IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Oleh
29.07.2024

Outline

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

Executive Summary

- Summary of methodologies
    - Data Collection: SpaceX API and Web Scraping
    - Data Wrangling
    - Exploratory Data Analysis with Visualizations and SQL
    - Interactive Map with Folium
    - Interactive Dashboard with Plotly|Dash
    - Predictive Analysis
- Summary of all results

Introduction

- Many companies provide space travels like Virgin Galactic, Rocket Lab, Blue Origin, and Space X.
- SpaceX is the most successful and it has the advantage that the first stage of the Falcon 9 rocket is reusable
- The cost of a launch is 62 million dollars, while other providers cost upwards of 165 million dollars each
- The lower cost is largely due to the reusability of the first stage
- Knowing the key factors which will guarantee a successful Landing of the first stage is crucial
- This in turn will impact on the overall average cost for each Launch

Section 1

Methodology

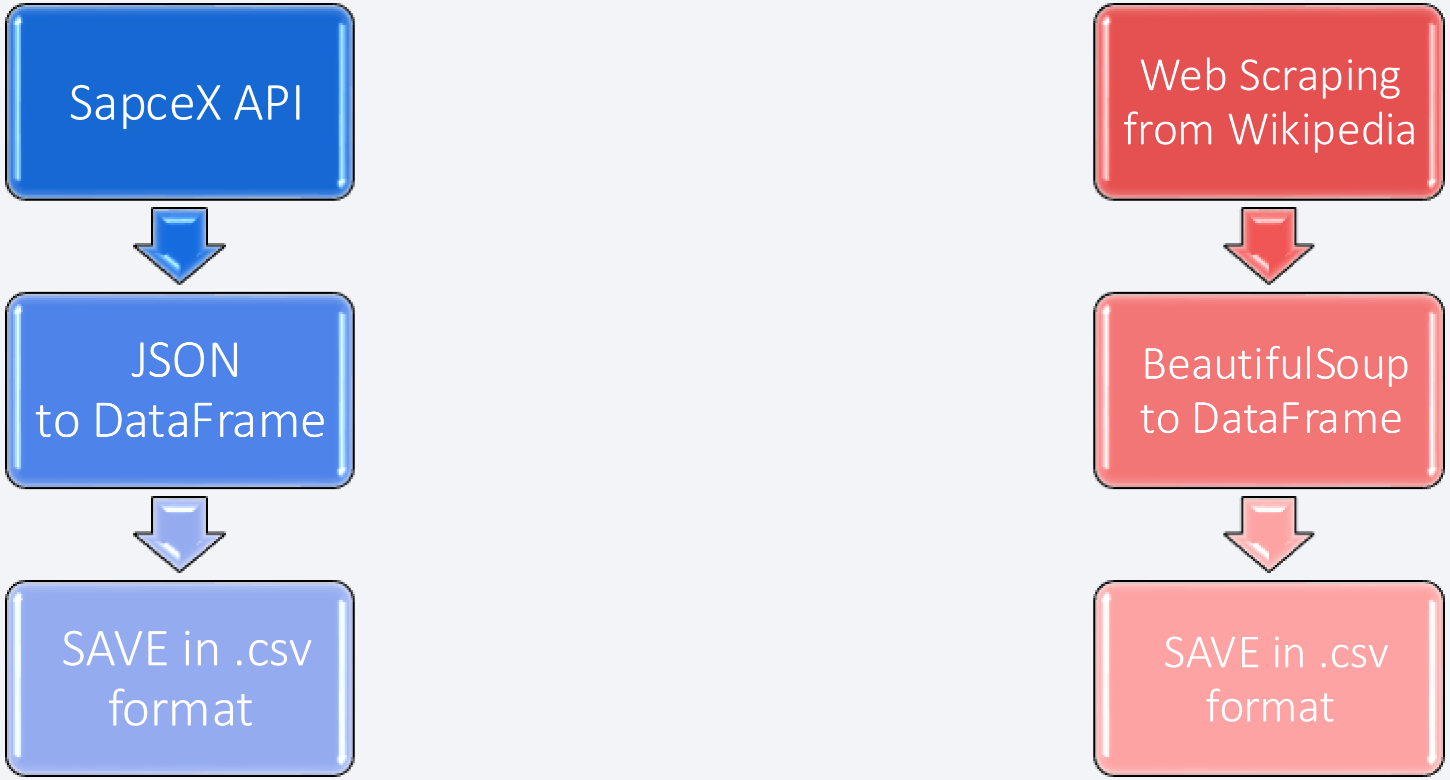
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and WebScaraping
- Perform data wrangling
  - Looking for null values, identifying categorical columns, on-hot encoding, ...
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Linear Multiple Regression, SVM, Classification Tree, and KNN models are trained and evaluated

Data Collection

- Data are collected with two methods as schematically shown below
- Further details are in the folowing slides



Data Collection – SpaceX API

- The SpaceX REST API is called several times with different parameters in order to get the right data in JSON format
- Data are converted into lists and then into a pandas Data Frame
- The parameters and the respective data are schematically shown in the right chart
- From the collected data only Falcon 9 rows are filtered and the resulting Data Frame is exported to a .csv file.

https://api.spacexdata.com/v4/

launches/past

- rocket
- payloads
- launchpad
- cores
- flight_number
- date_utc

core

- outcome of the landing
- type of the landing
- number of flights with a core
- etc.

rockets

- Booster name

launchpads

- Name of the launch site
- Longitude and Latitude

payloads

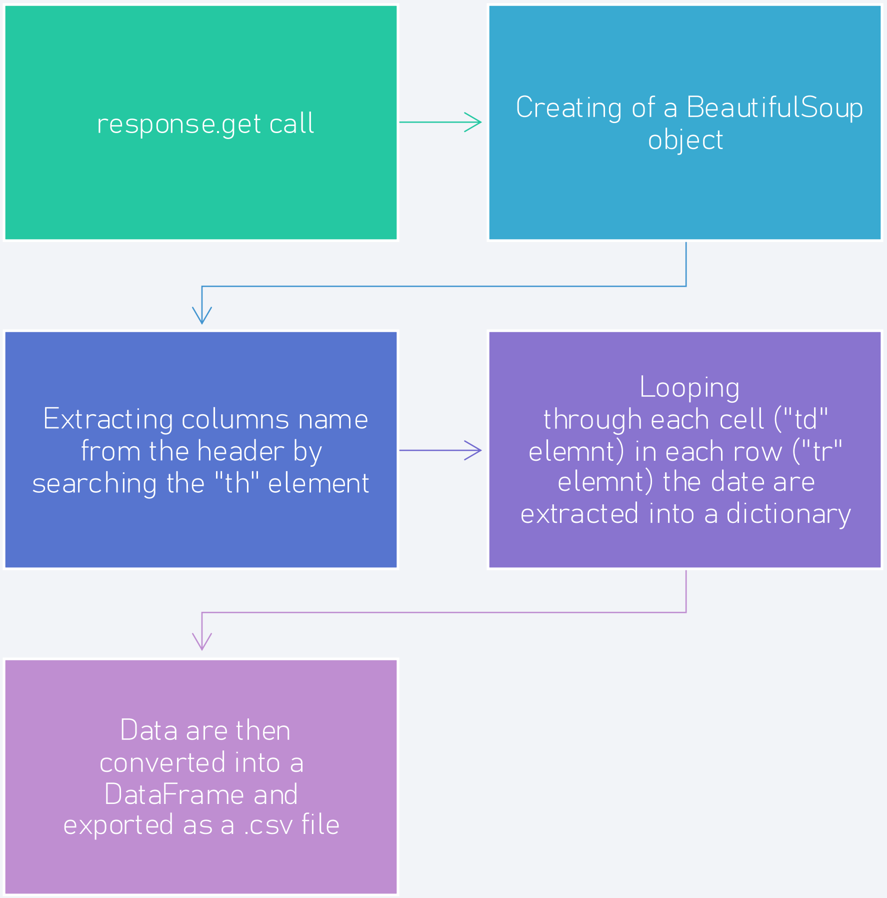
- Mass of the payload
- Orbit name

https://github.com/4ipsyki/SpaceY/blob/main/01%20-%20jupyter-labs-spacex-data-collection-

Data Collection - Scraping

https://github.com/4ipsyki/SpaceY/blob/main/02%20-%20jupyter-labs-webscraping.ipynb

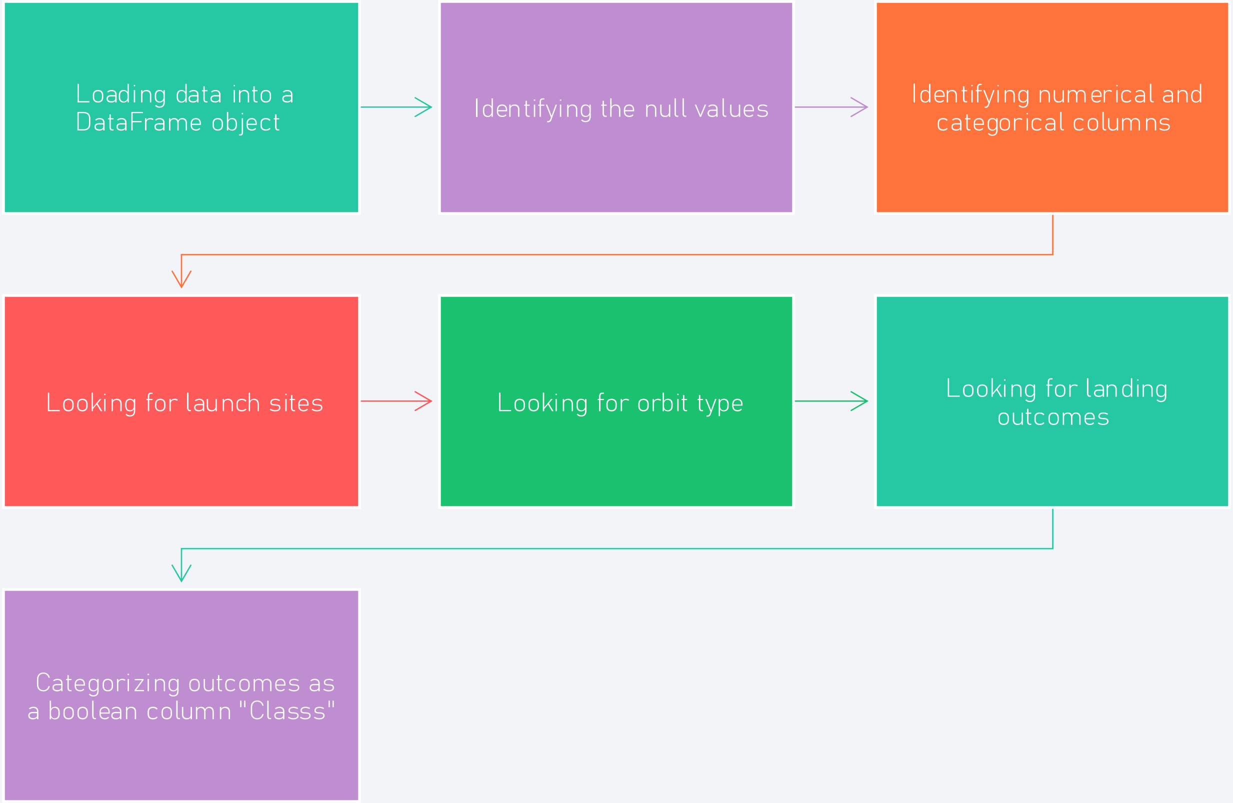
- Data are found in tables on Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- BeautifulSoup is used to extract the column names
- Data are extracted looping through the elements of each row of each table



https://github.com/4ipsyki/SpaceY/blob/main/02%20-%20jupyter-labs-webscraping.ipynb

Data Wrangling

- The mean landing outcome is Successful for 66.67 % of the cases



https://github.com/4ipsyki/SpaceY/blob/main/03%20-%20labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

The following plots were performed in order to identify preliminarily how the features influence the Class (Outcome):

- Scatter - FlightNumber vs. PayloadMass color-coded with Class
- Scatter - FlightNumber vs. LaunchSite color-coded with Class
- Scatter - PayloadMass vs. LaunchSite color-coded with Class
- Bar - Orbit Type vs. Success Rate
- Scatter - FlightNumber vs. Orbit Type color-coded with Class
- Scatter - PayloadMass vs. Orbit Type color-coded with Class
- Line - Year vs. Success Rate

As the last step, the categorical data were one-hot encoded into floats

https://github.com/4ipsyki/SpaceY/blob/main/05%20-%20jupyter-labs-eda-dataviz.ipynb

EDA with SQL

The data are loaded into a DataFrame object and converted into a DataBase.

The following sql queries were performed:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- 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
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/4ipsyki/SpaceY/blob/main/04%20-%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Each distinct Launch Site was marked on a Folium Map plot with a red Circle and the corresponding label
    - This is to understand the geographical locations of the launch sites
- For each record a Green (Successful Landing) and Red (Unsuccessful) Marker was added onto the map within a MarkerCluster object for a better visualization
    - This is to visually inspect how successful are each of the launch sites
- For a selected Launch Site, distances to the coastline, railroad, nearby city and highway were calculated and displayed on the map with a line and a label
    - This is to understand how close are the launch sites to public places and infrastructure

https://github.com/4ipsyki/SpaceY/blob/main/06%20-%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The following plots and features are created in the Dashboard:

- Success Rate vs. Launch Site discernable for Launch Site through a DropDown
- PayloadMass vs. Class color-coded for BoosterVersion and PayloadMass-windowed through a Slider

The second plot is also connected to the DropDown menu, thus allowing a better understanding of how the Class (Success Rate) is correlated with the considered features.

https://github.com/4ipsyki/SpaceY/blob/main/07%20-%20spacex_dash_app.py

Predictive Analysis (Classification)

The following Model are used:

- Linear Multiple Regression
- SVM
- Classification Tree
- K-Nearest Neighbour

Each model is trained on the Train set and evaluated on the Test set with metrices such as Accuracy and Confusion Matrix

Class converted into a Numpy Array Y → Features are Contained inside a DataFrame object X → X columns are standardized with SdandardScaler() from sklearn → X and Y are split into train/test sets as 80/20 → Mutliple models are trainde and tested → The results are summarized with the respective metrices

https://github.com/4ipsyki/SpaceY/blob/main/08%20-

Results

- CCAFS SLC 40 Launch Site is the most used, while KSC LC-39A is the most Successful for Landing
- Success Rate is increased for Launches with a Payload of more than 10000 kg
- SSO Orbit Type has 100% Success Rate
- After 2013 the Success Rate shows a steep improvement, with 2019 as the best year

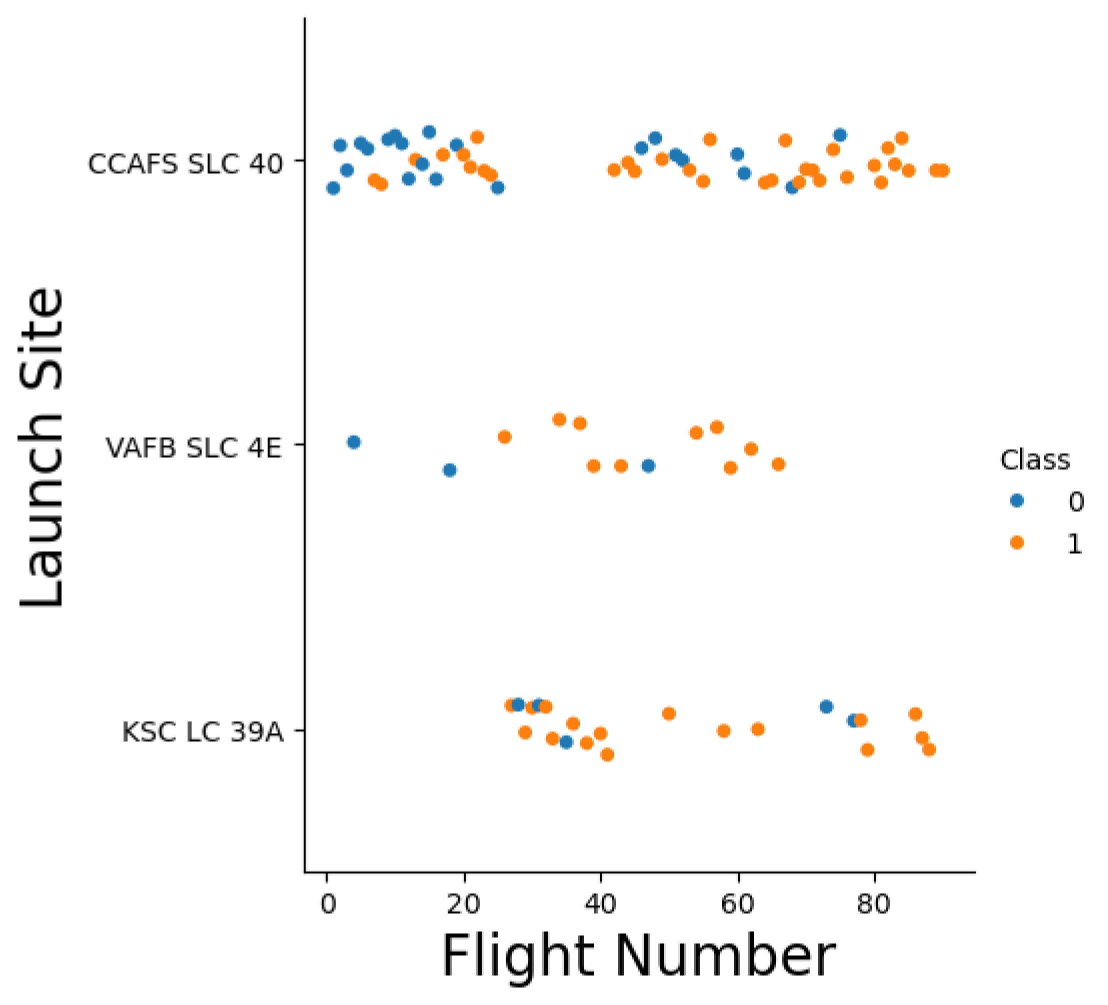
  with about 90 %
- Interactive analytics as Folium Map and Dashboard can be used to explore the data
- Predictive analysis shows that Classification Tree is the best model
- The main problem with predictive analytics is the inability to considerably improve the True Negatives
- True Positives are predicted correctly with all the four models

Section 2

Insights drawn from EDA

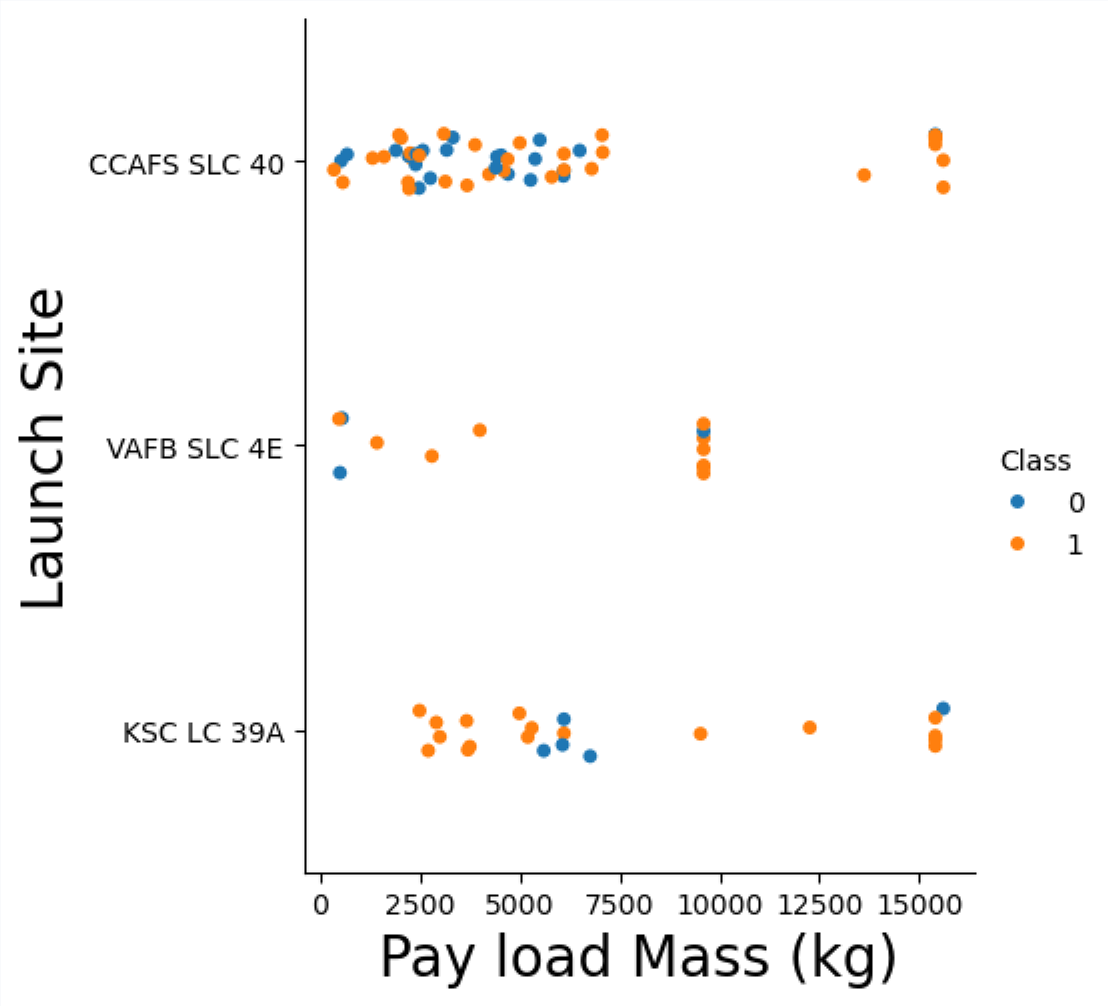
Flight Number vs. Launch Site

- There is a considerable preference for CCAFS SLC 40 Launch Site as compared to others
- This Launch site is especially predominant for about the first 20 launches where the majority of Unsuccessful landings have been obtained



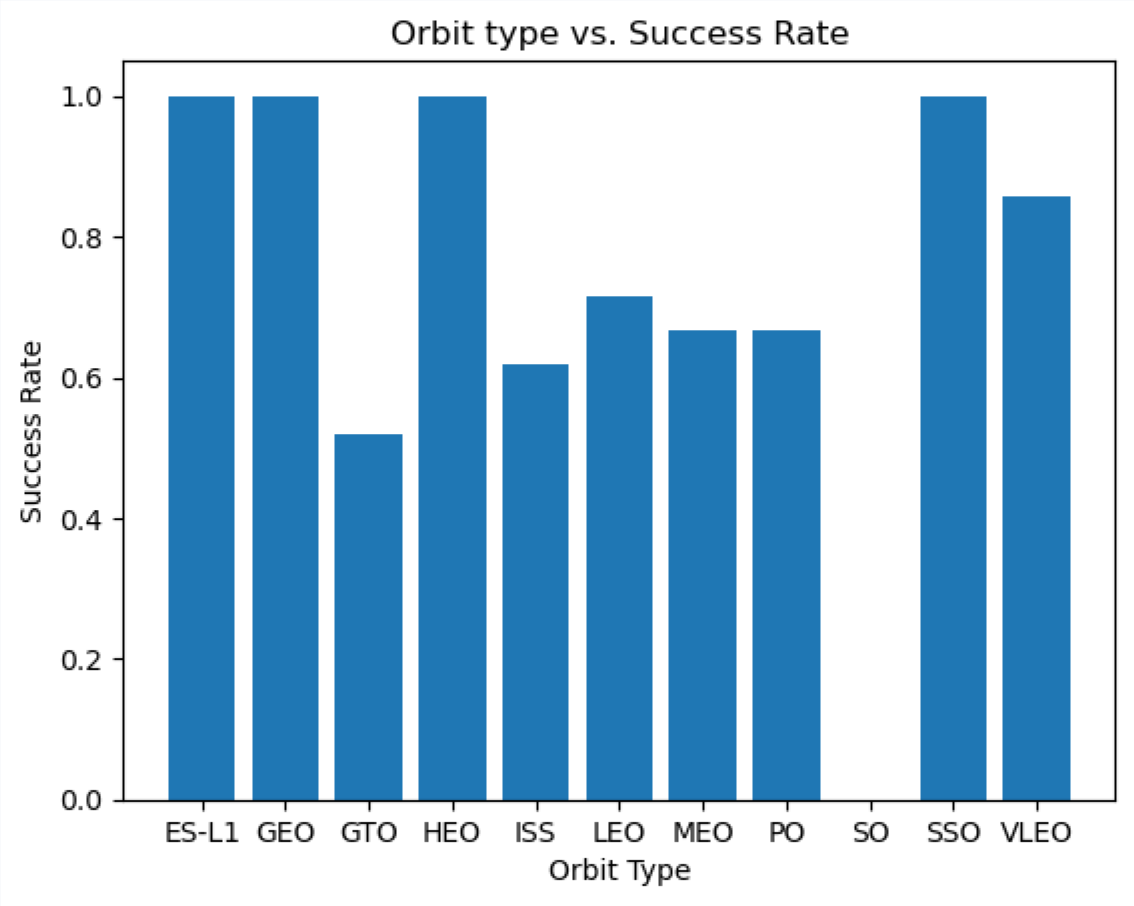
Payload vs. Launch Site

- The majority of Unsuccessful landings are obtained with a Payload below 7500 kg
- For Payloads of more than 10000 kg, the Success rate is greatly increased for the three Launch sites



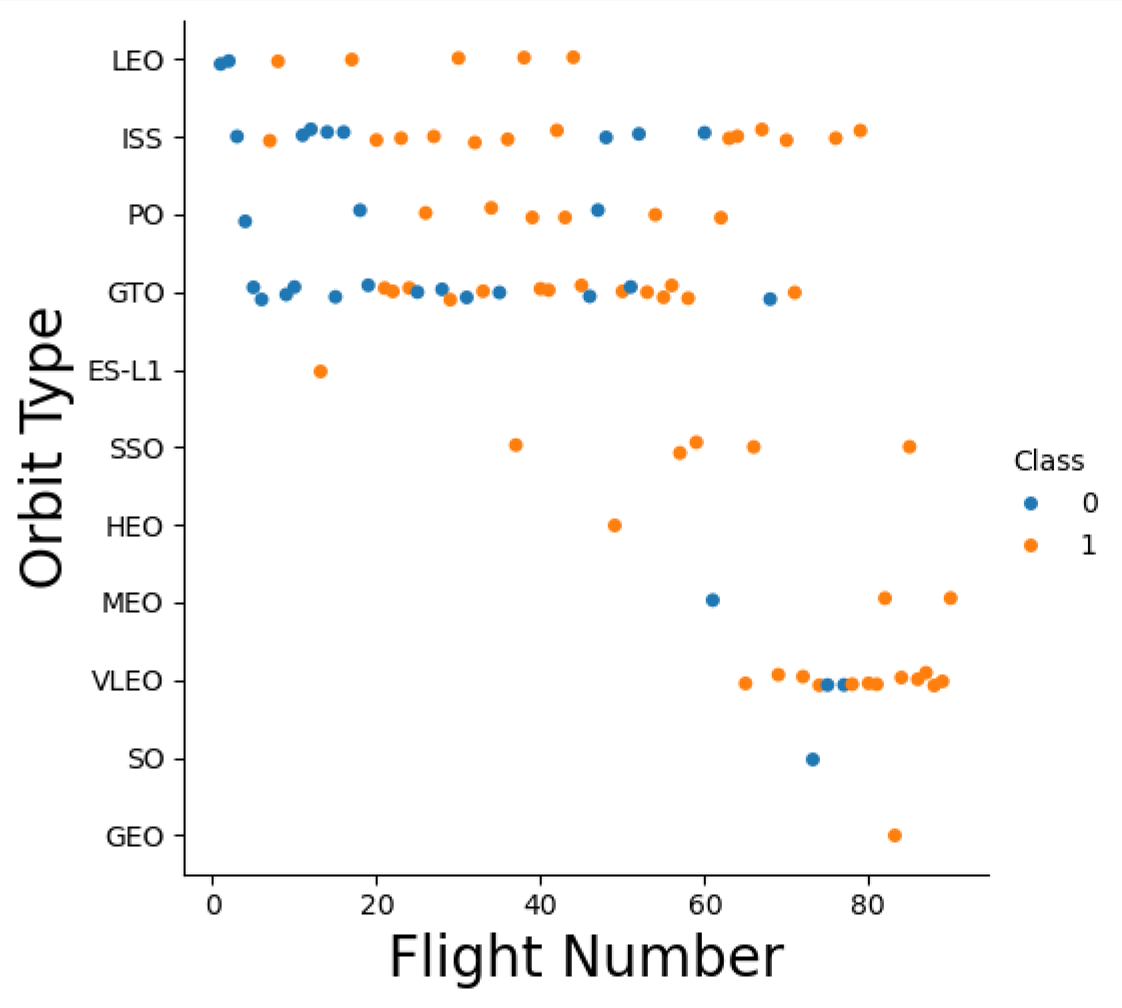
Success Rate vs. Orbit Type

- Four orbit types have a Succes rate of 100 %: ES-L1, GEO, HEO, and SSO
- Followed is VLEO with more than 80 %
- Five orbit types are between 50 and 70 %: ISS, LEO, MEO, PO, and GTO
- SO Orbit type is the most unsuccessful with 0 %



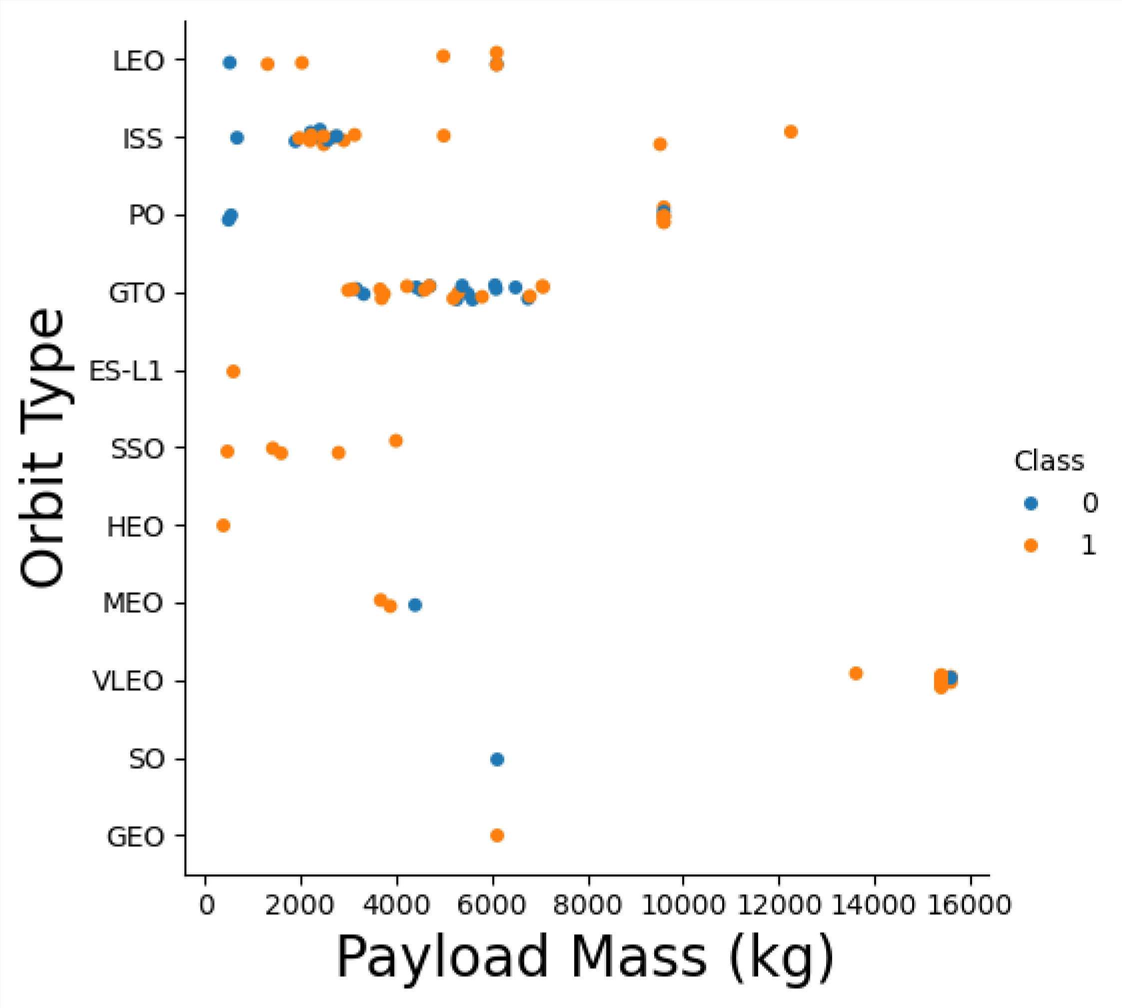
Flight Number vs. Orbit Type

- The SO, ES-L1, and GEO orbit types have been chosen only once, so the data are not representative
- Only ISS, PO, and GTO orbit types were chosen on a regular basis
- GTO orbit is the most frequent and thus has more unsuccessful landings
- SSO orbit type with five launches is the most successful



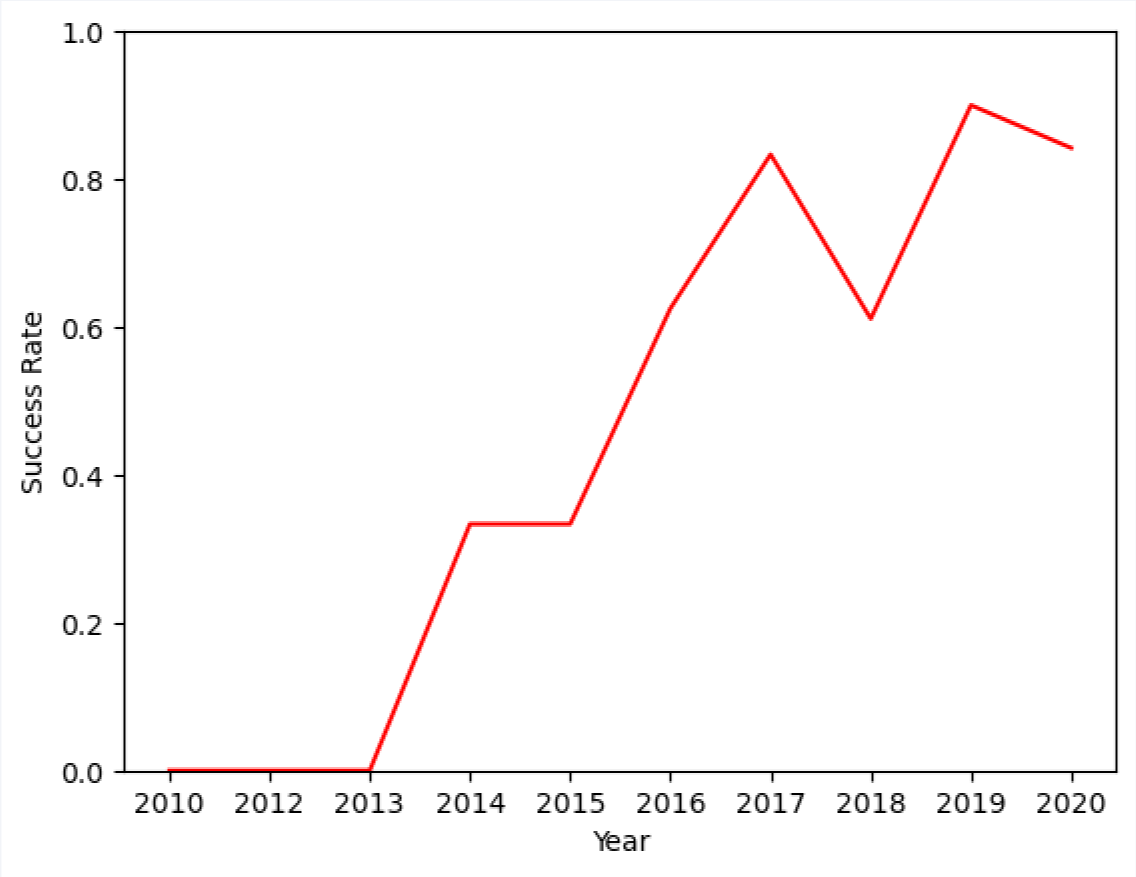
Payload vs. Orbit Type

- Contrary to slide 19 observations, the most successful type of orbit (SSO) was adopted with low PayloadMass
- The GTO orbit type is used for payloads between 3000 and 8000 kg and represents the most unsuccessful landings



Launch Success Yearly Trend

- Up to 2013 (for three years) the landings were unsuccessful
- In 2014 and 2015 the success was stable below 40%
- Up to 2017 the success was linearly growing
- The best year was in 2019 with about 90 %
- In 2018 there is a dip with a value similar to 2016



All Launch Site Names

- Launch Sites are extracted with an sql query by adopting the DISTINCT function
- It is found that there are only four Launch Sites

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

Launch Site Names Begin with 'CCA'

- The first five records where launch sites begin with `CCA` are shown below
- One can see that although the Landings were unsuccessful, the missions were successful

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

- The total payload carried by boosters from NASA is 45'596 kg
- The result is found by filtering the proper rows and using the SUM function for the calculation

Total_payload_mass_NASA(CRS)
45596

Average Payload Mass by F9 v1.1

- **The average payload mass carried by booster version F9 v1.1 is found to be 2534.67 kg**
- The result is found by filtering the proper rows and using the AVG function for the calculation

mean_payload_mass_carried_by_F9v1.1
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing on ground pad is on 22 Dec 2015

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- Here is the list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The booster versions are all of F9 FT type

Booster_Version	PAYLOAD_MASS__KG_	Mission_Outcome	Landing_Outcome
F9 FT B1022	4696	Success	Success (drone ship)
F9 FT B1026	4600	Success	Success (drone ship)
F9 FT B1021.2	5300	Success	Success (drone ship)
F9 FT B1031.2	5200	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Although there is a considerable amount of unsuccessful landings, we can clearly see that almost all of them were programmed
- Only one mission was Unsuccessful

Number_Success	Number_Failure
101	1

Boosters Carried Maximum Payload

- The names of the distinct boosters that have been used to carry the maximum payload mass are listed here
- One can see that all of them are of F9 B5 type

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

- The are two failed landing_outcomes in drone ship in year 2015
- One can see that they are three months apart and have the same BoosterVersion and LaunchSite

Month (in 2015)	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

- The most unsuccessful landing outcomes were the once where a landing was not programmed
- With 50/50 split of success is followed the landing on drone ship

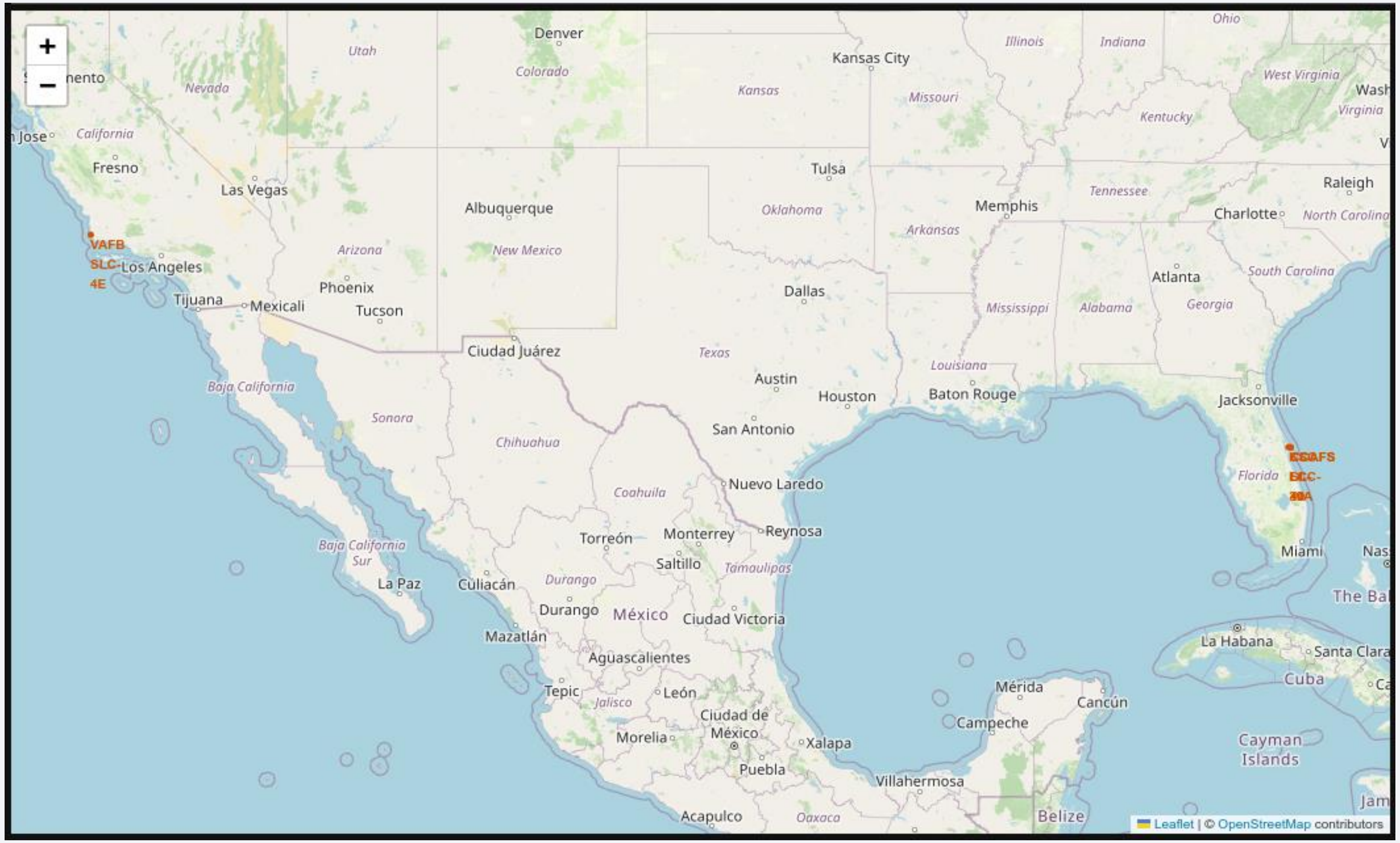
count(*)	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

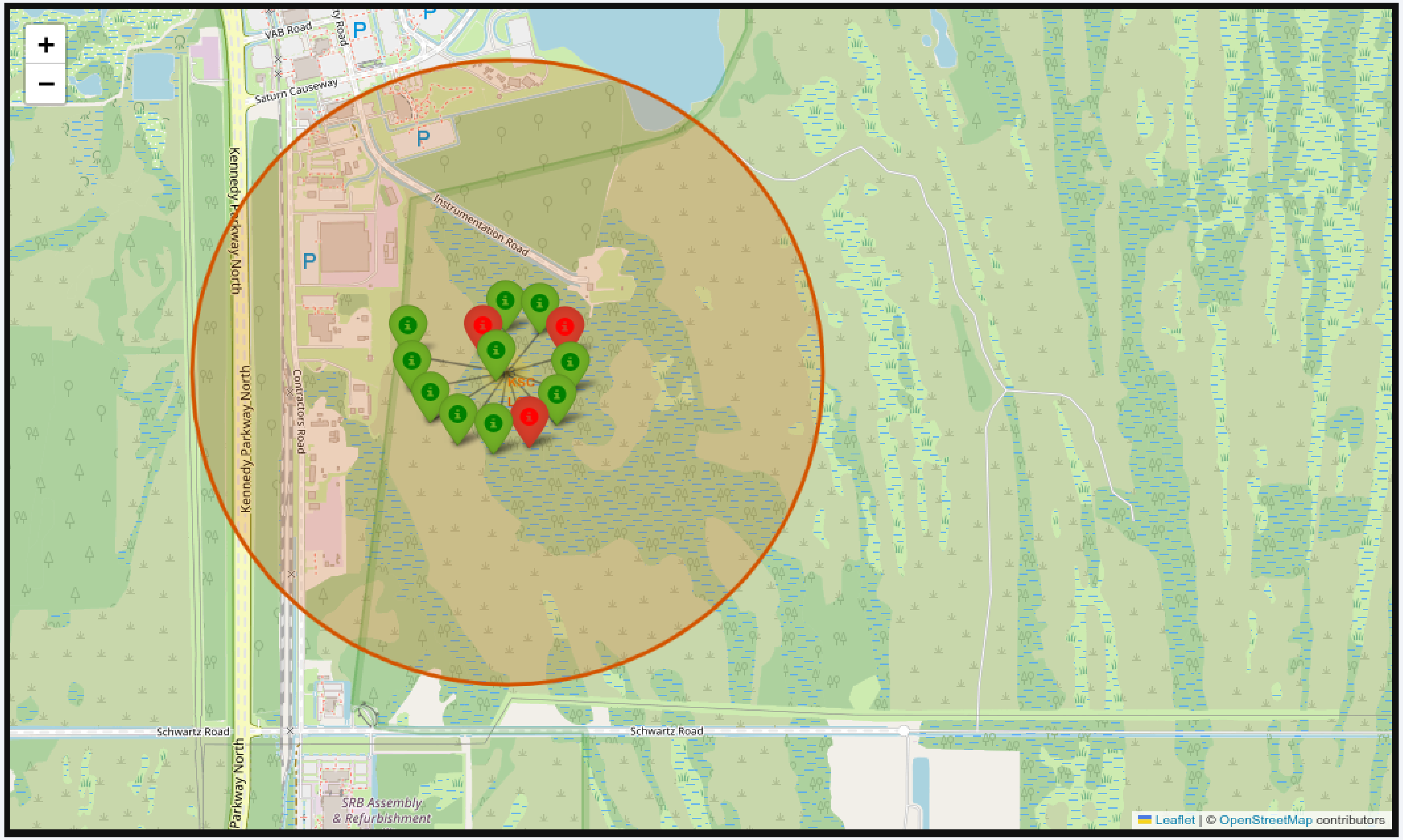
Launch Site Locations

- Launch sites are in the proximity of the coastline
- Launch sites are located in two regions, one for each coast in the USA



Landing Outcomes for each Launch Site

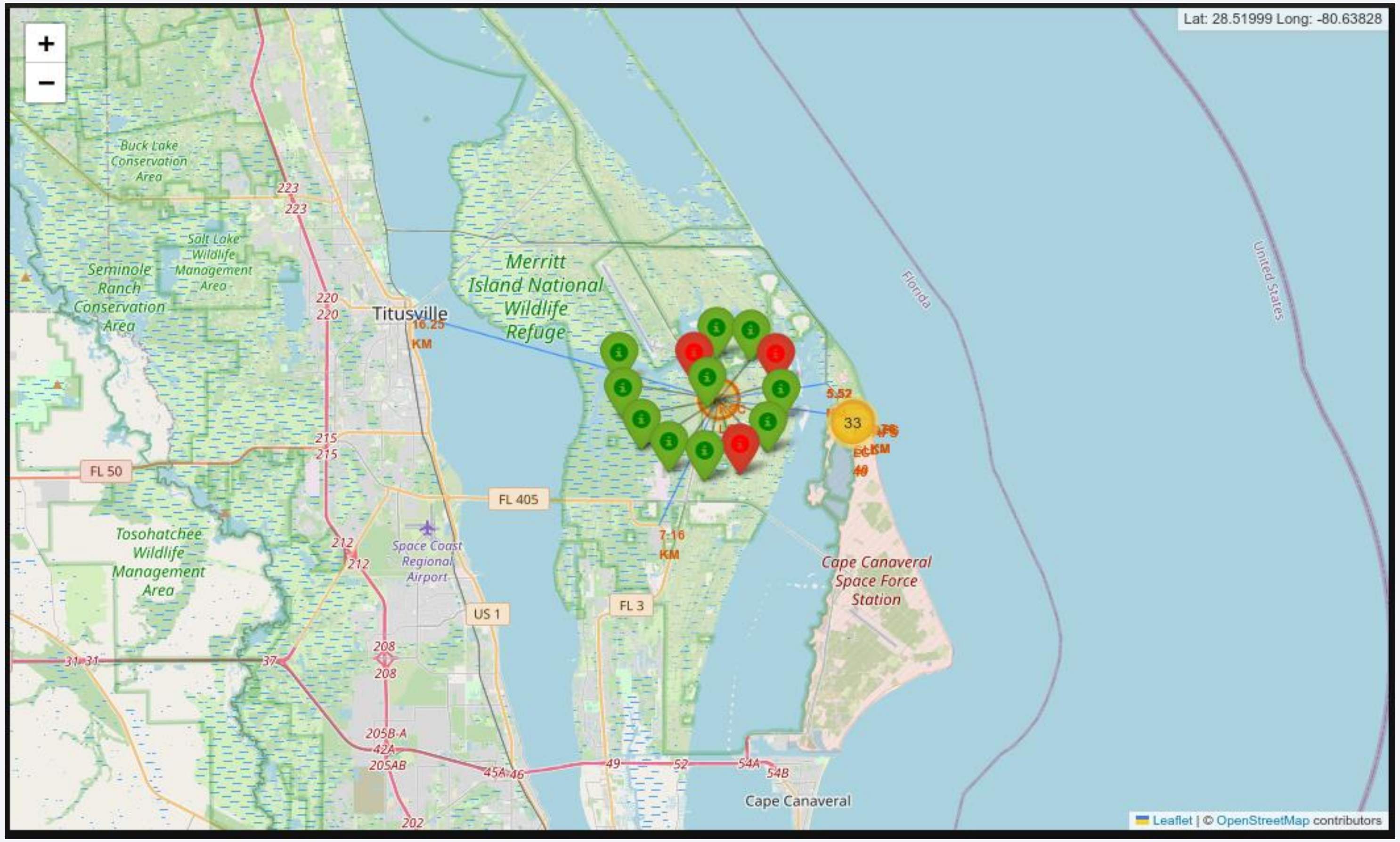
- For each Launch Site a MarkerCluster is added with Landing Outcome
- Green/Red marker represents successful/unsuccessful landing
- One can see that with increased amount of missions the landings are all successful



Proximities of KSC LC-39A Launch Site

- The distance of KSC LC-39A to the neighboring infrastructures and locations is shown on the map

- It can be seen that the nearby city is about 17 km away, while the closest freeway is only 8 km away

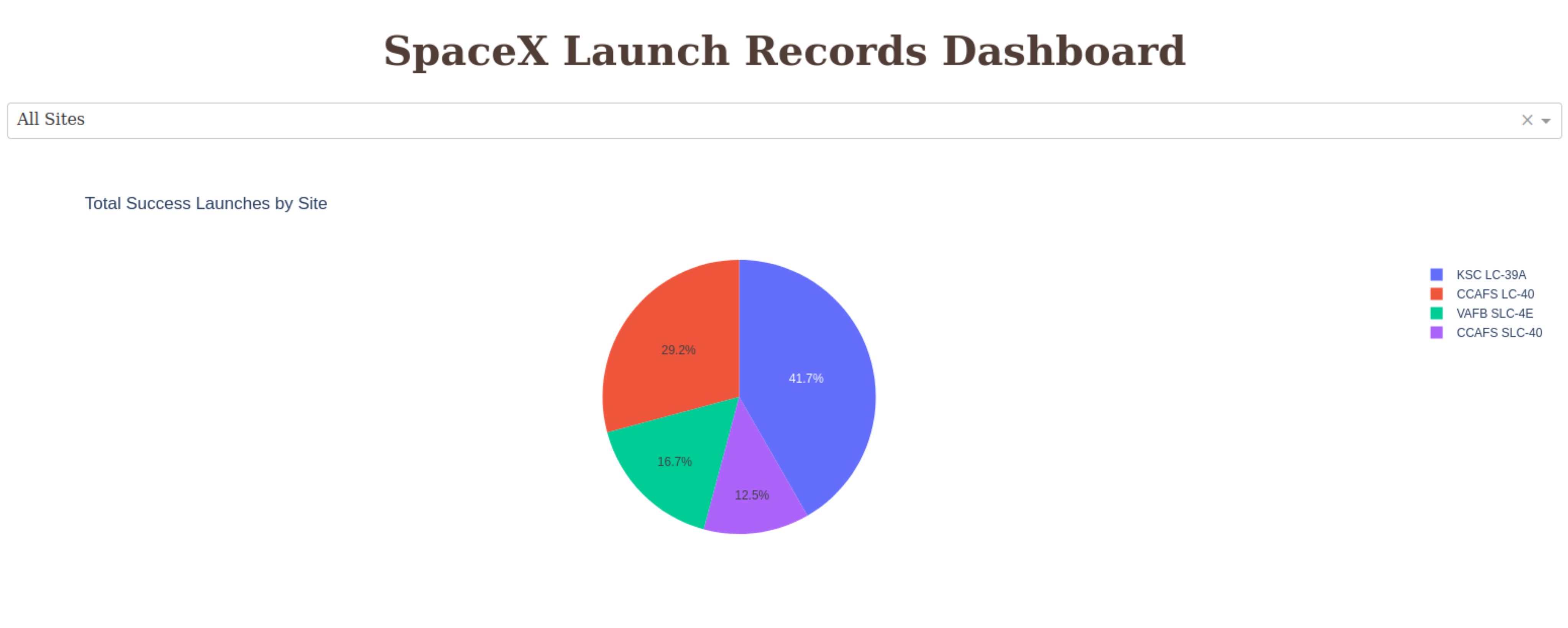


Section 4

Build a Dashboard with Plotly Dash

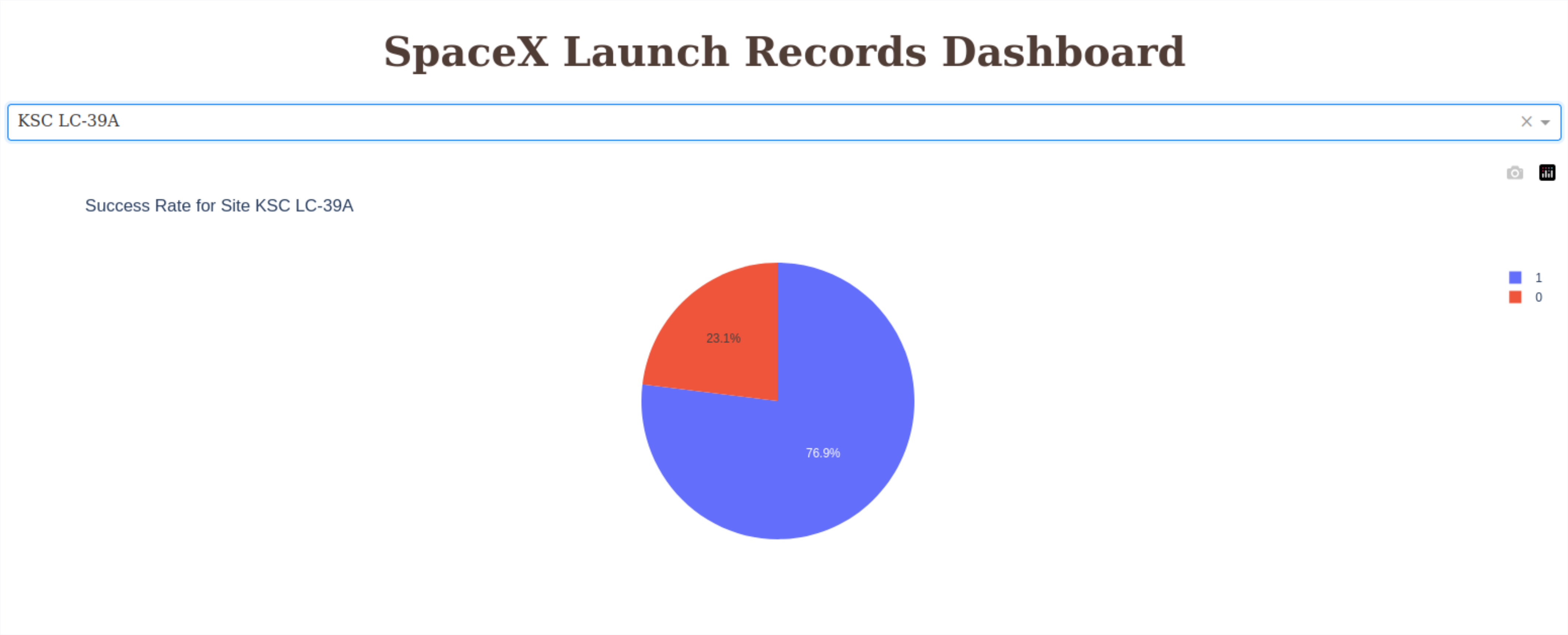
Dashboard: Success Rate for All Sites

- The most successful Launch Site, in terms of Landing outcome, is KSC LC-39A



Dashboard: KSC LC-39A Success Rate

- The success rate, in terms of Landing outcome, is about 77 %



Dashboard: Success Rate for 2000-7000 kg Payload

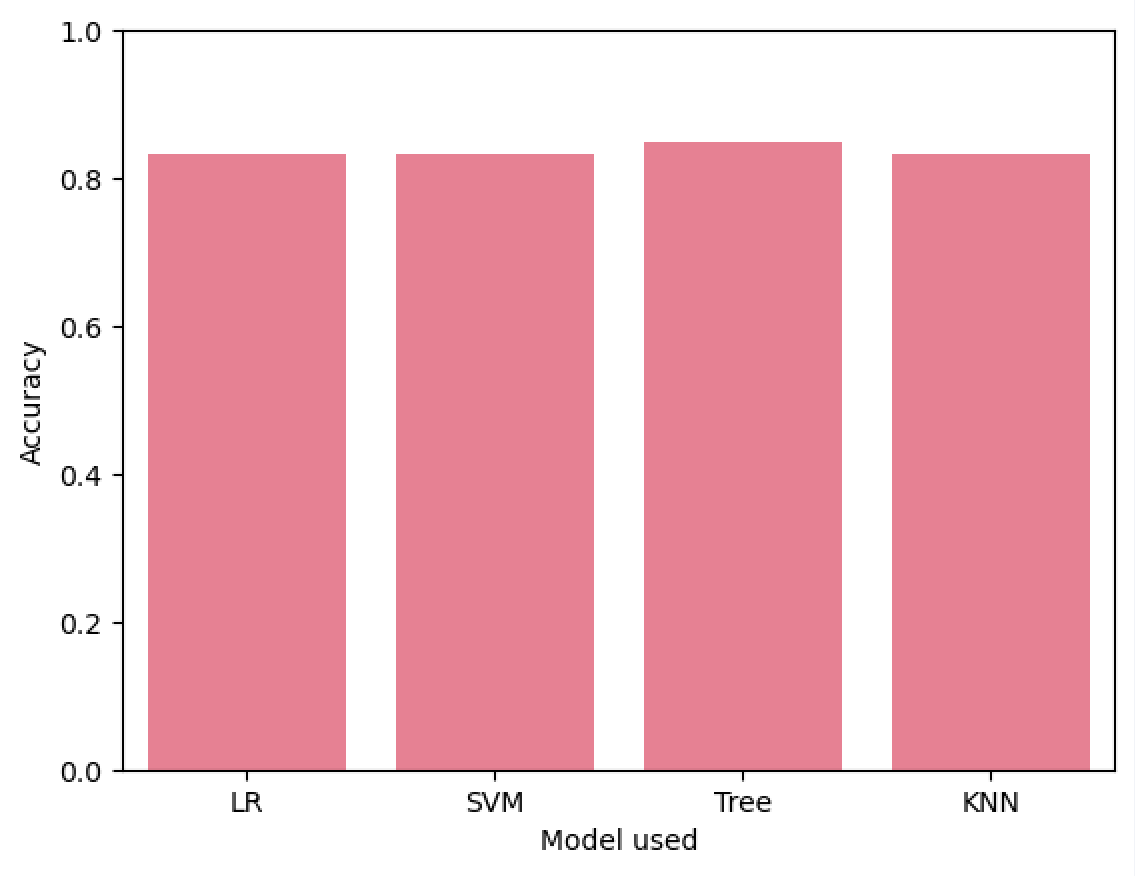
- The least successful landings are correlated with payload in the range 5500 - 7000 kg, where no successful ones are found
- The range of 2000 - 4000 kg is the most successful

Section 5

Predictive Analysis (Classification)

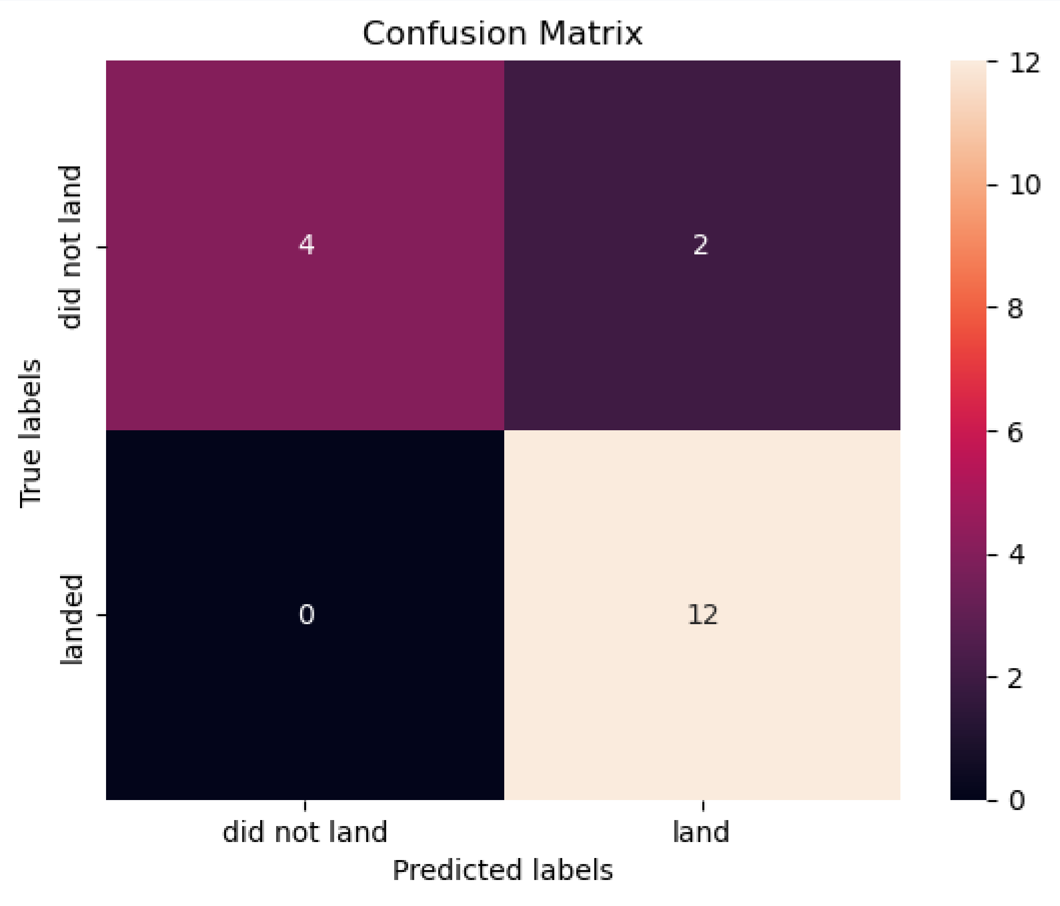
Classification Accuracy

- The best Accuracy is found to be for the Classification Tree model and about 85 %
- The other models performed identically to each other with 83 %



Confusion Matrix

- The Classification Tree performed better with True Negatives (4) with respect to the other models (3), thus reducing by one the False Positives (2)
- All the models performed identically with respect of True Positives, by predicting correctly all 12 cases



Conclusions

- In order to improve the Success rate the following Launch parameters should be adopted:
  - KSC LC-39A as the Launch Site
  - A Payload of more than 10000 kg
  - SSO Orbit Type
- During the first three Years of a hypothetically founded company will result in considerable costs due to Landing Failures
- Landing Failures serve as Learning material for perfecting the task
- A Success Rate of more than 90 % is achievable
- If the optimal Launch parameters are not available, a Classification Tree model can be used to predict the outcome of the Landing with 85 % Accuracy

Appendix 1

- The Notebooks, python code and Data are available at:

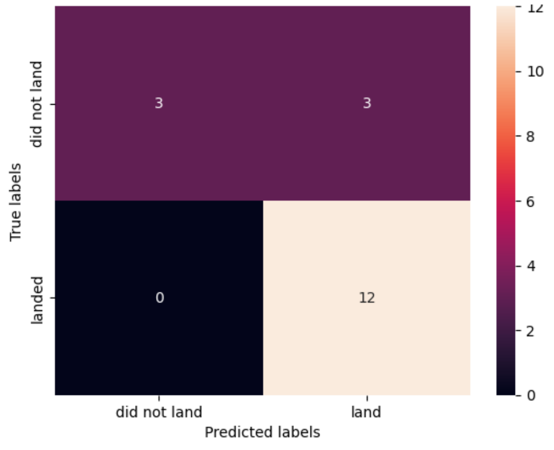
  https://github.com/4ipsyki/SpaceY

- Note 1: For a better understanding of a long-run costs for such a company, more recent data should be considered, together with the already available ones from 2017 and on

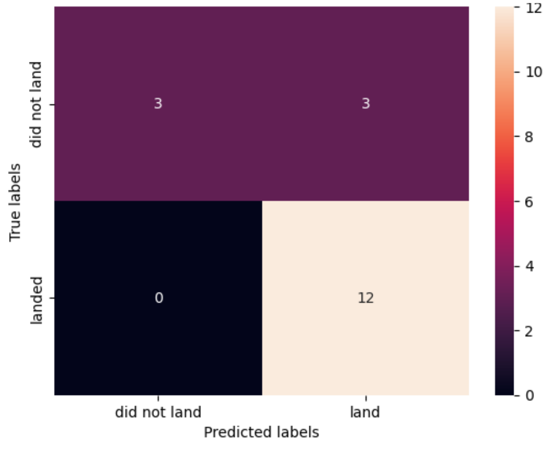
Appendix 2 - Predictive Analysis

- The Confusion Matrices for each model are shown on the right:
  - All the 12 True Positives are correctly predicted
  - False Positives are 3 for all models except Classification Tree

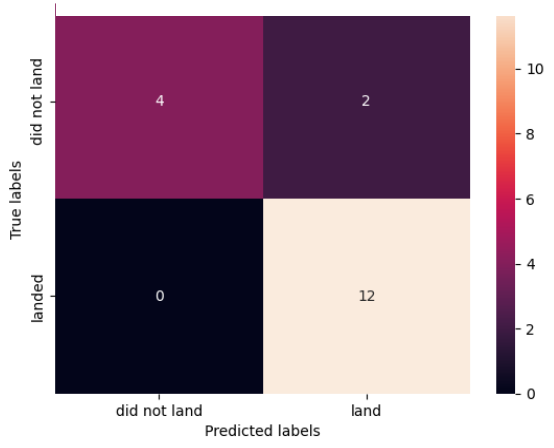
Multiple Linear Regression



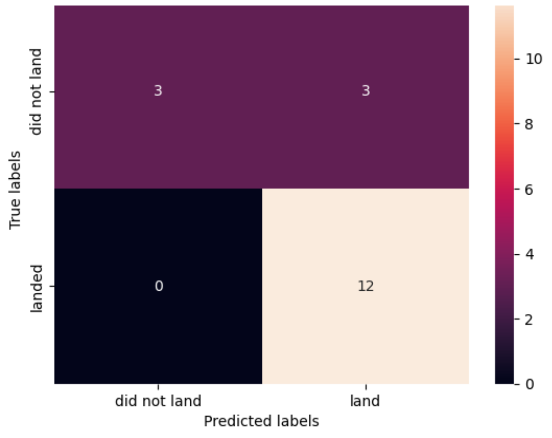
SVM



Classification Tree



KNN



Thank you!