IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

S.C.T.K
22 02 2022

Outline

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

Executive Summary

- Methodologies used
  - Data Collection: API, Web-Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA): SQL, Python
  - Interactive Visual Analysis (IVA): Folium, Plotly
  - Prediction: Machine Learning

- Results produced
  - EDA Results
  - IVA Results
  - Prediction Model

Introduction 1/2

- Background
  - Spacex prices Rocket Launches at a 100 million dollars less than competition
  - Savings come from reusing the expensive stage1 Rockets.
  - Unfortunately, not all Stage 1 Rockets return Safely
- Opportunity
  - Ability to predict IF a Stage 1 Rocket will land Safely
  - Enables competitors to profitably bid alongside Spacex

Introduction 2/2

- Project Output:
  - Machine Learning Model that predicts IF Stage 1 will land or not

- Project Stepping Stones
  - Identify Factors that influence successful Landing
  - Analyze interactions between factors and Success Rate

Section 1

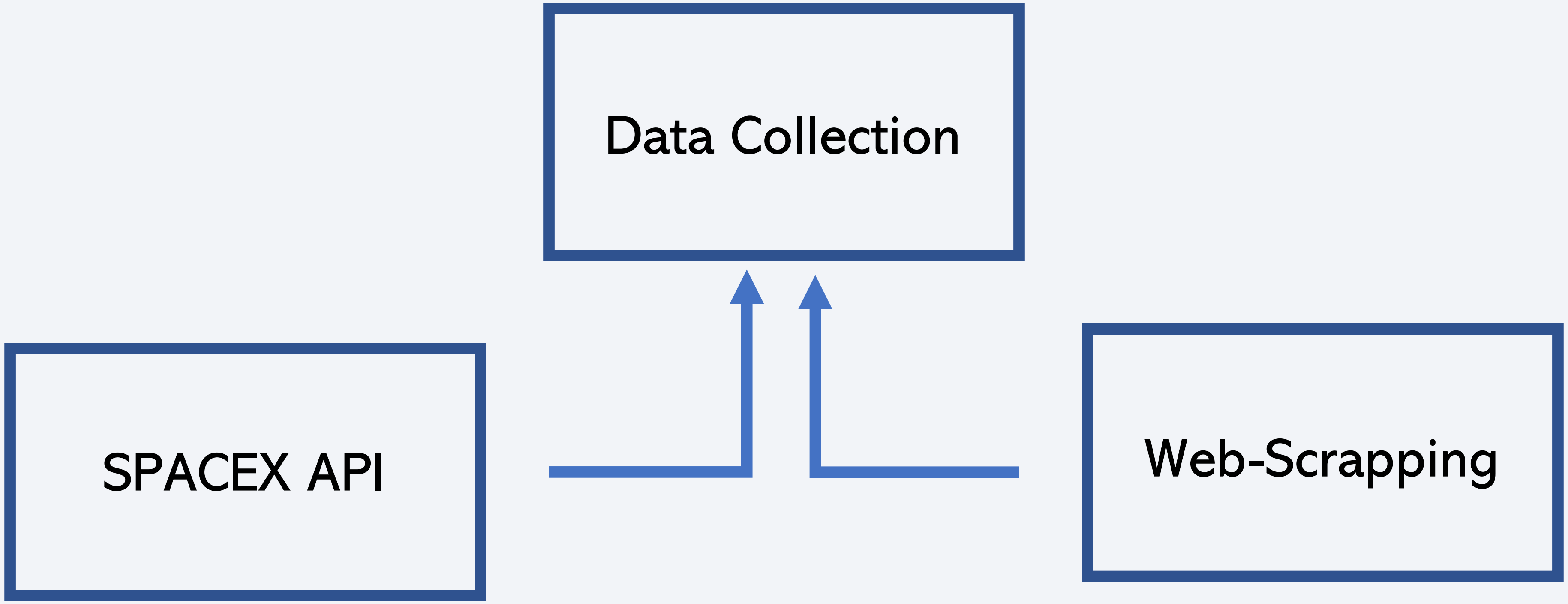
Methodology

Methodology

Executive Summary

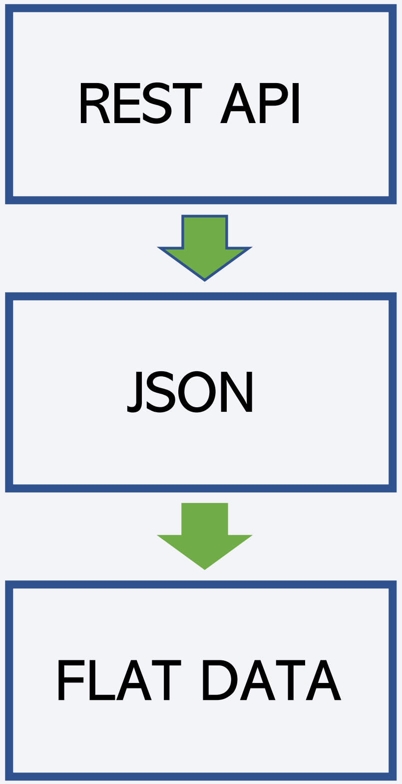
- Data collection methodology:
  - SpaceX API and Web Scraping from Wikipedia
- Perform data wrangling
  - Data Formatted, One Hot Encoded
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained Model on Training Data, Tested Model on Tested Data, Best Model Chosen.

Data Collection



Data Collection was performed in 2 ways.

Data Collection — Spacex API



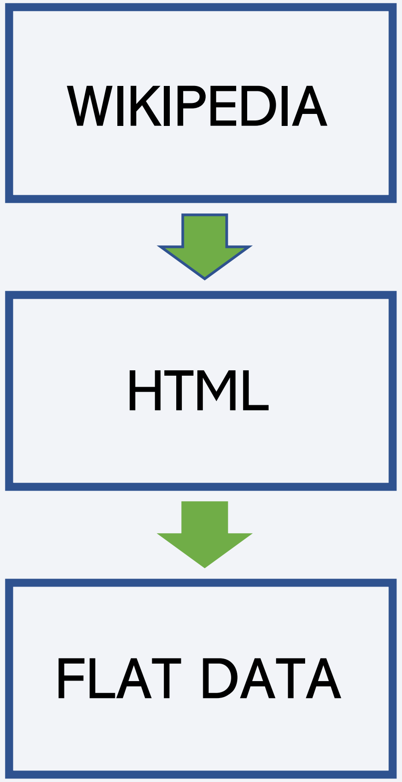

```python
[6]: spacex_url="https://api.spacexdata.com/v4/launches/past"

[7]: response = requests.get(spacex_url)
```

```python
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

[Python notebook: Data Collection using Spacex API]

Data Collection

Web-Scraping




```
[4]: static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

```
[5]: # use requests.get() method with the provided static_url
     # assign the response to a object
     response  = requests.get(static_url)
     print('Status of HTTP GET request : ',response.status_code)
     html_data  = response.text

     Status of HTTP GET request :  200
```

```
[6]: # Use BeautifulSoup() to create a BeautifulSoup object from a response text content
     soup = BeautifulSoup(html_data, 'html5lib')
```

```
[8]: # Use the find_all function in the BeautifulSoup object, with element type `table`
     # Assign the result to a list called `html_tables`
     html_tables = soup.find_all("table")
```

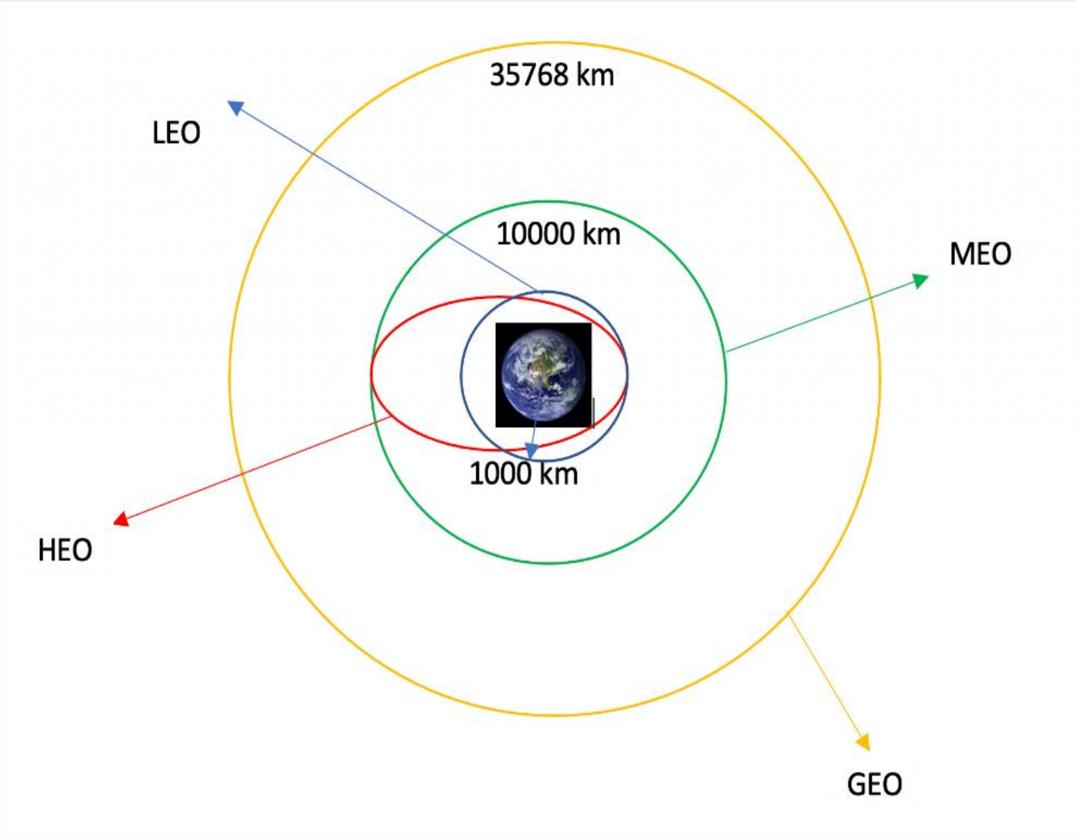
Starting from the third table is our target table contains the actual launch records.

```
[9]: # Let's print the third table and check its content
     first_launch_table = html_tables[2]
```

[Python notebook: Data Collection using Web-Scraping]

Data Wrangling

- Data Wrangling
  - Exploratory Data Analysis
    - Launches from each Site
    - Occurrence of each Orbit
    - Mission Outcome per Orbit Type
  - Determine Training Labels
    - Create Label – ”Landing Outcome”

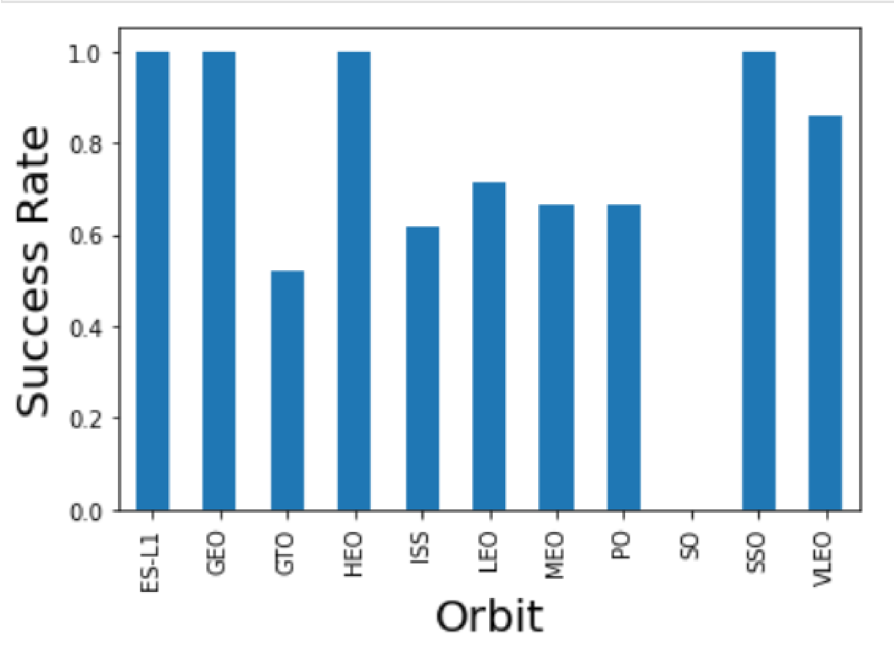


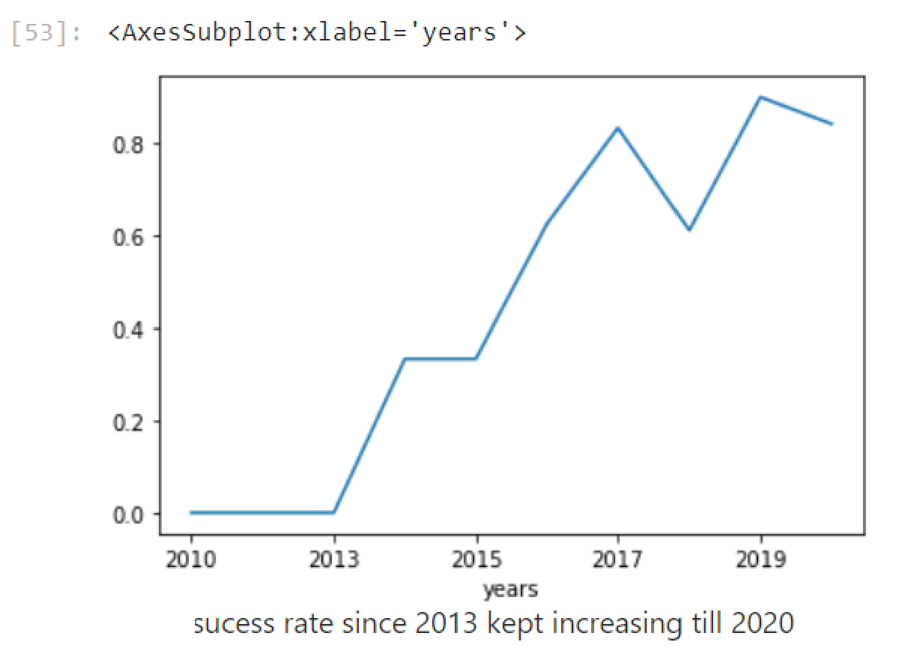
[Python notebook: Data Wrangling]()

EDA with Data Visualization

EDA (Data Vis)

- Flight number vs Launch site
- Payload vs Launch Site
- Success Rate vs orbit Type
- Flight number Vs Orbit type
- Payload vs Orbit Type
- Success Rate vs Year





[Python notebook: EDA DataViz]

EDA with SQL

EDA (SQL)

- List Unique Launch Sites
- 5 Launch Sites beginning with CCA
- Total Payload carried by NASA(CRS)
- Drone Ship Landings with Pyload between 4000, 6000
- Data of First successful Landing on ground pad
- Names of Boosters that carried Max Payload
- Rank Landing Outcome by occurrence

[49]:

booster_version	landing_outcome	payload_mass
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

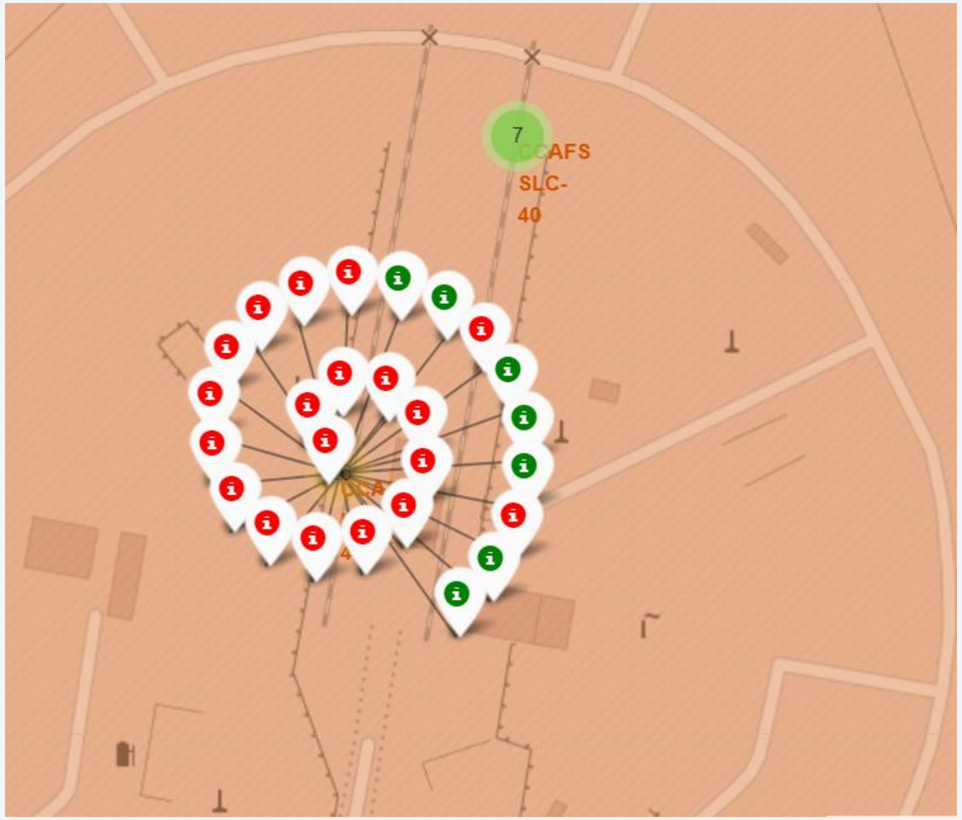
[53]:

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

[Python notebook: EDA (SQL)]

Build an Interactive Map with Folium

- Launch sites were marked and labelled on the map of USA
- Launch Outcomes (binary – Success, Failure) was set to Red and Green markers
- Number of Successes and Failures were marked for each Launch Site.
- Distances were calculated between launch site and Highways, Railways, Cities and Coastline.

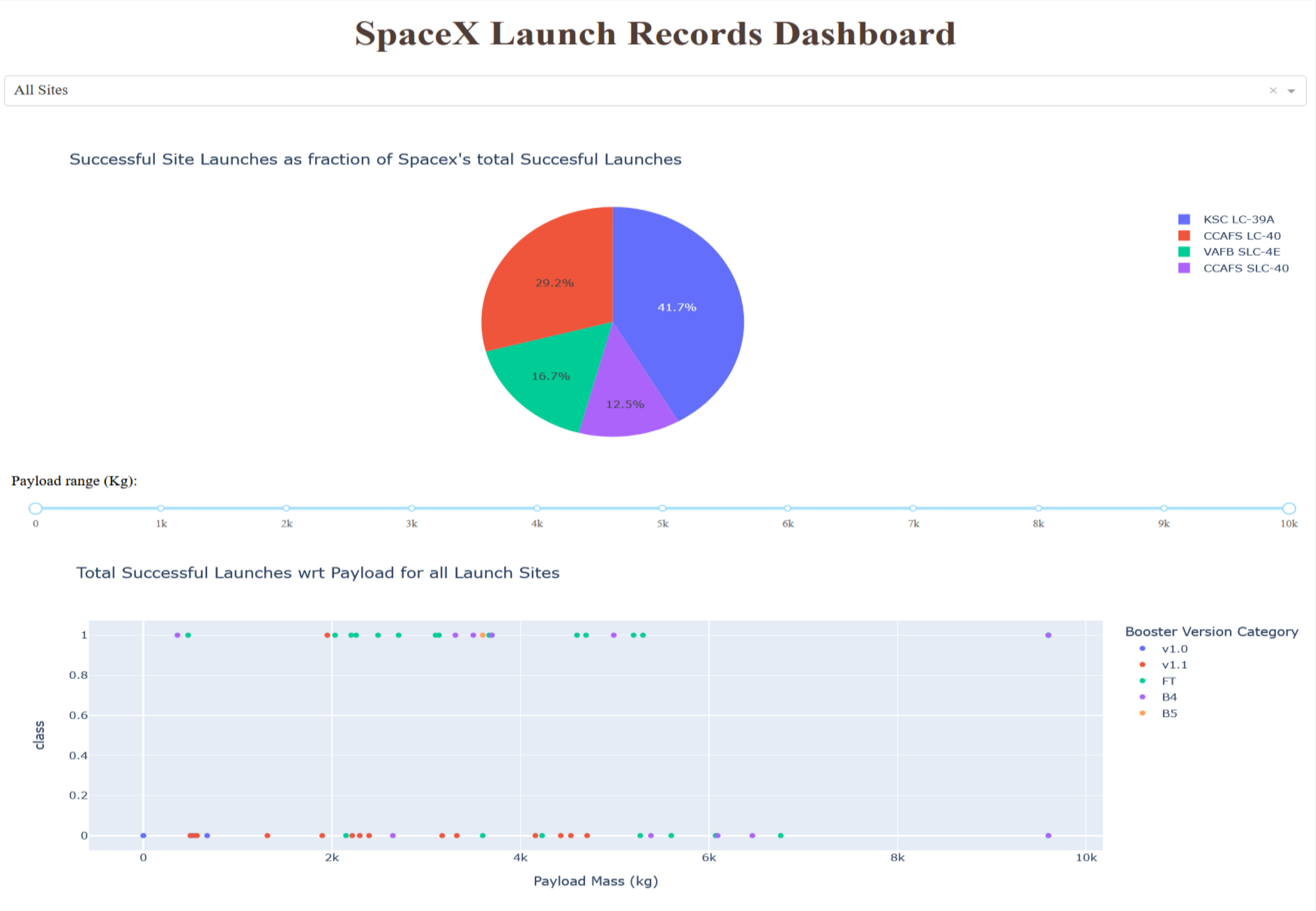


[Python notebook: Interactive Folium Map]

Build a Dashboard with Plotly Dash

- Pie Chart showing total launches per Site.
- Scatter plot of Landing outcome for different Boosters with a Payload Mass Slider

[Python notebook: Interactive Dashboard.]



Predictive Analysis (Classification)

- Processes used in Building Model
  - Clean and Normalize Data into Panda
  - One Hot Encode Categorical Fields
  - Split Data in Test and Train Set.
  - Run the data through various Machine Learning Algorithms.
- Processes used in Evaluating Model
  - Evaluate each model on accuracy of predictions on test set
  - Tune Hyperparameters for each modelYou need present your model development process using key phrases and flowchart
- Process used in Improving Model
  - Feature engineering

[Python notebook: Predicitive Analysis]

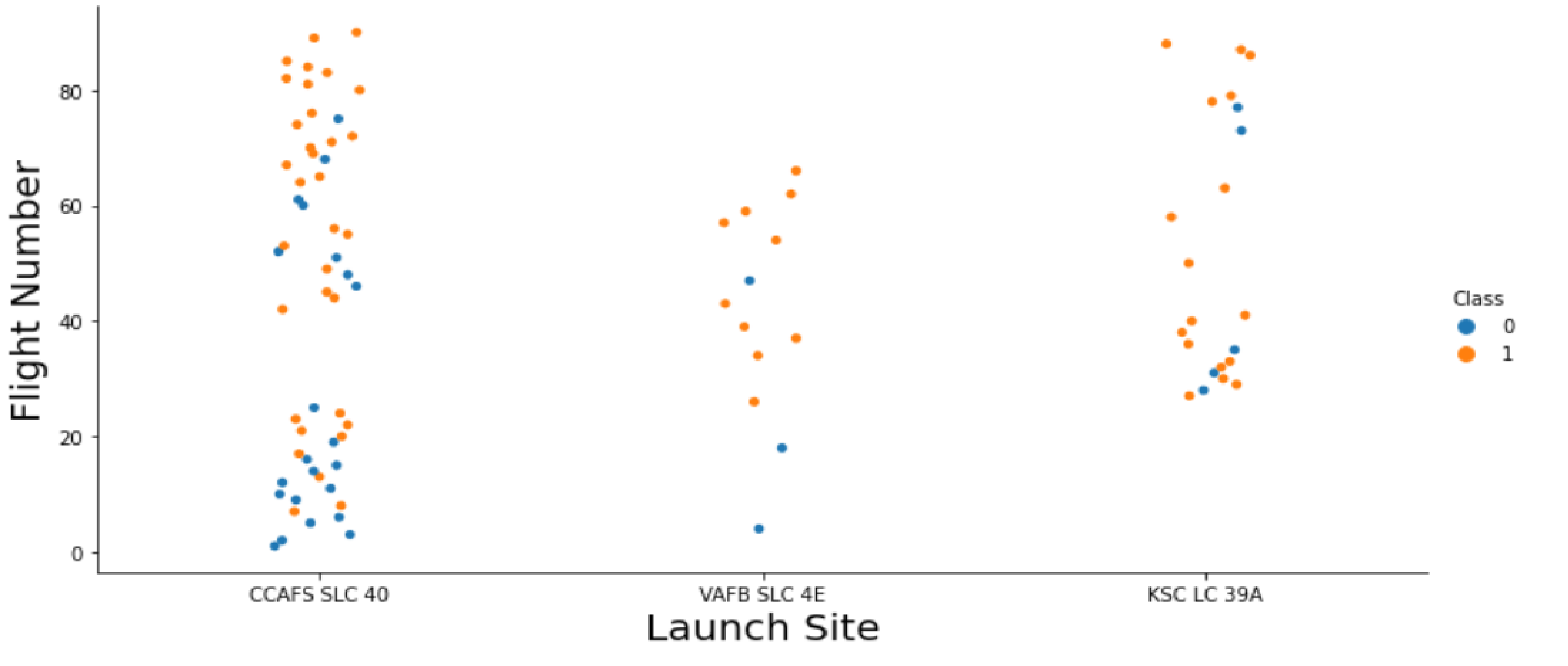
Results

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

Section 2

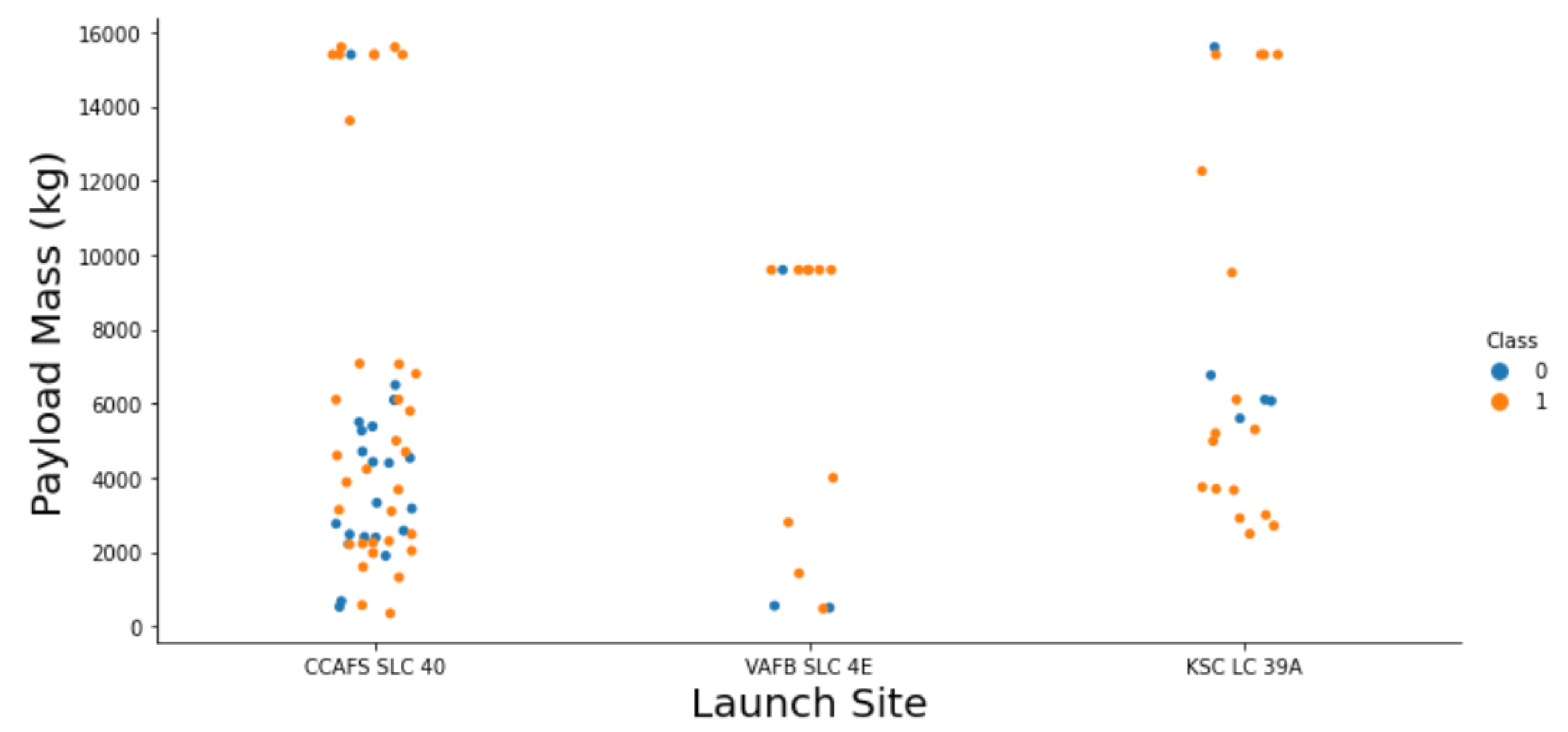
Insights drawn from EDA

Flight Number vs. Launch Site (Data Vis)



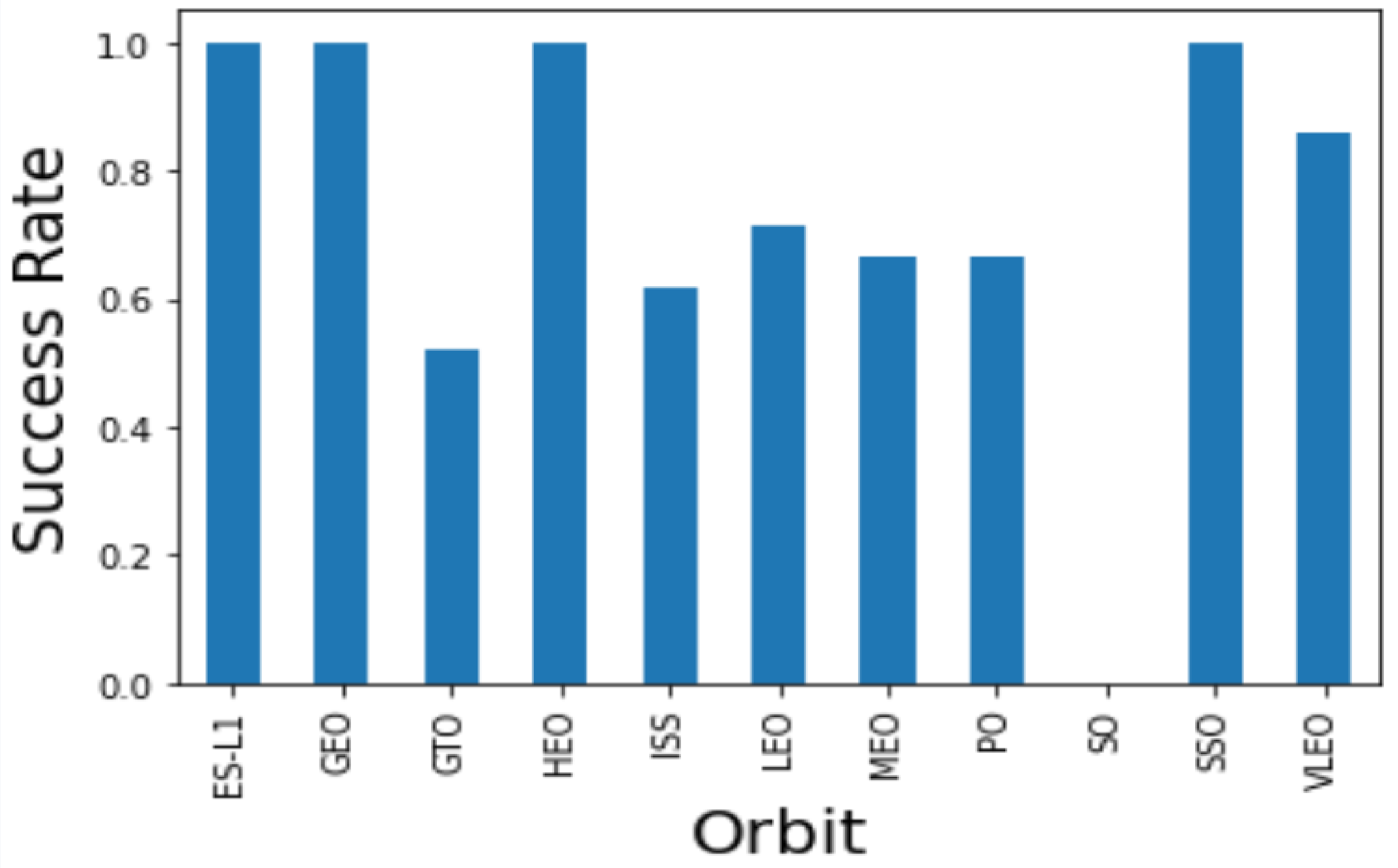
Lower Flight Numbers have higher failure rate on all sites

Payload vs. Launch Site (Data Vis)



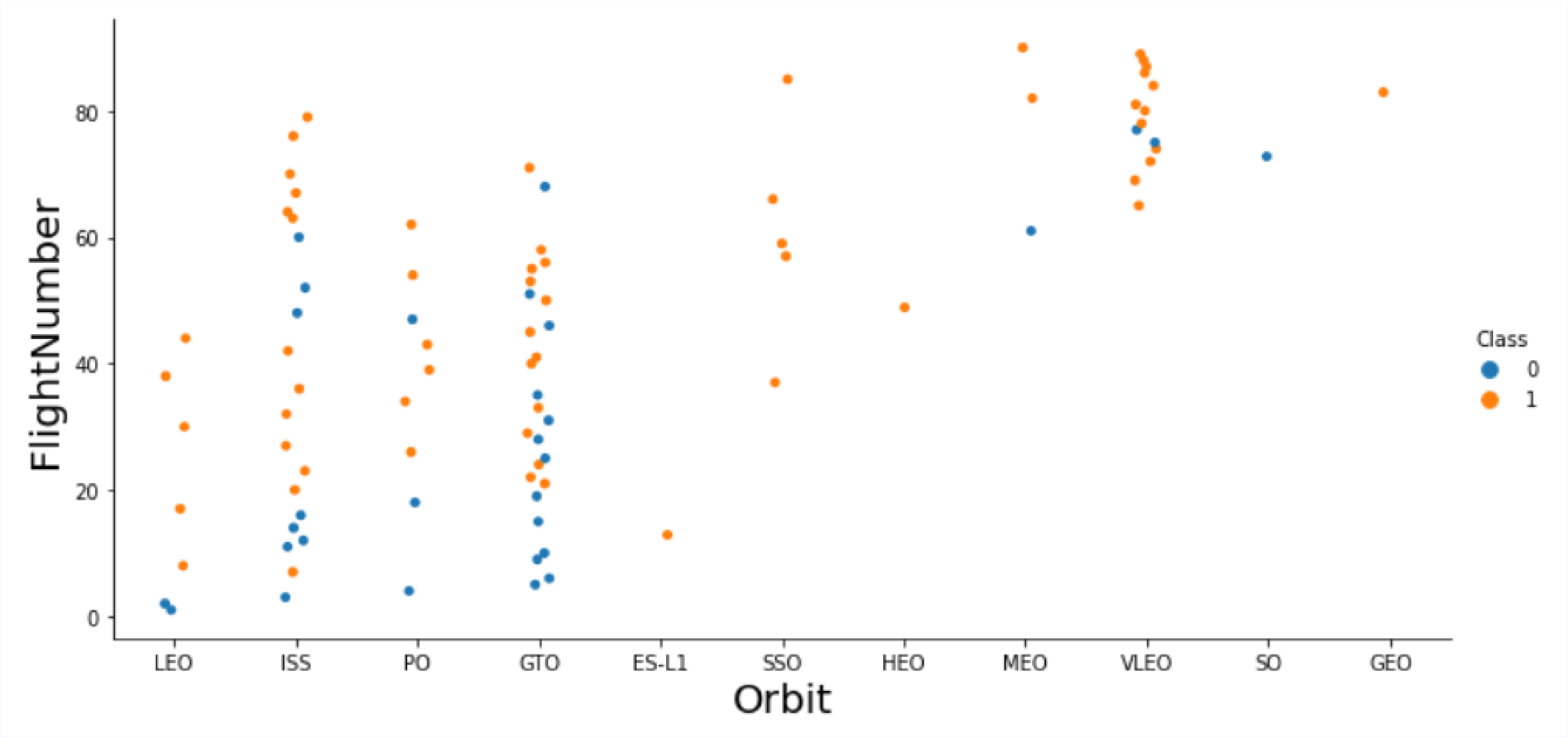
Higher Payloads have slightly lower Failure rate, especially at CCAFS SLC 40.

Success Rate vs. Orbit Type (Data Vis)



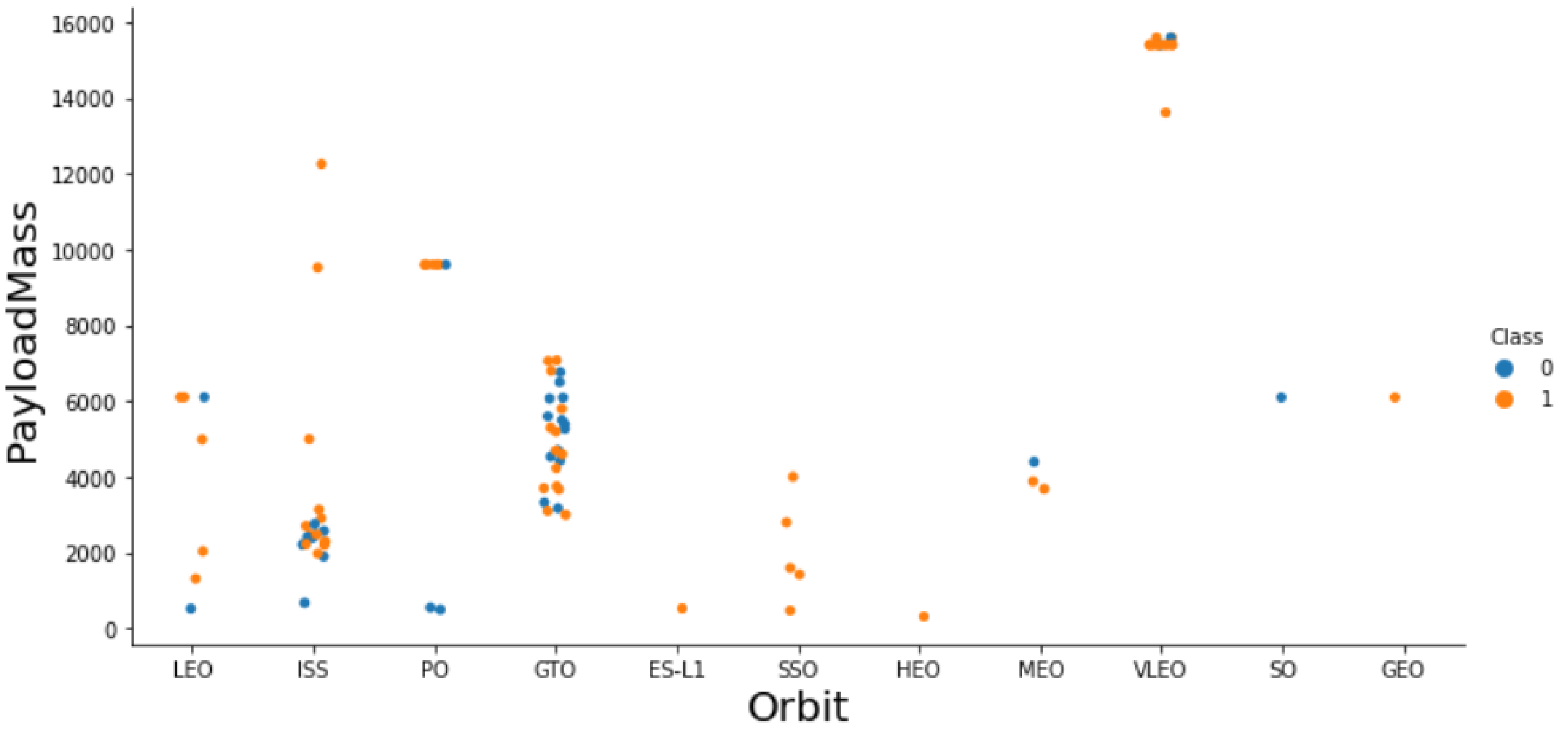
- ESL1, GEO, HEO & SSO have the highest Success Rate

Flight Number vs. Orbit Type (Data Vis)



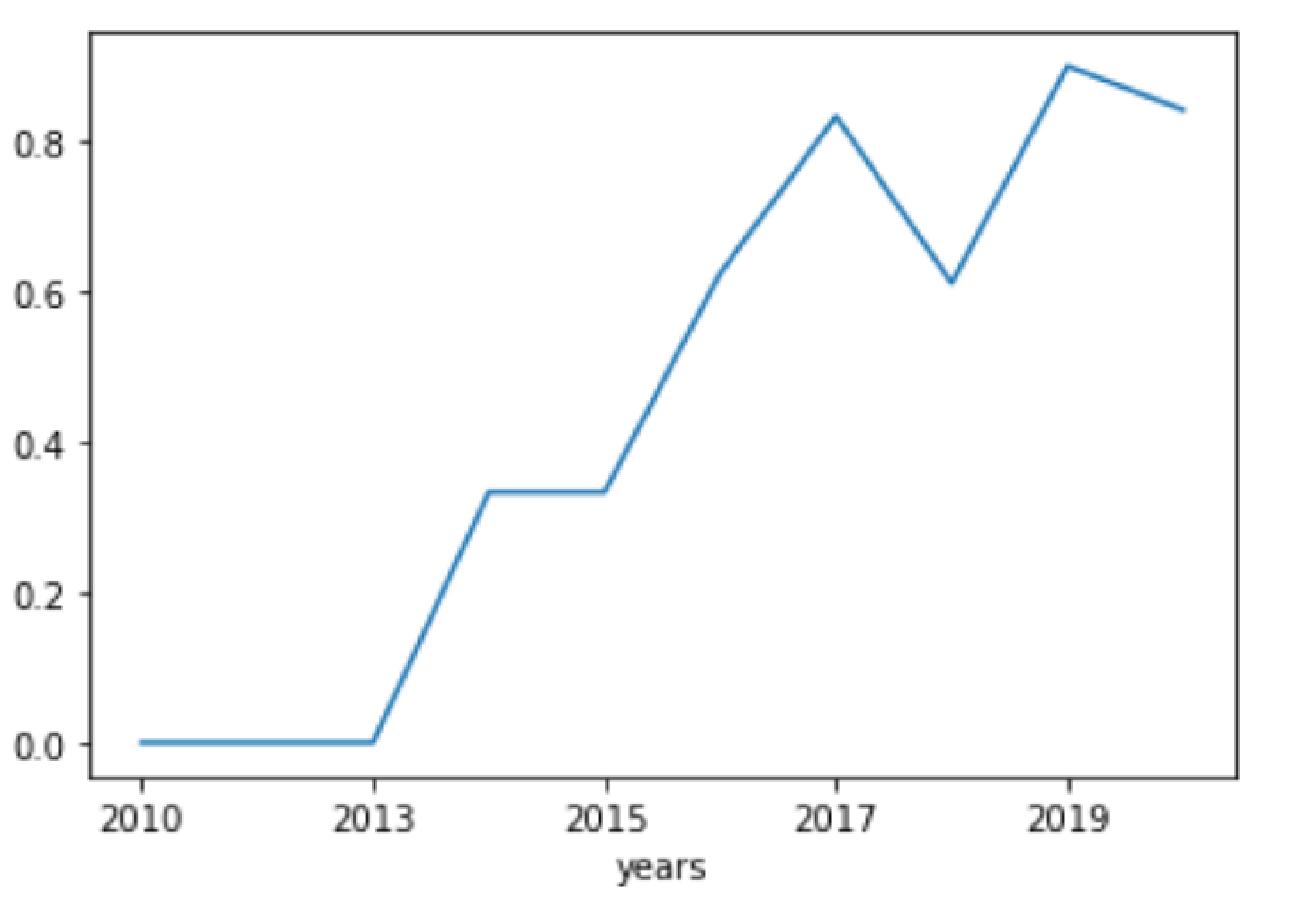
Correlation b/w Flight Number and Orbit is strong for Launch Site LEO, weak for Launch Site GTO

Payload vs. Orbit Type (Data Vis)



Correlation b/w PayloadMass and Orbit is Strong for Launch Site Leo, Weak for GTO

Launch Success Yearly Trend (Data Vis)



Success Rate has generally risen since 2013

All Launch Site Names (SQL)

Display the names of the unique launch sites in the space mission

[44]:

```
%%sql

SELECT DISTINCT Launch_Site from SPACEXTBL;
```

[44]:

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

Query Explanation

Keyword DISTINCT used to list only Unique Values in column launch_site

Launch Site Names Begin with 'CCA' (SQL)

Display 5 records where launch sites begin with the string 'CCA'

[45]:
```
%%sql

SELECT * from SPACEXTBL
    WHERE Launch_Site LIKE 'CCA%'
    LIMIT 5;
```

[45]:

DATE	time_utc	booster_version	launch_site	payload	payload_mass	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

Query Explanation

Keyword LIMIT used to list only 5 values where condition satisfied

Total Payload Mass (SQL)

Display the total payload mass carried by boosters launched by NASA (CRS)

[46]:
```
%%sql

SELECT SUM(Payload_Mass)
    from SPACEXTBL
    WHERE Customer = 'NASA (CRS)';
```

[46]:

1
45596

Query Explanation

Keyword SUM used to add up values in Payload_Mass where condition satisfied

Average Payload Mass by F9 v1.1 (SQL)

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

[47]:
```
%%sql

SELECT AVG(Payload_Mass)
    from SPACEXTBL
    WHERE Booster_Version LIKE 'F9 v1.1%';
```

[47]:

1
2534

Query Explanation

Keyword AVG used to return average of values in Payload_Mass column where condition satisfied

First Successful Ground Landing Date (SQL)

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

[48]:
```
%%sql

SELECT min(Date) as FirstDate
    from SPACEXTBL
    WHERE Landing_Outcome = 'Success (ground pad)';
```

[48]:

firstdate
2015-12-22

Query Explanation

Keyword MIN used to return Lowest Value of DATE which is the earliest date.

Successful Drone Ship Landing with Payload b/w 4000 & 6000 (SQL)

List the names of the boosters which have success in drone ship

and have payload mass greater than 4000 but less than 6000

[49]:
```
%%sql

SELECT Booster_Version, Landing_Outcome, Payload_Mass
    from SPACEXTBL
    WHERE Landing_Outcome LIKE '%Success (drone ship)%'
    AND Payload_Mass BETWEEN 4000 AND 6000;
```

[49]:

booster_version	landing_outcome	payload_mass
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Query Explanation

Keyword BETWEEN used to return values in between the two set limits and where conditions satisfied.

Total Number of Successful and Failure Mission Outcomes (SQL)

List the total number of successful and failure mission outcomes

[50]:
```
%%sql

SELECT Mission_Outcome, COUNT(Mission_Outcome) as Total
    from SPACEXTBL
    GROUP BY Mission_Outcome;
```

[50]:

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Query Explanation

Keyword COUNT used to return number of occurrences of unique values, Keyword Group by used to list values grouped by parameter passed, in this case Mission_Outcome.

Boosters Carried Maximum Payload (SQL)

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

[51]:
```
%%sql

SELECT DISTINCT Booster_Version, Payload_Mass
    from SPACEXTBL
    WHERE Payload_Mass=(SELECT MAX(Payload_Mass)
                        from SPACEXTBL);
```

[51]:

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

Query Explanation

Keyword MAX used to return maximum value of value in column Payload_Mass, for every DISTINCT (Unique value of) Booster Version.

2015 Launch Records (SQL)

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

[52]:
```
%%sql

SELECT Landing_Outcome, Booster_Version, Launch_Site
    from SPACEXTBL
    WHERE Landing_Outcome = 'Failure (drone ship)'
    AND YEAR(Date) = 2015
```

[52]:

landing_outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Query Explanation

Keyword YEAR() used to return year from the Column DATE.

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

```
%%sql

SELECT Landing_Outcome, COUNT(Landing_Outcome) as Total
    from SPACEXTBL
    WHERE Date BETWEEN '2010-06-04' and '2017-03-20'
    GROUP BY Landing_Outcome
    ORDER BY Total DESC
```

[53]:

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

Query Explanation

Keyword BETWEEN used to return entries with DATE Column values between limits. Group by and Order by used to Group and Order results.

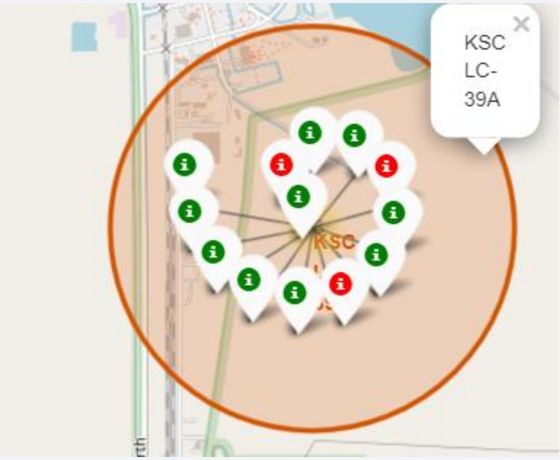
Section 3

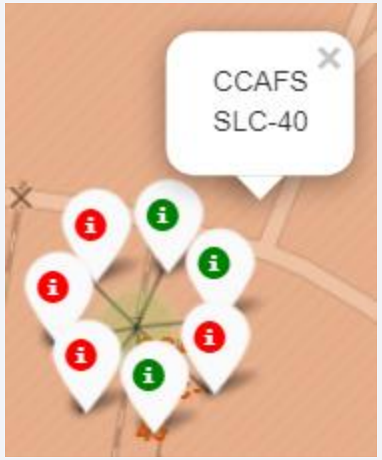
Launch Sites Proximities Analysis

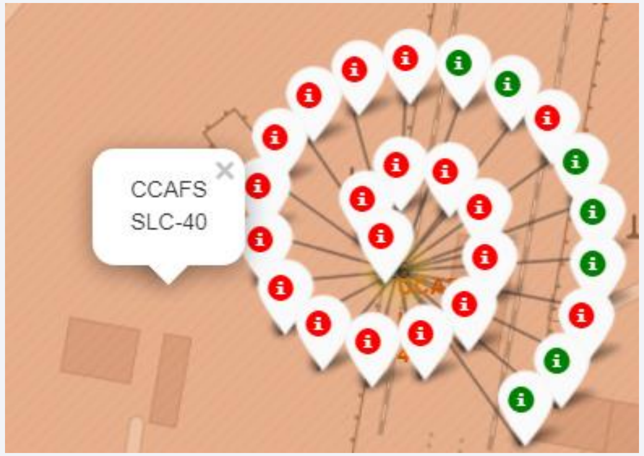
Spacex Launch Sites – World Map (Folium)

SpaceX Launch Sites are in the USA – Florida & California

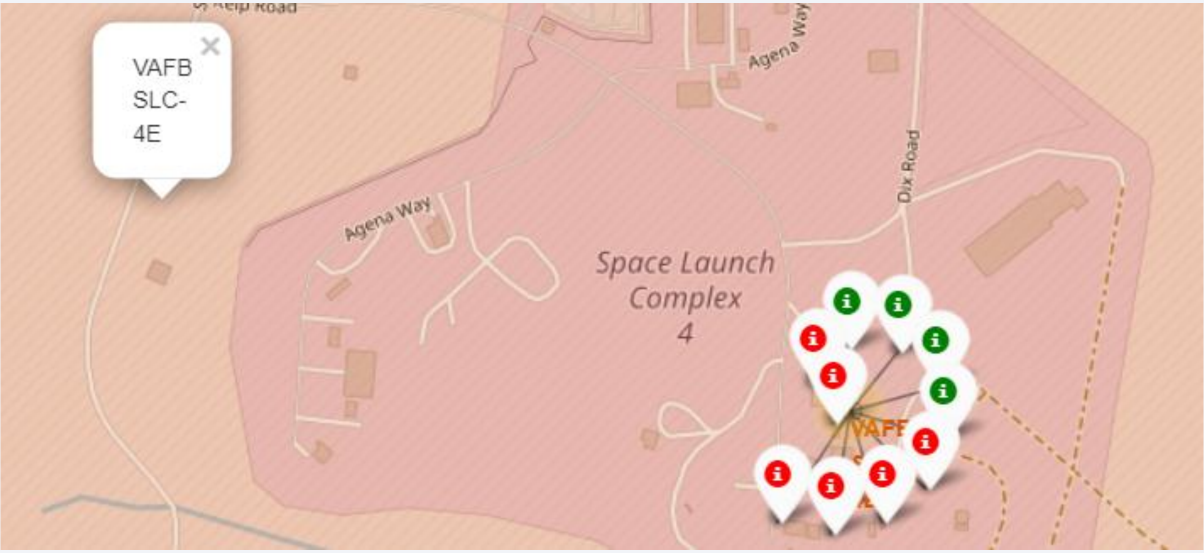
Launch Sites with Color Markers (Folium)







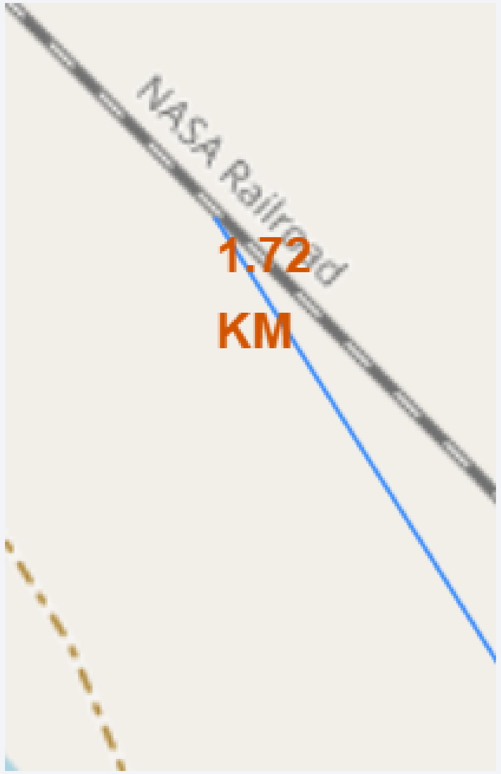
Florida Launch Sites

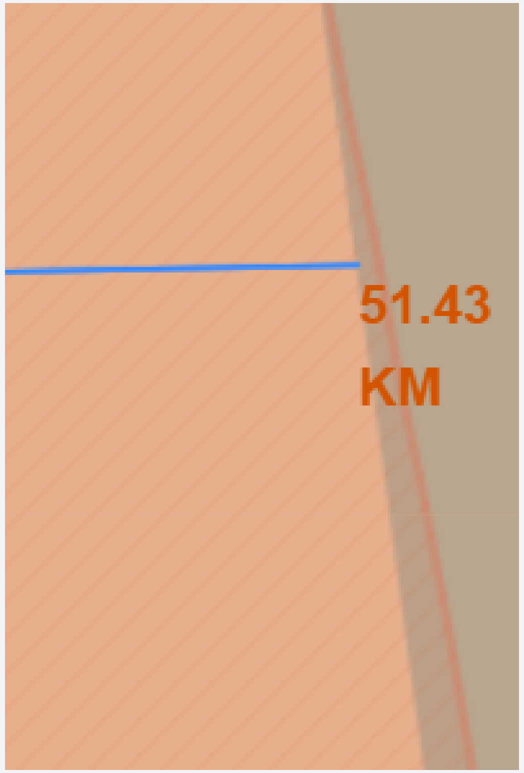


California Launch Sites

Green Markers = Successful Launch
Red Markers = Failed Launch

CAFS-SLC40 Distance to landmarks (Folium)



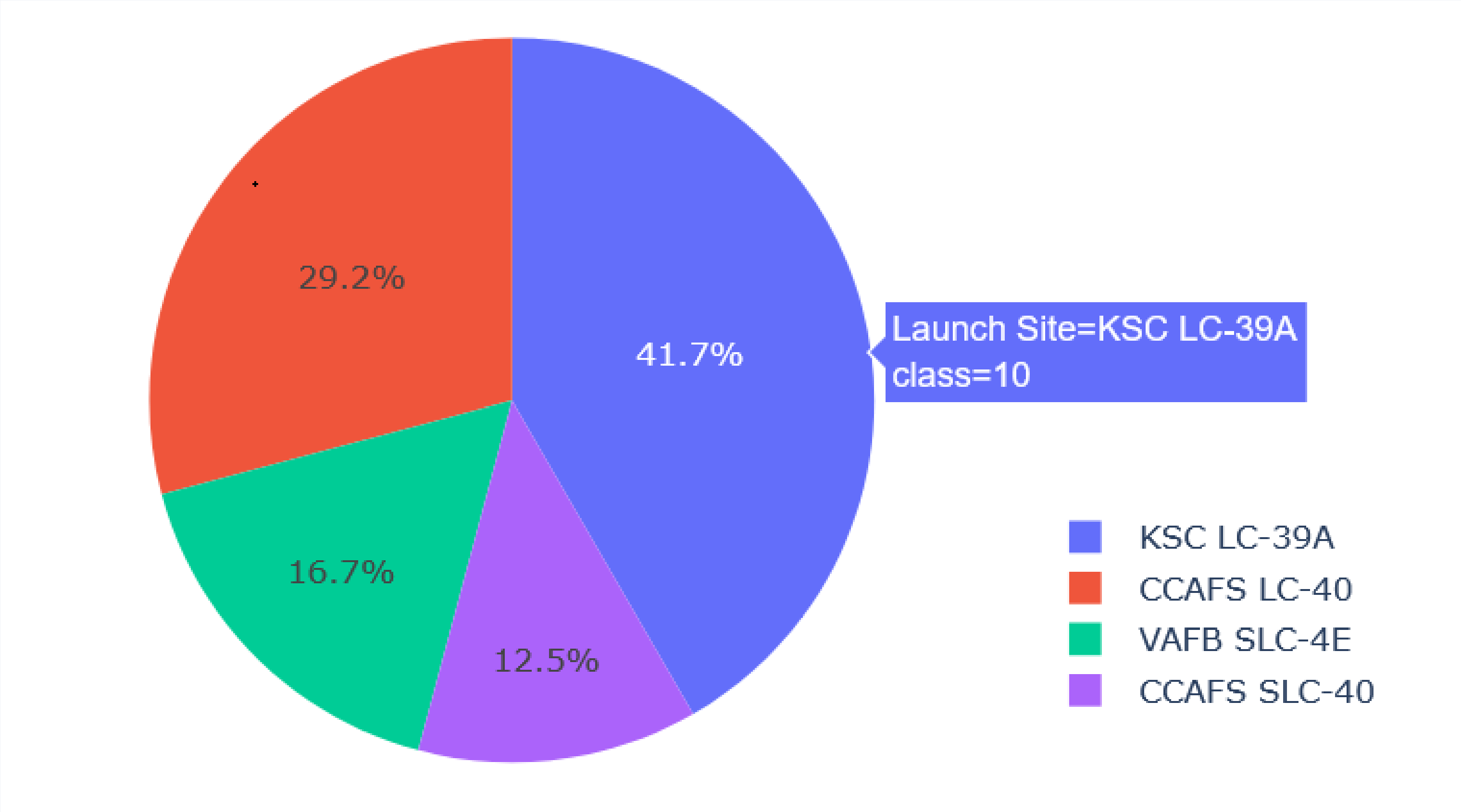



```
distance_highway = 0.5834695366934144  km
distance_railway = 1.7195668376581266  km
distance_city = 51.43416999517233  km
distance_coast = 0.8627671182499878  km
```

Section 4

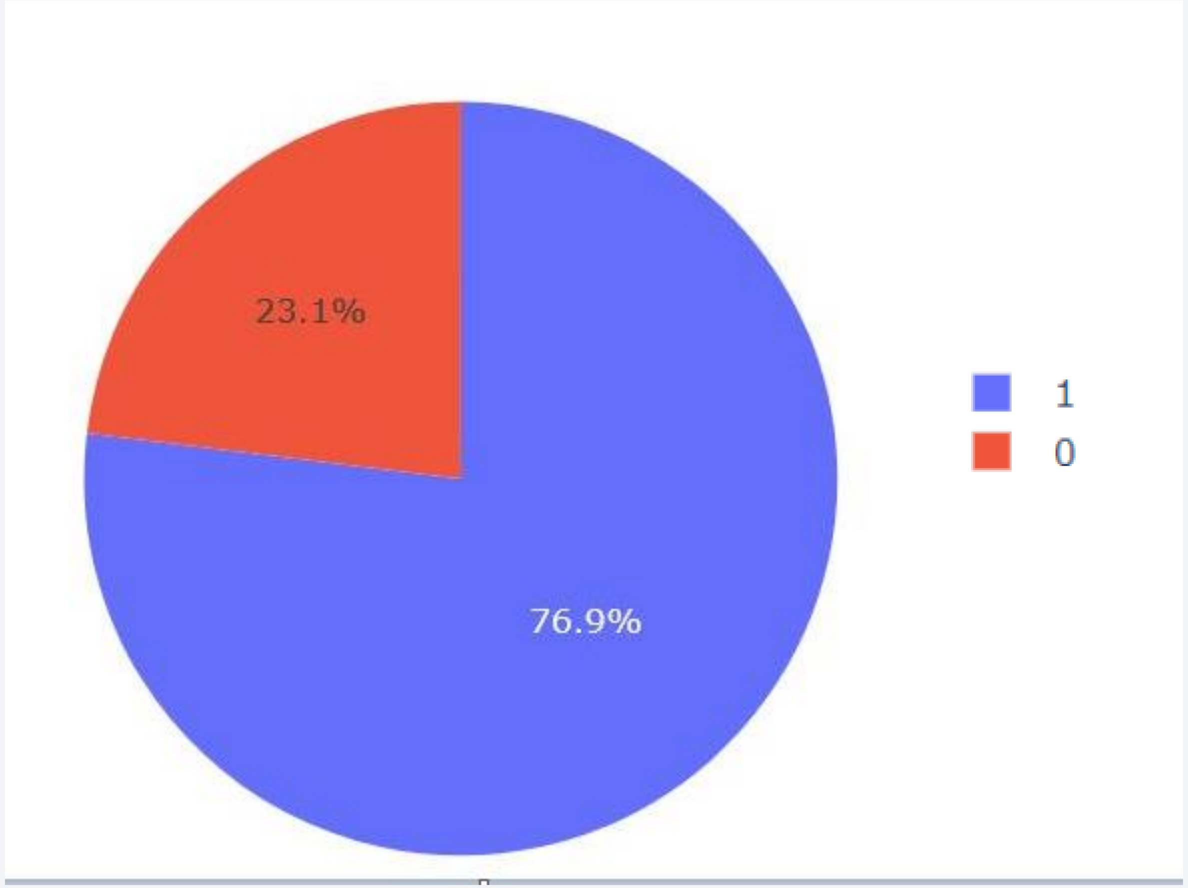
Build a Dashboard with Plotly Dash

% of total successful launches by Launch Site (plotly)



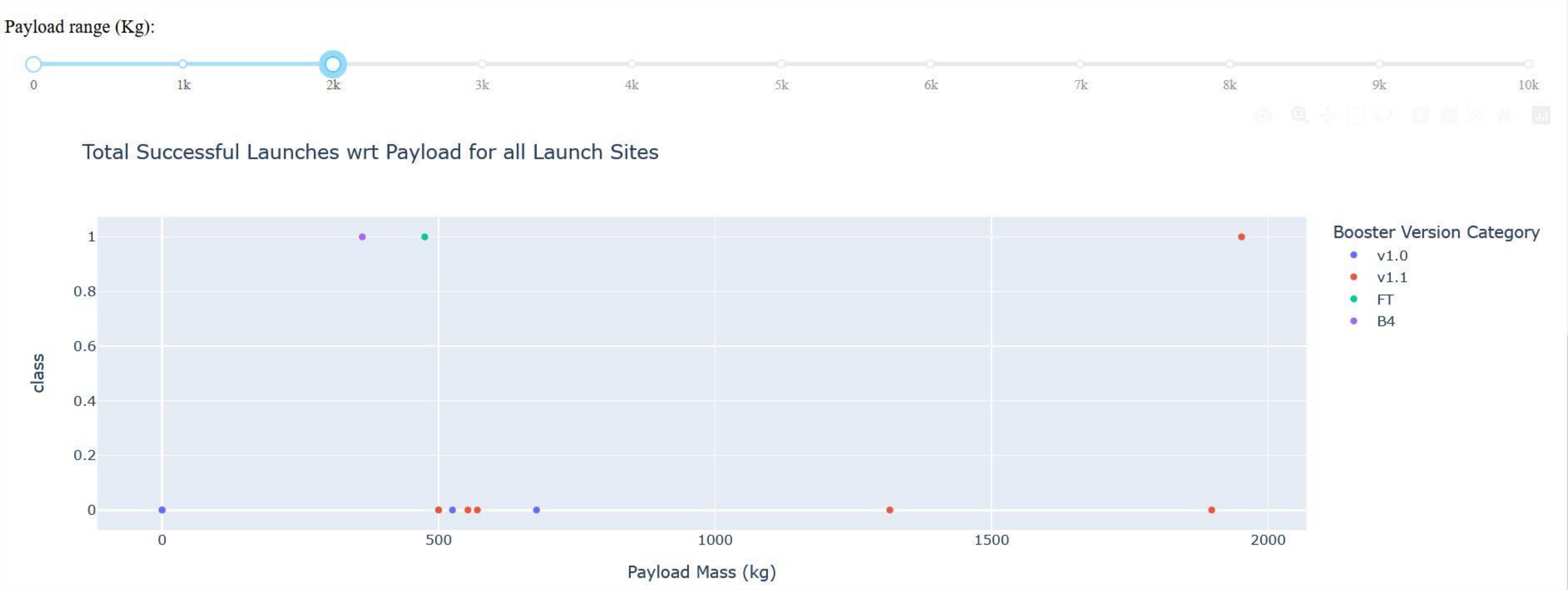
KSC LC – 39 A has the most successful launches as fraction of SpaceX's total Successful launches

Success ratio for most successful site. (plotly)



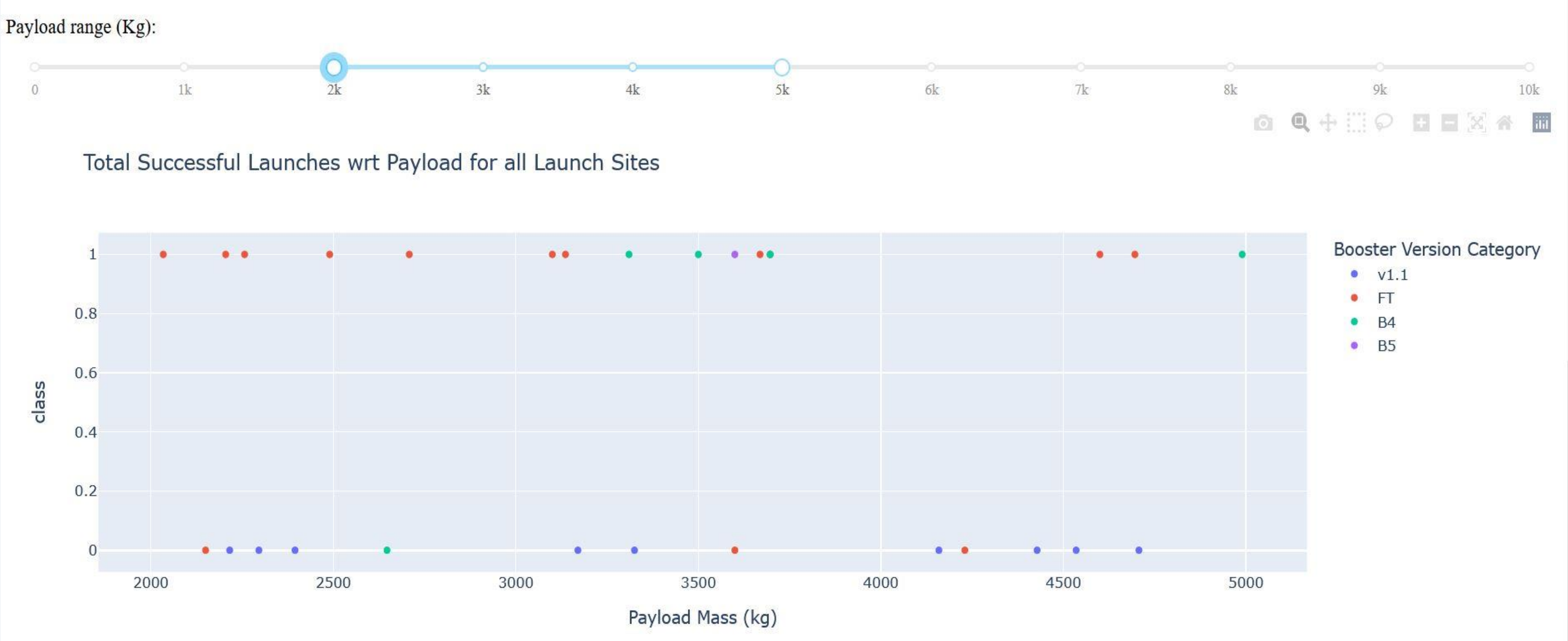
Total Success Launches for site KSC LC-39A

Successful Launches with 0<Payload<2000, All Sites (plotly)



High failure rate observed between Payloads of 0 and 2000 kg.

Successful Launches with 2000<Payload<5000, All Sites (plotly)



Inconclusive results for success for Payloads between 2000, 5000

Section 5

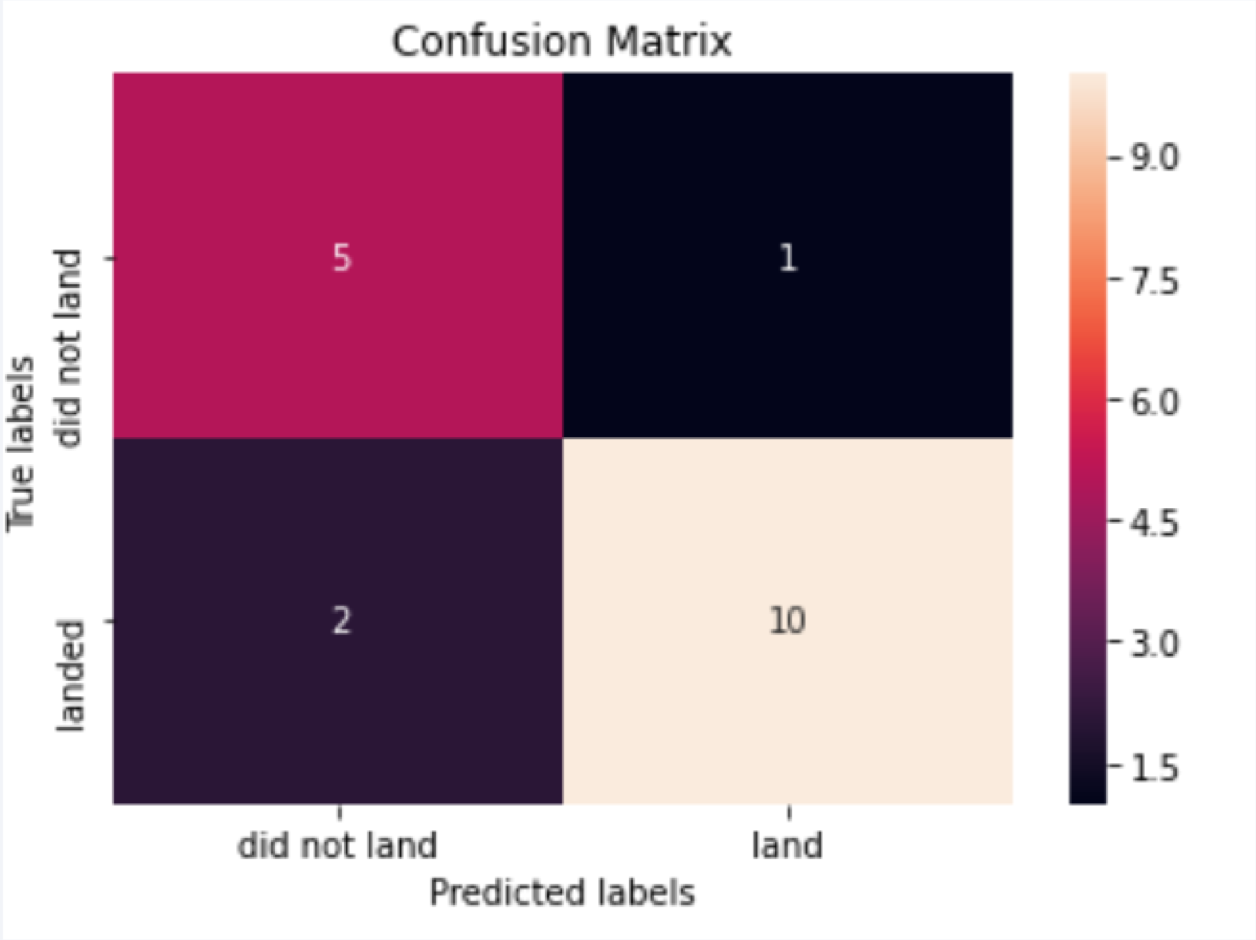
Predictive Analysis (Classification)

Classification Accuracy

```
Train-Set Accuracy
Logrithmic:  0.8464285714285713
SVM       :  0.8482142857142856
Tree      :  0.8892857142857142
KNN       :  0.8482142857142858


Test-Set Accuracy
Logrithmic:  0.8333333333333334
SVM       :  0.8333333333333334
Tree      :  0.8333333333333334
KNN       :  0.8333333333333334
```

Tree did better than the other 3 on Train Set, but did the same as other three on Test Set.

Confusion Matrix



Excluding 2 False Negatives, and 1 False Positive, the algorithm did very well

Conclusions

- Success rates of launch sites start increasing after year 2013.
- ESL1, GEO, HEO & SSO are the Orbits with the highest Success Rate
- KSC LC-39A had the most successful launches
- Decision tree had the highest accuracy of all the Machine Learning Algorithm
- It is possible to compete with SpaceX, given our ability to predict successful landing with reasonable accuracy given initial conditions

Appendix

[Github Repository with all Notebooks]

Thank you!