

Winning Space Race with Data Science

JY Chan
19th August 2022



Outline

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

Executive Summary

Objective:

To develop a predictive model to determine whether the SpaceX Falcon 9 stage 1 rocket will successfully land based on a given set of features such as launch site, booster version, orbit, payload mass etc.

Methodologies:

- Data collection via SpaceX API and web scraping of Wikipedia page.
- Exploratory Data Analysis (EDA) i.e. data wrangling, data analysis via SQL and data visualization, and building interactive visual analytics via Plotly Dash.
- Development of Machine Learning predictive model.

Results:

- Results from EDA revealed the relationship among the features and shed light on features that are ideal in predicting the success of landing.
- Machine learning predictive model is able to predict the outcome of rocket landing with good accuracy. 3

Introduction

1. Project Background and Context:

The goal of this project is to predict if the Falcon 9 first stage rocket will successfully land based on a given set of features. SpaceX claims that the Falcon 9 rocket launch costs \$62 million, which is less than half the cost of other provider i.e. north of \$165 million. The lower cost is attributed to the fact that SpaceX reuses the first stage of the rocket. Thus, by determining if the rocket landing will be successful or not, we can estimate the cost of rocket launch, and this information is crucial for anyone who wants to compete with SpaceX to bid for rocket launch project.

2. Problems Statement:

- What are the main factors and characteristic of a successful or failed landing?
- What are the correlations of each features and the success or failure of landing?
- What are the conditions that will yield the highest success rate of Falcon 9 rocket landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - SpaceX API
 - Web scraping of Wikipedia page
- Perform data wrangling
 - Removing unnecessary columns, replace empty cell, add target column in the data frame, and one hot encoding of categorical data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - Building, tuning, and evaluation of various classification models

Data Collection

The data sets were collected via 2 sources:

(1) SpaceX REST API (Source: <https://api.spacexdata.com/v4/rockets/>)

- The information retrieved from the API includes rocket booster type, launch location, payload information etc.

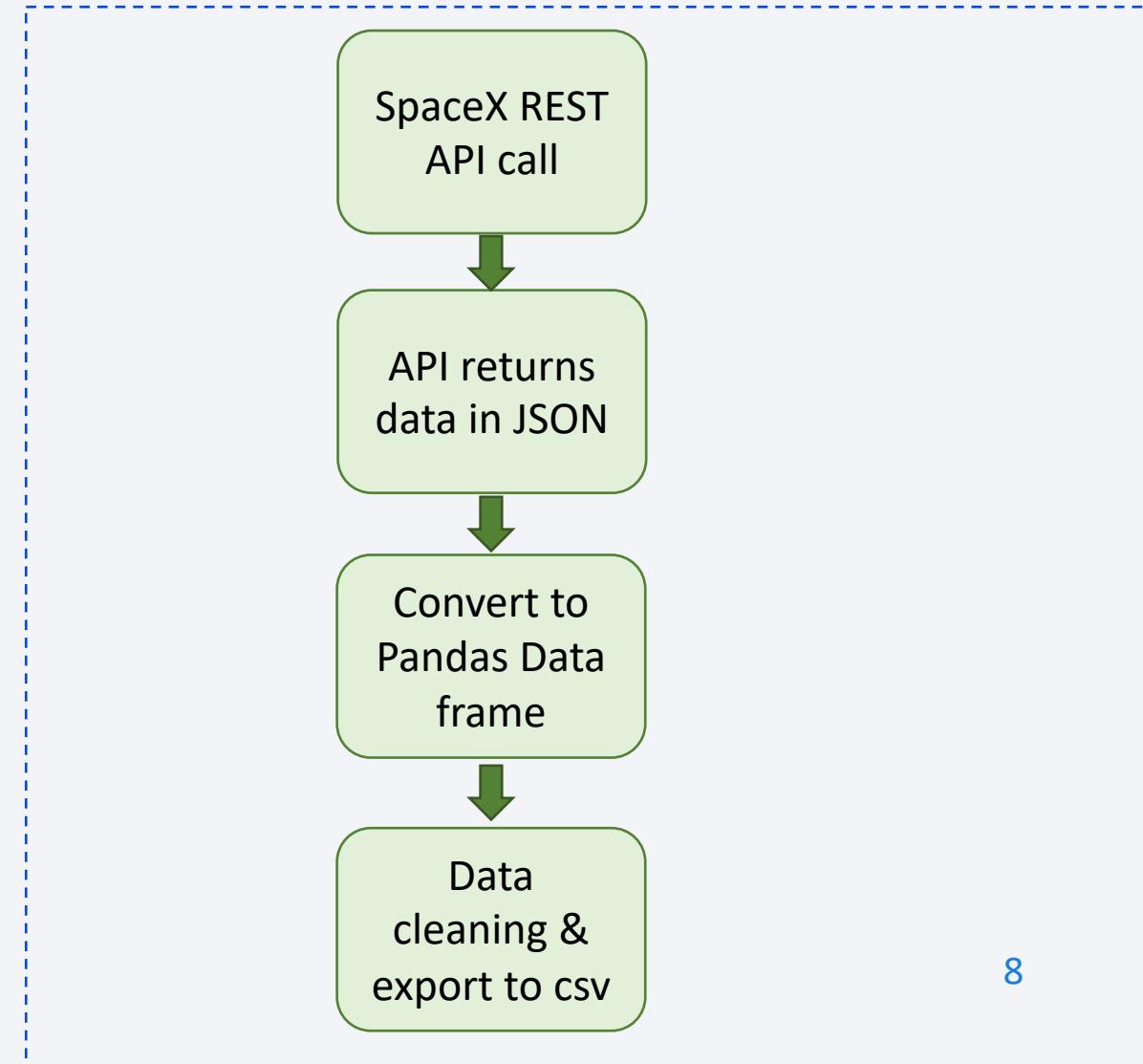
(2) Web scraping of Wikipedia page on SpaceX rocket launch

- Source URL https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Launches [edit]									
2010 to 2013 [edit]									
[hide] Flight No.	Date and time (UTC)	Version, Booster [a]	Launch site	Payload ^[5]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0 ^[2] B0003 ^[3]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit	No payload (excl. Dragon Mass)	LEO	SpaceX	Success	Failure ^{[4][5]} (parachute)
First flight of Falcon 9 v1.0. ^[6] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage. ^(more details) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even got to deploy. ^[7]									
2	8 December 2010, 15:43 ^[8]	F9 v1.0 ^[2] B0004 ^[3]	CCAFS, SLC-40	SpaceX COTS Demo Flight 1 (Dragon C101)	Unknown (excl. Dragon Mass)	LEO (ISS)	NASA (COTS) various others ^[9]	Success ^[4]	Failure ^{[4][10]} (parachute)
Maiden flight of SpaceX's Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and then reentry. ^[11] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, again before the parachutes were deployed. ^[7] ^(more details) It also included eight CubeSats ^[10] and a wheel of Brie cheese. Before the launch, SpaceX discovered that there was a crack in the nozzle of the 2nd stage's Merlin vacuum engine. SpaceX cut off the end of the nozzle and got NASA's approval to fly in this configuration. ^{[11][2]}									
3	22 May 2012, 07:44 ^[13]	F9 v1.0 ^[2] B0005 ^[3]	CCAFS, SLC-40	SpaceX COTS Demo Flight 2 ^[14] (Dragon C102)	525 kg (1,157 lb) ^[15] (excl. Dragon mass)	LEO (ISS)	NASA (COTS)	Success ^[16]	No attempt
The Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS. ^[13] ^(more details)									
8 October 2012, 00:35 ^[17]	F9 v1.0 ^[2] B0006 ^[3]	CCAFS, SLC-40	SpaceX CRS-1 ^[18] (Dragon C103)	4,700 kg (10,400 lb) (excl. Dragon mass)	LEO (ISS)	NASA (CRS)	Success	No attempt	
			Orbcomm-OG2 ^[19]	172 kg (378 lb) ^[20]	LEO	Orbcomm	Partial failure ^[21]		

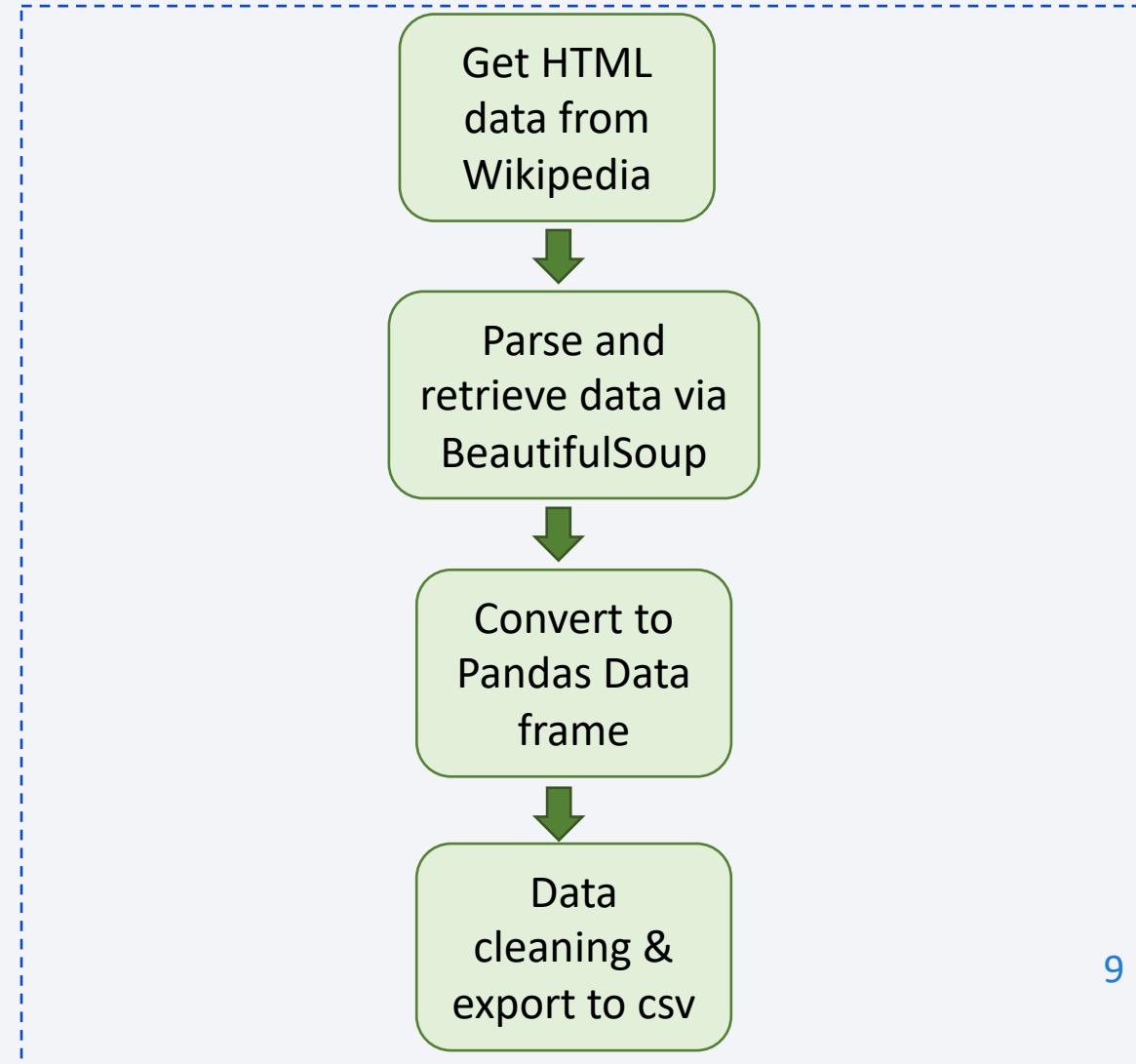
Data Collection – SpaceX API

- SpaceX public REST API allows the retrieval of launch data.
- The launch data is obtained via the Python Request package.
- Github URL: [Source Code](#)



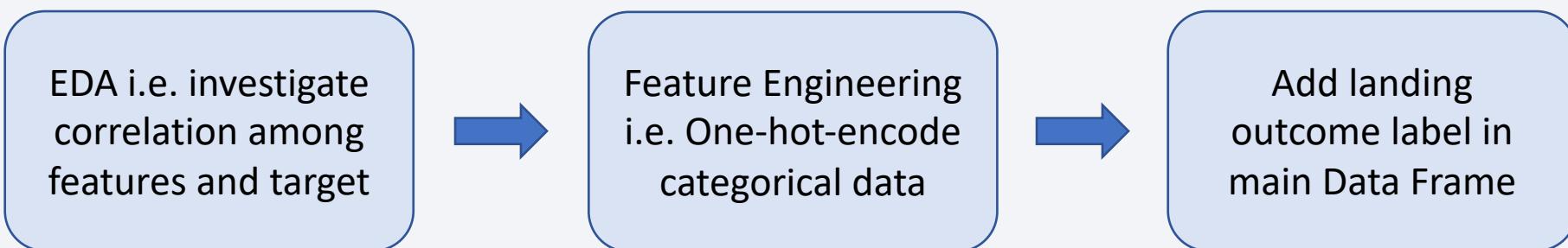
Data Collection - Scraping

- SpaceX rocket launch data is also available via web scraping of Wikipedia page.
- BeautifulSoup module is used to parse and retrieve the required info from the HTML text retrieved.
- Github URL: [Source Code](#)



Data Wrangling

- Exploratory Data Analysis (EDA) was performed on the launch dataset, which includes basic descriptive statistics on each features as well as exploration of correlation between the features and launch outcome.
- Key areas such as total launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were investigated.
- Categorical data such as launch site, orbit, Landing Pad and Serial were one-hot-encoded.
- Landing outcome label was created in the Outcome column on the main Data Frame whereby '1' represents successful landing and '0' represents fail landing.



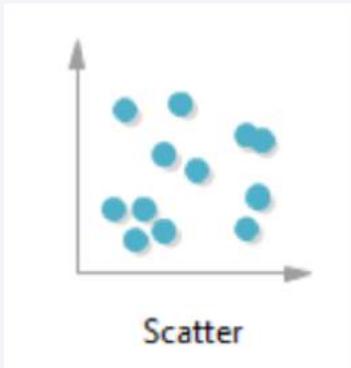
EDA with Data Visualization

Several charts were plotted with the data as summarized below:

(1) Scatter plot

- Flight number vs Payload mass
- Flight number vs launch site
- Orbit vs Flight number
- Payload vs Orbit type
- Orbit vs Payload Mass

*Scatter plot elicits the correlation among each of the features.



(2) Bar chart

- Landing success rate vs Orbit
- *Bar chart uncovers the relationship between numerical and categorical values.



(3) Line chart

- Landing success rate vs Year
- *Line chart depicts the trend of a data.



EDA with SQL

Summary of SQL queries performed on the dataset:

- Names of the unique launch sites in the space mission
- Top 5 launch sites whose name begin with the string 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- 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
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

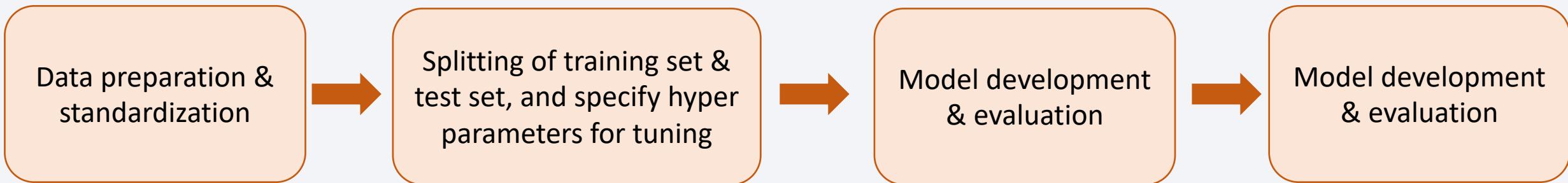
- Markers, circles, lines and marker clusters were used with Folium Maps.
- Markers are used to indicate points like launch sites.
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center or the launch site.
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site.
- Lines are used to indicate distances between two coordinates on the map.

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize the launch data:
 - Percentage of launches by site
 - Payload range
- These variables enabled one to quickly analyze the relationship between payloads and launch sites, thus helping to determine where is best site to launch the rocket according to payloads to secure a higher success rate.

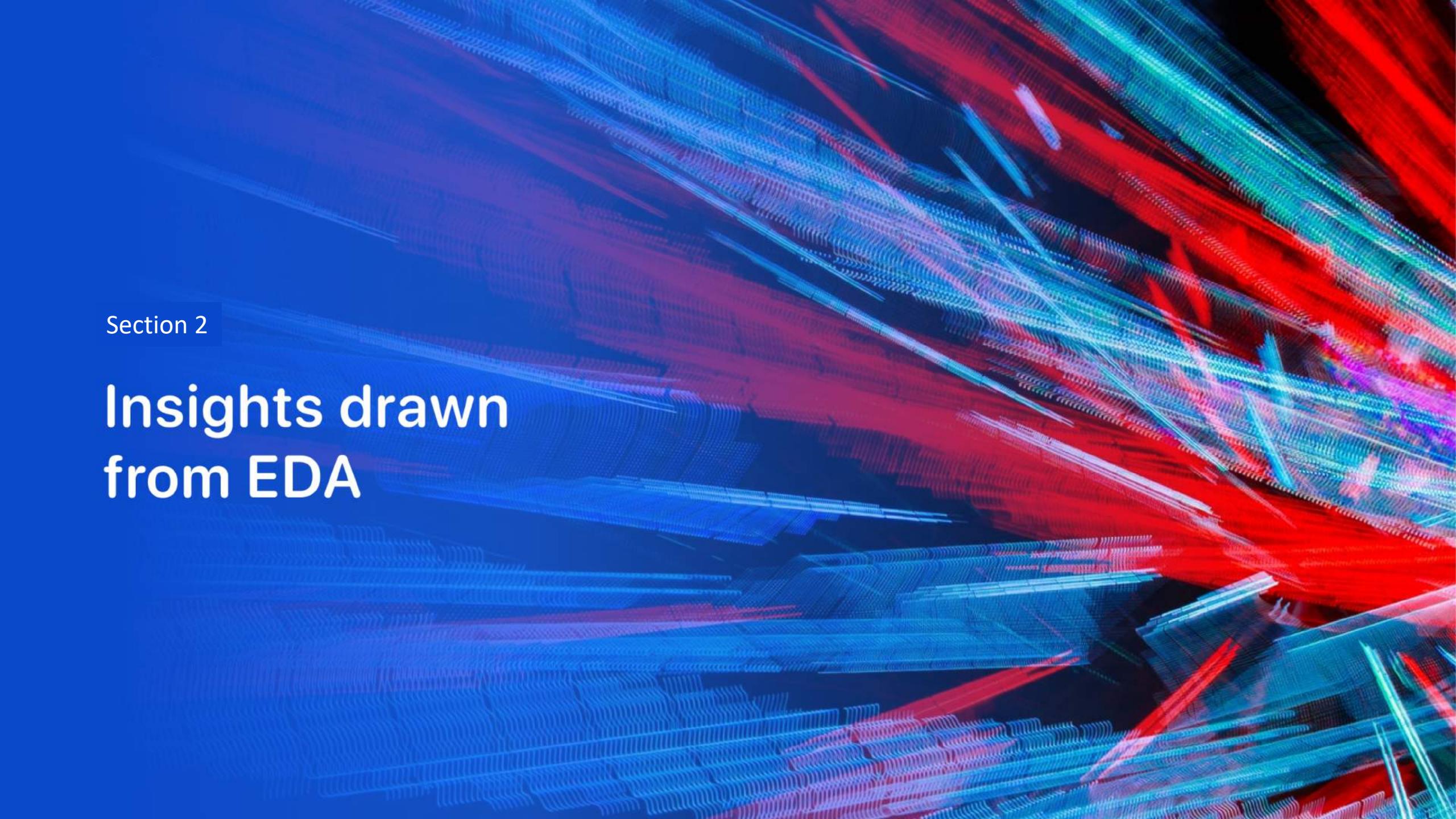
Predictive Analysis (Classification)

- 4 predictive models (classification) were developed to forecast the landing outcome, namely
 - (1) Logistic Regression
 - (2) Support Vector Machine (SVM)
 - (3) Decision Tree
 - (4) K-nearest neighbor
- The models were each trained on training set (80% of the main data), and the hyper parameters were tuned via Grid Search method.
- The models were evaluated based on the in-sample accuracy, accuracy on test data, as well as via a confusion plot to identify if there were any false positive result predicted by the models.



Results

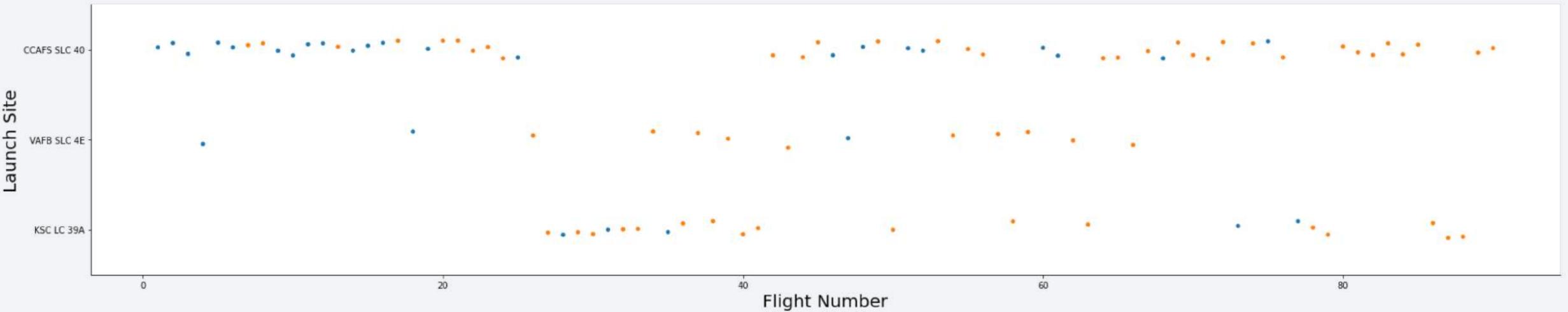
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that is more dense and vibrant towards the right side of the frame, while appearing more sparse and blue-tinted on the left. The overall effect is reminiscent of a high-energy particle simulation or a futuristic circuit board.

Section 2

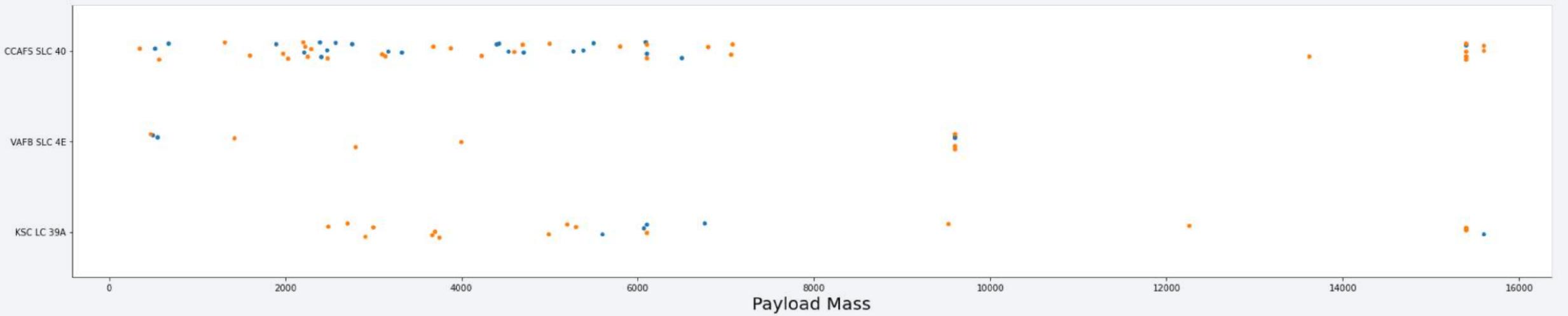
Insights drawn from EDA

Flight Number vs. Launch Site



- From the plot, site CCAFS SLC 40 has the most rocket launch, followed by KSC LC 39A and VAFB SLC 4E.
- Generally, the launch success rate for each launch site improves over time.

Payload vs. Launch Site



- Launch with payload of roughly 9,000 kg has higher success rate.
- Payload which is heavier than 10,000 kg is only launched on site CCAFS SLC 40 and KSC LC 39A.

Success Rate vs. Orbit Type

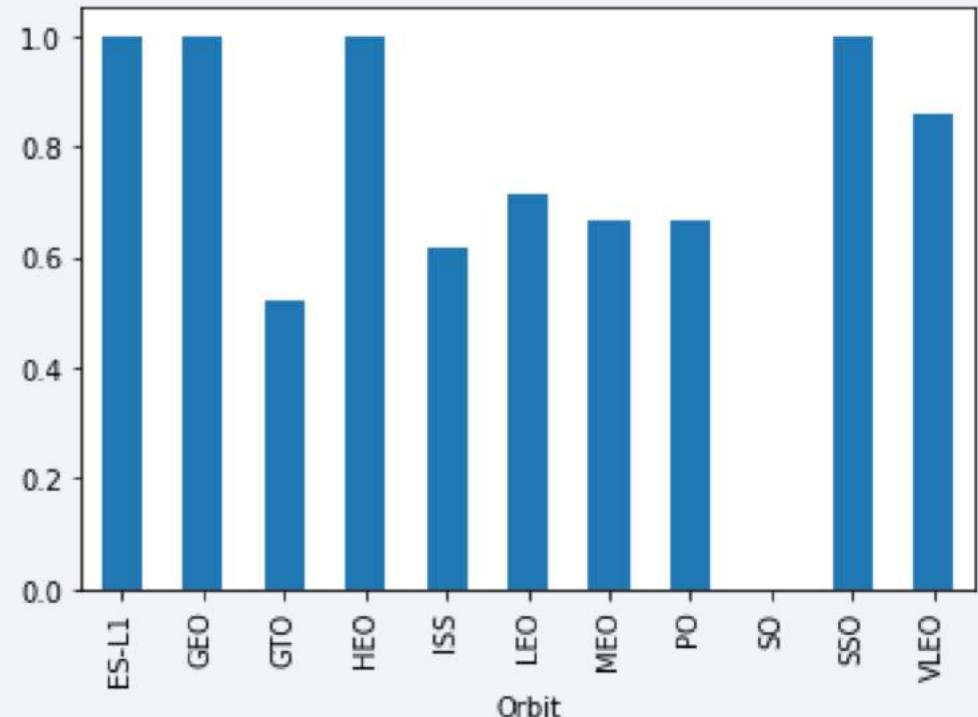
1. Orbits with the highest success rate (~100%) includes:

- ES-L1
- GEO
- HEO
- SSO

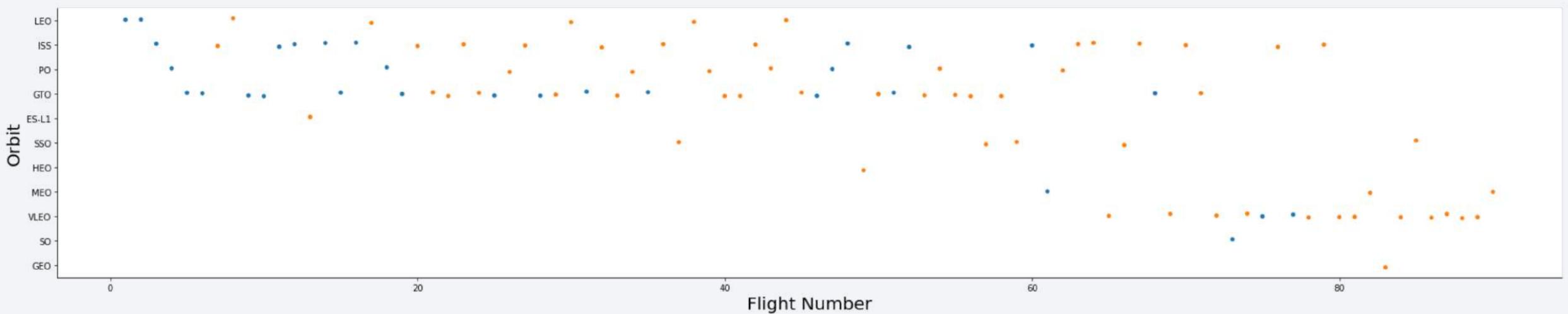
2. This is then followed by:

- VLEO (~80%)
- LEO (~70%)
- MEO & PO (~65%)
- ISS (~60%)
- GTO (~55%)

3. SO is the only orbit with 0% success rate.

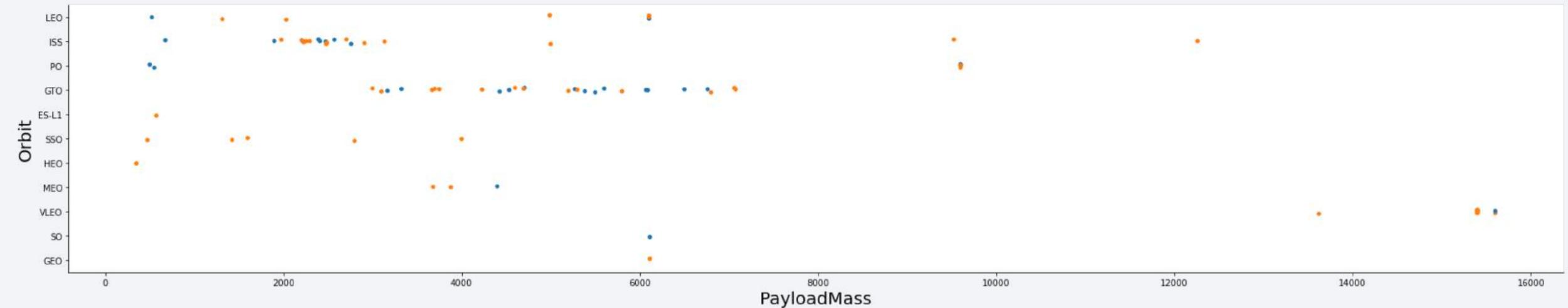


Flight Number vs. Orbit Type



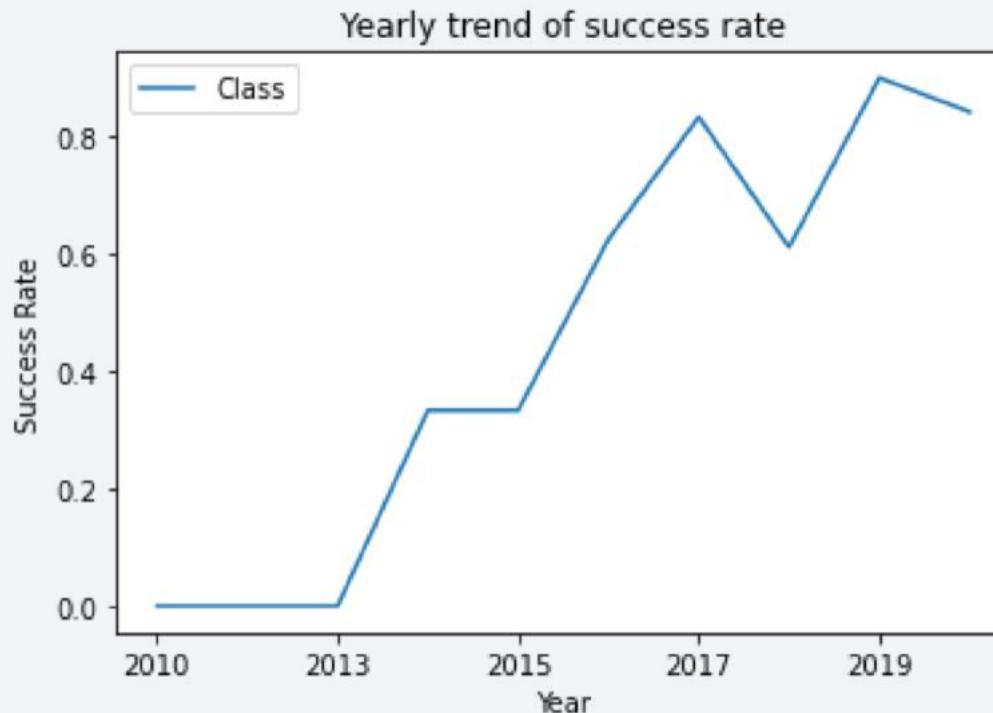
1. LEO Orbit's success rate is related to the number of flight i.e. the higher the flight number, the higher the success rate.
2. There seems to be no relationship between fight number and success rate for orbit GTO.
3. VLEO, SO and GEO are the orbits that were explored most recently, whereby SO has zero success rate (only 1 single launch) while VLEO has high success rate of 86%.
4. ISS and GTO are the orbits that were mostly explored with SpaceX Falcon 9 rocket launch.

Payload vs. Orbit Type



1. With heavy payloads, the successful landing are more for Polar, LEO and ISS orbits.
2. For GTO orbit, we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
3. ISS orbit has the widest range of payload launched.

Launch Success Yearly Trend



1. The mission success rate generally increases over the year, with highest success rate attained in year 2019.
2. In the first 3 years (2010 to 2013), the success rate is zero and this could be attributed to the fact that the team is experimenting, making adjustment and improve the technologies of the rockets.

All Launch Site Names

```
[8]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

```
Done.
```

```
[8]: Launch_Site
```

```
-----  
CCAFS LC-40
```

```
VAFB SLC-4E
```

```
KSC LC-39A
```

```
CCAFS SLC-40
```

1. As displayed in the query results above, there are 4 distinct launch site.
2. They are retrieved by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

```
[11]: %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
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

1. The 5 record with launch site started with “CCA” is displayed in the query results above.

2. They are retrieved by selecting the entries using the ‘WHERE’ and ‘LIKE’ operator to get the record that starts with ‘CCA’ and ‘LIMIT’ operator to limit the result to 5 entries.

Total Payload Mass

```
[13]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer='NASA (CRS)';

* sqlite:///my_data1.db
Done.

[13]: SUM(PAYLOAD_MASS__KG_)

45596
```

1. The total payload mass is 45,596 kg.
2. The value is retrieved using the SQL aggregate function i.e. ‘SUM’.

Average Payload Mass by F9 v1.1

```
[10]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1';  
* sqlite:///my_data1.db  
Done.  
[10]: AVG(PAYLOAD_MASS__KG_)  
_____  
2928.4
```

1. The average payload mass is 2,928.4 kg.
2. The value is retrieved using the SQL aggregate function i.e. 'AVG'.

First Successful Ground Landing Date

```
[41]: %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE `Landing _Outcome` = 'Success (ground pad)';  
* sqlite:///my_data1.db  
Done.  
[41]: MIN(DATE)  
01-05-2017
```

1. The first successful landing date is 1st May 2017.
2. The value is retrieved using the SQL aggregate function i.e. ‘MIN’ with ‘WHERE’ clause to specify the successful landing outcome

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[7]: %sql SELECT Booster_version FROM SPACEXTBL WHERE `Landing _Outcome`='Success (drone ship)' and (PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000)
* sqlite:///my_data1.db
Done.
[7]: Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2
```

1. There are 4 successful landing on drone ship with payload between 4,000 kg and 6,000 kg.
2. The value is retrieved using the SQL ‘WHERE’ clause to specify the successful landing outcome as well as the range of the payload.

Total Number of Successful and Failure Mission Outcomes

```
[9]: %sql SELECT Mission_Outcome ,COUNT(Mission_Outcome ) FROM SPACEXTBL GROUP BY Mission_Outcome;  
* sqlite:///my_data1.db  
Done.
```

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

1. The total number of successful mission is 100, and the total number of failure mission is 1.
2. The value is retrieved using the SQL aggregate function i.e. ‘COUNT’ with ‘GROUP BY’ clause to group the mission outcome.

Boosters Carried Maximum Payload

```
[47]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

* sqlite:///my_data1.db
Done.

[47]: 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
```

1. There are total 12 booster who carried the maximum payload.
2. The value is retrieved using subquery with SQL aggregate function i.e. ‘MAX’ applied within the subquery.

2015 Launch Records

```
[51]: %sql SELECT substr(Date,4,2) AS Month, `Landing _Outcome`, Booster_Version, Launch_Site FROM SPACEXTBL WHERE
```

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

```
Done.
```

```
[51]: 

| 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 |


```

1. There are 2 failure mission on drone ship in year 2015 i.e. in January and April, both happened on launch site CCAFS LC-40.
2. The value is retrieved using function i.e. ‘substr’ to retrieve the month and year from the ‘Date’ column.

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

Landing _Outcome	Counts
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

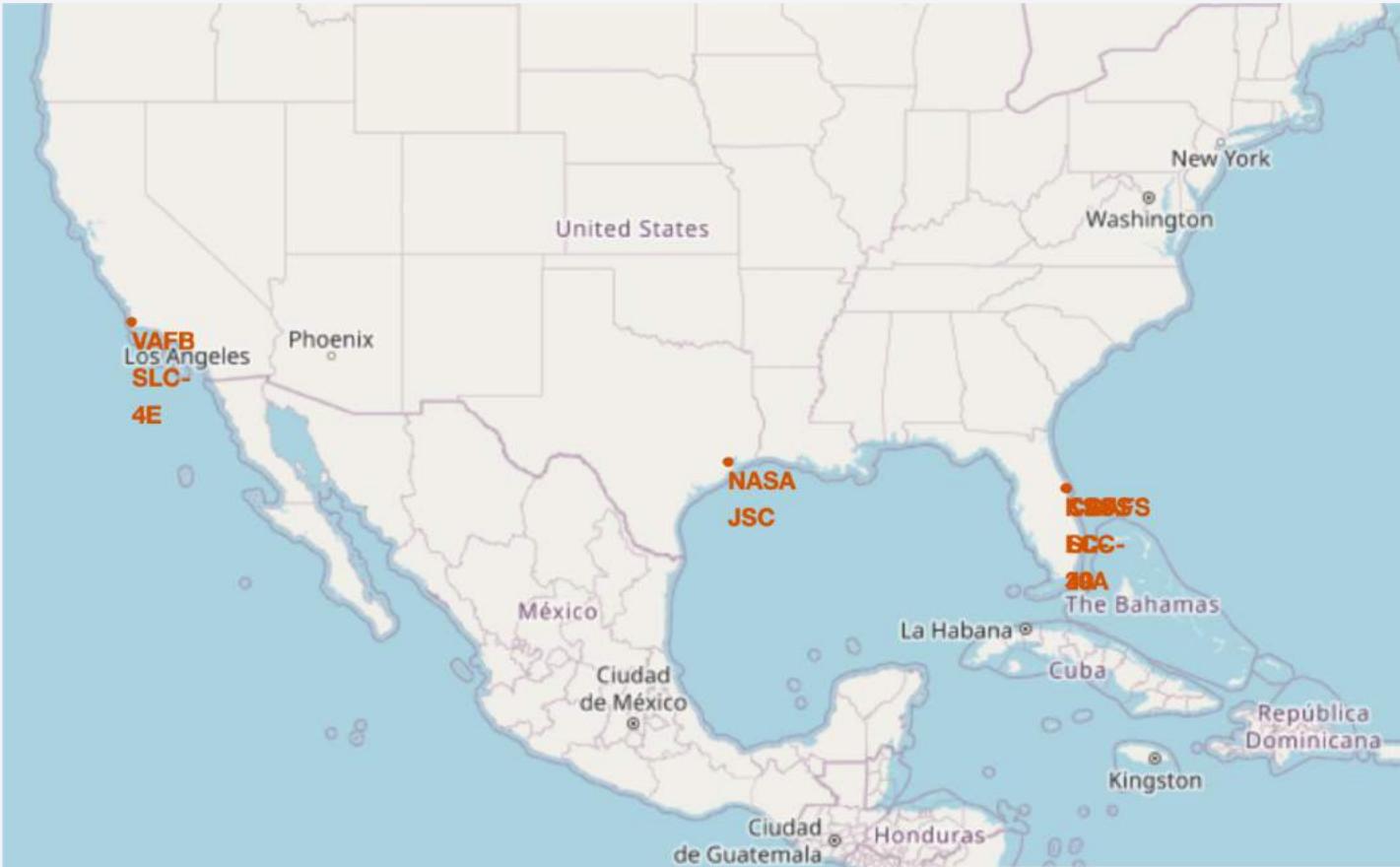
1. The ranking of the launch outcome between 2010-06-04 and 2017-03-20 are listed above.
2. The value is retrieved using the SQL aggregate function i.e. ‘COUNT’ with ‘WHERE’ clause to specify date range and ‘ORDER BY’ clause to rank the result in from highest to lowest.

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against a dark blue-black void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in coastal and urban areas. A thin, glowing horizontal band represents the atmosphere or the horizon where the Earth meets the sky.

Section 3

Launch Sites Proximities Analysis

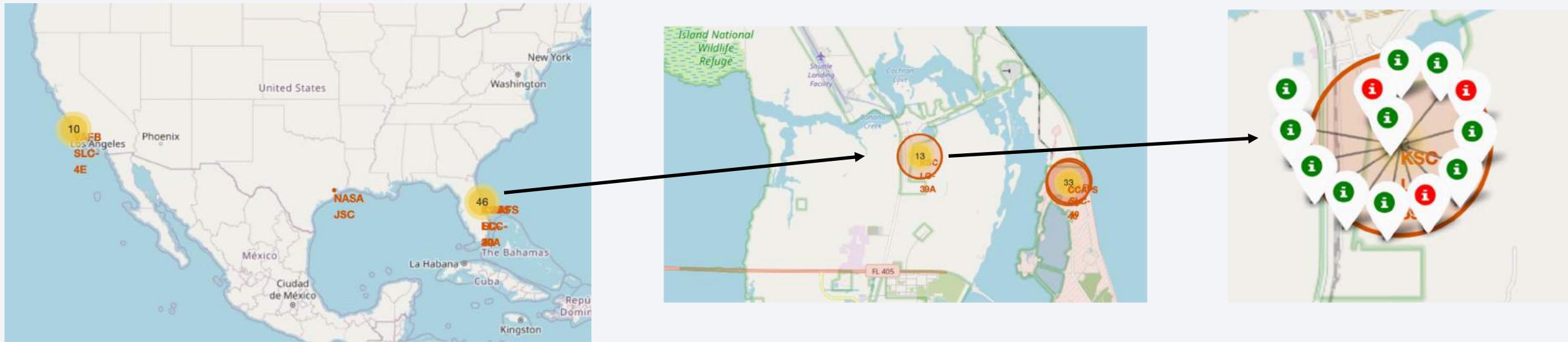
Folium Map – All Launch Sites



1. From the map, it is observed that all the launch sites are close to the sea which may be attributed to safety reason, yet not too far off from road and railroad.

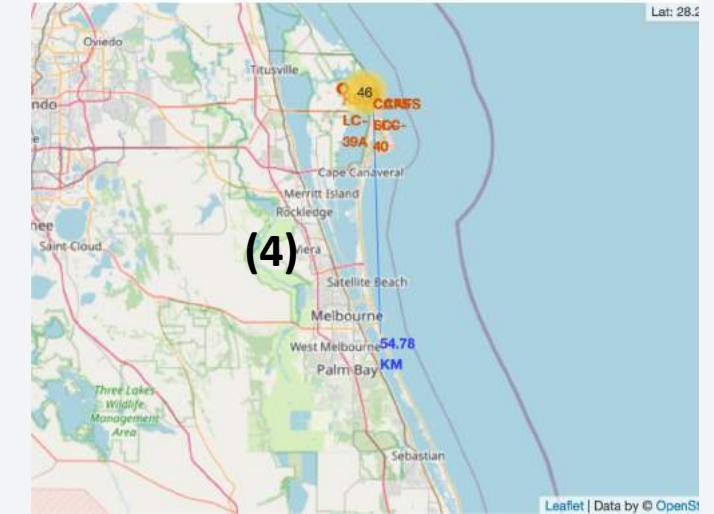
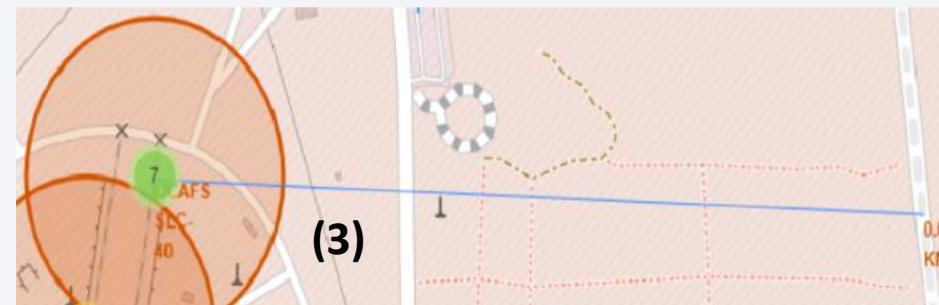
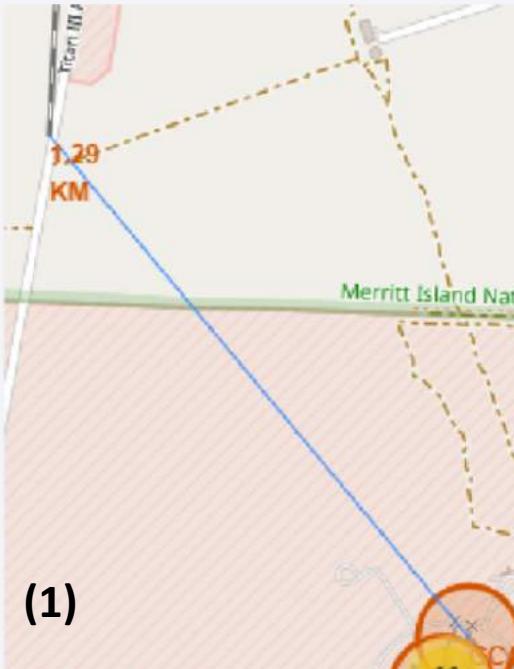
Folium Map – Launch Outcome by Site

Example for launch site KSC LC-39A launch outcomes.



1. Green marker indicates successful launch, and red marker indicates failed launch.

Folium Map – Safety and Logistic

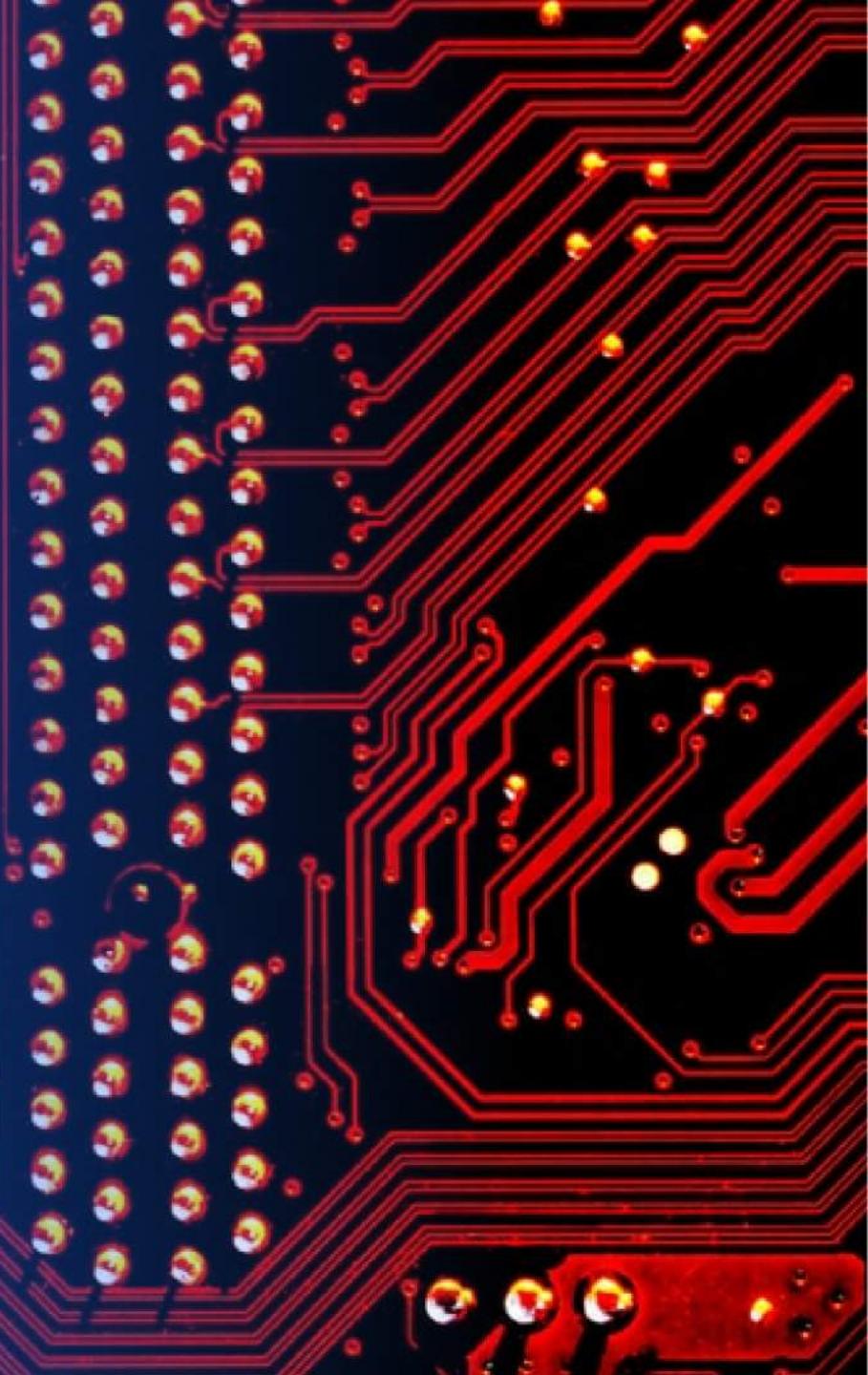


Note: (1) Railway, (2) Sea (3) Highway and (4) Nearest city

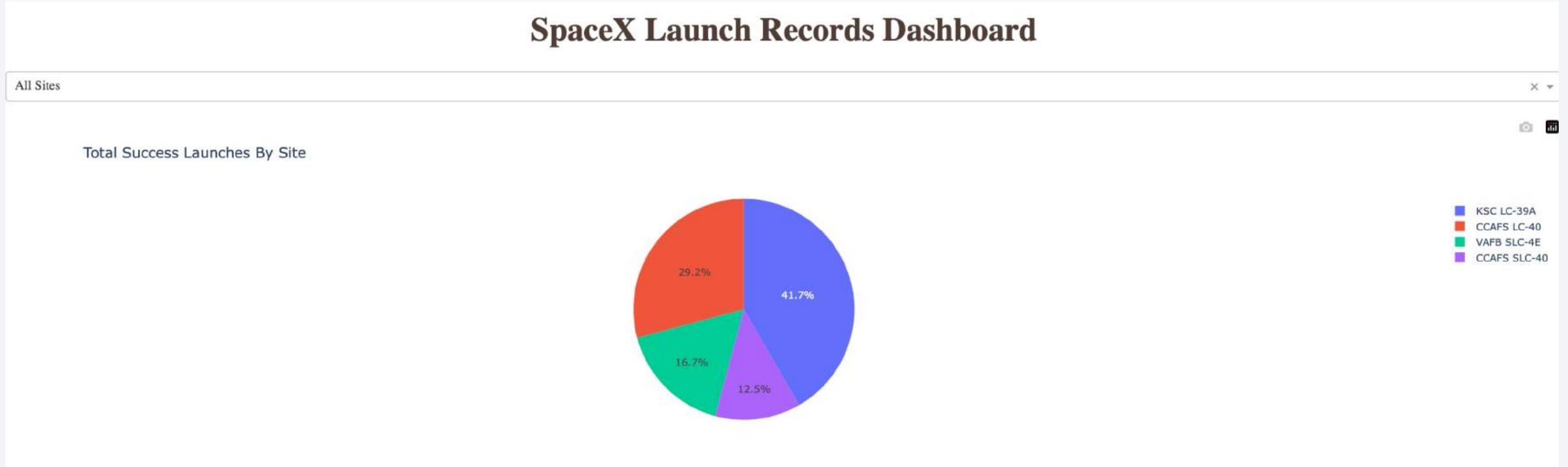
1. Looking at launch site CCAFS SLC-40, it is located in close proximity to railway, highway and sea, but far away from city.

Section 4

Build a Dashboard with Plotly Dash

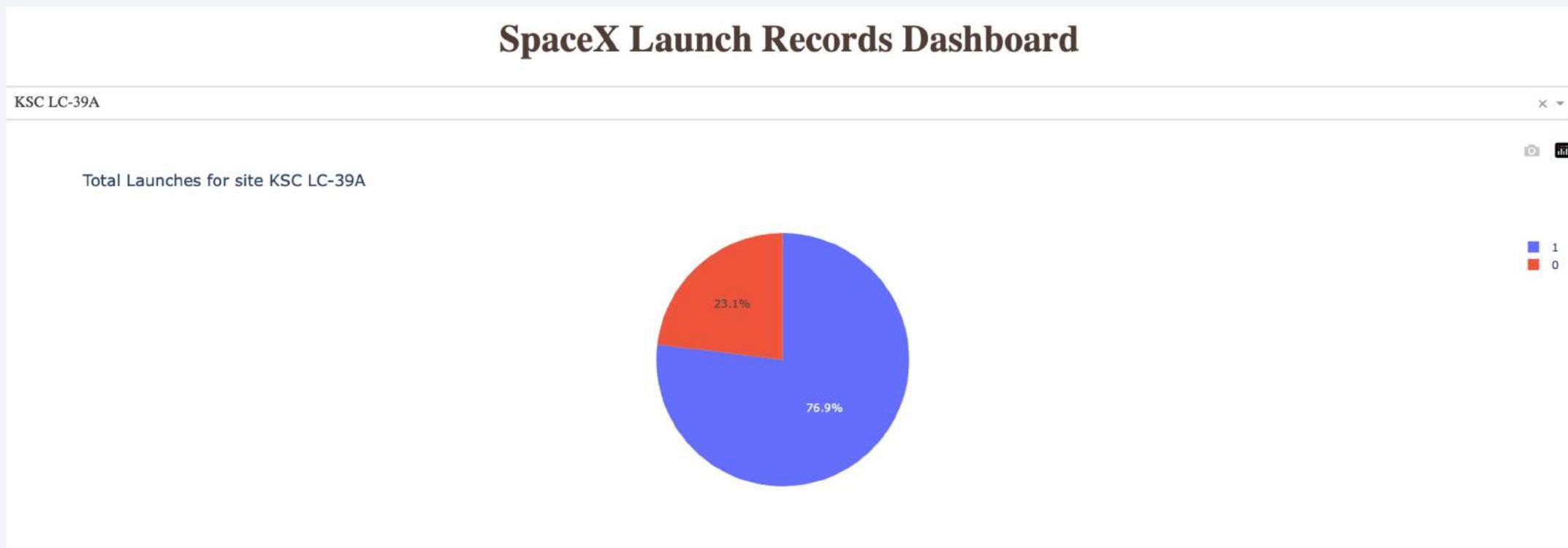


Dashboard - Successful Launches by Site



1. The launch location seems to correlate to the launch outcome.
2. Launch site **KSC LC-39A** has the highest launch success rate (41.7%), followed by CCAFS LC-40 (29.2%), VAFB SLC-4E (16.7%) and CCAFS SLC-40 (12.5%). The percentage value are proportion of each site with success launch relative to all the successful launch.

Dashboard - Launch Success Rate for KSC LC-39A



1. The launch success rate for KSC LC-39A is **76.9%**.

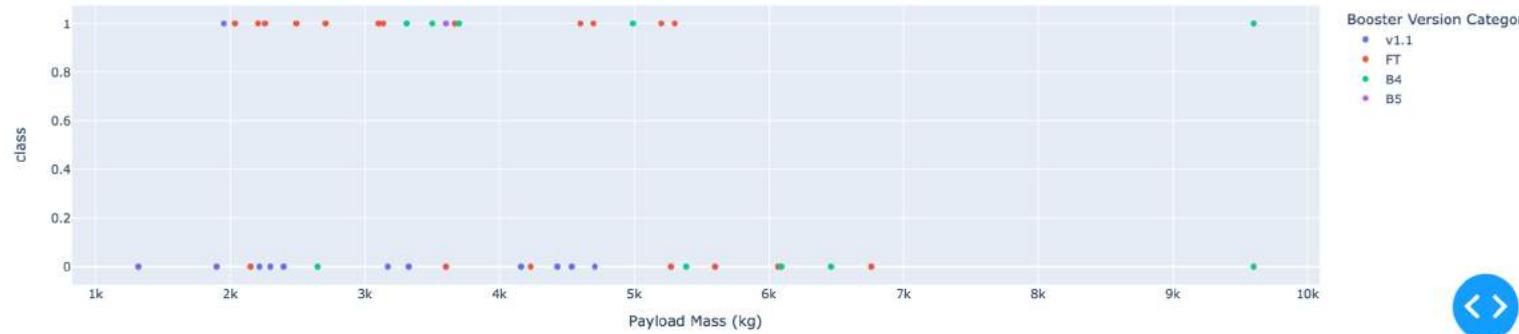
Dashboard - Payload vs Launch Outcome

Payload range (Kg):

0 100

All sites - payload mass between 1,000kg and 10,000kg

(1) Payload between 1,000 kg and 10,000 kg

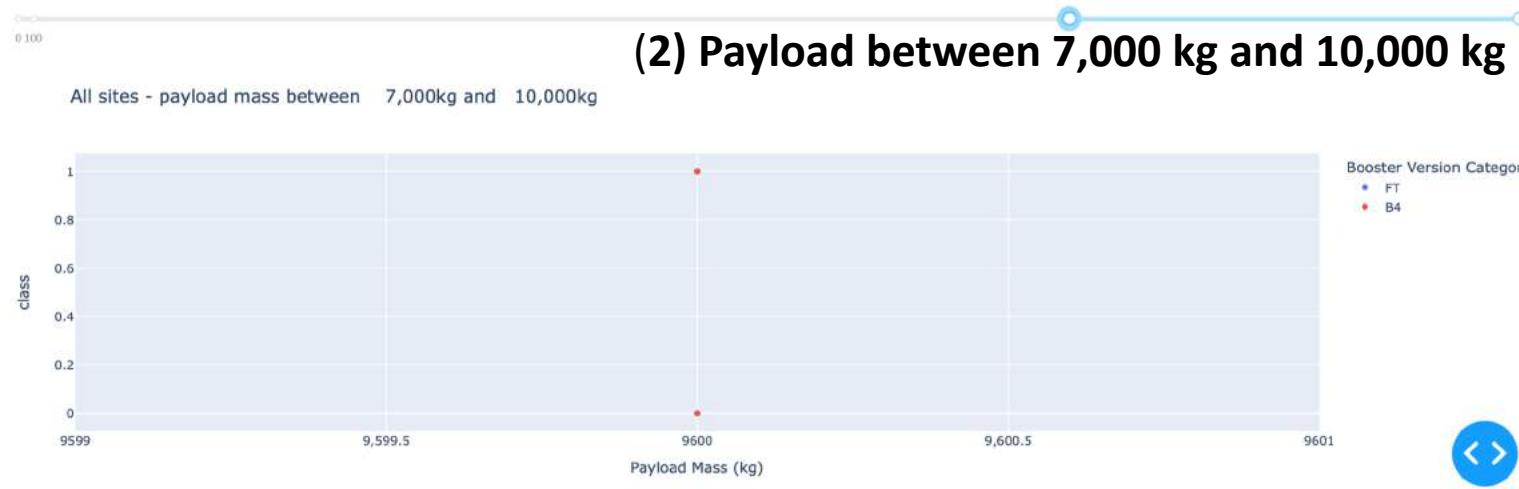


Payload range (Kg):

0 100

All sites - payload mass between 7,000kg and 10,000kg

(2) Payload between 7,000 kg and 10,000 kg



1). At low payload weight of below 6,000kg, FT booster rocket has higher success rate.

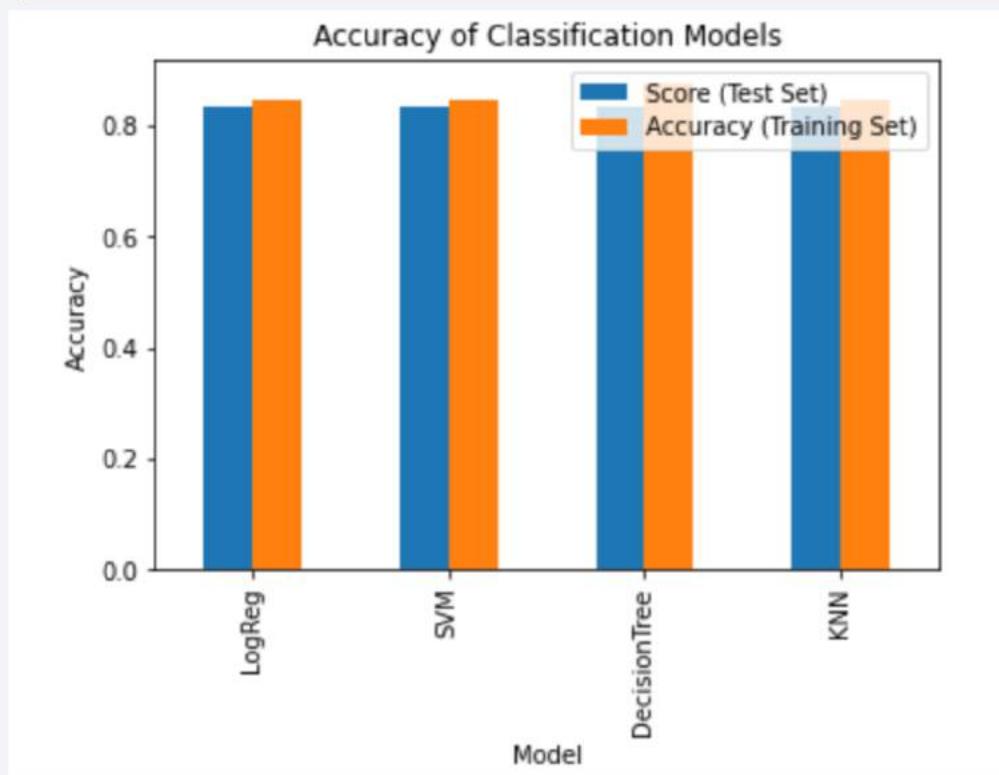
2). For payload weight greater than 7,000 kg, currently there is not enough data to estimate the success rate.

The background of the slide features a dynamic, abstract design. It consists of several thick, curved lines that transition in color from blue on the left to yellow on the right. These lines create a sense of motion and depth, resembling a tunnel or a stylized road. The overall effect is modern and professional.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



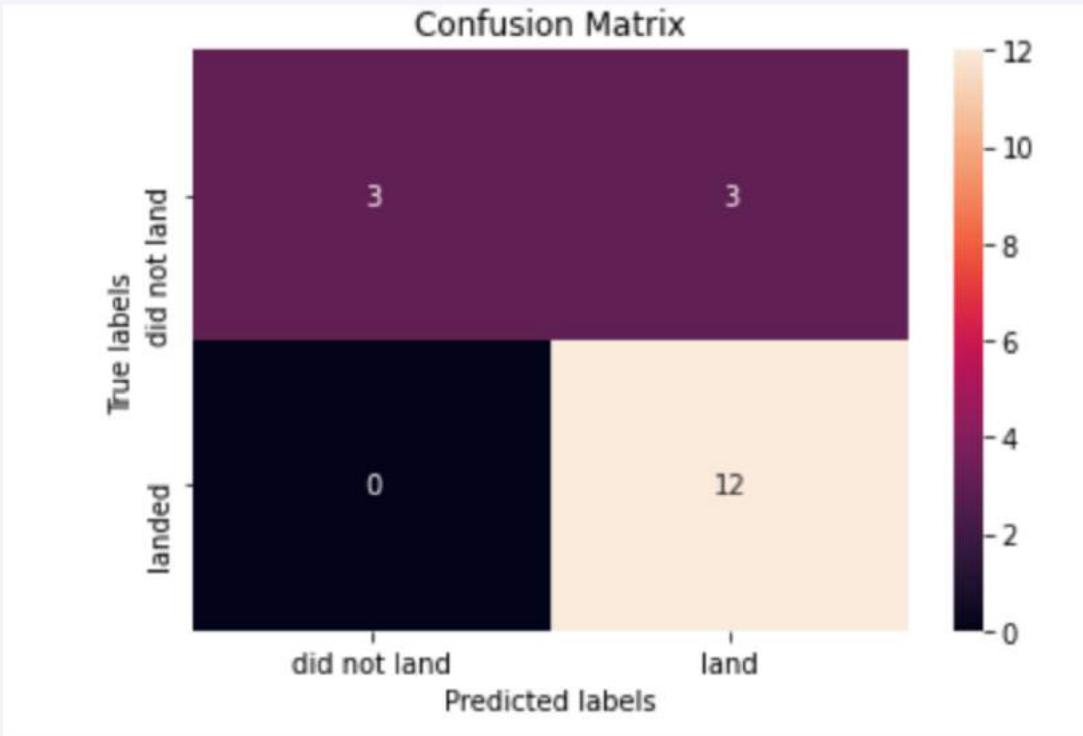
	Score (Test Set)	Accuracy (Training Set)
LogReg	0.833333	0.847222
SVM	0.833333	0.847222
DecisionTree	0.833333	0.875000
KNN	0.833333	0.847222

1). 4 classification models were evaluated on the data set and Decision Tree classifier has the highest accuracy (87.5%) on the training set.

2) When evaluated on test data, all models show equal accuracy of 83.33%. More test data is needed in order to better evaluate the accuracy of each model.

Confusion Matrix

The confusion matrix of Decision Tree Classifier (highest accuracy) is shown as below:



- 1). While Decision Tree Classifier exhibit the highest accuracy on training set, it has limitation of giving false positive result i.e. there are 3 cases where the model predicted 'Successful Landing' but in fact the actual outcome is failed landing.
- 2). The confusion matrix is the same for all 4 models.

Conclusions

- The success rate of SpaceX Falcon 9 stage 1 landing is dependent on several features including launch site, booster type, orbit and payload mass.
- From the EDA, it can be inferred that:
 - a) The orbits with the highest success rate are: GEO, HEO, SSO & ES-L 1.
 - b) The launch site with the highest success rate is KSC LC 39-A.
 - c) A combination of payload which weight less than 6,000 kg and FT booster rocket can achieve a higher mission success rate.
 - d) Depending on the orbit of flight, the payload mass has crucial impact on the success rate. A heavier payload mass yield higher success rate on certain orbits.
 - e) Generally the mission success rate improves over the year as the rocket launch technologies progress
- Albeit Decision Tree Classifier is selected as it has the highest accuracy on training set relative to the other models, they all share the same accuracy on the test set and hence more data is needed to better evaluate these models to determine the best predictive model.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!

