IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mikhail Krylov

22/03/2023

Outline

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

Executive Summary

The cost of the launch of the Falcon 9 rocket depends on the success or failure of the landing of the reuse of the first stage. The reuse of first stage allows to significantly decrease the cost of the launch and go ahead of competitors. However the ability of reuse of the first stage depends on its successful landing, which is not always the case. To better predict the costs of the flight it is important to be able to predict if the landing of the first stage would be successful.

In this report the summary of the research is presented. The research itself included historical data gathering and analysis. Based on this analysis the key factors that could affect the success of the landing process were determined. Based on the data several machine learning algorithms were trained including Logistic Regression, support vector machine, decision tree classifier ,and KNN. All models were trained with different set of parameters to find the best setting.
The models that we trained showed similar results (according to the confusion matric) with the score 0.833(3)

Introduction

The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX. SpaceX’s accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars. Other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse

The first stage does most of the work of getting the load to the orbit. This stage is quite large and expensive. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes the first stage does not land and it will crash. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer. Our task in this project is to determine the price of each launch. To do that we gather information about Space X and determine if SpaceX will reuse the first stage. In this project will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - To collect the data we used web scraping from publicly available data from wiki page
  - We also collected data using API
- Perform data wrangling
  - After data was scraped and put into data frame the data type for each column was checked and analysis against null values was performed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

To collect data we use several methods, including web scraping and use of SpaceX API

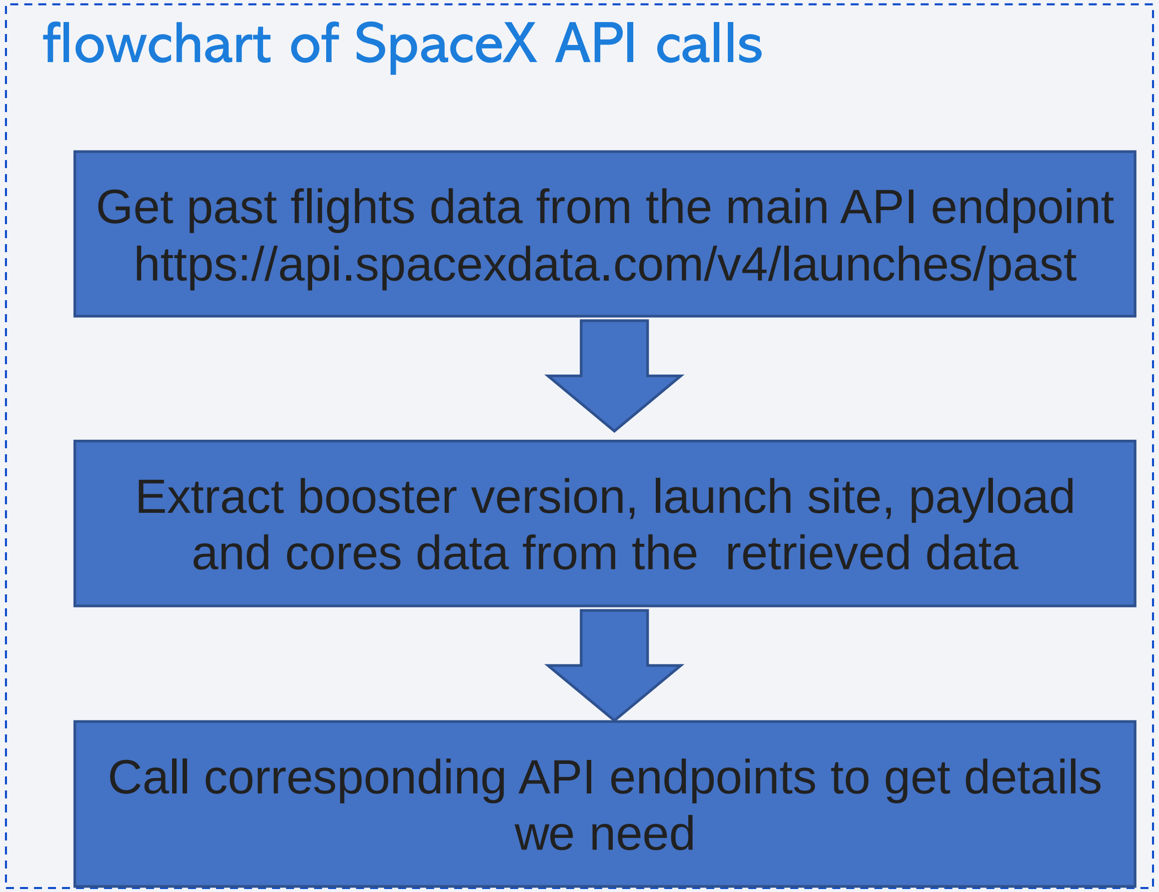
To collect data web scraped web page and used python beautiful soup library to parse the data, then the data e=was processed and put into the data frame

while using API provided by SpaceX several calls were made to extract the data we needed including booster version, launch site, payload mass, and version of core

Data Collection – SpaceX API

- To get data the API provided by SpaceX company was used. The process is displayed on the flowchart

- The completed notebook is available via link: https://github.com/krylov-mihail/capstone-ibm/blob/master/data%20import%20notebook.ipynb



flowchart of SpaceX API calls

Get past flights data from the main API endpoint https://api.spacexdata.com/v4/launches/past

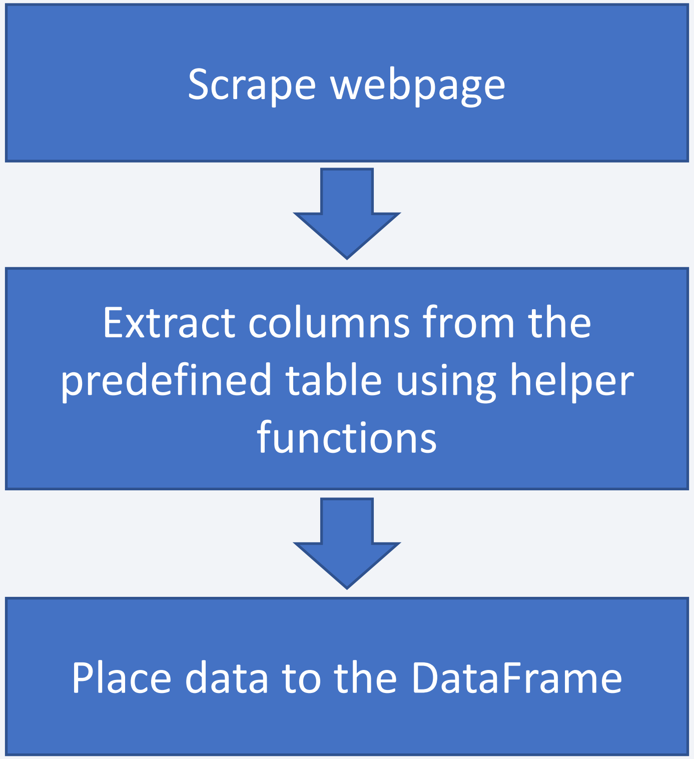
Extract booster version, launch site, payload and cores data from the retrieved data

Call corresponding API endpoints to get details we need

Data Collection - Scraping

- The notebook is available via link https://github.com/krylov-mihail/capstone-ibm/blob/master/jupyter-labs-webscraping.ipynb

Flowchart of web scraping process



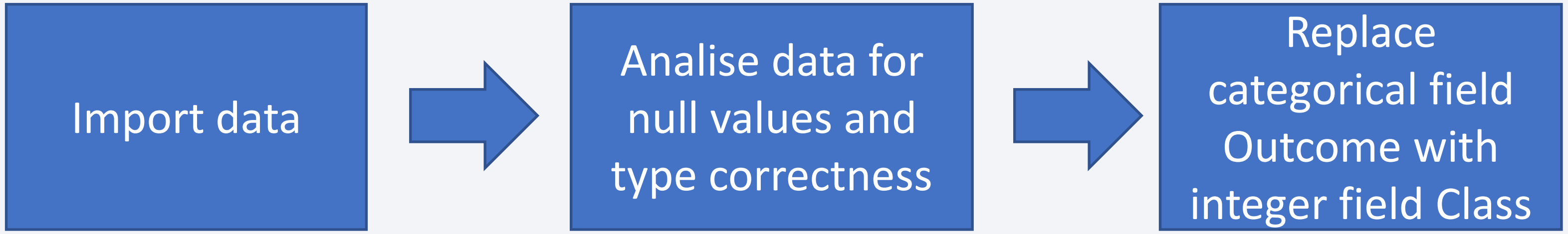
Scrape webpage

Extract columns from the predefined table using helper functions

Place data to the DataFrame

Data Wrangling

- The main goal of the data wrangling process we undertook was about generating a field Class with 2 possible outcomes – failure and success. In the initial data set there were different types of outcome and we reduced them to the targeted set of 0 and 1 and we also converted the type to integer



Import data

Analise data for null values and type correctness

Replace categorical field Outcome with integer field Class

- The notebook with data wrangling is available via link https://github.com/krylov-mihail/capstone-ibm/blob/master/Labs%20Spacex%20Data%20wrangling.ipynb

EDA with Data Visualization

To better understand the influence of the features on the outcome the EDA was performed for following pairs of data

- Payload Mass / Flight Number - to see how the payload change over time
- Launch Site / Flight Number - to see how locations changed over time
- Launch Site / Payload Mass – to see if there is a specialization between launch sited by payload mass
- Visualize the relationship between success rate of each orbit type
- Visualize the relationship between FlightNumber and Orbit type
- Visualize the relationship between Payload and Orbit type
- Visualize the launch success yearly trend – to see the improvement of the outcome

The notebook with EDA using viaual plots is available via link https://github.com/krylov-mihail/capstone-ibm/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

summary the SQL queries performed

- select DISTINCT launch_site from spacex
- select * from spacex where launch_site like 'CCA%' LIMIT 5
- select SUM(payload_mass__kg_) from spacex where customer = 'NASA (CRS)’
- select AVG(payload_mass__kg_) from spacex where booster_version = 'F9 v1.1’
- select min(DATE) from spacex where landing__outcome = 'Success (ground pad)’
- select distinct booster_version from spacex where landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ between 4000 AND 6000
- select count(mission_outcome), mission_outcome   from spacex group by mission_outcome
- select booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
- select landing__outcome, booster_version, launch_site, DATE from spacex where landing__outcome = 'Failure (drone ship)' and YEAR(DATE)=2015
- select count(landing__outcome), landing__outcome from spacex where DATE BETWEEN '2010-06-04 ' and '2017-03-20' GROUP BY landing__outcome ORDER BY count(landing__outcome) DESC

- Notebook is available via link https://github.com/krylov-mihail/capstone-ibm/blob/master/eda-sql.ipynb

Build an Interactive Map with Folium

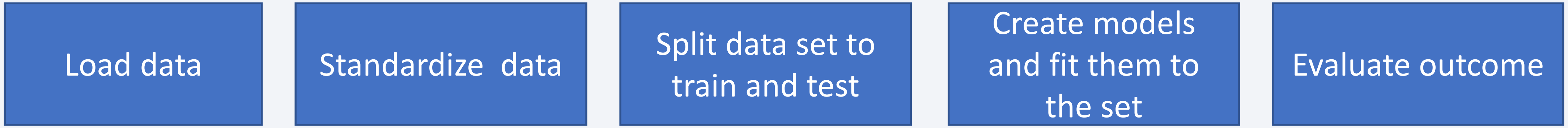
- We added coloured markers to show the failure/success events , circles to show locations of the launch sites, lines to draw the lines between launch sites and coast lines,
- Those object were added to help visualize the special data on the map and draw conclusions about the launch site location features
- The notebook is available via link: https://github.com/krylov-mihail/capstone-ibm/blob/master/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- To better understand relation ship between the number of success launches and launch place as well as the connection between the mass of the load and outcome of the mission two interactive plots were built. Both plots allow view the data for a specific launch site or across all launch sites.
- Here is a link to a dashboard file https://github.com/krylov-mihail/capstone-ibm/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- To predict the mission outcome with the Machine learning several models were trained, including Logistic Regression, support vector machine, decision tree classifier ,and KNN. For all models the GridSearchCV was used to fine tune the model and find the best set of the parameters. The undertaken steps are shown at the diagram:



Load data

Standardize data

Split data set to train and test

Create models and fit them to the set

Evaluate outcome

- The notebook with the results is available via link: https://github.com/krylov-mihail/capstone-ibm/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

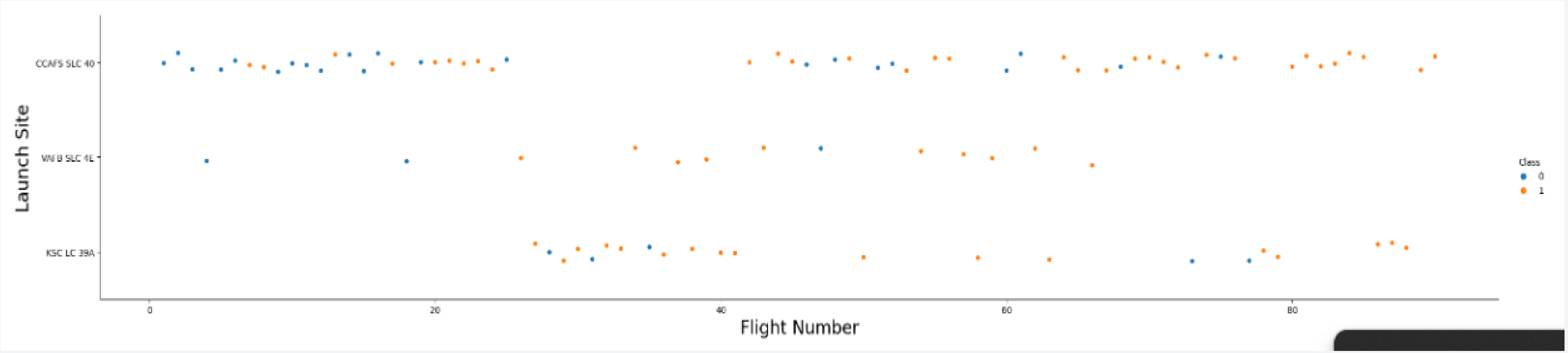
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

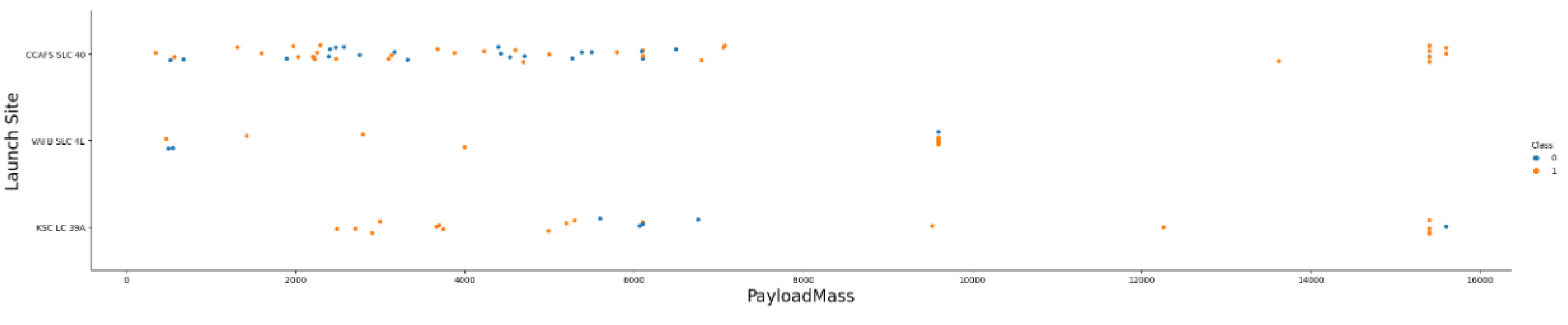
- Scatter plot of Flight Number vs. Launch Site



- We can see that early launches were mostly made from CCAFS SLC 40 launching site and they had higher failure rate

Payload vs. Launch Site

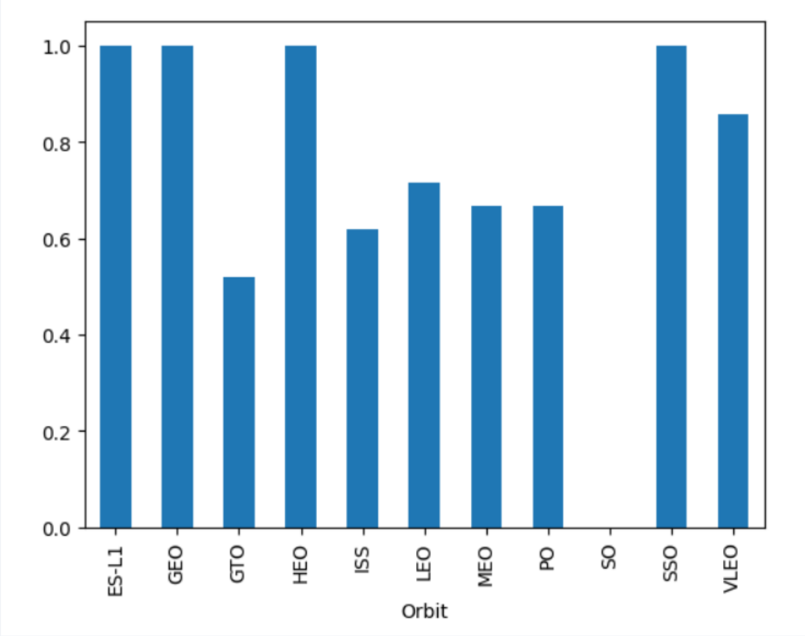
- Scatter plot of Payload vs. Launch Site



- From the plot we can see that heavy-mass loads were launched from CCAFS SLC 40 and KSC LC39A, while middle-mass loads were lunched from VAFB SLC 4E. Small-weight loads were launched from all three launch sites

Success Rate vs. Orbit Type

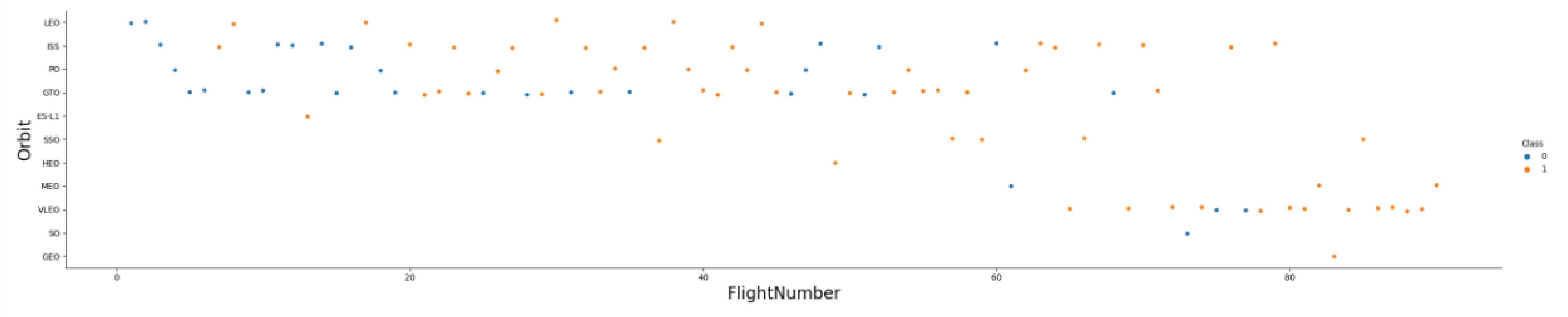
- A bar chart for the success rate of each orbit type



- We can see that for of the target orbits have success rate of 100% (ES-L1, GEO, HEO and SSO)

Flight Number vs. Orbit Type

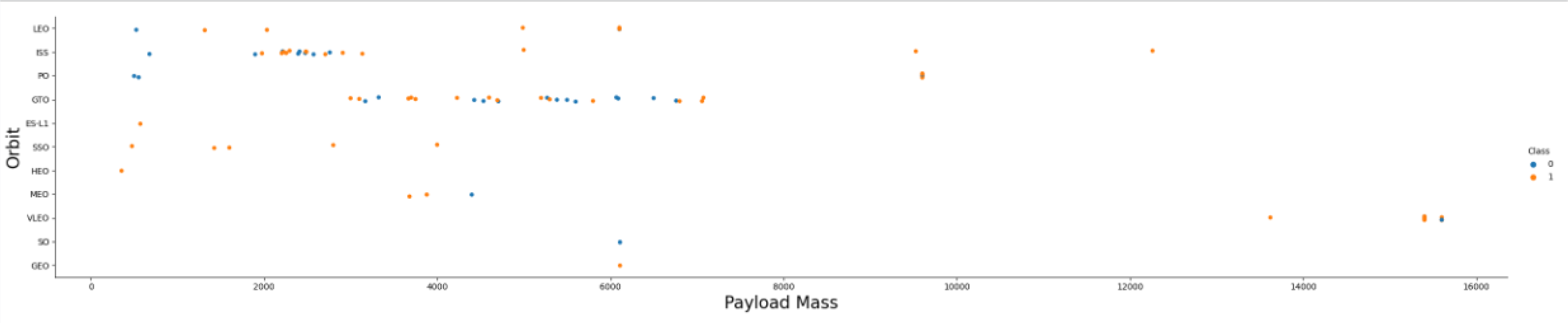
- A scatter point of Flight number vs. Orbit type



- From the scatter plot we can see that 100% success rate that we saw on previous bar chart can be explained by the low total number of launches. Also the plot clear states that launches some of the orbits (VLEO, SO) appeared later than others

Payload vs. Orbit Type

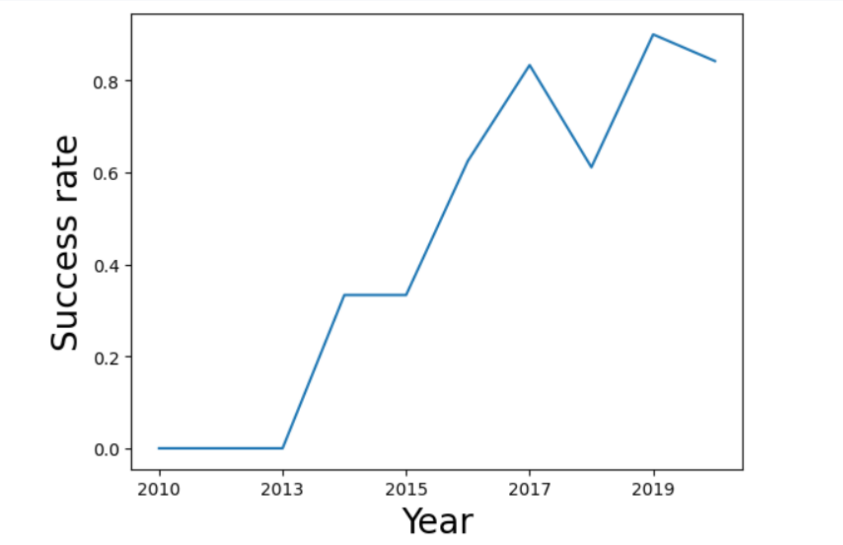
- scatter point of payload vs. orbit type



- From the scatter plot we can wee that there is a dependency between the orbit and the payload and the orbit type. The heaviest loads are sent to VLEO orbit

Launch Success Yearly Trend

- a line chart of yearly average success rate

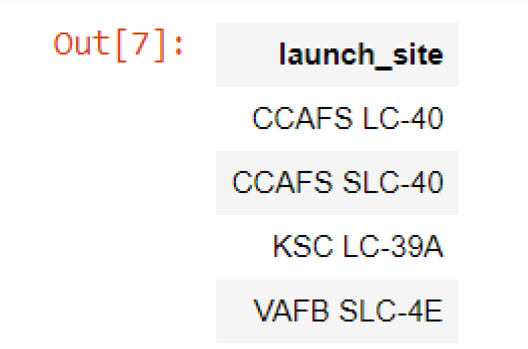


- From the chart above we can see that starting from 2013 the success outcome has improved

All Launch Site Names

- To get unique launching sites following query can be used
  select DISTINCT launch_site from spacex

  the result is shown on te following screenshot:



Out[7]:

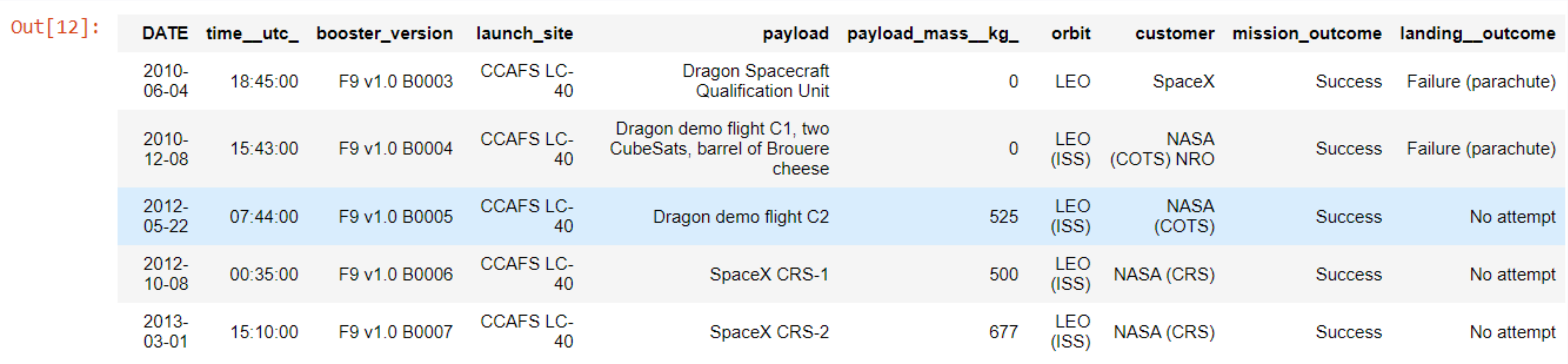
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- To Find 5 records where launch sites begin with `CCA` following query can be used

select * from spacex where launch_site like 'CCA%' LIMIT 5

- The result of the query can be seen at the screenshot below



Out[12]:

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

- To Calculate the total payload carried by boosters from NASA following query can be used

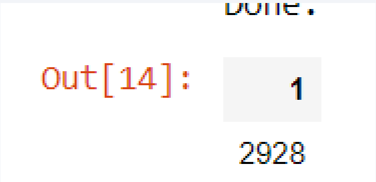
```
select SUM(payload_mass__kg_) from spacex  where customer = 'NASA (CRS)’
```

- The total payload mass is 45596 kg

Average Payload Mass by F9 v1.1

- To Calculate the average payload mass carried by booster version F9 v1.1 following query can be used:
- select AVG(payload_mass__kg_) from spacex where booster_version = 'F9 v1.1'

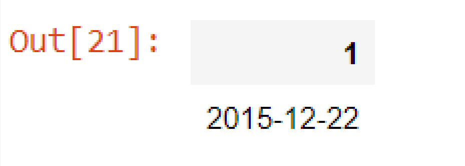
- The average payload equals 2928 kg as can be seen at the screenshot



```
         Done.
Out[14]:     1
          2928
```

First Successful Ground Landing Date

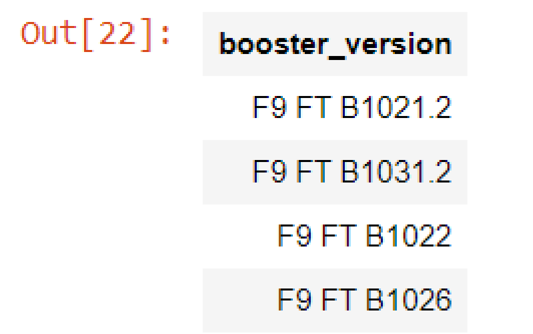
- To Find the dates of the first successful landing outcome on ground pad can be found using following query
- select min(DATE) from spacex where landing__outcome = 'Success (ground pad)’
- The result is presened on the screenshot below:



```
Out[21]:          1
         2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- To List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 following query can be used:
- select distinct booster_version from spacex where landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ between 4000 AND 6000
- The list of boosters is shown on the screenshot below



Out[22]:

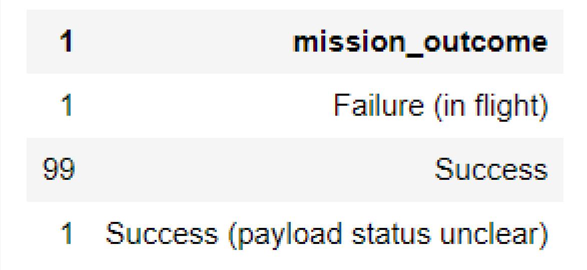
booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- To Calculate the total number of successful and failure mission outcomes following query can be used:

select count(mission_outcome), mission_outcome   from spacex group by mission_outcome

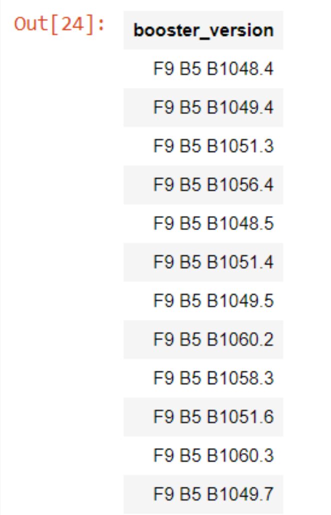
- Basically only one mission out of 100 failed



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

Boosters Carried Maximum Payload

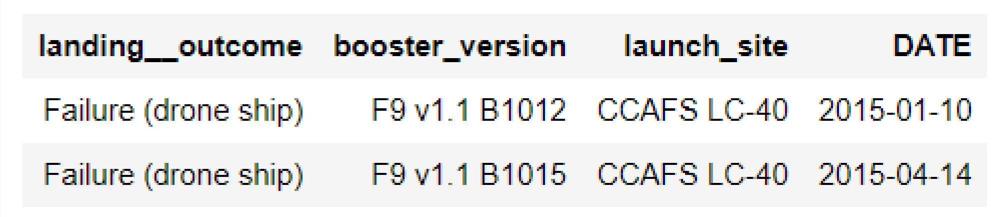
- We can find the List of the names of the booster which have carried the maximum payload mass using following query
- select booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
- From screen shot below we can see that different boosters were used



```
Out[24]:  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

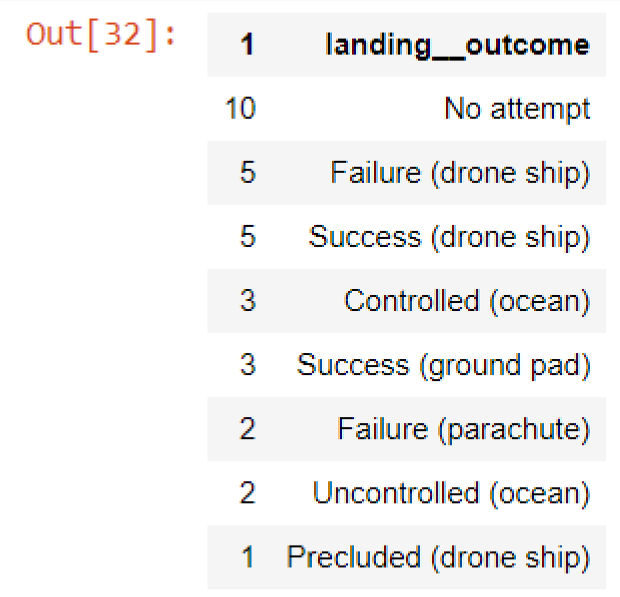
- To List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 we can use following query: select landing__outcome, booster_version, launch_site, DATE from spacex where landing__outcome = 'Failure (drone ship)' and YEAR(DATE)
- The result is shown on the screenshot



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

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

- Ranking of 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 can be done using the query
- select count(landing__outcome), landing__outcome from spacex where DATE BETWEEN '2010-06-04 ' and '2017-03-20' GROUP BY landing__outcome ORDER BY count(landing__outcome) DESC
- The results are shown on the print screen



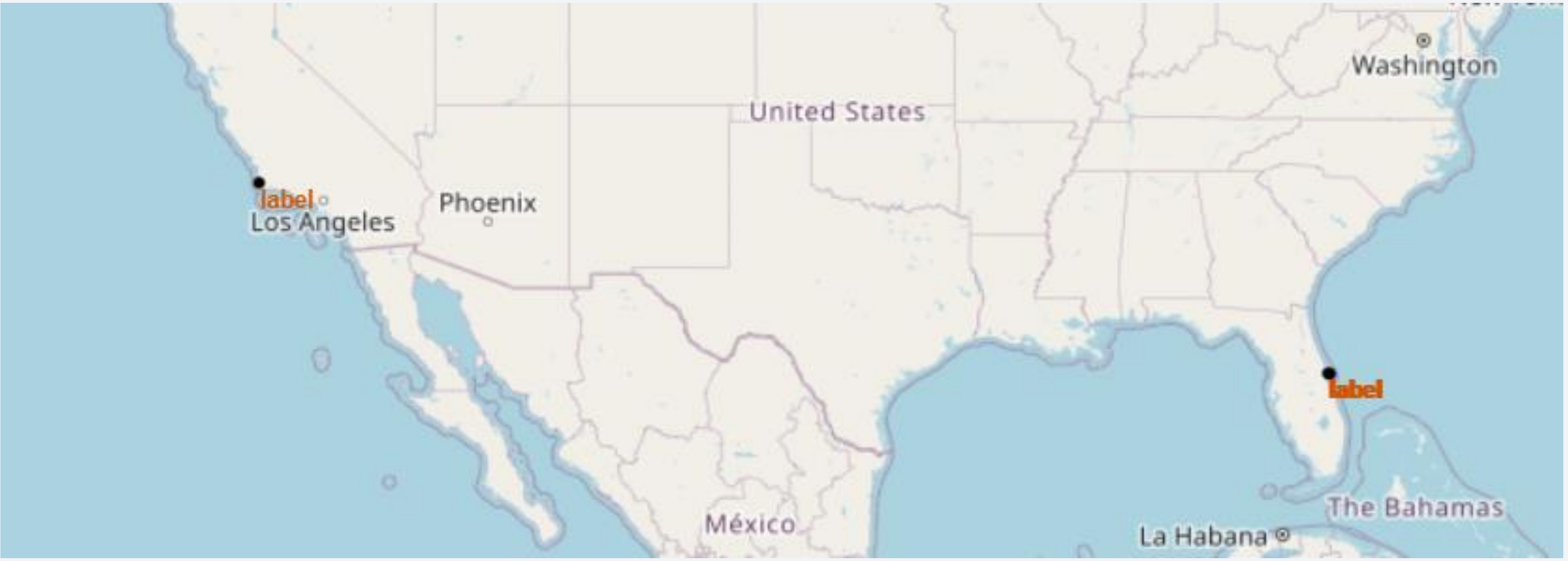
Out[32]:

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

Section 3

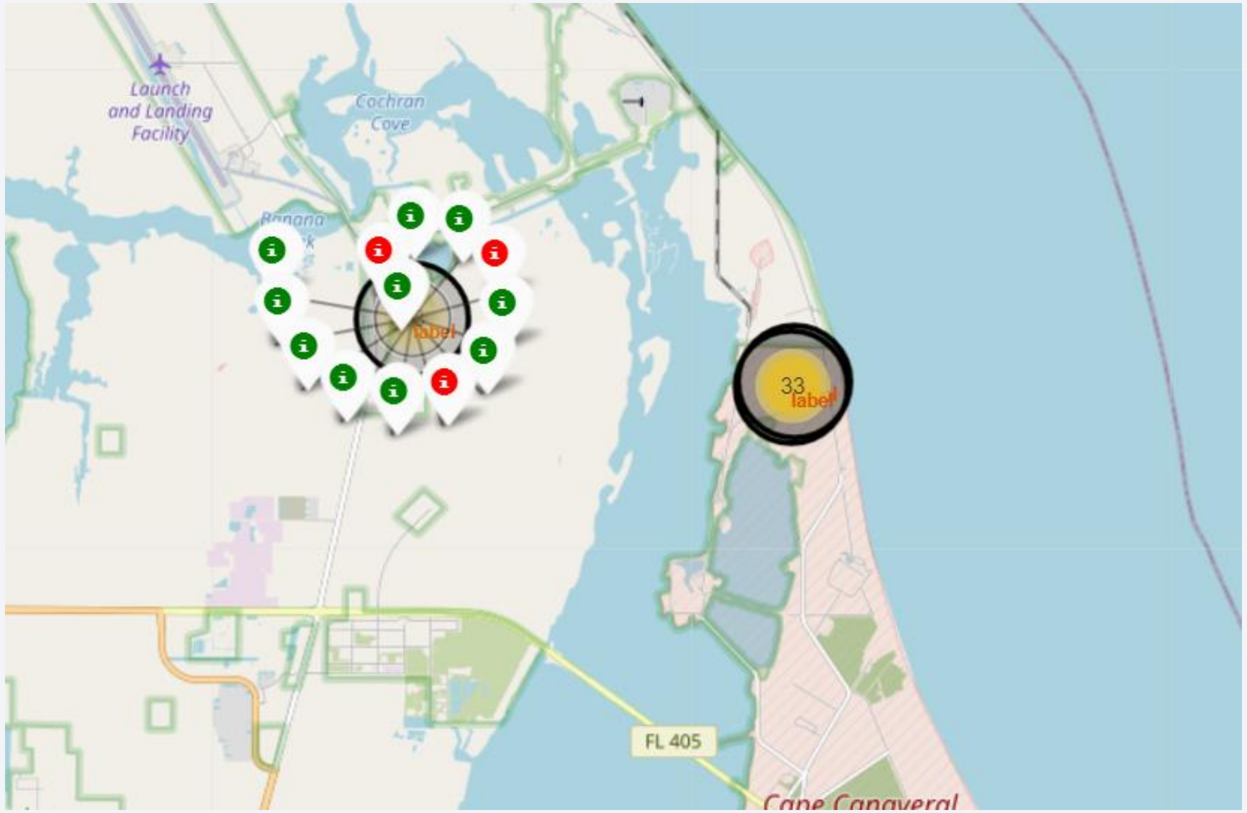
Launch Sites Proximities Analysis

Launch sites on a map



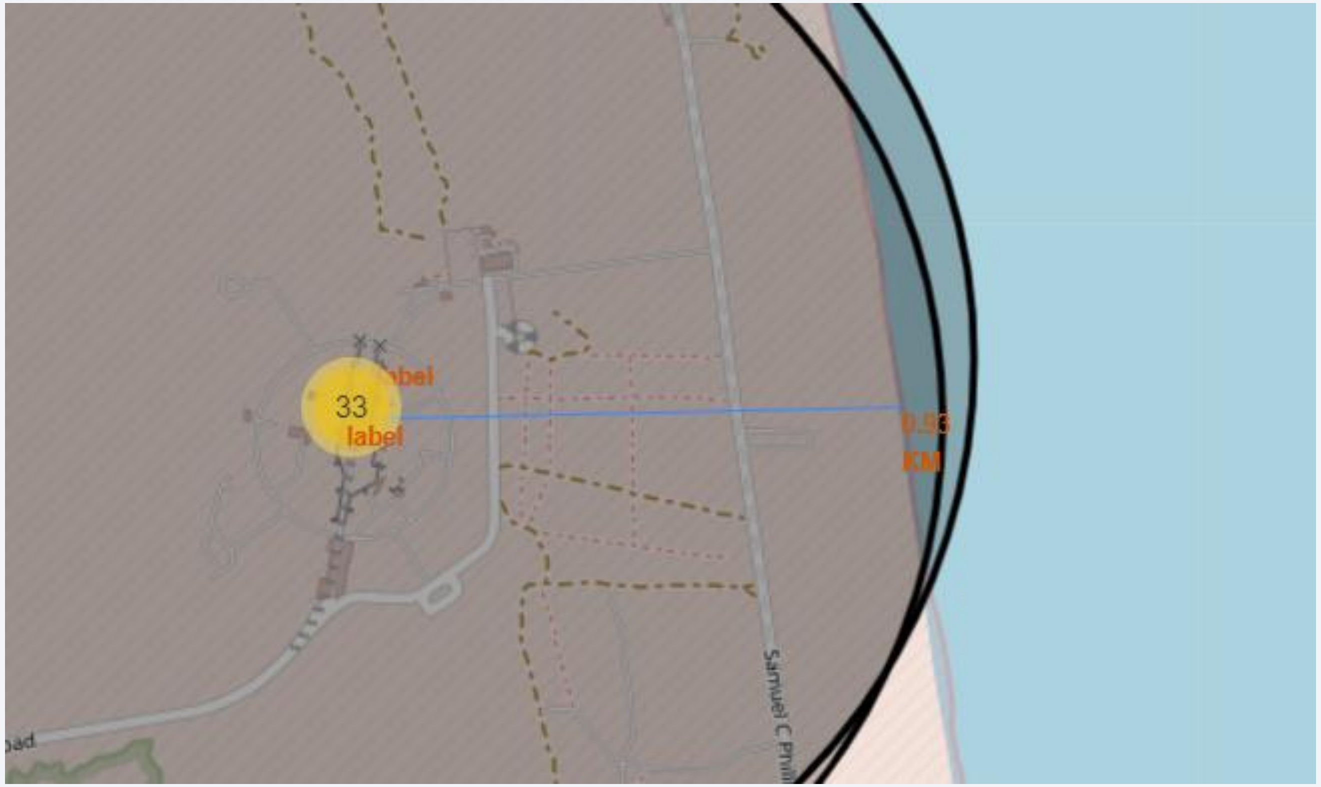
- Using a Folium library the launch sites were added to map

Success/failed launches



- Success / failed outcome are shown on the map for each Launch Site

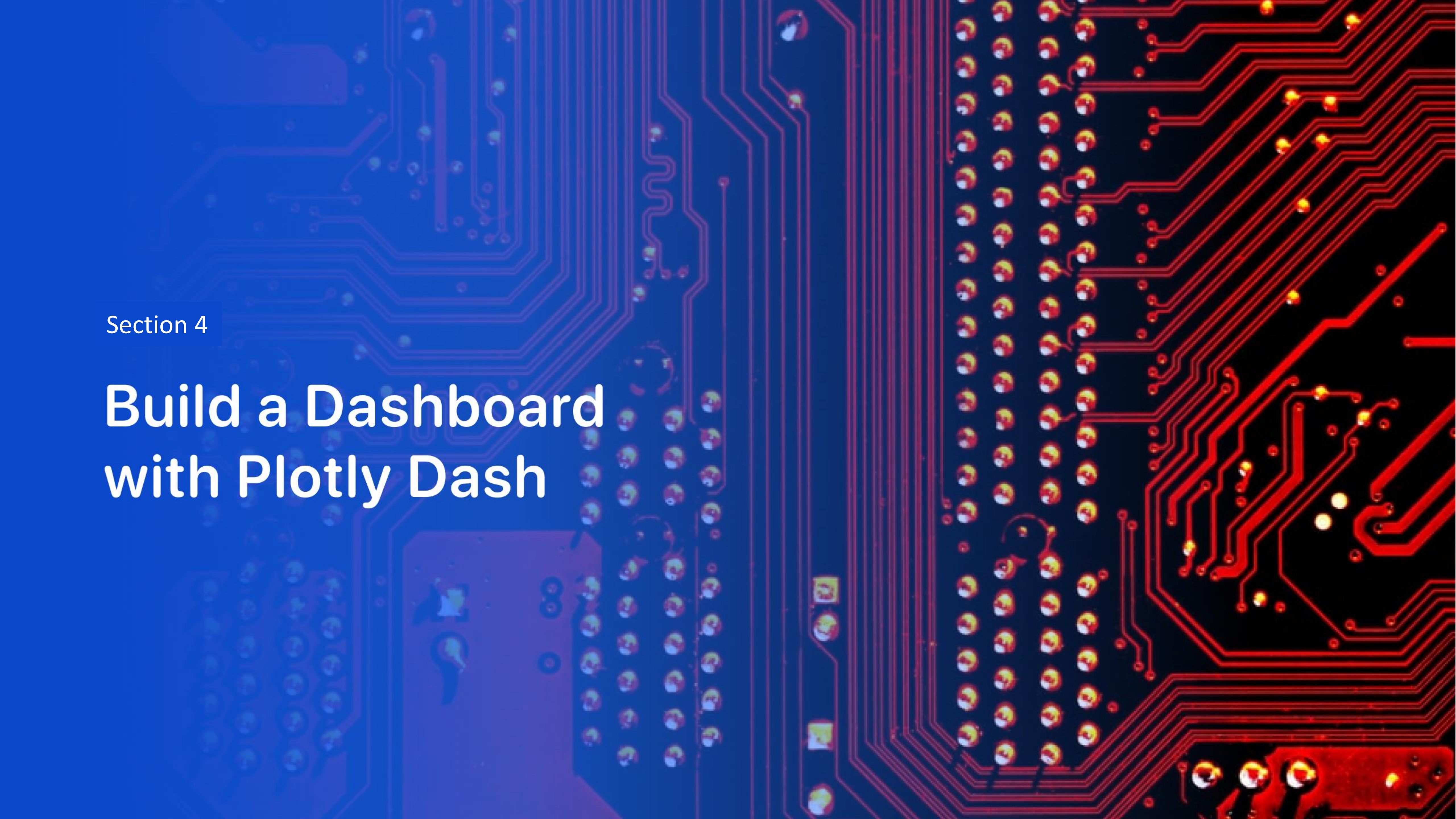
Coastline proximity to a Launching site



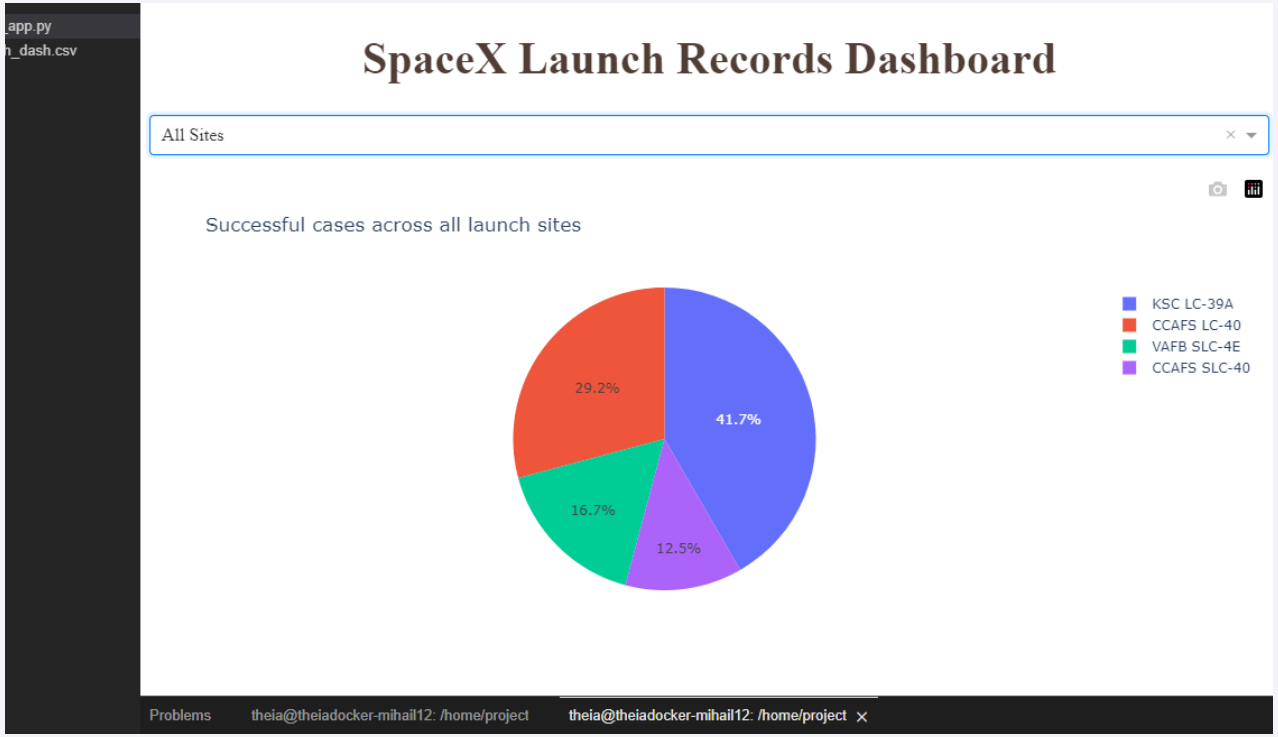
- Using Folium functionality we can put the distances of the different objects on to the map

Section 4

Build a Dashboard with Plotly Dash

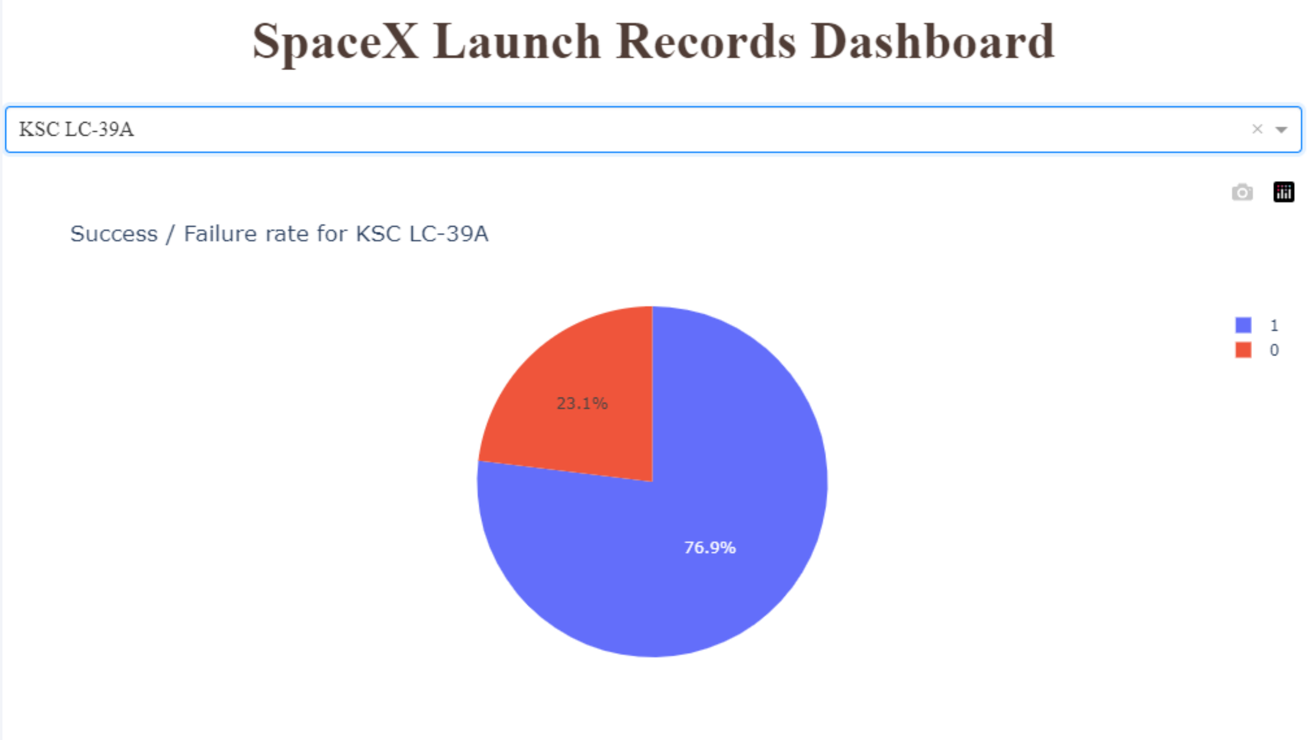


Successful outcomes across all launch


app.py
n_dash.csv
SpaceX Launch Records Dashboard
All Sites
Successful cases across all launch sites
41.7%
29.2%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Problems
theia@theiadocker-mihail12: /home/project
theia@theiadocker-mihail12: /home/project

- From the pie chart we can conclude that the most of th successful outcomes were made from KSC LC-39A

Success / failure outcomes for specific launch site



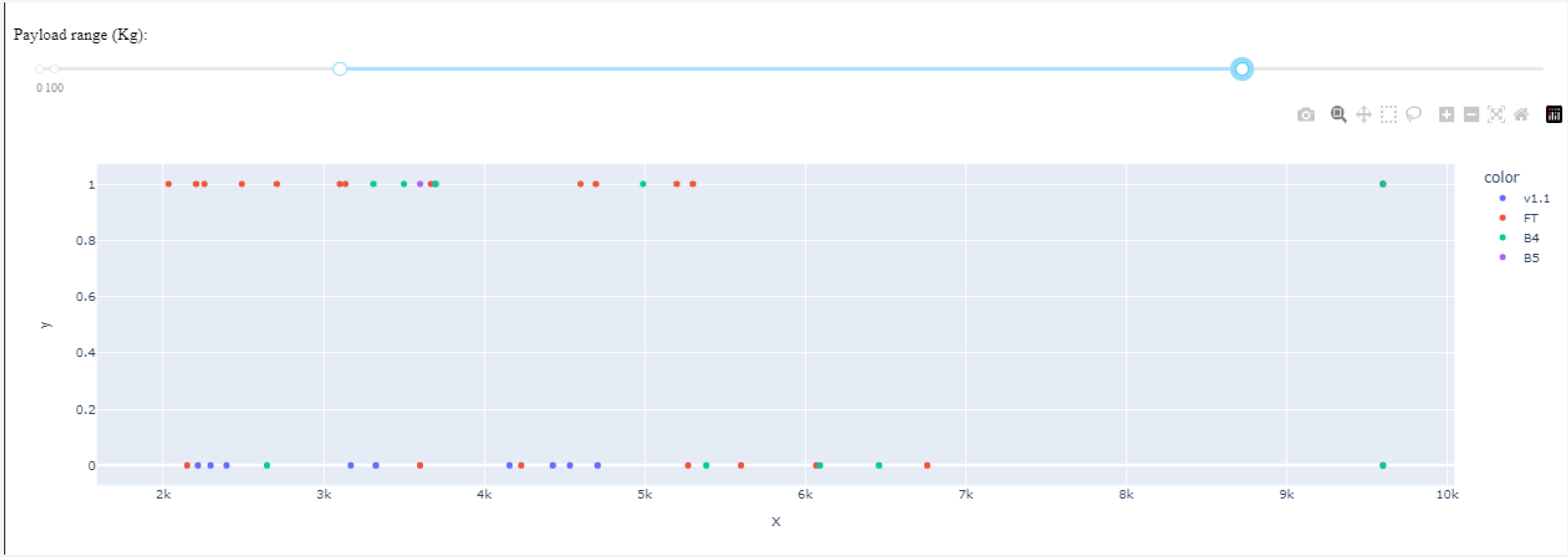
SpaceX Launch Records Dashboard

KSC LC-39A

Success / Failure rate for KSC LC-39A

- The generated chart also shows the outcomes per specific launch site. For example KSC LC-39A – success rate is 76,9%

Payload vs. Launch Outcome scatter

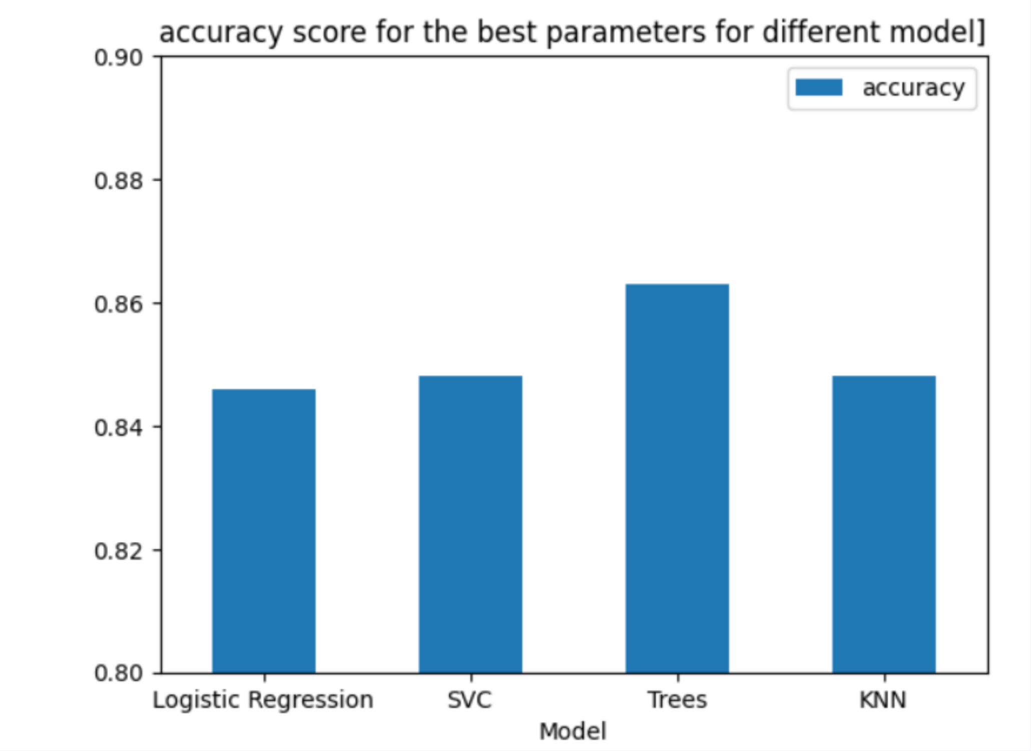


- The scatter plot shown above allows us to see the outcome of the returning of the first stage depending on the payload. From the chart we can estimate what was the mass of the payload that historically tended to have successful outcomes

Section 5

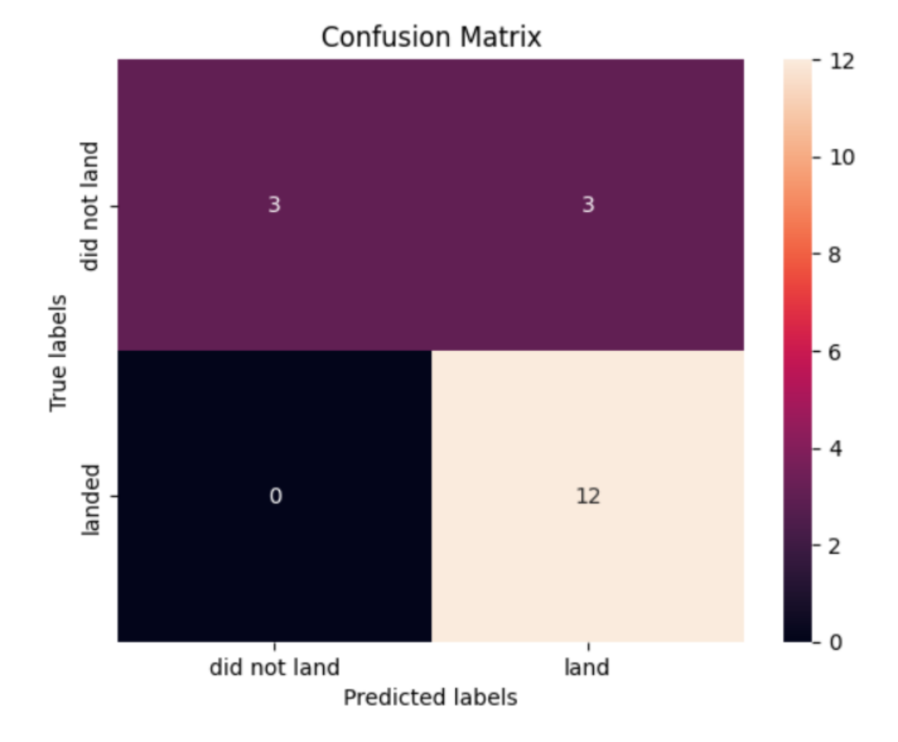
Predictive Analysis (Classification)

Classification Accuracy



- The Trees model gave us the highest value for the accuracy

Confusion Matrix



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

Conclusions

- During the project the data was gathered and prepared for analysis using word scraping and API calls
- Explanatory Data Analysis was performed to find the parameters that can be used as features to train our model. We used both visualization approach and data analysis using sql
- We have trained 4 Machine learning Model (Logistic Regression, SVC, Trees and KNN) with different parameters. Using test set we found out that the Modael that gives the highest accuracy (0.863) is Trees Classifier Model
- The parameters giving the best results are: 'criterion': 'gini', 'max_depth': 16, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'

Appendix

- The repository containing all used data and notebooks can be found via link

  https://github.com/krylov-mihail/capstone-ibm/tree/master

Thank you!