Winning Space Race with Data Science

Billy

8 November 2022

Outline

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

Executive Summary

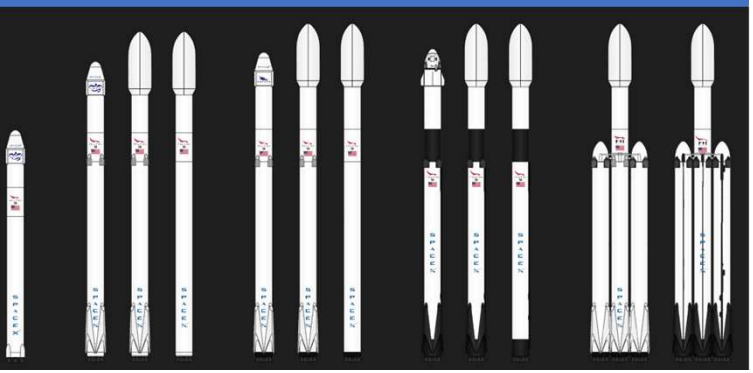
- Summary of methodologies

Create a machine learning pipeline to predict if the first stage will land given the data provided from a certain API or through data scraping. Falcon 9 rocket launches data will be analyzed to gain insights for the trend in success due to certain rocket operation variable.

- Summary of all results

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

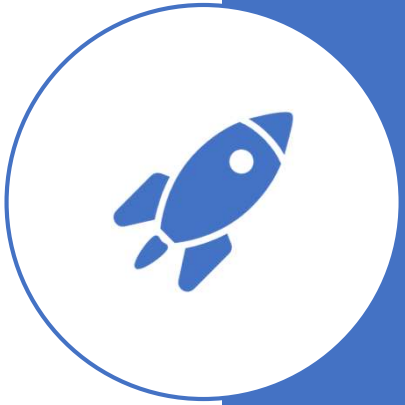
Introduction – Project background and context



The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is Space X. Space X's accomplishments include Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. 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. Spaces X's Falcon 9 launch like regular rockets. To help us understand the scale of the Falcon 9, we are going to use these diagrams from Forest Katsch, at zlsadesign.com. He is a 3D artist and software engineer. He makes infographics on spaceflight and spacecraft art. He also makes software. The payload is enclosed in the fairings. Stage two, or the second stage, helps bring the payload to orbit, but most of the work is done by the first stage. The first stage is shown here. This stage does most of the work and is much larger than the second stage. Here we see the first stage next to a person and several other landmarks. This stage is quite large and expensive. Unlike other rocket providers, Space X's Falcon 9 Can recover the first stage. Sometimes the first stage does not land. Sometimes it will. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

Introduction – Problems to answer

Determine the **trend** of space X to gain insight for the decision to **price** the commercial space travel by predicting if the landing will be **successful** or not based on a couple of independent variable.



Section 1

Methodology

Methodology

Executive Summary

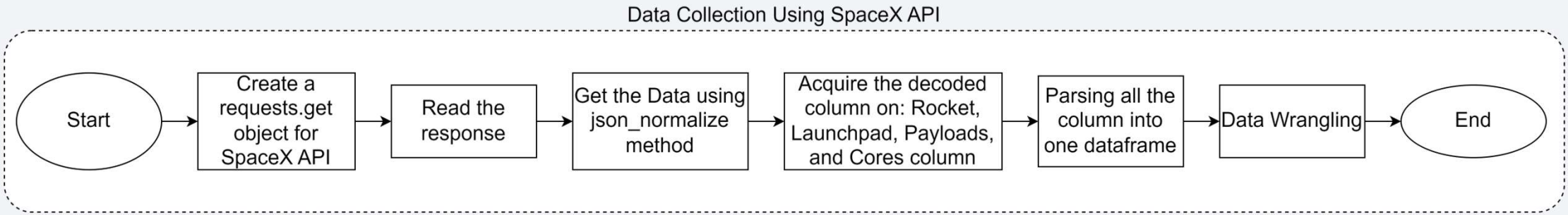
- Data collection methodology:
  - By using request from API
  - By using web scraping
- Perform data wrangling
  - Simple data analysis: value count, checking for nulls, determining data types
  - Create a landing outcome label from Outcome column as feature to predict
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using grid search cross-validation method on various classification machine learning method
  - Analyzing each confusion matrix
  - Determine which models to use based on the test score accuracy

Data Collection

Data collection process can be divided into two categories:

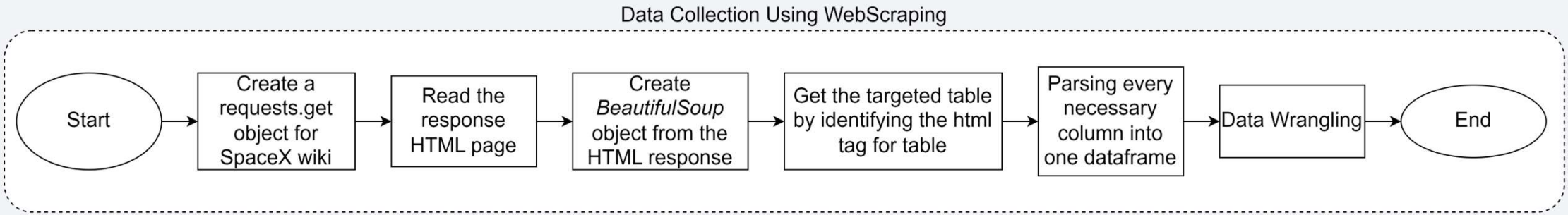
- Collection using requests.get from SpaceX API method,
- And web scraping using BeautifulSoup.

Data Collection – SpaceX API



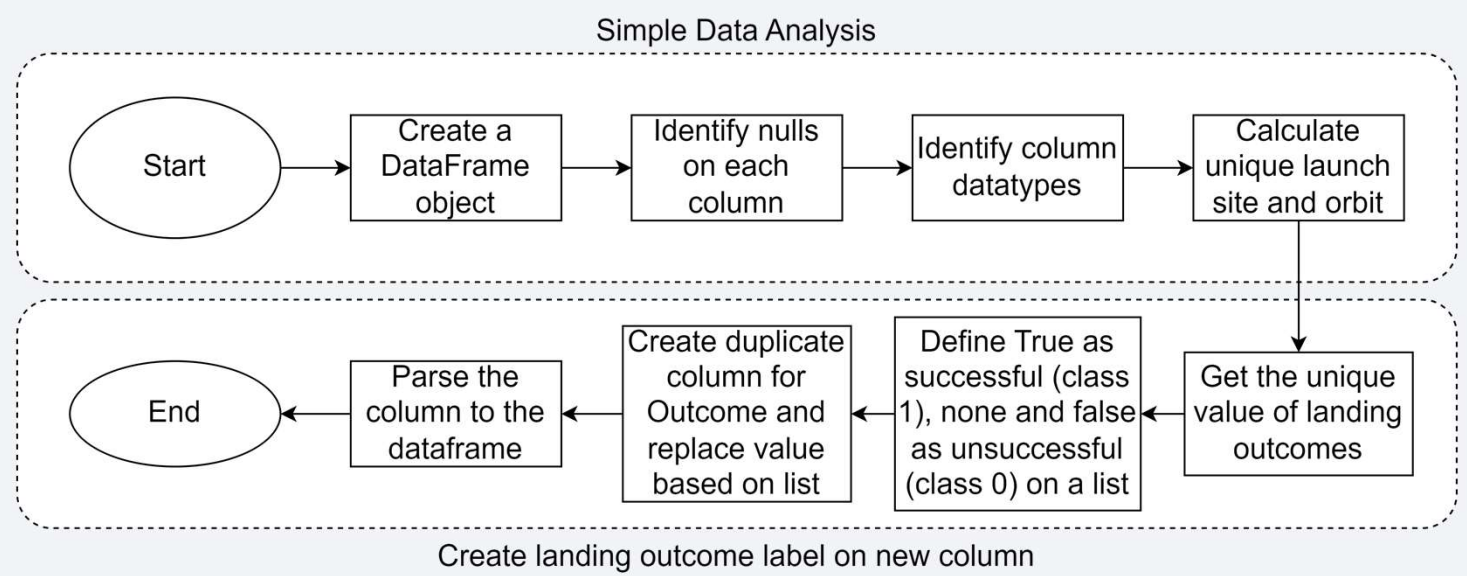
IBM-Data-Science-Applied-Data-Science-Capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · bil-lekid/IBM-Data-Science-Applied-Data-Science-Capstone (github.com)

Data Collection – Web scraping in wikipedia



IBM-Data-Science-Applied-Data-Science-Capstone/jupyter-labs-webscraping.ipynb at main · bil-lekid/IBM-Data-Science-Applied-Data-Science-Capstone (github.com)

Data Wrangling



[IBM-Data-Science-Applied-Data-Science-Capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · bil-lekid/IBM-Data-Science-Applied-Data-Science-Capstone (github.com)]

EDA with Data Visualization

Scatter Plot

- To analyze the relationship between 2 independent variable towards the success rate. Effectively visualizing the relationship between 3 variable using points and color.

Bar Chart

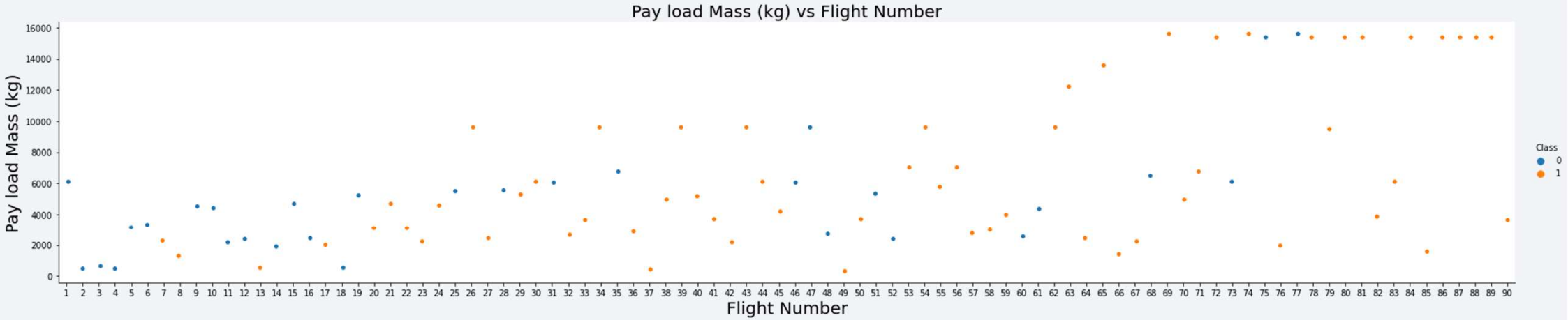
- Visualize success rate based on the orbit type

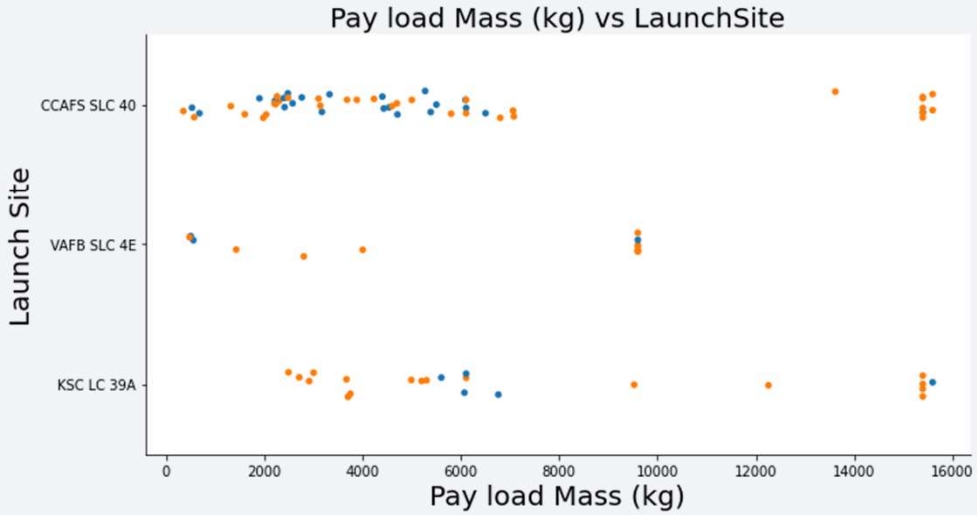
Line Chart

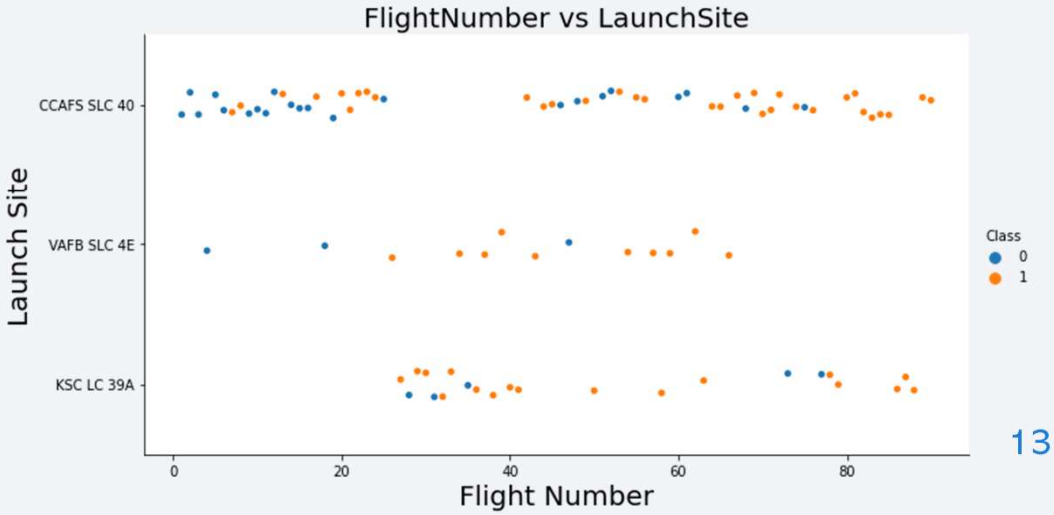
- Visualize success rate growth over the year from 2010 to 2020

IBM-Data-Science-Applied-Data-Science-Capstone/jupyter-labs-eda-dataviz.ipynb at main · bil-lekid/IBM-Data-Science-Applied-Data-Science-Capstone (github.com)

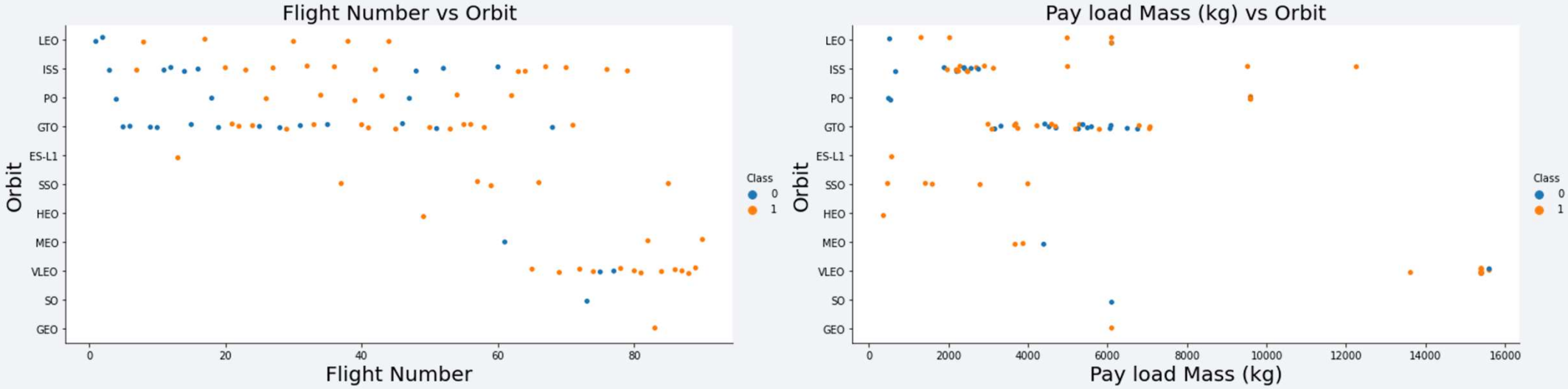
EDA with Data Visualization – Scatter Plot



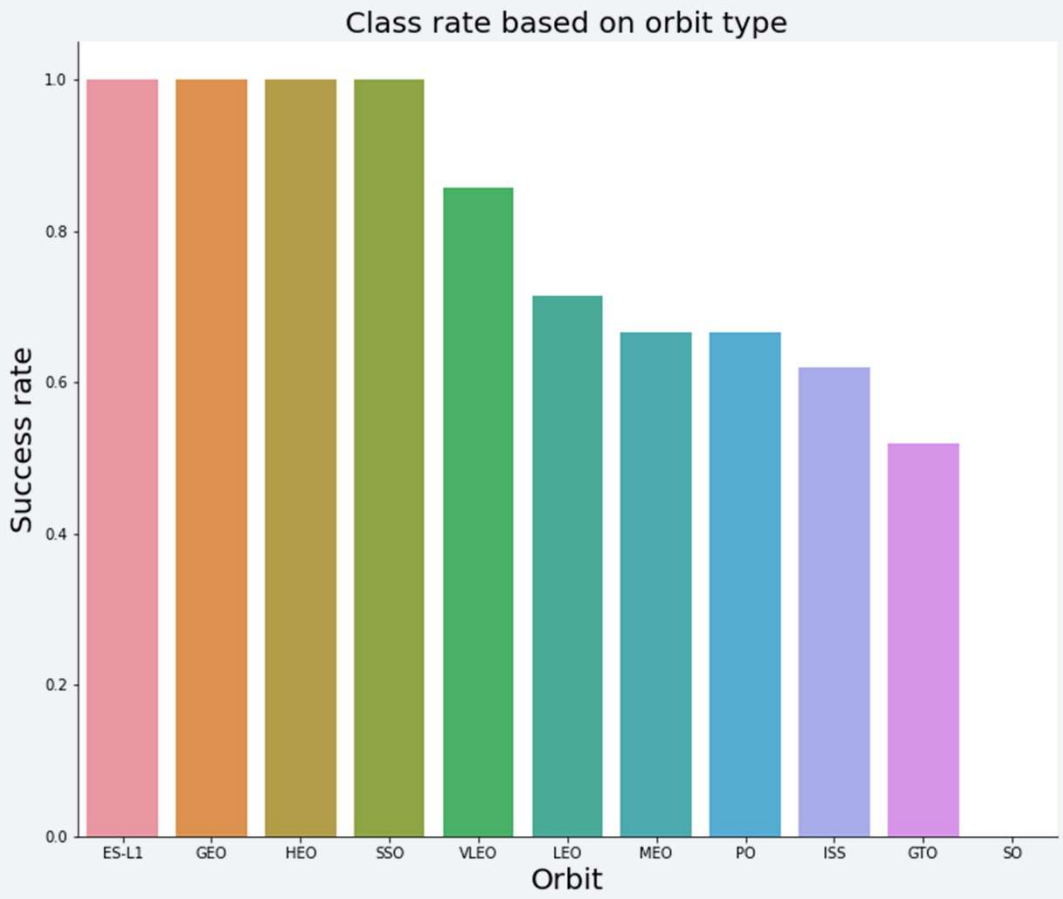




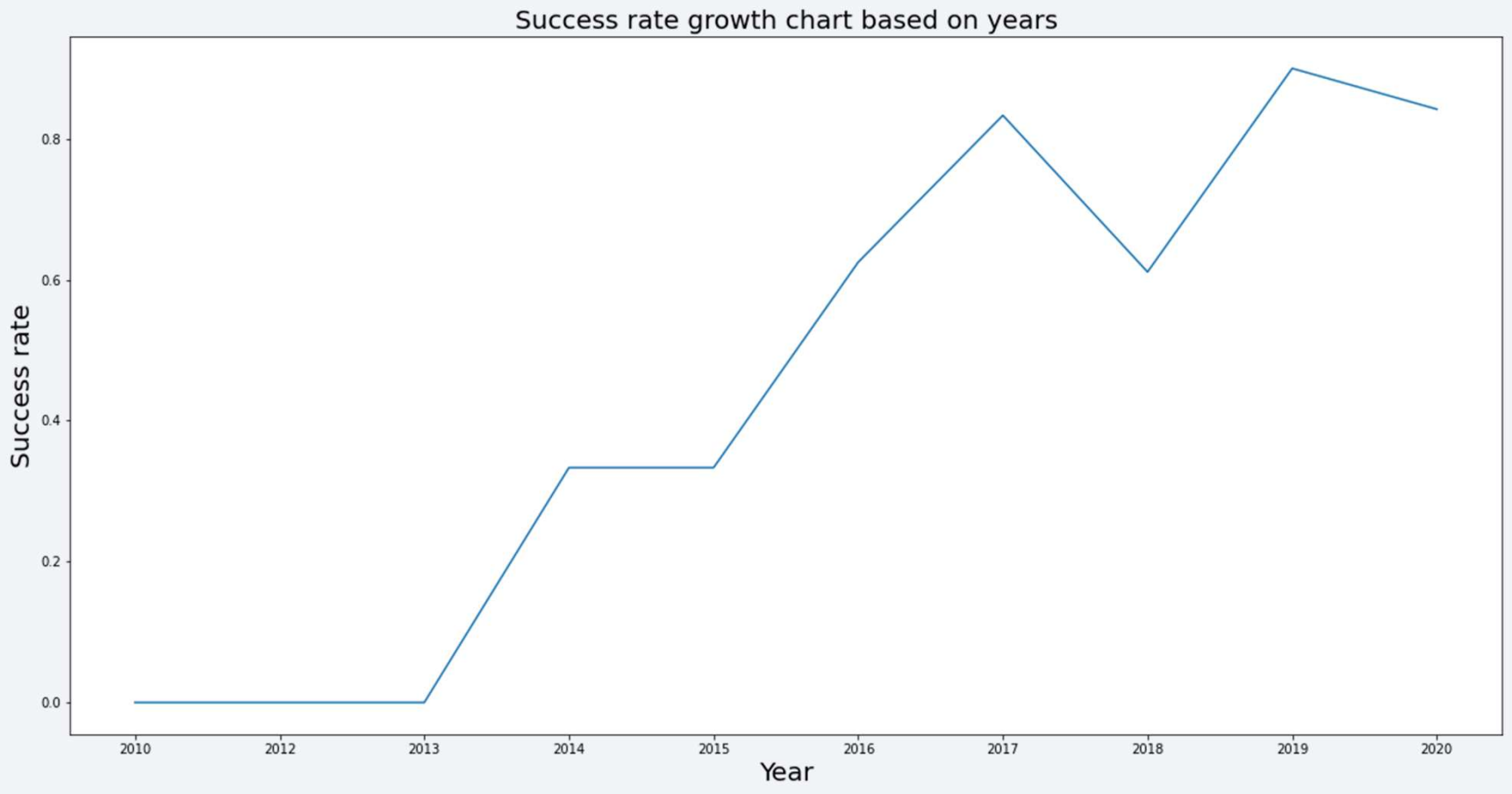
EDA with Data Visualization – Scatter Plot Continuation



EDA with Data Visualization – Bar Chart



EDA with Data Visualization – Line Chart

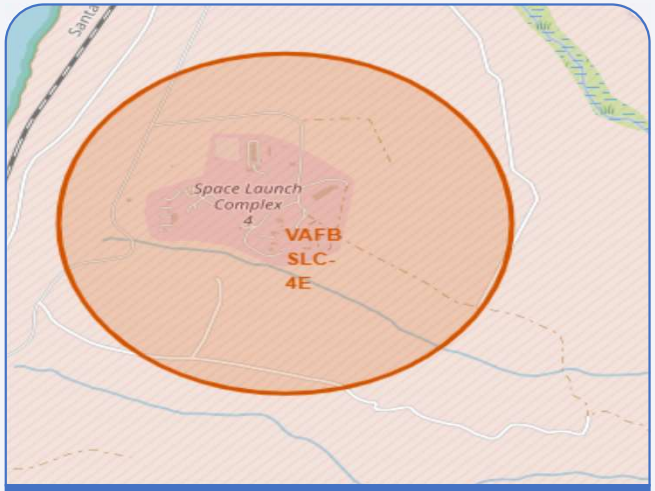


EDA with SQL

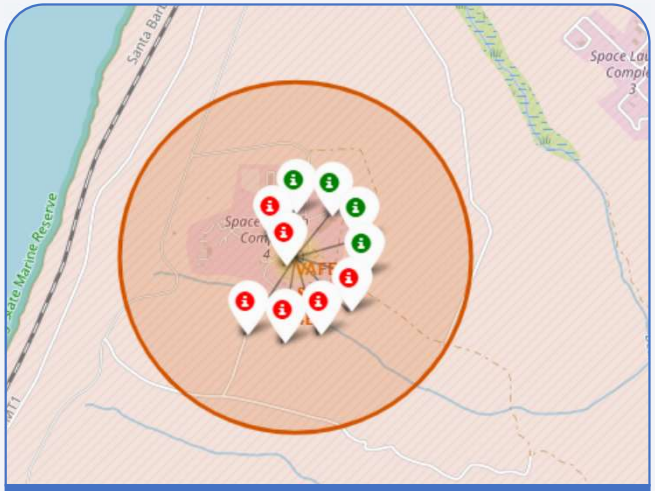
- Display names of unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was achieved.
- 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
- 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

IBM-Data-Science-Applied-Data-Science-Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · bil-lekid/IBM-Data-Science-Applied-Data-Science-Capstone (github.com)

Build an Interactive Map with Folium



Add Circle and Marker to mark launch sites



Add marker cluster to show the total of successful landing and unsuccessful landing (green = successful, red = unsuccessful)

Add line with popup and marker to mark distance to closest city, coast, railroad, and highway.

[IBM-Data-Science-Applied-Data-Science-Capstone/lab_jupyter_launch_site_location.ipynb at main · bil-lekid/IBM-Data-Science-Applied-Data-Science-Capstone (github.com)]

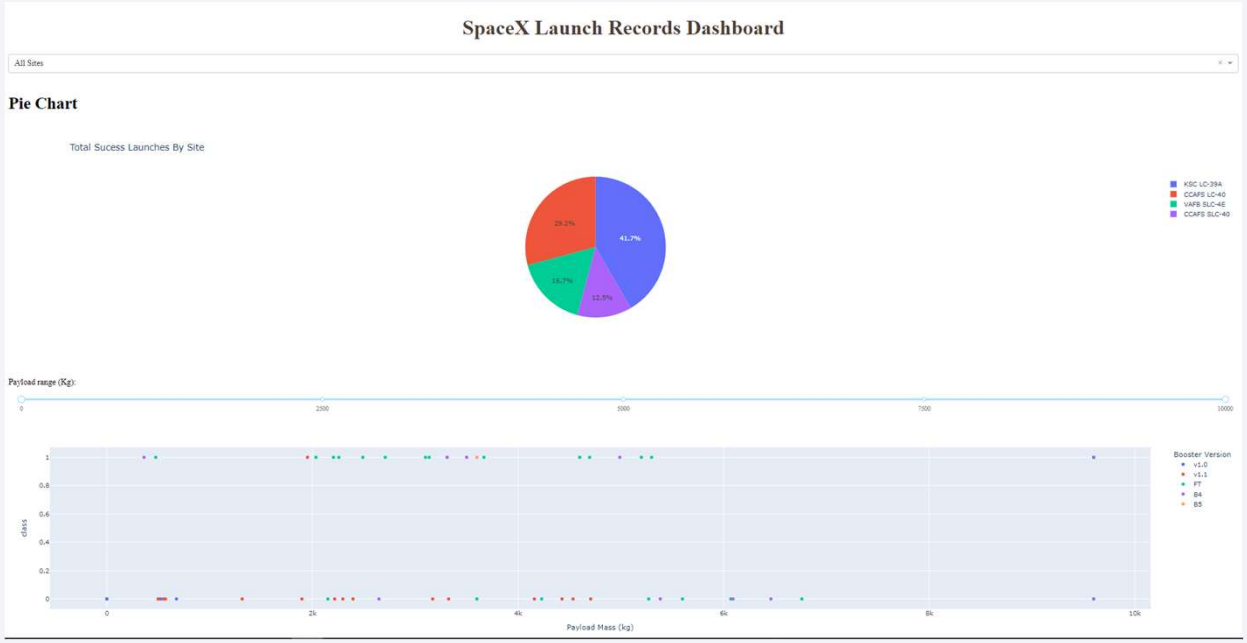
Build a Dashboard with Plotly Dash

Feature

- Launch site selection using dropdown
- Payload range (Kg) selection using slider

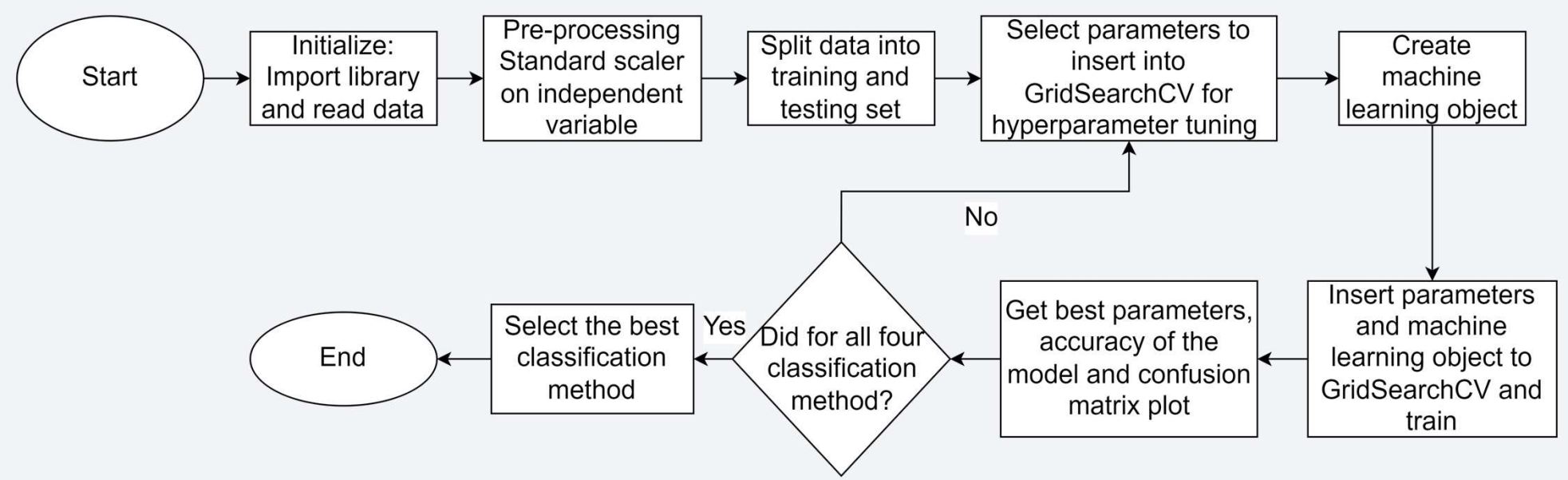
Chart

- Total success launches by site using pie chart
- Payload Mass (Kg) vs success rate with launch site color labeled using scatter plot

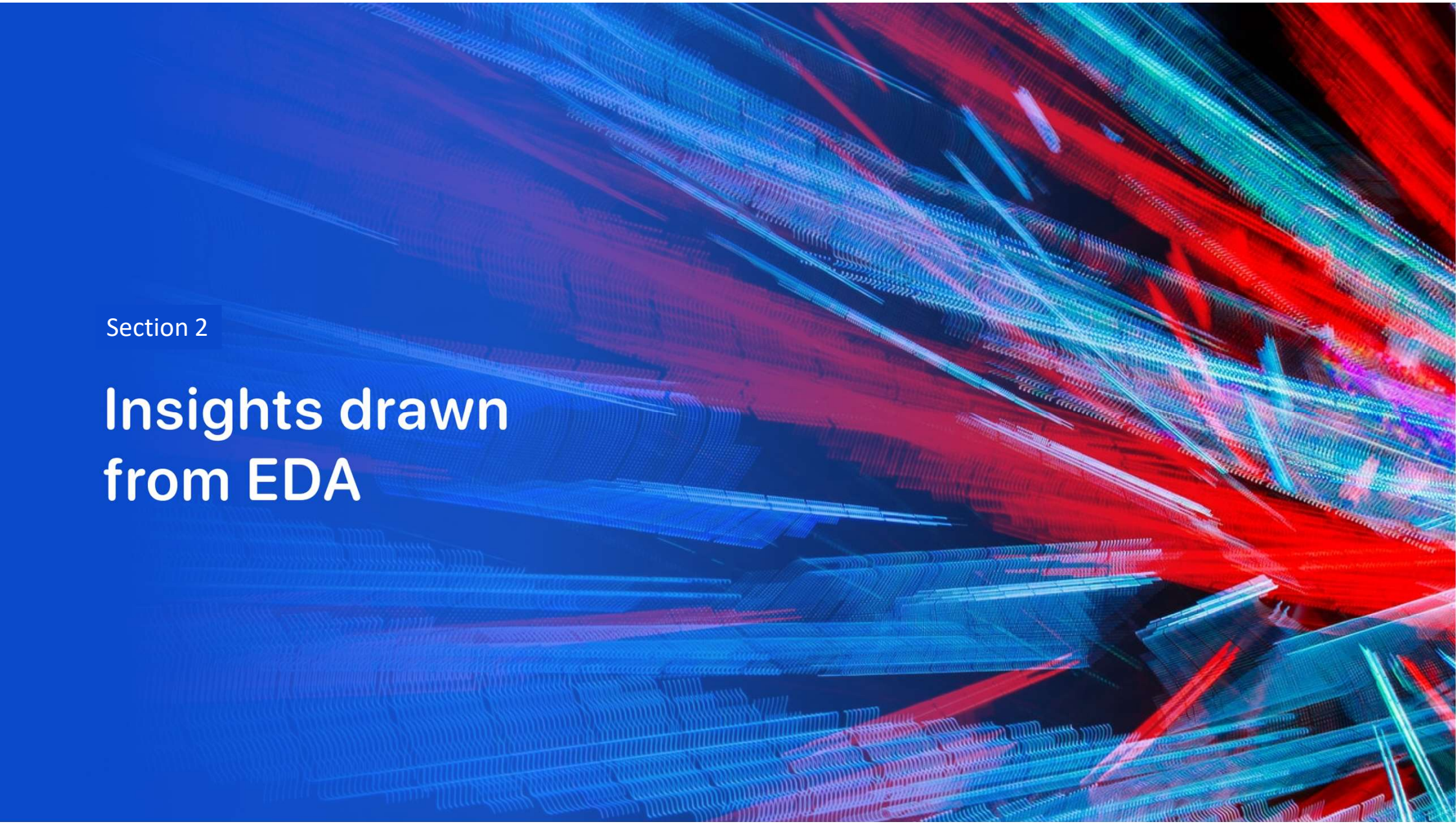


[IBM-Data-Science-Applied-Data-Science-Capstone/dashboard.py at main · bil-lekid/IBM-Data-Science-Applied-Data-Science-Capstone (github.com)]

Predictive Analysis (Classification)

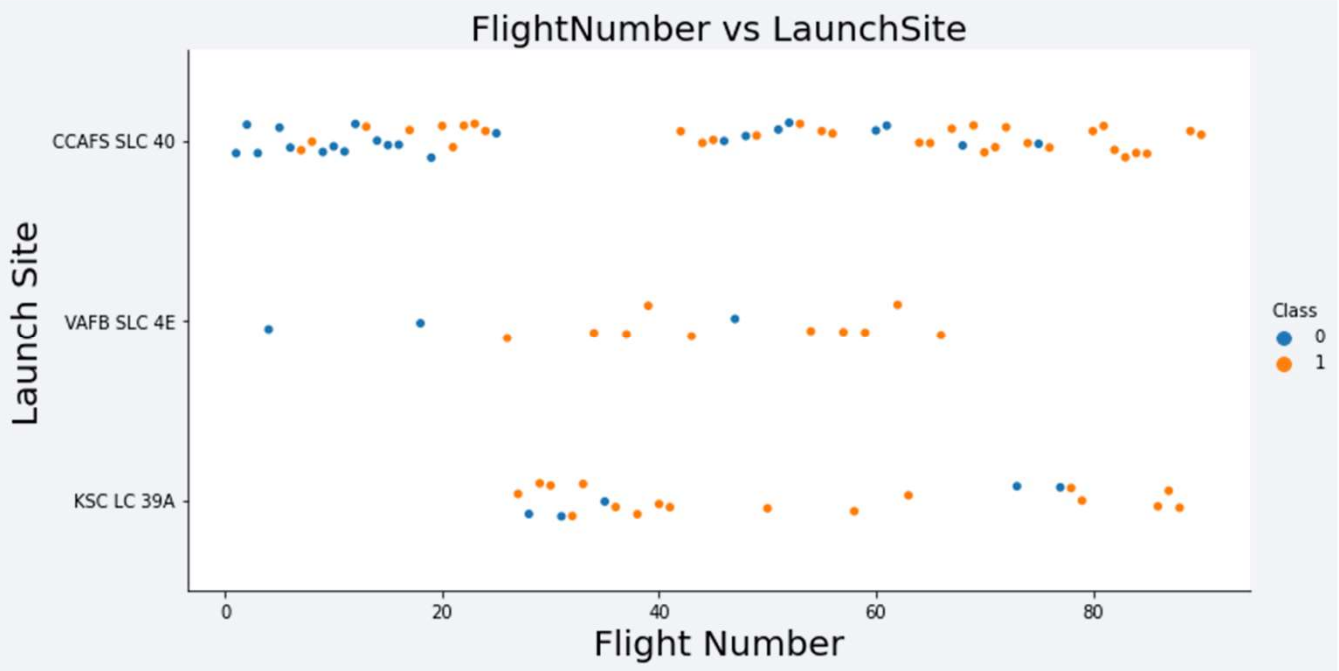


IBM-Data-Science-Applied-Data-Science-Capstone/dashboard.py at main · bil-lekid/IBM-Data-Science-Applied-Data-Science-Capstone (github.com)


Section 2
Insights drawn
from EDA

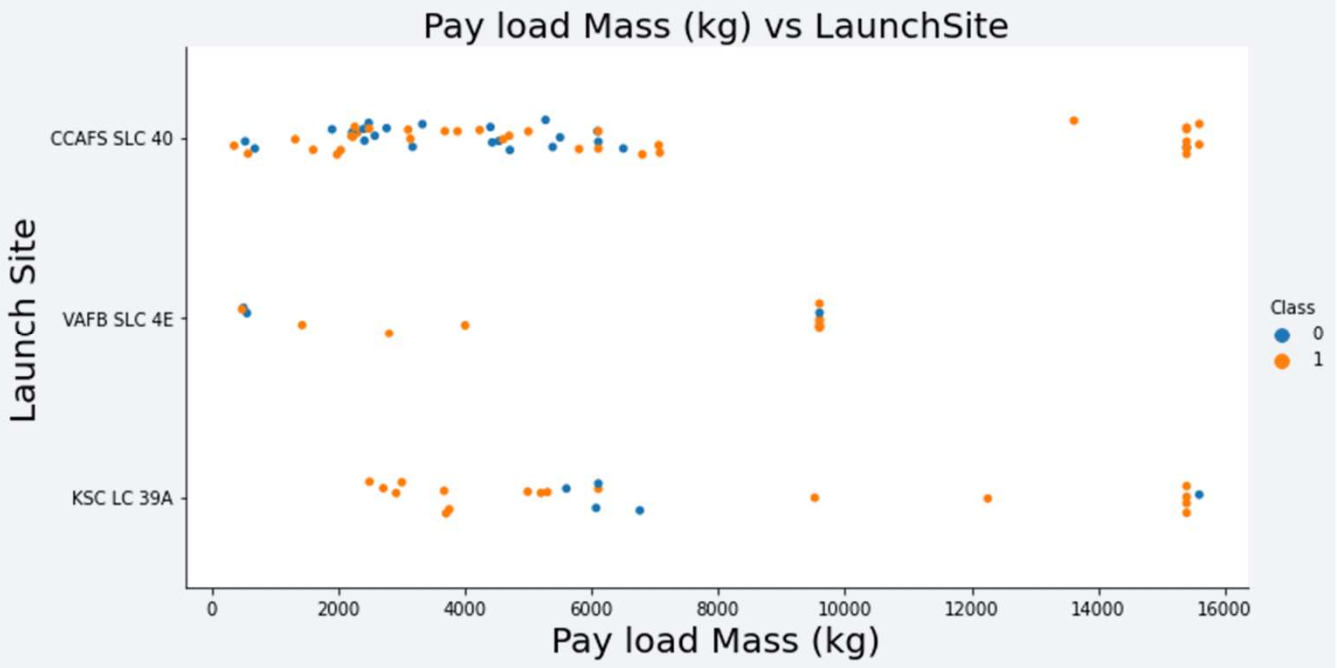
Flight Number vs. Launch Site

- For all the launch site as the flight number goes up, there are less unsuccessful landing.



Payload vs. Launch Site

- CCAFS SLC 40 launch site have a cluster of successful and unsuccessful landing in the range of 0 to 8000 kg payload mass, but at the range of 13000 to 16000 kg the success rate is 100 %
- For the other launch site there seems to be a random cluster of unsuccessful landing at different point. But as the payload mass goes up drastically, the success rate goes up.
- The sudden increase in payload mass like outlier might have been intentional since they finally found an optimal payload mass.

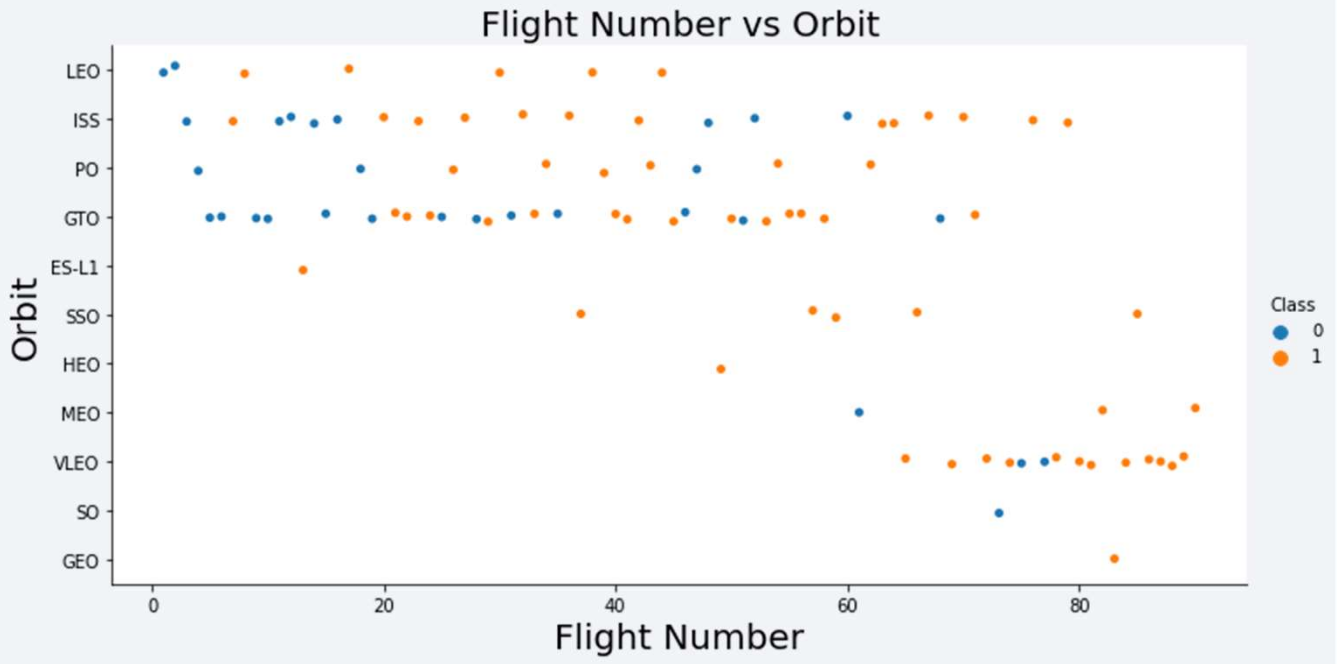


Success Rate vs. Orbit Type

- There have been 100 % success rate on ES-L1, GEO, HEO, and SSO orbit.
- Other 6 orbits had success rate in the range of 0.85 to around 0.5.
- SO is the only orbit that had 0 % success rate.

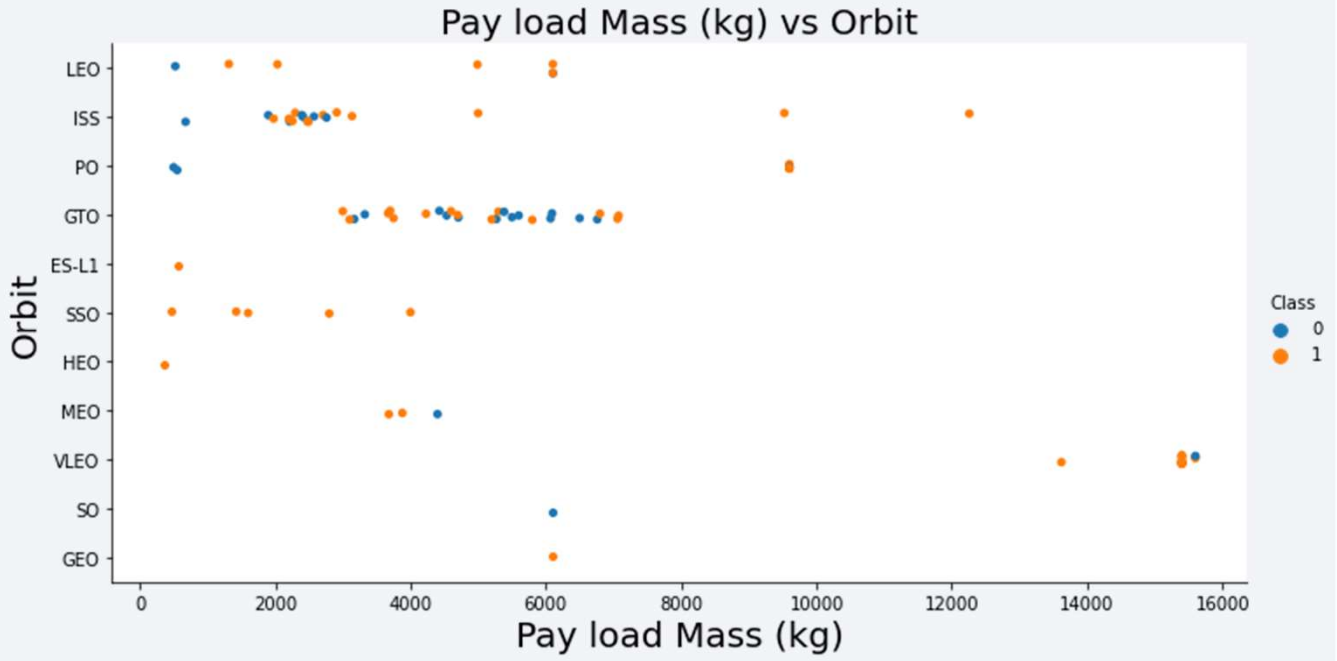
Flight Number vs. Orbit Type

- We can see SO orbit only had 1 attempt and it's unsuccessful, no wonder the success rate is 0 %.
- ES-L1, GEO, HEO, and SSO, have relatively low number compared to some orbits like GTO or ISS. To be exact ES-L1, GEO, and HEO only had 1 attempt and it is successful
- There seems to be a random cluster or points of unsuccessful mark placed in random places of the chart.
- Flight number and orbit doesn't have any correlation.



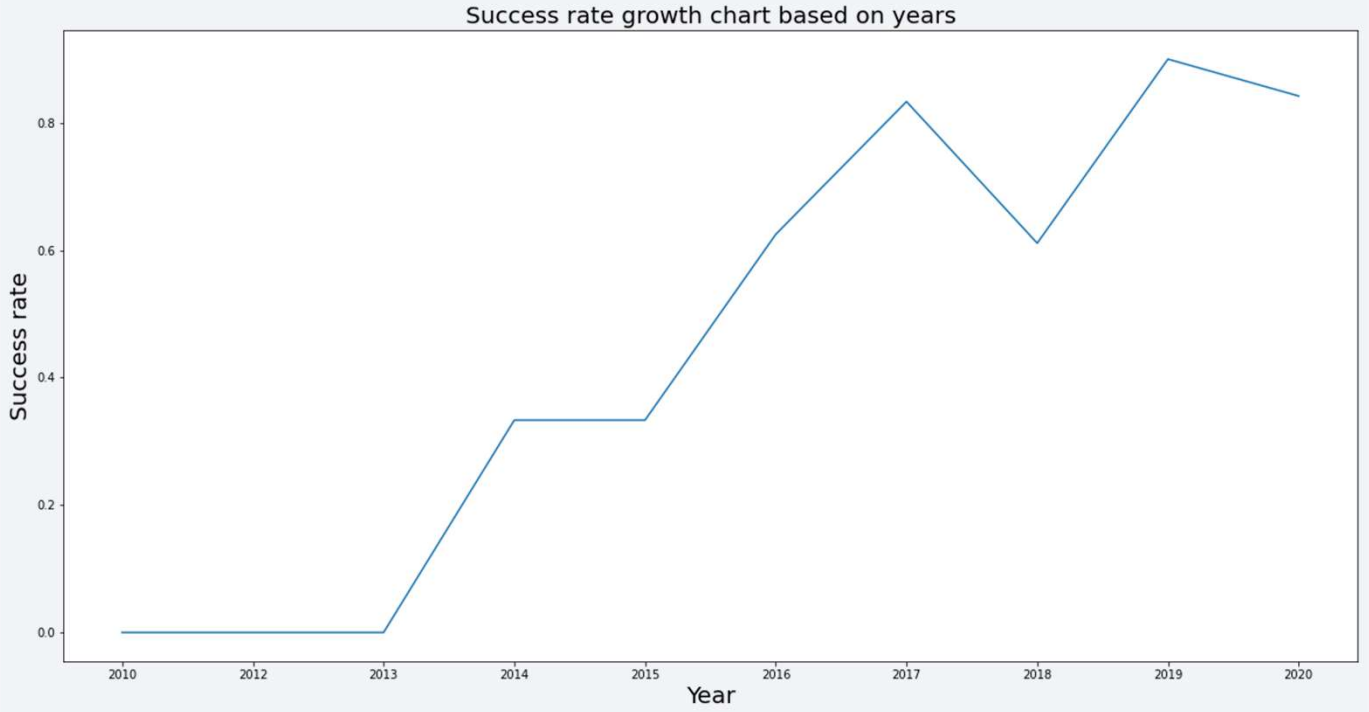
Payload vs. Orbit Type

- Again, there seems to be no connection. As soon as the payload surpasses 12000 kg, there seems to be a high success rate.



Launch Success Yearly Trend

- There are a positive correlation between the success rate as the year goes by. It reached around 90% success rate in 2019.



All Launch Site Names

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

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

```
%%sql
SELECT DISTINCT(Launch_Site) AS unique_launch_site
FROM SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

CCAFS LC-40

CCAFS SLC-40

Display 5 records where launch sites begin with the string '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

45596

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

2928.4

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"
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

01-05-2017

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)"
```

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

MIN("Date")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

F9 v1.1

F9 v1.1 B1011

F9 v1.1 B1014

F9 v1.1 B1016

F9 FT B1020

⋮

F9 B5B1062.1

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 PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2

Total Number of Successful and Failure Mission Outcomes

Success = 100

Failure = 1

List the total number of successful and failure mission outcomes

```
%%sql
SELECT DISTINCT(SELECT COUNT(Mission_Outcome)
        FROM SPACEXTBL
        WHERE Mission_Outcome LIKE "%Success%") AS Success,
        (SELECT COUNT(Mission_Outcome)
        FROM SPACEXTBL
        WHERE Mission_Outcome LIKE "%Failure%") AS Failure
FROM SPACEXTBL
```

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

Success	Failure
100	1

Boosters Carried Maximum Payload

- 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

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

+ Code + Markdown

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

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

01

02

03

04

04

06

12

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.

+ Code + Markdown

```
%%sql
SELECT SUBSTR(Date,4,2)
FROM SPACEXTBL
WHERE SUBSTR(Date,7,4)="2015"
```

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

SUBSTR(Date,4,2)
01
02
03
04
04
06
12

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

- Success (ground pad)
- No attempt
- Success
- Success
- Success (ground pad)
- Success (drone ship)
- Controlled (ocean)
- Failure
- Success
- Failure
- Failure (drone ship)
- Success
- ⋮
- Failure (parachute)

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

+ Code + Markdown

```
%%sql
SELECT "Landing _Outcome"
FROM SPACEXTBL
WHERE Date BETWEEN "04-06-2010" AND "20-03-2017"
ORDER BY Date DESC
```

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

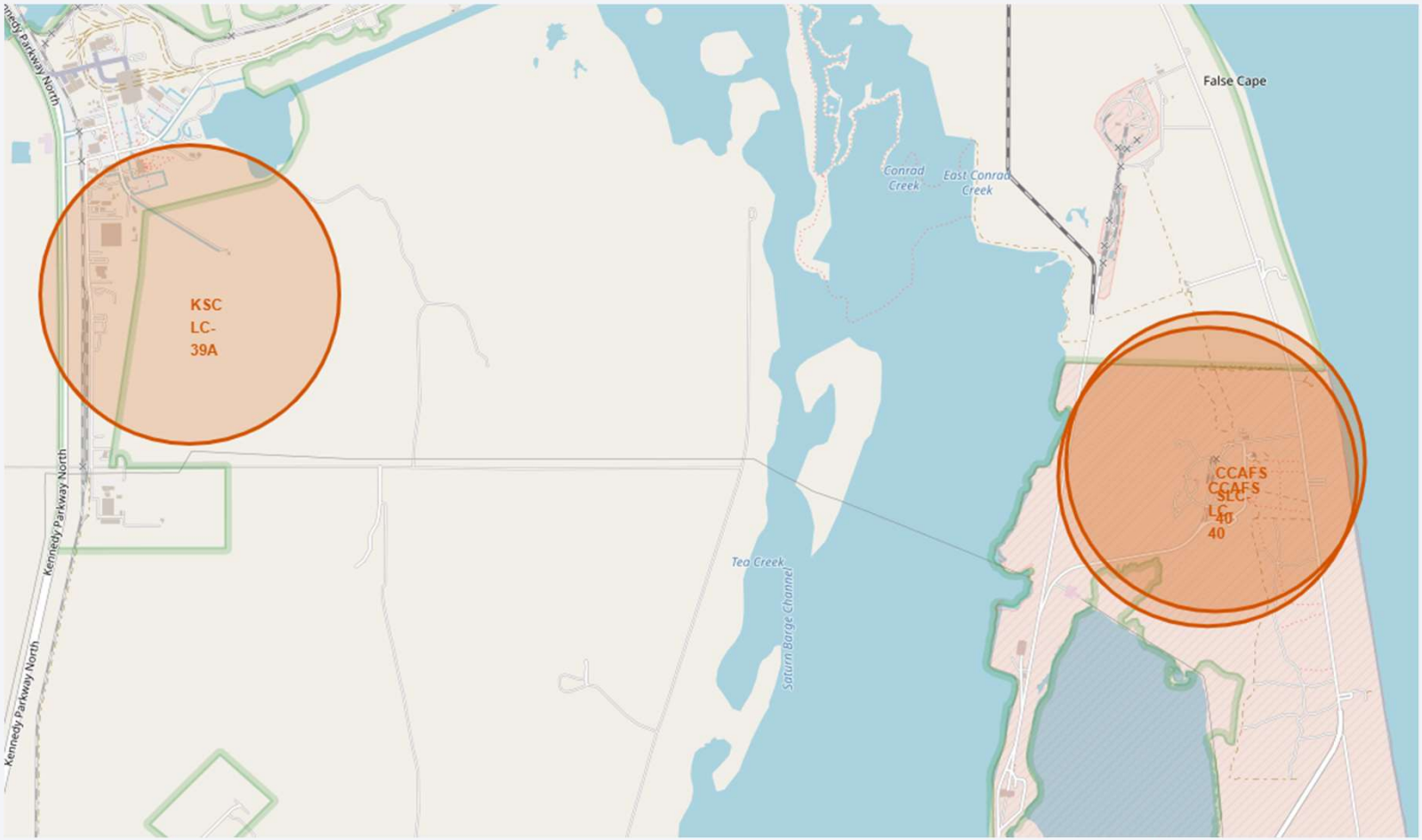
Landing _Outcome
Success (ground pad)
No attempt
Success
Success
Success (ground pad)
Success (drone ship)
Controlled (ocean)
Failure
Success
Failure
Failure (drone ship)
Success


Section 3
Launch Sites
Proximities Analysis

Mark all launch site on map

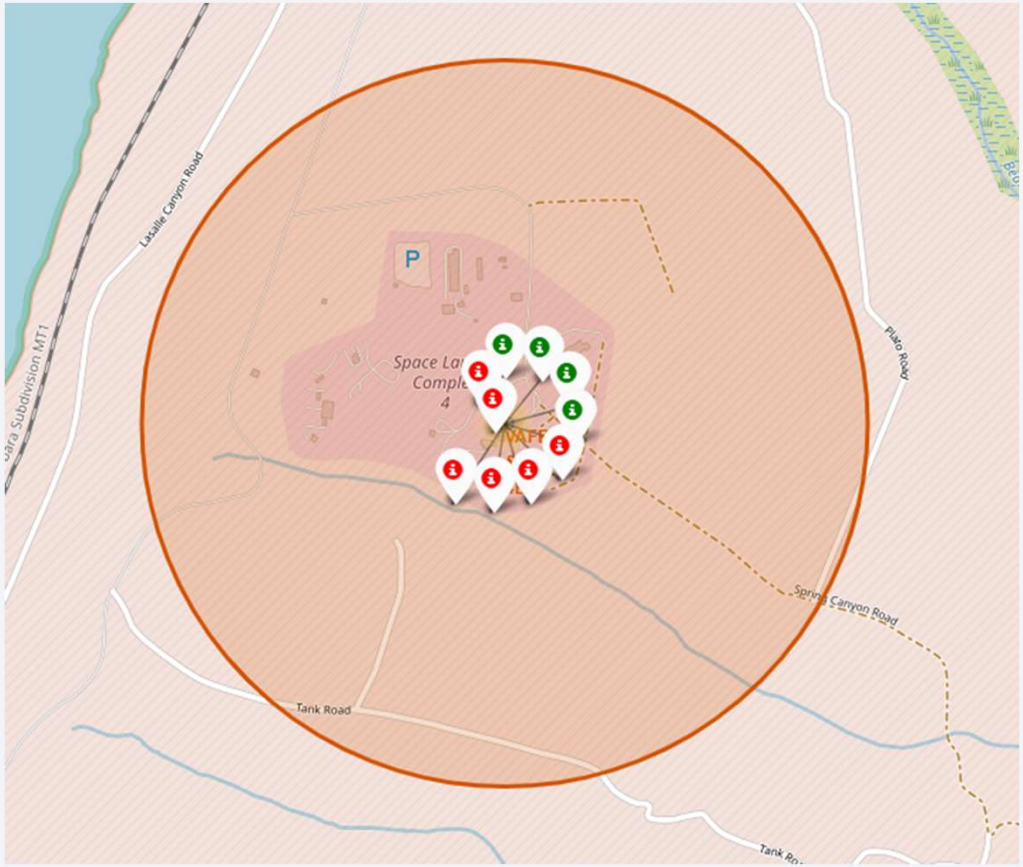
- VAFB SLC-4E Launch site is located at southwest of united states while the rest of the launch sites are located at southeast

Mark all launch site on map (Continuation)

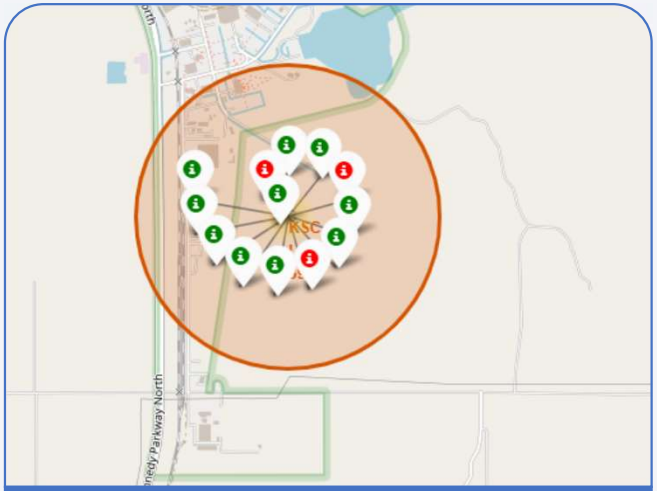


Marking the success/failed launches for each site on the map

- Marker cluster is placed on each launch sites based on the amount of unsuccessful and successful launches. The red marker indicates unsuccessful launches in the launch site while the green marker indicates the opposite.
- VAFB SLC-4E had 10 launches, 4 success and 6 fail

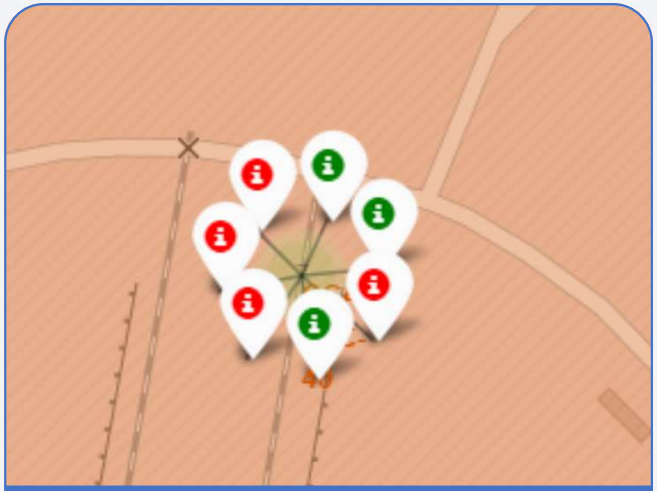


Marking the success/failed launches for each site on the map (Continuation)



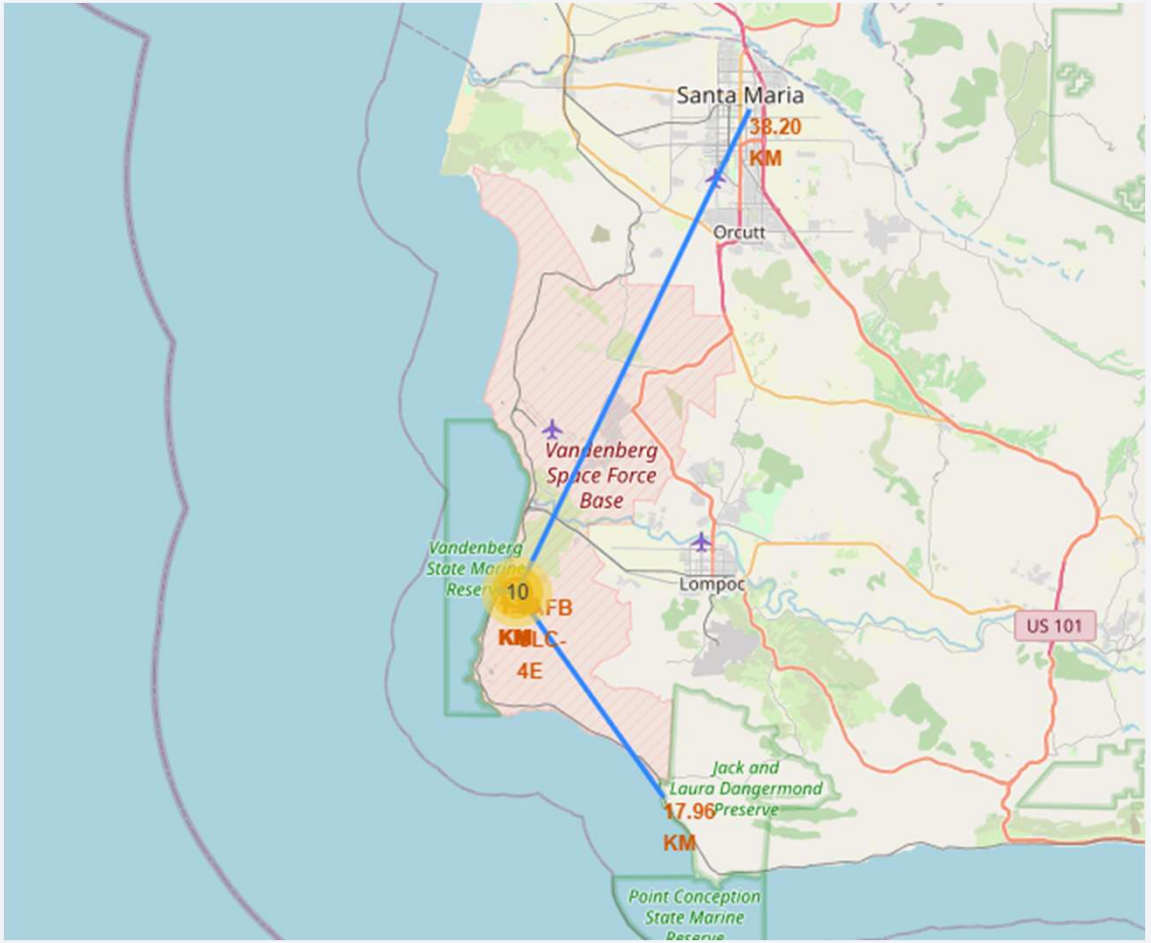
KSC LC-39A had 13 launches, 10 success and 3 failed

CCAFS LC-40 had 26 launches, 7 success and 19 failed

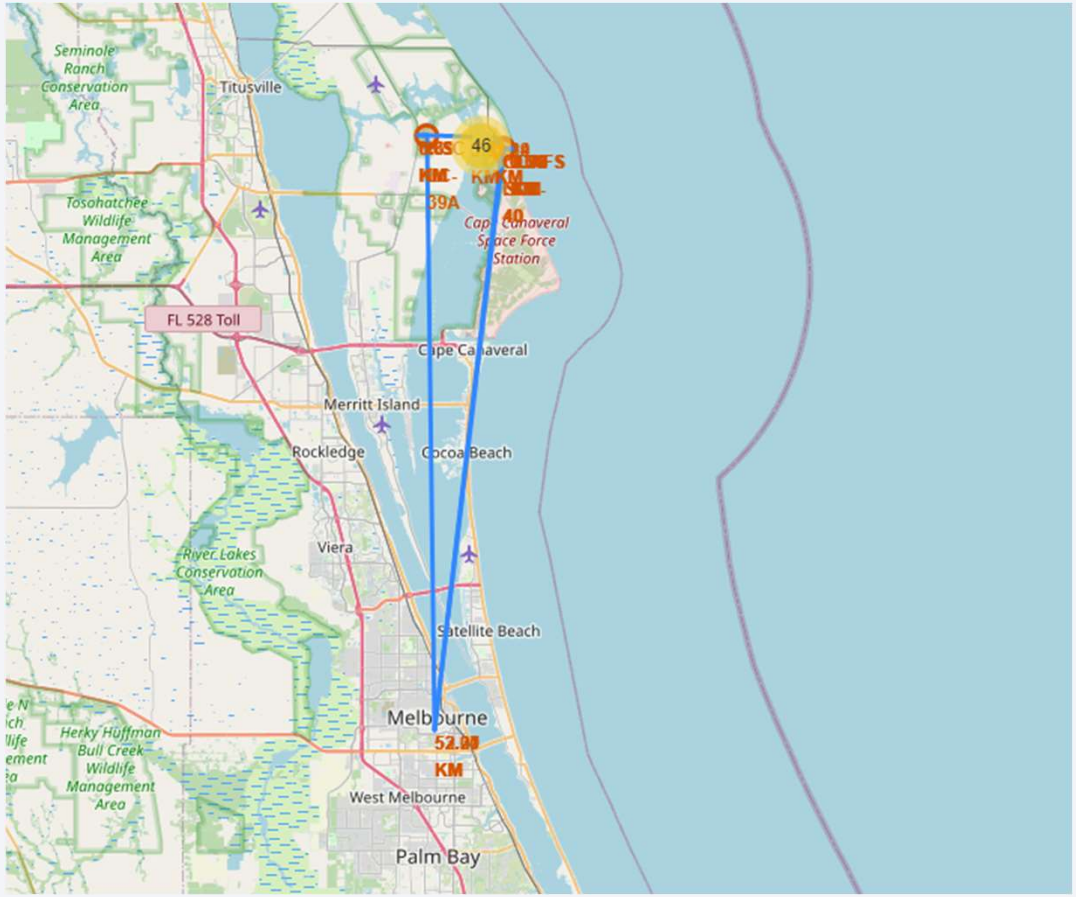


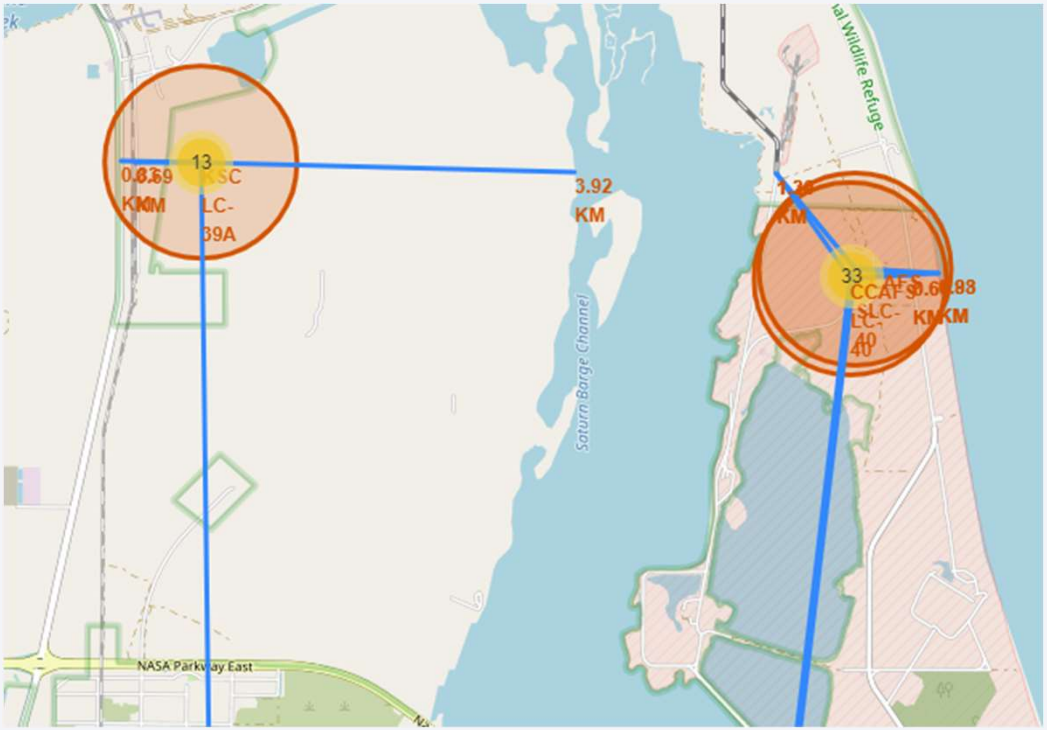
CCAFS SLC-40 had 7 launches, 3 success and 4 failed

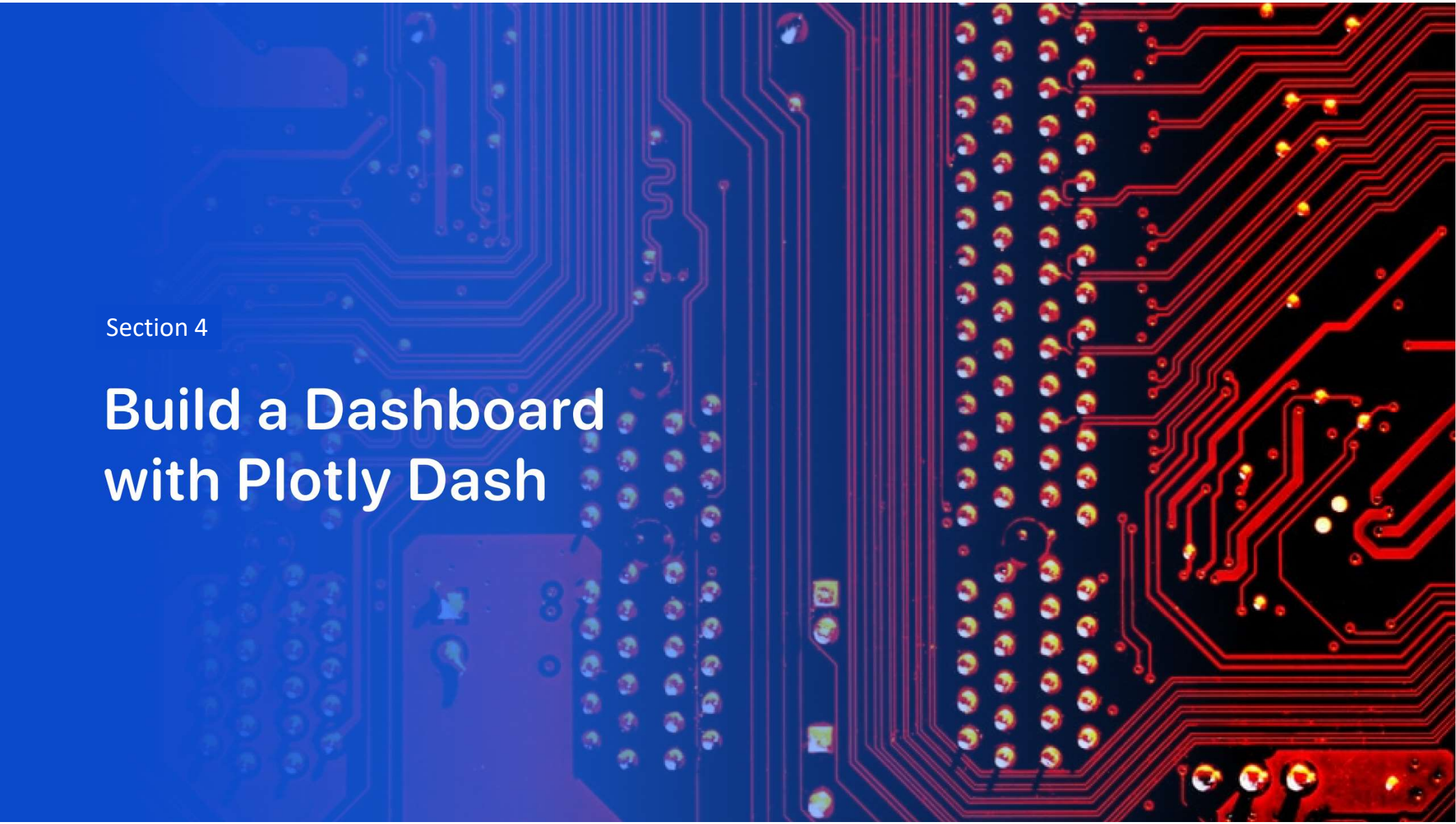
Create a distance marker from the launch site to the nearest city, coast, railway, and highway



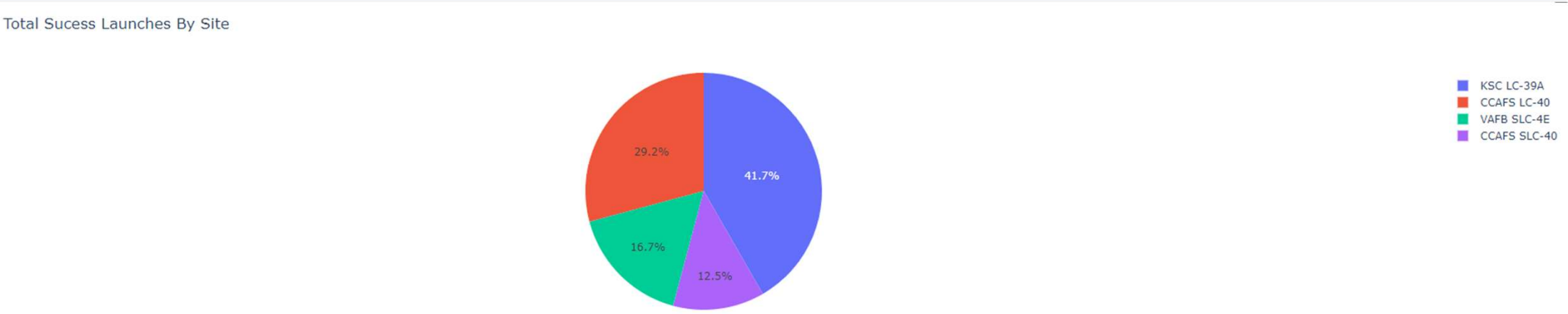
Create a distance marker from the launch site to the nearest city, coast, railway, and highway (Continuation)





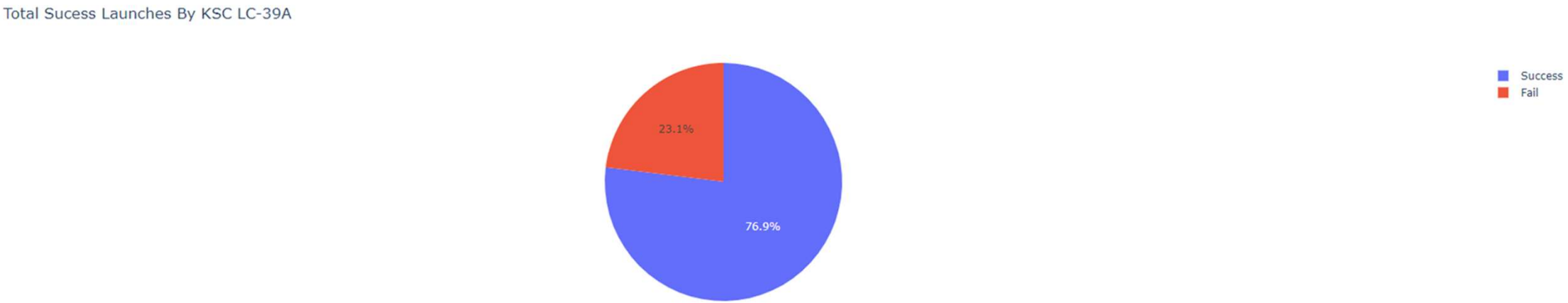

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by All Sites Pie Chart



- KSC LC-39A had the most success launches at 41.7 % of the pie, while CCAFS LC-40 had 29.2 %, VAFB SLC-4E had 16.7%, CCAFS SLC-40 had 12.5%.

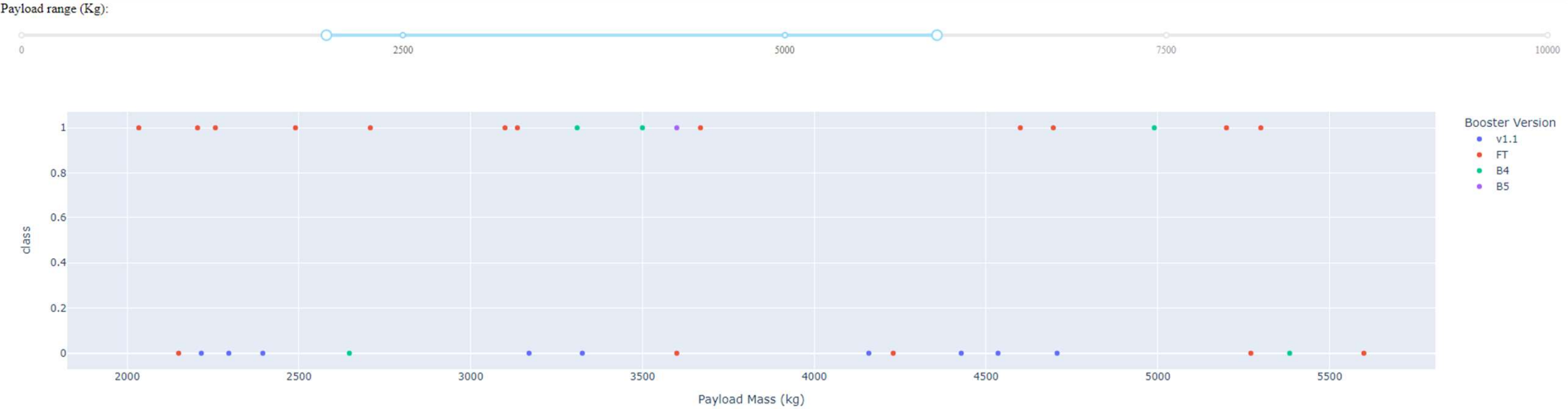
Launch site with the highest launch success ratio Pie Chart



- KSC LC-39A is the only launch site with success launches more than fail launches. The launch success ratio is 76.9 % successful launches.

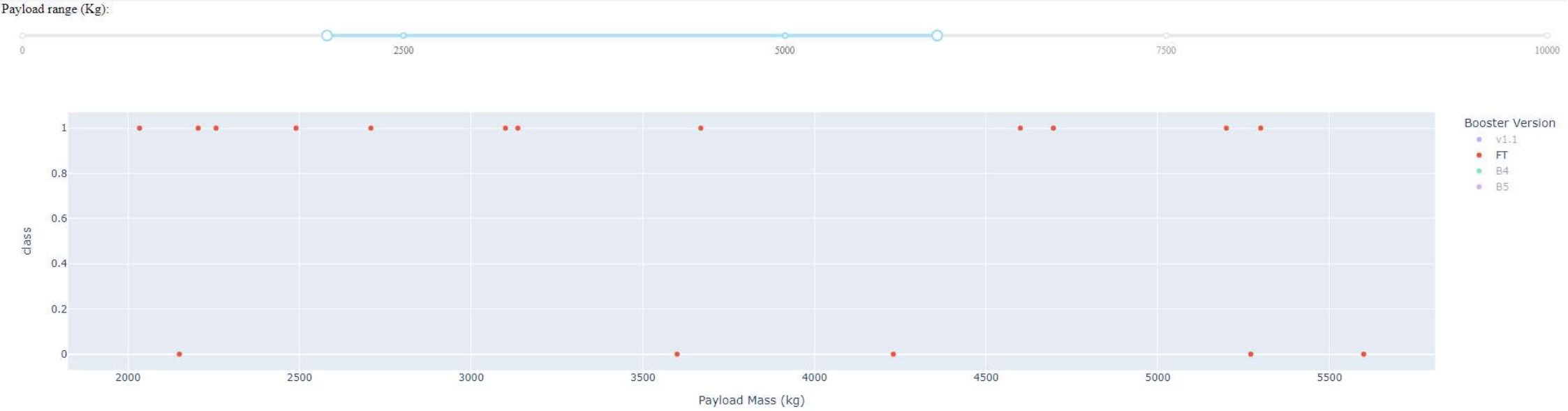
Payload vs Launch Outcome scatter plot

- The best payload range is 2000 to 6000 kg for success rate.



Payload vs Launch Outcome scatter plot (Continuation)

- The best booster version would be FT, excluding B5 booster since it is 100 % success rate but only with one sample

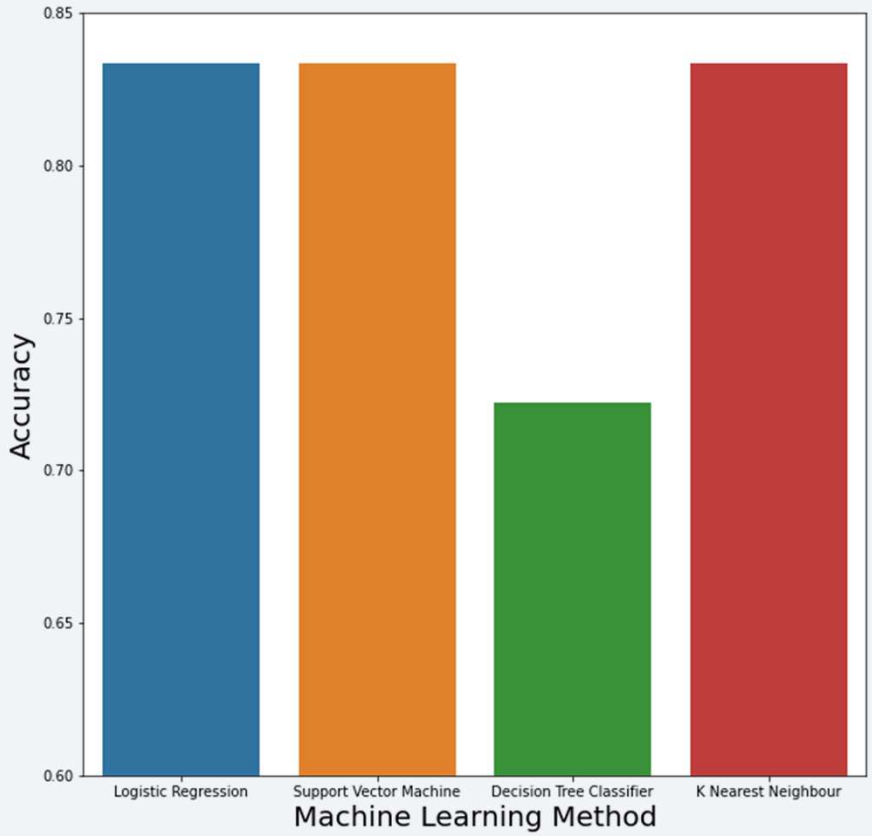


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistic regression, support vector machine, and K nearest neighbor all had the accuracy of 0.8334 after calculating the score using the test data.
- Decision tree classifier had accuracy score of 0.7222



Confusion Matrix

- True Positive (TP) = 12
- True Negative (TN) = 3
- False Positive (FP) = 0
- False Negative (FN) = 3
- The positive prediction had good accuracy since it successfully predicted all right.
- The negative prediction had mediocre and almost bad accuracy since the amount of false negative is equal to the true negative

Conclusions

- Data is collected by two methods, using SpaceX API and web scraping. The collected data will be wrangled before proceeding.
- There have been a relation between payload mass and success rate, which in specific huge payload mass the success based on orbit or launch site is relatively higher.
- There have been uptrend of success rate from 2010 to 2021
- By using SQL, it is possible to perform advanced exploratory data analysis on specific filter such as average of a certain column or data based on dates.
- By using folium and dash, it is possible to create a dashboard and interactive map chart to furthermore simplify data analysis for stakeholder presentation.
- The best classification method on test data based on accuracy is logistic regression, support vector machine, and K nearest neighbor. The models can predict if the launch is going to be successful or not based on 15 independent variable, the models had an accuracy of 0.8334.

Thank you!