

Conquering space through Data Science

Paolo Fania
19.10.2012

OUTLINE

Executive Summary

Introduction

Methodology

Results

Discussion

Conclusion

Appendix



EXECUTIVE SUMMARY

- The following methodologies were used to analyze data:

- Data collection using web scraping and the SpaceX API;
- Exploratory data analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
- Different machine learning algorithms

- Summary of all the results

- It was possible to collect insightful data from public sources;
- EDA allowed to identify which features were best suited to conclude if a launch would be successful;
- Machine learning showed which model (and which characteristics) are the most important to predict successful launchings and what can be improved to ensure a higher number of successes.

INTRODUCTION

- The objective is to evaluate if the Falcon 9 first stage will land successfully
- Desirable answers:
 - The best way to estimate (and hopefully optimize) the total costs for launches, by predicting successful landings of the first stage of rockets;
 - Where the best place to make launches is located.

Part 1

Methodology



METHODOLOGY - 1

Executive Summary

- Data collection:
 - Data from Space X was obtained from 2 sources:
 - Space X API (<https://api.spacexdata.com/v4/launches/past>)
 - WebScraping from Wikipedia
(https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Data wrangling:
 - Collected data was enriched by creating a landing outcome feature, a.k.a. “training labels”, that will be used by the machine learning algorithms
- Exploratory data analysis (EDA):
 - Data gathered through SQL and displayed in meaningful diagrams through the visualization libraries “matplotlib” and “seaborn”



METHODOLOGY - 2

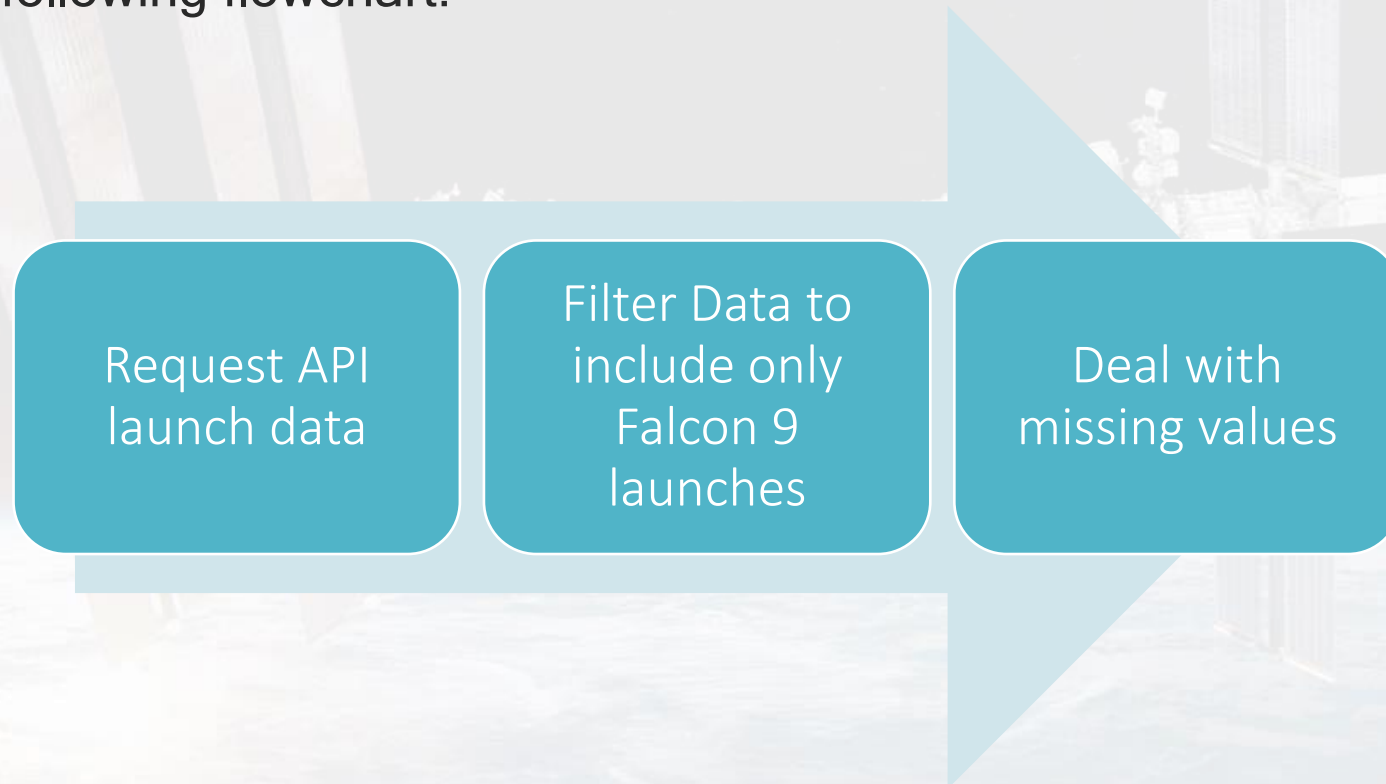
Executive Summary

- Interactive visual analytics:
 - The data was organized into an interactive dashboard using the Dash and Folium
- Predictive analysis:
 - The data collected so far was normalized, divided in training and test data sets and evaluated by four different classification models, with the accuracy of each model being also evaluated. Different combinations of parameters were adopted to improve the overall precision.



METHODOLOGY – DATA COLLECTION 1

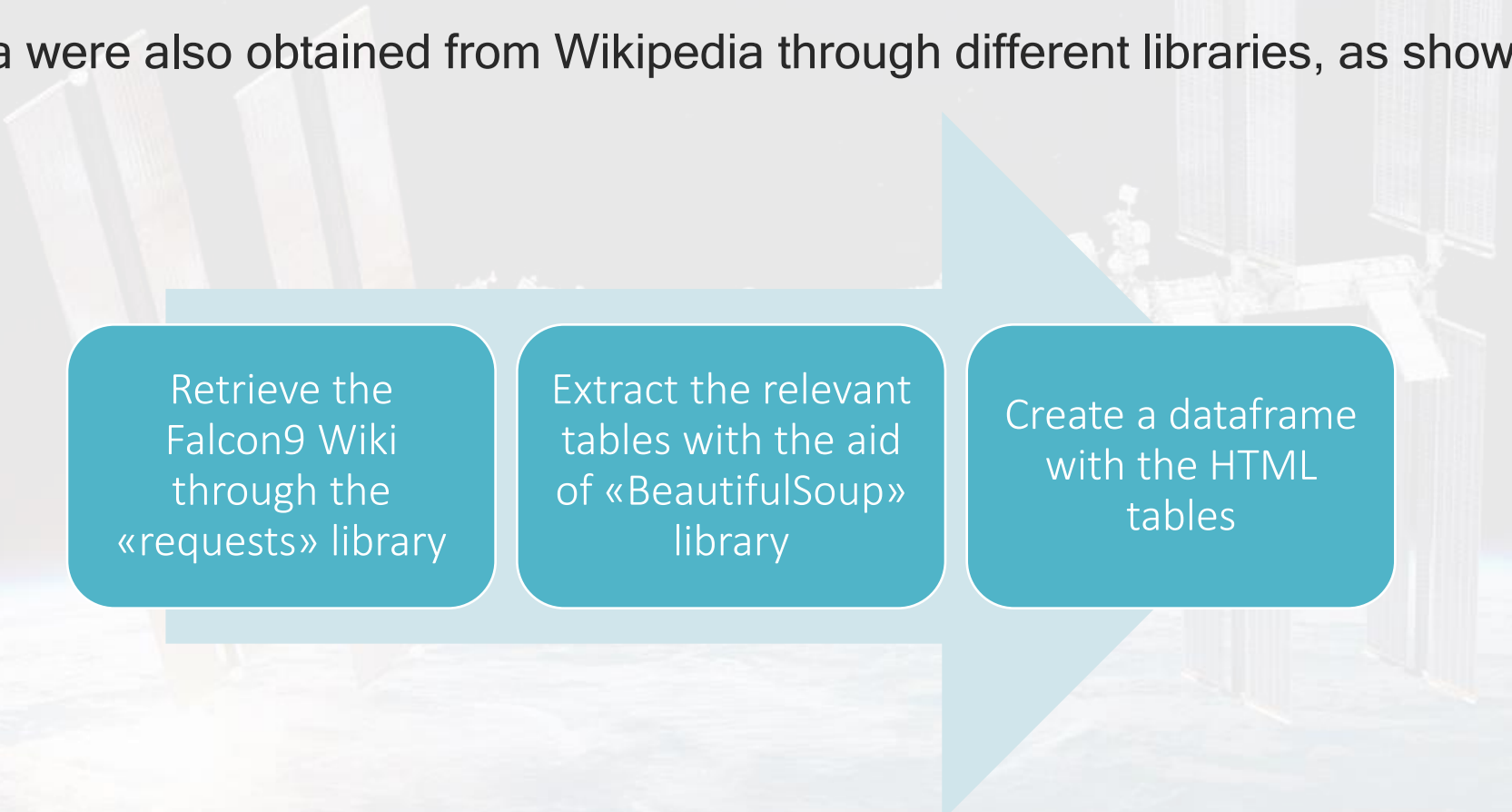
- Space X offers a public API, which contains data on past launches. The data were requested according to the following flowchart:



Source code: https://github.com/paofan86/IBM_Data_Science_Capstone/blob/main/01%20-%20Data%20Collection/jupyter-labs-spacex-data-collection-api-ELABORATED.ipynb

METHODOLOGY – DATA COLLECTION 2

- Space X data were also obtained from Wikipedia through different libraries, as shown below:



Retrieve the
Falcon9 Wiki
through the
«requests» library

Extract the relevant
tables with the aid
of «BeautifulSoup»
library

Create a dataframe
with the HTML
tables

Source code: https://github.com/paofan86/IBM_Data_Science_Capstone/blob/main/02%20-%20Data%20Wrangling/01%20-%20jupyter-labs-webscraping_ELABORATED.ipynb

METHODOLOGY – DATA WRANGLING

- Exploratory Data Analysis (EDA) was initially performed, in order to calculate the number of launches per site, the occurrences of each orbit and their mission outcomes.
- In the end, the landing outcome label was created from the Outcome column of the dataset.

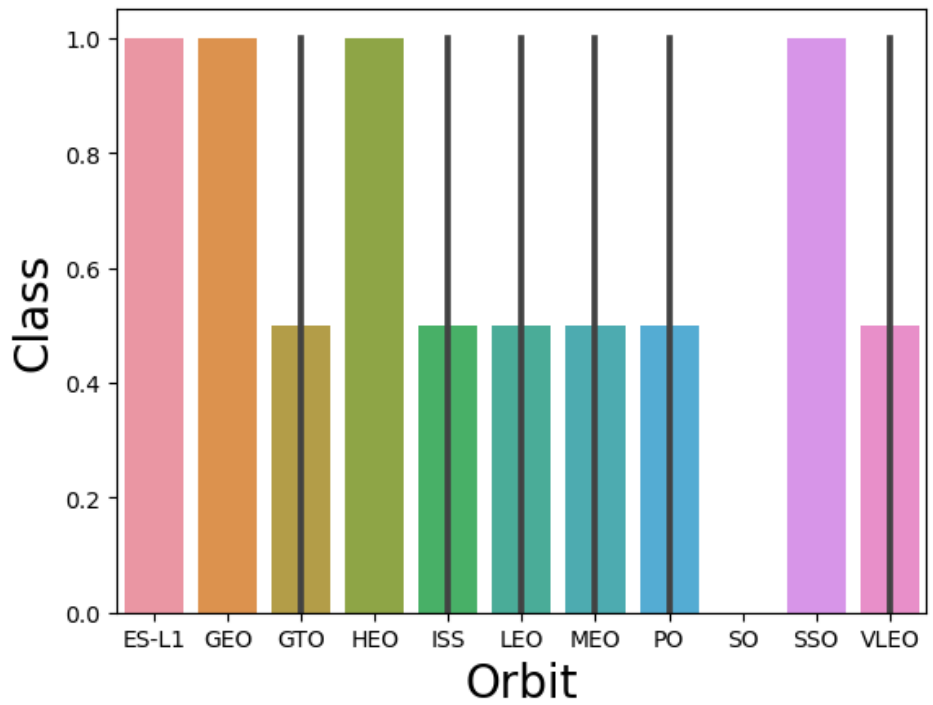


Source code: https://github.com/paofan86/IBM_Data_Science_Capstone/blob/main/02%20-%20Data%20Wrangling/02%20-%20labs-jupyter-spacex-Data%20wrangling_ELABORATED.ipynb

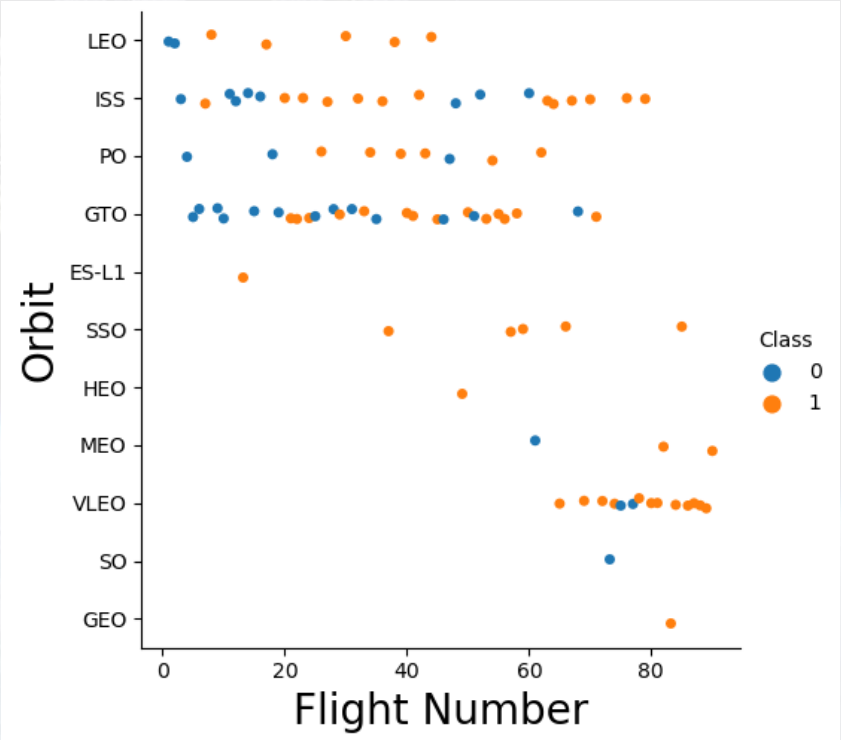
METHODOLOGY – EDA (WITH GRAPHS)

- To explore data, scatter plots and bar plots were used to visualize the relationship between pair of features, such as *Payload Mass / Flight Number*, *Launch Site / Flight Number*, *Launch Site X Payload Mass*, etc.

1) Relationship between success rate and orbit type



2) Relationship between Flight number and orbit type



METHODOLOGY – EDA (WITH SQL)

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites, whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched 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 between 4000 and 6000 kg;
- 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;
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

METHODOLOGY – INTERACTIVE ANALYTICS 1 (FOLIUM)

- Launch Sites Locations were analyzed geographically with the help of the library “Folium”:
 - 1) Launch sites are indicated through markers;
 - 2) Circles are used to highlight areas around specific coordinates, like NASA Johnson Space Center;
 - 3) Marker clusters are utilized to gather groups of events (i.e., launches in a launch site) and minimize the number of dots on the map, making it clearer.
 - 4) Lines are used to calculate distances between two coordinates or with the coastline.

METHODOLOGY – INTERACTIVE ANALYTICS 2 (Dashboard)

- Graphs and plots were used to visualize data according to the following features:
 - Percentage of launches by site
 - Payload range
- This allowed to quickly analyze the relation between payloads and launch sites, helping to identify the best place to successfully perform a launch

METHODOLOGY – PREDICTIVE ANALYSIS

- Four machine learning algorithms were adopted and compared:
 - logistic regression
 - support vector machine
 - decision tree
 - k nearest neighbors.

Data preparation
through
standardization

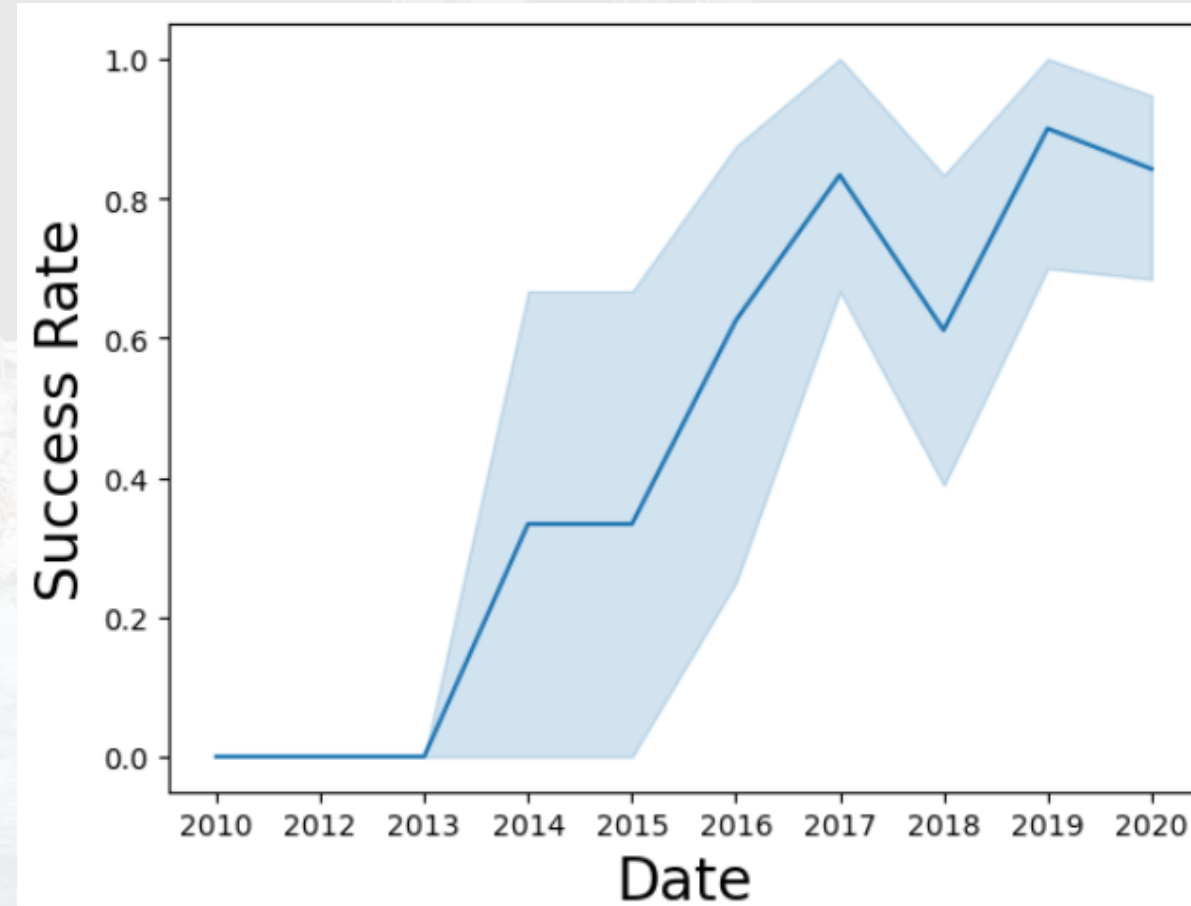
Testing of each
model (fine-tuning of
hyperparameters)

Comparison and
results

RESULTS

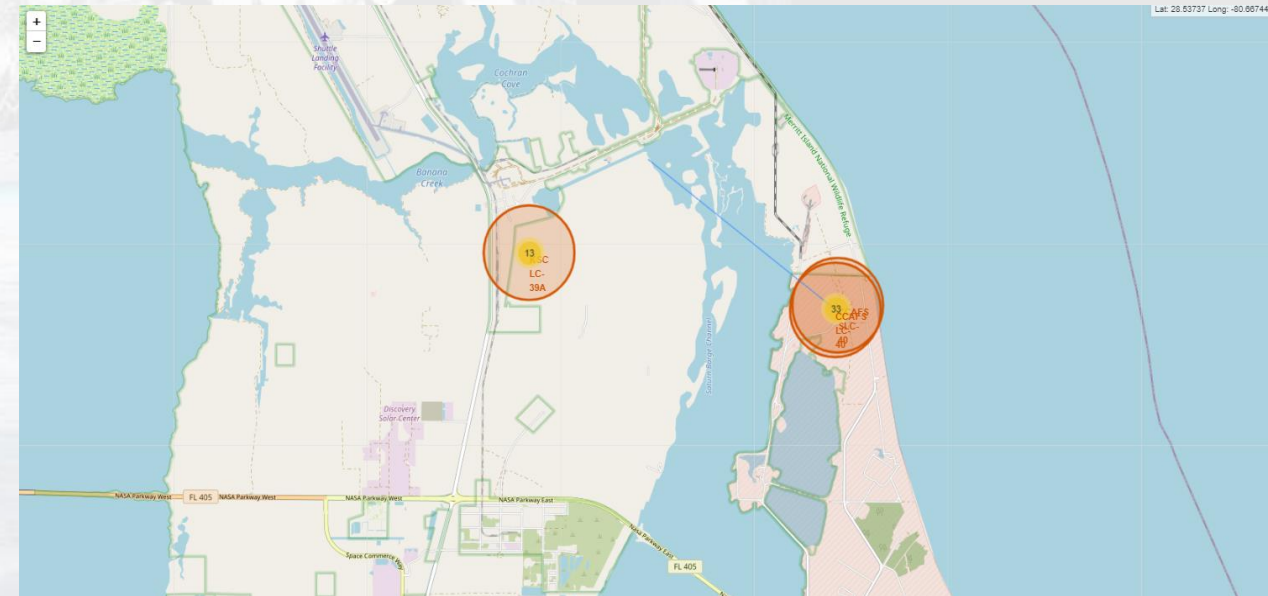
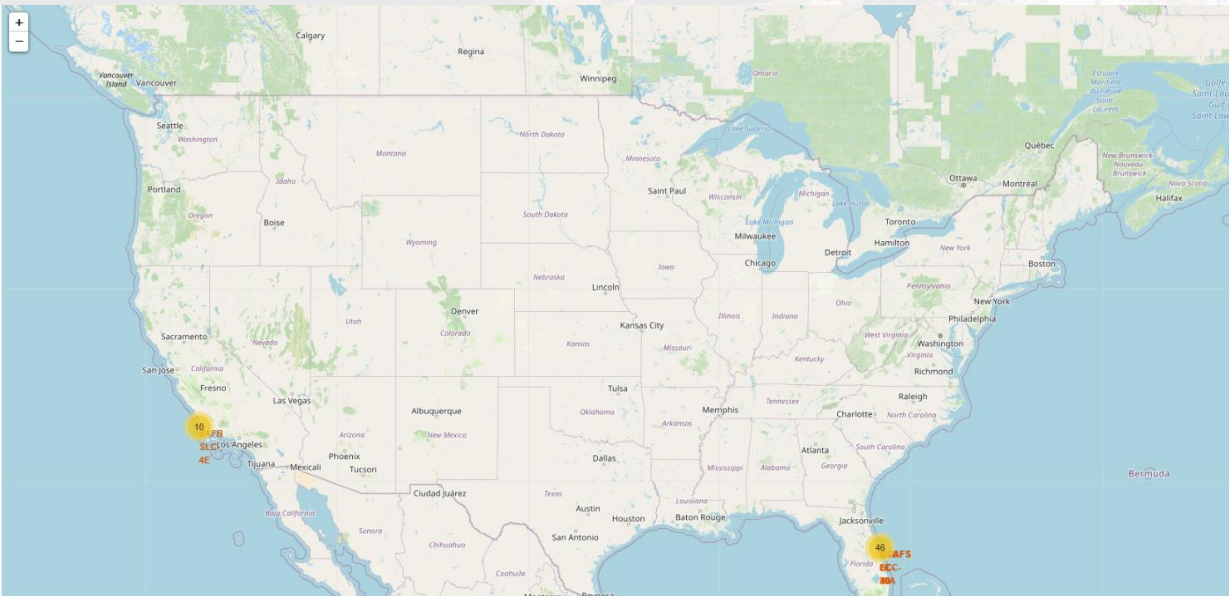
Exploratory data analysis results:

- Space X uses 4 different launch sites;
- The average payload of F9 v1.1 booster is ~6105 kg;
- The first success landing outcome happened in 2014, four years after the first launch (2010);
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- The number of landing outcomes, starting from 2014, became better and better as the years passed.



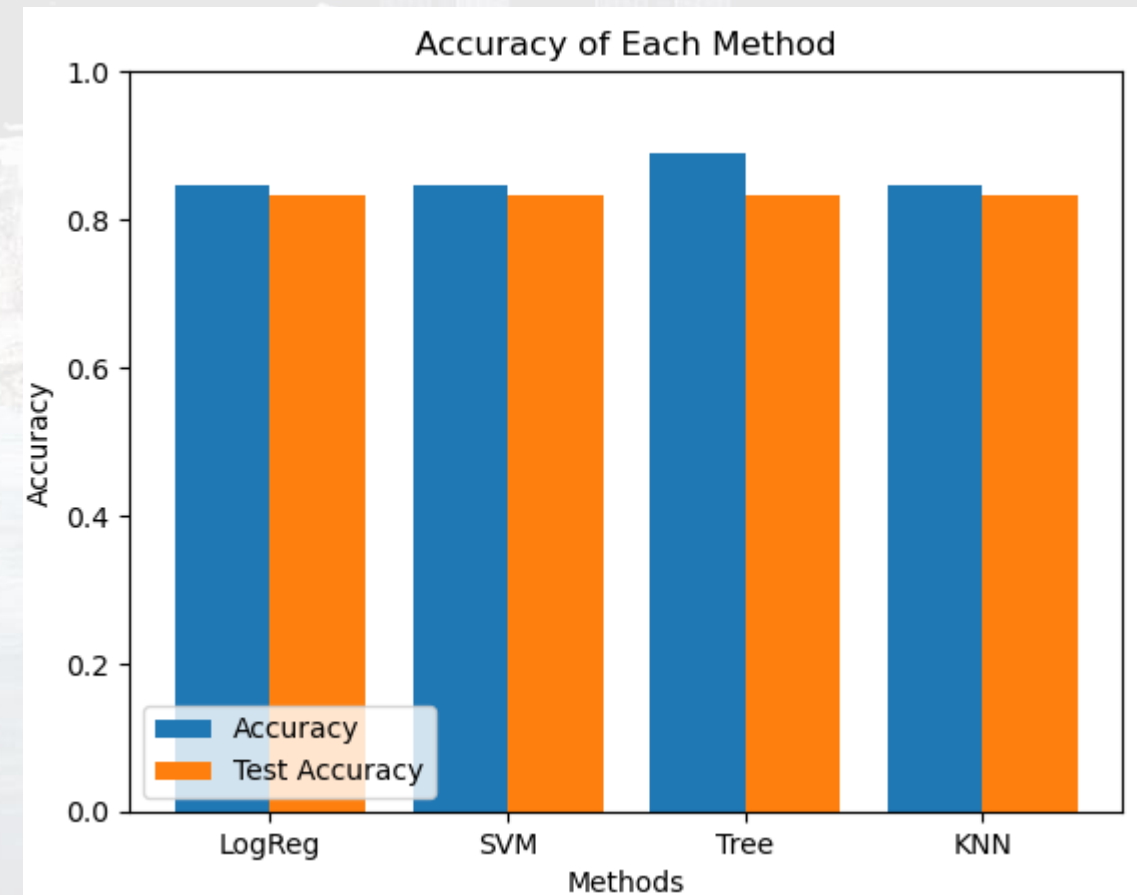
RESULTS

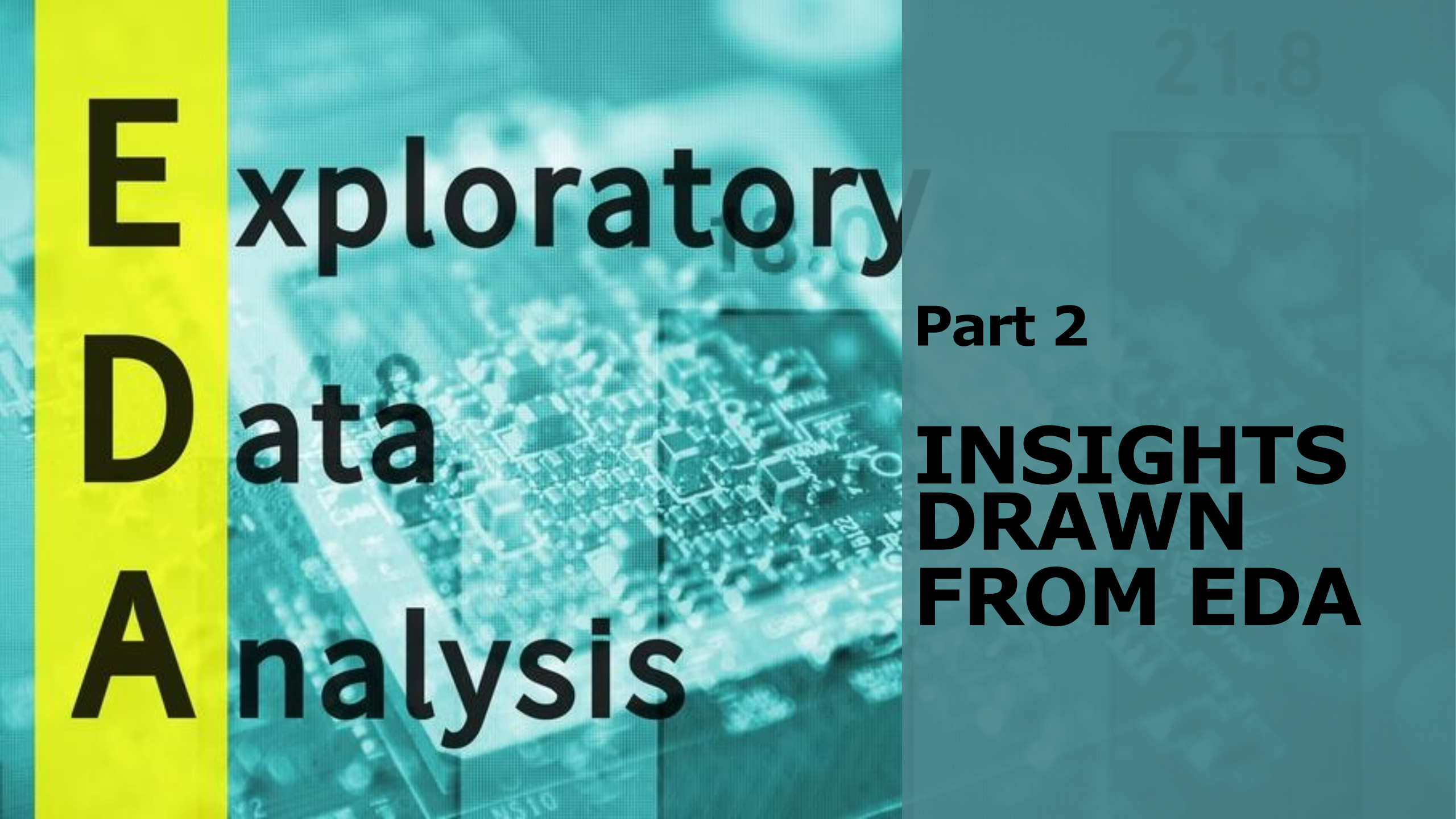
- Using interactive analytics allowed to identify that launch sites are usually:
 - located in isolated and safe places (in case of accidents)
 - near sea (for the need of having a good logistic infrastructure)
 - The closest possible to the equator in order to take optimum advantage of the Earth's substantial rotational speed.



RESULTS

- Predictive Analysis proved that the “decision tree classifier” is the best model to predict successful landings, having an accuracy of 89% and accuracy for test data over 83%



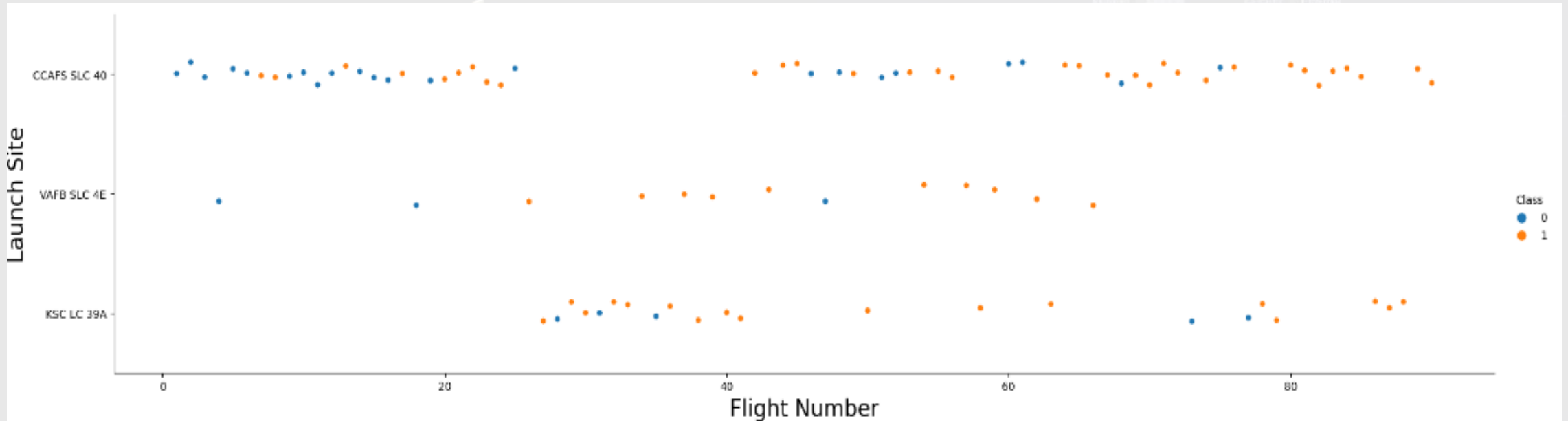


Exploratory **D**ata **A**nalysis

Part 2

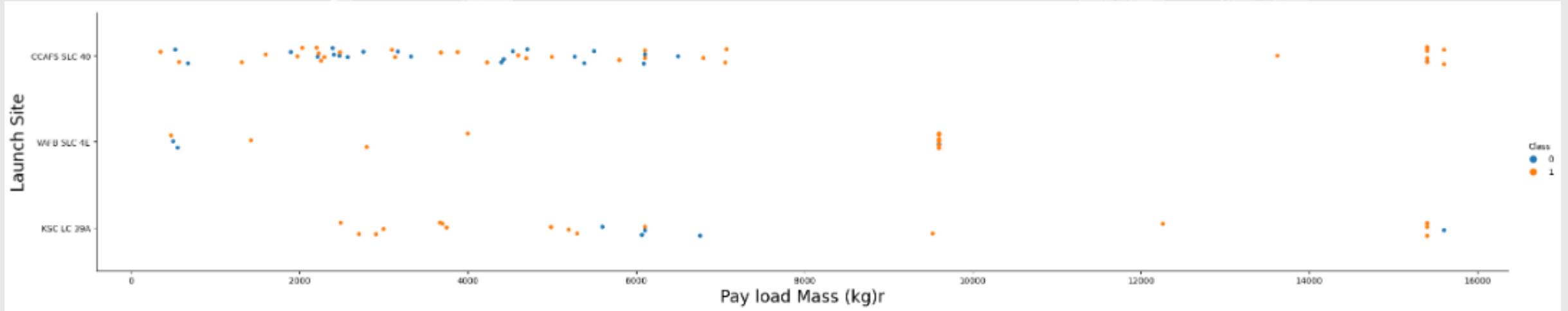
**INSIGHTS
DRAWN
FROM EDA**

Flight Number vs Launch Site



- According to the plot above, it is possible to determine that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful
- It is also worth noticing that the general success rate improved over time.

Payload vs Launch Site

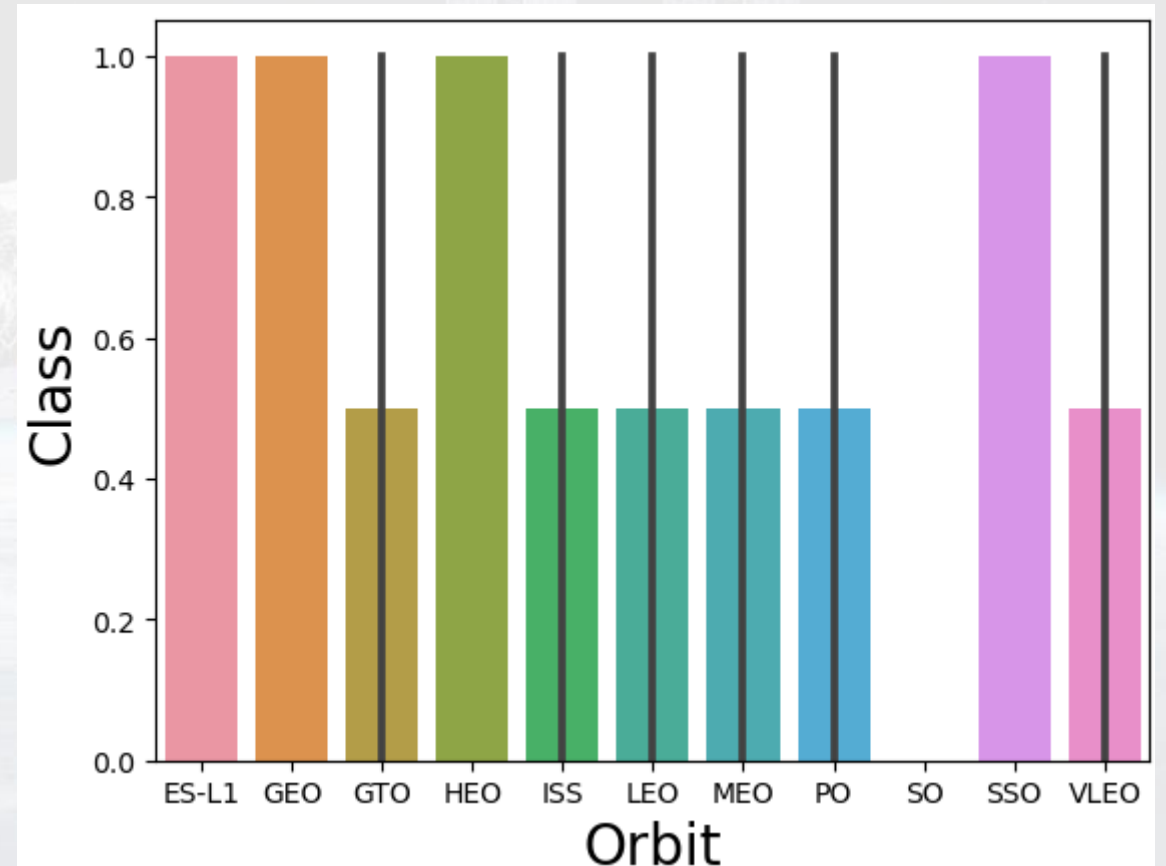


- Launches with payloads over 9 ton have excellent success rate (>90%);
- Launches with payloads over 12 ton seem to be possible only on site CCAFS SLC 40 and KSC LC39.

Success Rate vs Orbit Type

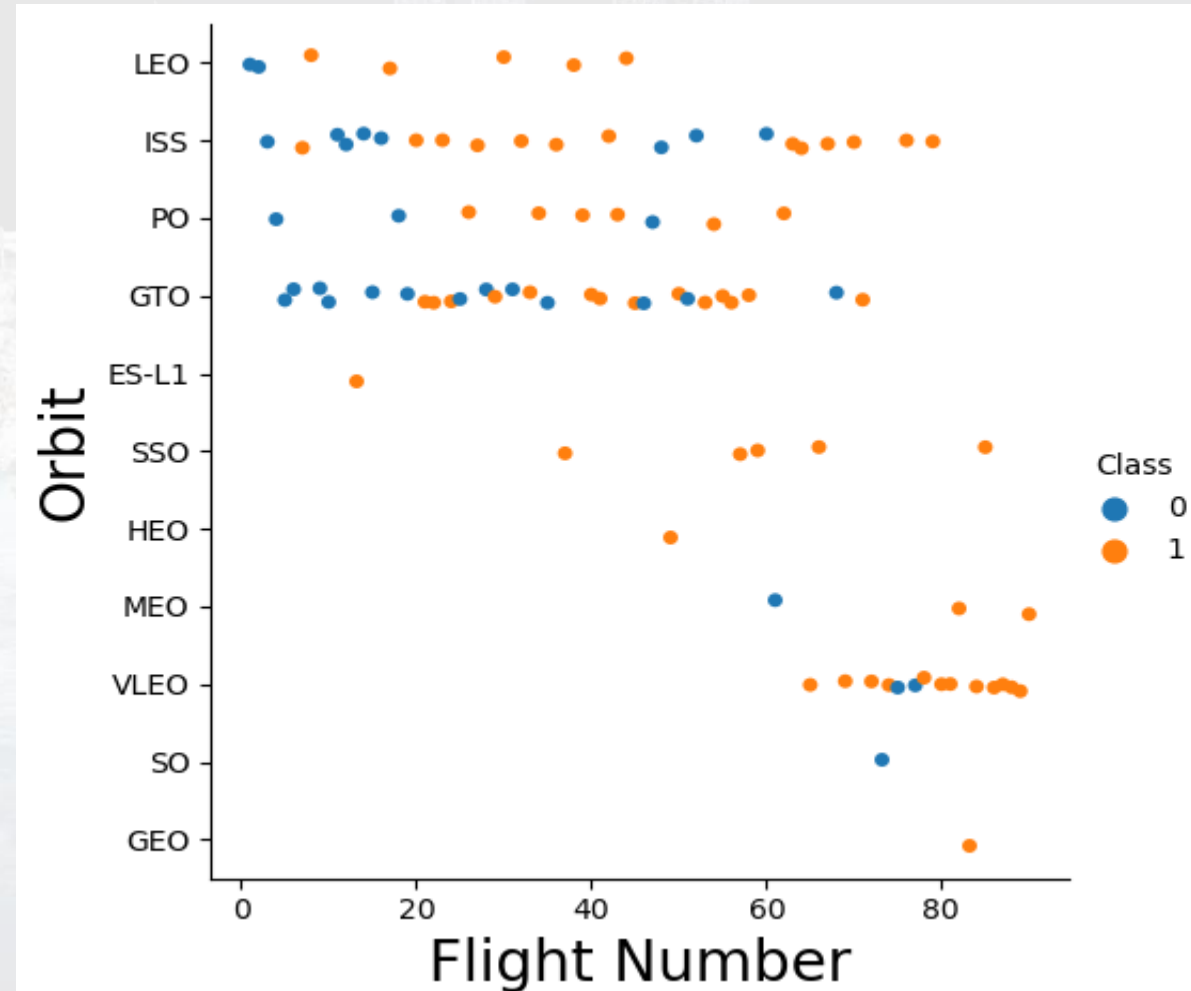
- The biggest success rates happens to orbits:

- ES L1
- GEO
- HEO
- SSO

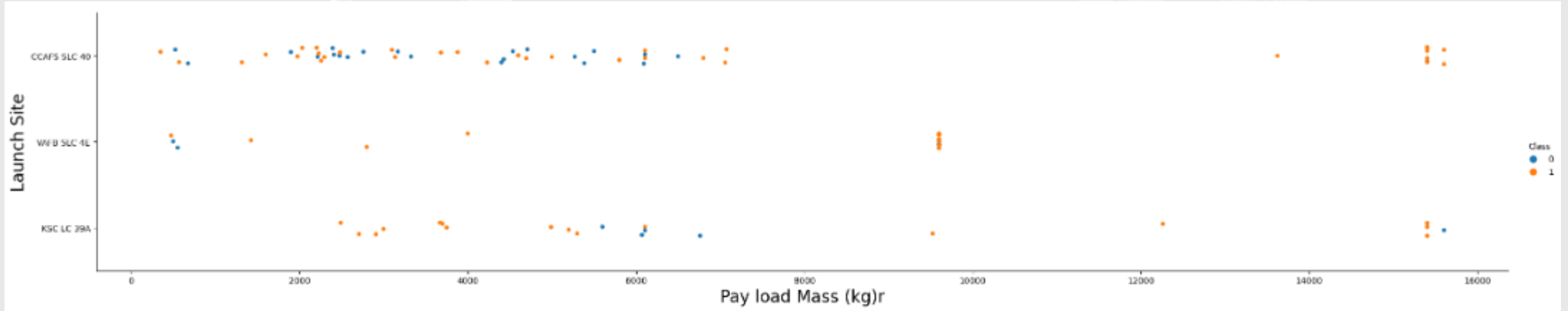


Flight Number vs Orbit Type

- As a tendency, the success rate improved over time for all orbits;
- VLEO orbit had a recent increase of its frequency, meaning this could be a new and more successful orbit.



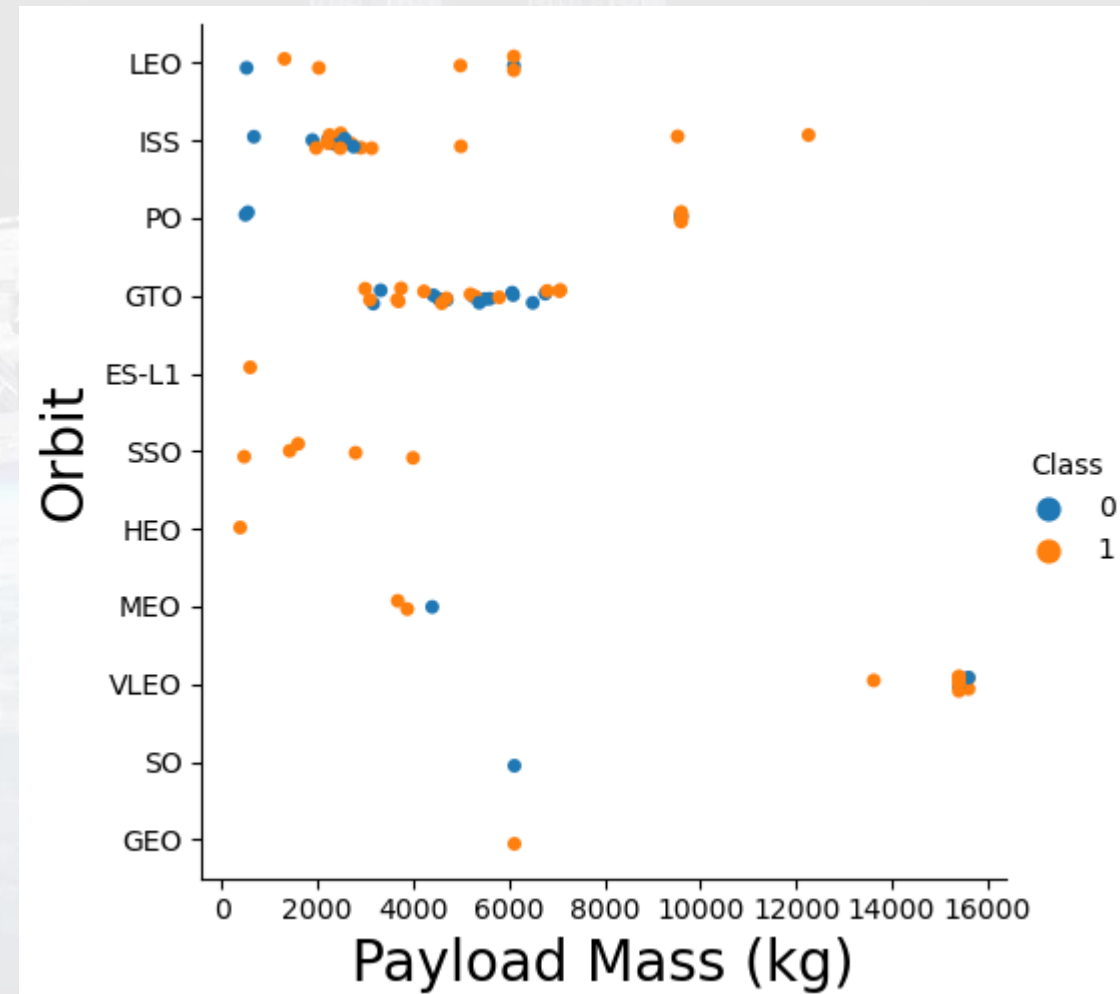
Payload vs Orbit Type



- Launches with payloads over 9 ton have excellent success rate (>90%);
- Launches with payloads over 12 ton seem to be possible only on site CCAFS SLC 40 and KSC LC39.

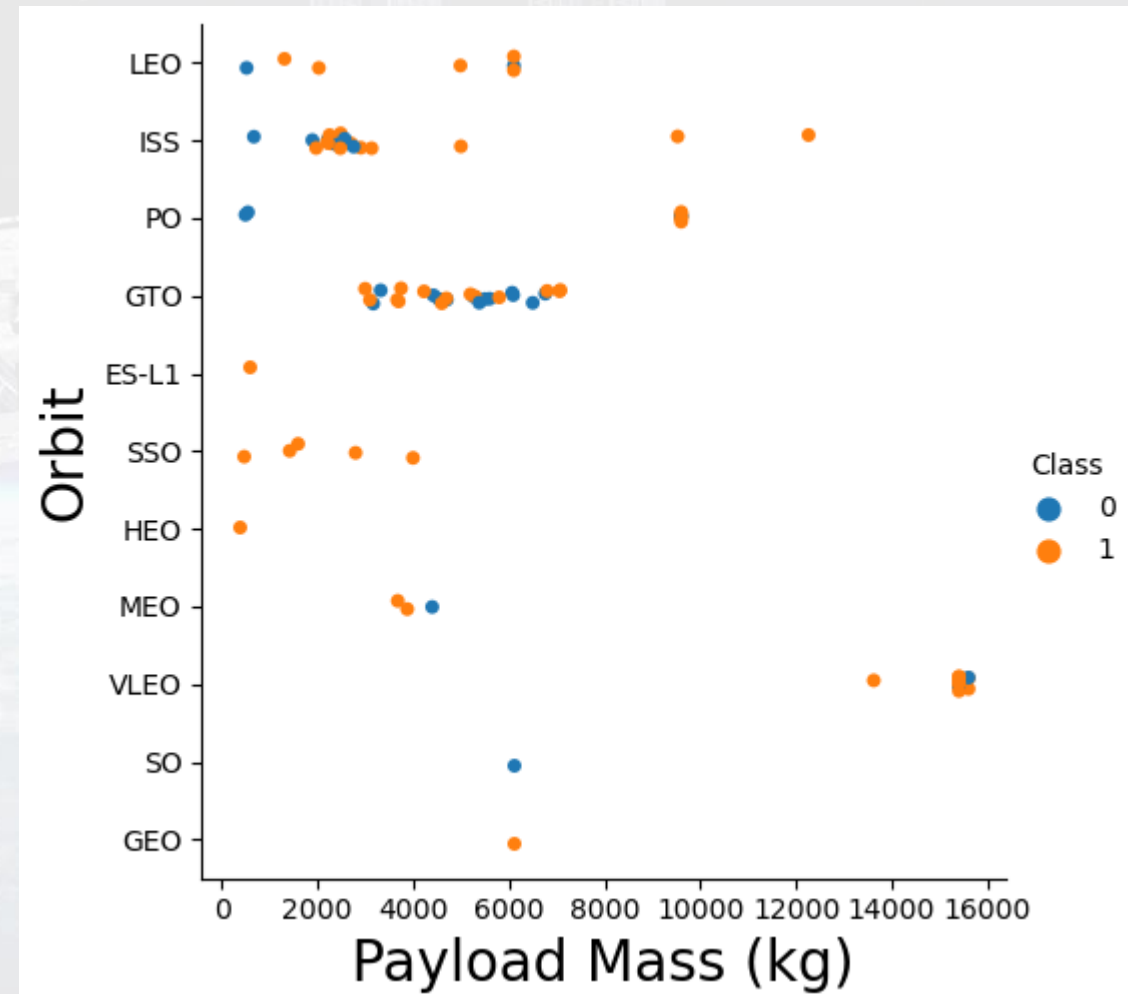
Payload vs Orbit Type

- No clear relation could be found between payload and success rate for the orbit GTO;
- ISS orbit has the widest range of payload and still manages to display a good rate of success;
- SO and GEO orbits have a significantly low number of flights



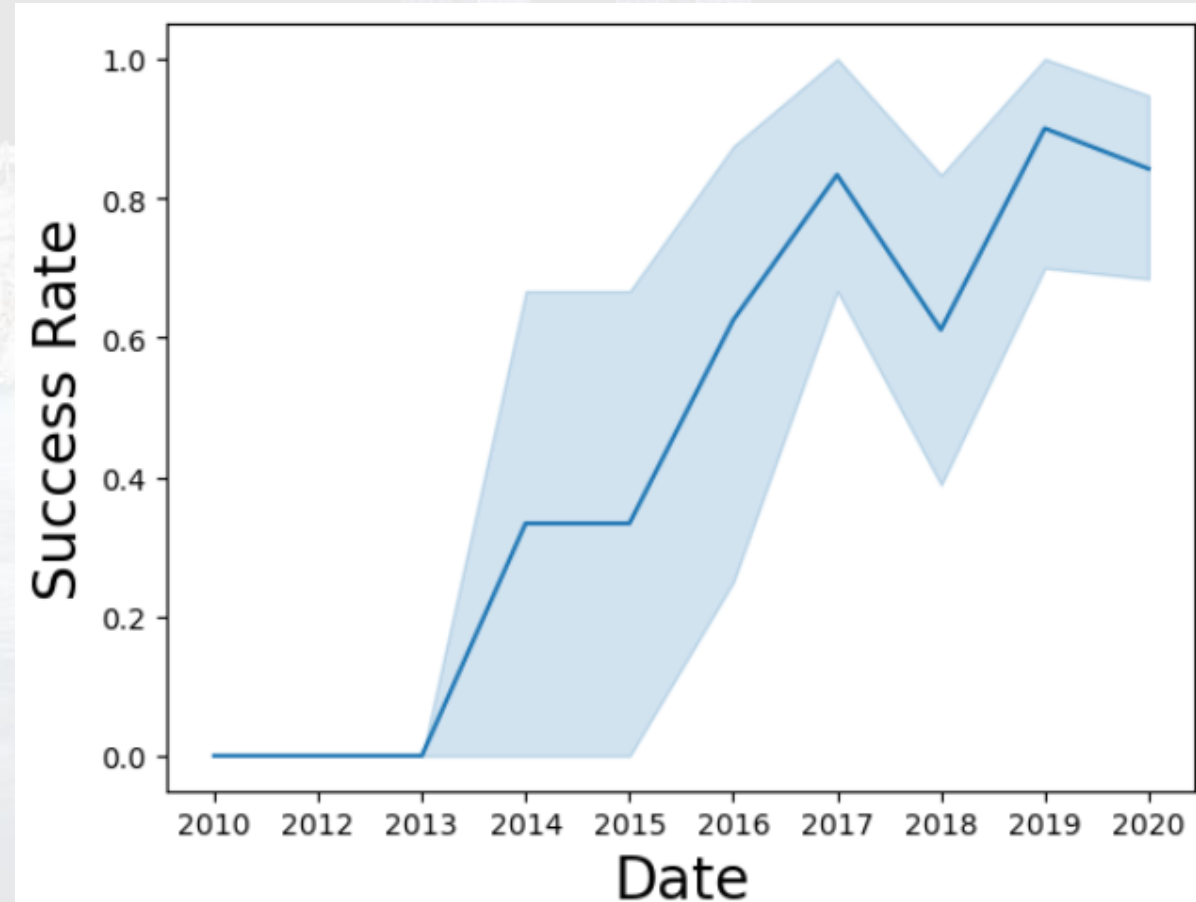
Payload vs Orbit Type

- No clear relation could be found between payload and success rate for the orbit GTO;
- ISS orbit has the widest range of payload and still manages to display a good rate of success;
- SO and GEO orbits have a significantly low number of flights



Yearly trend of successful launches

- The success rate started increasing consistently from 2013 up to 2020
- SpaceX was in the first three years clearly still in a research&development phase



SQL ANALYSIS – LAUNCH SITES NAMES

- Four launch sites were found in the data, by searching for unique occurrences of “launch_site” in the database:

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

SQL ANALYSIS – LAUNCH SITE ‘CCA’

- 5 sample records where “launch_site” column starts with ‘CCA’ (i.e. Cape Canaveral):

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

SQL ANALYSIS – TOTAL PAYLOAD MASS

- Total payload carried by boosters from NASA (obtained by summing all payloads whose code contains 'CRS'):

Total Payload (kg)

45596

SQL ANALYSIS – AVERAGE PAYLOAD MASS

- Average payload mass carried by booster version F9 v1.1:

The background of the slide features a faded image of a rocket launch. A rocket is shown ascending from the bottom left, with a large plume of fire and smoke. In the upper right, the structure of a space station or orbital platform is visible, consisting of several large rectangular panels and a central framework. Overlaid on this background is a semi-transparent data box with a teal header and a light blue body.

Avg Payload (kg)
2928

SQL ANALYSIS – DRONE SHIP LANDINGS

- Boosters which successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000 kg:

Booster Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

SQL ANALYSIS – NUMBER OF SUCCESSES AND FAILURES

- Number of successful and failure landing mission:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

SQL ANALYSIS – BOOSTERS WITH MAXIMUM PAYLOAD MASS

- Boosters which have carried the maximum payload mass (= 15600 kg):

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

SQL ANALYSIS – LAUNCH RECORDS OF 2015

- Failed landing outcomes in drone ship, their booster versions, launch site and dates for year 2015:

Date	Landing outcome	Booster version	Launch site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

SQL ANALYSIS – RANKING OF LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

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

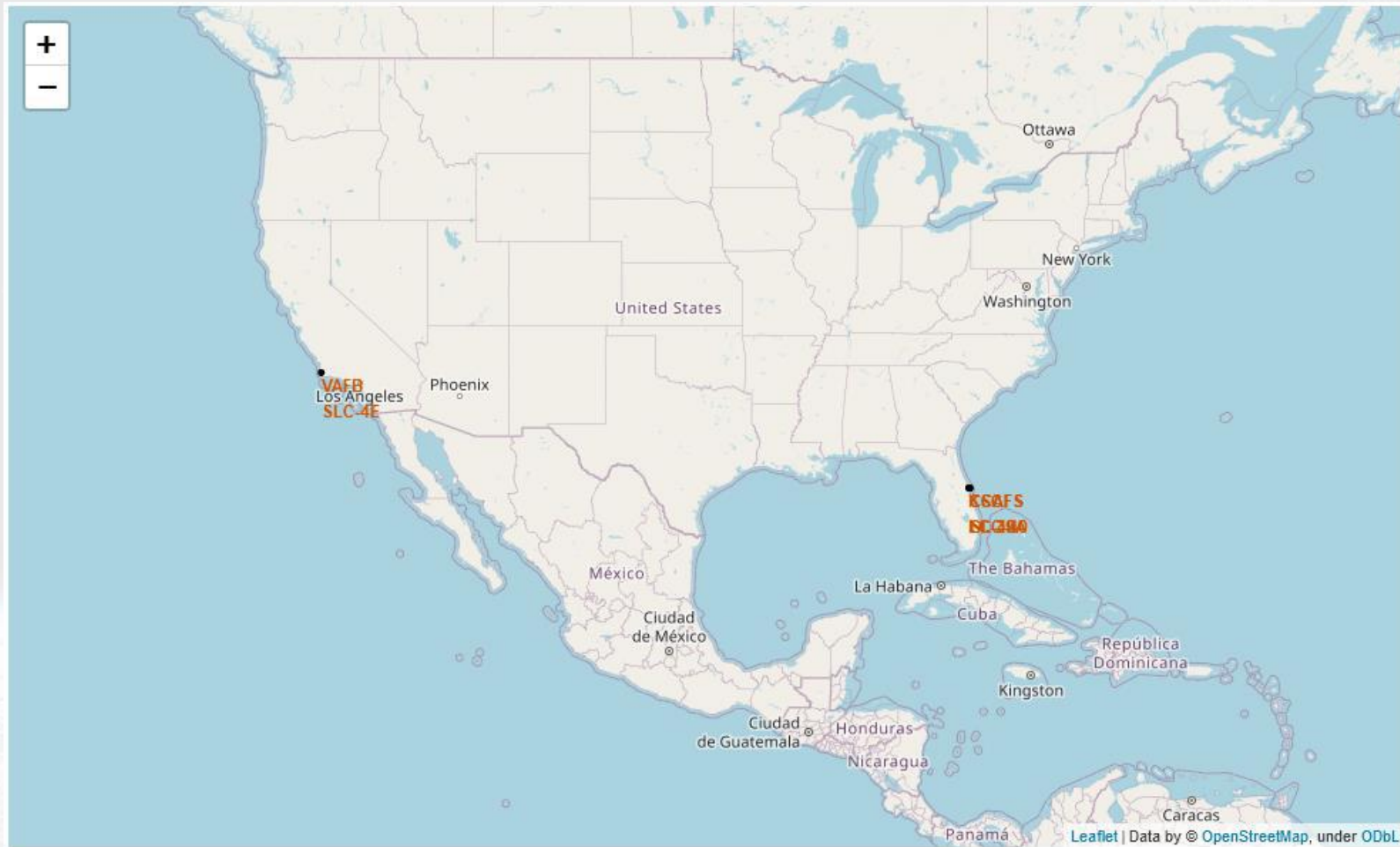
- This table draws the attention to the fact that “no attempt” must be taken into account moving forward.

A satellite image of Earth from space, showing the Italian peninsula and surrounding regions. The land is illuminated by sunlight, appearing in shades of brown and tan, while the surrounding oceans are dark blue. The curvature of the Earth is visible at the top of the frame.

Part 3

Analysis on launch site's positions

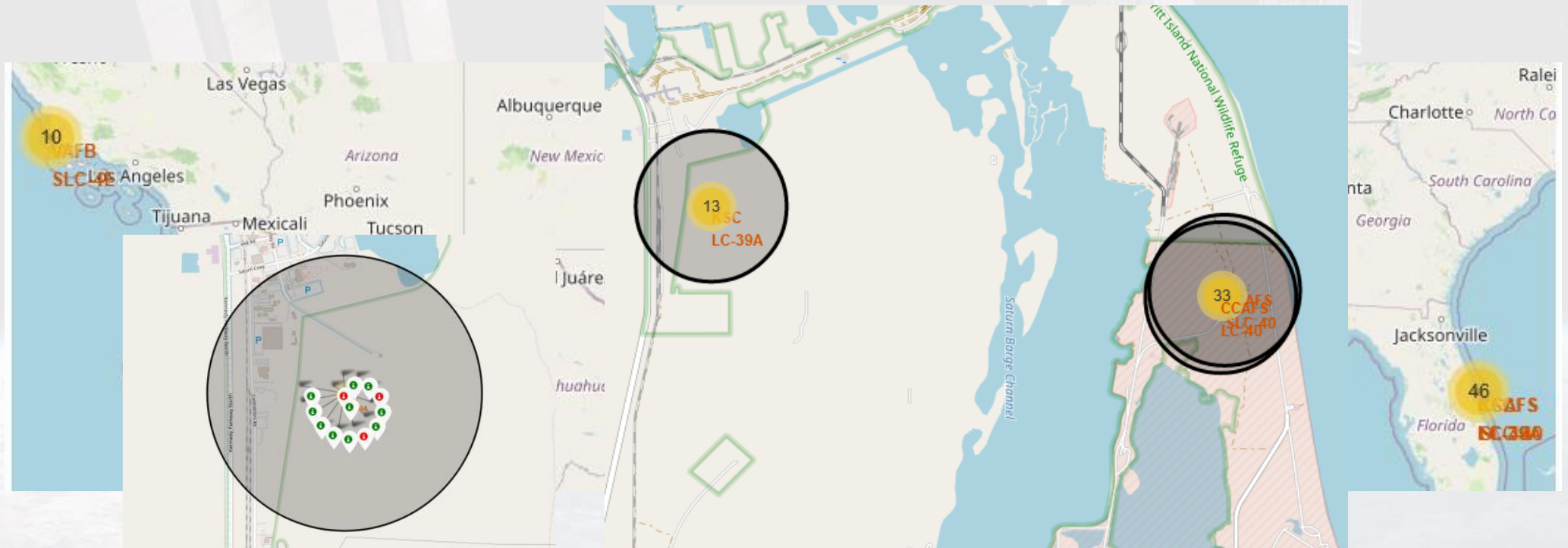
All available launch sites



Launch sites are near the sea for safety and logistics reasons, and are located closer to the equator to exploit Earth's rotational speed.

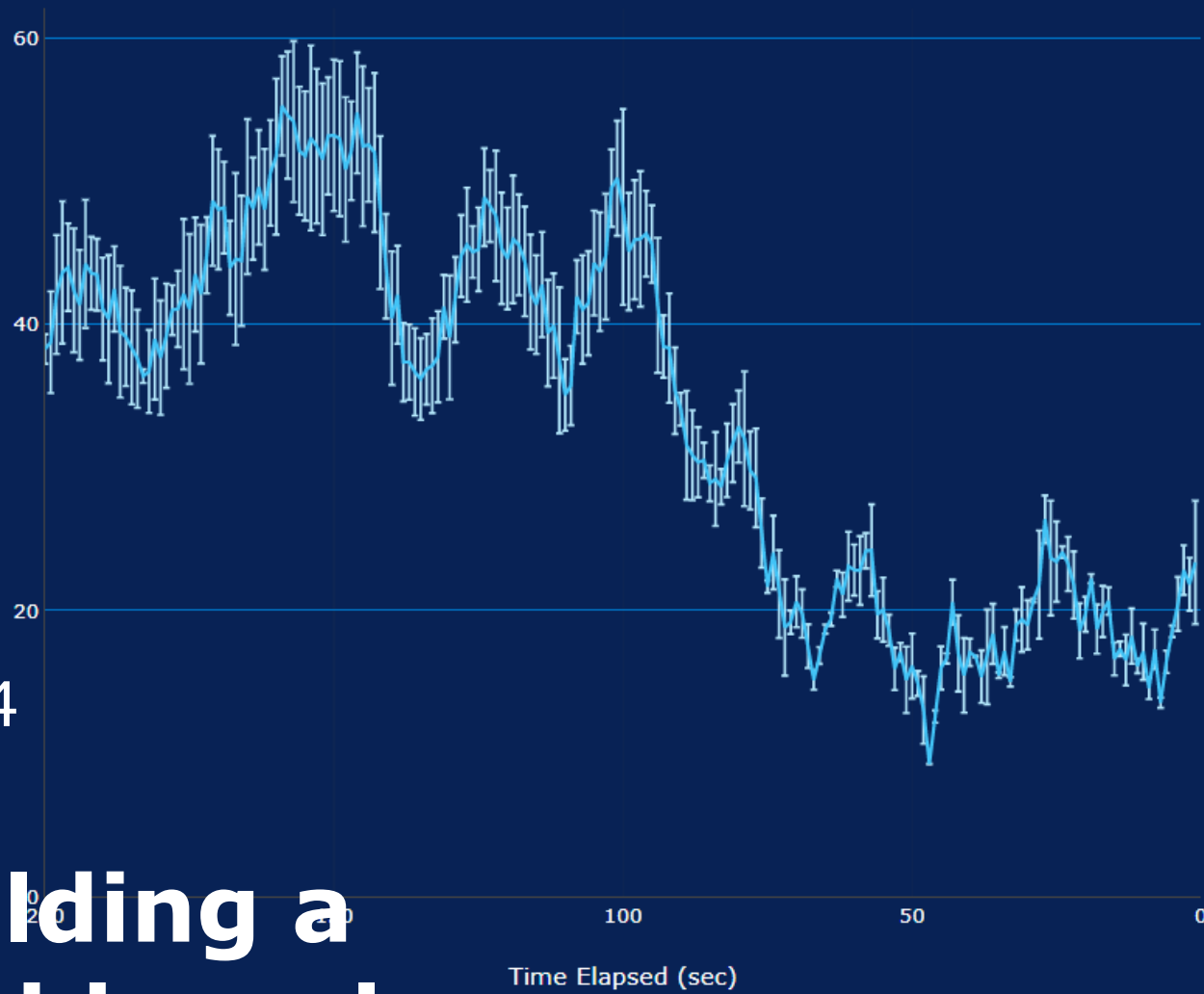
Launch outcomes by site

KSC LC-39A site is a very good one because of his proximity to the sea, being quite isolated from urban areas and being located to the south of the USA, closer to the equator:

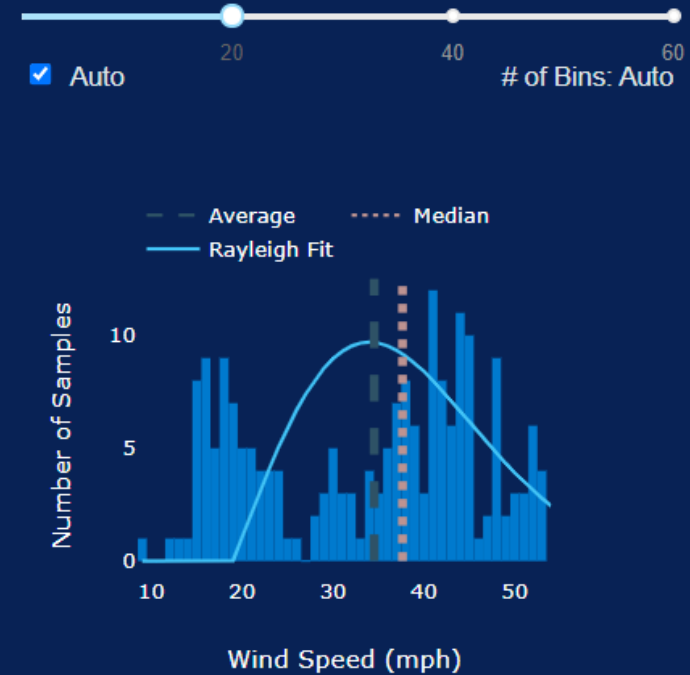


(green markers indicate success, red failure)

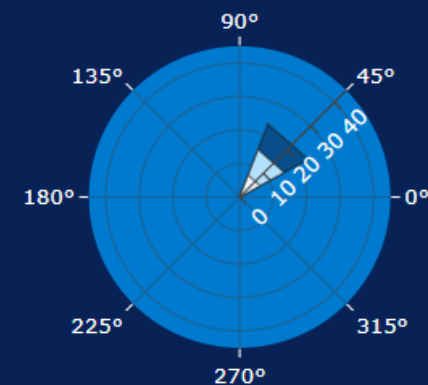
WIND SPEED (MPH)



WIND SPEED HISTOGRAM



WIND DIRECTION



Part 4

Building a
dashboard
with Dash

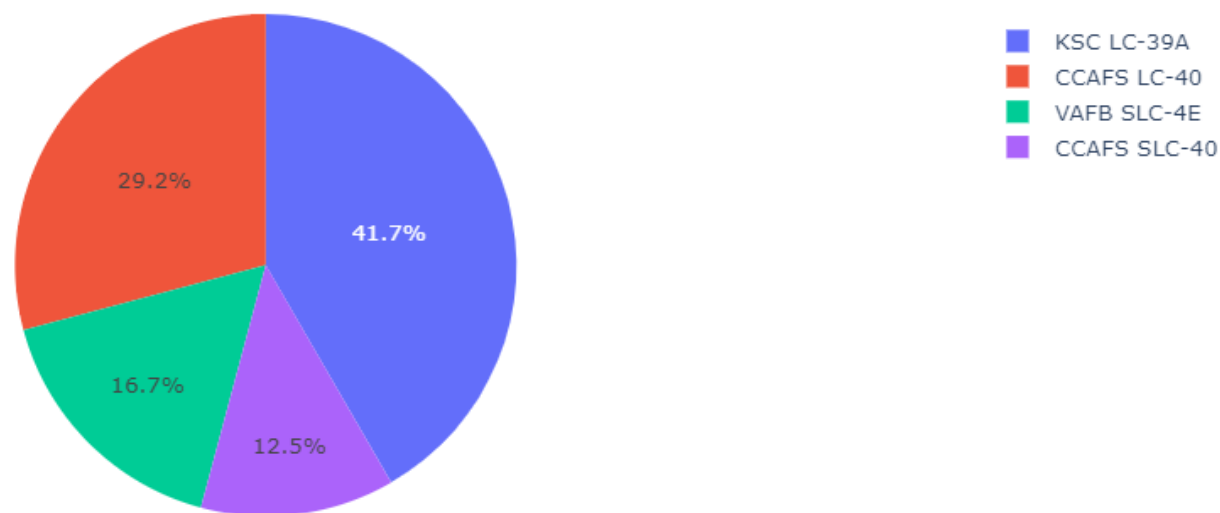
Successful Launches by Site

SpaceX Launch Records Dashboard

All Sites

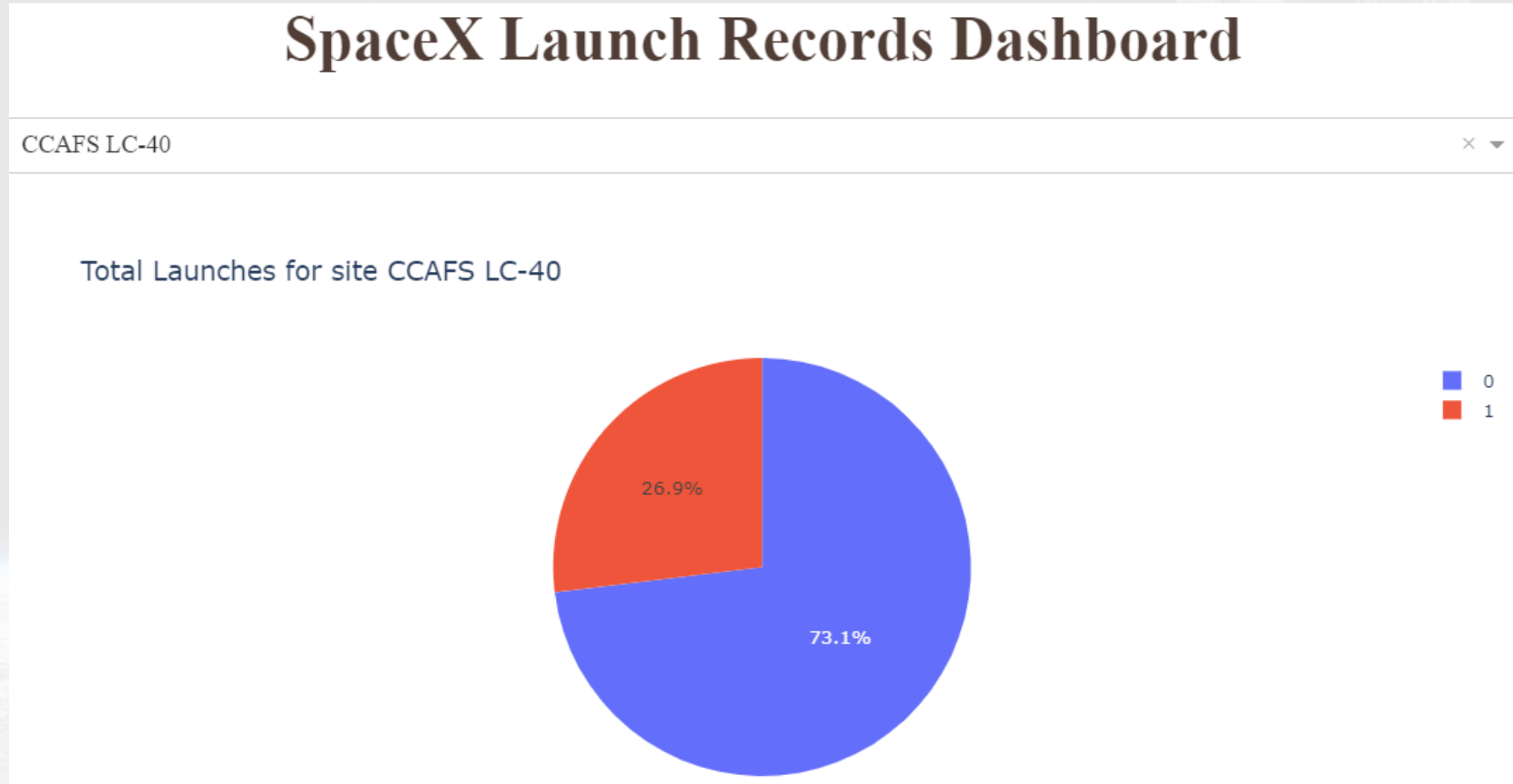


Total Success Launches By Site



- As already argued before, KSC LC-39A seems to be a very successful site

Launch success ratio for KSC LC-39A



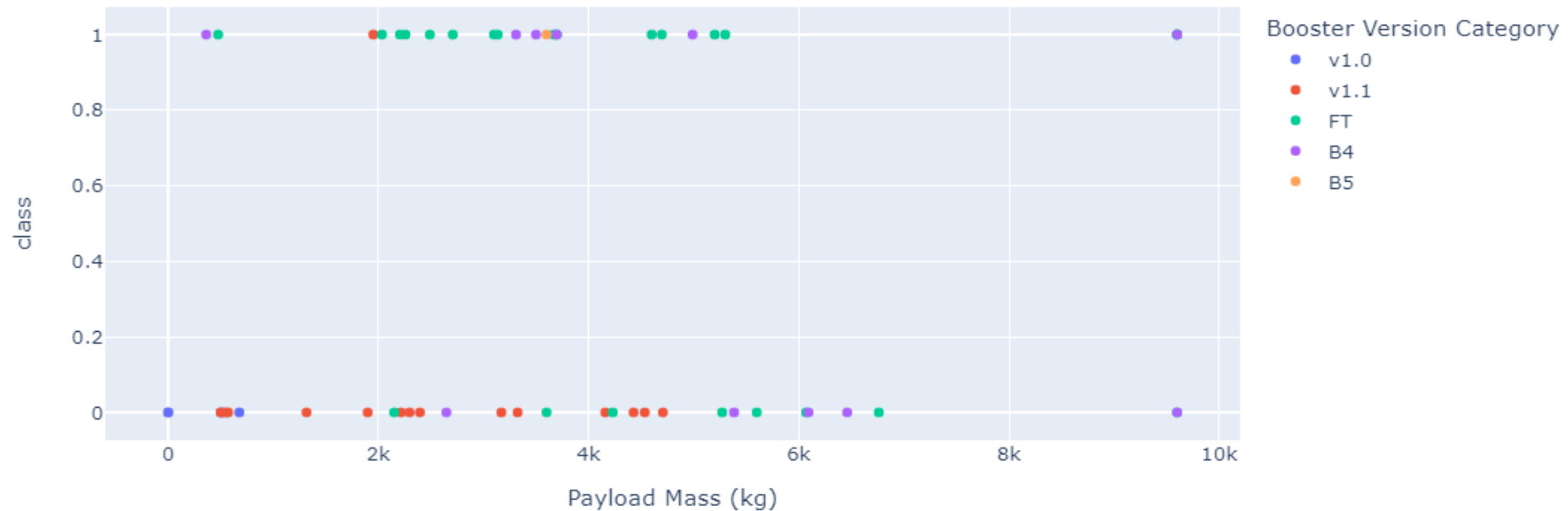
- 73.1% of launches are successful in this site.

Payload mass vs launch outcome

Payload range (Kg):



All sites - payload mass between 0kg and 9,600kg



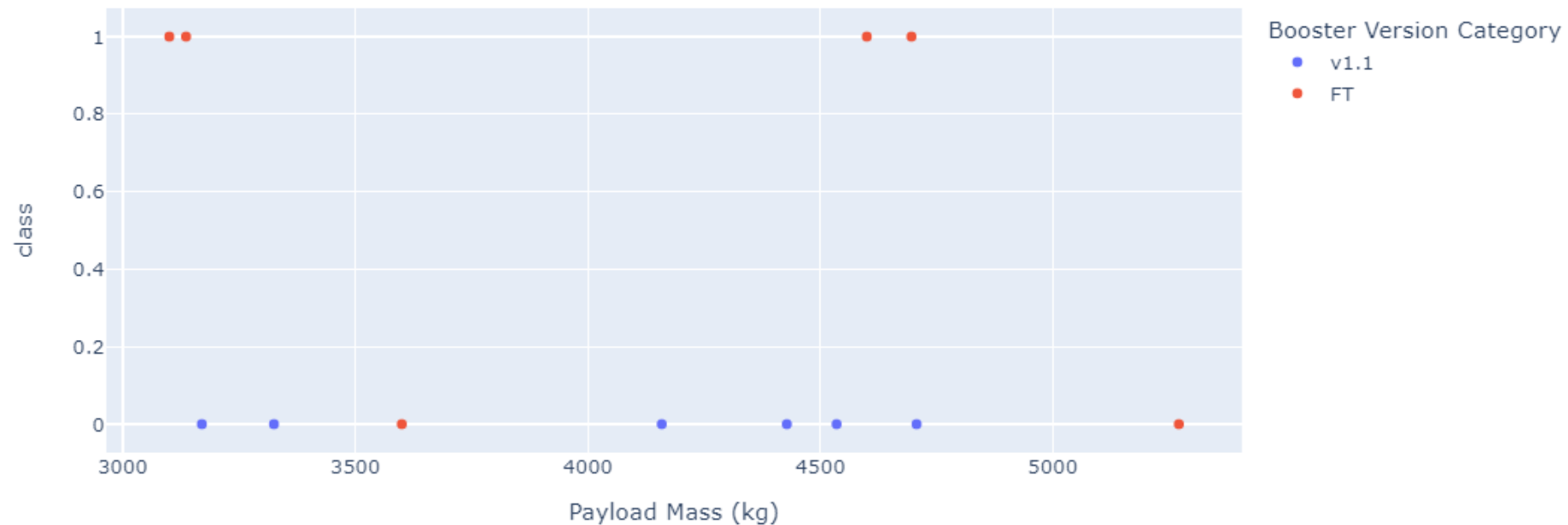
- Payloads under 6 ton paired with FT boosters seems to constitute the best possible recipe for a successful outcome

Payload mass vs launch outcome

Payload range (Kg):



Site CCAFS LC-40 - payload mass between 3,000kg and 7,000kg



- FT boosters proved to be the best when handling mid-range payloads (3 to 7 ton)

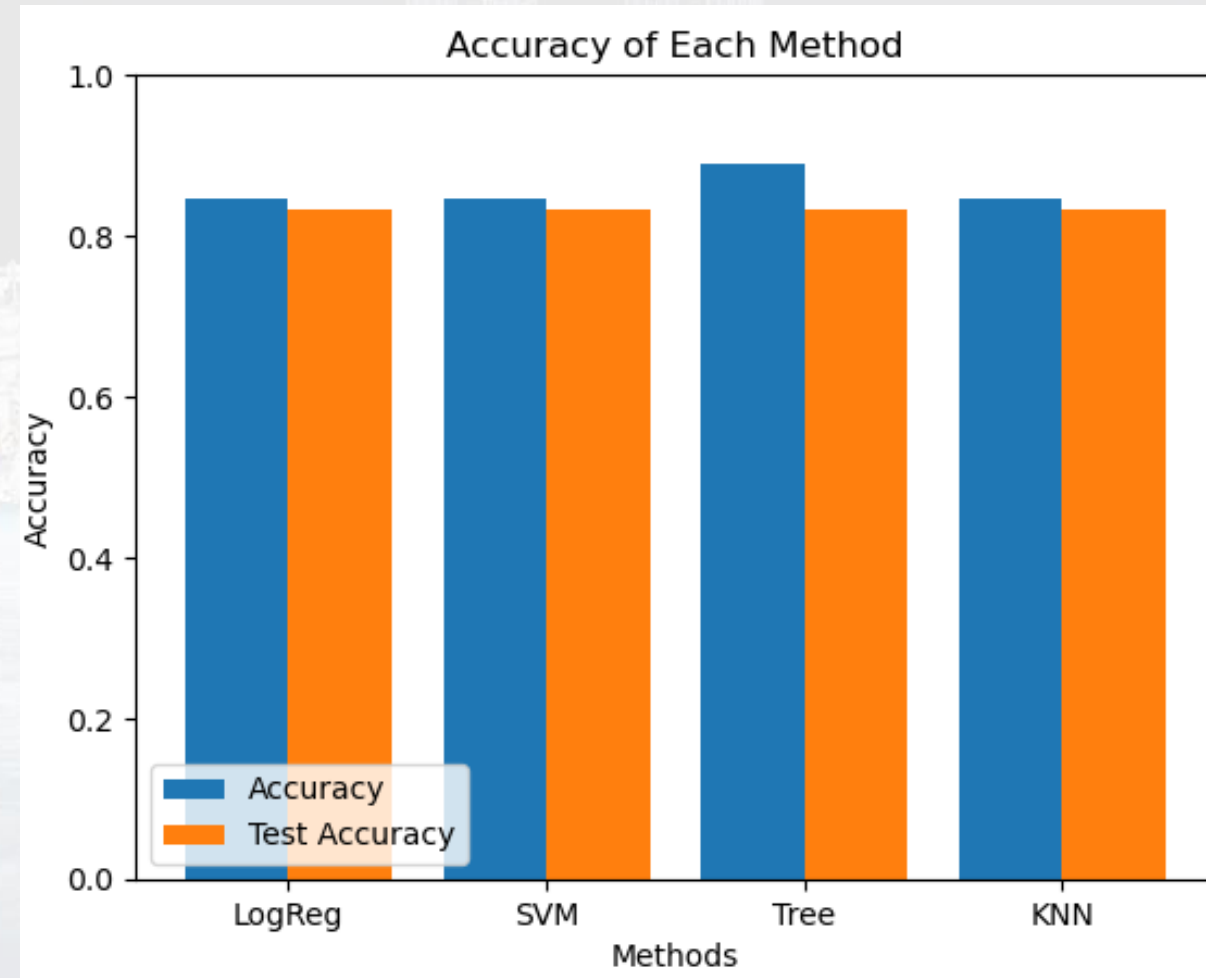
Part 5

Predictive Analysis (Comparison)

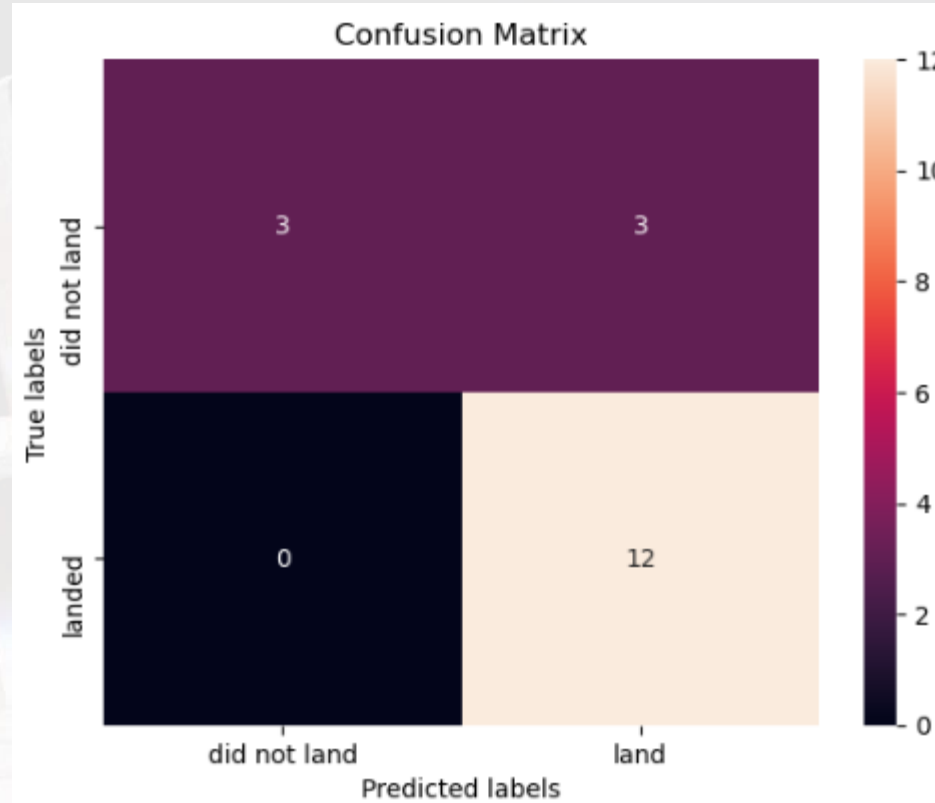


Classification accuracy

- Four classification models were trained (logistic regression, SVM, Tree classifier and KNN), and their accuracies are plotted beside;
- The model with the highest classification accuracy is **Decision Tree Classifier**, which has an accuracy of nearly 90%.



Confusion matrix of the decision tree classifier algorithm



- The confusion matrix proves its accuracy by showing that true positive and true negative are far more than the false ones.

Final table with all the ML algorithms

Model	Accuracy	TestAccuracy
LogReg	0.84722	0.83333
SVM	0.84722	0.83333
Tree	0.88889	0.83333
KNN	0.84722	0.83333

Conclusions

- Different data sources (SpaceX API, Wikipedia, etc.) were analyzed, refining conclusions as more and more data become clearer;
- The best launch site turned out to be KSC LC 39A because of its advantageous position (near the sea, well served by infrastructures and close to the equator);
- Launches above 7,000kg had a good success rate, but the best combination turned out to be FT booster with a payload mass not exceeding 6 ton;
- Although the majority of mission outcomes were a success,, successful landing outcomes seem to improve over time, most likely due to evolution of processes and the constant improvement of rockets;
- The *decision tree classifier* proved to be the best algorithm in predicting successful landings, and thus the ideal method to increase profits.

Appendix

- Folium graphs were not being showed in Github, so they were added here for clarity;
- All the codes and images were taken from the ipython scripts, which are referenced in the summary.

Thank you
for your
attention!

