IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Irina Glickman
6/20/2023

Outline

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

Executive Summary

- Methodologies:
  - ✓ Data collection with web scrapping
  - ✓ Exploratory data analysis including visualization
  - ✓ Data visualization
  - ✓ Building an interactive dashboard
  - ✓ Predictive analysis

- Results
  - ✓ The accuracy of the predictive models (logistic regression, SVM, decision tree, KNN): 83%

Introduction

- Our mission is to predict if the Falcon 9 first stage will land successfully in order to determine the cost of a launch.

- This information can be used to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Performing web scraping to collect Falcon 9 historical launch records from Wikipedia
- Perform data wrangling
  - Clean the requested data
  - Replace the missing values of the PayloadMass with the calculated mean
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a machine learning pipeline to predict if the first stage will land

Data Collection

- A get request to the SpaceX API

    Use the API to extract information using identification numbers in the launch data:

        We would like to learn:

    - ✓ The booster name
    - ✓ The name of the launch site being used, the longitude, and the latitude.
    - ✓ The mass of the payload and the orbit that it is going to
    - ✓ The outcome of the landing, the type of the landing, number of flights with specific core, and core info

- Data wrangling and formatting

Data Collection – SpaceX API

- Data collection with SpaceX REST
- GitHub URL of the completed SpaceX API calls notebook

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

requesting rocket launch data from SpaceX API with URL

response

decode the response content as a Json and turn it into a Pandas dataframe

use the API again to get information about the launches using the IDs given for each launch

construct our dataset using the data we have obtained and create a Pandas data frame

remove the Falcon 1 launches keeping only the Falcon 9 launches

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- GitHub URL of the completed web scraping notebook

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

perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response

Create a BeautifulSoup object from the HTML response

collect all relevant column names from the HTML table header

create an empty dictionary with keys from the extracted column names and conver it into a Pandas dataframe

export it to a CSV

Data Wrangling

- Describe how data were processed
  - ✓ Identify and calculate the percentage of the missing values in each attribute
  - ✓ Calculate the number of launches on each site
  - ✓ Calculate the number and occurrence of each orbit
  - ✓ Calculate the number and occurence of mission outcome per orbit type
  - ✓ Create a landing outcome label from Outcome column

- GitHub URL of your completed data wrangling related notebooks

https://github.com/IrinaGlickman/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite%20(1).ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - ✓ Flight Number vs. Payload Mass to overlay the outcome of the launch.
    We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
  - ✓ Flight Number vs. Launch Site
    Different launch sites have different success rates
  - ✓ Payload vs. Launch Site
  - ✓ Success rate vs. Orbit type
    To find which orbits have high success rate
  - ✓ Flight Number vs. Orbit type
  - ✓ Payload vs. Orbit type
  - ✓ Launch success yearly trend
- https://github.com/IrinaGlickman/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite%20(2).ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
  - ✓ Display the names of the unique launch sites in the space mission
  - ✓ Display 5 records where launch sites begin with the string 'CCA'
  - ✓ Display the total payload mass carried by boosters launched by NASA (CRS)
  - ✓ Display average payload mass carried by booster version F9 v1.1
  - ✓ List the date when the first succesful landing outcome in ground pad was achieved.
  - ✓ List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - ✓ List the total number of successful and failure mission outcomes
  - ✓ List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - ✓ List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - ✓ **Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.**
- https://github.com/IrinaGlickman/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
    - ✓ Circle - to add a highlighted circle area with a text label on a specific coordinate.
    - ✓ Marker for each launch site on the site map
    - ✓ Markers for all launch records
    - ✓ Mouse Position on the map to get coordinate for a mouse over a point on the map
    - ✓ Mouse Position to mark down a point on the closest coastline and to calculate the distance between the coastline point and the launch site.
    - ✓ Poly Line to draw a line between a launch site to the selected coastline point
- https://github.com/IrinaGlickman/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - ✓ Launch Site Drop-down Input Component
    to chose which launch site is shown on graphs
  - ✓ Success-pie-chart
    to show which site has the largest successful launches
  - ✓ Callback function to render success-pie-chart based on selected site dropdown
  - ✓ Range Slider to Select Payload
    Which payload range has the highest/lowest launch success rate
  - ✓ Success-payload-scatter-chart
    to show which F9 Booster version has the highest launch success rate
  - ✓ Callback function to render the success-payload-scatter-chart scatter plot
- https://github.com/IrinaGlickman/Applied-Data-Science-Capstone/blob/main/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- https://github.com/IrinaGlickman/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

Create a column for the class and standardize the data

Split the data into training and testing data

Create a LR/SVM/Decision Tree/KNN objects and fit them to find the best parameters from the dictionary parameters.

Calculate the accuracy of each model

Plot a confusion matrix for each model

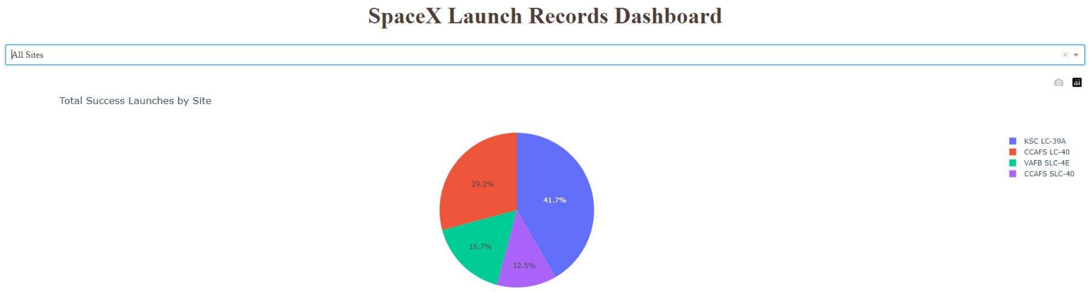
Compare the performances of each model

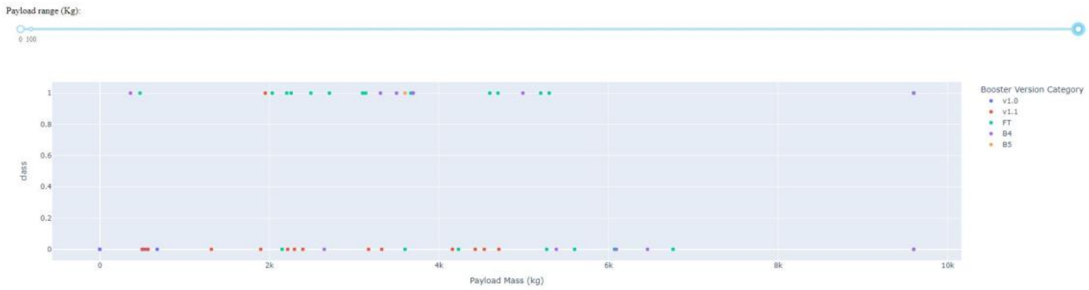
Results

- Exploratory data analysis results
  - ✓ As the flight number increases, the first stage is more likely to land successfully
  - ✓ The more massive the payload, the less likely the first stage will return.
  - ✓ CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
  - ✓ For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
  - ✓ In the LEO orbit the Success appears related to the number of flights, but there is no relationship between flight number when in GTO orbit.
  - ✓ With heavy payloads the positive landing rates are more for Polar, LEO and ISS.
  - ✓ The success rate since 2013 kept increasing till 2020
- Interactive analytics demo in screenshots
- Predictive analysis results

Results

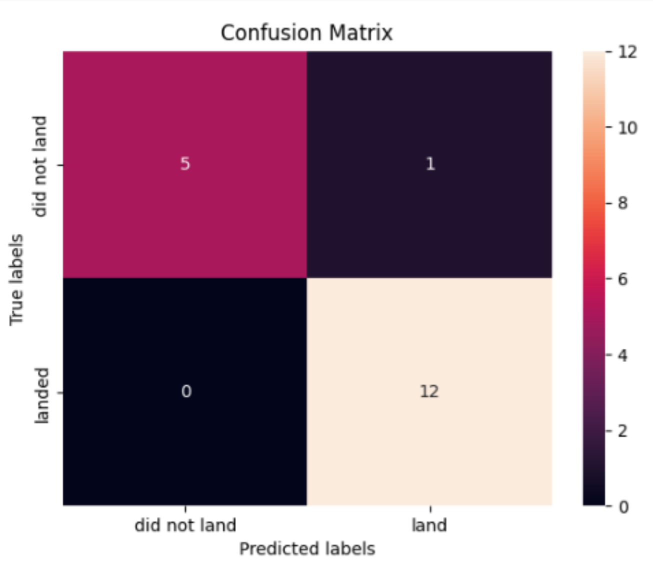
- Interactive analytics demo in screenshots





Results

- Predictive analysis results
  - ✓ The classification models LR, SVM, Decision Tree and KNN are performing with the same accuracy of 83%
  - ✓ The major problem of LR, KNN and SVM is false positives.
  - ✓ Decision Tree shows less false positives

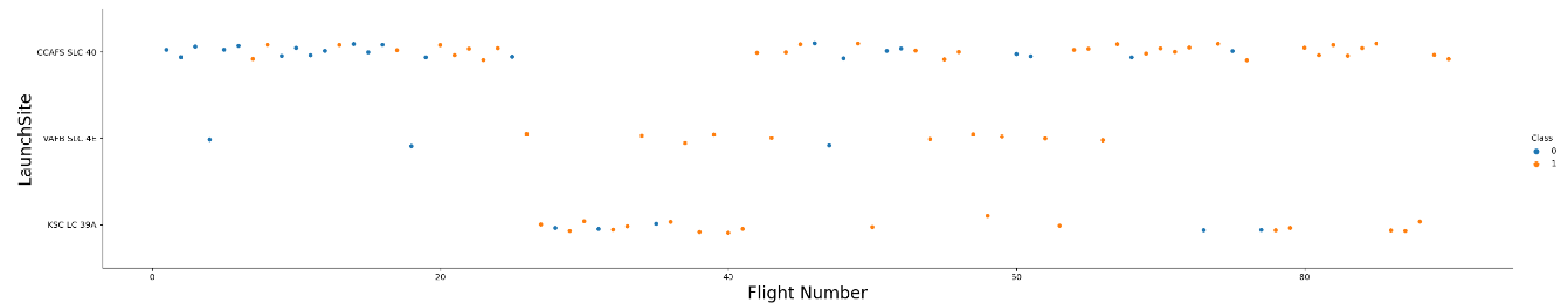


Section 2

Insights drawn from EDA

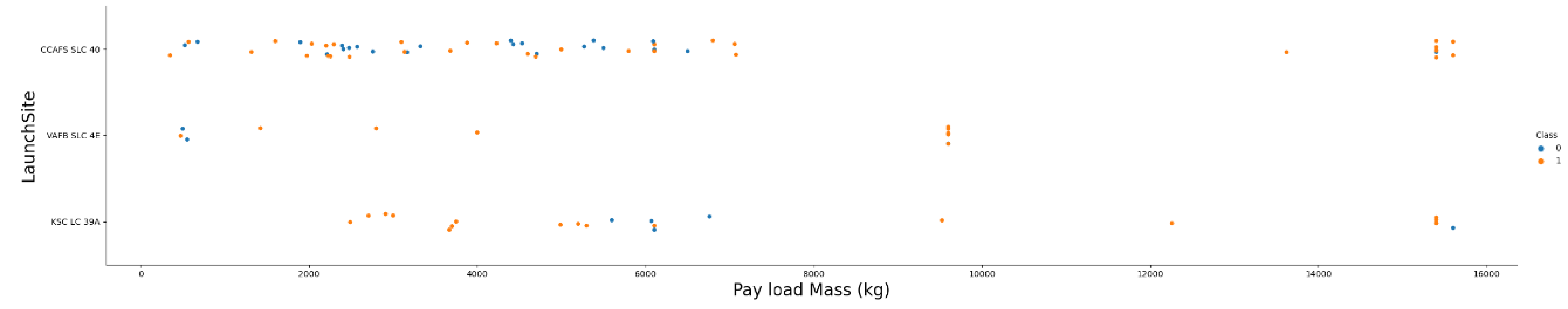
Flight Number vs. Launch Site

- A scatter plot of Flight Number vs. Launch Site



Payload vs. Launch Site

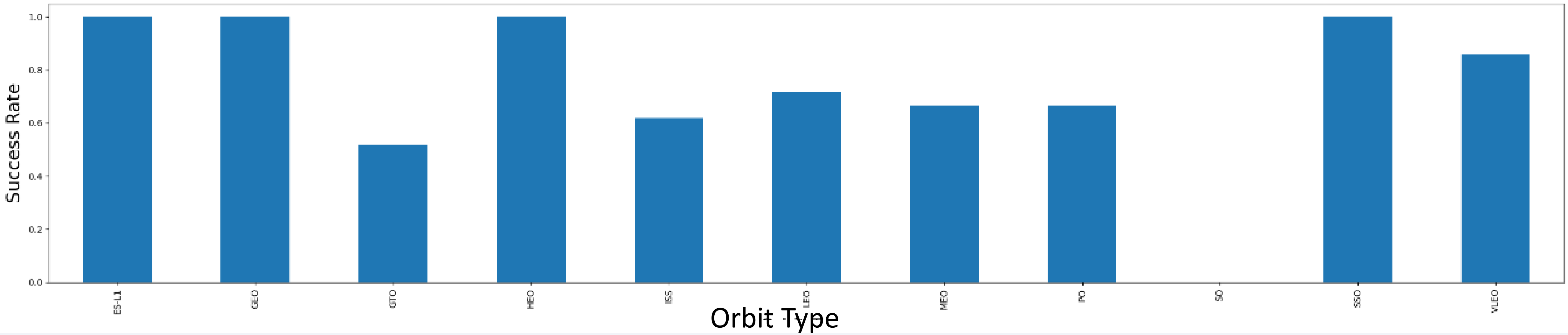
- A scatter plot of Payload vs. Launch Site



For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)

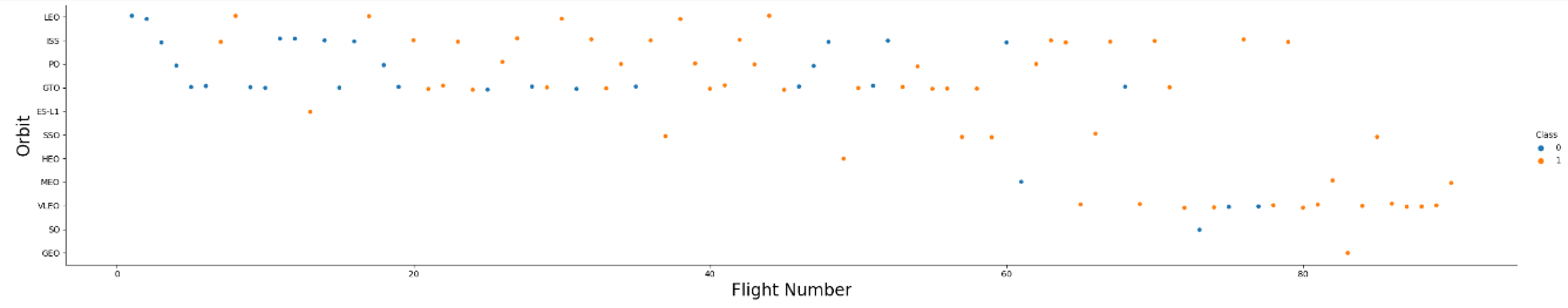
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type



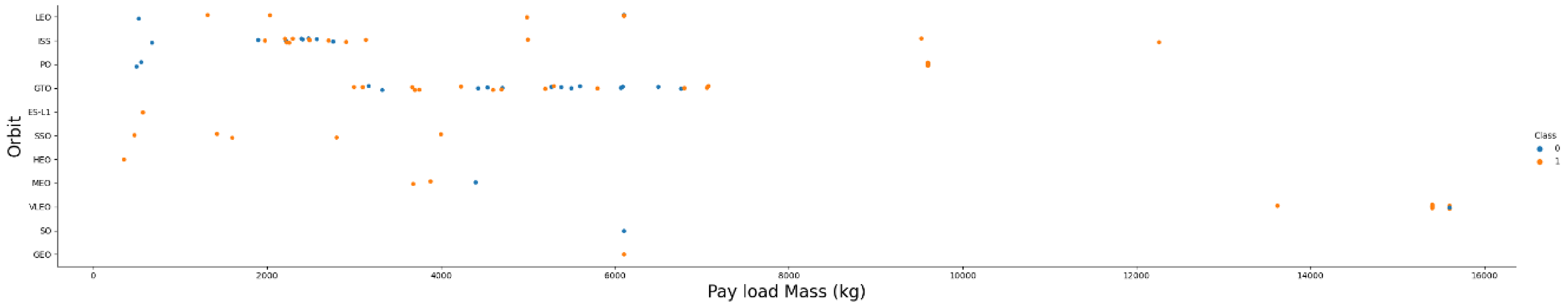
Flight Number vs. Orbit Type

- A scatter point of Flight number vs. Orbit type



Payload vs. Orbit Type

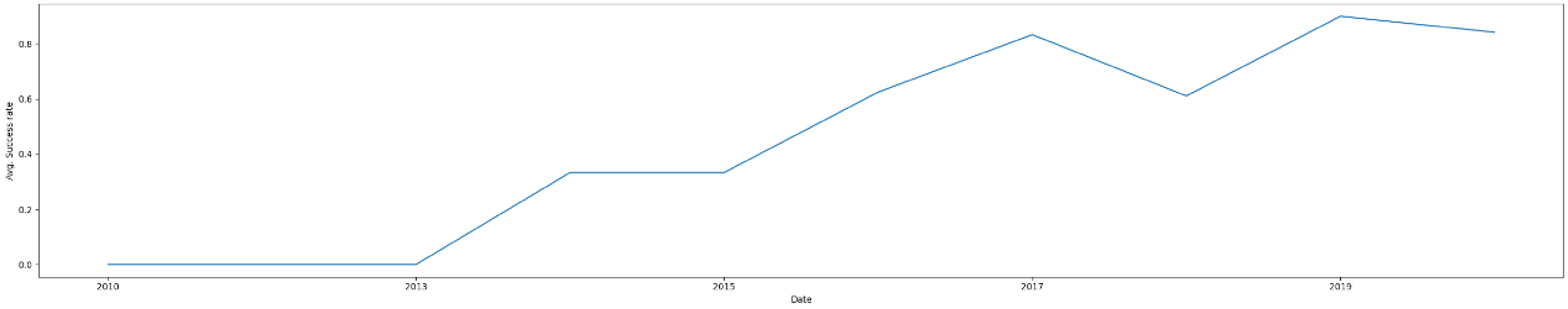
- A scatter point of payload vs. orbit type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



The sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites

Task 1

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

[63]:
```
%%sql
select distinct Launch_Site
from SPACEXTBL
;
```

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

[63]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

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

Task 2

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

```
%%sql
select *
from SPACEXTBL
where Launch_Site like 'CCA%'
limit 10
;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt
12/03/2013	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170.0	GTO	SES	Success	No attempt
01/06/2014	22:06:00	F9 v1.1	CCAFS LC-40	Thaicom 6	3325.0	GTO	Thaicom	Success	No attempt
18/04/2014	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296.0	LEO (ISS)	NASA (CRS)	Success	Controlled (ocean)
14/07/2014	15:15:00	F9 v1.1	CCAFS LC-40	OG2 Mission 1 6 Orbcomm-OG2 satellites	1316.0	LEO	Orbcomm	Success	Controlled (ocean)
08/05/2014	8:00:00	F9 v1.1	CCAFS LC-40	AsiaSat 8	4535.0	GTO	AsiaSat	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Task 3

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

```
%%sql

select total (PAYLOAD_MASS__KG_)
from SPACEXTBL
;
```

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

total (PAYLOAD_MASS__KG_)
619967.0

Average Payload Mass by F9 v1.1

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

Task 4

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

```
%%sql

select avg(PAYLOAD_MASS__KG_)
from SPACEXTBL
where Booster_Version like "F9 v1.1%"
;
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

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

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql

select min(Date)
from SPACEXTBL
where Landing_Outcome = 'Success (ground pad)'

;
```

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

min(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Task 6

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

```
%%sql

select distinct Booster_Version
from SPACEXTBL
where Landing_Outcome = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)

;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%%sql

select Mission_Outcome, count(Mission_Outcome) as Qty
from SPACEXTBL
group by Mission_Outcome
;
```

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

Mission_Outcome	Qty
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Task 8

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

```
%%sql

select distinct Booster_Version
from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

;
```

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

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql

select substr(Date, 4, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTBL
where substr(Date,7,4)='2015' and Landing_Outcome='Failure (drone ship)'

;
```

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

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

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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql

SELECT Landing_Outcome, COUNT(*) as 'Quantity', Date
from SPACEXTBL
where date between '04-06-2010' and '20-03-2017' and Landing_Outcome like '%Success%'
GROUP BY Landing_Outcome
ORDER BY 'Quantity'
;
```

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

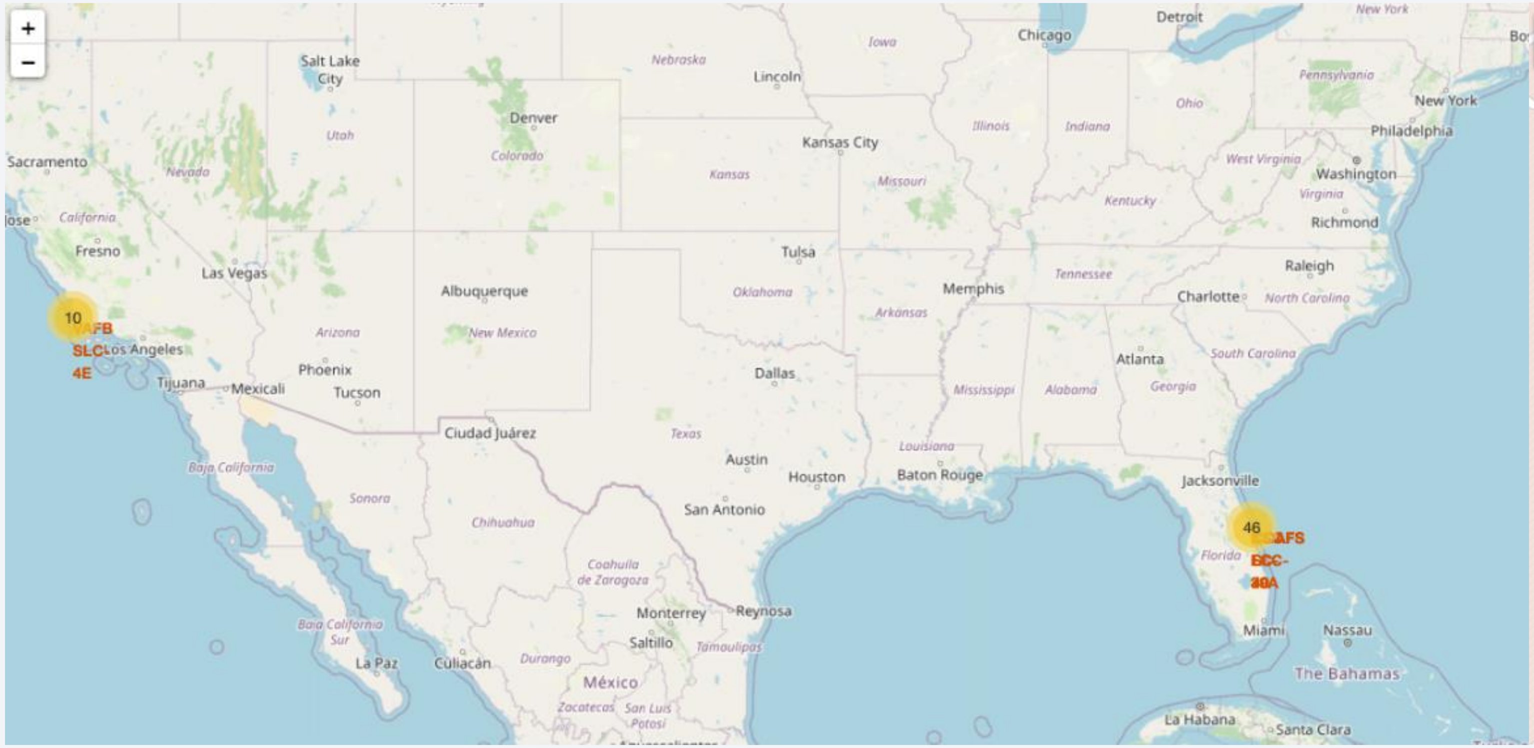
Landing_Outcome	Quantity	Date
Success	20	08/07/2018
Success (drone ship)	8	04/08/2016
Success (ground pad)	7	18/07/2016

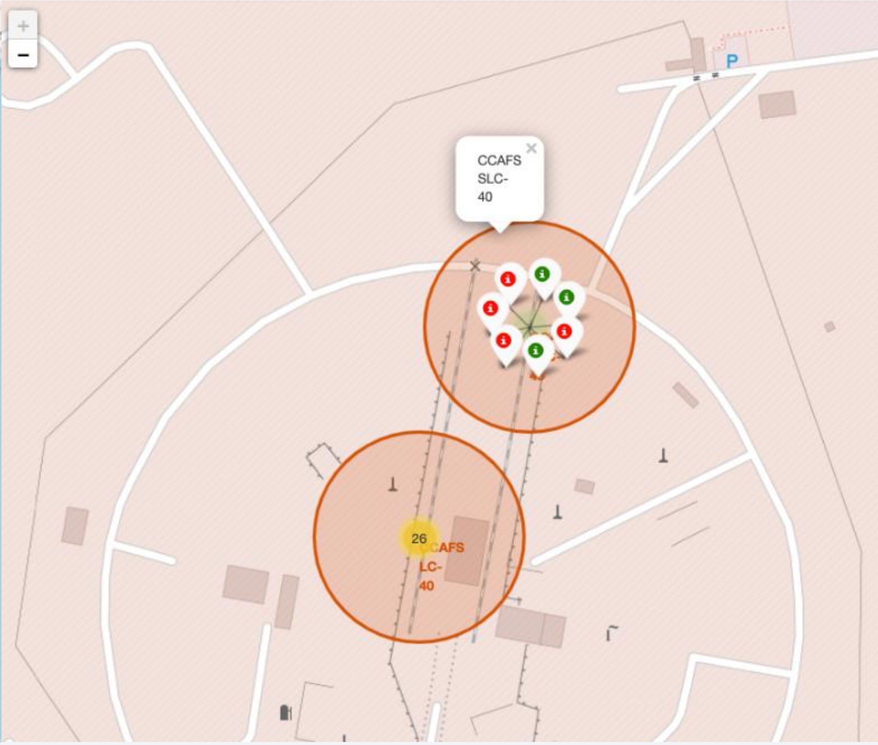
Section 3

Launch Sites Proximities Analysis

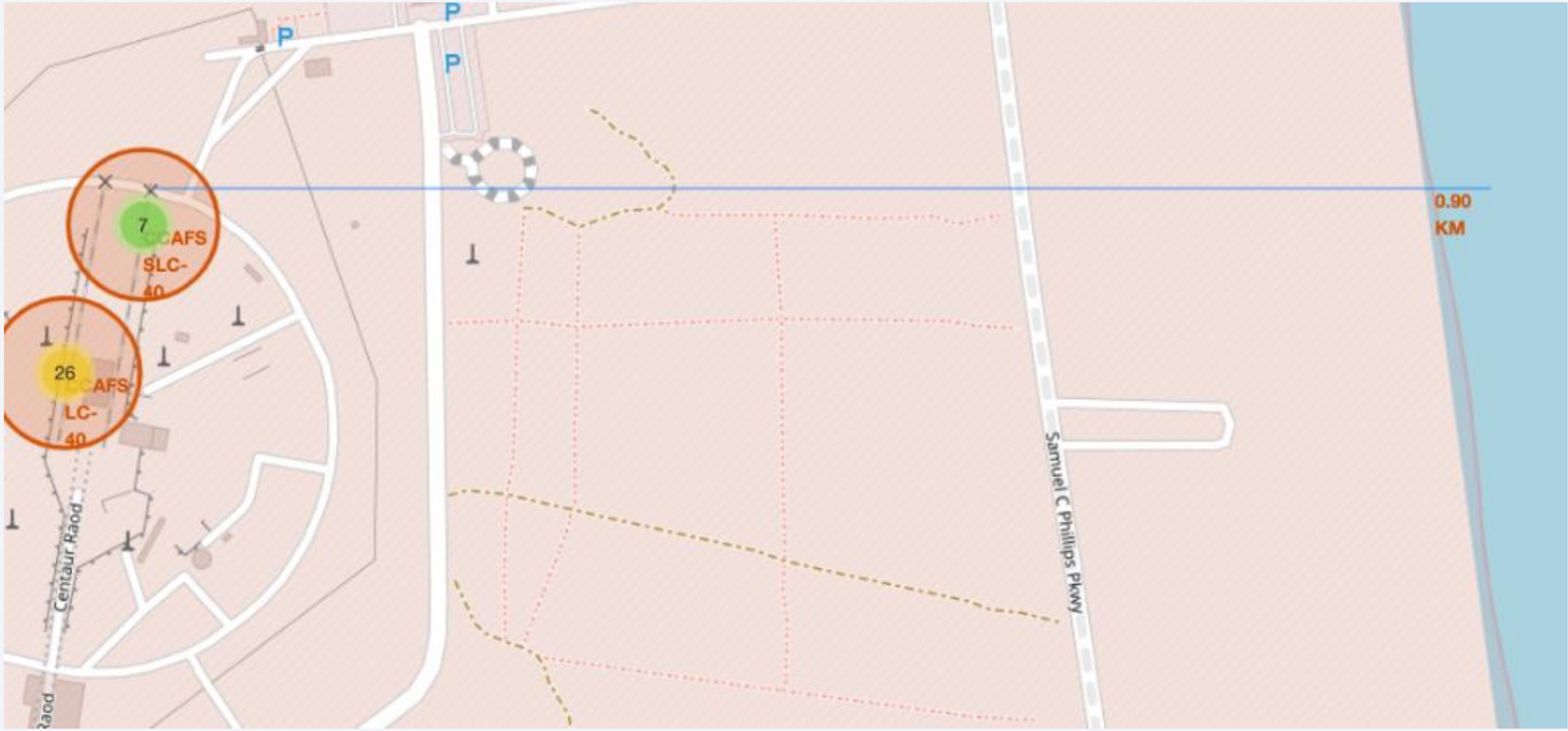
Launch Sites Locations

Launch Outcomes





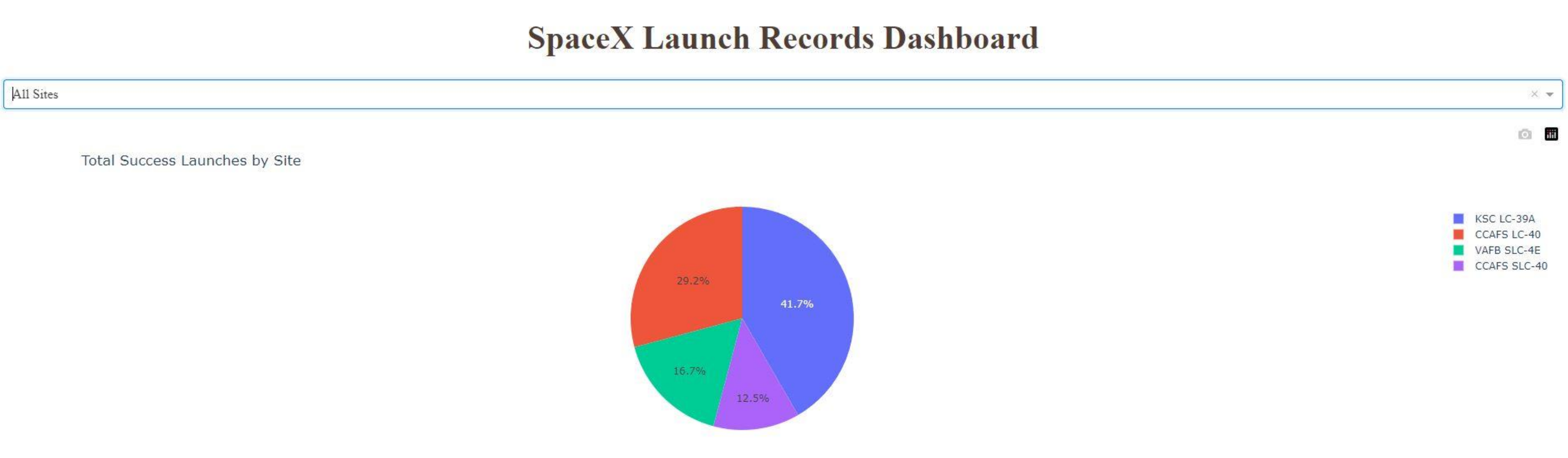
<Folium Map Screenshot 3>



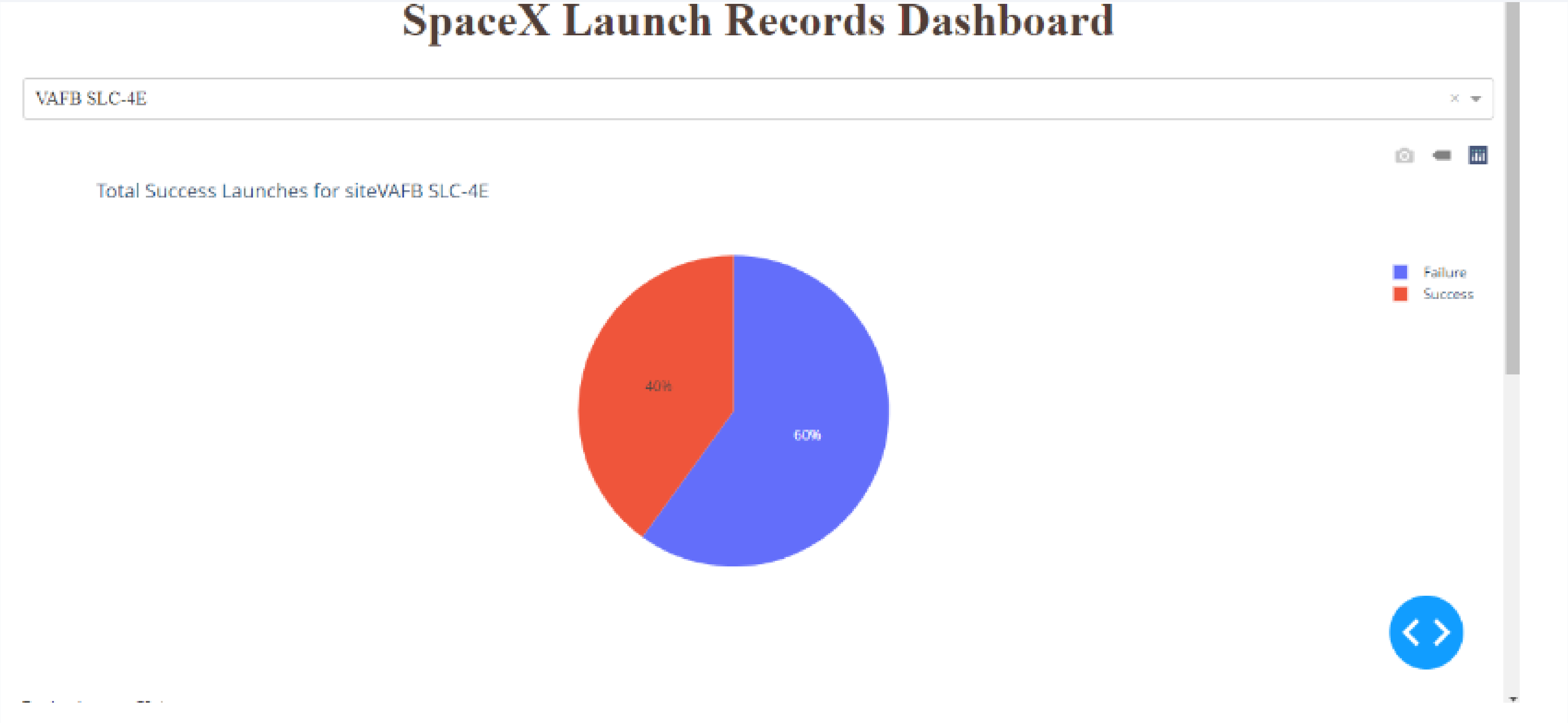
Section 4

Build a Dashboard with Plotly Dash

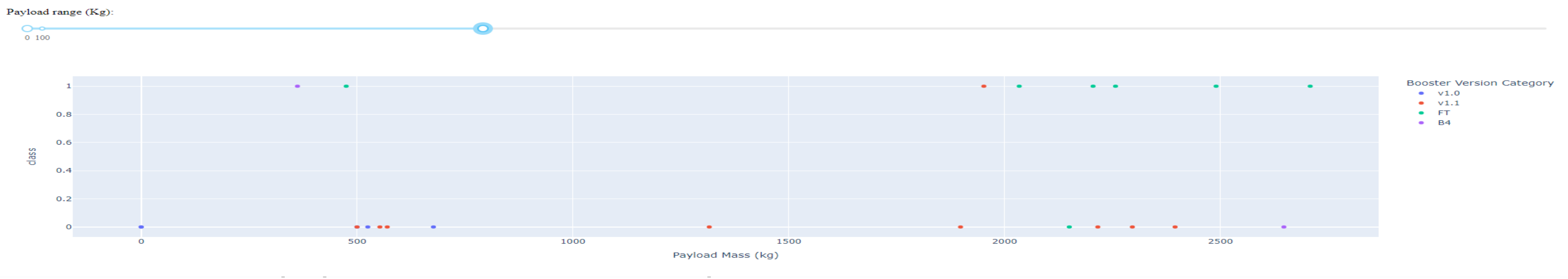
Total Success Launches by Site

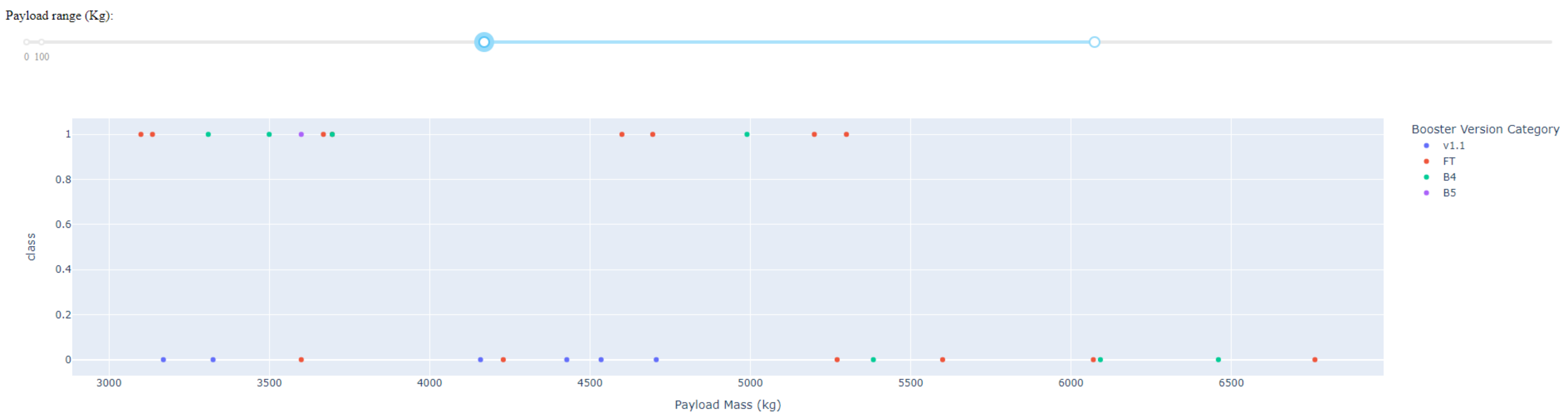


Success Ratio for VAFB SLC-4E Launch Site



<Dashboard Screenshot 3>



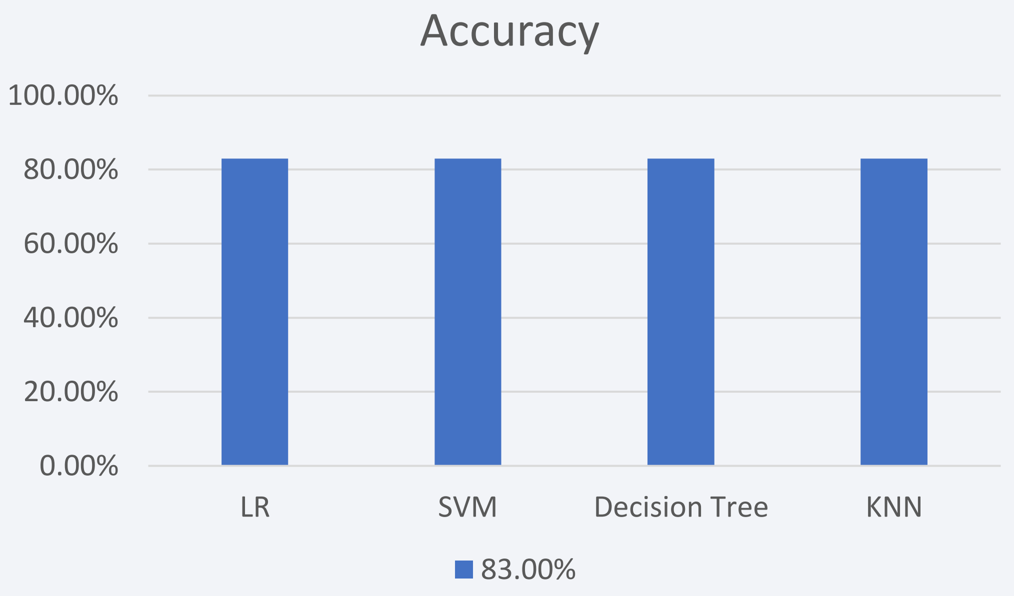


Section 5

Predictive Analysis (Classification)

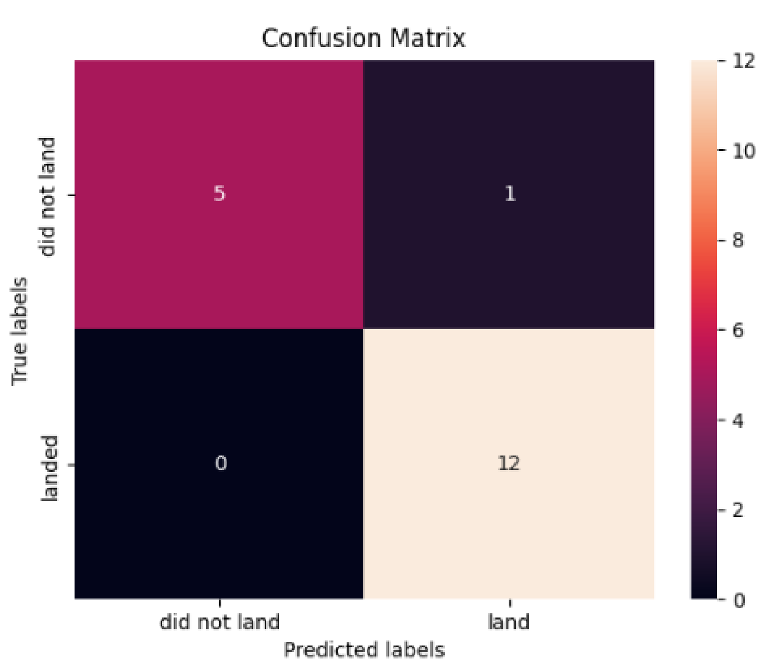
Classification Accuracy

- The classification models – LR, SVM, Decision Tree and KNN have the same accuracy of 83%



Confusion Matrix

- Decision Tree model performs better when dealing with false positive error



Conclusions

- As the flight number increases and the pay load is lighter, the first stage is more likely to land successfully.
- A better classification model to predict the success of a launch is a Decision Tree since it has a smaller false positive error and the accuracy of 83% (same accuracy of LR, SVM, KNN).
- KSC LC-39A and VAFB SLC 4E launch sites have a higher success ratio.
- The success rate since 2013 kept increasing till 2020

Thank you!