IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sunita Karmani
12/08/2023

Outline

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

Executive Summary

- **Summary of Methodologies:**
  - This research is done to identify the factors which determine the safe landing of the rocket Falcon 9 of SpaceX. The following methodology was used to determine the results:
    - **Data collection**:SpaceX REST API and webscraping tools.
    - **Data wrangling**: transforming data into meaningful data for further analysis.
    - **EDA**: Performing data visualization techniques to see the relation between the variables.
    - **Analysis**: Analysing the data using SQL queries to determine total number of successful landings.
    - **Model Building:** building models to predict the landing outcomes using different modelling techniques.
- **Summary of all Results:**
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES -L1, GEO, HEO, and SSO have a 100% success rate

Introduction

- **Project background and context:**
  - SpaceX is a recognized leader in commercial space exploration. It launches satellites, cargo and manned spacecraft. But the basis of its activities are launch vehicles, each of which is unique in some way. Falcon 9 is the second rocket created by SpaceX. Its development was announced back in 2005. Falcon 9 is the most massive commercial rocket in the world, and also one of the most reliable carriers in general. Its latest versions are capable of putting 22,800 kg of payload into low Earth orbit, and the first stage can make a controlled landing and then be reused.

- **Problems you want to find answers:**
  - The project is aimed to determine the various factors which affect the safe first landing of the rocket such as payload mass,launch site, number of flights and orbits.
  - To also determine the best predictive model for successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using Spacex REST API and Web Scraping related Wiki pages.
- Perform data wrangling
  - Data was filtered, missing values were handled, and one hot encoding was used to prepare data for further analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Various modelling techniques were used to determine the best predictive model for analysis.

Data Collection- SpaceX API

Request Data from : api.spacexdata.com/v4/launches/past

Github URL:

Used .json()	Filter Data	Null Values Handling
This method was used to create a JSON file. Then converted to a dataframe using JSON_normalize function.	In this step, data was filtered to include only Falcon 9 launches.	The last step involved the handling of null values which were present in PayLoad Mass column. They were replaced with the mean value.

Data Collection – Web Scraping using Wiki pages

Request Data	Data was requested from Wiki page of Falcon 9 launches
Beautiful soup creation and extraction of HTML tables	Using HTML response method, Beautiful Soup was created. Then columns were extracted from HTML table headers.
Dictionary and DataFrame Creation	HTML tables were parsed to create dictionary of the columns and its values. Then a DataFrame was formed using the Dictionary.

Data was then exported to a CSV file
Github URL:

Data Wrangling

- **Steps:**

1. EDA to determine Data Labels
2. Calculations:
   a. No. of launches of each site
   b. No. of orbit occurrence
   c. No. of mission outcome

3. Creation of binary outcome:

   0- unsuccessful landing

   1- successful landing

4. Export data to CSV file

Results:

- True Ocean: Successful landing on ocean
- False Ocean: ocean landing failed
- True RTLS: Successful landing on ground Pad
- False RTLS: Ground pad landing failed
- True ASDS: Successful drone ship landing
- False ASDS: Drone ship landing failed

The last step involved conversion of successful landing as a binary 1 and a failed landing as a binary 0.

Github URL:

EDA with Data Visualization

- In order to see if a relationship exists between various variables, **scatter plots** and **bar charts** were plotted.
- **Scatter plots:**
  - Flight number vs Launch Site
  - PayLoad vs Launch Site
  - Flight No. vs Orbit Type
  - Payload vs Orbit Type
- **Bar Chart:**
  - Success rate vs Orbit Type

GitHub URL:

EDA with SQL

- **Queries:**
  - Unique launch site names
  - 5 records where launch site begins with ‘CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date of first successful landing on ground pad
  - Names of boosters which had success landing on drone ship and have payload mass between 4,000 and 6,000.
  - Total number of successful and failed missions
  - Booster versions names which have carried the max payload
  - Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
  - Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

Github URL:

Build an Interactive Map with Folium

- **Markers Indicating Launch Sites:**
  - Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
  - Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates Map with Folium

- **Colored Markers of Launch Outcomes:**
  - Added green markers of successful and red unsuccessful launches at each launch site to show which launch sites have high success rates

- **Distances Between a Launch Site to Proximities:**
  - Added colored lines to show distance between launch site CCAFS SLC40 and its proximity to the nearest coastline, railway, highway, and city

- **GITHUB URL:**

Build a Dashboard with Plotly Dash

- **Dropdown List with Launch Sites:**
  - Select all launch sites or a specific launch site

- **Dashboard with Plotly Dash Slider of Payload Mass Range**
  - Select payload mass range

- **Pie Chart Showing Successful Launches:**
  - Percentage of successful and unsuccessful launches.

- **Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version:**
  - Correlation between Payload and Launch Success

Github URL:

Predictive Analysis (Classification)

1. Import all the relevant **libraries**
2. Create a NumPy Array from the column **Class**
3. Standardize the data with **StandardScaler.**
4. Split the data using **Train_Test_Split.**
5. Create a **GridSearchCV** object with cv=10
6. Apply the GridSearchCV to the various algorithms:
   a. Logistic regression
   b. Support vector machine
   c. Decision tree classifier
   d. K-Nearest Neighbor
7. Assess the **Confusion Matrix** of all models
8. Identify the best model using **Jaccard Score, F1 Score** and **Accuracy.**

Github URL:

Results

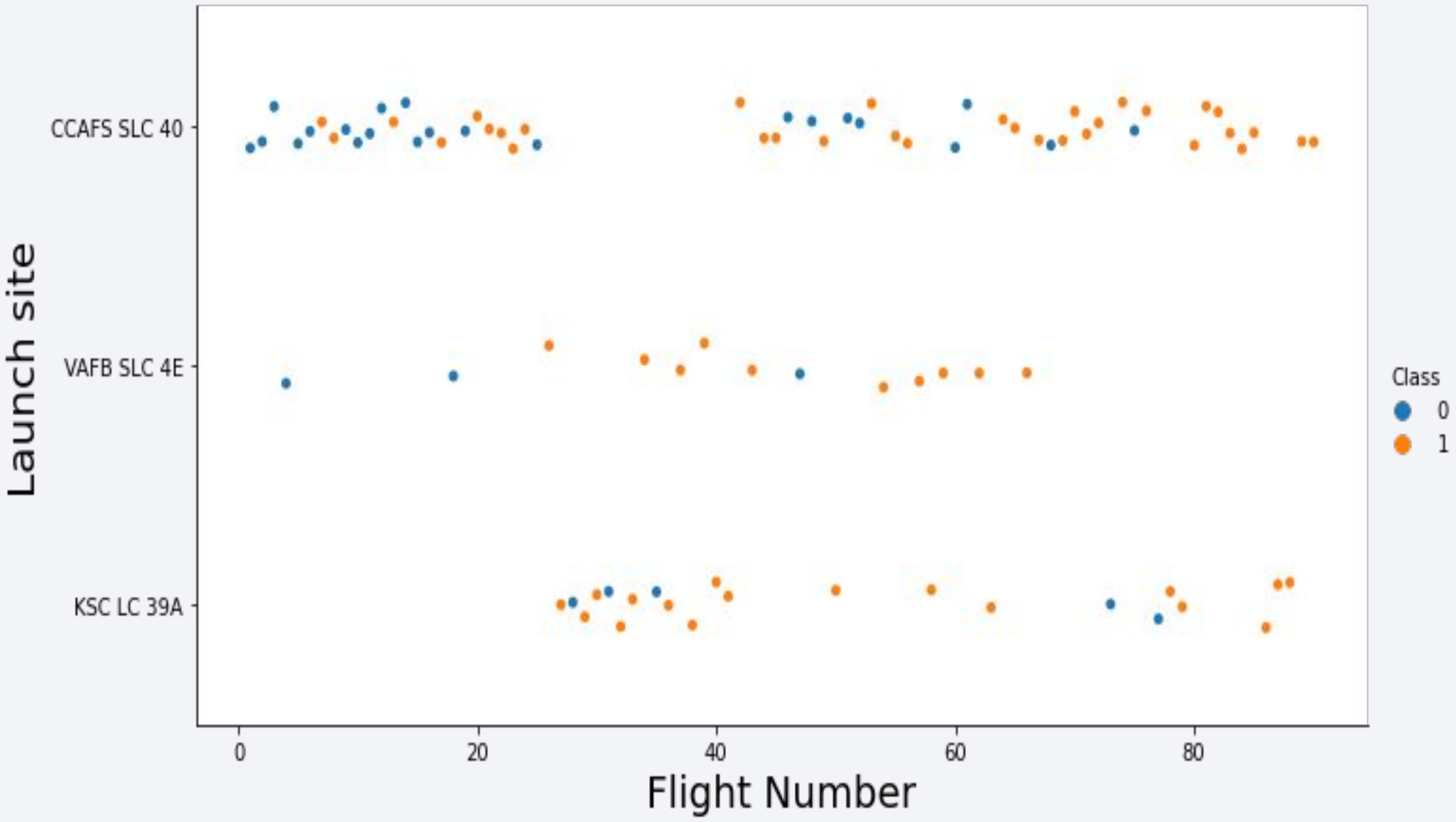
- **<u>Exploratory data analysis results:</u>**
    - KSC LC-39A has the highest success rate among landing sites
    - Launch success has improved over time.
    - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- **<u>Interactive analytics demo in screenshots:</u>**
    - Most launch sites are near the equator and coastline.
- **<u>Predictive analysis results</u>**
    - Decision Tree is the best model to predict

Section 2

Insights drawn from EDA

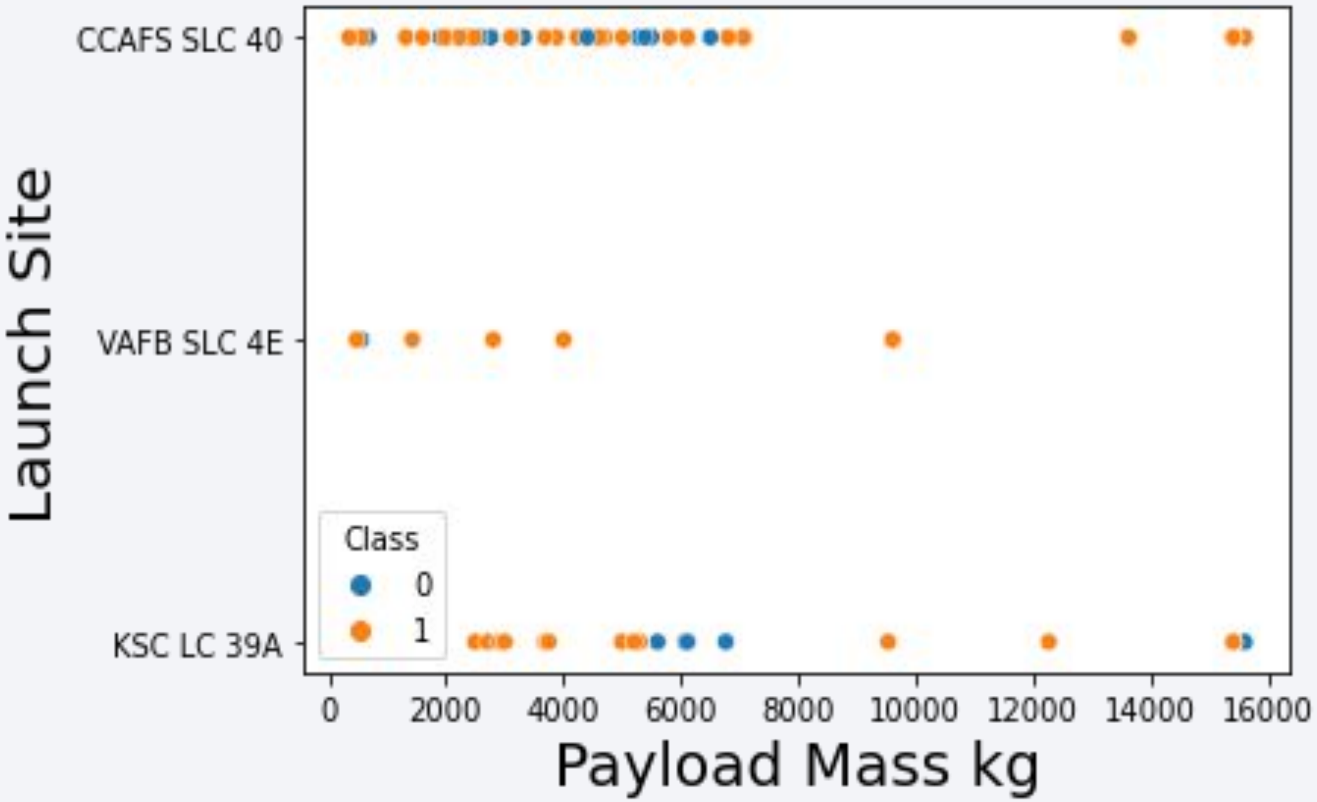
Flight Number vs. Launch Site

- Success rate increases as Flight number increases.
- **VAFB SLC 4E** and **KSC LC 39A** launch sites have higher success rates.
- CCAFS SLC 40 launch site report around 50% of launches.
- Blue dots represent failed
- Orange dots represent success



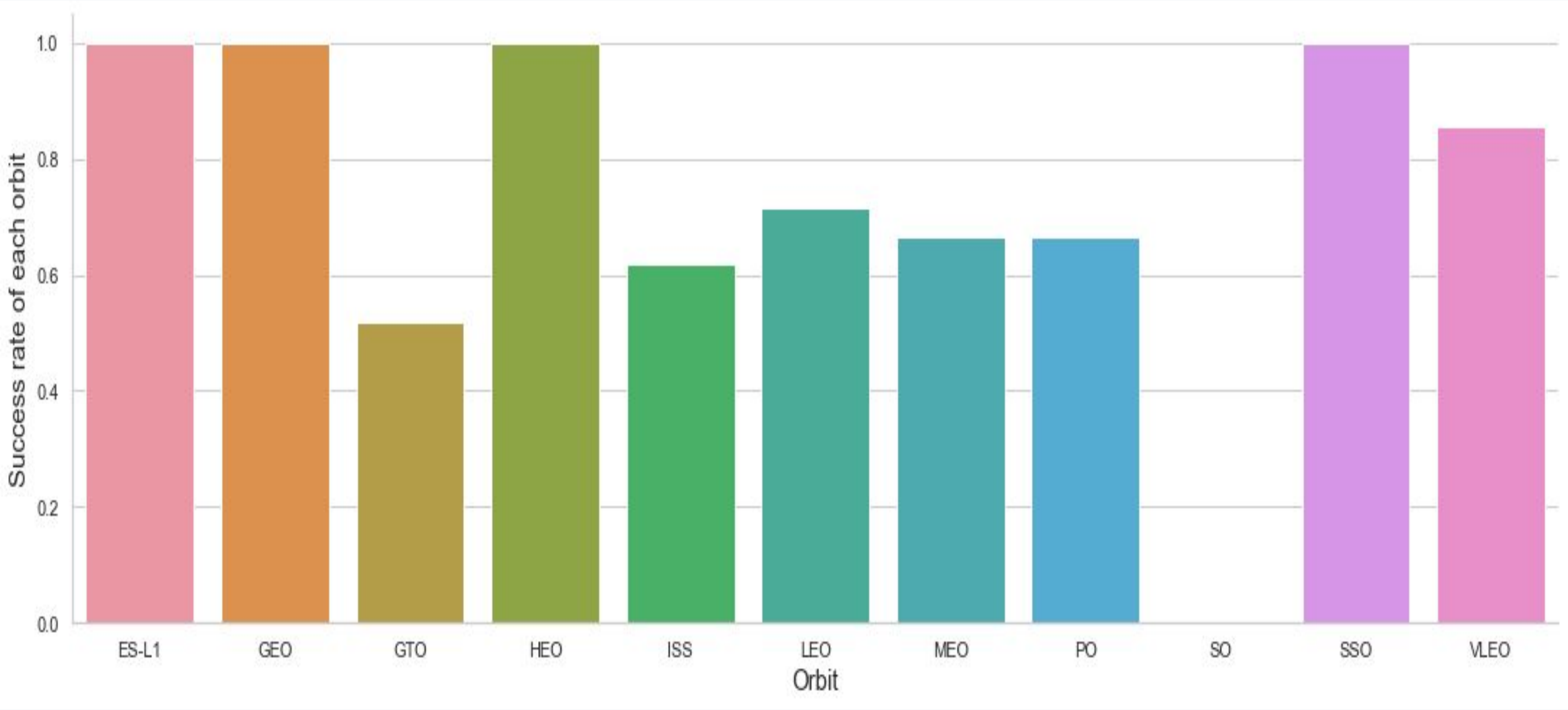
Payload vs. Launch Site

- Success rate is **directly proportional** to PayLoad Mass.
- Payload Mass of **7000kg** or more reported success.
- **KSC LC 39A** has a 100% success rate for launches less than 5,500 kg



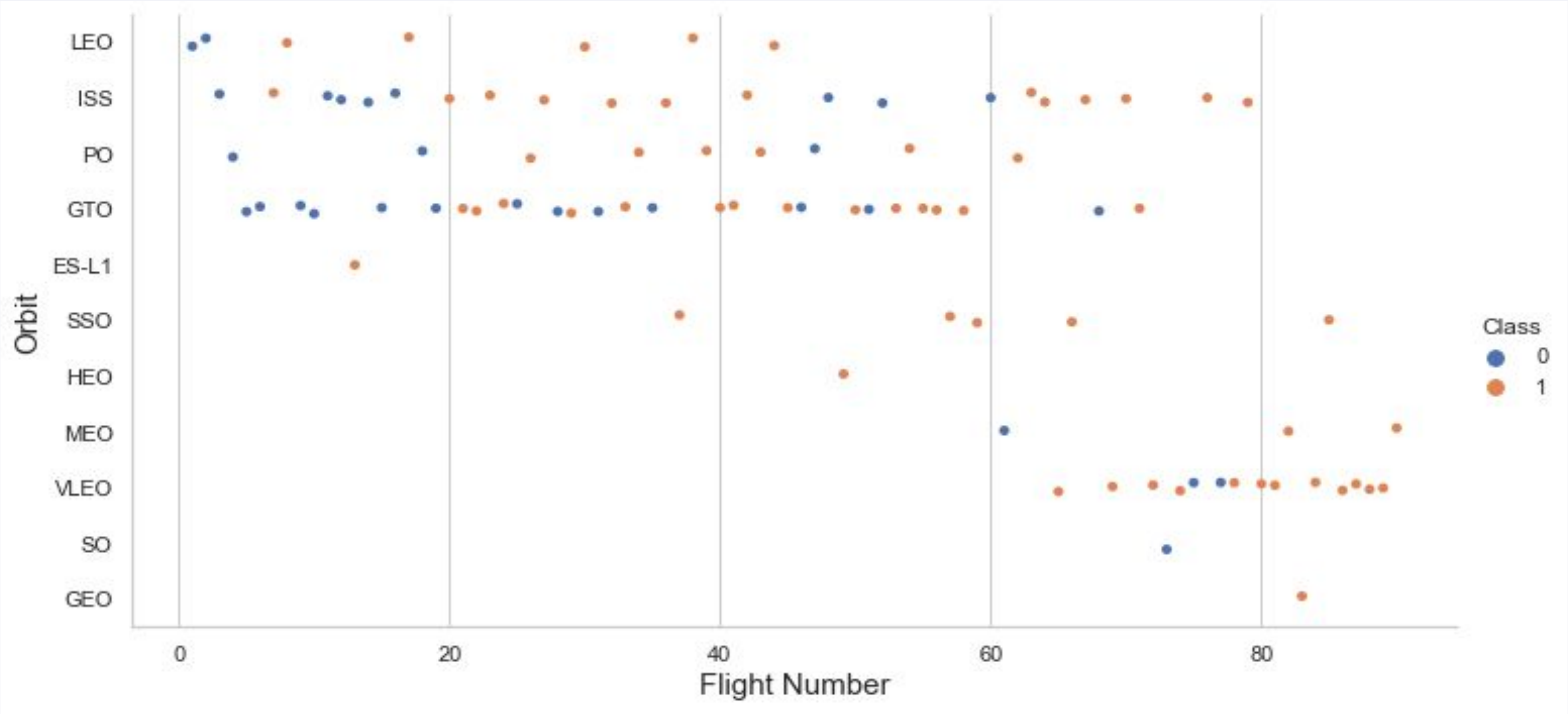
Success Rate vs. Orbit Type

- **100%** Success Rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- **50%-80%** Success Rate:
  - GTO
  - ISS
  - LEO
  - MEO
  - PO
- **0%** Success Rate:
  - SO



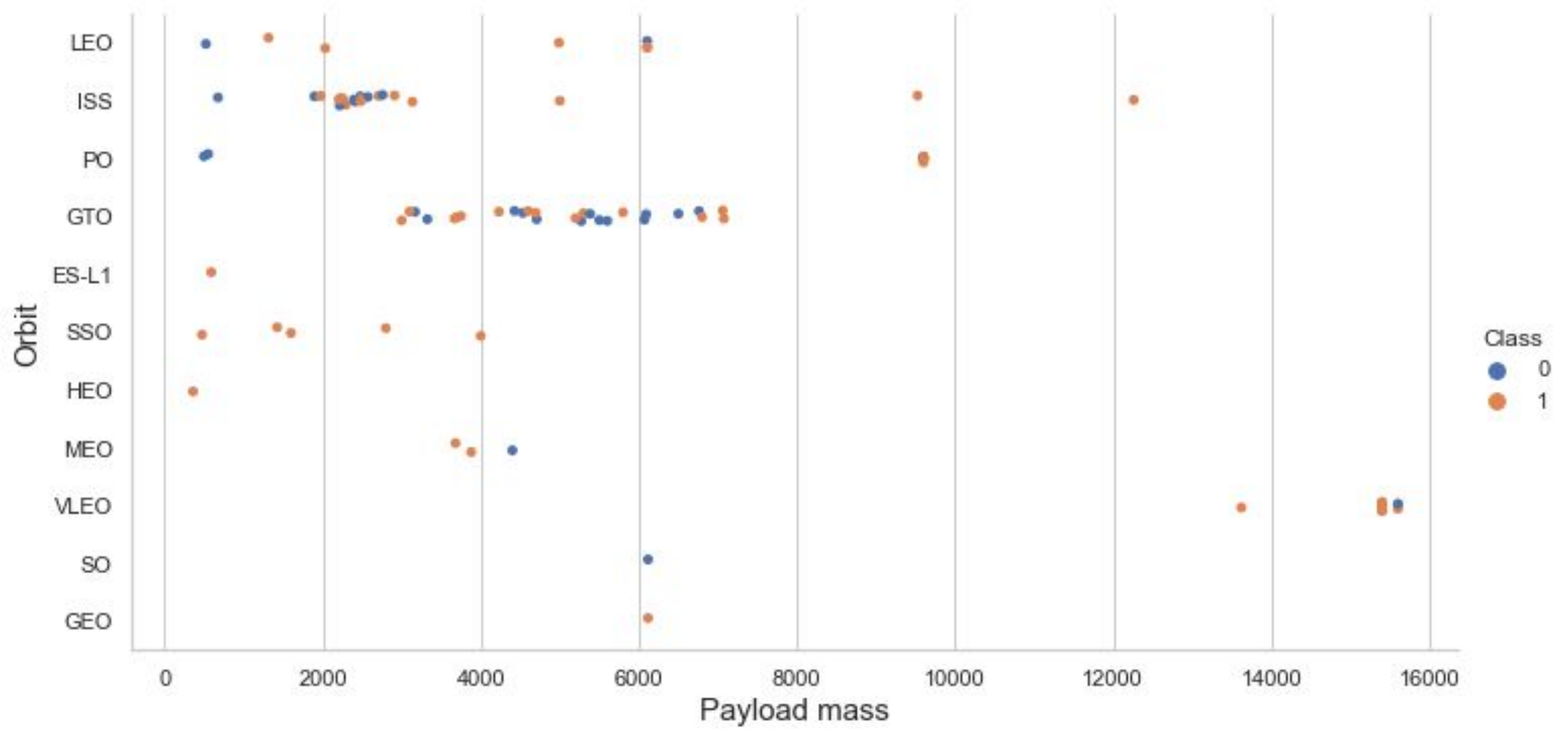
Flight Number vs. Orbit Type

- The success rate increases with the number of flights for each orbit
- This relationship can be seen for the LEO orbit
- The GTO orbit does not follow this trend



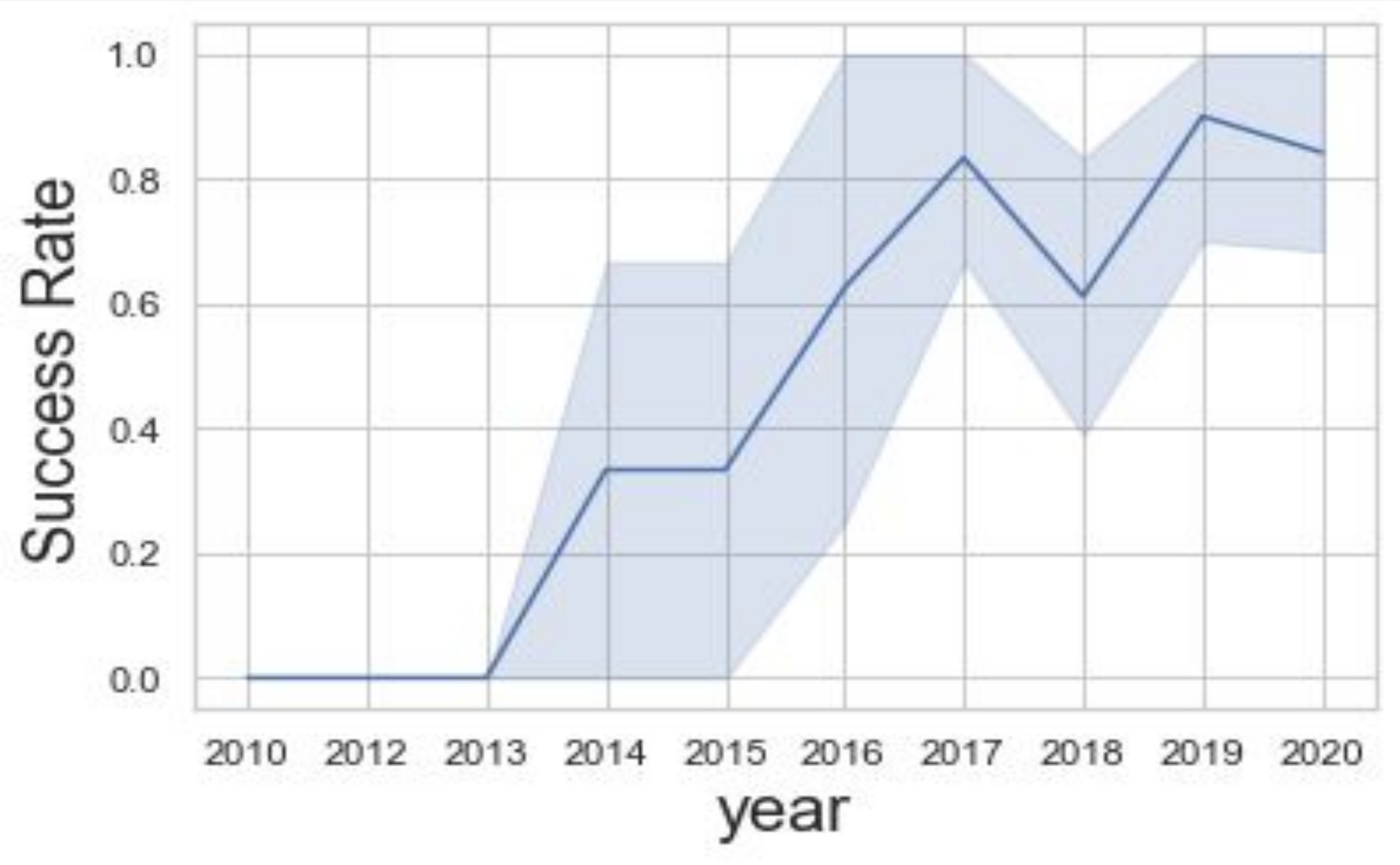
Payload vs. Orbit Type

- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

- The success rate has relatively improved after 2013.



All Launch Site Names

Launch Site Names:

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

```
In [8]: %sql select distinct("LAUNCH_SITE") from SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[8]:

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

Query to display Launch Sites starting with CCA

```
In [9]: %sql SELECT * \
            FROM SPACEXTBL \
            WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total Payload Mass :

- 45,596 kg (total) carried by boosters launched by NASA (CRS)

Average Payload Mass:

- 2,928 kg (average) carried by booster version F9 v1.1

```
In [10]:  %sql SELECT SUM(PAYLOAD_MASS__KG_) \
              FROM SPACEXTBL \
              WHERE CUSTOMER = 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.

Out[10]:
```

SUM(PAYLOAD_MASS__KG_)
45596

```
In [11]:  %sql SELECT AVG(PAYLOAD_MASS__KG_) \
              FROM SPACEXTBL \
              WHERE BOOSTER_VERSION = 'F9 v1.1';

 * sqlite:///my_data1.db
Done.

Out[11]:
```

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First Successful Ground Landing Date was 12/22/2015
- Query was used with the Minimum function

```
In [12]:  %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'

 * sqlite:///my_data1.db
Done.

Out[12]:
min(DATE)
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:

```
In [18]:  %sql select Booster_Version from SPACEXTBL \
          where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 AND 6000;

 * sqlite:///my_data1.db
Done.

Out[18]:
```

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

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes:

In [14]:

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass:

In [15]:
```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ =  (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

Out[15]:

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

2015 Launch Records

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

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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:

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

```
* sqlite:///my_data1.db
Done.
```

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

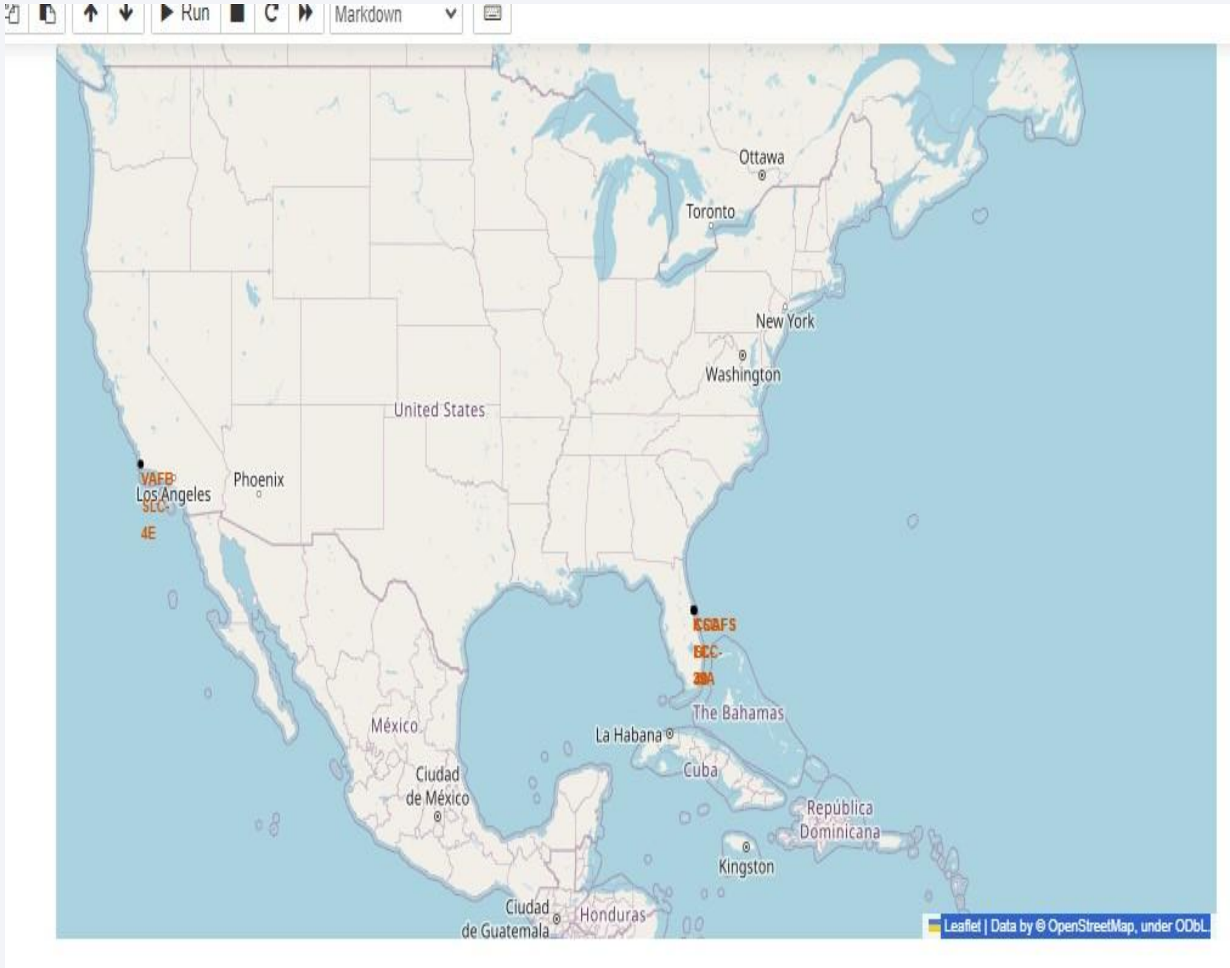
Section 3

Launch Sites Proximities Analysis

Launch Sites on Map

- Launch Sites are near Equator because the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit.

- Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



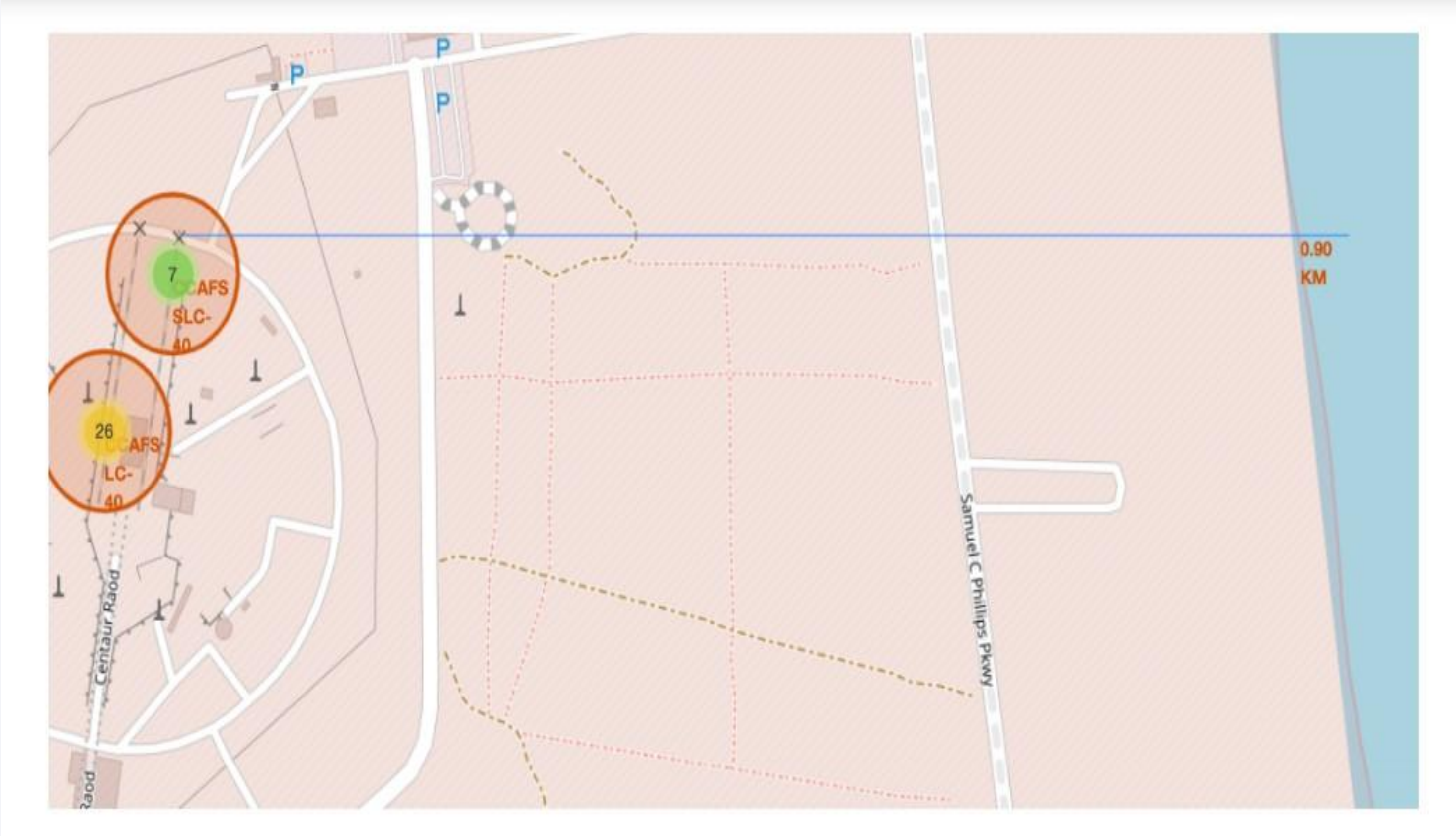
Success/failed launches for each site

- green markers = successful launch
- red marker = failed launch

Distances between a launch site to its proximities

CCAFS SLC-40:

- .86 km from nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway



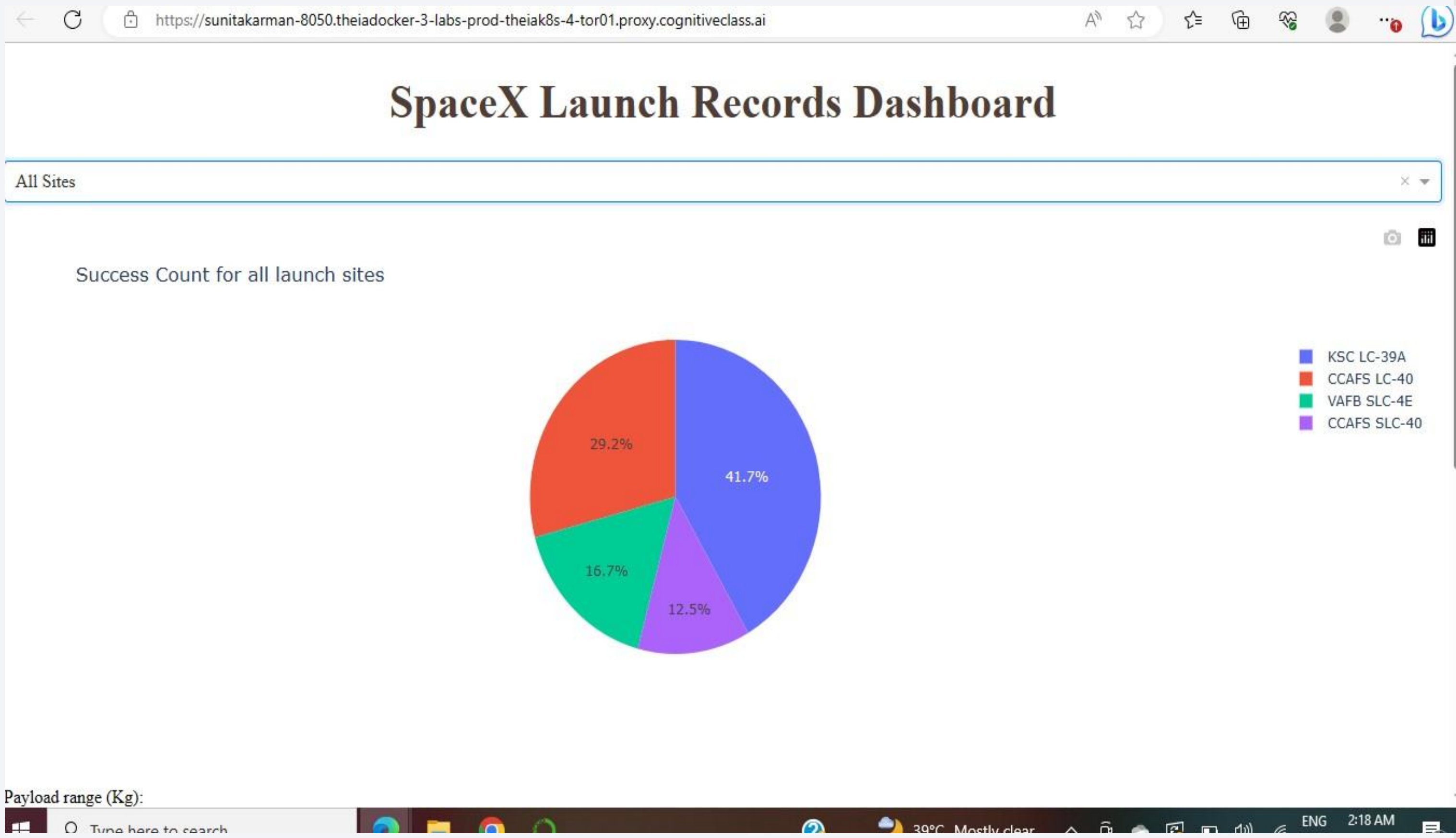
Section 4

Build a Dashboard with Plotly Dash

Success Launch for all Sites

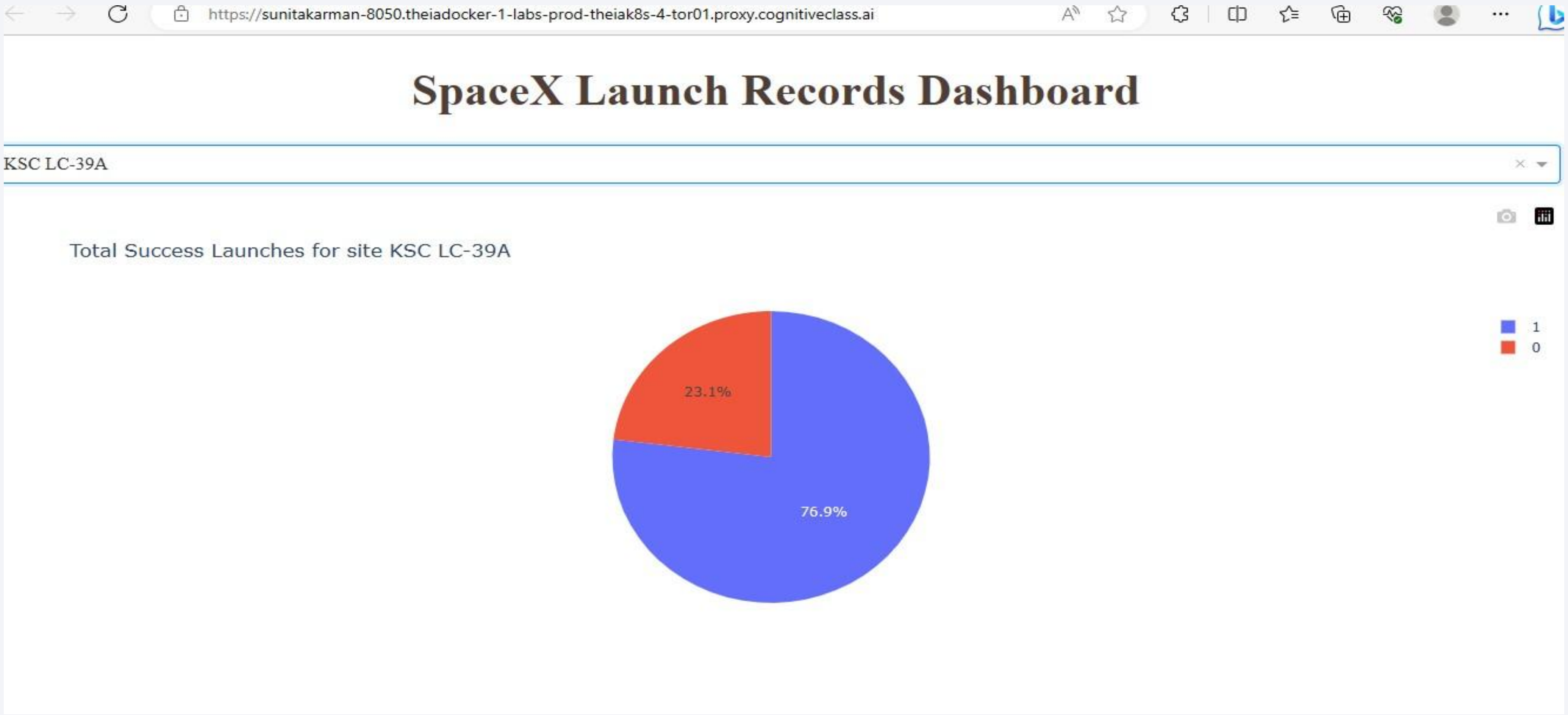
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)

- CCAFS SLC 40 has the least success launches (12.5%)



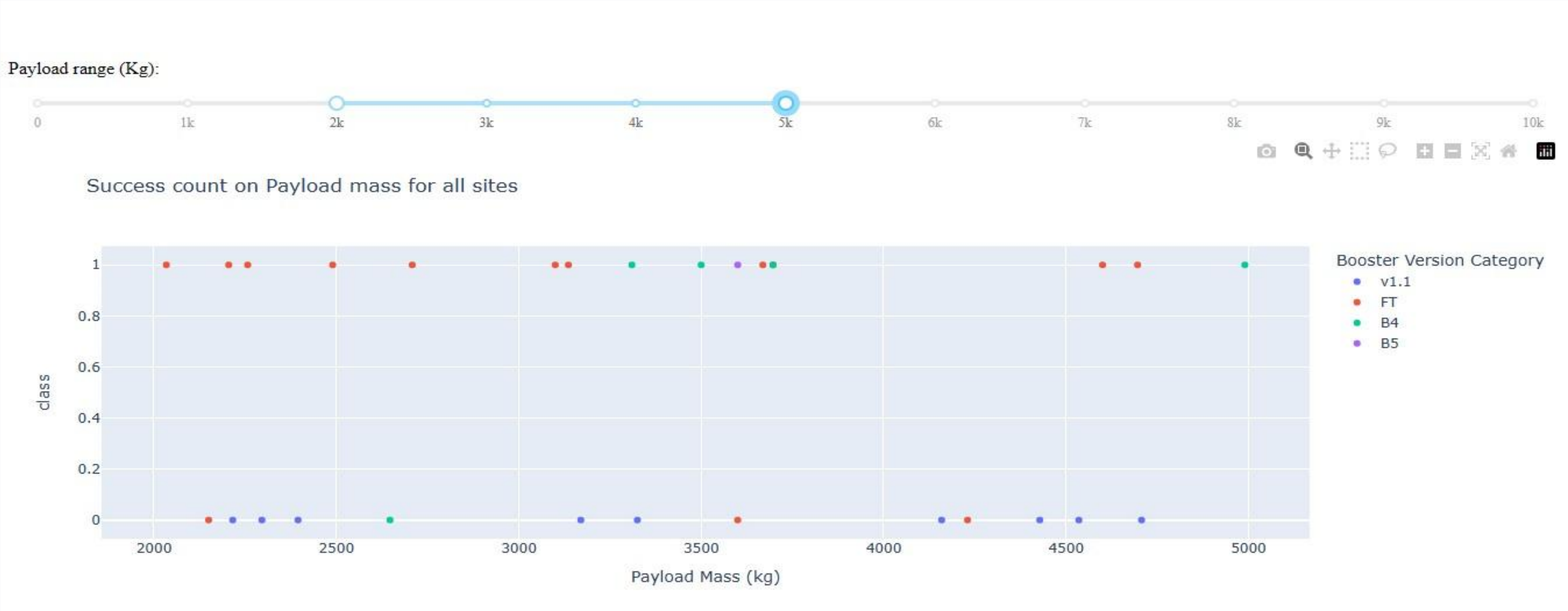
Launch Success (KSC LC-29A)

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches



Payload Mass and Success

Payload Mass between 2000kg and 5000kg have the highest success rate.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

The Decision Tree Classifier has the highest accuracy among all the models which is **88.57%**

	ML Method	Accuracy Score (%)
0	Support Vector Machine	84.821429
1	Logistic Regression	84.642857
2	K Nearest Neighbour	84.821429
3	Decision Tree	88.571429

Confusion Matrix

- This is a confusion matrix of Decision Tree Classifier. It shows that True positive are 3 which is the same for all models.

- While False Negative are 7 which is the least among all models hence implying that type 2 error is the least.

Conclusions

- The **decision tree model** was the best classification model.
- Most of the launch sites are near the **equator** for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- All the launch sites are close to the **coast**
- **KSC LC-39A** has the highest success rate among launch sites. It has a 100% success rate for launches less than 5,500 kg
- Orbits: **ES-L1, GEO, HEO, and SSO** have a 100% success rate
- Payload Mass: Across all launch sites, the **higher the payload mass** (kg), the **higher the success rate**

Appendix - Dataset used for Classification Techniques

```
In [6]: X.head(100)
```

Out[6]:

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	...	Serial_B1058	Serial_B1059	Se
0	1.0	6104.959412	1.0	1.0	0.0	0.0	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	...	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	...	0.0	0.0	
3	4.0	500.000000	1.0	1.0	0.0	0.0	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	...	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	...	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	...	1.0	0.0	
87	88.0	15400.000000	6.0	5.0	5.0	0.0	0.0	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	...	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	...	0.0	0.0	

90 rows × 83 columns

TASK 1

Thank you!