IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stas Shavlev
16/10/2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data wrangling
- Exploratory data analysis (including interactive elements and dashboards)
- Predictive modeling (classification)

Summary of all results

- Exploratory data analysis results
- Predictive modeling results (best model)

Introduction

SpaceY intends to compete with the existing companies in the field of reusable first stage rocket modules. The company wants to use existing data collected from SpaceX Falcon 9 rocket launches to predict the first stage landing successes and other commercial implications.

Topics explored in the project

- Impact of measurable launch parameters such as payload mass, launch sites, orbits etc. on the succuss rate of booster landings
- Predictive modeling of the landing outcome based on measured parameters.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - SpaceX Rest API
    - Web scraping from Wikipedia
- Perform data wrangling
    - Data filtering
    - Missing values replacement
    - Feature transformations: One-hot encoding, standardization etc.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Train models, optimize hyper-parameters, evaluate performance using a test set and select best performing model

Data Collection

The complete data set used in this project is a combination of two distinct data sets collected from SpaceX Rest API and from a Wikipedia detailing SpaceX launches, using web scraping.

SpaceX Rest API

- Collect data from SpaceX API
- Parse the json file to dataframe
- Data Wrangling (cleaning, NaN filling and filtering)
- Export to csv

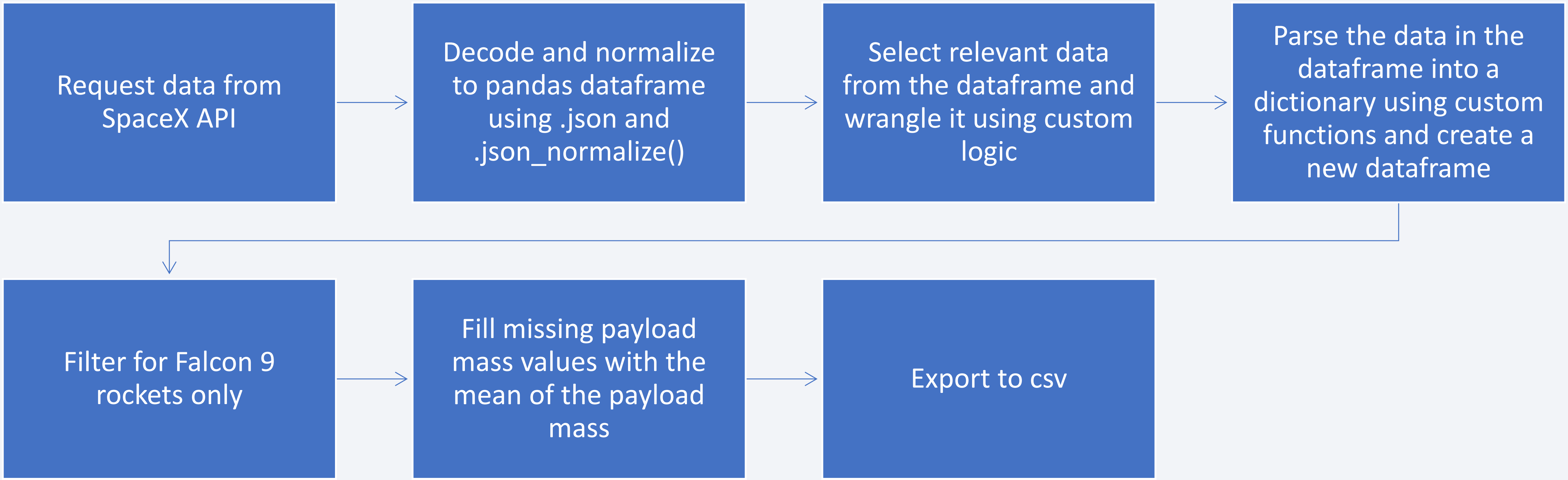
Web Scraping Wikipedia

- Scrap the Wikipedia page
- Using Beautiful Soup library extract relevant table elements and parse into a new dataframe
- Export the dataframe to csv

Combine Results

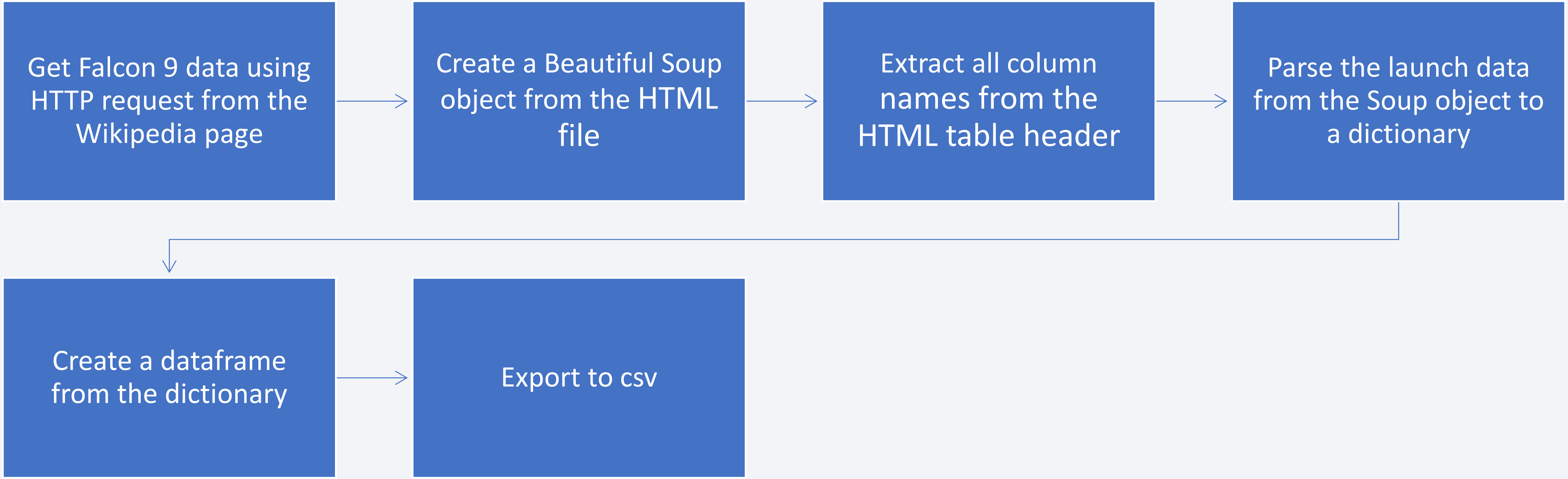
- Combine the two csv into a single working table

Data Collection – SpaceX API



The jupyter notebook can be found here: [SpaceX API Notebook]

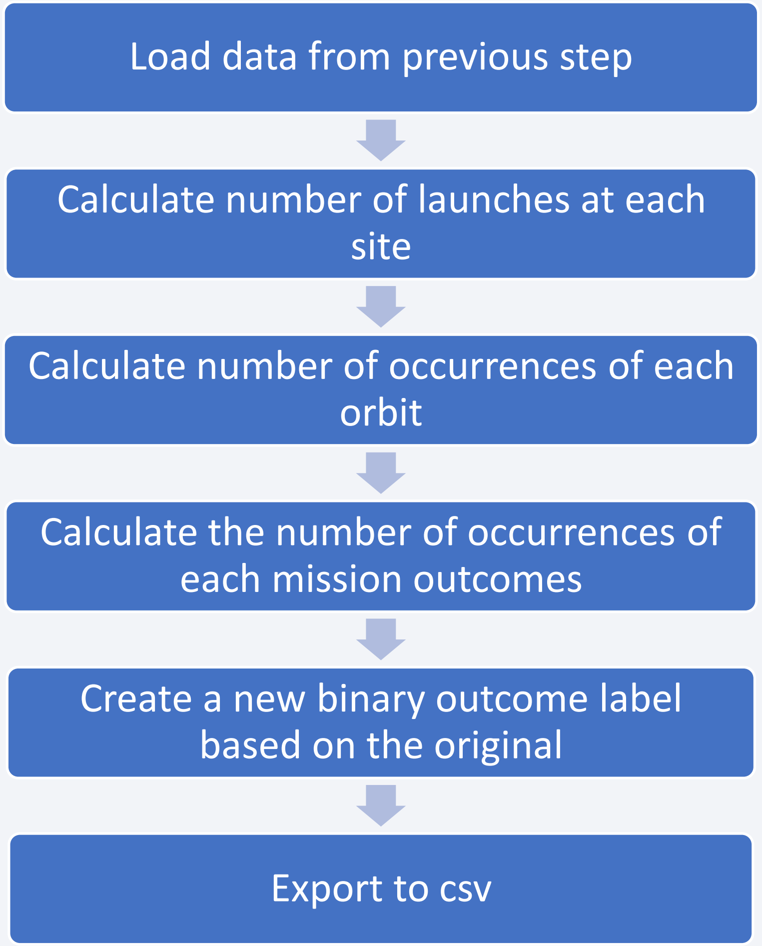
Data Collection – Web Scraping



The jupyter notebook can be found here: [Web Scraping Notebook]

Data Wrangling

- The most important part of this step is to convert to multi-labeled mission outcome into a binary label of 0 and 1
- The reason we are converting into a binary classification, is to allow us to model a classification model in the next stage
- The jupyter notebook can be found here: [Data Wrangling Notebook]



EDA with Data Visualization

- Charts that were used in the EDA, grouped by type of chart
  - Catplots/Scatterplots: Flight Number vs Launch Site, Payload Mass vs Launch Site, Flight Number vs Orbit Type, Payload Mass vs Orbit Type
  - Barplots: Success Rate by Orbit Type,
  - Lineplots: Success Rate by Year (Yearly trend)
- Catplots/Scatter plots show the relationship between variables
- Bar plots compare parameters by distinct groups of data, highlighting differences between groups
- Line Charts usually used for time trend presentation
- The jupyter notebook can be found here: [Data Visualizations]

EDA with SQL

- Performed SQL queries to obtain the following data
    - Unique launch sites
    - Top 5 records for launch site name starting with 'CCA'
    - Total payload mass of mission launched by 'NASA (CRS)'
    - Average payload mass of F9 v1.1 booster only
    - The date of the first successful mission with a ground pad landing
    - Names of boosters with successful drone ship landings with payloads in the range of 4k to 6k
    - Total number of failed and successful missions
    - Names of boosters that carried the maximum payload mass
    - Name, booster version, launch site and month for failed drone ship landings in 2015
    - Rank landing outcomes between 2010-06-04 and 2017-03-20 in descending order
- The jupyter notebook can be found here: [EDA with SQL]()

Build an Interactive Map with Folium

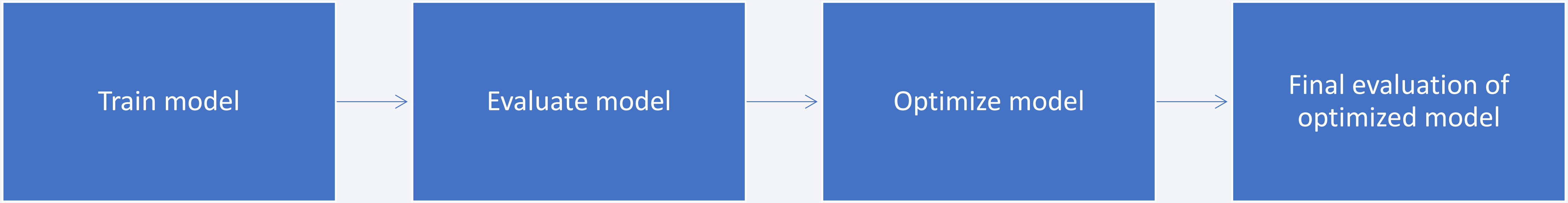
- Marked all site locations on a map generated by Folium, using their respective latitude and longitude coordinates to show their relative proximity to the equator
- Add marker cluster to each site to show all successful (green) and failed (red) landings, to visualize the success rate of each site
- The analysis then focused on the site VAFB SLC-4E (California), adding the distances to its closest landmarks such as coastline, highway, railway and city
- From the analysis we can see that the site is fairly close to major landmarks, so the stakes of failed mission is very high
- The jupyter notebook can be found here: [Interactive Map with Folium]

Build a Dashboard with Plotly Dash

- The dashboard was created using Dash Web and included the following features:
    - Launch Site drop down – allows the user to select all sites or a specific site for analysis
    - Pie Chart of successful launches – If all sites are selected it will show the total number of successful launches, and the breakdown of success/failures if a single site is selected
    - Payload mass slider – allows the user to limit the payload range and dynamically change the other charts
    - Scatter chart of Payload Mass vs. Success Rate of different booster versions
- The jupyter notebook can be found here: [Interactive Dashboard Code]

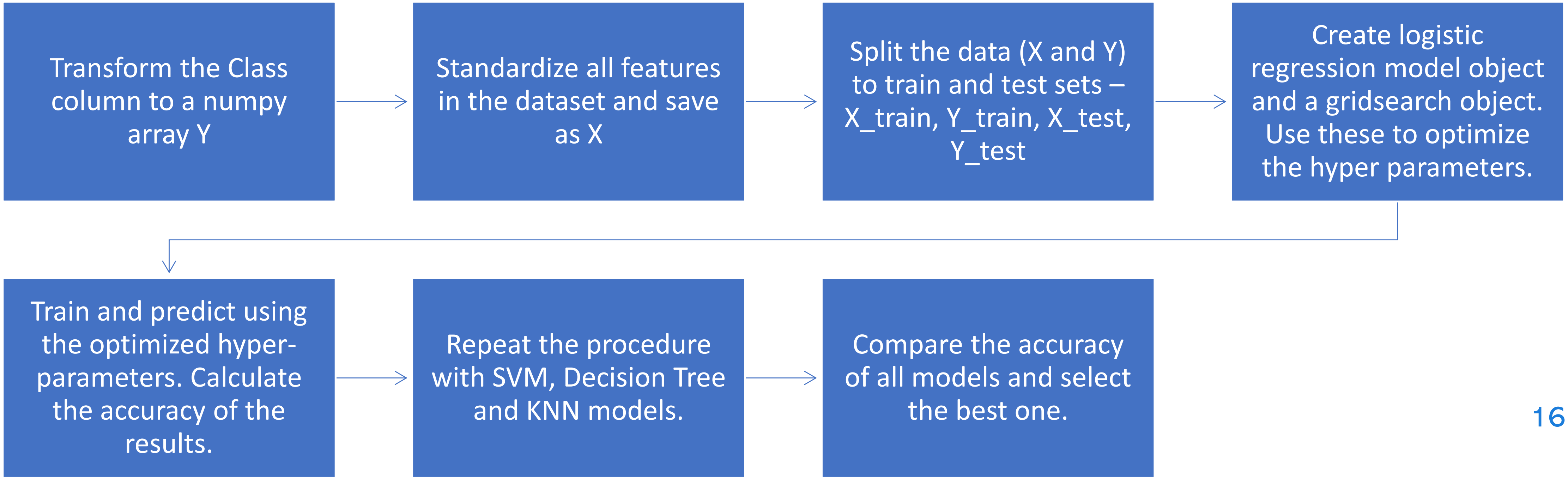
Predictive Analysis (Classification) - General

- We will follow the same general flow for all different models (see below)
- After completing this process for all models, we will select the best performing model based on the final evaluation metric



Predictive Analysis (Classification)

- The full modeling methodology steps are shown below
- Note – in order to keep the results reproducible, we set the random state of each model (where applied) to 42
- The jupyter notebook can be found here: [Classification Notebook]



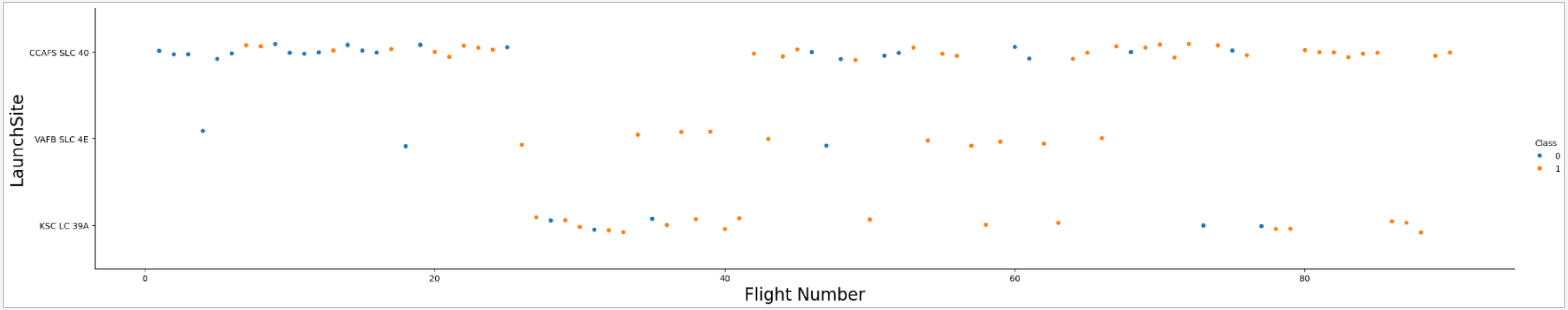
Results

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

Section 2

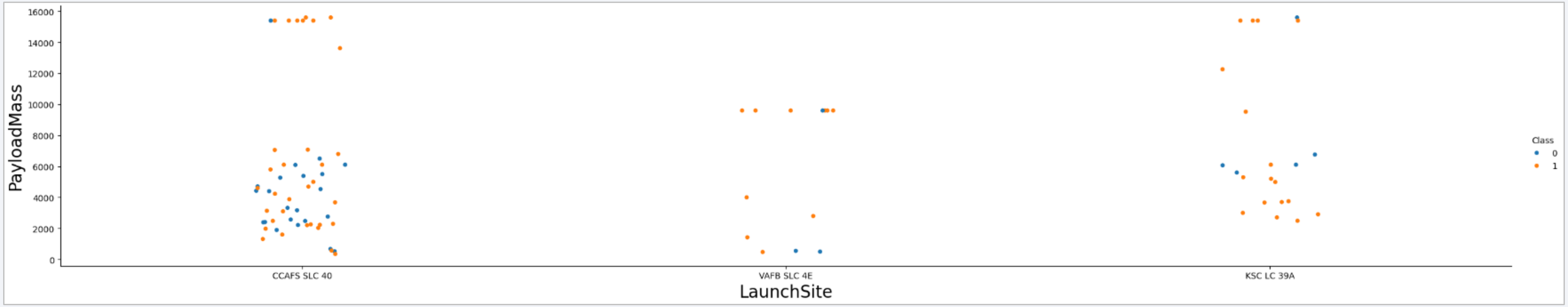
Insights drawn from EDA

Flight Number vs. Launch Site



- Increased success rate over time, leading to the assumption that new flights have higher chance of success
- CCAPS SLC 40 has more launches than the other sites, but with lower success rate, even at later dates

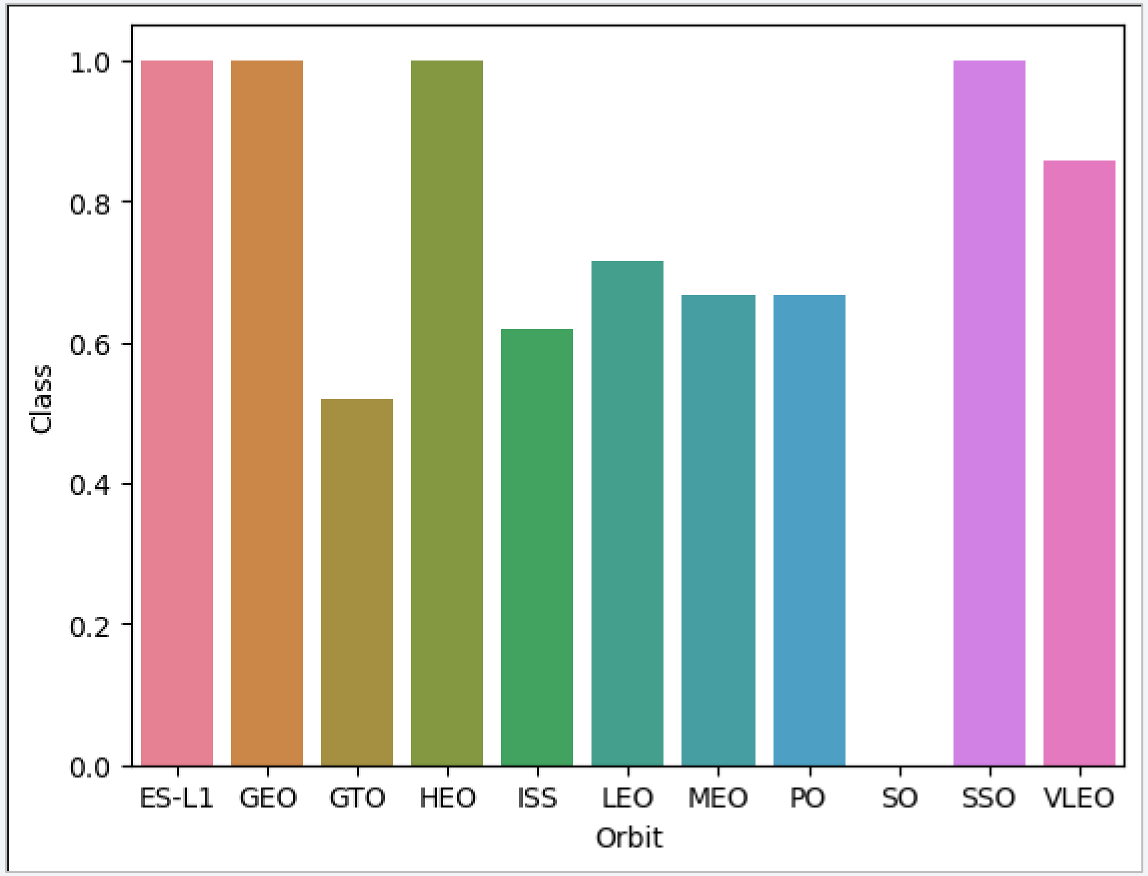
Payload vs. Launch Site



- Clear cut off at 7000kg with much higher success rates for heavier loads
- KSC LC 29A has 100% success rate for below 5500 payloads

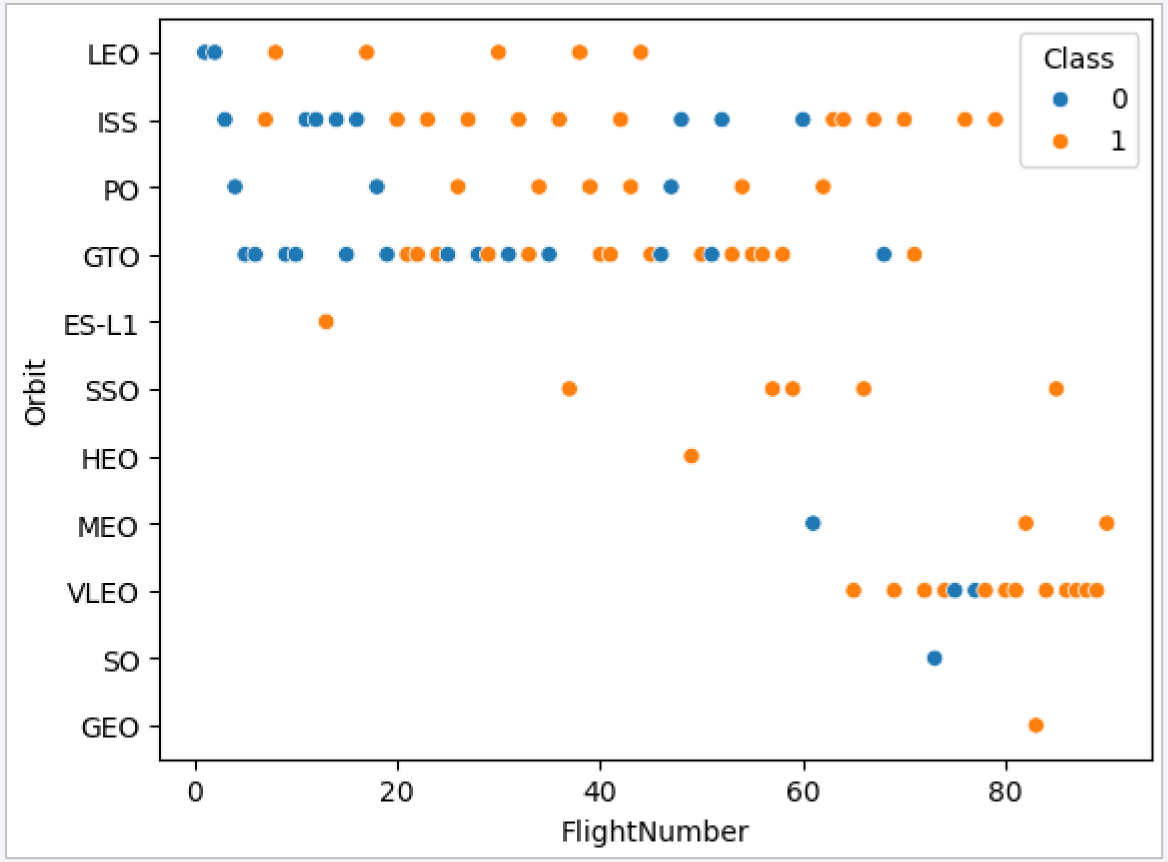
Success Rate vs. Orbit Type

- SO orbit has 0% success rate
- Orbits ES-L1, GEO, HEO and SSO show 100% success rate
- Other orbits are between 50% and 75% success rates



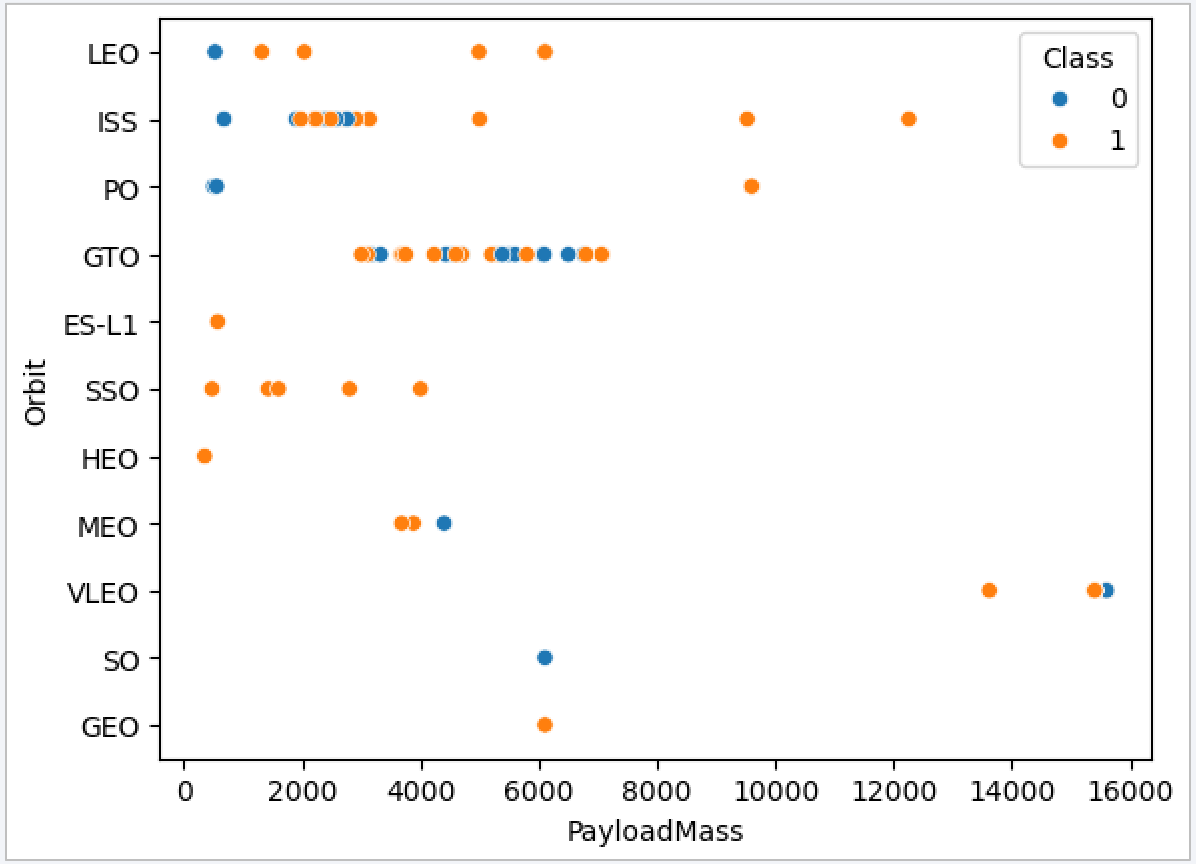
Flight Number vs. Orbit Type

- No clear connection between the flight number and orbit type
- ISS and GTO represent the majority of launches



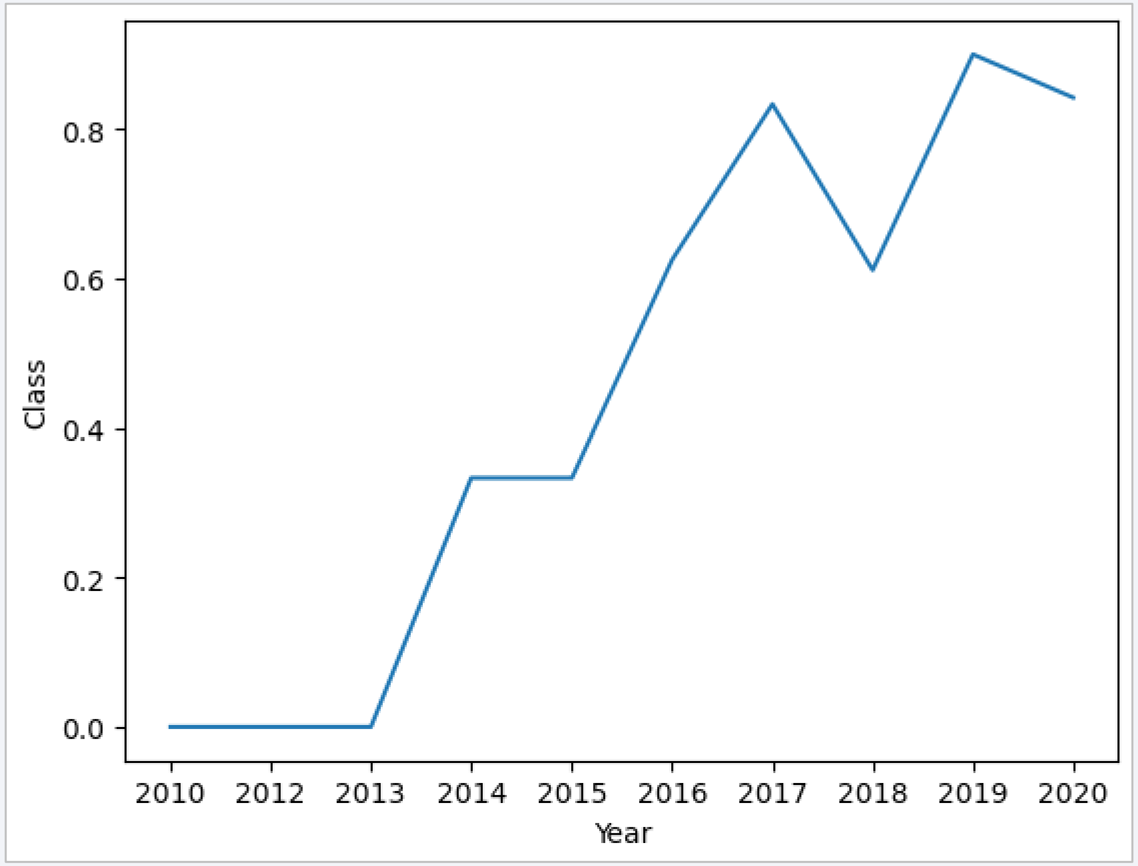
Payload vs. Orbit Type

- Payloads above 10k were only placed in specific orbits: ISS, PO and VLEO
- GTO orbit has more failures as the payload increases



Launch Success Yearly Trend

- There is a consistent increase in success rates since 2013
- In 2018 there was a drop in success rate of missions



All Launch Site Names

- The launch sites as queried from the database
- The distinct function was used to drop the repeating values

```
%sql select distinct launch_site from SPACEXDATASET;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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 was filtered to include only sites starting with ‘CCA’ in the WHERE clause
- The top 5 records are obtained using the LIMIT function

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
45596

- The query was filtered to include only customer called ‘NASA (CRS)’

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%'

 * sqlite:///my_data1.db
Done.
```

average_payload_mass
2534.6666666666665

- The query was filtered to include only booster version containing ‘F9 v1.1’

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing_outcome = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

first_successful_landing
2015-12-22

- The query was filtered to include only successful ground landings on pad
- The min function finds the minimal date

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXDATASET where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

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

- The query was filtered using two conditions
- Displaying the specific boosters in all matching records

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	total_number
Failure (in flight)	1
Success	98

- Total of 98 successful missions vs 1 failure in flight

Boosters Carried Maximum Payload

- The maximum payload was loaded 12 times
- A nested query was used to calculate the max payload for the where clause

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

* sqlite:///my_data1.db
Done.

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

```
%%sql select substr(date, 6,2) as month, date, booster_version, launch_site, landing_outcome from SPACEXDATASET
    where landing_outcome = 'Failure (drone ship)' and substr(date,0,5) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Only 2 flights that ended with failure to land on a drone ship were in 2015
- The substr function was needed as the existing sql database doesn’t support native date functions

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

- The landing outcomes between the provided dates are shown here
- The query is ordered and ranked

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXDATASET
    where date between '2010-06-04' and '2017-03-20'
    group by landing_outcome
    order by count_outcomes desc
```

 * sqlite:///my_data1.db
Done.

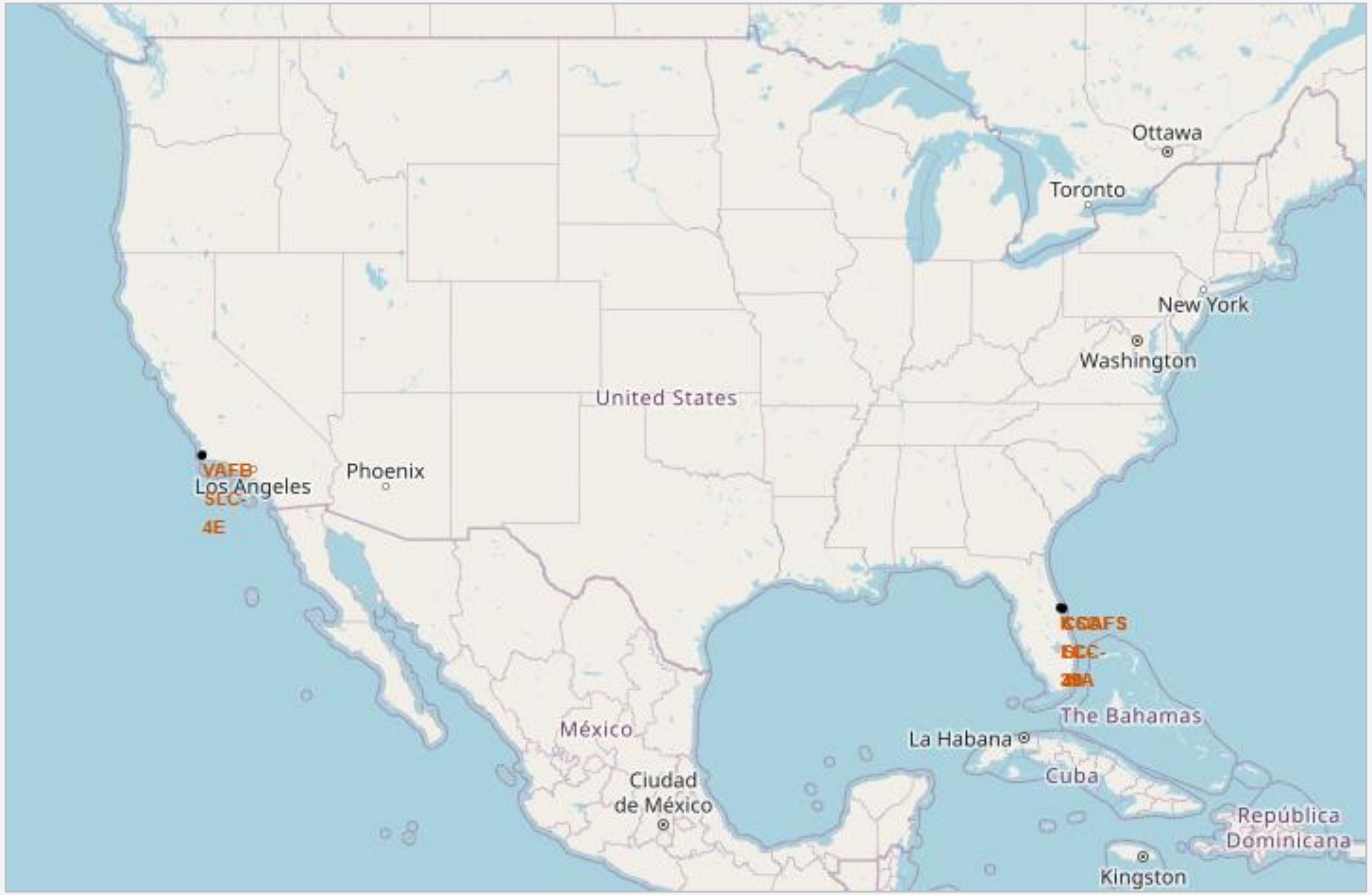
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

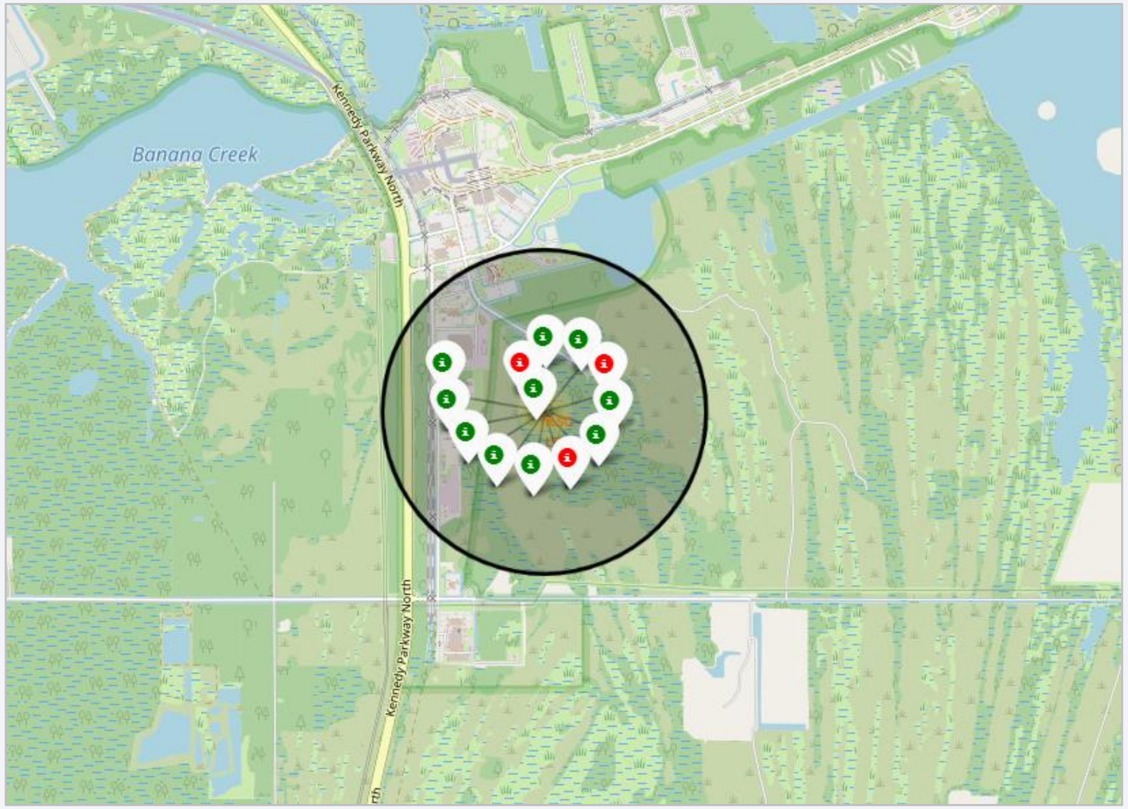
USA Launch Sites in California and Florida

- The two major locations for launching sites in the US are close to the equator
- The proximity to the equator is important as the earth's speed there is the highest thus helping the rockets reaching escape velocity
- The proximity to the ocean minimizes the risks to population and infrastructure in case of failures



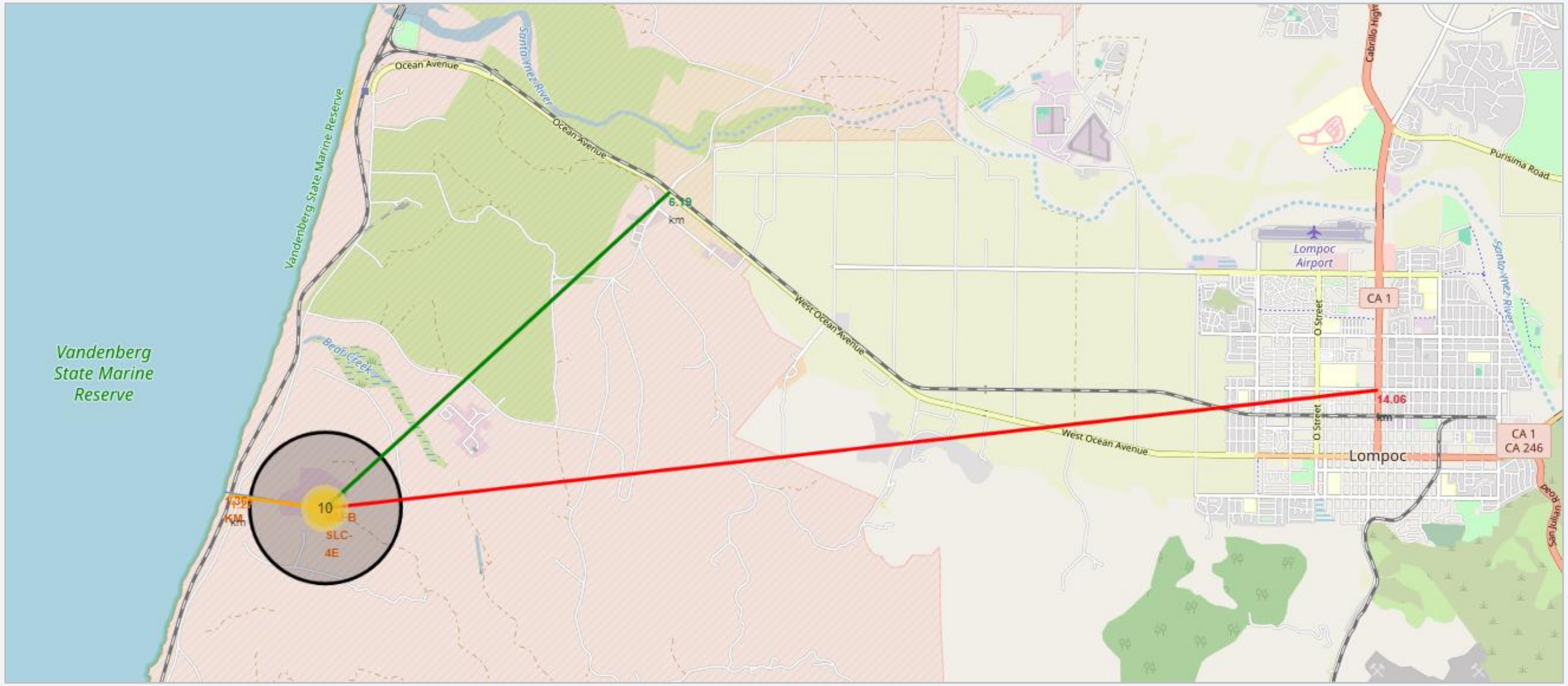
Color Labeled Launch Outcomes

- The marker cluster allows us to visually inspect the success rate of each launch site
- Green markers – Success
- Red markers – Failure
- The site shown here KSC LC 39A has a very high success rate for example ~76%



Distance of Landmarks and Infrastructure

- Relative distance of launch site VAFB SLC-4E to it's closest landmarks and infrastructures such as highway (6km) and railroad (1.2km)



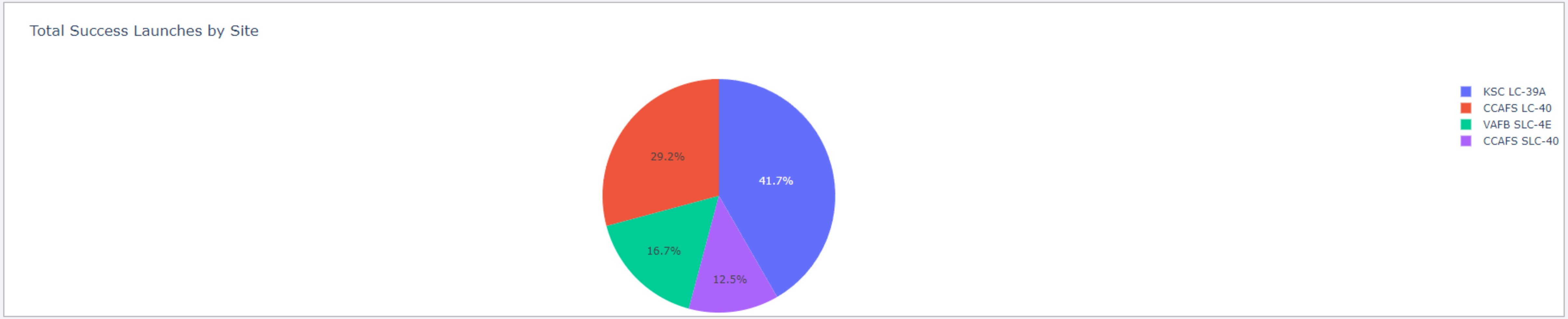
- The site is very close to the closest city (less than 15km), a distance a failed rocket can cover within seconds leading to some concern

Section 4

Build a Dashboard with Plotly Dash

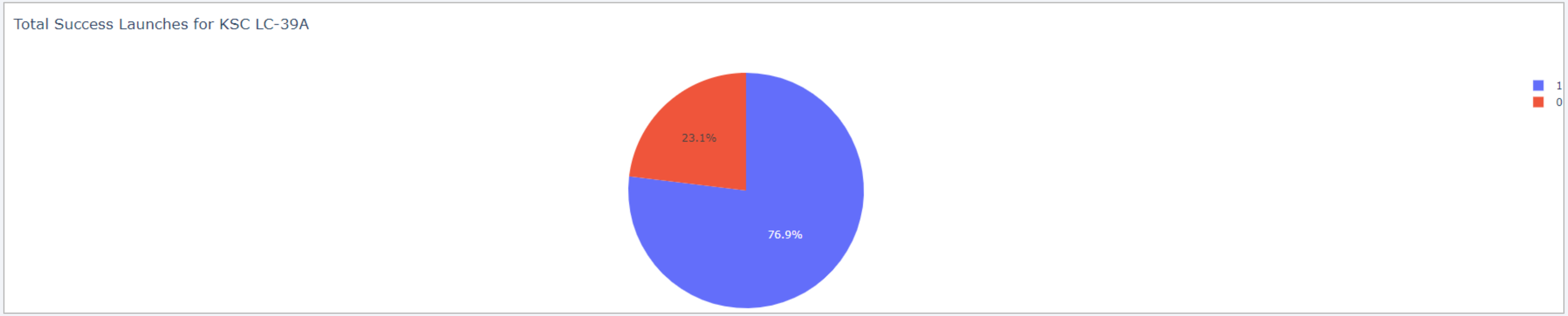
Total Launch Success for All Sites

- Site KSC LC-39A is clearly the leader in terms of successful launches with 41.7%

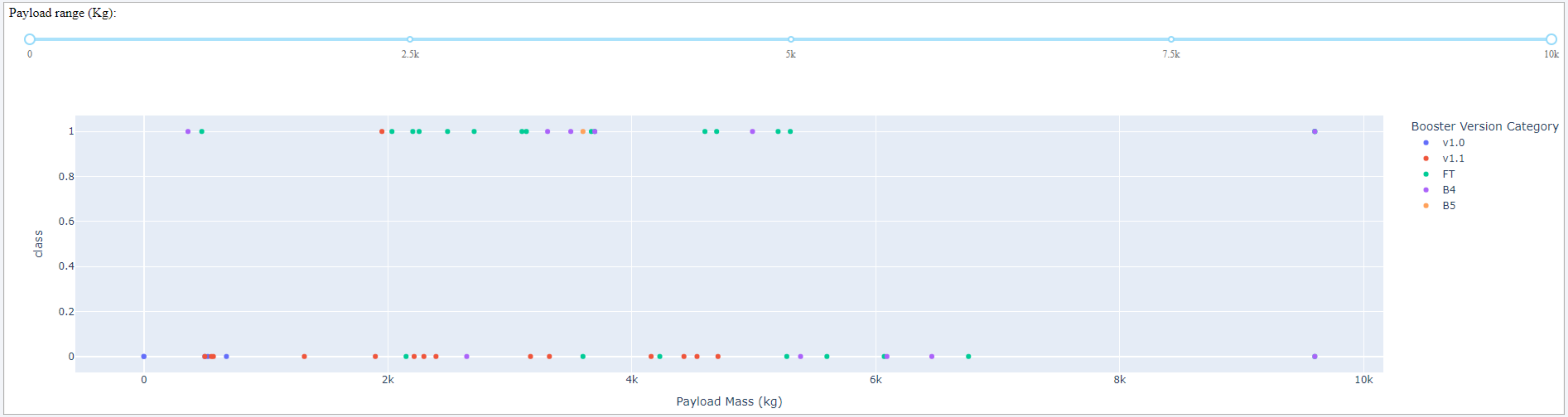


Launch Site With Highest Success Ratio

- Site KSC LC-39A is leader in success ratio (rate) among the four sites with 76.9%
- This is in addition to it being the site with highest count of successful launches



Payload vs. Launch Outcome for All Sites

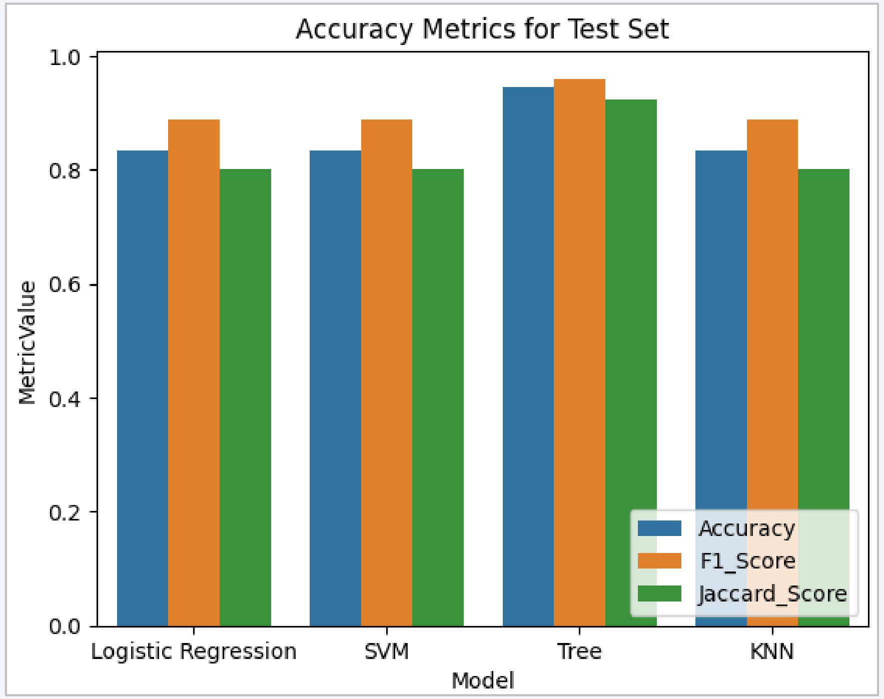


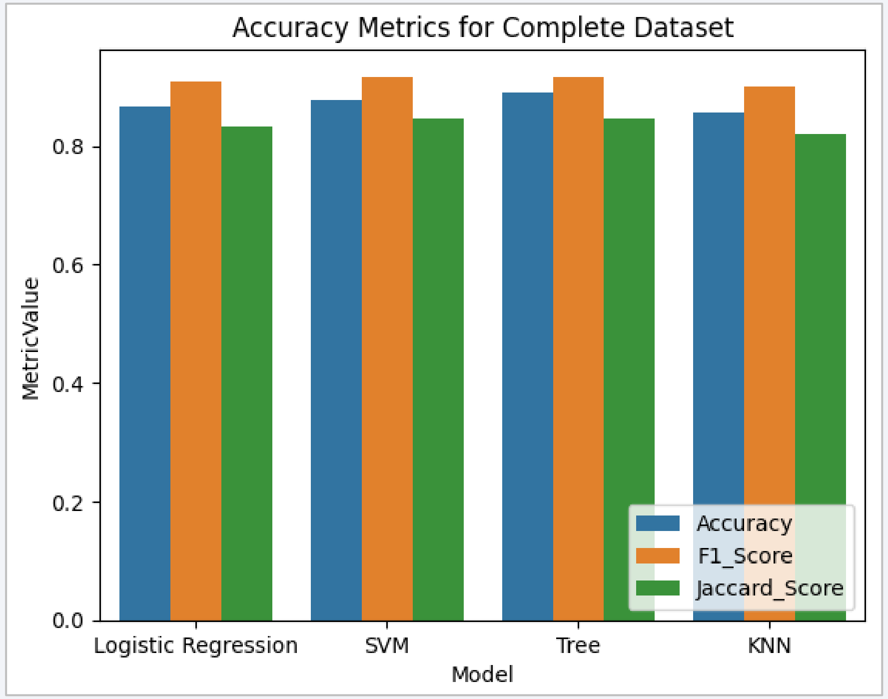
- Payloads in the range of 2000 kg and 5500 kg have much higher success rates
- Booster version v1.1 seems to have a very low success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

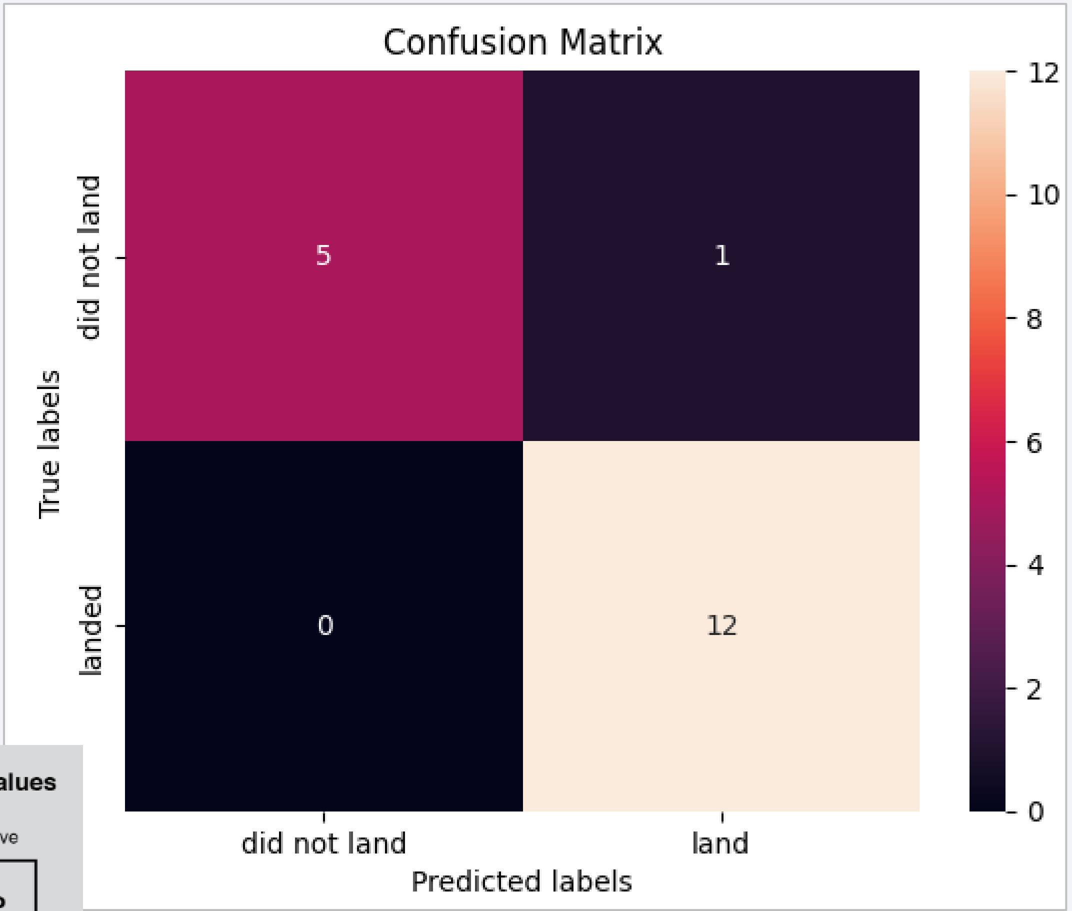


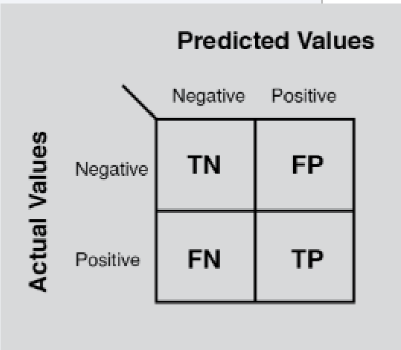


- The decision tree model shows better performance for both the test set and the complete dataset
- The random state of the models was set to 42, changing them does change the results slightly

Confusion Matrix

- The confusion matrix clearly shows that the decision tree predicts almost perfectly the results
- Only one False Positive result is present





Conclusions

- Using a decision tree model, we can very accurately predict the outcome of a launch
- There is a clear increase in launch success rate over the years peaking in 2019
- Some sites exhibit better success rates, such as KSC LC-39A
- Some orbits have 100% success rates, such as ES-L1, GEO, HEO and SSO

Appendix

- All data and notebooks used in the project can be found in the following git repository [Project GIT]

Thank you!