IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hélder G. V. Andrade
28/03/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection by API
  - Data collection with web scraping
  - Exploratory data with SQL
  - Exploratory data with data visualization
  - Machine learn prediction
- Summary of all results
  - Exploratory data analysis result
  - Predictive analytics result

Introduction

- Project background and context
  - Space X advertises Falcon 9 rocket launches on its website costing 62 million dollars.
  - The costs of its main competitors are considerably more expensive, part of the savings are due to the fact that.
  - Space X can reuse the first stage, but sometimes the landing of the first stage fails and it cannot be reused.
  - So, if we can determine whether the first stage will land, we can estimate the cost of a launch.
  - The goal of the project is to develop machine learning code to predict whether the first stage will be successful.
- Problems you want to find answers
  - What factors determine whether the rocket will land successfully?
  - Which of these factors should we work on to increase the chances of success?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceXAPI(https://api.spacexdata.com/v4/rockets/)Describe how data was collected
  - WebScraping
  - (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_l
  - aunches)
- 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
  - The collected data was normalized, divided into training and testing datasets and evaluated by four different models, with the accuracy of each model being evaluated using different combinations of parameters.

Data Collection

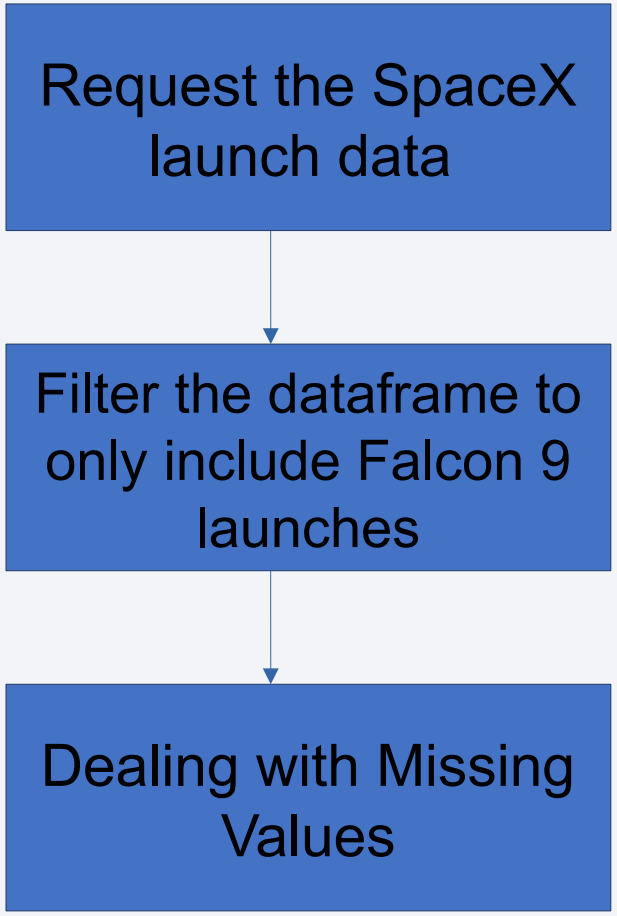
Describe how data sets were collected.

https://api.spacexdata.com/v4/rockets/

https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose
- https://github.com/he1der/Curso-basico/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

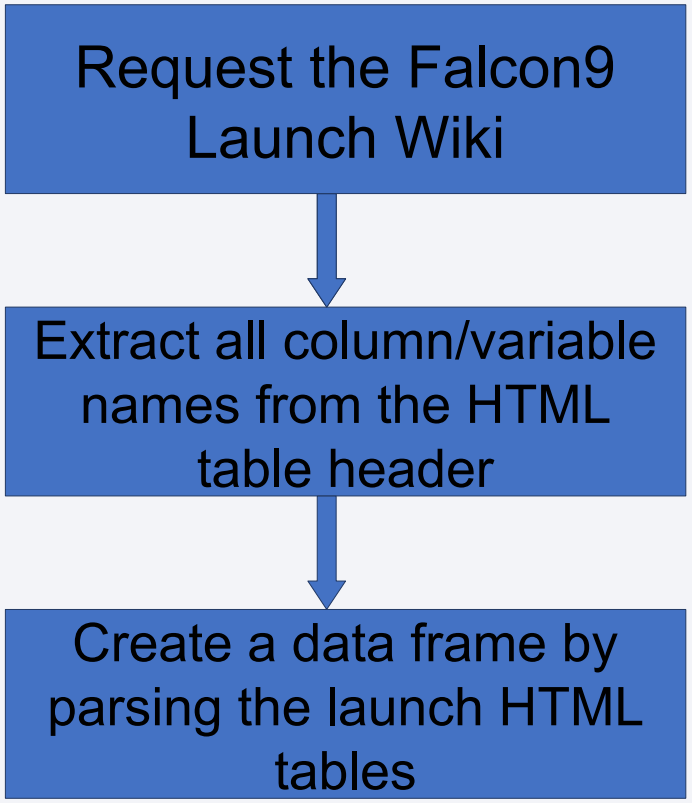


Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

. Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

https://github.com/he1der/Curso-basico/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

Describe how data were processed
You need to present your data wrangling process using key phrases and flowcharts
Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

https://github.com/he1der/Curso-basico/blob/main/labs-jupyter-spacex-Data%20wrangling(1).ipynb

EDA with Data Visualization

Summarize what charts were plotted and why you used those charts

Bar plot, acater plot and line plot

Show the relationship between success rate of each orbit type;

Show the relationship between FlightNumber and Orbit type;

Show the relationship between Payload and Orbit type;

Show the launch success yearly trend.

Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
https://github.com/he1der/Curso-basico/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Using bullet point format, summarize the SQL queries you performed

. Names of the unique launch sites in the space mission

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

. The total payload mass carried by boosters launched by NASA (CRS)

. Average payload mass carried by booster version F9 v1.1

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

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

. The total number of successful and failure mission outcomes

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

. 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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

https://github.com/he1der/Curso-basico/blob/7d9abc56c4e7fa1955095a8eccec5611d448a87b/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

. Markers to indicate points like launch sites

. Circles to indicate highlighted area around specifc cordinates

. Lines to indicate distances between two cordinates

Explain why you added those objects

Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

https://github.com/he1der/Curso-basico/blob/7d9abc56c4e7fa1955095a8eccec5611d448a87b/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

. Mark all launch sites on a map

. Mark the success/failed launches for each site on the map

. Calculate the distances between a launch site to its proximities

Explain why you added those plots and interactions

Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

lab_jupyter_launch_site_location.jupyterlite.ipynb

Predictive Analysis (Classification)

Summarize how you built, evaluated, improved, and found the best performing classification model

. I read the data from SpaceX launches, processed the data, divided the set into training and testing

. I trained different machine learning algorithms with different hyperparameters

. And used accuracy to determine the best model

You need present your model development process using key phrases and flowchart

Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

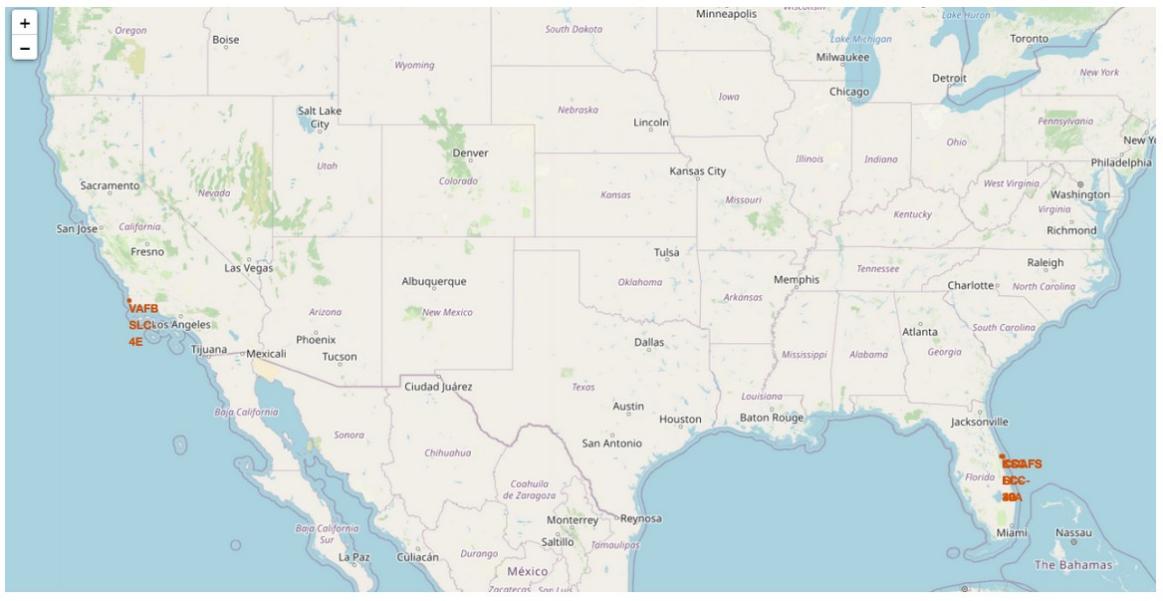
https://github.com/he1der/Curso-basico/blob/7d9abc56c4e7fa1955095a8eccec5611d448a87b/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
  - Space x uses differents launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
  - The total payload mass carried by boosters launched by NASA (CRS) was, 45.596 kg
  - The average payload mass carried by booster version F9 v1.1 was, 2534.666 kg
  - The first succesful landing outcome in ground pad was in 2018-07-22

Results

- Interactive analytics demo in screenshots



Results

- Predictive analysis results
  - Log Regression test acuracy: 0.8333333333333334
  - SVM test acuracy: 0.8333333333333334
  - tree test acuracy: 0.7777777777777778
  - KNN test acuracy: 0.8333333333333334
  -

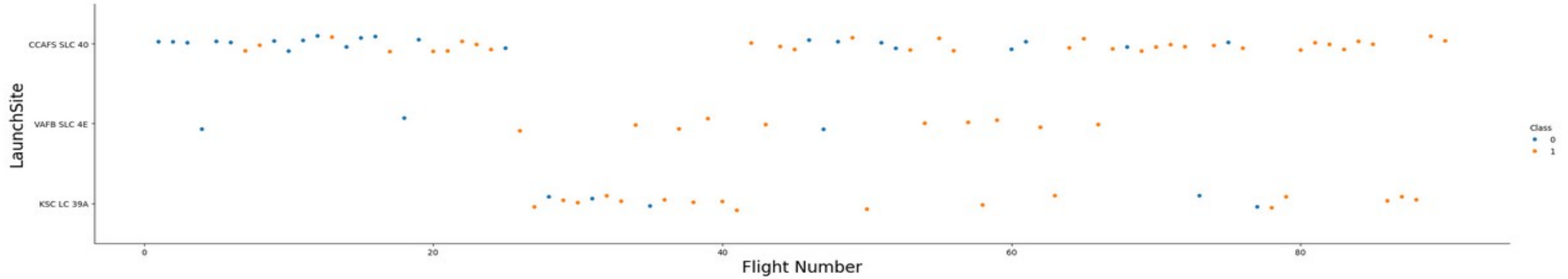
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site
- Show the screenshot of the scatter plot with explanations

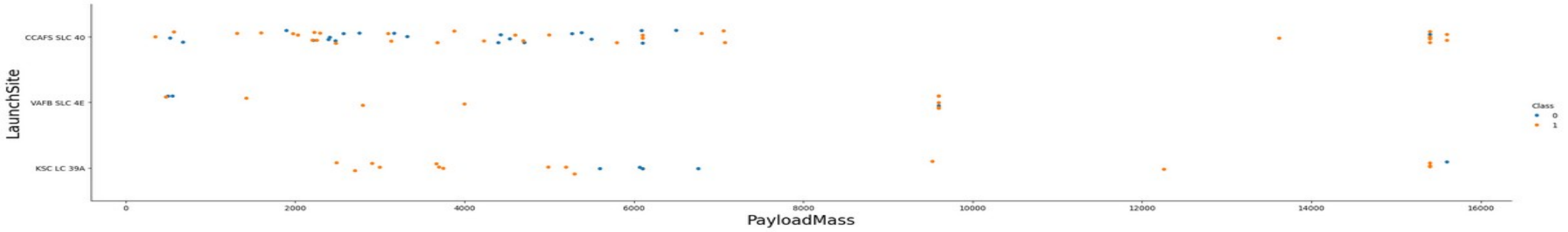
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hu
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```



Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site
- Show the screenshot of the scatter plot with explanations

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, a
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
plt.show()
```

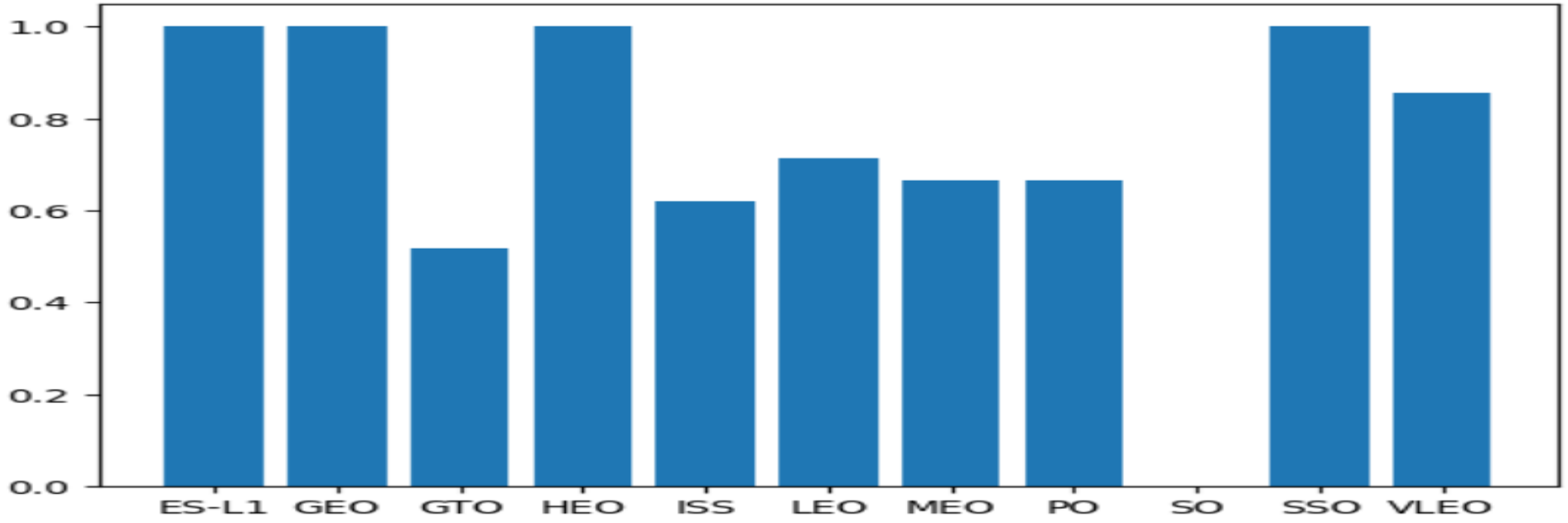


Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations

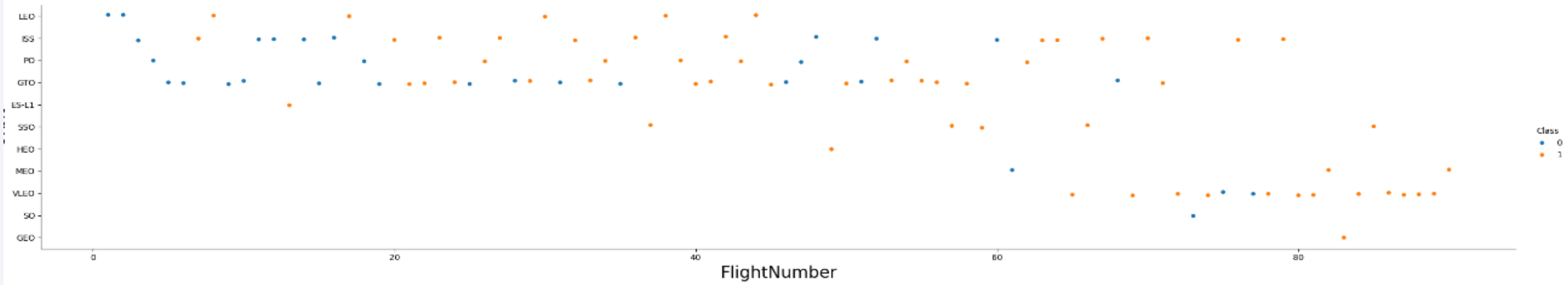
```
# HINT use groupby method on Orbit column and get the mean of Class column
orbit = df[['Orbit','Class']].groupby('Orbit').mean()

plt.bar(orbit.index.values, orbit['Class'])

plt.show()
```



Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type
- Show the screenshot of the scatter plot with explanations

```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the cla
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

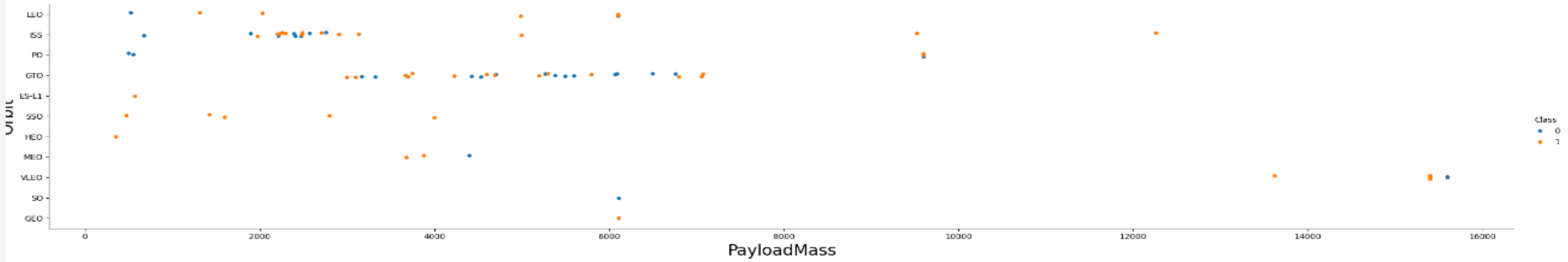


You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type
- Show the screenshot of the scatter plot with explanations

```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class va
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



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

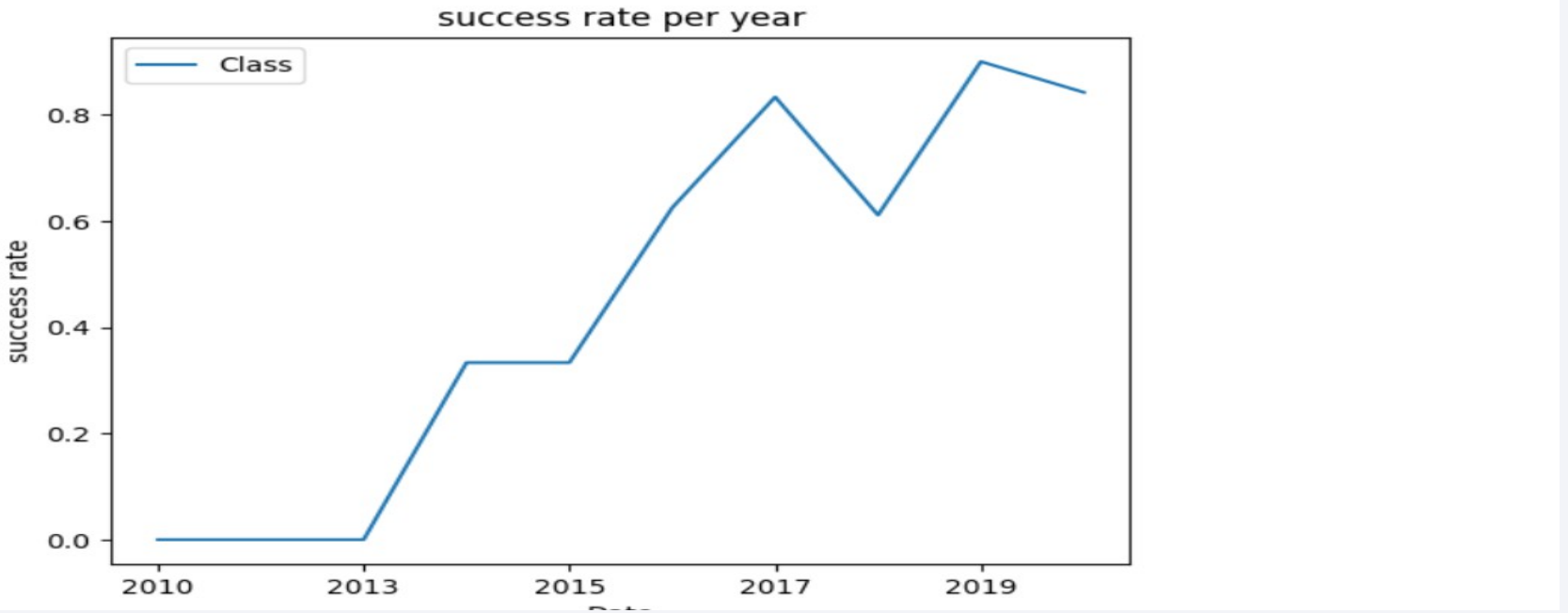
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
df[['Class','Date']].groupby('Date').mean().plot()
plt.title('success rate per year')
plt.ylabel('success rate')
plt.xlabel('Date')
```

```
Text(0.5, 0, 'Date')
```



All Launch Site Names

Find the names of the unique launch sites

Present your query result with a short explanation here

```
%sql select distinct  Launch_Site from SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`

Present your query result with a short explanation here

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

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

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

Total Payload Mass

Calculate the total payload carried by boosters from NASA

Present your query result with a short explanation here

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

Present your query result with a short explanation here

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE 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

Present your query result with a short explanation here

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success'
```

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

min(Date)
2018-07-22

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

Present your query result with a short explanation here

```
%sql select distinct Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success' and PAYLOAD_MASS__KG_ betv
```

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

Booster_Version
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

Present your query result with a short explanation here

```
%sql select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
```

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

Mission_Outcome	count(*)
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

Present your query result with a short explanation here

```
%sql select distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) fr
```

```
 * 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

Present your query result with a short explanation here

```
%sql select substr(Date, 6,2) as month, Booster_version, Launch_site, Landing_Outcome from SPACEXTABLE where Land
```

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

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

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

Present your query result with a short explanation here

```
%sql select Landing_Outcome, count(*) from SPACEXTABLE where Landing_Outcome in ('Failure (drone ship)', 'Success
```

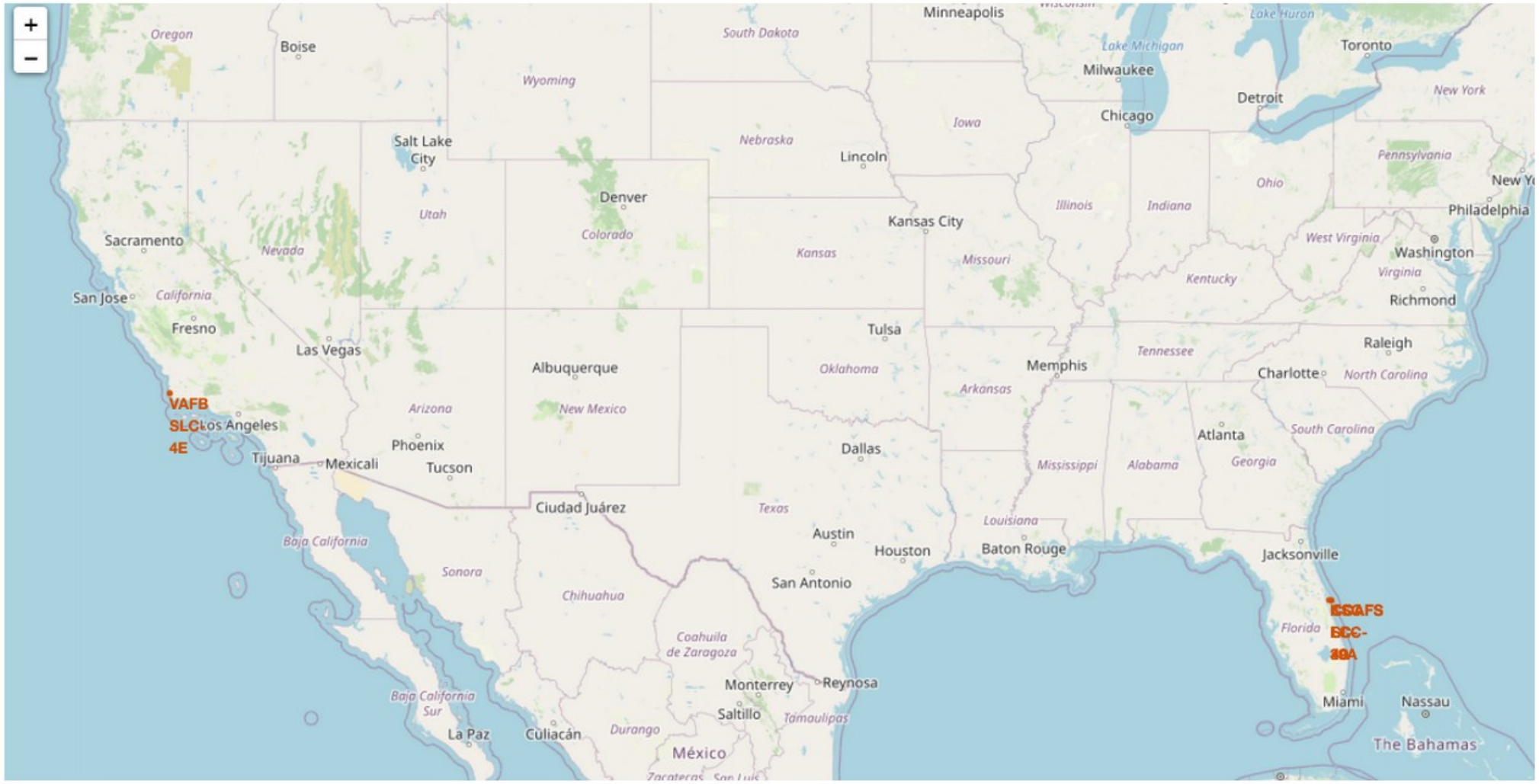
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count(*)
Failure (drone ship)	5
Success (ground pad)	3

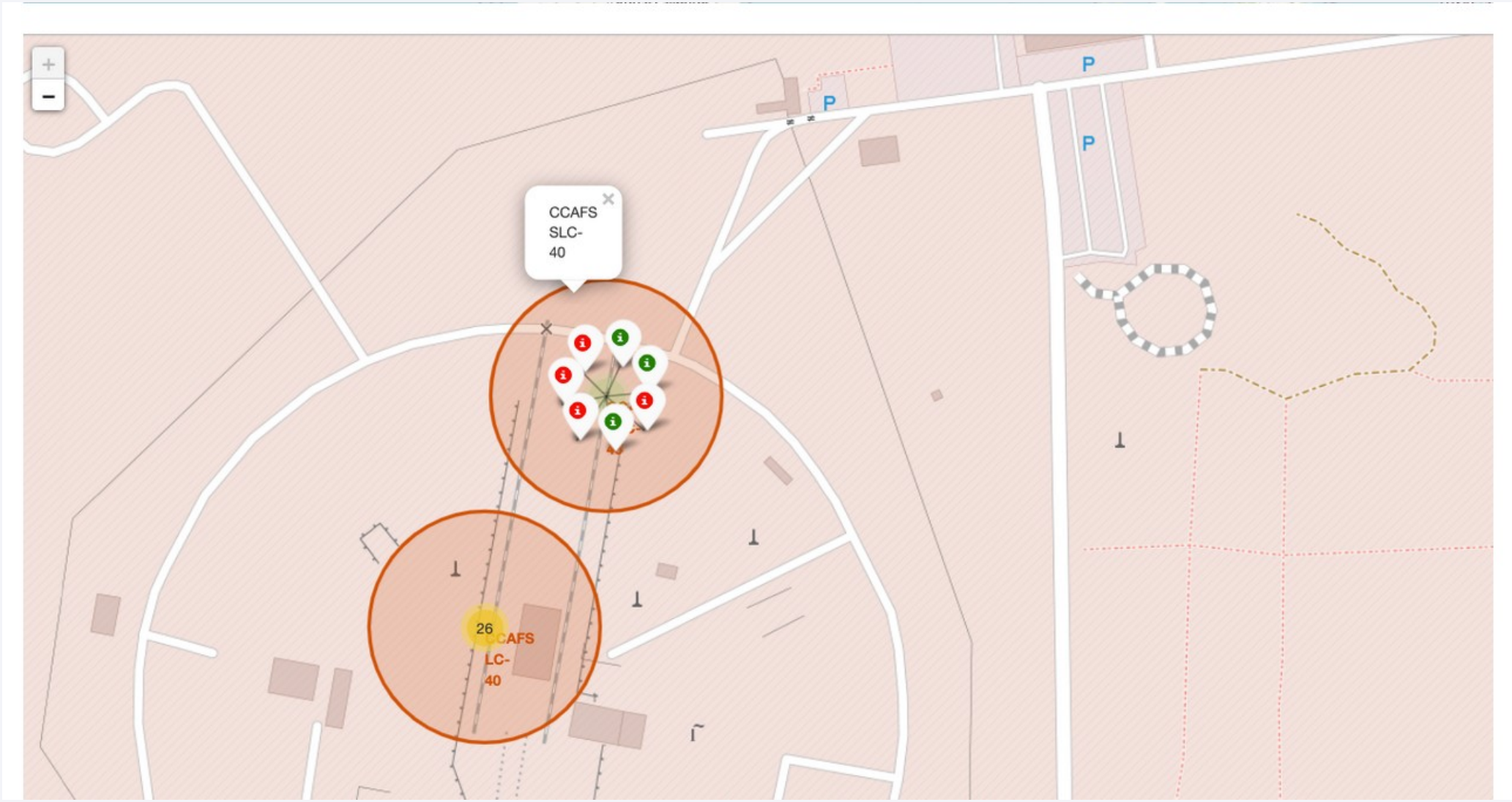
Section 3

Launch Sites Proximities Analysis

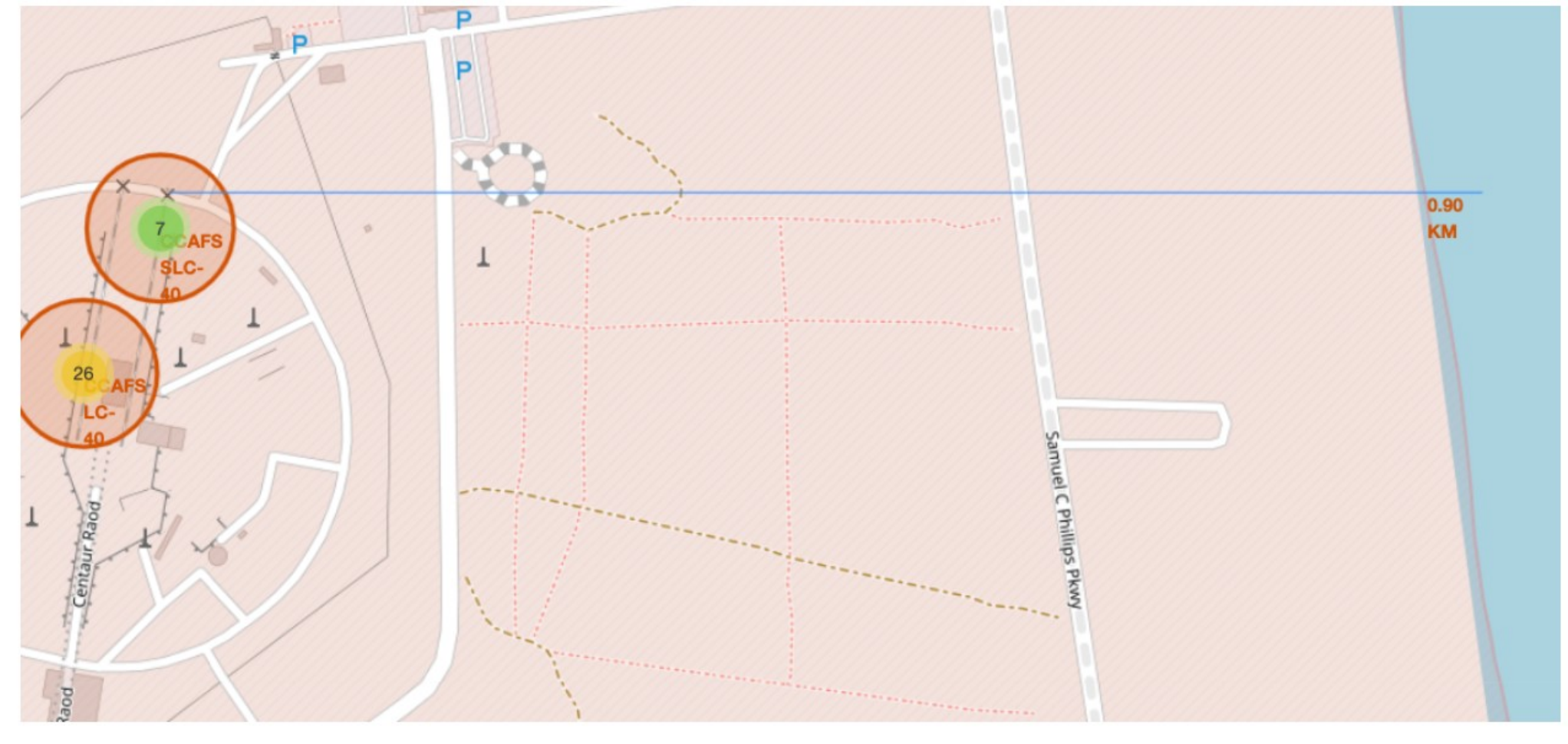
Folium Map marked launch sites

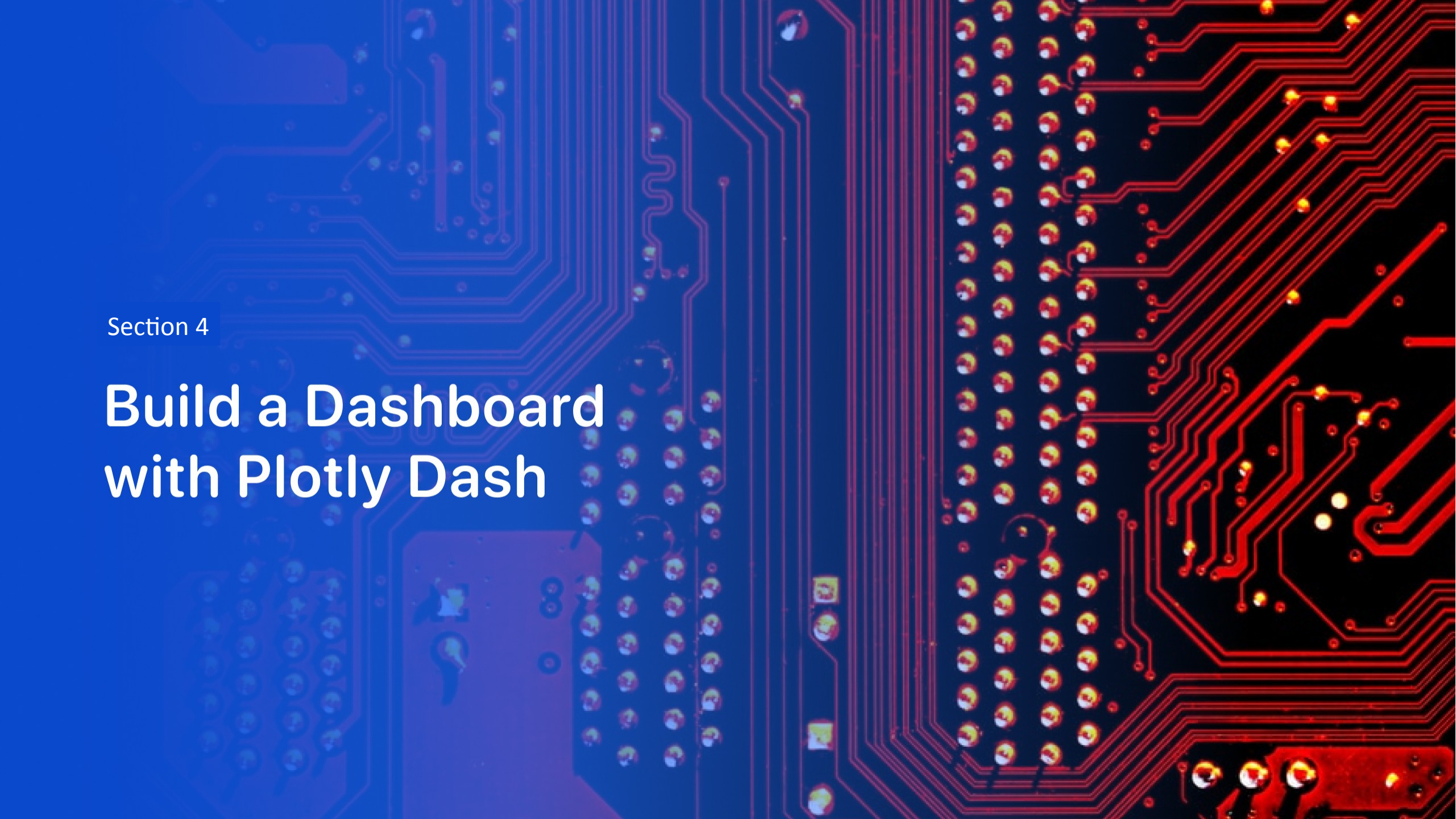


Folium Map launch sites have relatively high success rates.



Folium Map PolyLine between a launch site to the selected coastline point




Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

Replace <Dashboard screenshot 1> title with an appropriate title

Show the screenshot of launch success count for all sites, in a piechart

Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

Replace <Dashboard screenshot 2> title with an appropriate title

Show the screenshot of the piechart for the launch site with highest launch success ratio

Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

Replace <Dashboard screenshot 3> title with an appropriate title

Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

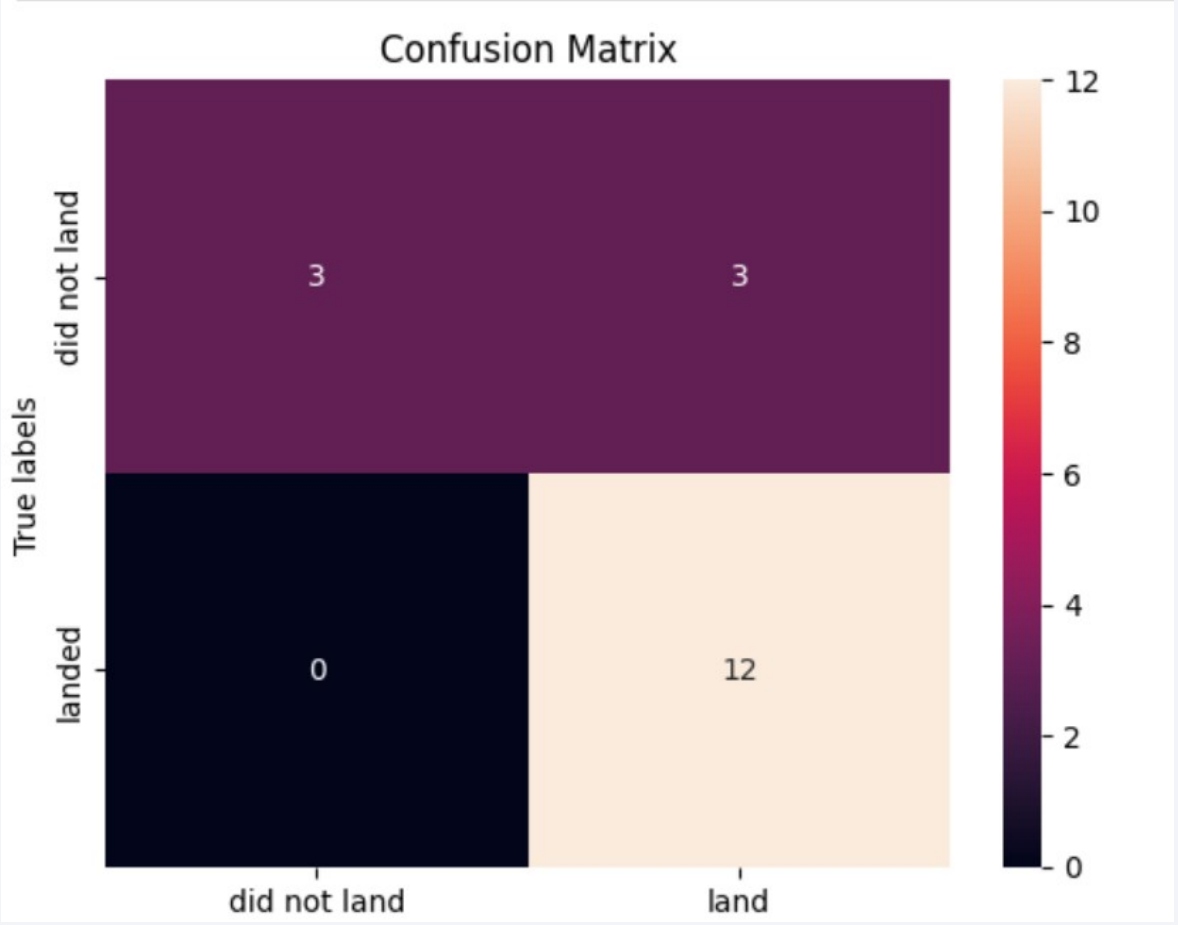
Predictive Analysis (Classification)

Classification Accuracy

- Log Regression test acuracy: 0.8333333333333334
- SVM test acuracy: 0.8333333333333334
- tree test acuracy: 0.7777777777777778
- KNN test acuracy: 0.8333333333333334

Confusion Matrix

```
yhat=logrec_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- The tree algorithm is the best machine learning for this data set
- KSC LC-39A have the most successful launches of every sites
- The low weighted payloads performed better than heavy payloads

Thank you!