IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eric Wong
April 7th, 2022

Outline

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

Executive Summary

- Data Collection done with API and Webscraping, then cleaned through Wrangling
- Visualization reveals the best orbit, flight number and payload mass combination to achieve launch success, and machine learning pipelines can be built to determine best prediction methods

Introduction

- SpaceY is looking to compete with SpaceX's rocket launches
- SpaceX's Falcon 9 rocket launches are occasionally able to reuse first stage of launch
  - First stage is most fuel-intensive and expensive stage
- Want to create model to predict whether a certain rocket launch can reuse first stage of launch

Section 1

Methodology

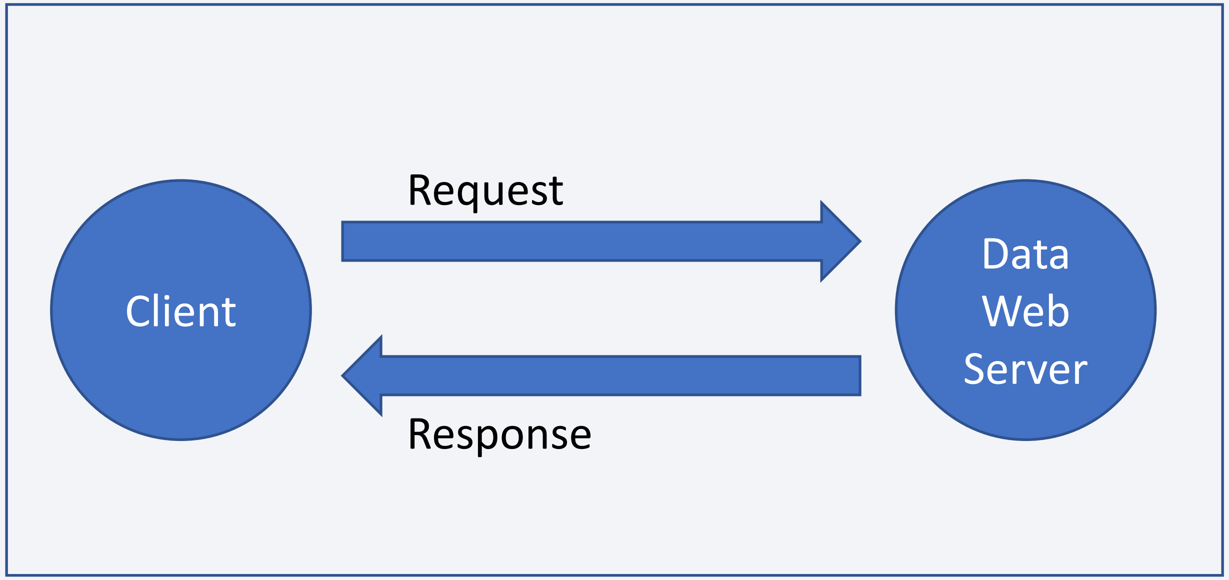
Methodology

Executive Summary

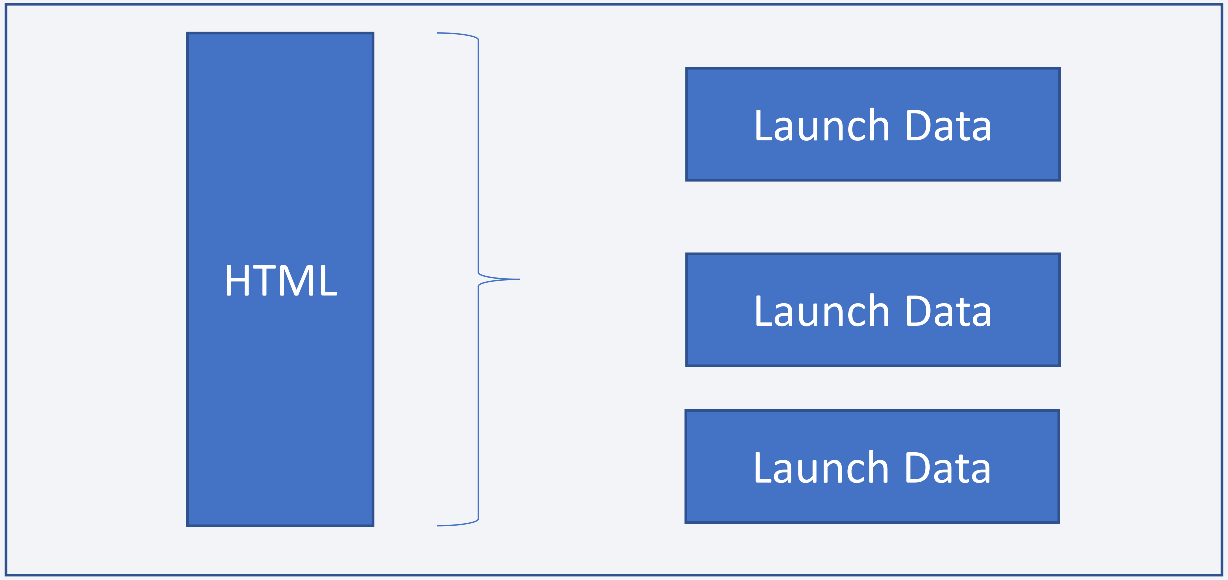
- Data collection methodology:
  - Data Collected with API calls/Web Scraping
- Perform data wrangling
  - Reclassified launches into binary successes and failures
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built machine learning pipeline to determine if launch results in success or failure for recovering first stage

Data Collection

- Data sets were collected using REST API Calls and Webscraping
- API
- Webscraping



API

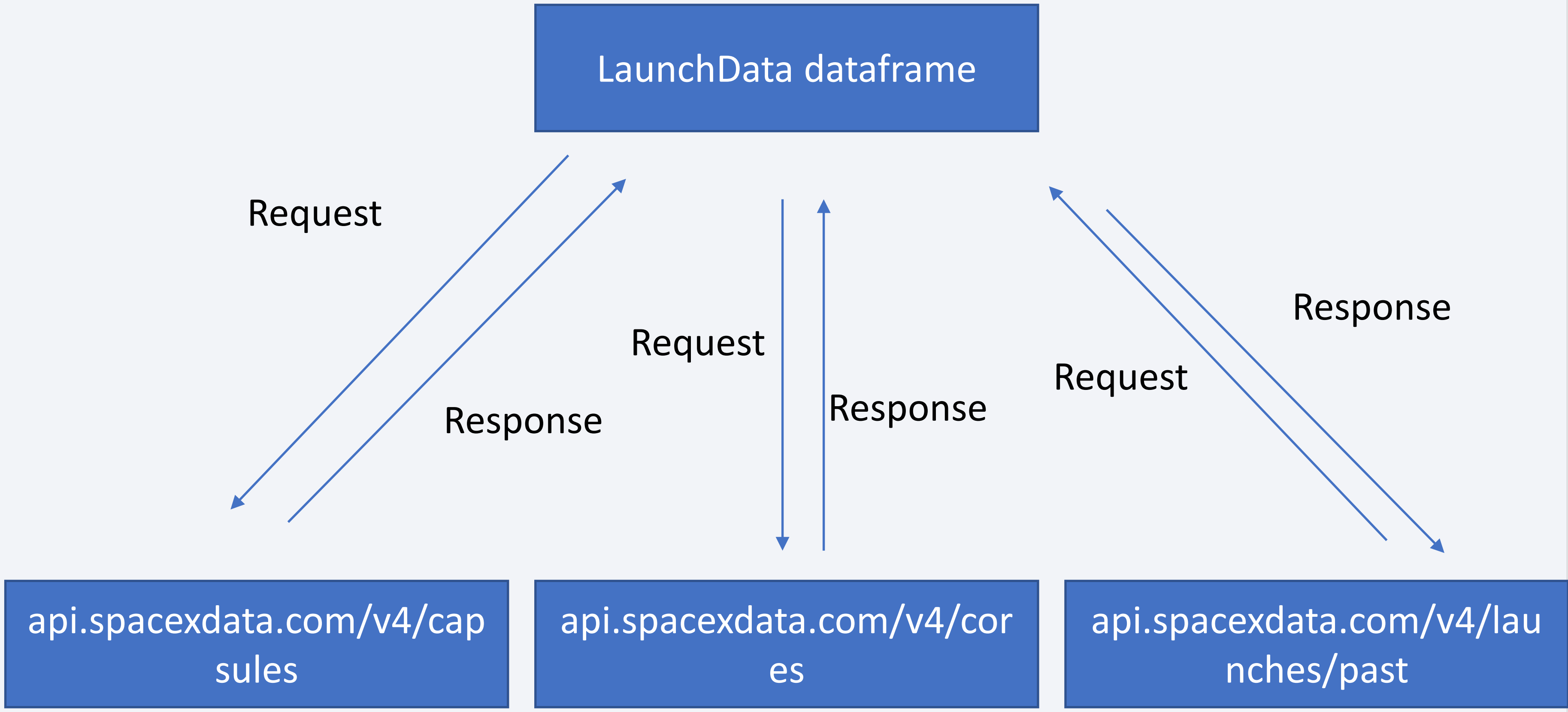


Webscraping

Data Collection – SpaceX API

- SpaceX Launch Data collected using a REST API

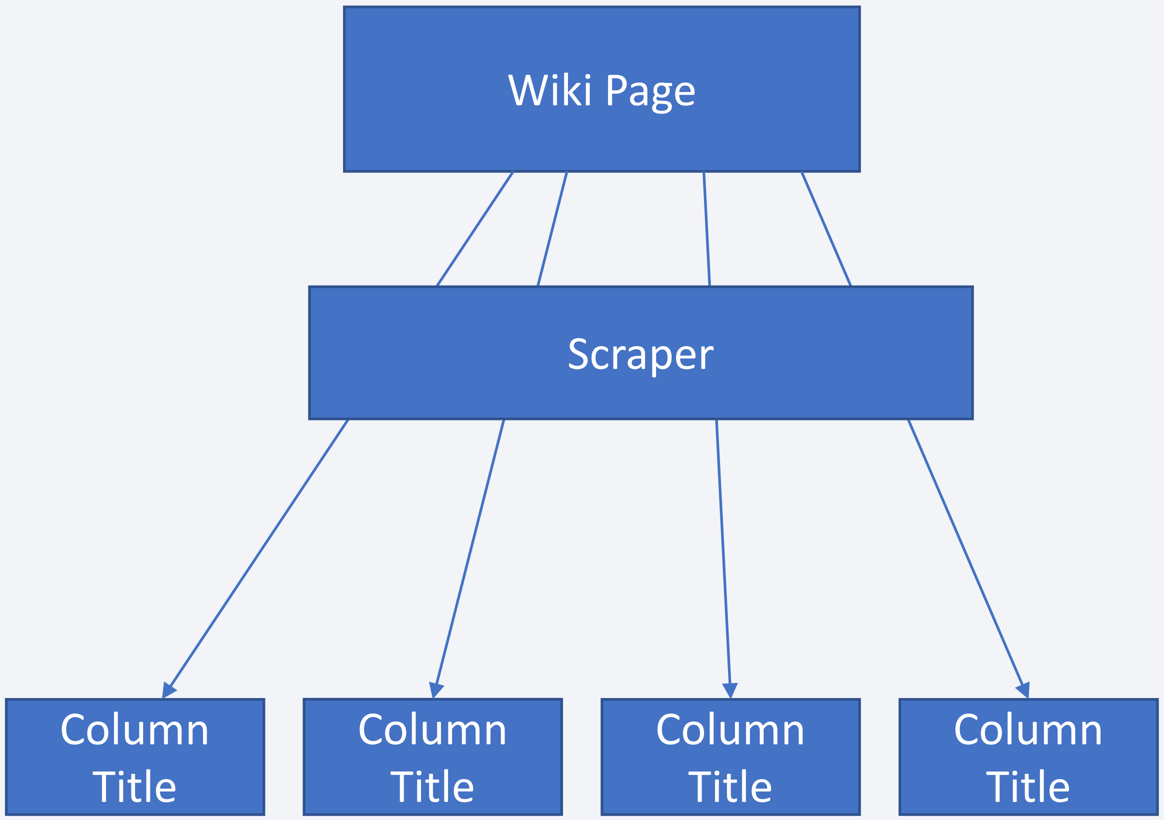
- Github Notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/Data%20Collection%20API.ipynb



Data Collection - Scraping

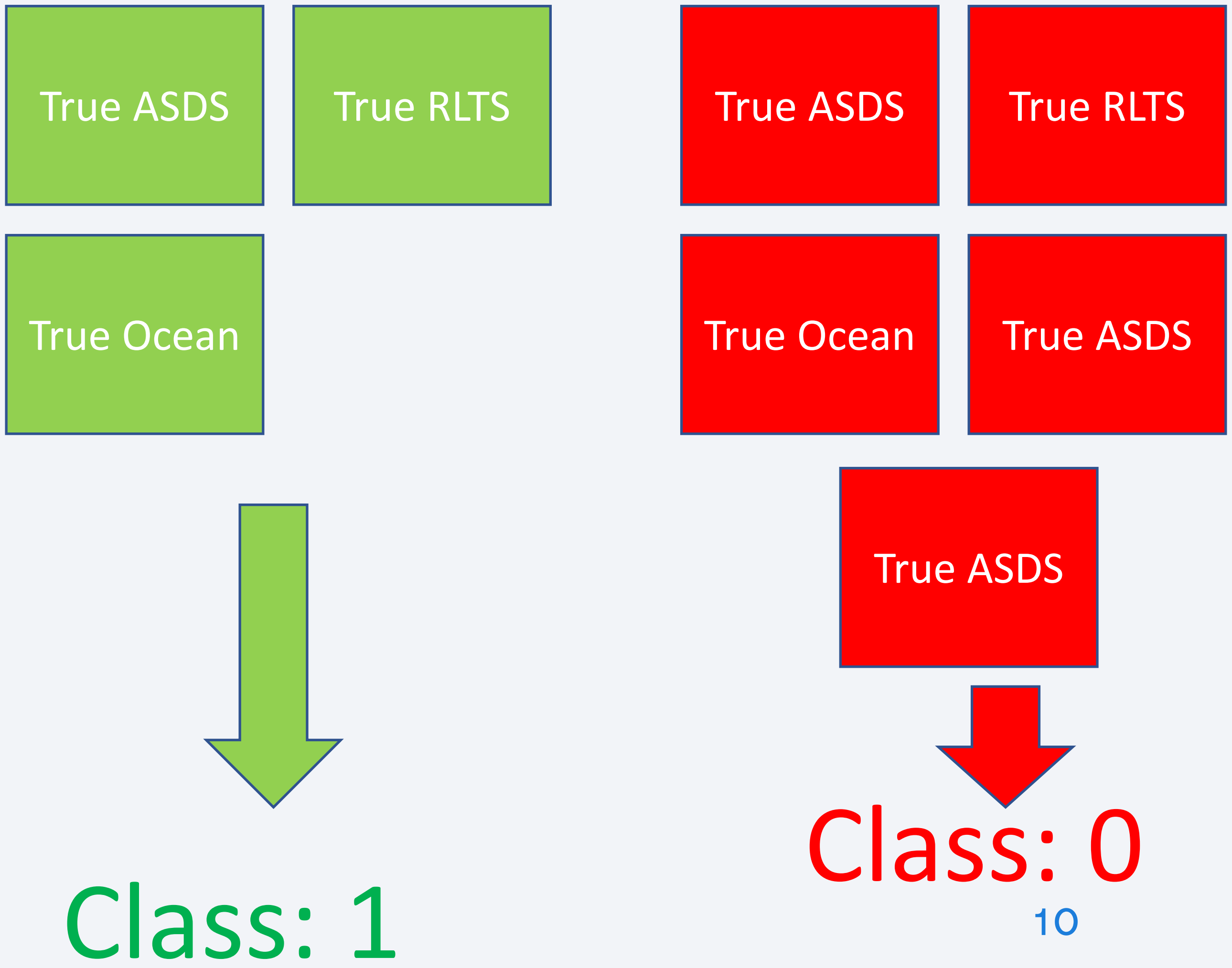
- BeautifulSoup Webscraper used to grab column titles and names from wiki page

- GitHub Notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Data is cleaned and unified before any data is processed
  - "Good" landings where first stage is recovered is converted into numerical variable "Class = 1"
  - "Bad" landings is converted into "Class = 0"
- GitHub notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/EDA.ipynb



EDA with Data Visualization

- Numerous charts plotted to examine data
  - "Payload Mass (kg) vs. Flight Number" – Track successes over increasing flight numbers, mass
  - "Flight Number vs. Launch Site" – Separate launches by launch site to see success trends
  - "Launch Site vs. Payload Mass" – Track success over Payload Mass
  - "Class vs. Orbit Type" – Bar Chart to see probability of success based on Orbit type
  - "Orbit vs. Flight Number" – Separate launches by orbit to see success over successive launches
  - "Orbit vs. Payload Mass" – Track success of different orbits over increasing Payload Mass
  - "Probability of Success vs. Year" – Track probability trend over increasing years
- Github Notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

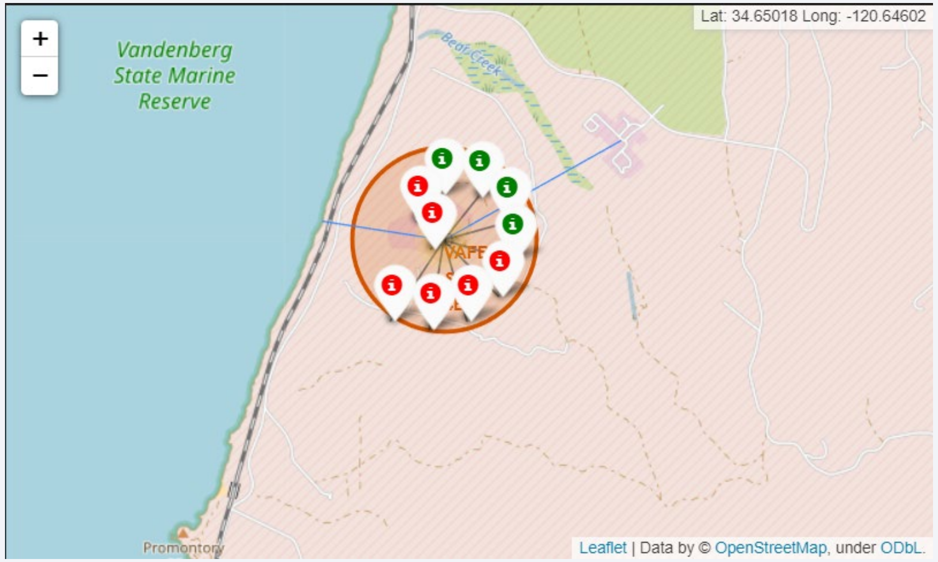
- SQL queries made to explore the data:
  - Displayed the names of unique launch sites
  - Displayed 5 records where the launch site began with the string "CCA"
  - Displayed Total Payload Mass carried by NASA (CRS) Launches
  - Displayed Average Payload Mass carried by Booster version F9 v1.1
  - Listed Earliest Date of successful Ground Pad Landing
  - Listed Booster Names of launches with success on Drone Ship and that has Payload Mass greater than 4000 and less than 6000
- GitHub Notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/EDA%20with%20SQL.ipynb

EDA with SQL

- SQL queries made to explore the data:
  - Listed total number of successful and failure mission outcomes
  - Listed the name of Boosters that carried maximum Payload Mass
  - Listed the booster version, launch site names and failed landing outcomes occurring on a drone ship in Year 2015
  - Listed the count of each landing outcome between 2010-06-04 and 2017-03-20, in descending order
- GitHub Notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/EDA%20with%20SQL.ipynb

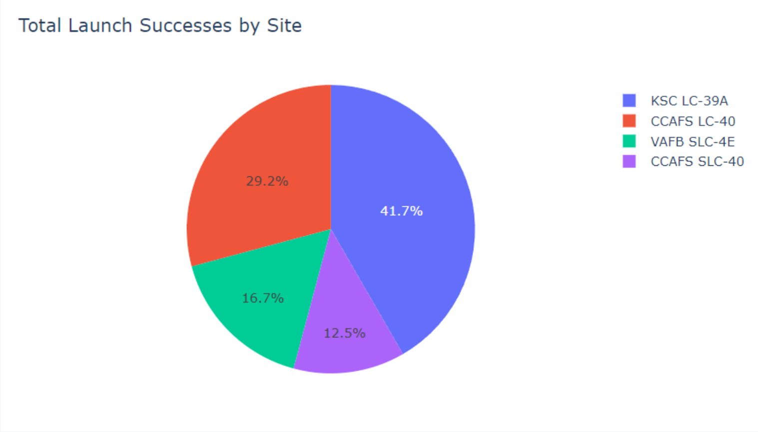
Build an Interactive Map with Folium

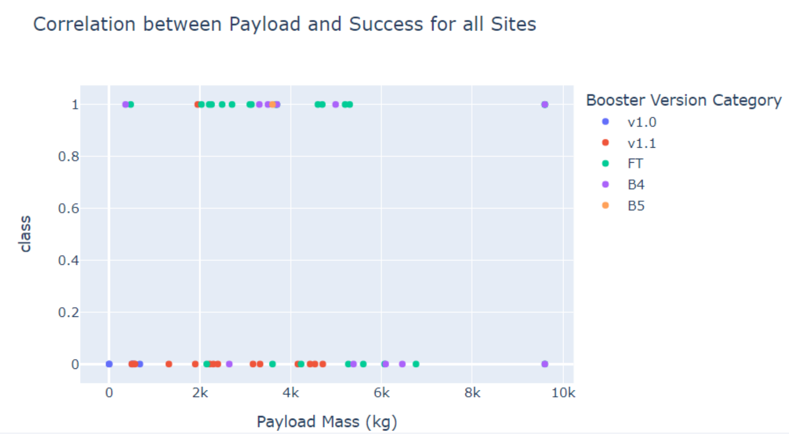
- Added in marker clusters to track geographical locations of all launches
    - These launches allow us to visually examine geographical trends
- GitHub Notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb



Build a Dashboard with Plotly Dash

- Dashboard used to quickly examine data trends
  - Pie chart used to examine successes broken down by launch pad names/Success vs. failure for each launch site
  - Line chart used to see Success/Failure vs. Payload Mass
- Plots/interactions created to see effects of launch sites and Payload Mass on mission success
- GitHub Notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/spacex_dash_app.py





Predictive Analysis (Classification)

- Classification model built in final steps:
  - “Class” data is transformed to Numpy data
  - Data is transformed and then a train/test split is created
  - GridSearchCV items are created for Logistic Regression, SVM, Decision Trees and K-Nearest Neighbors predictors
  - Accuracy scores and confusion matrices used to evaluate models
- GitHub Notebook: https://github.com/erictwong18/IBM_SpaceX_Project/blob/main/Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results

```python
df.head(5)
```

Python

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Results

- Predictive analysis results (using best parameters)

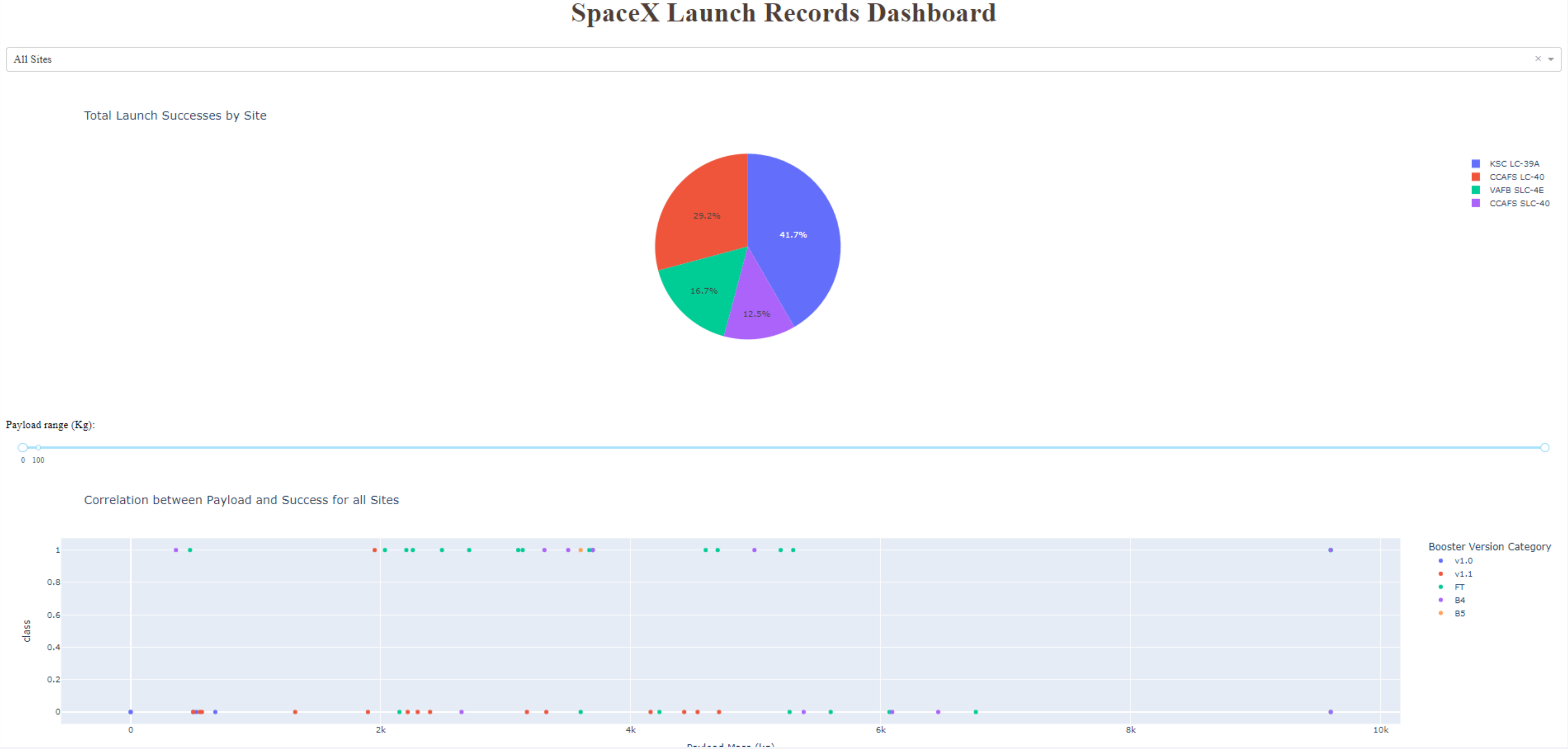
```
"""
testLogReg = LogisticRegression(C = logreg_cv.best_params_['C'], penalty = logreg_cv.best_params_['penalty'], solver = logreg_cv.best_params_['solver'])
testLogReg.fit(X_train, Y_train)
testLogReg.score(X_test, Y_test)
logreg_cv.best_params_
"""
print("Logistic Regression Best Score: " + str(logreg_cv.best_score_))
print("SVM Best Score: " + str(svm_cv.best_score_))
print("Decision Tree Best Score: " + str(tree_cv.best_score_))
print("K-Nearest Neighbors Best Score: " + str(knn_cv.best_score_))
```

[32] Python

```
Logistic Regression Best Score: 0.8464285714285713
SVM Best Score: 0.8482142857142856
Decision Tree Best Score: 0.8892857142857142
K-Nearest Neighbors Best Score: 0.8482142857142858
```

Results

- Interactive Analysis Dashboard Demo

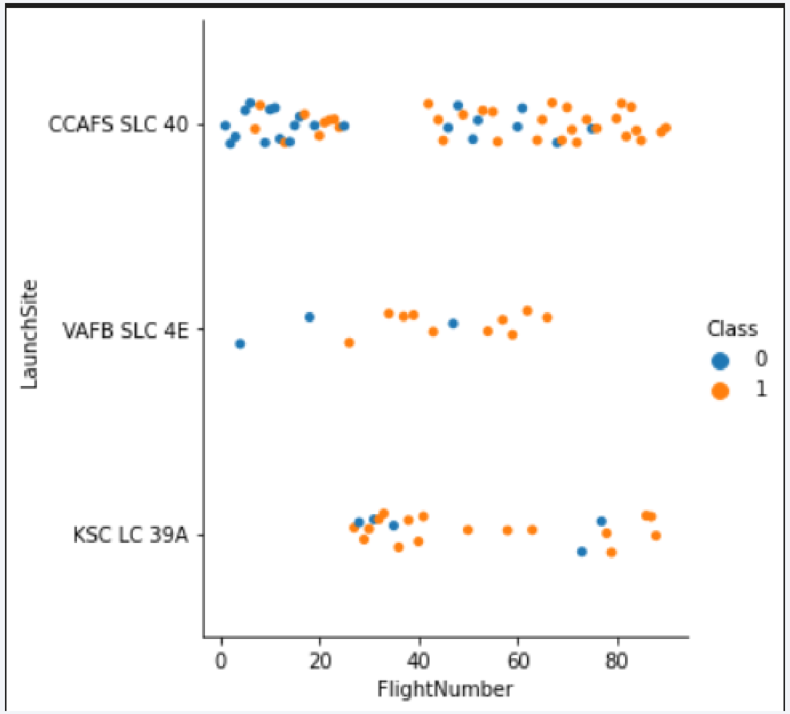


Section 2

Insights drawn from EDA

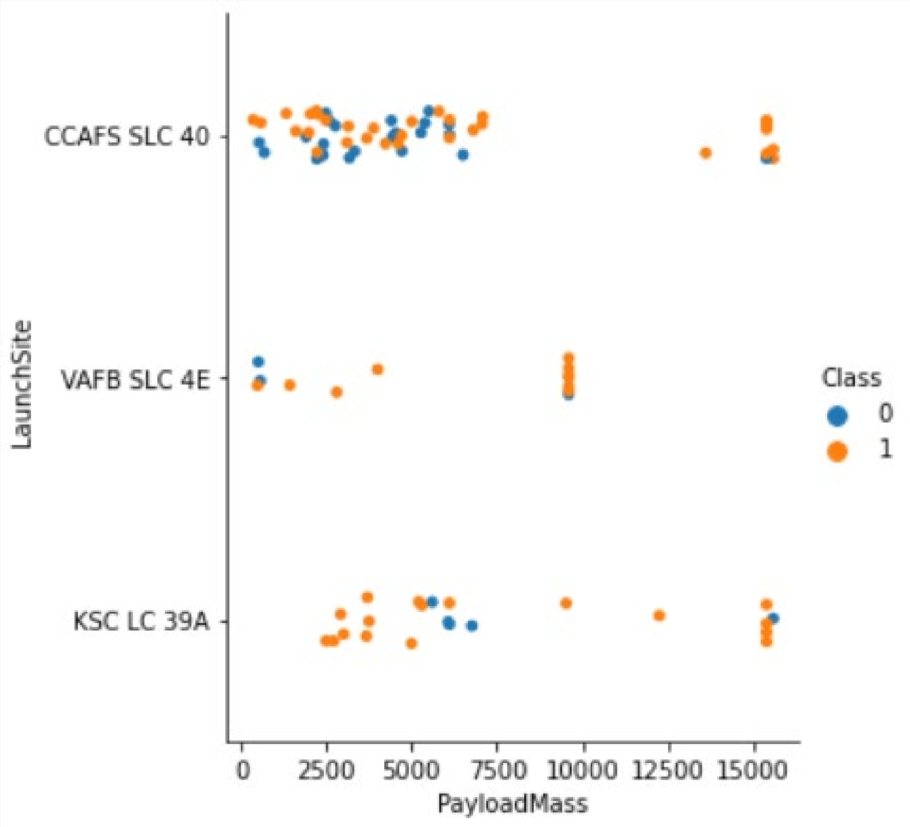
Flight Number vs. Launch Site

- Success across the Launch Sites increases with flight number



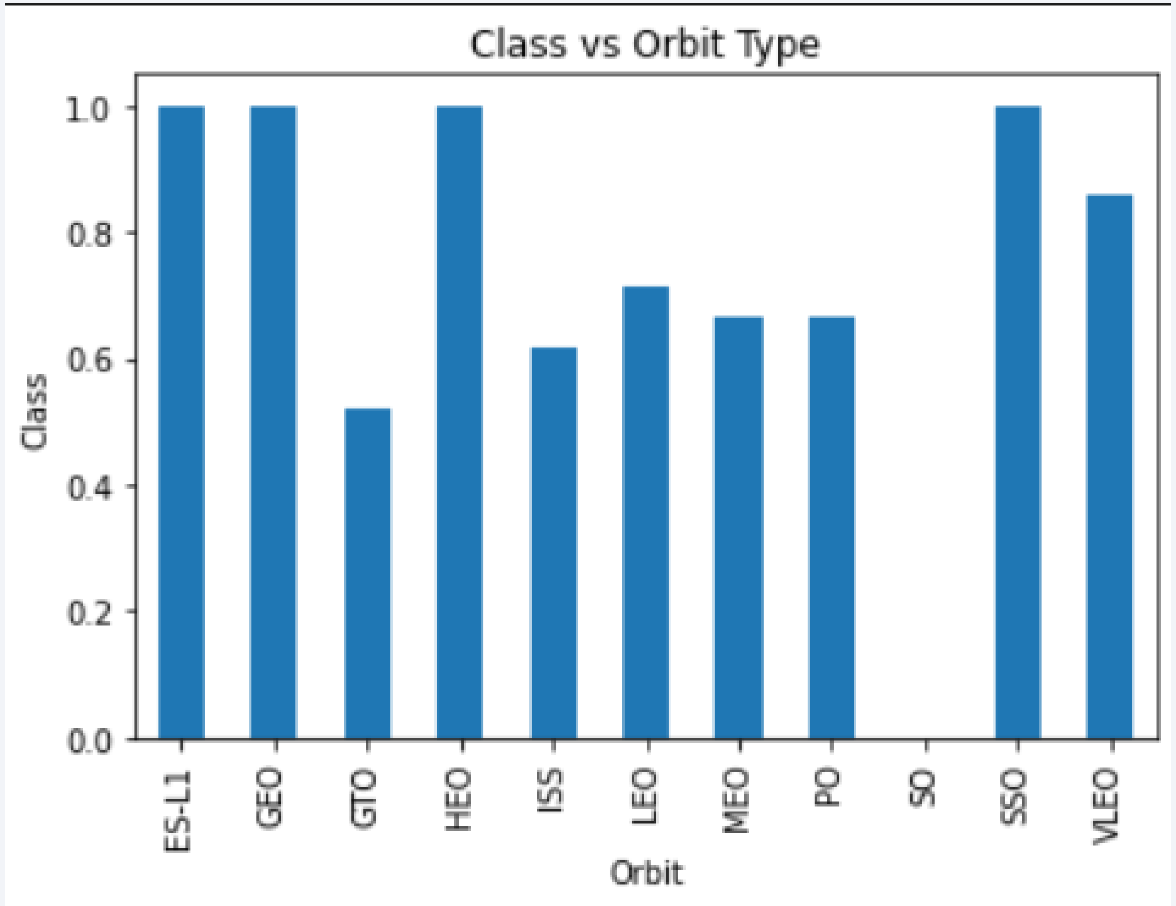
Payload vs. Launch Site

- Success and failures varies with each payload mass
    - VAFB SLC 4E limits the Payload Mass at 10000



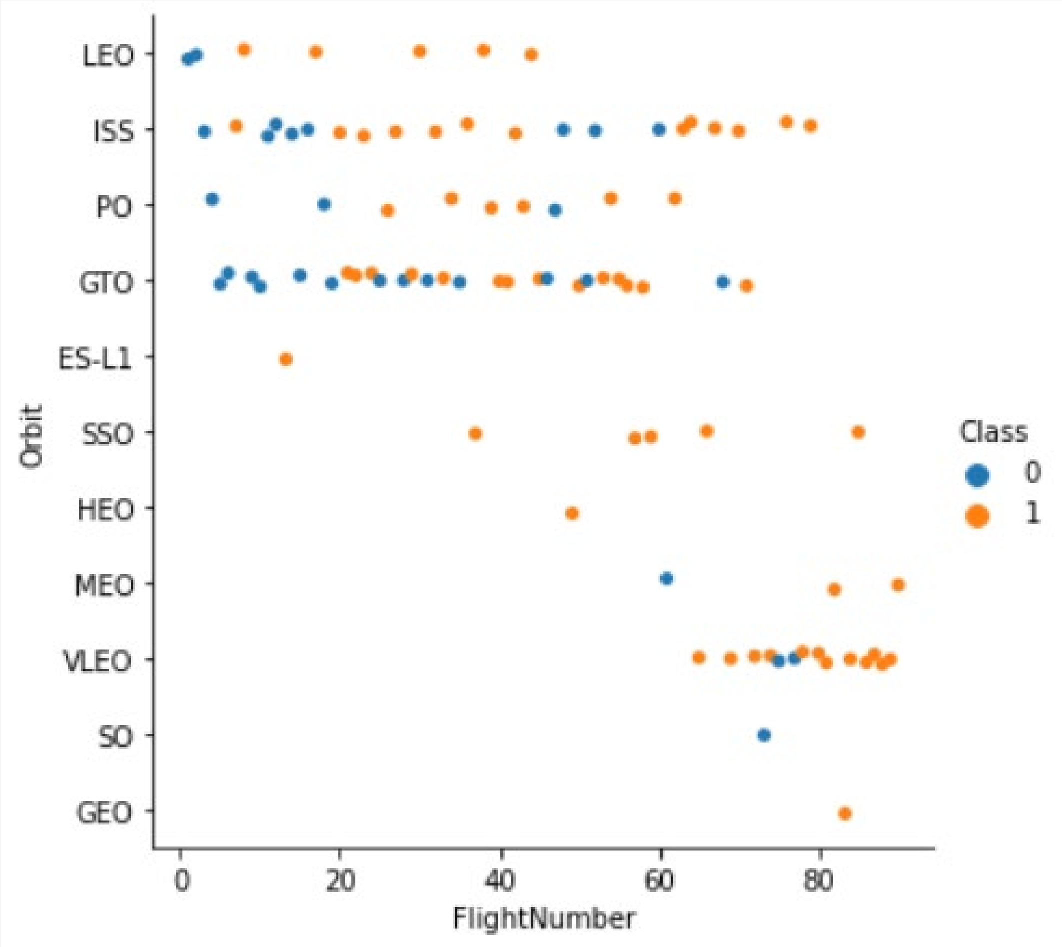
Success Rate vs. Orbit Type

- Bar Chart displays probability of success (on scale from 0.0 to 1.0) for various orbits
- ES-L1, GEO, HEO and SSO all have 100% probability of success



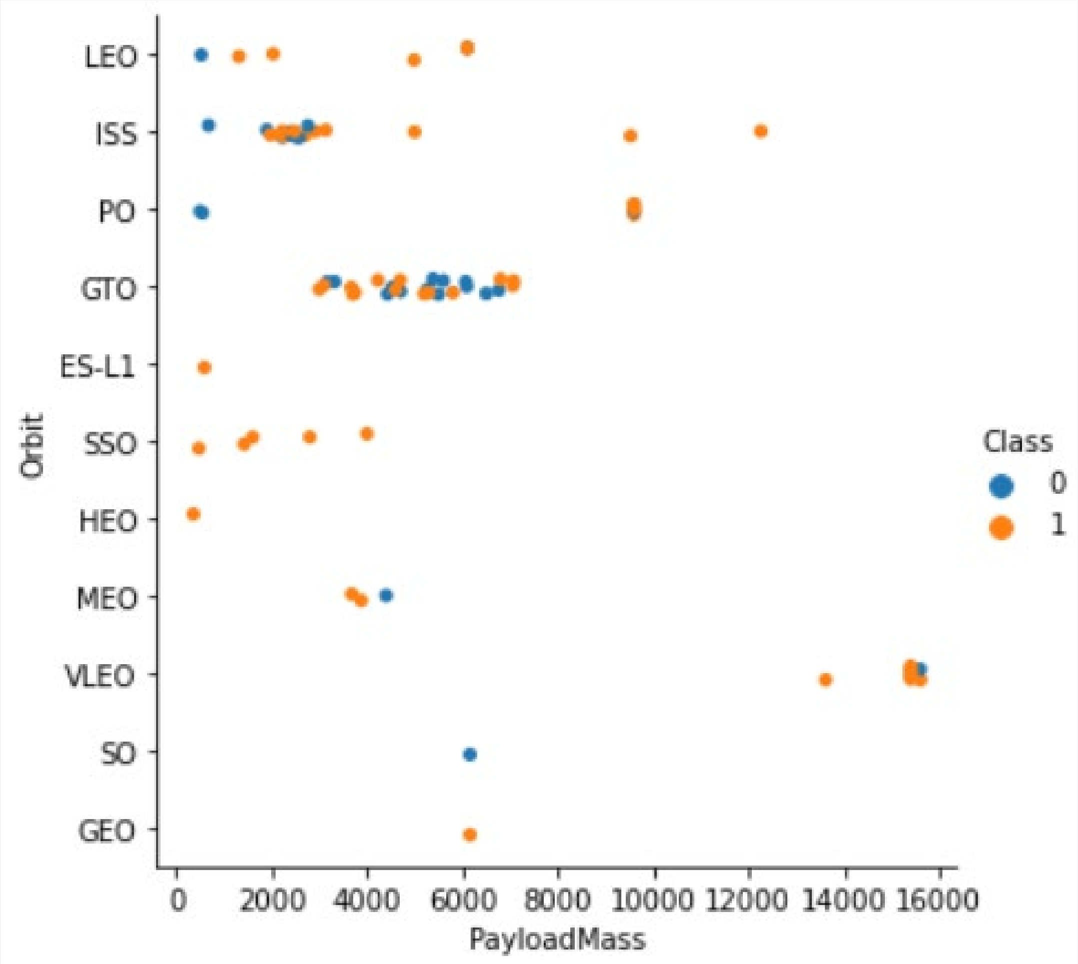
Flight Number vs. Orbit Type

- Success/Failures of launches are divided by successive flight numbers and orbit types



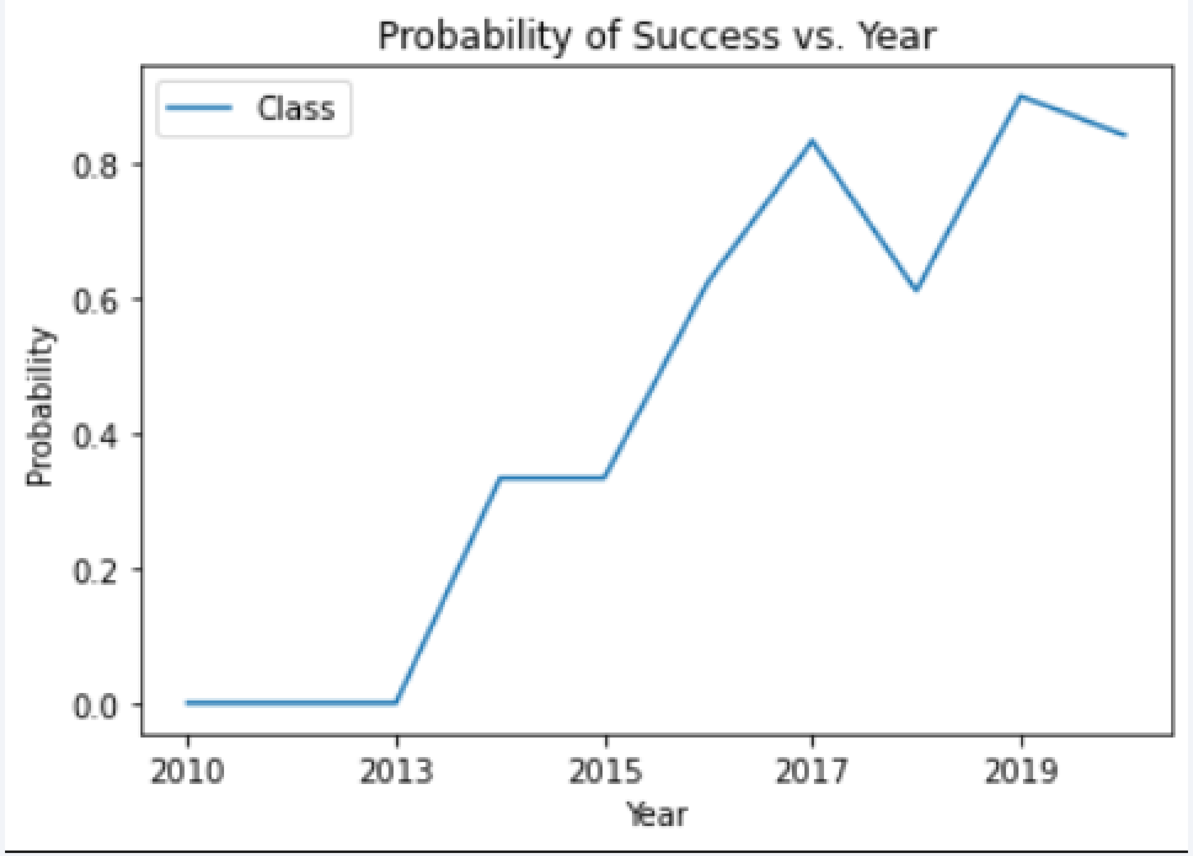
Payload vs. Orbit Type

- Success/Failures of launches are divided by successive payload mass and orbit types



Launch Success Yearly Trend

- Line Chart tracking the overall probability of success for launches over years
    - Success as a whole generally increases over time



All Launch Site Names

- SQL Query used to determine all the names of launch sites

```
%sql SELECT DISTINCT launch_site FROM SPACEXDATASET
```

Python

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

Launch Site Names Begin with 'CCA'

- SQL Query used to find 5 records where launch sites begin with string ‘CCA’

```
%sql SELECT * FROM SPACEXDATASET WHERE launch_site LIKE 'CCA%' LIMIT 5
```

Python

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 used to find the total payload mass for boosters launched by NASA (CRS)

```
#%sql SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET WHERE customer LIKE 'NASA (CRS)'
%sql SELECT SUM(payload_mass__kg_) FROM (SELECT * FROM SPACEXDATASET WHERE customer LIKE 'NASA (CRS)')
```

Python

1
45596

Average Payload Mass by F9 v1.1

- SQL query to find the average payload mass carried by F9 v1.1 boosters

```
%sql SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET WHERE booster_version LIKE 'F9 v1.1'
```

1
2928

First Successful Ground Landing Date

- SQL Query to find the earliest successful landing outcome for ground pad launches

```
%sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE landing__outcome LIKE 'Success (ground pad)'
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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 DISTINCT booster_version FROM SPACEXDATASET WHERE (payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000) AND landing__outcome LIKE 'Success (drone ship)'
```

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 Query to list the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(DATE), landing__outcome FROM SPACEXDATASET GROUP BY landing__outcome
```

1	landing__outcome
5	Controlled (ocean)
3	Failure
5	Failure (drone ship)
2	Failure (parachute)
22	No attempt
1	Precluded (drone ship)
38	Success
14	Success (drone ship)
9	Success (ground pad)
2	Uncontrolled (ocean)

Boosters Carried Maximum Payload

- SQL Query to list the names of Boosters carrying the maximum payload mass.

```
%sql SELECT DISTINCT booster_version FROM SPACEXDATASET WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET)
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- SQL Query to find the failed landing outcomes in the drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT landing__outcome, booster_version, launch_site FROM SPACEXDATASET WHERE landing__outcome LIKE 'Failure (drone ship)' AND DATE LIKE '2015%'
```

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

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

- SQL Query to count the landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

```
LECT landing__outcome, COUNT(booster_version) as outcome_count FROM SPACEXDATASET WHERE DATE > '2010-06-04' AND Date < '2017-03-20' GROUP BY landing__outcome ORDER BY outcome_count
```

Python

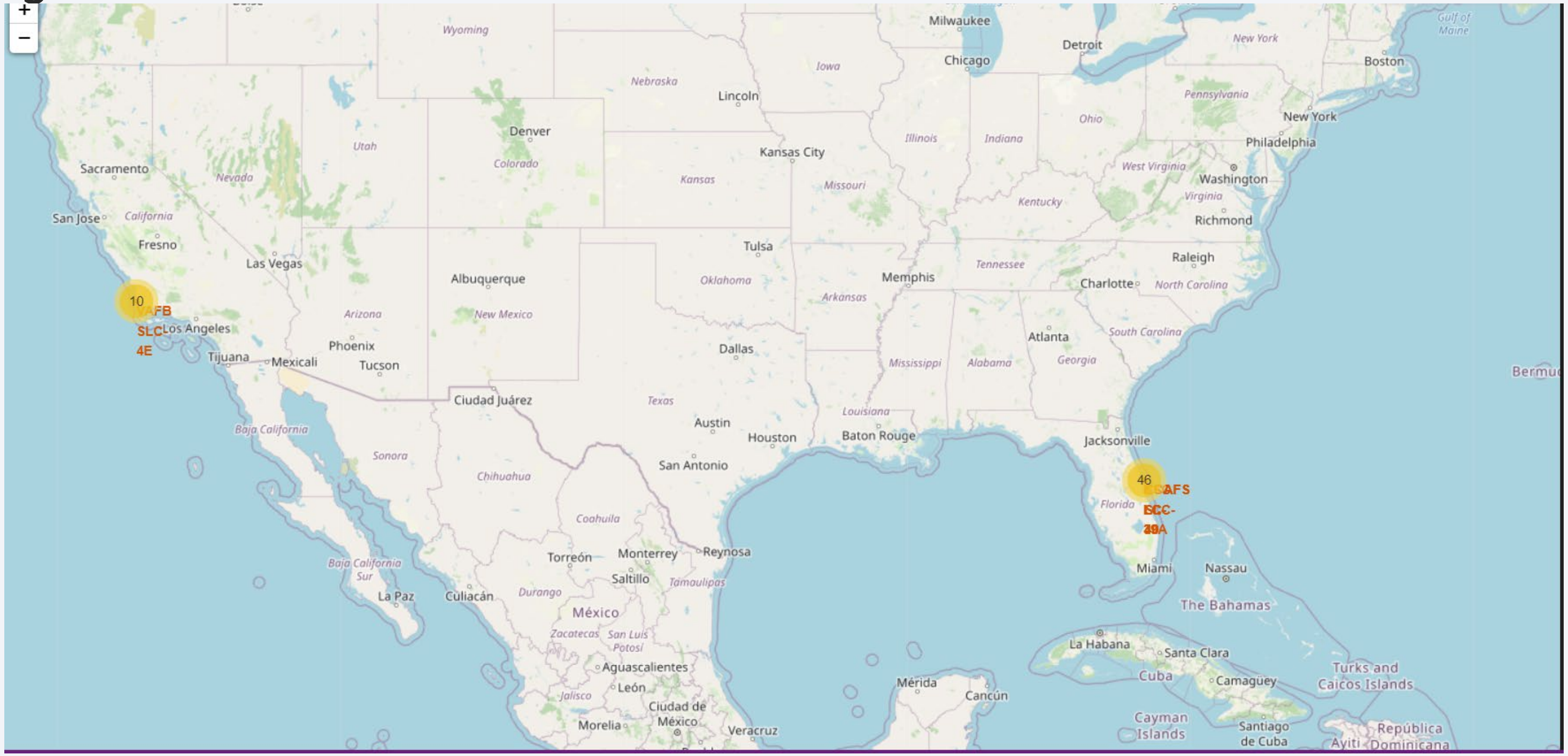
landing__outcome	outcome_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

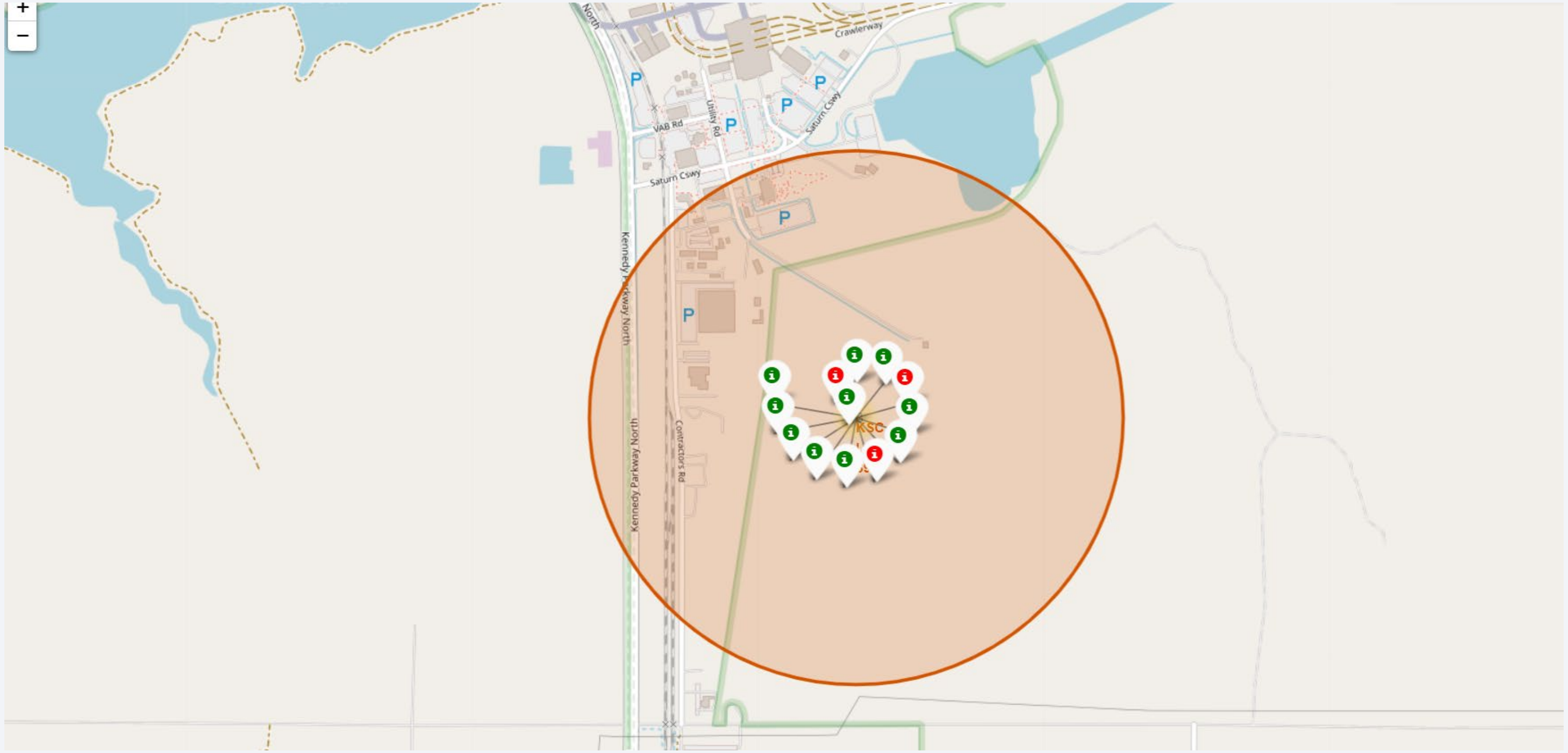
Folium: Launch Site Clusters

- Launch Site map shown on a national/global level that holds the map clusters housing the locations of all the launch sites



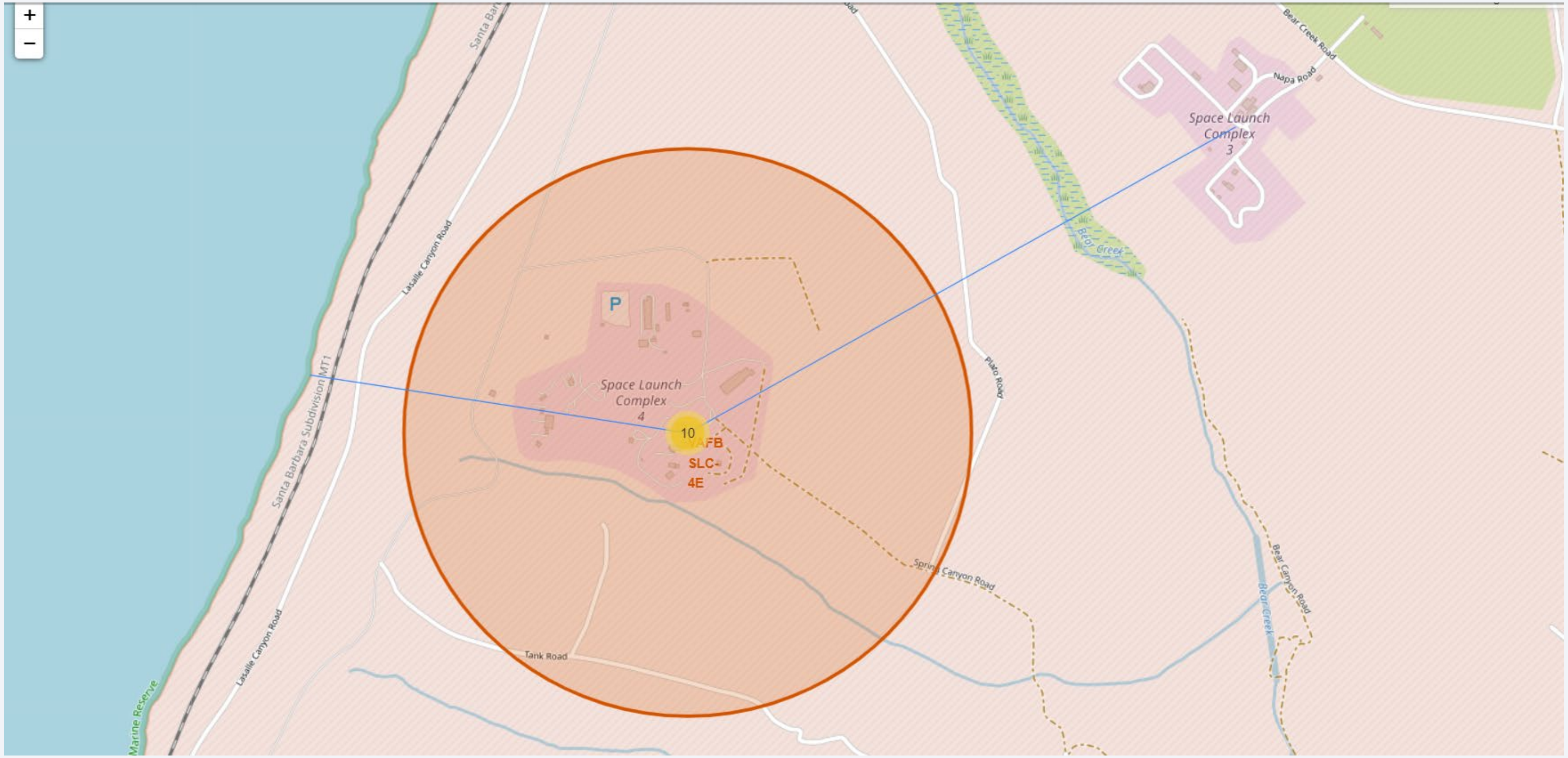
Folium: Color-coded markers

- Launch map, when clicked on, shows color-coded markers for each launch tied to the cluster, where green is success and red is failure



Folium: Features Map

- Folium map allows for lines to be drawn to determine distance between certain features

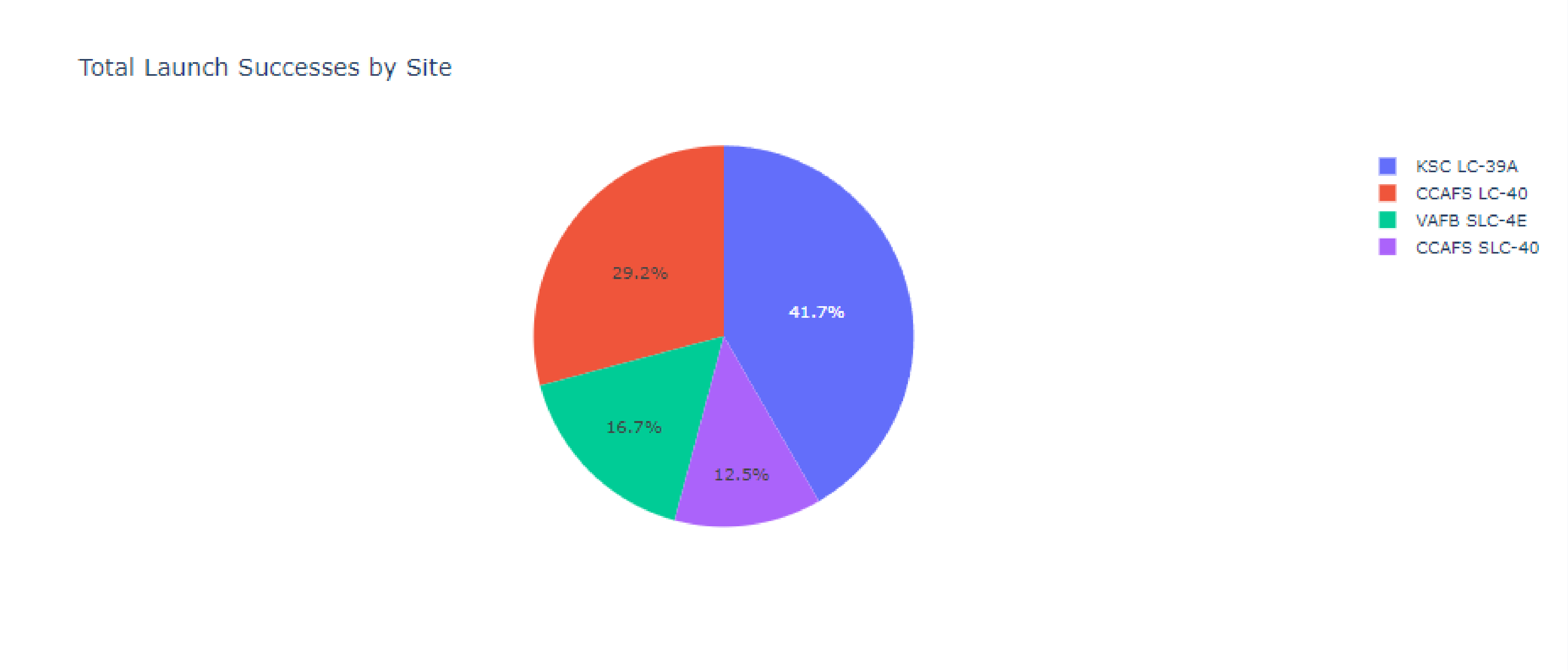


Section 4

Build a Dashboard with Plotly Dash

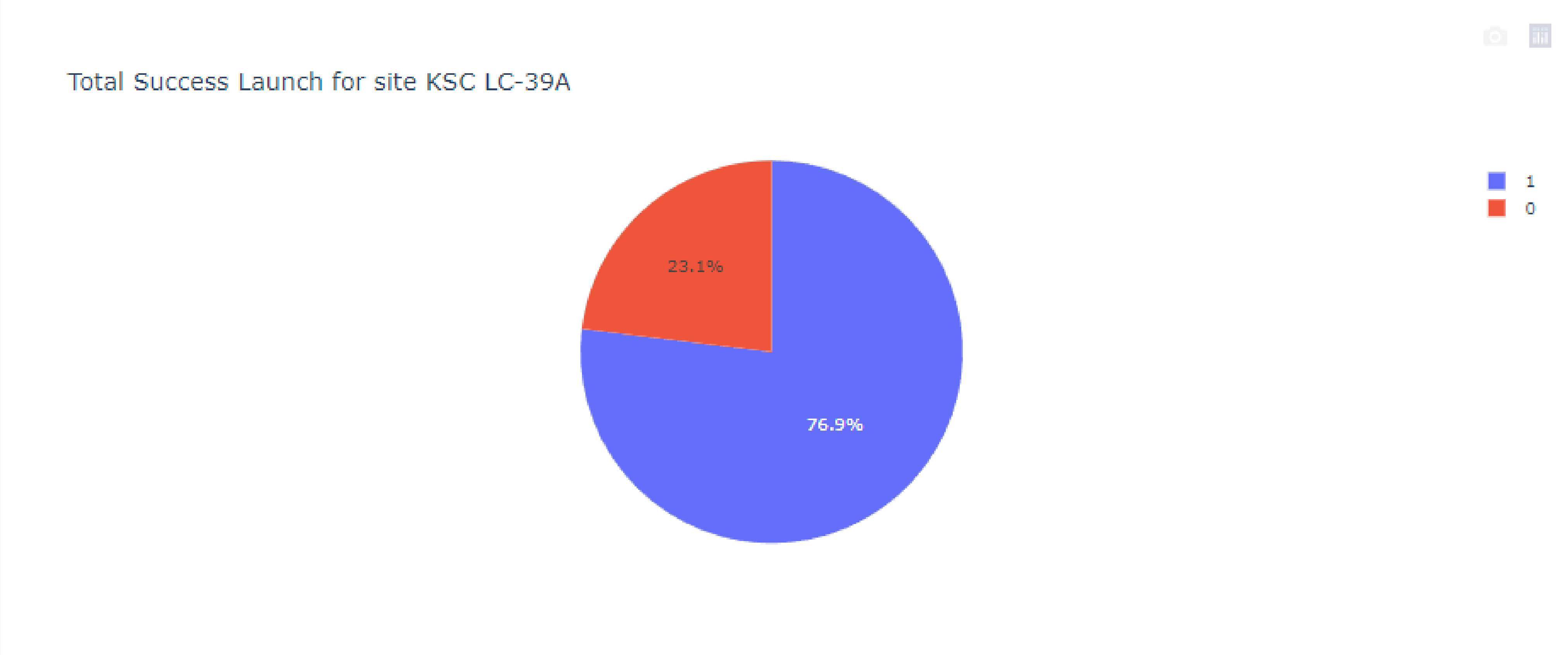
Pie Chart: All Sites

- Pie chart breaking down all the successful launches, divided by the different launch sites



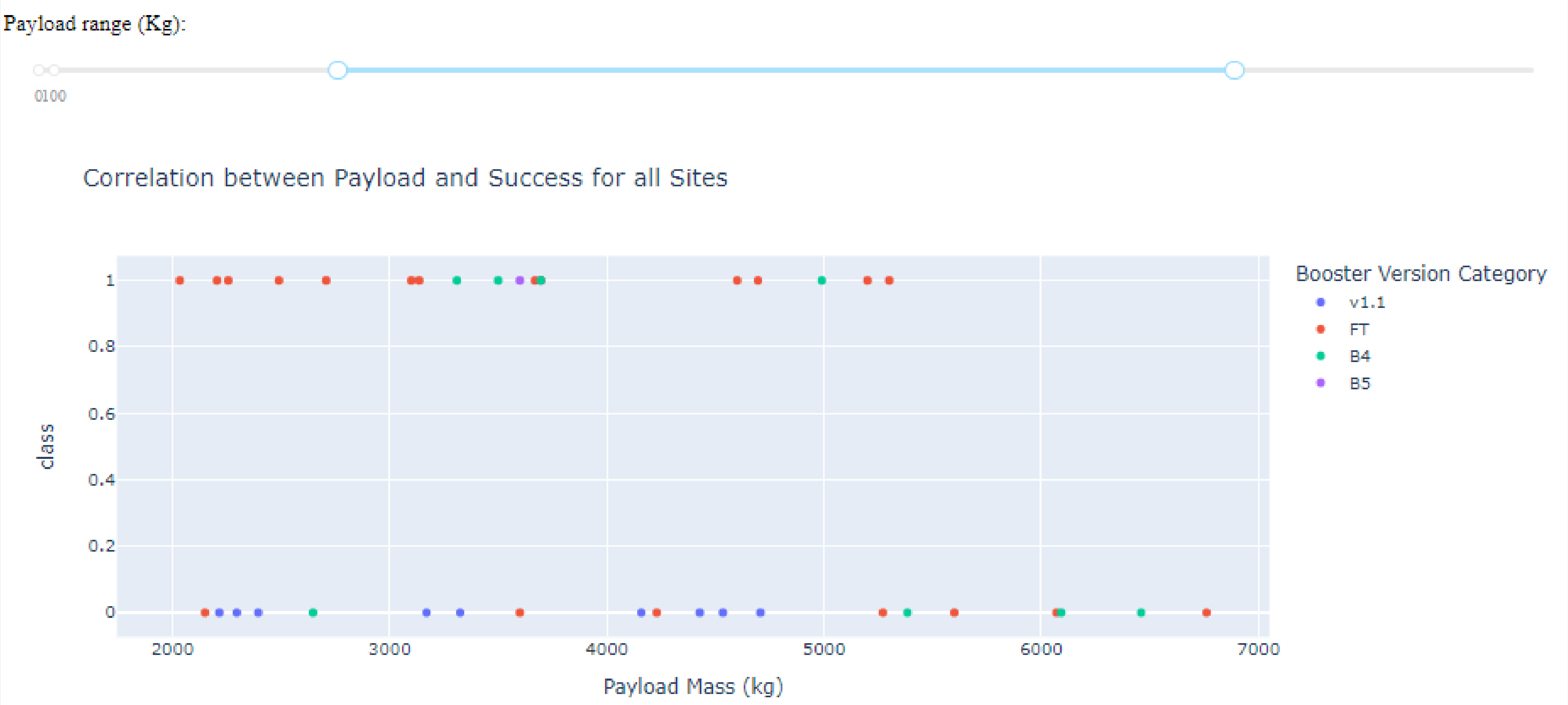
Pie Chart: KSC LC-39A

- Pie Chart demonstrating the total success vs. failure breakdown for KSC LC-39A, the highest percentage of success vs. failure



Scatter Plot: Payload vs. Launch Outcomes

- Payload vs. Launch Outcome scatter plot shows success (1) vs. failure (0) between all different Booster Version Categories between 2000-7000 kg

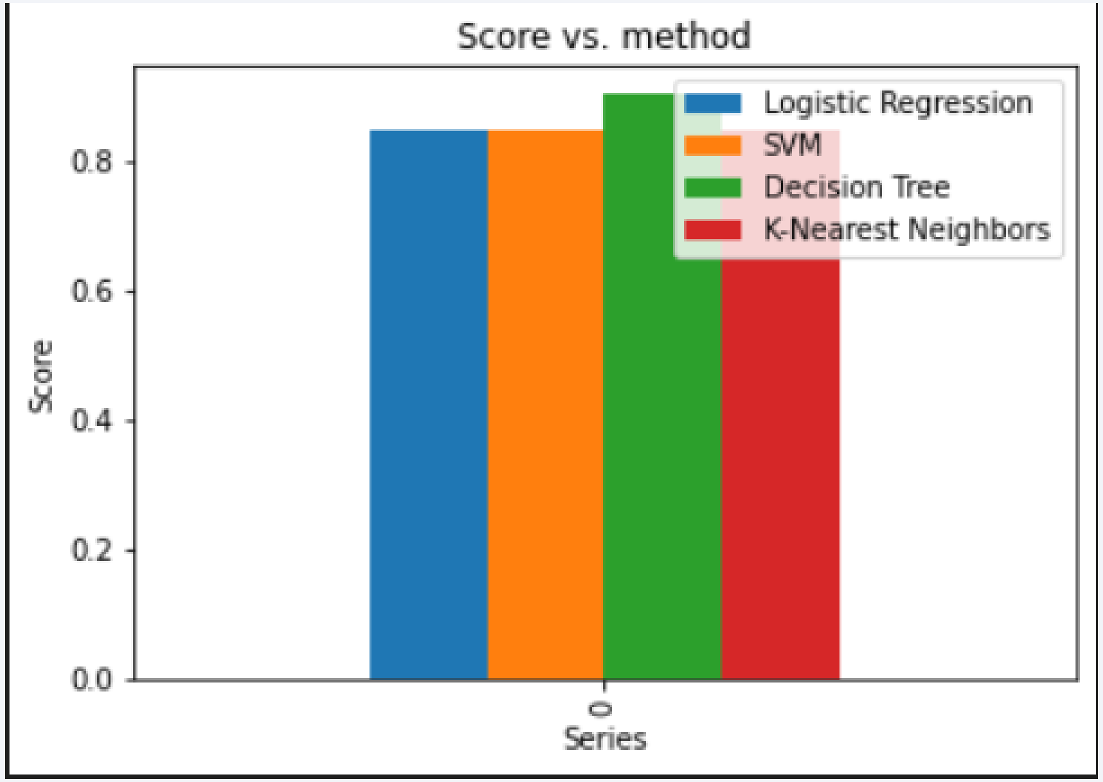


Section 5

Predictive Analysis (Classification)

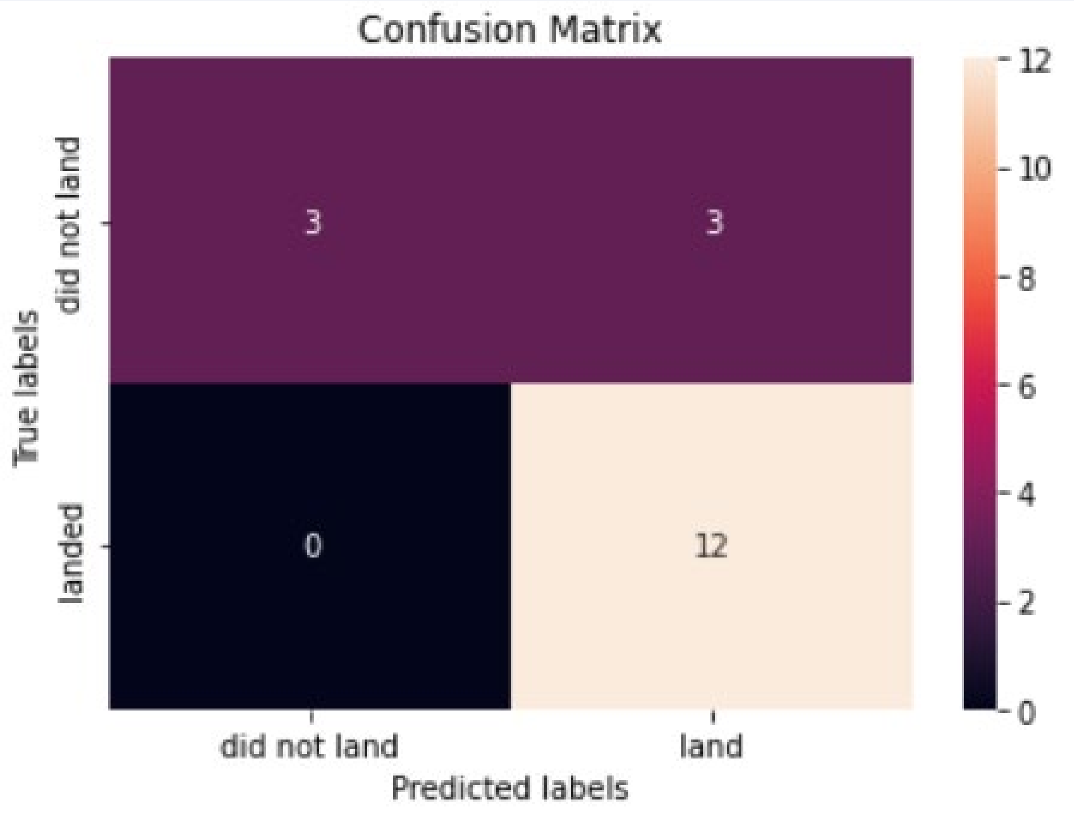
Classification Accuracy

- Based off accuracy scores for the four selected prediction methods, Decision Tree was the method with the highest score



Confusion Matrix

- Confusion Matrix for the "Decision Tree" method, which shows that "Landed" with true is largely accurate



Conclusions

- KSC LC-39A is the most successful launch site
- Decision tree is the most successful prediction method
- As flight numbers increase, the chance of success increases
- Payload Mass is generally successful around 10000 kg
- ES-L1, GEO, HEO and SSO are most successful orbits

Appendix

- Github Repository: https://github.com/erictwong18/IBM_SpaceX_Project

Thank you!