Winning Space Race with Data Science

Gabriel Nascimento Serafim

May 21, 2022

Outline

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

Executive Summary

- Project's methodology followed the sequence below:
  - Data collection using SpaceX API and web scraping;
  - Exploratory Data Analysis (EDA), using techniques such as data cleaning, data wrangling and data visualization;
  - Use of 4 different Machine Learning model for prediction of successful launch.
- Summary of all results:
  - Data collection process was efficient, allowing the collection of valuable information from different sources;
  - EDA allowed to identify relevant features to predict the desired outcome;
  - Machine Learning results showed important features and parameters to predict successful launches.

Introduction

- As a potential competitor to SpaceX, Space Y has the desire to understand the viability of being a rocket launch company.

- We want to know:
  - How to estimate cost for launches by predicting successful and unsuccessful launches;
  - Best location to perform successful launches.

Section 1

Methodology

Methodology

Executive Summary

- Data from Space X were obtained from 2 different sources:
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - Web scraping from: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- Perform data wrangling
  - A database containing the desired features and outcome was created after performing data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Four different models (logistic regression, SVM, decision tree and KNN) were trained after data normalization, split between training and test sets and use o grid search to find best parameters combination for each model

Data Collection

- Data from Space X were obtained from 2 different sources:
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - Web scraping from: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

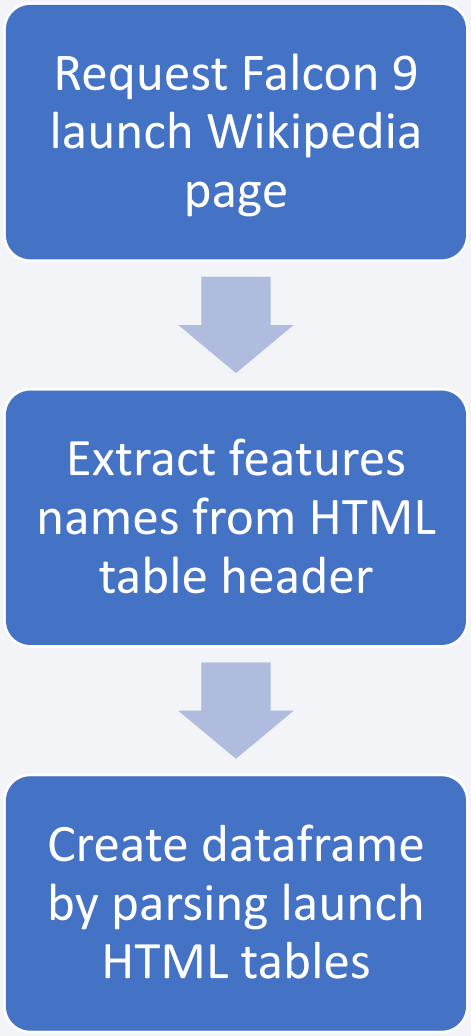
Data Collection – SpaceX API

- SpaceX offers a public API that allows to access and retrieve desired data
- This API was used according to the flowchart beside and then data is persisted.

- Source code: https://github.com/gnserafim/applied-data-science-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.



- Source code: https://github.com/gnserafim/applied-data-science-capstone/blob/main/jupyter-labs-webscraping

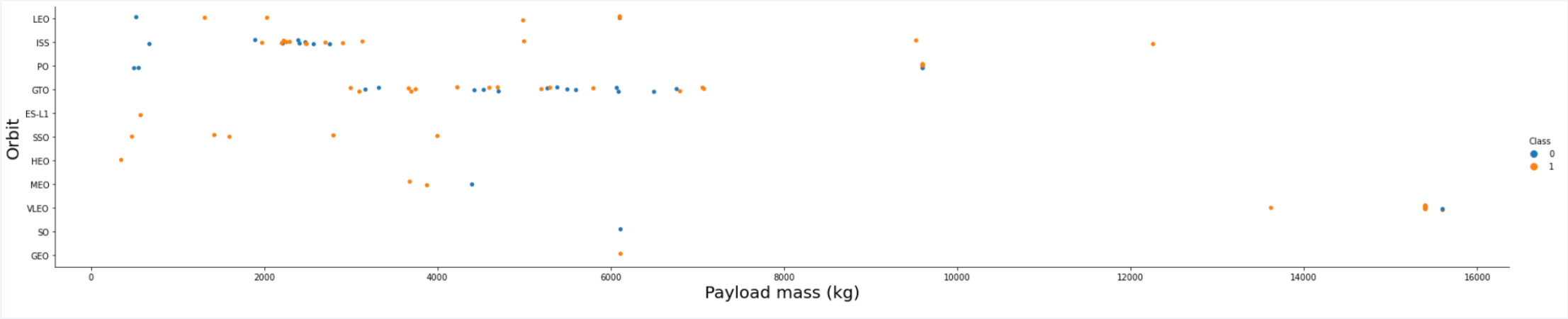
Data Wrangling

- Initial EDA was performed for understand of dataset
- Summary launches per site, occurrence of each orbit and occurrence of mission outcome per orbit
- Definition of landing outcome label

- Source code: https://github.com/gnserafim/applied-data-science-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



- Source code: https://github.com/gnserafim/applied-data-science-capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

Source code: https://github.com/gnserafim/applied-data-science-capstone/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site;
  - Lines are used to indicate distances between two coordinates.

Source code: https://github.com/gnserafim/applied-data-science-capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Html view: https://htmlpreview.github.io/?https://github.com/gnserafim/applied-data-science-capstone/blob/main/lab_jupyter_launch_site_location.html

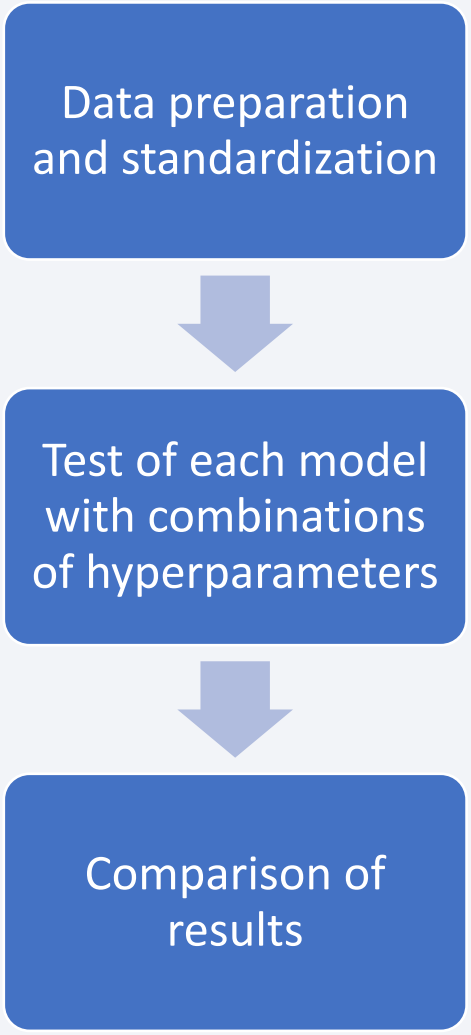
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Source code: https://github.com/gnserafim/applied-data-science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: **Logistic Regression, SVM, Decision Tree and K-nearest Neighbors.**



Source code: https://github.com/gnserafim/applied-data-science-capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

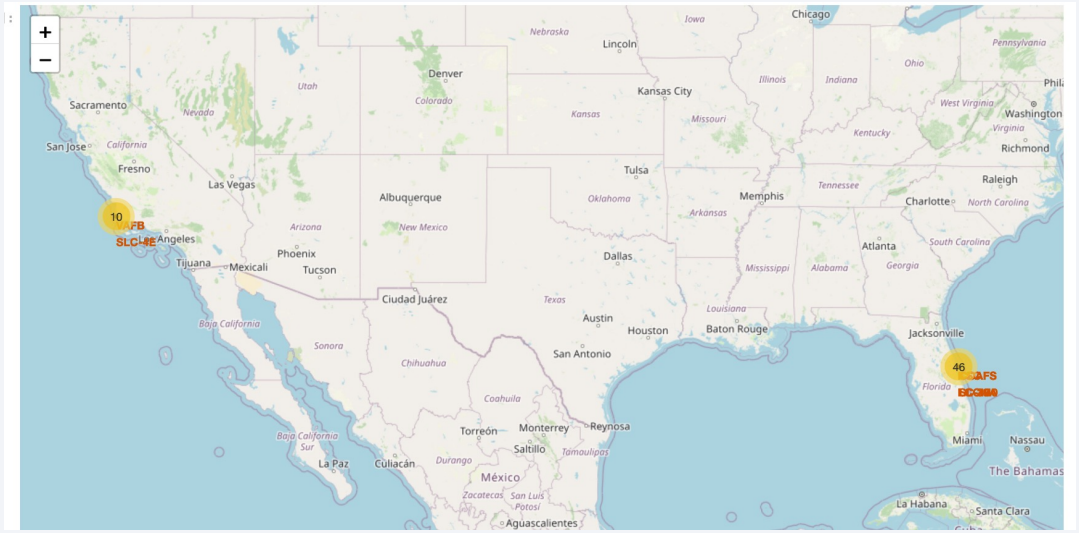
Results

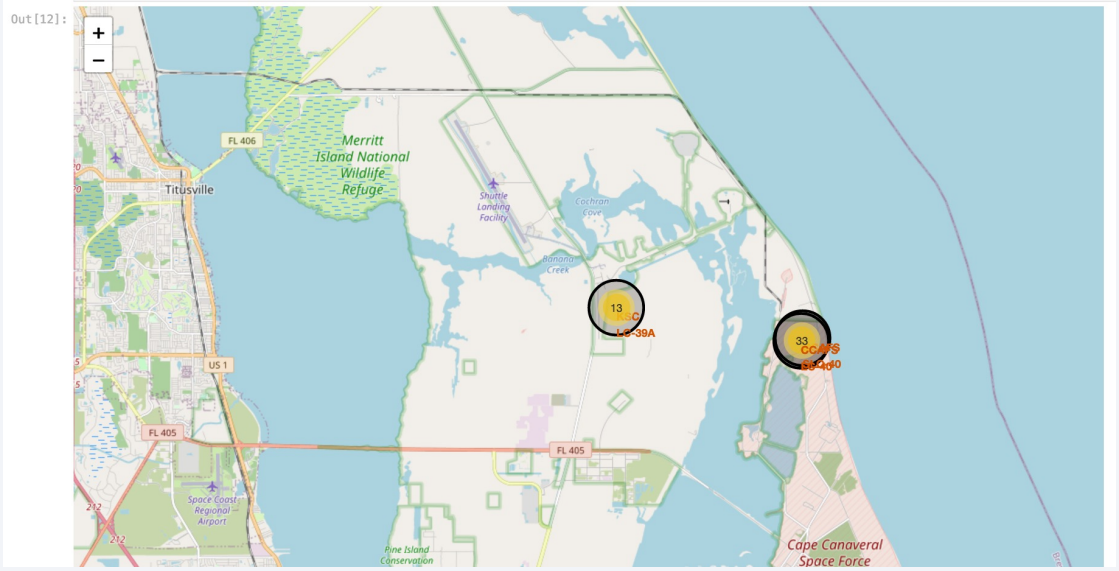
Exploratory data analysis results can be summarized as:

- Space X uses 4 different launch sites;
- Majority of flights took off from CCAFS SLC 40 - first launch site available
- VAFB site has the least amount of flight, with almost 77% of success rate, same as KSC
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- Success rate since 2013 kept increasing till 2020.

Results

- Using interactive analytics, itwas possible to identify that launch sites are situated around places with access to the sea and that have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.

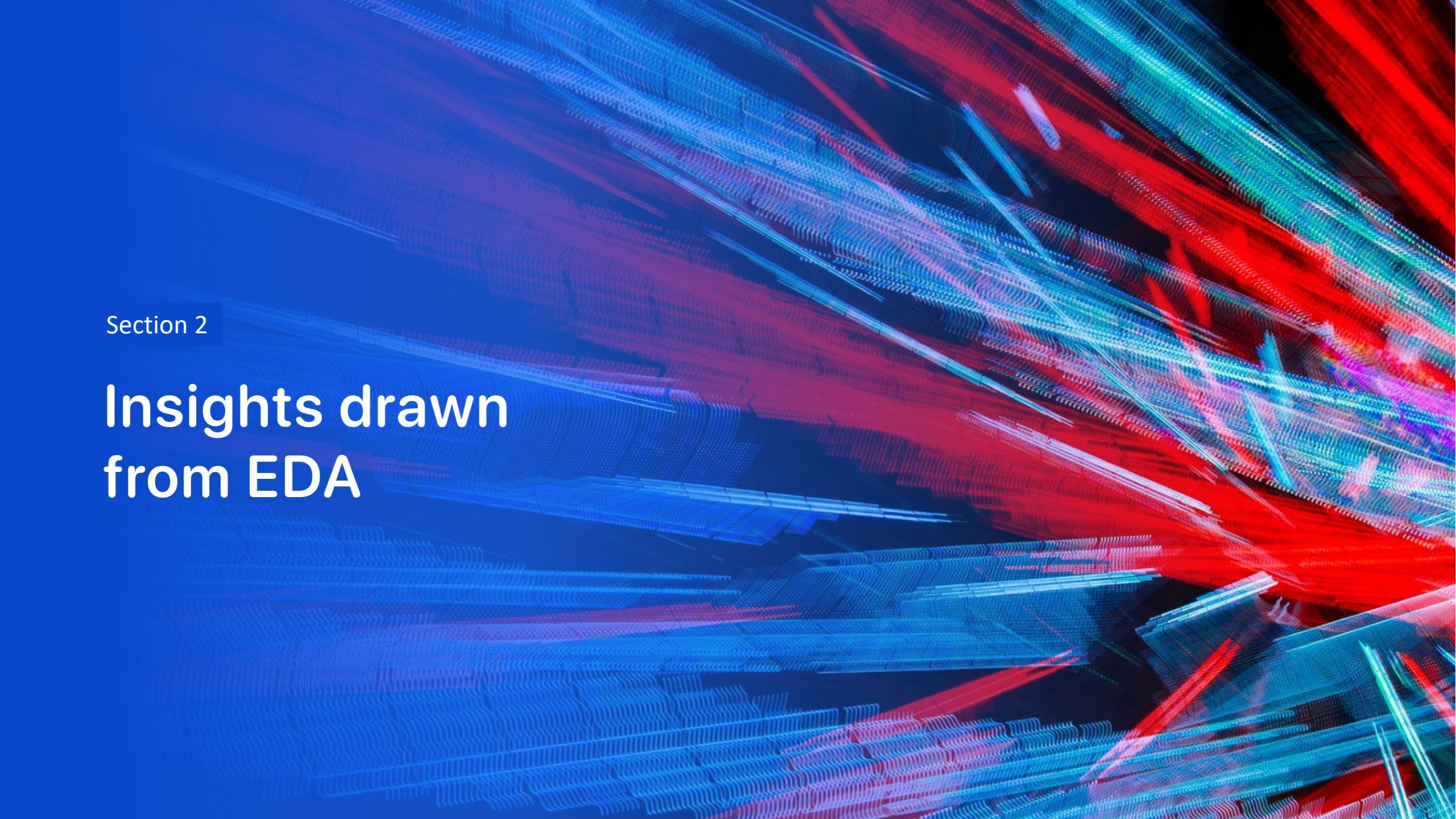




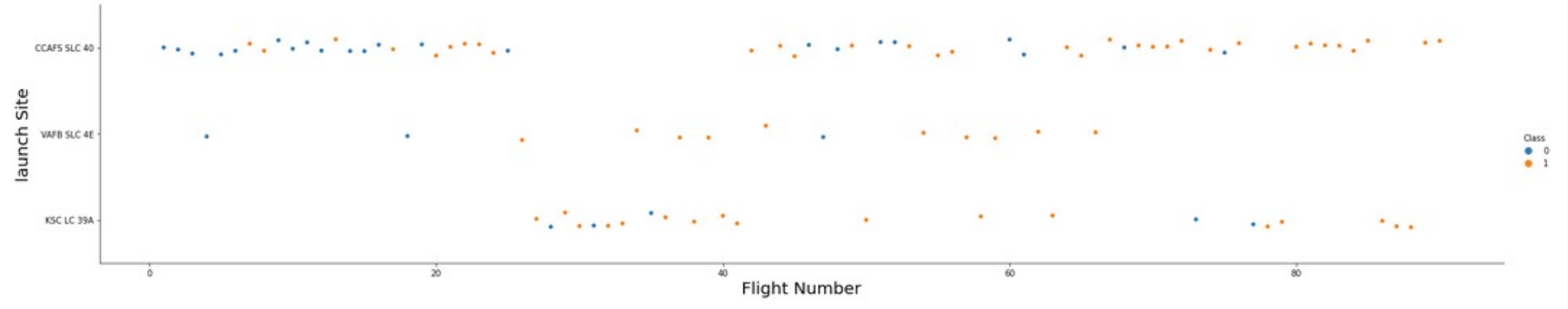
Results

- Predictive Analysis showed that SVM is the best model to predict successful landings, having train accuracy over 86% and test accuracy for test data over 83%.

```
logistic regression
tuned hpyerparameters :(best parameters)  {'C': 1, 'penalty': 'l2', 'solver': 'lbfgs'}
train accuracy : 0.8214285714285714
test accuracy : 0.8333333333333334


SVM
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
train accuracy : 0.8482142857142858
test accuracy : 0.8333333333333334


decisiton tree
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}
train accuracy : 0.8625
test accuracy : 0.6666666666666666


KNN
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 6, 'p': 1}
train accuracy : 0.8339285714285714
test accuracy : 0.8333333333333334
```

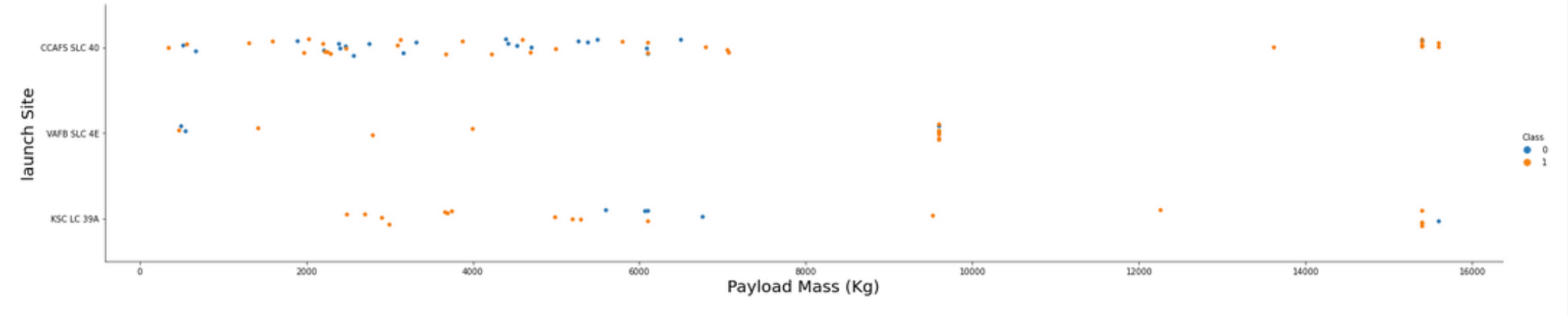

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Majority of flights took off from CCAFS SLC 40 - first launch site available;
- VAFB site has the least amount of flight, with almost 77% of success rate, same as KSC
- Considering the last 20 flight, the current success rate is 85%

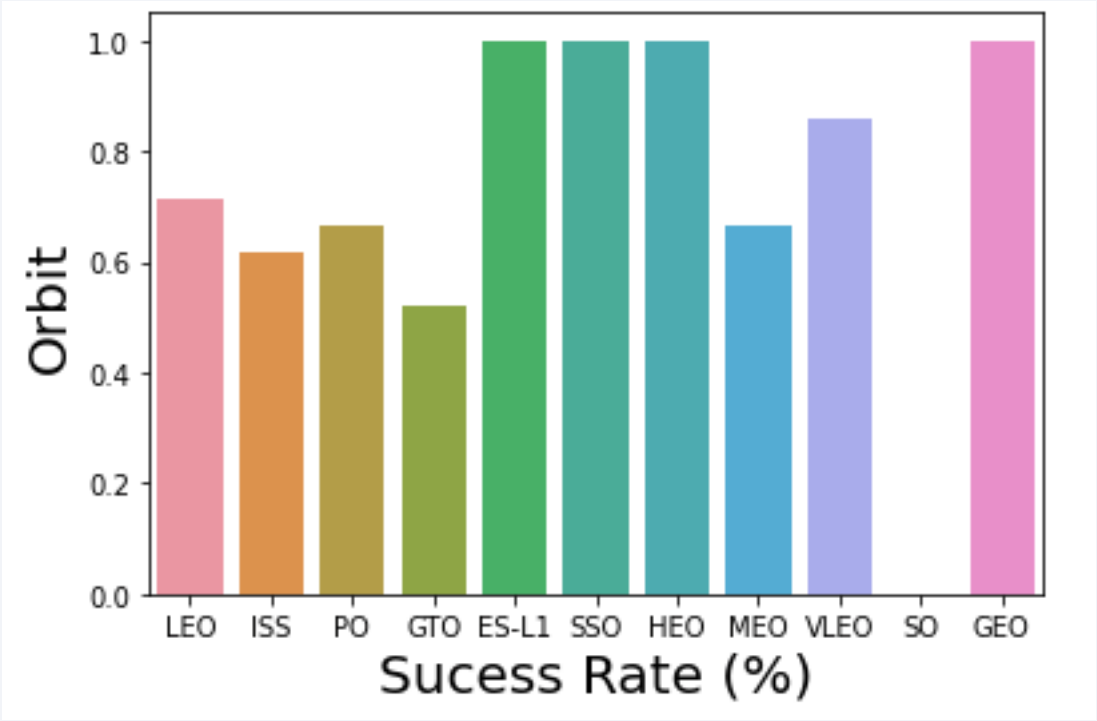
Payload vs. Launch Site



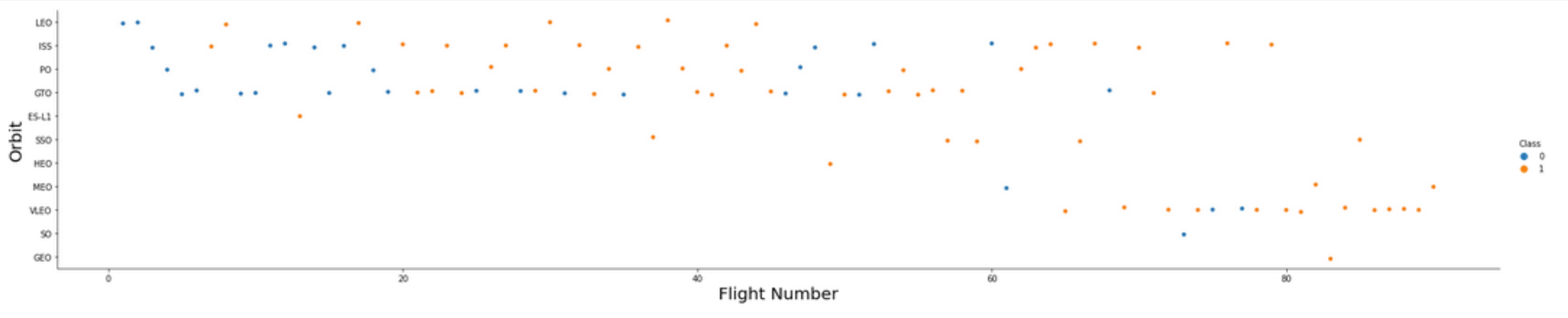
- There are no rockets launched with heavy payload (greater than 10K Kg) for VAFB-SLC launch site;
- Launches with payload mass equal or greater than 9K kg have a success rate around 87%

Success Rate vs. Orbit Type

- All Orbit have a success rate greater than 50%
- Orbits SSO, HEO, MEO and GEO have a success rate of 100%;
- The only launch for orbit SO was not successful

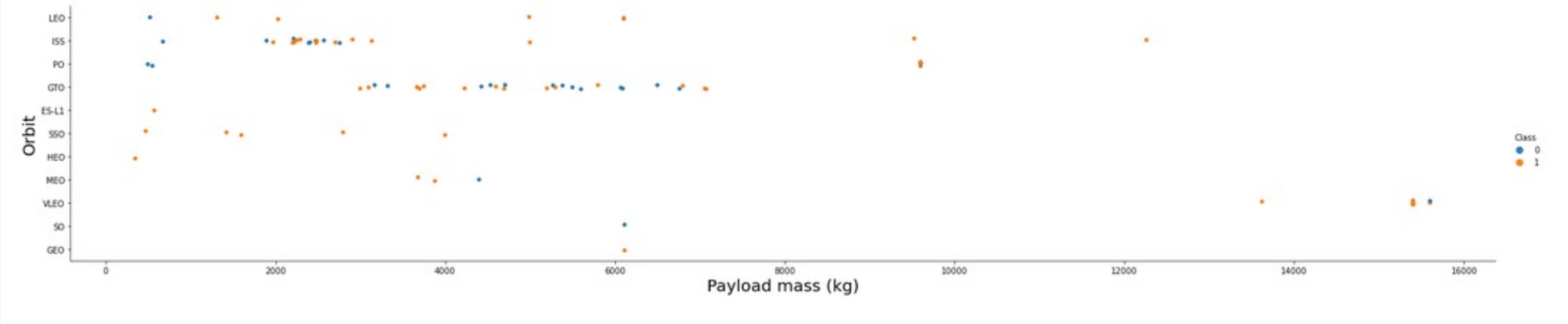


Flight Number vs. Orbit Type



- Success rate has been improving over time for all orbits;
- LEO orbit success may be related to the number of flights, while the same can't be assumed for GTO orbit
- The number of VLEO orbit flights have been growing significantly

Payload vs. Orbit Type

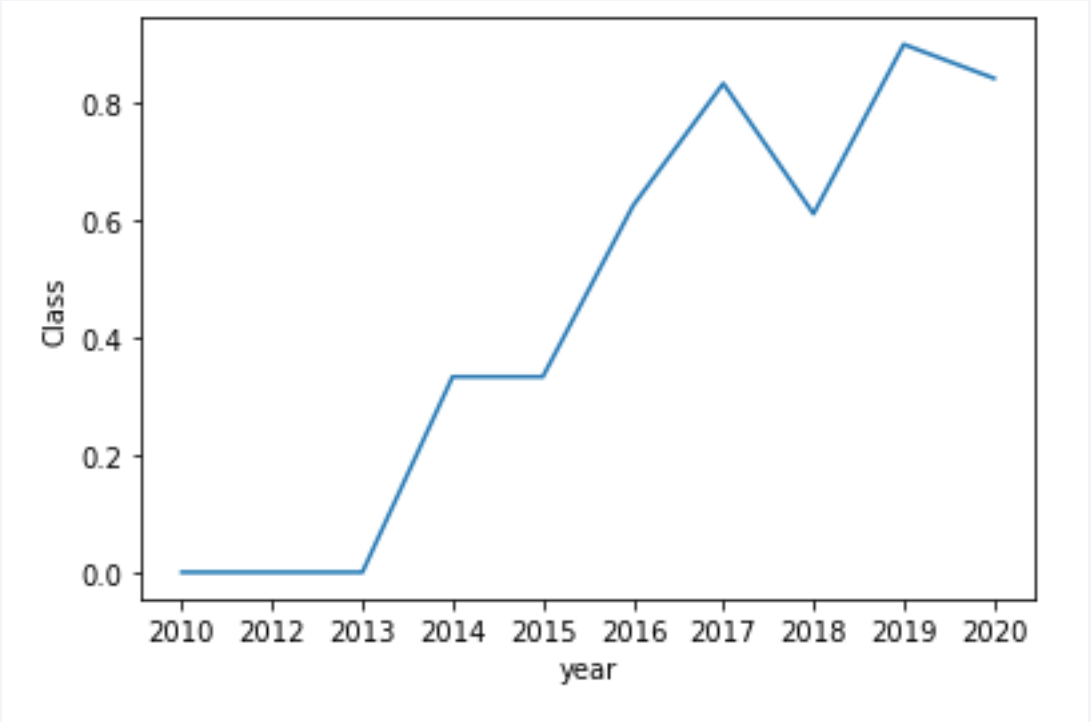


- Polar, LEO and ISS orbit have higher success rate for launches high heavy ;
- It is not possible to distinguish a relation between payload and success rate to orbit GTO;
- SO and GEO orbit don't have enough launches to be able to take any insights from it

Launch Success Yearly Trend

- the success rate since 2013 kept increasing till 2020;

- First 3 year can be considered as test period for adjustment and improving of technology



All Launch Site Names

Unique launch sites were obtained selecting the distinct values from ‘launch_site’ features of the dataset

In [5]:
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX ORDER BY 1;
```

```
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b
Done.
```

Out[5]:

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

Launch Site Names Begin with 'CCA'

In [6]:
```
sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[6]:

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

Results above is a 5-sample size of flights that took of from Cape Canaveral launch site

Total Payload Mass

- Result is the sum of payload mass where customer correspond to ‘NASA (CRS)’

```
In [8]: sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)'
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu
Done.
Out[8]:    1
       45596
```

Average Payload Mass by F9 v1.1

- Result is the average of payload mass after filtering booster version like ‘F9 v1.1’

```
In [13]: sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE BOOSTER_VERSION LIKE '%F9 v1.1%'
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.c
Done.
Out[13]:    1
            2534
```

First Successful Ground Landing Date

- Result was obtained by getting the minimum value of date after filtering landing outcome for successful ground pad landings

```
In [14]: sql SELECT MIN(DATE) FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)'
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde0
Done.
Out[14]:        1
        2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Selecting distinct values of booster version where feature payload mass has values between 4K and 6K Kg

In [15]:
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship
```

```
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[15]:

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

Total Number of Successful and Failure Mission Outcomes

- By grouping mission outcome and selecting the count of total lines, it is possible to get the amount of successful and failed mission

In [21]:
```
sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEX GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tg
Done.
```

Out[21]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Using a sub-query to get the maximum payload mass from the dataset, the different version of a F9 booster are the distinct values that carried this respective maximum payload

```
In [23]: sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/b
Done.
```

Out[23]:

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

2015 Launch Records

- Landing outcomes in drone ship, their booster versions and launch site names that occurred in year 2015

In [28]:
```
sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 20
```

```
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[28]:

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

- Rank, from top to bottom, the count of landing outcomes between the date 2010-06-04 and 2017-03-20

In [30]:
```
sql SELECT LANDING__OUTCOME, COUNT(*) FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY 2 D
```

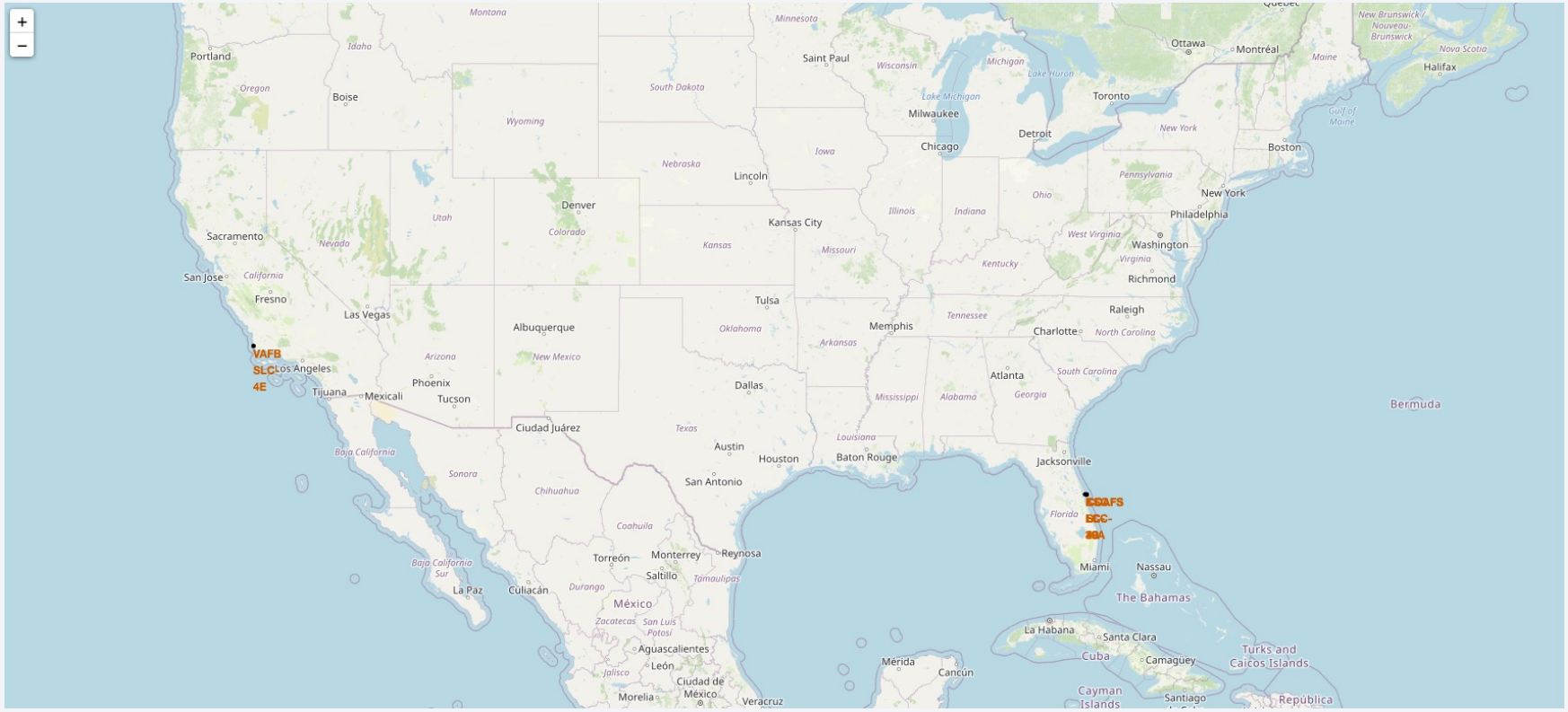
```
 * ibm_db_sa://jzl40730:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[30]:

landing__outcome	2
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

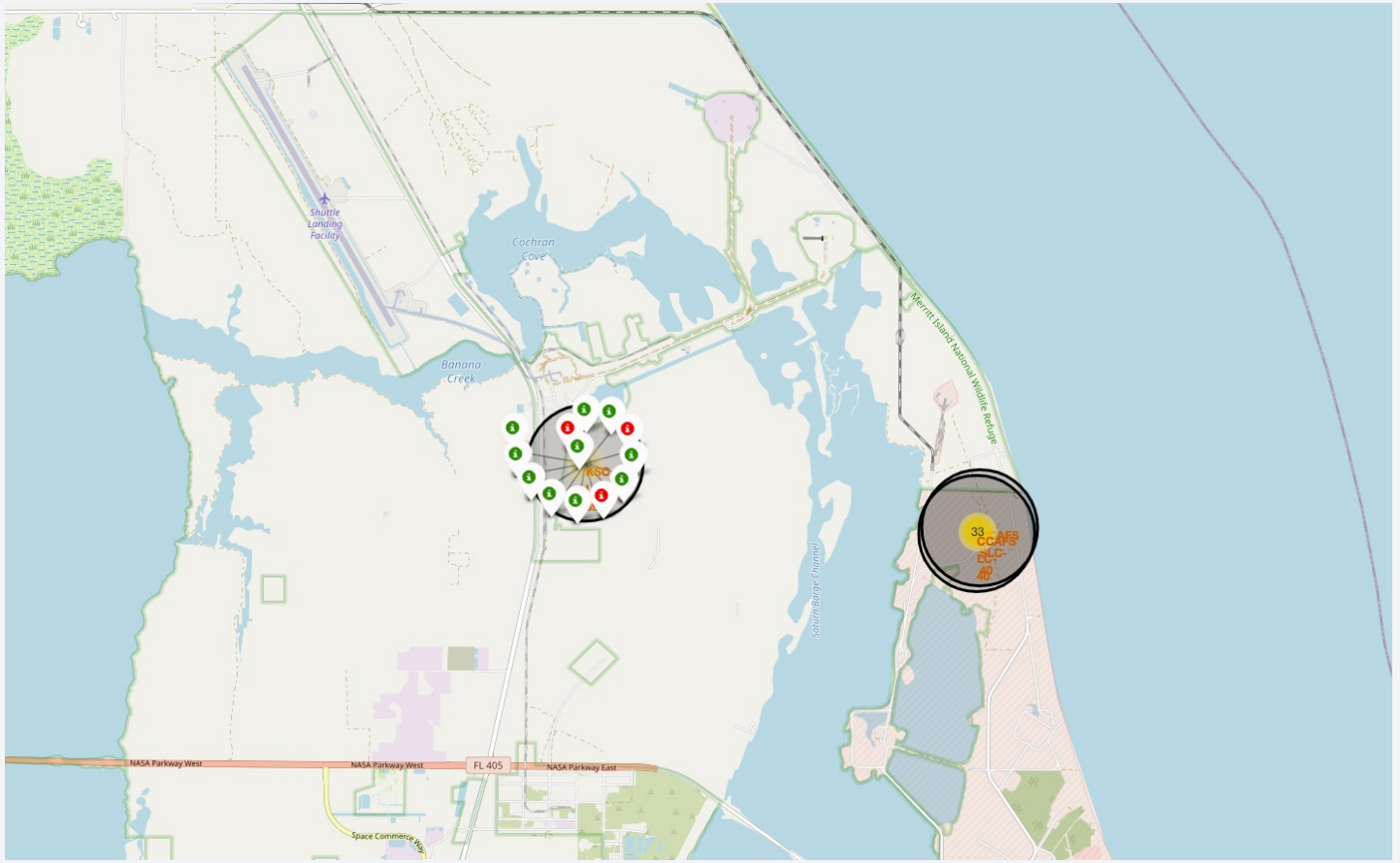
All Sites Location



All 4 sites are located close to the sea

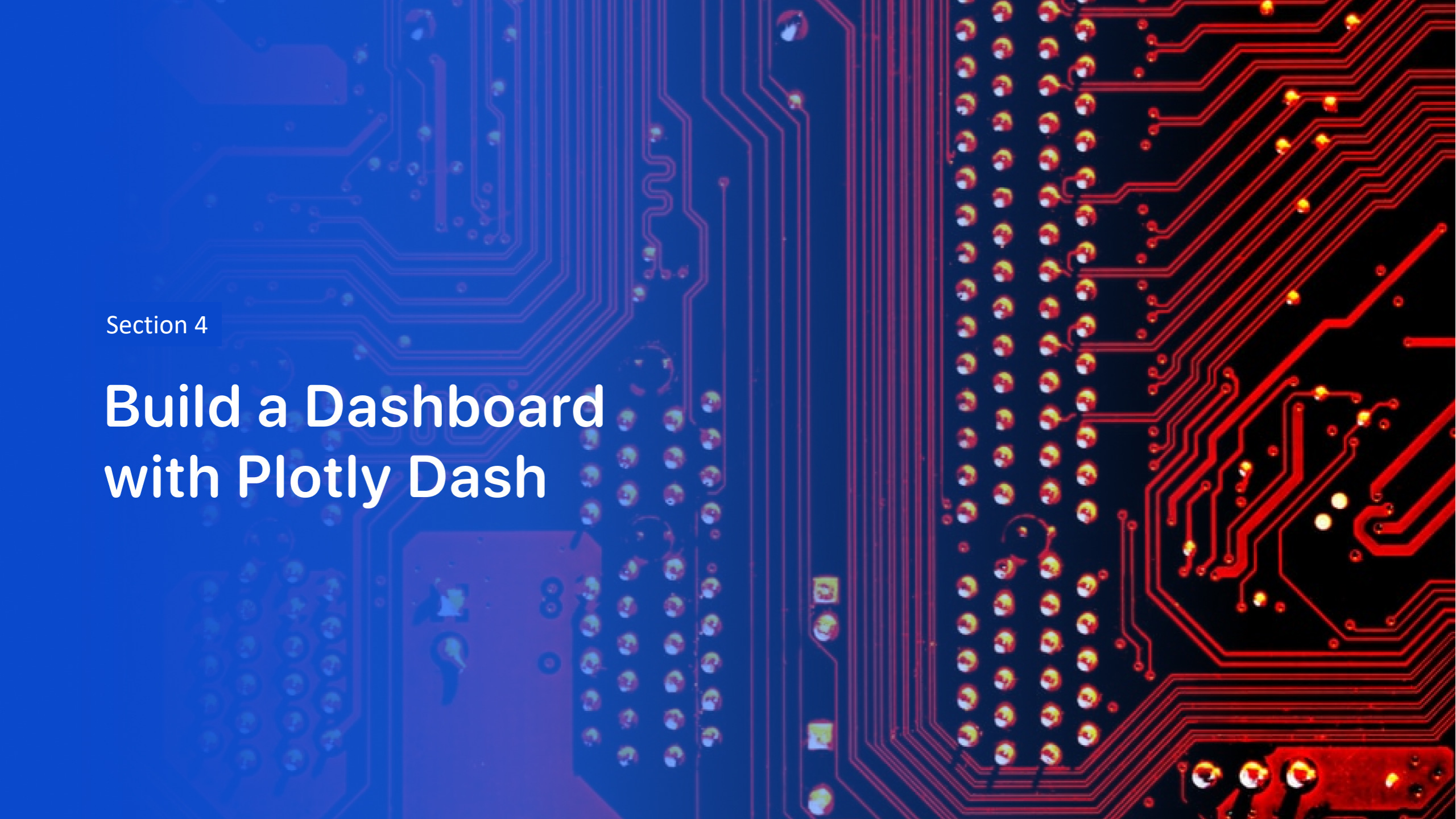
Launch Outcome by Site

- Using launch site KSC LC-39A as an example, green and red markers indicated successful and failed launches, respectively, for that specific site

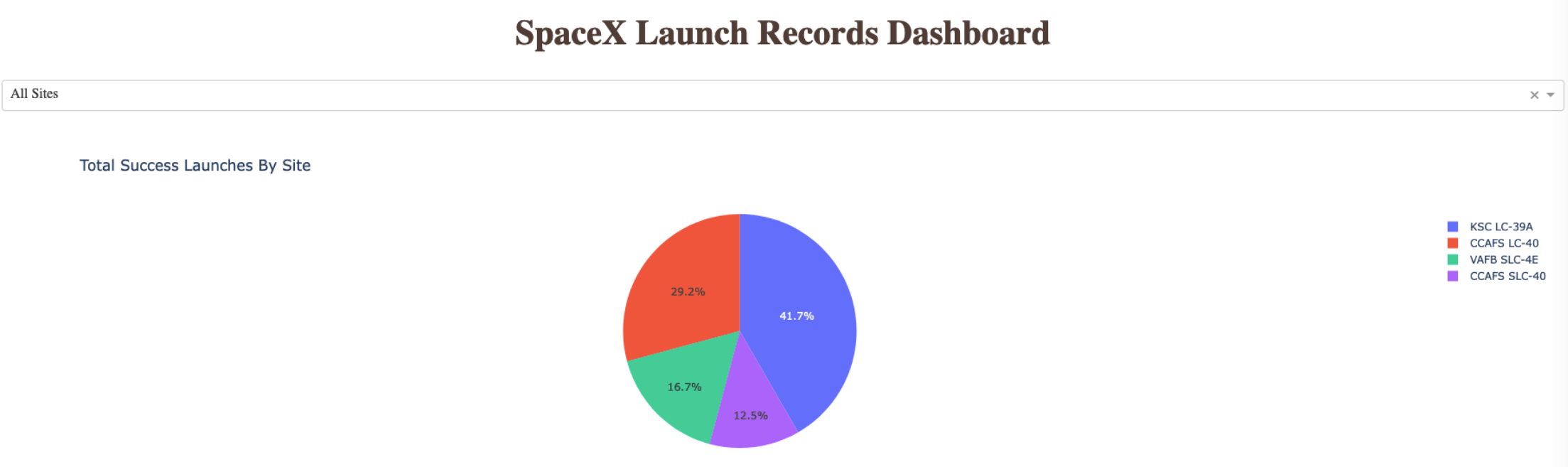


Nearby infrastructure

Site CCAFS CLS-40 is less than 1km apart from the coastline

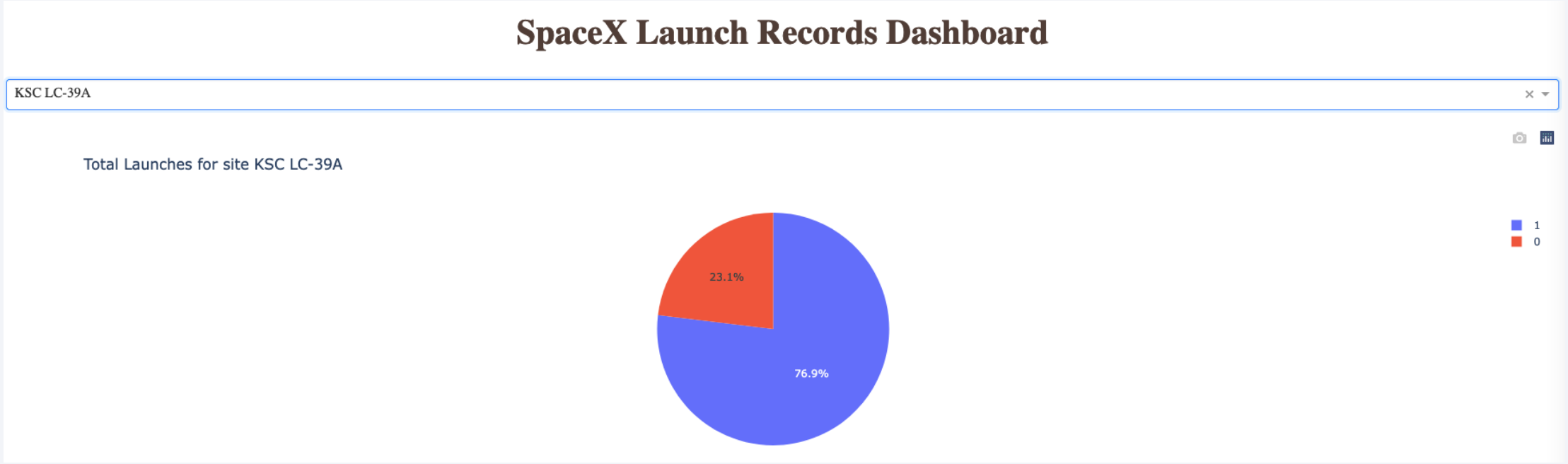

Section 4
Build a Dashboard with Plotly Dash

Successful Launches - Piechart



Due to difference in success ratio between launch site, location may have a significant impact on launches outcome

KSC LC-39A – Launch Success Ratio



- Almost 77% of launches performed form this site were successful

Payload vs. Launch Outcome – Scatter Plot

- Almost all launch that took off from and had payload equal or less than 2000 kg have failed CCAFS LC-40

Payload vs. Launch Outcome – Scatter Plot

Majority of successful launches had payload mass between 2000 and 5800 kg

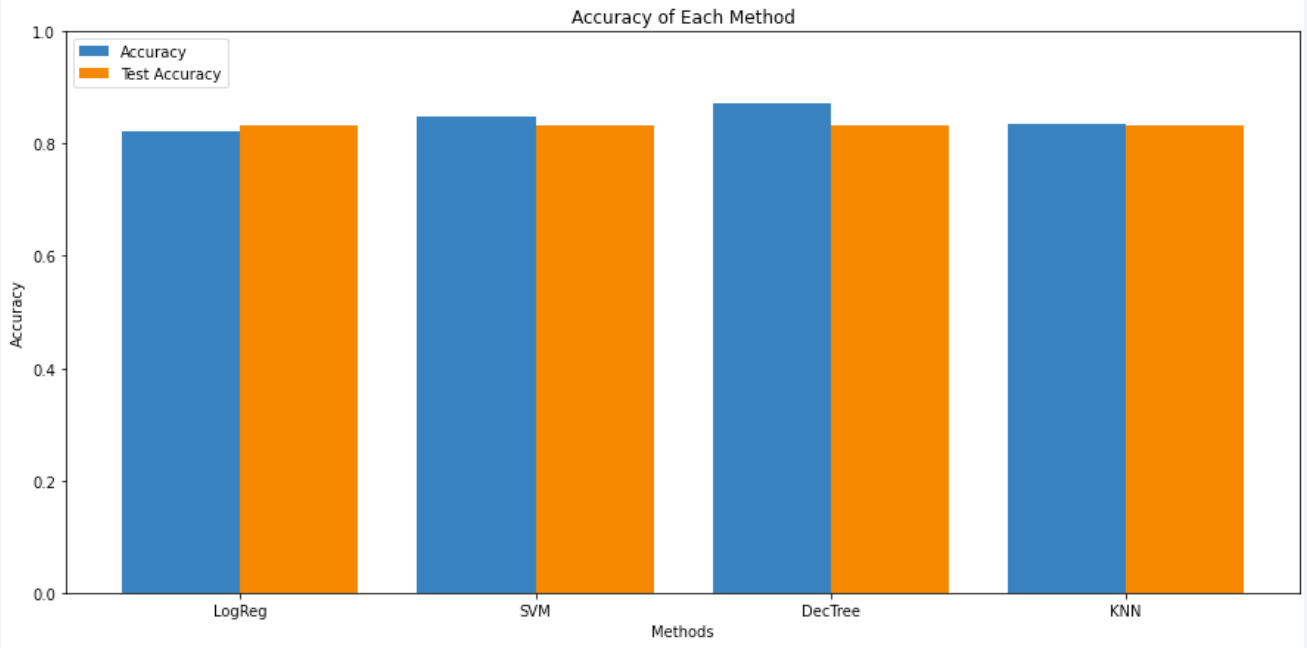
Section 5

Predictive Analysis (Classification)

Classification Accuracy

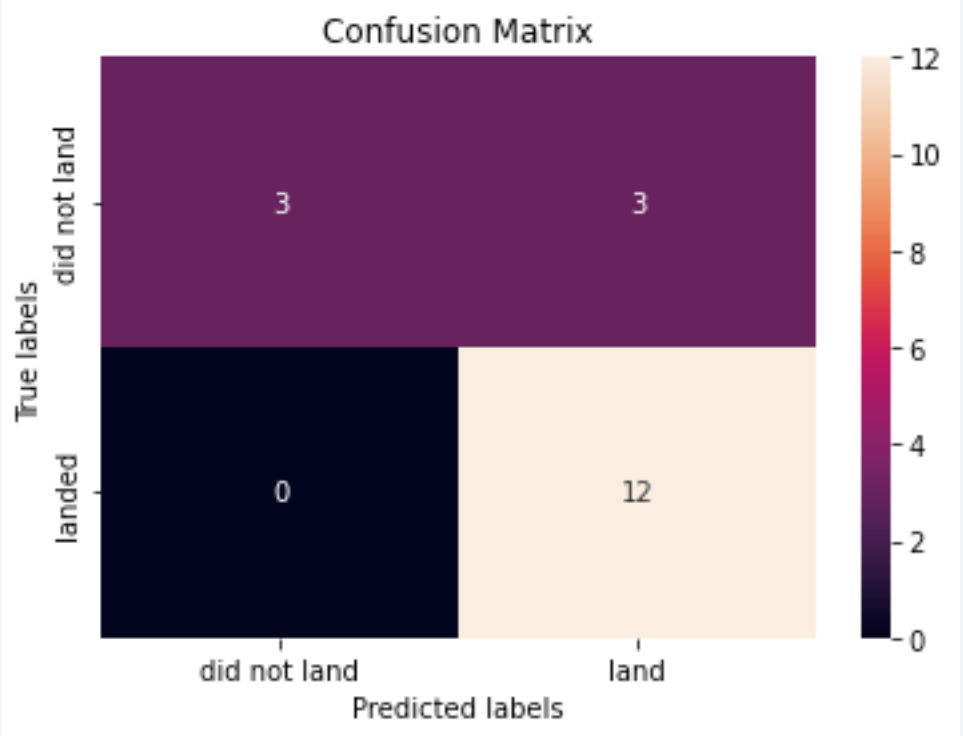
- Train and test accuracy from all four classification models tested

- The model with the highest train accuracy (around 87%) is the Decision Tree Classifier



Confusion Matrix

- Confusion matrix shows that the Decision Tree model can correctly predict all successful landed launch, but have a median performance to correctly predict failed launches



Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!