IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dimas Jackson
May 1, 2023

Outline

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

Executive Summary

Summary of methodologies

- Data collection from Space X database and web
- Descriptive data analysis
- Predictive analysis with Machine Learning

Summary of all results

- The launch site KS LC-39A, payload mass between 2500-5000 kg and orbit type SSO have highest success rate.
- We can predict if a launch will be successful or fail with 87% accuracy in test data

Introduction

- In this project, I predicted if the Space X Falcon 9 first stage will land successfully

- I collected, processed and analyzed data and built, validated and compared Machine Learning models to understand the most relevant variables related to the improving of success rate.

Section 1

Methodology

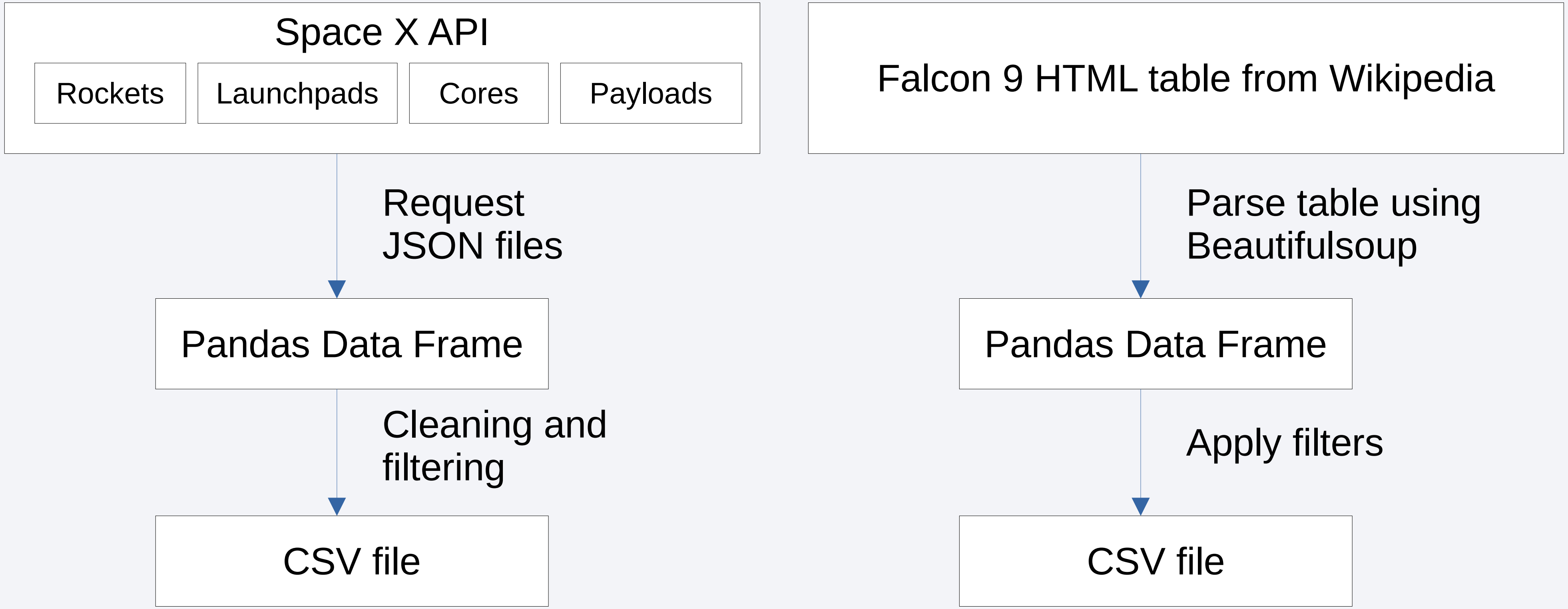
Methodology

Executive Summary

- Data collection methodology:
  - Data extraction from Space X API and webscraping from Wikipedia
- Perform data wrangling
  - Missing values treatment and class labeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium & Plotly Dash
- Perform predictive analysis using classification models
  - Hyperparameter tuning for SVM, Classification Trees and Logistic Regression

Data Collection

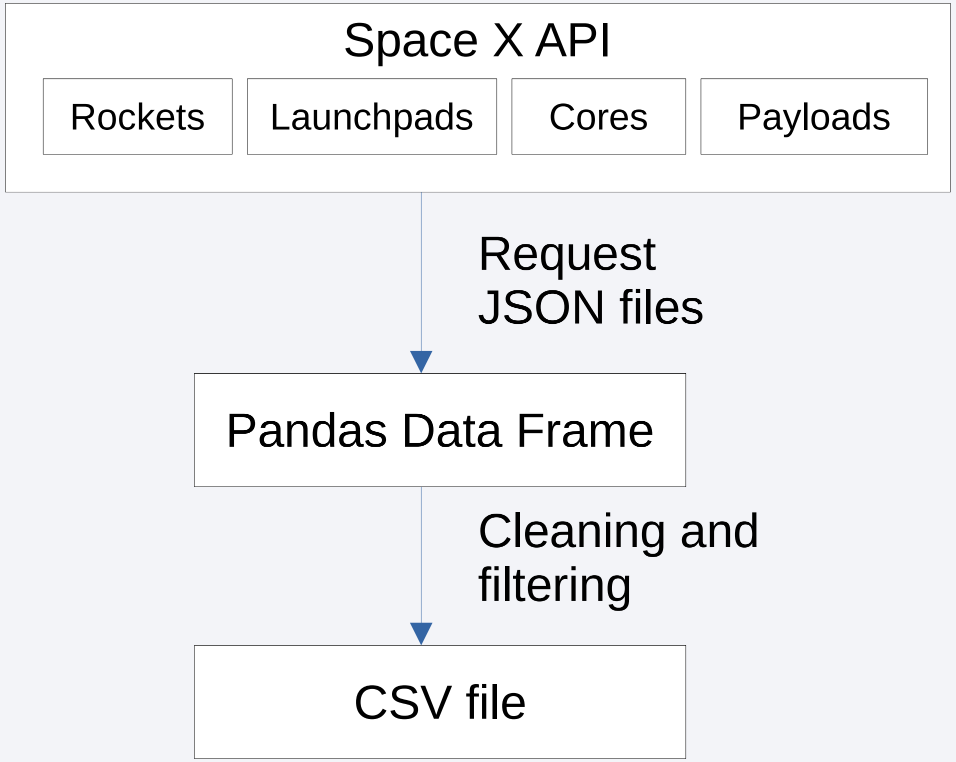
Space X API data request and webscraping from Wikipedia:



Data Collection – SpaceX API

- Extract and concatenate past launch data using ‘requests’ library
- Replace missing payload mass by the column average
- Filter by the relevant features to analysis

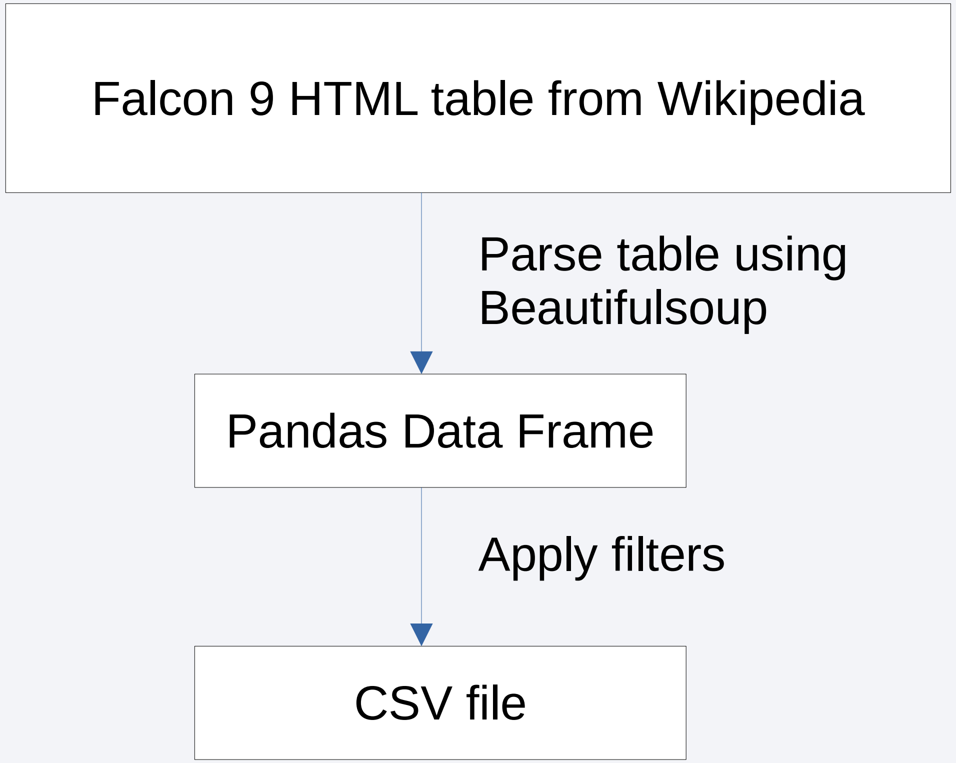
Data Collection Github link



Data Collection - Scraping

- Store the html table response in a txt file
- Parse headers, rows and coluns
- Create a Pandas data frame
- Export a CSV file

Webscraping Git Hub link



Data Wrangling

- Import the CSV files created in data collection
- Calculate the number of launches for each site
- Find the number of occurrences for each orbit
- Create a landing outcome label from Outcome column

Data Wrangling Git Hub link

EDA with Data Visualization

- Scatter plot FlightNumber vs. PayloadMass and Launch Site, to understand how these variables affect the launch outcome

- Use barplots to understand the relationship between success rate of each orbit type

- Visualize the launch success yearly trend using line chart

Data Vizualization Git Hub link

EDA with SQL

- Display average payload mass carried by booster version F9 v1.1
- 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 names of the booster_versions which have carried the maximum payload mass.
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

EDA with SQL Git Hub Link

Build an Interactive Map with Folium

- Mark the success/failed launches for each site on the map
- Mark down a point on the closest coastline and calculate the distance between the coastline point and the launch site.
- Draw a line between a launch site to its closest city, railway and highway

Folium Map Analysis Git Hub Link

Build a Dashboard with Plotly Dash

- Pie chart to show the total successful launches count for all sites
- Scatter chart to show the correlation between payload and launch success.
- Pie chart to show the Success vs. Failed counts for the site

Dashboard Git Hub Link

Predictive Analysis (Classification)

- Standardize the data
- Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression using cross validation
- Find the method performs best using test data

Machine Learning Analysis Git Hub Link

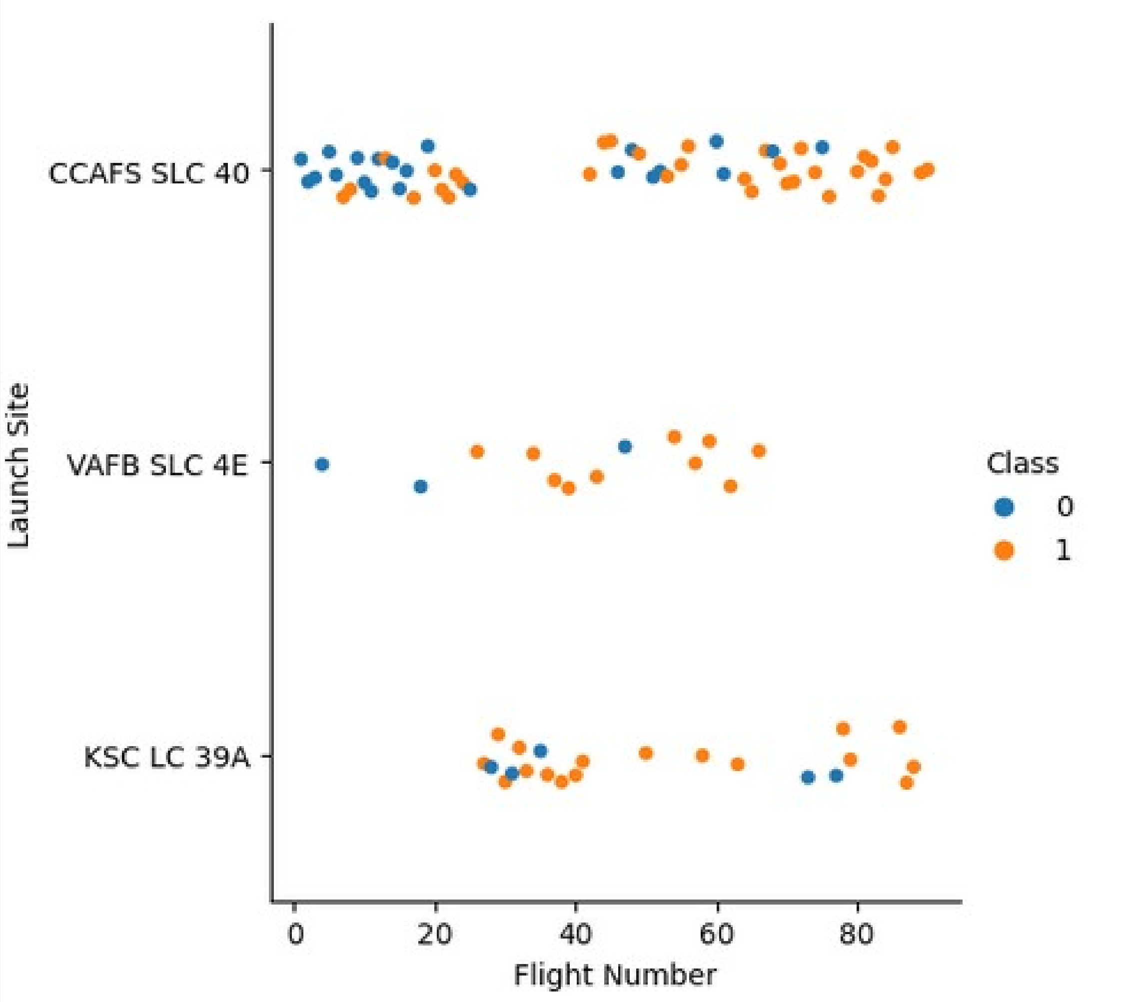
Results

- Success rate improve with flight number for all sites.
- The orbits ES-L1, GEO, HEO and SSO have high success rate.
- The success rate since 2013 kept increasing till 2020
- Payload mass range 2500-5000 kg has the highest launch success rate
- We can predict the landing success with 87% accuracy using Decision Tree model

Section 2

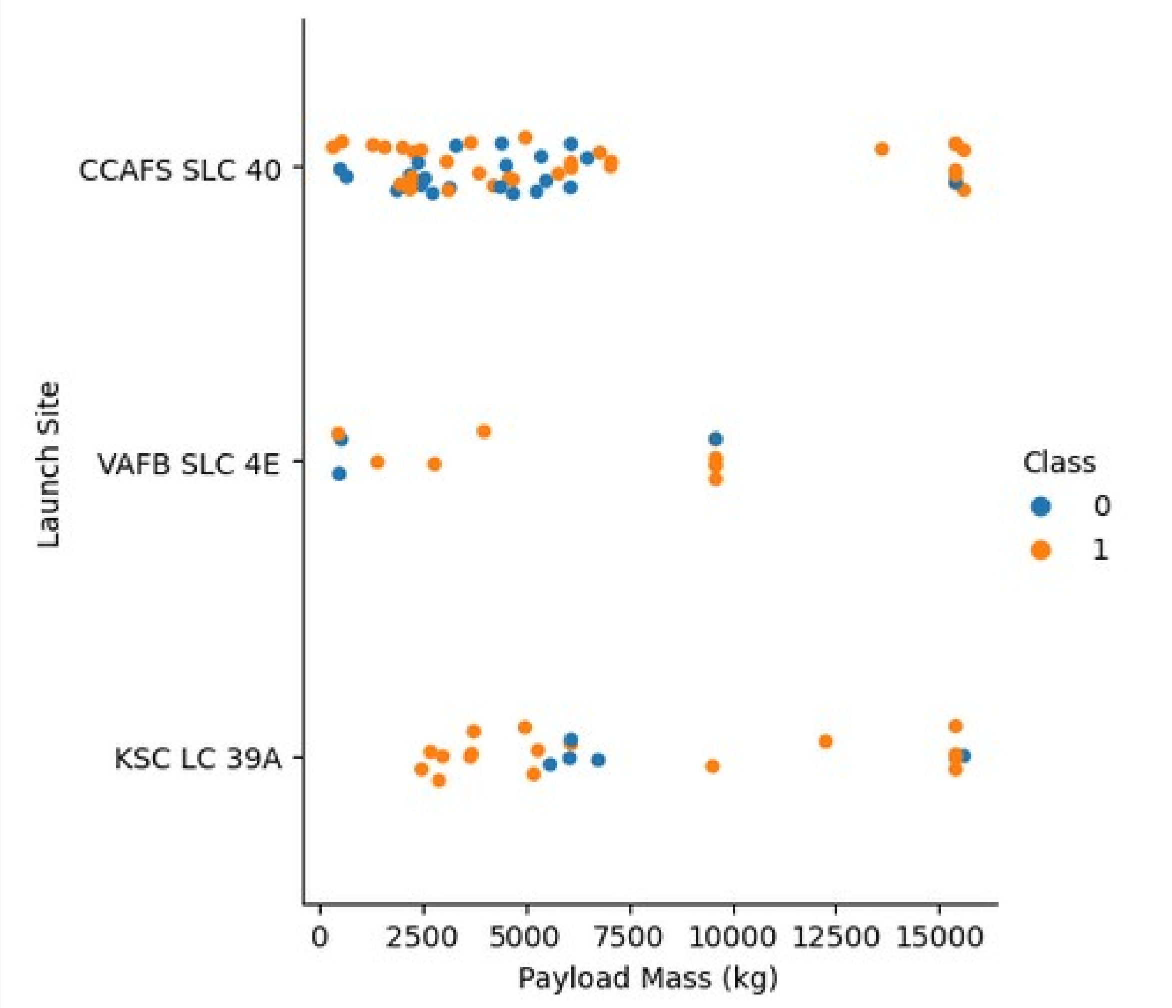
Insights drawn from EDA

Flight Number vs. Launch Site



- Success rate improve with flight number for all sites
- CCAFS SLC 40 had much more launches

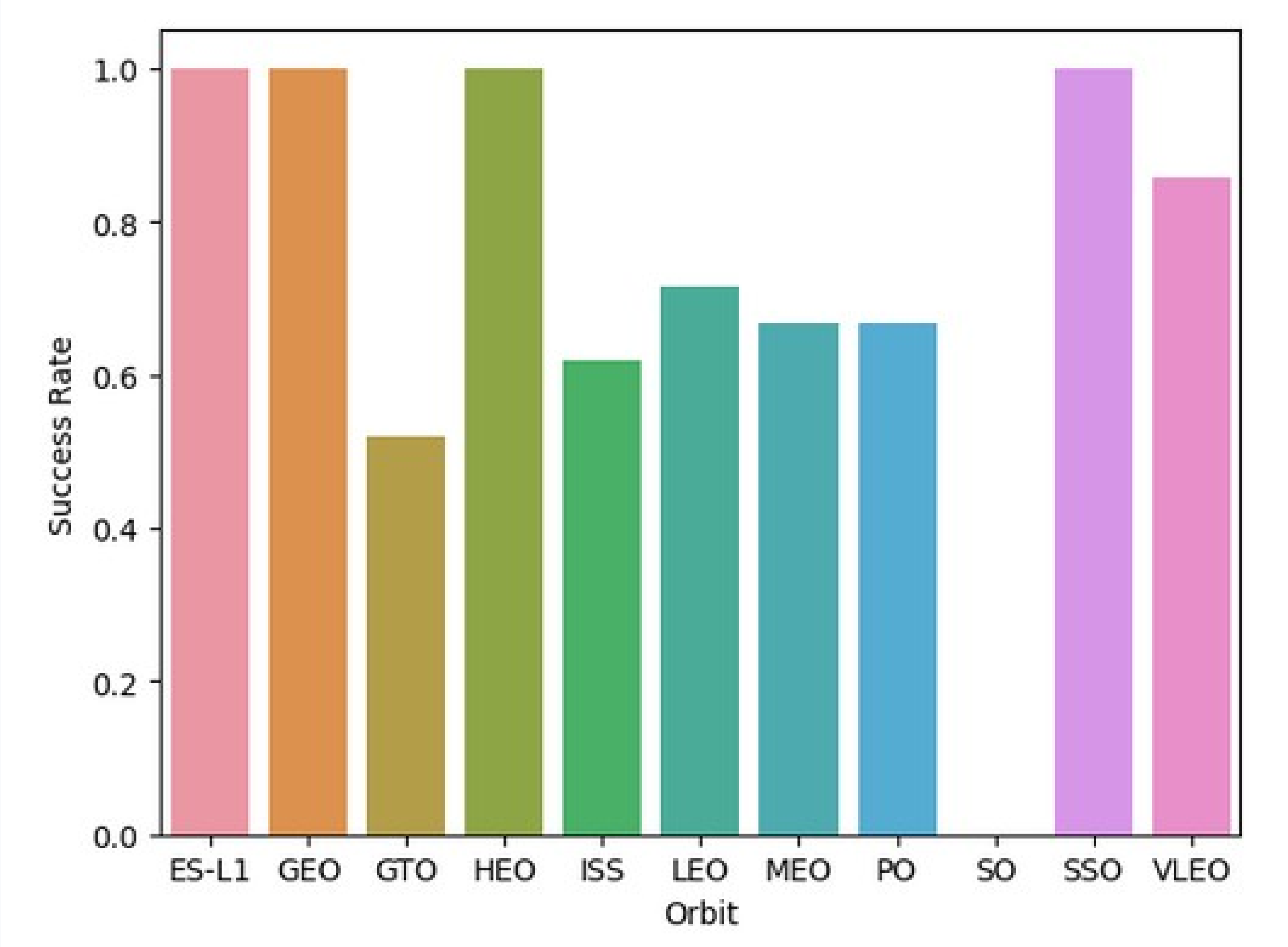
Payload vs. Launch Site



- There are more launches for lighter stages

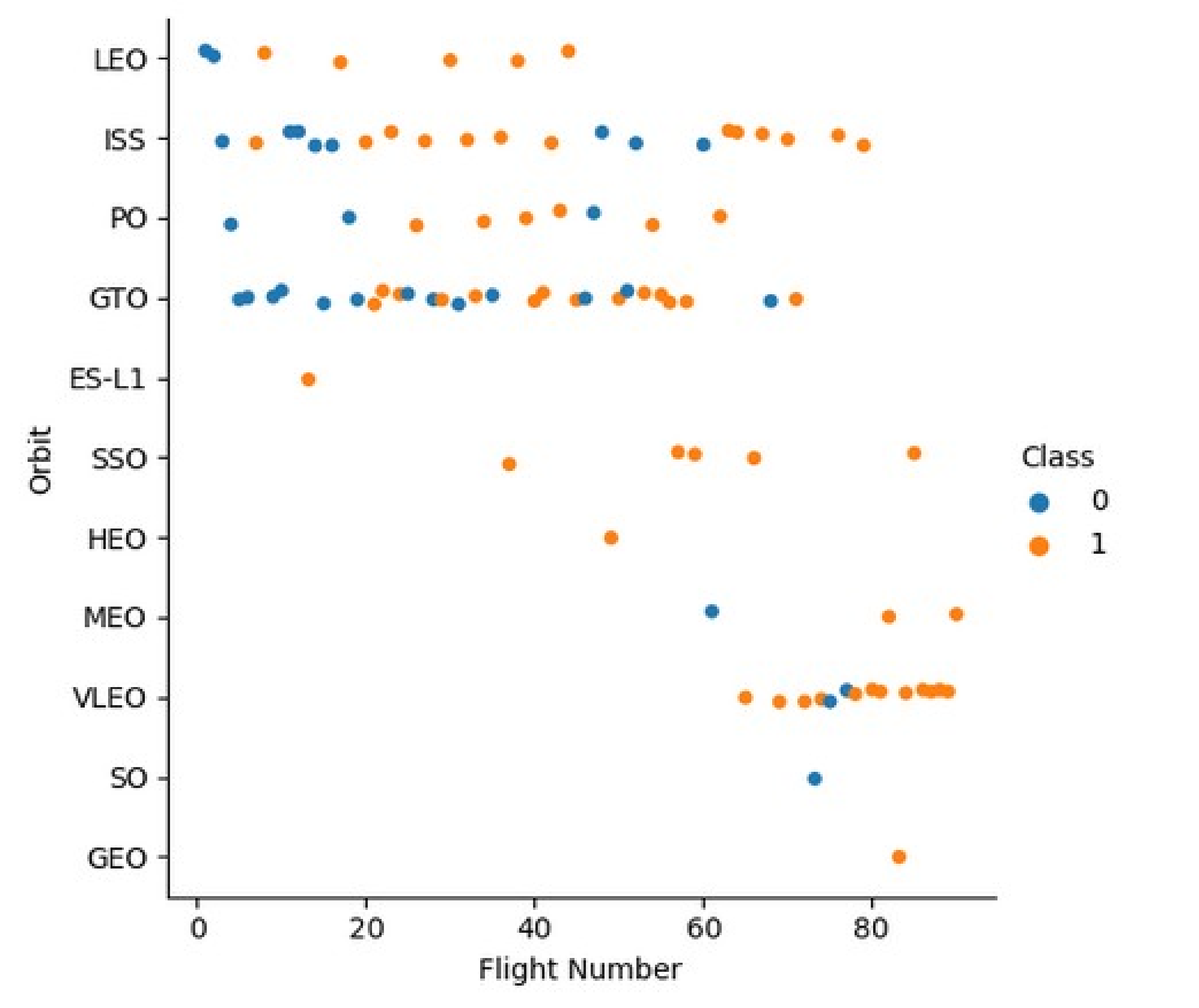
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



- The orbits ES-L1, GEO, HEO and SSO have high success rate
- SO orbit success rate is zero

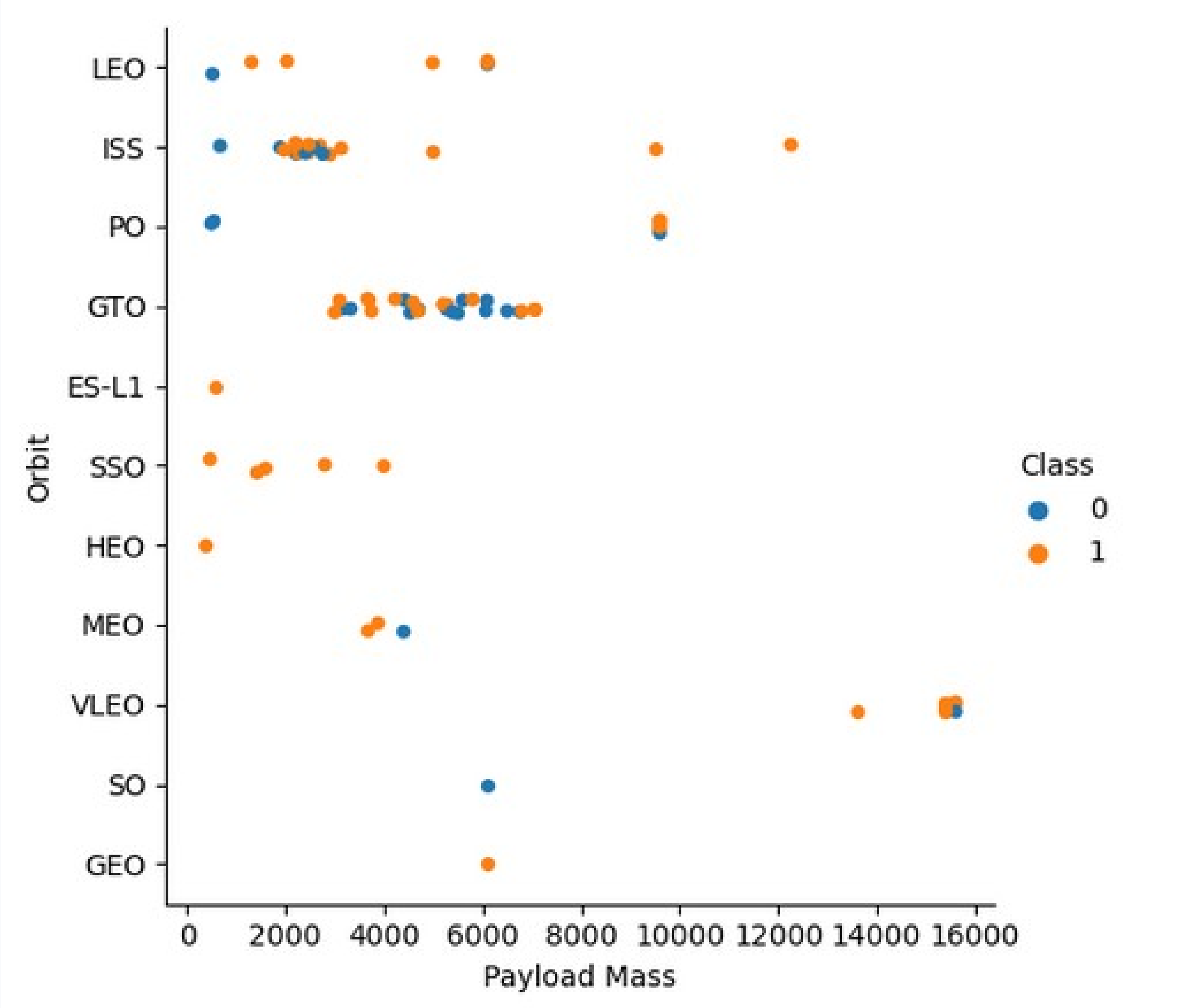
Flight Number vs. Orbit Type



- Although the success rate being 1, GEO, ES-L1 and HEO have just one launch

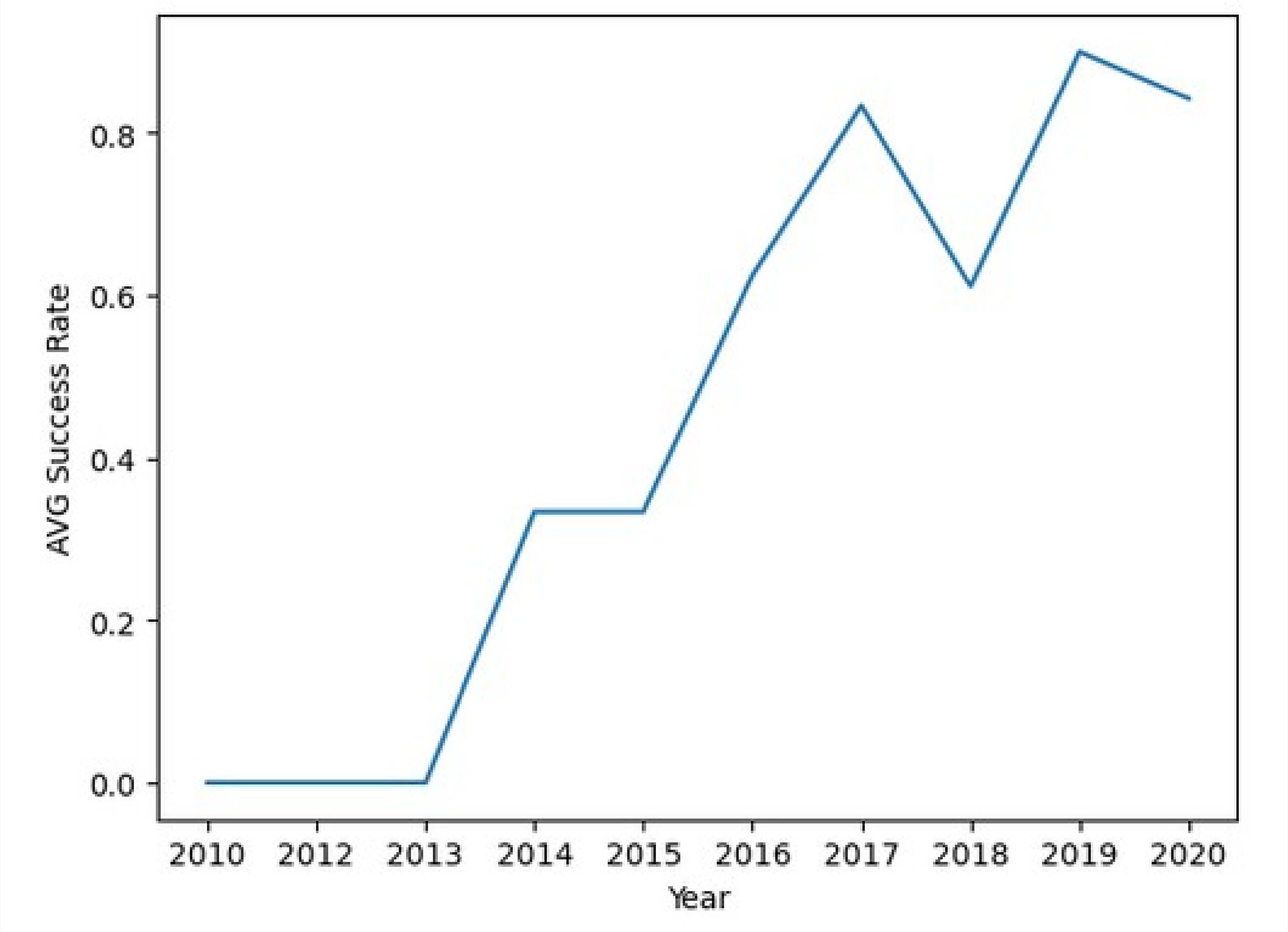
- The success rate seems Show the screenshot of the scatter plot with explanationsto increase with flight number for LEO but there are no relationship for GTO
- SSO looks the best choice

Payload vs. Orbit Type



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

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

The SQL query to find the distinct launch sites:

```
sql = """
        SELECT DISTINCT launch_site FROM launch;
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

	0
0	CCAFS SLC-40
1	KSC LC-39A
2	CCAFS LC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

SQL query to find 5 records where launch sites begin with `CCA`:

```
sql = """
        SELECT * FROM launch
        WHERE launch_site LIKE 'CCA%'
        LIMIT 5;
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

0	2010-06-04	None	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	None	18:45:00	Failure (parachute)
1	2010-12-08	None	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0.0	LEO (ISS)	NASA (COTS) NRO	Success	None	15:43:00	Failure (parachute)
2	2012-05-22	None	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	None	07:44:00	No attempt
3	2012-10-08	None	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	None	00:35:00	No attempt
4	2013-03-01	None	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	None	15:10:00	No attempt

Total Payload Mass

SQL query to calculate the total payload carried by boosters from NASA

```
sql = """
        SELECT SUM(payload_mass__kg_) FROM launch
        WHERE payload LIKE '%CRS%';
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

	0
0	111268.0

Average Payload Mass by F9 v1.1

SQL query to calculate the average payload mass carried by booster version F9 v1.1

```
sql = """
    SELECT AVG(payload_mass__kg_) FROM launch
    WHERE booster_version LIKE '%F9 v1.1%';
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

	0
0	2534.666667

First Successful Ground Landing Date

SQL query to find the dates of the first successful landing outcome on ground pad

```
sql = """
    select MIN(date) from launch2
    where landing_outcome like '%Succes%' and landing_outcome like '%ground%';
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

	0
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
sql = """
    select booster_version from launch2
    where landing_outcome like '%Succes%' and landing_outcome like '%drone%'
        and payload_mass__kg_ between 4000 and 6000;
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

	0
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

SQL query to calculate the total number of successful and failure mission outcomes

```
sql = """
    select count(1) from launch2 l
    where mission_outcome like '%Success%';
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

	0
0	100

```
sql = """
    select count(1) from launch2 l
    where mission_outcome like '%Failure%';
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

	0
0	1

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

```
sql = """
    SELECT Booster_Version FROM launch2
    WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM launch2
);
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

List of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
sql = """
    select landing_outcome, booster_version, launch_site, "date" from launch2
    where landing_outcome like '%Fail%' and landing_outcome like '%drone%'
    and extract(year from "date") = 2015;
"""
cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

	0	1	2	3
0	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
1	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

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

```
sql = """
    select count(1) as count_o, landing_outcome from launch2
    where "date" between '2010-06-04' and '2017-03-20'
    group by landing_outcome
    order by count_o desc;
"""

cursor.execute(sql)
data = cursor.fetchall()
pd.DataFrame(data)
```

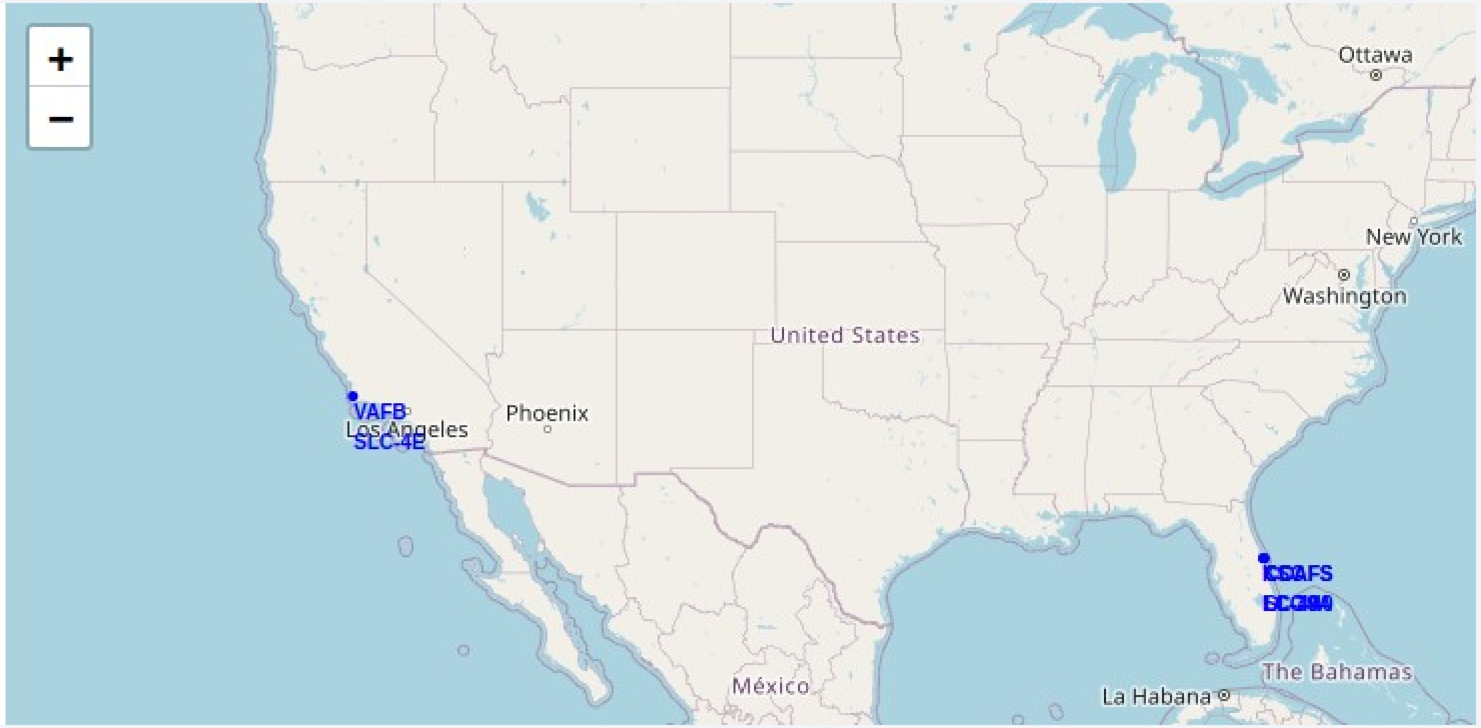
0	10	No attempt
1	5	Failure (drone ship)
2	5	Success (drone ship)
3	3	Success (ground pad)
4	3	Controlled (ocean)
5	2	Uncontrolled (ocean)
6	2	Failure (parachute)
7	1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

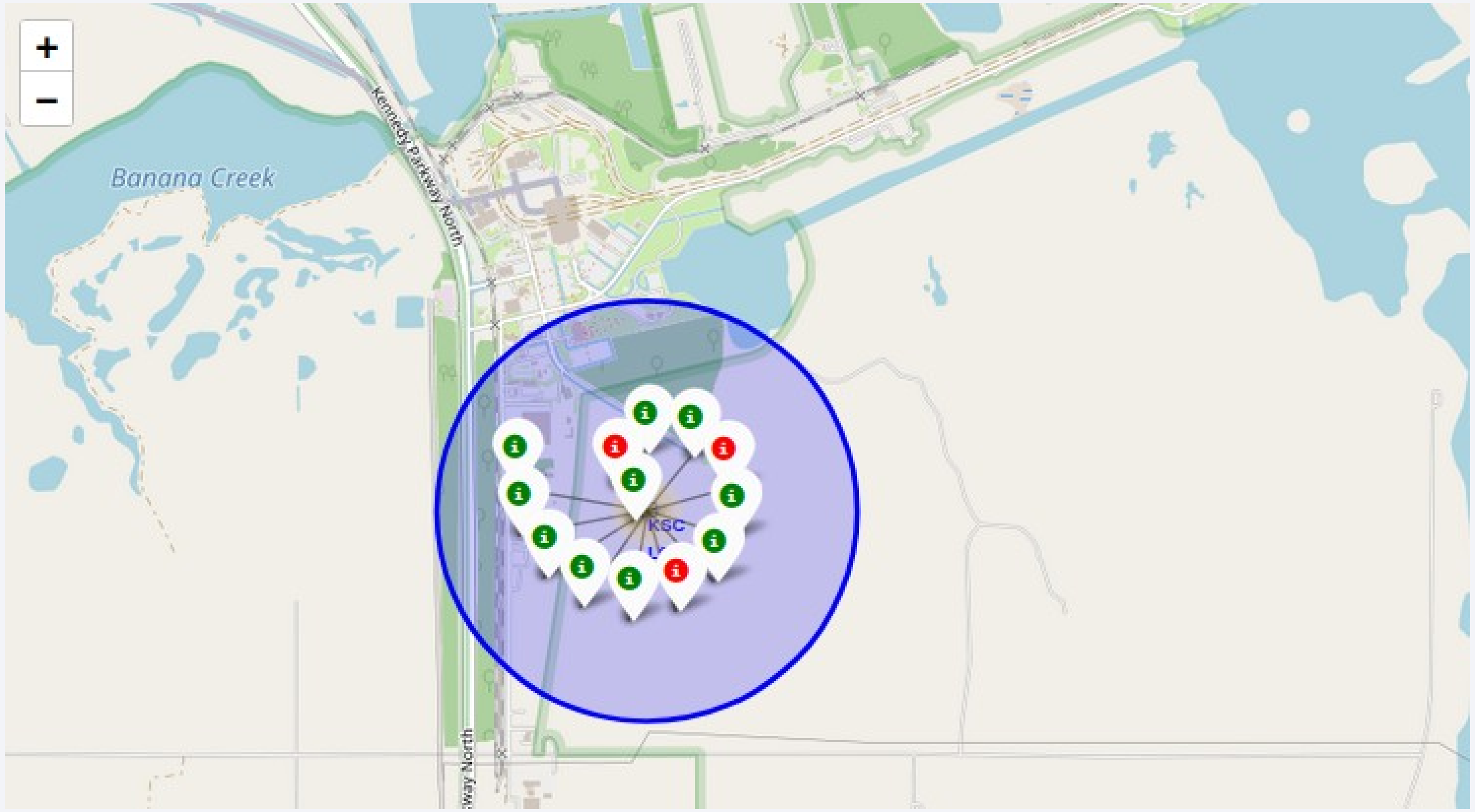
<Folium Map Screenshot 1>

Generated folium map with all launch sites’ location markers on a global map



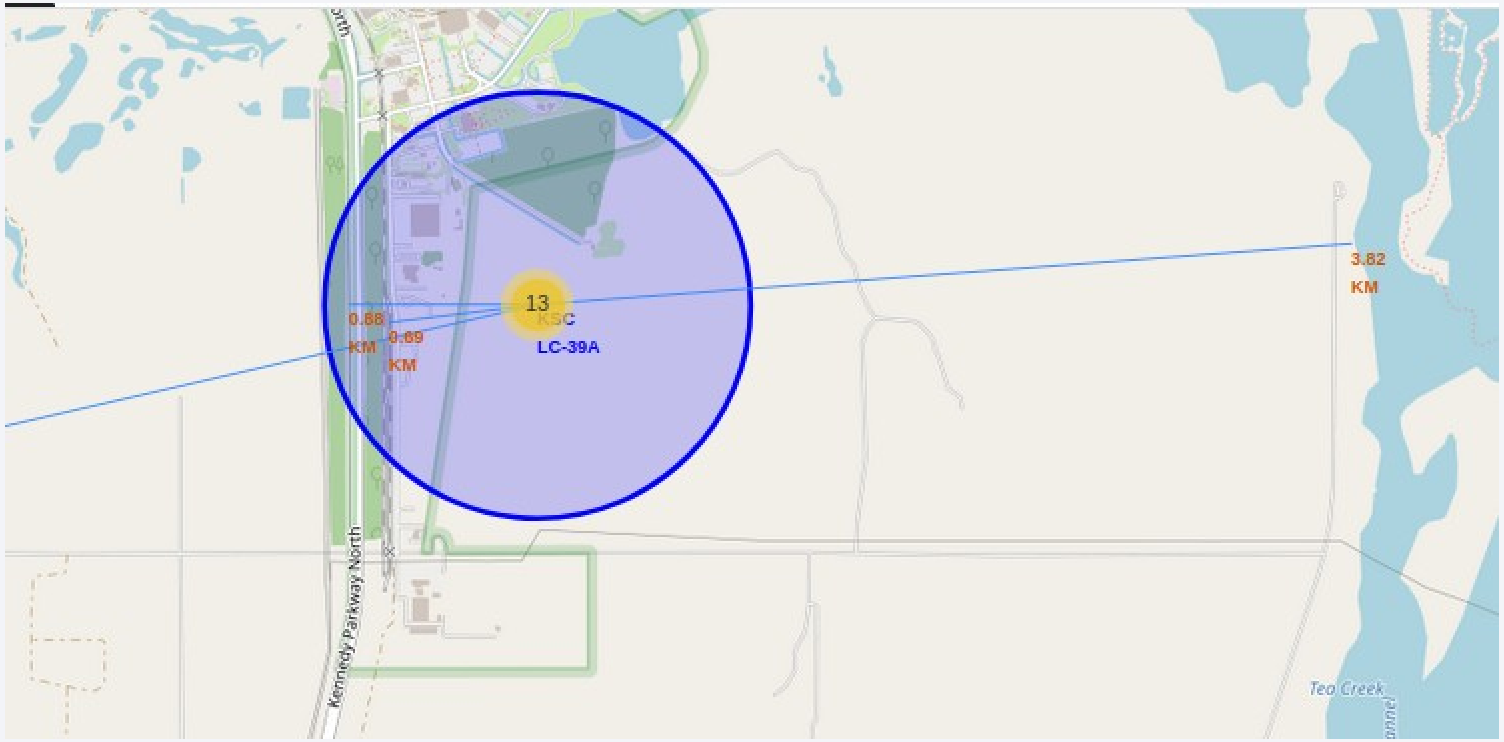
<Folium Map Screenshot 2>

Color-labeled launch outcomes in KSC LS: success (green) failure (red)



<Folium Map Screenshot 3>

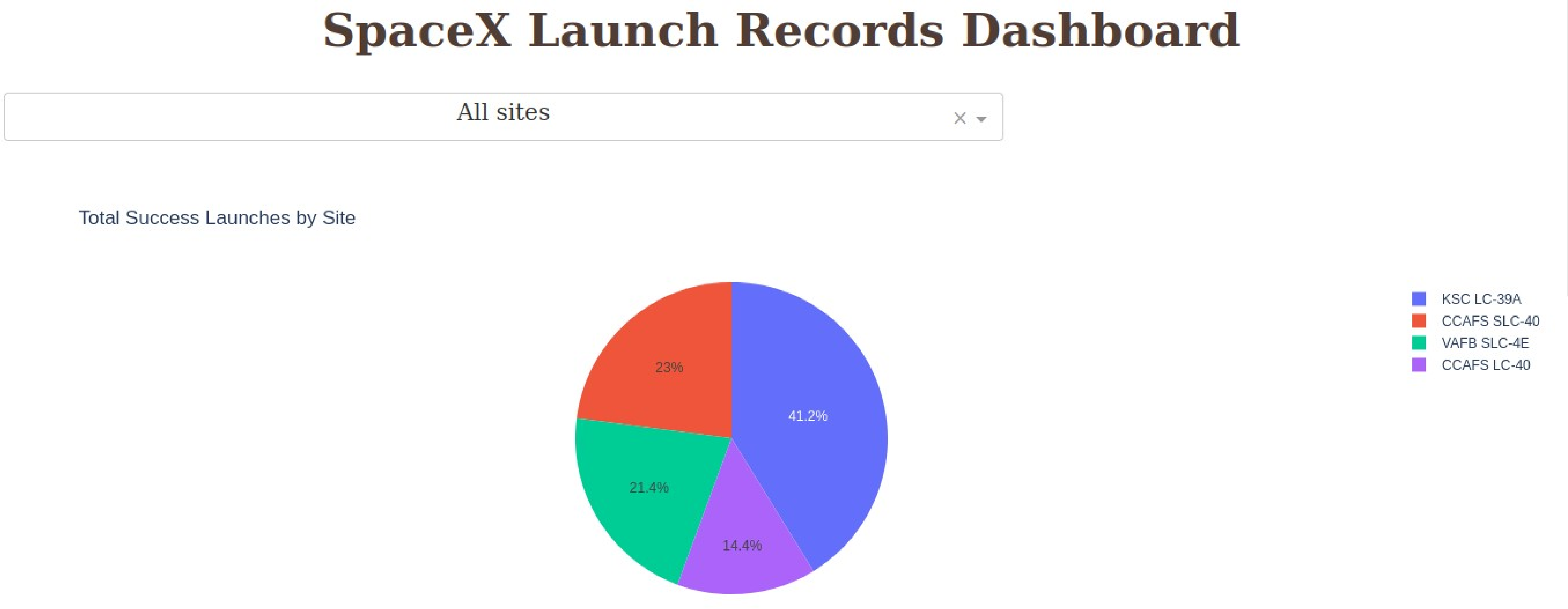
Launch site to its proximity railway, highway, coastline, with distance



Section 4

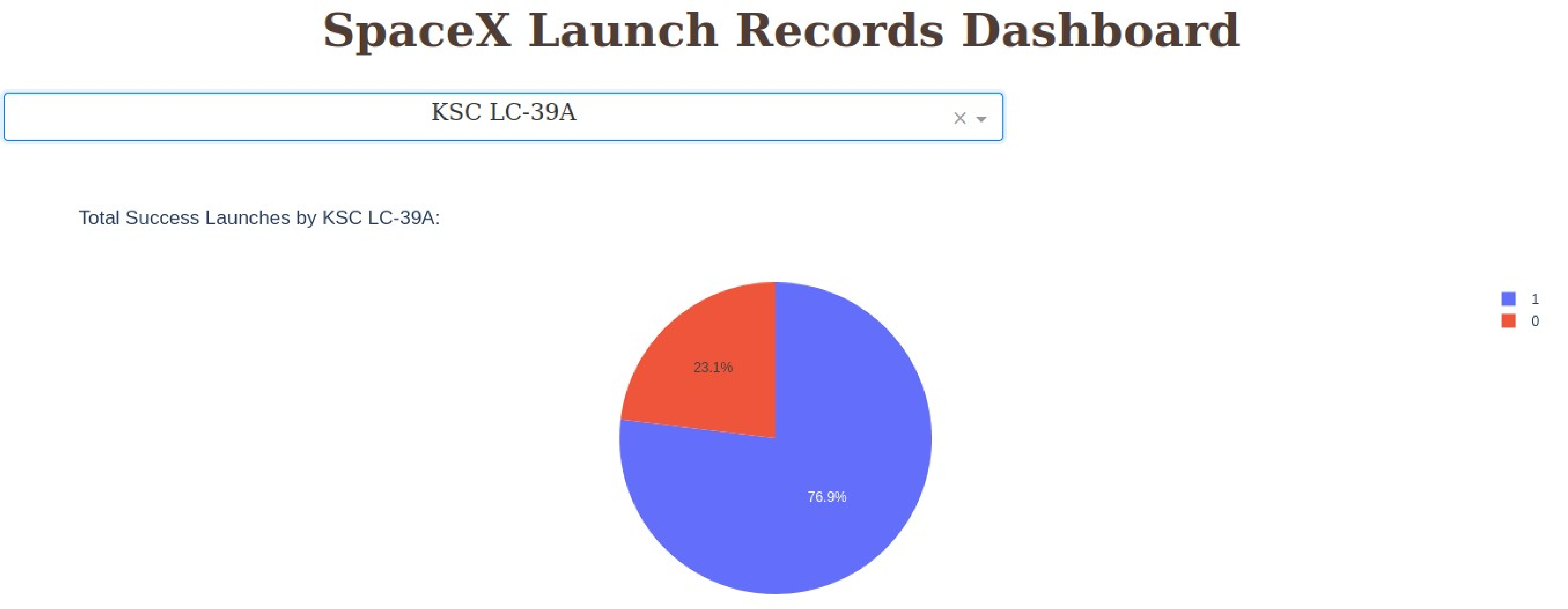
Build a Dashboard with Plotly Dash

Interactive Plot Success Launches by Site



The KSC LC-39A site had the highest number of success launches

Interactive Plot Launch Site with highest success



76.9% of launches had success in site KSC LC-39A

Interactive Plot Payload Mass vs. Success Class

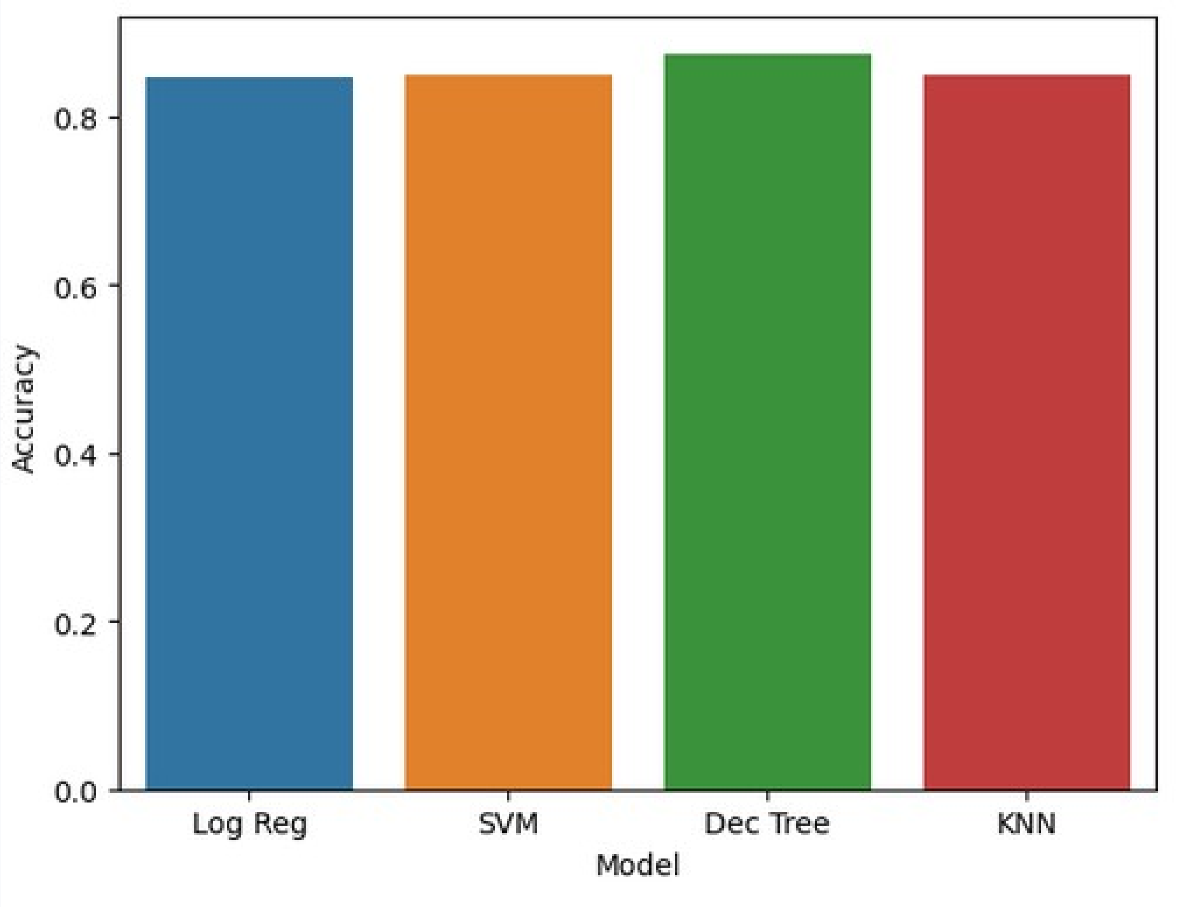
The best payload mass range was 2500-5000 kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

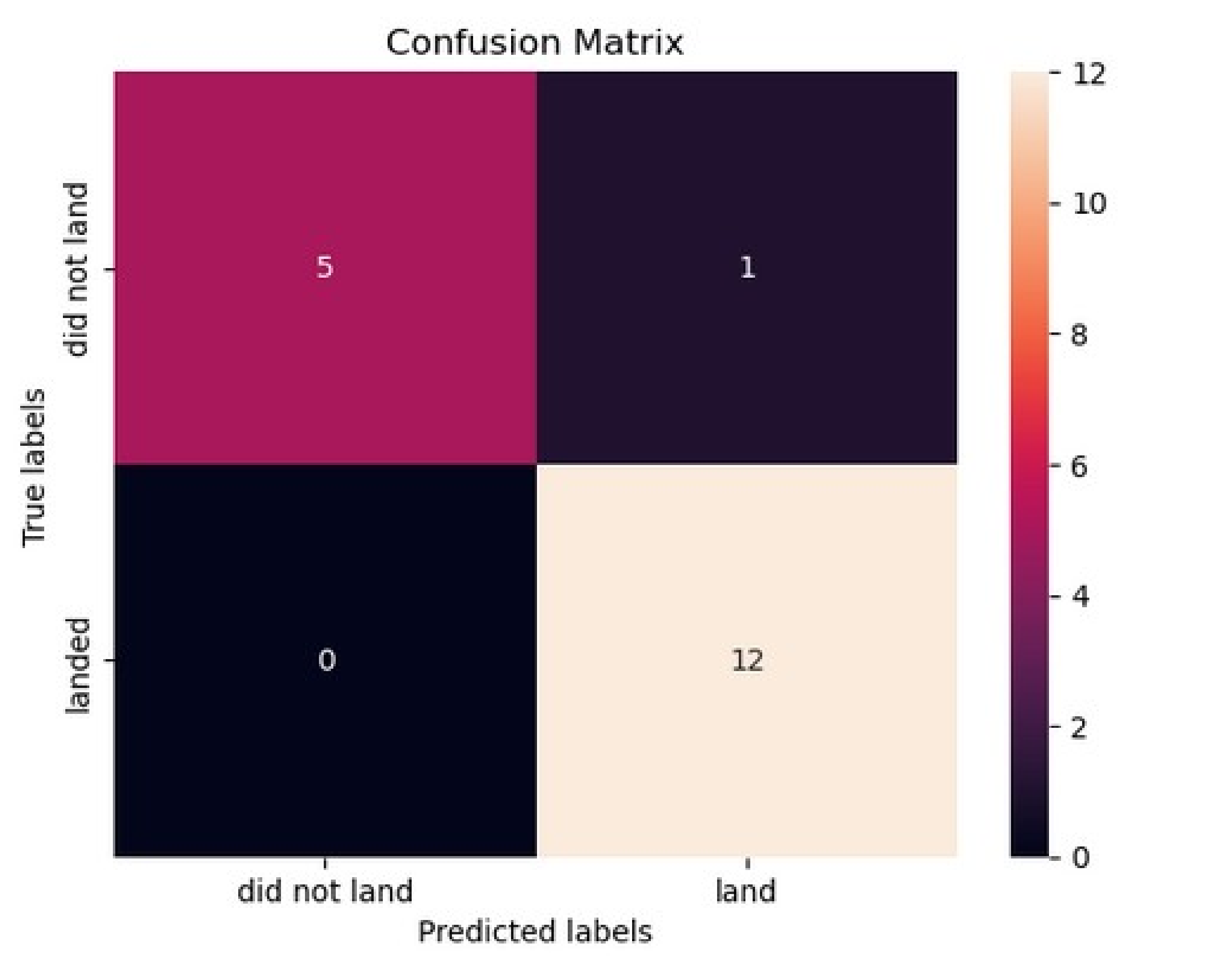
- All models considered have accuracy greater than 80%
- The best classifier was Decision Tree, with 87.5% accuracy



Confusion Matrix

- The Decision Tree model performed well in predicting the successful landing

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18



Conclusions

The better choice to improve the success rate is: Launch site KS LC-39A, payload mass between 2500-5000 kg, orbit type SSO.

Given the payload mass, launch site, and orbit type we can predict if a launch will be successful or fail with 87% accuracy in validation data

Space X may save USD 100 million by reusing the first stage of Falcon 9, so this findings could improve the company profit

Appendix

Data set created during this project before and after one hot encoding

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	...
0	1.0	6104.959412	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	2.0	525.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	3.0	677.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	...
3	4.0	500.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	5.0	3170.000000	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...
...	...	...	...	...	...	...	...	...	...	...	...
85	86.0	15400.000000	2.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...
86	87.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...
87	88.0	15400.000000	6.0	5.0	5.0	0.0	0.0	0.0	0.0	0.0	...
88	89.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...
89	90.0	3681.000000	1.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	...

Thank you!