Winning Space Race with Data Science

Kwabena Baah-Boakye
October 22, 2022

Outline

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

Executive Summary

Methodology

- Data Collection using web scraping and Space X API;
- Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive Visual analytics with SQL, matplotlib, seaborn, folium and plotly dash
- Machine Learning Prediction with different classification algorithms

Summary of all results

- Pay load has an effect on the success of a mission
- There is a 60% probability of successful landing

Introduction

Background and Context

The aim is to determine if Space Y can compete with Space X in the space race.

This conclusion can be drawn by determining if a mission will be successful based on factors like landing site and payload.

The results will inform Space Y for instance of which landing sites to use and Payload to send for the first stage of a mission to be successful and ships to be reusable.

Problems

Which Payload gives the best chance of success in stage one

Where should a rocket launch from

Where should a rocket land

Section 1

Methodology

Methodology

Data collection methodology:	• Secondary data on previous space missions were collected as follows: • Source: Wikipedia , Method: web scraping • Source: Space X's website , Method: API
Perform data wrangling	• One-hot encoding was used for feature engineering
Interactive visual analytics	• Folium and Plotly Dash
Exploratory data analysis (EDA)	• Visualization matplotlib and seaborn python libraries • SQL
Predictive analysis using classification models	• KNN, Logistic Regression, Decision Tree Classifier and Support Vector Machine

Data Collection

Data collection was done using get request to the SpaceX API.

- decoded using .json() function and converted to pandas dataframe using .json_normalize().

The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas data frame for future analysis.

In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.

Data was cleaned checked for missing values and missing values replaced

Data Collection – SpaceX API

Import Libraries and Create Helper Functions → Request and parse the SpaceX launch data → Filter out Falcon 9 Launches → Preliminary Data Wrangling

Major Libraries

- Requests (SpaceX Source Page)
- Beautiful Soup
- Pandas

Data Collection notebook available here

Data Collection - Scraping

Create Helper Functions → Created Response Object → Create Beautiful Soup Object from Response Object → Extract Column/Variable Names → Create data frame by parsing the launch HTML tables

Major Libraries

- Requests (Wiki Page)
- Beautiful Soup
- Pandas

Web scraping notebook available here

Data Wrangling

Exploratory Data Analysis to determine features and label

Summarization based on features/ columns

Landing outcome was created as a Label

Major Libraries Used

- Numpy
- Pandas
- Pandas

[Data Wrangling notebook available here]

EDA with Data Visualization

Scatterplots and bar plots for visualization of relationships between	flight number and launch Site,
	payload and launch site,
	success rate of each orbit type
	flight number and orbit type,
	the launch success yearly trend

EDA with Visualization Notebook Link

EDA with SQL

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string ‘CCA’
- Total payload mass carried by boosters launched by NASA
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

[EDA with SQL Github Link]

Build an Interactive Map with Folium

Markers, circles, lines and marker clusters were used with Folium Maps

- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicate groups of events in each coordinate, like launches in a launch site;
- Lines are used to indicate distances between two coordinates.

[Source Code for Folium Maps]

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data

- Pie charts for Probability of success per launches by site
- Scatter Plots to analyze relationship between payload mass and outcome

This application allowed to interactively analyze the relation between payloads and launch sites

[Source Code for Plotly Dash App]

Predictive Analysis (Classification)

Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

Data Preparation and standardization → Split data into train and test datasets → Determine the optimal hyper parameters with GridSearchCV → Iteratively train each model with training dataset → Iteratively test each model with test dataset → Compare model scores to identify the bets classifier

[Source Code for Predictive Analysis]

Results

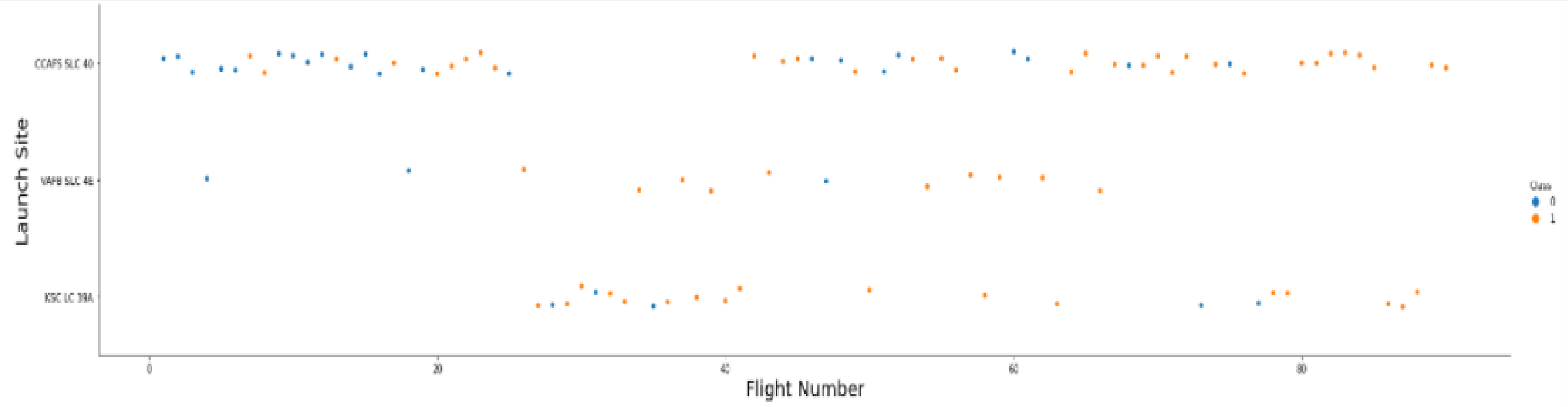
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

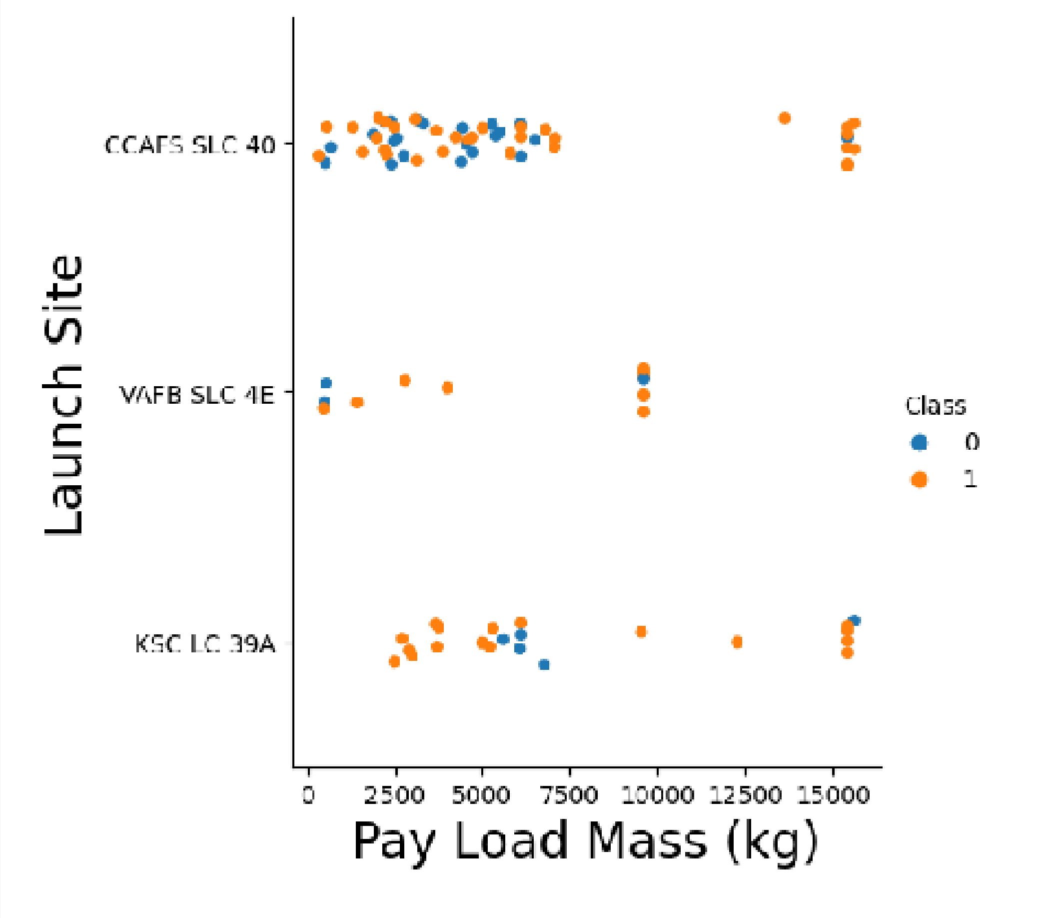
Insights drawn from EDA

Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.

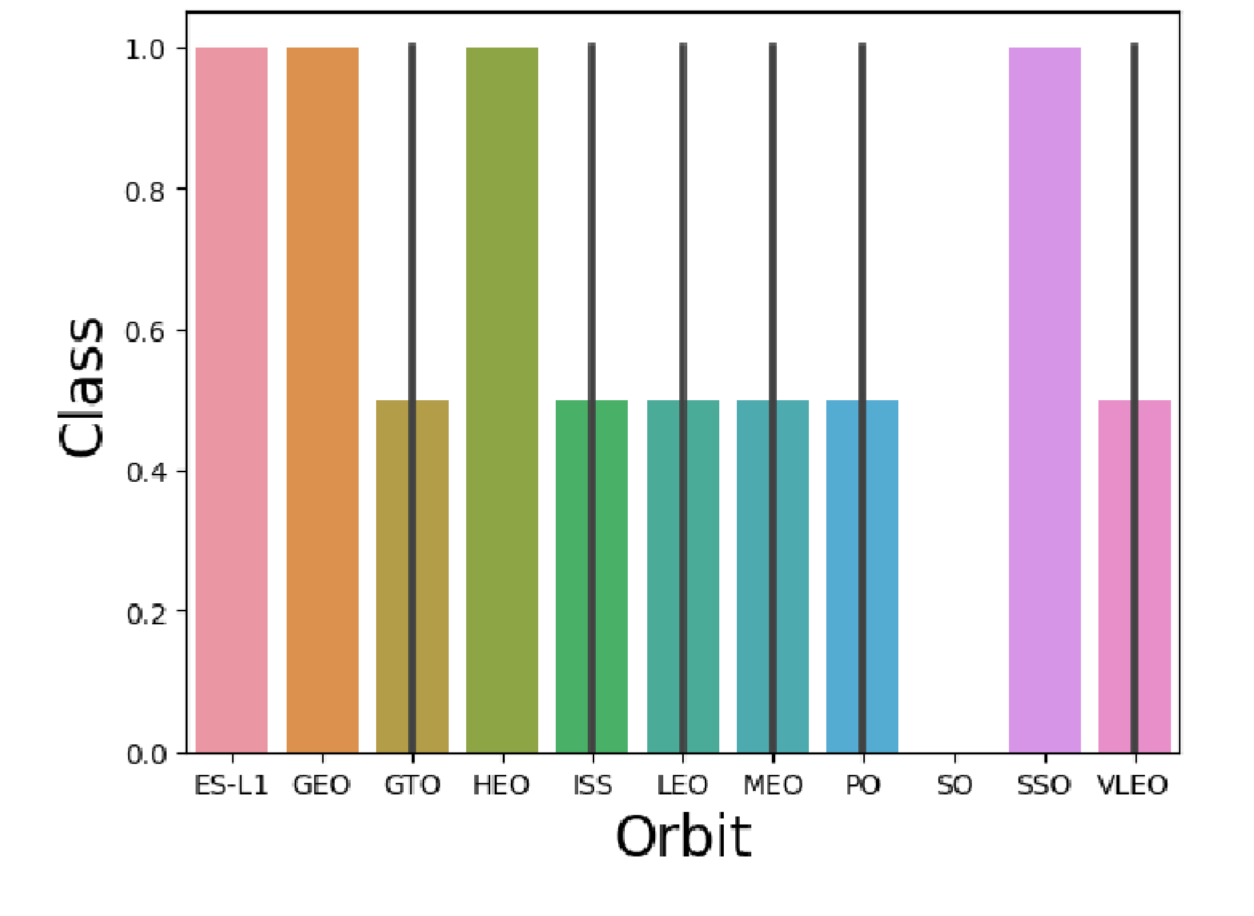


Payload vs. Launch Site



For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)

Success Rate vs. Orbit Type



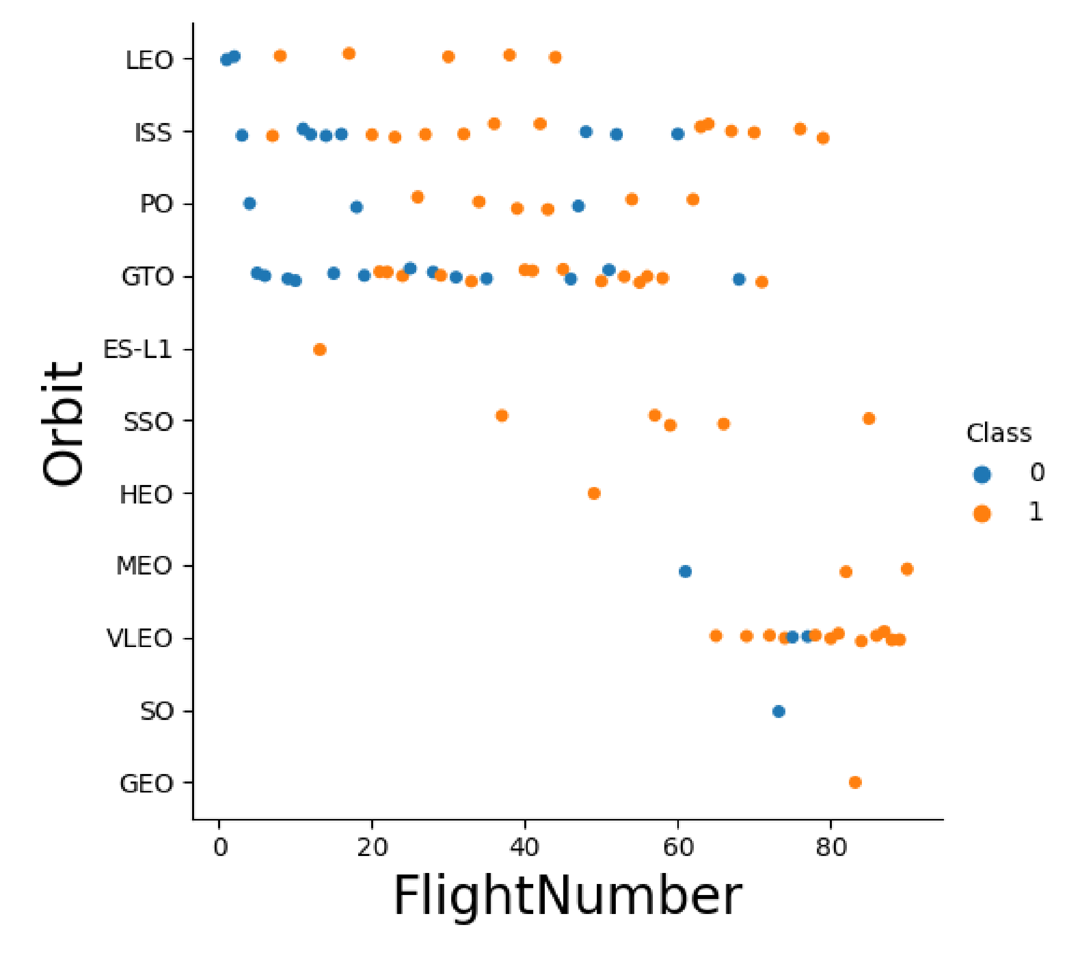
Most success rate

- ES-L1,
- GEO,
- HEO
- SSO,

Flight Number vs. Orbit Type

In the **LEO** orbit the Success appears related to the number of flights;

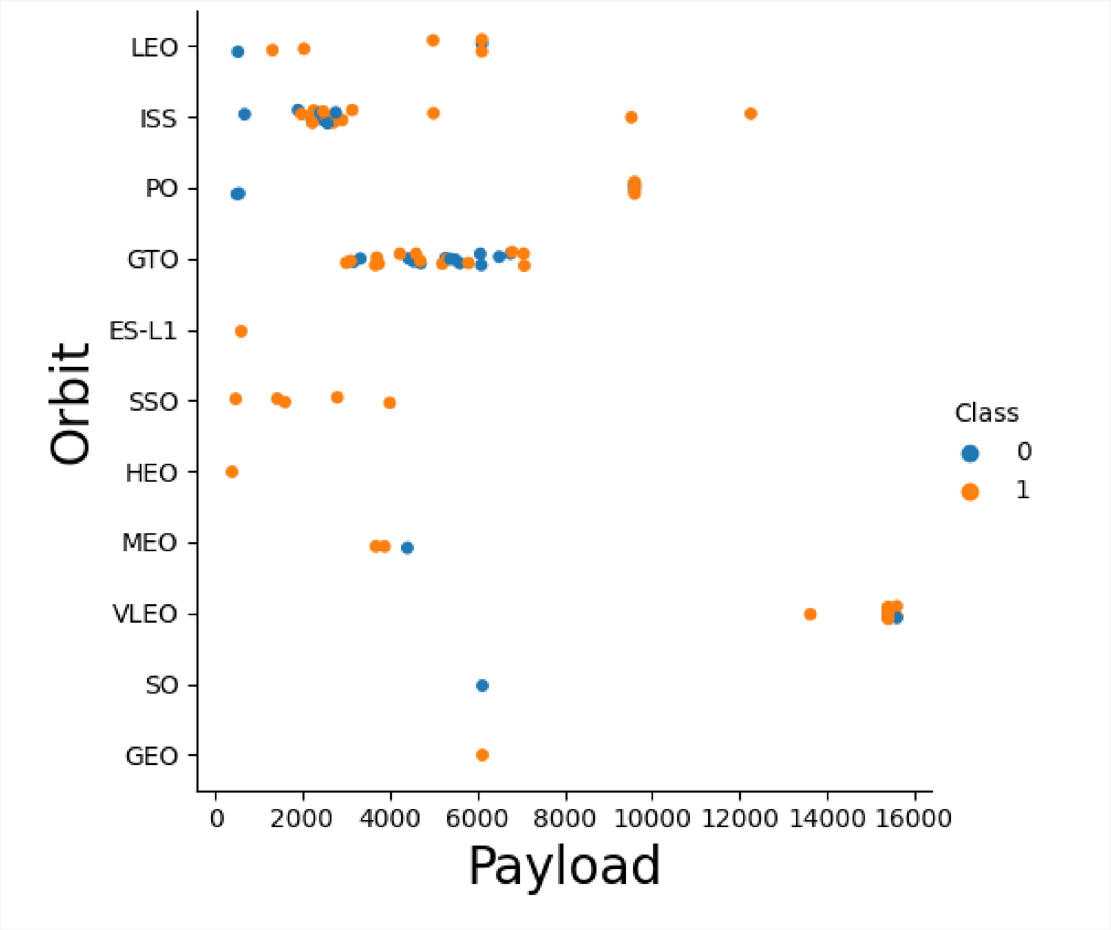
on the other hand, there seems to be no relationship between flight number when in **GTO** orbit



Payload vs. Orbit Type

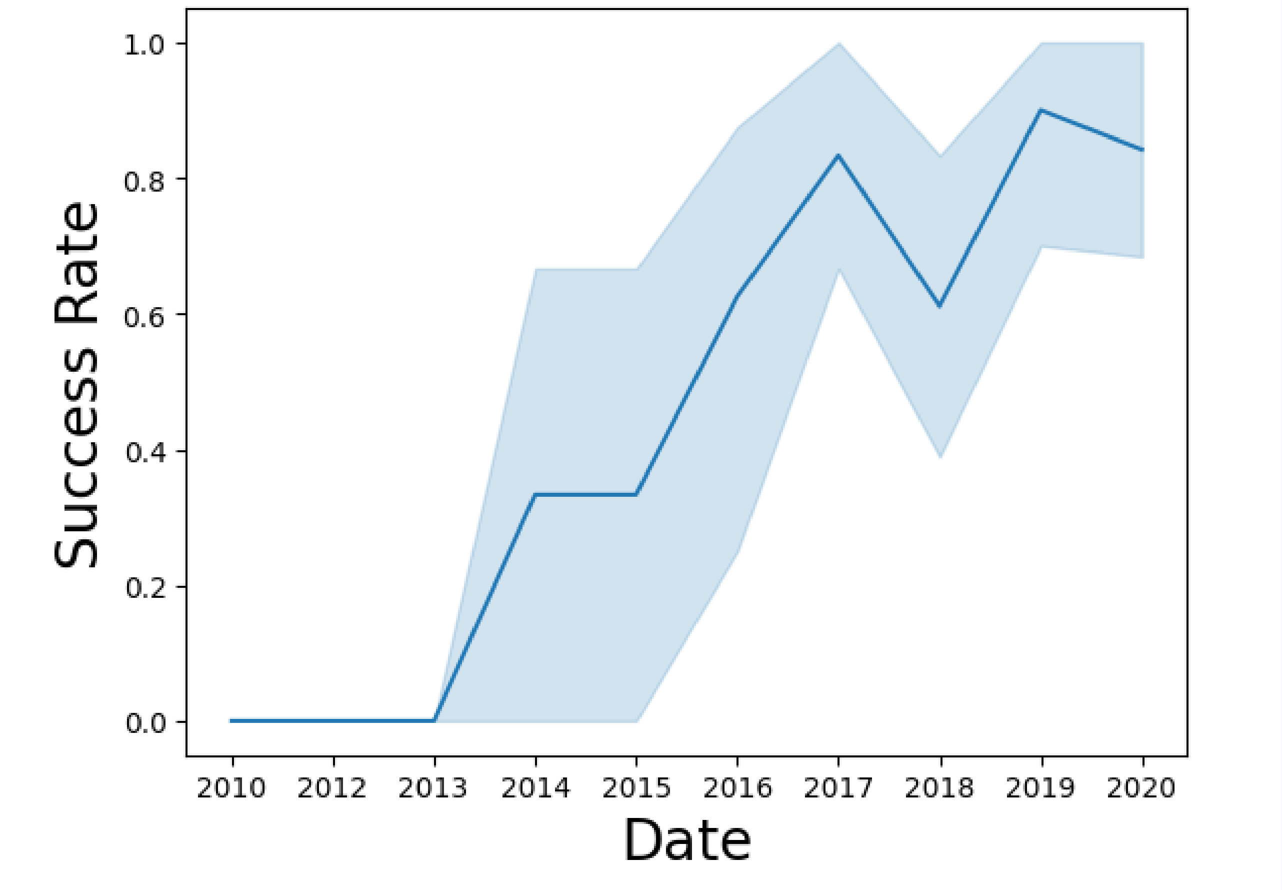
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However 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

The success rate in 2013 kept increasing until 2020



All Launch Site Names

The 4 launch sites are:

- Cape Canaveral Launch Complex 40 (CAFS LC-40)
- Cape Canaveral Space Launch Complex 40 (CCAFS SLC-40)
- Kennedy Space Center Launch Complex 39A (KSC LC-39A)
- Vandenberg Air Force Base Space Launch Complex (VAFB SLC-4E)

```
%sql select DISTINCT(Launch_Site) from SPACEXTBL
```

```
 * 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'

5 launch site names beginning with CCA

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	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 select CUSTOMER, SUM(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE CUSTOMER IN ( 'NASA (CRS)')
```

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

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

The total payload Mass carried by NASA(CRS) – 45,596KG

Average Payload Mass by F9 v1.1

```
%sql select booster_version, avg(payload_mass__kg_) from spacextbl where booster_version like 'F9 v1.1%'
```

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

Booster_Version	avg(payload_mass__kg_)
F9 v1.1 B1003	2534.6666666666665

- The average payload mass carried by booster version F9 V1.1 – 2,534.66kg

First Successful Ground Landing Date

```
%sql select min(date), launch_site, booster_version from spacextbl where [Landing _Outcome] = 'Success (ground pad)'
```

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

min(date)	Launch_Site	Booster_Version
01-05-2017	KSC LC-39A	F9 FT B1032.1

The first successful landing outcome in ground pad was achieved om 01-05-2017 with booster version F9 FT B1032.1 at launch site KSC LC-39A

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from spacextbl where [landing _outcome] like 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000
```

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

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

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:	
	F9 FT B1022
	F9 FT B1026
	F9 FT B1021.2
	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(mission_outcome) as total from spacextbl group by mission_outcome
```

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

Mission_Outcome	total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Total number of successful and failure mission outcomes

Mission Outcome	Total
• Failure (in flight)	1
• Success	98
• Success	1
• Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%sql select booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

The booster which have carried the maximum payload mass – 15,600KG

2015 Launch Records

```
%sql select substr(Date, 4, 2) as Month, [landing _outcome], booster_version, launch_site from spacextbl where substr(Date, 7, 4) = '2015' and [landing _outcome] like 'Failure (drone ship)%'
```

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

Month	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

Month	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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))

In [19]:
```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

Out[19]:

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

No attempt was highest for the period

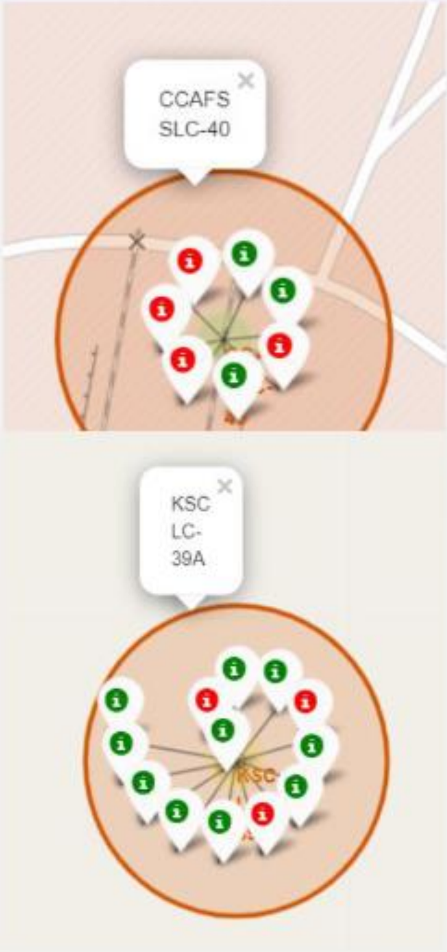
Section 3

Launch Sites Proximities Analysis

Global Map of Launch Sites

All landing Sites are located in the United States of America

Close –up with Launch sites

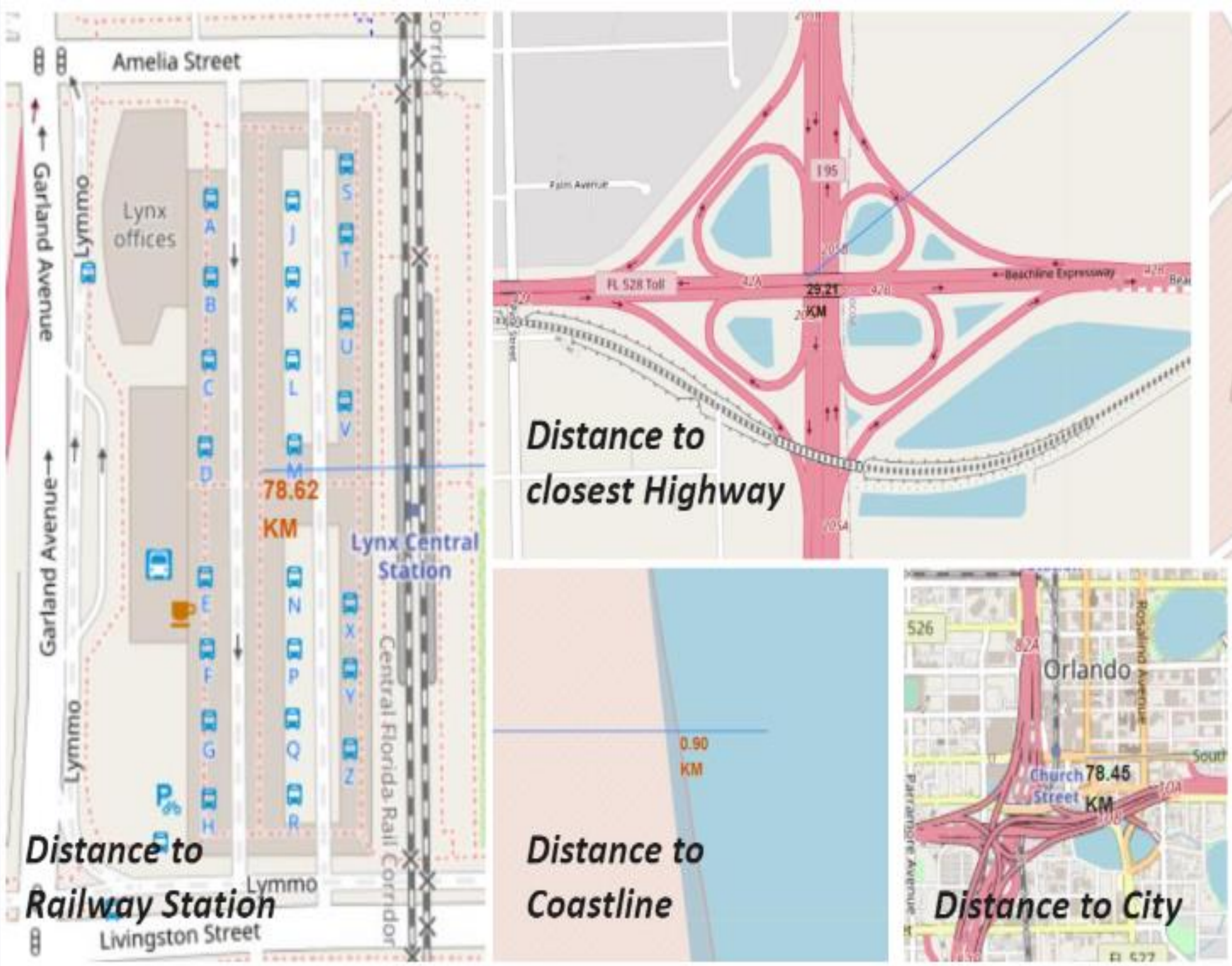


Florida Launch Sites

Green Marker shows successful Launches and Red Marker shows Failures

California Launch Site

Proximities to Launch Sites

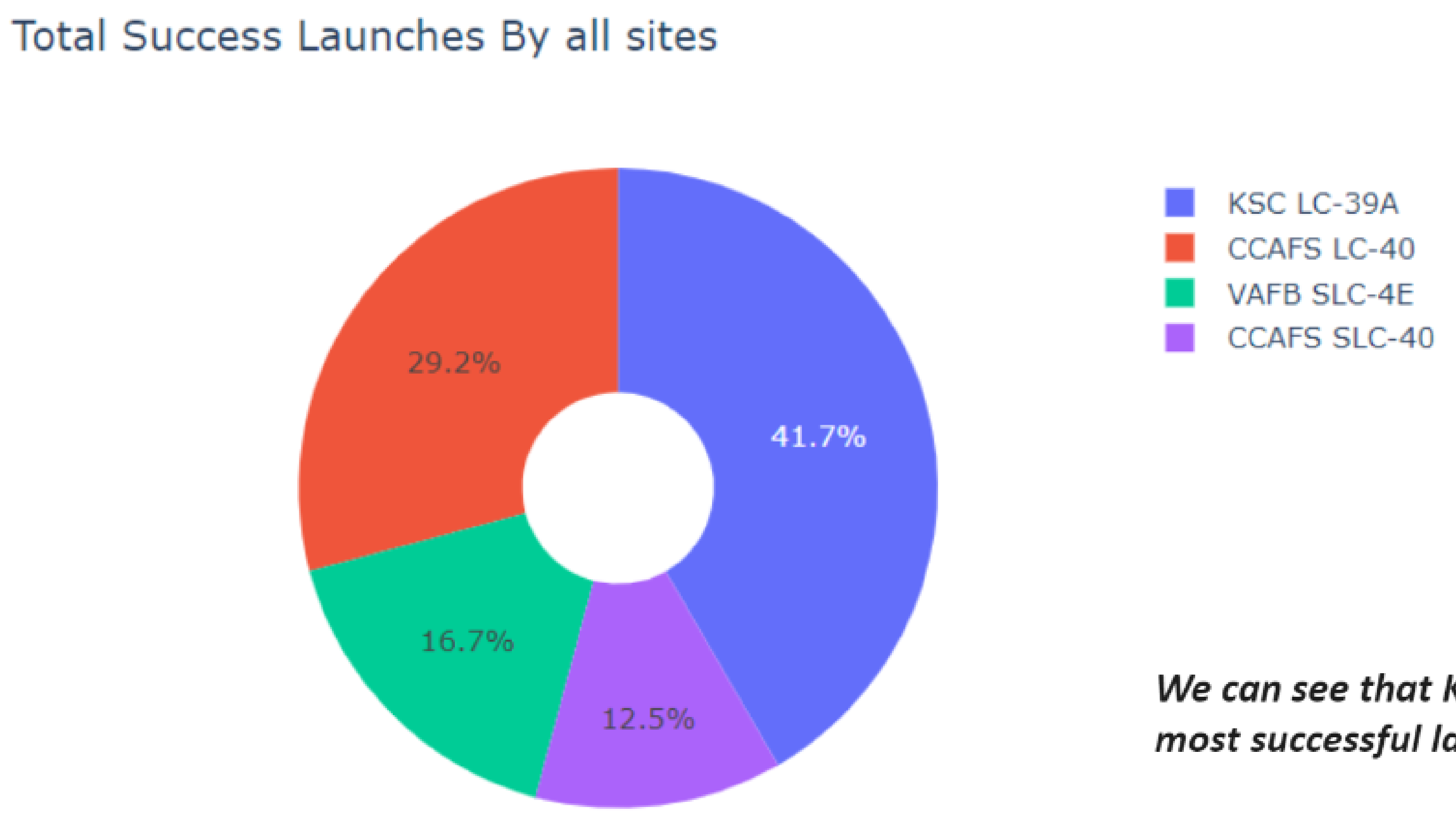


•Are launch sites in close proximity to railways? No
•Are launch sites in close proximity to highways? No
•Are launch sites in close proximity to coastline? Yes
•Do launch sites keep certain distance away from cities? Yes

Section 4

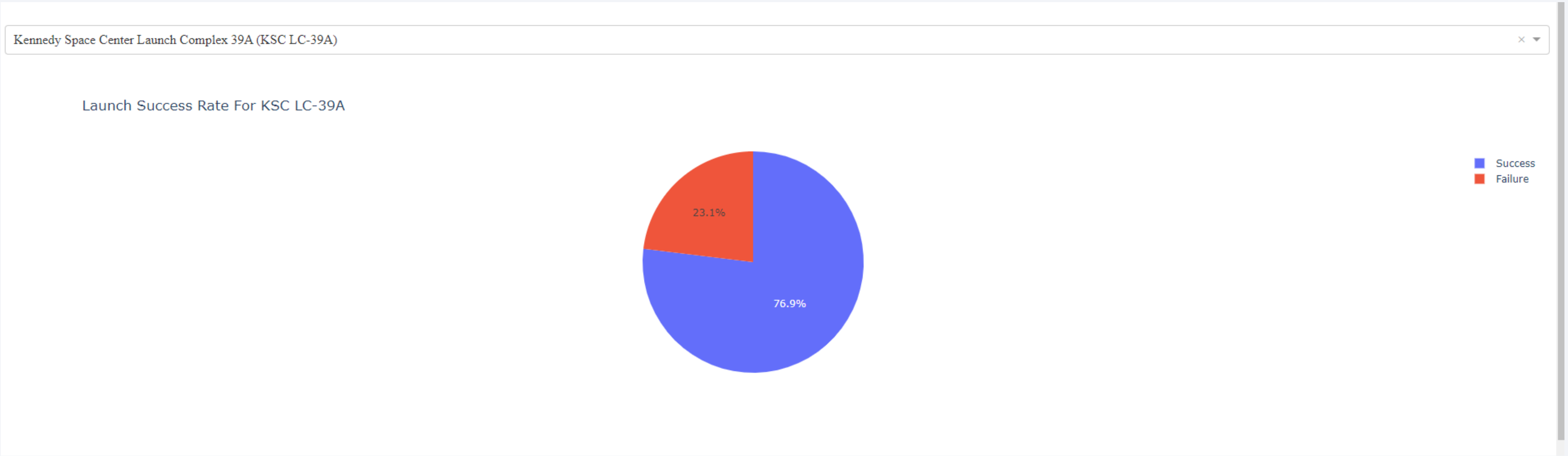
Build a Dashboard with Plotly Dash

Success of Launches



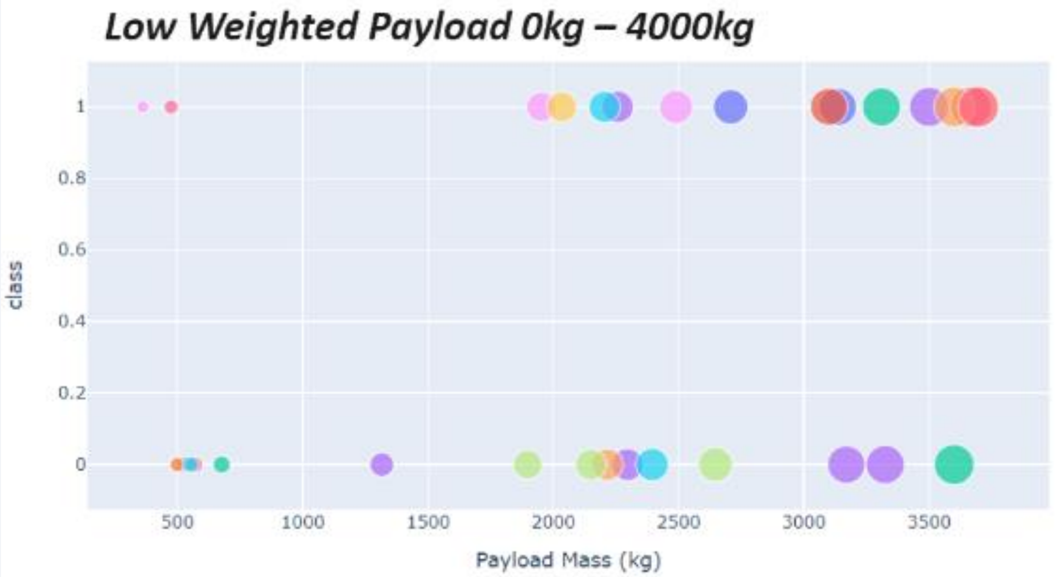
We can see that KSC LC-39A had the most successful launches from all the sites

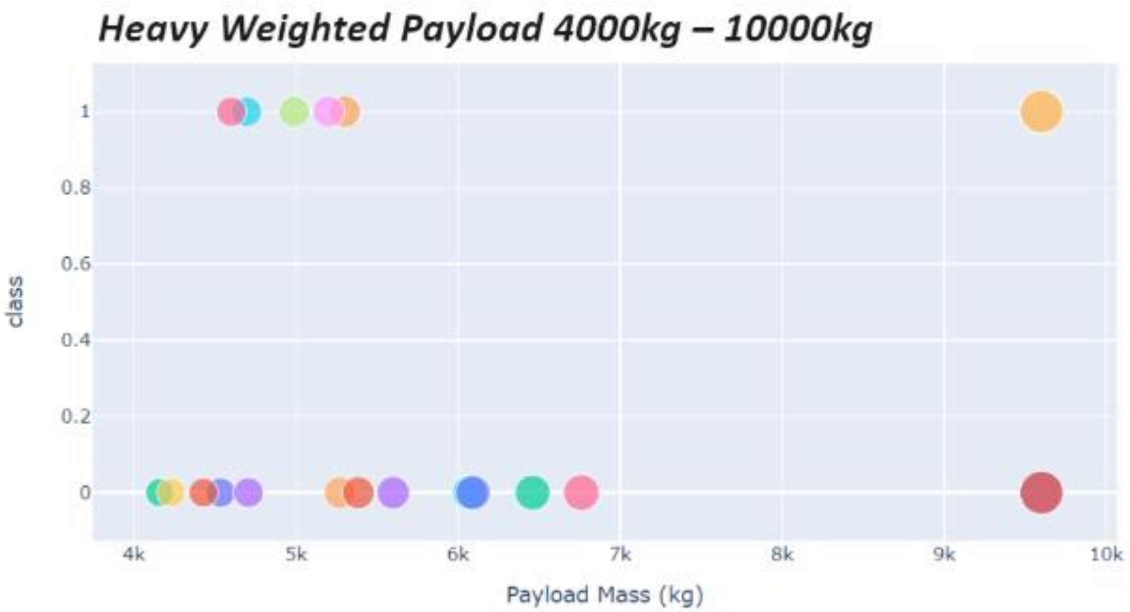
Success for KSC LC-39A



Success rate is 76.9%

Payload Vs Launch Outcome for all sites



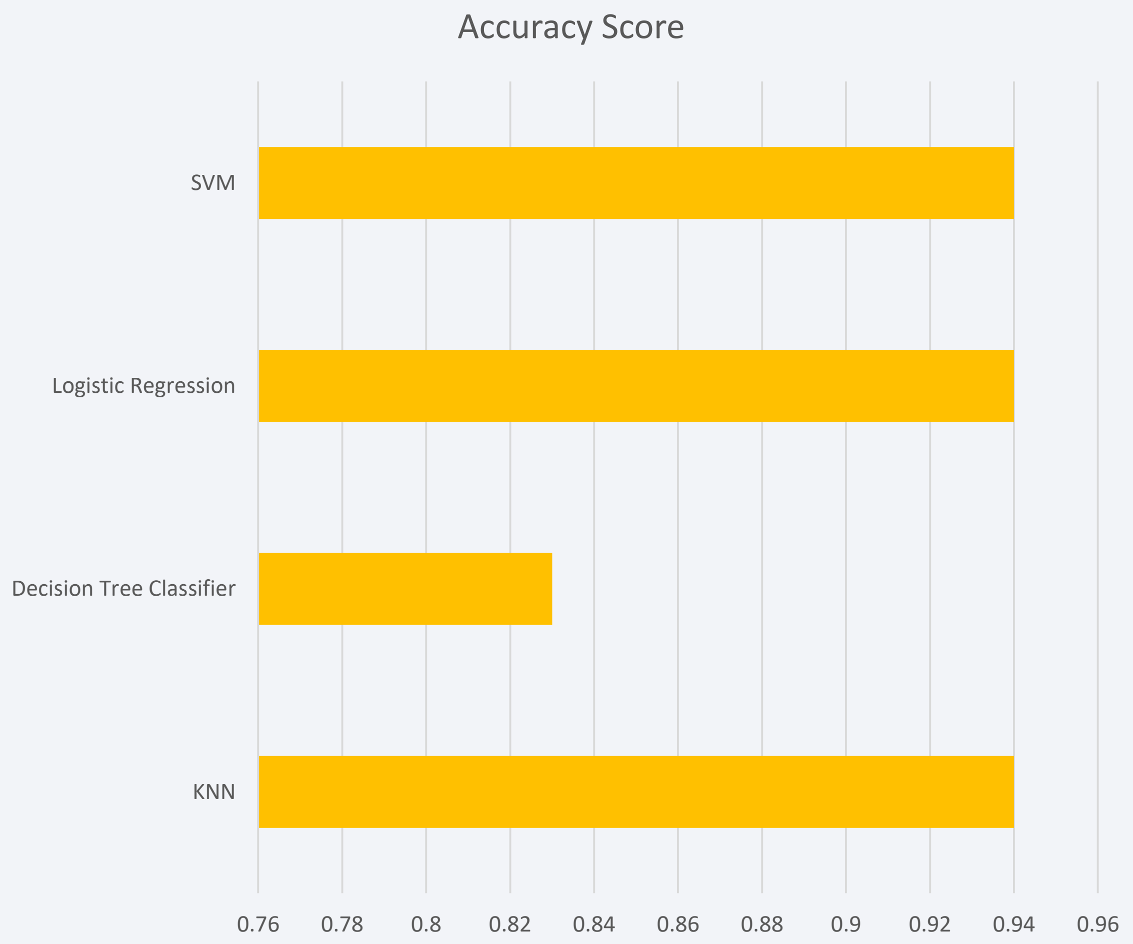


We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

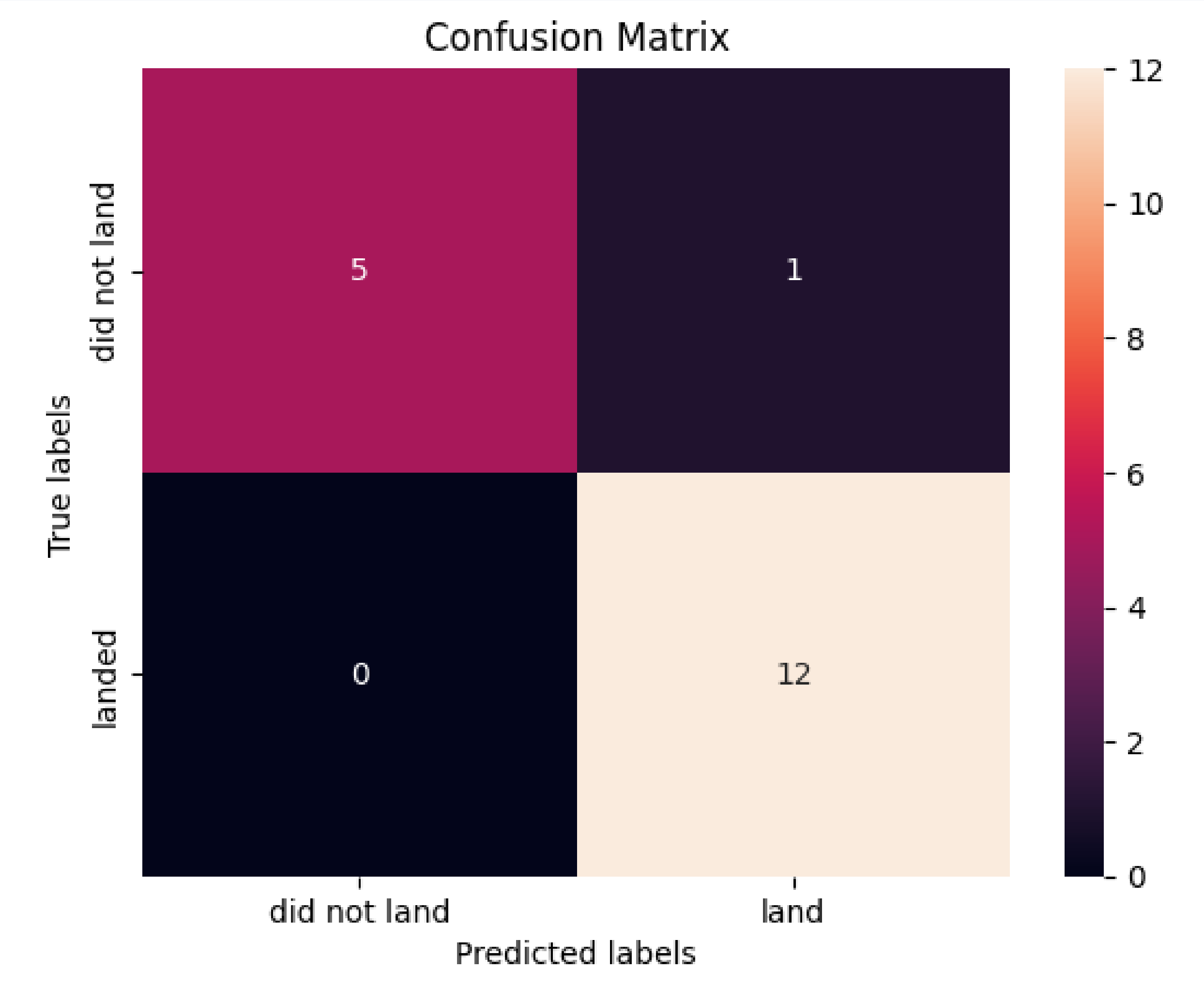
Classification Accuracy



Three models have the highest accuracy of 0.94;

- Support Vector Machine Classifier
- Logistic Regression
- K-Nearest Neighbour

Confusion Matrix



The false positive figure is one

Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.

Thank you!