IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Karolien Driesen>
<7/08/2024>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Aim

- Building a machine learning model to predict if SpaceX Falcon 9 first stage will land successfully

Methodology

- **Data collection**

  Data on Falcon 9 (F9) launches was gathered using SpaceX Rest API and web scraping.

- **Data wrangling**

  Data were converted to a Pandas dataframe and transformed into a clean and meaningful dataset.

Executive Summary

Methodology (continued)

- **Exploratory data analysis**

  Data were explored to gain insight into underlying patterns and trends in landing the first stage of F9 with graphs and SQL.

- **Interactive visual analysis**

  - Launch sites and landing outcome of F9 first stage were marked on a map using Folium.
  - An interactive dashboard was built to further gain insight in the relationship between launch sites, booster version, payload and succesrate of landing the F9 rocket with Plotly Dash.

Executive Summary

Methodology (continued)

- **Predictive analysis**

  Classification models to predict the landing outcome of F9 first stage were built and evaluated.

Results

- **Exploratory data analysis**
  - Landing successrate increases with the number of flights.
  - No heavy payloads (>10000 kg) were launched from launch site VAFB-SLC.

Executive Summary

Results (continued)

- **Exploratory analysis**
  - ES-L1, GEO, and HEO have a 100% success rate in landings, while other orbits like LEO and GTO show variable success rates.
  - Success in LEO seems to be related to the number of flights; while in GTO orbit, there is no clear relationship between the number of flights and landing successrate.
  - Heavy payloads are more successful in Polar, LEO and ISS.
  - Success rate have rised since 2013 to 2020.

Executive Summary

Results (continued)

- **Exploratory analysis**
  - Early boosterversions (v1.0 and v1.1) have a lower succesrate in landing first stage.
  - Succesrate of landings and payload capacity increases with later booster versions (B5 version has 100% succesrate).
  - The technological advancements of SpaceX's Falcon 9 rocket are clearly visible. Each new version brought improvements in both, reusibility and payload capacity, which has significantly reduced the cost of acces to space.

Executive Summary

Results (continued)

- **Predictive Analysis**
  - Highest classification accuracy is 83% with Logistic Regression, SVM and KNN.
  - The Logistic Regression model seems more reliable for new data compared to SVM and KNN, which may overfit the training data.
  - Further refinement and improvement of the model is possible with new data from SpaceX

Introduction

Project background and context

- Commercial Space Industry: SpaceX is setting benchmarks in launch efficiency and cost-effectiveness.
- Space Y's Vision: if we can predict the success of landing the first stage of Falcon 9, we can determine the cost of a launch and compete with SpaceX

Questions to answer

- What factors impact the successful landing of SpaceX's F9 first stage?
- Can we accurately predict if the first stage of F9 rockets will land successfully, based on available data?

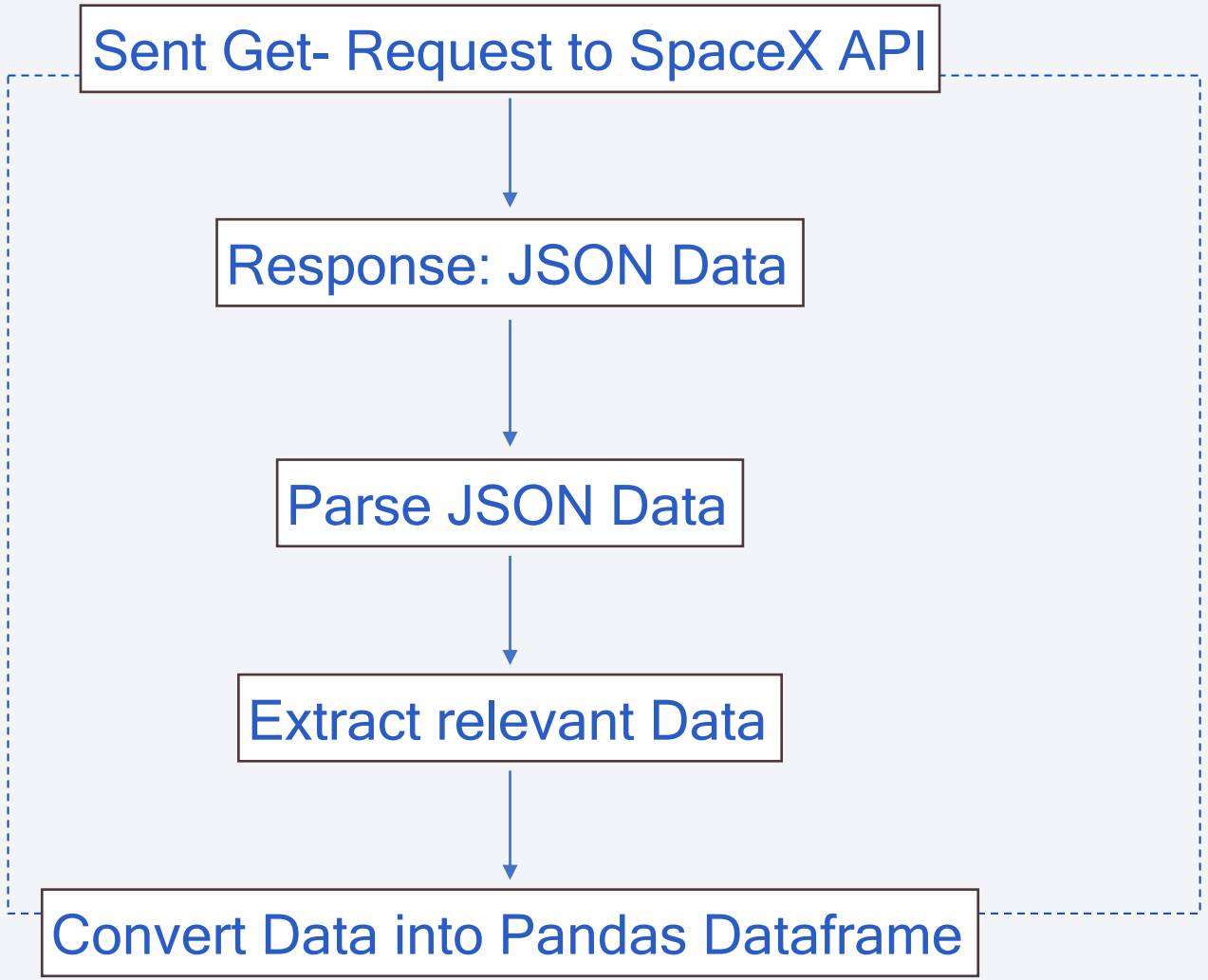
Section 1

Methodology

Methodology

- For each past Falcon 9 launch, data was gathered using the SpaceX Rest API (api.spacexdata.com/v4/) with various endpoints (‘Past Launches', 'Capsules', and ‘Cores’)
- Data from more historical Falcon 9 Launch Records were collected from a Wikipedia page using web scraping with Beautifulsoup (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Data relevant in predicting a successful or failed landing were extracted and converted into a Pandas Dataframe. Included in the dataframe were variables such as: flight number, launch date, orbit type, booster version, payload, launch pad and landing outcome.

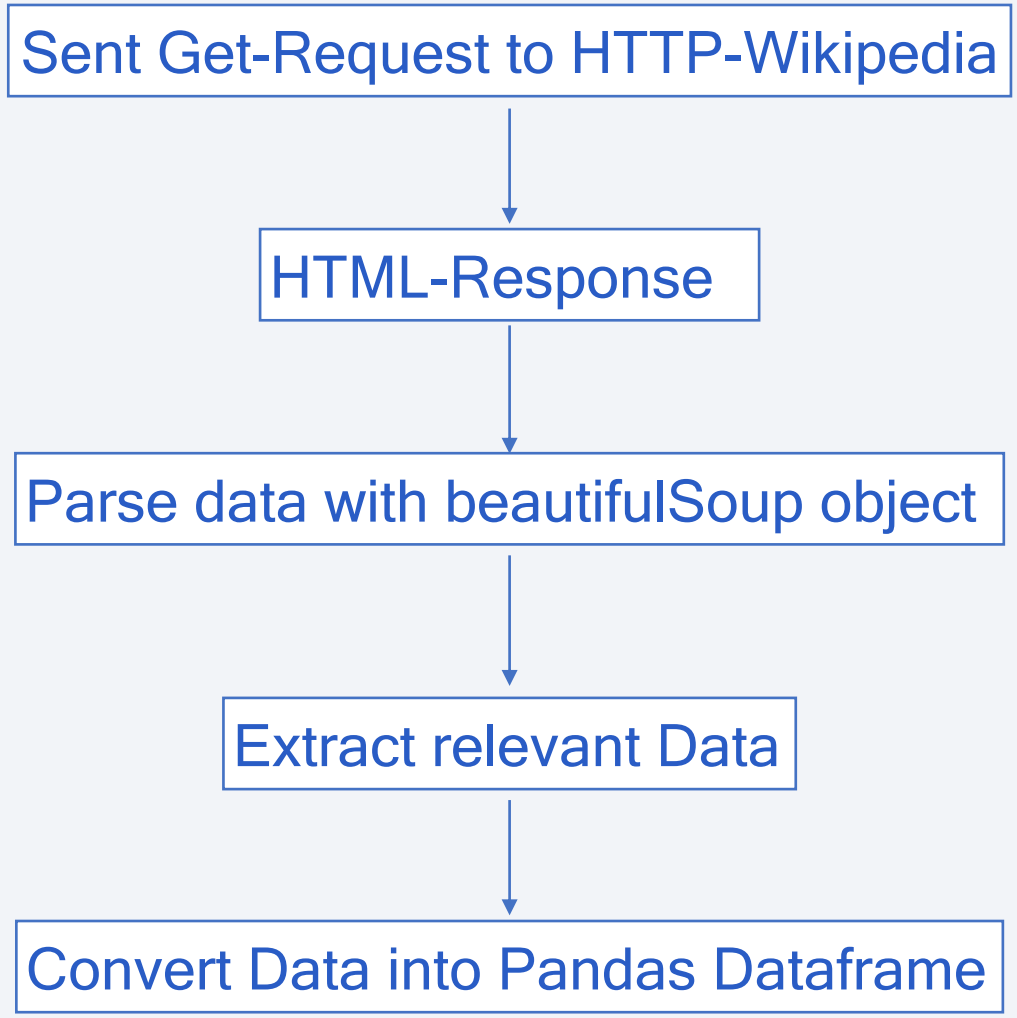
Data Collection – SpaceX API



GitHub URL of SpaceX API calls notebook:

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/jupyter-labs-spacex-data-collection-api%20(1).ipynb

Data Collection - Scraping



GitHub URL of web scraping notebook:

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/jupyter-labs-webscraping.ipynb

Data Wrangling

Transform raw data into clean, meaningful dataset

Filter/sample data → Exclude data records of Falcon 1 - launches

Replace missing/null values → Replace PayloadMass if missing by Mean

Make a (new) numerical label from the categorical variable ‘landing outcome’ → Create column ‘class’ :
‘0’ = failed landing
‘1’ = successful landing

GitHub URL of data wrangling notebook:

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Exploratory analysis

- Flight Number vs. Launch Site (scatter plot)
- Payload Mass vs. Launch Site (scatter plot)
- Landing Success Rate vs Orbit Type (bar chart)
- Flight Number vs Orbit type (scatter plot)
- Payload Mass vs. Orbit Type (scatter plot)
- Annual Trend in Landing Success (line chart)

GitHub URL of EDA with data visualization notebook:

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/edadataviz.ipynb

Is there a correlation between these variables?

EDA with SQL

- Unique launch sites
- Total payload mass for NASA (CRS) launches
- Average payload mass for Falcon 9 v1.1 rockets
- Date of the first successful landing on a ground pad
- Booster versions with successful drone ship landings
- Number of successful and failed missions
- Booster versions with maximum payload mass
- Failed drone ship landings in 2015 with month names
- Count and rank landing outcomes between specific dates

GitHub URL of EDA with SQL notebook:

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/jupyter-labs-eda-sql-coursera_sqllite%20(3).ipynb

Can we gain any significant insights from these SQL - selected data?

Building an Interactive Map with Folium

1. Circles for each launch site

   To visually mark the exact locations of the launch sites on a map

2. Markers for each launch site

   To display the names of the launch sites on the map for quick identification

3. Marker clusters grouping all launches

   To keep the map organized when there are many markers at the same location

4. Colored markers (green for successful landings, red for failed landings)

   To visualize quickly the success or failure of landings per location

Building an Interactive Map with Folium (Continued)

5. Mouse position

   To show the exact coordinates of a point on the map when moving the mouse

6. Marker for proximities to a launch site

   To indicate nearest coastline, railway, highway, city to a launch site

7. PolyLine between a launch site and its proximities

   To represent the distance between the launch site and nearest coastline, railway, highway and city

GitHub URL of interactive map with Folium notebook

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/lab_jupyter_launch_site_location%20(1)%20(1).ipynb

Building a Dashboard with Plotly Dash

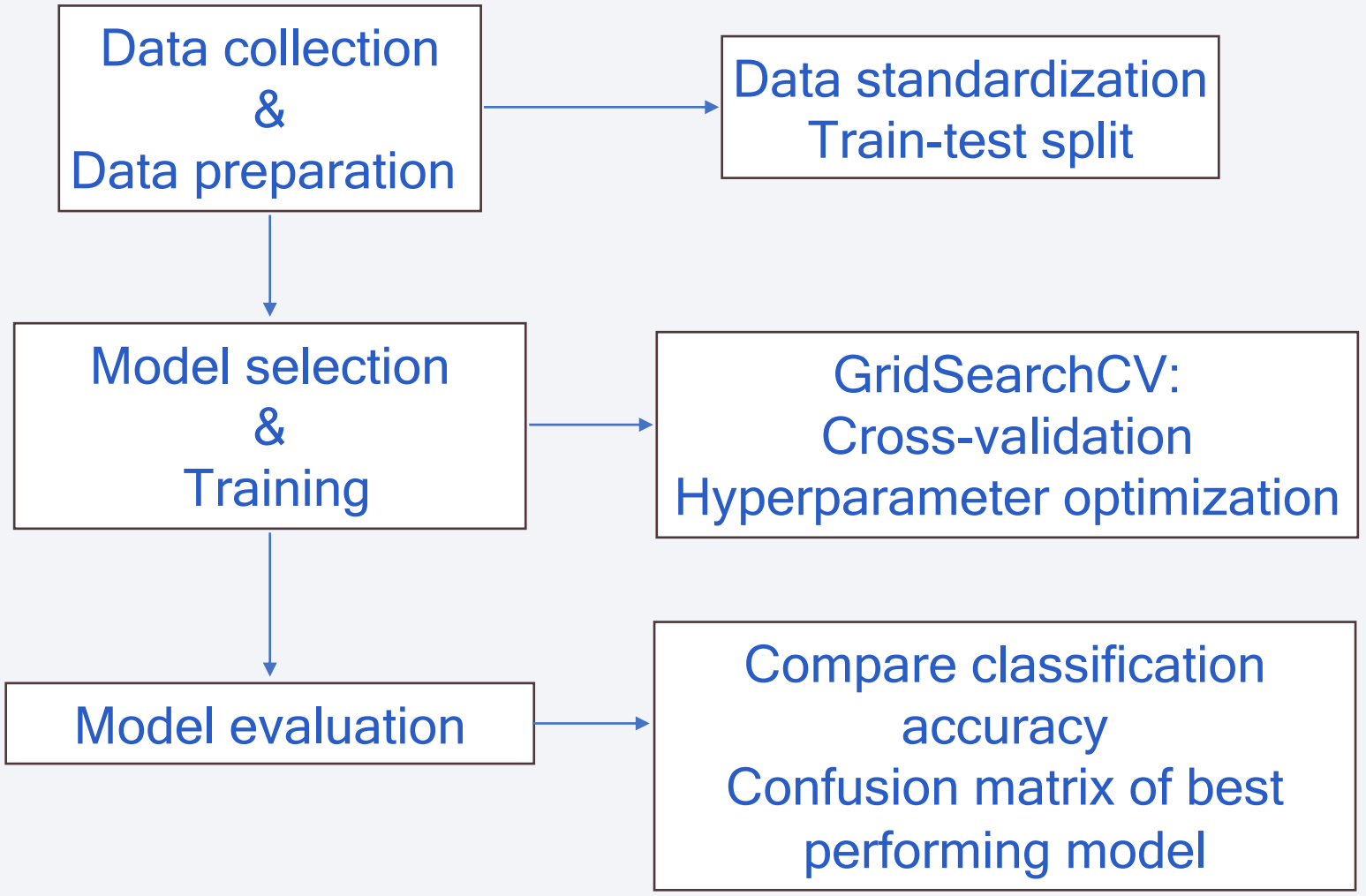
- Create a drop-down menu for launch sites (select-options: ‘all launch sites’ or select launch site separately)
- Make a pie chart of the landing outcome (succeeded or failed)
- Add a slider for payload mass
- Add scatterplot for payload, booster version and landing outcome

- To compare landing performance for different launch sites
- To analyze if there is a correlation between payload and landing outcome, distinguishing between booster versions

GitHub URL of Plotly Dash notebook:

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/spacexdashapp

Predictive Analysis (Classification)

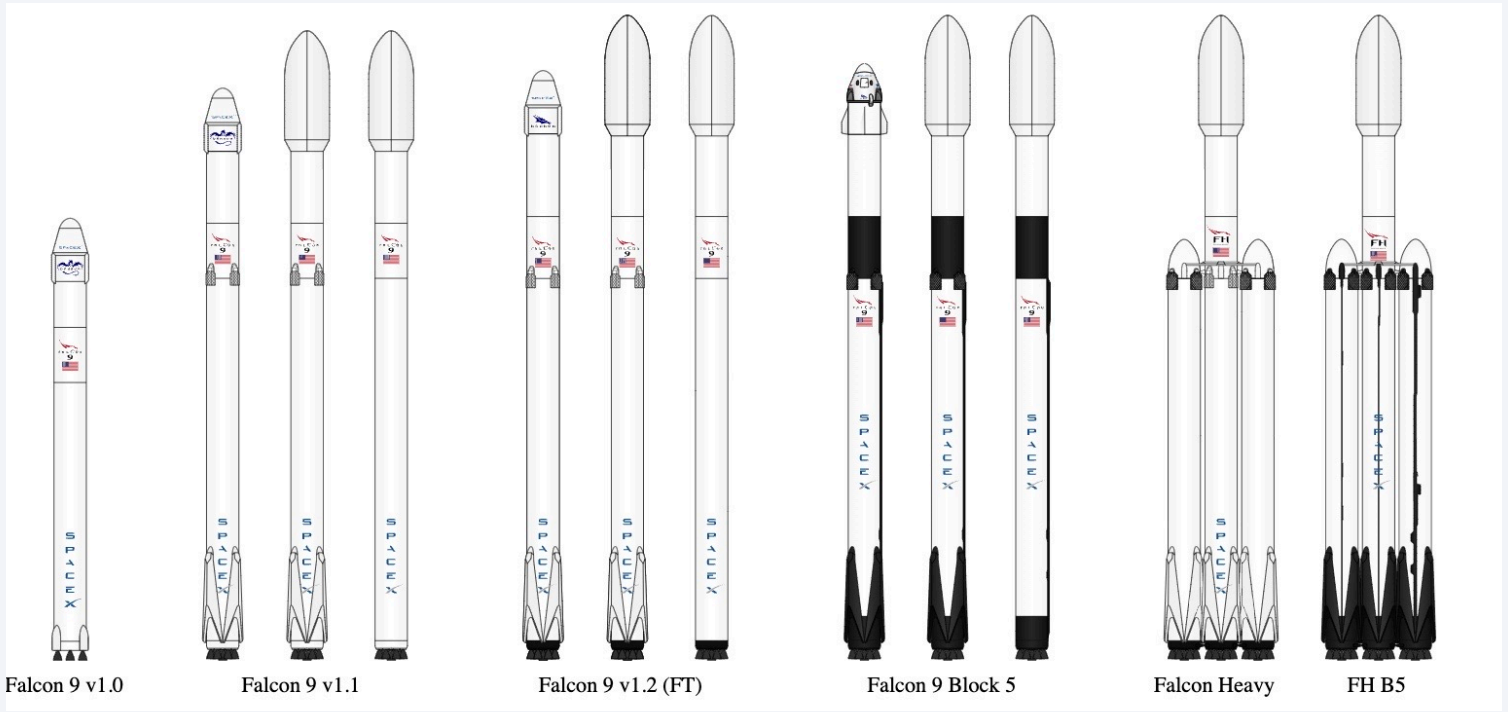


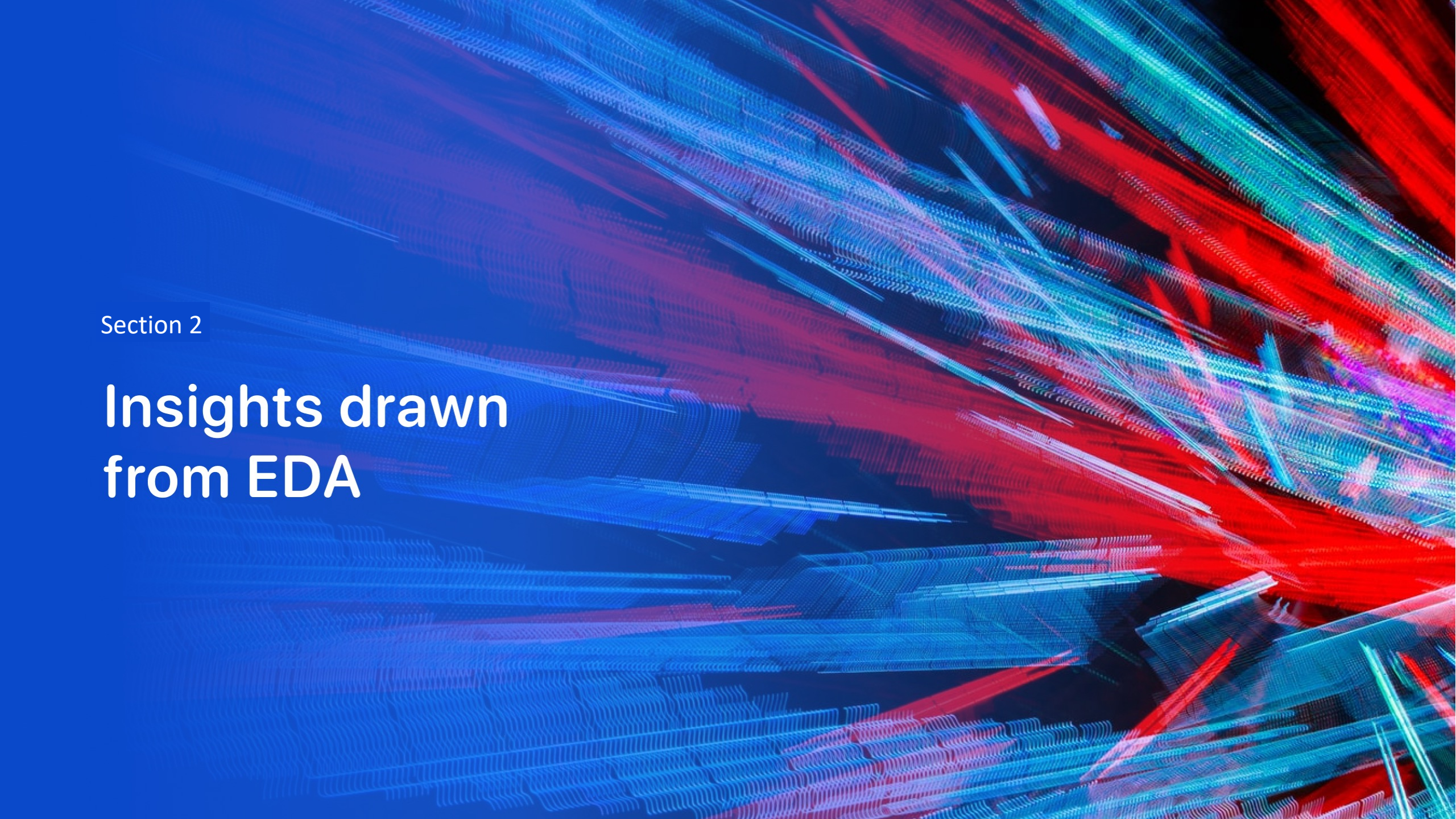
GitHub URL of predictive analysis notebook

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/SpaceX_Machine%20Learning%20Prediction_Part_5%20(4).ipynb

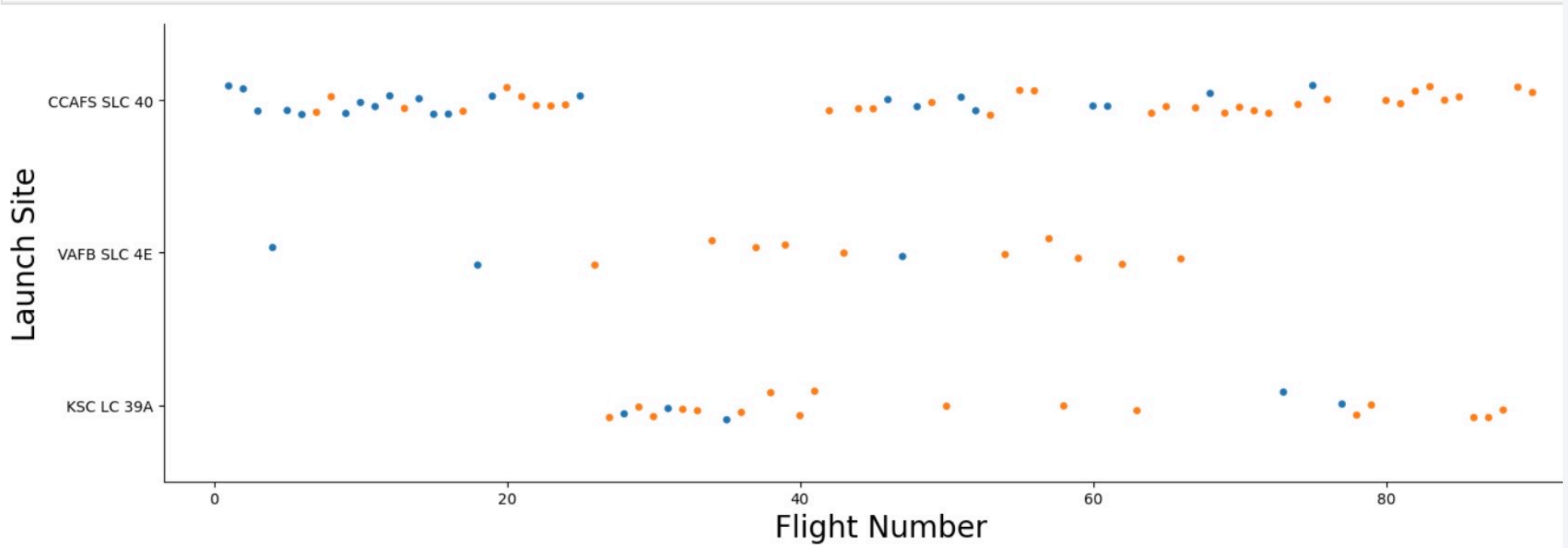
Results

- Section 2: Exploratory data analysis
- Section 3: Launch Sites and Proximities analysis
- Section 4: Building a dashboard with Plotly
- Section 5: Predictive analysis and conclusions



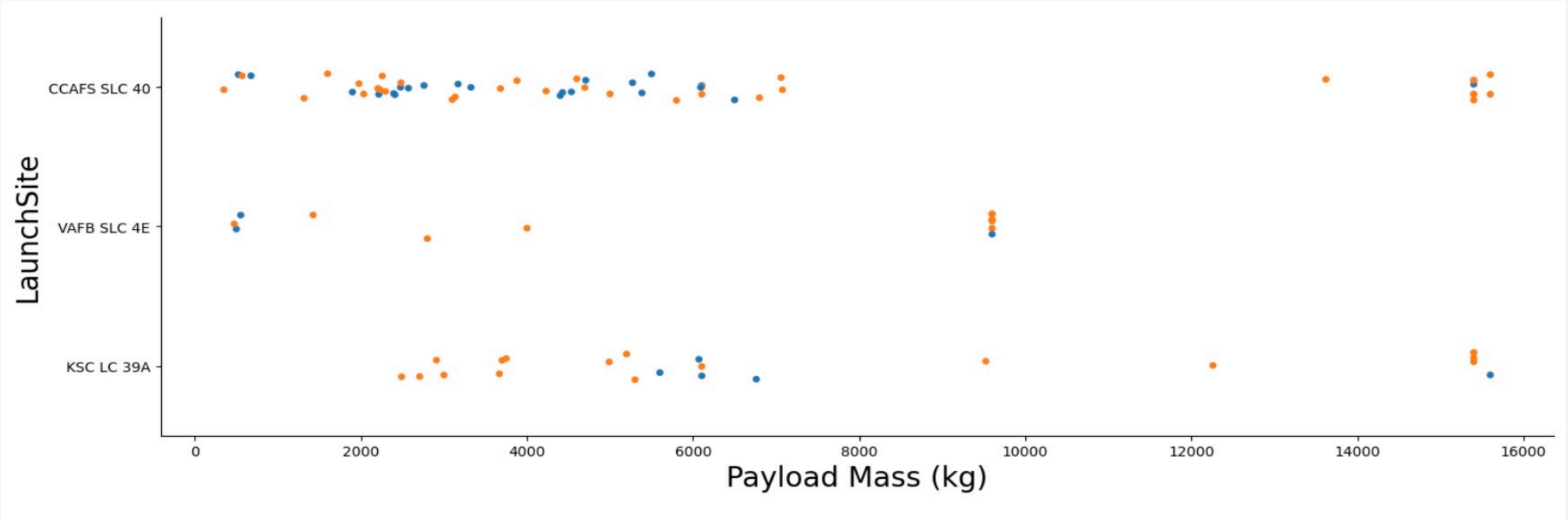

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



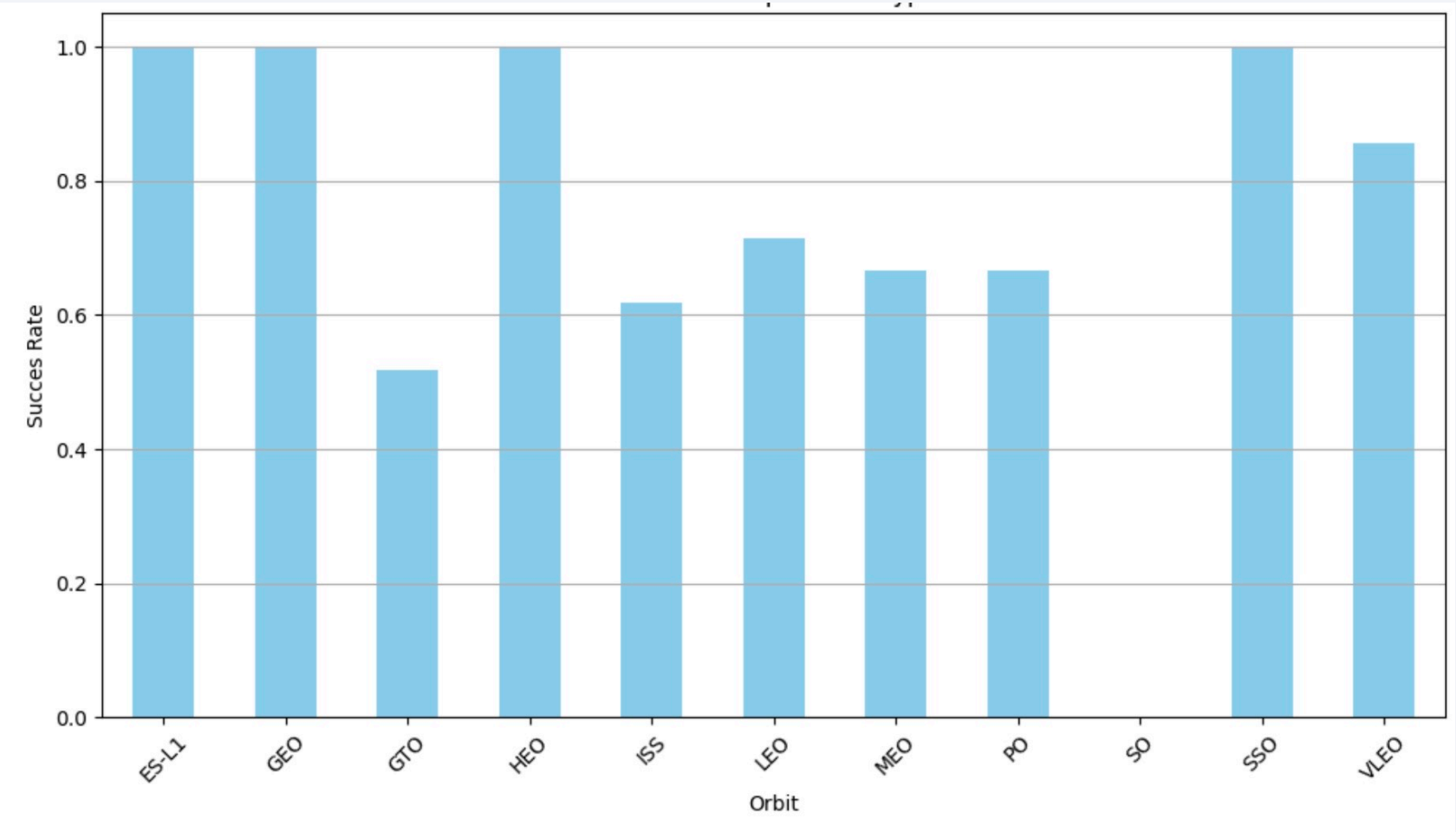
- As the flight number increases, the first stage is more likely to land successfully.

Payload vs. Launch Site



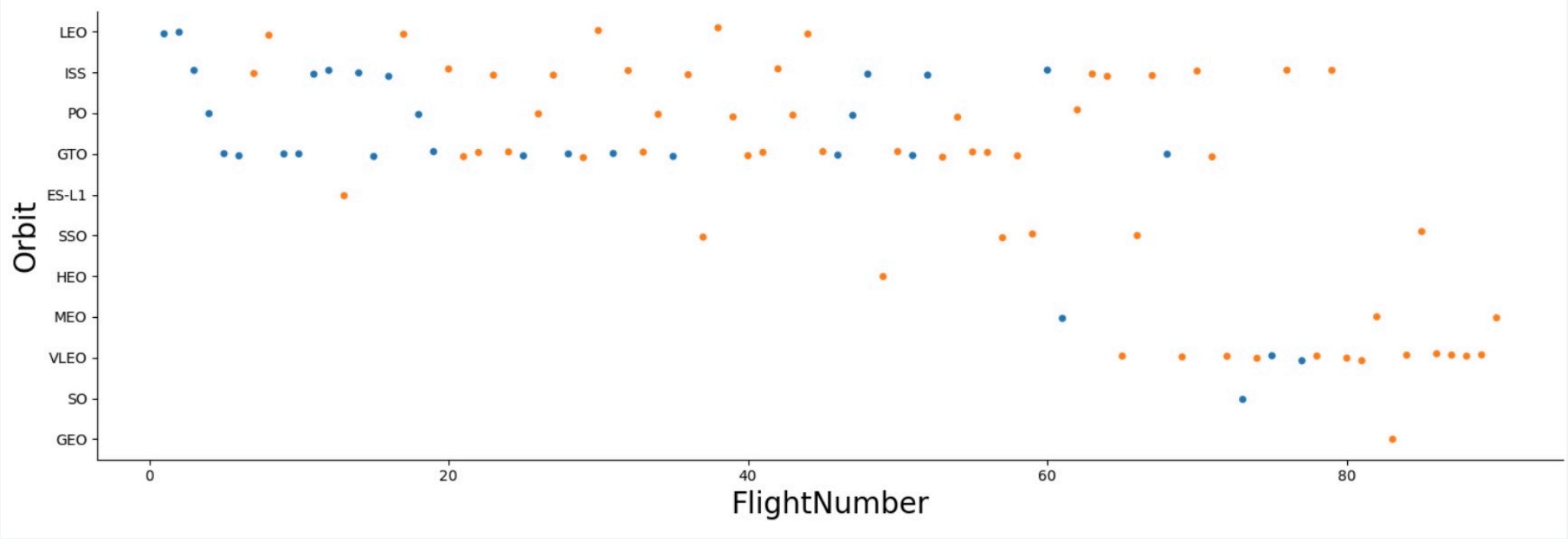
- No clear correlation between payload and succesrate of landing first stage.
- The number of successful and failed landings is evenly distributed for very heavy payloads.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000 kg).

Success Rate vs. Orbit Type



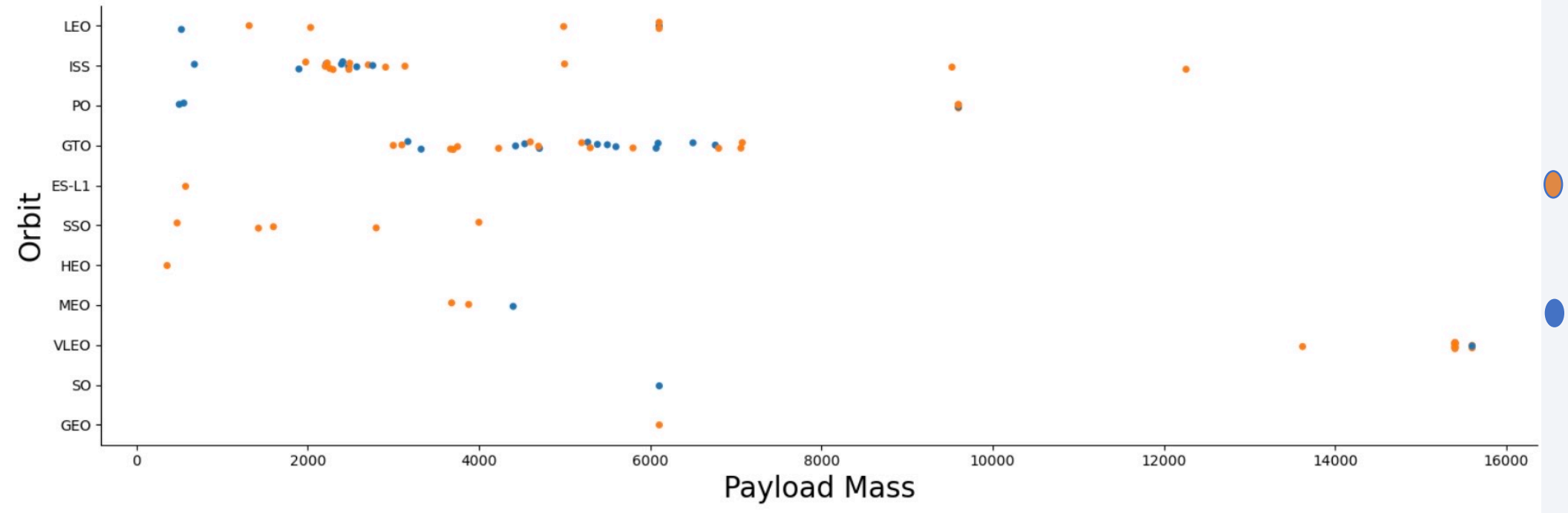
- ES-L1, GEO, and HEO all have a 100% success rate (1.0 on y-axis)

Flight Number vs. Orbit Type



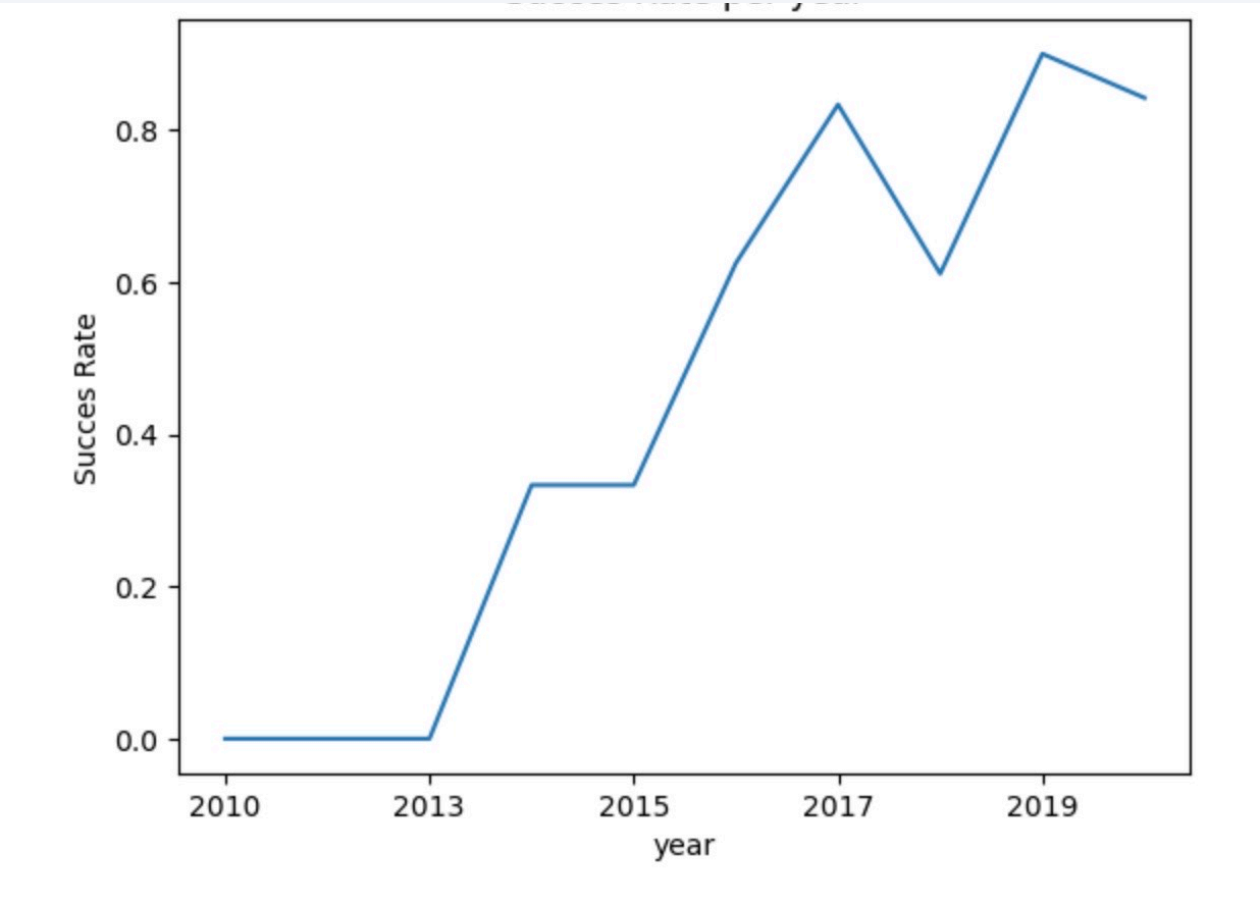
- Success seems to be related to the number of flights in LEO orbit
- In GTO orbit, there appears to be no relationship between flight number and success
- Overall, succesrate increases with number of flights

Payload vs. Orbit Type



- Heavy payloads are more frequent in Polar, LEO and ISS

Launch Success Yearly Trend



- The success rate of landings since 2013 kept increasing till 2020

All Launch Site Names

CCAFS LC-40	Cape Canaveral Air Force Station Launch Complex 40 (Florida, East Coast)
CCAFS SLC-40	Cape Canaveral Air Force Station Space Launch Complex 40 (Florida, East Coast)
KSC LC-39A	Kennedy Space Center (Florida, East Coast)
VAFB SLC-4E	Vandenberg Air Force Base (California, West Coast)

- These are the launch sites for Falcon 9 in the US.
- In 2000, The CCAFS LC-40 was renamed to SLC-40 (transition from military-focused to more commercial and government space missions)

5 Launch Site Names 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	F9 v1.0 B0003	CCAF S LC-40	Dragon Space craft Qual. U.	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43	F9 v1.0 B0004	CCAF S LC-40	Dragon demo flight C1	0	LEO (ISS)	NASA(COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44	F9 v1.0 B0005	CCAF S LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS) NRO	Success	No attempt
2012-10-08	0:35	F9 v1.0 B0006	CCAF S LC-40	Space X CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10	F9 v1.0 B0007	CCAF S LC-40	Space X CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Overview of the variables included in the dataset
- First 5 Launches of Falcon 9 Rocket
- First two attempts landing the first stage were unsuccessful.
- No landing attempt was made for the last three missions.

Total Payload Mass

Sum(PAYLOAD_MASS_KG)	Booster_Version	Customer
45596	F9 v1.0 B0006	NASA (CRS)

- The ability to deliver this amount of payload for a key customer like NASA demonstrates the economic viability of SpaceX's approach.

Average Payload Mass by F9 v1.1

Booster_Version	AVG(PAYLOAD_MASS_KG_)
F9 v1.1 B1003	2534.66

- This is the average amount of payload that a newer version of the Falcon 9 booster was able to deliver.
- It gives an idea of how many launches it would take for the newer version to match the total payload of earlier versions missions.
- Increasing payload capacity reduces the number of launches needed for similar missions.

First Successful Ground Landing Date

min(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

- First time a Falcon 9 first stage successfully landed on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster Version	PAYLOAD_MASS__KG	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- All these boosters have successfully landed on a drone ship.
- Later versions have higher payload capacity.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- (Almost) all mission outcomes were succesful even if a landing dit not succeed.

- A failure in landing has no impact on the mission outcome

Boosters Carried Maximum Payload

Date	Booster_version	PAYLOAD_MASS_KG
2019-11-11	F9 B5 B1048.4	15600
2020-01-07	F9 B5 B1049.4	15600
2020-01-29	F9 B5 B1051.3	15600
2020-02-17	F9 B5 B1056.4	15600
2020-03-18	F9 B5 B1048.5	15600
2020-04-22	F9 B5 B1051.4	15600
2020-06-04	F9 B5 B1049.5	15600
2020-09-03	F9 B5 B1060.2	15600
2020-10-06	F9 B5 B1058.3	15600
2020-10-18	F9 B5 B1051.6	15600
2020-10-24	F9 B5 B1060.3	15600
2020-11-25	F9 B5 B1049.7	15600

- When maximum payload was delivered, all launches used the B5 version of the Falcon 9.
- It clearly shows reuse of these boosters, as seen in the numbers after the decimal point. Some boosters are used more frequently than others.

2015 Launch Records

month	year	Booster_Version	Launch_Site	Landing_Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Earlier in 2015, 2 attempts for landing F9 first stage on a drone ship failed.
- Different boosters were used for both launches

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

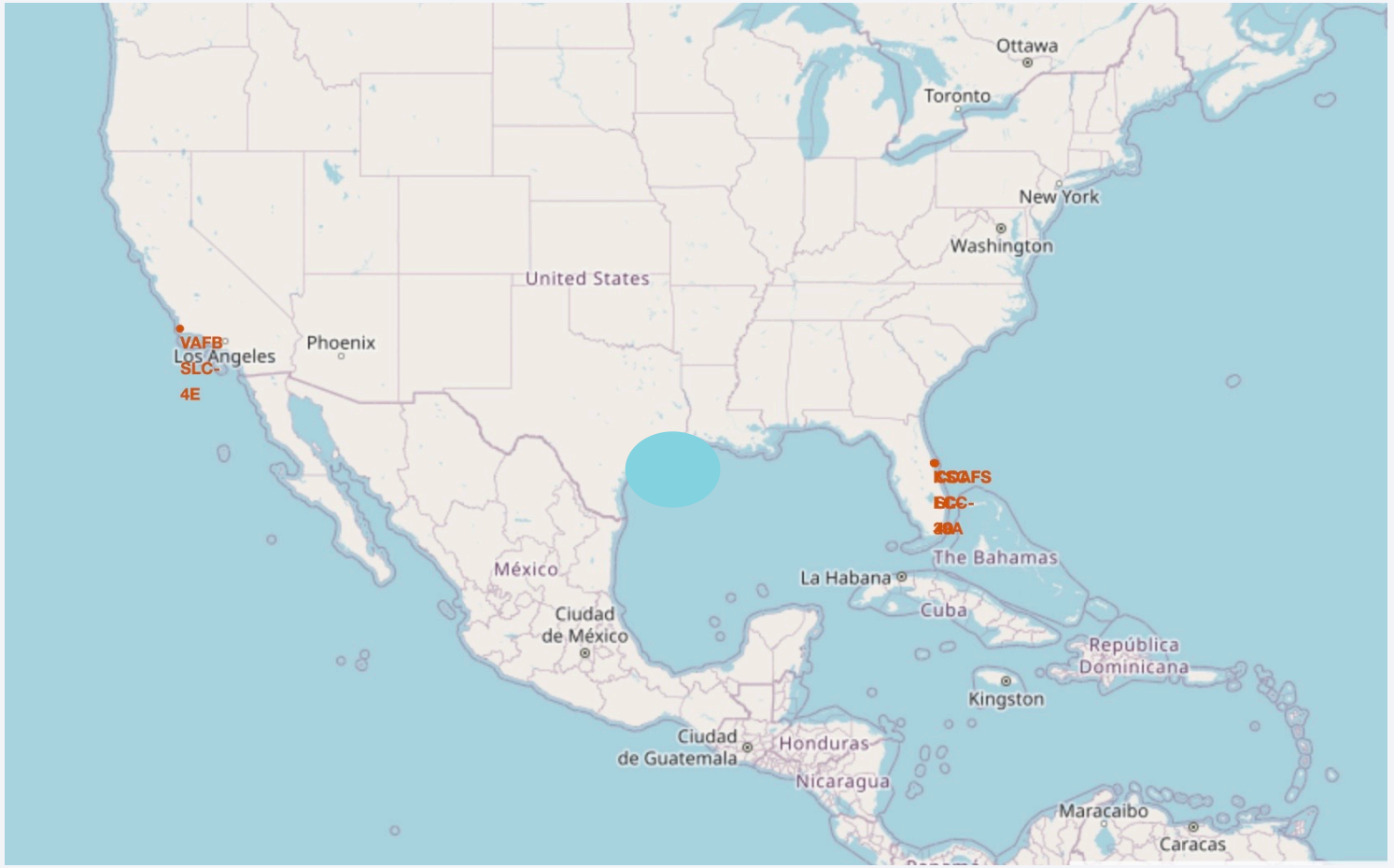
Date	Landing Outcome	Outcome count
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3

- The first successful landing on a drone ship occurred on April 8, 2016
- During this time period there were 5 failed attempts to land on a drone ship, starting from January 10, 2015.
- The number of successful landings on a drone ship (5) is equal to the number of failures and probably illustrate the progress in technology.
- Ground pad landings seem to occur less frequently then drone ship landings (3) in this time period) but were succesful

Section 3

Launch Sites Proximities Analysis

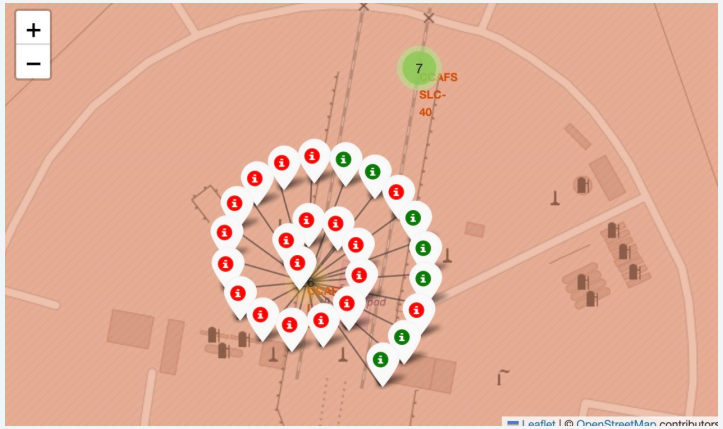
Space X Falcon 9 launch sites



The locations [Longitude, Latitude] that facilitate Space X Falcon 9 missions are pinned and marked with the name of the launch site

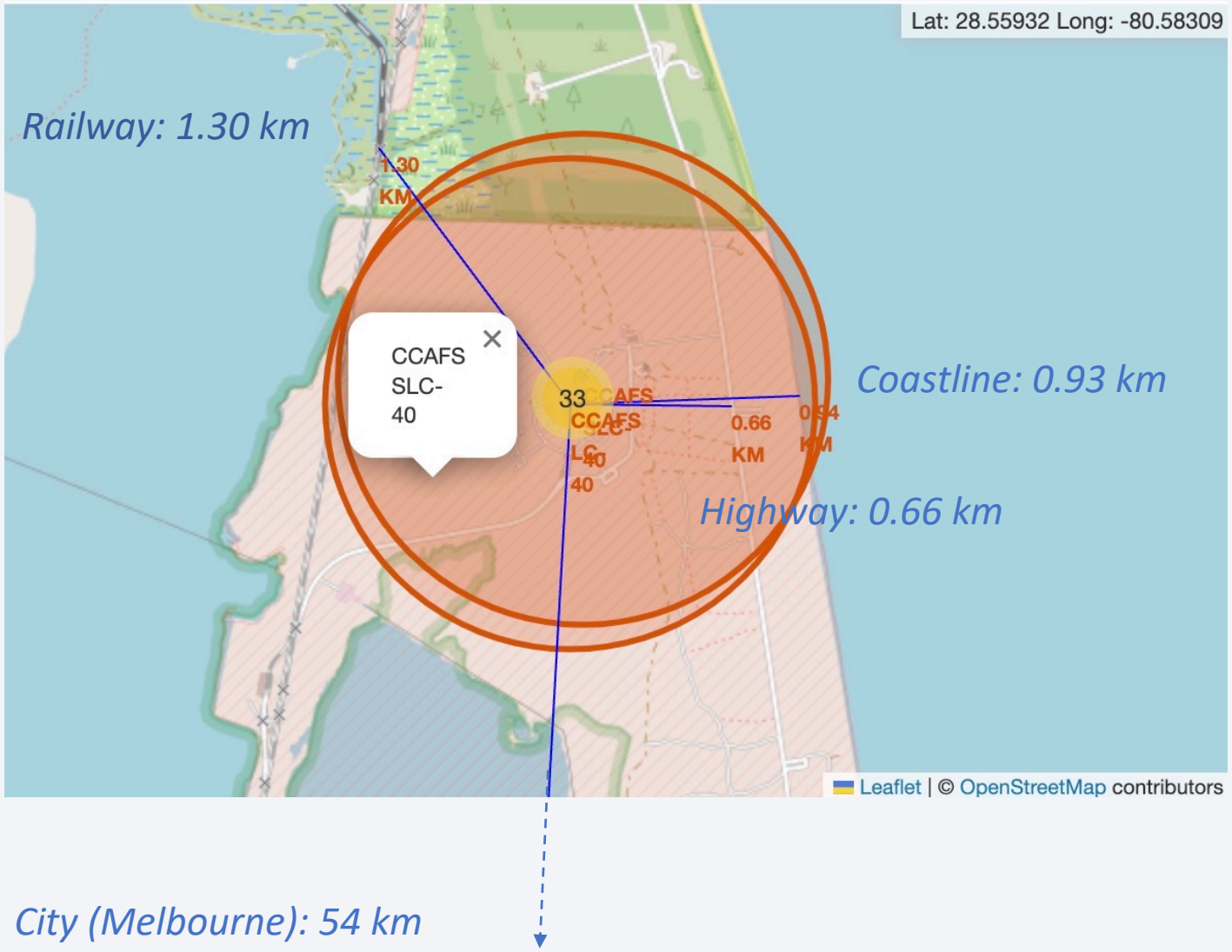
Number of launches and Landing outcome by Site

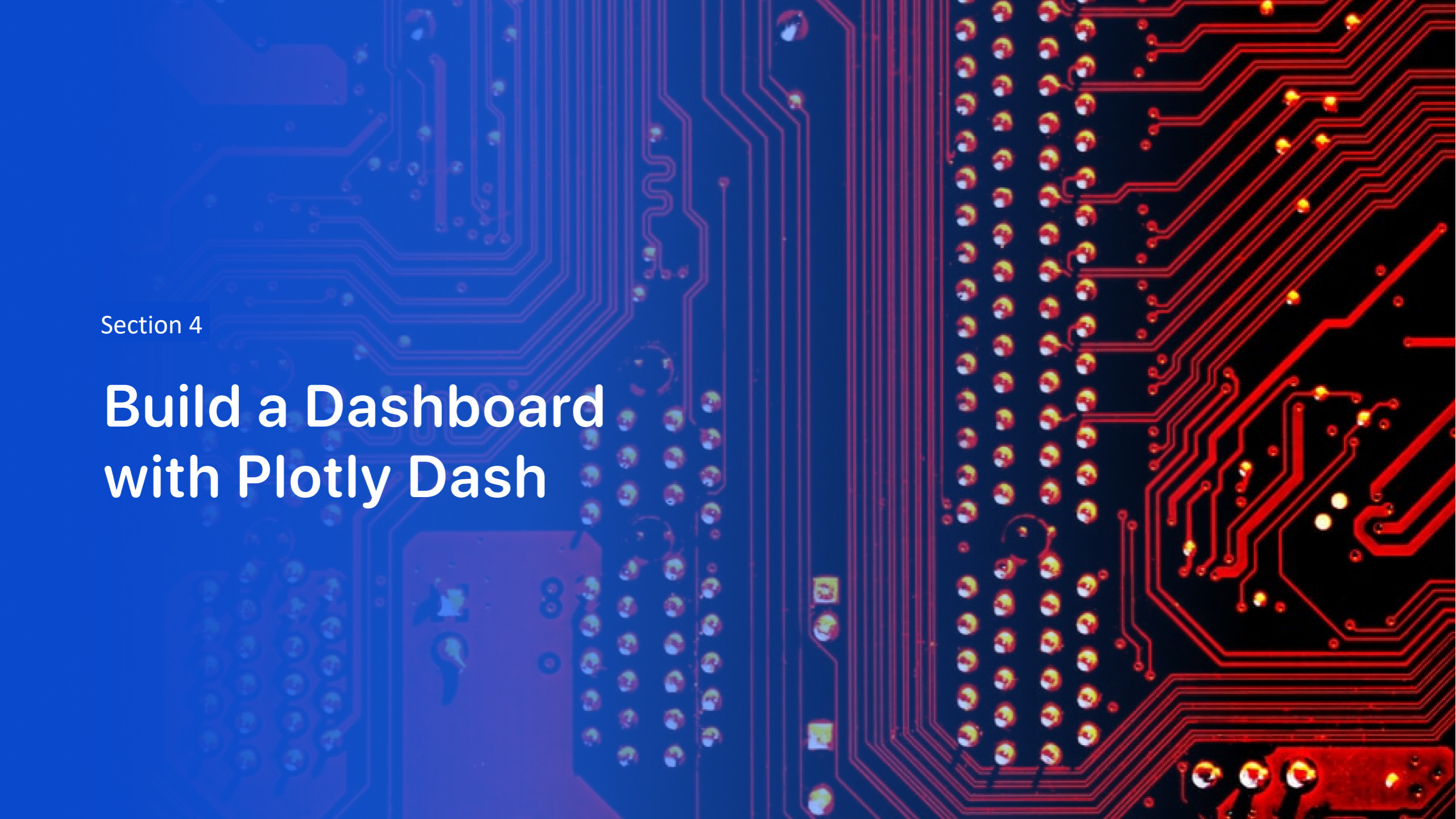
Distribution of number of launches between East and West Coast



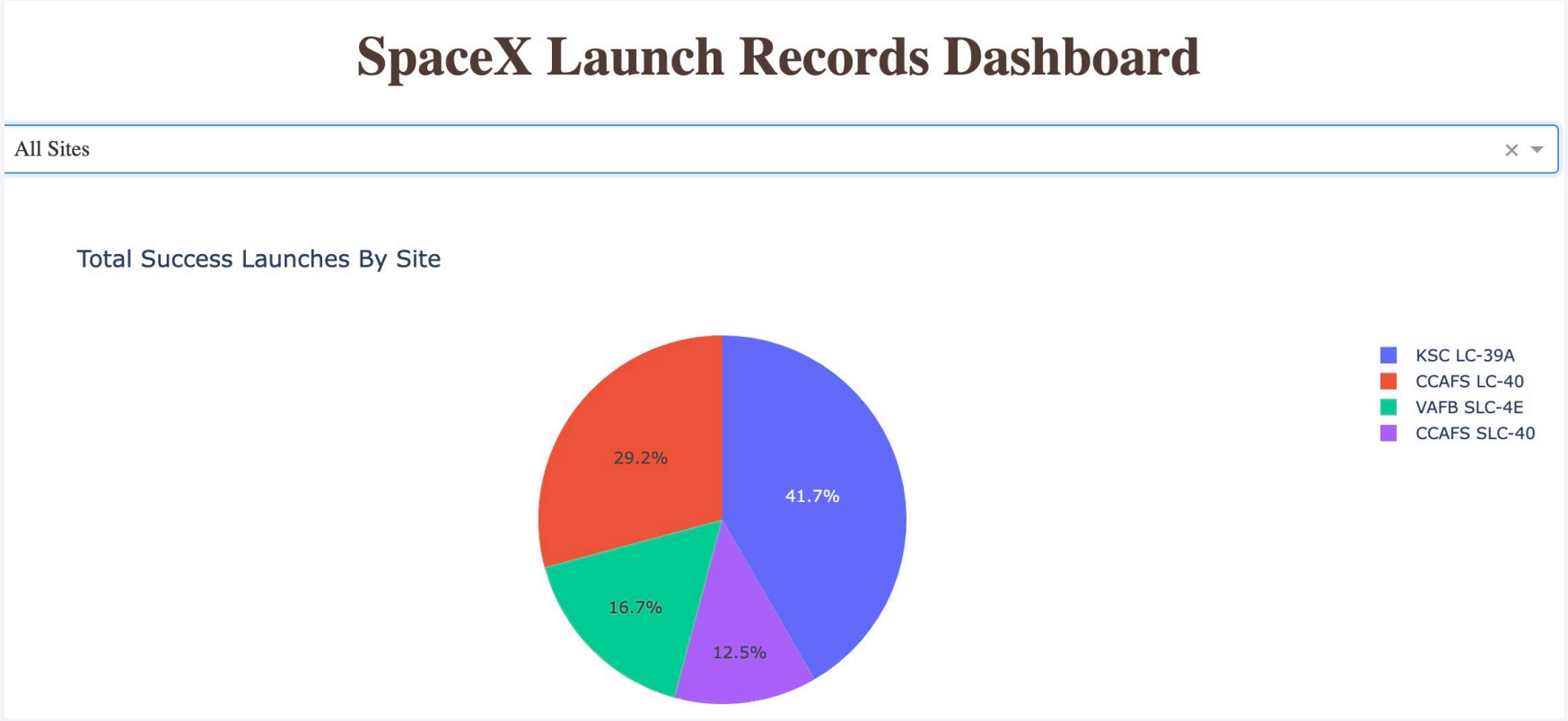
CCAFS SLC-40: failed & succeeded landings (green=successful)

Launch Sites and its Proximities (km): CCAFS SLC-40



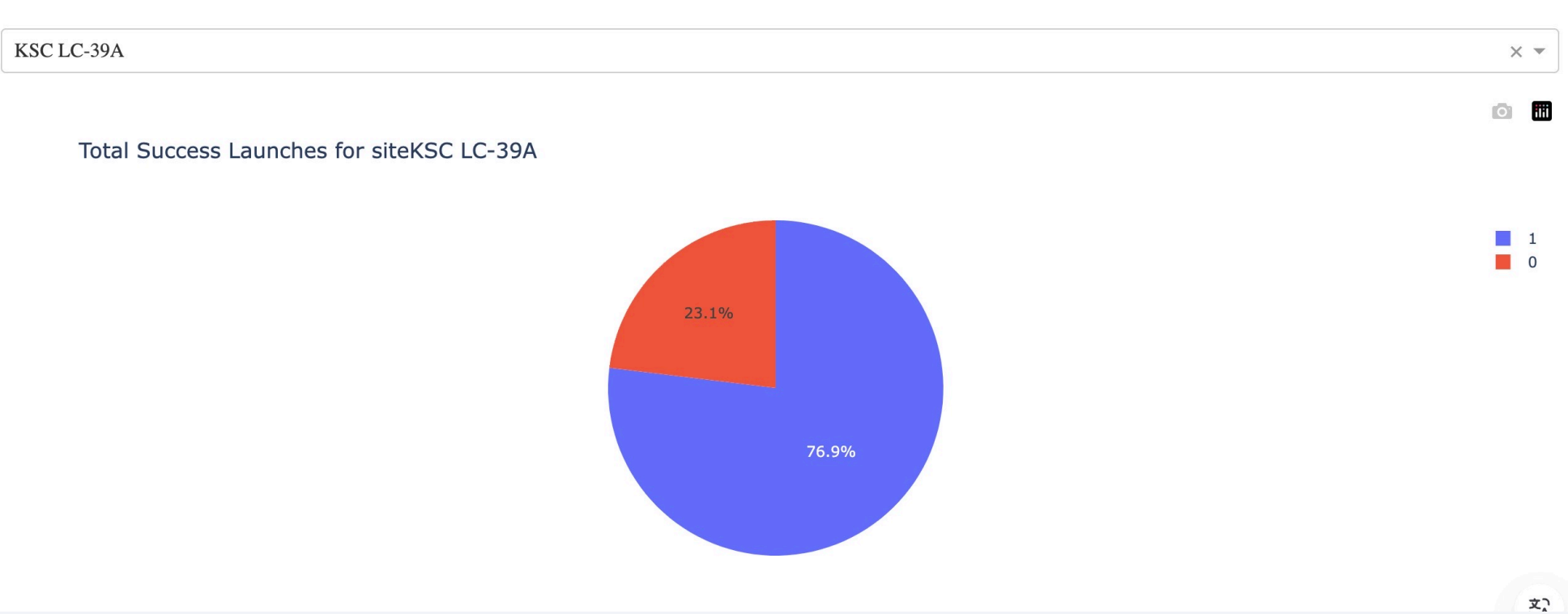

Section 4
Build a Dashboard with Plotly Dash

Succes rate landings for all sites



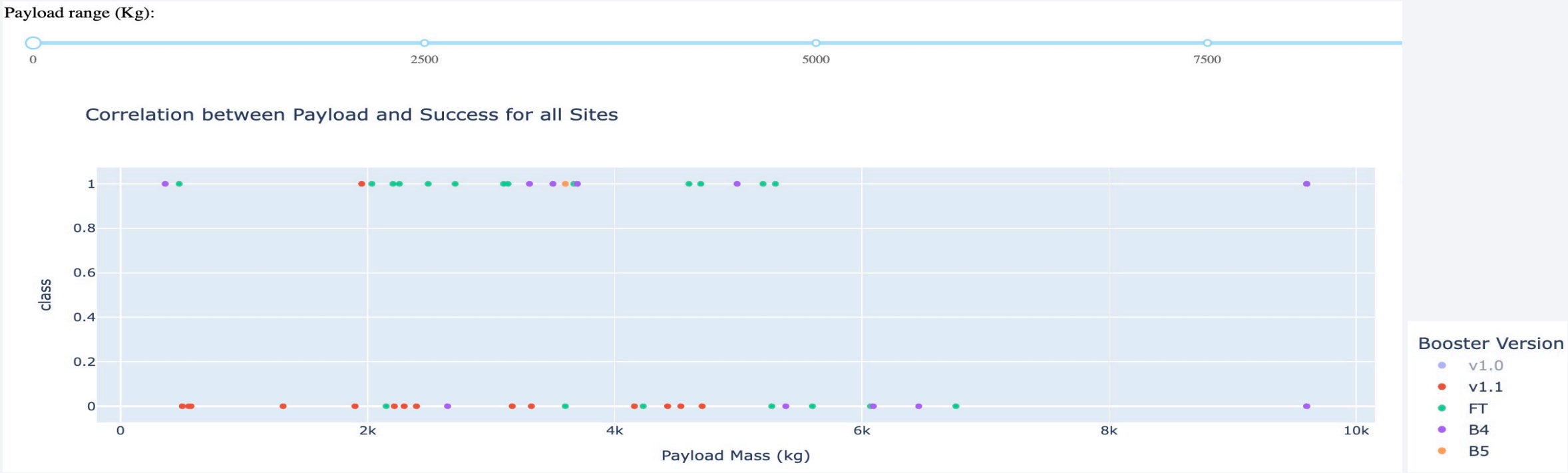
- Launch Complex 39A at Kennedy Space Center (KSC LC-39A) takes the lead when it comes to successful Falcon 9 landings

KSC LC-39A: Highest succes rate of landing first stage



- 76.9% of Space X landings at KSC LC-39A were succesful

Correlation Payload vs. Landing Outcome (all sites)

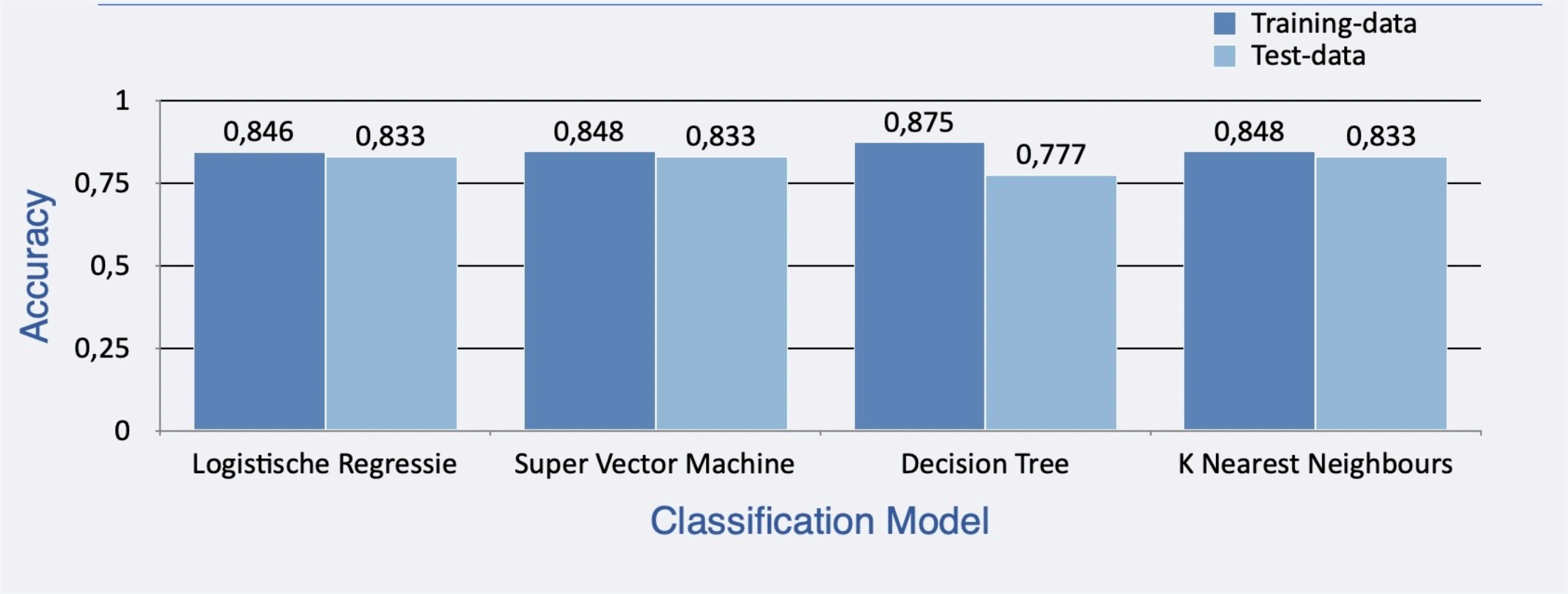


- Early boosterversions (v1.0 and v1.1) have a lower succesrate in landing first stage.
- Succesrate of landings and payload capacity increases with later booster versions (B5 version has 100% succesrate).

Section 5

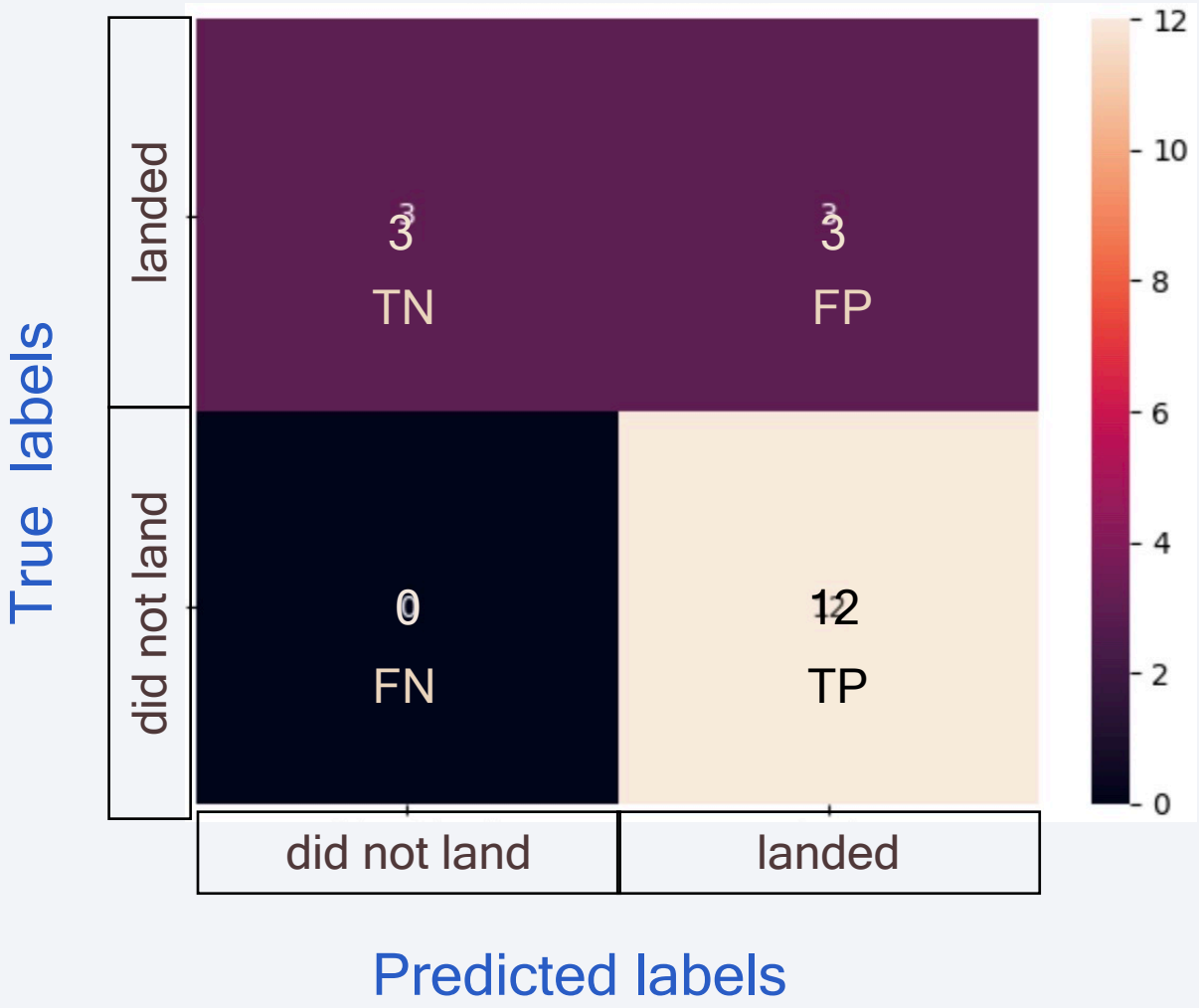
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix for Logistic Regression Model

Predicted vs Actual classification of Space X Landings: Key Metrics



- **Accuracy:**

  (TP + TN)/(TP + TN +FP + FN) = 83.33%

- **Precision:**

  (TP)/(TP+FP) = 80%

- **Recall:**

  `(TP)/(TP+FN) = 100%

- **F1-Score:**

  (2*(precision*Recall)/(Precision + Recall)

  = 88.89%

Conclusions

- Factors such as flight number, booster version, orbit type, payload mass and launch site were identified as important factors for machine learning prediction.
- Multiple classification models were tested, including Logistic regression, Support vector machines (SVM), decision tree and K-nearest neighbors (KNN)
- **83%** is highest classification accuracy, found in **Logistic Regression**, SVM - as well as in KNN
- SVM and KNN are probably over-fitting training data (higher training accuracy)

Conclusions (continued)

- Logistic regression is likely to be more reliable on new data
- Performance Metrics for Logistic Regression: **Accuracy**: 83.33%, **Precision**: 80%, **Recall:** 100%, **F1-score**: 88.89%
- With an accuracy of 83%, there's still a margin for error
- Further refinement and improvement of model is possible with new data from SpaceX

Appendix

In this section, you will find a list of all the links to the original queries, graphs, and codes used throughout the presentation. These materials provide additional context and support for the analyses discussed.

Data Collection (data collected using SpaceX Rest API)

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/jupyter-labs-spacex-data-collection-api%20(1).ipynb

Data Collection Webscraping (data collected using Wikipedia_page)

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/jupyter-labs-webscraping.ipynb

Appendix

Data wrangling

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/labs-jupyter-spacex-Data%20wrangling.ipynb

Data exploratory analysis (EDA vizualisation: charts)

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/edadataviz.ipynb

Data exploratory analysis (sql-queries)

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/jupyter-labs-eda-sql-coursera_sqllite%20(3).ipynb

Appendix

Visual data analysis (Folium)

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/lab_jupyter_launch_site_location%20(1)%20(1).ipynb

Code used to built interactive dashboard with Plotly

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/spacexdashapp

Predictive analysis

https://github.com/35en/applied-data-science-capstone/blob/acd1fc0c7ef28ac0ff5320fd4b7aeac440e6e391/SpaceX_Machine%20Learning%20Prediction_Part_5%20(4).ipynb

Thank you!