IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gabriela Oliveira
08/05/2022

Outline

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

Executive Summary

Summary of methodologies

Data was collected using web scraping. In jupyter notebook this data was cleaned and prepared. Further EDA were made using python, SQL and data visualization tools. Finally classification models were build.

Summary of all results

It was possible to draw conclusions over the outcomes accordingly to the launching parameters.

Classification models were successfully build to predict a launching outcome with ~83% of accuracy.

Introduction

Companies are making space travel affordable, and one of the most successful between them is SpaceX, which was founded in 2002 by Elon Musk. SpaceX advertises Falcon9 launches on its website by less then other providers. The reason this is possible is because SpaceX can reuse the first stage of the launchings. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

As a data scientist of an alternate company, Space Y, we want to determine if it's possible to reuse the first stage depending of the launching setup.

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

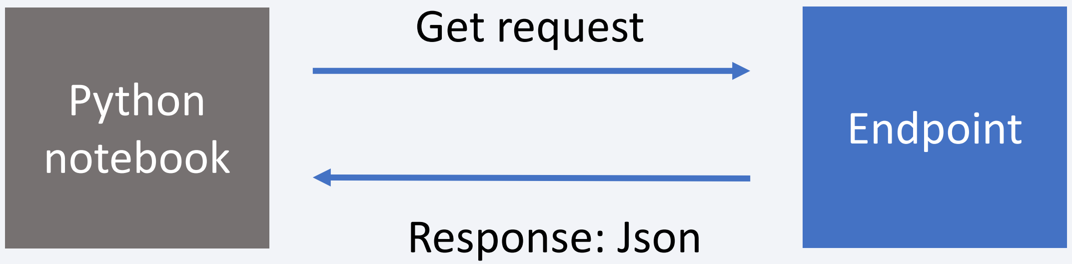
Data Collection

Data was collected using and SpaceX API, the endpoints starts with api.spacexdata.com/v4/. The main endpoint used to get launch data was api.spacexdata.com/v4/launches/past by performing a get request using the requests library. A list of json files was obtained and to convert them to a dataframe, a json_normalize function was used. Also, functions were provided to access additional information

Another part of the data was collected by web scraping Wikipedia Pages with tables containing information on Falcon9 launches using the Beautiful Soup packages.

Data Collection – SpaceX API

Data collection scheme by using SpaceX Endpoint



api.spacexdata.com/v4/launches/past

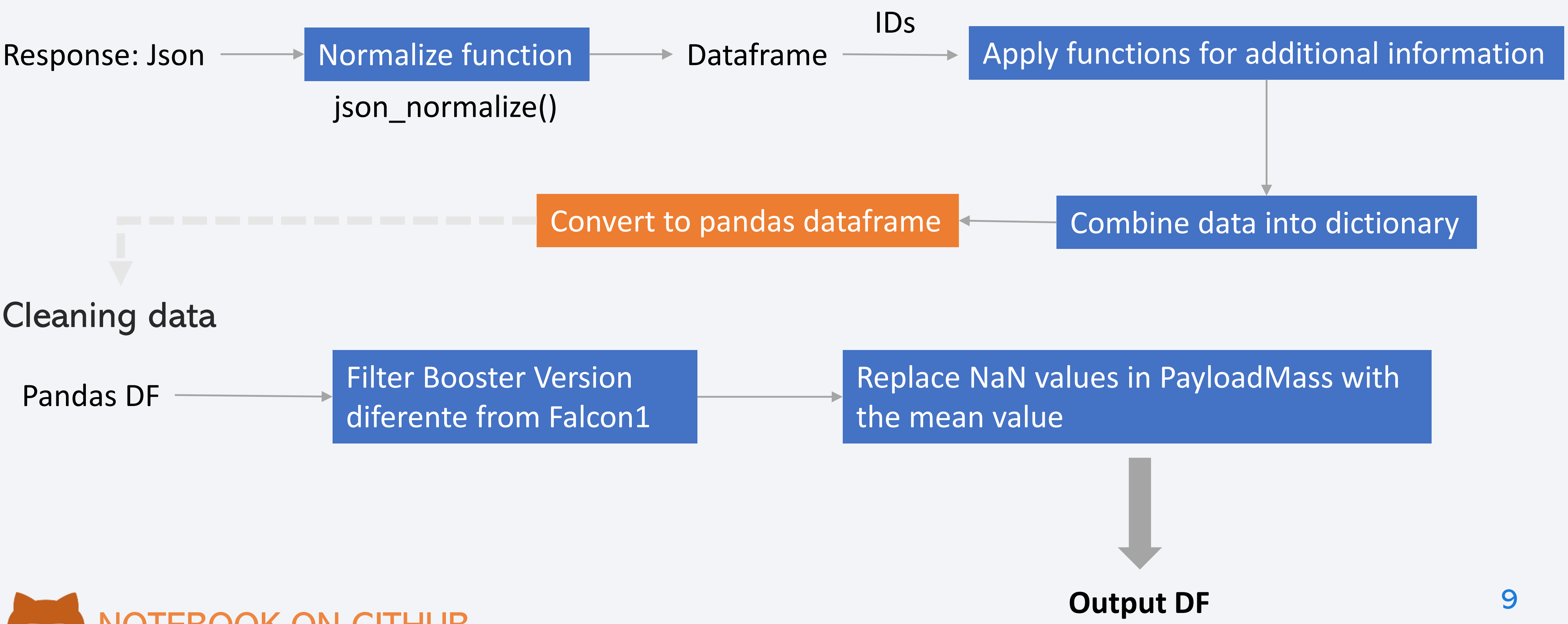
Additional information

Endpoint	Function
https://api.spacexdata.com/v4/rockets/	getBoosterVersion
https://api.spacexdata.com/v4/launchpads/	getLaunchSite
https://api.spacexdata.com/v4/payloads/	getPayloadData
https://api.spacexdata.com/v4/cores/	getCoreData

NOTEBOOK ON GITHUB

Data Collection – SpaceX API

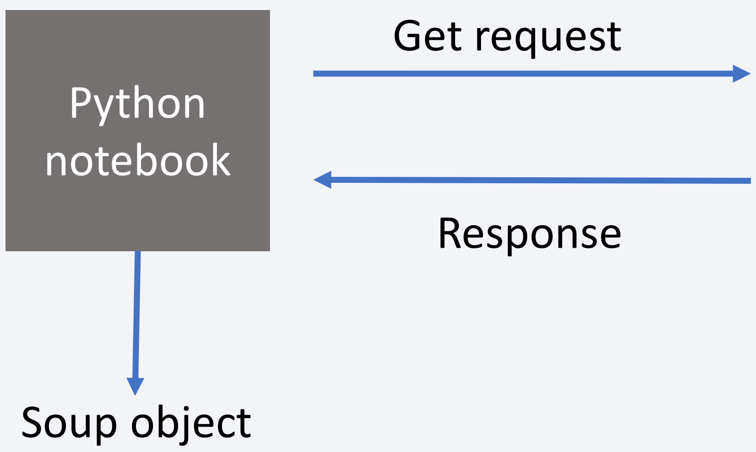
Process to obtain a Pandas dataframe



NOTEBOOK ON GITHUB

Data Collection - Scraping

Data collection scheme by web scraping Wikipedia pages



Past launches

2010 to 2013

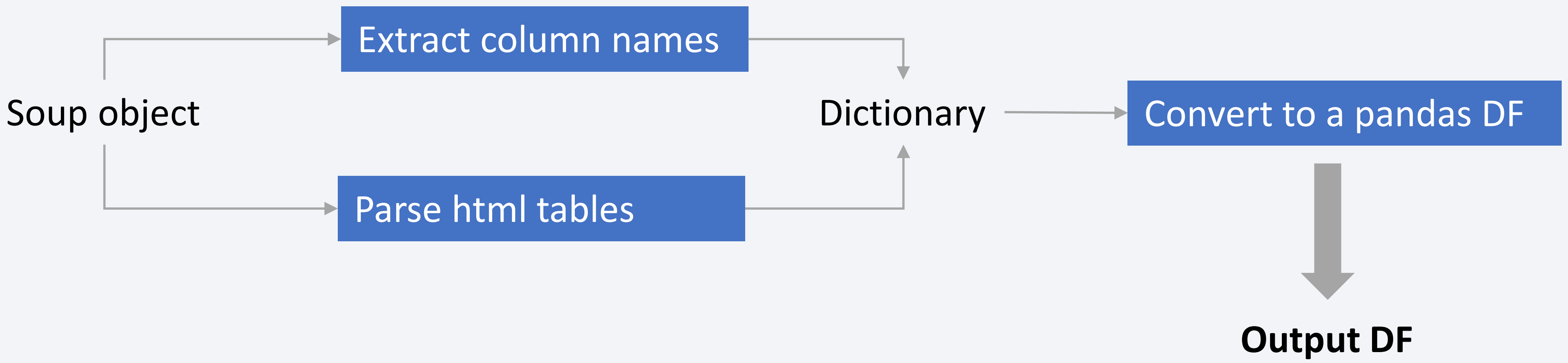
[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								
4	8 October 2012, 00:35[21]	F9 v1.0[7] B0006.1[8]	CCAFS, SLC-40	SpaceX CRS-1[22] (Dragon C103)	4,700 kg (10,400 lb)	LEO (ISS)	NASA (CRS)	Success	No attempt
				Orbcomm-OG2[23]	172 kg (379 lb)[24]	LEO	Orbcomm	Partial failure[25]	
	CRS-1 was successful, but the secondary payload was inserted into an abnormally low orbit and subsequently lost. This was due to one of the nine Merlin engines shutting down during the launch, and NASA declining a second reignition, as per ISS visiting vehicle safety rules, the primary payload owner is contractually allowed to decline a second reignition. NASA stated that this was because SpaceX could not guarantee a high enough likelihood of the second stage completing the second burn successfully which was required to avoid any risk of secondary payload's collision with the ISS.[26][27][28]								

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

NOTEBOOK ON GITHUB

Data Collection - Scraping

Obtaining data in a pandas dataframe



NOTEBOOK ON GITHUB

Data Wrangling

In data wrangling an Exploratory data analysis was performed.

Also, a interpretation of the outcomes was made, creating a column where succeful landings were interpreted as 1 and failed landings as 0, to be used in further analysis.

NOTEBOOK ON GITHUB

Data Wrangling

Launches of each site

Dataframe Launch Sites Column → Apply value_counts() → Number of Launches for each site

Launches of each orbit type

Dataframe Orbit column → Apply value_counts() → Number of Laucnhes for each orbit

Type of outcomes

Dataframe Outcomes column → Apply value_counts → Number of Launches for each outcome → Save in variable landing_outcomes

Create list with bad outcomes → Iterate the values and append to list if it exists in the bad outcomes list → DF column class

NOTEBOOK ON GITHUB

EDA with Data Visualization

Scatter plots

In the EDA with data visualization the preferred chart type was the scatter plot (since it is great to identify patterns and relationships) for the following analysis:

Flight Number x Launch Site

Payload mass x Launch Site

Flight Number x Orbit type

Payload mass x Orbit type

Bar plot

A bar plot was used to analyze success rate and orbit type, since it is a great approach to compare categorical variables

Line plot

In the end, a line plot (ideal to plot a variable behavior on a time line) was used to analyze the success rate over the years.

NOTEBOOK ON GITHUB

EDA with SQL

To EDA with SQL the following queries were performed:

- Unique launch sites names;
- 5 records with “CCA” in the launch site name;
- Total payload mass carried by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;
- Total number of successful and failures on mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
- 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.

NOTEBOOK ON GITHUB

Build an Interactive Map with Folium

In the map visualizations, a mark and a circle were created for every launch site using folium. To add that to the map, the method add_child was used. Also the outcome of each launch was added to its location on the map using green (successful) and red(failed) markers.

After that, a MousePosition was added as well, so the latitude and longitude could be seen. A function was given to calculate the distance between the launch site and a chosen position on the map. In the end, a line was added to mark this calculated distance from the site and the coast, a railway, a highway and a city.

NOTEBOOK ON GITHUB

Build a Dashboard with Plotly Dash

A pie chart was added to visualize the amount of launchings of each site. A dropdown list was also added to display the success and failure rate of each location.

A scatter plot was added at the bottom containing payload mass in the x axis and outcome rate in the y axis. The legend divided the points accordingly to the booster version. An iteration with a payload range selector was also added. This view made it possible to analyze which booster version had the best outcome and which payload mass was loaded.

NOTEBOOK ON GITHUB

Predictive Analysis (Classification)

Data visualization lab

Create dummy variables

Predictive Analysis

Cleaned DF

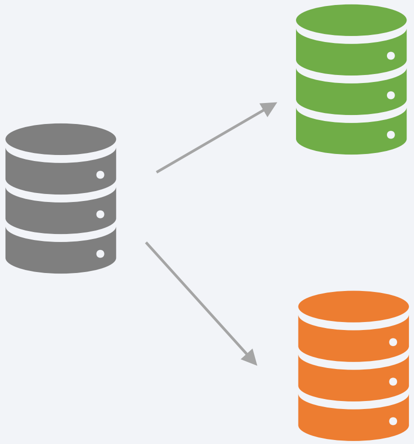
Creater Target variable array Y (class) using numpy

Create X array and standerdize it

Train/test split

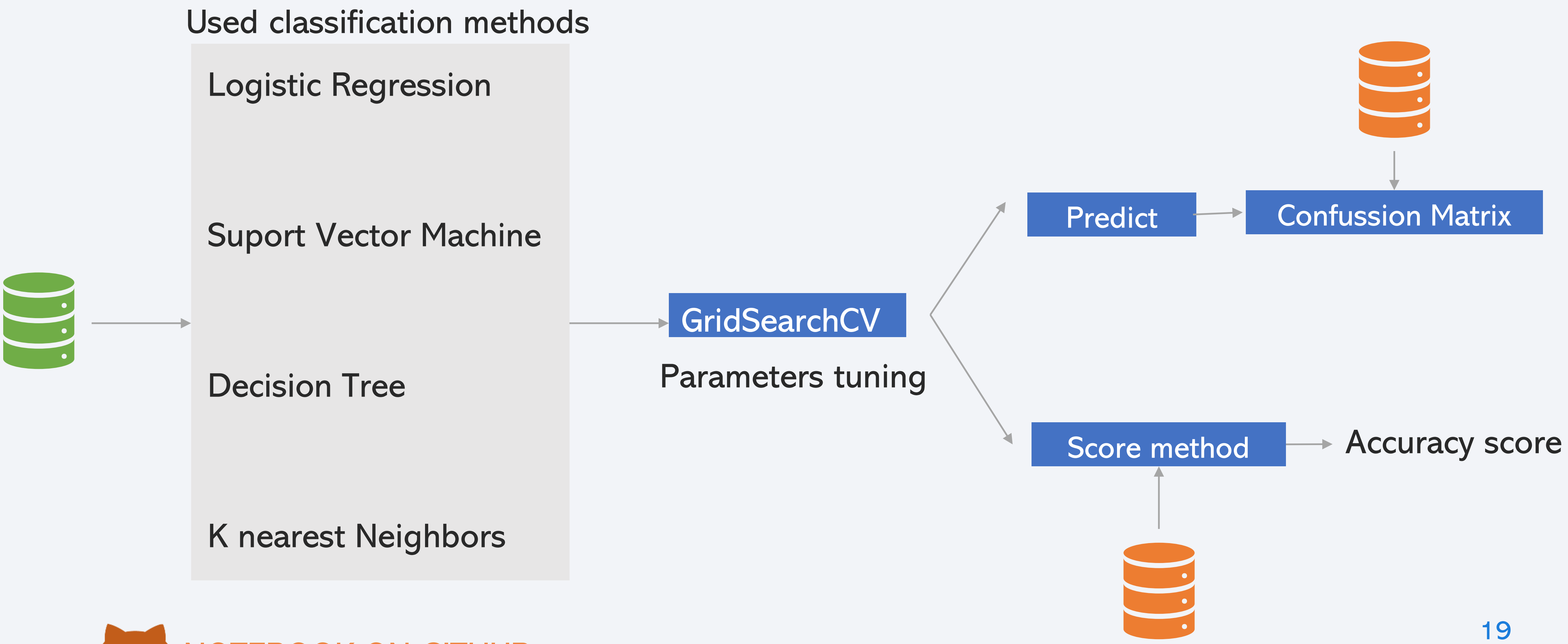
test_size=0.2

Random_state=2



NOTEBOOK ON GITHUB

Predictive Analysis (Classification)



Results

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

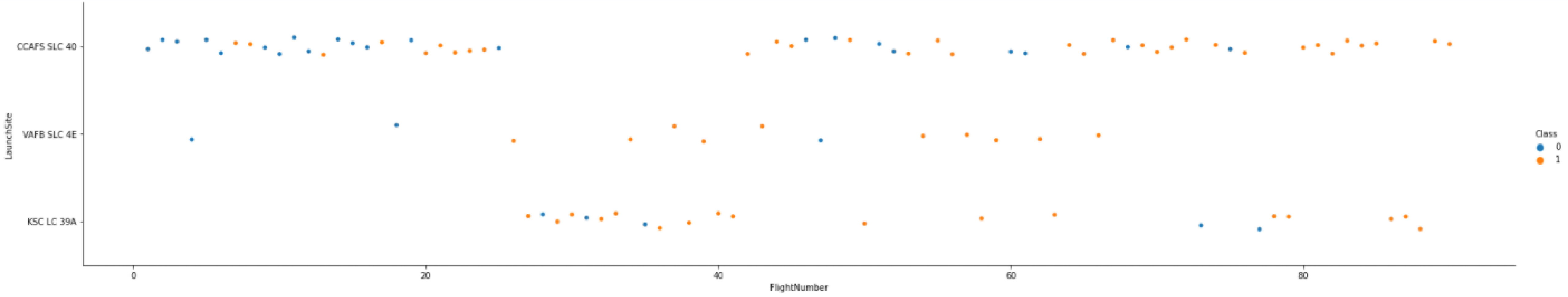
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Launch sites have a different amount of launch events and different success rates.

The first flight numbers had a lower success rate compared to the last ones.

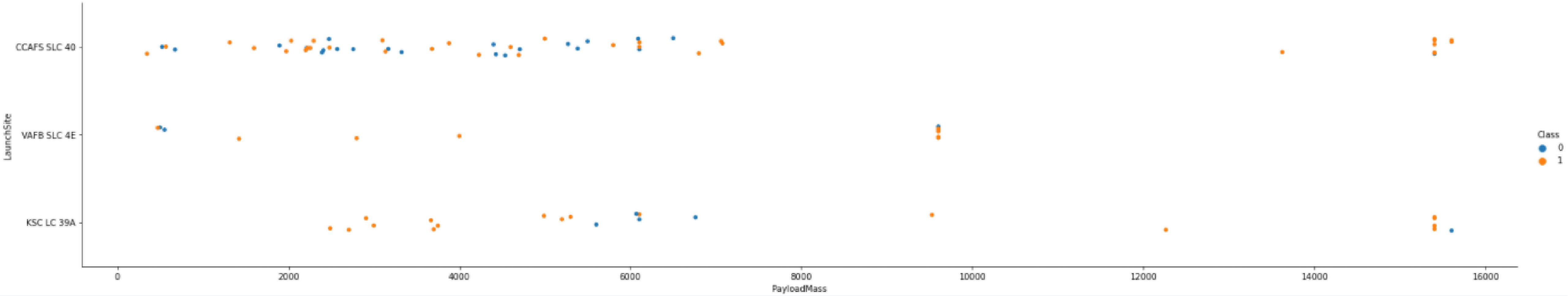


Payload vs. Launch Site

The payload for the most launch events are between 500 and 7000.

There isn't launches with more then 10000 of payload for VAFB SLC 4E site.

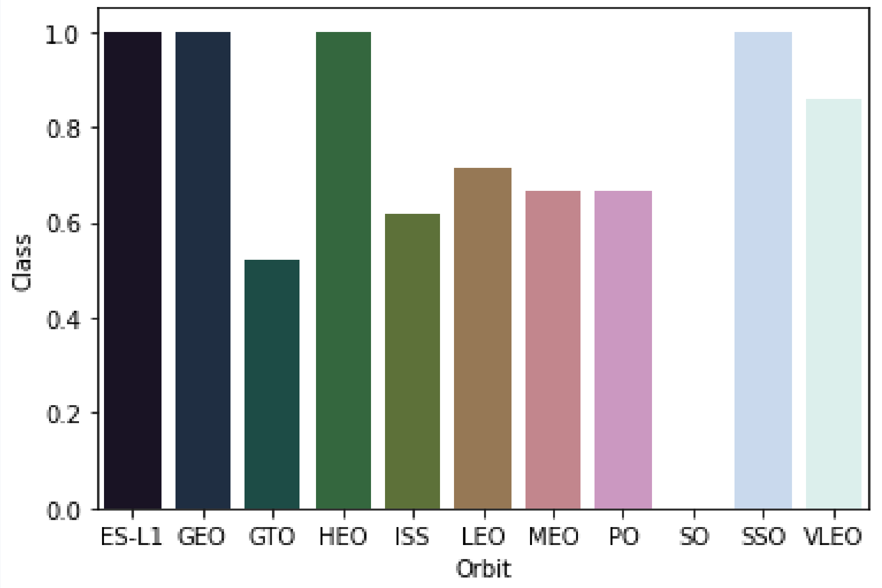
High payload seems to have less failure events.



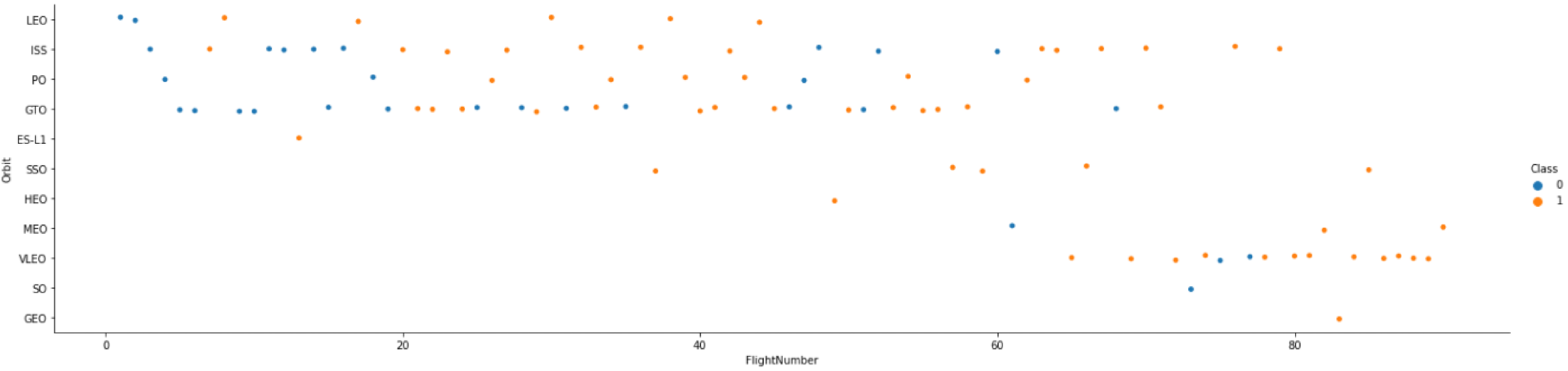
Success Rate vs. Orbit Type

The most successful rates are for the orbits ES-L1, GEO, HEO and SSO.

They are followed by VLEO orbit, with 0.8 rate.



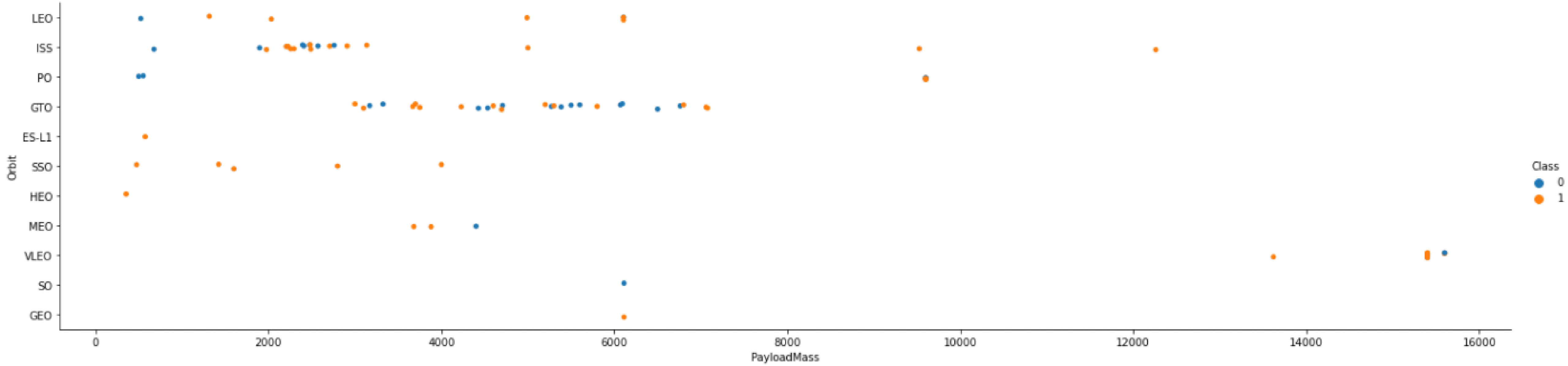
Flight Number vs. Orbit Type



The orbits ES-L1, GEO and HEO have high success rates, but a very low sampling. The SSO orbit has a slightly larger sample and very high successful rates.

The VLEO with 0.8 success rate has a significant sample size, so its probably the most reliable one.

Payload vs. Orbit Type



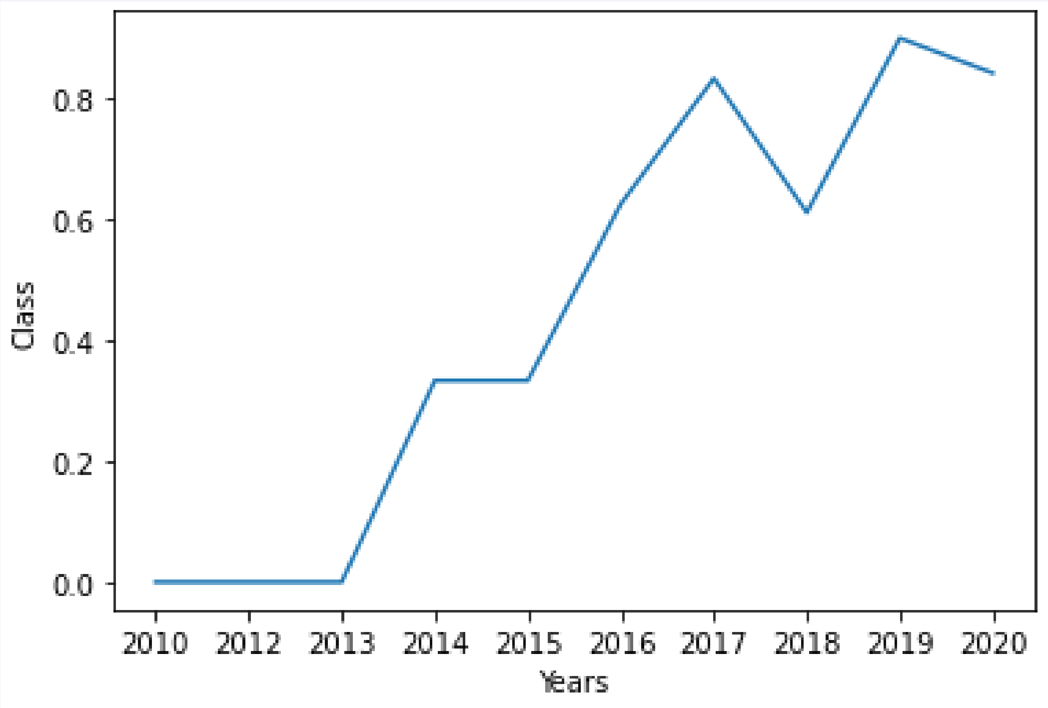
The payload for most orbit types is around 500 and 7000.

The VLEO is the only one with launches with a payload higher then 13000.

The GTO orbit has payload mass around 1000 and 7000, while ISS has payload mass around 2000 and 3500.

Launch Success Yearly Trend

The success rate has grown over the years, which makes sense, as the project has probably been improved by the performance from the launching events.



All Launch Site Names

There's 4 unique launch sites

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

Launch Site Names Begin with 'CCA'

The Launch site which begins with the word CCA is CCAFS LC-40.

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	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total payload carried by boosters from NASA

Out[23]:

1
45596

Average Payload Mass by F9 v1.1

The average payload mass by F9 v1.1

avg(payload_mass__kg_)
2928.4

This is a small amount of mass compared to others

First Successful Ground Landing Date

The first successful landing was in 2017 in the KSC LC-39A site with 5300 of pay load mass.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

The drone ship successful landing happened in 2016, before ground landing.

After that, it happened again in 2017 carrying a larger amount of payload mass

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
11-10-2017	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

There's a total of 100 positive mission outcomes and 1 failure. So the mission outcome does not affect on the reuse of the first stage.

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

12 booster versions carried the maximum payload mass

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

The failed lading outcomes in drone ship in 2015 used different booster versions but happened in the same launch site.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
10-01-2015	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
14-04-2015	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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

Between 2010-06-04 and 2017-03-20 there were 38 success outcomes, and considering drone ship and ground pad there were 53 successfull landing outcomes.

Landing_Outcome	count(*)
Success	38
No attempt	11
Success (drone ship)	9
Success (ground pad)	6
Failure	3
Controlled (ocean)	2
No attempt	1

Additional analysis: Max Payload

Analysing the launches after 2018, the combination of launch sites and orbit and what was the maximum payload mass.

Launch_Site	Orbit	max(PAYLOAD_MASS__KG_)
CCAFS SLC-40	LEO	15600
KSC LC-39A	LEO	15600
KSC LC-39A	LEO (ISS)	12530
VAFB SLC-4E	Polar LEO	9600
CCAFS SLC-40	GTO	7075
KSC LC-39A	GTO	5300
CCAFS SLC-40	MEO	4311
VAFB SLC-4E	SSO	4200
CCAFS SLC-40	SSO	3130
CCAFS SLC-40	LEO (ISS)	2617
VAFB SLC-4E	LEO	1192
CCAFS SLC-40	HEO	362

Additional analysis: Outcome and hour

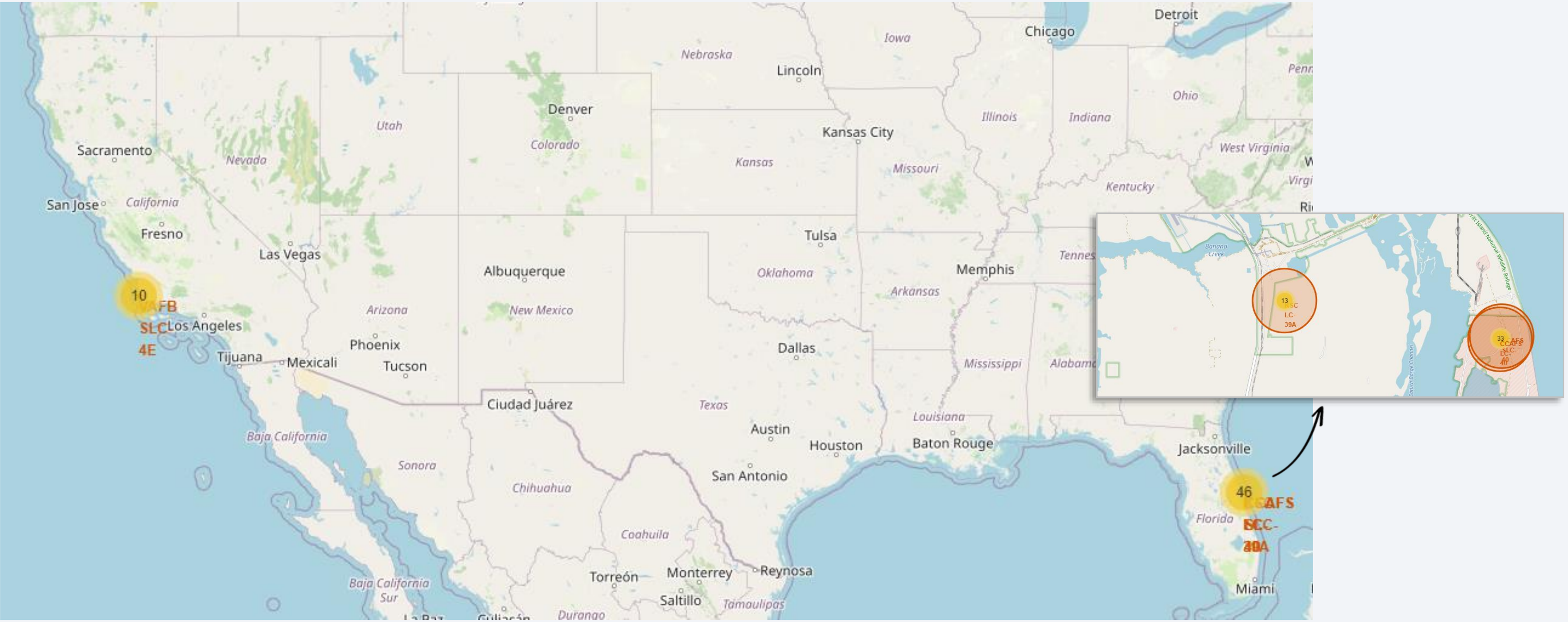
The hour of the launching doesnt seem to affect it, once there's similar number of positive outcomes in different times of the day

HOUR	SUM(OUTCOME)
14	6
20	5
05	5
22	4
19	4
02	4
01	4
21	3
17	3
15	3
12	3

Section 3

Launch Sites Proximities Analysis

All launch sites on map



The launch sites area all very close to the coast.
There’s more launch events in the east coast.

Outcomes of each site

East cost

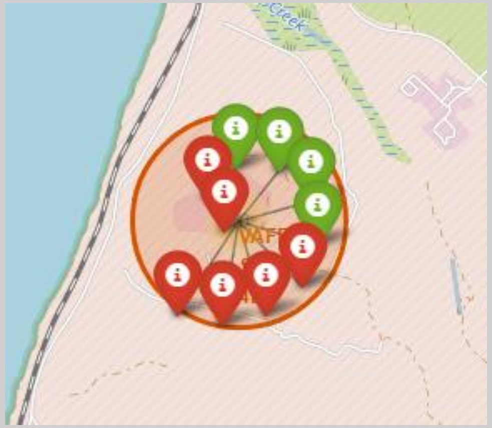
West cost

CCAFS SLC-40

CCAFS LC-40

KSC LC-39A

VAFB SLC-4E



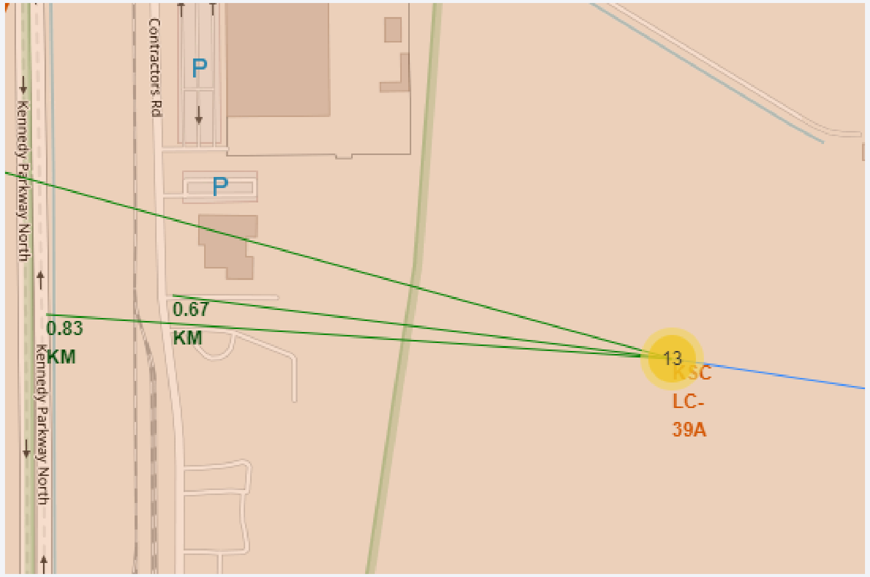
Its possible to compare the surroundings of each site and maybe suggest some hypothesis. There's more successful launches in the east coast.

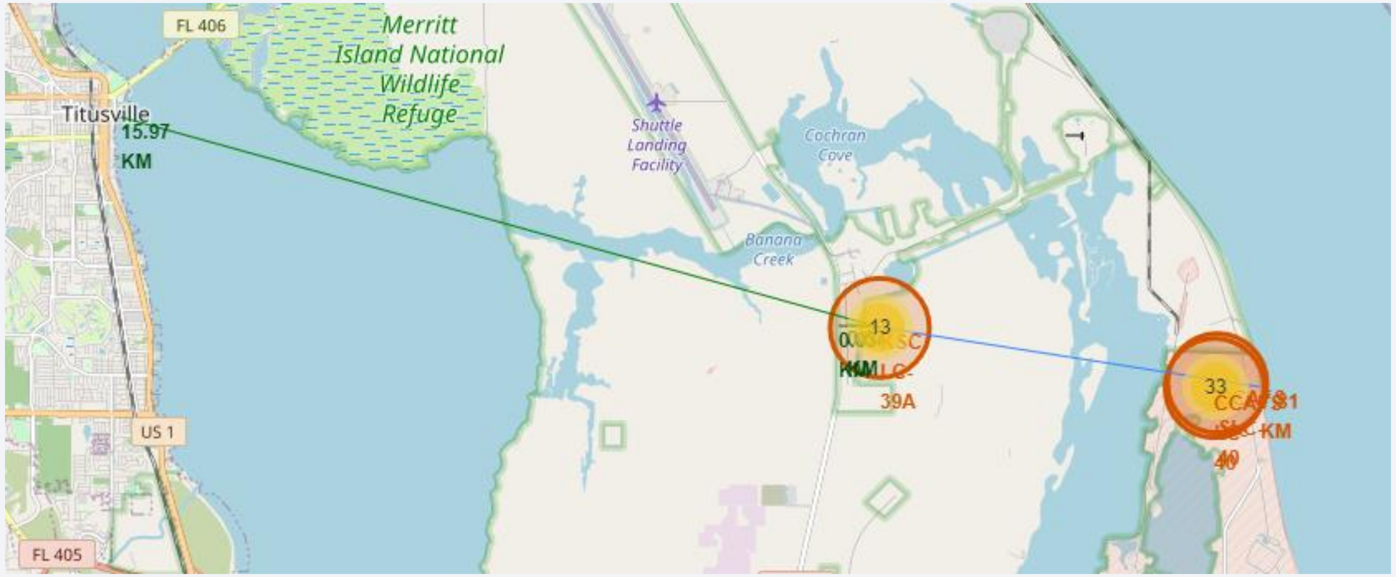
The given dataset for this lab had a different number of launches.

"The following dataset with the name `spacex_launch_geo.csv` is an augmented dataset with latitude and longitude added for each site."

Surroundings of KSC LC-39A

The chosen site was KSC LC-39A



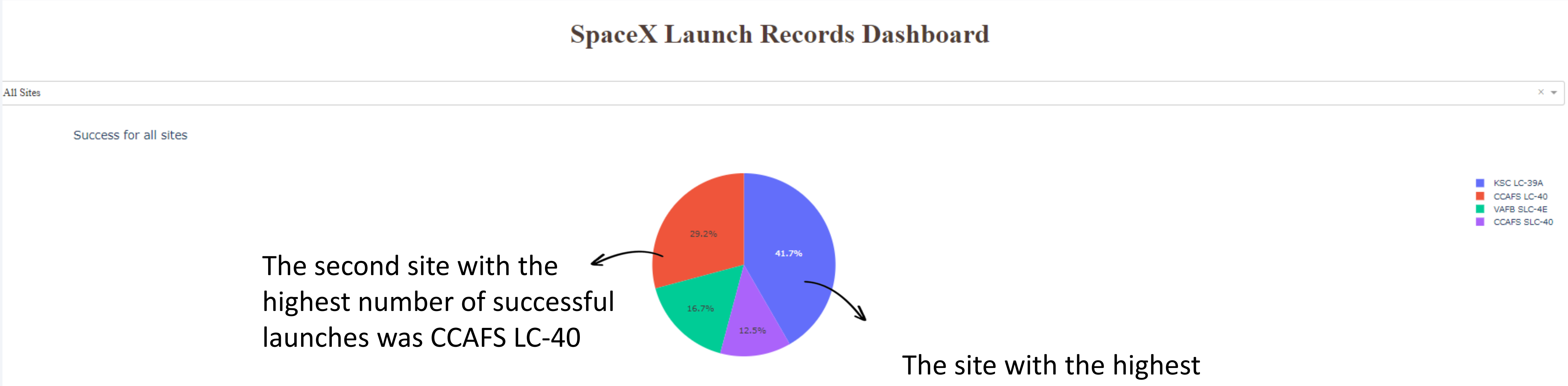


The chosen site is very close to a highway and a railway, and relatively close to a city (Titusville) with a distance of around 15.97 km.

Section 4

Build a Dashboard with Plotly Dash

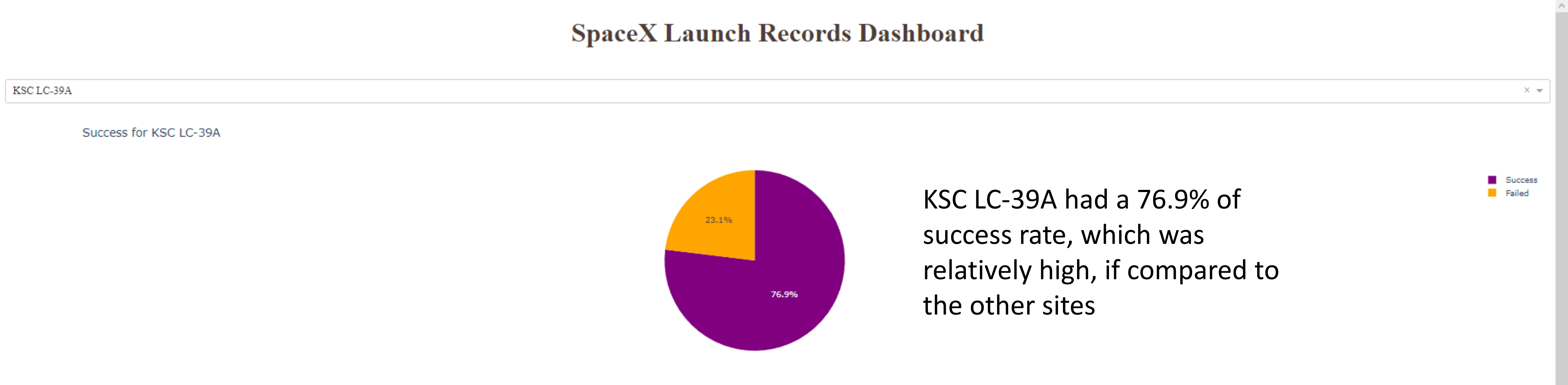
Success rate of each site



The second site with the highest number of successful launches was CCAFS LC-40

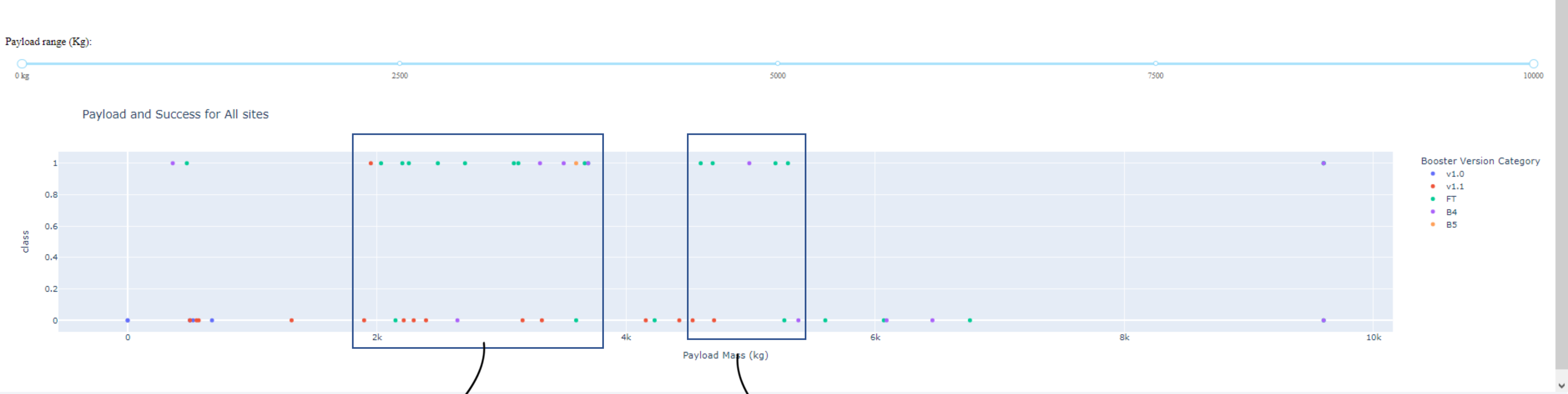
The site with the highest number of successful launches was KSC LC-39A

Success/Failure rate for KSC LC-39A



KSC LC-39A had a 76.9% of success rate, which was relatively high, if compared to the other sites

Payload vs. Launch Outcome scatter plot for all sites

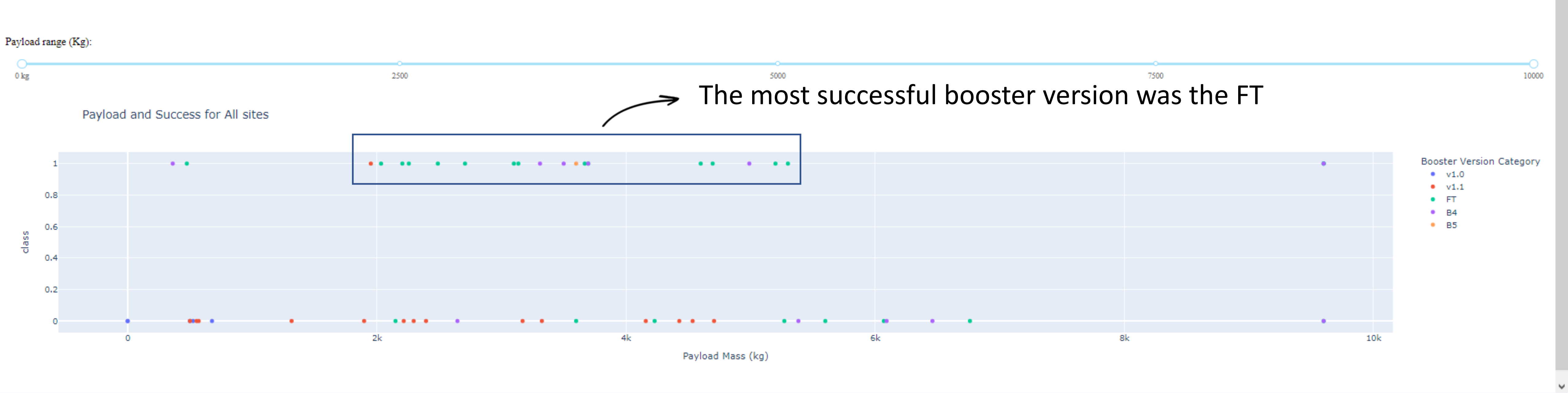


The range between around 1900kg and 3900kg have around 60% of success rate.

The range between around 4500kg and 5500kg have around 55% of success rate.

It's possible to identify in this visualization that range between 2000kg and 5500kg is were lies the highest amount of successful events. Especially between 1900kg and 3900kg

Payload vs. Launch Outcome scatter plot for all sites



Section 5

Predictive Analysis (Classification)

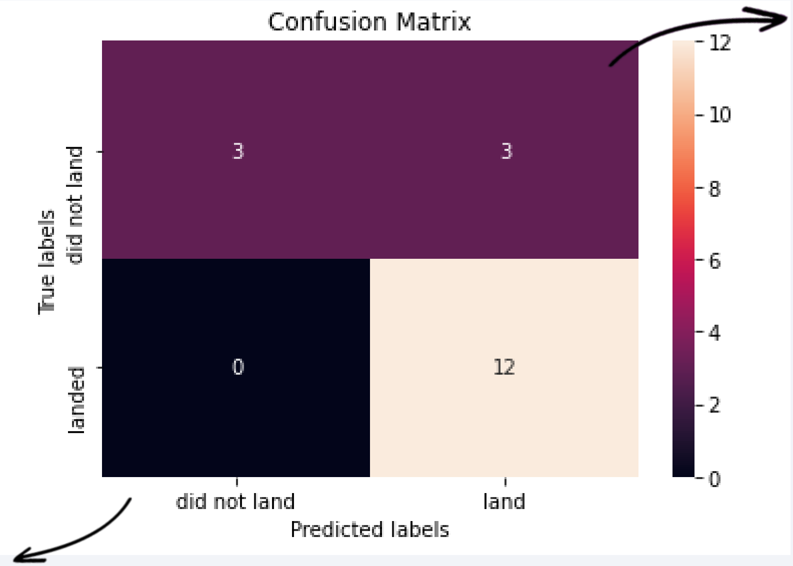
Classification Accuracy

All models reached the same accuracy, they all performed the same.

Model	Accuracy
Logistic Regression	0.8333333333333334
Support Vector Machine	0.8333333333333334
Decision Tree	0.8333333333333334
K Nearest Neighbors	0.8333333333333334

Confusion Matrix

Since all the models performed the same, they all have the following confusion matrix:



3 failed events were predicted as successfull ones

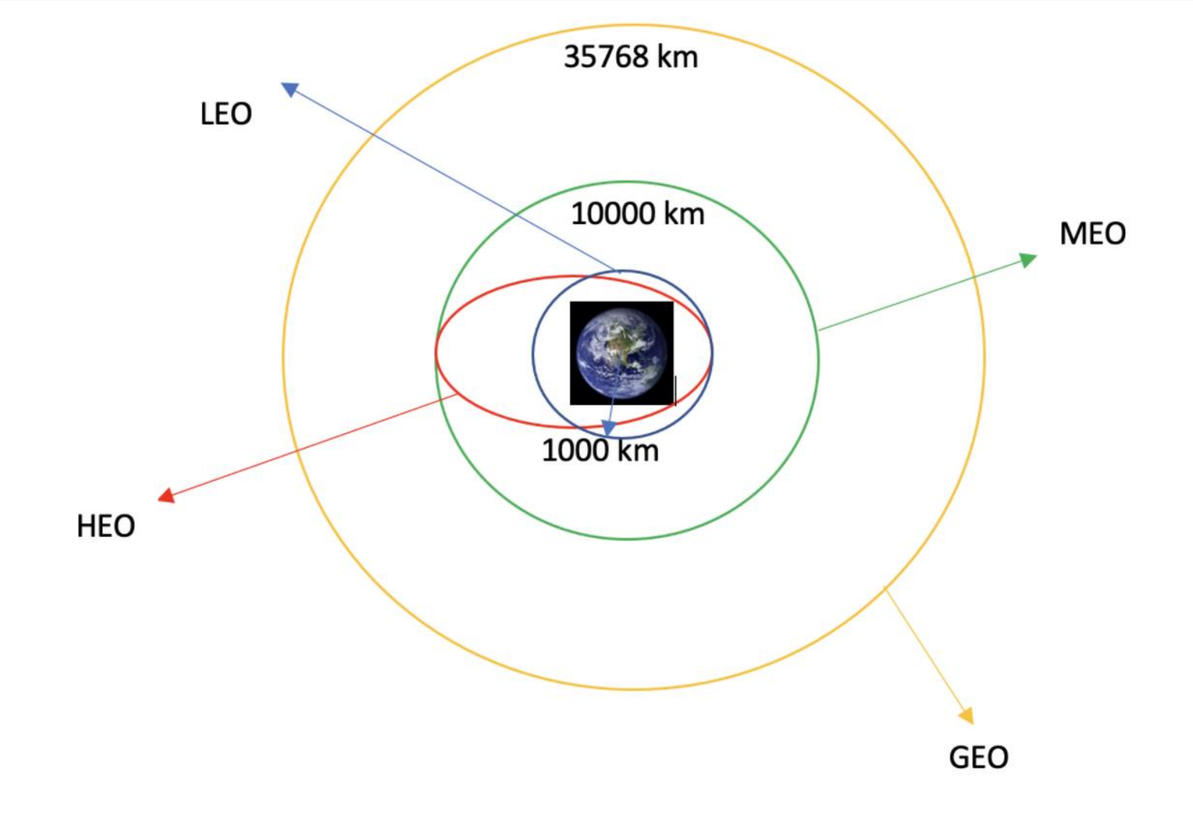
There were no “landed” events predicted as “did not landed”.

Conclusions

- Successful launches are increasing over time as the project receives improvement based on lessons learned from previous launches.
- Considering the amount of launch events of each orbit, VLEO and LEO seems to be a good orbit choice.
- The most successful booster was the FT, but it seems that it can't have a pay load mass over 5300 kg.
- It is possible to have high pay load mass depending on the parameters.
- KSC LC-39A is the most successful launch site.
- The methods have good performance on predicting positive outcomes.
- Even though the model accuracy was high, it is not good in predicting negative outcomes, since from 6 events 3 were predicted incorrectly.

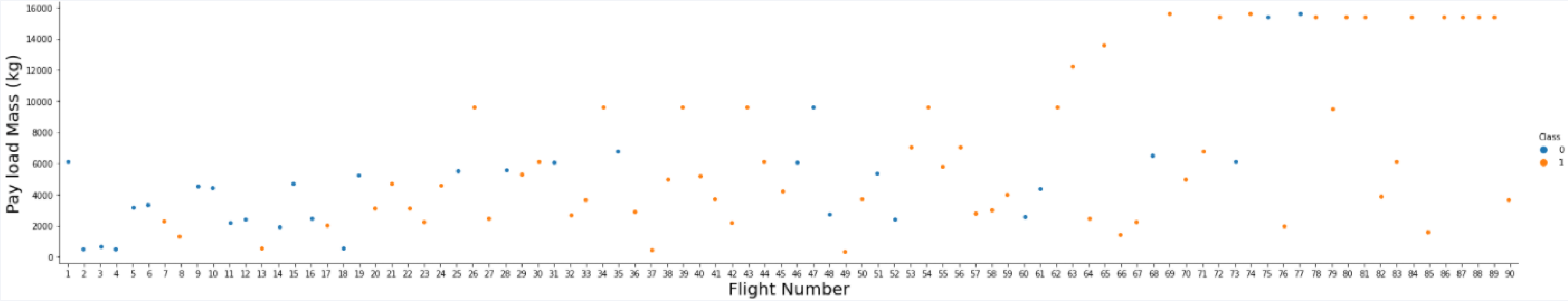
Appendix

- Orbits illustration



Appendix

- VLEO description: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation [2].
- Payload mass and FlighNumber



Appendix

- SQL query for the maximum payload analysis

```
%%sql
select Launch_Site, Orbit, max(PAYLOAD_MASS__KG_)  from SPACEXTBL
where 1=1
and cast(substr(DATE, 7,4) as text)  >= '2018'
and upper(Landing_Outcome) like '%SUCCESS%'
group by Launch_Site, Orbit
order by 3 desc
```

Appendix

- SQL query for the hour and outcome analysis

```
%%sql
SELECT HOUR, SUM(OUTCOME)
FROM(
SELECT
SUBSTR(TIME_UTC,1,2) AS HOUR
, CASE WHEN upper(Landing_Outcome) LIKE '%SUCCESS%' THEN 1
    ELSE 0
    END OUTCOME
FROM SPACEXTBL)
GROUP BY HOUR
ORDER BY 2 DESC
```

Appendix

- Grid Search CV parameters

Logistic Regression
tuned hpyerparameters :(best parameters) {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713

Support Vector Machine
tuned hpyerparameters :(best parameters) {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy : 0.8482142857142856

Decision Tree
tuned hpyerparameters :(best parameters) {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'random'}
accuracy : 0.8767857142857143

K nearest neighbors
tuned hpyerparameters :(best parameters) {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858

Appendix

- Repository URL

Thank you!