IBM Developer
SKILLS NETWORK

Winning the Space Race with Data Science

Devid Mazzaferro
January 10, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection was done through web scraping and using SpaceX's API;
  - A first pass of Exploratory Data Analysis followed, with basic data wrangling, data visualization and interactive analytics through the use of a dashboard;
  - A launch site exploration was achieved through the use of Folium maps.
  - Finally, to achieve the results we sought after, we applied Machine Learning algorithms to predict the outcomes of SpaceX's launches.
- Summary of all results
  - The Exploratory Data analysis allowed us to identify the best features to predict the success of a SpaceX launch;
  - We analyzed the importance of a proper launch site with access to commodities and far from cities;
  - Through Machine Learning, we found the best algorithms to predict the success of a launch.

Introduction

- Our objective for this analysis was to evaluate the viability of our company, Space Y as a valid competitor to the market leader SpaceX.

- We tried to find answers to the following questions:
  - What is the best estimator for launch success?
  - Where are the best locations to launch from?
  - Can Machine Learning help us in predicting launch success?

Section 1

Methodology

Methodology

- Data collection methodology:
    - Data for Space X launches was collected from SpaceX's API (https://api.spacexdata.com/v4/launches/past)
    - Success data was web scraped from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Data wrangling:
    - Data wrangling required us to create a landing outcome label called "class" defining good and bad outcomes numerically

Methodology

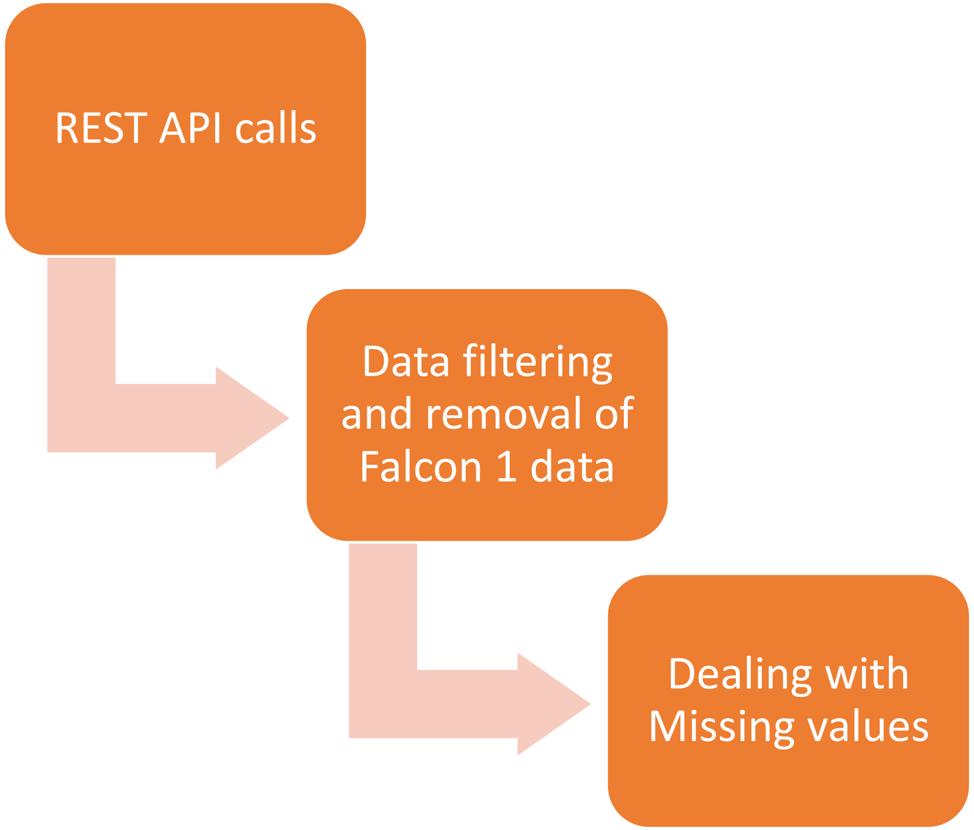
- Exploratory data analysis (EDA):
  - EDA was achieved through data visualization and SQL queries.
- Interactive Visual Analysis:
  - Interactivity was achieved through a Plotly Dash interactive dashboard.
- Predictive Analysis using classification models:
  - Various predictive models were trained and tested in order to find the best model at predicting launch success.

Data Collection

- Data was collected both through public API and Wikipedia:
  - Data for Space X launches was collected from SpaceX's API (https://api.spacexdata.com/v4/launches/past)
  - Success data was web scraped from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX offers publicly available data through its own API;
- We removed data pertaining Falcon 1 launches, as our interest was in Falcon 9's success rate.
- Missing values in Payload Mass were replaced with the mean value of Payload Masses.



REST API calls

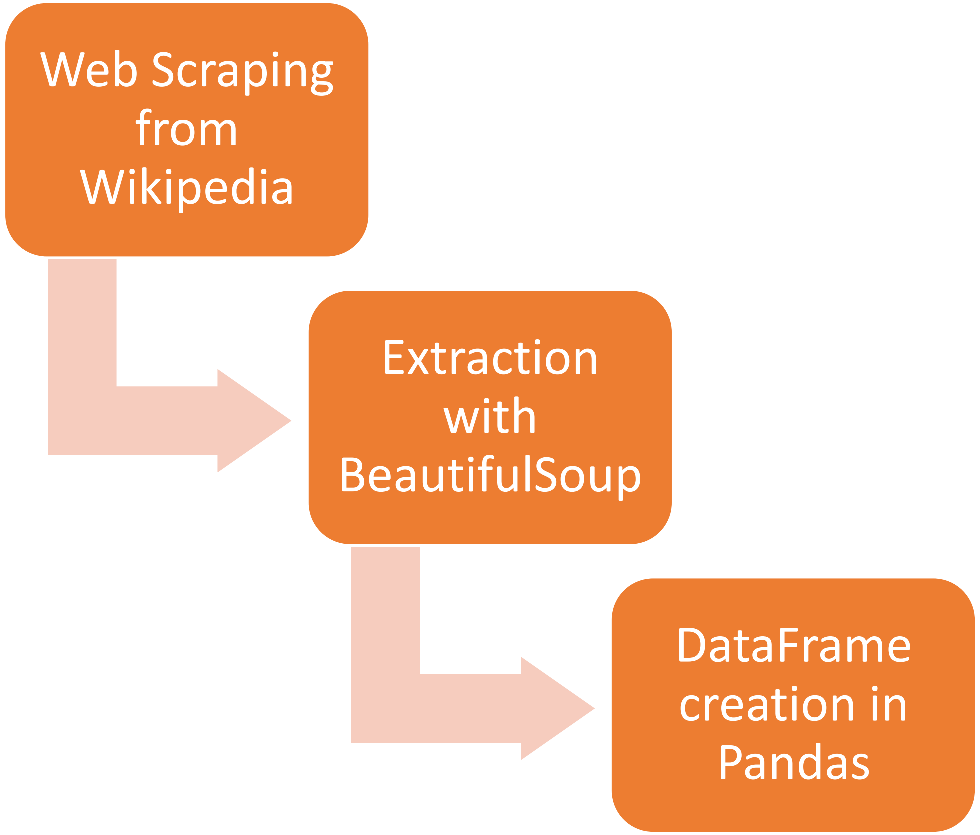
Data filtering and removal of Falcon 1 data

Dealing with Missing values

Complete code: https://github.com/kurasahakai/testrepo/blob/master/Applied%20Data%20Science%20Capstone/Final_notebook1.ipynb

Data Collection – Web Scraping

- Web scraped data was obtained from Wikipedia;
- HTML table data was extracted using BeautifulSoup;
- Dictionaries were compiled to create a complete DataFrame.


Web Scraping from Wikipedia
Extraction with BeautifulSoup
DataFrame creation in Pandas

Complete code: https://github.com/kurasahakai/testrepo/blob/master/Applied%20Data%20Science%20Capstone/jupyter-labs-webscraping.ipynb

Data Wrangling

- Exploratory Data Analysis was performed on the scraped DataFrame;

- Summarization of key quantities:
  - Number of launches per site;
  - Occurrences of different orbital launches;
  - Mission outcome per orbit type;

- A landing outcome label was created from the Outcome column.

EDA with Data Visualization

- To explore basic connections, various scatterplots and bar charts were explored:
  - Flight Number VS Launch Site: we saw Launch Site VAFB SLC 4E has been phased out, whereas Launch Sites CCAFS SLC 40 and KSC LC 39A are still active and successful;
  - Payload VS Launch Site: we found no payload greater than 10 000 kg has been launched from Launch Site VAFB SLC 4E;
  - Success rate of orbit types: HEO and SSO orbits have an enormous success rate, on the flip side, GTO, ISS, PO orbit launches have low success rate. SO, ES-L1, GEO and MEO orbit have only one launch each, not enough to draw conclusions;
  - Flight number VS Orbit type: LEO's orbit success rate has increased with flight number, whereas GTO's success rate seems uncorrelated to flight number.
  - Yearly Launch trend: SpaceX's success has steadily increased throughout the years.

Complete code: https://github.com/kurasahakai/testrepo/blob/master/Applied%20Data%20Science%20Capstone/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Explorative Data Analysis was performed with SQL with these queries:
  - Names of unique launch sites;
  - 5 Launch sites with name beginning in “CCA”;
  - Total Payload carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster F9 v1.1;
  - First ground pad successful landing date;
  - List of boosters which have succeeded in drone ship landing;
  - Total number of successful and failed mission outcomes;
  - Boosters that carried the maximum payload mass;
  - Month, failure landing outcome, booster version and launch sites for the year 2015;
  - Successful landings between 04/06/2010 and 20/03/2017

Complete Code: https://github.com/kurasahakai/testrepo/blob/master/Applied%20Data%20Science%20Capstone/jupyter-labs-eda-sql-coursera_sqllite(3).ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used in Folium maps to do some launch site location analysis:
  - Markers were used to indicate locations;
  - Circles were used to highlight areas around specific coordinates;
  - Marker clusters were used to indicate events, such as launches, these were color coded to represent failed and successful launches;
  - Lines were used to draw distances between locations.

Complete code: https://github.com/kurasahakai/testrepo/blob/master/Applied%20Data%20Science%20Capstone/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A Plotly Dash interactive dashboard was created with the following elements:
  - A dropdown menu to select between different Launch Sites or All Launch Sites;
  - A pie chart to indicate success rate per selected Launch Site;
  - A slider to select launches by payload;
  - An interactive scatter plot with data per Launch Site and filtered by payload.

Complete Code: https://github.com/kurasahakai/testrepo/blob/master/Applied%20Data%20Science%20Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

- Different classification models were compared to find the best:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbour

Complete Code: https://github.com/kurasahakai/testrepo/blob/master/Applied%20Data%20Science%20Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results:
  - SpaceX launches from 4 different launch sites;
  - Throughout the years, the success rate for launches has increased;
  - The first successful landing happened in 2015;

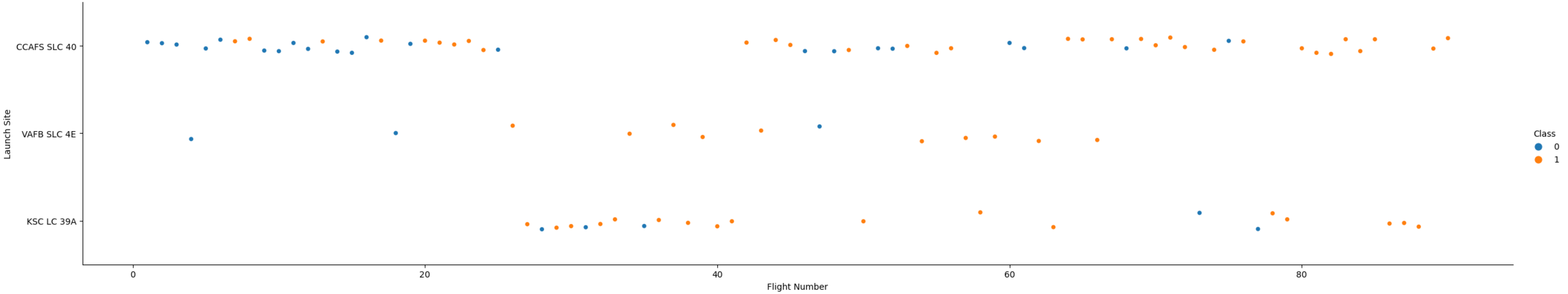
- Predictive analysis results:
  - The best predictor for landing success in the Decision Tree model.

Section 2

Insights drawn from EDA

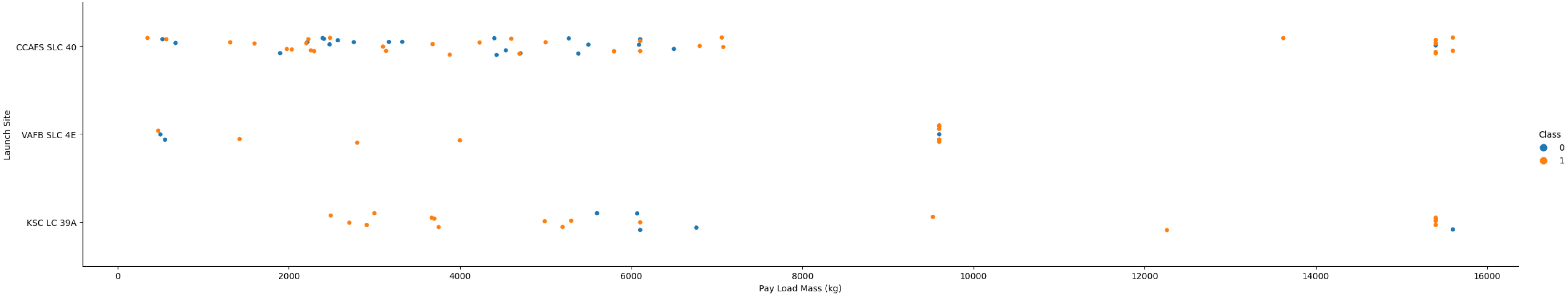
Flight Number vs. Launch Site

- Launch Site VAFB SLC 4E has been phased out;
- Launch Sites CCAFS SLC 40 and KSC LC 39A continue operations with great results.



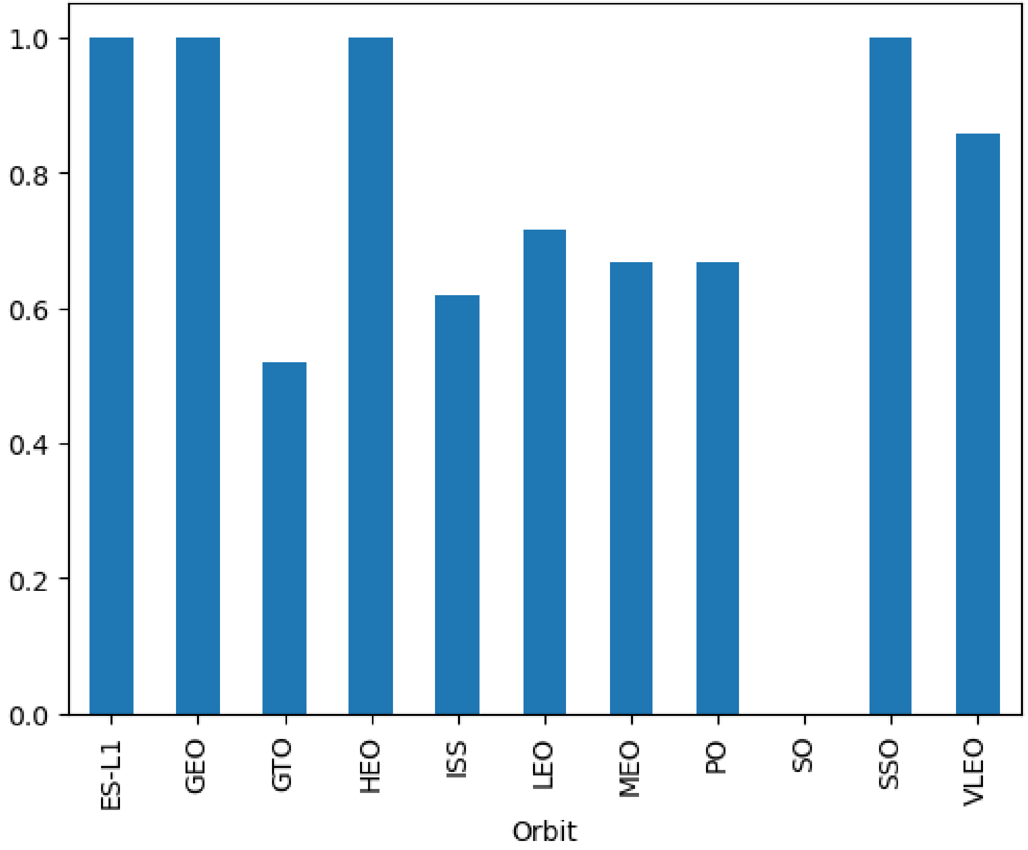
Payload vs. Launch Site

- No launch over 10 000 kg of payload has been done from Launch Site VAFB SLC 4E



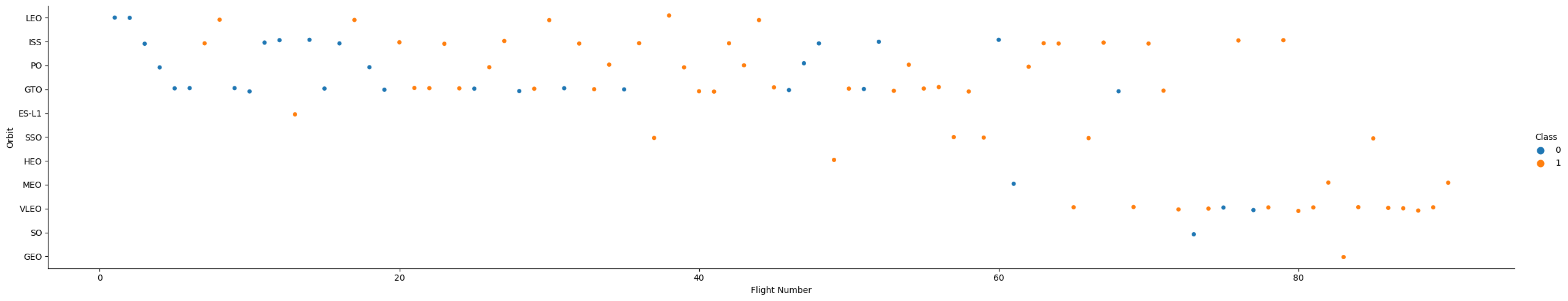
Success Rate vs. Orbit Type

- HEO and SSO orbits have an enormous success rate;
- On the flip side, GTO, ISS, PO orbit launches have low success rate.
- SO, ES-L1, GEO and MEO orbit have only one launch each, not enough data to draw conclusions;



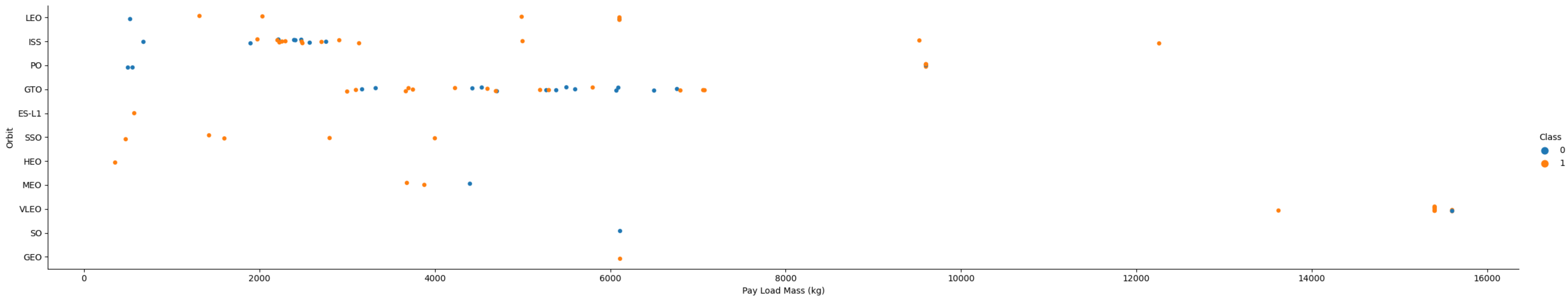
Flight Number vs. Orbit Type

- In LEO orbit launches, the success seems to depend on the number of flights;
- In GTO orbit launches this correlations seems to not be important.



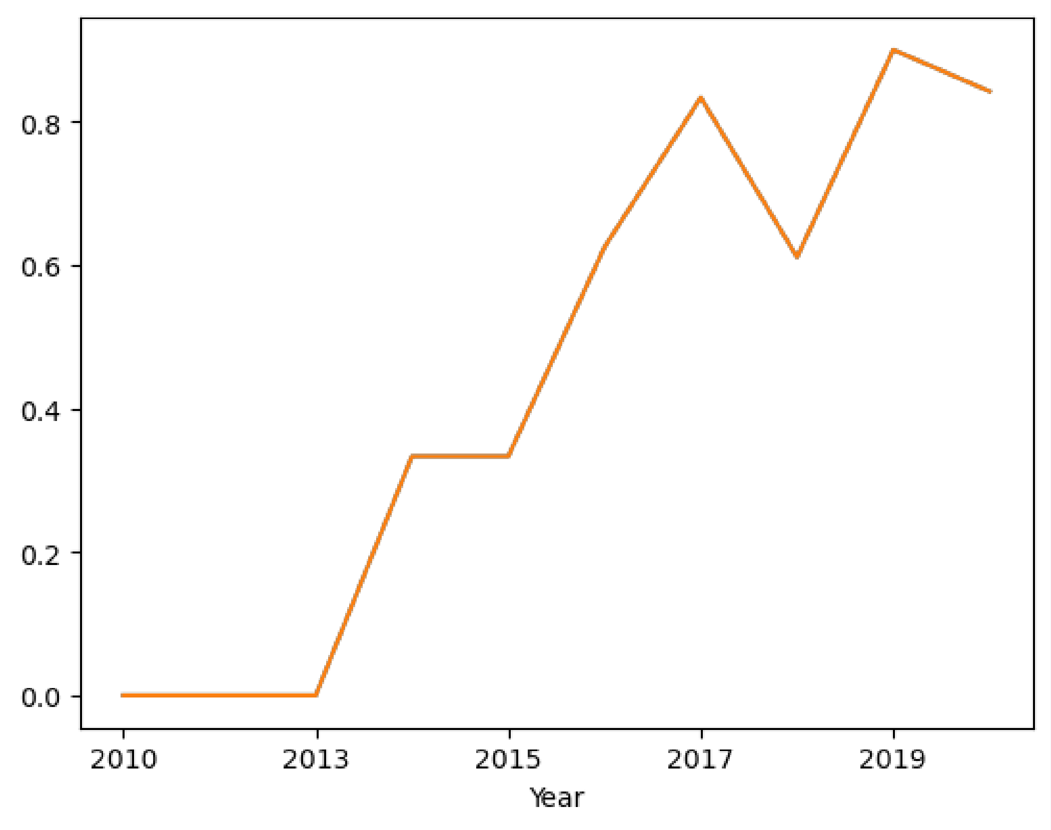
Payload vs. Orbit Type

- When looking at successful landings, Polar LEO and ISS orbits are best;
- In GTO orbits it’s hard to draw conclusions on success rates.
- VLEO orbits seem to be heavily skewed towards heavy payloads (10 000+ kg)



Launch Success Yearly Trend

- The trend for success for launches has been upwards since 2013.
- In 2019 there was a peak of successful landings at over 90%.

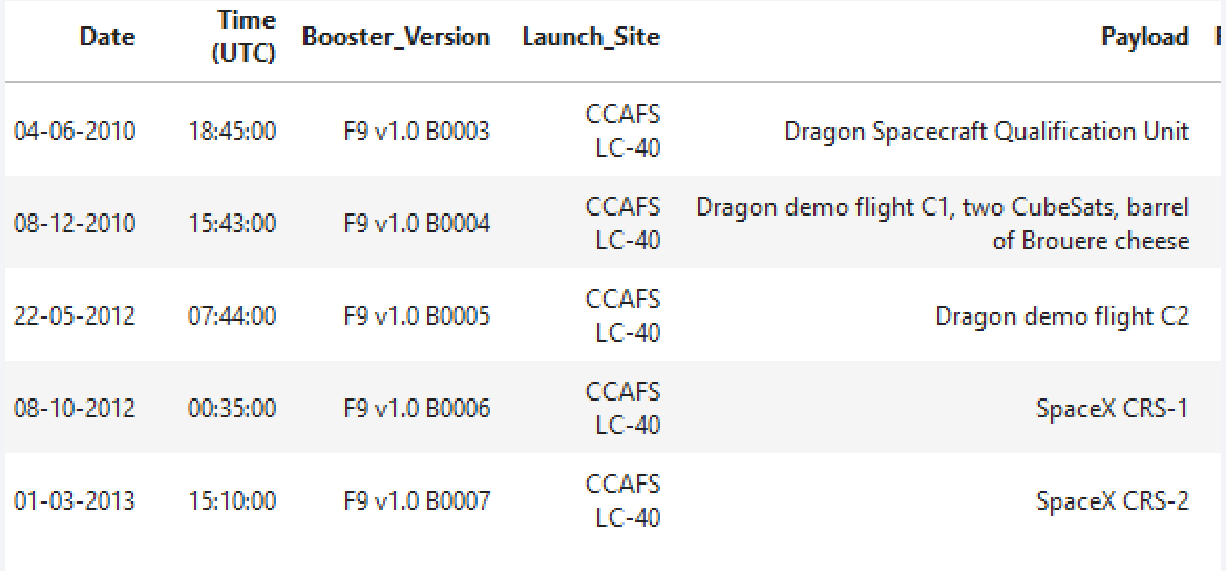


All Launch Site Names

- According to our databases, there are four Launch Sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- These were queried using:
  - %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL

Launch Site Names Begin with 'CCA'

- 5 records where launch sites began with “CCA”.
- These are Cape Canaveral sites.
- They were queried using:
  - %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE “CCA%” LIMIT 5



Date	Time (UTC)	Booster_Version	Launch_Site	Payload
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

Total Payload Mass

- The total payload carried by boosters from NASA (CRS) was 48213 kg;
- This was queried using:
    - %sql SELECT SUM(PAYLOAD_MASS__KG_ AS TOT_PL FROM SPACEXTBL WHERE “Customer” LIKE “%CRS%”

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.
- This was queried using:
  - %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PL WHERE BOOSTER_VERSION = “F9 v1.1”

First Successful Ground Landing Date

- The first successful landing on Ground pad was achieved on 22-12-2015.
- This was obtained querying:
  - %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE “Landing _Outcome” = “Success (ground pad)”

Successful Drone Ship Landing with Payload between 4000 and 6000

- Four boosters landed on a droneship with payload mass between 4000 kg and 6000 kg:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- These were obtained querying:
  - %sql SELECT “Booster_Version” FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND “Landing _Outcome” = “Success (drone ship)”

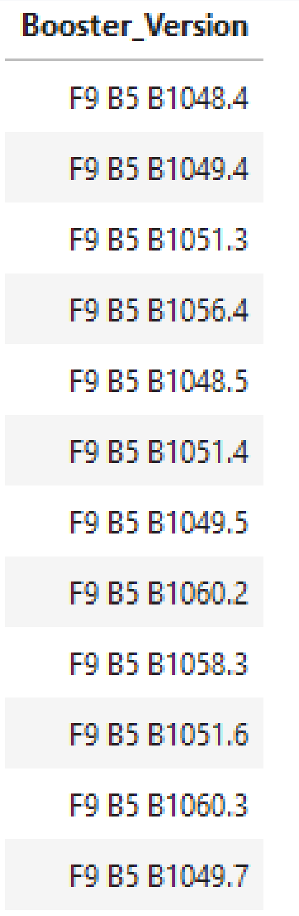
Total Number of Successful and Failure Mission Outcomes

- There were a total of 99 Successes, 1 Success with payload status unclear and 1 failure in flight.
- These were queried with:

  -
    ```
    %sql SELECT MISSION OUTCOME, COUNT(*) AS TOT FROM SPACEXTBL GROUP BY MISSION_OUTCOME
    ```

Boosters Carried Maximum Payload

- Plenty of boosters carried the maximum payload mass.
- These were queried with:
  - %sql SELECT DISTINCT “Booster_Version” FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)



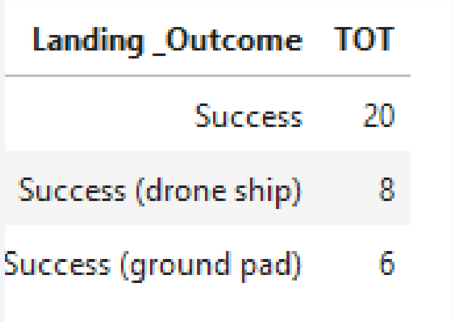
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- There were two failed landing outcomes on drone ships in 2015, their booster version and launch sites are
  - F9 V1.1 B1012 – CCAFS LC-40
  - F9 V1.1 B1015 – CCAFS LC-40
- These were queried with:
  - %sql SELECT substr(Date,4,2) as month, “Landing _Outcome”, LAUNCH_SITE FROM SPACEXTBL WHERE substr(Date,7,4) = “2015” AND “Landing _Outcome = “Failure (drone ship)”

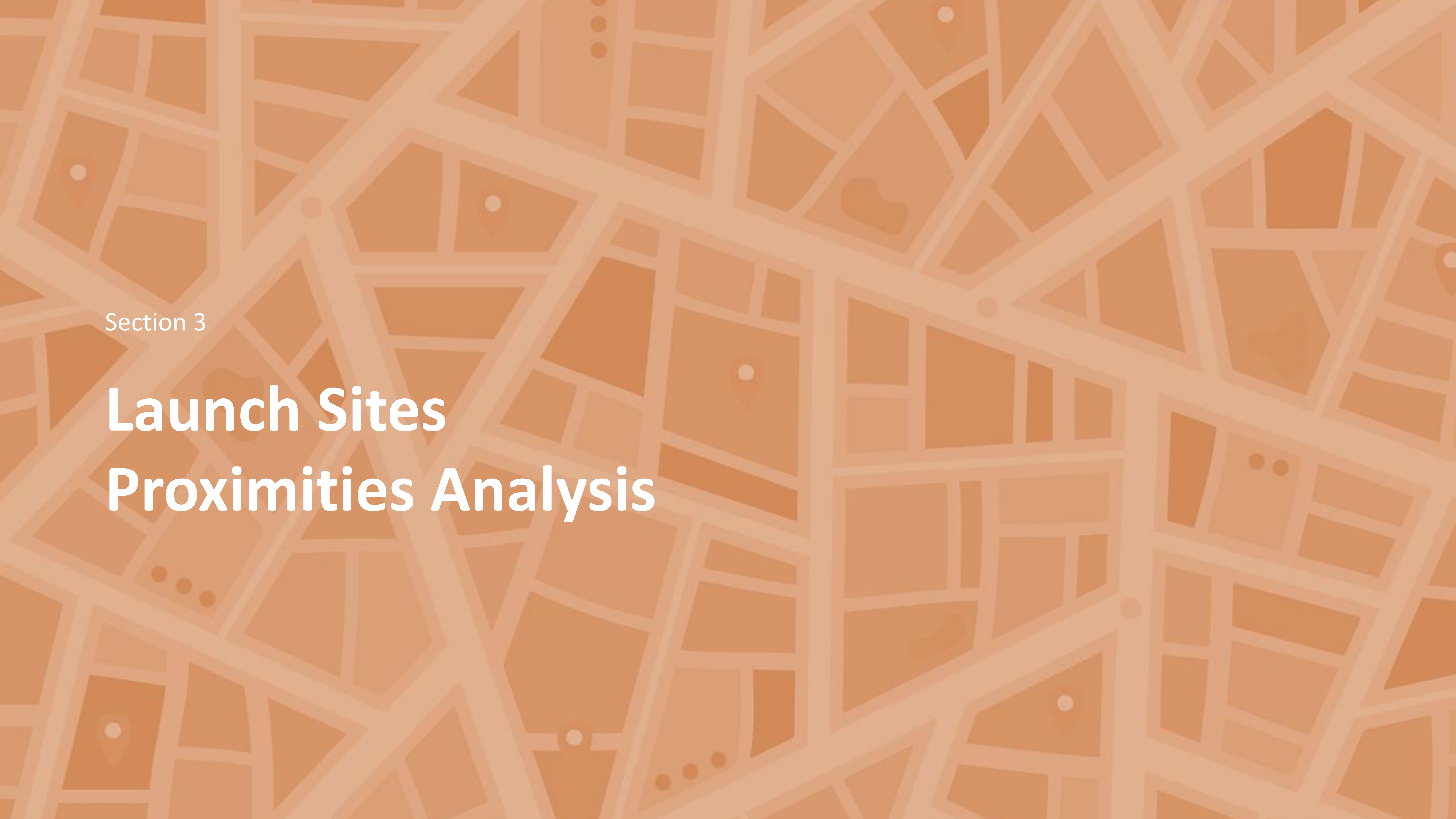
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Ranking of successful landing outcomes between 04-06-2010 and 20-03-2017 in descending order:



Landing _Outcome	TOT
Success	20
Success (drone ship)	8
Success (ground pad)	6

- These were queried with:
  - %sql SELECT “Landing _Outcome”, COUNT(*) AS TOT FROM SPACEXTBL WHERE DATE BETWEEN “04-06-2010” AND “20-03-2017” AND “Landing _Outcome” LIKE “%Success%” GROUP BY “Landing _Outcome” ORDER BY TOT DESC

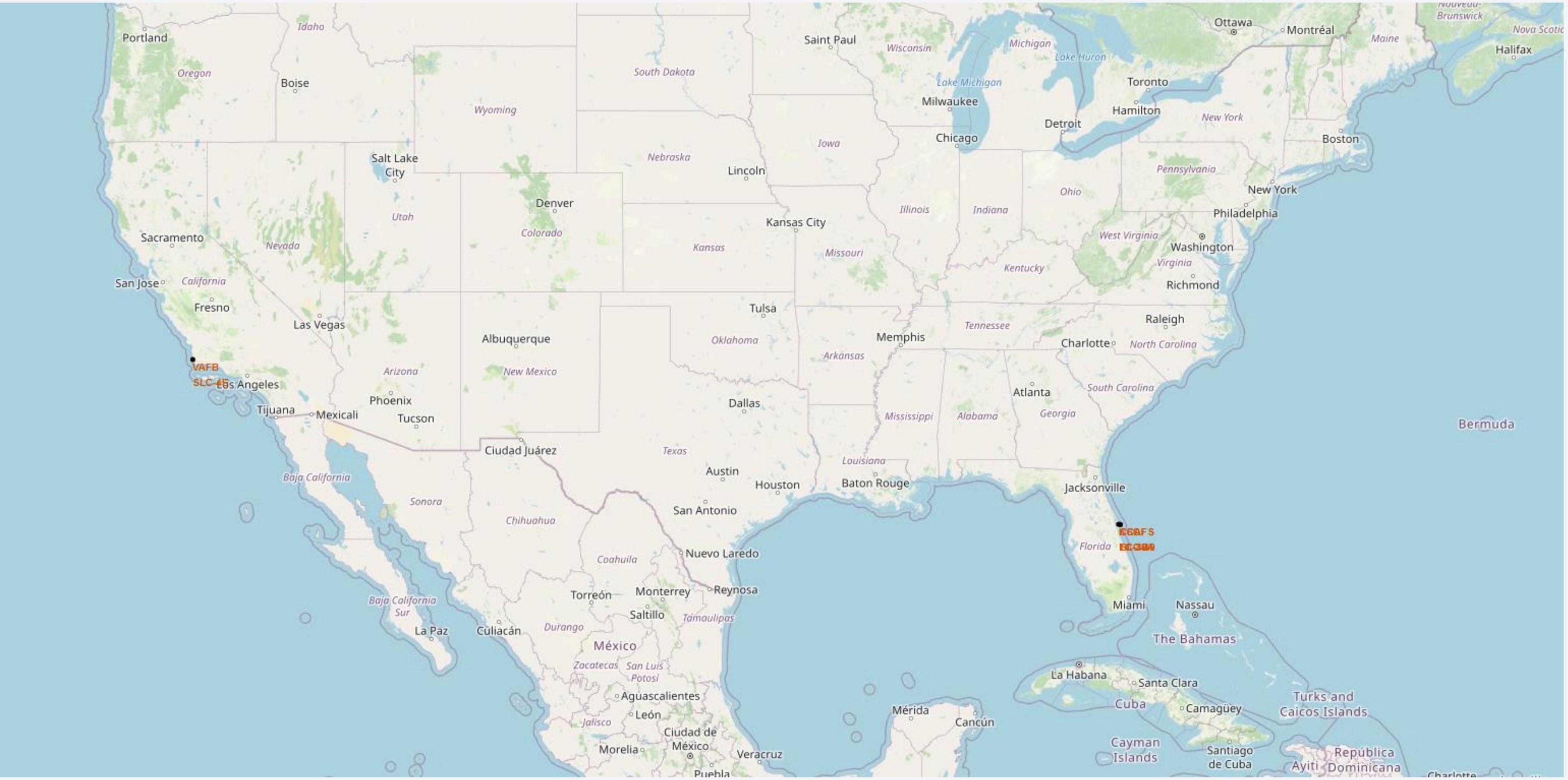


Section 3

Launch Sites Proximities Analysis

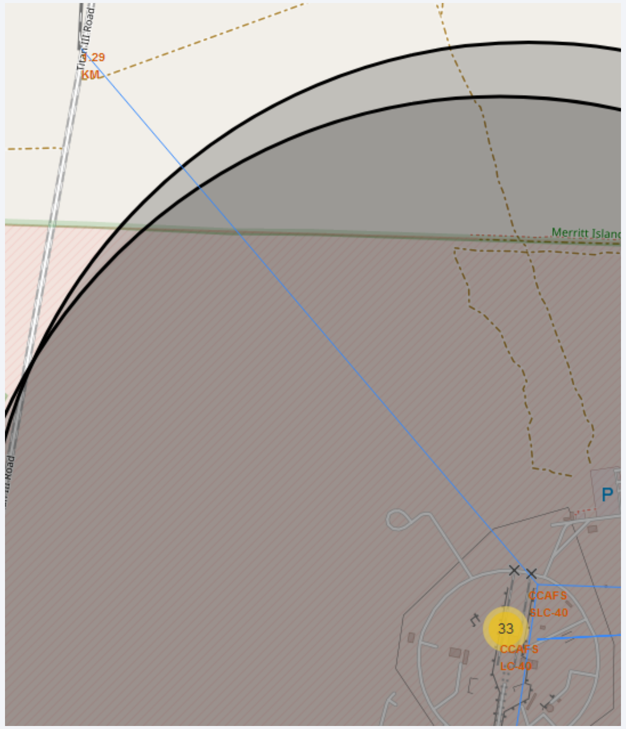
Launch Site proximity to coastlines

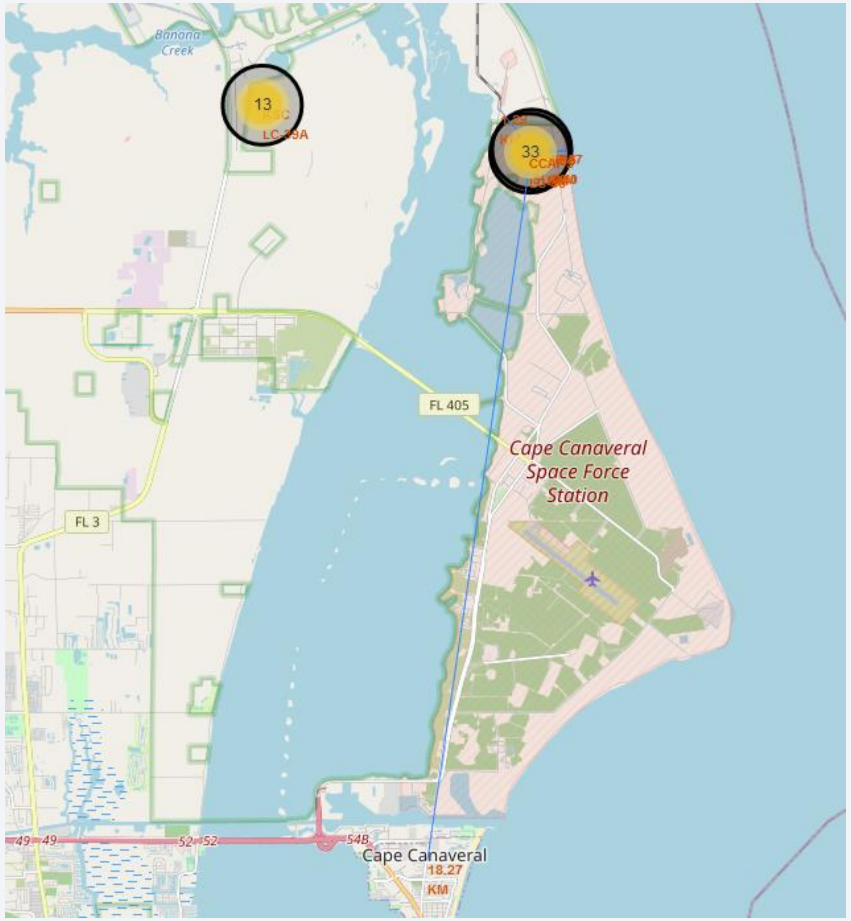
- All Launch Sites are near a coastline



Launch Site proximity to railways and distance from cities

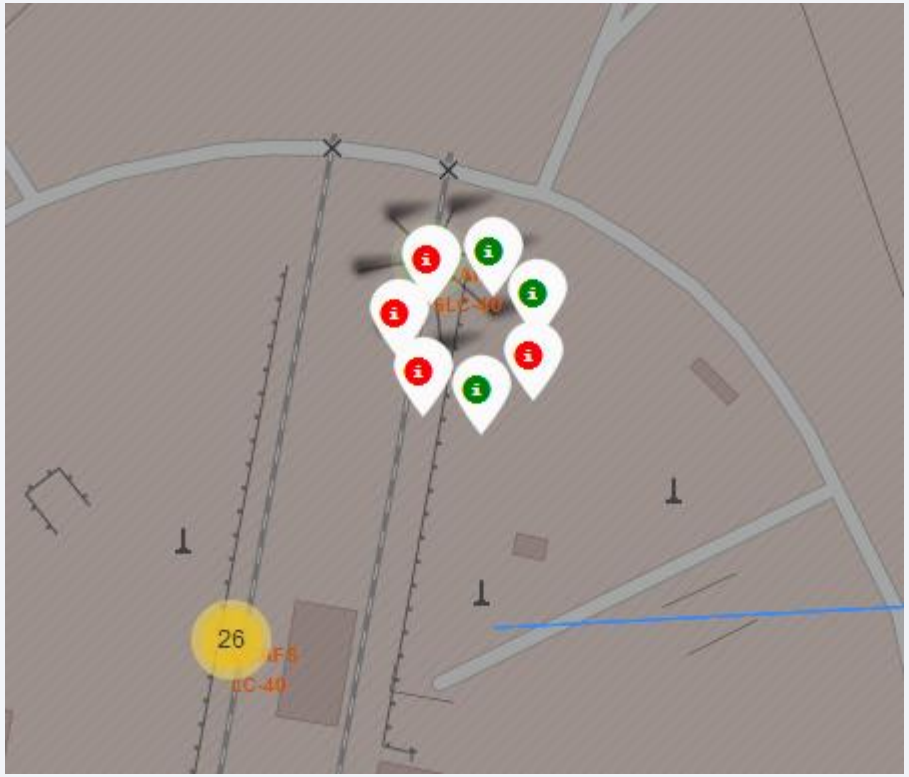
- Another important proximity is railways, this is useful to move equipment in and out of the Launch Site.
- Cities tend to be relatively far from Launch Sites.

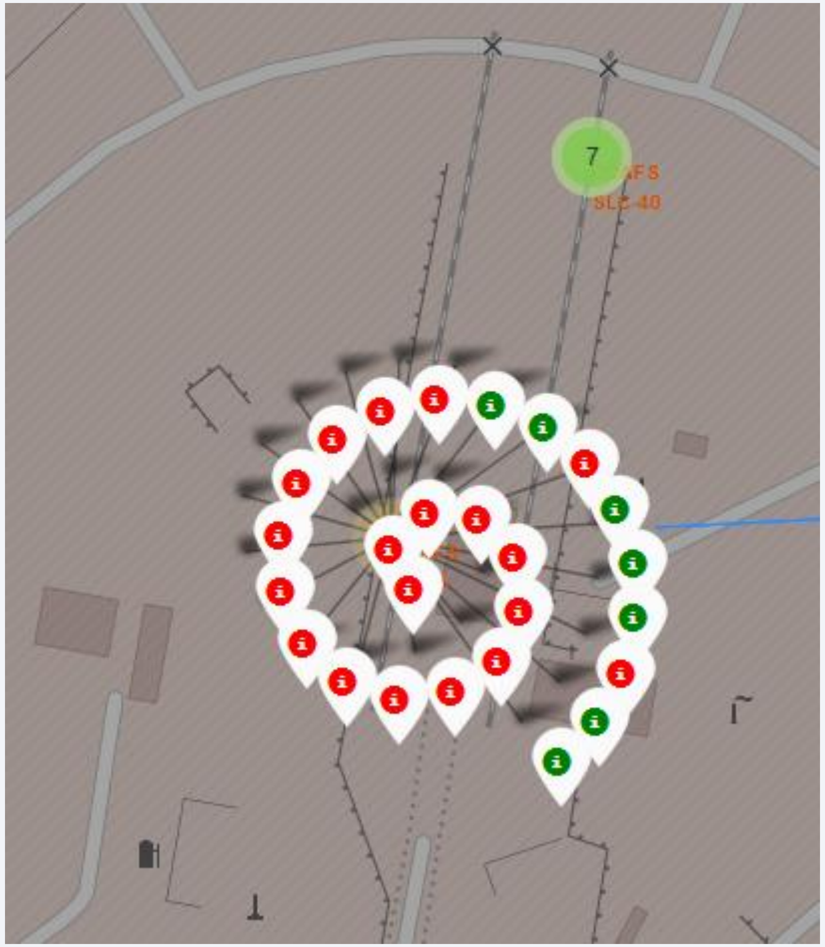




Success Rate Color Coding

- Launch results were color coded: green means it was successful, red that it was unsuccessful.

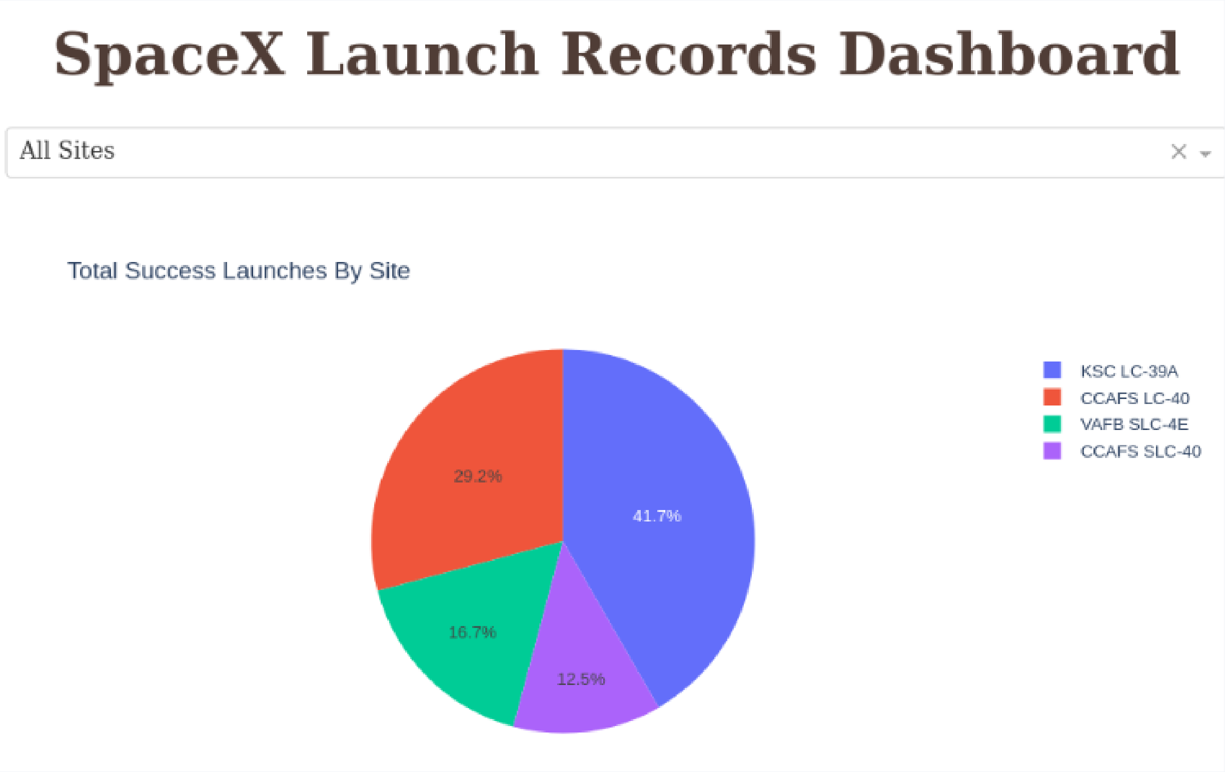




Section 4

Build a Dashboard With Plotly Dash

Success of Launches by Site



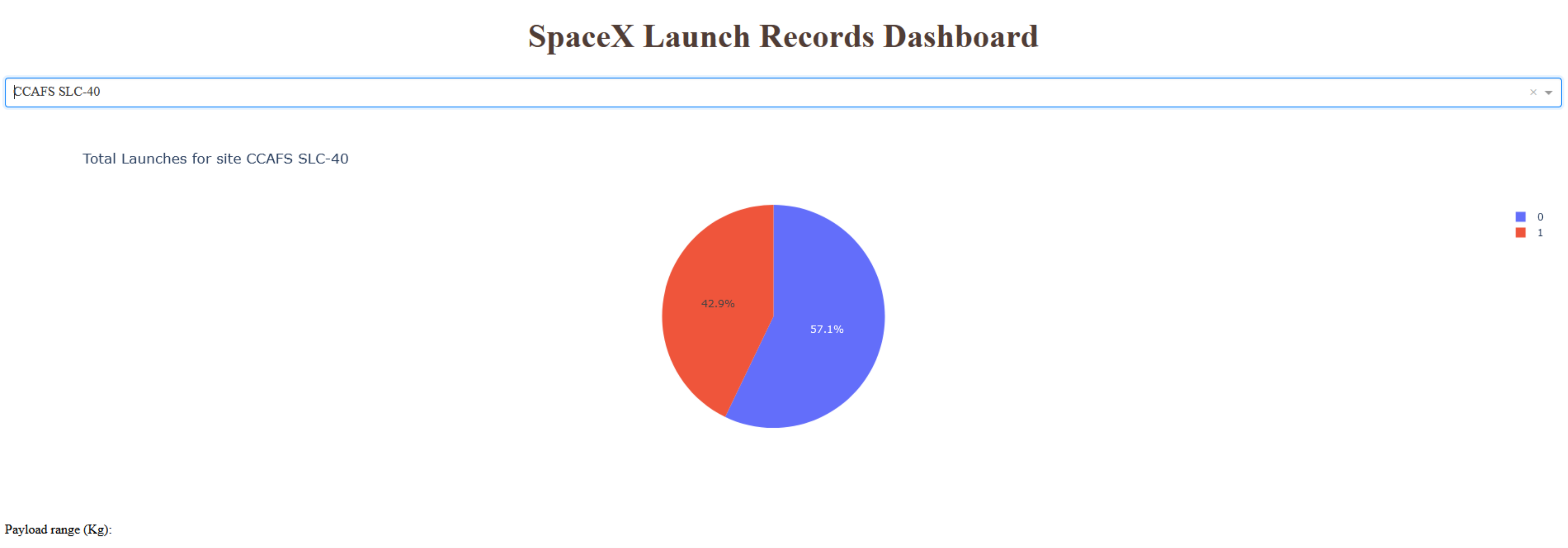
SpaceX Launch Records Dashboard

All Sites

- Launch Site seems to be a huge factor in launch success, with KSC LC-39A with over 76% of success rate.

Success Rate for site

- CCAFS SLC-40 has a 57% failure rate.



SpaceX Launch Records Dashboard

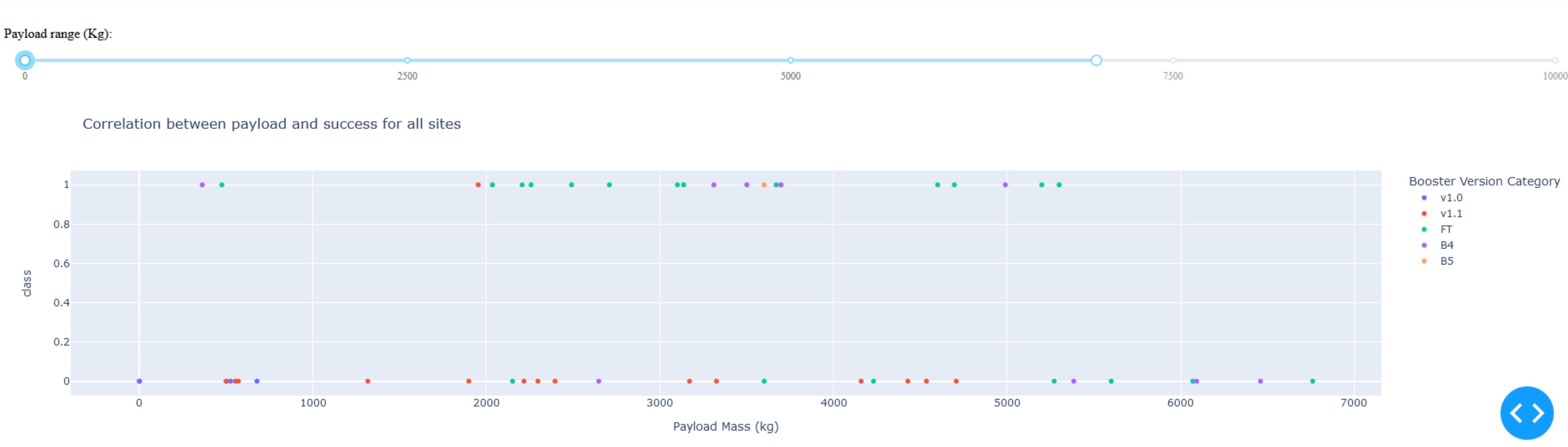
CCAFS SLC-40

Total Launches for site CCAFS SLC-40

Payload range (Kg):

FT Boosters’ Success Rate

- FT boosters have and enormous success rate

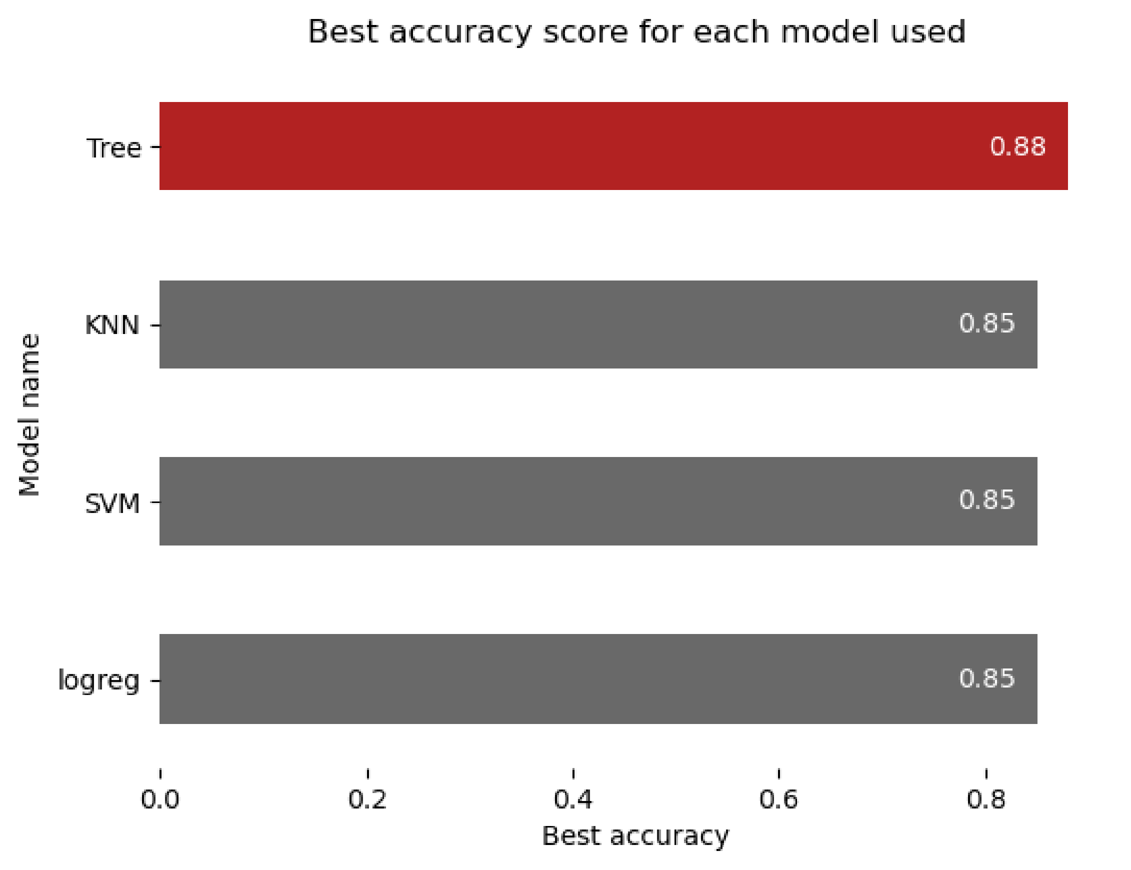


Section 5

Predictive Analysis (Classification)

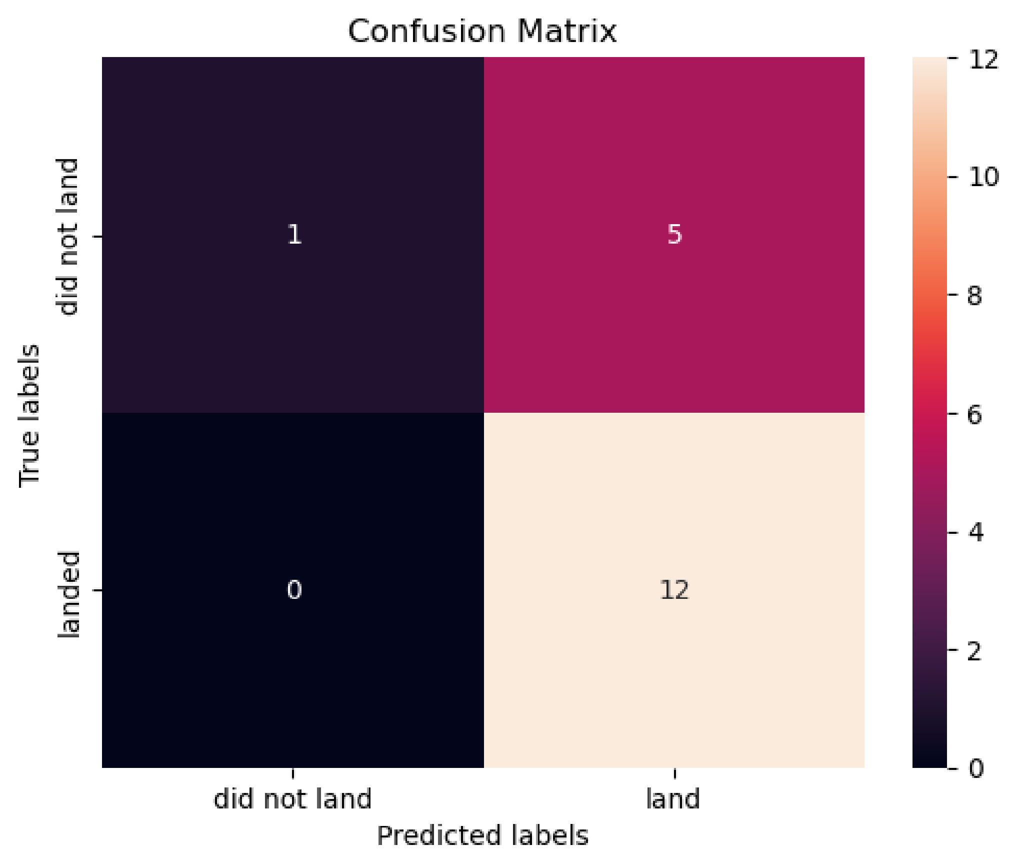
Classification Accuracy

- The model that yielded the best accuracy score was the Decision Tree



Confusion Matrix

- Its Confusion Matrix yielded the best prediction for True Positives



Conclusions

- The best launch site appears to be KSC LC-39A;
- CCAFS SLC-40 has a 57% failure rate;
- In case of new Launch Site exploration, coastal sites with access to railroads and highways are to be preferred.
- FT Boosters seem to work very well below 6000 kg of payload;
- The only way to succeed seems to be try and try again, as proved by the meteoric rise in success after a rocky start from SpaceX;
- Decision Trees classifiers seem to be the most apt at describing the success of a mission.

Appendix

- All the relevant code can be found at: https://github.com/kurasahakai/testrepo/tree/master/Applied%20Data%20Science%20Capstone

- Cover/last page art designed by pikisuperstar / Freepik
- Section 3 art designed by pikisuperstar / Freepik
- Section 4 art designed by starline / Freepik

Thank you!