IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Min Chae
25 June 2022

Outline

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

Executive Summary

- Summary of methodologies
- Using exploratory analysis and machine learning, we can conclude the most prominent factors in predicting whether or not a mission's rocket will succeed is the flight number and that a decision tree is the best classifier for determining whether or not SpaceX will be able to reuse a rocket for a future mission.

Introduction

- This project explores the many factors that dictate whether or not SpaceX launches will be able to reuse rockets for further flights in order to keep costs low
- This project wants to understand what factors best predict a successful rocket launch mission and what model can be used to predict future outcomes

Section 1

Methodology

Methodology

Executive Summary

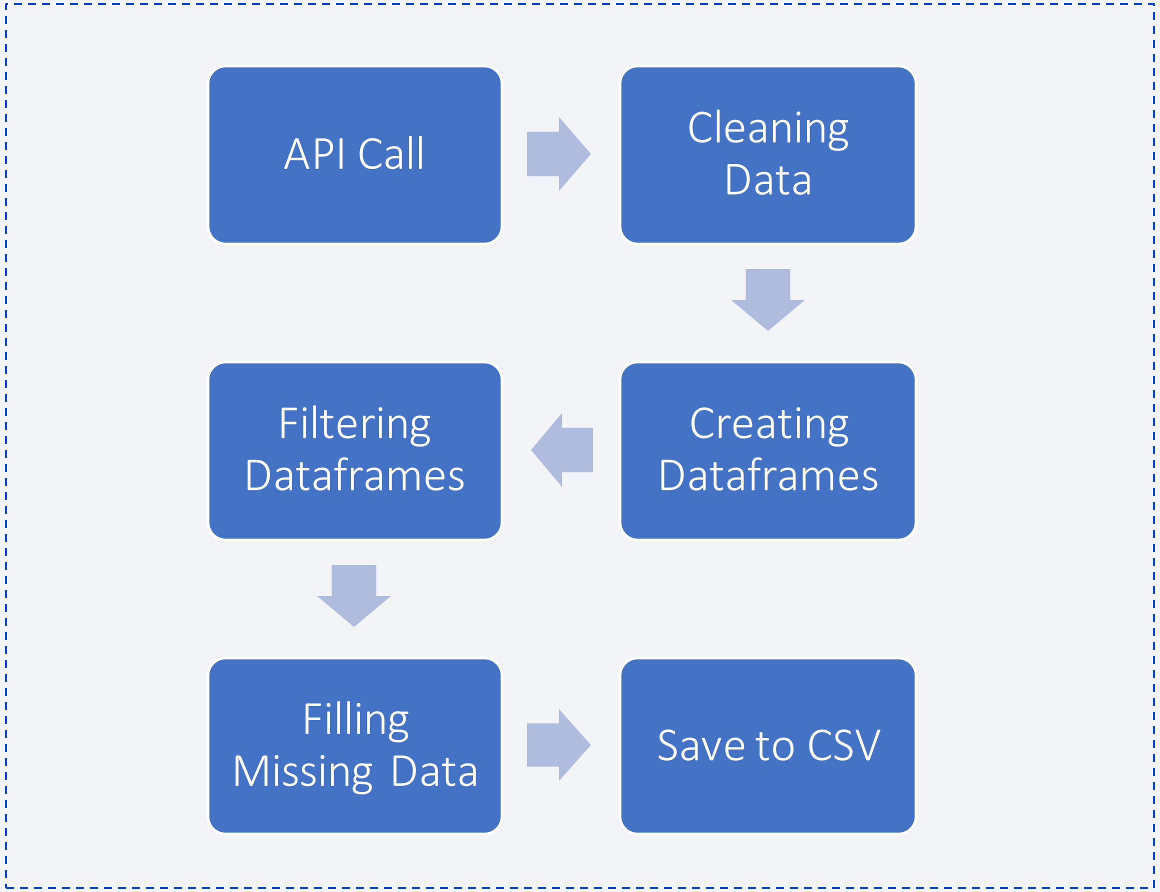
- Data collection methodology:
  - Data was collected through web scraping and SpaceX API calls
- Perform data wrangling
  - Incomplete or missing values were retrofitted with appropriate values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building classification models and their evaluations

Data Collection

- The data was obtained through two methods:
  - API calls
  - Web scraping

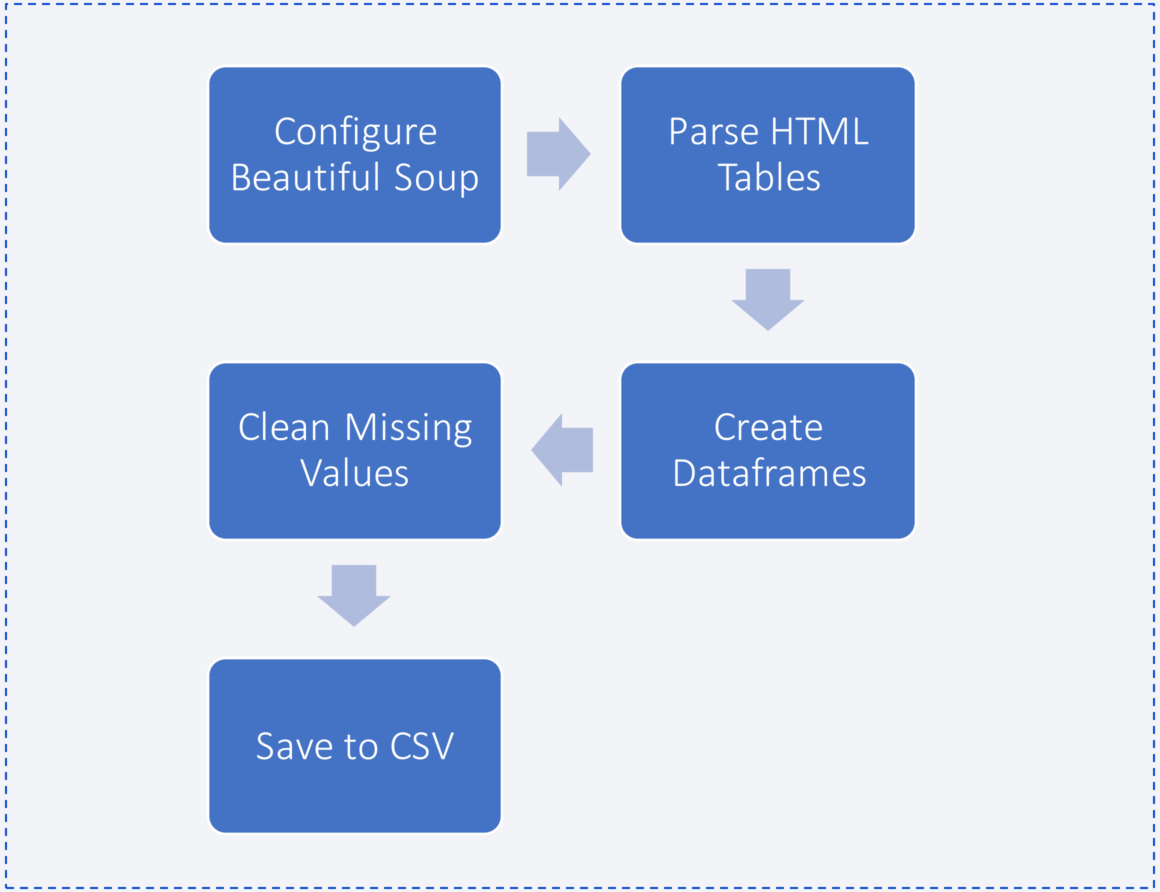
Data Collection – SpaceX API

- [ibm-data_sci/api_and_wrangling.ipynb at main · Minwook-Chae/ibm-data_sci (github.com)]



Data Collection - Scraping

- ibm-data_sci/webscraping.ipynb at main · Minwook-Chae/ibm-data_sci (github.com)



Data Wrangling

- Here, I added a one-hot encoding column corresponding to whether or not the landing was successful
- 1 was a success
- 0 was a failure
- ibm-data_sci/data_wrangling.ipynb at main · Minwook-Chae/ibm-data_sci (github.com)

EDA with Data Visualization

- Early data analysis measured metric such as how the payload masses, preferred launch sites, orbit types, and success rate changed overtime
- Interestingly, payloads with larger masses were more successful than lighter payloads
- ibm-data_sci/eda_data_visualization.ipynb at main · Minwook-Chae/ibm-data_sci (github.com)

EDA with SQL

- SQL queries include:
  - Selecting unique launch sites
  - Total payload masses
  - Counts of specific boosters
  - Counts of successes and failures
  - Sorting missions based on metrics such as weights and dates
- ibm-data_sci/sql_exploratory.ipynb at main · Minwook-Chae/ibm-data_sci (github.com)

Build an Interactive Map with Folium

- Using Folium, markers such as lines, circles, and text were added
- These were useful for labeling and measuring distances
- ibm-data_sci/launch_sites_geo.ipynb at main · Minwook-Chae/ibm-data_sci (github.com)

Build a Dashboard with Plotly Dash

- First, there is a pie chart featuring the distribution of launches between certain launch sites
- Also the pie chart breaks down specific launch sites to give their rates of success
- Then, there is a scatter plot showing the numerous launches and their outcomes, which can be filtered by payload sizes and specific launch sites
- ibm-data_sci/spacex_dash_app.py at main · Minwook-Chae/ibm-data_sci (github.com)

Predictive Analysis (Classification)

- After scaling the data, GridSearchCV was used to find the best parameters between different types of models
- From there, the results of each model was plotted in a confusion matrix and the accuracy was recorded
- Based on the results, the best performing model was a decision tree with an accuracy of up to 94%

- [ibm-data_sci/ml_predictions.ipynb at main · Minwook-Chae/ibm-data_sci (github.com)]

Results

- From EDA, we can see landings become more successful overtime

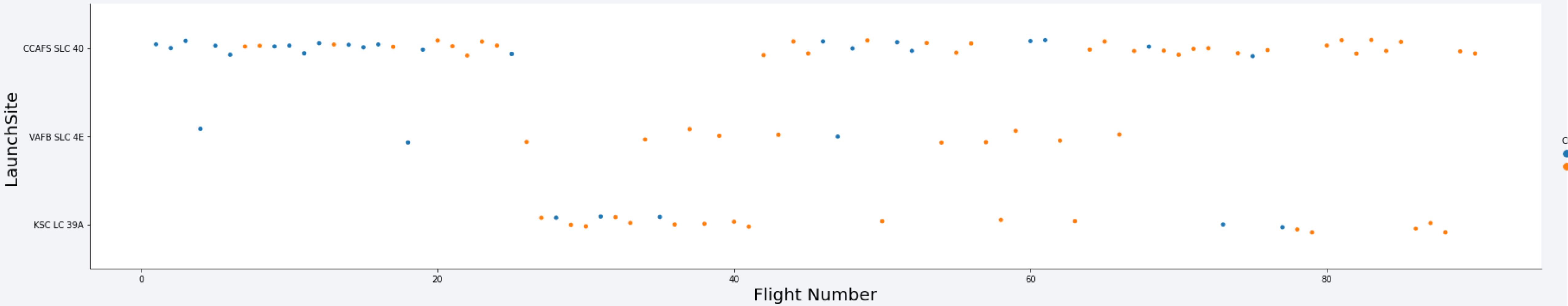
- Decision Trees are the best model to predict successes

Section 2

Insights drawn from EDA

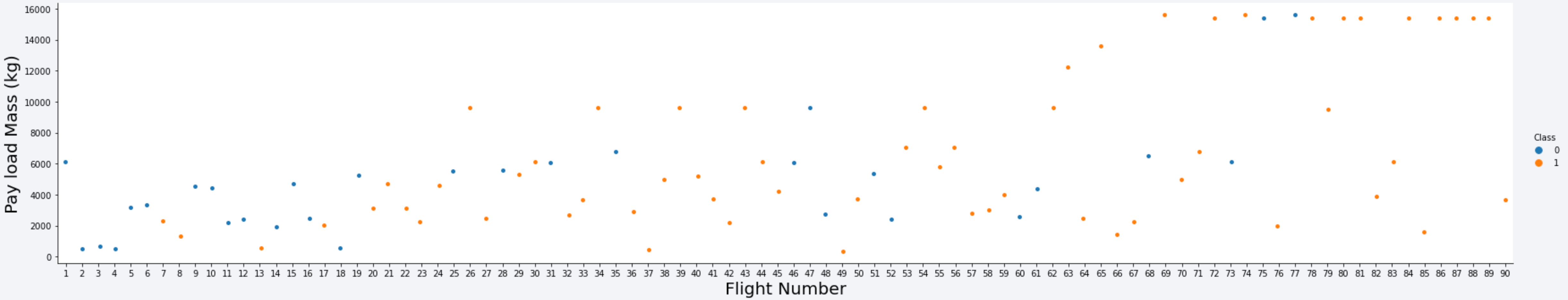
Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



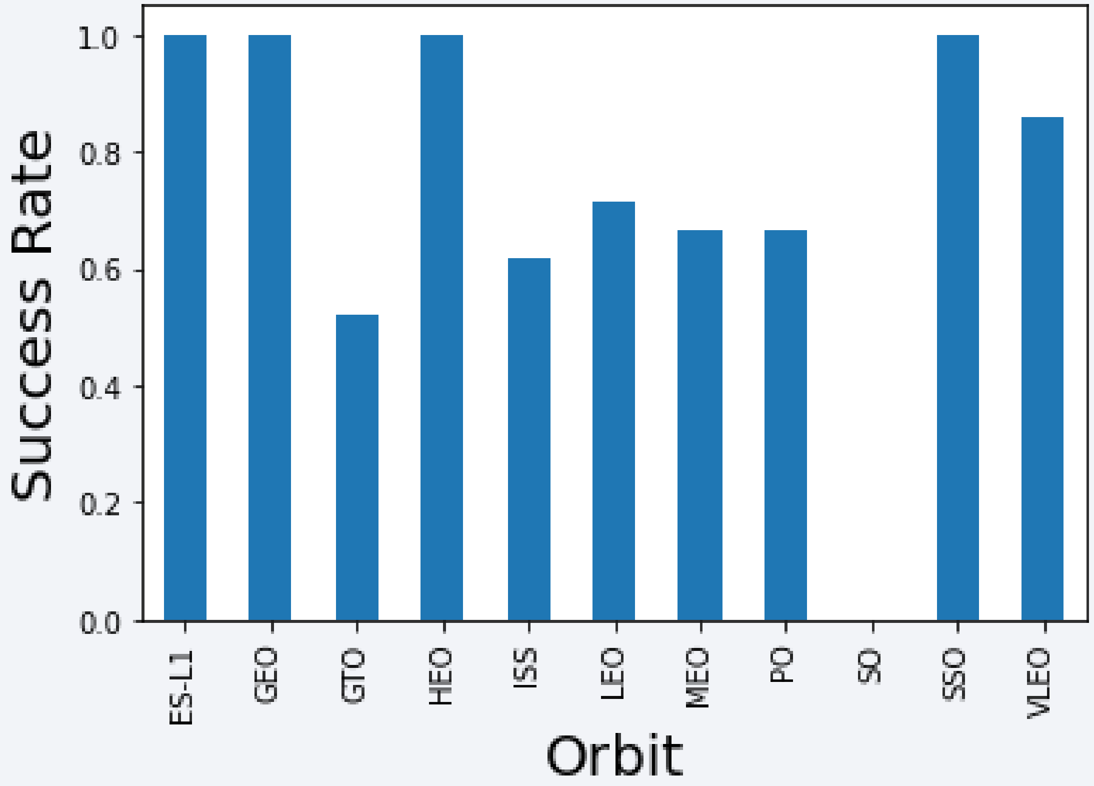
- We can see that SpaceX favors CCAFS and KSC in later flights
- Also, launches are more successful in the later flights

Payload vs. Launch Site



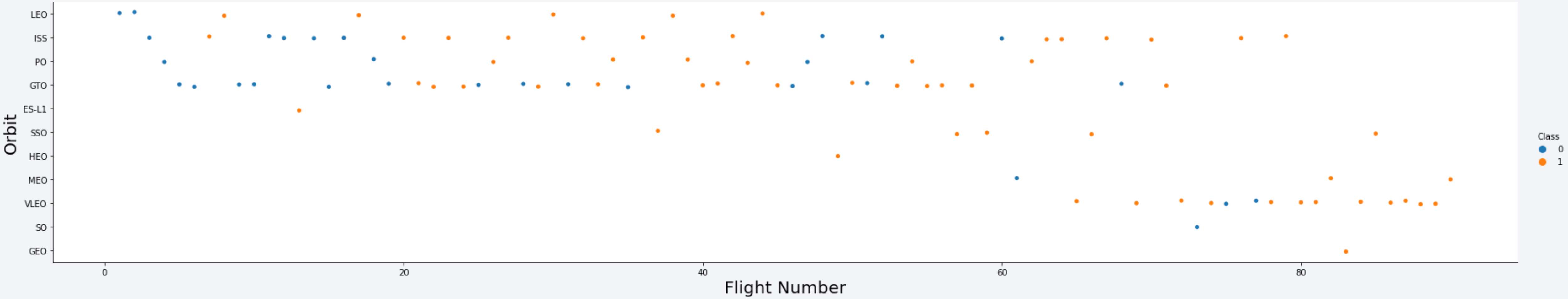
- As the number of flights progresses, SpaceX sees more successes and heavier payloads

Success Rate vs. Orbit Type



- We can see the orbits with the highest success rate are ES-L1, GEO, HEO, and SSO
- The worst rates belongs to SO

Flight Number vs. Orbit Type

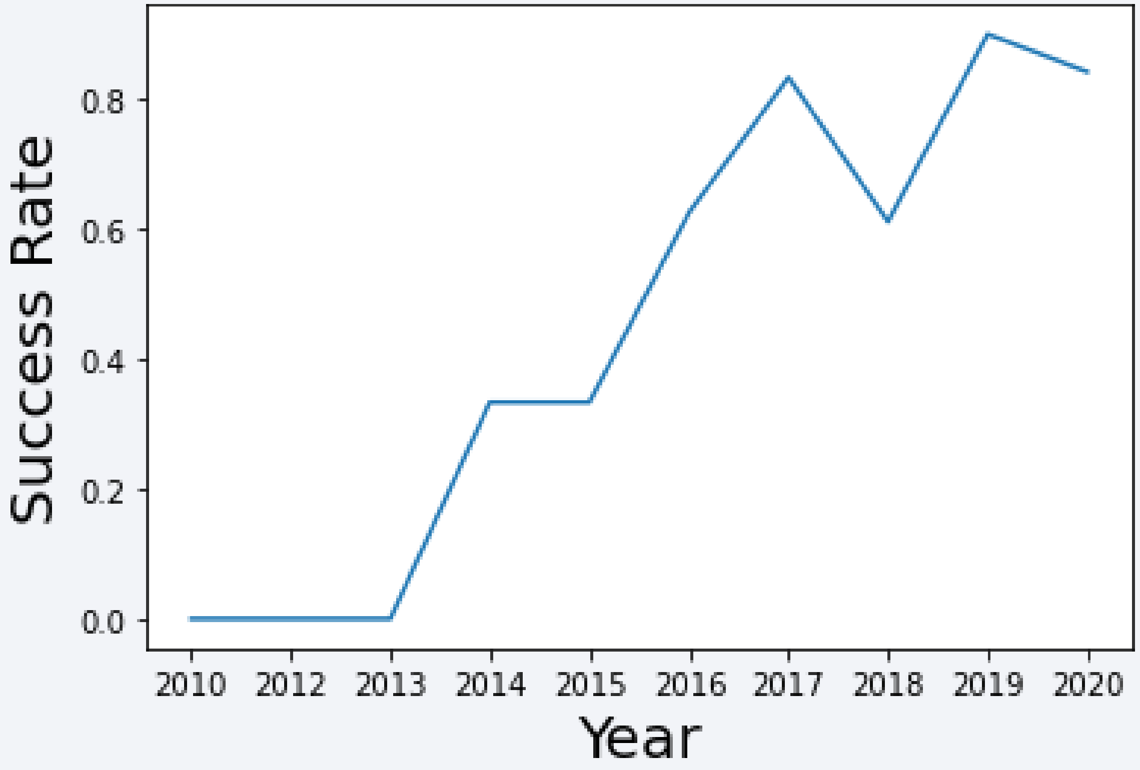


- We can see SpaceX gets better as launches overtime and favors the VLEO orbit lately

Payload vs. Orbit Type

- We can see that the VLEO orbit has flights with very large payloads
- Also, higher payloads have a higher rate of success

Launch Success Yearly Trend



- SpaceX success rates have been increasing on a year-to-year basis

All Launch Site Names

SELECT DISTINCT Launch_Site

FROM SPACEXTBL

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

Launch Site Names Begin with 'CCA'

SELECT *

FROM SPACEXTBL

WHERE Launch_Site LIKE "CCA%"

LIMIT 5

Unnamed: 0	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	08-12-2010	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)
2	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

SELECT Customer, sum(PAYLOAD_MASS__KG_)

FFROM SPACEXTBL

sum(PAYLOAD_MASS__KG_)
619967

Average Payload Mass by F9 v1.1

```
SELECT Booster_Version, avg(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version = "F9 v1.1"
```

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

First Successful Ground Landing Date

```
SELECT Date

FROM SPACEXTBL

WHERE Date = (SELECT min(Date) FROM SPACEXTBL WHERE "Landing _Outcome" =
   'Success (ground pad)')
```

Date
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_
   BETWEEN 4000 AND 6000
```

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

```
SELECT Mission_Outcome, count(Mission_Outcome)
FROM SPACEXTBL
GROUP BY Mission_Outcome
```

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

Boosters Carried Maximum Payload

```
SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM
   SPACEXTBL)
```

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

```
SELECT substr(Date, 4, 2) as Month, "Landing _Outcome", Booster_Version,
   Launch_Site
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Failure (drone ship)' AND substr(Date,7,4)='2015'
```

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

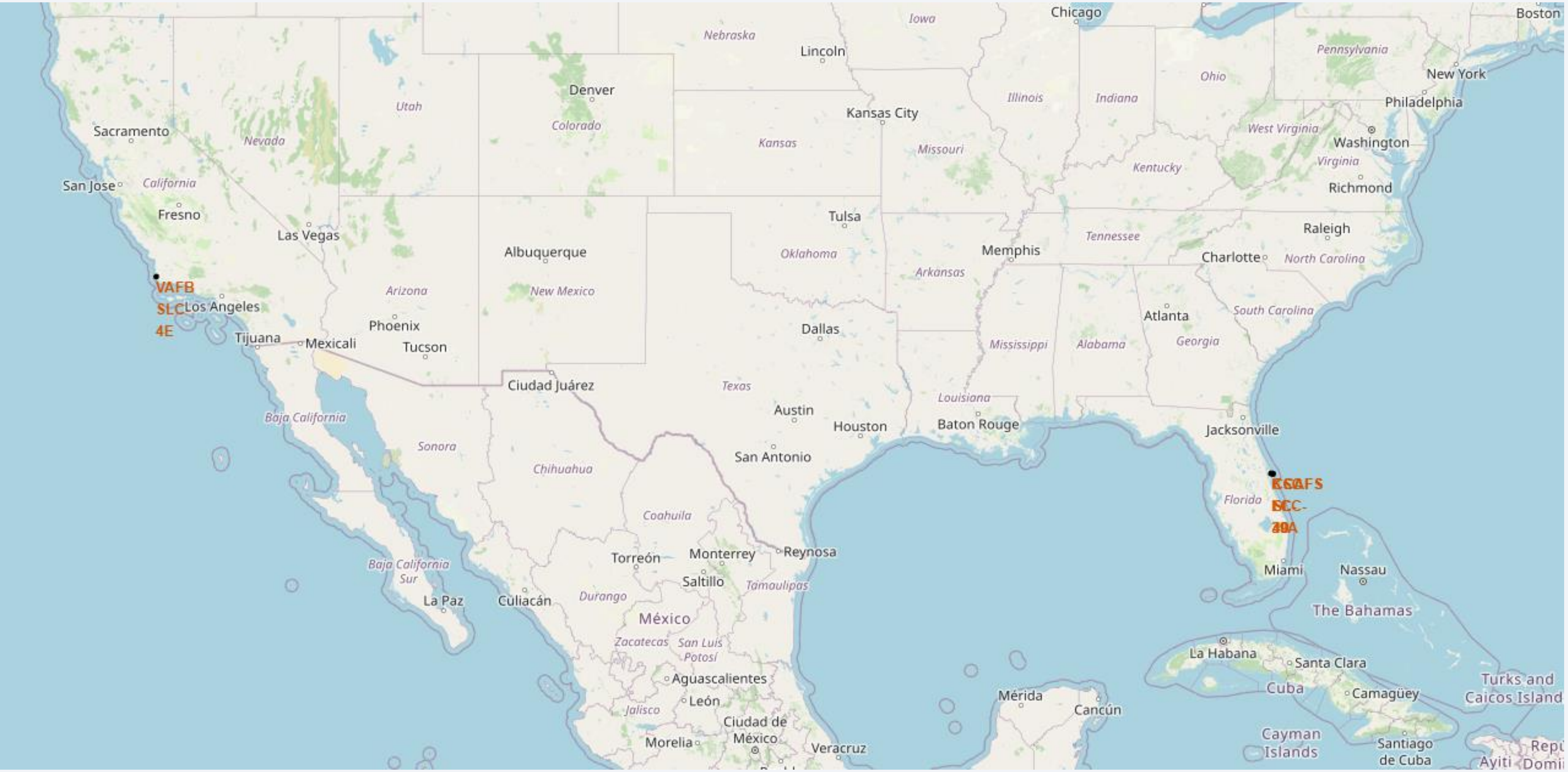
```
SELECT "Landing _Outcome", count(*) as counts
FROM SPACEXTBL
WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY "Landing _Outcome"
ORDER BY count(*) DESC
```

Landing _Outcome	counts
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

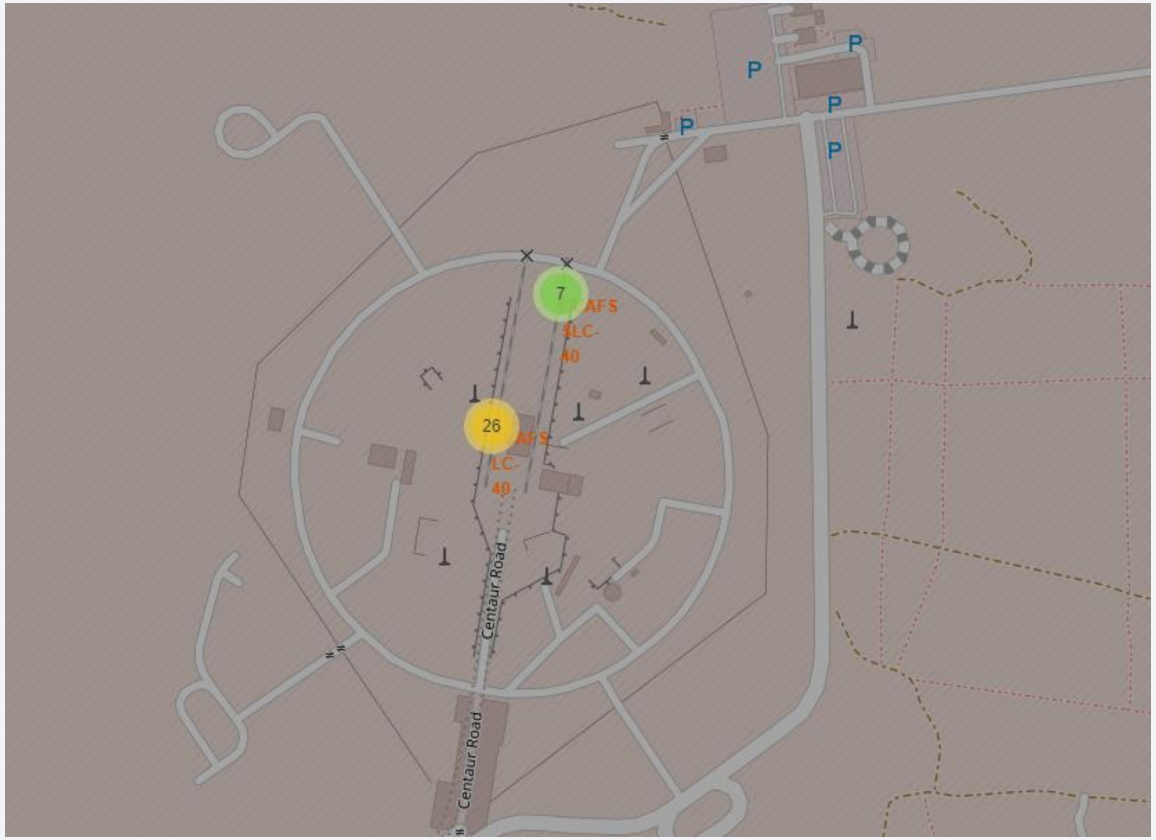
Section 3

Launch Sites Proximities Analysis

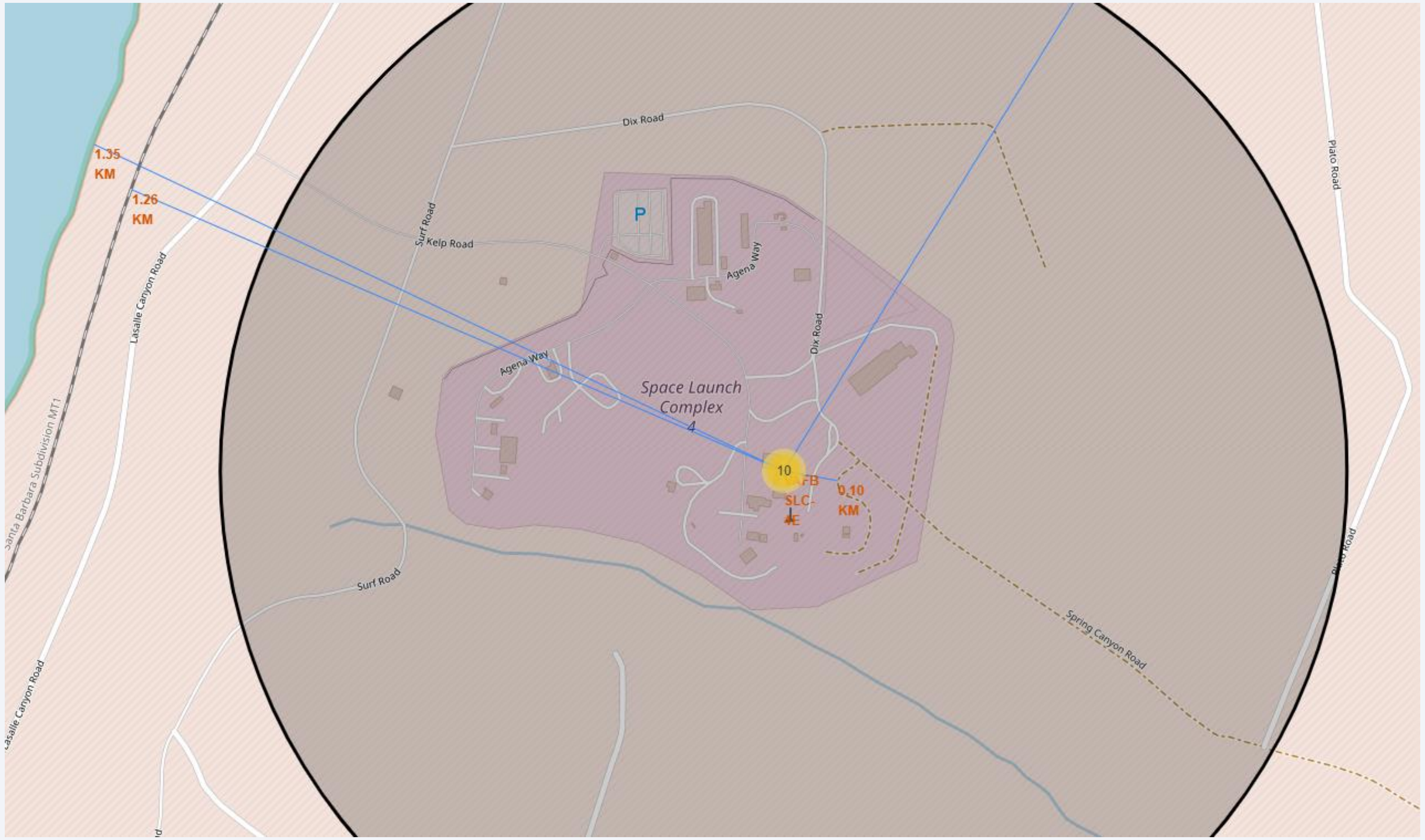
Folium Map of All Launch Sites



Folium Map with color-coded launch sites



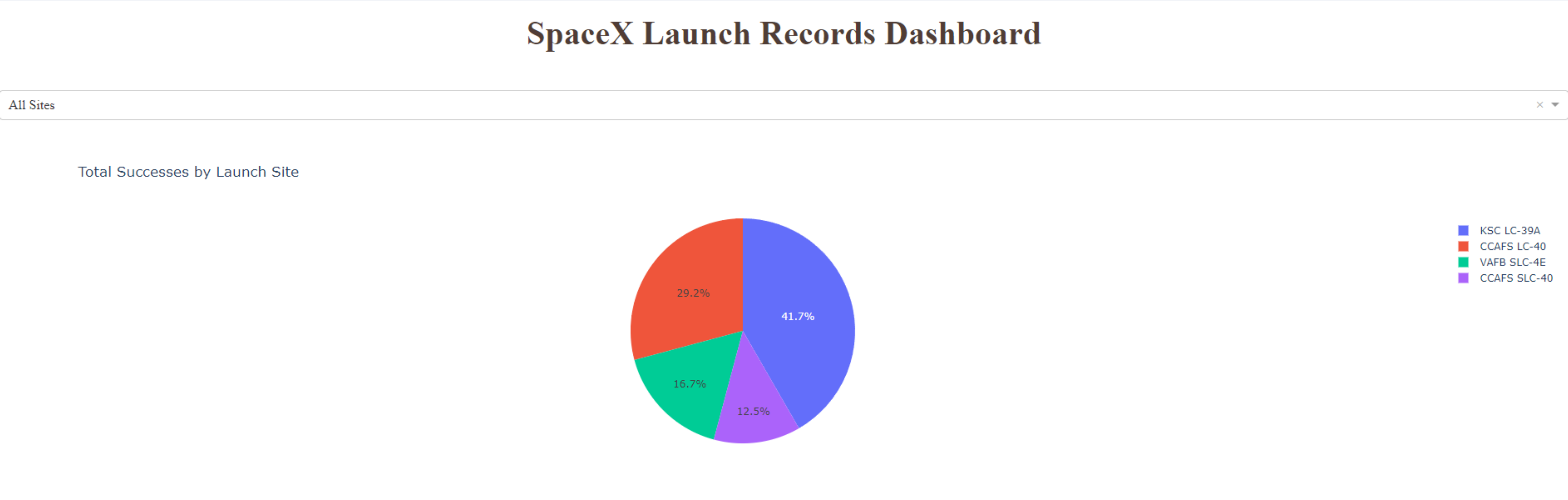
Folium Map with distance markers



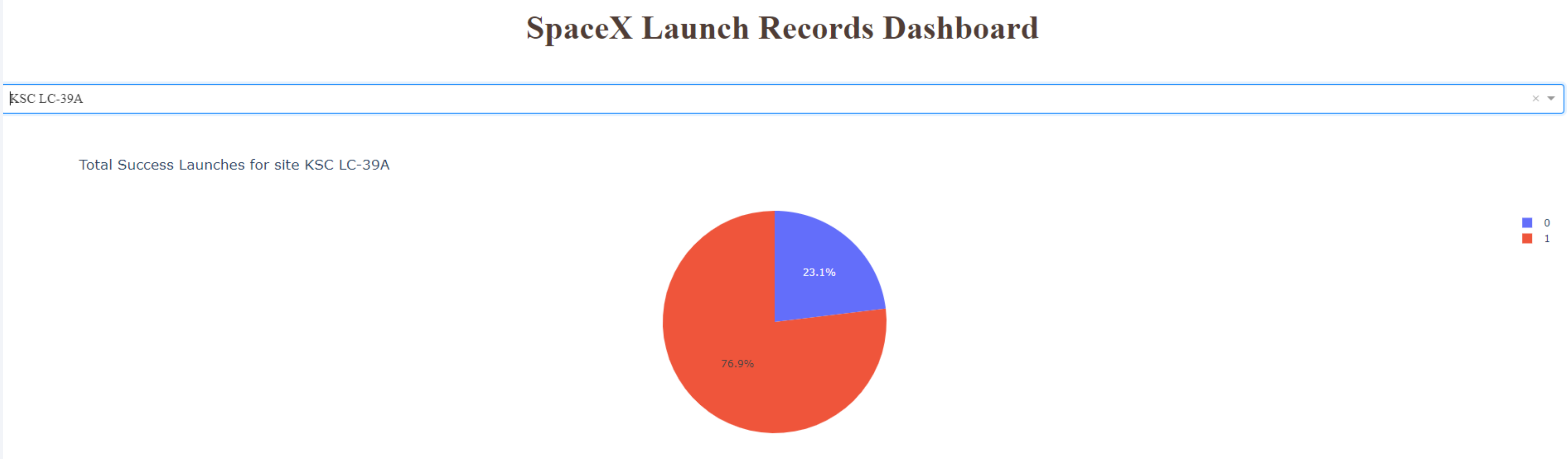
Section 4

Build a Dashboard with Plotly Dash

Dash screenshot, successes for each launch site



Dash screenshot, highest success rate launch site

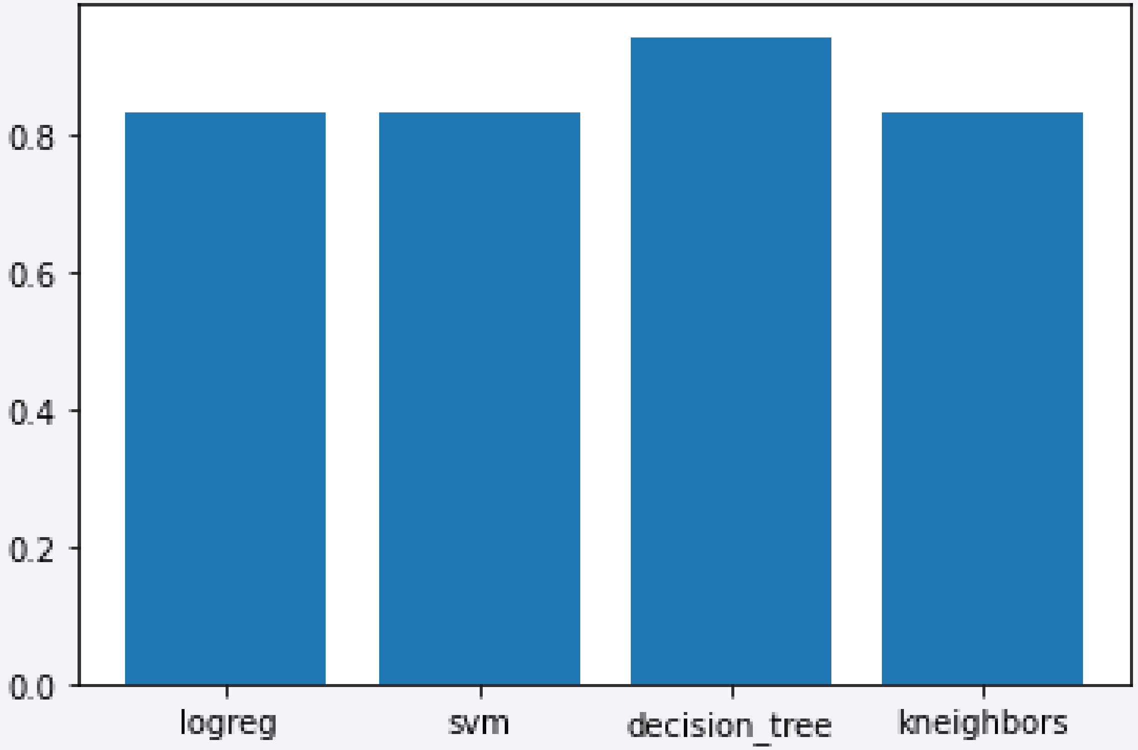


Dash screenshot, scatterplot payload mass vs success

Section 5

Predictive Analysis (Classification)

Classification Accuracy of various models



- The decision tree model has the best accuracy

Highest Accuracy Confusion Matrix

- Decision Tree Confusion Matrix

Conclusions

- SpaceX has gotten much better at launching reusable rockets overtime, with success rates reaching highs in the nineties
- SpaceX recently favors CCAFS SLC-40 and KSC LC-39A launch sites and VLEO orbits
- KSC LC-39A is the most successful launch site
- Use the Decision Tree model to predict outcomes

Thank you!