IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeff Bleaney
April 8, 2023

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection using SpaceX Rest API and Web Scrapping
  - Exploratory Data Analysis using data wrangling, data visualization, and an interactive dashboard
  - Predictive Analysis using Machine Learning
- Summary of all results:
  - Successfully collected data from public sources
  - Determined key factors in identifying a successful landing
  - Identified the best predictive model

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, while other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. So, if we can determine if the first stage will land, we can determine the cost of a launch
- We aim to identify the key factors affecting the success of the first stage landing, and use this information to predict the outcomes of future landings

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Utilized SpaceX REST API to extract data from SpaceX website (https://api.spacxdata.com/v4/rockets/)
  - Srapped data from spacex wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Filled missing values using attribute mean, and utilized One Hot Encoding for categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built and compared Logistical Regression, Support Vector Machine, Decision Tree Classification, and k-Nearest Neighbors models

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

Using the public API offered by SpaceX, data was obtained as follows:

- Perform a request to the API
- Decode the response as a Json, and convert into a Pandas dataframe
- Filter the data frame to the relevant column, and eliminate rows with multiple payloads
- Extract dates from the data, and limit our dataset to launches occurring prior to Nov. 13, 2020
- Extract the relevant information from each column, and construct a new dataframe with the extracted data
- Filtered data frame to only include Falcon 9 launches
- Filled missing value with the mean of each column

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/Space%20X%20Falcon%209%20First%20Stage%20Landing%20Prediction.ipynb

Data Collection - Scraping

Additional information pertaining to the SpaceX Launches was obtained from Wikipedia as follows:

- Perform a request to the Wikipedia page
- Create a BeautifulSoup object with the response
- Collect data by parsing the HTML tables
- Convert table data into a Pandas dataframe

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/Space%20X%20Falcon%209%20Web%20Scraping.ipynb

Exploratory Data Analysis - Data Wrangling

Exploratory Data Analysis was performed to obtain the following information:

- Number of launches from each site
- Number of occurrences of each orbit type
- Number and occurrence of each mission outcome per orbit type
- Landing Class labels, which were appended to the dataframe

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/Space%20X%20data%20wrangling.ipynb

EDA with Data Visualization

Exploratory Data Analysis is performed with the use of scatter plots, line charts, and bar charts to visualize the relationship between the following variables:

- Payload mass and flight number
- Launch site and flight number
- Launch site and payload mass
- Orbit type and flight number
- Payload mass and orbit type
- Year and Success rate

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/Space%20X%20EDA%20with%20Visualization.ipynb

EDA with SQL

Exploratory Data Analysis is performed using SQL queries to obtain the following:

- Names of the unique launch sites in the space mission
- Total payload mass carried by boosters launched by NASA
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful ground pad landing was achieved
- Names of boosters which have success in drone ship landings with a payload mass between 4000, and 6000 kg
- Total number of successful and unsuccessful mission outcomes

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/SpaceX%20EDA%20with%20SQL.ipynb

EDA with SQL (Cont'd)

- Names of the booster versions which have carried the maximum payload mass
- A list of records displaying the month, failure landing outcomes in drone ship, booster version, and launch site for the year 2015
- A ranked list for the count of successful landing outcomes between the dates 04-06-2010 and 20-03-2017, in descending order

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/SpaceX%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Exploratory Data Analysis is performed by adding the following features to a folium map:

- Circle markers of the launch sites with popup labels and text labels
- Colored markers of the launch outcomes for each launch site grouped in marker clusters
- Lines showing the distance between each launch site and various points of interest

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/SpaceX%20Data%20Visualization%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Exploratory Data Analysis is performed by creating a Plotly Dashboard with the following features:

- Dropdown list to select each launch site
- Pie charts showing the success rate of each site (or all sites)
- Slider to select a range for the payload mass
- Scatter plot showing the relation between payload mass and success rate for each site

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

Four classification models were built and tested: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and k-Nearest Neighbors

Predictve analysis is performed with these models using the following steps:

- Data preparation and standardization
- Split the data into training and test sets
- Apply the classification model
- Obtain the accuracy of each model
- Plot a confusion matrix for each model
- Compare results across all models

SOURCE: https://github.com/jableaney/IBM-Data-Science-Capstone-Project/blob/master/SpaceX%20Landing%20Prediction%20Models.ipynb

Results

- Exploratory data analysis results
  - Visualization
  - SQL
  - Folium Map
  - Dashboard
- Predictive analysis results
  - Predict landing outcomes

Section 2

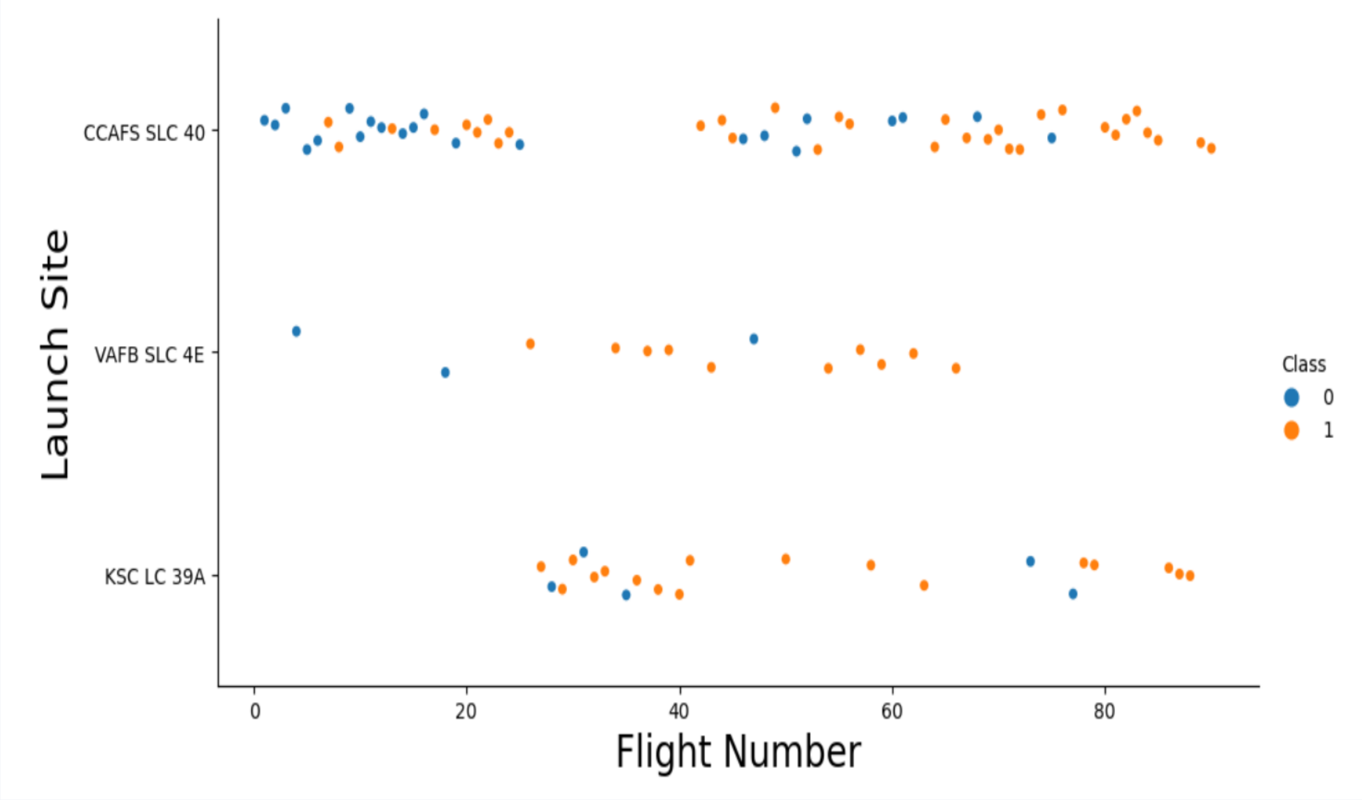
Insights drawn from EDA

Flight Number vs. Launch Site

Key observations:

- The site with the most launches is CCAFS SLC 40
- VAFB SLC 4E has been used for the fewest number of launches
- Success rate improves over time across all launch sites
- KSC LC 39 appears to have the highest success rate

Flight Number vs. Launch Site

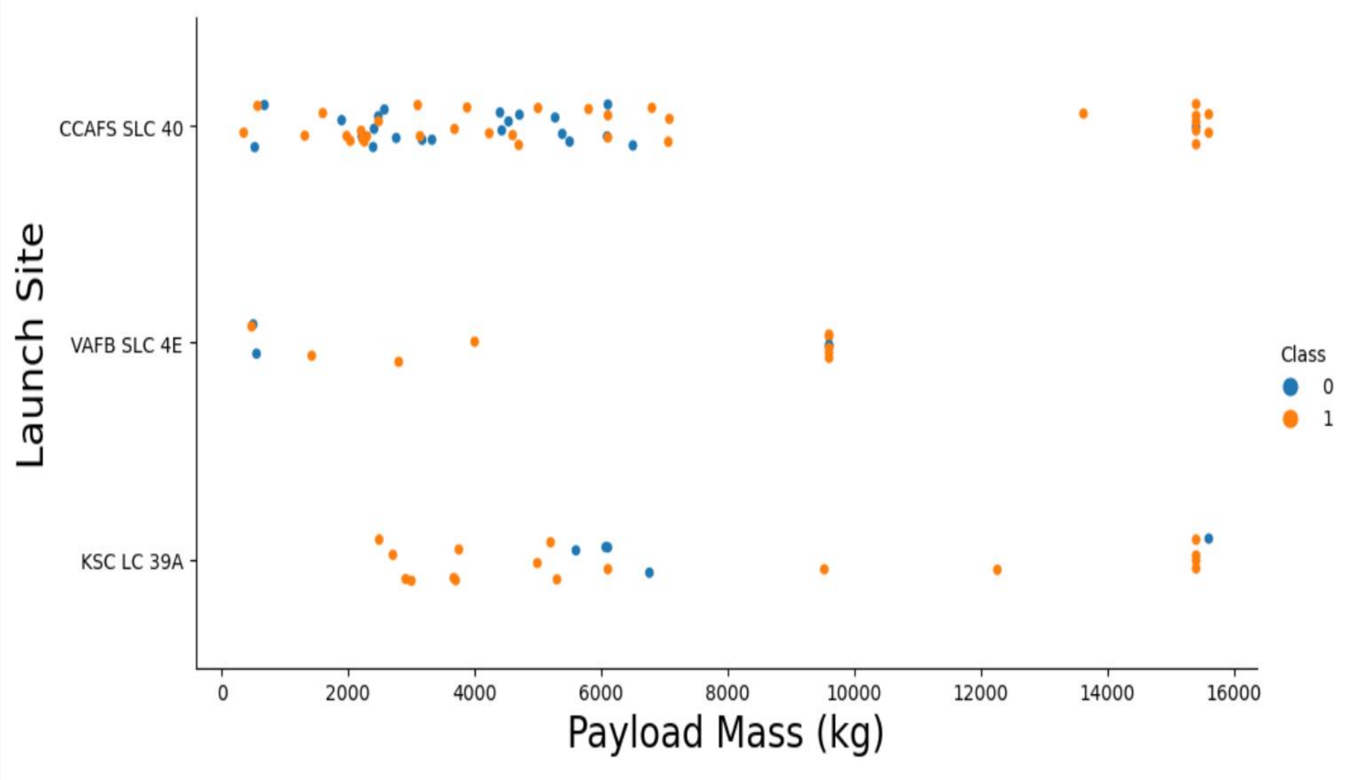


Payload vs. Launch Site

Key observations:

- Payloads over 10 000 kg then to have a higher success rate
- Launches with the maximum payload have primarily been launched from KSC LC 39A and CCAFS SLC 40
- KSC LC 39A has the highest success rate for launches with a payload under 6 000 kg

Payload Mass (kg) vs. Launch Site

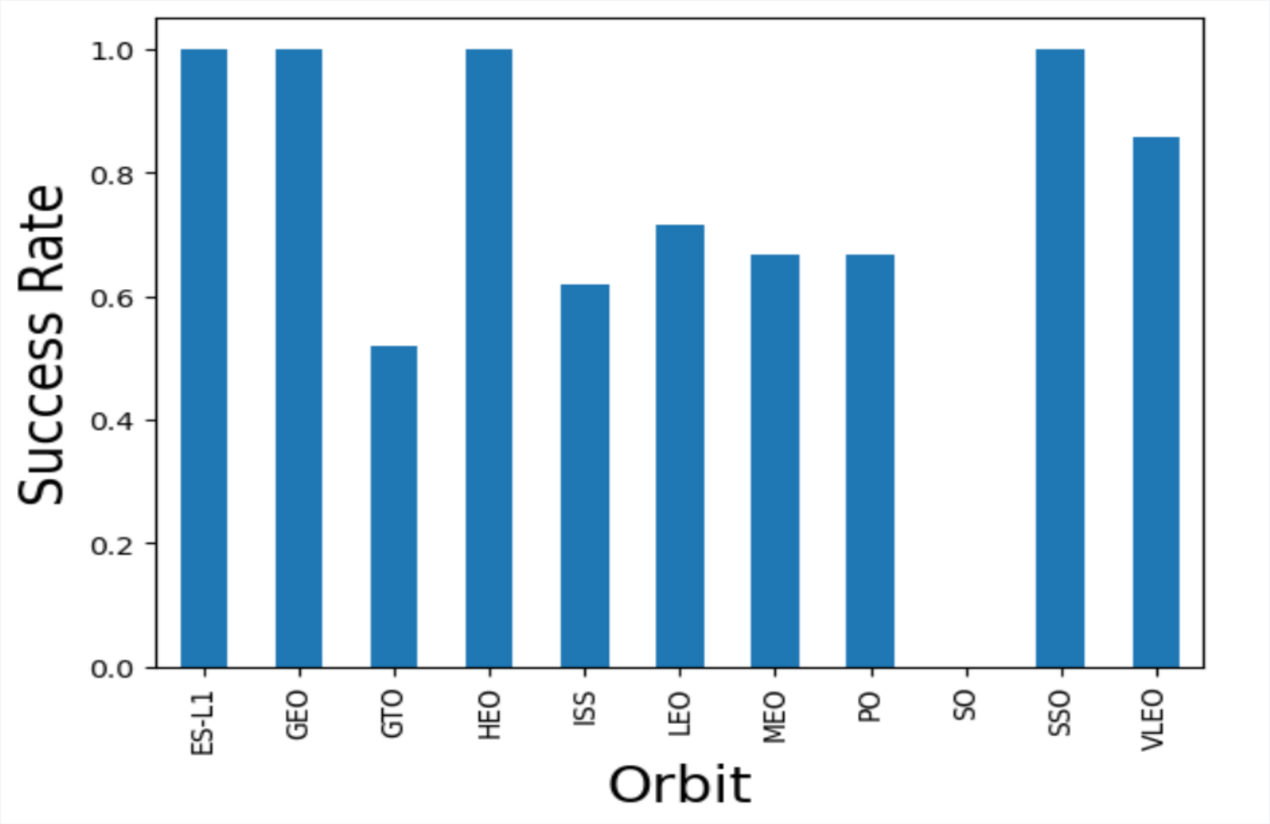


Success Rate vs. Orbit Type

Key observations:

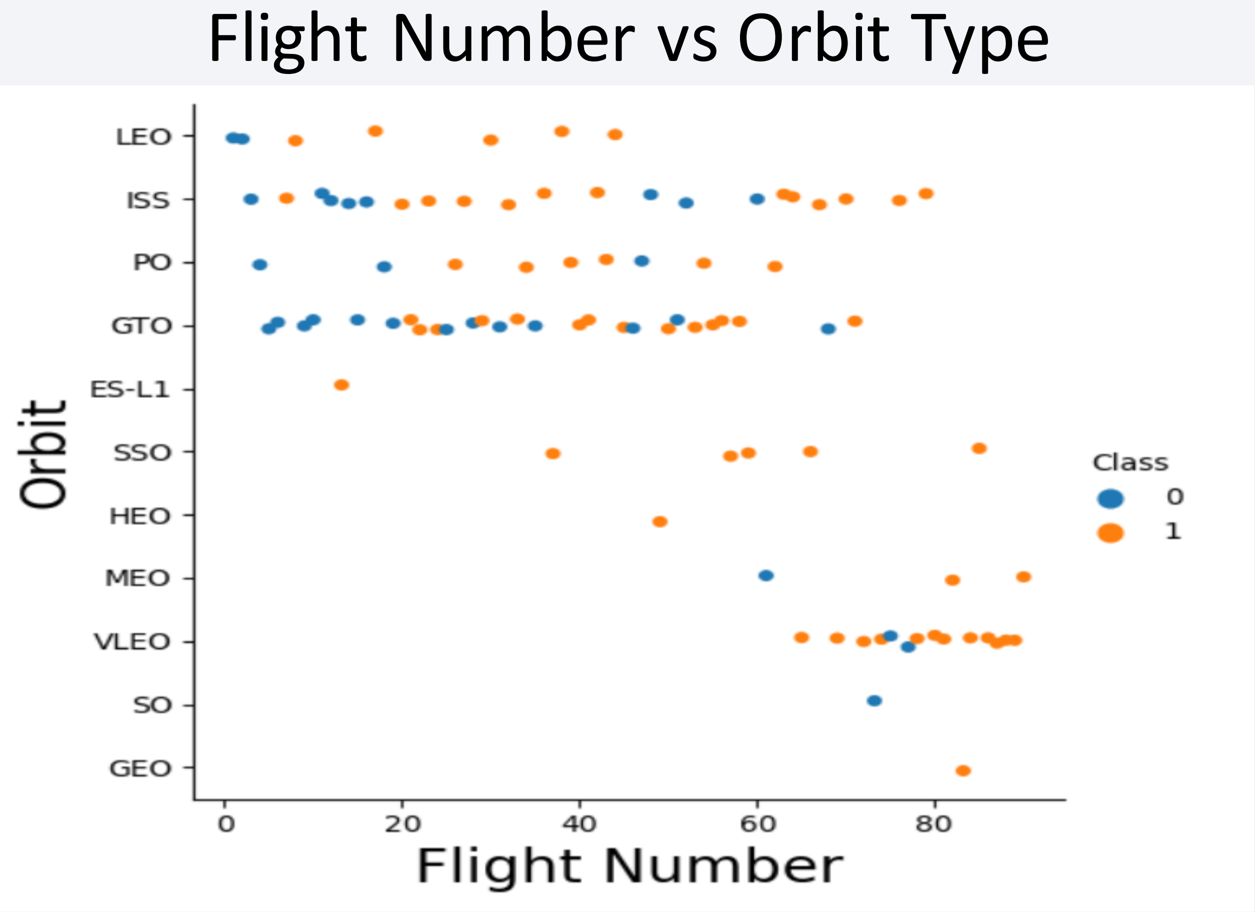
- Rockets launched into ES-L1, GEO, HEO and SSO have a 100% success rate
- There have been no successful launches with a SO destination
- GTO, ISS, LEO, MEO, PO have success rates between 50% and 70%

Success Rate Per Orbit Type



Flight Number vs. Orbit Type

- Key observations
  - Most launches have had either an ISS, GTO, or VLEO destination
  - There have only been one launch for each of the ES-L1, HEO, SO, or GEO destinations
  - LEO has had a 100% success rate after the first two launches

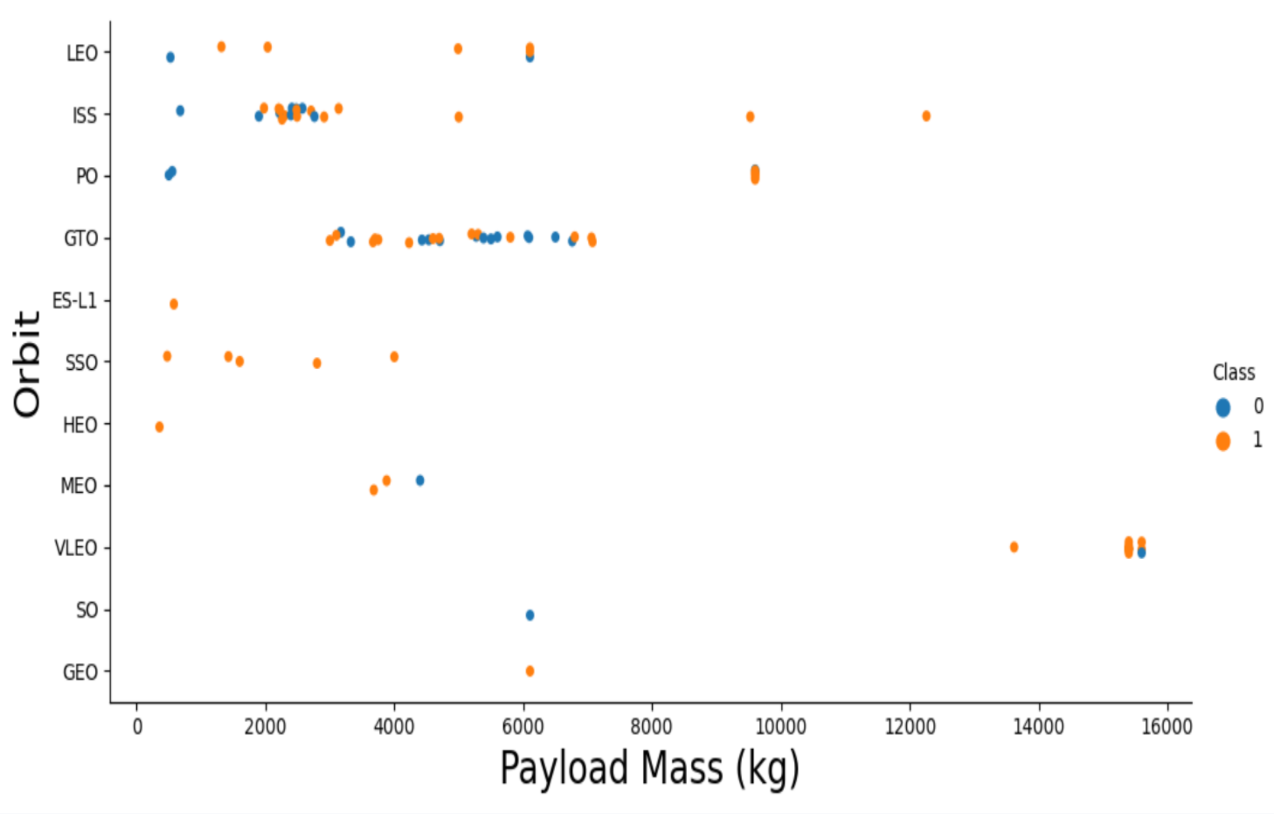


Payload vs. Orbit Type

Key observations:

- All payloads in excess of 13 000 kg have been destined for a VLEO orbit
- There have been a large number of launches destined for a GTO orbit, with payloads entirely between 3 000 kg and 8 000 kg
- ISS has the widest range of payload mass

Payload Mass (kg) vs Orbit Type

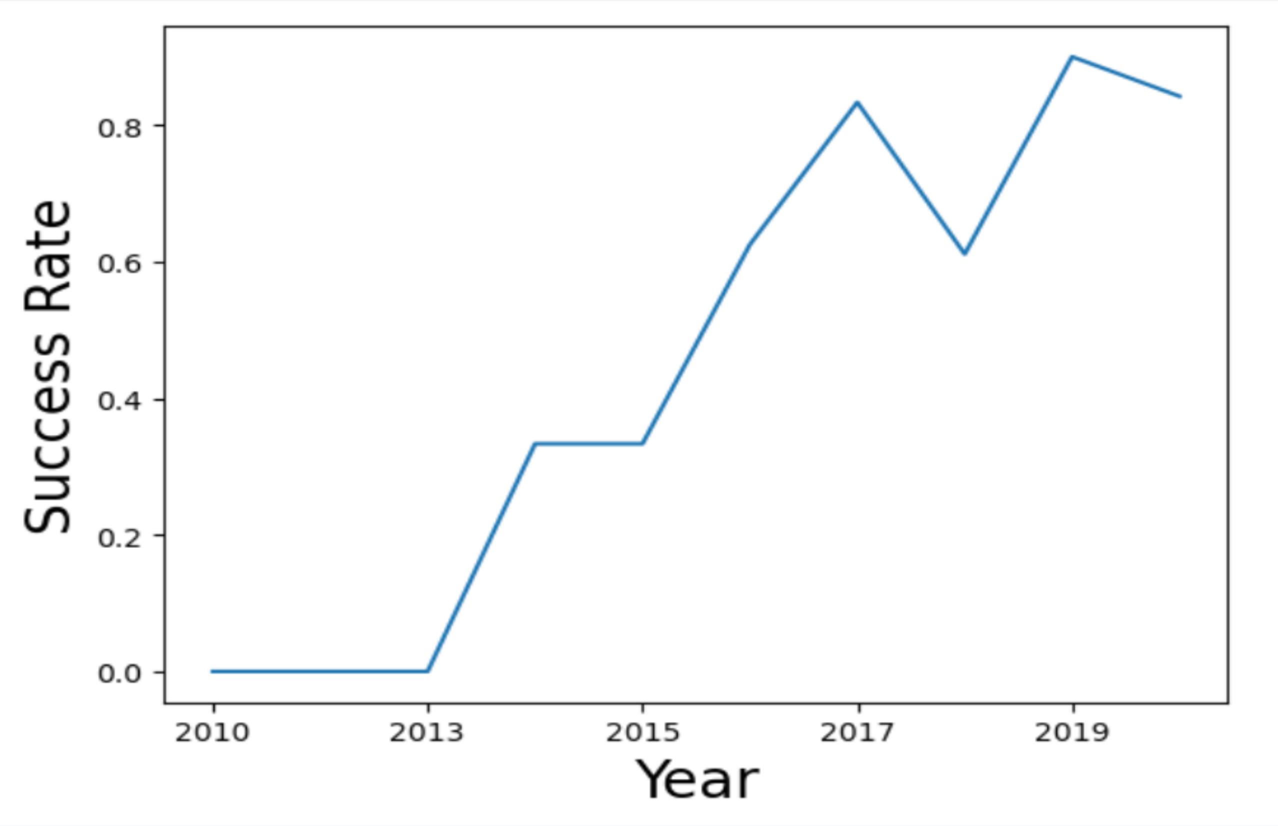


Launch Success Yearly Trend

Key observations:

- Success rate continually increased from 2013 to 2017
- The highest success rate was achieved in 2019
- 2018 and 2020 showed a small reduction in success rate compared to the previous year

Success Rate Per Year (2010 – 2020)



All Launch Site Names

Distinct Launch Site Names

In [12]:

```
%%sql

select distinct("Launch_Site")
from SPACEXTBL
```

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

Out[12]:

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

Launch Site Names Begin with 'CCA'

5 Records for Launch Sites Beginning with ‘CCA’

In [15]:

```
%%sql

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

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

Out[15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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 Launched For NASA

```
%%sql

select sum(PAYLOAD_MASS__KG_) as "Total_Payload_Mass"
from SPACEXTBL
where Customer = 'NASA (CRS)'
```

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

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

Average Payload Carried by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%%sql

select avg(PAYLOAD_MASS__KG_) as "Average_Payload_Mass"
from SPACEXTBL
where Booster_Version like "F9 v1.1%"
```

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

Average_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

First Successful Ground Landing

```
%%sql

SELECT min(Date)
from SPACEXTBL
where "Landing _Outcome" = 'Success (ground pad)'
```

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

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster with Successful Drone Ship Landings for Payloads Between 4 000 and 6 000 kg

```
%%sql

select distinct(Booster_Version)
from SPACEXTBL
where PAYLOAD_MASS__KG_ between 4000 and 6000 and "Landing _Outcome" = 'Success (drone ship)'
```

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

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

Total Number of Successful and Failure Mission Outcomes

Count of Each Mission Outcome Type

```
%%sql

select Mission_Outcome, count(*) as "Count"
from SPACEXTBL
group by Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

Boosters Used For Max Payload

```
%%sql

select distinct(Booster_Version), PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ in
    (select max(PAYLOAD_MASS__KG_)
    from SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Month of Failed Missions (2015)

```
%%sql

select substr(Date, 4,2) as Month, "Landing _Outcome", Booster_Version, Launch_Site
from SPACEXTBL
where substr(Date, 7,4) = '2015' and "Landing _Outcome" = 'Failure (drone ship)'
```

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

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

Count of Successful Landings (2010-06-04 to 2017-03-20)

```
%%sql

select "Landing _Outcome", count("Landing _Outcome") as "Count"
from SPACEXTBL
where Date between '04-06-2010' and '20-03-2017' and "Landing _Outcome" like 'Success%'
group by "Landing _Outcome"
order by "Count" desc
```

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

Landing _Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

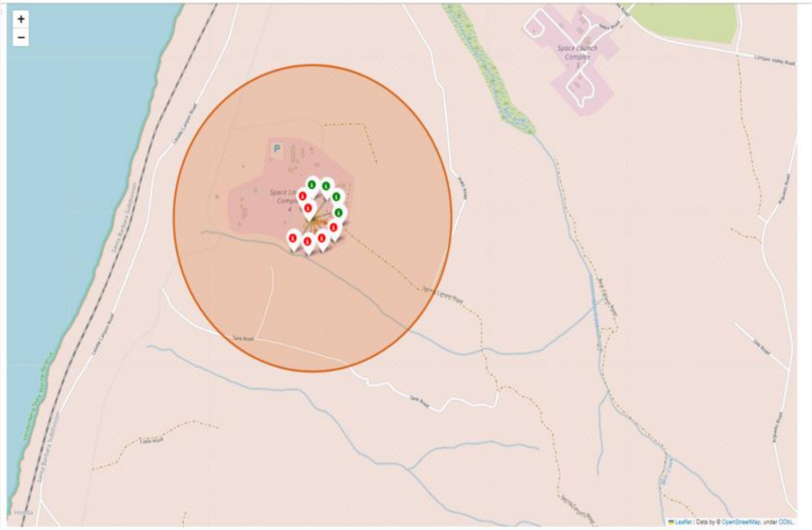
Launch Sites Proximities Analysis

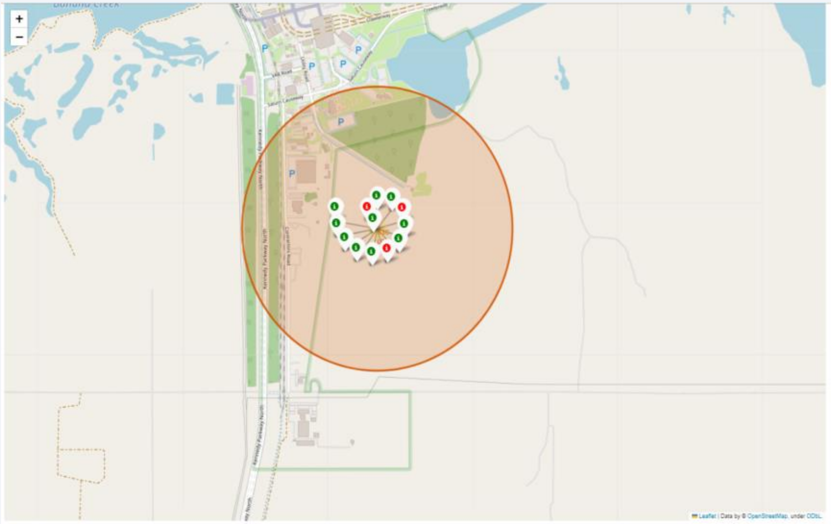
All Launch Sites – Folium Map

```python
# Initial the map
site_map = folium.Map(location=nasa_coordinate, zoom_start=5)

# For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name as a popup label

for index, site in launch_sites_df.iterrows():
    circle = folium.Circle((site['Lat'],site['Long']), radius=1000, color='#d35400', fill=True).add_child(folium.Popup(site['Launch Site']))
    marker = folium.map.Marker(
        (site['Lat'],site['Long']),
         icon=DivIcon(
             icon_size=(20,20),
             icon_anchor=(0,0),
             html='<div style="font-size: 12; color:#d35400;"><b>%s</b></div>' % site['Launch Site'],
         ))

    site_map.add_child(circle)
    site_map.add_child(marker)

site_map
```

Launch Outcomes – Folium Map

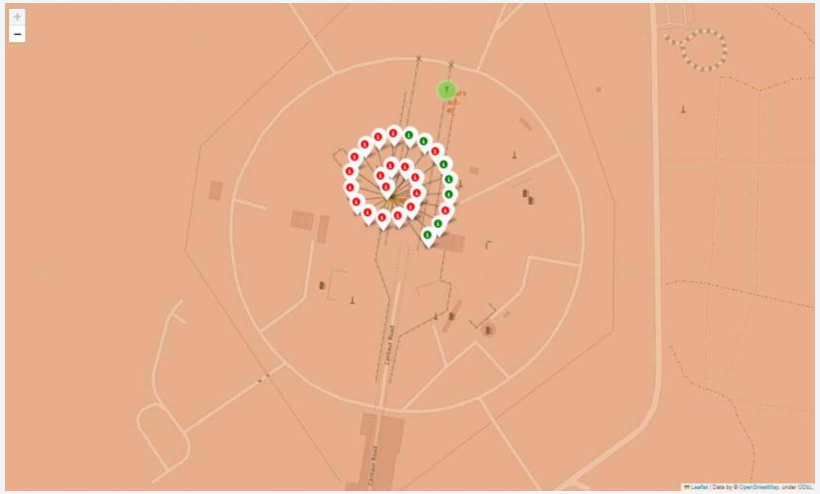
VAFB SLC-4E

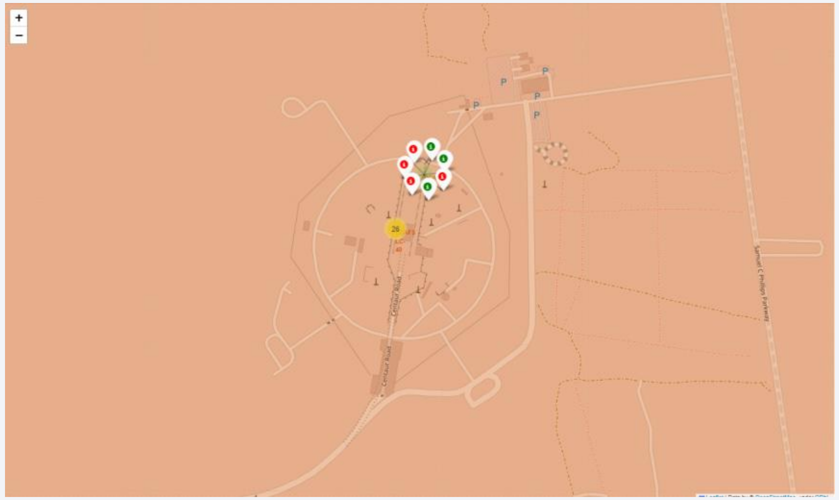




KSC LC-39

CCAFS LC-40



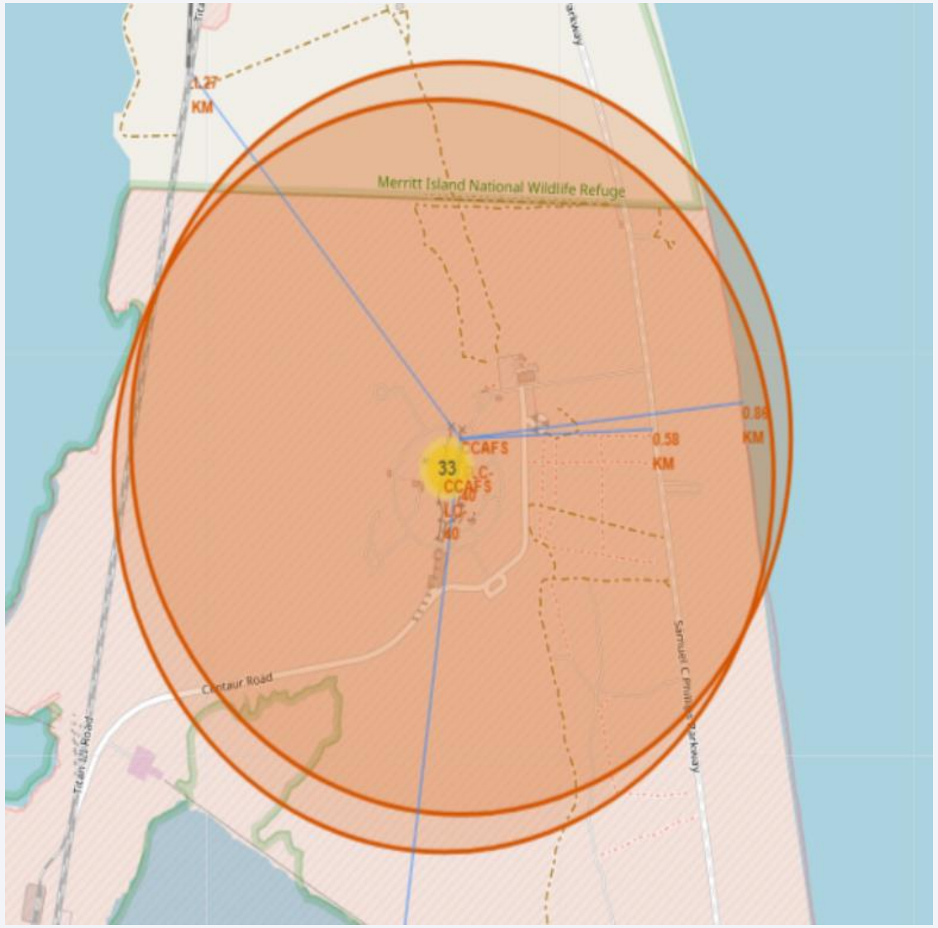


CCAFS SLC-40

Proximity to Points of Interest – Folium Map

Distances

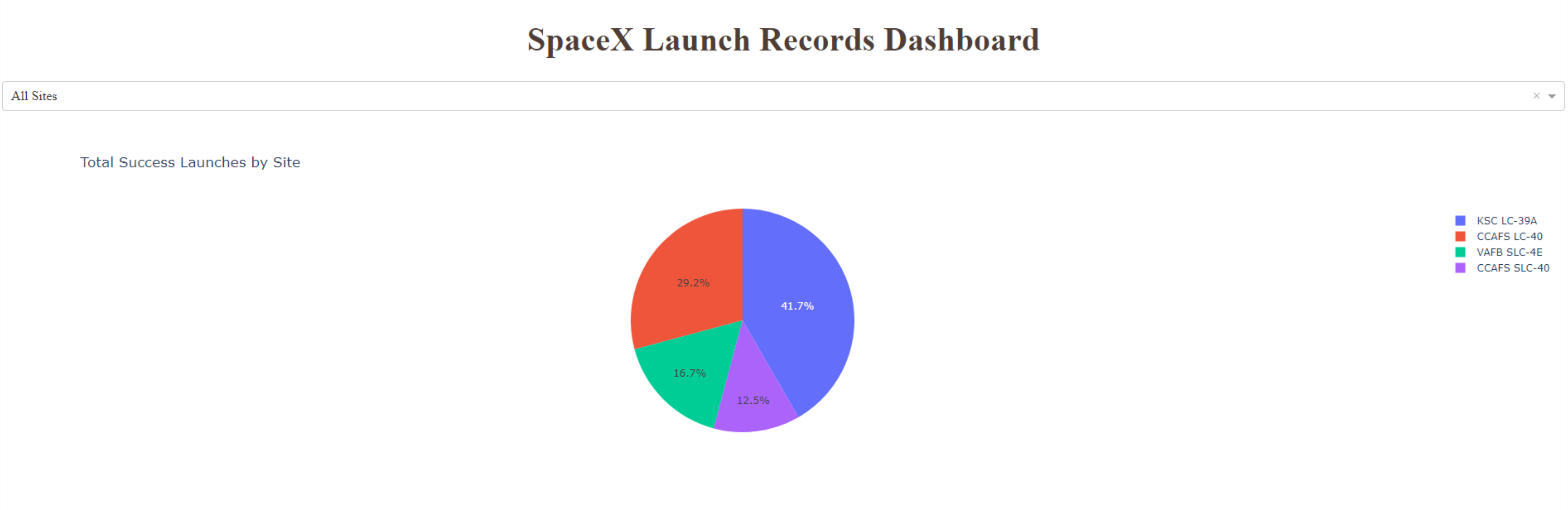
- Coast: 0.86 km
- Highway: 0.58 km
- Railway: 1.27 km
- City (Melbourne): 52.64 km



Section 4

Build a Dashboard with Plotly Dash

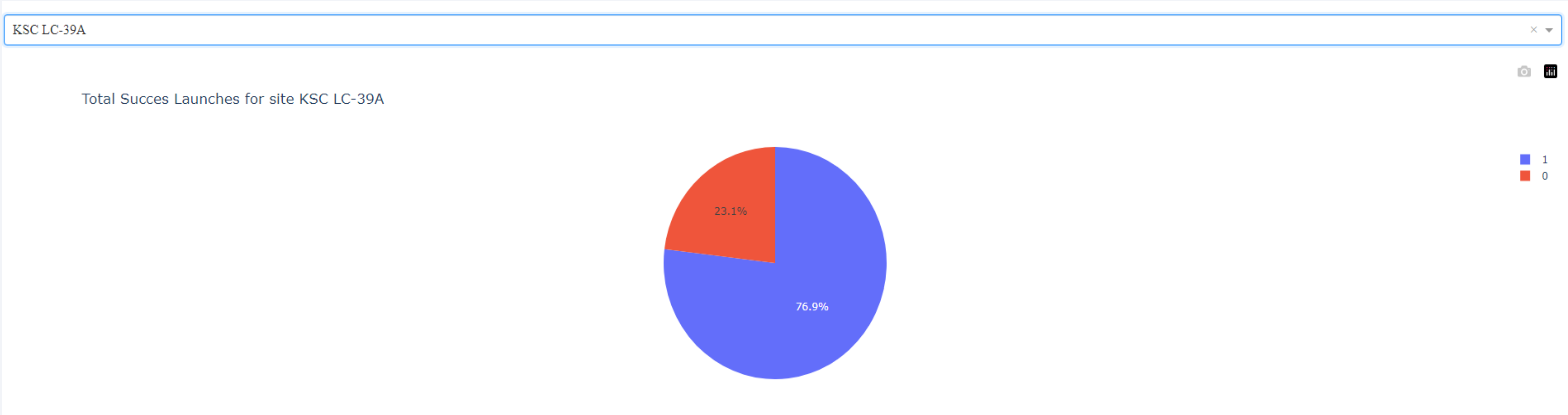
Plotly Dash – Successful Launches (All Sites)



Observations:

- KSC LC-39A has the most successful launches
- CCAFS SLC-40 has the fewest successful launches

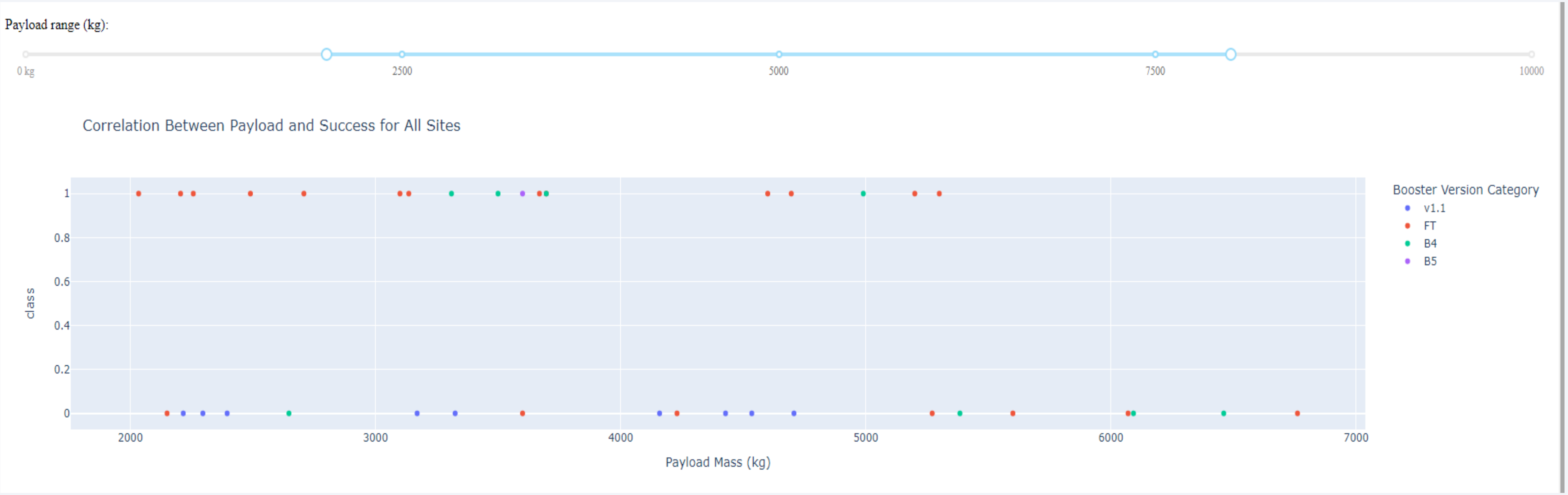
Plotly Dash – Success Ratio (KSC LC-39A)



Observations:

- Launch site KSC LC-39A has a 76.9% success rate

Plotly Dash – Payload vs Outcome (All Sites)



Observations:

- There have been no successful launches with payloads between 5,500 kg and 8,000 kg

Section 5

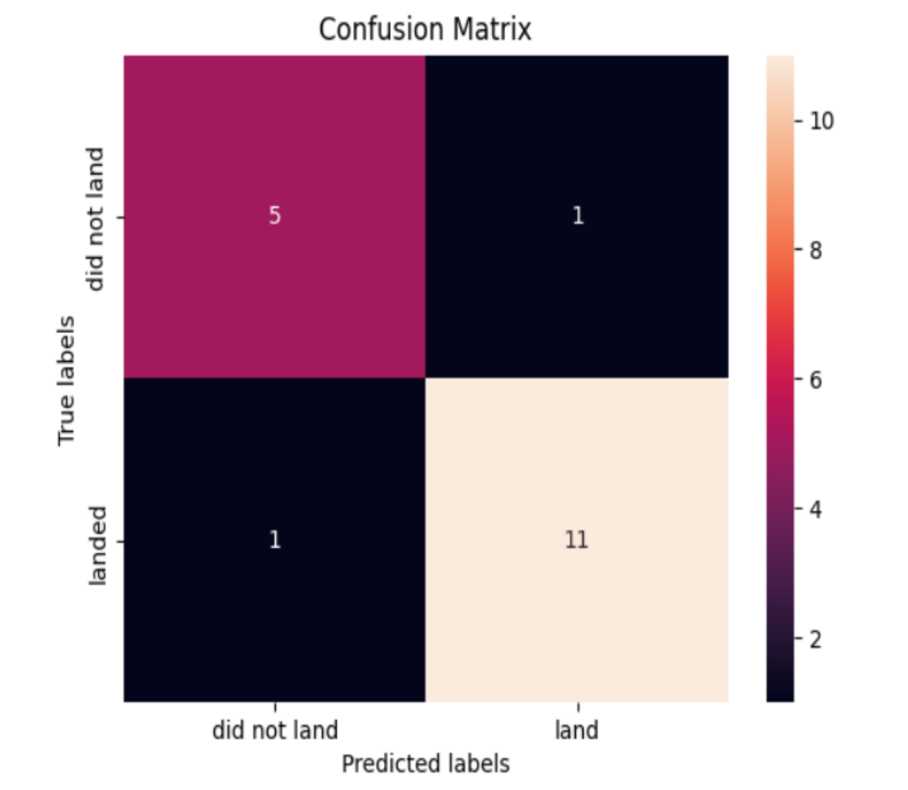
Predictive Analysis (Classification)

Confusion Matrix

Observations:

- Decision tree model yielded a single false positive and 11 true positives from the test set
- Decision tree model yielded a single false negative, and 5 true negatives from the test set

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- The site with the most successful landing is KSC LC-39A
- No good estimations can be made with the provided information for payloads greater than 6 000 kg
- Success rate for all types of orbits improved over time
- Decision tree model provide the optimal classification model

Thank you!