Winning Space Race with Data Science

DAT LE

October 19th 2021

Outline

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

Executive Summary

- ❖ Summary of methodologies
  - ✓ Data collection
  - ✓ Data wrangling
  - ✓ EDA with data visualization
  - ✓ EDA with SEQ
  - ✓ Data visualization with Folium
  - ✓ Data visualization with Plotly Dash
  - ✓ Predictive method using ML classification
- ❖ Summary of all results
  - ✓ EDA results
  - ✓ Data analysis through different types of plots
  - ✓ Predictive analysis

Introduction

- ❖ Project background
  - ✓ We will predict if the Falcon 9 first stage will land successfully
  - ✓ Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars
  - ✓ Other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage
- ❖ Problems you want to find answers
  - ✓ If we can determine if the first stage will land, we can determine the cost of a launch
  - ✓ The correlation of multiples factors (payload, launch sites, launching year, etc.) on the success rate of the landing of the first stage of Falcon 9
  - ✓ Identify the requirements of launch sites (close to coastline, or the residential area?)
  - ✓ Which ML model provides the best prediction for the success rate of the rocket launch

Section 1

Methodology

Methodology

Executive Summary

- ❖ Data collection methodology:
  - ✓ Using Space X TEST API
  - ✓ Web Scrapping from Wikipedia
- ❖ Perform data wrangling
  - ✓ Describe how data was processed (convert training labels to 1- booster successfully landing-, and 0-unsuccessful landing)
- ❖ Perform exploratory data analysis (EDA) using visualization and SQL
- ❖ Perform interactive visual analytics using Folium and Plotly Dash
- ❖ Perform predictive analysis using classification models
  - ✓ How to build, tune, evaluate classification models

Data Collection

- ❖ We collect data illustrating information related to the rocket launching of the Space X project such as payload, locations of launch sites, type of launch pads, rocket types, etc.

- ❖ The collected data will be then used to predict whether the landing of first stage of future Falcon 9 launches successful or not

- ❖ Required data were collect using 2 methods
  - ✓ Space X REST API
  - ✓ Web scrapping Wikipedia using Beautiful Soup

Data Collection – SpaceX API

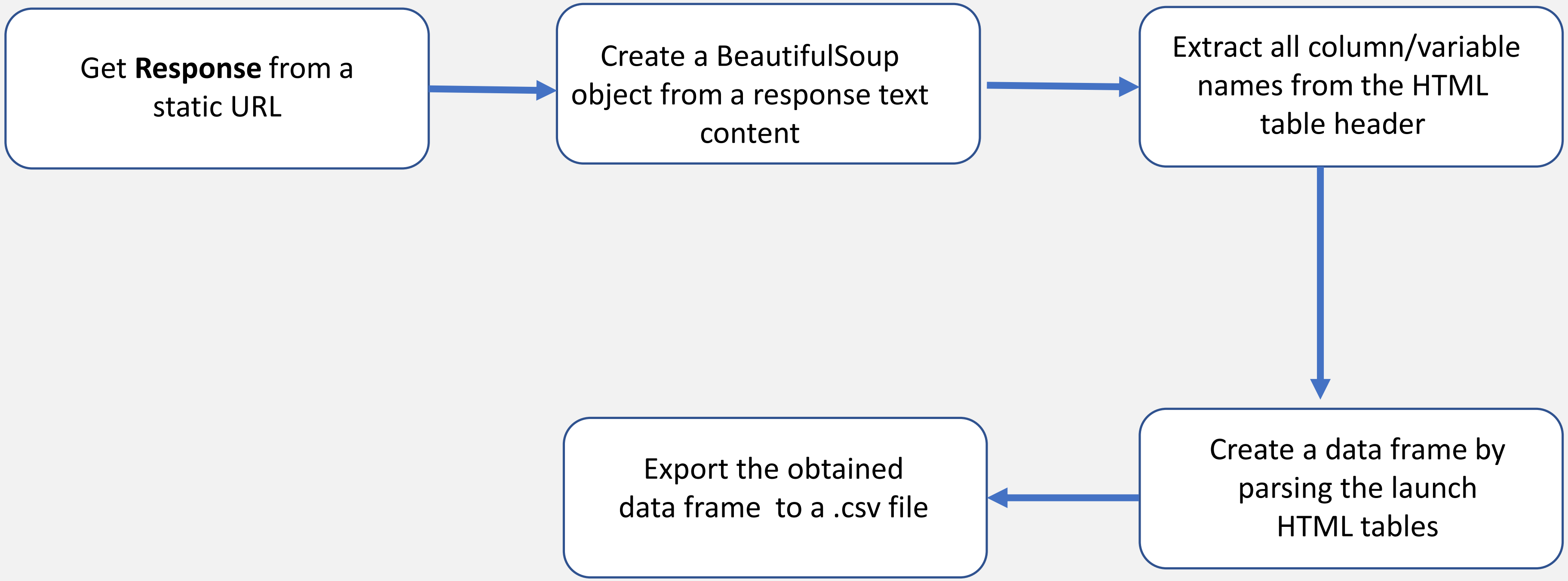
The flowchart of the data collection with Space X REST calls

1. Getting **Response** from Space X REST API
2. Convert **Reponse** to .JSON result, then formalize it into a data frame
3. Using custom functions to extract required information from the data frame
4. Using custom functions to extract information from the data frame
5. Use extracted information to create a data dictionary → new data frame
6. Filter the data frame to obtain Falcon 9 data, and export it to a .csv file

Github link

Data Collection – Web Scrapping

The flowchart of the data collection REST calls



[Github link]

Data Wrangling

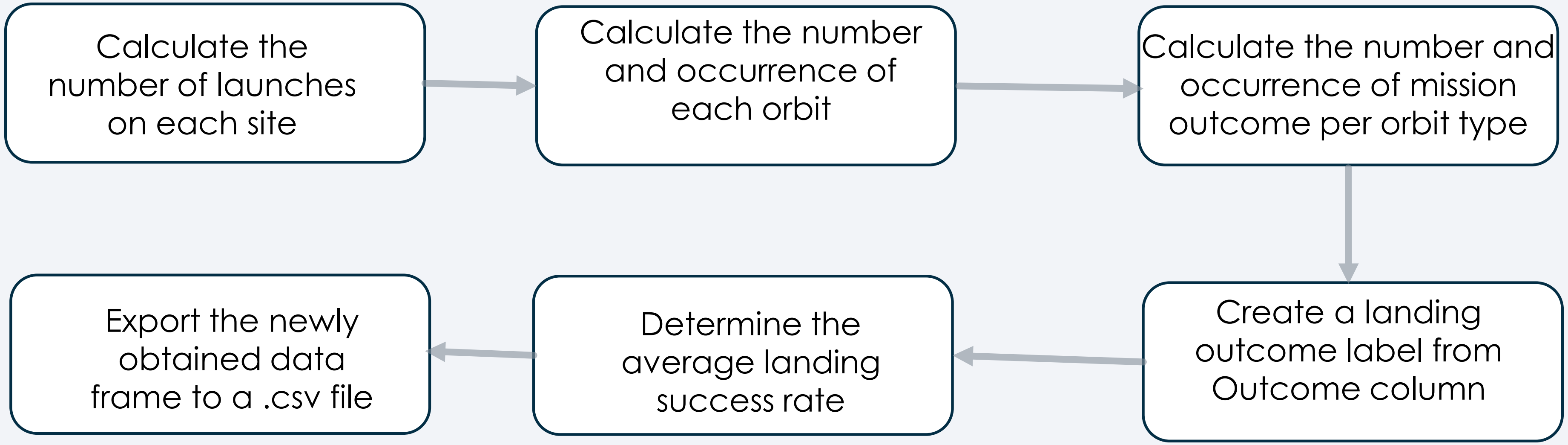
- ❖ We need to find some patterns in the data and determine what would be the label for training supervised models.

- ❖ In the data set, there are several different cases where the booster did not land successfully
  - ✓ *Ex:*
    - o ***False Ocean*** - unsuccessfully landing to the ocean/ ***True Ocean*** - successfully landing
    - o ***False RTLS*** - unsuccessfully landing to a ground pad/ ***True RTLS*** - successfully landing

- ❖ We will mainly convert the outcomes into Training Labels with **1** means the booster successfully landed, **0** means it was unsuccessful.

Github link

Data Wrangling Workflow



EDA with Data Visualization

[Github link]

- Scatter graphs is used to identify the relationship of one variable with another (i.e., correlation or trend patterns) It also helps in detecting outliers in the plot
  - ✓ *Flight number vs Payload Mass*
  - ✓ *Flight number vs Launch site*
  - ✓ *Payload vs Launch site*
  - ✓ *Orbit type vs Flight number*
  - ✓ *Payload vs Orbit type*
  - ✓ *Orbit type vs Payload mass*
- Bar Graph is suitable in the case when one axis of the chart shows the specific categories being compared, and the other axis represents a measured value
  - *Mean vs Orbit*
- Line graph is used to track changes over short and long periods of time, and hence clearly showing an increasing/decreasing trend.
  - *Success rate vs Year*

EDA with SQL

[Github link]

- **Using SQL queries to obtain necessary information from the data set**
  - ✓ *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 in year 2015*
  - ✓ *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*

Build an Interactive Map with Folium

- Circle Markers are used to highlight locations of the launch sites on a map with text labels. The location of a launch site is identified thanks to the site's latitude and longitude coordinates

- Markers are created for all launch records. If a launch was successful (class=1), then we use a green marker and if a launch was failed, we use a red marker (class=0)

- A launch only happens in one of the four launch sites, which means many launch records will have the exact same coordinate. Marker clusters is used to simplify a map containing many markers having the same coordinate

- Polyline markers are used to display the distances between a launch site to its nearest coastline, railway, highway, and city

Github link

Build a Dashboard with Plotly Dash

- ❖ Pie graph is used to illustrate
  - ✓ The numbers of successful launches at all launch sites
  - ✓ The radio between the numbers of successful and unsuccessful launches at a selected site
- ❖ Scatter graph is used to illustrate the relationship between the Payload Mass (Kg) and the Outcome given different Booster versions
  - ✓ Different Boosters are displayed using different colors (legends) in the scatter graph

Github link

Predictive Analysis (Classification)

❖ ***BUILDING MODEL***

- ✓ Create a data frame X and a Numpy array Y representing the labels
- ✓ Standardize the data frame X
- ✓ Split data X into training and test data sets
- ✓ Check how many samples in the test data set
- ✓ Use 4 different classification algorithms: ***Logistic regression, SVM, Decision Tree, and KNN***
- ✓ Create a **GridSearchCV** object with the corresponding parameters dictionary for each classification algorithm
- ✓ Fit our data sets into the **GridSearchCV** and train our data sets

❖ ***Evaluation Phase***

- ✓ Check accuracy of each model
- ✓ Get the best hyper parameters for all of the 4 classification algorithms
- ✓ Check accuracy using the test data sets
- ✓ Plot confusion matrix with each algorithm
- ✓ Find the algorithm yielding the best accuracy scores

Github link

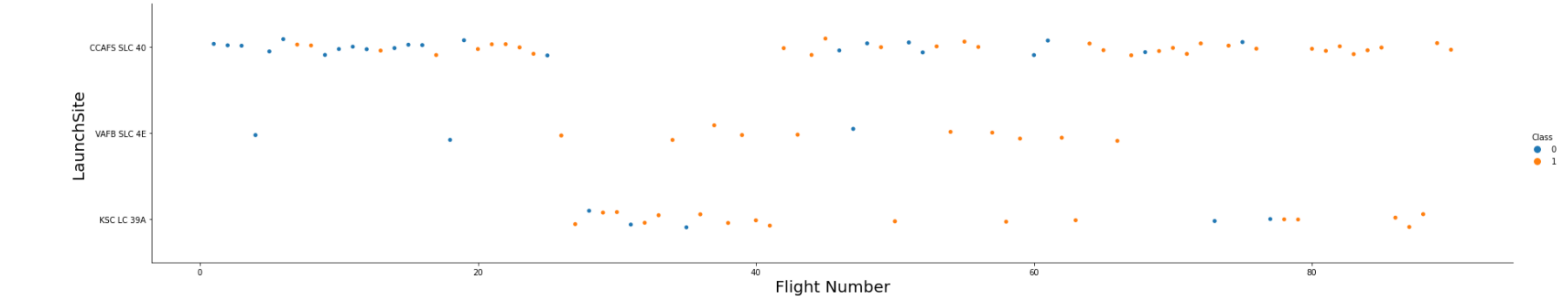
Results

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

Section 2

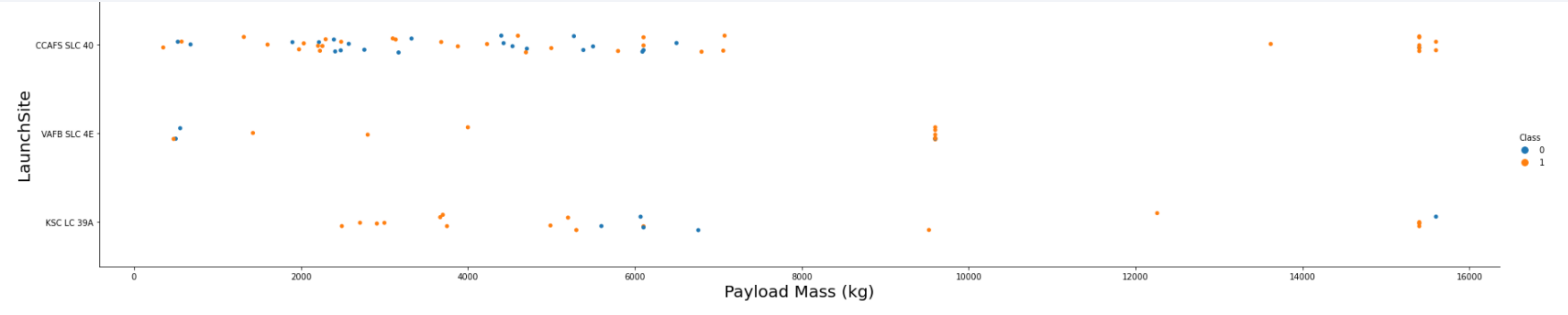
Insights drawn from EDA

Flight Number vs. Launch Site



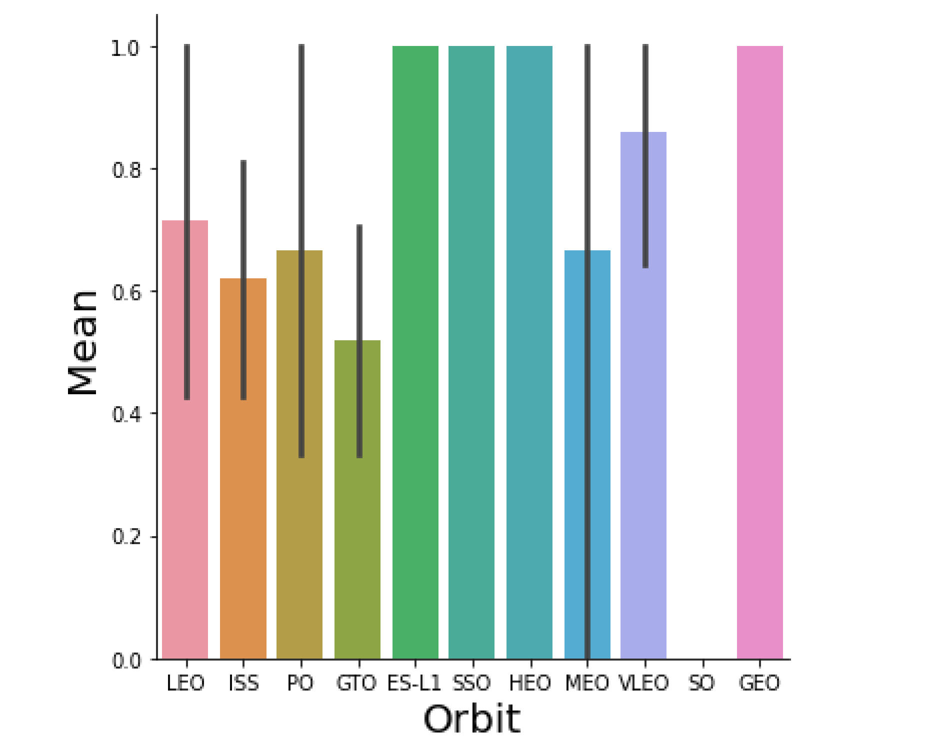
- As the flight number increases, the landing of the 1st stage is more successful
- The KSC-39A and VAFB SLC 4E have more successful landings than the CCAFS LC-40

Payload vs. Launch Site



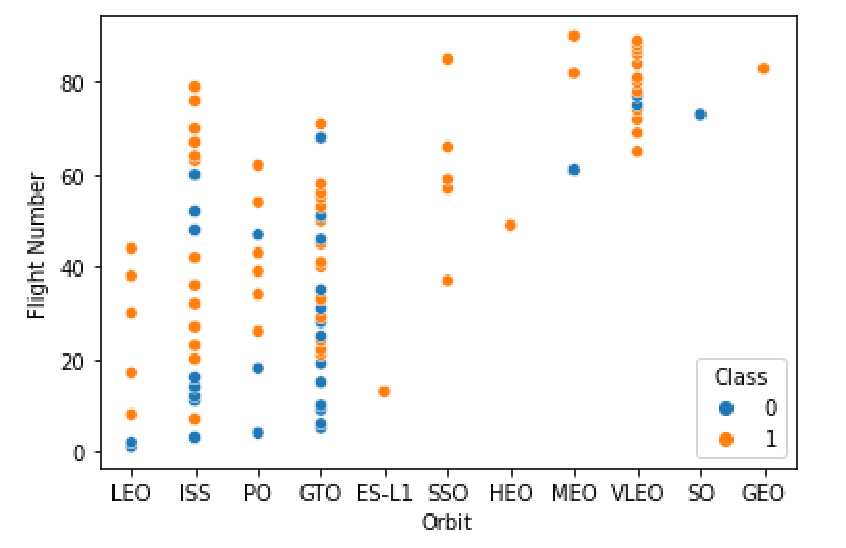
- ❖ For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000)
- ❖ As payload mass increases at Launch Site CCAFS SLC 40, the success rate goes up

Success Rate vs. Orbit Type



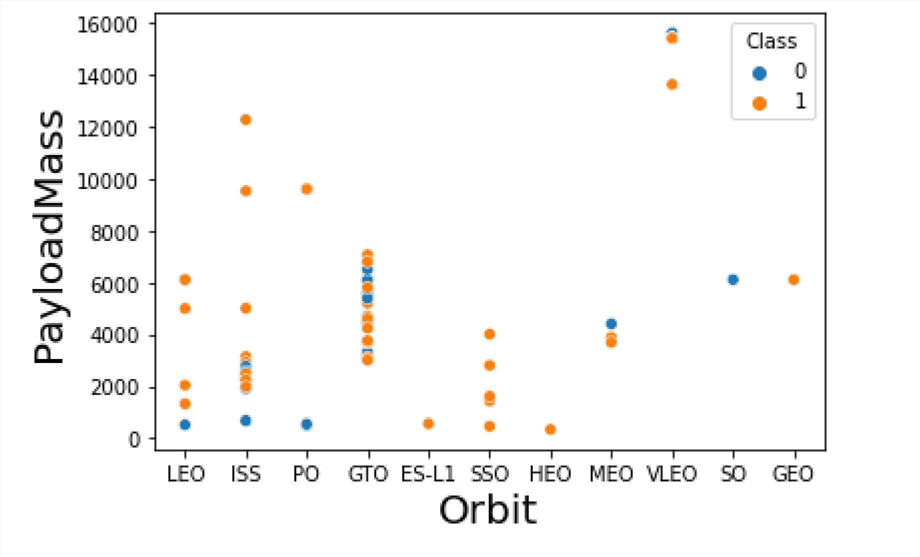
❖ Orbit GEO,HEO,SSO,ES-L1 has the best success rate

Flight Number vs. Orbit Type



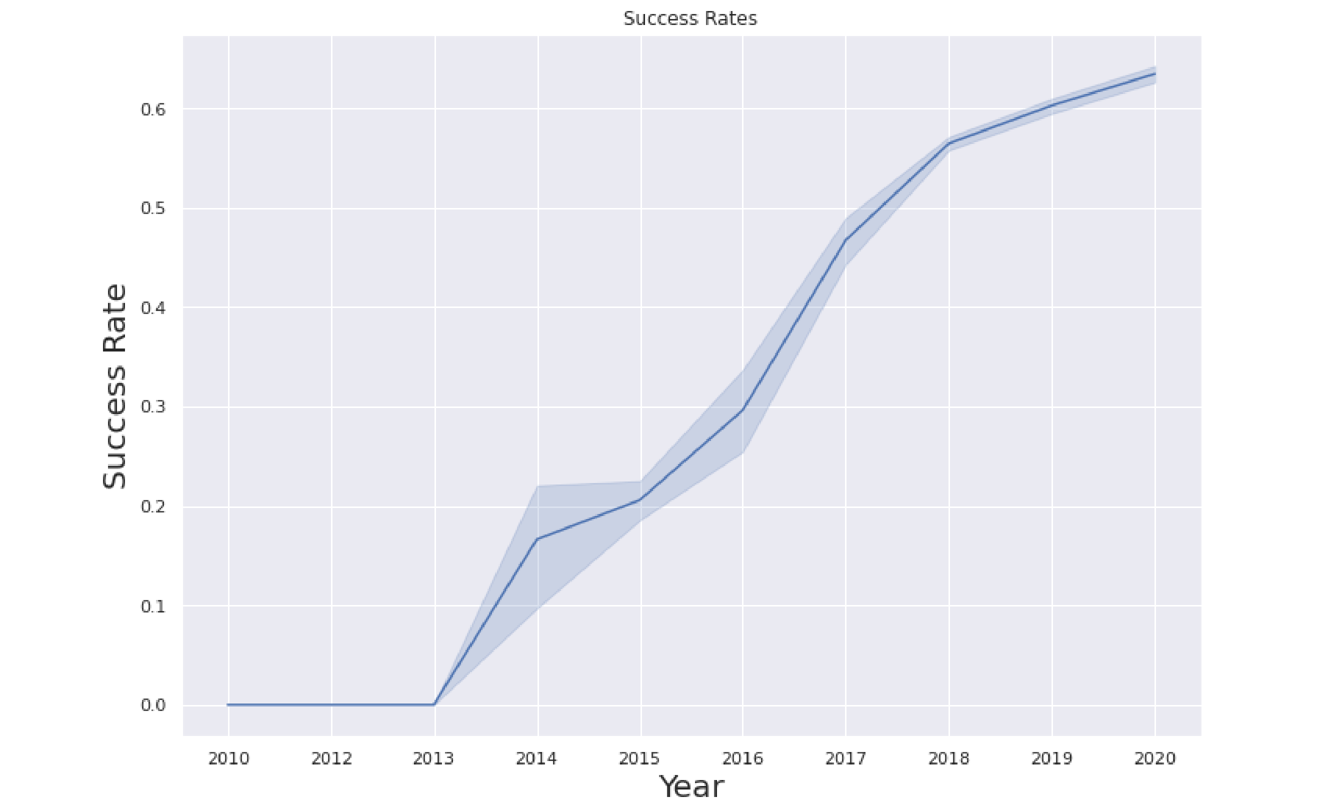
- You should see that in the LEO orbit the Success appears related to the number of flights
- It seems that there is no relationship between flight number when in GTO orbit

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



❖ The success rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

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

❖ Using DISTINCT to get unique launch site names from SPACXBTL

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

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

- Using **Like '%CCA%** to display all the launch site starting with 'CCA'
- Using **limit 5** to only display 5 records

Total Payload Mass

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer='NASA (CRS)'
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
45596

- Using **sum** on **Payload_mass__kg_** to calculate total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
2534

- Using **avg** on **payload_mass__kg_** to calculate the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql select min(DATE) from SPACEXTBL where landing__outcome like 'Success (ground pad)'
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

1
2015-12-22

- Using **min** on **DATE** to display the the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTBL where landing__outcome like 'Success (drone ship)' and payload_mass__kg_
between 4000 and 6000
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

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

- ❖ List the names of the boosters which have success in drone ship (**like 'Success (drone ship)'**) and have **payload_mass__kg_** greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(mission_outcome) from SPACEXTBL
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
101

❖ Using **count** on **mission_outcome** to calculate the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTBL)

 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

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

- List the names of the **booster_version** which have carried the maximum payload mass.
- A sub query is used to obtain the maximum payload mass

2015 Launch Records

```
%sql select booster_version,launch_site from SPACEXTBL where YEAR(DATE)='2015' and landing__outcome='Failure (drone shi
p)'
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- List 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 count(landing__outcome) from SPACEXTBL where landing__outcome='Failure (drone ship)' and (DATE between '201
0-06-04' and '2017-03-20')
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

1
5

```
%sql select count(landing__outcome) from SPACEXTBL where landing__outcome like '%Failure%' and (DATE between '2010-06-0
4' and '2017-03-20')
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

1
7

```
%sql select count(landing__outcome)  from SPACEXTBL WHERE (landing__outcome  like '%Success%') and (DATE between '2010-
06-04' and '2017-03-20')
```

```
 * ibm_db_sa://xhz67687:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:3273
3/bludb
Done.
```

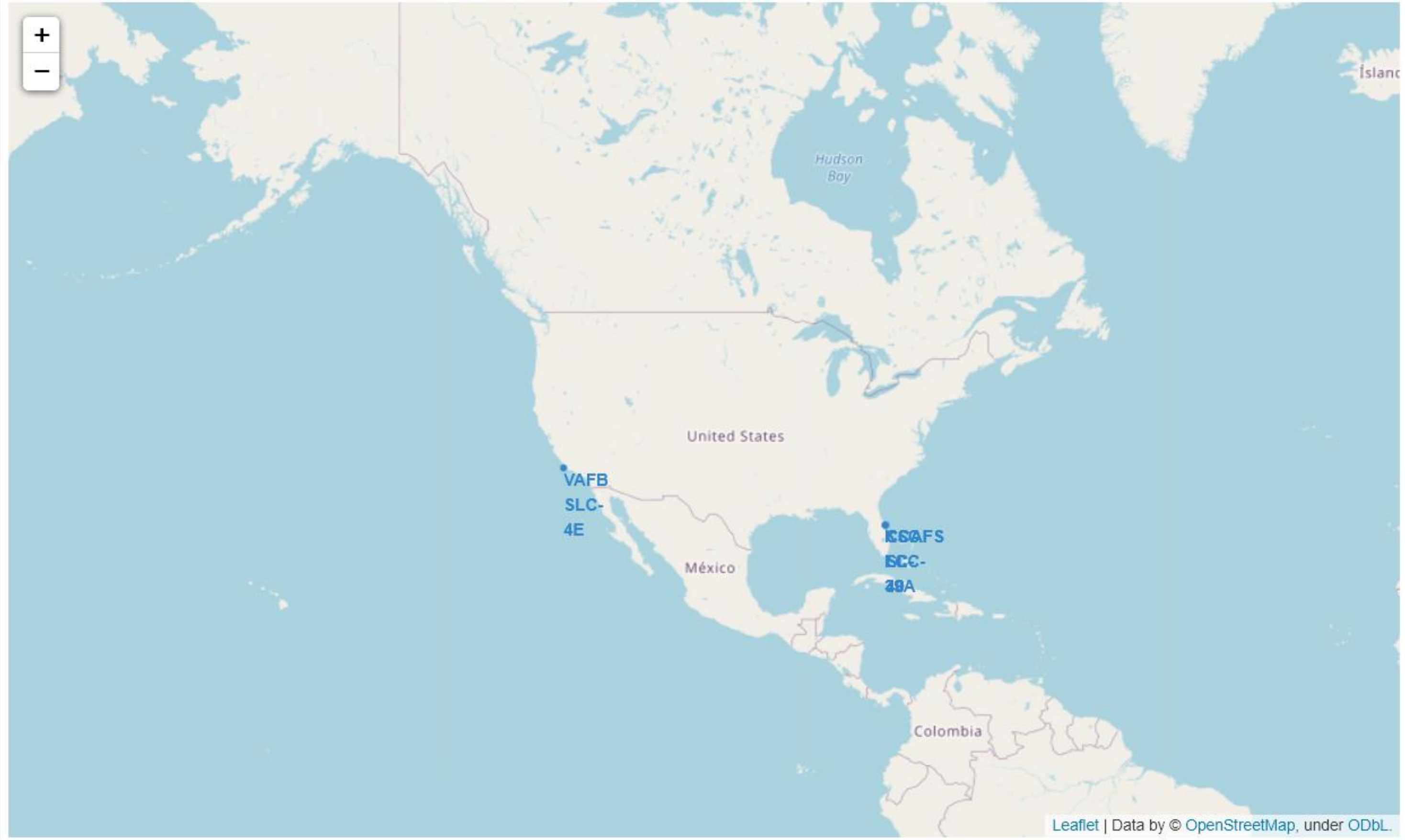
1
8

- Rank the count of landing outcomes (such as **Failure (drone ship)** or **Success**) between the date 2010-06-04 and 2017-03-20, in descending order

Section 4

Launch Sites Proximities Analysis

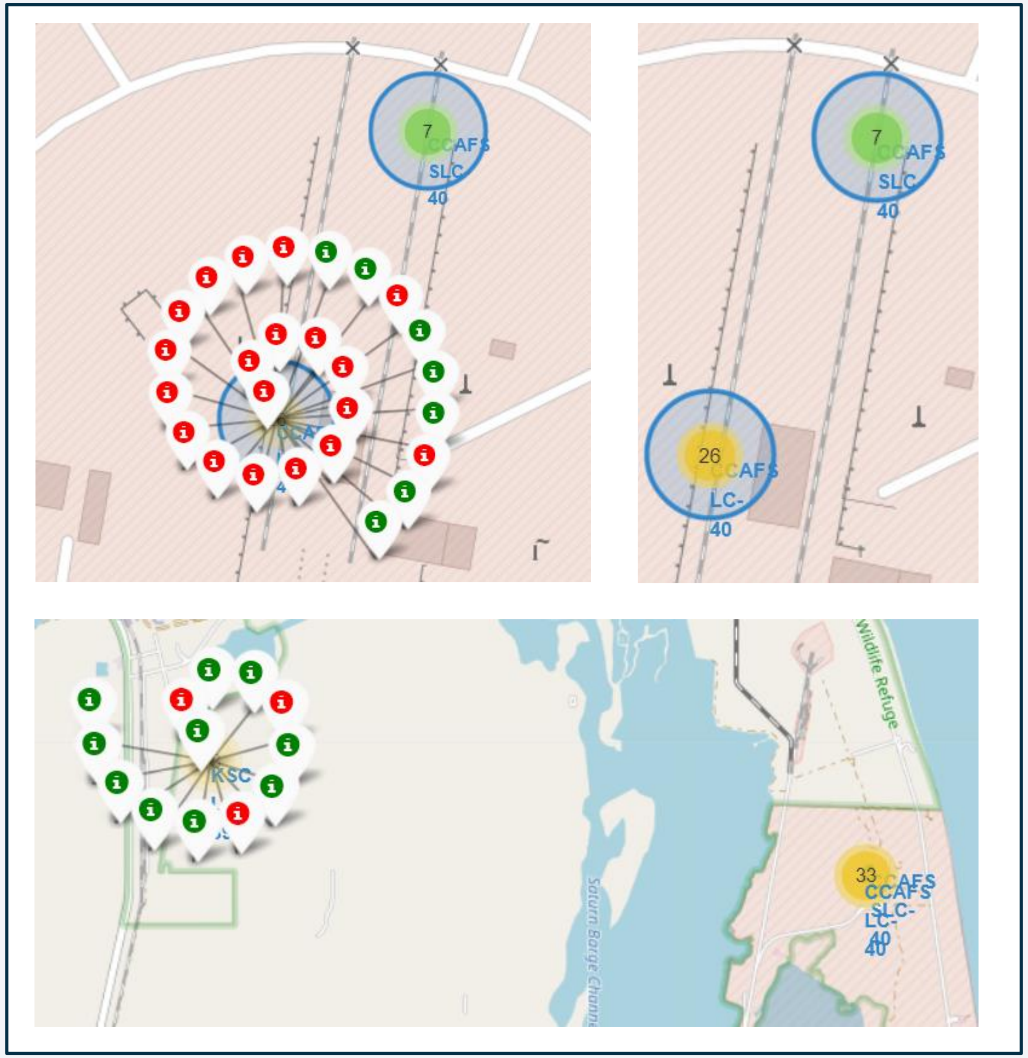
Launching Sites of SPACE X



- ❖ We can see that all the launching sites are located along the coastline (West and East coasts)

Launch-site labels and markings

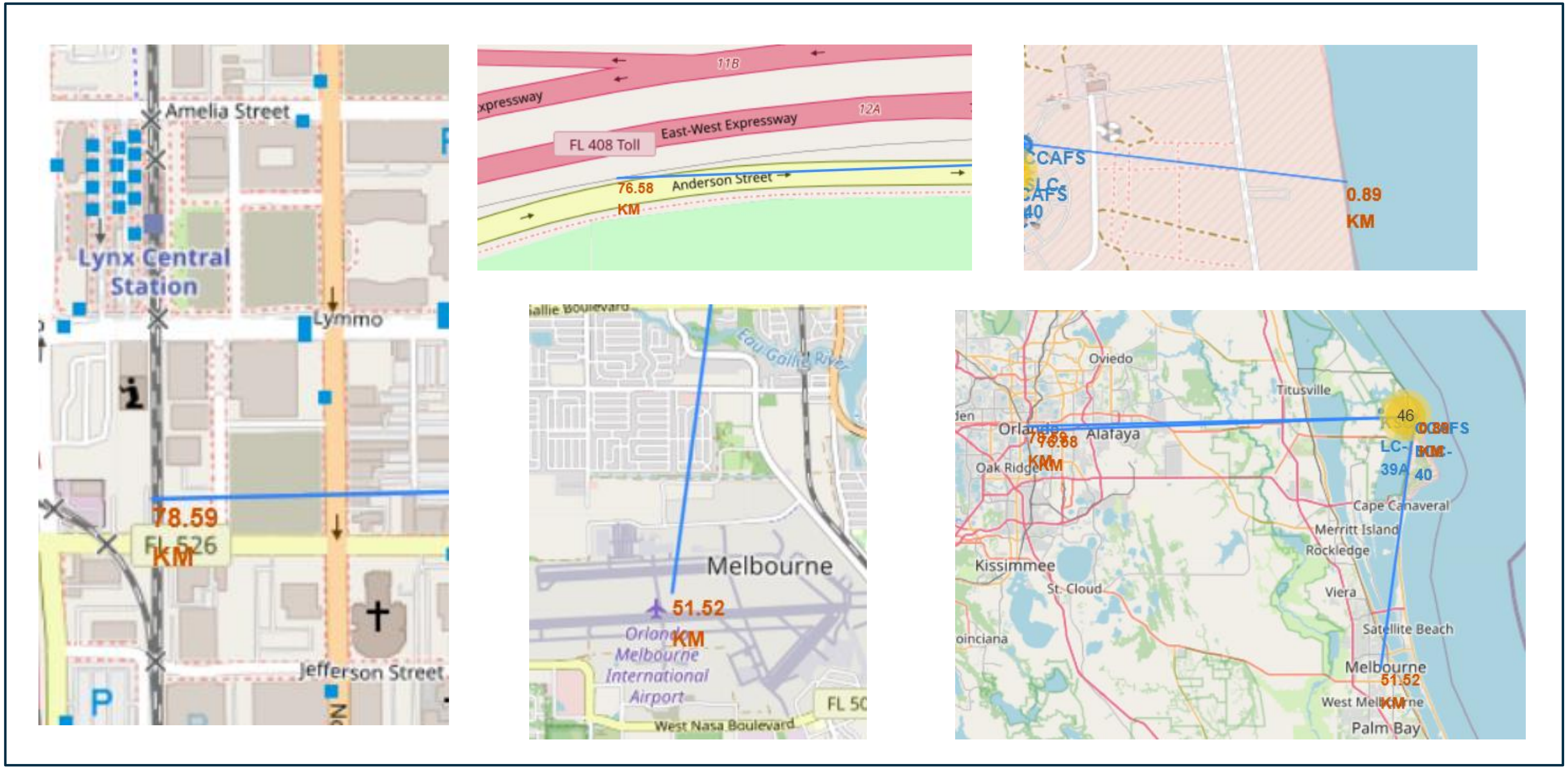
East coast



West coast

- ✓ There are more launch sites on the East coast than the West coast
- ✓ Blue marker means successful landing while red marker represents a failed landing

Characteristics of launch site locations

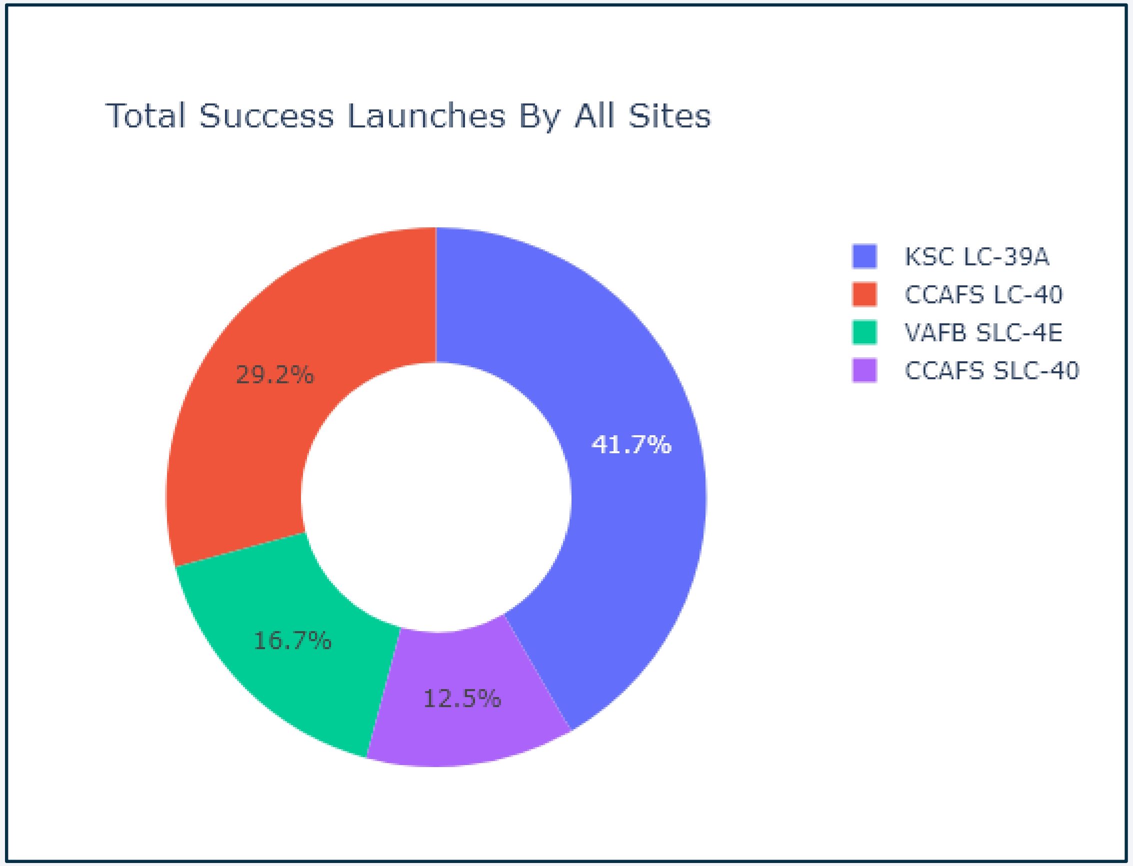


- The launch sites should be located near the coastlines and distanced from railways, highways, and residential areas

Section 5

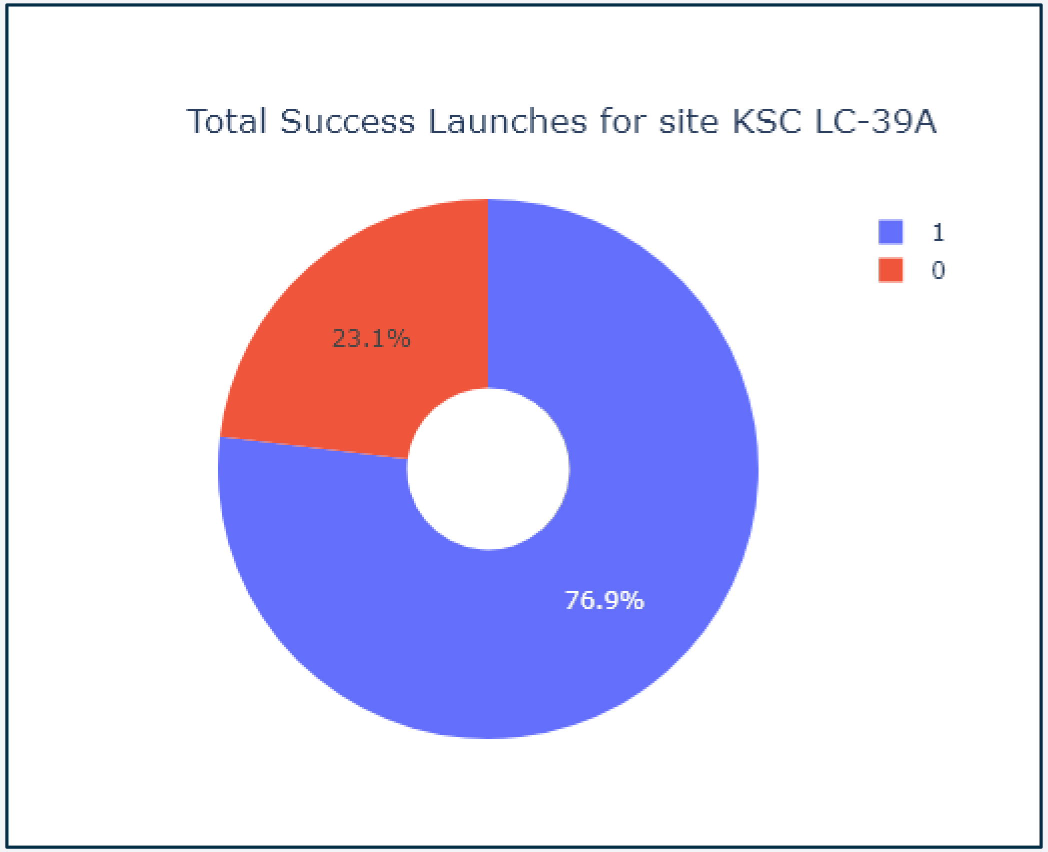
Build a Dashboard with Plotly Dash

Launch success count for all sites



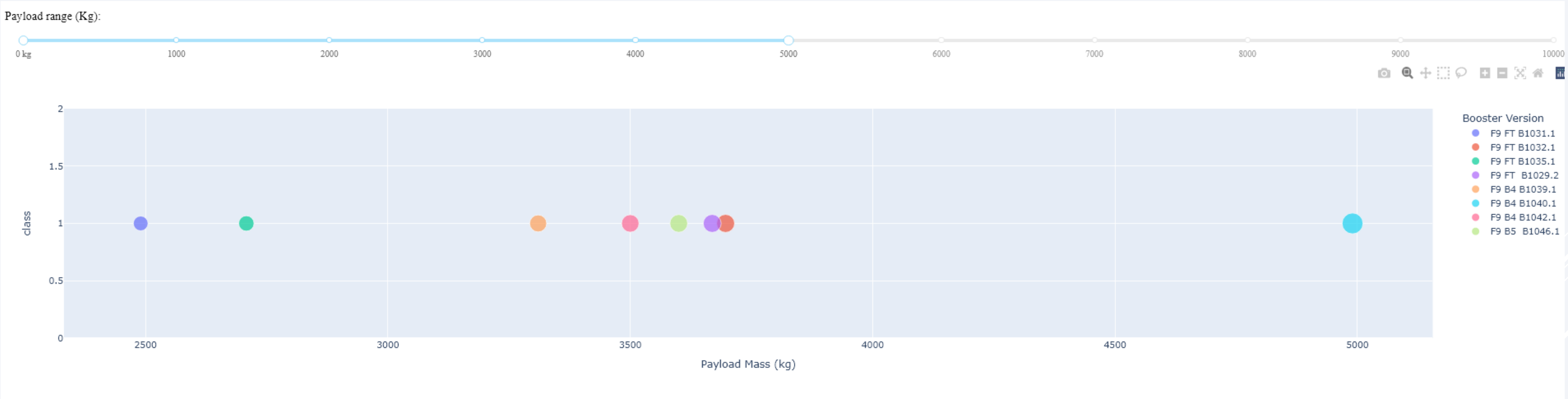
- *Launch site KSC LC-39A had the most successful launches from all the sites*

Launch site with highest launch success



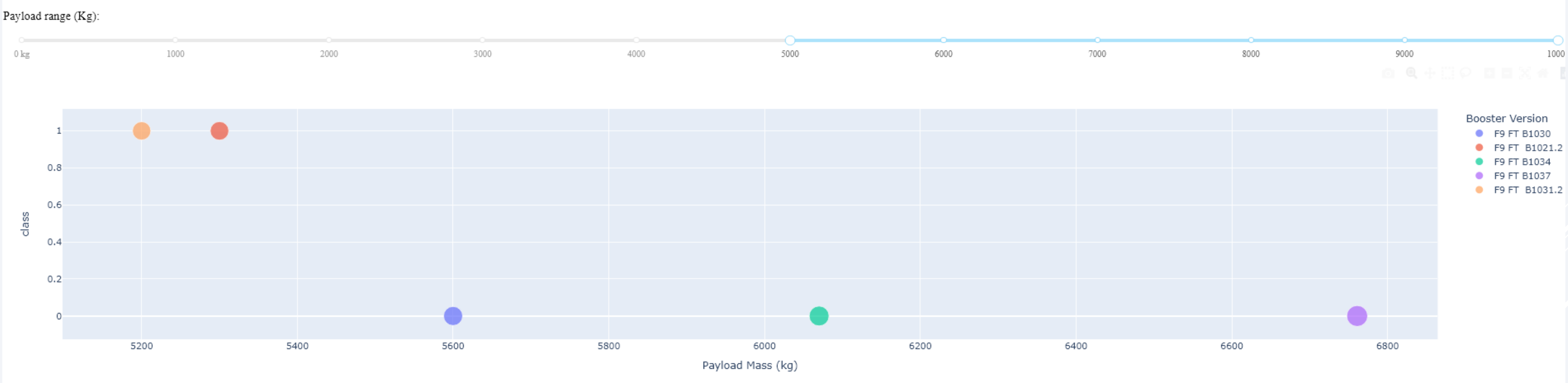
❖ *KSC LC-39A has the highest success rate 76.9% amongst all launch sites*

Payload vs. Launch Outcome



- ❖ Scatter graph for Payload vs. Launch Outcome with the payload in the range of [0:5000] kg

Payload vs. Launch Outcome



✓ Scatter graph for Payload vs. Launch Outcome with the payload in the range of [5000:10000] kg

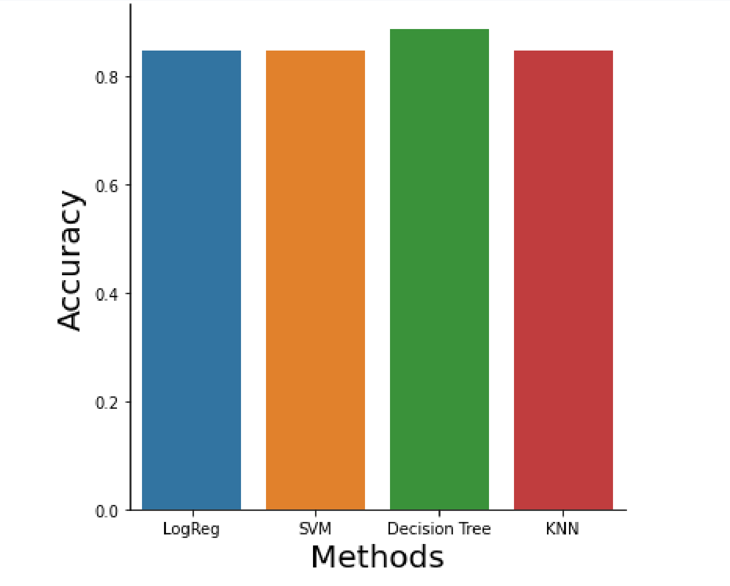
→ The lighter the payload of the rocket is, the more success rate the launch gets

Section 6

Predictive Analysis (Classification)

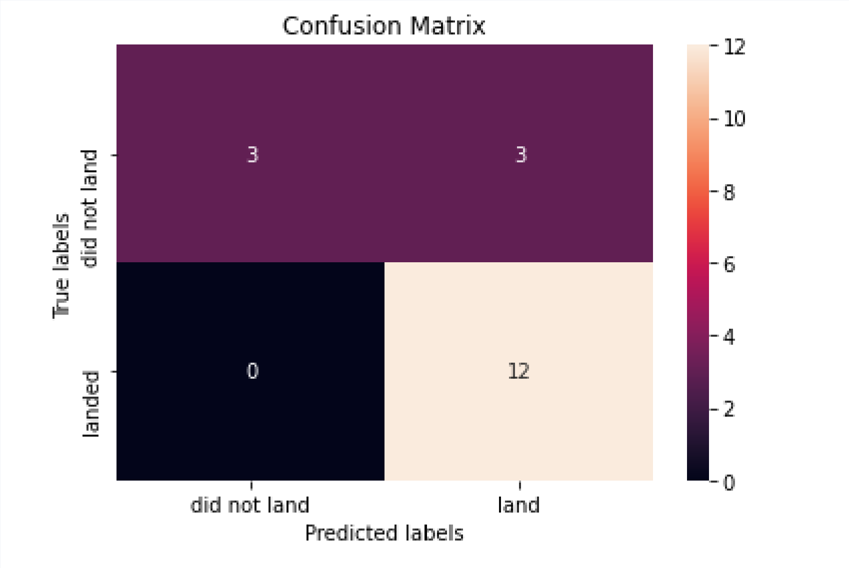
Classification Accuracy

	Methods	Accuracy
0	LogReg	0.846429
1	SVM	0.848214
2	Decision Tree	0.889286
3	KNN	0.848214



- ❖ From the bar graph, we can see that the Decision Tree method yielded the best Accuracy with the training and validation data sets
- ❖ The Decision Tree methods had the success rate of 83.33% with the test data set

Confusion Matrix



- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes

- We see that the major problem is false positives, which means that the true label is "**did not land**" but predicted as "**land**'

Conclusions

- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate
- *Launch site KSC LC-39A had the most successful launches from all the sites*
- *KSC LC-39A has the highest success rate 76.9% amongst all launch sites*
- The lighter the payload of the rocket is, the more success rate the launch gets
- Decision Tree method yielded the best Accuracy with the training and validation data sets

Appendix

- ✓ Data Analysis with Python – Final Assignment
- ✓ Machine Learning with Python – Final Assignment
- ✓ Data Science Project

Thank you!