Winning Space Race with Data Science

Zhuoyi Zhao

September 18, 2021

Outline

Executive Summary

Introduction

Methodology

Results

Conclusions

Executive Summary

- In this project, we collected data from SpaceX REST API and performed data wrangling on the data.
- We also conducted exploratory data analysis (EDA) using visualization and SQL and performed interactive visual analytics using Folium and Plotly Dash.
- After standardizing and creating one-hot encoding, we performed predictive analysis using classification models including Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors models.
- The results are derived and presented through exploratory data analysis, interactive analytics demo as well as predictive analysis.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, which is much less expensive than the cost of other providers, 165 million dollars. Such substantial savings is due to the reuse of the first stage.
- This makes the determination of whether the first stage will land successfully or not significant since it corresponds to the estimation of the cost of a launch.
- In this research, we aim to predict whether the Falcon 9 first stage will land successfully or not.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Request and parse from SpaceX REST API, filter data frame, deal with missing values

Perform data wrangling

- Calculate the number of launches on each site, each orbit, mission outcome per orbit type, respectively
- Create a landing outcome label from Outcome column

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Separate the features and target, standardize the features, split into training and testing data, train classification models, tune hyperparameters, calculate the accuracy, select the best performing model

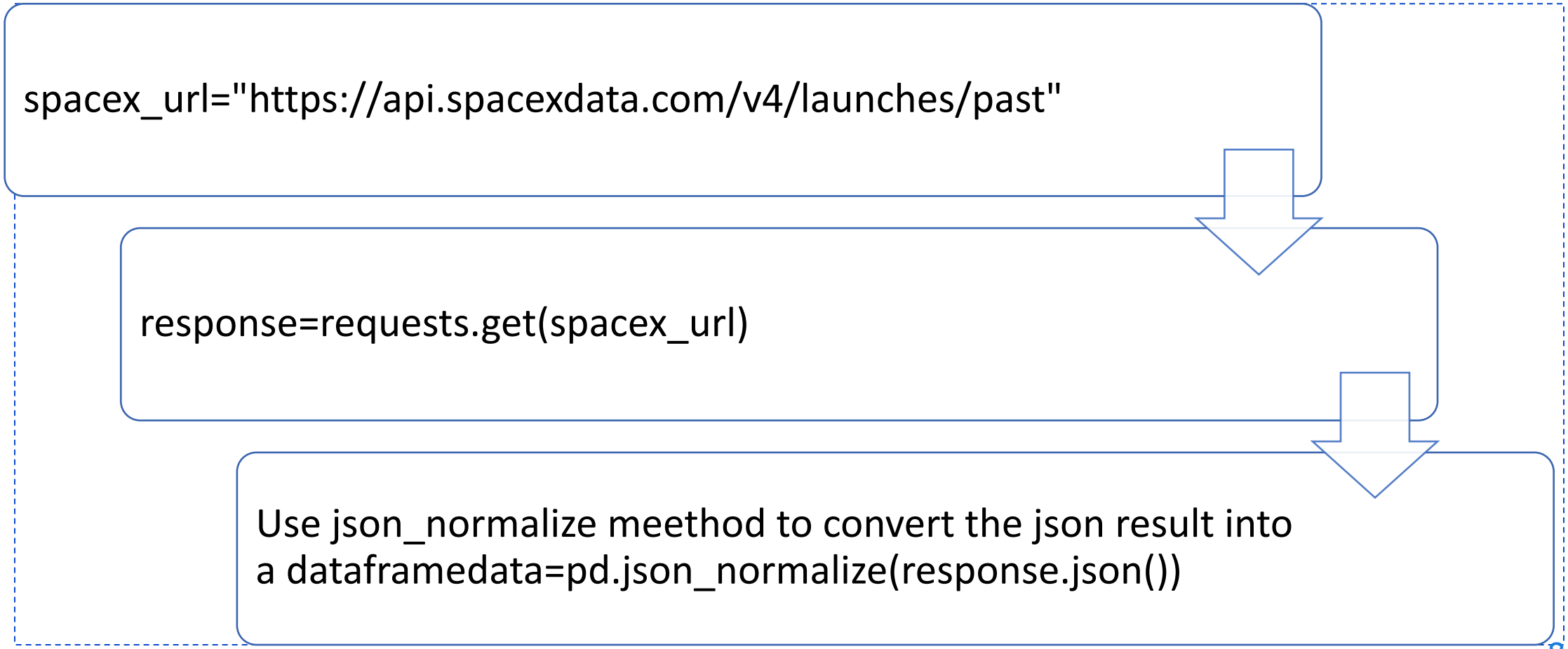
Data Collection

Request and parse the SpaceX launch data using the GET request

Filter the dataframe to only include Falcon 9 launches

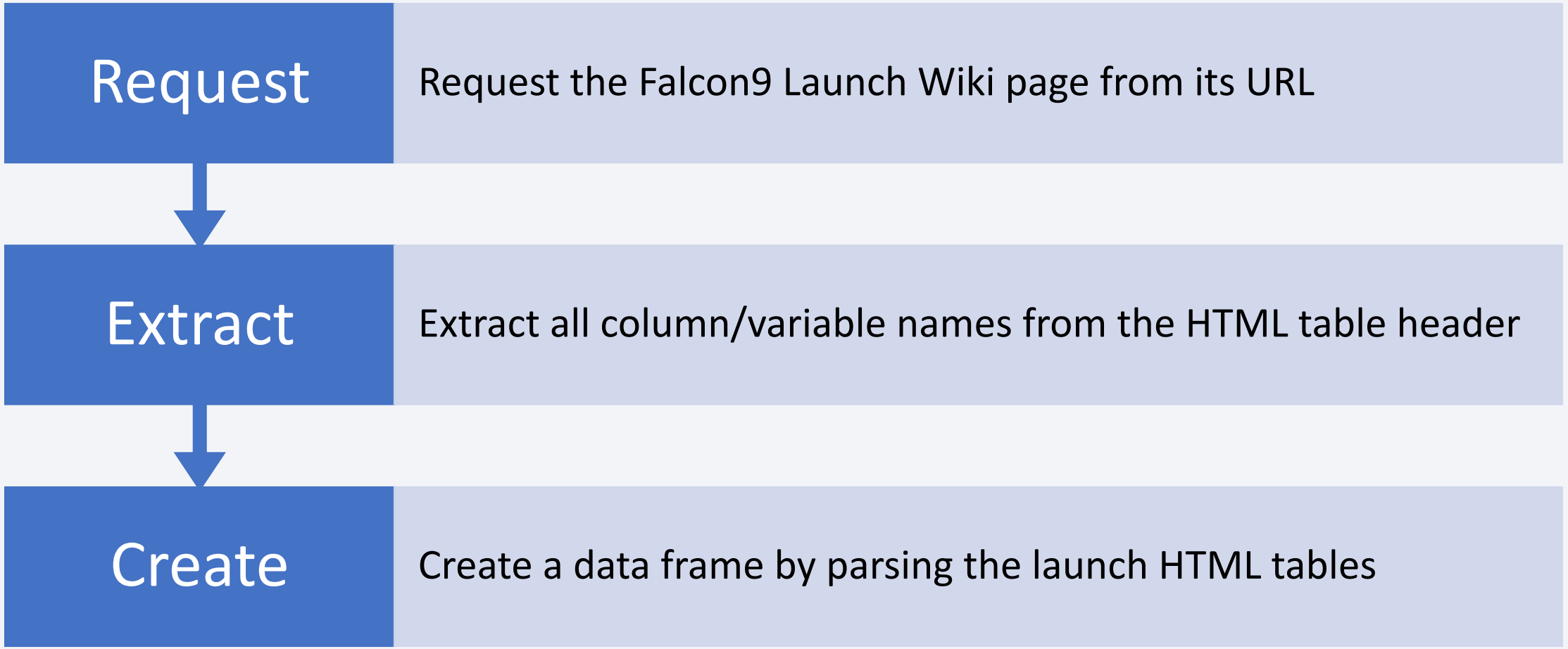
Dealing with Missing Values

Data Collection – SpaceX API



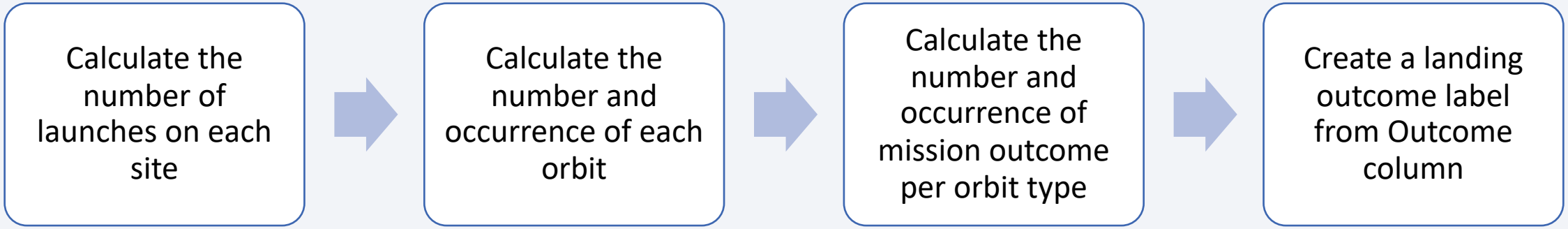
GitHub URL of the Data Collection API notebook: https://github.com/zhuoyizhao/public-ibm-applied-data-science-capstone/blob/master/Week%201%20-%20Data%20Collection%20API.ipynb

Data Collection - Scraping



GitHub URL of the Data Collection with Web Scraping notebook: https://github.com/zhuoyizhao/public-ibm-applied-data-science-capstone/blob/master/Week%201%20-%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling



GitHub URL of the EDA notebook: https://github.com/zhuoyizhao/public-ibm-applied-data-science-capstone/blob/master/Week%201%20-%20EDA.ipynb

EDA with Data Visualization

- To obtain preliminary insights about how each important variable would affect the success rate so as to select the features used for success prediction, we plotted the following charts:

  - Scatter plot showing the relationship between Flight Number and Launch Site
  - Scatter plot showing the relationship between Payload and Launch Site
  - Bar plot showing the relationship between success rate of each orbit type
  - Scatter plot showing the relationship between FlightNumber and Orbit type
  - Scatter plot showing the relationship between Payload and Orbit type
  - Line plot showing the launch success yearly trend

GitHub URL of the EDA with Data Visualization notebook: https://github.com/zhuoyizhao/public-ibm-applied-data-science-capstone/blob/master/Week%202%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was acheived.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. 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

GitHub URL of the EDA with SQL notebook: https://github.com/zhuoyizhao/public-ibm-applied-data-science-capstone/blob/master/Week%202%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- The following map objects were created and added to a folium map
  - Markers - to show the location of each launch site
  - Circles - to highlight the location of each launch site and the name of each launch site
  - Mouse Position – to get the coordinate for a mouse over a point on a map
  - Lines - to show the distance between two locations

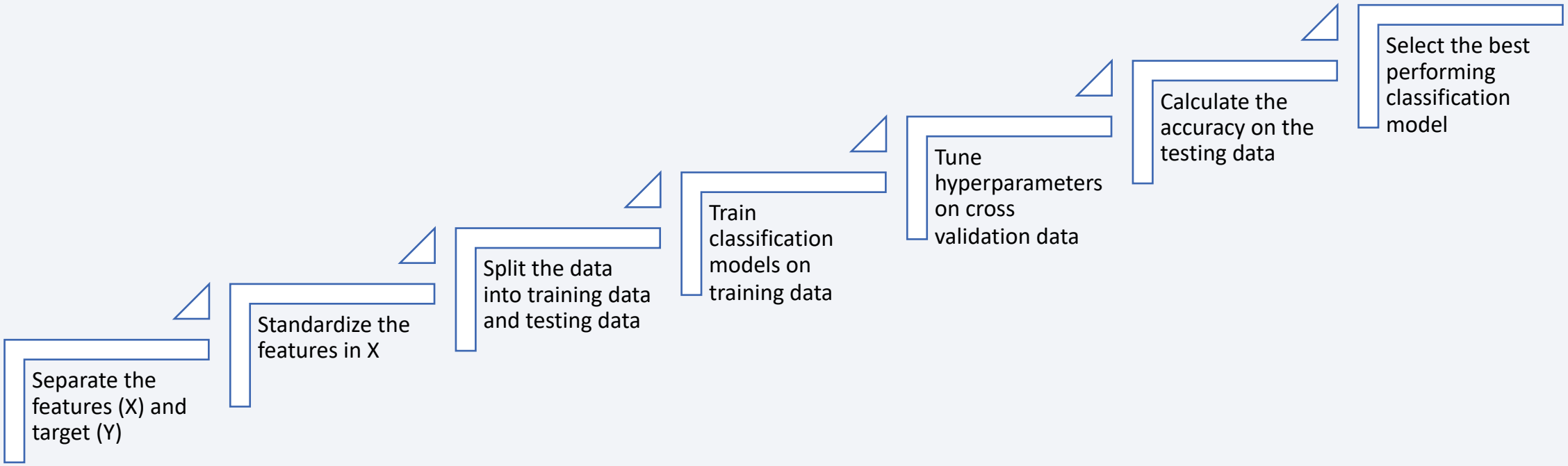
GitHub URL of the Interactive Visual Analytics with Folium notebook: https://github.com/zhuoyizhao/public-ibm-applied-data-science-capstone/blob/master/Week%203%20-%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- A pie chart that shows
  - The percentages of total success launches of each launch site if no specific launch site is selected in the dropdown box
  - The percentages of success and failure for the selected launch if a specific launch site is selected in the dropdown box
- A scatter plot that shows the relationship between payload mass and class colored by booster versions for all launch sites or for a given launch site. The range of payload mass can be adjusted by the range slider.

GitHub URL of the Plotly Dash Lab: https://github.com/zhuoyizhao/public-ibm-applied-data-science-capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)



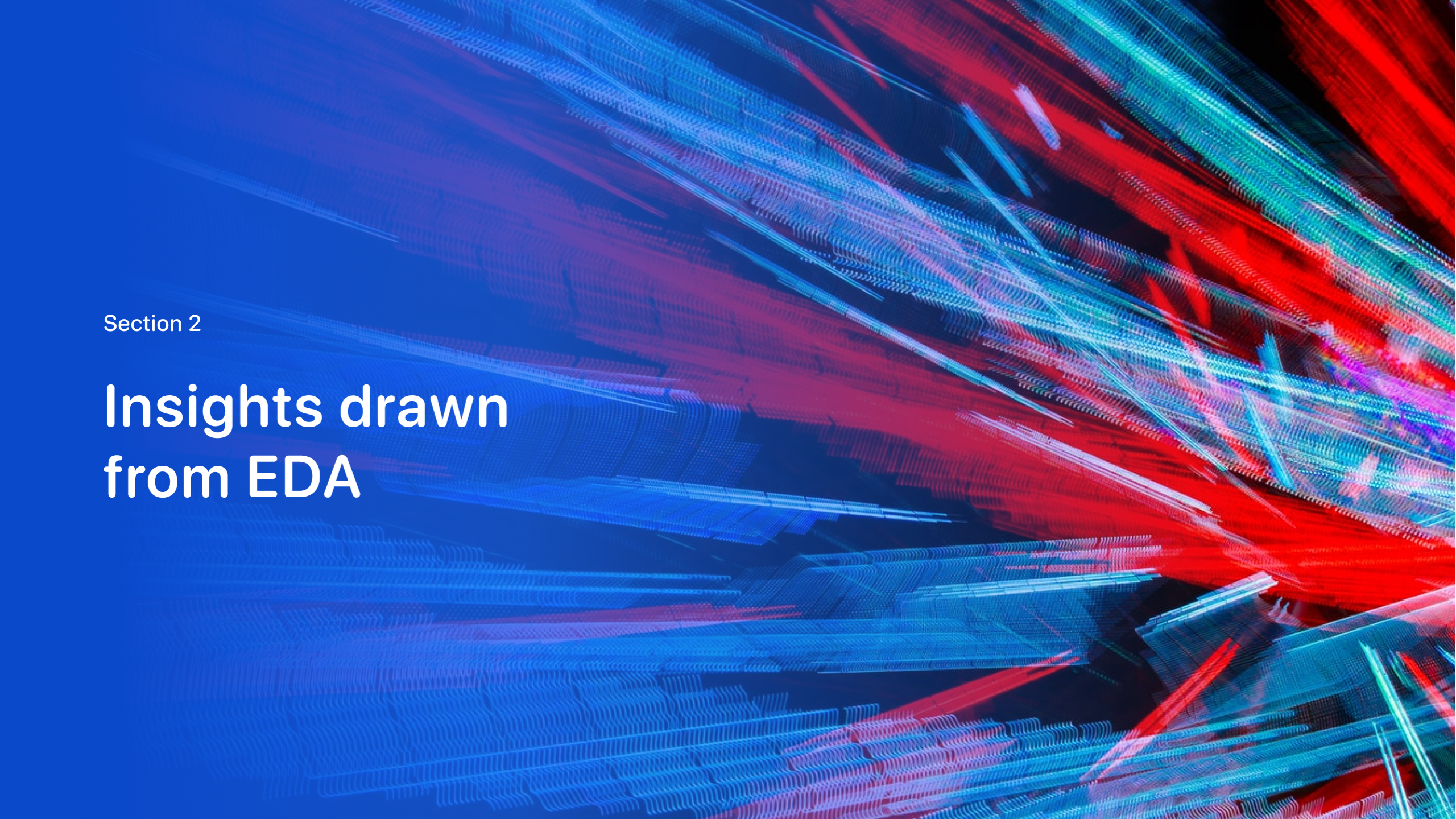
GitHub URL of the Machine Learning Prediction notebook: https://github.com/zhuoyizhao/public-ibm-applied-data-science-capstone/blob/master/Week%204%20-%20Machine%20Learning%20Prediction.ipynb

Results

Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

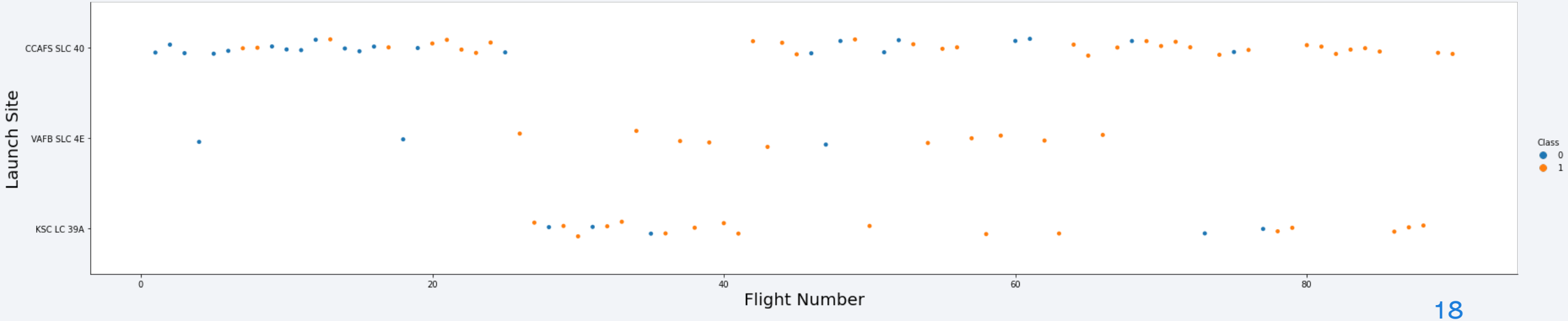


Section 2

Insights drawn from EDA

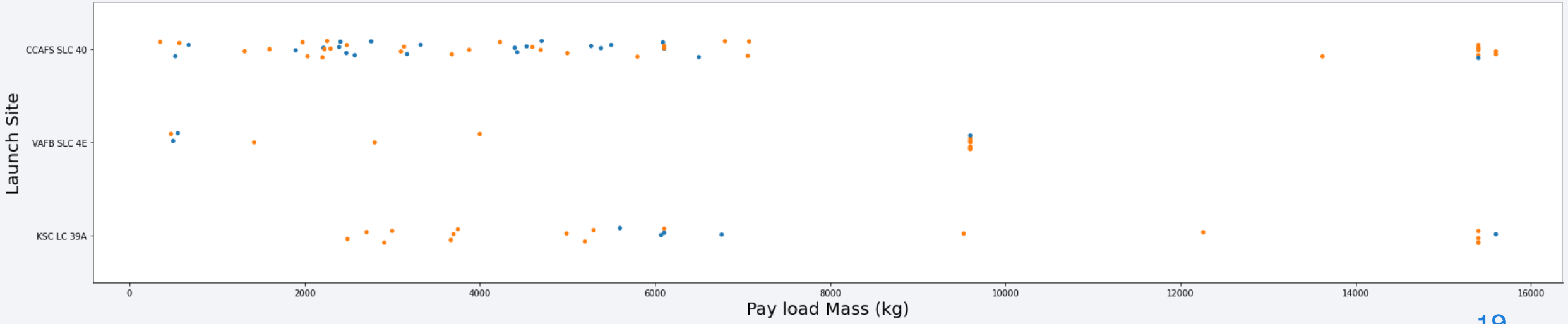
Flight Number vs. Launch Site

- The success rate increases as time goes by for each of the launching sites in general.
- Rockets with a flight number greater than 65 are mainly launched by CCAFS SLC 40 and KSC LC 39A.



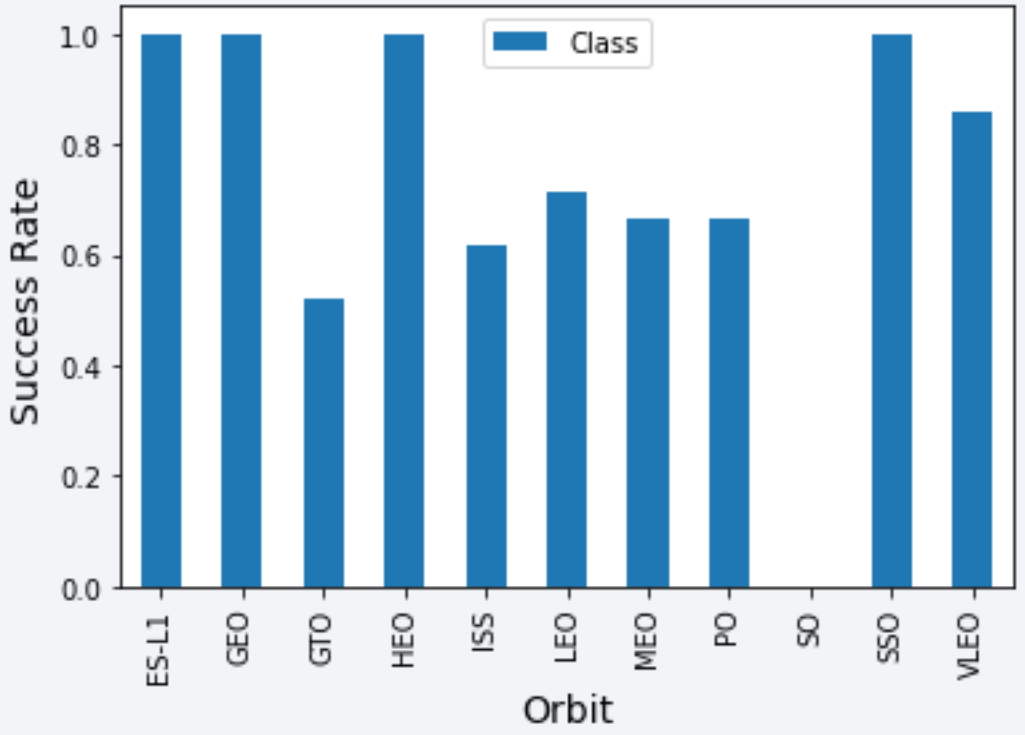
Payload vs. Launch Site

- Heavy payloads are mainly launched by CCAFS SLC 40 and KSC LC 39A and have a higher success rate.



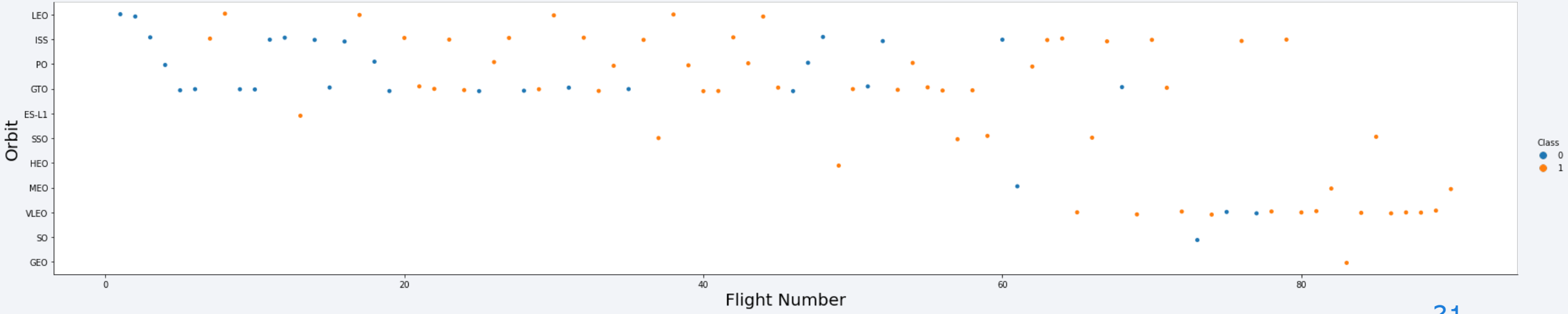
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, and SSO have a high success rate of 100%.
- The success rate of the rest orbits are between 50% and 70% except for orbit SO.
- Orbit SO has the lowest success rate, 0%.



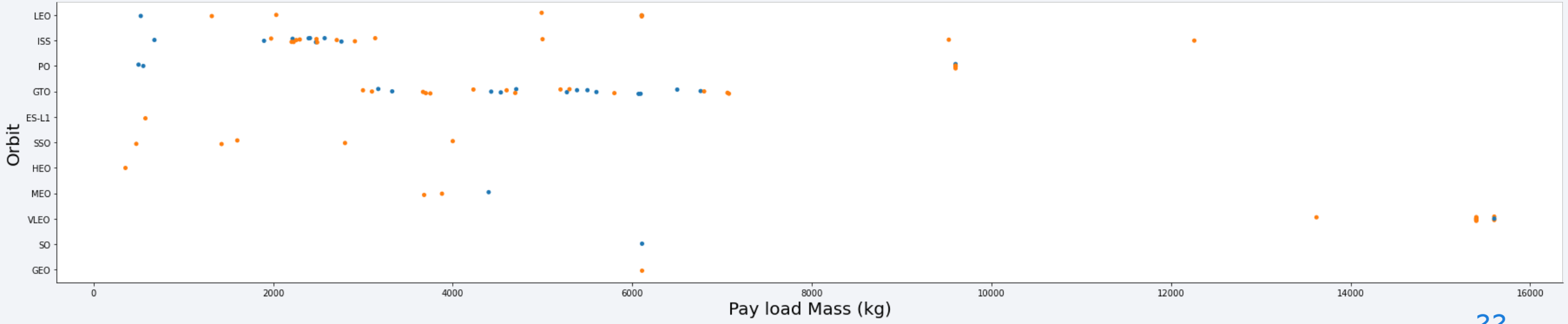
Flight Number vs. Orbit Type

- LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.



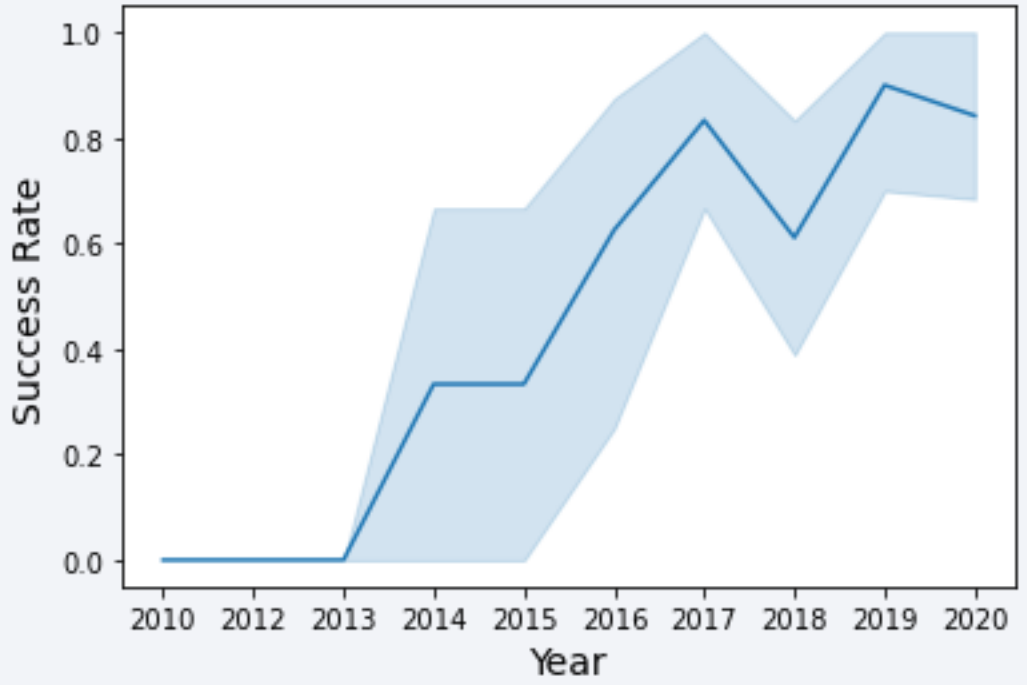
Payload vs. Orbit Type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- The success rate has been increasing between 2013 and 2020 in general, but there was a sudden drop in year 2018.



All Launch Site Names

- There are four launch sites in total, namely, CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

Display the names of the unique launch

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

```
 * ibm_db_sa://hsp36639:***@ba99
Done.
```

Out[4]:

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

Launch Site Names Begin with “CCA”

- The examples of launches from launch sites beginning with “CCA” are as follows:

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

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE launch_site LIKE 'CCA%'
LIMIT 5
```

```
 * ibm_db_sa://hsp36639:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Total Payload Mass

- The total payload carried by boosters from NASA (CRS) is 45,596 kg.

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

```
%%sql
SELECT SUM(payload_mass__kg_)
FROM SPACEXTBL
WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://hsp36639:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a0
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,534 kg.

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

```
%%sql
SELECT AVG(payload_mass__kg_)
FROM SPACEXTBL
WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://hsp36639:***@ba99a9e6-d59e-4883-8fc0-c
Done.
```

1
2534

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was on December 22, 2015.

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

Hint:Use min function

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

```
 * ibm_db_sa://hsp36639:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clo
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were: JCSAT-14, JCSAT-16, SES-10, and SES-11 / EchoStar 105.

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

```
%%sql
SELECT payload
FROM SPACEXTBL
WHERE (landing__outcome = 'Success (drone ship)') AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
```

```
 * ibm_db_sa://hsp36639:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdom
1/bludb
Done.
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Among all the mission outcomes, 99 of them were successful, 1 of them was successful but payload status unclear, and 1 failed in flight.

List the total number of successful and failu

```
%%sql
SELECT mission_outcome, COUNT(*)
FROM SPACEXTBL
GROUP BY mission_outcome
```

```
 * ibm_db_sa://hsp36639:***@ba99a9e
1/bludb
Done.
```

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

Booster Versions Carried Maximum Payload

- The names of the booster version which have carried the maximum payload mass are shown on the right.

List the names of the booster_versions which

```
%%sql
SELECT DISTINCT booster_version
FROM SPACEXTBL
WHERE payload_mass__kg_ =
(SELECT MAX (payload_mass__kg_)
FROM SPACEXTBL)
ORDER BY booster_version
```

```
 * ibm_db_sa://hsp36639:***@ba99a9e6-
1/bludb
Done.
```

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

- In year 2015, 2 launches failed in drone ship. They were both launched at site CCAFS LC-40. One of them had a booster of version F9 v1.1 B1012, and the other had F9 v1.1 B1015.

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 DATE LIKE '2015%'
```

```
 * ibm_db_sa://hsp36639:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.da
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

- Between the date 2010-06-04 and 2017-03-20, 10 launches had no attempt, 5 failed and succeeded in drone ship, respectively, 3 launches had controlled (ocean) and success (ground pad), respectively, 2 launches had failure (parachute) and uncontrolled (ocean), respectively, and finally, 1 was precluded (drone ship).

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

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

```
 * ibm_db_sa://hsp36639:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

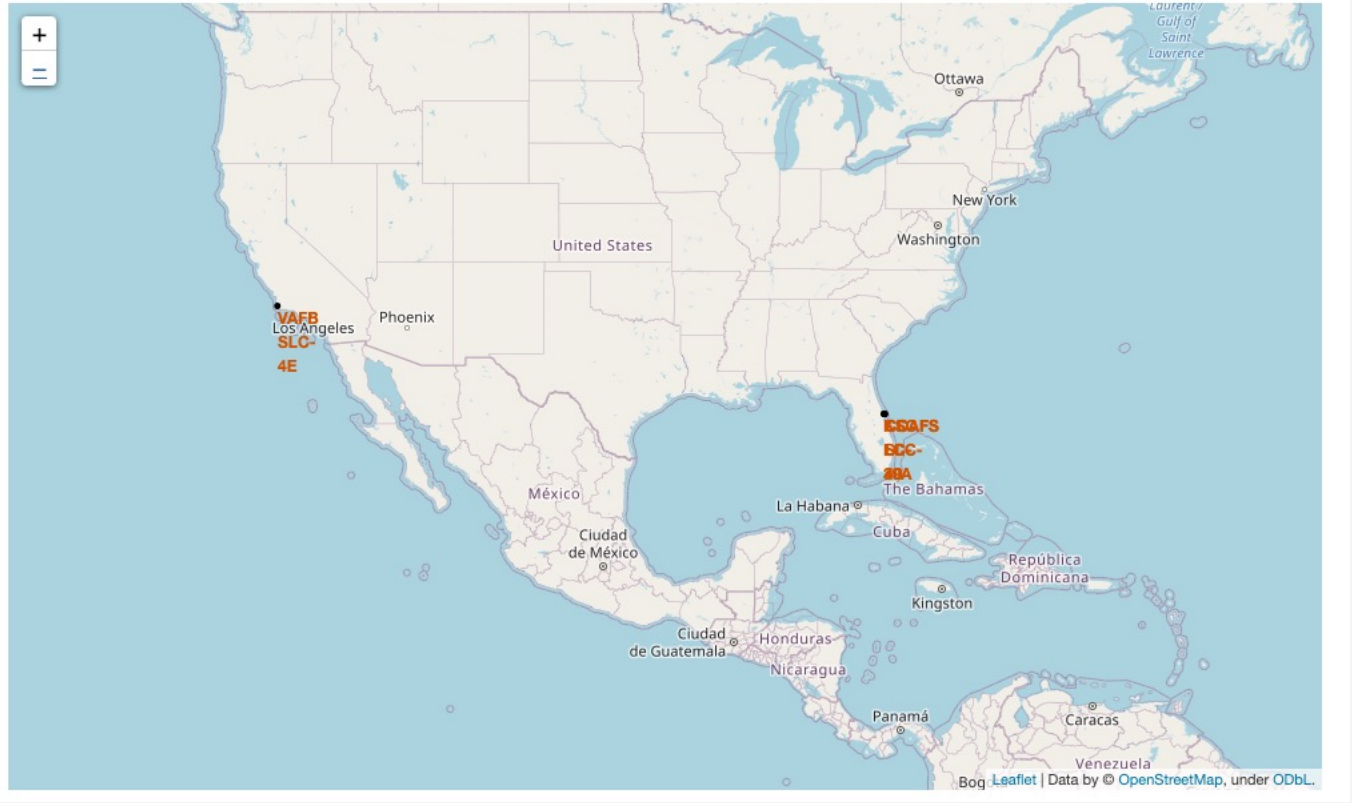
landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

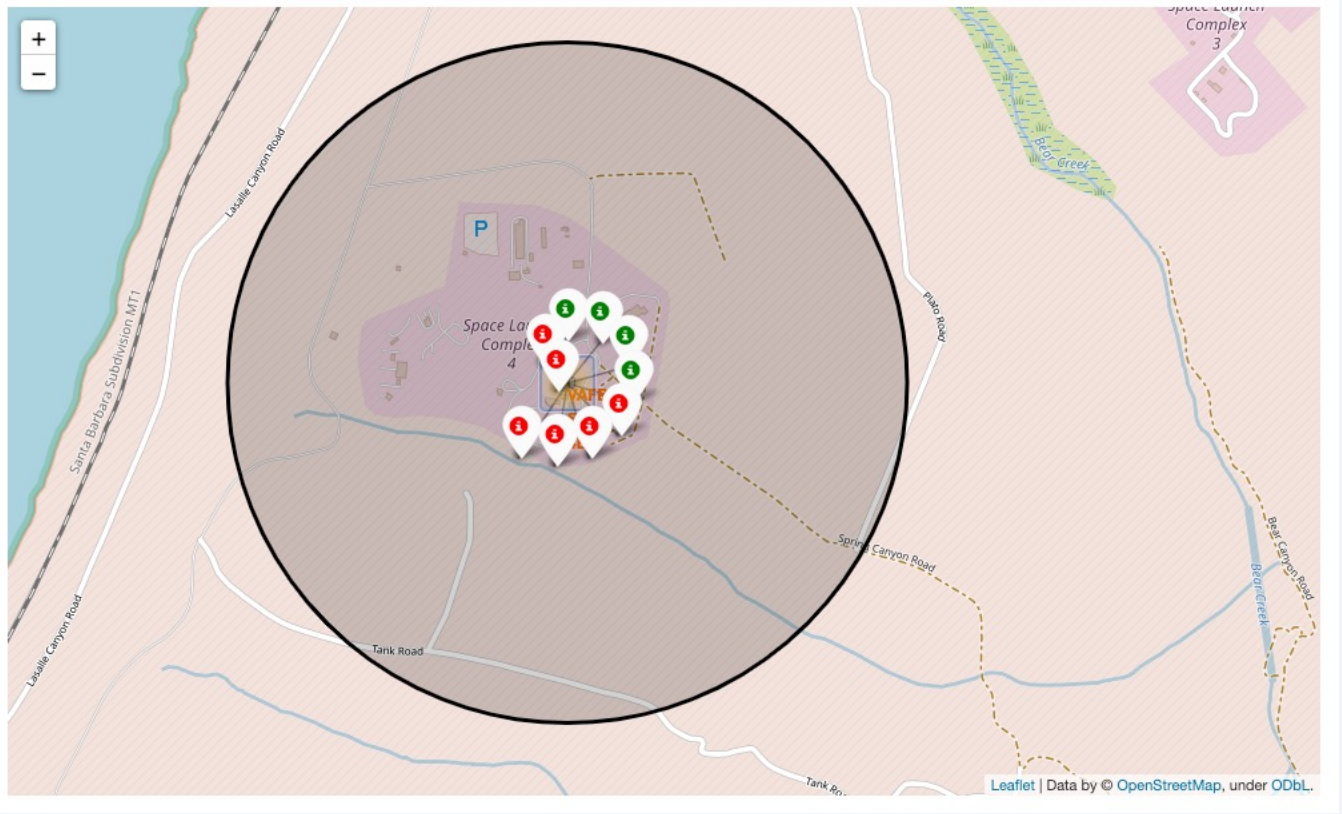
Locations of Launch Sites

- Among the four launch sites, one is in California (VAFB SLC-4E), and three are in Florida (CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A). In particular, CCAFS LC-40 and CCAFS SLC-40 are very close to each other.



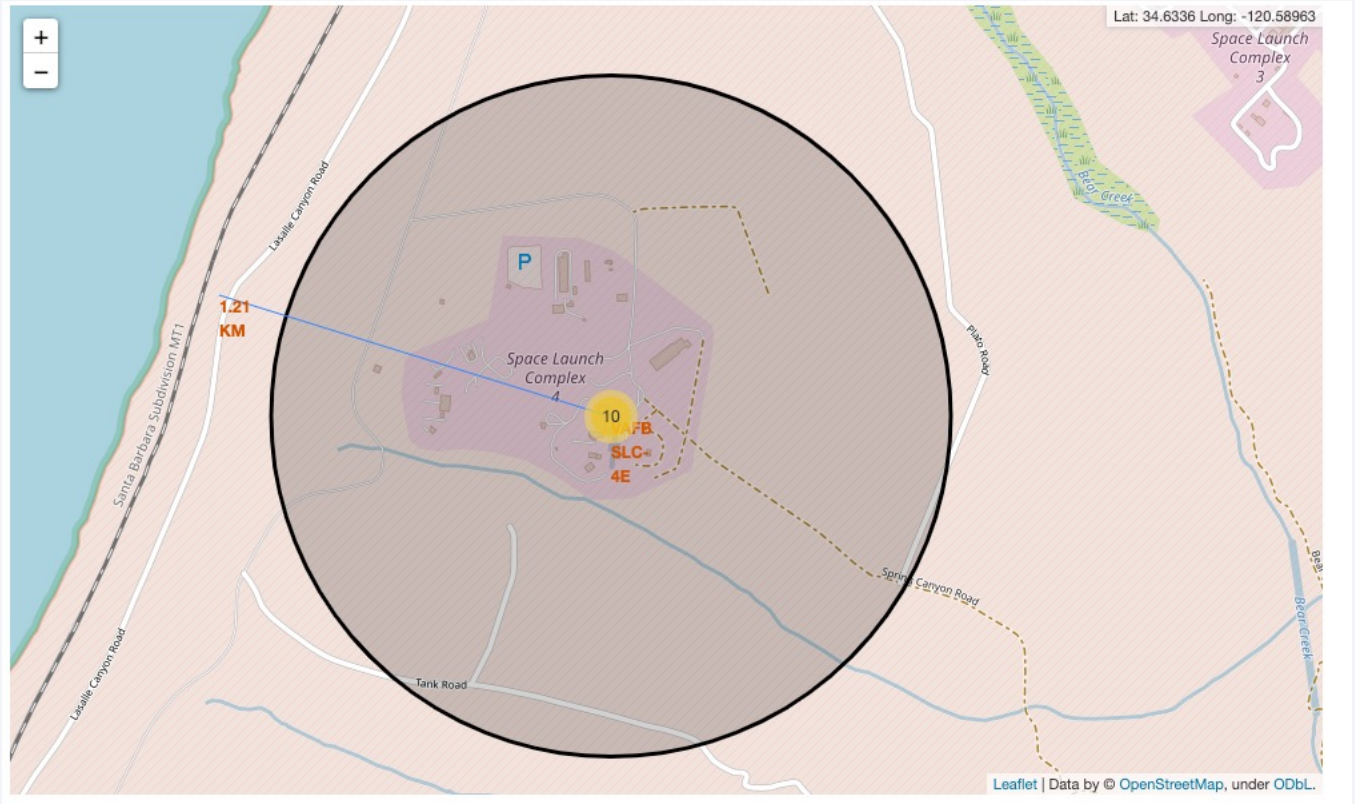
Color-Labeled Launch Outcomes for Site VAFB SLC-4E

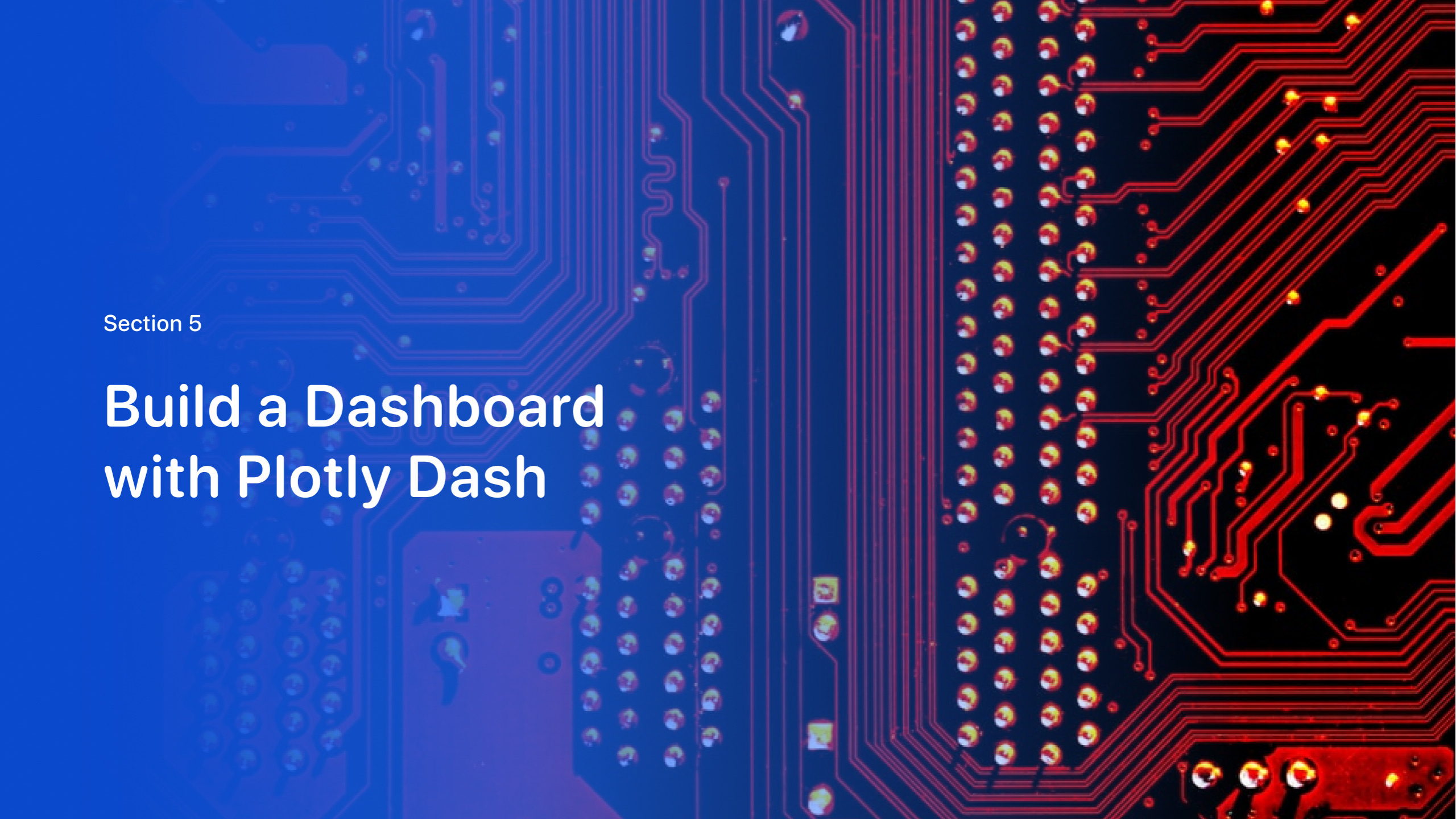
- For site VAFB SLC-4E, 4 out of 10 launches were successful (in green), and 6 out of 10 launches were failed (in red).



Distance from Site VAFB SLC-4E to Closest Railway

- The distance from Site VAFB SLC-4E to the closest railway is round 1.21 km.



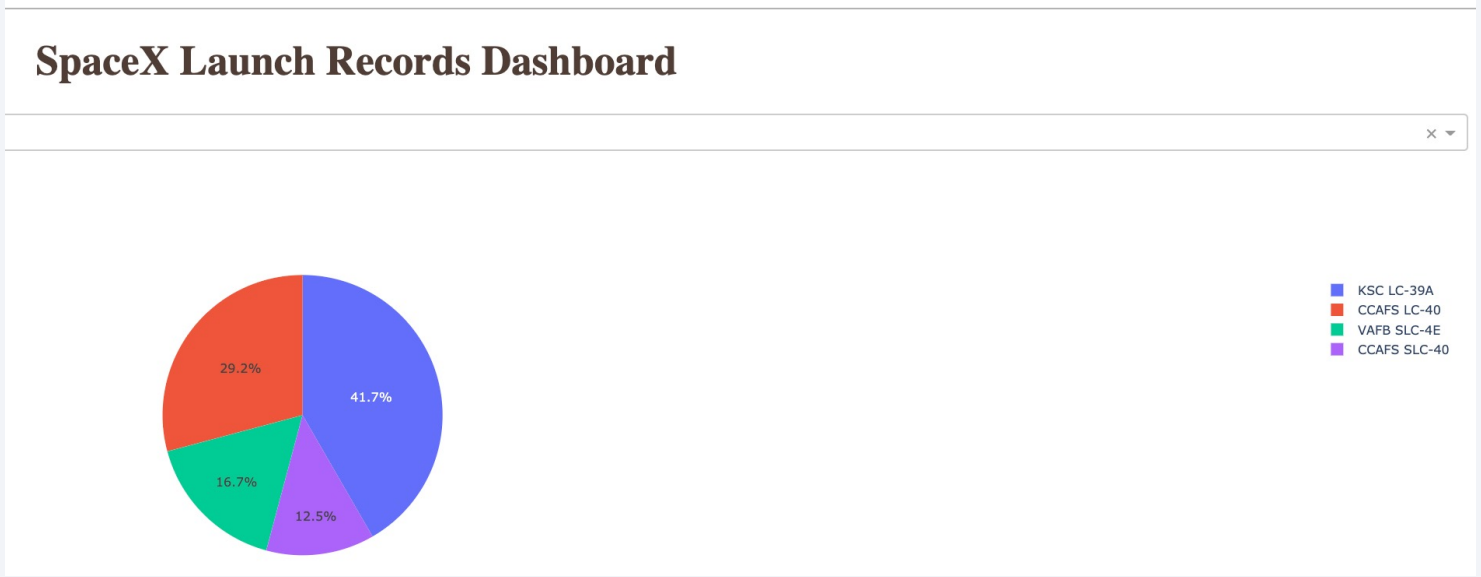


Section 5

Build a Dashboard with Plotly Dash

Total Success Launches By Site

- Site KSC LC-39A had the most total successful launches, taking 41.7% of all the successful launches.
- The second and third most sites were CCAFS LC-40 (29.2%) and VAFB SLC-4E (16.7%), respectively.
- Site CCAFS SLC-40 has the least, taking 12.5%.



Total Success Launches for Site KSC LC-39A

- Site KSC LC-39A has the highest launch success ratio, 76.9%.

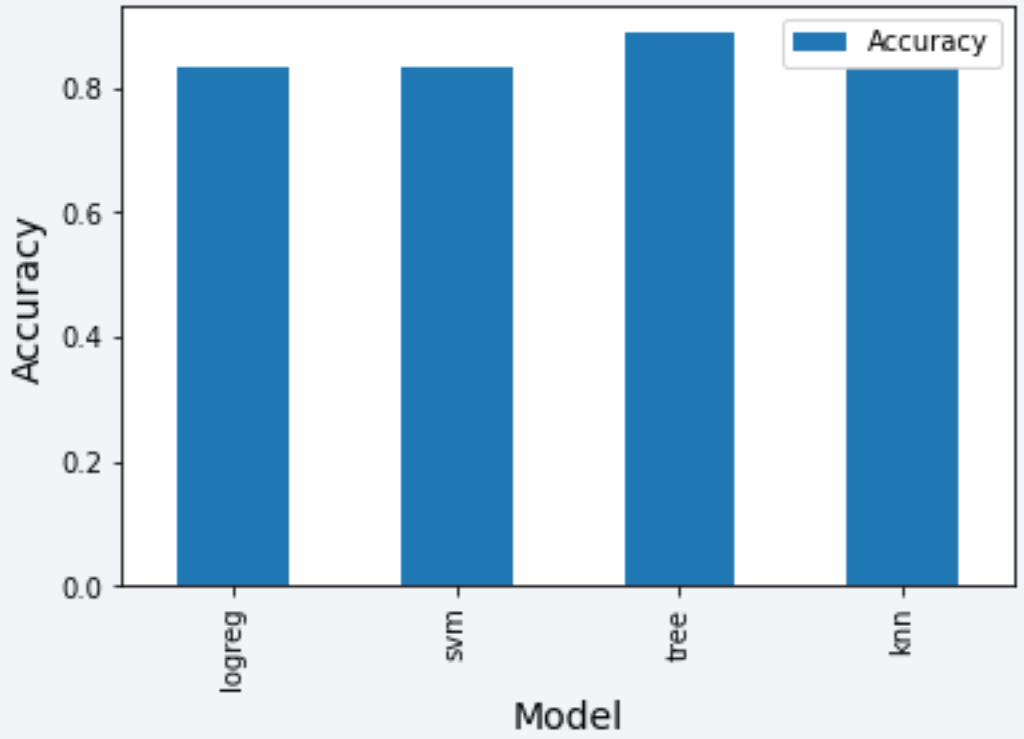
Correlation between Payload and Launch Outcome

- For payloads within range 2,000 kg – 5,500 kg, booster version FT had the largest success rate.

Section 6

Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree model has the highest classification accuracy, 88.89%.

Confusion Matrix

- Best performing model: Decision Tree
- The confusion matrix shows that it correctly predict 11 launches that landed, and 5 launched that did not land. The false positive and false negative categories had 1, respectively.

Conclusions

- There are four launch sites in total, namely, CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E. Site KSC LC-39A has the highest launch success ratio, 76.9%.
- Orbits ES-L1, GEO, HEO, and SSO have a high success rate of 100%.
- The success rate has been increasing between 2013 and 2020 in general, but there was a sudden drop in year 2018..
- For payloads within range 2,000 kg – 5,500 kg, booster version FT had the largest success rate.
- Decision Tree model performs best on the prediction of whether the Falcon 9 first stage will land successfully or not, with an accuracy around 88.89%.

Thank you!