



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

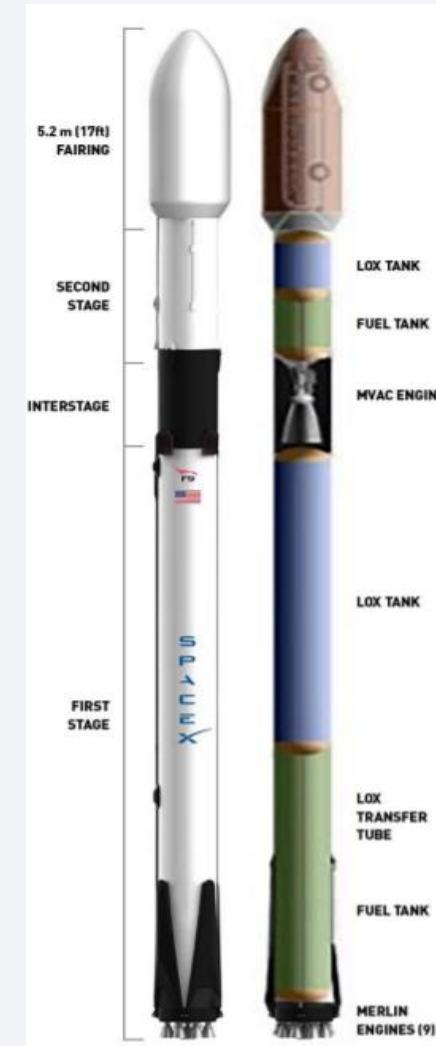
Vanessa De Oliveira
09/16/2021



Outline

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

Fig. 1 - Falcon 9 Block 5 overview



Source (Falcon 9 User's Guide)

Executive Summary

- Summary of methodologies

In this project, data extracted from SpaceX REST API and webscraped from the Falcon 9 Wikipedia page are used to investigate whether a Falcon 9 rocket will successfully land or not. After cleaning, wrangling, formatting, visualizing (on interactive folium and dashboard) and extensively analyzing the data in Python and SQL, different classification models were trained and tested. The models tested were Decision Tree, KNN, Logistic Regression and SVM.

- Summary of all results

All 4 models trained in this project presented the same evaluations, very likely due to the lack of observations. In general, the models performed well, correctly predicting all the successful landings (Precision =1); The EDA analyses show that CCAFS SLC-40 is the most popular site, but KSC LC-39A has the highest successful rate; Overtime, more technologies have improved landings performance and the last updated version of Falcon 9, Block 5, shows the highest successful rate among all versions and the highest payload capacity, which characterize large commercial advantage.

Introduction

- SpaceX is the only company to return a spacecraft from low-earth orbit;
- SpaceX's Falcon 9 rocket launches with a cost of US\$ 62 million while other competitors cost upward of US\$ 165 million;
- Reusing the first stage is the key determinant of SpaceX highly competitive operational cost;
- Knowing the likelihood of first stage successfully land or not could support competitors bidding against SpaceX for a rocket launch;
- Objective: classify whether the first stage will successfully land or not.

Section 1

Methodology

Methodology

Executive Summary

- Data Collection: SpaceX API and Webscraping Falcon 9 historical launch records from Wikipedia page titled List of Falcon 9 and Falcon Heavy launches;
- Data wrangling: Found patterns and converted landing outcomes into training labels with 1 means the booster successfully landed 0 means it was unsuccessful;
- Performed exploratory data analysis (EDA) using visualization and SQL;
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
 - Built, tuned, evaluated the classification models

Data Collection

- Data was collected from 2 main sources:
 - SpaceX REST API: Main data source used for investigation, model training, test and prediction;
 - Webscraping : Collected Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches. This is a secondary data source used for the application of webscraping technique and further application of magic SQL method for investigation purpose.

Data Collection – SpaceX API

Overview:

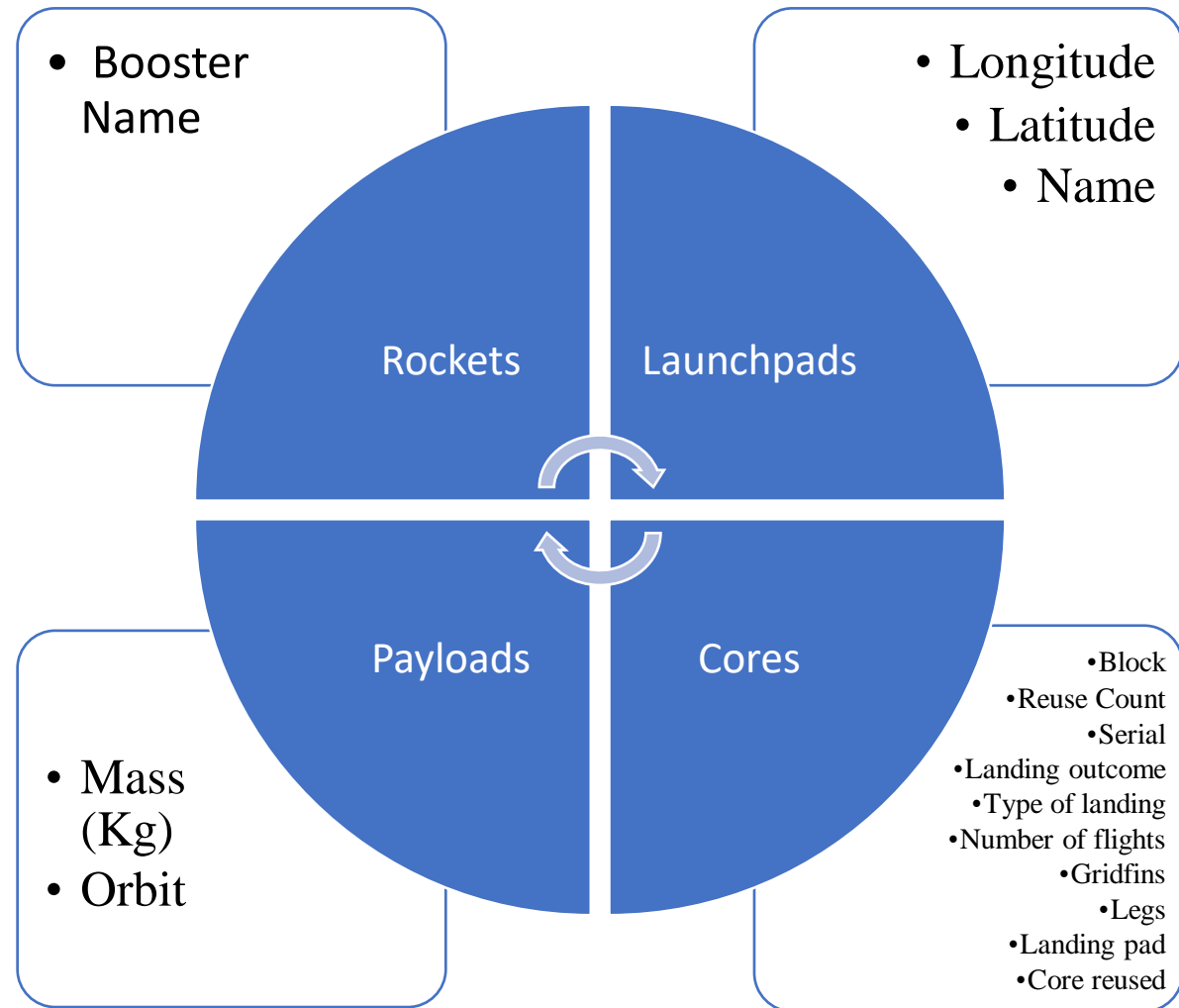
- Data Collection from SpaceX REST API;
- Defined functions to extract information from data;
- Requested and parsed data using GET request;
- Filtered data of interest, dropped rows with multiple cores/ payloads and formatted datetime;
- Defined lists to accommodate data extraction;
- Used functions to decode dataset and extract it to the lists;
- Created dictionary and assigned lists to correspondent keys, then save in Pandas;
- Filtered only Falcon 9 observations;
- Replaced payload missing values with the mean;

Code: <https://github.com/vantoks/Spacex/blob/master/space.ipynb>

Data Collection – SpaceX API

Using Get request to extract data from:

- Rockets
- Payloads
- Launchpads
- Cores



Code:

<https://github.com/vantoks/Spacex/blob/master/space.ipynb>

Data Collection - Scraping

Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia

Request Falcon 9 Launch Wiki page from its URL



Extract all column/variable names from the HTML table header



Create a data frame by parsing the launch HTML tables

Code: <https://github.com/vantoks/SpaceX/blob/master/webscraping.ipynb>

Data Wrangling

Exploratory Data Analysis (EDA)

- Information to explore:
 - Number of launches on each site
 - Number and occurrence of each orbit
 - Number of mission outcome per orbit type

Create a landing outcome

- Converted categorical landing outcomes into binary labels where 1 means the booster successfully landed and 0 means it was unsuccessful

Code: <https://github.com/vantoks/SpaceX/blob/master/wrangling.ipynb>

EDA with Data Visualization

Visualize the following relationships:

- Flight Number and Launch Site
 - Insight about sites usage over the years;
- Payload and Launch Site
 - Is there any preferable site for different ranges of payload?
- Success rate of each orbit type
 - Can the choice of orbit characterize different challenges for the success landing?
- Flight Number and Orbit type
 - Have the destination orbit changed overtime?
- Payload and Orbit type
 - Is there any correlation between the payload mass the orbit of destination?
- Launch success yearly trend
 - Investigate launching progress overtime;

Code: <https://github.com/vantoks/SpaceX/blob/master/dataviz.ipynb>

EDA with Data Visualization

Features Engineering:

- Selected only features;
- Transformed all categorical variables in dummies;
- Casted all columns to float64.

EDA with SQL

SQL was used to explore the webscraped Falcon 9 data:

- Unique launch sites in the space mission;
- Visualize records where launch sites 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;
- 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;
- 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;
- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20.

Code: <https://github.com/vantoks/Spacex/blob/master/eda-sql.ipynb>

Build an Interactive Map with Folium

Points on interactive maps:

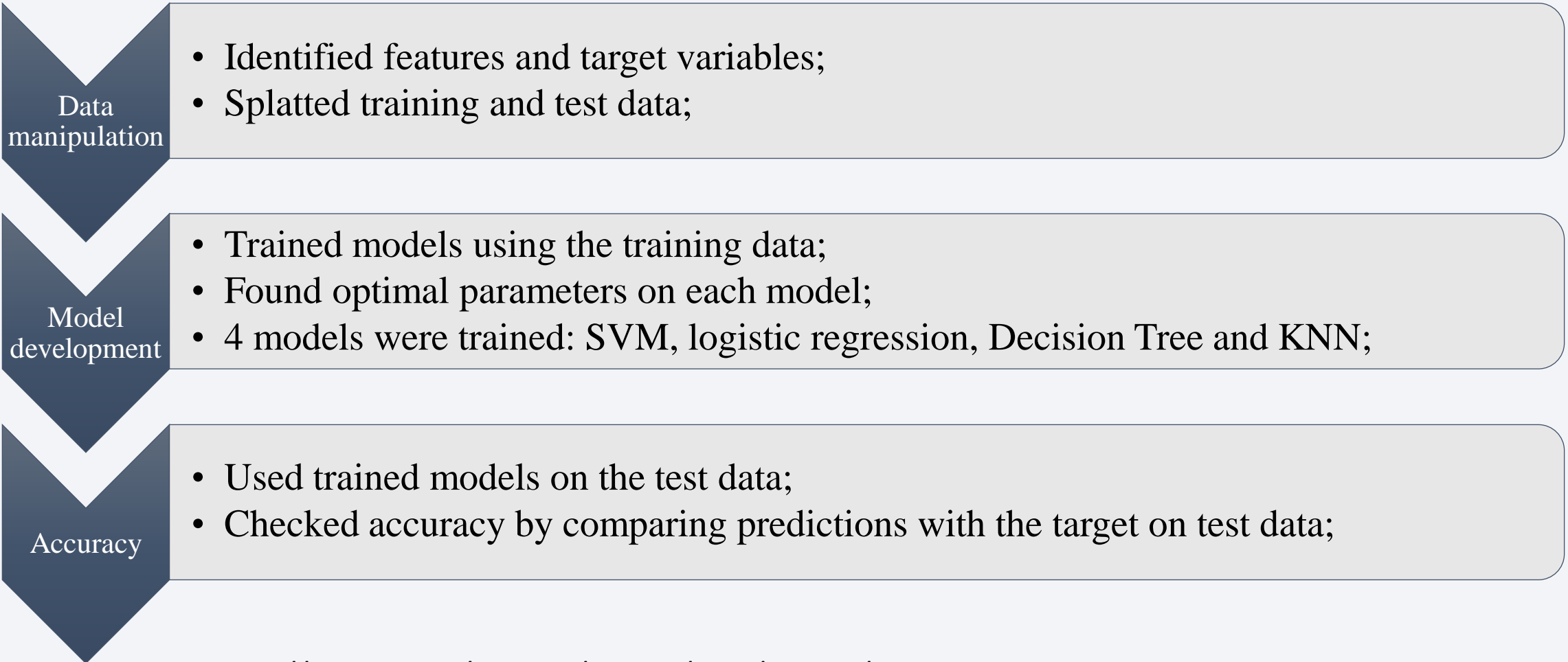
- All launch pads:
 - Circles are shown near the U.S. Atlantic and Pacific coasts;
- Success/failed launches for each site:
 - Quickly visualize launches outcome as markers on each site;
- Distance between CCAFS-SLC and the Atlantic ocean:
 - Blue line connects CCAFS-SLC (which has the greatest number of flights) and the Atlantic ocean. The proximity to the ocean is important since it is the destination of several landings;
- Distance between CCAFS-SLC site and Orlando, FL:
 - The location also allows to leverage aerospace talent pool available in Florida.

Build a Dashboard with Plotly Dash

Dashboard includes:

- Pie Chart of total success launches;
 - Blue: Represents the outcome 1, for successful landing;
 - Red: Outcome 0, for failed landing;
- Scatter Plot of success count of payload mass
 - Payload range select the payload range of interest;
 - Visualize 3 options of boosters: FT, B4, B5
 - Attempt to visualize patterns in the data that could indicate potential correlation between booster and payload mass;
- Code: https://github.com/vantoks/Spacex/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)



Code: <https://github.com/vantoks/Spacex/blob/master/Prediction.ipynb>

Results

- CCAFS SLC 40 hosted most part of the flights over the years and VAFB SLC 4E is the least popular site,
- VAFB SLC 4E hasn't hosted flights with payloads greater than 10,000 kg, KSC LC-39A successfully landed all flights with payload less than approximately 5,500 Kg, CCAFS SLC 40 seems to be fore successful landing heavier payloads;
- ES-L1, GEO and SSO seems to have the highest success rate while GTO presents the lowest;
- Overall, higher success rate can be found on higher flight numbers, showing that most recent flights are more successful than older ones;
- There seems to be no relationship between flight number and the GTO orbit;
- Heavier payloads (>9,500 kg) headed VLEO, ISS and PO orbits and they were mostly successful;
- GTO does not seem to have a correlation with Payload mass;
- Success rate has increased since 2013;
- Launches sites seem to be strategically located near on the ocean and relatively high populated areas;
- All the 4 models presented similar results and accuracies, potentially due to the small dataset used.

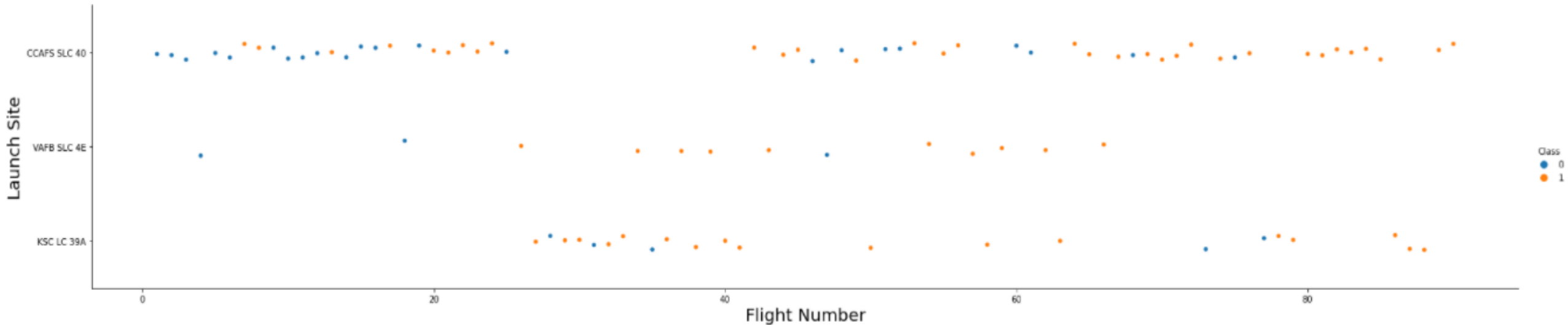
The background of the slide is a complex, abstract composition. It features a dark blue base color on the left, which transitions into a vibrant, multi-colored area on the right. This transition is achieved through a series of diagonal, overlapping bands and streaks in shades of red, teal, and light blue. A fine, white grid pattern is visible throughout the image, particularly in the darker areas, giving it a digital or data-driven appearance. The overall effect is one of dynamic movement and high-tech aesthetics.

Section 2

Insights drawn from EDA

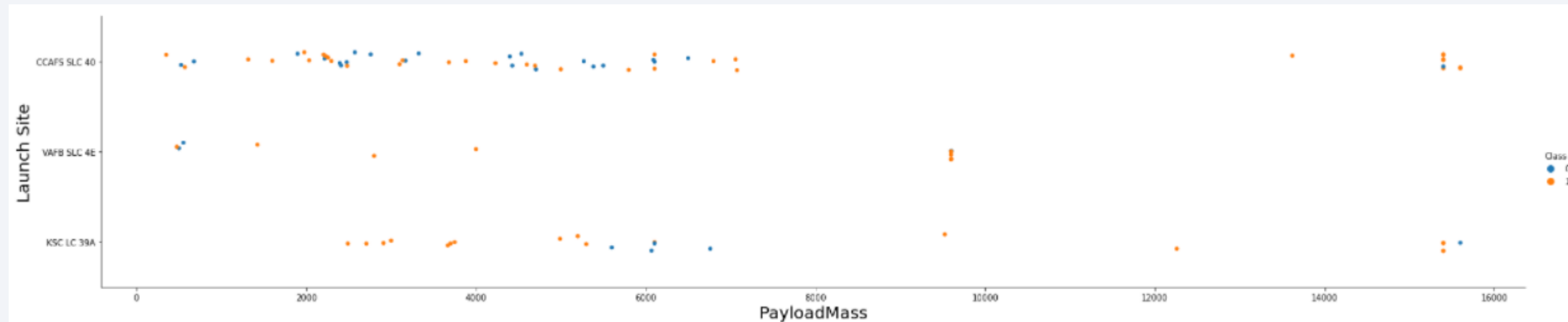
Flight Number vs. Launch Site

- Most of the launches happened from CCAFS SLC 40
- VAFB SLC 4E is the least popular site;
- Higher flight numbers present higher likelihood of successful land, so flights are getting better at landing overtime;
- KSC LC 39A became more popular around flight 25;



Payload vs. Launch Site

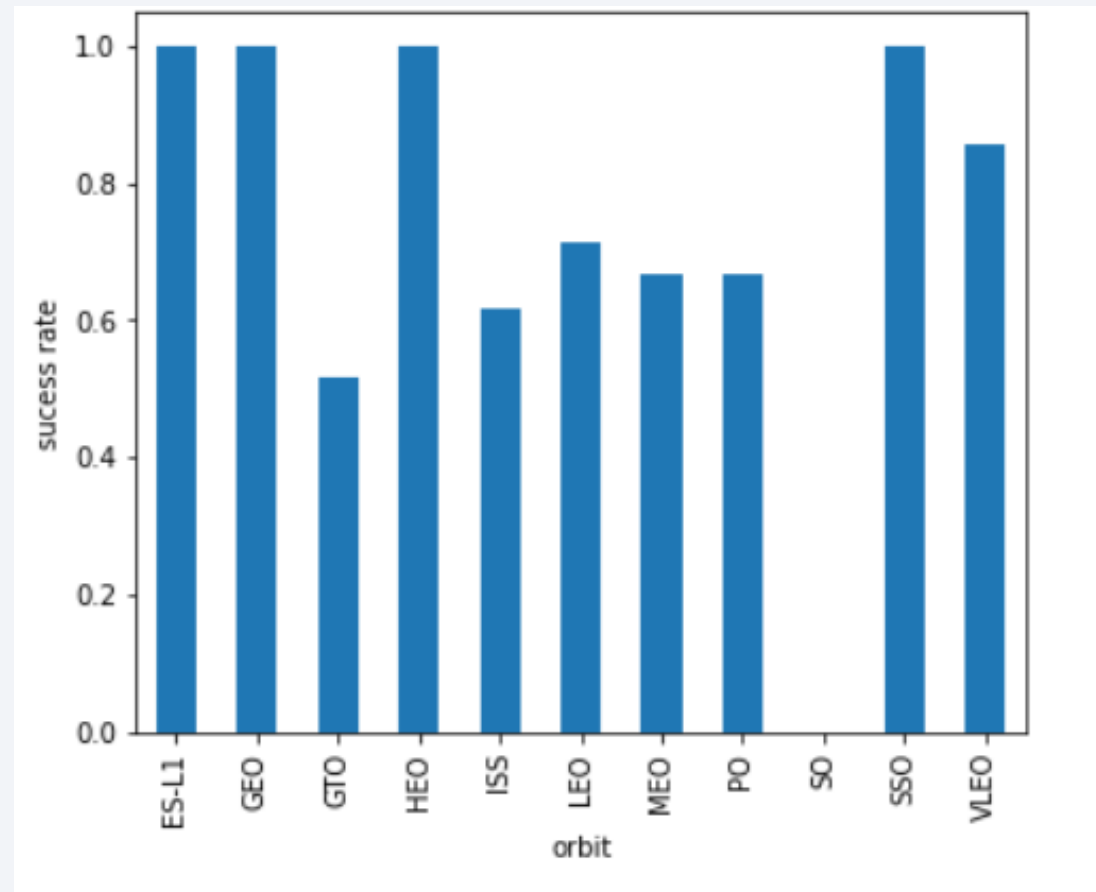
- Most launches presented payload mass under 7,000 kg;
- Heavier payload mass were more likely to successfully land;
- KSC LC-39A launches between 5,000 kg and 7,000 kg were less likely to succeed;
- Overall, payloads heavier than 7,000 kg seem to present higher success rate.



Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO seem to be associated with the highest success rates
- GTO has the lowest success rate;

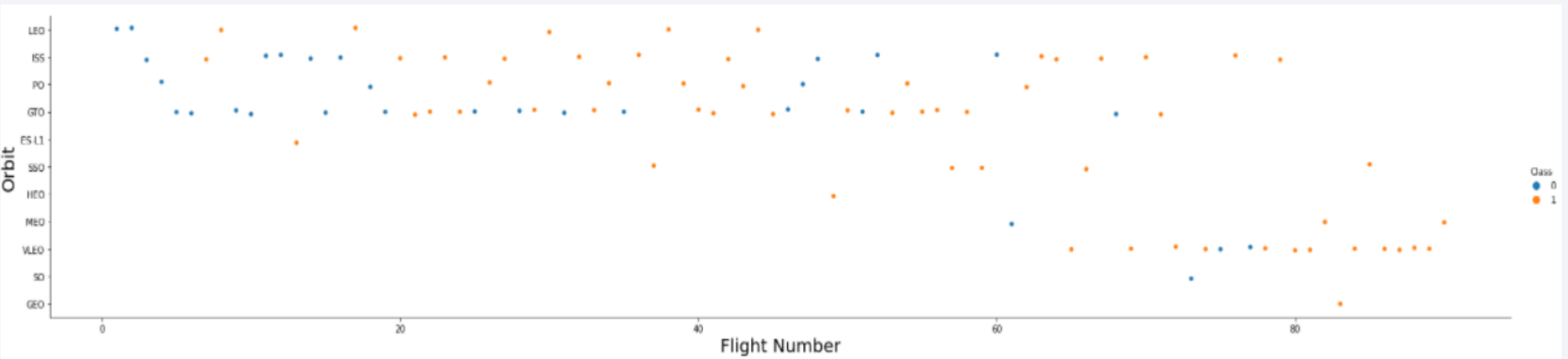
Fig. 2 – Success rate by orbit type



Flight Number vs. Orbit Type

- There is a potentially positive relationship between LEO orbit and the success rate;
- There seems to be no relationship between flight number and the GTO orbit.

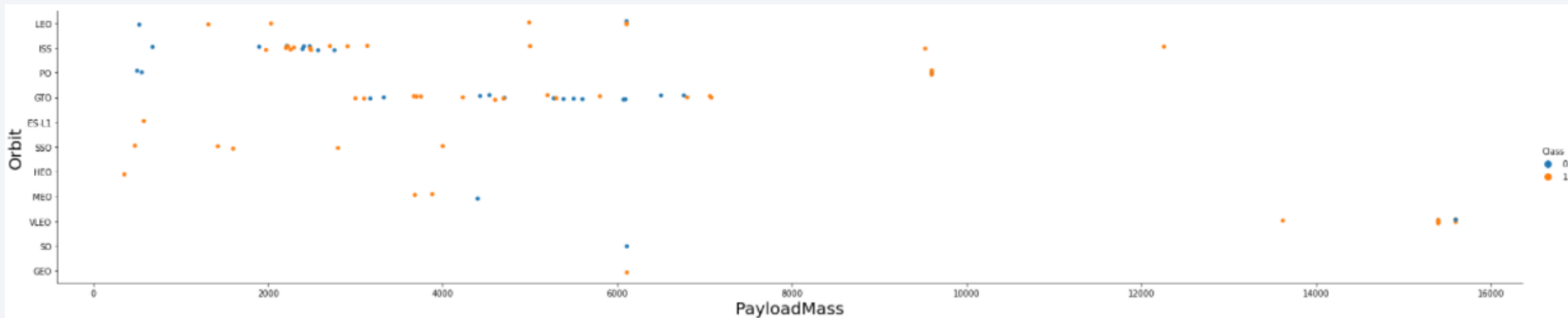
Fig. 3 – Flight Number vs. Orbit Type



Payload vs. Orbit Type

- Heavier payloads (>9,500 kg) headed VLEO, ISS and PO orbits and they were mostly successful;
- GTO does not seem to have a correlation with Payload mass;

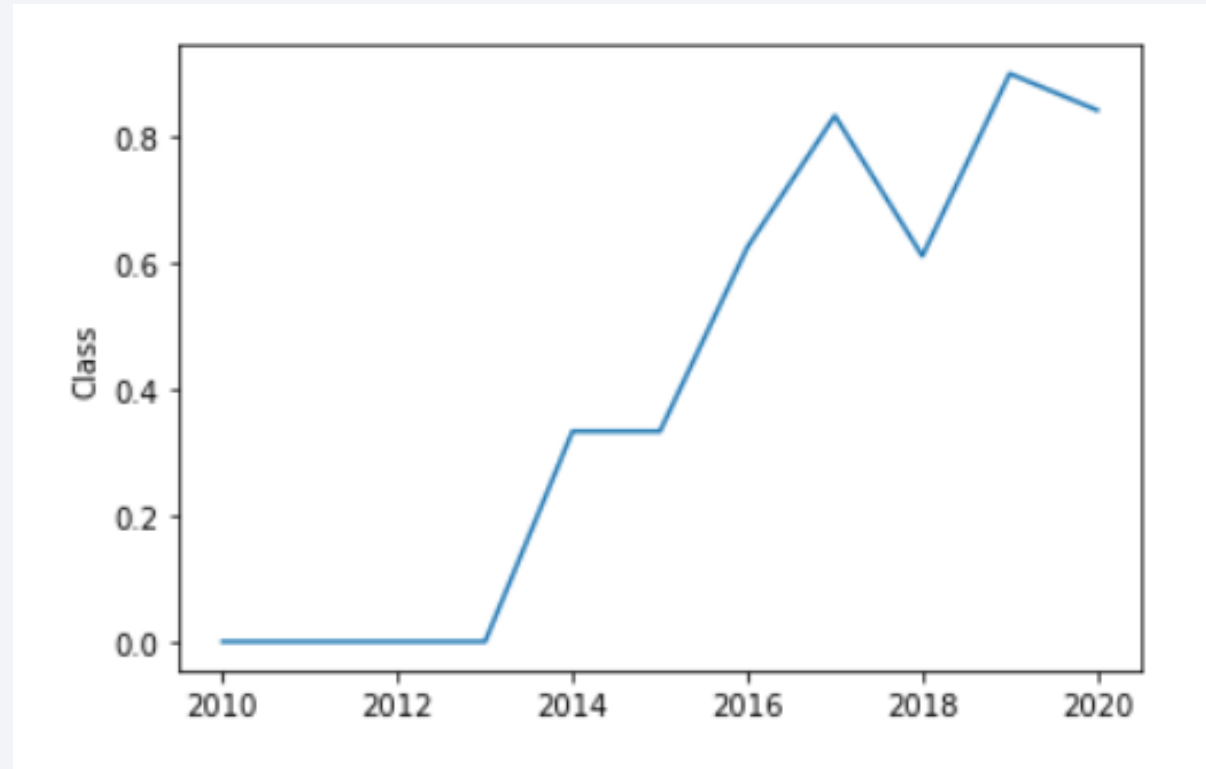
Fig. 4 – Payload vs. Orbit Type



Launch Success Yearly Trend

- Since 2013 the success rate has increased;

Fig. 5 – Launch Success Yearly Trend



All Launch Site Names

- Figure 6 shows the unique launch sites identified on the webscraped data from Falcon 9 Wikipedia page. CCAFS LC-40 is the older name of CCAFS SLC-40 site in Florida. KSC LC-40 is also located in Florida and VAFB SLC-4E is the launch site in California;

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

Figure 6 – All launch site names
Source: Outcome generated in SQL

Launch Site Names Begin with 'CCA'

Figure 7 shows that the older launches from Cape Canaveral are all described as CCAFS LC-40, which is the previous name of CCAFS SLC-40,

Fig. 7 – First 5 observations of CCAFS LC-40 and CCAFS SLC-40 sites

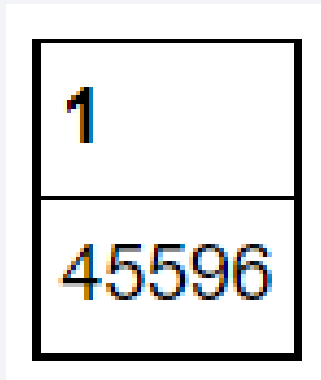
DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Source: Outcome generated in SQL

Total Payload Mass

- NASA selected Falcon 9 launch vehicle and Dragon spacecraft to transport cargo and crew to the International Space Station (ISS), and other LEO destinations, and for Cargo Resupply Services (CRS). The contract was for a guaranteed minimum of 20,000 kg to be carried to the ISS. (Falcon 9 Launch Vehicle Payload User's Guide)
- Figure 8 shows that the total payload carried by boosters from NASA is 45,596 kg until July 2021.

Fig. 8 – Total Payload Mass carried by NASA




Source: Outcome generated in SQL

Average Payload Mass by F9 v1.1

- Falcon 9 can accommodate typical payloads with mass from 1360 to 6800 kg. (Falcon 9 User's Guide)
- Figure 9 shows that the average payload mass carried by booster version F9 v1.1 is 2,534 kg;

Fig. 9 – Average Payload Mass by F9 v1.1



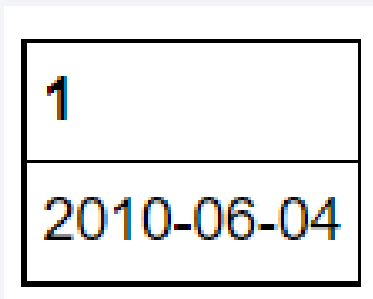
1
2534

Source: Outcome generated in SQL

First Successful Ground Landing Date

- Figure 10 shows that the first successful ground landing happened on April 6, 2010.
- Successful ground landings have the advantage to present reduced costs of spacecraft recover;

Fig. 10 – First Successful Ground Landing Date



1	2010-06-04
---	------------

Source: Outcome generated in SQL

Successful Drone Ship Landing with Payload between 4000 and 6000

- Figure 11: Boosters that successfully landed on drone ship and had payload mass between 4000 and 6000;

Fig. 11 - Boosters names

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Source: Outcome generated in SQL

Total Number of Successful and Failure Mission Outcomes

- Figure 12 shows that the webscraped data present 3 different mission outcomes;
- Out of the 101 observations, the majority, 99 counts, were successful missions;

Fig. 12 - Total success count by mission outcome

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

Source: Outcome generated in SQL

Boosters Carried Maximum Payload

- Figure 13 shows that the boosters which have carried the maximum payload mass are mainly the F9 B5 versions;

Fig. 13 – Booster versions carrying 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

Source: Outcome generated in SQL

2015 Launch Records

- There were 2 failed landing outcomes in drone ship in 2015. They are both F9 v1.1 booster versions, and both launched from CCAFS LC-40 site;

Fig. 14 – Failure landings in drone ship in 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

Source: Outcome generated in SQL

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

- Figure 15 ranks the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order;
- Retrieval attempts only happen on missions where the launcher has enough leftover fuel to accomplish the re-entry maneuvers necessary to safely return a rocket. Thus, no attempt outcome is associated with SpaceX decision of do not try to recover the booster under certain flying conditions;
- Drone ship landing total counts of failure and success are the same;

Fig. 15 – Count per Landing Outcomes from 2010-06-04 to 2017-03-20

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

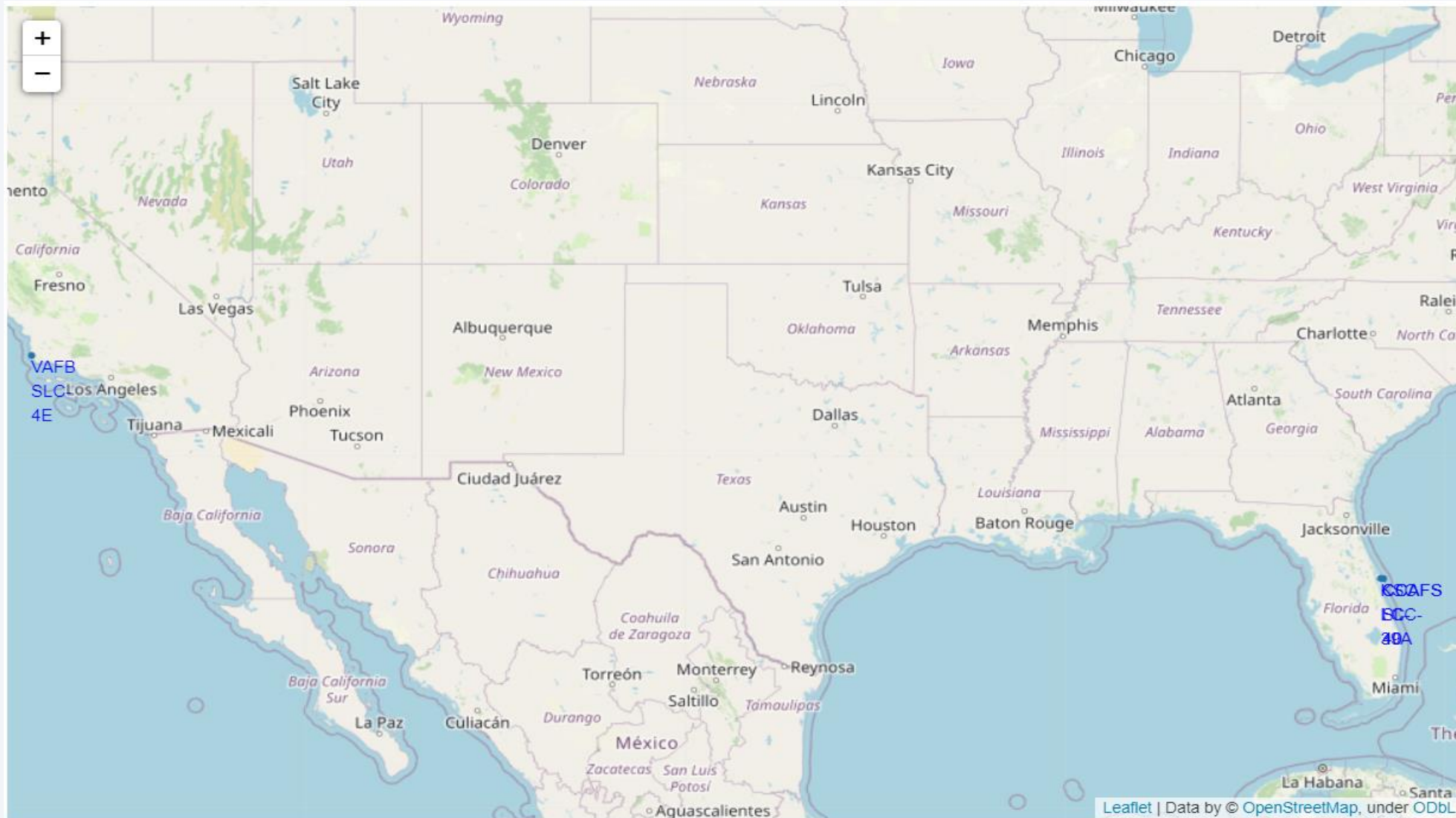
Source: Outcome generated in SQL

Section 4

Launch Sites Proximities Analysis

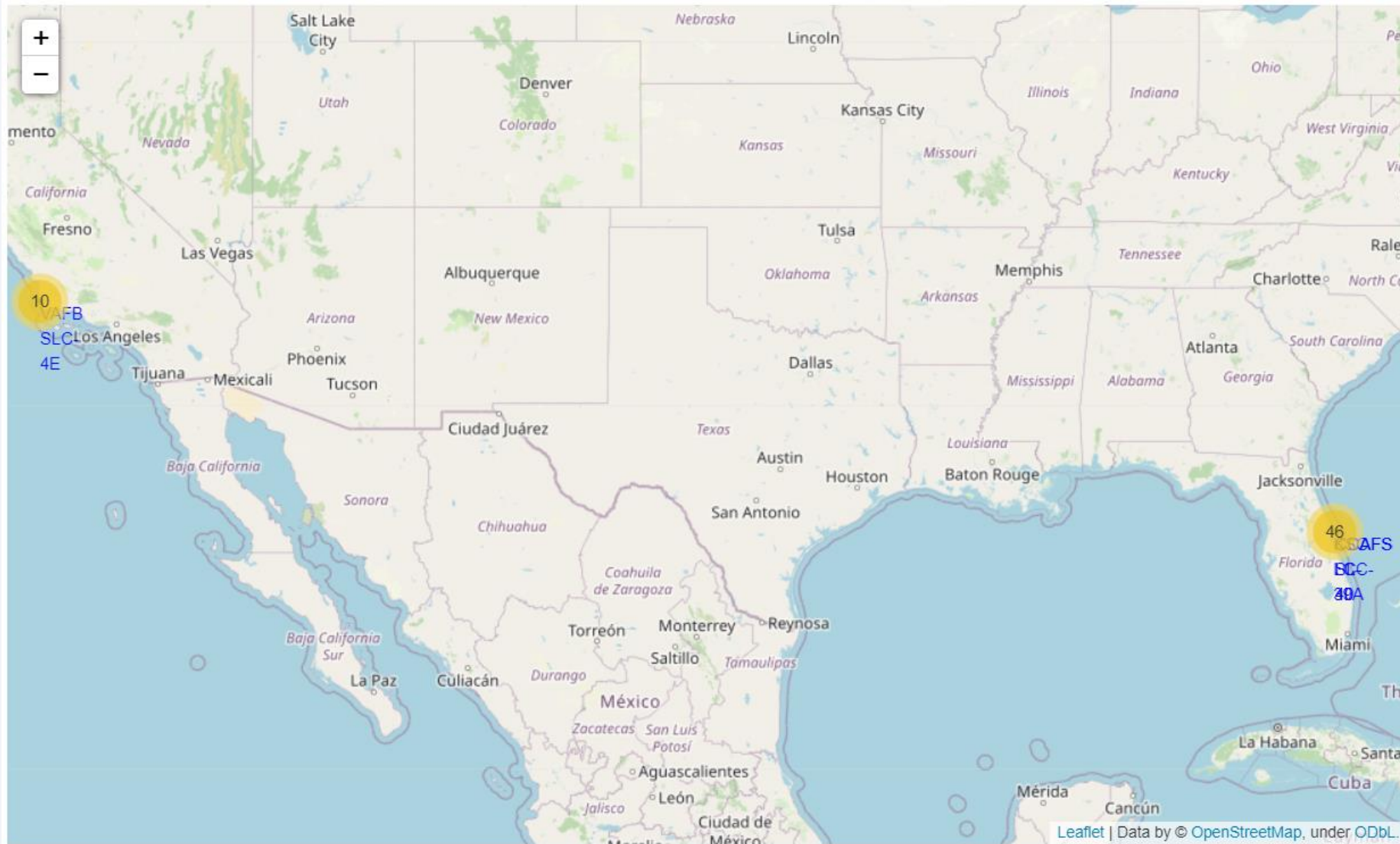


All launch sites



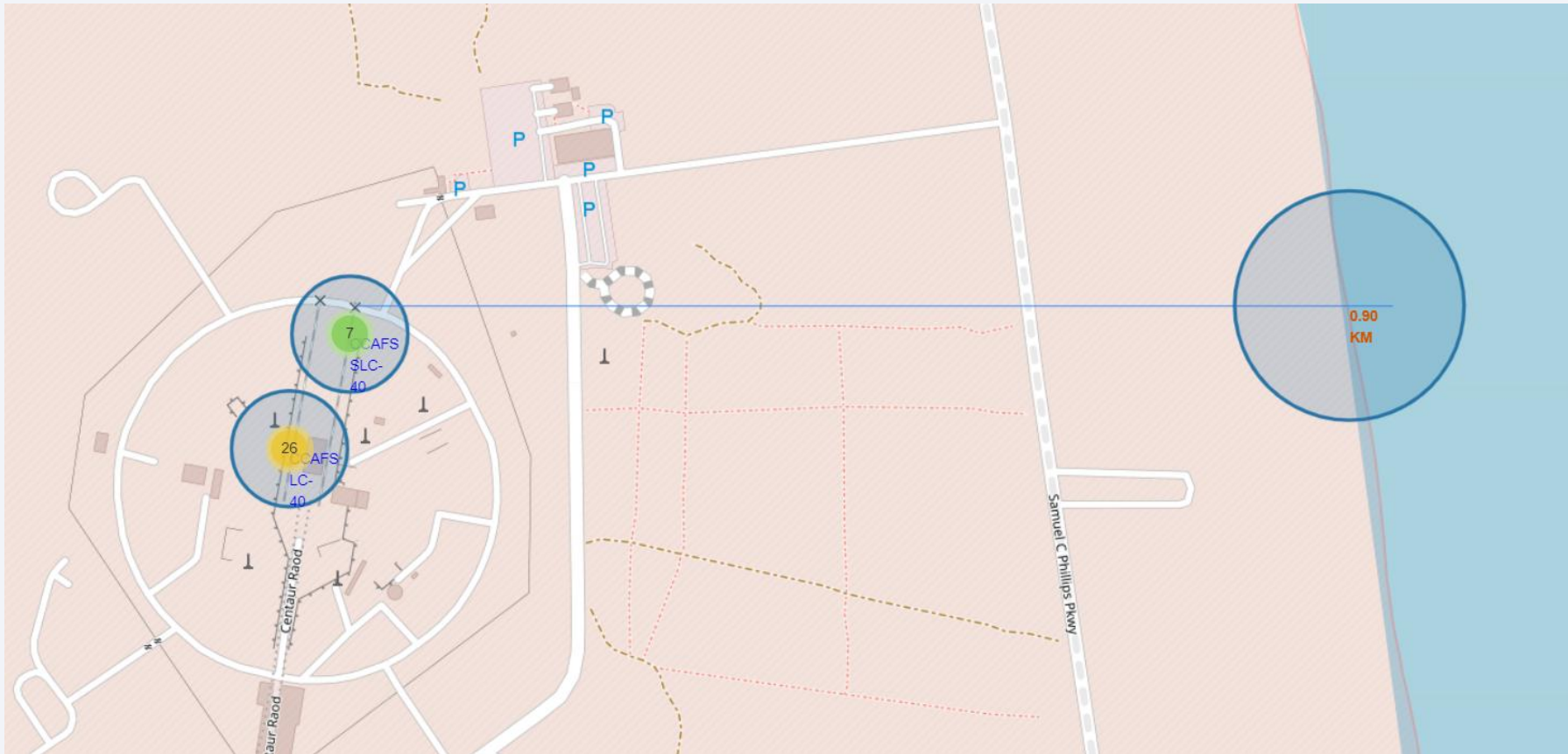
- Blue circles represent each launch site in the U.S.A.

Launch outcomes for each site



- Yellow circles represent the cluster markers on each launch;
- California's site have 10 launches marked;
- Florida's sites have a total of 46 launches marked;

Distance between CCAFS SLC-40 to the Atlantic



- Blue line connects the 0.9 Km of distance between CCAFS SLC-40 and the ocean;
- Most unsuccessful landings are controlled and planned to land in the oceans;



Section 5

Build a Dashboard with Plotly Dash

Launch success count for all sites

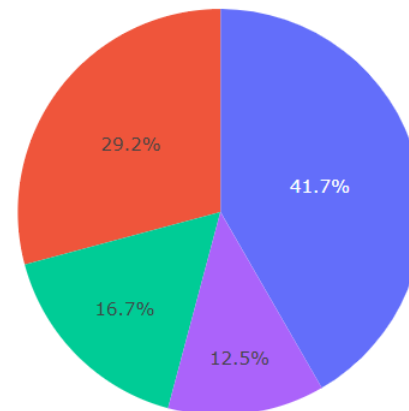
- KSC LC-39A has a total of 10 successful flights, the highest count, representing 41.7% of the total successful launches
- 2nd success count happened on CCAFS LC-40
- The least successful site is CCAFS SLC-40

SpaceX Launch Records Dashboard

All Sites



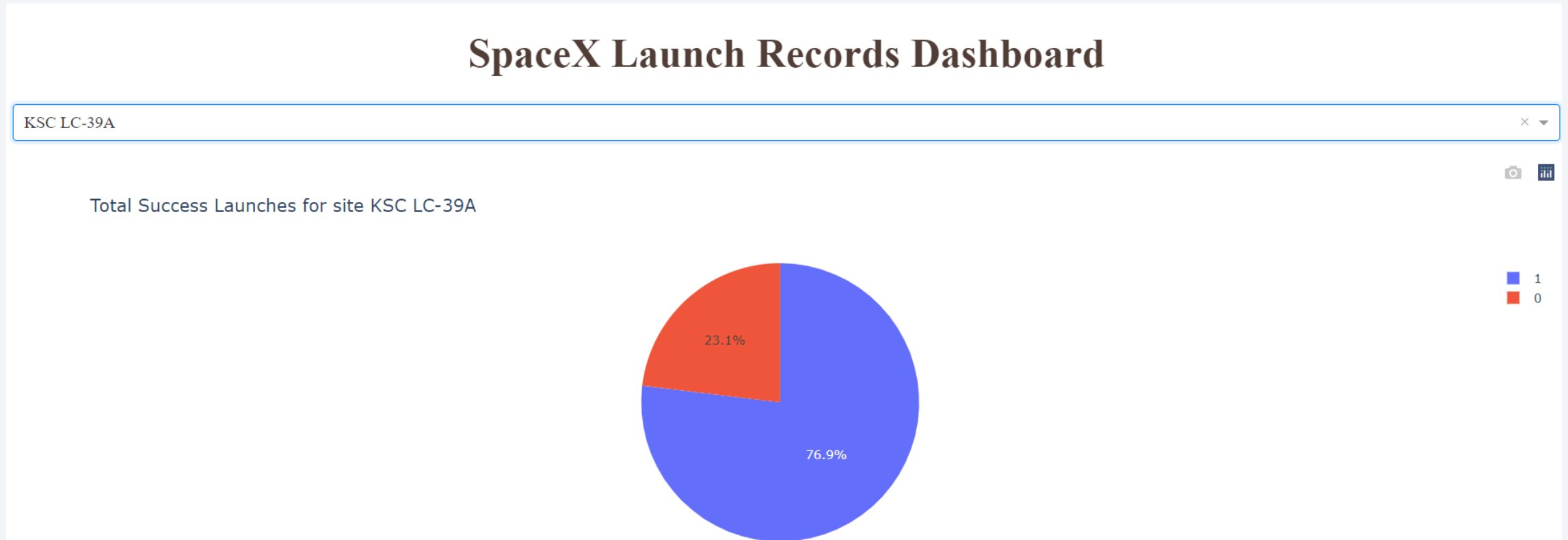
Success Count for all launch sites



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

Launch site with highest launch success ratio

- KSC LC-39A has the highest successful rate of 76.9%



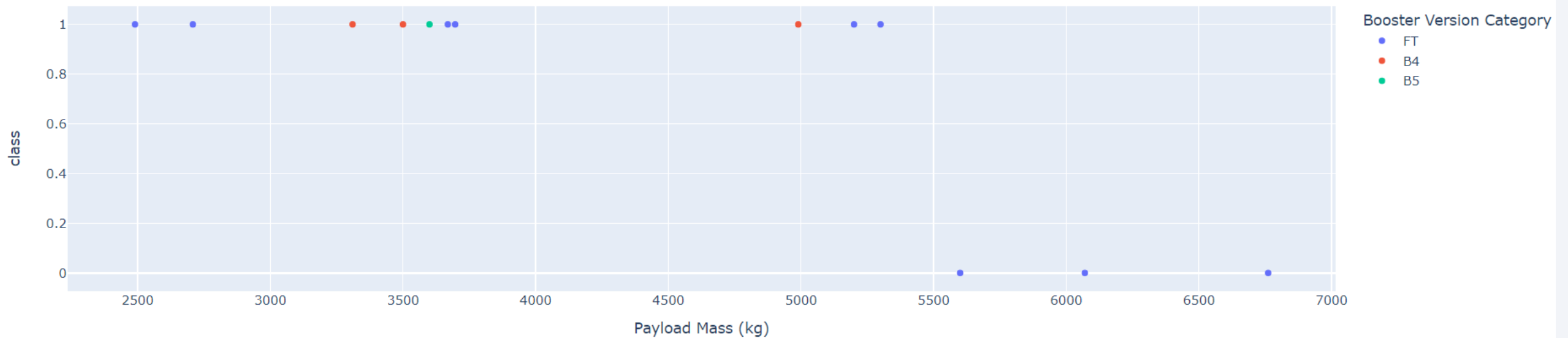
Payload vs. Launch Outcome scatter plot for all sites

- FT booster versions on larger payloads were more likely to fail

Payload range (Kg):



Success count on Payload Mass for KSC LC-39A





Section 6

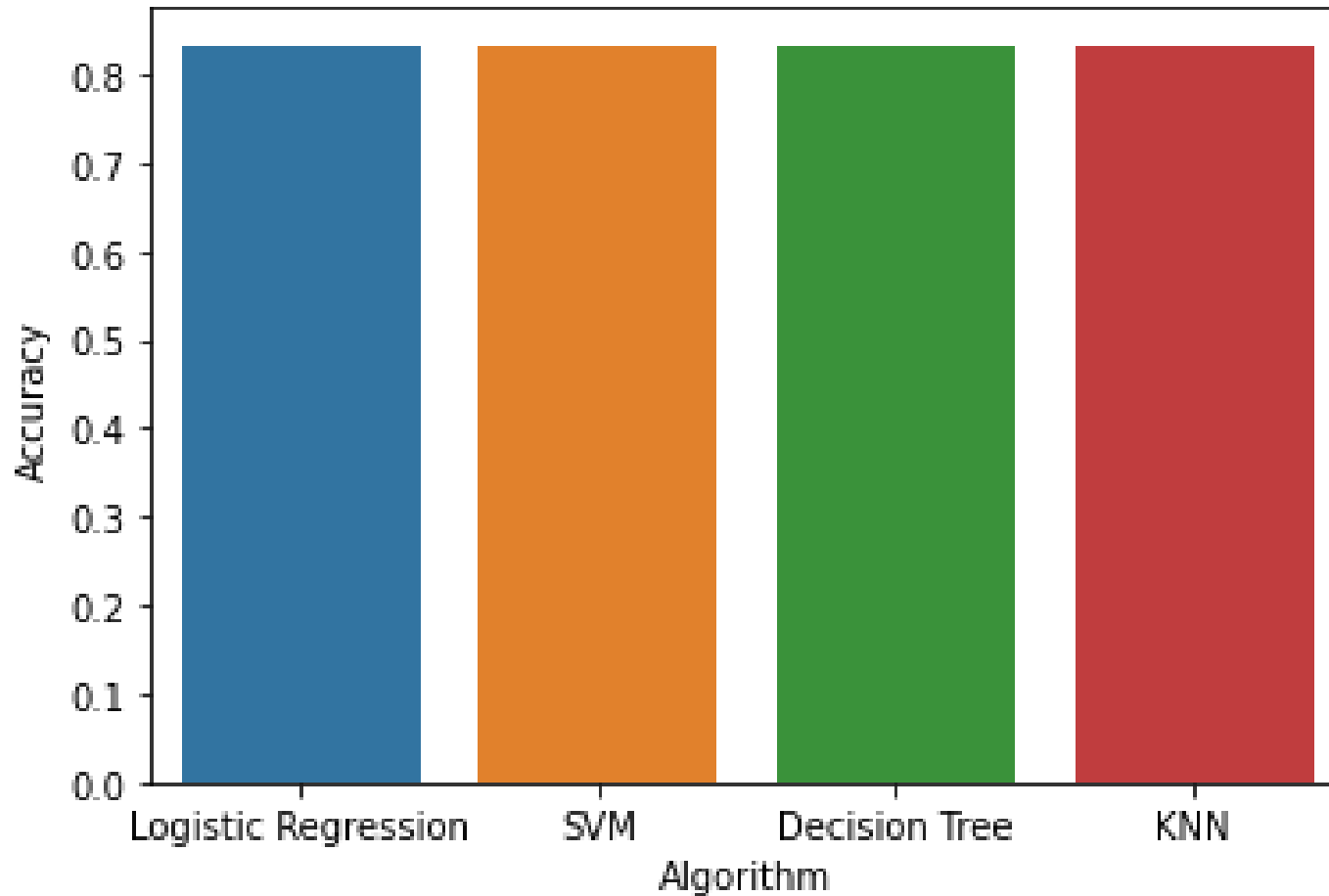
Predictive Analysis (Classification)

Classification Accuracy

Algorithm	Jaccard	F1-score	LogLoss
Logistic Regression	0.800000	0.814815	0.478667
SVM	0.800000	0.814815	N/A
Decision Tree	0.800000	0.814815	N/A
KNN	0.800000	0.814815	N/A

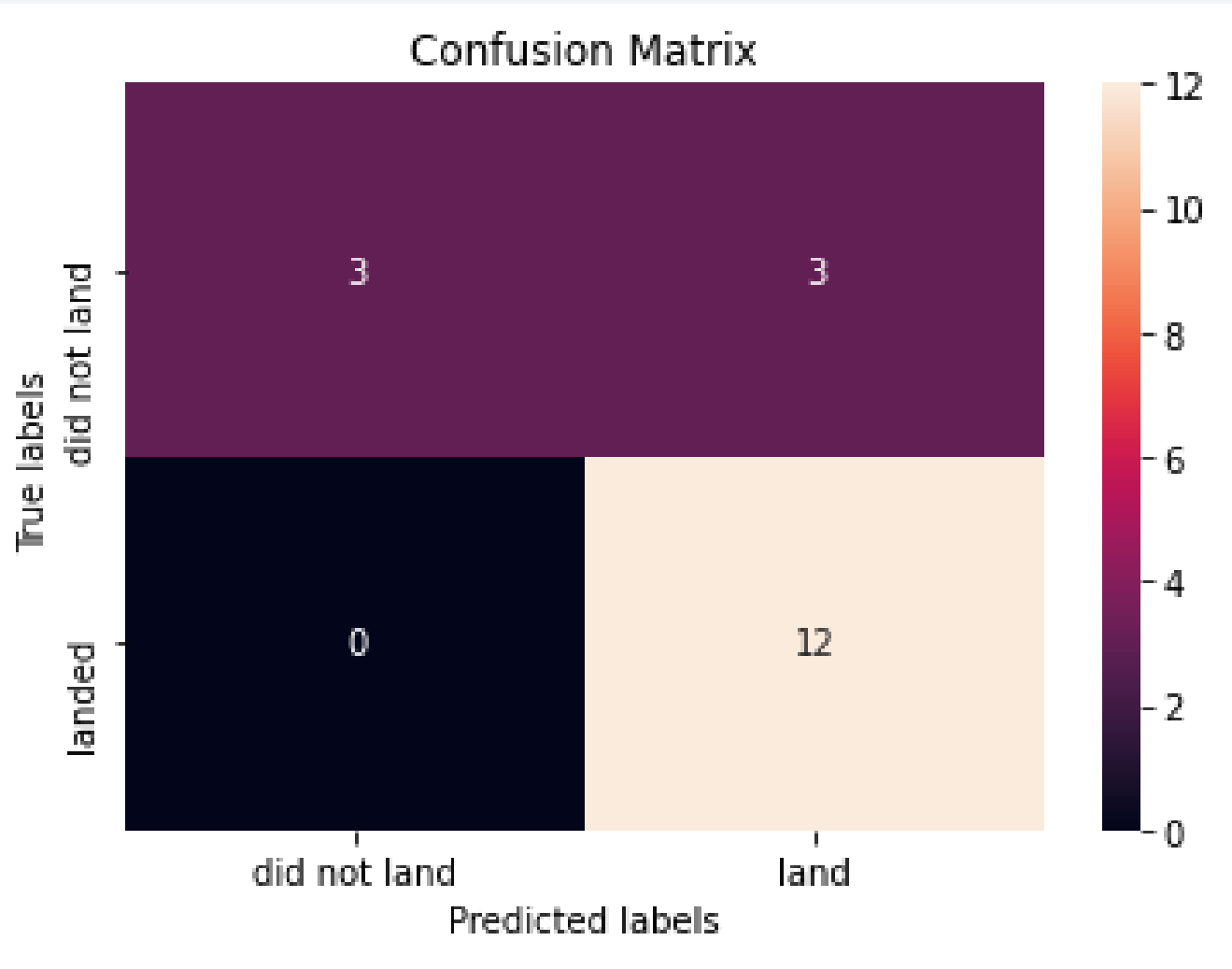
- All models present same test accuracy;
- Jaccard score of 0.8 and F1 score of 0.81 indicate that any of the 4 models are appropriate;

Classification Accuracy



- All models present a test accuracy of >0.83 , indicating that any model trained is appropriate;
- The convergency in accuracy is most likely due to the small dataset;

Confusion Matrix



- The 4 models presented the same Confusion Matrix as well;
- Out of the 6 flights that did not successfully land, the models predicted 50% correctly;
- Out of the 12 that successfully landed, all of them were correctly classified;
- Thus, these model have higher propensity to commit Type II error;
- Precision = 1
- Recall = 0.5

Conclusions

- Most flights have departed from CCAFS SLC 40, but KSC LC-39A is the site with highest successful rate.
- Different orbits to present conditions that implicate in different levels of challenge at landing. GEO, HEO, ES-L1 and SSO are associated with the most successful rates;
- Overtime, improvements to the Falcon 9 have been made and new versions created, which has led to an increase in successful landings;
- The 4 classification models developed to predict whether a flight will successfully land show high performance, F1 score = 0.81, specially when predicting the successfully landing outcome (all 'successful landed' observations were correctly classified).
- Increasing data observations could support improvements and eventually create dissimilarity among the model results;

Appendix

- Python codes:
 - Data collection: <https://github.com/vantoks/Spacex/blob/master/space.ipynb>
 - Webscraping: <https://github.com/vantoks/Spacex/blob/master/webscraping.ipynb>
 - Data wrangling: <https://github.com/vantoks/Spacex/blob/master/wrangling.ipynb>
 - EDA with SQL: <https://github.com/vantoks/Spacex/blob/master/eda-sql.ipynb>
 - Data visualization: <https://github.com/vantoks/Spacex/blob/master/dataviz.ipynb>
 - Interactive folium: <https://github.com/vantoks/Spacex/blob/master/interactive-folium.ipynb>
 - Dashboard: https://github.com/vantoks/Spacex/blob/master/spacex_dash_app.py
 - Prediction: <https://github.com/vantoks/Spacex/blob/master/Prediction.ipynb>

Thank you!

