Winning Space Race with Data Science

Ghanshyam Paunikar
March 06, 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- Problems you want to find answers

  How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing? - Does the rate of successful landings increase over the years? - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

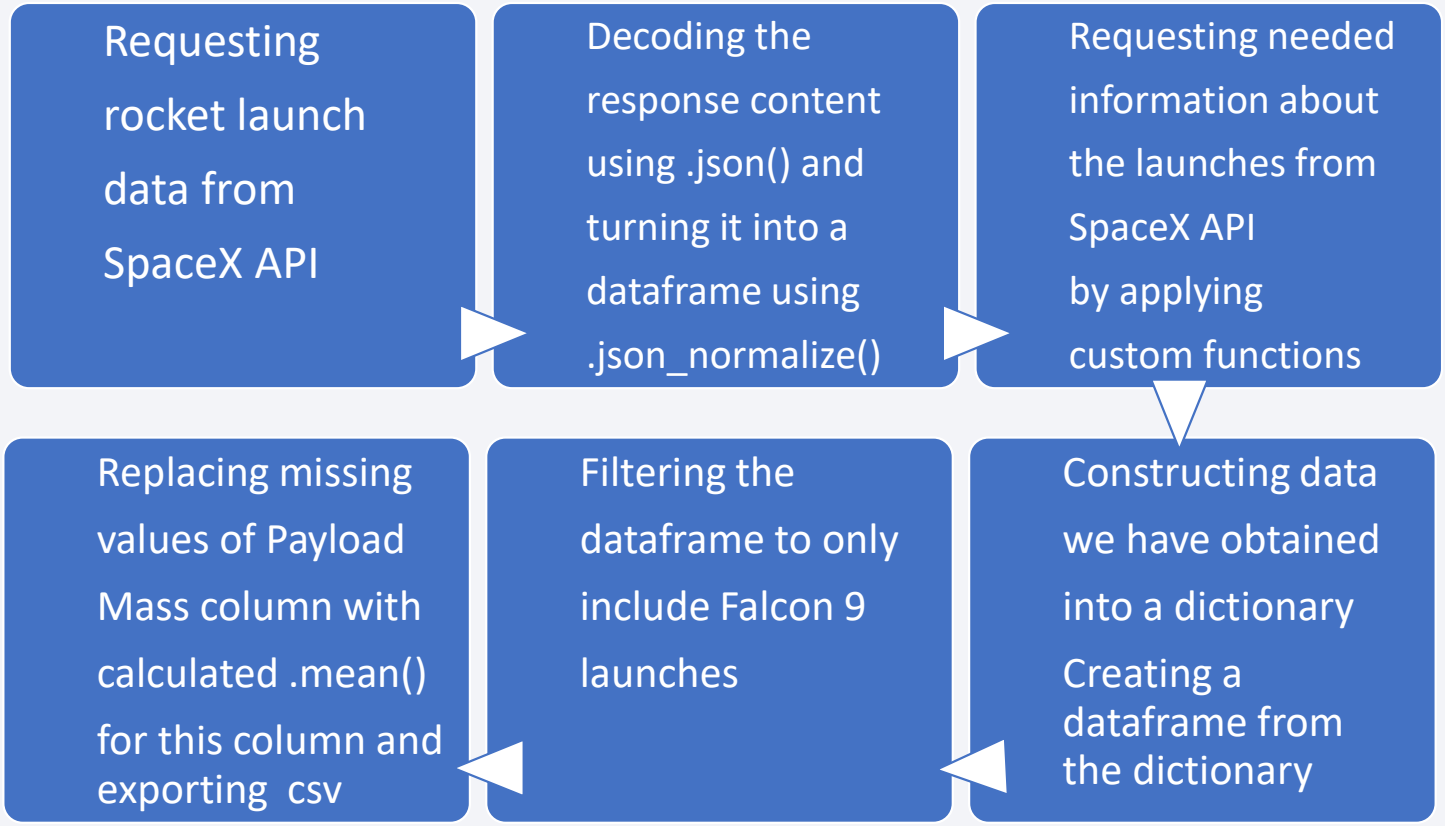
Executive Summary

- Data on the SpaceX Falcon 9 first stage landings was collected from a public API, unaffiliated with SpaceX, and from a Wikipedia article. Additional data sets were provided with the course in CSV file format.
- Datawas wrangled/cleaned in preparation for visualizations, queries, and machine learning model training.
- Exploratory Data Analysis (EDA) was performed using data visualizations and SQL.
- Interactive data visualizations were created using Folium and Plotly Dash.
- Predictive analysis using classification models was done using machine learning models.

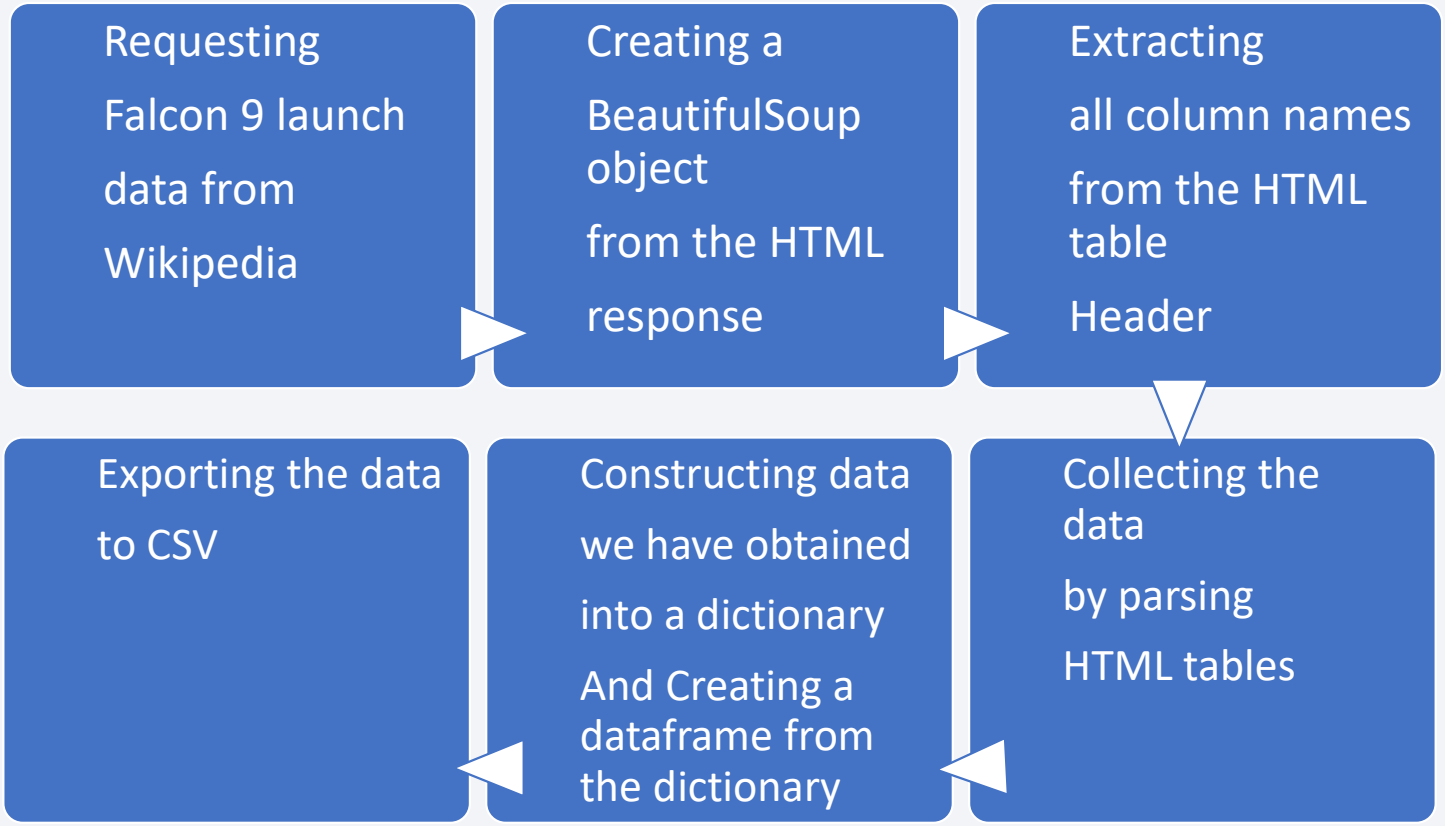
Data Collection

- The data sets were collected from:
- An IBM copy of a response from a publicly accessible API with launch data in JSON format.
- A permanently-linked Wikipedia page with launch data in HTML tables.
- Additional data sets were provided with the course in CSV file format (Highlighted as the darker green CSV files in the top row of the diagram below). See appendix for links.

Data Collection – SpaceX API



Data Collection - Scraping



Data Wrangling

- The CSV file from the first section contained the data in need of cleaning/wrangling.
- The launch sites, orbit types and mission outcomes were processed and reformatted.
- The mission outcome types were converted to a binary classification (one-hot encoding) where 1 represented the Falcon 9 first stage landing being a success and 0 represented a failure.
- The new mission outcome classification column was added to the DataFrame.

EDA with Data Visualization

- Charts were 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).

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

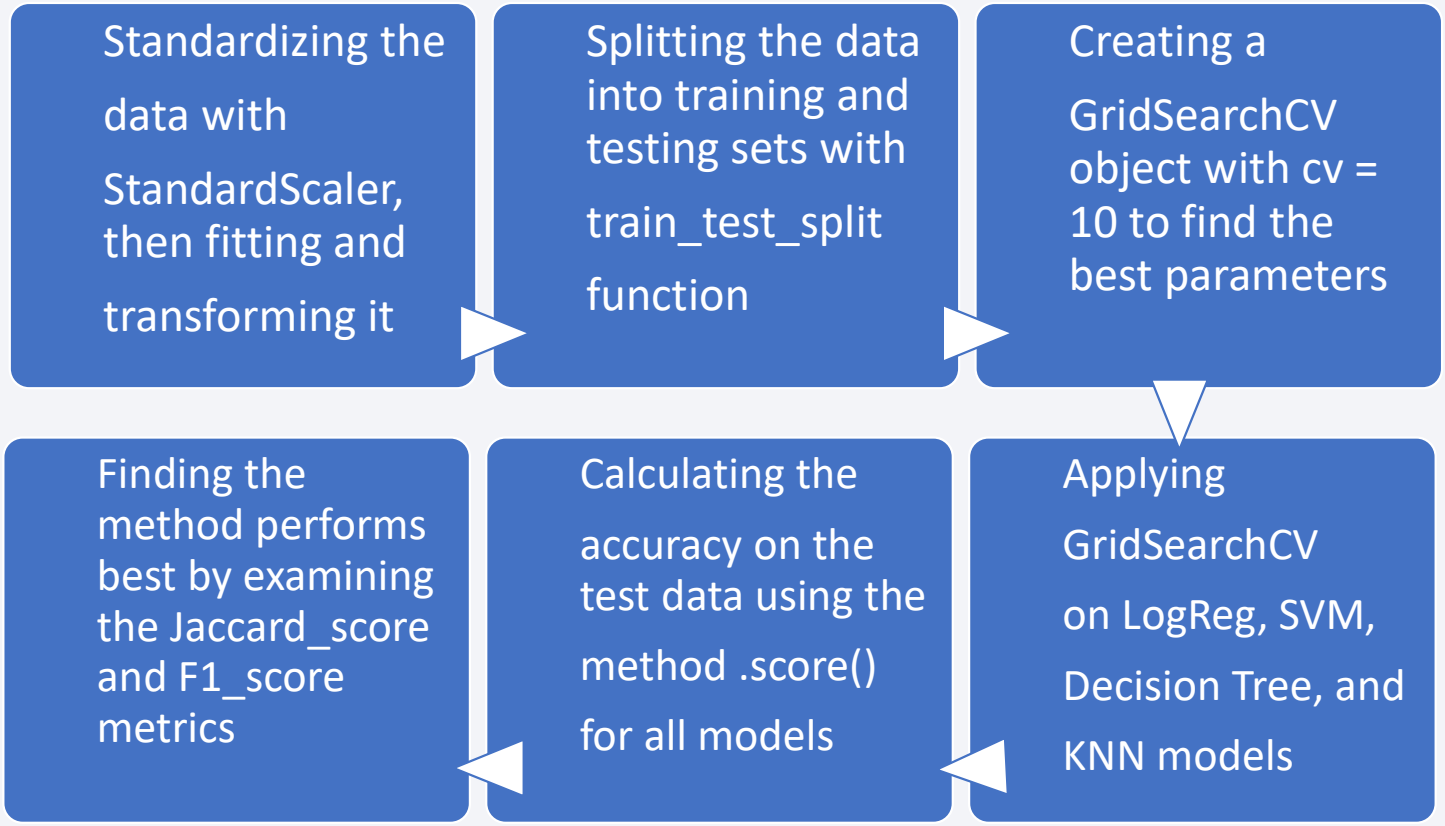
Build an Interactive Map with Folium

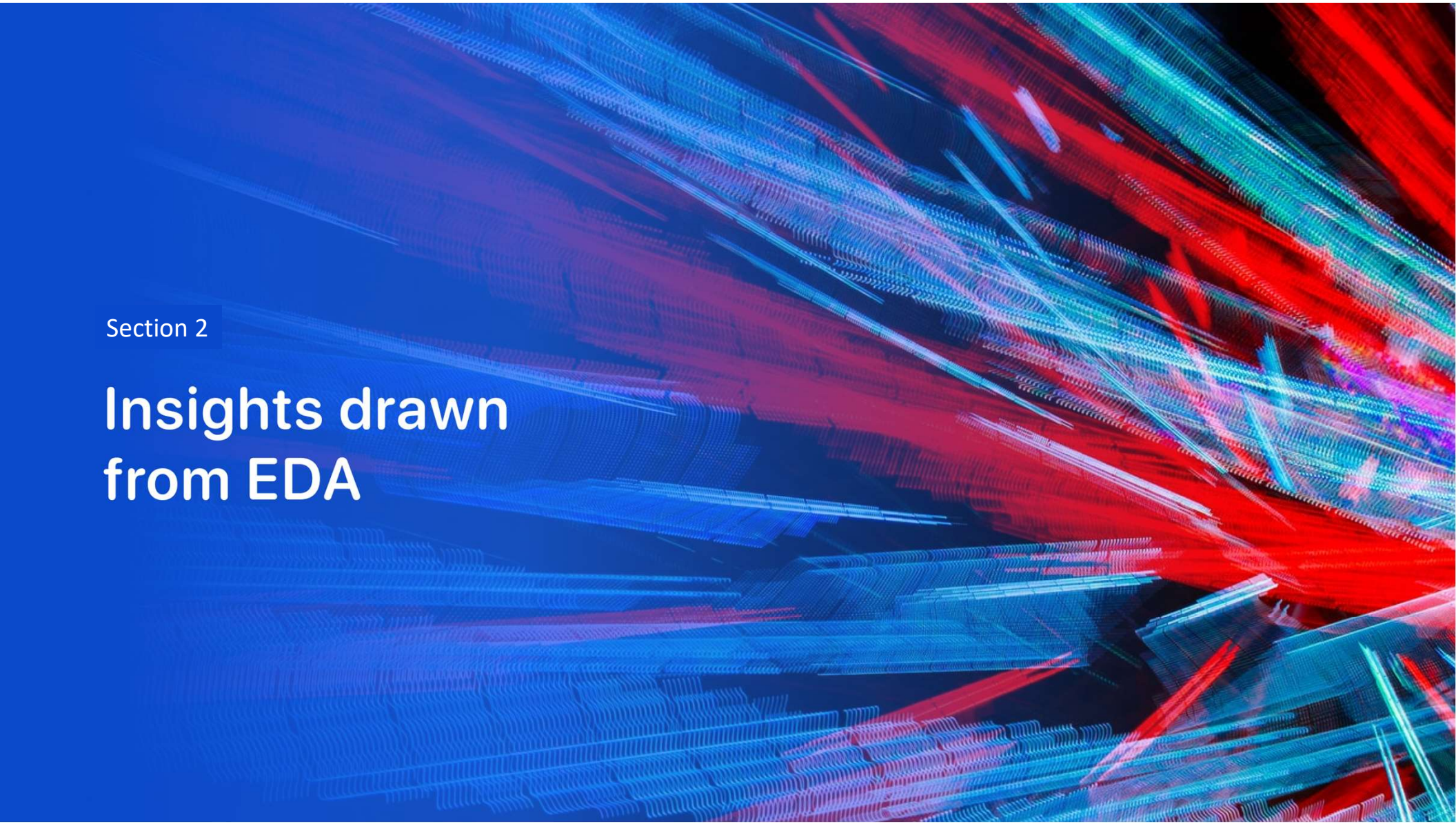
- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:

Build a Dashboard with Plotly Dash

- The Plotly Dash dashboard included a dropdown input to select data from ‘one’ or ‘all’ launch sites to display on the pie chart and scatterplot.
- For ‘one’ launch site, the pie chart displayed the distribution of successful and failed Falcon 9 first stage landings for that site.
- For ‘all’ launch sites, the pie chart displayed the distribution of successful Falcon 9 first stage landings between the sites.
- The input slider is used to filter the payload masses for the scatterplot.
- The scatterplot displayed the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category.

Predictive Analysis (Classification)





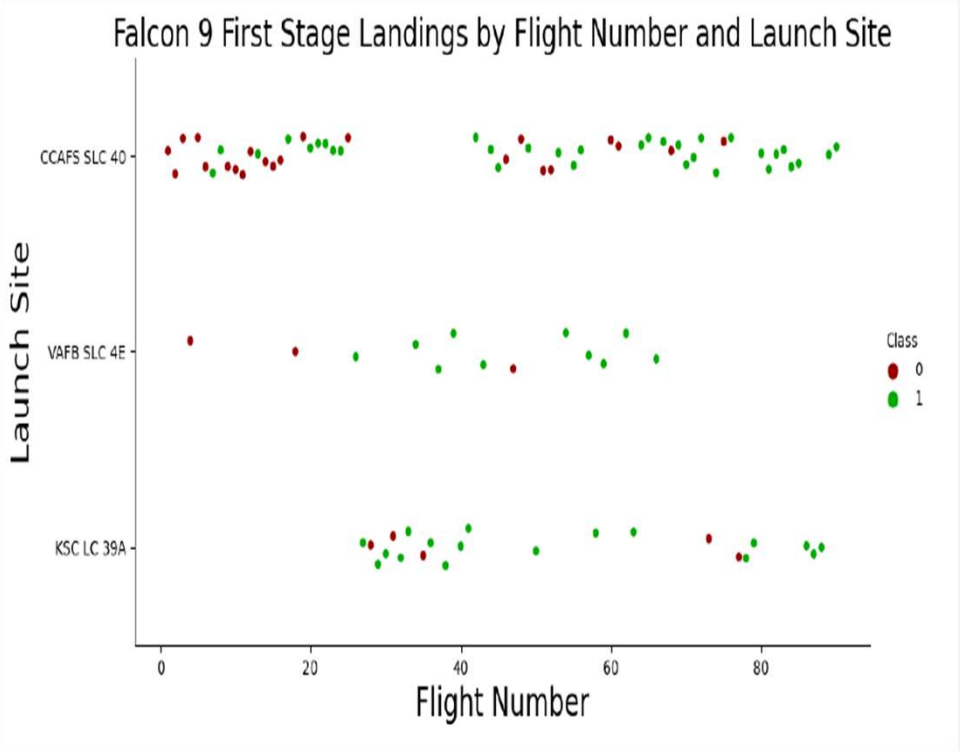
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Falcon 9 first stage failed landingsare indicated by the ‘0’ Class (●red markers) and successful landings by the ‘1’ Class (●green markers).

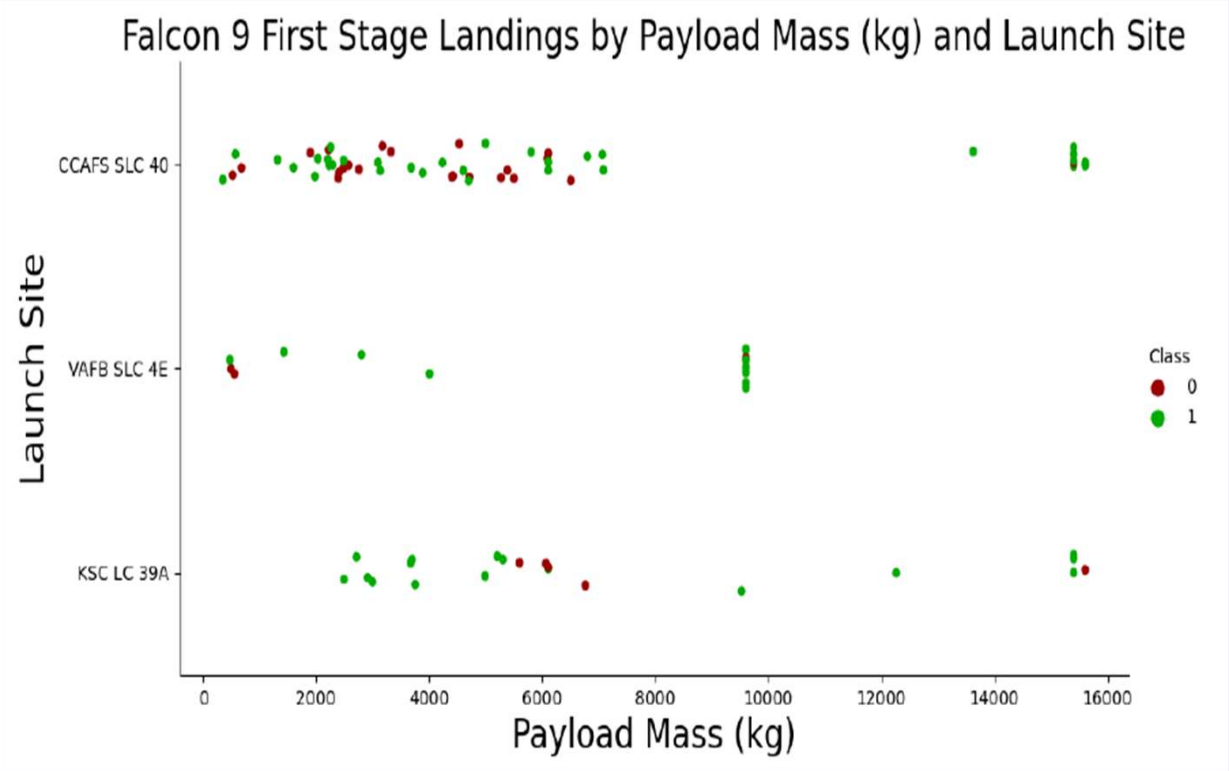
- Successful Falcon 9 first stage landings appear to become more prevalent as the flight number increases.



Payload vs. Launch Site

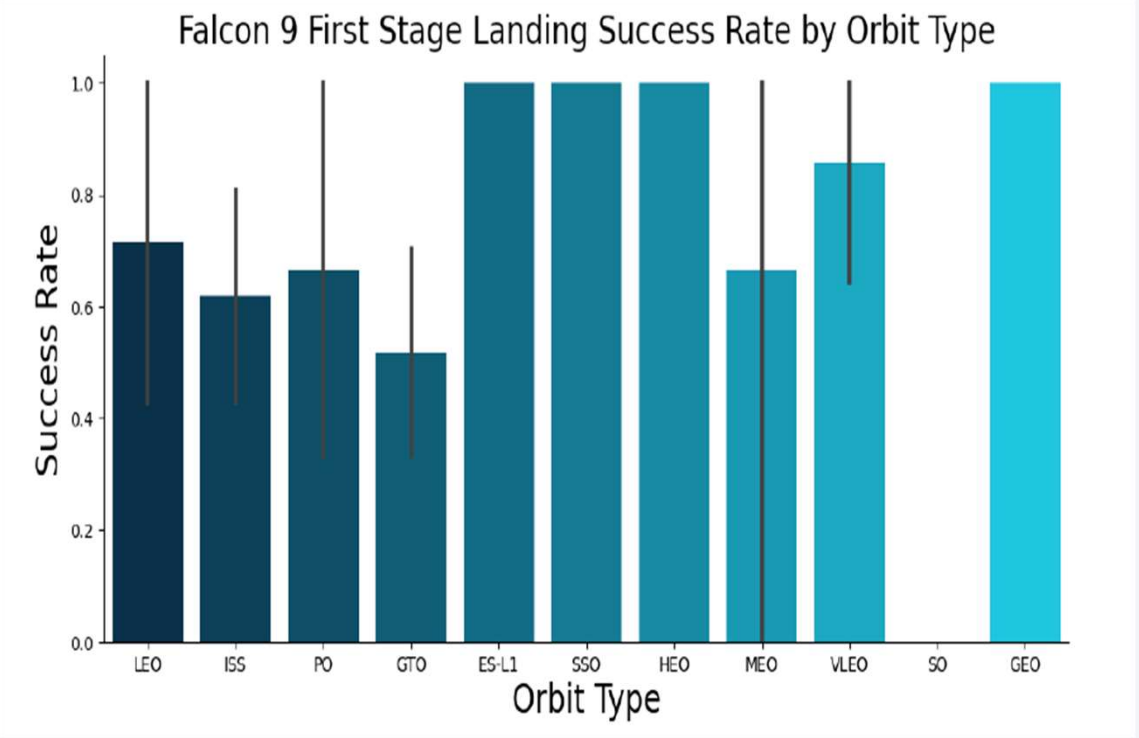
- Falcon 9 first stage failed landingsare indicated by the ‘0’ Class (●red markers) and successful landings by the ‘1’ Class (●green markers).

- For the CCAFS SLC 40 launch site, the payload mass and the landing outcome appear to not be strongly correlated.

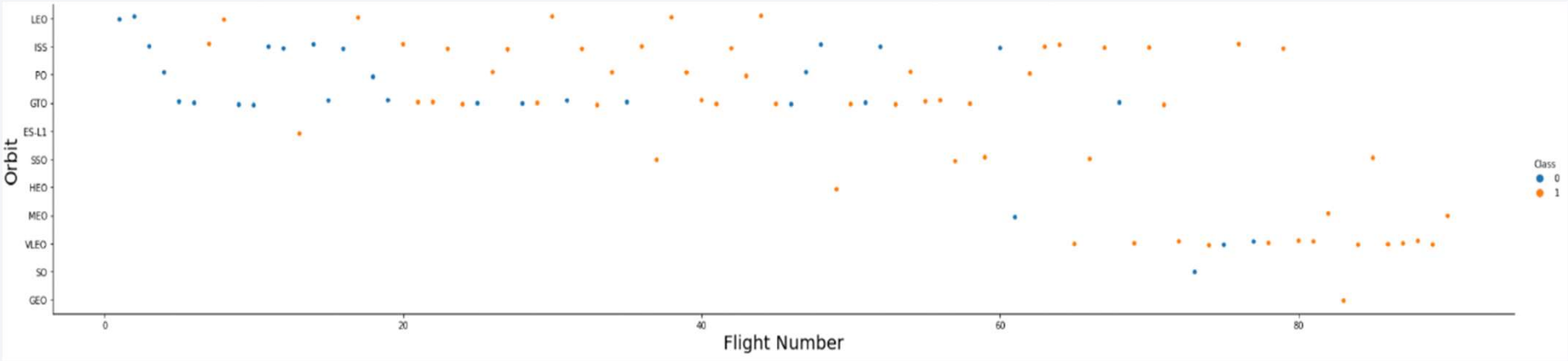


Success Rate vs. Orbit Type

- ES-L1, SSO, HEO and GEO orbits have no failed first stage landings.
- SO orbits have no successful first stage landings.

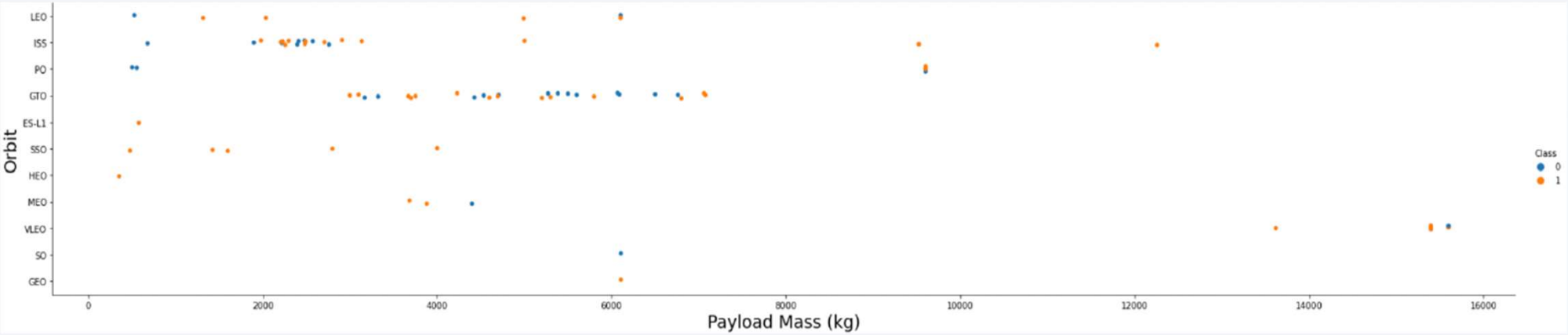


Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

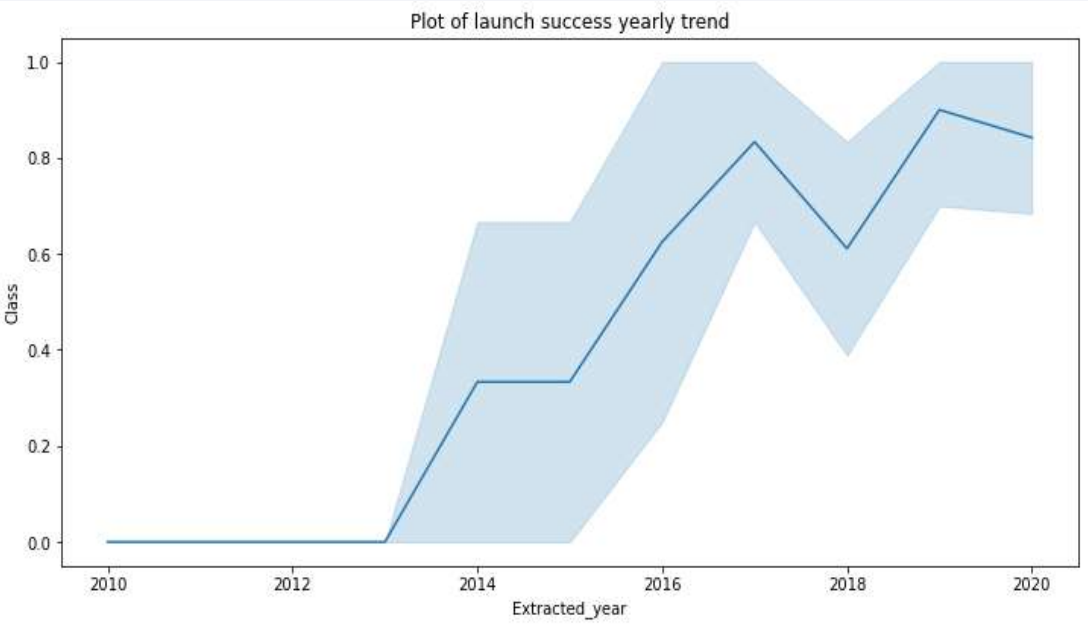
Payload vs. Orbit Type



- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020.



All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

- Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

- Finding 5 records where launch sites begin with `CCA`

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Calculating the total payload carried by boosters from NASA

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

First Successful Ground Landing Date

- Finding the dates of the first successful landing outcome on ground pad

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
```

Out[9]:

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

Total Number of Successful and Failure Mission Outcomes

- Listing the total number of successful and failure mission outcomes.

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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

- We used a combinations of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
         where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- 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

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
             where date between '2010-06-04' and '2017-03-20'
             group by landing__outcome
             order by count_outcomes desc;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

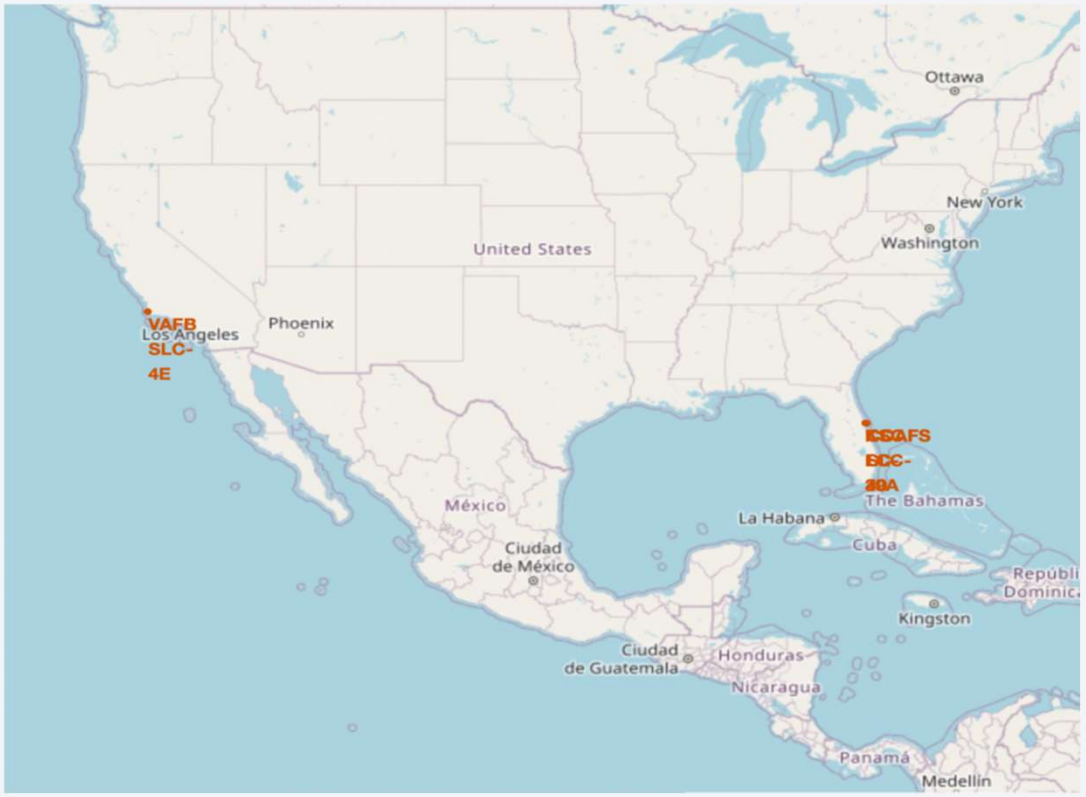
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

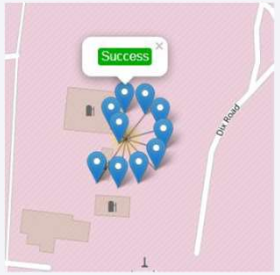
Falcon 9 Launch Site Locations

Most of Launch sites are in proximity to the Equator line.
The land is moving faster at the equator than any other place on the surface of the Earth.
Anything on the surface of the Earth at the equator is already moving at 1670 km/hour.
If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching.
This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
All launch sites are in very close proximity to the coast; while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

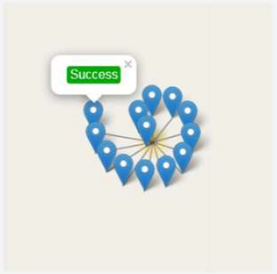


Map Markers of Success/Failed Landings

VAFB SLC-4E

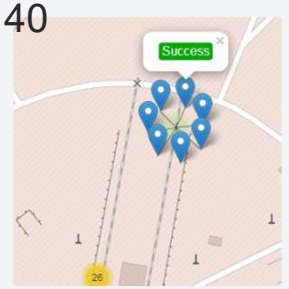


KSC LC-39A



CCAFS LC-40

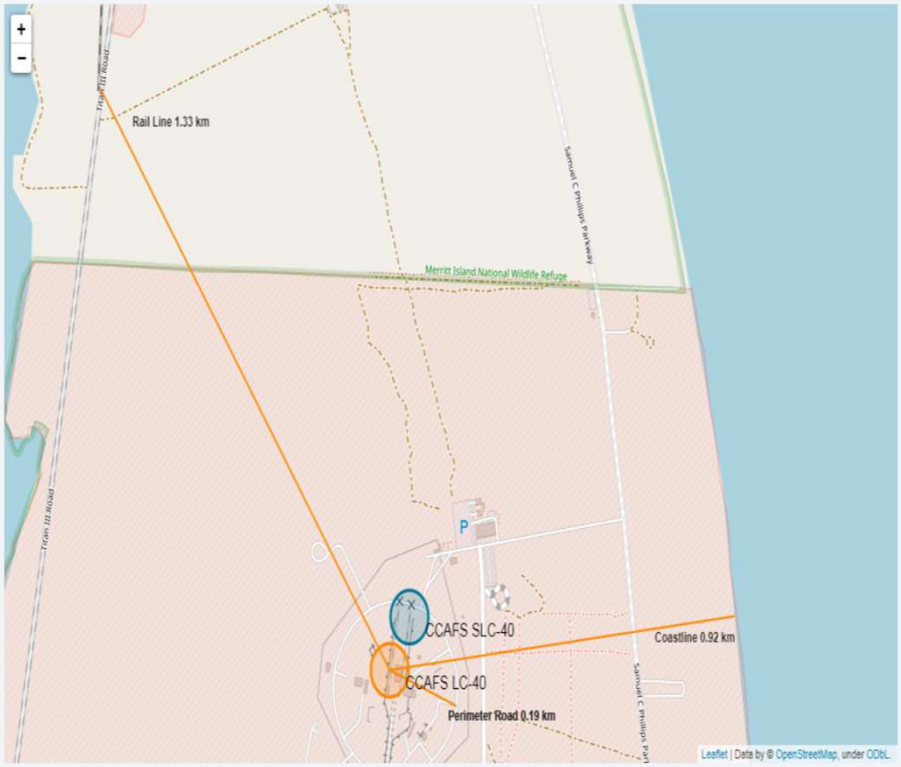
CCAFS SLC-40

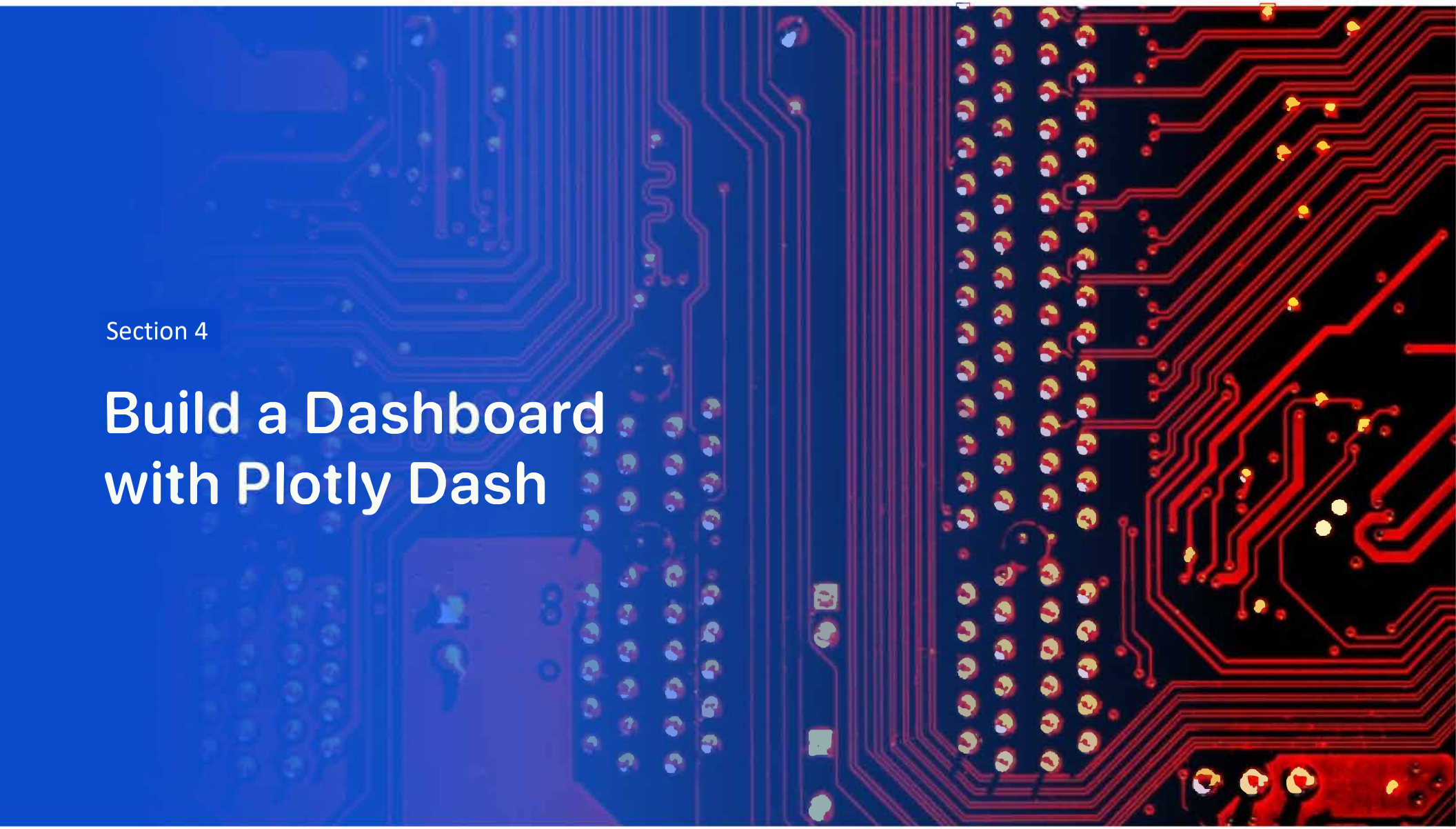


- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.
- A sense of a launch site's success rate for Falcon 9 first stage landings can be gleaned from the relative number of green success markers to red failure markers.

Distance from Launch Site to Proximities

- The CCAFS LC-40 and CCAFS SLC-40 launch sites have coordinates that are close to being, but are not exactly, right on top of each other.
- The perimeter road around CCAFS LC-40 is 0.19 km away from the launch site coordinates.
- The coastline is 0.92 km away from CCAFS LC-40.
- The rail line is 1.33 km away from CCAFS LC-40.



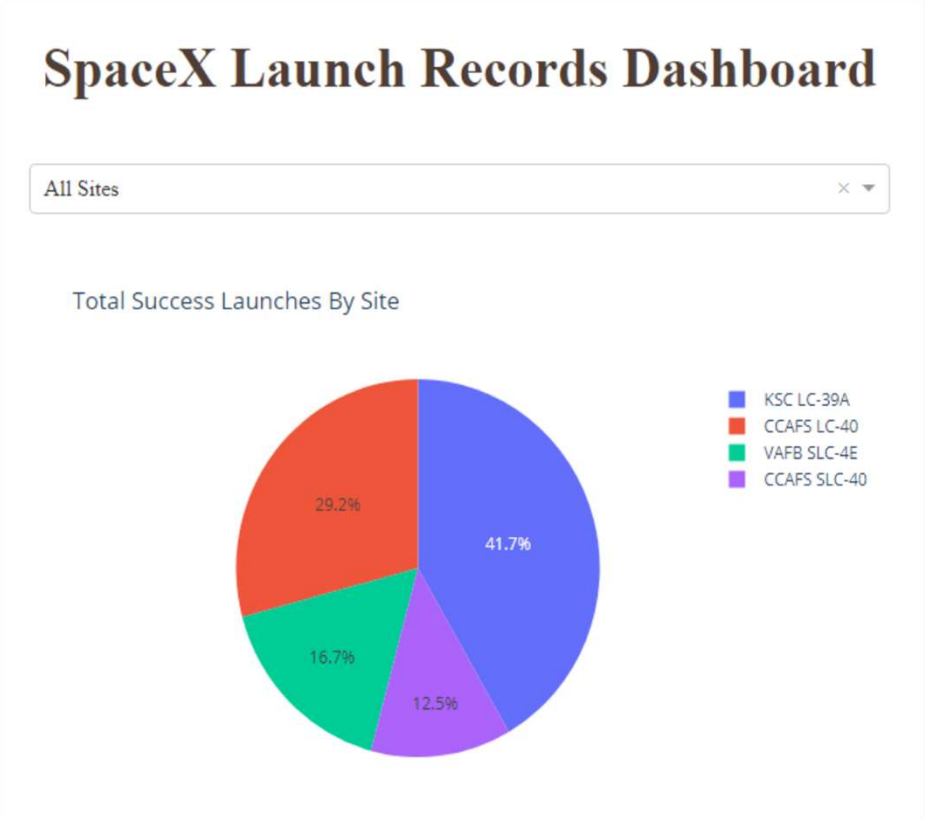


Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

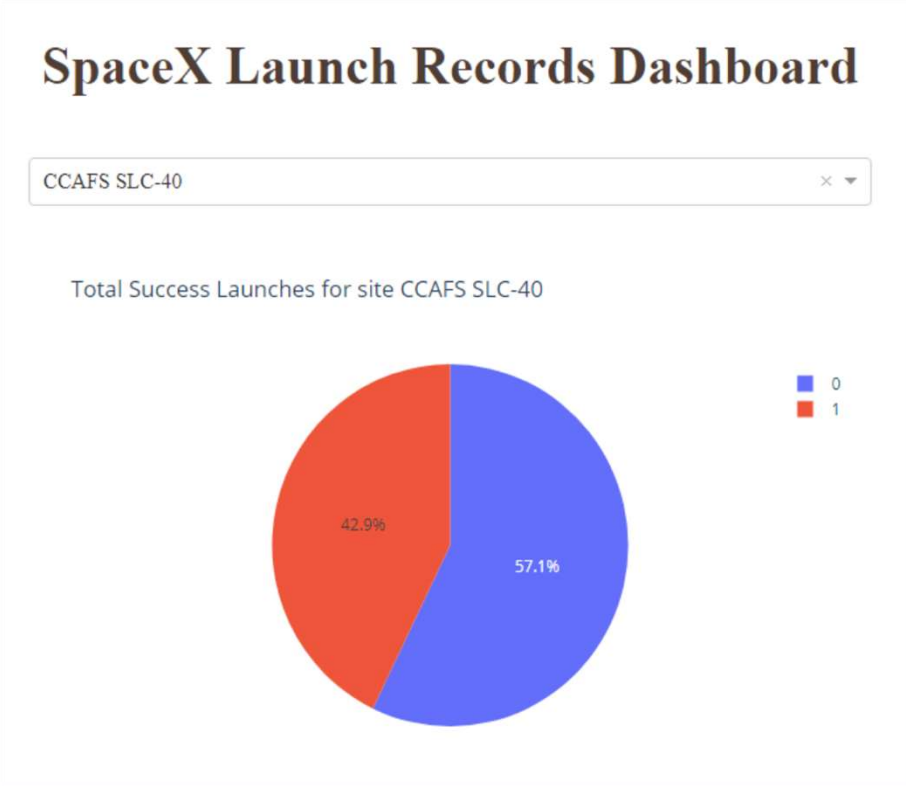
- The dropdown menu allowed the selection of one or all launch sites.
- With all launch sites selected, the pie chart displayed the distribution of successful Falcon 9 first stage landing outcomes between the different launch sites.
- The greatest share of successful Falcon 9 first stage landing outcomes (at 41.7% of the total) occurred at KSC LC-39A.



Launch Site with Highest Launch Success Ratio

- Falcon 9 first stage failed landings are indicated by the ‘0’ Class (■blue wedge in the pie chart) and successful landings by the ‘1’ Class (■red wedge in the pie chart).

- CCAFS SLC-40 was the launch site that had the highest Falcon 9 first stage landing success rate (42.9%).

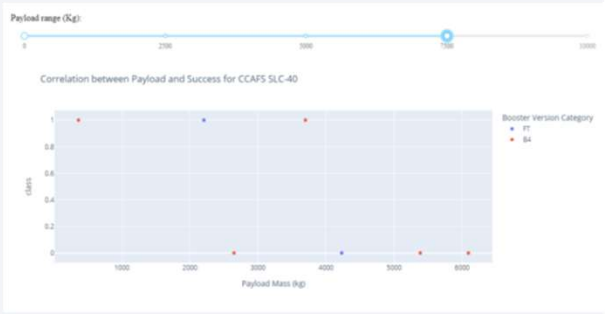


Payload vs. Launch Outcome

- These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider.
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.
- The ‘FT’ booster version category has the largest success rate.

CCAFS LC-40

CCAFS SLC-40



KSC LC-39A

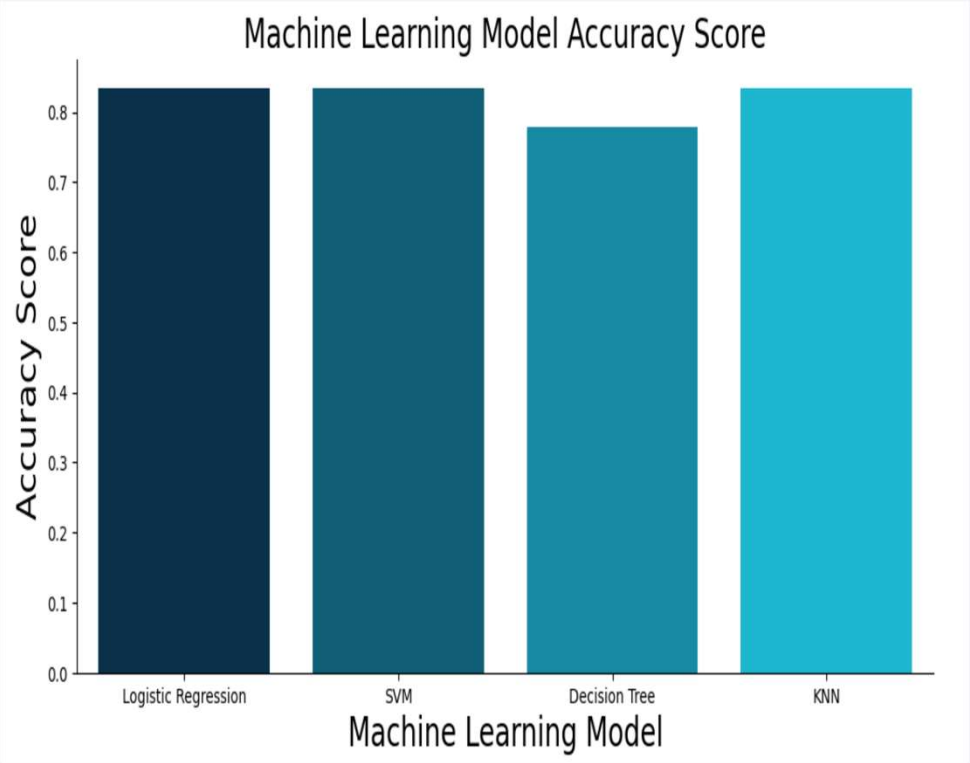
VAFB SLC-4E

Section 5

Predictive Analysis (Classification)

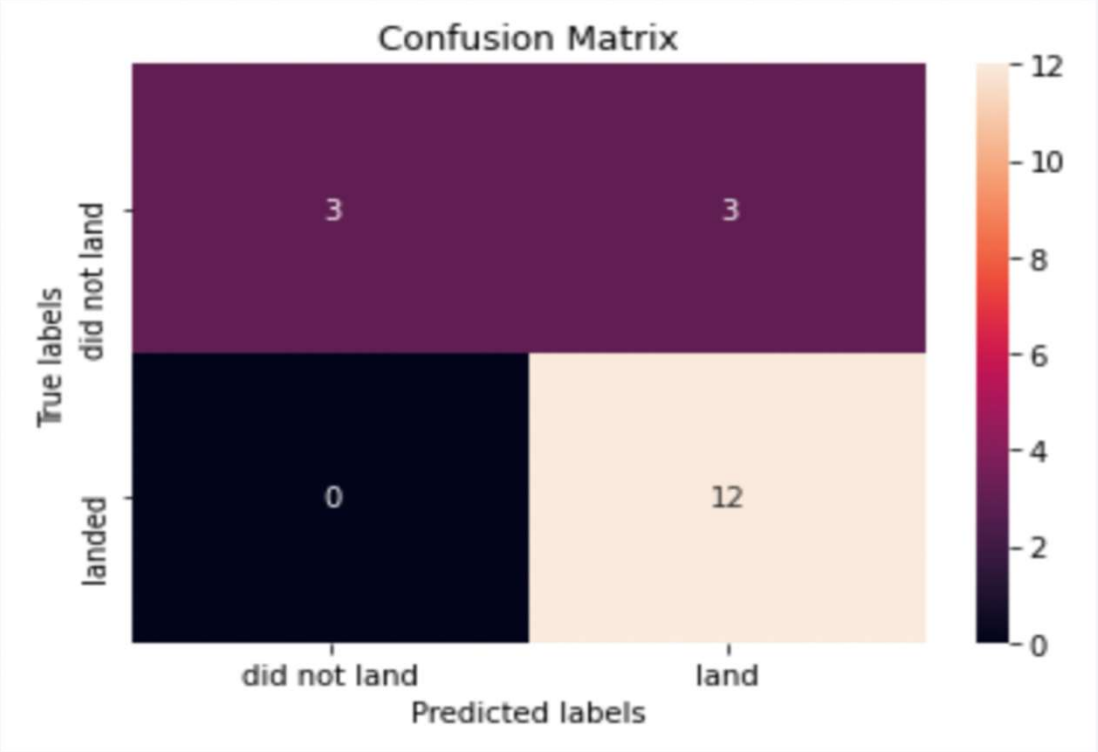
Classification Accuracy

- All models performed equally well except for the Decision Tree model which performed poorly relative to the other models.



Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



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.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Thank you!