IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Porter Dosch
1/7/24

Outline

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

Introduction

- Rocket Launches typically expensive (~$165 million)
- SpaceX able to reduce costs by reusing first stage of rocket
- How to determine success/failure of landing?

Executive Summary

- Data was collected with an API
- To analyze the data, multiple python libraries, including pandas and matplotlib, were used
- Interactive visual analytics were used to determine relations among variables
- Several factors, including flight numbers and launch site, exhibit strong correlation with success of a launch

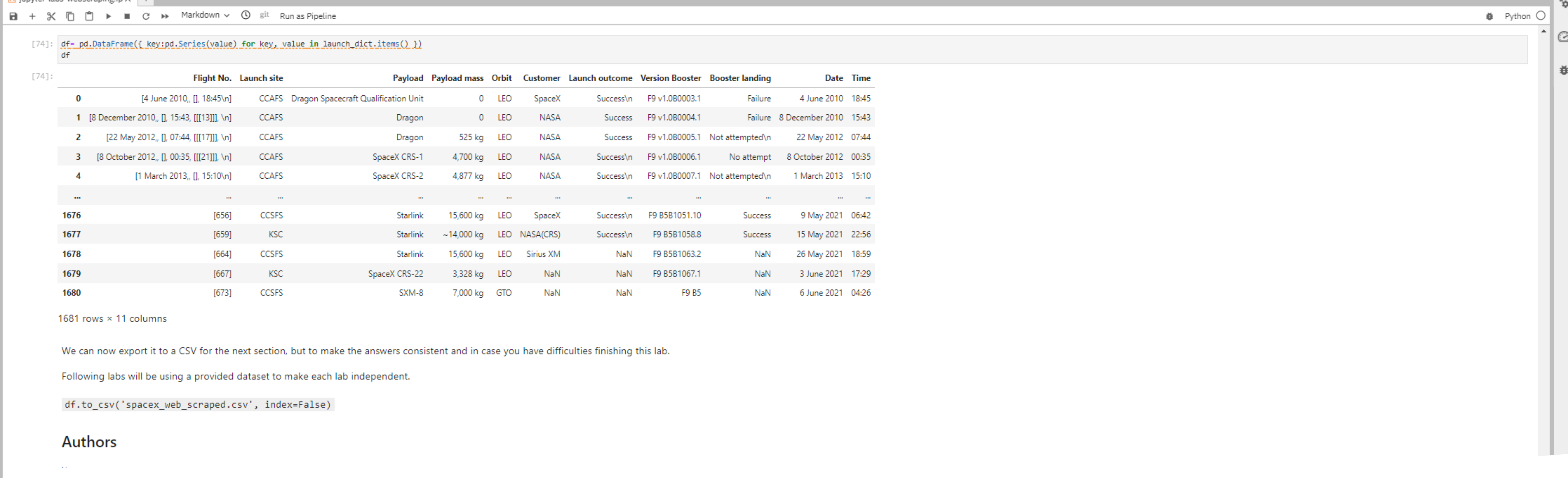
Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was scraped using BeautifulSoup from the SpaceX launch website
- Perform data wrangling
  - Data was processed to include only necessary columns in a workable format
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was split into training and testing sets to evaluate models on accuracy



[74]:
```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
df
```

[74]:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	[4 June 2010,, [], 18:45\n]	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	[8 December 2010,, [], 15:43, [[[13]]], \n]	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	[22 May 2012,, [], 07:44, [[[17]]], \n]	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	Not attempted\n	22 May 2012	07:44
3	[8 October 2012,, [], 00:35, [[[21]]], \n]	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	[1 March 2013,, [], 15:10\n]	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	Not attempted\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
1676	[656]	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
1677	[659]	KSC	Starlink	~14,000 kg	LEO	NASA(CRS)	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
1678	[664]	CCSFS	Starlink	15,600 kg	LEO	Sirius XM	NaN	F9 B5B1063.2	NaN	26 May 2021	18:59
1679	[667]	KSC	SpaceX CRS-22	3,328 kg	LEO	NaN	NaN	F9 B5B1067.1	NaN	3 June 2021	17:29
1680	[673]	CCSFS	SXM-8	7,000 kg	GTO	NaN	NaN	F9 B5	NaN	6 June 2021	04:26

1681 rows × 11 columns

We can now export it to a CSV for the next section, but to make the answers consistent and in case you have difficulties finishing this lab.

Following labs will be using a provided dataset to make each lab independent.

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

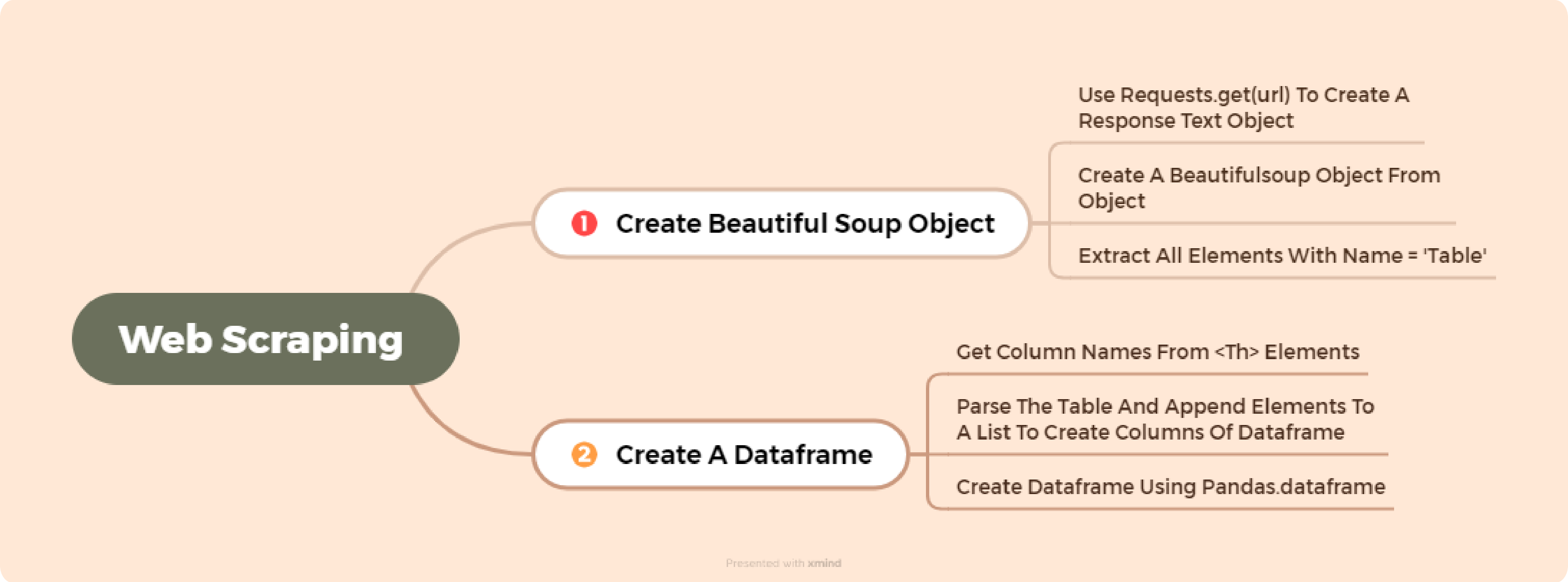
Authors

Data Collection

- Data sets were collected using a web scraper (BeautifulSoup)
- Parsed through HTML of SpaceX page
- Found tables with information relevant to launch

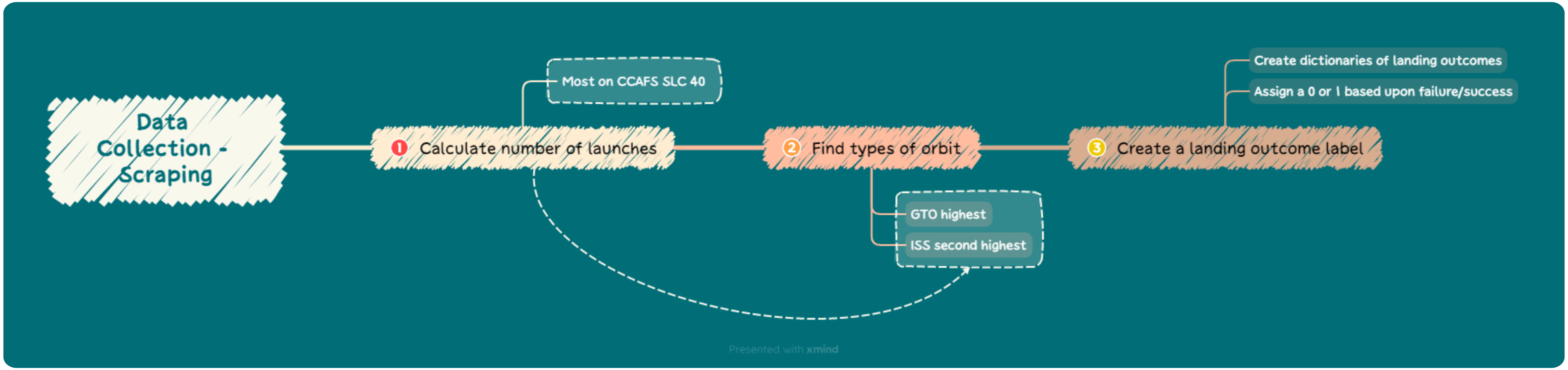
Data Collection – SpaceX API

- Call made to SpaceX API using requests.get() function
- Response was decoded as a JSON and turned into a Pandas DataFrame
- [Notebook SpaceX API Github]



Data Collection - Scraping

- Webscraping primarily done through BeautifulSoup
- Enabled searching through HTML of webpage to find data
- [Webscraping Notebook]



Data Wrangling

- Problematic column was launch type – needed categorical variables
- Changed form of data to include a categorical "Class" column describing success or failure of launch
- [Data Wrangling Notebook]

EDA with Data Visualization

- Several charts plotted to determine potential causal relationships between variables
- Attempted to find a correlation between Success (class) and Flight Number, Payload Mass, Orbit Type, or Launch Site
- All were chosen because of potential to influence launch outcome (eg. As Flight Number increased, the company became more experienced, and would consequently expect an increase in success rate)
- EDA with Data Visualization Notebook

EDA with SQL

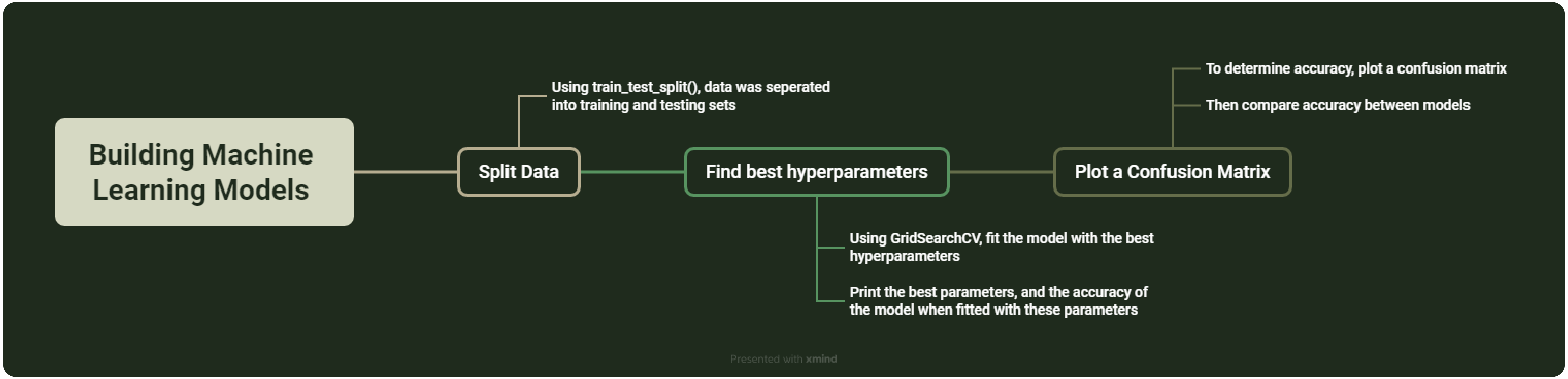
- Query 1: Displaying unique launch sites
- Query 2: Finding 5 records beginning with "CCA"
- Query 3: Calculating total payload mass from NASA-launched boosters
- Query 4: Average payload mass from booster version F9 v1.1
- Query 5: Find date of first successful launch in ground pad
- Query 6: List names of boosters with success in drone ship in a certain range of payload mass
- Query 7: List total number of successful and failed mission outcomes
- Query 8: Find booster versions which have carried maximum payload mass
- Query 9: Display records and month names for drone failures in year 2015
- Query 10: Rank count of landing outcomes between range of dates in descending order.
- EDA with SQL Notebook

Build an Interactive Map with Folium

- Added circles and markers to denote individual launch sites
- Created MarkerCluster objects to show launches without cluttering map
- Added PolyLines to determine proximity of launch sites to key points (railways, rivers, etc.)
- Interactive Folium Map Notebook

Build a Dashboard with Plotly Dash

- Plotly dashboard that contains pie charts with number of successful launches per site, or success rate per site
- Scatter plot of payload mass and launch success, with an adjustable slider for payload mass
- These plots were added to demonstrate correlations between variables in the dataset and success of the launch
- Plotly Dash Python File



Predictive Analysis (Classification)

- Trained and evaluated the accuracy of several models
- Logistic Regression, Support Vector Machine, Decision Tree, and k-Nearest Neighbors
- Predictive Analysis Notebook

Results

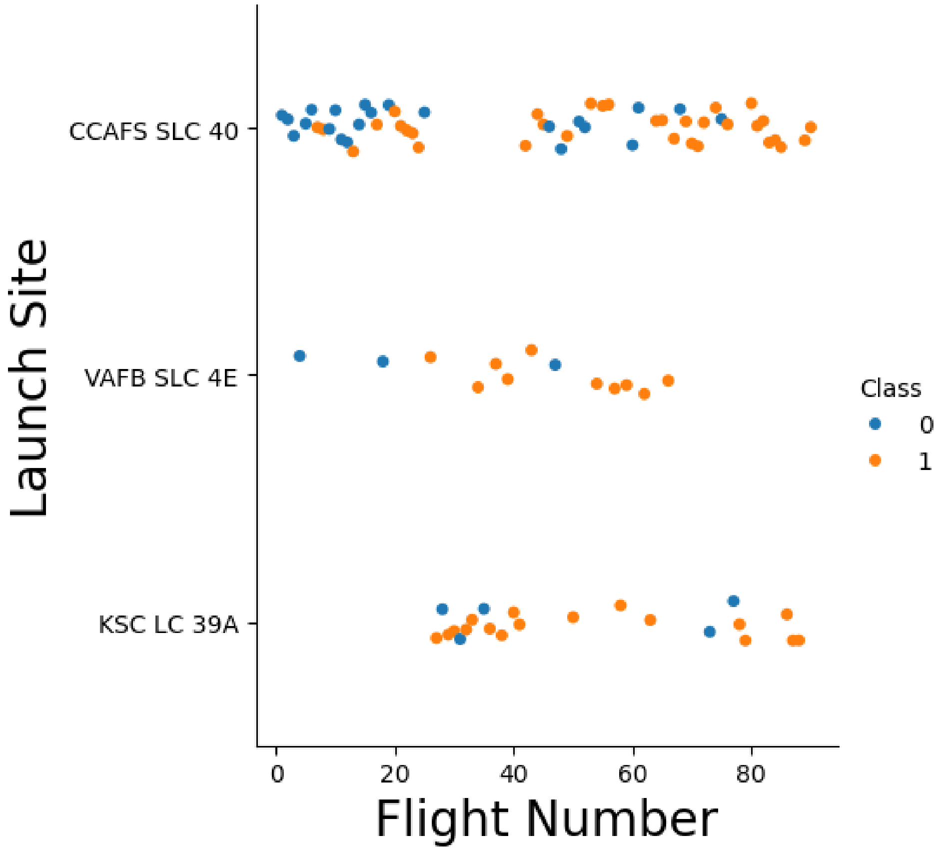
- Insights drawn from EDA
- Interactive analytics (Folium, Plotly) demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

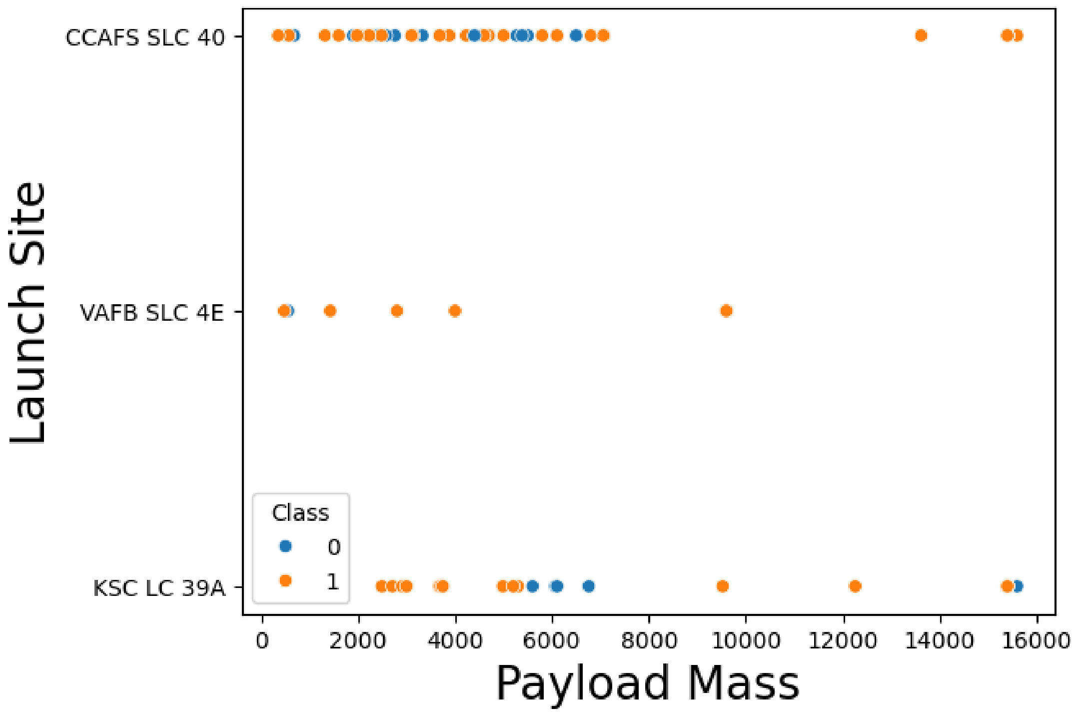
- There appears to be a lower success rate at launch site CCAFS SLC 40 compared to the others
- Generally, an increase in flight number is accompanied by an increase in successes



Payload vs. Launch Site

- Initially, one would believe that an increase in payload mass would correspond to a higher likelihood of failure

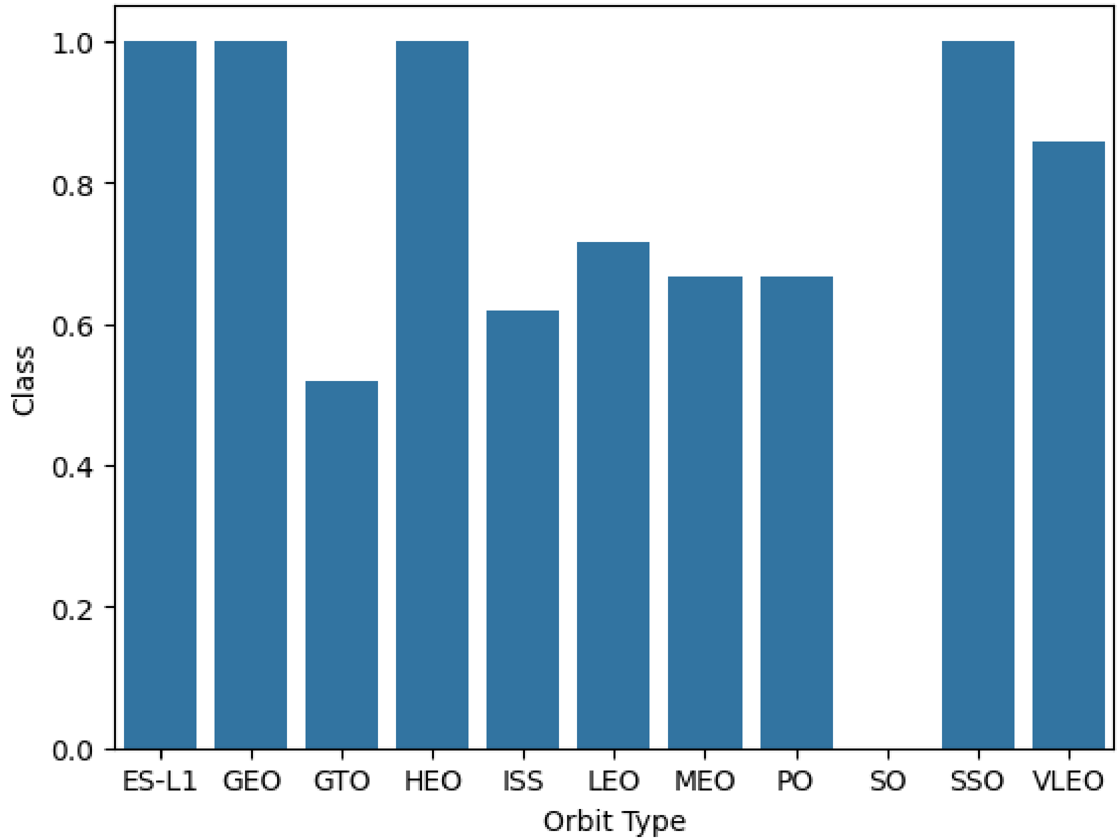
- Failures seem to be distributed fairly evenly among high and low payload masses, with site CCAFS SLC 40 still displaying the highest failure rate



Success Rate vs. Orbit Type

- Success rates appear to differ among orbit types

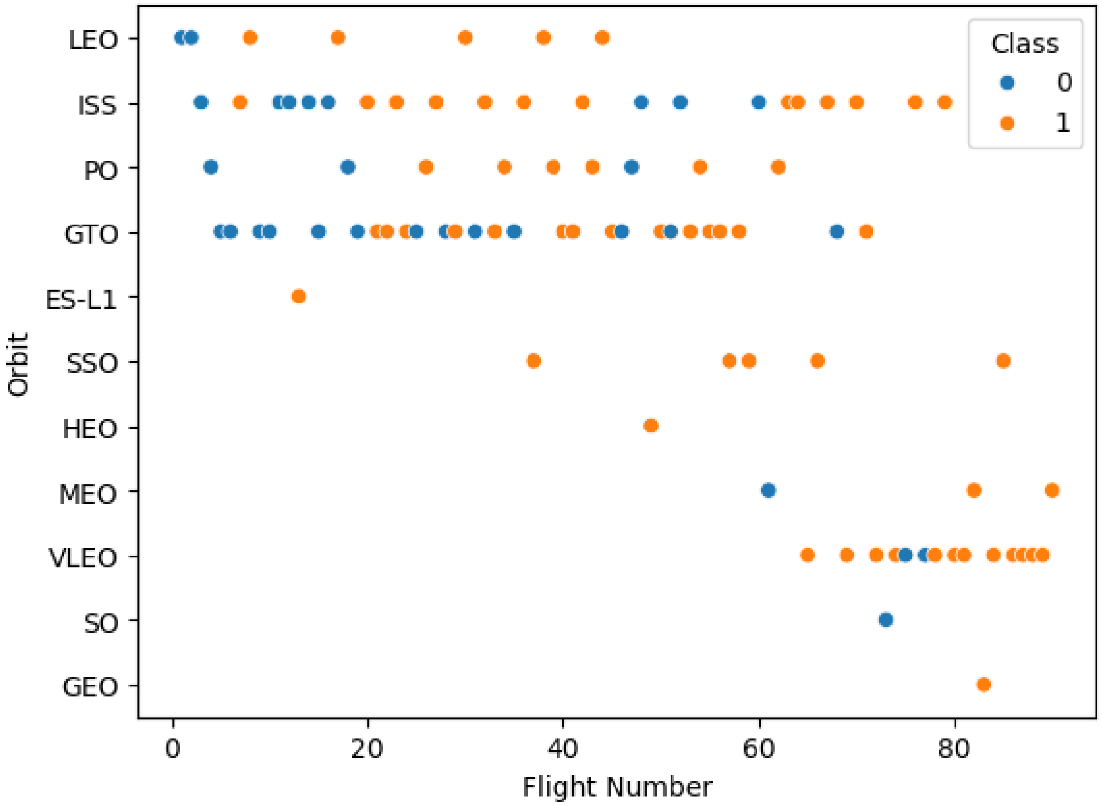
- ES-L1, GEO, HEO, and SSO orbit types have the highest success rates



Flight Number vs. Orbit Type

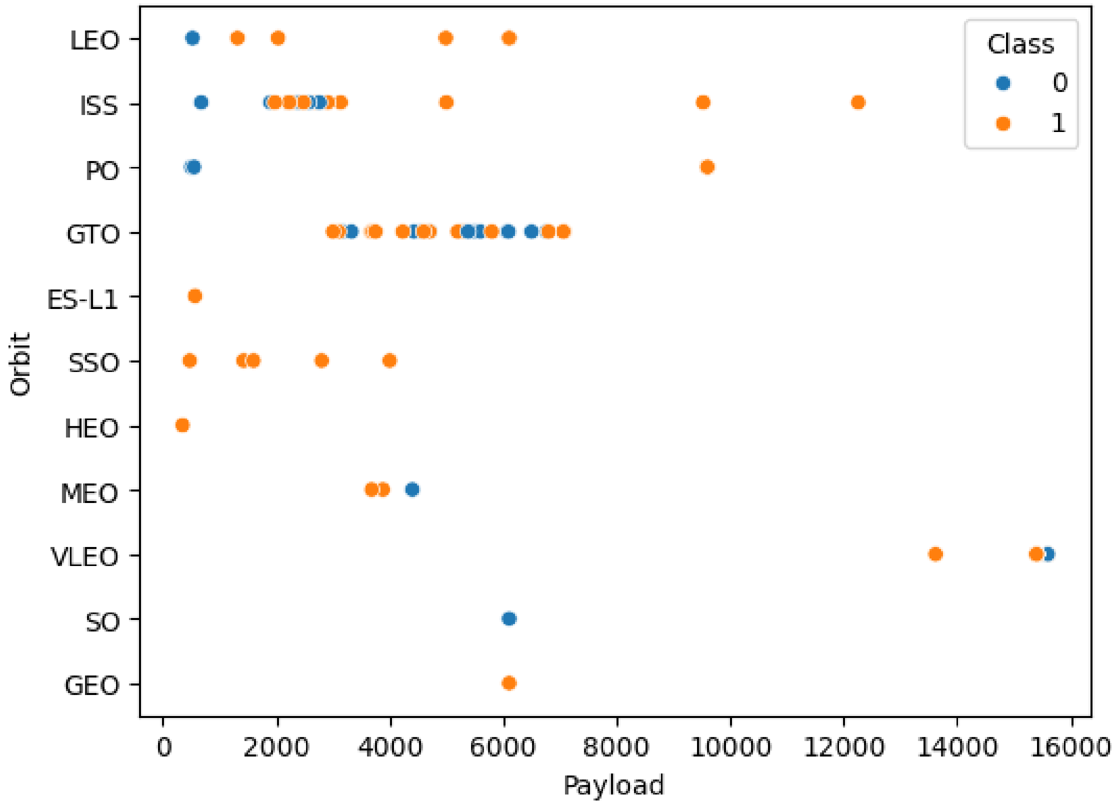
- The previously identified orbit types had a high (100%) success rate because of only having one data point

- Even across generally unsuccessful orbit types, as flight number increased, so did likelihood of success



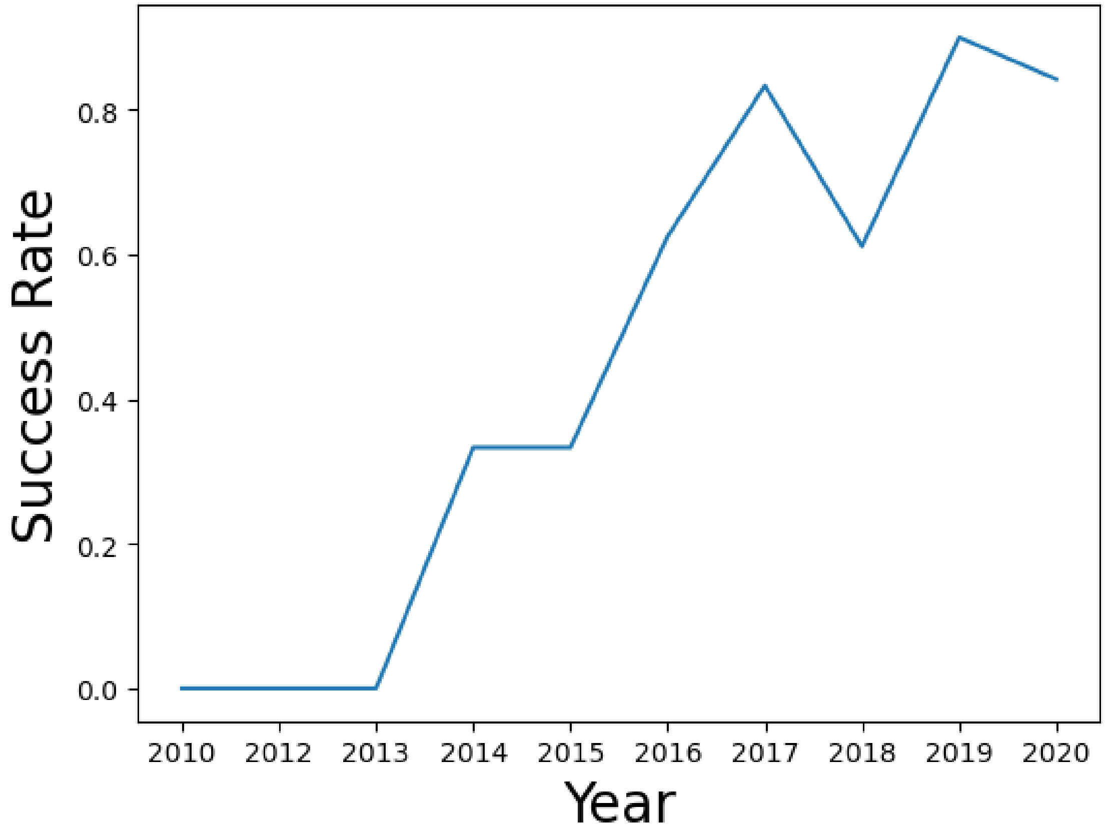
Payload vs. Orbit Type

- Payload mass again seems to be independent of outcome
- Although success rates differed among orbit types, there are both successful and failed launches with varied payload mass



Launch Success Yearly Trend

- In general, as year increases, the success rate for that year also increases
- Similar to flight number, this is likely due to the company becoming more experienced with rocket launches, and less prone to error



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

```
[49]: %sql select distinct "Launch_Site" from SPACEXTABLE
```

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

[49]:

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

All Launch Site Names

- Task: Find the names of the unique launch sites
- There are four unique launch sites in the data set

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

```
[19]: %sql select * from SPACEXTABLE where "Launch_Site" like "CCA%" LIMIT 5

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

[19]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Launch Site Names Begin with 'CCA'

- Task: Find 5 records where launch sites begin with `CCA`
- Within the data set, many entries have a launch site beginning with 'CCA'

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

```
[18]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" = "NASA (CRS)"

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

[18]:

sum(PAYLOAD_MASS__KG_)
45596

Total Payload Mass

- Task: Calculate the total payload carried by boosters from NASA
- NASA Boosters have carried a total of 45,596 kg in payload mass

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

```
[11]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" = "F9 v1.1"
 * sqlite:///my_data1.db
Done.
```

[11]:

avg(PAYLOAD_MASS__KG_)
2928.4

Average Payload Mass by F9 v1.1

- Task: Calculate the average payload mass carried by booster version F9 v1.1
- The average mass carried by booster F9 v1.1 was roughly 3000 kg

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[13]: %sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)"

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

[13]:

min("Date")
2015-12-22

First Successful Ground Landing Date

- Task: Find the date of the first successful landing outcome on ground pad
- The first successful landing outcome on ground pad was December 22, 2015

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[14]: %sql select "Booster_Version" from SPACEXTABLE where ("Landing_Outcome" = "Success (drone ship)" and "PAYLOAD_MASS__KG_" between 4000 and 6000)

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

[14]:

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- Task: List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- There were four boosters that had success in drone ship within the specified payload mass range

List the total number of successful and failure mission outcomes

```
[28]: %sql select "Mission_Outcome", count("Mission_Outcome") from SPACEXTABLE GROUP BY "Mission_Outcome"
```

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

[28]:

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

Total Number of Successful and Failure Mission Outcomes

- Task: Calculate the total number of successful and failure mission outcomes
- Almost every mission was a success, since failed landings were planned, and were thus considered a successful mission

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[35]: %sql select "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

[35]:

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

Boosters Carried Maximum Payload

- Task: List the names of the booster which have carried the maximum payload mass
- There were 12 boosters that carried the maximum payload mass

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
[40]: %sql select "Booster_Version", "Landing_Outcome", "Launch_Site", substr("Date", 6, 2) as month, substr(Date,0,5) as year from SPACEXTABLE where ("Landing_Outcome" = "Failure (drone ship)" and year = "2015")
```

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

[40]:

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

2015 Launch Records

- Task: List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- There were 2 instances of a failed landing in drone ship in 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.

```
[48]: %sql select "Landing_Outcome", count("Landing_Outcome") as "Outcome_Count" from SPACEXTABLE WHERE ("Date" BETWEEN '2010-06-04' and '2017-03-20') GROUP BY "Landing_Outcome" ORDER BY "Outcome_Count" DESC

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

[48]:

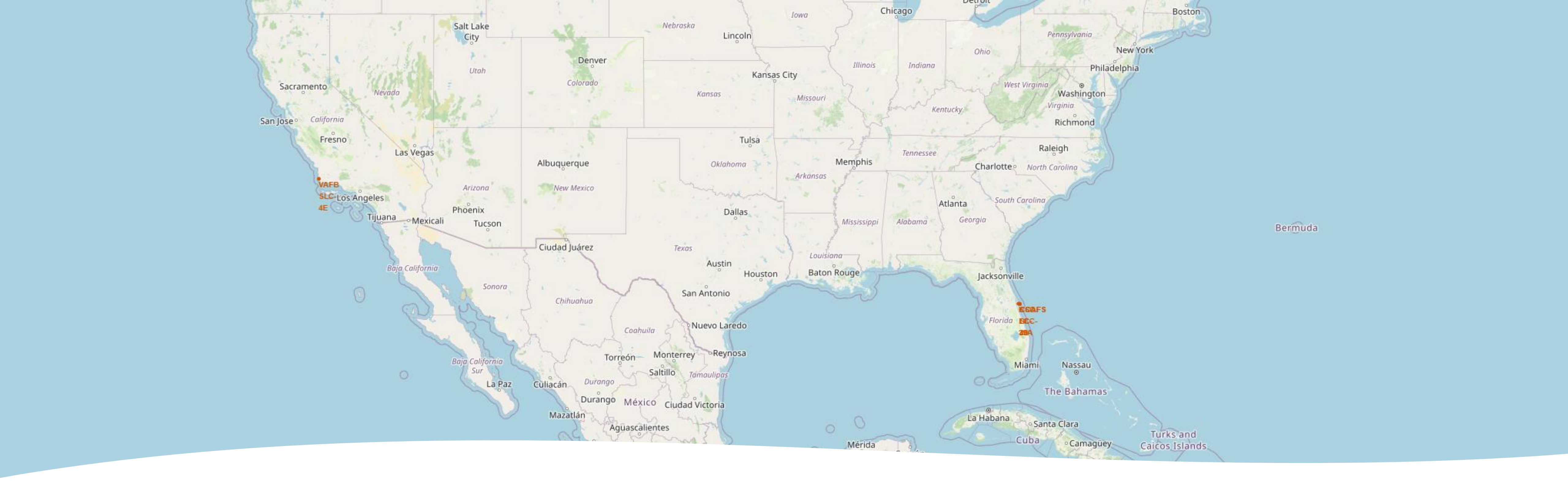
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

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

- Task: 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
- The most common landing outcome was no attempted landing, and the least common was precluded in drone ship

Section 3

Launch Sites Proximities Analysis

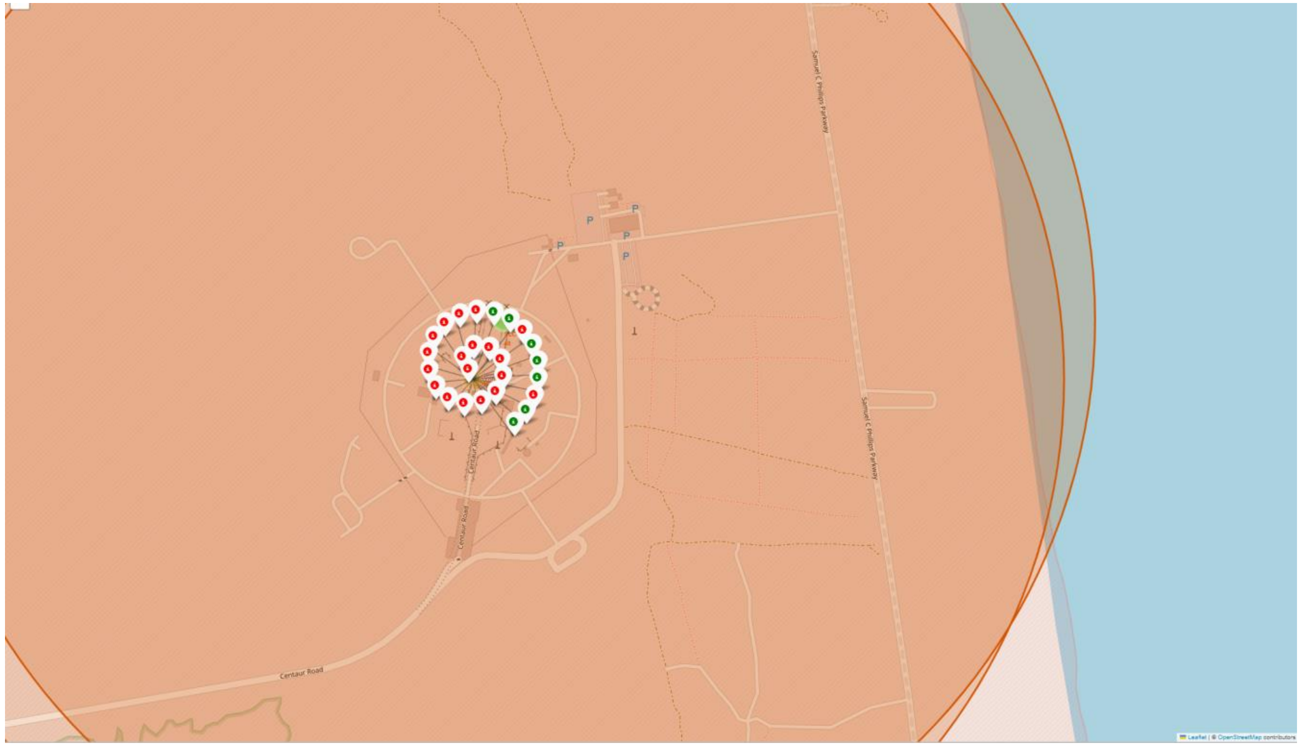


Launch Sites

- Task 1: Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- One launch site in California, three in Florida

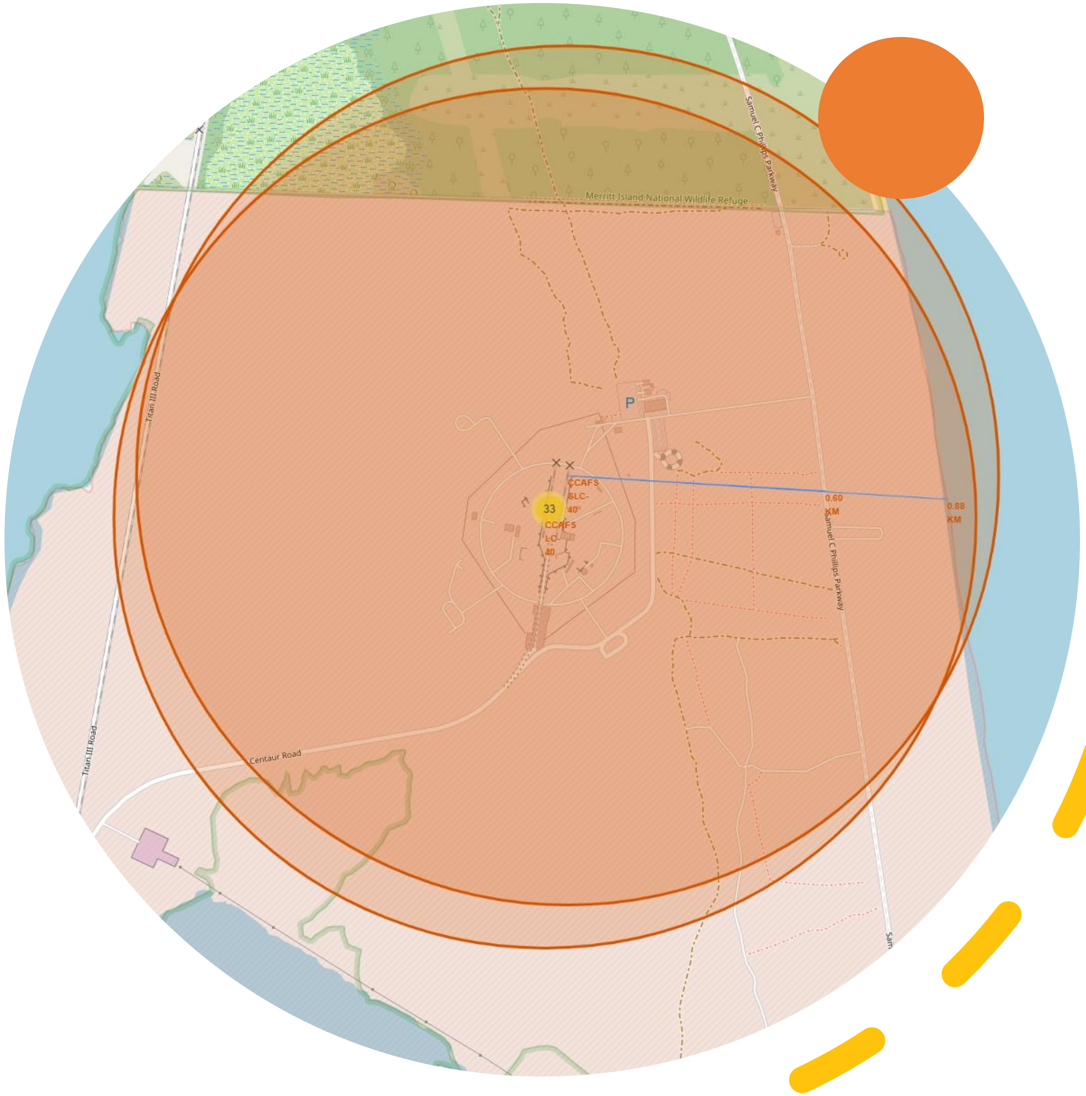
Launch Clusters

- Successful (green) and failed (red) launches are shown within the marker clusters on the map
- CCAFS LC-40 and its launch results are shown in picture



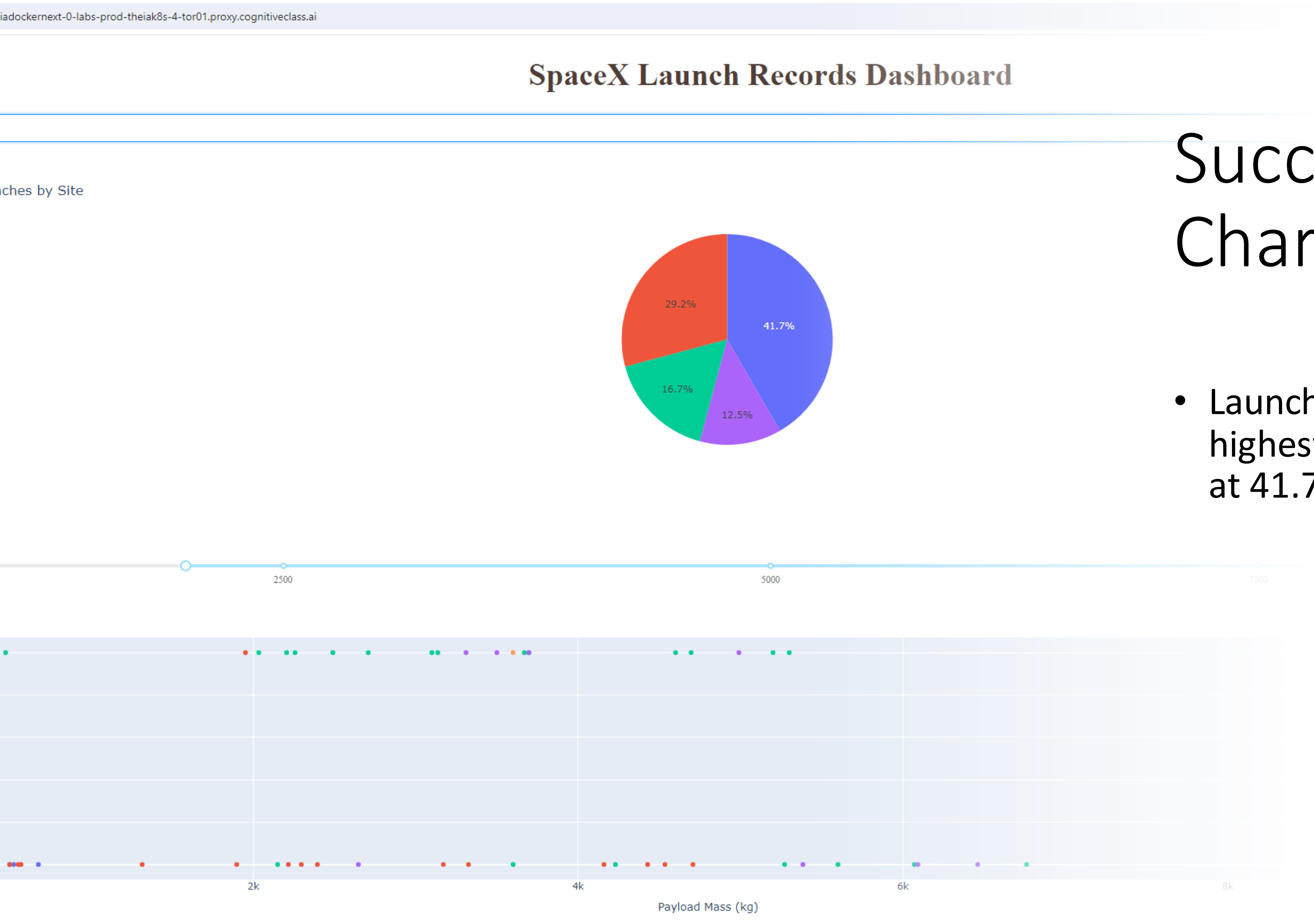
Distance to Points of Interest

- Using PolyLine objects, distance to points of interest was plotted
- Launch sites appear to maintain close proximity to coastlines and highways
- Likely for safety of launches and ease of access to materials (respectively)
- They remain farther from major cities



Section 4

Build a Dashboard with Plotly Dash

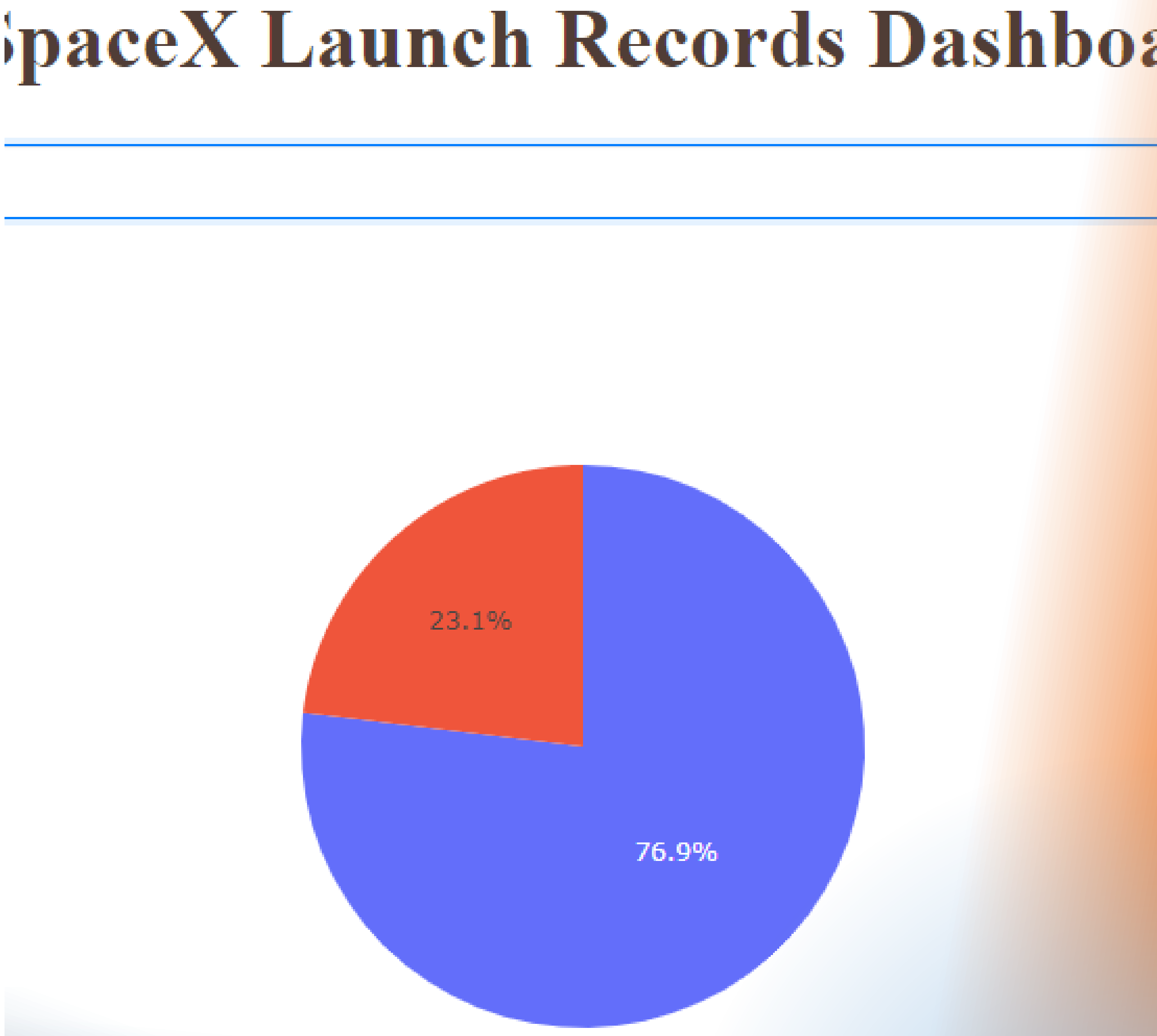


Success Rate by Chart

- Launch site KSC LC-39A has the highest percentage of successes at 41.7%

Most Successful Launch Site

- 76.9% of launches at site KSC LC-39A were successful
- This is by far the highest success rate of all launch sites, with CCAFS SLC-40 having the second highest at 42.9%



Payload Mass vs. Launch Outcome

- Payload mass does not appear from the scatter plot to have a significant effect on the success of a launch

- However, booster version does appear to have an effect
- For example, booster FT has a significantly higher success rate than other types

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Accuracy for each trained model shown in figure 1
- Decision Tree had the highest accuracy out of the models trained

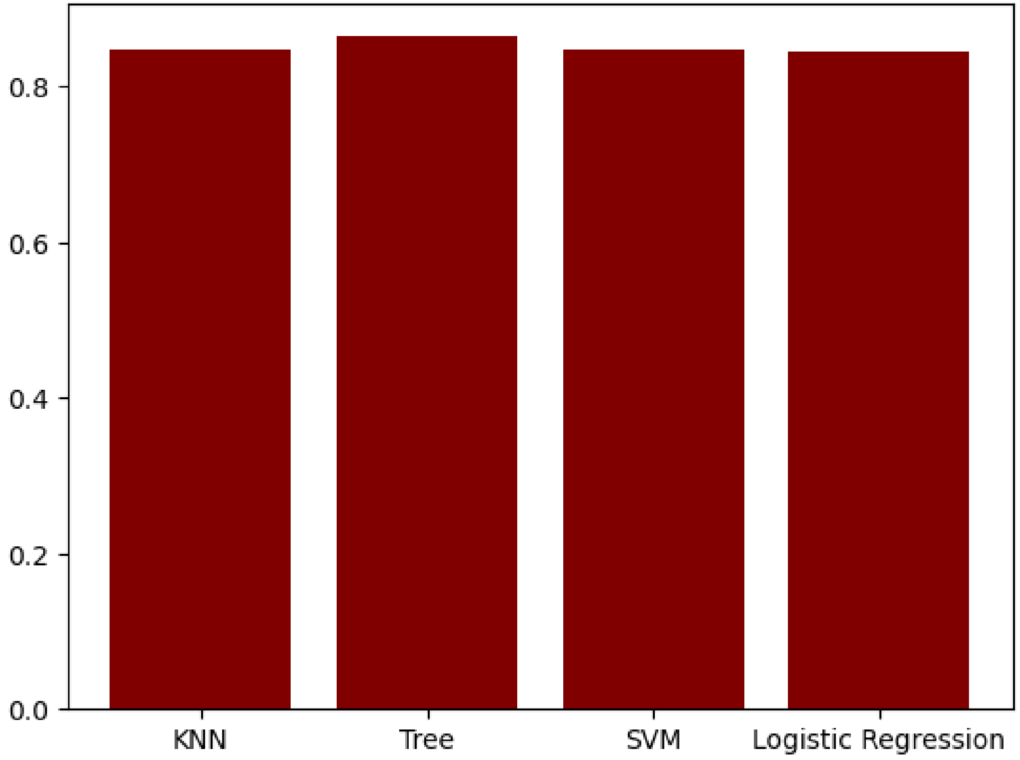
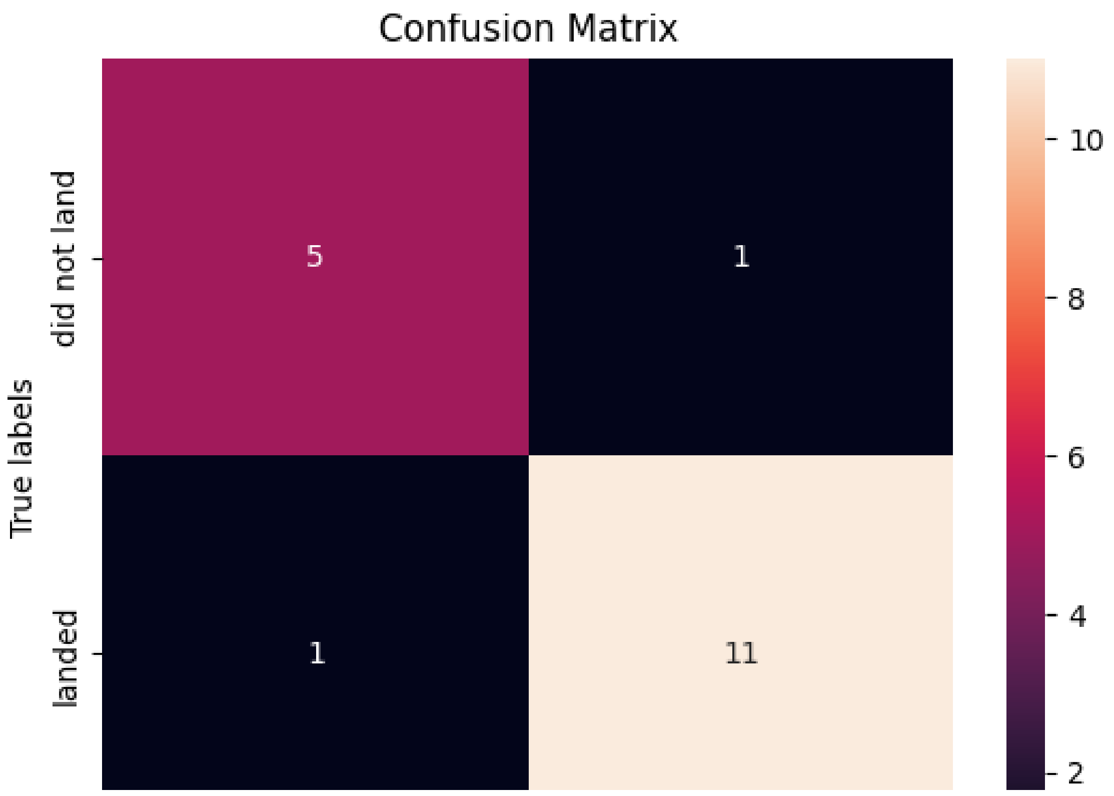


Figure 1

Confusion Matrix



- Whereas the other models suffered from false positives, the Decision Tree model did not have a significant number of false positives or negatives
- In addition, it had a higher accuracy than the other model types

Conclusions

- As flight number increases, success rate also tends to increase
- Launch sites CCAFS SLC-40 had a lower success rate than the other two
- Certain orbit types appeared to have a higher success rate (ES-L1, GEO, HEO, SSO)
- However, this was only because of few data points with that orbit type
- Payload mass seems to be independent of launch success
- Launch sites seem to be positioned close to coastlines and highways, but farther from major cities

Appendix

- Github Repository

Thank you!