IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wanjie Feng
July/15/2023

Outline

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

Executive Summary

- Summary of methodologies
  - SpaceX REST API, Web scraping → Read historical SpaceX data
  - Data Wrangling → Clean and label data
  - Exploratory data analysis using visualization and SQL → key variables
  - Interactive visual analytics using Folium and Ploty Dash → Launch sites and success rate
  - Predictive analysis using classification models → Best data model
- Summary of all results
  - Launch Site, Landing site or method, Payload Mass, Booster Version, Year, orbit type are key factors
  - Decision Tree is the most accurate algorithm.

Introduction

- **Project background and context**
  - As a commercial space company, SpaceX perhaps is the most successful one.
  - One key reason is Falcon 9 reusable first stage, which makes Falcon 9 relatively inexpensive and more competing.
  - To compete with SpaceX, SpaceY needs to analyze why Falcon 9 can successfully recover first stage.

- **Problems you want to find answers**
  - Based on SpaceX Falcon 9 historical data, SpaceY needs to find what key factors are on the success of first stage recovering.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API, Python Requests and JSON libraries
  - Web scraping related Wiki pages
- Perform data wrangling
  - Missing value percentage; Data types; Values counts on Launch Site, Orbit, Outcome; label landing outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - GridSearchCV find best parameters for LR, SVM, Decision Tree and KNN

Data Collection

- Describe how data sets were collected.
  - SpaceX REST API, Requests
  - Web Scraping, BeautifulSoup

Data Collection – SpaceX API

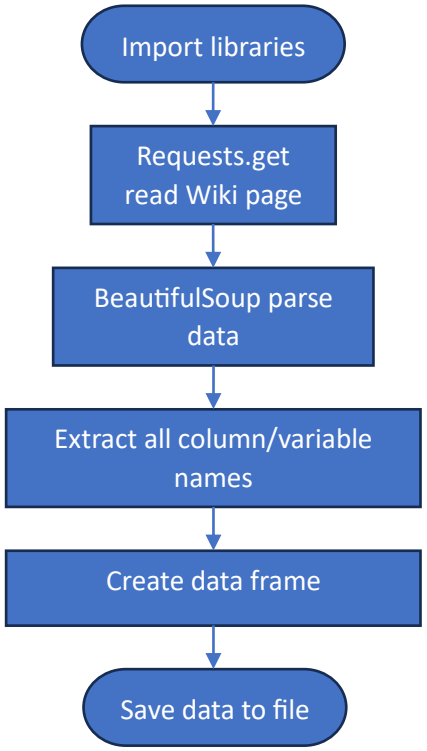
- SpaceX API and Requests to read in data → decode and normalize → select required data for this project →clean them and save to file.

- https://github.com/Aaron2014/IBM_Data_Science_Practice/blob/9bf8c275db97aa370604724486cd34b03adc6815/SpaceX%20-%20Collecting%20the%20data.ipynb

Data Collection - Scraping

- Requests read Wiki page → BeautifulSout parse data → Create data frame → Save

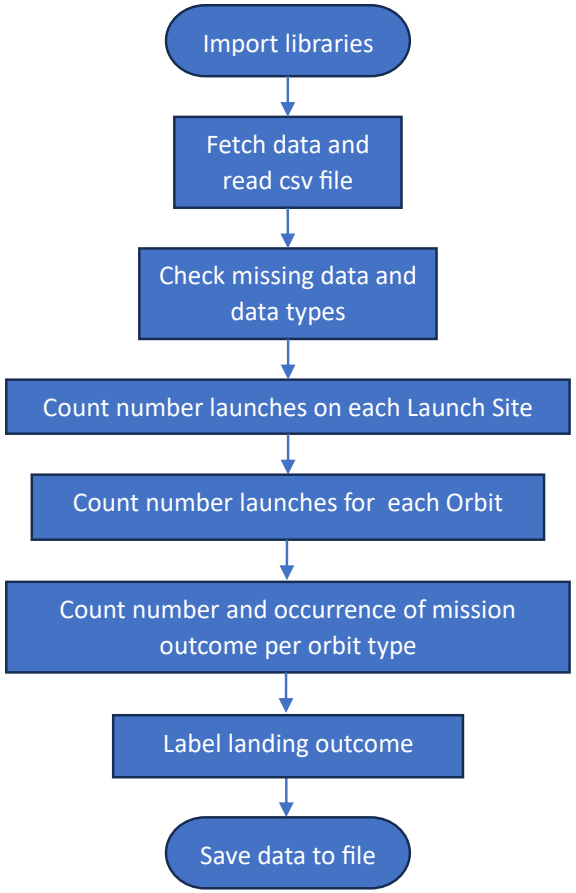
- https://github.com/Aaron2014/IBM_Data_Science_Practice/blob/693b48abb95cf5b398a9630bfe252ff88a7c2443/SpaceX%20-%20Web%20Scraping%20from%20Wiki.ipynb



Data Wrangling

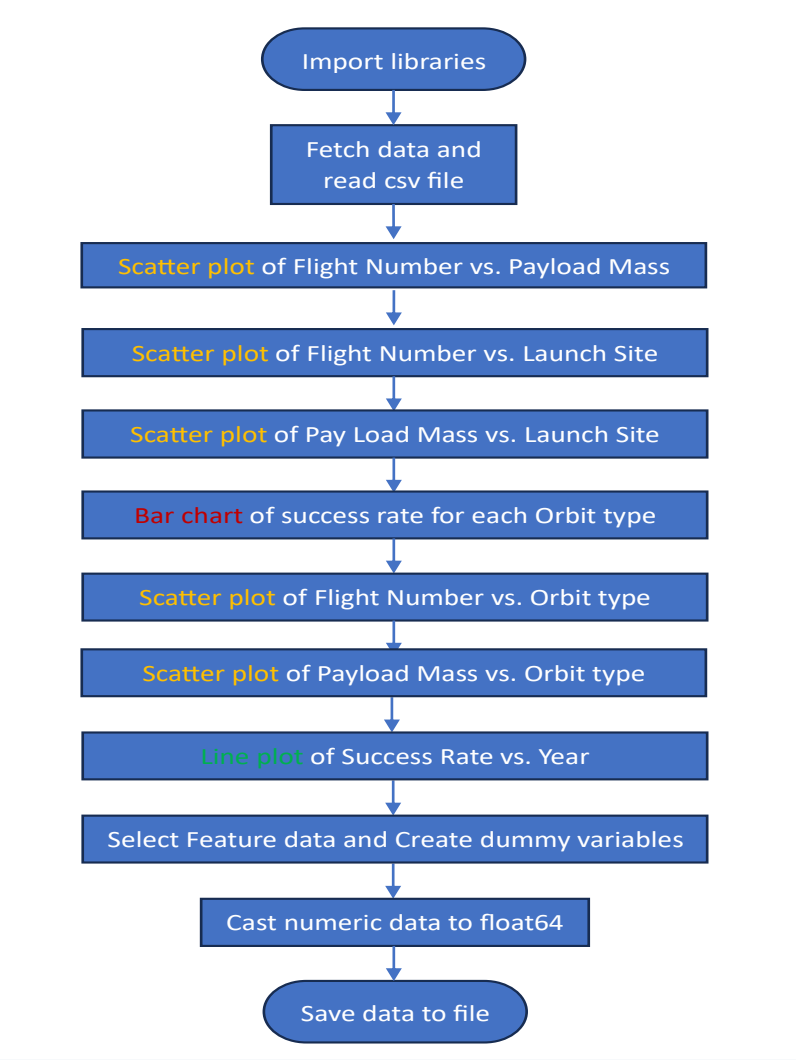
- Data were read → missing data and data type were checked → launches on each launch site and for each orbit were counted → landing outcome were labeled.

- https://github.com/Aaron2014/IBM_Data_Science_Practice/blob/693b48abb95cf5b398a9630bfe252ff88a7c2443/SpaceX%20-%20Data%20Wrangling.ipynb



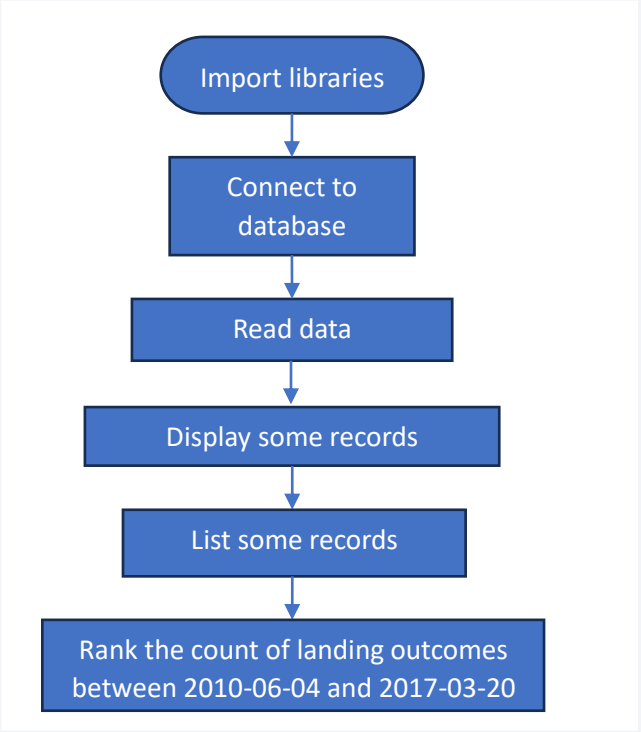
EDA with Data Visualization

- Scatter plots: find relationships between variables
- Bar chart: visualize relationship between success rate of each orbit type
- Line plot: view launch success yearly trend
- https://github.com/Aaron2014/IBM_Data_Science_Practice/blob/693b48abb95cf5b398a9630bfe252ff88a7c2443/SpaceX%20-%20Exploring%20and%20Preparing%20Data.ipynb



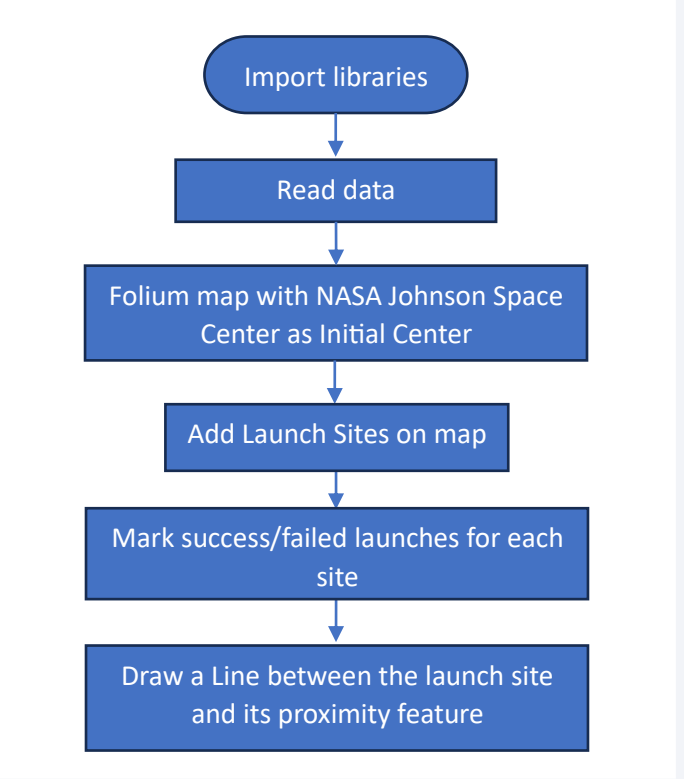
EDA with SQL

- Display:
  - ➢the names of the unique launch sites in the space mission
  - ➢5 records where launch sites begin with the string 'CCA'
  - ➢the total payload mass carried by boosters launched by NASA (CRS)
  - ➢ average payload mass carried by booster version F9 v1.1
- List:
  - ➢ the date when the first successful landing outcome in ground pad
  - ➢ the names of the boosters succeeded in drone ship with payload mass >4000 & < 6000
  - ➢ the total number of successful and failure mission outcomes
  - ➢ the names of the booster_versions carried the maximum payload mass
  - ➢ the month, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Rank: the count of landing outcomes between the date 2010-06-04 and 2017-03-20
- https://github.com/Aaron2014/IBM_Data_Science_Practice/blob/693b48abb95cf5b398a9630bfe252ff88a7c2443/SpaceX%20-%20Overview%20of%20Dataset.ipynb



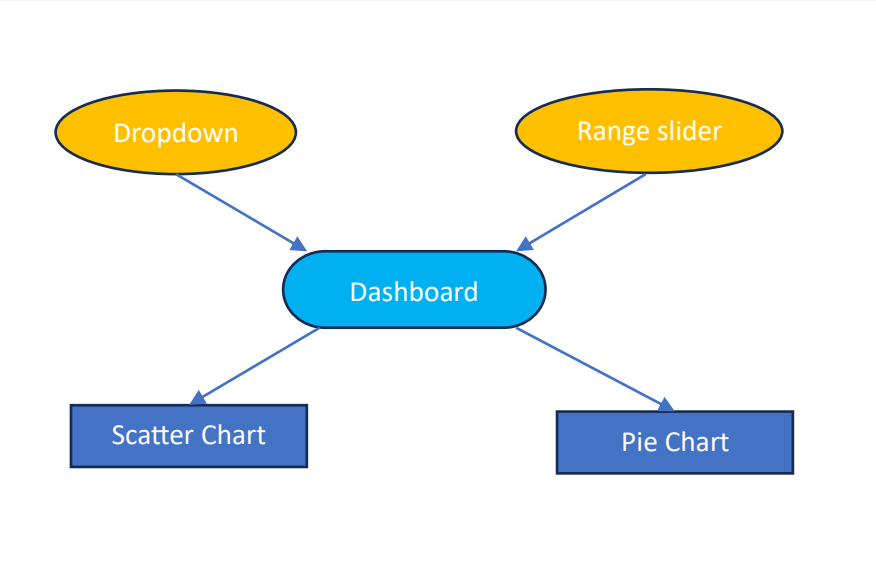
Build an Interactive Map with Folium

- Folium Map with NASA Johnson Space Center as initial center → Add Launch Sites → Mark success/failed launches for each site → Draw a line between the launch site and its proximity feature.
- Where launch sites locate?
- Which sites have high success rates?
- What features near each launch site?
- https://github.com/Aaron2014/IBM_Data_Science_Practice/blob/693b48abb95cf5b398a9630bfe252ff88a7c2443/SpaceX%20-%20Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb



Build a Dashboard with Plotly Dash

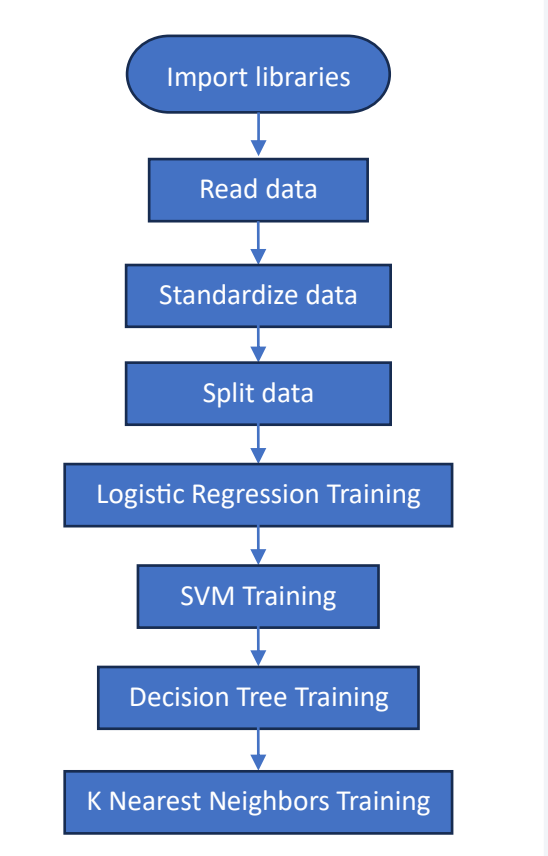
- Dropdown of launch site
- Range slider of payload mass
- Pie chart of success launch by each site
- Scatter chart of success launch vs. payload mass



- https://github.com/Aaron2014/IBM_Data_Science_Practice/blob/693b48abb95cf5b398a9630bfe252ff88a7c2443/spacex_dash_app.py

Predictive Analysis (Classification)

- Standardize data
- **LR, SVM, Decision Tree, KNN** algorithms with **GridSearchCV** to find best parameters.
- https://github.com/Aaron2014/IBM_Data_Science_Practice/blob/693b48abb95cf5b398a9630bfe252ff88a7c2443/SpaceX%20-%20Machine%20Learning%20Prediction.ipynb



Results

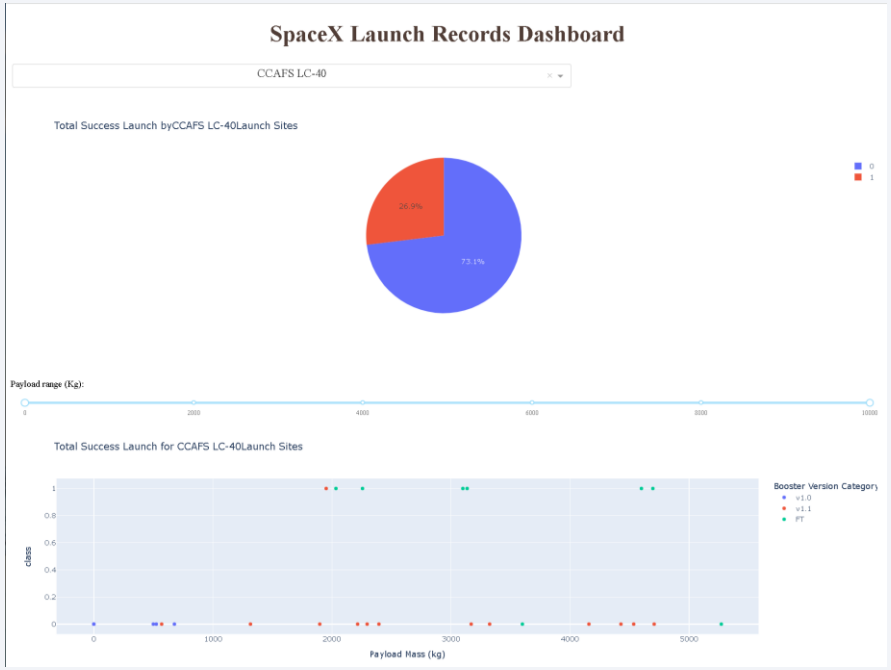
- Exploratory data analysis results
  - ❖ Success rate increase with flight number, year.
  - ❖ Success rate decrease with payload mass.
  - ❖ ES-L1, GEO, HEO, SSO orbits have higher success rate
  - ❖ Different launch sites have different success rates.
  - ❖ Not all orbits have success rate related to flight number
  - ❖ Payload mass don’t have clear relationship with Orbit

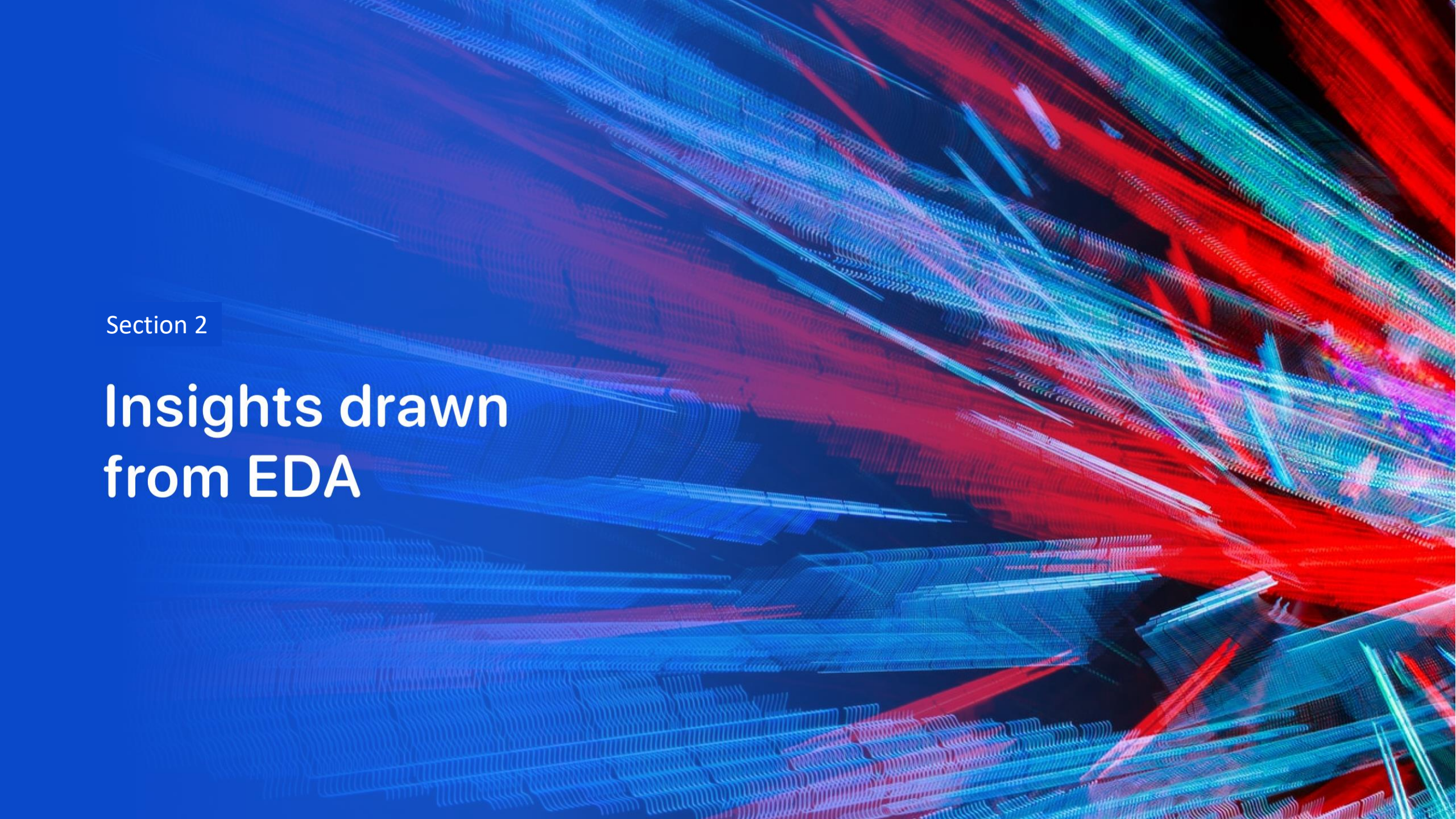
- Predictive analysis results

[34]:

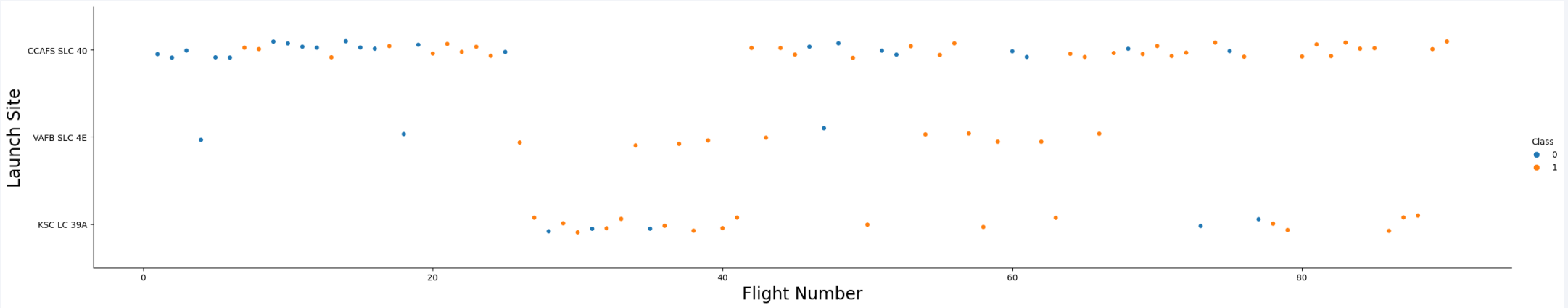
	Algorithm	Accuracy	F1-Score	Score
0	LR	0.846429	0.888889	0.833333
1	KNN	0.848214	0.888889	0.833333
2	Decision Tree	0.875000	0.888889	0.833333
3	SVM	0.848214	0.888889	0.833333

- Interactive analytics demo



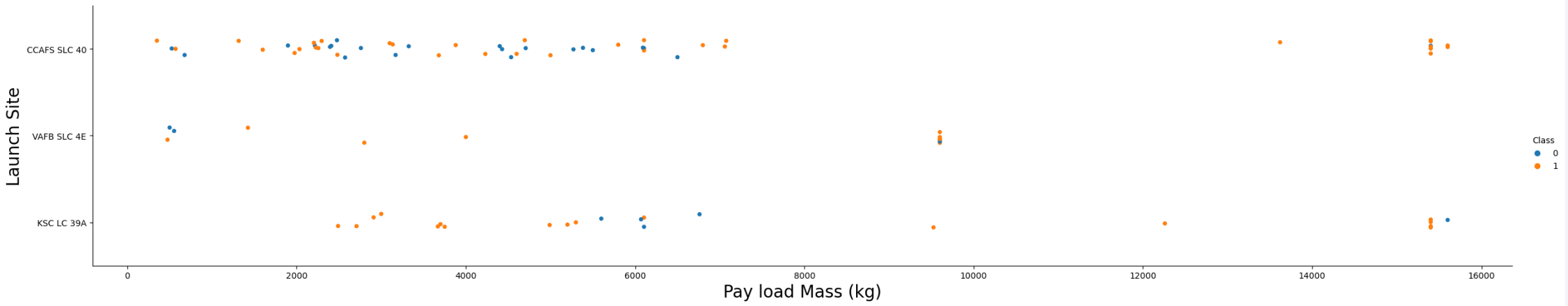

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Different launch sites have different success rate
- **CCAF SLC 40** : 60%
- **KSC LC 39A** and **WAFB SLC 4E**: 77%

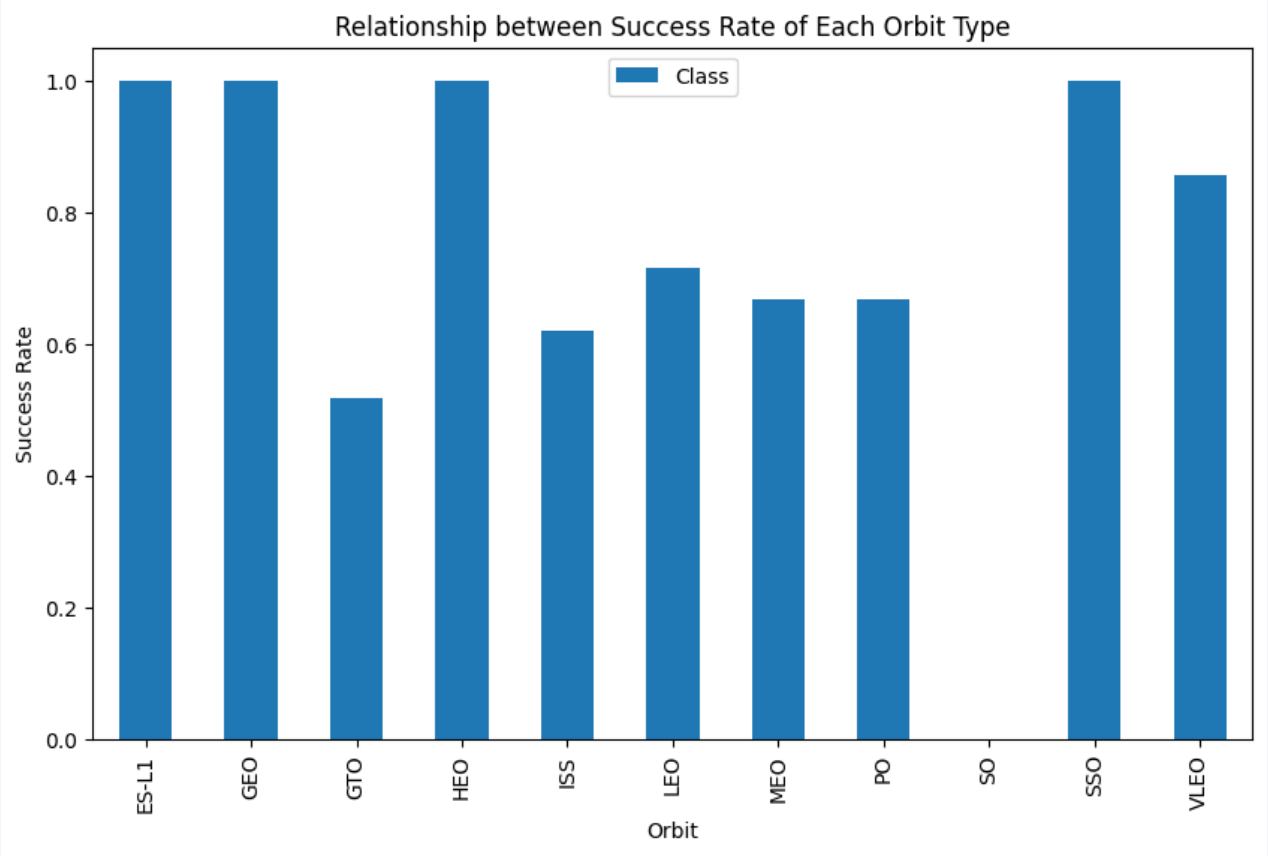
Payload vs. Launch Site



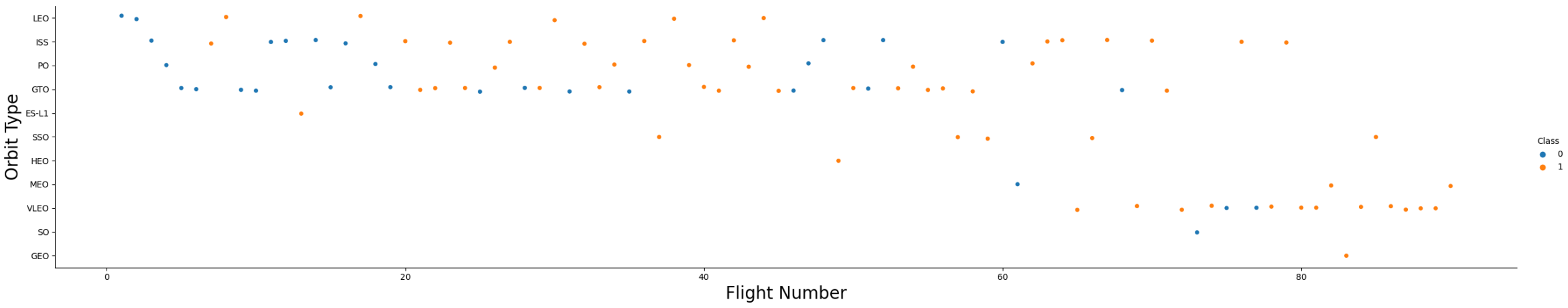
- **WAFB SLC:** no heavy payload mass (>10000kg)

Success Rate vs. Orbit Type

- ES-L1 (1), GEO (1), HEO (1) and SSO (5): higher success rate
- SO (1) : lowest success rate
- GTO (27) : relative lowest rate

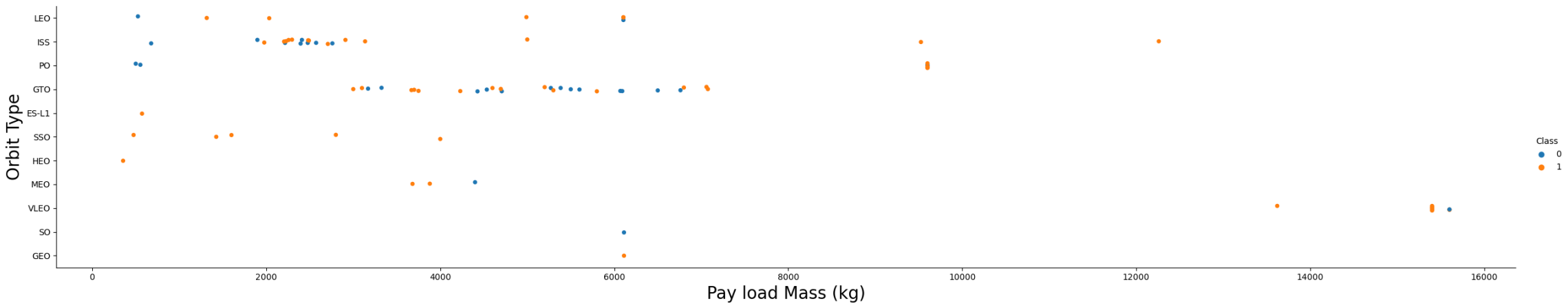


Flight Number vs. Orbit Type



- LEO: positive relationship
- GTO and others: no clear relationship

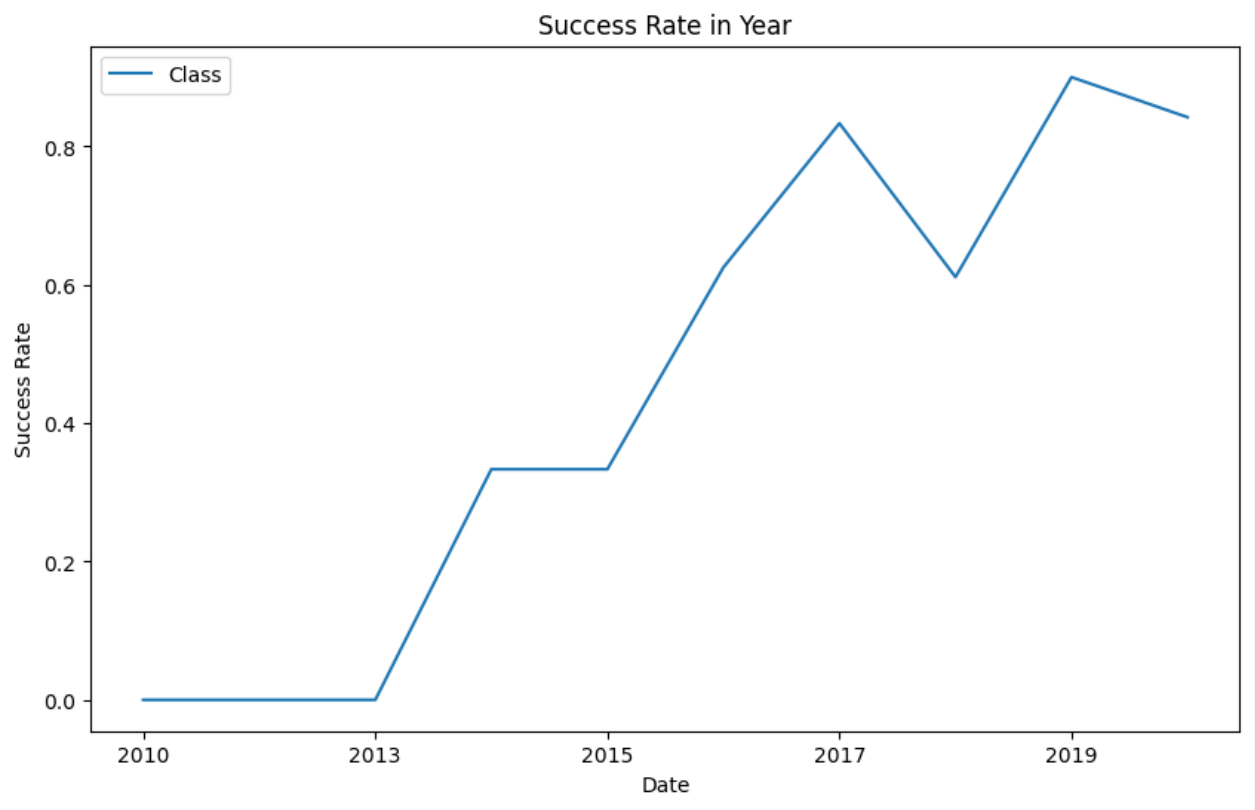
Payload vs. Orbit Type



- PO, LEO, ISS: positive relationship
- GTO and others: no clear relationship

Launch Success Yearly Trend

- Success rate increase with year



All Launch Site Names

- Four Launch Site:
  - CCAFS LC-40
  - WAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

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

[7]:
```
%%sql

SELECT DISTINCT (Launch_Site)
FROM SPACEXTBL
WHERE Launch_Site is not null
```

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

[7]:

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

Launch Site Names Begin with 'CCA'

- Use LIKE ‘CCA%’

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

[8]:
```
%%sql

SELECT *
FROM SPACEXTBL
WHERE Launch_Site like 'CCA%'
LIMIT 5
```

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

[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- WHERE Customer = ‘NASA (CRS)’

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

```
[9]: %%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass_By_NASA
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)'
```

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

[9]:

Total_Payload_Mass_By_NASA
45596.0

Average Payload Mass by F9 v1.1

- WHERE

Booster_Version LIKE ‘F9 v1.1%’

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

```
[10]: %%sql

SELECT ROUND(AVG(PAYLOAD_MASS__KG_),2) AS AVG_Payload_Mass_KG
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

[10]:

AVG_Payload_Mass_KG
2534.67

First Successful Ground Landing Date

- MIN(Date)

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

Hint:Use min function

[11]:
```
%%sql

SELECT MIN(Date)
FROM SPACEXTBL
WHERE Mission_Outcome = 'Success'
```

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

[11]:

MIN(Date)
01/06/2014

Successful Drone Ship Landing with Payload between 4000 and 6000

- WHERE

"Landing_Outcome" = "Success (drone ship)" and ("PAYLOAD_MASS__KG_" BETWEEN 4000 and 6000)

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[14]:
```
%%sql

SELECT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE  "Landing_Outcome" = "Success (drone ship)" and ("PAYLOAD_MASS__KG_" BETWEEN 4000 and 6000)
```

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

[14]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696.0
F9 FT B1026	4600.0
F9 FT B1021.2	5300.0
F9 FT B1031.2	5200.0

Total Number of Successful and Failure Mission Outcomes

- In 102 records, there are 99 success, 1success with payload status unclear and 1 failure.

Task 7

List the total number of successful and failure mission outcomes

[20]:
```
%%sql

SELECT Mission_Outcome, COUNT(*) AS 'Number'
FROM SPACEXTBL
WHERE Mission_Outcome IS NOT null
GROUP BY Mission_Outcome
```

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

[20]:

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

Boosters Carried Maximum Payload

- **F9 B5** carried the maximum payload mass

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

[16]:
```
%%sql

SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

[16]:

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

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

[17]:
```
%%sql

SELECT substr(Date, 4, 2) AS 'Month in 2015', Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' and substr(Date,7,4)='2015'
```

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

[17]:

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

- In April and October 2015, there were failures landing on drone ship.

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

- No Attempt: 10, highest

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.

[18]:
```
%%sql

SELECT  Landing_Outcome, COUNT(*) AS Number
FROM SPACEXTBL
WHERE (substr(Date, 7, 4)||substr(Date,4,2)||substr(Date,1,2)) BETWEEN '20100604' and '20170320'
GROUP BY Landing_Outcome
ORDER BY Number DESC
```

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

[18]:

Landing_Outcome	Number
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

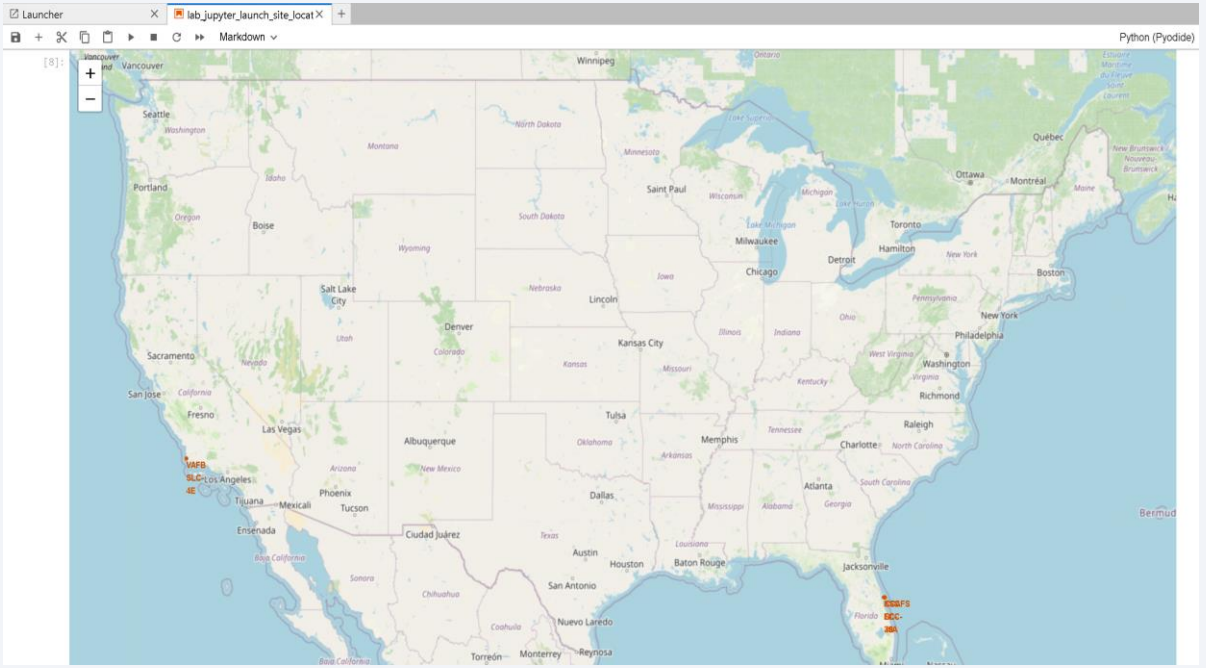
Section 3

Launch Sites Proximities Analysis

Explore Data using Folium Map

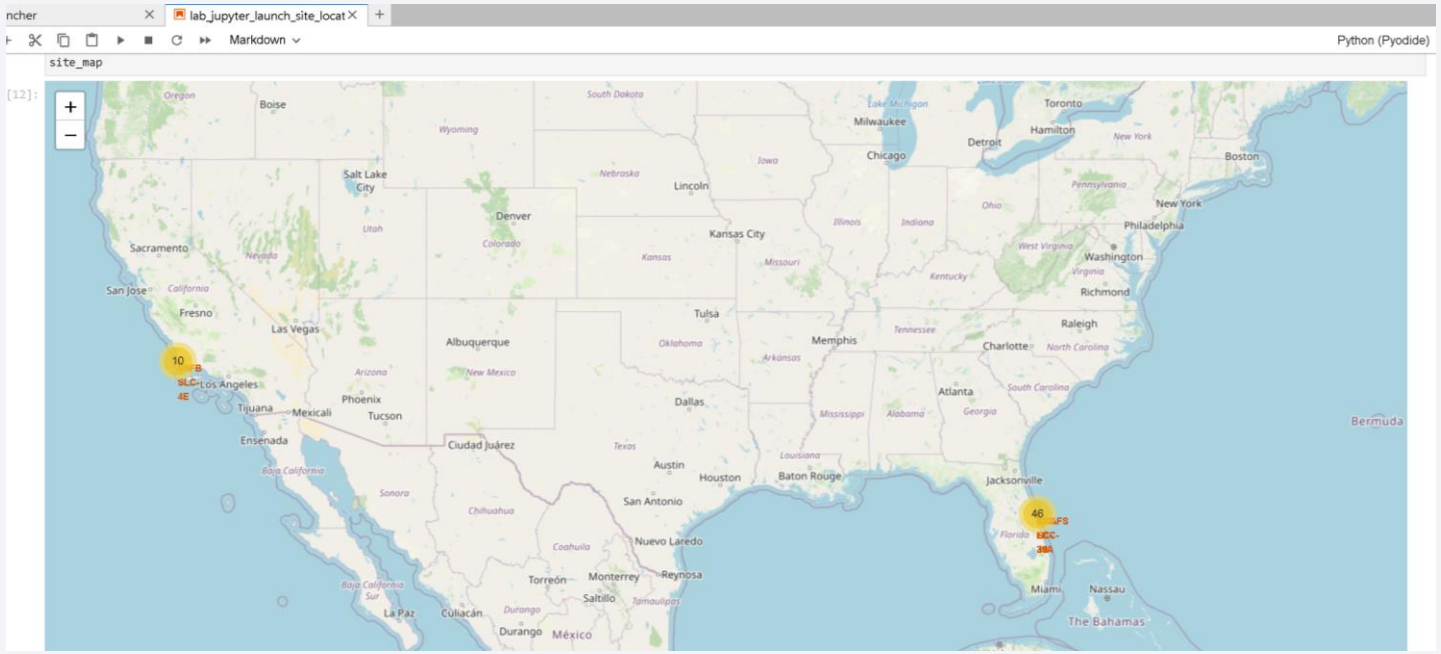
- VAFB SLC-4E in CA, West Coast

- KSC LC-39A, CCAFS SLC-40, CCAFS LC-40 in FL, East Coast



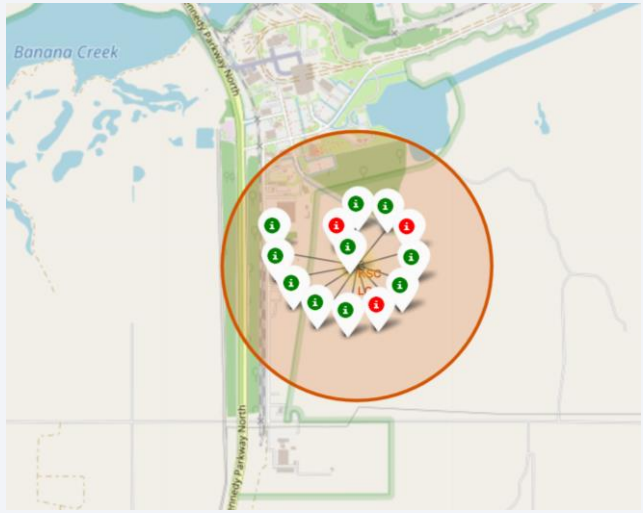
Explore Data using Folium Map

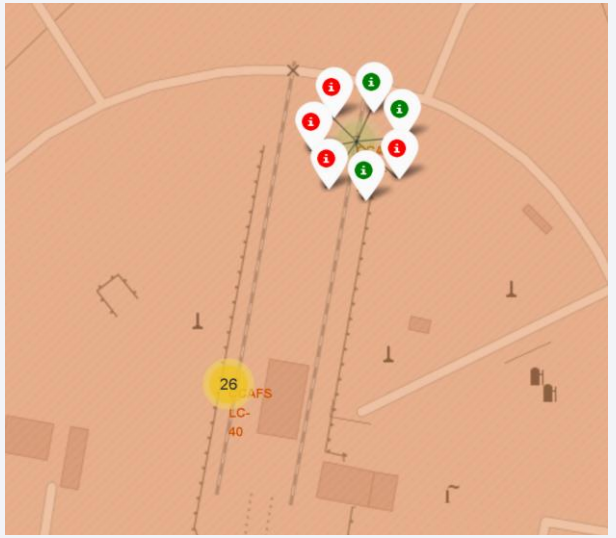
- 56 launches in total
- 10 in CA
- 46 in FL

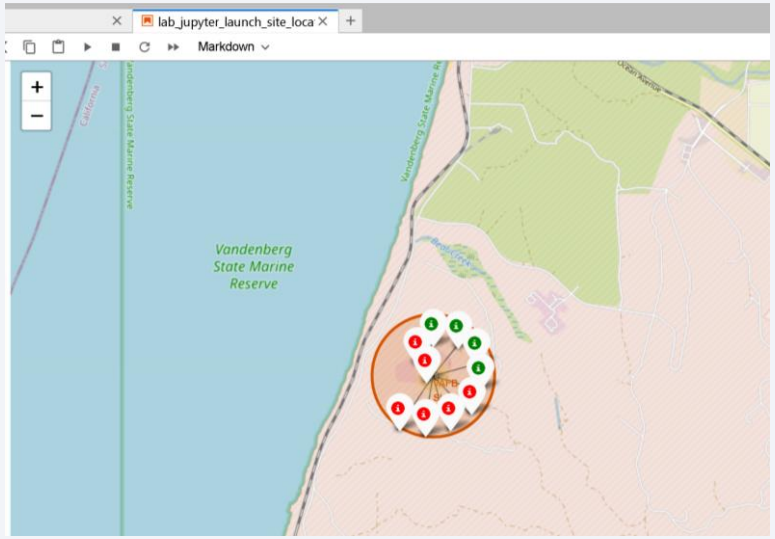


Explore Data using Folium Map

- VAFB SLC-4E: 4 of 10 success
- KSC LC-39A: 10 of 13 success
- CCAFS LC-40: 7 of 26 success
- CCAFS SLC-40: 3 of 7 success

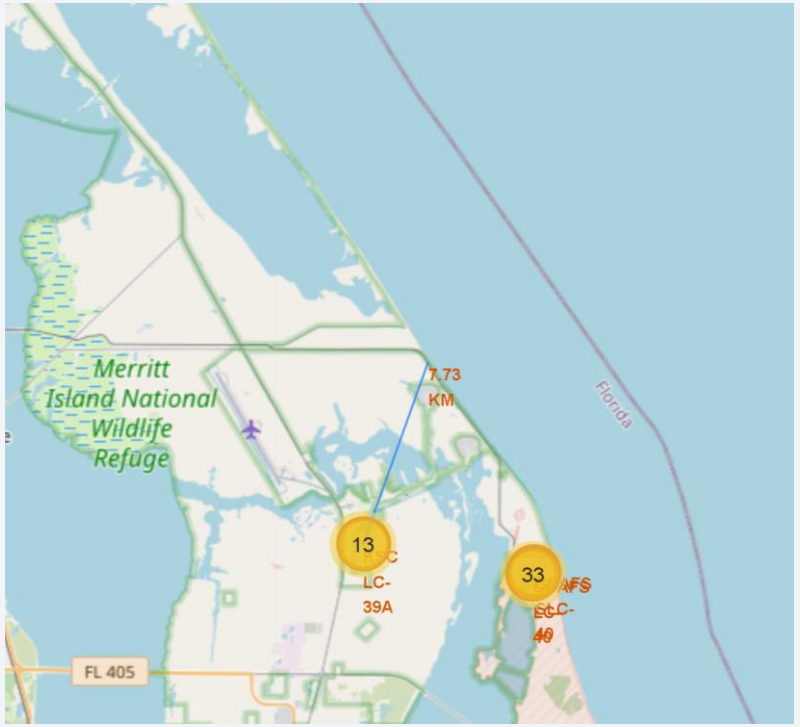


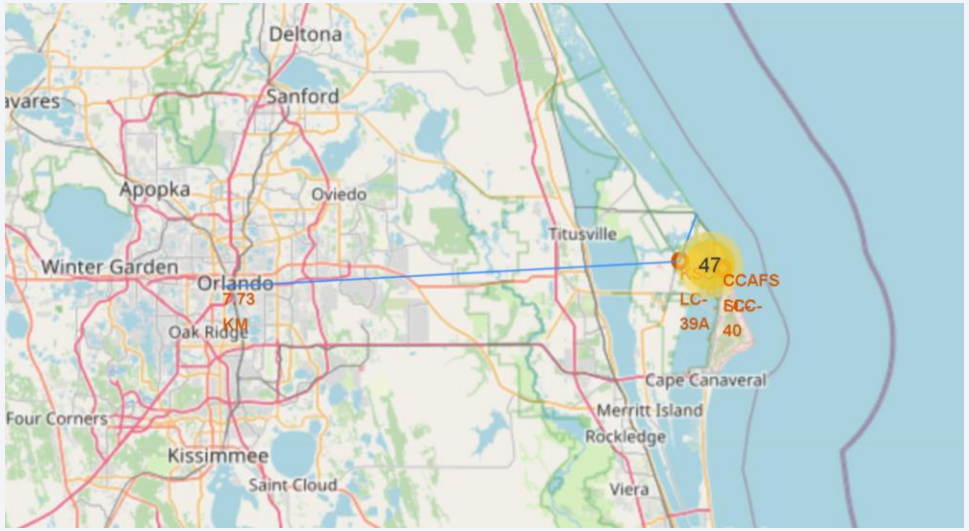


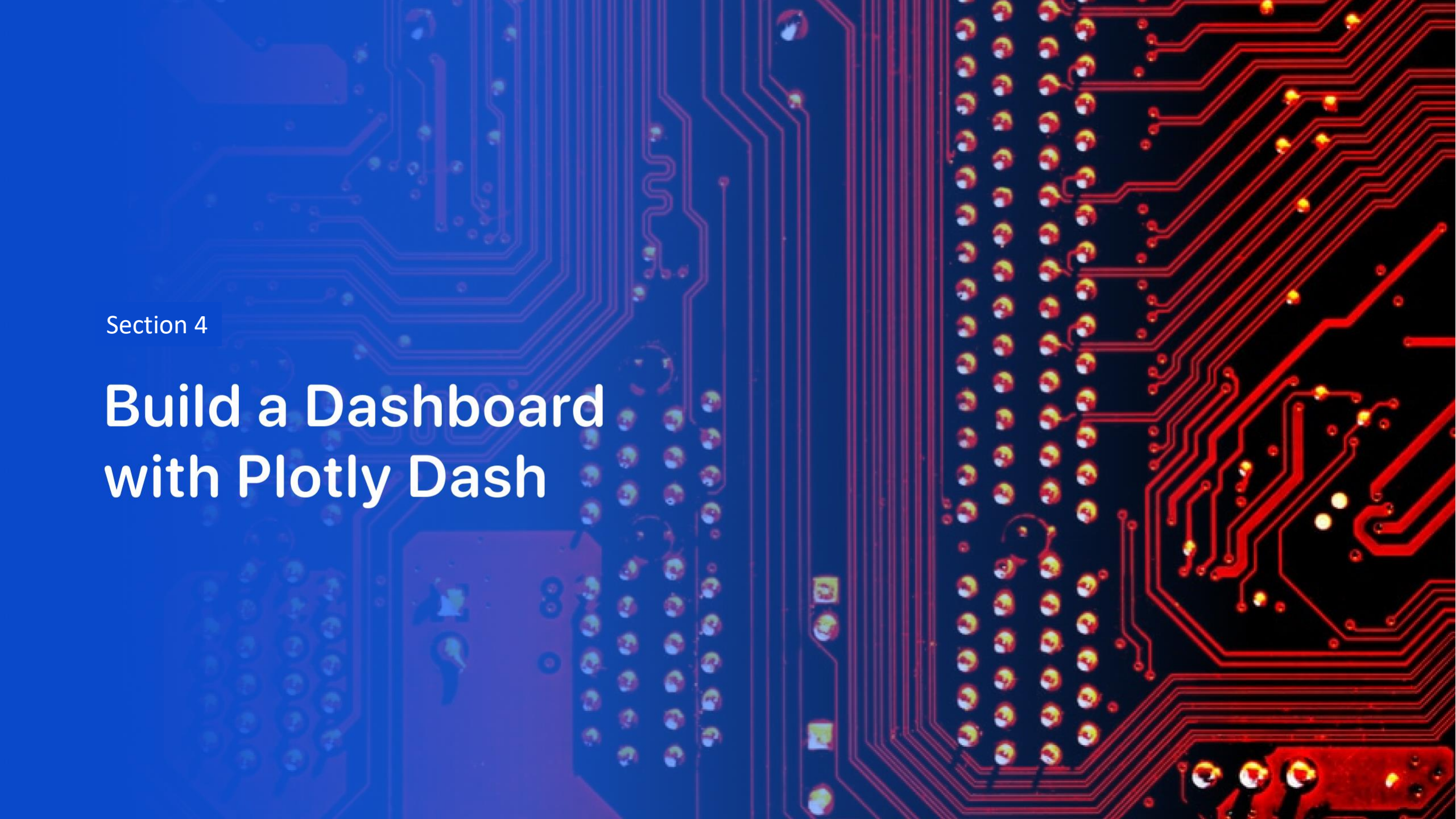


Explore Data using Folium Map

- Marked coast point is 7.73km to KSC LC-39A site
- Orlando is 773km to KSC LC-39A site

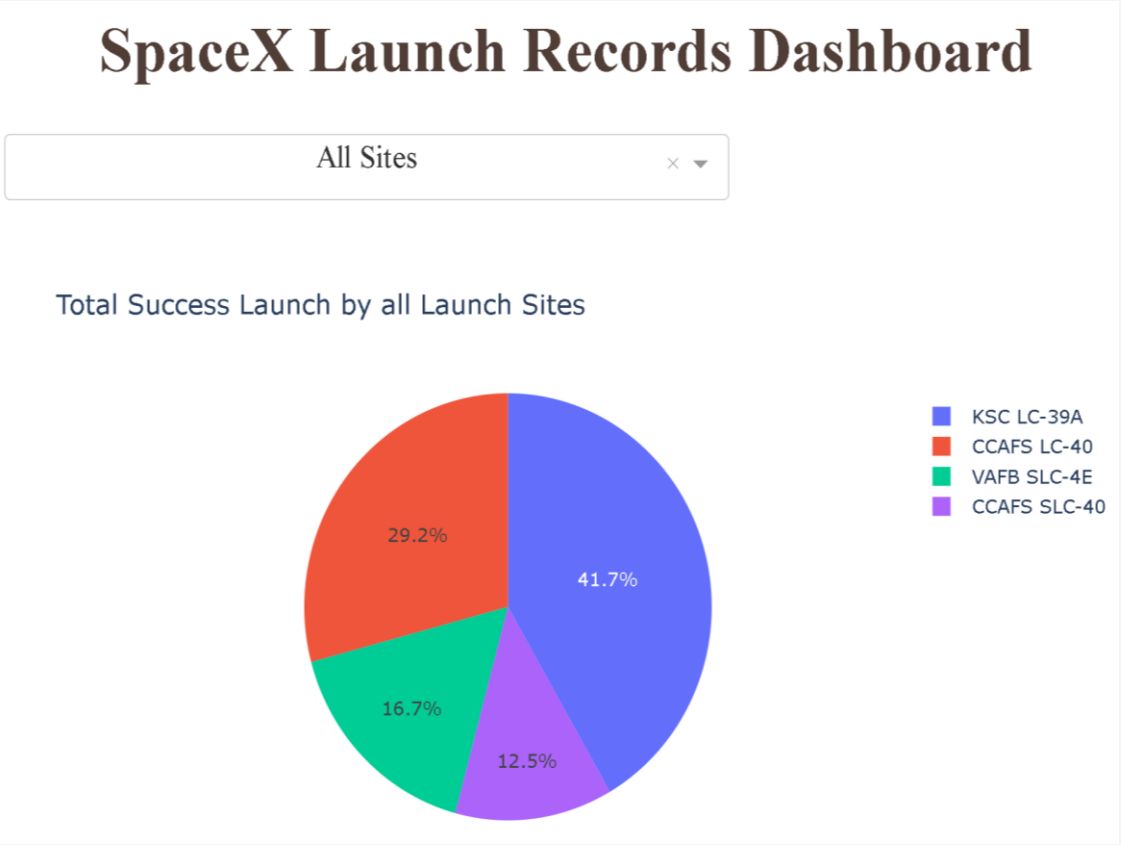





Section 4
Build a Dashboard with Plotly Dash

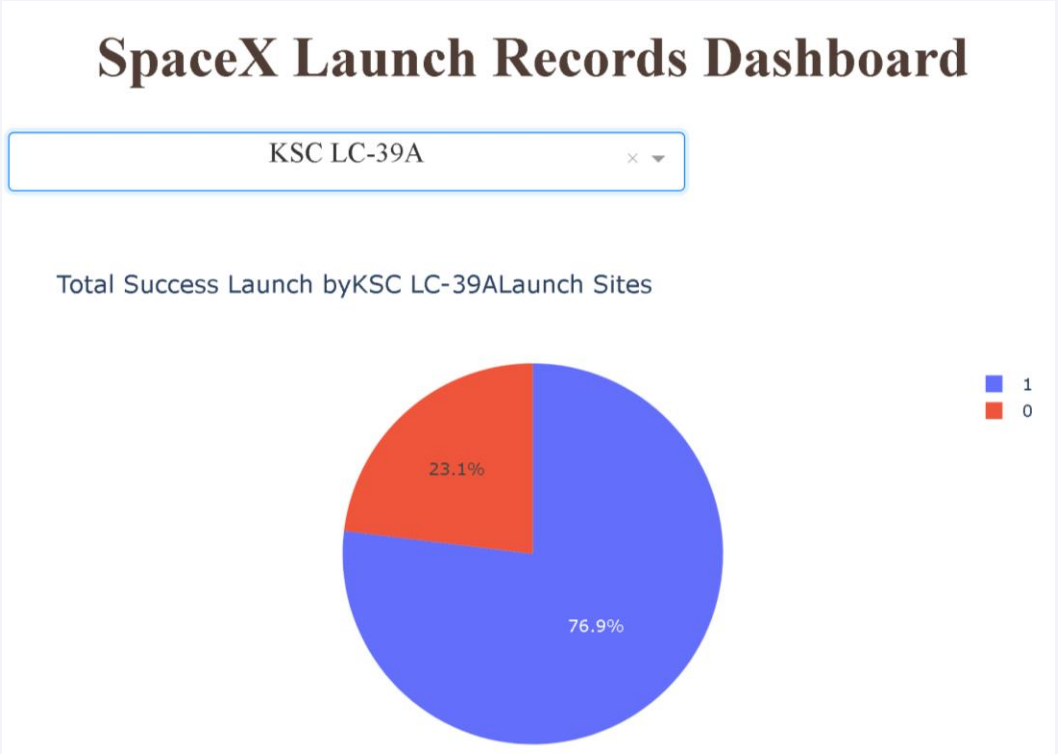
Explore Data with Dashboard

- KSC LC-39A: 41.7%
- CCAFS LC-40: 29.2%
- VAFB SLC-4E: 16.7%
- CCAFS SLC-40: 12.5%



Explore Data with Dashboard

- KSC LC-39A has 76.9% success rate



Explore Data with Dashboard

- Version FT, B4, B5 take most of success launches

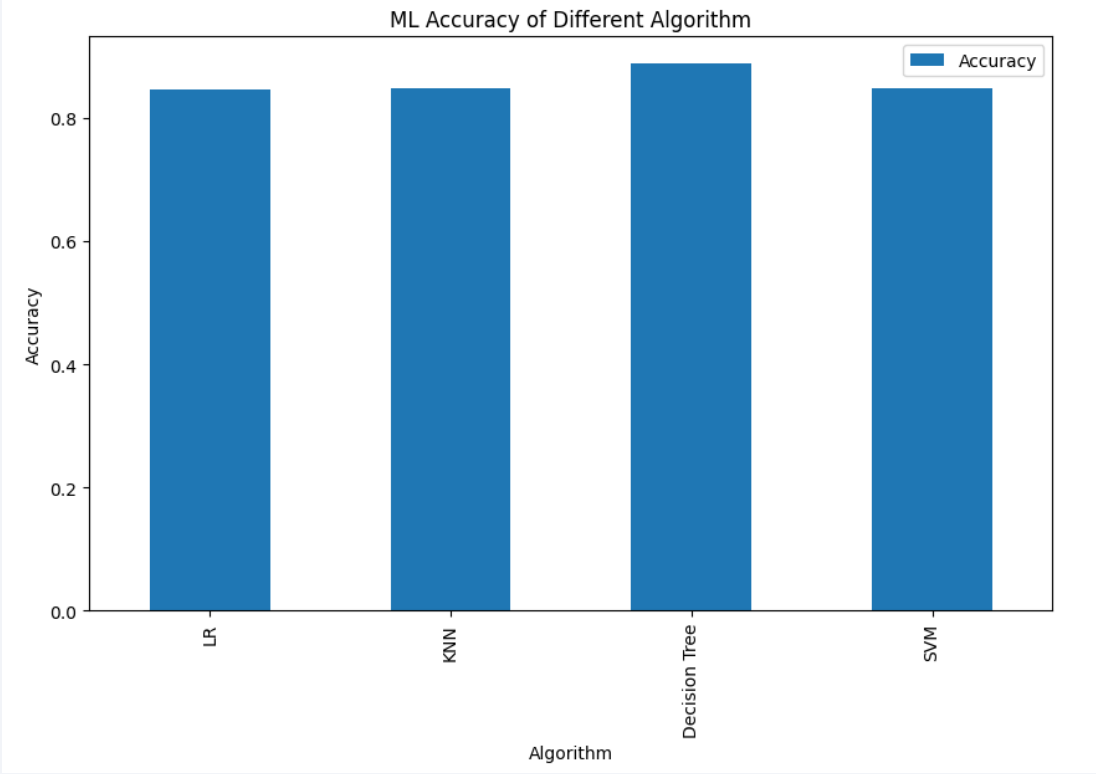
- Version v1.0, v1.1 have most of failures

Section 5

Predictive Analysis (Classification)

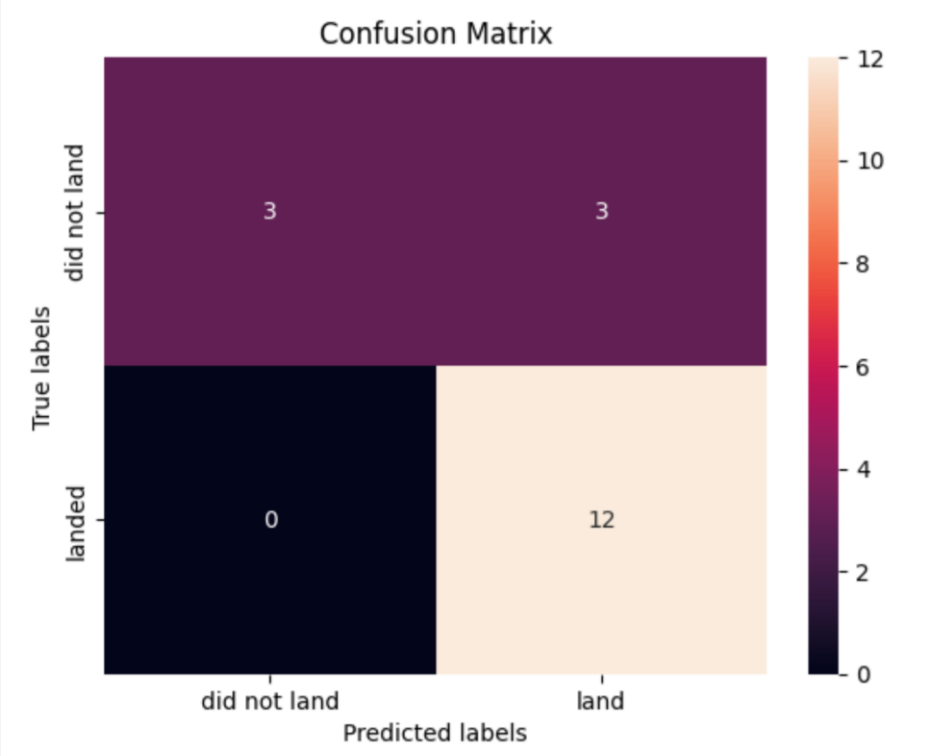
Classification Accuracy

- Decision Tree has relativly highest accuracy



Confusion Matrix

- 12 landed predicted as landed
- 3 failures predicted as success



Conclusions

- Different launch sites have different success rate. KSC LC 39A and WAFB SLC 4E have success rate of 77%
- Success rate increase with Flight Number and Year
- Success rate decrease with payload mass
- ES-L1, GEO, HEO, SSO orbits have higher success rate
- Not all orbits have success rate related to flight number
- Payload mass don’t have clear relationship with Orbit
- 4 launch sites are in CA and FL. 10 of 56 launches are in CA and the rest are in FL
- Booster version of FT, B4, B5 have higher success rate
- Decision Tree is more accurate than LR, SVM and KNN

Appendix

- All Jupyter Notebooks and python code have been uploaded to Github and their links have been pasted in this presentation.
- https://github.com/Aaron2014/IBM_Data_Science_Practice

Thank you!