IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Sandoval H.
12 – may - 2022

Outline

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

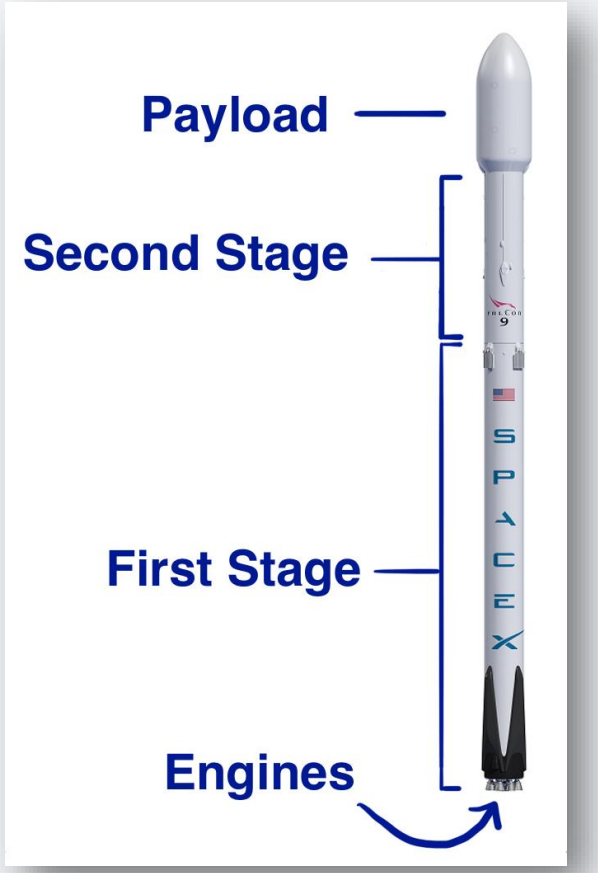
Executive Summary

- Summary of methodologies
- Summary of all results

To help determine the cost of rocket launches. We will predict if the Falcon 9 first stage will land successfully. We used open data from a SpaceX API and web scraping Wikipedia. Analize the relationships between the data available plotting charts, maps, and an interactive dashboard. Then selected the best features to test various machine learning models. The best performing model gave us an accuracy of about 0.84.

Introduction

- Project background and context
- Problems you want to find answers



The rocket launch have faced a huge cost reduction by reusing the first stage of the rocket. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

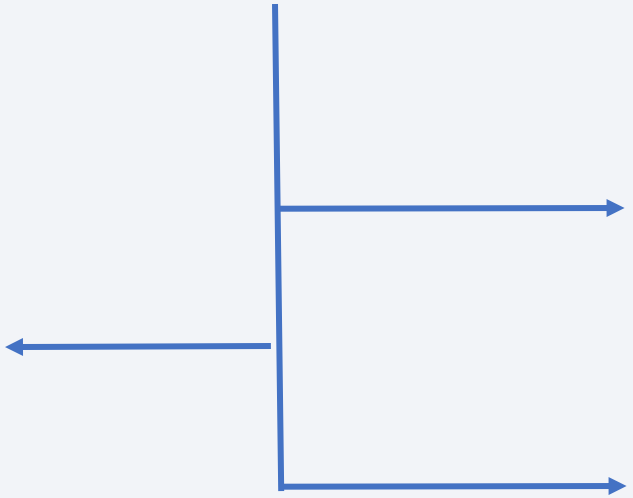
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

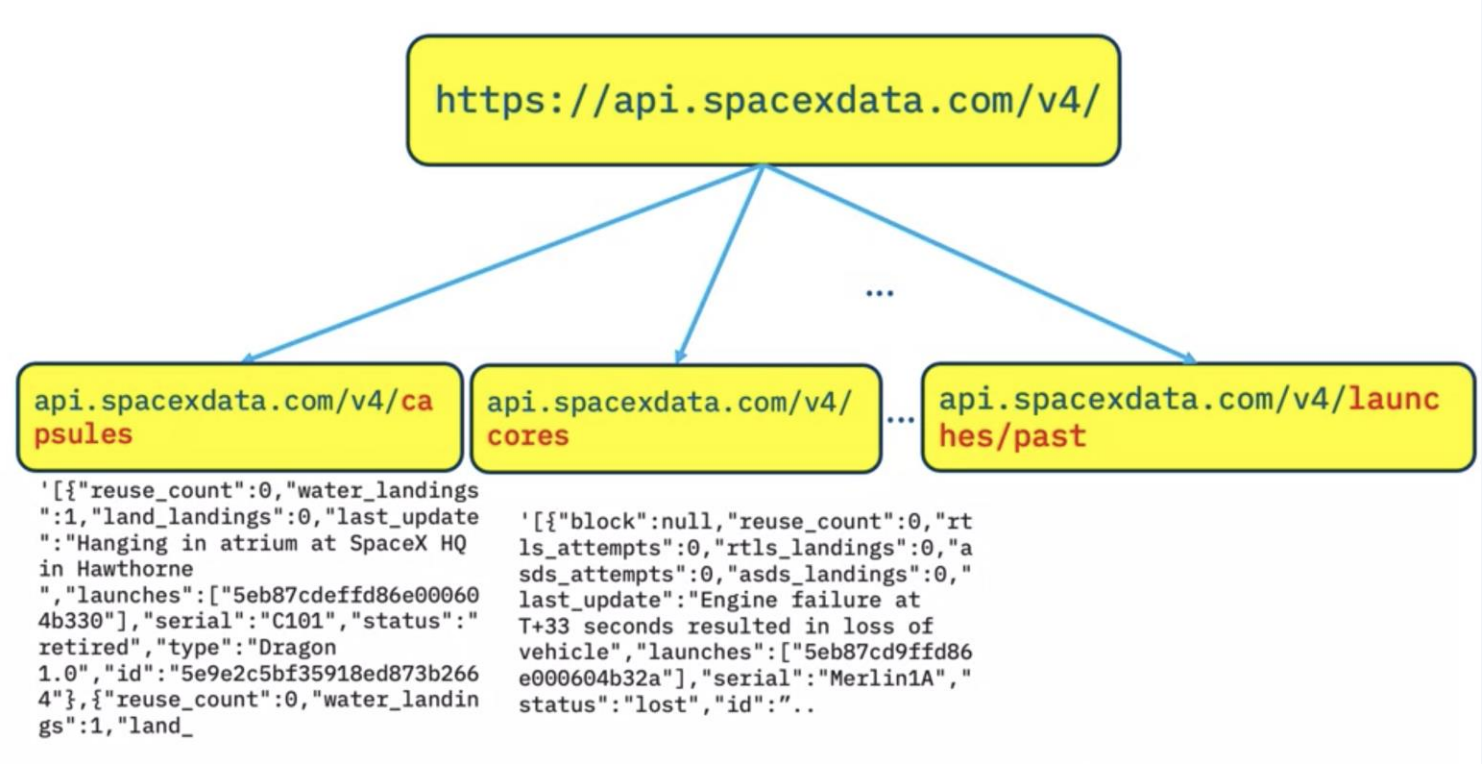
r-spacex/**SpaceX-API**

🚀 Open Source REST API for SpaceX launch, rocket, core, capsule, starlink, launchpad, and landing pad data.

83 Contributors | 15 Issues | 1 Discussion | 9k Stars | 798 Forks



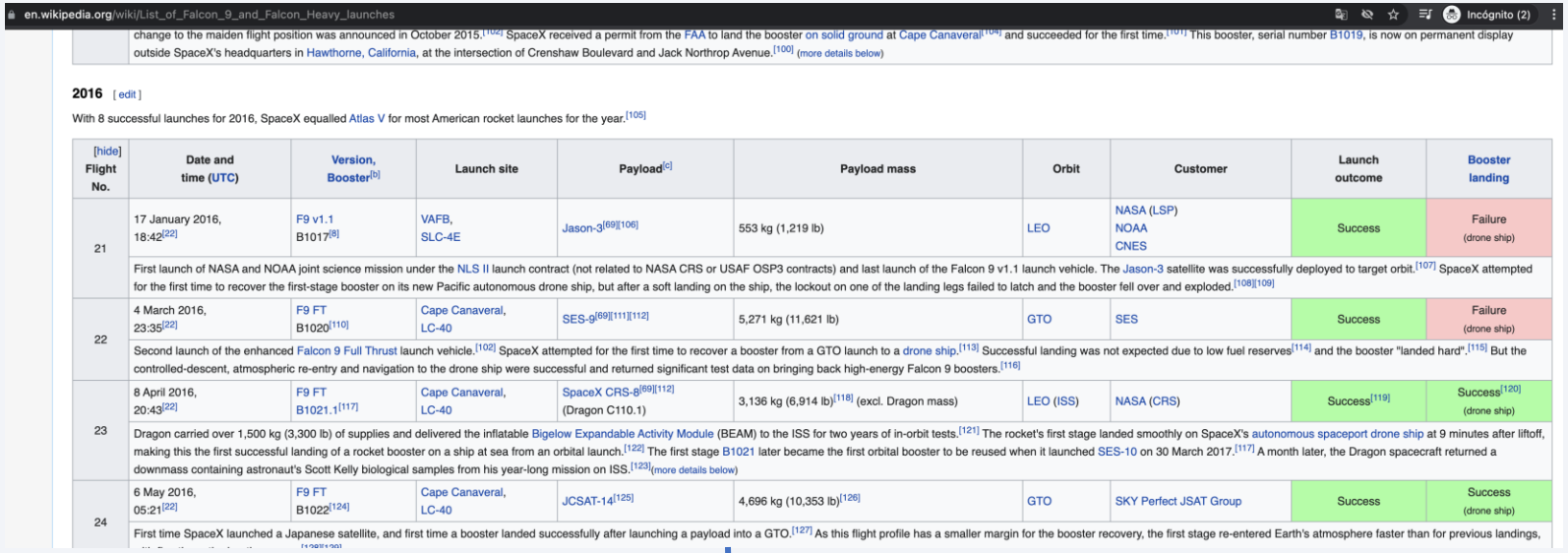
Data Collection – SpaceX API



[Click here to go to Notebook explaining this process]

(https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/1_1_data_collection_api.ipynb)

Data Collection – Scraping - Wikipedia



en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

change to the maiden flight position was announced in October 2015.[102] SpaceX received a permit from the FAA to land the booster on solid ground at Cape Canaveral[104] and succeeded for the first time.[101] This booster, serial number B1019, is now on permanent display outside SpaceX's headquarters in Hawthorne, California, at the intersection of Crenshaw Boulevard and Jack Northrop Avenue.[100] (more details below)

2016 [edit]

With 8 successful launches for 2016, SpaceX equalled Atlas V for most American rocket launches for the year.[105]

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
21	17 January 2016, 18:42[22]	F9 v1.1 B1017[8]	VAFB, SLC-4E	Jason-3[69][106]	553 kg (1,219 lb)	LEO	NASA (LSP) NOAA CNES	Success	Failure (drone ship)
	First launch of NASA and NOAA joint science mission under the NLS II launch contract (not related to NASA CRS or USAF OSP3 contracts) and last launch of the Falcon 9 v1.1 launch vehicle. The Jason-3 satellite was successfully deployed to target orbit.[107] SpaceX attempted for the first time to recover the first-stage booster on its new Pacific autonomous drone ship, but after a soft landing on the ship, the lockout on one of the landing legs failed to latch and the booster fell over and exploded.[108][109]								
22	4 March 2016, 23:35[22]	F9 FT B1020[110]	Cape Canaveral, LC-40	SES-9[69][111][112]	5,271 kg (11,621 lb)	GTO	SES	Success	Failure (drone ship)
	Second launch of the enhanced Falcon 9 Full Thrust launch vehicle.[102] SpaceX attempted for the first time to recover a booster from a GTO launch to a drone ship.[113] Successful landing was not expected due to low fuel reserves[114] and the booster "landed hard".[115] But the controlled-descent, atmospheric re-entry and navigation to the drone ship were successful and returned significant test data on bringing back high-energy Falcon 9 boosters.[116]								
23	8 April 2016, 20:43[22]	F9 FT B1021.1[117]	Cape Canaveral, LC-40	SpaceX CRS-8[69][112] (Dragon C110.1)	3,136 kg (6,914 lb)[118] (excl. Dragon mass)	LEO (ISS)	NASA (CRS)	Success[119]	Success[120] (drone ship)
	Dragon carried over 1,500 kg (3,300 lb) of supplies and delivered the inflatable Bigelow Expandable Activity Module (BEAM) to the ISS for two years of in-orbit tests.[121] The rocket's first stage landed smoothly on SpaceX's autonomous spaceport drone ship at 9 minutes after liftoff, making this the first successful landing of a rocket booster on a ship at sea from an orbital launch.[122] The first stage B1021 later became the first orbital booster to be reused when it launched SES-10 on 30 March 2017.[117] A month later, the Dragon spacecraft returned a downmass containing astronaut's Scott Kelly biological samples from his year-long mission on ISS.[123](more details below)								
24	6 May 2016, 05:21[22]	F9 FT B1022[124]	Cape Canaveral, LC-40	JCSAT-14[125]	4,696 kg (10,353 lb)[126]	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
	First time SpaceX launched a Japanese satellite, and first time a booster landed successfully after launching a payload into a GTO.[127] As this flight profile has a smaller margin for the booster recovery, the first stage re-entered Earth's atmosphere faster than for previous landings,								

Selecting useful features

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

[Click here to go to Notebook explaining this process]

(https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/1_2_Webscraping.ipynb)

Data Wrangling

- Checked for null values and oddities.
- Got a sense of features caracteristics such as feature type, count per category, mean and others.

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

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

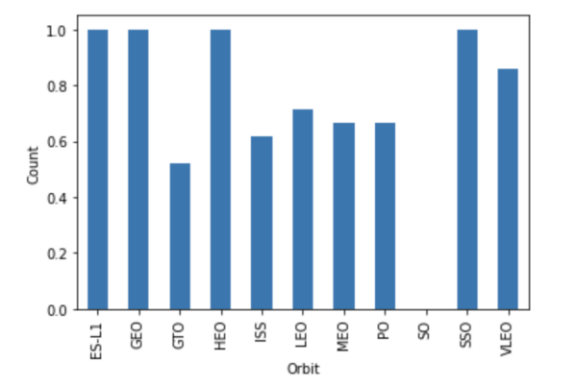
```
[ ]:  Orbit  Outcome
      ES-L1  True Ocean      1
      GEO    True ASDS       1
      GTO    False ASDS      1
             None ASDS       1
             None None      11
             True ASDS      13
             True Ocean      1
      HEO    True ASDS       1
      ISS    False ASDS      2
             False Ocean     1
             False RTLS      1
             None ASDS       1
             None None       3
             True ASDS       5
             True Ocean      1
             True RTLS       7
      LEO    None None       2
             True Ocean      1
             True RTLS       4
      MEO    None None       1
             True ASDS       2
```

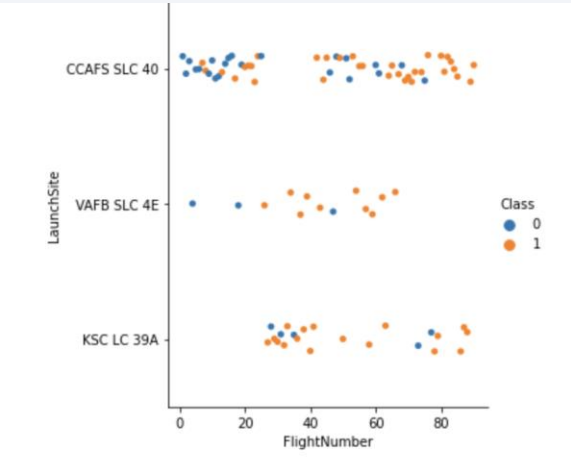
Click here to go to Notebook explaining this process

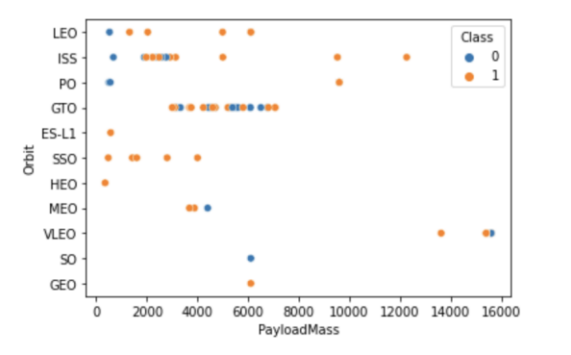
https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/1_3_Data_wrangling.ipynb

EDA with Data Visualization

- We use scatterplot (catplot), bar charts and line chart to get a look at how data is distributed. In these plots was also used combinations of features.







[Click here to go to Notebook explaining this process]

https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/2_2_Eda_Visualization.ipynb

EDA with SQL

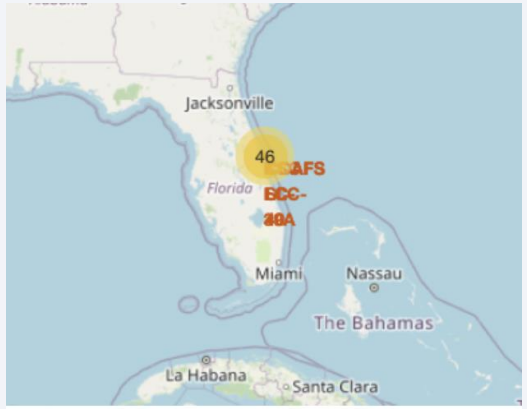
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for 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

Click here to go to Notebook explaining this process

https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/2_1_Exploratory_Data_Analysis_SQL_(RESUELTO_IBM_CLOUD)_.ipynb

Build an Interactive Map with Folium

Ubicated the location sites on a interactive folium map to get a sense of where the launches are from. At the map was also marked each launch in its corresponding place and were given a color different color depending if the recovery of the first stage succeed. Finally, we looked at the lauch sites proximities.

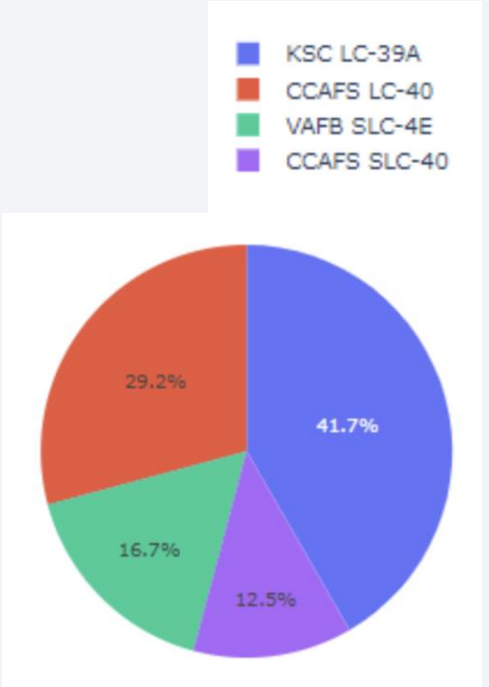


[Click here to go to Notebook explaining this process]

https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/3_1_Launch_site_location_MAPS.ipynb

Build a Dashboard with Plotly Dash

- To get a better sense of the relationship between success rate, payload mass and booster type, a interactive dashboard was build.
- It have pie chart that changes acording to the launch site. If all sites are selected, it show the proportioin each site contribute to all the success launches. Otherwise it will show the proportion of success (1) launches and failure launches (1) of the corresponding location.



Click here to go to Notebook explaining this process

https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/3_2_Interactive_Dashboard_DASH.ipynb

Build a Dashboard with Plotly Dash

A Scatterplot wich change acording to the selected site and the selected payload mass range. It will show you if the distribution of successes and failures of launches and its correlation with the payload mass, aditionally you may get some insight on the relationships between those two variables and the Booster Version Category.

Click here to go to Notebook explaining this process

https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/3_2_Interactive_Dashboard_DASH.ipynb

Predictive Analysis (Classification)

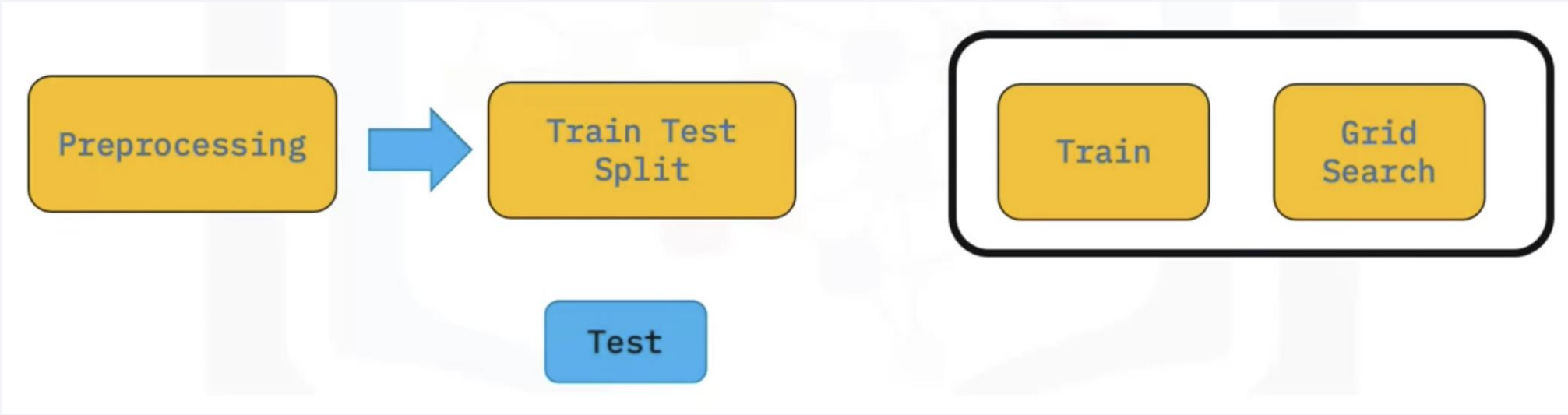
- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Click here to go to Notebook explaining this process

https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/4_1_Machine_Learning_Prediction.ipynb

Predictive Analysis (Classification)

- First all the features were preprocessed to fit the models and were split into a train set to train the algorithm and a test set to see how accurate it would be in real life

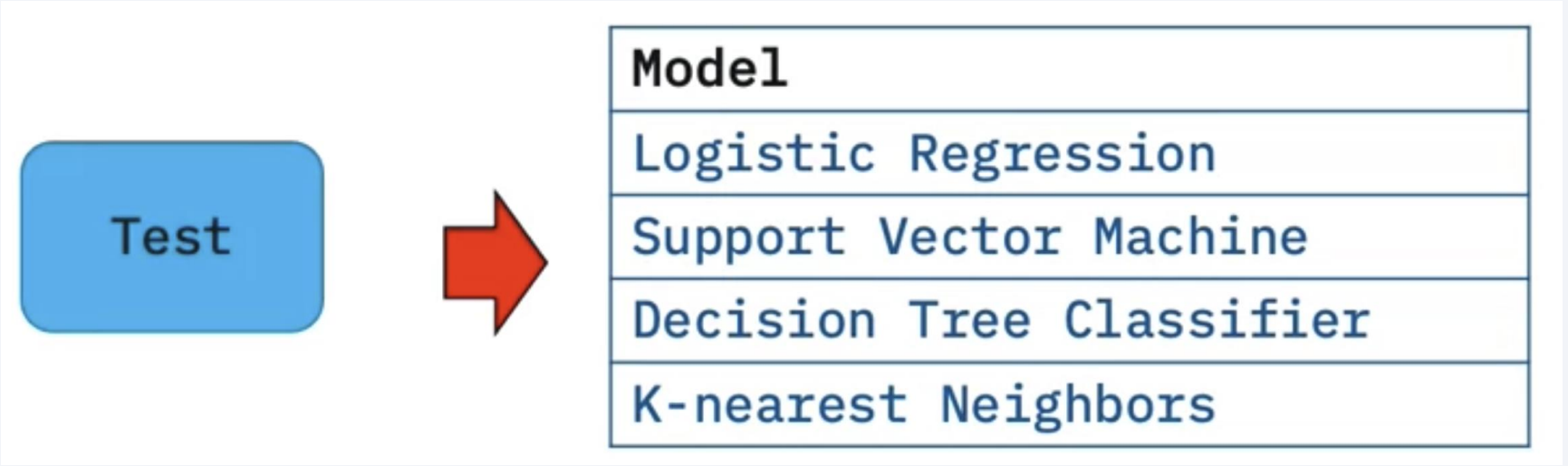


Click here to go to Notebook explaining this process

https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/4_1_Machine_Learning_Prediction.ipynb

Predictive Analysis (Classification)

- Grid Search was used to select the best parameter for each model. Then each model was tested on the test set and the model with best accuracy was selected.



Click here to go to Notebook explaining this process

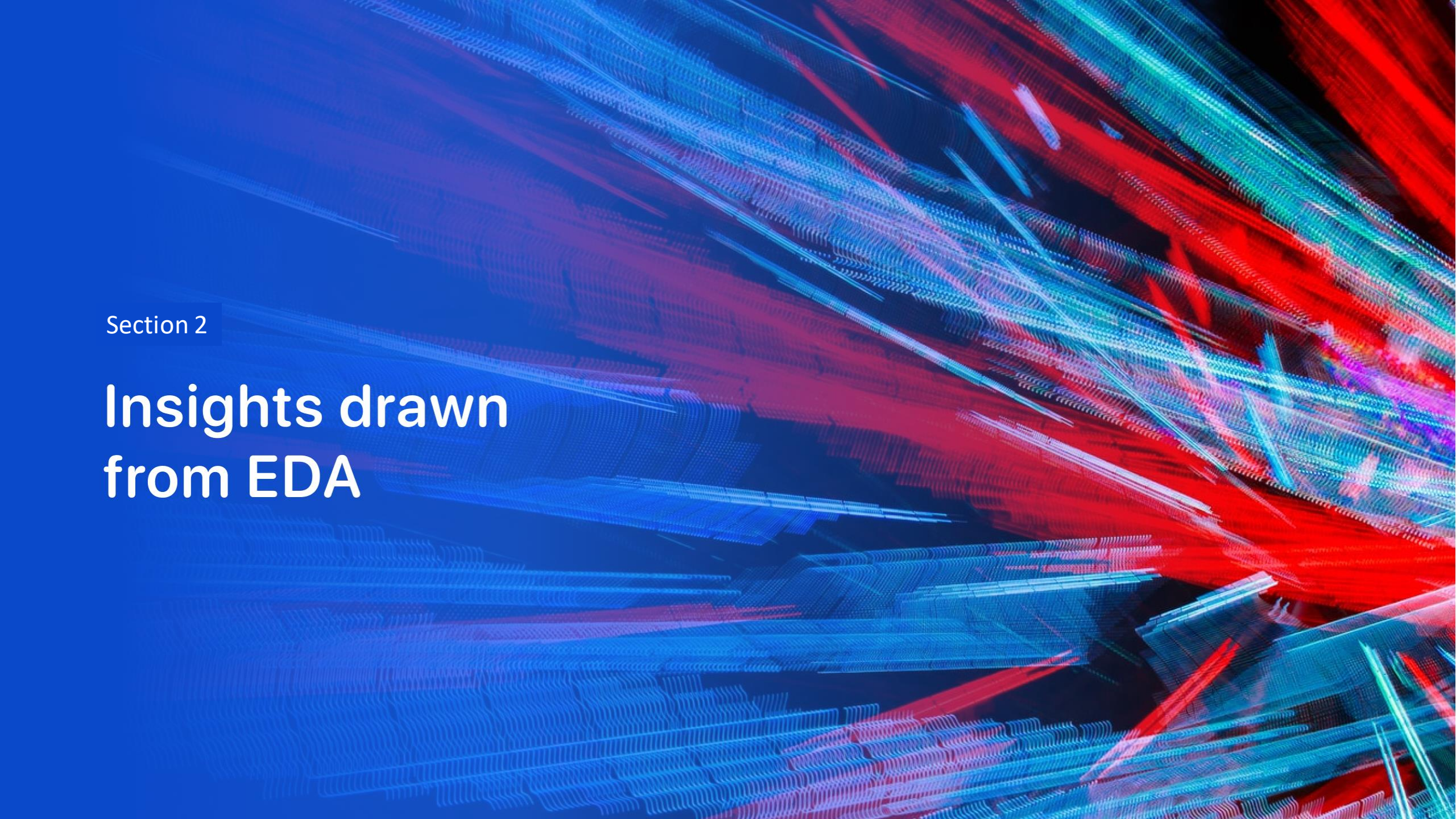
https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/blob/main/10_SpaceX_Capstone/4_1_Machine_Learning_Prediction.ipynb

Results

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

Section 2

Insights drawn from EDA

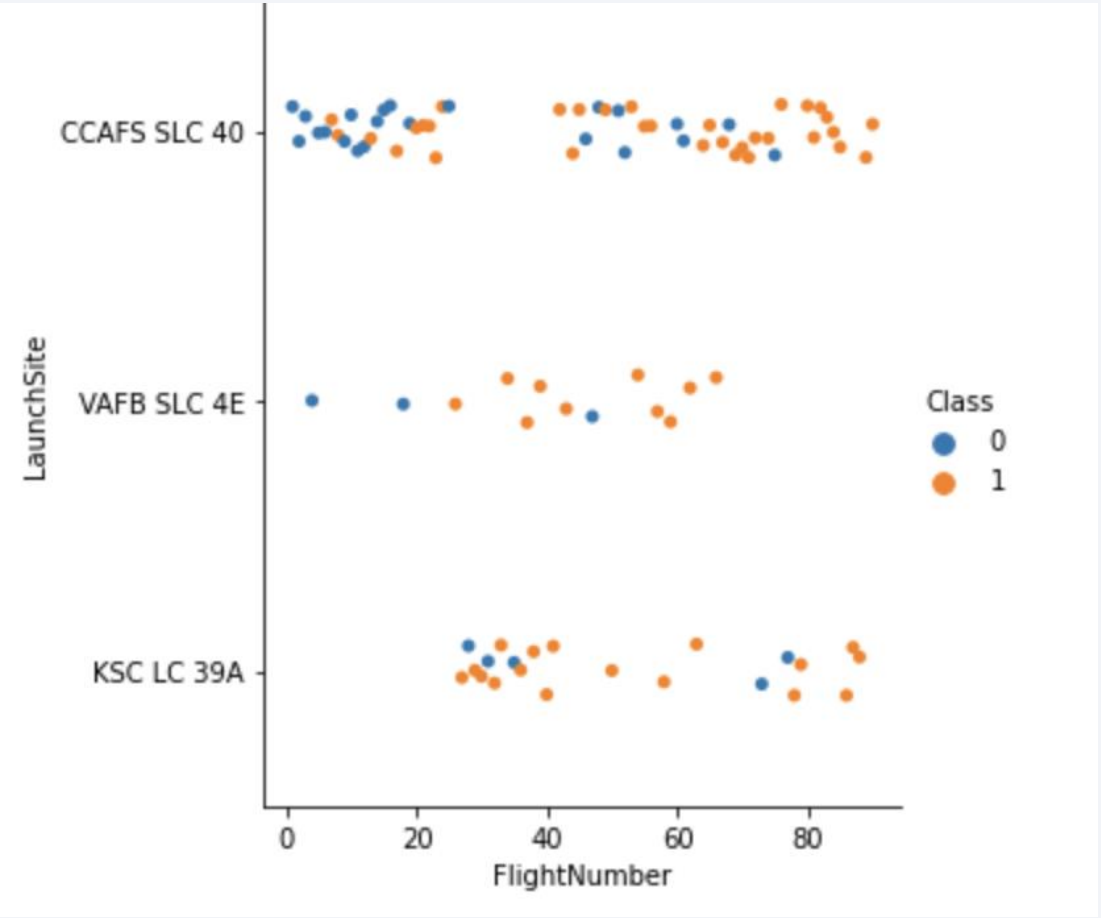


Flight Number vs. Launch Site

- We can observe that as 'FlightNumber' increases the 'Class' tend to converge to 1. (This may be because the people working on the rockets had gain more expertise from the previous launches.

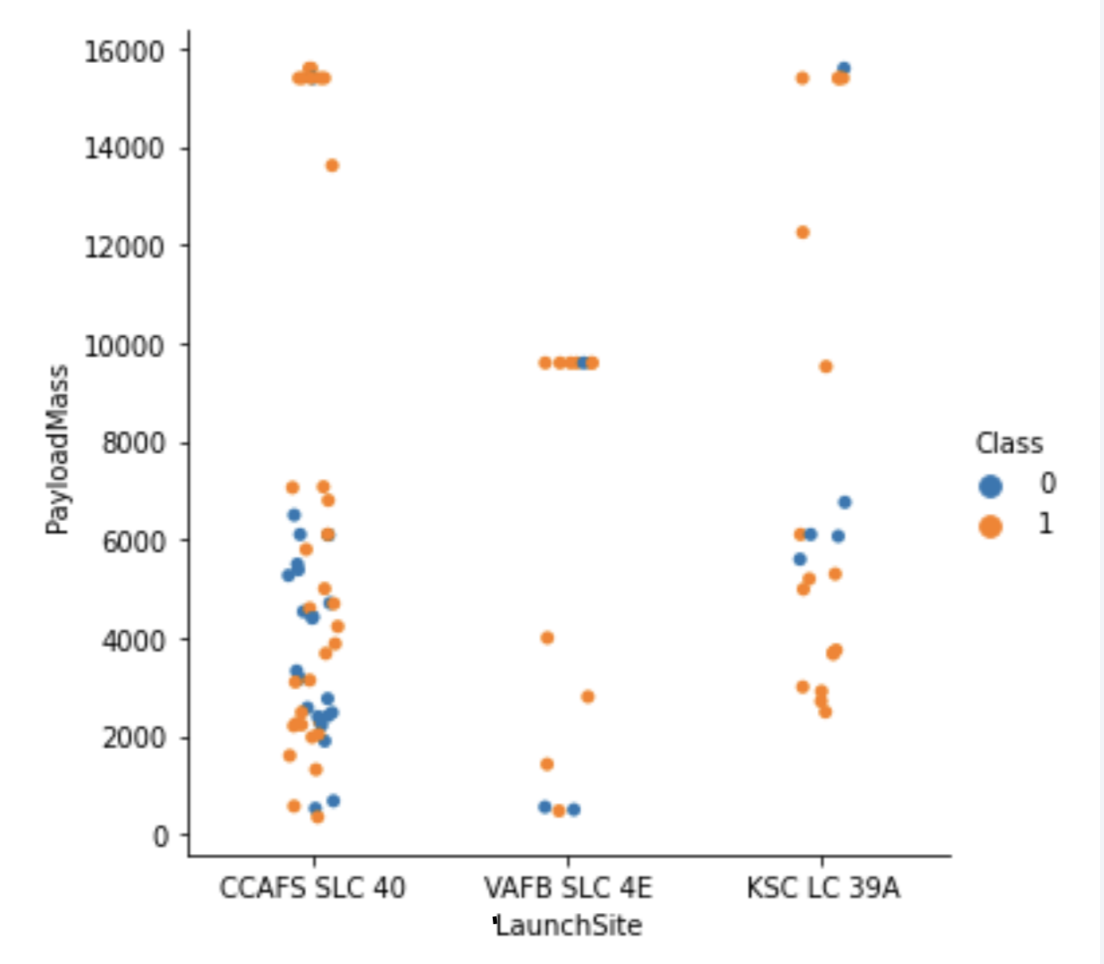
- I was shown that we have more examples of 'CCAFS SLC 40' launches Sites than in the other 2.

- Note: the calss 0 means the first stage was not recovered and the 1 mean it was recovered



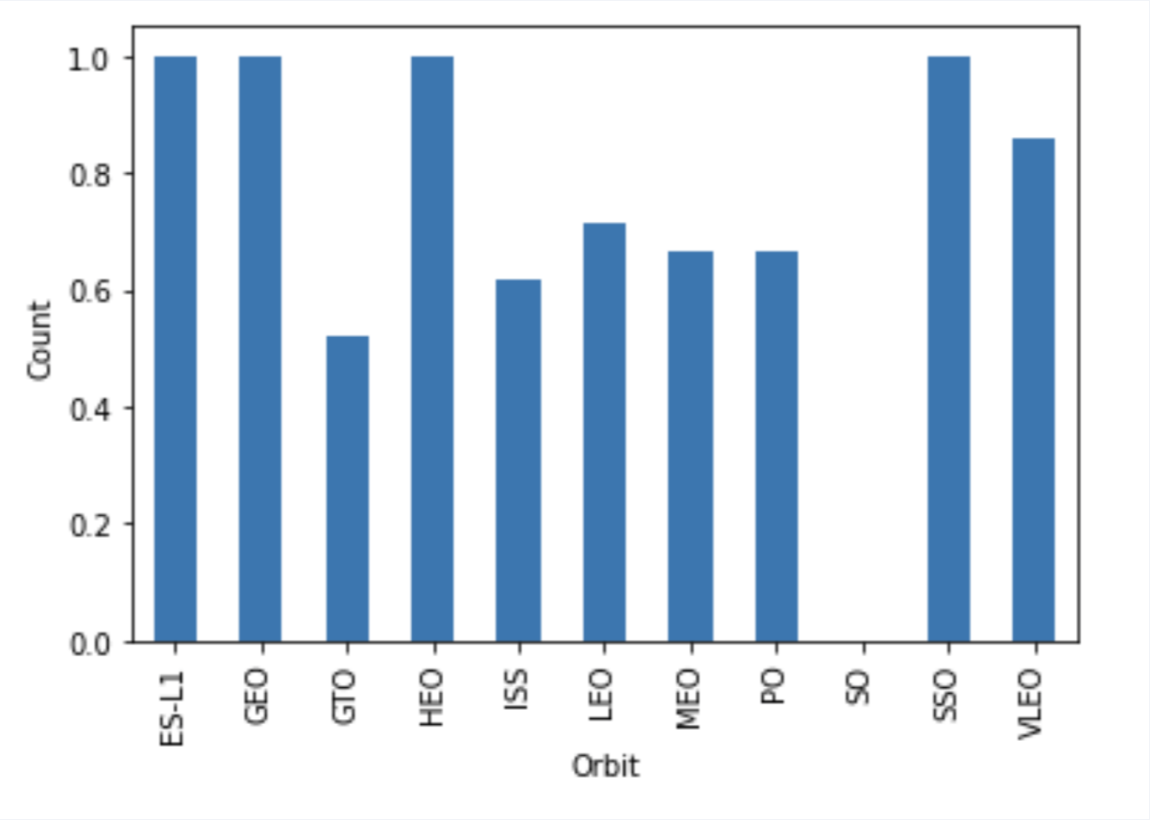
Payload vs. Launch Site

- VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000) and in the other two tend to have a payload Mass of less than 8000.



Success Rate vs. Orbit Type

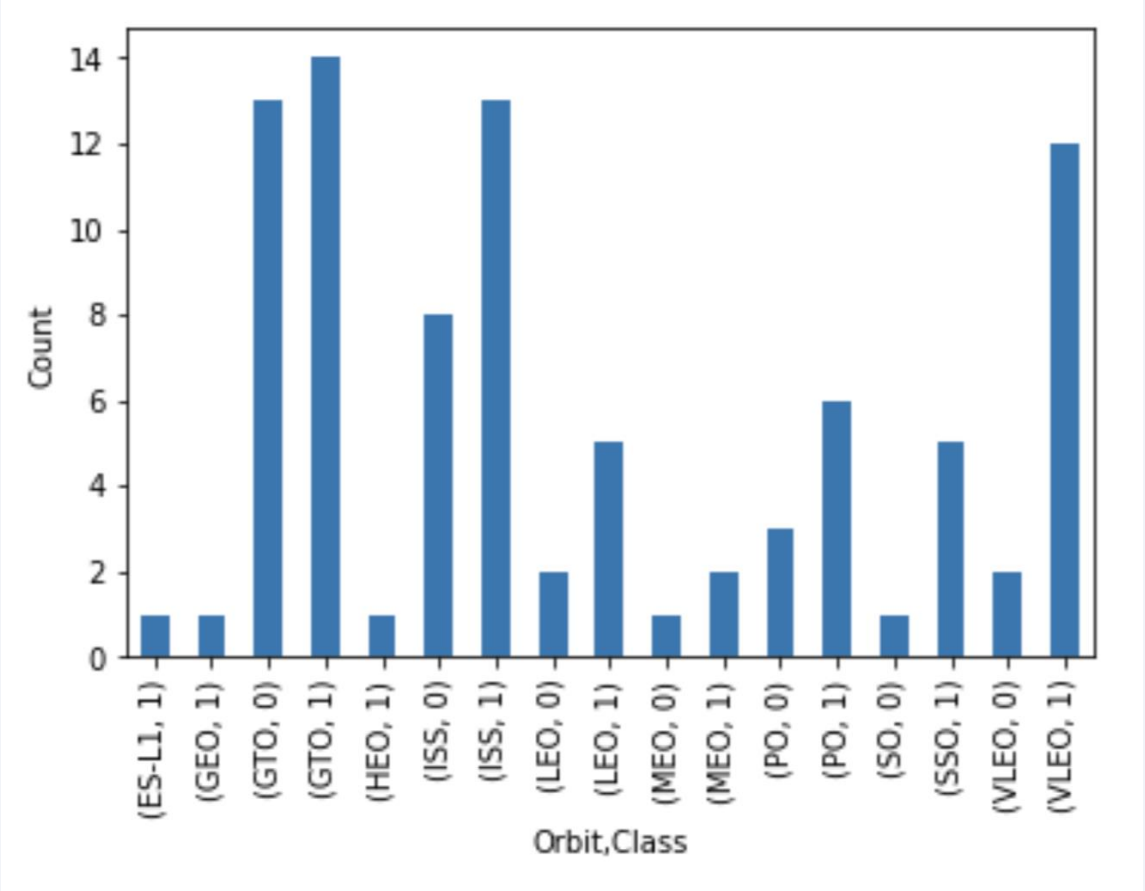
- 1 represents a 100%



- we have some orbits that have an almost perfect success rate

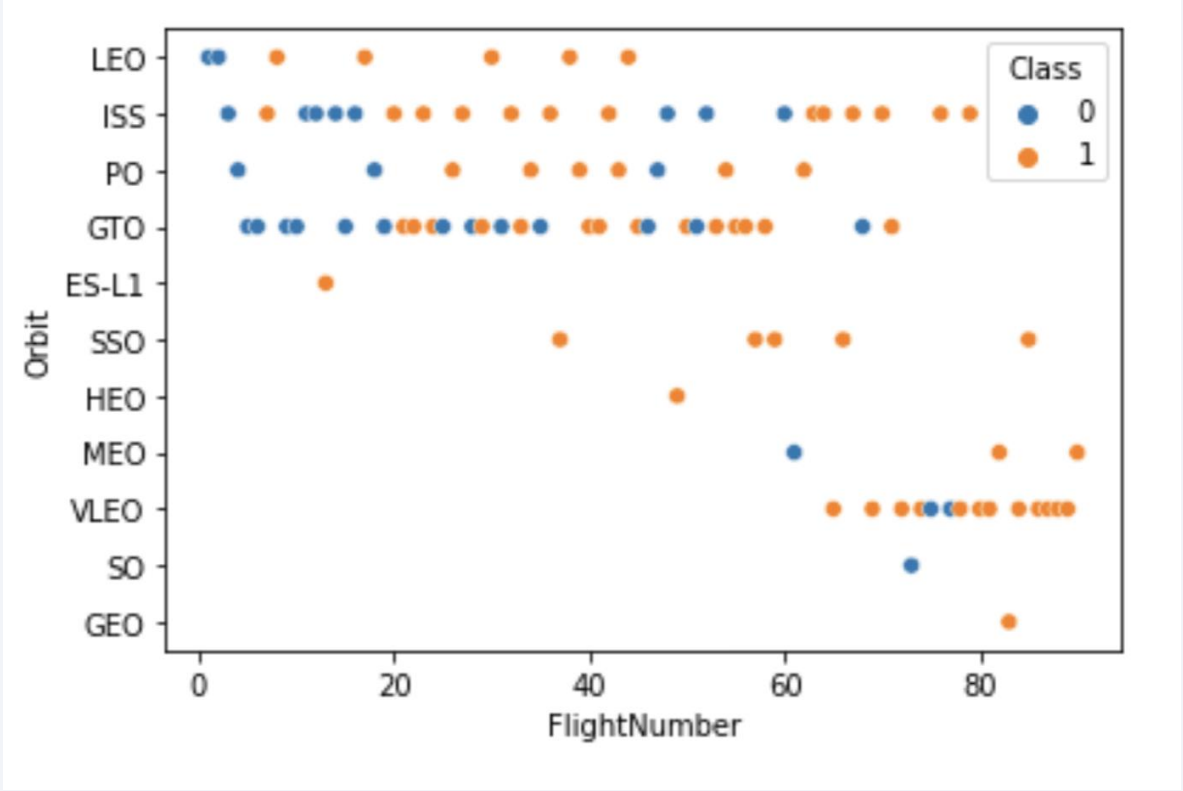
Success Rate vs. Orbit Type

- This chart complement the previous one.
- we have a very different launches amount for each orbit.
- Generally, the Orbits where we have a lot of launches do not have a high success rate



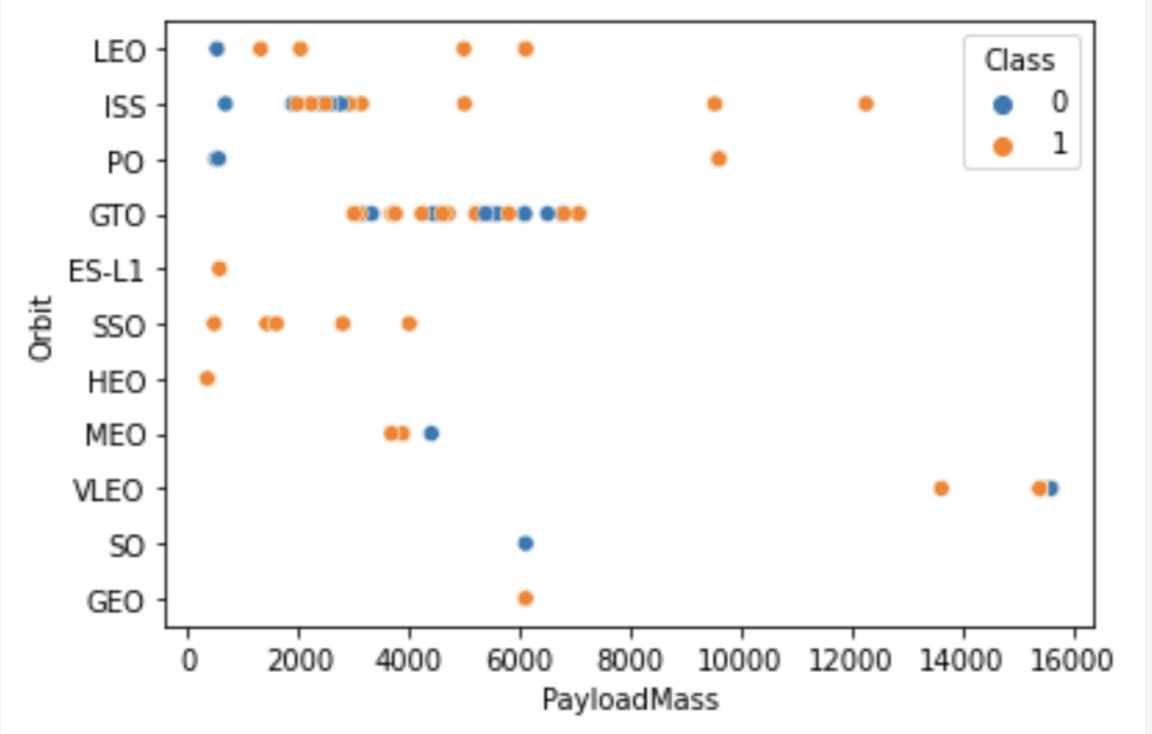
Flight Number vs. Orbit Type

- 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.
- As the FlightNumber increases we aso see more lauches to dirrerent types of orbits



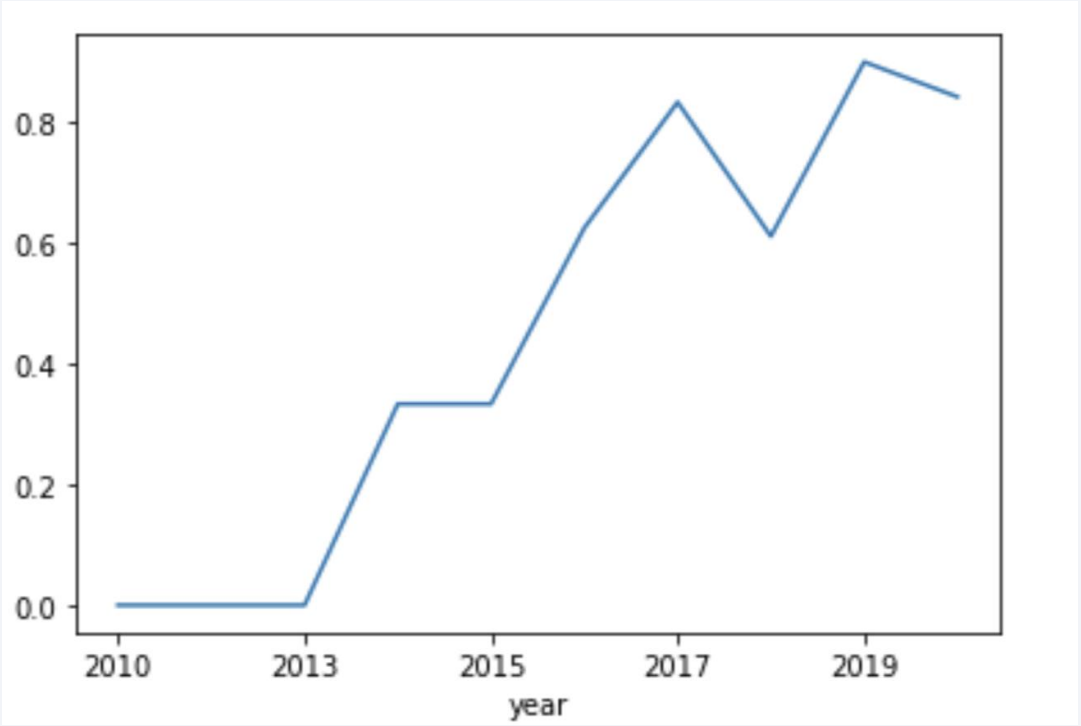
Payload vs. Orbit Type

- 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.
- It is curious that GTO and ISS ORBITS tend to have a very grouped payload mass.



Launch Success Yearly Trend

- At the y axis 1 is the success of first stage recovery.
- We observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

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

The launches were only develop at 4 sites.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEX where  launch_site Like 'CCA%' limit 5
```

```
 * ibm_db_sa://dgc21618:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

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

5 of the launch records which launch site begin with CCA

Total Payload Mass

```
%sql select sum(payload_mass__kg_) from SPACEX where  customer='NASA (CRS)'
```

```
 * ibm_db_sa://dgc21618:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqr
Done.
```

1
45596

- In total we have 45596 kg of payload

Average Payload Mass by F9 v1.1

```
%sql select avg(DATE) from SPACEX where  booster_version='F9 v1.1'
```

```
 * ibm_db_sa://dgc21618:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3r
Done.
```

1
2928

- In average the payload mass of ships with booster version F9 v1.1 is 2928 kg.

First Successful Ground Landing Date

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

- (2015-12-22) First successful outcome in ground pad landing we have record.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select  booster_version from SPACEX where landing__outcome='Success (drone ship)' and payload_mass__kg_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://dgc21618:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/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

```
%sql select  count(mission_outcome) from SPACEX where mission_outcome='Success'
```

```
 * ibm_db_sa://dgc21618:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u9
Done.
```

1
99

- Had 99 success outcomes.

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

```
%sql select booster_version from (select * from SPACEX ORDER BY payload_mass__kg_ DESC) LIMIT 10
```

```
 * ibm_db_sa://dgc21618:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.app
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

2015 Launch Records

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

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

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

- 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

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


Section 3
Launch Sites
Proximities Analysis

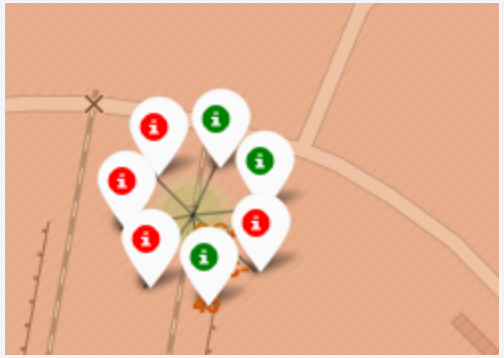
General Location Observation

We can observe that the launch sites are at United States Coasts near the Equator line. At the image at the right we have 3 sites very near. One of those is almost on top of other.

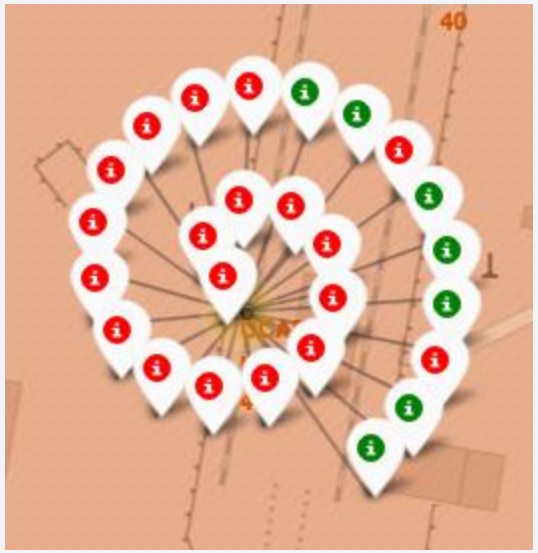
Success and Failures

VAFB SLC-4E

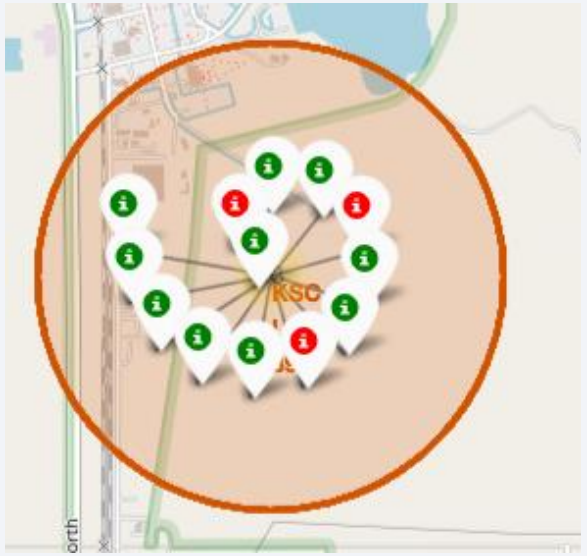
CCAFS SLC-40



CCAFS LC-40



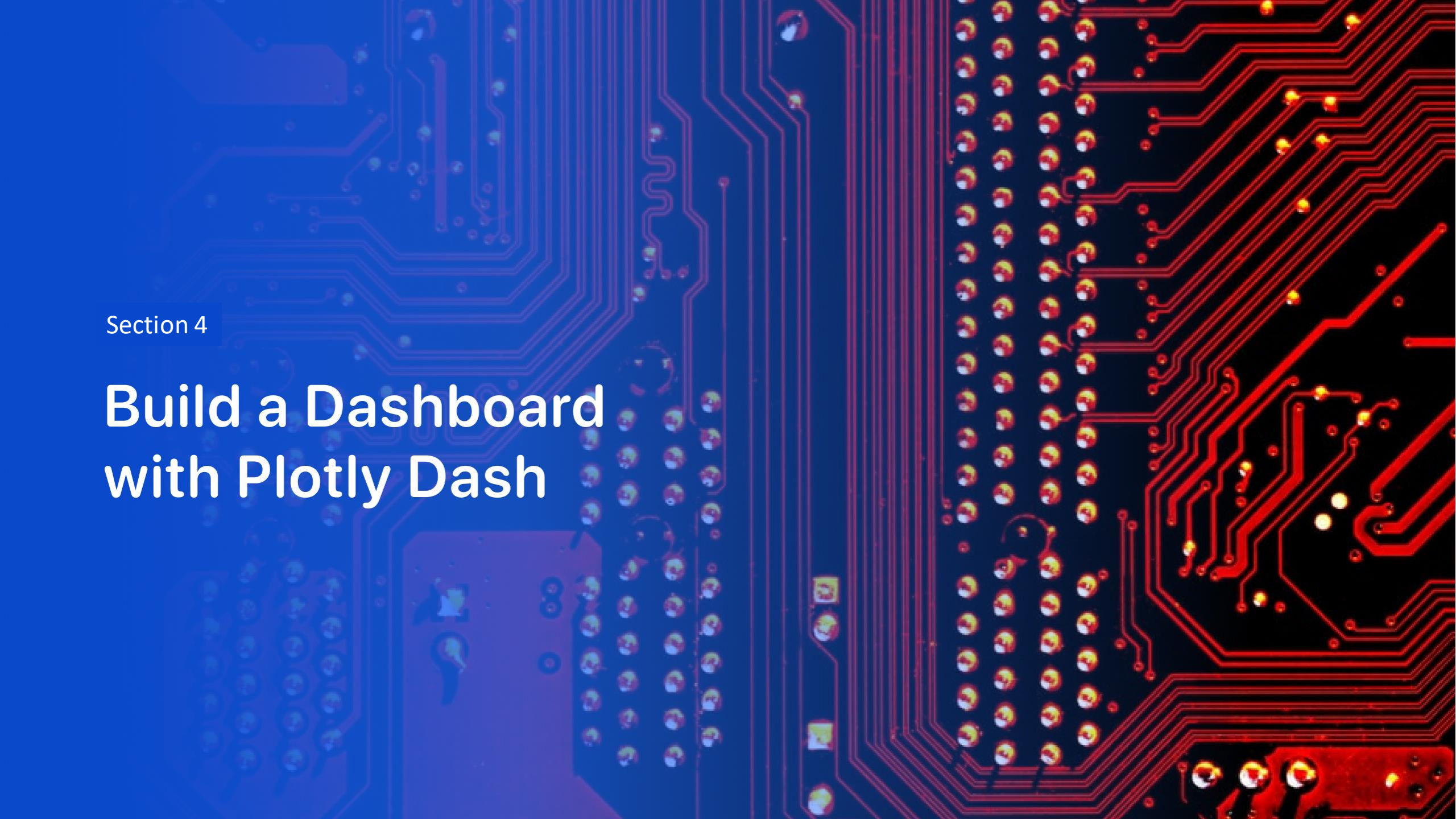
KSC LC-39A



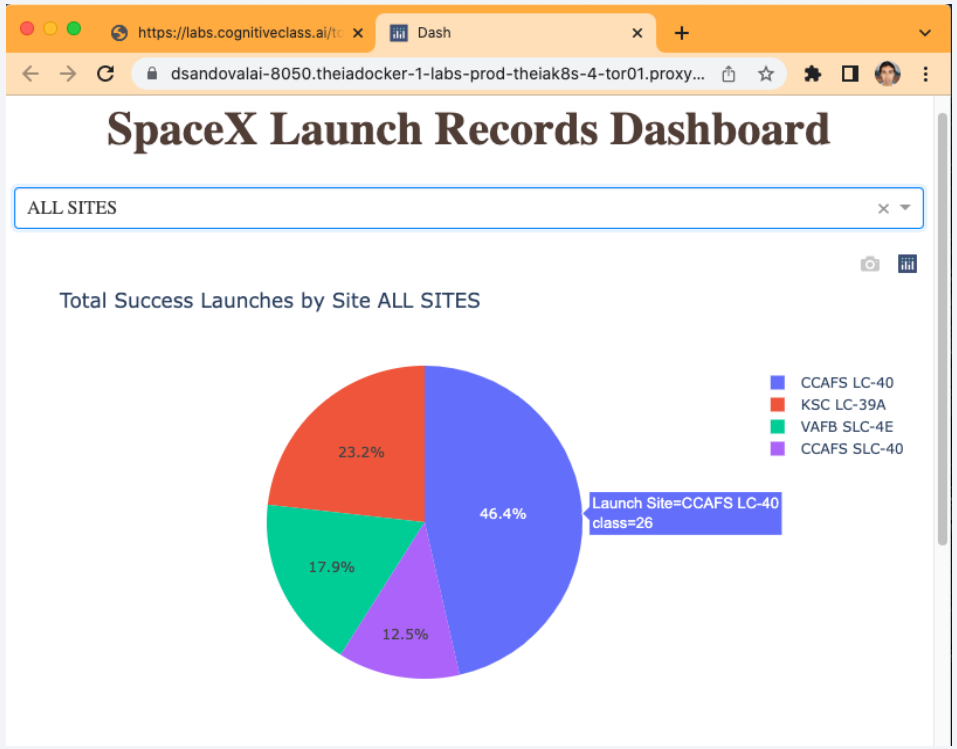
- On this images we observe the success (green) and failure (red) of first stage of launches of each site.

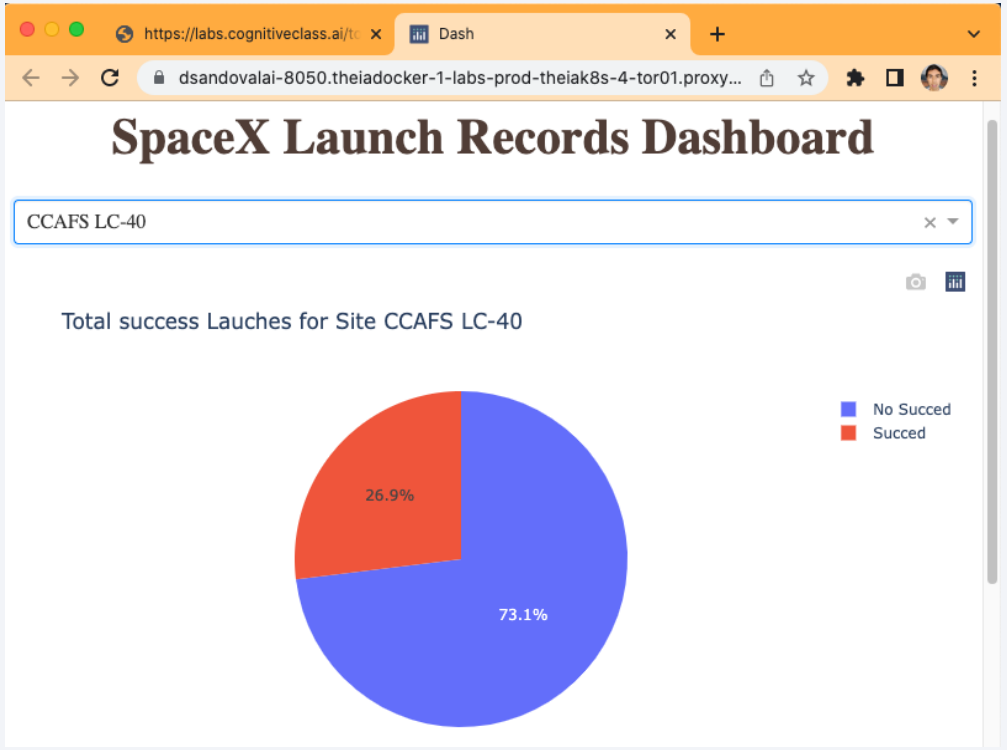
Proximities

At the maps we were able to see how the proximities of each launch sites, such as closest city, railway, highway, etc.


Section 4
Build a Dashboard with Plotly Dash

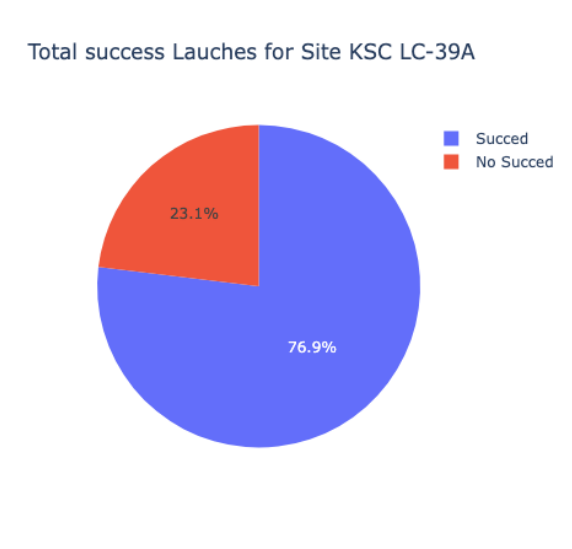
Dashboard Pie Chart Success / Launch Site

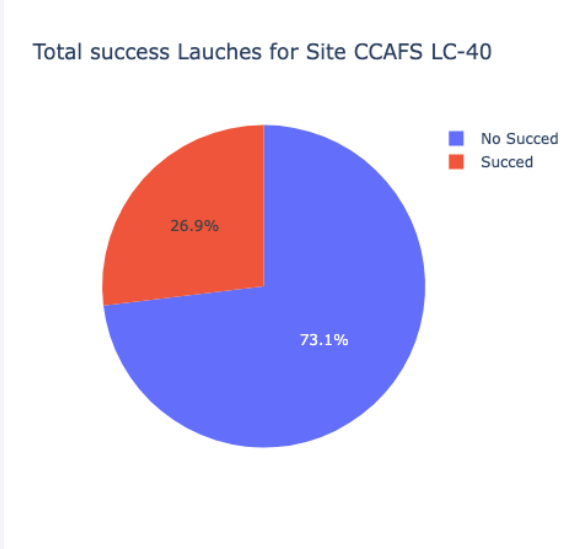




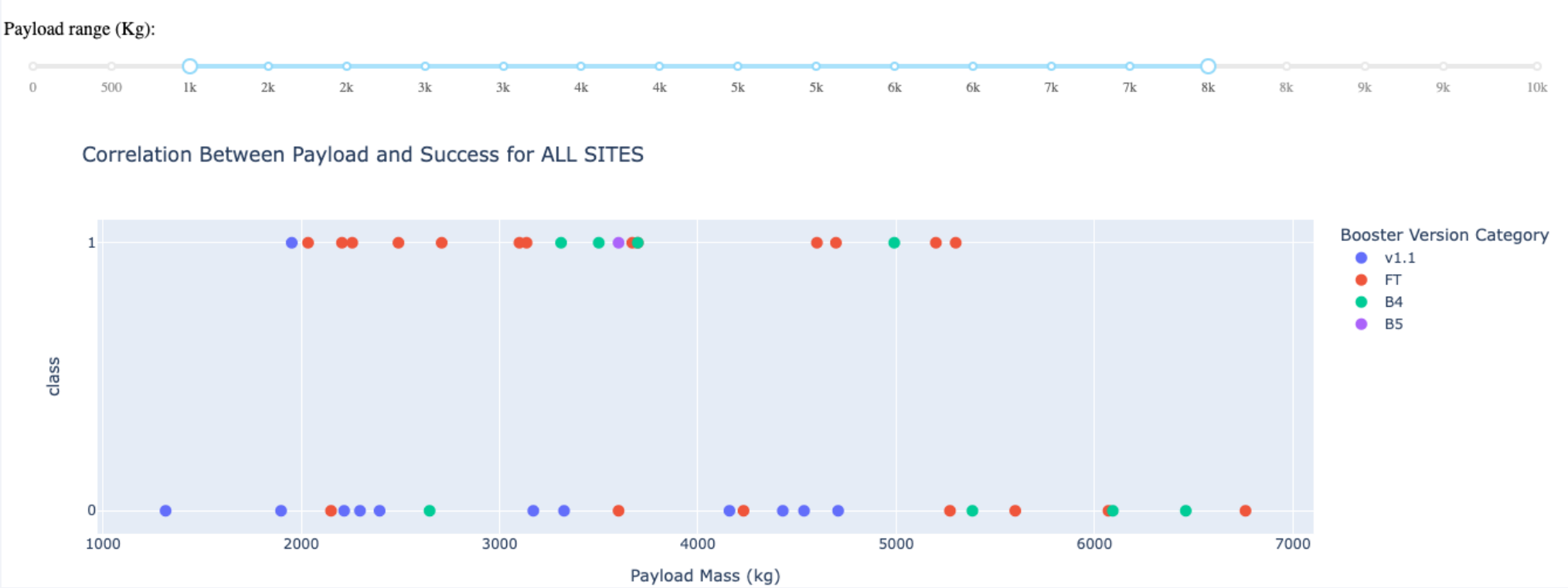
We have launch success principally from one site. However, it's also observed that between the site launches the success rate is about only ¼.

Launches at different Sites





<Dashboard Screenshot 3>



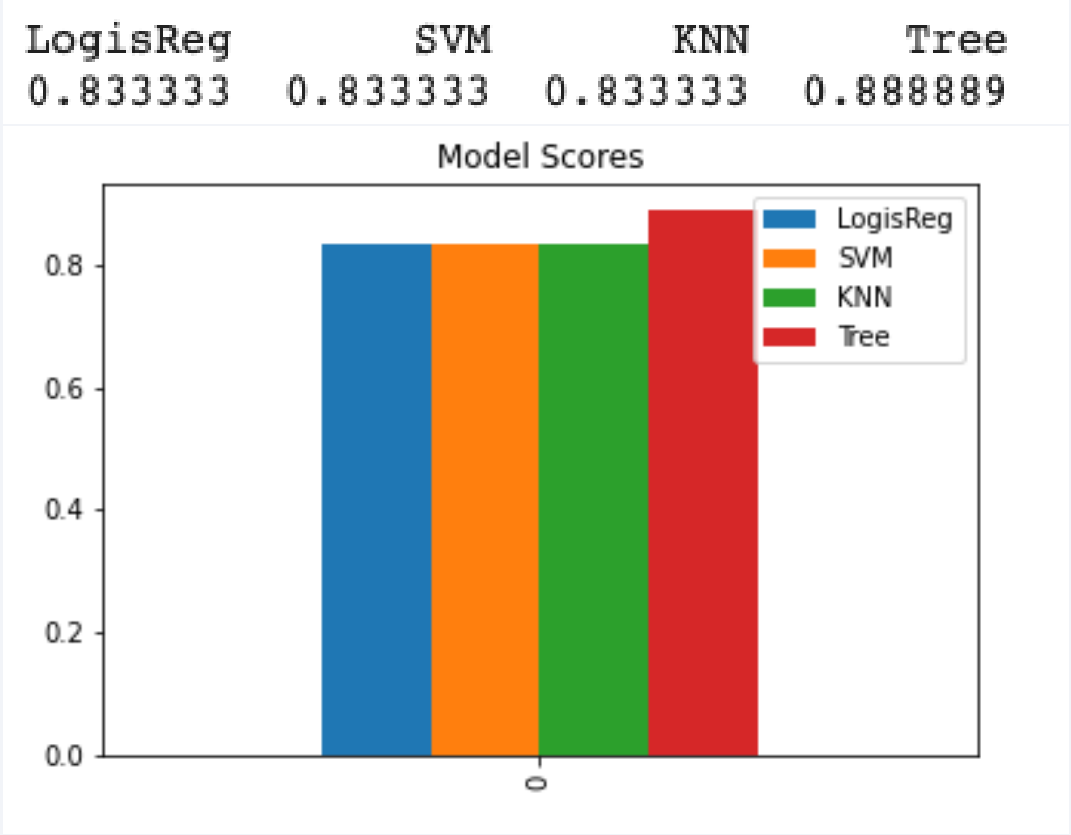
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

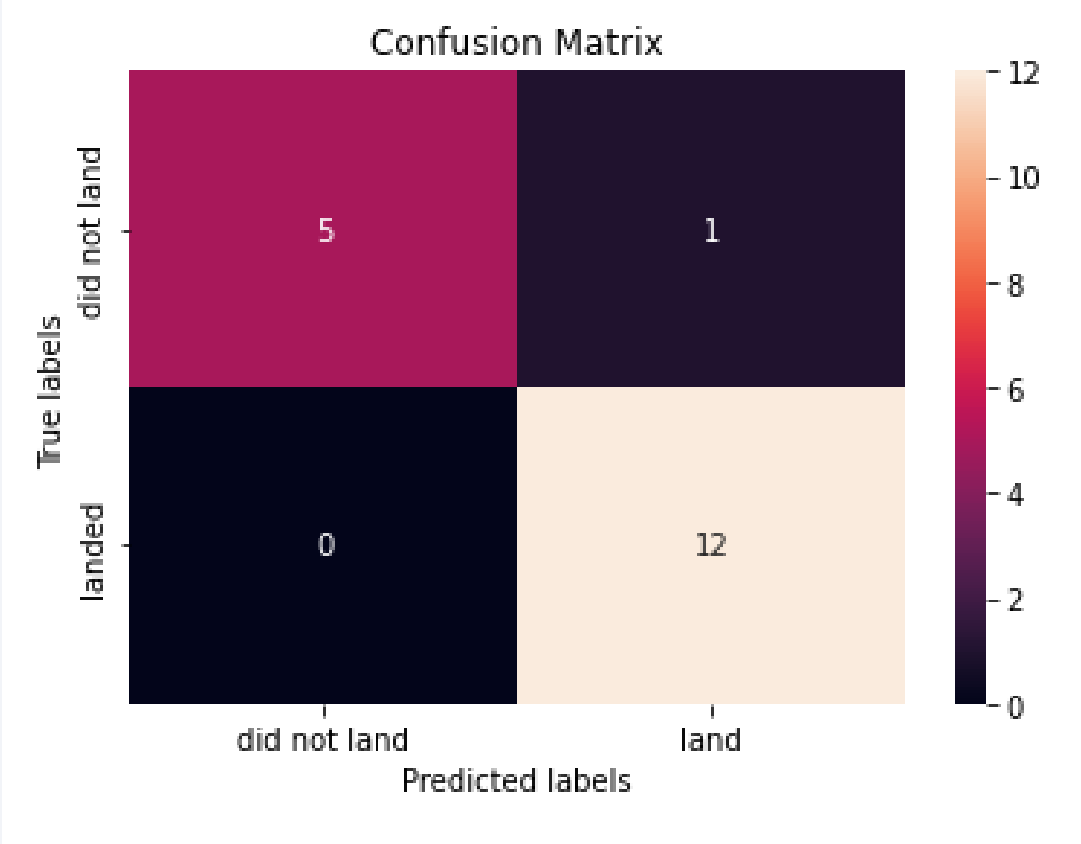
Classification Accuracy

LogisReg	SVM	KNN	Tree
0.833333	0.833333	0.833333	0.888889



The decision tree model was the best, performing with an accuracy on the test data of 88.8%

Confusion Matrix of Decision Tree Clasificator



- We can see that the predictions are almost perfect however we had a false positive.

Conclusions

- We can make a quite good prediction of the first stage recovery success (88.8%), it may be used to increase profits.
- If this model were mean to be deployed in real life, we should consider technological advances and new knoledge on the field, so it would be a countinuously updating model.
- This model was created using a very small set of data, future research may include other domain expertise features such as purpuose of launch, climate variables, etc.

Appendix

- This slides are part of the Ibm Data Science Capstone: www.coursera.org/learn/applied-data-science-capstone
- You can find the repository with detailed notebooks of this project at: https://github.com/danisshai/IBM-DATA-SCIENCE-PROGRAM/tree/main/10_SpaceX_Capstone

Thank you!