IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Juan Carlos Garzon Pico
03 April 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection API
  - Data Collection with Web Scraping
  - Data Wrangling
  - EDA with SQL
  - EDA with Data Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Ploty Dash
  - Machine Learning Prediction
- Summary of all results
  - EDA results
  - Visual Data Analytics and Dashboards
  - Machine Learning (Classification)

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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.

- Problems you want to find answers

Determine the price of each launch. You will do this by gathering information about Space X and creating dashboards for your team. You will also determine if SpaceX will reuse the first stage. Instead of using rocket science to determine if the first stage will land successfully, you will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data were obtained from 2 resources:
    - Space X API: https://api.spacexdata.com/v4
    - Web scraping: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Perform exploratory Data Analysis and determine Training Labels. Create a column for the class, and standardize the data. After, split into training data and test data to find the best Hyperparameter for SVM, Classification Trees, and Logistic Regression. Finally, find the method that performs best using test data.

Data Collection

- Objectives

To make a get request to the SpaceX API. You will also do some basic data wrangling and formating.

  - Request to the SpaceX API
  - Clean the requested data

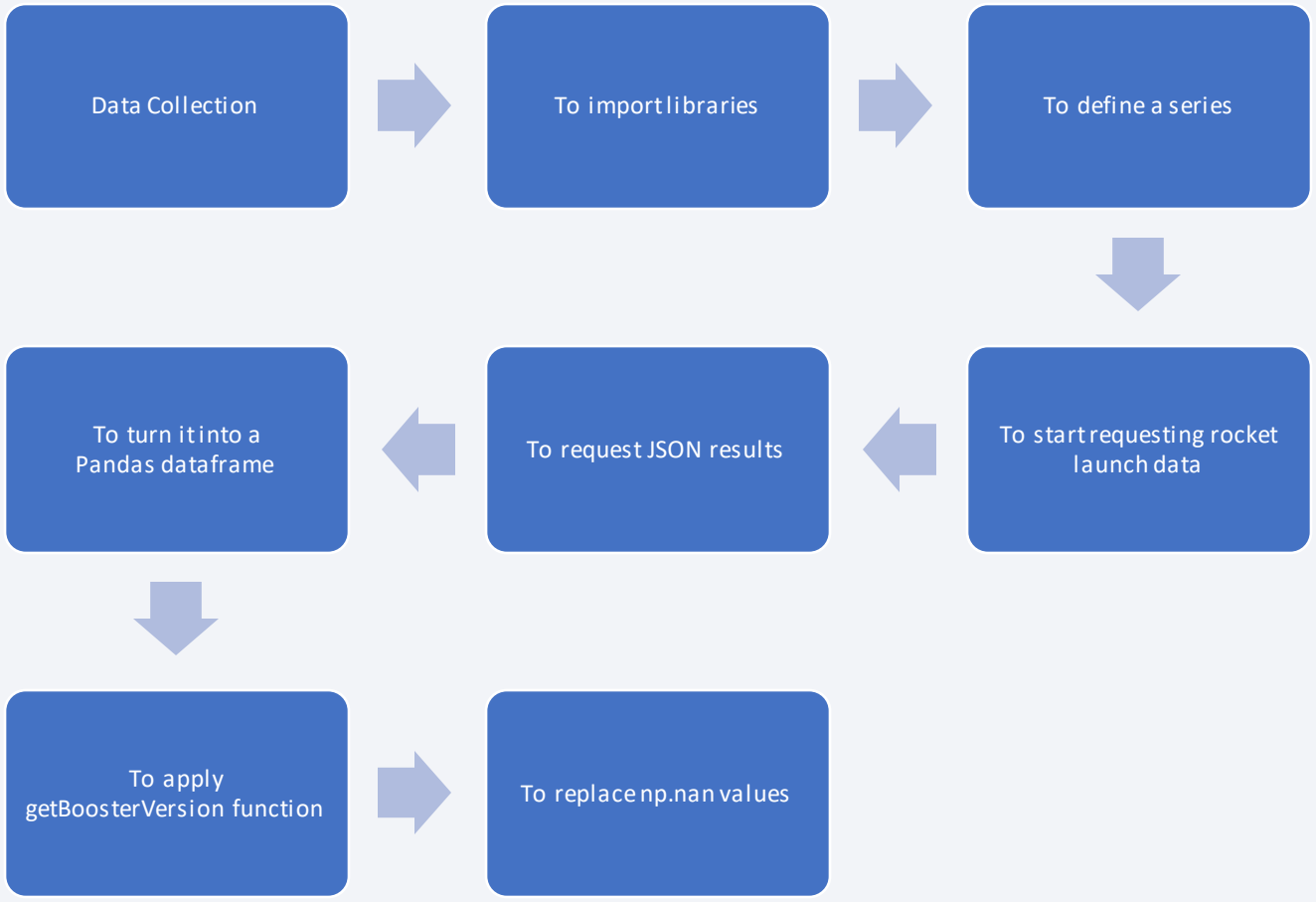
- Description of how data was collected:

  - We imported libraries into the lab, defined a series of helper functions that helped us use the API to extract information using identification numbers in the launch data, then started requesting rocket launch data from SpaceX API with the space data URL.
  - Secondly, we requested JSON results more consistent, we decoded the response content as a Json using .json() and turned it into a Pandas dataframe using .json_normalize(). After we applied getBoosterVersion function method to get the booster version. Finally, we let construct our dataset using the data we have obtained.
  - Also, we calculated below the mean for the PayloadMass using the .mean(). Then we used the mean and the .replace() function to replace np.nan values in the data with the mean you calculated.

Data Collection – SpaceX API

- Here is the GitHub URL of the completed notebook:

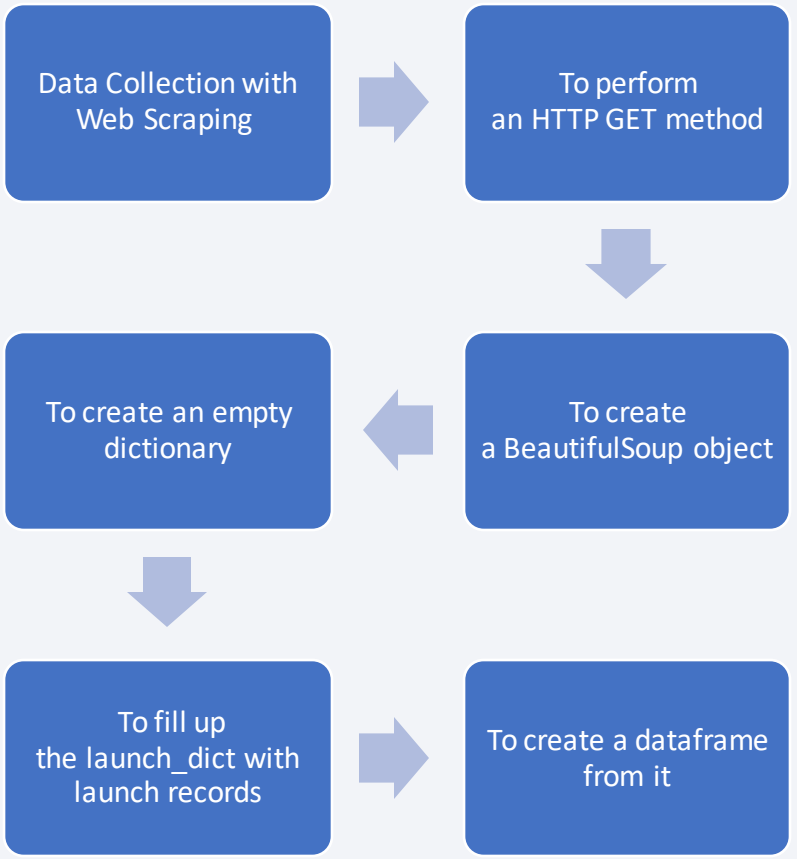
  https://github.com/Juank0621/Applied-Data-Science-Capstone/blob/374c84395434dd9b5ba6adaecc90f582217774c9/1.%20SpaceX%20Falcon%209%20Data%20Collection%20API.ipynb



Data Collection - Scraping

- Web scrap Falcon 9 launch records with BeautifulSoup:
  - Extract a Falcon 9 launch records HTML table from Wikipedia
  - Parse the table and convert it into a Pandas data frame
- Here is the GitHub URL of the completed notebook:

  https://github.com/Juank0621/Applied-Data-Science-Capstone/blob/374c84395434dd9b5ba6adaecc90f582217774c9/2.%20Space%20X%20Falcon%209%20Data%20Collection%20with%20Web%20Scraping.ipynb

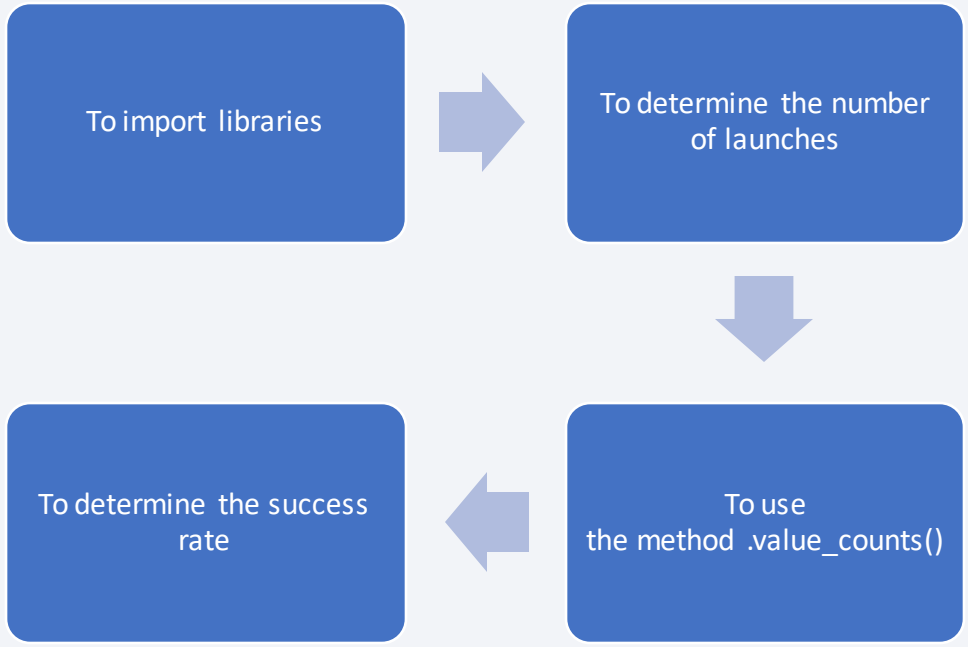


Data Wrangling

- We imported libraries such as panda and numpy, after, we determined the number of launches on each site, we also used the method .value_counts() to determine some patterns in the data and finally we determined the success rate.

- Here is the GitHub URL of the completed notebook:

  https://github.com/Juank0621/Applied-Data-Science-Capstone/blob/909a868bc4e307761ddb05f573cd14aeb22e8f9f/3.%20Space%20X%20Falcon%209%20Data%20Wrangling.ipynb

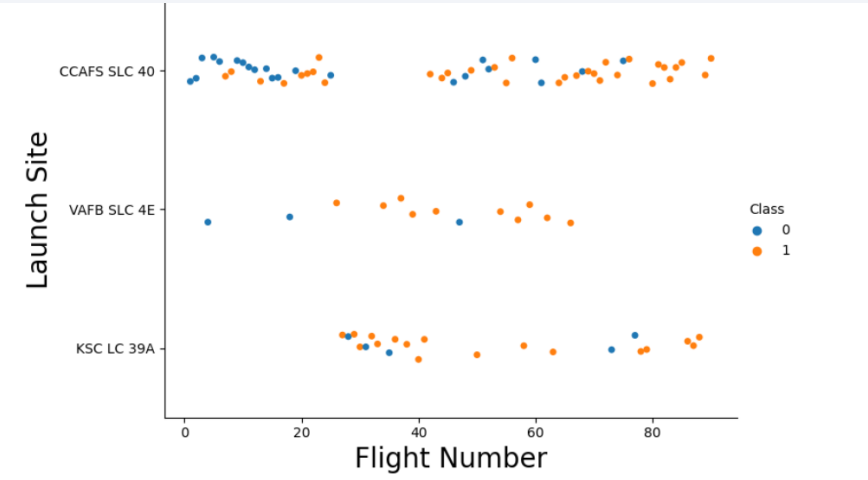


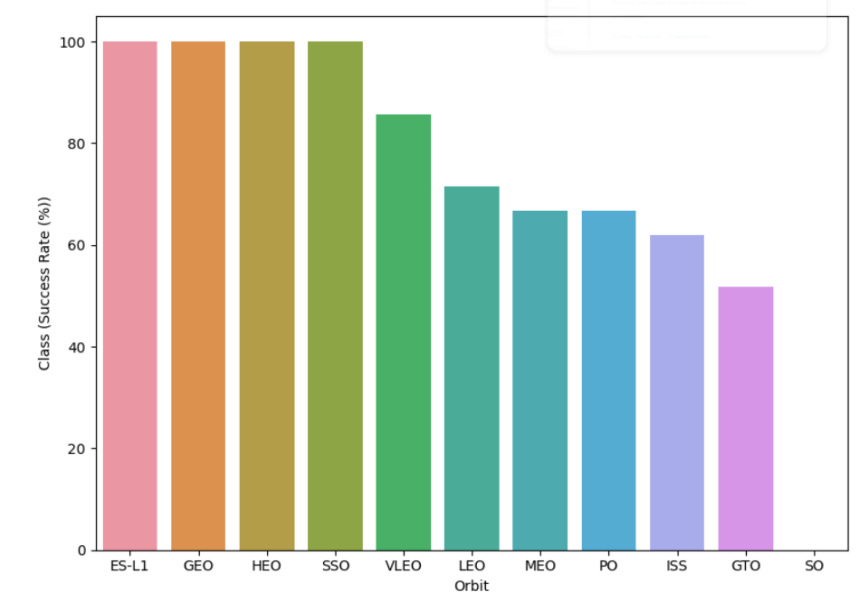
EDA with Data Visualization

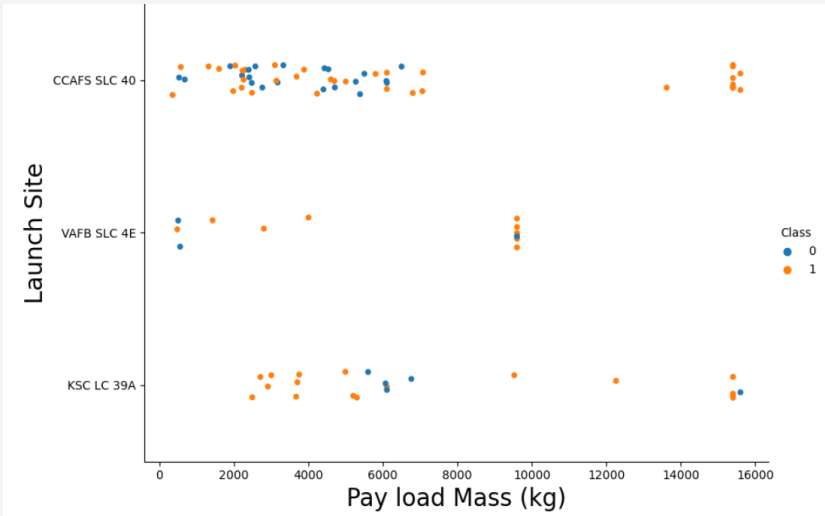
- Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
  - Exploratory Data Analysis
  - Preparing Data Feature Engineering
- First of all, we used **scatter plots** to visualize the relationship between Flight Number and Launch Site, Payload and Launch Site, FlightNumber and Orbit type, and Payload and Orbit type.
- Secondly, we used **bar chart** to visual the relationship between the success rate of each orbit.
- Thirdly, we used **line plot** to visualize the launch success yearly trend.
- Here is the GitHub URL of the completed notebook:

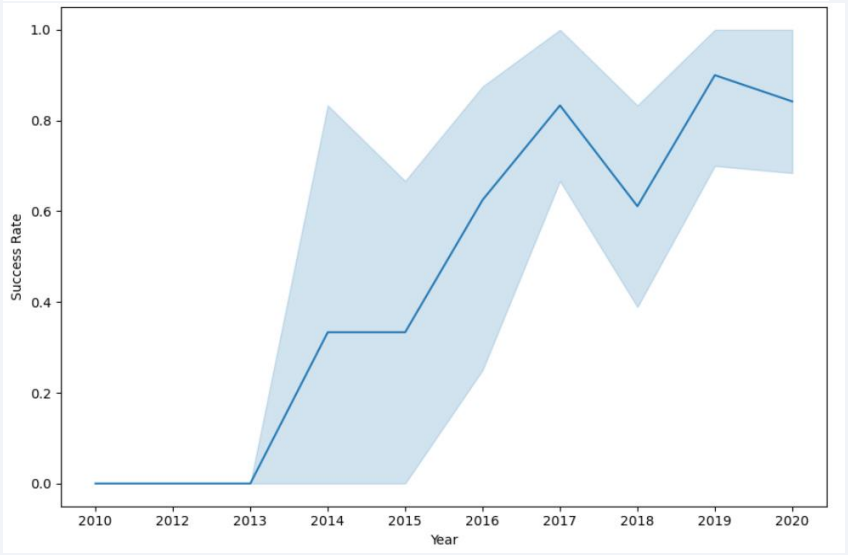
  https://github.com/Juank0621/Applied-Data-Science-Capstone/blob/909a868bc4e307761ddb05f573cd14aeb22e8f9f/5.%20Space%20X%20Falcon%209%20EDA%20with%20Data%20Visualization.ipynb

EDA with Data Visualization









EDA with SQL

- SQL queries for EDA:
  - Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

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

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%';
```

EDA with SQL

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

```
%sql SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE "Landing _Outcome" = "Success (ground pad)";
```

- 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 DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOA
```

- List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

```
%sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

EDA with SQL

- 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.

```
%sql SELECT substr(Date,7,4), substr(Date, 4, 2),"Booster_Version", "Launch_Site", Payload, "PAYLOAD_MASS__KG_", "Mission_Outcome", "Landing _Outcome"
```

- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%sql SELECT "Landing _Outcome", COUNT(*) AS QTY FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' AND (Date BETWEEN '04-06-2010' AND '20-03-2017
```

- Here is the GitHub URL of the completed notebook:

https://github.com/Juank0621/Applied-Data-Science-Capstone/blob/f410e5d5ae161cf25ced60fd11cb4908b49edf45/4.%20Space%20X%20Falcon%209%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines, and marker clusters were created and added to a folium map
  - Marker for each launch site on the site map.
  - Use a circle to add a highlighted circle area with a text label on a specific coordinate.
  - Marker Cluster indicates many launch records which have the exact same coordinate.
  - Lines were used to indicate distances between two coordinates.
- Here is the GitHub URL of the completed notebook:

  https://github.com/Juank0621/Applied-Data-Science-Capstone/blob/095c595557e3c02b2eaa289fe1b665003008868e/6.%20Space%20X%20Falcon%209%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

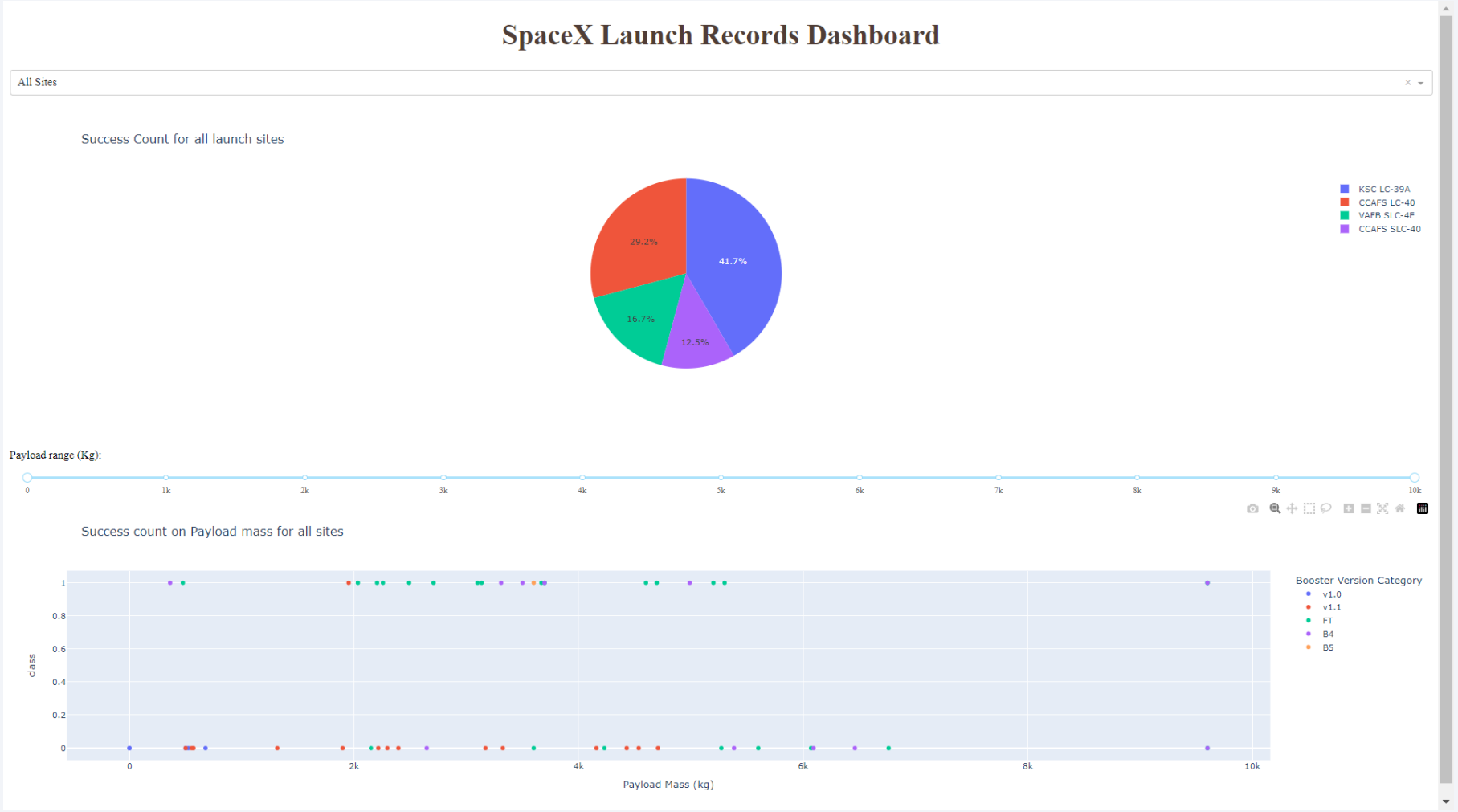
Build a Dashboard with Plotly Dash

- Build and interactive dashboard application with Plotly dash:
  - Add a Launch Site Drop-down Input Component
  - Add a callback function to render success-pie-chart based on selected site dropdown
  - Add a Range Slider to Select Payload
  - Add a callback function to render the success-payload-scatter-chart scatter plot

- Here is the GitHub URL of the completed notebook:

  https://github.com/Juank0621/Applied-Data-Science-Capstone/blob/909a868bc4e307761ddb05f573cd14aeb22e8f9f/7.%20Space%20X%20Falcon%209%20Interactive%20Dashboard%20with%20Ploty%20Dash.py

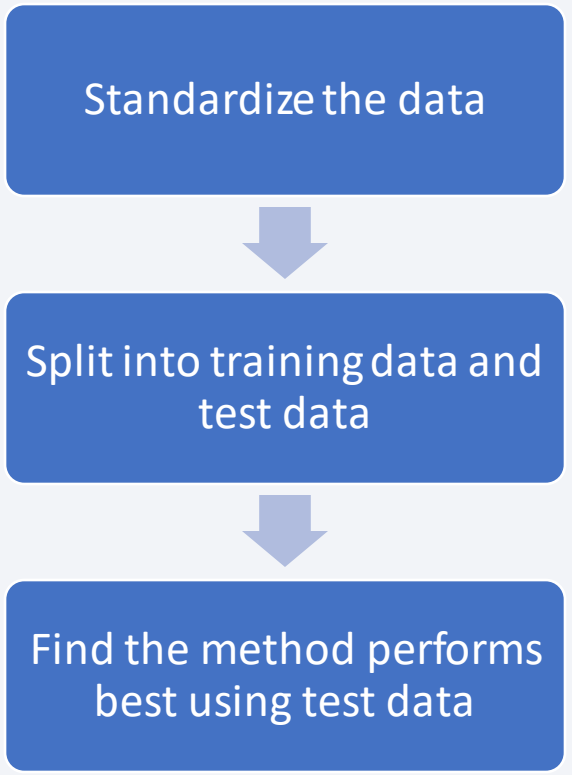
Build a Dashboard with Plotly Dash



Predictive Analysis (Classification)

- We compared 4 classification models (SVM, Classification Trees, Logistic Regression, and k nearest neighbors) in order to compare the results and find the method that performs best.
- Here is the GitHub URL of the completed notebook:

  https://github.com/Juank0621/Applied-Data-Science-Capstone/blob/80f97151abb13ed51bf0e3343cf1fff1b88b1081/8.%20Space%20X%20Falcon%209%20Machine%20Learning%20Prediction.ipynb

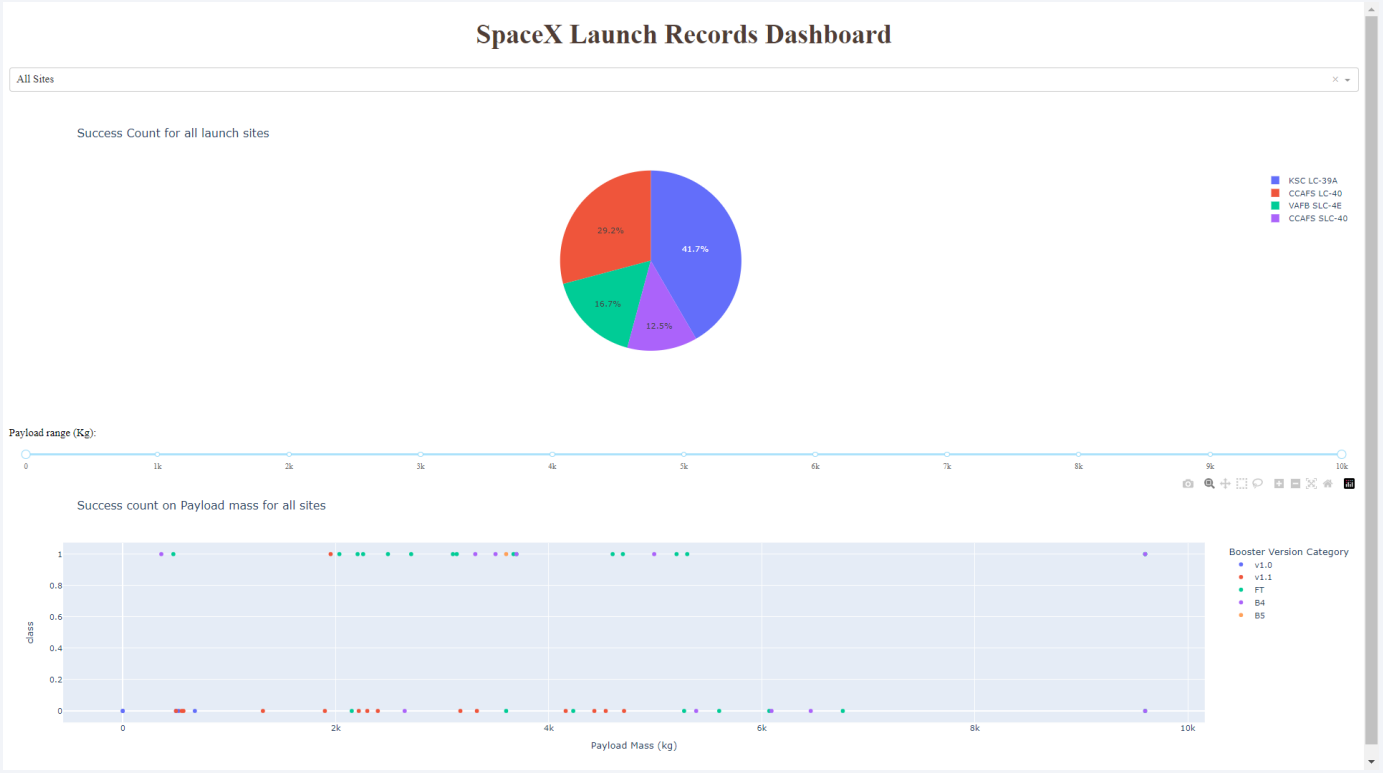


Results

- Exploratory data analysis results:
  - Space X uses four different launch sites (CCAFS LC-40, VAFB SLC-4, EKSC LC-39A, CCAFS SLC-40)
  - The number of landing outcomes was increasing as the years passed.
  - All launch sites are located in North America more exactly US. Also, we can observe that the launch sites are close to the coast.

Results

- Space X Launch Records Dashboard



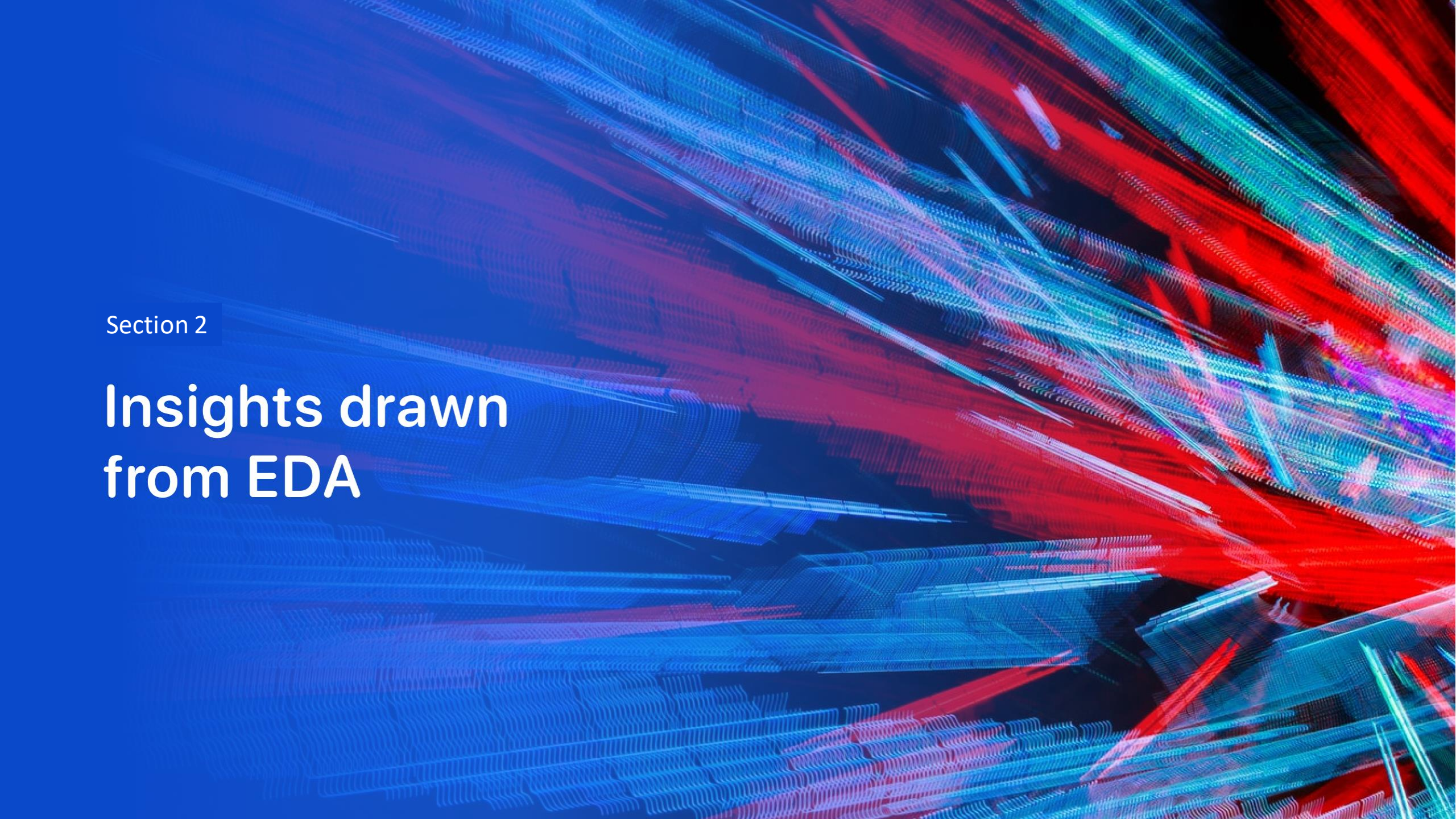
Results

- Predictive analysis results showed that all the models had an accuracy of over 83%.

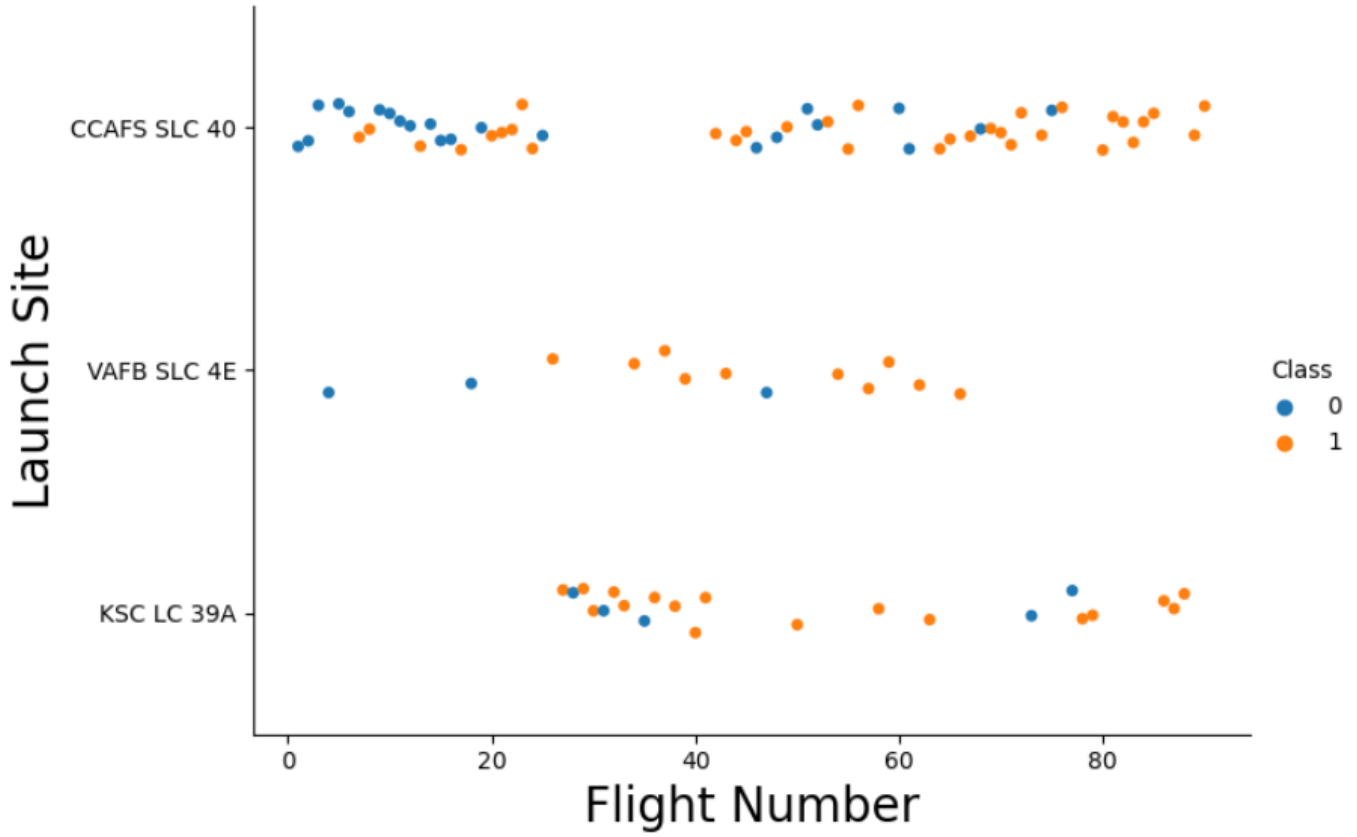
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

Section 2

Insights drawn from EDA



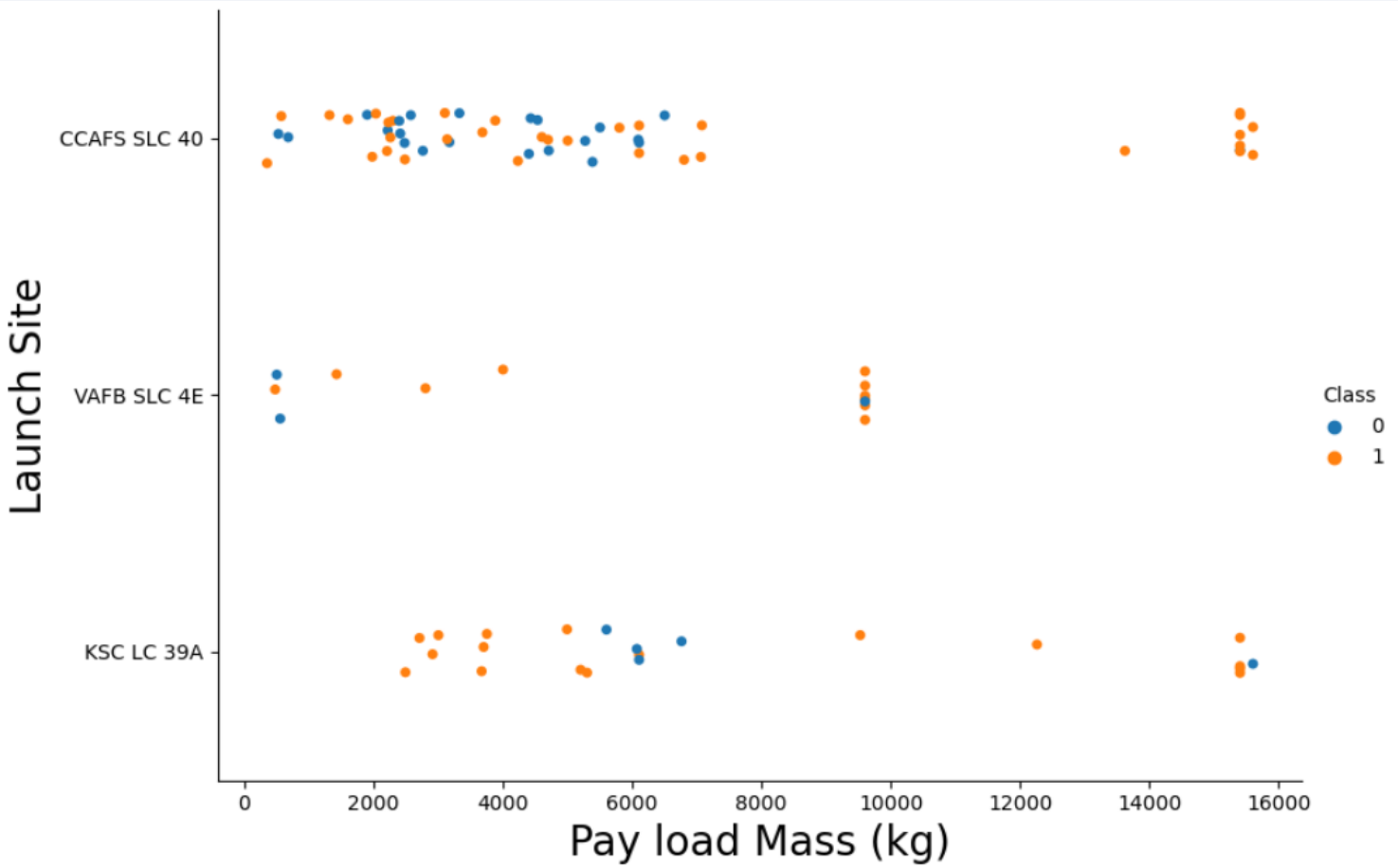
Flight Number vs. Launch Site



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

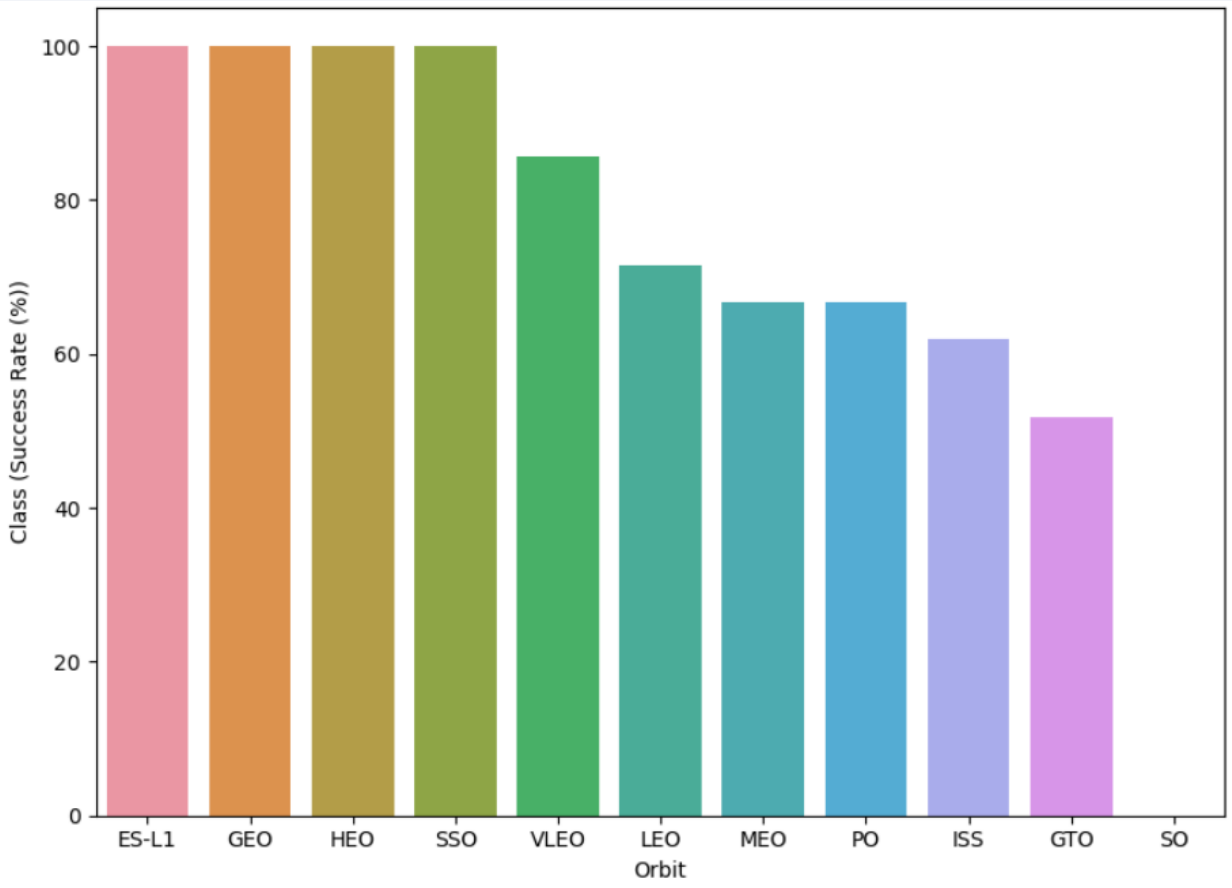
We can observe that, as the flight number increases in each of the 3 launch sites, so does the success rate. The success rate for the VAFB SLC 4E launch site is 100% after Flight number 50. Both KSC LC 39A and CCAFS SLC 40 have 100% success rates after the 80th flight.

Payload vs. Launch Site



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

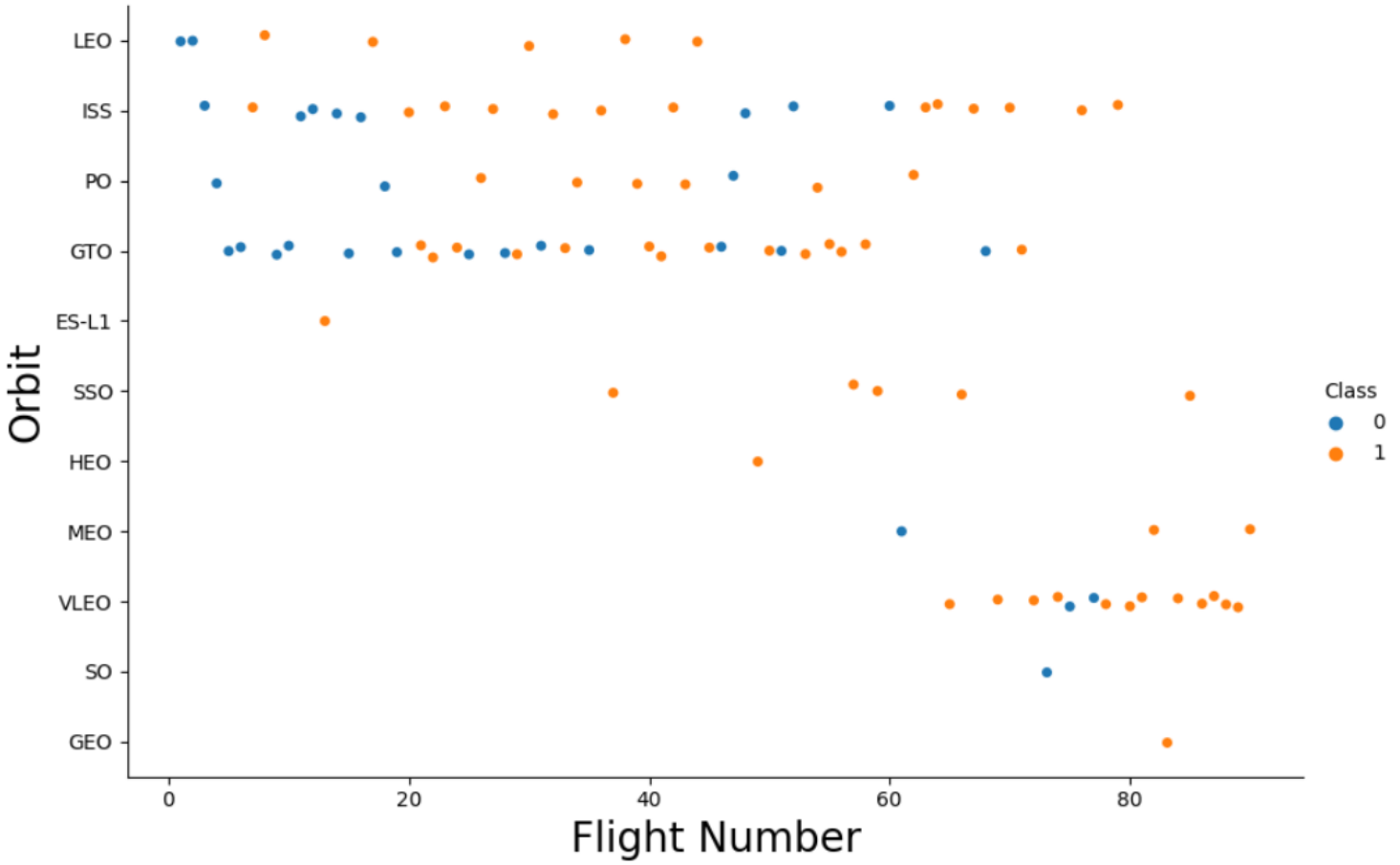
Success Rate vs. Orbit Type



Analyze the ploted bar chart try to find which orbits have high sucess rate.

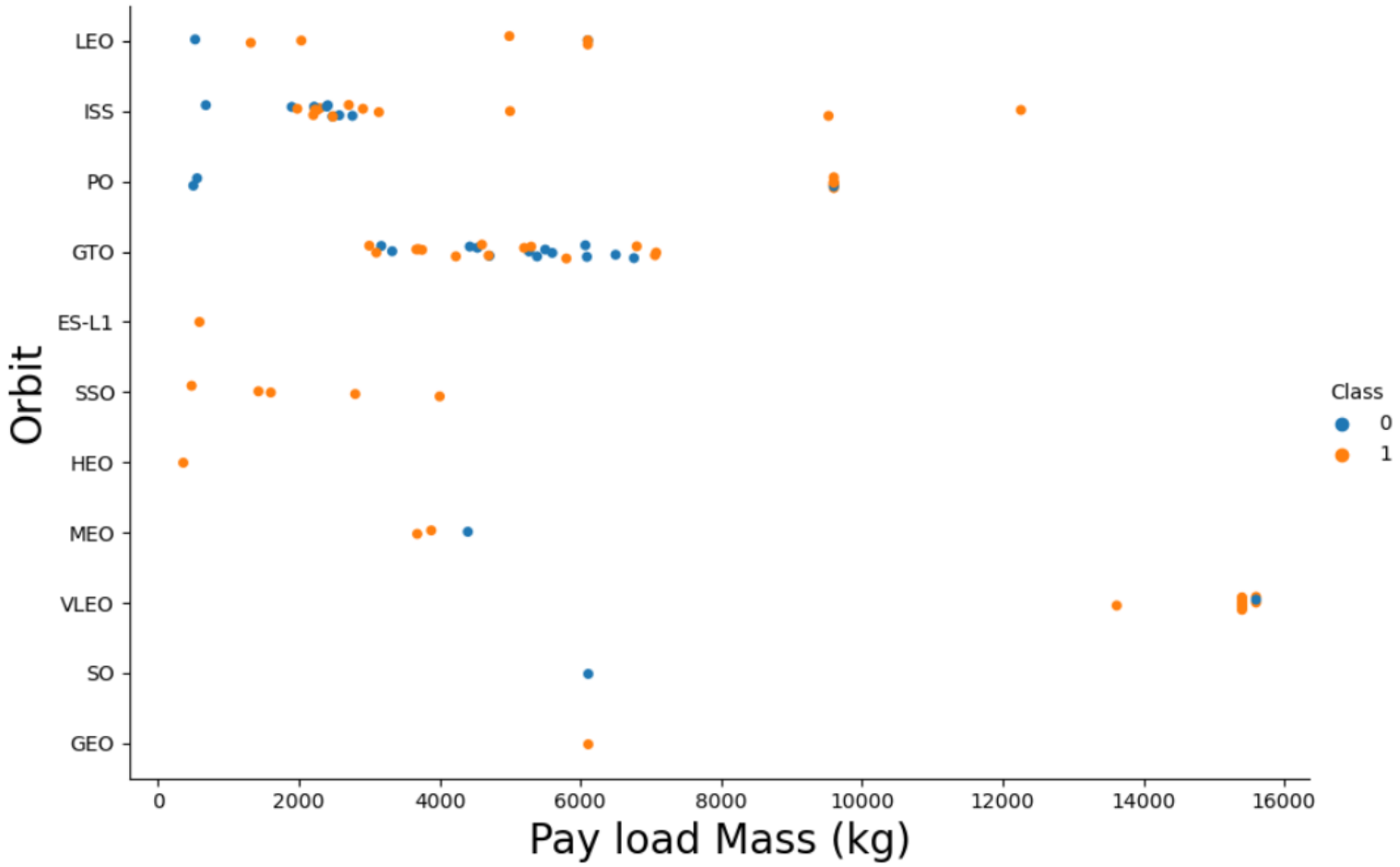
Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate

Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

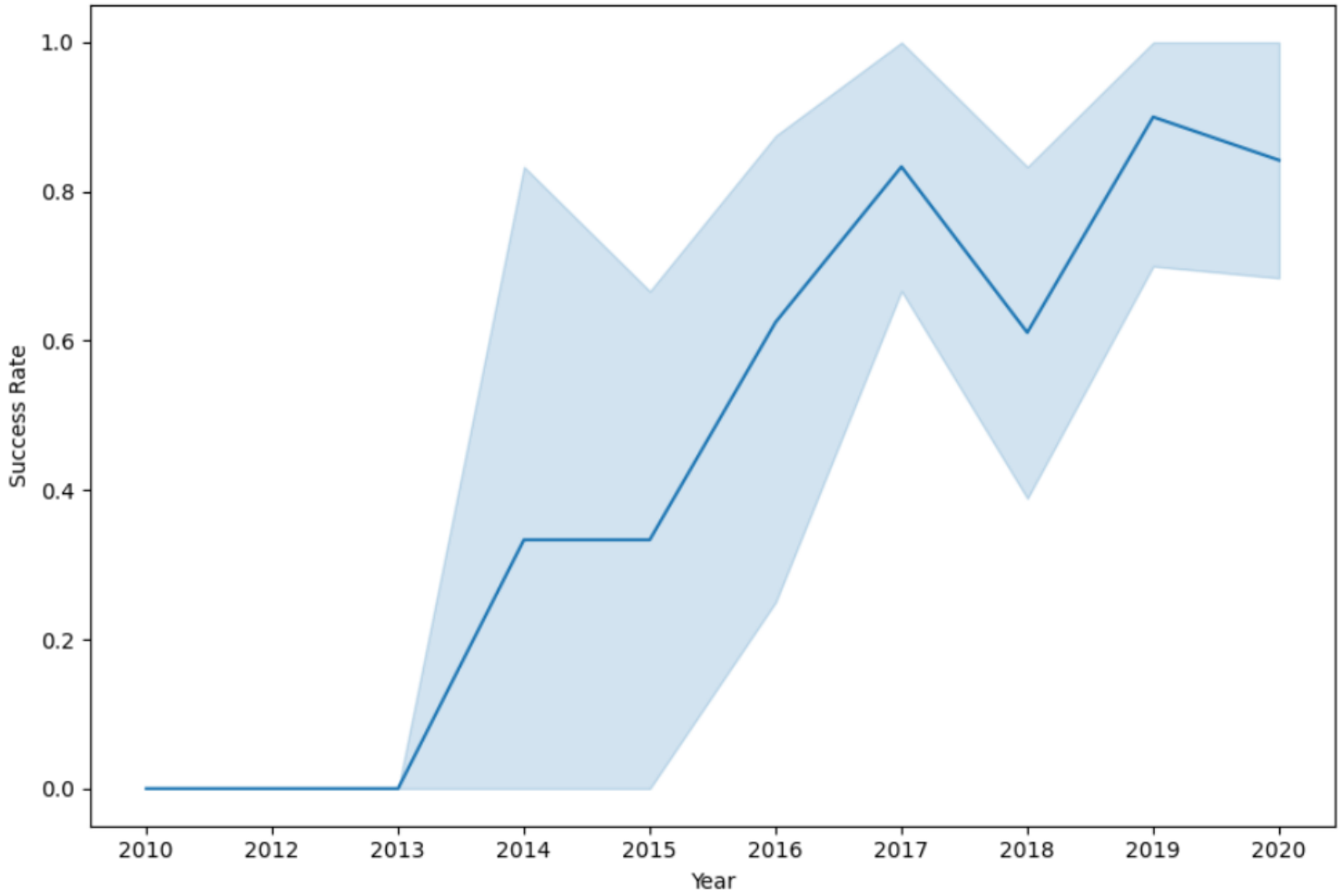
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Find the names of the unique launch sites

- We obtained four launch sites by selecting unique occurrences of LAUNCH_SITE.

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

In [7]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

Out[7]:

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

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with 'CCA'

- Use LIKE command with '%' wildcard in WHERE clause to select and display a table of all records where launch sites begin with the string 'CCA' and LIMIT 5 to only show the first 5 rows. We can observe 5 samples of CCA.

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Calculate and display the total payload carried by boosters from NASA.

- Use the SUM () function to return and display the total sum of 'PAYLOAD_MASS_KG' column for Customer 'NASA (CRS)'.
- The total payload was 45596 KG.

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

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

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

- Use the AVG () function to return and display the average payload mass carried by booster version F9 v1.1.
- The average payload mass for booster version F9 v1.1 was 2928.4 KG.

Task 4

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1';
```

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

Payload Mass Kgs	Customer	Booster_Version
2928.4	SES	F9 v1.1

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

- Use the MIN () function to return and display the first (oldest) date when the first successful landing outcome on the ground pad was achieved.
- The first successful landing outcome on the ground pad success on 01-05-2017.

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

Hint:Use min function

In [15]:
```
%sql SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE "Landing _Outcome" = "Success (ground pad)";
```

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

Out[15]:

MIN(DATE)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- Use the SELECT DISTINCT statement to return and list the unique names of boosters with operators > 4000 and < 6000 to only list boosters with payloads between 4000-6000 with the landing outcome of 'Success (drone ship)'.
- We can obtain 4 results of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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 [16]:
```
%sql SELECT DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOA
```

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

Out[16]:

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

- Use the COUNT () function together with the GROUP BY statement to return the total number of mission outcomes.

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

- Use a subquery to return and pass the Max Payload that has carried max payload.
- This is the list of the booster which have carried the maximum payload mass.

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

```
%sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

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

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

2015 Launch Records

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

- The list only has two occurrences.

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.

```
%sql SELECT substr(Date,7,4), substr(Date, 4, 2),"Booster_Version", "Launch_Site", Payload, "PAYLOAD_MASS__KG_", "Mission_Outcome", "Landing _Outcome"
```

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

substr(Date,7,4)	substr(Date, 4, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing _Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	Success	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	Success	Failure (drone ship)

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- We obtained 34 successful landing_outcomes in total between the date 04-06-2010 and 20-03-2017

```
%sql SELECT "Landing _Outcome", COUNT(*) AS QTY FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' AND (Date BETWEEN '04-06-2010' AND '20-03-2017
```

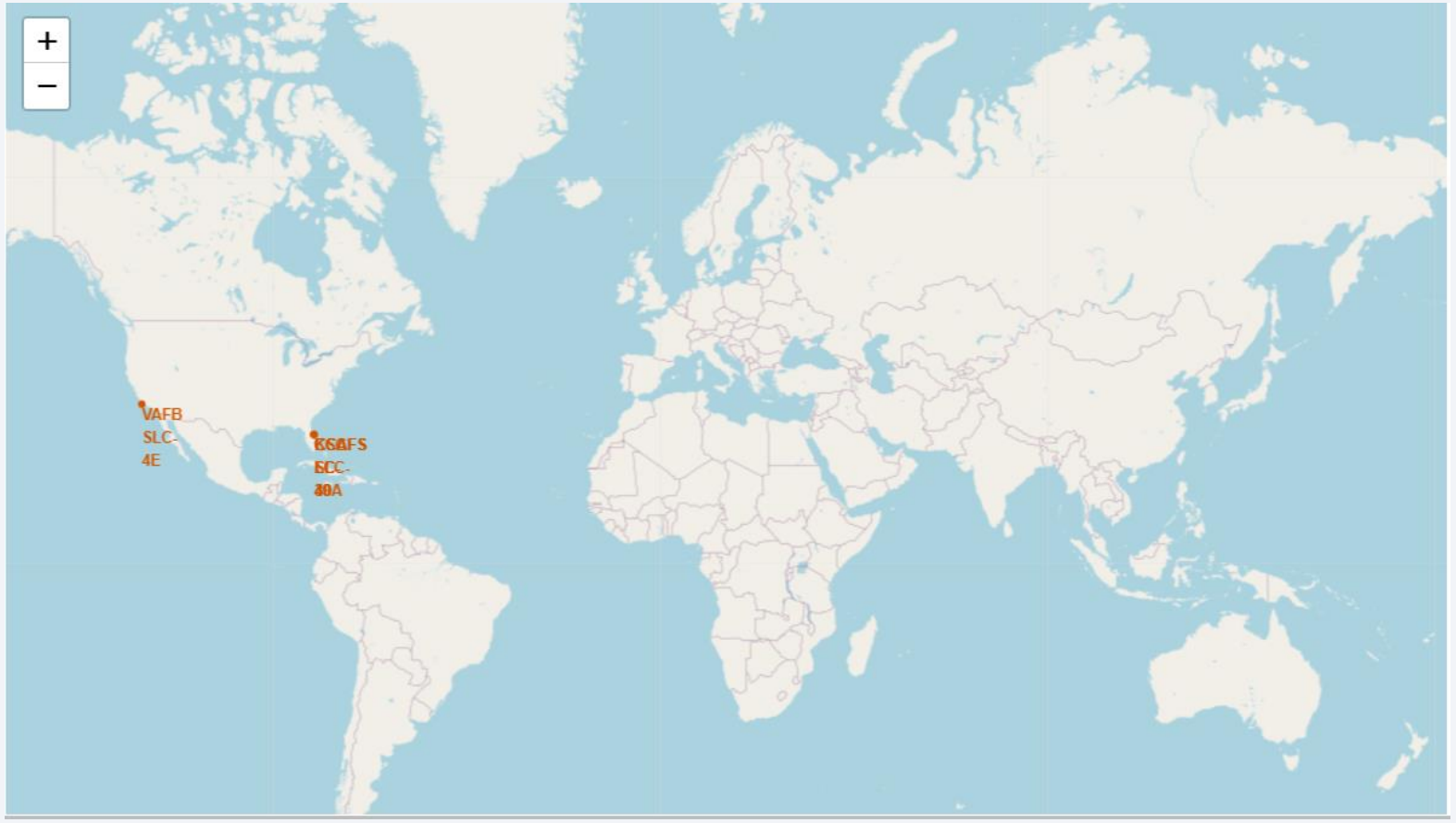
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	QTY
Success	20
Success (drone ship)	8
Success (ground pad)	6


Section 3
Launch Sites
Proximities Analysis

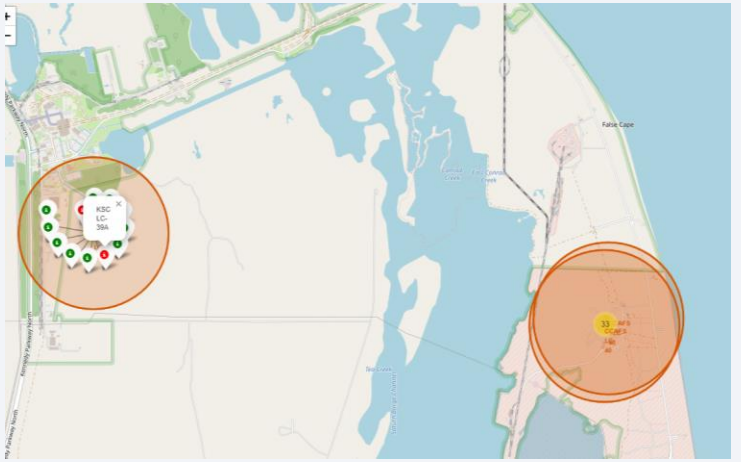
All launch sites locations on a global map

- All launch sites are located in North America more exactly US. Also, we can observe that the launch sites are close to the coast.



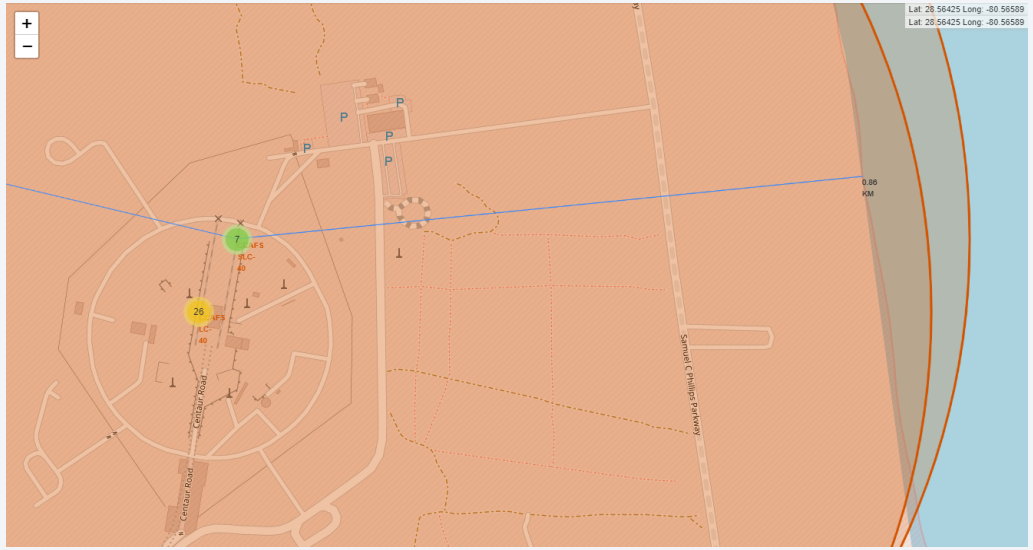
Launch Sites Outcomes

- The green markers indicate successful launches and the red ones indicate failure.
- Launch site KSC LC-39A has relatively high success rates than CCAFS SLC-40 and CCAFS LC-40.

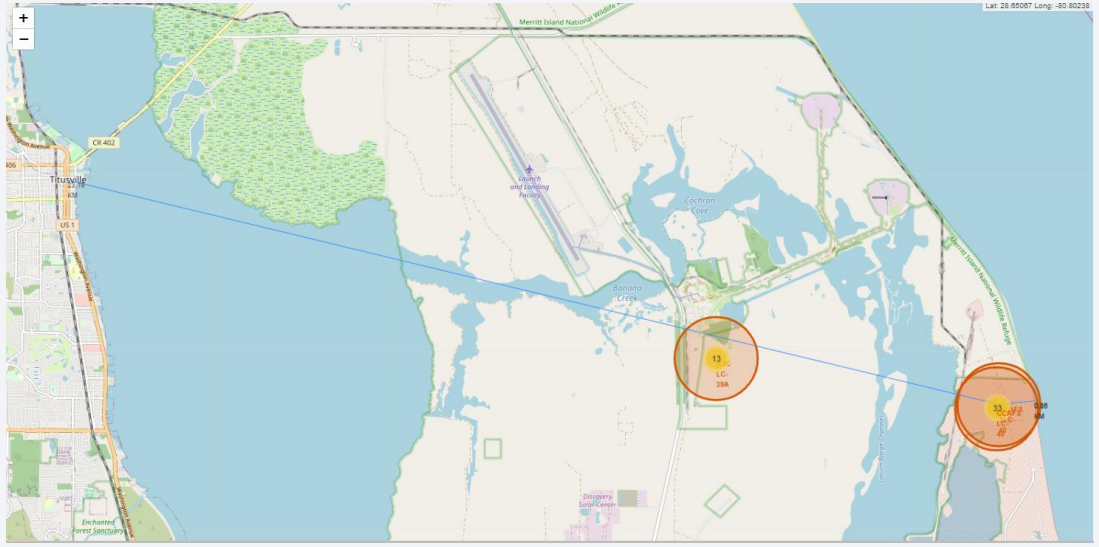


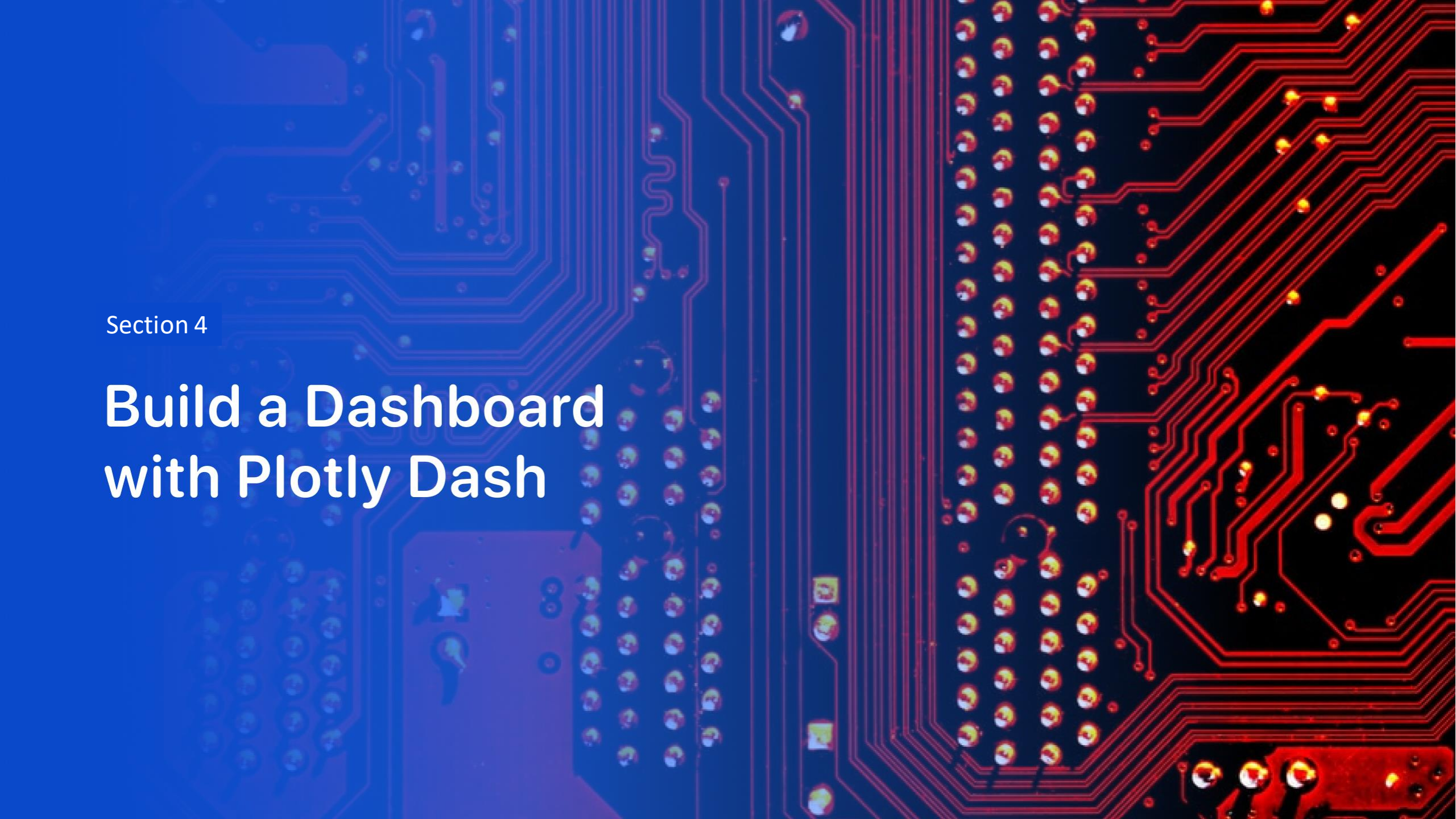
Launch Sites Proximities

- Launch site CCAFS SLC-40 has a proximity to the coastline of 0.86 km.



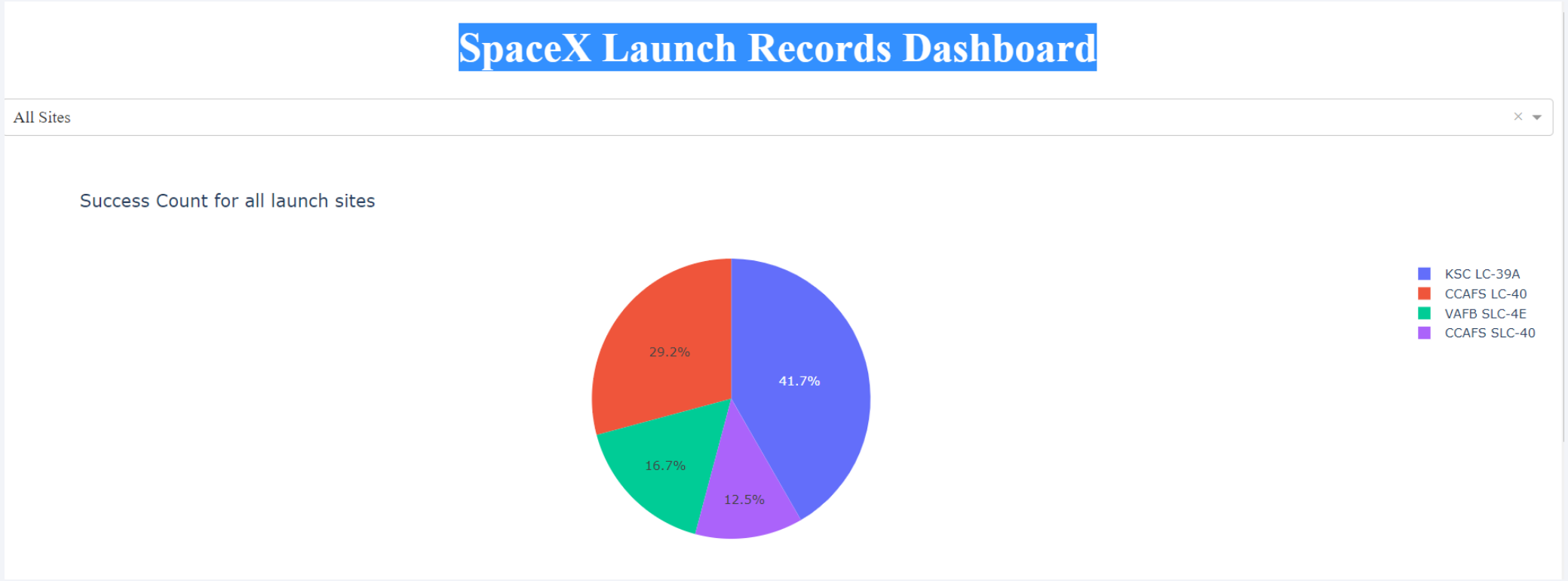
- Launch site CCAFS SLC-40 has a proximity to highway (Washinton Avenue) of 23.19 km.




Section 4
Build a Dashboard with Plotly Dash

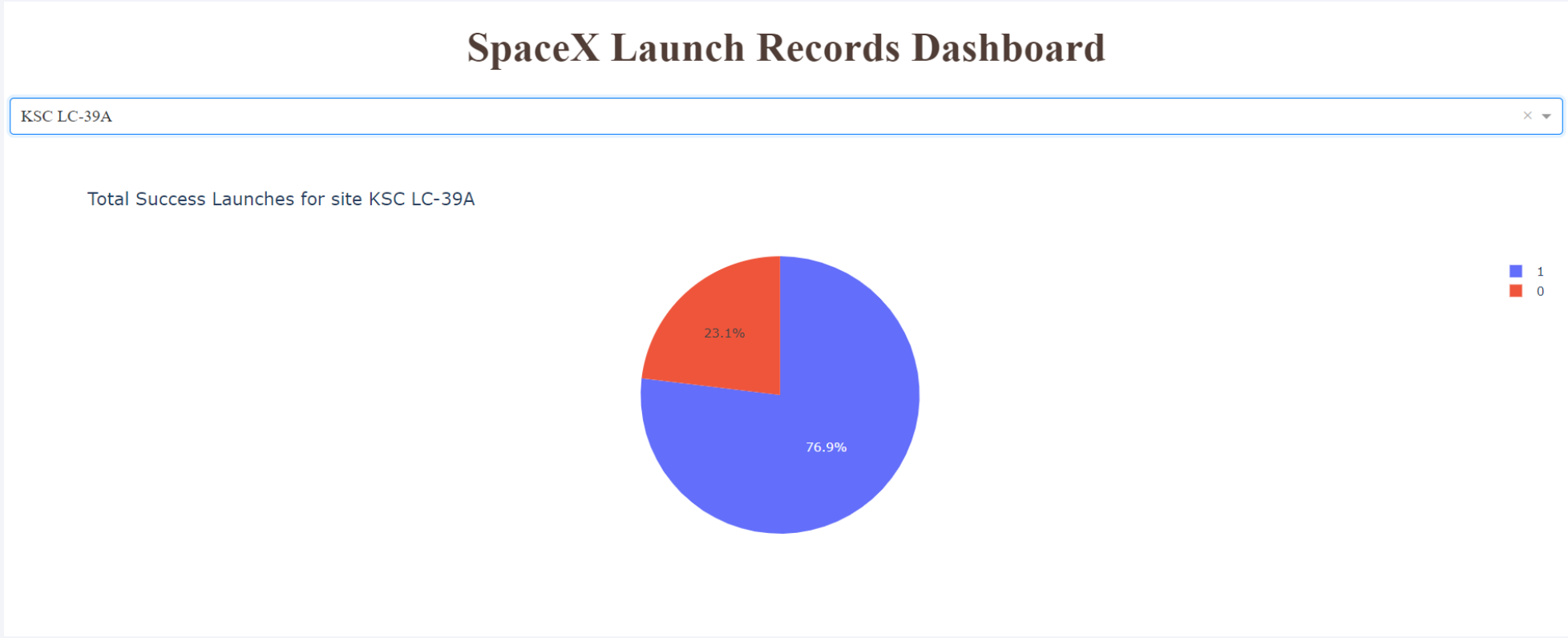
Launch success for all sites

- Launch site KSC LC-39A had the highest launch success rate at 41.7% followed by CCAFS LC-40 at 29.2%, VAFB SLC-4E at 14.7%, and finally CCAFS SLC-40 at 12.5%.



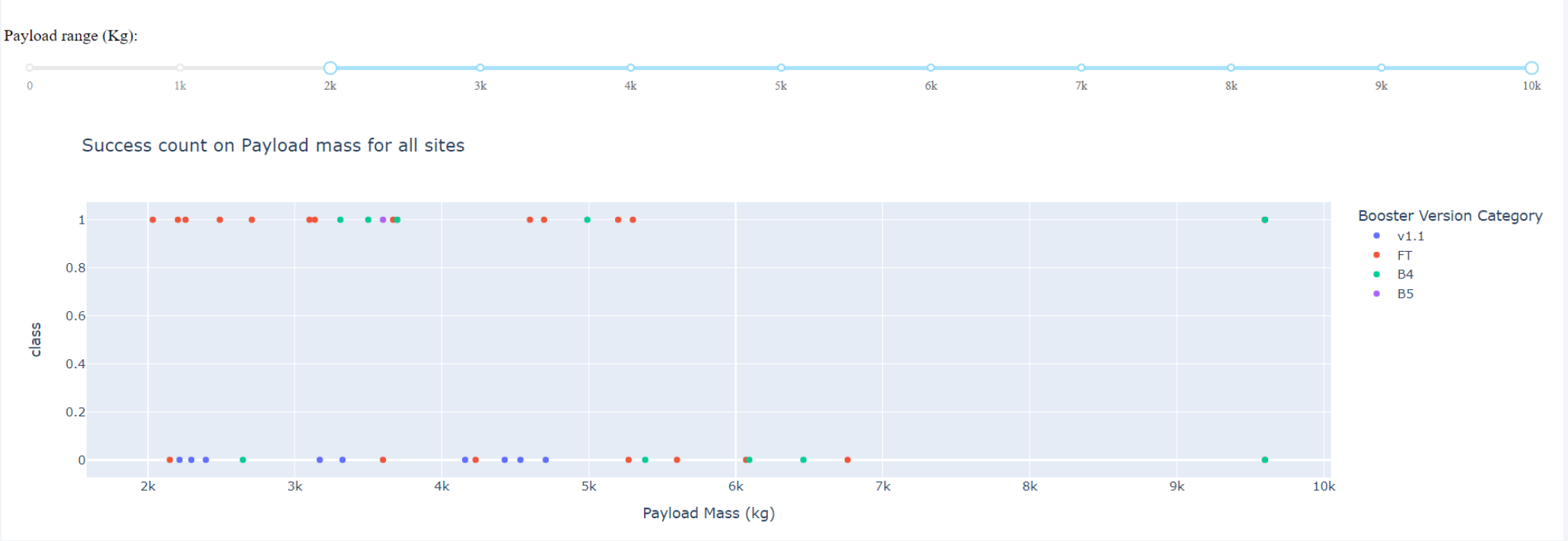
Launch site with the highest launch success ratio

- Launch site KSC LC-39A had the highest success ratio of 76.9% against 23.1% of failed launches.



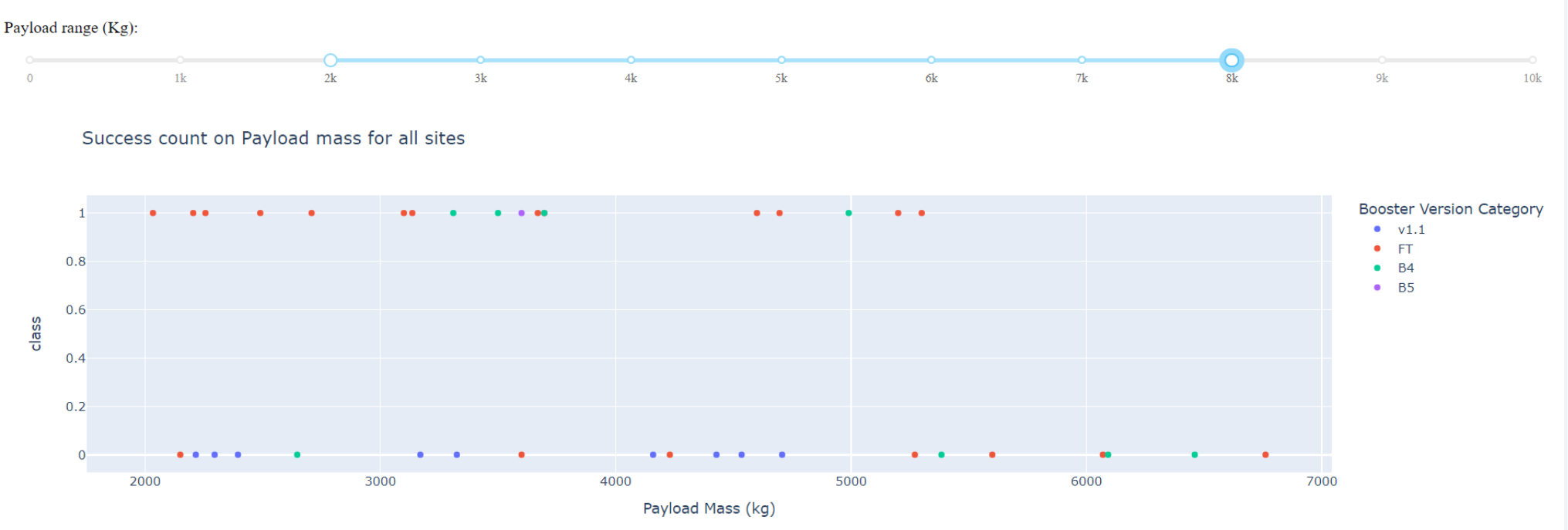
Payload vs. Launch Outcome for all sites

- We can see the payload between 2K to 10k and observe that the booster version v1.1 has the smallest successful rate.



Payload vs. Launch Outcome for all sites

- We can see the payload between 2K to 8k and observe that the booster version FT has the largest success rate.



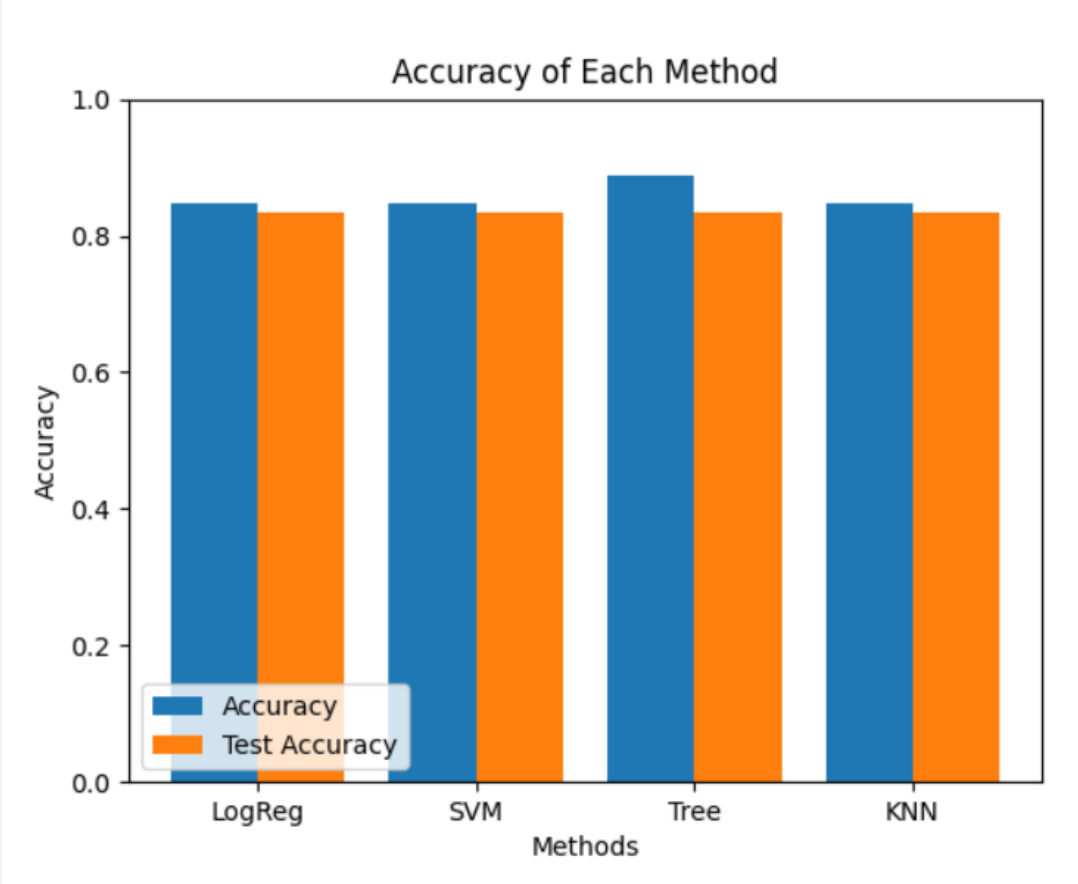
Section 5

Predictive Analysis (Classification)

Classification Accuracy

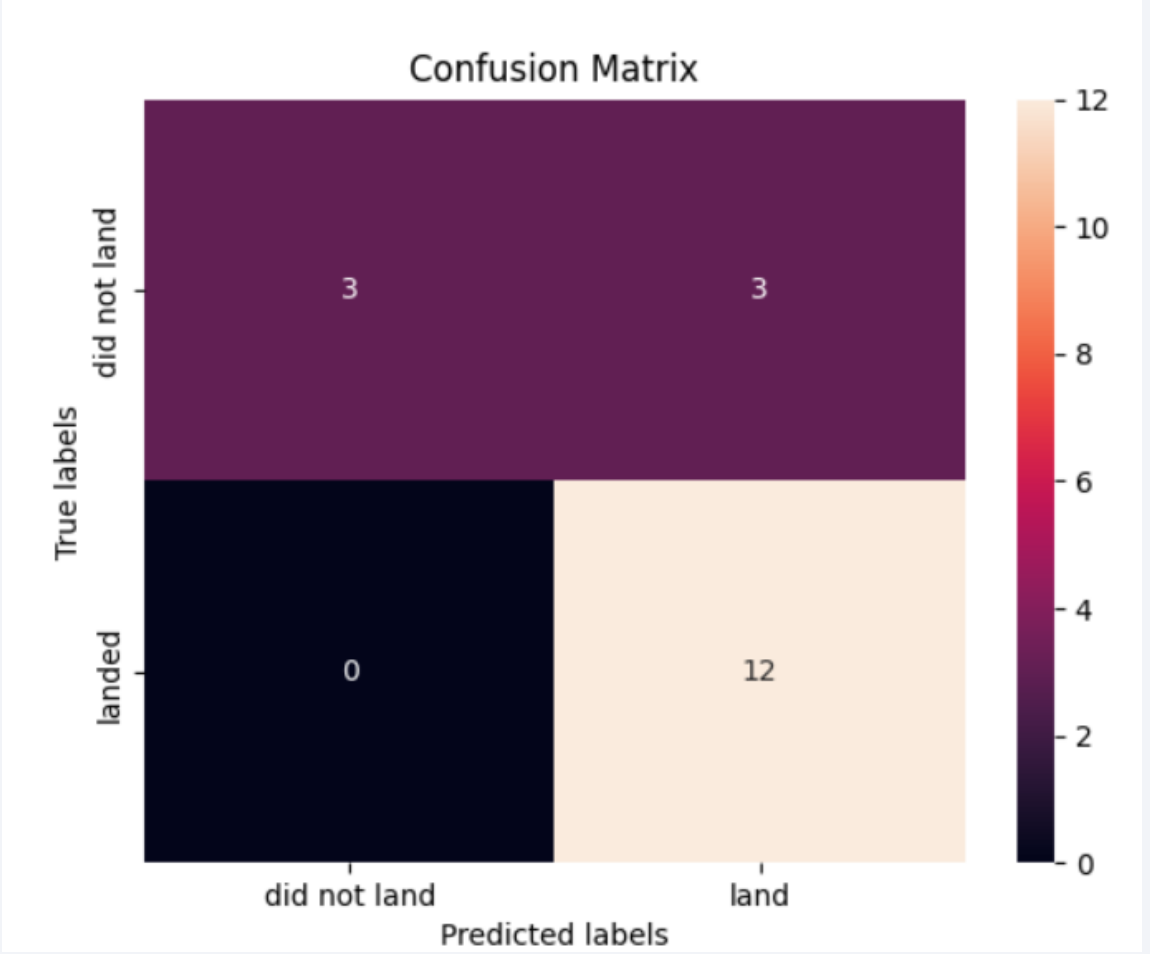
- We compared 4 classification models (SVM, Classification Trees, Logistic Regression, and k nearest neighbors)

- The model with the highest classification accuracy was the Decision Tree Classifier with an accuracy of 88%.



Confusion Matrix

- All 4 classification models obtained the same confusion matrixes and were able equally to distinguish between the different classes.



Conclusions

- All launch sites are located in North America more exactly US. Also, we can observe that the launch sites are close to the coast.
- Launch site KSC LC-39A had the highest launch success rate at 41.7% followed by CCAFS LC-40 at 29.2%, VAFB SLC-4E at 14.7%, and finally CCAFS SLC-40 at 12.5%.
- Launch site KSC LC-39A had the highest success ratio of 76.9% against 23.1% of failed launches.
- The model with the highest classification accuracy was the Decision Tree Classifier with an accuracy of 88%.
- All 4 classification models obtained the same confusion matrixes and were able equally to distinguish between the different classes.
- The number of landing outcomes was increasing as the years passed.

Appendix

- We can't observe maps from Folium on Github.

Thank you!