IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zina Mahefarivo RAKOTOMAVONJATOVO
12/04/2024

Outline

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

Executive Summary

- Summary of methodologies

  - Data collection using API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL and visualization
  - Interactive Visual Analytics
  - Dashboarding
  - Predictive analytics

- Summary of all results

  - Identify the key features that affect the result of the launch
  - An interactive dashboard for real-time analysis has been developed
  - Produce a model capable of predicting the outcome of rocket launch with an accuracy of 83.8%

Introduction

- Project background and context

  SpaceX advertises Falcon 9 rocket launches at a cost of 62 million dollars, compared to 165 million dollars for other providers

  This cost saving is essentially due to the fact that SpaceX reuses its rocket first stage

  Being able to predict the likelihood of a successful landing gives the advantage, to an alternate company, to estimate more precisely the launch costs, and apply a better bid strategies when competing against SpaceX

  And this is the aim of this analysis

- Problems you want to find answers
  - What are the factors that determine the successful landing of a rocket?
  - Can we predict the likelihood of a successful landing?
  - What is the maximum accuracy of the prediction?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Calling SpaceX REST API
  - Web Scraping on Wikipedia pages
- Perform data wrangling
  - Data Inspection, Data Cleaning, Encoding Categorical values, Data Normalization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models from Scikit-Learn library
  - Find best parameters using *GridSearchCV*

Data Collection

We have 2 different data sources involving 2 different methods:

- **SpaceX REST API**

  - Call the API using *requests* object

  - Parse the JSON response to Pandas *DataFrame*

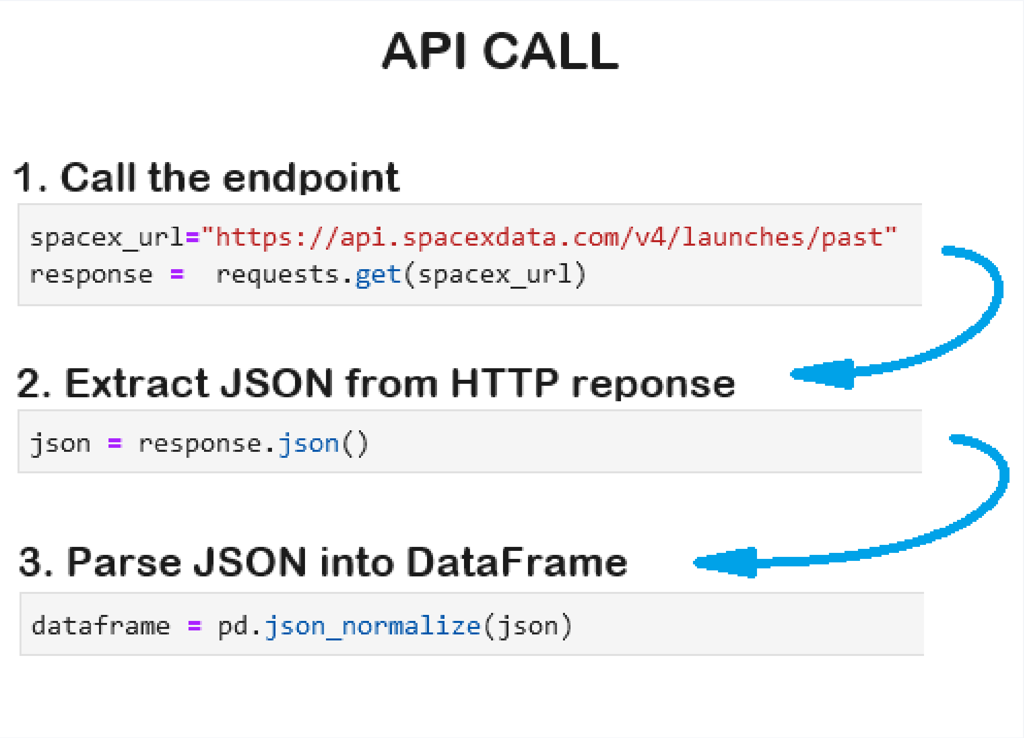
- **Web Scraping**

  - *BeautifulSoup* is used to extract data from Wikipedia web page

  - SpaceX launch records contained in HTML Table are parsed and converted to Pandas *DataFrame*

Data Collection – SpaceX API

- Need to call multiple endpoints to get the whole data set
- Start requesting rocket launch historical data
- From *launchpad* , request for launch site name and location
- From *payloads* , request for payload weight and orbit
- From *cores* , request for rocket details
- Compile all into a single DataFrame



- Link to notebook: https://github.com/zina-mahefarivo/spacex-capstone-project/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Obtain page content by using HTTP get
- Create BeautifulSoup object from HTML page
- Select Launch HTML table
- Iterate on rows
- Extract data from each cell
- Create DataFrame from all result lists

WEB SCRAPING

1. Get page content

```
static_url = "https://en.wikipedia.org/w/index.php"
            +"?title=List_of_Falcon_9_and_Falcon_Heavy_launches"
            +"&oldid=1027686922"
response = requests.get(static_url)
```

2. Create soup

```
soup = BeautifulSoup(response.content, "html.parser")
```

3. Iterate on rows

```
for table_number,table in enumerate(
    soup.find_all('table',"wikitable plainrowheaders collapsible")
):
```

4. Extract data from cell

```
            launch_site = row[2].a.string
            launch_dict["Launch site"].append(launch_site)
```

5. Create DataFrame

```
df = pd.DataFrame(
    { key:pd.Series(value) for key, value in launch_dict.items() }
)
```

- Link to notebook: https://github.com/zina-mahefarivo/spacex-capstone-project/blob/main/2_jupyter-labs-webscraping.ipynb

Data Wrangling

- Clean data
  - Identify missing values
  - Replace missing *PayloadMass* by mean value
- Ensure Data Consistency
  - Check data types
- Explore Data
  - Nb of launches per site
  - Nb of occurence of each orbit
  - Nb of occurence of outcome

OUTCOME	COUNT	SITE	LANDING	CLASS
True ASDS	41	Drone Ship	Success	1
None None	19	Not Attempted	Not Attempted	0
True RTLS	14	Ground Pad	Success	1
False ASDS	6	Drone Ship	Fail	0
True Ocean	5	Ocean	Success	1
False Ocean	2	Ocean	Fail	0
None ASDS	2	Drone Ship	Not Attempted	0
False RTLS	1	Ground Pad	Fail	0

- Transform Data
  - From *Outcome*, extract/encode the *Class* of the launch
- Link to notebook: https://github.com/zina-mahefarivo/spacex-capstone-project/blob/main/3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Plotted Charts

- **Flight Number VS Launch Site:** To verify the pattern between the 2 variables and the outcome
- **Payload VS Launch Site:** To visualize which site launch heavy payload
- **Orbit VS Success Rate:** To identify which orbits have high landing success rate
- **Orbit VS Flight Number:** To check the relationship between Flight Nb and orbit
- **Orbit VS Payload:** To verify to which orbit heavy/light payloads are sent
- **Year VS Success Rate:** To emphasize from which year success rate started to increase
- **Link to notebook:** https://github.com/zina-mahefarivo/spacex-capstone-project/blob/main/4-jupyter-lab-data-visualization.ipynb

EDA with SQL

- List of unique launch sites in the space mission
- List 5 records where launch sites begin with 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing on ground pad
- Boosters with payload between 4000 and 6000 kg that successfully landed on drone ship
- Total number of successful and failure mission outcomes
- Names of boosters that carried the maximum payload mass
- Count of landing outcomes between 2 dates in descending order
- Link to notebook: https://github.com/zina-mahefarivo/spacex-capstone-project/blob/main/5_jupyter-lab-eda-with-sql.ipynb

Build an Interactive Map with Folium

- Circle - Used to indicate a launch site
- Marker - Combined with a Circle, is used to indicate a launch site
- Marker - Also used to indicate the outcome of each launch in each site
- MarkerCluster - Used to display a constellation of Markers
- MousePosition - Indicates the latitude and longitude of the mouse pointer
- Line - Used to indicate distance

- Link to notebook: https://github.com/zina-mahefarivo/spacex-capstone-project/blob/main/6_jupyter-lab_launch-site-location.ipynb

Build a Dashboard with Plotly Dash

INPUTS

- Drop-down list - Allows to select different launch sites
- Range slider - Allows to select a range of payload weight

CHARTS

- Pie chart - Shows the proportion of success landing for all or the selected site
- Scatter plot - Displays the correlation between the Payload and the Outcome

- Link to notebook: https://github.com/zina-mahefarivo/spacex-capstone-project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Algorithms: Logistic Regression, SVM, Decision Tree, KNN
- Use GridSearchCV object to find the best parameters
- Display confusion matrix
- Calculate accuracy on training and test sets
- Compare scores to identify the best model

- Link to notebook: https://github.com/zina-mahefarivo/spacex-capstone-project/blob/main/7_jupyter-lab_SpaceX_Machine_Learning_Prediction.ipynb

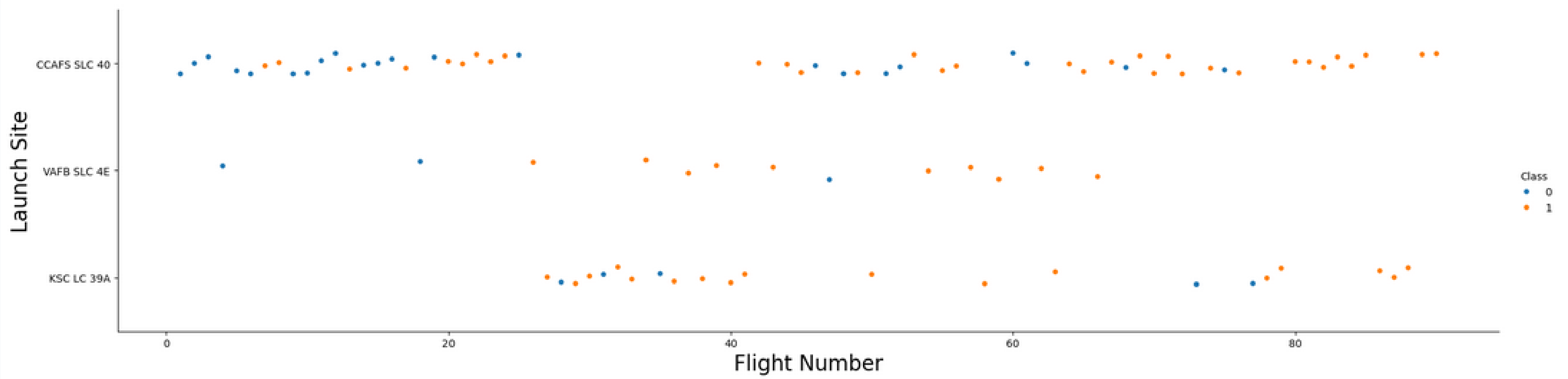
Results

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

Section 2

Insights drawn from EDA

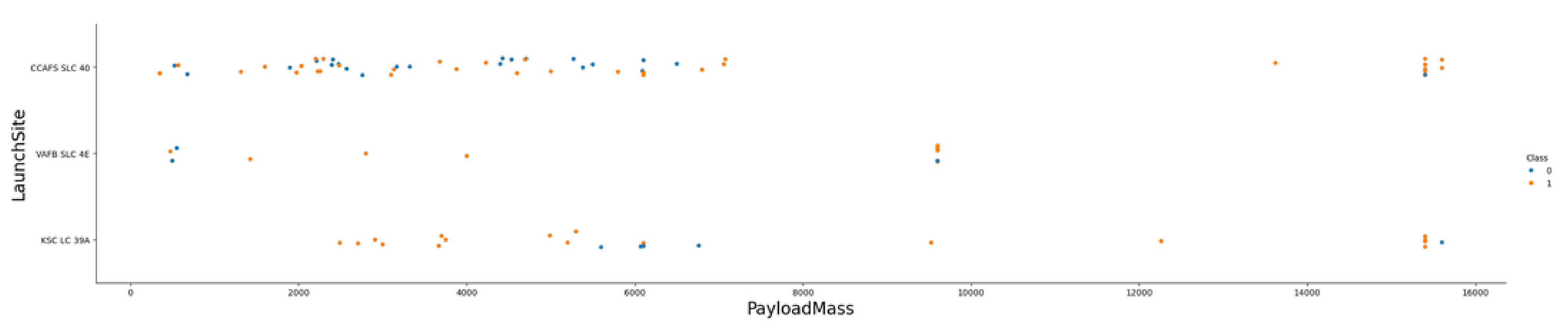
Flight Number vs. Launch Site



Insights:

- There are more successful landing starting from Flight Number 25
- With more attempts, we have more successful outcome

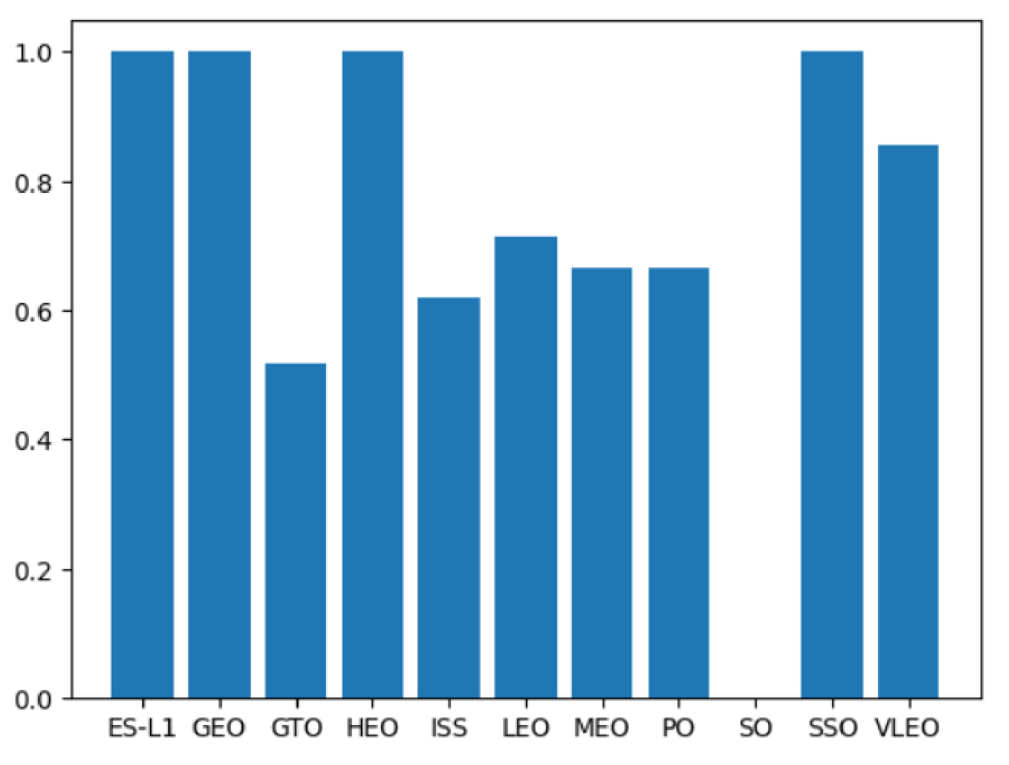
Payload vs. Launch Site



Insights:

- No heavy launches has been sent from **VAFB SLC 4E**
- There are more failed landing attempts on **CCAFS SLC 40** than on other sites

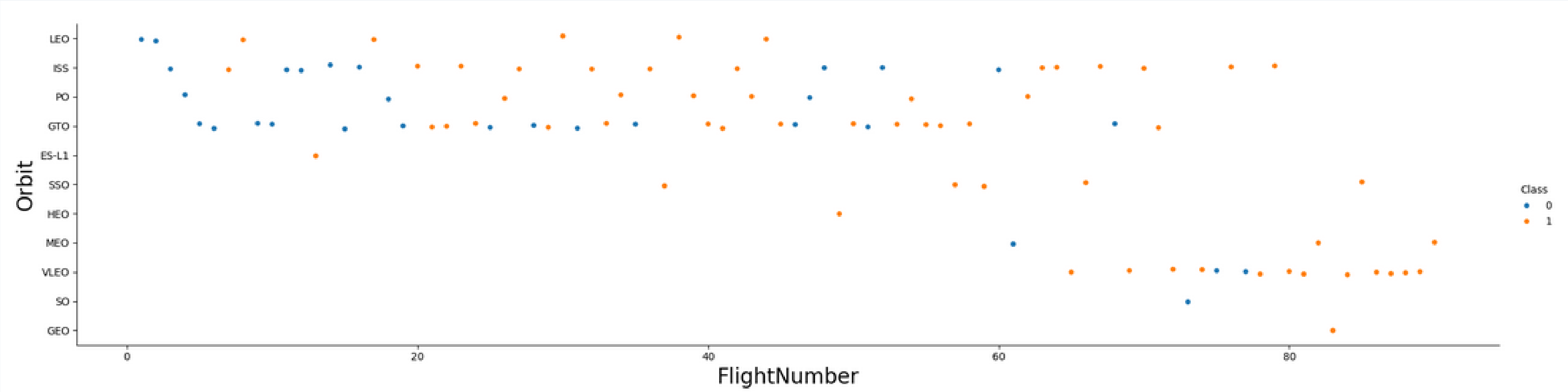
Success Rate vs. Orbit Type



Insights:

- Rockets sent to **ES-L1**, **GEO**, **HEO**, **SSO** and **VLEO** have high success rate
- No rockets sent to **SO** have landed successfully

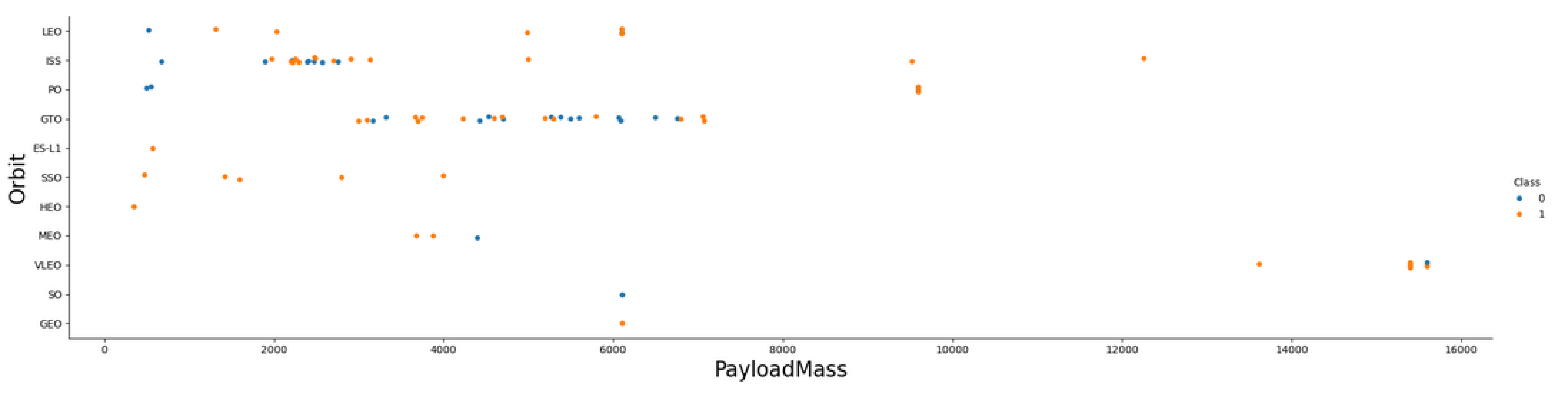
Flight Number vs. Orbit Type



Insights:

- For LEO orbit, the outcome appears related to the Flight Number
- No rockets sent to **SO** have landed successfully

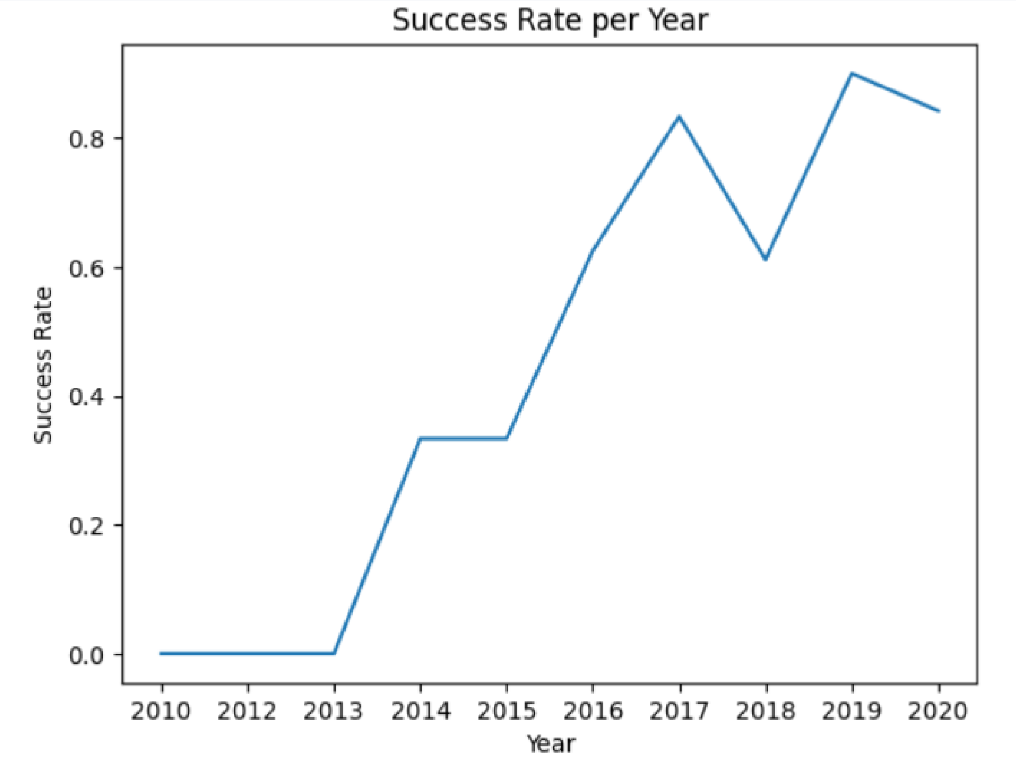
Payload vs. Orbit Type



Insights:

- For heavy payloads, there is more successful landing for **ISS**, **PO**, and **VLEO**
- Successful and failed landing are mixed for **GTO**

Launch Success Yearly Trend



Insights:

- The landing success rate kept increasing since 2013 till 2020

All Launch Site Names

In [19]:
```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

Out[19]:

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

- Use of the keyword DISTINCT to return unique Launch Site Name
- There is only 4 different launch Site used by SpaceX

Launch Site Names Begin with 'CCA'

In [18]:

```
%%sql

select * from SPACEXTABLE
where Launch_Site like 'CCA%'
limit 5;
```

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

Out[18]:

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

- `where Launch_Site like 'CCA%'` indicates that Launch Site must start with 'CCA'
- `limit 5;` restricts the results to 5 records only

Total Payload Mass

```
In [23]: %%sql

         select sum(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS
         from SPACEXTABLE
         where Customer = 'NASA (CRS)';

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

Out[23]:

TOTAL_PAYLOAD_MASS
45596

- The total payload carried by boosters from NASA is **45596 kg**

Average Payload Mass by F9 v1.1

In [26]:
```
%%sql

select avg(PAYLOAD_MASS__KG_) as AVERAGE_PAYLOAD_MASS
from SPACEXTABLE
where Booster_Version = 'F9 v1.1'
```

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

Out[26]:

AVERAGE_PAYLOAD_MASS
2928.4

- Boosters version F9 v1.1 carry in average a payload of **2928.4 kg**

First Successful Ground Landing Date

```
In [30]:  %%sql

          select min(Date) as FIRST_SUCCESS_LANDING
          from SPACEXTABLE
          where Landing_Outcome = 'Success (ground pad)'

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

Out[30]:

FIRST_SUCCESS_LANDING
2015-12-22

- The first successful Ground Landing was on 22 December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

In [33]:

```
%%sql

select Booster_Version
from SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Out[33]:

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

- There are 4 different booster versions that have successfully landed on drone ship with payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

In [34]:

```
%%sql

select Mission_Outcome, count(Mission_Outcome)
from SPACEXTABLE
group by Mission_Outcome
```

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

Out[34]:

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

- Only 1 Failure Mission Outcome has been found in the dataset

Boosters Carried Maximum Payload

In [37]:

```
%%sql

select Booster_Version
from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

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

- There are 12 different Boosters that carried the maximum payload

2015 Launch Records

In [41]:
```
%%sql

select substr(Date, 6,2) as month, Landing_Outcome, Booster_Version
from SPACEXTABLE
where substr(Date,0,5)='2015'
and Landing_Outcome = 'Failure (drone ship)'
-- limit = 1
```

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

Out[41]:

month	Landing_Outcome	Booster_Version
01	Failure (drone ship)	F9 v1.1 B1012
04	Failure (drone ship)	F9 v1.1 B1015

- There are 2 failed landing_outcomes on drone ship in year 2015

  First was on January and the second in April

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

In [45]:

```
%%sql

select Landing_Outcome as OUTCOME, count(Landing_Outcome) as COUNT
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by 2 desc
```

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

Out[45]:

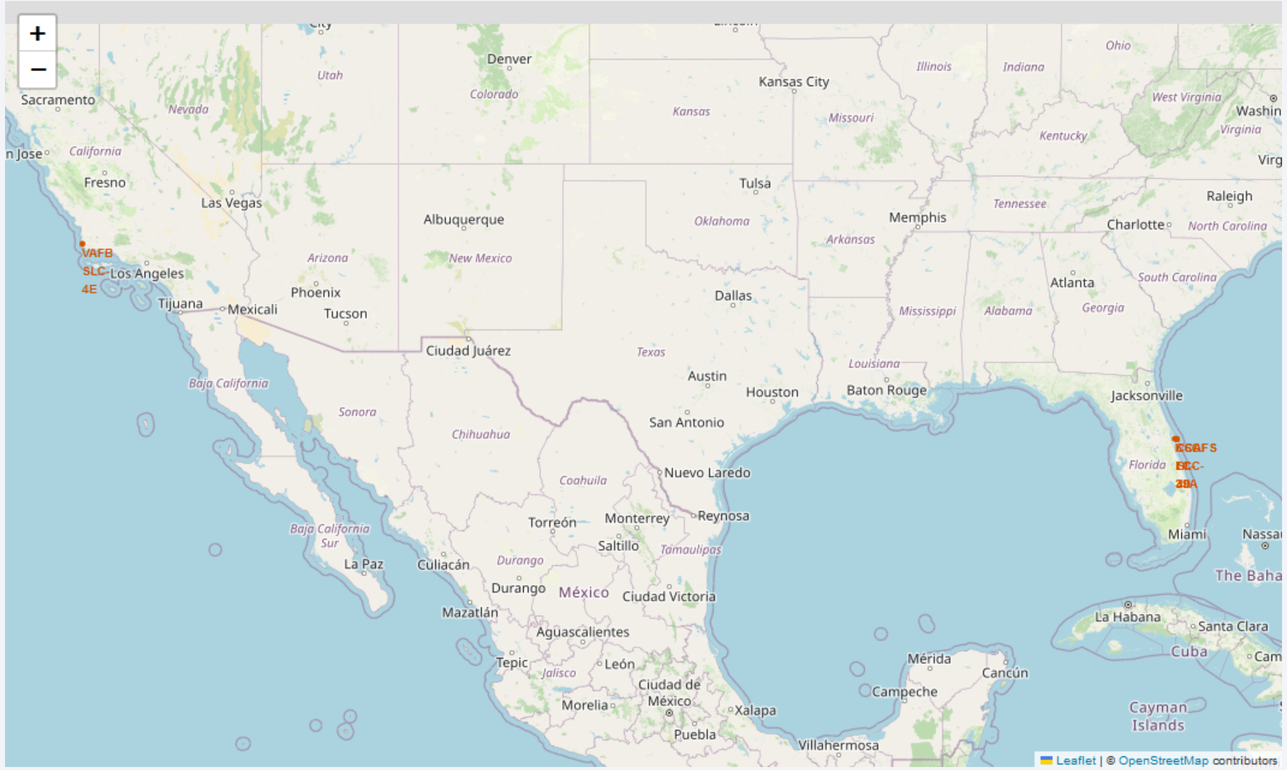
OUTCOME	COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- The most frequent outcome is No attempt
- The least recurring outcome is Precluded on Drone Ship

Section 3

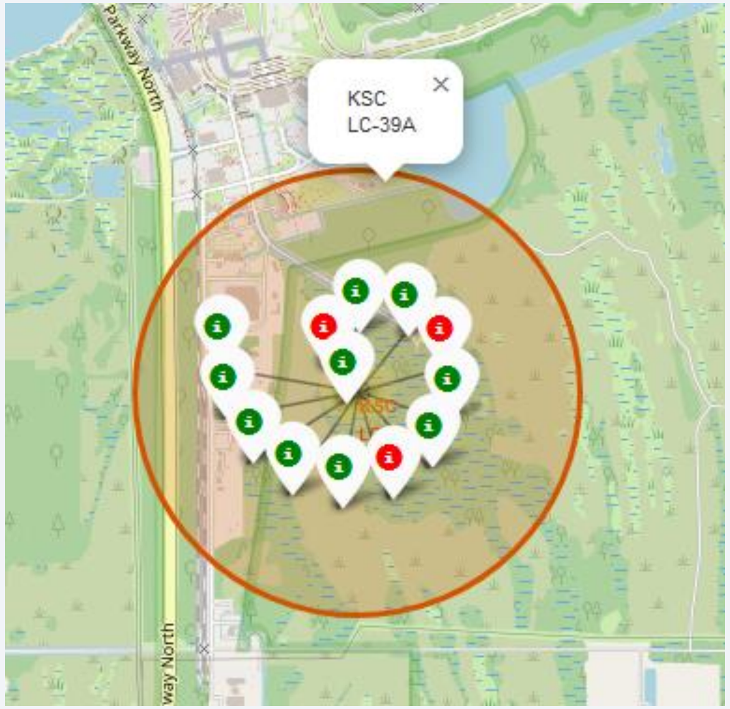
Launch Sites Proximities Analysis

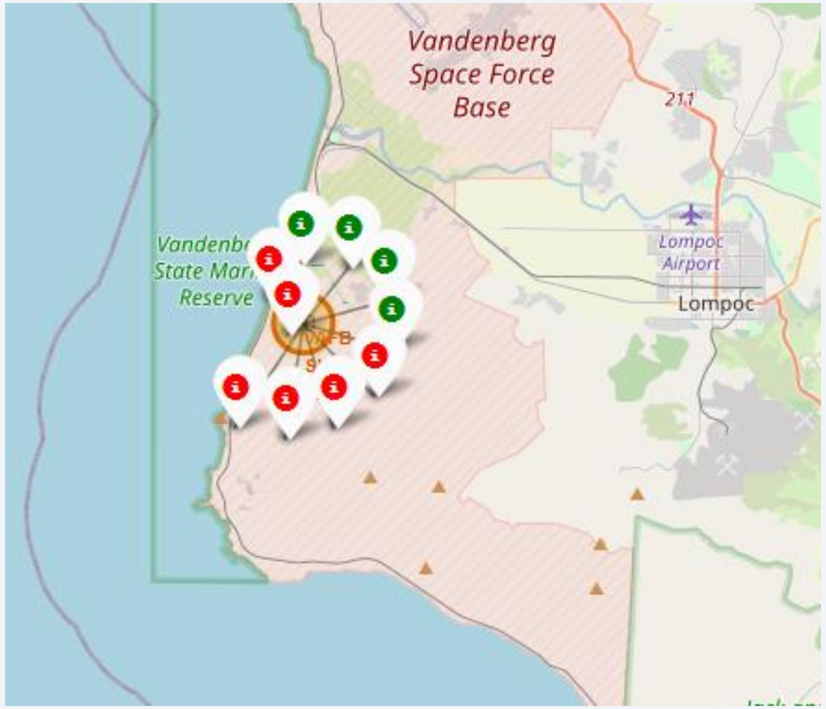
SpaceX Launch Sites



- All launch sites are located on American coasts
- Only VAFB SLC 4E is located on west coast. All the others are on east coast

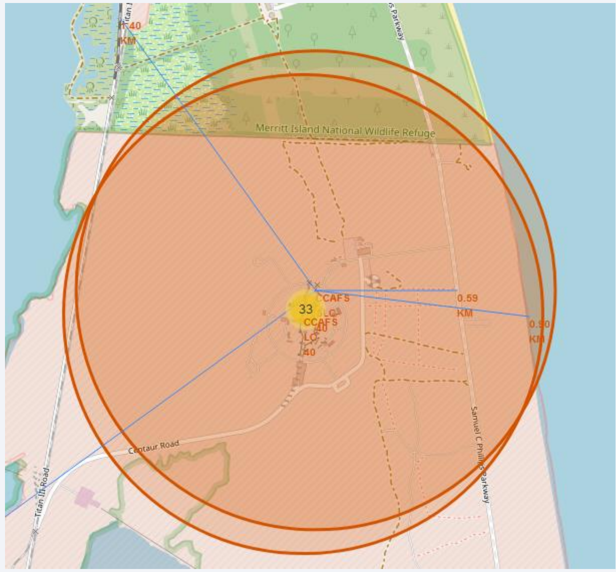
Landing outcome for each Launch Site

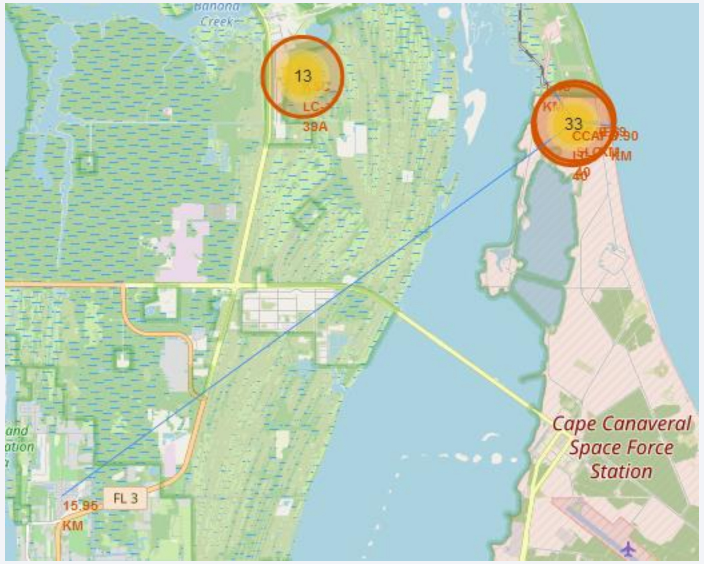




- A successful landing is represented by a green Icon, And Red Icon for a failure
- Launch sites with high success rates can be easily identified

Launch Site Proximities



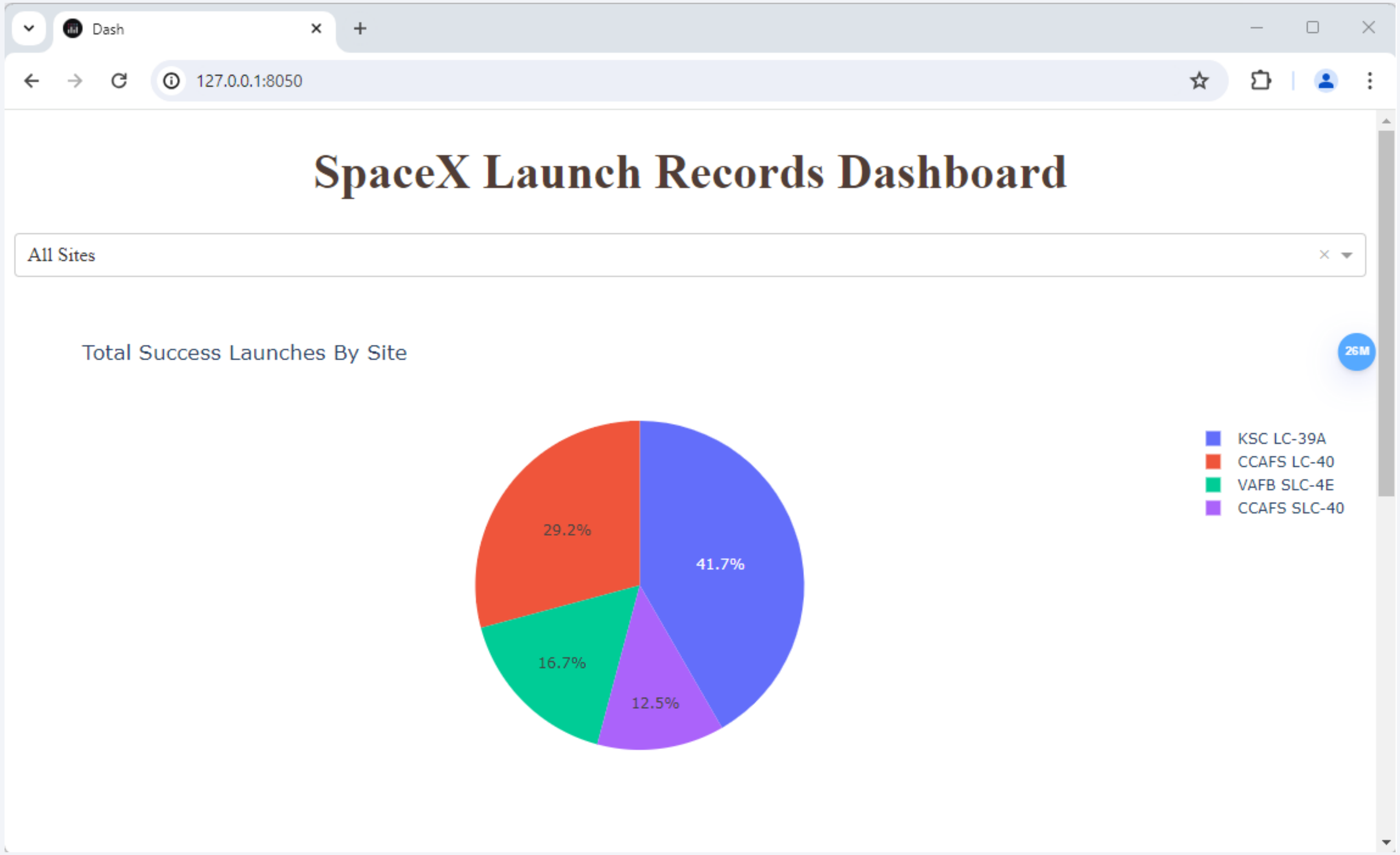


- Space launch sites are located near the equator
- Launch sites are in close proximity to railways to facilitate material (heavy cargo) transportation
- Launch sites are in close proximity to highways to facilitate people and material transportation
- Launch sites are in close proximity to coastline as water is safety buffer in case of launch failure
- Launch sites are not in close proximity to cities to ensure the safety of the population

Section 4

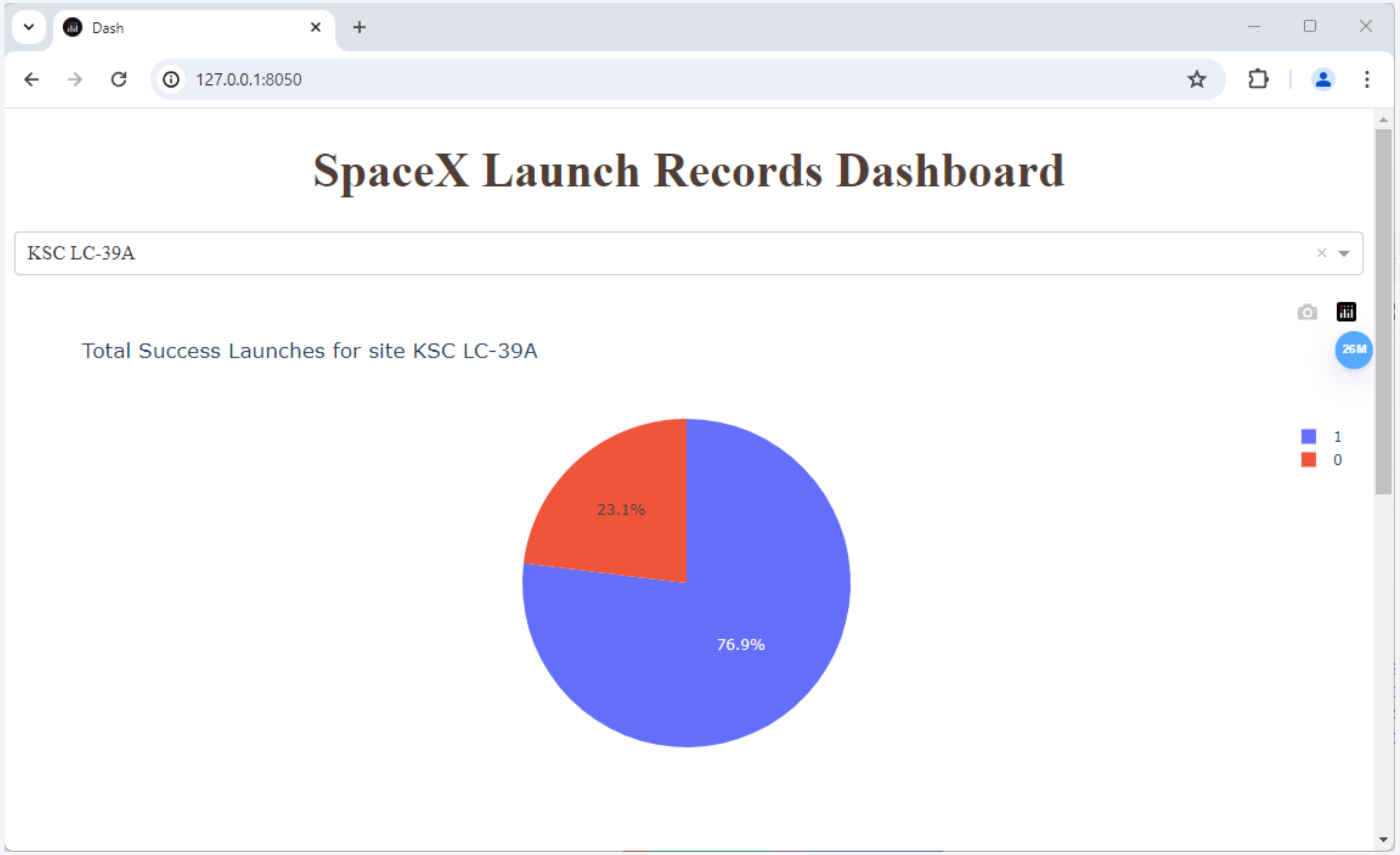
Build a Dashboard with Plotly Dash

Dashboard – Success Launches by Site



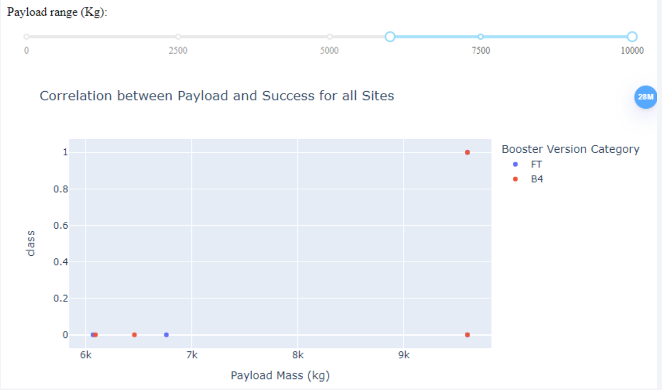
- The dropdown list allows to select launch Site
- KSC LC-39A has the highest success rate
- CCAFS SLC-40 has the lowest success rate

Dashboard - Site with Highest Success Rate



- The proportion of success landing for KSC LC-39A is 76.9%
- The ratio of failed landing for KSC LC-39A is 23.1%
- KSC LC-39A is selected in the dropdown list

Dashboard – Correlation between Payload and Success

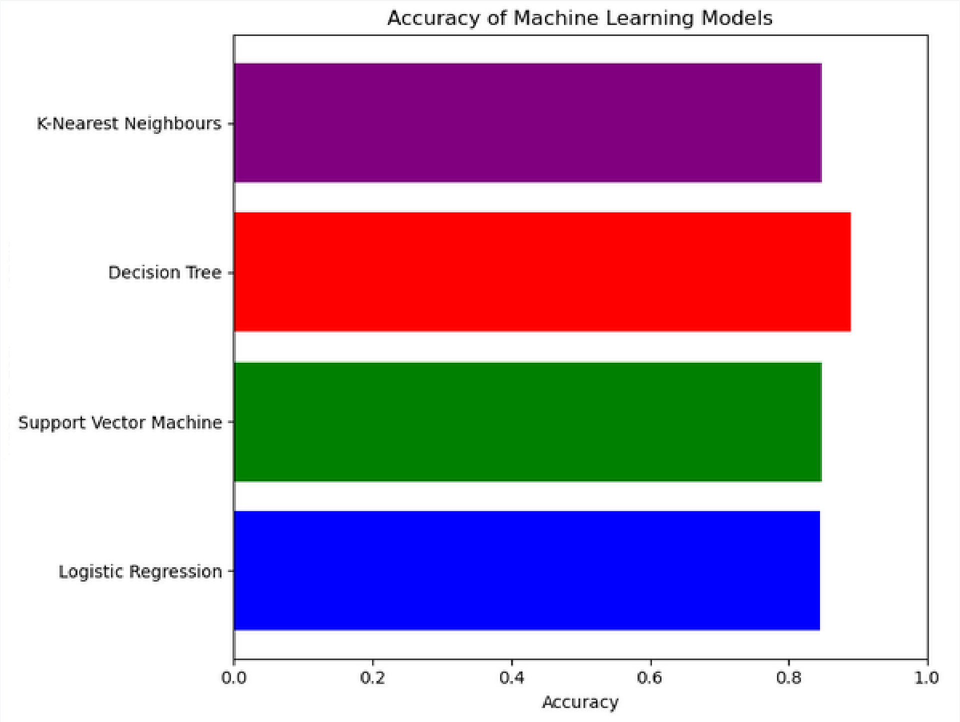


- For light payload, the boosters B4, FT and V1.1 provide a better success rate
- For medium payload, the boosters B4, FT provide a better results
- For heavy payload, only B4 provides a good success rate
- The 1st stage has higher chance to land if they carried medium payload

Section 5

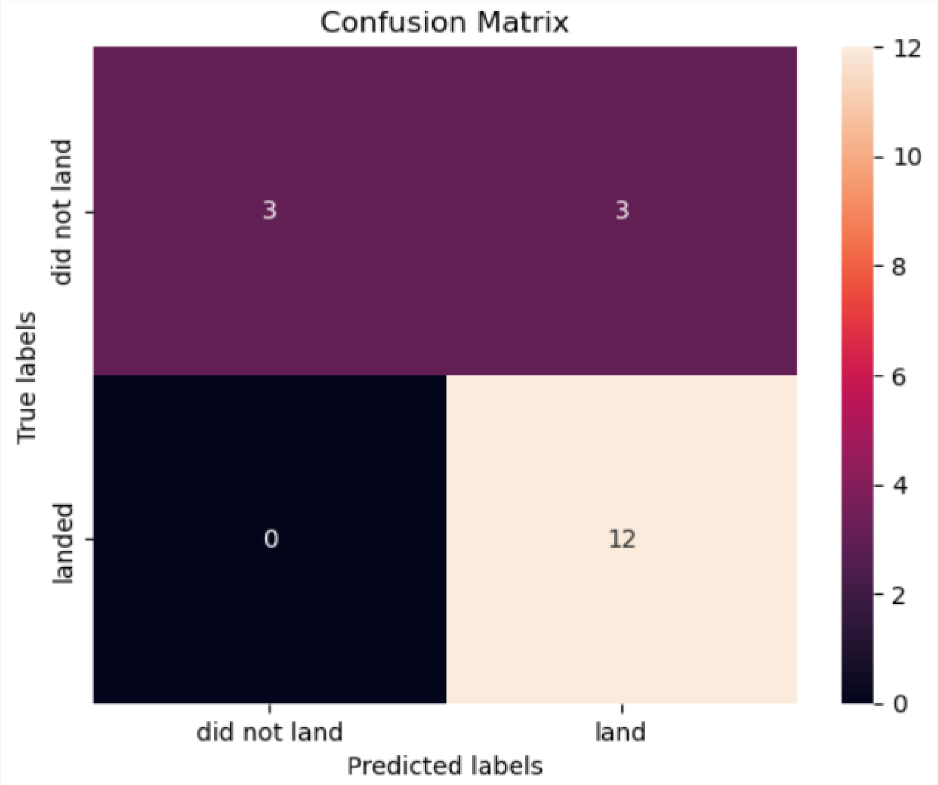
Predictive Analysis (Classification)

Classification Accuracy



- The Decision Tree model has the highest classification accuracy

Confusion Matrix



- The major problem is false positives
- 3 unsuccessful landing predicted as successful by the classification model

Conclusions

- The success rate kept increasing since 2013 till 2020
- Rockets sent to orbits ES-L1, GEO, HEO, SSO and VLEO have high success rate
- KSC LC-39A has the highest success rate among launch sites
- The booster version B4 provides good results for light, medium, and heavy payloads
- The Decision Tree classifier performs slightly better on training set, but all models perform the same on unseen data (test set)
- The accuracy of the classification model is 83.33%

Thank you!