IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SHIV
15-May-2024

Outline

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

Executive Summary

- **Summary of methodologies**
- Data Collection and Wrangling
- EDA with Visualization and SQL
- Folium Map and Plotly Dashboard Construction
- Predictive Analysis

- **Summary of all results**
- EDA and Interactive Analysis
- Predictive Analysis

Introduction

- Project background and context

"SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage."

- We want to predict the outcome of a launch depending on factors
- Factors like Orbit, Payload etc.
- Generally success rate increases over time
- We have used 4 types of ML models to determine the outcome. Spoiler: decision tree worked the best.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  SpaceX REST API and Web Scraping

- Perform data wrangling

  Filter, Impute and Encode the data for classification

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  Initialize the model object, train, test and validate to properly tune the hyperparameters for optimal results

Data Collection

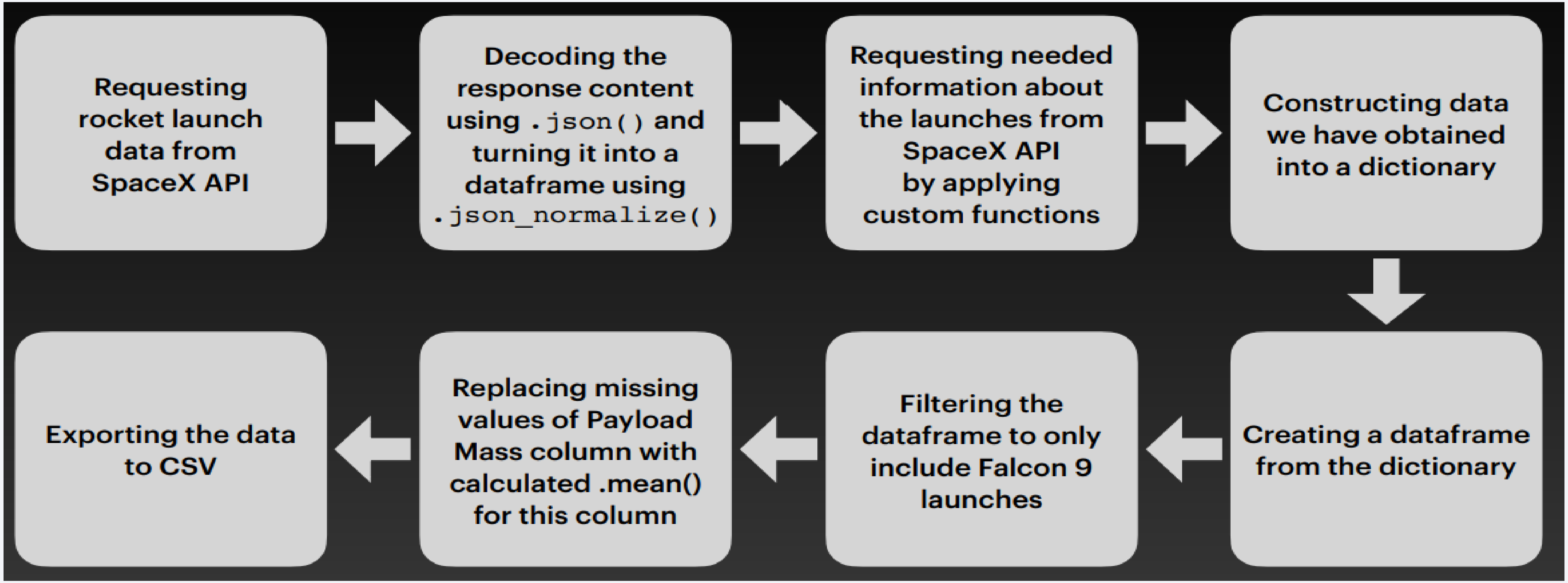
- Involves data collection from both the api and web scraping to complete the dataset
- Schema obtained by API

  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- Schema obtained by web scraping

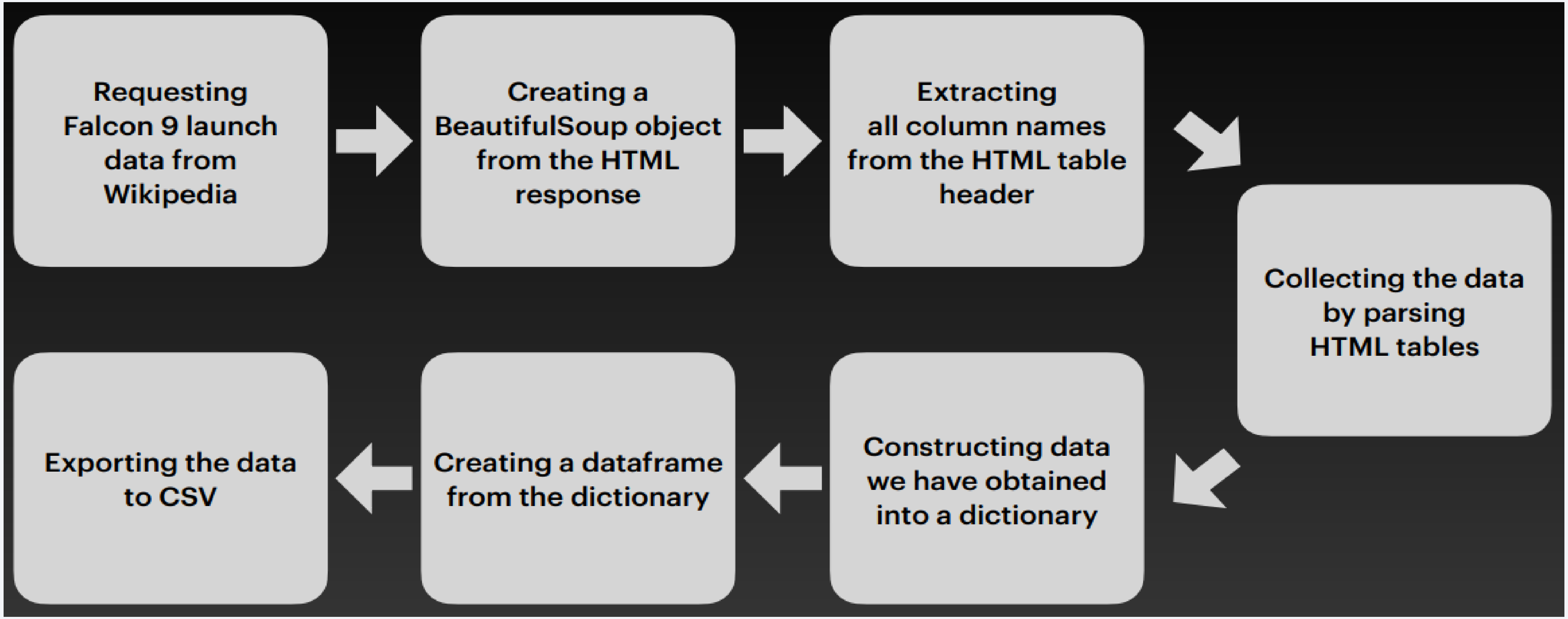
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



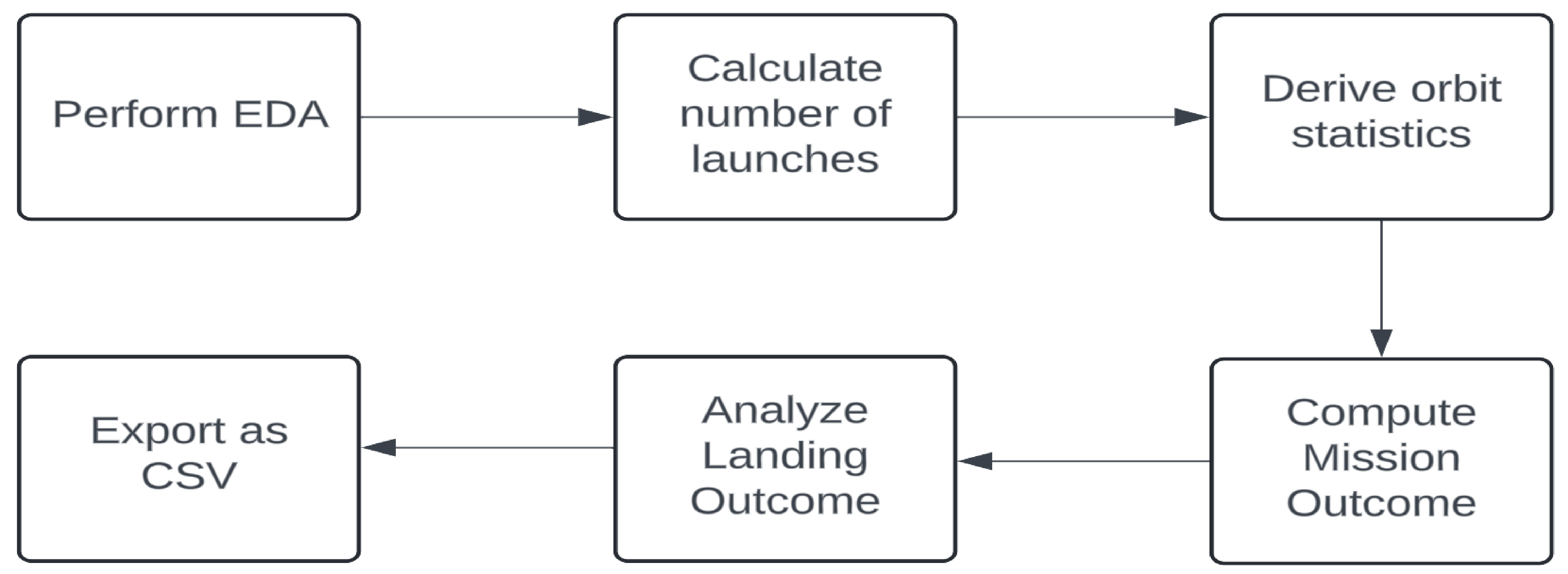
[GH WEB API]

Data Collection - Scraping



GH WEBSCRAPE

Data Wrangling



[GH DATA WRANGLING]

EDA with Data Visualization

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
Line charts show trends in data over time

Charts were plotted:
Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

GH EDA VISUAL

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

GH EDA SQL

Build an Interactive Map with Folium

1) Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center and all the launch centers using its latitude and longitude coordinates as a start location
2) Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
3) 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

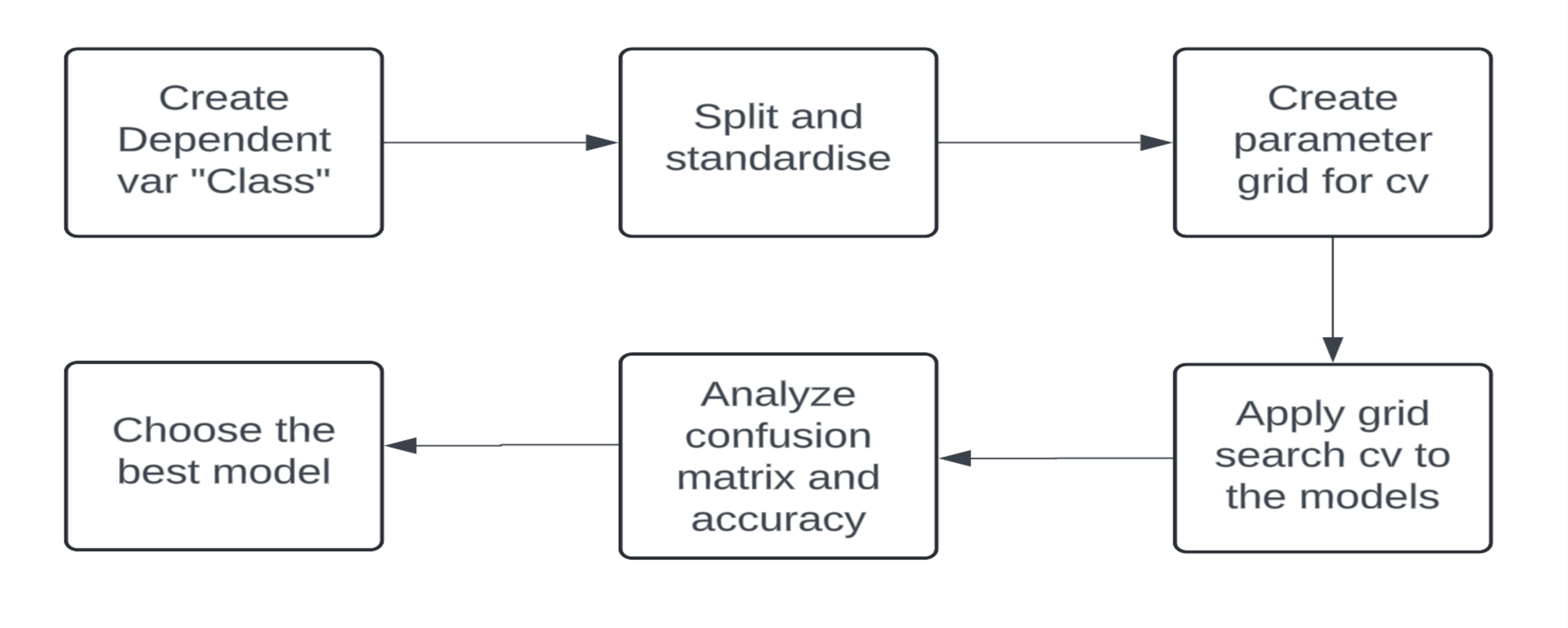
GH FOLIUM

Build a Dashboard with Plotly Dash

1) Pie Chart showing Successful Launches
2) Launch Sites Dropdown List
3) Slider of Payload Mass Range
4) Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions

GH DASHBOARD

Predictive Analysis (Classification)



[GH ML PREDICTION]

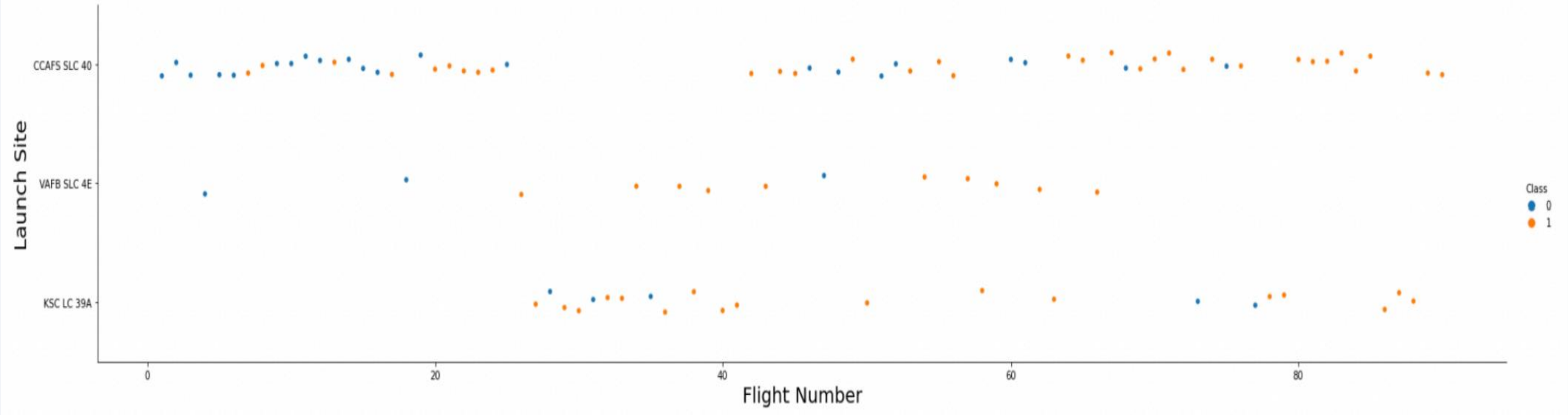
Results

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

Section 2

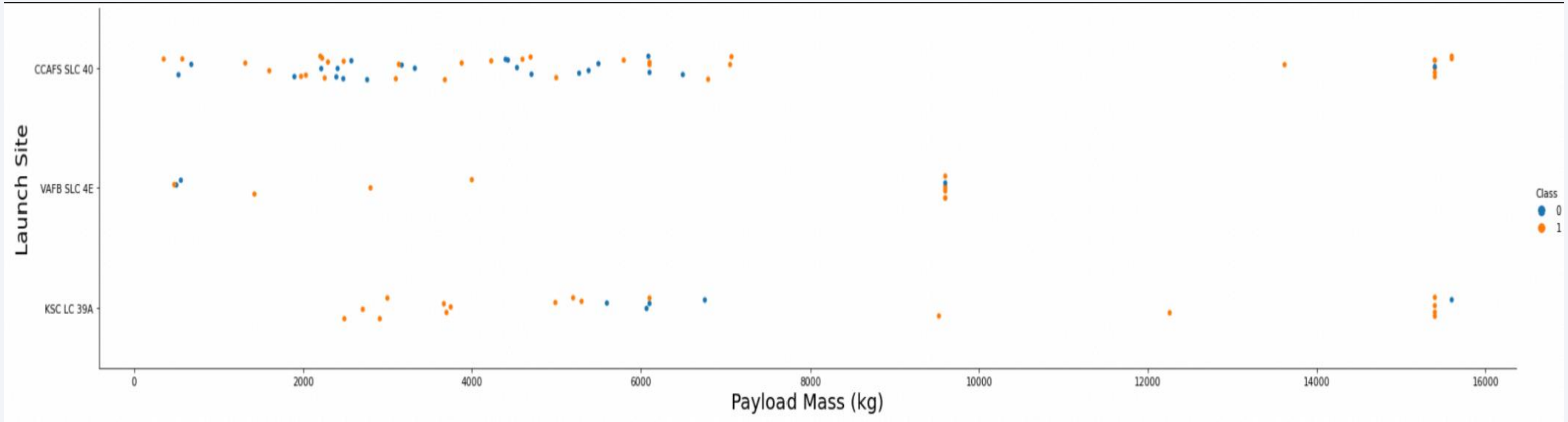
Insights drawn from EDA

Flight Number vs. Launch Site



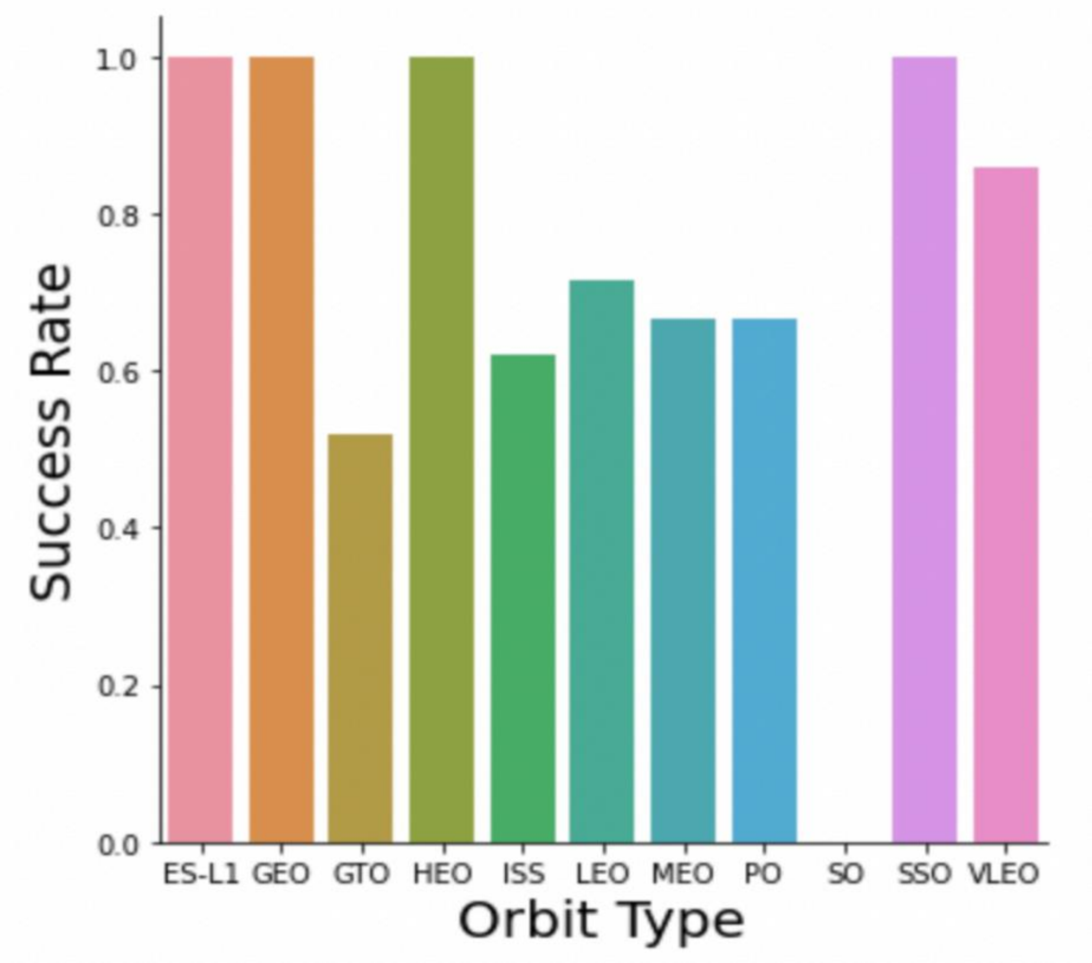
Scatter plot of flights are shown here, earlier flights have a low success rate which implies that success rate improves with time.

Payload vs. Launch Site



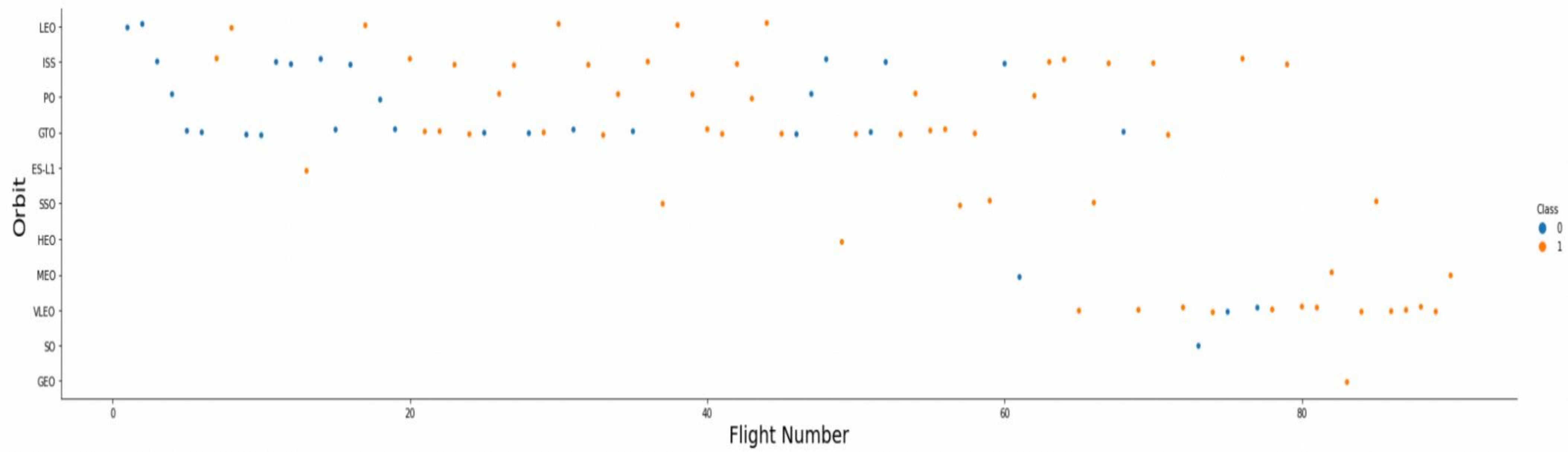
Generally, higher the payload, greater the chances of success. However, KSC LC 39A Has a high success rate.

Success Rate vs. Orbit Type



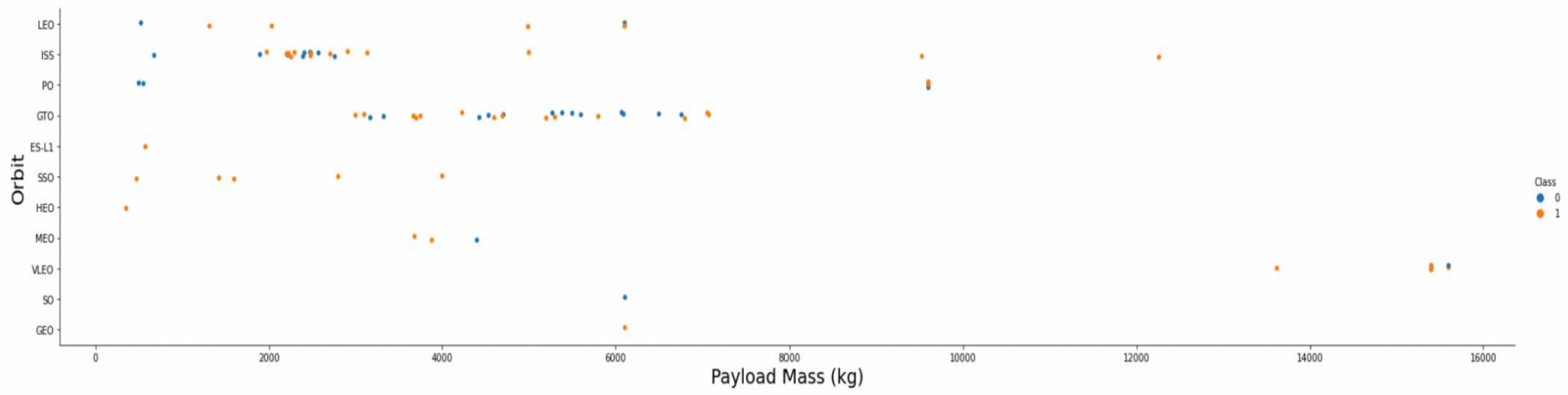
- SO has 0% success rate
- ES-L1, GEO, HEO, SSO have a 100% success rate

Flight Number vs. Orbit Type



Its difficult to fix a relationship when the orbit is GTO. In LEO the success rate seems to be increasing with flight number.

Payload vs. Orbit Type

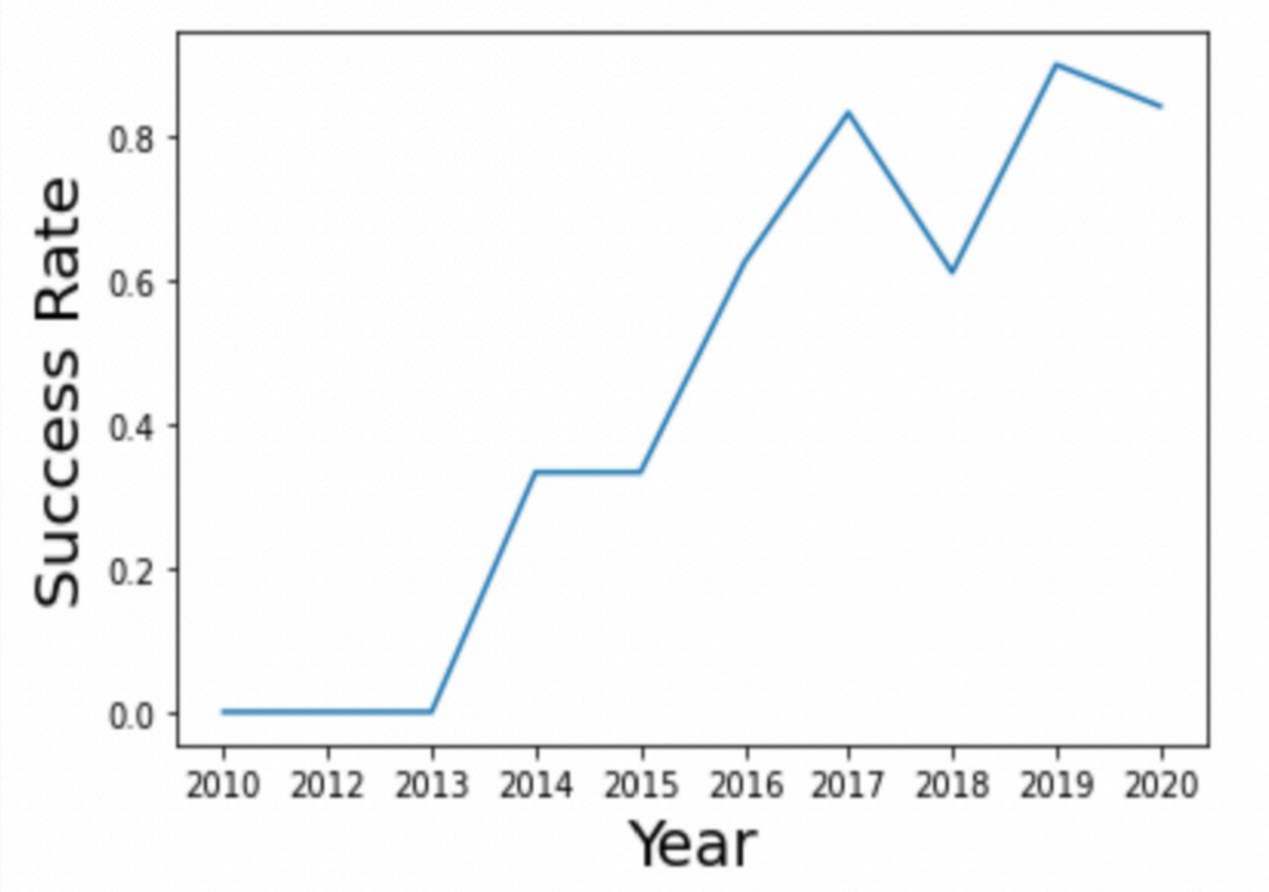


Generally heavy payloads imply a higher success rate in Polar LEO, however that changes in GTO.

Launch Success Yearly Trend

Success rate is generally increasing year-on-year rapidly from 2013-2017

A sharp drop occurs on 2018 and a slight drop again on 2020



All Launch Site Names

Display unique names of launch sites using the distinct keyword in query. A total of 4 launch sites exist.

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

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

- Finding 5 records where launch sites begin with `CCA`

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

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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	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

- Calculated the total payload carried by boosters from NASA

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

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud
Done.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Calculated 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%';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

average_payload_mass
2534

First Successful Ground Landing Date

- Found 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)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bl
Done.
```

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Lists 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;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculated the total number of successful and failure mission outcomes

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

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdo
Done.
```

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

Boosters Carried Maximum Payload

- Lists the names of the booster which have carried the maximum payload mass

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

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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

- Lists 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;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3119
Done.
```

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

- Ranking 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;
```

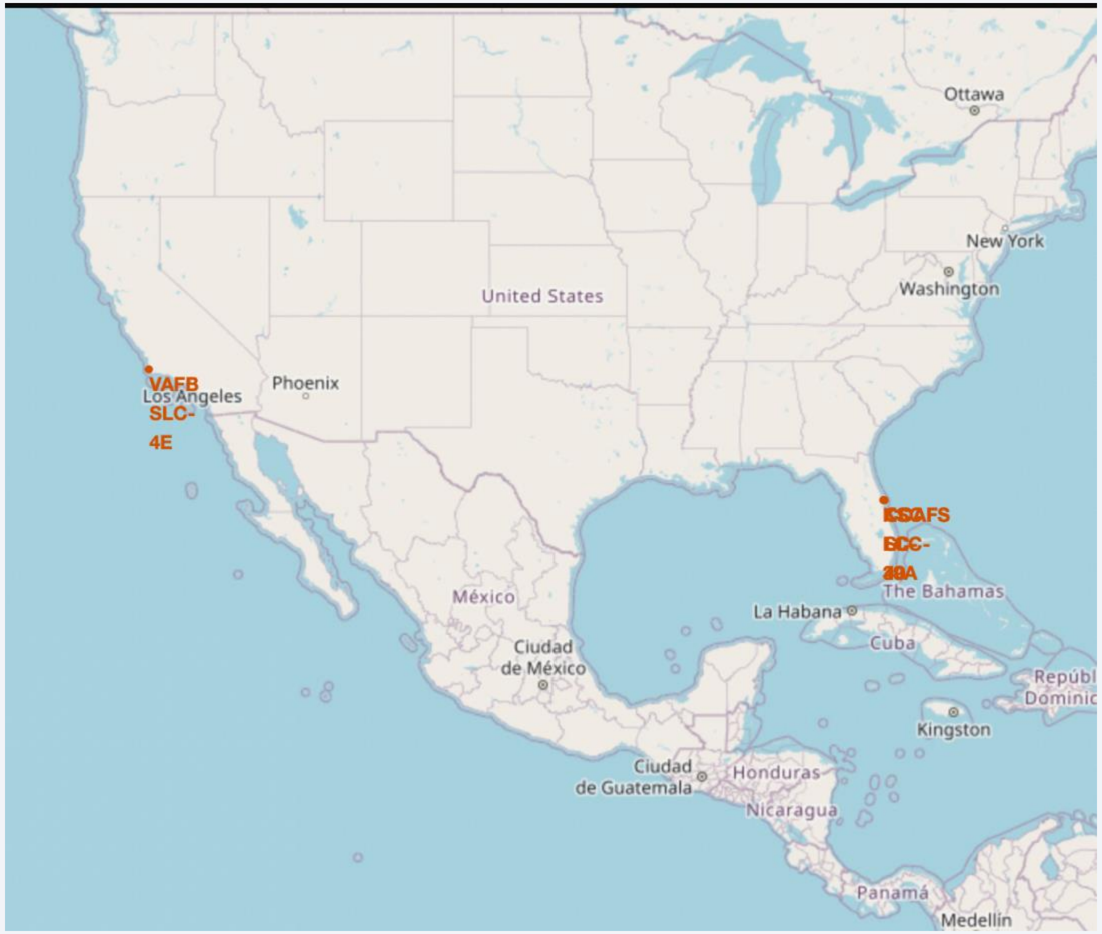
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb
Done.
```

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

Launch Sites

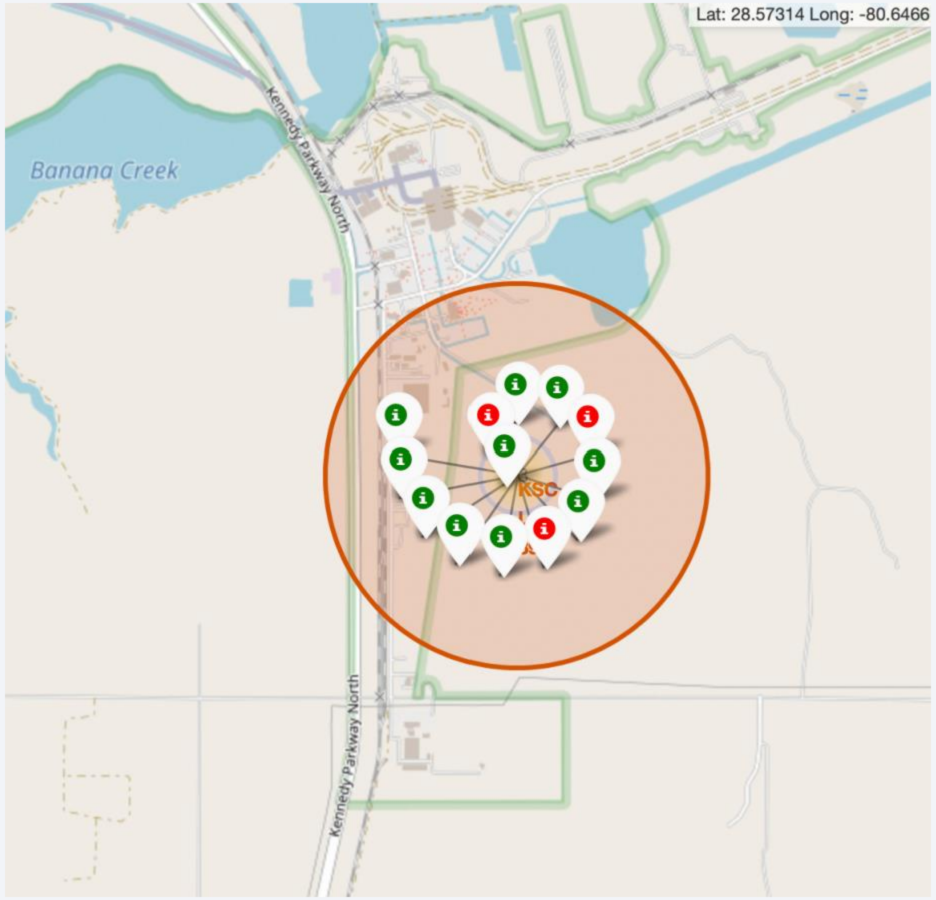


Most of Launch sites are in proximity to the Equator line and the coast to minimise risks of crashes. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

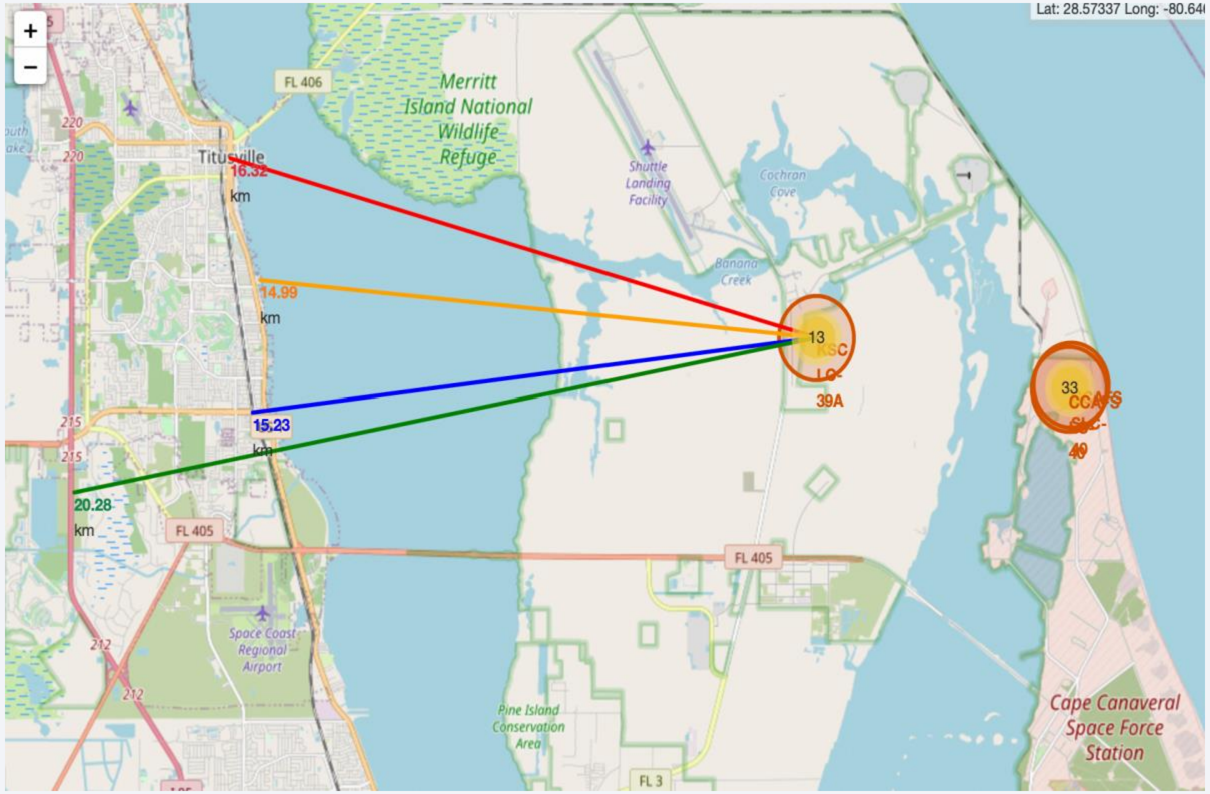
Launch Records

From the colour-labelled markers we should be able to easily identify which launch sites have relatively high success rates. –

Green Marker = Successful Launch
Red Marker = Failed Launch



Proximity of the Launch Site



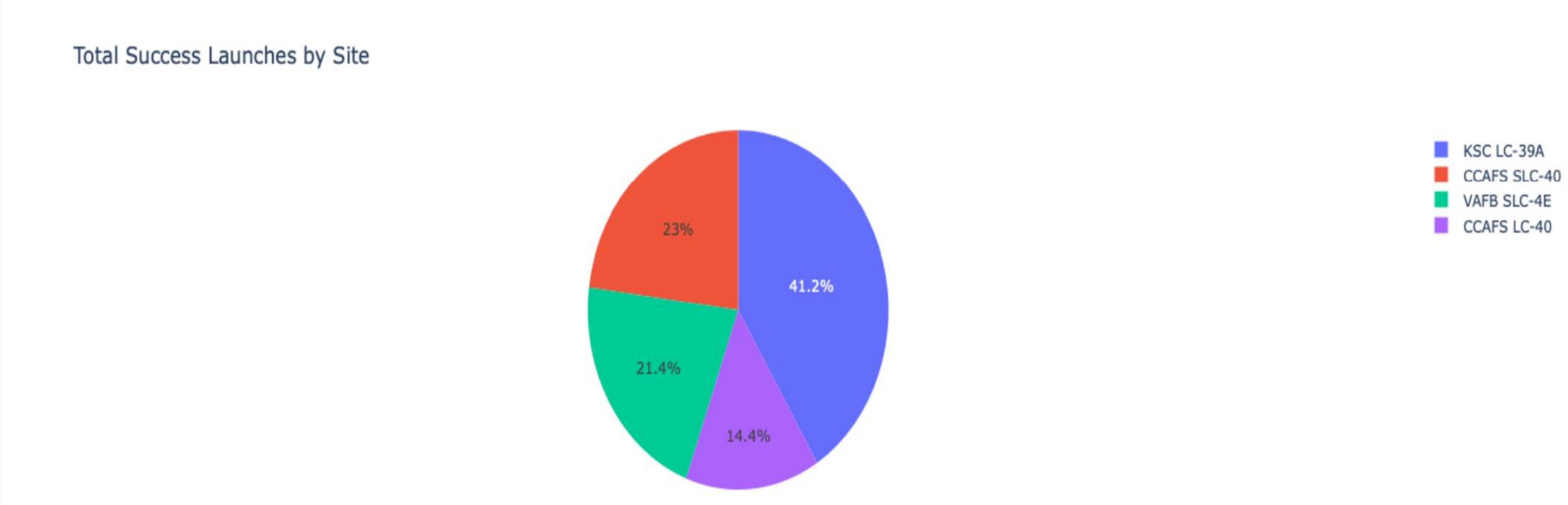
- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
- relatively close to railway (15.23 km)
- relatively close to highway (20.28 km)
- relatively close to coastline (14.99 km)

Section 4

Build a Dashboard with Plotly Dash

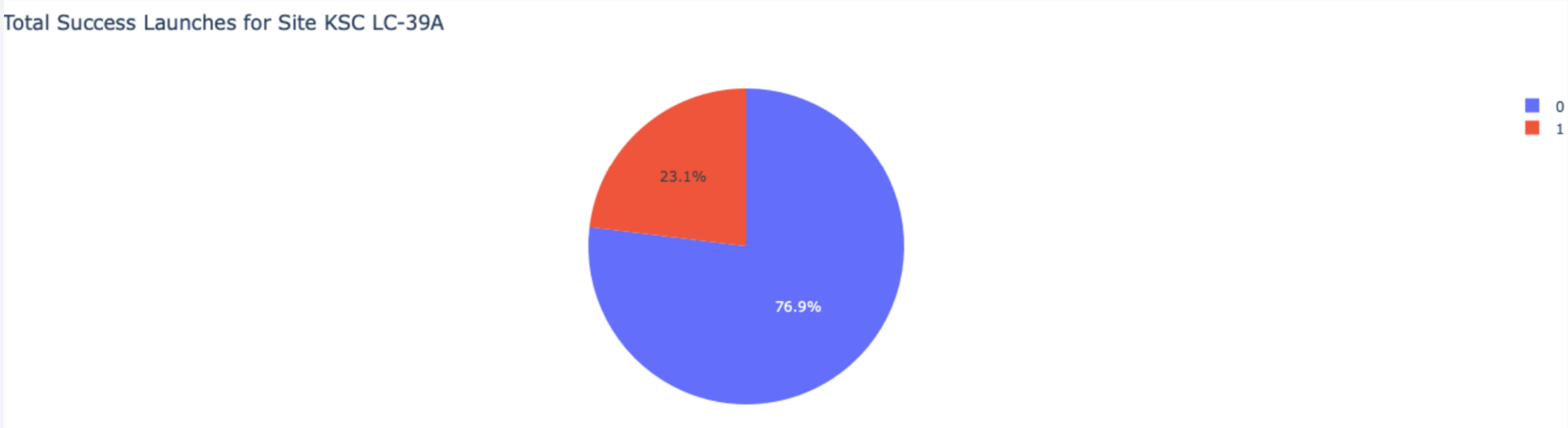
Success Share

- KSC LC-39A has the highest share of successful landings
- Whereas CCAFS LC-40 has the lowest



Stats of KSC LC-39A

KSC LC-39A is a reliable launch site with a decent success rate of nearly 77%



Payload vs Outcome

- We can see the relation between Payload Mass vs Launch Outcome for all sites. Most successes are in the range 2000kg-5000kg

Section 5

Predictive Analysis (Classification)

Classification Accuracy

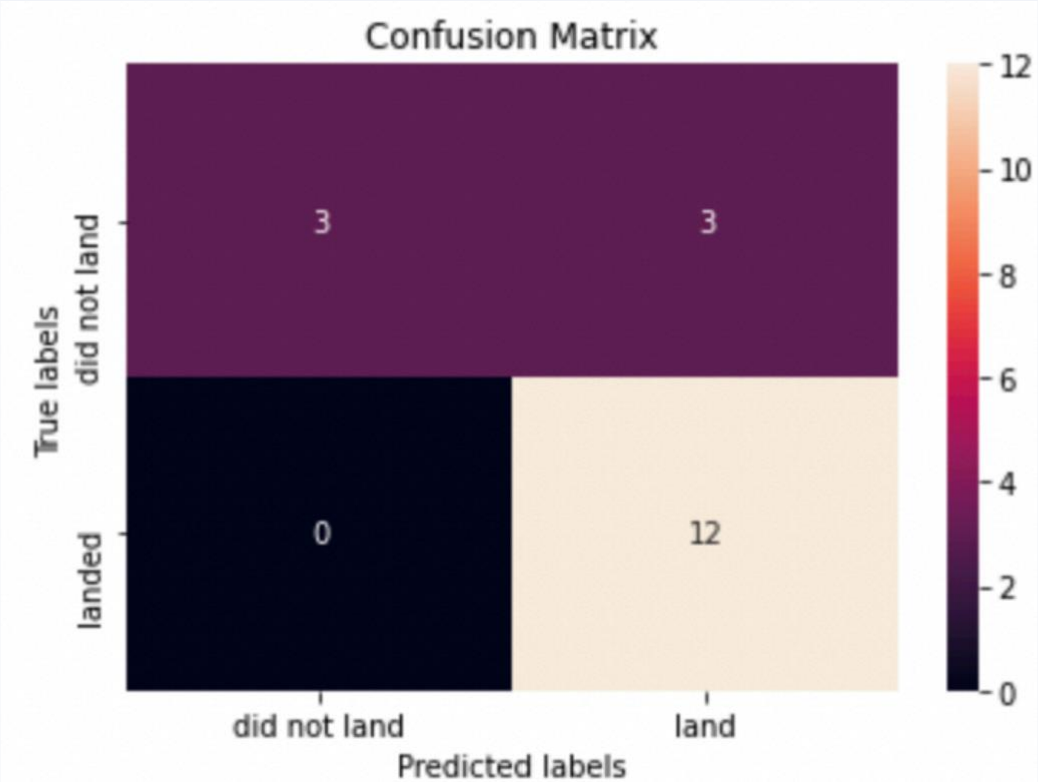
- 4 ML models – Logistic Regression, Support Vector Machine, Decision Tree and K-nearest Neighbours are built and validated

- The best accuracy is chosen from the grid search cv.

```
Log reg accuracy : 0.8464285714285713
SVM accuracy : 0.8482142857142856
Tree accuracy : 0.8892857142857145
KNN accuracy : 0.8482142857142858
```

Confusion Matrix

- A relatively high False Positive Rate is evident in the model



Conclusions

- Decision Tree has the best performance
- Launches with payloads in a specific range of payload(2000-5000kg) have a higher success rate
- As time goes by, the success rate of launches and landings is increasing
- KSC LC-39A is the most reliable launch site
- In the orbits ES-L1, GEO, HEO and SSO the success rate is a perfect 100%

Appendix

Resources used are majorly provided by:

https://www.ibm.com/

https://www.coursera.org/

Applied Data Science Capstone - Introduction - Week 1 | Coursera

SpaceX

Thank you!