IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Peter Suh

November 2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection - Request and parse Space X launch data using a GET request and filter Data & Web Scraping
  - Data Wrangling - deal with missing values
  - Exploratory Data Analysis (EDA) - calculate the number of launches on each site, and figure the number and occurrence of each orbit; understanding key values and information to create a dataset.
  - EDA - Data Analysis using SQL
  - EDA - Visual data analysis
    - using Pandas and Matpoltlib
    - Using Folium Lab
    - Using Ploty Dash
  - Predictive Data Analysis Using Machine Learning - standardize data and split training data and test data
    - Use a variety of machine learning models to predict landing. Models used are support vector machine, logistic regression, decision tree, and K-nearest neighbor
    - Evaluate the quality of each of the models used
- **Summary of all results**
  - Model Prediction performance - all models performed the same, not one model was better than another
  - Analytics - launch sites similar traits being near on coast
  - EDA
    - KSC LC-39A site has highest success rate
    - Orbits ES_L!, GEO, HEO and SSO have 100% completion success rate

Introduction

- Project background and context
  - Use existing data of Space X launch data about launches, rockets used, payload delivered, launch specifications, landing specifications, and land outcomes.
- Problems you want to find answers
  - The goal is to use data to predict whether SpaceX will attempt to land a rocket or not.
  - Analyze how data, and features effect or do not affect rocket landing success
  - Determine which models (machine learning) can predict successful landing

Section 1

Methodology

Methodology

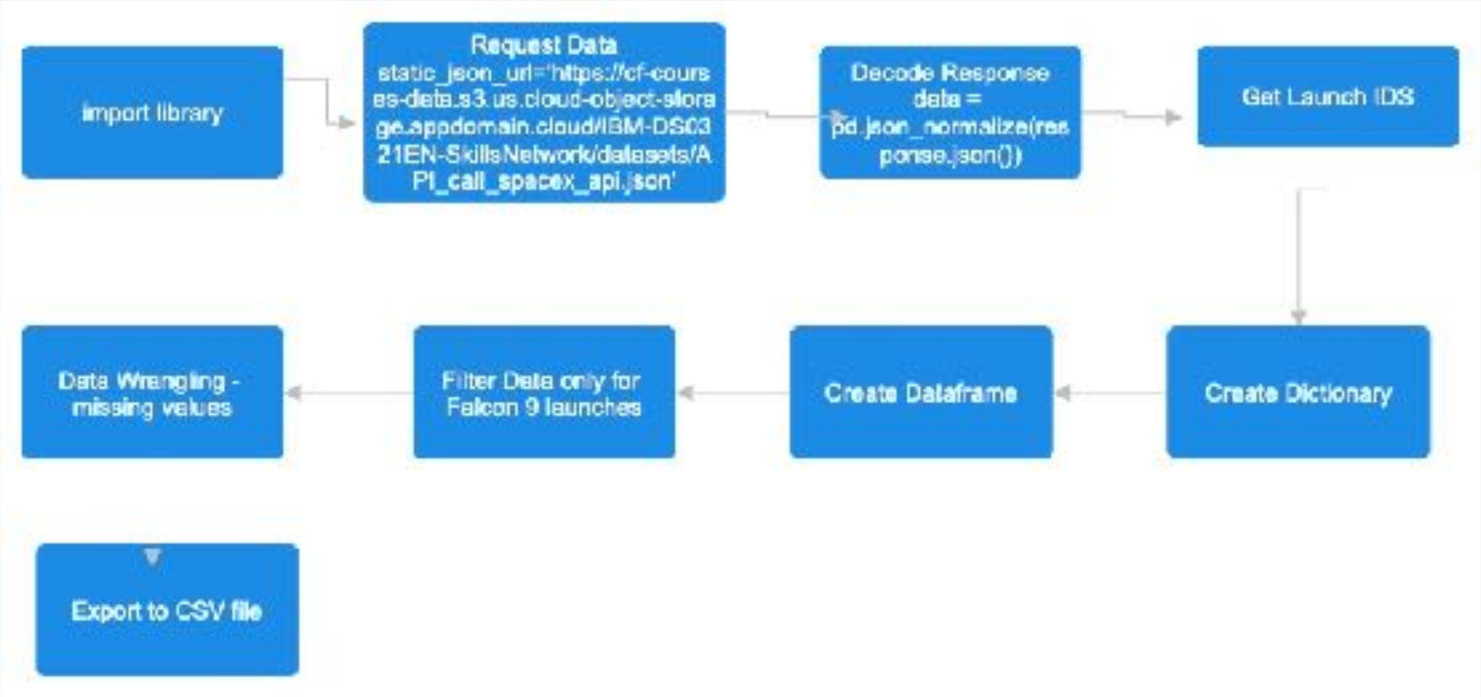
Executive Summary

- Data collection methodology:
  - Data collected using GET request - JSON file format using SpaceX Rest API and web scraping to obtain data from Wikipedia
- Perform data wrangling
  - Filtering data: identify and handle missing data values apply one hot encoding for non-numeric data.
  - Create data sets/data frames for analysis and visualization
- 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, and evaluate classification models
- Build Predictive Models

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

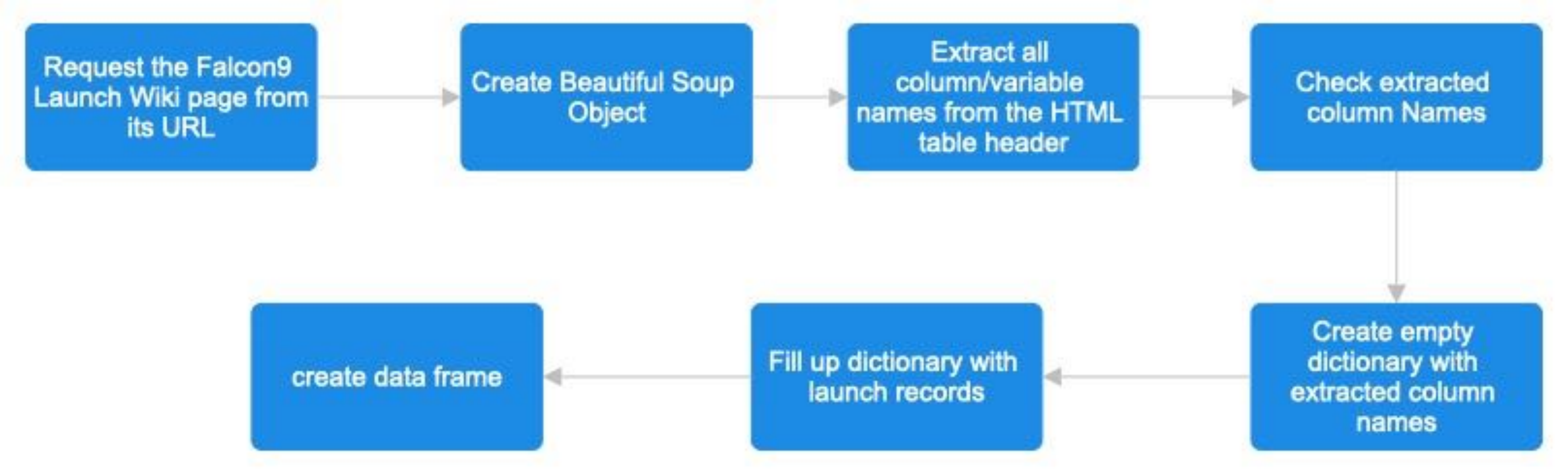
Data Collection – SpaceX API



Github Url

https://github.com/psuhm1/DS_IBM_Course/blob/main/01.a.jupyter-labs-spacex-data-collection-api.ipynb

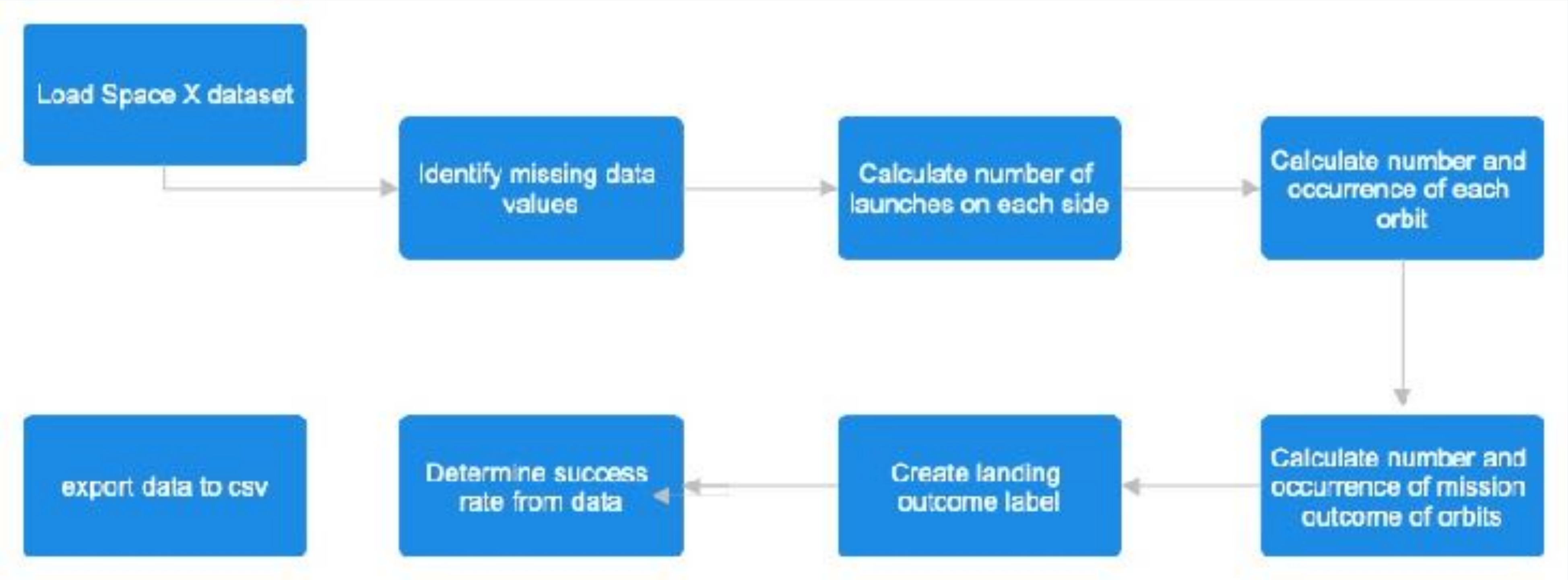
Data Collection - Scraping



Github url

https://github.com/psuhm1/DS_IBM_Course/blob/main/01.b.jupyter-labs-webscraping.ipynb

Data Wrangling



Github URL

https://github.com/psuhm1/DS_IBM_Course/blob/main/2.labs-jupyter-spacex-Data wrangling.ipynb

EDA with Data Visualization

Summarize what charts were plotted and why you used those charts

1. Scatter Plot Charts
    1. Flightnumber vs PayloadMass. The first stage is more likely to land successfully as flight number increases. As payload mass increases, the first stage is expected to land successfully.
    2. Visualize the relationship between the flight number and the launch site.
    3. Visualize the relationship to the Payload and Launch Site
    4. See the relationship between FlightNumber and Orbit type
2. Bar Chart
    1. Relationship of Orbit type vs success rate
3. Lineplot
    1. See the class vs. date relationship

Github URL

https://github.com/psuhm1/DS_IBM_Course/blob/main/2.A. jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Using bullet point format, summarize the SQL queries you performed

- Display unique launch sites
- Display 5 records where launch sites begin with string CCA
- Display total payload mass carried by boosters launched by NASA CRS
- Display average payload mass carried by booster version F9 v1.1
- List names of boosters that have success in drone ships and have payload mass greater than 4000 but less than 600
- List the total number of successful and failed mission outcomes
- List names of booster_versions that have carried maximum payload mass
- Last records which will display monthnames, failure landing outcomes in drone ship, booster version, and launch site in the year 2015
- Rank count of landing outcomes or success between date dates.

https://github.com/psuhm1/DS_IBM_Course/blob/main/2.B.jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Summarize map objects

1. Identify launch sites - identify where the launch sites are on the map
2. Create folium circle - circle NASA Johnson Space Center
3. Add marker cluster - distinguish by color
4. Add mouse position to coordinate for launch site
5. Add poly line - Calculate the distance of the launch site to the coast and create
6. Add markers with distance closet to city, highway and railway.

https://github.com/psuhm1/DS_IBM_Course/blob/main/
3.A.module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Summarize what plots/graphs and interactions you have added to a dashboard

1. Pie chart for all launch sites selected
2. Pie chart for launch site selected
3. Slide ranger to select payload for launch
4. Scatter Plot to show how payload is related to mission outcomes
5. Github URL
    1. https://github.com/psuhm1/DS_IBM_Course/blob/main/3.B.dash_interactivity.py

Predictive Analysis (Classification)

https://github.com/psuhm1/DS_IBM_Course/blob/main/4.A.module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

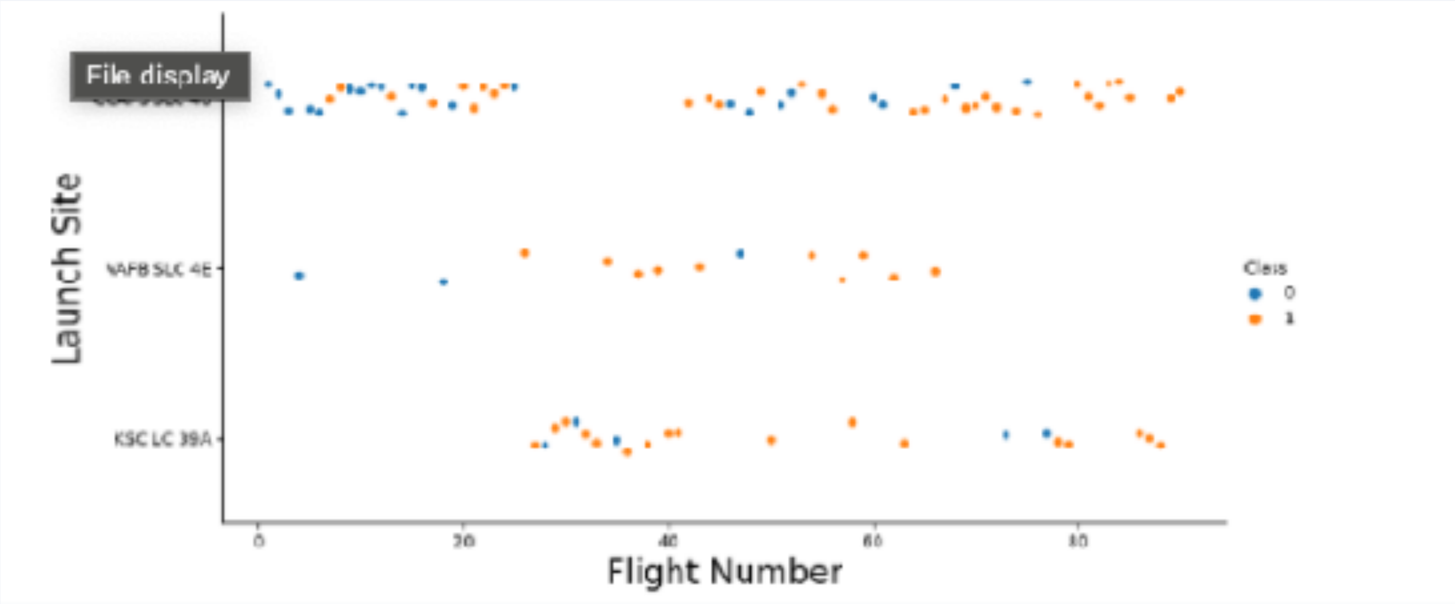
- Exploratory data analysis results
    - Orbit ES-L1, GEO, HEO AND SSO have 100% completion rate
    - Launches more successful with more time / launches
    - KSC LC-39A highest success rate for launch site
- Predictive analysis results
    - Model performance was the same/simliar for all models trained and predicted

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

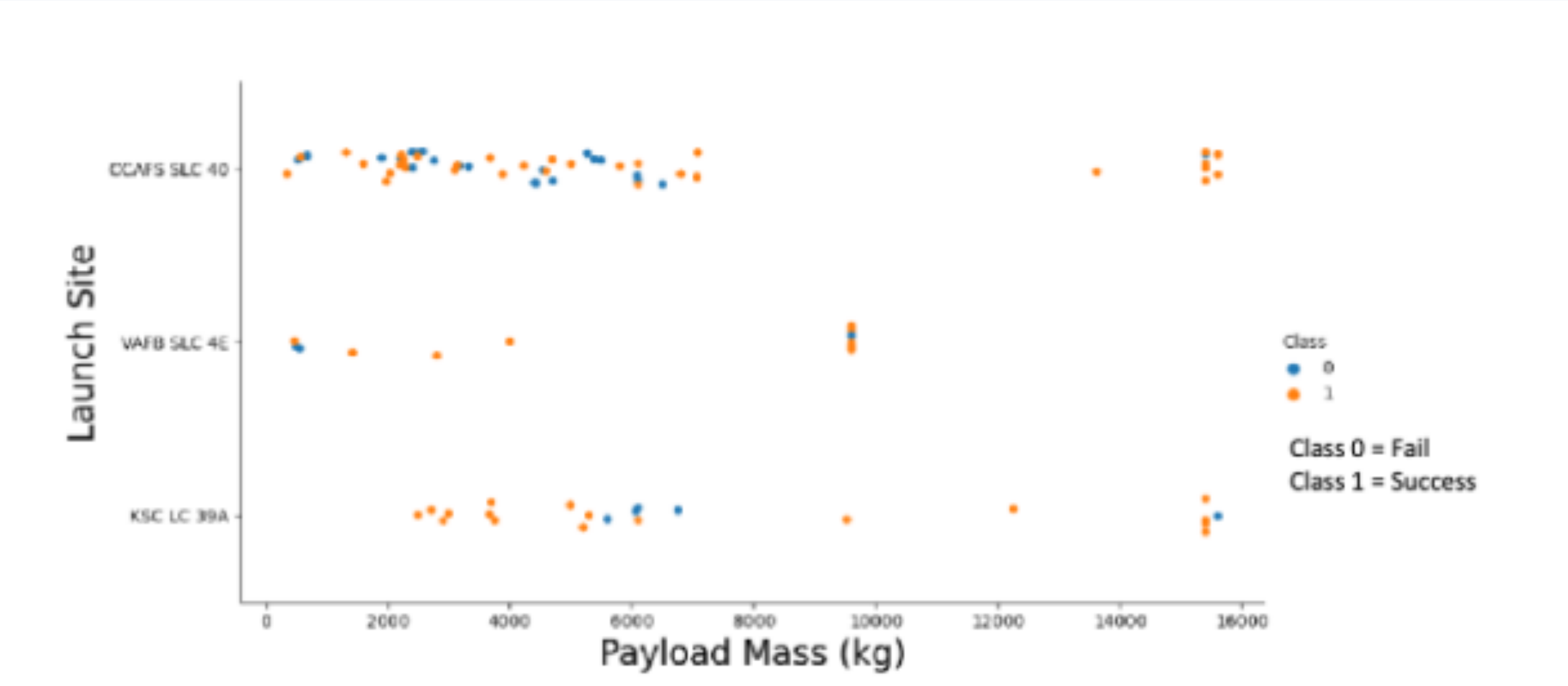
- Later flights had higher success rates than earlier flights
- VAFB SLC and KSC LC launch sites have higher success rates



Payload vs. Launch Site

Greater the payload mass - launch success increases

Payload mass above 8000kg were mostly successfuly

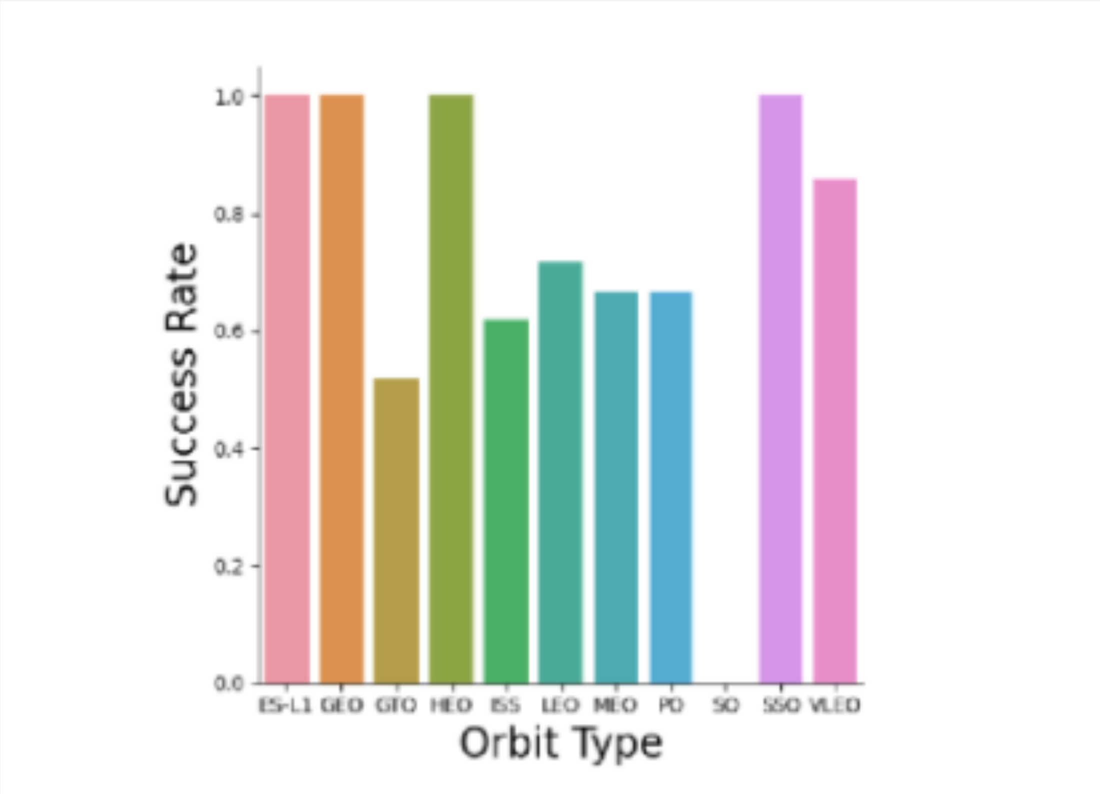


Success Rate vs. Orbit Type

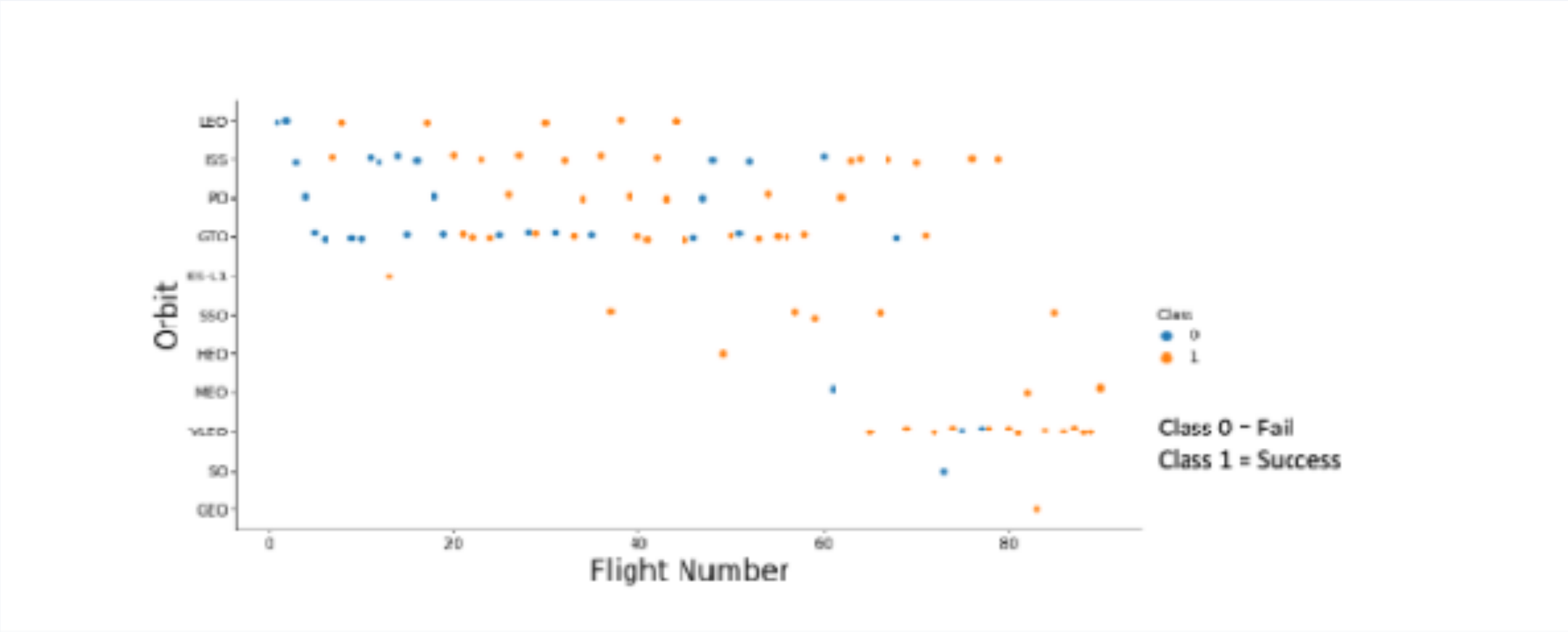
ES-L1, GEO, HEO, SSO = 100% SUCCESS RATE

GTO, ISS, LEO, MEO AND PO = 50%+ SUCCESS RATE

SO = 100% FAILURE

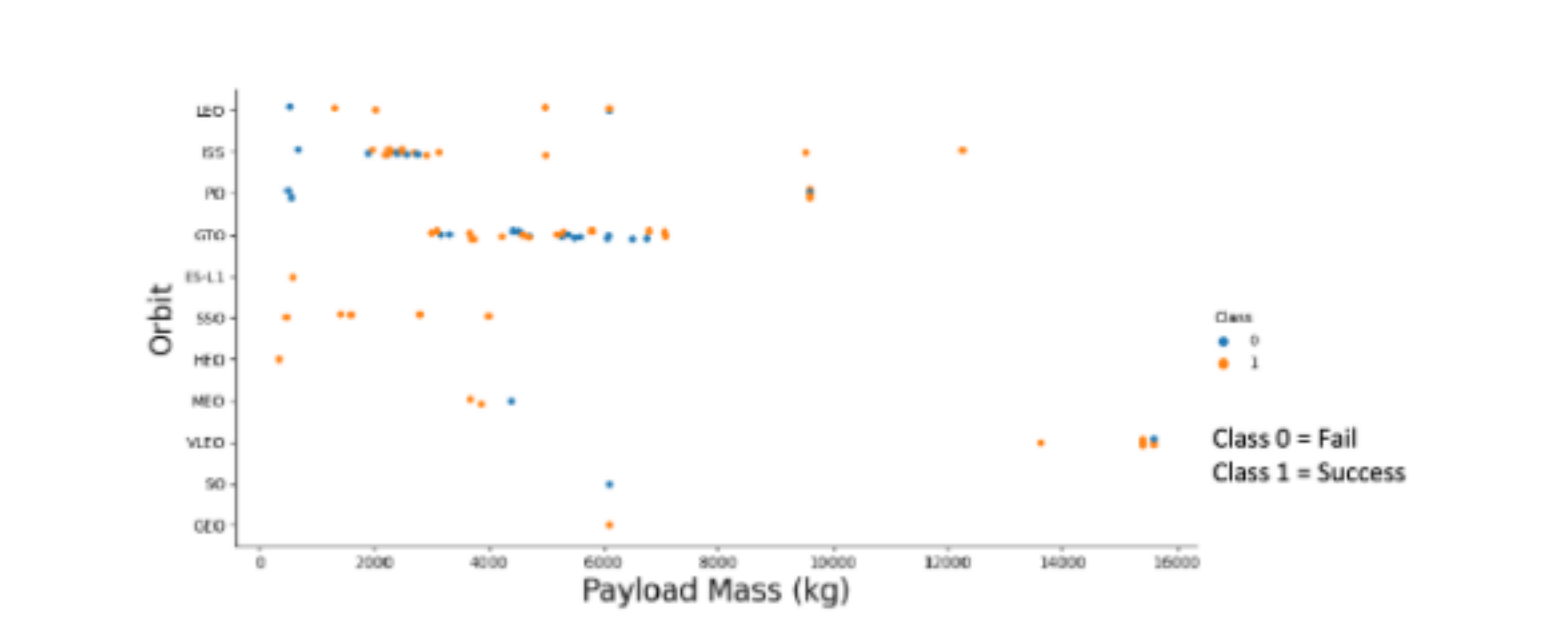


Flight Number vs. Orbit Type



Success rate improves as the number of flights increase for each orbit

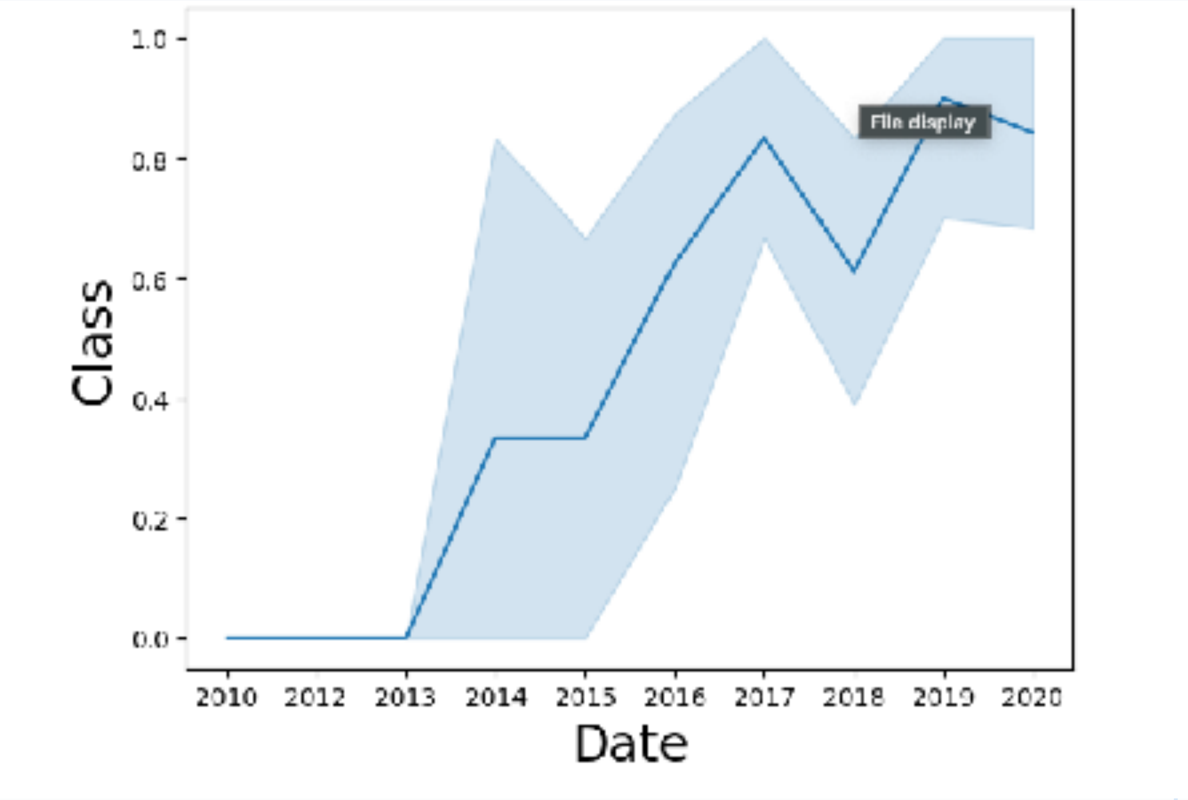
Payload vs. Orbit Type



Heavier payload increases success - shown by orbits LEO, ISS and PO

Launch Success Yearly Trend

- Launch success has improved from zero to 80%+
- 2019 had best performing year
- 2010-2013 had zero or poor performing years.



All Launch Site Names

```
In [13]: %sql select DISTINCT(LAUNCH_SITE) from SPACEXTBL;

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

Out[13]:

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [15]:
```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

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

Out[15]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

Total payload mass = 619967

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [22]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS payloadmass FROM SPACEXTBL;

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

Out[22]:

payloadmass
619967

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [24]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS payloadaverage FROM SPACEXTBL;
```

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

Out[24]:

payloadaverage
6138.287128712871

First Successful Ground Landing Date

In [30]:

```
%sql SELECT min(Date) FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success%';
```

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

Out[30]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [36]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE LANDING_OUTCOME='Success (drone ship)' AND PAYLOAD_MASS__KG_>400

 * sqlite:///my_data1.db
Done.

Out[36]:
```

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

Total Number of Successful and Failure Mission Outcomes

In [38]:
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS missionoutomes FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

Out[38]:

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

Boosters Carried Maximum Payload

In [49]:
```
%sql SELECT BOOSTER_VERSION as boostversion FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_= (SELECT MAX(PAYLOAD_MASS__KG_
```

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

Out[49]:

boostversion
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

```
In [76]: %sql SELECT substr(Date,6,2) as month, DATE, BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPACEXTBL WHERE L

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

Out[76]:

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

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

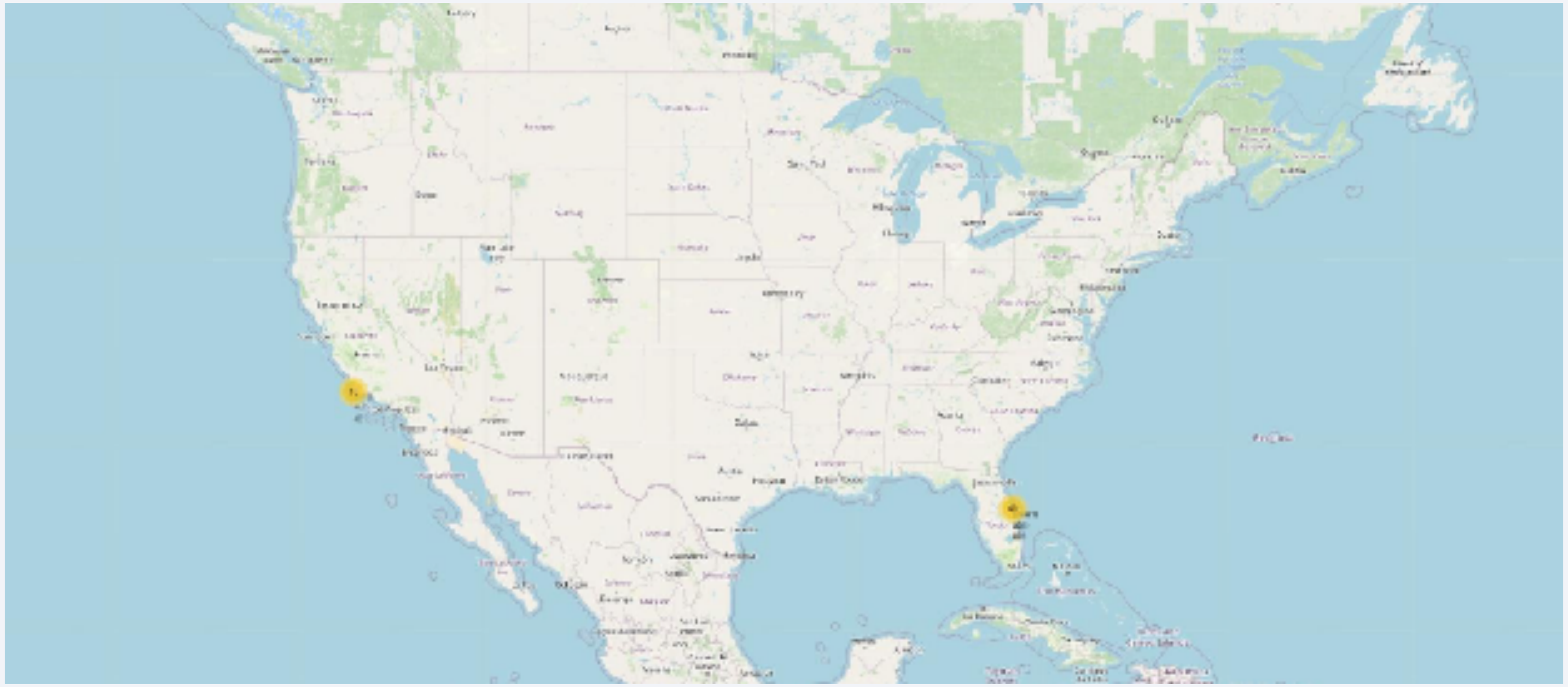
```
In [79]: %sql SELECT LANDING_OUTCOME, COUNT(*) AS COUNT_OUTCOMES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-

 * sqlite:///my_data1.db
Done.
Out[79]:
```

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

Section 3

Launch Sites Proximities Analysis

Launch Sites



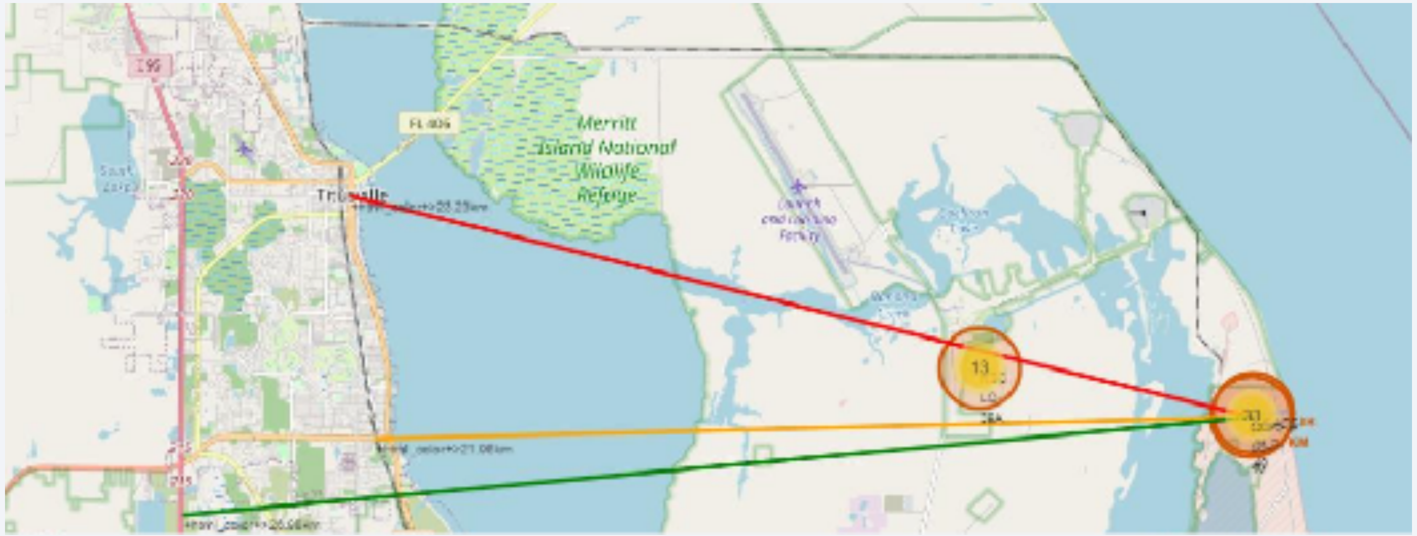
- Launch sites close to coast
- Launch sites close to equator
- Launch sites located where winter weather shouldn’t impact launch months

Launch Outcomes

Red dots denote not successful launches green shows successful launches

Distance to Proximities

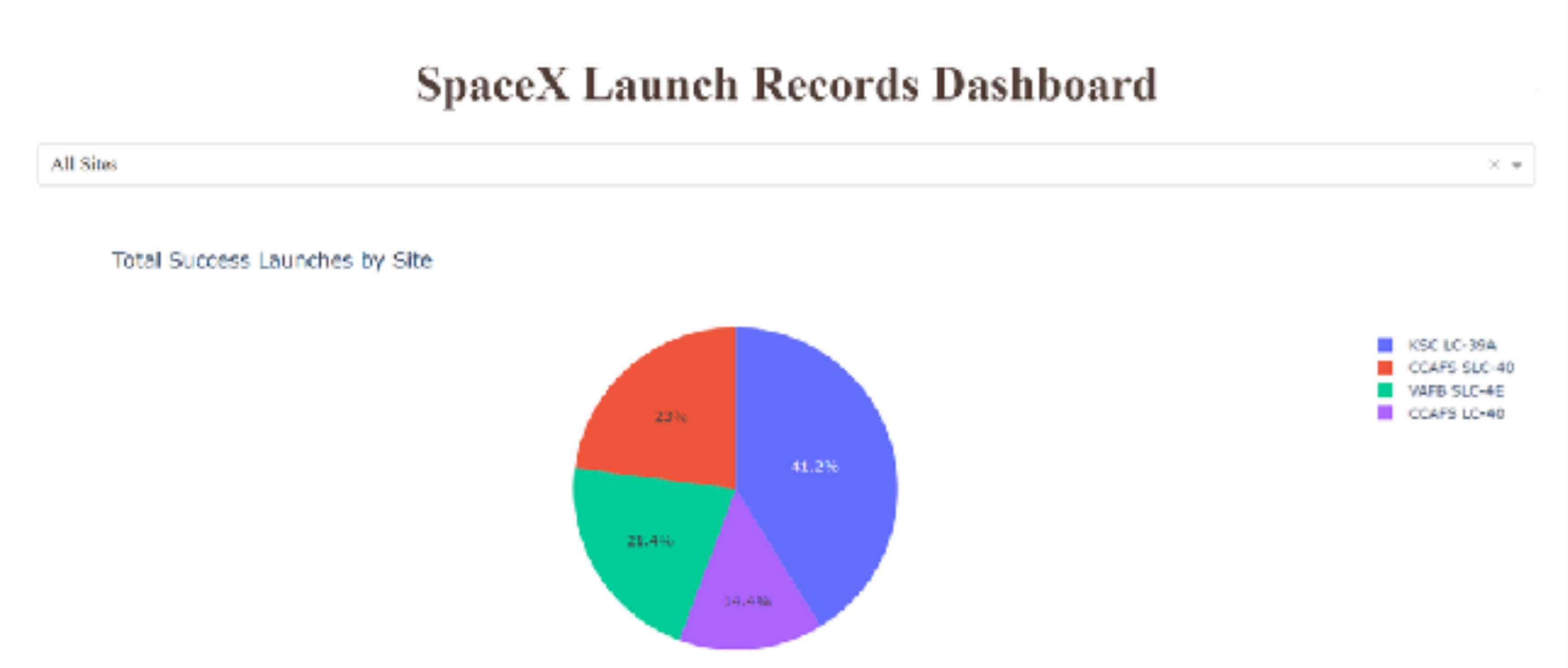
- CCAFS SLC-40 Launch Site
  - 27km to the closest highway
  - 23.2km to the nearest city
  - 22 to the closest railway
  - Less than 1km to the coast



Section 4

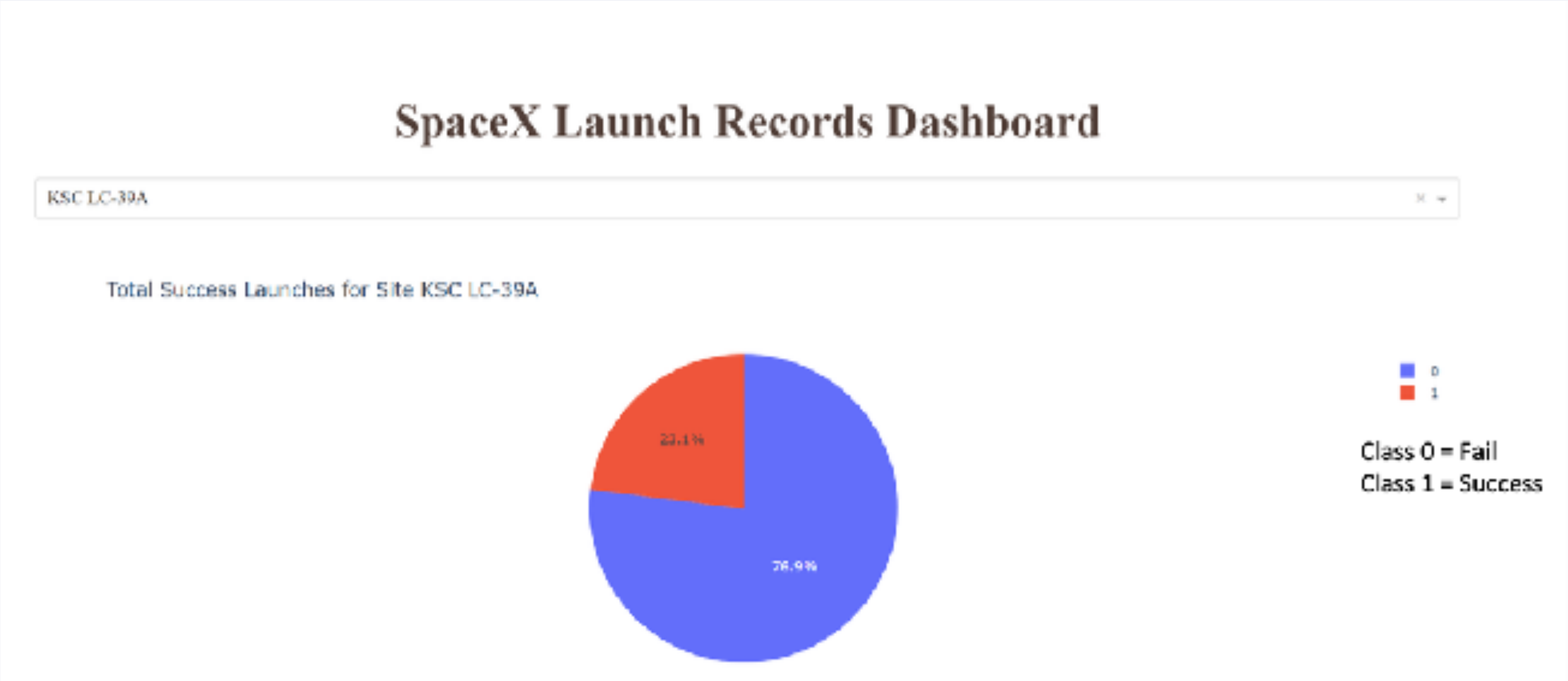
Build a Dashboard with Plotly Dash

Launch Success % by Launch Site



Highest success launch site KSC LC-29A

- 76.9% of launches were successful
- 23.1% of launches were not successful



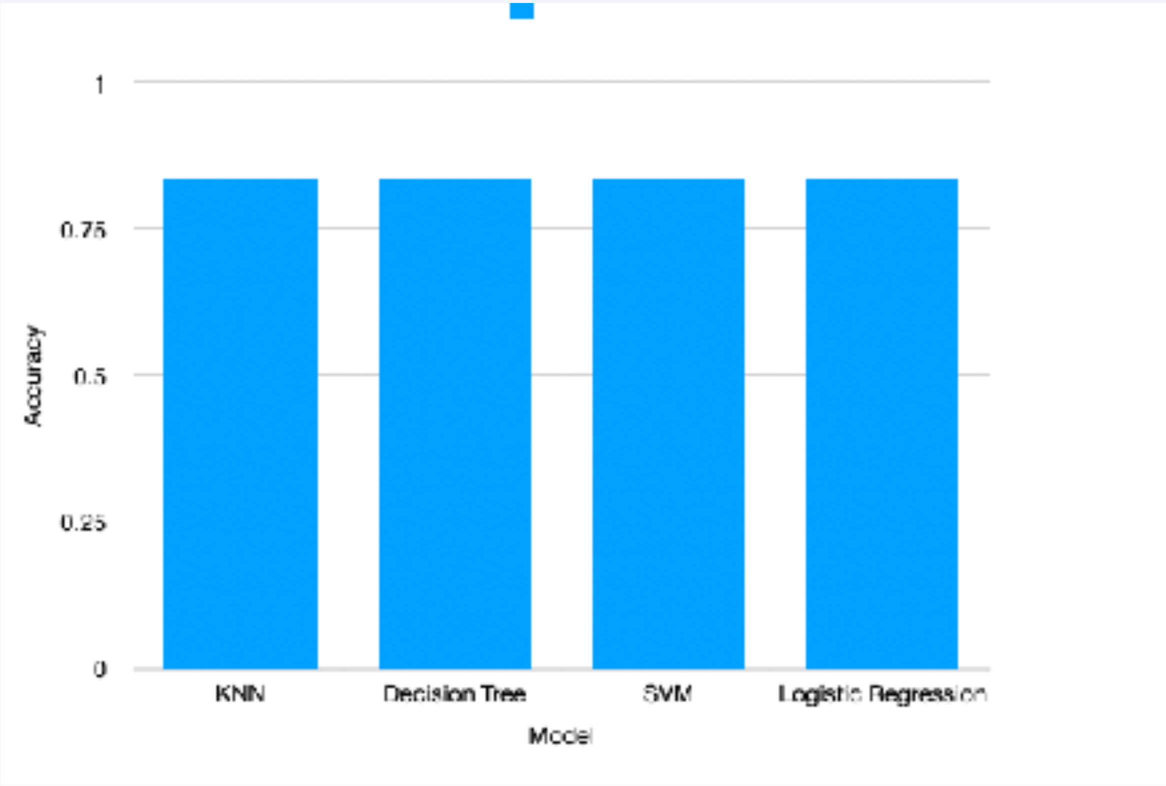
Compare payload mass versus launch success

Highest success with launches that had payloads between 2K kg and 5kg

Section 5

Predictive Analysis (Classification)

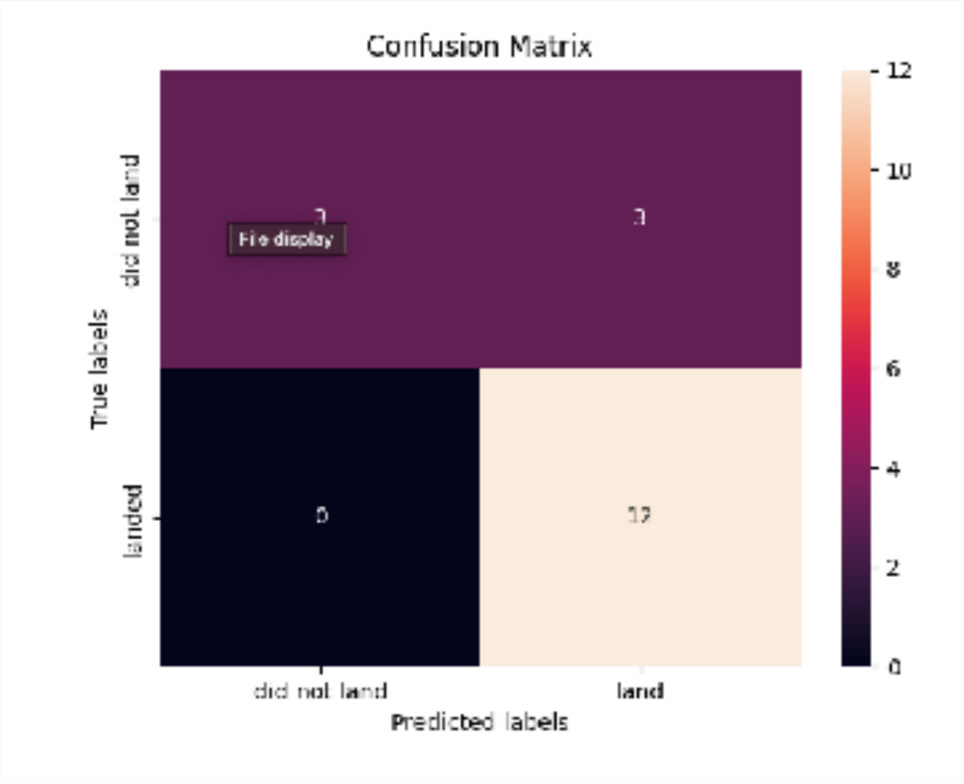
Classification Accuracy



Each of models had similar accuracy

Of 83.3%

Confusion Matrix



All confusion matrices for the data looked at the s

same. Not one of the matrix different data accurac

performance

Conclusions

- Model performance using the test data showed no model was better than others. All had 83.33% success rate.
- Overall launch success increases over time, and with more launches
- Launch sites have similar characteristics in that are near the coast, close to the equator, and far away from cities
- Payload mass does affect launch success; more payload mass has a higher launch success rate
- Launch site selection is a factor where KSC LC-39A has the highest success rate.
- Additional newer data/dataset show results that differ from the current data set. Hypothesis should be examed

Thank you!