IBM Developer
SKILLS NETWORK

Winning Space Race With Data Science

Alberto Arancet
25-07-2025

Outline

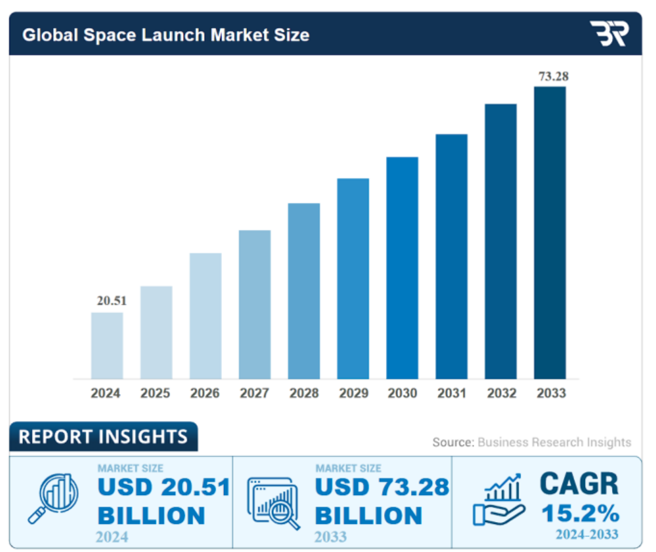
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- SpaceX have being at the frontier in the space business thanks to the revolutionizing methodology of reusable rockets (in this case boosters and launchers).
- The recovery of their rockets had made them a very profitable since their reduce their cost in a big way, of course the recovery of each rocket depend on different factors, in this report we can get a prediction of the model of a 94,4%, depending of course of the payload, and orbit

Introduction – Project Background and context

- Space, specially the business related has being growing, the global space launch market as being on the rise in a very fast pace, this growth is being lead by the increase interest of the private sector in the areas of satellite launchers, space exploration and advance in technologies like propulsion



[Space Launch Market Size, Growth | Report [2025-2033]]

[Rocket Propulsion Market Size, Share & Forecast Report - 2032]

Introduction – Problem we want to tag

- SpaceX had being very successful in the business of space, specially in frontier of the idea of recovery or reusable rockets. Falcon 9 rocket launches with a cost of 62 million dollars when the first stage of their rockets can be reused.
- The success recovery of the booster is one of the key factor for the profitability of SpaceX, this is the most important cost reduction in their line.
- This report is being made with the mission of predict the success or like hood of the first stage rocket landing for use recovery.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected.
  - The data was collected from different sources, in this case open source Space X Rest API, Web Scraping Falcon9 launch data from Wikipedia

Data Collection – SpaceX API

1. Import libraries and define auxiliary functions
2. Request the data from SPaceX api
3. Request and parse the SpaceX launch data using the GET request
4. Filter the dataframe to only include Falcon 9 launches
5. Use Data Wrangling to deal with Missing Values
6. Export to a **CSV**

[capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · AlbertoArancetMerino/capstone]

Data Collection - Scraping

Extract a Falcon 9 launch records HTML table from Wikipedia and parse the table and convert it into a Pandas data frame

Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Export to a **CSV**

[capstone/jupyter-labs-webscraping.ipynb at main · AlbertoArancetMerino/capstone]

Data Wrangling

Explored data to determine the label for training supervised models

Calculated the number of launches on each site

Calculated the number and occurrence of each orbit

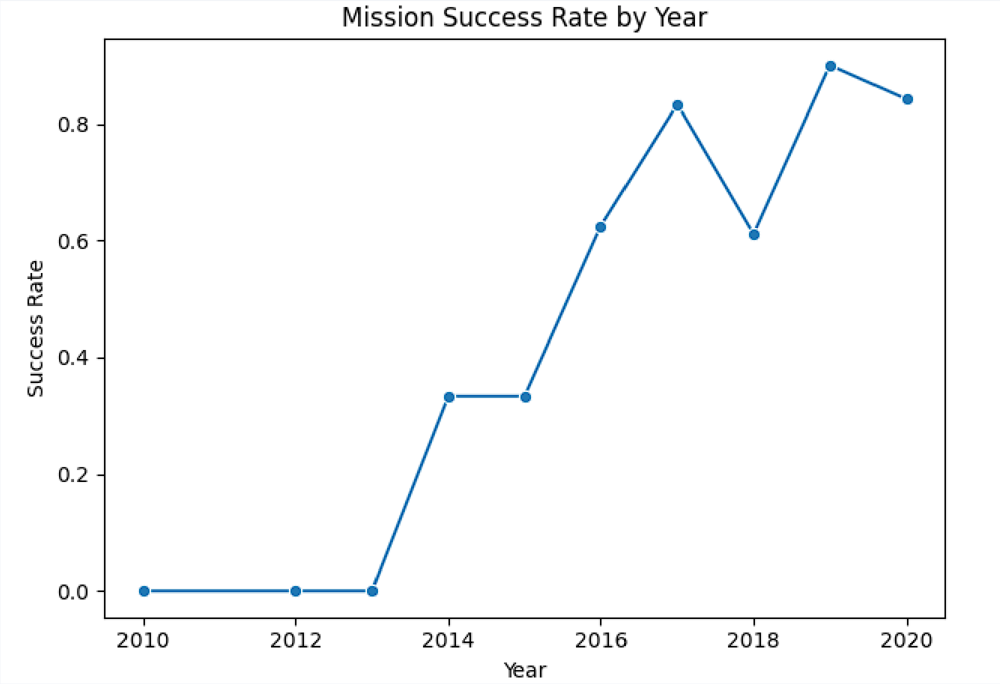
Training label: 'Class'

Calculated the number and occurrence of mission outcome per orbit type

Created a landing outcome training label from 'Outcome' column

capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · AlbertoArancetMerino/capstone

EDA with Data Visualization



Loaded data into an IBM DB2 Instance

Run diferente SQL queries to extract information about:

- **Diferent Launch sites**
- **Payload masses**
- **Mission outcomes**
- **Booster landings**

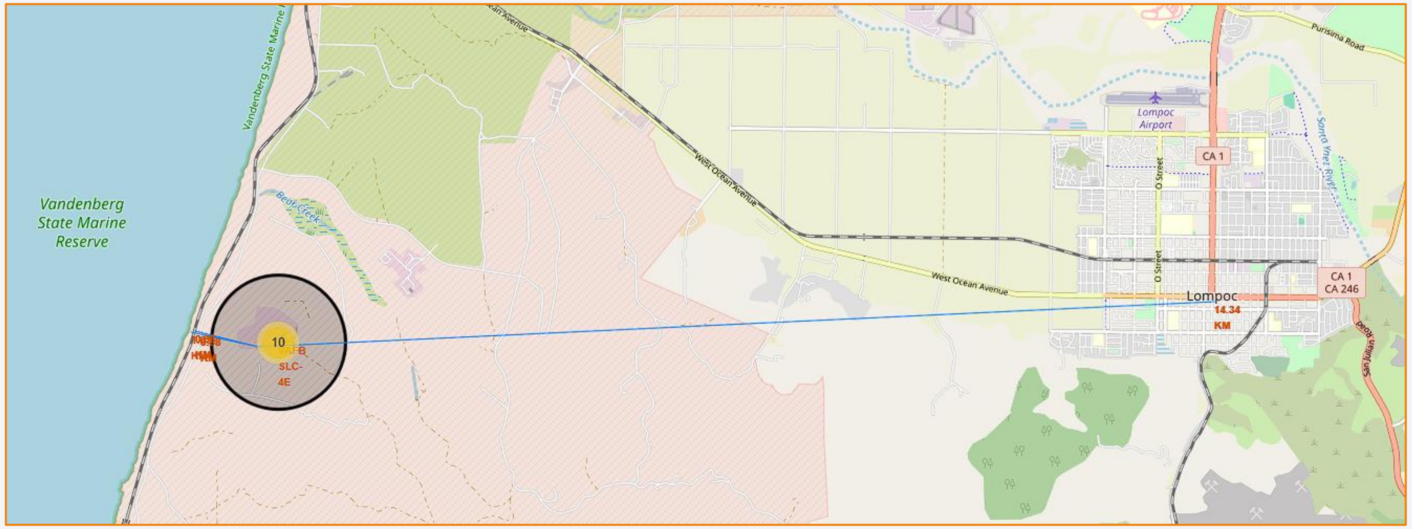
[capstone/edadataviz.ipynb at main · AlbertoArancetMerino/capstone]

EDA with SQL

- We run SQL queries to get and list the different information about
- Launch sites
- Payload masses
- Booster versions
- Mission outcomes
- Failure outcomes
- Booster landins
- Landings outcoms

capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · AlbertoArancetMerino/capstone

Build an Interactive Map with Folium



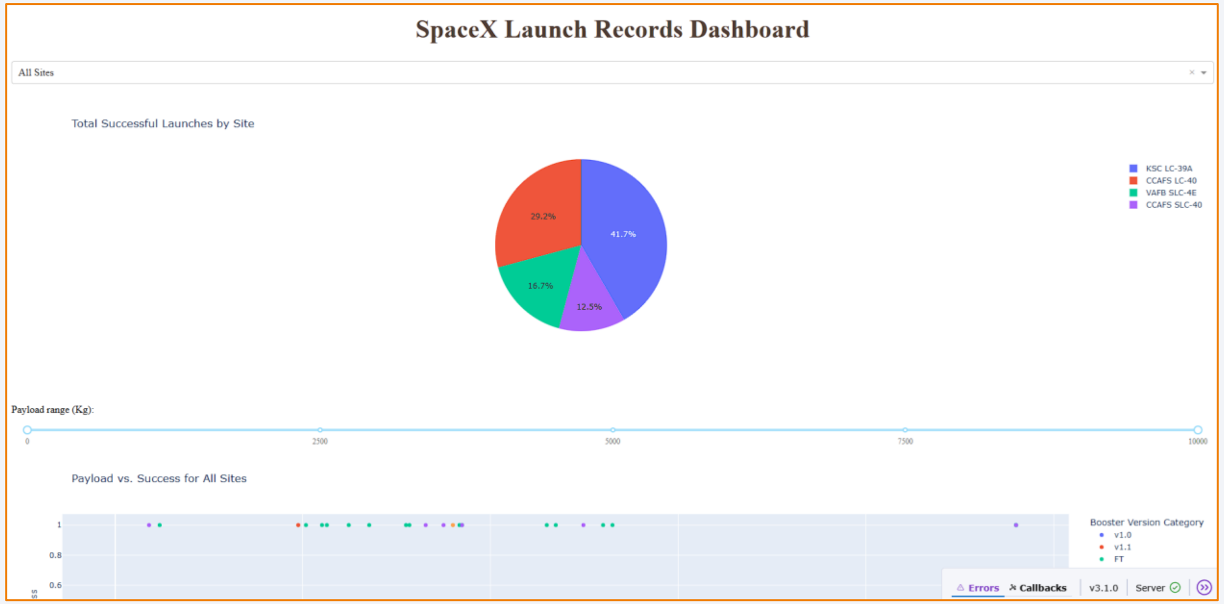
Screenshot of interactive Folium map showing proximity from VAFB-SLC 4E launch site to nearby railway, highway, and coastline

Data Visualization:

- Launch Sites Location Analysis
- Used Python interactive mapping library called Folium
- Marked all launch sites on a map
- Marked the successful/failed launches for each site on map
- Mark Railways / Highways / Coastlines close to the launch site
- Calculated the distances between a launch site to its proximities Cities

capstone/lab_jupyter_launch_site_location.ipynb at main · AlbertoArancetMerino/capstone

Build a Dashboard with Plotly Dash

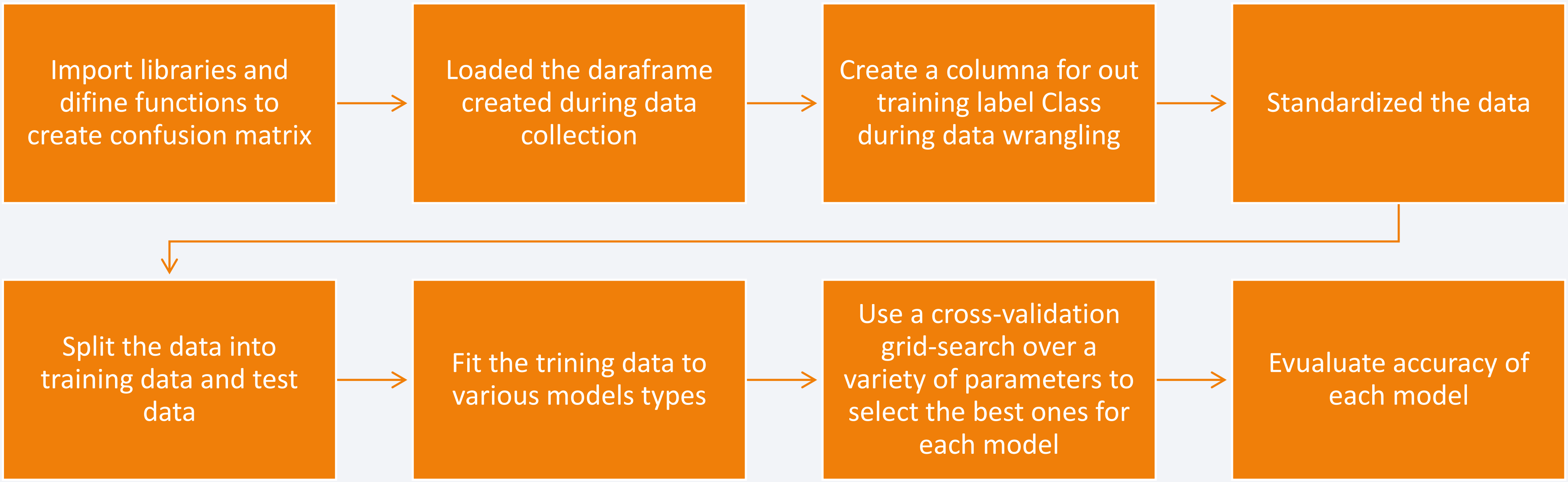


The dashboard was made with Plotly this include a :

- Dropdown menu for selecting launch sites Pie charts displaying success rate.
- Scatter chart displaying launch site, payload mass, success/failure Range with a slider for selecting range of payload mass (kg)

capstone/spacex-dash-app.py at main · AlbertoArancetMerino/capstone

Predictive Analysis (Classification)



capstone/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · AlbertoArancetMerino/capstone

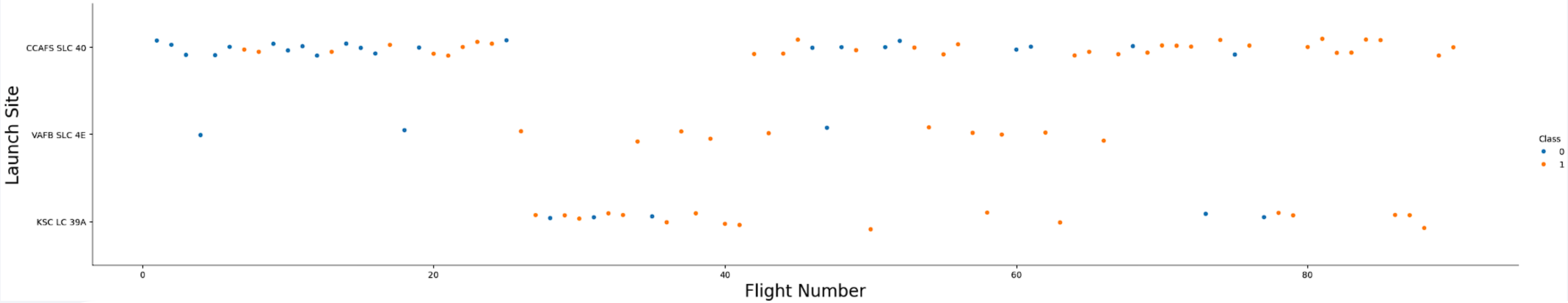
Results

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

Section 2

Insight drawn From EDA

Flight Number vs. Launch Site



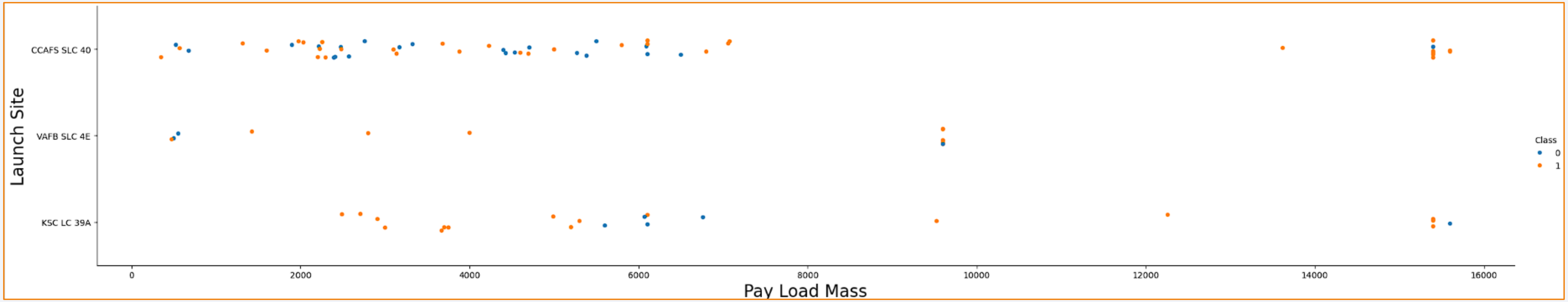
Show a scatter plot of Flight Number vs. Launch Site

- This chart shows the values of Launch Sites, Flight Numbers and the success/failure of each launch

For what we can gather:

- CCAFS-SLC have the highest launch and also is the site with the more failures

Payload vs. Launch Site



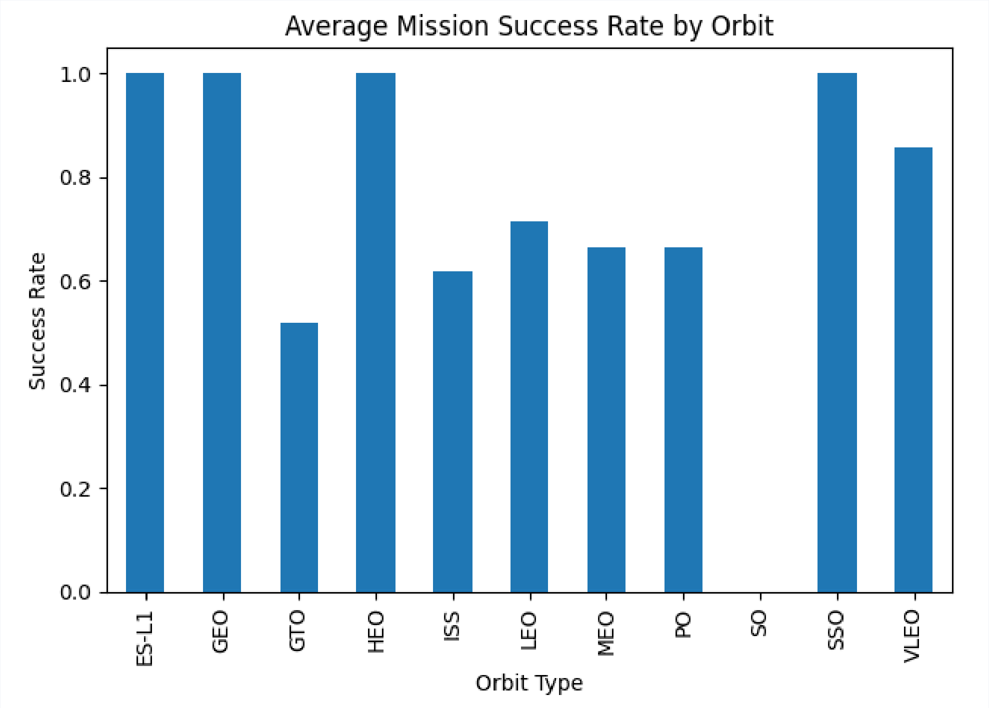
Show a scatter plot of Payload vs. Launch Site

- This chart shows the payload mass per launch site and the success/failure of each launch

For what we can gather:

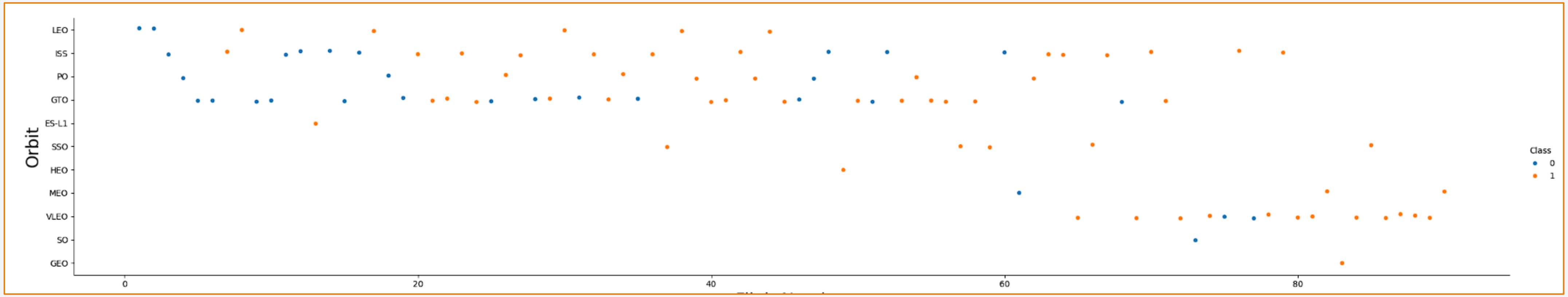
- The heavy payload have less chance of failure

Success Rate vs. Orbit Type



- We can see that GTO have the lowest success rate and SSO the highest
- The success rate is very dependent on the payload mass and orbit

Flight Number vs. Orbit Type



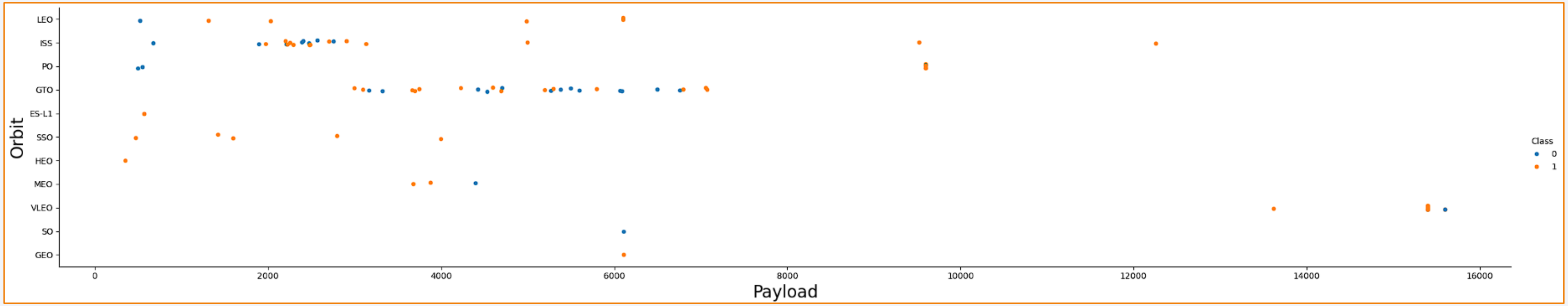
Flight Number vs. Orbit Type

- This chart info is about orbit vs flight number, it shows the success rate vs flight, orbit.

For what we can gather:

Lower orbits are a lower risk

Payload vs. Orbit Type



Payload vs. Orbit Type

- This chart info is about orbit vs playload, it shows the success rate vs payload, orbit.

For what we can gather:

Independently of the diferente payload mass, GTO is a very risky orbit affecting the diferente missions success rate, even when the falcon 9 reliability improves there still failed booster recovery

Launch Success Yearly Trend

Launch Succsess Yearly trend

- In this chart we can se how the Falcon 9 reliability increase or improves over time.

For what we can gather:

- Success rate and recovery depend on the payload mass and orbit
- The falcon 9 average recovery rate is of 66%

All Launch Site Names

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

Query result

In [20]:
```
%sql SELECT DISTINCT Launch_Site From SPACEXTABLE
```

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

Out[20]:

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

Launch Site Names Begin with 'CCA'

- Launches with CCA

In [23]:
```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTABLE
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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

Out[23]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
In [24]:  %%sql
          SELECT SUM(PAYLOAD_MASS__KG_)
          FROM SPACEXTABLE
          WHERE Customer = 'NASA (CRS)';

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

Out[24]:

SUM(PAYLOAD_MASS__KG_)
45596

We use a SQL quary to to calculate the total payload mass in KG, of all the records in the table SPACEX where the Customer is “NASA (NRS)”

- “Select Sum(PAYLOAD_MASS__KG_)” is the total sum of the payload carried by the rocket
- “FROM SPAXTABLE” this speciefies the table where the data is stored
- “WHERE Customer = ‘NASA (NRS)´” is the aplied filter so it can only sum a specific set in the case the NASA (NRS)

Average Payload Mass by F9 v1.1

```
In [13]: %%sql
         SELECT AVG(PAYLOAD_MASS__KG_)
         FROM SPACEXTBL
         WHERE Booster_Version LIKE 'F9 v1.1%';

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

Out[13]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

We use a SQL quary to to calculate the avarage payload mass in KG from the table, where only include records of the Boster version F9 V1.1

- From this we can see that the payload mass is 2534,7 Kg

First Successful Ground Landing Date

In [31]:

```
%%sql
SELECT MIN(Date)
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Out[31]:

MIN(Date)
2015-12-22

We use a SQL query to find the first or realiest launch date from the SPACEX table where the rocket had a successful landing on a graund pad

- “SELECT MIN(Date)” take the minimum or earliest value from the date comun
- “FROM SPACEXTABLE” use the data from the SPACEX Table
- “WHERE Landing_Outcome = 'Success (ground pad)'” this is the filter , for those records where the landing was successful and hapend on a ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

Out[33]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

In [33]:

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

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

We use a SQL query to find the booster Version for a launcher that landed successfully in a drone ship and carried a payload masss between 4000 kg and 6000 kg

- “SELECT Booster_Version FROM SPACEXTABLE” List the booster types from the table SPACEX
- “WHERE Landing_Outcome = 'Success (drone ship)’” filter the success landing on adrone ship
- “AND 4000 < PAYLOAD_MASS__KG_ < 6000” make another filter por the payload range

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and fai
- Present your query result with a short expla

```
In [35]:  %%sql
          SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
          FROM SPACEXTABLE
          GROUP BY MISSION_OUTCOME;

           * sqlite:///my_data1.db
          Done.
Out[35]:
```

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

We use a SQL query to find and count the total number of successful and failure misión outcomes

- “SELECT MISSION_OUTCOME” this retrieves the mission result
- “COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER” this caunt how many times the outcome hapend"
- “FROM SPACEXTABLE GROUP” this take the data from the SPACEX table
- “BY MISSION_OUTCOME” this group the different records by outcome type

Boosters Carried Maximum Payload

Out[37]:

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

In [37]:

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

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

We use a SQL query to find the diferent/unique booster that carried the maximum payload

- "SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL " this give us the unique or distinct booster types from the SPACEX table
- "WHERE PAYLOAD_MASS__KG_ =" this filter the payload mass equals the maximum mass found in the table
- (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)" an then this select the maximum payload

2015 Launch Records

```
In [22]: %%sql
PRAGMA table_info(SPACEXTBL);
SELECT Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date, 1, 4) = '2015';

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

Out[22]:

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

We use a SQL query to launch fails in the year 2015

- PRAGMA table_info(SPACEXTBL); this give us list of column details from the
- SELECT Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL this give us how it landed (Landing_Outcome), the rocket version (Booster_Version), and where it launched (Launch_Site)
- WHERE Landing_Outcome = 'Failure (drone ship)’ this Filters only those missions that had a failed landing on a drone ship
- AND substr(Date, 1, 4) = '2015’; This take data for launches in 2015

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

In [24]:

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Total_Number
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Total_Number DESC;
```

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

Out[24]:

Landing_Outcome	Total_Number
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

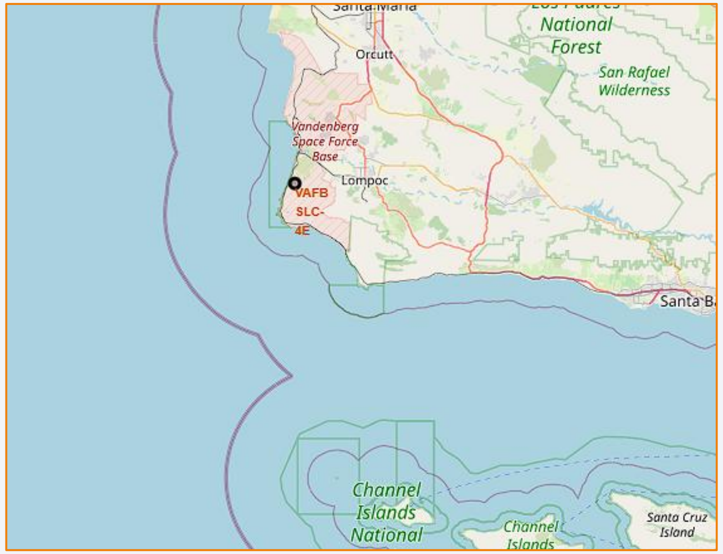
We use a SQL query to count the total landin outcomes

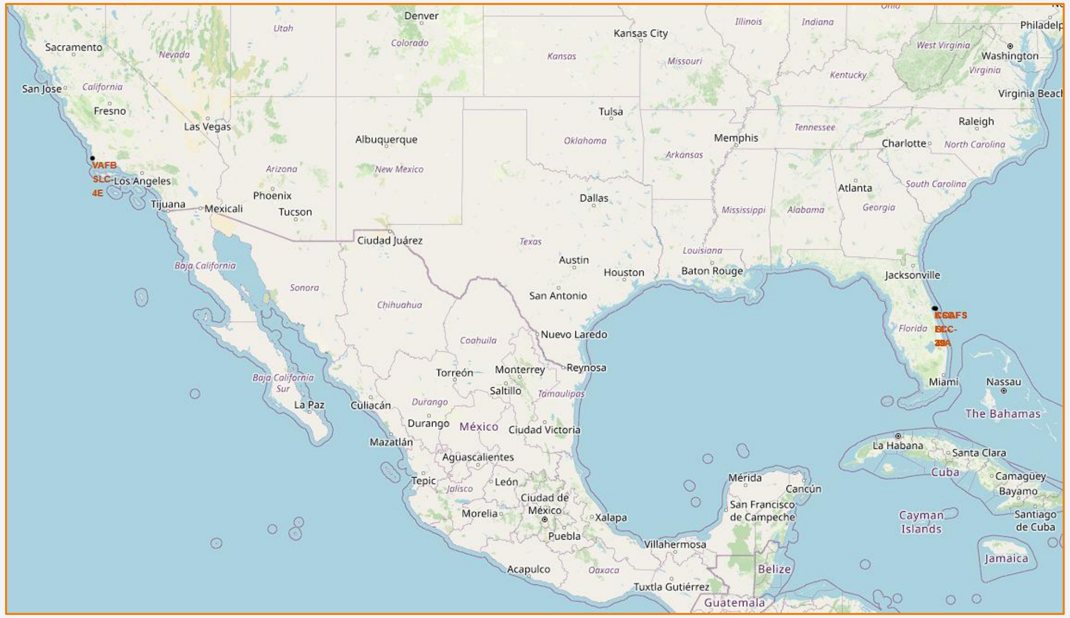
- “SELECT Landing_Outcome COUNT(Landing_Outcome) AS Total_Number FROM SPACEXTBL” this take each landing result and times it occurred.
- “WHERE Date BETWEEN '2010-06-04' AND '2017-03-20’” this Filters missions only within that time window.
- GROUP BY Landing_Outcome ORDER BY Total_Number DESC; Groups by each distinct landing result and sort it

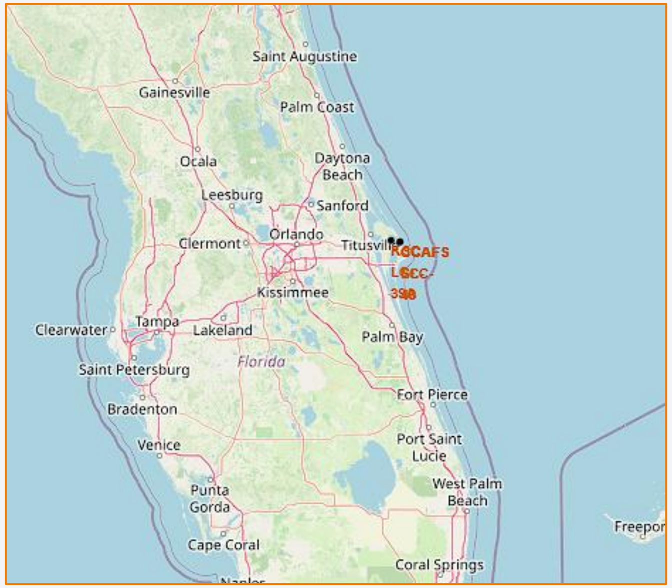
Section 3

Launch Site Proximities Analysis

Launch location analysis







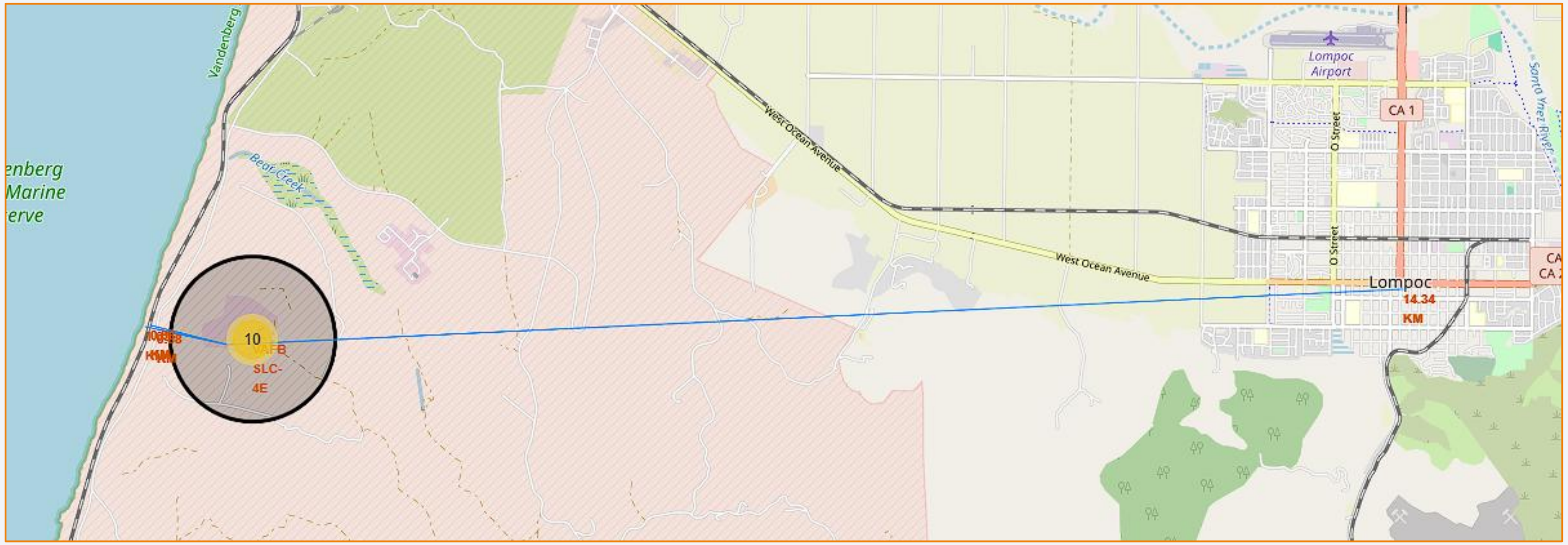
Launch sites locations

VAFB SLC-4E, KSC-LC29A, CCAFS-LC4O,CCAF-SLC40

VAFB-SLC 4E launch site

Visualizing the booster landing outcomes for each launch site highlights which launch sites have relatively high success rate this case the selected launch site , being VAFB-SLC 4E, have a very low success rate

Distance and landmarks

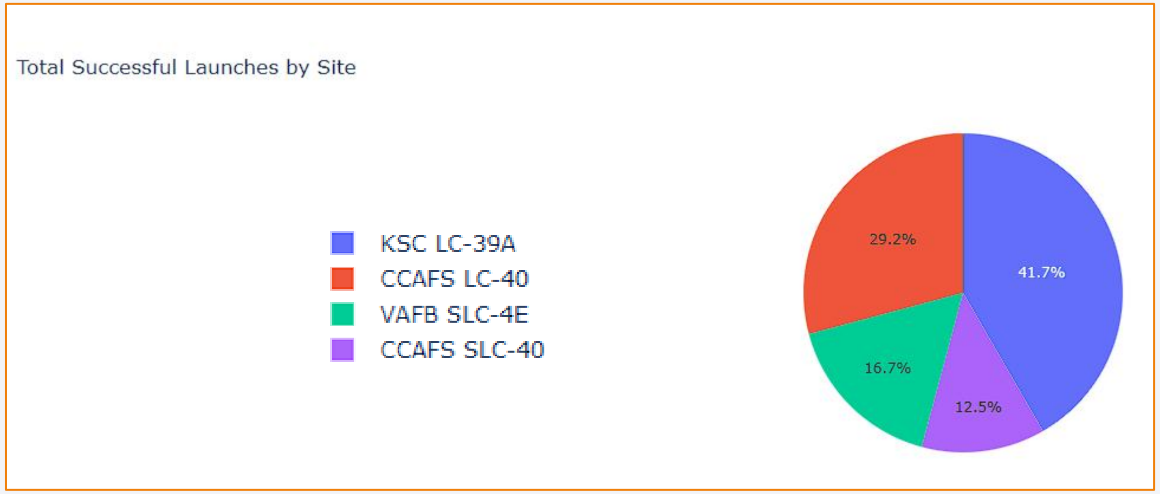


Distance city = 14.339346934328596 km
Distance railroad = 0.976072106423253 km
Distance highway = 0.8771731735070982 km

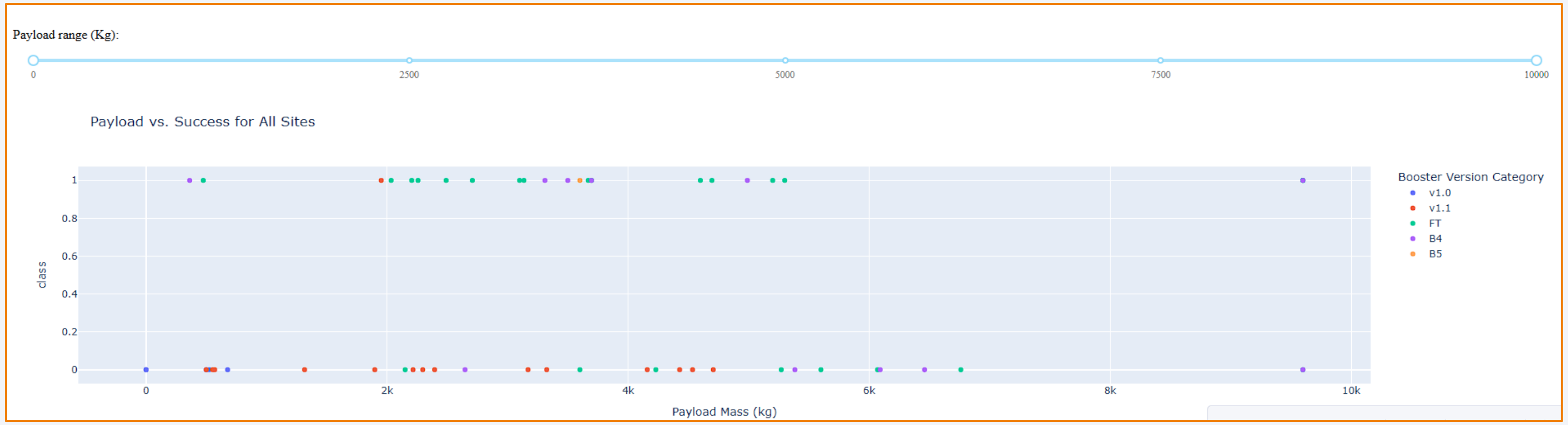
Section 4

Build a Dashboard With Plotly Dash

Dashboard

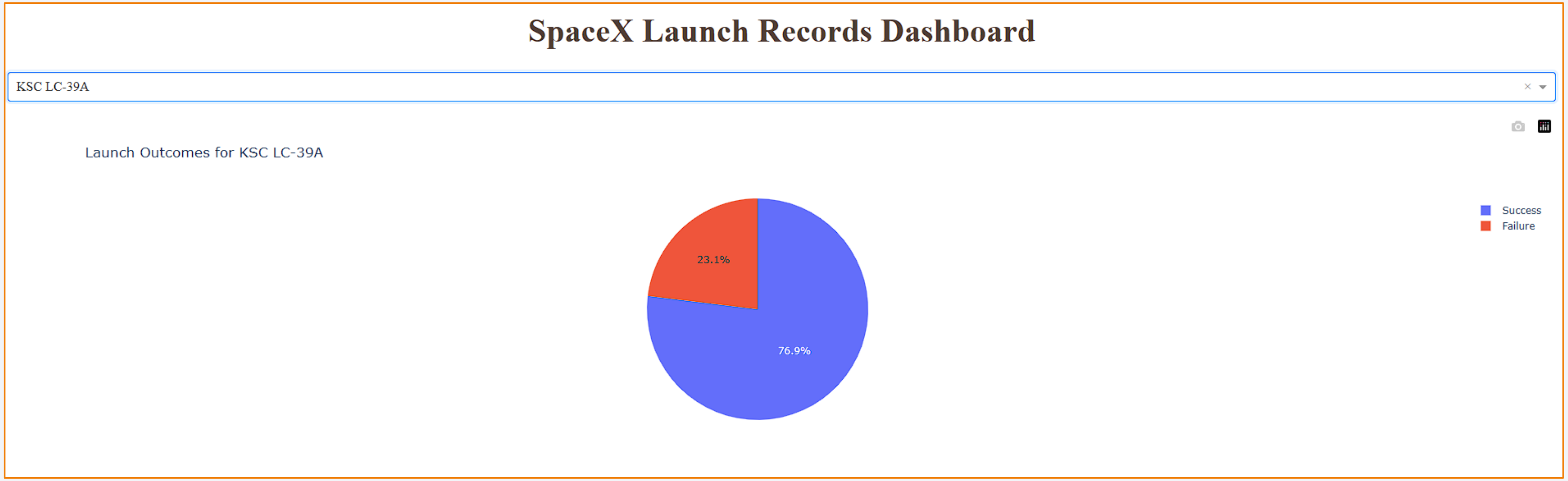


Launch Site	class	
CCAFS LC-40	0	19
	1	7
CCAFS SLC-40	0	4
	1	3
KSC LC-39A	0	3
	1	10
VAFB SLC-4E	0	6
	1	4



Theis dashboard allaws interactive visualization and análisis of the Falcon 9 successful launches, it is made of an pie chart where we can see the diferente launchsites and the successfull launches of each site

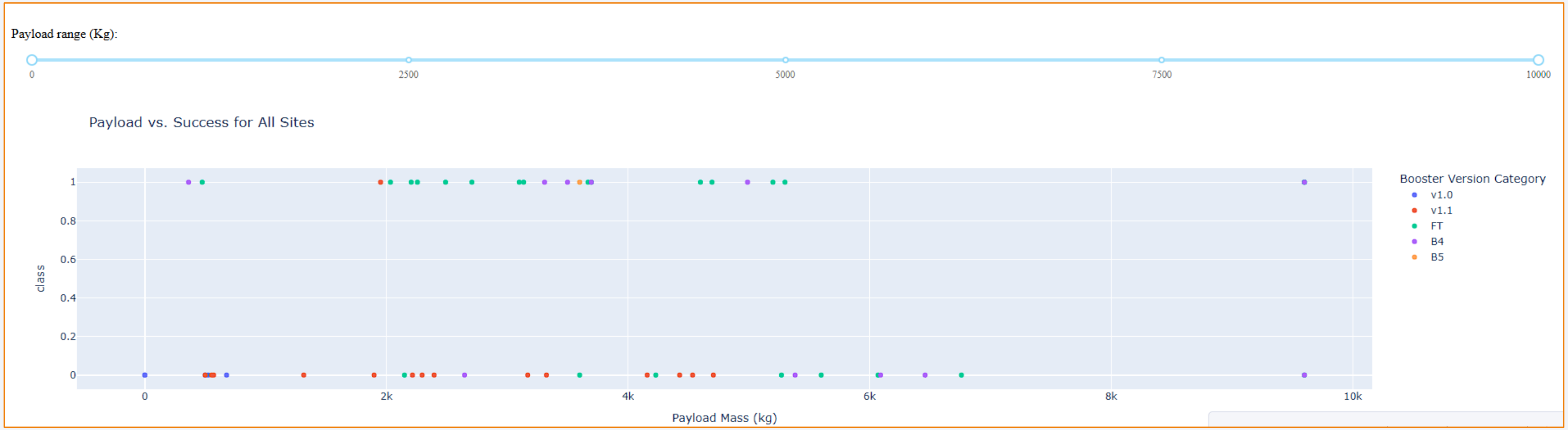
Dashboard Pie Chart



We can see themost succesful launchsite, in this case is KSC LC-39ª , with 13 flights of wich 10 where successfull

We can determinate, that heavy payload are have a higer risk, success does not depend sole it in the booster

Dashboard Scatter Plot



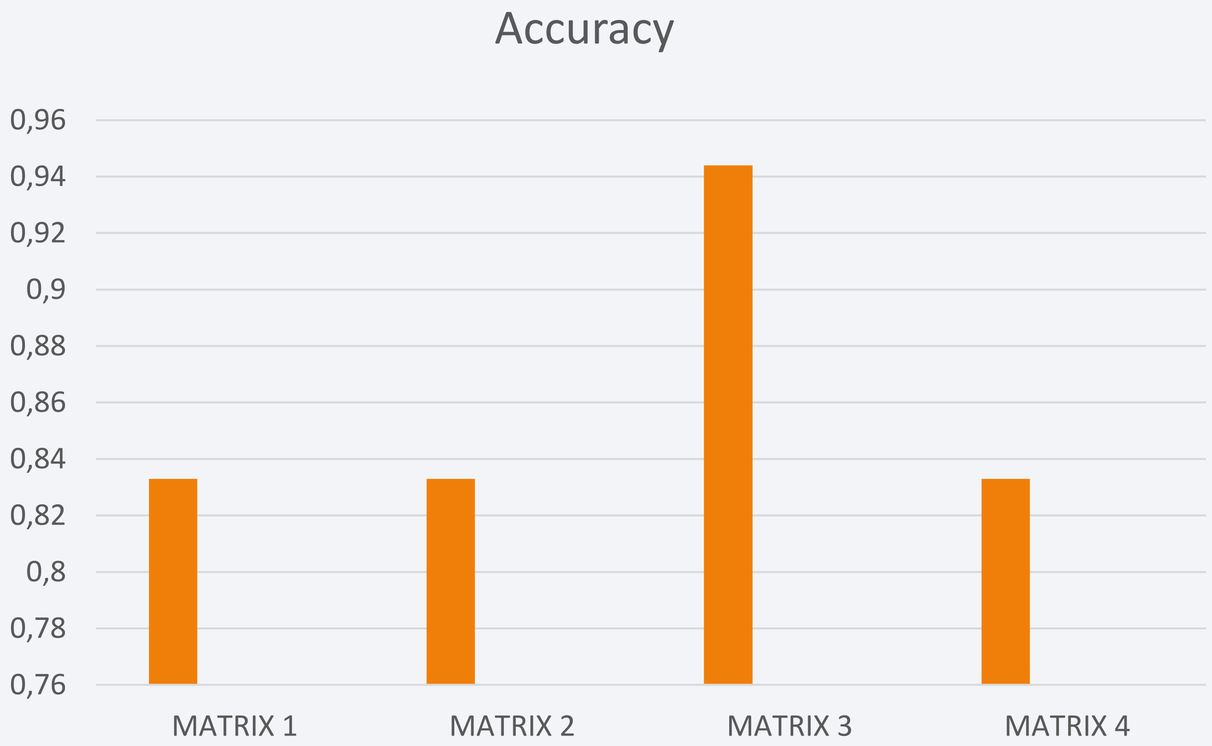
In the scater plot we can conclude that:

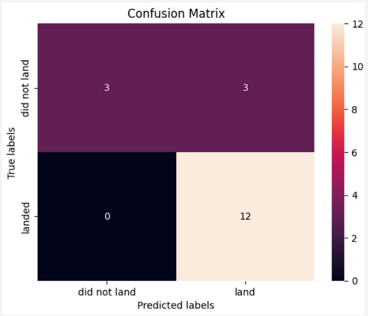
- v1, and 1.1 are one of the flyes with a very low realibility
- Ft (full thrist) is the one with the highest success rate

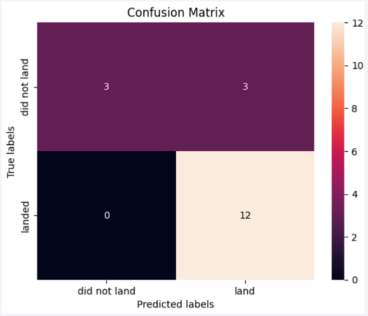
Section 5

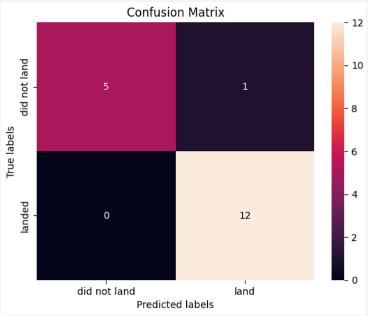
Predictive Analysis (Classification)

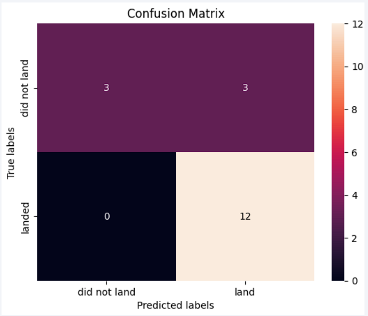
Classification Accuracy







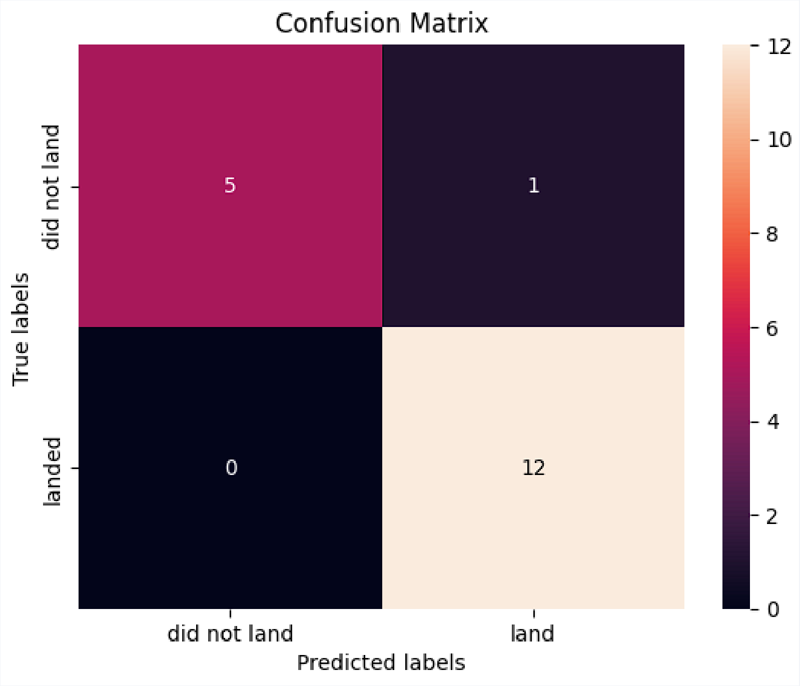


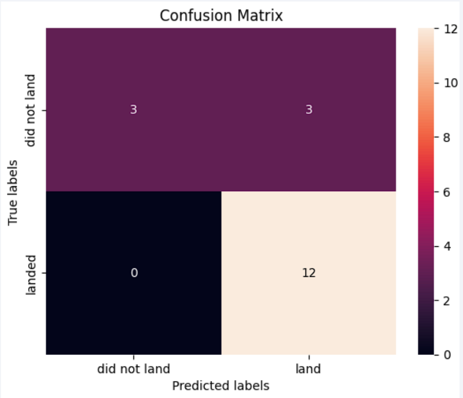


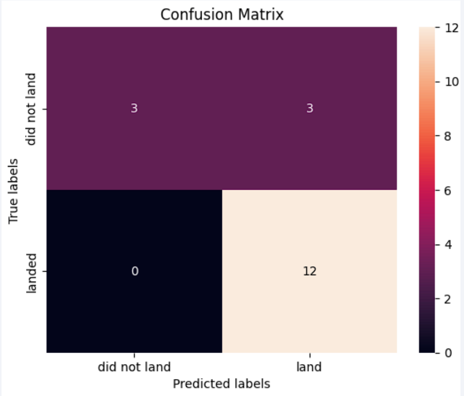
We can se the accuarcy of the 4 models, as we can se the third one has the most accuracy

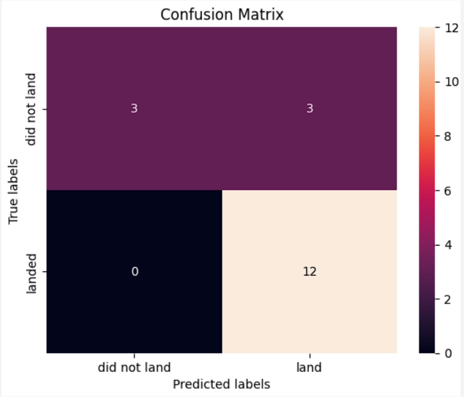
Confusion Matrix

- The confusion matrix of the 3rtd model had the was the best performing model while the other 3 where the same









Conclusions

- SpaceX could enhance the success rate of its GEO/GTO launches by adopting an equatorial sea-based launch approach, similar to the past “Sea Launch” platform.
- Their Falcon 9 FT boosters have shown reliability by landing on drone ships, which supports the viability of ocean-based operations.
- GEO/GTO launches are inherently riskier due to high energy requirements at lift-off and potential vibrational damage to booster electronics and control systems.
- SpaceX has publicly stated that each first-stage booster costs more than $15 million, highlighting the financial impact of launch failures
- Predictive models were used to estimate SpaceX’s ability to successfully land the first-stage booster.
- Three out of four models showed an average accuracy of 83.3%, while the third model reached an impressive 94% accuracy.
- Heavier payloads are associated with higher risk of failure, indicating more stress on the system.
- Success is not solely dependent on the booster—other factors (like payload mass, launch trajectory, or environmental conditions) also play significant roles.

Appendix

Notebooks to recreate dataset, analysis, and models:

capstone/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · AlbertoArancetMerino/capstone

capstone/dashboard.png at main · AlbertoArancetMerino/capstone

capstone/edadataviz.ipynb at main · AlbertoArancetMerino/capstone

capstone/edadataviz.ipynb at main · AlbertoArancetMerino/capstone

capstone/jupyter-labs-spacex-data-collection-api.ipynb at main · AlbertoArancetMerino/capstone

capstone/jupyter-labs-webscraping.ipynb at main · AlbertoArancetMerino/capstone

capstone/lab_jupyter_launch_site_location.ipynb at main · AlbertoArancetMerino/capstone

capstone/labs-jupyter-spacex-Data wrangling.ipynb at main · AlbertoArancetMerino/capstone

capstone/spacex-dash-app.py at main · AlbertoArancetMerino/capstone

Images sources:

Space Launch Market Size, Growth | Report [2025-2033]

Rocket Propulsion Market Size, Share & Forecast Report - 2032

Thank you!