IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Flavio Santos Moraes
16th Mars 2022

Outline

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

Executive Summary

- ## Summary of methodologies

1. Parsing and cleaning data using python language with Jupyter Notebook IDE
2. Transform data from web sites using “webscrapping – Beautifulsoup” libraries from Python
3. Data wrangling data to get key information from previous data
4. Data base manipulation with Python and SQL
5. Explore and prepare data using Matplotlib and Seaborn Python’s libraries
6. Using Folium to make interactive maps visualizations
7. Using Dash to create interactive visual analytics dashboard
8. Use of predictive machine learnings tools to analyze the problem

Executive Summary

- Summary of results

The final model was able to predict with 83.33% of accuracy when a Falcon 9 rocket will successfully land based on historic data and many key factors

Introduction

- Project background and context

The SpaceX is an aerospace manufacturer founded by Elon Musk with the goal of reducing space transportation cost.

The company has successfully managed to recover the first stage of a three rockets stages after its launch. The first stage is the biggest and most expensive part of the rocket so that’s one of the reasons SpaceX has managed to reduce the launch cost.

- Problems you want to find answers

Using data analysis, we’ll try to predict when a rocket from SpaceX will successfully land based in some of the information collect on the internet from Wikipedia and others open-source information sites.

Section 1

Methodology

Methodology

Executive Summary

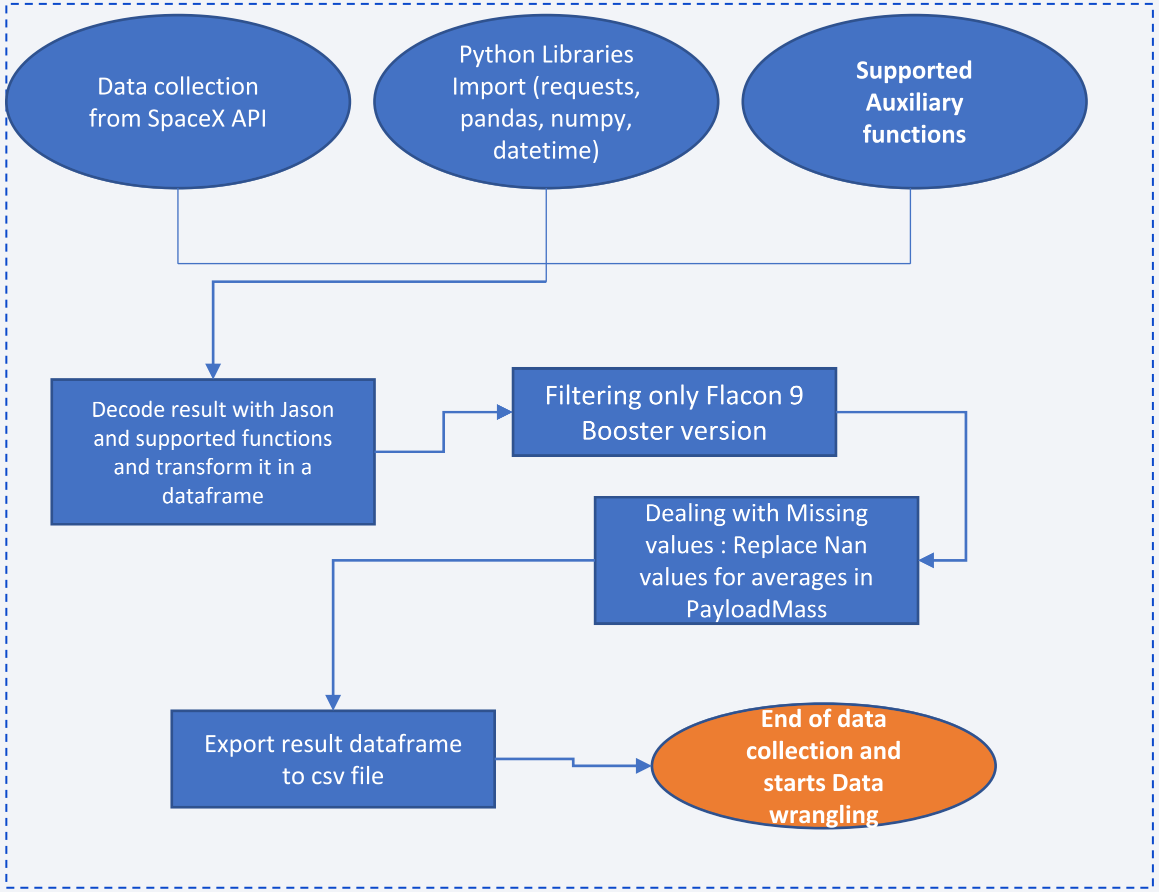
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using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

Data Collection

- The next two slides will show how the data has been collected and the main sources
- Also, the link for the Jupyter Notebook with the code behind the data collection is provided through GitHub

Data Collection – SpaceX API

- Flowchart Data collection



Flowchart 1: Data Collection

- Github link

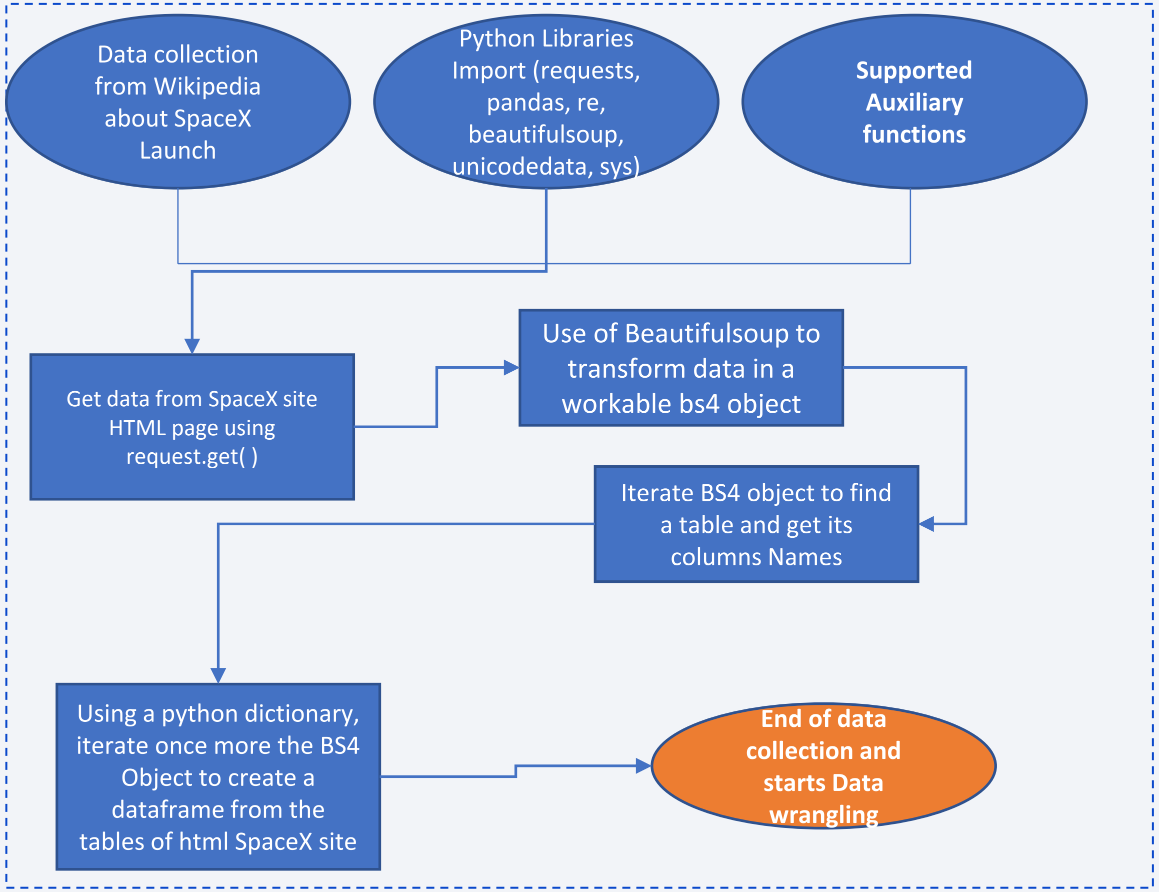
https://github.com/fsmoraes78/applied-data-science/blob/2b540d05fc592d75e16e9f583cfaedcbc6df8dc2/jupyter-labs-spacex-data-collection-api.ipynb

- SpaceX API link

https://api.spacexdata.com/v4/rockets/

Data Collection - Scraping

- Flowchart Data collection –web scrapping



Flowchart 2: Data Collection web scrapping

- Github link

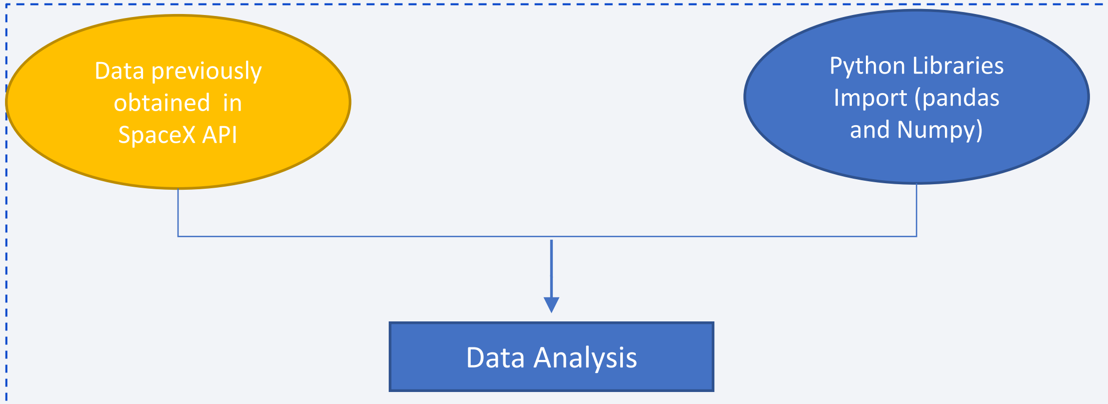
https://github.com/fsmoraes78/applied-data-science/blob/abc2b9f5f1bacc696f68f11124f1d446828ba26c/jupyter-labs-webscraping.ipynb

- Wikipedia link

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

Data Wrangling

- Exploratory Data Analysis



From DataFrame:

- Identify nulls values
- Check types
- Calculate launches per Site
- Calculate numbers and occurrences of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column

Flowchart 3: Data Wrangling

Launch per sites results:

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

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

Fig1: Launch per sites

Orbits:

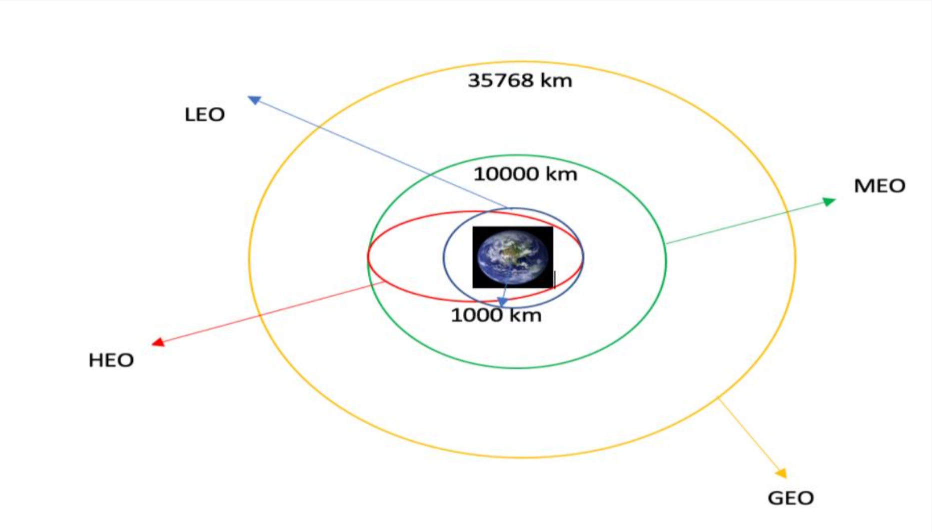


Fig2: Orbits

Data Wrangling

- Determine Training Labels

Final training label created according to success/fail landing.

```
display(df.head(5))
```

erVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

- Github link:

https://github.com/fsmoraes78/applied-data-science/blob/11adc5bc9f5f27db97bb95fd60016fafa0f1b896/Applied%20Data%20Science_wrangling.ipynb

EDA with Data Visualization

All the graphs and results will be present later in this presentation. Below you can find the github link for the Jupyter notebook

- *GitHub link:*

  https://github.com/fsmoraes78/applied-data-science/blob/e8fa09d0f64ae6a860758def5f32143fe5624233/Applied%20Data%20Science_eda_matplolib.ipynb

EDA with SQL – Tasks and SQL performed

The queries statements and the results will be present later in this presentation.
Below you can find the github link for the Jupyter notebook

- *GitHub link:*

  https://github.com/fsmoraes78/applied-data-science/blob/1db62d5fbefe7845a4dee1e6e371030ec211d53a/Applied%20Data%20Science_sql.ipynb

- *Watson studion link: (since bd not working from Github):*

  https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/521aa1a9-6182-43c0-9f82-586144c854ba/view?access_token=da96fa00d0ea33351c7501230eb872358089e15631666a47baa72985466229c1

Build an Interactive Map with Folium

- We’ve built an interactive map to analyze the launch site proximity with Folium
- We’ve created markers to sign all launch sites on a map
- We’ve created Circles to highlight and label an area in the map
- We’ve created markers_clusters to indicate in the map the number of launches where the first stage of the rocked has succeed to land
- We’ve used mousepoint to find the Latitude and Longitude of some point in the Map and calculate its distance from Launch site.
- *GitHub link:*

  https://github.com/fsmoraes78/applied-data-science/blob/a01f874701fe58cf01c556b4401ecedec336af85/lab_jupyter_launch_site_location_(folium%20MAP).ipynb

Build a Dashboard with Plotly Dash

- Using Dash library, we are able to create interactive dashboard

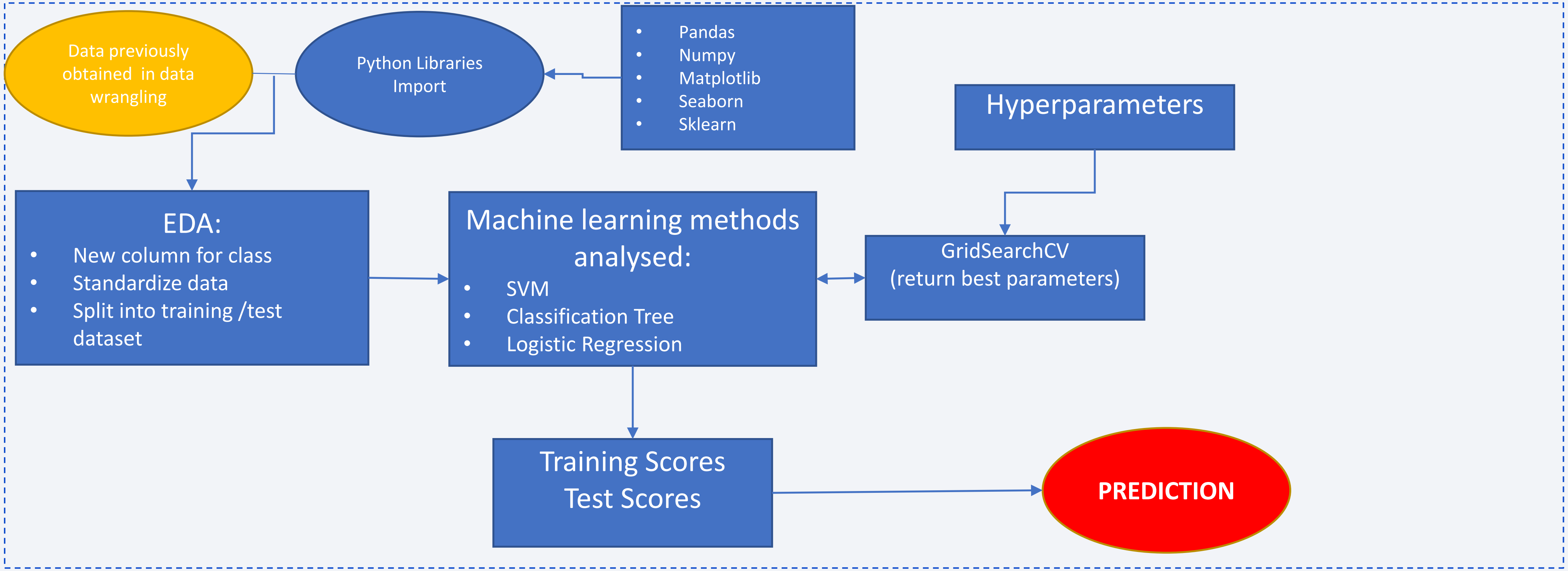
- We have created two main graphs:
  - A pie chart showing the number of launches per site and the number of successful lands per site once the it is selected by a dropdown combo.

  - The second graph is a scatter graph showing the relation between success land and Payload mass per site. We have added a slider bar to filter the Payload range also we can see the results per site by selecting the site in the previous dropdown selector.

- *GitHub link:*

  https://github.com/fsmoraes78/applied-data-science/blob/a01f874701fe58cf01c556b4401ecedec336af85/spacex_dash_app.ipynb

Predictive Analysis (Classification)



Flowchart 4: Flow of predictive analyses

- *GitHub link:*

  https://github.com/fsmoraes78/applied-data-science/blob/1db62d5fbefe7845a4dee1e6e371030ec211d53a/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

In the next slides you will find:

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

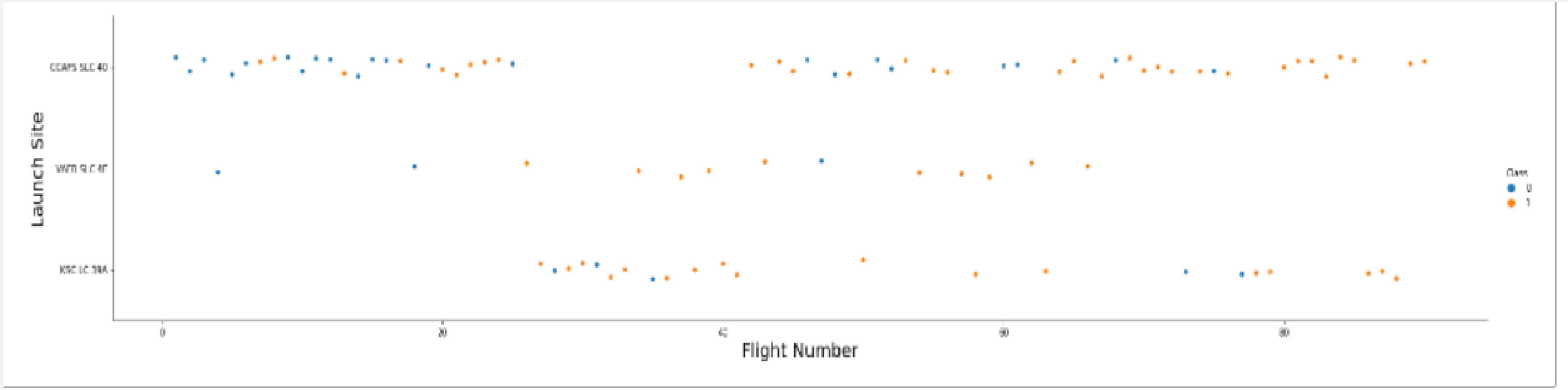


Fig3. Success landing (orange points) according to launch site vs Fligh Number

“ We see that as the flight number increases, the first stage is more likely to land successfully ”

Payload vs. Launch Site

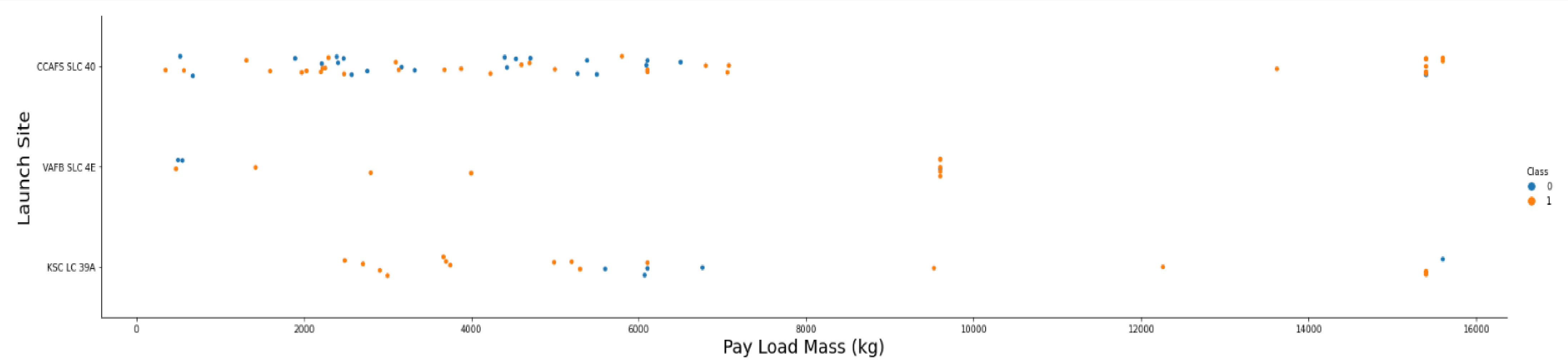


Fig4. Success landing (orange points) according to launch site and Payload Mass

“ We can see the payload mass is also important; it seems the more massive the payload, the more likely the first stage will return. ”

Success Rate vs. Orbit Type

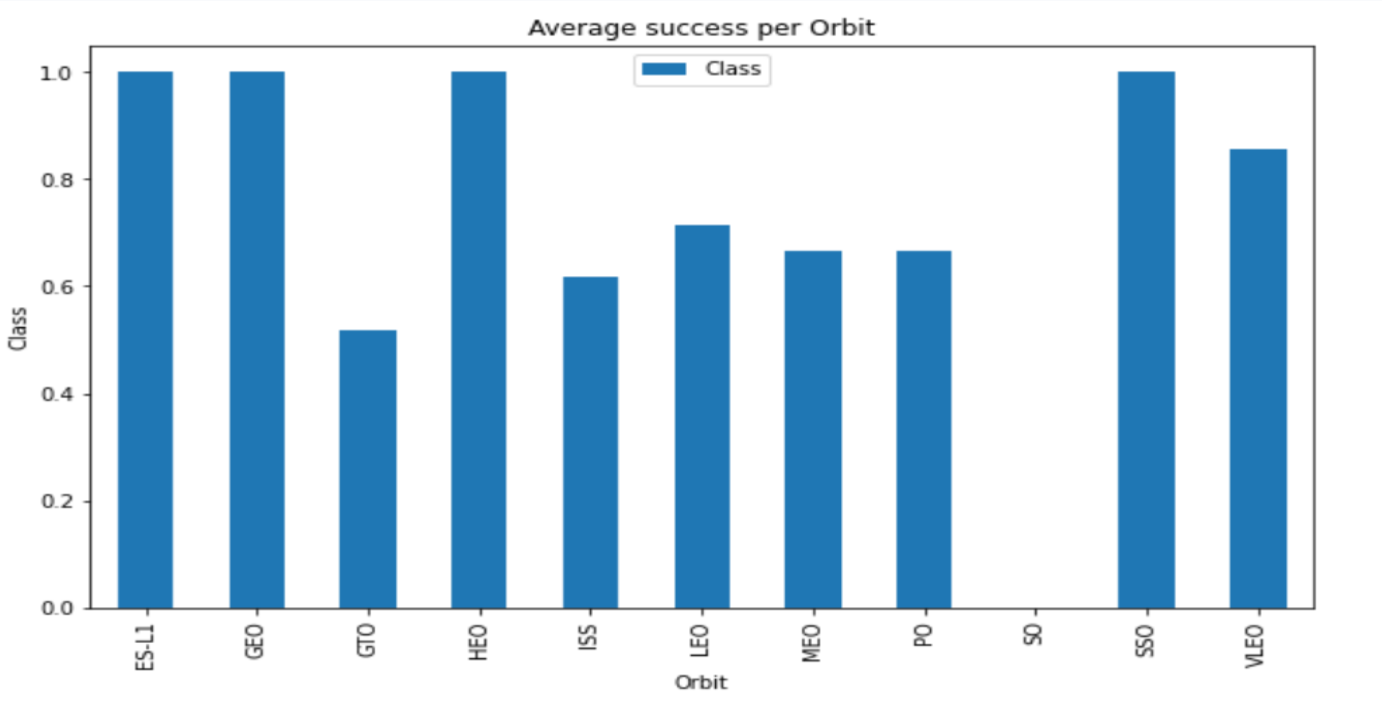


Fig5. Average success per Orbit

The Orbits ES-L1 GEO HEO SSO had 100% return success

Flight Number vs. Orbit Type

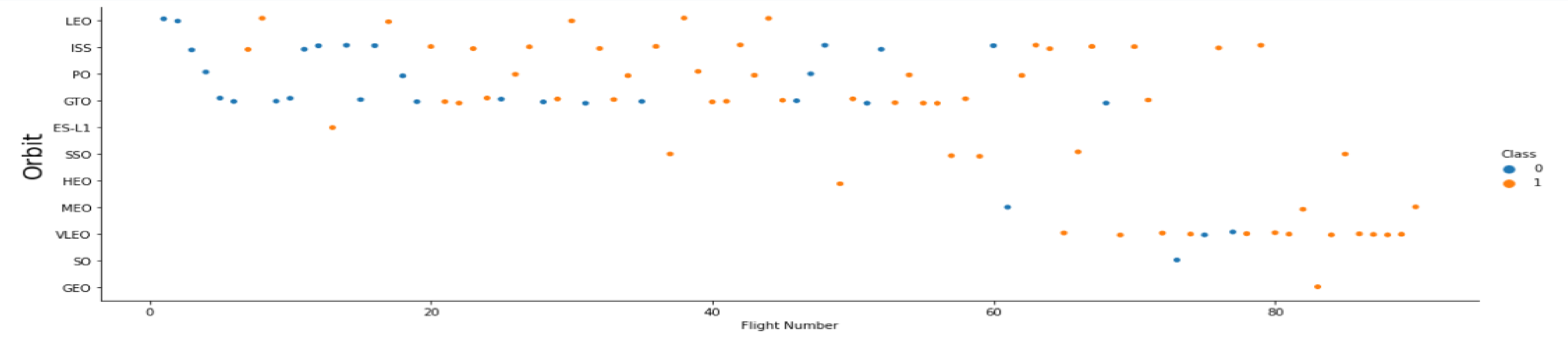


Fig6. Relation Orbit vs Flight Number

“ We can see that 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 after the first success ”

Payload vs. Orbit Type

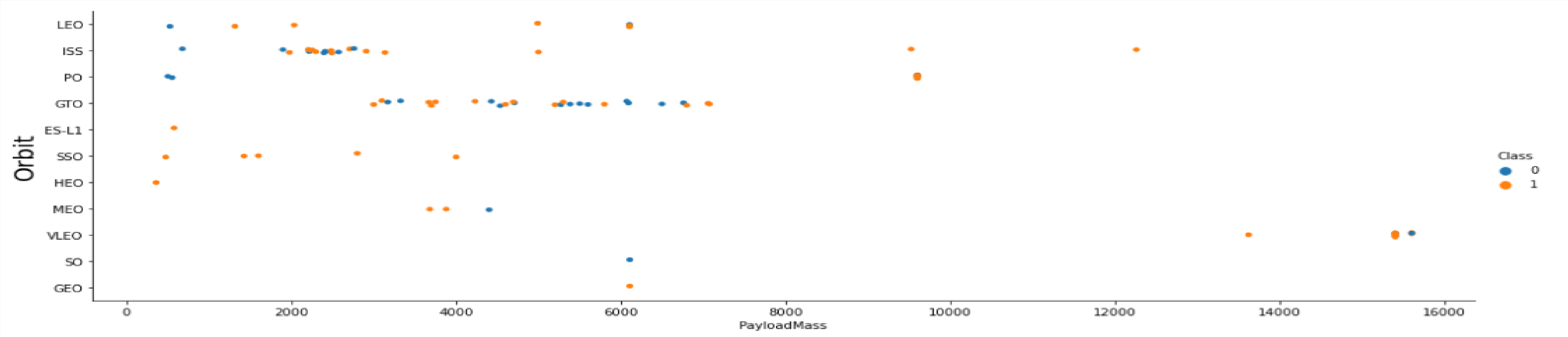


Fig7. Relation Orbit vs Payload Mass

“ 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

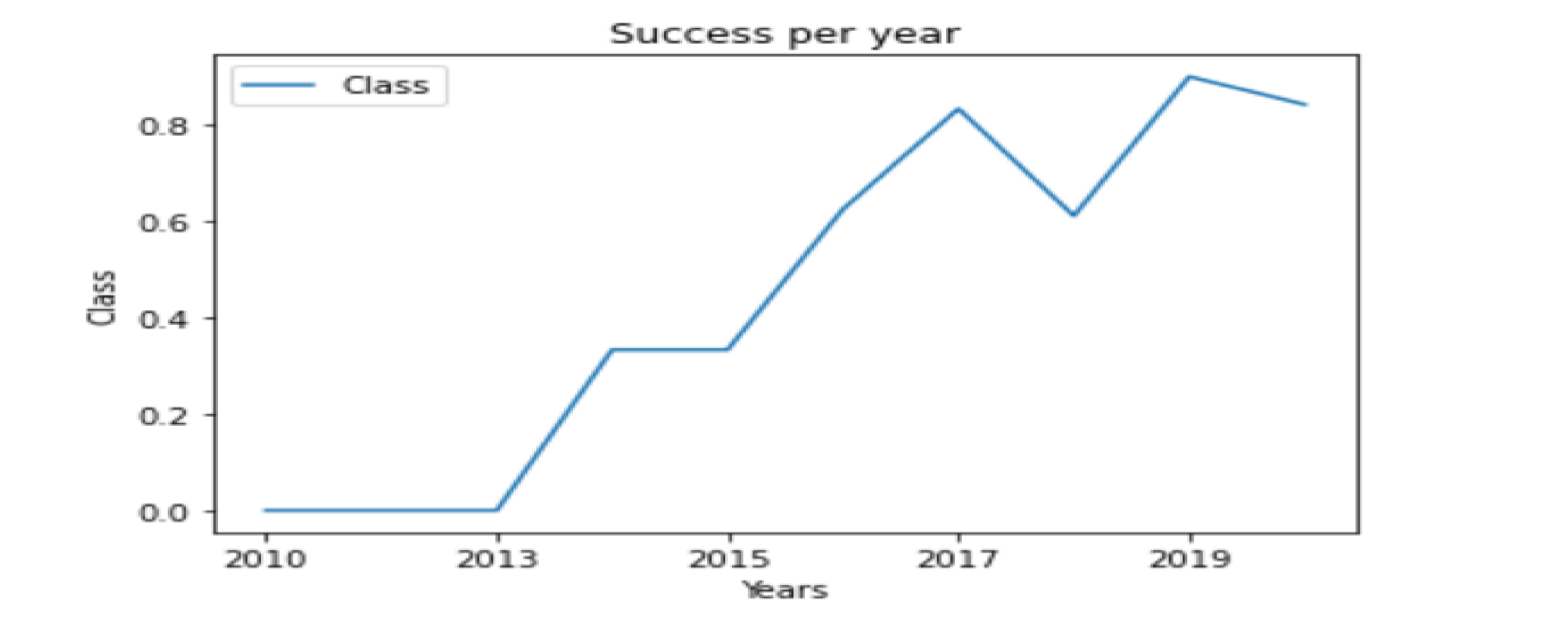


Fig8. Graph line success per year

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

All Launch Site Names

- SELECT DISTINCT launch_site FROM SPACEXTBL

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

Fig9. Result of select distinct query (to avoid duplicates)

Launch Site Names Begin with 'CCA'

- SELECT * FROM SPACEXTBL WHERE launch_site LIKE '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

Fig10. Partial view of conditional select query with part of string

Total Payload Mass

- SELECT booster_version, SUM(payload_mass__kg_) as total_pl FROM SPACEXTBL GROUP BY booster_versionPresent your query result with a short

```
 * ibm_db_sa://phy09264:***@5
Done.
Out[23]:
```

booster_version	total_pl
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092

Fig11. Partial view of the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

- SELECT booster_version, AVG(payload_mass__kg_) as avg_pl FROM SPACEXTBL GROUP BY booster_version

booster_version	avg_pl
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092

Fig12. Partial view of the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

- SELECT min(DATE) as st from SPACEXTBL where mission_outcome = 'Success'

st
2010-06-04

Fig13. Result query showing date of the first success mission that landed back in the ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT DISTINCT booster_version,payload_mass__kg_ from SPACEXTBL where mission_outcome = 'Success' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000

booster_version	payload_mass__kg_
F9 B4 B1040.2	5384
F9 B4 B1040.1	4990
F9 B5 B1046.2	5800
F9 B5 B1047.2	5300
F9 B5 B1048.3	4850
F9 B5 B1051.2	4200
F9 B5 B1058.2	5500
F9 B5B1054	4400
F9 B5B1060.1	4311
F9 B5B1062.1	4311
F9 FT B1021.2	5300
F9 FT B1031.2	5200
F9 FT B1032.2	4230
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1030	5600
F9 FT B1032.1	5300
F9 v1.1	4535
F9 v1.1 B1011	4428
F9 v1.1 B1014	4159
F9 v1.1 B1016	4707

Fig14. Result query showing the booster names that had success in drone ship and have payload between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

- SELECT count(mission_outcome) as total_success, mission_outcome FROM SPACEXTBL GROUP BY mission_outcome

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

Fig15. Result query showing the number of success of mission_outcomes (not included the landing proccess)

Boosters Carried Maximum Payload

- SELECT booster_version, payload_mass__kg_ FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXTBL)

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Fig16. Result query showing the boosters with highest payload mass (of 15 600 kg)

2015 Launch Records

- SELECT DATE, landing__outcome, booster_version FROM SPACEXTBL WHERE landing__outcome LIKE '%drone%' and year(DATE)=2015

DATE	landing__outcome	booster_version
2015-01-10	Failure (drone ship)	F9 v1.1 B1012
2015-04-14	Failure (drone ship)	F9 v1.1 B1015
2015-06-28	Precluded (drone ship)	F9 v1.1 B1018

Fig17. Result query showing the drone ship fails landing in 2015

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

- SELECT count(landing__outcome) as total_success, landing__outcome FROM SPACEXTBL WHERE date >='2010-06-04'and date <='2017-03-20' GROUP BY landing__outcome ORDER BY total_success DESC

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

Fig18. Result query ranking the outcomes between 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

Launch Sites Localization Map

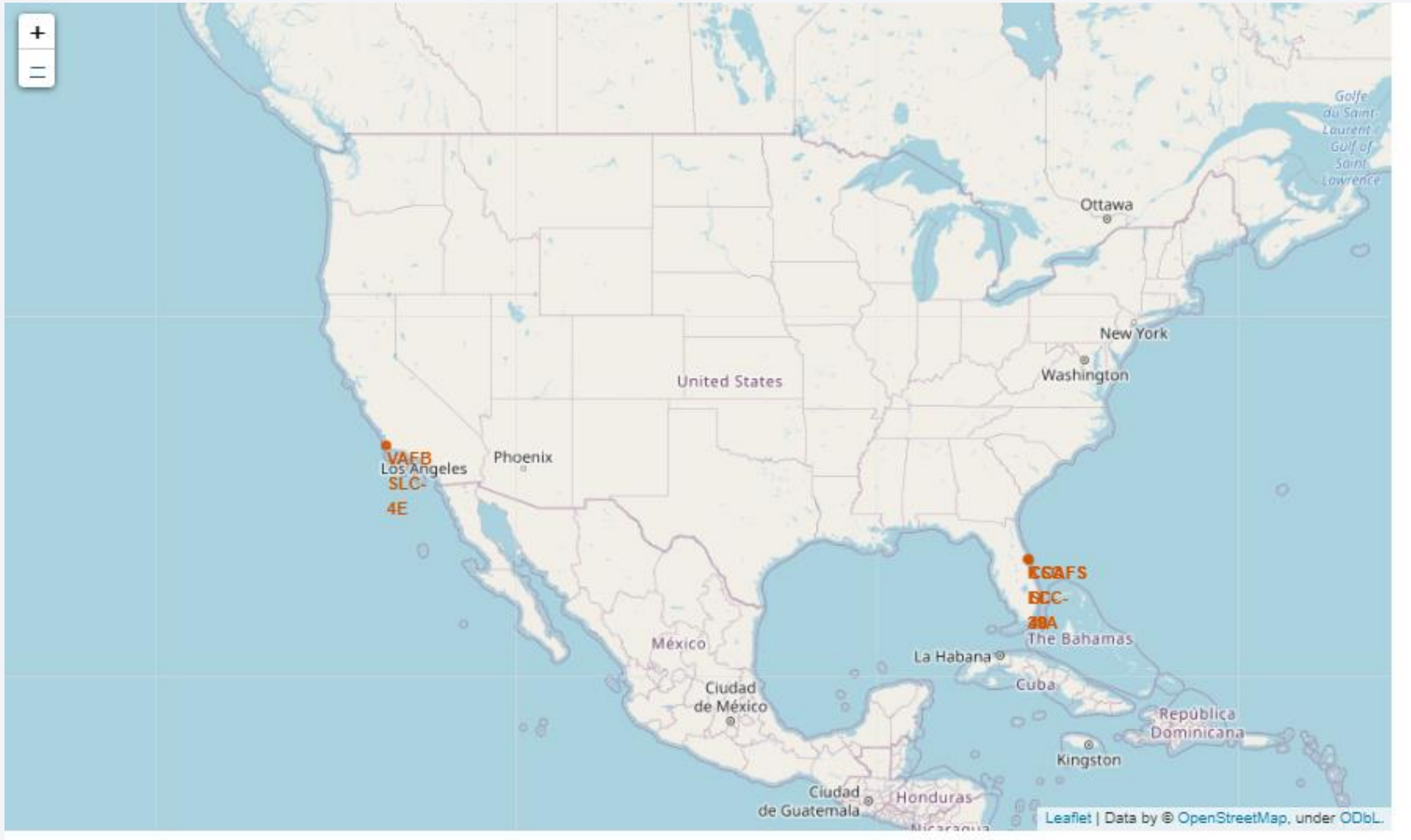


Fig19. Map showing the space X launch sites

Marker–clusters showing success landing outcome

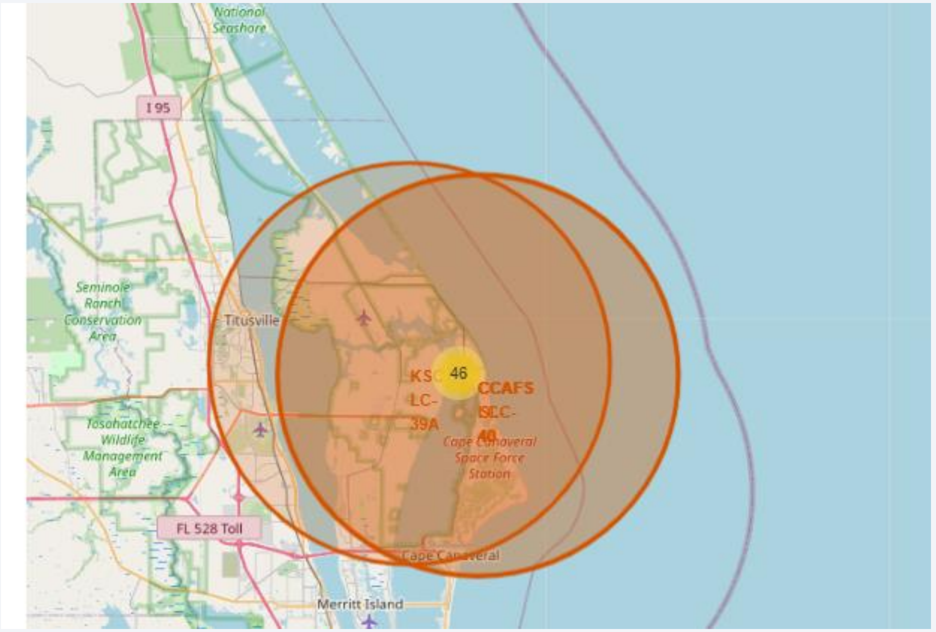


Fig20. Success landing outcomes

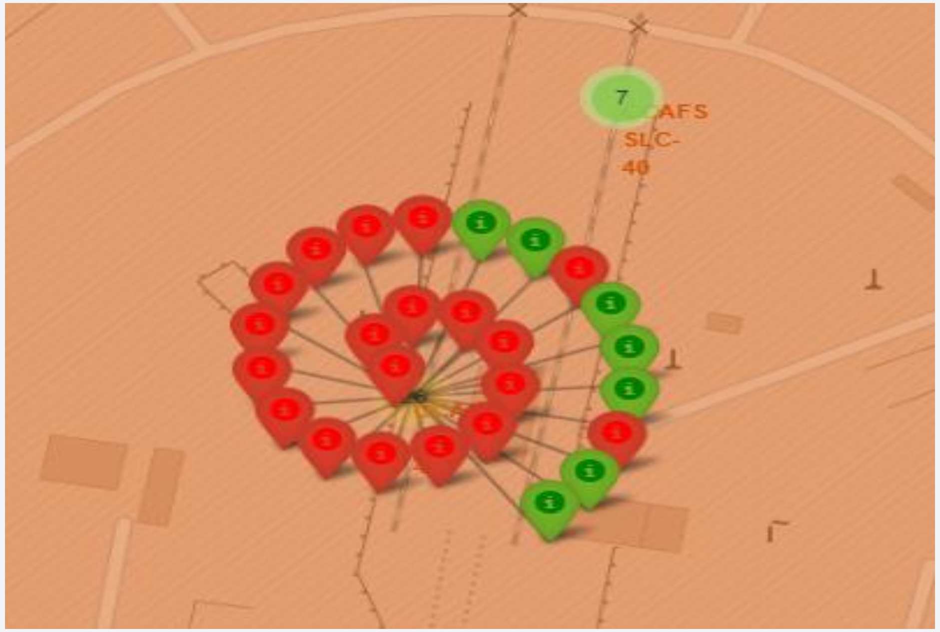


Fig21. Success landing outcomes -details

Distance line between Sta-Maria City and launch site

Fig22. Line showing the distance between the launch site VAFB SLC-4E and the Santa Maria downtown (39 km – straight line)

Section 4

Build a Dashboard with Plotly Dash

Dashboard with Dash – Launch counts

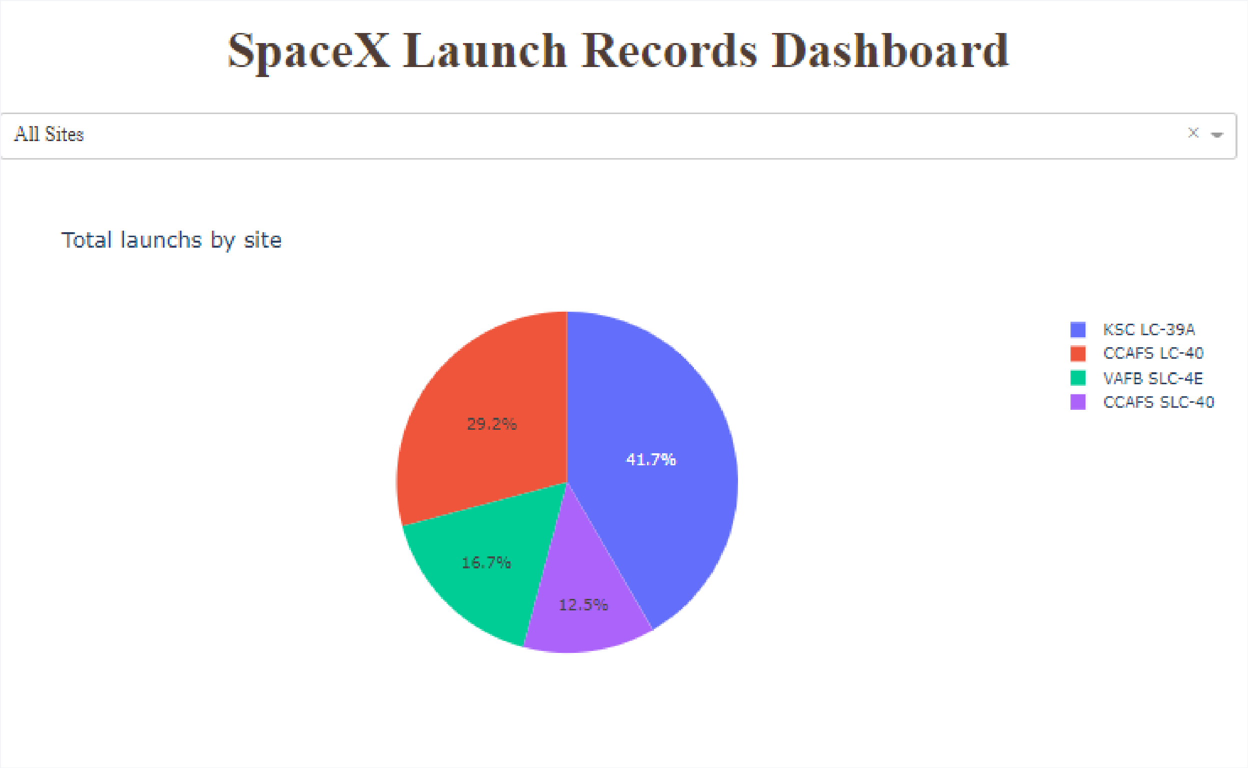


Fig23. Pie Chart showing the number of launchs per site.

Dashboard with Dash – Success per site

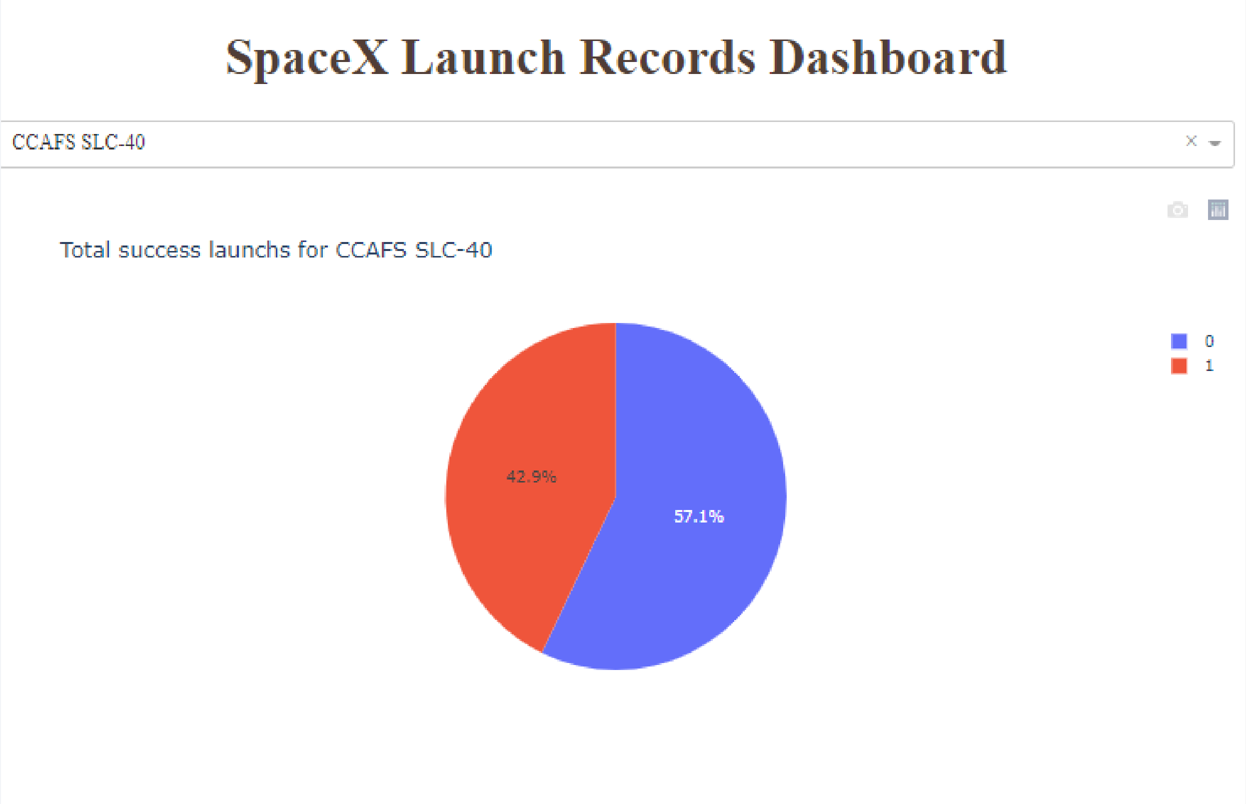


Fig24. Pie Chart showing the number of success from the site CCAFS SLC-40 .

Dashboard with Dash – Scatter graph and slider bar

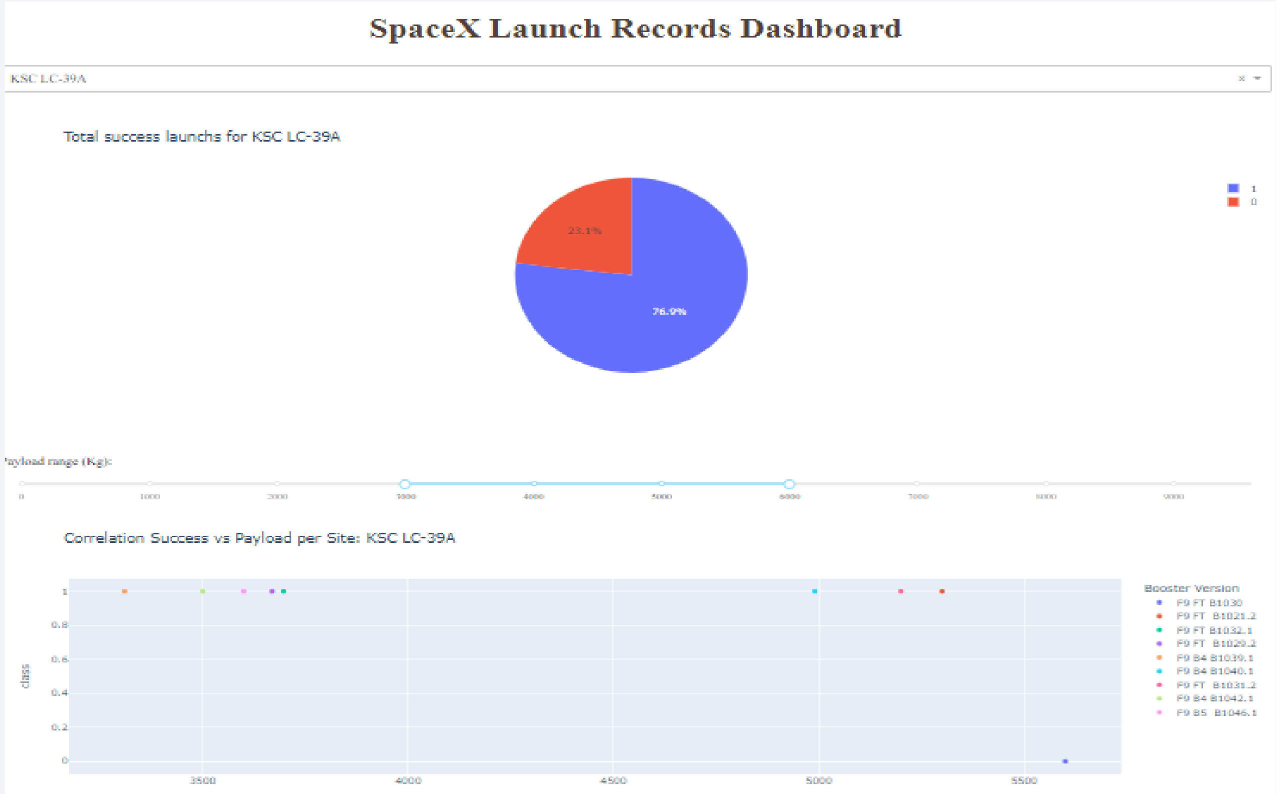


Fig25. Slider bar to select payload mass range according to selected site

Dashboard with Dash – Scatter graph and slider bar

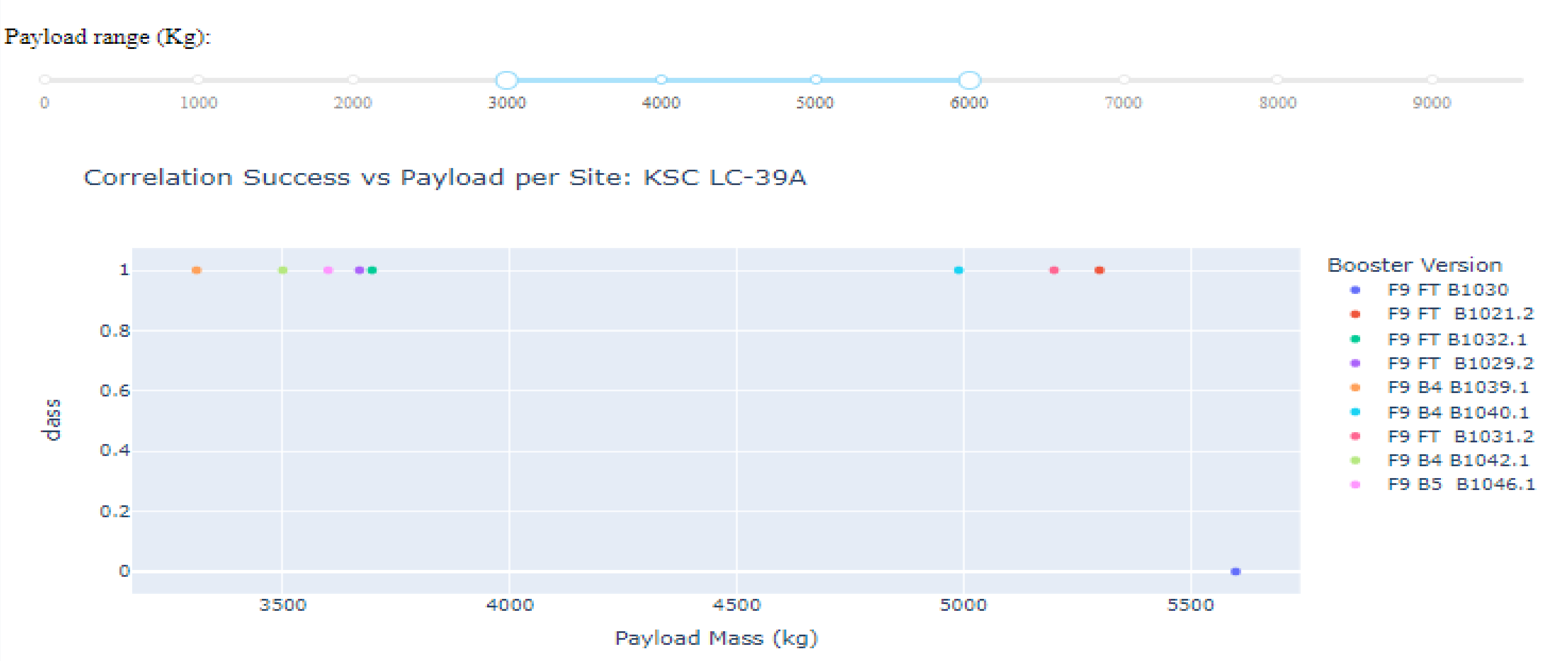


Fig26. Slider bar to select payload mass range according to selected site (Zoom) –Success per Booster/Paylaod mass

Section 5

Predictive Analysis (Classification)

Classification Accuracy

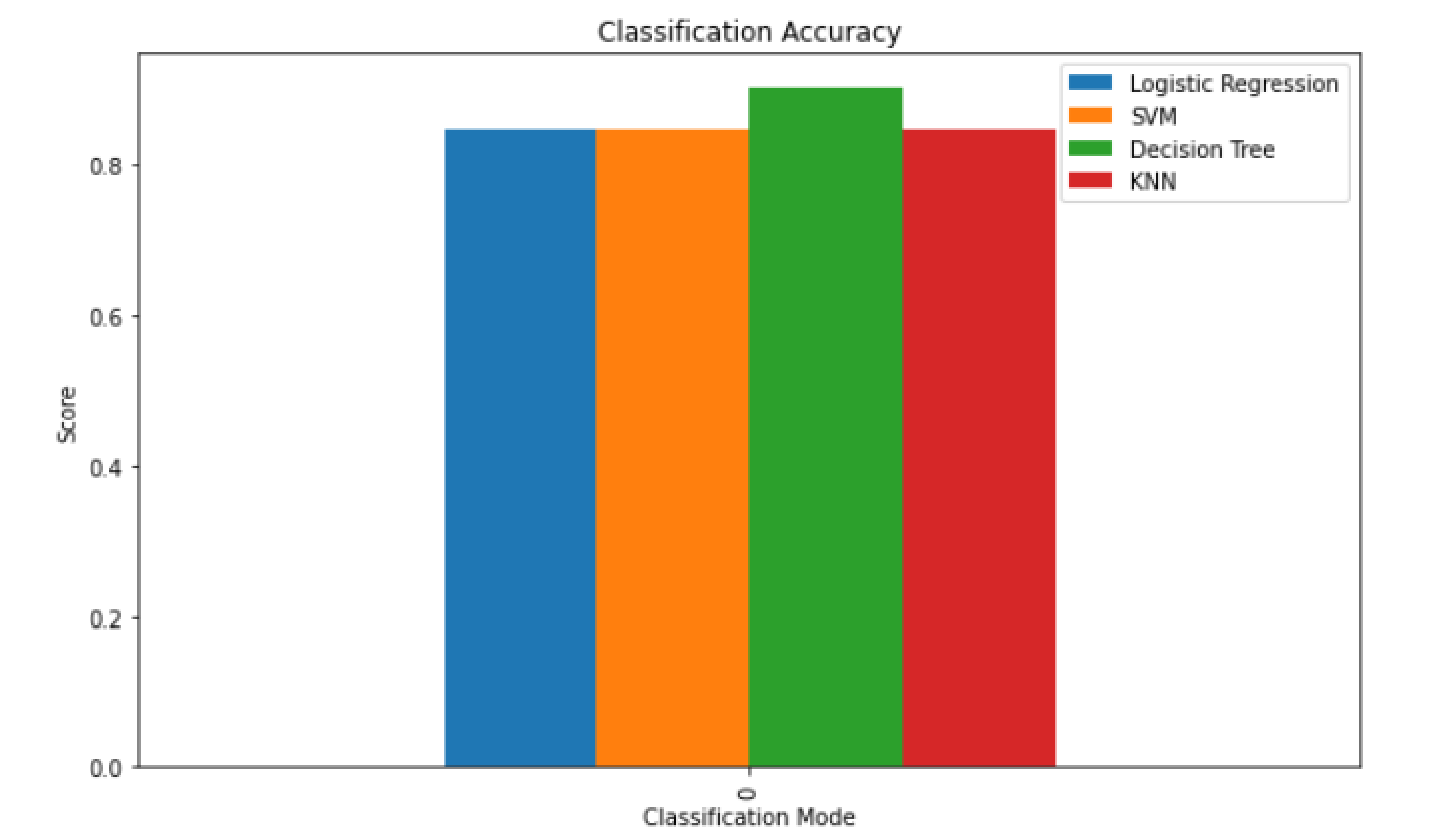


Fig27. Bar Chart – Best Model selection
Decision Tree – 90.35% accuracy

Confusion Matrix

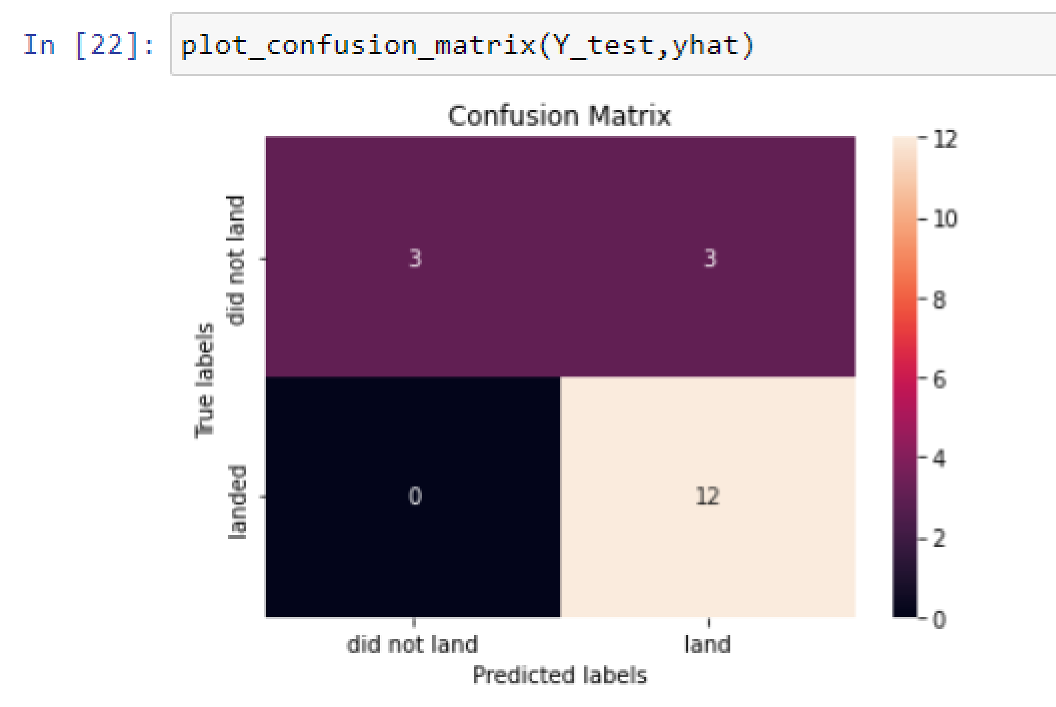


Fig27. Confusion Matrix from Decision Tree Test settings results – Predicted 15 out of 18 correctly

Conclusions

- The use of machine learning is an excellent and easy way to create a tool to predict an event, in this case the success of landing a rocket from Space X
- Personally, I think the main risk of the Machine Learning approach is to not understand what are being analyzed. That's the reason I think EDA is even more important than the Machine learning process
- The main advantage of using the machine learning is using the power of the computer to analyze a very big quantity of data and avoid to use complicate calculus
- With about 90% of certain of prediction we could go deep in the exploratory data analysis to make the prediction tool even more accurate

Appendix

- Github general link:

  https://github.com/fsmoraes78/applied-data-science

Thank you!