Winning Space Race with Data Science

Nazim

31st May 2023

Outline

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

Executive Summary

- ## Summary of Methodologies:

This project follows these steps:

- ❑ Data Collection
- ❑ Data Wrangling
- ❑ Exploratory Data Analysis
  - ❑ Using Relational Database and CSV Files
- ❑ Interactive Visual Analytics
  - ❑ Dashboards
- ❑ Predictive Analysis (Classification)
- ❑ Using Panda Libraries extensively

- ## Summary of Results:

This project produced the following outputs and visualizations:

- ❑ Exploratory Data Analysis (EDA) results
- ❑ Geospatial analytics
- ❑ Interactive dashboard
- ❑ Predictive analysis of classification models

Introduction

- Project background and context
- One of the major issue in Space launches has been the cost of a launch , Space X has shown that the cost can be considerably reduced . Space X launches at the cost of 62m vs the competitors cost of 162m. Much of the savings are achieved by SpaceX as it can land, and then re-use the first stage of the rocket.

- Problems you want to find answers
- Critical information we need to find out is how can we re-use the first stage , we need to identify the causes of the failures and the success factor. So that we can use this information for our company **Space Y** to successfully use first stage and cut cost and be competitive with Space X

Section 1

Methodology

Methodology

1. Data Collection
   - Collection data from spacexdata.com
   - Web Scraping from List of Falcon 9 and Falcon Heavy launches (2010–2019) - Wikipedia
2. Data Wrangling
   - Scrubbing , Organizing the data into proper columns removing blank data and filling it with mean values.
   - Hot encoding
3. Exploratory Data Analysis
   - Using SQL queries to manipulate and evaluate the SpaceX dataset
   - Using Pandas and Matplotlib to visualize relationships between variables, and determine patterns
4. Interactive Visual Analytics
   - Geospatial analytics using Folium
   - Creating an interactive dashboard using Plotly Dash
5. Data Modelling and Evaluation
   - Using Panda libraries to:
     - Pre-process (standardize) the data
     - Split , test and data using train_test_split
     - Usingdifferent classification models
     - Find hyperparameters using GridSearchCV
   - Plotting confusion matrices
   - Assessing the accuracy of each classification model

Data Collection

Data was collected from spacexdata.com and List of Falcon 9 and Falcon Heavy launches (2010–2019) - Wikipedia

Spacexdata collected through API contains the historical data of the launches with information pertaining to site of the launch, payload, launch specification, its landing outcome and other details

Used

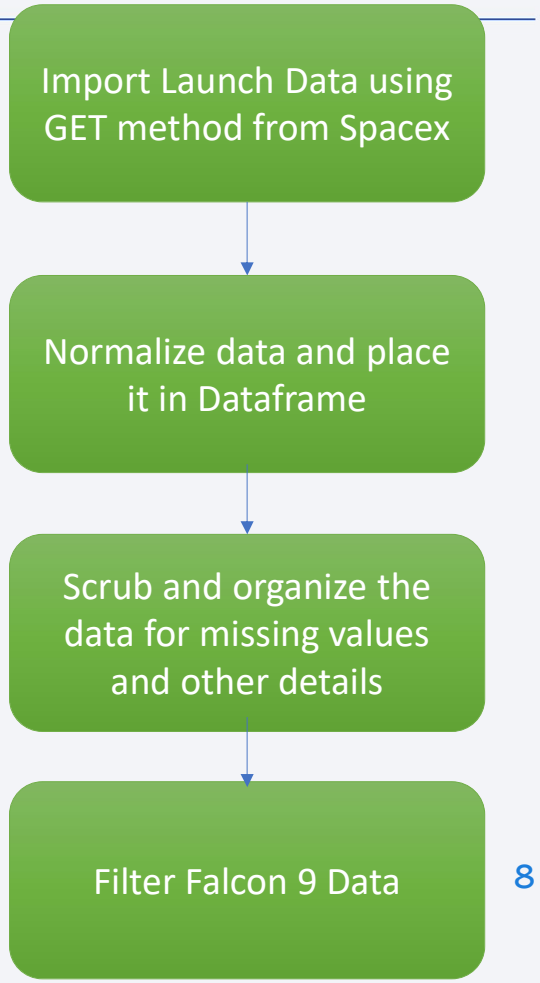
- GET response
- Stored in Dataframes
- Used Sql and Data Manipulation Language
- Focus on Falcon 9 launches

Data Collection – SpaceX API

Space X provides its launch data on its website

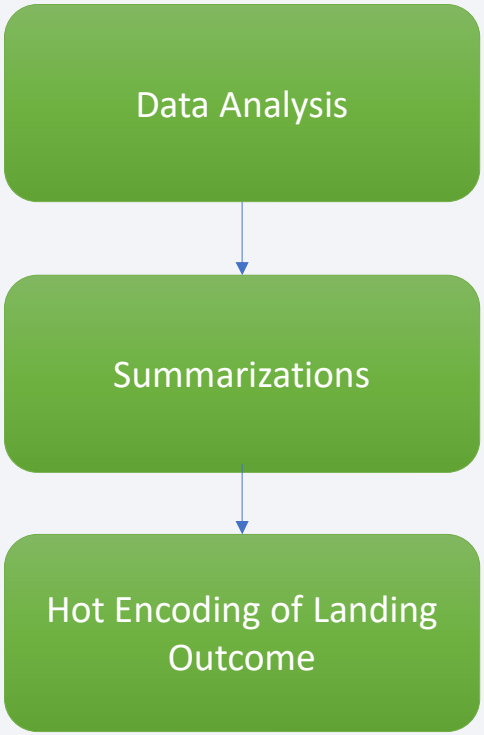
```
# Takes the dataset and uses the rocket column to call the API and append the data to the list
def getBoosterVersion(data):
    for x in data['rocket']:
       if x:
        response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
        BoosterVersion.append(response['name'])
```

- IBMCertProject/01.data-collection-api.ipynb at main · inazim/IBMCertProject (github.com)



Data Collection - Wrangling

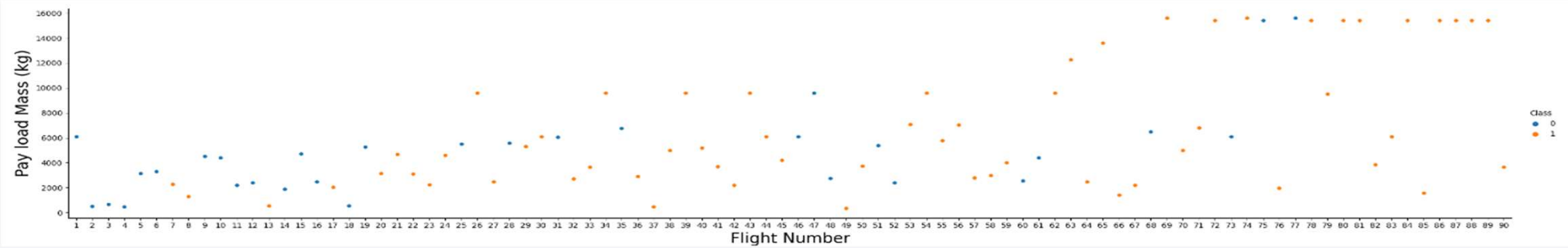
- Started with Data Analysis
- Calculated the Summary values
  - # of Launches on Eachsite
  - # of Occurrences on Each Orbit
  - Calculated # of occurrences of mission outcome per orbit
- Hot encoding the Landing outcome

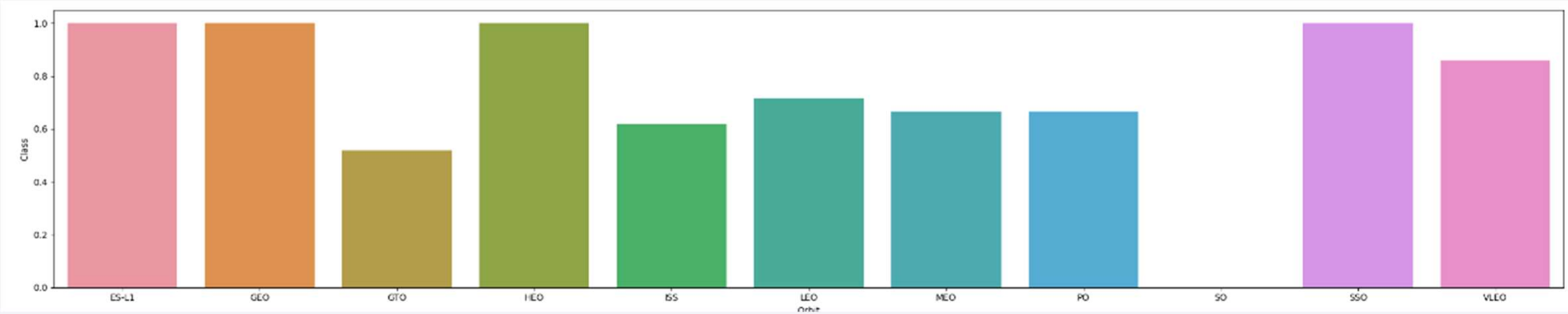


IBMCertProject/02.data_wrangling_jupyterlite.jupyterlite.ipynb at main · inazim/IBMCertProject (github.com)

EDA with Data Visualization

- Data was visualized using Scatter Charts, Line Charts and Bar Charts

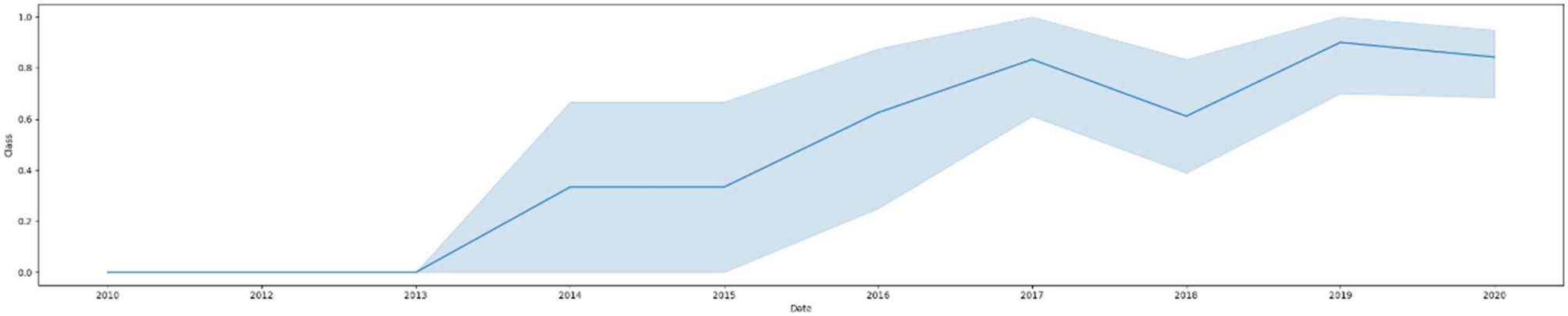




Next slide please for the URL

EDA with Data Visualization

- Data was visualized using Scatter Charts, Line Charts and Bar Charts



IBMCertProject/04.eda-dataviz.ipynb.jupyterlite.ipynb at main · inazim/IBMCertProject (github.com)

EDA with SQL

- The SQL queries performed on the data set were used to:
  1. Display the names of the unique launch sites in the space mission
  2. Display 5 records where launch sites begin with the string 'CCA'
  3. Display the total payload mass carried by boosters launched by NASA (CRS)
  4. Display the average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome on a ground pad was achieved
  6. List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
  7. List the total number of successful and failed mission outcomes
  8. List the names of the booster versions which have carried the maximum payload mass
  9. List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015
  10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

IBMCertProject/03.eda-sql-coursera_sqllite.ipynb at main · inazim/IBMCertProject · GitHub

Build an Interactive Map with Folium

- Folium Maps were used as below
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
  - Lines are used to indicate distances between two coordinates.

IBMCertProject/05.launch_site_location.jupyterlite.ipynb at main · inazim/IBMCertProject (github.com)

Build a Dashboard with Plotly Dash

- Pie Chart and Scatter graph were used with added interaction

1. Pie chart was used plotting total successful launches per site
   - This gives visual view of the most successful launch site
   - Dropdown box was used to see the success/failure ratio for an individual site

2. Scatter graph to depict the relation between outcome (success or not) and payload mass (kg)
   - Filter added for ranges of payload masses
   - Filtered added for booster version

IBMCertProject/05b. Plotly Dash dashboard_
spacex_dash_app.py at main · inazim/IBMCertProject · GitHub

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

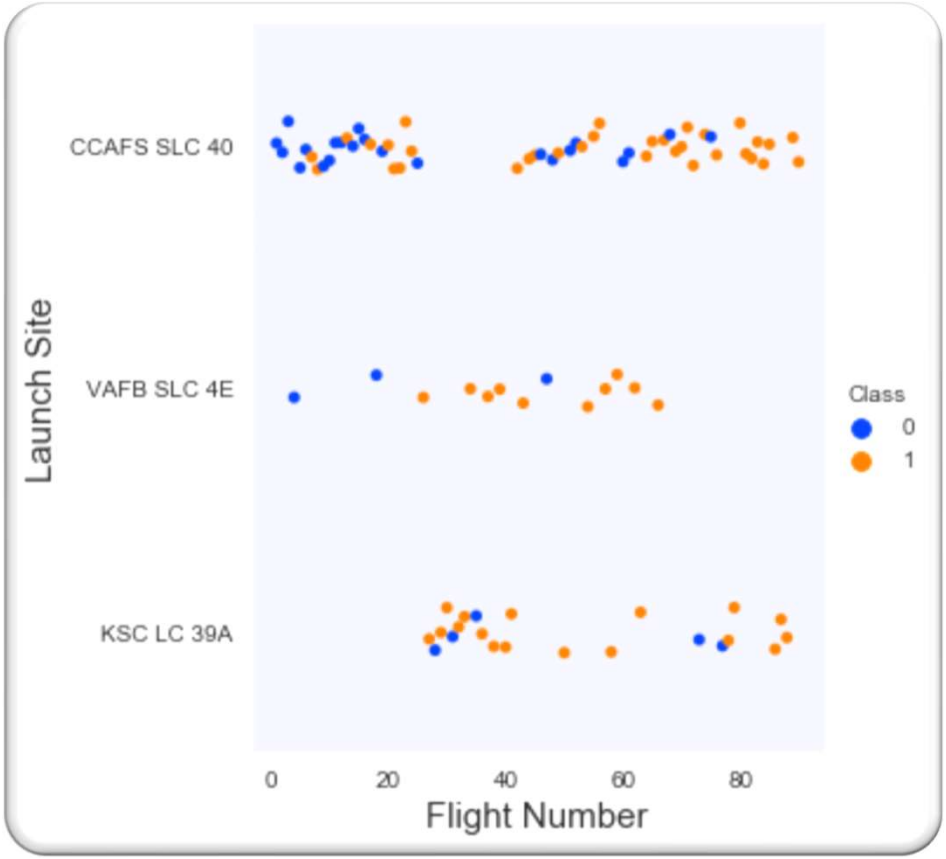
IBMCertProject/06.SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · inazim/IBMCertProject (github.com)

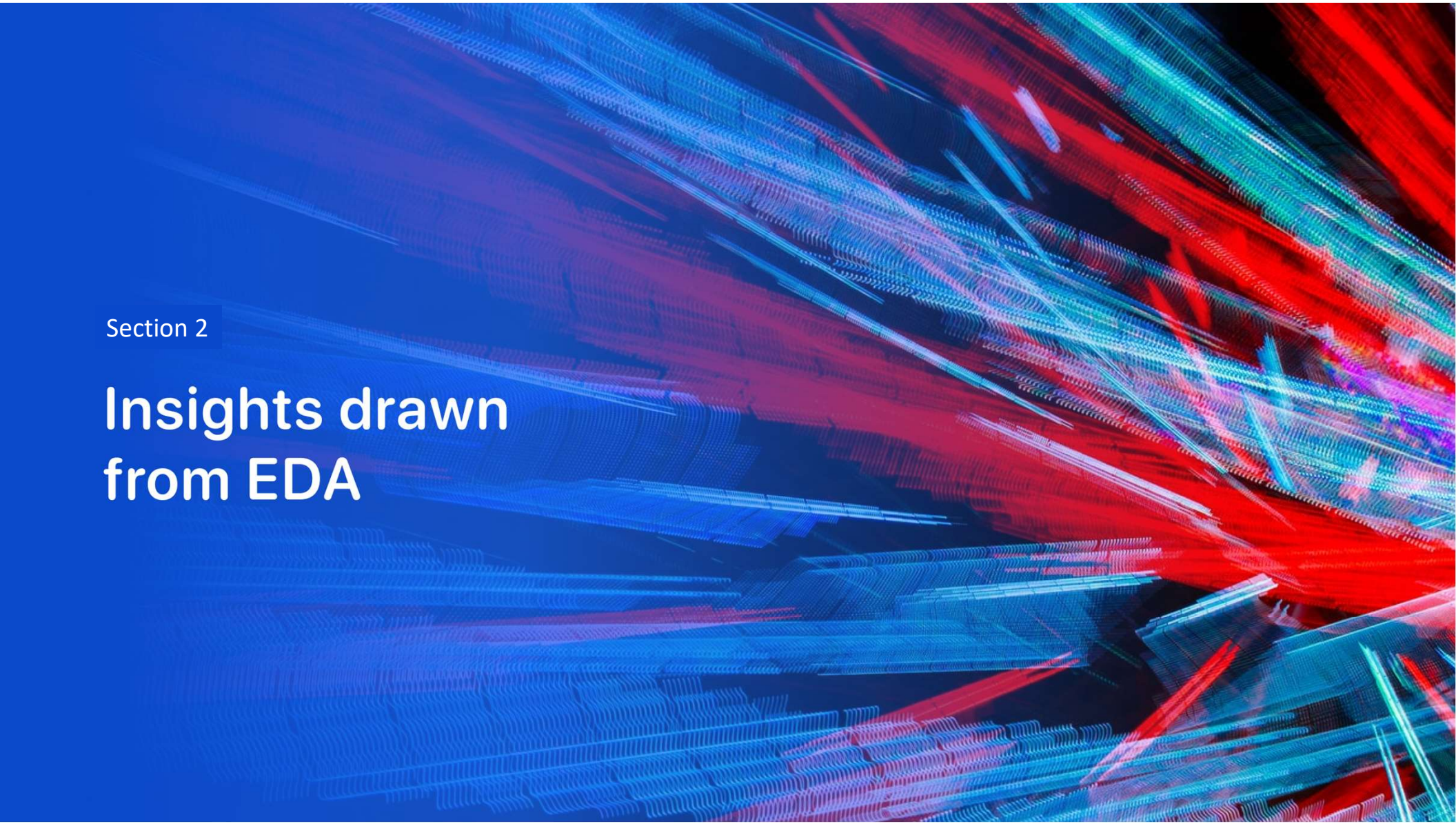
Results

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

Exploratory data analysis results

- The scatter plot of Launch Site vs. Flight Number shows that:
- As the number of flights increases, the rate of success at a launch site increases.
- Most of the early flights (flight numbers < 30) were launched from CCAFS SLC 40, and were generally unsuccessful.
- The flights from VAFB SLC 4E also show this trend, that earlier flights were less successful.
- No early flights were launched from KSC LC 39A, so the launches from this site are more successful.
- Above a flight number of around 30, there are significantly more successful landings (Class = 1).

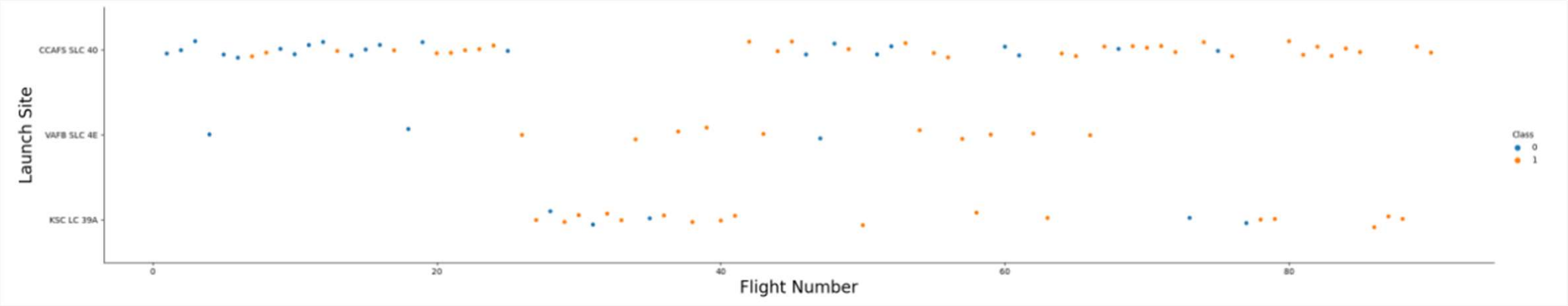




Section 2

Insights drawn from EDA

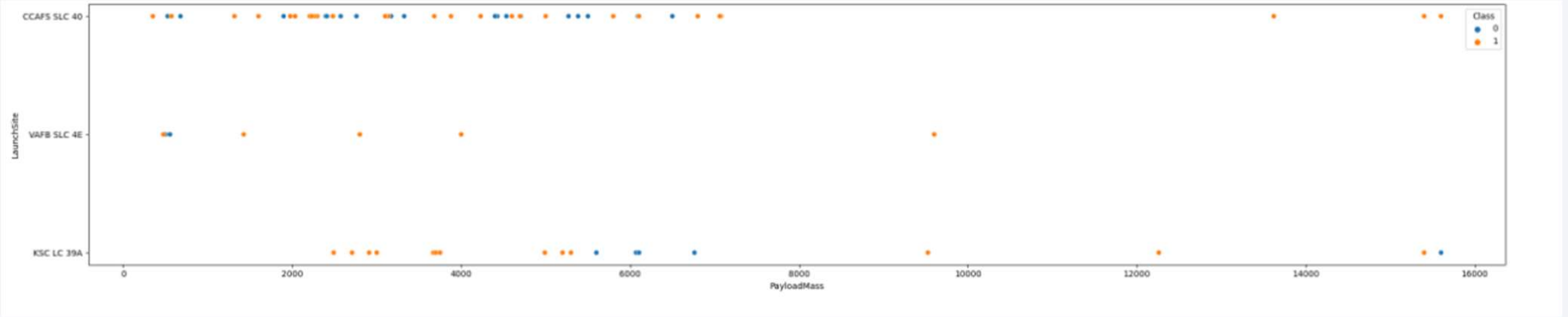
Flight Number vs. Launch Site



The scatter plot of Launch Site vs. Flight Number shows that:

- Increase in flight increases the Success rate.
- The flights from VAFB SLC 4E also show this trend, that earlier flights were less successful.
- Above a flight number of around 30, there are significantly more successful landings (Class = 1).

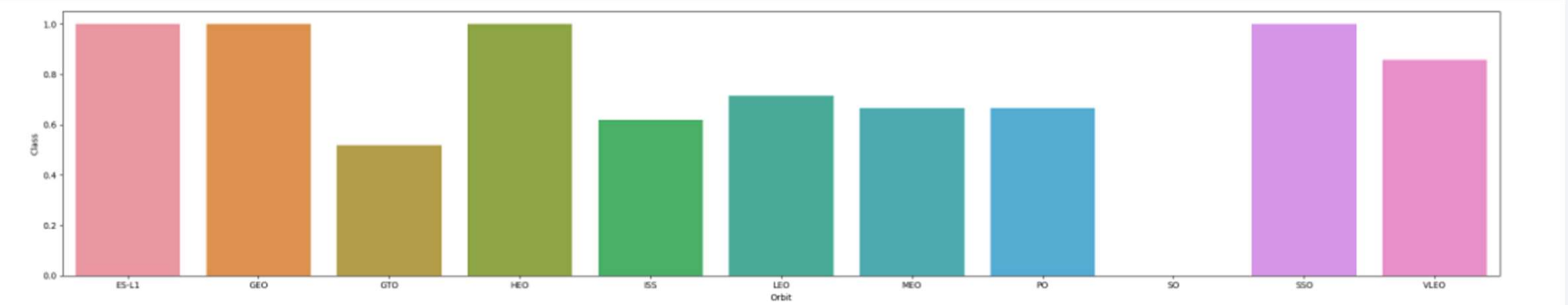
Payload vs. Launch Site



Results of Payload vs Launch site as as follows

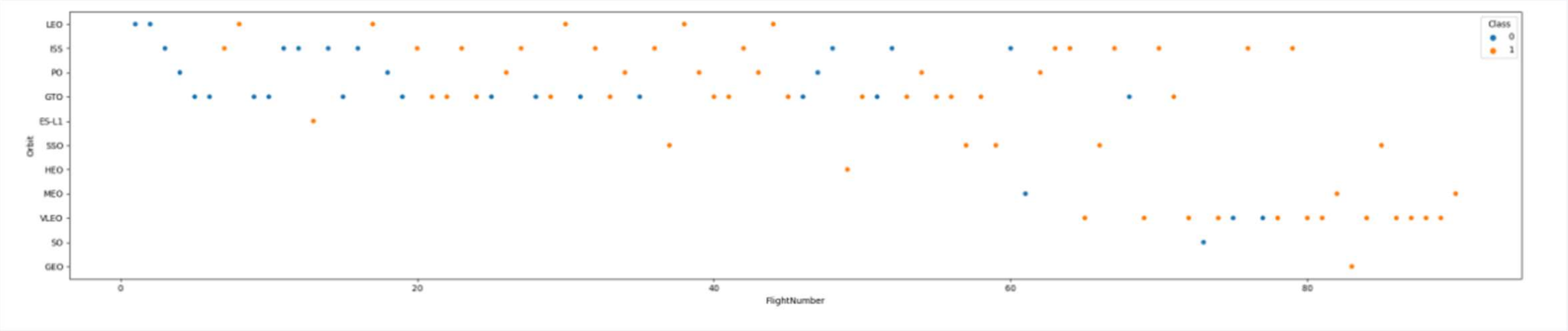
- Above a payload mass of around 10000 kg there are very few unsuccessful landings, but there are few launches with this payload
- All sites launched a variety of payload masses, with most of the launches from CCAFS SLC 40 being comparatively lighter payloads with few exceptions.

Success Rate vs. Orbit Type



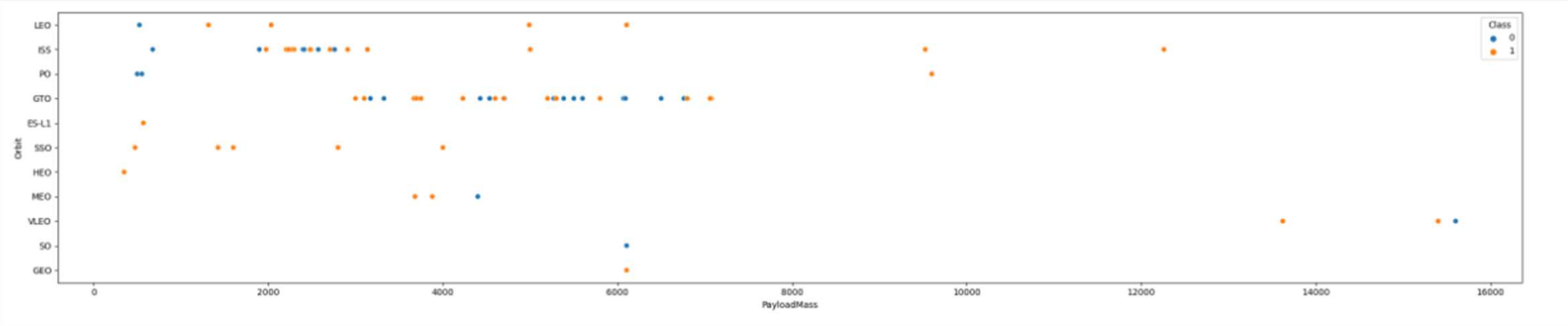
- Success rate for ES-11, GEO, HEO and SSO is 100%
- SO has no success.

Flight Number vs. Orbit Type



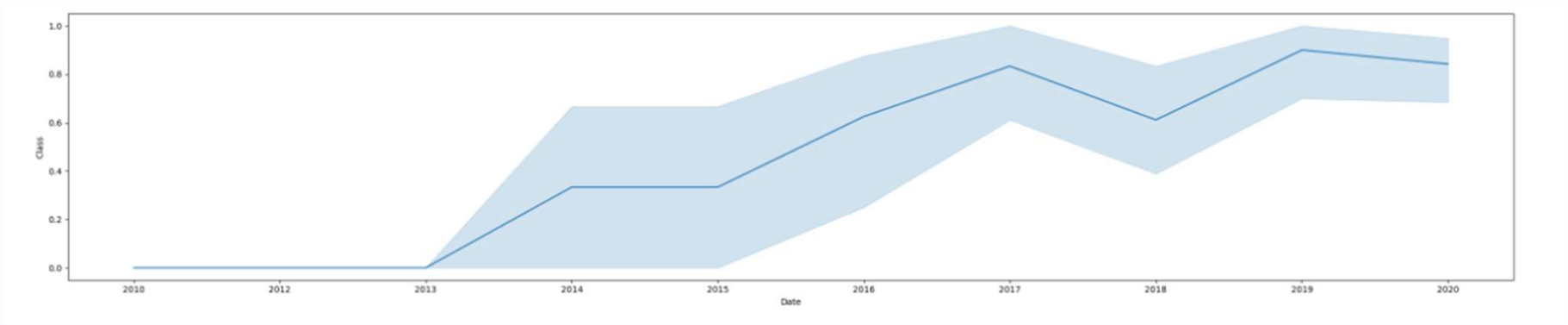
- SSO has 5 flight with 100% success where as GEO, HEO, and ES-L1 orbits have only 1 flight, we need more data to ascertain their success.
- Increase in flight's increases the probability of success

Payload vs. Orbit Type



- VLEO – Heavier payload
- PO,ISS and LEO has more success rate as per the available data

Launch Success Yearly Trend



- 2010 to 2013 no success
- From 2013 the success rate has increased progressively
- From 2019 the success rate is considerably high reaching to its peak

All Launch Site Names

```
]: %sql SELECT distinct LAUNCH_SITE FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
]:
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

- Distinct clause gives the Unique values in the given column(s)

Launch Site Names Begin with '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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- * is a wild card characters representing multiple characters ;
- Limit 5 – Gives only 5 rows

Total Payload Mass

```
]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
]:
```

TOTAL_PAYLOAD_MASS
45596.0

- Sum function adds all the column values given , the rows are filtered by the where clause

Average Payload Mass by F9 v1.1

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

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

AVERAGE_PAYLOAD_MASS
2928.4

- Avg function finds the average of the given data
- Where clause filters the record based on the given condition ; here it is for Booster_Version='F9 v1.1'

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

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

FIRST_SUCCESSFUL_GROUND_LANDING
01/08/2018

- Min function finds the earliest (in case of date) or the least lesser value (in case of numbers)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE (Landing_Outcome = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

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

- Between clause searches for values between a given Range

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome,COUNT(*) AS NO_OF_MISSIONS FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

Mission_Outcome	NO_OF_MISSIONS
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Group by clause Groups the data as per the given column name, Count(*) – Counts the number of Rows

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT 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

- Maximum finds the highest value in the given query; the Subquery gives the value that can be used by the main query to filter data

2015 Launch Records

```
%sql SELECT substr(Date, 4, 2) MONTHNAMES,BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL \
    WHERE (Landing_Outcome = 'Failure (drone ship)') AND (substr(Date,7,4)='2015' );
```

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

MONTHNAMES	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- Substring function is used to extract part of a string or date.
- And clauses is used to combine multiple conditions. If both are true then the rows that match both the condition are shown

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

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS TOTAL_NUMBER FROM SPACEXTBL \
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' and Landing_Outcome='Succ%' GROUP BY Landing_Outcome \
    ORDER BY TOTAL_NUMBER DESC;
```

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

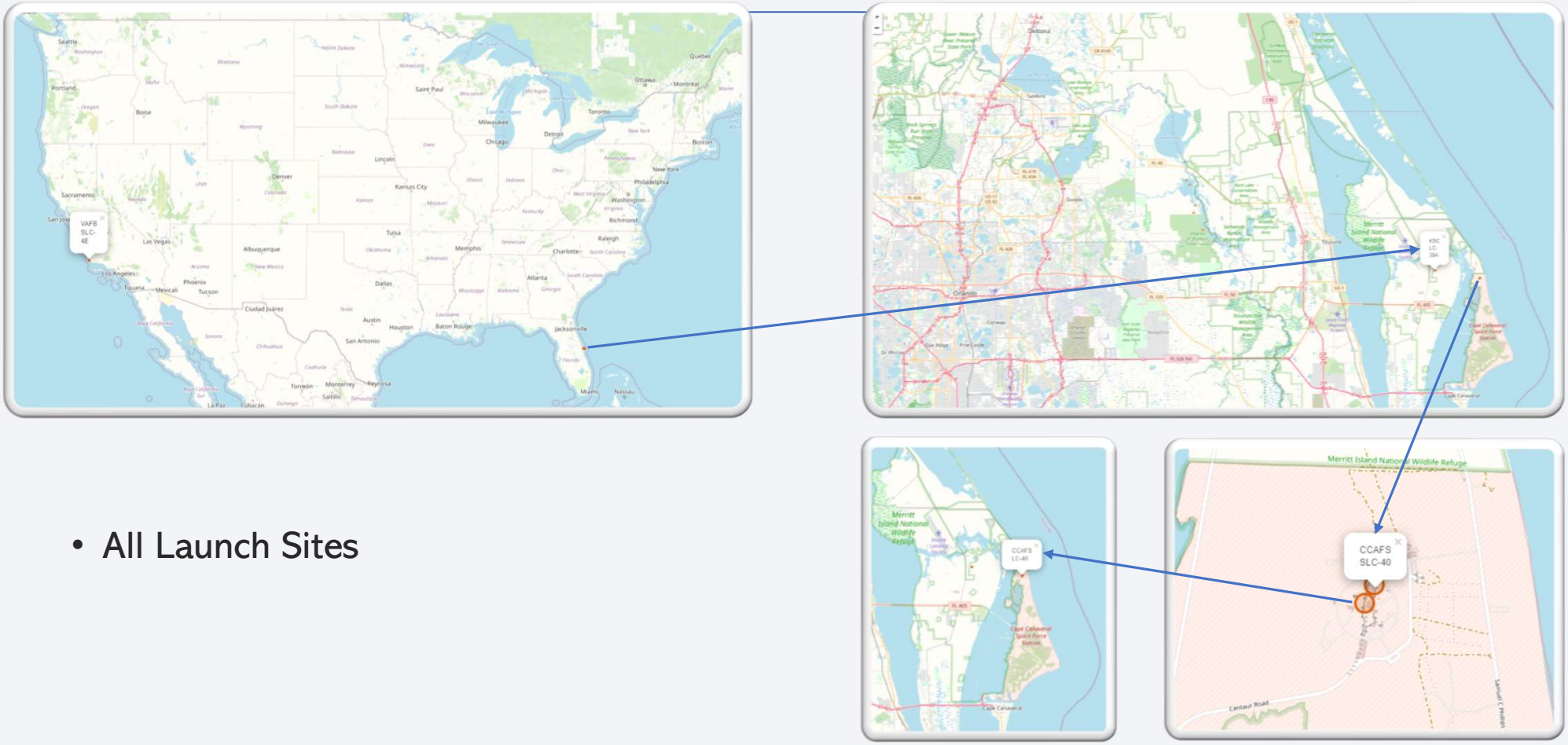
Landing_Outcome	TOTAL_NUMBER

- Order by clauses sorts the data by the given column(s) . Desc clause with the Order by sorts the data in descending order

Section 3

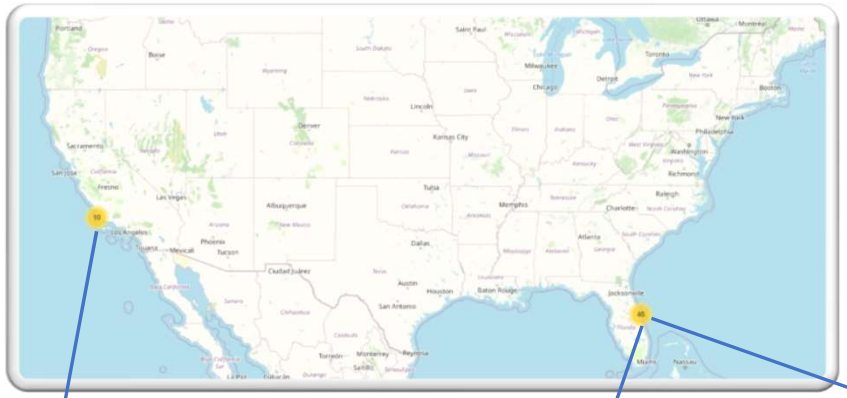
Launch Sites Proximities Analysis

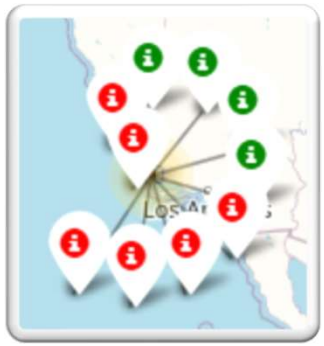
All Launch Sites Location

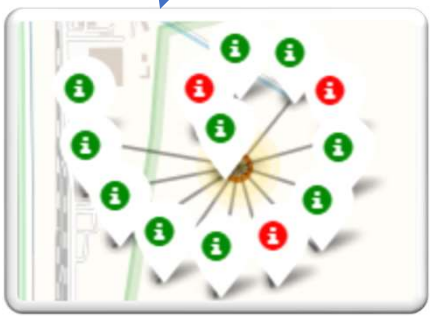


- All Launch Sites

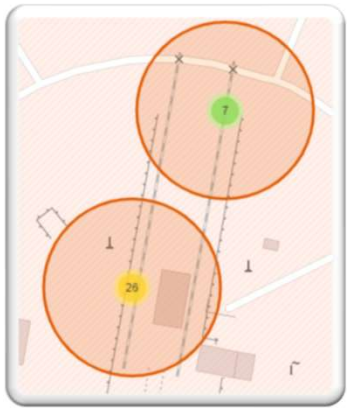
SUCCESS/FAILED LAUNCHES FOR EACH SITE



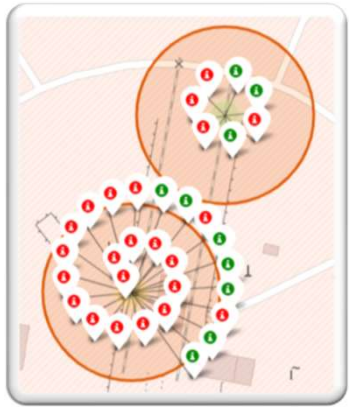


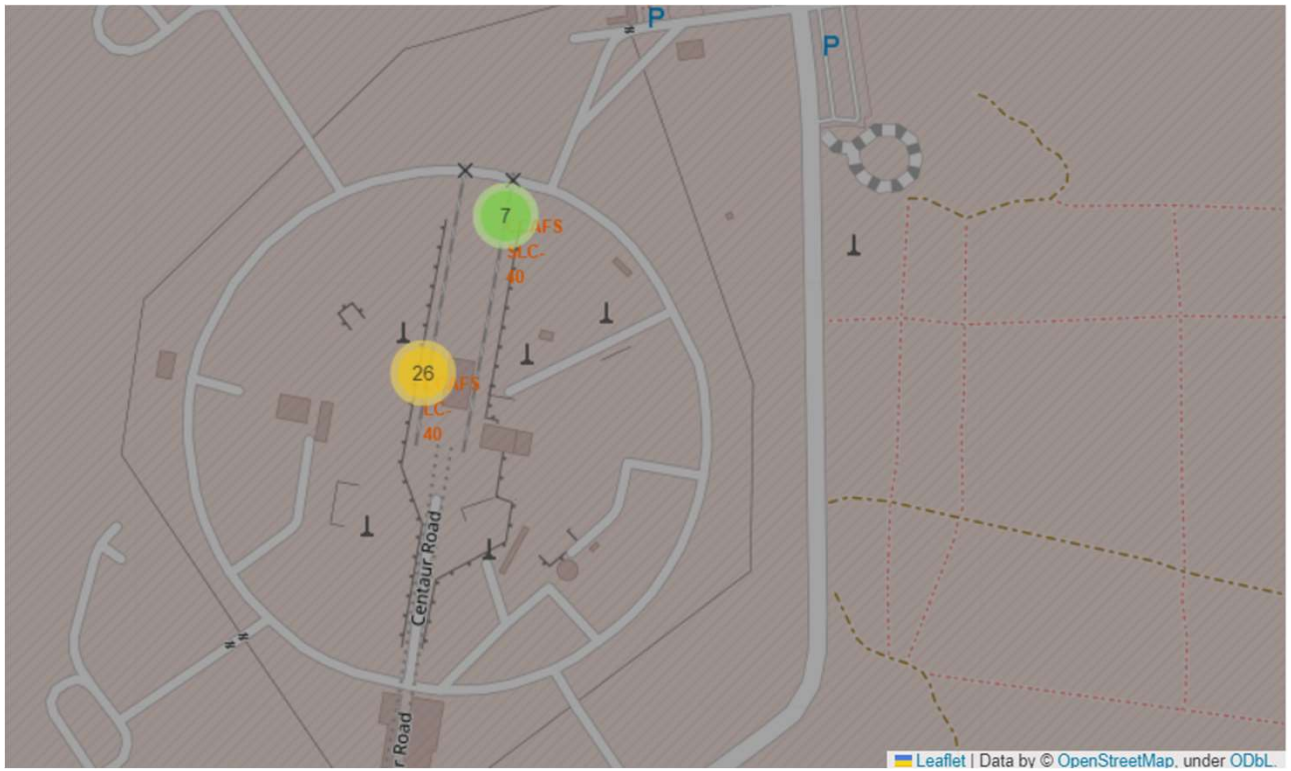


- Clusers used to group Launches,
- Green icons – Success
- Red icons - Failure.

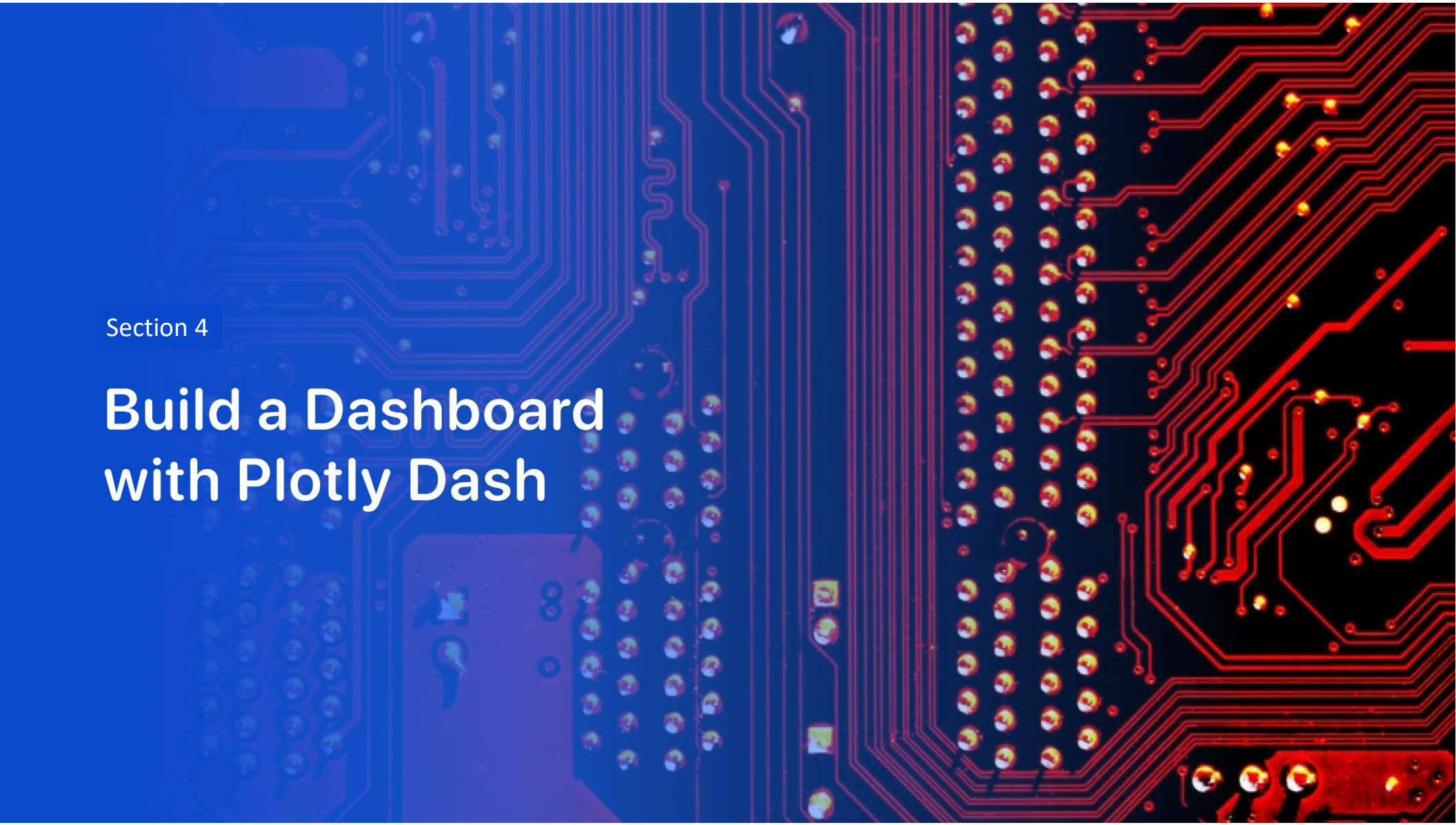


=



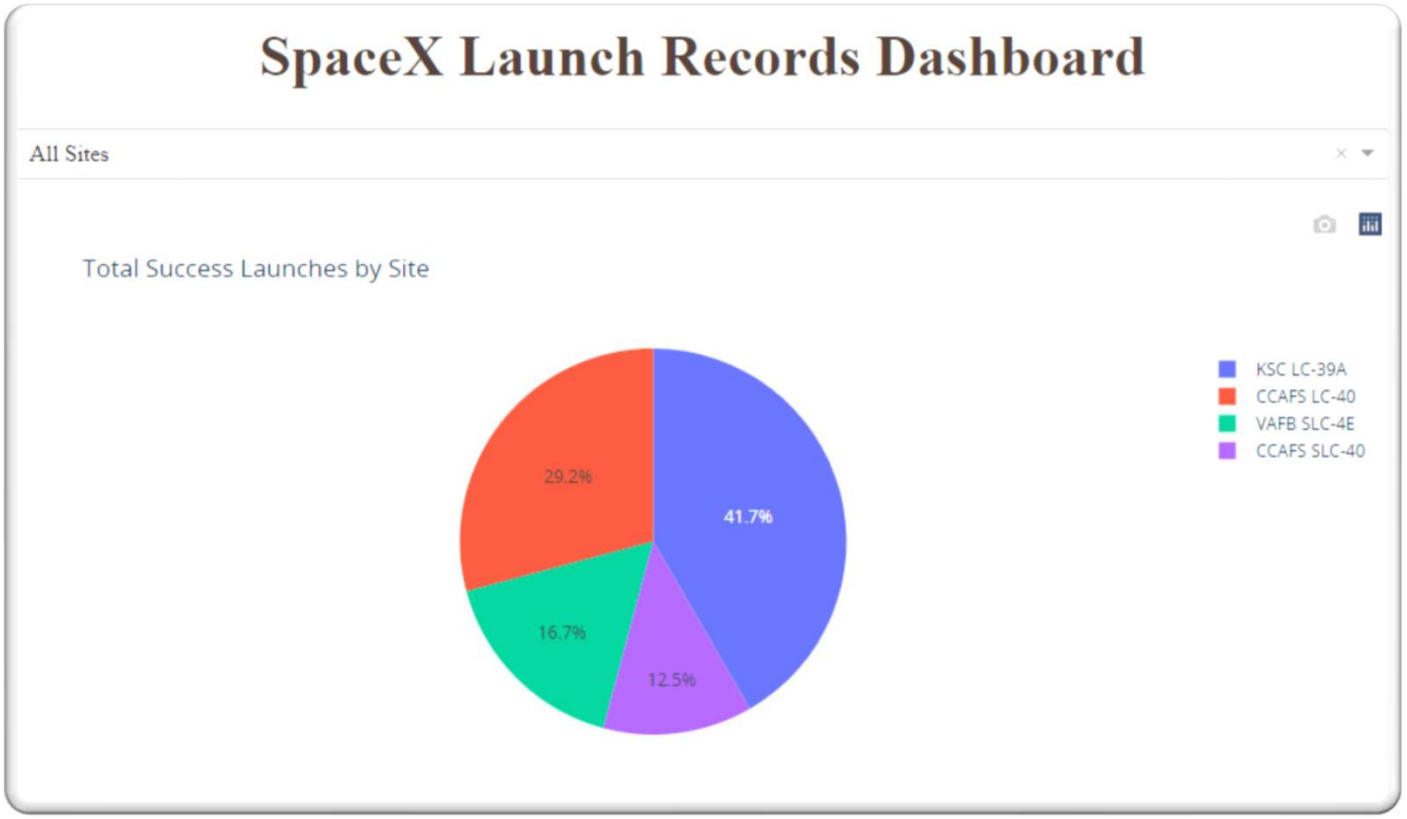


- It is Near to the Road



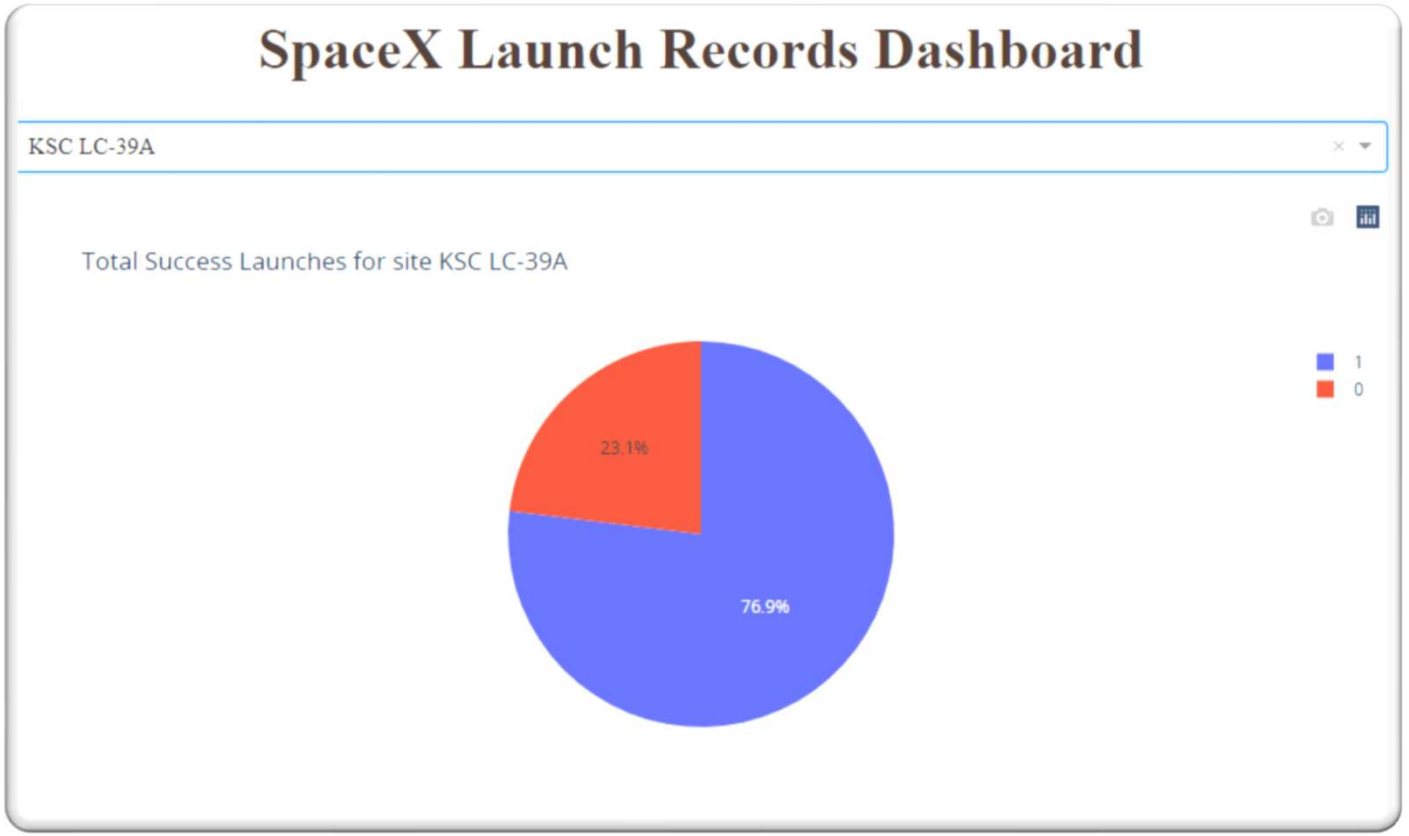
Section 4

Build a Dashboard with Plotly Dash



- KSC LC-39 A is the most successful with 41.7% Successes

The launch site KSC LC-39 A also had the highest rate of successful launches, with a 76.9% success rate.



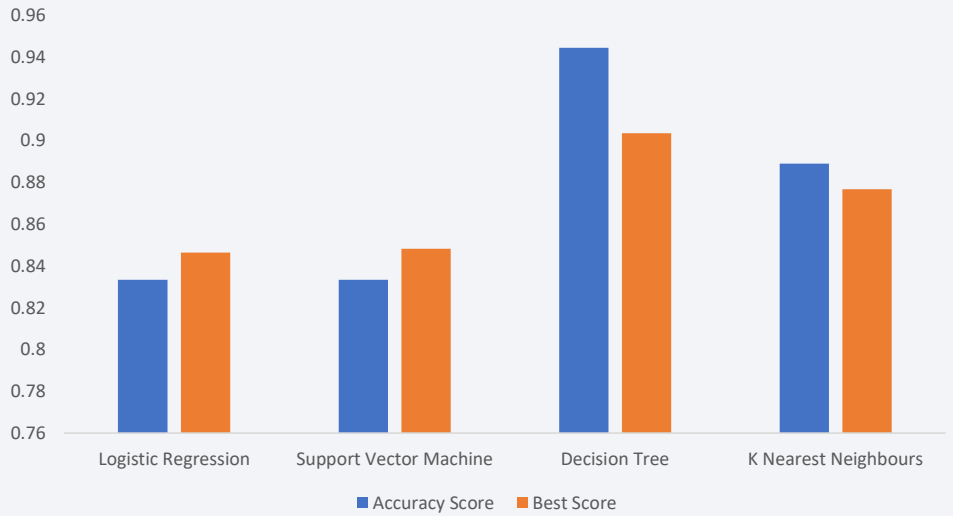
- Payloads are of different types Low, Medium to Huge

Section 5

Predictive Analysis (Classification)

Classification Accuracy

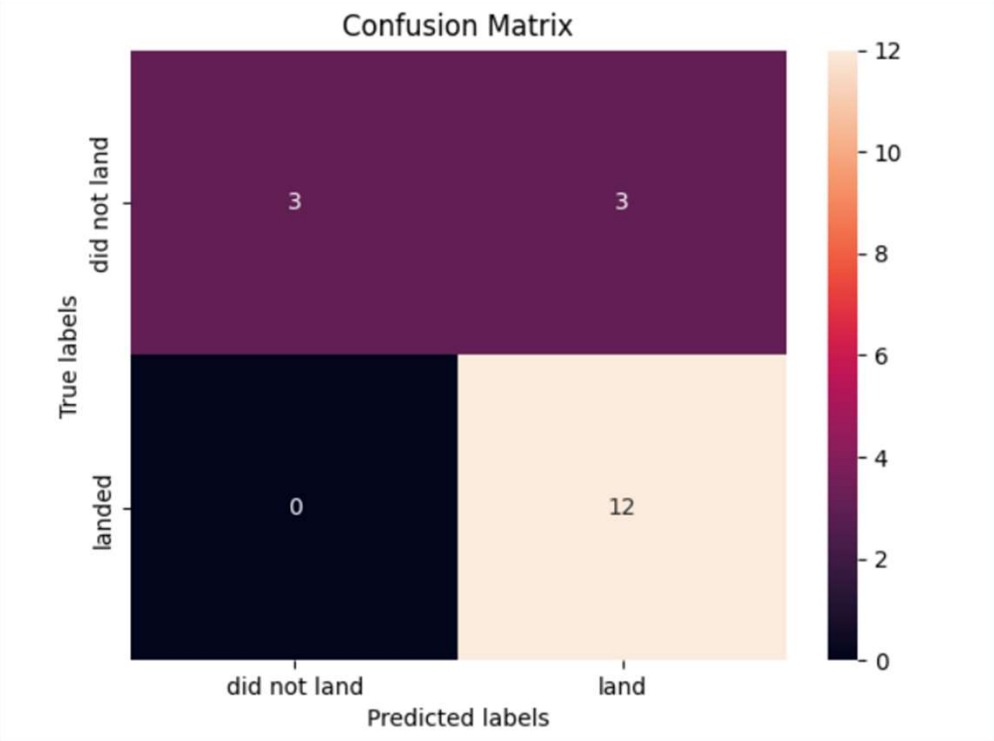
Method	Accuracy Score	Best Score
Logistic Regression	0.833333	0.846429
Support Vector Machine	0.833333	0.848214
Decision Tree	0.944444	0.903571
K Nearest Neighbours	0.888889	0.876786



- Decision Tree is given the best – Accuracy and Score

Confusion Matrix

- The Big number of True positive(12) gives the high accuracy level.



Conclusions

- There is a gradual progression of Launch success, from 2019. It shows that there has been a good learnings from the earlier failures
- The Launch site KSC LC-39 A is the most successful with 41.7% success rate
- Success rate for ES-11, GEO, HEO and SSO (with 5 launches) is 100%
- Decision Tree classification gives the highest accuracy and Best Score.

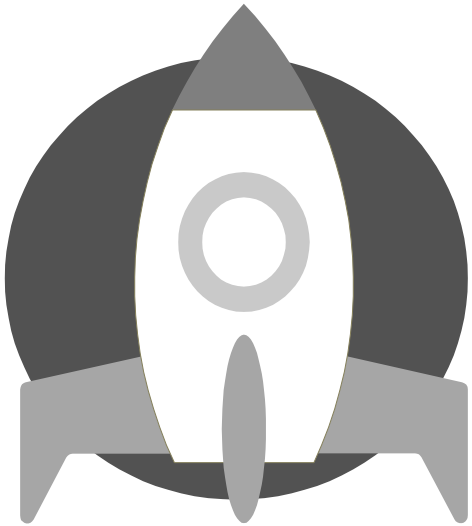
Appendix

Resources:

Python For Beginners | Python.org

Machine learning, explained | MIT Sloan

Exploratory Data Analysis Results



- Success rate is increasing with increase in number of flights launched
- Above 7000kg the success rate decreases
- 100% success rate for ES-L1,GEO,HEO and SSO; Lowest rate is SO
- Success started from 2013 and has been progressively increasing
- Average payload for Falcon 9 v1.1 booster is 2,928kg
- Success rate of Falcon 9 was better in landing on Drone ships with above the average payload

Thank you!