IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dashuai Chen
03-17-2023

Outline

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

Executive Summary

- Summary of methodologies

➢ Data preparation
  - ✓ Data collection from API and web page
  - ✓ Data wrangling, transforming and mapping

➢ Data exploration
  - ✓ Exploratory data analysis (EDA) using SQL
  - ✓ Data visualization using python
  - ✓ Interactive visualization analysis with dashboards

➢ Predictive analysis
  - ✓ Machine learning model training
  - ✓ Prediction using trained Model
  - ✓ Results analysis

- Summary of all results

Introduction

Background: SpaceX-commercial space age

- Sending spacecraft to the International Space Station
- Starlink, a satellite internet constellation providing satellite Internet access
- Sending manned missions to Space.

Objective:

- To determine the price of each launch.
- To determine if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

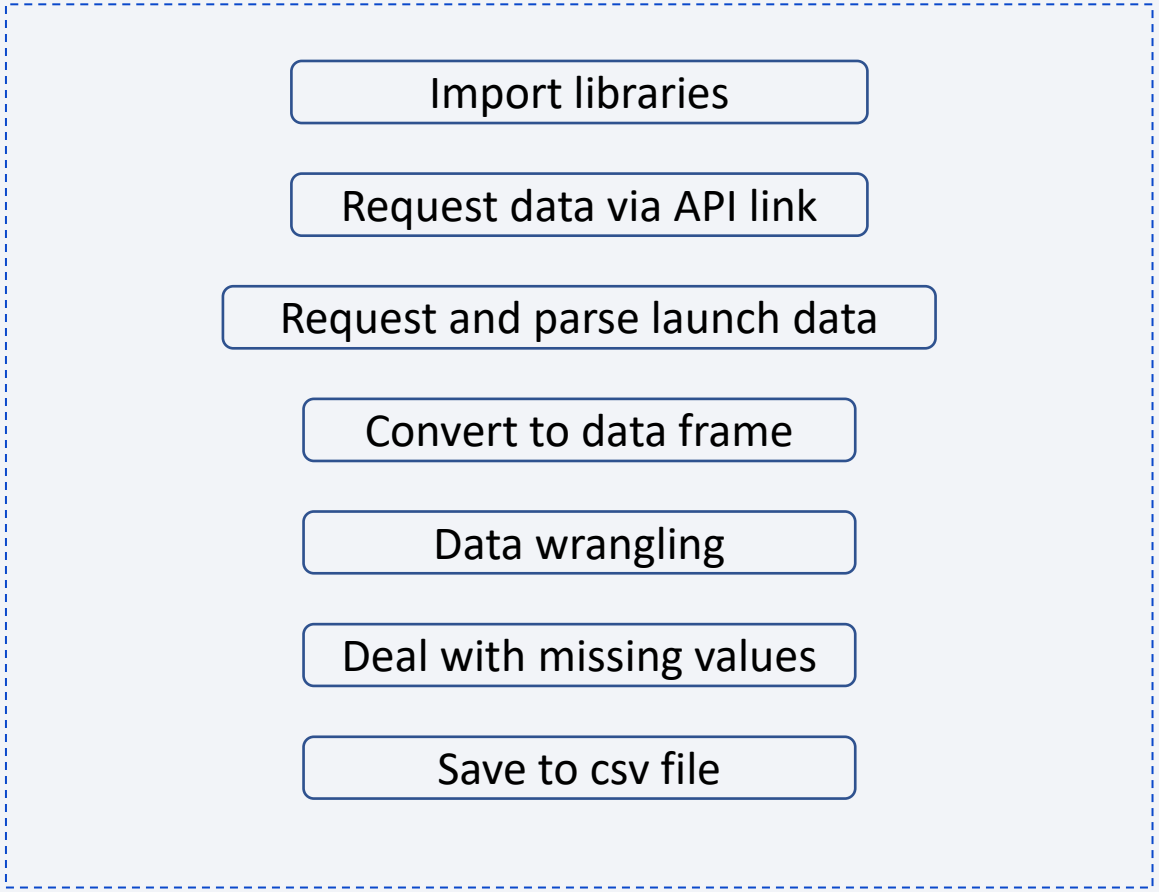
- Data collection methodology:
  - Require data from SpaceX REST API
  - Collect data from Wikipedia page
- Perform data wrangling
  - Use python to find relationship and pattern of labels
  - Find the labels to be used for machine learning training
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, support vector machine, decision tree, and KNN are built via Sklearn
  - GridSearchCV are used to choose the fitting models with best hyperparameters

Data Collection

- SpaceX REST API:
  - ➢ url = https://api.spacexdata.com/v45/launches/past
    - ✓ Require data from the SpaceX API
    - ✓ Clean the data
- Web scraping:
  - ➢ url = https://en.Wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - ✓ Extract the Faclon 9 launch records HTML table from Wikipedia
    - ✓ Parse the table, convert it into Pandas dataFrame

Data Collection – SpaceX API

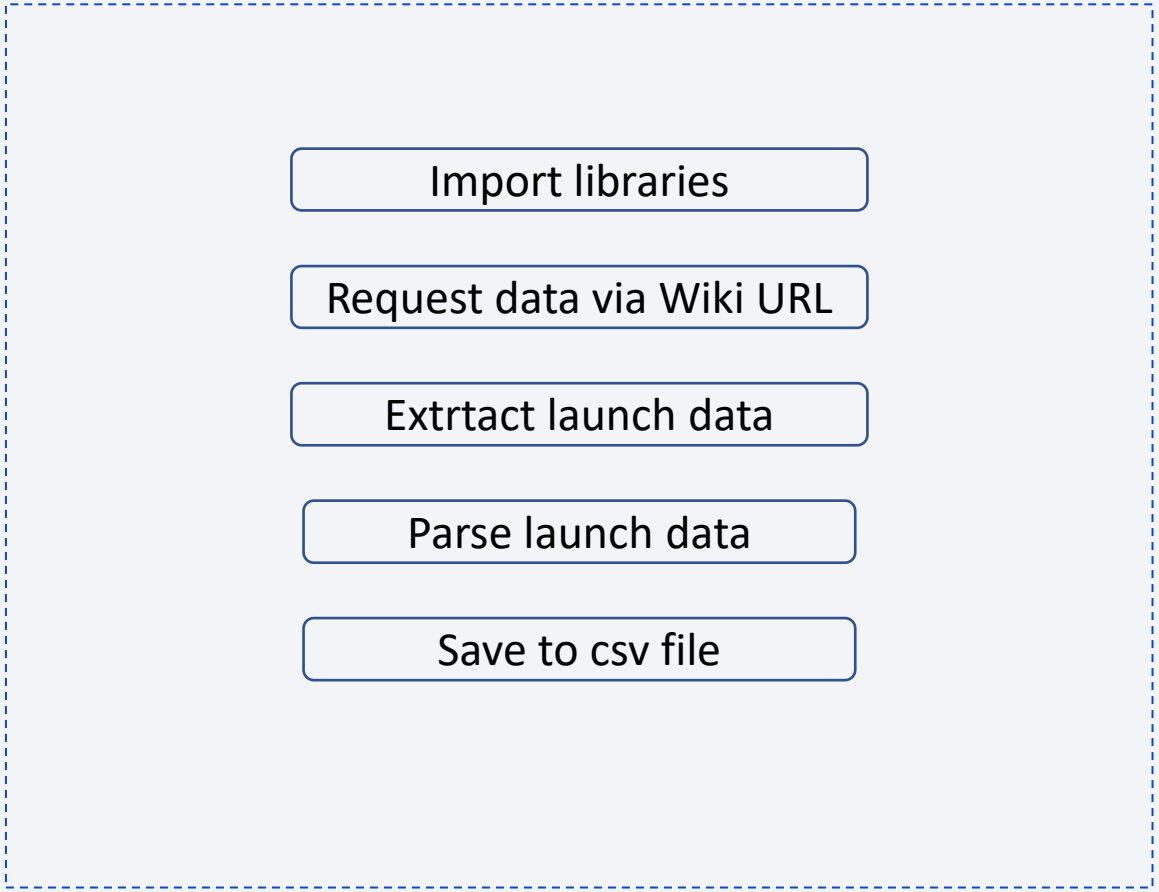
- Request data from SpaceX API
- Pandas is used to convert the json data into dataFrame
- Find and save the Falcon 9 launches data
- Export the dataframe to .csv file
- GitHub URL:
- https://github.com/general-chen/Coursera/blob/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone/week1/Lab1_collecting_the_data.ipynb



Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- Export to .csv file
- Link:

https://github.com/general-chen/Coursera/blob/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone/week1/jupyter-labs-webscraping.ipynb



Data Wrangling

- Import Libraries and Define Auxiliary Functions
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Link:
- https://github.com/general-chen/Coursera/blob/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone/week1/lab2_data_wrangling.ipynb

EDA with Data Visualization

- A scatter point chart: visualize the relationship between Flight Number and Launch Site
- A scatter point chart: visualize the relationship between Payload and Launch Site
- A bar chart: visualize the relationship between success rate of each orbit type
- A scatter point chart: visualize the relationship between FlightNumber and Orbit type
- A scatter point chart: visualize the relationship between Payload and Orbit type
- A line chart: visualize the launch success yearly trend

https://github.com/general-chen/Coursera/blob/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone/week2/Lab4-jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

https://github.com/general-chen/Coursera/blob/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone/week2/Lab3-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities
  - To see whether it is close to a coast
  - To see whether it is close to a railway
  - To see whether it is close to a highway
  - To see whether it is close to a city
- To find some geographical patterns about launch sites.

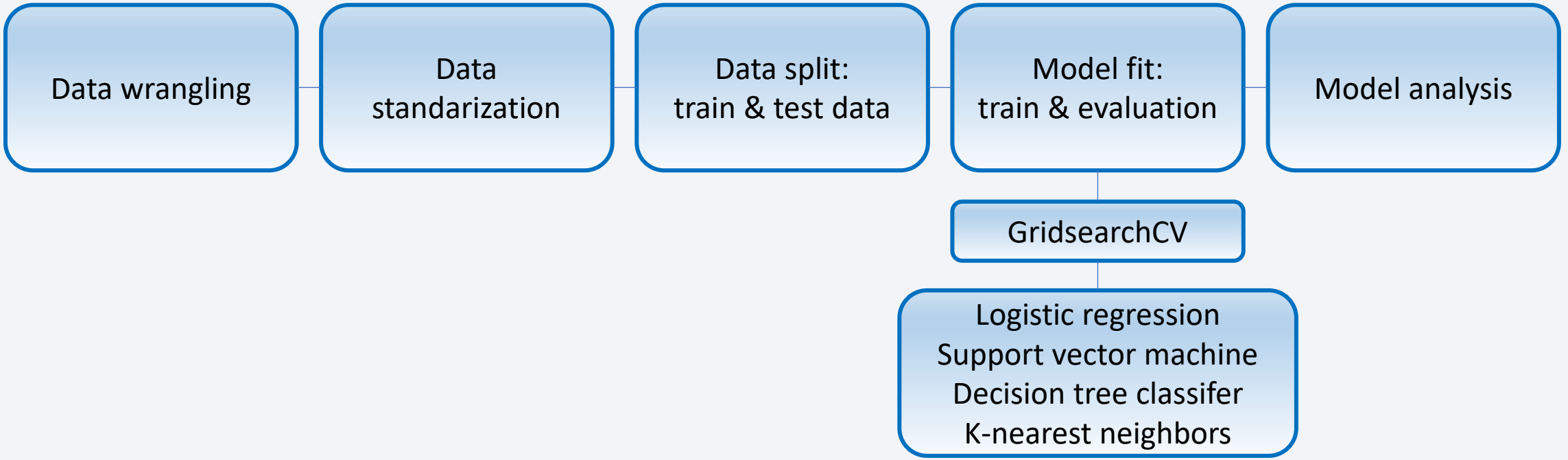
https://github.com/general-chen/Coursera/blob/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone/week3/Lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- a Launch Site Drop-down Input Component
- a callback function to render success-pie-chart based on selected site dropdown
- a Range Slider to Select Payload
- a callback function to render the success-payload-scatter-chart scatter plot
- To find a pattern of success rate between launch site and payload
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

https://github.com/general-chen/Coursera/blob/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone/week3/spacex_dash_app.py

Predictive Analysis (Classification)



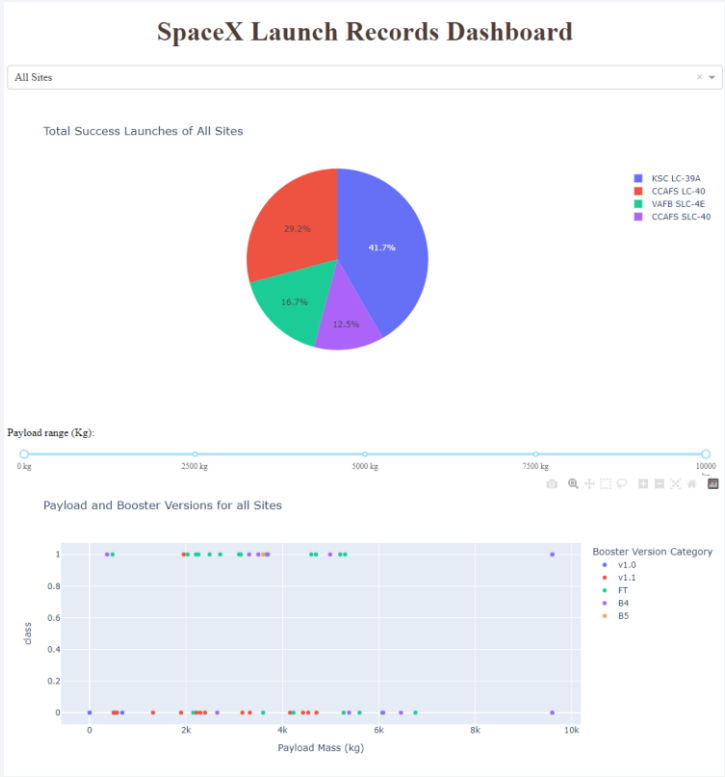
https://github.com/general-chen/Coursera/blob/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone/week4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

EDA

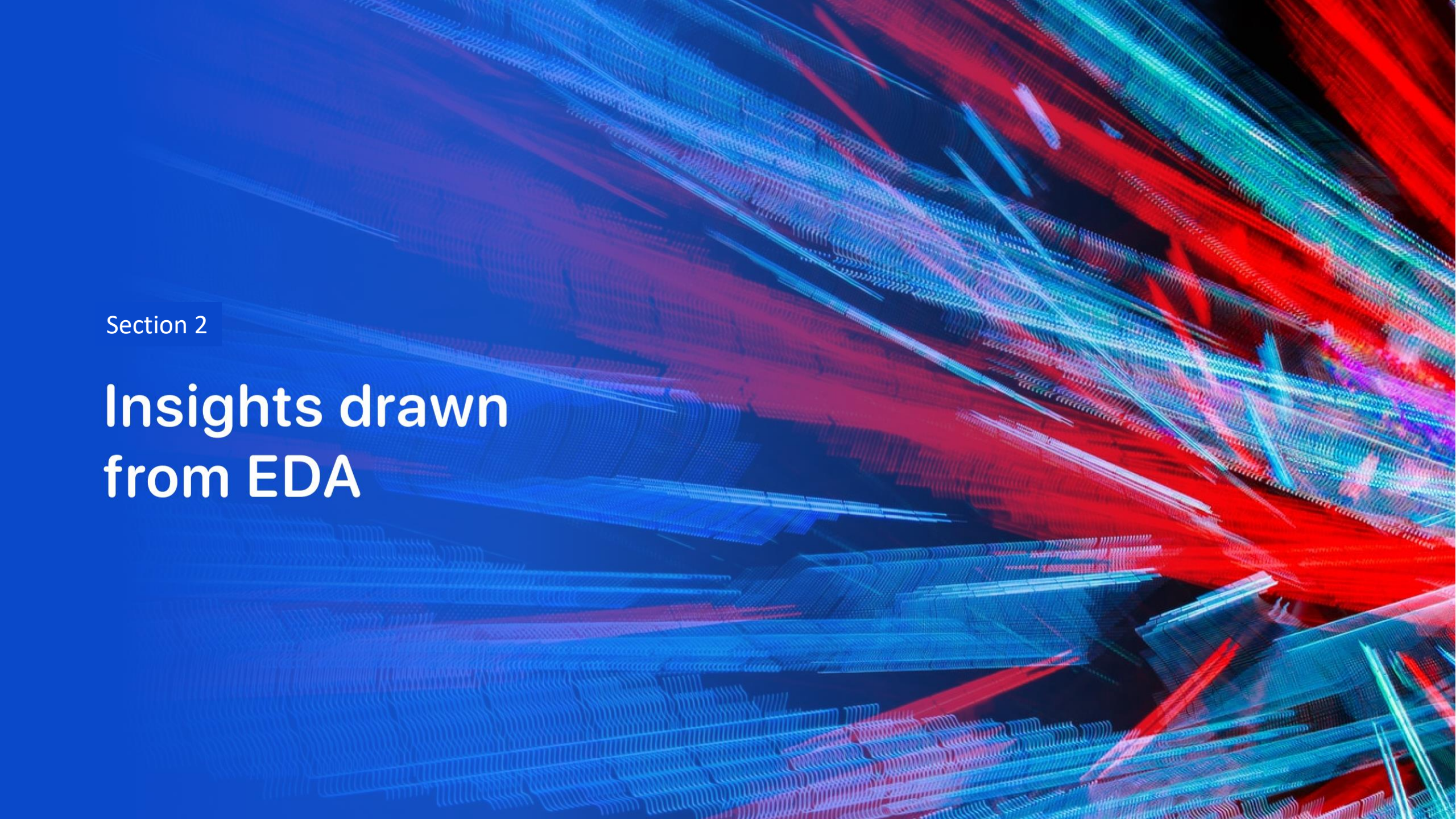
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- the sucess rate since 2013 kept increasing till 2020

Interactive analysis

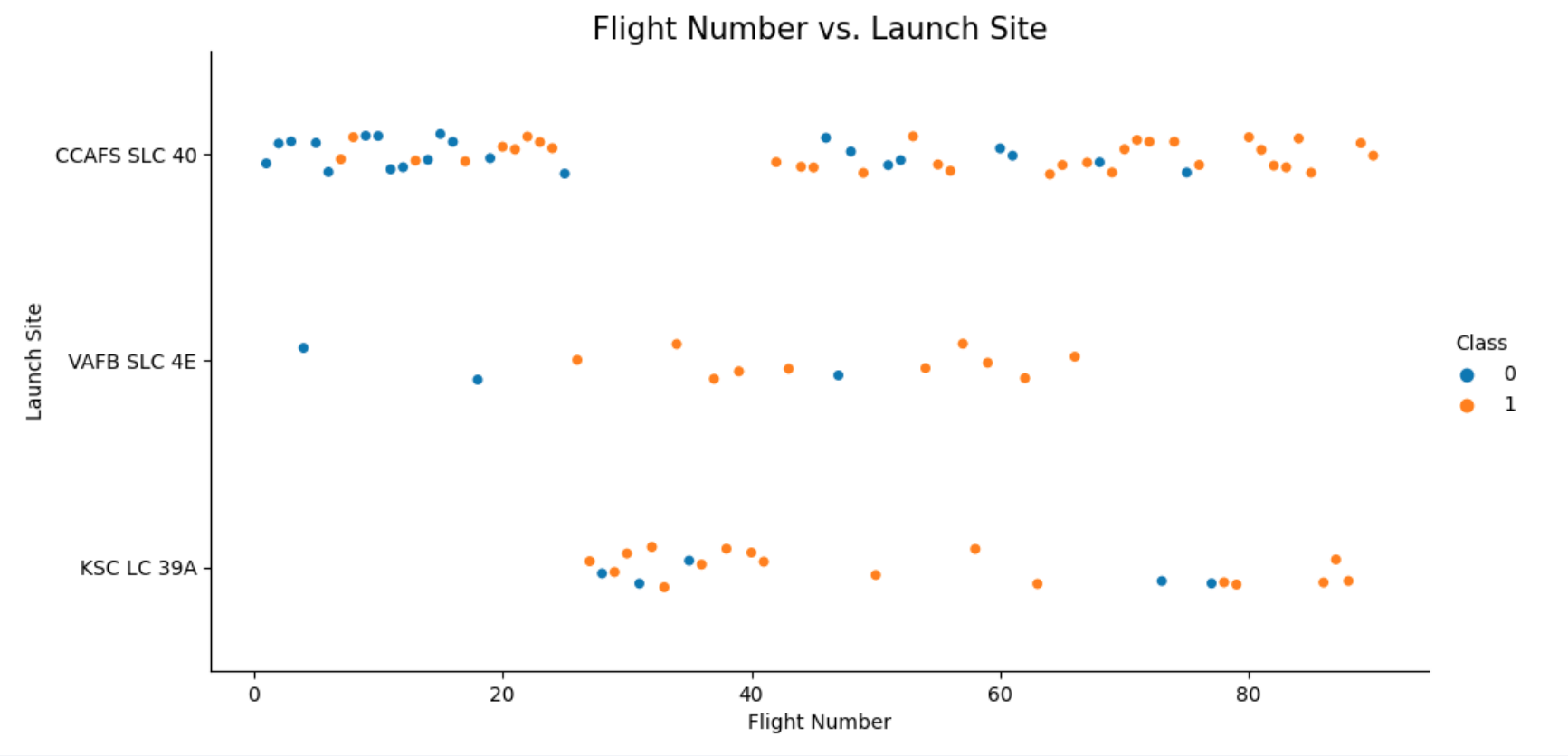


Predictive analysis

Based on the decision tree classifier model, there will be 5 true positive, 1 false positive, 1 false positive, and 11 true negative. The accuracy of this model is about 88.9%.

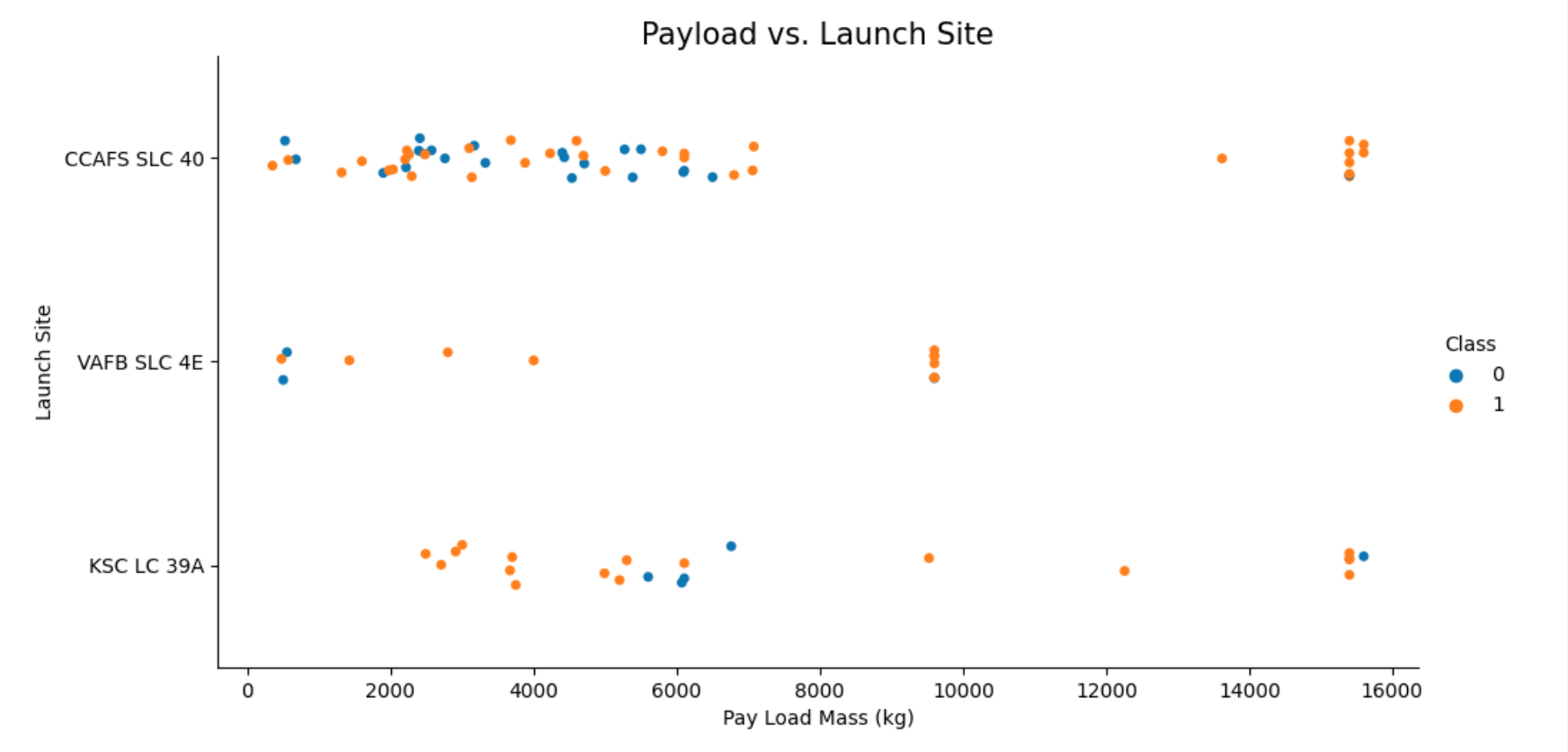

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



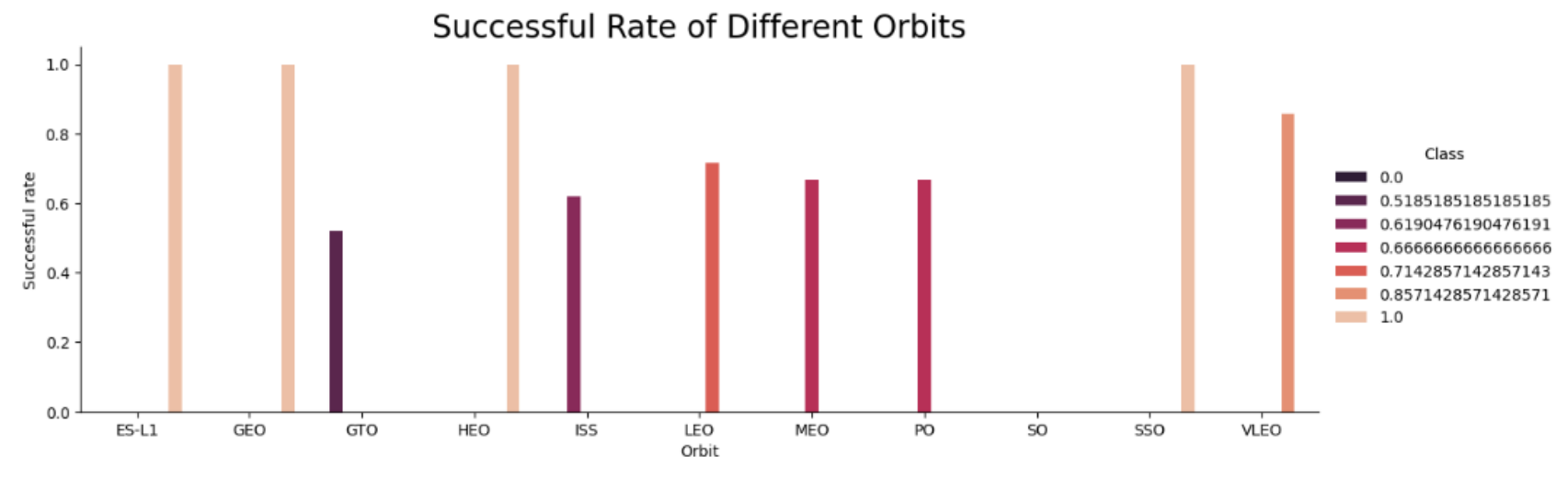
CCAFS SLC has the lowest successful rate, while KSC LC 39A has the highest successful rate.

Payload vs. Launch Site



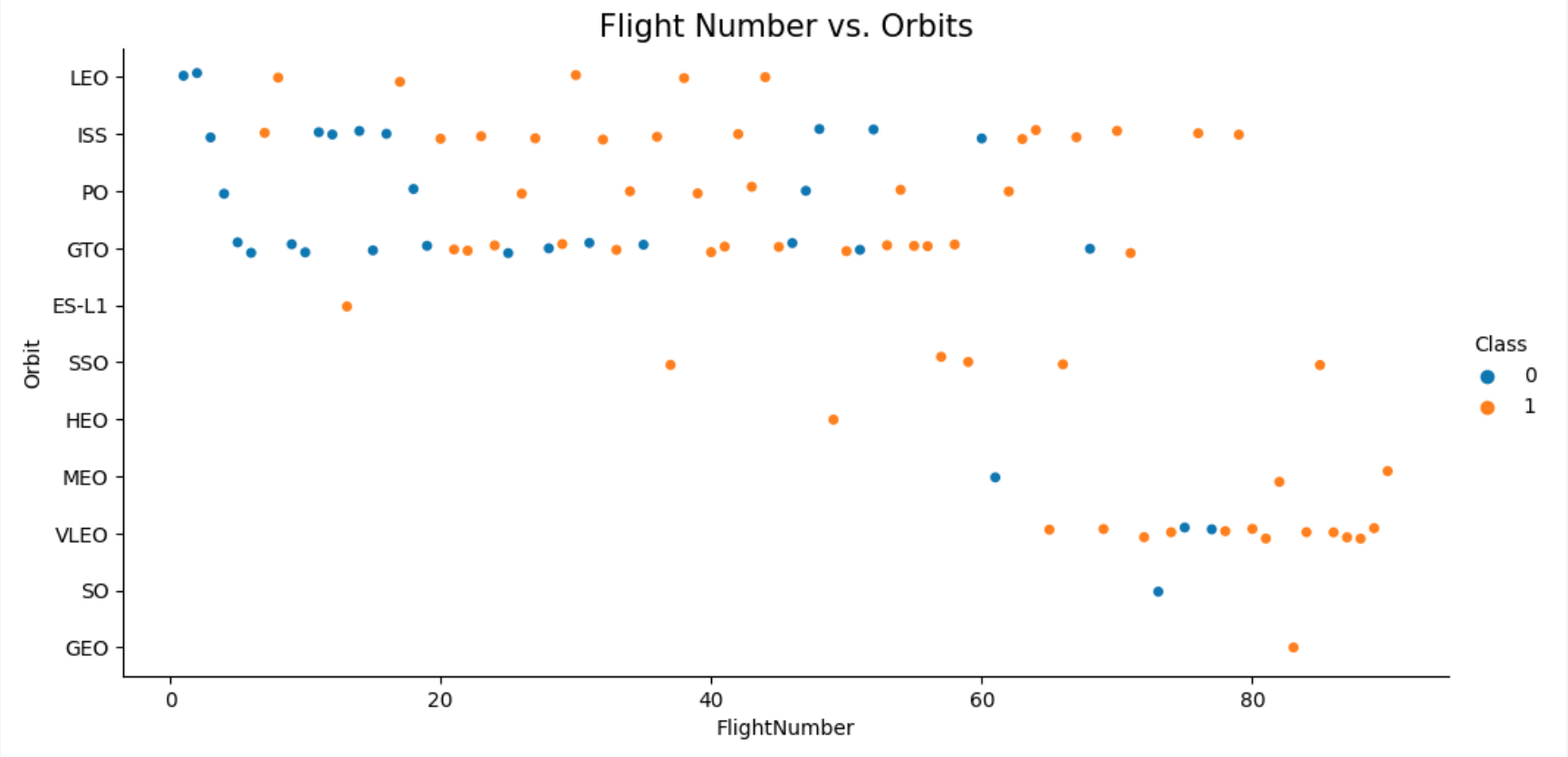
For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



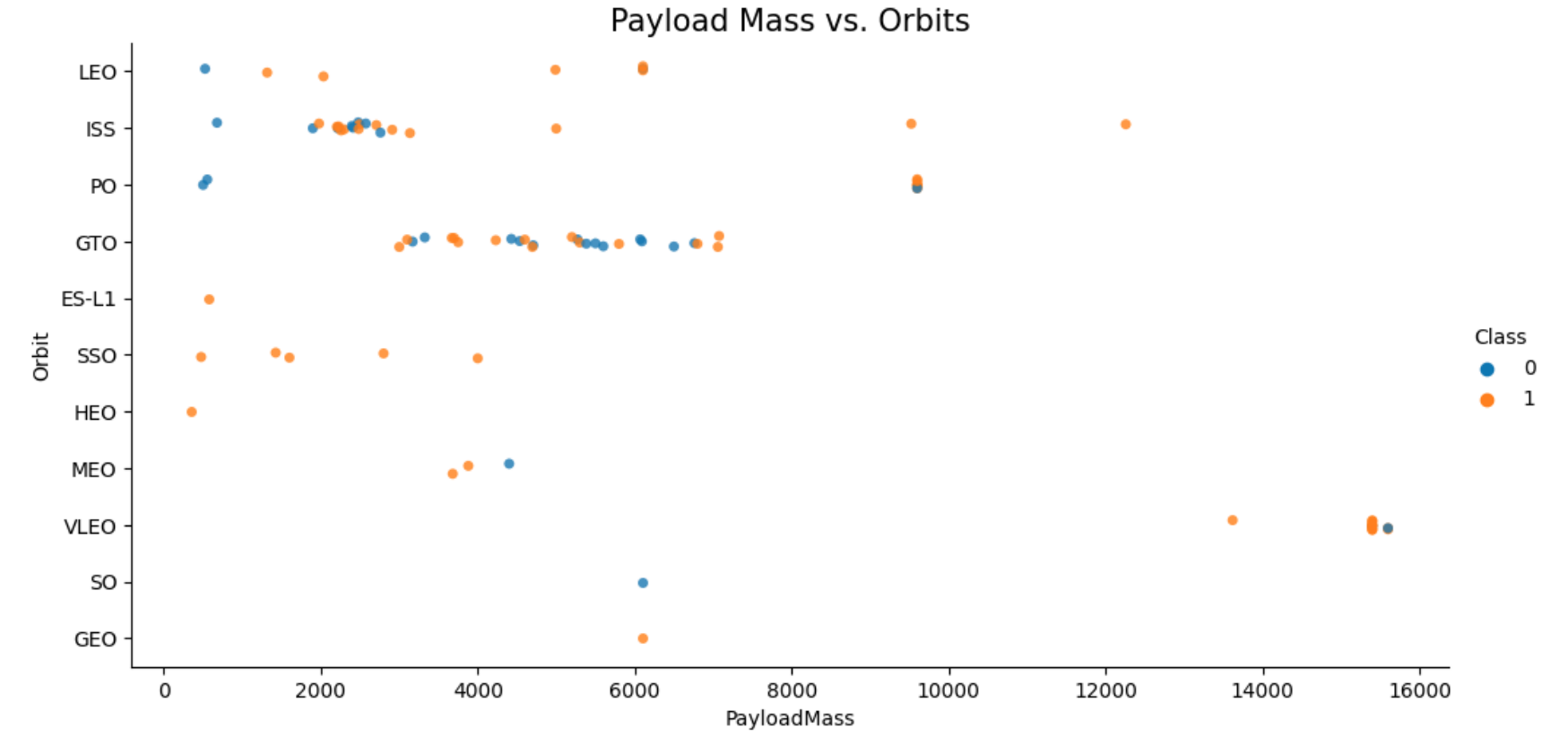
ES-L1, GEO, HEO, SSO have highest successful rate.

Flight Number vs. Orbit Type



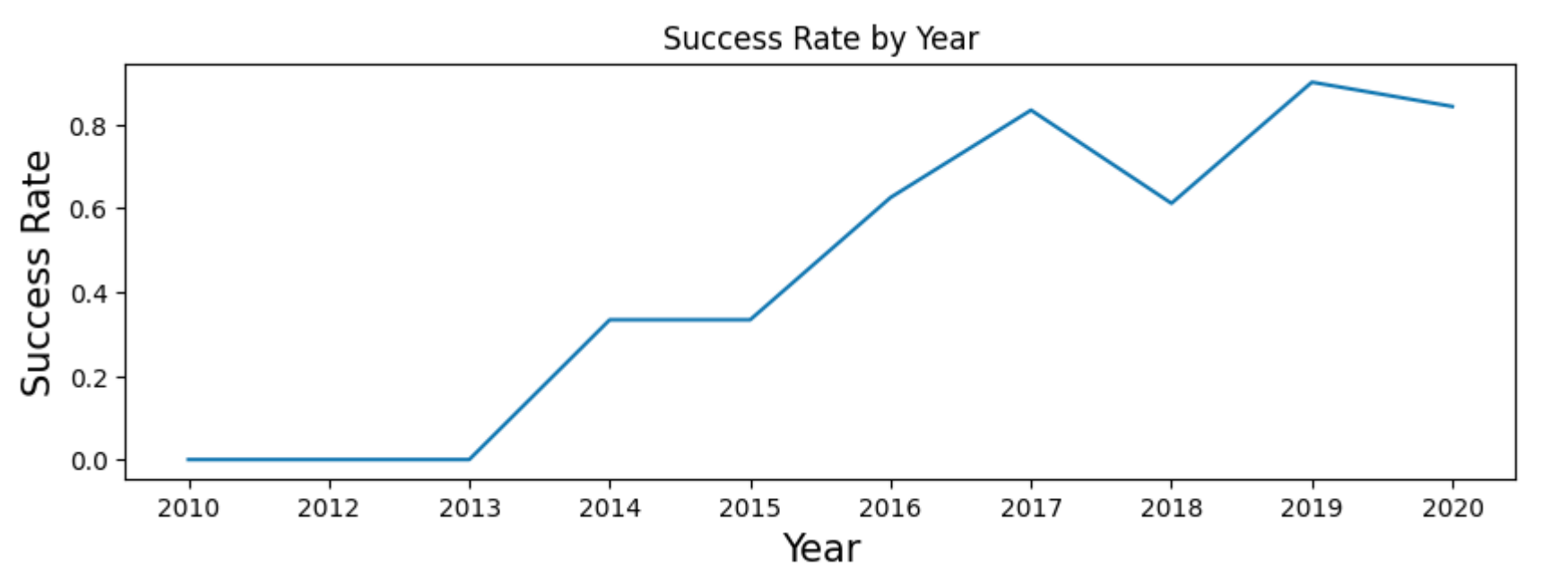
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

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



The success rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

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

There are 4 different launch sites in total

Launch Site Names Begin with 'CCA'

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

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

The landing outcomes are all not successful

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)';
```

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

total_payload
45596

The total payload mass by NASA is 45596 kg

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(PAYLOAD_MASS__KG_) AS Avg_Payload FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1';
```

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

Avg_Payload
2928.4

The average payload mass by F9 v1.1 is 2928.4kg

First Successful Ground Landing Date

```
%sql SELECT min(date) AS Early_Date from SPACEXTBL where Landing_Outcome LIKE 'Success (ground pad)'
```

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

Early_Date
01-05-2017

The first successful ground landing data is at 01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT Customer, Landing_Outcome,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE Landing_Outcome ='Success (drone ship)' AND PAYLOAD_MASS__KG_ BE
 * sqlite:///my_data1.db
Done.
```

Customer	Landing_Outcome	PAYLOAD_MASS__KG_
SKY Perfect JSAT Group	Success (drone ship)	4696
SKY Perfect JSAT Group	Success (drone ship)	4600
SES	Success (drone ship)	5300
SES EchoStar	Success (drone ship)	5200

%sql SELECT DISTINCT Customer, Landing_Outcome,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE Landing_Outcome ='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

Landing by drone ship is the most successful

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, Count(*) AS Numbers FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

There is 1 failure in flight, 99 successes and 1 success with payload status unclear

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version, Max_Payload FROM (SELECT Booster_Version, MAX(PAYLOAD_MASS__KG_) AS Max_Payload FROM SPACEXTBL GROUP BY Booster_Version)
```

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

Booster_Version	Max_Payload
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092
F9 B4 B1045.1	362
F9 B4 B1045.2	2697
F9 B5 B1046.1	3600
F9 B5 B1046.2	5800

Different booster version have various max payload mass

2015 Launch Records

```
%sql SELECT SUBSTR(Date,4,2) AS Month, Booster_Version, Launch_site FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%drone%' AND SUBSTR(Date,7,4) =
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

%sql SELECT SUBSTR(Date,4,2) AS Month, Booster_Version, Launch_site FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Failure%drone%' AND SUBSTR(Date,7,4) = '2015'

In 2015, January and April, two launches failure with booster B1012 and B1015

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

```
%sql SELECT Landing_Outcome, COUNT(*) AS Numbers FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success%' AND Date BETWEEN '04-06-2010' AND '20-03-2017' G
```

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

Landing_Outcome	Numbers
Success	20
Success (drone ship)	8
Success (ground pad)	6

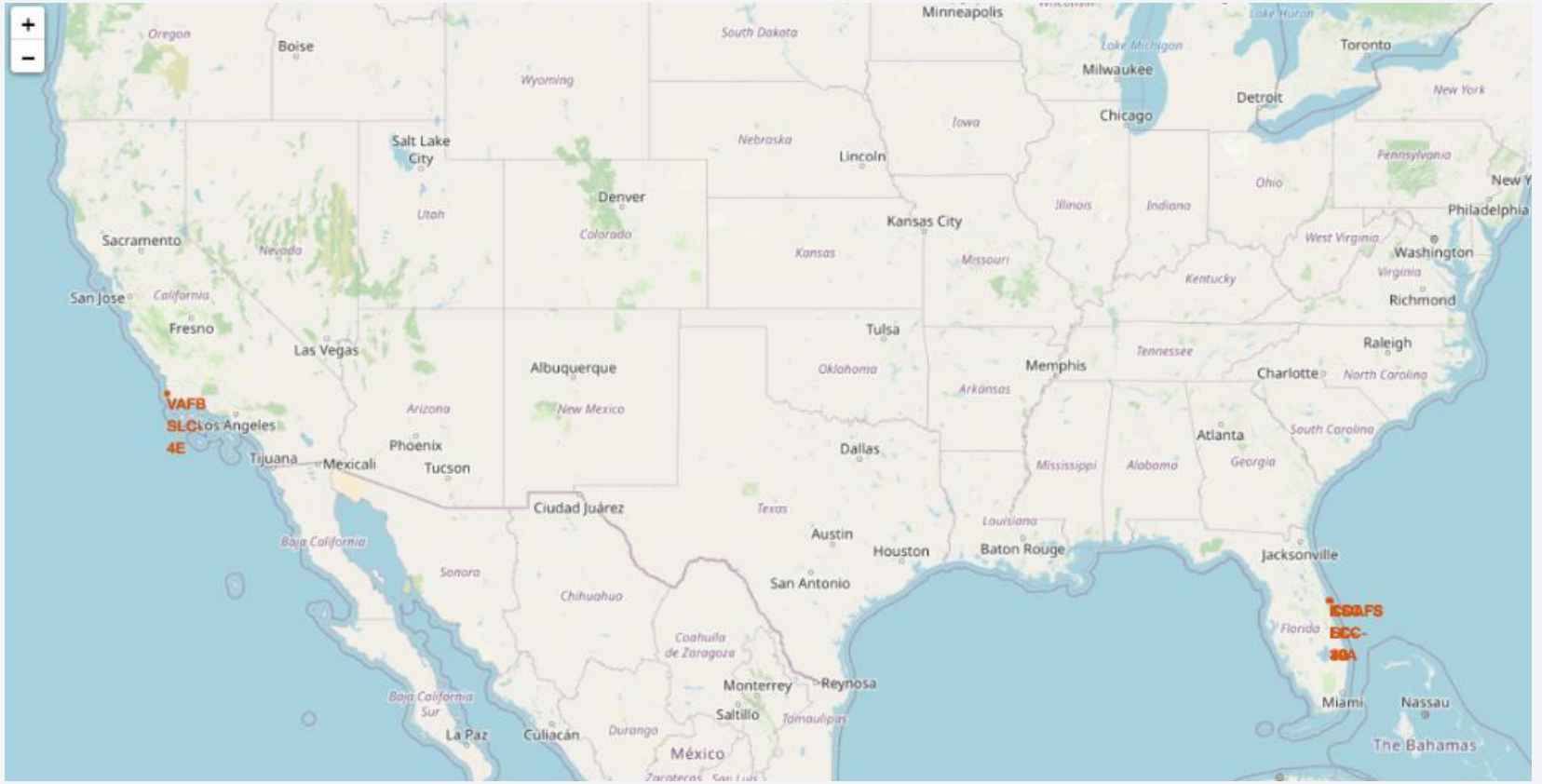
%sql SELECT Landing_Outcome, COUNT(*) AS Numbers FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success%' AND Date BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY Landing_Outcome ORDER BY Numbers DESC;

There are 20 successful landings between 2010-06-04 and 2017-03-20, 8 successful drone ship landing and 6 successful ground pad landing

Section 3

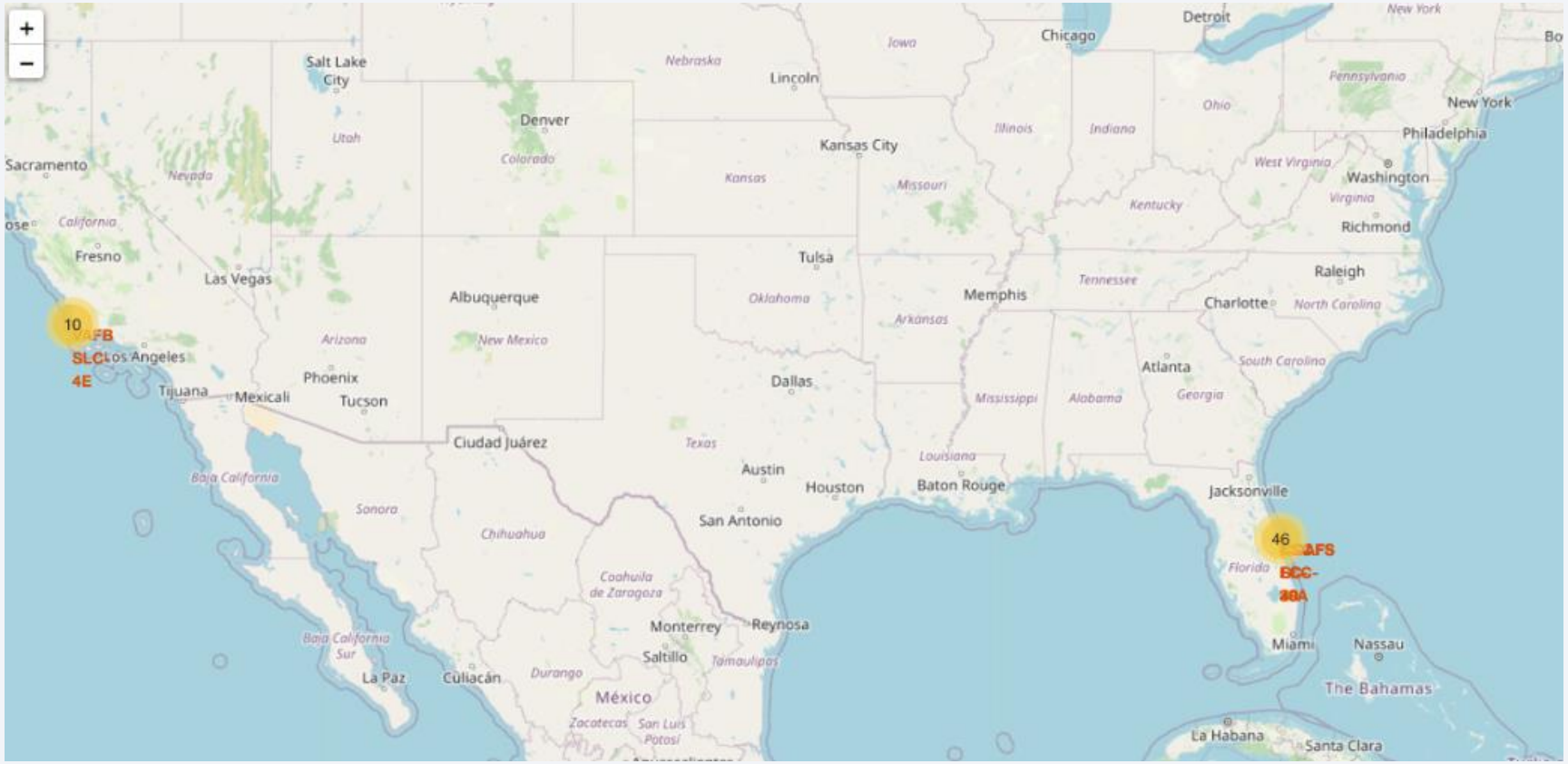
Launch Sites Proximities Analysis

<All launch sites on a map>



- All launch sites in proximity to the Equator line
- All launch sites in very close proximity to the coast

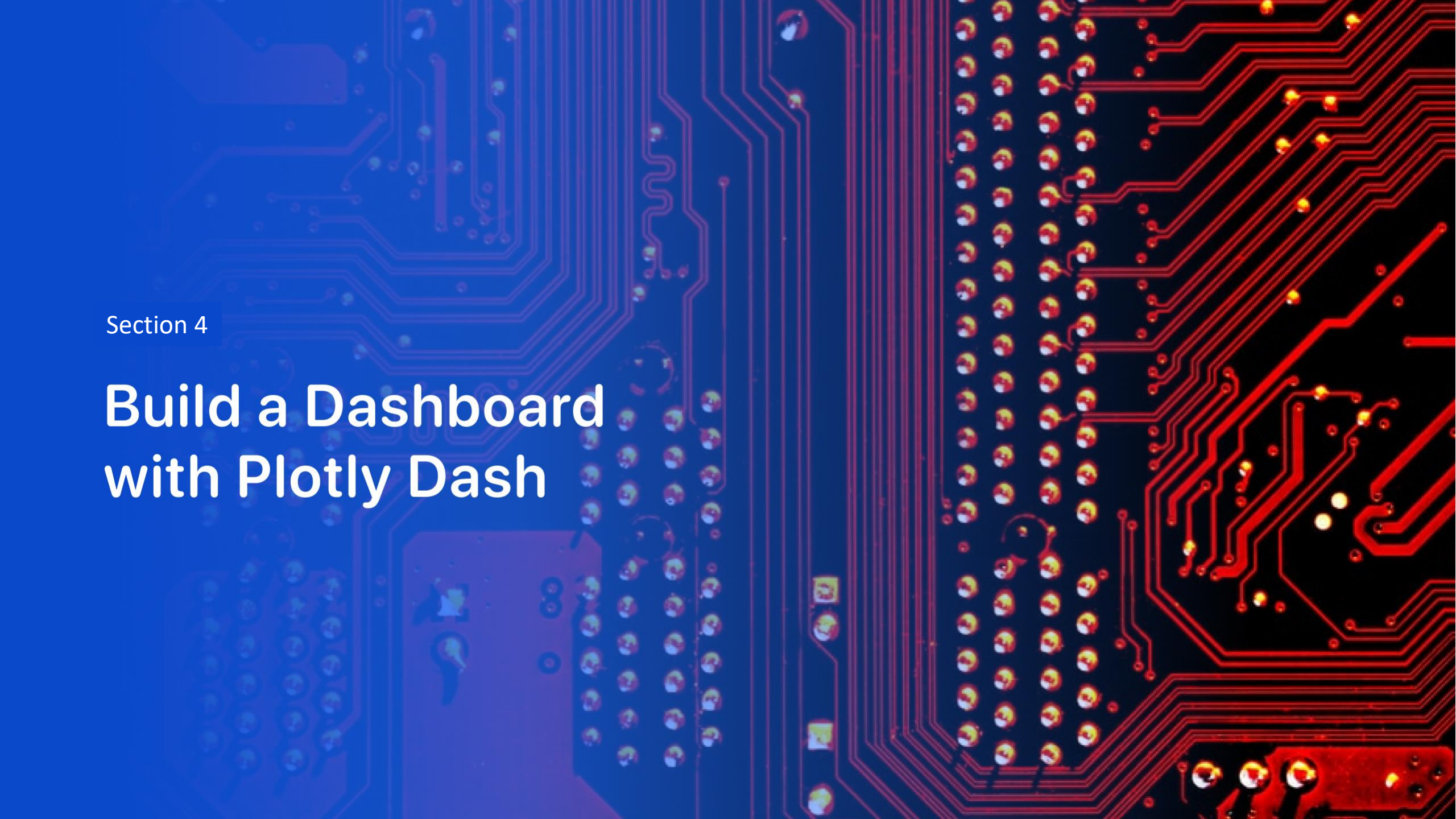
<Launch outcomes of different site>



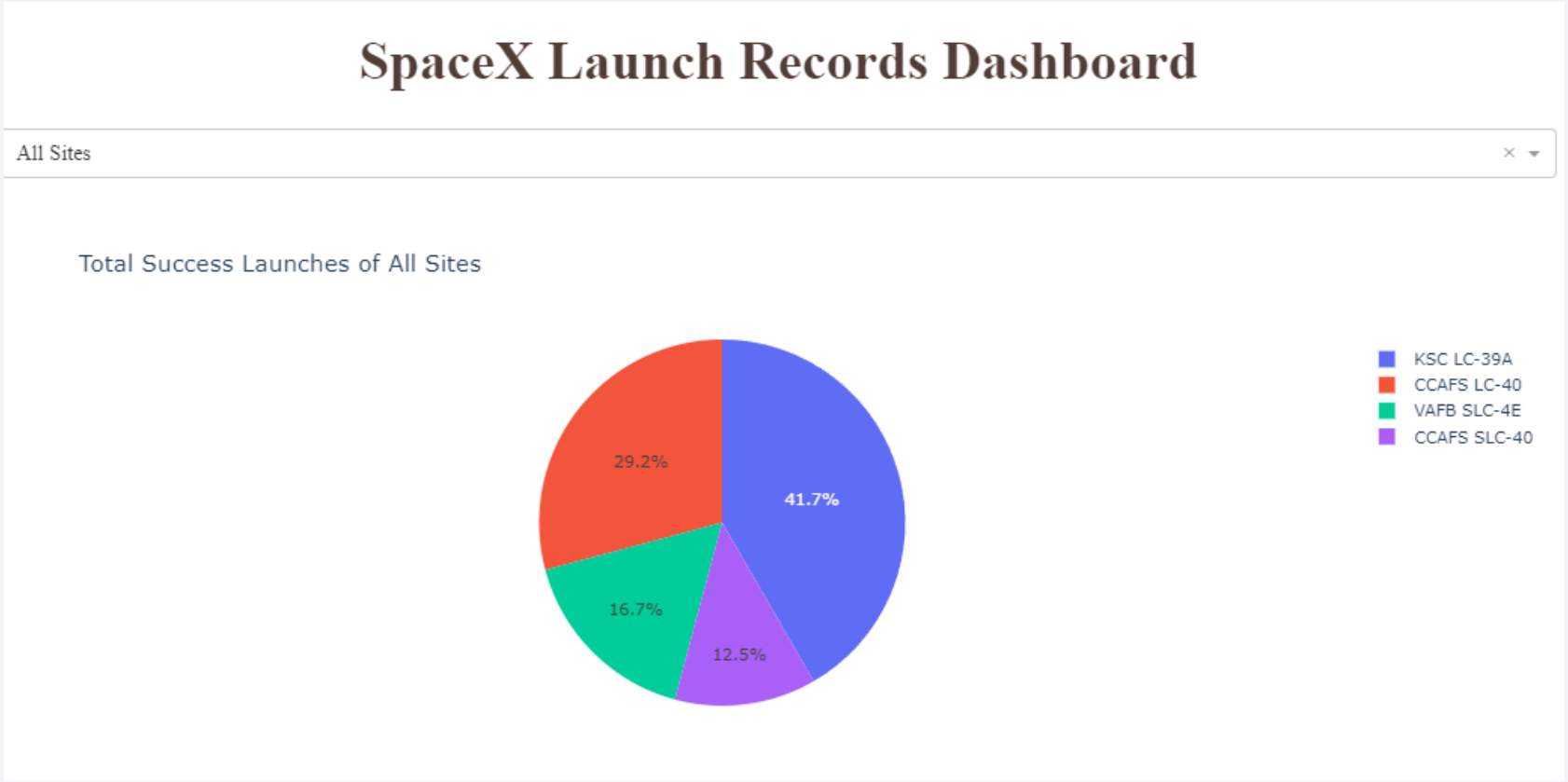
- At the left coast site, there are 10 trails, and 46 trails at the right coast site

<The proximity of the launch sites>

- CCAFS SLC-40 is 0.9km far away from the nearest coast

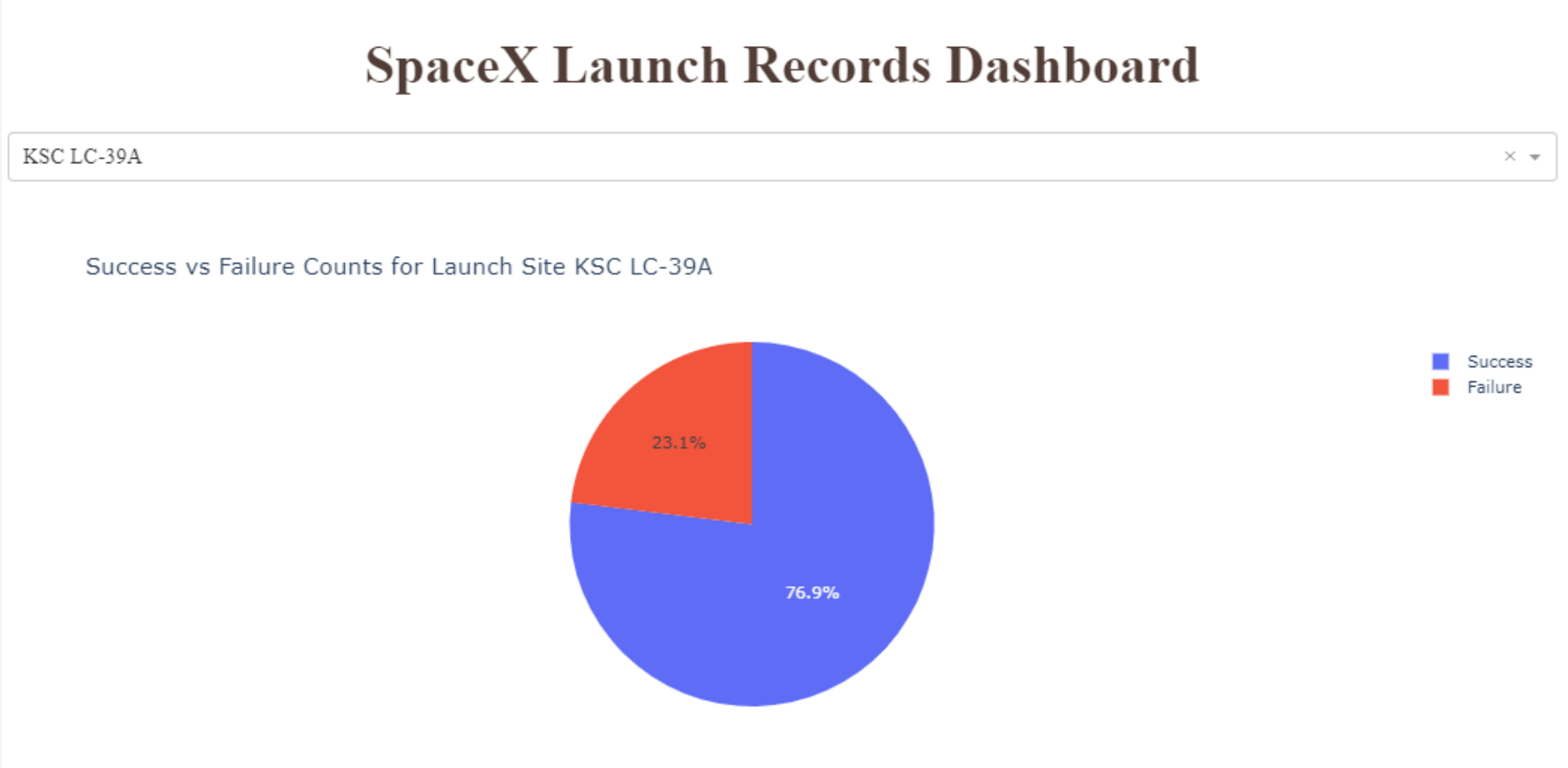

Section 4
Build a Dashboard with Plotly Dash

<All site launch records>



- KSC LC-39A: 41.7%
- CCAFS LC-40: 29.2%
- VAFB SLC-4E: 16.7%
- CCAFS SLC-40: 12.5%

<Launch site with highest launch success ratio>



- KSC LC-39A: success launch ratio: 76.9%

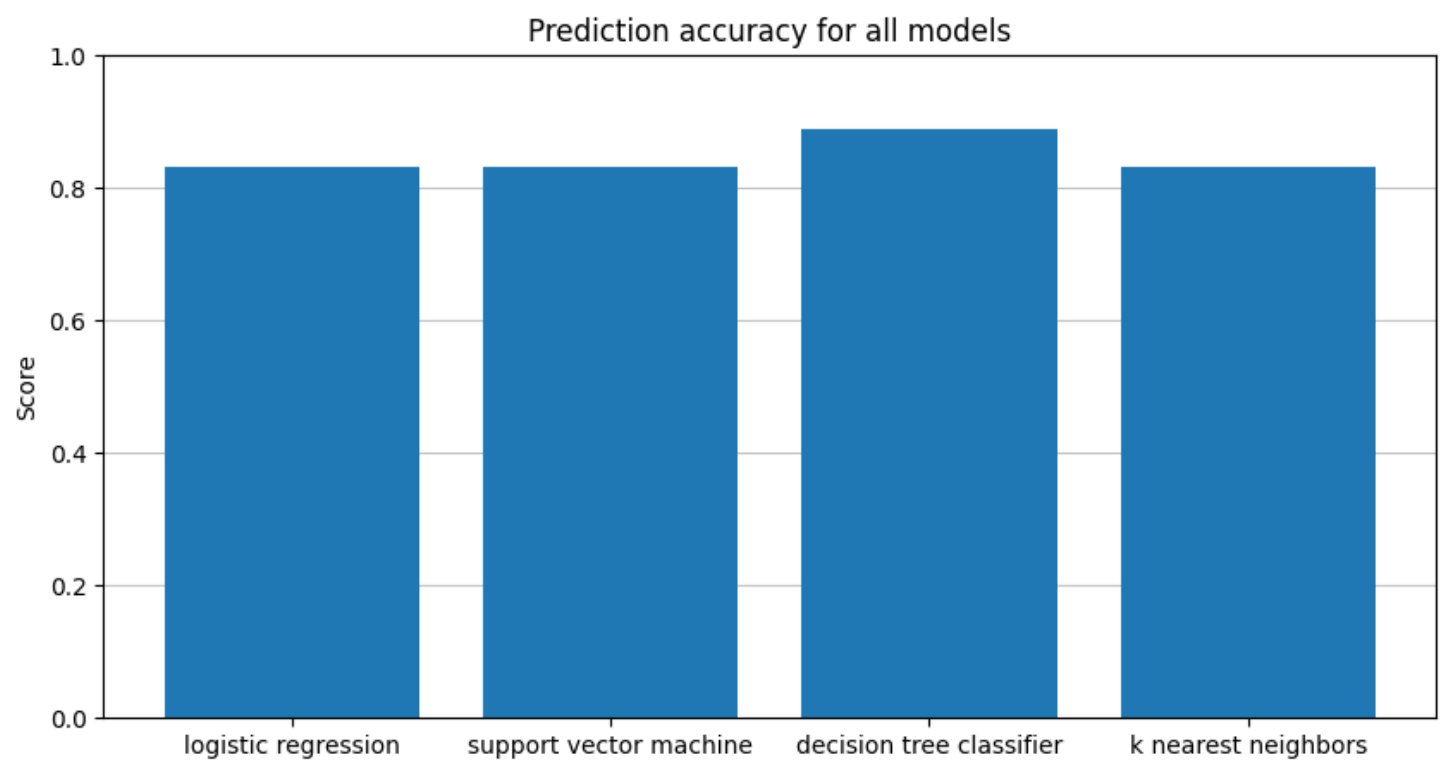
<Payload vs. Launch Outcome>

- Payload mass with 2k to 5k has success landing outcome
- V1.0 can take the heaviest payload
- FT has the most successful landing

Section 5

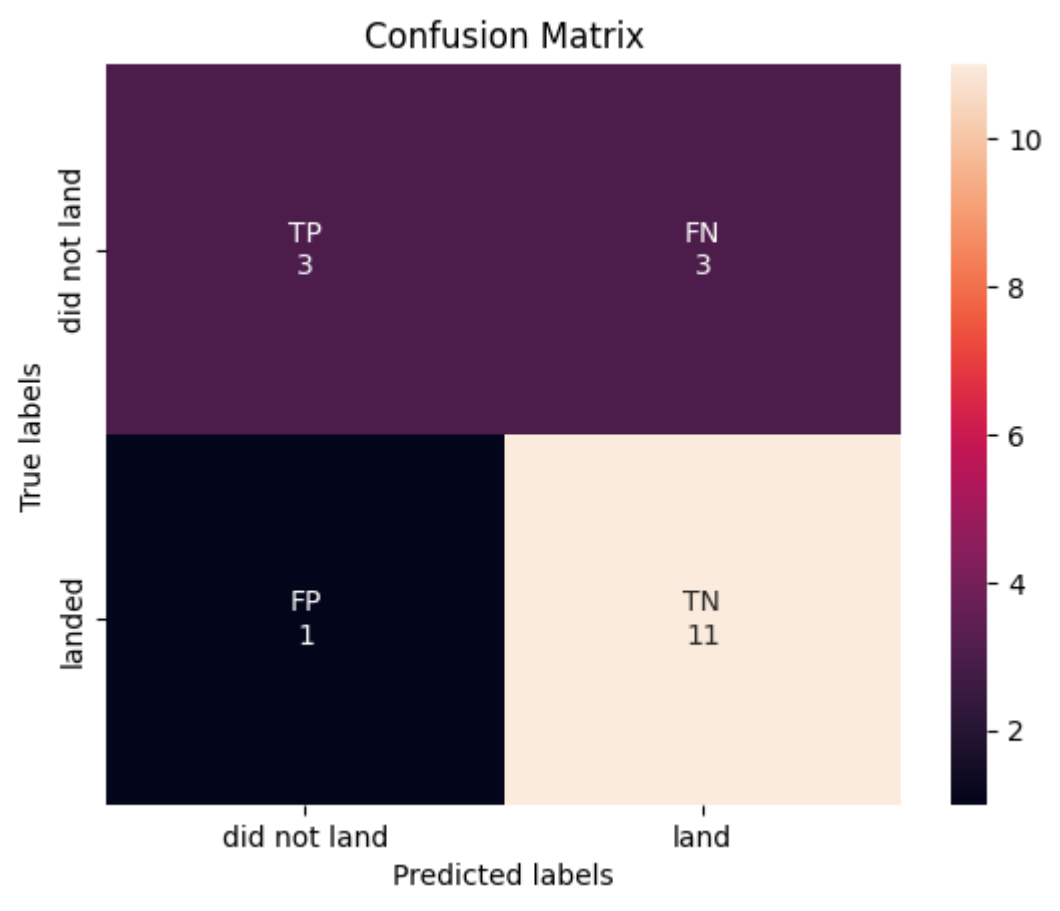
Predictive Analysis (Classification)

Classification Accuracy



- Decision tree classifier has the highest score

Confusion Matrix



- Decision tree predicts that there will be:
  - 3 true positive
  - 11 true negative
  - 3 false positive
  - 1 false positive

Conclusions

- CCAFS SLC has the lowest successful rate, while KSC LC 39A has the highest successful rate
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- ES-L1, GEO, HEO, SSO have highest successful rate.
- 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.
- The success rate since 2013 kept increasing till 2020
- Decision tree classifier gives the best prediction with accuracy of 88.9%

Appendix

- https://www.coursera.org/learn/applied-data-science-capstone/home/info
- https://github.com/general-chen/Coursera/tree/main/IBM_Data_Science_Professional_Certificate
- https://github.com/general-chen/Coursera/tree/main/IBM_Data_Science_Professional_Certificate/Applied_Data_Science_Capstone

Thank you!