IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SIDDARTH S
26/11/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using API and Web Scraping
  - Data wrangling
  - Exploratory Data Analysis (EDA)
  - Interactive Dashboard
  - Machine learning models
- Summary of all results
  - Results from EDA
  - Plotly Dashboard
  - Predictive Analysis

Introduction

SpaceX's revolutionary idea to reuse the first stage of the rocket in order to reduce the overall cost of the rocket launch has been successful. But there are various factors affecting the successful landing of the rocket.

- ❑Will the Falcon 9 successfully land?
- ❑What factors affect the successful landing?
- ❑Has the landing success increased over time?
- ❑Best model that can predict the success of the landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web Scraping from Wikipedia and Data using SpaceX API
- Perform data wrangling
  - Dealing with missing values and using One Hot Encoding to prepare the data for 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
  - Logistics Regression, Support Vector Machine (SVM), Decision Tree, K Nearest Neighbors (KNN)

Data Collection

- Data collection involved API requests from SpaceX REST API
- Web scraping table from SpaceX Wikipedia page

<u>Data from SpaceX REST API</u>

FlightNumber, PayloadMass, Date, BoosterVersion, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, Block, Longitude, Latitude, ReusedCount, Serial

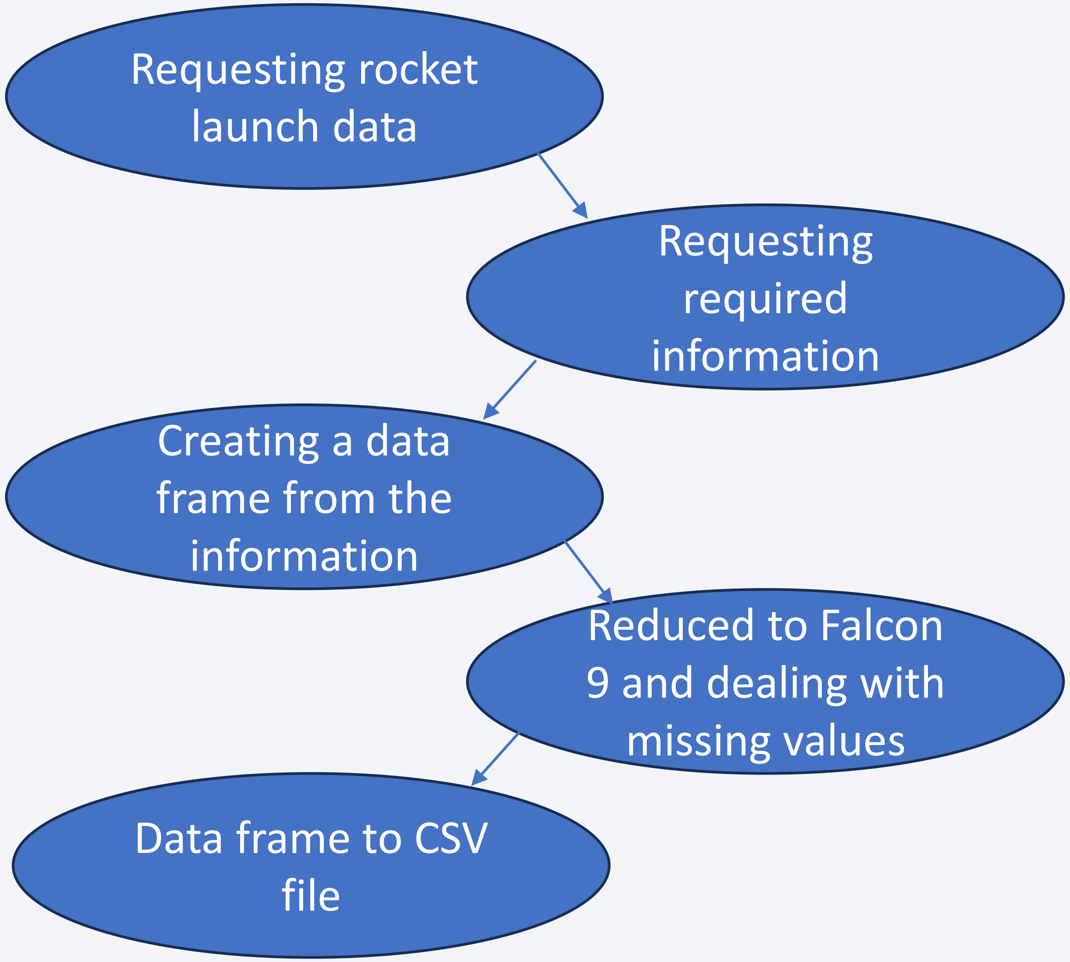
<u>SpaceX Wikipedia page</u>

Flight No., Launch site, Payload, Payloadmass, Orbit, Launch outcome, Customer, Booster landing, Version Booster, Date, Time

Data Collection – SpaceX API

- Request data from SpaceX REST
- Clean the data
- Export data to CSV file
- GitHub URL

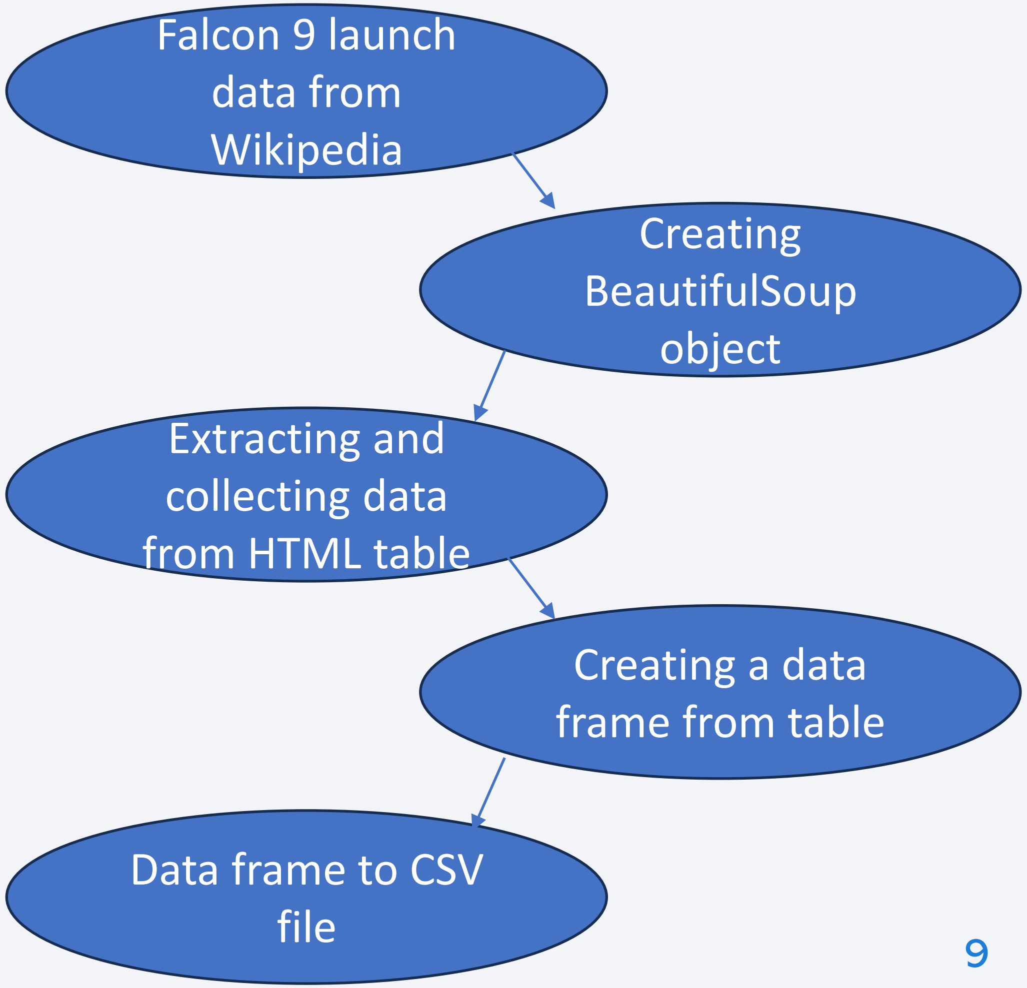
https://github.com/SidduS-code/IBM-Datascience/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

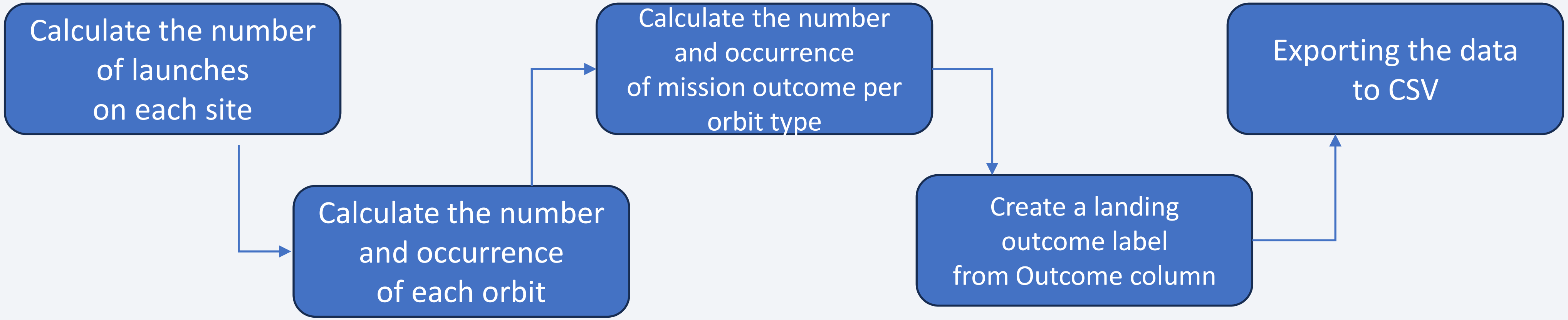
- Falcon 9 launch data from the Wikipedia URL
- Extract Falcon 9 launch records HTML table
- Convert the table to dataframe
- Export data to CSV file
- GitHub URL

https://github.com/SidduS-code/IBM-Datascience/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Exploratory data analysis was performed to find patterns in the data and determine the label for supervised models.



- GitHub URL

https://github.com/SidduS-code/IBM-Datascience/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts plotted
  - Scatter plot
  - Bar chart
  - Line chart
- Scatter plots show the relationship between variables. If a relationship exists, could be used in machine learning model.
- GitHub URL

https://github.com/SidduS-code/IBM-Datascience/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

SQL Queries

- 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 ships 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 records that will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in the year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order.

GitHub URL

https://github.com/SidduS-code/IBM-Datascience/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers of all Launch Sites
- Coloured Markers of the launch outcomes for each Launch Site
- Distances between a Launch Site to its proximities

GitHub URL

https://github.com/SidduS-code/IBM-Datascience/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

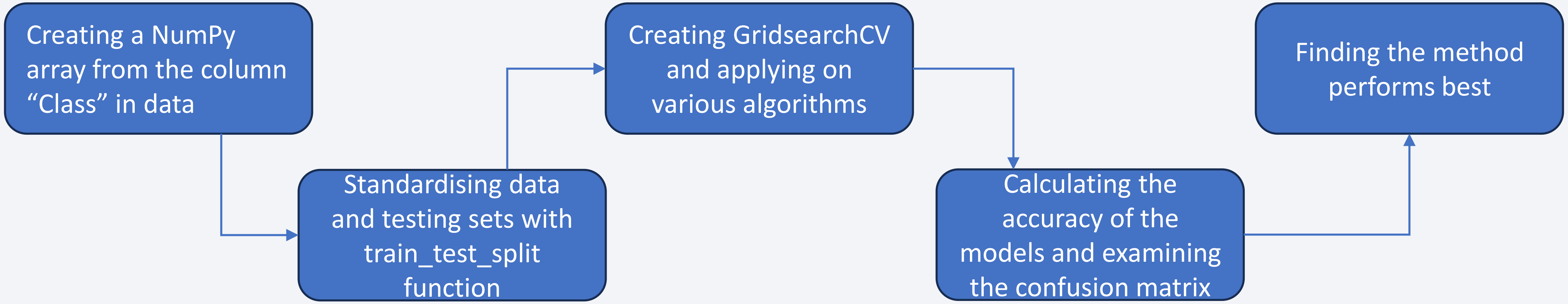
- Launch Sites Dropdown List
- Pie Chart showing Success Launches
- Slider of Payload Mass Range
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions

GitHub URL

https://github.com/SidduS-code/IBM-Datascience/blob/main/Dash%20app.ipynb

Predictive Analysis (Classification)

- The accuracy of 4 methods are calculated:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Classification Trees
    - K Nearest Neighbors (KNN)



GitHub URL

https://github.com/SidduS-code/IBM-Datascience/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

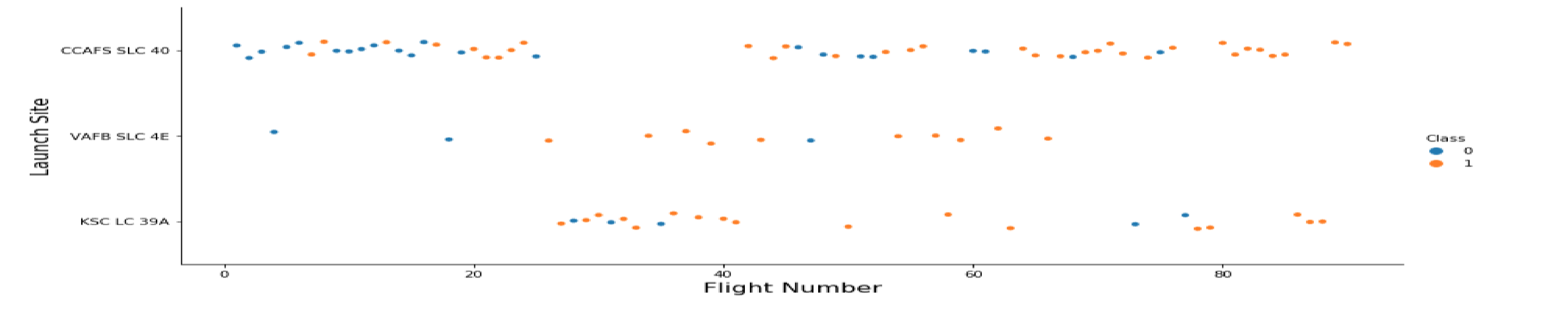
Results

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

Section 2

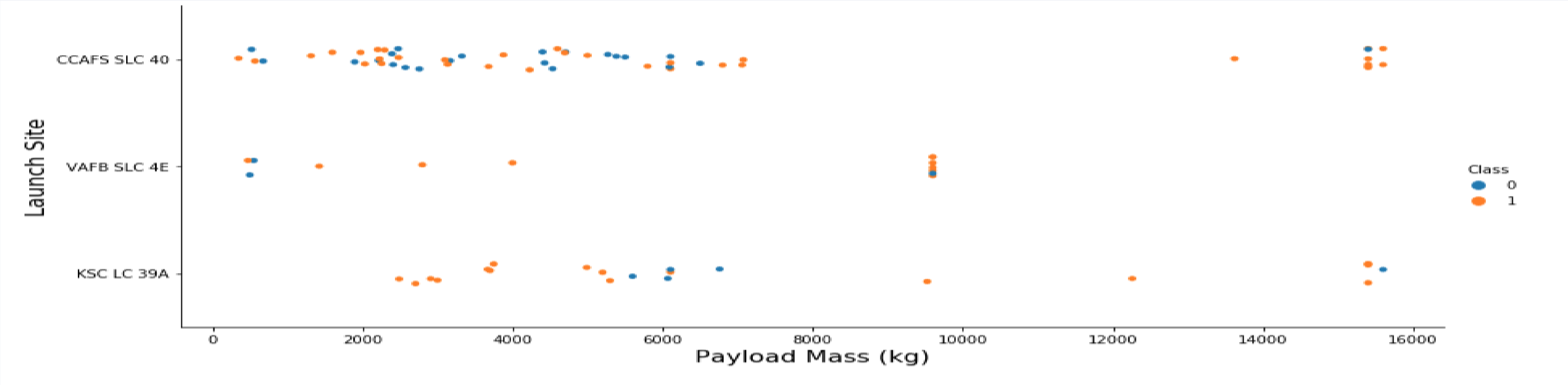
Insights drawn from EDA

Flight Number vs. Launch Site



- The earliest flights all failed to perform while the latest flights all succeeded
- The CCAFS SLC 40 launch site has about 50% of the launches
- Charts shows that each new launch has a higher rate of success

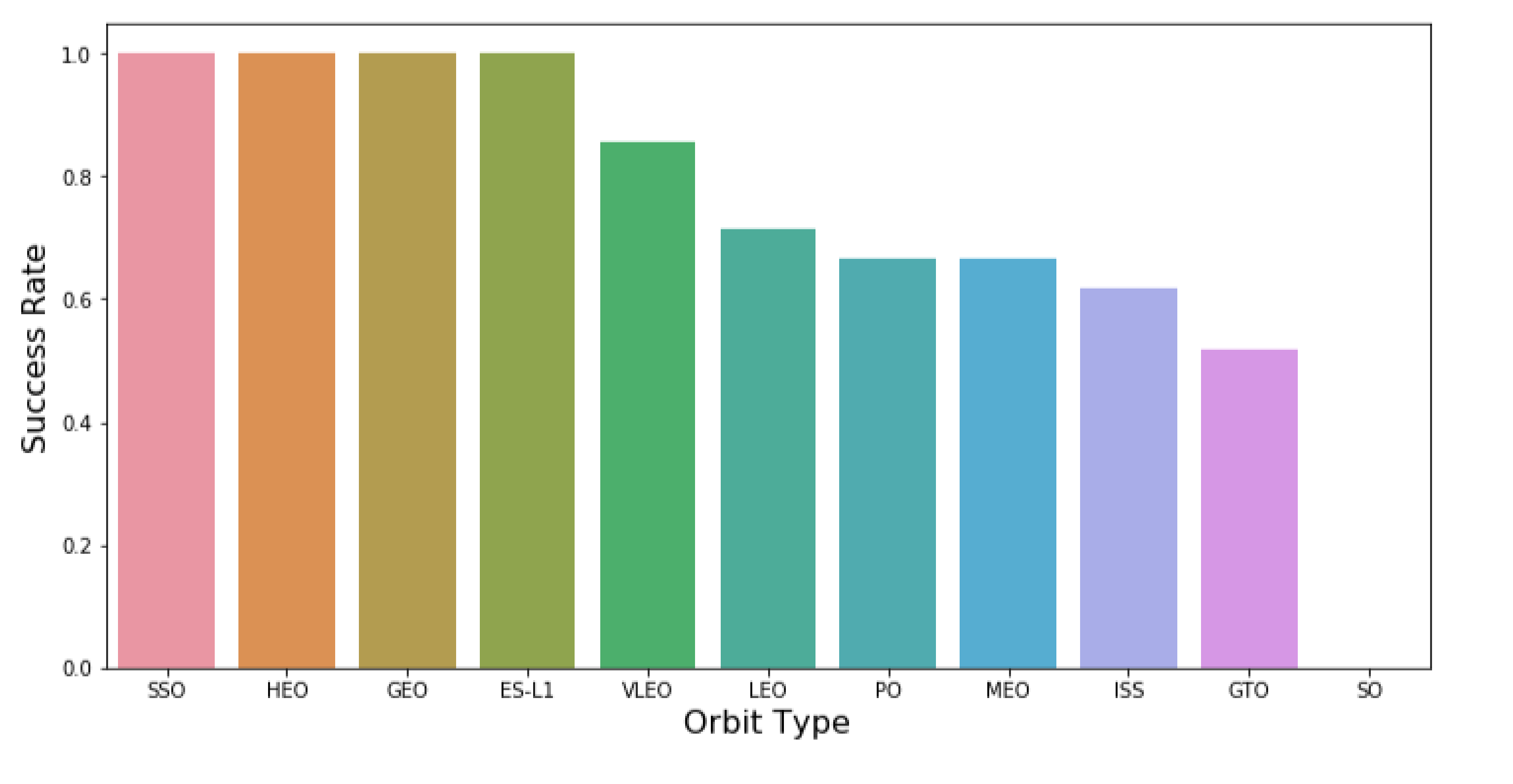
Payload vs. Launch Site



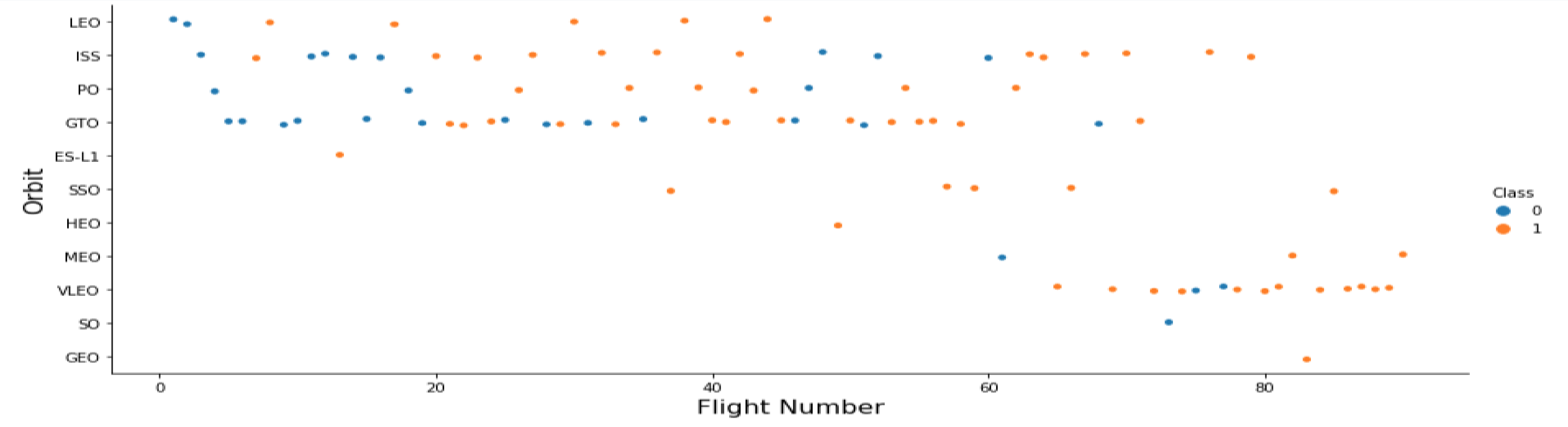
- The success rate is high for payload mass over 9000kg
- All the lauchsites holds most launches at lower payload mass
- KSC LC 39A has mostly failed between payload mass of 5500 to 7000 and performed exceptionally well in any other payload mass
- In-depth study required on VAFB SCL 4E as it has failed to perform only thrice

Success Rate vs. Orbit Type

- SSO, HEO, GEO and ES-L1 has 100% success rate
- Whereas SO holds 0%
- VLEO has good record of 85% success rate
- LEO, PO, MEO, ISS and GTO requires more study on higher failure rate

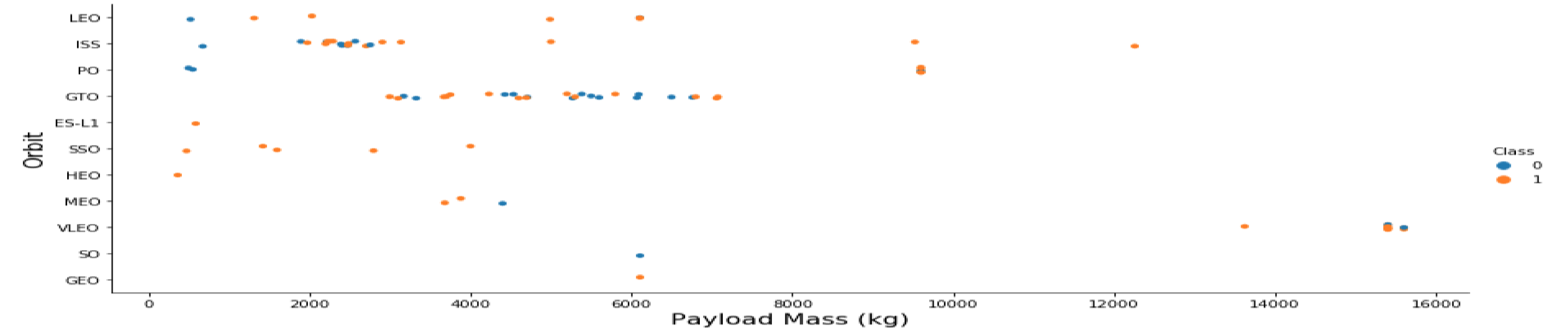


Flight Number vs. Orbit Type



- Plots shows the success and failure of the flight number against the Orbit type
- ISS has the most flight numbers varying in higher range

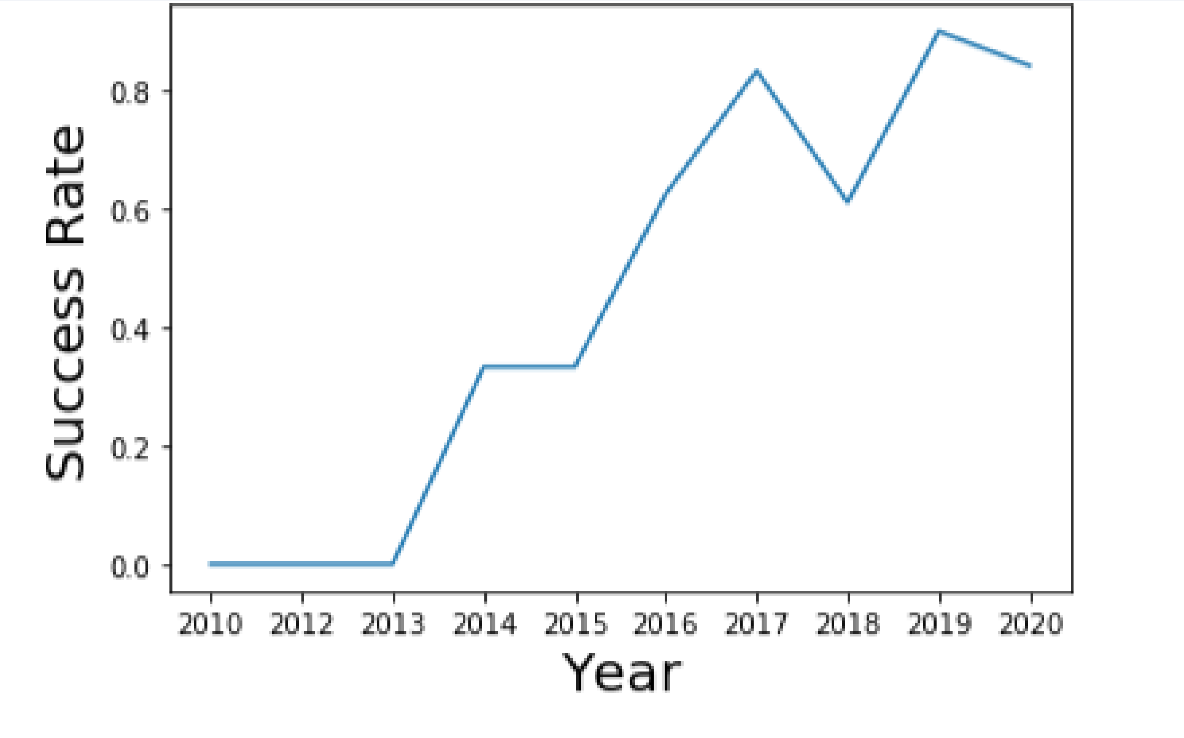
Payload vs. Orbit Type



- GEO and SO shares same payload mass whereas GEO succeed and SO failed
- For Payload mass above 9000kg, orbit type PO and ISS is a safer option
- LEO success rate is higher for Payload mass above 1000kg
- Despite varying mass in SSO, the success rate is always high

Launch Success Yearly Trend

- The success rate has been steadily increasing along the years



All Launch Site Names

```
In [6]: %sql select distinct Launch_Site from SPACEXTABLE;

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

Out[6]:

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

CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E are the four launch sites

Launch Site Names Begin with 'CCA'

In [7]:
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5;
```

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

Out[7]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

The above query was used to find the sites that begins with `CCA`

Total Payload Mass

In [8]:
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)';
```

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

Out[8]:

sum(PAYLOAD_MASS__KG_)
45596

Total mass of 45596 kg payload has been carried by boosters

Average Payload Mass by F9 v1.1

```
In [9]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1';

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

Out[9]:

avg(PAYLOAD_MASS__KG_)
2928.4

On average 2928.40kg payload mass is carried by booster version F9

First Successful Ground Landing Date

```
In [10]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'

 * sqlite:///my_data1.db
Done.

Out[10]:
```

min(Date)
2015-12-22

The first successful landing occurred on 22nd December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below 4 booster versions are observed to be successful when payload mass is between 4000 and 6000.

```
In [11]: %sql select Booster_Version from SPACEXTABLE\
         where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000;

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

Out[11]:

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

Total Number of Successful and Failure Mission Outcomes

- 100 missions had been successful whereas 1 failed.
- Group by used to count the mission outcome

List the total number of successful and failure mission outcomes

In [12]:
```
%sql SELECT Mission_Outcome, COUNT(*) as count from SPACEXTABLE\
GROUP BY Mission_Outcome ORDER BY count DESC;
```

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

Out[12]:

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

Boosters Carried Maximum Payload

- We observed the below booster version carried a high payload mass

```
In [13]: %sql select Booster_Version from SPACEXTABLE\
         where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

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

Out[13]:

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

year.

In [14]:
```
%sql select substr(Date, 6, 2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE\
where substr(Date, 1, 4) = '2015' and Landing_Outcome = 'Failure (drone ship)'
```

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

Out[14]:

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

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

- 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

In [15]:
```
%sql select Landing_Outcome, count(*) as count_outcomes \
from SPACEXTABLE\
where Date between '2010-06-04' and '2017-03-20'\
group by Landing_Outcome\
order by count_outcomes Desc;
```

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

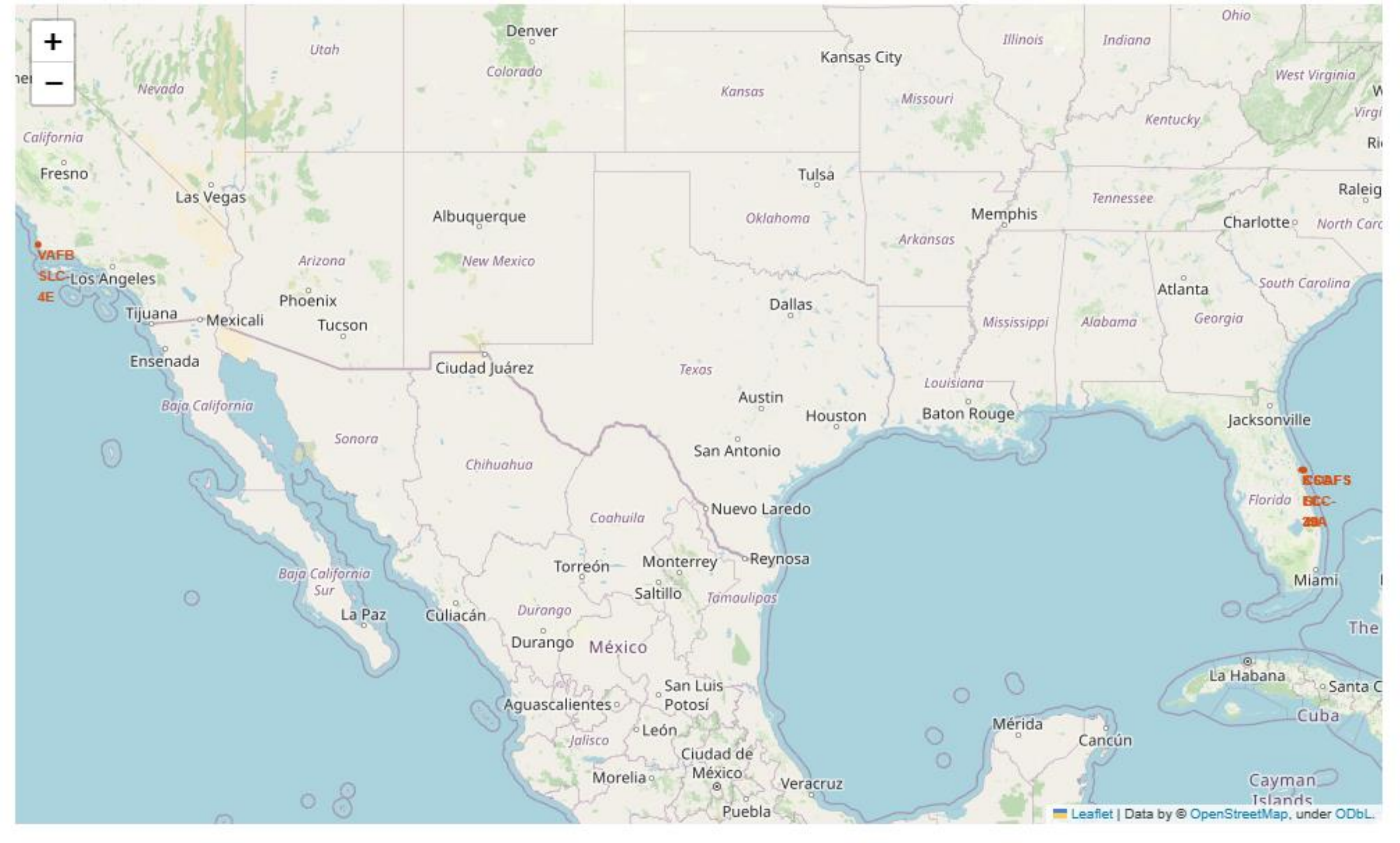
Out[15]:

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

Section 3

Launch Sites Proximities Analysis

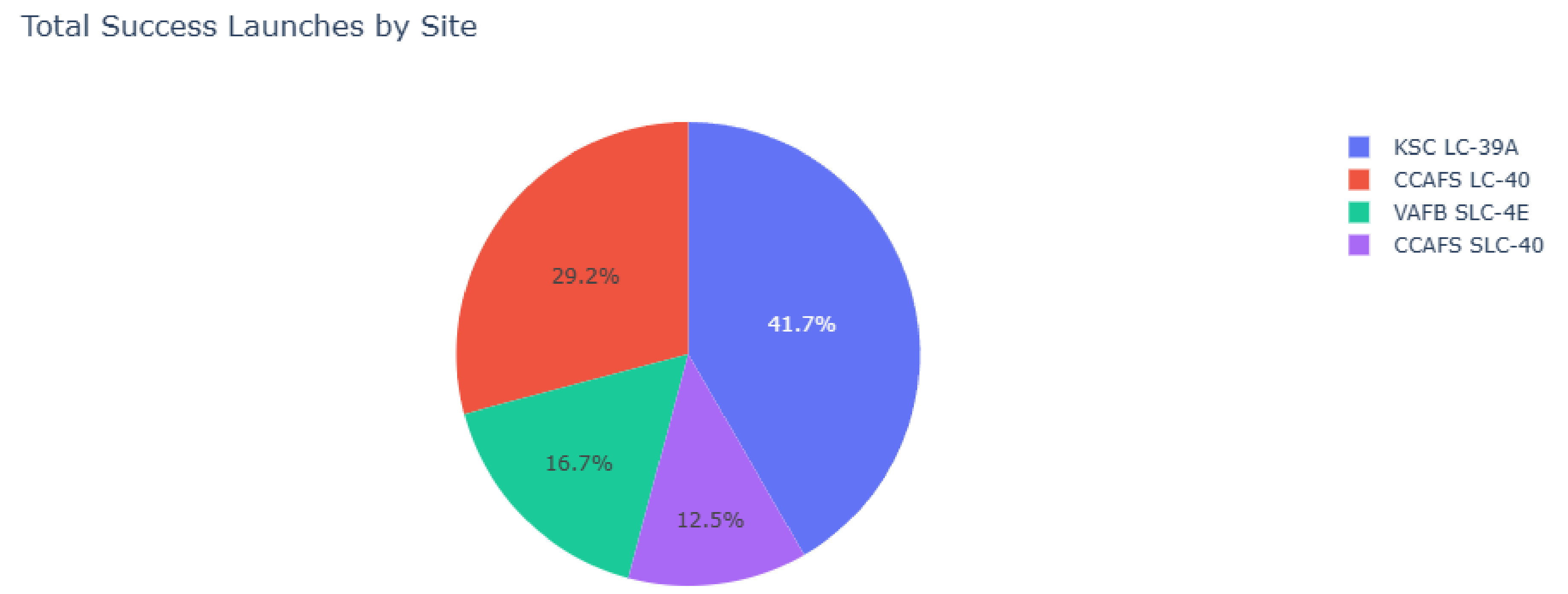
Four Launch site location



Section 4

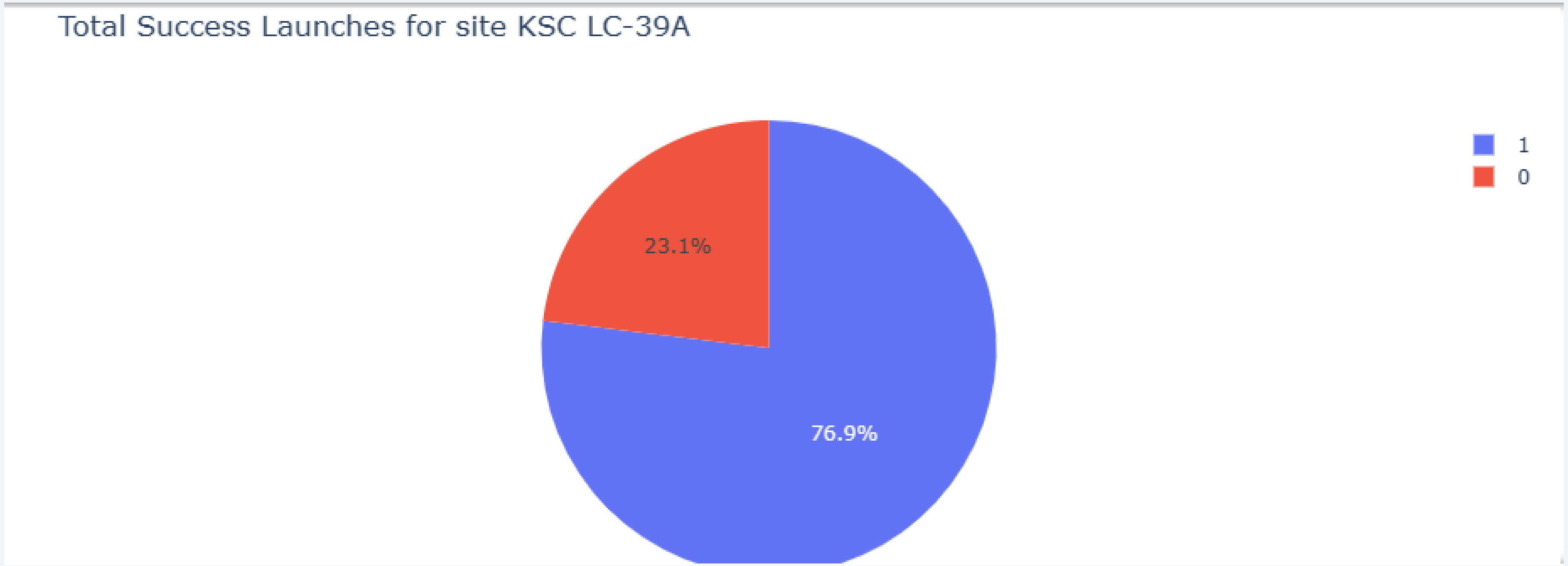
Build a Dashboard with Plotly Dash

Launch success count for all sites



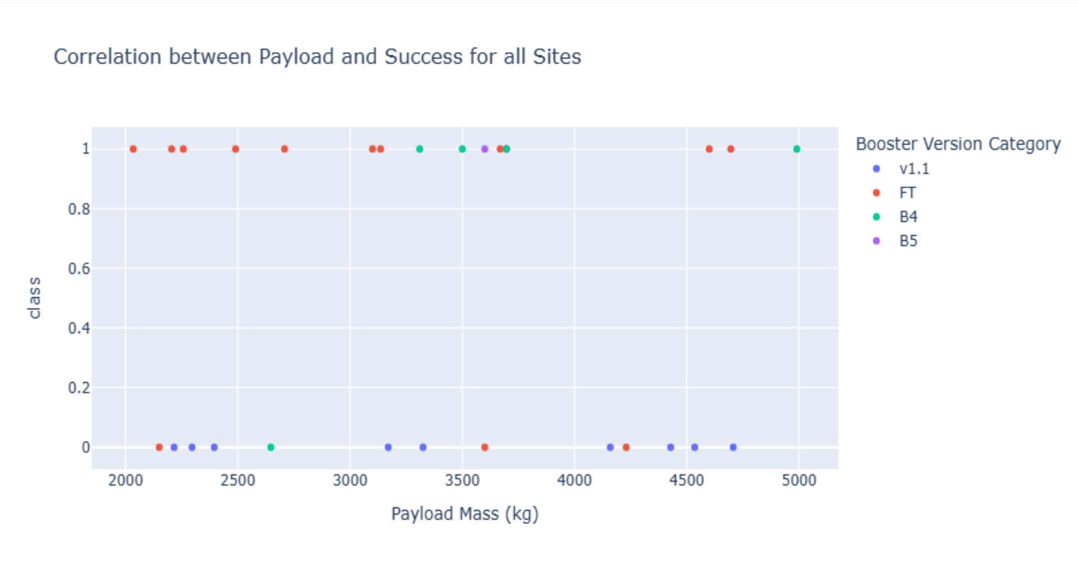
- KSC LC – 39A had the most successful launches with 41.7%

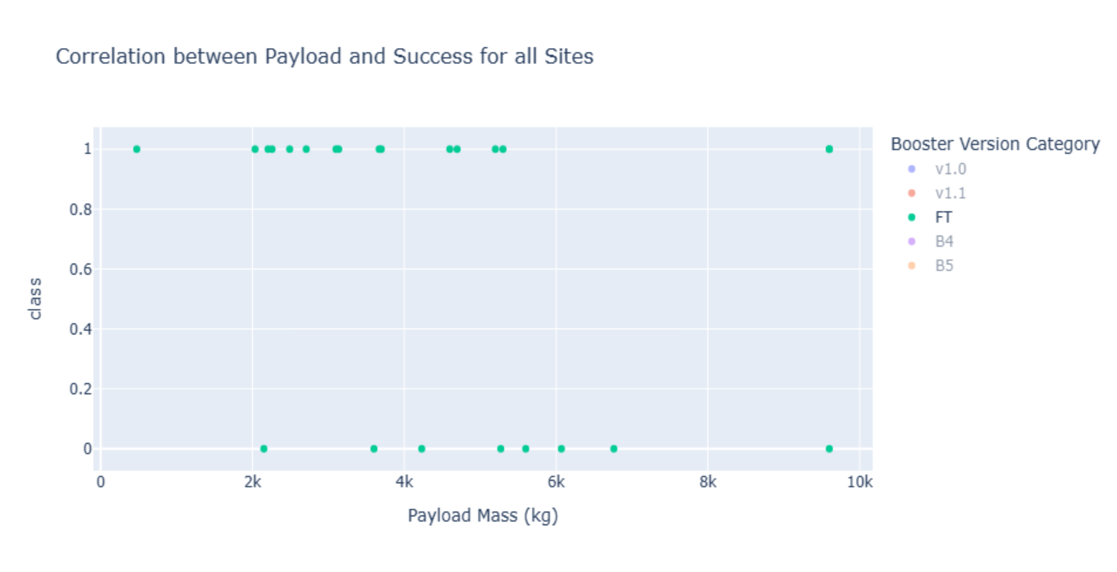
Highest launch success ratio



KSC LC – 39A has a success rate of 76.9% of all the launches which is highest among all the launch sites

Payload vs Launch Outcome





- For the payload mass of less than 5000kg, the success rates have been low.
- Booster version FT carries a higher success rate invariably at any payload mass.

Section 5

Predictive Analysis (Classification)

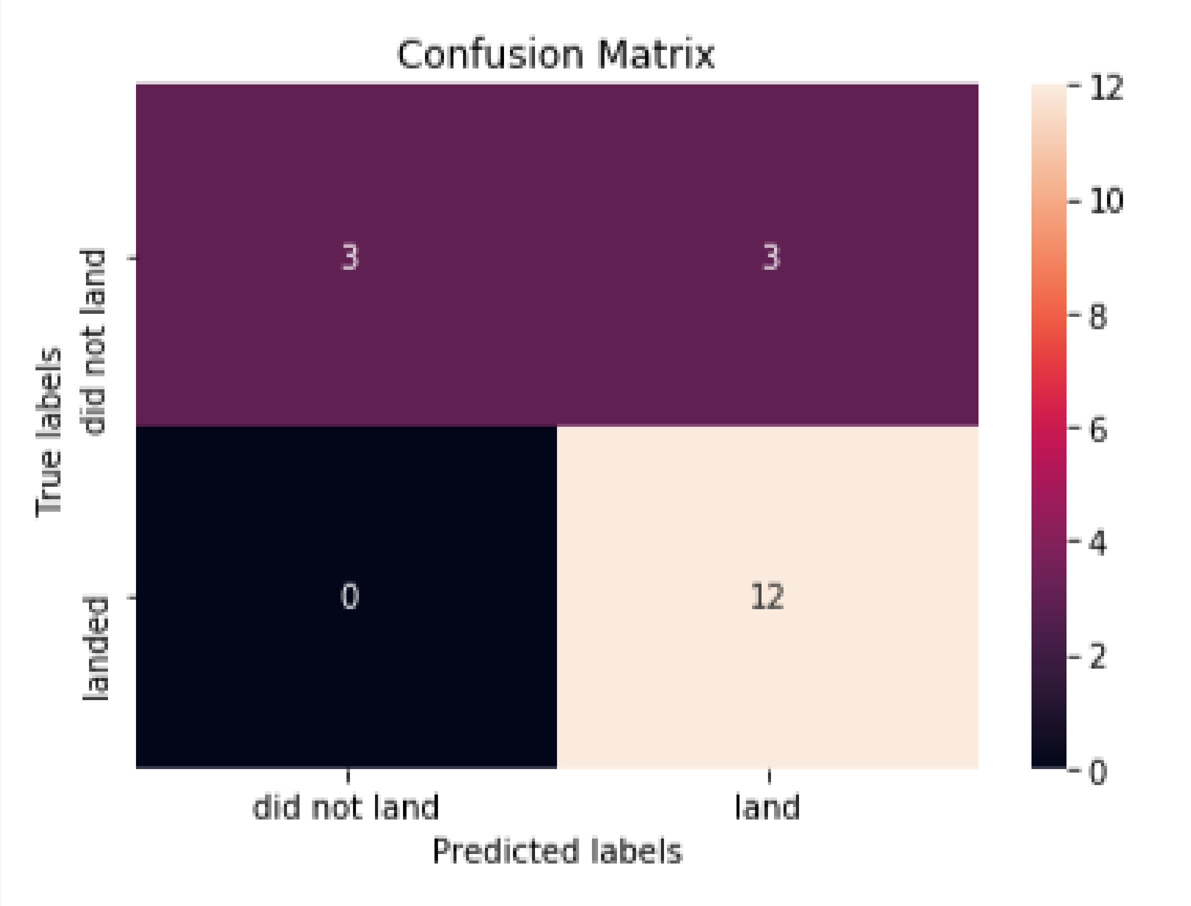
Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.750000	0.800000
F1_Score	0.888889	0.888889	0.857143	0.888889
Accuracy	0.833333	0.833333	0.777778	0.833333

- Based on the scores of the Test Set, we can confirm that classification tree performance is not at par with rest of the models
- LogReg, SVM and KNN performs almost the same

Confusion Matrix

- Results of the confusion matrix for SVM, Logistic regression and KNN were same
- Classification tree holds higher false positive
- Since the flase negative is zero the model may tend to overfit



Conclusions

For successful landing

- Launch site: KSC LC-39A had the highest success rate
- **Orbit types:** EL-L1, GEO, HEO and SSO
- Higher payload mass is much preferable

Machine learning prediction

Logistic regression, SVM and KNN are the preferred model over other

Appendix

In [46]:
```
from sklearn.metrics import jaccard_score, f1_score

# Examining the scores from Test sets
jaccard_scores = [
                  jaccard_score(Y_test, logreg_yhat, average='binary'),
                  jaccard_score(Y_test, svm_yhat, average='binary'),
                  jaccard_score(Y_test, tree_yhat, average='binary'),
                  jaccard_score(Y_test, knn_yhat, average='binary'),
                 ]

f1_scores = [
             f1_score(Y_test, logreg_yhat, average='binary'),
             f1_score(Y_test, svm_yhat, average='binary'),
             f1_score(Y_test, tree_yhat, average='binary'),
             f1_score(Y_test, knn_yhat, average='binary'),
            ]

accuracy = [logreg_score, svm_score, tree_score, knn_score]

scores_test = pd.DataFrame(np.array([jaccard_scores, f1_scores, accuracy]), index=['Jaccard_Score', 'F1_Score', 'Accuracy'] , co
scores_test
```

Out[46]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.750000	0.800000
F1_Score	0.888889	0.888889	0.857143	0.888889
Accuracy	0.833333	0.833333	0.777778	0.833333

Thank you!