IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Trinh Bao Khanh Huyen
19 Feb 2024

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

- Project background and context

  Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollar.
  Other providers cost upward of 165 million dollars each
  Much of the savings is because Space X can reuse the first stage.
  If it is possible to predict the outcome of first stage, the cost of a launch is determined. It is useful if an alternate company wants to bid against space X for a rocket launch.
  This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully

- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web scraping: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
  - Requesting rocket launch data from SpaceX API: "https://api.spacexdata.com/v4/launches/past"
- Perform data wrangling:
  - Collect data and enriched using outcome label based on outcome data.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - At this step, data is normalized, divided into training and test data sets.
  - These data set are tested on 4 different models to find the method performs best using test data.
  - Different parameters are used to find the most suited model.

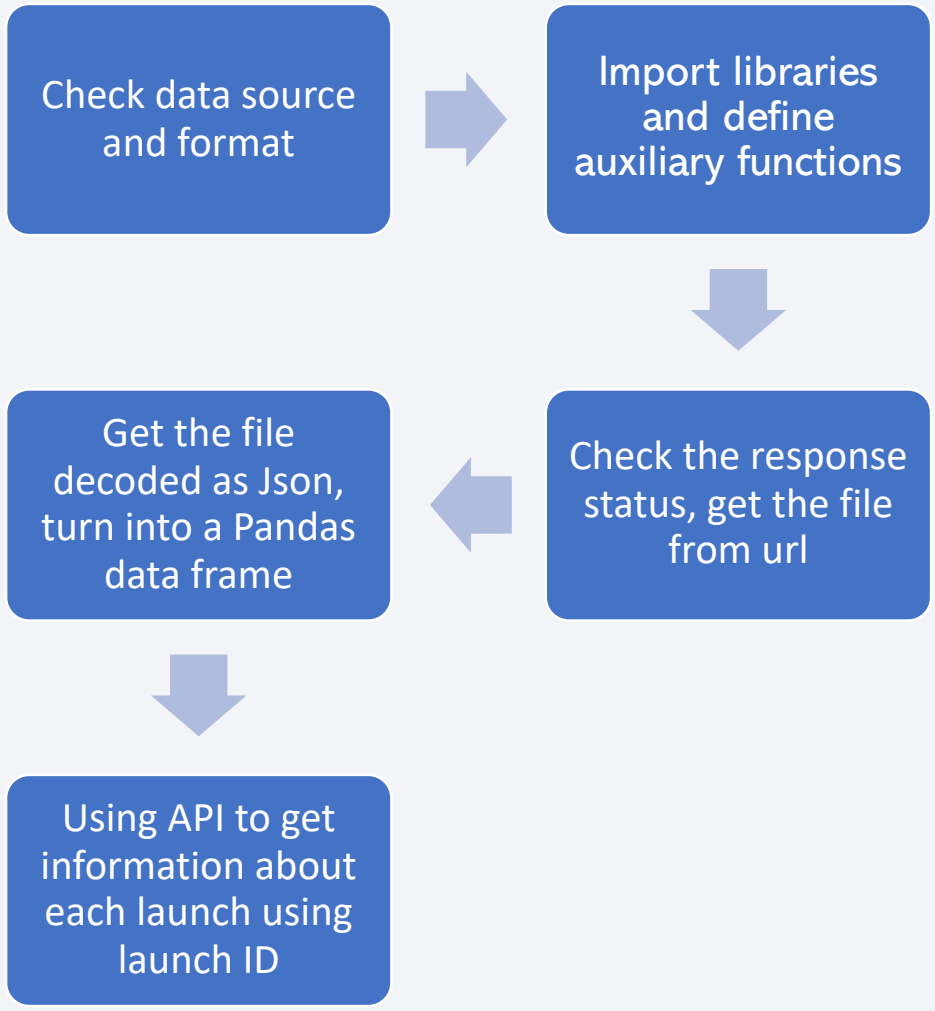
Data Collection

- Describe how data sets were collected.

Datasets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

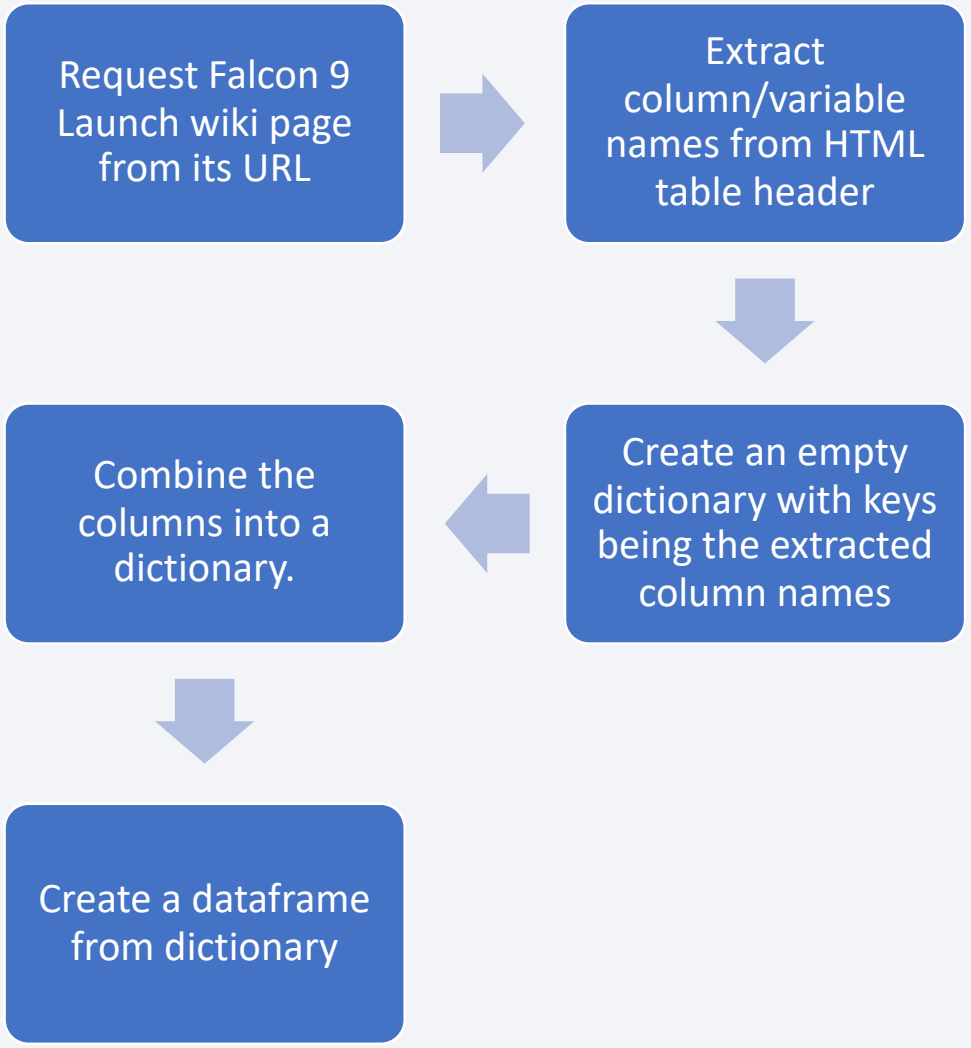
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;
- • This API was used according to the flowchart beside and then data is persisted.
- GitHub URL: https://github.com/JennTrinh/Data-Sci--Coursera---Capstone/blob/a5f256a2454bf095fdc4bf3fac6fb987ccbdf180/Wk1-labs-spacex-data-collection-api.ipynb



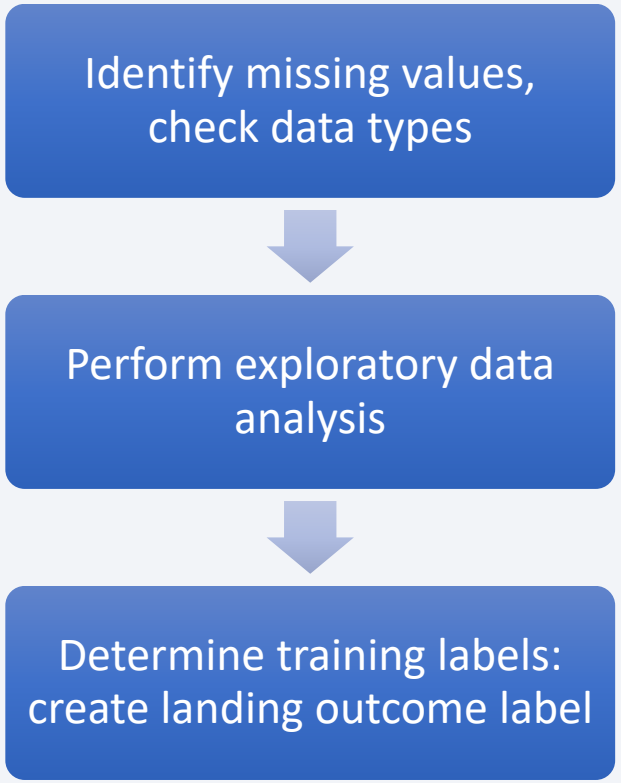
Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- GitHub URL: https://github.com/JennTrinh/Data-Sci--Coursera---Capstone/blob/a5f256a2454bf095fdc4bf3fac6fb987ccbdf180/Wk1-labs-webscraping.ipynb



Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summary launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.
- GitHub URL: https://github.com/JennTrinh/Data-Sci--Coursera---Capstone/blob/a5f256a2454bf095fdc4bf3fac6fb987ccbdf180/Wk1-labs-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts:
  - Scatter point chart with Flight Number and the launch site: Visualize the relationship between Flight Number and Launch Site.
  - Scatter point chart with Pay Load Mass (kg) and the launch site: observe if there is any relationship between launch sites and their payload mass.
  - Bar chart for the success rate of each orbit: visually check if there are any relationship between success rate and orbit type to find which orbits have high success rate.
  - Scatter point chart with Flight Number and the Orbit: check if there is any relationship between Flight Number and Orbit type.
  - Scatter point chart with Payload and the Orbit: reveal the relationship between Payload and Orbit type.
  - Line chart with year and the success rate: to get the average launch success trend.
- GitHub URL: https://github.com/JennTrinh/Data-Sci--Coursera---Capstone/blob/a5f256a2454bf095fdc4bf3fac6fb987ccbdf180/Wk2-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begins with the string 'CCA’
  - Total pay load mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Success/Failed landing out comes in droneship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (droneship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20
- GitHub URL: https://github.com/JennTrinh/Data-Sci--Coursera---Capstone/blob/a5f256a2454bf095fdc4bf3fac6fb987ccbdf180/Wk2-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Mark the launch sites on map:

  - Adding folium.Marker to mark the area and adding folium.Circle to add a highlighted circle area with a text label on a specific coordinate.

- Mark the success / failure launches for each site on map:

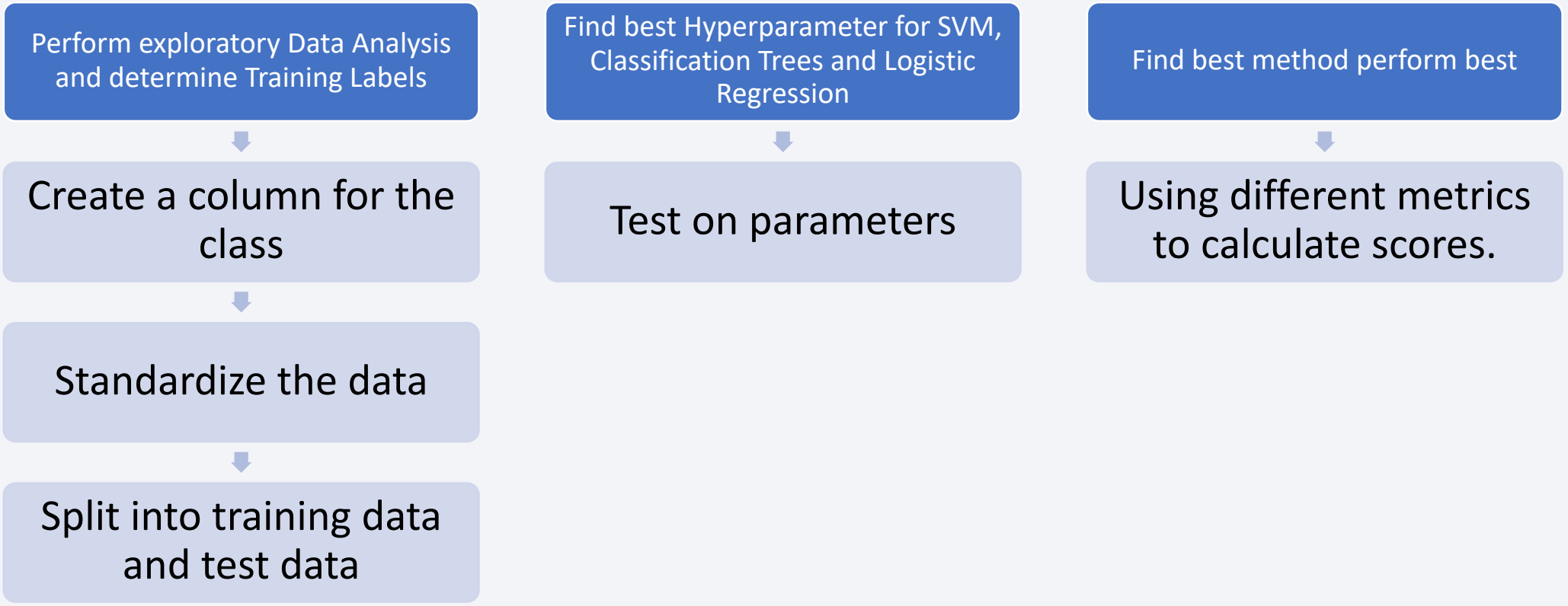
  - Using the green marker for success and red marker for failure launches.
  - Creating a MarkerCluster object to simplify a map containing many markers having the same coordinate.
  - Adding a folium.Marker to marker cluster object for each launch site result to easily identify which launch sites have relatively high success rates.

- Github url: https://github.com/JennTrinh/Data-Sci--Coursera---Capstone/blob/a5f256a2454bf095fdc4bf3fac6fb987ccbdf180/Wk3-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard:
  - Pie chart to show the total successful launches count for all sites and selected site: the proportion of success rate from all launch sites to observe which launch site counted largest percentage as well as the success rate within the selected launch site.
  - Scatter chart to show the correlation between payload and launch success: whether there is a relationship between payload and success rate of booster versions and which booster version perform best.
- Add the GitHub URL: https://github.com/JennTrinh/Data-Sci--Coursera---Capstone/blob/a5f256a2454bf095fdc4bf3fac6fb987ccbdf180/Wk3-spacex_dash_app.py

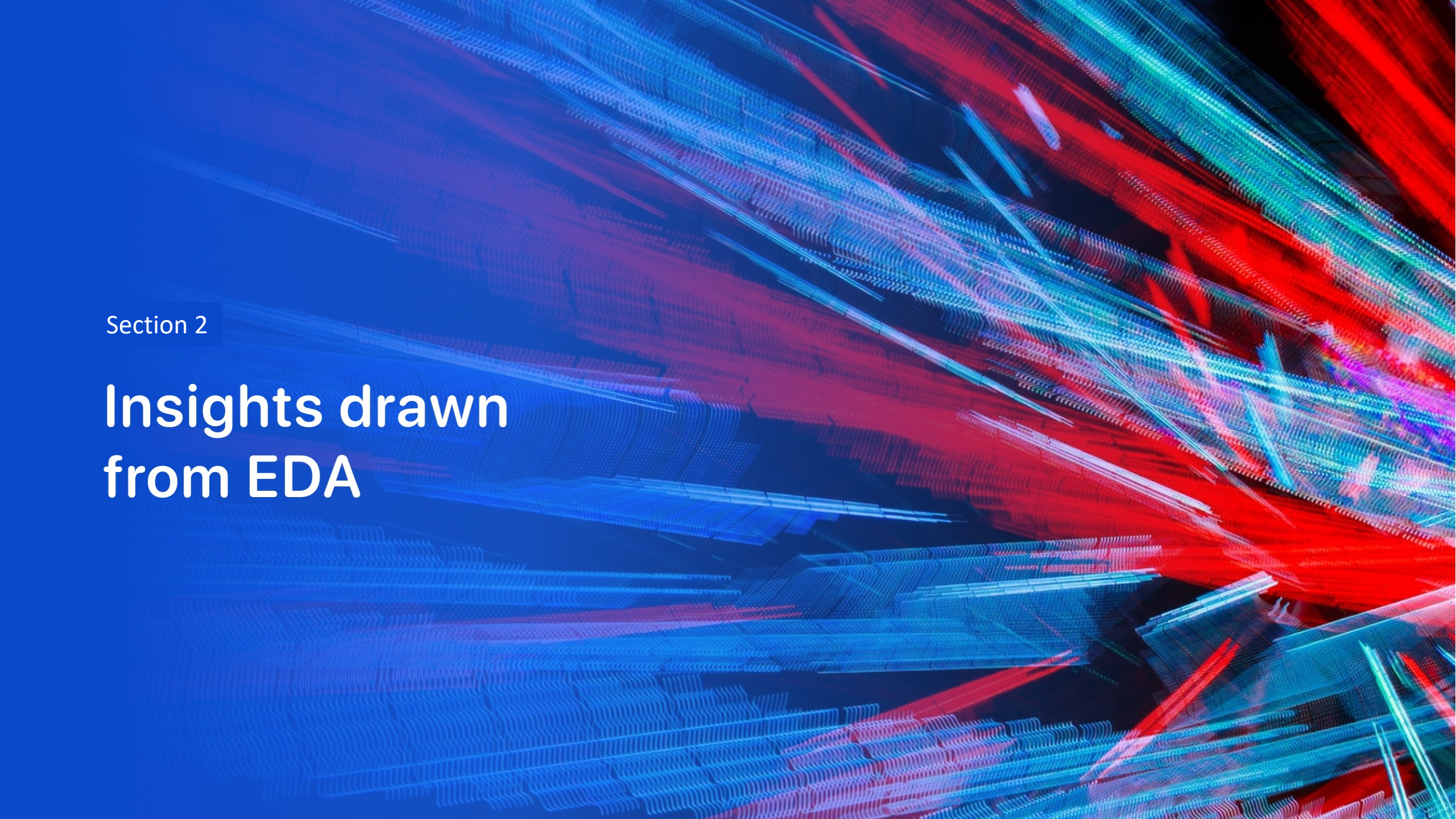
Predictive Analysis (Classification)



- GitHub URL: https://github.com/JennTrinh/Data-Sci--Coursera---Capstone/blob/a5f256a2454bf095fdc4bf3fac6fb987ccbdf180/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

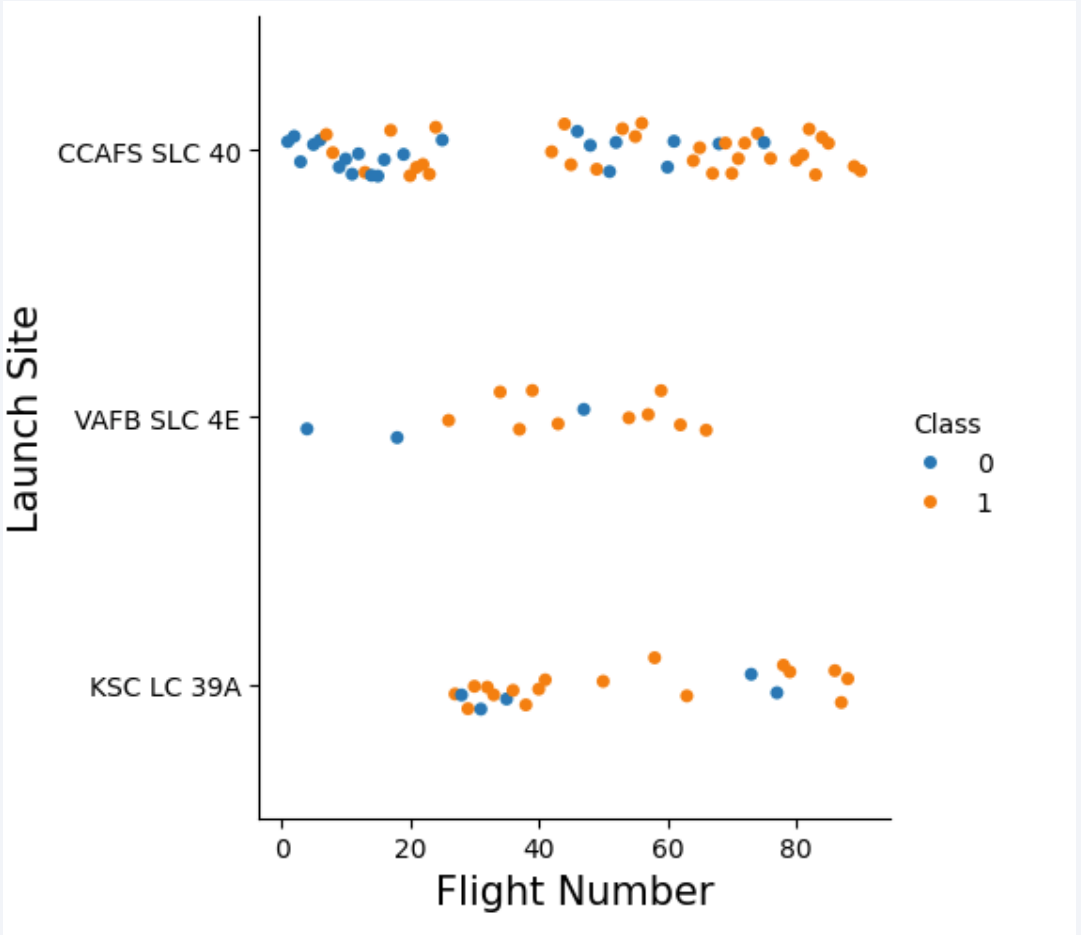
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Discovery if there is any relationship between the success rate, flight number and launch site.

After flight number 40, there are higher success rate in all 3 launch sites.

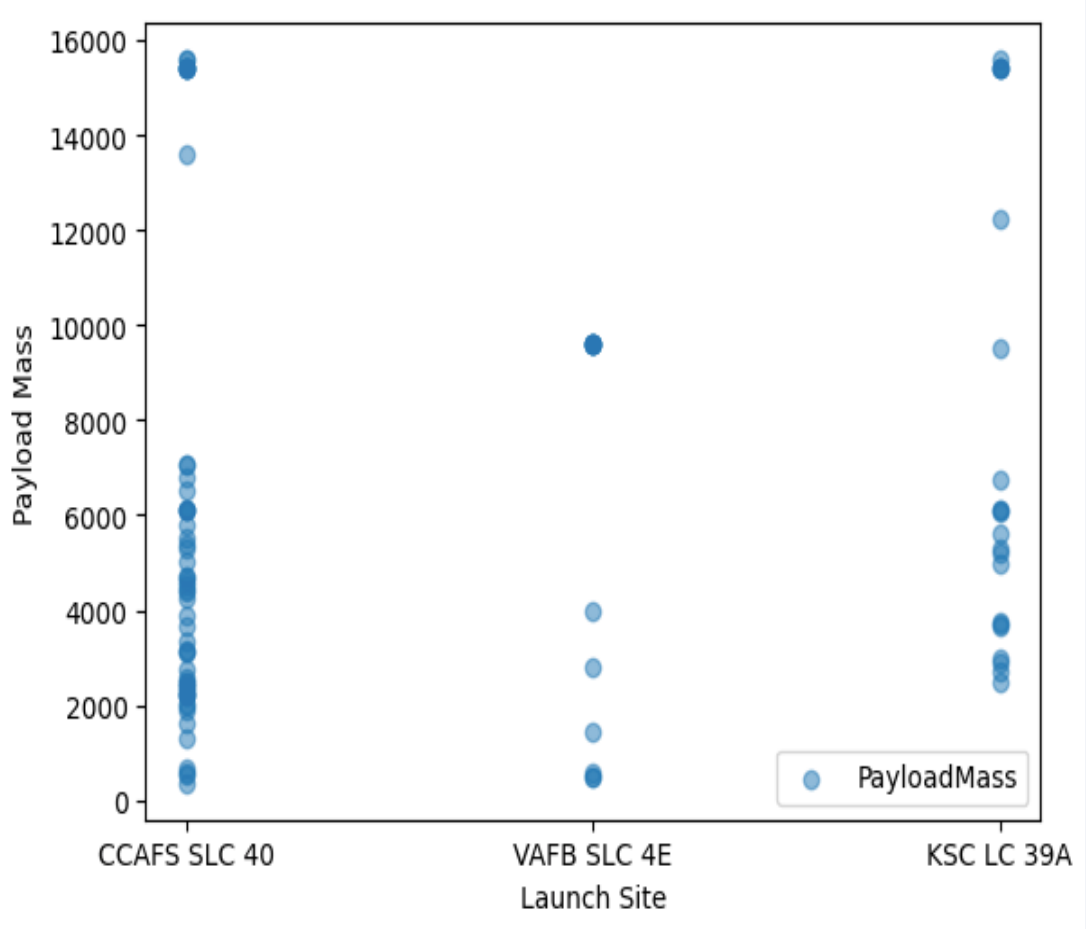


Relationship between Flight Number and Launch Site

Payload vs. Launch Site

Discovery if there is any relationship between launch sites and their payload mass

At the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (>10000)

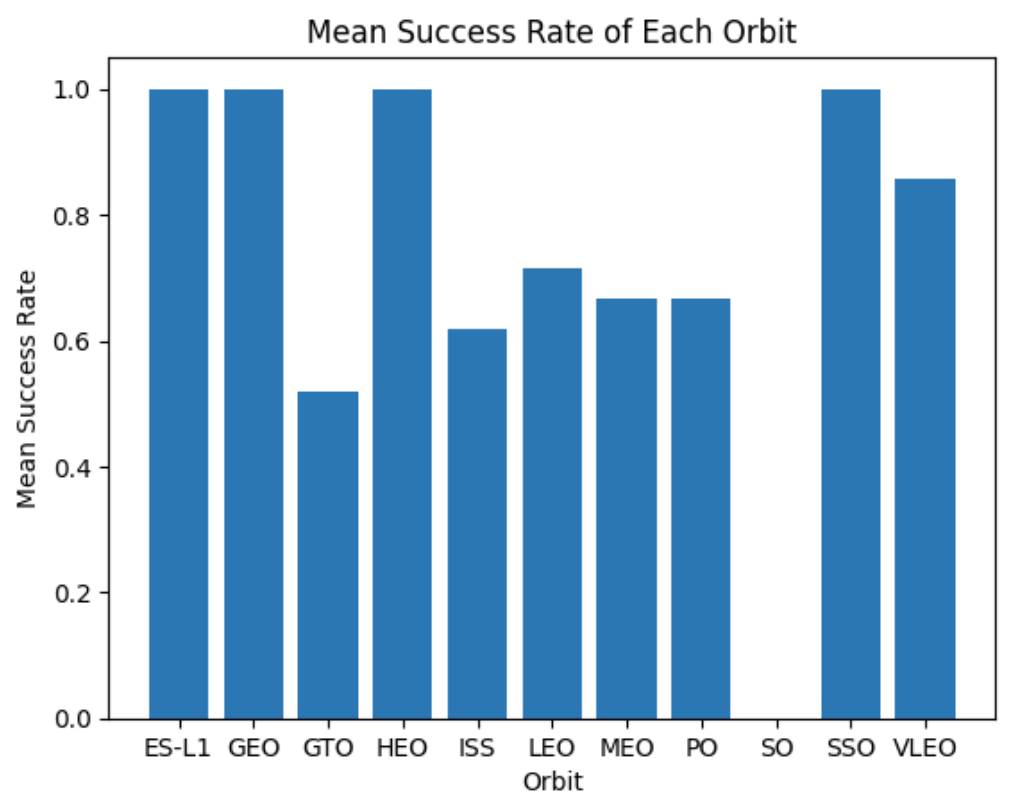


Relationship between Payload and Launch Site

Success Rate vs. Orbit Type

Discovery if there is any relationship between success rate and orbit type

SO orbit has no success rate, orbits having the highest success were ES-L1, HEO, SSO and VLEO

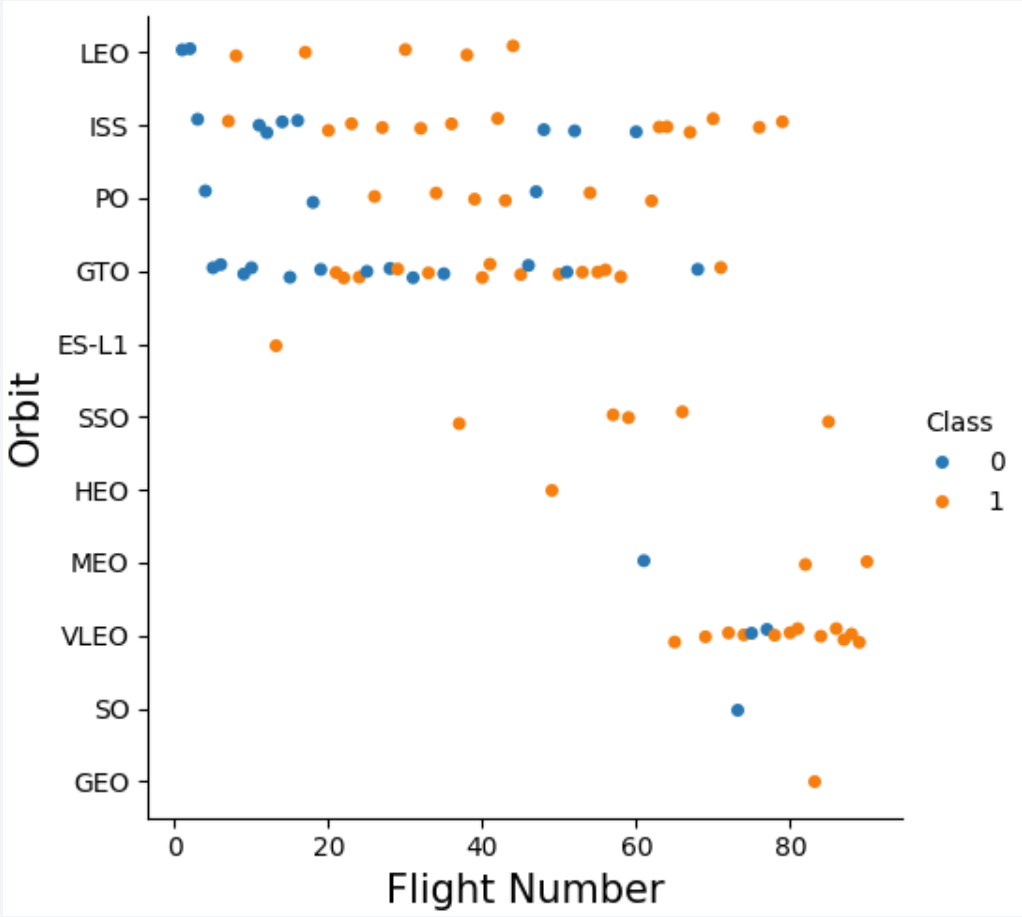


Relationship between Success rate and Orbit type

Flight Number vs. Orbit Type

Discovery if there is any relationship between Flight number and Orbit type

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

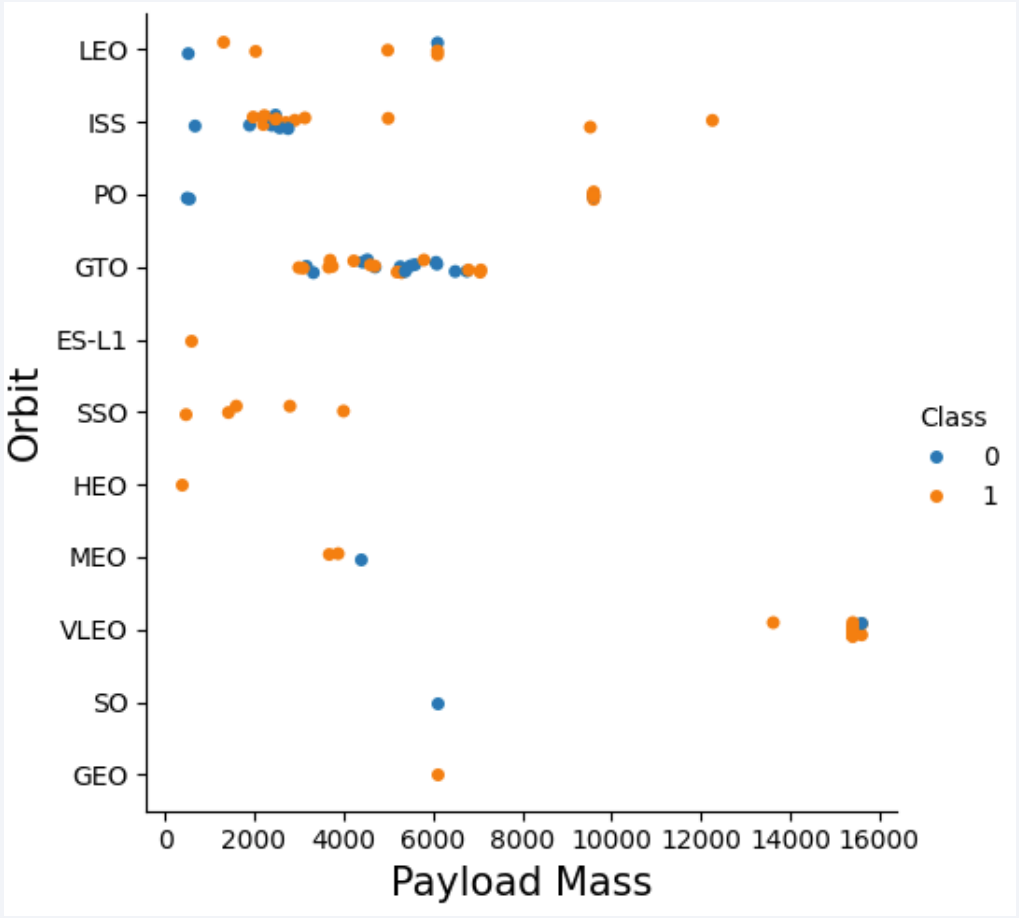


Relationship between Flight number and Orbit type

Payload vs. Orbit Type

Discovery if there is any relationship between Payload and Orbit type

With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
At the GTO launch site, both success and unsuccessful mission cannot be distinguished well.

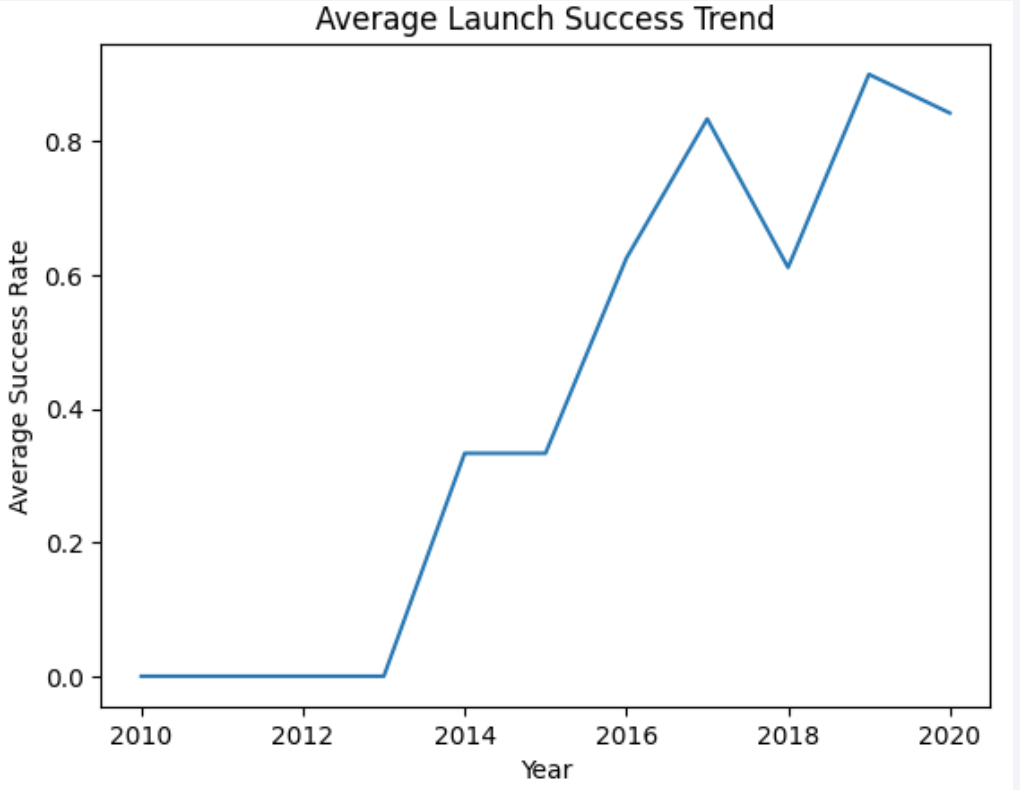


The relationship between Payload and Orbit type

Launch Success Yearly Trend

Visualizing trending of the launch success yearly

There was a short increasing in success rate from 2013 to 2014.
Success rate increased again in 2015 after stabilizing in 2014 and kept the trend till 2017.



Visualization of average launch success trend yearly

All Launch Site Names

- Get all the launch site name using **distinct combine with select** function

%sql select distinct Launch_Site FROM SPACEXTABLE;

```
%sql select distinct Launch_Site from SPACEXTABLE;
```

Running query in 'sqlite:///my_data1.db'

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`: the records are chosen using *** to collect all information from every column**, adding condition **'CCA%'** and **limit 5.**

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5;
```

Running query in 'sqlite:///my_data1.db'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- Calculate the total payload carried by boosters from NASA: calculation is made using **SUM** with the **PAYLOAD_MASS__KG_** column with condition as **Customer = 'NASA (CRS)'**

%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)';

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

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)';
```

Running query in 'sqlite:///my_data1.db'

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1: calculation is made using **avg on PAYLOAD_MASS__KG_** column and condition is added as **where Booster_Version like 'F9 v1.1%'.**

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%';
```

Running query in 'sqlite:///my_data1.db'

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad: get the smallest date using **Min** on **Date** column, with condition **where Landing_Outcome like 'Success (ground pad)'**

```
]:  %sql select Min(Date) from SPACEXTABLE \
        where Landing_Outcome like 'Success (ground pad)';
```

Running query in 'sqlite:///my_data1.db'

]:

Min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000: using **select** function on multiple columns to satisfy the requirement by adding condition **where Landing_Outcome like 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000**

  %sql select Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome like '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

```
%sql select Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome like 'Succ
    and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

Running query in 'sqlite:///my_data1.db'

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes: adding multiple conditions **Mission_Outcome like 'Success%' or Mission_Outcome like 'Fail%' group by Mission_Outcome,** and adding a new column to the result table as **Count** being the result from **count** function of Mission_Outcome column

%sql select Mission_Outcome , count (Mission_Outcome) as Count from SPACEXTABLE where Mission_Outcome like 'Success%' or Mission_Outcome like 'Fail%' group by Mission_Outcome;

```
%sql select Mission_Outcome , count (Mission_Outcome) as Count from SPACEXTABLE where Mission_Outcome like 'Succe
    or Mission_Outcome like 'Fail%' \
        group by Mission_Outcome;
```

Running query in 'sqlite:///my_data1.db'

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass: using the sub query as the condition to combine with the main query after **where** clause

%sql select Booster_Version, max(PAYLOAD_MASS__KG_) as max_payload_mass from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) as max_payload_mass from SPACEXTABLE);

```
%sql select max(PAYLOAD_MASS__KG_) as max_payload_mass, Booster_Version from SPACEXTABLE group by Booster_Versior
```

Running query in 'sqlite:///my_data1.db'

max_payload_mass	Booster_Version
2647	F9 B4 B1039.2
5384	F9 B4 B1040.2
9600	F9 B4 B1041.2
6460	F9 B4 B1043.2
3310	F9 B4 B1039.1
4990	F9 B4 B1040.1
9600	F9 B4 B1041.1
3500	F9 B4 B1042.1
5000	F9 B4 B1043.1
6092	F9 B4 B1044

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.: using multiple conditions and function **substr** to extract the year from **Date** column as part of condition after **where** clause.

%sql select substr(Date, 6,2) as Month, Booster_Version, Launch_Site, Landing_Outcome from SPACEXTABLE where Landing_Outcome like '%Failure (drone ship)%' and substr(Date,0,5) = '2015' ;

```
%sql select substr(Date, 6,2) as Month, Booster_Version, Launch_Site, Landing_Outcome \
    from SPACEXTABLE where \
    Landing_Outcome like '%Failure (drone ship)%' and \
        substr(Date,0,5) = '2015' ;
```

Running query in 'sqlite:///my_data1.db'

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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: using multiple condition in **where** clause and combine **order by, group by** to rank the count in descending order.

```
%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE \
where Date BETWEEN '2010-06-04' AND '2017-03-20' \
     group by Landing_Outcome \
        ORDER BY COUNT(Landing_Outcome) DESC;
```

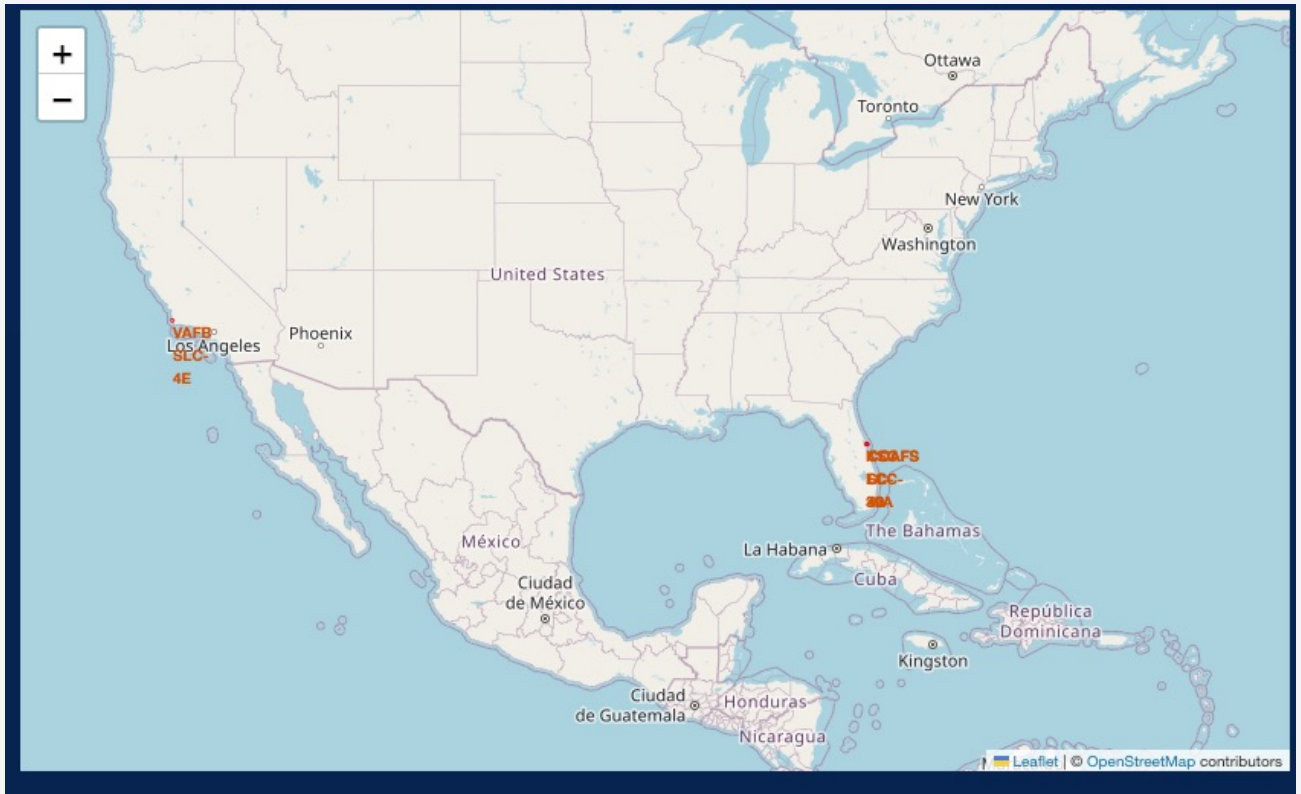
Running query in 'sqlite:///my_data1.db'

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


Section 3
Launch Sites
Proximities Analysis

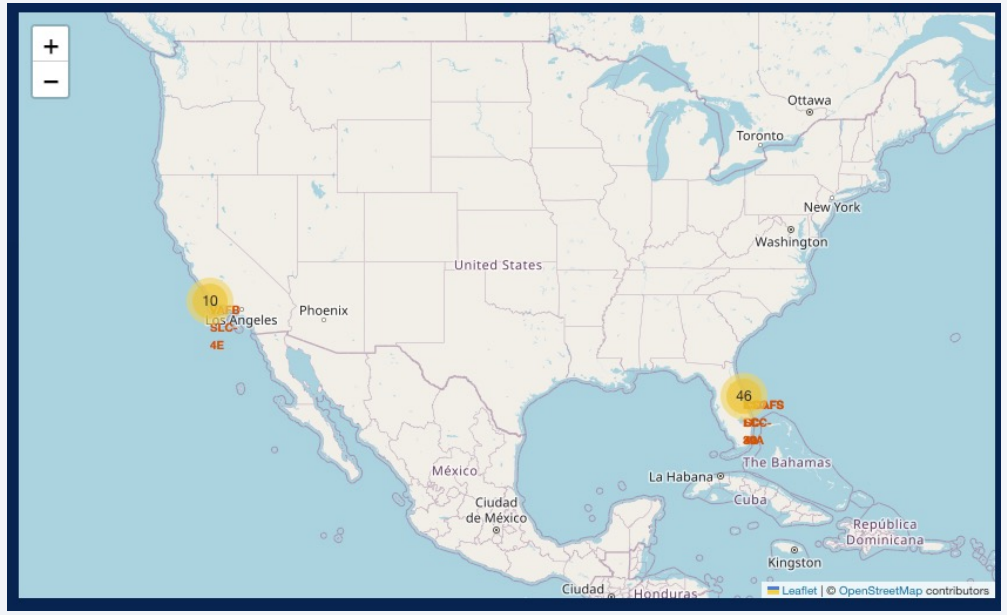
Launch sites marked on Map

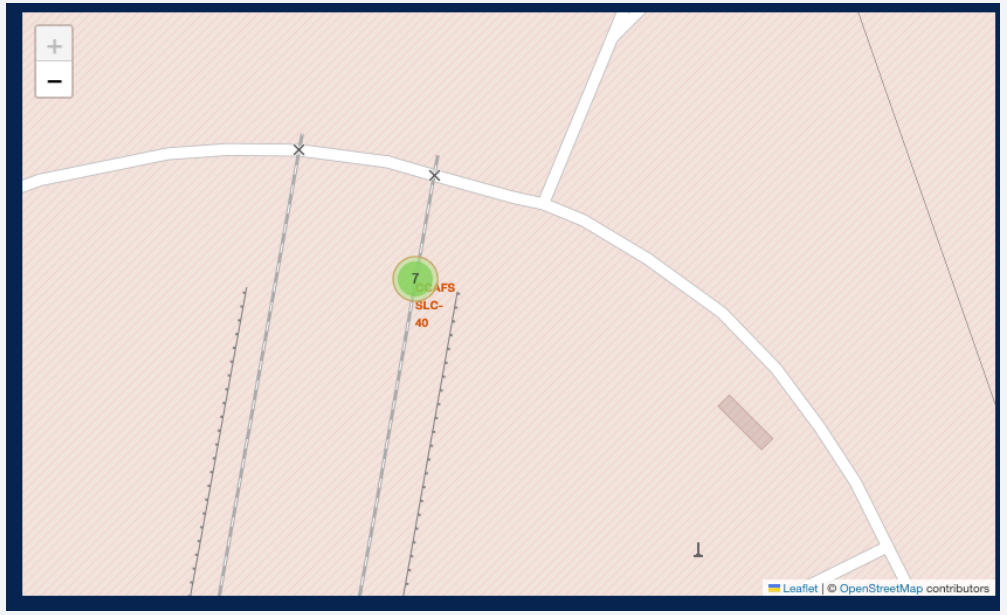
- The locations are marked using red color with names using coordinate from data set, the markers were made by using folium.Marker and folium.Circle.
- From the map, it is possible to observe that all the launch site locations are relatively close to the coast and proximity to the Equator line.



Success/Failed launches for each site on Map

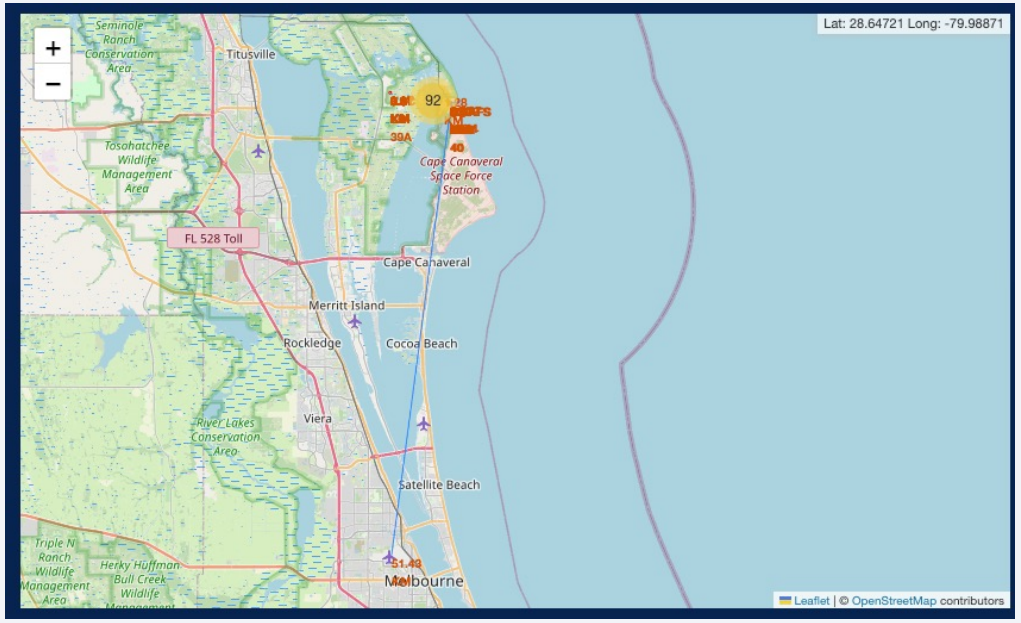
- The launch sites are marked on map along with outcomes of each launch.
- The success launches are marked with green and red represents failed launch.
- From the map, it is possible to get the information of the sites having high success rates.

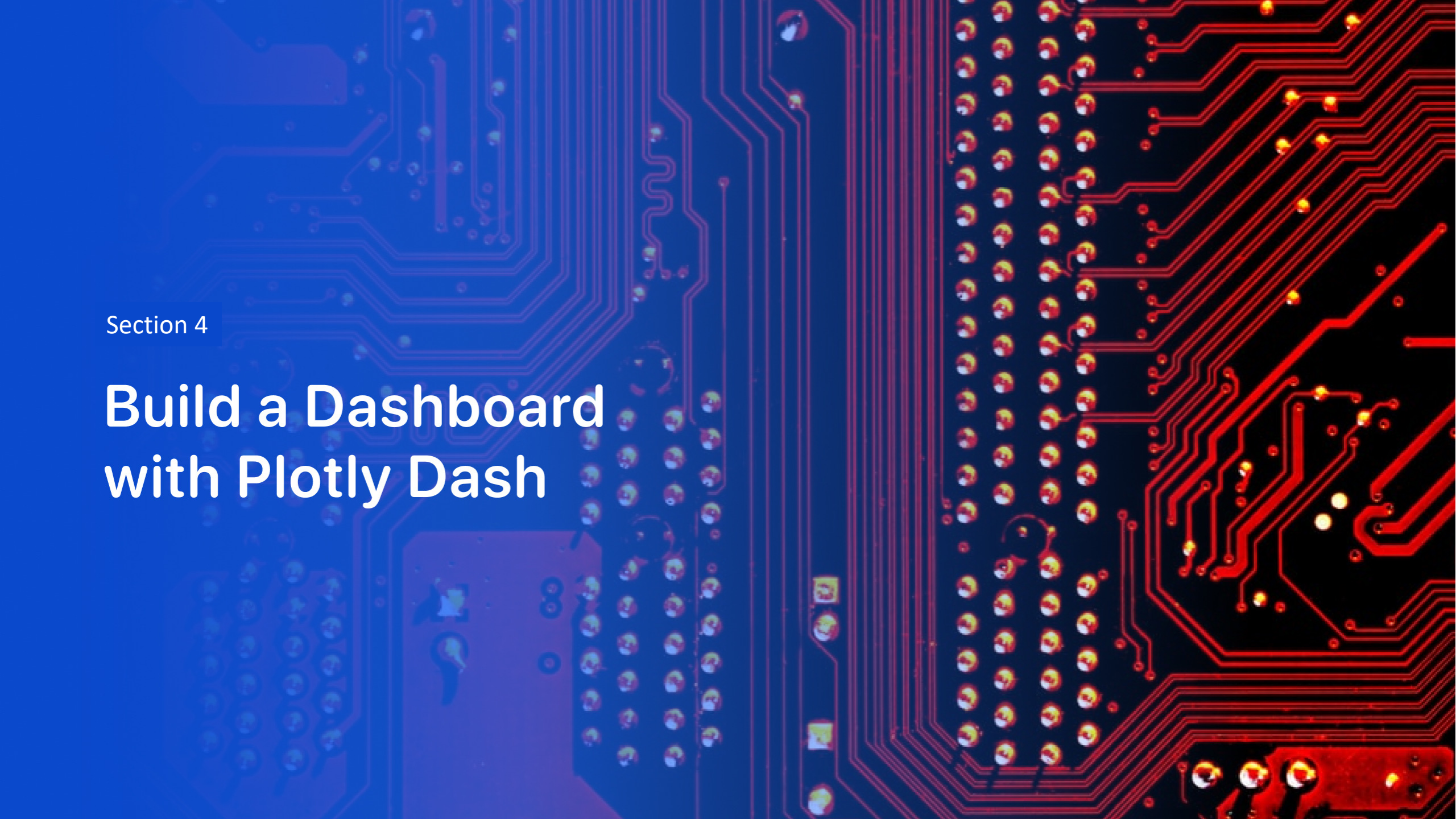




Distance between a launch site to its proximities

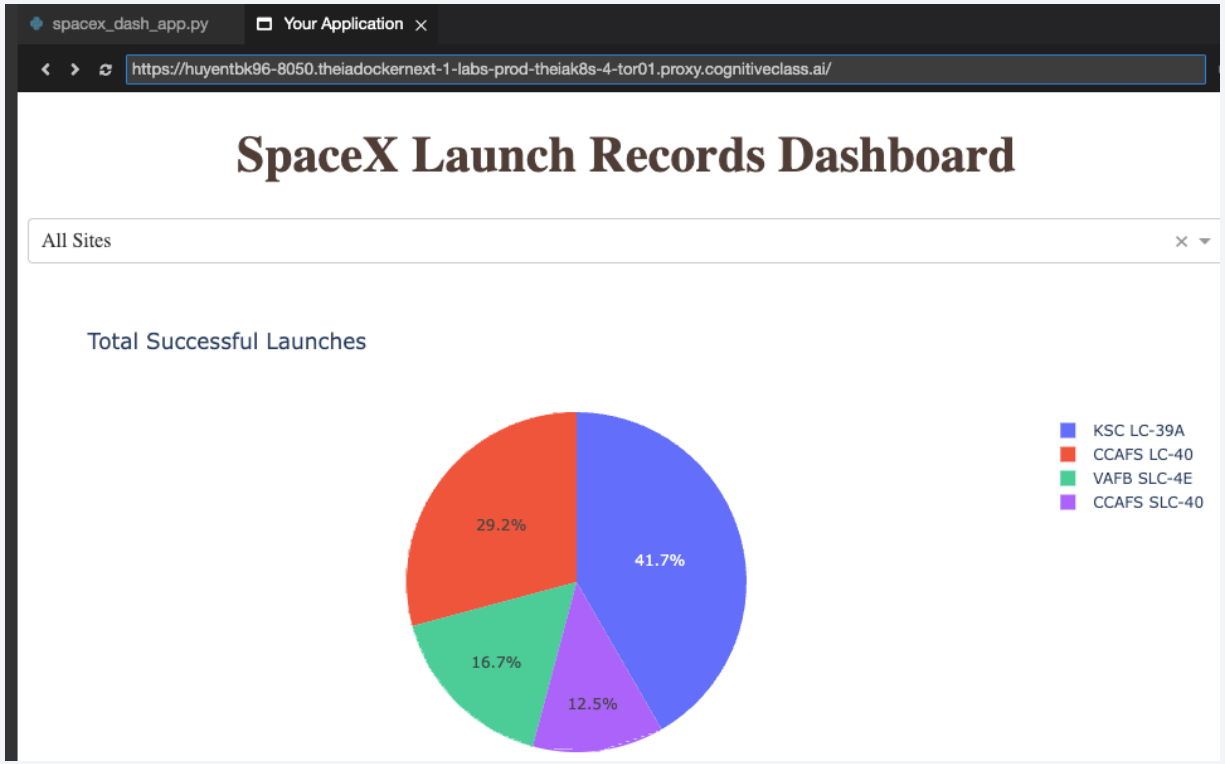
- The distance lines to the proximities are shown by a blue line.
- Between the launch site and travelling stations is relatively close, these stations are railways, highways, airport.
- Also, it is possible to notice that the launch site locations are not close the nearby major cities.
- The findings suggest that there are some considerations about the location of these launch sites to make it efficient to travel and to comply with safety standard.




Section 4
Build a Dashboard with Plotly Dash

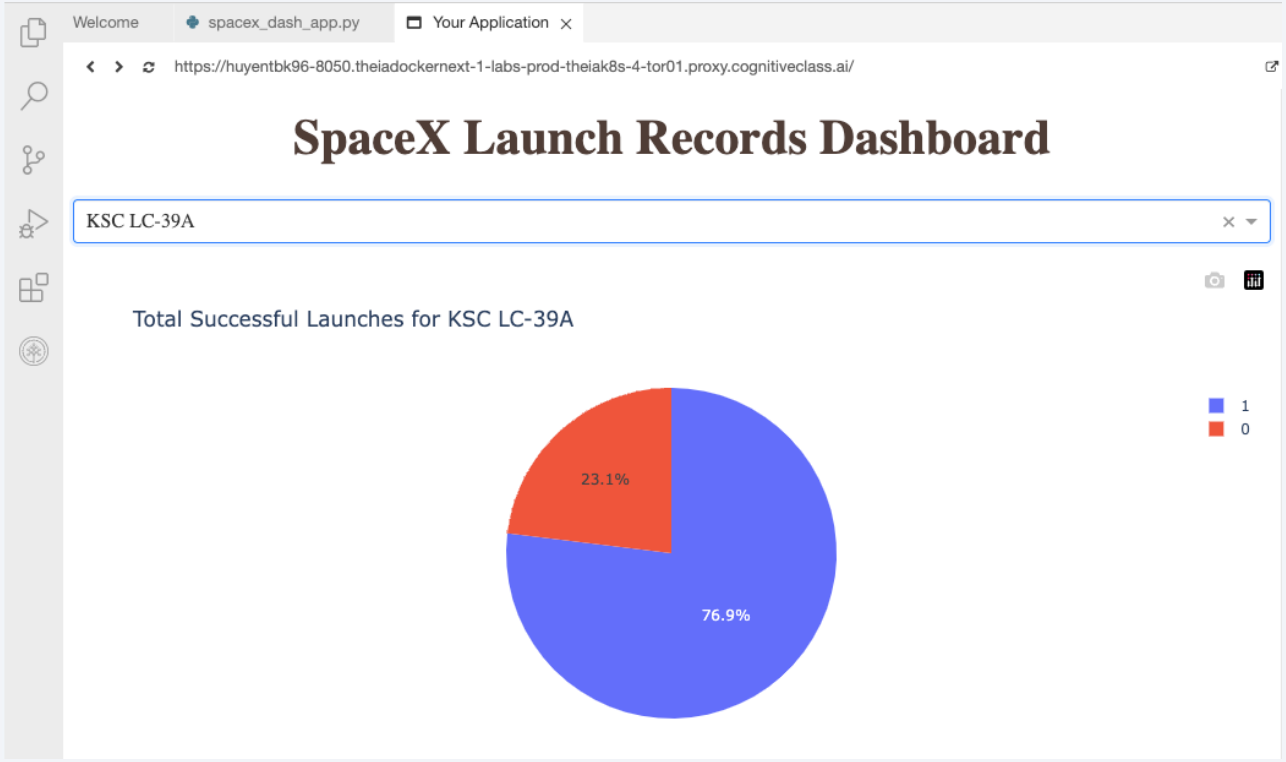
Launch success count for all sites

- All the launch sites shown in the pie chart is colored along with the success rates.
- The launch site with highest success count is KSC LC-39A



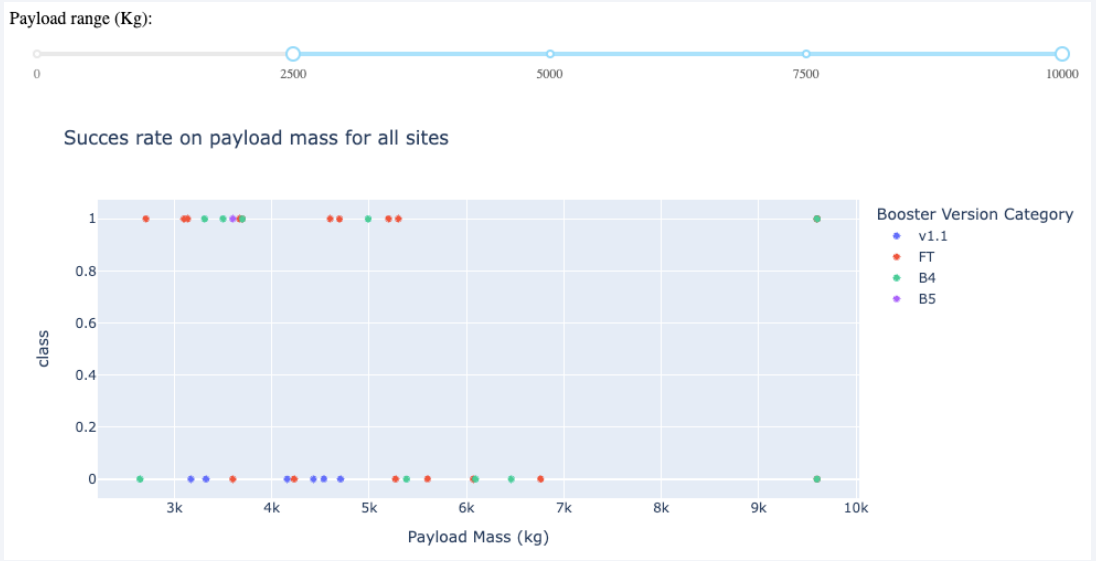
Launch site with highest launch success ratio

- From previous plot, it is easy to observe that KSC LC-39A has the highest success rate.
- Even that, this launch site still has that proportion of failure around 23.1 %.



Payload may be correlated with mission outcomes for all sites

- All over, The booster version with the highest success rate belongs to FT, still it is worth to mention B4 as well.
- But in considerations with different payload range, it is possible to see how well other booster versions perform.

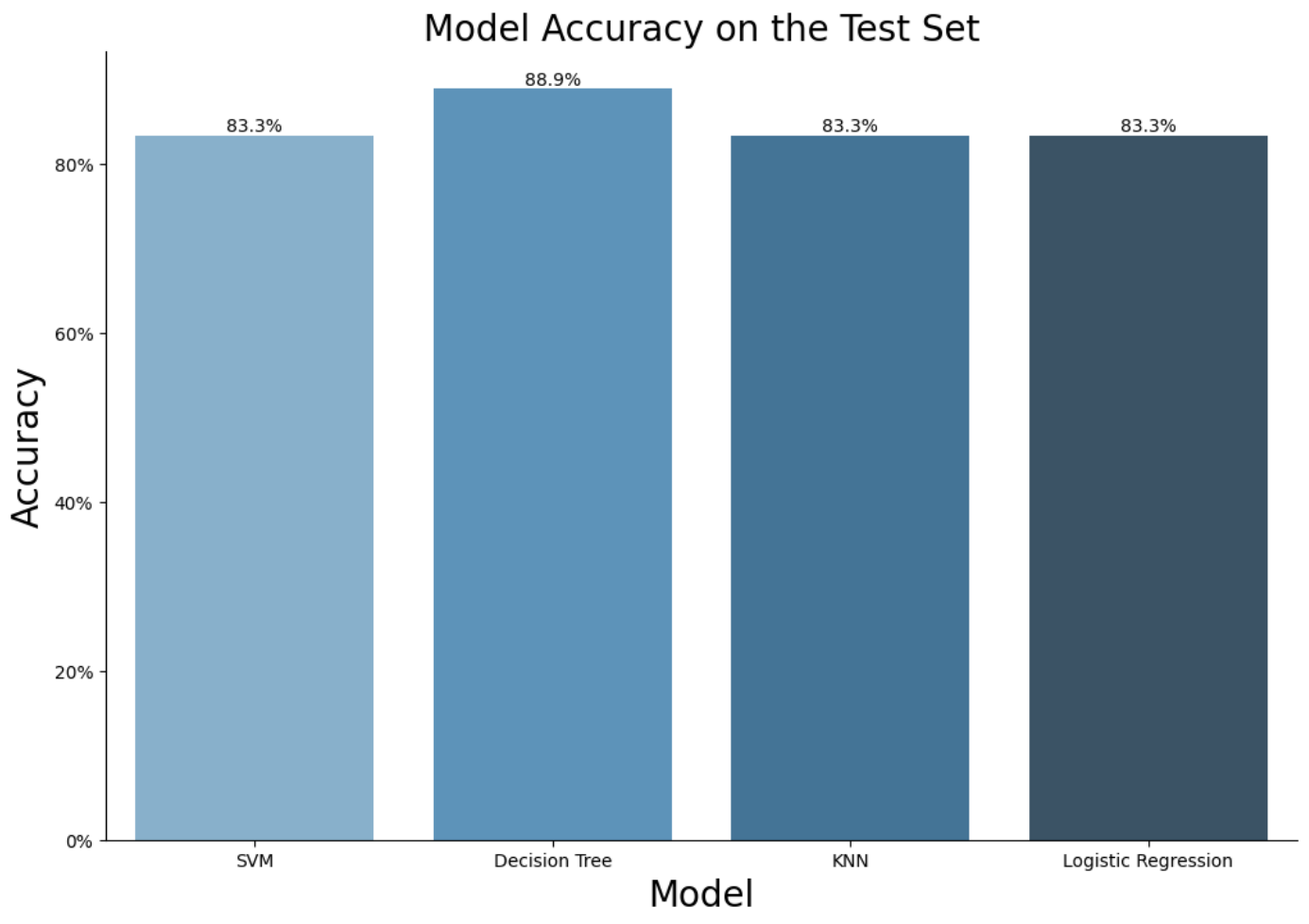


Section 5

Predictive Analysis (Classification)

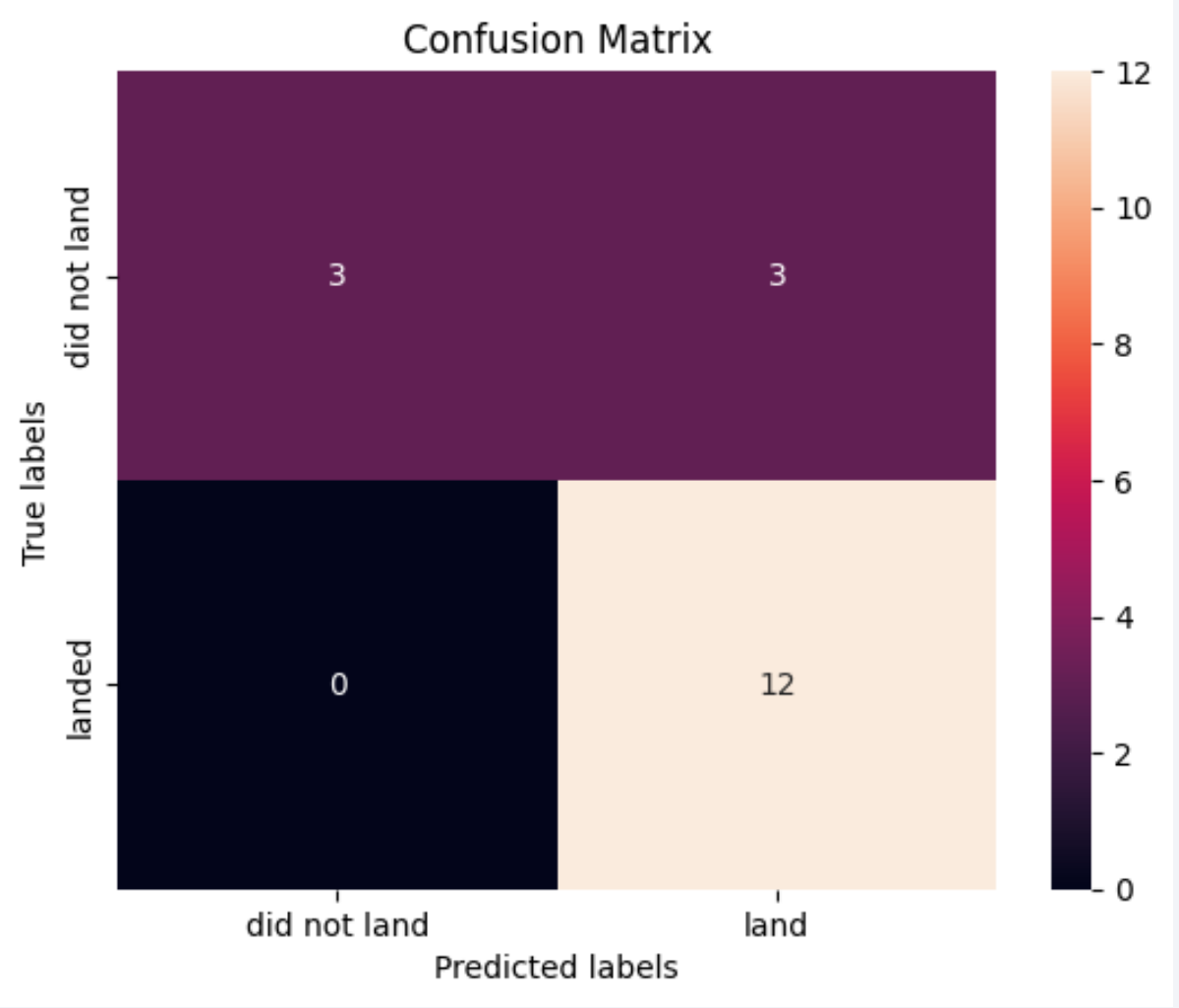
Classification Accuracy

- The model has the highest accuracy score is Decision Tree



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

- 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!