IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Narimane Dahou
Aug 6th, 2023

Outline

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

Executive Summary

- **The following methodologies were used to analyze data:**
  - Data Collection using web scraping and SpaceX API;
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
  - Machine Learning Prediction.
- **Summary of all results**
  - It was possible to collected valuable data from public sources;
  - EDA allowed to identify which features are the best to predict success of launchings;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with Space X.

- **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?
  - Where is the best place to make launches?

Section 1

Methodology

Methodology

- **Data collection methodology:**
    - SpaceX Rest API
    - Web Scraping
- **Performed data wrangling**
    - Filtered the data, dealt with missing values, used 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**
    - Built, tuned, evaluated classification models to ensure the best results

Data Collection

- **Data sets were collected from:**
- Space X API (https://api.spacexdata.com/v4/rockets/)
- Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) using web scraping techniques.

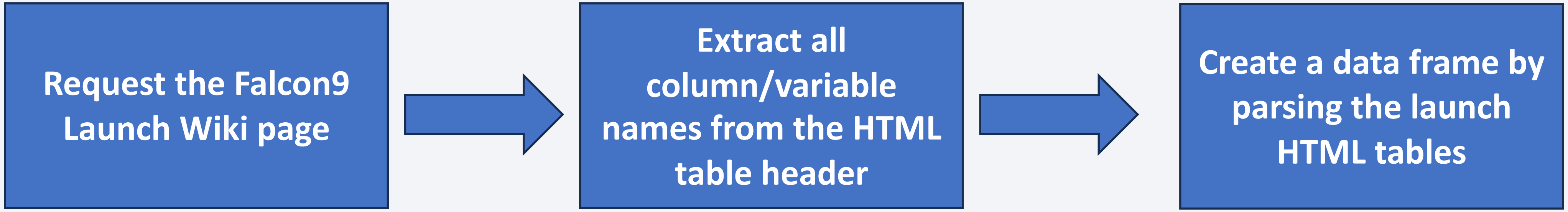
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;

- https://github.com/psychstat/Coursera-Capstone-Assignment/blob/main/1.%20jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

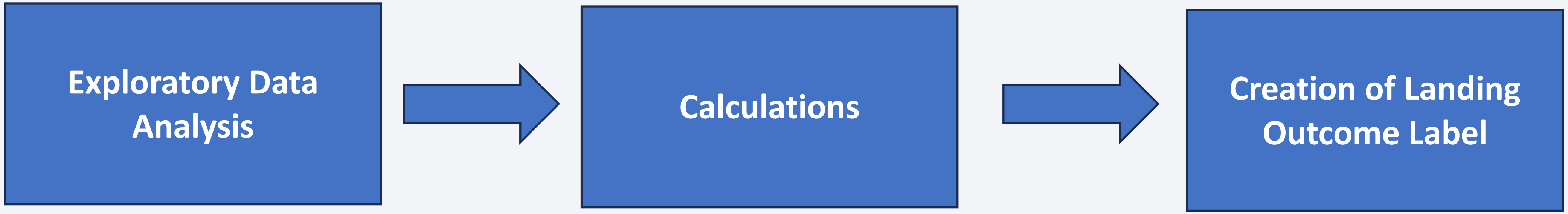
- Data from SpaceX launches are obtained from Wikipedia



- https://github.com/psychstat/Coursera-Capstone-Assignment/blob/main/2.%20jupyter-labs-webscraping.ipynb

Data Wrangling

- Exploratory Data Analysis (EDA) was first performed on the dataset.
- Then launches per site, occurrences of each orbit, and occurrences of mission outcome per orbit type were calculated.



- https://github.com/psychstat/Coursera-Capstone-Assignment/blob/main/3.%20IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- **To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:**

  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

- https://github.com/psychstat/Coursera-Capstone-Assignment/blob/main/5.%20IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- **The following SQL queries were performed:**
  - Displayed names of unique launch sites in the space mission;
  - Displayed 5 records where launch sites began with the string “CCA”
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed the average payload mass carried by booster version F9 v1.1
  - Listed the date of the first successful landing outcome
  - Displayed the names of the boosters which have success in the drop ship and have payload mass between 4000 and 6000kg
  - Listed the total number of successful and failed mission outcome
  - Displayed the names of the booster versions which have carried the maximum payload mass.
  - Listed the months, failed landing outcomes in drone ship, booster versions, and launch site names of year 2015.
  - Ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20

- https://github.com/psychstat/Coursera-Capstone-Assignment/blob/main/4.%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Markers, circles, lines and marker clusters were used with Folium Maps**
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates.

- https://github.com/psychstat/Coursera-Capstone-Assignment/blob/main/6.%20IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

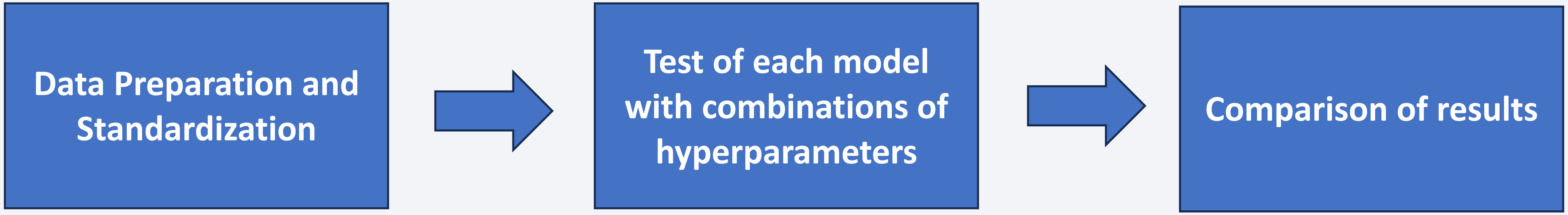
Build a Dashboard with Plotly Dash

- **The following graphs and plots were used to visualize data:**
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- https://github.com/psychstat/Coursera-Capstone-Assignment/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



- https://github.com/psychstat/Coursera-Capstone-Assignment/blob/main/7.%20IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

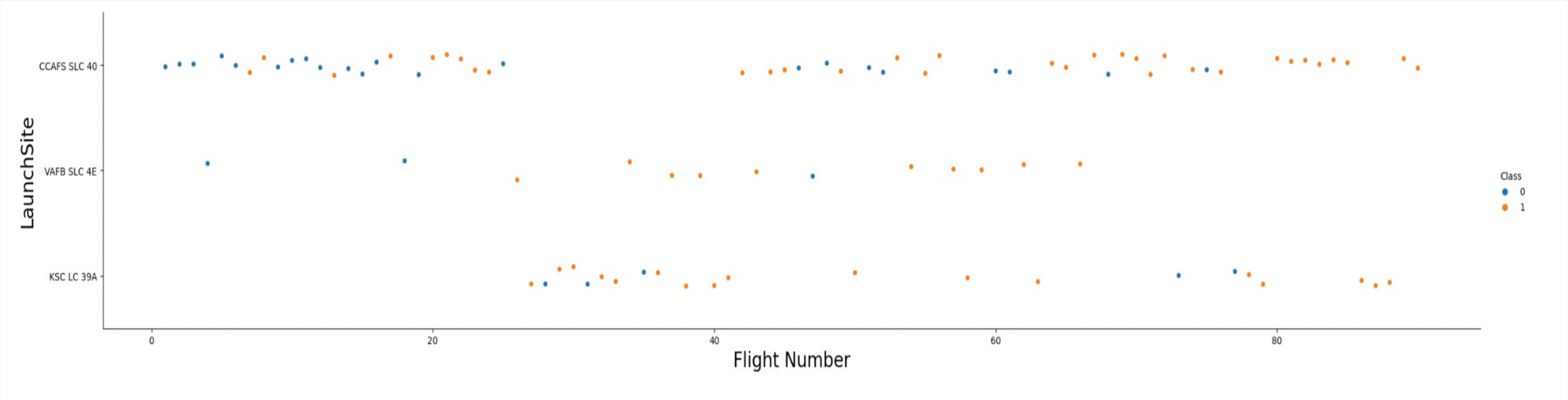
Results

- Exploratory data analysis results:
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

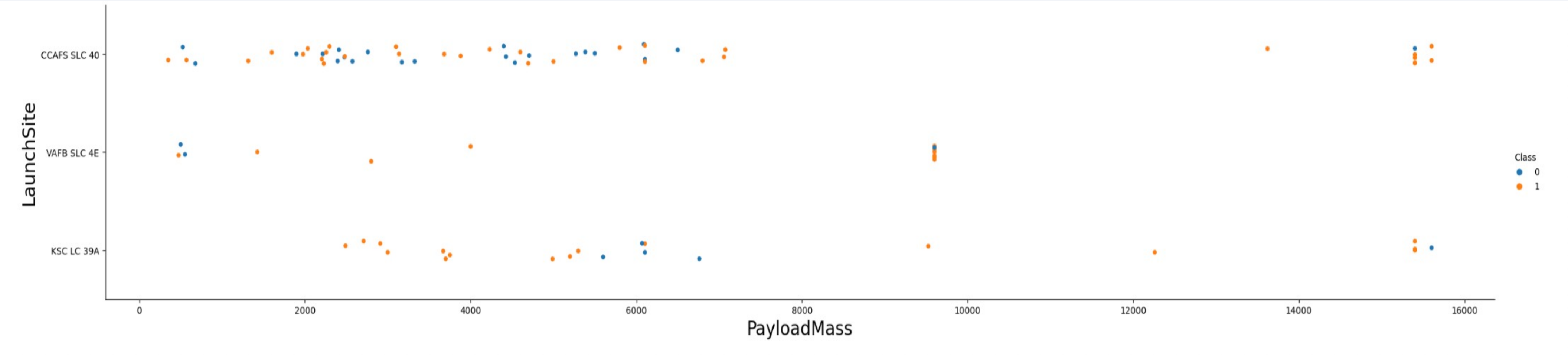
Insights drawn from EDA

Flight Number vs. Launch Site



- According to this scatterplot, the general success rates increased over time.

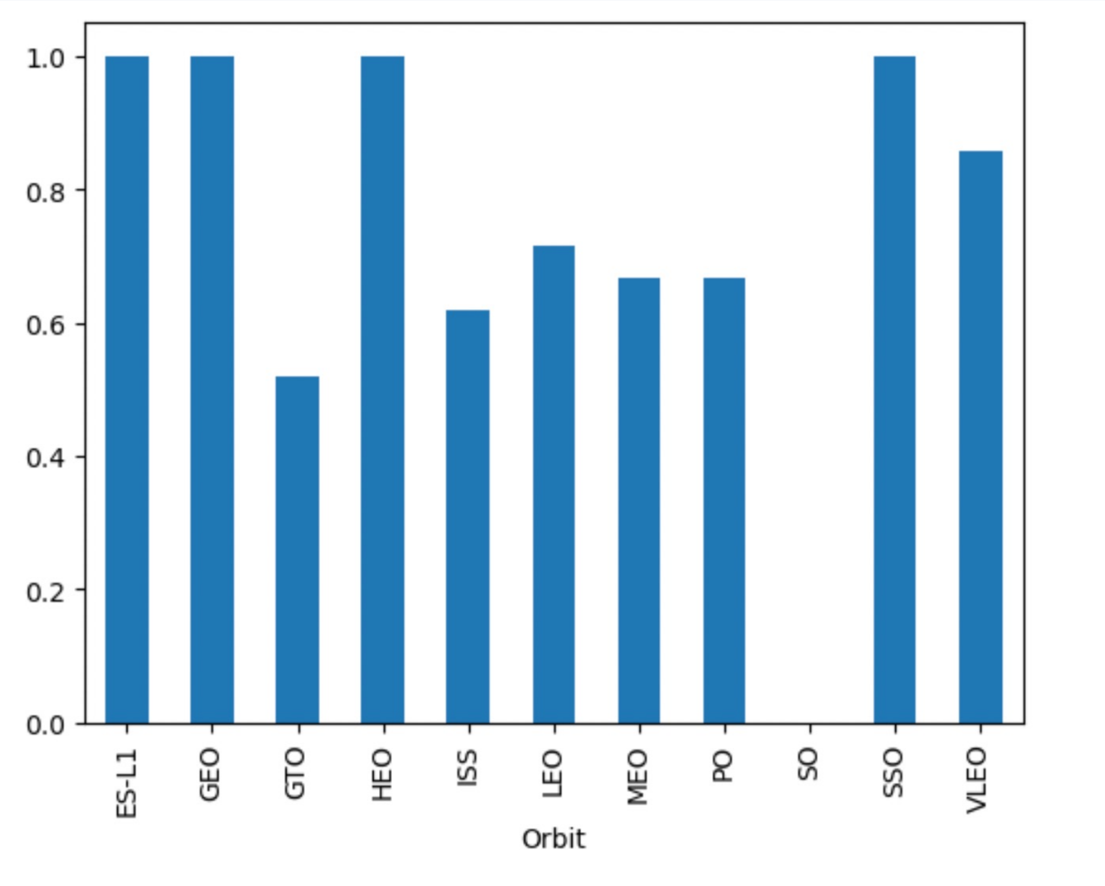
Payload vs. Launch Site



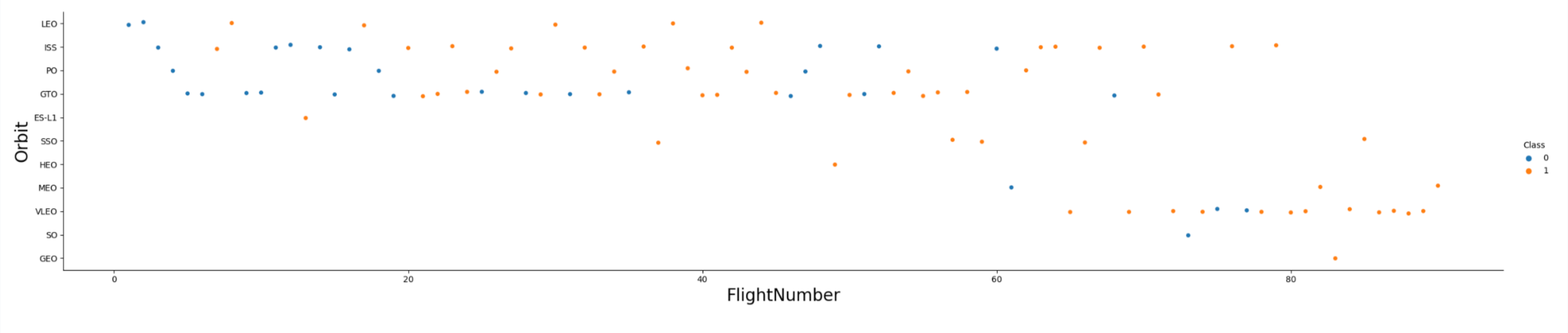
- Payloads over 9000kg have the best success rates for all Launch sites.
- Launch site CCAFS SLC 40 has awful success rates for payloads less than 7000kg.
- Launch site KSC 39A does not do very well with payloads around 6000kg.

Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, SSO, and VLEO have the best success rates, overall.
- Orbits GTO, ISS, LEO, MEO, and PO have medium success rates.
- Orbit SO has the worst success rate at 0%.

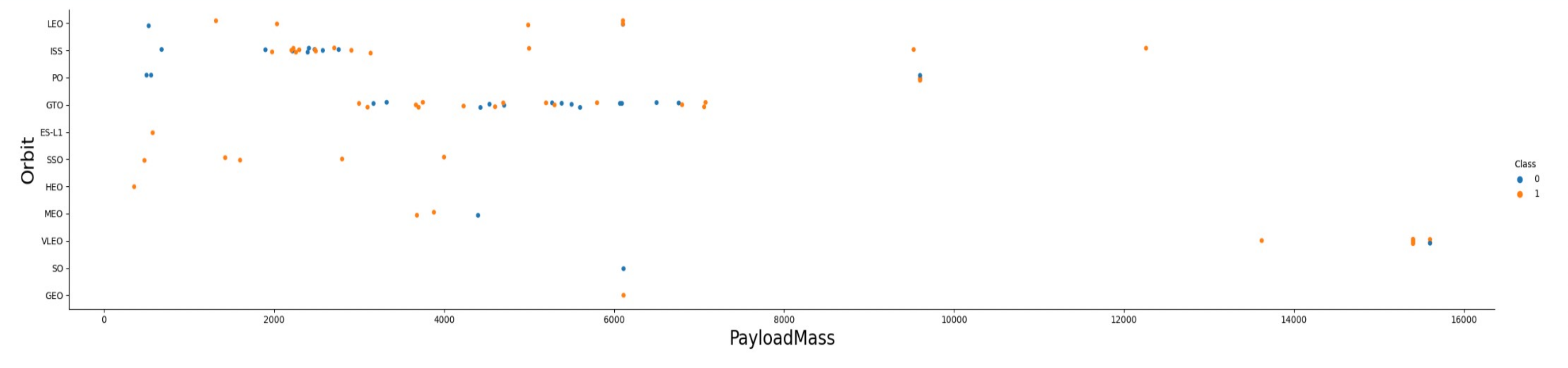


Flight Number vs. Orbit Type



- 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.

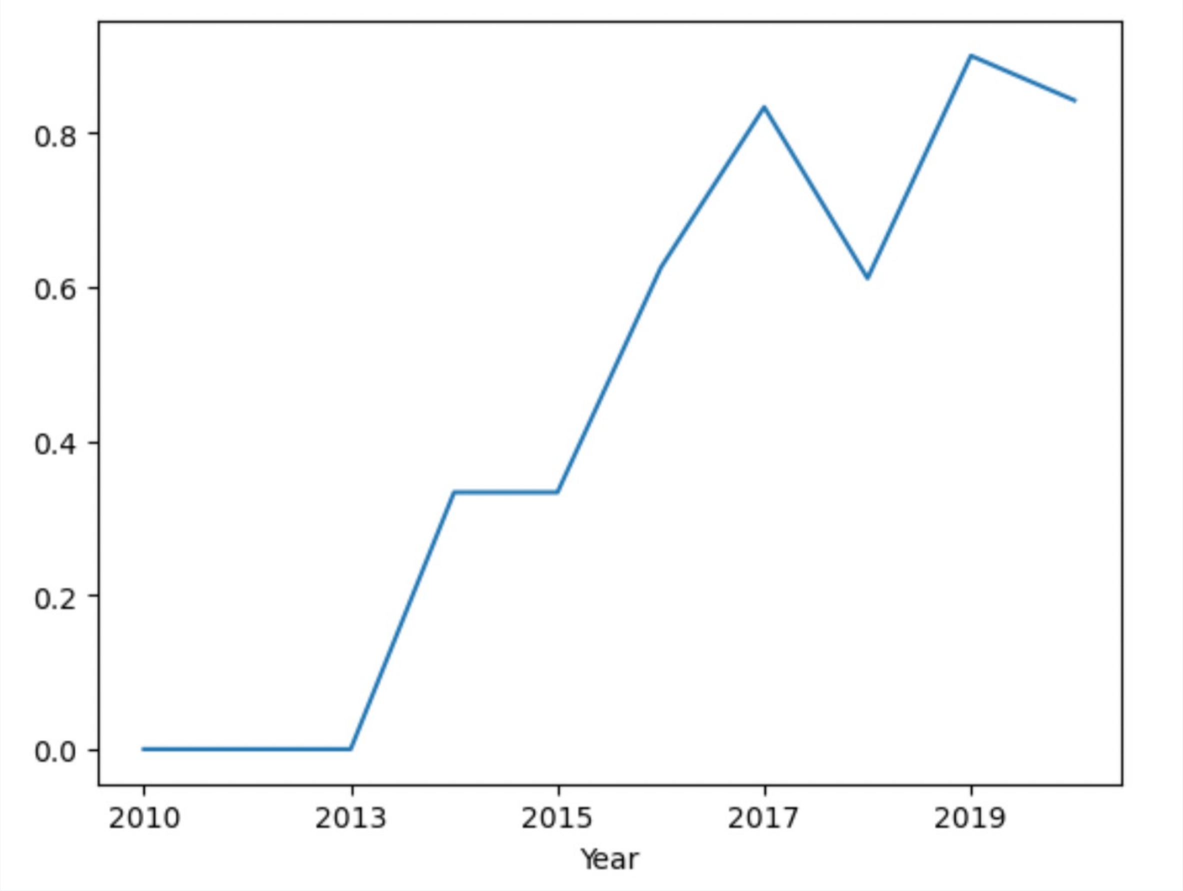
Payload vs. Orbit Type



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

Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

- Displaying the names of the unique launch sites.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (
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 (
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

- Displaying the 5 records where launch sites begin with `CCA`

Total Payload Mass

```
[11]:  %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';

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

[11]:

TOTAL_PAYLOAD
111268

- Displaying the total payload carried by boosters from NASA.

Average Payload Mass by F9 v1.1

```
[12]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

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

[12]:

AVG_PAYLOAD
2928.4

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

First Successful Ground Landing Date

```
[13]: %sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';

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

[13]:

FIRST_SUCCESS_GP
2015-12-22

- Displaying the date of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUT
```

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

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

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

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

- Displaying the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

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

2015 Launch Records

```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (
```

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

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

- List of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

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

```
%sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROU
```

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

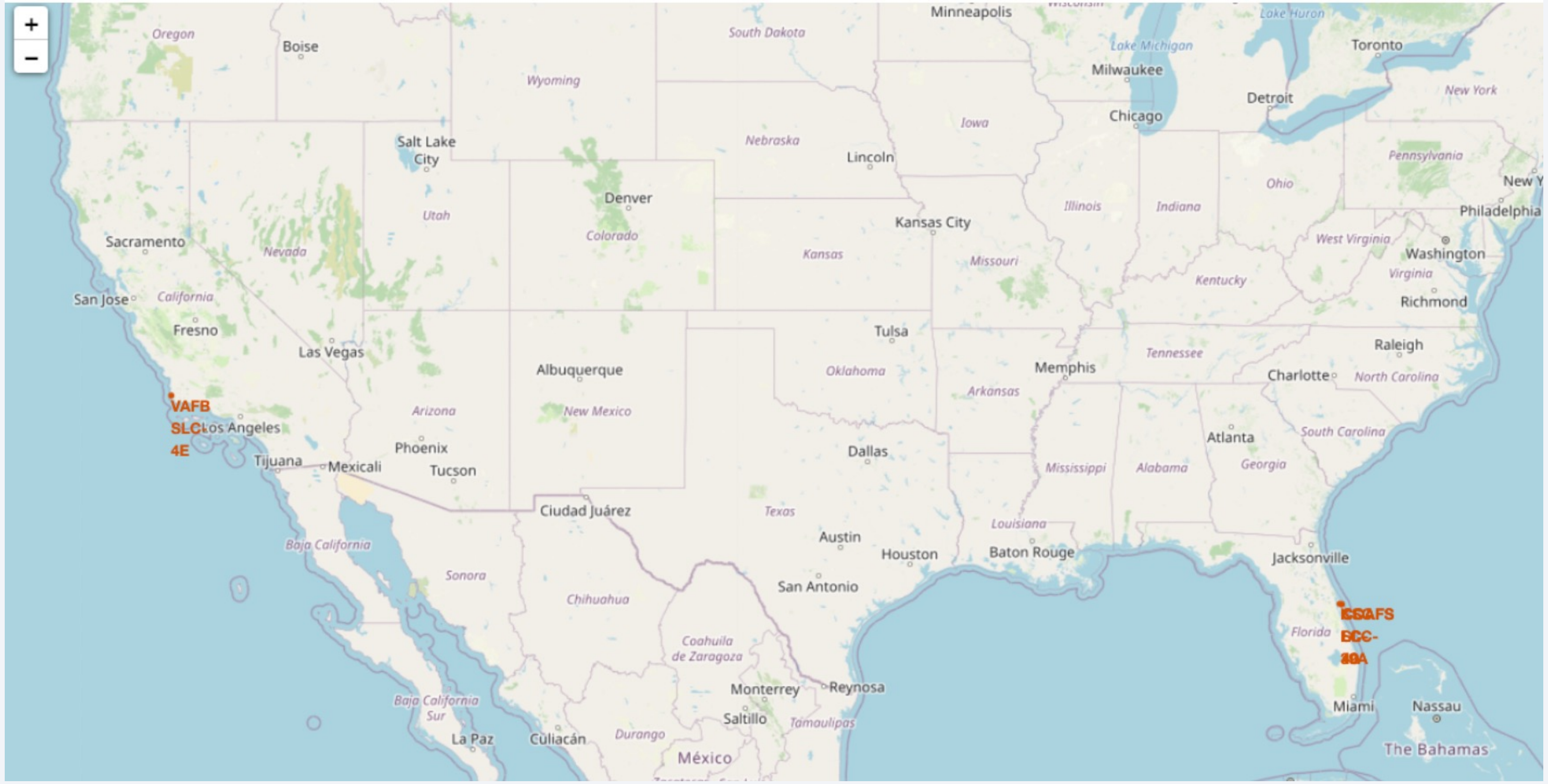
Landing_Outcome	QTY
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- The ranking 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

Section 3

Launch Sites Proximities Analysis

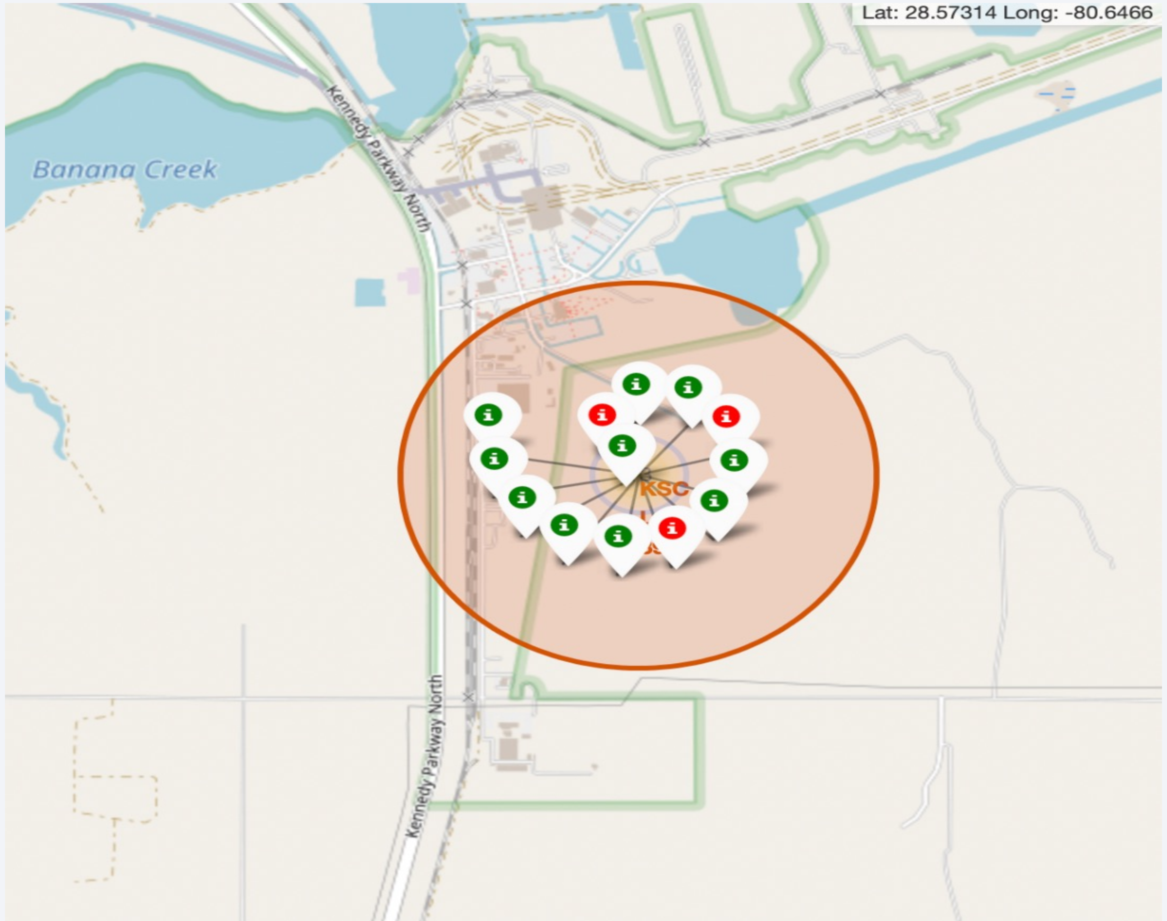
All Launch Sites



- Launch sites’ location markers on a global map. All launch sties are in very close proximity to the coast.

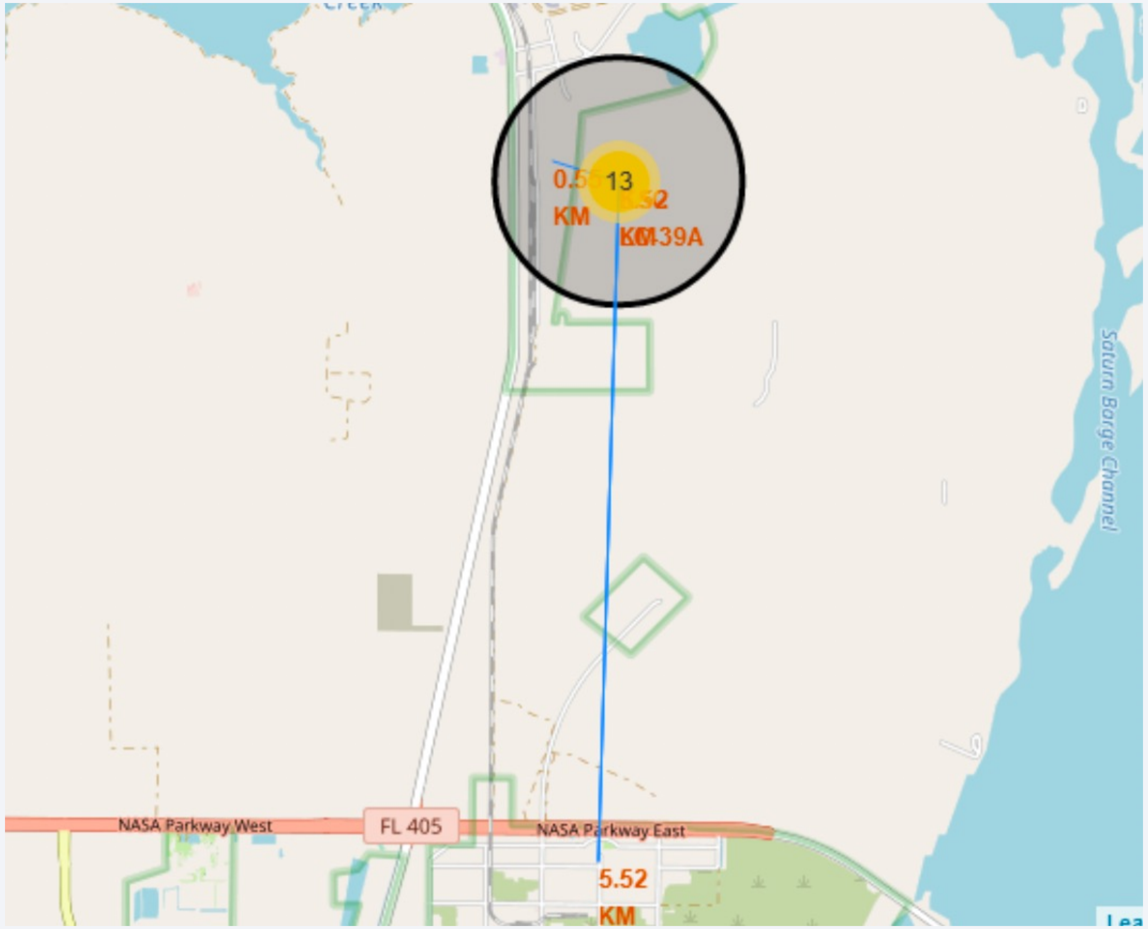
Launch Outcomes by Site

- Green markers indicate successful launches while red markers indicate failed launches.



Distance of Launch Site KSC LC-39A from its Proximities

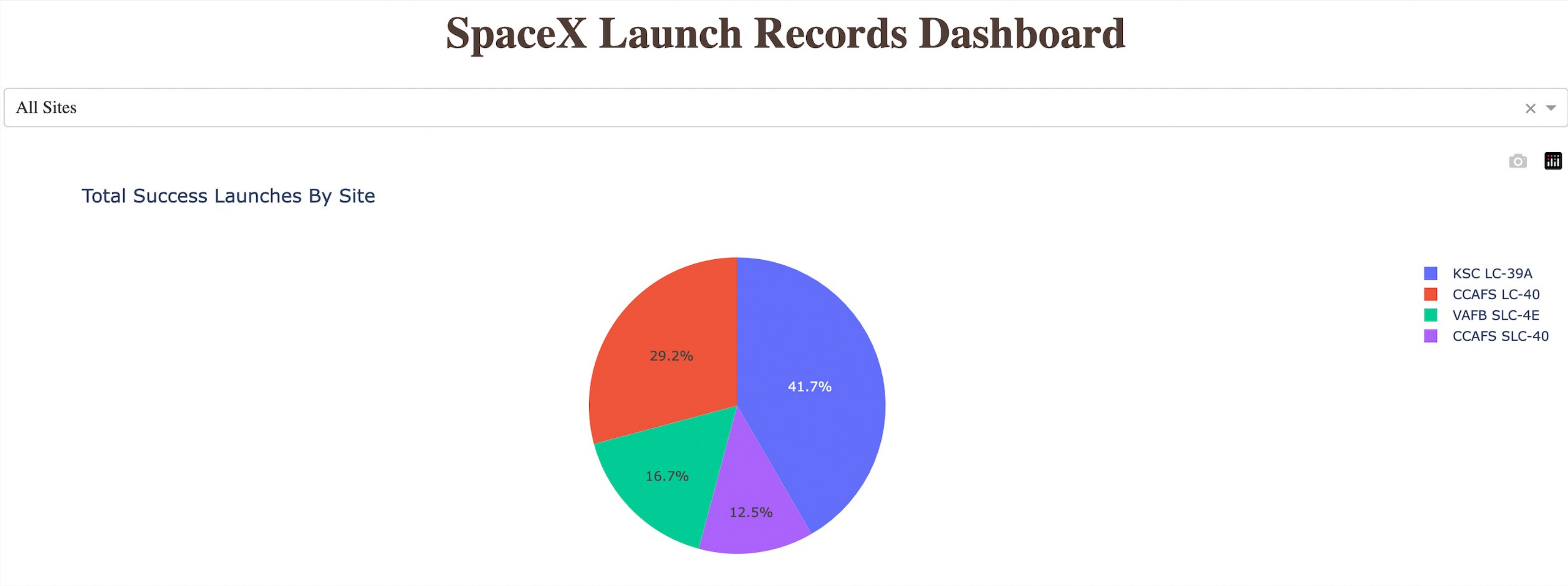
- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.



Section 4

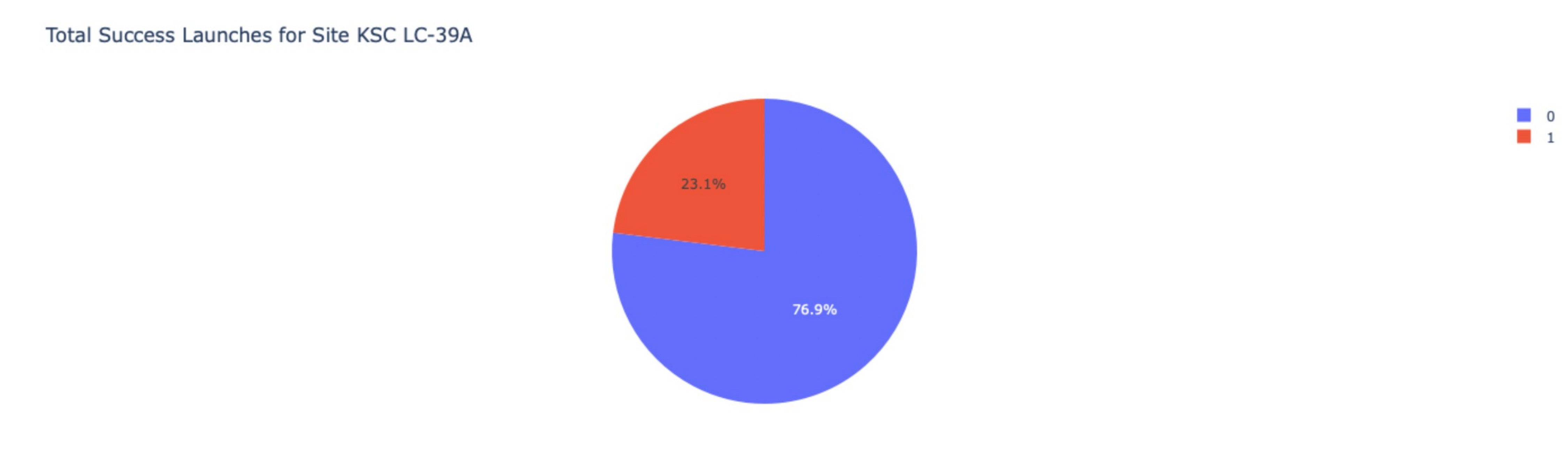
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- From the chart, we can see that launch site KSC LC-39A (blue) has the most successful launches.

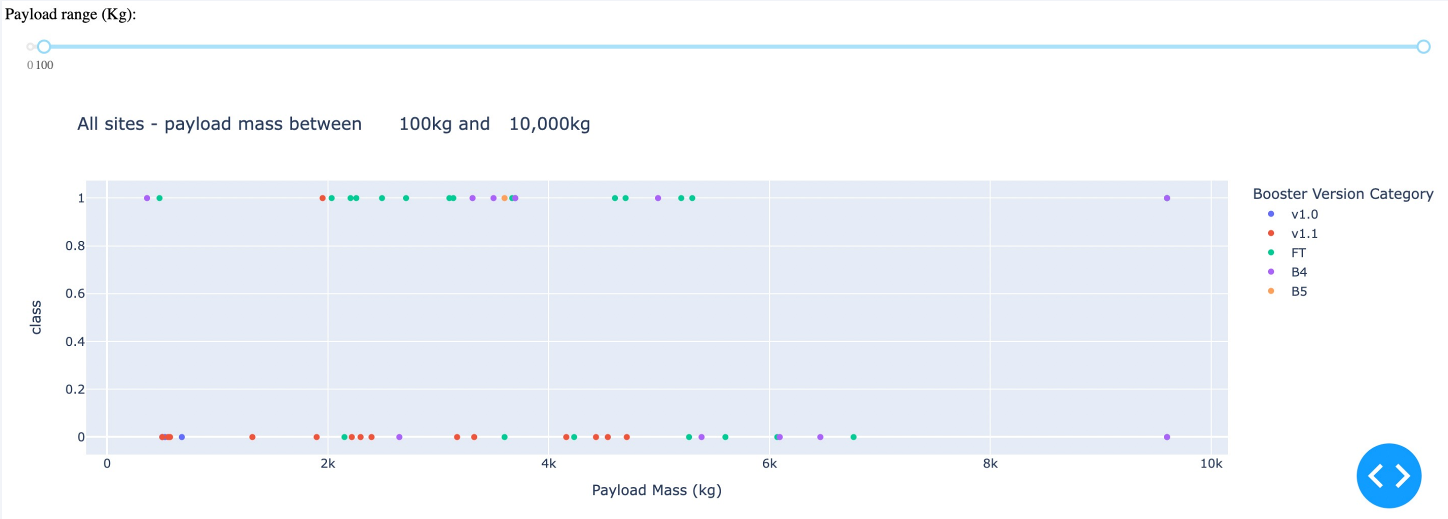
Total Success Lauches for Site KSC LC-39A

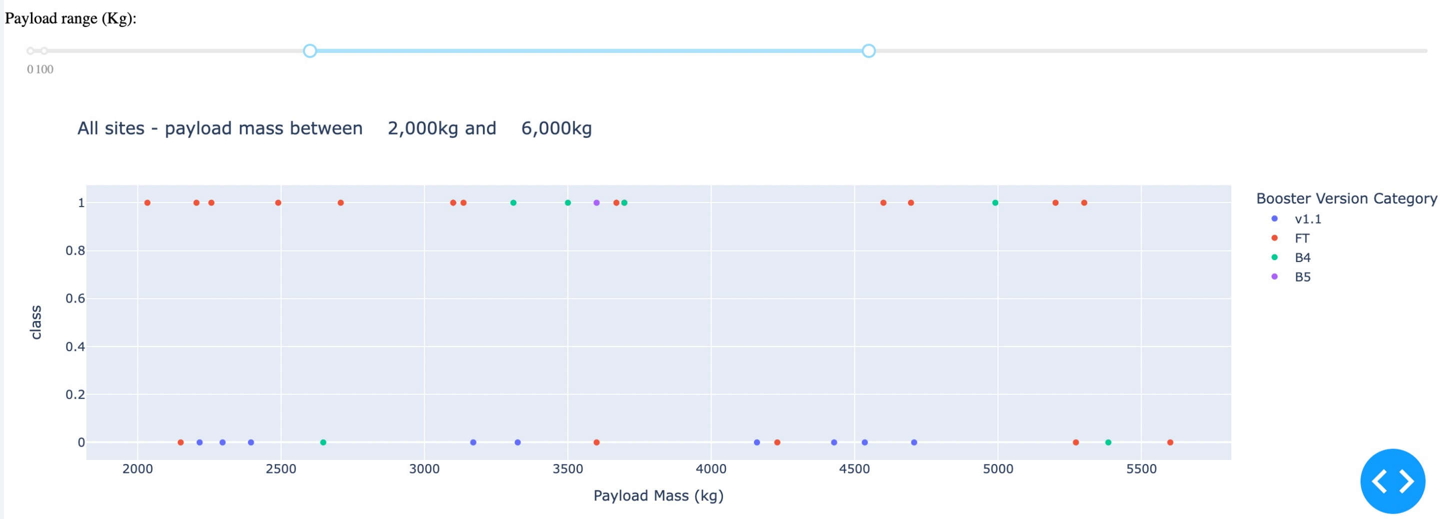


- 76.9% of launches from site KSC LC-39A were successful.

Payload vs. Launch Outcome for all Sites

- According to the chart, payloads between 1900kg and 5500kg have the highest success rate.





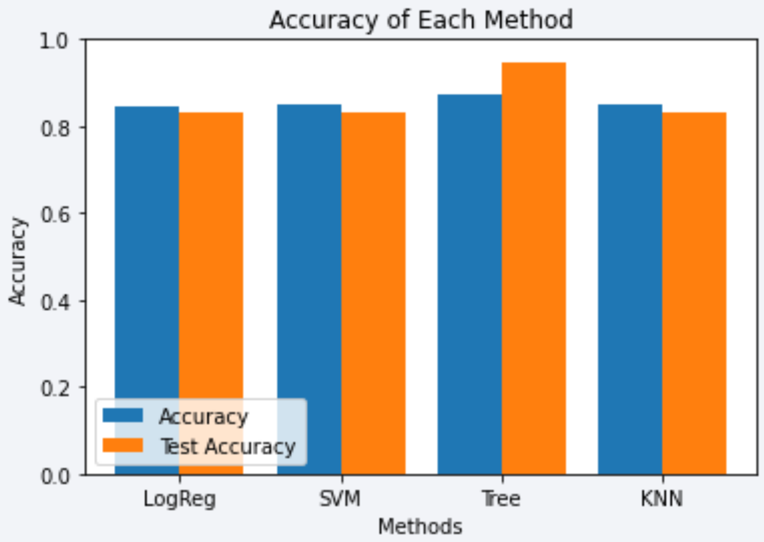
Section 5

Predictive Analysis (Classification)

Classification Accuracy

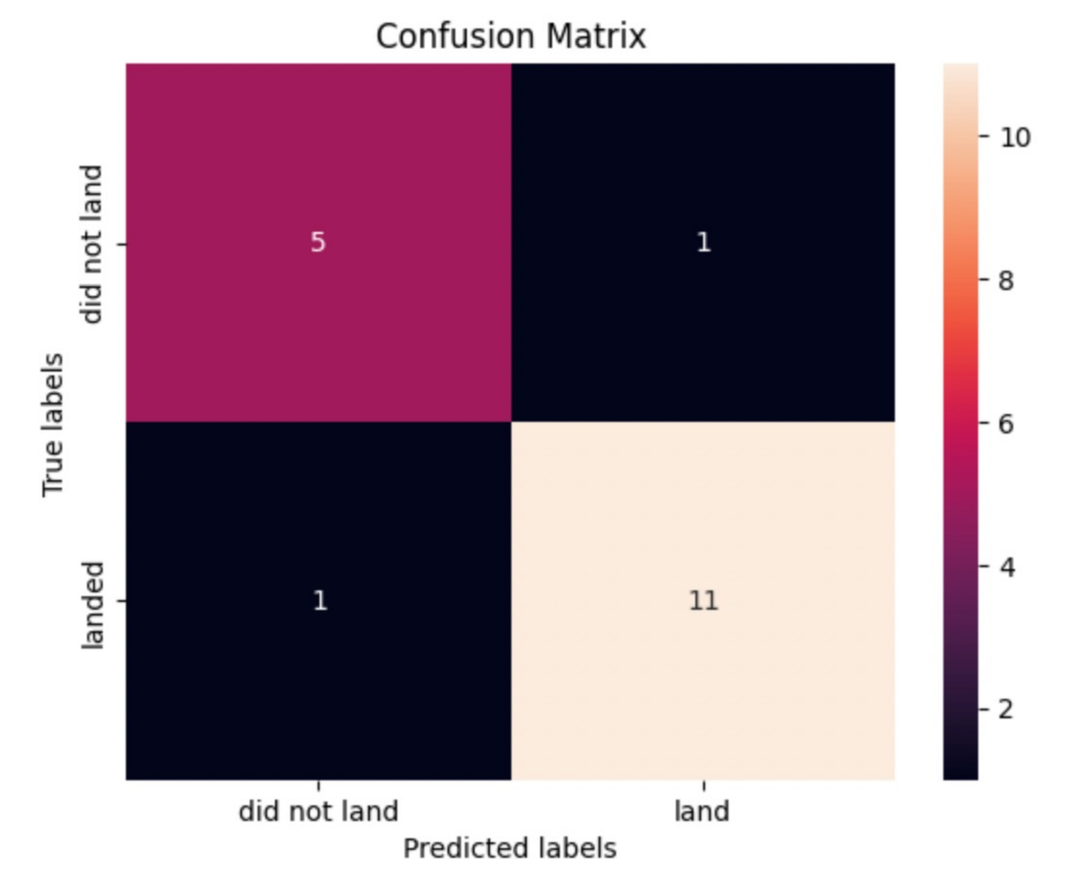
Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.88889
KNN	0.84821	0.83333

- The Decision Tree Classifier has the highest accuracy at 87.5%



Confusion Matrix of Decision Tree Classifier

- The Confusion Matrix demonstrates that the model can correctly predict between launches that landed and launches that did not land.



Conclusions

- Launch site KSC LC-39A has the best success rate of all launch sites.
- Most of the launch sites are near the equator and the coastline.
- Launches success rates have increased over the years.
- Orbits ES-L1, GEO, HEO, SSO, and VLEO have the best success rates, overall.
- Payloads over 9000kg have the best success rates for all Launch sites.
- The Decision Tree Classifier is the best algorithm to use on this dataset.

Appendix

- None for this presentation.

Thank you!