IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Clifford D Ngwarati
28/12/2022

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project background and context

- The SpaceX Falcon 9 rocket launch cost 62 million dollars whereas its competitors cost upward of 165 million dollars each. The price difference is explained by the fact that SpaceX can reuse the first stage. If a competitor determines if the SpaceX first stage will land, the competitor can go on and determine the cost of the launch. This information is valuable to competitors if it wants to compete with SpaceX for a rocket launch contract

Questions to be answered

- To determine how variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing
- To determine the most suitable machine learning to predict the success of the first stage landing
- To determine the conditions that facilitate the best landing success rate

Section 1

Methodology

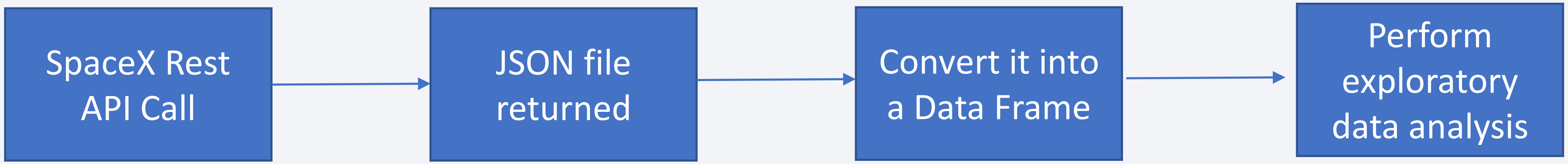
Methodology

Executive Summary

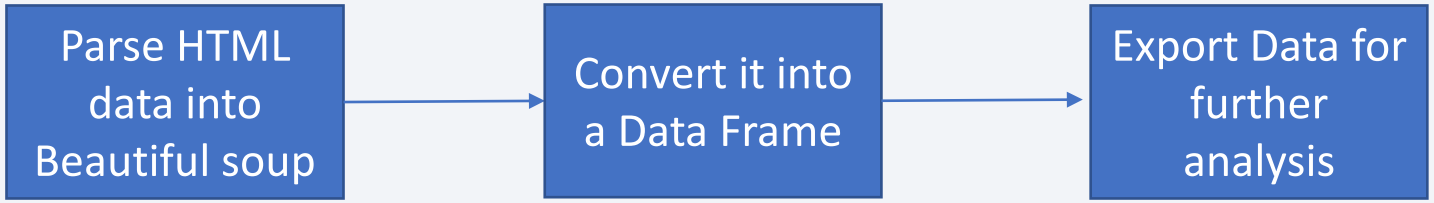
- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dealing with missing values
  - Dropping unimportant columns
  - Hot encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected via the SpaceX Rest API and web scrapping on Wikipedia
- You need to present your data collection process use key phrases and flowcharts. The high-level data collection flow chart for the SpaceX Rest API is shown below

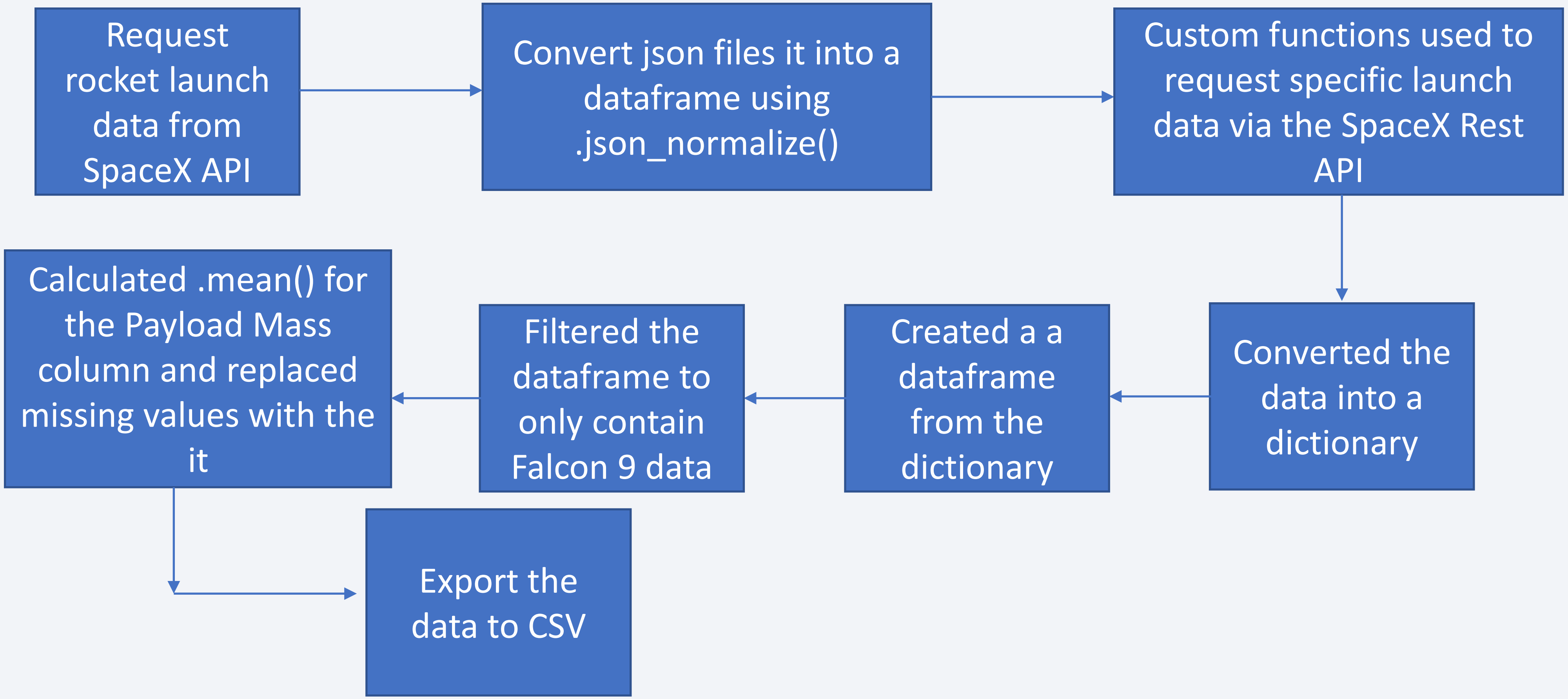


- The high-level data collection flow chart for web scrapping on Wikipedia is shown below



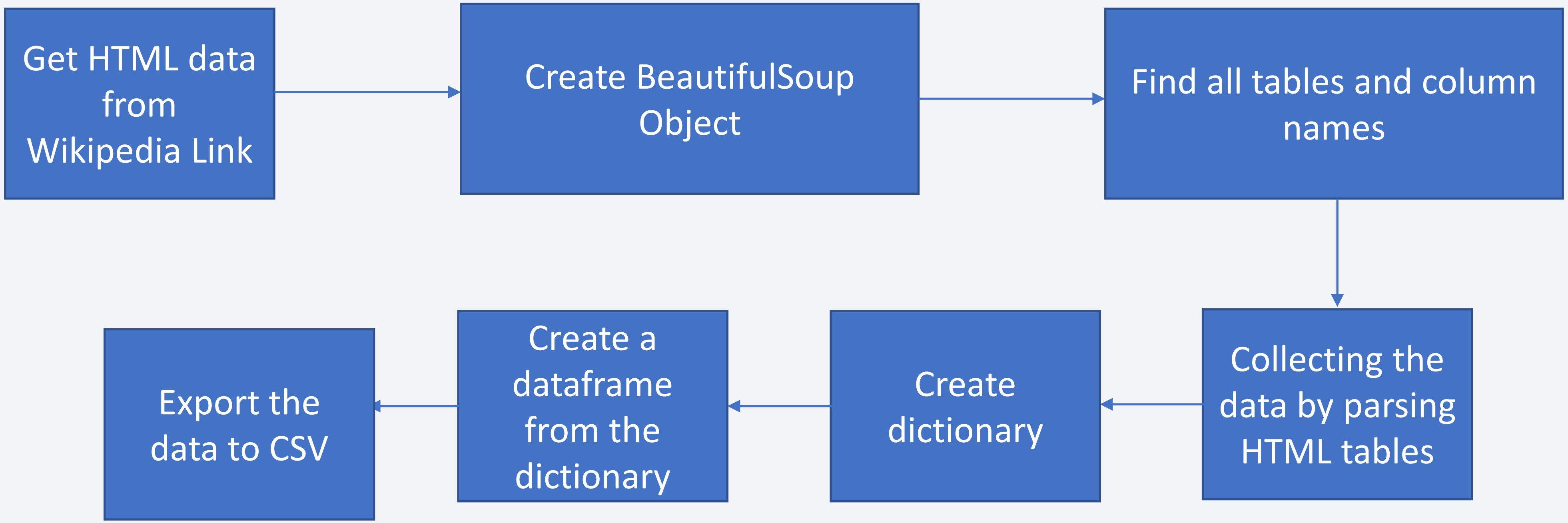
Data Collection – SpaceX API

- The detailed data collection SpaceX REST API calls flowchart is as follows



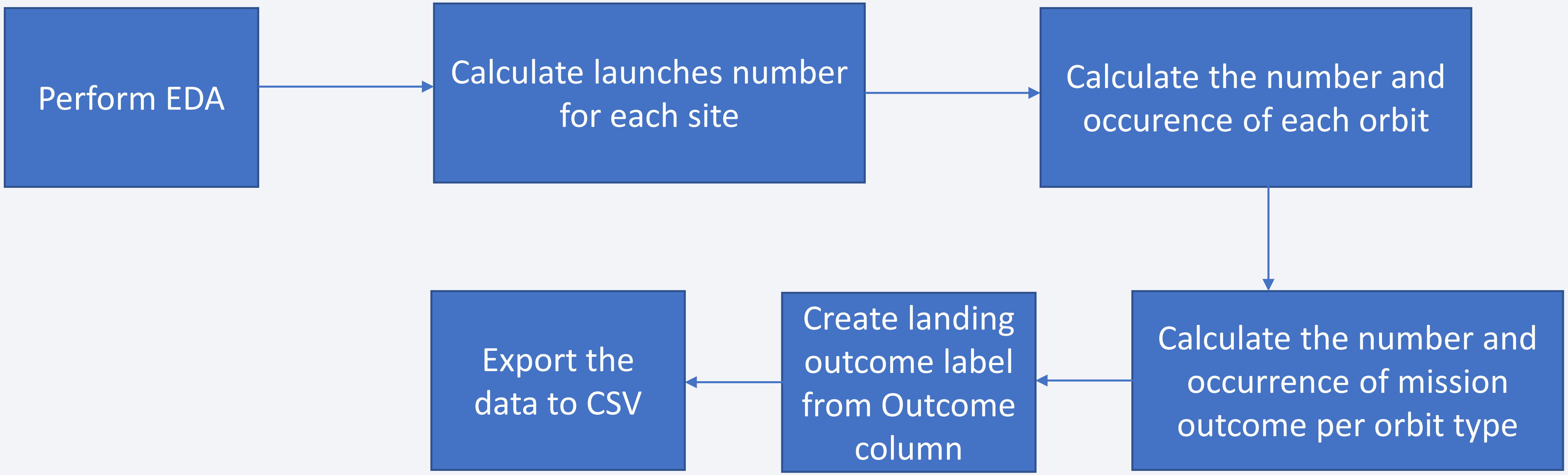
Data Collection - Scraping

- The detailed data collection web scrapping flowchart is as follows



Data Wrangling

- We had situations in which the first stage booster did not land successfully. In order to simplify the analysis, we decided to transform string variables into categorical variables, where 1 means the mission was a success and 0 means the mission was a failure
- The data wrangling process was as follows



EDA with Data Visualization

The following Scatter Plots were created

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Orbit vs. Flight Number
- Payload vs. Orbit Type
- Orbit vs. Payload Mass

Scatter plots show relationship between variables.

The following bar and line graphs were created respectively

- Success rate vs. Orbit
- Success rate vs. Year

Bar graphs show the relationship between numeric and categoric variables whereas show trends in data over time (time series).

EDA with SQL

The following SQL queries were performed

- Displaying the names of the unique lauunch 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, faiilure landing_ouutcomes in drone ship, booster versions, launch_site for the months in year 2015.
- •Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Build an Interactive Map with Folium

Markers of all Launch Sites:

- - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using
- its latitude and longitude coordinates as a start location.
- - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude
- and longitude coordinates to show their geographical locations and proximity to Equator and
- coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to
- identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- - Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an
- example) and its proximities like Railway, Highway, Coastline and Closest City

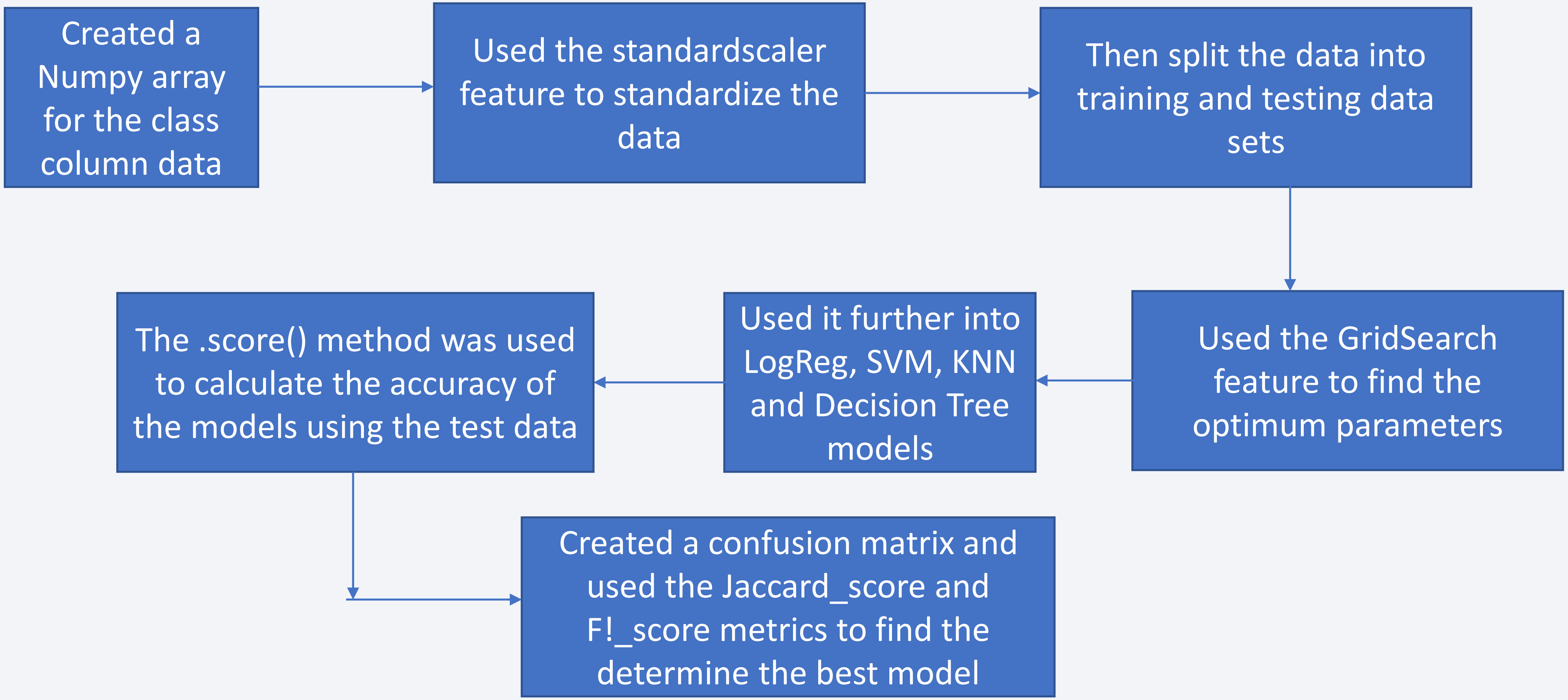
Build a Dashboard with Plotly Dash

These are the components of the dashboard: dropdown, pie chart, rangeslider and scatter plot components

- Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown).
- Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (plotly.express.pie).
- Rangeslider allows a user to select a payload mass in a fixed range(dash_core_components.RangeSlider).
- Scatter chart shows the relationship between two variables, in particular Success vs Payload Mass (plotly.express.scatter).

Predictive Analysis (Classification)

- Below shows how classification model was built, evaluated and improved,



Results

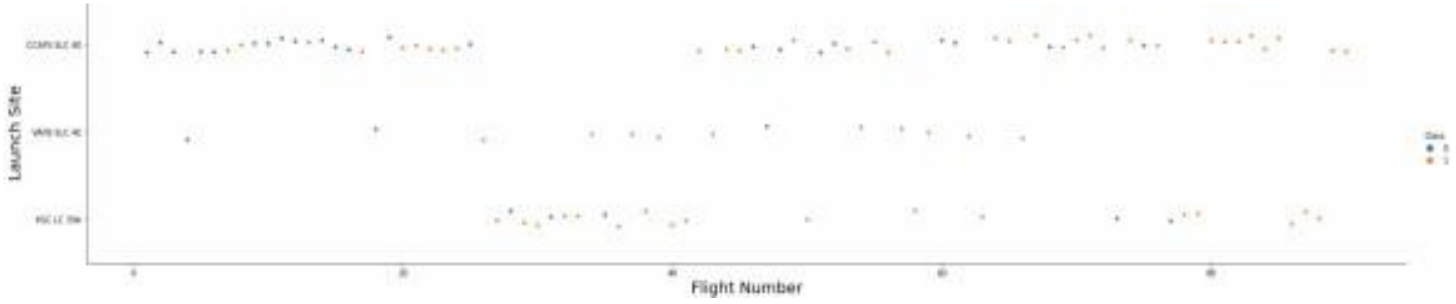
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

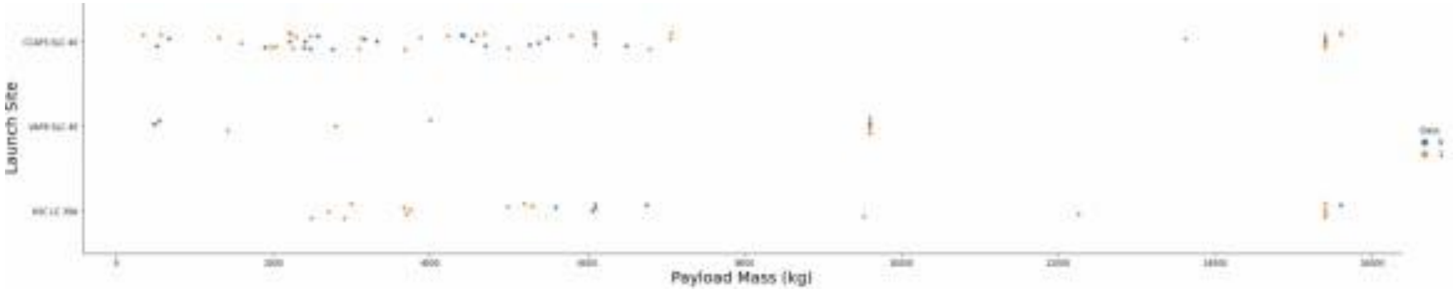
Flight Number vs. Launch Site

- The earlier flights failed whereas the latest flights succeeded which makes sense because SpaceX learnt from earlier failures for future success
- VAFB SLC 4E and KSC LC 39A have better success rates than other launch sites



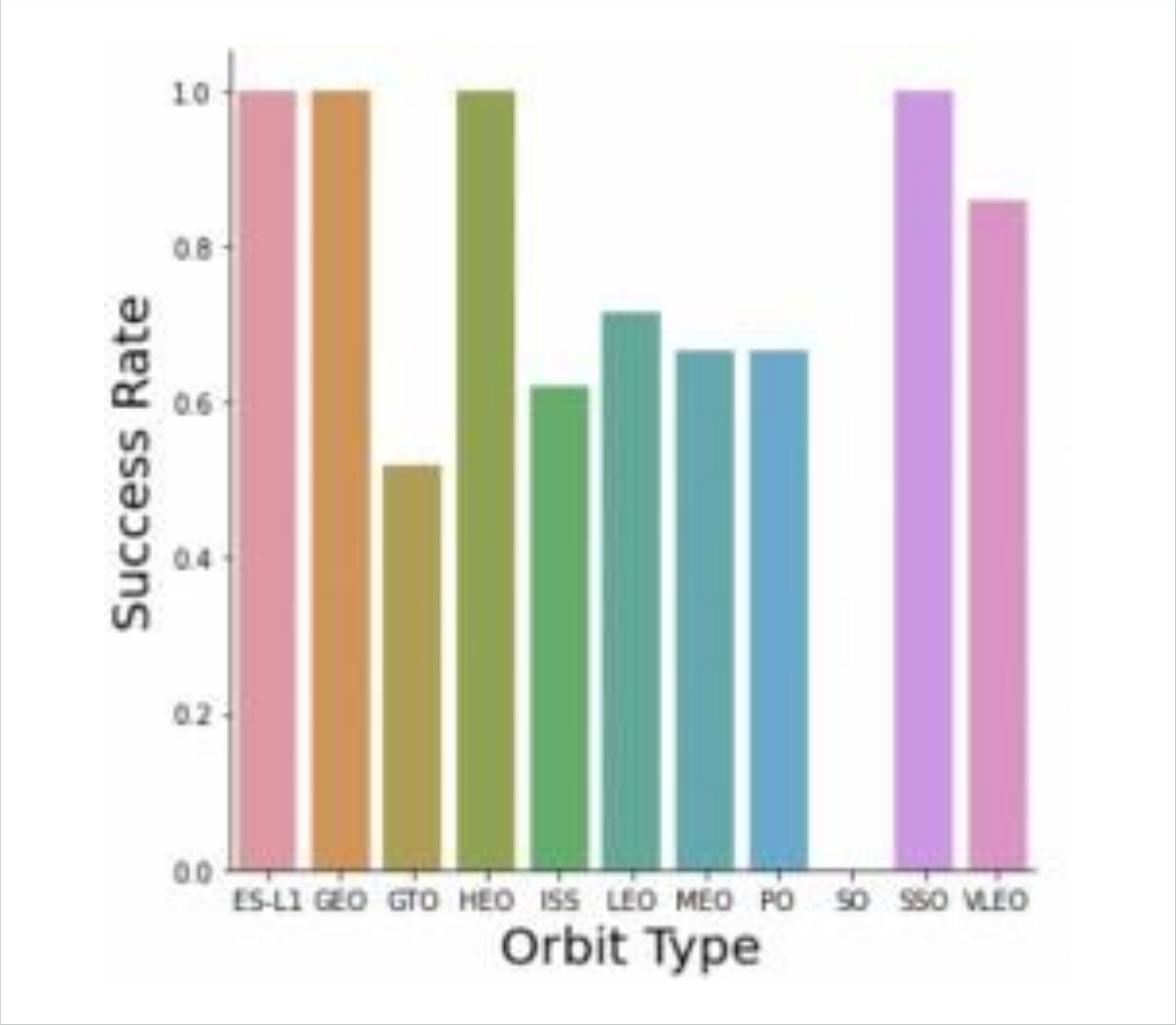
Payload vs. Launch Site

- For almost all the launch sites the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg



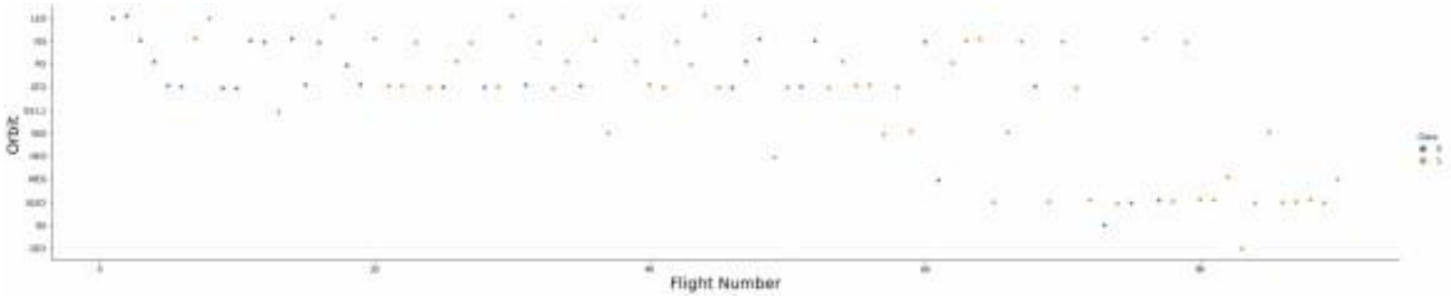
Success Rate vs. Orbit Type

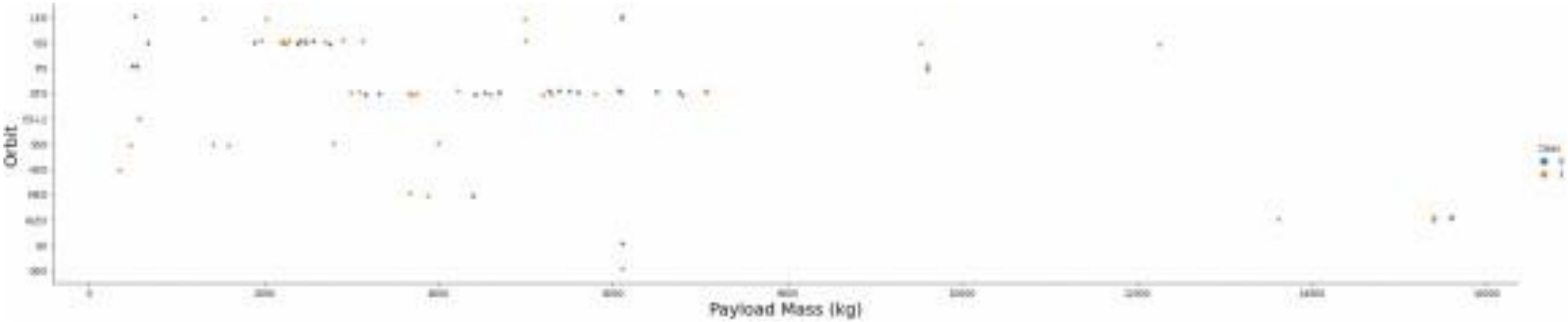
- Orbits with 100% success rate are: ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate is SO
- Orbits with success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO



Flight Number vs. Orbit Type

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



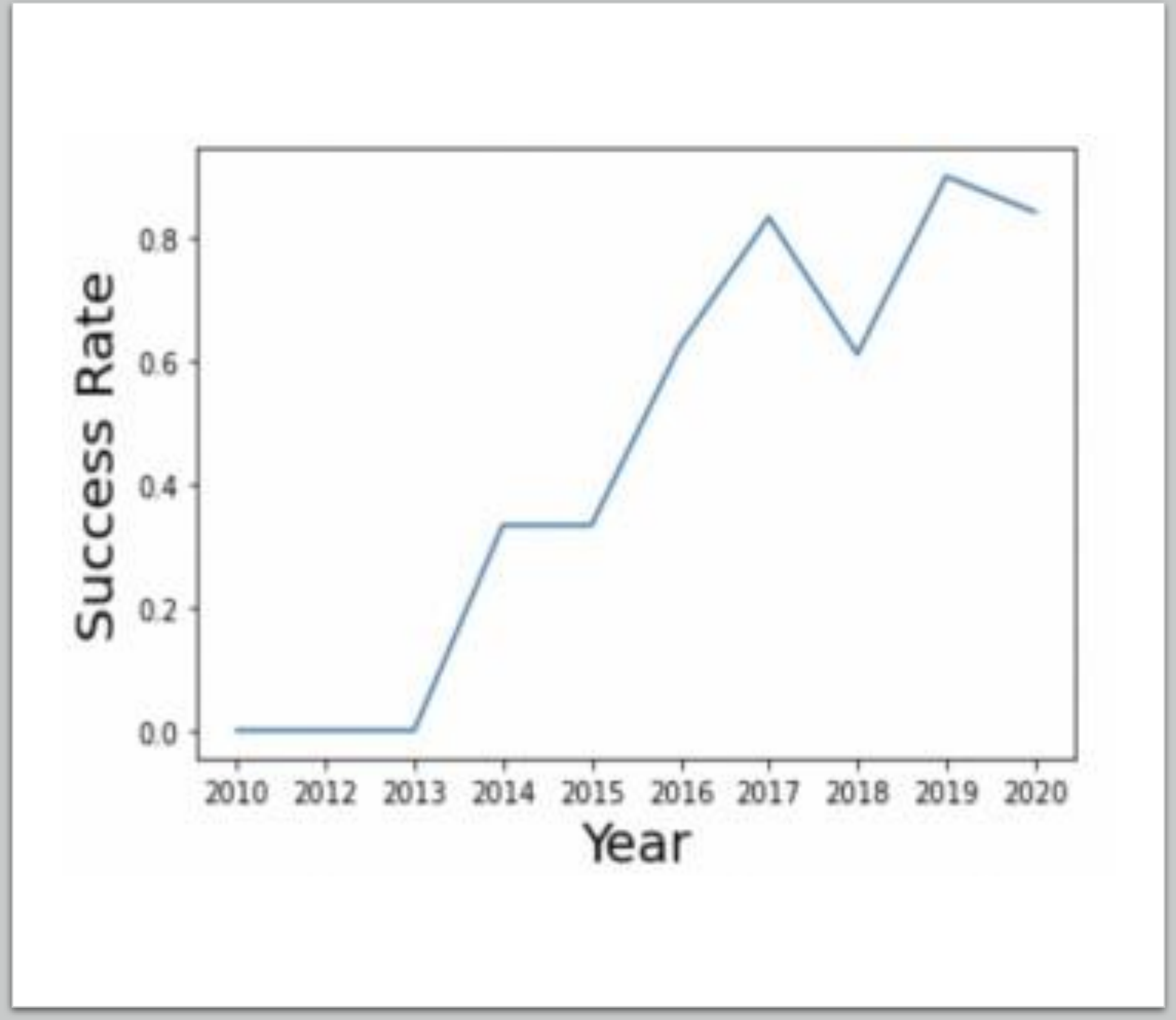


Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- SQL query displays the unique launch sites

```
%sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

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

Launch Site Names Begin with 'CCA'

- SQL query below displays 5 records where launch sites begin with `CCA`

```
%sql select * from SPACEXDATASET 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- SQL query calculates the total payload carried by boosters from NASA

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- SQL query to calculate the average payload mass carried by booster version F9 v1.1

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

average_payload_mass
2534

First Successful Ground Landing Date

- SQL query to find the dates of the first successful landing outcome on ground pad

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

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL query 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 booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 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

- SQL query to calculate the total number of successful and failure mission outcomes

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

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- SQL query to list the names of the booster which have carried the maximum payload mass

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

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 query list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- SQL query 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(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

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

Section 3

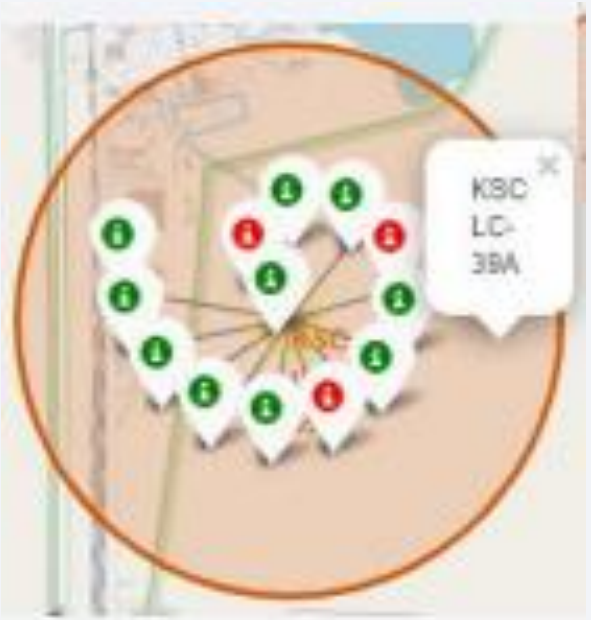
Launch Sites Proximities Analysis

Folium Map – Launch Sites

- Launch site can be close to the equator because any object launched from the equator already moves at maximum speed which is why it makes sense for rockets to launch close to the equator. Its also evident from the map that rockets are launched close to the sea/ocean in order to minimize debris interfering with normal way of life for people.

Color Labeled Markers

- Green marker represents successful launches. Red marker represents unsuccessful launches. We note that KSC LC-39A has a higher launch success rate.



Distances between CCAFS SLC-40 and its proximities

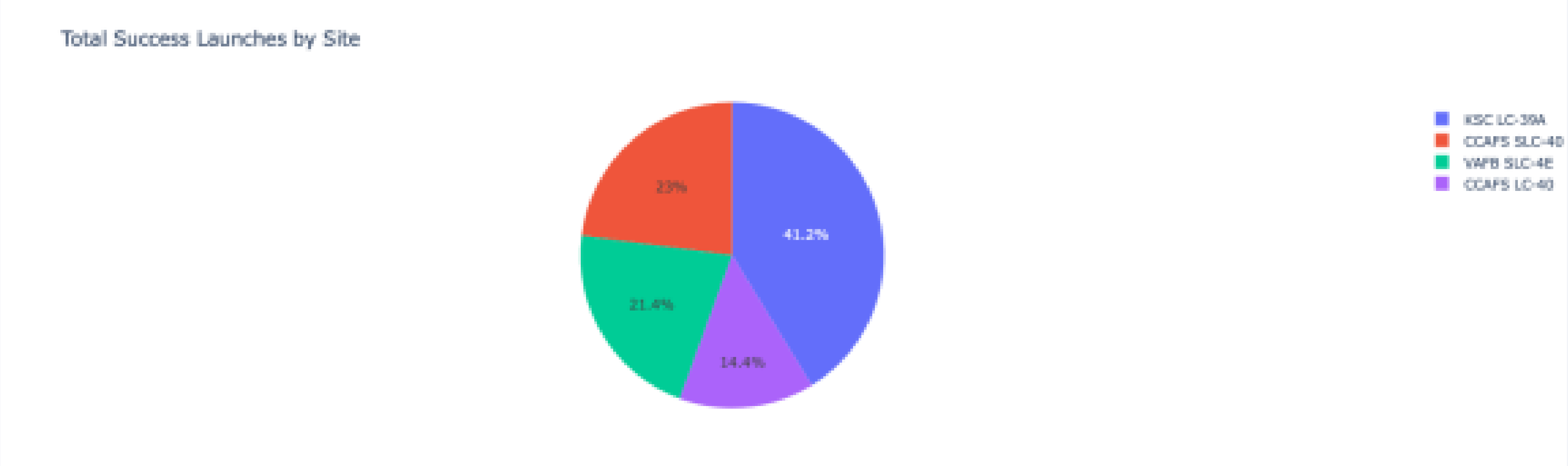
- Is CCAFS SLC-40 in close proximity to railways ? Yes
- Is CCAFS SLC-40 in close proximity to highways ? Yes
- Is CCAFS SLC-40 in close proximity to coastline ? Yes
- Do CCAFS SLC-40 keeps certain distance away from cities ? No

Section 4

Build a Dashboard with Plotly Dash

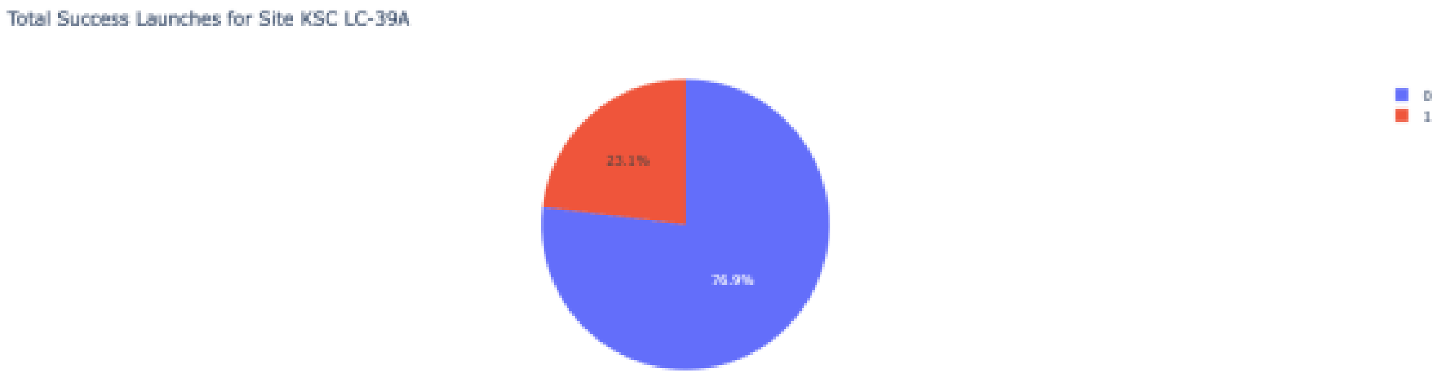
Launch success count

KSC LC-39A has the best success rate of launches.



Launch site with highest launch success ratio

- that KSC LC-39A has achieved a 76.9% success rate while getting a 23.1% failure rat



Payload Mass vs. Launch Outcome for all sites

- Low weighted payloads have a better success rate than the heavy weighted payloads

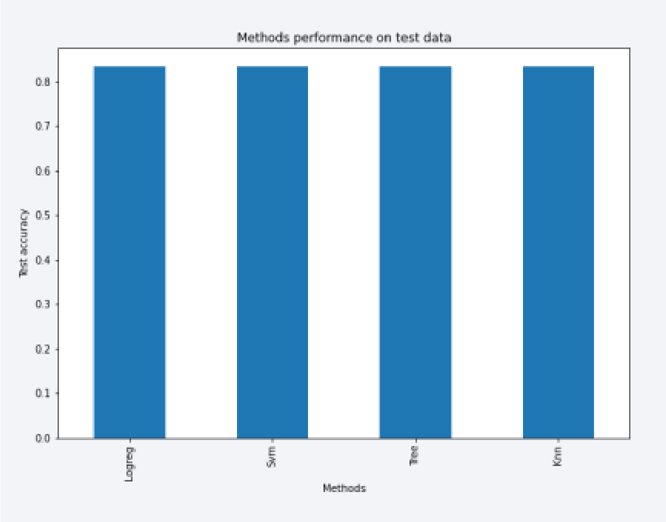
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- As we can see we could not separate the models based on the accuracy of the test data which could be due to the small test sample size

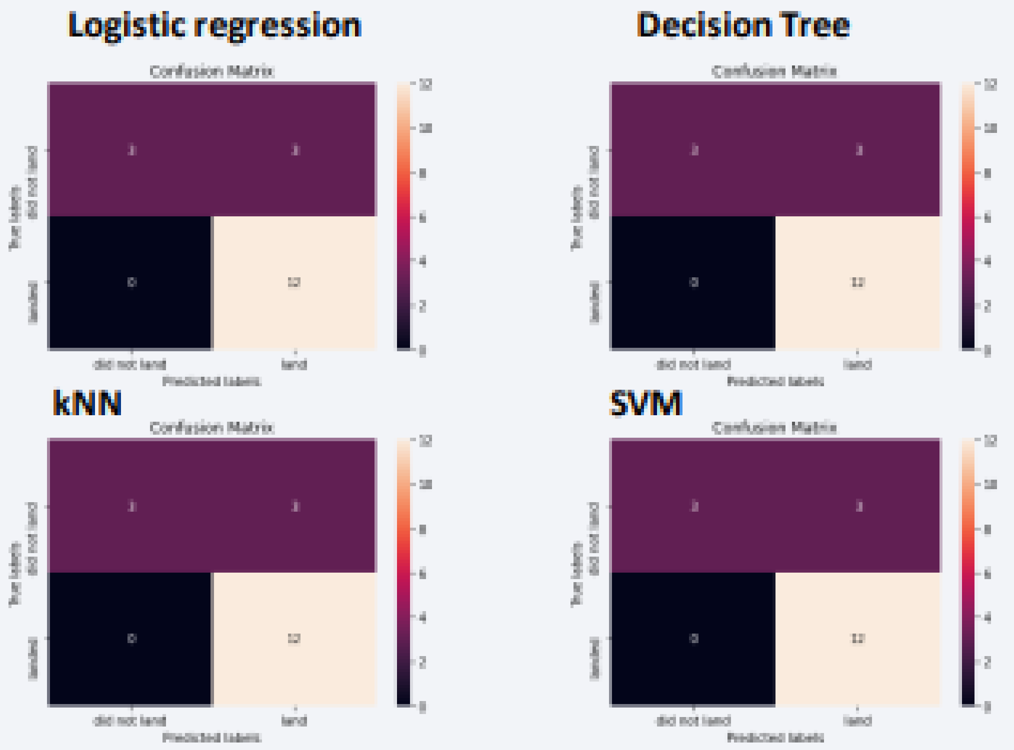
- When the entire data set is used the decision tree gives us better accuracy



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

- The logistic regression can distinguish between the different classes the major problem is false positives across all model



Conclusions

- The decision tree model is the most suitable algorithm for this project
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1
- Low weighted payloads have a better success rate than the heavy weighted payloads
- KSC LC-39A has the highest success rate when compare with other launch sites
- Launch sites are generally built far away from people close to the Equator line or close to the coast

Thank you!