IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Federico Hattab

18th February 2026

Outline

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

Executive Summary

- The commercial space age is here, and companies are competing in driving down the costs of launching rockets. In this landscape SpaceX is capable of reusing the first stage of the rocket, arguably the largest contributor to the overall rocket cost. In order to assess the competitivity of other companies, we want to be able to predict if the Falcon 9 first stage will land successfully.

- To reach our goal, we first collected data using SpaceX API and web scraping. Data was then cleaned and some Exploratory Data Analysis (EDA) was done to find patterns in the data and determine what would be the label for training supervised models. Using Jupyter notebook and SQL queries on a Db2 database we gained a deeper insight into the data set. The EDA phase also involved using Pandas, Matplotlib and Seaborn to visualize relationships between feature and perform Feature Engineering, in preparation of the training of ML models. Further visual analytics was performed with Folium, studying data related to the geographic location of the launches. Furthermore a Dash app was build to provide interactive data exploration capabilities. Finally, data was standardized, and 4 ML methods (Logistic Regression, SVM, Classification Trees and KNN) were used to build prediction models.

- Results show that Falcon 9 launch success rate varies across the 4 launch sites used. Depending on the weight of the payload different launch sites were used. No launch with payload heavier than 10000 kg from the VAFB-SLC site has been done. Different target orbits have different success rates. Launches to the LEO orbit started to be all successful after the initial launches. Overall, the success rate since 2013 kept increasing till 2020. The site with the largest number of successful launches and the highest launch success rate is KSC LC-39A. In successful launches B4 and FT boosters were predominantly used, and the booster selection changes with site and payload mass. The ML method with the best performance among the ones tested is the Tree Classifier, with a score of 88.89% on test data, using tuned hyperparameters.

Introduction

Background and context

- The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful company poised to lead this new era is SpaceX. SpaceX's accomplishments include sending spacecraft to the International Space Station.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage of the rocket.
- In order to do make strategic decision and gain more insight into the industry, our company, Space Y, wants to gain deeper insight into SpaceX launches

Problem we want to find an answer to:

- The goal of this work is to gain more insight into SpaceX historical launch data and to be able to determine if SpaceX will reuse of not the first stage in future launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and web scraping
- Perform data wrangling
  - Pre-processing of data with Pandas and feature engineering
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data preparation and split between train and test set
  - Training of different models and use of Grid Search for parameters tuning
  - Evaluation of performances

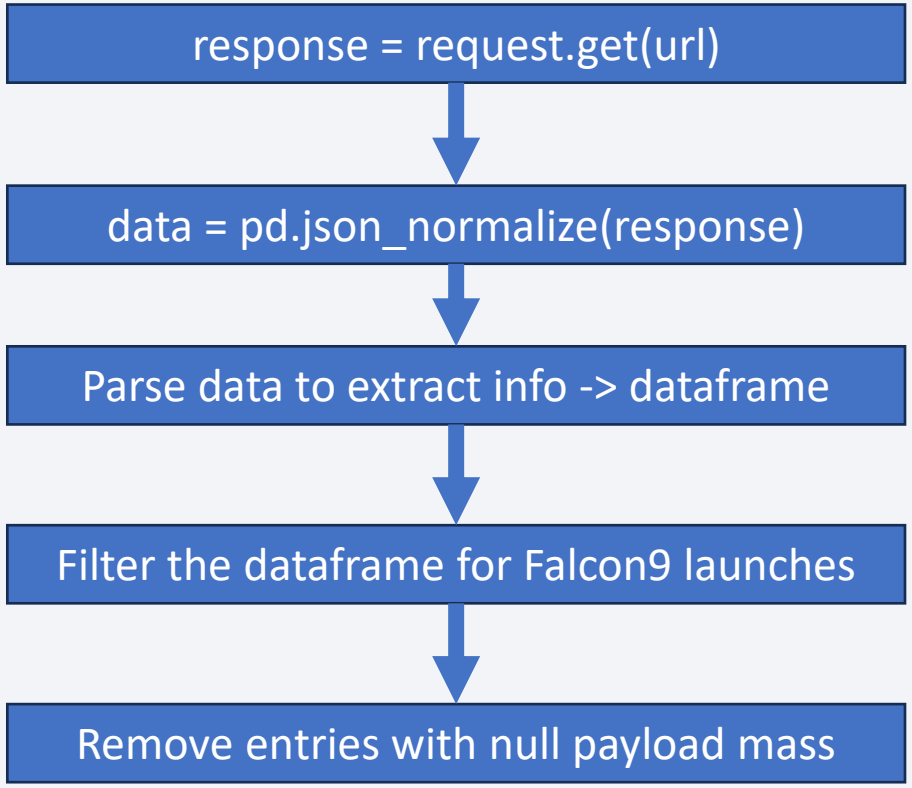
Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

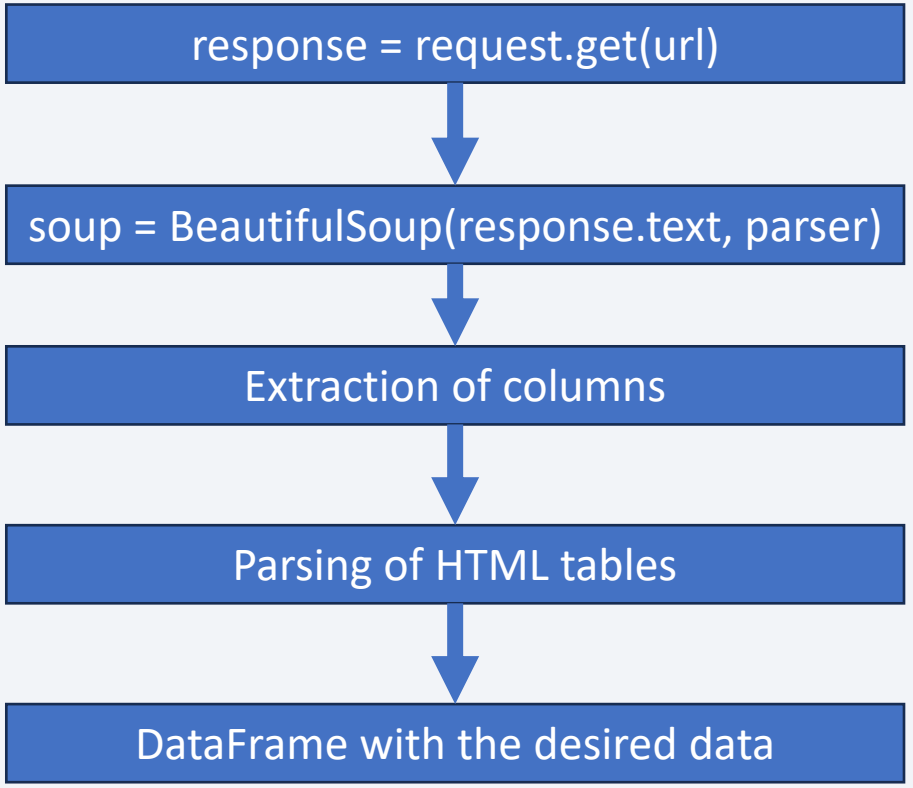
- Use of API to get json file
- Parsing of the json file to extract relevant information to a DataFrame
- Cleanup of the DataFrame to include only Falcon9 launches and non null payload masses

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/fe-htt/testrepo/blob/main/Module%201/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from its URL and create a BeautifulSoup object from the HTML response
- Extract all column/variable names from the HTML table header and Create a data frame by parsing the launch HTML tables
- GitHub URL of the completed web scraping notebook: https://github.com/fe-htt/testrepo/blob/main/Module%201/jupyter-labs-webscraping.ipynb

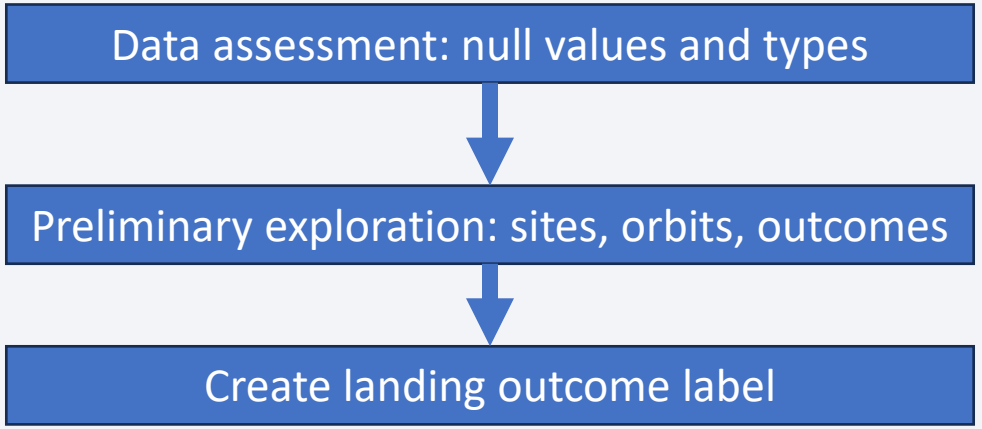


Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

Data Wrangling

- Assess the presence of null values and the type of each field
- Calculate the number of launches on each site, to each orbit and mission outcomes for each orbit
- Create a Class column that tells if the landing was successful or not
- GitHub URL of the completed web scraping notebook: https://github.com/fe-htt/testrepo/blob/main/Module%201/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

EDA with Data Visualization

- The charts plotted are:
  - Scatter plot Flight Number vs Payload Mass: to see how the continuous launch attempts would affect the launch outcome for different payload masses
  - Scatter plot Flight Number vs Launch Site: to see if there are differences between sites in terms of launch trends and success rate
  - Scatter plot Payload Mass vs Launch Site: to observe if there is any relationship between launch sites and their payload mass
  - Bar chart of the relationship between success rate of each orbit type: to visually check if there are any relationship between success rate and orbit type
  - Scatter plot Flight Number vs Orbit Type: to see if there is any relationship between flight number and orbit type
  - Scatter Plot Payload Mass vs Orbit Type: to reveal the relationship between payload mass and orbit type
  - Line plot of launch success yearly trend: to observe the average success rate trend over the years
- GitHub URL of the completed EDA with data visualization notebook: https://github.com/fe-htt/testrepo/blob/main/Module%202/edadataviz.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

EDA with SQL

- Display the names of the unique launch sites in the space mission: %sql select distinct "Launch_Site" from SPACEXTABLE
- Display 5 records where launch sites begin with the string 'CCA': %sql select * from SPACEXTABLE where "Launch_Site" like "CCA%" limit 5
- Display the total payload mass carried by boosters launched by NASA (CRS): %sql select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where Customer like "NASA (CRS)"
- Display average payload mass carried by booster version F9 v1.1: %sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" like "F9 v1.1"
- List the date when the first succesful landing outcome in ground pad was achieved: %sql select MIN(Date) from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000: %sql select "Booster_Version" from SPACEXTABLE where ("Landing_Outcome" like "%drone ship%") and ("Mission_Outcome" = "Success") and ("PAYLOAD_MASS__KG_" between 4000 and 6000);
- List the total number of successful and failure mission outcomes: %sql select "Mission_Outcome", COUNT("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome";
- List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function: %sql select distinct "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") from SPACEXTABLE);
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015: %sql select substr(Date, 6,2), "Landing_Outcome", "Booster_version", "Launch_Site" from SPACEXTABLE where (substr(Date,0,5)='2015') and ("Landing_Outcome" = "Failure (drone ship)");
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order: %sql select "Landing_Outcome", count("Landing_Outcome") as Count from SPACEXTABLE where (Date between "2010-06-04" and "2017-03-20") group by "Landing_Outcome" order by Count desc;
- GitHub URL of the completed EDA with SQL notebook: https://github.com/fe-htt/testrepo/blob/main/Module%202/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Addition of a circle (folium.Circle) and marker (folium.Marker) for each launch site: to visualize the location of the launch sites and evaluates things such as proximity to the equator and to the cost
- Marking of each launch and its outcome, using a marker cluster and folium.Marker: to easily identify success rates on each test site
- MousePosition: to get the coordinates of places on the map and evaluate proximity of poins of interest
- Markers and Polylines to connect launch sites and points of interest to easily identify distance from locations of interest such as the coast, cities, railways and highways
- GitHub URL of the completed interactive map with Folium map: https://github.com/fe-htt/testrepo/blob/main/Module%203/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

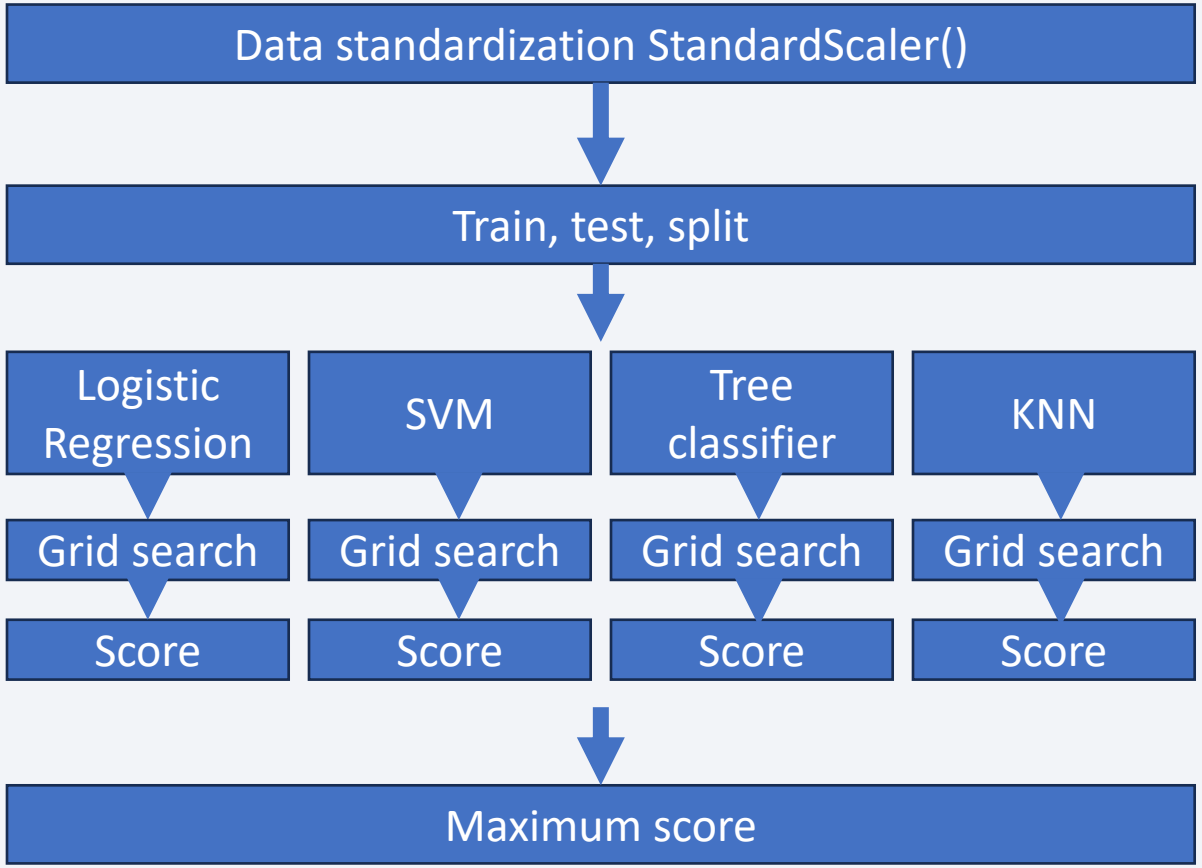
- Drop down menu with callback to charts: to choose if you want to analyze a specific launch site or all launches from all sites together. The callback is to display the content of the selection on the dashboard charts.
- Pie chart: to visualize the relative launch success rate among the launch sites and to visualize the repartition between success and failure of launches for any given site.
- RangeSlider with callback to scatter chart: to select which payload mass range to visualize for easier interpretation of the chart
- Scatter chart of launch result vs payload mass: to investigate the relationship between the payload mass and the outcome of launches.
- GitHub URL of the completed Plotly Dash lab: https://github.com/fe-htt/testrepo/blob/main/Module%203/spacex-dash-app.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

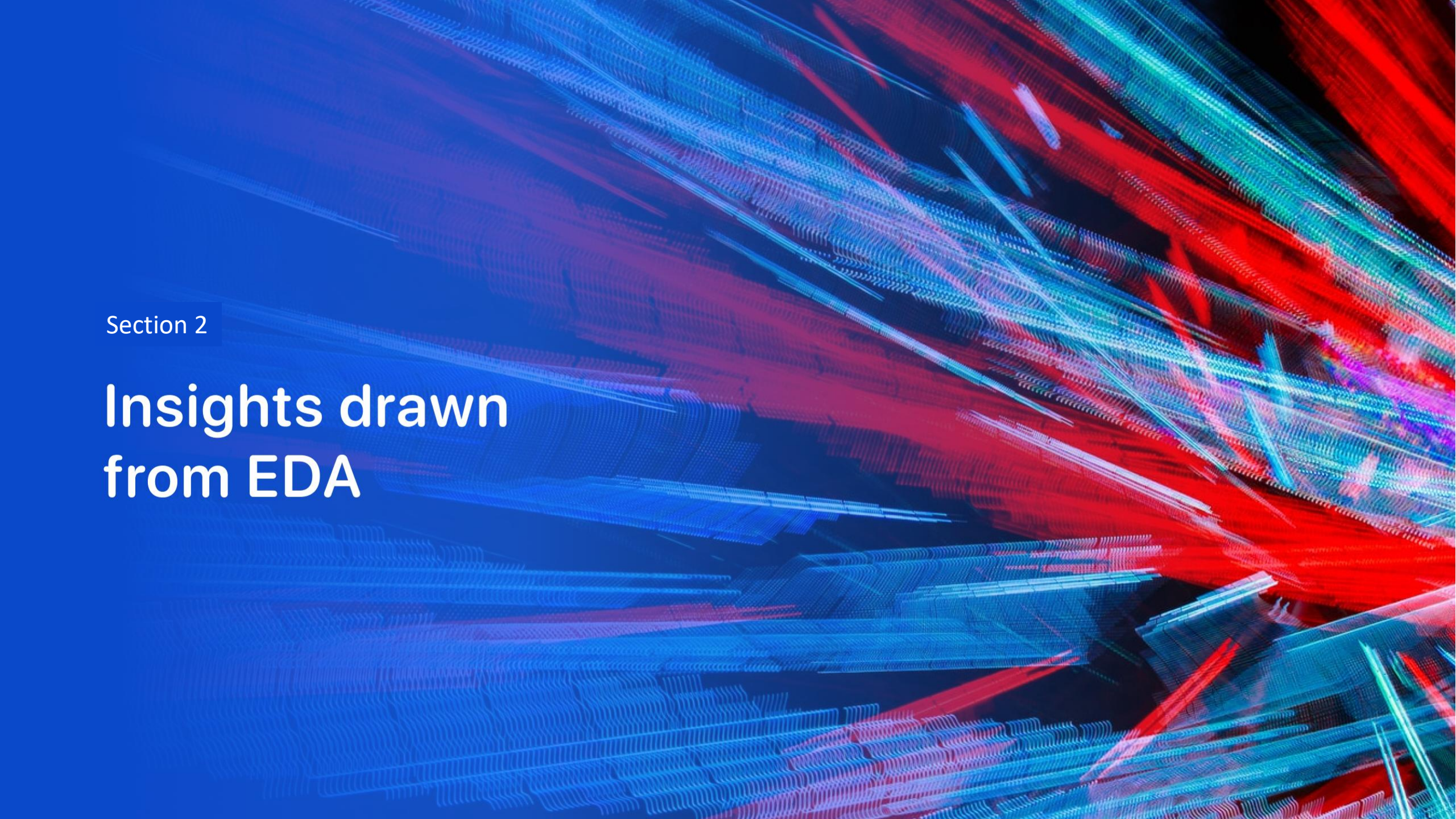
Predictive Analysis (Classification)

- Data standardization with StandardScaler()
- Splitting of data in training and testing datasets with train_test_split
- Creation of logicsti regression obect and cridSearchCV object, fitting of the model> finding of the best hyperparamenters and of the model Score on test data
- Repetition of the last step on SVM, Tree classifier and KNN models.
- Finding of model with maximum score on test data
- GitHub URL of the completed predictive analysis lab: https://github.com/fe-htt/testrepo/blob/main/Module%204/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

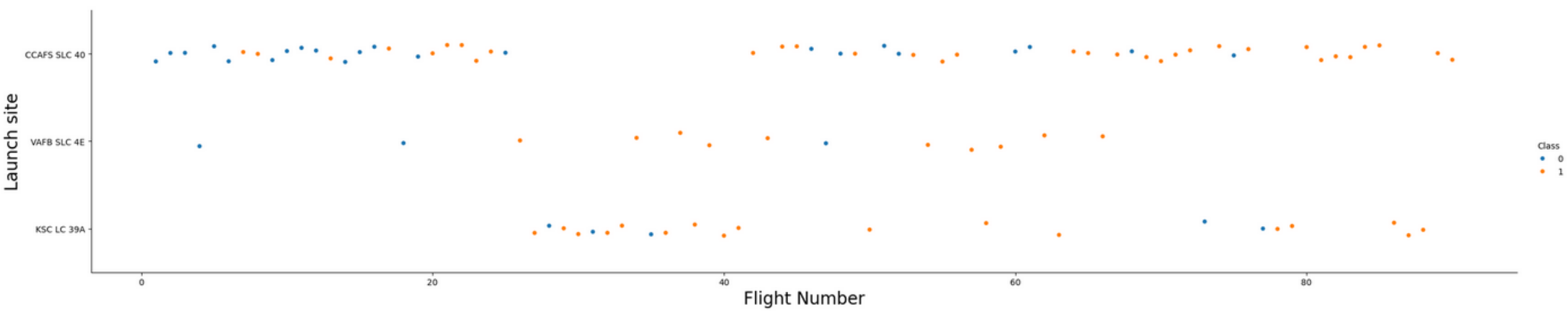


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

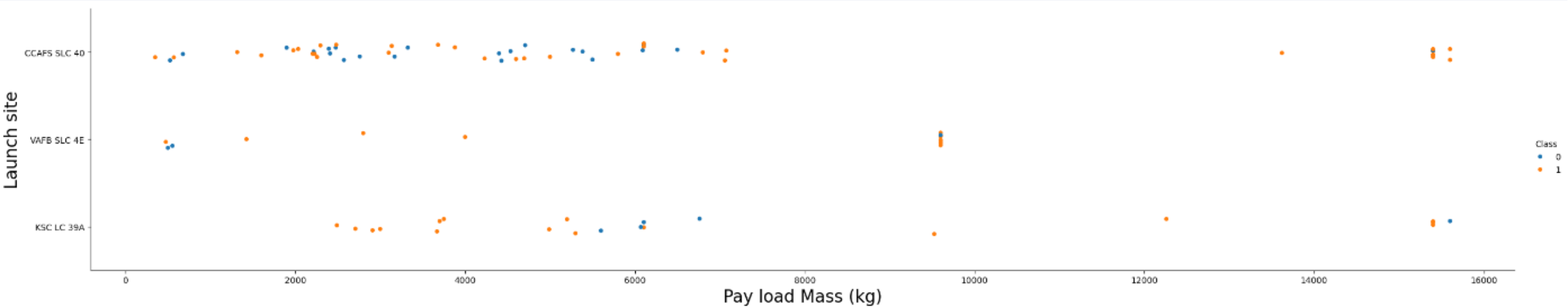

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



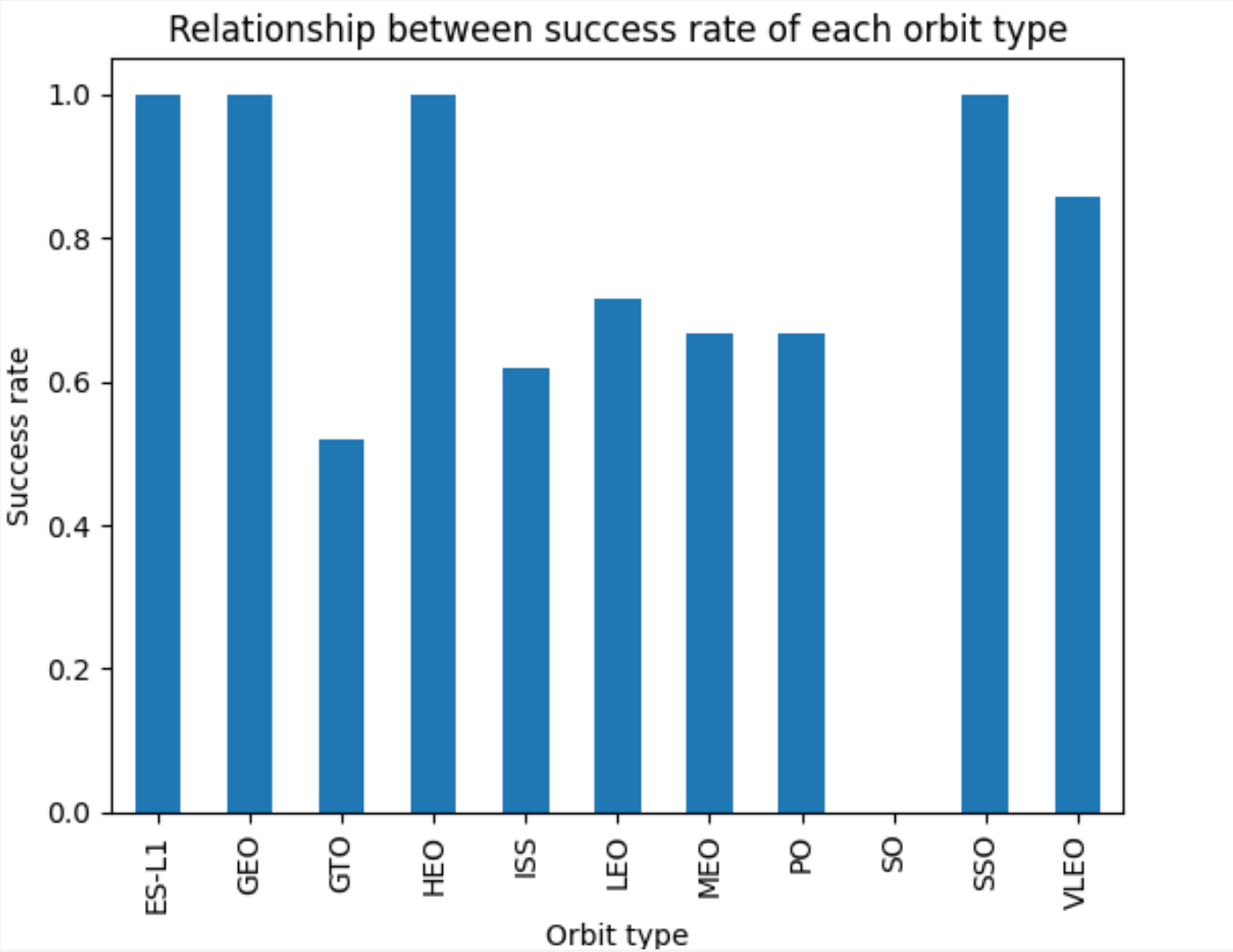
- It is clear how in the beginning launches were made predominantly at the CCAFS SLC 40 site, and lates a few sequences of launches took place over all sites. A few launch campaign can be visually identified and it can also be seen of the success rate tends to increase with the increase of flight number

Payload vs. Launch Site



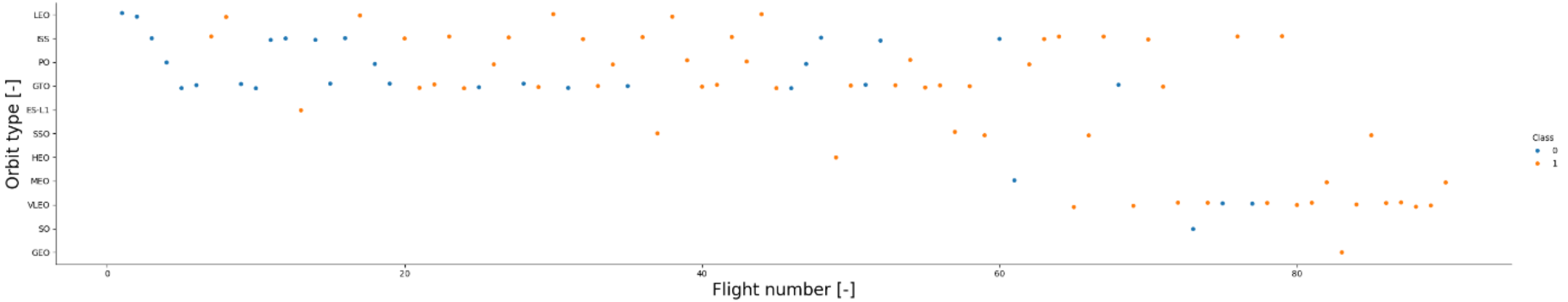
- For the VSFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000 kg). It is also evident that the most used launchsite is CCAFS SCL 40

Success Rate vs. Orbit Type



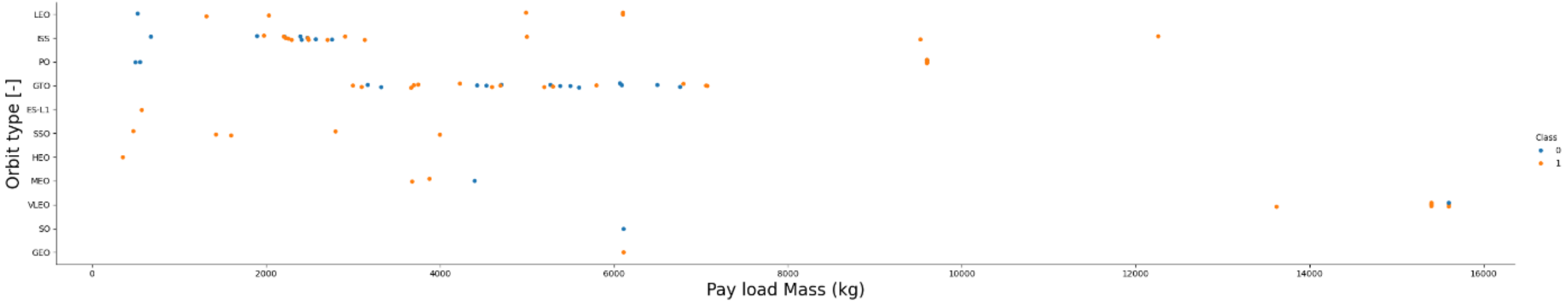
- There is no clear trend that emerges from this plot. What can be said is that for orbits ES-L1, GEO, HEO and SSO the success rate is close to 1

Flight Number vs. Orbit Type



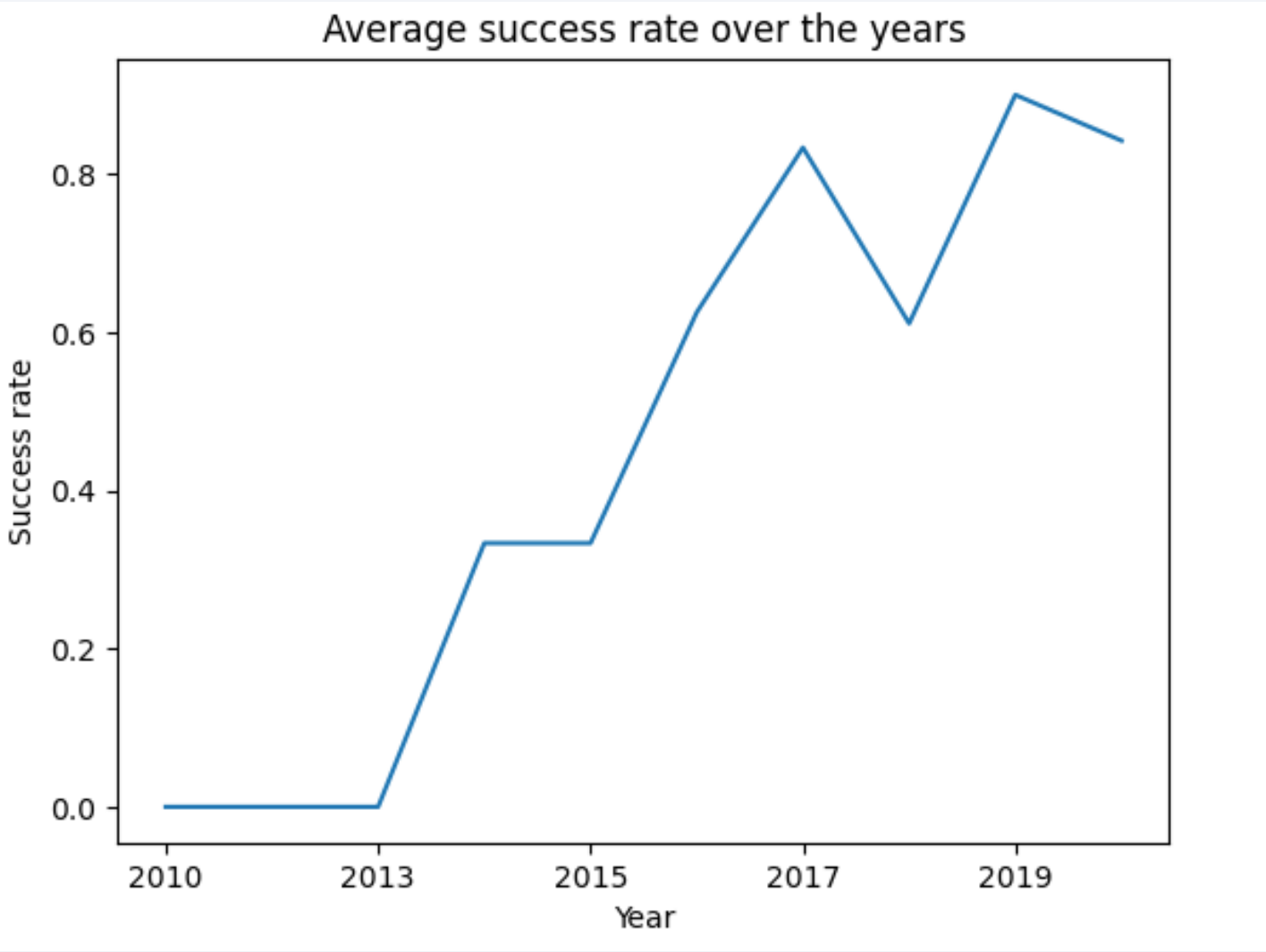
- In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

All Launch Site Names

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

- The unique launch sites are obtained from the SPACEXTABLE by querying it using the distinct keyword. $ launchsites are found.
- %sql select distinct "Launch_Site" from SPACEXTABLE

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	18:45:00	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- The query uses the keywork where in combination with like to locate rows that contain the string CCA. Also the wildcard character % is used to catch all instances of “CCA”. The keywork limit followed by 5 limits the rows displayed to 5. All of the 5 launches selected were successful and directed to LEO orbit.
- %sql select * from SPACEXTABLE where "Launch_Site" like "CCA%" limit 5

Total Payload Mass

SUM("PAYLOAD_MASS__KG_")
45596

- The total payload mass is obtained by summing the payload mass of all launches related to NASA (CRS). This total mass is just above 45000 kg.
- %sql select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where Customer like "NASA (CRS)"

Average Payload Mass by F9 v1.1

AVG("PAYLOAD_MASS__KG_")
2928.4

- The average payload for the booster version F9 v1.1 is obtained by selecting the average of the payload mass for the rows where the booster version is F0 v1.1. This average payload mass is just below 300 kg.
- %sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" like "F9 v1.1"

First Successful Ground Landing Date

MIN(Date)
2015-12-22

- The dates of the first successful landing outcome on ground pad is found by selecting among the records of successful ground pad landings the one with the smallest date. Said date is the 22nd of December 2015.
- %sql select MIN(Date) from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- By selecting the entries that contain the characters ‘drone ship’ in the Landing_Outcome field, that have “Success” in the “Mission_Outcome” field and that have a payload mass between 400 and 6000 is possible to extract the desired list of booster versions. There are 5 booster version that satisfy these conditions.
- %sql select "Booster_Version" from SPACEXTABLE where ("Landing_Outcome" like "%drone ship%") and ("Mission_Outcome" = "Success") and ("PAYLOAD_MASS__KG_" between 4000 and 6000);

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

- 4 types of mission outcomes are possible. Among these the overwhelming majority of instances fall into Success, while only one falls into Failure (in flight)
- %sql select "Mission_Outcome", COUNT("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome";

Boosters Carried Maximum Payload

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

- There are 12 booster versions that have carried the maximum payload mass.
- %sql select distinct "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") from SPACEXTABLE);

2015 Launch Records

substr(Date, 6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Two failures are found, both from the launch site CCAFS LC-40

- %sql select substr(Date, 6,2), "Landing_Outcome", "Booster_version", "Launch_Site" from SPACEXTABLE where (substr(Date,0,5)='2015') and ("Landing_Outcome" = "Failure (drone ship)");

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

]:

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

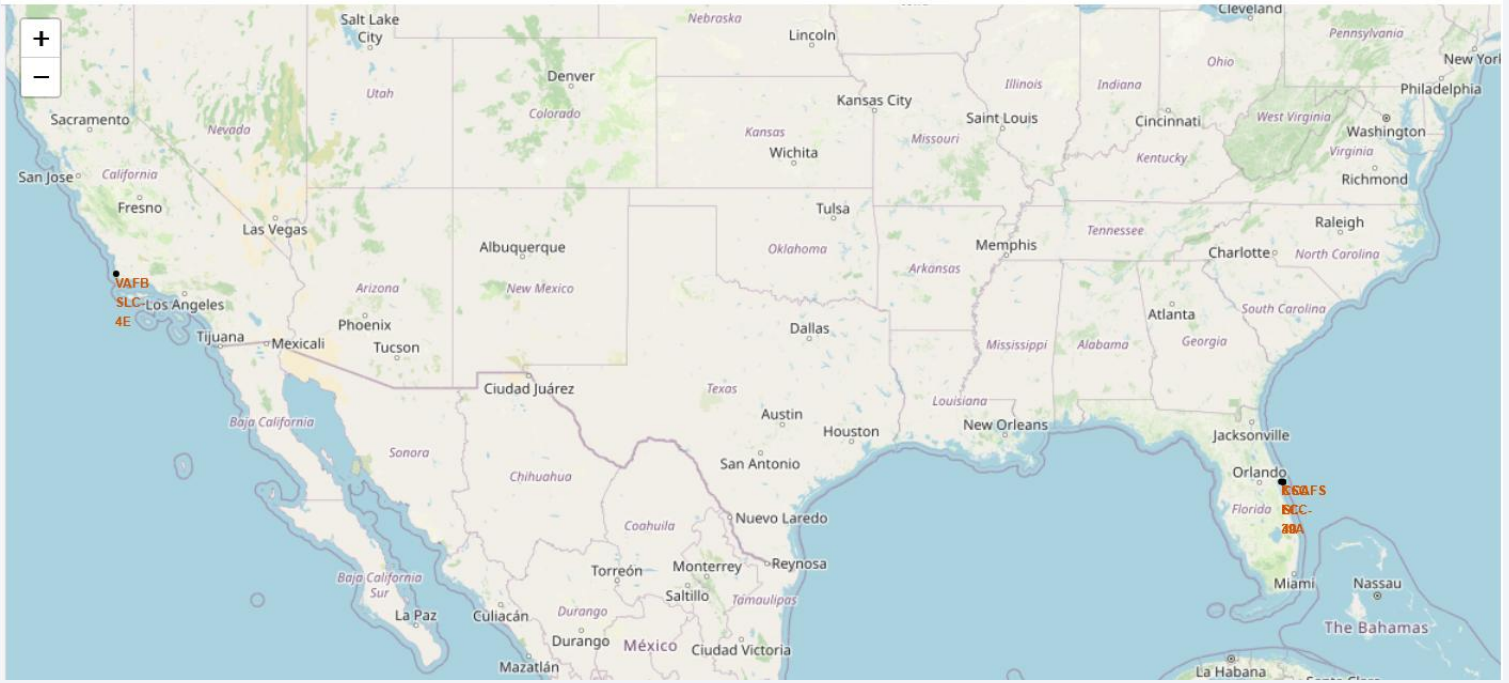
- The most commond outcome is no attempt, followed by Succes (drone ship)

- %sql select "Landing_Outcome", count("Landing_Outcome") as Count from SPACEXTABLE where (Date between "2010-06-04" and "2017-03-20") group by "Landing_Outcome" order by Count desc;

Section 3

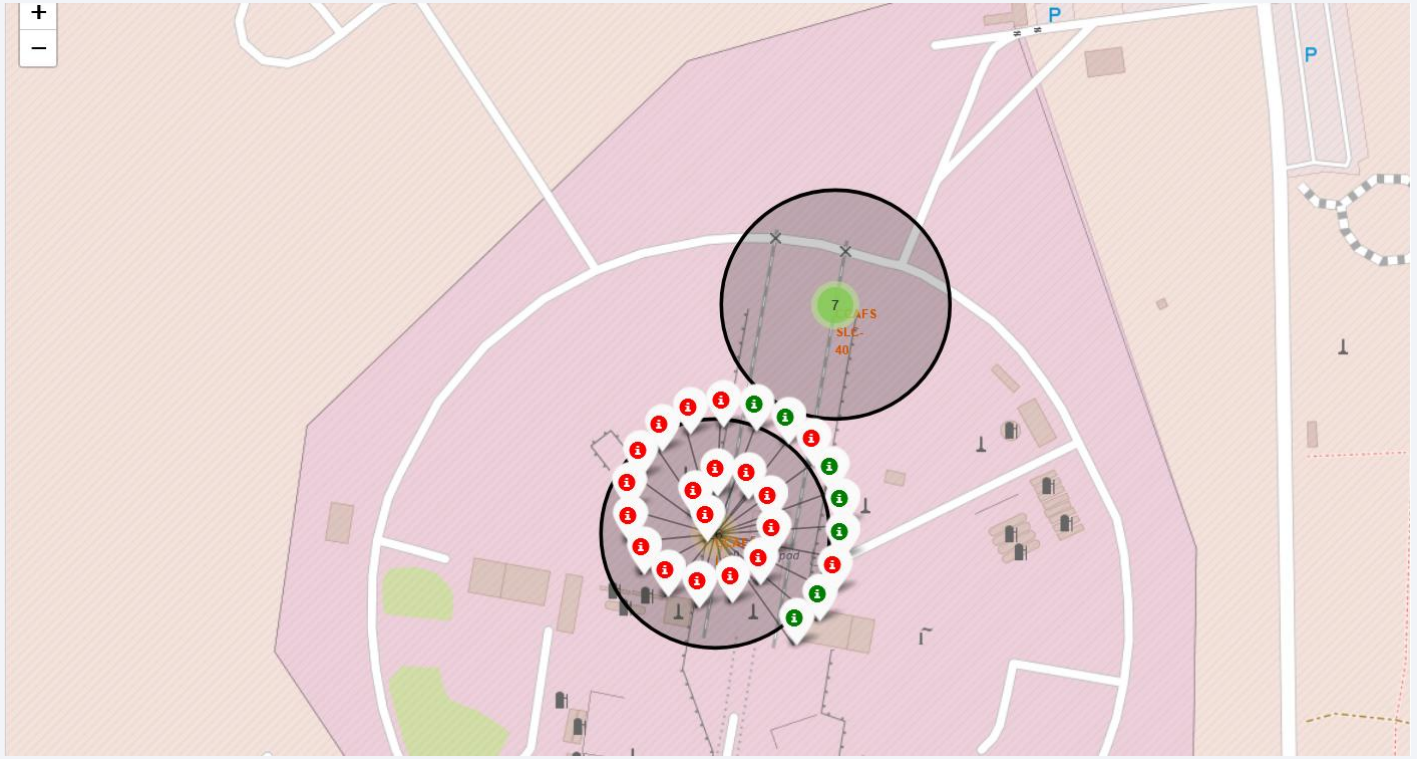
Launch Sites Proximities Analysis

Visualization of launch sites on a map



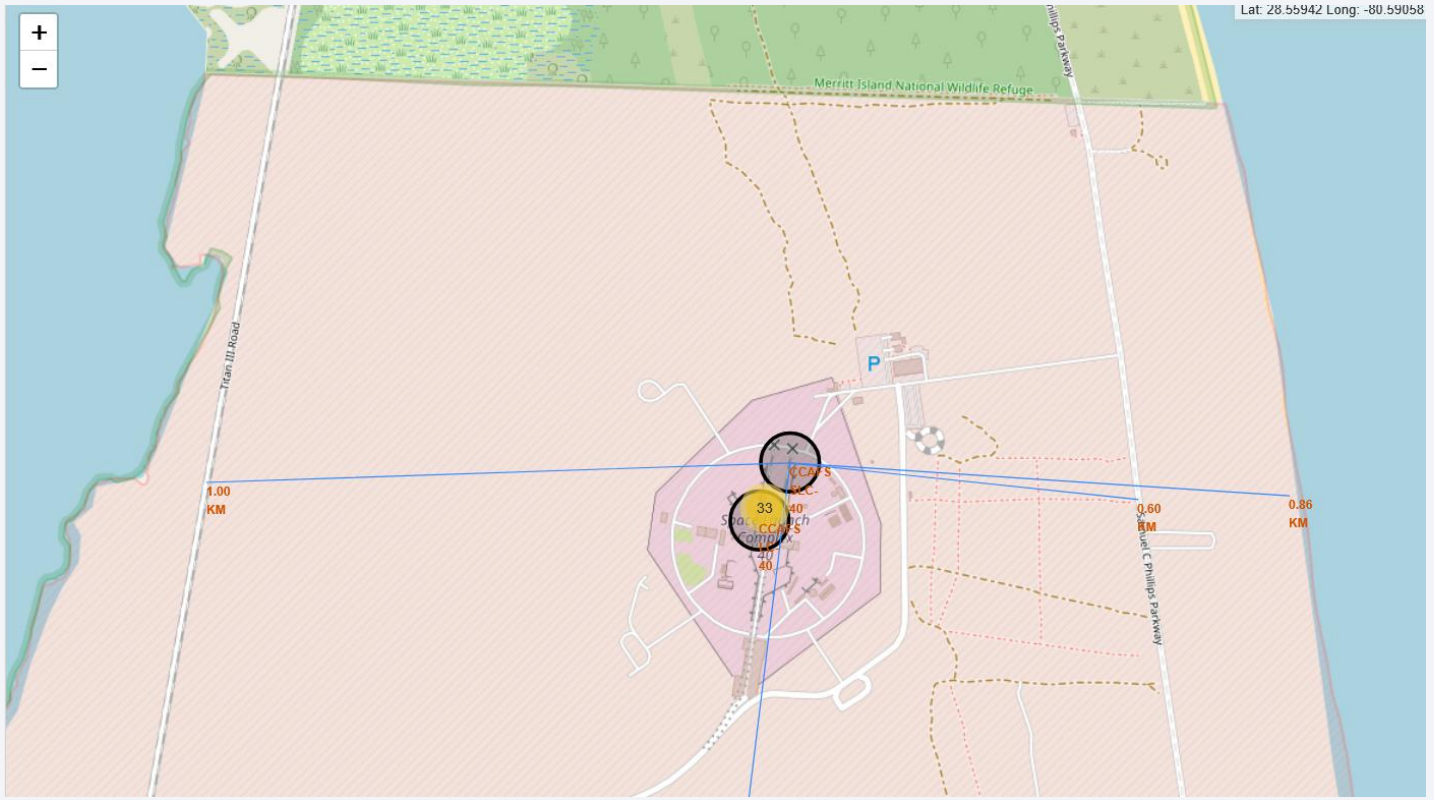
- Among the sites, three are located on the east cost, in Florida, while one is located on the West coast, in California. All sites are in close proximity to the sea.

Marking of success/failed launches for each site on the map

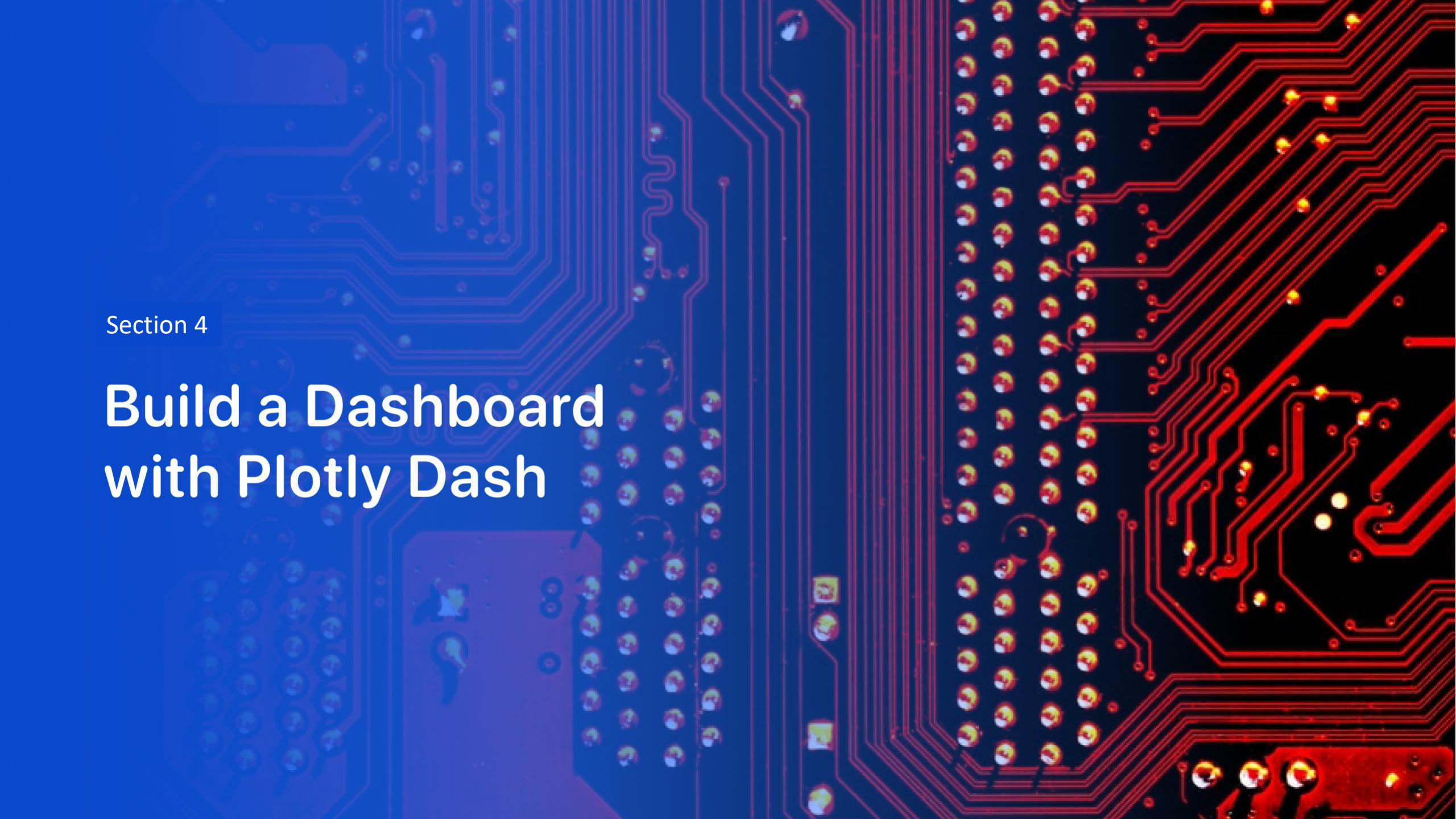


- The Vanderberg launch site is a relatively low success rate, with more failed lunches that successful ones.
- The Cape Canaveral 39A is quite successful, with very few failed compared to the successful ones
- The Cape Canaveral SLC40 launch site is a relatively low success rate, with more failed lunches that successful ones.
- The Cape Canaveral LC40 launch site is a relatively low success rate, with more failed lunches that successful ones. Also, this site has many launches.

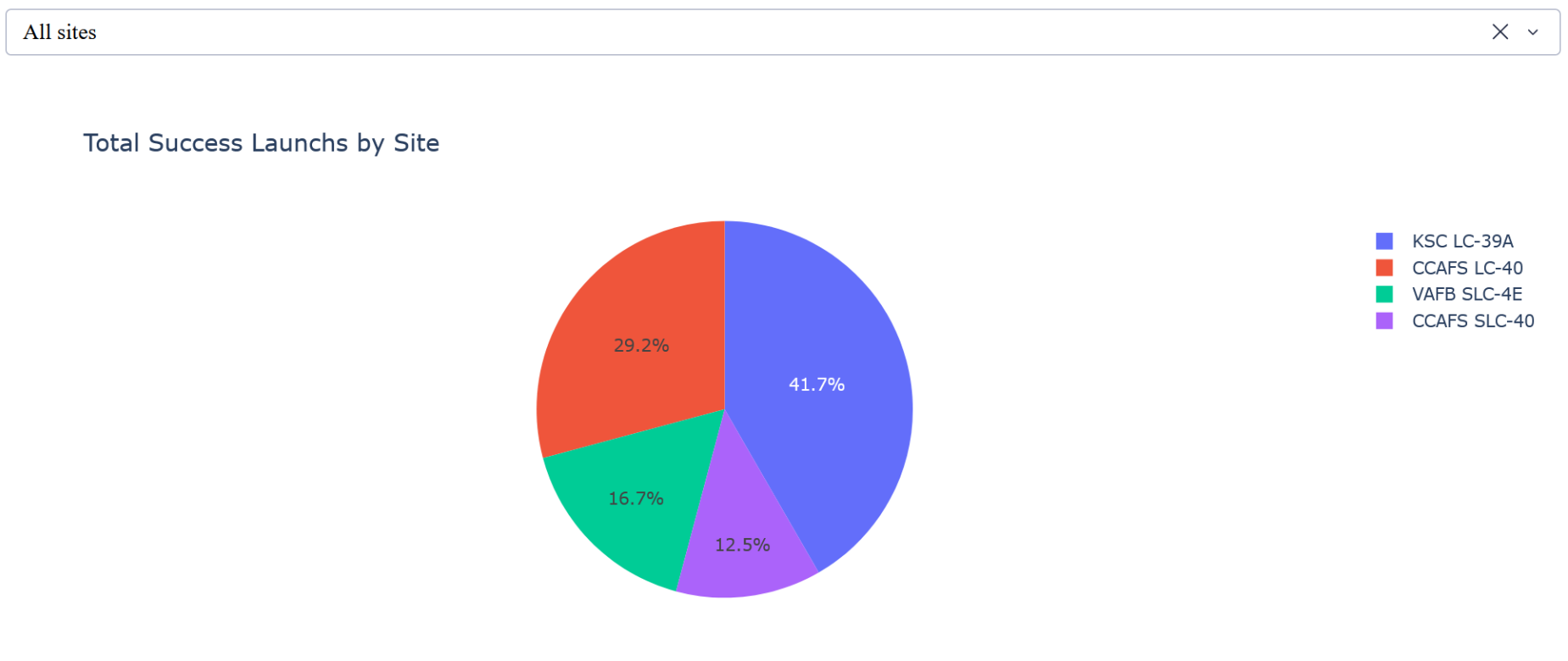
Visualization of distances between a launch site and points of interest



- xplain the important elements and findings on the screenshot
- launch sites are in close proximity to railways (in the case of Cape Canaveran, about 1 km)
- launch sites are in close proximity to highways (in the case of Cape Canaveran, less than 1 km)
- launch sites are in close proximity to coastline (in the case of Cape Canaveran, less than 1 km)
- launch sites keep certain distance away from cities (in the case of Cape Canaveran, it is about 55 km away rfom Melbourne)

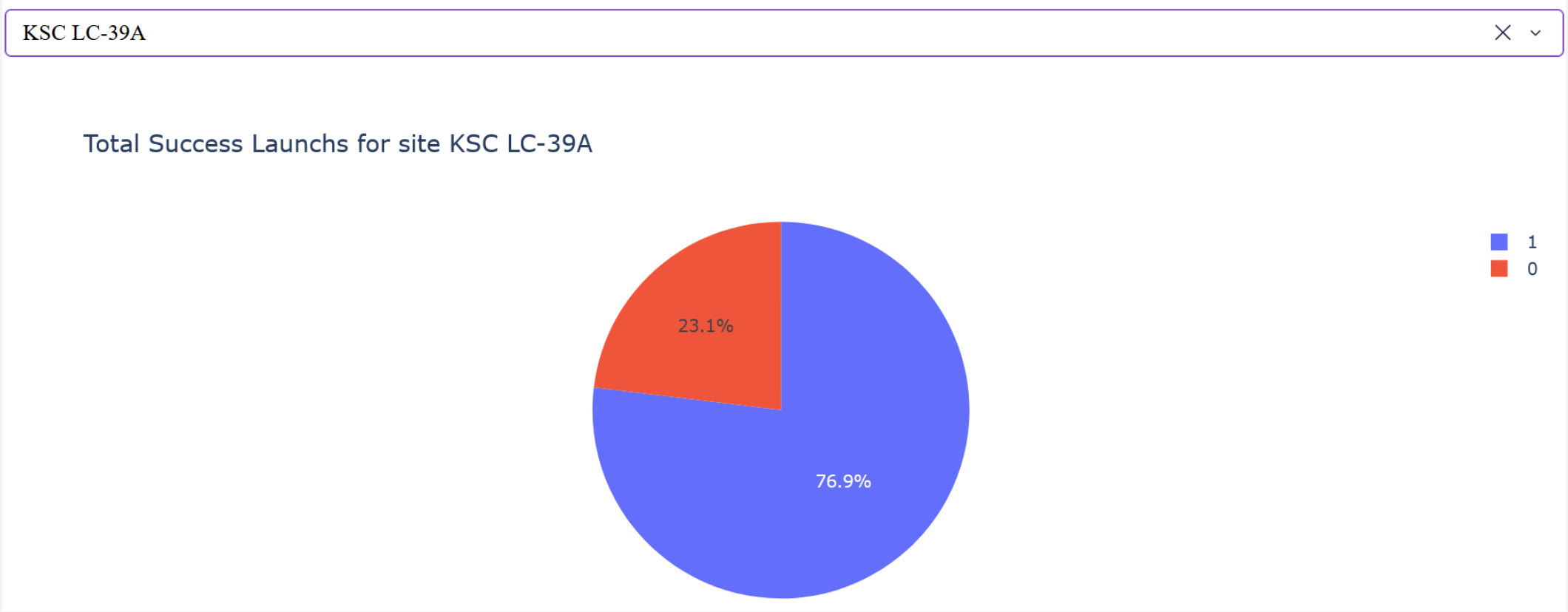

Section 4
Build a Dashboard with Plotly Dash

All sites: launch success count



- The piechart shows the launch success count for all sites. It is evident how almost half of all successful launches have been made at the KSC LC-39A site, followed by the CCAFS LC-40 accounting for just below one third of total successful launches

KSC LC-39A launch success ratio



- The pie chart shows the launch success ratio of the site KSC LC-39A. This site is the one with the highest ratio, with 76.9% of successful launches overall!

Correlation between payload and success for all sites

- In the figures a scatter plot showing the relationship between payload mass and result of the mission is shown. From the top plot we can wee how most successful launches have been made in the range of 2000 to 6000 kg, while above 6000 kg there is only 1 successful launch. Zooming in the range 2000 to 4000, it can be seen how the booster version category BT is the most successful one, accounting for the overwhelming majority of successful launches

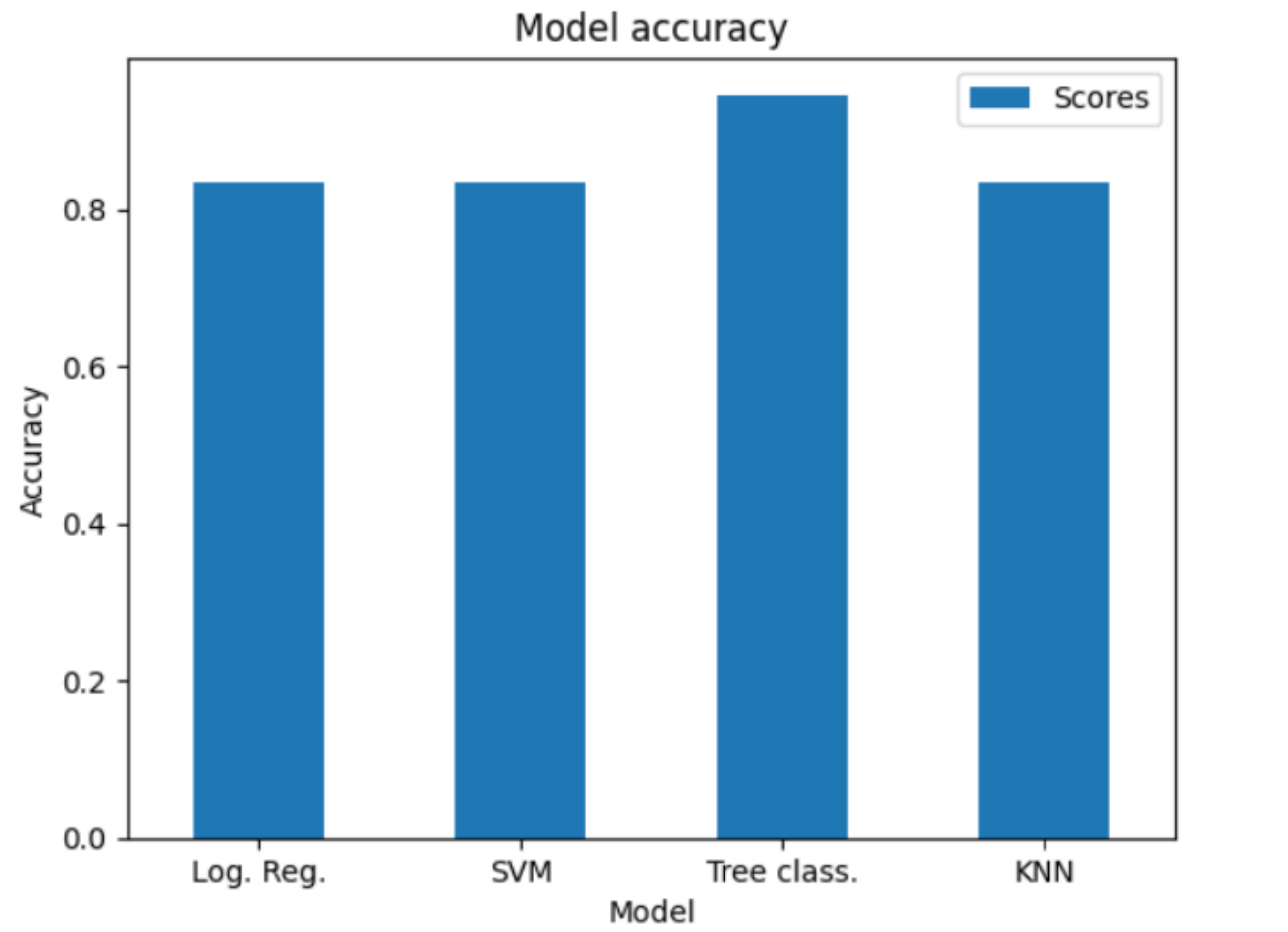
Correlation between payload and success for all sites

- Zooming in the range 6000 to 1000, we can see how very few launches have been made and the FT booster version was never successful, while the B4 version managed to deliver a successful mission close to 1000 kg.

Section 5

Predictive Analysis (Classification)

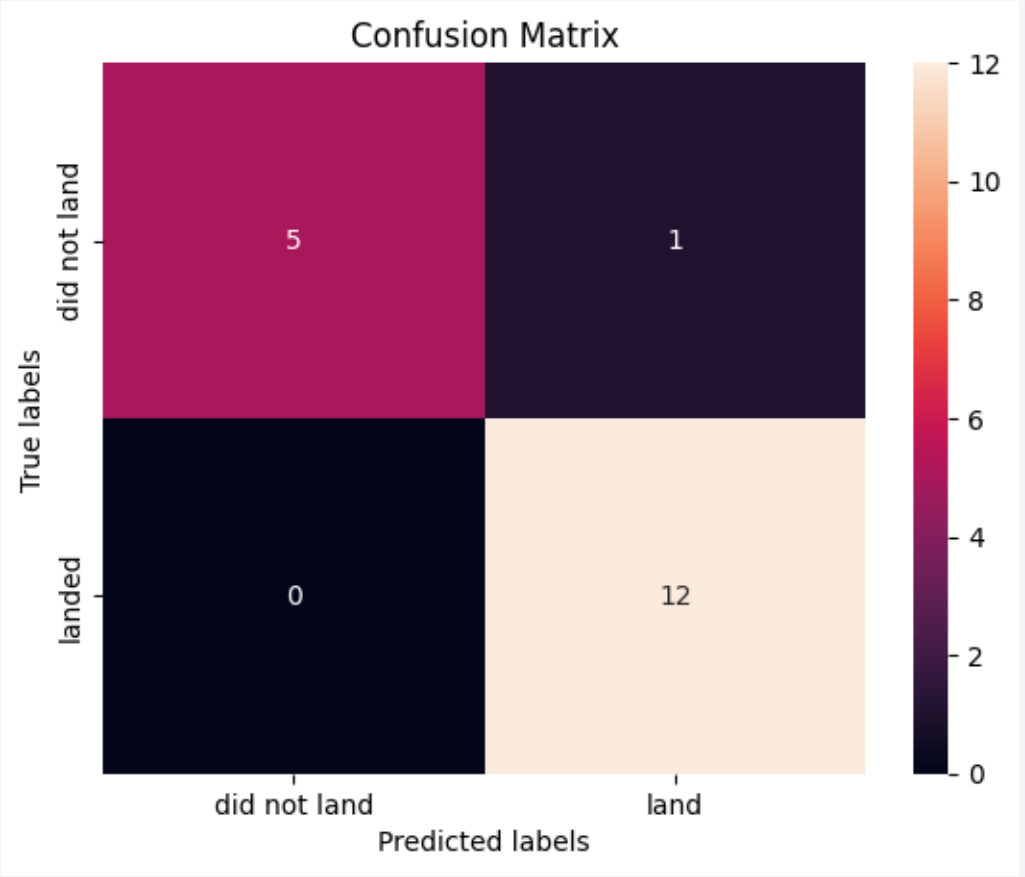
Classification Accuracy



- The tree classifier model has the highest accuracy, while the logistic regression, SVM and K-Nearest Neighbor perform all equally worse.

Confusion Matrix

- The figure shows the confusion matrix for the Tree Classifier model
- We see that the model can distinguish between the different classes, with only one false positive and no false negatives.



Conclusions

- The vast majority of launches have been made with payloads below 7000 kg
- The success of launches steadily increased from 2013 to 2020
- Launch sites are located in close proximity to the coast and to critical infrastructures such as railways and highways. They keep a certain distance from the cities
- Among the launch sites, the one with the highest success ratio is KSC LC-39A (76.9%), which is also the site where the majority of launches have been made (41.7%)
- 4 models have been trained on the SpaceX data, among these the tree classifier performs best with an accuracy of 83.34% on the test data. Only one false positive was predicted during testing.
- We gained insight into SpaceX rocket launch operations and are able to make informed prediction on the result of future launches

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!