IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohamed Khaled Attia
18/1/2023

Outline

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

Executive Summary

- Using EDA and Machine Learning
- for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)

Introduction

- The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

Section 1

Methodology

Methodology

Executive Summary

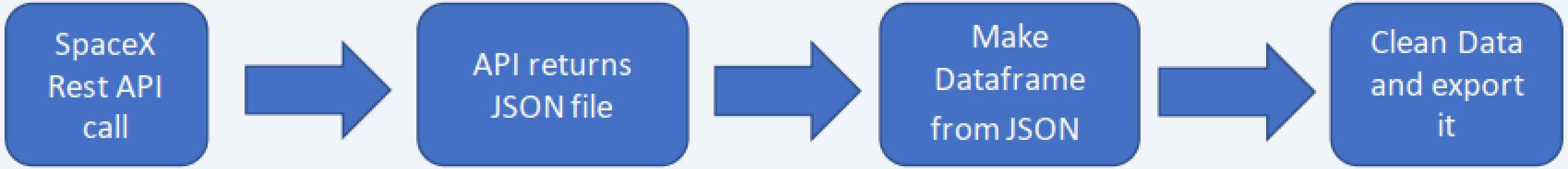
- Data collection methodology:
  - Space X Falcon 9 First Stage Landing Prediction is extracted from URL by pandas library
- Perform data wrangling
  - perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models, there are several different cases where the booster did not land successfully.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Perform exploratory Data Analysis and determine Training Labels, Find the method performs best using test data

Data Collection

- Data is collected using pandas libraries by read_csv Function from
- "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv". URL
- Data is collected using the following code:
- URL2 = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_3.csv'
- resp2 = await fetch(URL2)
- text2 = io.BytesIO((await resp2.arrayBuffer()).to_py())
- X = pd.read_csv(text2)

Data Collection – SpaceX API

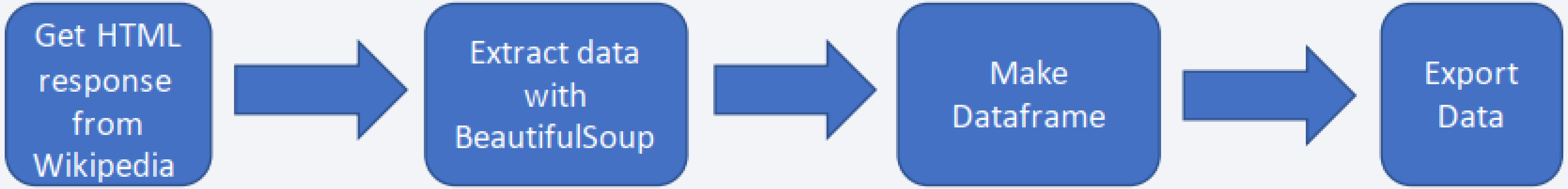
- URL2 = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_3.csv'
- resp2 = await fetch(URL2)
- text2 = io.BytesIO((await resp2.arrayBuffer()).to_py())
- X = pd.read_csv(text2)



Data Collection - Scraping

- We Used .get() method and also used the BeatifulSoup object to contain the content on the HTML web API

- For EX:
- results = requests.get(url).json()
- Results
- {'meta': {'code': 200, 'requestId': '5d31e36c018cbb00396d4086'},
- 'response': {'venues': [{'id': '4fa862b3e4b0ebff2f749f06',
- 'name': "Harry's Italian Pizza Bar",
- 'location': {'address': '225 Murray St',
- 'lat': 40.71521779064671,
-

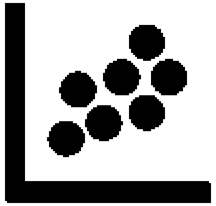


Data Wrangling

- In the dataset, there are several cases where the booster did not land successully.
- There is True Ocean,RTLS, ASDS means the mission has been successful.
- And there is False Ocean, RTLS, ASDS means the mission was a failure.
- We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.

EDA with Data Visualization

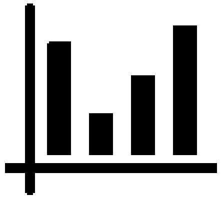
- **Scatter Graphs**
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Orbit vs. Flight Number
  - Payload vs. Orbit Type
  - Orbit vs. Payload Mass



Scatter plots show relationship between variables. This relationship is called the correlation.

- **Bar Graph**
  - Success rate vs. Orbit

Bar graphs show the relationship between numeric and categoric variables.



- **Line Graph**
  - Success rate vs. Year

Line graphs show data variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data.

EDA with SQL

- We performed SQL queries to gather and understand data from dataset:
- Displaying 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 successful 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.

Build an Interactive Map with Folium

Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas

- Red circle at NASA Johnson Space Center's coordinate with label showing its name rdinates with label showing launch
- folium.features.DivIcon).
- The grouping of points in a cluster to display multiple (folium.plugins.MarkerCluster).
- Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing.
- (folium.map.Marker, folium.Icon).
- Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them.
- (folium.map.Marker, folium.PolyLine, folium.features.DivIcon)

Build a Dashboard with Plotly Dash

Dashboard has dropdown, pie chart, rangeslider and scatter plot components
• Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown).
• Pie chart shows the total success dropdown component (plotly.express.pie).
• Rangeslider allows a user to select a payload mass in a fixed range

• Scatter chart shows the relationship between two variables, in particular Success vsPayload Mass (plotly.express.scatter).

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels →
- create a column for the class →
- Standardize the data →
- Split into training data and test data →
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression →
- Find the method performs best using test data.

Results

- The best model is logesticRegression model with score = 0.94
- For Example:

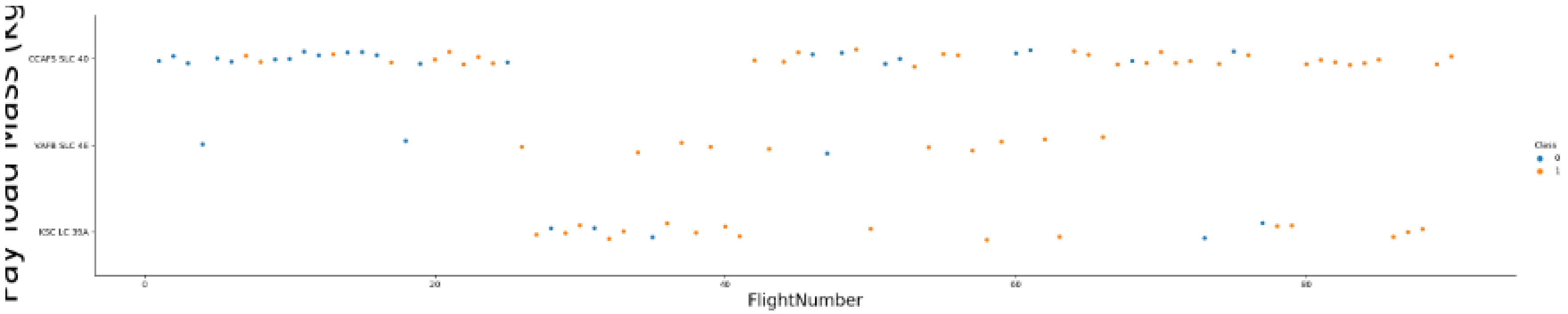
```
[88]: predictors = [knn_cv, svm_cv, logreg_cv, tree_cv]
best_predictor = ""
best_result = 0
for i, predictor in enumerate(predictors):
    #print(predictors[i], end='')
    j = predictor.score(X_test, Y_test)
    print(j)

0.7222222222222222
0.8888888888888888
0.9444444444444444
0.9444444444444444
```

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

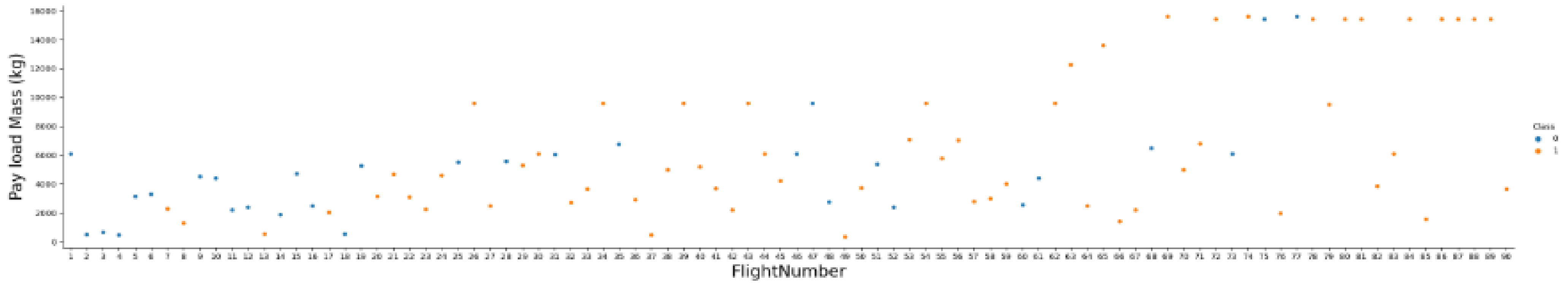
```
[99]: sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
      plt.xlabel("FlightNumber",fontsize=20)
      plt.ylabel("Pay load Mass (kg)",fontsize=40)
      plt.show()
```



- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

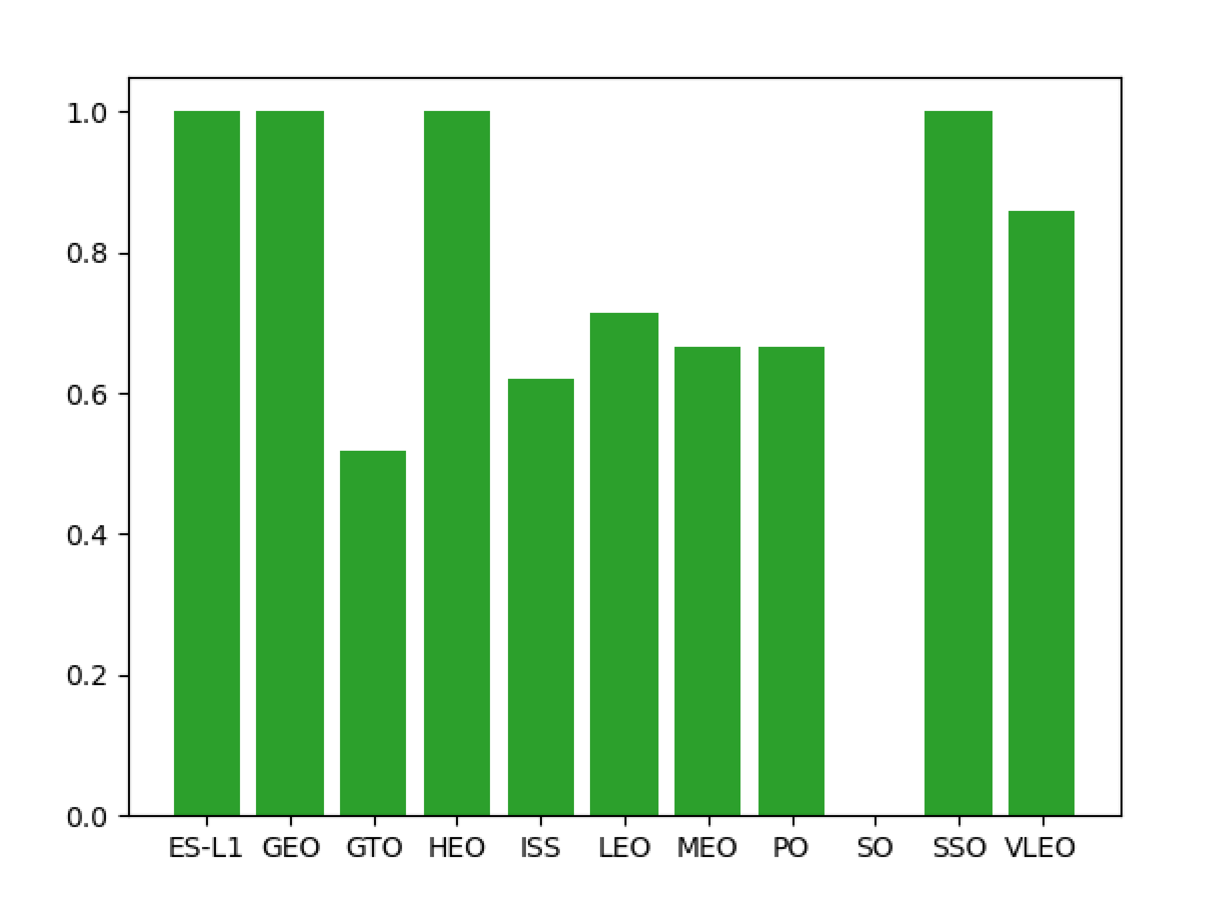
Payload vs. Launch Site

```
[6]: sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)
     plt.xlabel("FlightNumber",fontsize=20)
     plt.ylabel("Pay load Mass (kg)",fontsize=20)
     plt.show()
```

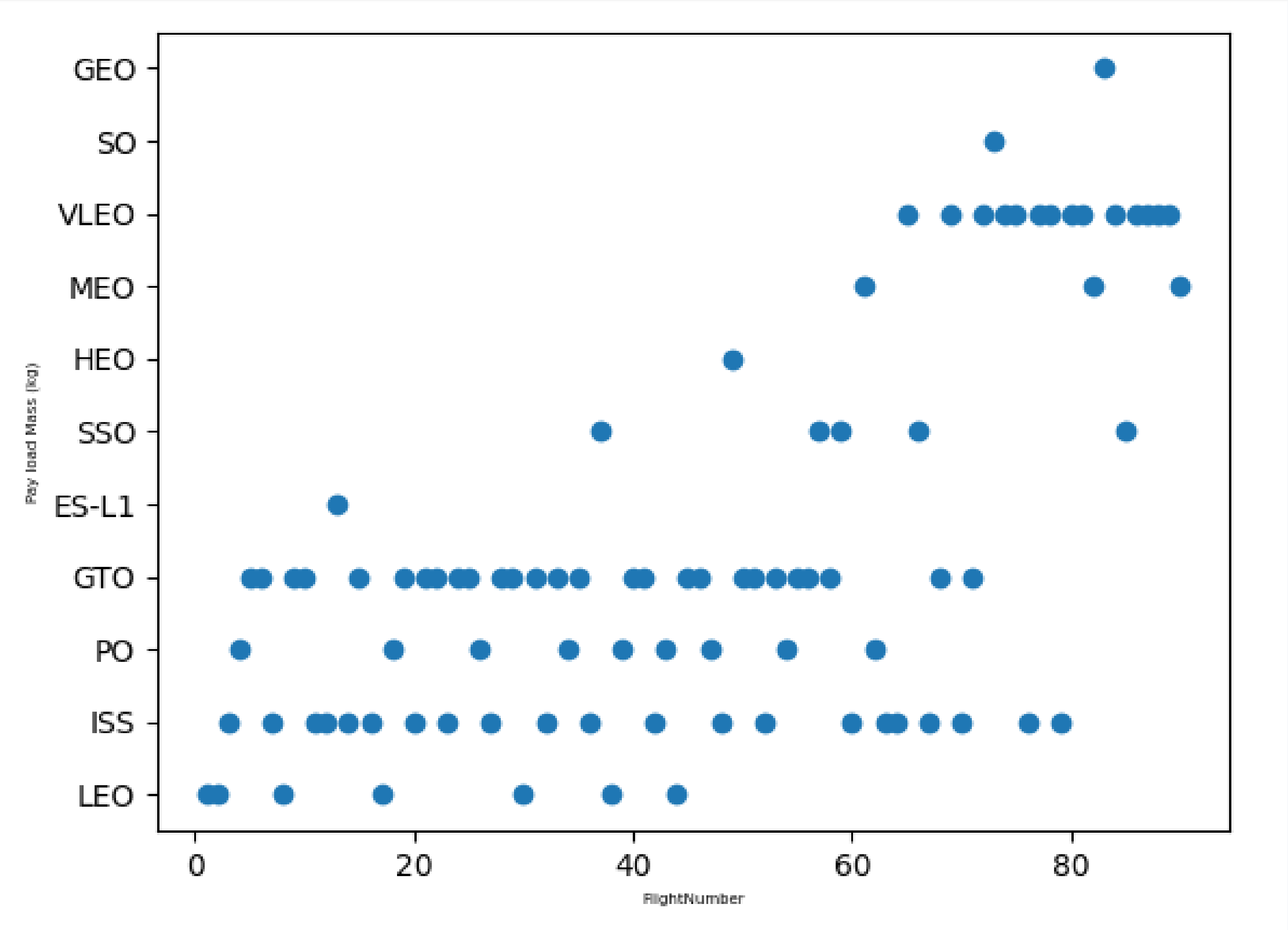


We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

Success Rate vs. Orbit Type

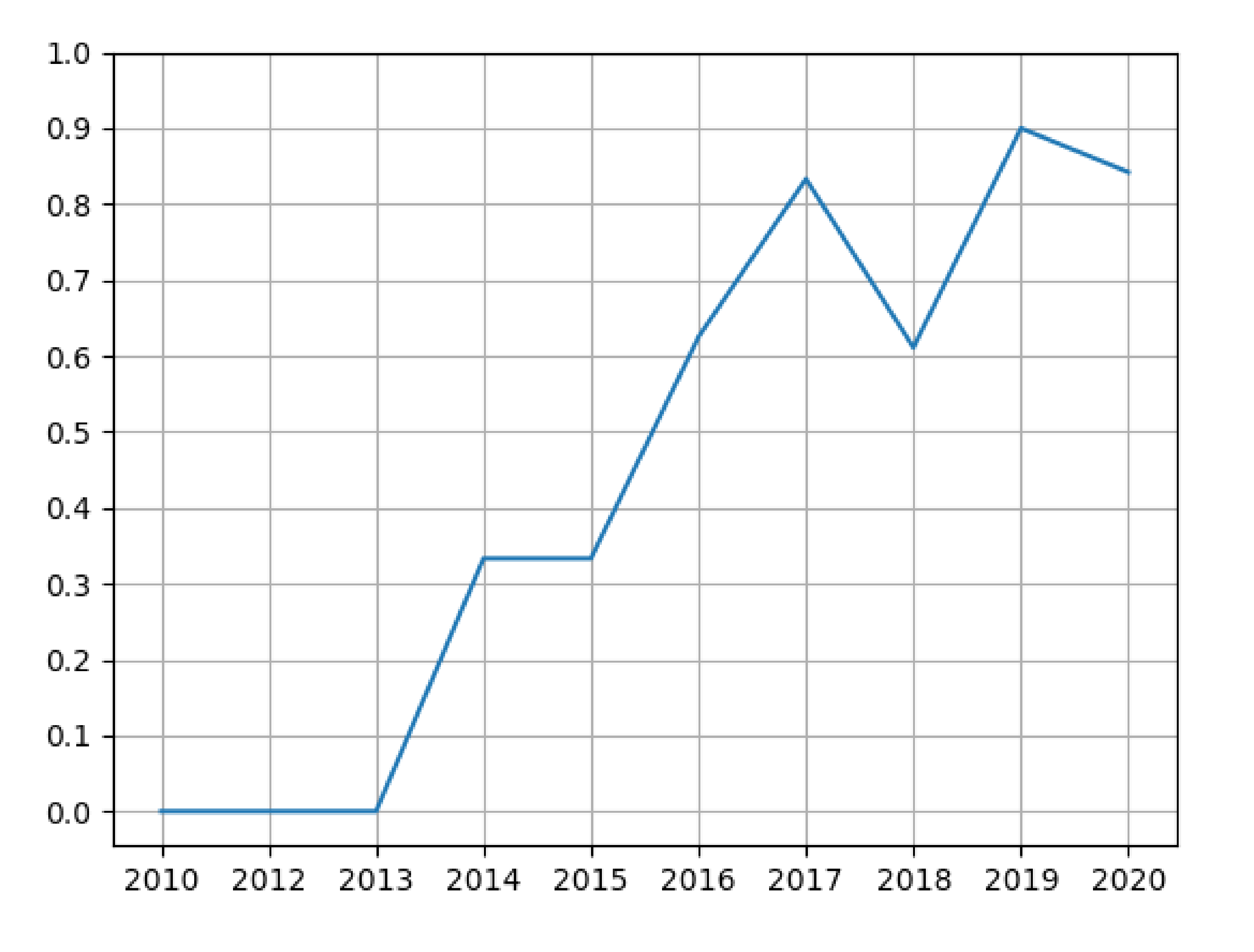


Flight Number vs. Orbit Type



Launch Success Yearly Trend

- It seems in 2019 there is the most likely to success of the class



All Launch Site Names

The most launch Site trips is CCAFS SLC 40

```
[7]: # Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

```
[7]: CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Launch Site Names Begin with 'CCA'

```
SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

[25]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

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

[25]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

[35]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1';
```

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

[35]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

[58]:
```
%%sql
SELECT "Date", "Time (UTC)", "Landing _Outcome" FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE '%success%'
LIMIT 1;
```

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

[58]:

Date	Time (UTC)	Landing _Outcome
22-12-2015	01:29:00	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

[59]:

```
%%sql
SELECT * FROM SPACEXTBL LIMIT 2
```

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

[59]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)

Total Number of Successful and Failure Mission Outcomes

[61]:
```
%%sql
SELECT * FROM SPACEXTBL
WHERE "Mission_Outcome" = 'Success'
AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000 LIMIT 3;
```

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

[61]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
05-08-2014	08:00:00	F9 v1.1	CCAFS LC-40	AsiaSat 8	4535	GTO	AsiaSat	Success	No attempt
07-09-2014	05:00:00	F9 v1.1 B1011	CCAFS LC-40	AsiaSat 6	4428	GTO	AsiaSat	Success	No attempt
02-03-2015	03:50:00	F9 v1.1 B1014	CCAFS LC-40	ABS-3A Eutelsat 115 West B	4159	GTO	ABS Eutelsat	Success	No attempt

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

[64]:
```
%%sql
SELECT * FROM SPACEXTBL
WHERE Mission_Outcome LIKE '%FAILURE%'
```

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

[64]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
28-06-2015	14:21:00	F9 v1.1 B1018	CCAFS LC-40	SpaceX CRS-7	1952	LEO (ISS)	NASA (CRS)	Failure (in flight)	Precluded (drone ship)

Boosters Carried Maximum Payload

```
[91]: %%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

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

[91]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600

2015 Launch Records

```
[101]: %%sql
SELECT substr("Date", 4, 2) AS 'Month', "Landing _Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL
WHERE substr("Date",7,4)='2015'
AND substr("Date", 4, 2)
AND "Landing _Outcome" LIKE "%Failure%";
```

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

[101]:

Month	Landing _Outcome	Booster_Version	Launch_Site
01	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

```
[122]: %%sql
SELECT "Date", "Landing _Outcome", COUNT("Landing _Outcome") AS LandingOutcomeCount, "Booster_Version", "Launch_Site"
FROM SPACEXTBL
WHERE "Date" BETWEEN '04-06-2010' and '20-03-2017'
AND "Landing _Outcome" LIKE '%success%'
GROUP BY "Landing _Outcome"
ORDER BY COUNT("Landing _Outcome");

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

[122]:

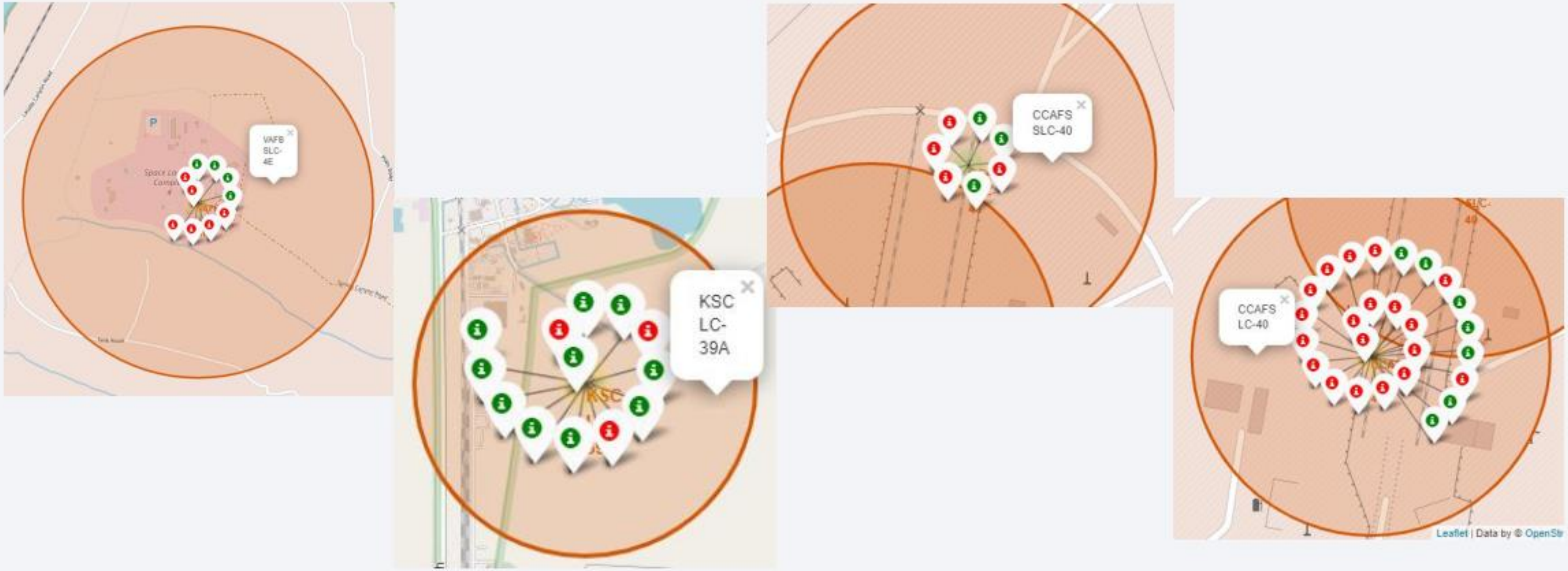
Date	Landing _Outcome	LandingOutcomeCount	Booster_Version	Launch_Site
18-07-2016	Success (ground pad)	6	F9 FT B1025.1	CCAFS LC-40
08-04-2016	Success (drone ship)	8	F9 FT B1021.1	CCAFS LC-40
07-08-2018	Success	20	F9 B5 B1046.2	CCAFS SLC-40

Section 3

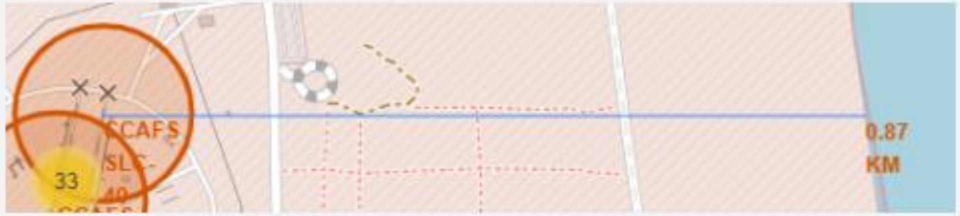
Launch Sites Proximities Analysis

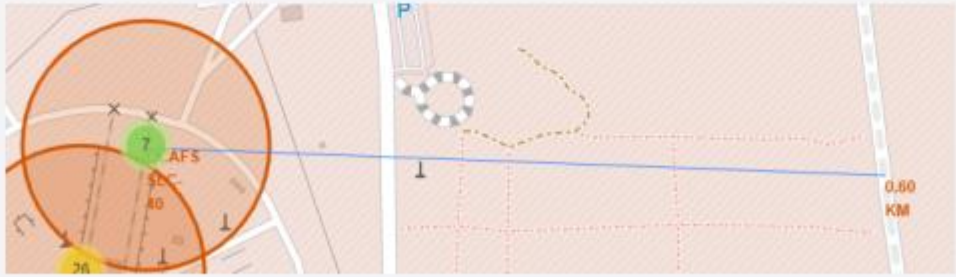
Folium map / Ground stations

Folium map / Color Labeled Markers



Folium Map / Distances between CCAFS SLC-40 and its proximities





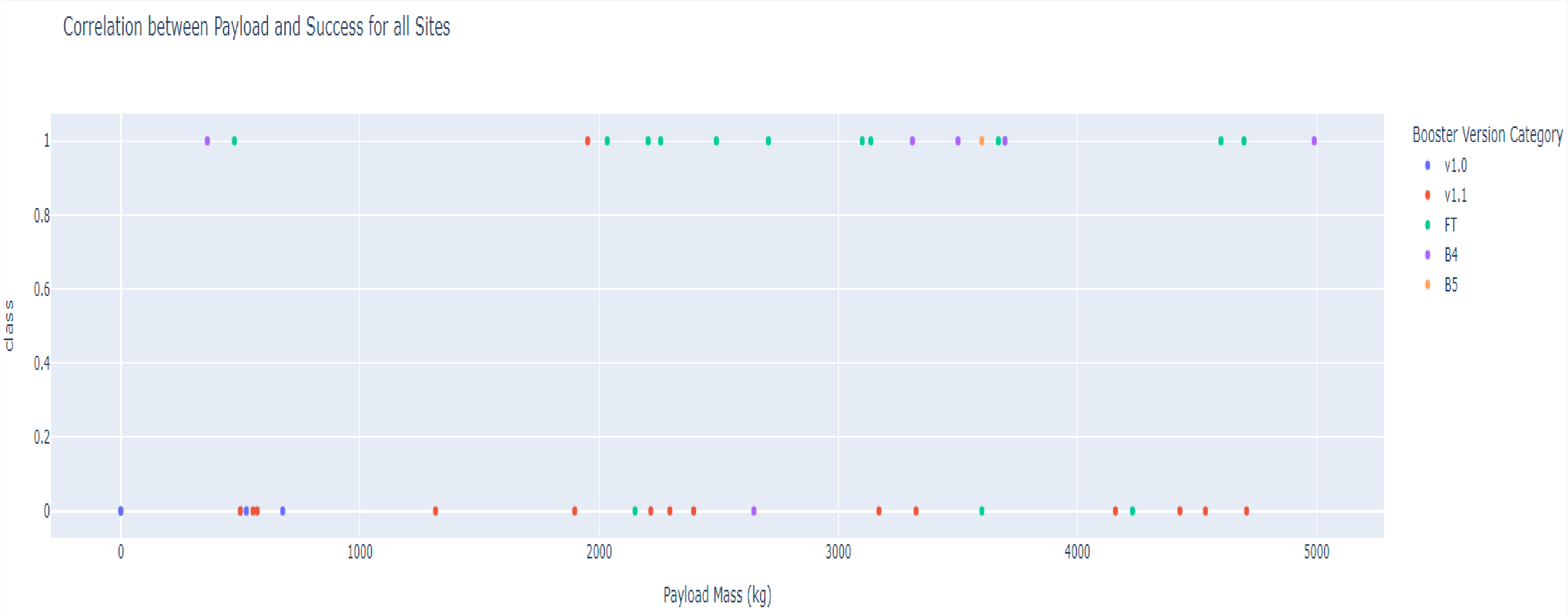
Section 4

Build a Dashboard with Plotly Dash

Dashboard / Total success by Site

Dashboard / Total success launches for Site KSC LC-39A

Dashboard – Payload mass vs Outcome for all sites with different payload mass selected

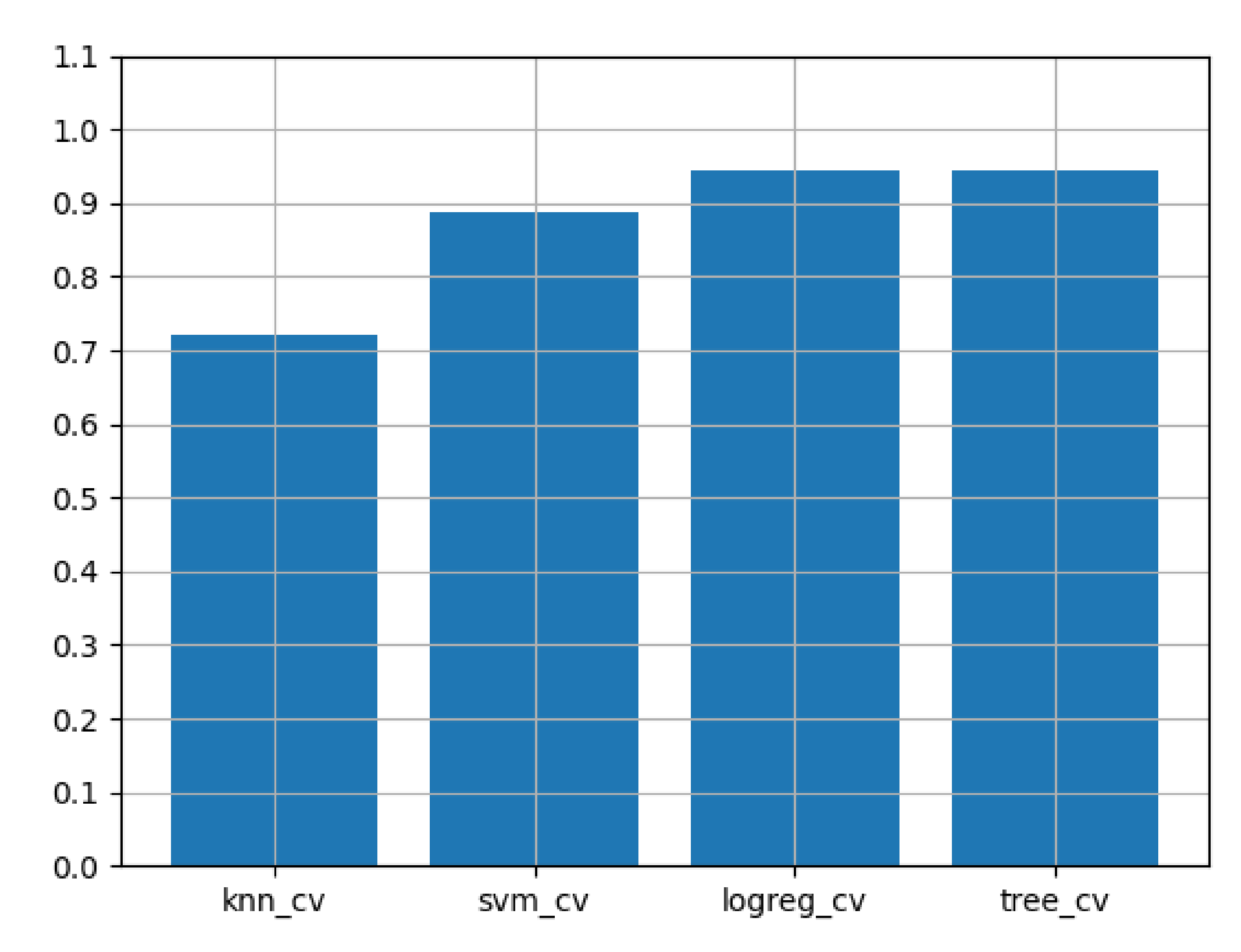


Section 5

Predictive Analysis (Classification)

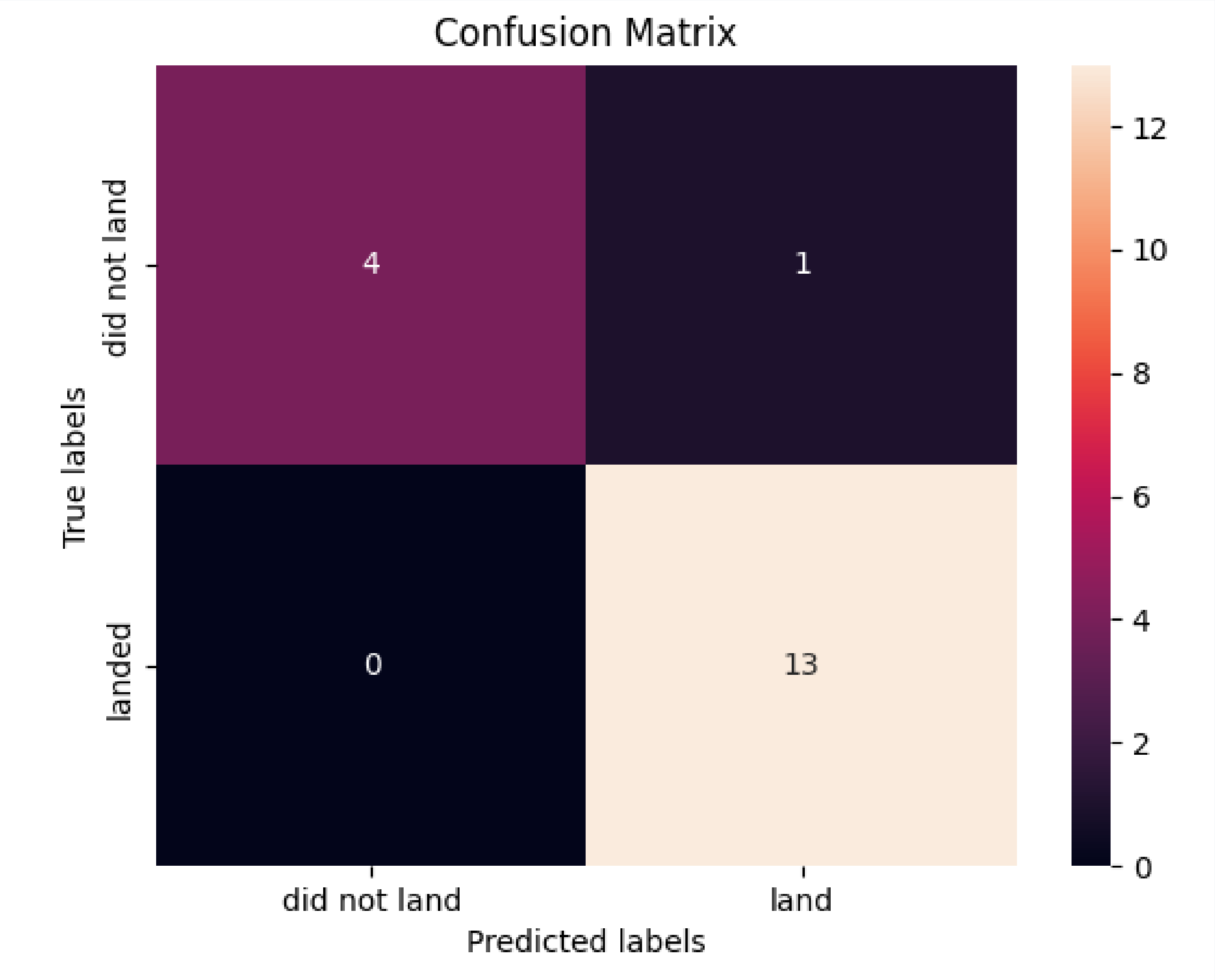
Classification Accuracy

- The best are log_reg_cv
- and Tree_cv models



Confusion Matrix

- Only one element from the whole sample is predicted wrong



Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and
  especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
  • The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
  • Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
  • For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

Appendix

- https://github.com/MKhElhalawany/Coursera

Thank you!