IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

David Heck
3/9/2023

Outline

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

Executive Summary

- Based on the analysis of prior SpaceX launch and landing data we are able to run predictive models to tell us which attributes of the process are important to the outcome. We utilized public SpaceX data, Python Visualization tools, and Python Machine Learning algorithms to arrive at our results. While many factors can contribute to a successful launch and landing, we found these to be optimum:
  - Site KSC LC-39A in Florida
  - Payload Mass <= 7000 kg
  - Booster Type FT
  - Low Earth Orbits

Introduction

- Project Overview
  - As a research effort for prospective aerospace companies interested in developing their space programs, we would like to look at the SpaceX Falcon 9 model. By being able to land and reuse their booster rockets, the cost savings for this venture is a big draw. By analyzing historical launch and landing data from SpaceX, we hope to find which attributes of the process are more likely to lead to a successful landing, and thus cost savings for the prospective aerospace company.
- Questions to be Answered
  - As mentioned, by analyzing prior launch/landing data we want to know what attributes of the process are likely to contribute to a successful landing. For example, are certain launch locations, payload mass or orbit type likely to lead to a successful landing? Let's find out!

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launch data pulled from SpaceX API via Python interface
- Perform data wrangling
  - Cleaned data using categorical means to replace missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Utilized data visualization tools and aggregate SQL functions to further analyze data
- Perform interactive visual analytics using Folium and Plotly Dash
  - Performed geographical analytics and interactive dash graphs to investigate data
- Perform predictive analysis using classification models
  - Ran various models to find the most accurate prediction of mission success/failure

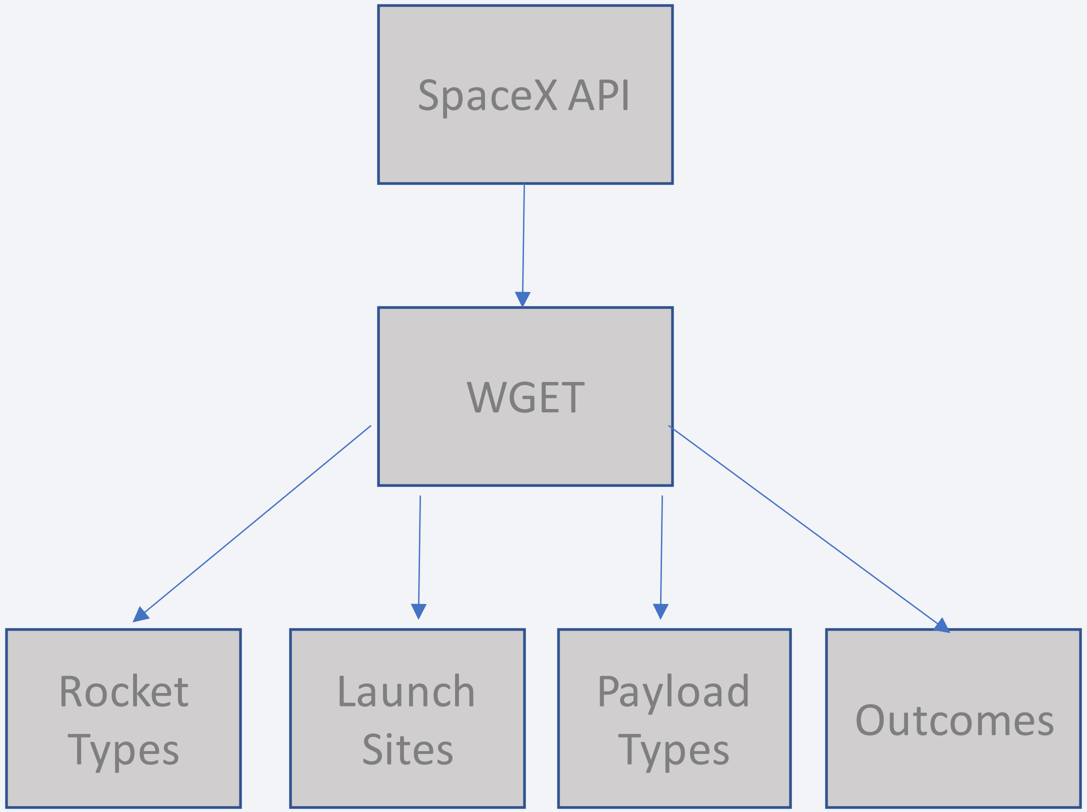
Data Collection

- Data was pulled from the SpaceX wiki page which consists of launch and landing data from prior missions. Additional data was pulled using the Python get command, making calls to the following SpaceX API URL's. Once downloaded the JSON objects are converted to usable Pandas data frames.
  - Rockets, Payloads, Launchpads, Outcome data

Data Collection – SpaceX API

- Data was pulled via the SpaceX API engine to gather launch/outcome data. The dataset was further enchance by grabbing data on the Launch Sites, Rockets Types and Payload Types

- https://github.com/dheck500/coursera-data-science-capstone/blob/master/labs-01-spacex-data-collection-api.ipynb



Data Collection - Scraping

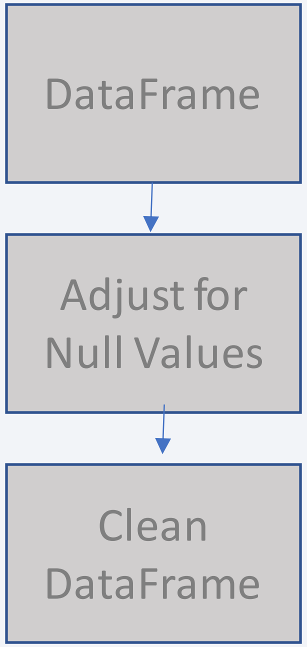
- Falcon 9 SpaceX data was pulled from the SpaceX Wiki page using WGET. Data was then reformatted using BeautifulSoup where HTML items could be parsed into Dataframes.

- https://github.com/dheck500/coursera-data-science-capstone/blob/master/labs-02-webscraping.ipynb

Data Wrangling

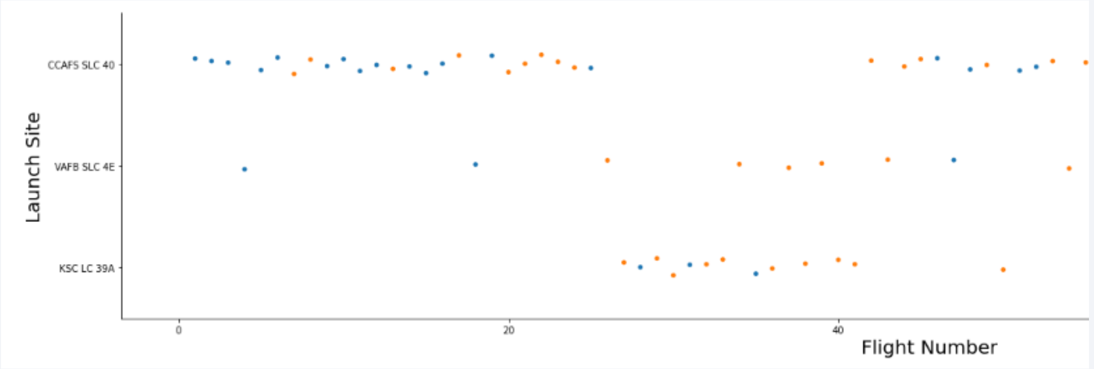
- Once imported we found that several attributes were blank. These were filled in by:
  - Average Payload Mass
  - Nan Values

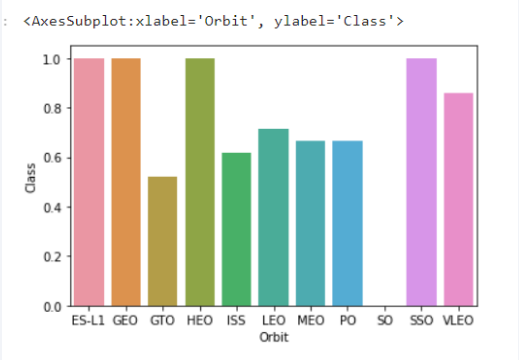
- https://github.com/dheck500/coursera-data-science-capstone/blob/master/labs-03-spacex-Data%20wrangling.ipynb

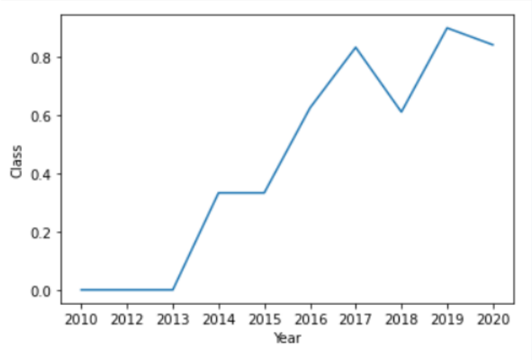


EDA with Data Visualization

- To analyze the data we looked at Scatter charts for Payload Mass, Flight Number, Launch Site.
- We also looked at mission success rates by Orbit Type and over time using bar and line graphs.
- https://github.com/dheck500/coursera-data-science-capstone/blob/master/labs-05-eda-dataviz.ipynb







EDA with SQL

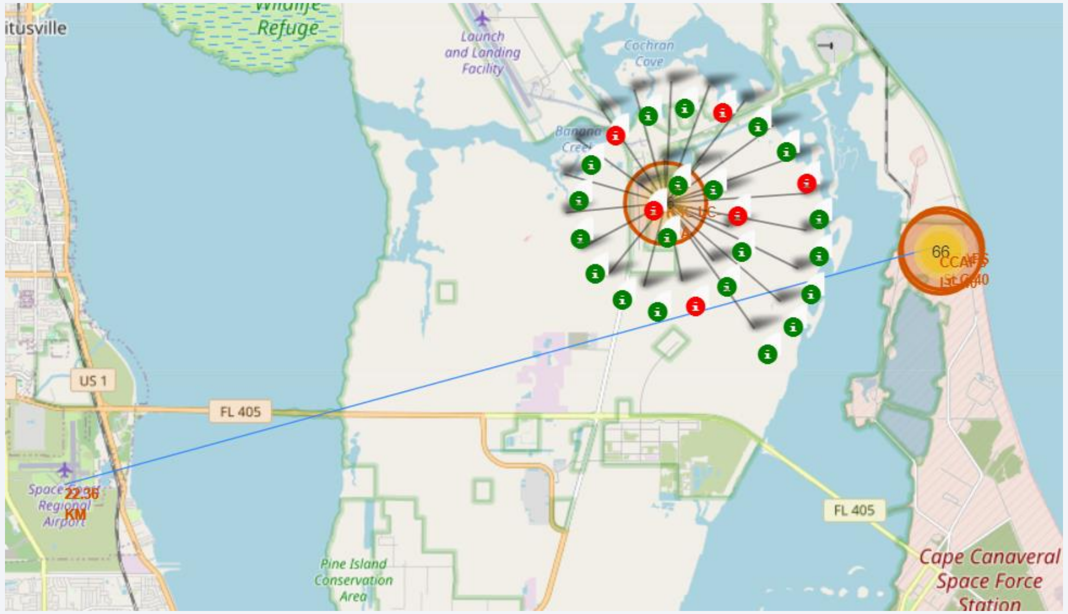
- Further Data Analysis was performed using SQL
  - Unique Names of Launch Sites
  - Average Payload Mass by Booster
  - Distribution of most commonly seen Outcomes

- https://github.com/dheck500/coursera-data-science-capstone/blob/master/labs-04-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

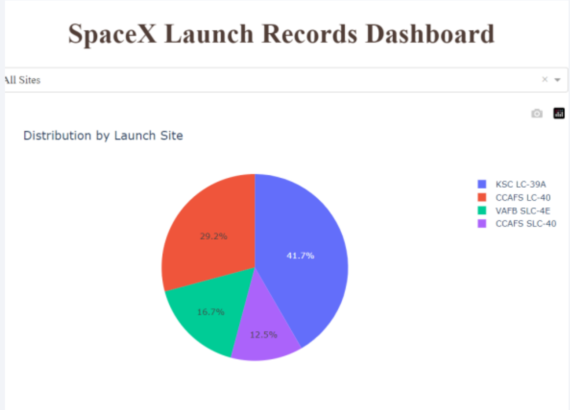
- Markers were added for each launch site. By incorporating the color coded outcomes we are able to see which sites are more successful.

- https://github.com/dheck500/coursera-data-science-capstone/blob/master/labs-06-launch-site-location.ipynb

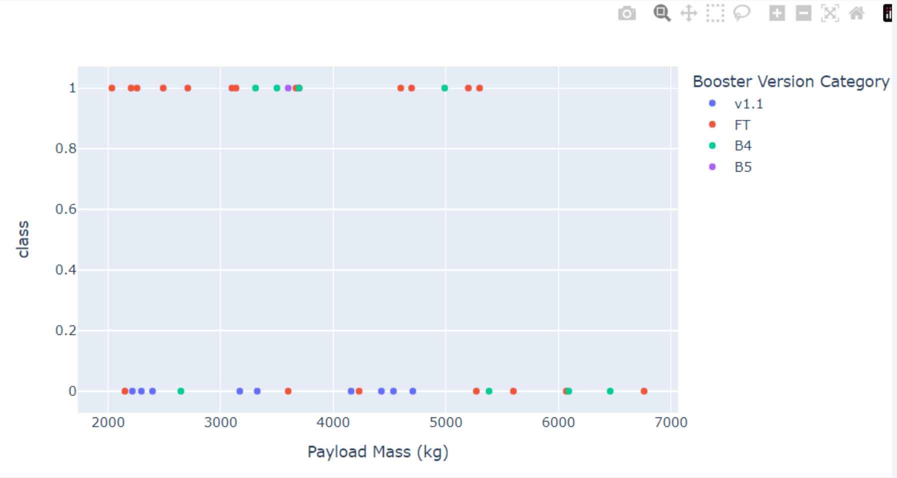


Build a Dashboard with Plotly Dash

- Using Plotly Dash we have an interactive set of graphs that we can further analyze outcome data by site.



- https://github.com/dheck500/coursera-data-science-capstone/blob/master/spacex_dash_app.py



Predictive Analysis (Classification)

- Using the launch data attributes and the binary classification of outcome 0 (failure) or 1 (success), we ran predictions using these models:
  - Logistical Regression
  - SVM
  - Decision Tree
  - K-Nearest Neighbor

- https://github.com/dheck500/coursera-data-science-capstone/blob/master/labs-07-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

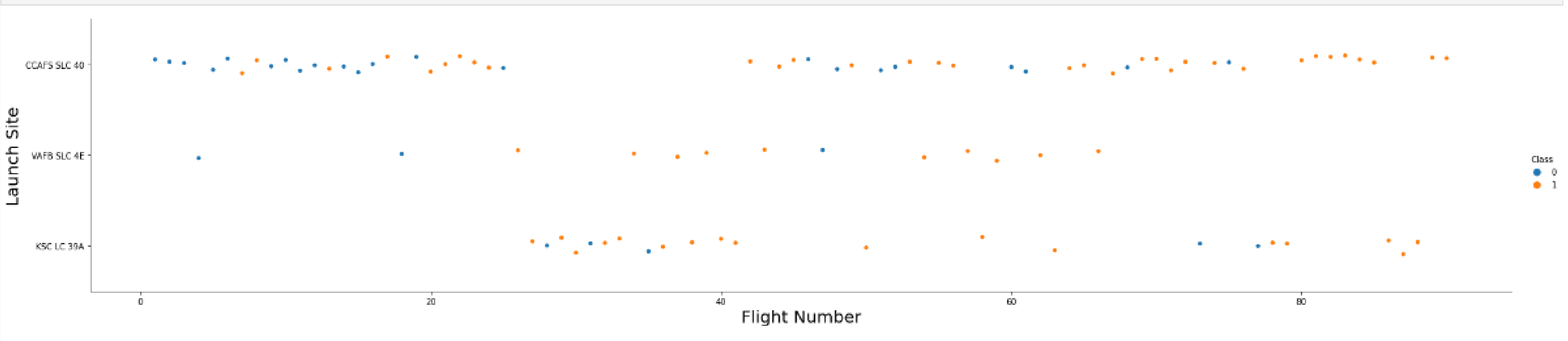
Results

- Based on the analysis we found that results are more favorable for success when:
  - Site KSC LC-39A in Florida
  - Payload Mass <= 7000 kg
  - Booster Type FT
  - Low Earth Orbits
- For the models we analyzed we found the Decision Tree model returned the best accuracy for predicting outcomes.

Section 2

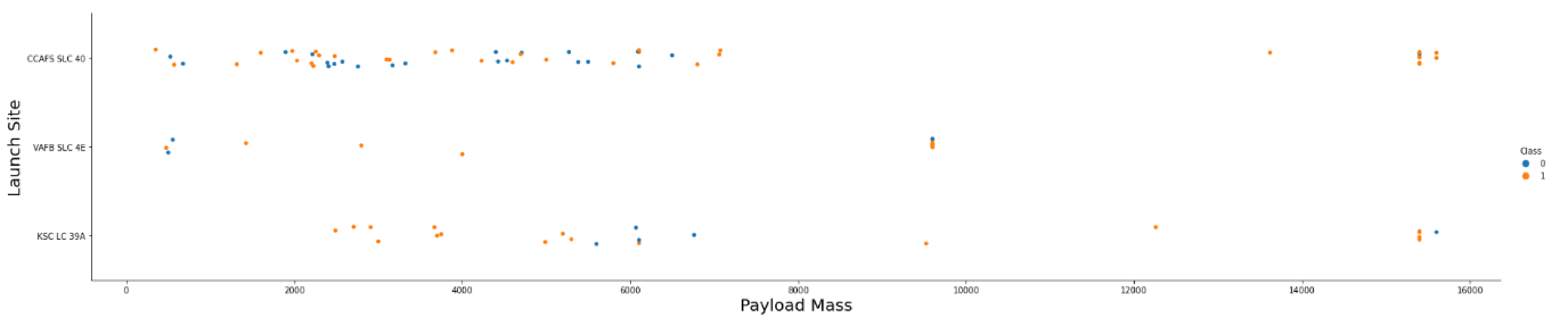
Insights drawn from EDA

Flight Number vs. Launch Site



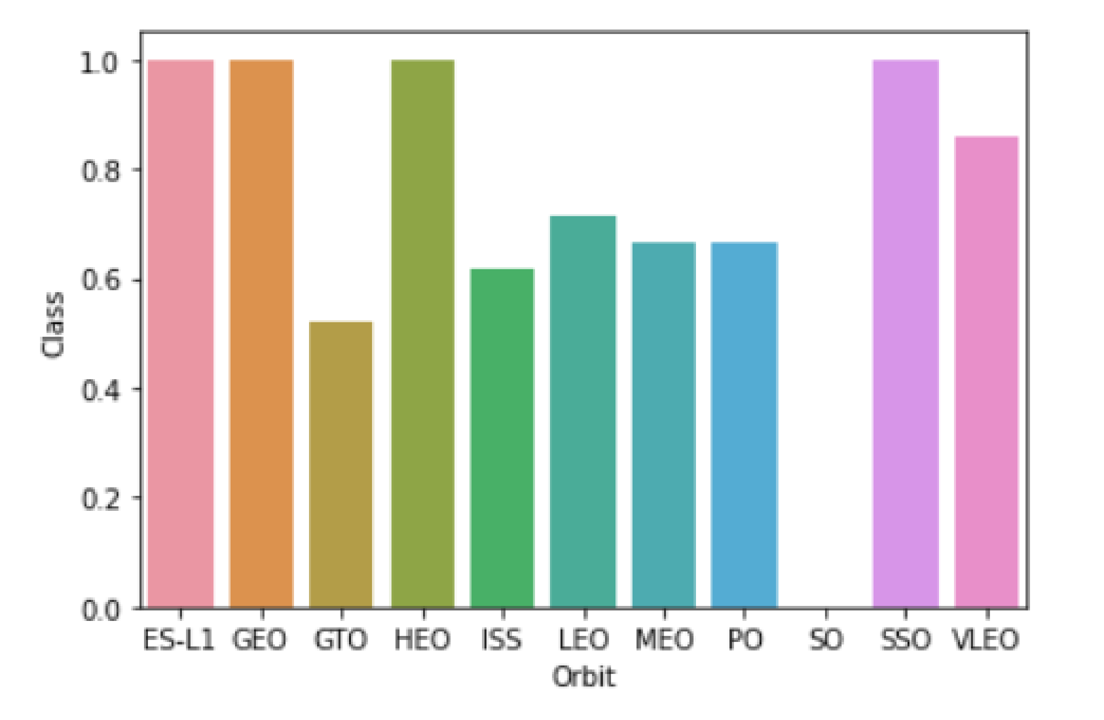
Based on the scatter plot above, we can see that SpaceX launches were more successful the further along in the program they got, regardless of the launch site. We can also see that site CCAFS SLC 40 was the most commonly utilized.

Payload vs. Launch Site



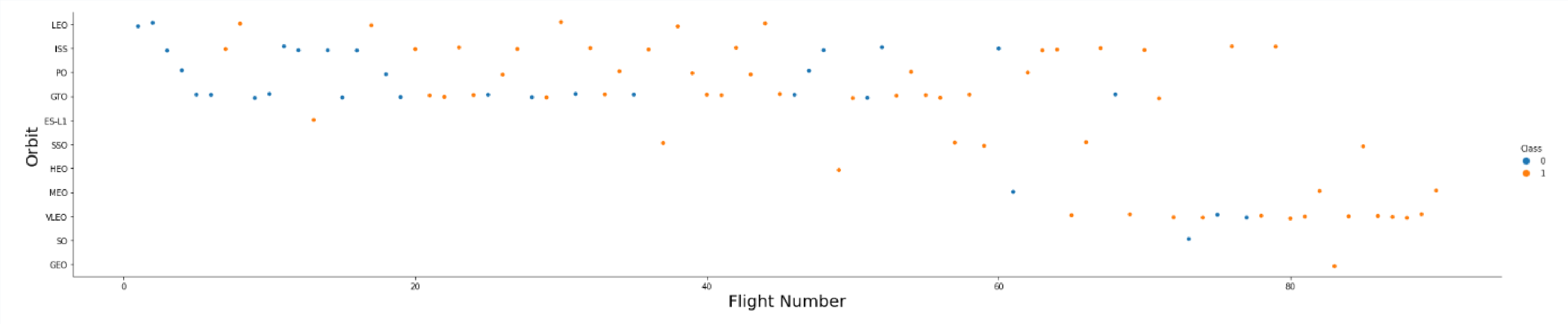
Based on the scatter plot above, we can see that SpaceX launches typically have a payload mass less than 7000 kg. The outcomes under this threshold have shown mixed results. It appears that flights with a higher payload, while not as common, have shown to be more successful, across all launch sites.

Success Rate vs. Orbit Type



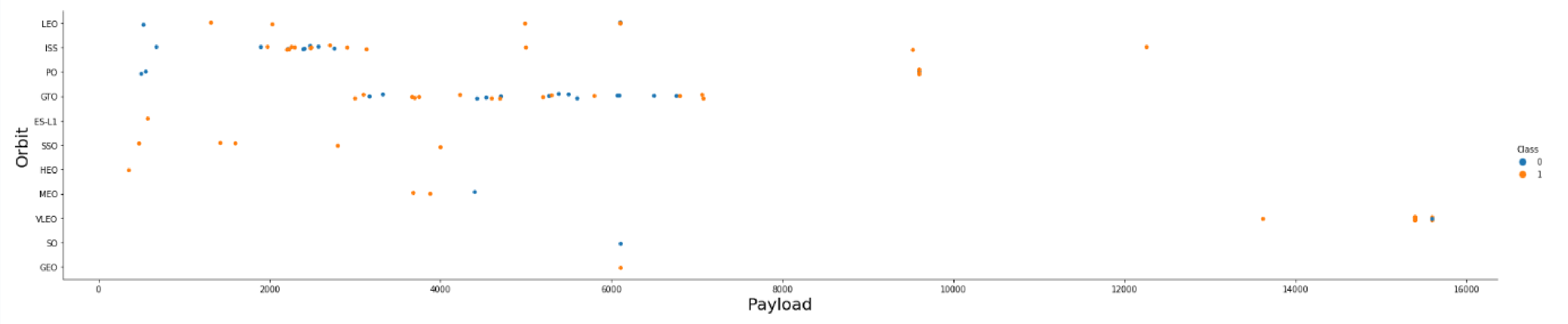
Based on the bar graph above, we can see which SpaceX orbit types are more likely to be successful. On average the lower distance orbits tend to be less successful.

Flight Number vs. Orbit Type



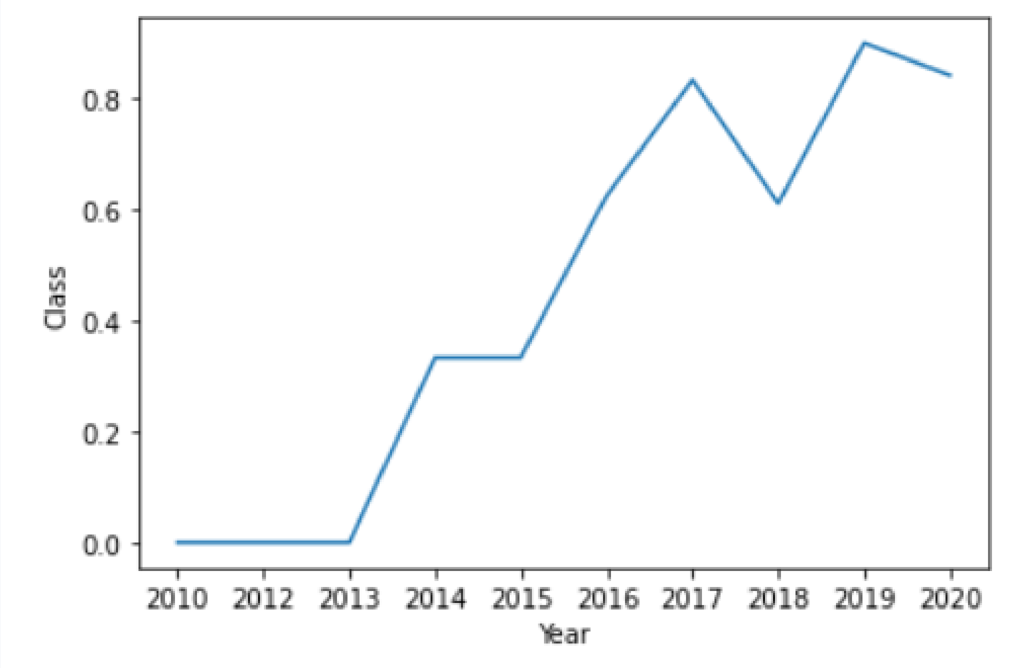
Based on the scatter plot above, we can see which SpaceX orbit types are more likely to be successful as the program advanced. We can see most of the earlier mission focused on low earth orbits. More recently many of the missions have been very low earth orbits which have been very successful.

Payload vs. Orbit Type



Based on the scatter plot above, we can see that payloads under 7000 kg and low earth orbits are more frequently used, with mixed outcomes. The higher payloads, while more rarely used do appear to be more successful.

Launch Success Yearly Trend



Based on the line graph above, we can that the success rate of the program has risen over the years, despite a dip in performance in 2018.

All Launch Site Names

```
sql = "select distinct launch_site from spacextbl order by launch_site"
df = pd.read_sql(sql, conn)
df
```

	LAUNCH_SITE
0	CCAFS LC-40
1	CCAFS SLC-40
2	KSC LC-39A
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
sql = "select * from spacextbl where launch_site like 'CCA%' limit 5;"
df = pd.read_sql(sql, conn)
df
```

	DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD	PAYLOAD_MASS__KG_	ORBIT	CUSTOMER	MISSION_OUTCOME	LANDING__OUTCOME
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	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	2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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 = "select customer,sum(PAYLOAD_MASS__KG_) payload_mass from spacextbl where customer='NASA (CRS)' group by customer"
df = pd.read_sql(sql, conn)
df
```

	CUSTOMER	PAYLOAD_MASS
0	NASA (CRS)	45596

Average Payload Mass by F9 v1.1

```
sql = """
select booster_version,avg(PAYLOAD_MASS__KG_) payload_mass
from spacextbl where booster_version='F9 v1.1' group by booster_version
"""
df = pd.read_sql(sql, conn)
df
```

	BOOSTER_VERSION	PAYLOAD_MASS
0	F9 v1.1	2928

First Successful Ground Landing Date

```
sql = """
select count(*),min(date)
from spacextbl
where landing__outcome='Success (ground pad)'
"""
df = pd.read_sql(sql, conn)
df
```

	1	2
0	9	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
sql = """
select distinct booster_version
from spacextbl
where landing__outcome='Success (drone ship)' and
PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
"""
df = pd.read_sql(sql, conn)
df
```

	BOOSTER_VERSION
0	F9 FT B1021.2
1	F9 FT B1031.2
2	F9 FT B1022
3	F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

```
sql = """
select mission_outcome,count(*)
from spacextbl
group by mission_outcome
"""
df = pd.read_sql(sql, conn)
df
```

	MISSION_OUTCOME	2
0	Failure (in flight)	1
1	Success	99
2	Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
sql = """
select distinct booster_version from spacextbl
where PAYLOAD_MASS__KG_ in (
select max(PAYLOAD_MASS__KG_)
from spacextbl)
"""
df = pd.read_sql(sql, conn)
df
```

	BOOSTER_VERSION
0	F9 B5 B1048.4
1	F9 B5 B1048.5
2	F9 B5 B1049.4
3	F9 B5 B1049.5
4	F9 B5 B1049.7
5	F9 B5 B1051.3
6	F9 B5 B1051.4
7	F9 B5 B1051.6
8	F9 B5 B1056.4
9	F9 B5 B1058.3
10	F9 B5 B1060.2
11	F9 B5 B1060.3

2015 Launch Records

```python
sql = """
select landing__outcome,booster_version, launch_site, date
from spacextbl
where landing__outcome in ('Failure (drone ship)') and
date between '2015-01-01' and '2015-12-31'
"""
df = pd.read_sql(sql, conn)
df
```

	LANDING__OUTCOME	BOOSTER_VERSION	LAUNCH_SITE	DATE
0	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
1	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

```
sql = """
select landing__outcome,count(*)
from spacextbl
where date between '2010-06-04' and '2017-03-20'
group by landing__outcome
order by count(*) desc
"""
df = pd.read_sql(sql, conn)
df
```

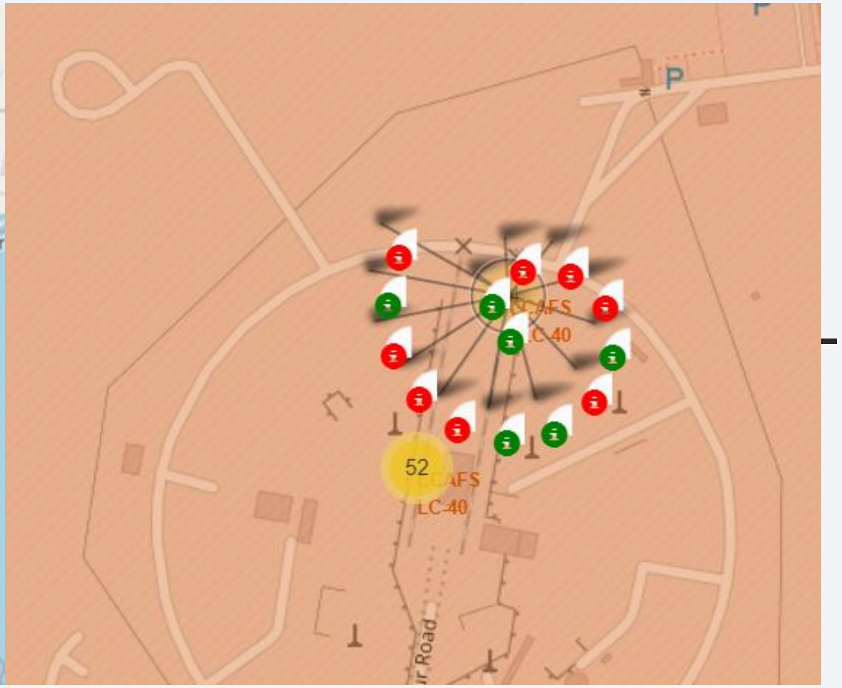
	LANDING__OUTCOME	2
0	No attempt	10
1	Failure (drone ship)	5
2	Success (drone ship)	5
3	Controlled (ocean)	3
4	Success (ground pad)	3
5	Failure (parachute)	2
6	Uncontrolled (ocean)	2
7	Precluded (drone ship)	1

Section 3

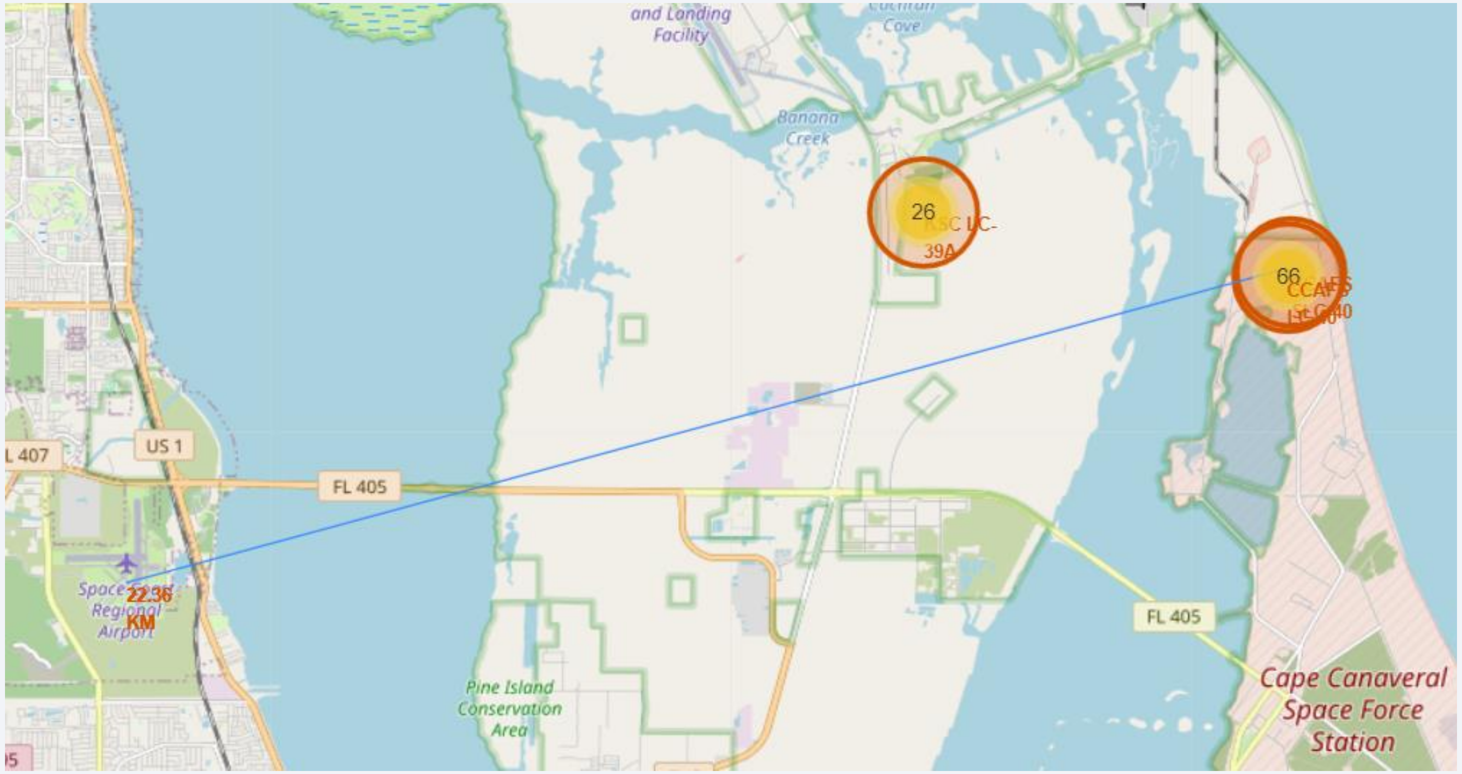
Launch Sites Proximities Analysis

Folium Map of Launch Sites

Folium Interactive Map with Site Outcomes



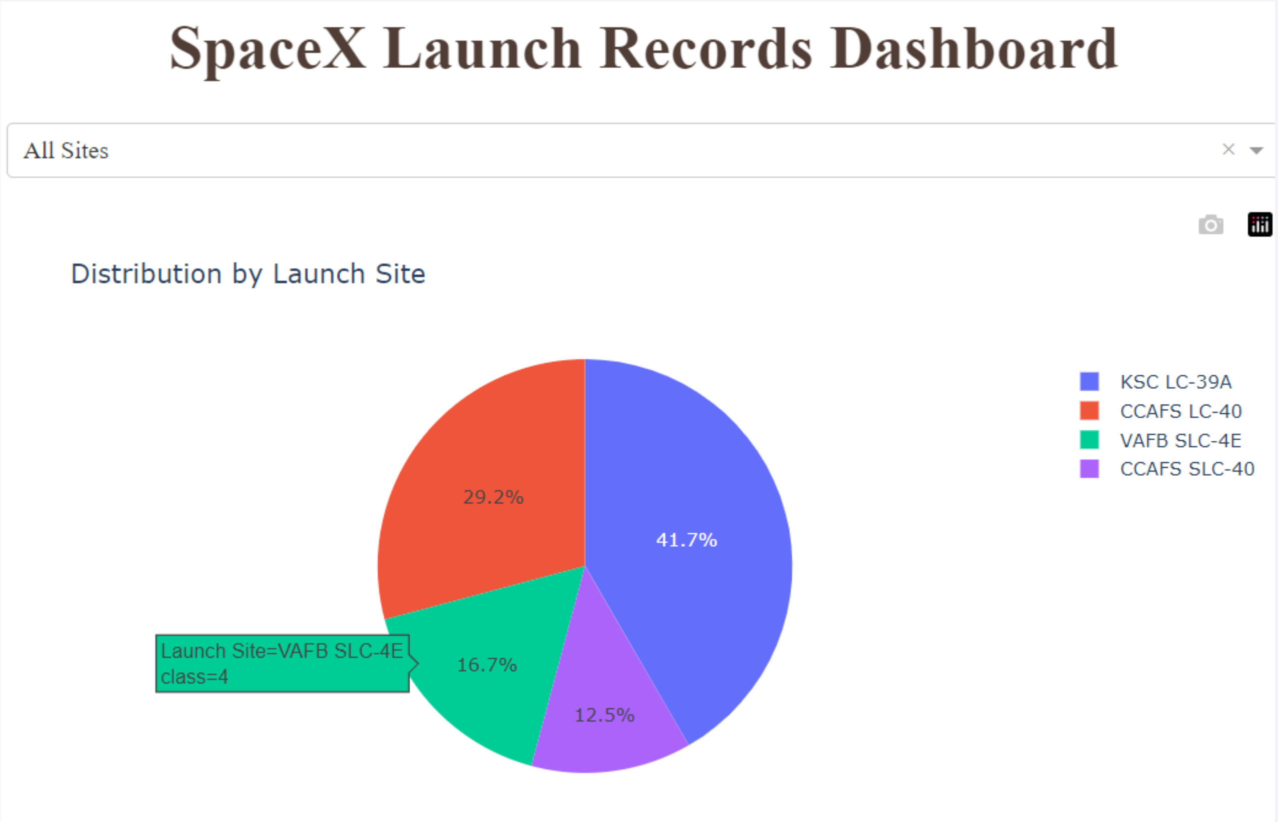
Folium Map of Launch Sites – Distance to Landmarks



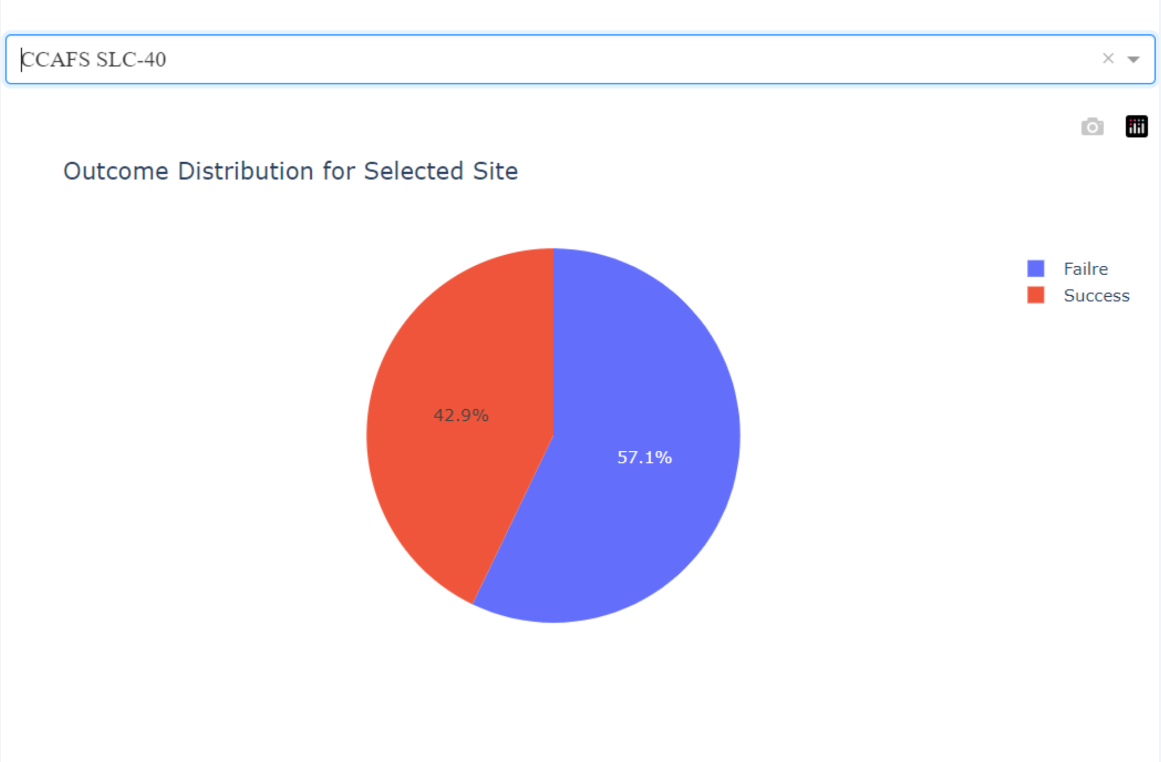
Section 4

Build a Dashboard with Plotly Dash

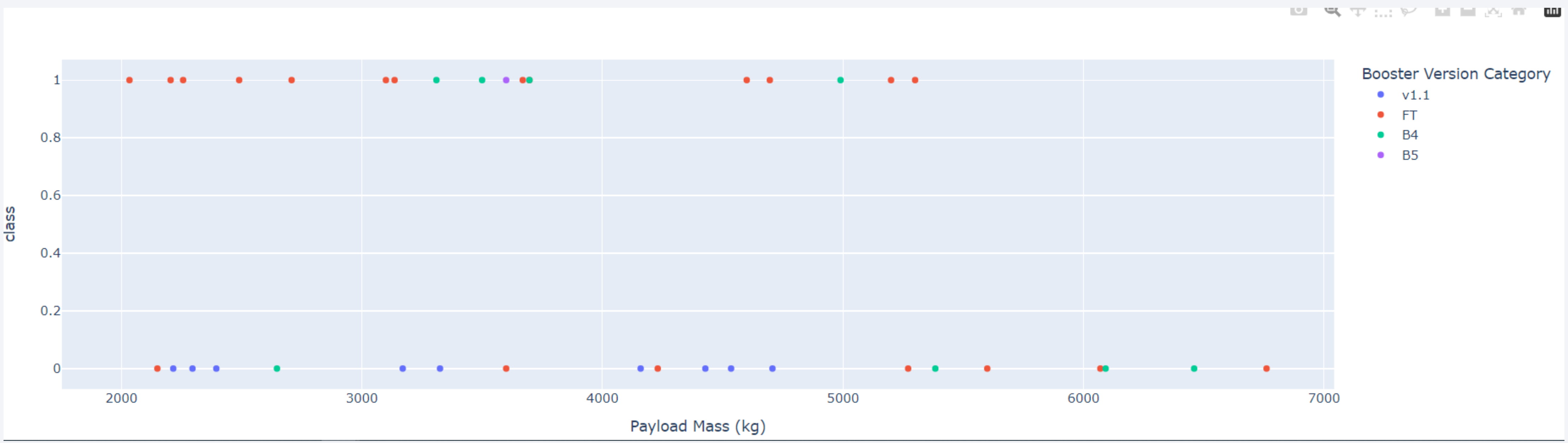
Dash - Distribution of Missions by Launch Site



Dash – Success Rate by Launch Site



Dash – Outcomes by Payload and Booster

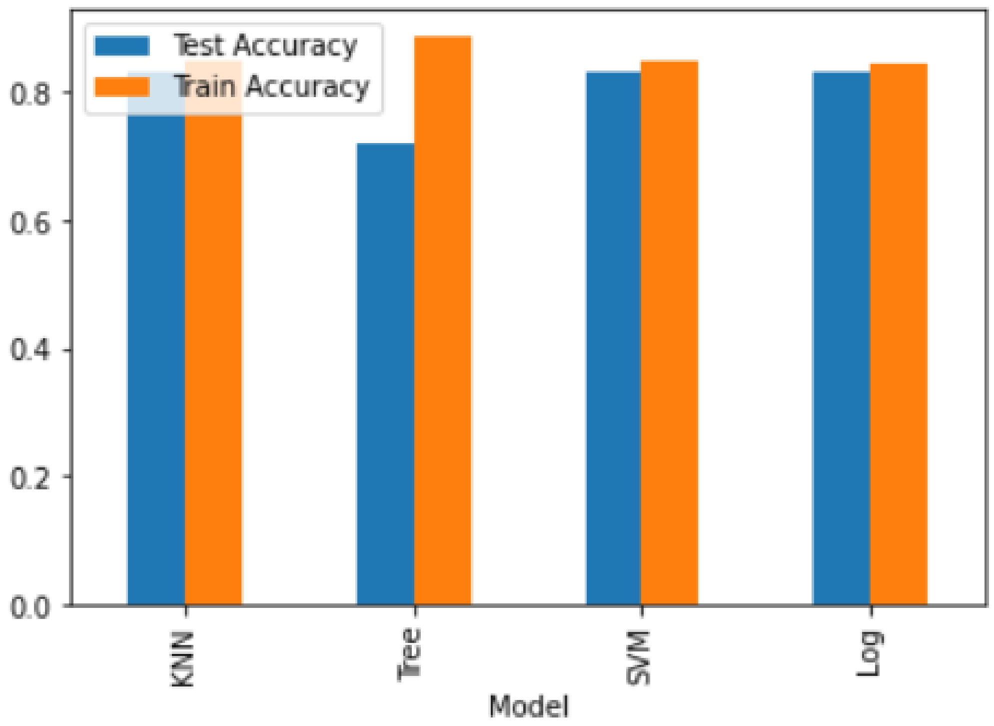


Section 5

Predictive Analysis (Classification)

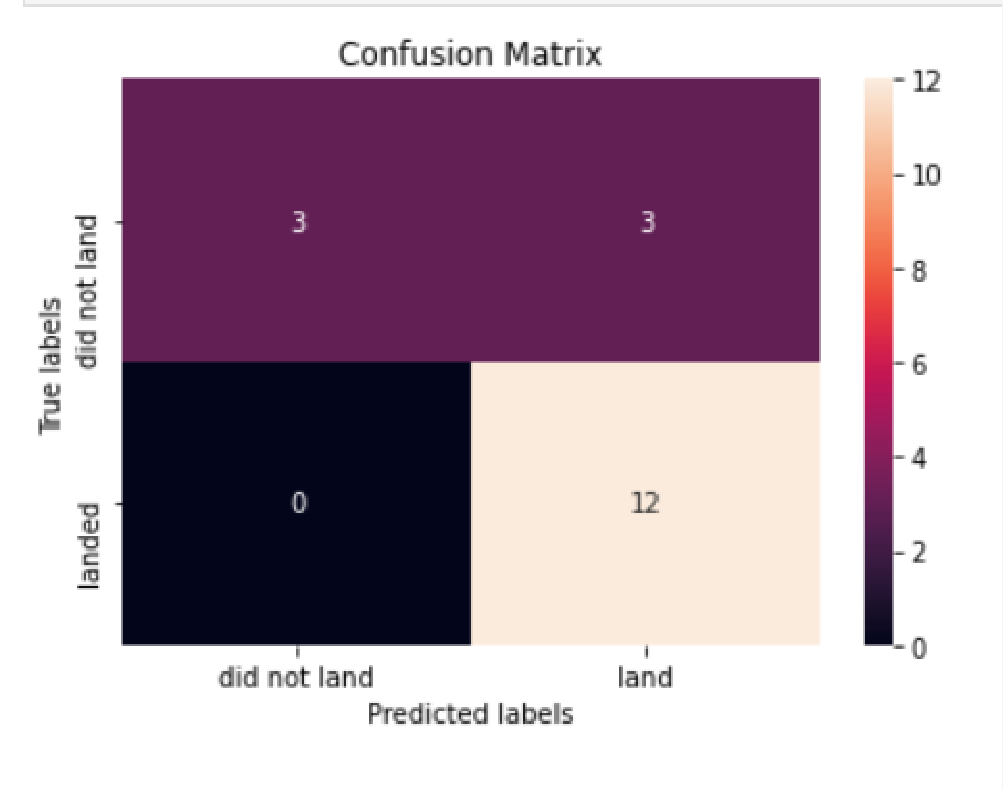
Classification Accuracy

<AxesSubplot:xlabel='Model'>



Outcomes of the 4 Machine Learning models we tested. All did quite well, but the Decision Tree model was most effective.

Confusion Matrix



The results of running our test set thru the Decision Tree model which was the most effective of the 4 models we analyzed.

Conclusions

- Again while many factors can lead to the success of a mission, we found that results were most favorable with:
    - Site KSC LC-39A in Florida
    - Payload Mass <= 7000 kg
    - Booster Type FT
    - Low Earth Orbits

Appendix

- https://github.com/dheck500/coursera-data-science-capstone
- https://api.spacexdata.com/v4/launches/past
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Thank you!