Winning Space Race with Data Science

Giuseppe Rodolfo P. Lomanto
29/01/2024

Outline

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

Executive Summary

- Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization tools
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis

- Summary of all results

- Exploratory Data Analysis results
- Predictive analysis results

Introduction

- Project background and context:

  In the business of commercial space exploration, SpaceX stands out as the most accomplished enterprise, revolutionizing space travel by making it economically viable. The company prominently showcases Falcon 9 rocket launches on its website, boasting a price tag of $62 million while other providers have a demand of $165 million for similar services. A significant factor contributing to this cost disparity lies in SpaceX's innovative approach of reusing the first stage of the rocket. Consequently, the ability to predict the successful landing of the first stage holds the key to estimating the overall launch cost. Using both available data information and machine learning models, our objective is to predict whether SpaceX will reuse the first stage.

- Problems to find answers:

  - How do variables such as payload mass, launch site, number of flights, and orbit types impact the likelihood of a successful first stage landing?

  - What the rate of successful landings over time can tell us?

  - What is the most effective classification algorithm for predicting the reuse of the first stage in this specific context?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - - Using SpaceX Rest API
  - - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - - Filtering the data and dealing with missing values
  - - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using SQL and visualization tools (Matplotlib and Seaborn)
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Defining the most accurate model

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX’s Wikipedia entry.

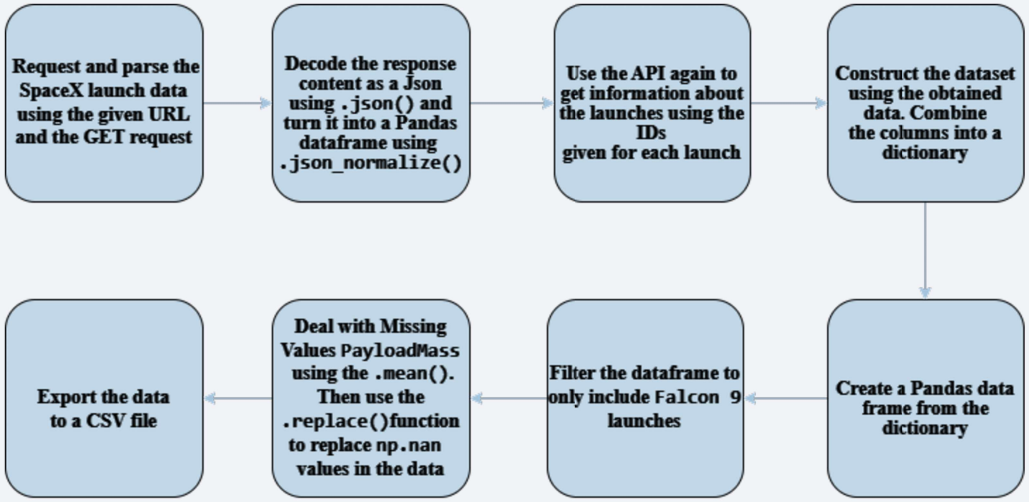
The use of both collection methods were necessary in order to get complete information about the launches for a more detailed analysis.

Data Collection – SpaceX API

- Make a get request to the SpaceX API with some basic data wrangling and formating.

  - Request to the SpaceX API

  - Clean the requested data



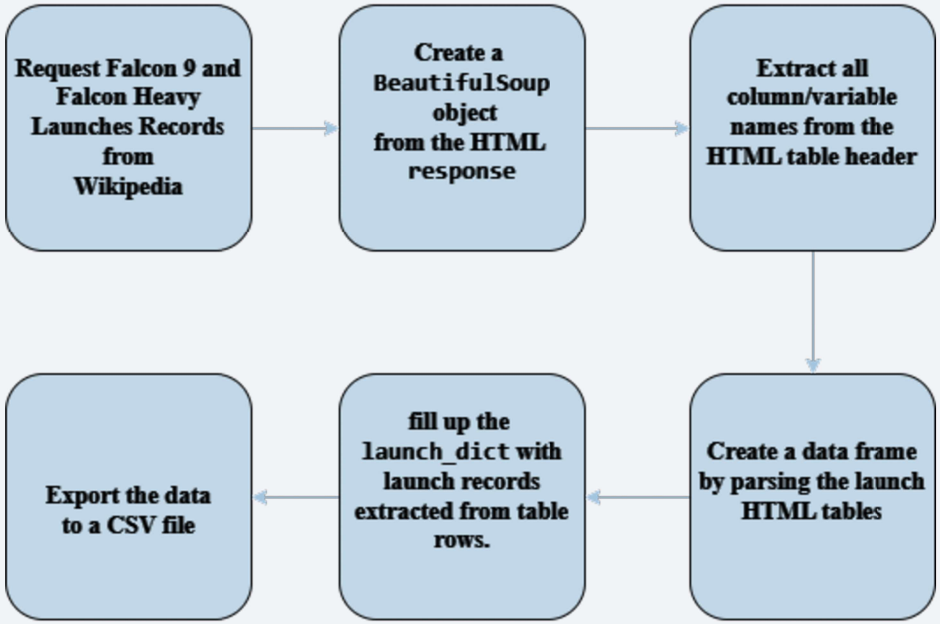
https://github.com/rodolfoplng/IBM-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Web scrap Falcon 9 launch records with BeautifulSoup:

  - Extract a Falcon 9 launch records HTML table from Wikipedia

  - Parse the table and convert it into a Pandas data frame



https://github.com/rodolfoplng/IBM-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

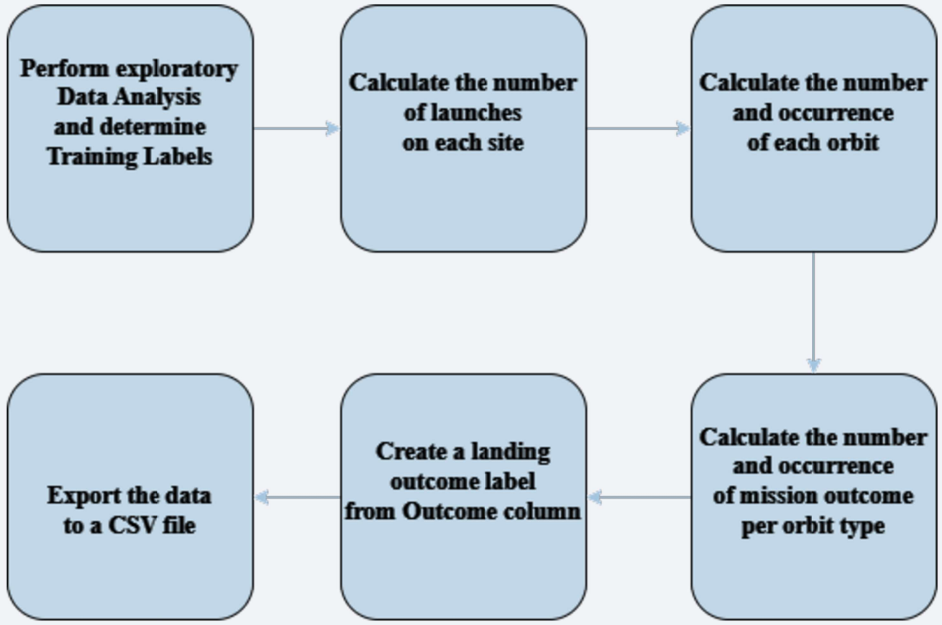
Data Wrangling

- Perform exploratory Data Analysis and determine Training Labels

- Exploratory Data Analysis

- Determine Training Labels

https://github.com/rodolfoplng/IBM-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Charts plotted:

- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series).

https://github.com/rodolfoplng/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Performed SQL queries:

• Displaying the names of the unique launch sites in the space mission

• Displaying 5 records where launch sites begin with the string ‘CCA'

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

• Displaying average payload mass carried by booster version F9 v1.1

• Listing the date when the first successful landing outcome in ground pad was achieved

• Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

• Listing the total number of successful and failure mission outcomes

• Listing the names of the booster versions which have carried the maximum payload mass

• Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

• Ranking 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

https://github.com/rodolfoplng/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

https://github.com/rodolfoplng/IBM-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the

Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

https://github.com/rodolfoplng/IBM-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels

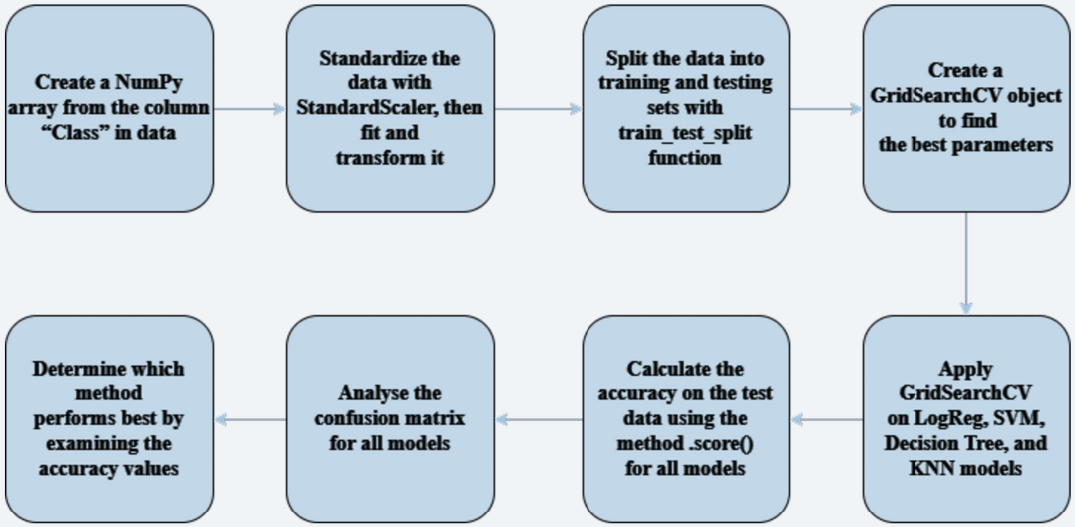
- Create a column for the class

- Standardize the data

- Split into training data and test data

-Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

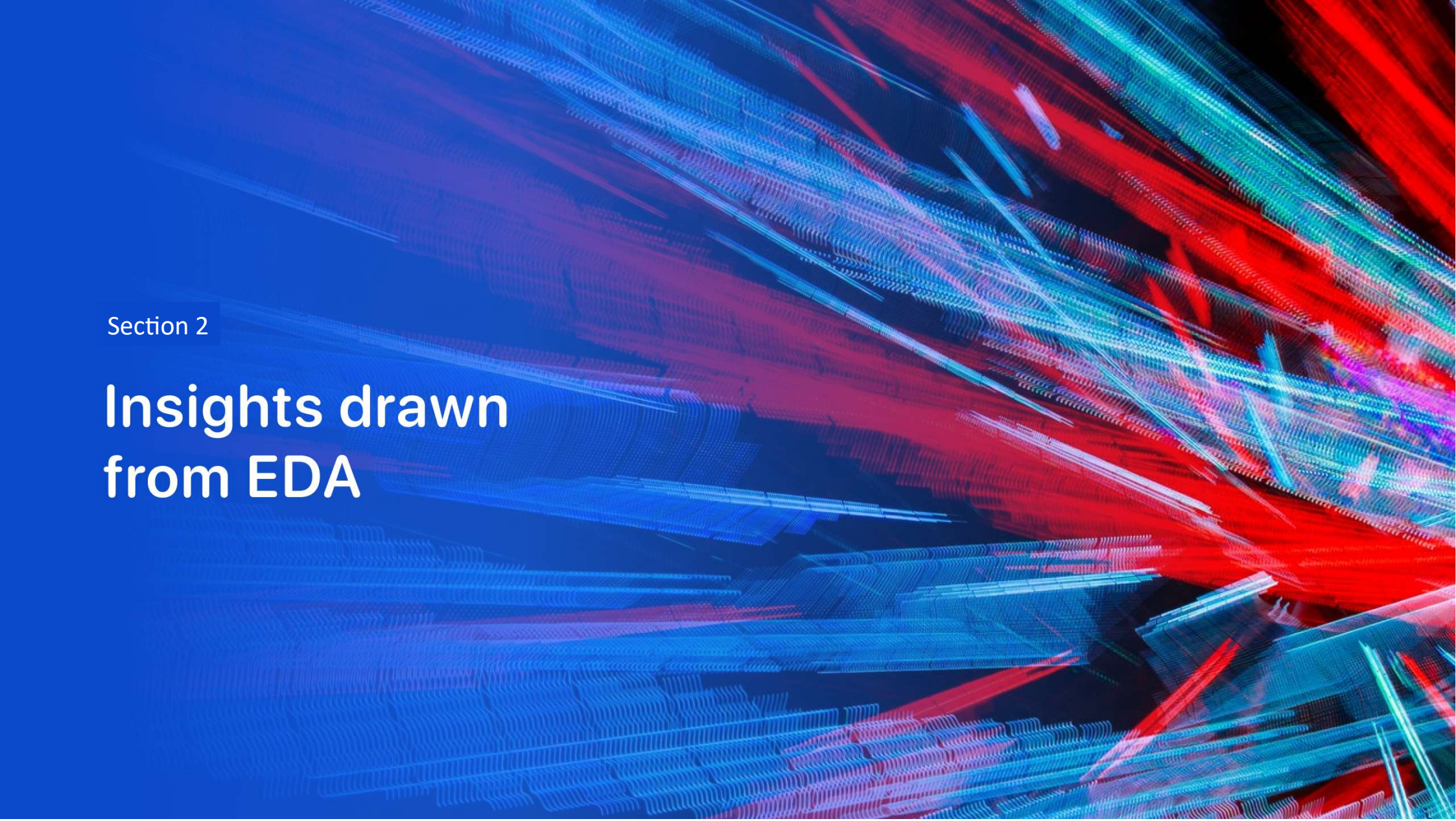
- Show the method that performs best using test data



https://github.com/rodolfoplng/IBM-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

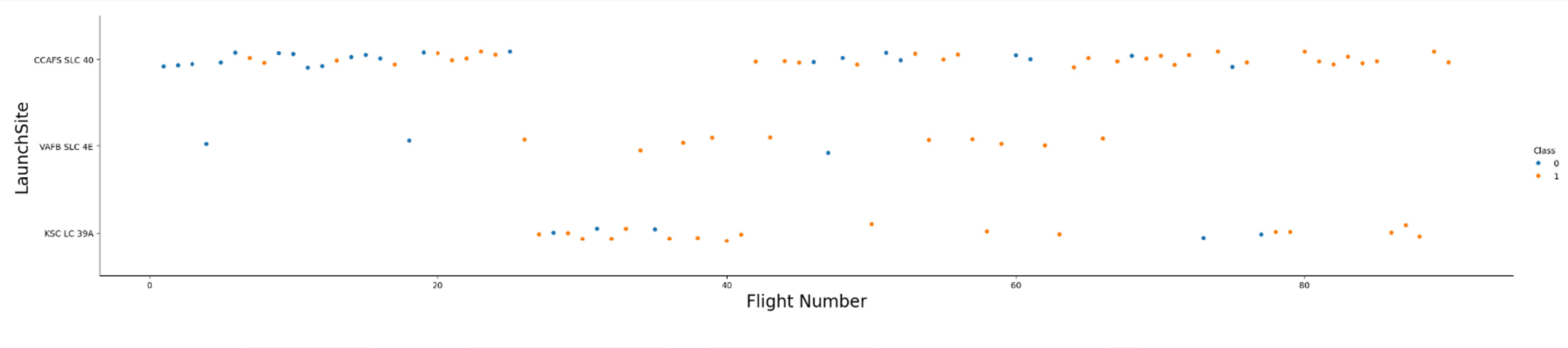
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

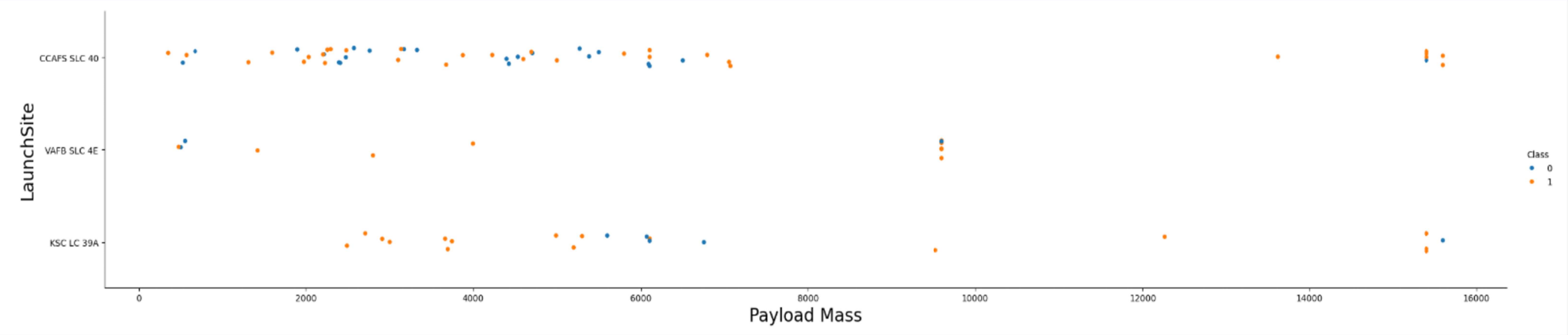
Insights drawn from EDA

Flight Number vs. Launch Site



- First flight numbers had a higher failure rate.
- The CCAFS SLC 40 launch site has the most amount of launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates, but were not used at earlier flights.
- Each new launch has a higher chance of success.

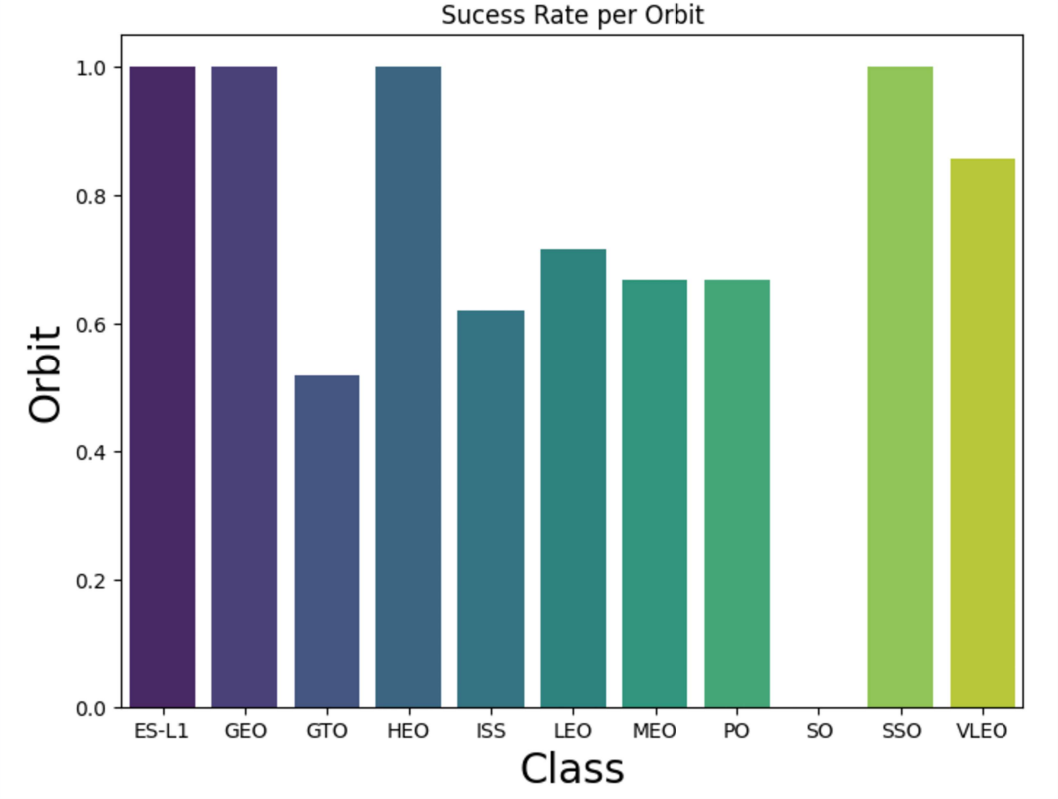
Payload vs. Launch Site



- Most of the launches with payload mass over 7000 kg were successful.
- All launches on KSC LC 39A were successful for payload mass under 5500 kg.

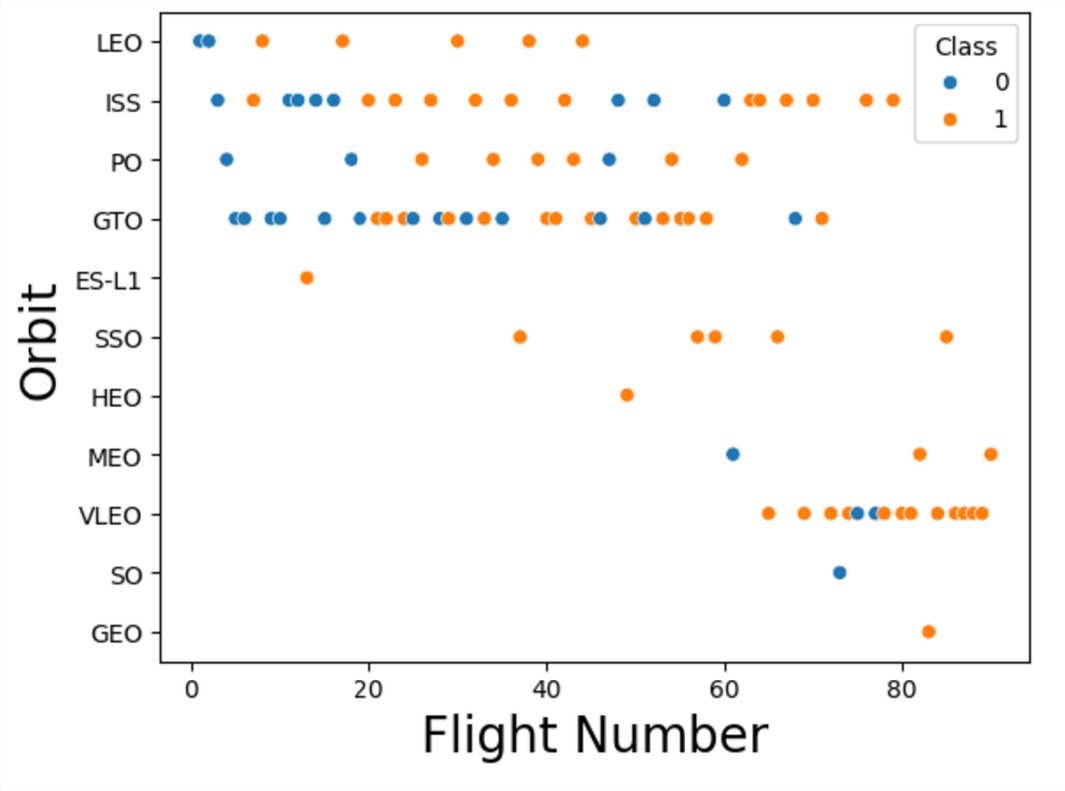
Success Rate vs. Orbit Type

- Orbit types and altitudes doesn’t necessarily have influence on success rate.
- Low , medium and high orbits have similar success rates.



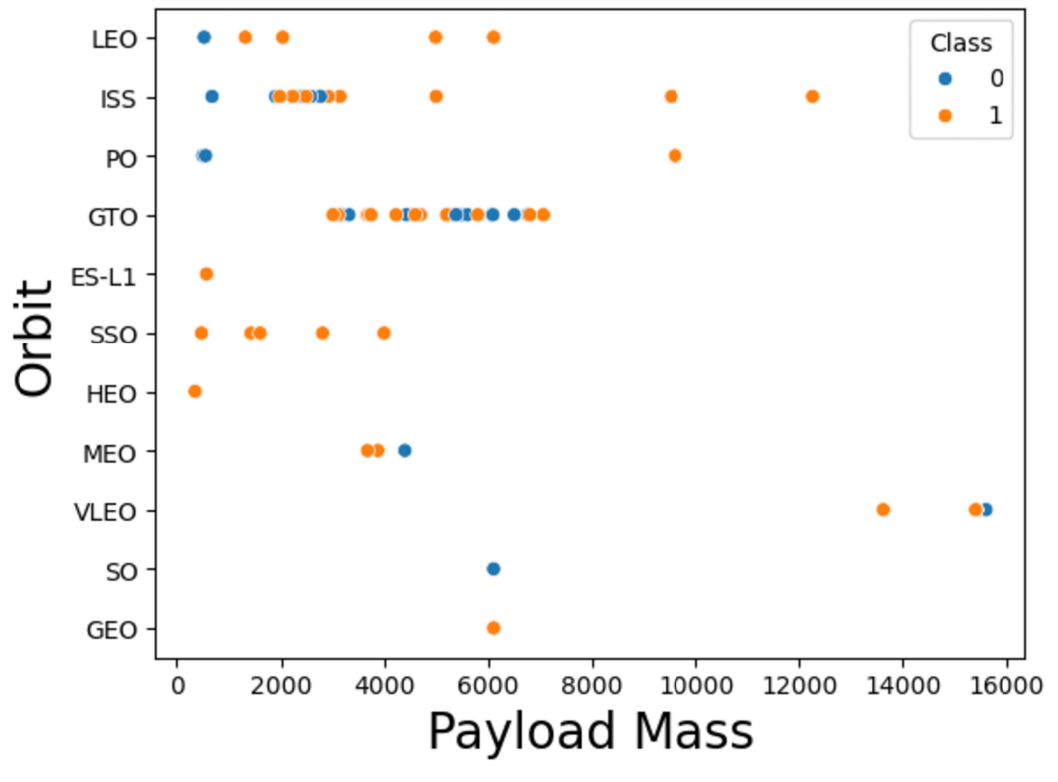
Flight Number vs. Orbit Type

- First flight numbers had a higher failure rate.
- Earlier flights were launched on most common orbits (LEO, ISS, Polar and GTO)



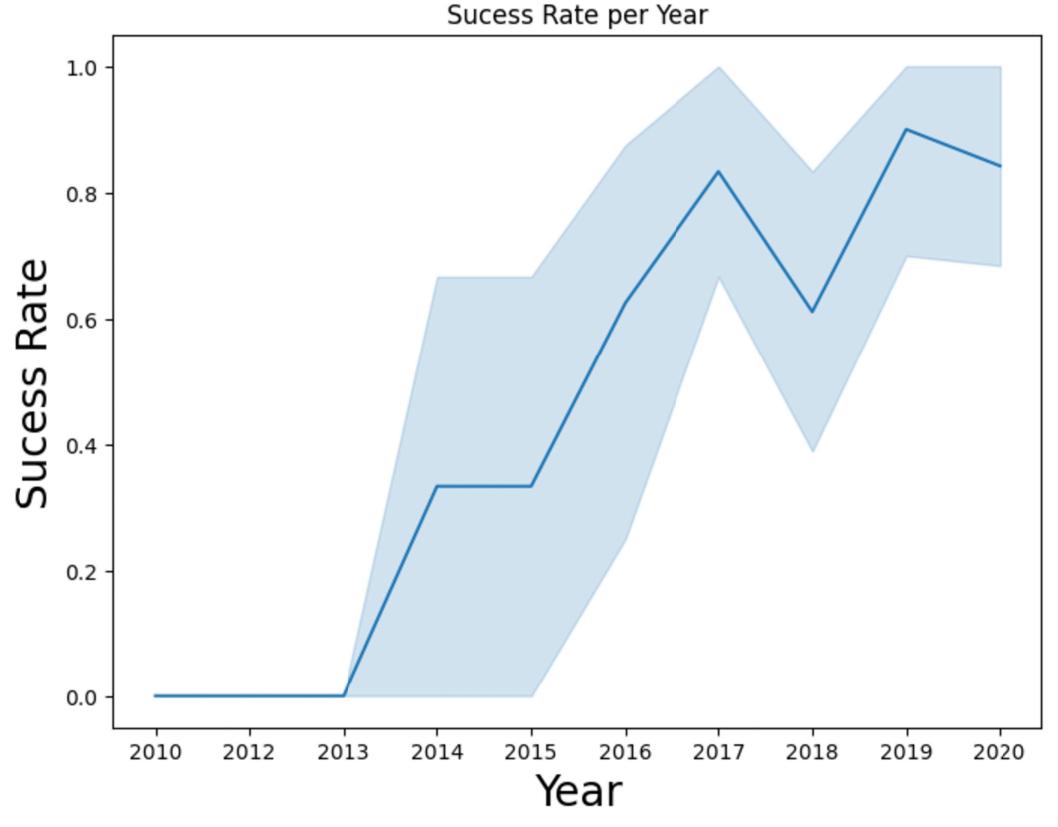
Payload vs. Orbit Type

- Heavier payloads were launched on lower orbits (LEO and ISS), with one failure case.
- Lighter payloads were more distributed across orbit types, with a balanced sucess rate at GTO orbit.



Launch Success Yearly Trend

- Sucess rate kept increasing since 2013 till 2020, with a slight fall in 2018.



All Launch Site Names

- Four distinct launch sites used by SpaceX.

Task 1

Display the names of the unique launch sites in the space mission

In [13]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

Out[13]:

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

Launch Site Names Begin with 'CCA'

- Displaying 5 records where launch sites begin with the string 'CCA'.

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

```
In [15]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

```
In [17]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)";
 * sqlite:///my_data1.db
Done.
Out[17]: SUM(PAYLOAD_MASS__KG_)
                          45596
```

- Total payload mass carried by boosters launched by NASA (CRS): 45596 kg.

Average Payload Mass by F9 v1.1

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

```
In [18]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1";
 * sqlite:///my_data1.db
Done.
```

Out[18]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average payload mass carried by boosters version F9 v1.1: 2928,4 kg.

First Successful Ground Landing Date

```
In [23]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)";
 * sqlite:///my_data1.db
Done.
Out[23]: MIN(Date)
         2015-12-22
```

- First successful ground landing date: 22/12/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [26]:
```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ B
```

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

Out[26]:

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

- Listing the names of the boosters which have success in drone ship landing and with payload mass between than 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [29]:
```
%sql SELECT COUNT(*) FROM SPACEXTABLE WHERE Mission_Outcome IN ('Success', 'Failure');
```

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

Out[29]:

COUNT(*)
98

- Total number of successful and failure mission outcomes: 98.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
In [31]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXT
 * sqlite:///my_data1.db
Done.
```

Out[31]:

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

- Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

```
In [40]: %sql SELECT substr(Date, 6,2) as Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE substr(Date,0,
 * sqlite:///my_data1.db
Done.
```

Out[40]:

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

- Failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

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

In [54]:
```
%sql SELECT Landing_Outcome, COUNT(*) as "COUNT" FROM SPACEXTABLE WHERE Date BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY
```

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

Out[54]:

Landing_Outcome	COUNT
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

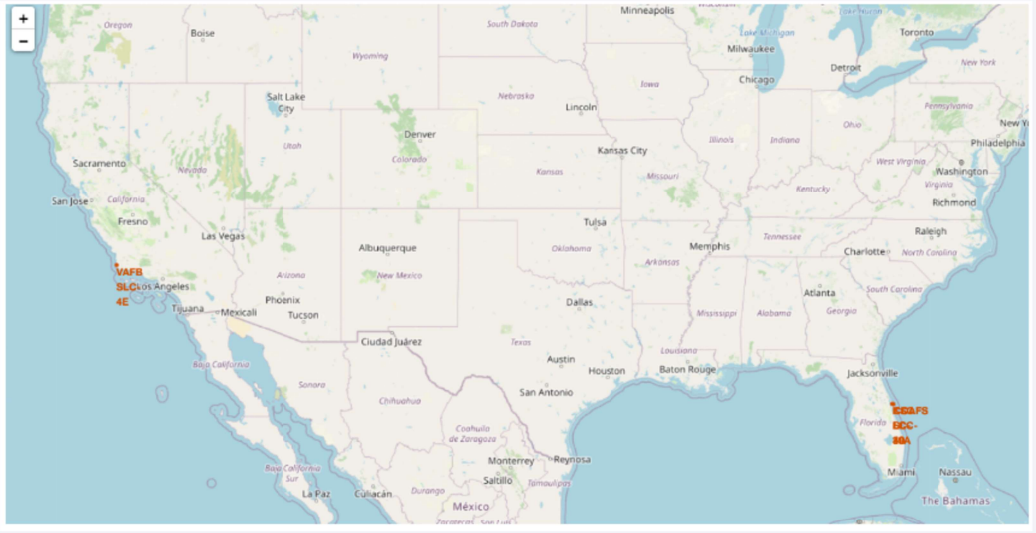
- Count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Section 3

Launch Sites Proximities Analysis

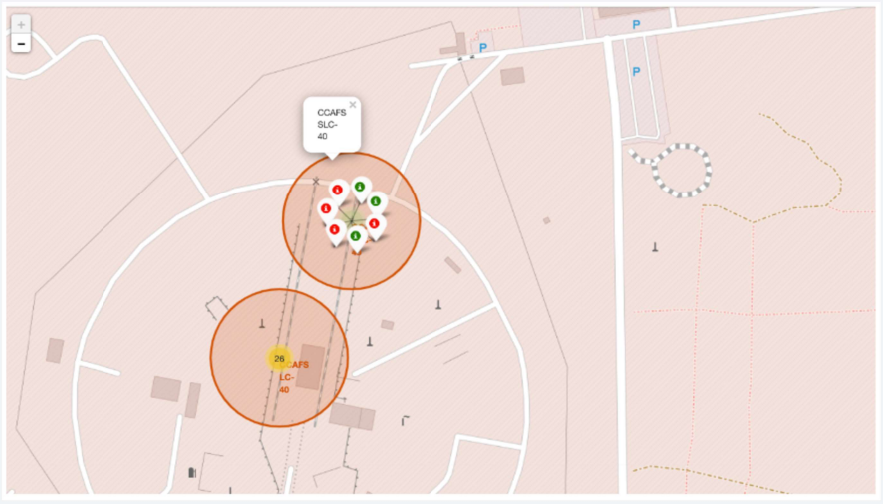
Launch sites locations

- All launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast.



Launch records on the map

- Green Marker = Successful Launch
- Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.

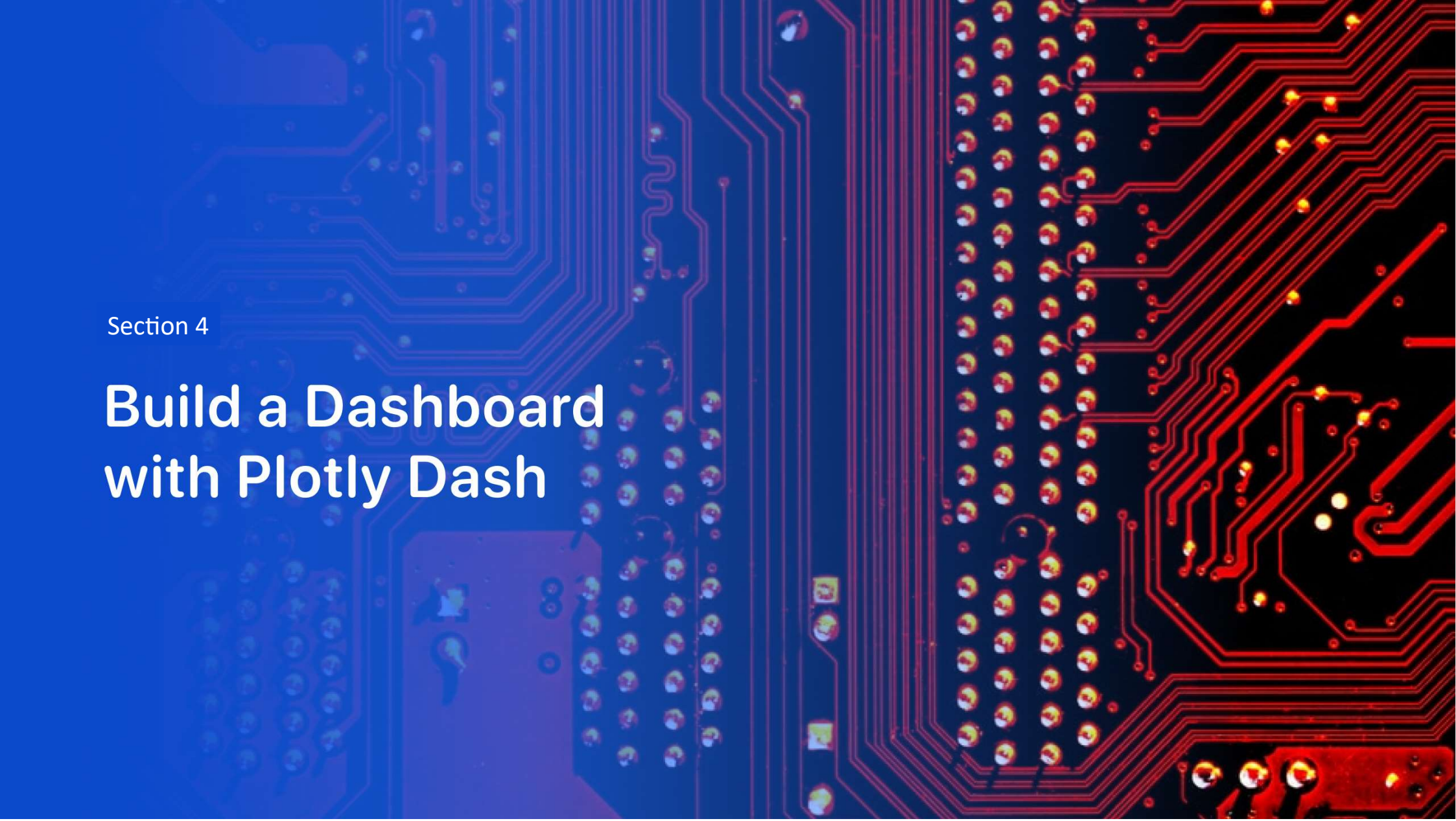


Distances from launch site KSC LC-39A

Replace <Folium map screenshot 3> title with an appropriate title

Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

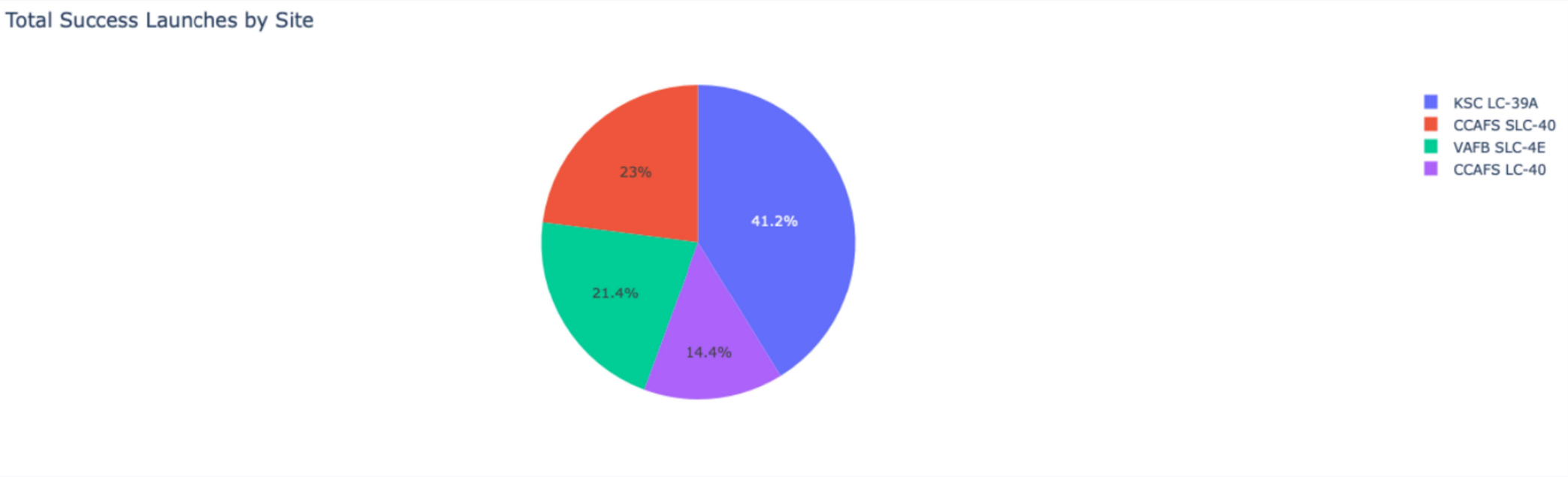
Explain the important elements and findings on the screenshot



Section 4

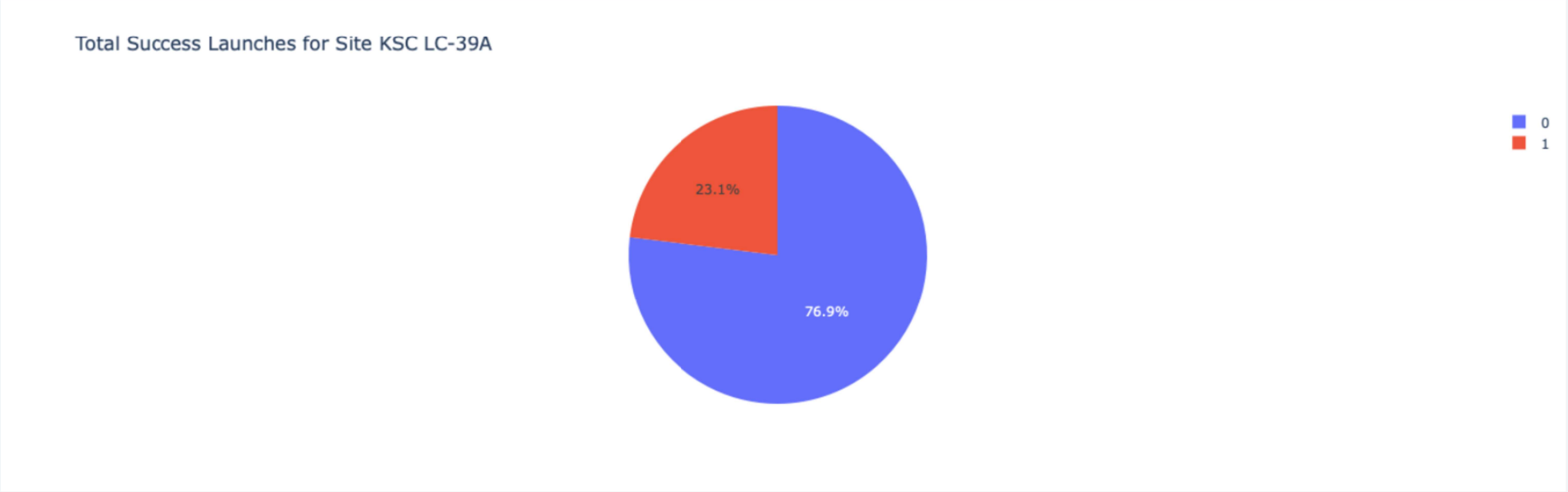
Build a Dashboard with Plotly Dash

Total launch success for all sites



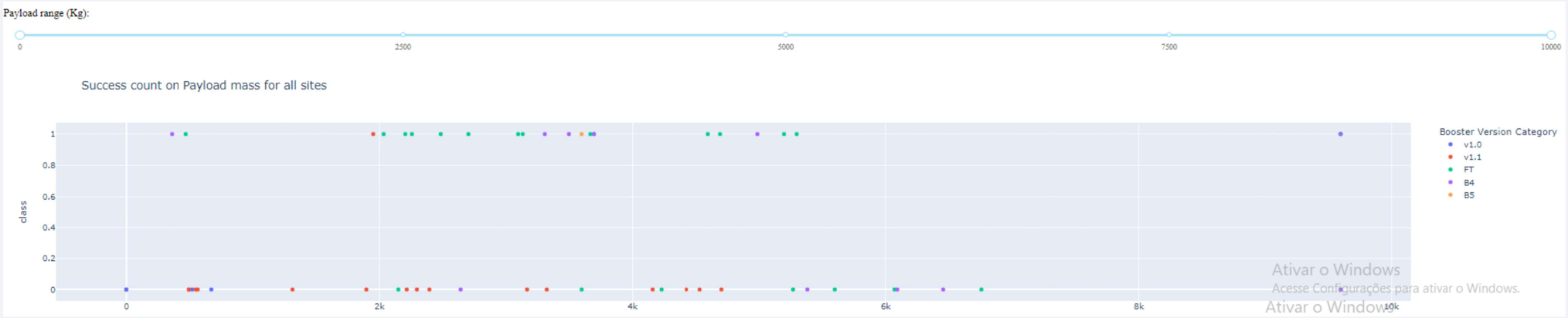
- Launch site KSC LC-39A has the most absolute successful launches.

Launch site with highest success ratio



• KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites



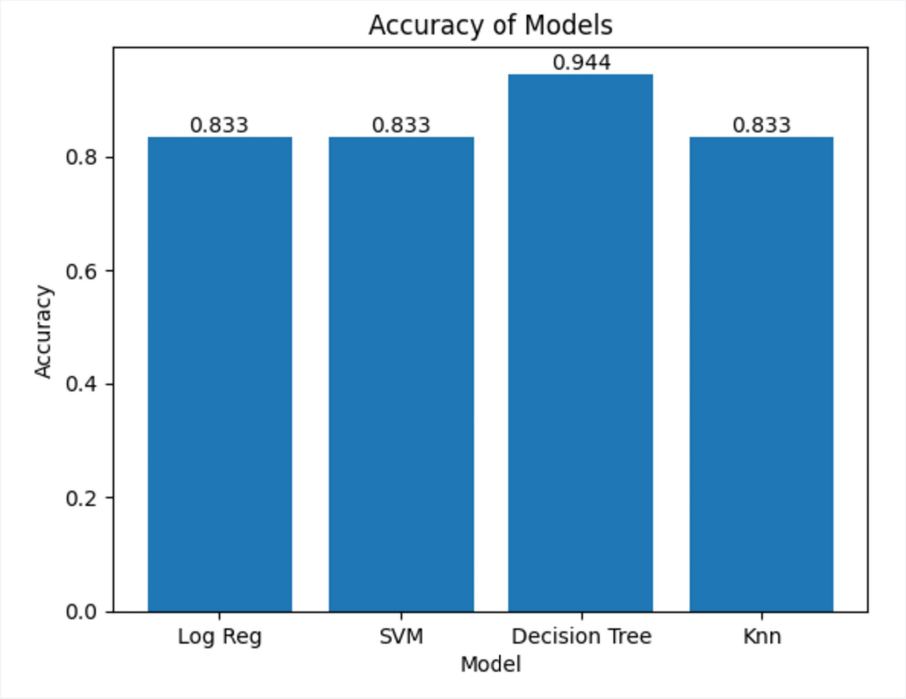
- Boosters v1.0 and v 1.1 have higher failure rate

Section 5

Predictive Analysis (Classification)

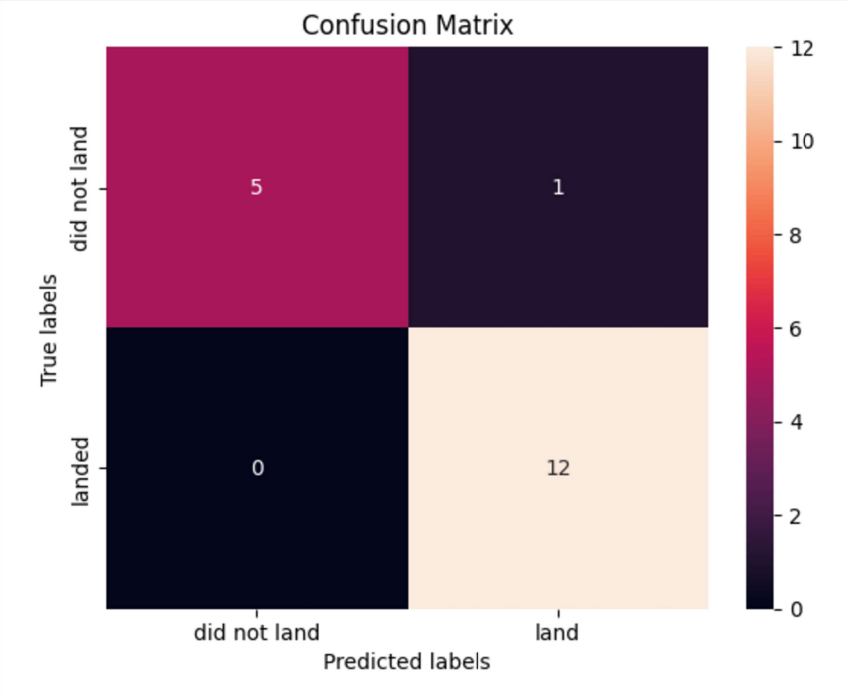
Classification Accuracy

- The Decision Tree model is the most accurate one.



Confusion Matrix

- The confusion matrix for the Decision Tree model shows no false negatives and only one true positive.



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- The success rate of launches increased over the years.
- KSC LC-39A launch site has the highest success rate.

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!