IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sayak Majumder
08.10.2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **Summary of methodologies**
  - ❑ Data collection via SpaceX API and web-scraping
  - ❑ Data wrangling and pre-processing
  - ❑ Exploratory data analysis with SQL and interactive visualizations (Folium, Plotly, Dash)
  - ❑ Predictive analysis using classification machine learning models
- **Summary of all results**
  - ❑ The results of the EDA are showcased in Sections 2, 3, 4 and 5
  - ❑ Section 6 focuses on the results of the predictive analysis.

Introduction

- **Project Background**

  SpaceX, substantially inexpensive compared to other spacecraft engineering companies, is able to save millions of dollars by reusing the first stage of its Falcon 9 rockets. With the help of data science, this project aims to help a competing company determine the cost of each launch by assuring that the first stage has a successful launch or an unsuccessful one.

- **Research Questions**

  i. Which factors correspond to a higher success rate of a launch?

  ii. What are the effects of each relationship between spacecraft features on the launch outcome?

  iii. Is there any trend that can help to predict the likelihood of reusing the first stage?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected via API requests and web-scraping.
- Perform data wrangling
  - From the API response content, auxiliary functions have been used to filter Falcon 9 data and its features (flight number, payload mass, outcome, etc.)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, Support Vector Machines, K-Nearest Neighbors, and Decision Trees have been used for predictive analysis and confusion matrices to evaluate said classifiers.

Data Collection

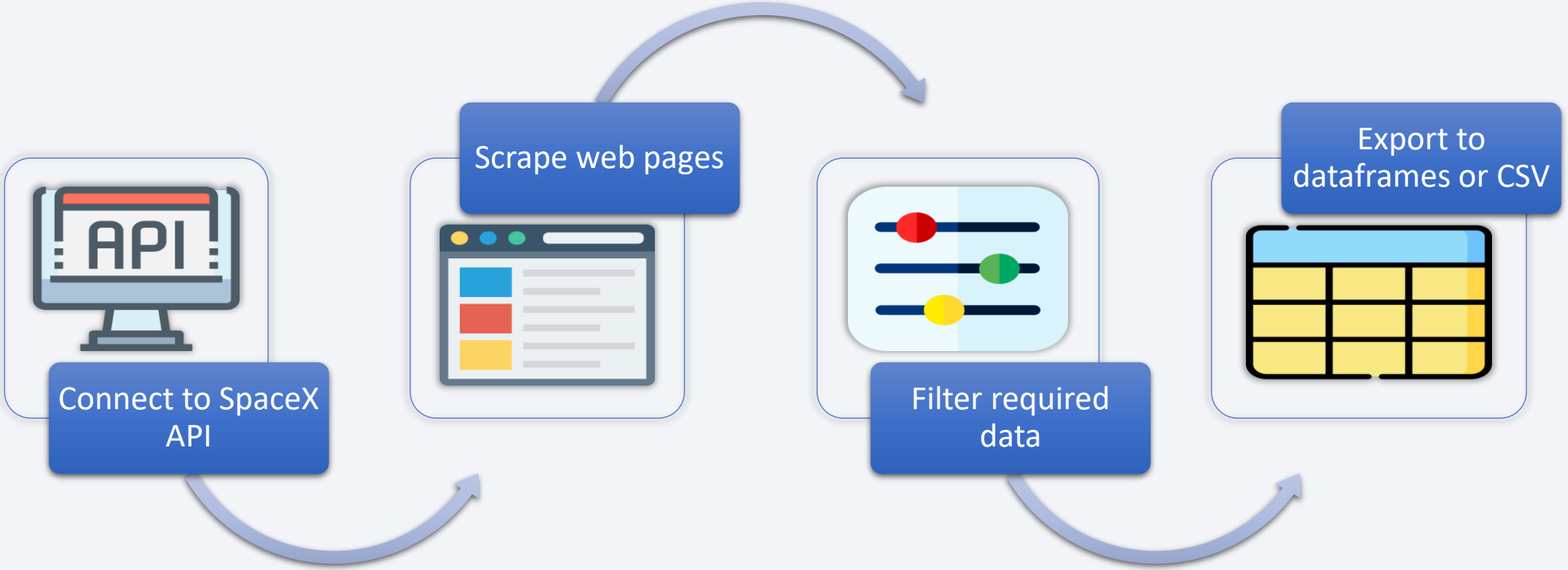


image: Flaticon.com

Data Collection – SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

Target a specific endpoint of SpaceX API

Fetch Response from API and convert response into json

```
response = requests.get(spacex_url)

data = pd.json_normalize(response.json())
```

Normalise json object and convert into dataframe

Use customized functions to filter dataframe

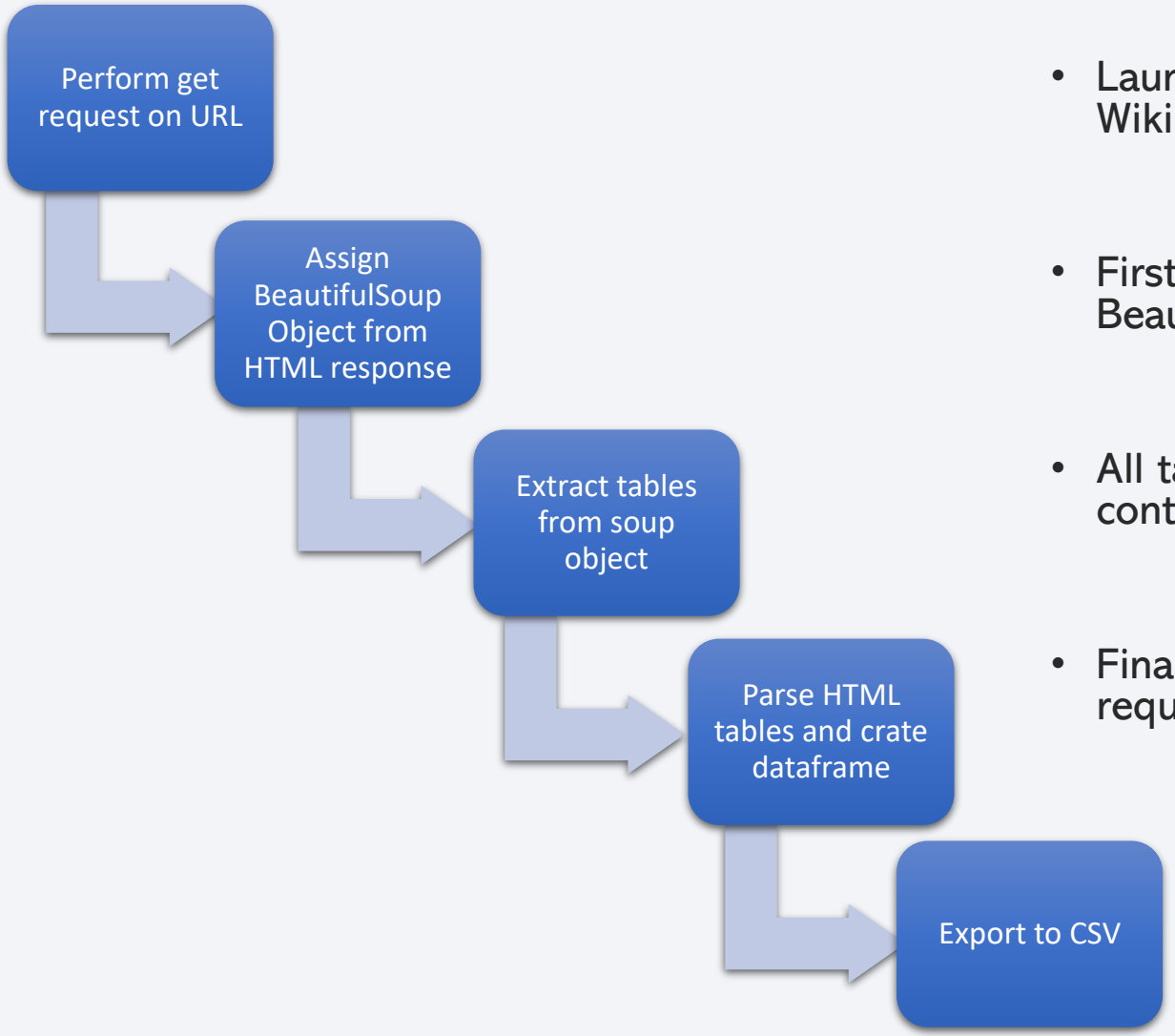
```
# Call getBoosterVersion
getBoosterVersion(data)

# Call getLaunchSite
getLaunchSite(data)

# Call getPayloadData
getPayloadData(data)

# Call getCoreData
getCoreData(data)
```

Export dataframe to csv

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40

GitHub link

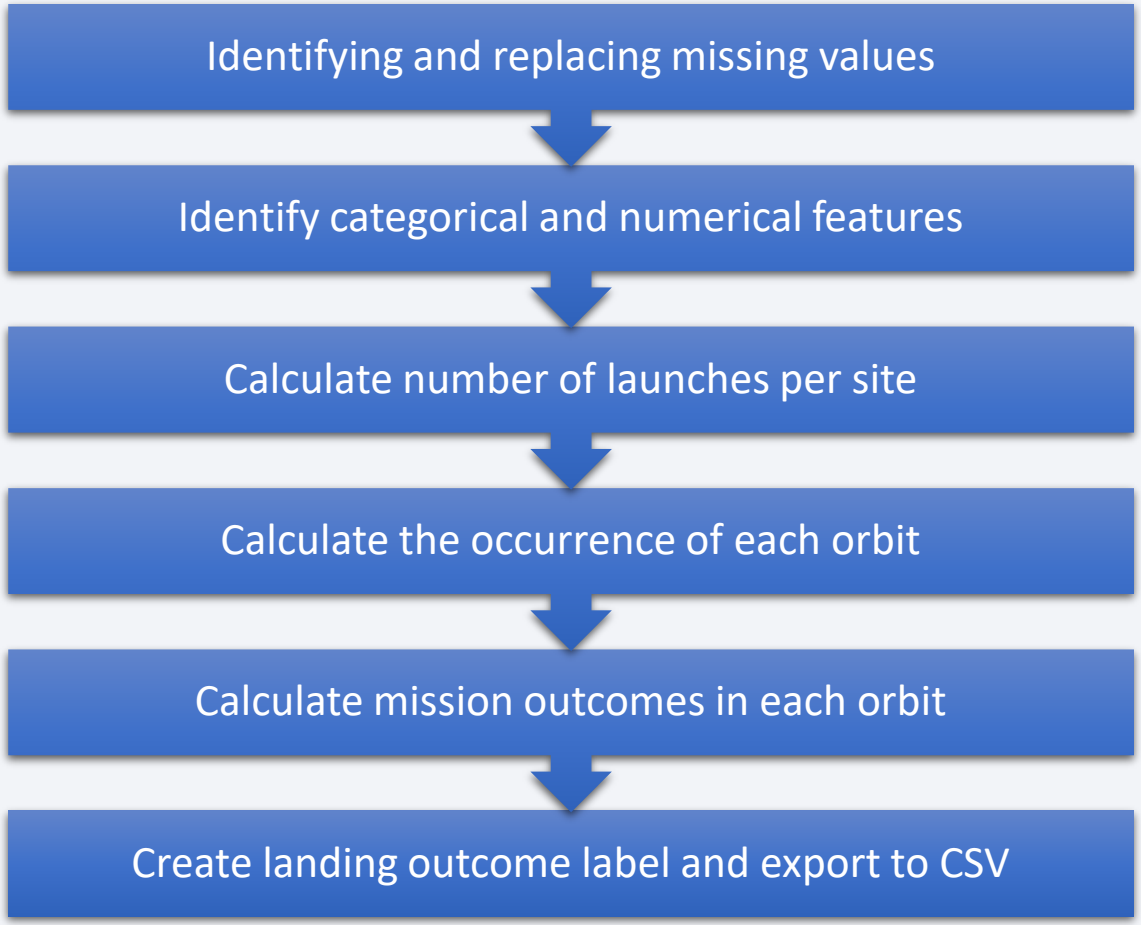
Data Collection - Scraping



- Launch records of Falcon 9 are web-scraped from a Wikipedia page using BeautifulSoup.
- Firstly, the get-request response is assigned to a BeautifulSoup object.
- All tables are extracted from the HTML response content.
- Finally, the HTML tables are parsed to create the required dataframe.

GitHub link

Data Wrangling



- Missing values are identified and cleaned by replacing or dropping them.
- Value counts of launch sites and orbits based on mission outcome are determined.
- Finally, training labels are created as per the Outcome column.

GitHub link

EDA with Data Visualization

- **Categorical Plots/ Scatter plots –**

Four plots are used to analyze the effect of launch outcome on

i. *Flight Number vs Payload Mass*
ii. *Flight Number vs Launch Site*
iii. *Flight Number vs Orbit Type*
iv. *Payload Mass vs Orbit Type*

- **Line plot –**

A line chart is used to analyze the yearly launch success trend.

- **Bar Graph –**

A bar graph is used to determine the success rate of each orbit type.

GitHub link

EDA with SQL

SQL queries have been used to determine the following –

- Names of unique launch sites
- 5 records where launch sites begin with ‘CCA’
- Total payload mass carried by boosters of NASA(CRS)
- Average payload mass carried by booster version F9v1.1
- Date of first successful landing date on Ground Pad
- Booster names, with successful landings on drone ships, having payload mass between 4000 and 6000
- Total successful and failed mission outcomes
- Booster versions that have carried maximum payload mass
- Failed landings, their booster versions and launch sites in 2015
- Ranking count of successful landings between 04-06-2010 and 20-03-2017

GitHub link

Build an Interactive Map with Folium

Using the latitude and longitude coordinates for each launch site, several Folium map objects are used to gain more insight into the data.

- **Circles** – Launch site locations are circled, with popups showing the names of the site.
- **Map Marker** – Locations of launch sites are marked using this object
- **Marker Color** – Red is used for failed launch outcomes while green indicates success.
- **Marker Cluster** – Launch outcomes are grouped using a cluster
- **Polylines** – Lines drawn from one location to another are used to signify the distance between locations and their importance

GitHub link

Build a Dashboard with Plotly Dash

Dash Components

- **Drop-down** – To select between different launch sites
- **Range Slider** – To select a payload mass range

Dash Callbacks

Multiple callbacks are used to interact with each dash component.

Plotly graphs

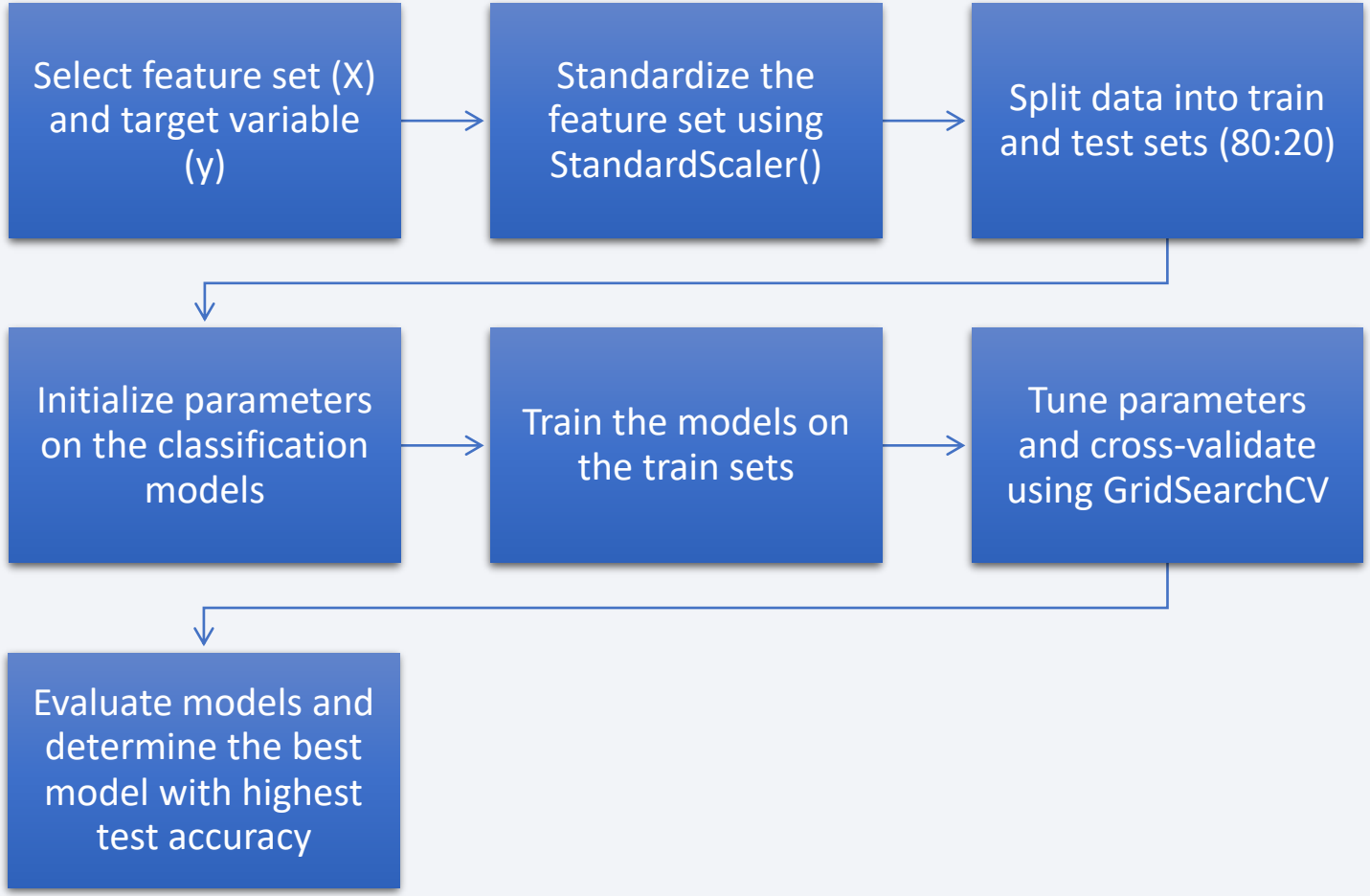
- **Pie chart** – To determine count of successful launches and success rates of launch sites
- **Scatter chart** – To determine correlation between payload successful launches for each booster version

GitHub link

Predictive Analysis (Classification)

Classification Models Used –

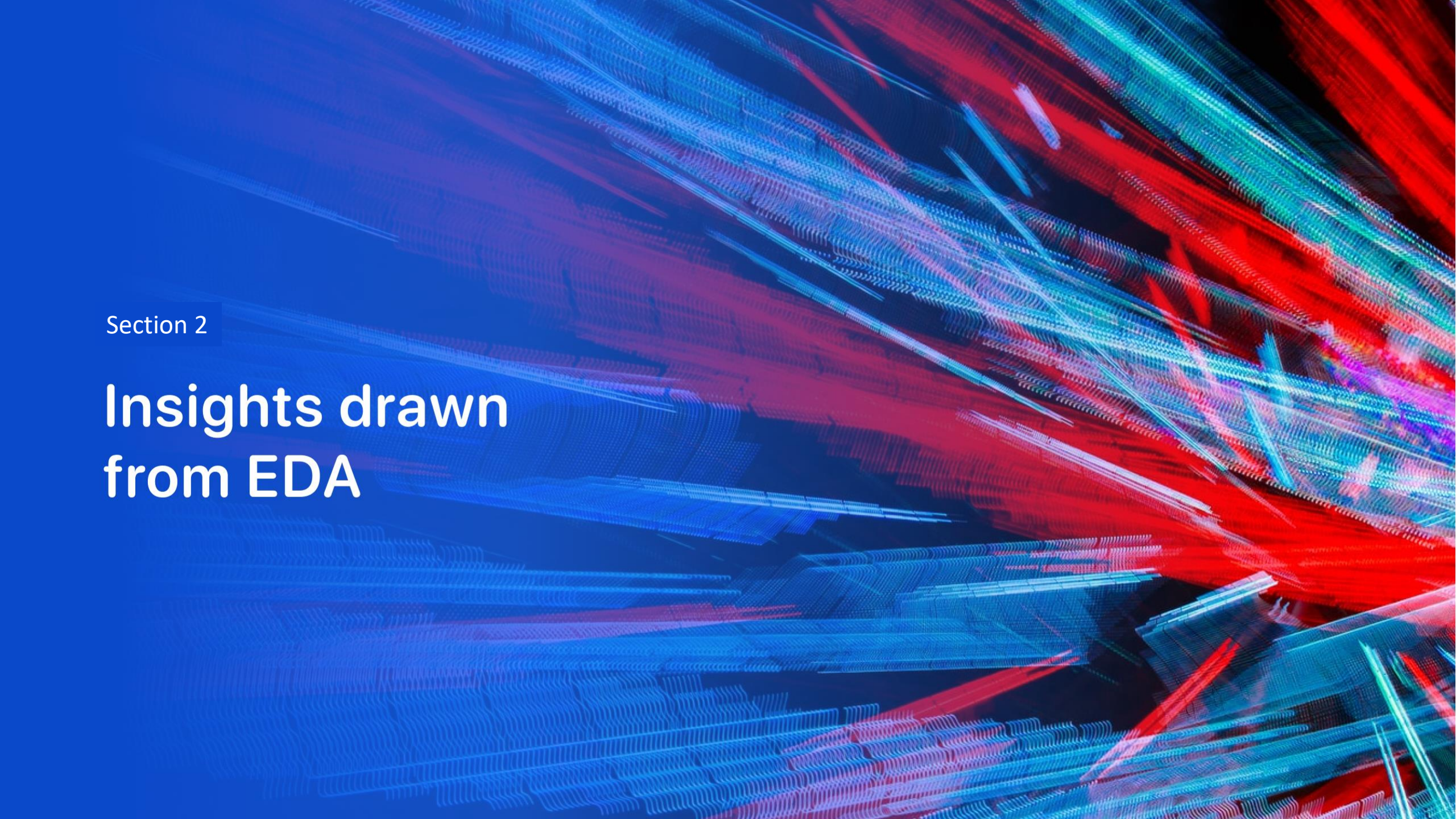
- Logistic Regression
- KNN
- SVM
- Decision Trees



GitHub link

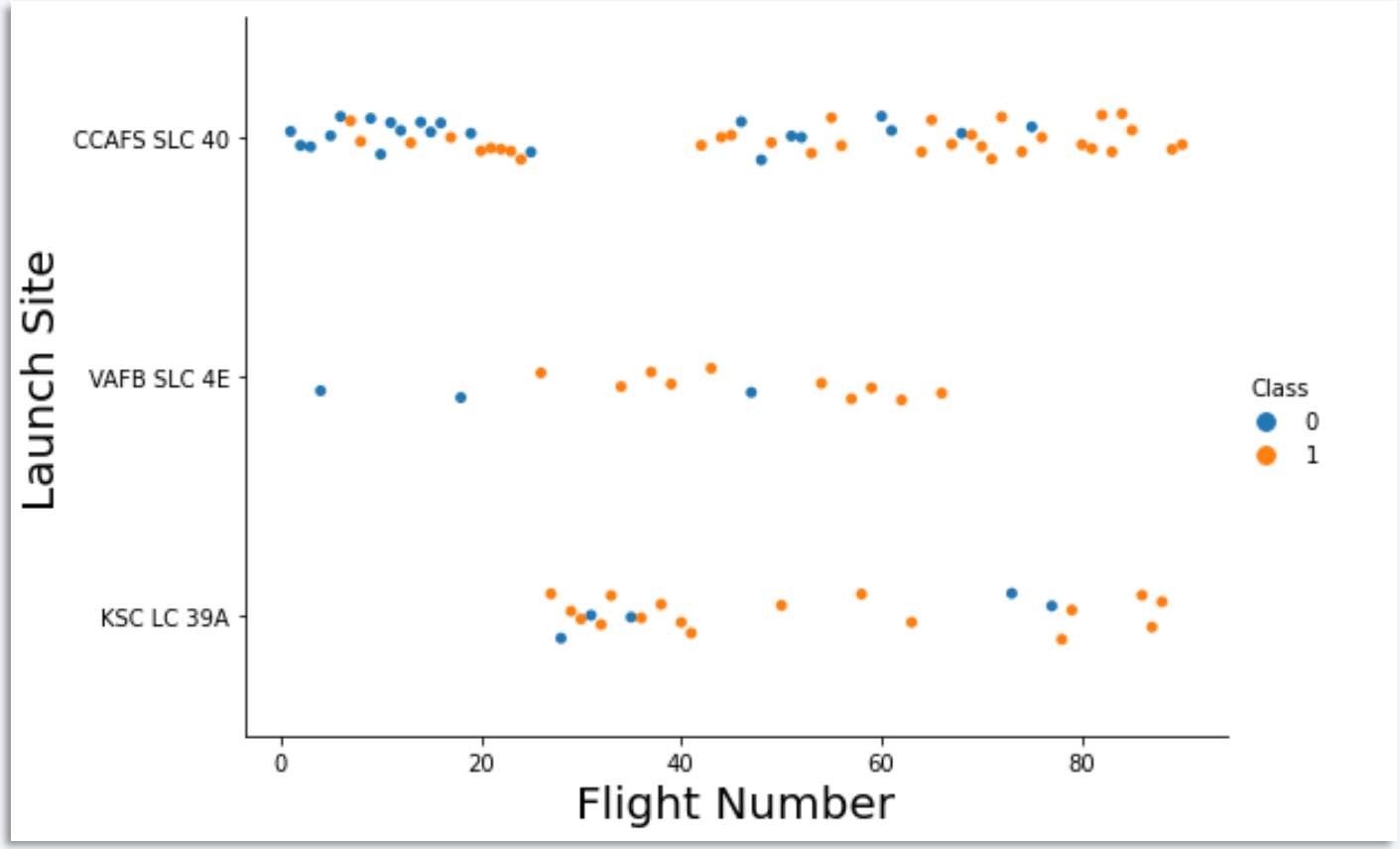
Results

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


Section 2
Insights drawn from EDA

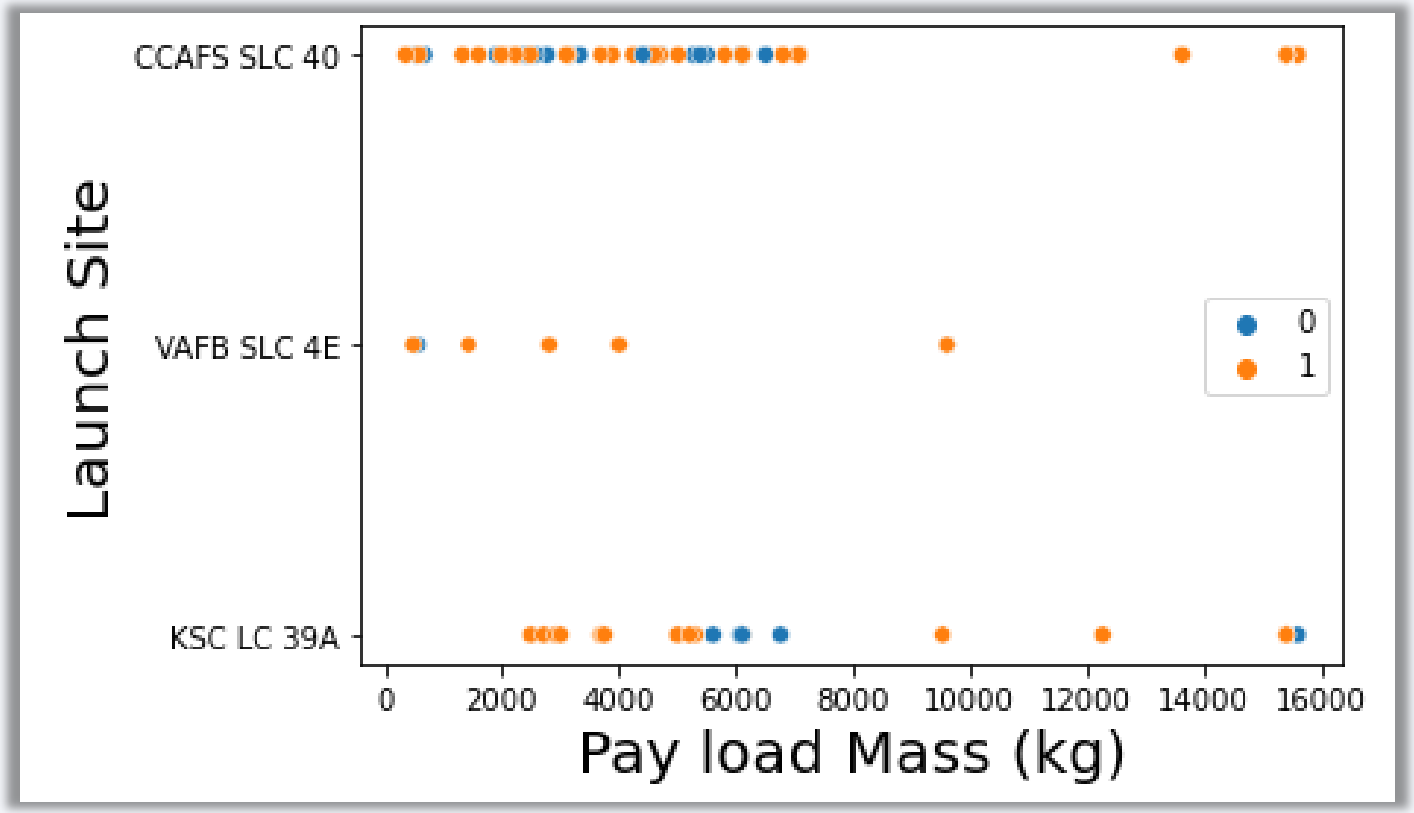
Flight Number vs. Launch Site

- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E have a success rate of 77%
- As Flight Number increases, the first stage is more likely to land
- Launch site CCAFS LC-40 has the most launches, while VAFB SLC 4E has not been used for the last 20 launches.



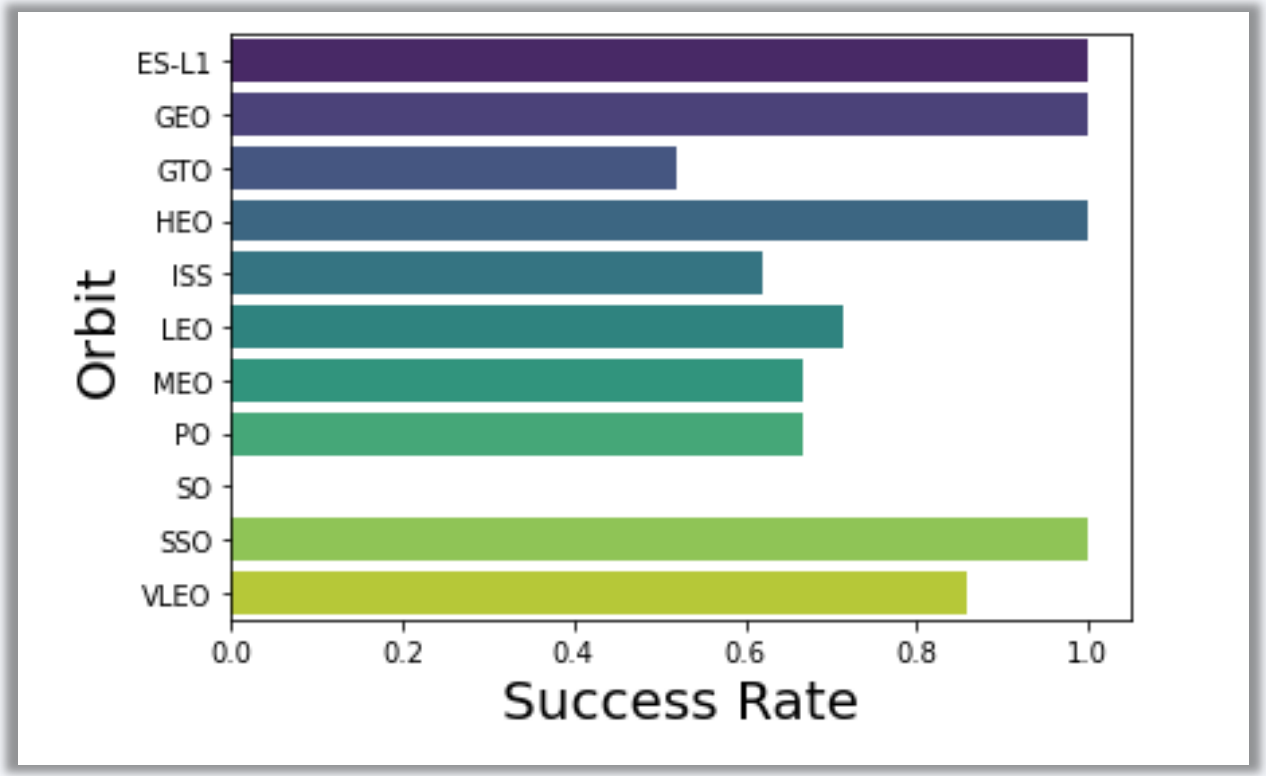
Payload vs. Launch Site

- In the case of VAFB-SLC, no rockets have been launched having payload mass greater than 10000.



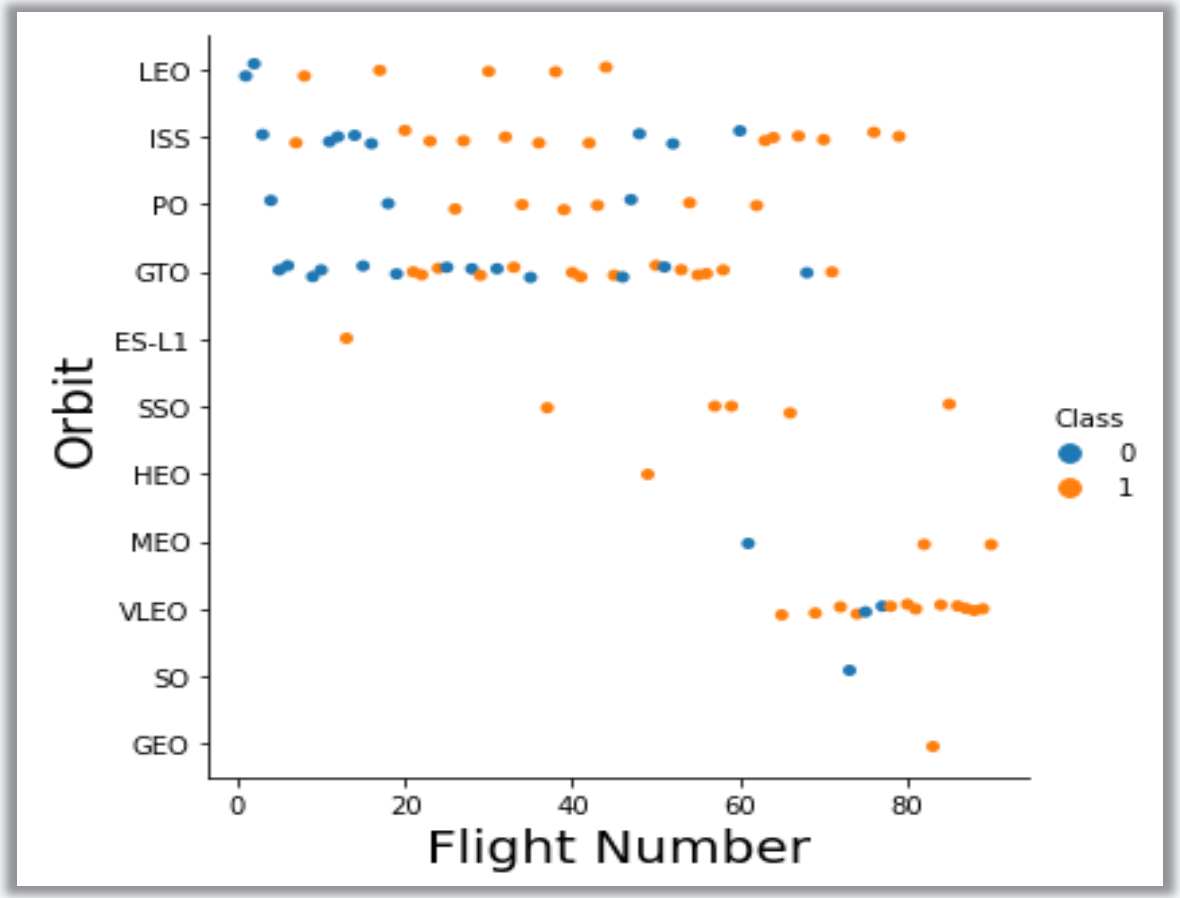
Success Rate vs. Orbit Type

- Orbit SO has a 0% success rate while ES-L1, GEO, HEO and SSO have a 100% success rate.



Flight Number vs. Orbit Type

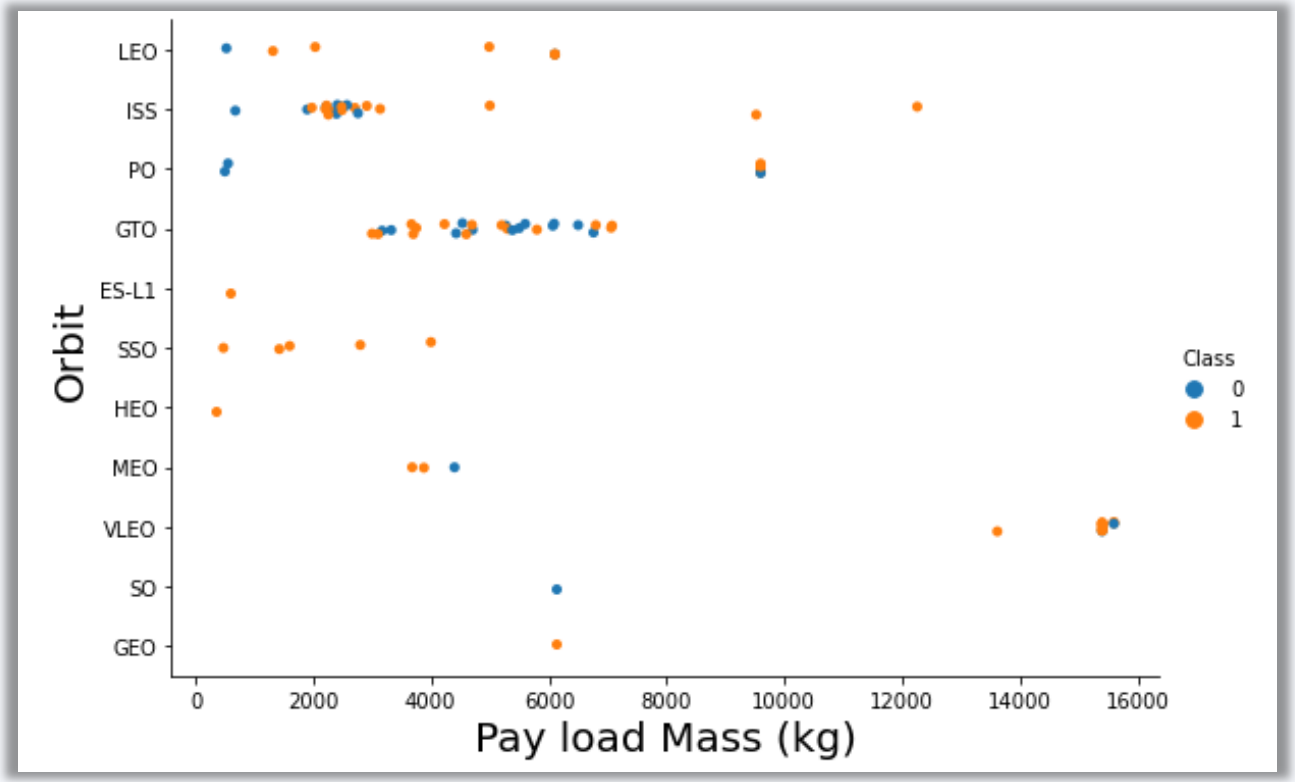
- In LEO orbit, as Flight Number increases, the success rate is high.
- For GTO orbit, there seems to be no relation with flight number



Payload vs. Orbit Type

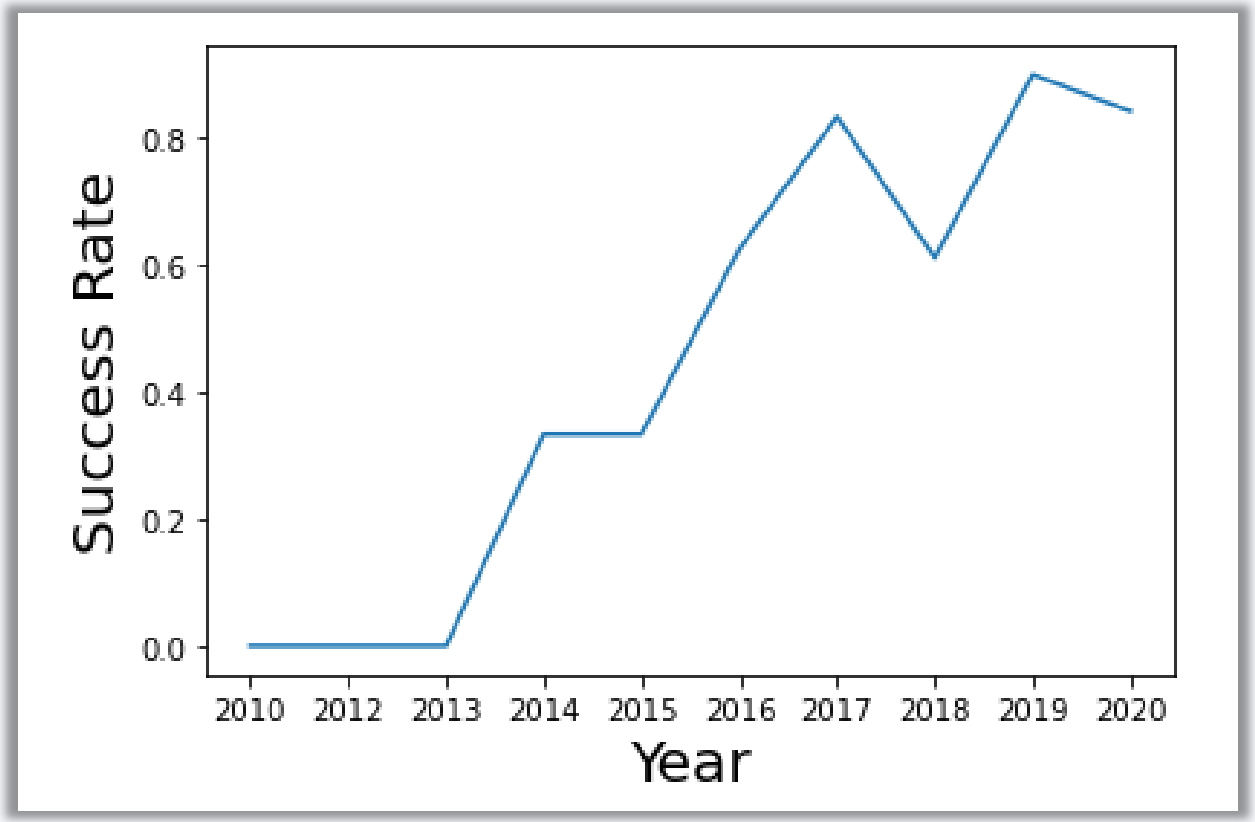
- In the case of heavy payloads, the success rate is comparatively higher for Polar, ISS and LEO orbit types.

- No relation can be found in the case of GTO orbit.



Launch Success Yearly Trend

- Since 2013, the success rate has been improving till 2020, except for a minor drop in 2018.



All Launch Site Names

- SQL Query

```
select distinct("Launch_Site") from SPACEXTBL;
```

- Description

  The distinct function helps to retrieve unique values from Launch_Site column, which is from table SPACEXTBL

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

Launch Site Names Begin with 'CCA'

- SQL Query

```
select * from SPACEXTBL
where Launch_Site like "CCA%"
limit 5;
```

- Description

  The where clause is used to filter launch sites starting with a given string while the limit function is used to display a given number of records.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass of NASA (CRS) Boosters

- SQL Query

```
select sum(PAYLOAD_MASS__KG_) as " Total payload mass carried by NASA (CRS) boosters" from SPACEXTBL
where Customer = "NASA (CRS)";
```

- Description

  The sum function adds up the values of the payload mass column where the Customer was NASA (CRS) boosters.

Total payload mass carried by NASA (CRS) boosters
45596

Average Payload Mass by F9 v1.1

- SQL Query

```
select avg(PAYLOAD_MASS__KG_) as "Average payload mass carried by booster version F9 v1.1" from SPACEXTBL
where Booster_Version = "F9 v1.1";
```

- Description

  The average function finds the average of the values of payload mass column filtered for F9 v1.1 booster version.

Average payload mass carried by booster version F9 v1.1
2928.4

First Successful Ground Landing Date

- SQL Query

```
select min(Date) as "First Succesful Landing - Ground Pad" from SPACEXTBL
where "Landing _Outcome" = "Success (ground pad)";
```

- Description

  The min function helps to retrieve the minimum value of the Date column (earliest in this case) where the Landing_Outcome value shows Success and landing type to be Ground Pad.

Average payload mass carried by booster version F9 v1.1
2928.4

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL Query

```
select distinct(Booster_Version) as "Booster Ver Succesful Landing - Drone Ship"
from SPACEXTBL
where "Landing _Outcome" = "Success (drone ship)"
and
PAYLOAD_MASS__KG_ between 4000 and 6000;
```

- Description

  The and operator helps to add two filters involving Landing_Outcome column as Success on Drone Ship landing and payload mass column as a value between 4000 and 6000 in order to find unique booster versions.

Booster Ver Succesful Landing - Drone Ship
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- SQL Query

```
select sum(Mission_Outcome like "Success%") as "Mission Outcomes - Successful",
       sum(Mission_Outcome like "Failure%") as "Mission Outcomes - Failure"
from SPACEXTBL;
```

- Description

  The like function helps to filter with string indexing by finding values that start or end with given characters before or after a % sign respectively.

Mission Outcomes - Successful	Mission Outcomes - Failure
100	1

Boosters Carried Maximum Payload

- SQL Query

```
select distinct(Booster_Version) from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_)
                           from SPACEXTBL);
```

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

- Description

  The where function filters booster versions whose payload is a value which in turn is retrieved by another query (sub-query) that helps to find the maximum payload mass value for all booster versions.

2015 Launch Records

- SQL Query

```
select  substr(Date, 4, 2) as "Month",
        "Landing _Outcome" as 'Landing Outcome',
        Booster_Version, Launch_Site
from SPACEXTBL
where "Landing _Outcome" like "Failure (drone ship)"
and
Date like "%2015";
```

- Description

  The substr function finds a subset of a string, such as retrieving month values (05) from Date (01-05-2005). The and operator adds two filters to find month, launch sites and booster versions of failed landings on drone ships in 2015.

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

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

- SQL Query

```
select "Landing _Outcome", count("Landing _Outcome") as "Count"
from SPACEXTBL
where "Landing _Outcome" like "Success%"
and
substr(Date,7,4)||substr(Date,4,2)||substr(Date,1,2) between '20100604' and '20170302'
group by "Landing _Outcome"
order by count("Landing _Outcome") desc;
```

- Description

  The group by clause groups the table by landing outcome values in order to aggregate the total count of successful landing outcomes between given dates sorted in descending order. The substr function concatenates year, month and day strings to filter the date range.

Landing _Outcome	Count
Success (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

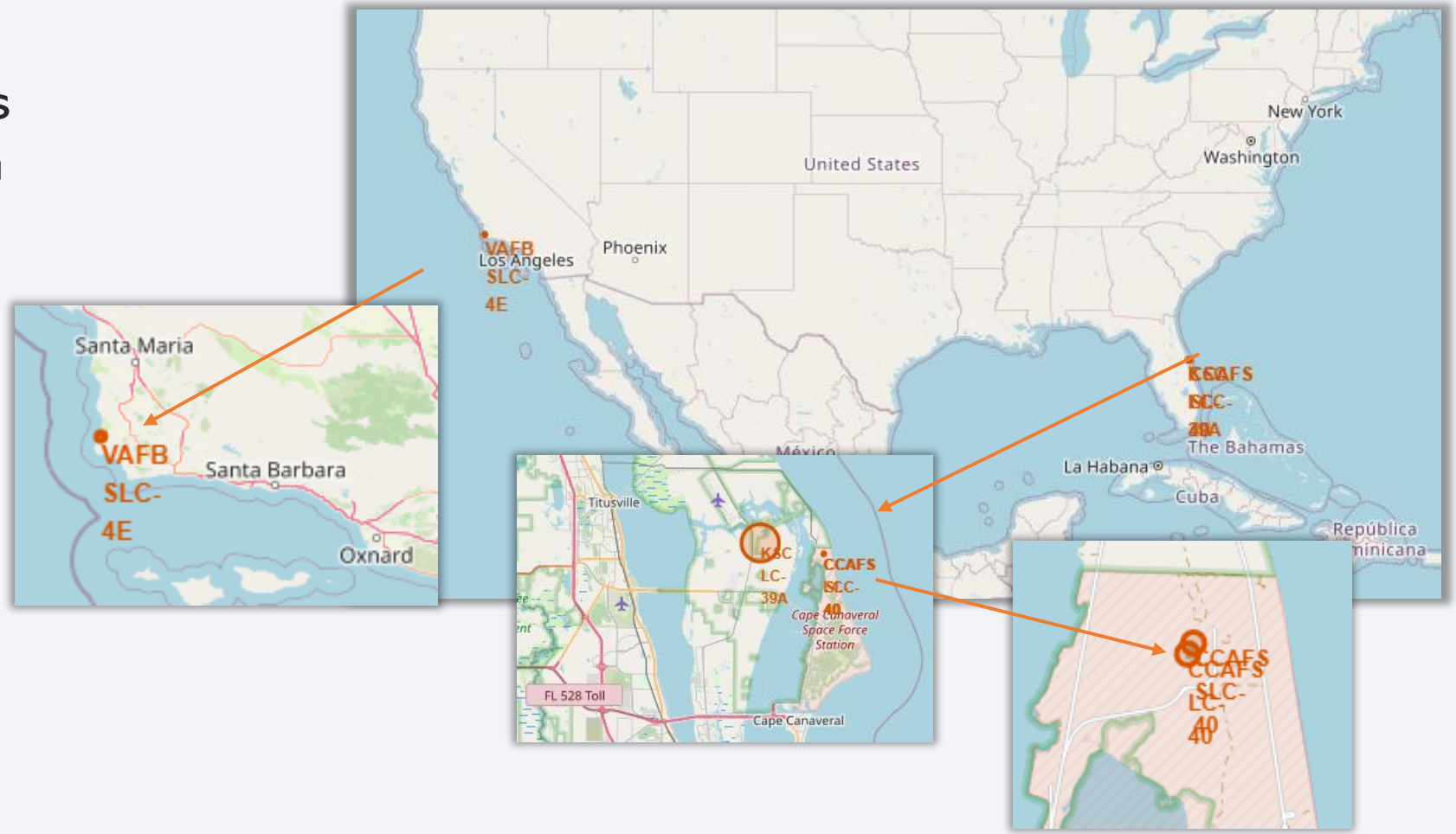
- Total of 4 Launch Sites locations visualized on the global map.

VAFB SLC-4E – California

KSC LC-39A – Florida

CCAFS SLC-40 – Florida

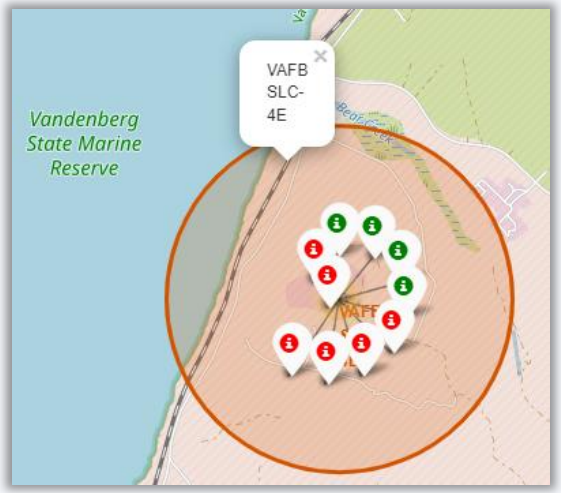
CCAFS LC-40 – Florida

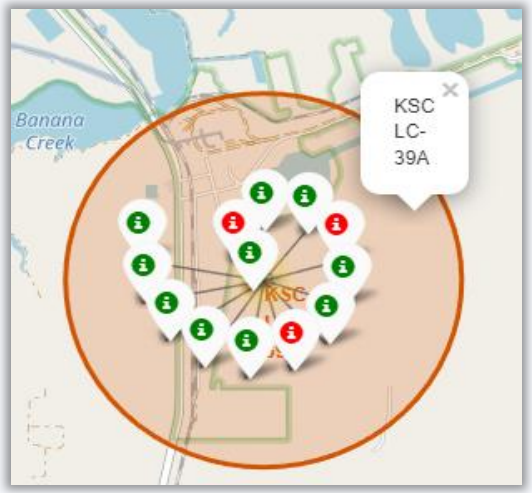


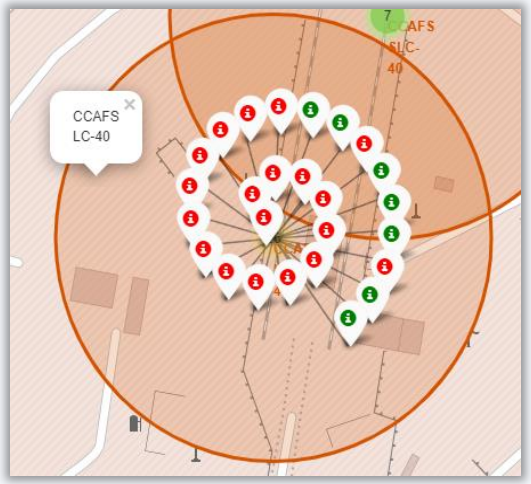
Color-Labeled Launch Outcomes

Green marker indicates successful launch outcomes while red marker indicates failures.

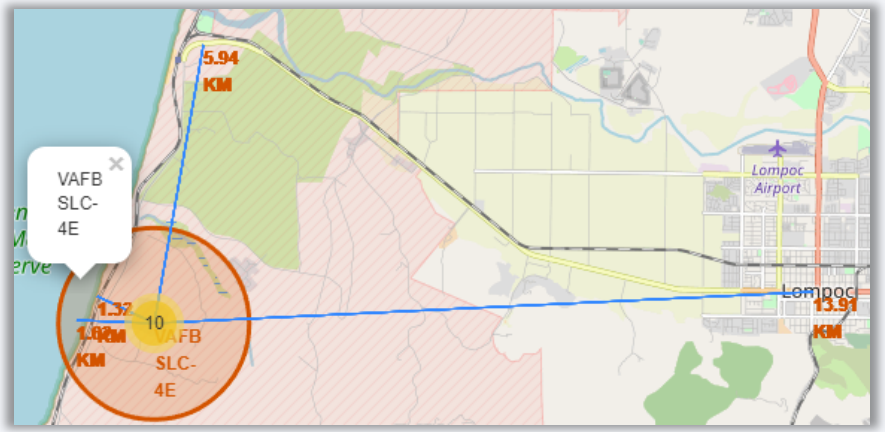
Results indicate that KSC LC-39A has a higher success rate.



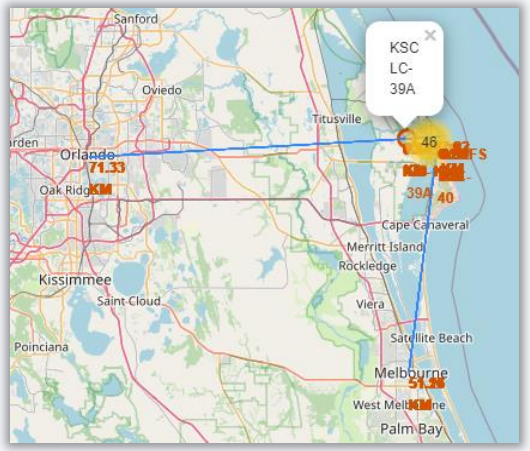


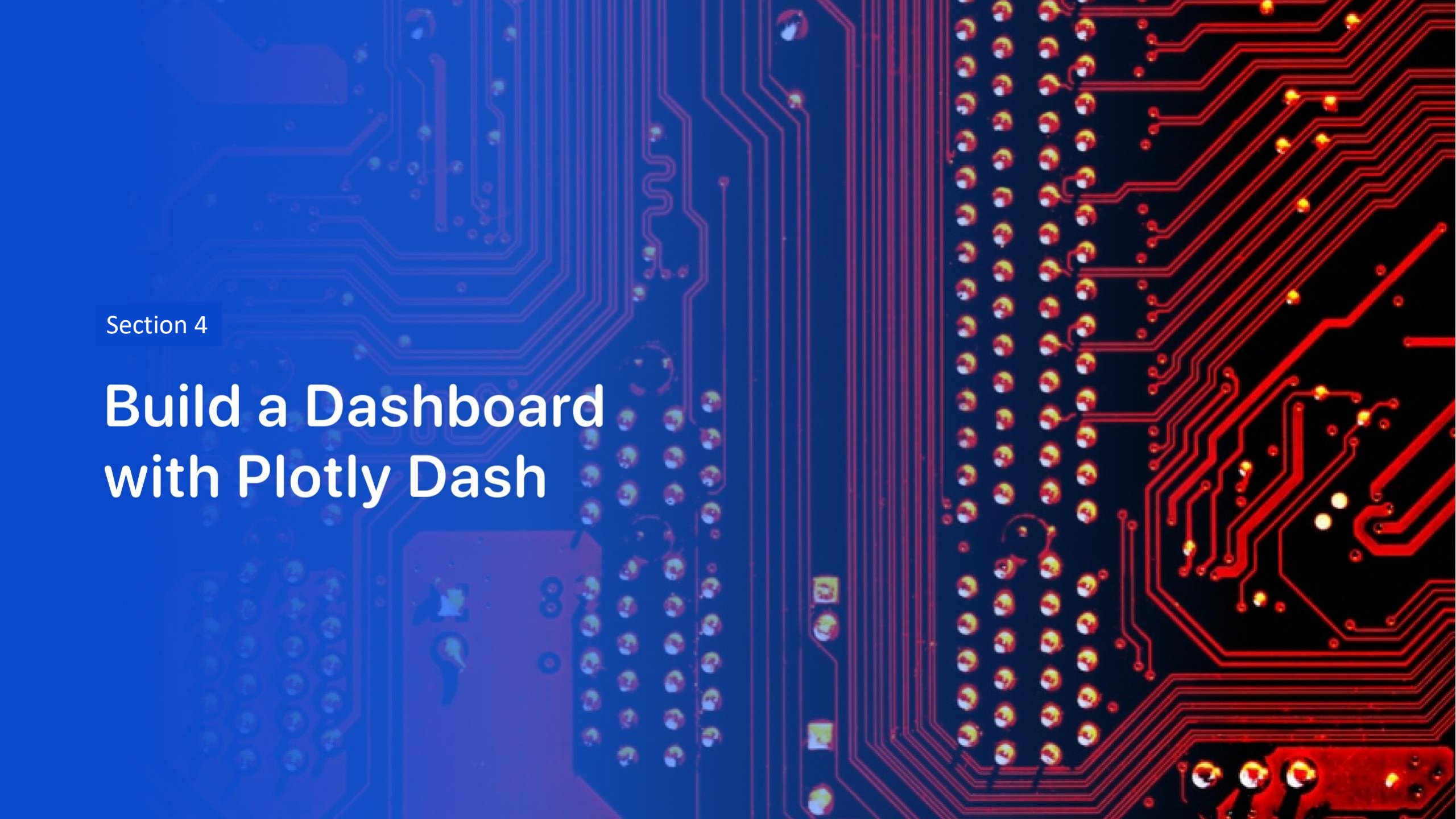


Distance from Site to Close Proximities

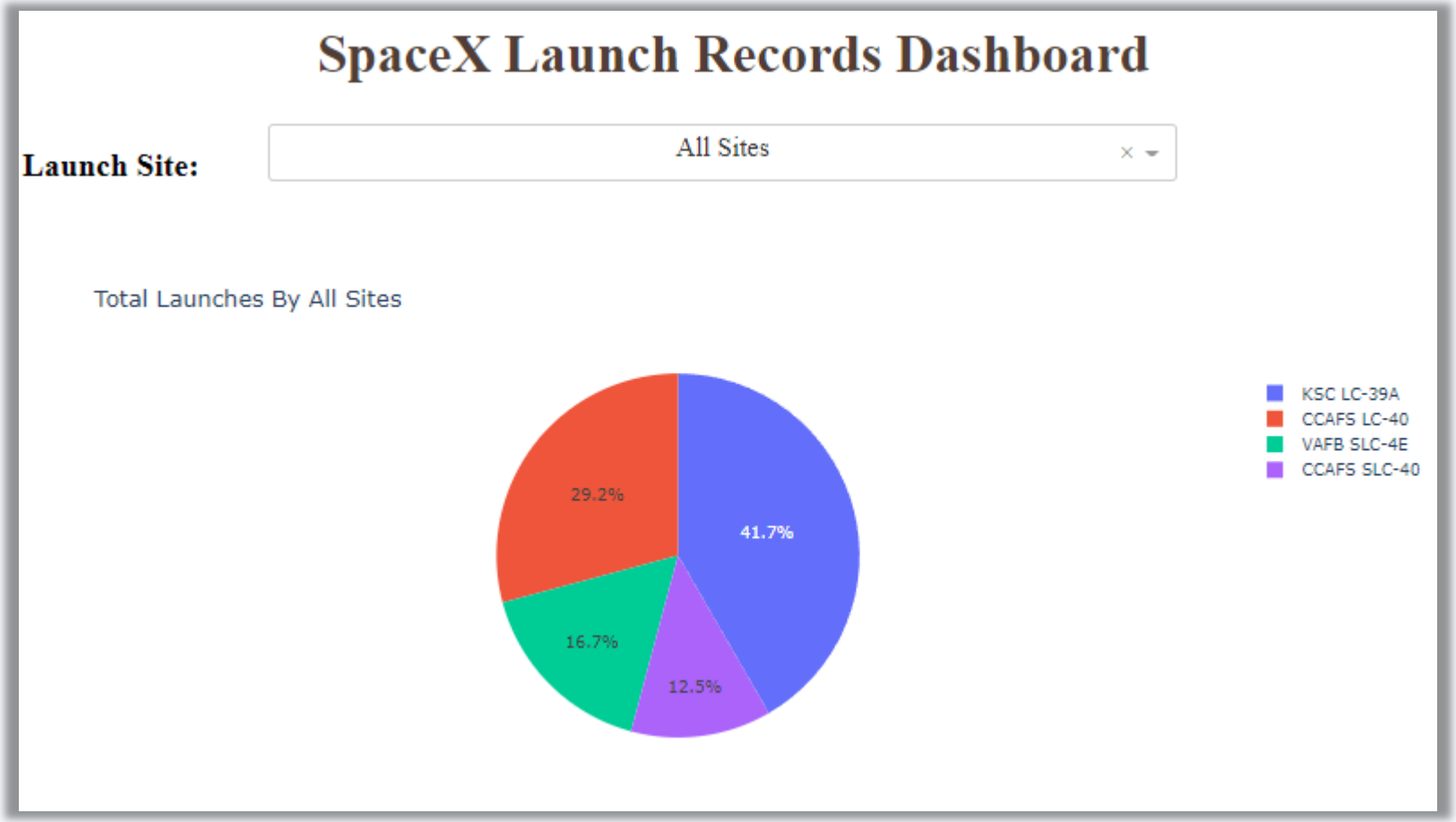


- Railways are in close proximity, roughly within 1.5 km.
- Highways are in close proximity, roughly within 1 to 6 km.
- Coastlines or any water body are in close proximity, roughly within a kilometer.
- Launch sites are far away from cities, with proximity ranging from 14 to 70 km.



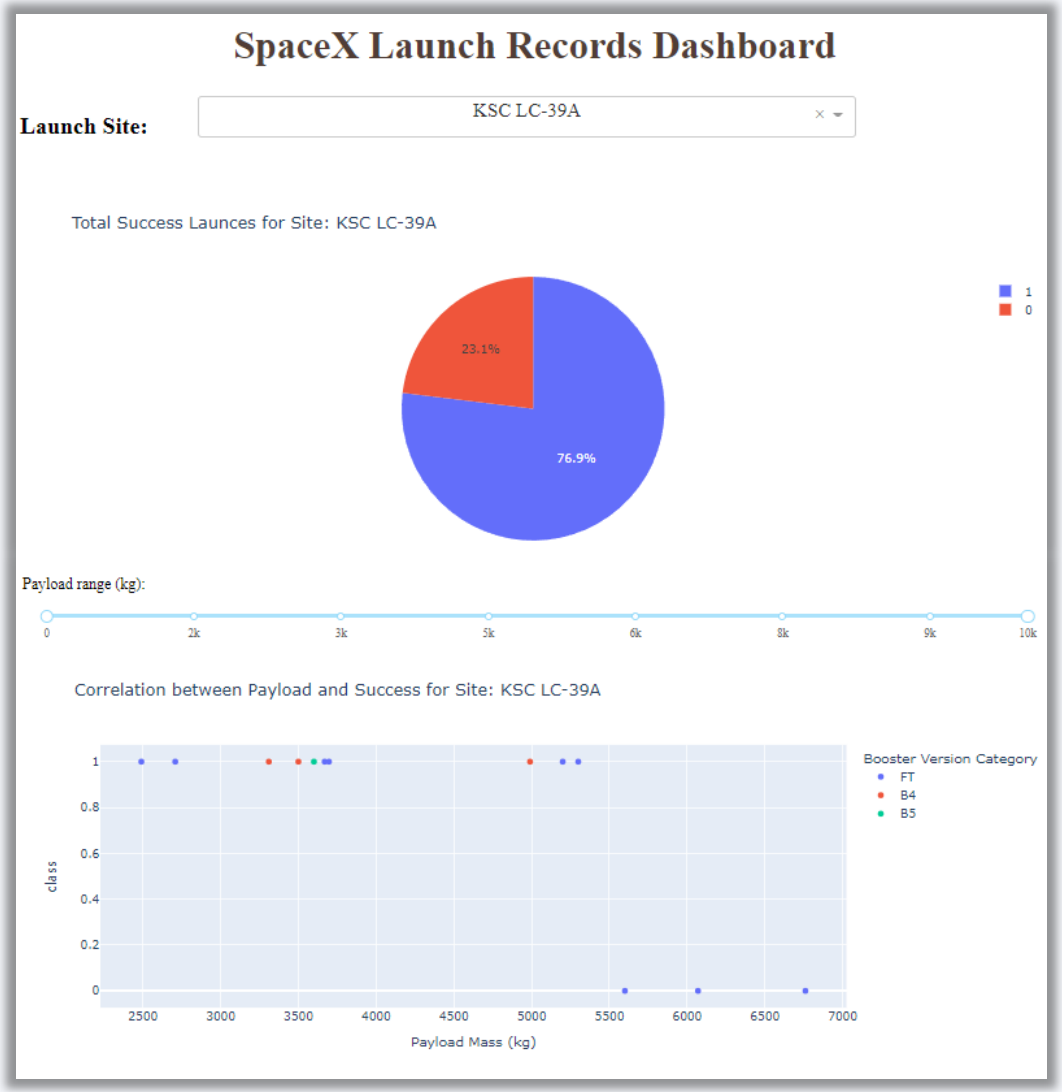

Section 4
Build a Dashboard with Plotly Dash

Launch Success Rate for All Sites



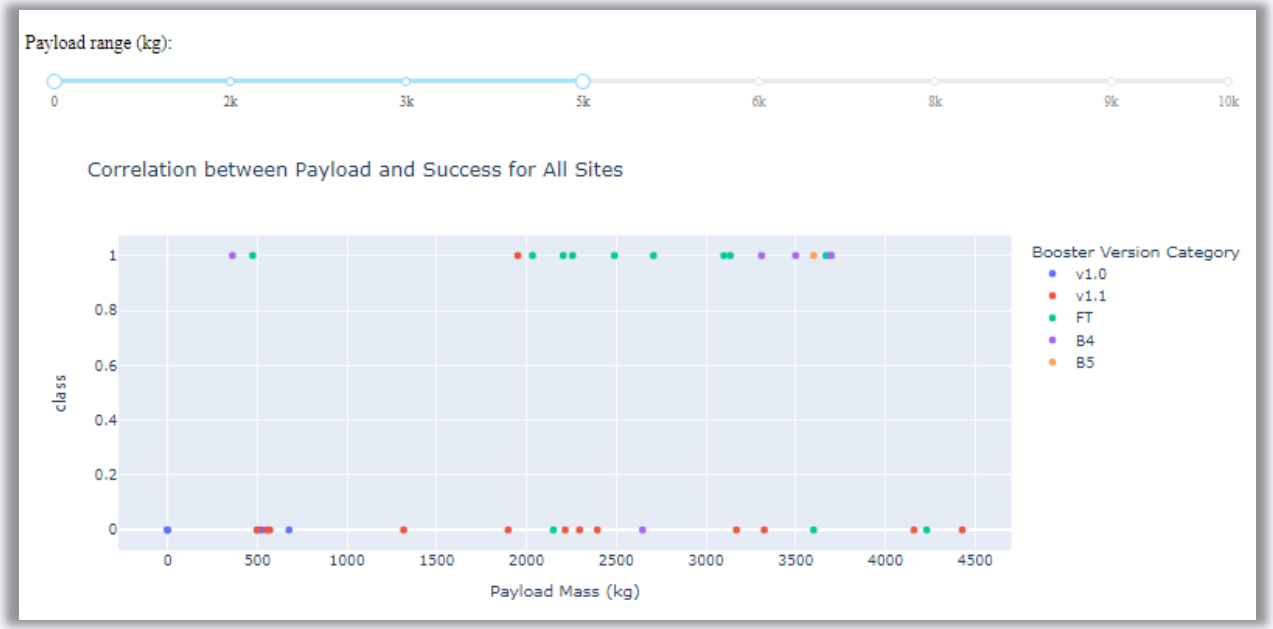
- Among all sites, KSC LC-39A had the most successful launches

Launch Site with Highest Success Ratio



- KSC LC-39A had a 76.9% success rate.
- All failed launches had –
  i. Payload mass above 5500 kg
  ii. FT booster version category

Payload Mass vs Launch Outcome



- Heavier payloads (5000 – 1000 kg) have a lower success rate than lighter payloads (0 – 4000 kg).

Section 5

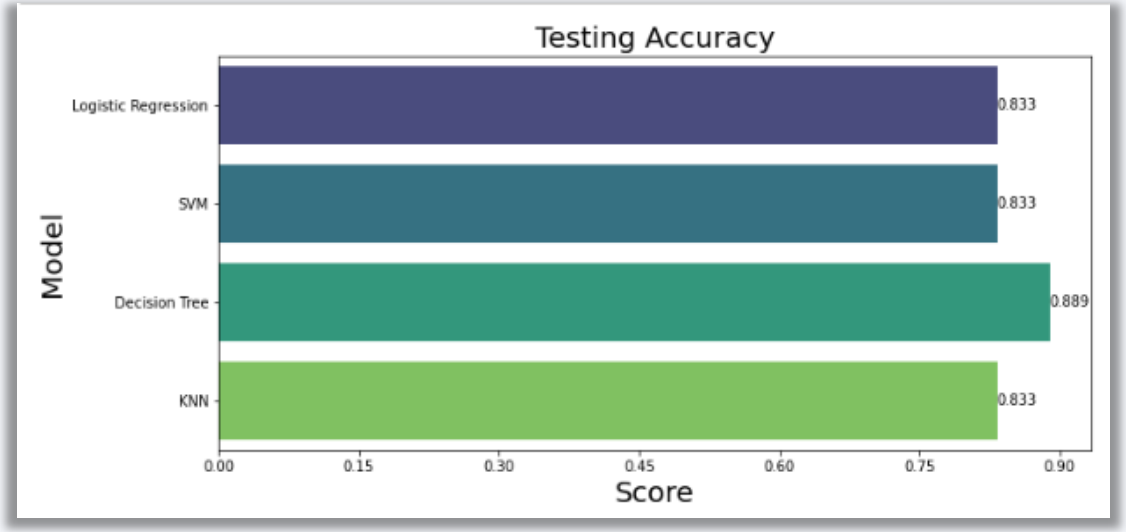
Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree offers the best results with a testing accuracy of approx. 88.9%

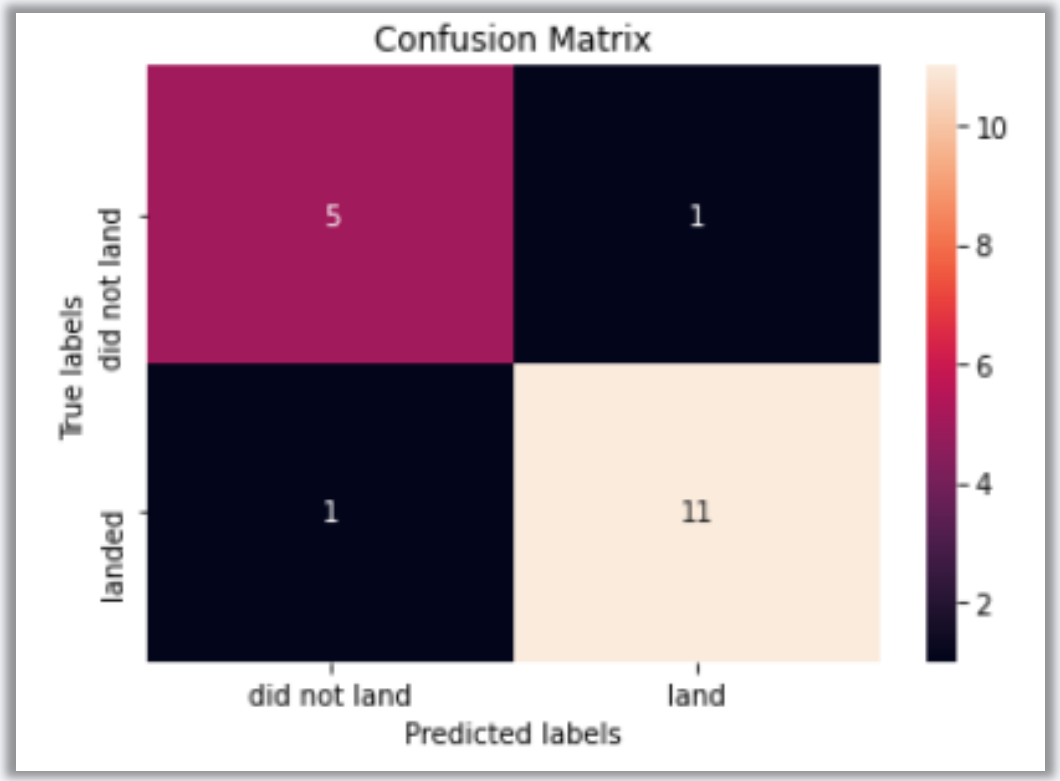
- Parameters:

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}
```



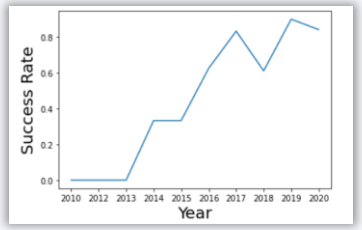
Confusion Matrix

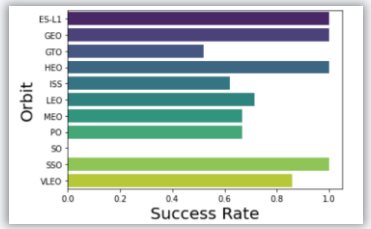
- Confusion Matrix of the Decision Tree indicates –
  i. Accuracy: (TP+TN)/Total = (11+5)/18 = 0.888
  ii. Misclassification Rate: (FP+FN)/Total = (1+1)/18 = 0.111
  iii. Precision: TP/(TP+FP) = 11/12 = 0.916
  iv. Recall: TP/(TP+FN) = 11/12 = 0.916

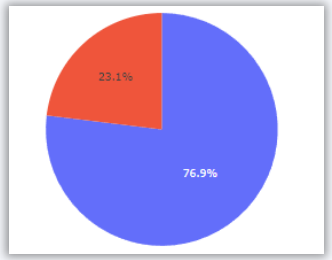


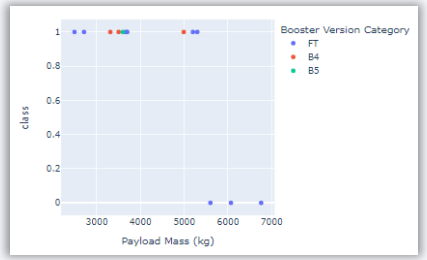
Conclusions

- Launch Success Rate has been improving each passing year (except 2018).
- Orbit SP is yet to register a successful launch outcome.
- KSC LC-39A launch site had the most success with lighter payloads.
- Decision Tree Classifier offers the best results in predicting launch outcomes from historic data.

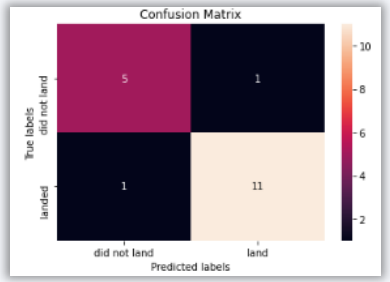








Model	Train Accuracy	Test Accuracy
Logistic Regression	0.846	0.833
SVM	0.848	0.833
Decision Tree	0.877	0.889
KNN	0.848	0.833



Thank you!