Winning Space Race with Data Science

DATA SCIENCE CAPSTONE PROJECT

VALENTINE ZUH NJUNG

February 18, 2022

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- **Summary of all results**
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

Project background and context

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; The insurance rate on a Falcon 9 is about 4% currently, That's the same rate as competitors' similarly-capable rockets, such as the European launcher Arianespace's Ariane 5 or U.S. rocket builder United Launch Alliance's (ULA) Atlas V. But Ariane 5 and Atlas V launches go for upwards of $165 million each, meaning a Falcon 9 premium is about $2.5 million while its competitors' premiums would be in the range of $7 million.much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

Problems you want to find answers

-What factors determine if the rocket will land successfully?

-The interaction amongst various features that determine the success rate of a successful landing.

-What operating conditions needs to be in place to ensure a successful landing program

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - ❖ Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - ❖ One-hot encoding was applied to categorical features. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means Clustering Algorithm, all unique items under Venue Category are 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
  - ❖ tuning and evaluation of classification models to ensure the best results

Data Collection

- *Describe how data sets were collected.*

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API:

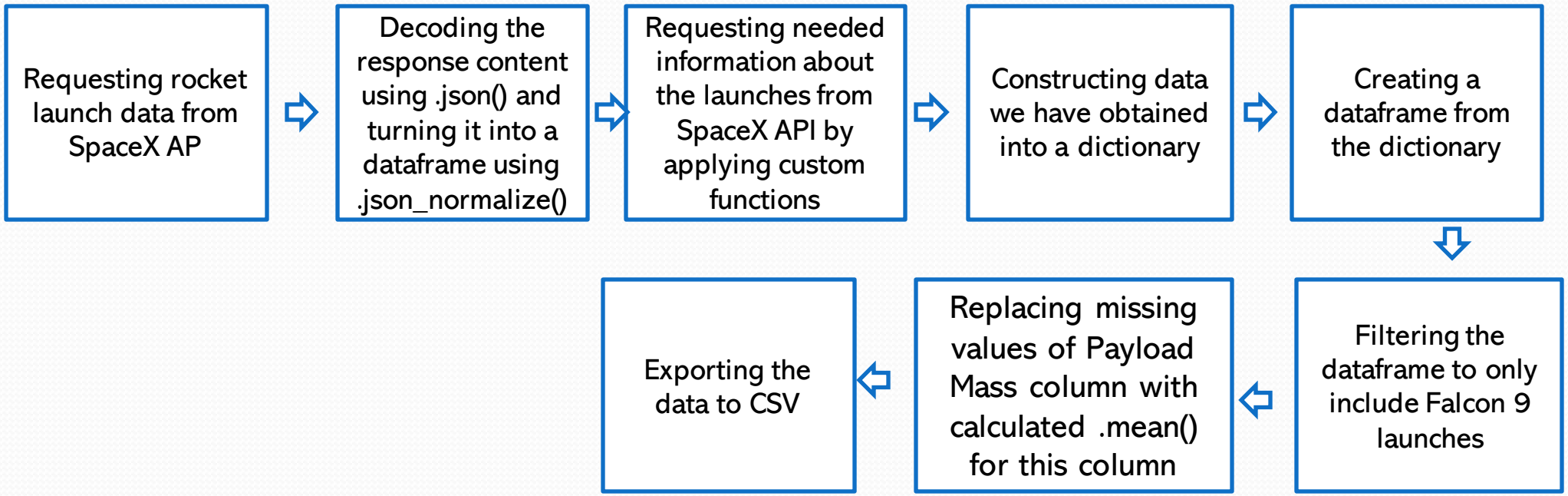
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

Data Columns are obtained by using Wikipedia Web Scraping:

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

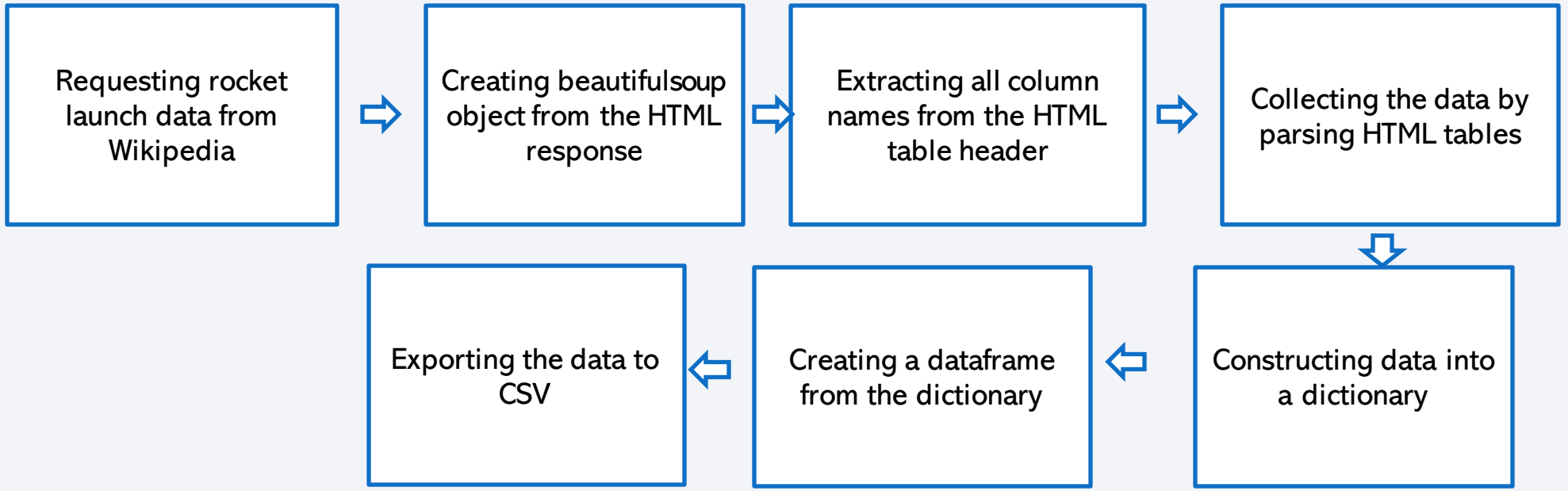
Data Collection – SpaceX API

Flowchart of Spacex API Calls



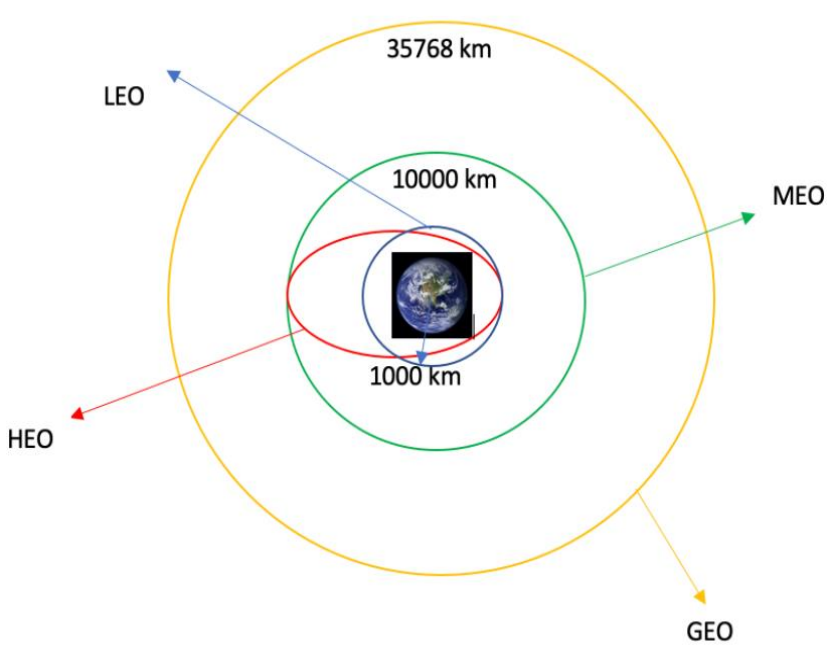
GitHub URL of the completed SpaceX API calls notebook: https://github.com/Njung-Valentine/DATA-SCIENCE-PROFESSIONAL-CERTIFICATE/blob/main/data%20collection%20using%20API.ipynb

Data Collection - Scraping



GitHub URL of the completed web scraping notebook: https://github.com/Njung-Valentine/DATA-SCIENCE-PROFESSIONAL-CERTIFICATE/blob/main/data%20collection%20using%20API.ipynb

Data Wrangling



We performed exploratory data analysis and determined the training labels

Calculate the number of lunches on each side

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome per orbit type

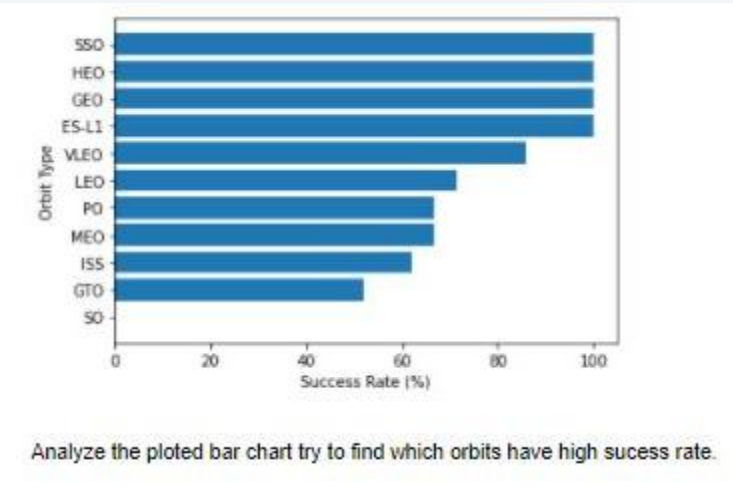
Create a landing outcome label from Outcome column

Exporting the data to CSV

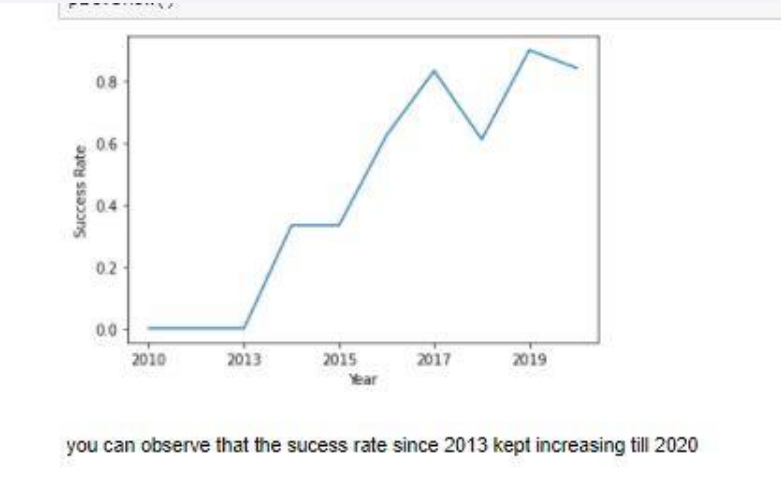
- GitHub URL of your completed data wrangling related notebooks:https://github.com/Njung-Valentine/DATA-SCIENCE-PROFESSIONAL-CERTIFICATE/blob/main/Data%20Wrangling%20.ipynb

EDA with Data Visualization

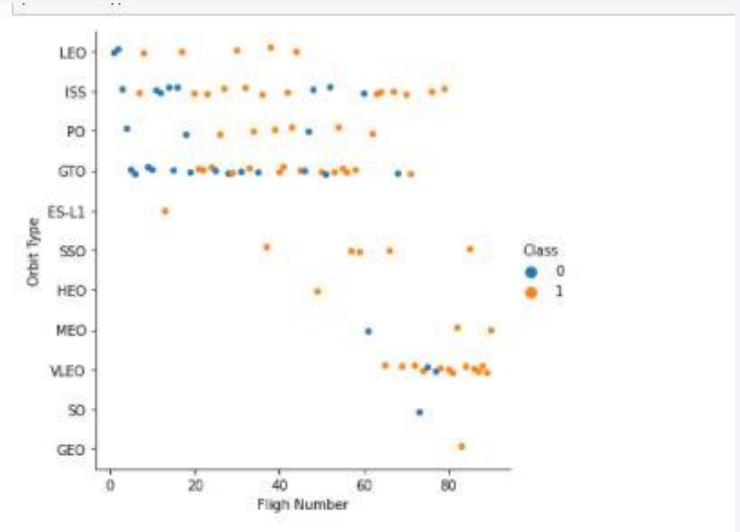
- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.



Bar chat shows comparison among discrete categories
This is to show the categories being compared and a measured value



Line chat show trends in data with respect to time



Scatter plot shows relationship between variables. If such a relationship exist then it can be use to build a machine learning model.

- GitHub URL: https://github.com/Njung-Valentine/DATA-SCIENCE-PROFESSIONAL-CERTIFICATE/blob/main/EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

Performed SQL queries

- The names of unique launch sites in the space mission.
- Displaying 5 records where lunch site begin with string 'CCA'
- The total payload mass carried by boosters launched by NASA (CRS)
- The 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*
- The total number of successful and failure mission outcomes
- *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*

Add the GitHub URL: https://github.com/Njung-Valentine/DATA-SCIENCE-PROFESSIONAL-CERTIFICATE/blob/main/DEA%20with%20SQL.ipynb

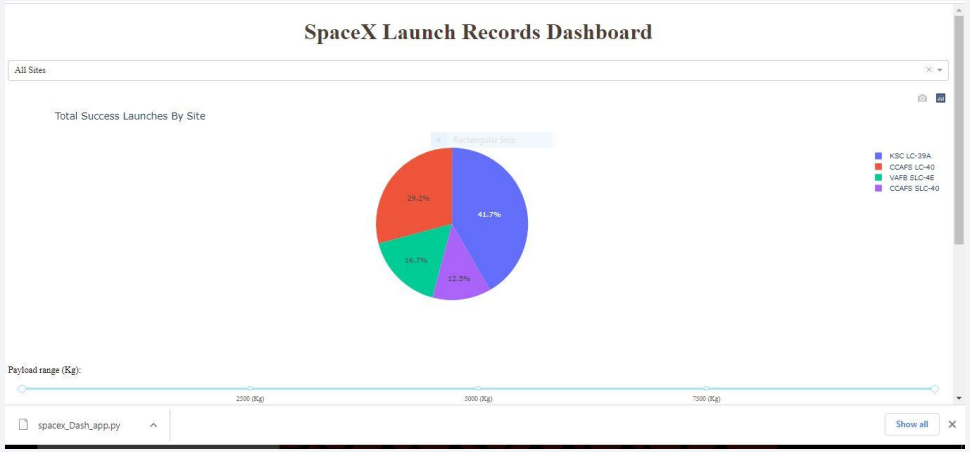
Build an Interactive Map with Folium

- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. Like railways, highways and coastlines.

Add the GitHub URL: https://github.com/Njung-Valentine/DATA-SCIENCE-PROFESSIONAL-CERTIFICATE/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab%20(1).ipynb

Build a Dashboard with Plotly Dash

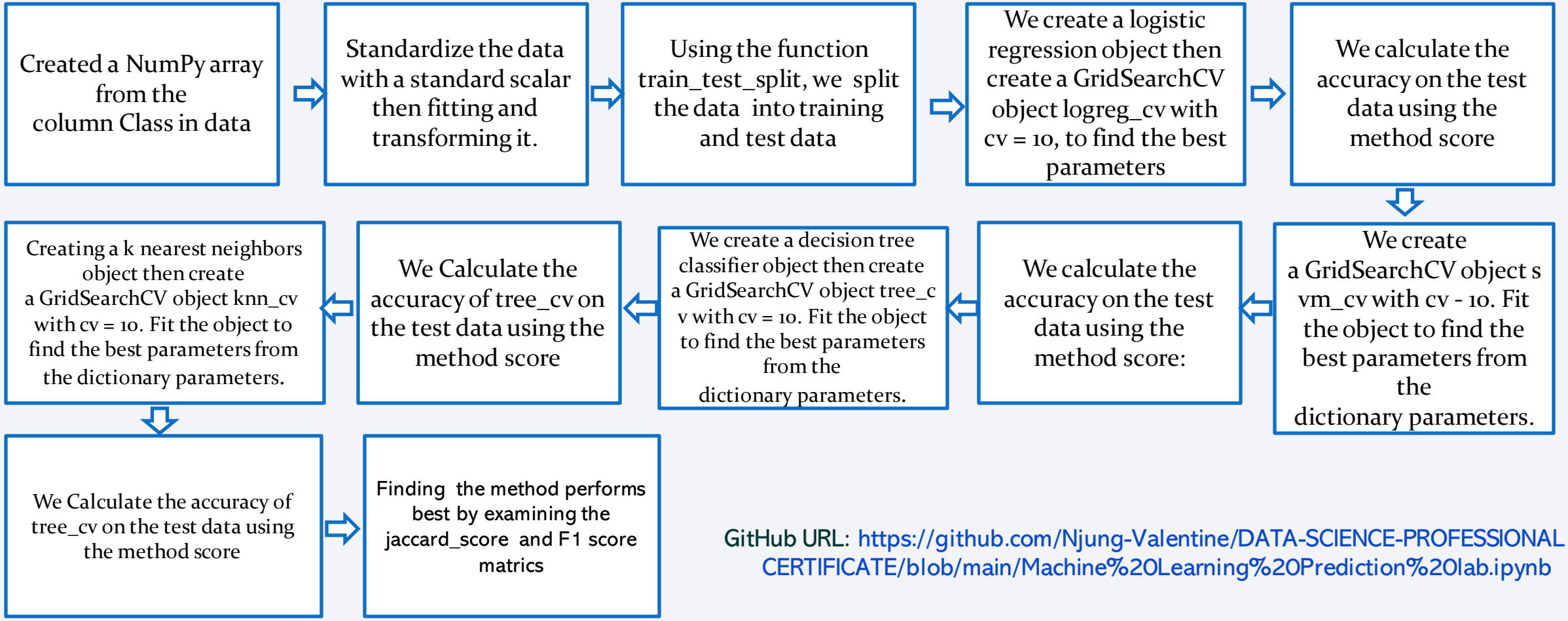
- We created a dropdown list to enable lunch site selection.
- We Added a pie chart to show total successful lunch count for all sites and the success vs failed count of a site if a particular lunch site was selected
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.



Add the GitHub URL: https://github.com/Njung-Valentine/DATA-SCIENCE-PROFESSIONAL-CERTIFICATE/blob/main/spacex_Dash_app%20(2).py

Create a k nearest neighbors object then create a `GridSearchCV` object `knn_cv` with cv = 10. Fit the object to find the best parameters from the dictionary `parameters`.

Predictive Analysis (Classification)



GitHub URL: https://github.com/Njung-Valentine/DATA-SCIENCE-PROFESSIONAL CERTIFICATE/blob/main/Machine%20Learning%20Prediction%20lab.ipynb

Results

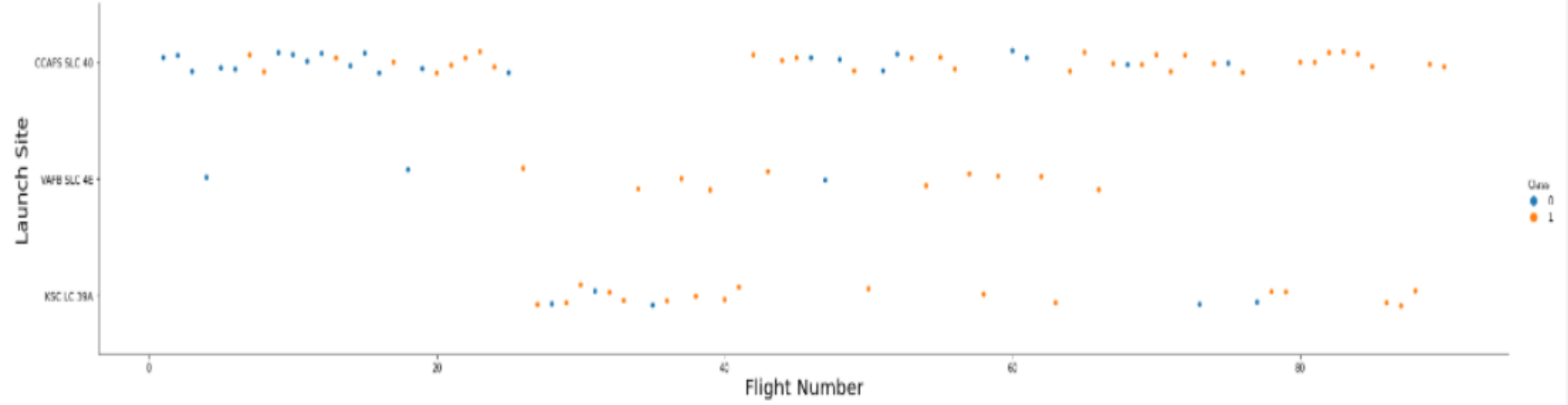
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

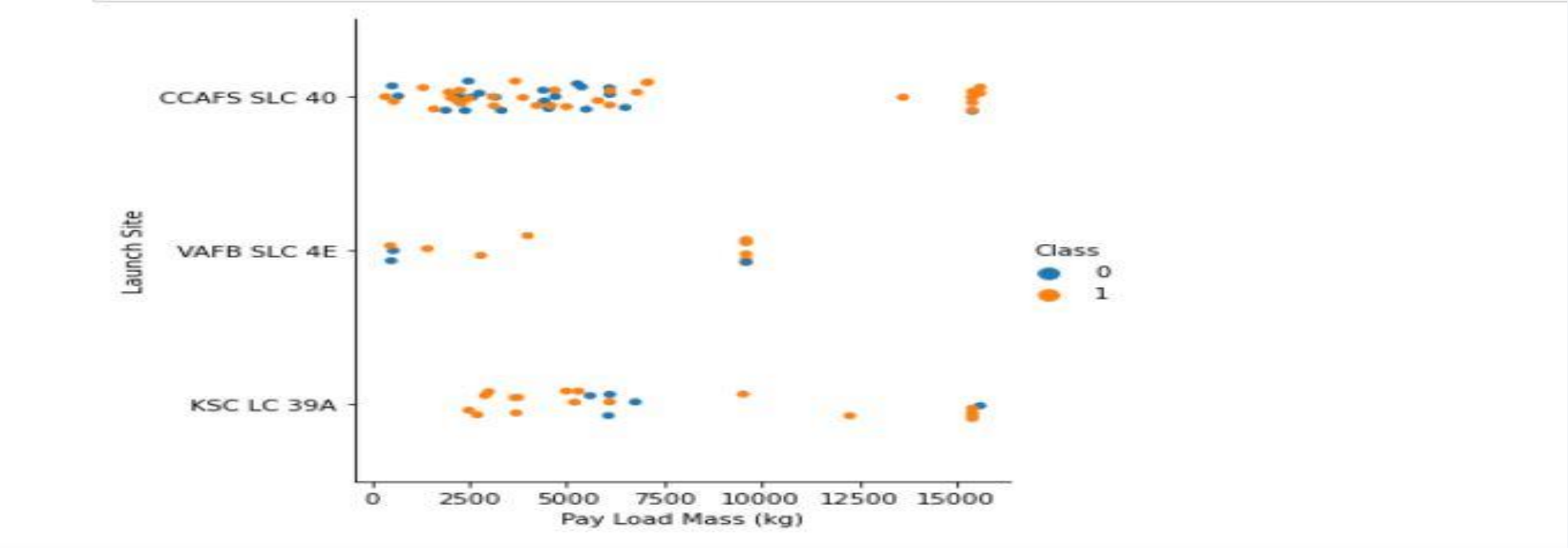
Launch Site vs Flight Number



- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.
- The earliest flights all failed while the later flights all succeeded

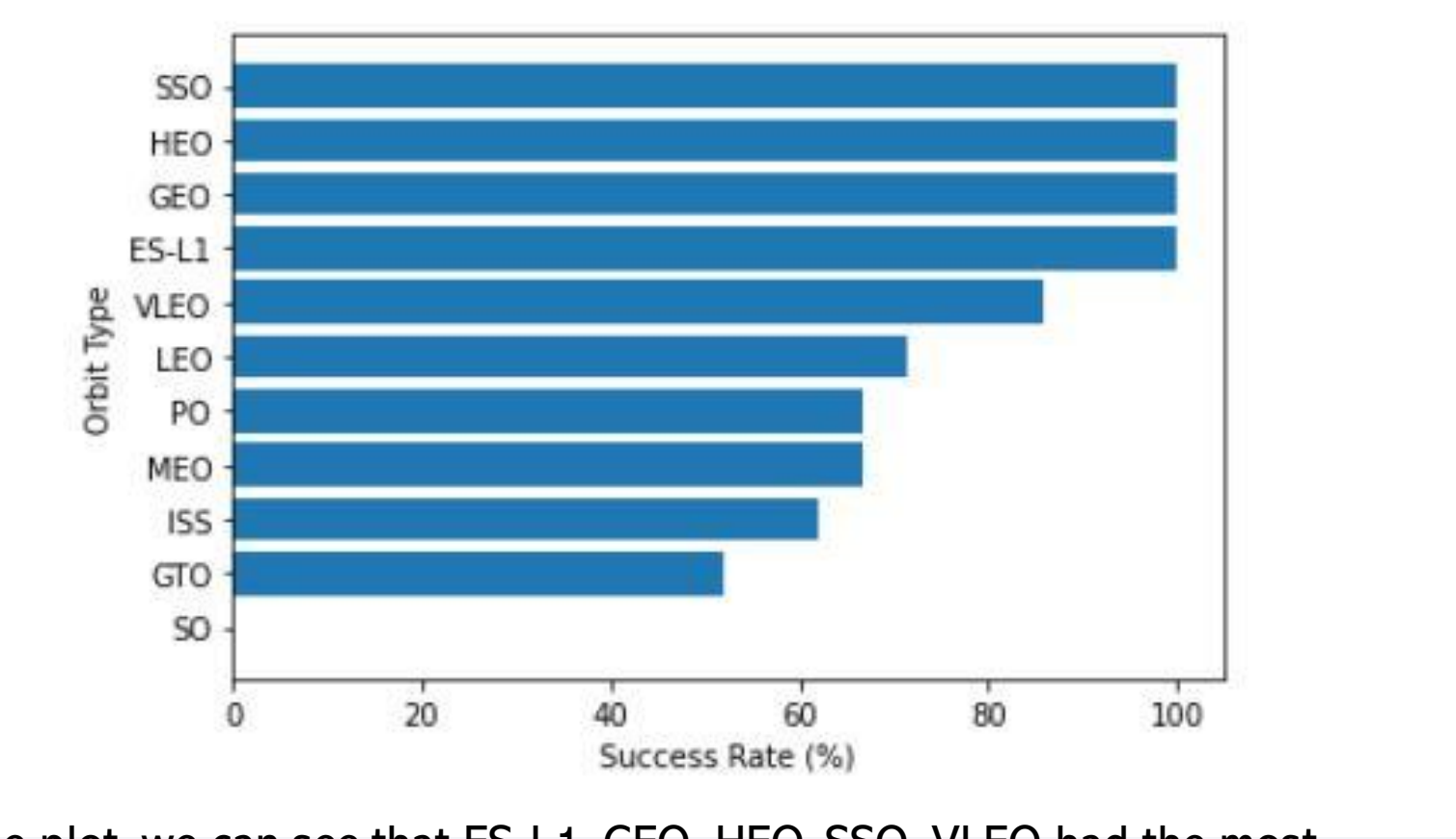
Payload vs. Launch Site

Payload vs. Launch Site



- For every lunch site, the higher the payload mass the higher success.
- Most of the lunches with payload mass over 7000 were successful

Success Rate vs. Orbit Type

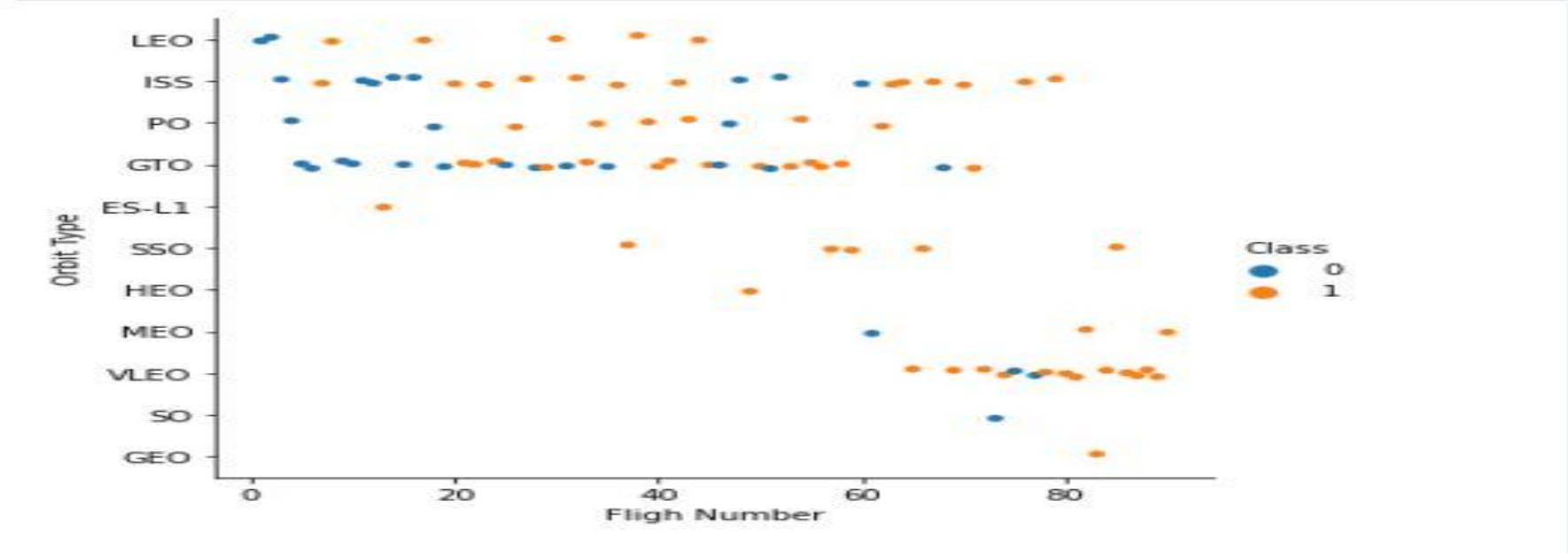


•From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

•Orbit with 0% success rate SO

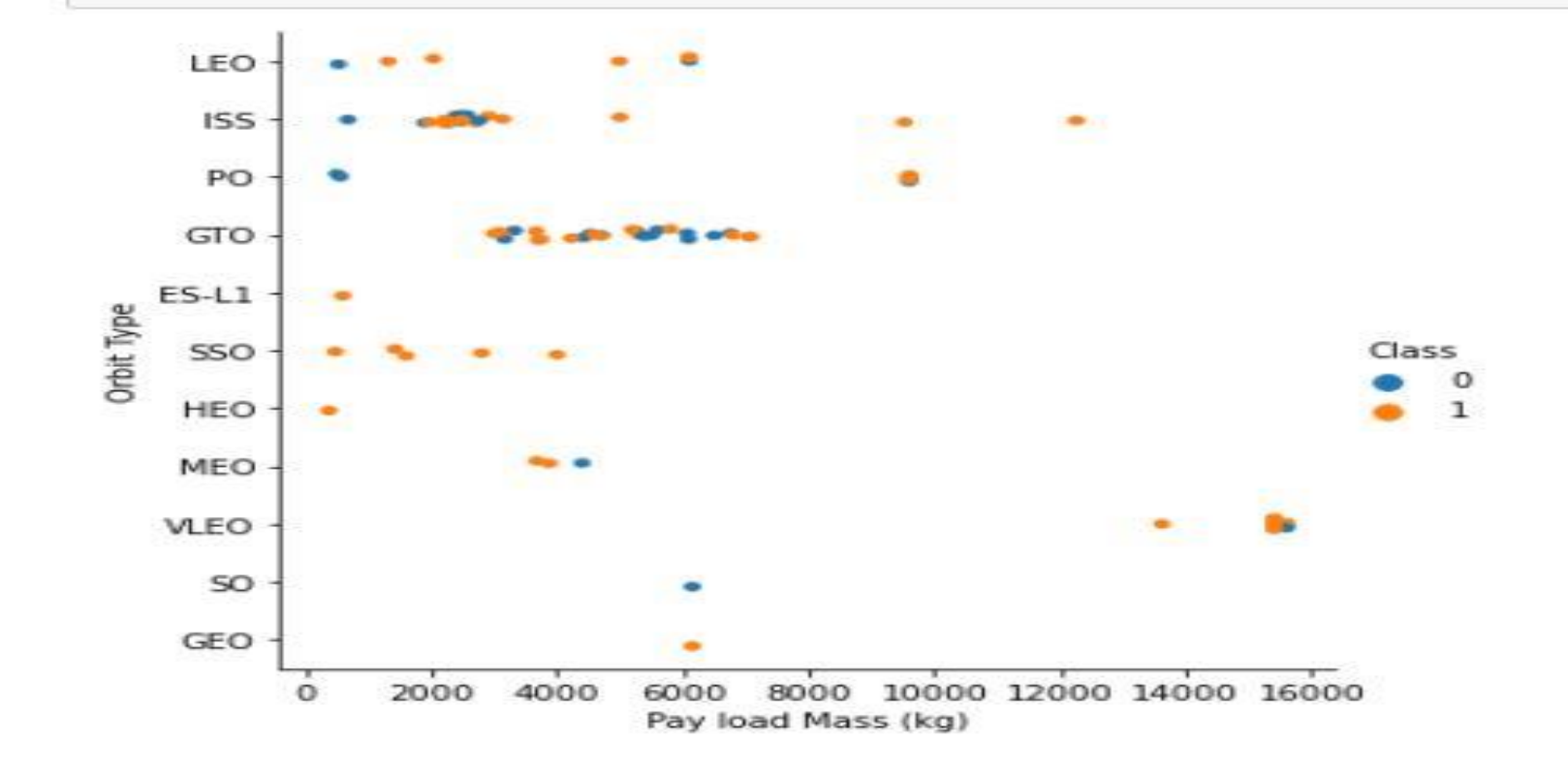
Flight Number vs. Orbit Type

Orbit Type vs Flight Number



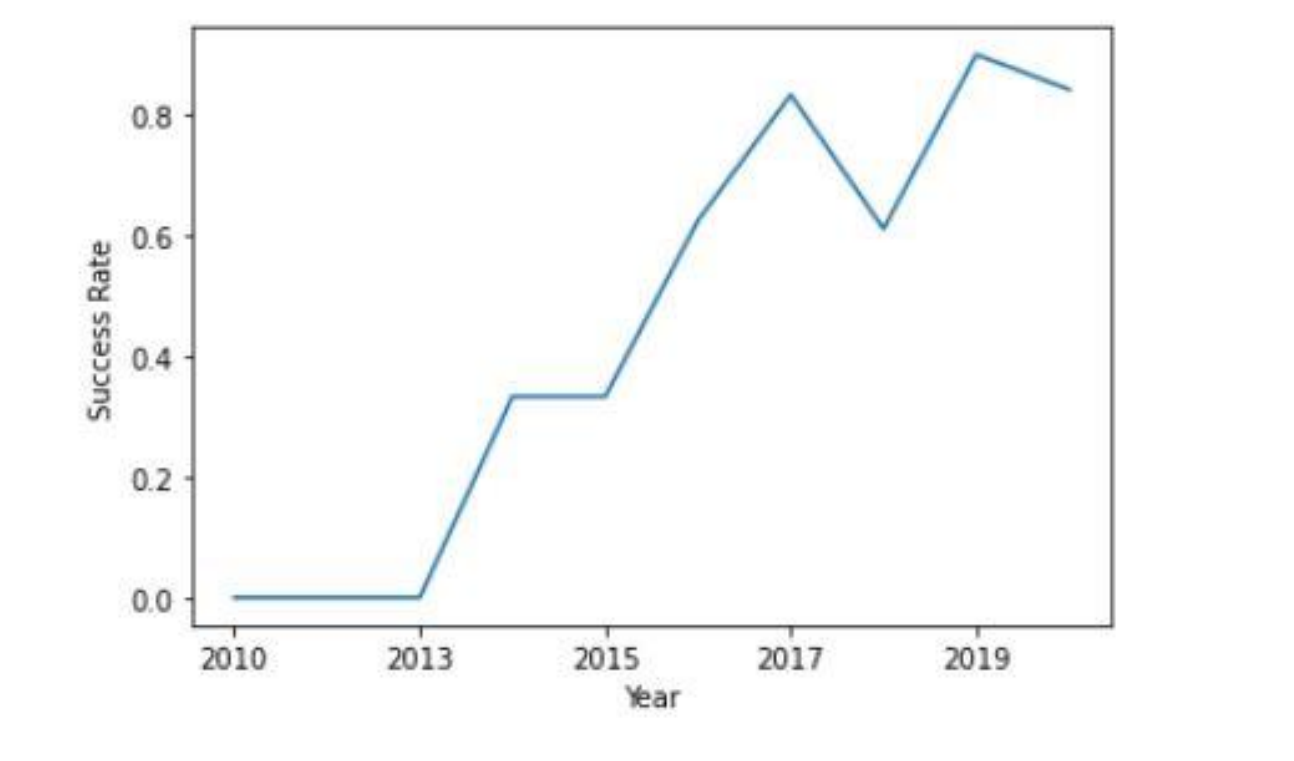
We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

Payload vs. Orbit Type



• with heavy payloads have negative influence on GTO orbit and positive on polar LEO orbits. the successful landing are more for PO, LEO and ISS orbits.

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 20120

All Launch Site Names

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

In [4]:
```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

```
In [5]: %%sql
        SELECT * FROM SPACEXTBL
        WHERE LAUNCH_SITE LIKE 'CCA%'
        LIMIT 5
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

Out[5]:

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

- We used the query above to display 5 records where launch sites begin with 'CCA'

Total Payload Mass

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

In [6]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_kg
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

Out[6]:

total_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

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

In [7]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass_kg
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

Out[7]:

avg_payload_mass_kg
2928

First Successful Ground Landing Date

In [21]:

```
%%sql
SELECT MIN(DATE) AS first_successful_landing_date
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

Out[21]:

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload 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

In [22]:
```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

Out[22]:

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [14]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS total_number
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

Out[14]:

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

Boosters Carried Maximum Payload

In [15]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

Out[15]:

booster_version	payload_mass__kg_
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

2015 Launch Records

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

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = '2015'
```

```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

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

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

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS total_number
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY total_number DESC
```

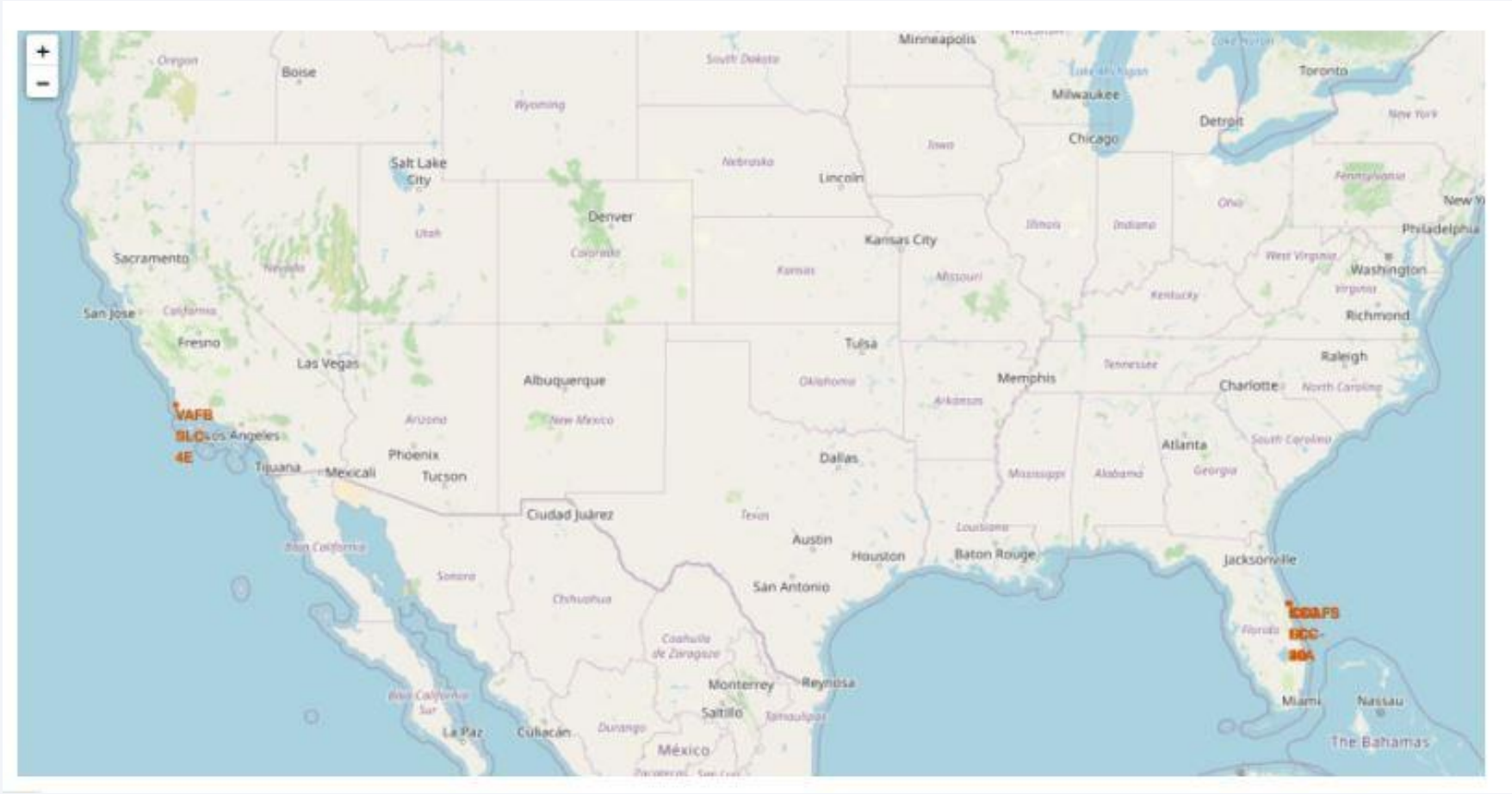
```
 * ibm_db_sa://xsb78386:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/bludb
Done.
```

landing__outcome	total_number
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

Section 3

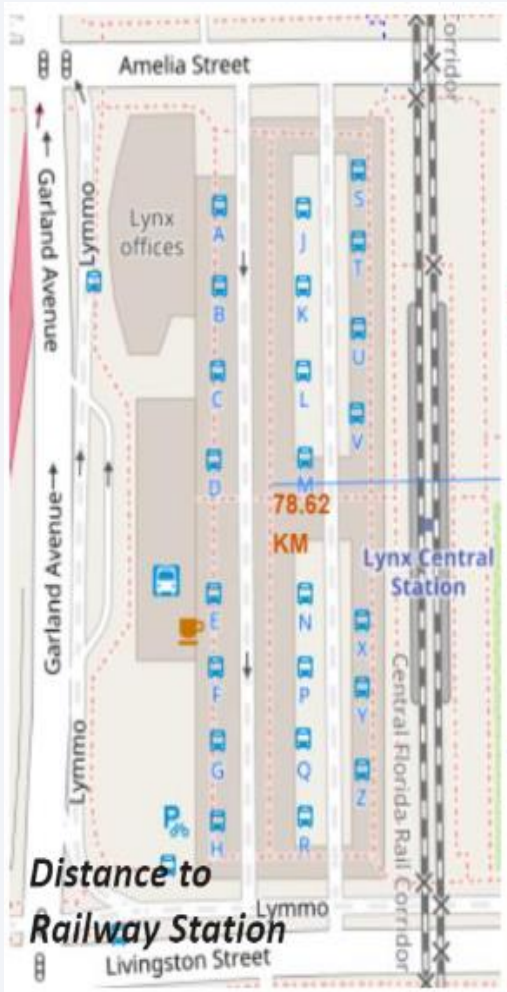
Launch Sites Proximities Analysis

All launch sites global map markers



Markers showing launch sites with color labels

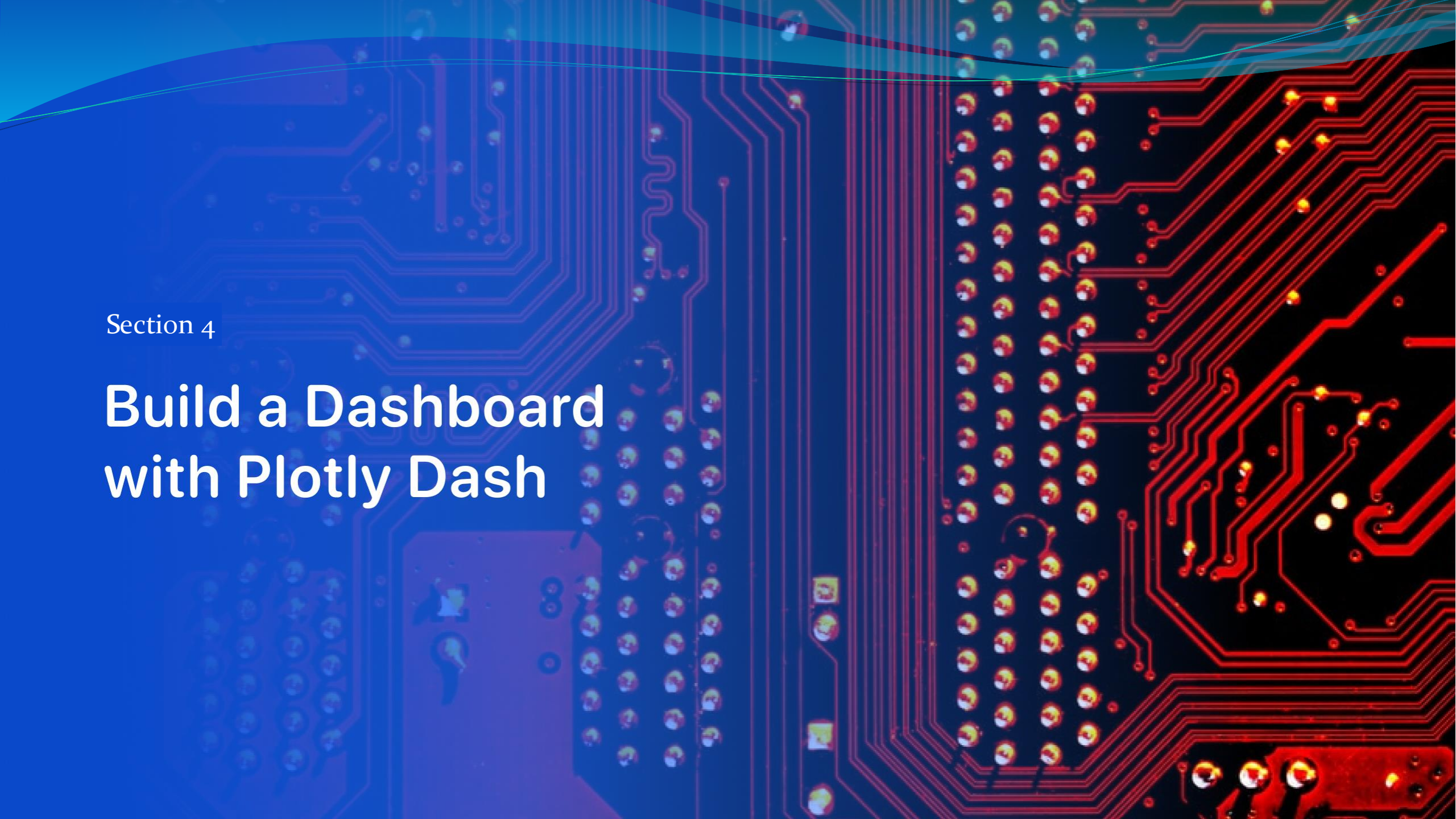
Launch Site distance to landmarks



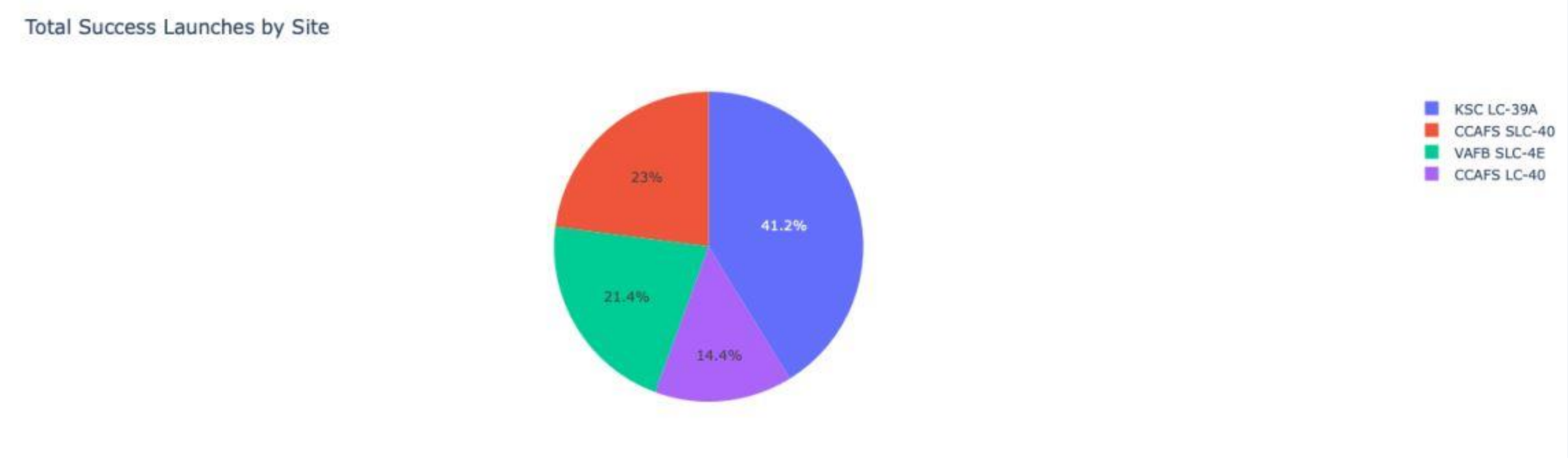
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 4

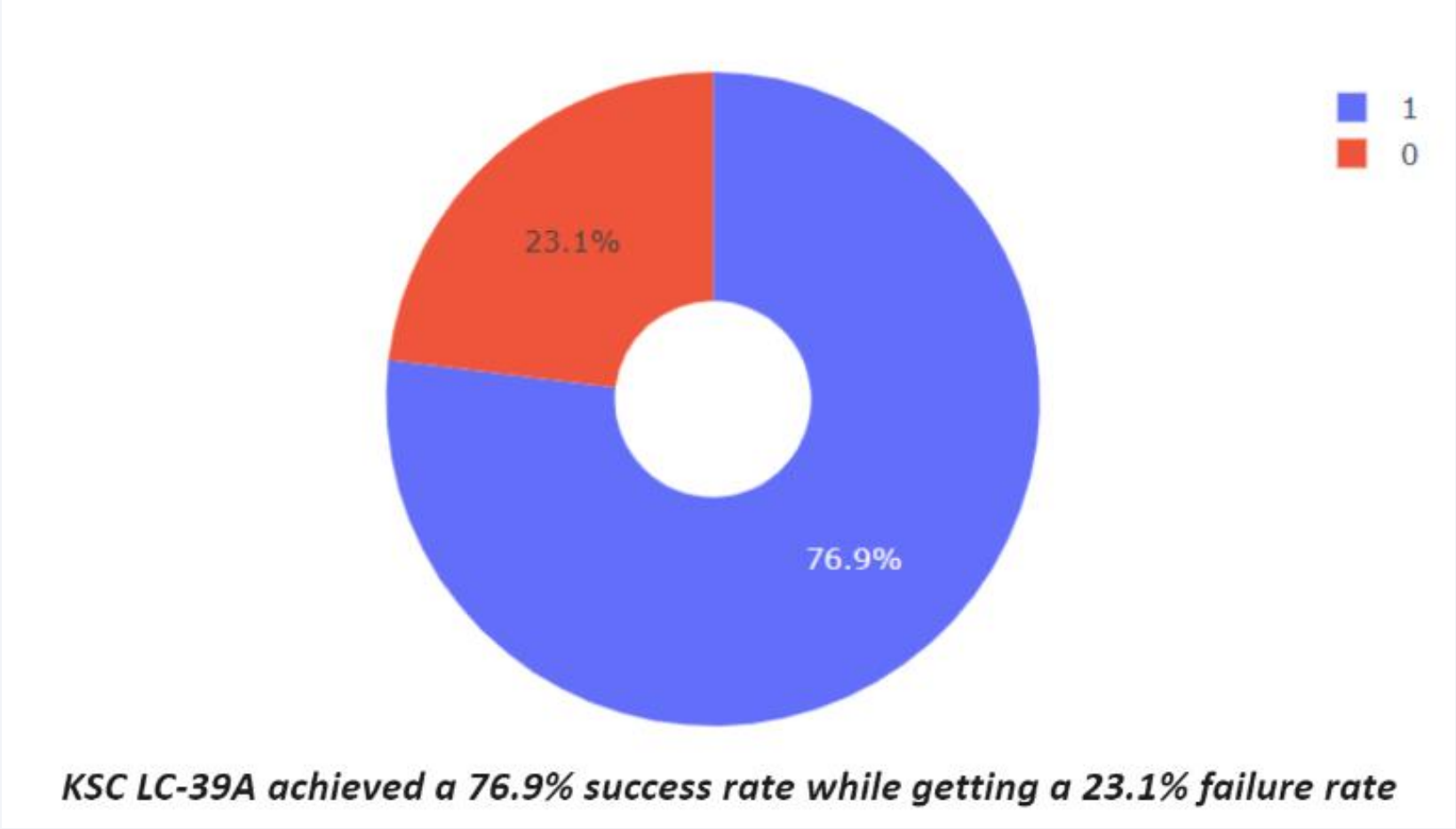
Build a Dashboard with Plotly Dash



Pie chart showing the success percentage achieved by each launch site

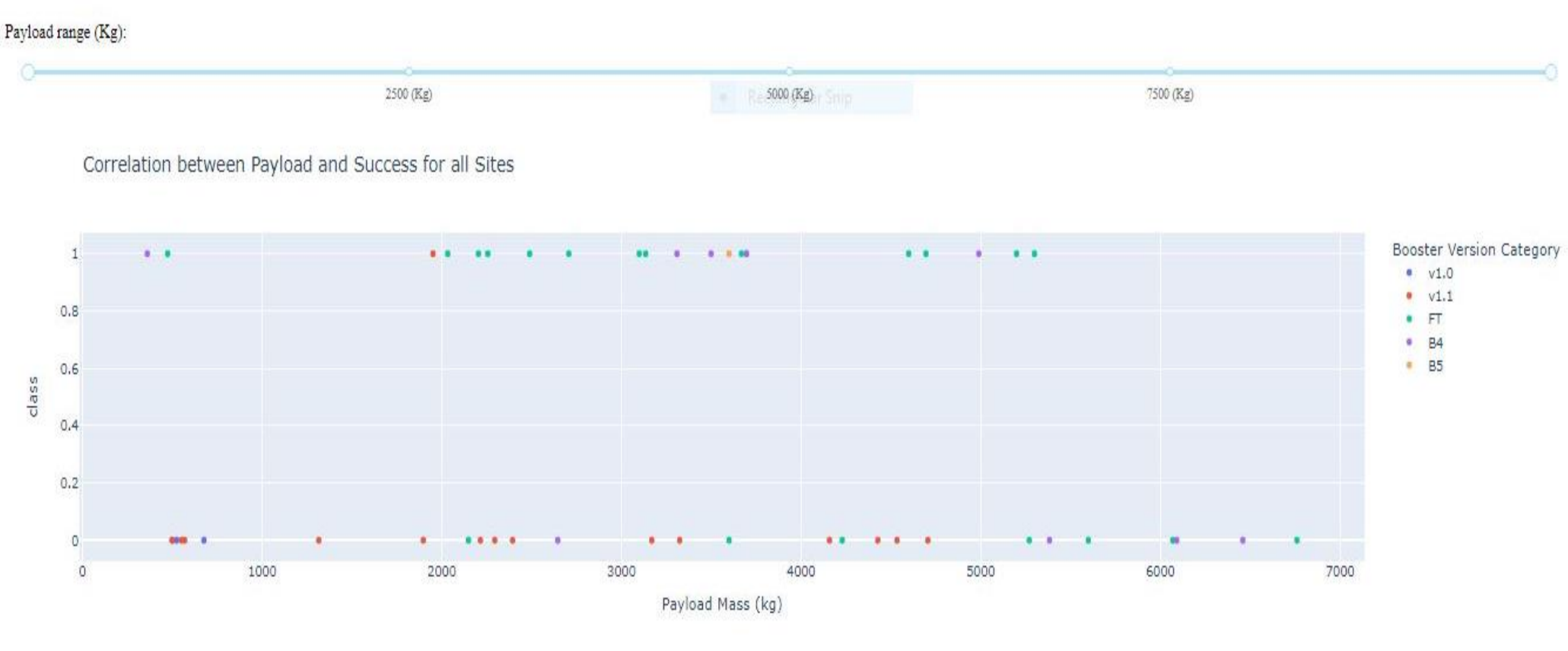


Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

correlation between payload mass and lunch outcome for each site



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and accuracy of test set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

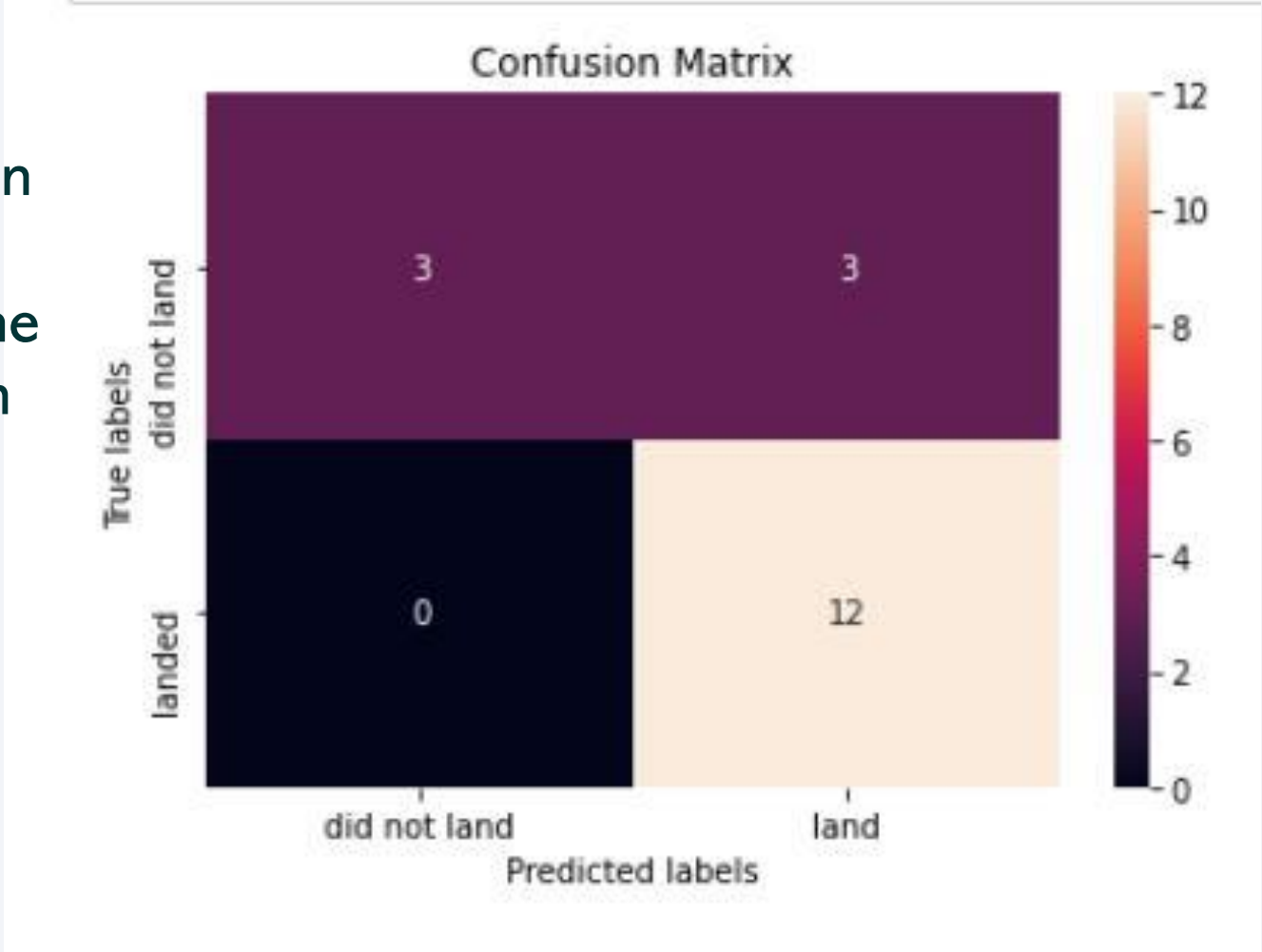
Scores and accuracy of the entire data set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

The scores of the entire data set confirms that the best model is the decision tree model. This model has both higher score and accuracy .

Confusion Matrix

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

- The decision tree model is the best algorithm for this dataset.
- Launches with low payload mass show better results than launches with larger payloads..
- Most of launches are in proximity to the equator line and all the sites are in very close proximity to the coast.
- Over the years there is an increase the success rate of the launches.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites

Thank you!