Winning Space Race with Data Science

Isabel Yong
12/11/2023

Outline

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

Executive Summary

- Summary of methodologies: Data was gathered both through an API call and through web scraping
- Summary of all results: We will derive results from exploratory data analysis, data visualization and predictive analysis

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- We will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We worked with SpaceX launch data that is gathered from the SpaceX REST API and we also collected data on Wiki pages through web scraping.
- Perform data wrangling
  - We dealt with missing values by replacing them with the mean value and we created a column for the outcome of the launch.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We compared classification models using functions in the sklearn library.

Data Collection

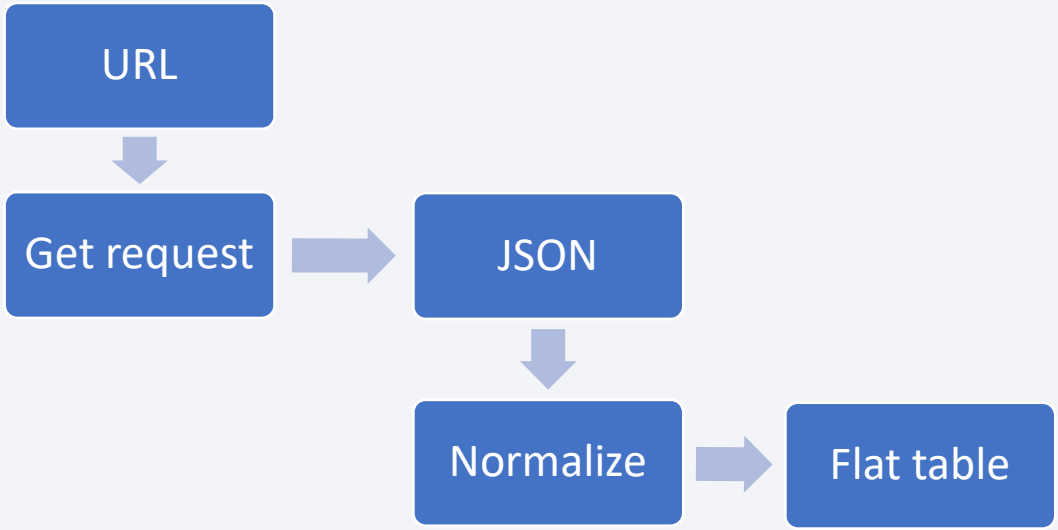
- API call to https://api.spacexdata.com/v4/rockets/

- Web scraping the WIKI page https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

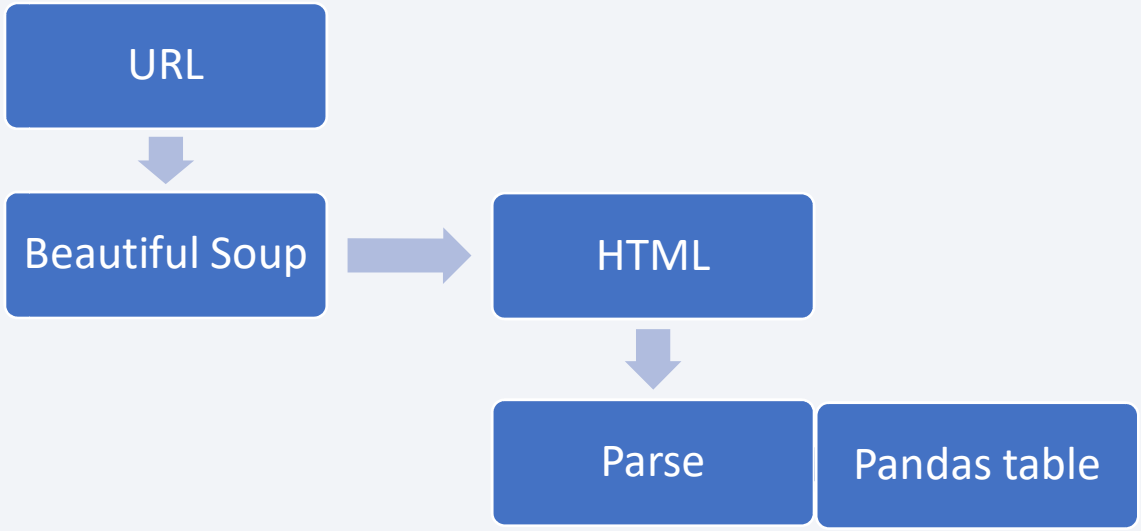
Data Collection – SpaceX API

- We use the Request library to get the data from the provided URL in form of a JSON file.
- We normalize the JSON file to obtain a flat table.
- https://github.com/isabelyong/Capstone_Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- We use the BeautifulSoup library to web scrape the wiki page provided and get an HTML document.
- We parse it to obtain the desired table: Falcon 9 rocket launches.
- https://github.com/isabelyong/Capstone_Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

Dealing with NULL values

- When the payload mass was unknown we replaced it with the mean value

Adding a column for Class

- We added a column for the outcome of the launch: 1 for Success; 0 for Failure.

- https://github.com/isabelyong/Capstone_Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- To understand the relationship between variables, we created scatter plots:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Launch Site vs. Payload Mass
  - Flight Number vs. Orbit Type
  - Payload Mass vs. Orbit Type
- To visualize the success rate per orbit type we used a bar chart
- To visualize the yearly trend of the success rate we used a line chart
- https://github.com/isabelyong/Capstone_Project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- We computed SQL queries to obtain the following information:
  - What are the launch sites?
  - What is the total and average Payload Mass per Booster Version?
  - When was the first successful landing on a ground pad?
  - Which Booster Versions have carried the maximum Payload Mass?
  - What is the total number of successful and failure mission outcomes?
- https://github.com/isabelyong/Capstone_Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We used Folium maps and highlighted Launch Sites with circles and markers.
- We used Marker Clusters to indicate the launches from each site (color-coded based on success/failure)
- We draw lines to mark the distances and analyze the proximities of launch sites.
- link

Build a Dashboard with Plotly Dash

- We created a dashboard to monitor successful/unsuccessful launches from each site
- Though a dropdown menu it’s possible to choose which launch site to analyze
- A pie chart shows the distribution of successful launches
- With a slider we can choose the payload mass range
- The scatter plot shows payload mass vs outcome
- link

Predictive Analysis (Classification)

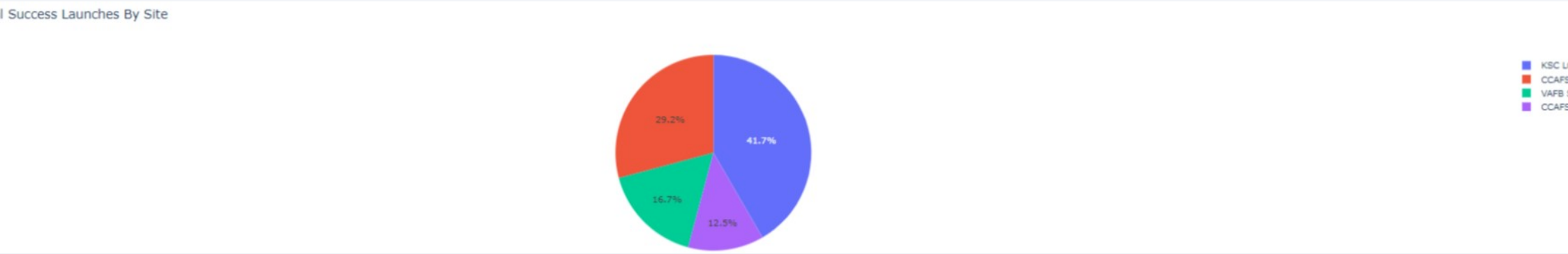
- We evaluated the following classification models:
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree classifier
  - K nearest neighbors (KNN)

- Evaluation Methods:
  - Accuracy
  - Confusion matrix

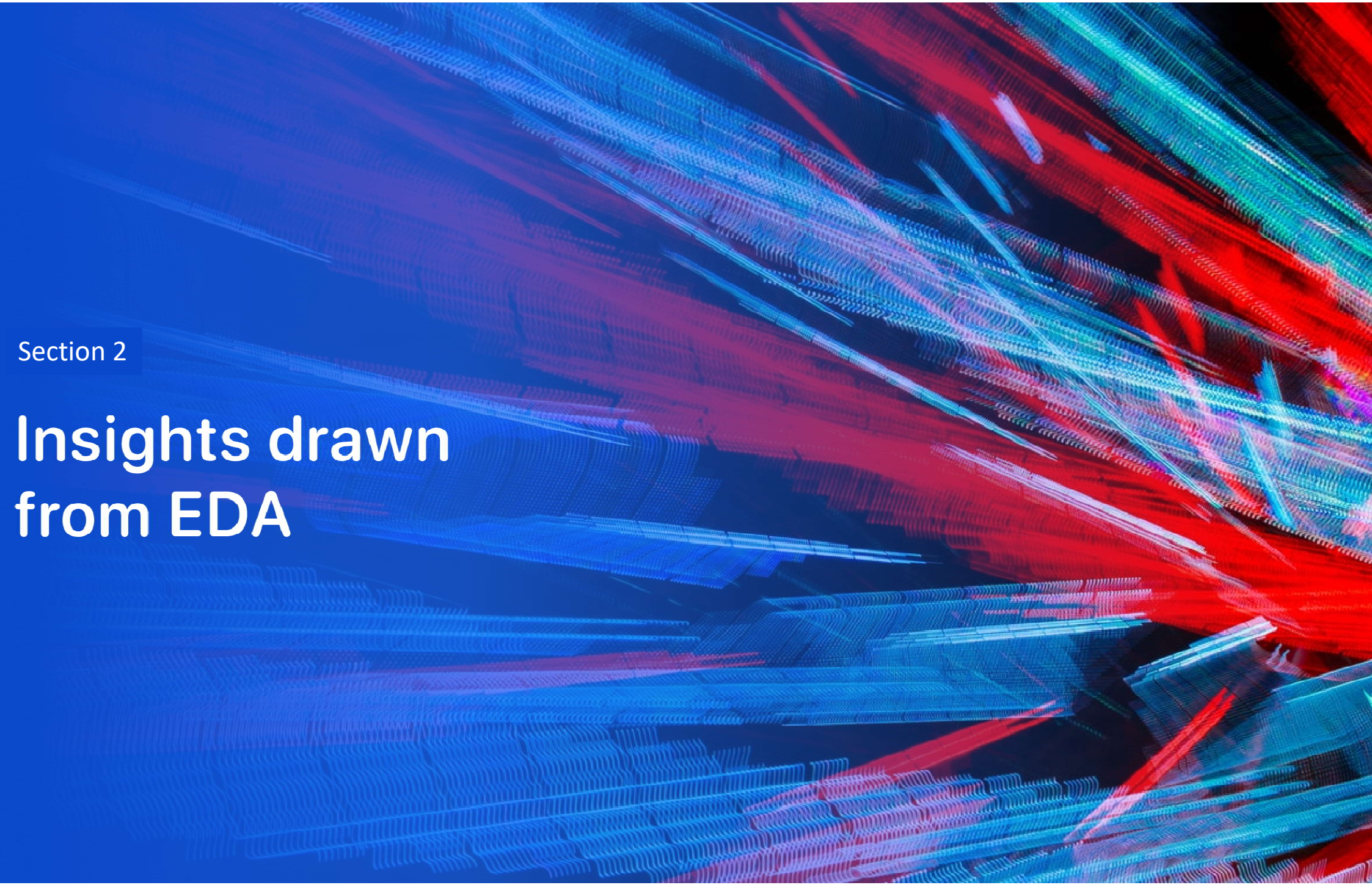
Results

- Different launch sites have different success rates:
  - CCAFS LC-40, has a success rate of 60%.
  - KSC LC-39A and VAFB SLC 4E have a success rate of 77%.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10 000).
- Some orbit types have a success rate of 100%: ES-L1, GEO, HEO, SSO
- Generally, the success rate has increased over the years

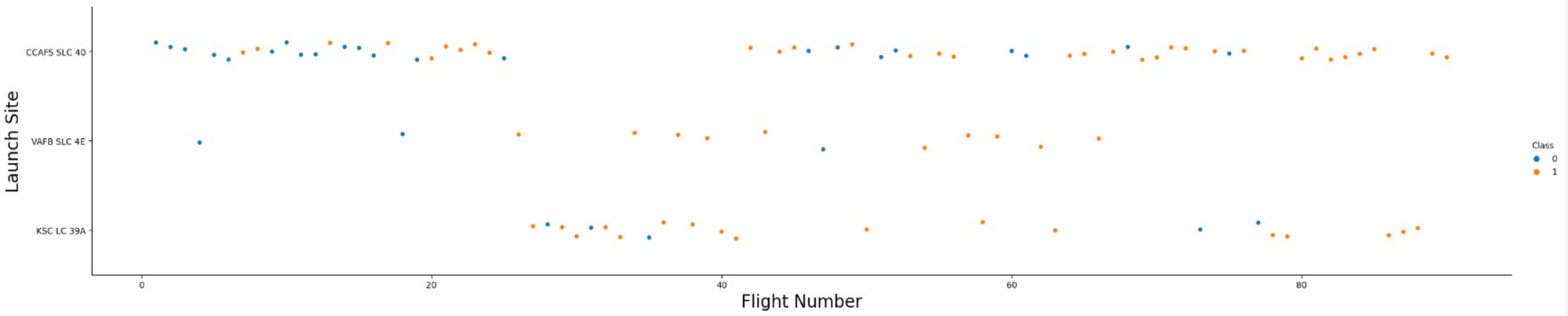
Results



- Out of all the launch sites, KSC LC-39A has the most successful launches
- The decision tree classifier is the best predictive model

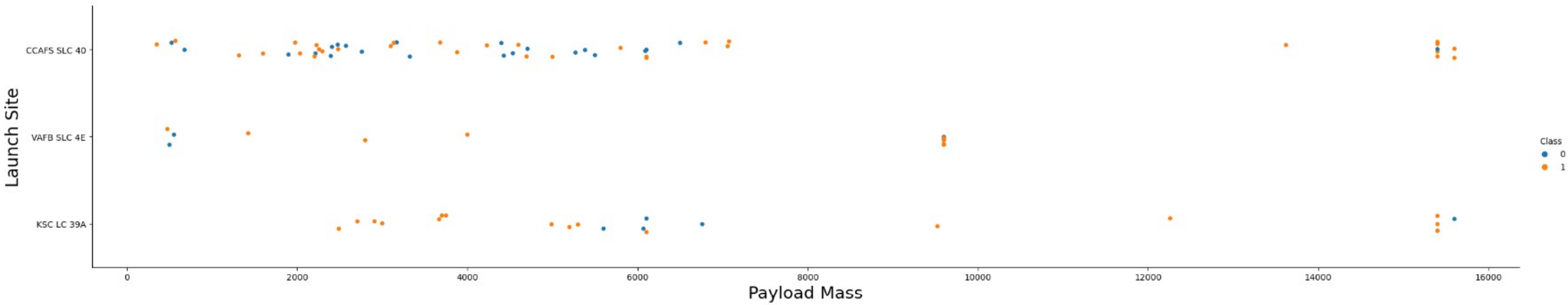

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



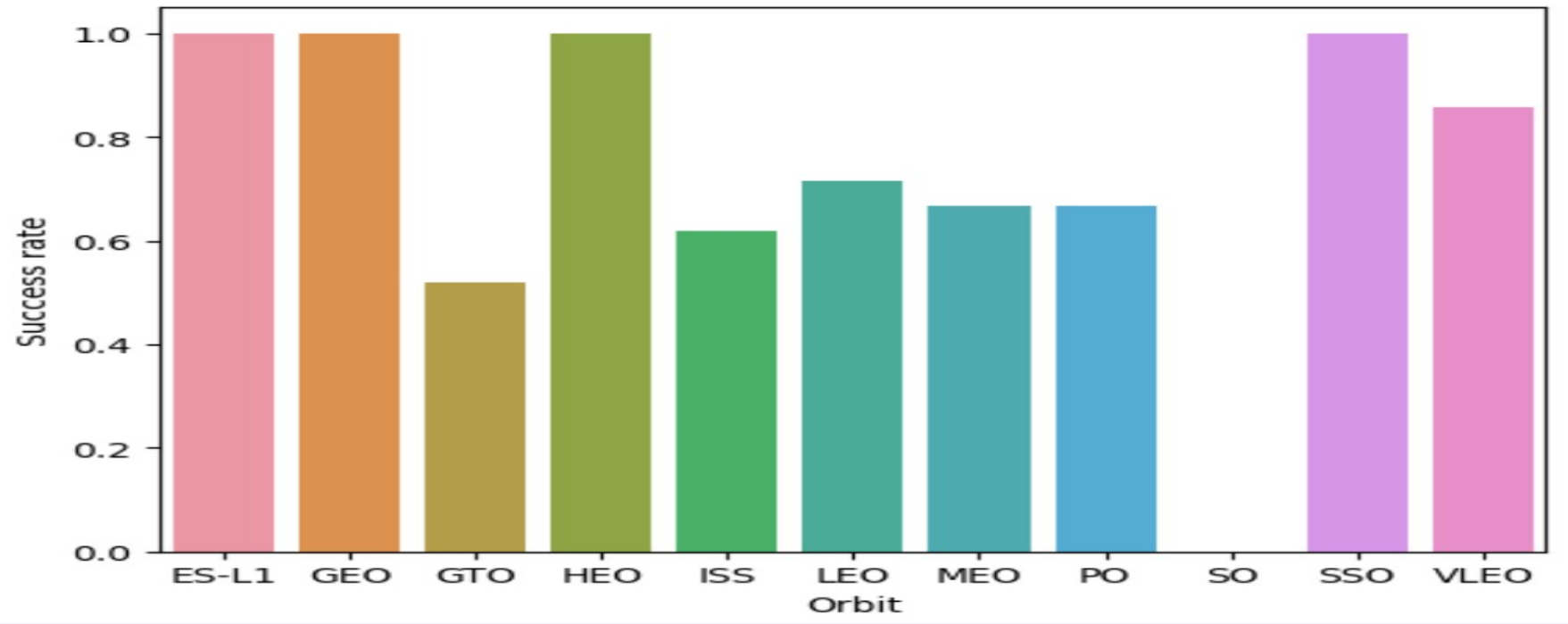
- CCAFS SLC 40 has the most launches.
- For all launch site, the success rate increases with flight number

Payload vs. Launch Site



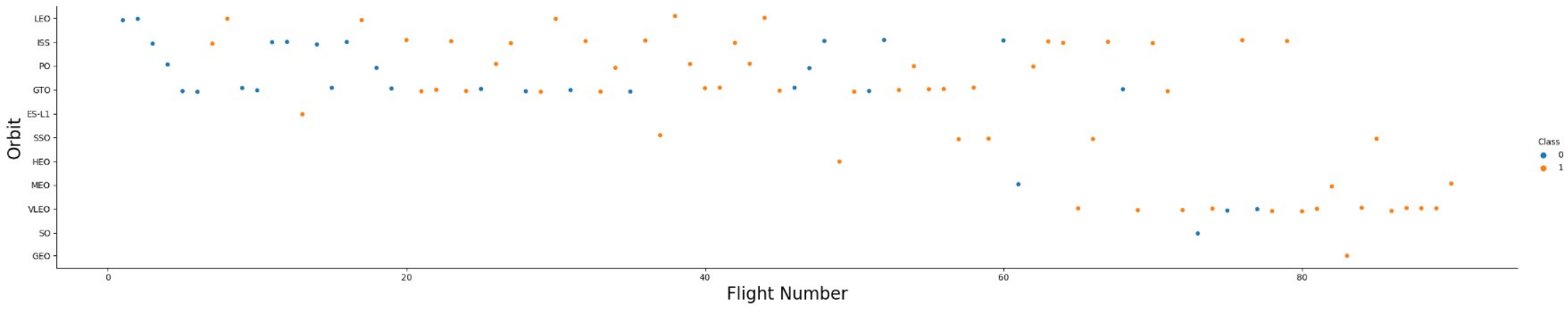
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10 000 kg)
- Most launches are below 8000 kg
- Heavier launches have a higher success rate

Success Rate vs. Orbit Type



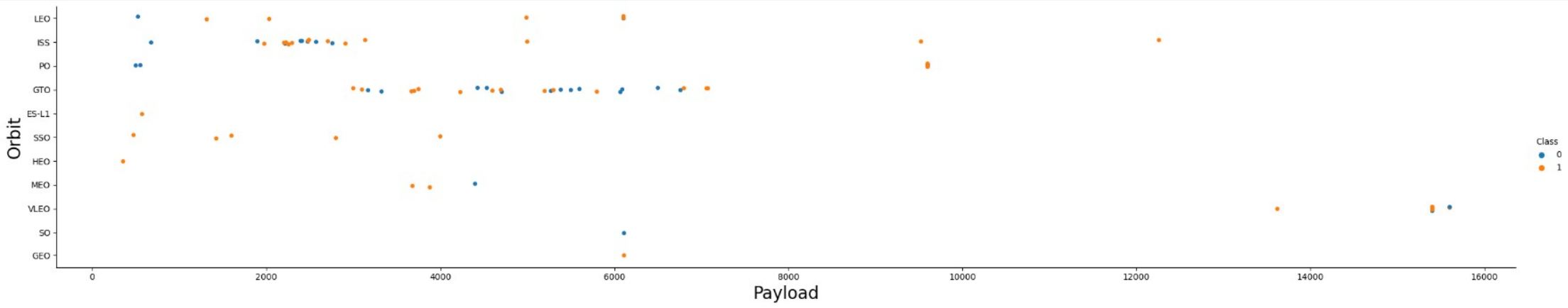
- ES-L1, GEO, HEO, SSO have a success rate of 100%
- SO has a success rate of 0%
- Orbit type strongly influences the outcome of the launch

Flight Number vs. Orbit Type



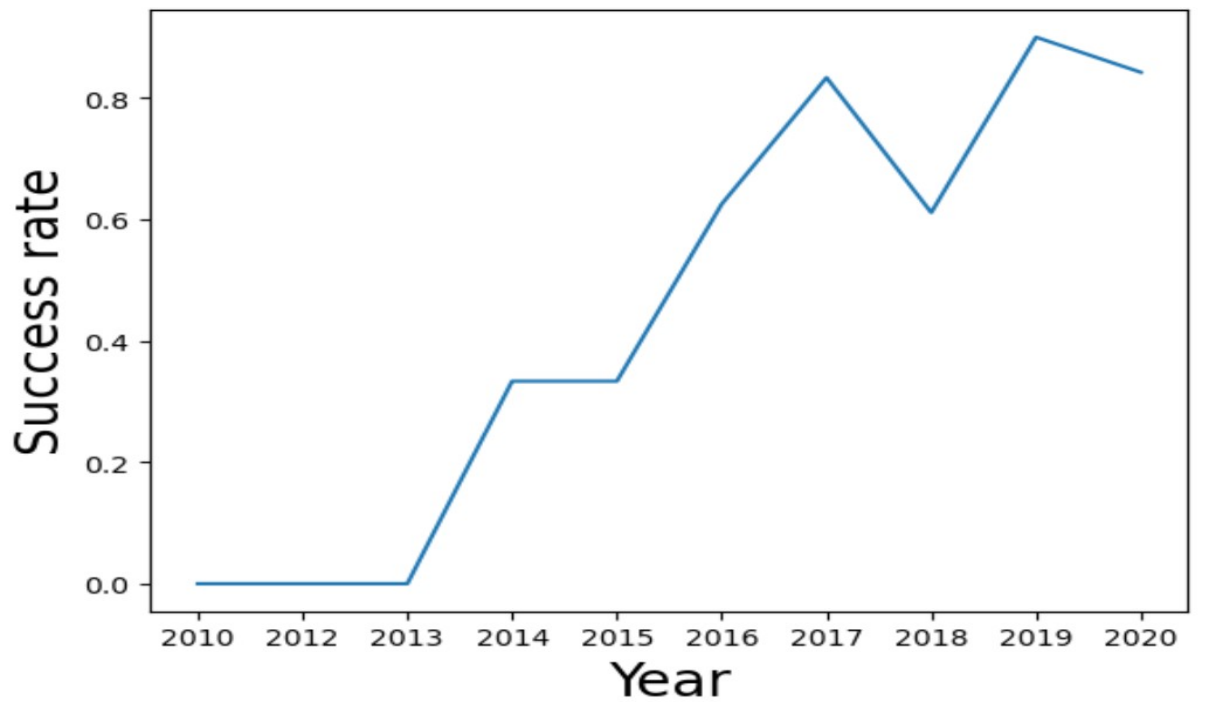
- In the LEO orbit the success appears related to the number of flights
- There is no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



- With heavy payloads, the success rate is higher for PO and LEO
- We cannot establish a clear relationship between Payload and success rate for GTO orbit

Launch Success Yearly Trend



- the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing again

All Launch Site Names

- These are the Launch Sites involved in our study

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

Launch Site Names Begin with 'CCA'

- These are 5 examples of records for launch sites beginning with ‘CCA’

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcor
6/4/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachu
2/8/2010	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 (parachu
05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
0/8/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
3/1/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Total Payload Mass

- This is the total payload mass carried by boosters launched by NASA (CRS)

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- This is the average payload mass carried by booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- This is the date of the first successful launch

```
Date    2015-12-22
dtype: datetime64[ns]
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- These are the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000¶

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

Total Number of Successful and Failure Mission Outcomes

- This is a list of the total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

- These are the booster versions which have carried the maximum payload mass

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

2015 Launch Records

- This is a list of the records which displays the months, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

	Month	Landing_Outcome	Booster_Version	Launch_Site
13	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
14	11	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
15	2	No attempt	F9 v1.1 B1014	CCAFS LC-40
16	4	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
17	4	No attempt	F9 v1.1 B1016	CCAFS LC-40
18	6	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
19	12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

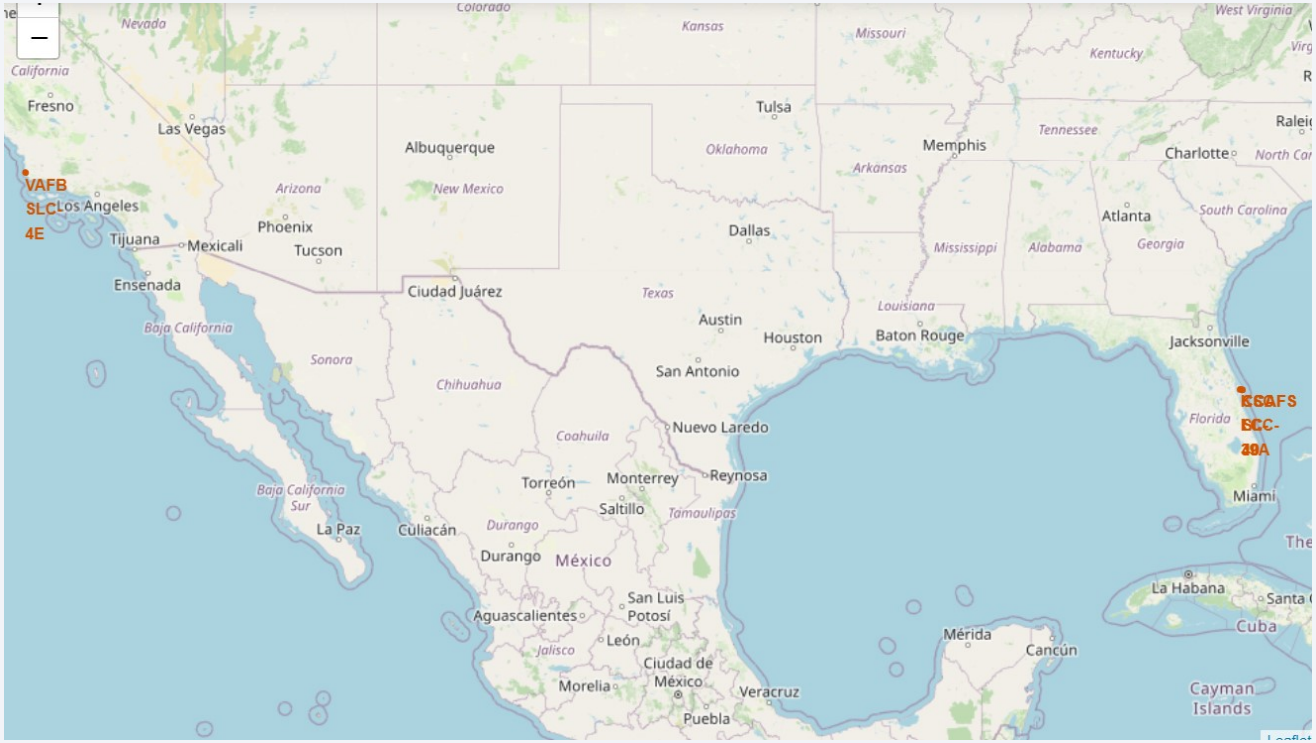
- This is a ranking of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

	Landing_Outcome
3	No attempt
1	Failure (drone ship)
5	Success (drone ship)
6	Success (ground pad)
0	Controlled (ocean)
7	Uncontrolled (ocean)
2	Failure (parachute)
4	Precluded (drone ship)


Section 3
Launch Sites
Proximities Analysis

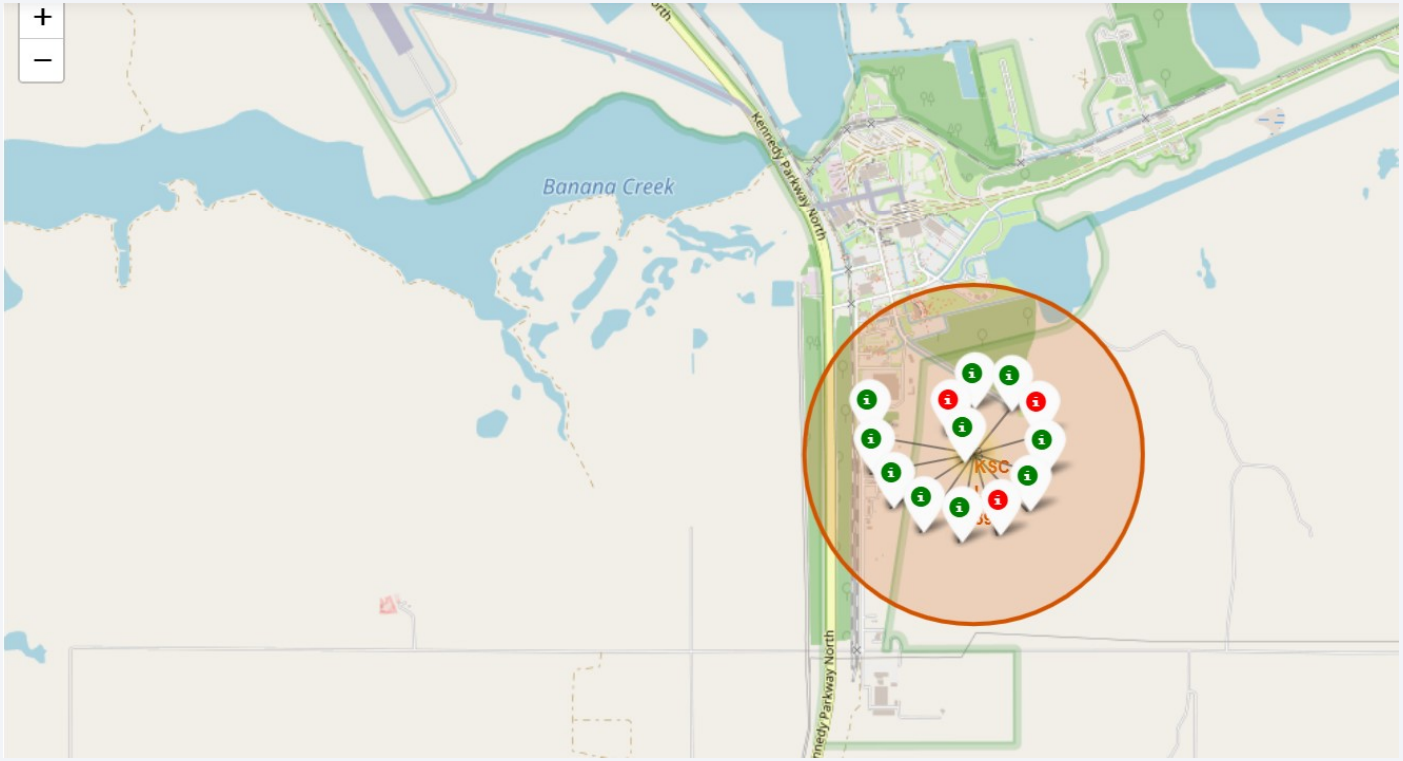
Launch Site locations

- On this map we highlighted the four launch sites with markers and circles



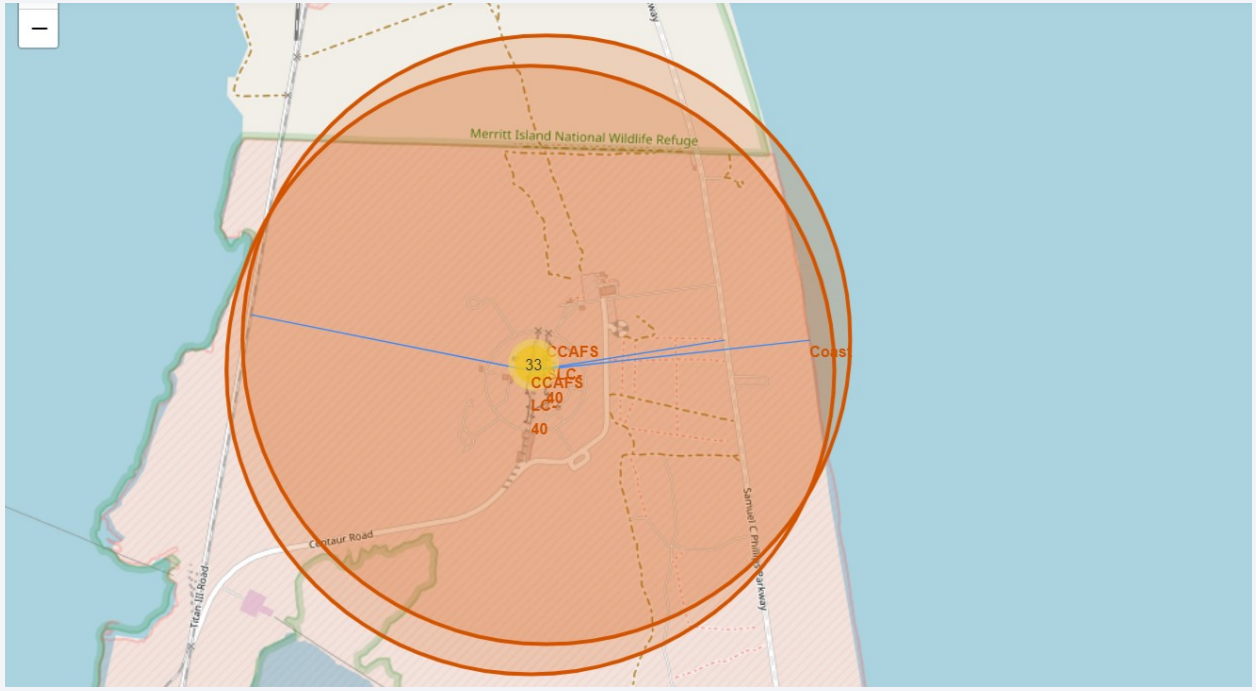
Outcomes of the launches on each site

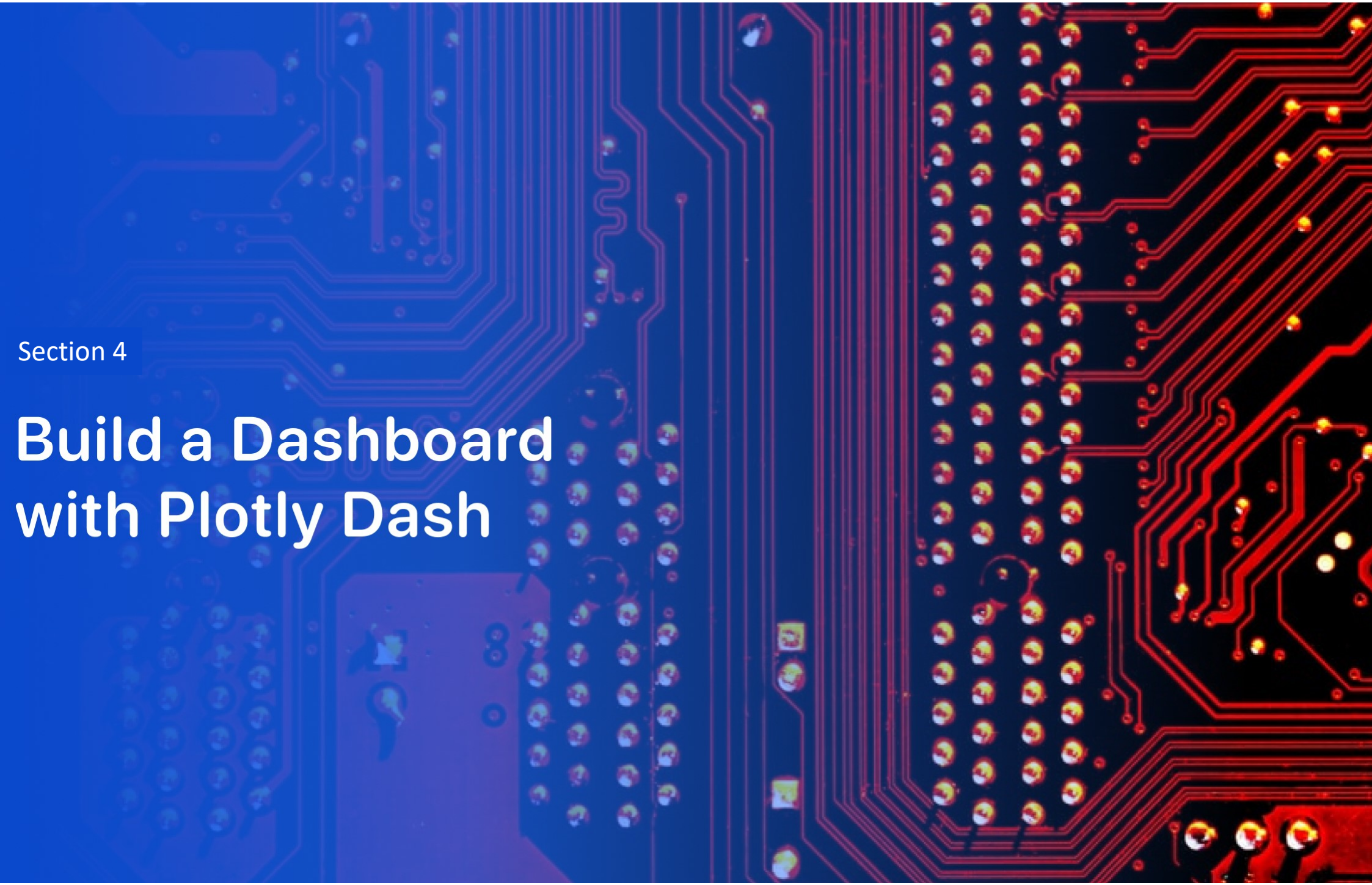
- This screenshot shows the outcomes of the launches from one of the sites



<Folium Map Screenshot 3>

- The blue lines highlight the distance to the coast/railway/highway from one of the sites



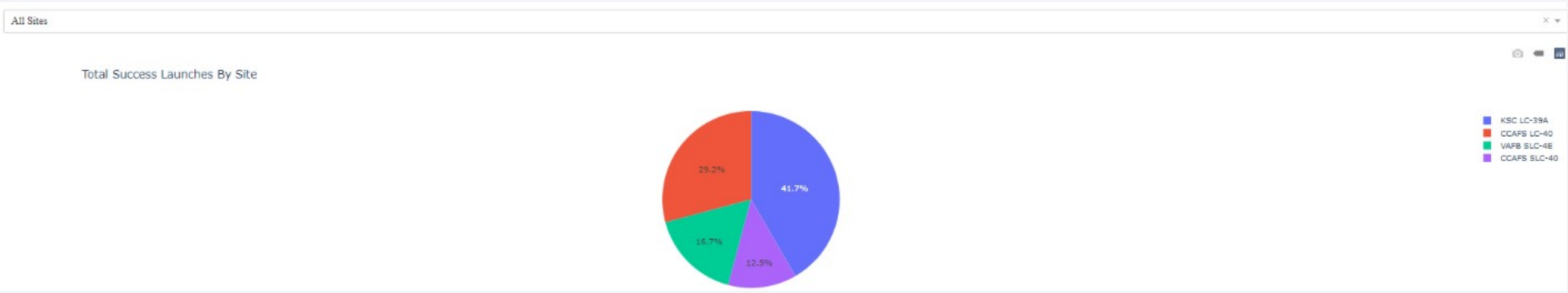


Section 4

Build a Dashboard with Plotly Dash

Launch success count for all sites

- The pie chart represents the share of successful launches for each site.
- KSC LC-39A has the highest number of successful launches
- CCAFS SLC-40 has the lowest number of successful launches



Success ratio for CCAFS LC-40

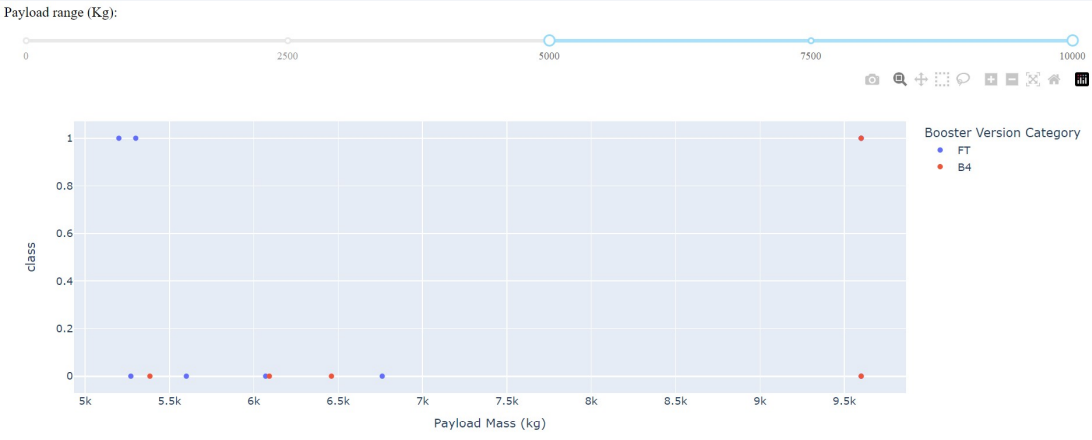
- CCAFS LC-40 is the site with highest success ratio
- 73.1% of its launches are successful

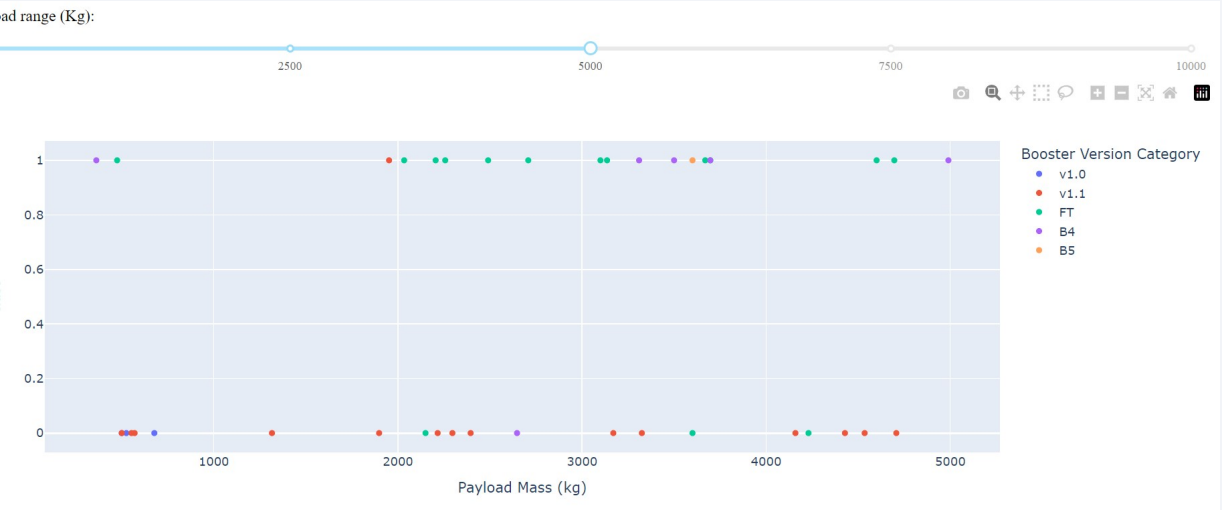
AFS LC-40

Total Success Launches for site CCAFS LC-40

Payload vs. Launch Outcome

The screenshot on the right shows launch outcomes for heavier payloads.





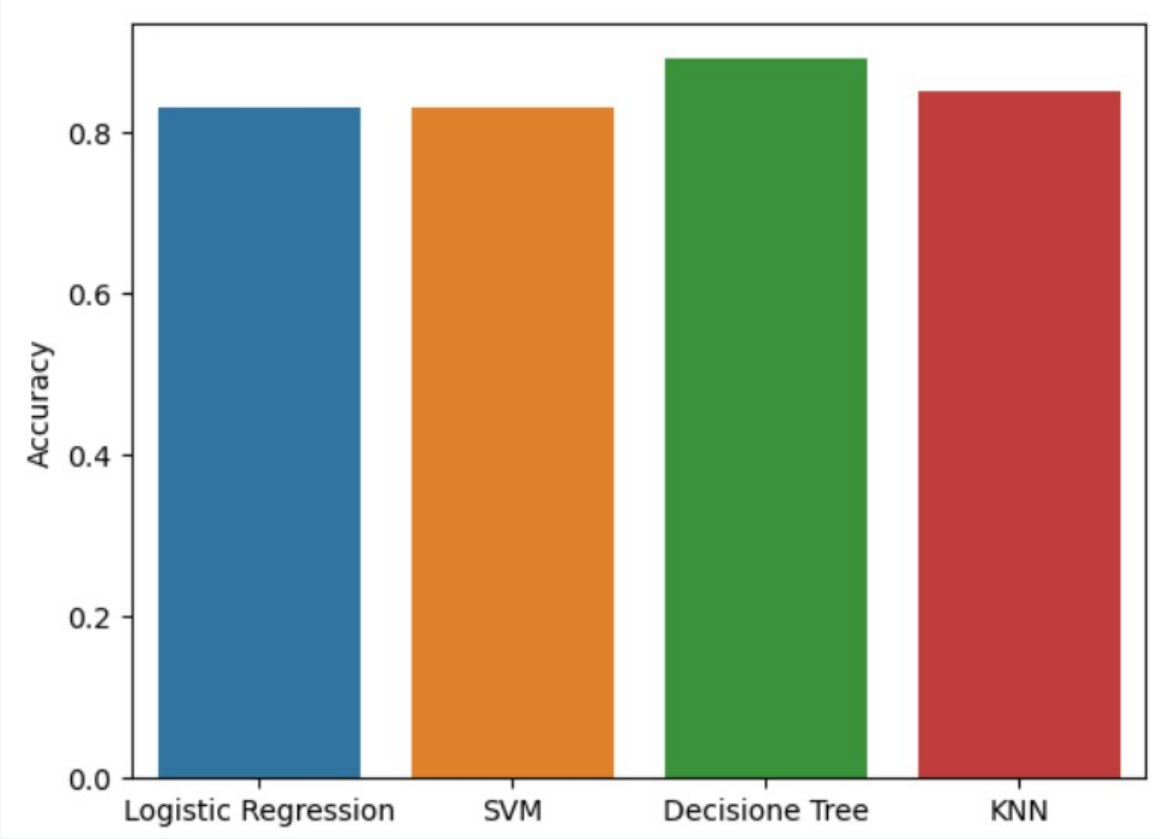
The screenshor on the left shows launch outcom for lighter payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

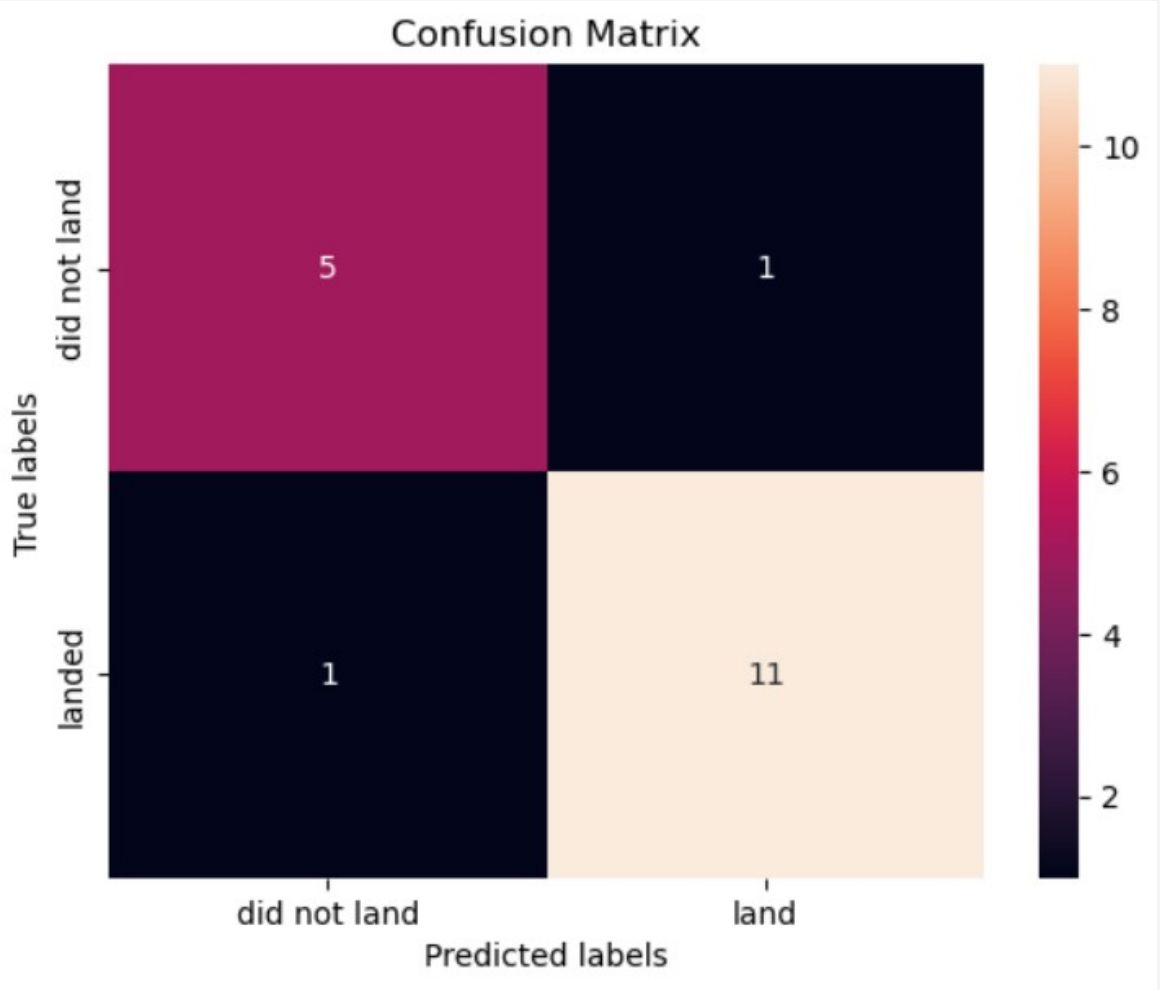
- The Decision Tree has the highest accuracy



Confusion Matrix

This is the confusion matrix for the decision tree classifier.

The values outside the diagonal are small, which indicates that the model generally predicts the outcome correctly



Conclusions

- Launch Site, Flight Number, Payload Mass and Orbit Type all influence the outcome of the launch. Especially the Orbit Type.
- KSC LC-39A is the most successful Launch Site, both in terms of total number of successful launches and success rate
- The general trend of the success rate through the years is increasing
- The decision tree classifier is the best model out of the four tested models for this dataset

Appendix

```
In [7]: # EX
        # landing_outcomes = values on Outcome column
        landing_outcomes = df[['Outcome']].value_counts()
        landing_outcomes
```

```
Out[7]: Outcome
        True ASDS      41
        None None      19
        True RTLS      14
        False ASDS      6
        True Ocean      5
        False Ocean     2
        None ASDS       2
        False RTLS      1
        dtype: int64
```


```
In [7]: # EX
        # landing_outcomes = values on Outcome column
        landing_outcomes = df[['Outcome']].value_counts()
        landing_outcomes
```

```
Out[7]: Outcome
        True ASDS      41
        None None      19
        True RTLS      14
        False ASDS      6
        True Ocean      5
        False Ocean     2
        None ASDS       2
        False RTLS      1
        dtype: int64
```

Thank you!