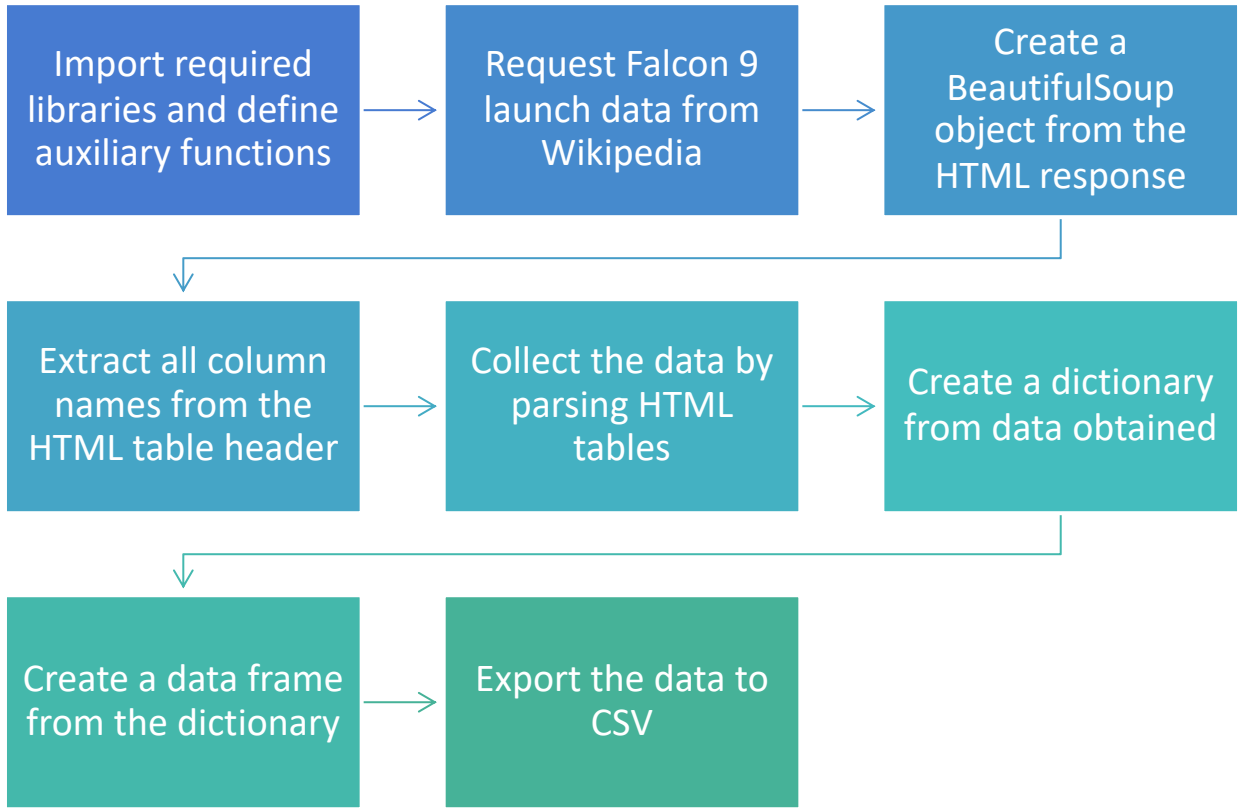
The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas data frame for future analysis.

DATA COLLECTION – SPACEX API



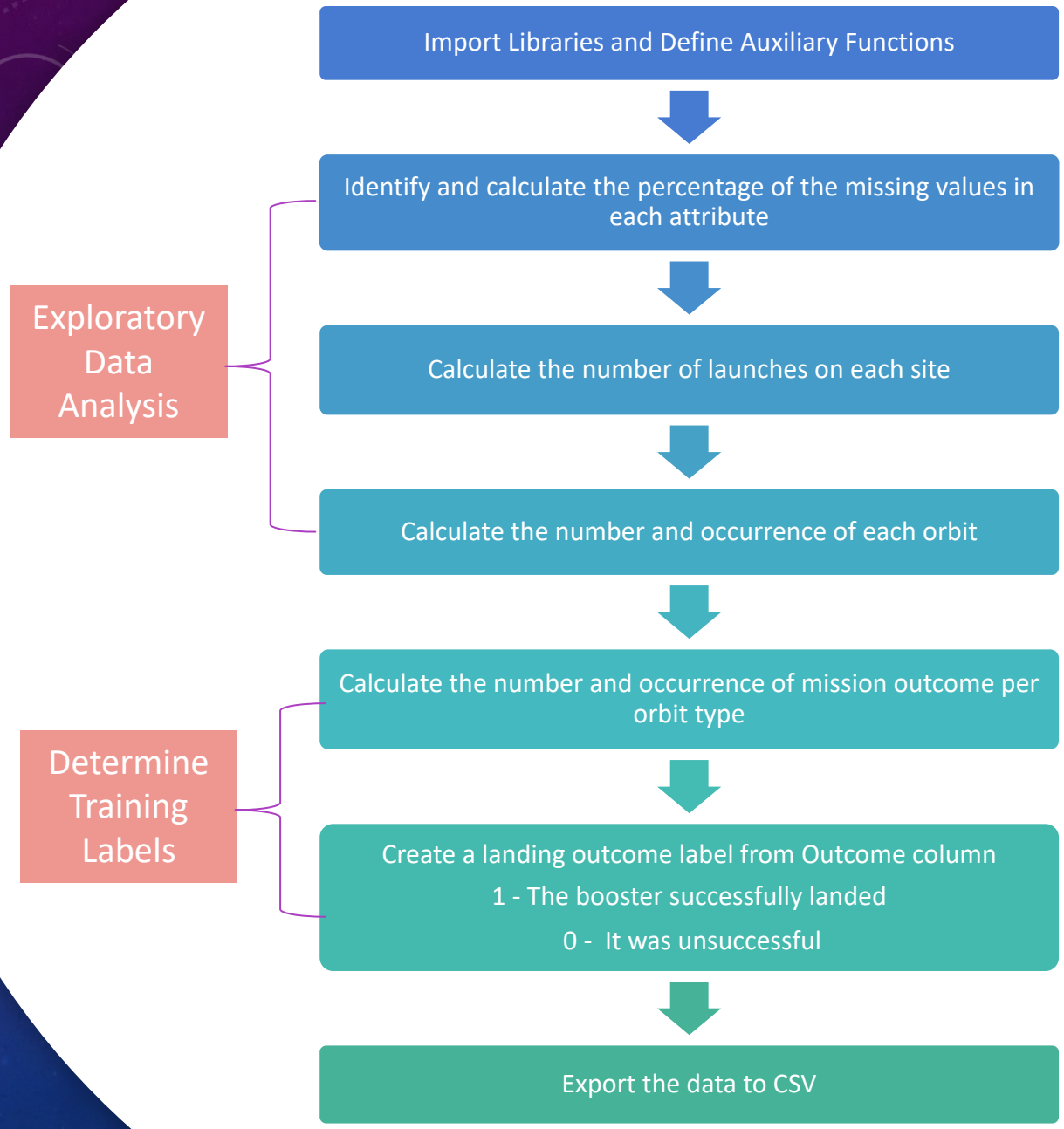
Github URL-Data Collection – SpaceX API

DATA COLLECTION - SCRAPING



Github URL-Data Collection - Scraping

DATA WRANGLING



Github URL-Data Wrangling

EDA WITH DATA VISUALIZATION

EDA with Data Visualization

Chart Type	Goal	Charts Plotted
Scatter plot	To find the relationship between variables. If a relationship exists, they could be used in machine learning model	FlightNumber vs PayloadMass
		FlightNumber vs LaunchSite
		PayloadMass vs Launch Site
		FlightNumber vs Orbit type
		Payload vs Orbit type
Bar plot	To compare different categorical or discrete variables for measured value.	Success rate of each orbit type
Line plot	To show trends in data over time (time series)	Yearly Trend of Launch success rate

Feature Engineering

Select the features that will be used in success prediction → Create dummy variables to categorical columns → Cast all numeric columns to float64

Github URL-EDA with Data Visualization

EDA WITH SQL

Performed SQL queries:

- Download the datasets
- Connect to the database
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string ‘CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass. Use a subquery.
- List the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 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 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.

Colored Markers of the launch outcomes for each Launch Site:

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

Github URL-Build an Interactive Map 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 (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:

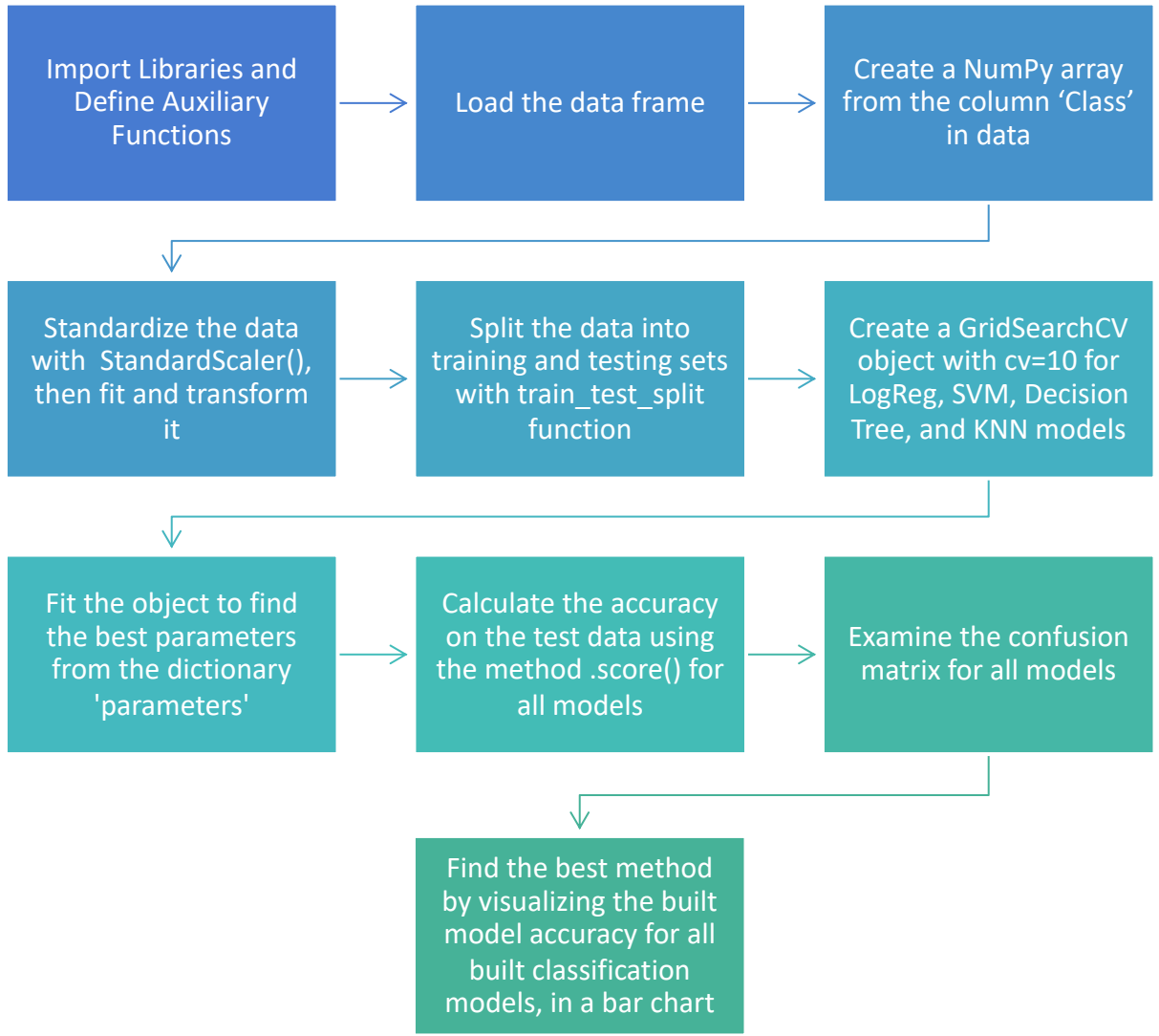
- Added a slider to select Payload range.

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.

Github URL-Build a Dashboard with Plotly Dash

PREDICTIVE ANALYSIS (CLASSIFICATION)



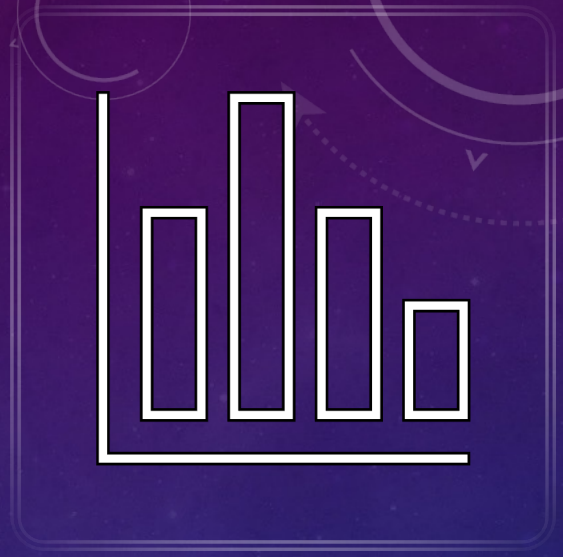
Github URL-Predictive Analysis (Classification)

RESULTS

EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

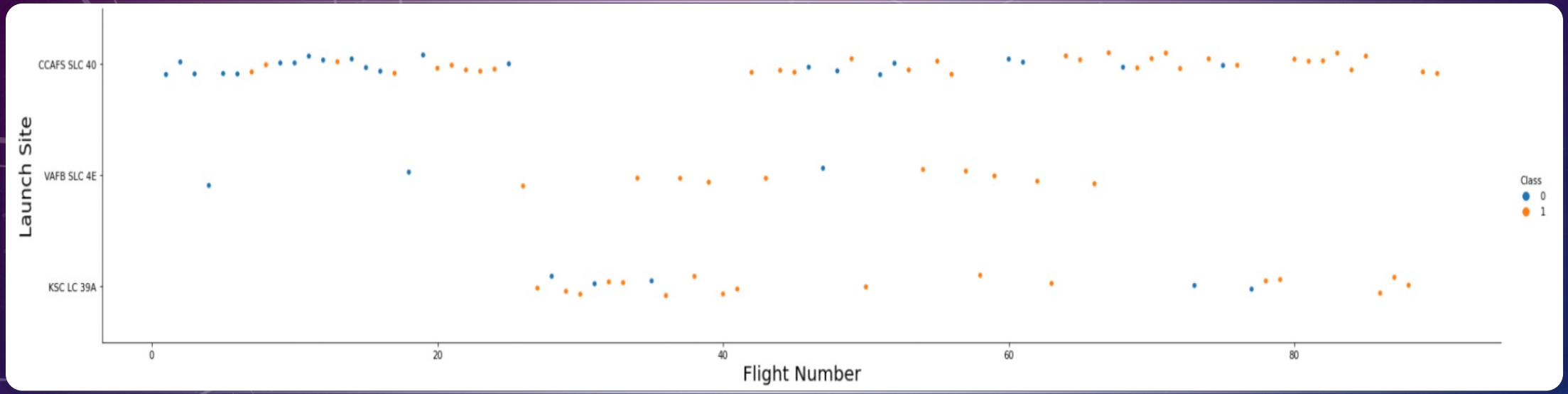
PREDICTIVE ANALYSIS RESULTS



INSIGHTS DRAWN FROM EDA

SECTION 2

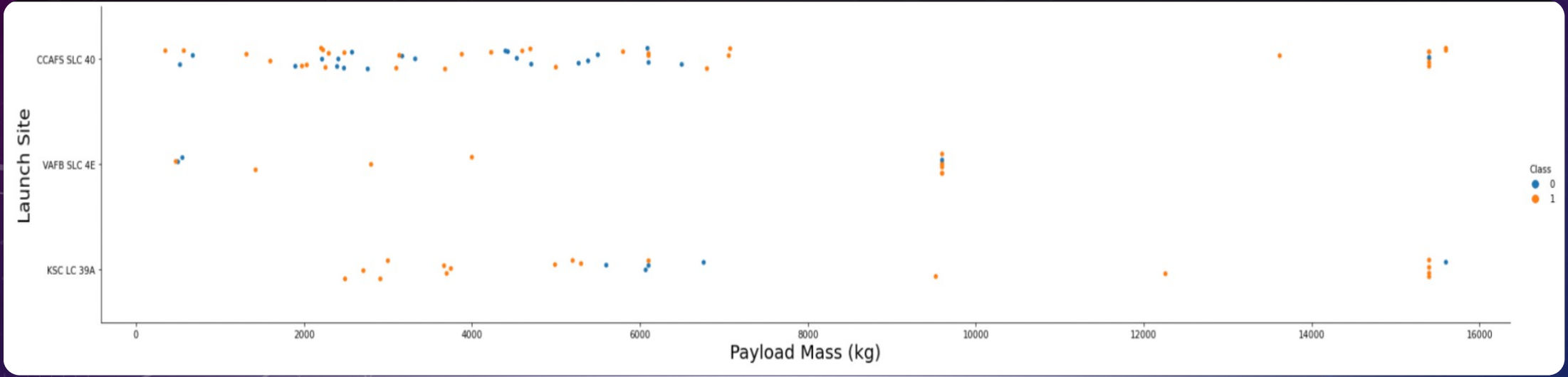
FLIGHT NUMBER VS. LAUNCH SITE



Explanation:

- The increase in success rate over time (indicated in Flight Number).
- Likely a big breakthrough around flight 20 which significantly increased success rate.
- CCAFS SLC 40 appears to be the main launch site as it has the most volume.
- 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



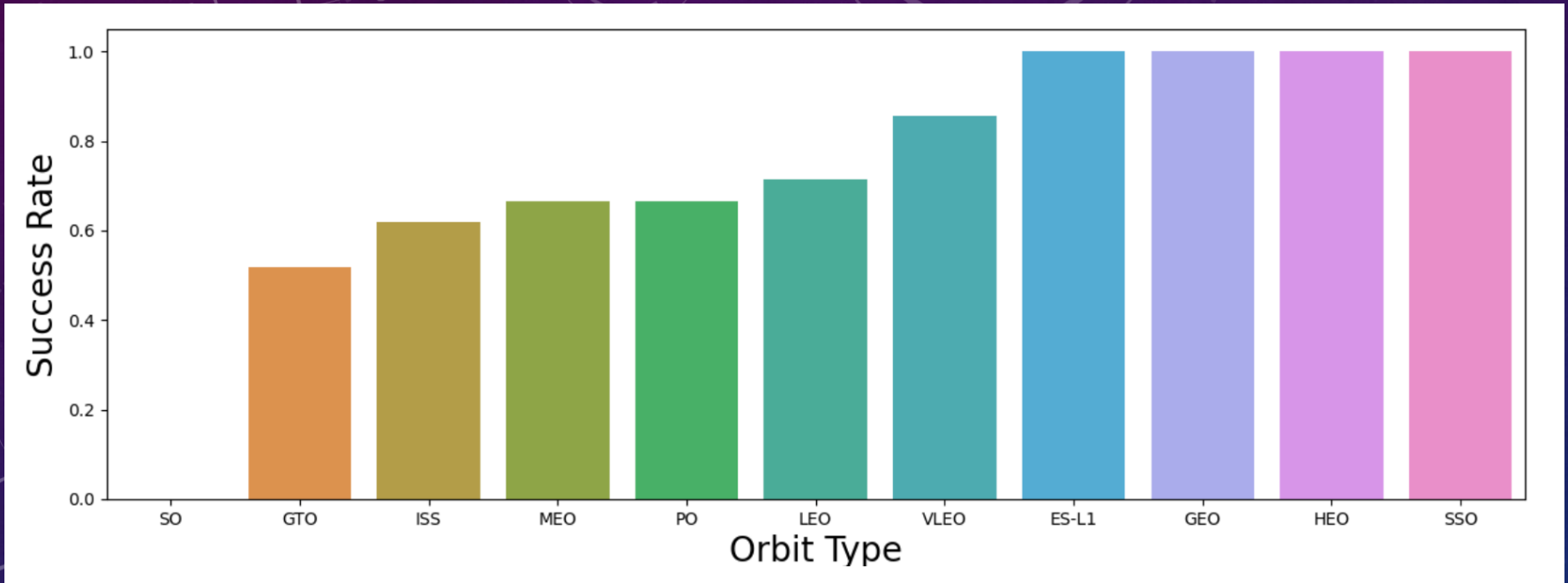
Explanation:

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



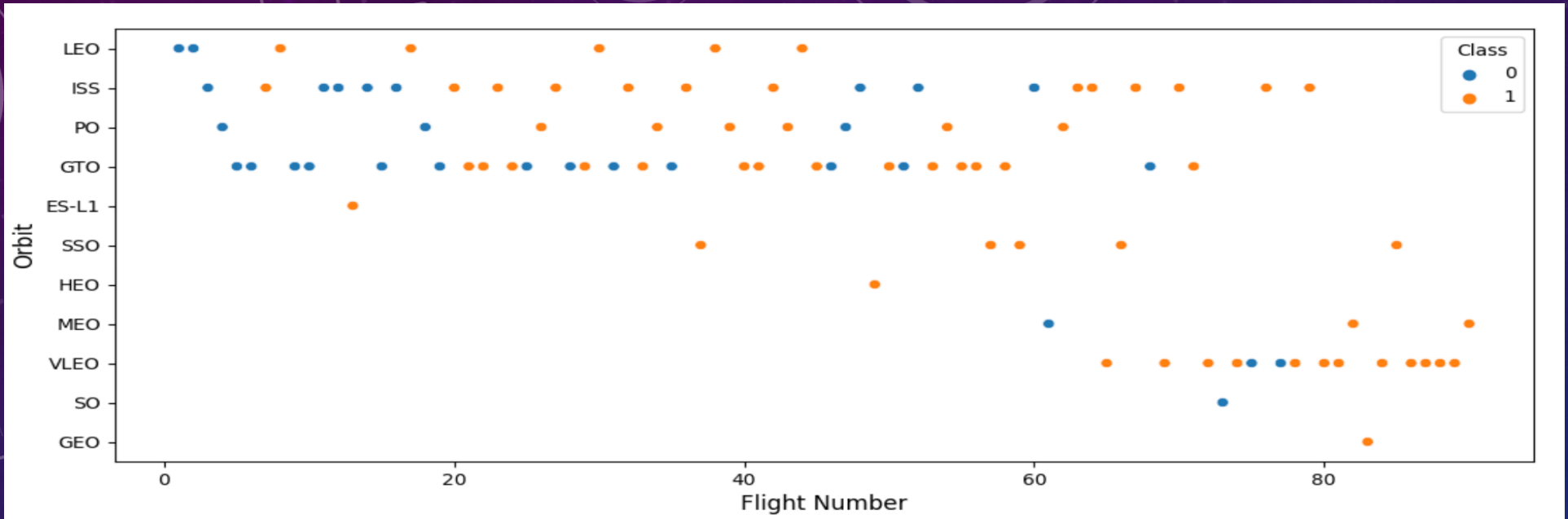
Explanation:

Orbits with 100% 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

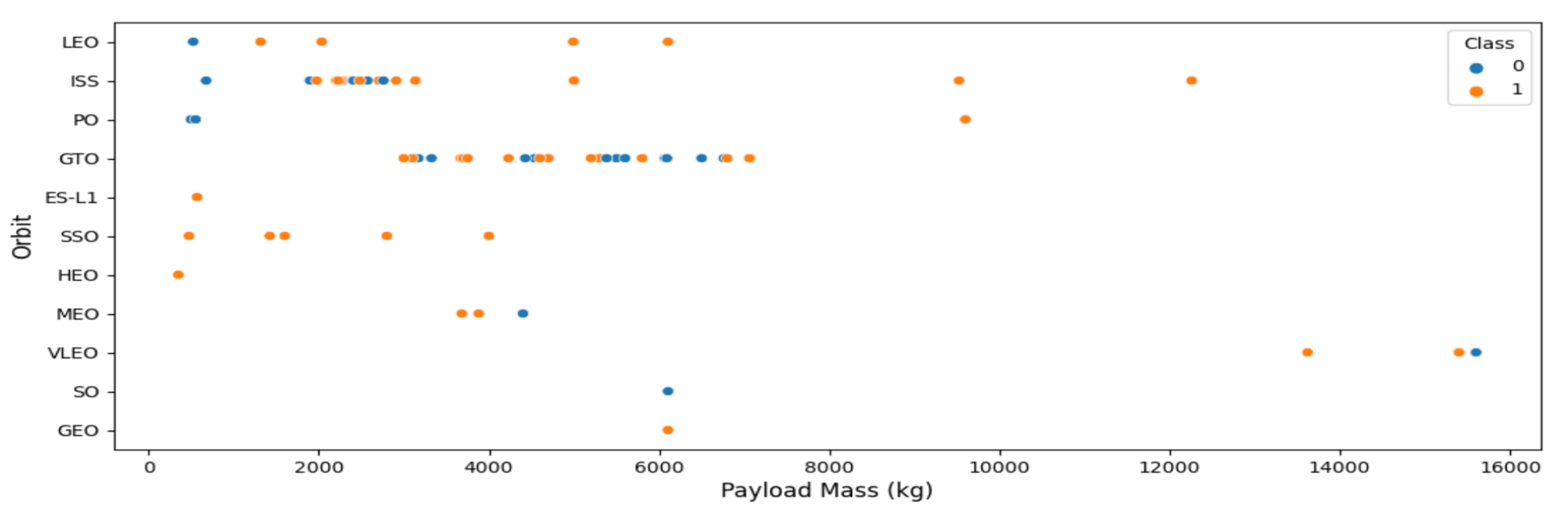
FLIGHT NUMBER VS. ORBIT TYPE



Explanation:

- Launch Orbit preferences changed over Flight Number
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

PAYLOAD VS. ORBIT TYPE



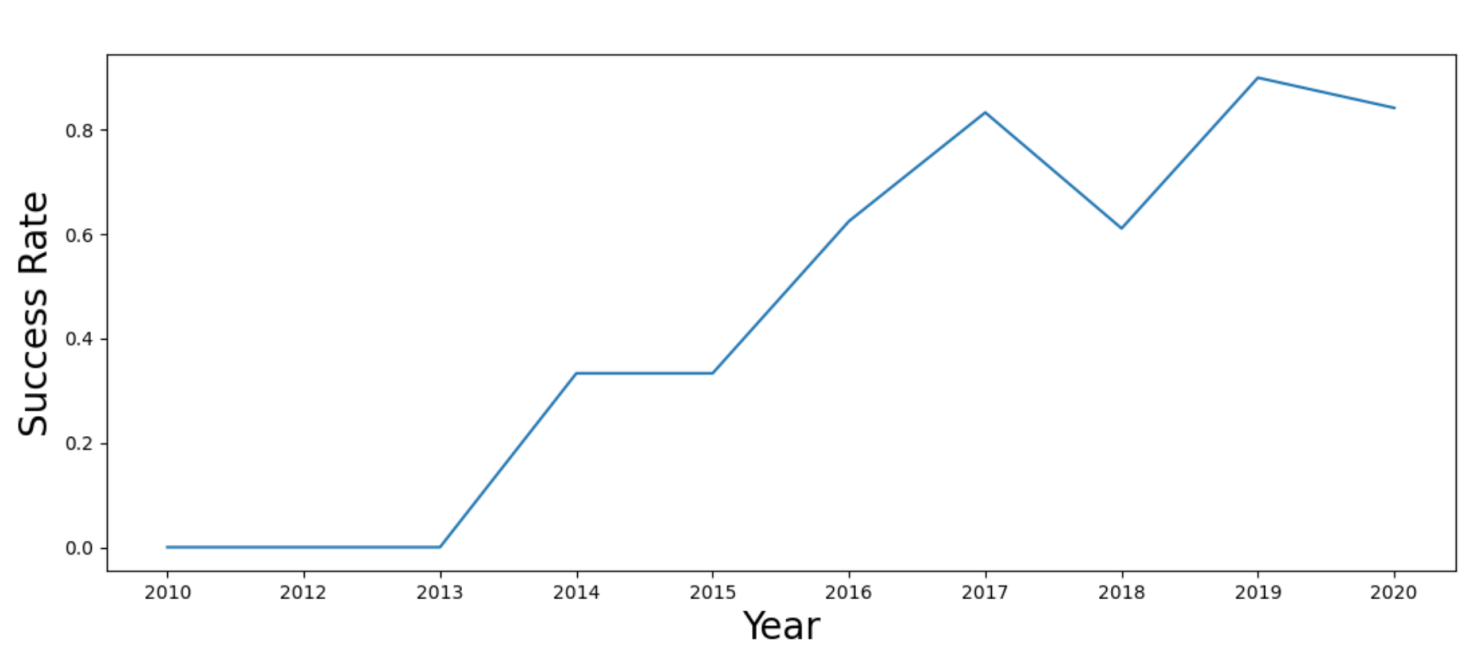
Explanation:

With heavy payloads the successful landing are more for LEO and ISS

With payloads less than 4000 kg the successful landing are more for SSO,HEO and MEO.

For GTO we cannot distinguish this well as both positive landing rate and negative landing are both there.

LAUNCH SUCCESS YEARLY TREND



Explanation: Success rate since 2013 kept increasing with a slight dip in 2018 till 2020

ALL LAUNCH SITE NAMES

```
In [4]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[4]:

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

Explanation:

- Displaying the names of the launch sites in the space mission

LAUNCH SITE NAMES BEGIN WITH 'CCA'

In [5]:
```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/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

Explanation:

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

TOTAL PAYLOAD MASS

```
In [6]: %sql SELECT SUM (PAYLOAD_MASS__KG_) as Total_Payload_Mass FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
Out[6]: total_payload_mass
        45596
```

Explanation:

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 SPACEX WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
Out[7]: average_payload_mass
                        2534
```

Explanation:

- 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 SPACEX where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
Out[8]:
```

first_successful_landing
2015-12-22

Explanation:

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

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

In [9]:
```
%sql SELECT BOOSTER_VERSION FROM (SELECT * FROM SPACEX WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000) WHERE landing__outcome = 'Success (
```

```
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[9]:

booster_version
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

```
In [10]: %sql SELECT DISTINCT MISSION_OUTCOME,COUNT(MISSION_OUTCOME) FROM SPACEX GROUP BY MISSION_OUTCOME;
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[10]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

Listing the total number of successful and failure mission outcomes

BOOSTERS CARRIED MAXIMUM PAYLOAD

In [11]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);
```

```
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/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

Explanation:

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

2015 LAUNCH RECORDS

In [12]:
```
%%sql select year(date) as year, monthname(date) as month, landing__outcome, booster_version, launch_site
from SPACEX where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[12]:

YEAR	MONTH	landing__outcome	booster_version	launch_site
2015	January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanation:

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

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

[13]:
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEX
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

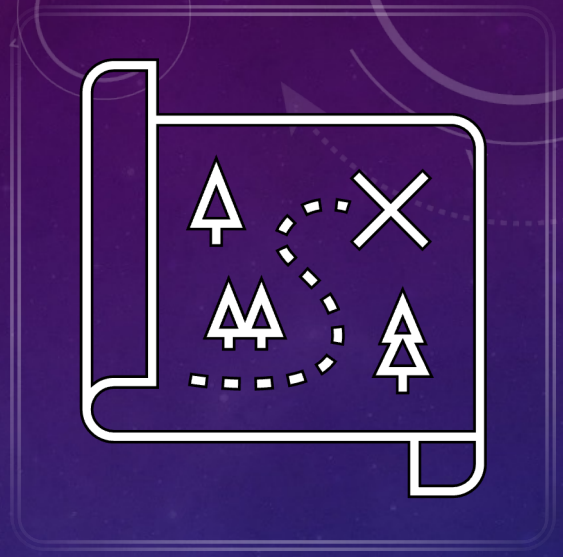
```
 * ibm_db_sa://fkk36439:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

[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

Explanation:

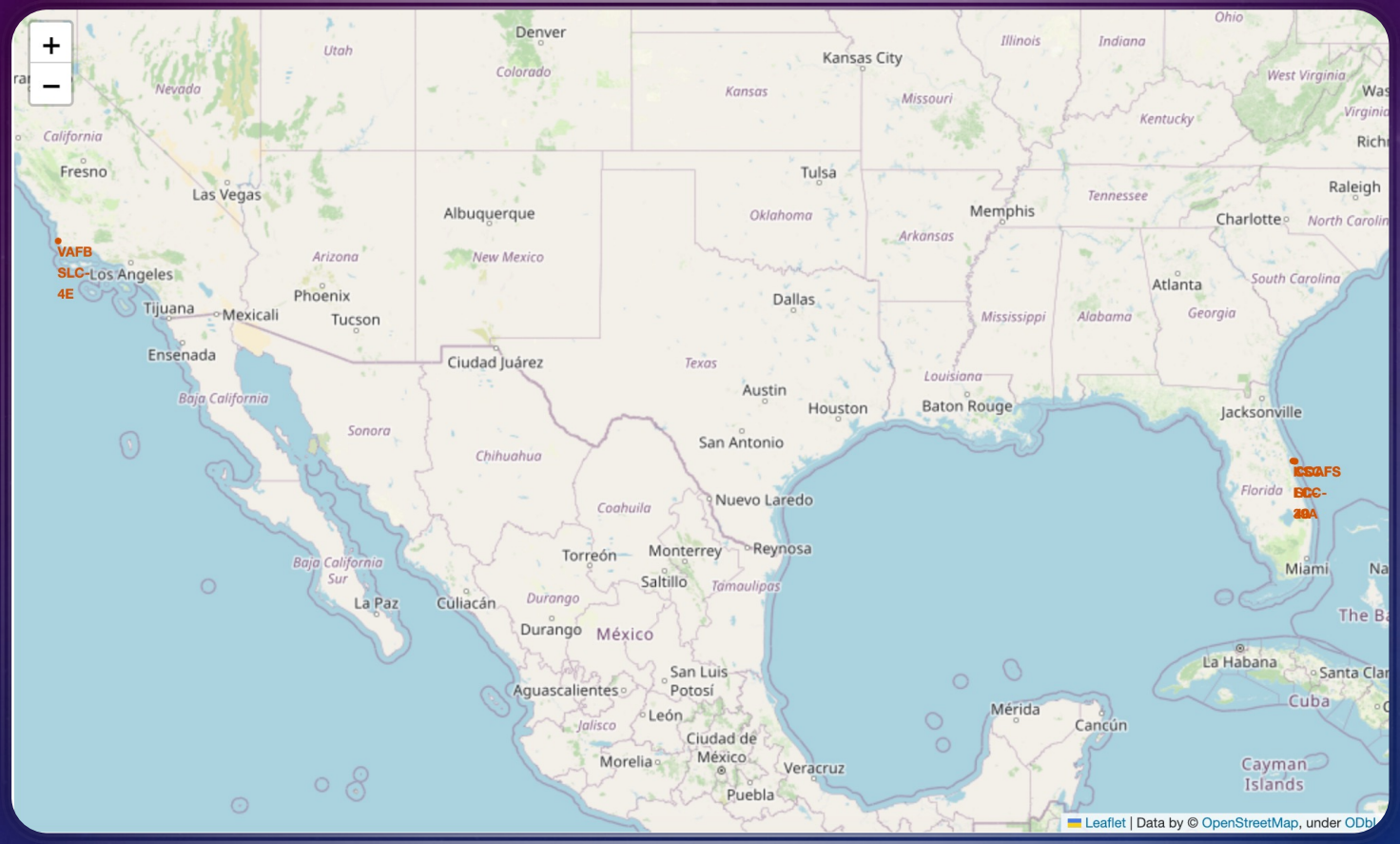
Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.



LAUNCH SITES PROXIMITIES ANALYSIS

SECTION 3

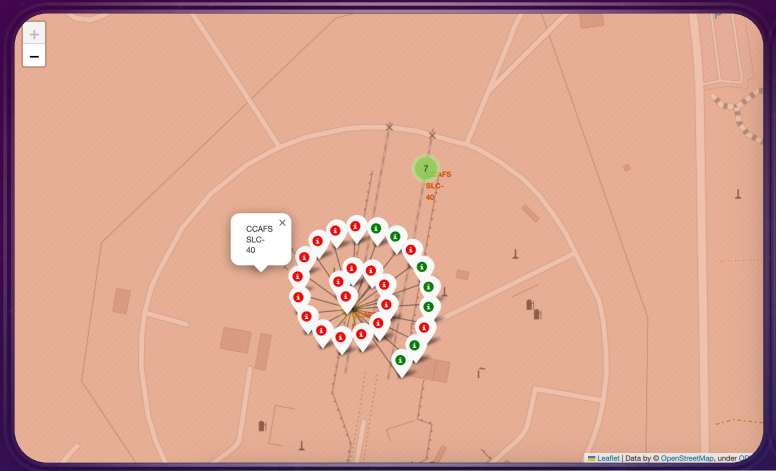
ALL LAUNCH SITES' LOCATION

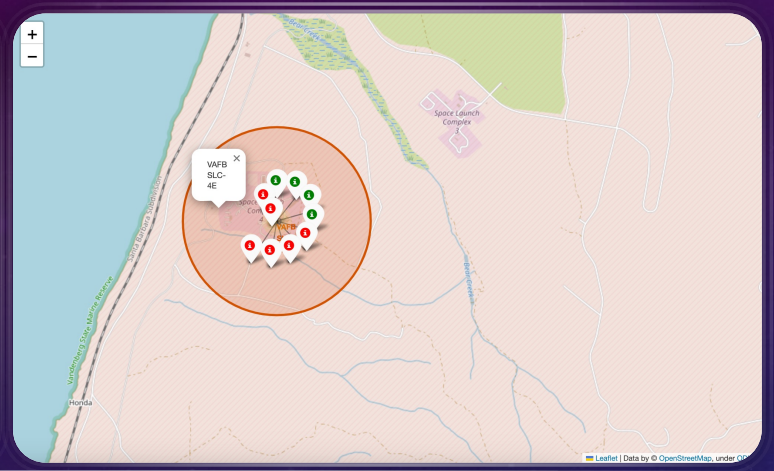


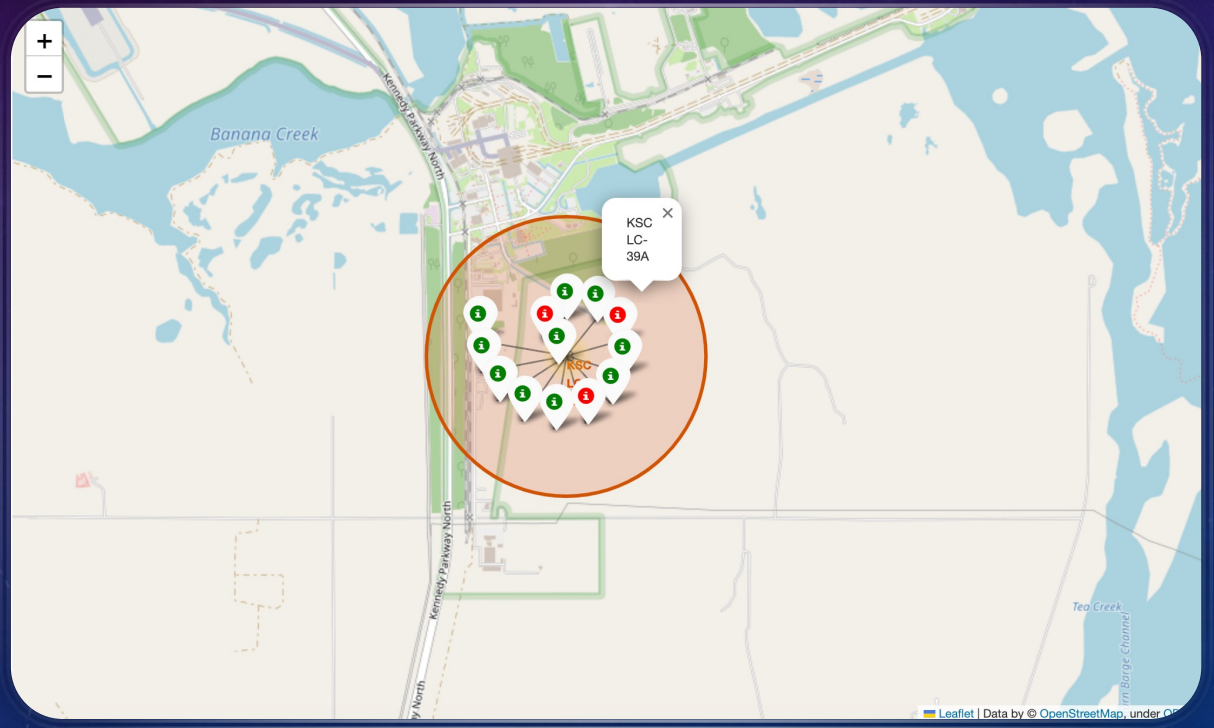
Explanation:

- Most of Launch sites are in proximity to the Equator line
- All launch sites are very close to the coast, which reduces the risk of debris falling or exploding near people.

COLOR-LABELED LAUNCH OUTCOMES ON THE MAP



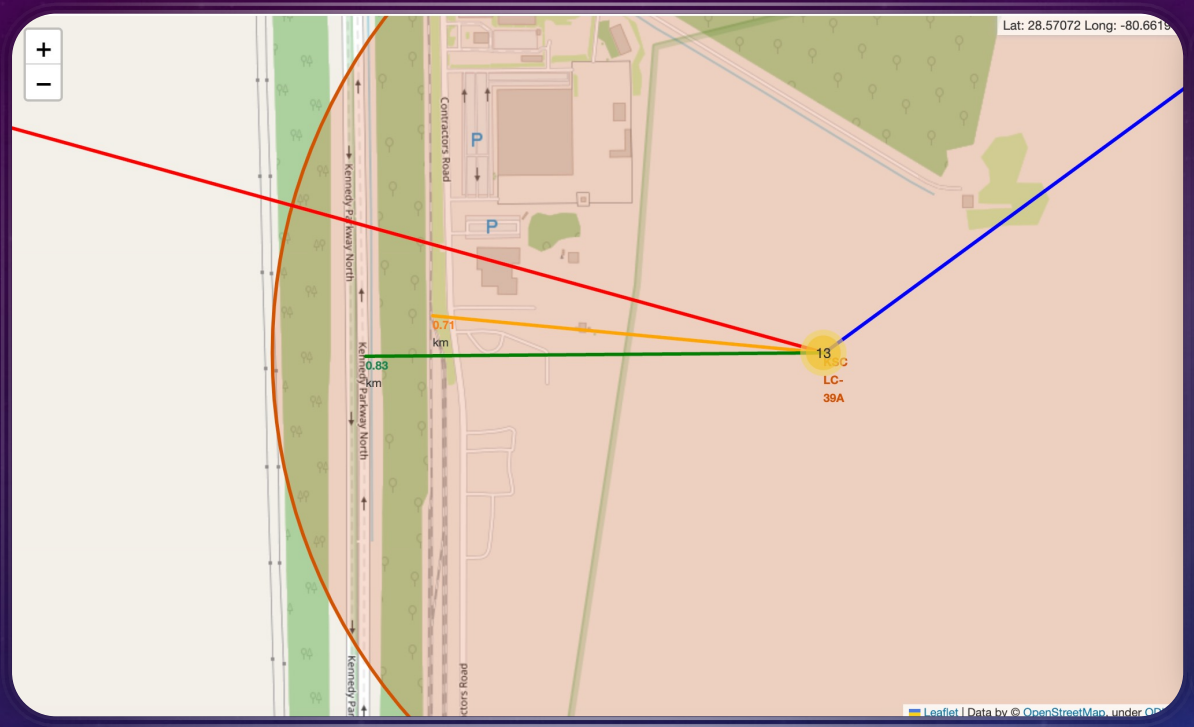


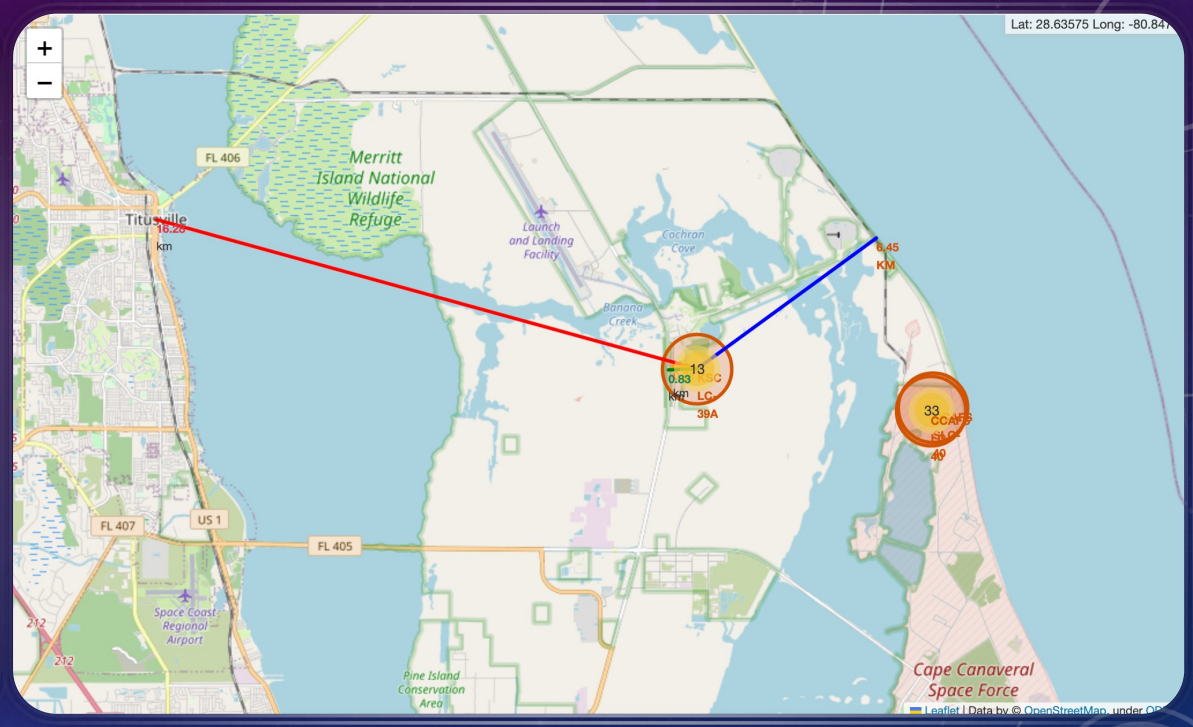


Explanation:

- From the color-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





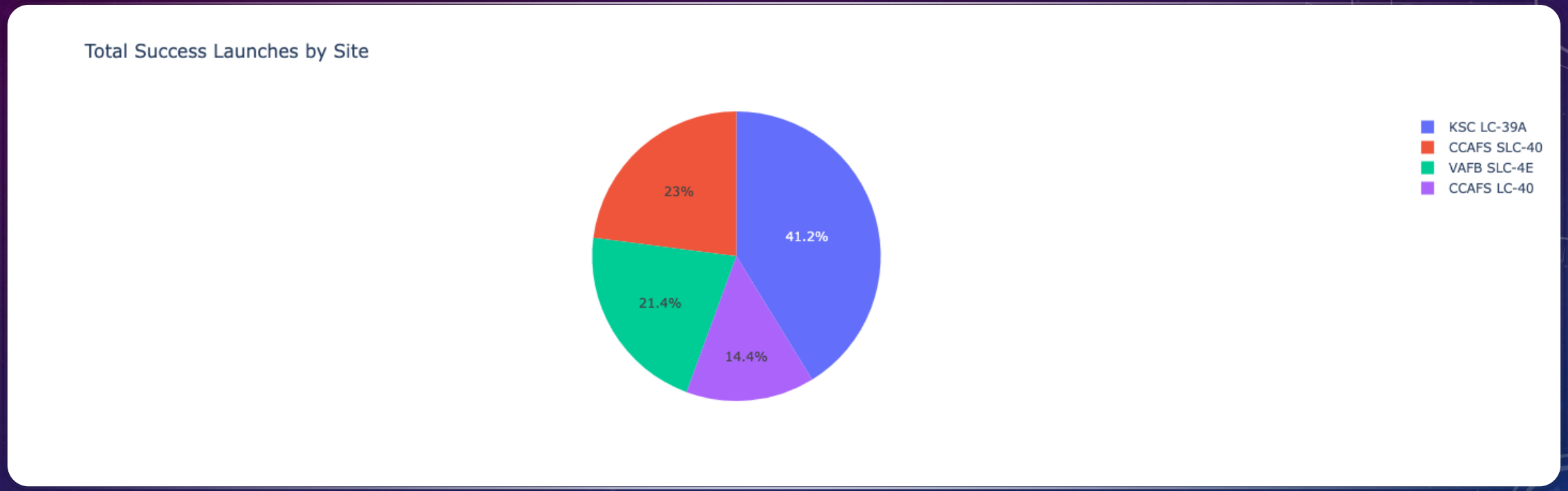
Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly see that the distance from:
  - Railway is 0.71 km
  - Highway is 0.83 km
  - Coastline is 6.45 km
  - Closest City is 16.26 km

BUILD A DASHBOARD WITH PLOTLY DASH

SECTION 4

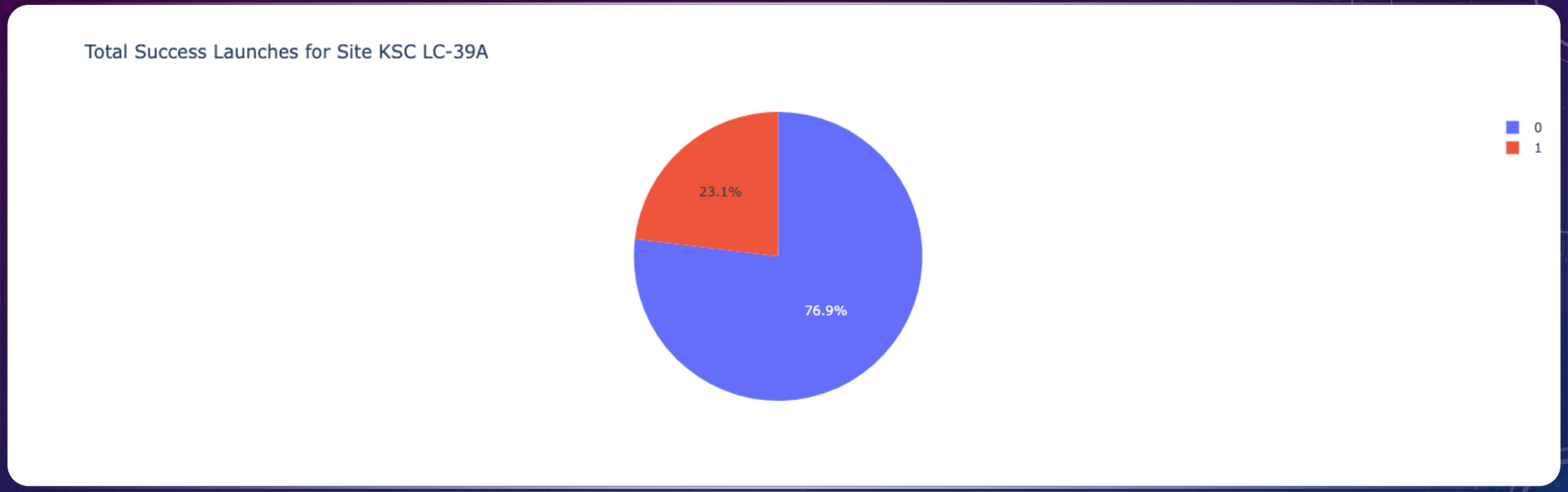
LAUNCH SUCCESS RATIO OF ALL SITE



Explanation:

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

LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO



Explanation:

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

PAYLOAD VS. LAUNCH OUTCOME

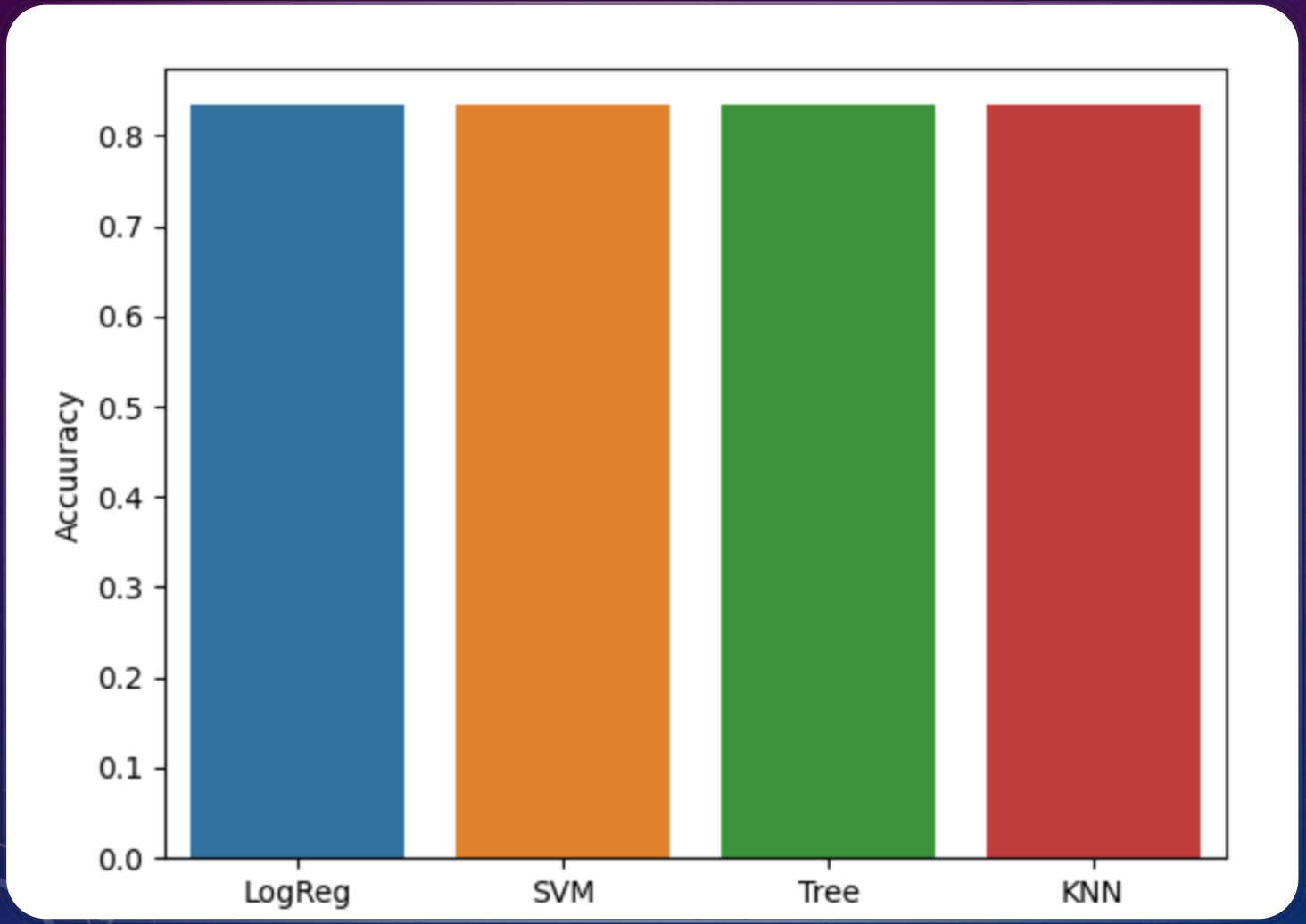
Explanation:

- Class indicates 1 for successful landing and 0 for failure.
- In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.
- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

PREDICTIVE ANALYSIS (CLASSIFICATION)

SECTION 5

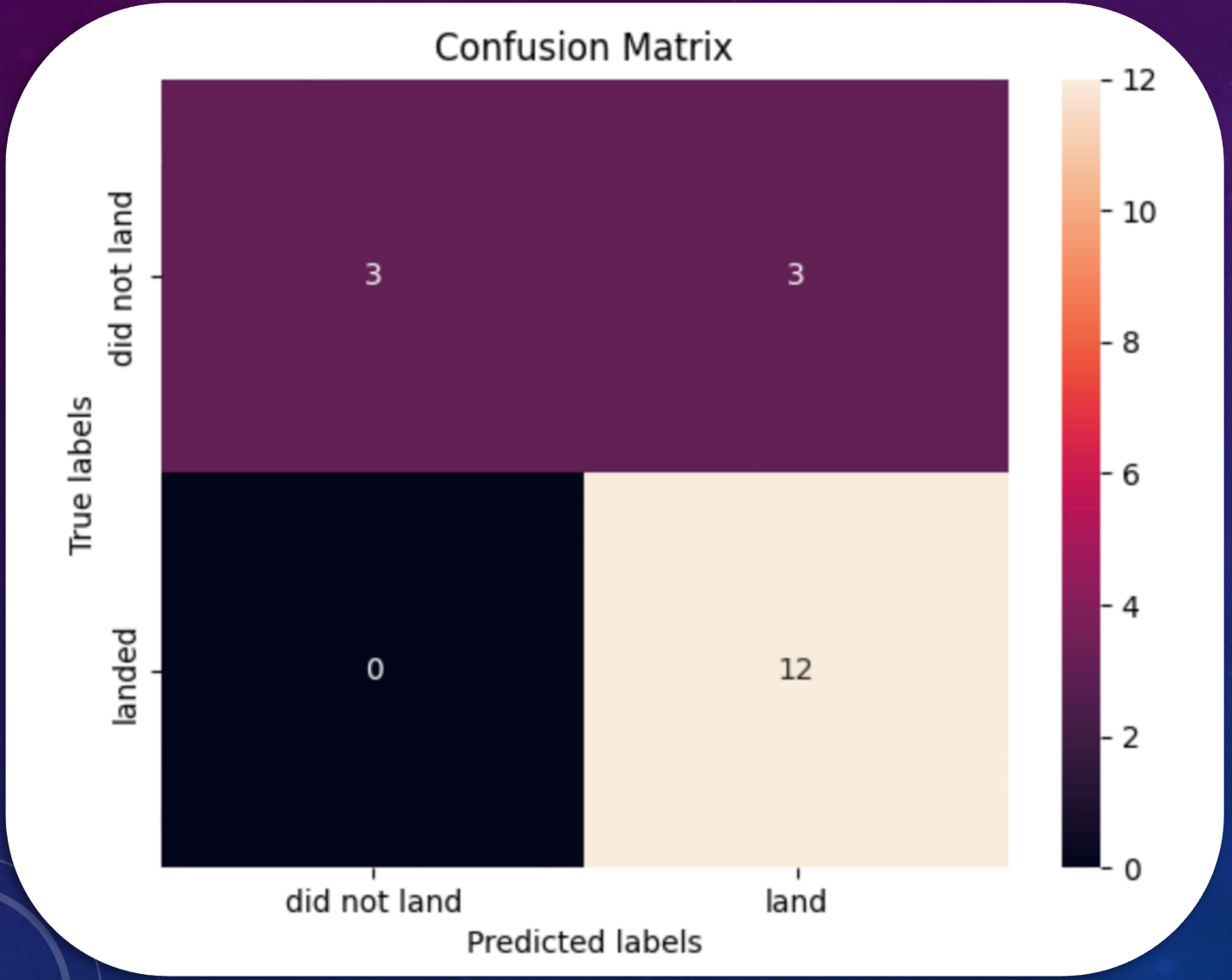
CLASSIFICATION ACCURACY



Explanation:

- All models had virtually the same accuracy on the test set at 83.33% accuracy.
- It should be noted that test size is small at only sample size of 18.
- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.

CONFUSION MATRIX



Explanation:

- Correct predictions are on a diagonal from top left to bottom right.
- Since all models performed the same for the test set, the confusion matrix is the same across all models. 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).
- Our models over predict successful landings

CONCLUSIONS

The success rate of launches increases over the years.

For every launch site the higher the payload mass, the higher the success rate.

All launch sites are very close to the coast, railway and highway. Furthermore, all launch sites are located at a safe distance from cities.

KSC LC-39A has the highest success rate of the launches from all the sites.

With payloads less than 4000 kg the successful landing are more for orbits SSO,HEO and MEO.

With heavy payloads the successful landing are more for orbits LEO and ISS.

Orbits ES-L1, GEO, HEO and SSO have 100% success rate

All models had virtually the same accuracy on the test set at 83.33%. We likely need more data to determine the best model.

APPENDIX

GitHub repository URL :
https://github.com/Pankati/Capstone

Special Thanks to:
Instructors
Coursera
IBM

THANK YOU