Winning Space Race with Data Science

Henry Fistgerald Rojas Monteza
04/27/2022

OUTLINE

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

Executive Summary

Summary of Methodologies

The project has followed the next steps:

1. Data Collection using the API from SpaceX and Wikipedia.
2. Preprocessing and cleaning data.
3. Data Exploration using SQL, visualization, maps and interactive visualization.
4. find best parameters for machine learning models
5. Visualize accuracy score of all models.

Summary of Results

We have gotten exploratory data visualizations and four machine learning models with average accuracy of 83.33%

INTRODUCTION

Launching a rocket cost approximately $165m therefore, SpaceX has been launches Falcon 9 rockets at a cost of around $62m and much of the savings are because SpaceX can land, and then re-use the first stage of the rocket.

Apace Y tasks us to train a machine learning model to predict successful Stage 1 recovery.

Section 1

Methodology

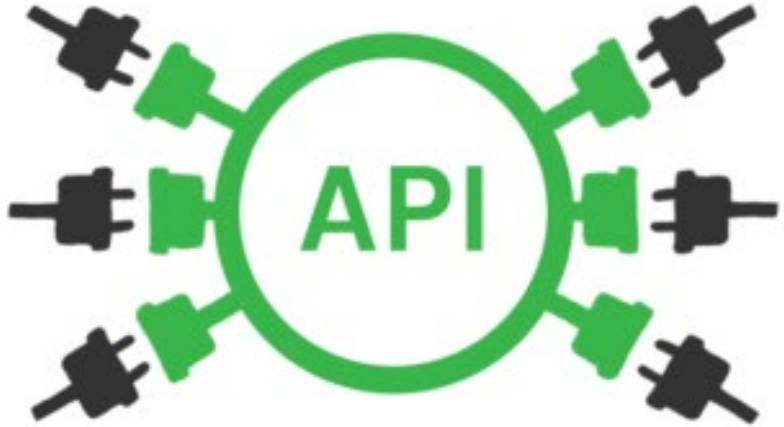
Methodology

Executive Summary

- Data collection methodology:
  - The data was got from SpaceX public API and using web scraping.
- Perform data wrangling
  - Classifying landings as successful and unsuccessful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

DATA COLLECTION

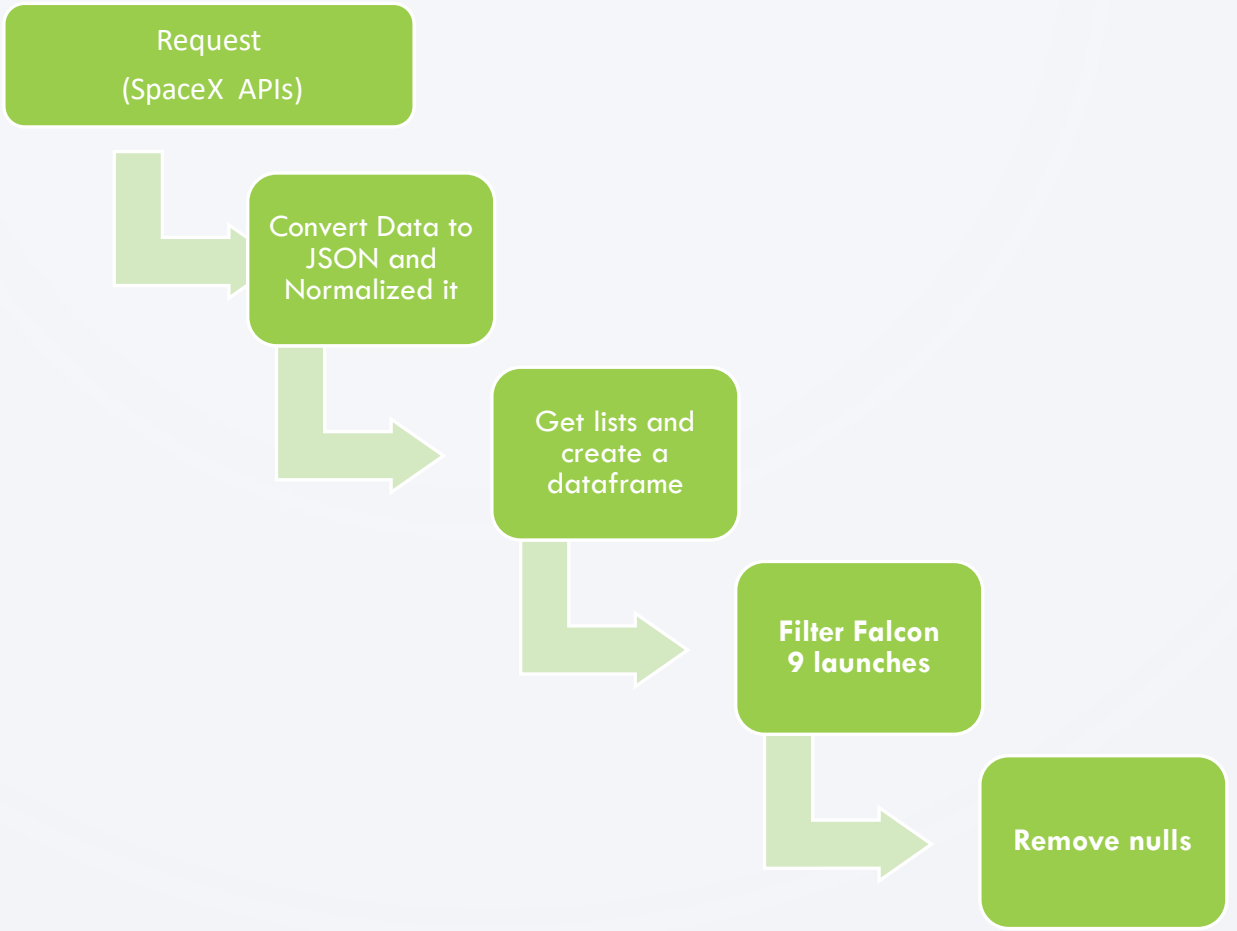
- The data was got making requests from Space X public API and making web scraping from a table on Wikipedia.
- The process was made using python and pandas.



SPACEX

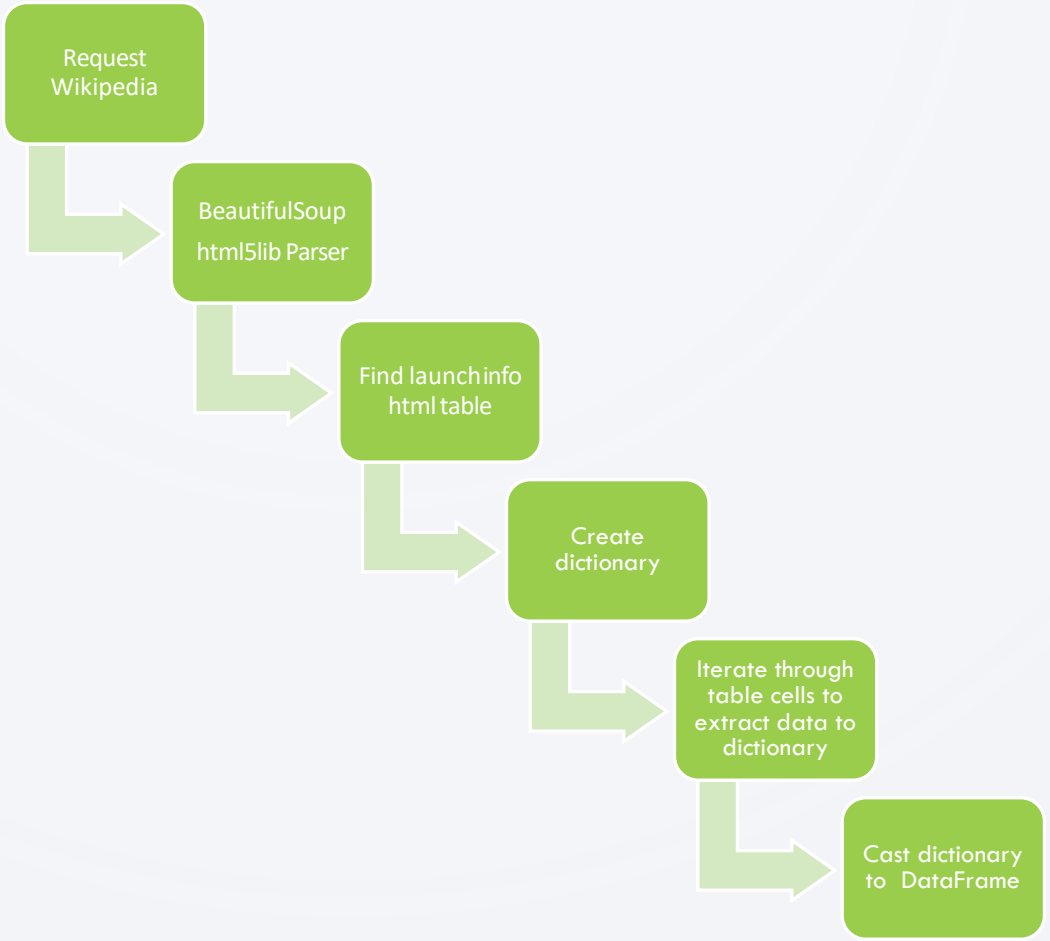
Data Collection – SpaceX API

- GitHub Repository



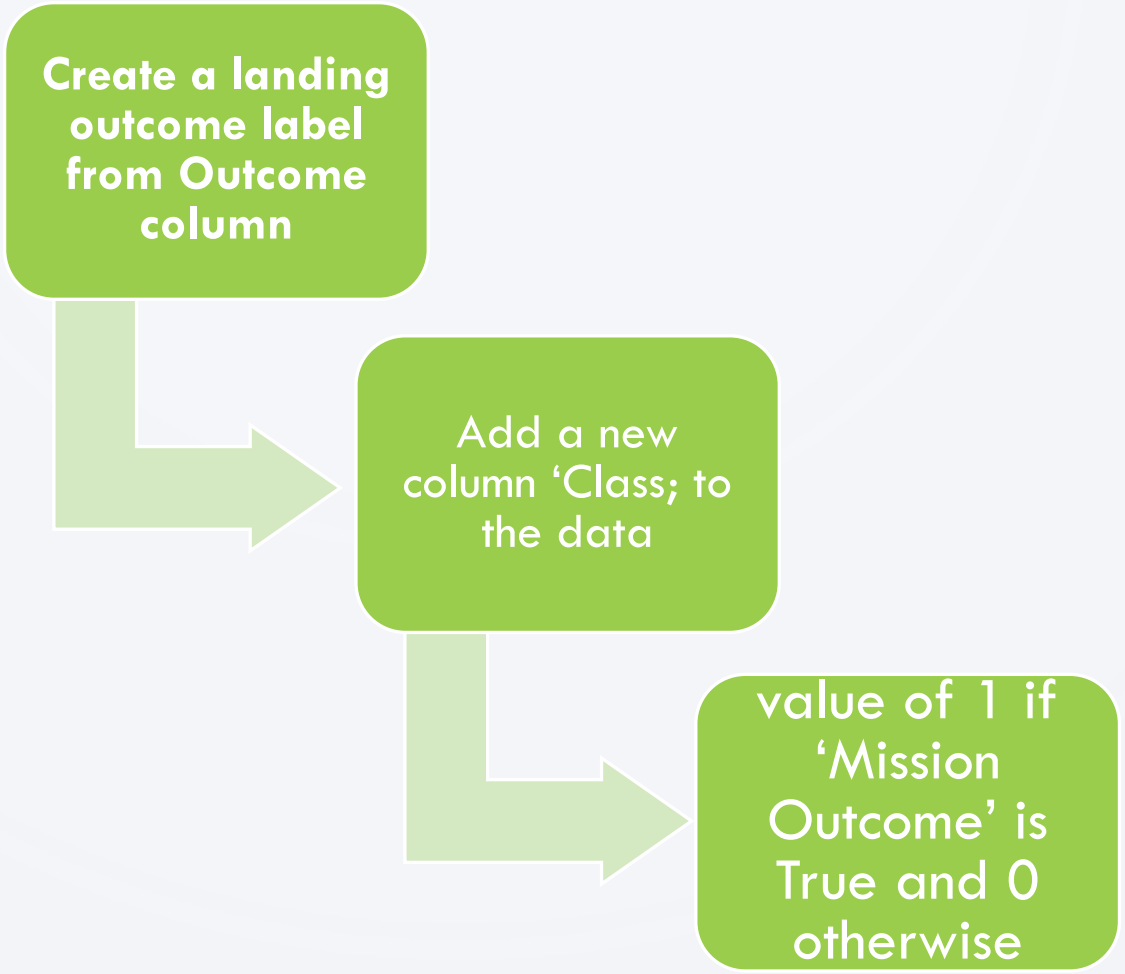
Data Collection - Scraping

- GitHub Repository



Data Wrangling

- GitHub repository



EDA with Data Visualization

- GitHub Repository
- **Scatter chart:** This kind of chat was used to compare Flight Number and Launch Site, Payload and Launch Site, Orbit Type and Flight Number and Payload and Orbit Type
- **Bar Chart:** This kind of chart was used to visualize the relationship betwen Success Rate and Orbit Type.
- **Line Chart:** This kind of chat was used to visualize the relationship betwen Success Rate and Year

EDA with SQL

- GitHub Repository

- 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 the average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome on a ground pad was achieved
- List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
- List the total number of successful and failed mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015
- 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

Build a Dashboard with Plotly Dash

- GitHub Repository

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and
- booster version category.

Predictive Analysis (Classification)

GitHub Repository

Get and clean data

Transform features using standard scaler

train data

Score models on split test set

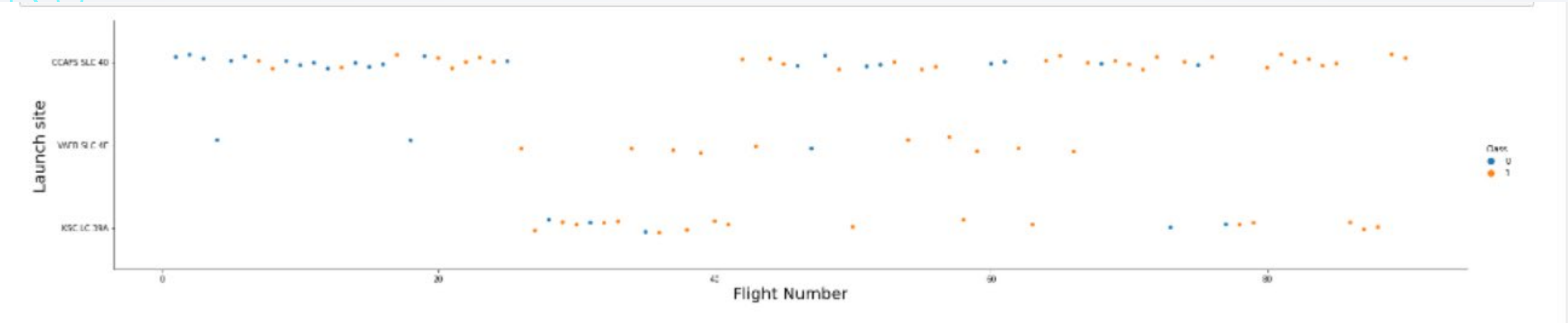
Use GridSearchCV on all the models

Use Confusion Matrix for all models

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

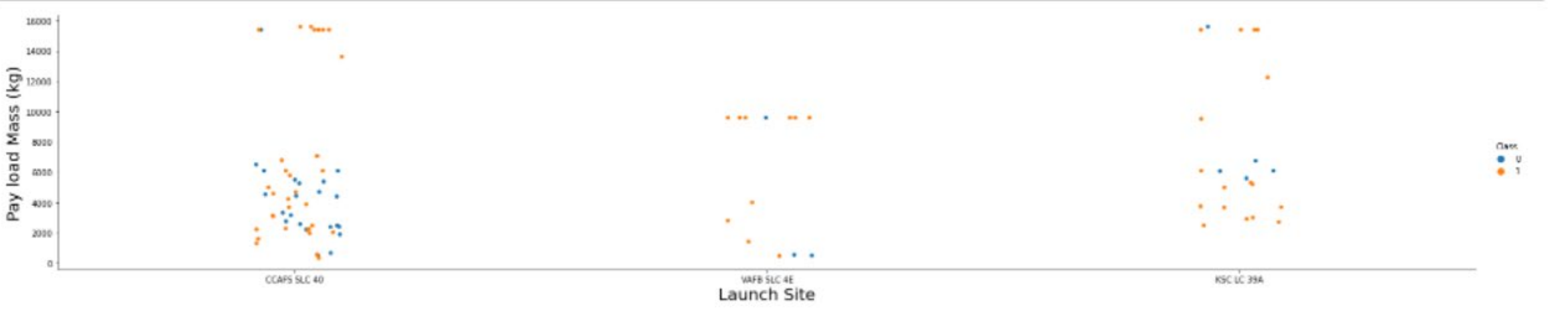


Graphic suggests an increase in success rate over time (indicated in Flight Number).

Likely a big breakthrough around flight 20 which significantly increased success rate.

CCAFS appears to be the main launch site as it has the most volume.

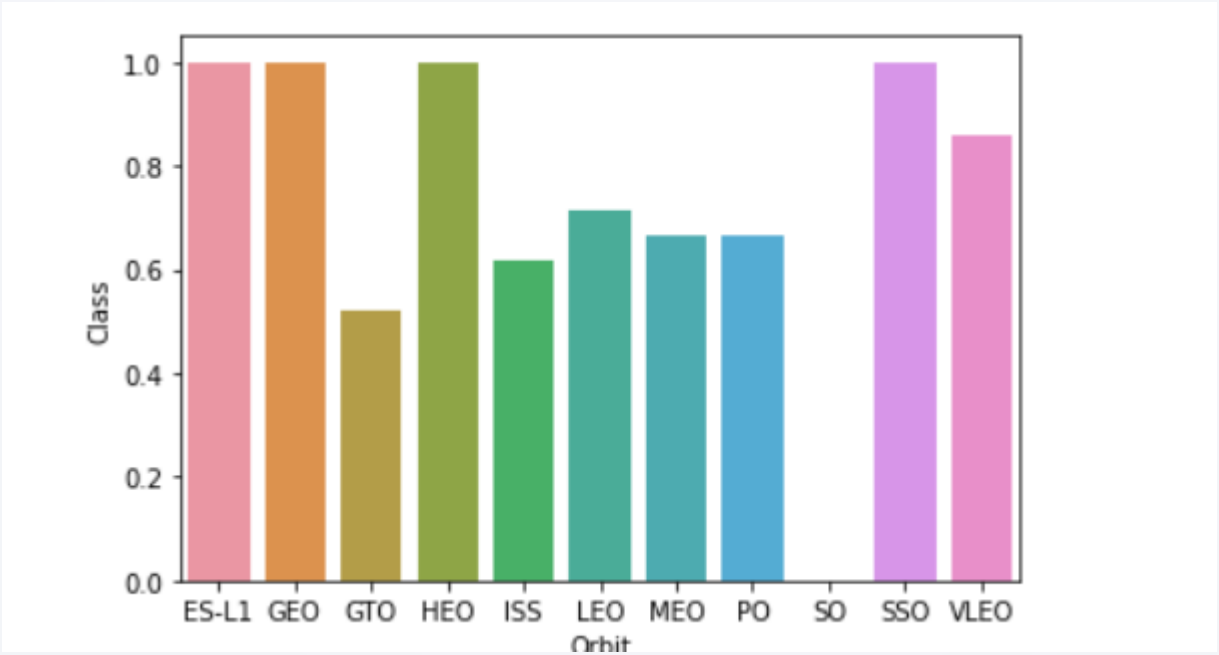
Payload vs. Launch Site



Payload mass appears to fall mostly between 0-6000 kg.

Different launch sites also seem to use different payload mass.

Success Rate vs. Orbit Type



ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)

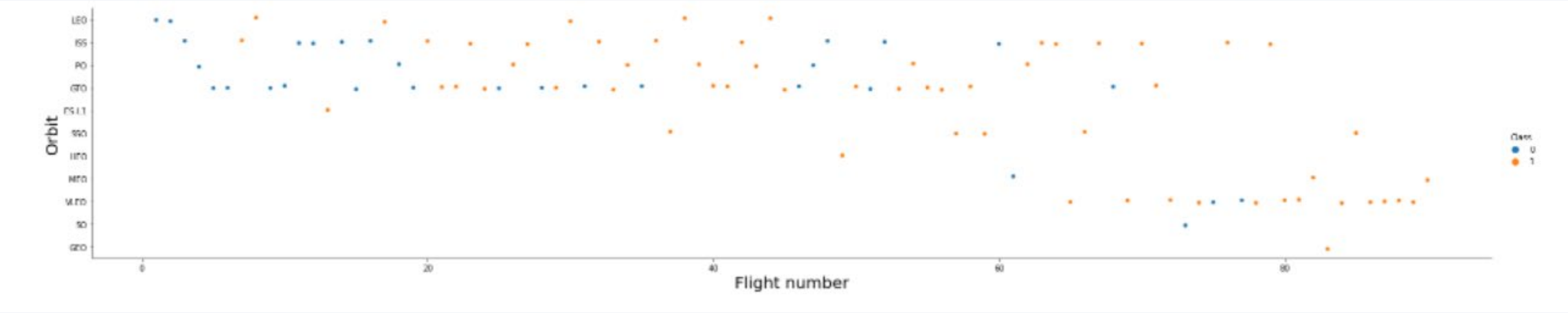
SSO (5) has 100% success rate

VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate but largest sample

Flight Number vs. Orbit Type



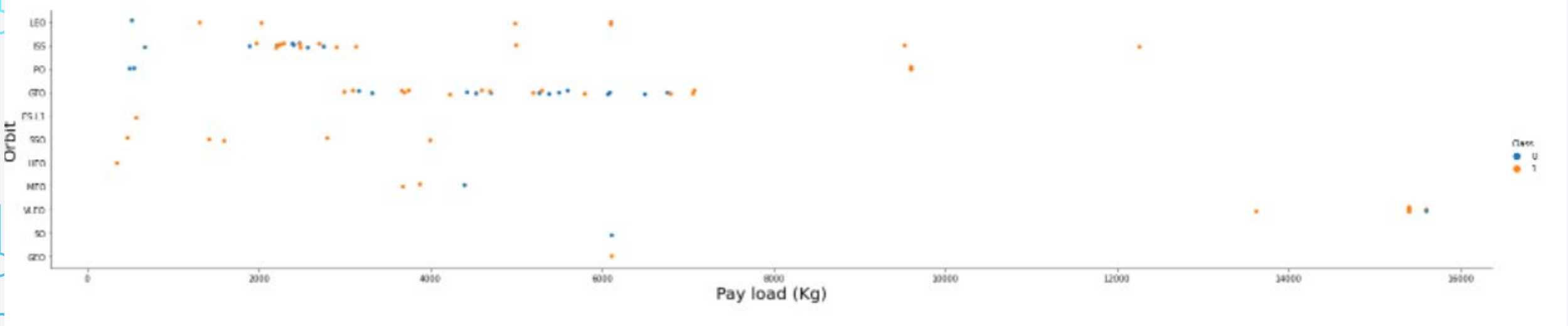
Launch Orbit preferences changed over Flight Number.

Launch Outcome seems to correlate with this preference.

SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches

SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type



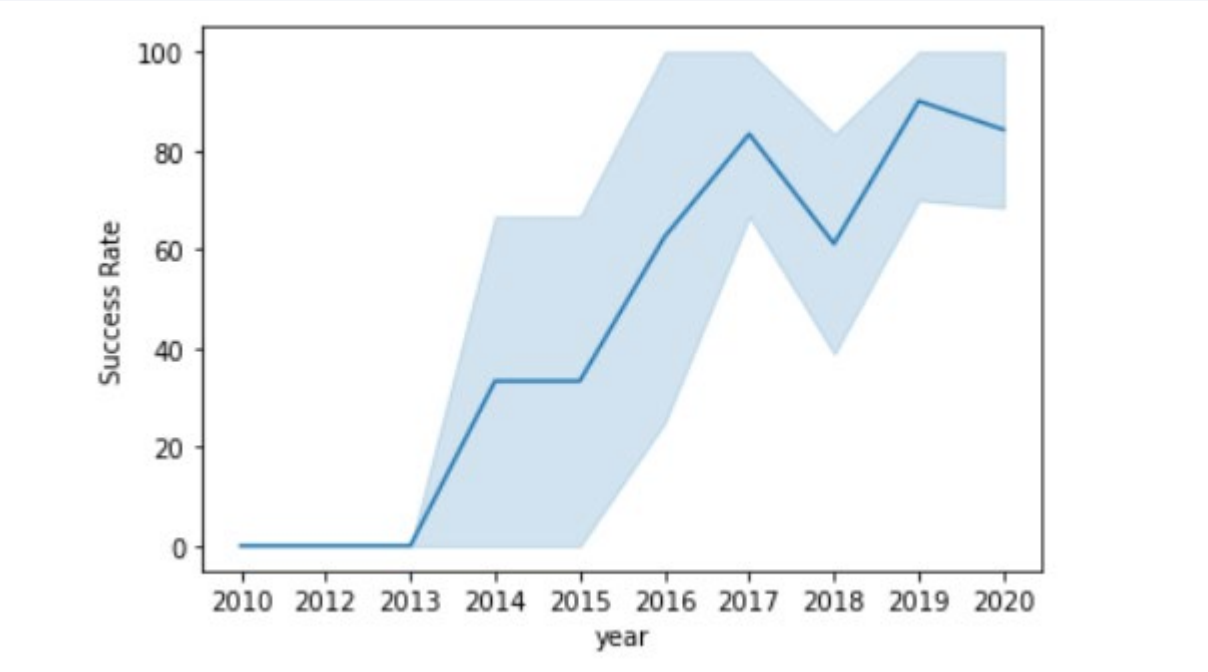
Launch Orbit preferences changed over Flight Number.

Launch Outcome seems to correlate with this preference.

SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches

SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Launch Success Yearly Trend



Success generally increases over time since 2013 with a slight dip in 2018

Success in recent years at around 80%

All Launch Site Names

Display the names of the unique launch sites in the space mission

In [6]:
```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
```

```
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffa
Done.
```

Out[6]:

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

Launch Site Names Begin with 'CCA'

```
]: %sql select LAUNCH_SITE from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.
Done.
```

]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
]: %sql select SUM(payload_mass__kg_) as total from SPACEXDATASET where customer = 'NASA (CRS)'

 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.app
Done.
```

]:

total
45596

Average Payload Mass by F9 v1.1

```
15]: %sql select AVG(payload_mass__kg_) as average from SPACEXDATASET where booster_version like 'F9 v1.1%'
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.clo
Done.
```

15]:

average
2534

First Successful Ground Landing Date

```
In [18]: %sql select min(DATE) as date from SPACEXDATASET where mission_outcome='Success'
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98
Done.
```

Out[18]:

DATE
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

In [21]:
```
%sql select unique(booster_version) from SPACEXDATASET where payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000 and landing__
outcome = 'Success (drone ship)'
```

```
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

Out[21]:

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

Total Number of Successful and Failure Mission Outcomes

```
]: %sql select mission_outcome, count(mission_outcome) as total from SPACEXDATASET GROUP by mission_outcome
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30
Done.
```

]:

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

Boosters Carried Maximum Payload

```
maxm = %sql select max(payload_mass__kg_) from SPACEXDATASET
maxv = maxm[0][0]
%sql select booster_version from SPACEXDATASET where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/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

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site  from SPACEXDATASET where DATE like '2015%'
AND landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%sql select landing__outcome, count(*) as count from SPACEXDATASET where Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP by landing__outcome ORDER BY count Desc
```

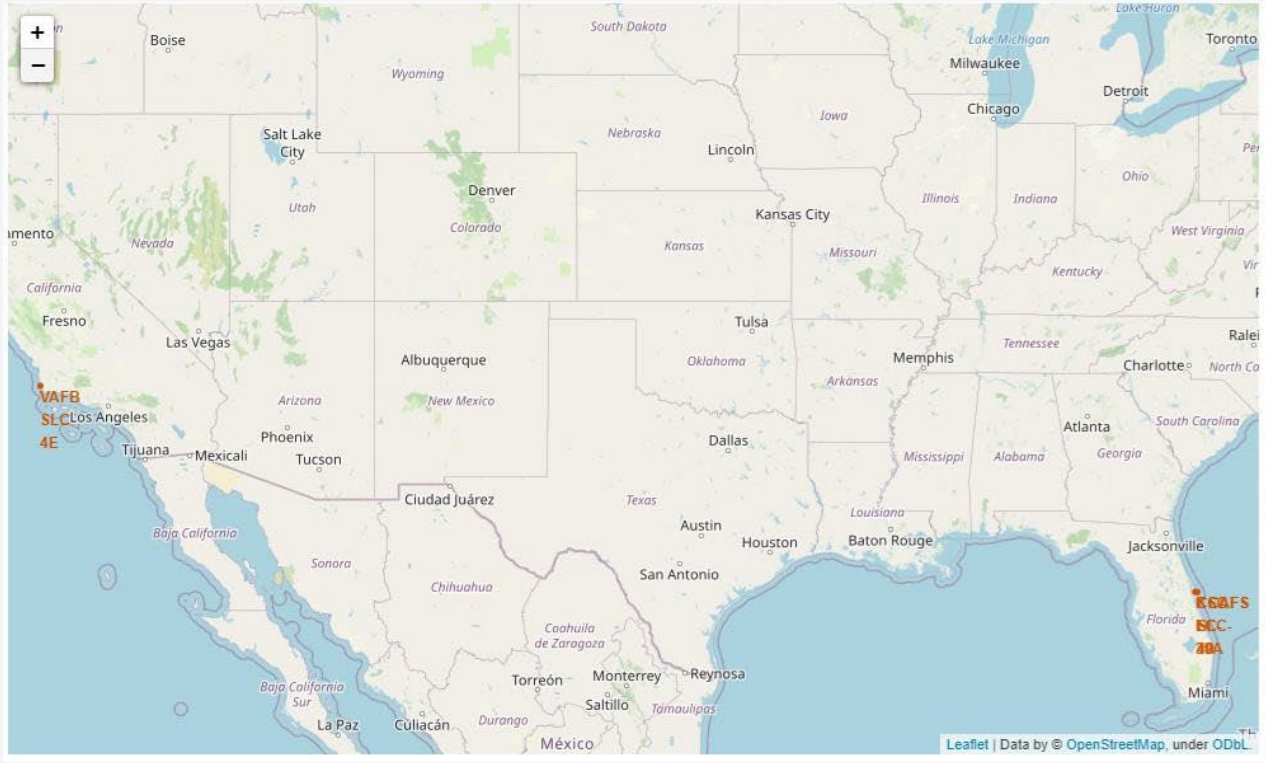
```
 * ibm_db_sa://gdb80748:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

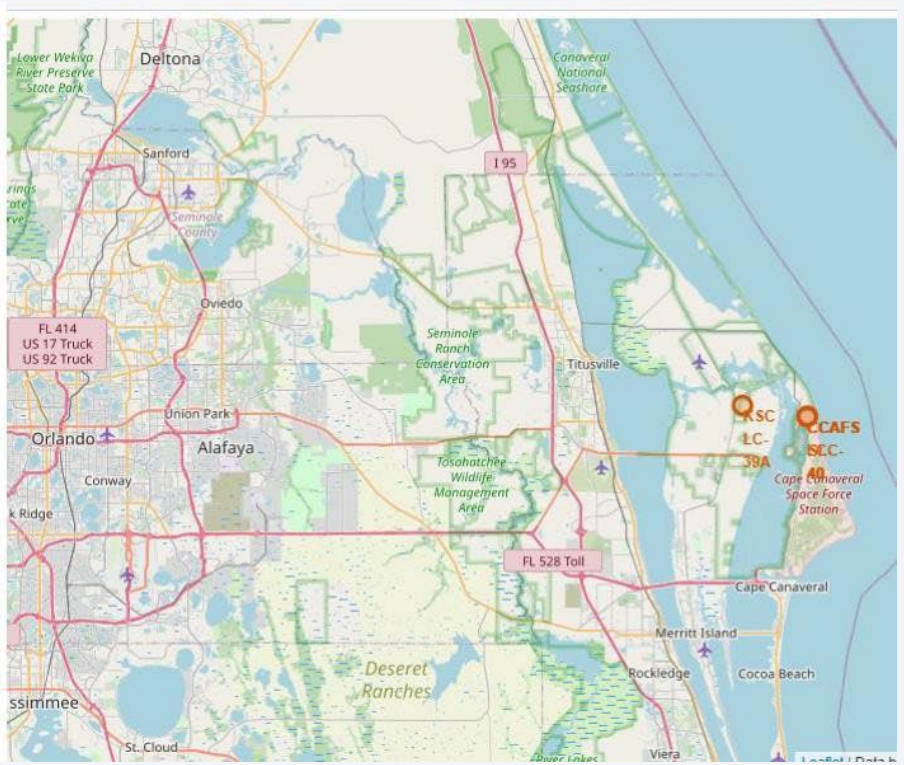
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

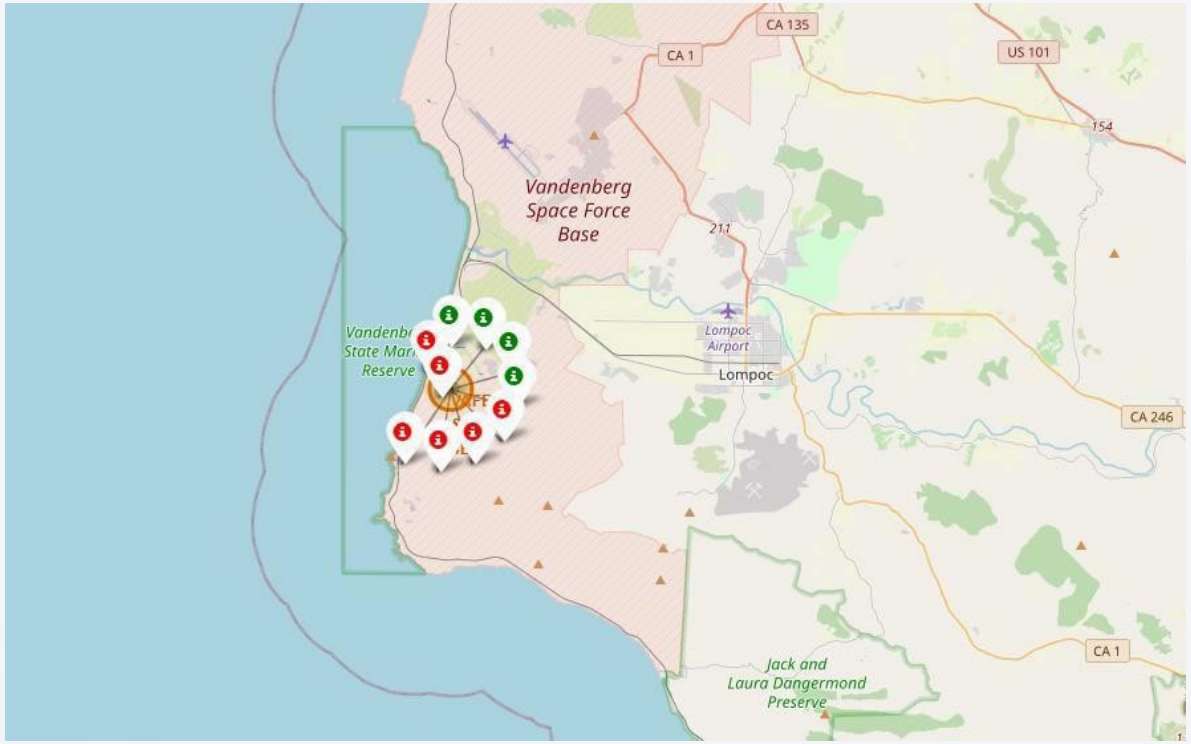
Launch Site Locations





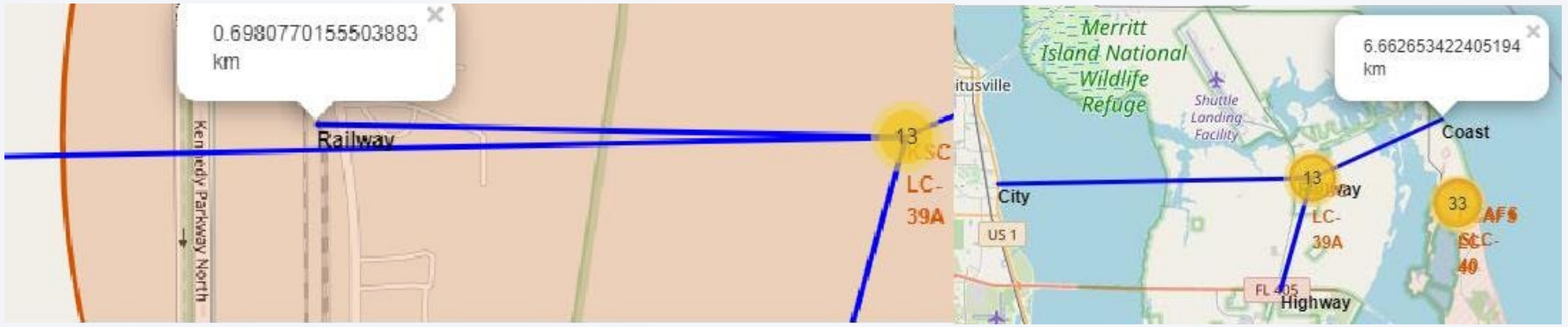
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color-Coded Launch Markers



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

Key Location Proximities

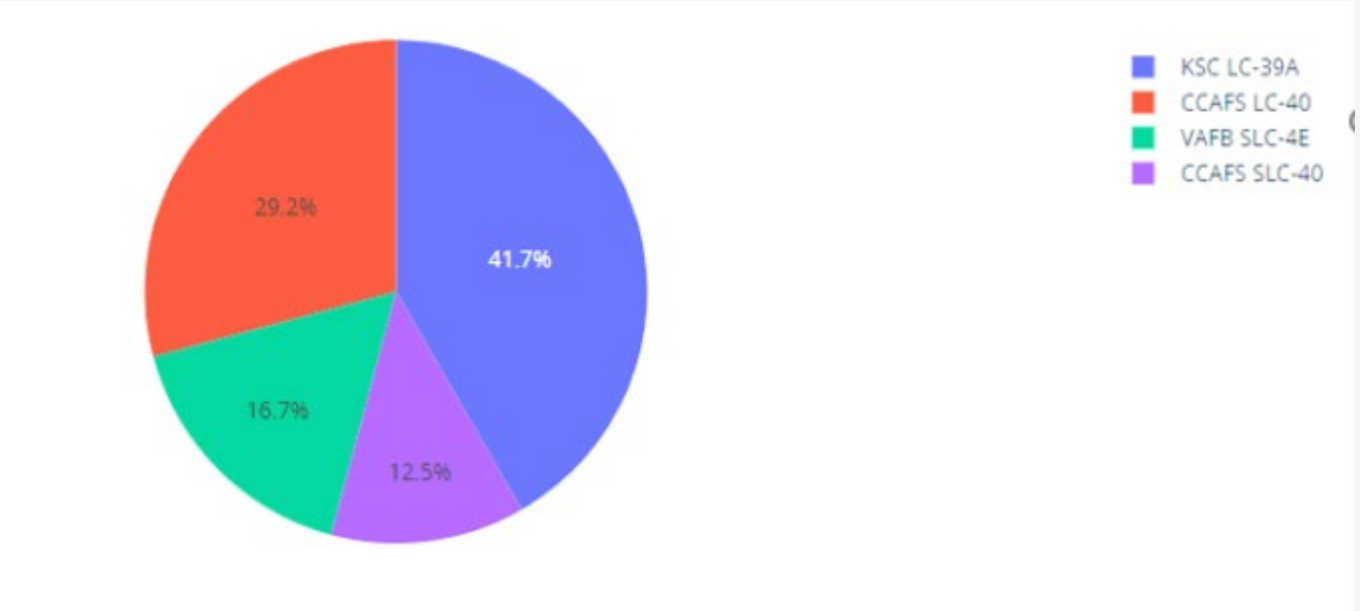


Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

Section 4

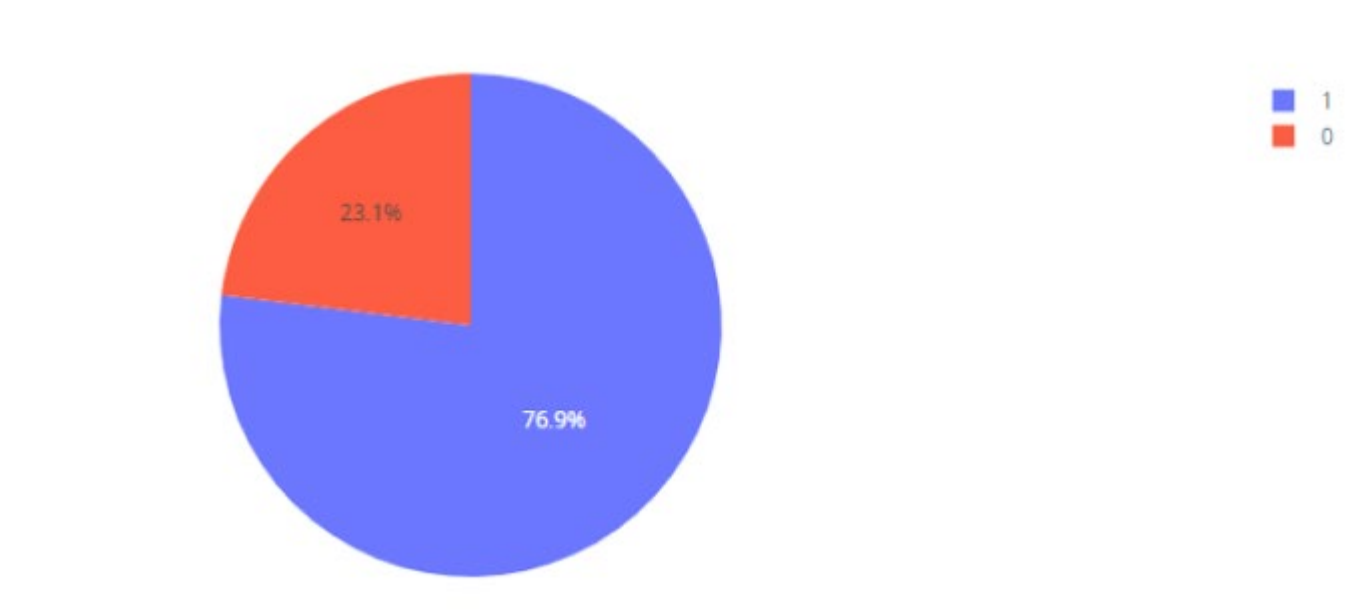
Build a Dashboard with Plotly Dash

Launch success count for all sites



The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches.

Highest Success Rate Launch Site

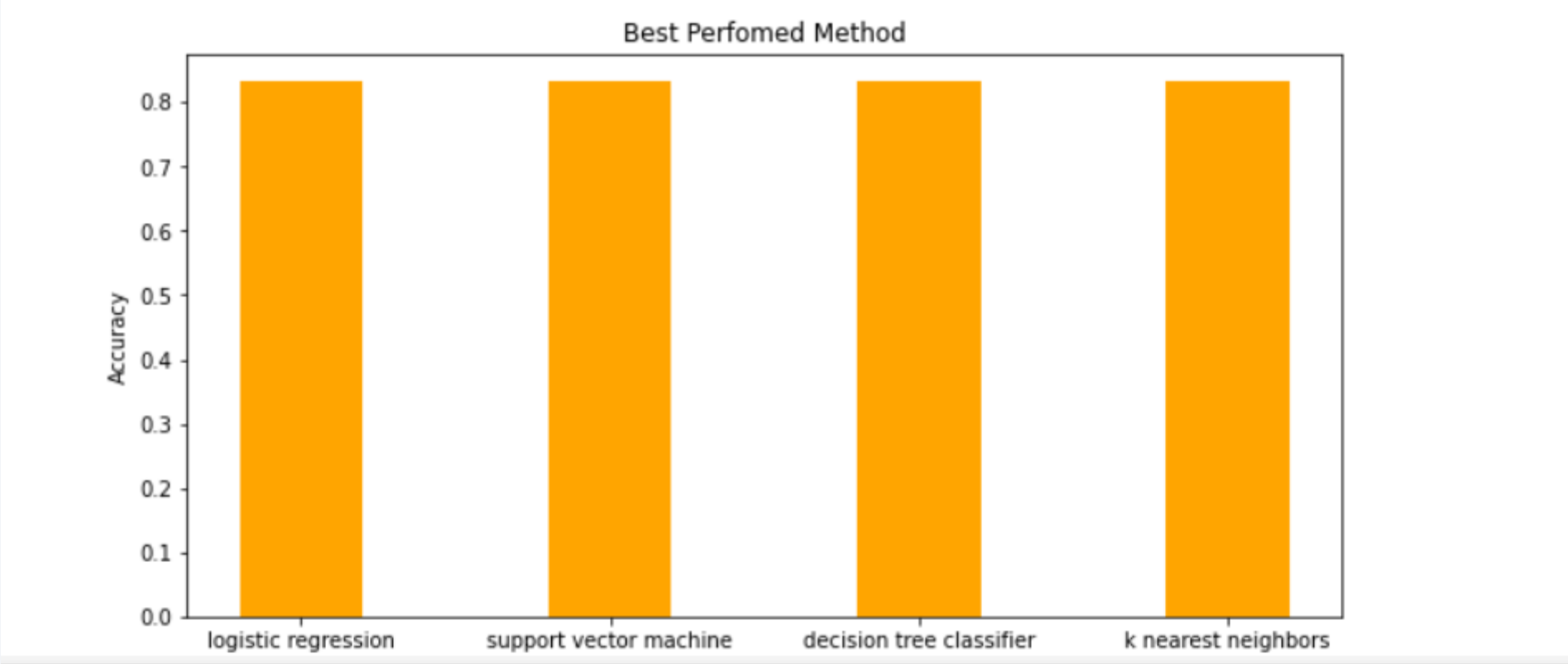


The launch site KSC LC-39 A had the highest rate of successful launches, with a 76.9% success rate.

Section 5

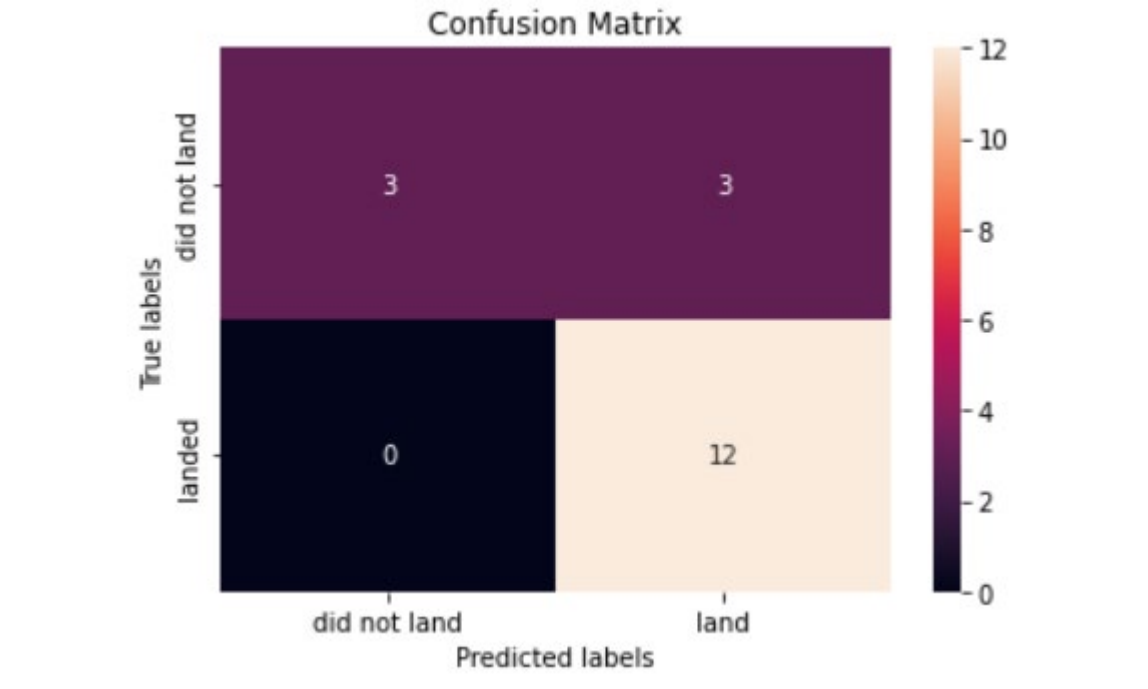
Predictive Analysis (Classification)

Classification Accuracy



All the model obtained best accuracy with 83.33%

Confusion Matrix



- All the models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.

Conclusions

- As the number of flights increases, the rate of success at a launch site increases,
- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.
- The success for massive payloads (over 4000kg) is lower than that for low payloads.
- All the model obtained an accuracy with 83.33%

Appendix

- GitHub Repository

Thank you!