IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sopagnaphea Kim
Date: 28 May, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis (EDA) using Visualization and SQL
  - Interactive Visual Analytics using Folium and Plotly Dash
  - Predictive Analysis using Classification Models
- Summary of all results
  - Data Analysis Results
  - Data Visualization & Interactive Dashboards
  - Predictive Model Analysis Result

Introduction

- Project Background:
  - The commercial space age is here, companies are making space travel affordable for everyone and among them ‘Space X’ is the most successful so far. ‘Space X’ is relatively inexpensive due to reusing first stage of launching. As a result, a new rocket company ‘Space Y’ has been created to compete with ‘Space X’.

- Problem Statement:
  - Determine if ‘Space X’ will reuse their first stage.
  - Determine the price of each launch from ‘Space X’.

Section 1

Methodology

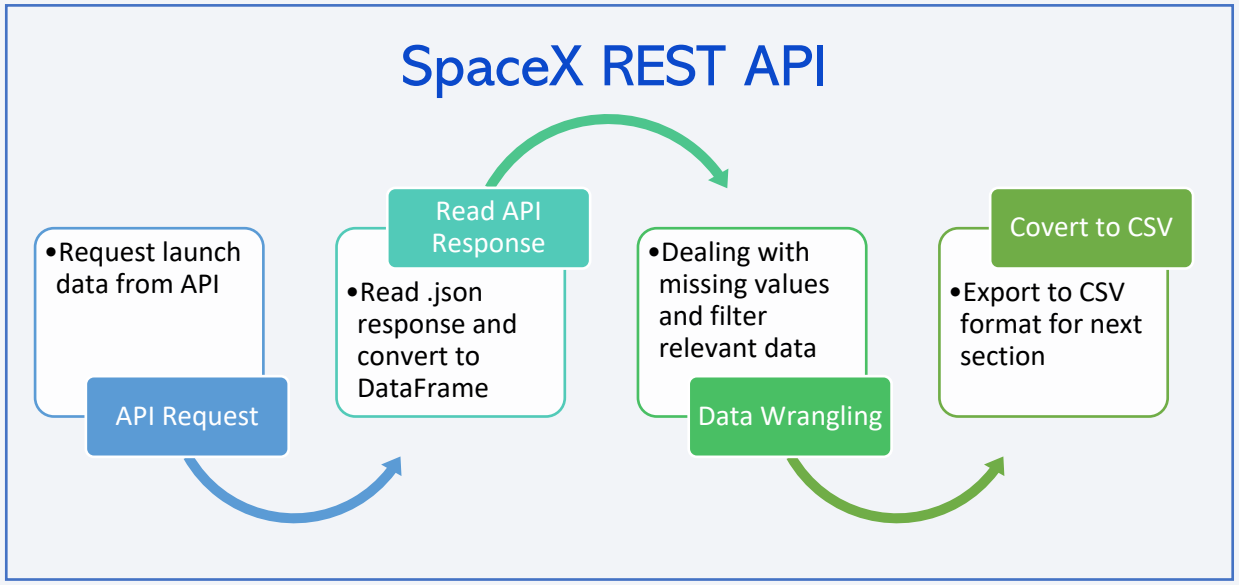
Methodology

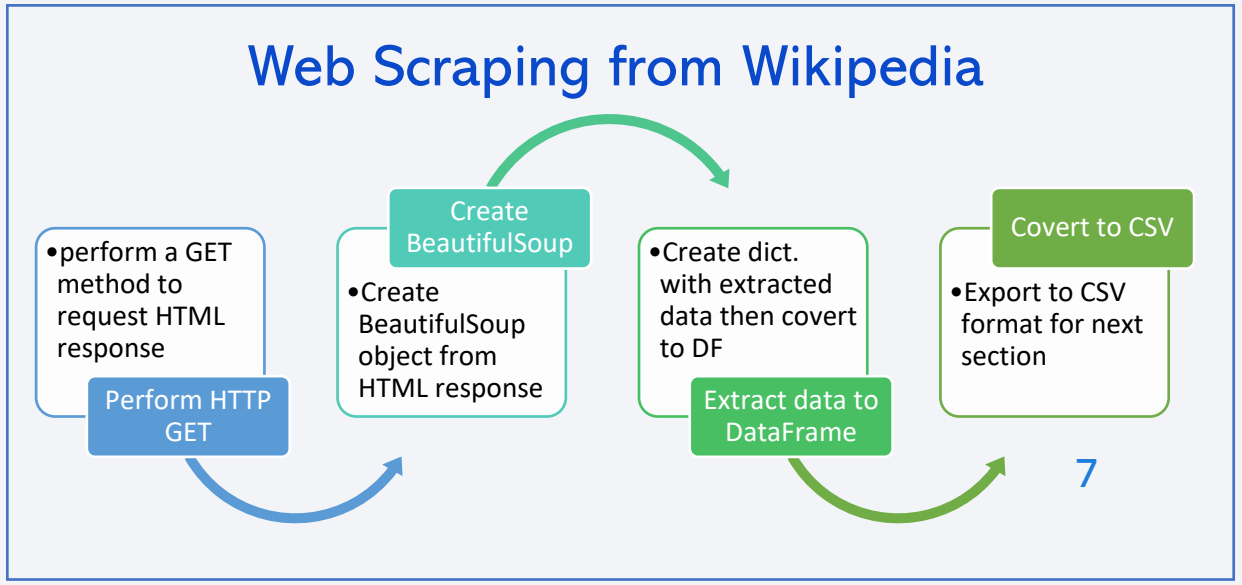
Executive Summary

- Data collection methodology:
  - Data was collected through: SpaceX API & web scraping Falcon 9 and Falcon Heavy Launches records from Wikipedia.
- Perform data wrangling
  - Convert the 'Landing Outcome' into new column 'Class' (1 = Success, 0 = Failure)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find the best performing method (among Logistic Regression, SVM, Classification Tree & KNN) using test data

Data Collection

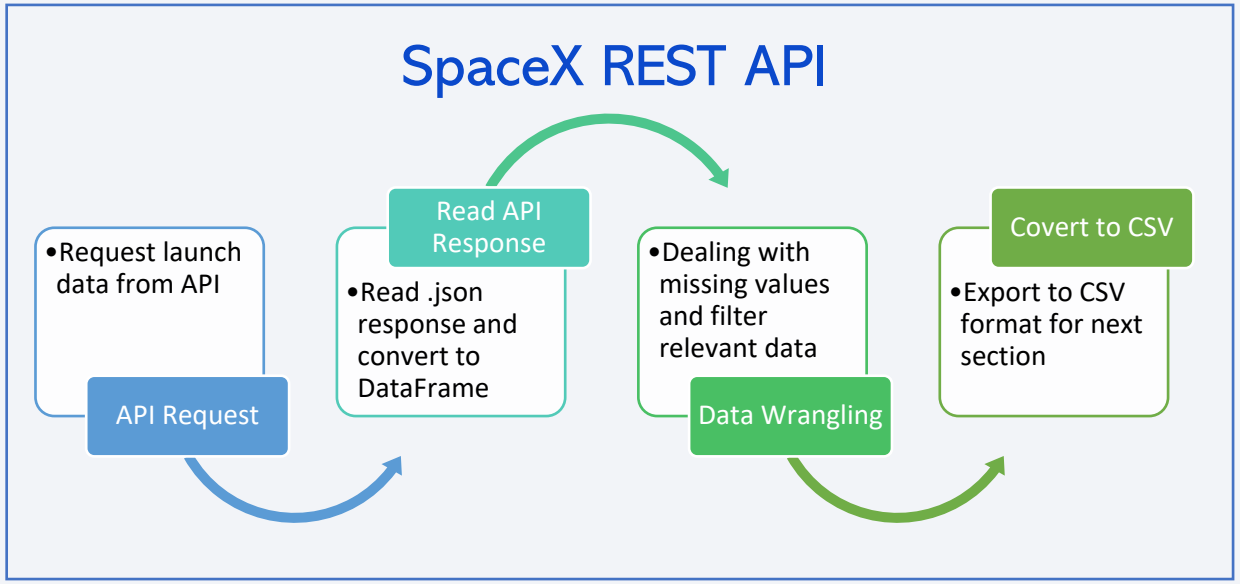
- Our Data have been collected through 2 methods:
  - SpaceX REST API: will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome. This data will be used to predict whether SpaceX will attempt to land a rocket or not.
  - Web scraping Falcon 9 Launch data from related Wiki pages: scrape some HTML tables that contain valuable Falcon 9 launch records.





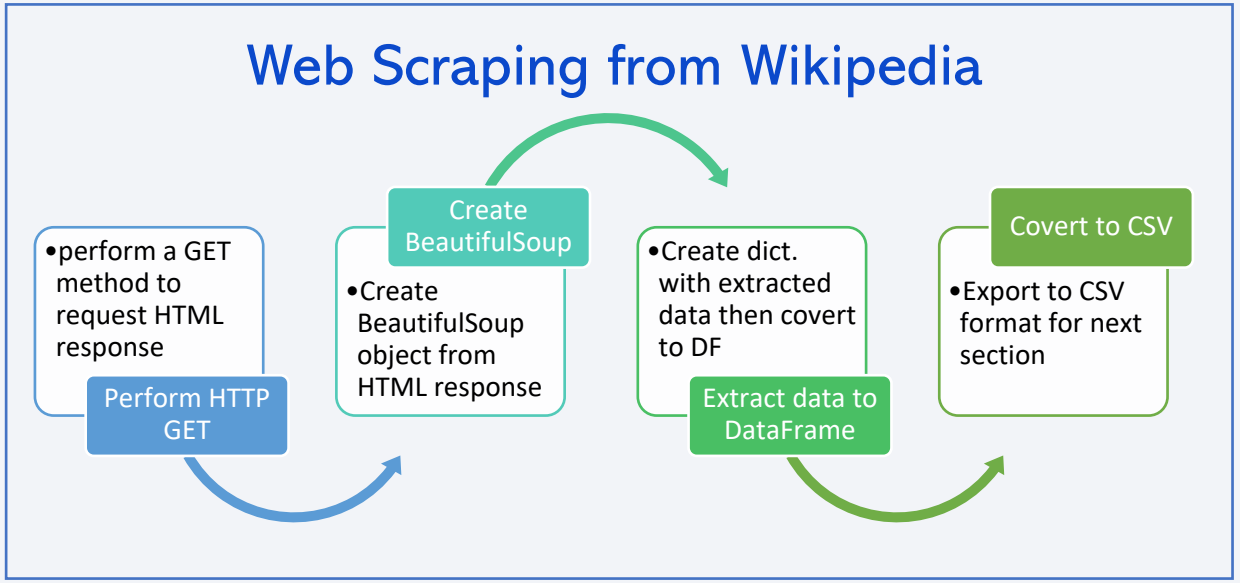
Data Collection – SpaceX API

- Steps of Using SpaceX API:
  1. API Request: requesting rocket launch data from SpaceX API.
  2. Read API Response: decode the response content and turn it into dataframe using .json_normalize().
  3. Data Wrangling: remove the Falcon 1 launches & replace nan values with mean.
  4. Convert to CSV: export file to a CSV for the next section
- GitHub URL of SpaceX API: Link



Data Collection - Scraping

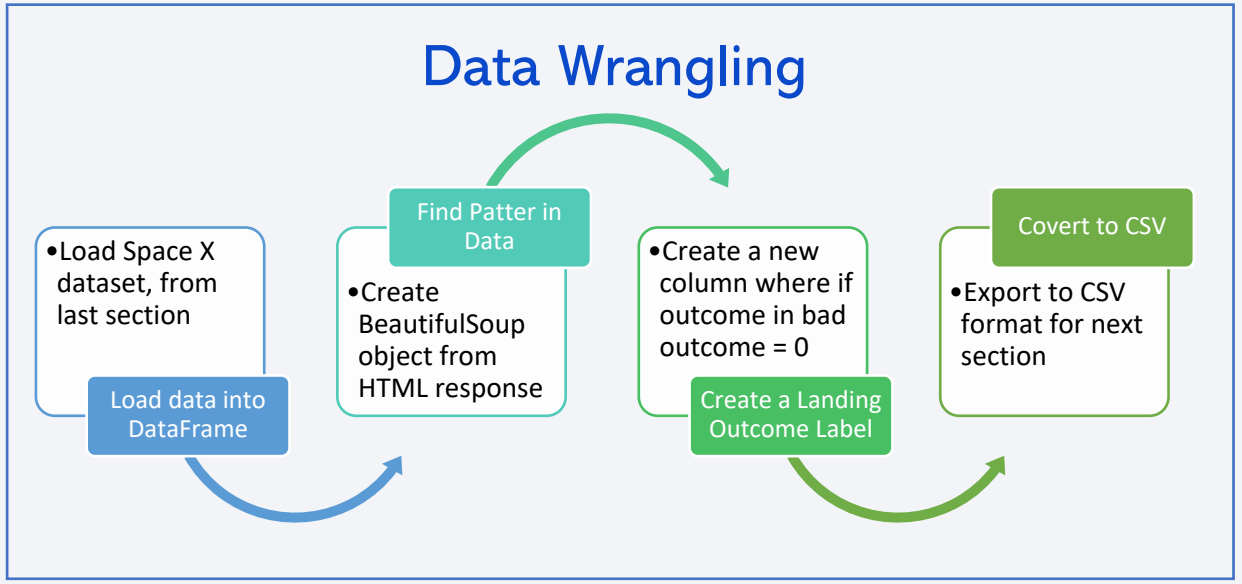
- Step of Using Web Scraping:
    1. Perform HTTP GET: perform an HTTP GET method to request the Falcon9 Launch HTML page.
    2. Create BeautifulSoup Object: Create a BS object from the HTML response.
    3. Extract data to DataFrame: collect all relevant column names from the HTML table header/values and convert into DataFrame.
    4. Convert to CSV: export file to a CSV for the next section



- GitHub URL of web scraping: Link

Data Wrangling

- Perform Exploratory Data Analysis (EDA) and determine Training Labels.
- Steps of Using EDA:
  1. Load Data:
  2. Find Pattern:
     - **Calculate the number of launches on each site**
     - **Calculate the number of launches**
     - **Calculate the number and occurrence of each orbit**
  3. Create a 'Landing Outcome' Label: create a list where the element is zero if the corresponding row in Outcome is in the set bad_outcome.
  4. Convert to CSV: export file to a CSV for the next section
- GitHub URL: Link



EDA with Data Visualization

- Following charts have been utilized to gain further insights of data:
  - Scatter Plot: used to visualize:
    - Relationship between **PayloadMass** vs. **FlightNumber**
    - Relationship between **LaunchSite** vs **FlightNumber**
    - Relationship between **LaunchSite** vs **PayloadMass**
    - Relationship between **OrbitType** vs **FlightNumber**
    - Relationship between **OrbitType** vs **PayloadMass**
  - Bar Chart: used to visualize:
    - Relationship between Success Rate of Each Orbit Type
  - Line Chart: used to visualize:
    - Launch Success Yearly Trend
- GitHub URL of data visualization: Link

EDA with SQL

- To further gain insight of data, following SQL queries have been performed:
  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 average payload mass carried by booster version F9 v1.1
  5. List the date when the first succesful landing outcome in ground pad was acheived.
  6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  7. List the total number of successful and failure mission outcomes.
  8. List the names of the booster_versions which have carried the maximum payload mass.
  9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  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.

- GitHub URL of SQL: Link

Build an Interactive Map with Folium

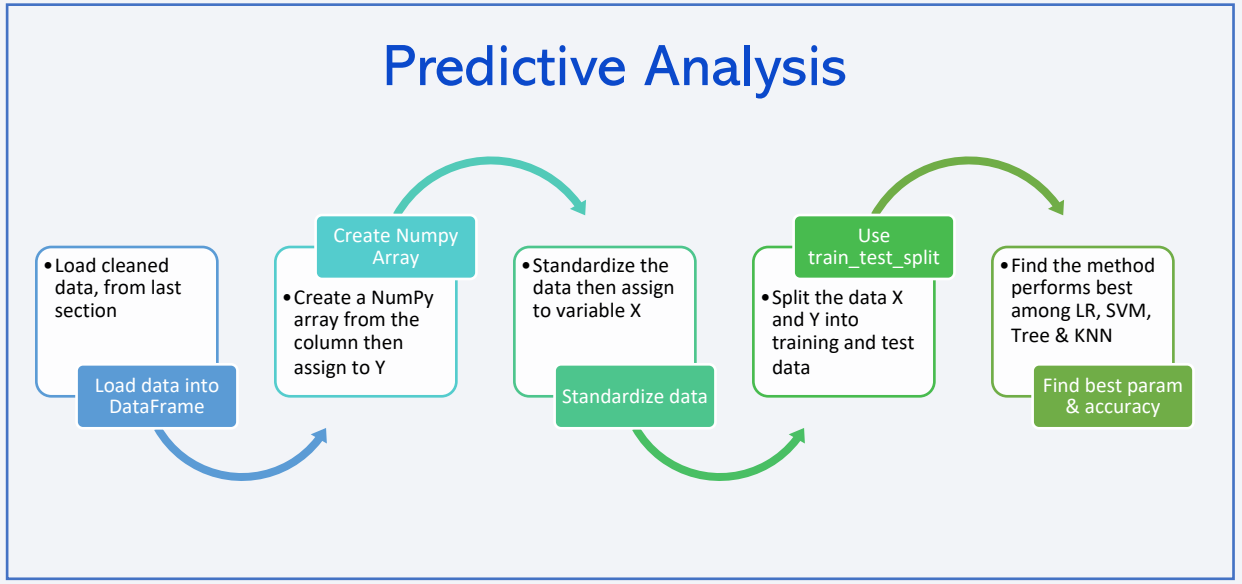
- Following are map objects that have been created and added to a folium map:
  - folium.Circle ( ): use to add a highlighted circle area with a text label on a specific coordinate.
  - folium.Marker ( ): use to mark all launch sites on a map.
  - MarkerCluster ( ): use to simplify a map containing many markers having the same coordinate.
  - Calculation the distances between a launch site to its proximities, we utilize:
    - MousePosition ( ): use to get coordinate for a mouse over a point on the map.
    - folium.Marker ( ): mark down a point on the closest coastline using MousePosition and calculate the distance between the coastline point and the launch site.
    - PolyLine ( ): draw a PolyLine between a launch site to the selected coastline point.
- Questions that we can answer from this:
  - Are launch sites in close proximity to railways? **YES**
  - Are launch sites in close proximity to highways? **YES**
  - Are launch sites in close proximity to coastline? **YES**
  - Do launch sites keep certain distance away from cities? **YES**
- GitHub URL of Interactive Map: Link

Build a Dashboard with Plotly Dash

- Following are plots/graphs and interactions that have added to dashboard:
  - Drop-down: add a Launch Site Drop-down Input Component.
  - Pie Chart & Callback: add a callback function to render success-pie-chart based on selected site dropdown
  - Range Slider: add a Range Slider to Select Payload.
  - Scatter Plot & Callback: add a callback function to render the success-payload-scatter-chart scatter plot.
- Questions that we can answer from this:
  - Which site has the largest successful launches? **KSC LC-39A (41.7%)**
  - Which site has the highest launch success rate? **KSC LC-39A (76.9%)**
  - Which payload range(s) has the highest launch success rate? **2K – 4K Payload Mass (KG)**
  - Which payload range(s) has the lowest launch success rate? **6K – 8K Payload Mass (KG)**
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? **FT**
- GitHub URL of Plotly Dash: Link

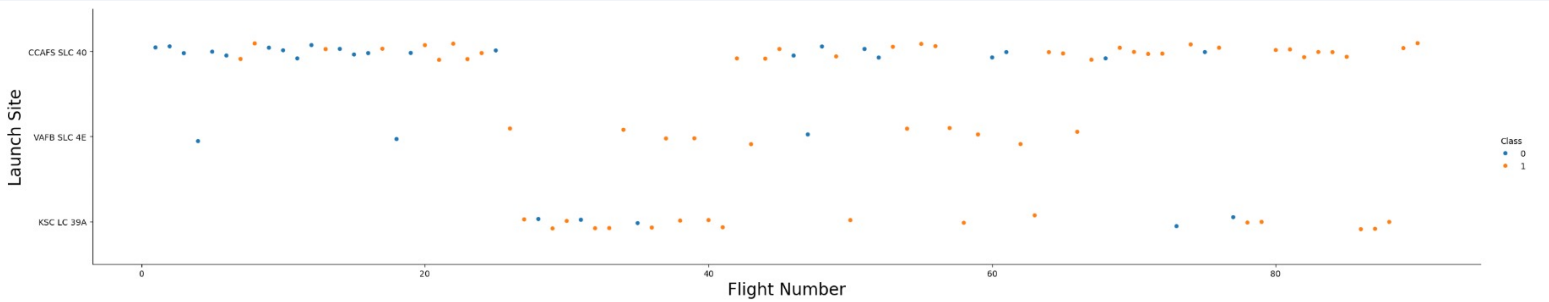
Predictive Analysis (Classification)

- Steps of Using Predictive Analysis:
    1. Load Data into DataFrame (DF).
    2. Create Numpy Array then assign to Variable Y.
    3. Standardize data then assign to Variable X.
    4. Use train_test_split to split data into training and testing data.
    5. Calculate 'Best Params' & Accuracy then compare accuracy among (Logistic Regression, SVM, Decision Tree, and KNN) to find best performing method.
- GitHub URL of predictive analysis: Link

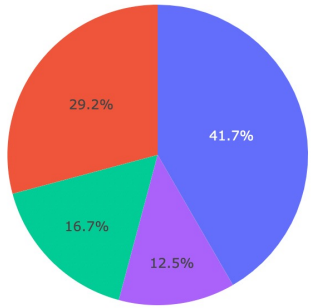


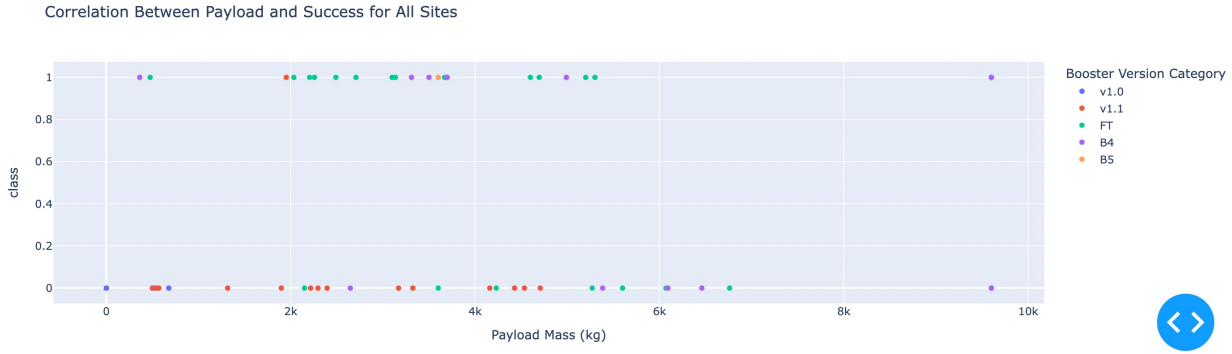
Results

- Exploratory data analysis results



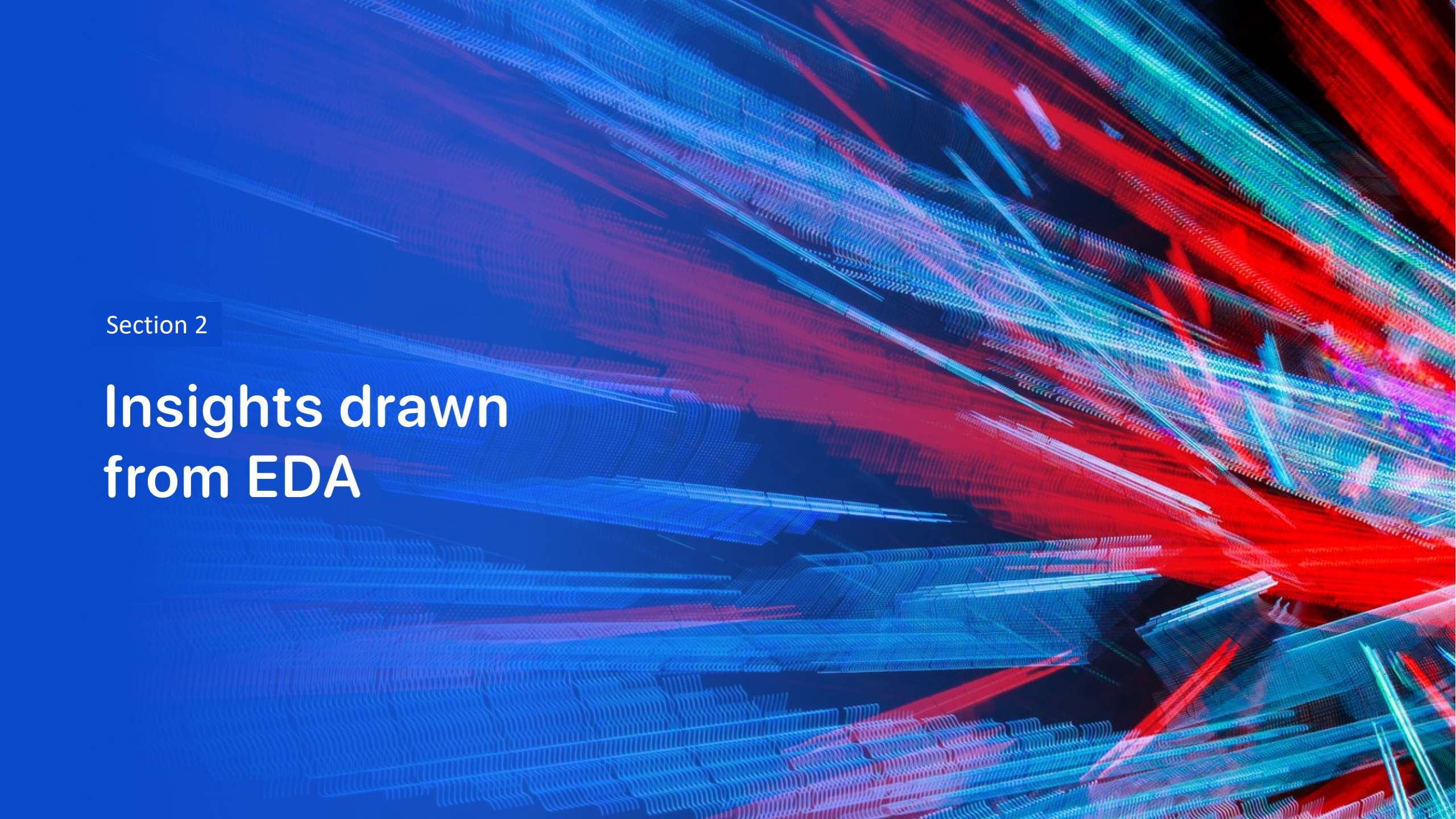
- Interactive analytics demo in screenshots



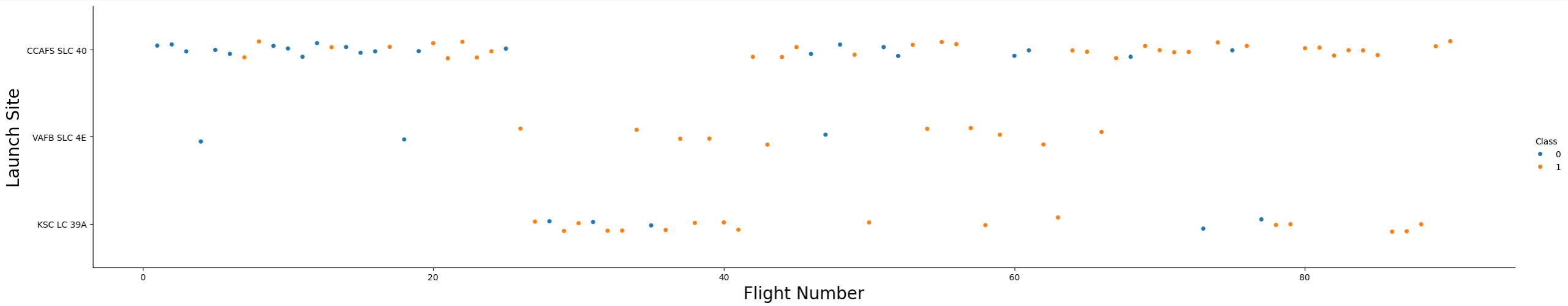


- Predictive analysis results

	ML Method	Accuracy Score (%)
0	SVM	83.333333
1	Log Reg	83.333333
2	KNN	83.333333
3	Tree	88.888889

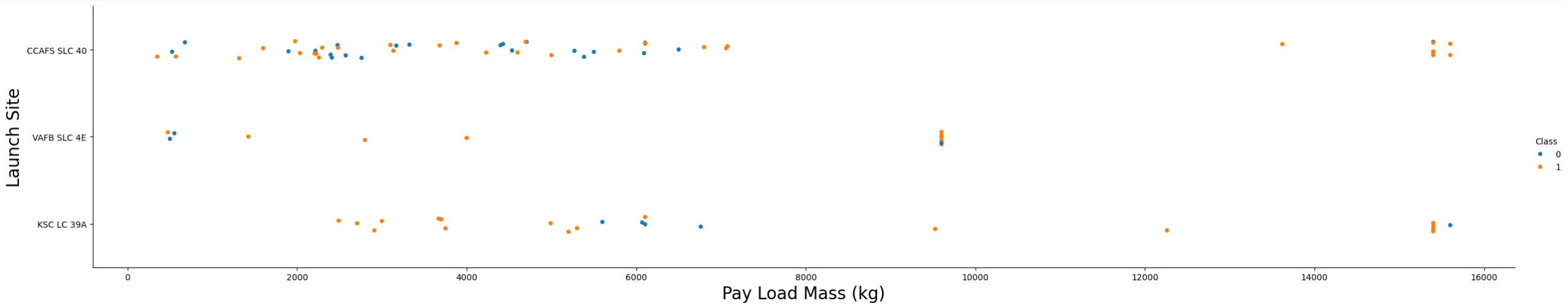

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- For launch site VAFB SLC, we can observe that it take around 25 Flights attempt to be successfully landed.
- For launch site KSC LC, we can observe that it takes around 25 Flights attempt to be successfully launched & landed.
- The more Flight attempts, the more Success Rate (Class = 1) increased.

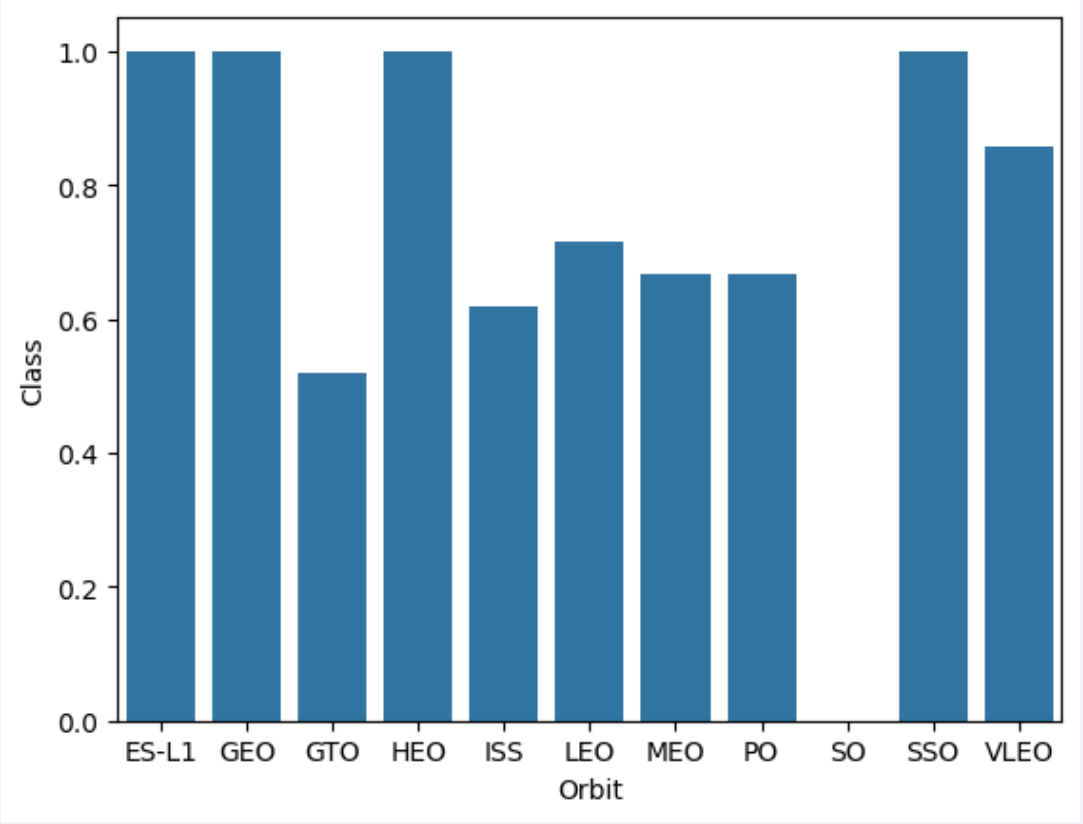
Payload vs. Launch Site



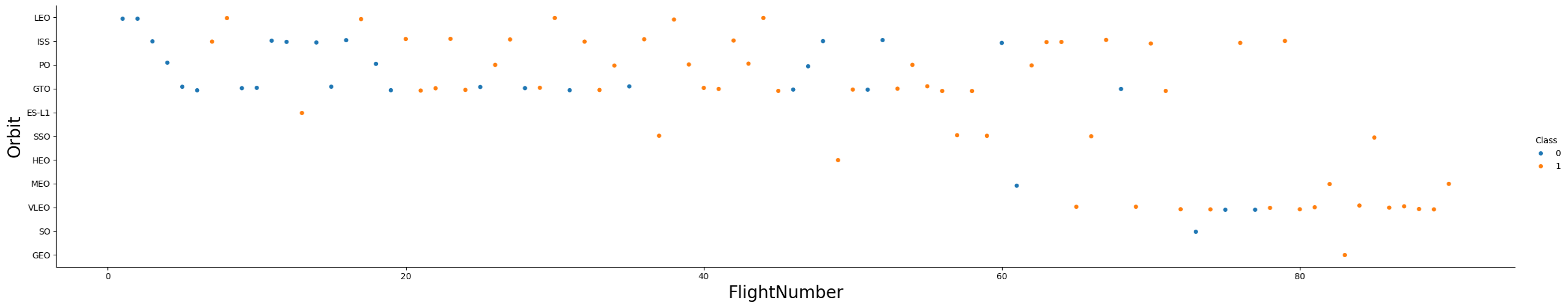
- For launch site VAFB-SLC, we can observe that there are no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type

- Orbit (ES-L1, GEO, HEO & SSO) has the highest Success Rate.
- There is 0% success rate for Orbit SO.

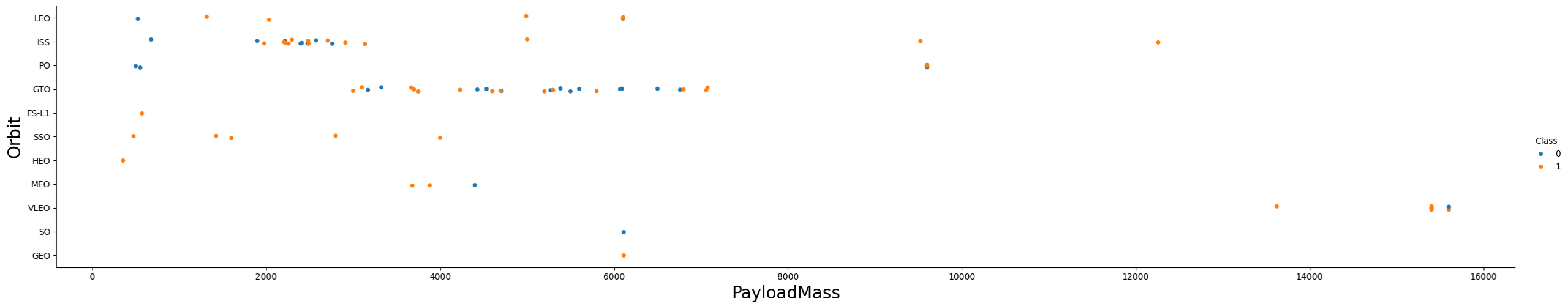


Flight Number vs. Orbit Type



- In the LEO orbit, the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.

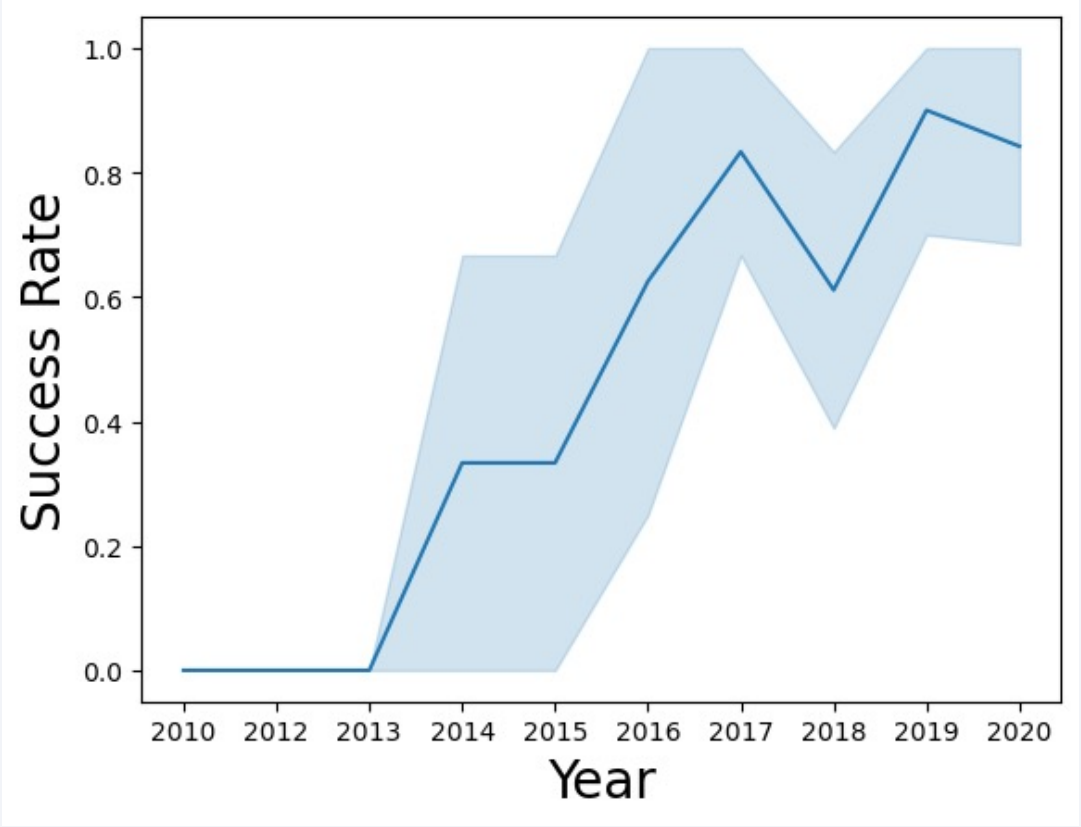
Payload vs. Orbit Type



- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO, we cannot distinguish this well as both positive landing rate and negative landing are both here and there.

Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

- Query:

```
%sql select Launch_Site, count(Launch_Site) from SPACEXTBL group by Launch_Site
```

- Description: Display the names of the unique launch sites in the space mission

- Result:

Launch_Site	count(Launch_Site)
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

Launch Site Names Begin with 'CCA'

- Query:

```
%sql select * from SPACEXTBL where Launch_Site LIKE "CCA%" limit 5
```

- Description: Display 5 records where launch sites begin with the string 'CCA'

- Result:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Query:

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer IN ("NASA (CRS)")
```

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

- Result:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Query:

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version IN ("F9 v1.1")
```

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

- Result:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Query:

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome IN ("Success (ground pad)")
```

- Description: List the date when the first successful landing outcome in ground pad was archived.

- Result:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome IN ("Success (drone ship)")
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_< 6000
```

- Description: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

- Result:

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

Total Number of Successful and Failure Mission Outcomes

- Query:

```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTBL group by Mission_Outcome
```

- Description: List the total number of successful and failure mission outcomes

- Result:

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

Boosters Carried Maximum Payload

- Query:

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ IN
(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

- Description: List the names of the booster_versions which have carried the maximum payload mass.

- Result:

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

2015 Launch Records

- Query:

```
%sql select substr(Date, 6,2) as Month_Names, Landing_Outcome, Booster_Version,
Launch_Site from SPACEXTBL where substr(Date,0,5)='2015' AND Landing_Outcome LIKE "Failure%"
```

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

- Result:

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

- Query:

```
%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTBL
where Date between ("2010-06-04") and ("2017-03-20") group by Landing_Outcome order by count(Landing_Outcome) desc
```

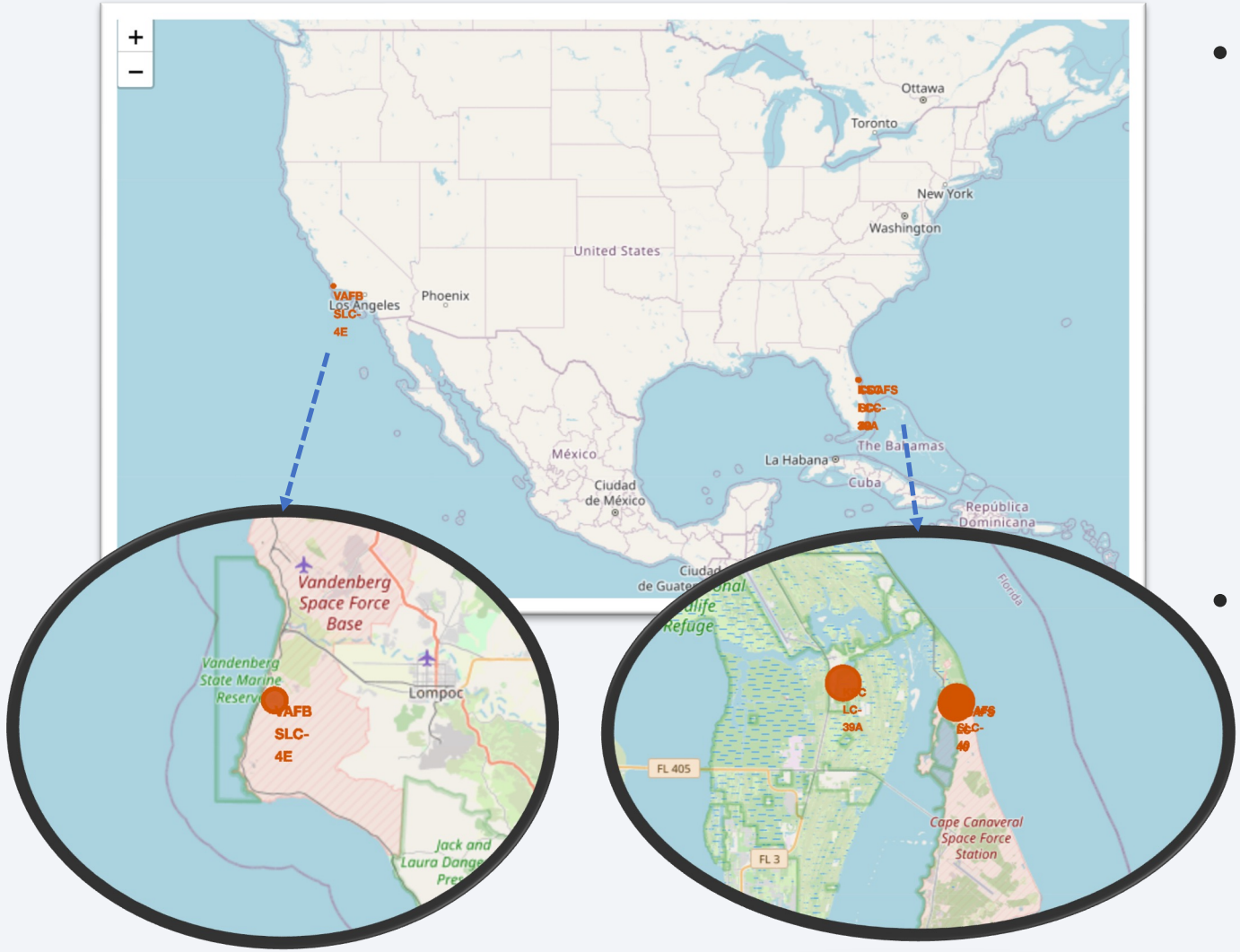
- Description: Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

- Result:

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

SpaceX Launch Sites

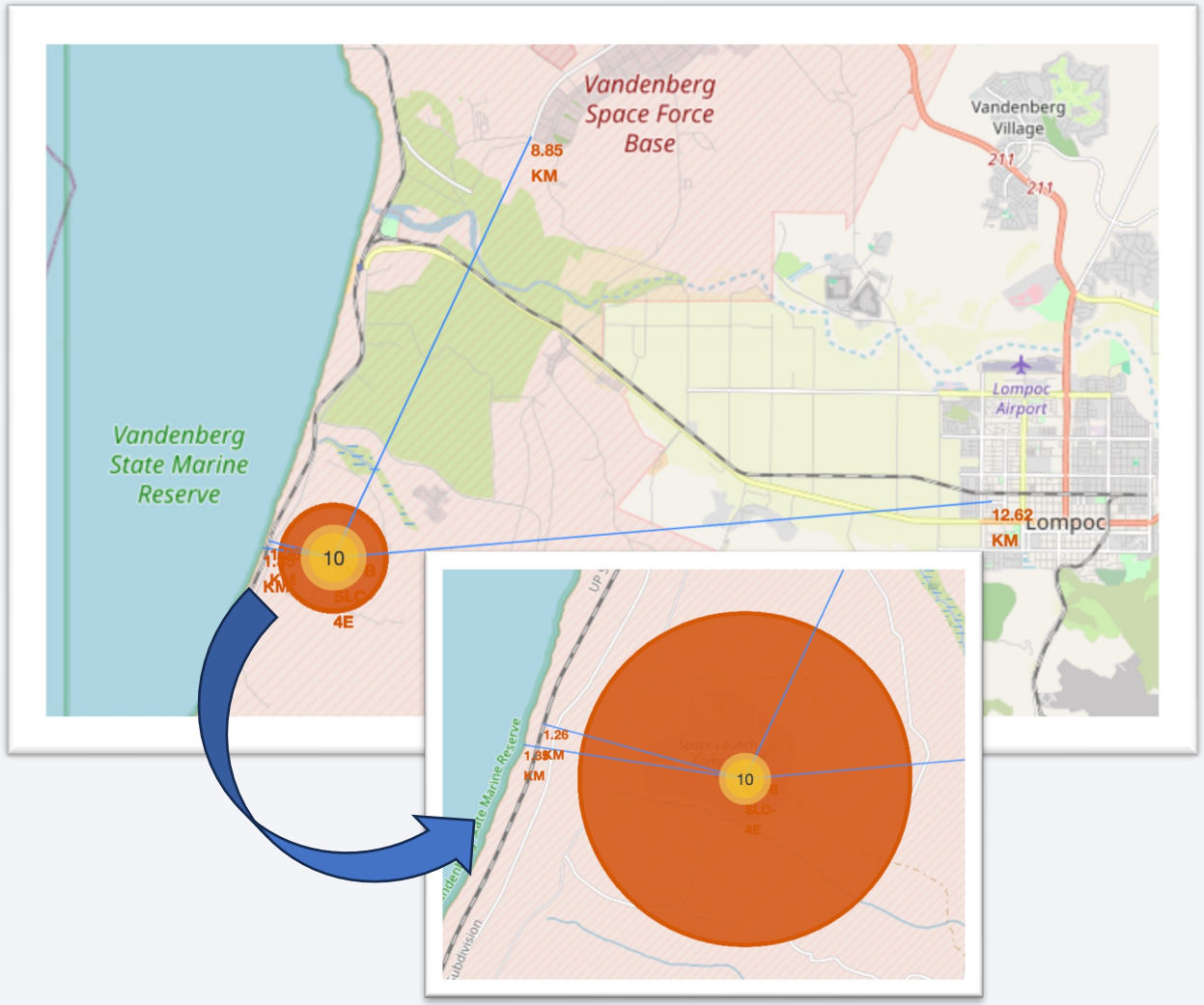


- In the main map, it show us that the launch sites are mainly based in United States.
  - In West part of US, launch site that is presented: VAFB SLC-4E
  - In East part of US, these are launch sites that are presented including: CCAFS LC-40, CCAFS SLC-40 & KSC LC-39A
- Questions that we can answer from this:
  - Are all launch sites in proximity to the Equator line? **YES**
  - Are all launch sites in very close proximity to the coast? **YES**

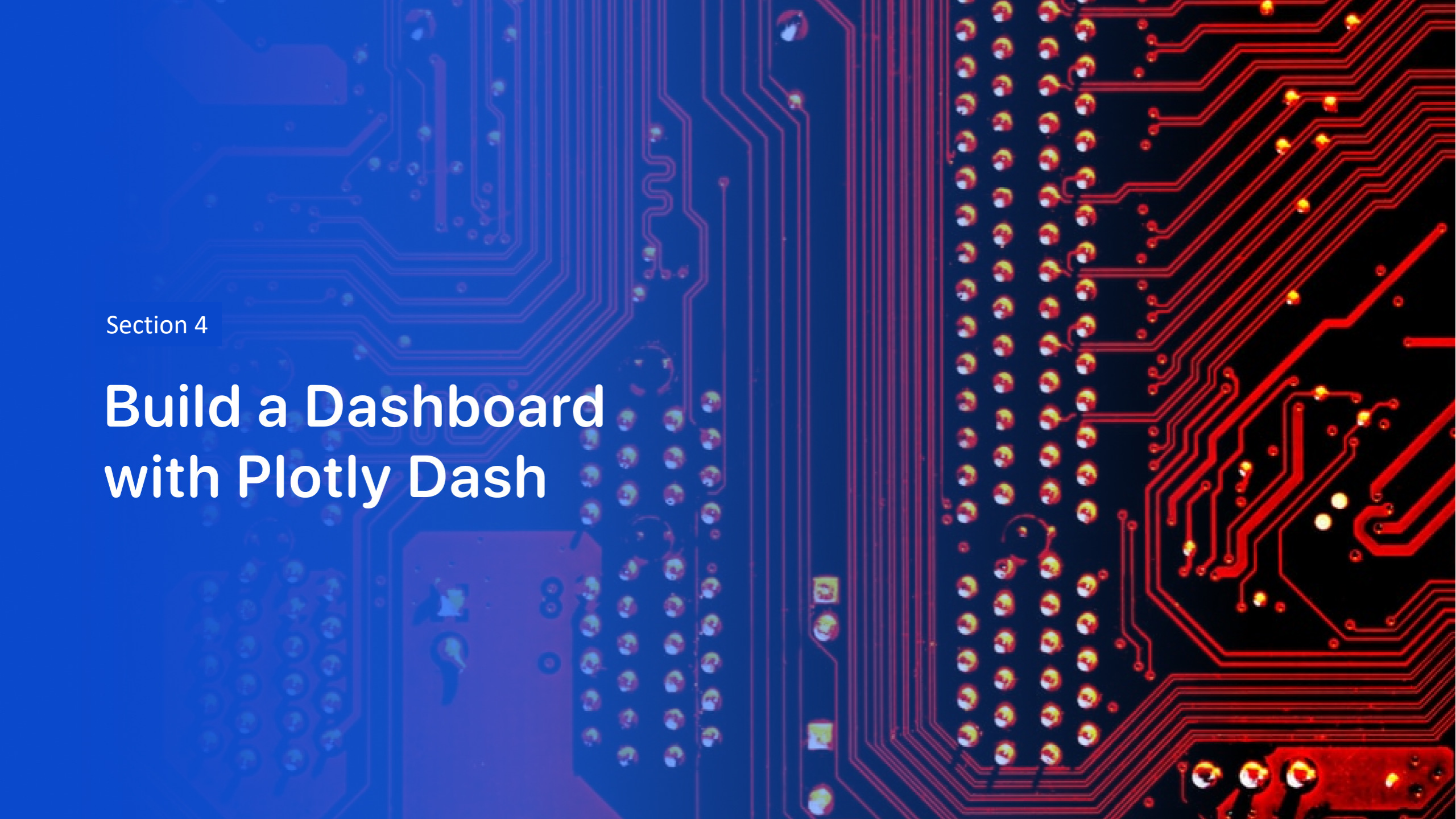
SpaceX – Success/Failure for all Launch Sites

- In the main map, it show the total attempt (including successful & failure) for all launch sites.
- In drill down images, we can see that KSC LC-39A display most positive (green) markup which corresponding to higher success rate, whereas CCAFS SLC-40 has lowest success rate.

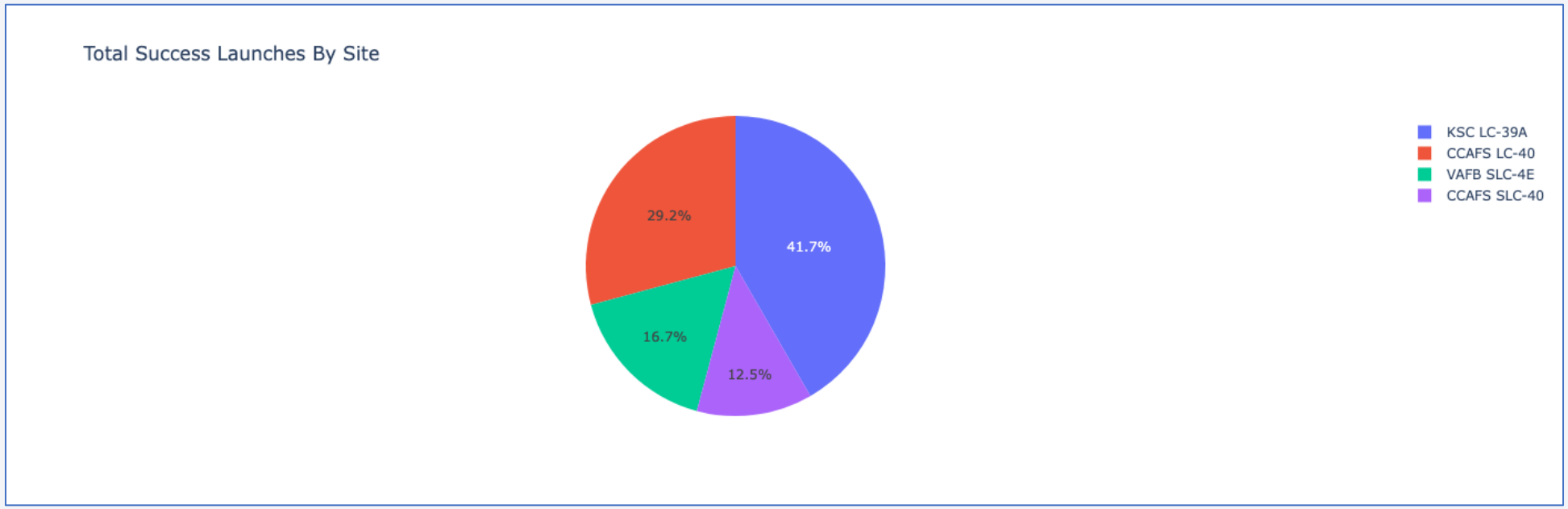
SpaceX – Launch Site to Proximity Distance



- This map show that launch site seems to locate quite far from city (8.85 KM from Space Force Base & 12.62KM from Airport); however, best strategic location for launch site is mostly close to coastline area, railways and highways (1.26 – 1.34 KM) as shown in drill down image.
- Questions that we can answer from this:
  - Are launch sites in close proximity to railways? YES
  - Are launch sites in close proximity to highways? YES
  - Are launch sites in close proximity to coastline? YES
  - Do launch sites keep certain distance away from cities? YES

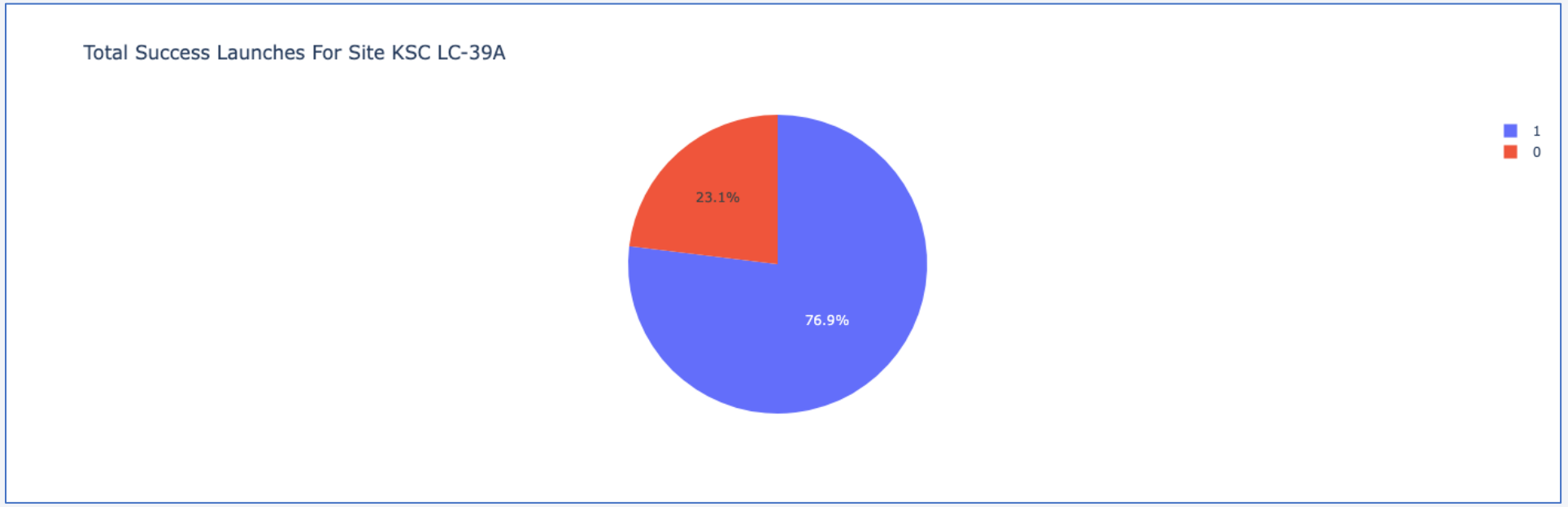

Section 4
Build a Dashboard with Plotly Dash

Total Success Rate By Each Launch Sites



- As shown, launch site KSC LC-39A has the highest success rate (41.7%).
- As shown, launch site CCAFS SLC-40 has the lowest success rate (12.5%).

Success Rate for Launch Site KSC LC-39A



- As shown, the success rate of launch site KSC LC-39A is 76.9% from all its Flight Attempt.

Correlation Between Payload and Success Rate

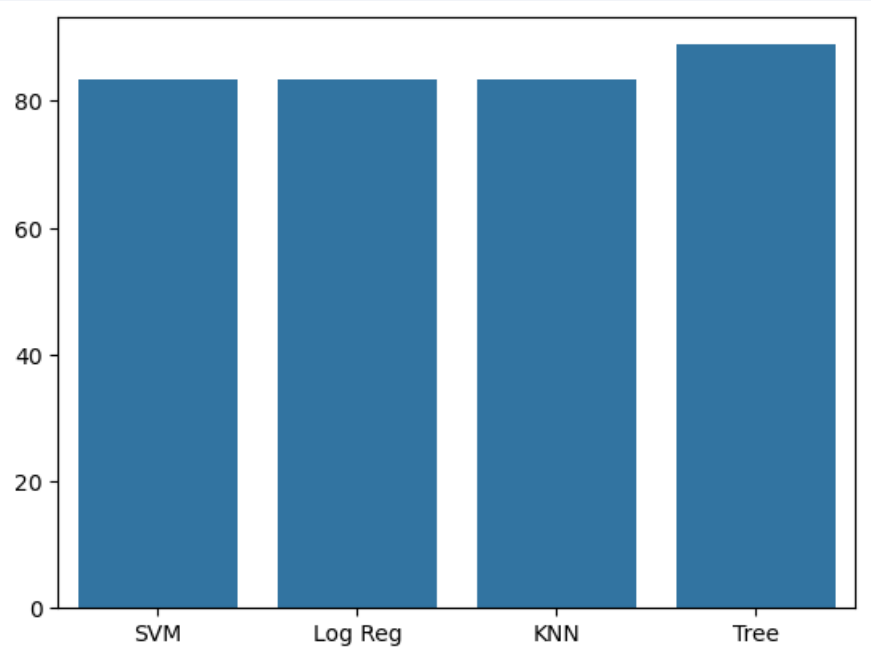
- As shown, Booster Version Category ‘FT’ has the most successful launch.
- As shown, most of successful launch are in range between 2K – 6K Payload Mass (KG).

Section 5

Predictive Analysis (Classification)

Classification Accuracy

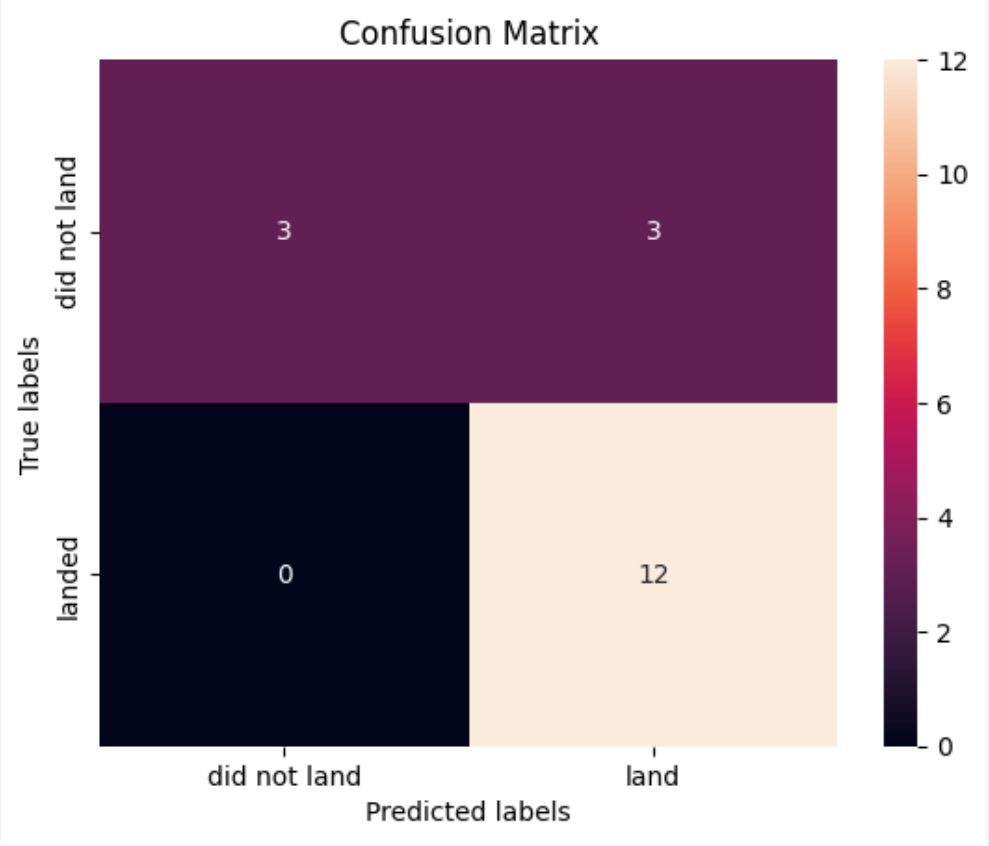
- Based on Bar Chart from Accuracy Test, we can see that Decision Tree method has the highest classification score at 88.88%.
- Whereas, it is the same score (83.33%) for other remaining methods.



	ML Method	Accuracy Score (%)
0	SVM	83.333333
1	Log Reg	83.333333
2	KNN	83.333333
3	Tree	88.888889

Confusion Matrix

- The confusion matrix is same for most of the models (LR, SVM, KNN).
- Per the confusion matrix, the classifier made 18 predictions.
- 12 scenarios were predicted Yes for landing, and they did land successfully (True positive).
- 3 scenarios (top left) were predicted No for landing, and they did not land (True negative).
- 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive).



Conclusions

- As the numbers of flights increase, the first stage is more likely to land successfully.
- Launch success rate increased up to around 80% from 2013 to 2020.
- Launch Site ‘KSC LC-39A’ has the highest launch success rate and Launch Site ‘CCAFS SLC40’ has the lowest launch success rate.
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest.
- Lunch sites are located strategically away from the cities and closer to coastline, railroads, and highways.
- The best performing Machine Learning Classfication Model is the Decision Tree with an accuracy of about 88.8%. When the models were scored on the test data, the accuracy score was about 83% for most of models.

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!