IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name: Tung Tran>
<Date: 2024-09-29>

Outline

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

Executive Summary

- Summary of methodologies
    - Data Collection
    - Data Wrangling
    - Exploratory Analysis (SQL & Matplotlib)
    - Interactive Analytics (Folium & Plotly dash)
    - Predictive Analysis (Classiffication techniques)
- Summary of all results
    - This presentation summarizes the data collection, exploratory analysis and predictive analysis results associated with SpaceX Falcon 9 case study. The purpose of this presentation is to let readers understand how the data collected and cleansed, then some primary correlation of features will be inferred by corresponding data through charts and interactive analytics; finally, classiffication techniques used during the predictive analysis to train and test model.

Introduction

- Project background and context
  - SpaceX 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 SpaceX 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 SpaceX for a rocket launch.
- Problems you want to find answers
  - In this presentation, readers will be walked through on how the author predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Request api and webscraping
- Perform data wrangling
  - Convert outcomes into training labels (1 and 0)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearchCV to find the best estimator and parameter with several methods such as KNN, SVM, Logistic Regression, etc.

Data Collection

- How data sets were collected.
  - Calling API
  - Webscraping from wikipedia
- Data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Github URL: https://github.com/Tung-depressedsuperman/IBM-DataScience_SpaceX-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Request API → Check status code → Decode the response contentas a json

Data Collection - Scraping

- Github URL: https://github.com/Tung-depressedsuperman/IBM-DataScience_SpaceX-project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- How data were processed
    - In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident. Therefore, we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed, and 0 means it was unsuccessful.
- GitHub URL: https://github.com/Tung-depressedsuperman/IBM-DataScience_SpaceX-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Count values of the outcome column → Group bad and good outcomes into 2 different groups → Create outcome label (0 & 1) from outcome column

EDA with Data Visualization

- Charts were plotted and why used those charts
  - Scatter plot. In order to understand the relationships of 2 or more features.
  - Bar chart. In order to compare the success rate of different orbit type.
  - Line chart. In order to know the launch success yearly trend over period of time.
- GitHub URL: https://github.com/Tung-depressedsuperman/IBM-DataScience_SpaceX-project/blob/main/edadataviz.ipynb

EDA with SQL

- The SQL queries performed
  - select distinct Launch_Site from SPACEXTBL
  - select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
  - select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer == 'NASA (CRS)'
  - select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version == 'F9 v1.1'
  - select min(Date) from SPACEXTBL where Landing_Outcome == 'Success (ground pad)'
  - select Booster_Version, PAYLOAD_MASS__KG_, Landing_Outcome from SPACEXTBL where Landing_Outcome == 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4001 and 5999
  - select Landing_Outcome, count(*) from SPACEXTBL where Landing_Outcome like 'Success%' or Landing_Outcome like 'Failure%' group by Landing_Outcome
  - select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
  - select substr(Date,6,2) as 'Month', Date, Booster_Version, Landing_Outcome, Launch_Site from SPACEXTBL where Landing_Outcome == 'Failure (drone ship)' and substr(Date,1,4) == '2015'
  - select Landing_Outcome, count(*) as Total_outcome from SPACEXTBL where Landing_Outcome == 'Success (ground pad)' or Landing_Outcome like 'Failure (drone ship)' and Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Total_outcome desc
- GitHub URL: https://github.com/Tung-depressedsuperman/IBM-DataScience_SpaceX-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map objects are created and added to a folium map
  - Circle. To add a highlighted circle area.
  - Marker. To mark on specific location with text label
  - MarkerCluster. To group all the markers which point to the same location into one cluster.
  - PolyLine. To draw a line beween 2 locations.
- GitHub URL: https://github.com/Tung-depressedsuperman/IBM-DataScience_SpaceX-project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions have been added to a dashboard
  - Pie chart. To compare the success percentages of all sites.
  - Scatter chart. To understand the correlation between Payload and Class.
- GitHub URL: https://github.com/Tung-depressedsuperman/IBM-DataScience_SpaceX-project/blob/main/spacex_dash_app.py

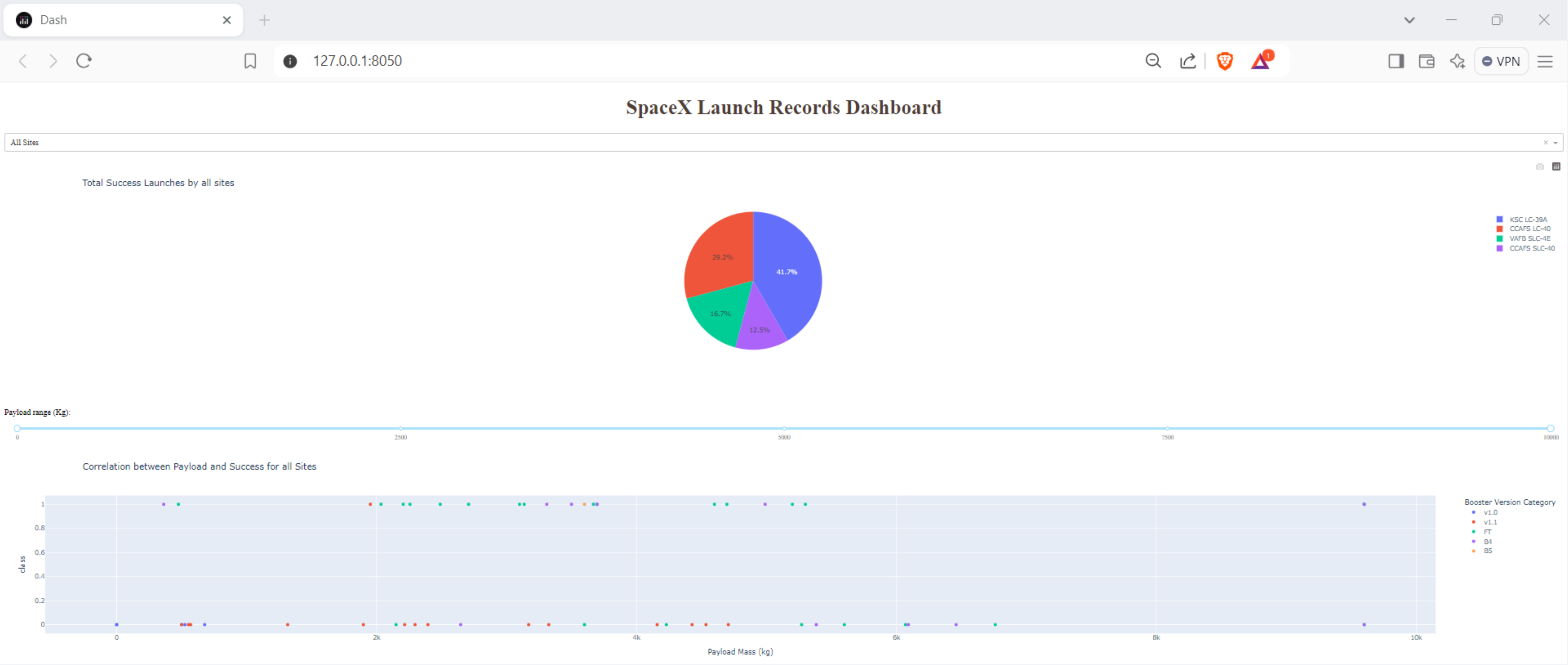
Predictive Analysis (Classification)

- GitHub URL: https://github.com/Tung-depressedsuperman/IBM-DataScience_SpaceX-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Create numpy array from column 'Class' → Standardize the data → Split the data into train, test data → Create GridSearchCV object → Create Logistic Regression, SVM, DecisionTreeClassifier, KNN object → Fit the training data with different parameters → Calculate the score of the best estimator for each method → Draw the confusion matrix → Compare all scores of all methods

Results

- Predictive analysis results: Logistic Regression and Decision Tree Classifier provide the best score.
- Exploratory data analysis results

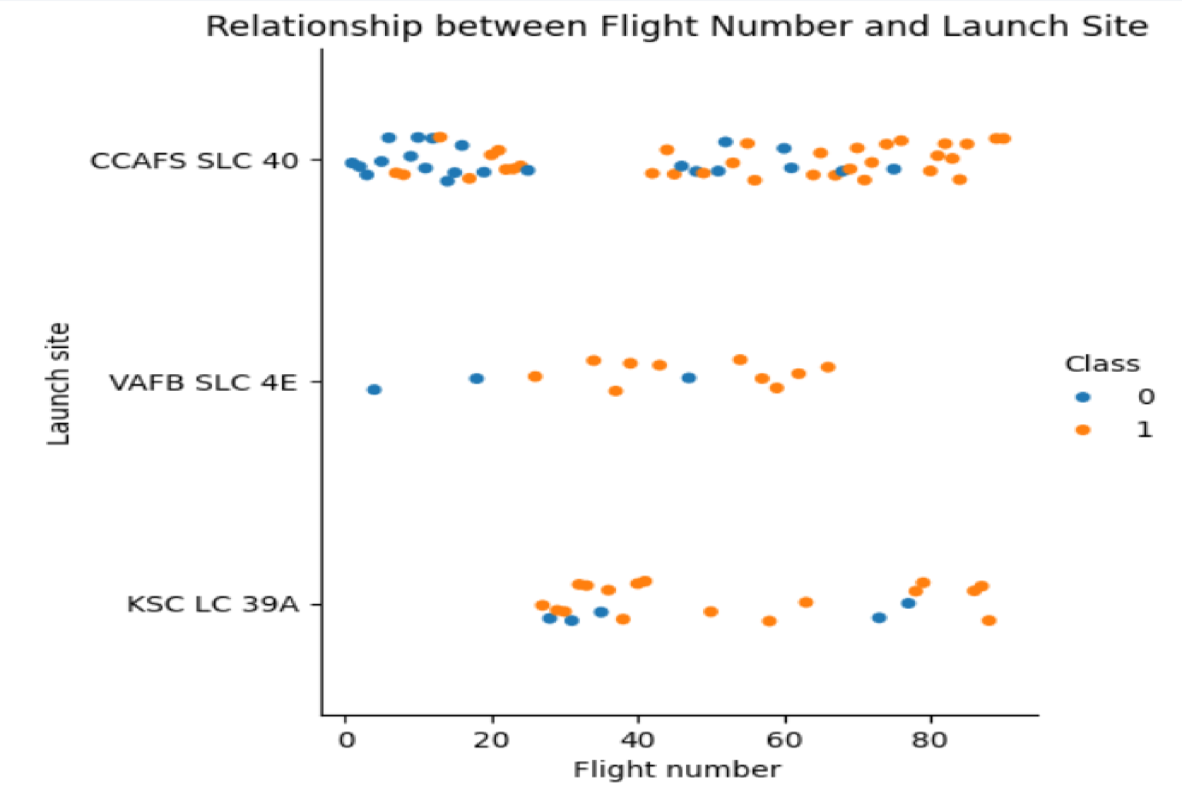


Section 2

Insights drawn from EDA

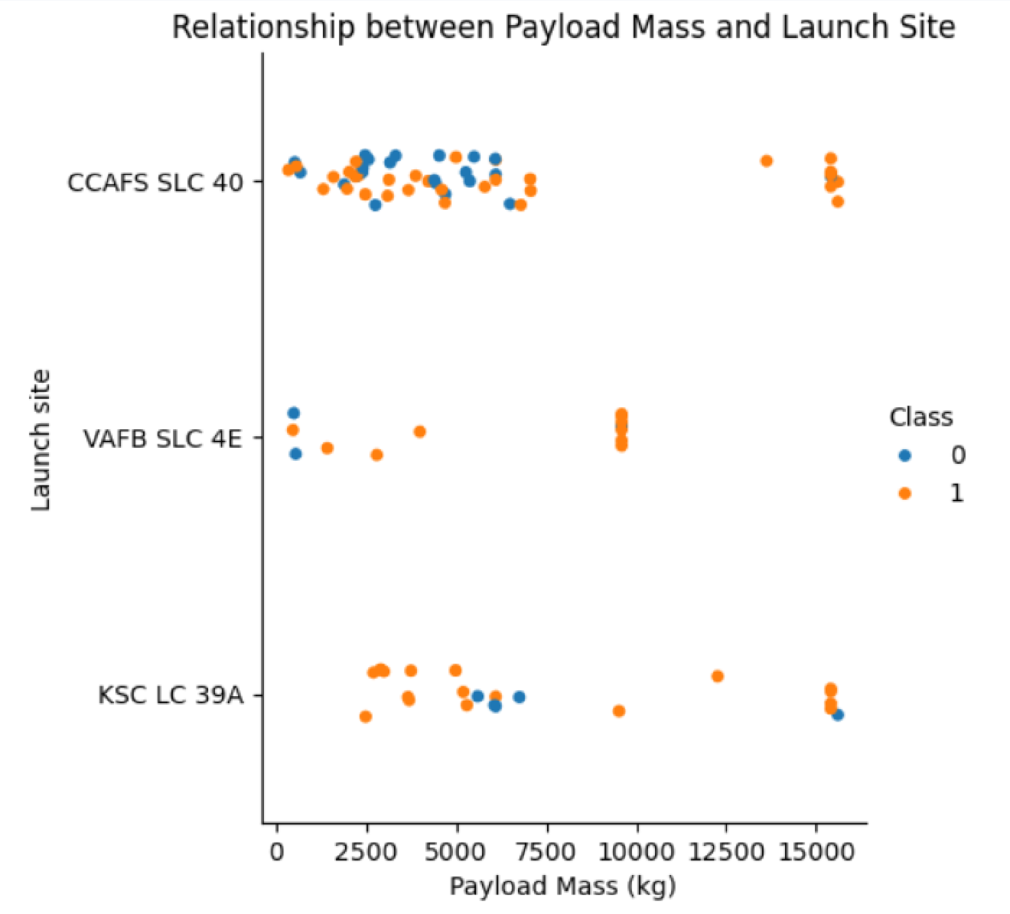
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site
  - Overall, the more rockets were launched at site VAFB SLC 4E and KSC LC 39A, the more successful the flight gained.
  - The above result is not totally correct with site CCAFS SLC 40. There are still unsuccessful flight, when flying more rockets at this site.

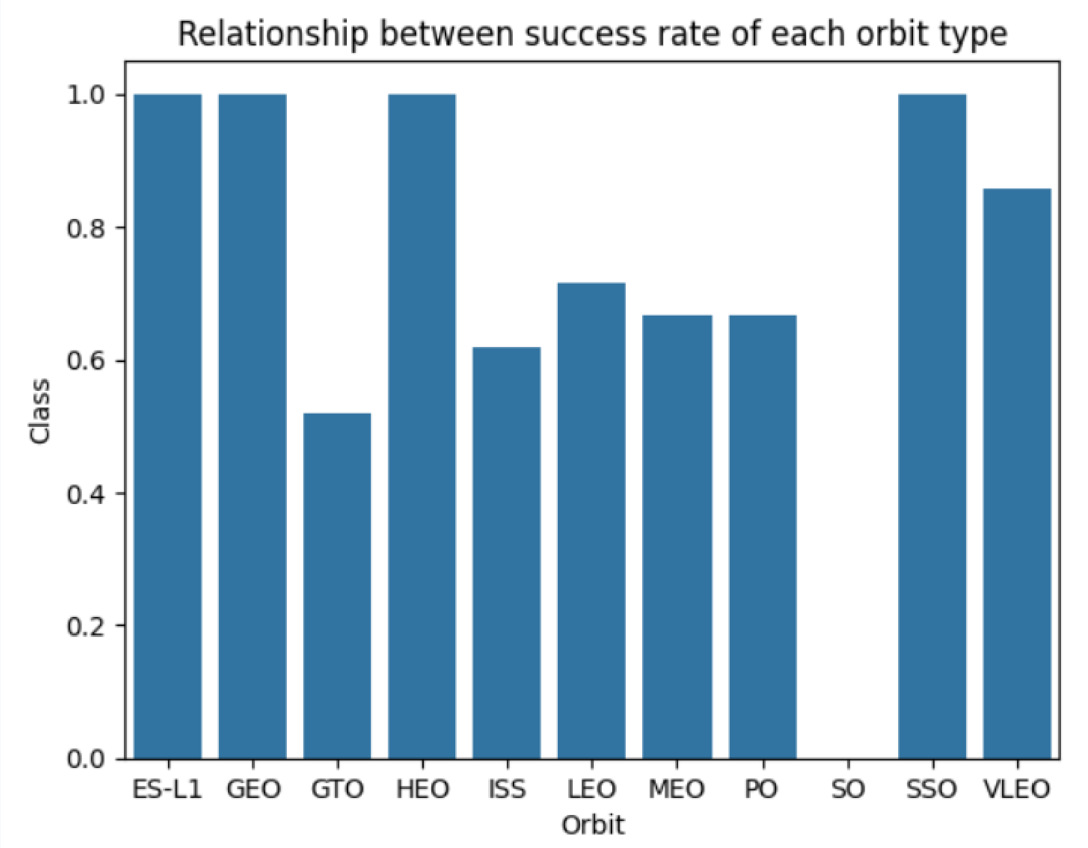


Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site
  - When the payload mass was getting heavier, the successful flight was getting more often at CCAFS SLC40 and VAFB SLC 4E sites.
  - Although the successful rate also got higher while payload getting heavier at KSC LC 39A, a few of unsuccessful flights occurred.



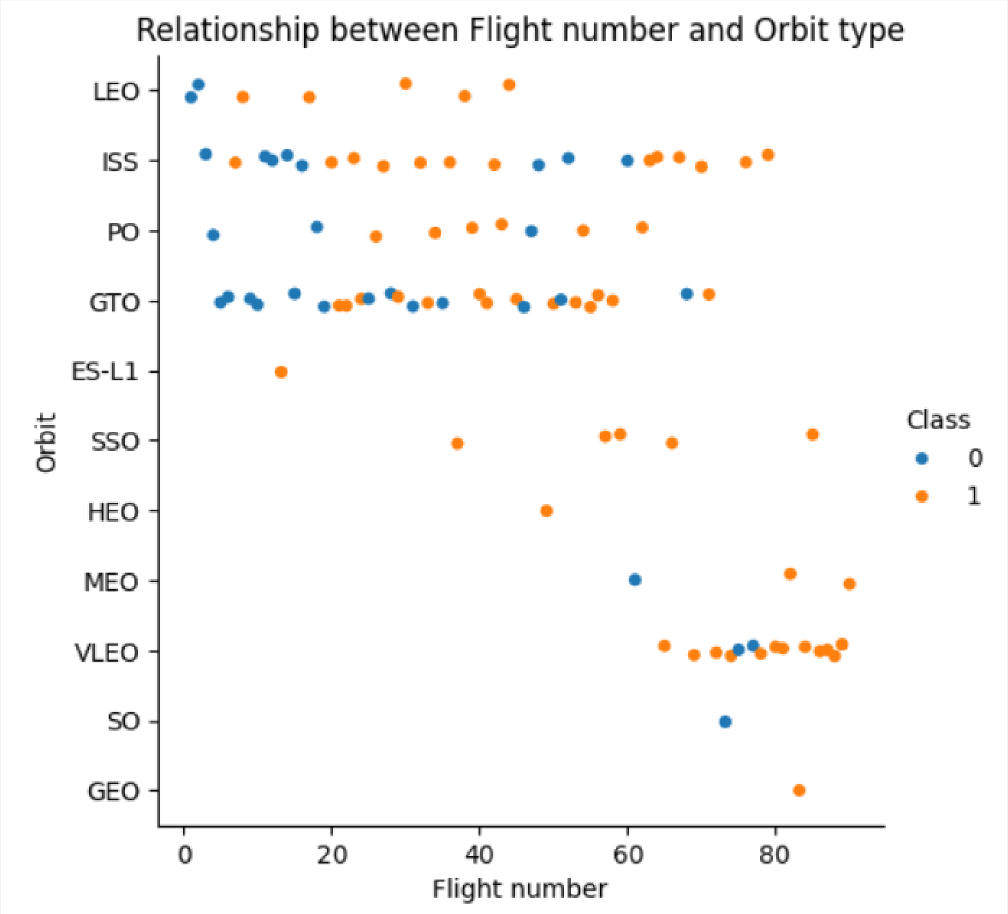
Success Rate vs. Orbit Type



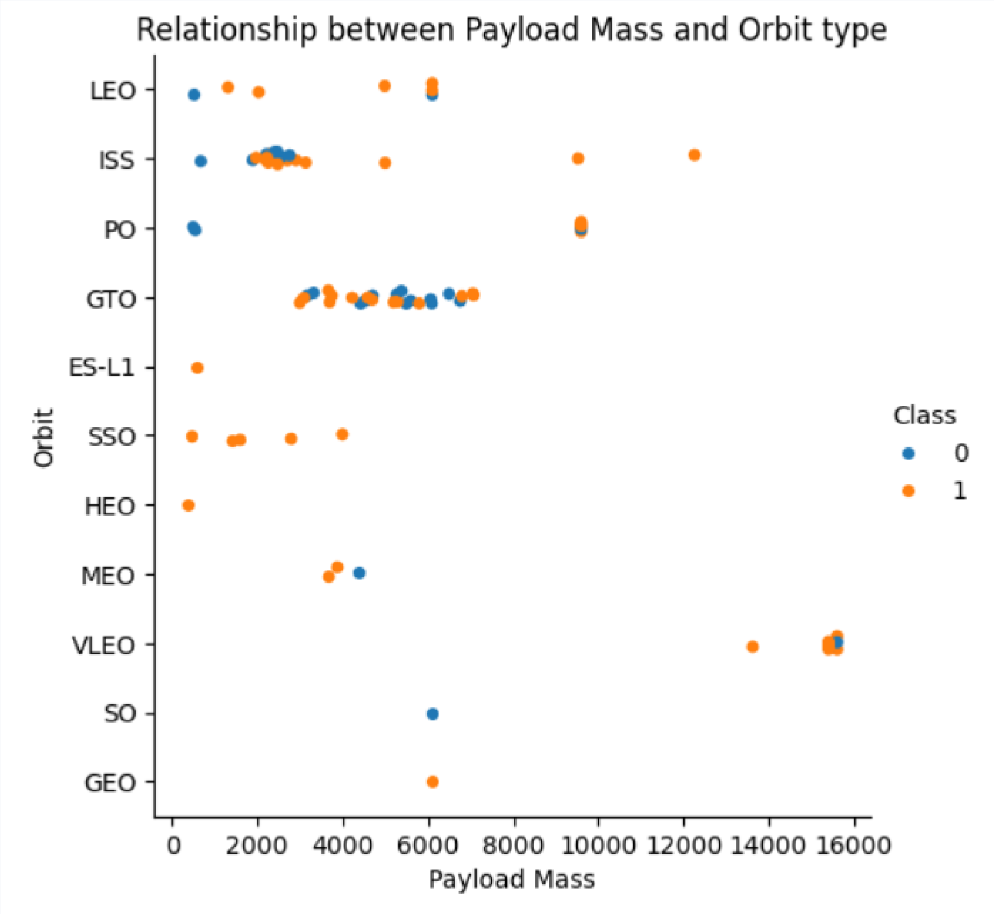
- Bar chart for the success rate of each orbit type
  - ES-L1, GEO, HEO, SSO and VLEO had the successful rate quite high
  - The rest stayed stably at around 0.6
  - The orbit type SO is the only one got none successful flight.

Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type
  - For LEO, ISS and VLEO orbit types, the more flight were launched, the more successful rate got.
  - Although the success rate of HEO, ES-L1 and GEO got maximum 1, number of flight were launched not much. Just only 1 flight was launched for each orbit.



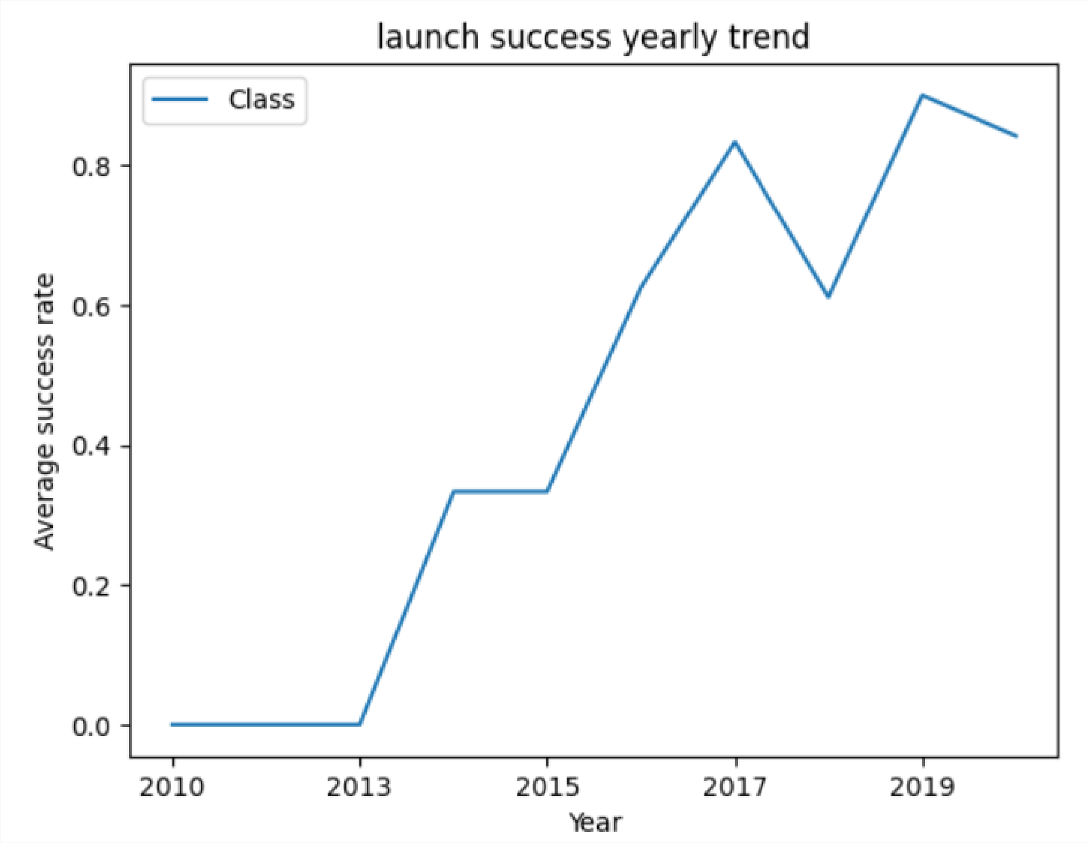
Payload vs. Orbit Type



- Scatter point of payload vs. orbit type
  - Overall, the relationship between payload mass and orbit type is not very clear.
  - For most of the orbit types, the number of successful and unsuccessful flights were recorded quite the same.

Launch Success Yearly Trend

- Line chart of yearly average success rate
  - Overall, the launch success rate increased significantly since 2013.
  - There was a small drop in 2018. And after that, the rate raised up fast.



All Launch Site Names

- There are 4 unique launch sites in total

In [10]:

```
%sql select distinct Launch_Site from SPACEXTBL
```

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

In [12]:
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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

Out[12]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Total Payload Mass

- The total payload carried by boosters from NASA (CRS) is 45596 kg

```
In [13]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer == 'NASA (CRS)'

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

Out[13]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.40 kg

In [14]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version == 'F9 v1.1'
```

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

Out[14]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first successful date of landing outcome on ground pad is 2015-12-22

In [15]:
```
%sql select min(Date) from SPACEXTBL where Landing_Outcome == 'Success (ground pad)'
```

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

Out[15]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below figure lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
In [18]: %sql select Booster_Version, PAYLOAD_MASS__KG_, Landing_Outcome from SPACEXTBL where Landing_Outcome == 'Success (drone shi

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

Out[18]:

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are listed in the figure below

In [31]:
```
me, count(*) from SPACEXTBL where Landing_Outcome like 'Success%' or Landing_Outcome like 'Failure%' group by Landing_Outcome
```

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

Out[31]:

Landing_Outcome	count(*)
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
Success	38
Success (drone ship)	14
Success (ground pad)	9

Boosters Carried Maximum Payload

- Below figure list the names of the booster that have carried the maximum payload mass which is 15600 kg

In [40]:

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from
```

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

Out[40]:

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

- There are 2 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [47]:

```
%sql select substr(Date,6,2) as 'Month', Date, Booster_Version, Landing_Outcome, Launch_Site from SPACEXTBL where Landing_Ou
```

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

Out[47]:

Month	Date	Booster_Version	Landing_Outcome	Launch_Site
01	2015-01-10	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	2015-04-14	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

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

- There are 5 Failure (drone ship) and 9 Success (ground pad) landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

In [59]:
```
%sql select Landing_Outcome, count(*) as Total_outcome from SPACEXTBL where Landing_Outcome == 'Success (ground pad)' or Lar
```

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

Out[59]:

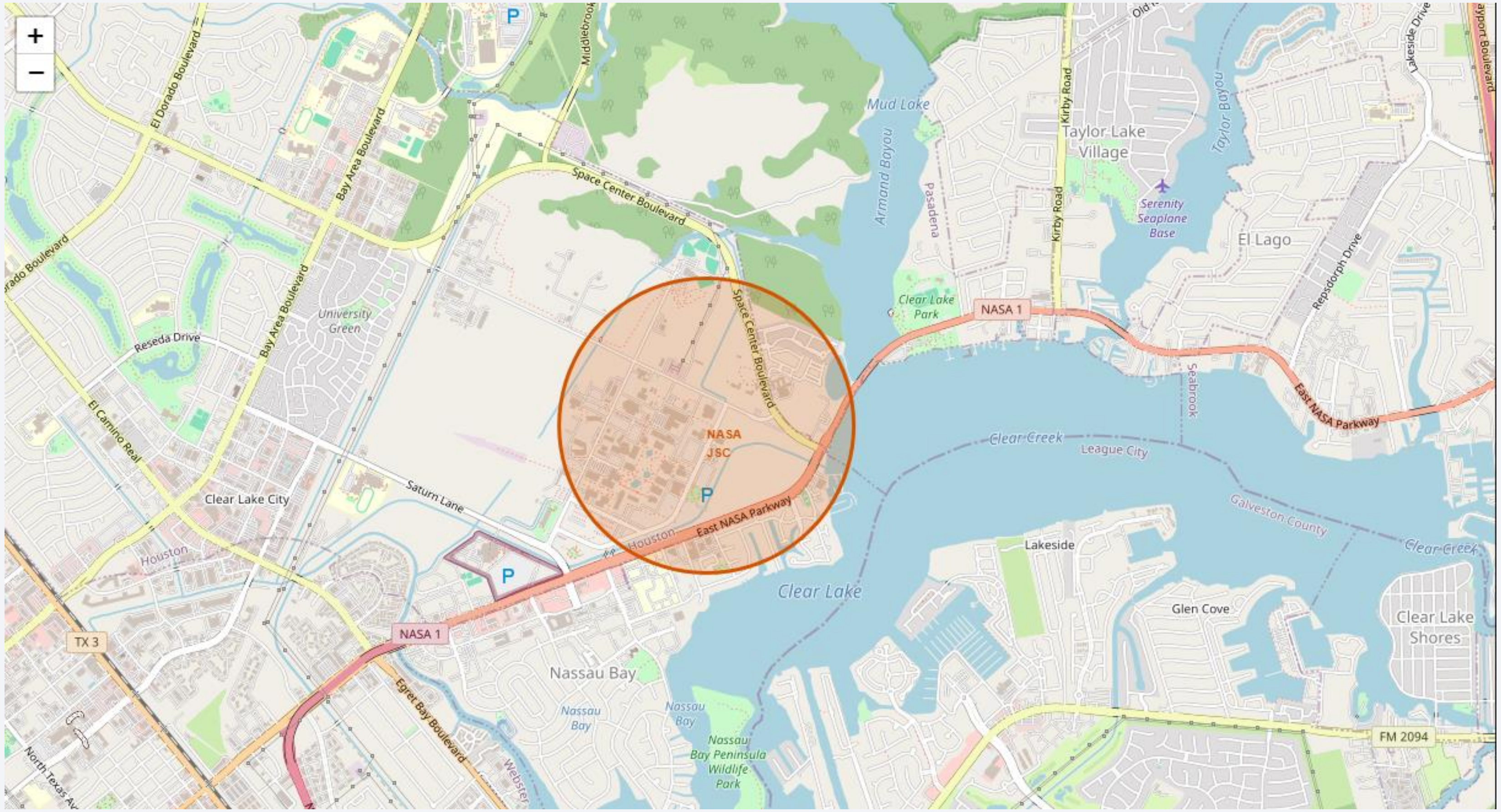
Landing_Outcome	Total_outcome
Success (ground pad)	9
Failure (drone ship)	5

Section 3

Launch Sites Proximities Analysis

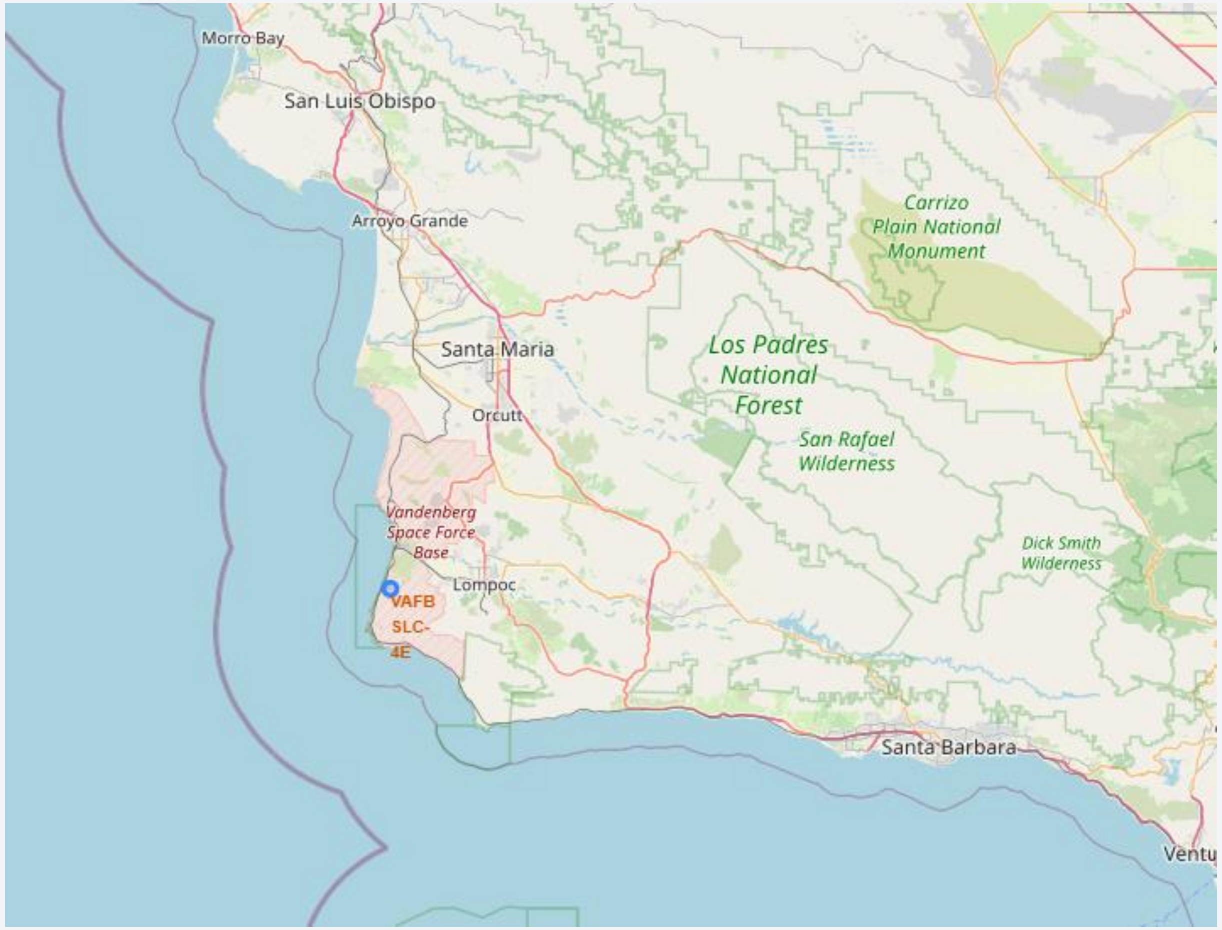
<NASA Johnson Space Center>

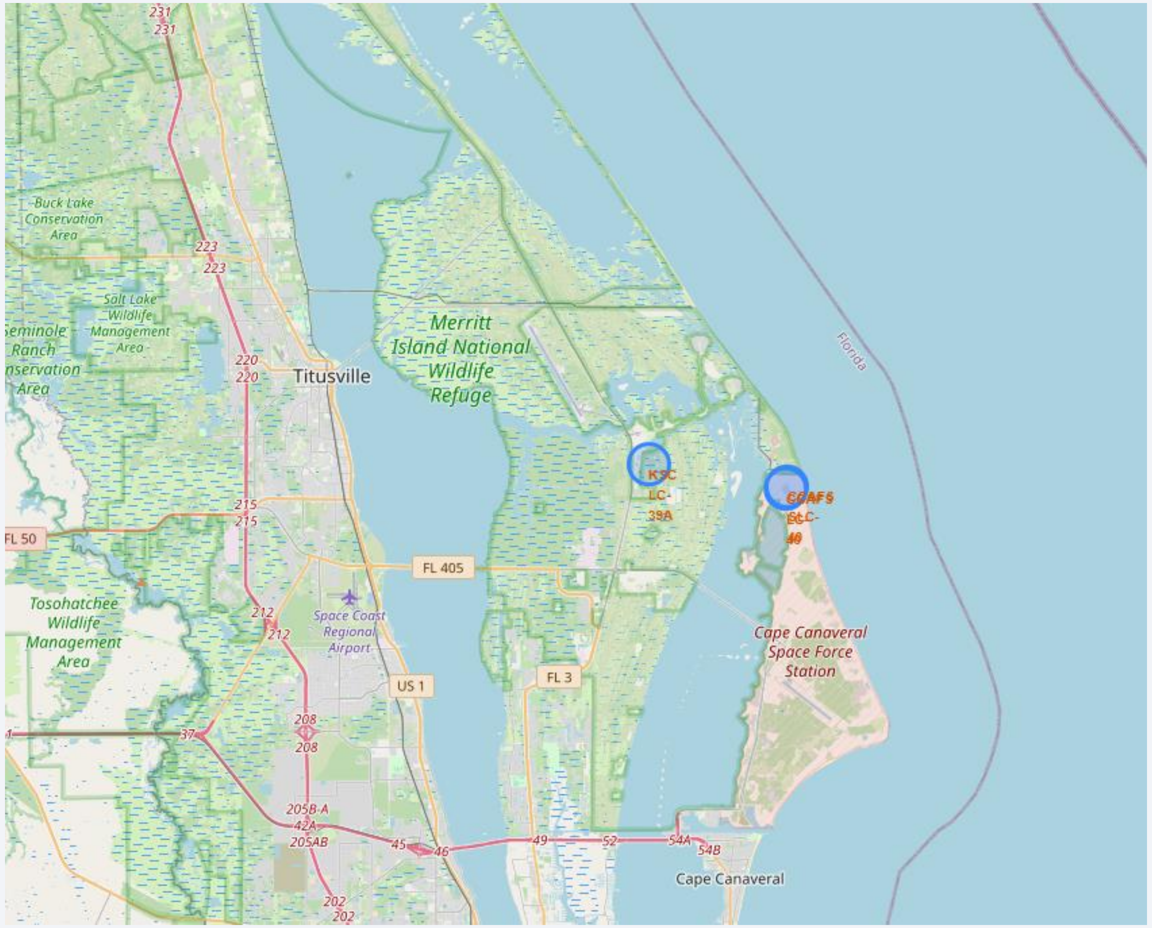
- The center is close to the coastal side.



<Launch sites>

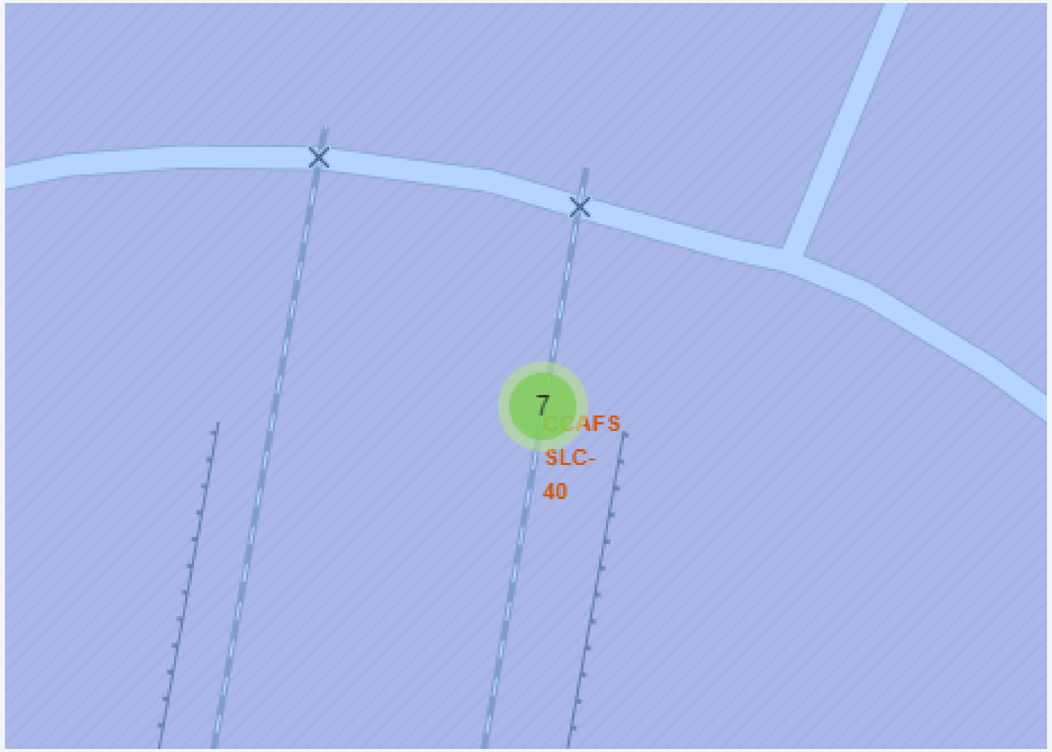
- There are 3 launch sites located in Florida state and just only one stayed in California state.





<Launch site CCAFS SLC-40 >

- There are 7 flights which were launched at CCAFS SLC-40 site, and 3 of them are successful and the rest are failed.

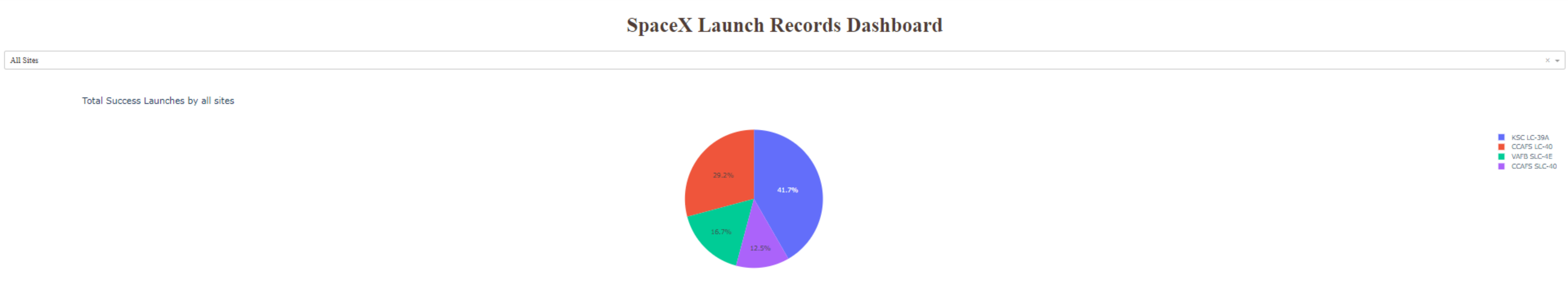


Section 4

Build a Dashboard with Plotly Dash

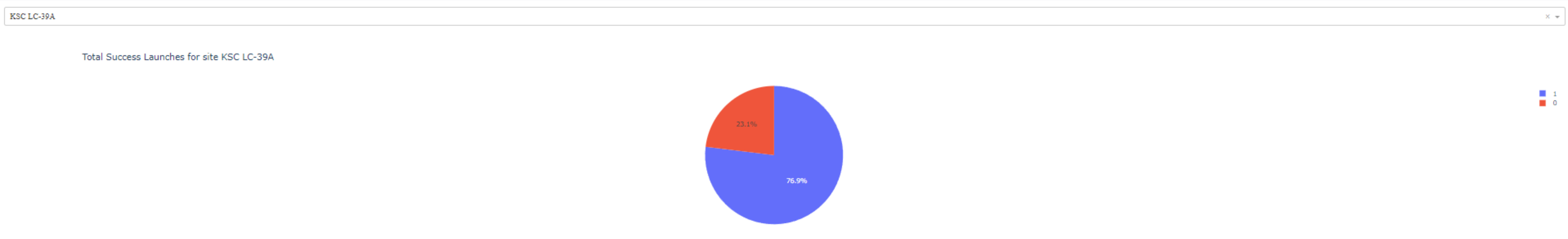
<Success Launches by all sites>

- The highest success launches belonging to KSC LC-39A with 41.70%, and CCAFS SLC-40 has the lowest success rate with just only 12.50%
- The success rate of CCAFS LC-40 and VAFB SLC-4E recorded 29.20% and 16.70% respectively



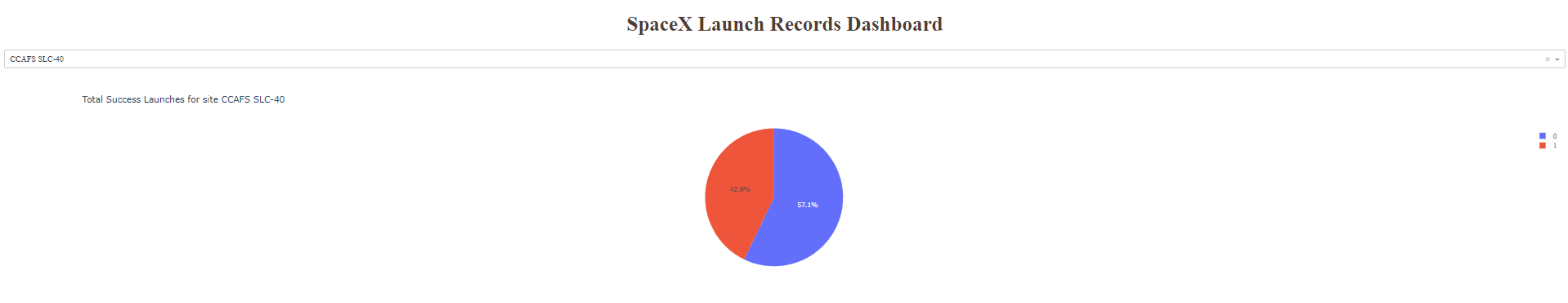
<Success ratio of KSC LC-39A site>

- The success rate occupied up to 77% approximately, and the rest are uncessful flight (23.10%)



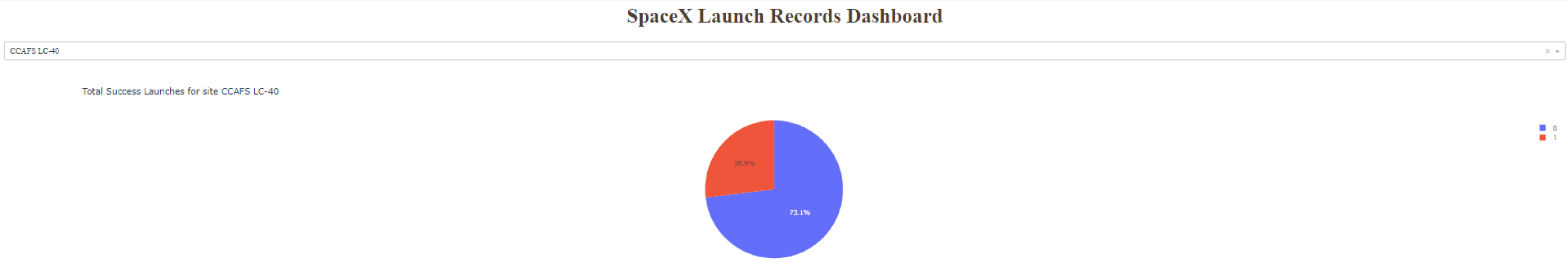
<Success ratio of CCAFS SLC-40 site>

- The success rate occupied up to 57% approximately, and the rest are uncessful flight around 43%



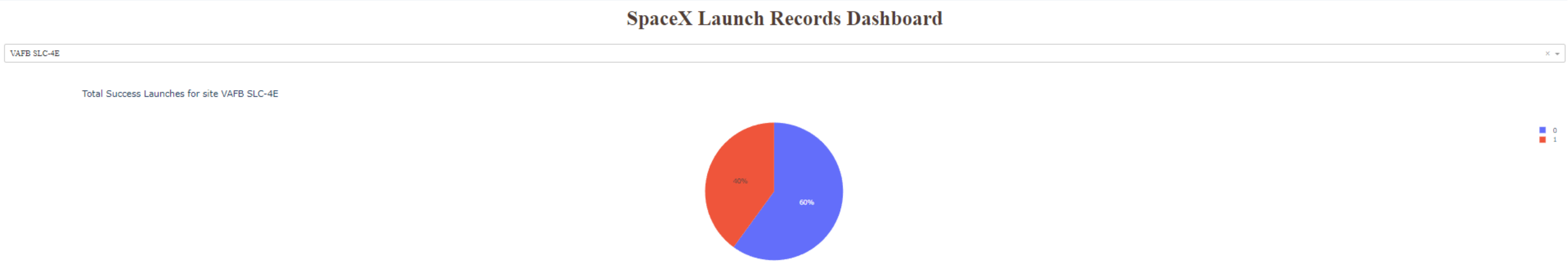
<Success ratio of CCAFS LC-40 site>

- The success rate occupied up to 73% approximately, and the rest are uncessful flight around 27%



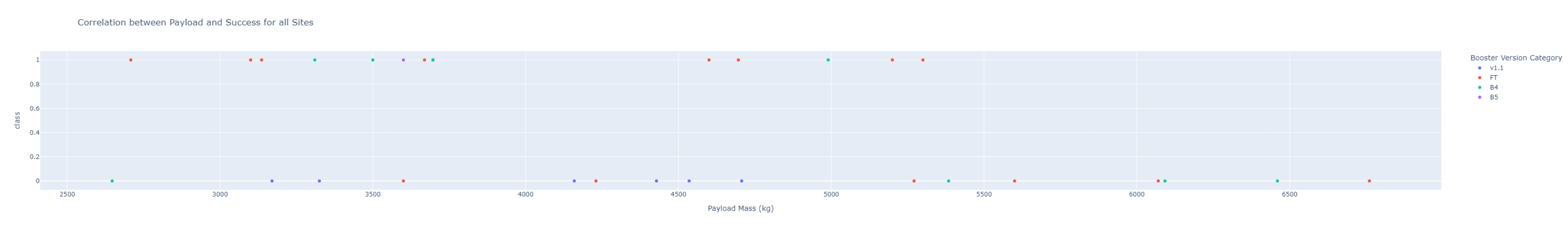
<Success ratio of CCAFS SLC-40 site>

- The success rate occupied up to exact 60%, and the rest are uncessful flights 40%



<Scatter plot of Payload vs. Launch Outcome>

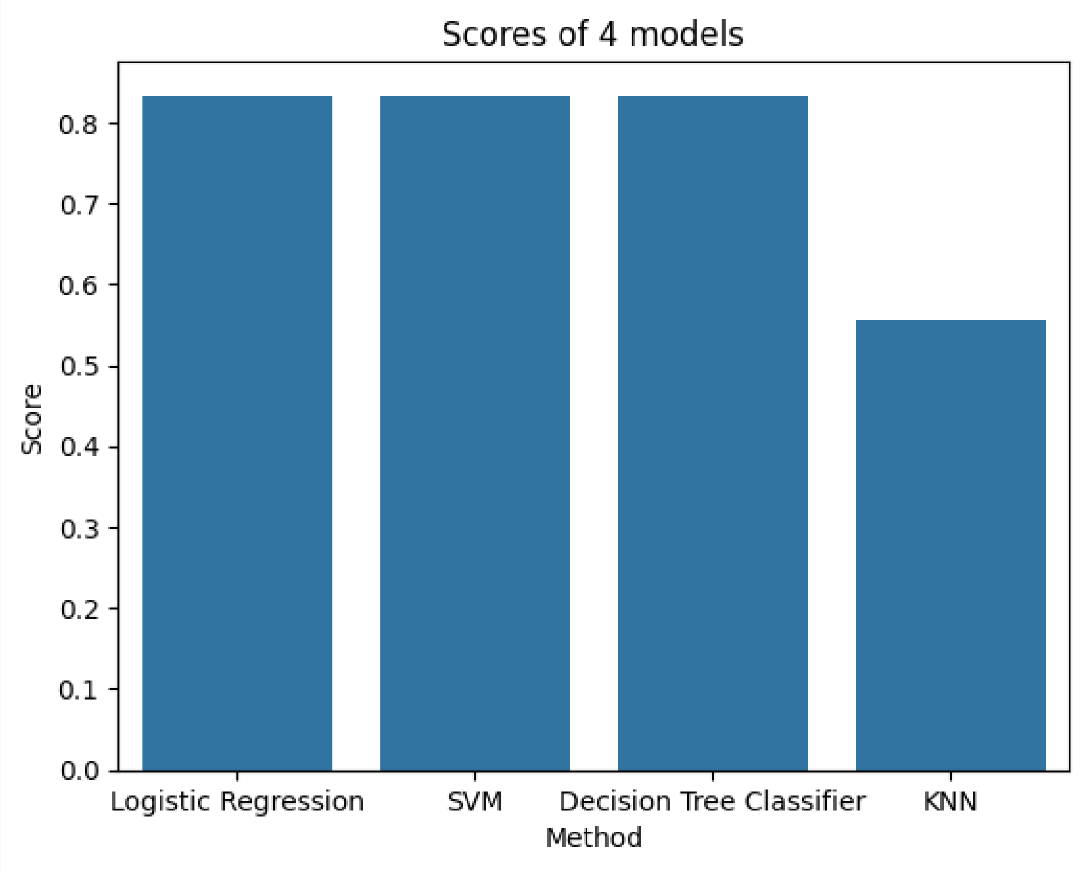
- Below figure shows Payload vs. Launch Outcome scatter plot for all sites, with payload range 2500-7500
- Booster version FT has the highest number of succesful flight
- Booster version v1.1 has the highest number of unsuccessful flight.
- When the payload mass got heavier than 5500 kg, there were none successful flights recorded.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

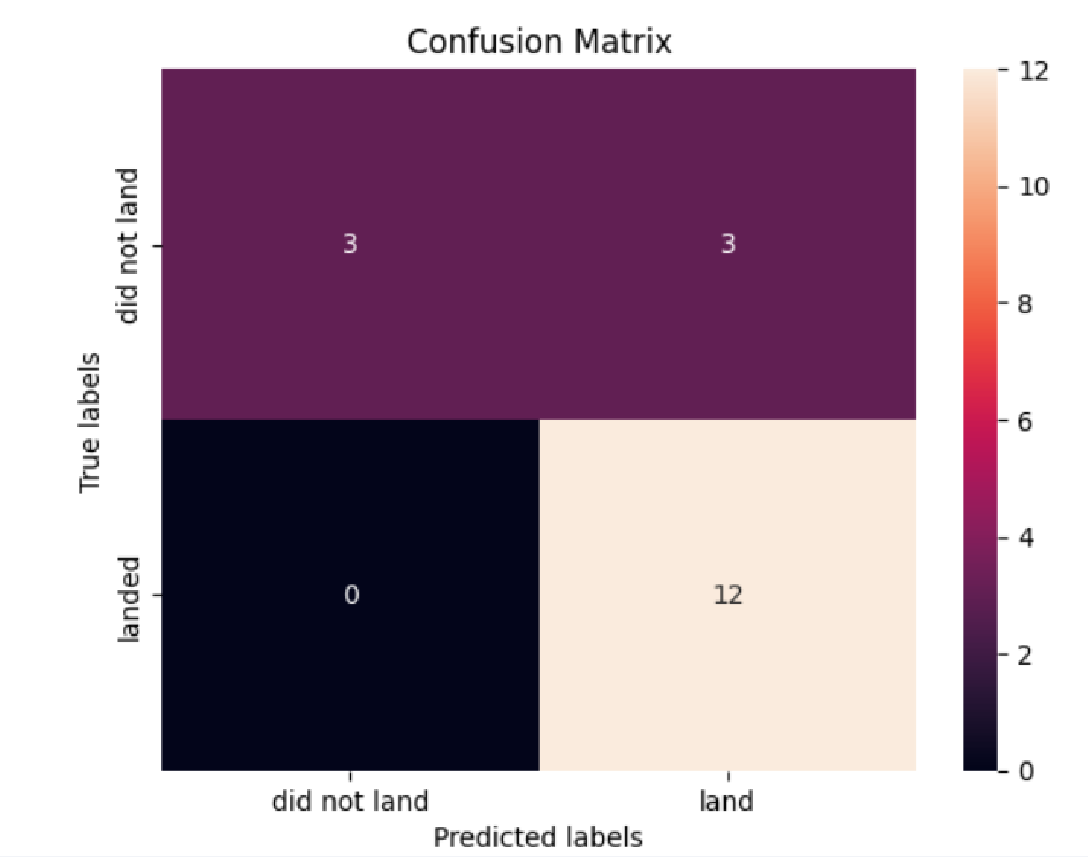


- There are 3 methods, which are Logistic Regression, SVM and Decision Tree Classifier, scoring the same and that one is the highest score on testing data.
- The KNN is the method should not be considered to train and test model. Because the score offered from this method is abit more than 0.5.

Confusion Matrix

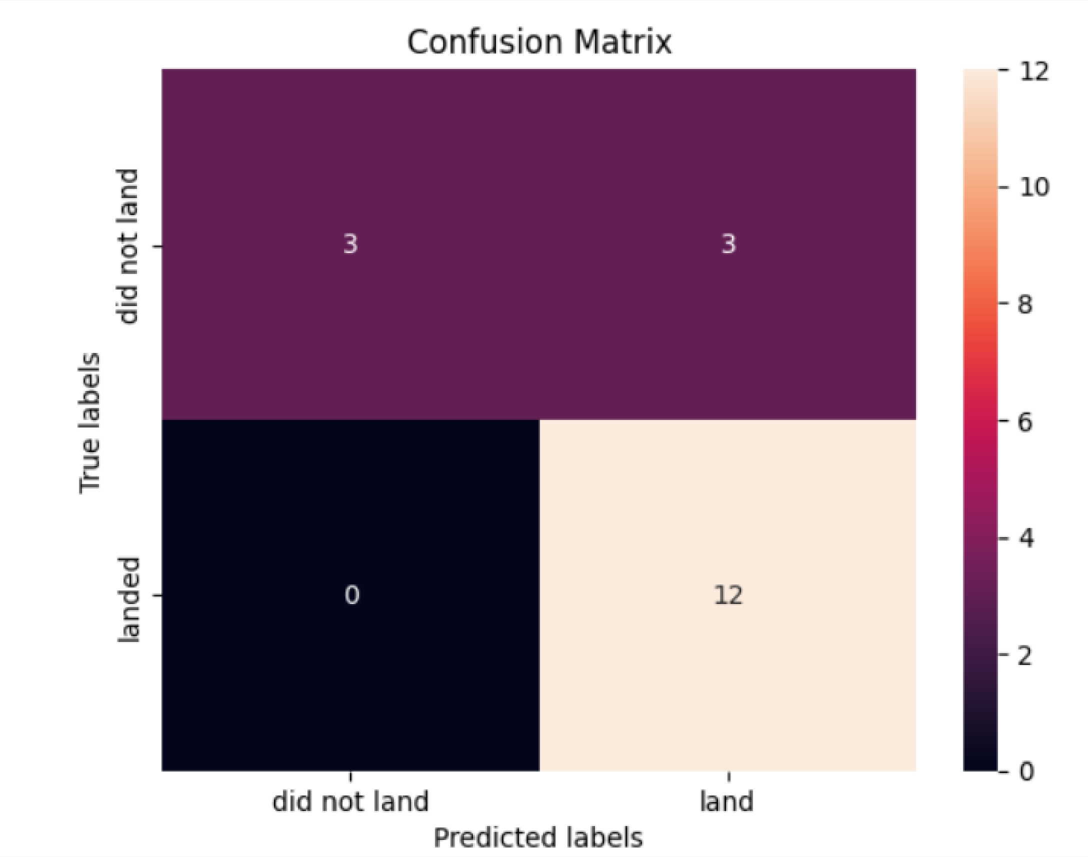
- As mentioned from previous slide, there are 3 methods scoring the same and also that score is the highest. As a result, the confusion matrixes of the best performing model of Logistic Regression, Decision Tree Classifier and KNN are displayed in the next slide.

Confusion Matrix of Logistic Regression



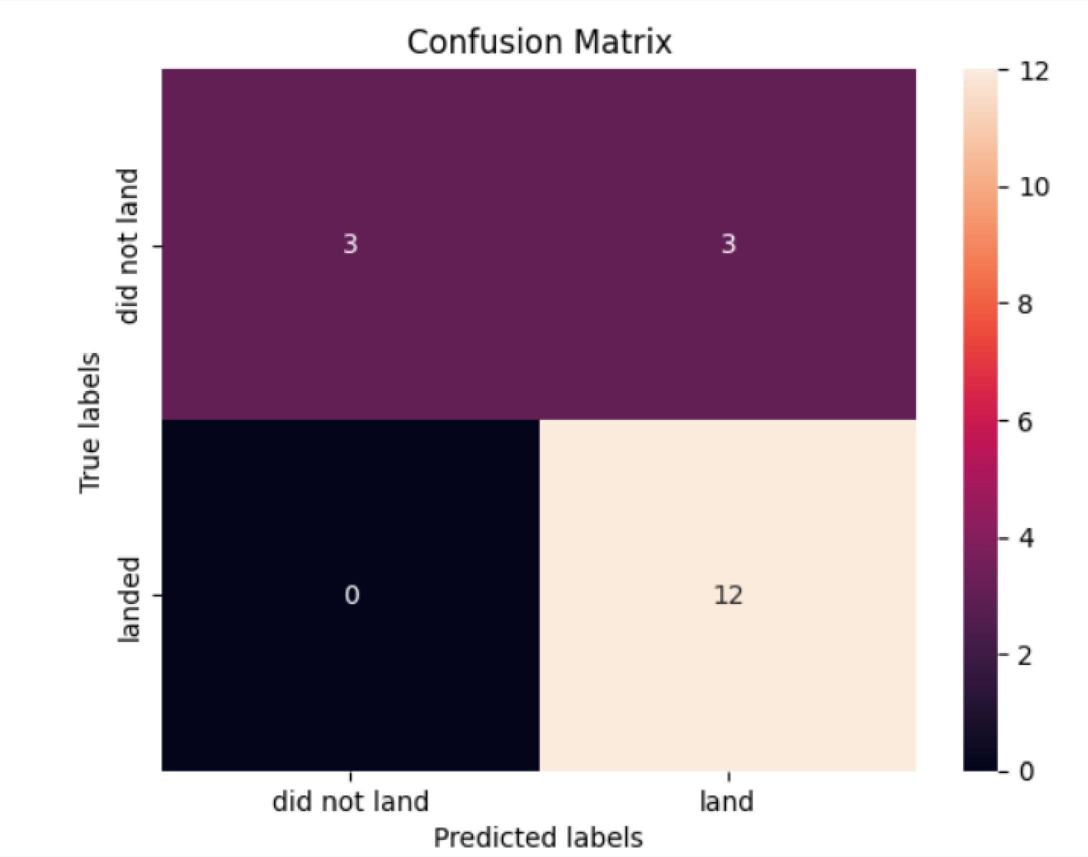
- This Logistic Regression method predict 12 launches land successfully and they did land without fails, and 3 launches land unsuccessfully and they did not land. Unfortunately, there were 3 launches this model predicted the flight land, but these flights did not land.

Confusion Matrix of Decision Tree Classifier



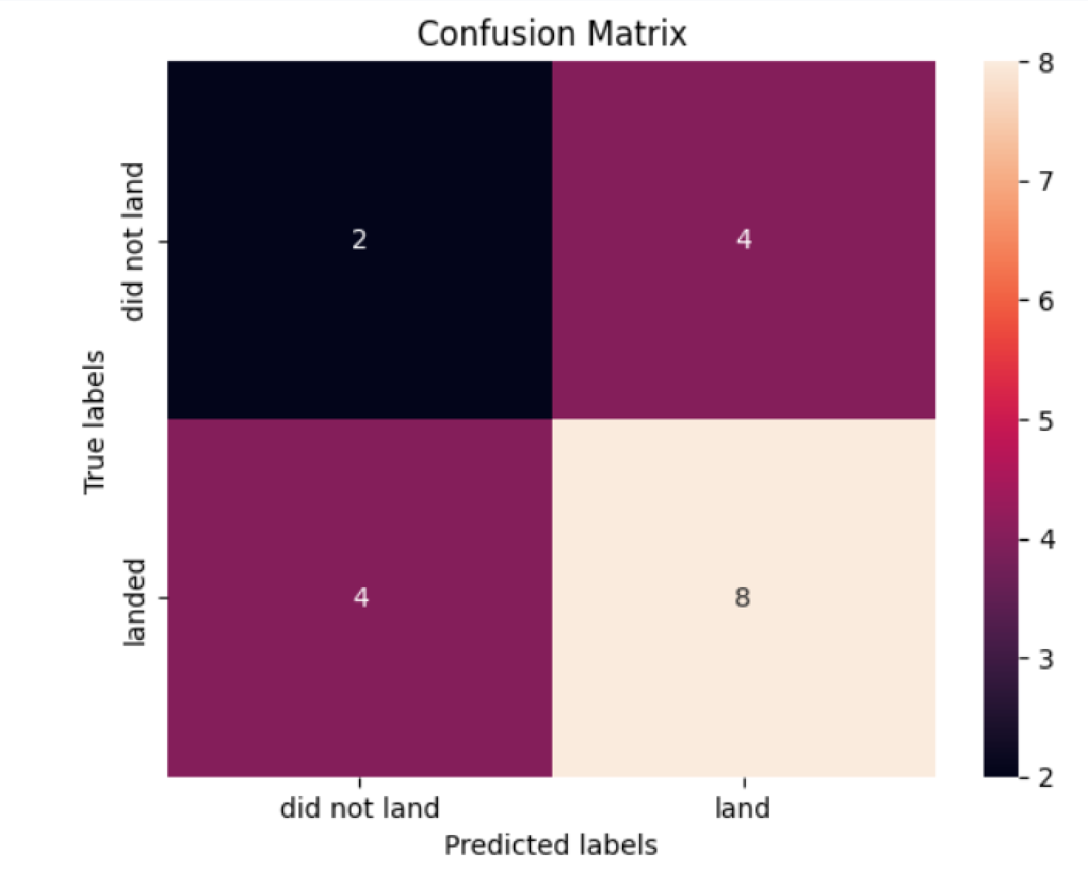
- This Decision Tree Classifier method predict 12 launches land successfully and they did land without fails, and 3 launches land unsuccessfully and they did not land. Unfortunately, there were 3 launches this model predicted the flight land, but these flights did not land.

Confusion Matrix of Support Vector Machine



- This Support Vector Machine method predict 12 launches land successfully and they did land without fails, and 3 launches land unsuccessfully and they did not land. Unfortunately, there were 3 launches this model predicted the flight land, but these flights did not land.

Confusion Matrix k Nearest Neighbors



- This k Nearest Neighbors method predict 8 launches land successfully and they did land without fails, and 2 launches land unsuccessfully and they did not land. Unfortunately, there were 4 launches this model predicted the flight land, but they did not land; and there were also 4 launches this model predicted the flight do not land, but they did land.

Conclusions

- Flight number and Payload has positive linear ratio with success rate. So, increasing number of flights and payload mass can be considered.
- The launch rate at KSC LC-39A can be operated more often, because the success rate when launching at this rate is quite high.
- The booster version FT should be chosen to utilize, because it held the best successful records.
- The launch sites should be located near the costal side, because it can handle the failed flight effectively.
- The method of Logistic Regression, SVM or Decision Tree Classifier is the sufficient method to train and test data.

Appendix

- Relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that have been created during this project are attached in this presentation.

Thank you!