IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Florian Breut
09/03/2025

Outline

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

Executive Summary

- This study is exploiting data from the SpaceX Rocket API and Wikipidia to highlight relashionships between booster launch parameters and probability of success of the mision
- One outcome of the study is that the success rate of SpaceX mission has been improving with flights and years, and the reusable rocket launch buisness seems viable after 2020 as the probability of success approaches 80%
- This study show that using the available data, we can predict the outcome of a mission with an accuracy close to 87%
- This study also highlight that there is still improvement to find in the launching of 5000-9000kg payloads where there are no example of successful mission

Introduction

- SpaceX has successfully designed and tested rockets with a reusable first stage.
  - It's a ground breaking technical solution
  - Dropping launching costs from 165 to 65 million dolllars
- Using the open data provided by SpaceX about its Falcon 9 rockets, we would like to
  - Understand what factors are contributing to a successful mission
  - Train a model to predict the success of a specific mission
  - Evaluate the profitability of such a buisness model

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how the data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The SpaceX Rocket Launch data API https://api.spacexdata.com/v4 was used to collect detailed information about each SpaceX Launch

Call SpaceX API → Convert to Python Data Frame → Filter Falcon 9 Data → Save as dataset_part_1.csv

- Webscrapping was used to extract data from the Falcon 9 Launch wikipedia page https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

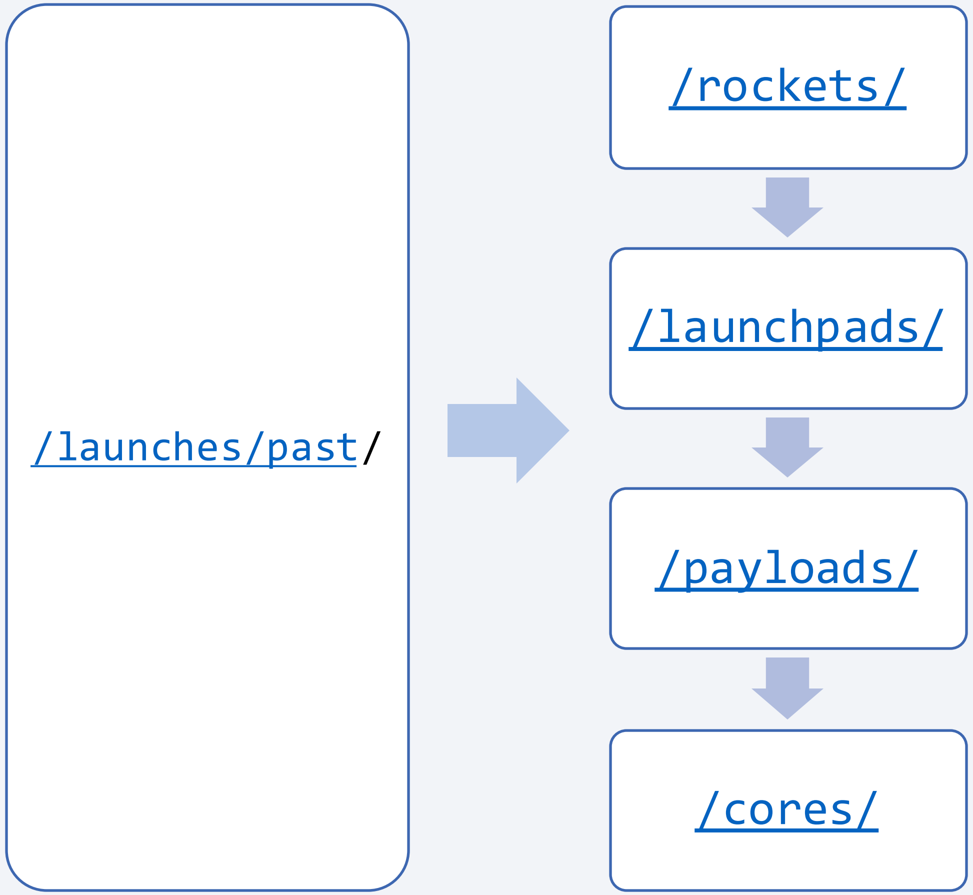
Request Falcon 9 Wiki page html → Webscraping Wiki page → Convert to Python Data Frame → Save as spacex_web_scraped.csv

- Data were processed using Python and saved as csv tables

Data Collection – SpaceX API

- We used the https://api.spacexdata.com/v4 API
- A call to the launches/past endpoint was made to collect data about each launch
- To get more details about the features returned by the launches/past endpoint we made calls to other endpoints.

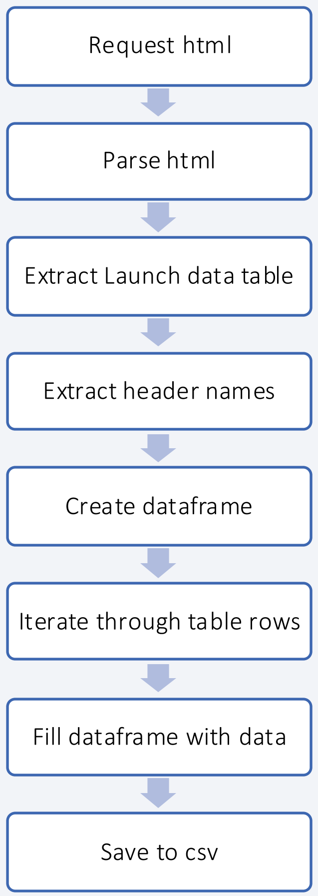
https://github.com/FlorianBreut/ads_capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- We used the Request library to get the html content of the Falcon 9 Launch wiki page
- We used BeautifulSoup to parse html and extract information from tables
- We filled a Pandas Dataframe with the extracted data and exported as a CSV file

https://github.com/FlorianBreut/ads_capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- The dataset was preprocessed to better understand the data, evaluate its quality and the number of missing values
- A new feature "Class" was created to represent the binary outcome of a mission (Success or Failure)

Calculate the number of missing values per feature → Visualize each feature type → Count launch site use → Count orbit use → Classify landing outcomes → Create the Class column → Assign landing outcome to the Class Column

https://github.com/FlorianBreut/ads_capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plot were used to visualize the influence of the payload mass, orbit type, flight number and launch site on the success rate of the mission
- Bar plot was used to visualize the success rate for each of the Orbit types

https://github.com/FlorianBreut/ads_capstone/blob/main/edadataviz.ipynb

EDA with SQL

Using SQL we explored the SpaceX launch dataset and gained more insights about the influence of the launch site and Booster version

- We listed the name of each launchsite
- We displayed the first 5 records for a CCA launch site
- We found the total payload mass that was launched by SpaceX
- We found the average payload mass carried by the F9 v1.1 Booster
- We found the first successful landing on a ground pad
- We listed the Boosters that successfully landed on a drone ship
- We found how many missions were successful and how many failed
- We found which Booster Version were able to carry the maximum payload

https://github.com/FlorianBreut/ads_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Using Folium, we displayed a Map with the folium.Map object
- We diplayed the different SpaceX launch sites using Marker and Cirle objects
  - The marker color indicates if the mission was a success or not
  - The name of the Launching pas is displayed
- We added a MousePosition object to get Real time Latitude and Longitude data on when moving the cursor on the Map in order to find Map features such as Highways, Coastlines, or cities
- We used the PolyLine Object to Visualize a line between the launch pads and a closest object of a specific type

https://github.com/FlorianBreut/ads_capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The Dashboard features a pie chart showing how much each launching site contributed to the successful missions
- A dropdown menu allow the user to select a specific launching site and visualize the success rate of this site
- A scatter plot shows the relationship between the payload mass and the succes of the mission
- A Slider allows the user to filter datapoints by selecting the minimum and maximum payload to display

https://github.com/FlorianBreut/ads_capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- A Model was built to predict the success of a mission base on the set of features present in the dataset_part_2 dataset
- We converted the features and labels to a format suitable for training with GridSearchCV
  - Label : The Class column is converted to a numpy array
  - Features : The dataset is scaled with the fit_transform function
- We trained the data for different models : SVM, KNN, Classification Tree, and Logistic Regression
- For each model architecture we found the best hyperparameter using a grid of parameters
- To find the best model we compared the accuracy on the test dataset and the confusion matrix

https://github.com/FlorianBreut/ads_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

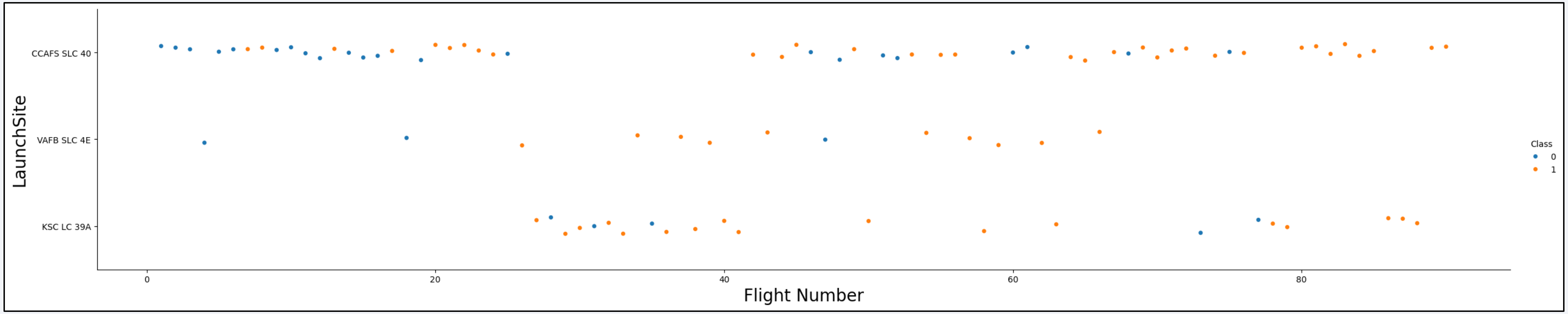
Results

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

Section 2

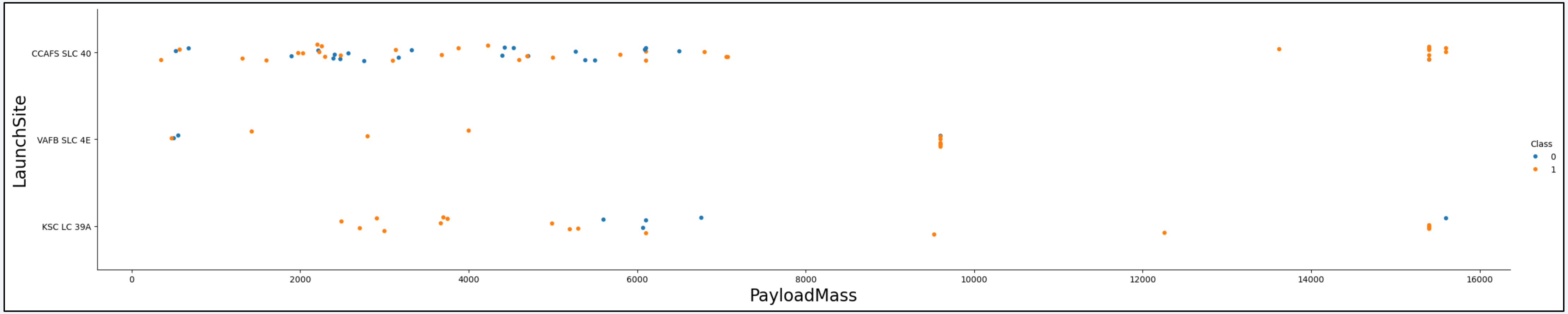
Insights drawn from EDA

Flight Number vs. Launch Site



- Most Failed mission have been launched from CCAFS SLC 40 and correspond mostly to the early tests (first 20 launches)
- The success rate increases with the number of the flight independantly from the launch site

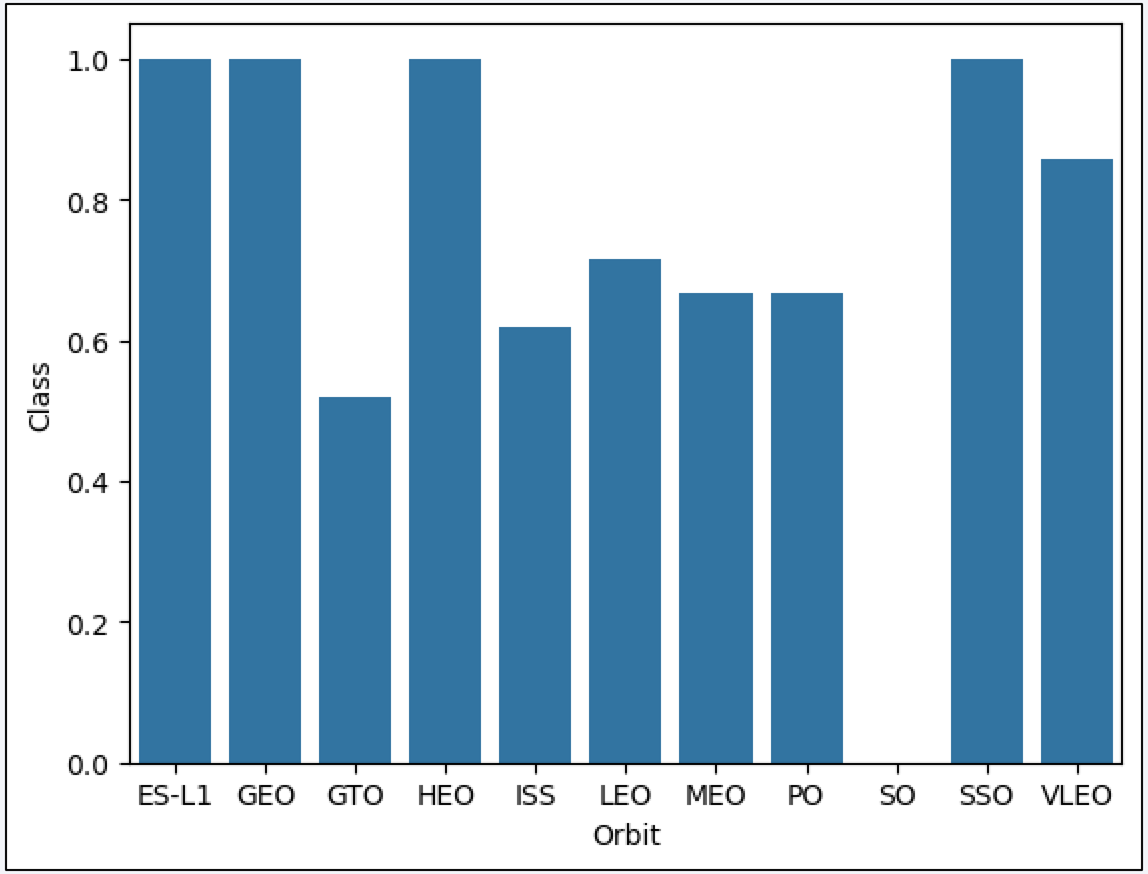
Payload vs. Launch Site



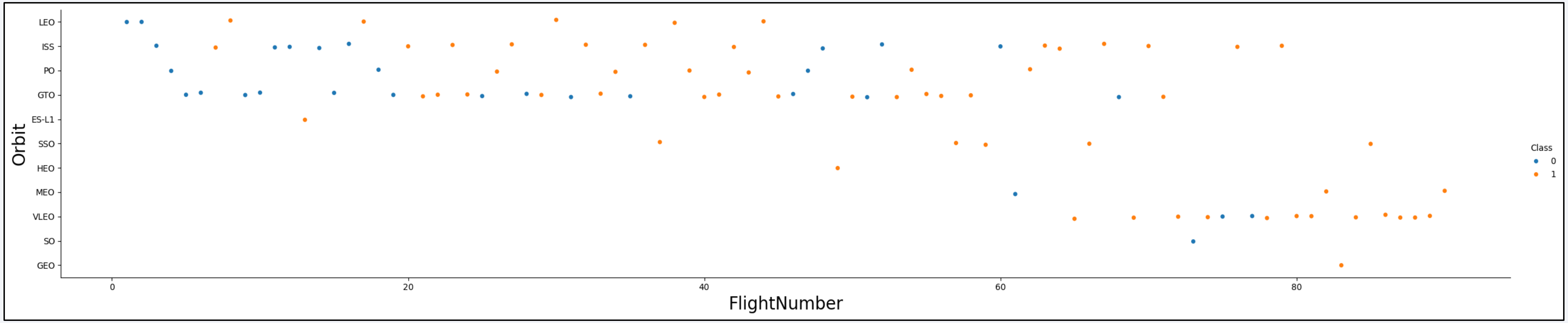
- Most mission with a payload greater than 8000kg were successful
- No payload above 10000kg were launched from VAFB-SLC
- Most failures for the KSC LC launch site are for 6000kg payloads

Success Rate vs. Orbit Type

- No successful missions were conducted for a SO orbit
- ES-L1, GEO, HEO, SSO all have a very high success rate
- For the GTO, ISS, LEO, MEO and PO orbits the probability of success is moderate

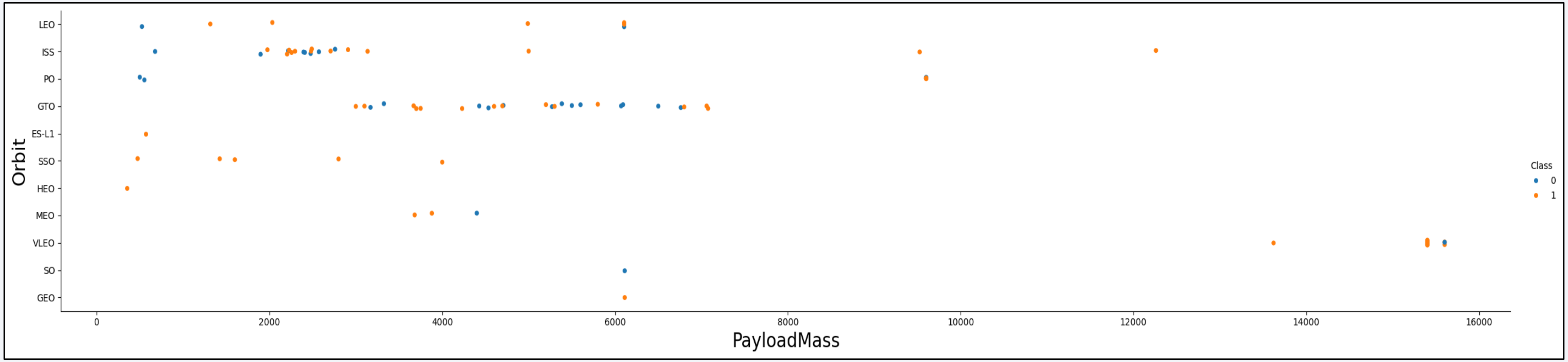


Flight Number vs. Orbit Type



- Early tests were mostly conducted for LEO, ISS, PO, GTO and ES-L1
- SSO, HEO, MEO, VLEO, SO and GEO Orbit were launched in higher flight numbers

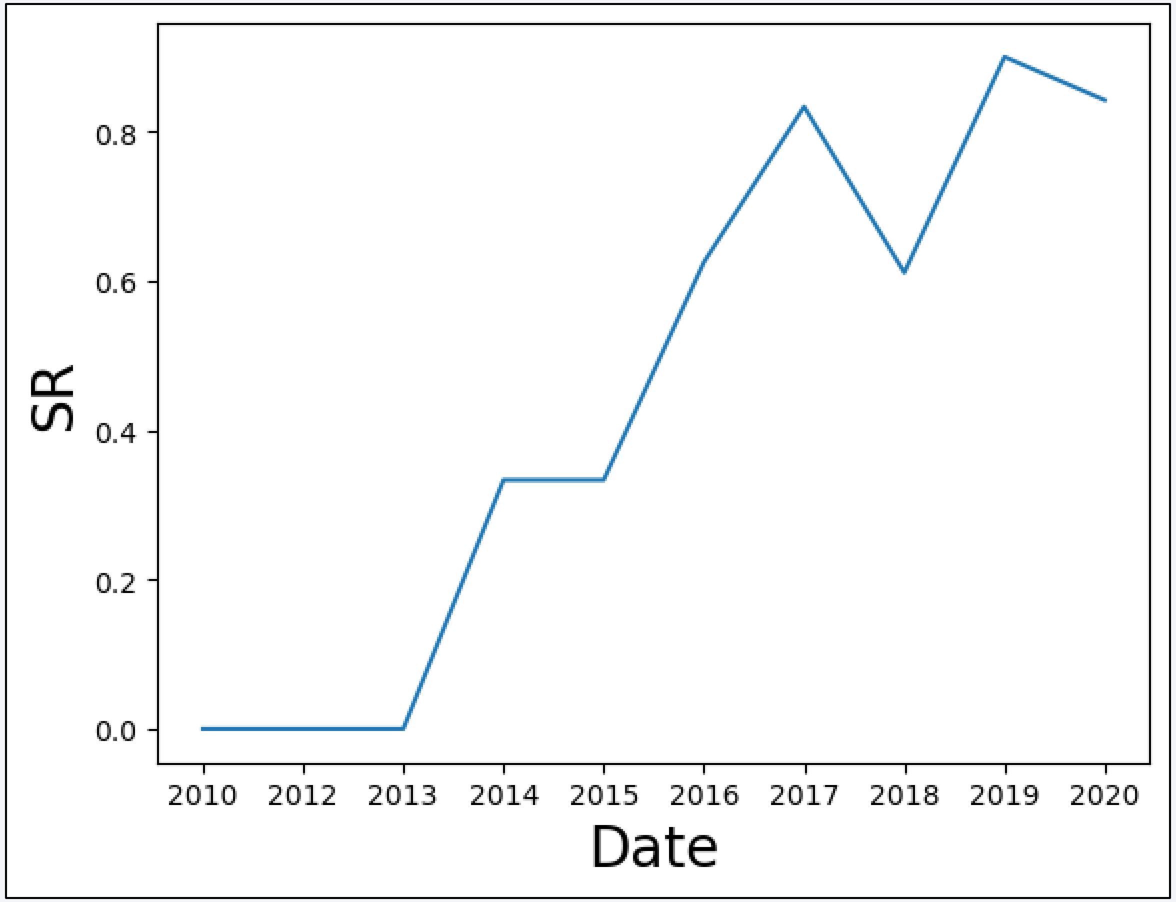
Payload vs. Orbit Type



- 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

- This trend clearly shows that the success rate increases with time
- However 2018 and 2020 doesn not comply with this trend



All Launch Site Names

- We extracted the unique launch site names using SQL command
  - `%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE`

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

Launch Site Names Begin with 'CCA'

- We displayed 5 records where the launch sites begin with CCA using :
  - `%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5`

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

Total Payload Mass

- We displayed the total payload mass carried by NASA boosters using
  - `%sql SELECT SUM(CAST("PAYLOAD_MASS__KG_" AS FLOAT)) FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';`

SUM(CAST("PAYLOAD_MASS__KG_" AS FLOAT))
45596.0

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1 :
  - ```
    %sql SELECT AVG(CAST("PAYLOAD_MASS__KG_" AS FLOAT)) FROM SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 V1.1%';
    ```

AVG(CAST("PAYLOAD_MASS__KG_" AS FLOAT))
2534.6666666666665
    ```

First Successful Ground Landing Date

- To find the dates of the first successful landing outcome on ground pad:
  - `%sql SELECT * FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (ground pad)" ORDER BY "Date" LIMIT 1`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
    - ```
      %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE ("Landing_Outcome"="Success (drone ship)" AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000)
      ```

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

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failure mission outcomes:
  - `%sql SELECT COUNT(*) FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "Success%";`
  - `%sql SELECT COUNT(*) FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "Failure%";`

COUNT(*)
10

Boosters Carried Maximum Payload

- To list the names of the booster which have carried the maximum payload mass:
  - ```
    %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)
    ```

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

- To list the failed landing_outcomes in drone ship, their booster versions, and launch site names for 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" LIKE "Failure (drone ship)")`

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

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

- To 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(*) FROM SPACEXTABLE WHERE ("Date" > "2010-06-04" AND "Date" < "2017-03-20") GROUP BY "Landing_Outcome" ORDER BY COUNT(*) DESC

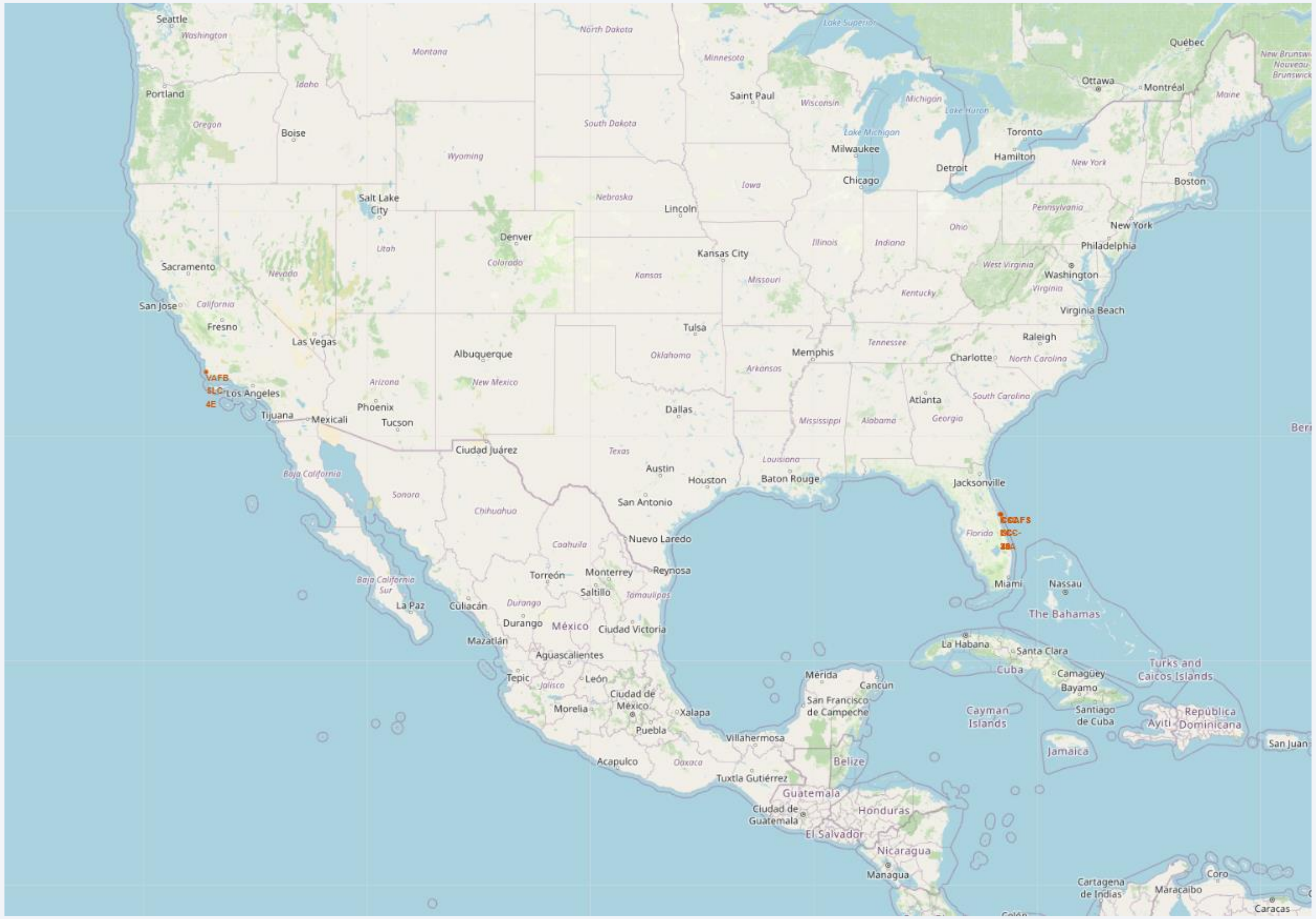
Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

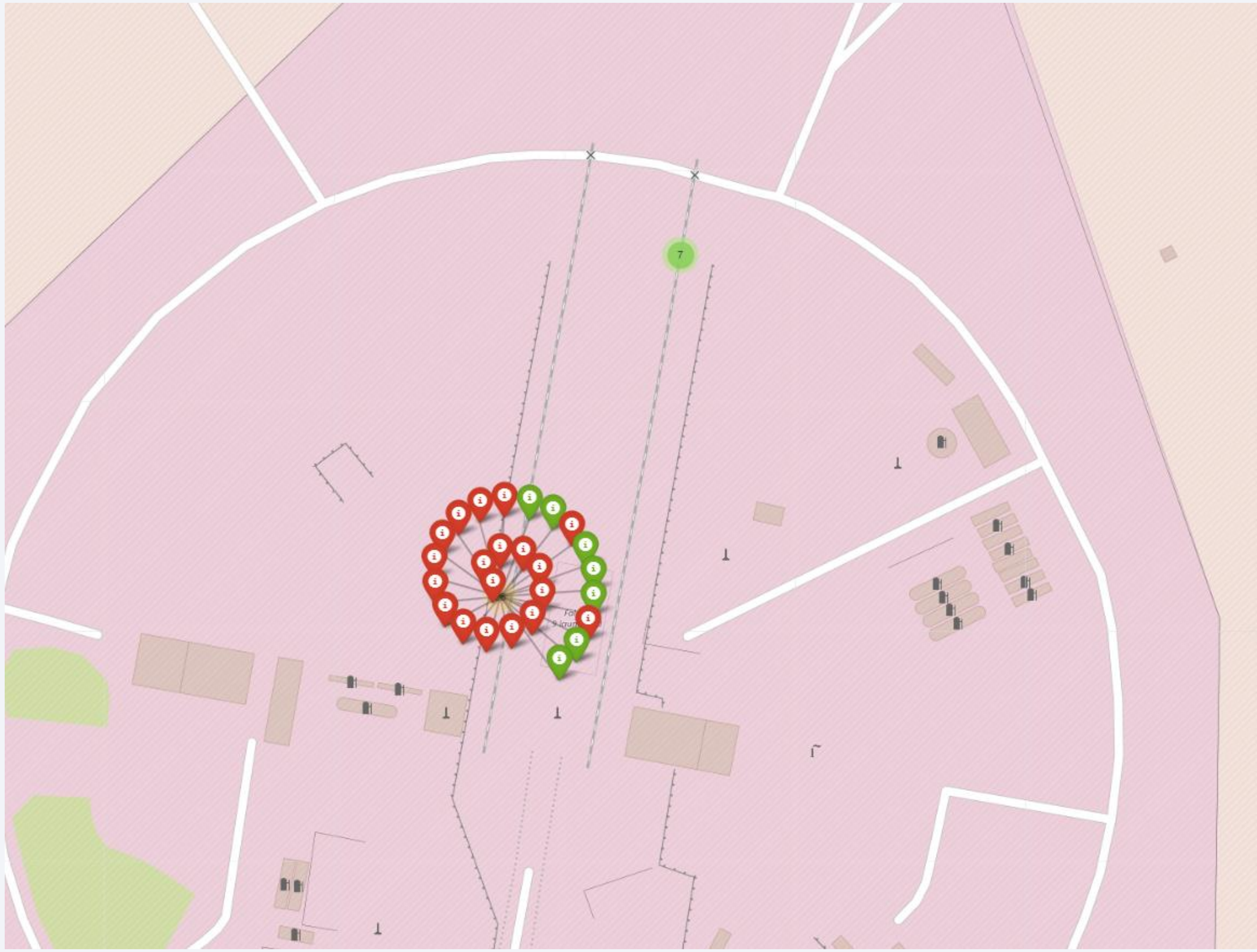
SpaceX Launch Sites

- VAFB SLC-4E is located in california
- All other launch sites are located in Florida



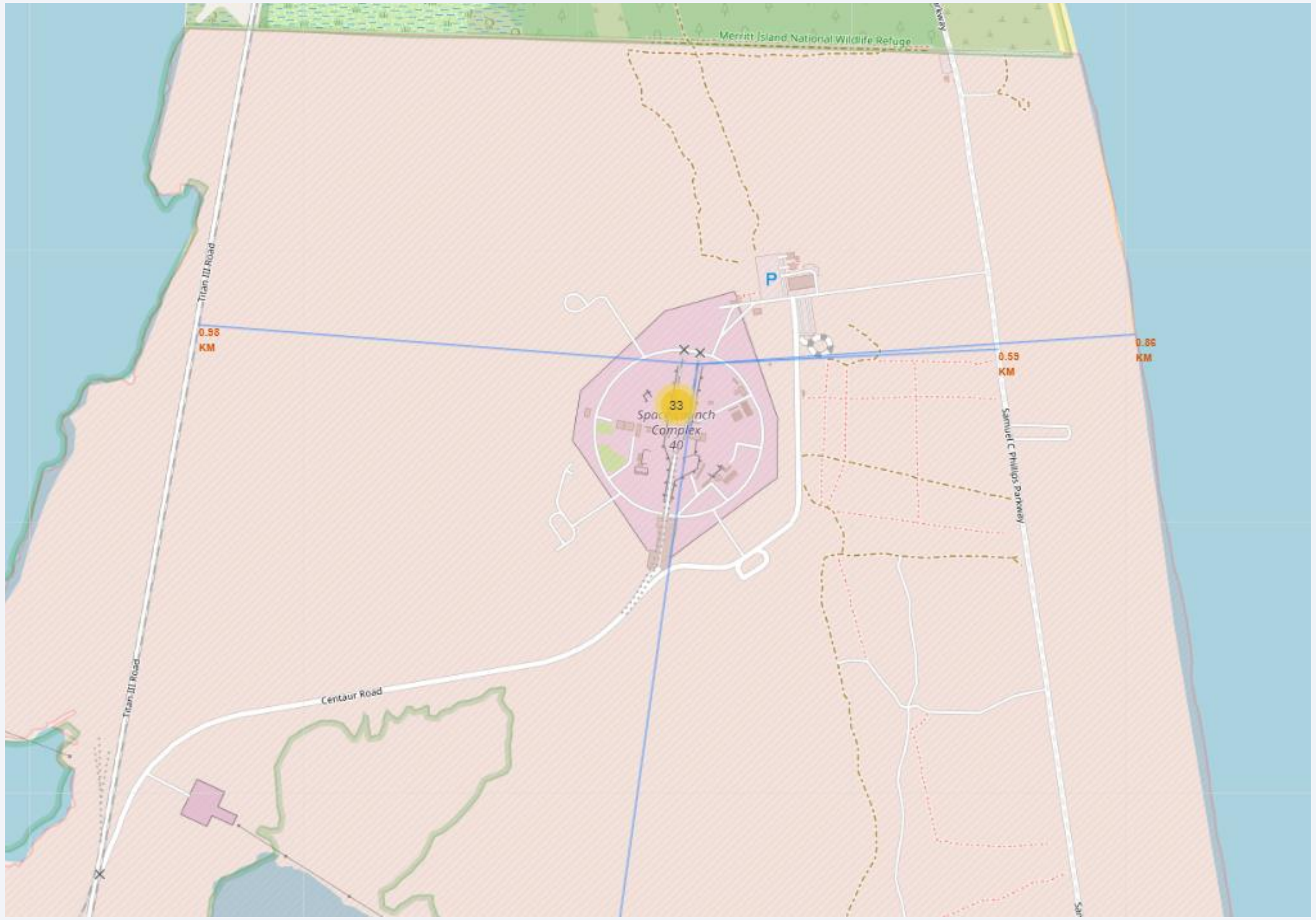
CCAFS LC-40 Mission Success

- KSC LC-39A has the highest success rate among all the launching sites



CCAF SLC-40 Landmark Proximity

- CCAF SLC-40 is located:
  - 0.98 km from the NASA Railway
  - 0.59km from the Samuel C Phillip Parkway
  - 0.86km from the coastline
  - 19km from Cape Canaveral

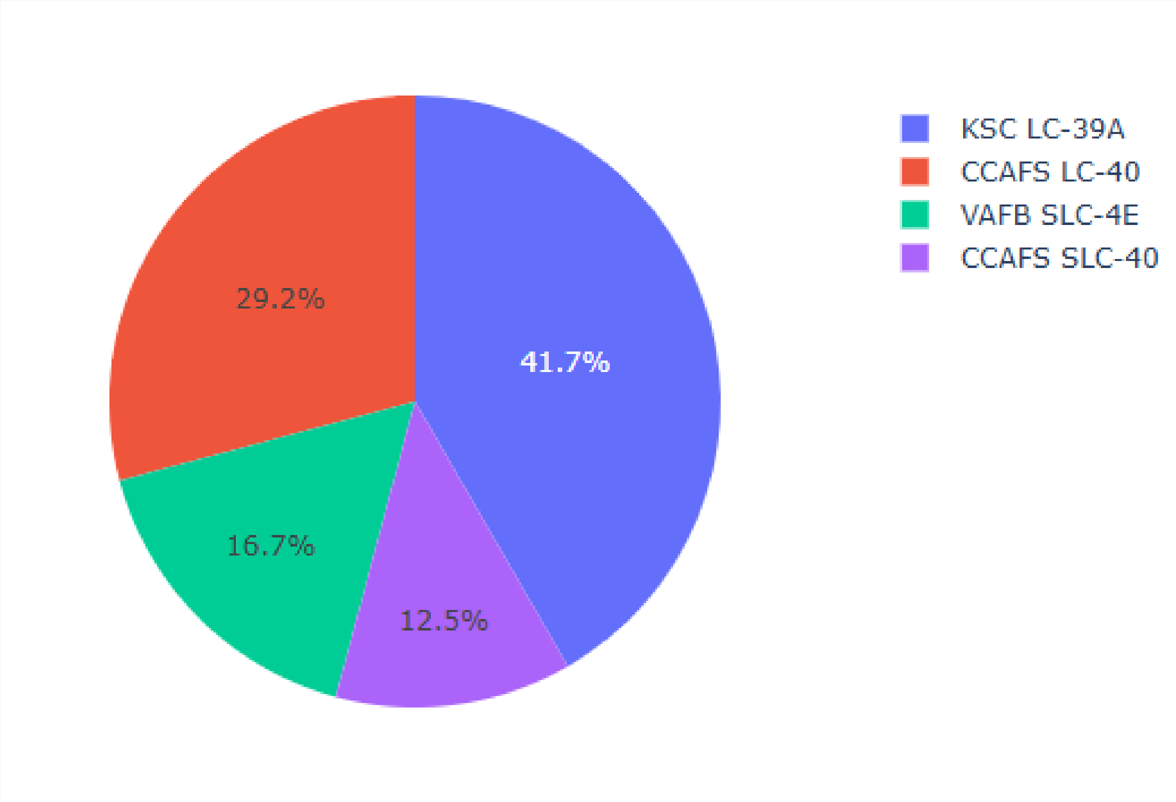


Section 4

Build a Dashboard with Plotly Dash

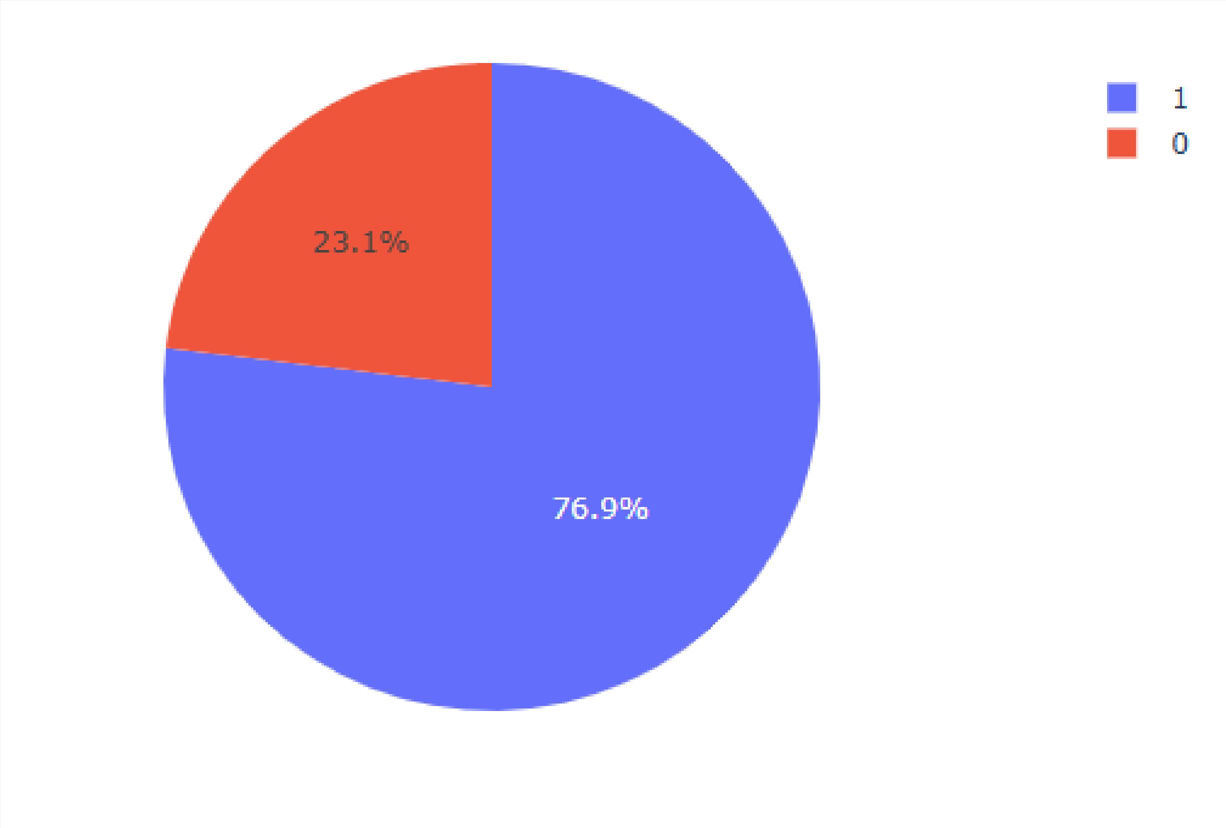
Success Contribution per Launching site

- KSC LC-39A has the highest success rate among all the launching sites and represents 42% of the successful missions

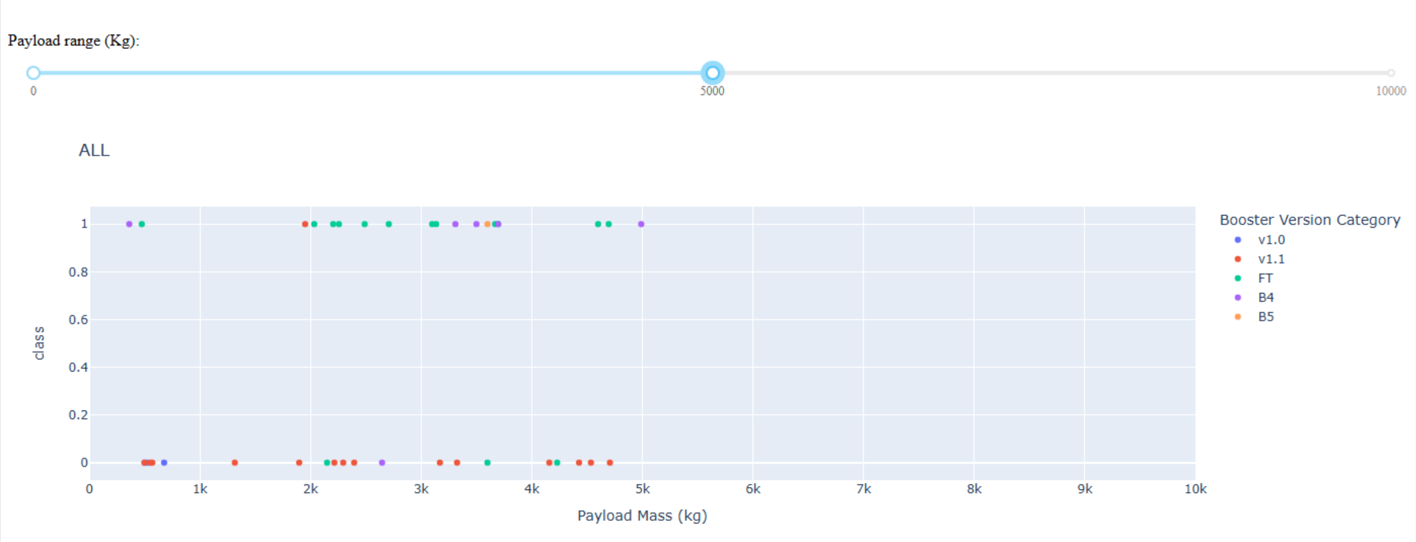


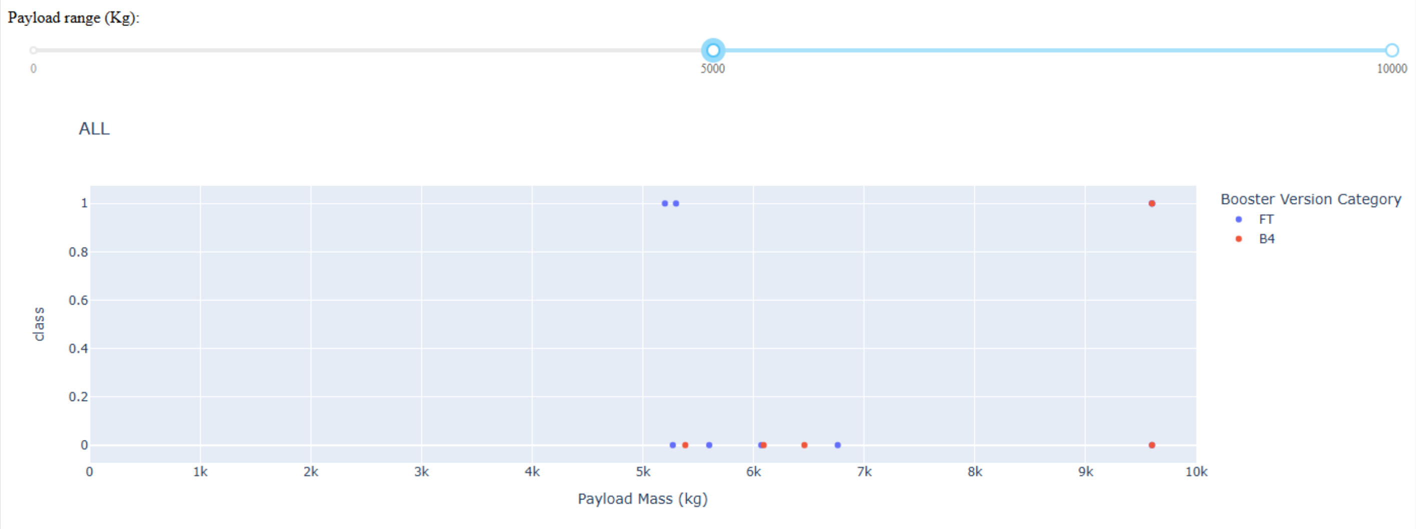
KSC LC-39A Success rate

- KSC LC-39A has the highest success rate among all launching sites



<Dashboard Screenshot 3>



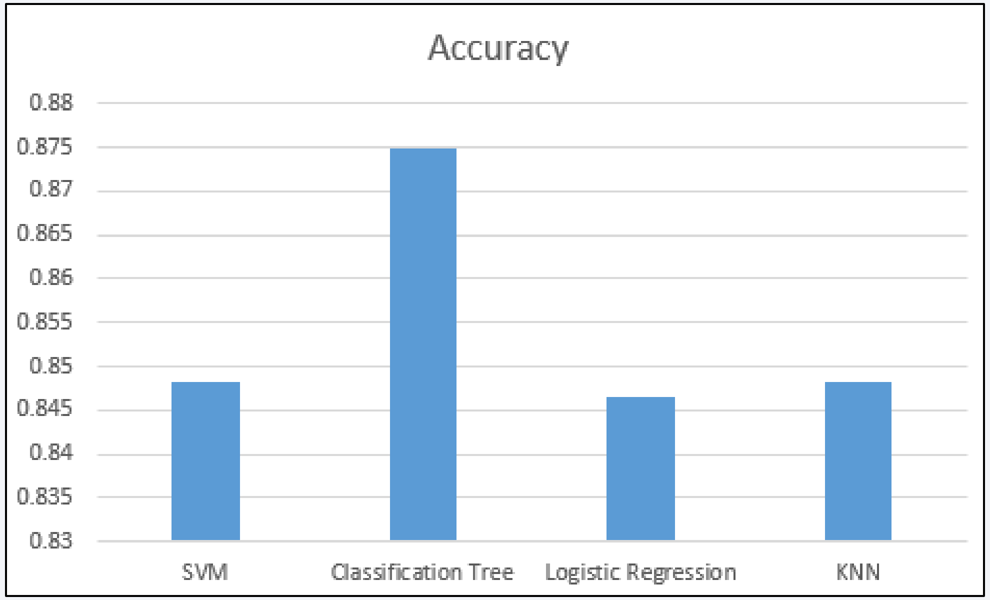


- FT and B4 Boosters where exclusively used for heavy payload
- Ther is no successful mission with a 5000 to 9000 kg mission

Section 5

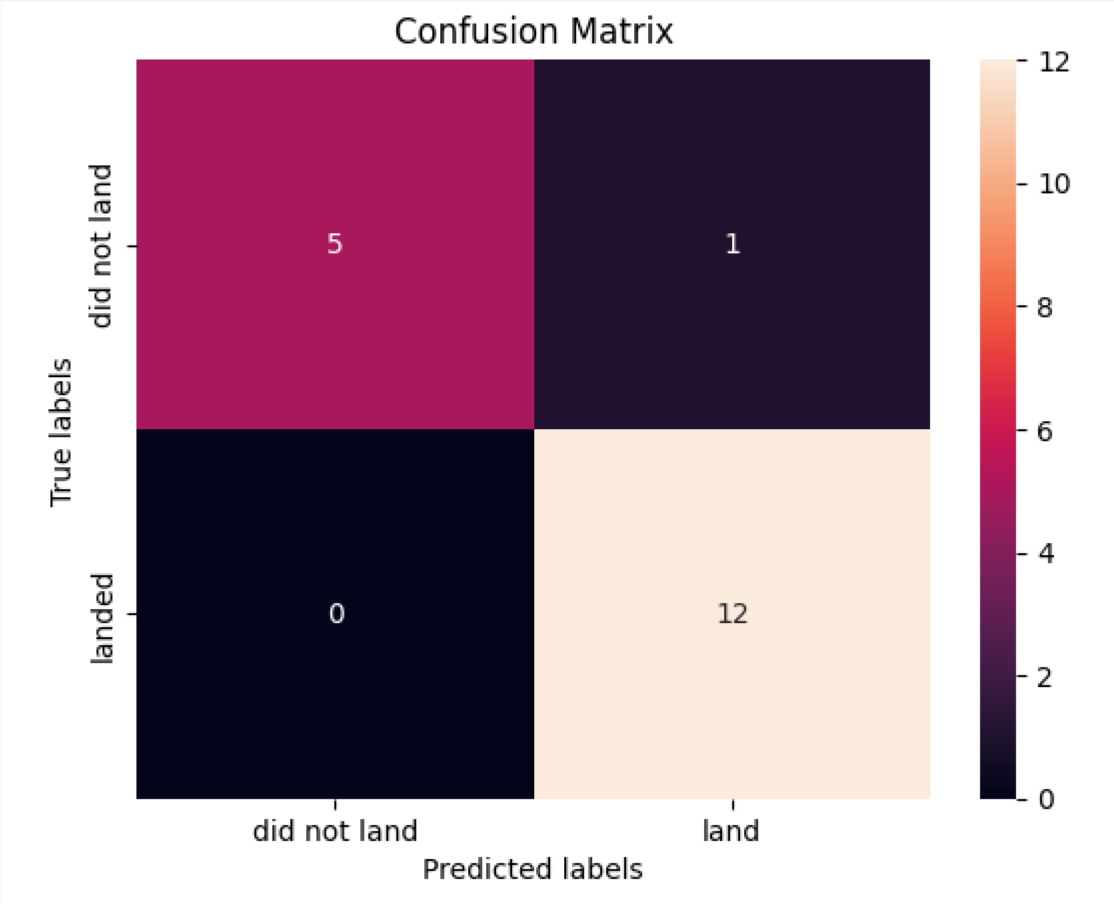
Predictive Analysis (Classification)

Classification Accuracy



- The Decision Tree Model gives the best accuracy (87.5%)
- The best hyper parameters for this model are `{'criterion': 'gini', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'best'}`

Confusion Matrix



Confusion matrix for the Decision Tree Model

Conclusions

- ES-L1, GEO, HEO, SSO orbits all have a very high mission success probability
- Early tests were mostly conducted for LEO, ISS, PO, GTO and ES-L1 which probably explains the lower success probability
- The launch site with the highest contribution to the mission success and the highest probability of mission success is the KSC LC-39A
- The success rate of SpaceX mission has increase between 2013 and 2020 up to 80% in 2020
- There is no successful mision for 5000 to 9000kg payloads, it could be an opportunity for a competitor to land on this segment
- To predict the outcome of a mission the best model is the Decision Tree, and its accuracy is 87.5%

Thank you!