IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Franky Leonardo Prieto
2024-Oct-15

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection through SpaceX API and web scraping
  - Data wrangling
  - Exploratory data analysis with SQL
  - Exploratory data analysis with data visualization
  - Interactive visual analytics with folium
  - Predictive analysis (classification with machine learning)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics in screenshots
  - Predictive analytics results

Introduction

- Project background and context

  SpaceX advertises Falcon 9 rocket launches on its website a cost of 62 million dollars; meantime other providers cost over 165 million 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- Problems to find answers

  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - How the rate of successful landings evolve in time?
  - What is the best algorithm to predict this goal?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected through SpaceX API and Web scraping from Wikipedia
- Perform data wrangling
  - Data was processed with pandas filtering, dealing with missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - building, tuning, evaluation of classification models.

Data Collection

- Data collection process involved two principal sources: SpaceX API and Wikipedia scraping. To collect this was used resquest and beautifulsoup python packages
- Data columns obtained from SpaceX REST API:

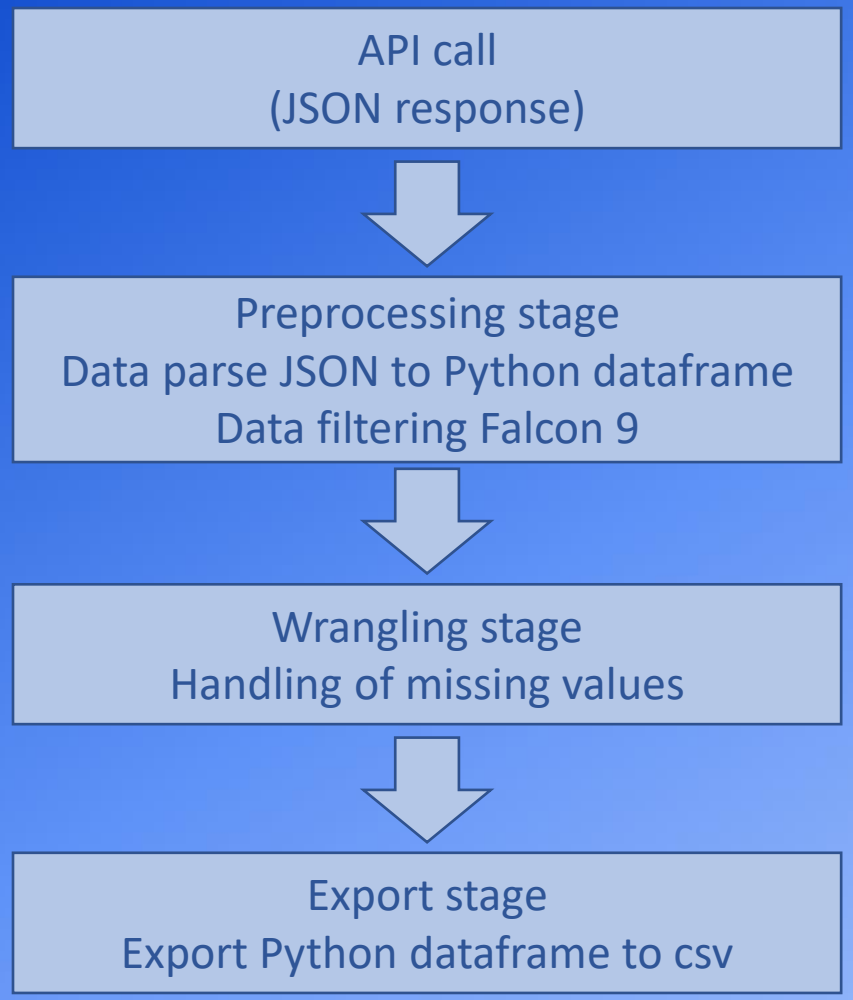
  booster version, payload mass, orbit, launch site, outcome, flights, grid, if is reused, legs, landing pad, block, reused count, serial, longitude, latitude

- Data columns obtained from Wikipedia, using web scrapìng:

  Flight number, lauch site, payload, payload mass, orbit, customer, launch outcome, date, time

Data Collection – SpaceX API

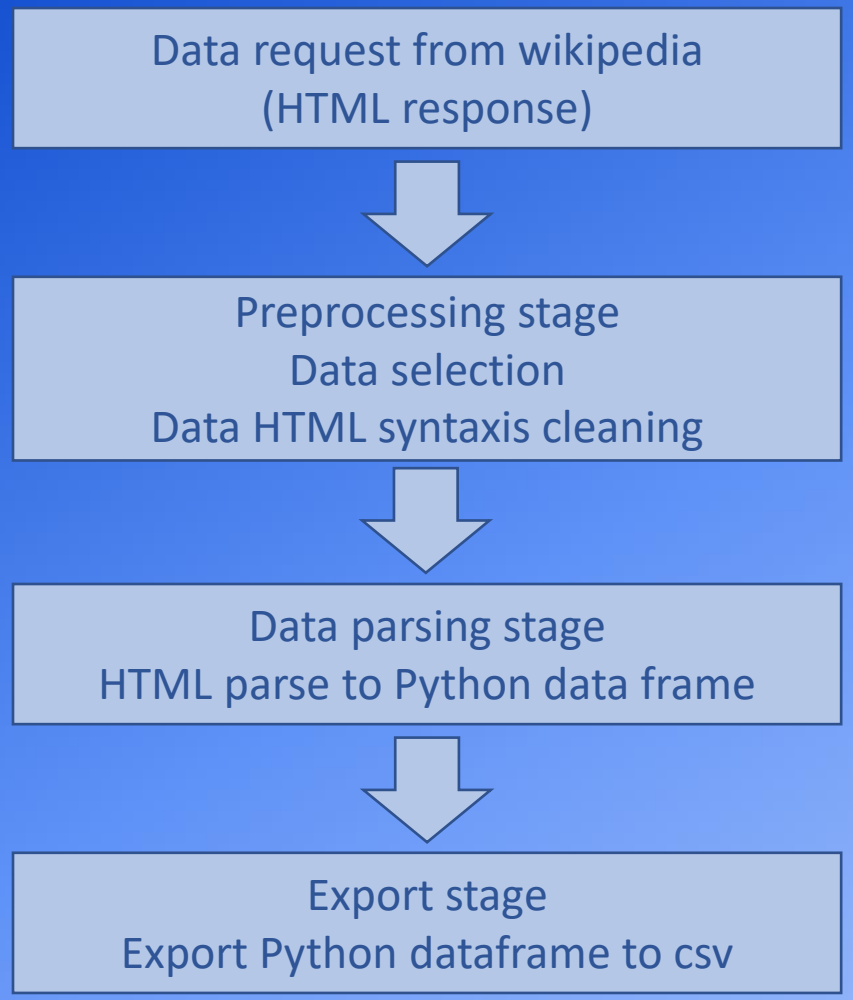
Was used get request to the SpaceX API to collect data, clean the requested data and some basic data wrangling and formatting.



https://github.com/pancenu/IBM-DS-CapstoneProject/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

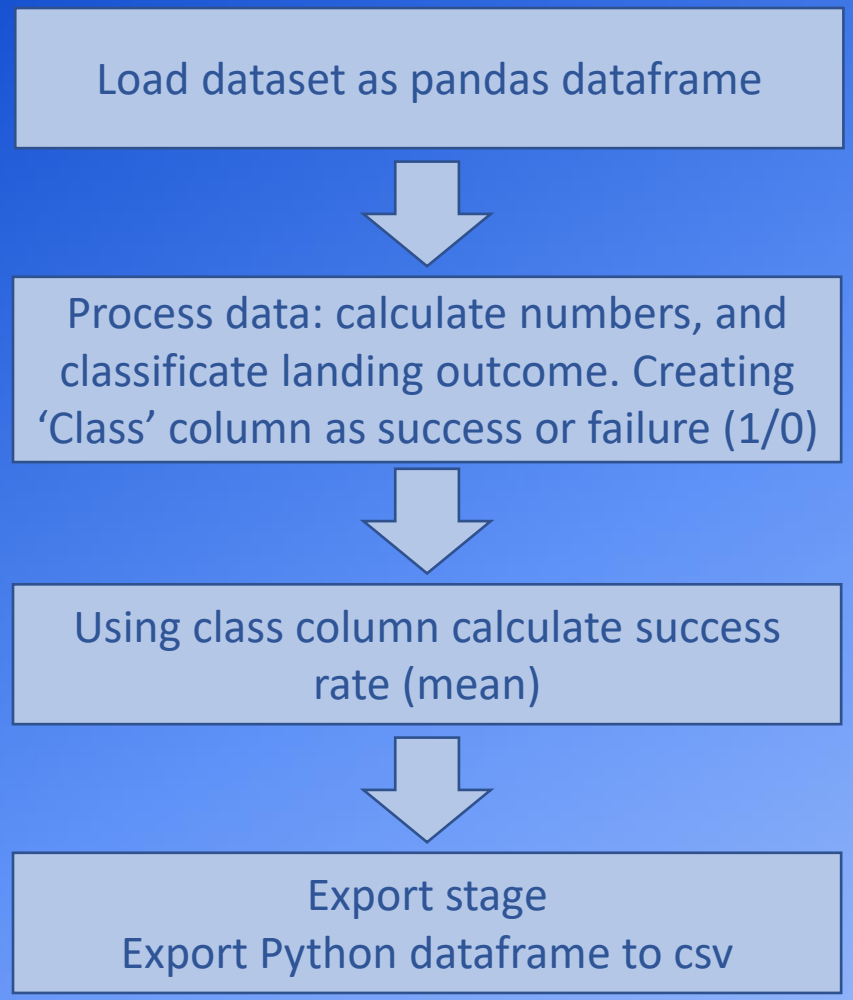
- Present your web scraping process using key phrases and flowcharts



https://github.com/pancenu/IBM-DS-CapstoneProject/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Really data wrangling starts in data collection from API, where missing data was processed, in this secton several number were calculated:
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome of the orbits
- Create a landing outcome label from Outcome column. Creating this label implies the classification of each landing as either a success or a failure (1 / 0).
- Finally export the dataframe as csv file.



https://github.com/pancenu/IBM-DS-CapstoneProject/blob/main/jupyter-labs-spacex-Data-wrangling.ipynb

EDA with Data Visualization

- There were five scatter plots shows relationships between a pair of variables and class (success vs failure) flight number vs payload mass, flight number vs launch site, launch site vs payload mass, flight number vs orbit type and orbit vs payload mass. Column plot success rate vs orbit shows differences between orbits. And finally line plot success rate vs year shows the evolution in time of success rate.

https://github.com/pancenu/IBM-DS-CapstoneProject/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Data was loaded into an sqlite database to use SQL queries
- Queries were written to obtain, among others, the following data:
  - Names of unique launch sites
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v 1.1
  - Total number of successful and failure missions outcomes
  - Date of first successful landing outcome in ground pad was acheived
  - Failed landing outcomes in 2015

https://github.com/pancenu/IBM-DS-CapstoneProject/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Circles and markers were created to identify launch sites on the map, indicating whether the launches were successful (green) or failures (red).
- Several places were located (nearest city, railroad, coastline) distances were calculated and added to the map connected by lines

https://github.com/pancenu/IBM-DS-CapstoneProject/blob/main/jupyter-labs-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

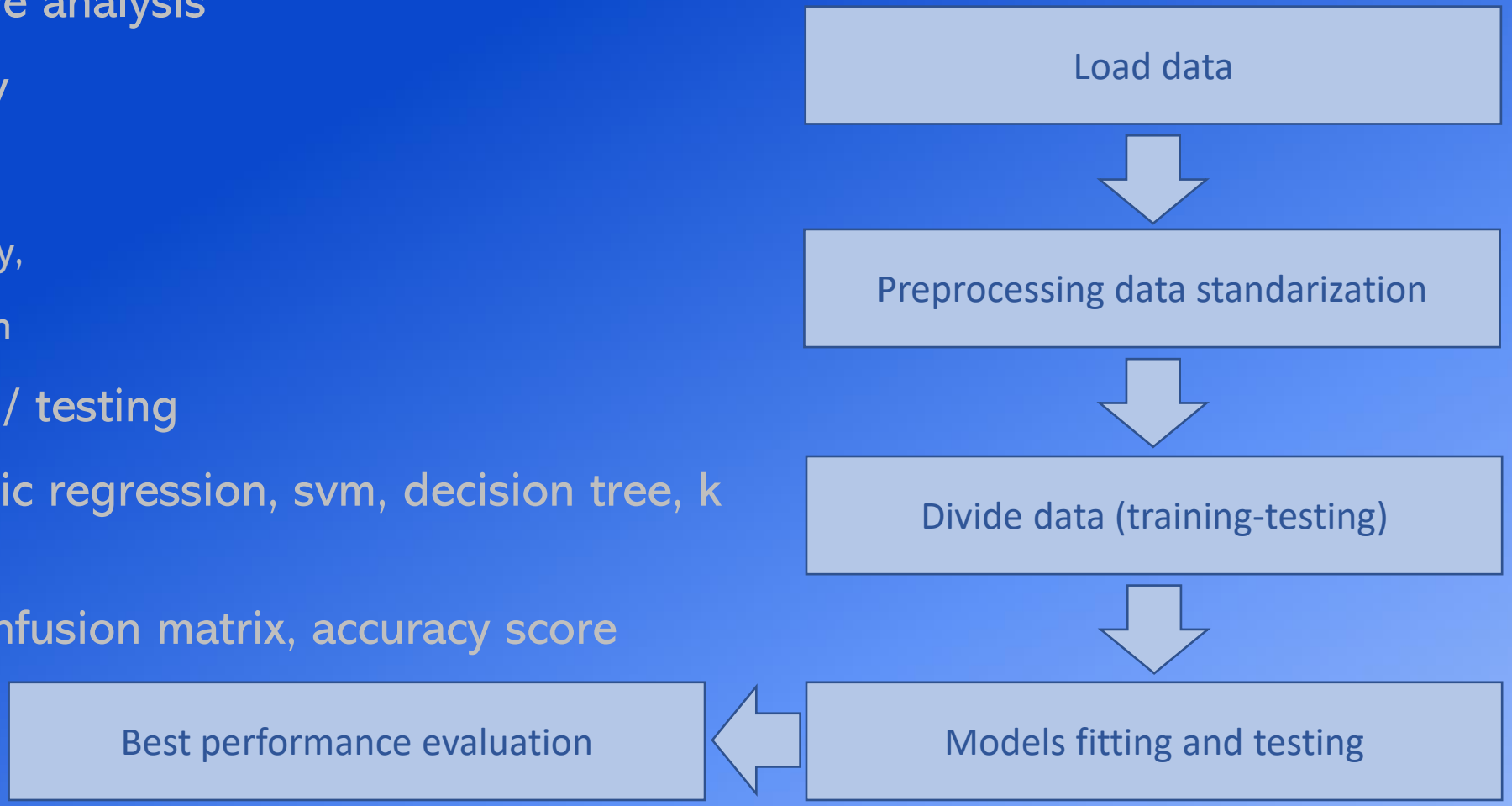
The dashboard built using plotly and dash has two sections:

- A dropdown list where you can select all sites or individual sites, and a pie chart that displays the success/failure proportion.
- A scatter plot showing payload mass vs. class (1 = success, 0 = failure), which can be filtered by payload mass range using a slider.

https://github.com/pancenu/IBM-DS-CapstoneProject/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Process for predictive analysis
- Import data from csv
- Preprocessing data:
  - Class to numpy array,
  - Data standardization
- Divide data training / testing
- Fiting models: logistic regression, svm, decision tree, k nearest neighbors
- Evaluate models: confusion matrix, accuracy score

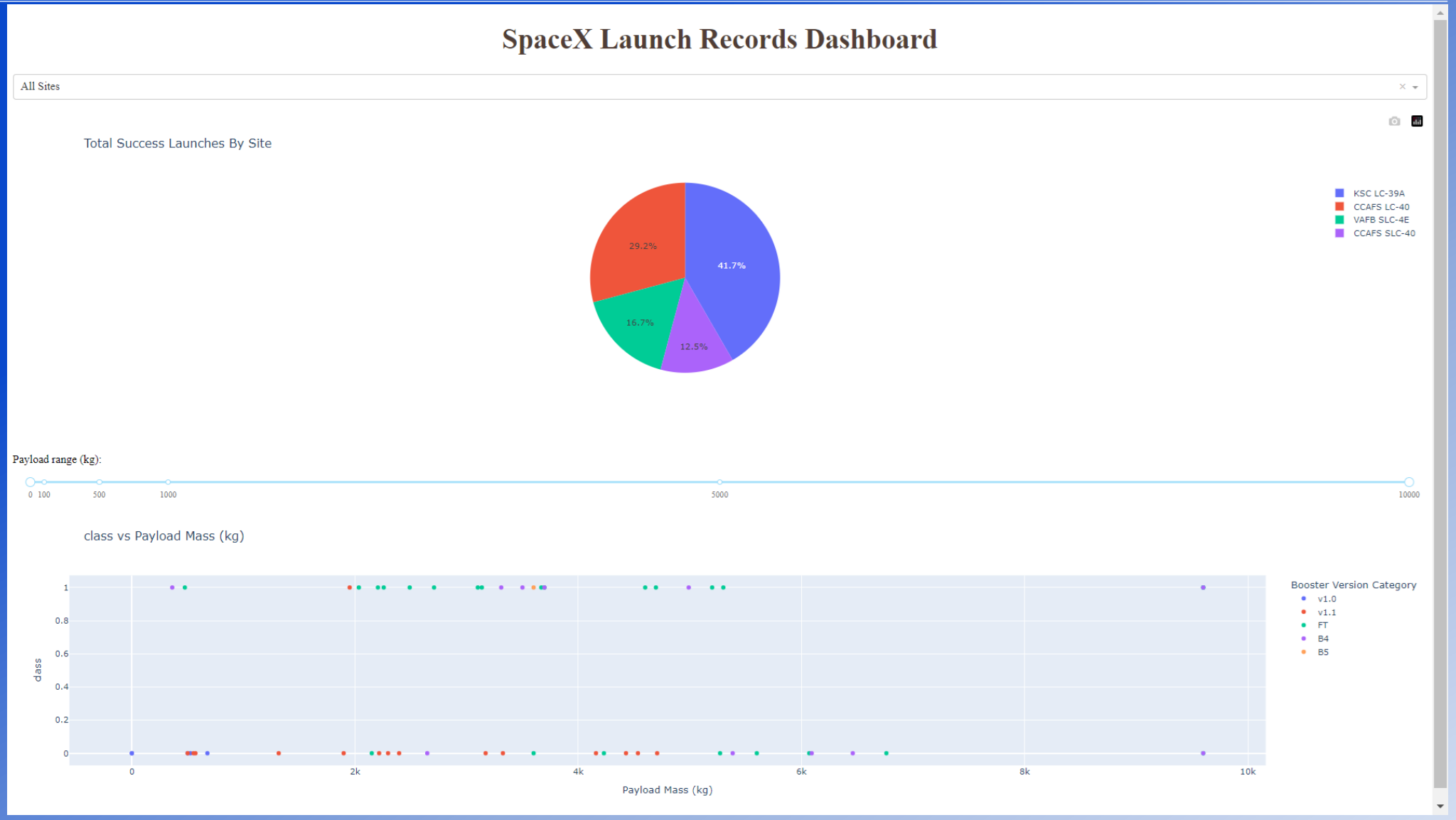


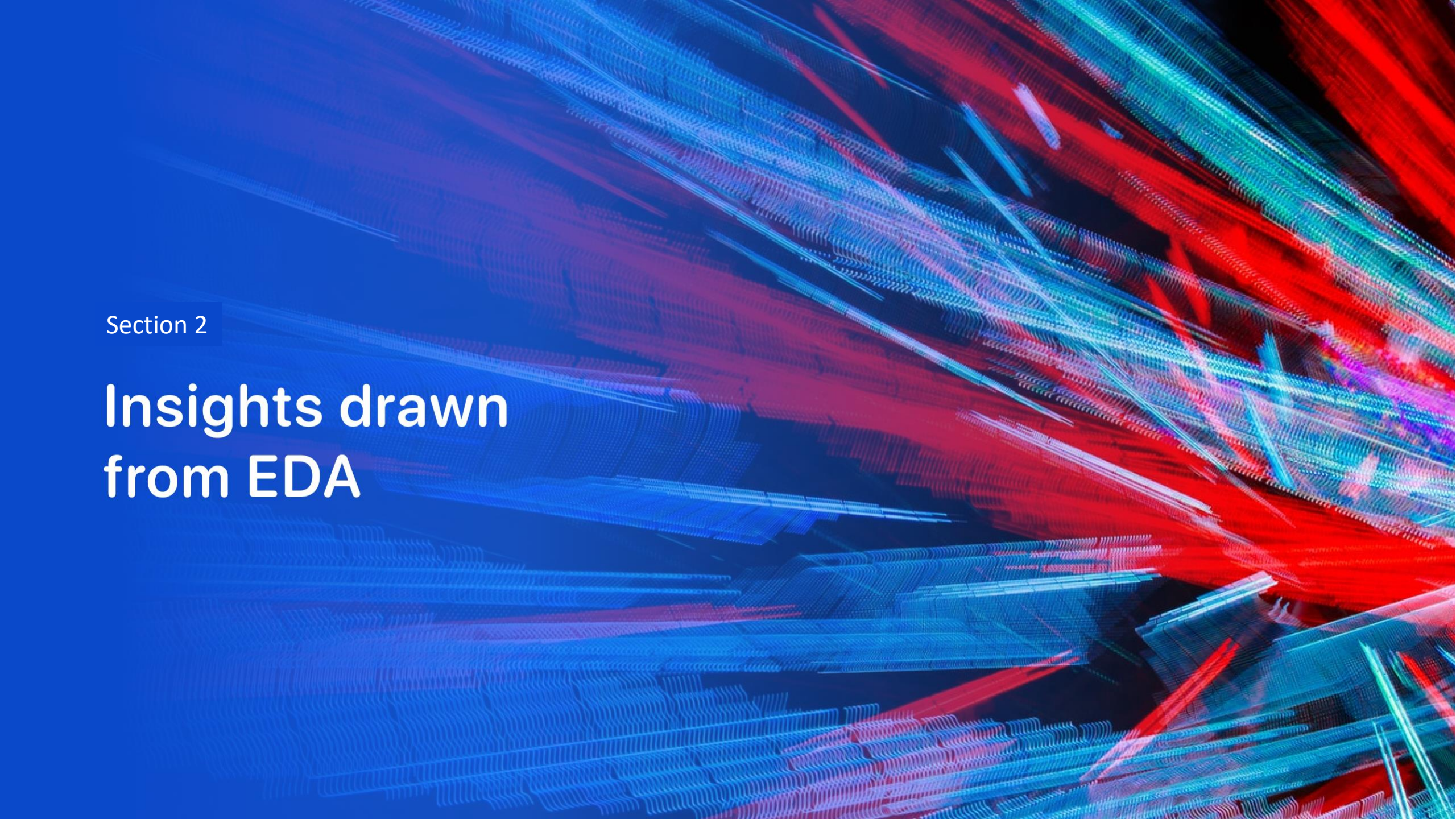
https://github.com/pancenu/IBM-DS-CapstoneProject/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

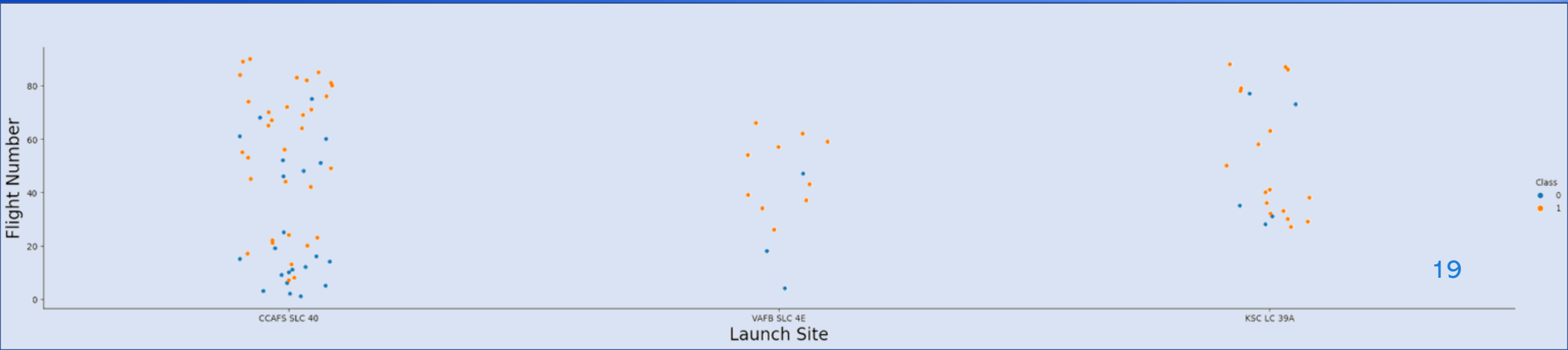
Results


SpaceX Launch Records Dashboard
All Sites
Total Success Launches By Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (kg):
0
100
500
1000
5000
10000
class vs Payload Mass (kg)
class
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5


Section 2
Insights drawn from EDA

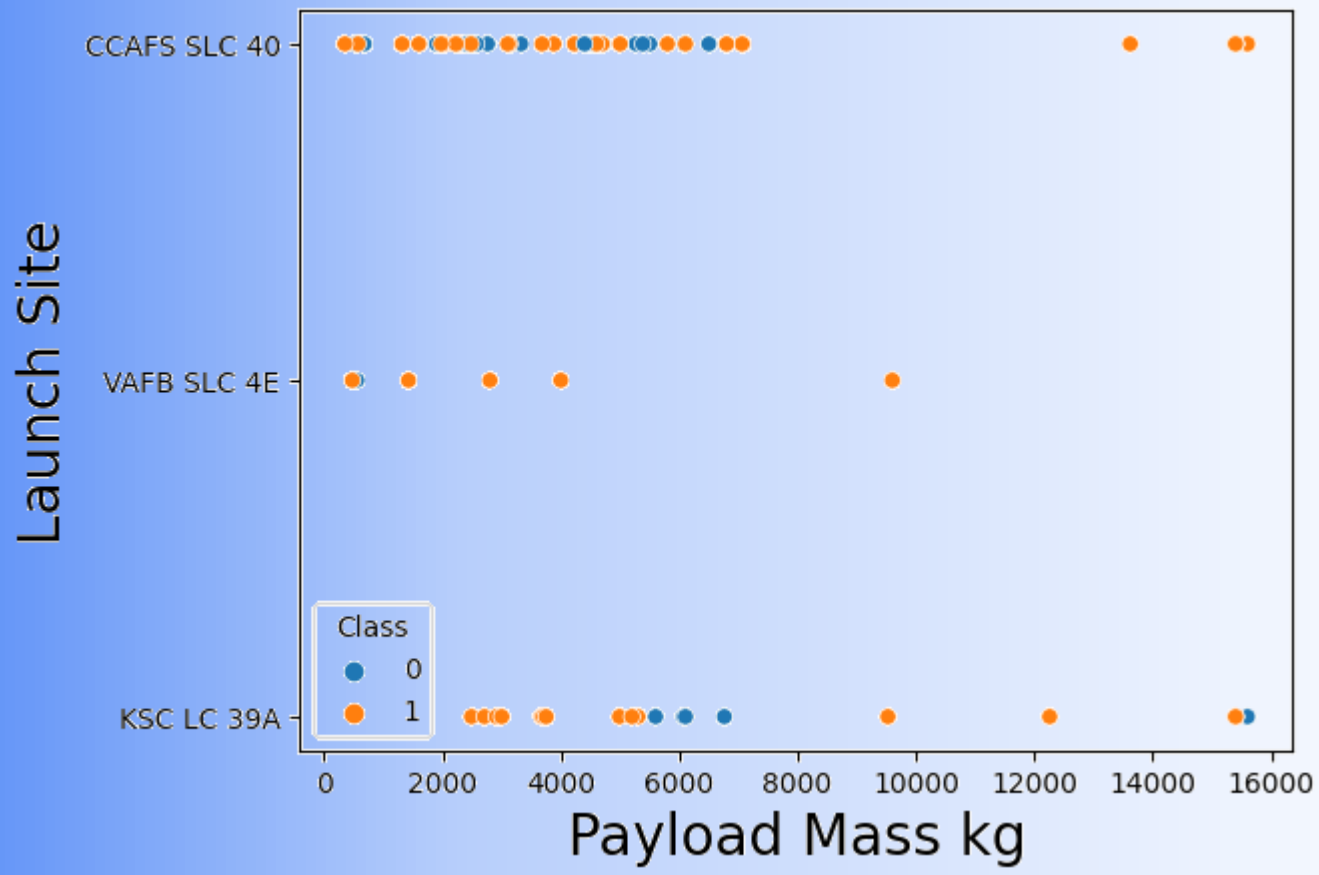
Flight Number vs. Launch Site

- This plot shows that the most of the rockets have been launched in CCAFS
- The recent rocket launches tend to be successful



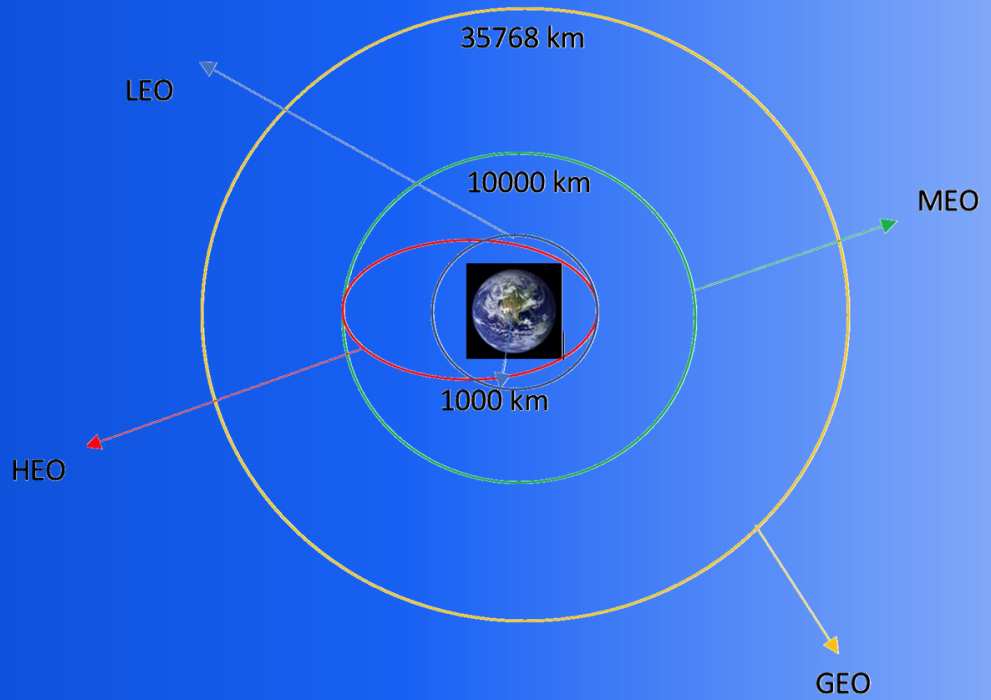
Payload vs. Launch Site

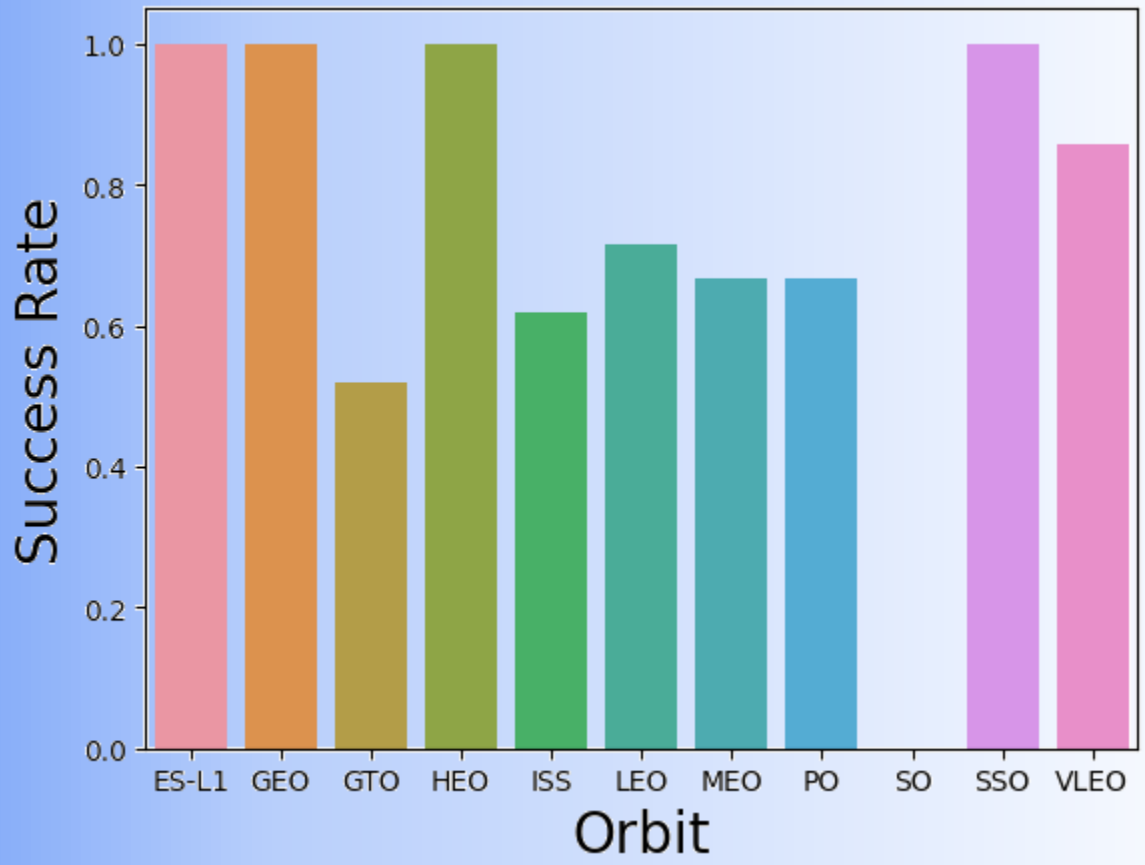
- This plot besides shows:
  - The most of rockets payload mass are under 8 000 kg
  - Besides the most of the heaviest ones have been successful



Success Rate vs. Orbit Type

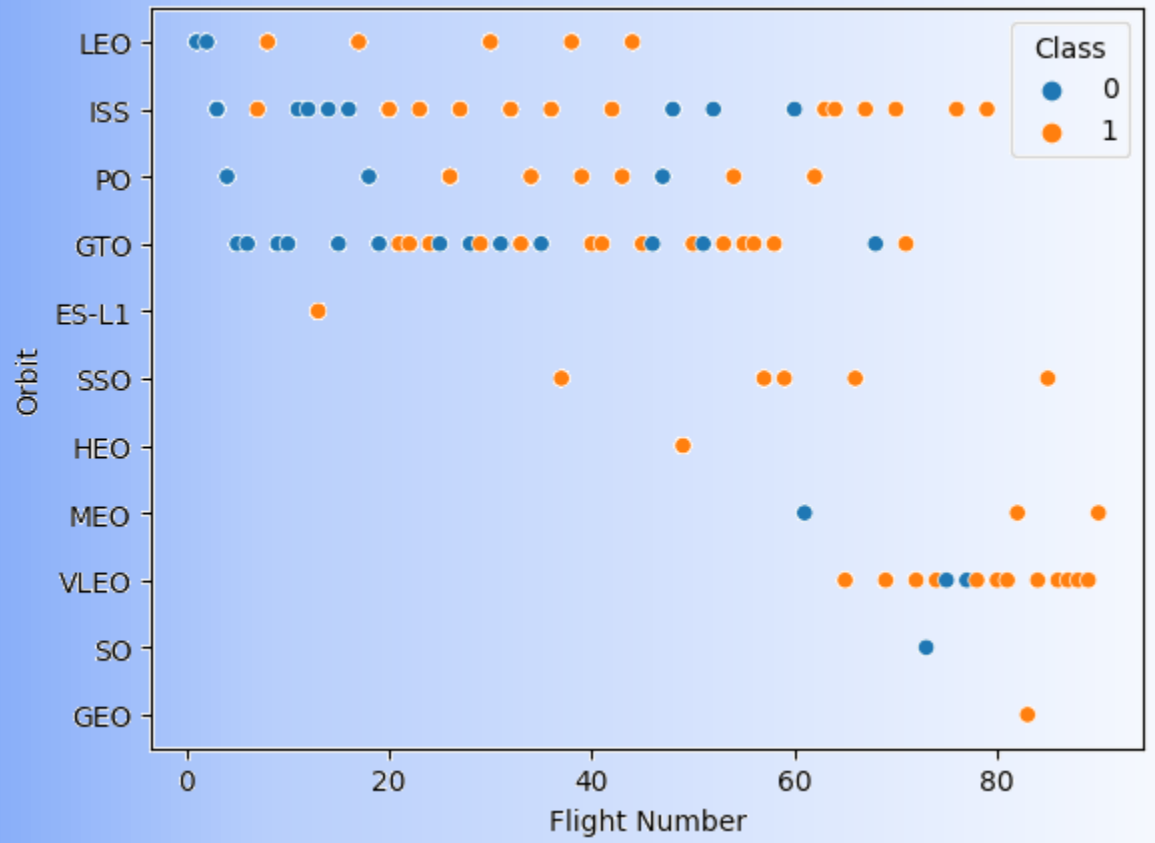
- The chart shows success rate that has been full successful for four orbits, SO has not any success and the rest have values between 50% and 90%





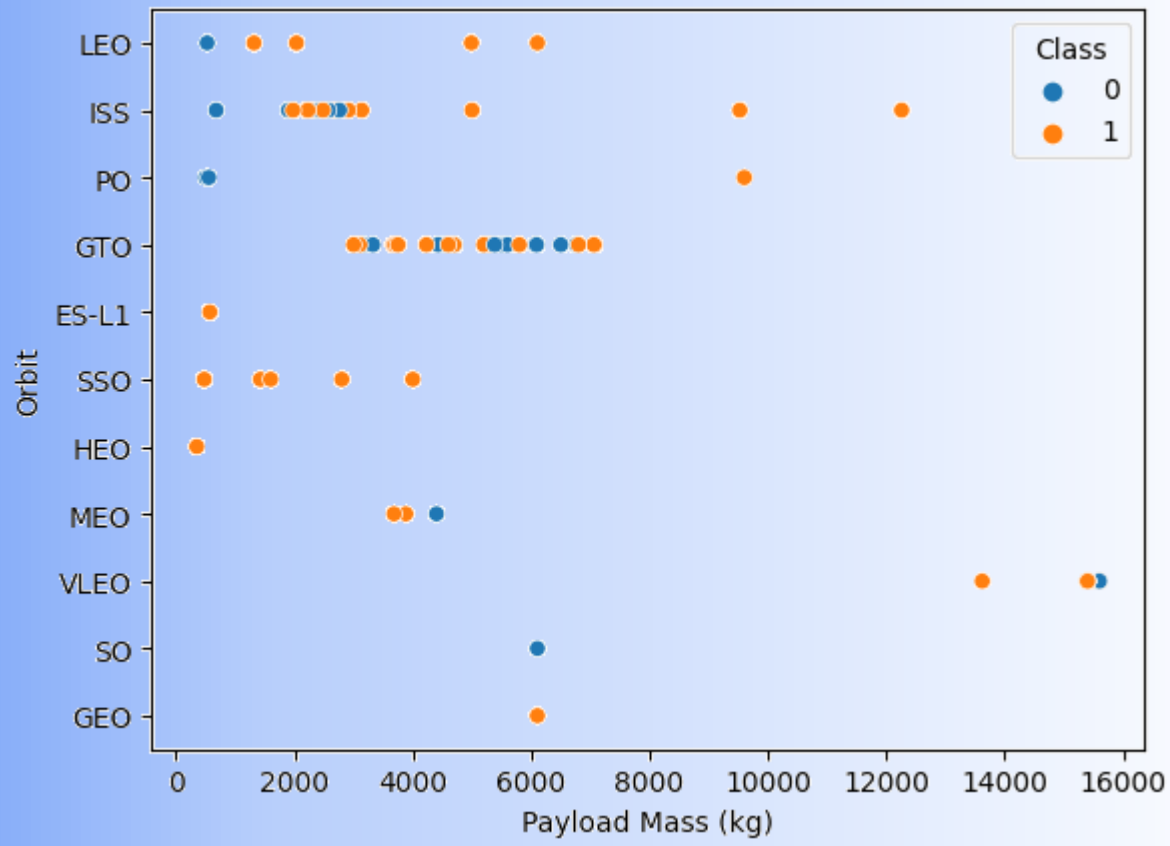
Flight Number vs. Orbit Type

- The scatter plot of Flight number vs. Orbit type shows the evolution, the first ones were unsuccessful and the most of the last ones have been successful.
- There are some orbits with full successful rate but with little launches.



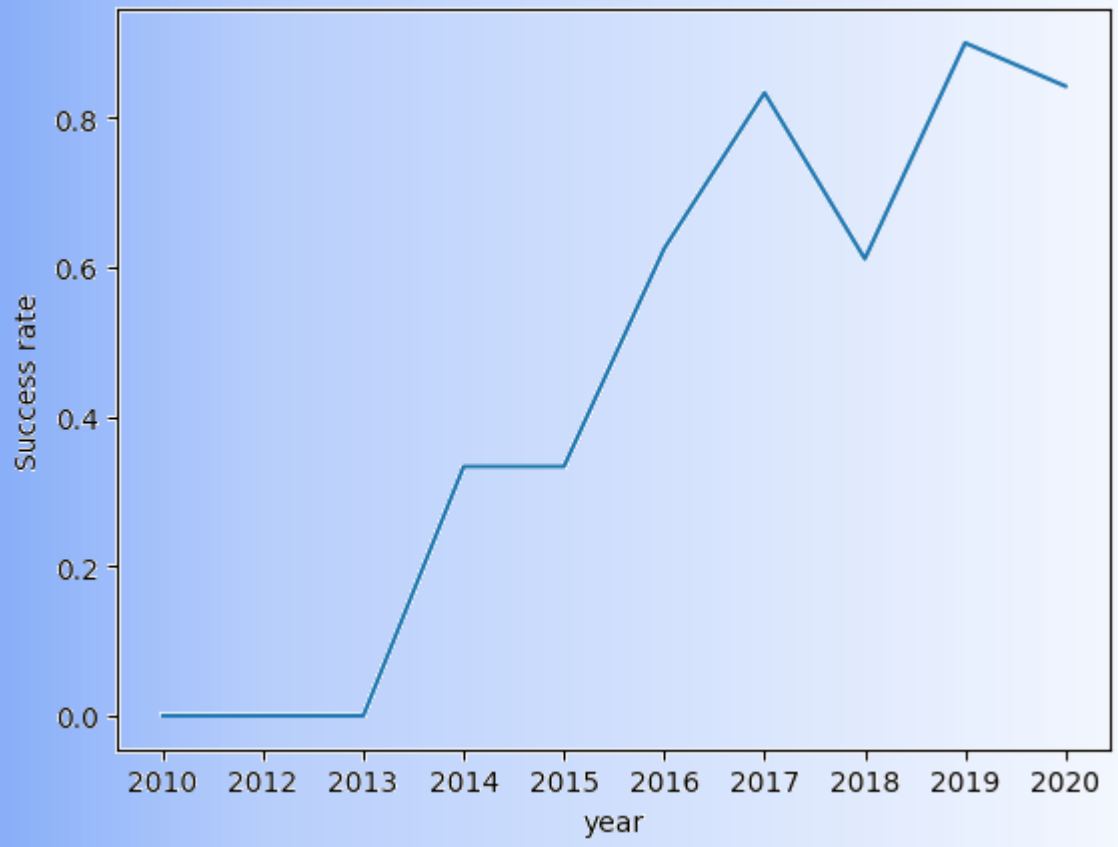
Payload vs. Orbit Type

- Scatter plot of payload vs. orbit type shows that each orbit seems to have a range payload mass.
- ISS has a range and some outliers.



Launch Success Yearly Trend

- The line chart of the yearly average success rate shows an upward trend over time.



All Launch Site Names

There are only four lanch sites:

- CCAFS LC-40:

  Cape Canaveral Launch Complex 40

- CCAFS SLC-40:

  Cape Canaveral Space Launch Complex 40

- VAFB SLC-4E:

  Vandenberg Space Force Base Space Launch Complex 4E

- KSC LC-39[a]

  Kennedy Space Center Launch Complex 39A (LC-39A)

Really the first two are the same place that change the name:

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

[10]

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

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

Launch Site Names Begin with 'CCA'

- This query returns 5 records where the launch sites begin with 'CCA'

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

[11]

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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 is 45 596 kg

```
%%sql
SELECT SUM( PAYLOAD_MASS__KG_ )
FROM SPACEXTABLE
WHERE Customer LIKE "NASA (CRS)"
```

[12]

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload has been 2 534,67 kg

```
[13]
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Booster_Version LIKE "F9 v1.1%"
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The first successful landing outcome on ground pad occurred on December 22th, 2015

```
[14]  %%sql
      SELECT min(Date)
      FROM SPACEXTABLE
      WHERE Landing_Outcome LIKE "Success (ground pad)"
```

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

```
...  min(Date)
     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 only four F9 FT B1022, B1026, B1021.2 and B1031.2

```
[15]
%%sql
SELECT Booster_Version, Payload
FROM SPACEXTABLE
WHERE (Landing_Outcome LIKE "Success (drone ship)")
AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

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

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

- According with this query, the most of the missions have been successful (100/101)

```
[16]
%%sql
SELECT
COUNT(CASE WHEN Mission_Outcome LIKE "Success%" THEN 1 END)
    AS Success_count,
COUNT(CASE WHEN Mission_Outcome LIKE "Failure%" THEN 1 END)
    AS Failure_count
FROM SPACEXTABLE;
```

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

Success_count	Failure_count
100	1

Boosters Carried Maximum Payload

- This query returns the list the names of the booster which have carried the maximum payload mass

```
%%sql
SELECT Booster_Version, Payload, PAYLOAD_MASS__KG_
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ =
    (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
    );
```

[17]

```
 * 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

- This query returns the List of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
SELECT
    CASE SUBSTR(Date,6,2)
        WHEN "01" THEN "January"
        WHEN "02" THEN "February"
        WHEN "03" THEN "March"
        WHEN "04" THEN "April"
        WHEN "05" THEN "May"
        WHEN "06" THEN "June"
        WHEN "07" THEN "July"
        WHEN "08" THEN "August"
        WHEN "09" THEN "September"
        WHEN "10" THEN "October"
        WHEN "11" THEN "November"
        WHEN "12" THEN "December"
    END AS month,
    Booster_Version,
    Landing_Outcome,
    Launch_Site
FROM SPACEXTABLE
WHERE
    ((Landing_Outcome LIKE "Failure (drone ship)")
    AND (SUBSTR(Date,0,5)='2015'))
```

[18]

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

month	Booster_Version	Landing_Outcome	Launch_Site
January	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
April	F9 v1.1 B1015	Failure (drone ship)	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

```
[11]
%%sql
SELECT Date, Landing_Outcome, COUNT(payload) as COUNT
FROM SPACEXTABLE
WHERE Date BETWEEN "2010-06-04" AND "2017-03-20"
GROUP BY Landing_Outcome
ORDER BY COUNT DESC
```

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

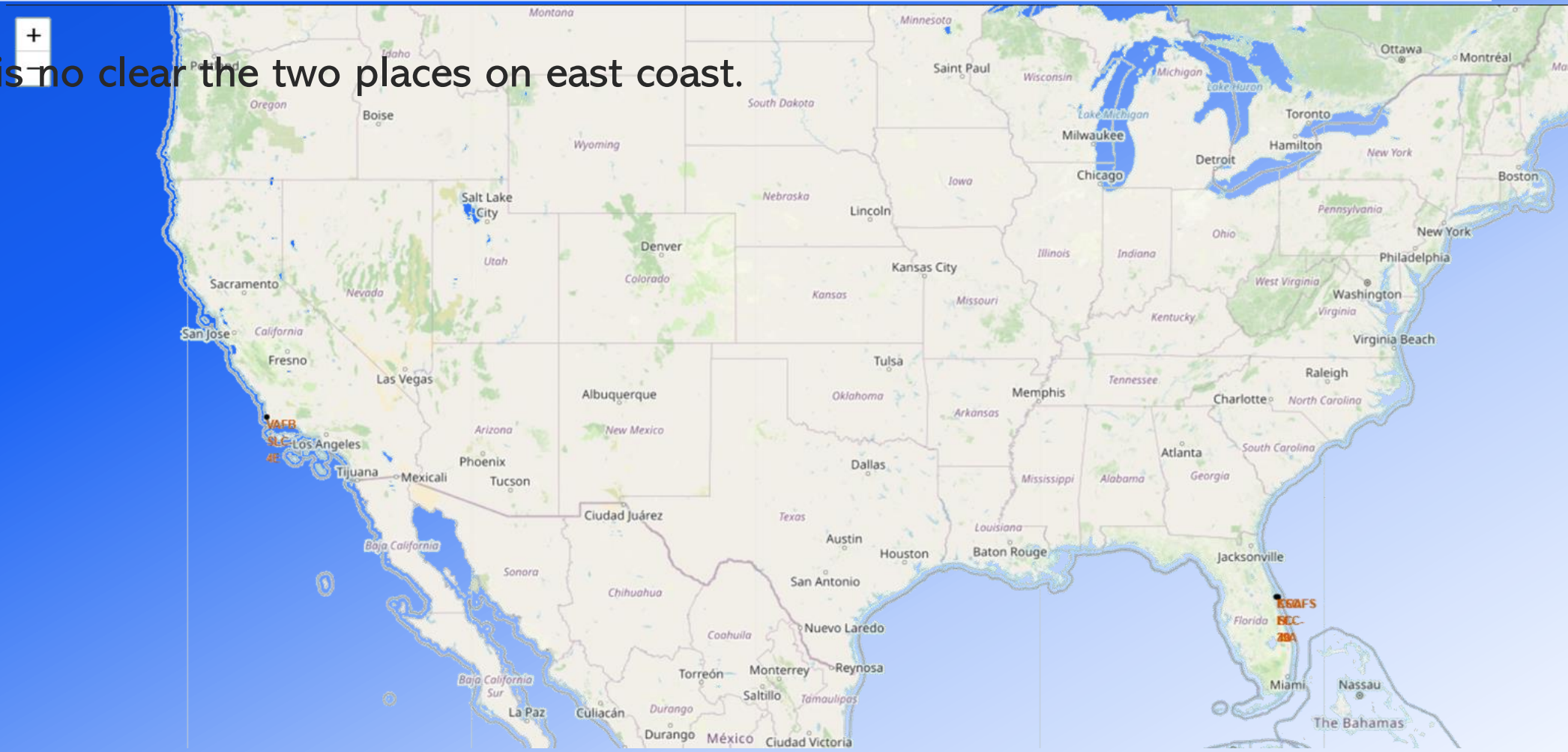
Date	Landing_Outcome	COUNT
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

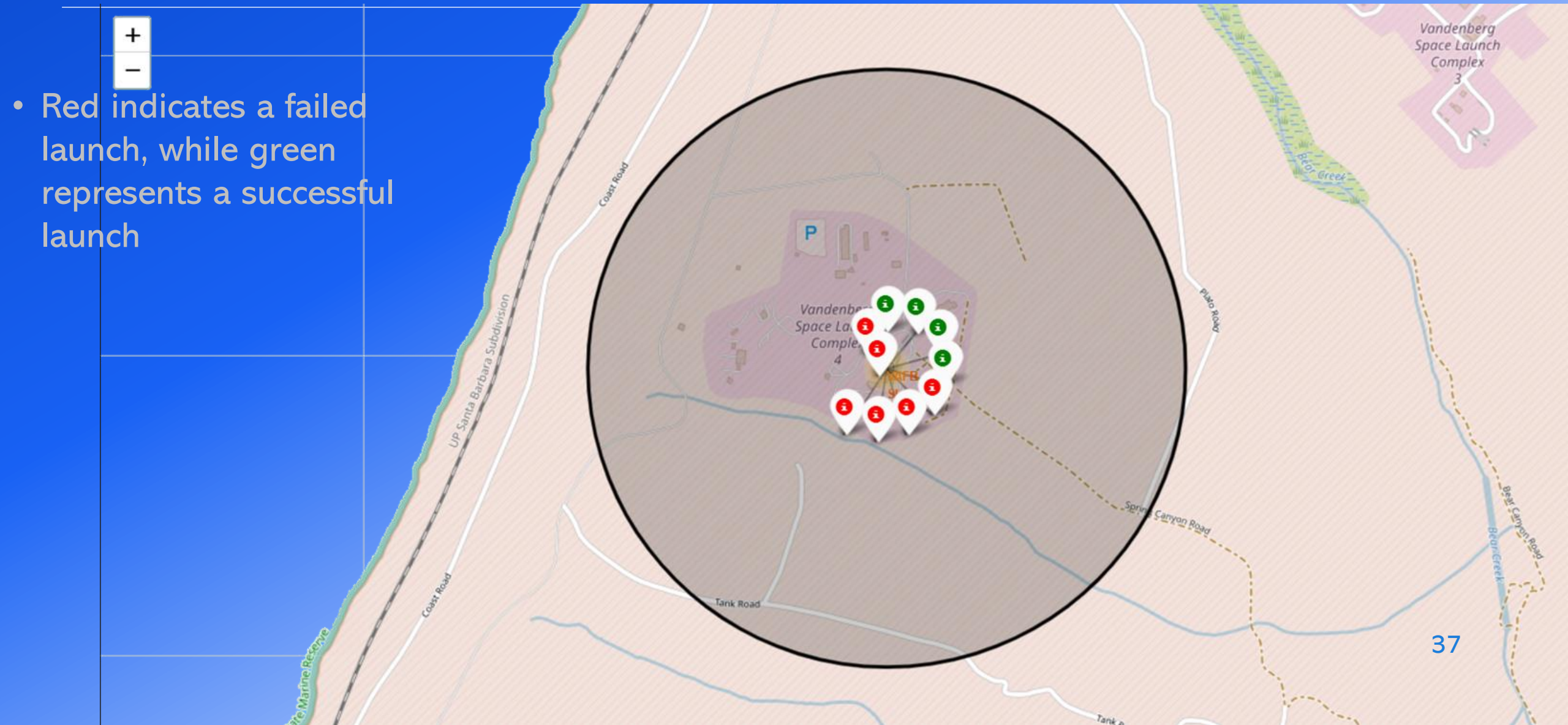
Launch sites location

- Launch sites (it is no clear the two places on east coast.



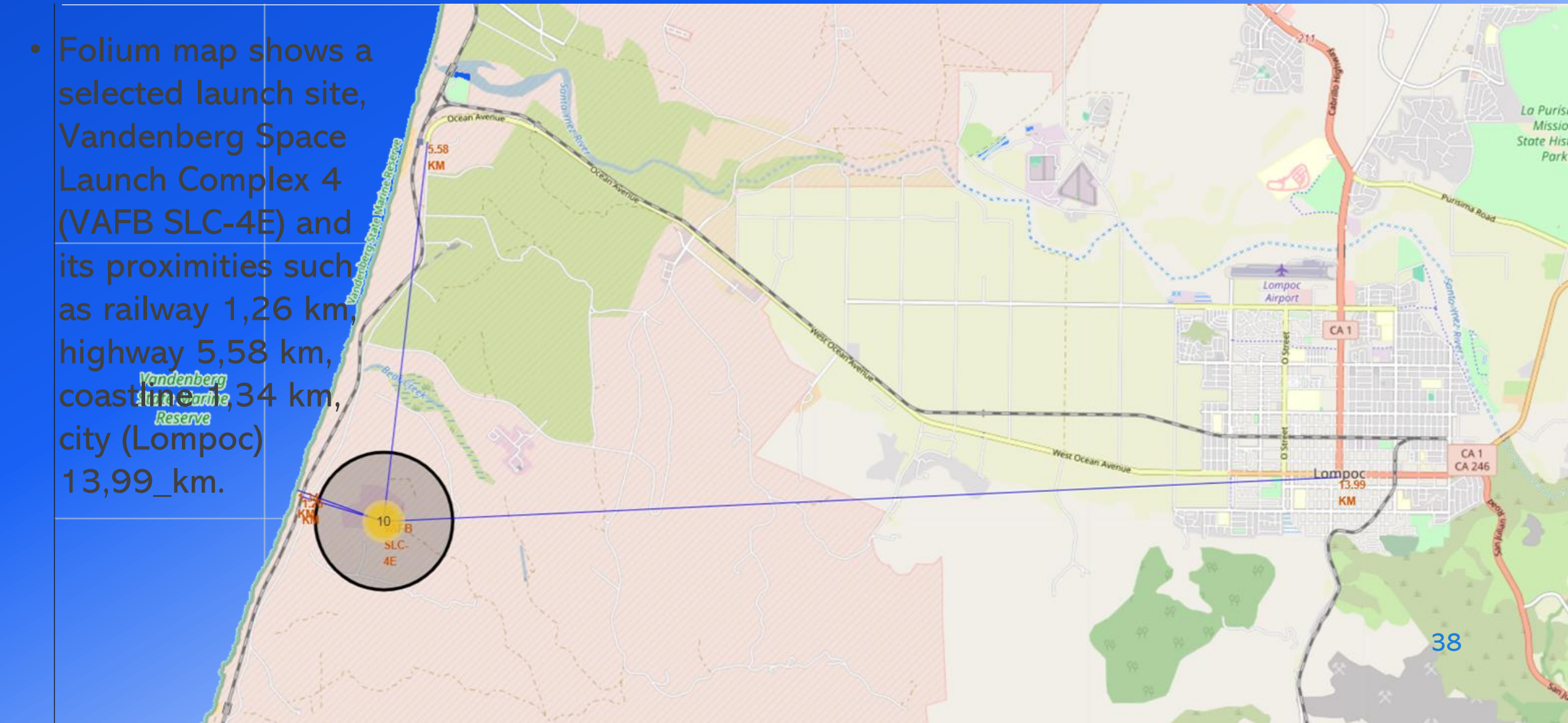
Color labeled launch outcomes on the map

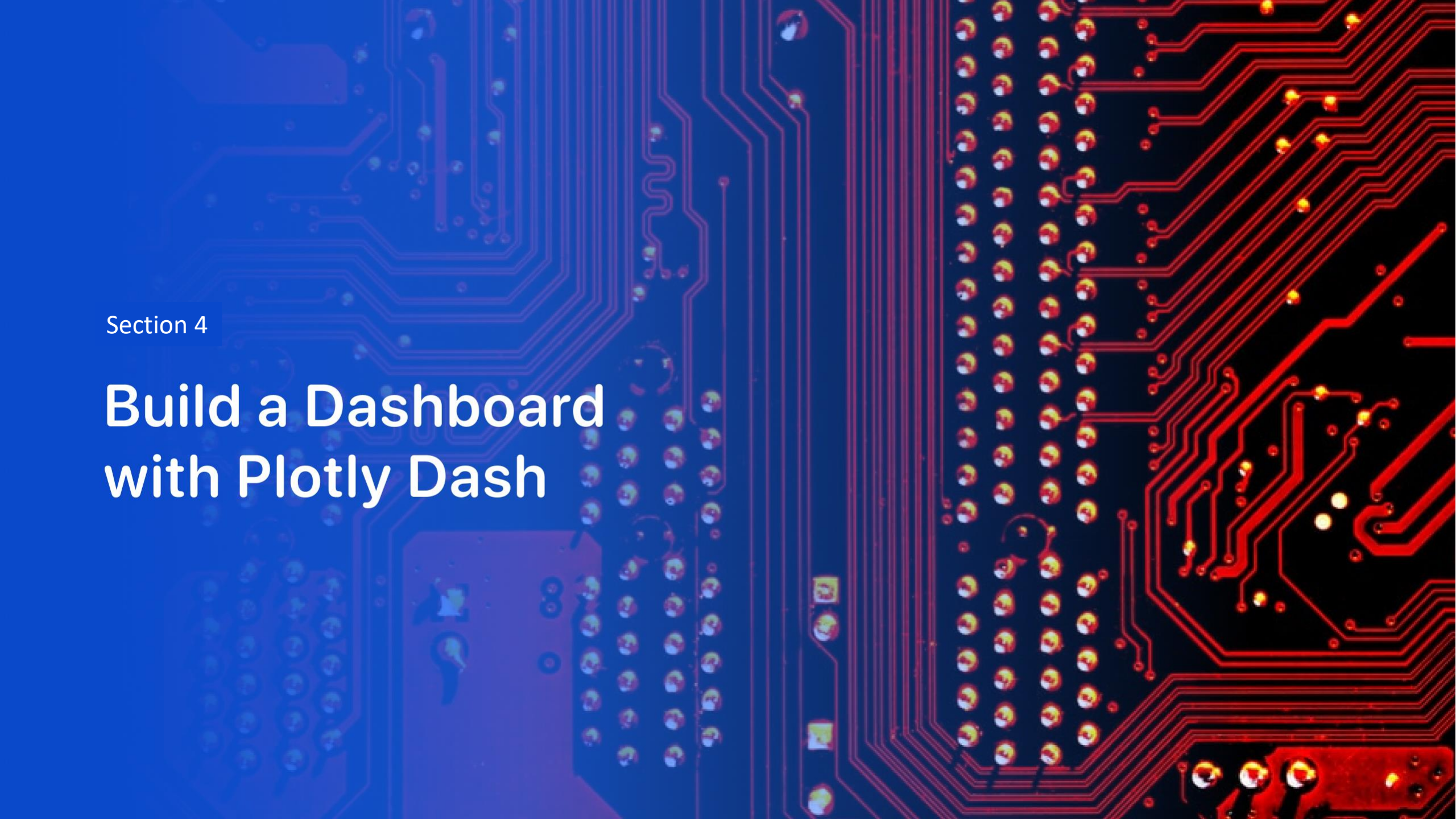
- Red indicates a failed launch, while green represents a successful launch



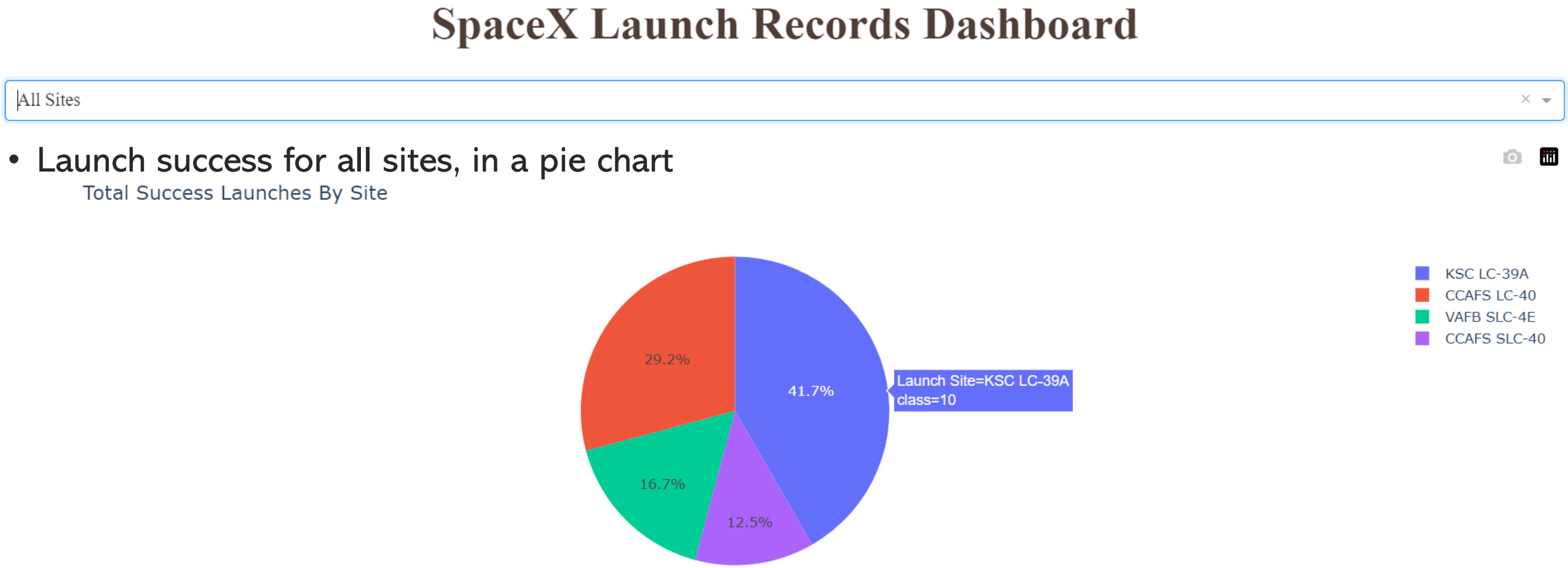
Vandenberg Space Launch Complex Proximities

- Folium map shows a selected launch site, Vandenberg Space Launch Complex 4 (VAFB SLC-4E) and its proximities such as railway 1,26 km, highway 5,58 km, coastline 1,34 km, city (Lompoc) 13,99_km.



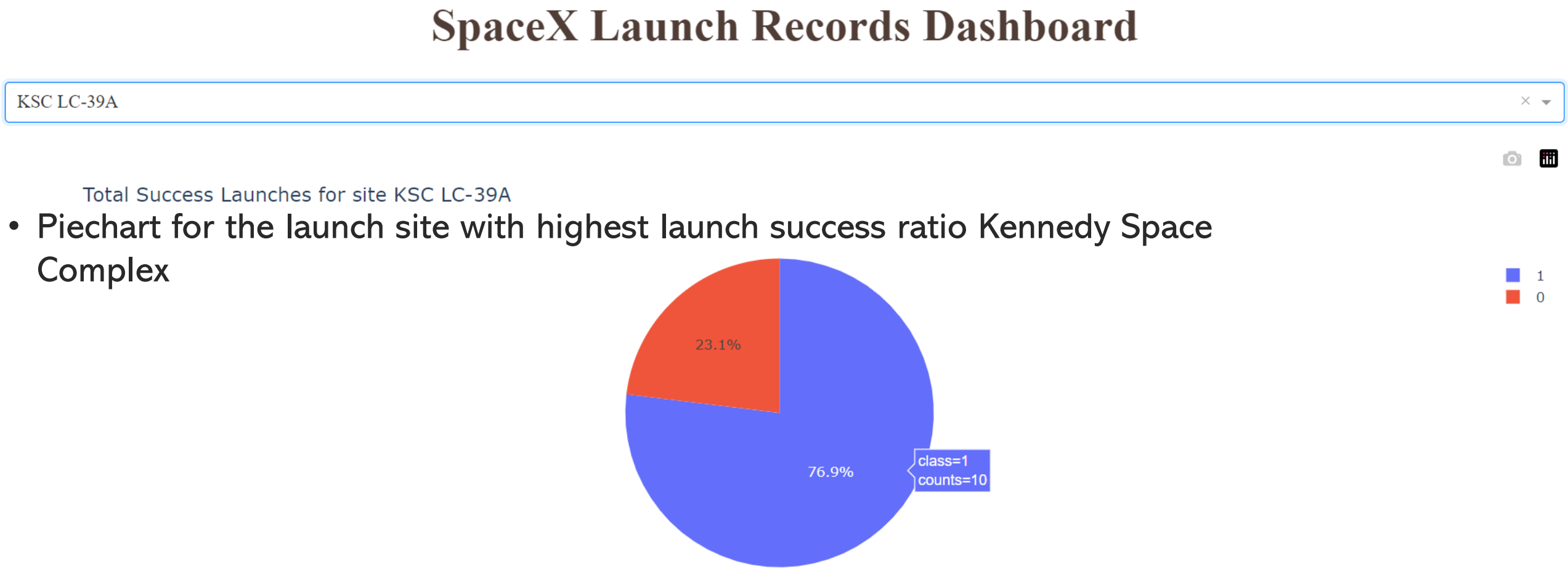

Section 4
Build a Dashboard with Plotly Dash

Launch success for all sites



- Launch success for all sites, in a pie chart

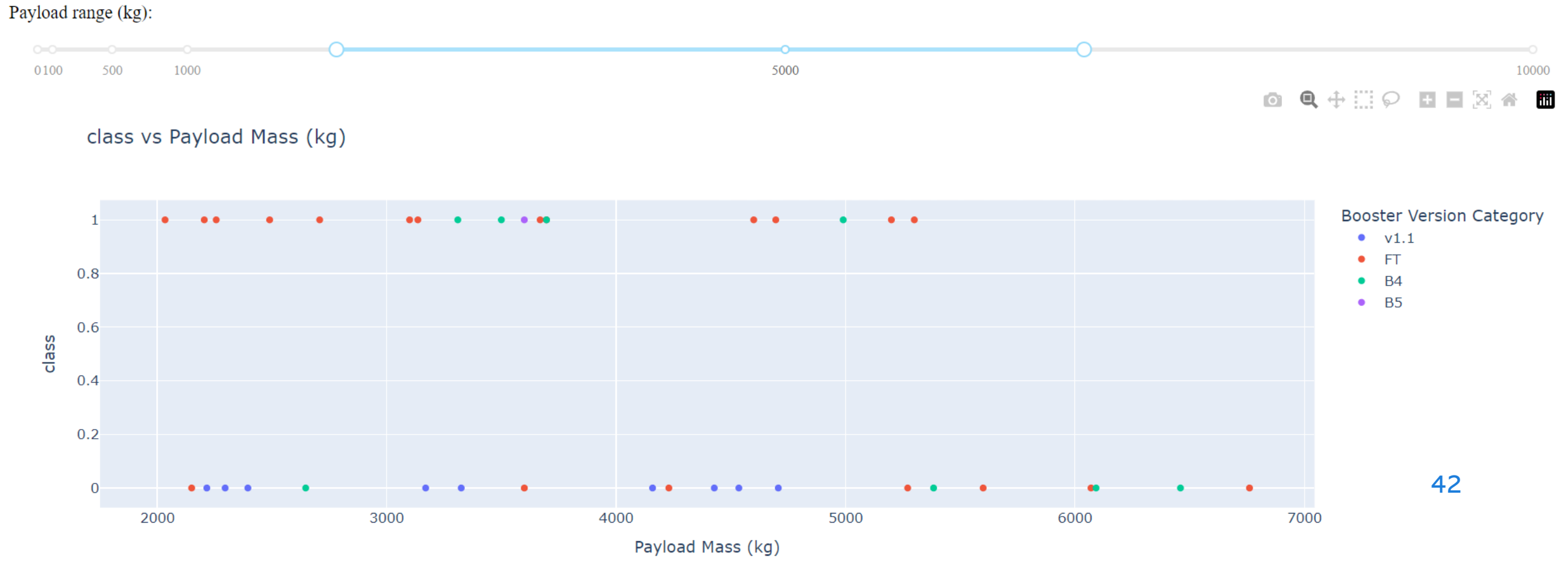
Highest launch success



- Piechart for the launch site with highest launch success ratio Kennedy Space Complex

Class vs payload mass (2 000 – 7 000) kg

- Screenshot shows Launch Outcome vs Payload mass scatter plot for all sites, with payload between 2 000 kg and 7 000 kg.

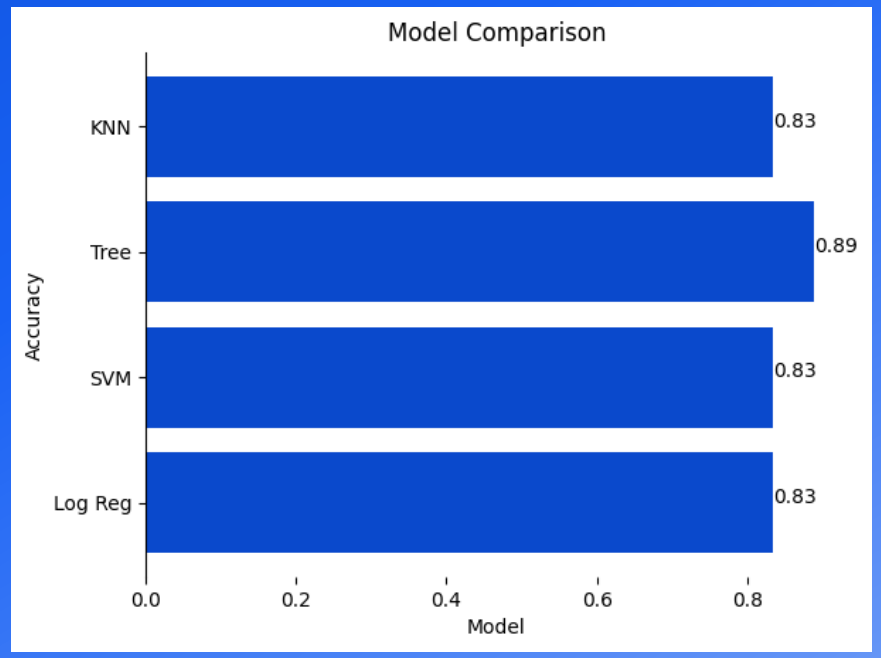


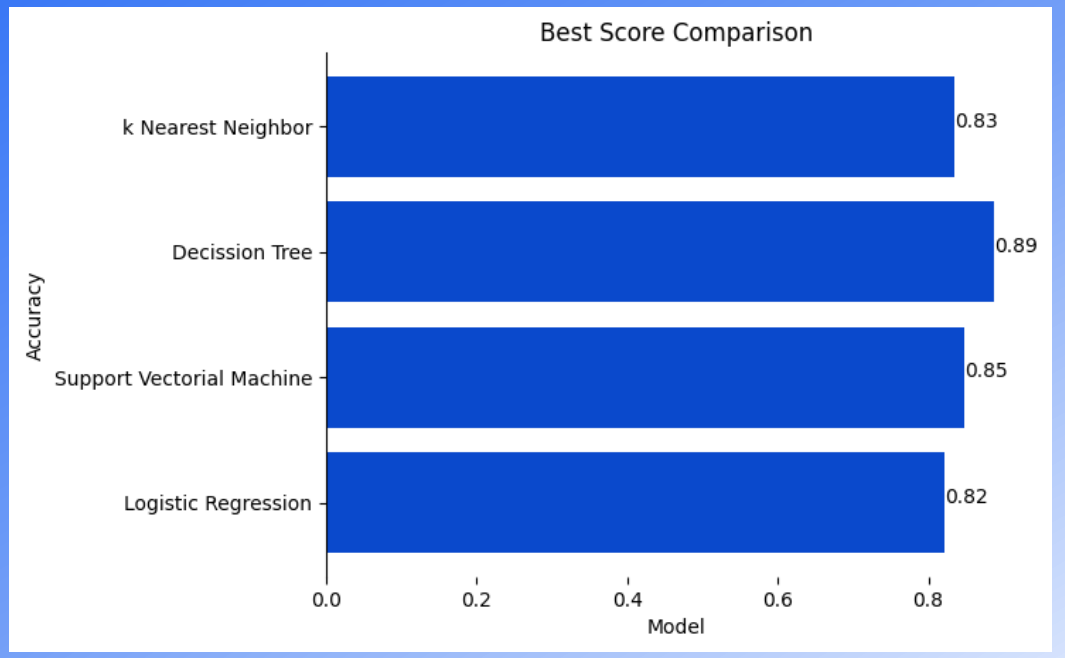
Section 5

Predictive Analysis (Classification)

Classification Accuracy

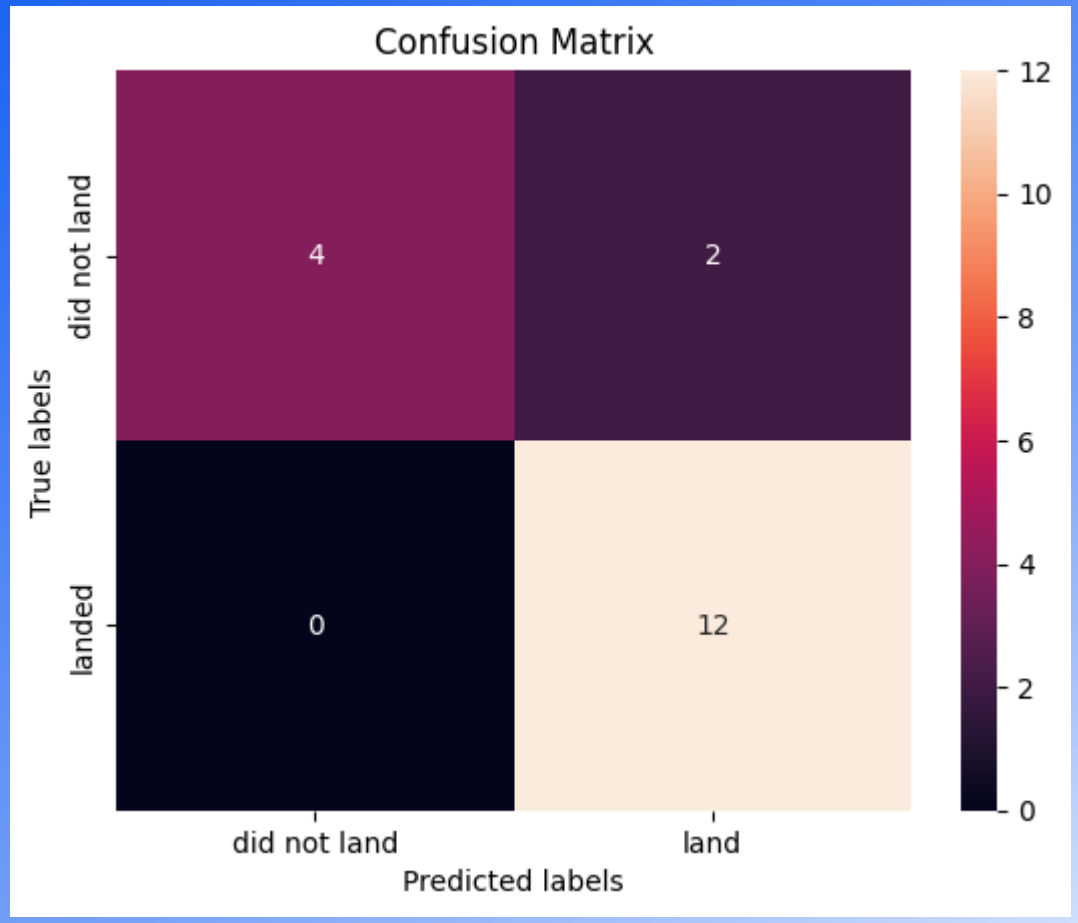
- The models evaluated shows the same accuracy and confusion matrix, except decision tree. Here best scores are compared, that shows that Decision Trees would have the best score.





Confusion Matrix

- This is the confusion matrix of tree decision model shows accuracy as 16/18 = 0.8889
- Precision = 12/14 = 0.857
- Recall = 12/12 = 1
- F1 score = 2*0.857*1 / (1.857) = 0.9231



Conclusions

- The success rate has been raising through the time.
- The decision tree classifier seems to be the best machine learnig model for the prediction of success launches.
- Kennedy Space Complex has been the most successful launch site.

Appendix

- All the project can be found here:

  https://github.com/pancenu/IBM-DS-CapstoneProject/

Thank you!