IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

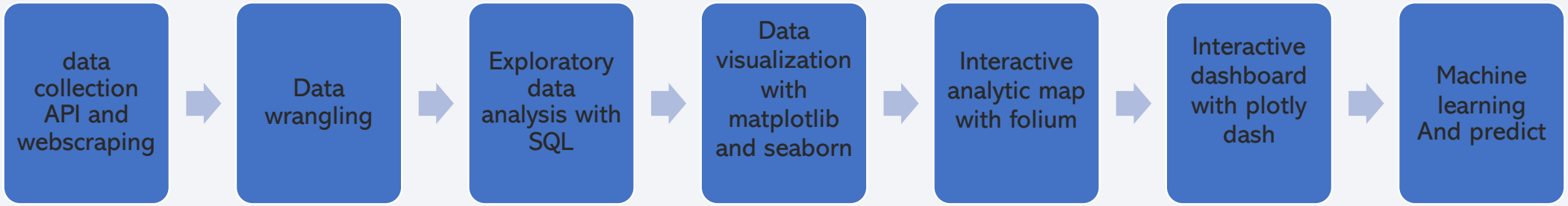
Feifei Wang
05/21/2022

Outline

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

Executive Summary

- Summary of methodologies



- Summary of all results

➢ flight number, payload mass, orbit type are related to the outcome.

➢ The launch sites were close to coastline, highway and railway, but far from cities.

➢ Site KSC LC-39A has the highest success rate and CCAFS SLC-40 has the least success launches.

➢ A smaller payload (e.g., 0-6000) has the largest success rate. FT booster also has the largest success rate.

➢ The machine learning process showed that since the dataset is limited, all the modals showed the same accuracy.

Introduction

- Project background and context

SpaceX, as the only commercialized space tech company, advertised that their Falcon 9 rocket launches are much less expensive than other providers because the first stage can be reused.

This is based on the success in the landing of the first stage.

- Problems you want to find answers

How the features of a launch affect the outcomes? Can we use the dataset of previous launches of SpaceX Falcon 9 to predict whether a launch will succeed or fail?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected via data collection API and webscraping using GET request and beautifulsoup and parsed into dataframe.
- Perform data wrangling
  - Missing values were replaced and landing outcome labels were created.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use scikit learn model functions(logistic regression, decision tree, KNN and svm) to build models, use gridsearch to find the best hyperparameter, and evaluate the models using accuracy and confusion matrix

Data Collection

REST API

GET request
from spaceX API

Parse content into Jason file

keep relevant features and Falcon 9 in dataframe

Scraping

GET request falcon 9 data from HTML page store in beautifulsoup

Extract columns and variables from the html table

parse into dataframe and keep relevant features

Data Collection – SpaceX API

GET request
from spaceX API

```
In [6]: spacex_url="https://api.spacexdata.com/v4/launches/past"

In [7]: response = requests.get(spacex_url)
```

Parse content into Jason file and normalize it

```
In [9]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spa

In [11]: # Use json_normalize meethod to convert the json result into a dataframe
         r1=response.json()
         data=pd.json_normalize(r1)
         data
```

keep relevant features and Falcon 9 in dataframe

```
In [21]: # Use the find_all function in the BeautifulSoup object, with element type `table`
         # Assign the result to a list called `html_tables`
         html_tables=BeautifulSoup.find_all('table')

In [22]: # Let's print the third table and check its content
         first_launch_table = html_tables[2]
         print(first_launch_table)
```

Github: https://github.com/w234feifei/Applied-Data-Science-Capstone/blob/main/FFjupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

GET request falcon 9 data from HTML page store in beautifulsoup

Extract columns and variables from the html table

parse into dataframe and keep relevant features

Github: https://github.com/w234feifei/Applied-Data-Science-Capstone/blob/main/FFjupyter-labs-webscraping.ipynb

In [17]:
```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

In [18]:
```
response=requests.get(static_url)
# assign the response to a object
# use requests.get() method with the provided static_url
```

In [19]:
```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
BeautifulSoup=BeautifulSoup(response.text)
```

In [21]:
```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables=BeautifulSoup.find_all('table')
```

In [26]:
```
column_names = []

# Apply find_all() function with `th` element on first_launch_table
th=first_launch_table.find_all('th')
# Iterate each th element and apply the provided extract_column_from_header() to get a column name
for i in range(len(th)):
    name=extract_column_from_header(th[i])
    if (name is not None and len(name)>0):
        column_names.append(name)

# Append the Non-empty column name (`if name is not None and len(name) > 0`) into a list called column_names
```

In [38]:
```
df=pd.DataFrame({key:pd.Series(value) for key,value in launch_dict.items()})
```

Data Wrangling

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Github: https://github.com/w234feifei/Applied-Data-Science-Capstone/blob/main/FFlabs-jupyter-spacex-Data%2520wrangling.ipynb

EDA with Data Visualization

Categorical plots of flight number vs. payload mass, flight number vs. launch site, payload vs. launch site, flight number vs. orbit type, payload vs. orbit type,

- Visualize the relationship between the two variables.

Bar plot of success rate vs. orbit type

- Visualize the relationship between success rate and orbit type.

Line plot of success rate vs. year

- Visualize the launch success yearly trend

Github: https://github.com/w234feifei/Applied-Data-Science-Capstone/blob/main/FFjupyter-labs-eda-dataviz.ipynb

EDA with SQL

```
SELECT DISTINCT Launch_Site from SPACEXTBL;
```

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

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

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

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer='NASA (CRS)';
```

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

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version='F9 v1.1';
```

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

```
SELECT MIN(Date) FROM SPACEXTBL WHERE Landing__Outcome= 'Success (ground pad)';
```

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

```
SELECT Booster_Version FROM SPACEXTBL WHERE (Landing__Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

```
SELECT SUM(Mission_Outcome LIKE 'Success%') AS TOTAL_SUCCESS, SUM(Mission_Outcome LIKE 'Failure%') as TOTAL_FAILURE FROM SPACEXTBL;
```

•List the total number of successful and failure mission outcomes

```
SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

•List the names of the booster_versions which have carried the maximum payload mass.

```
SELECT Booster_Version, Launch_Site FROM SPACEXTBL WHERE (Landing__Outcome LIKE 'Failure (drone ship)' AND Date LIKE '2015%');
```

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

```
SELECT COUNT(*) as COUNTS, Landing__Outcome FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing__Outcome ORDER BY COUNTS DESC;
```

•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: https://github.com/w234feifei/Applied-Data-Science-Capstone/blob/main/FFjupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Circles were added to show the highlighted area locations of launch sites
- Markers were added to show the names of launch sites
- Marker clusters were added to show the successful and failed launches in different colors for each sites
- Mouse position was added to show the coordinates of selected location
- Lines and makers were added to show the proximities of launch sites to coastline, cities, railway and highway.
- Github: https://github.com/w234feifei/Applied-Data-Science-Capstone/blob/main/FFlab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A pie chart was created. When “All” is selected in the dropdown box, the pie chart shows the total success launches for each site. When a single site is selected, the pie charts shows the success and failure launches for the selected sites. The pie chart is used to visualize the proportion of success and failure launches for single site and which sites have the highest and lowest success launches.
- A scatter plot was created to show the relationship between payload mass and outcomes for all sites and each site respectively responding to the dropdown menu selection. The payload mass range shown can be changed by the rangeslider. The scatter plot is used to show the how the success and failure are related to the payload mass.
- GitHub URL: https://github.com/w234feifei/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

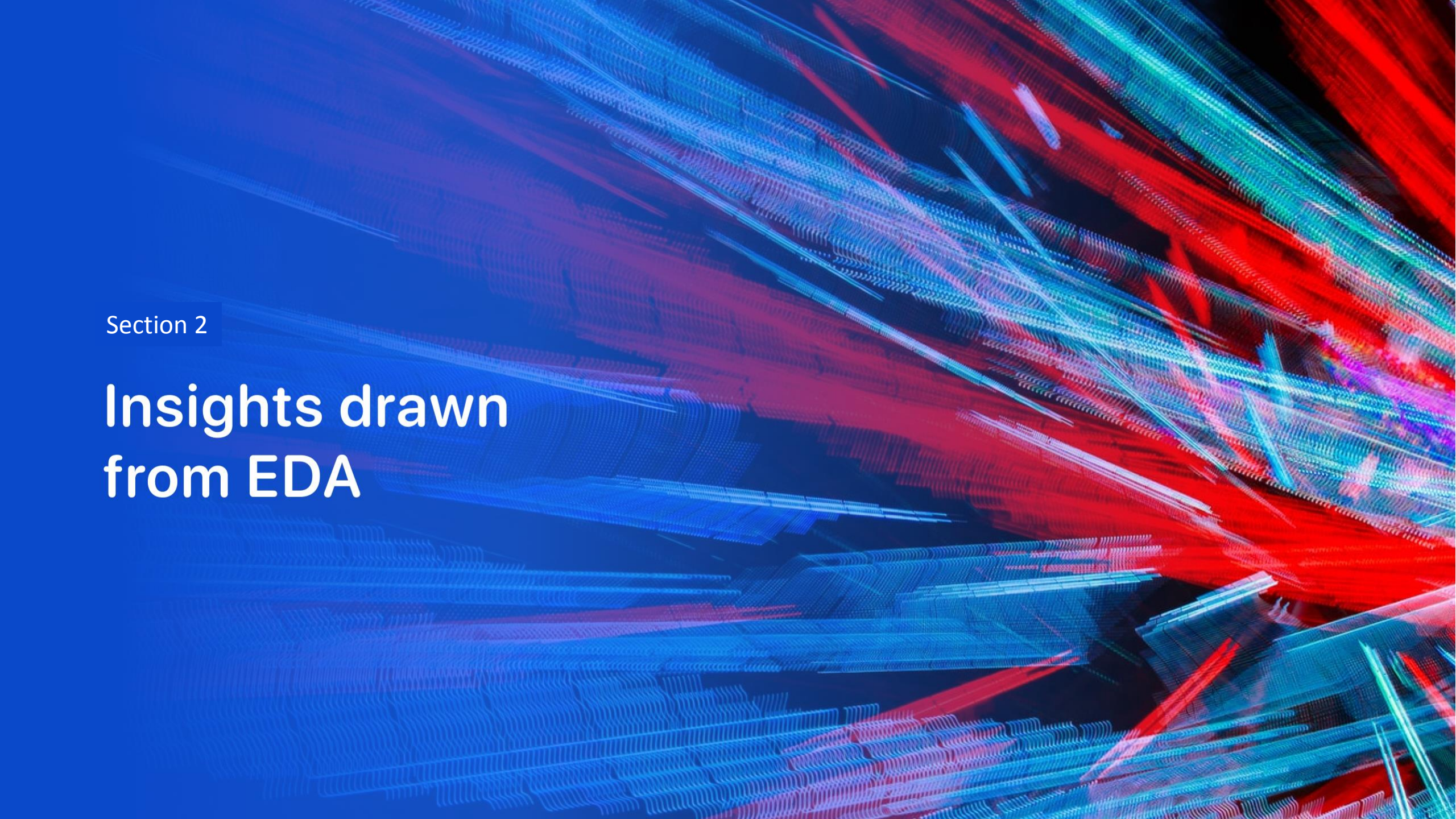
Predictive Analysis (Classification)

Logistic regression	Decision tree	KNN	SVM
Gridsearch and find the best parameters (C: 0.01; penalty: 12, solver: lbfgs)	Gridsearch and find the best parameters (criterion:entropy; max_depth:2, max_features:auto and min_samples_leaf)	Gridsearch and find the best parameters (algorithm:auto; n_neighbors: 9; p: 1)	Gridsearch and find the best parameters C: 1.0; gamma: 0.316, kernel: sigmoid
Find the accuracy on the validation data and test data	Find the accuracy on the validation data and test data	Find the accuracy on the validation data and test data	Find the accuracy on the validation data and test data
Predict with the trained model and generate a confusion matrix	Predict with the trained model and generate a confusion matrix	Predict with the trained model and generate a confusion matrix	Predict with the trained model and generate a confusion matrix

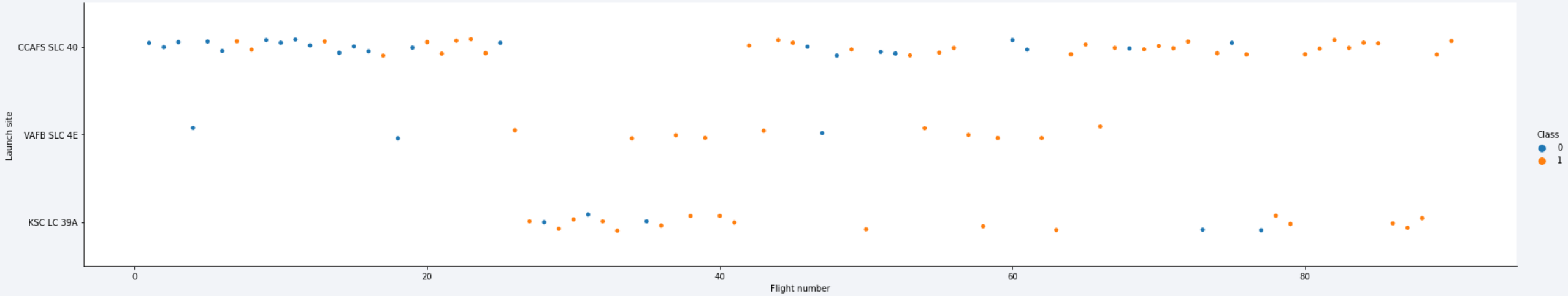
- GitHub URL: https://github.com/w234feifei/Applied-Data-Science-Capstone/blob/main/FFSpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Obtained exploratory data analysis results
- Obtained interactive analytics demo in screenshots
- Obtained predictive analysis results

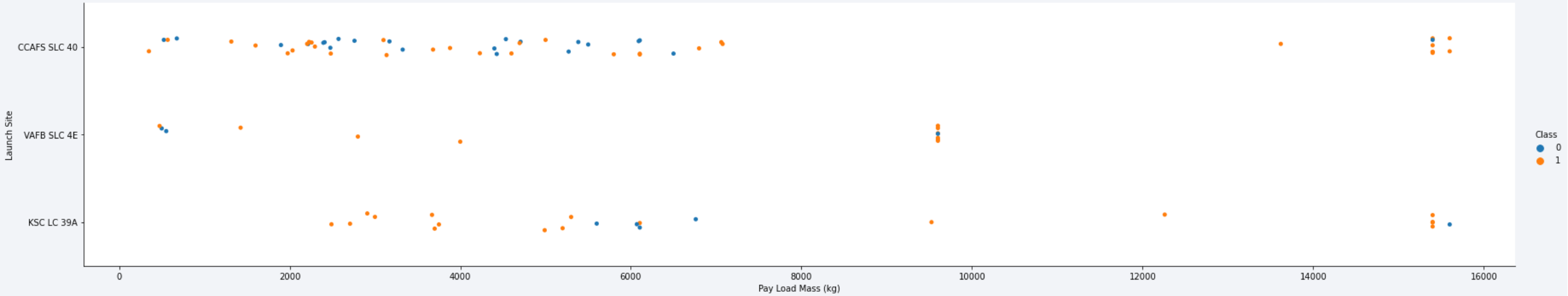

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 have flight number at 0-25 and 40-100, VAFB SLC 4E have flight number at 0-70 and KSC LC 39A have flight number at 25-100. VAFB SLC 4E has the least launches. The launch success rates of KSC LC 39A and VAFB SLC 4E are higher than those of CCAFS SLC 40.

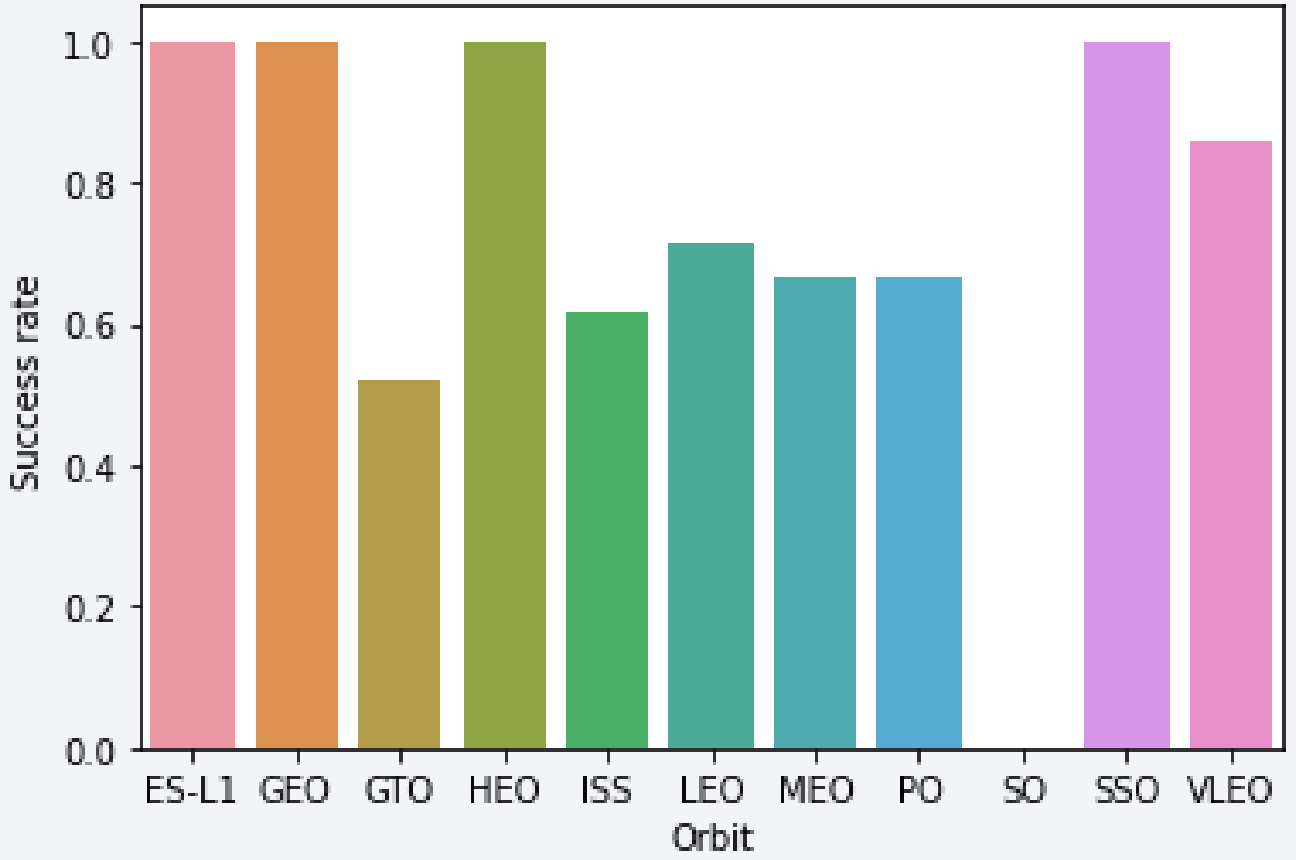
Payload vs. Launch Site



- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- For CCAFS SLC 40, there are no launches with payload mass at 8000-12000.
- For KSC LC 39A, there are no launches with payload mass at 0-2000.

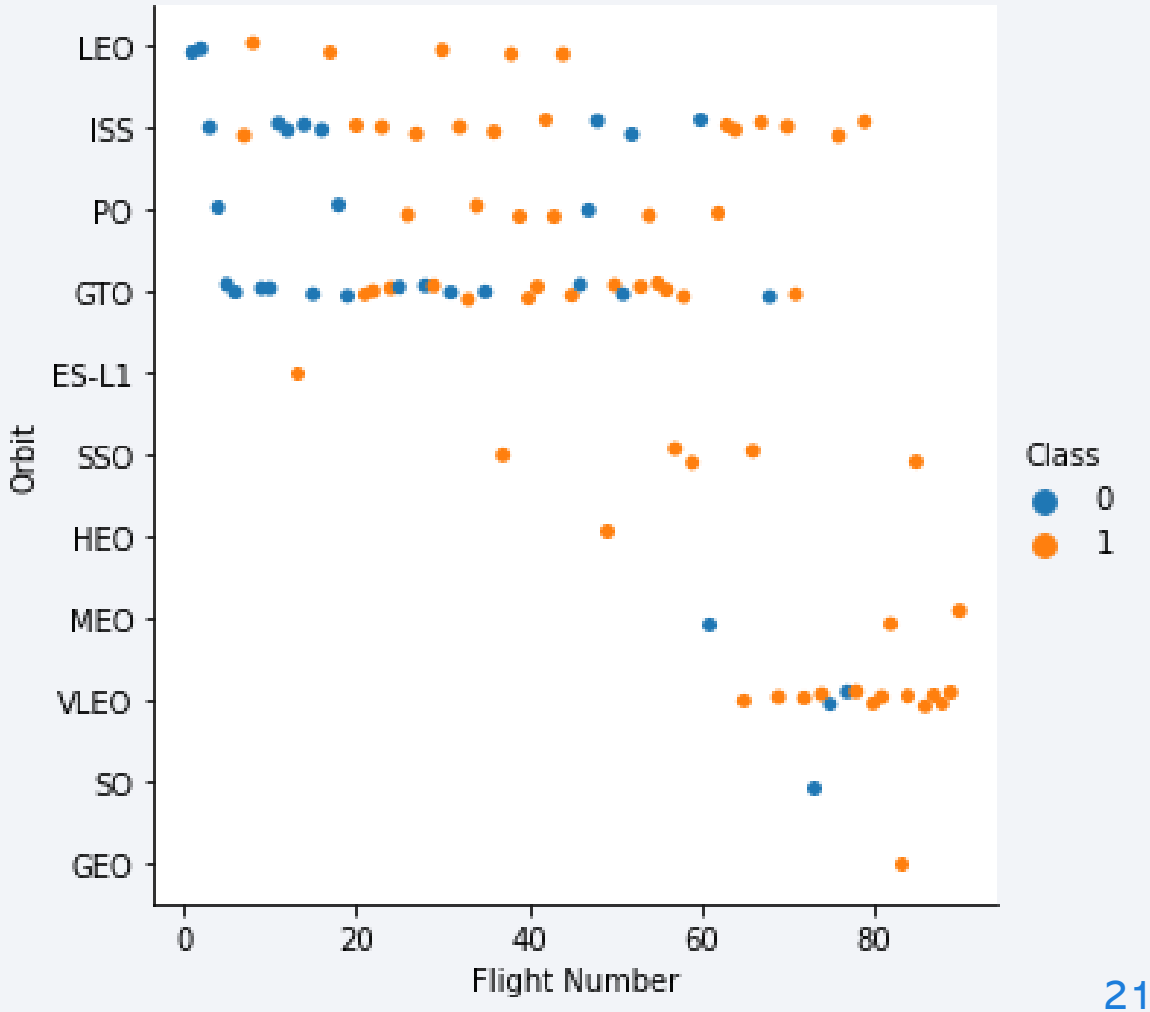
Success Rate vs. Orbit Type

- The orbits ES-L1, GEO, HEO, SSO AND VLEO have the highest success rate of 1. The SO has the lowest success rate of 0.



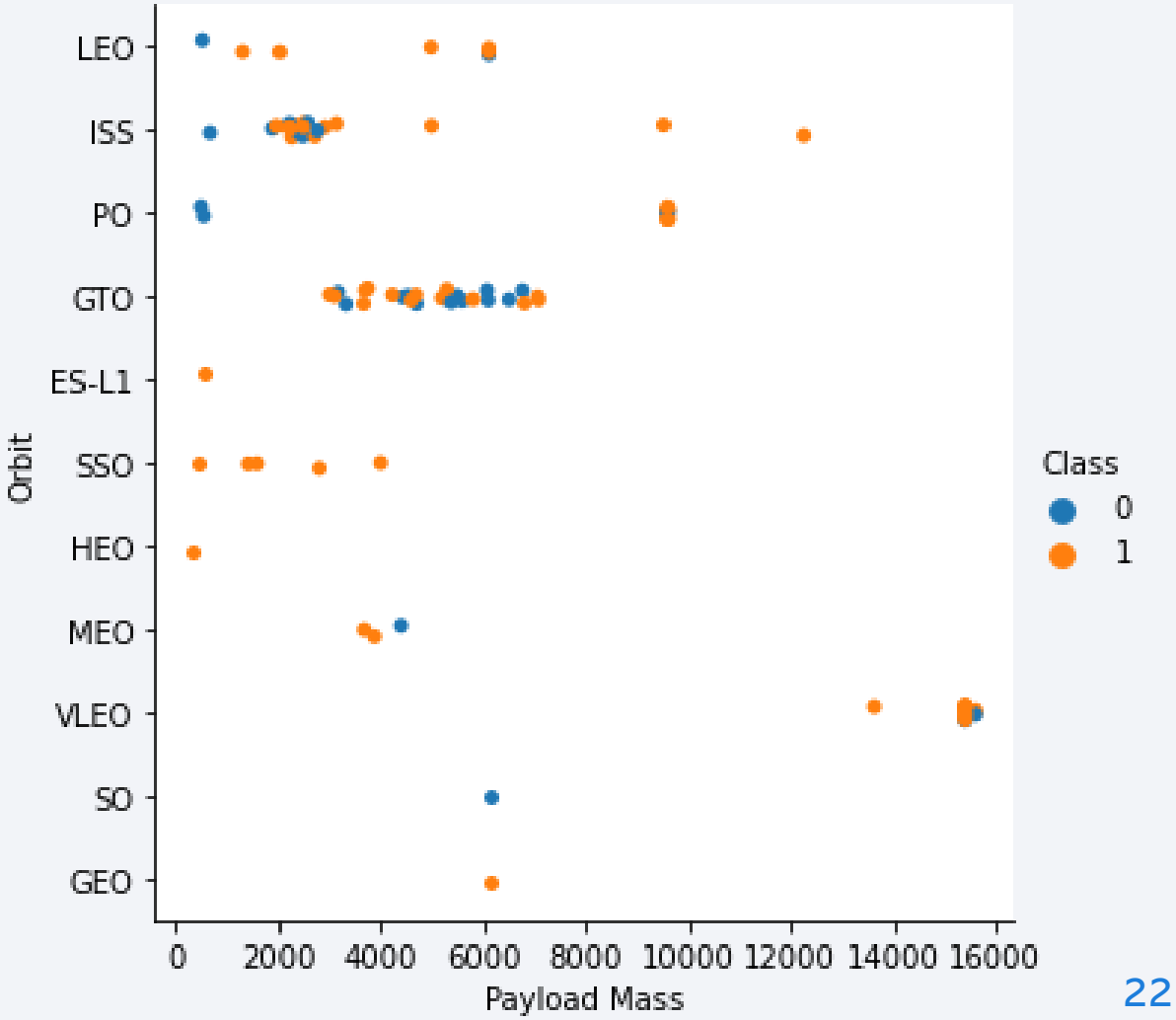
Flight Number vs. Orbit Type

- 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.



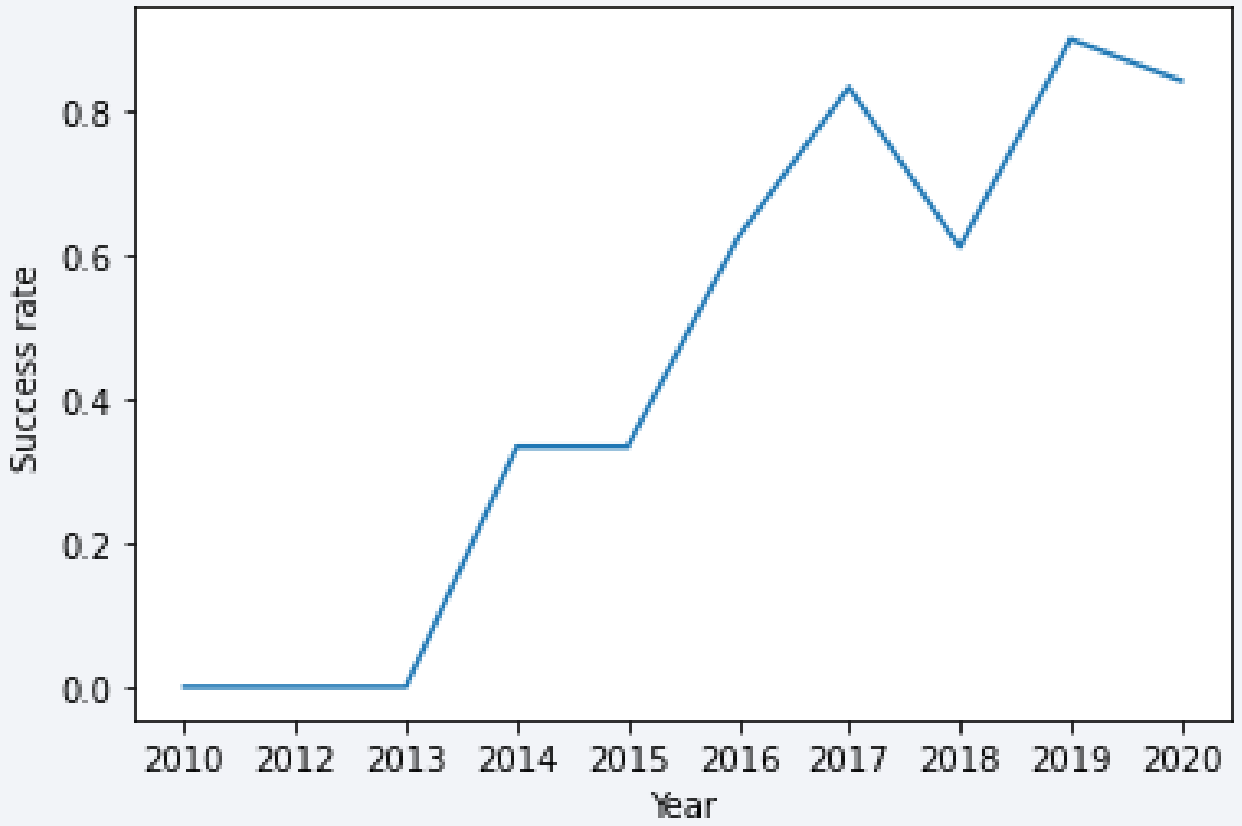
Payload vs. Orbit Type

- 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

- Find the names of the unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39, VAFB SLC-4E.

Task 1

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

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

```
 * ibm_db_sa://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`: find the top 5 records for launch site CCAFS LC-40.

Task 2

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

In [6]:
```
%%sql
SELECT * from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/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

Total Payload Mass

- The total payload carried by boosters from NASA is calculated to be 45596.

Task 3

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://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is calculated to be 2928

Task 4

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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version='F9 v1.1';
```

```
 * ibm_db_sa://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

1
2928

First Successful Ground Landing Date

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

Task 5

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://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
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 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Task 6

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 Booster_Version FROM SPACEXTBL WHERE (Landing__Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

```
 * ibm_db_sa://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are 100 and 1, respectively.

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT SUM(Mission_Outcome LIKE 'Success%') AS TOTAL_SUCCESS, SUM(Mission_Outcome LIKE 'Failure%') as TOTAL_FAILURE
FROM SPACEXTBL;
```

```
 * ibm_db_sa://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

total_success	total_failure
100	1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are F9 B5 B1048.4, F9 B5 B1049.4, F9 B5 B1051.3, F9 B5 B1056.4, F9 B5 B1048.5, F9 B5 B1051.4, F9 B5 B1049.5, F9 B5 B1060.2, F9 B5 B1058.3, F9 B5 B1051.6, F9 B5 B1060.3, F9 B5 B1049.7.

Task 8

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

```
%%sql
SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

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

2015 Launch Records

- The failed landing outcomes in drone ship in year 2015 occurred at CCAFS LC-40 with booster F9 v1.1 B1012 and F9 v1.1 B1015.

Task 9

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

```
%%sql
SELECT Booster_Version, Launch_Site FROM SPACEXTBL
WHERE (Landing__Outcome LIKE 'Failure (drone ship)' AND Date LIKE '2015%');
```

```
 * ibm_db_sa://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- 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: No attempt, failure(drone ship), success(drone ship), controlled(ocean), success (ground pad), failure (parachute). Uncontrolled (ocean), precluded (drone ship).

Task 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

```
%%sql
SELECT COUNT(*) as COUNTS, Landing__Outcome FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing__Outcome
ORDER BY COUNTS DESC;
```

```
 * ibm_db_sa://nqz00110:***@9938aec0-8105-433e-8bf9-0fbb7e483086.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32459/bludb
Done.
```

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

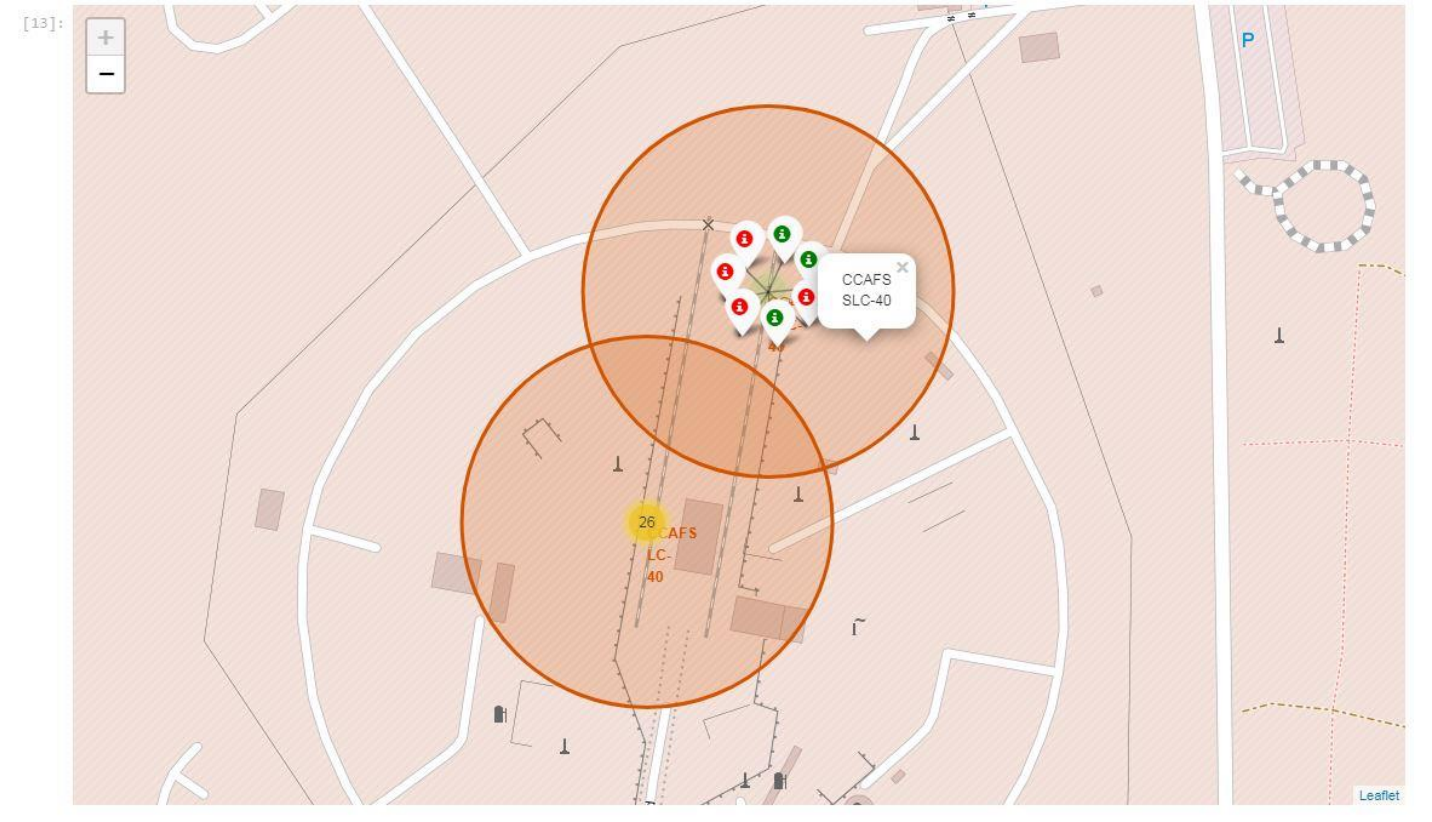

Section 3
Launch Sites
Proximities Analysis

Folium Map of all launch sites

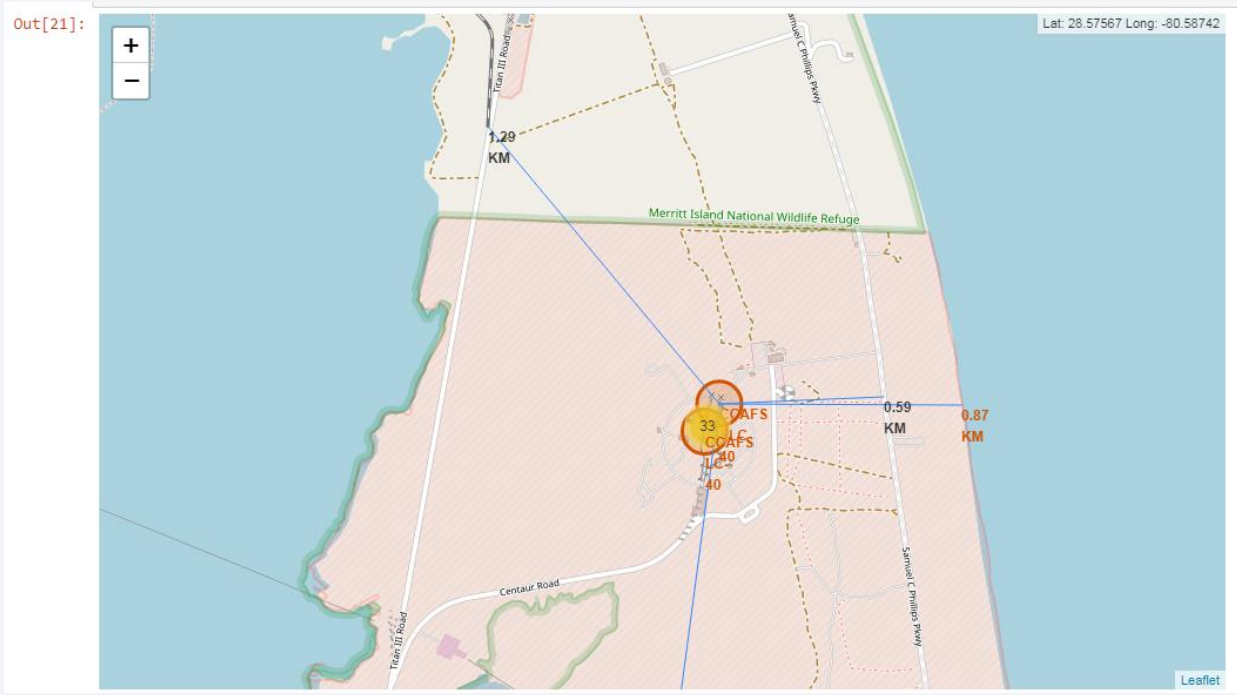
- All launch sites are marked on the global map and the names are shown in text.
- Most of the launch sites are proximity to the Equator line.
- All launch sites are in very close proximity to the coast.

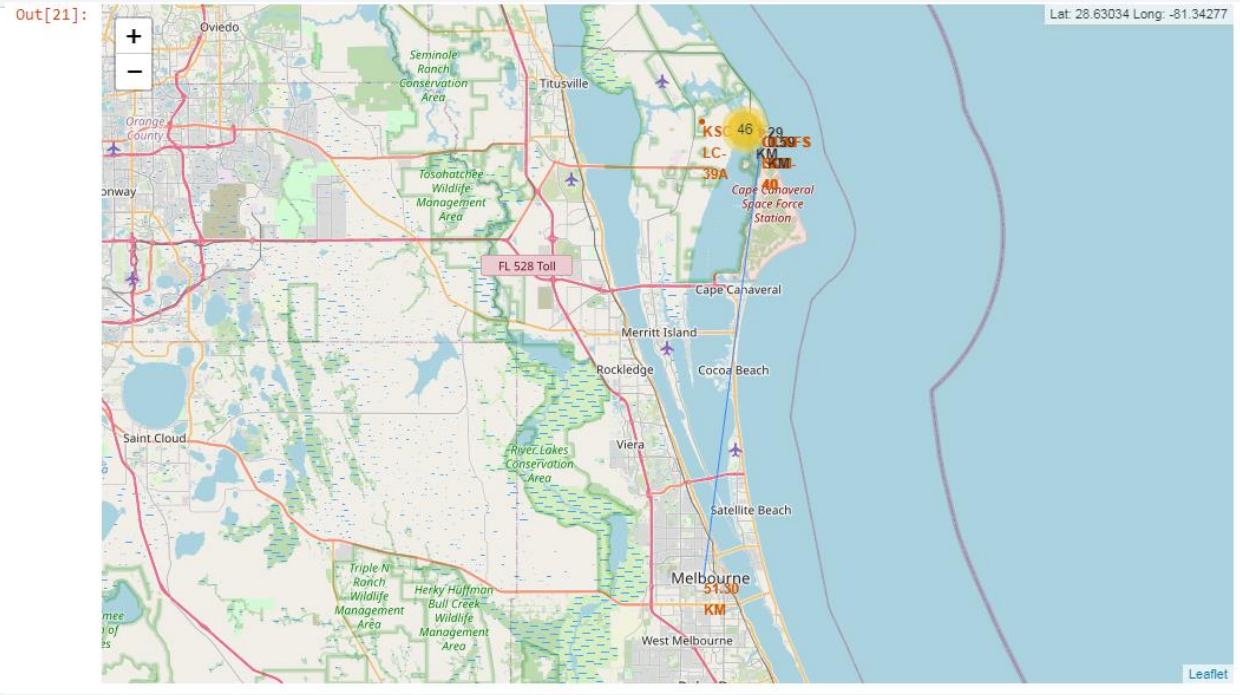
Folium Map of success/failed launches for each site

- The number of total launches are labeled in maker clusters for each site. And the success and failed launches are labeled in green and red, respectively, for each site.
- KSC LC-39A has the highest success rate.

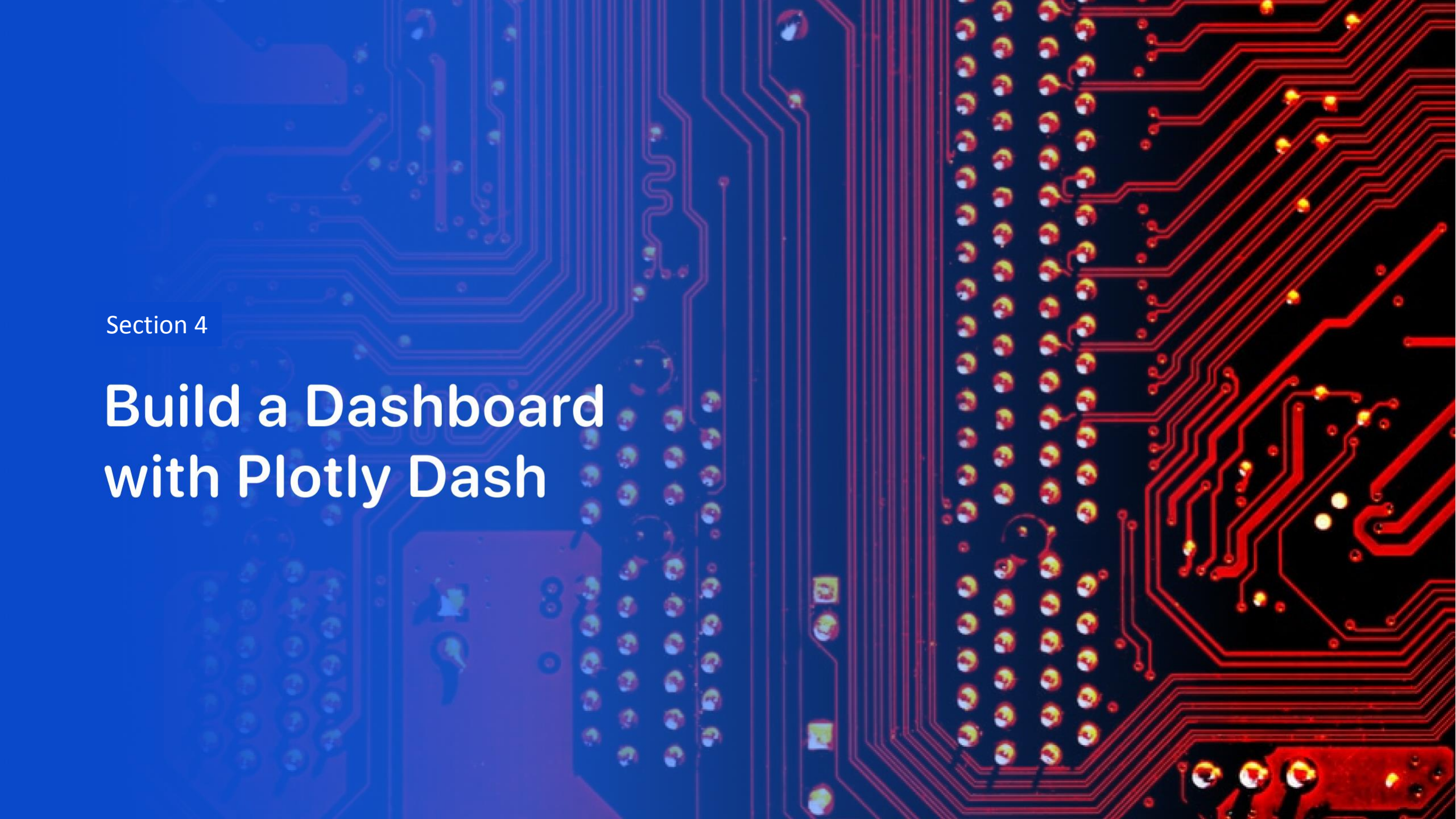


Folium Map of the proximities of launch site

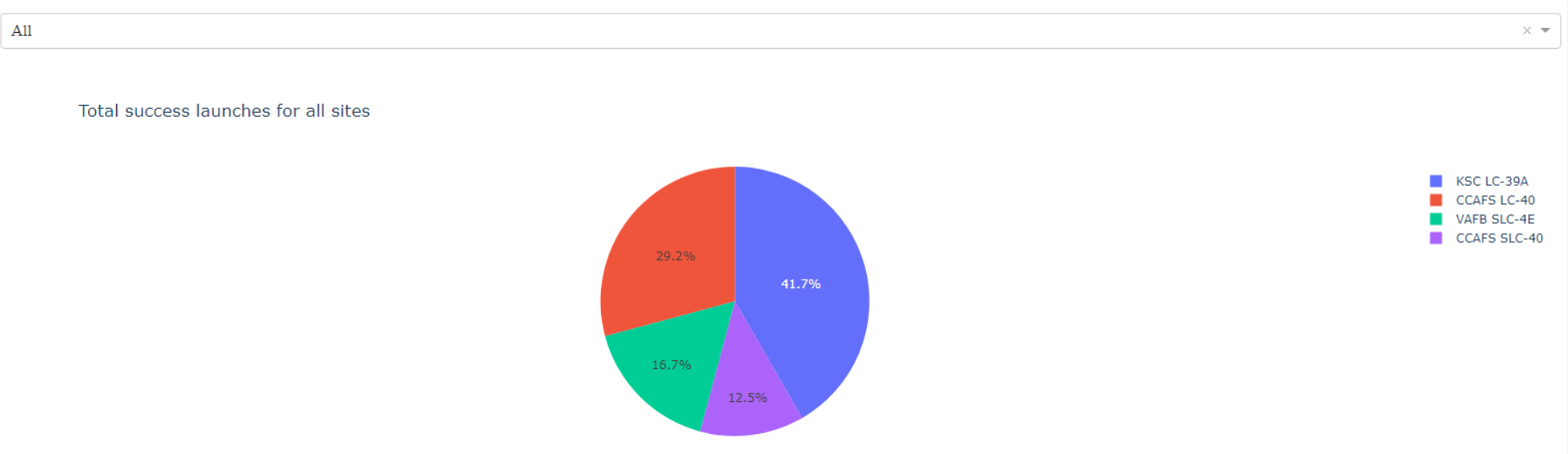




- Lines connected the launch site to the proximities such as coastline, railway, highway and city are drawn. The distance are calculated and shown beside the proximities.
- The launch sites are in close proximity to railways, highways and coastline, but keep certain distance away from cities.

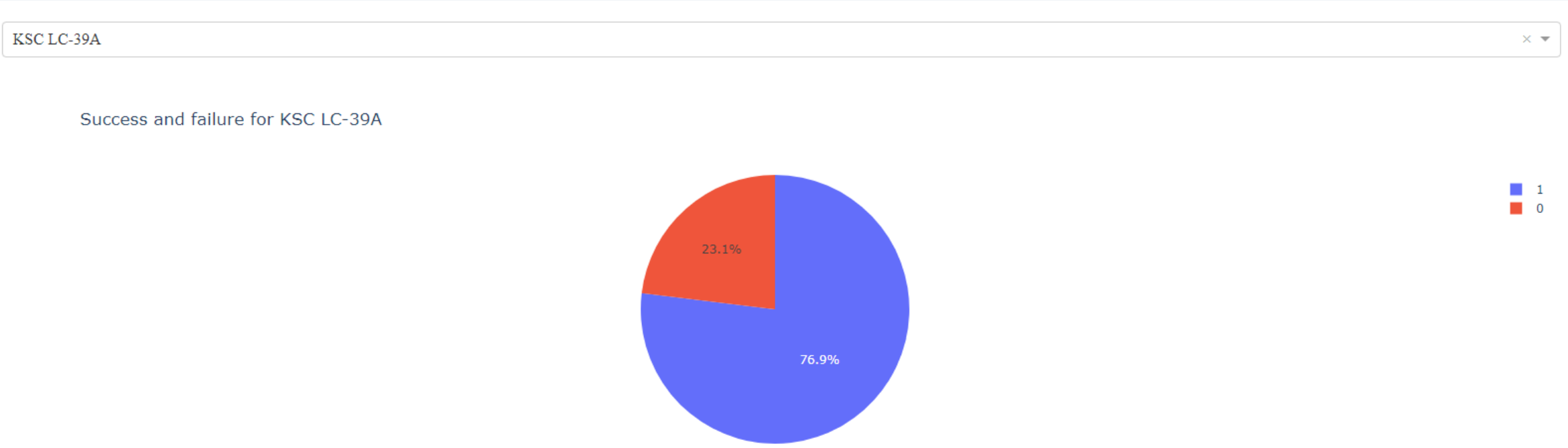

Section 4
Build a Dashboard with Plotly Dash

Pie chart of total success launches for all sites



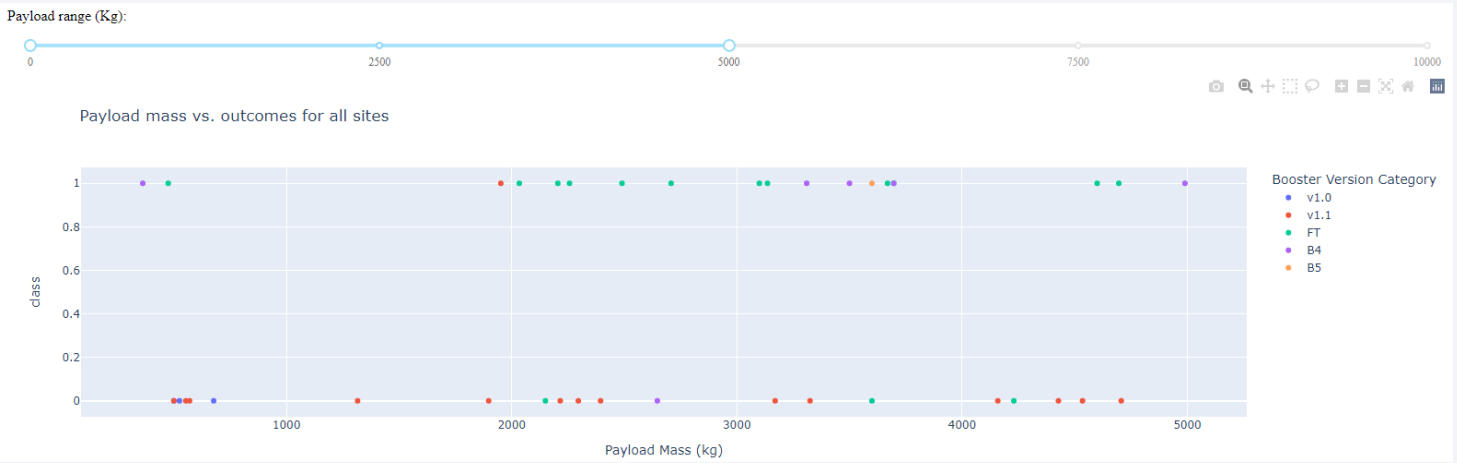
- The success for each site are proportioned in the pie chart.
- KSC LC-39A has the most success launches.
- CCAFS SLC-40 has the least success launches.

Pie chart for the launch site with highest launch success ratio

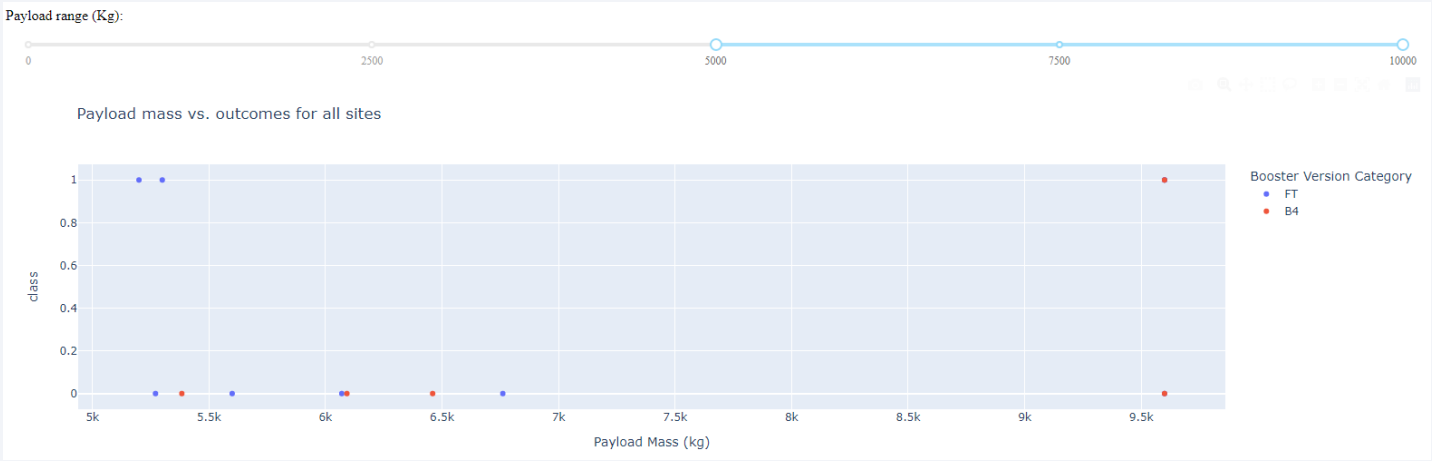


- The success and failure for KSC LC-39A, which has the highest launch success ratio, are proportioned into a pie chart.
- This site has success rate of 76.9% and failure rate of 23.1%.

Payload vs. Launch outcome scatter plot for all sites



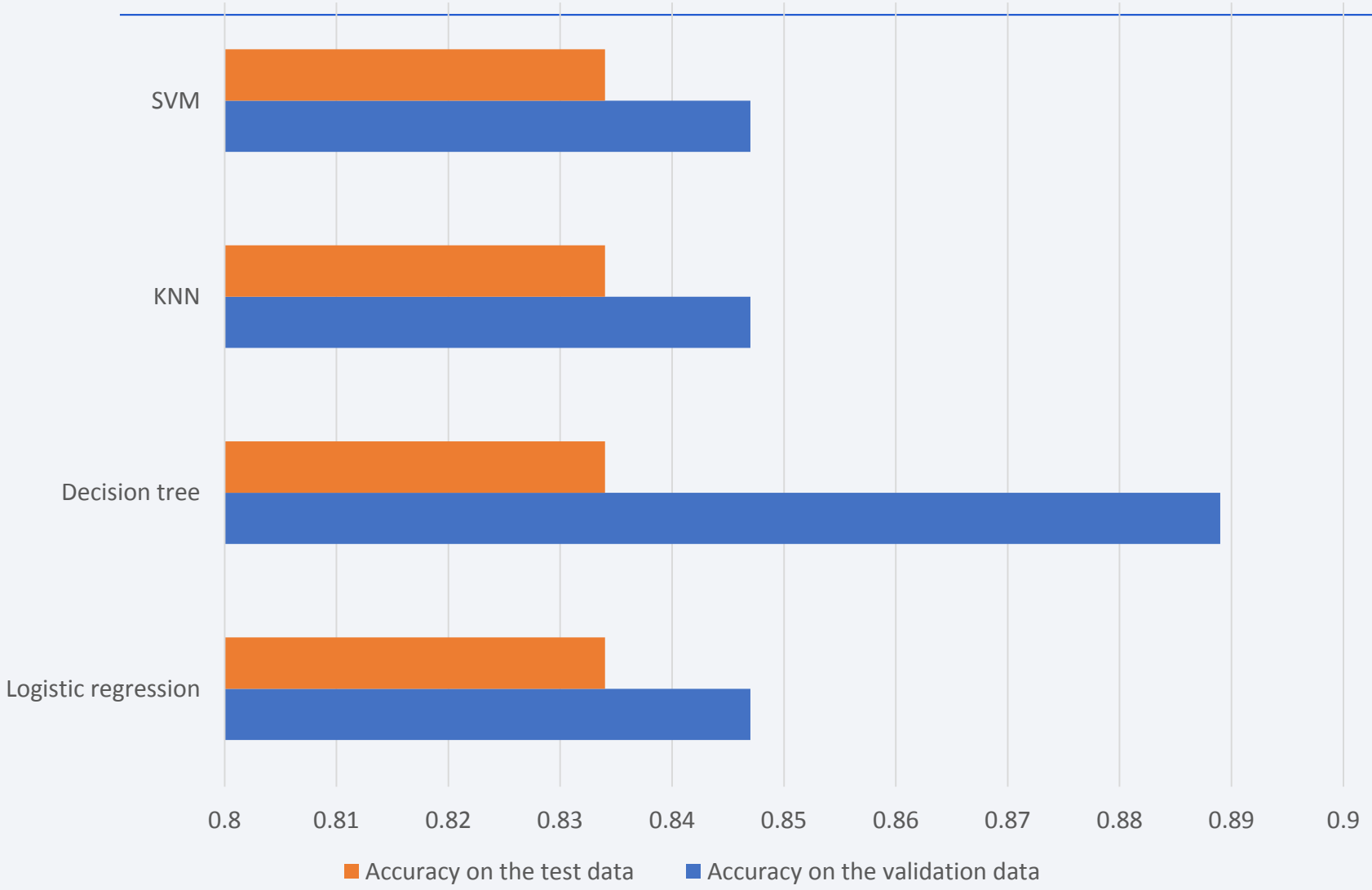
- The range of the payload mass of the scatter plot of payload vs. launch outcomes can be changed and different booster version are marked for the data points.
- A smaller payload (e.g., 0-6000) has the largest success rate. FT booster also has the largest success rate.



Section 5

Predictive Analysis (Classification)

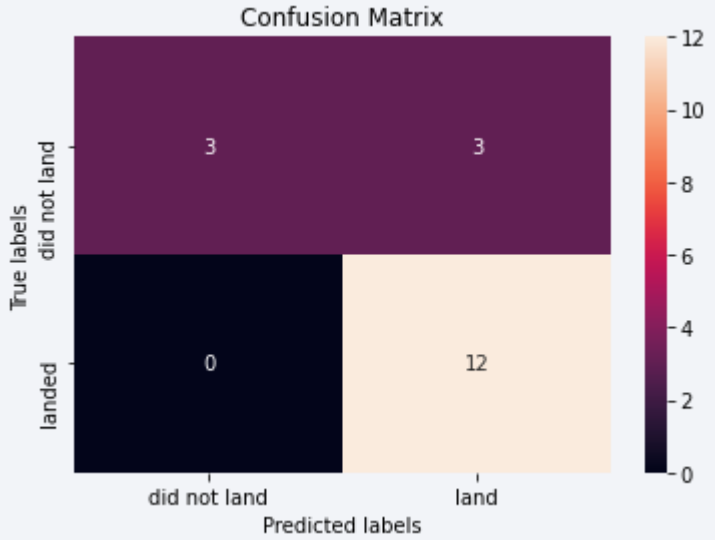
Classification Accuracy



- The accuracy on the validation data for the four models are the same, while the accuracy on the test data shows highest for the decision tree model.
- Since the dataset is too small, this relative high accuracy on the validation data is not convincing enough. Therefore, the four models have similar accuracy.

Confusion Matrix

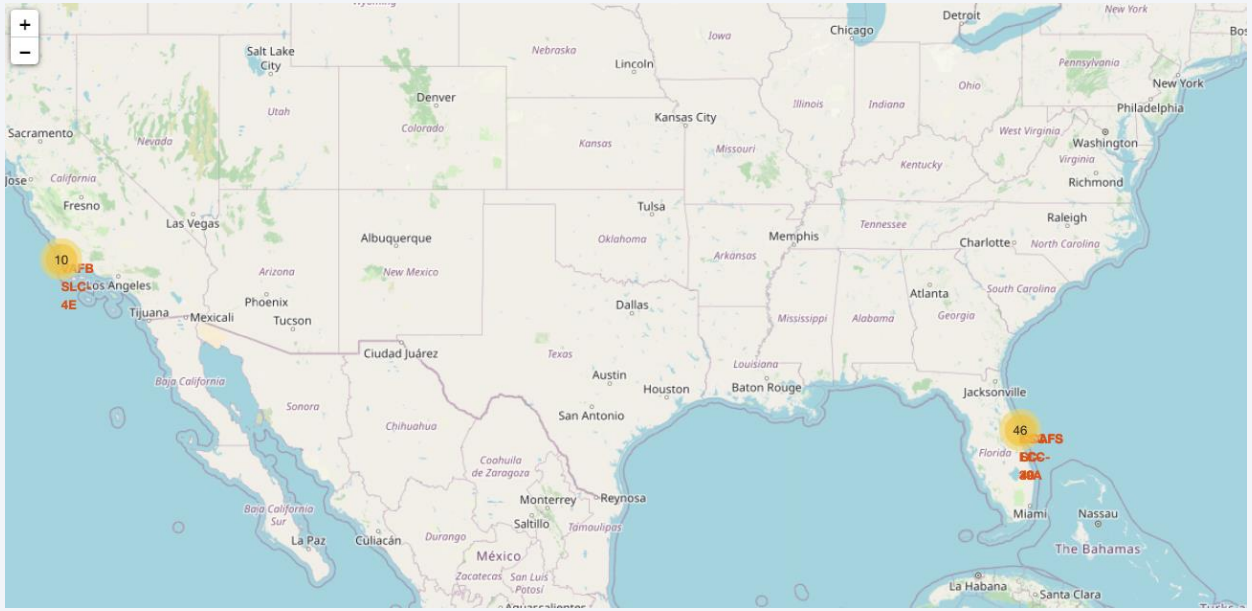
- All four models have the same confusion matrix where the true positive is 12, true negative is 3, false positive is 3 and false negative is 0. The major problem is false positive.



Conclusions

- flight number, payload mass, orbit type are related to the outcome.
- The launch sites were close to coastline, highway and railway, but far from cities.
- Site KSC LC-39A has the highest success rate and CCAFS SLC-40 has the least success launches.
- A smaller payload (e.g., 0-6000) has the largest success rate. FT booster also has the largest success rate.
- The machine learning process showed that since the dataset is limited, all the modals showed the same accuracy.

Appendix



- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!