IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Francisco Roman >
<12.04.2023>

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection through API.
  - Data Collection with Web Scraping.
  - Data Wrangling.
  - Exploratory Data Analysis with SQL.
  - Exploratory Data Analysis with Data Visualization.
  - Interactive Visual Analytics with Folium.
  - Machine Learning Prediction.
- Summary of all results:
  - Exploratory Data Analysis result.
  - Interactive analytics in screenshots.
  - Predictive Analytics result.

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with SpaceX.
- Problems you want to find answers:
  - The best way to estimate the total cost for launches,by predicting successful landings of the first stage ofrockets?
  - Where is the best place to make launches?.

Section 1

Methodology

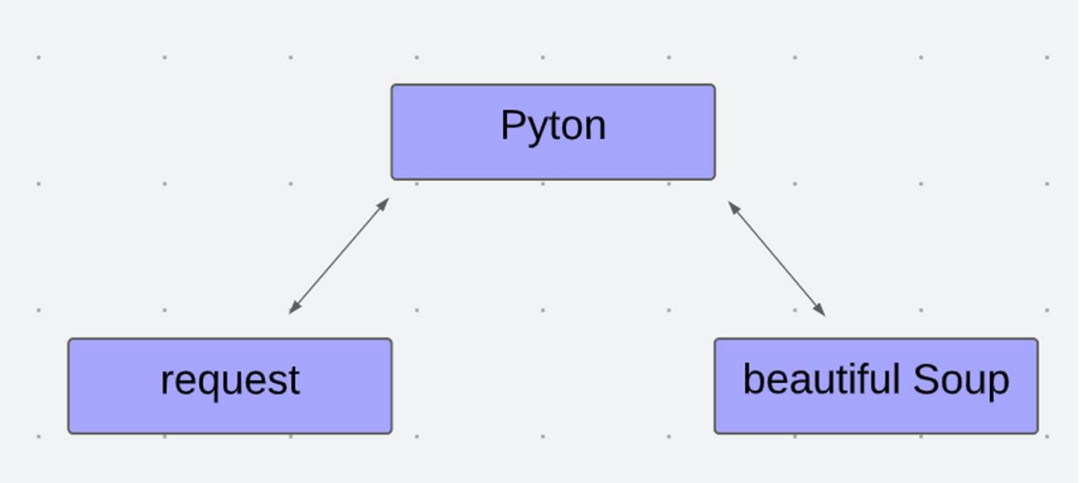
Methodology

Executive Summary

- Data collection methodology:
  - Space X data was collected from two sources, SpaceX API and Web Scraping
- Perform data wrangling
  - In the data we create a landing outcome label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We loaded the data and transformed to use in our training and testing with different model of machine learning (logistic regression, support vector machine, decision tree and k nearest neighbors), to this we use GridSearchCV with various hyperparameters. We used accuracy of ours models to compare their values and obtain the best model.

Data Collection

- Describe how data sets were collected.
  - Data was collecting using API request to SpaceX API (https://api.spacexdata.com/v4/rockets/) and using web scraping to Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scrapingtechnics.



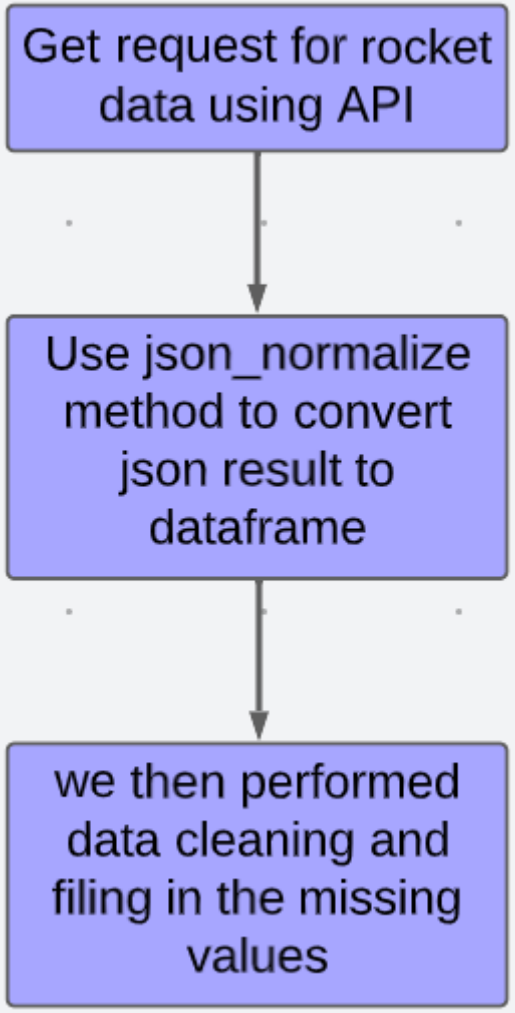
Flow chart of web scraping

Flow chart of API method

Data Collection – SpaceX API

- Data can be obtained from SpaceX API , this data is collected by a API request. The data is cleaned and formatting.

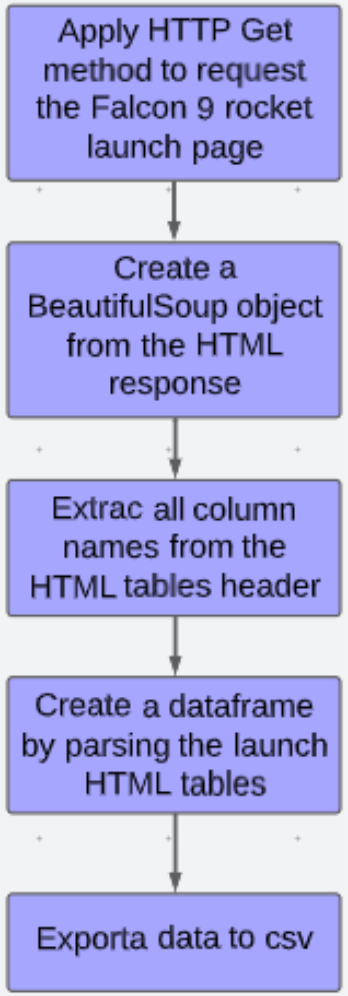
- https://github.com/PanchitoAureo/Applied-Data-Science-Capstone-IBM-data-science-professional-certificate/blob/main/jupyter-labs-spacex-data-collection-api%20(3).ipynb



Data Collection - Scraping

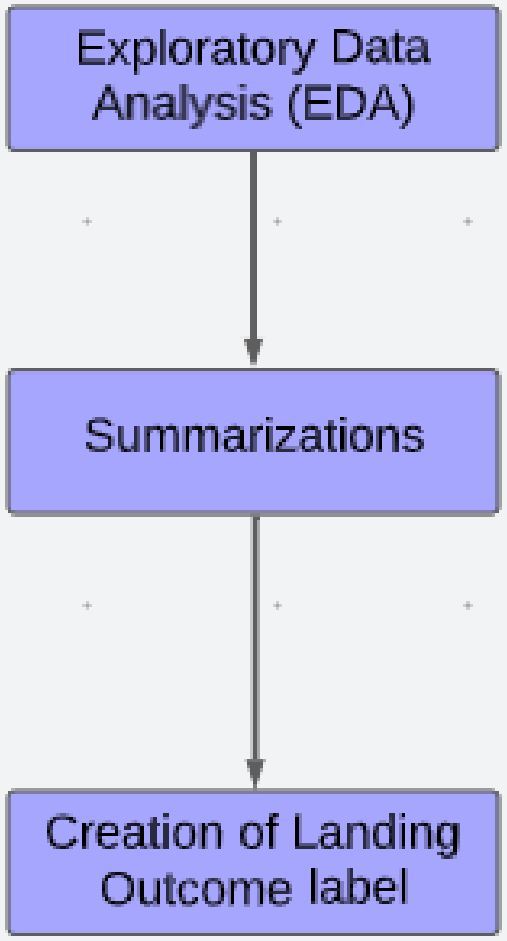
- Webscraping is applied to Wikipedia page, we recollect only the data from Falco 9 with beatifulSoup. The table is converted into a pandas dataframe.

- https://github.com/PanchitoAureo/Applied-Data-Science-Capstone-IBM-data-science-professional-certificate/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- On the data ser was performed Exploratory Data Analysis (EDA)
- Calculated the number of launches per site and occurrence and number of each orbits
- Create a outcome label from outcome column and export the results to csv
- https://github.com/PanchitoAureo/Applied-Data-Science-Capstone-IBM-data-science-professional-certificate/blob/main/jupyter-labs-webscraping%20(1).ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- We used Catplot because the variables show the relationship between a numerical and one or more categorical variables using one of several visual representations. We used Catplot to visualize the relationship between:
  - Flight Number vs Payload.
  - Flight Number vs Launch.
  - Flight Number vs Orbit type.
  - Payload vs Launch Site.
  - Payload vs Orbit Type.
- We Used BarChart to visualize the relationship between each Orbit type.
- We Used Line Chart to visualize the launches success yearly trend.
- https://github.com/PanchitoAureo/Applied-Data-Science-Capstone-IBM-data-science-professional-certificate/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

SQL queries performed:

- Display the names of the unique launch sites in the space mission:
  SELECT DISTINCT(launch_site) FROM SPACEXTBL;
- Display 5 records where launch sites begin with the string ‘CCA’:
  SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;
- Display the total payload mass carried by boosters launched by NASA (CRS):
  SELECT SUM(payload mass __kg)AS TOTAL_PAYLOAD_ MASS FROM SPACEXTBL WHEREcustomer="NASA (CRS);
- Display average payload mass carried by booster version FO v1.1:
  SELECT AVG{(payload_mass__kg_) AS AVG_PAYLOAD_MASS FROM SPACEXTBL WHEREbooster_version="F9v1.1’;
- List the date when the first successful landing outcome in ground pad was achieved:
  SELECT MIN(DATE) AS first_successful_landing FROM SPACEXTBL WHERE (landing _outcome)=Successrae)
- https://github.com/PanchitoAureo/Applied-Data-Science-Capstone-IBM-data-science-professional-certificate/blob/main/EDA%20with%20SQL%20lab.ipynb

Build an Interactive Map with Folium

Summary of map objects that were created and added to the Folium map

- folium.Circle and folium.Marker to add a highlighted circle area with a text label on a specific coordinate for each launch site on the site map.*
- MarkerCluster object for simplify a map containing many markers having the same coordinate.
- MousePosition on the map to get coordinate for a mouse over a point on the map.
- folium.PolyLine object to draw a line between a launch site to its closest city, railway and highway.
- https://github.com/PanchitoAureo/Applied-Data-Science-Capstone-IBM-data-science-professional-certificate/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

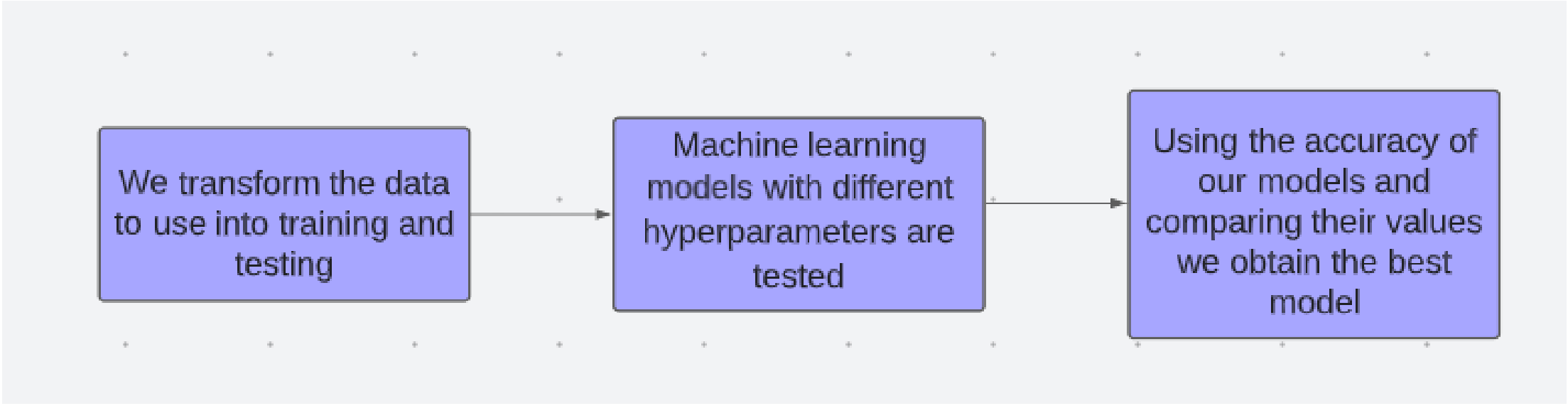
Build a Dashboard with Plotly Dash

- In this section, an interactive dashboard will be shown where the launch percentages for different launch sites will be graphed. One pie chart will be displayed for all sites and one figure for each site.
- In addition, a scatter plot will be shown where different booster combinations are bought with their payload mass.
- The importance of the pie charts is that they quickly allow us to see which launch site has the highest and lowest success rate. The scatter plot allows us to know which booster combination is the most effective and also in which weight ranges the shots are successful.

- https://github.com/PanchitoAureo/Applied-Data-Science-Capstone-IBM-data-science-professional-certificate/blob/main/spacex_dash_app.py

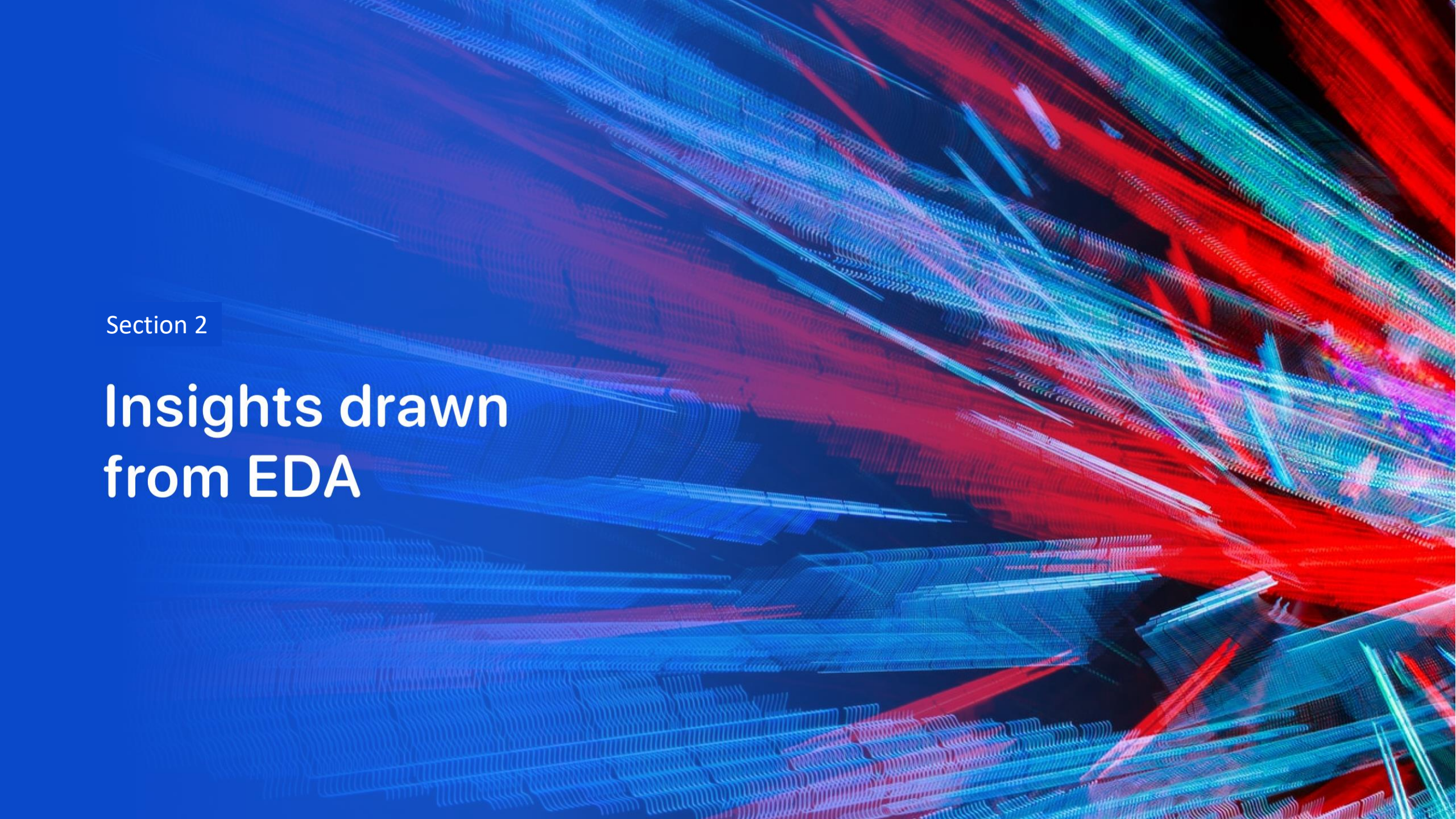
Predictive Analysis (Classification)

- We loaded the data and transformed to use in our training and testing with different model of machine learning (logistic regression, support vector machine, decision tree and k nearest neighbors), to this we use GridSearchCV with various hyperparameters. We used accuracy of ours models to compare their values and obtain the best model.
- https://github.com/PanchitoAureo/Applied-Data-Science-Capstone-IBM-data-science-professional-certificate/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



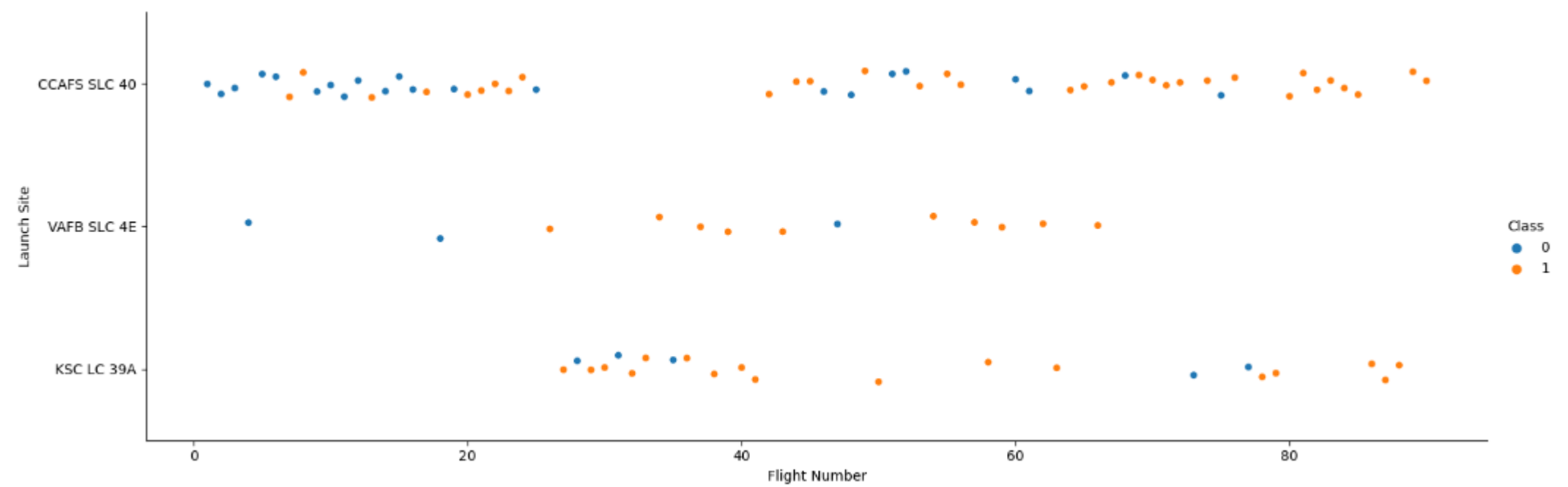
Results

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


Section 2
Insights drawn from EDA

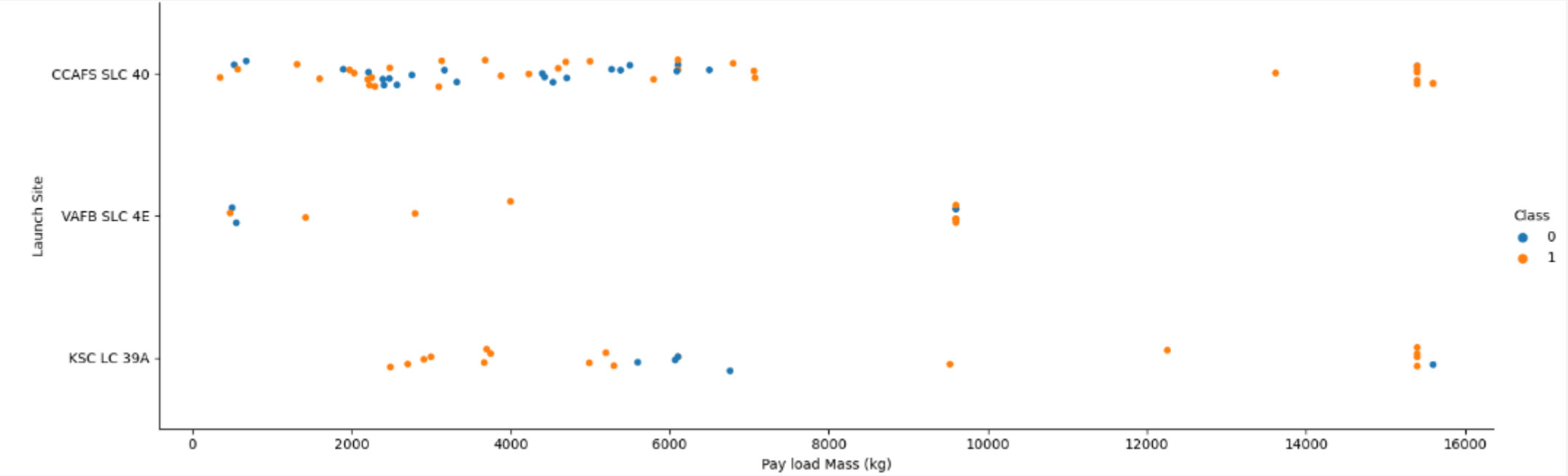
Flight Number vs. Launch Site

- The launch site is CCAF5 SLC40 concentrated the majority of the launches
- The launch site is CCAF5 SLC40 has the worst ratio of successful
- The successful rate has increased for every Launch Site



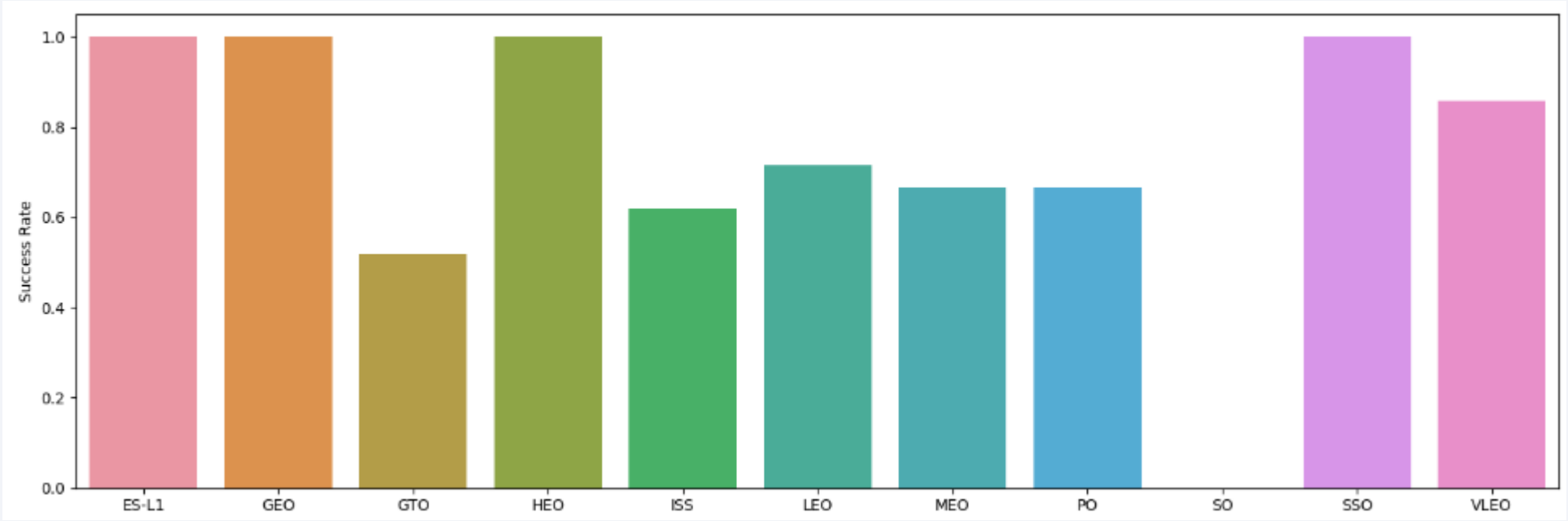
Payload vs. Launch Site

- Under 8000 kg concentrated the majority of the launches.
- Over 9000 kg the launches has the best ratio of successful.



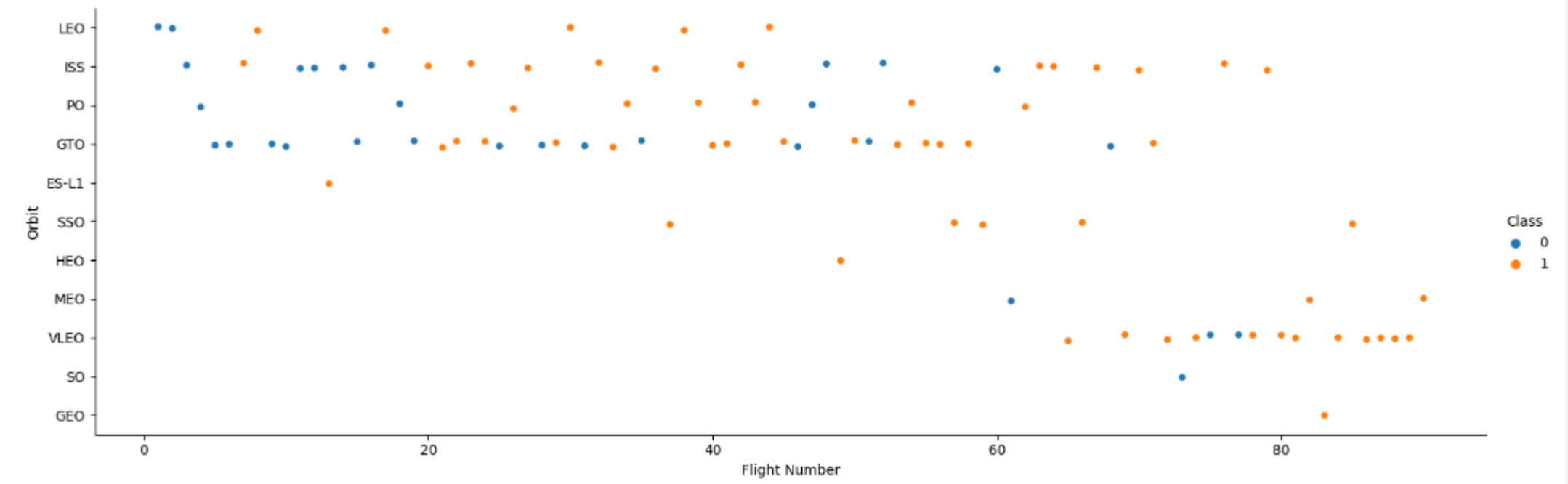
Success Rate vs. Orbit Type

- The biggest success ratios happens to orbits ES-L1, GEO, HEO and SSO.
- The lowest success ratios happens to orbit GTO.



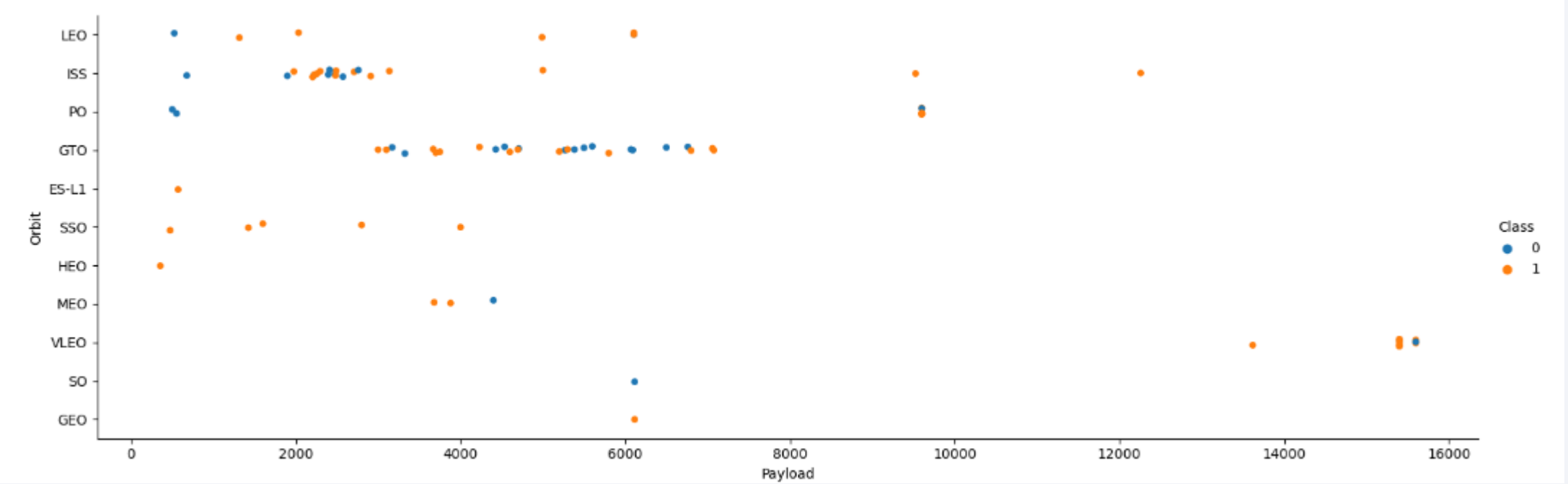
Flight Number vs. Orbit Type

- The lowest success ratios happens to orbits GTO and ISS
- The lowest success ratios happens in the beginning of the Fligth numbers, After the 40 Launches the ratio of successful grow up.



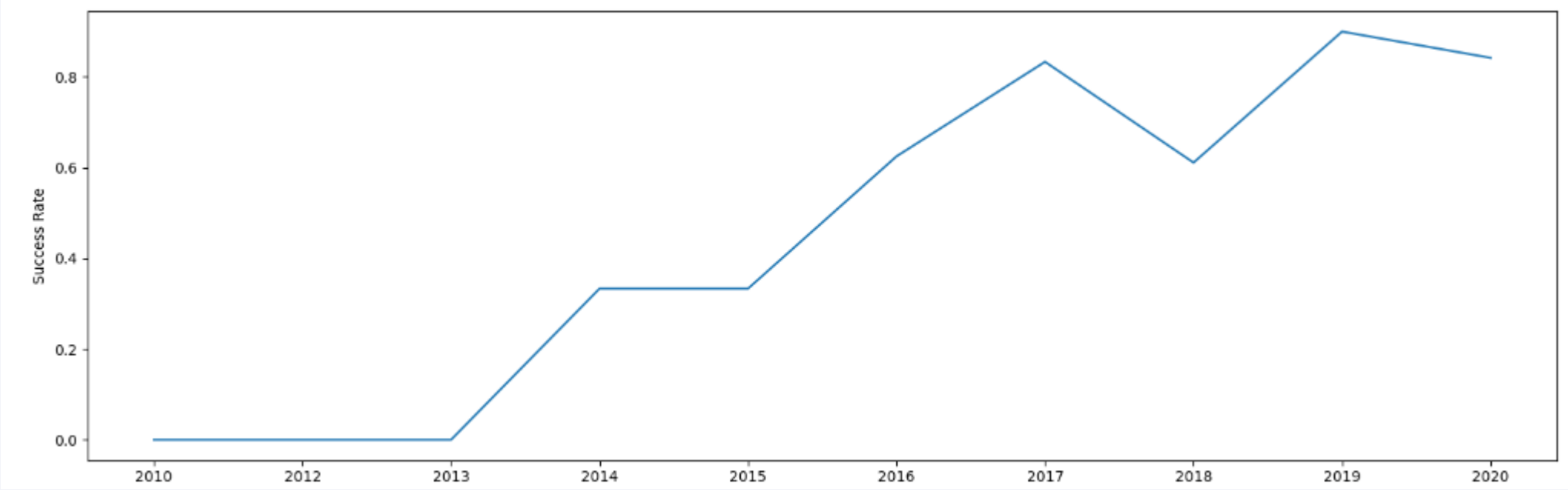
Payload vs. Orbit Type

- Over 9000 kg the launches has the best ratio of successful.
- Apparently, there is no relation between payload and success rate to orbit GTO.
- There are one launch to the orbits SO and GEO.



Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020.
- the first 3 years failed the vast majority of launches.
- In year 2018 was the only down on the success rate.



All Launch Site Names

- There are four launchsites
- The command DISTINCT show only unique launch sites from the SpaceX data.

```
In [9]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[9]:

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

Launch Site Names Begin with 'CCA'

- We used querys like WHERE , LIKE and LIMIT

In [32]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[32]:

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

- We used querys like SUM (summarized) AS (give a name to SUM variable),, WHERE, LIKE
- The total payload mass carried by boosters launches by NASA (CRS) was 48123 KG

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

In [37]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE CUSTOMER LIKE '%CRS%';
```

```
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[37]:

total_payload
48213

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928 KG
- We used querys like AVG (Average) AS (give a name to AVG variable), FROM , WHERE

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

In [34]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[34]:

avg_payload
2928

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad was 2015-12-22
- We used querys like MIN(Minimum) AS (give a name to MIN variable), FROM , WHERE

```
In [38]: %sql SELECT MIN(DATE) AS FIRST_SUCCESFUL_LANDING FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
Out[38]: first_succesful_landing
         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 was F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022 and F9 FT B1026
- We used querys like DISTINCT (show only unique variable), FROM , WHERE, BETWEEN (used to test whether an expression is within a range of values)

In [39]:
```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';
```

```
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[39]:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful is 99 and failure mission outcomes is 1
- We used querys like COUNT() AS (give a name to MIN variable), FROM , GROUP BY

List the total number of successful and failure mission outcomes

In [42]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[42]:

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- We used subquerys and querys like DISTINCT (show only unique variable), FROM , WHERE, BETWEEN (used to test whether an expression is within a range of values)
- The names of the booster which have carried the maximum payload mass is:

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

In [46]:
```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[46]:

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

2015 Launch Records

- We used subquerys and querys like TO_DATE, WHERE
- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 is:

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [88]:
```
#%sql SELECT DATE , LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE Landing__Outcome = 'Failure (drone ship)' AND YEAR(DATE) = 201
#%sql SELECT substr(Date, 4, 3) , LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE Landing__Outcome = 'Failure (drone ship)' AND YE
%sql  SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE Landing
```

```
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

Out[88]:

month_name	landing__outcome	booster_version	launch_site
JANUARY	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
APRIL	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order is:

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

In [92]:
```
%sql SELECT DATE, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND LANDING__OUTCOME LIKE '%
```

```
 * ibm_db_sa://tpn93678:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/BLUDB
Done.
```

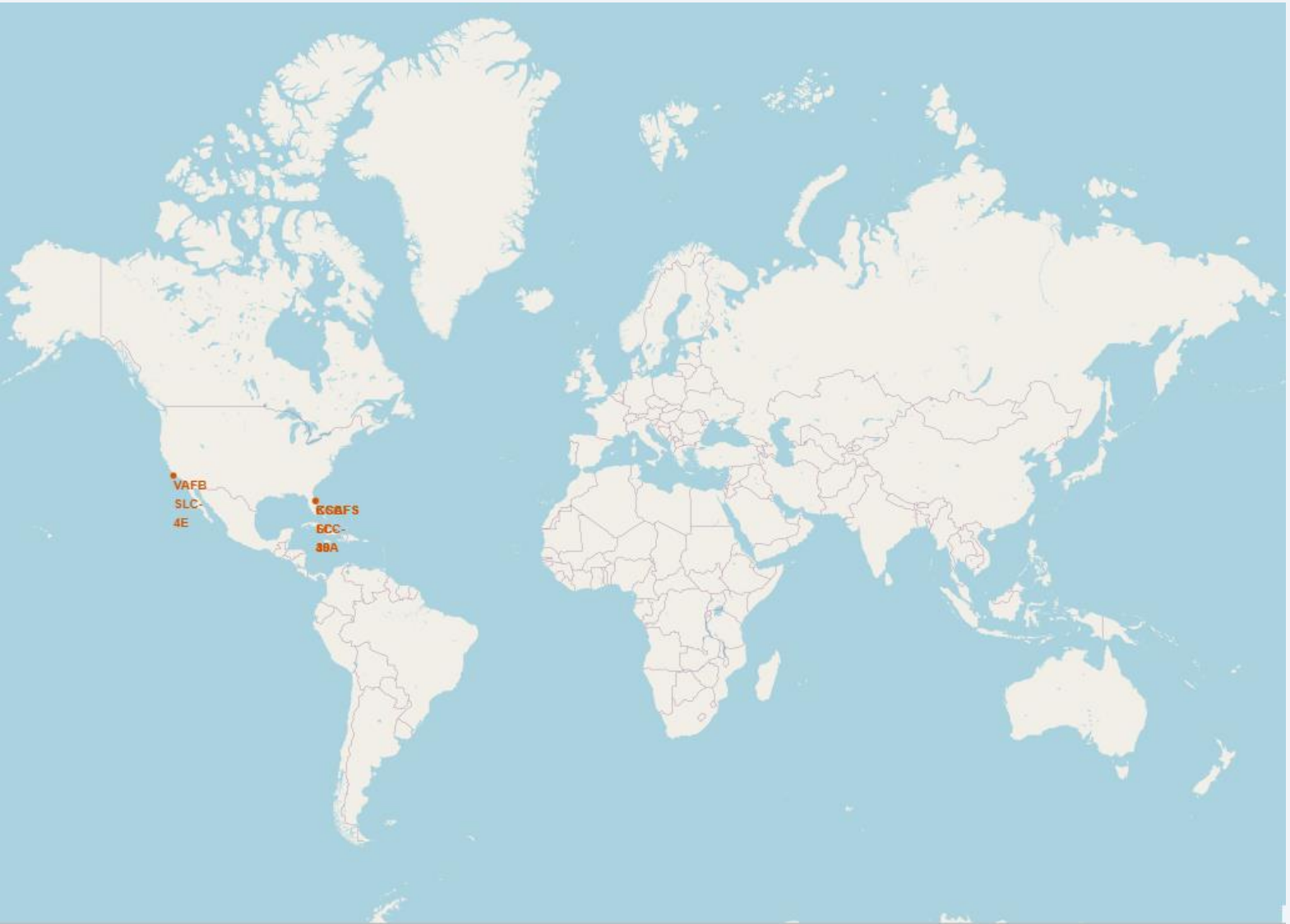
Out[92]:

DATE	total_number
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

Section 3

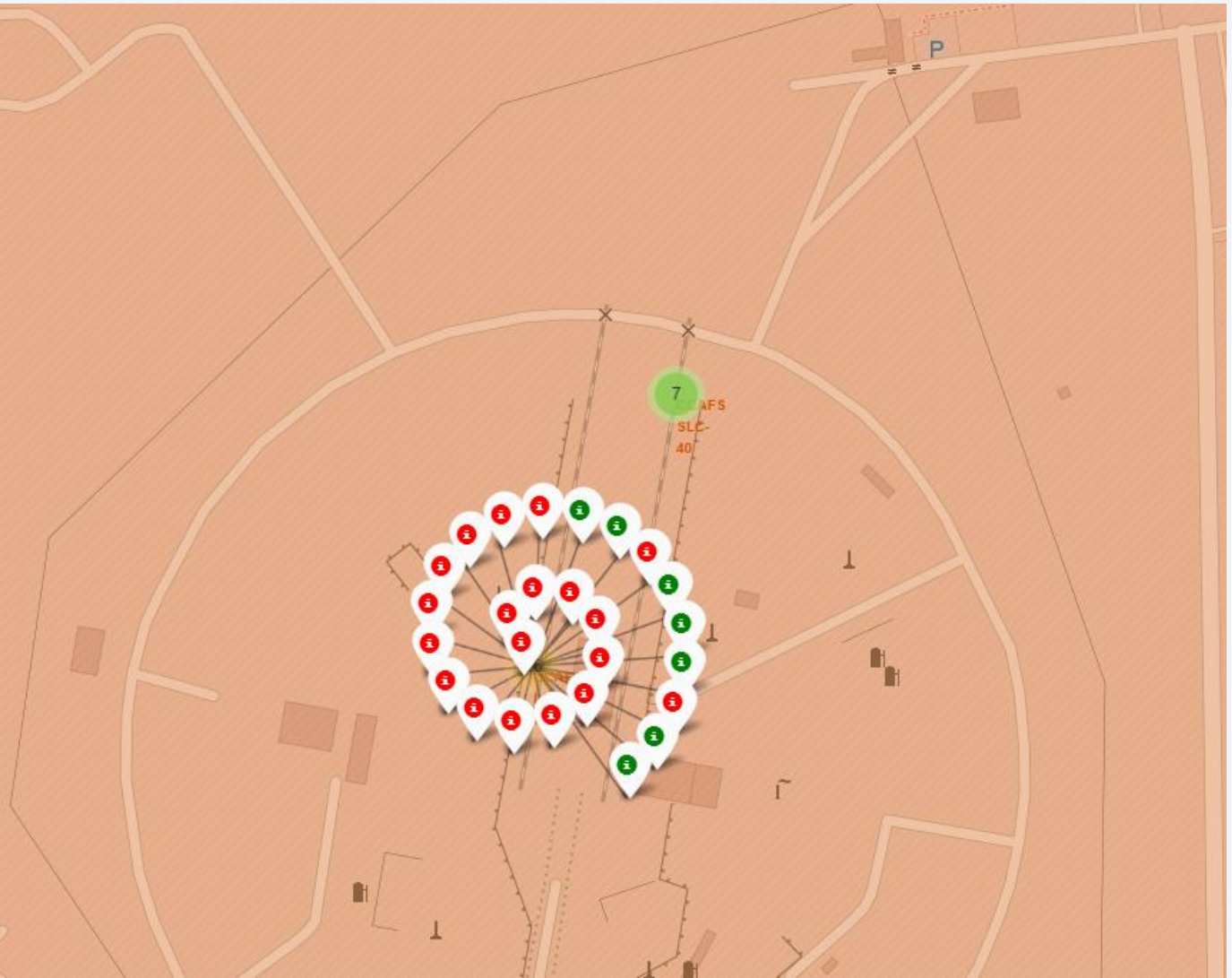
Launch Sites Proximities Analysis

All launch sites



- Lauch cities are near the sea.
- All launch sites are in USA.

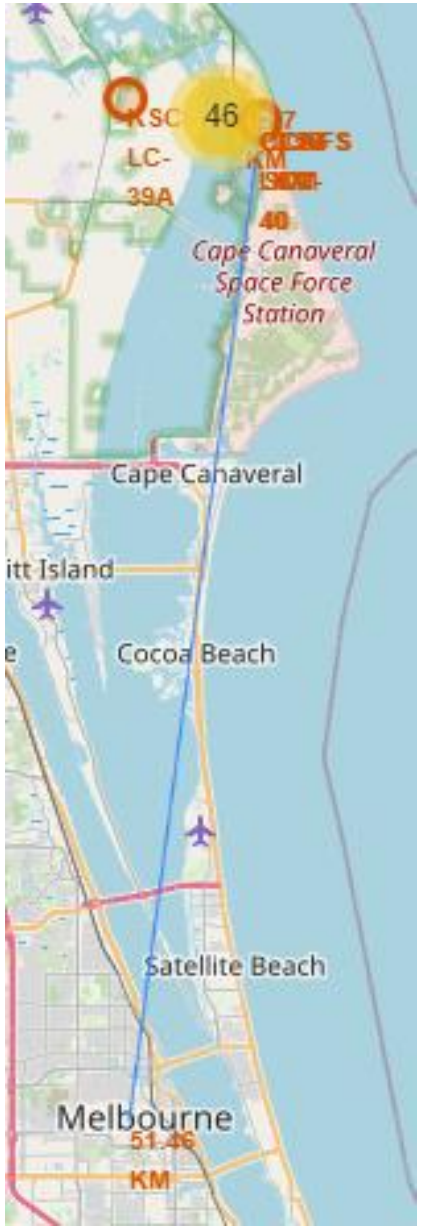
Markers showing launch sites with color labels

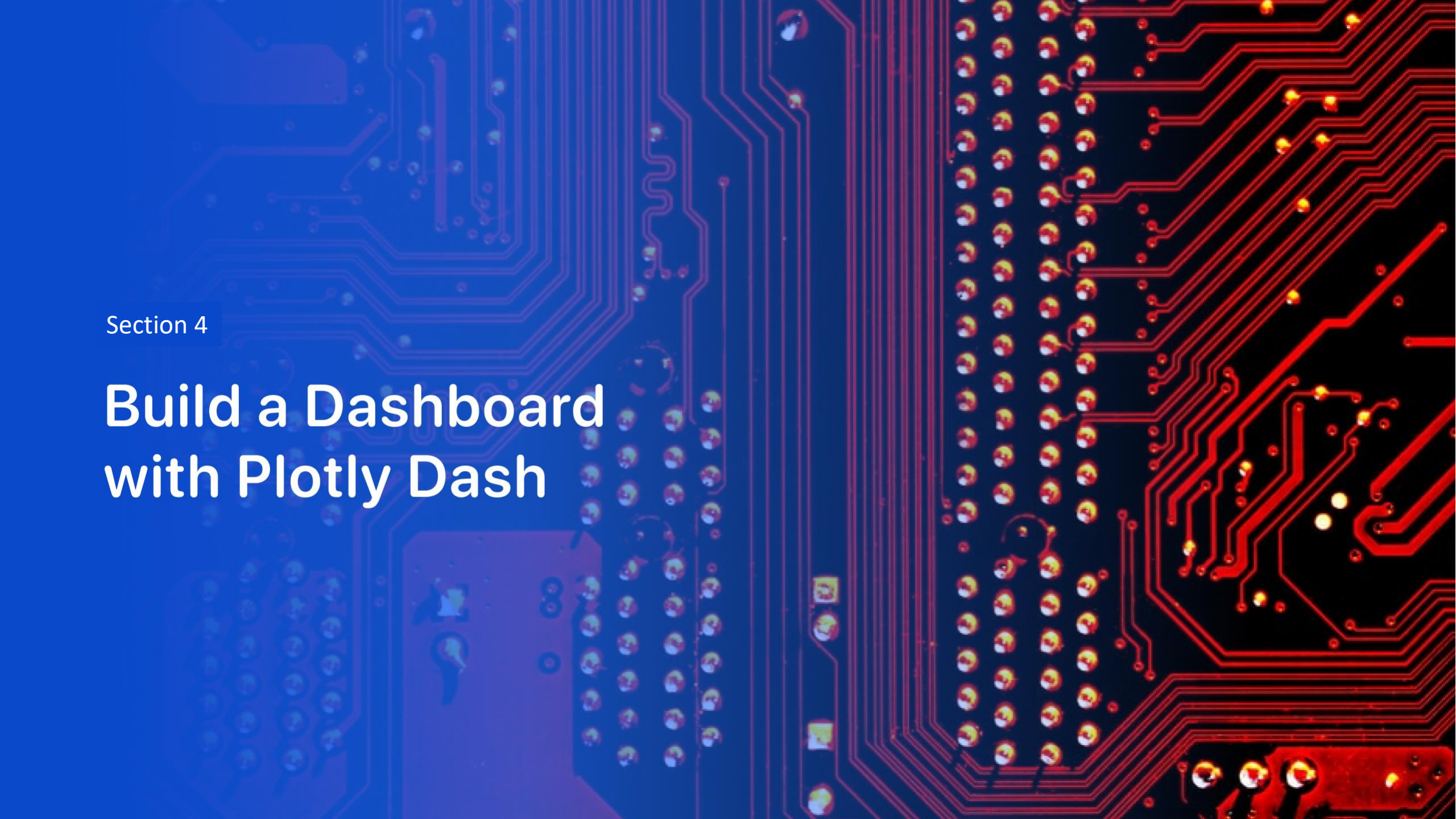


- Example of CCAFS SLC 40 and CCAFS LC 40 launch sites and launch outcomes.
- Green Marker shows successful Launches and Red Marker shows Failures.

Launch Site distance to landmarks

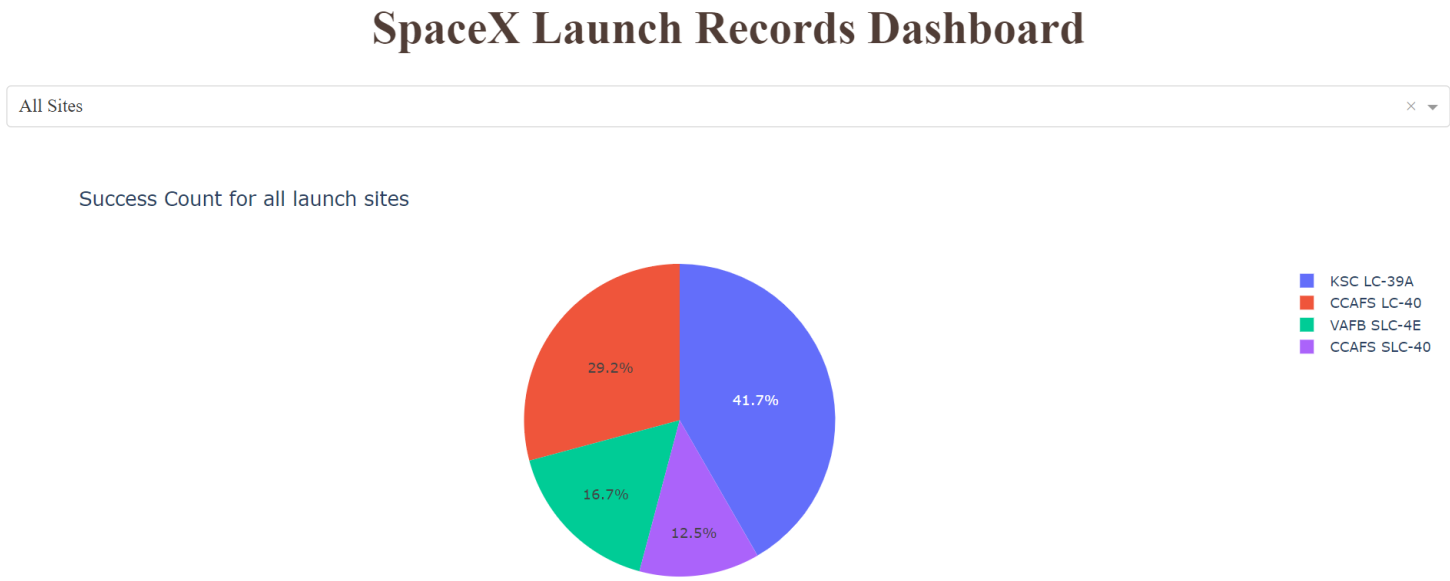
In CCAFS SLC 40 and CCAFS LC 40 launch sites to its proximities such as railway, highway, coastline, with distance calculated and displayed




Section 4
Build a Dashboard with Plotly Dash

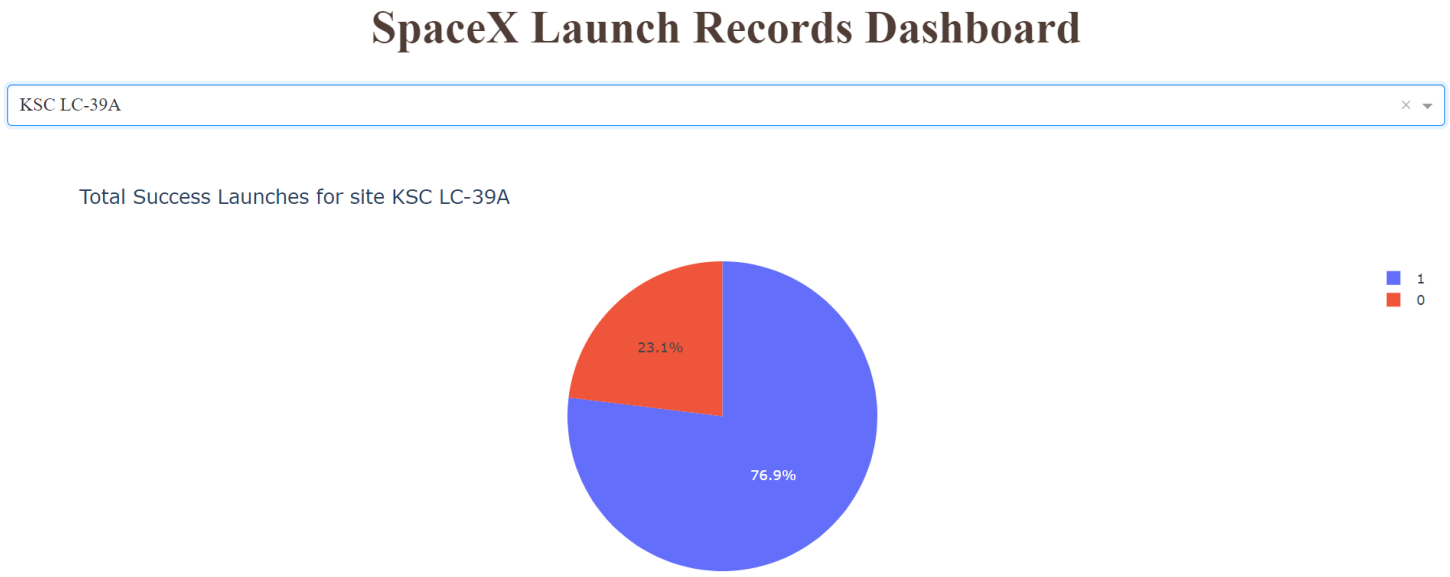
Launch success count for all sites

- In the figure show launch success count for all sites, in a pie chart.
- The importance of the graph shown is that we can make a quick comparison of which launch site has the highest and lowest launch success.



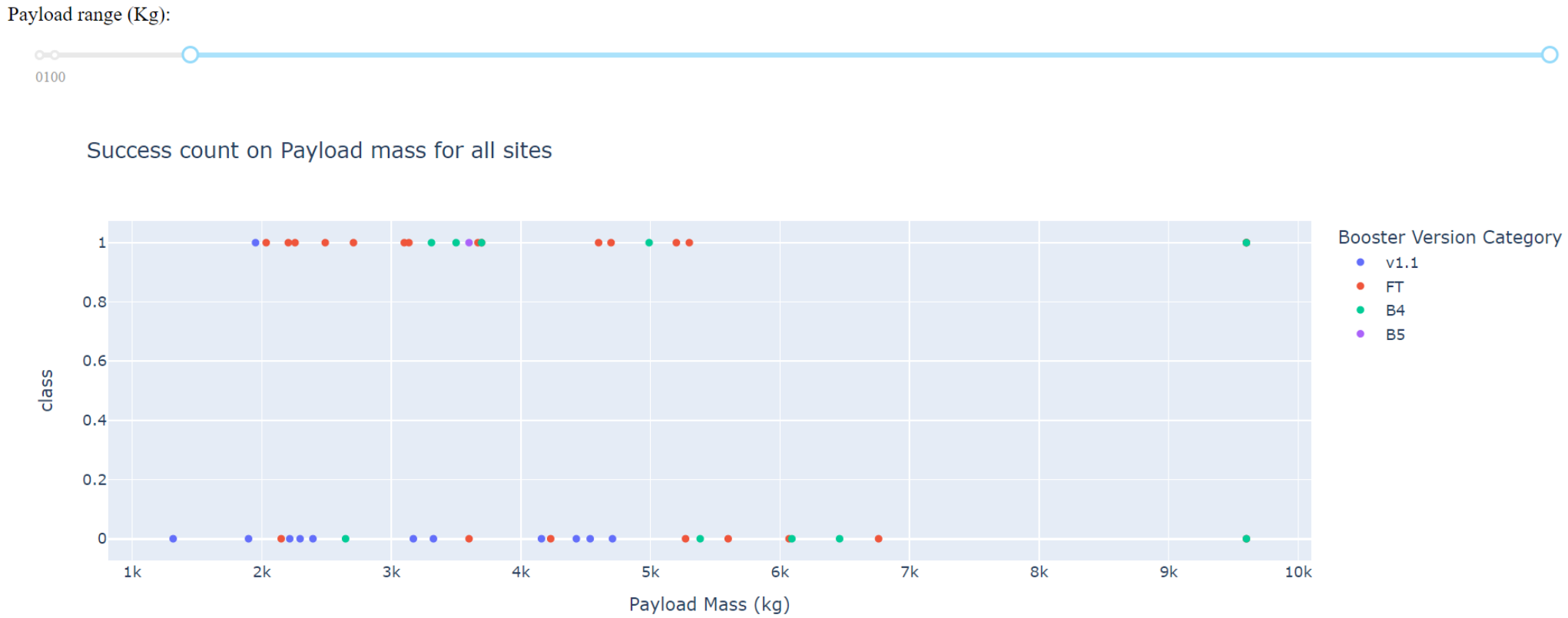
Highest Launch success is KSCLC-39A

- In the figure show launch success count for site KSCLC-39A, in a pie chart.
- The graph shows the percentage of successful launch (1) vs failed launch (0). This launch site is the one with the highest success rate with 76.9%



Payload vs. Launch Outcome

- In the figure show a Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- The importance of this figure is that we can conclude that the vast majority of rockets to have a successful launch must weigh less than 6000 kg. It is also appreciated that the most successful combination is the FT boosters

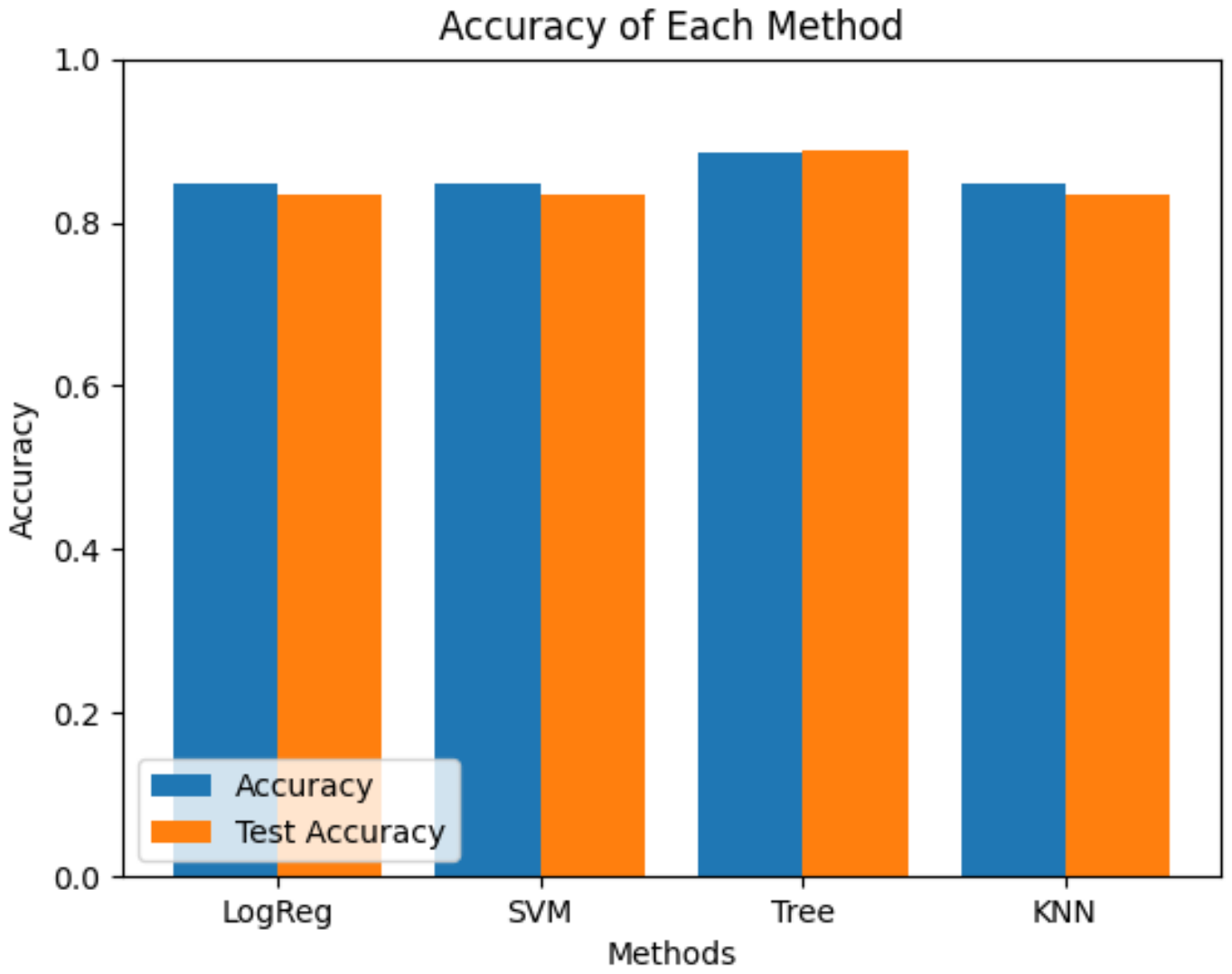


Section 5

Predictive Analysis (Classification)

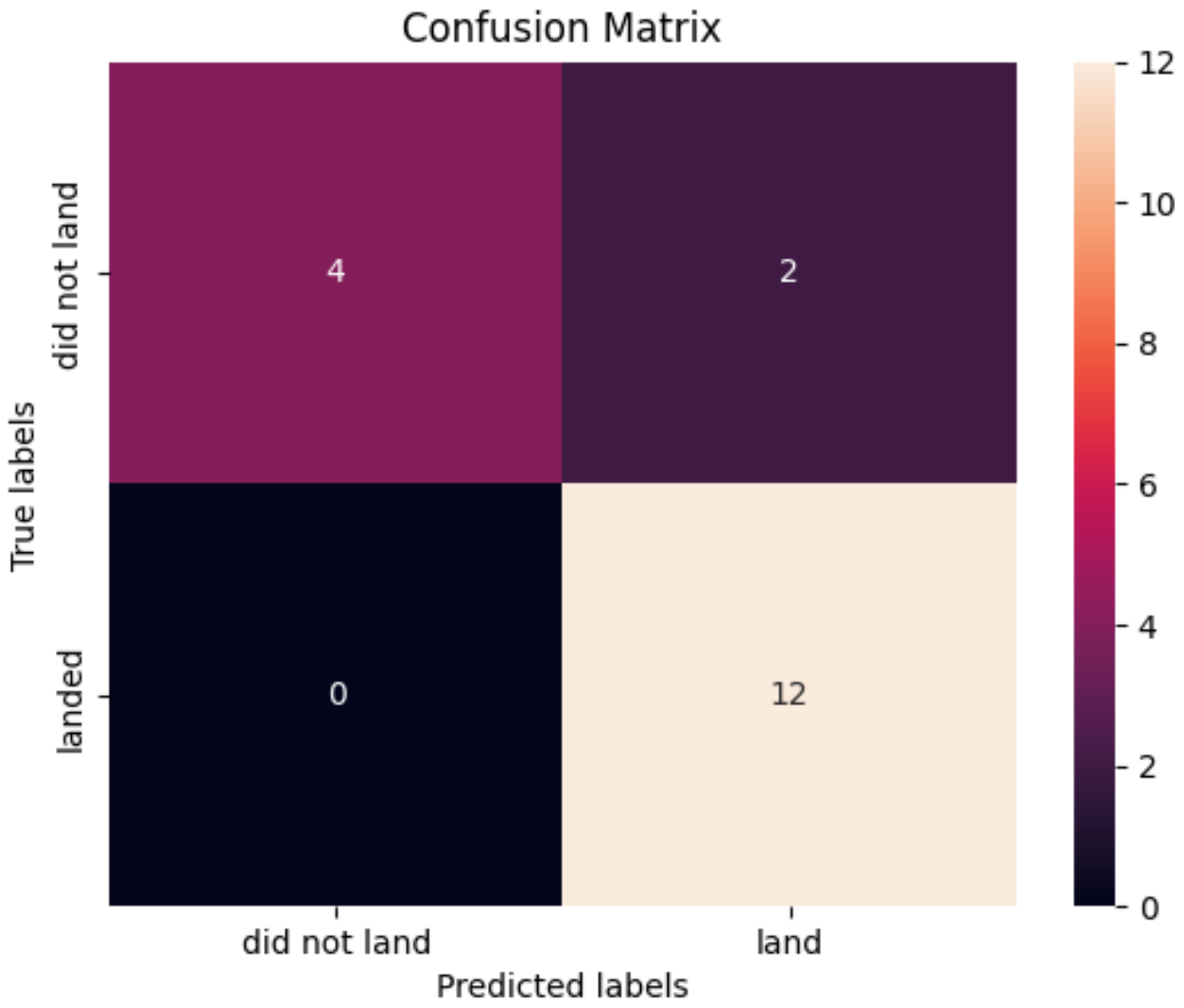
Classification Accuracy

- Here we visualize the built model accuracy for all built classification models, in a bar chart
- The tree model has 88% of accuracy, this is the highest classification accuracy



Confusion Matrix

- Confusion matrix of Decision Tree Classifier give us the big accuracy, showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- We can conclude that:
- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!