IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vaiva Petrikaite
May 8, 2022

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

SpaceY aims to compete successfully in the market of space journeys. This requires to estimate the costs of rocket-launching .

This project uses the publicly available data on the launches of SpaceX to determine the probability of the reuse of the first-stage.

The former is is of essence in minimising launching costs.

Section 1

Methodology

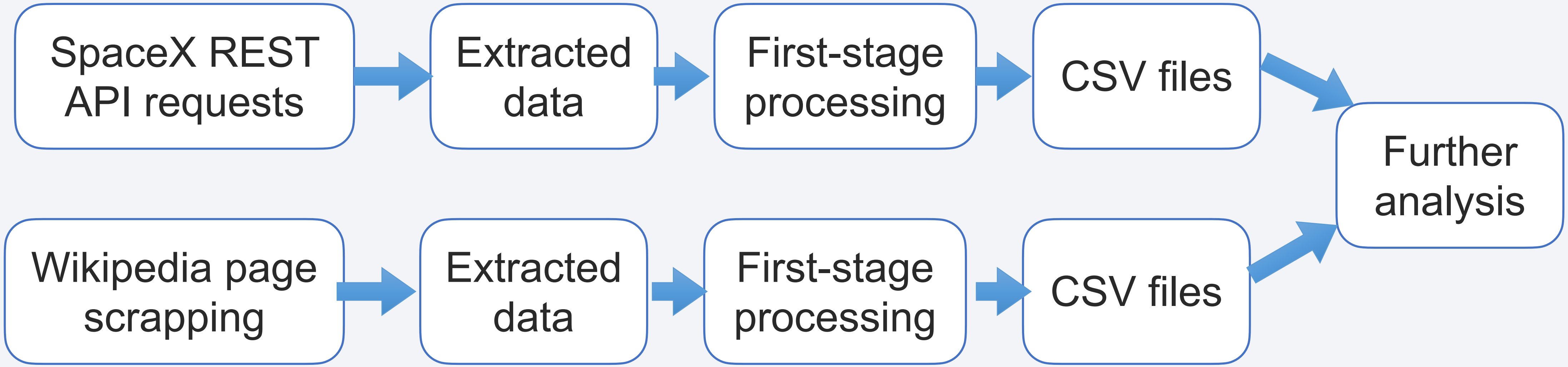
Methodology

Executive Summary

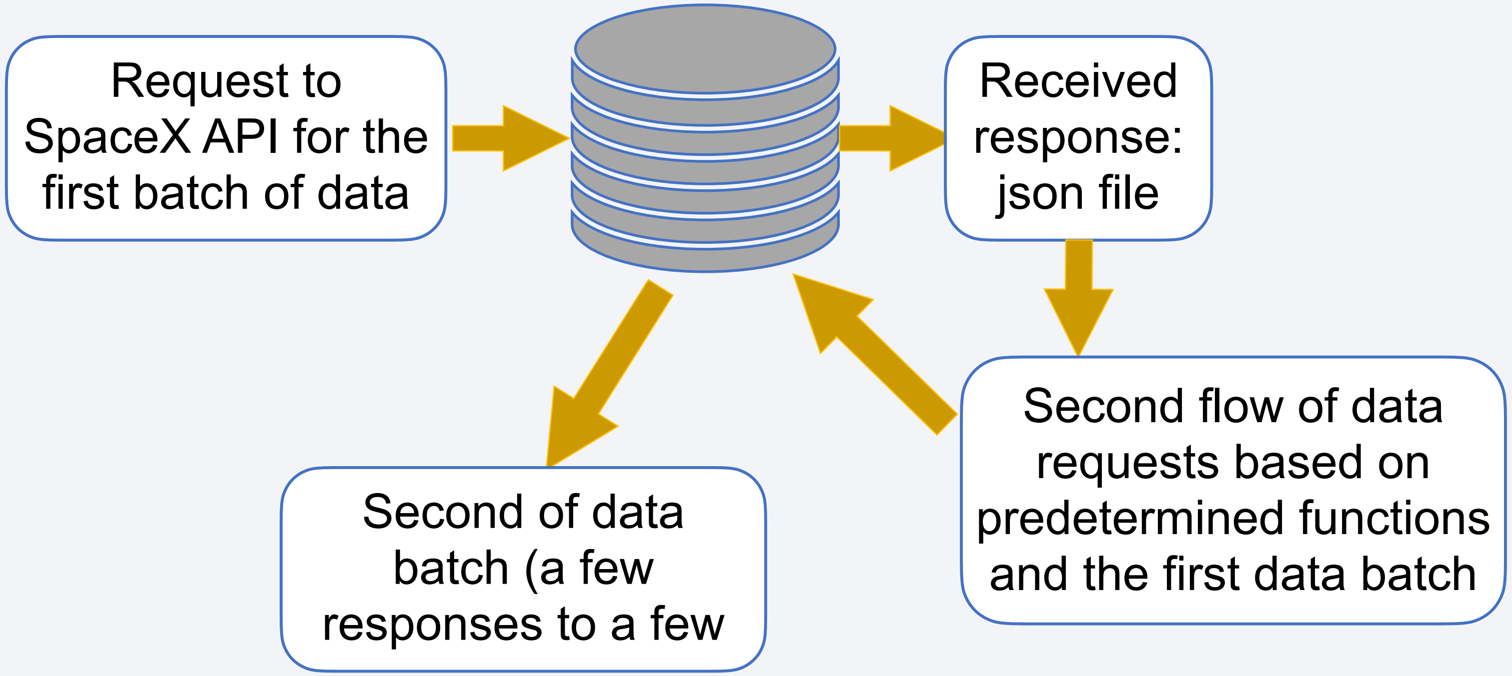
- Data collection methodology:
  - The data was collected from two sources: SpaceX REST API and Wikipedia pages on Falcon 9 historical launch records.
- Perform data wrangling
  - The data was merged into the usable datasets, categorical variables converted to appropriate dummies, and specific null values replaced by appropriate artefacts.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - There have been several classification models uses (KNN, SVM, LR, DT) and picked the one with the best predictive features.

Data Collection

- The data was collected by using two sources: SpaceX REST API and scrapping a Wikipedia page on the launches of Falcon 9.



Data Collection – SpaceX API



The Jupiter notebook is available at:
https://github.com/vaivapetrikaite/data_science_ibm/blob/68e8586bca7c3f1a9d72949971aecacdf530ea78/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

The Jupiter notebook is available at:
https://github.com/vaivapetrikaite/data_science_ibm/blob/58a3f545dc53cf44bb4055d1c3d165040f3d05b6/jupyter-labs-webscraping.ipynb

Data Wrangling

json imported to panda’s df

html file parsed and imported to panda’s df

One hot encoding for categorical variables

Replacing some missing values by means

Determining good and bad landing outcomes and coding to 0 and 1

The Jupiter notebook is available at:
https://github.com/vaivapetrikaite/data_science_ibm/blob/bd8f190f14322da722b2a8ec7753872a5fb0d701/labs-jupyter-spacex-Data%20wrangling.ipynb

Folium map

- To see whiter all necessary infrastructure was at hand, what hindrances, e.g. cities, were nearby the launching site, a few Folium pats were drawn.
- There were a few steps in the map drawing:
  - All launch sites market carefully
  - Successful and unsuccessful launches colour-coded in the data frame and marked on the map.
  - Distance to the nearest infrastructure (rails, highway) and other objects like coastline and cities were calculated and the lines were drawn

The Jupiter notebook is available at:
https://github.com/vaivapetrikaite/data_science_ibm/blob/c8411ce7a711777466d7f3522098ba79abdfcac4/lab_jupyter_launch_site_location.ipynb

EDA with Data Visualization

- There were three types of charts drawn:
  - To see some relationships between variables (between Flight Number and Launch Site, between Payload and Launch Site, between FlightNumber and Orbit type, between Payload and Orbit type);
  - The distribution of the success for different orbits
  - The trend of the dependent variable

The Jupiter notebook is available at:
https://github.com/vaivapetrikaite/data_science_ibm/blob/c94f4e8d18adc94447fb6a4682b30342ecf4d3e0/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed. To display:
  - the 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 successful landing outcome in ground pad was achieved;
  - 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;
  - the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015;
  - the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

The Jupiter notebook is available at:
https://github.com/vaivapetrikaite/data_science_ibm/blob/f0415379d14219378077f0c14f1fc0e1ad2dcf63/jupyter-labs-eda-sql-coursera.ipynb

Build a Dashboard with Plotly Dash

- The dashboard contains two figures:
  - The distribution of successful and unsuccessful launch overall and according to the site.
  - The scatter-plot showing correlation between Payload and Success.
- The plots help to see whether there is any relationship between the payload and the successful launches, whether it is a general trend of related to specific launch sites (if so, then must be some other characteristics determining the probability of success).

The Jupiter notebook is available at:
https://github.com/vaivapetrikaite/data_science_ibm/blob/4c59f08e76215d9bfd92973e4b32845e2ff62106/Dash%20dashboard.ipynb

Predictive Analysis (Classification)

- To find the best prediction methods a few classification methods were tried (KNN, Logistic regression, Decision tree and SVM).
- The process of estimation for each method:

Split the dataset into the training set and test set → Create the regression (classifier) object and run GridSearchCV for the best parameters. → Estimate the parameters, run the testing, compute the accuracy score and plot the confusion matrix

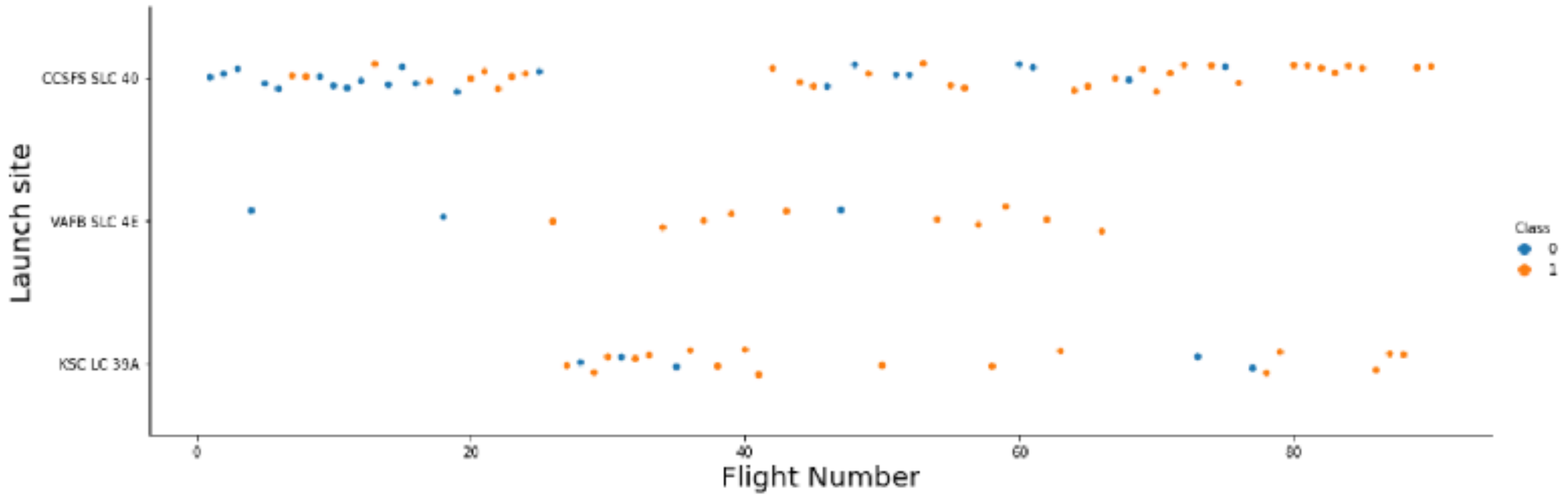
Later, compare all the methods and pick the best one.

The Jupiter notebook is available at:
https://github.com/vaivapetrikaite/data_science_ibm/blob/4c59f08e76215d9bfd92973e4b32845e2ff62106/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Section 2

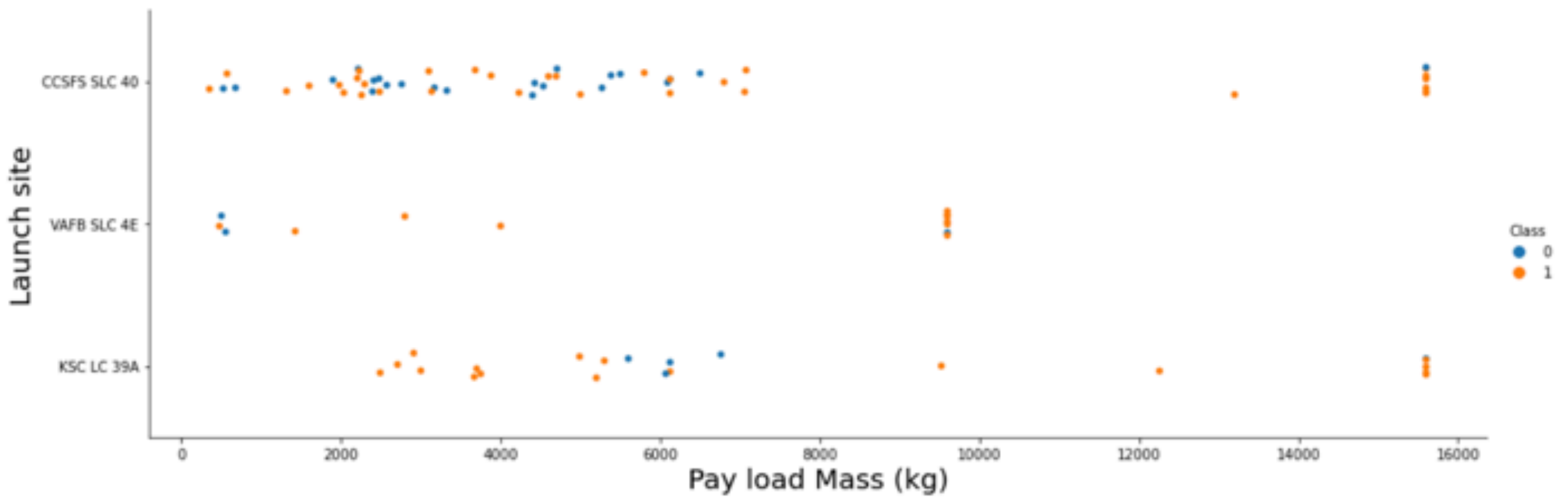
Insights drawn from EDA

Flight Number vs. Launch Site



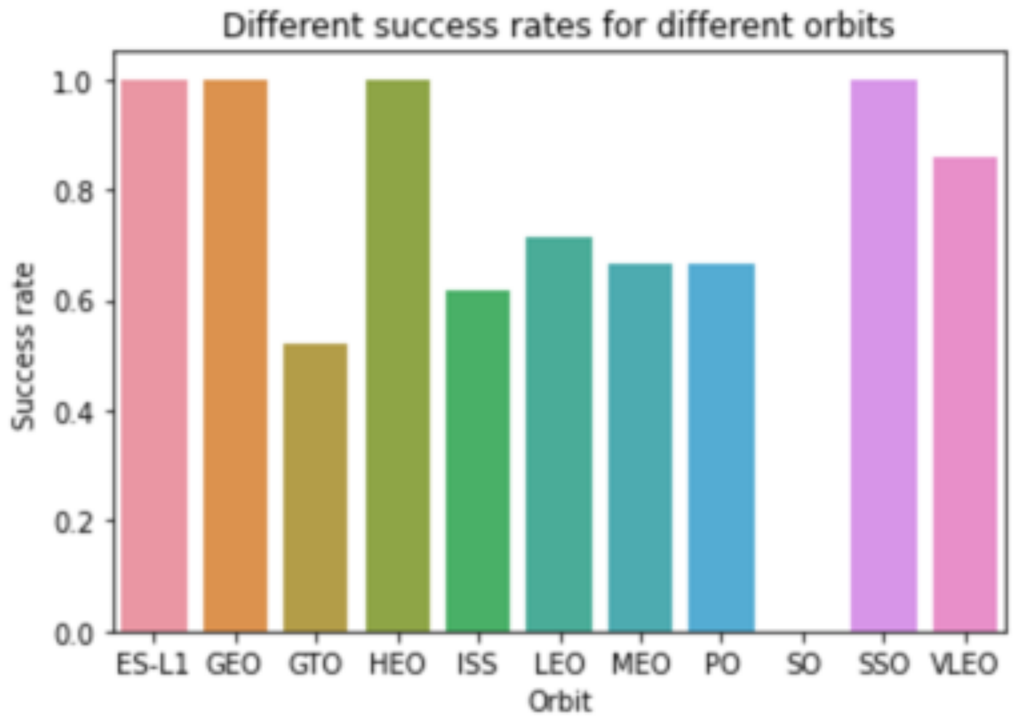
Higher flight numbers have tendency to be more successful.

Payload vs. Launch Site



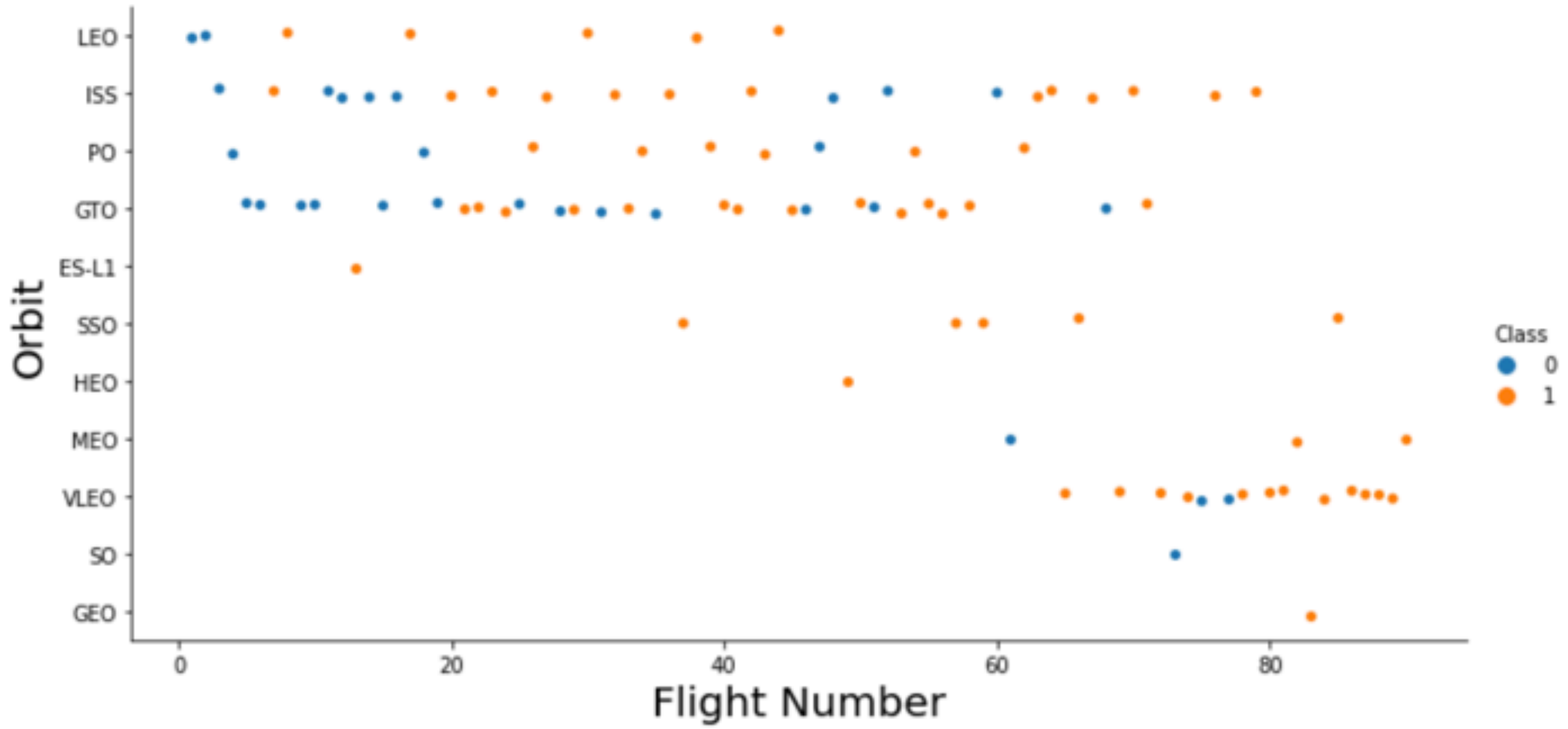
It is apparent that some sites do not launch heavy rockets. However, success and failure rates are more or less the same in all clusters.

Success Rate vs. Orbit Type



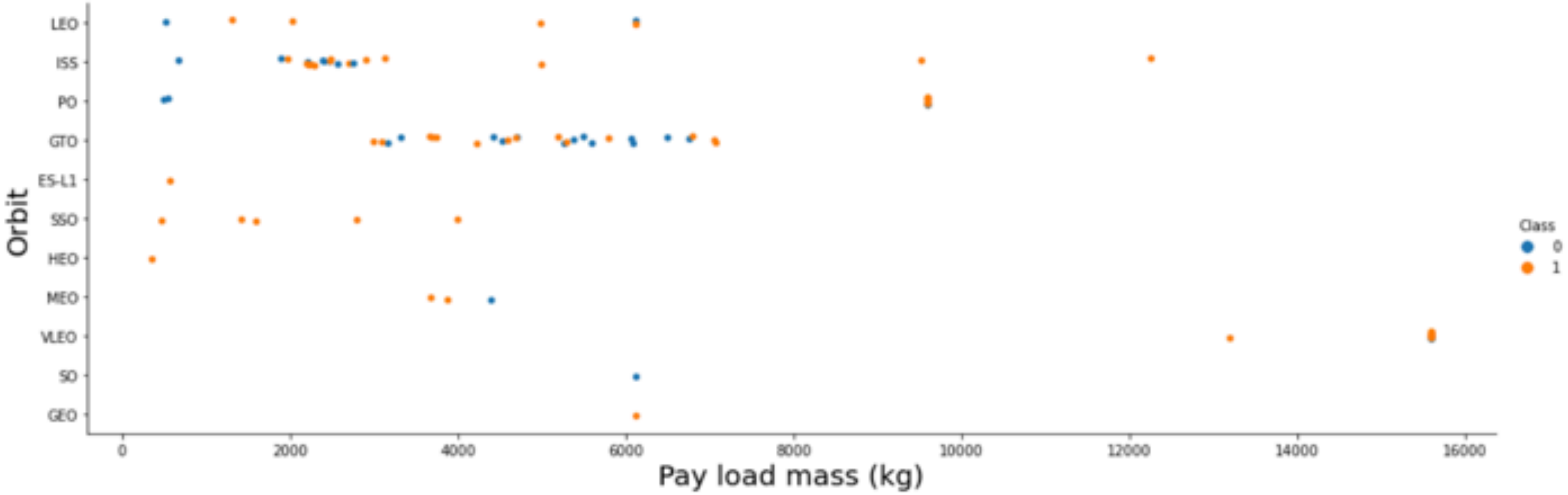
ES_L1, GEO, HEO and S5O have the highest success rates.

Flight Number vs. Orbit Type



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.

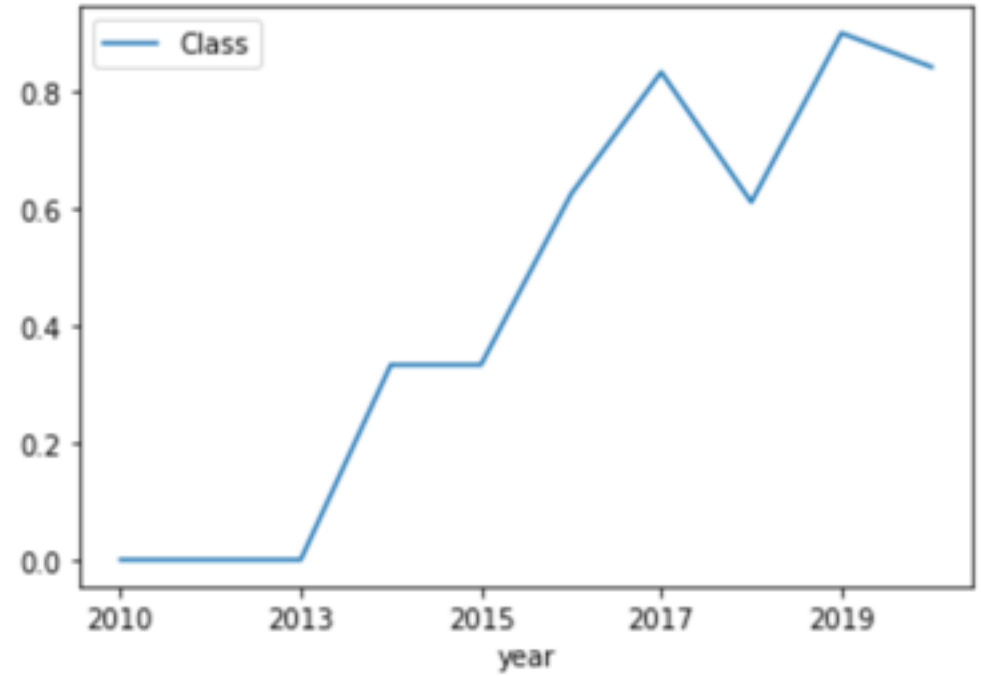
Payload vs. Orbit Type



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



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

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

```
%sql select distinct LAUNCH_SITE from SPACEXTBL
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where LAUNCH_SITE like '%CCA%' limit 5;
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	S
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	S
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	S
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	S
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	S

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER like '%NASA (CRS)%';
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

1
48213

Average Payload Mass by F9 v1.1

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%';
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

1
2534

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql select min(DATE) from SPACEXTBL where LANDING__OUTCOME='Success (ground pad)'
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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 BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)'
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
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

List the total number of successful and failure mission outcomes

```
%sql select count(MISSION_OUTCOME)  from SPACEXTBL where MISSION_OUTCOME like '%Success%'
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

1
100

```
select count(MISSION_OUTCOME)  from SPACEXTBL where MISSION_OUTCOME  not like '%Success%'
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

1
1

Boosters Carried Maximum Payload

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
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

[21]:
```
%sql select BOOSTER_VERSION,LANDING__OUTCOME, LAUNCH_SITE, DATE from SPACEXTBL where DATE<'01-01-2016' and DATE>'12-31-2014'
            and LANDING__OUTCOME='Failure (drone ship)'
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

[21]:

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

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

```
]: %sql select LANDING__OUTCOME as outcome, count(LANDING__OUTCOME) as count from SPACEXTBL where DATE<'03-20-2017'
and DATE>'06-04-2010' group by LANDING__OUTCOME order by count desc;
```

```
 * ibm_db_sa://jdm16963:***@21fecfd8-47b7-4937-840d-d791d0218660.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31864/BLUDB
Done.
```

]:

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

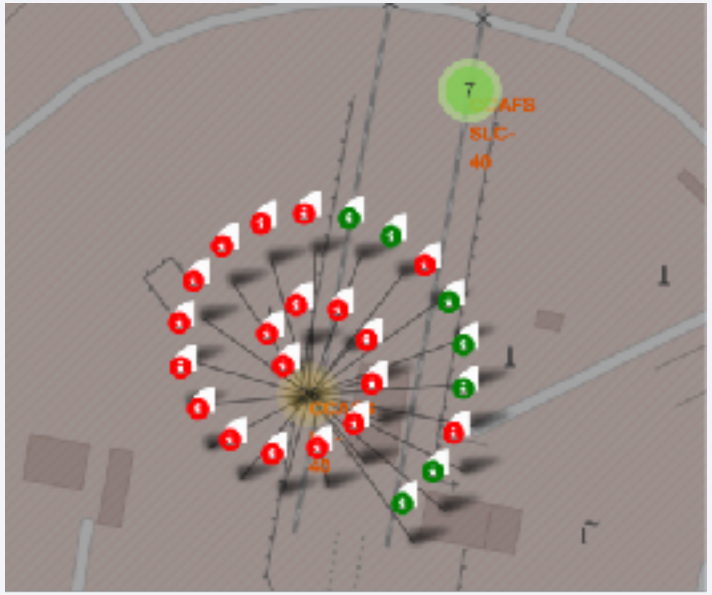
Section 3

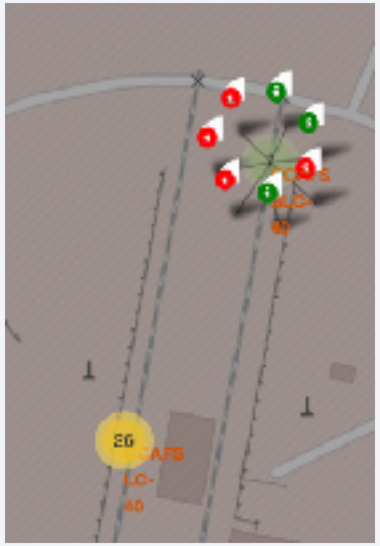
Launch Sites Proximities Analysis

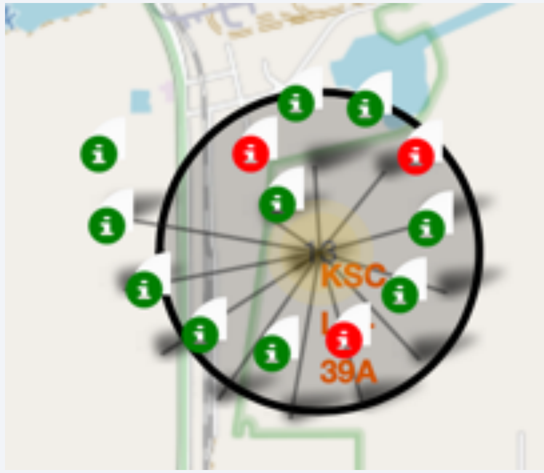
Launching locations

There are two clusters of launching sites: on the East in Florida and on the west in California.

Launching outcomes examples



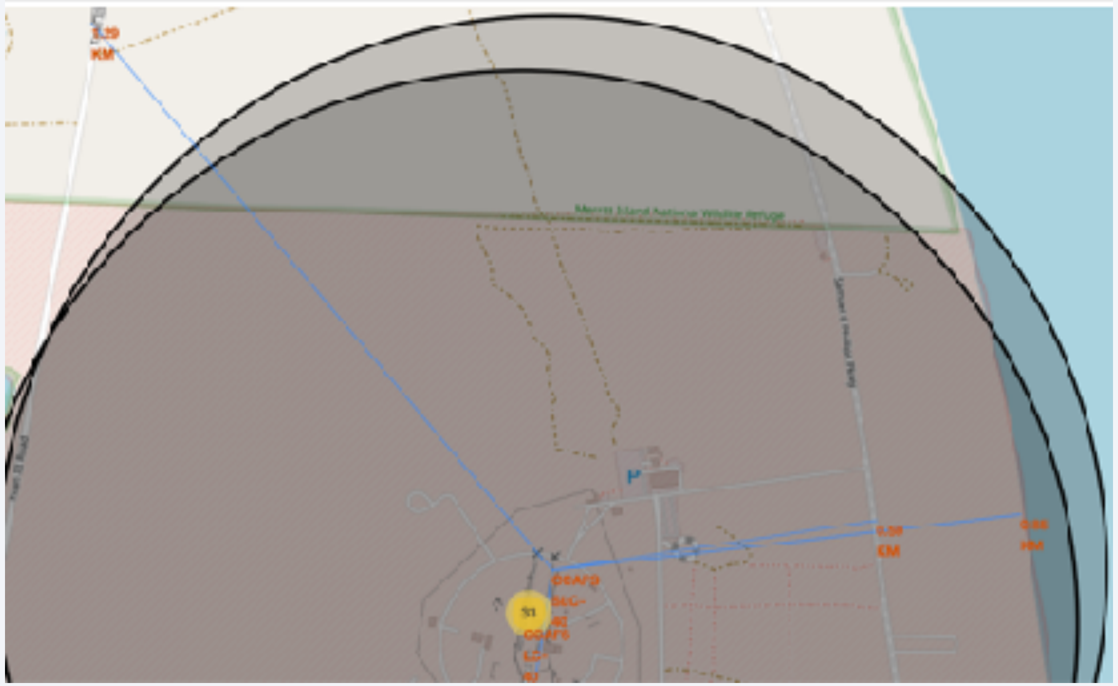




Different sites have different number of launches. However, the share of successful launches (green color) seems to be higher in KSC LC-39A. The properties of that launching site must be investigated.

Distances example

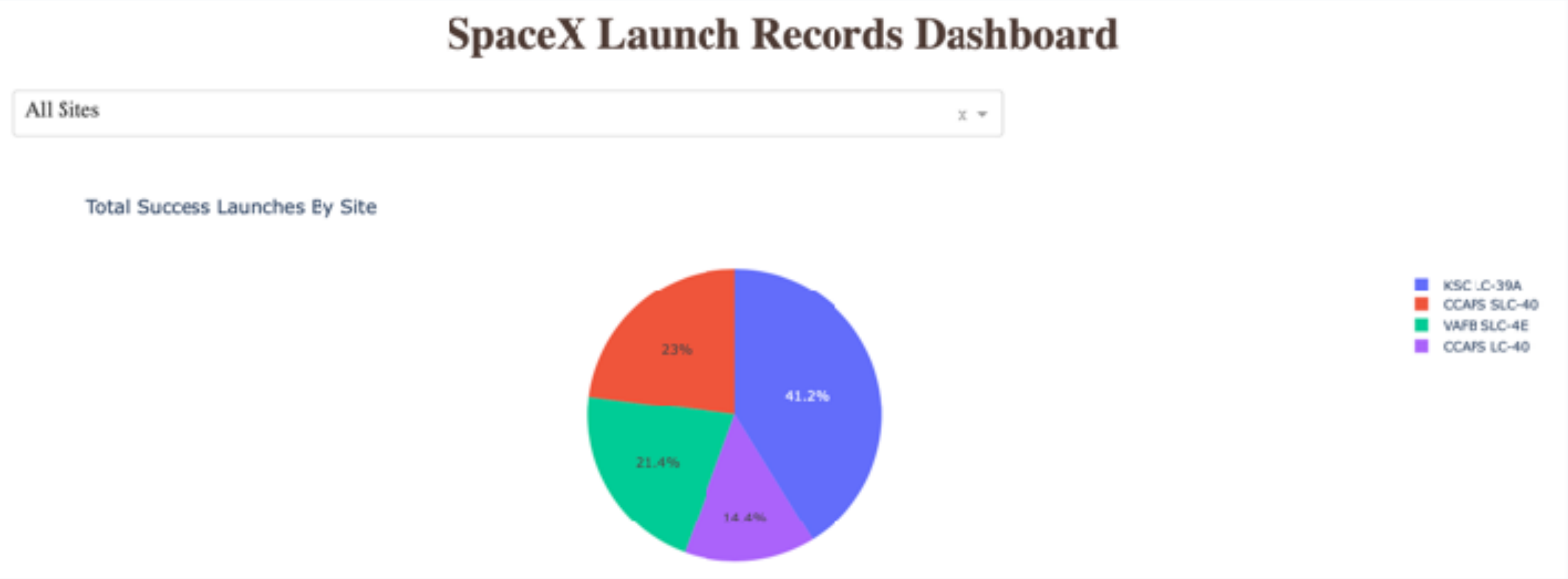
- The launching sites seem to be located close to the necessary infrastructure (highways, rail lines) and far away form the cities.
- E.g. CCAFS SLC-40 is located at:
- 1.29 km from the railways;
- 0.85 km form the coast;
- 0.69 km form the highway and 18.45 km form the Capee Canaveral city.



Section 4

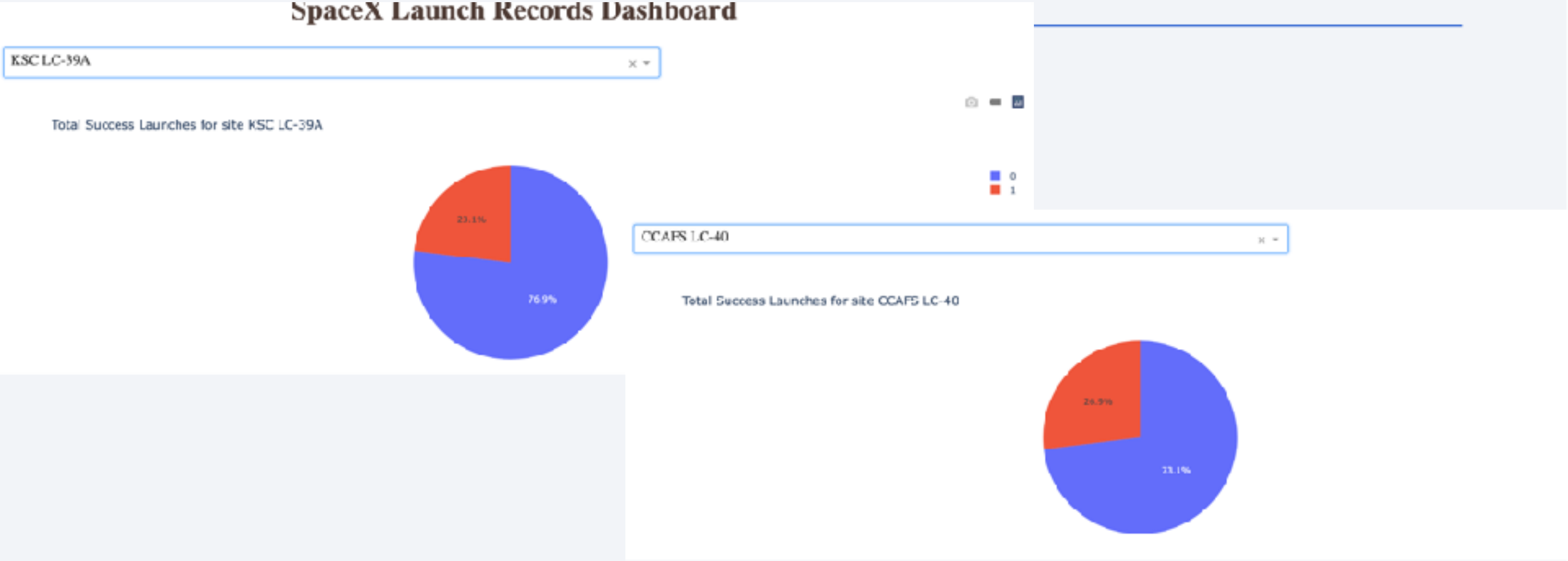
Build a Dashboard with Plotly Dash

Success rate distribution across sites



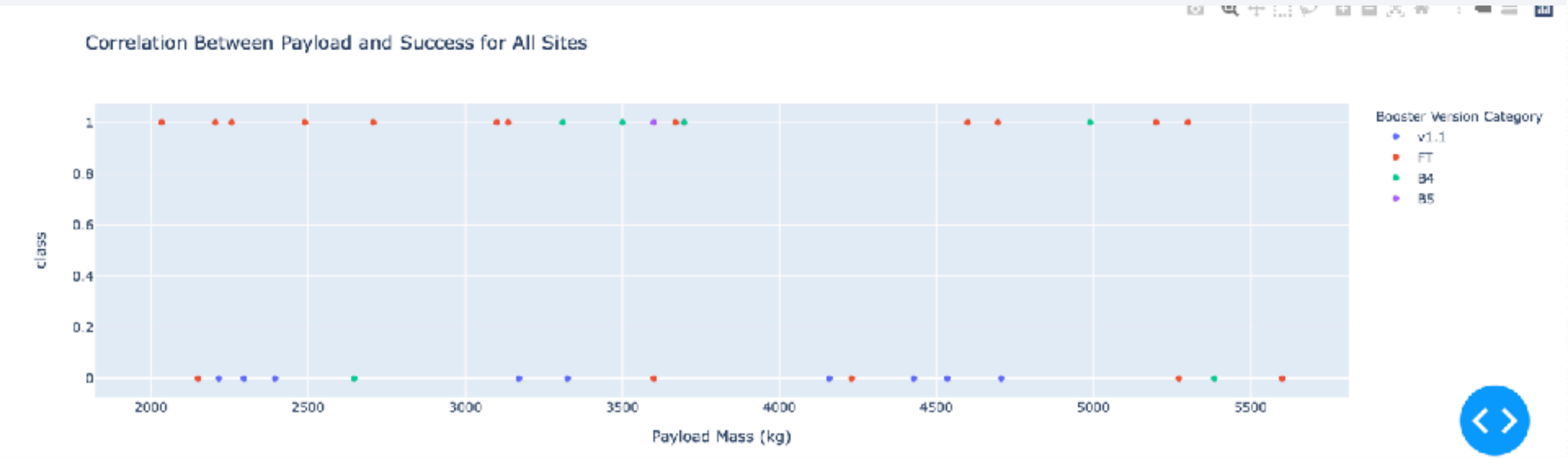
The most successful launches were at KSC-LC-39A

The most successful site



Although CCAFS LC-40 seems to be raging well behind in the previous graph, its share of successful launches is not the much beyond KSC LC-39A. The difference in the previous graph is related to the difference in the total number of launches

Payload and success

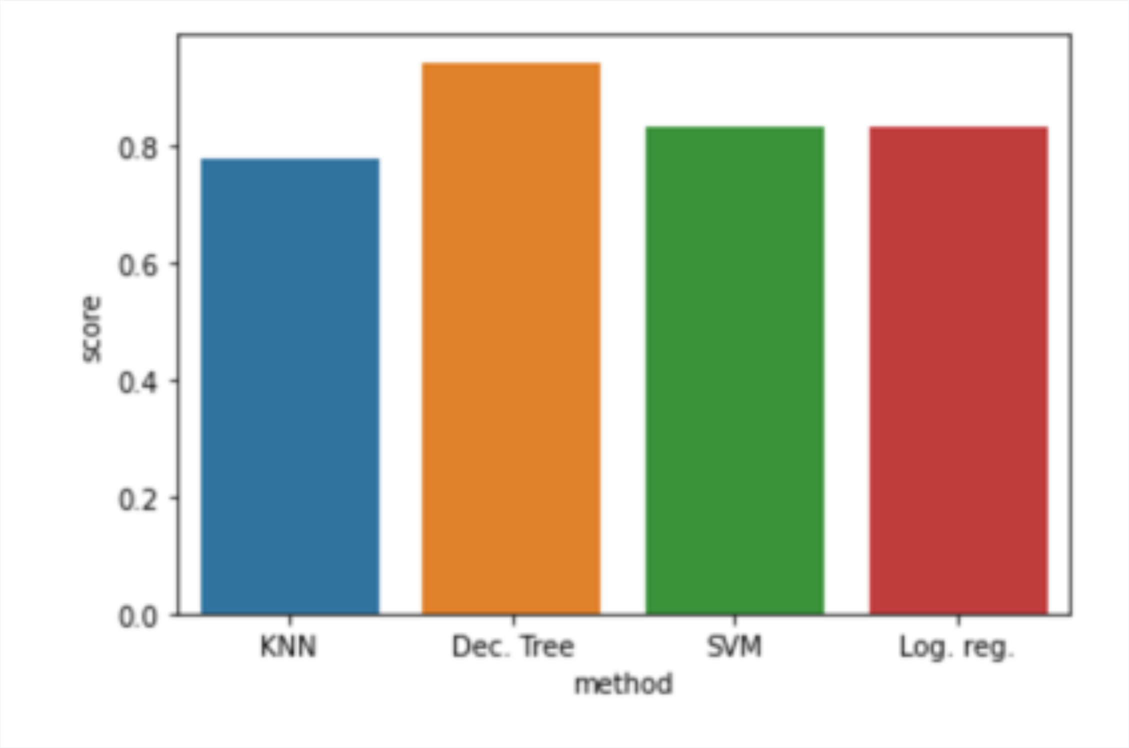


The strongest relationship between the Payload mass and success seems to be with FT booster version

Section 5

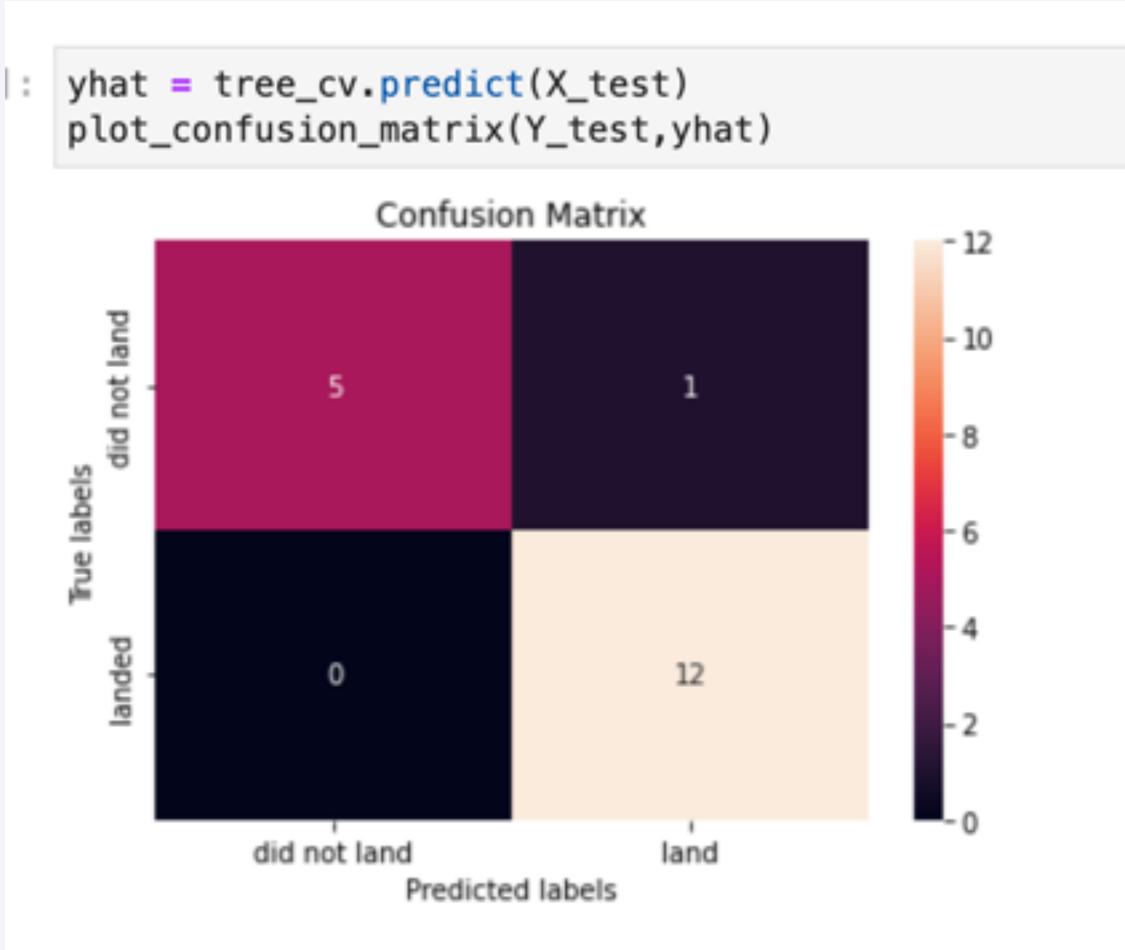
Predictive Analysis (Classification)

Classification Accuracy



Decision tree seems to be the best for predictions

Confusion Matrix



It is quite good: only one false positive :)

Conclusions

- Launching sites differ in the number of launches and success rates. The most successful site is KSC LC-39A: SpaceY must focus on there. Maybe some specific properties of this site must be investigated.
- The best classification method to be used for predictions is Decision tree.

Thank you!