IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lucas Argento

January 5th, 2022

Outline of the Project

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

Executive Summary

1

SpaceY is a new commercial rocket launcher that wants to bid against SpaceX. They needed to understand how SpaceX can manage to keep their costs so low.

2

Analyzing this data and building classification Machine Learning Models we managed to understand which are some of the key parameters for a successful landing and to predict if a new given launch will be able to launch its first stage or not.

3

Knowing that is is mainly because of the re-usability of their rockets, mission parameters were extracted from SpaceX public databases for later study.

4

We are now able to predict a successful landing with an accuracy level of around 83%. We are also providing SpaceY handy BI tools such as online dashboards for better understanding of future data.

Introduction - Some Context

SpaceY is a new commercial rocket launcher that wants to bid against SpaceX

SpaceX actually offers a mindblowing low cost for launches at less than half the price of its competitors.

They are able to achieve this because *most* of their launches manage to *land the first stage* of the rocket and reuse it.

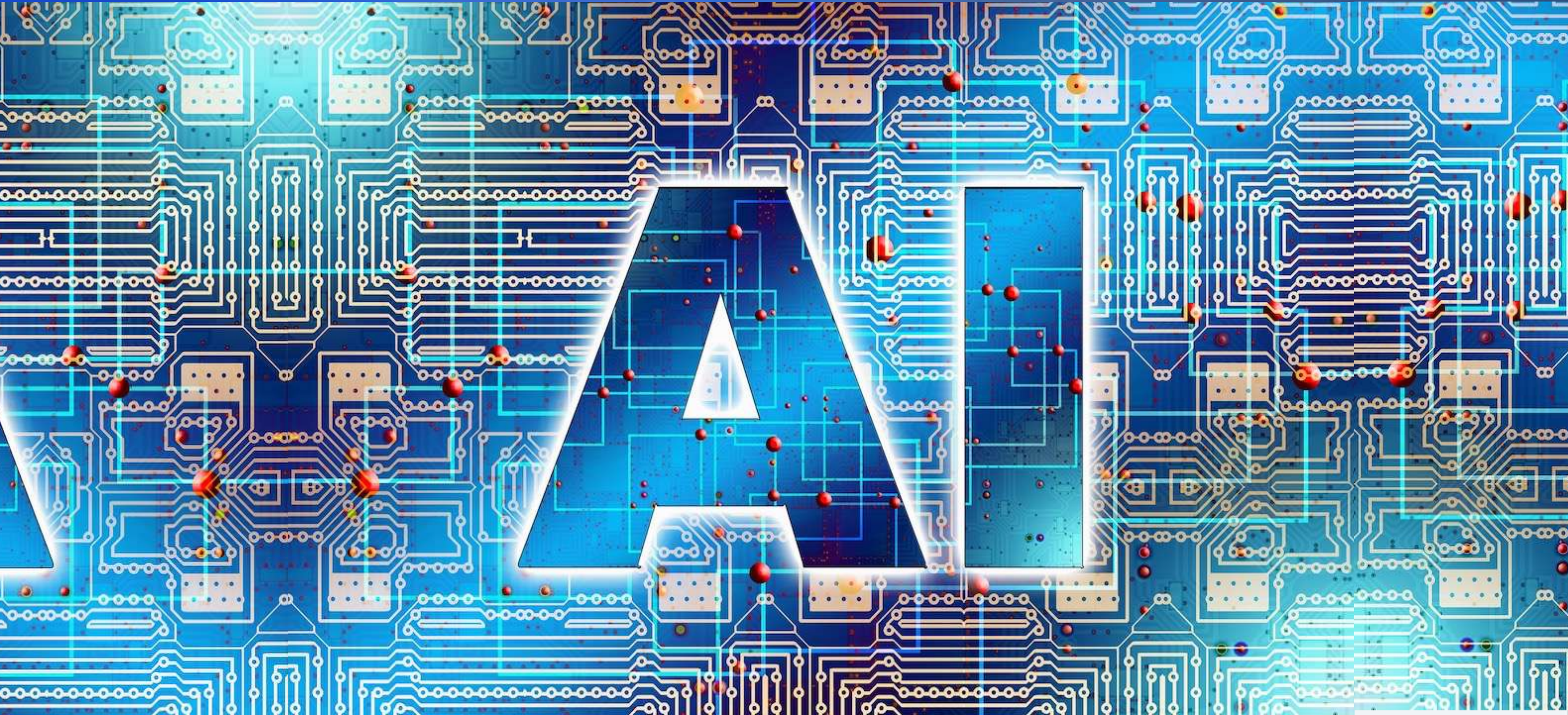
Introduction – The Problem

Our Client, SPACEX

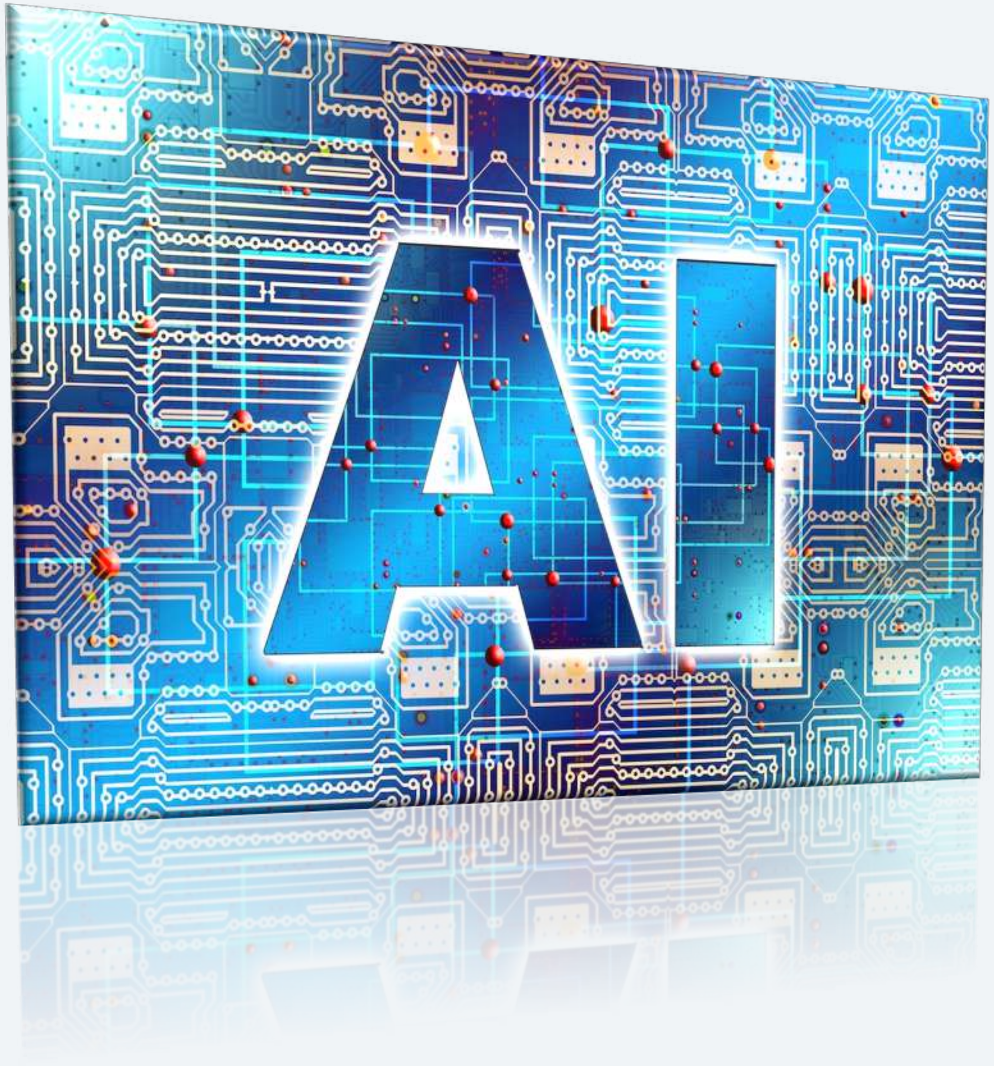
Needs to UNDERSTAND how SpaceX manages to land their first stages.

And replicate it, of course

Introduction – The Solution



Introduction – The Solution



We are proposing an AI Model that will allow SpaceY

To understand which are the key aspects of launches and decisions that allow SpaceX to perform first-stage landings that well.

To predict which future launch configurations are optimal and have the biggest chances of landing first-stages.

Section 1

Methodology

Methodology

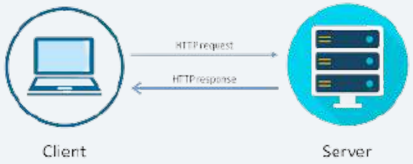
1. Data Collection
2. Data Wrangling
3. Exploratory Data Analysis
4. Data Visualization
5. ML Model Development
6. Insight Reporting

Methodology – Data Collection

Find both notebooks in the appendix section

Data Source A

SpaceX REST A.P.I Calls



GET Request

Jupyter Notebook For Analysis

Data Source B

Wikipedia "List of Falcon 9 and Falcon Heavy Launches" Web Scrapping

Beautiful Soup

Jupyter Notebook For Analysis

Methodology – Data Wrangling

Find Data Wrangling notebook in the appendix section

4 Key Steps were performed

1. Nº launches per site
2. Occurrence of each orbit in launches
3. Occurrence of mission outcome
4. Creation of "landing outcome" label in the dataset

CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

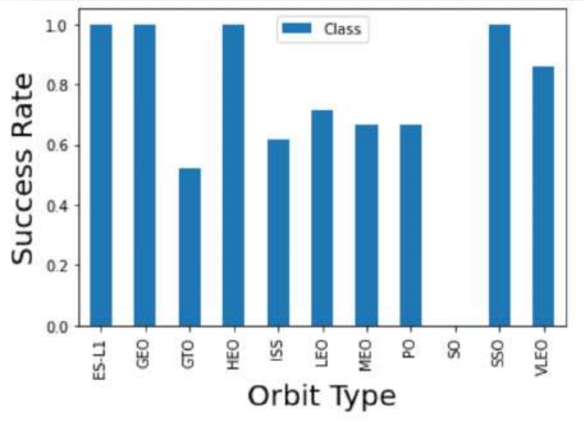
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
ES-L1	1
GEO	1
SO	1
HEO	1

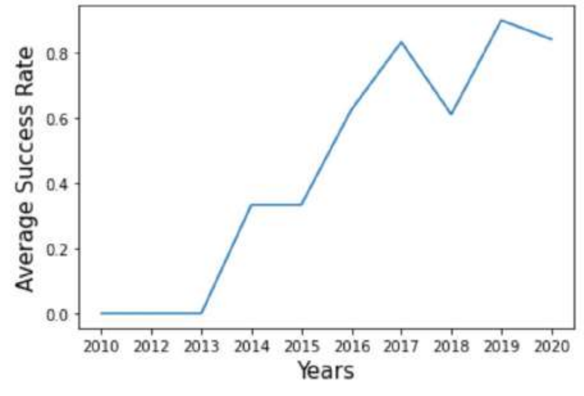
True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
None ASDS	2
False Ocean	2
False RTLS	1

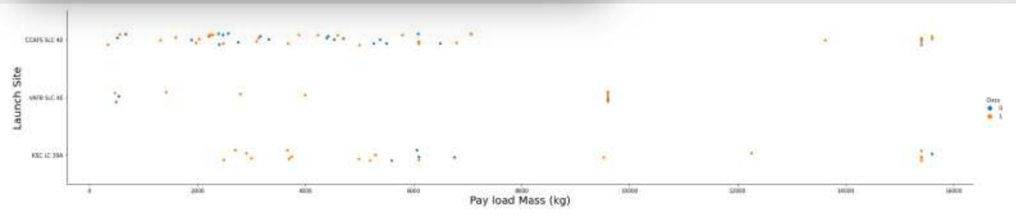
	Outcome	Class
0	None None	0
1	None None	0
2	None None	0
3	False Ocean	0
4	None None	0

Methodology – E.D.A with Data Visualization

Find E.D.A Viz notebook in the appendix section







Matplotlib and Seaborn were used to identify patterns and relationships between the Features and the Mission Outcome:

- Orbit type vs Success Rate
- FlightNumber vs Orbit Type
- Payload vs Orbit Type
- Year vs Success Rate
- Payload vs Launch Site
- FlightNumber vs Launch Site
- FlightNumber vs Payload

Methodology – E.D.A with SQL

Find EDA SQL notebook in the appendix section

missionoutcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

```
%%sql
SELECT sum (payload_mass__kg_) FROM SPACEXTBL as db
WHERE db.customer LIKE ('NASA (CRS)')
LIMIT 5;
```

```
%%sql SELECT booster_version,payload_mass__kg_ FROM SPACEXTBL
WHERE payload_mass__kg_ =  (SELECT max (payload_mass__kg_) from SPACEXTBL);
```

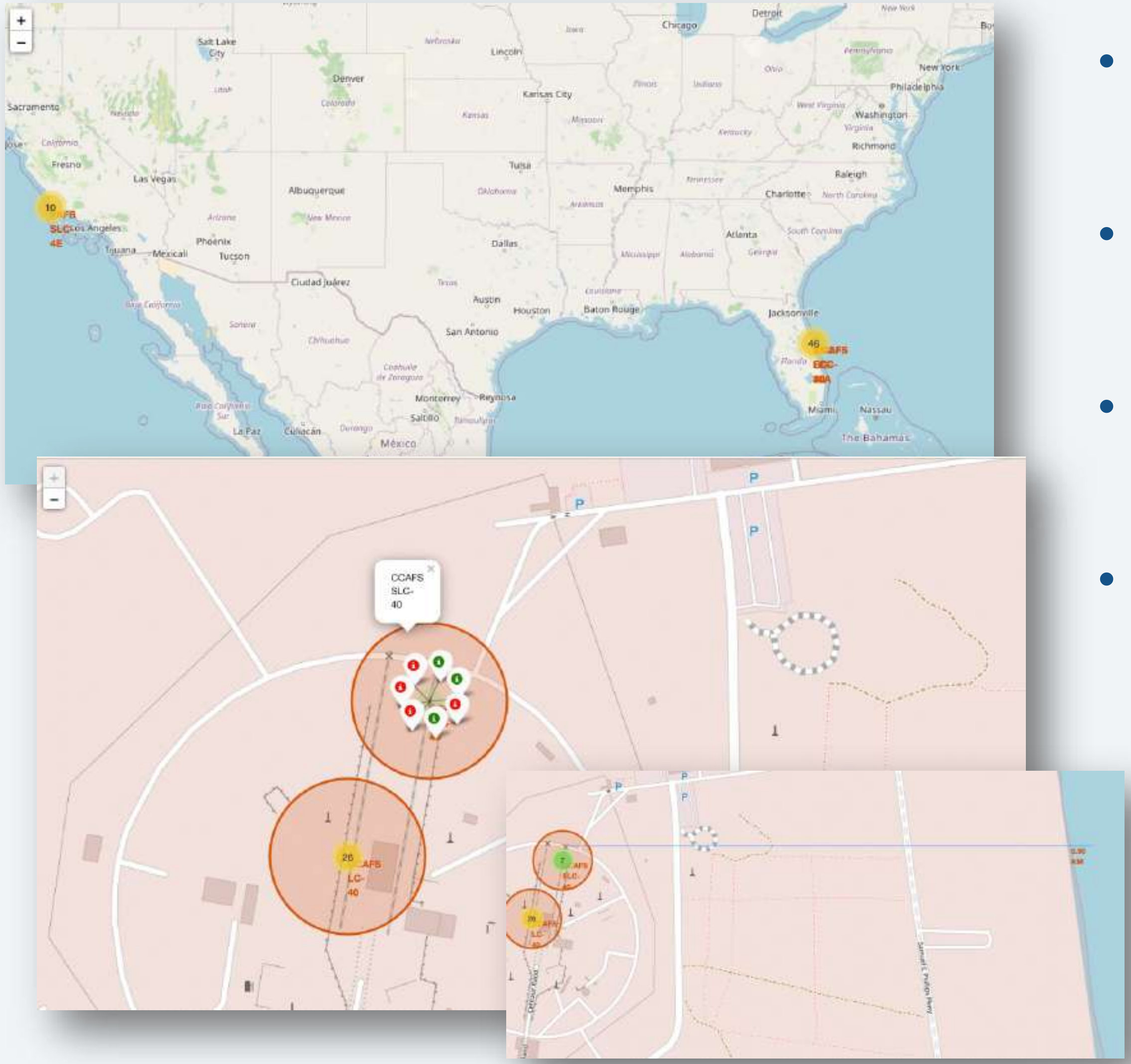
```
%%sql SELECT booster_version, landing__outcome,payload_mass__kg_ FROM SPACEXTBL
WHERE landing__outcome LIKE ('Success (drone ship)') and payload_mass__kg_ between 4000 and 6000;
```

An instance of an IBM DB2 database was created to run SQL queries on. Some of the exploration was oriented to:

- Launch Sites
- Payload Masses
- Booster Versions
- Mission Outcome
- Booster landings

Methodology – Folium Interactive Map

Find the Folium notebook in the appendix section



- We used the Folium Library to create an interactive map that shows all SpaceX's launching sites.
- This allowed us and the client to better understand which are the optimal locations to launch.
- Launching Sites where grouped into clusters for easier visualization and understanding.
- Important metrics such as launch site distance to cities, railways, highways and coastlines where plotted too.

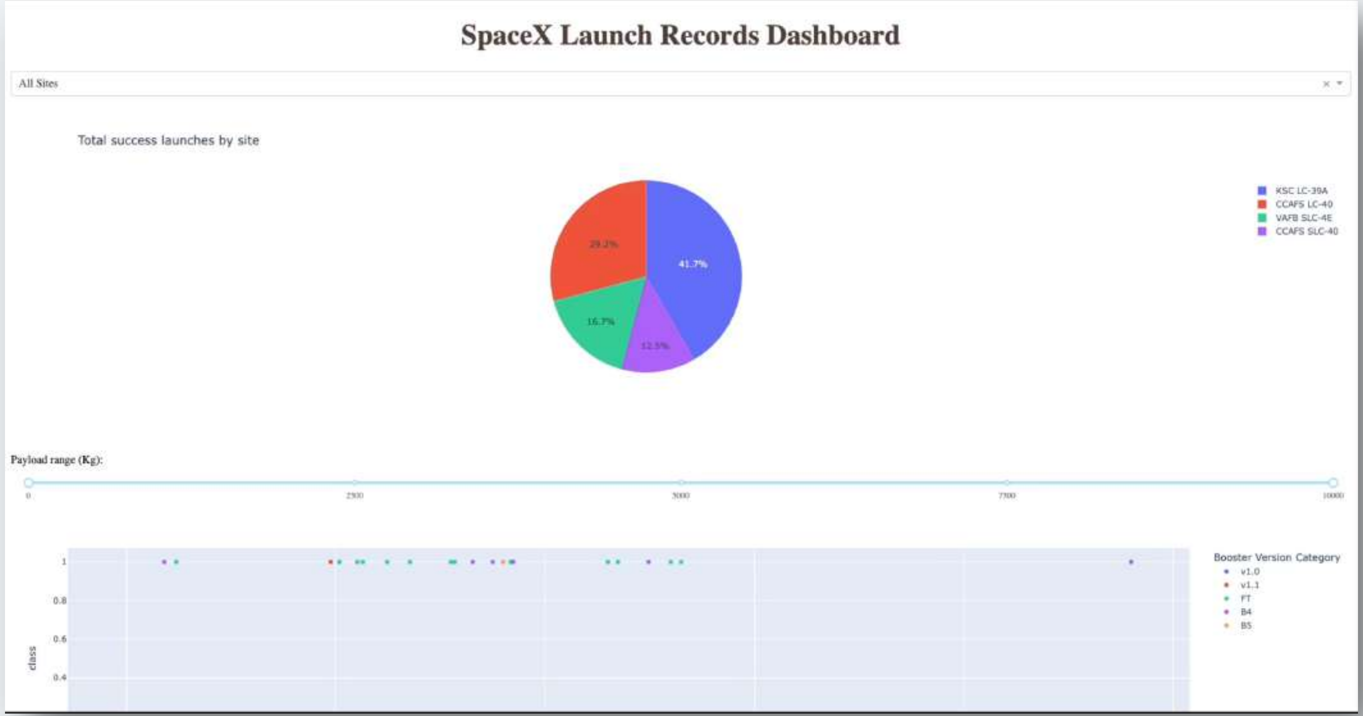
A Green-Red Successful-Unsuccessful color code was implemented to understand the data at a glance.

Methodology – B.I Dashboard

Find the Dash python file in the appendix section

A Dash dashboard was created for the SpaceY team to better understand the data and to continuously keep an eye on it. Using this web app, the team can find information about:

- Success rate for each launch site
- How the payload affects the success rate per launch site



Methodology – ML: Classification

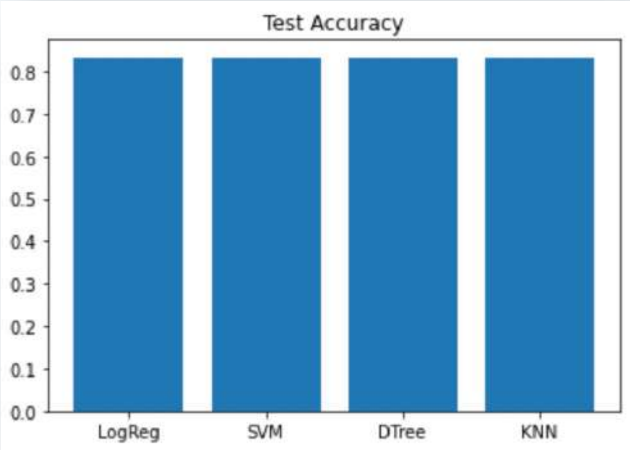
Find the ML notebook in the appendix section

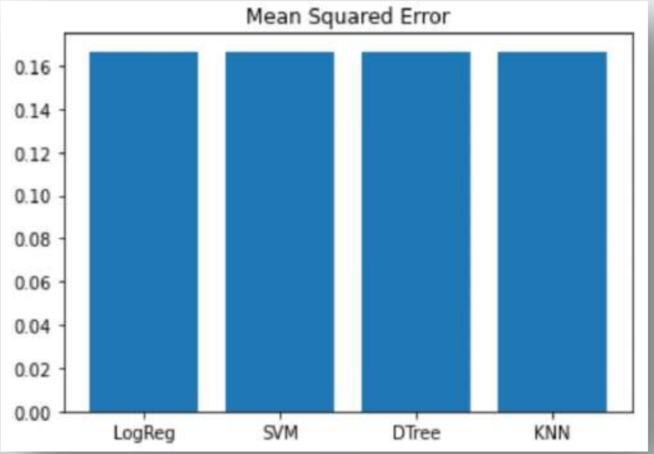
Data was split into training and testing datasets and four classification models were built.

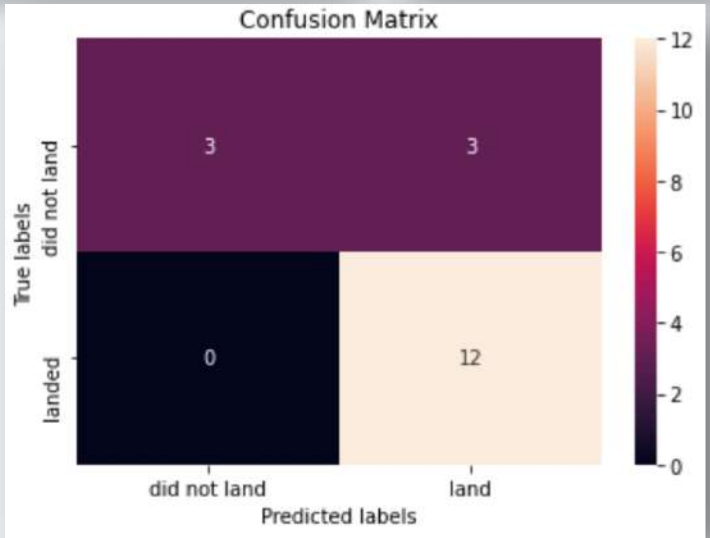
The models proposed were: SVM, Decision Trees, KNN and Logistic Regression. All of them were cross validated using 10 folds and sklearn's GridSearchCV function for hyperparameter tunning.

This four models were evaluated with testing data, calculating their accuracy

Our models can predict the mission outcome of future launches with an accuracy of 84%, only having bad results in false positive cases.



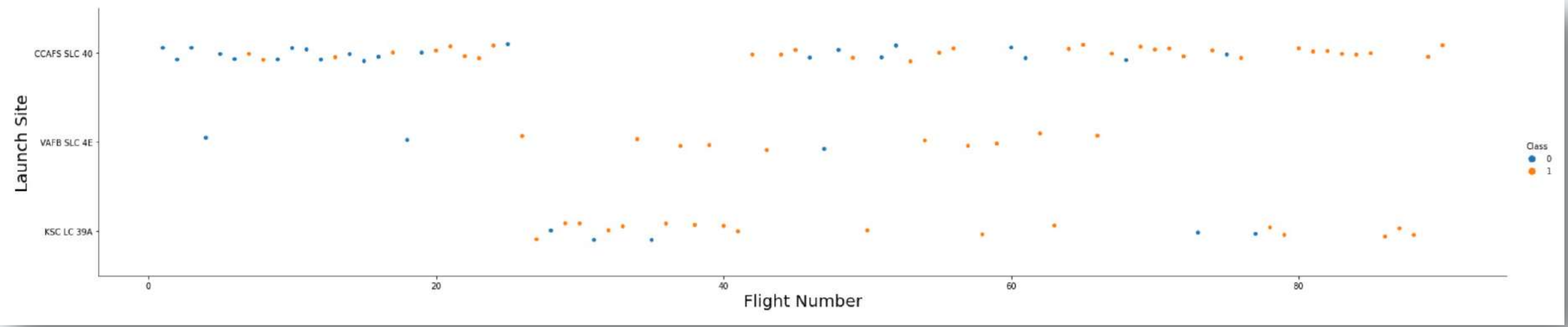




Section 2

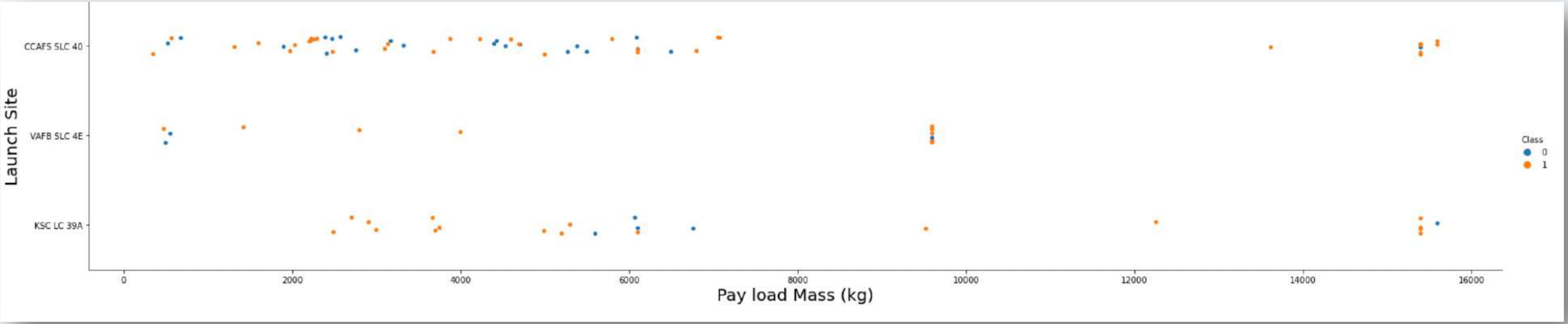
Insights drawn from EDA

Insights – Flight N° vs. Launch Site



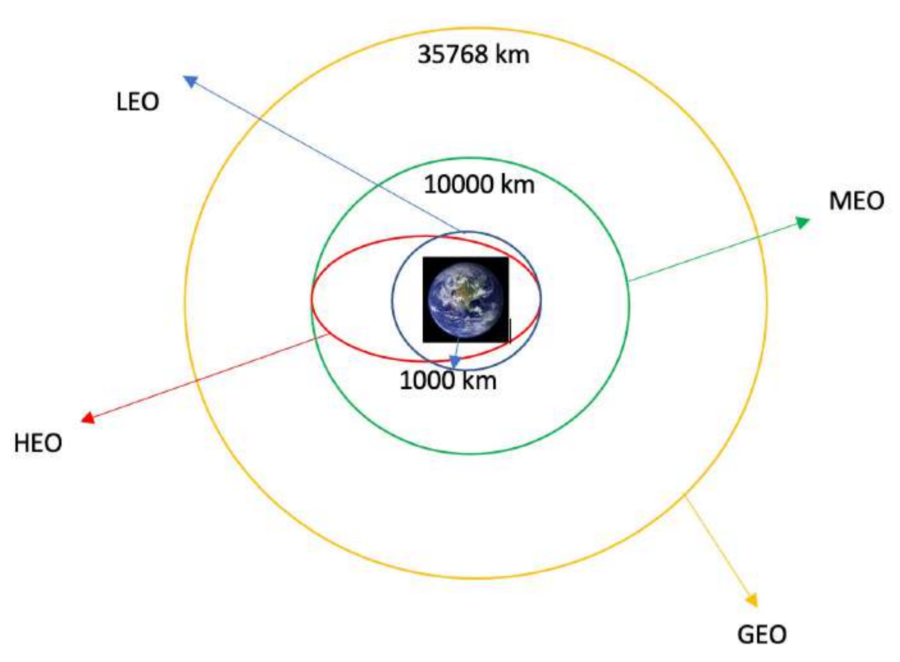
We can see that first launches (most of the launched from CCAFS SLC 40) failed. From launch 25th the success rate goes up for all 3 launch sites. Still having the biggest amount of failures in CCAFS SLC 40. Nevertheless this happens to be the site with more launches and most iterations over a short period of time.

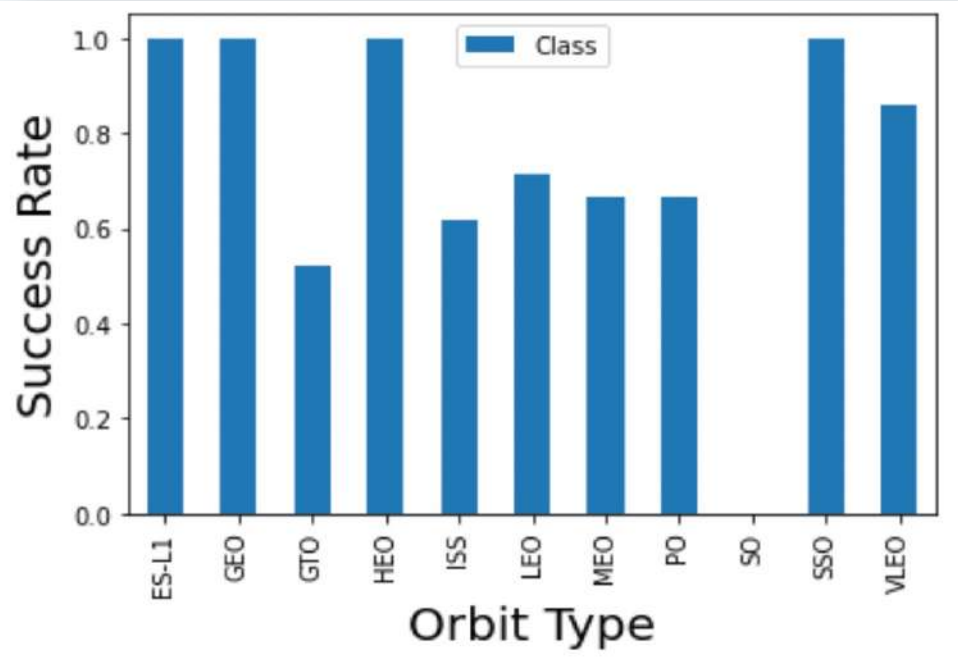
Insights – Payload vs. Launch Site



Most of the heavy (>8000kg.) payloads succeeded at landing. This may suggest that SpaceX first iterated with light payloads and then deployed the heavy ones. This can be seen as well noticing that there were far more light payloads launched than heavy ones.

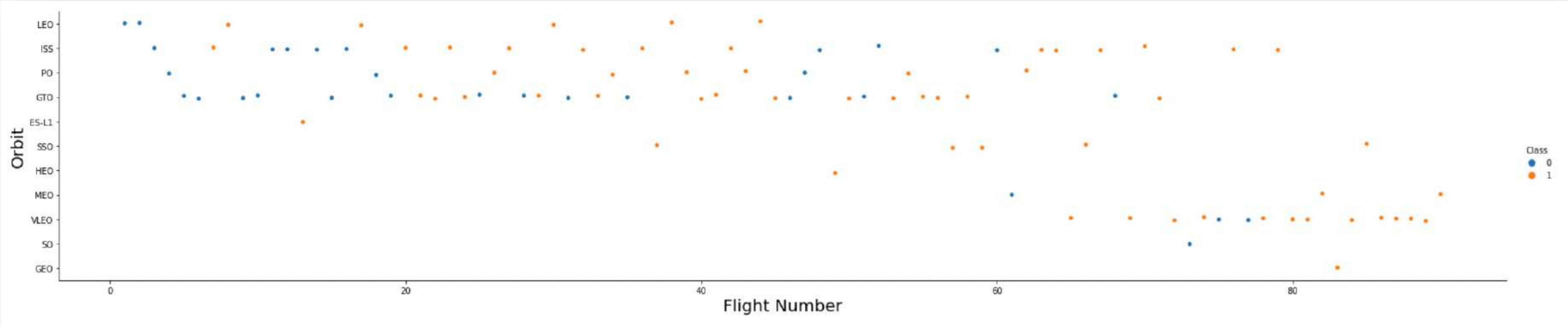
Insights – Success Rate vs. Orbit Type





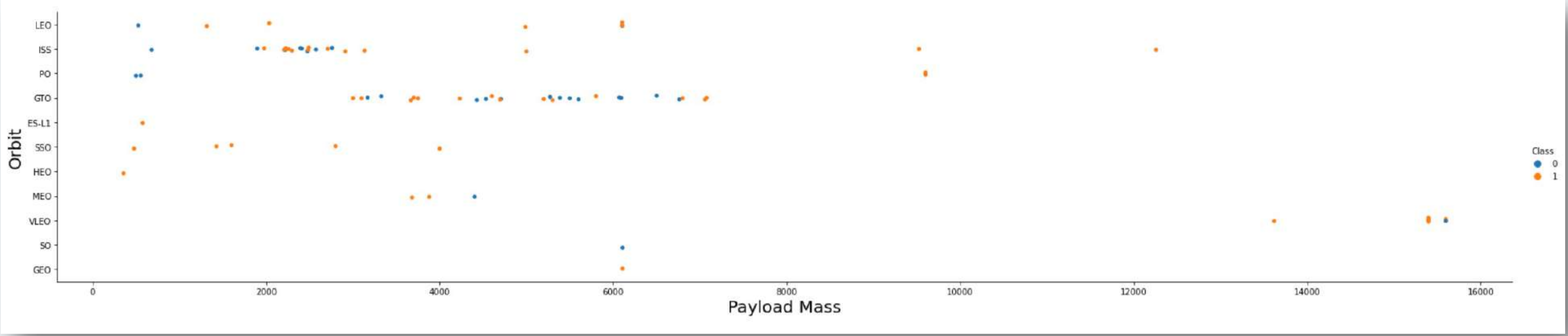
SO orbit was the only one that didn't have any successful launch, the team might want to avoid this orbit until understanding its correlation with landing better. Instead, they might be interested in these other orbits: most successful orbits were ES-L1, GEO, HEO and SSO.

Insights – Flight N° vs. Orbit Type



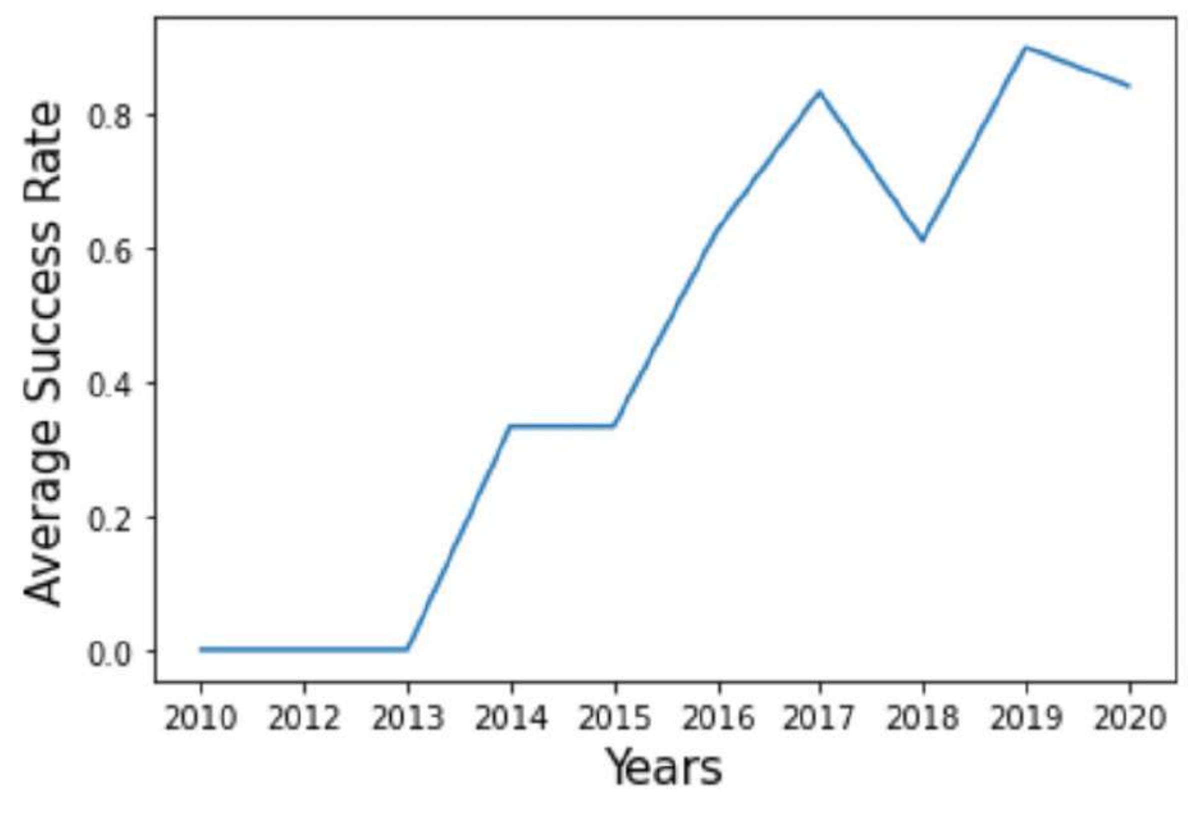
Looking at this scatter plot we can see the trend in SpaceX's orbit choosing.

Insights – Payload vs. Orbit Type



This plot suggests which amount of payload is reasonable to launch to each one of the listed orbits with SpaceX-like rockets.

Insights – Success Rate YoY



SpaceX has seen a huge improvement in success rate YoY and the overall trend is very good.

Insights – Launch Site Names

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

Insights – Launch Site Names that 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

Insights – Total Payload Mass

45596 kg

Insights – Average Payload Mass carried by Booster F9 v1.1

2928 kg

Insights – First Successful Ground Landing Date

December 22nd, 2015

Insights – Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version	landing__outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Insights – Total Number of Successful and Failure Mission Outcomes

missionoutcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Insights – Boosters that carried the maximum payload

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

Insights – 2015 Launch Records

1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

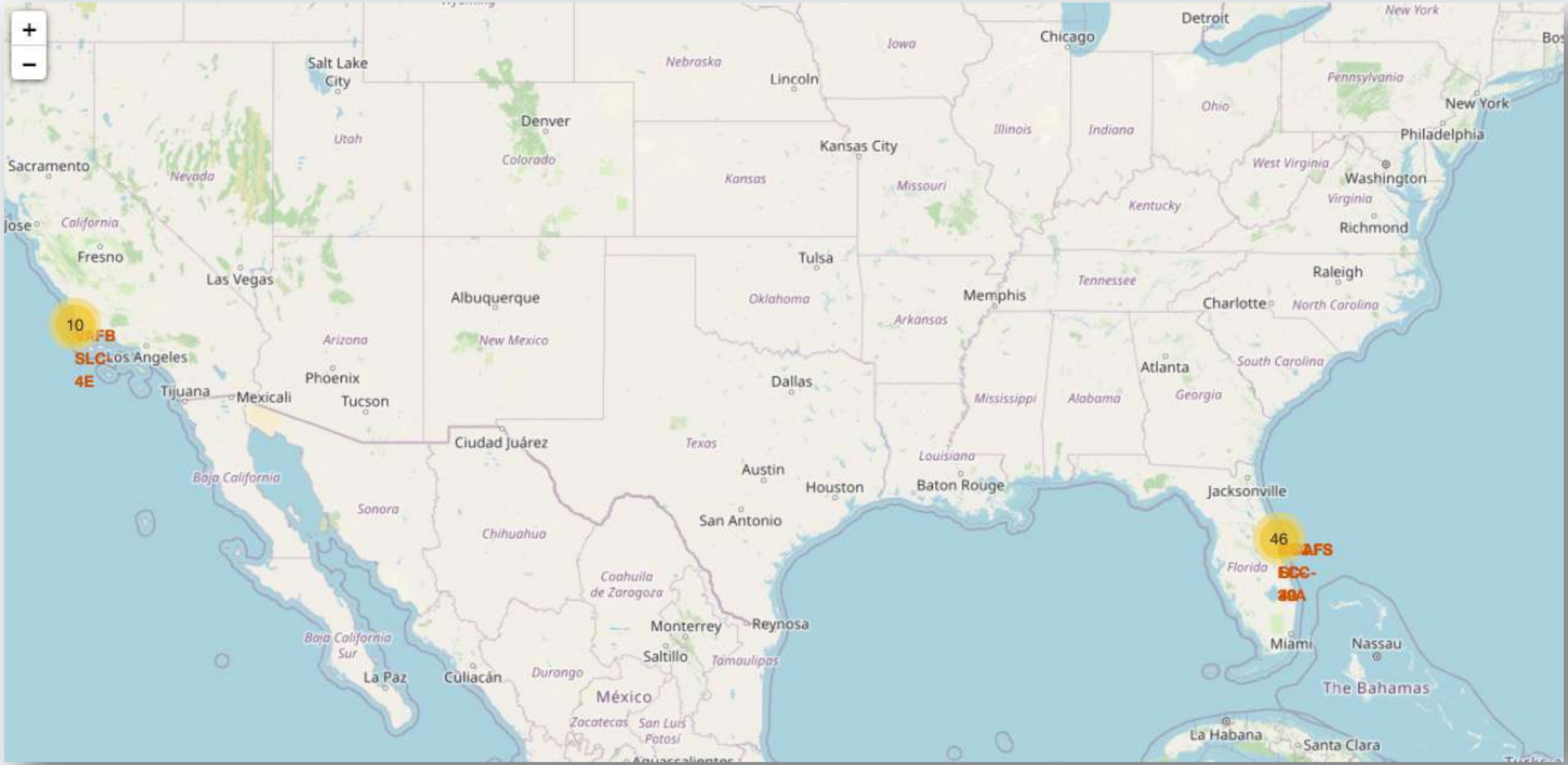
Insights – Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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

Section 4

Launch Sites Proximities Analysis

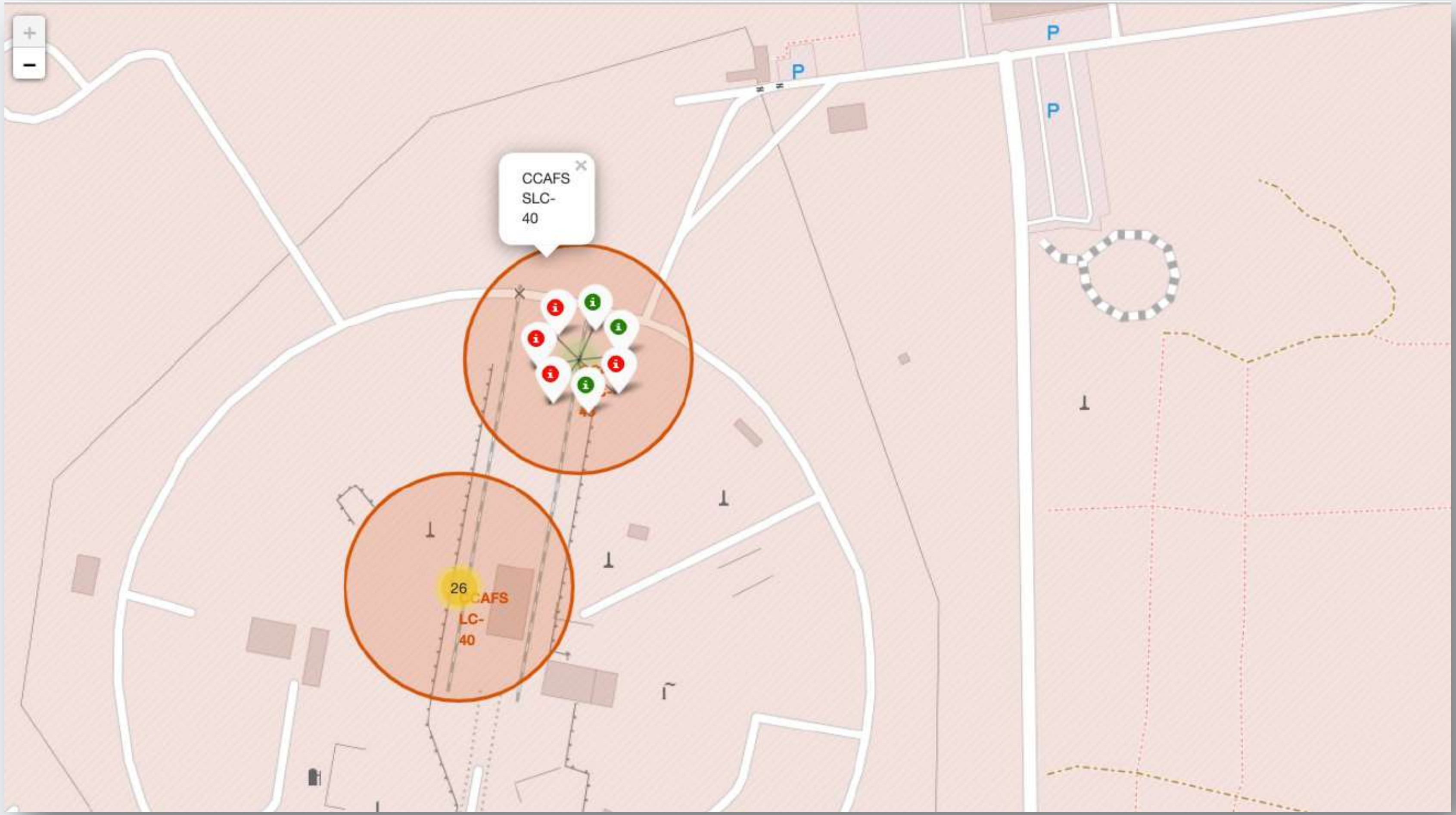
Launch Sites Locations



It is very difficult to understandt geospatial data without a map or visualization.

We can se how the optimal launch sites are close to the ecuator line

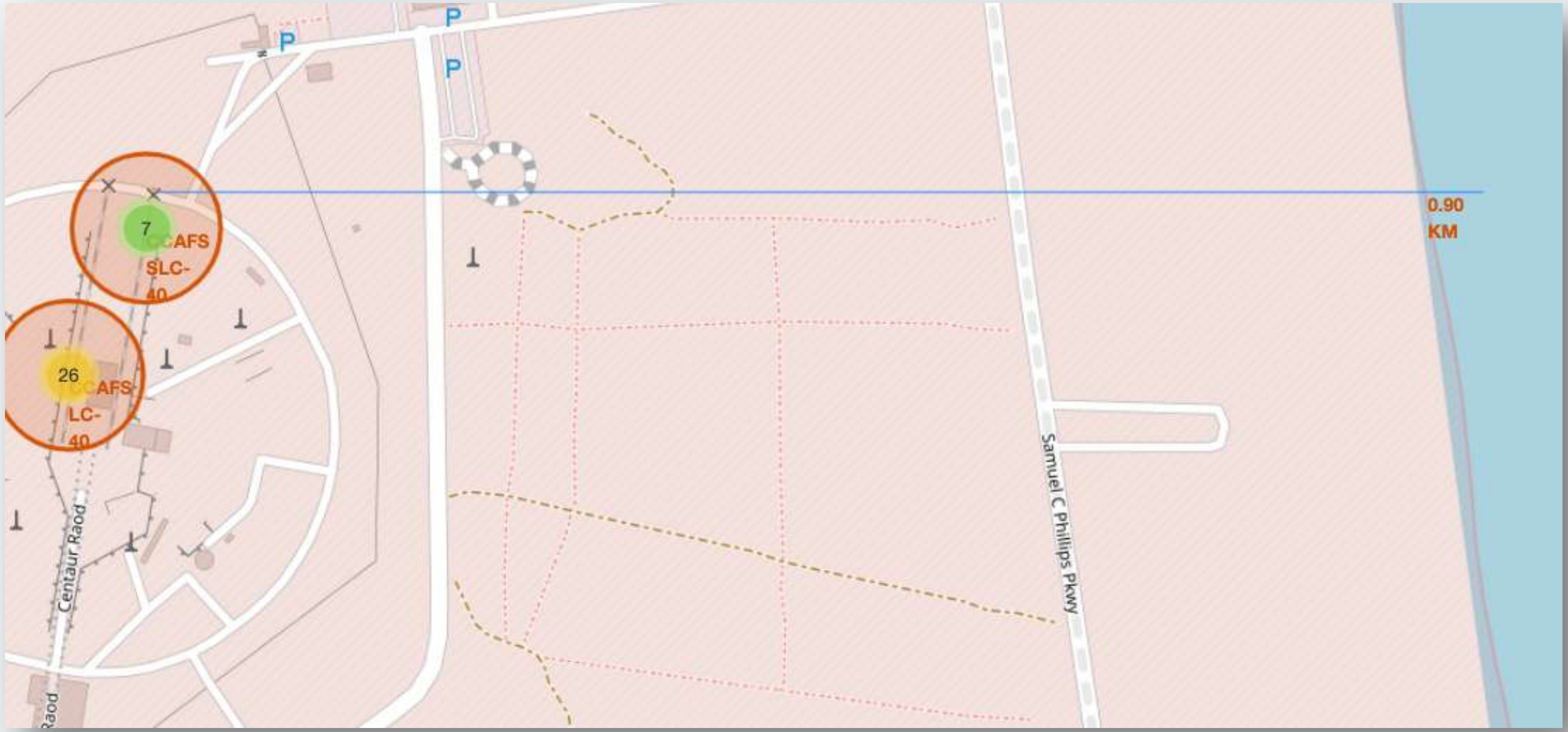
Map Color Code



Color coding the launch sites helps understanding wich places are better for launching

Green = Succesfull
Red = Unsuccesfull

On-map distance measurement



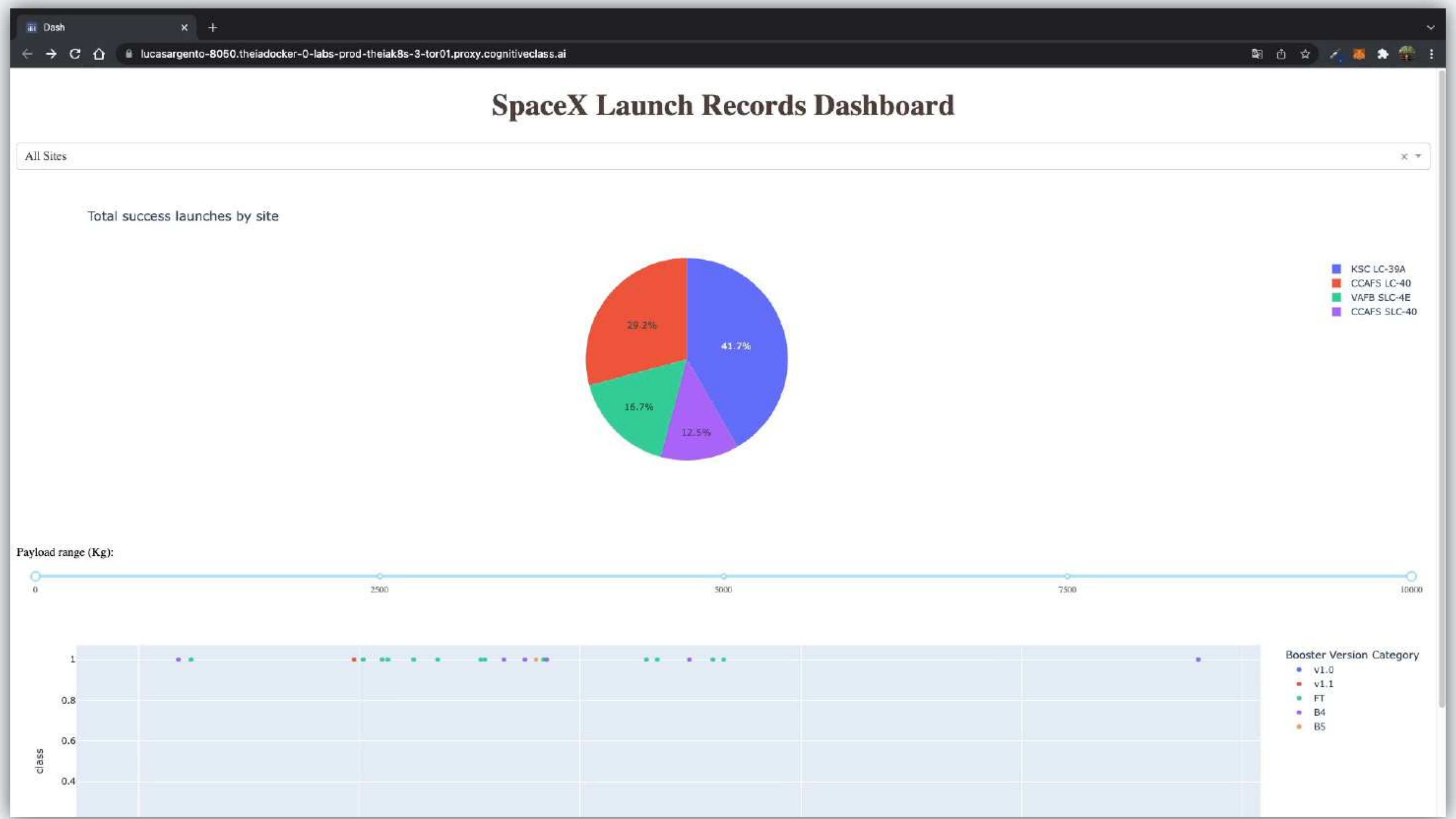
Measuring and plotting distances to relevant geographical places helps understanding where theese launch sites must be situated.

We found out it's a good practice to launch near railways and highways to easily transport heavy cargo. The site must be close to the sea to mitigate any on-flight emergency. Sites must be far away from cities to protect people and buildings.

Section 5

Build a Dashboard with Plotly Dash

B.I Dashboard. Video Proof of Concept.



If you are reading this report on pdf format, the video won't load.

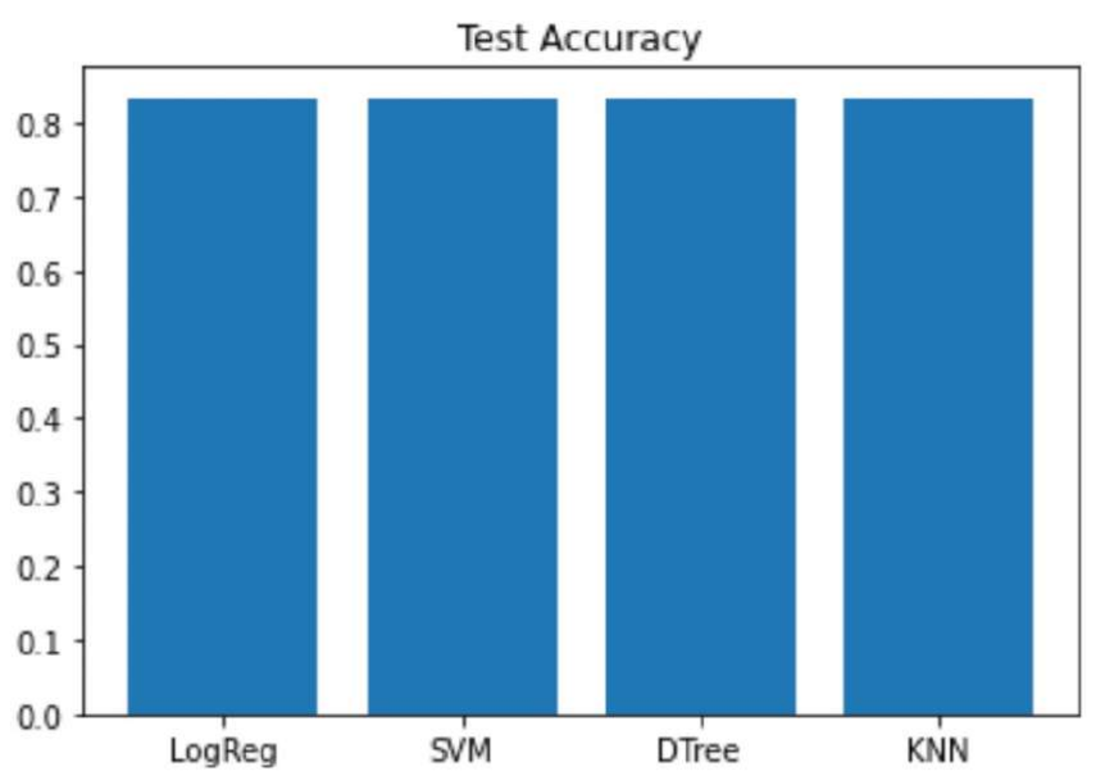
Please head up to the Appendix section and open and run the dash Python file.

Feel free to explore the Dashboard yourself!

Section 6

Predictive Analysis (Classification)

Classification Accuracy



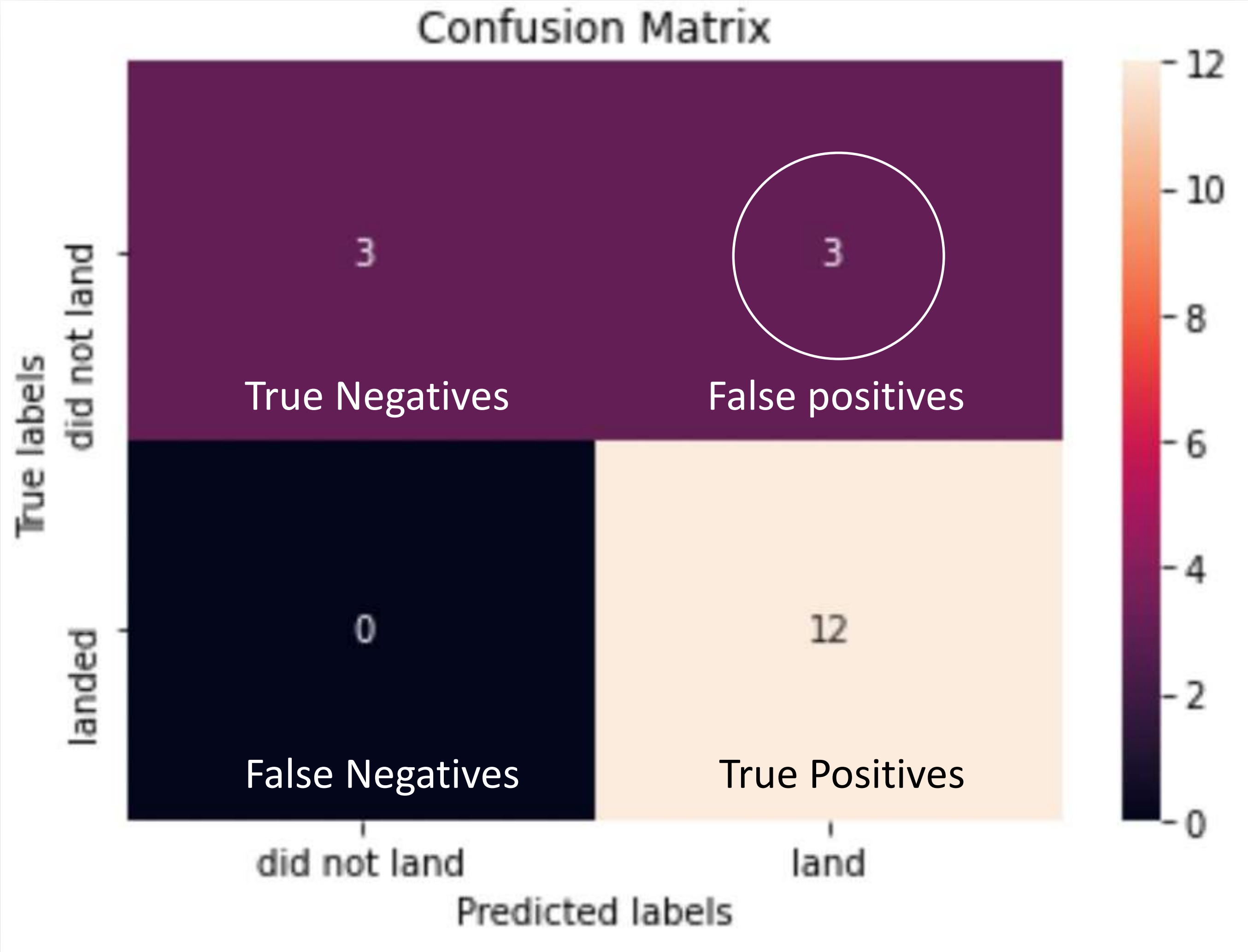
Model accuracy turned out almost the same for all four models.

After some research, we found out that this is caused because of the dataset's size. A bigger size would cause differences in accuracy, standing out which is the best performing model.

Until SpaceY gets even more data, we recommend to continuosly measure model accuracy and decide which one is performing better as the dataset grows.

This implies that SpaceY *needs* to keep collecting data to keep perfectioning the models and predictions.

Confusion Matrix



Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Good things:

- 3 true negatives
- 12 true positives
- 0 false negatives

Bad things:

- 3 false positives (upper right corner)

Conclusions

- SpaceY is now able to understant basic correlations between different launching conditions and a succesfull mission.
- This will allow them to design and define better Space Exploration Missions.
- SpaceY, after the mission-dessign process, can now validate their decisions based on our AI Classification model. They are now able to predict with 83% accuracy if their rocket will succesfully land, saving huge amounts of money.

Appendix

- **Data Collection:**
  - **API Calls notebook:** https://github.com/lucasargento/DataScienceCapstone/blob/master/DataCollection/Data%20Collection%20with%20SpaceX%20API.ipynb
  - **WebScrapping notebook:** https://github.com/lucasargento/DataScienceCapstone/blob/master/DataCollection/Data%20Collection%20with%20Web%20Scraping.ipynb
- **Data Wrangling:**
  - https://github.com/lucasargento/DataScienceCapstone/blob/master/DataWrangling/Data%20Wrangling.ipynb
- **EDA:**
  - **Visualization**: https://github.com/lucasargento/DataScienceCapstone/blob/master/EDA/EDA%20Dataviz.ipynb
  - **SQL**: https://github.com/lucasargento/DataScienceCapstone/blob/master/EDA/EDA%20with%20SQL.ipynb

Appendix

- **Data Visualization:**
  - **Folium notebook:** https://github.com/lucasargento/DataScienceCapstone/blob/master/Visualization/launch_site_location.ipynb
  - **Dash python file:** https://github.com/lucasargento/DataScienceCapstone/blob/master/Visualization/spacex_dash_app.py
- **Data Modelling (Classification):**
  - https://github.com/lucasargento/DataScienceCapstone/blob/master/ModellingAI/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Thank you!