IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

LAGIER Raphaël
22/10/2023

Outline

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

Executive Summary

- We were asked to predict the rate of successful landing of SpaceX Falcon 9 . Therefore, we analyzed the impact of several variable on the results such as Payload Mass, Launch site, Date, Booster Version...
- To achieve our goal, we used Python and its libraries for :
    - Data collection
    - Data wrangling and formatting
    - Data visualization
    - Machine Learning prediction
- All our analyses an researches lead to a prediction of succefull landing rate of 83.33%.

Table of Contents

Table of Contents

- Summary of all results

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. 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. In this module, you will be provided with an overview of the problem and the tools you need to complete the course.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X API (using requests)
  - Web Scrapping (using requests & BeautifulSoup)
- Perform data wrangling (with pandas & numpy)
  - Work on values quality : Identify missing values and type of values
  - Analyze value: number of launch per site, per orbit, successful launch per orbit

Methodology

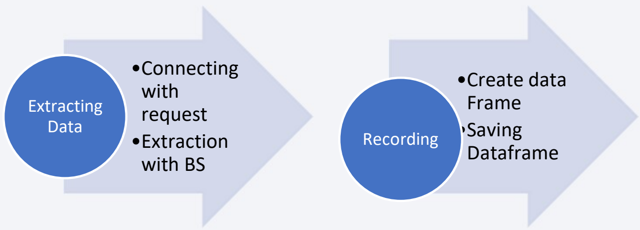
Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize data with StandardScaler
  - Split data for training (80%) and testing(20%)
  - Set and evaluate 4 classification methods : SVM, Logistic regression, KNN, Decision tree
  - Using predict, confusion matrix, score.

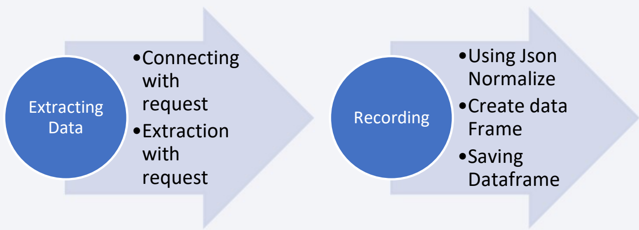
Data Collection

- Data were collected by 2 ways:
  - Using requests on SPACE X API [Space X api link]
  - Using requests and BeautifulSoup (BS) on Wikipedia site. [Wikipedia Link]

Web scraping with BeautifulSoup



API Data collection with Request



Data Collection – SpaceX API

- Create a connection to Space X API to withdraw data required to perform a data analyse.
- The Data Frame contains 17 features and 90 records after removing Falcon 1 launch.
- We decide to keep empty Landing pad with 26 null values to count launch with no landing.
- We replaced 5 missing value for payload mass with the means of payload mass recorded.

FlightNumber	Date	BoosterVersi	PayloadMas	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCour	Serial	Longitude	Latitude
6	04/06/2010	Falcon 9	6123.547647	LEO	CCSFS SLC 4	None None	1	False	False	False		1.0	0	B0003	-80.577366	28.5618571
8	22/05/2012	Falcon 9	525.0	LEO	CCSFS SLC 4	None None	1	False	False	False		1.0	0	B0005	-80.577366	28.5618571
10	01/03/2013	Falcon 9	677.0	ISS	CCSFS SLC 4	None None	1	False	False	False		1.0	0	B0007	-80.577366	28.5618571
11	29/09/2013	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False		1.0	0	B1003	-120.610829	34.632093
12	03/12/2013	Falcon 9	3170.0	GTO	CCSFS SLC 4	None None	1	False	False	False		1.0	0	B1004	-80.577366	28.5618571
13	06/01/2014	Falcon 9	3325.0	GTO	CCSFS SLC 4	None None	1	False	False	False		1.0	0	B1005	-80.577366	28.5618571
14	18/04/2014	Falcon 9	2296.0	ISS	CCSFS SLC 4	True Ocean	1	False	False	True		1.0	0	B1006	-80.577366	28.5618571
15	14/07/2014	Falcon 9	1316.0	LEO	CCSFS SLC 4	True Ocean	1	False	False	True		1.0	0	B1007	-80.577366	28.5618571
16	05/08/2014	Falcon 9	4535.0	GTO	CCSFS SLC 4	None None	1	False	False	False		1.0	0	B1008	-80.577366	28.5618571
17	07/09/2014	Falcon 9	4428.0	GTO	CCSFS SLC 4	None None	1	False	False	False		1.0	0	B1011	-80.577366	28.5618571
18	21/09/2014	Falcon 9	2216.0	ISS	CCSFS SLC 4	False Ocean	1	False	False	False		1.0	0	B1010	-80.577366	28.5618571
19	10/01/2015	Falcon 9	2395.0	ISS	CCSFS SLC 4	False ASDS	1	True	False	True	5e9e3032383	1.0	0	B1012	-80.577366	28.5618571

- Ghithub link to Api Collection Notebook

Create functions

Create get request

Normalize data

Create a dataframe with pertinent data

Data Collection - Scraping

- Create a request to https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- The Data Frame contains 11 features and 121 records.
- The data Booster Version is not very relevant due to the lack of 116 records.
- GitHub Webscraping notebook

Create functions

Create get request

Extr act the data from URL with BeautifSoup

Create a dataframe with pertinent data

Data Wrangling

- We analyze the data to determine the number of launch per orbit and the successful landing rate
- We had to affect 0 or 1 to bad and successful landing to calculate the average.
- 66,6% of landing were succeeded

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
1	04/06/2010	Falcon 9	6104.959411764706	LEO	CCAFS SLC 40	None None	1	False	False	False		1.0	0	B0003	-80.577366	28.5618571	0
2	22/05/2012	Falcon 9	525.0	LEO	CCAFS SLC 40	None None	1	False	False	False		1.0	0	B0005	-80.577366	28.5618571	0
3	01/03/2013	Falcon 9	677.0	ISS	CCAFS SLC 40	None None	1	False	False	False		1.0	0	B0007	-80.577366	28.5618571	0
4	29/09/2013	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False		1.0	0	B1003	-120.610829	34.632093	0
5	03/12/2013	Falcon 9	3170.0	GTO	CCAFS SLC 40	None None	1	False	False	False		1.0	0	B1004	-80.577366	28.5618571	0
6	06/01/2014	Falcon 9	3325.0	GTO	CCAFS SLC 40	None None	1	False	False	False		1.0	0	B1005	-80.577366	28.5618571	0
7	18/04/2014	Falcon 9	2296.0	ISS	CCAFS SLC 40	True Ocean	1	False	False	True		1.0	0	B1006	-80.577366	28.5618571	1
8	14/07/2014	Falcon 9	1316.0	LEO	CCAFS SLC 40	True Ocean	1	False	False	True		1.0	0	B1007	-80.577366	28.5618571	1
9	05/08/2014	Falcon 9	4535.0	GTO	CCAFS SLC 40	None None	1	False	False	False		1.0	0	B1008	-80.577366	28.5618571	0
10	07/09/2014	Falcon 9	4428.0	GTO	CCAFS SLC 40	None None	1	False	False	False		1.0	0	B1011	-80.577366	28.5618571	0
11	21/09/2014	Falcon 9	2216.0	ISS	CCAFS SLC 40	False Ocean	1	False	False	False		1.0	0	B1010	-80.577366	28.5618571	0
12	10/01/2015	Falcon 9	2395.0	ISS	CCAFS SLC 40	False ASDS	1	True	False	True	5e9e3032383ecb761634e7cb	1.0	0	B1012	-80.577366	28.5618571	0
13	11/02/2015	Falcon 9	570.0	ES-L1	CCAFS SLC 40	True Ocean	1	True	False	True		1.0	0	B1013	-80.577366	28.5618571	1
14	14/04/2015	Falcon 9	1898.0	ISS	CCAFS SLC 40	False ASDS	1	True	False	True	5e9e3032383ecb761634e7cb	1.0	0	B1015	-80.577366	28.5618571	0
15	27/04/2015	Falcon 9	4707.0	GTO	CCAFS SLC 40	None None	1	False	False	False		1.0	0	B1016	-80.577366	28.5618571	0
16	28/06/2015	Falcon 9	2477.0	ISS	CCAFS SLC 40	None ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	1.0	0	B1018	-80.577366	28.5618571	0
17	22/12/2015	Falcon 9	2034.0	LEO	CCAFS SLC 40	True RTLS	1	True	False	True	5e9e3032383ecb267a34e7c7	1.0	0	B1019	-80.577366	28.5618571	1
18	17/01/2016	Falcon 9	553.0	PO	VAFB SLC 4E	False ASDS	1	True	False	True	5e9e3033383ecbb9e534e7cc	1.0	0	B1017	-120.610829	34.632093	0
19	04/03/2016	Falcon 9	5271.0	GTO	CCAFS SLC 40	False ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	1.0	0	B1020	-80.577366	28.5618571	0
20	08/04/2016	Falcon 9	3136.0	ISS	CCAFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	2.0	1	B1021	-80.577366	28.5618571	1
21	06/05/2016	Falcon 9	4696.0	GTO	CCAFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	2.0	0	B1022	-80.577366	28.5618571	1
22	27/05/2016	Falcon 9	3100.0	GTO	CCAFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	2.0	1	B1023	-80.577366	28.5618571	1
23	18/07/2016	Falcon 9	2257.0	ISS	CCAFS SLC 40	True RTLS	1	True	False	True	5e9e3032383ecb267a34e7c7	2.0	1	B1025	-80.577366	28.5618571	1
24	14/08/2016	Falcon 9	4600.0	GTO	CCAFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	2.0	0	B1026	-80.577366	28.5618571	1
25	01/09/2016	Falcon 9	5500.0	GTO	CCAFS SLC 40	None ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	3.0	0	B1028	-80.577366	28.5618571	0
26	14/01/2017	Falcon 9	9600.0	PO	VAFB SLC 4E	True ASDS	1	True	False	True	5e9e3033383ecbb9e534e7cc	3.0	1	B1029	-120.610829	34.632093	1
27	19/02/2017	Falcon 9	2490.0	ISS	KSC LC 39A	True RTLS	1	True	False	True	5e9e3032383ecb267a34e7c7	3.0	1	B1031	-80.6039558	28.6080585	1

- [GitHub Data Wrangling notebook]

- Calculated the % of messing value for each attribute
- Control the data type of each attribute
- Determine the landing for each orbit
- Create a list of bad outcome name
- Create a column for bad outcome class (bad=0, Good=1) to calculate the average of success

EDA with Data Visualization

- We used scatter plot to visualize the relation ship between:
  - Flight number vs payload mass : as the flight number increase, the first stage has more chance to land. Moreover, we saw that more massive load has less chance to return.
  - Flight Number and launch site: we saw that , for each launch site, the success increase with flight number.
  - Flight Number and orbit: We saw that orbit has 100% of success.
  - Payload and orbit : analyze the relationship between payload and orbit.
- We used bar chart to visualize the relation ship between:
  - Payload and Launch site: VAFB SLC did not launch heavy payload.
- We used plot line to visualize the launch success yearly trend.
- Github DataViz notebook.

EDA with SQL

- We used several SQL queries to determine:
  - Unique launch site names.
  - Total payload mass carried by boosters launched by NASA .
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was acheived.
  - 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 failure 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

  - Github SQL Notebook

Build an Interactive Map with Folium

- We made a map with two launch sites and Nasa location.
- We create marker with number of launch
- We create specific icon for succeed launch (green) and failed launch (red)
- We calculate distance between launch site and facilities and mark them on the map

➔ Our goals was to have a good visualization of the location with their activity and we wanted to know their access to infrastructures.

- GitHub Visualization link

Build a Dashboard with Plotly Dash

- We made two graph:
  - A pie chart which show success rate
  - A scatter plot which show outcome per payload mass
- Each graph allow us to focus on one site. Thanks to these graph we can first have a general overview of the situation and then focus on the launch site we want to analyze.

- Github link to Dashboard

Predictive Analysis (Classification)

- We load dependent variable in a Numpy array.
- We transform the independent variables to obtain standardize result for each datas. Our goal is to give the same scale for each data.
- We split our datas in 2 groups : Training datas (80%) and test datas (20%)
- We decide to test 4 machine learning Algorithm:
  - Support Vector Machine
  - Logisitc regression
  - K Nearest Neighbours
  - Decision Tree
- We apply the methodology below for each model.

Predictive Analysis (Classification)

Parameters	•Create a dictonnary with parameters
Model	•Create the model with the parameters
GridSearchCV	•Create a grideSearch Cv on models •Fit it on train data set •Find the bests parameters •Calculate the grid best score
Accuracy	•Calcul the accuracy result with test data set •Create the confusion matrix

GitHub Machine Learning Notebook

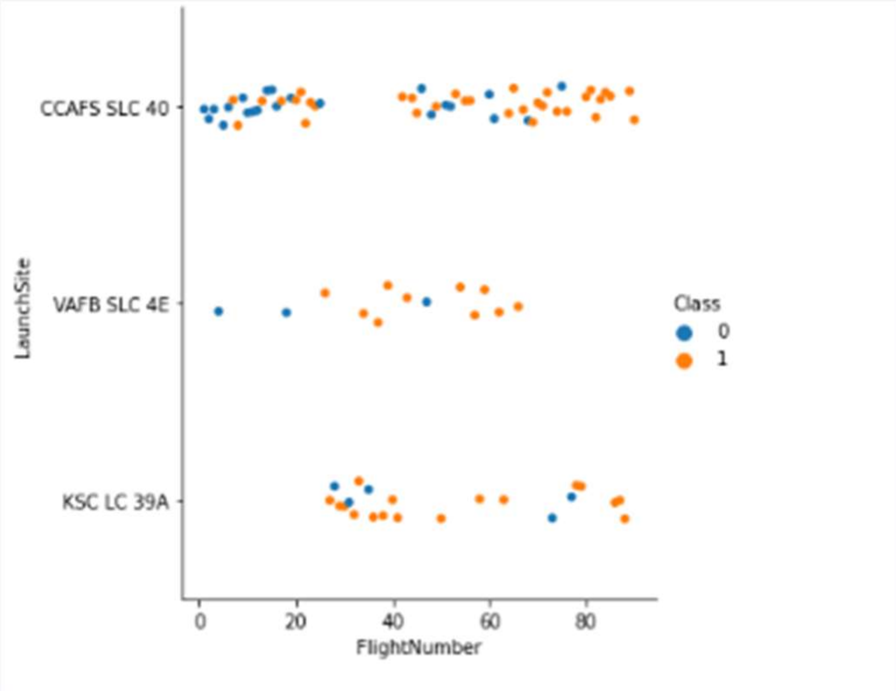
Results

- Exploratory data analysis highlight the relevant cause on successful launch
- Predictive analysis show that we can predict the result of a launch with an accuracy of 83.3%

Section 2

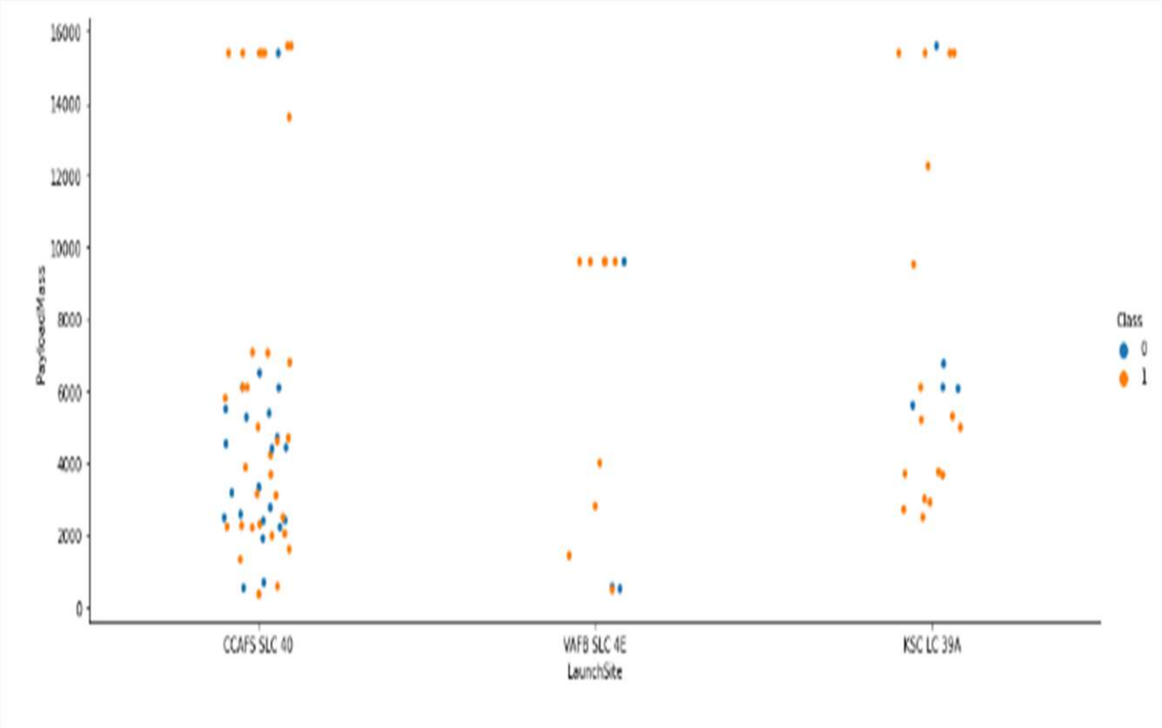
Insights drawn from EDA

Flight Number vs. Launch Site



We can see that most of bad landing are on the first flight . Experience increase the success.
We also can see than we have more bad landing on CCAFS SLC 40 regardless of the experience effect.

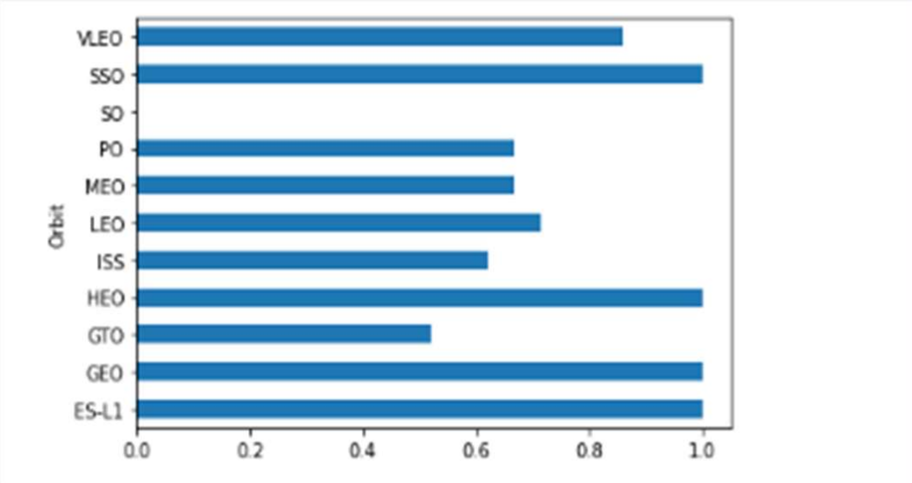
Payload vs. Launch Site



The VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

CCAFS SLC 40 is the most used launch site. He ‘s more efficient on the launch of heavy payload.

Success Rate vs. Orbit Type



Four orbites have 100% of success (SSO,HEO,GEO,ES-L1)

One orbite have more than 80% of success (VLEO)

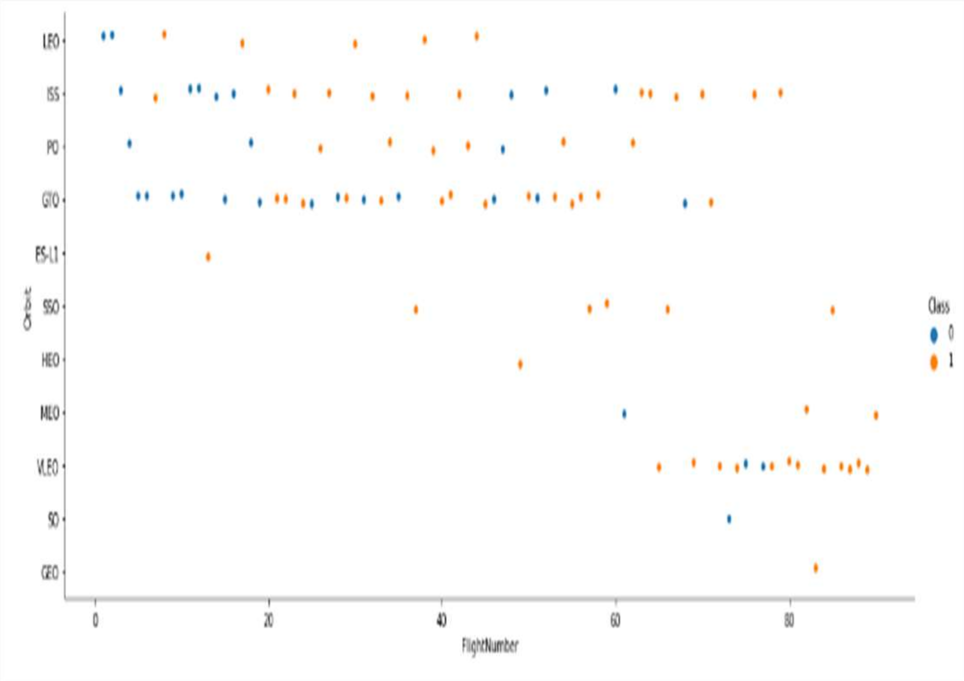
Four orbite have arround 60% of success (PO,MEO,LEO,ISS)

GTO have approximatly 50% of succes.

SO have no succes

We have to be careful with the chart because we don't look at the number of attempt per orbite at that point.

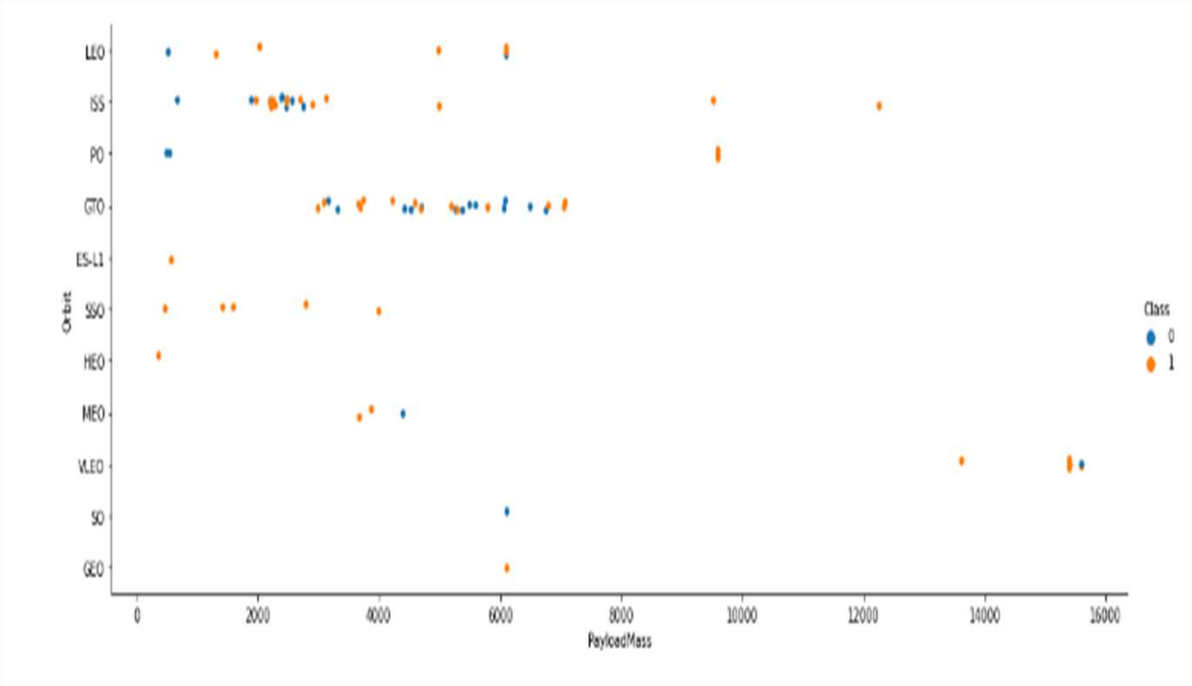
Flight Number vs. Orbit Type



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.

SO and GEO had only on attempt.

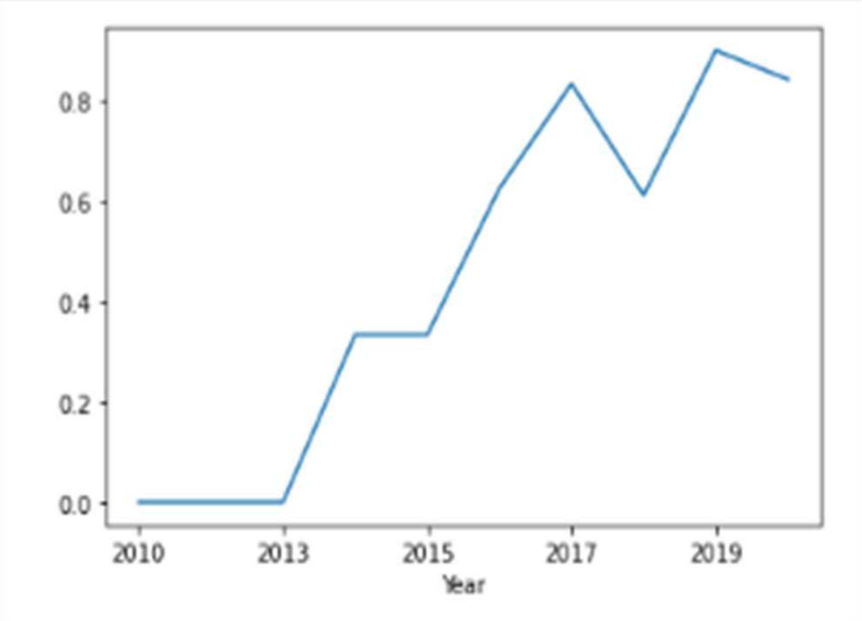
Payload vs. Orbit Type



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



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

All Launch Site Names

- The unique launch site are:
  - CCAFS SLC 40
  - VAFB SLC 4E
  - KSC LC 39A
- Using the DF.unique() method we saw that we have 3 launch site.

All Launch Site Names

- We have 4 launch site used.

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

Launch Site Names Begin with 'CCA'

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005

- To find records with launch site which begins with “CCA” I use the code below. The result is the 5 first records which meet the requirement.
  - df[df.LaunchSite.str.startswith('CCA')].head()

Launch Site Names Begin with 'CCA'

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
2013-12-03	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

- The total payload carry by boosters launched by NASA (CRS) is 20.
- It seems that NASA launched 20 tons of materials.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928 kgs.

- We must not confuse with median weight.

First Successful Ground Landing Date

- The first successful landing outcome on ground pad became the 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

F9 B4 B1040.2
F9 B4 B1040.1
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1054
F9 B5B1060.1
F9 B5B1062.1
F9 FT B1021.2

F9 FT B1031.2
F9 FT B1032.2
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1032.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

- The names of the booster which have 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 landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

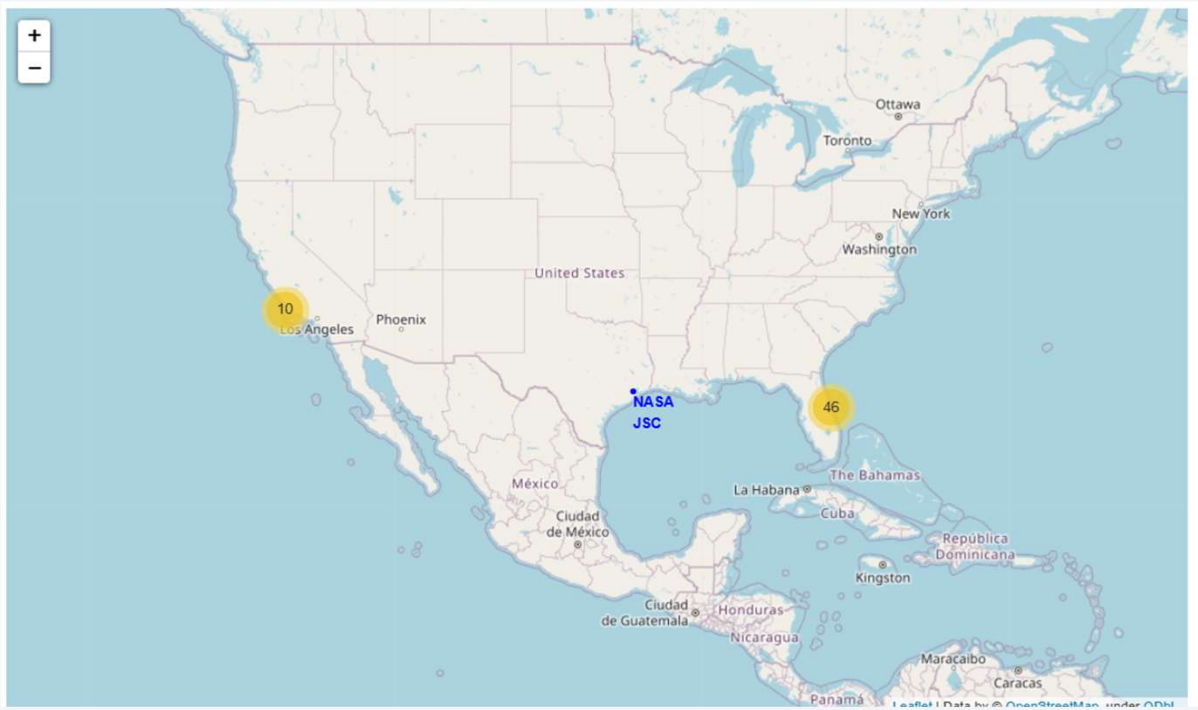
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

landing__outcome	DATE
No attempt	2017-03-16
Success (ground pad)	2017-02-19
Success (drone ship)	2017-01-14
Success (drone ship)	2016-08-14
Success (ground pad)	2016-07-18
Failure (drone ship)	2016-06-15
Success (drone ship)	2016-05-27
Success (drone ship)	2016-05-06
Success (drone ship)	2016-04-08
Failure (drone ship)	2016-03-04
Failure (drone ship)	2016-01-17
Success (ground pad)	2015-12-22
Precluded (drone ship)	2015-06-28
No attempt	2015-04-27
Failure (drone ship)	2015-04-14
No attempt	2015-03-02
Controlled (ocean)	2015-02-11
Failure (drone ship)	2015-01-10
Uncontrolled (ocean)	2014-09-21
No attempt	2014-09-07
No attempt	2014-08-05
Controlled (ocean)	2014-07-14
Controlled (ocean)	2014-04-18
No attempt	2014-01-06
No attempt	2013-12-03
Uncontrolled (ocean)	2013-09-29
No attempt	2013-03-01
No attempt	2012-10-08
No attempt	2012-05-22
Failure (parachute)	2010-12-08
Failure (parachute)	2010-06-04

Section 3

Launch Sites Proximities Analysis

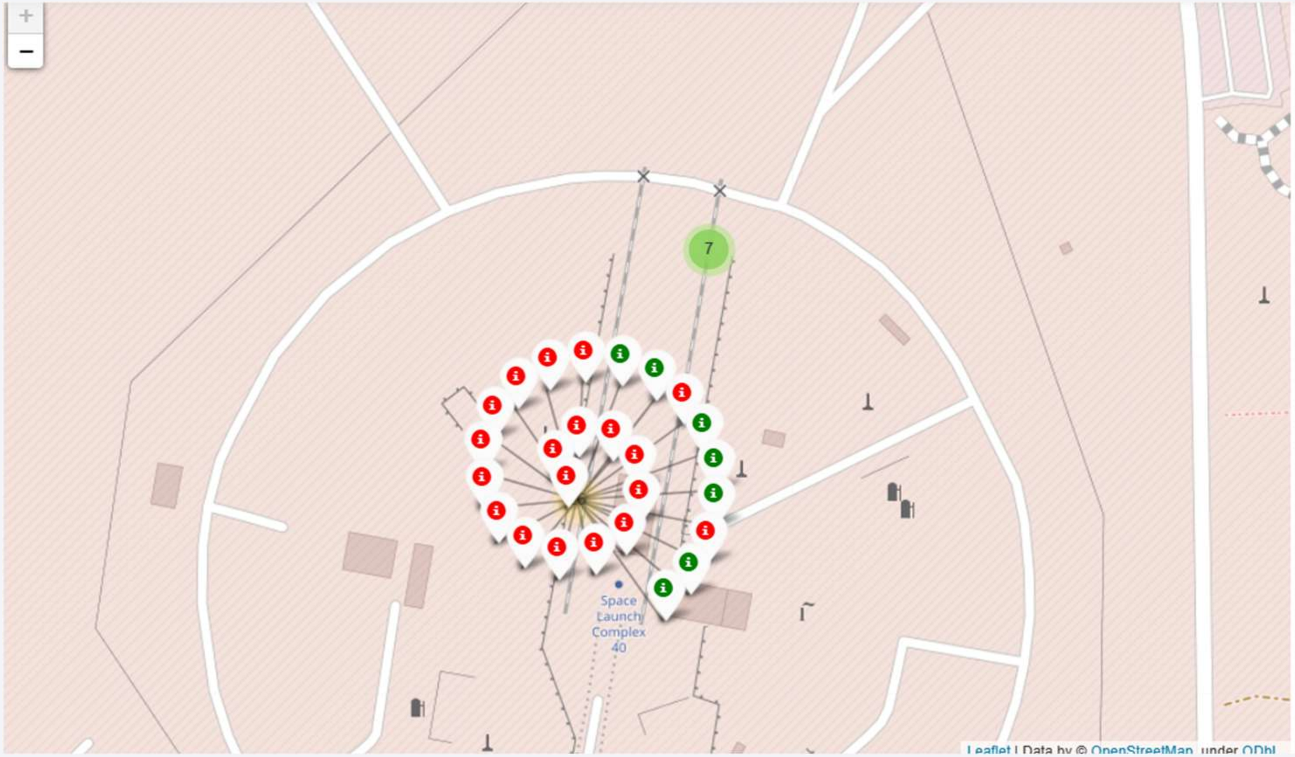
Launch Site localization



We can see launch sites are on the two costs of USA.

The east site launch almost 5 times more engine than the west one.

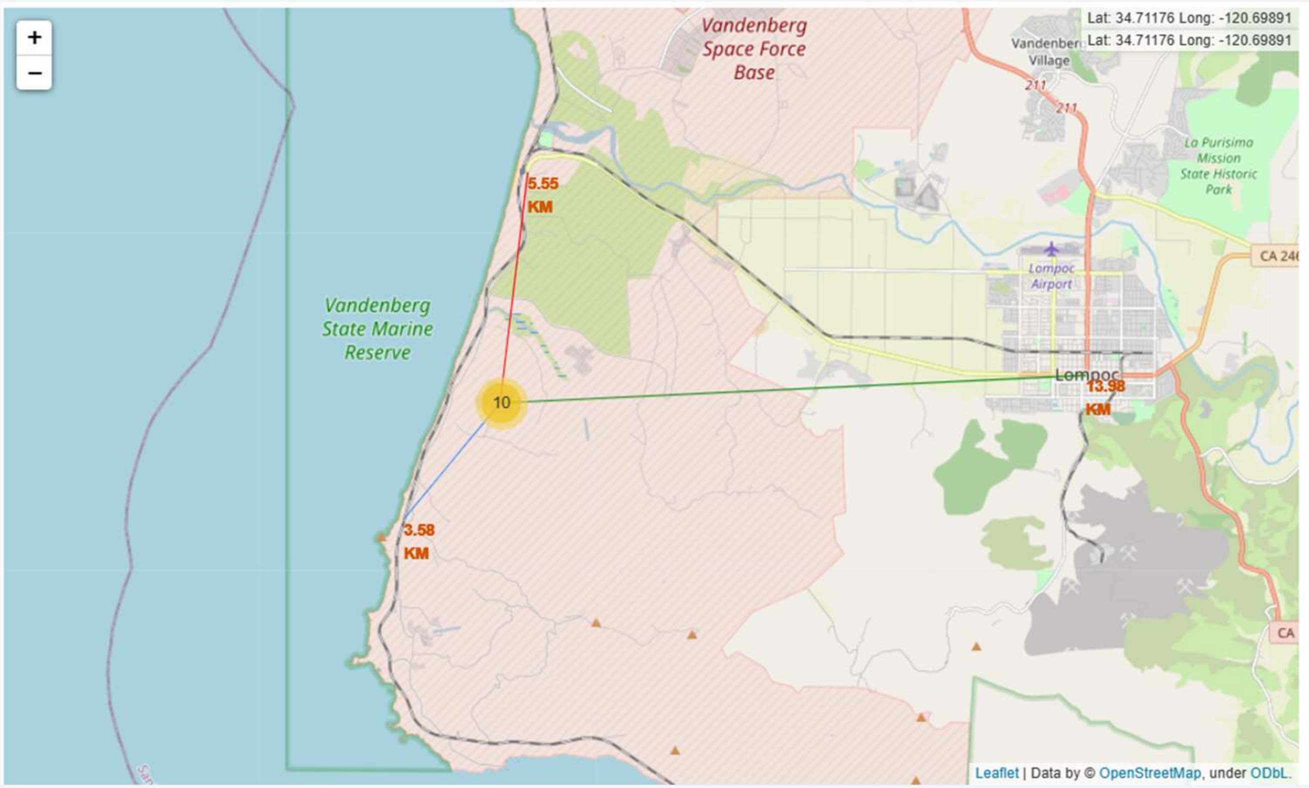
Outcome visualization per site



This map show us that only 7 launch success for 26 attempts.

We can focus on this details for each site.

Facilities locations



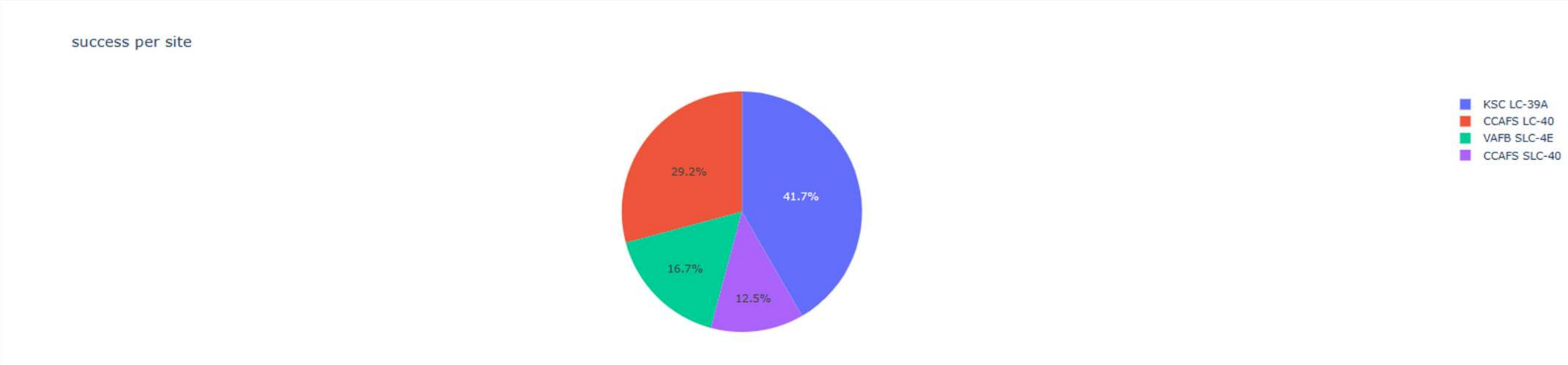
We can see geographic location with all the facilities for each site.

This visualization is very useful for logistic issues.

Section 4

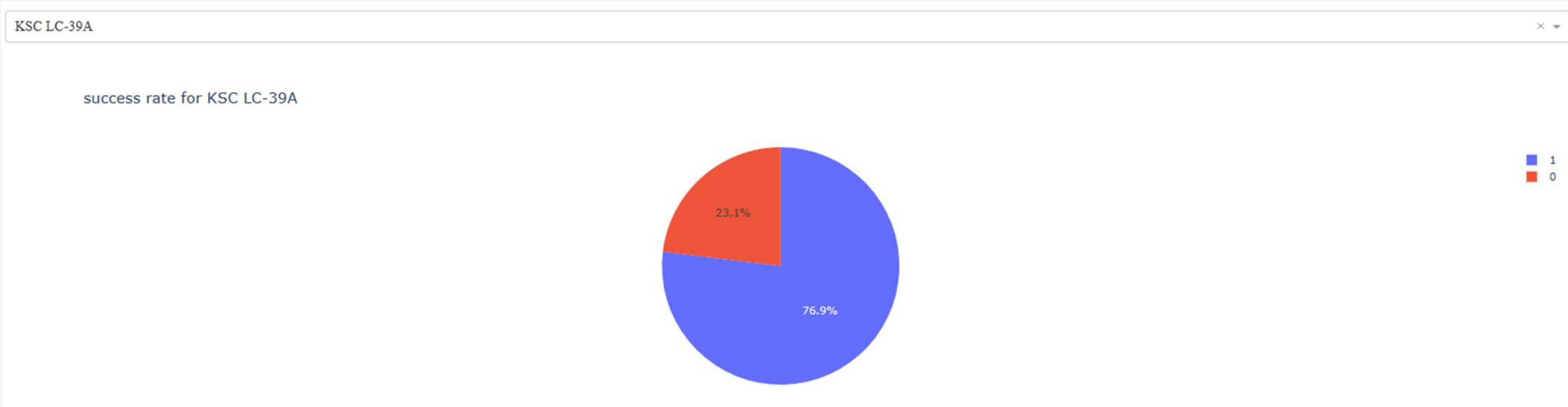
Build a Dashboard with Plotly Dash

Distribution by site of successful launches



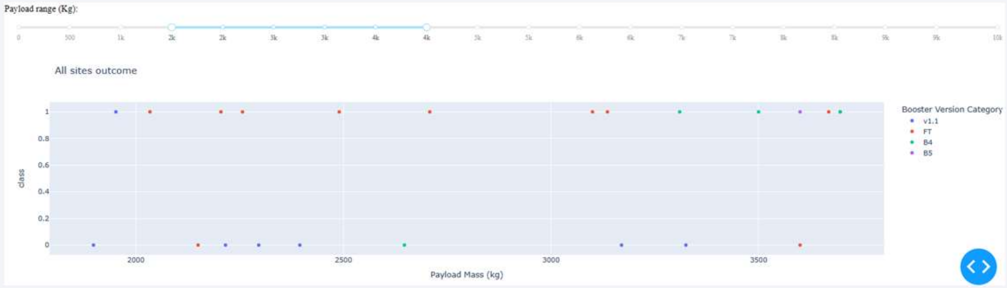
- Almost half of the success launch has been made from KSC LC-39A.
- A third of the success launch has been made from CCAFS LC-40

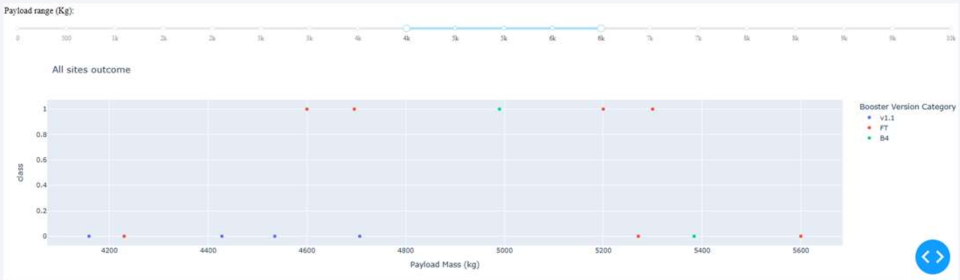
KSC LC-39A: launch outcome distribution.



- 76,9% of the launch has succeeded

Launch outcome by Payload Mass.



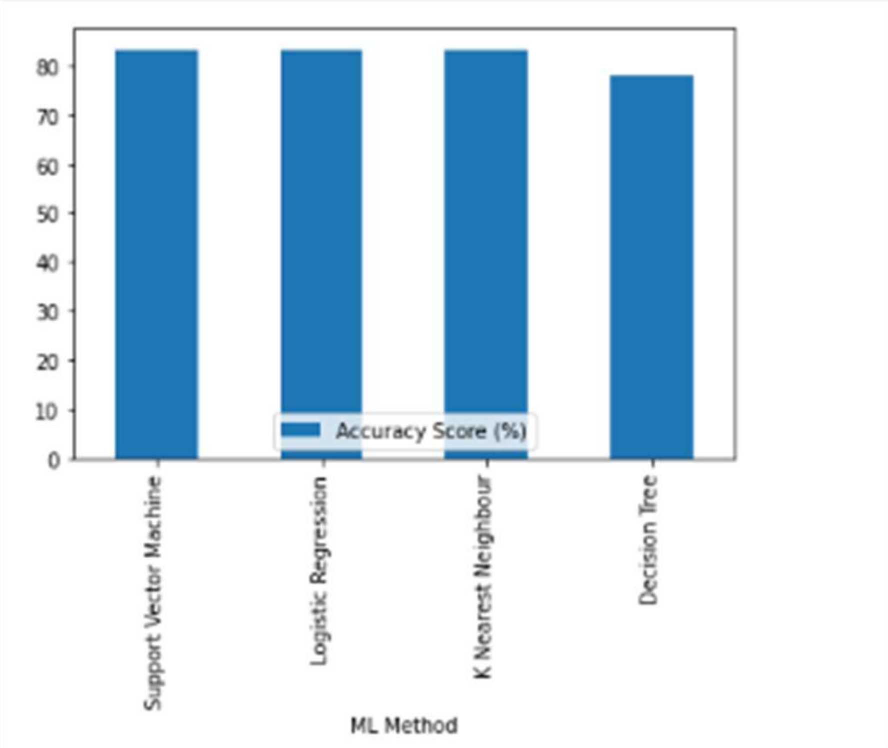


The first graph show us that most of the launch has been made for a payload between 2000k and 6000k. We can also see that FT booster is the most efficient.
The two other graph focuses on Payload (2000k increment). We can see that booster version V1.1 is not efficient.

Section 5

Predictive Analysis (Classification)

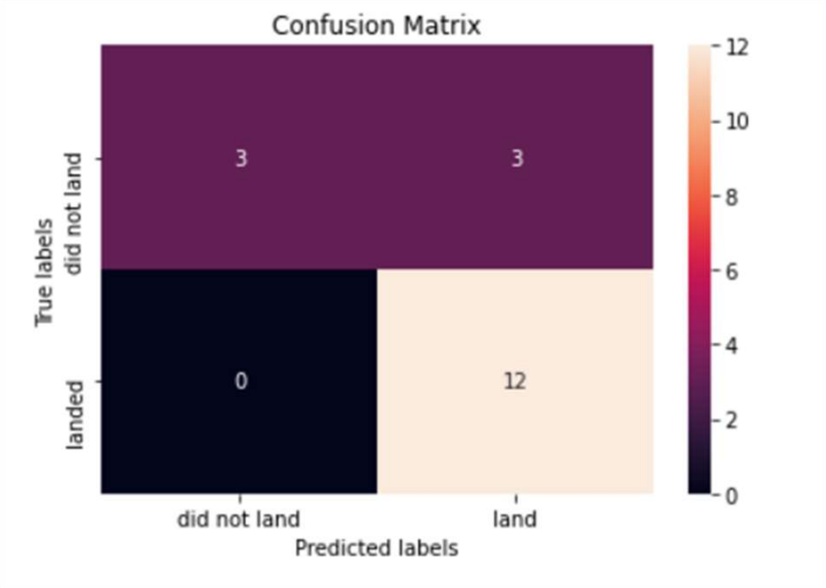
Classification Accuracy



- SVM, Logisitic Regression and K Nearest Neighbours have the same accuracy on test data set.

- The score is 83,33%

Confusion Matrix



- Our model predict good result for 100% of the real landed mission.

- Our model predict good result for 50% of the real crashed missions.

Conclusions

- Experience has an important impact on result.
- Results are better for smaller payload
- Orbite have an impact on result
- Booster version have a very important impact on result
- Predicitve models give us a good result at 83%.

Thank you!