CAPSTONE ASSIGNMENT

Dylan Ogden

05/03/2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis (EDA) with SQL
  - EDA Data visualization
  - Build an interactive map with Folium
  - Build an interactive dashboard with Plotly Dash
  - Employ classification models for predictive analysis
- Results
  - EDA results
  - Visualization results
  - ML results

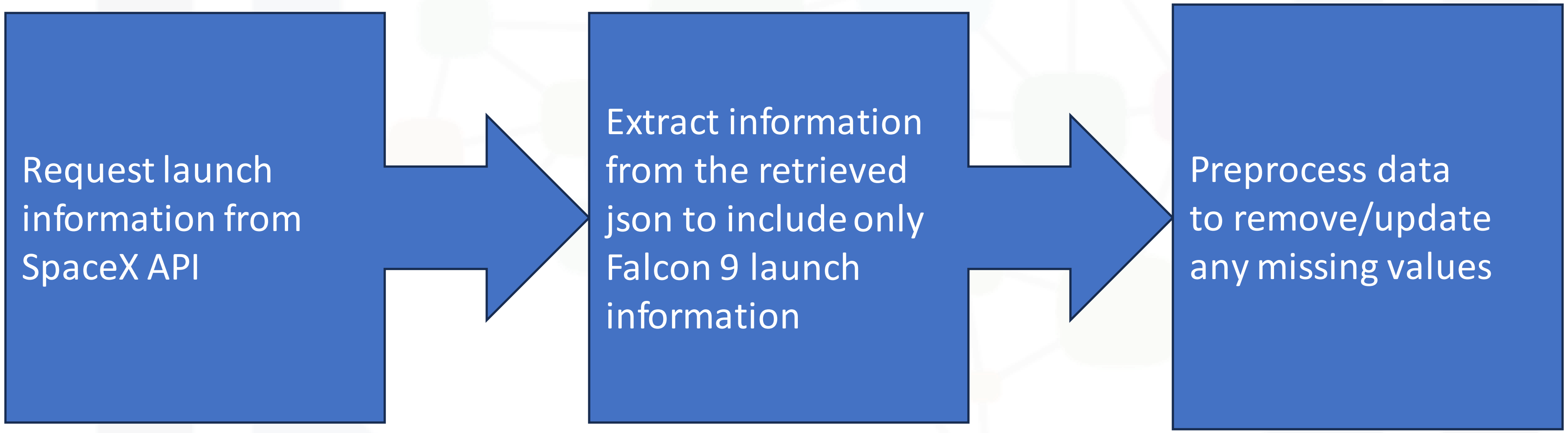
INTRODUCTION

- Overview
  - The modern space race is upon us and many companies are taking part in it. The most successful is SpaceX. SpaceX's accomplishments include: sending spacecrafts to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.
- Important Questions
  - What characteristics of the launches will help us in determining whether a mission will be successful?
  - Are the launches becoming more and more successful as the years progress?
  - Which model can we use to accurately predict the outcome of launches?

METHODOLOGY

- Data Collection
  - SpaceX API
  - Webscraping from Wikipedia
- Data Wrangling
- Exploratory Data Analysis
- Visual Analytics using Folium and Plotly Dash
- Predictive analysis with classification models

Data Collection using API



Data Collection using Webscraping

Request launch information for Falcon 9 from Wikipedia

→

Extract all relevant information from the retrieved HTML response.

→

Parse into a Pandas DataFrame object

→

Preprocess data to remove/update any missing values

Data Wrangling

Perform initial EDA

Summarize launches per site, occurrences of each orbit, and occurrences of mission outcome per orbit type were calculated.

Creation of the class label from the outcome column. Class label: outcome label

EDA Visualization

- EDA was performed to observe relationships between the different features of the SpaceX data.
- Relationships
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
- The types of plots include: line plots, scatter plots, bar charts, etc.

EDA with SQL

- Computed the following queries
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begin with "CCA"
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date of the first successful landing outcome
  - Names of the boosters which had success in drone ship and had payload mass in range of 4000 and 6000 kg
  - Total number mission outcomes for success and failure
  - Names of the booster versions which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, booster versions, and launch site names for year 2015
  - Rank of the count of landing outcomes between the dates 2010-06-04 and 2017-03-20

Visualization with Folium

- Added markers for launch sites.
- Added colored markers at the launch sites indicated mission outcomes.
- Computed the distances between the launch sites based on proximity.

Plotly Visualization

- Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches
- Slider of Payload Mass Range
- Added a scatter chart to show the correlation between payload and launch success.

Predictive Analysis

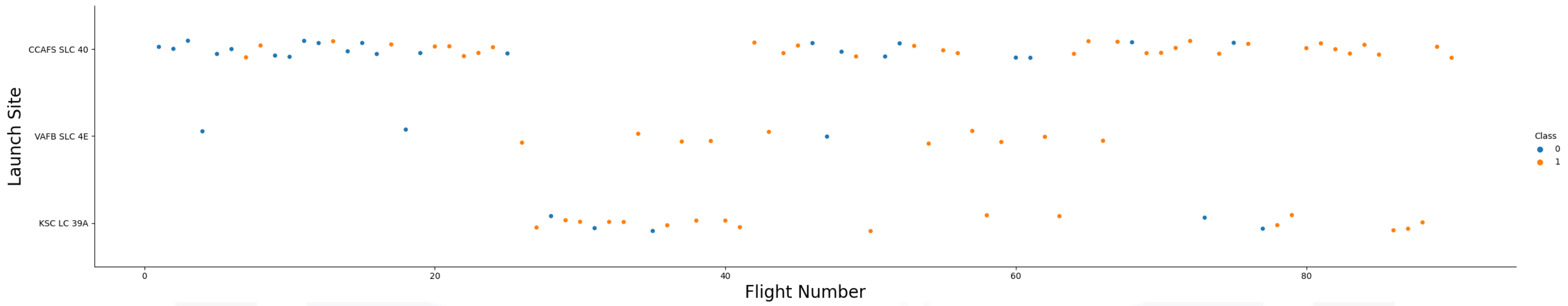
Four classification models were compared

1. Logistic Regression,
2. Support vector machine
3. Decision tree
4. K-nearest neighbors

Data preparation and standardization → Test of each model with combinations of hyperparameters → Comparison of results

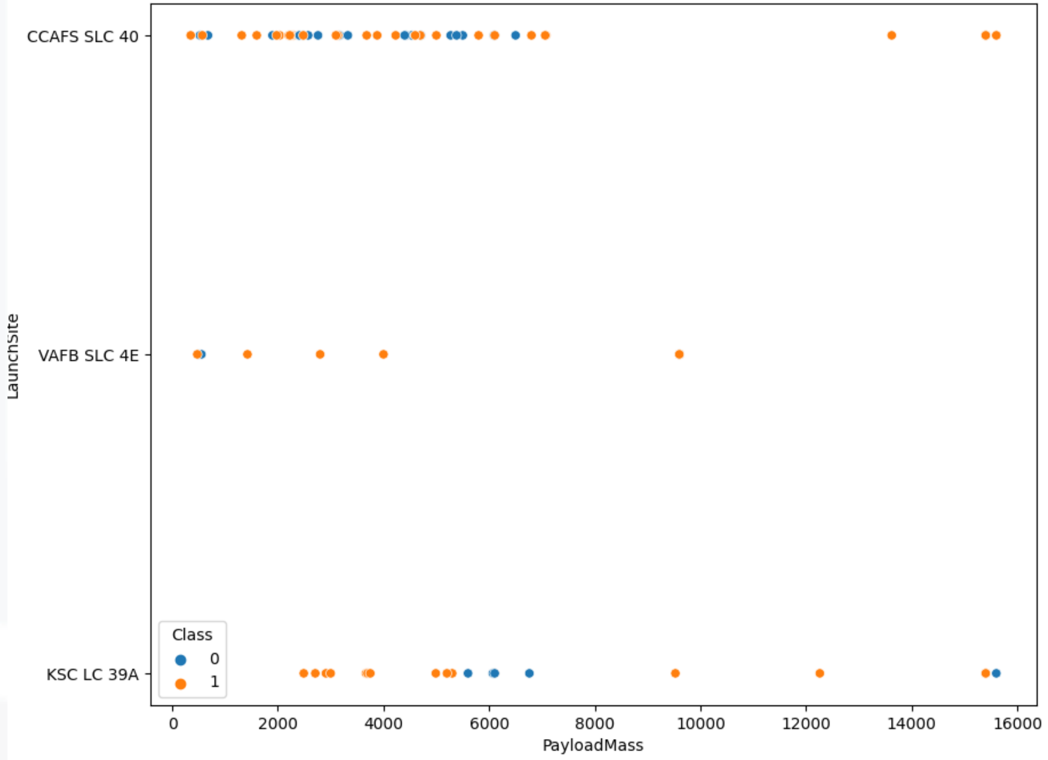
Results

Flight number based on launch site

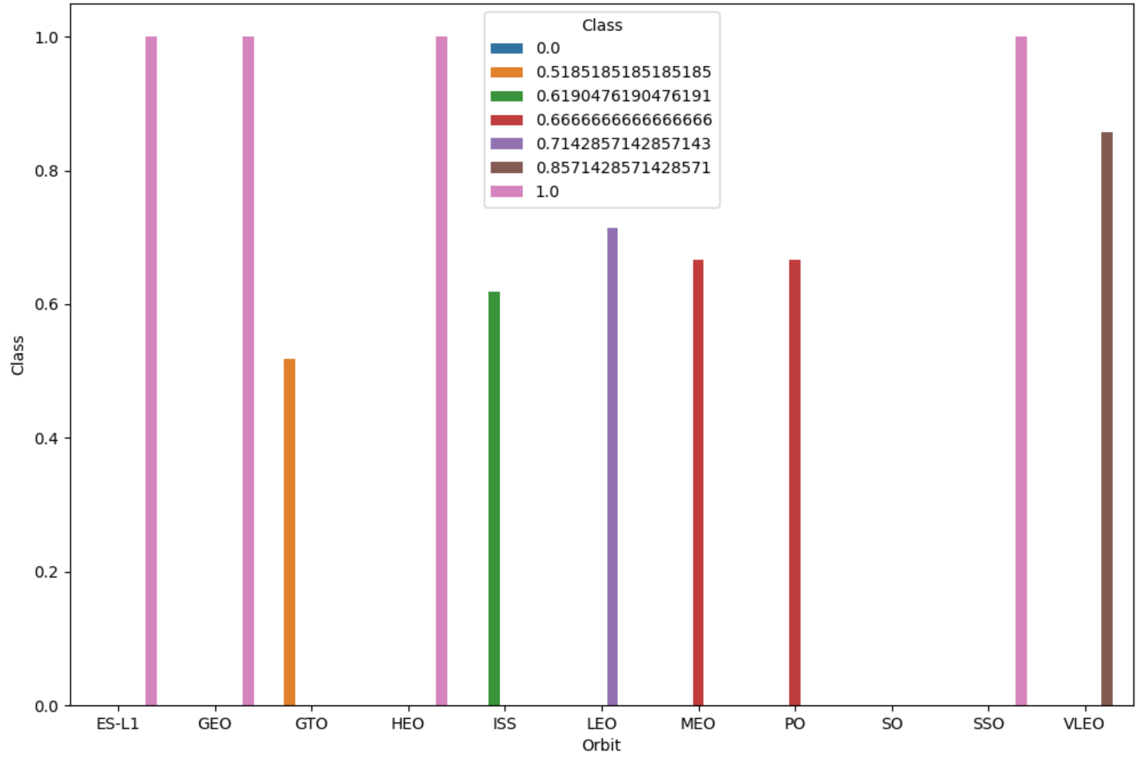


Flight success improved over time at all three locations. More flights are also conducted at site CCAFS SLC 40.

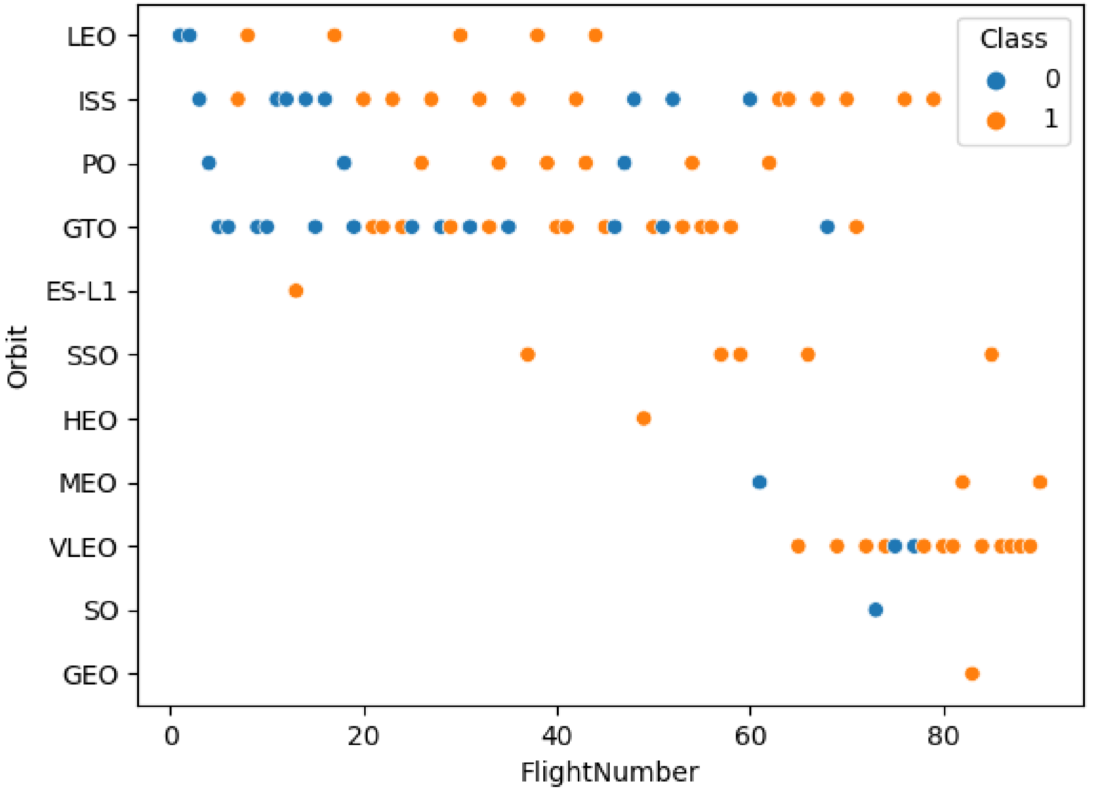
Payload mass at the launch sites



Orbit based on class success

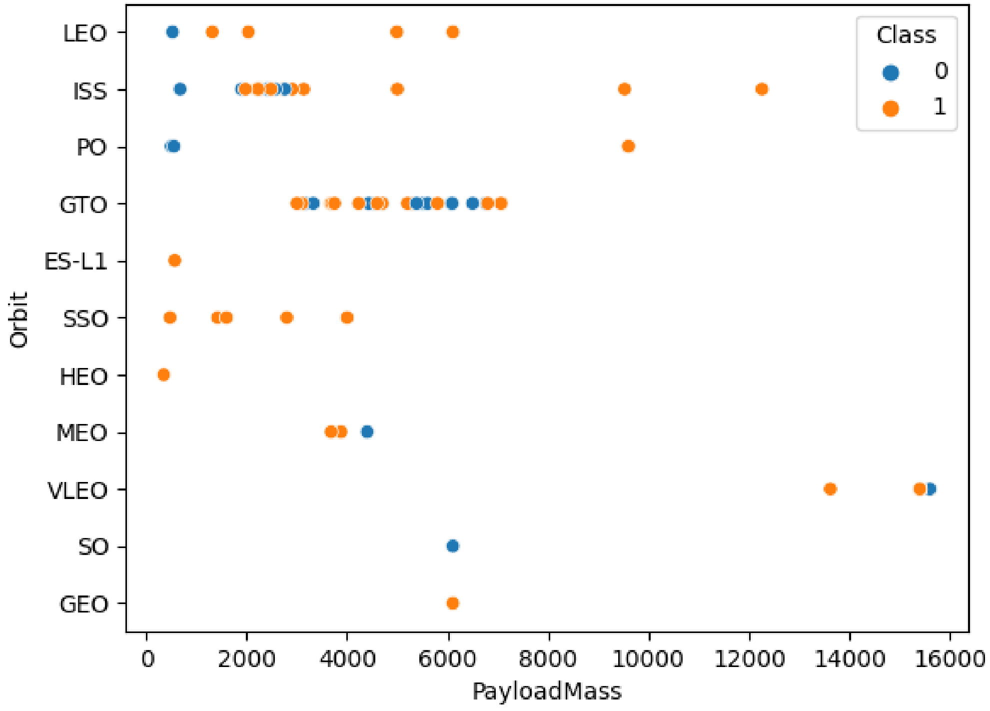


Flightnumber and orbit

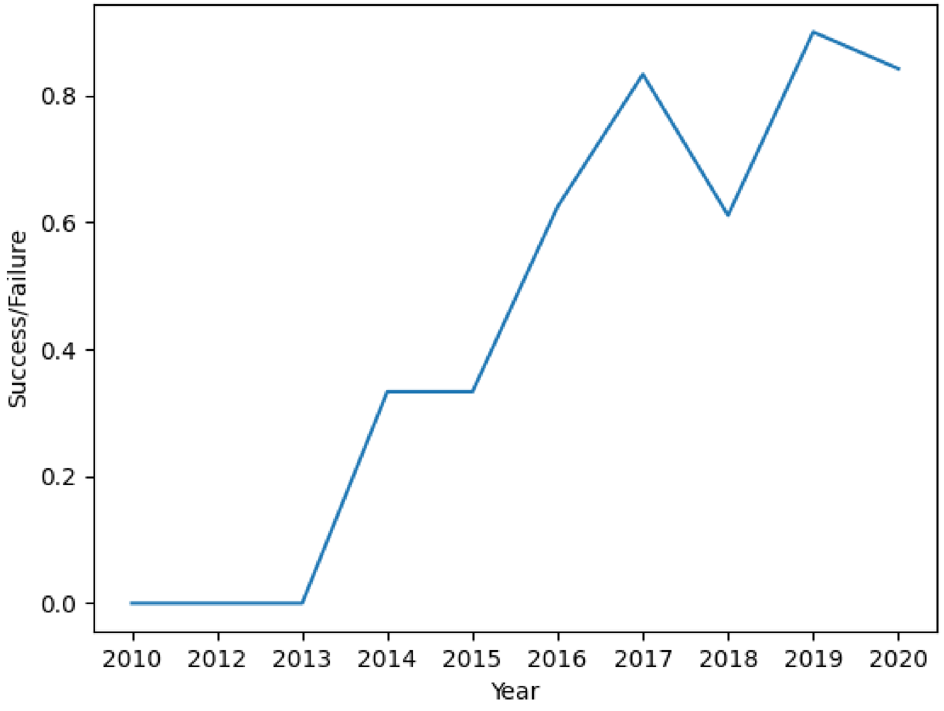


Payload mass and orbit

- There are heavy groupings of success and failures for both ISS and GTO
- GTO seems almost independent of payload mass in getting to orbit dude to success at 8000 kg
- ISS has greater number of successful missions than failures, based on the density of class 1 points.



Success rate over time



- The success rate improved over time.
- The first few years could be explained as the initial trial/prep stages going into later 2013 and into 2014.

Distinct Launch Sites

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
```

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

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

Info for Launch Sites - "CCA"

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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)
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
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
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 for NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_Mass FROM SPACEXTBL WHERE CUSTOMER LIKE 'NASA%'
```

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

Total_Mass
99980

Average Payload Mass for Falcon9

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_MASS FROM SPACEXTBL WHERE Booster_Version like 'F9 v1.1%'
```

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

AVG_MASS
2534.6666666666665

Earliest Success for Ground Pad

```
%sql SELECT MIN(Date) AS Early_Success FROM SPACEXTBL WHERE [Landing _Outcome] IS 'Success (ground pad)'
```

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

Early_Success
01-05-2017

Booster Version for Specific Payload Range

```
%sql SELECT Booster_Version AS Booster, [PAYLOAD_MASS__KG_] AS Payload FROM SPACEXTBL WHERE [Landing _Outcome] \
    LIKE 'Success (%drone%)' and [PAYLOAD_MASS__KG_] BETWEEN 4000 AND 6000
```

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

Booster	Payload
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Mission Outcomes

```
%sql SELECT COUNT(*) AS TOTAL, Mission_Outcome AS OUTCOMES FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

TOTAL	OUTCOMES
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Booster Versions at Max Payload

```
# Have to determine the payload mass before
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

Launch Records

```
%sql SELECT SUBSTR(Date, 4, 2) AS MONTH, [Landing _Outcome], Booster_Version, Launch_Site \
FROM SPACEXTBL WHERE [Landing _Outcome] IS 'Failure (drone ship)' LIMIT 2
```

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

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

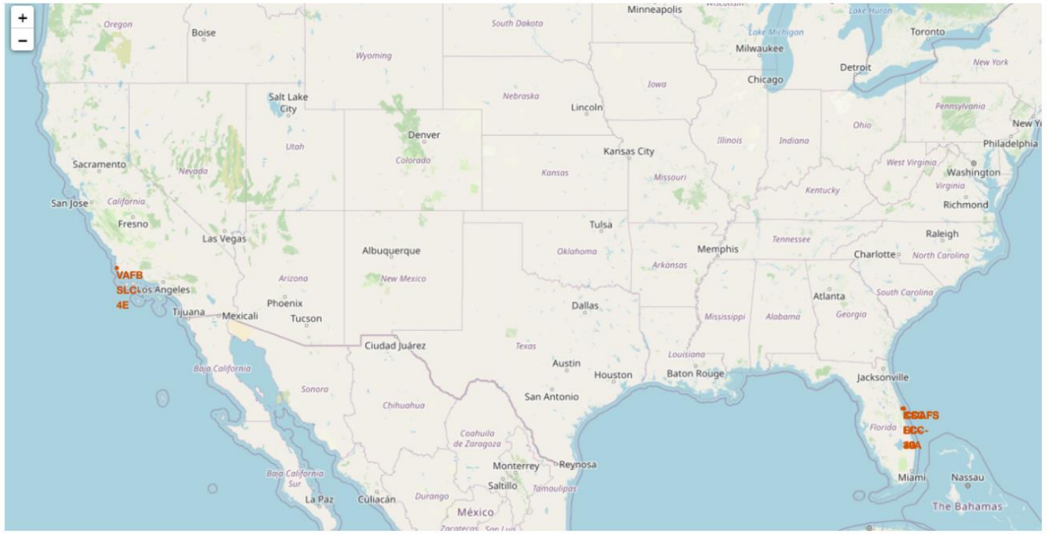
Landing Outcomes in Date Range

```
%sql SELECT [Landing _Outcome] AS Landing_Outcome, COUNT(*) AS N_Counts FROM SPACEXTBL WHERE DATE BETWEEN "04-06-2010" AND "20-03-2017" \
    GROUP BY [Landing _Outcome] ORDER BY N_Counts DESC
```

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

Landing_Outcome	N_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

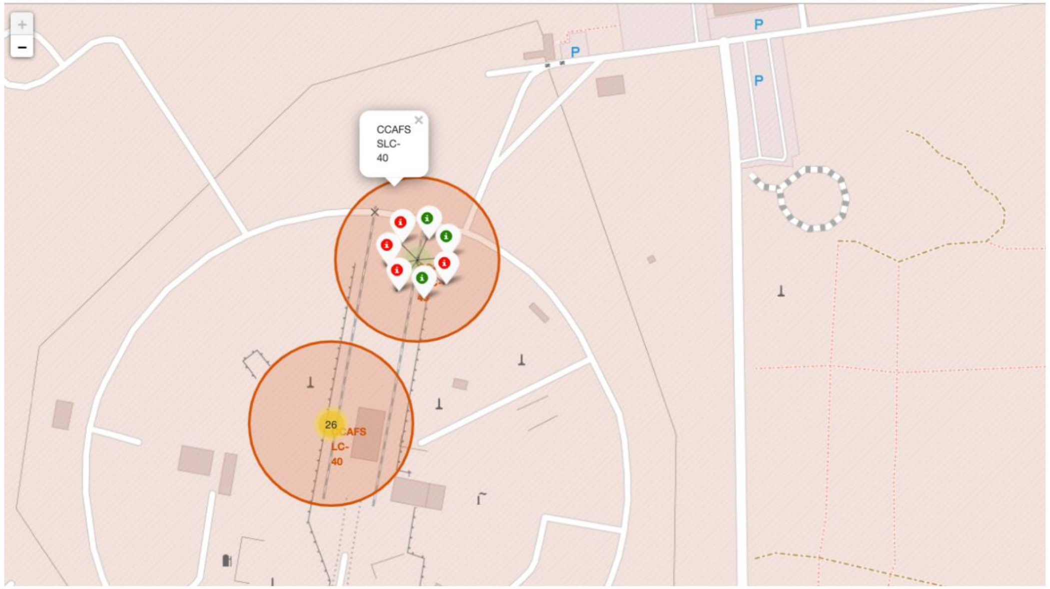
Launch Sites in Folium



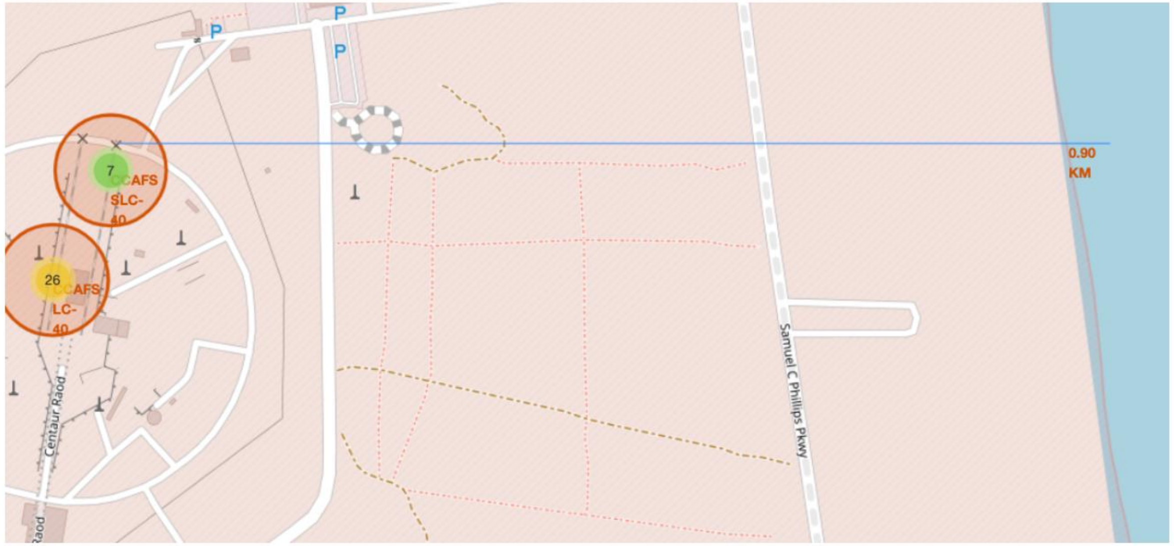
- The launch sites are located close to the ocean.
- That's typical for most rocket launches.
- As payloads are launched, re-entry, etc. Are typically crashed into the ocean due to being much safer than crashing around populations.

Color coded launch records

- The color coded points indicated on the map, provide insights into the successes of launches at the specified location.
- The one indicated here shows an equal amount of green (success) and red (failure) at the launch location: CCAFS SLC-40



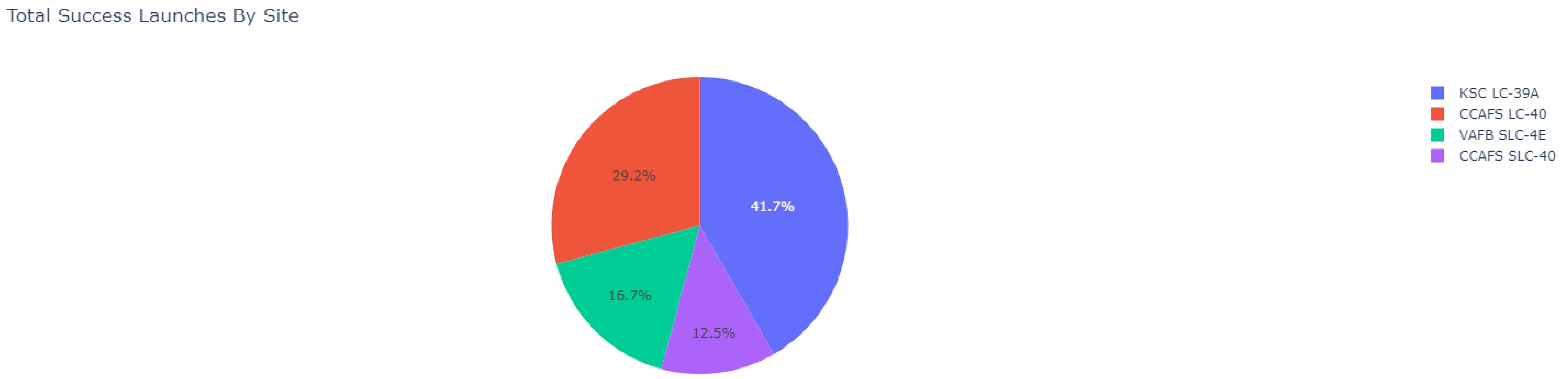
Distance to Launch Site



- At the same location as specified from the previous slide, this launch location is approximately 1 km away from the sea and located quite closely to a few side roads that lead and travel in proximity to the launch site.

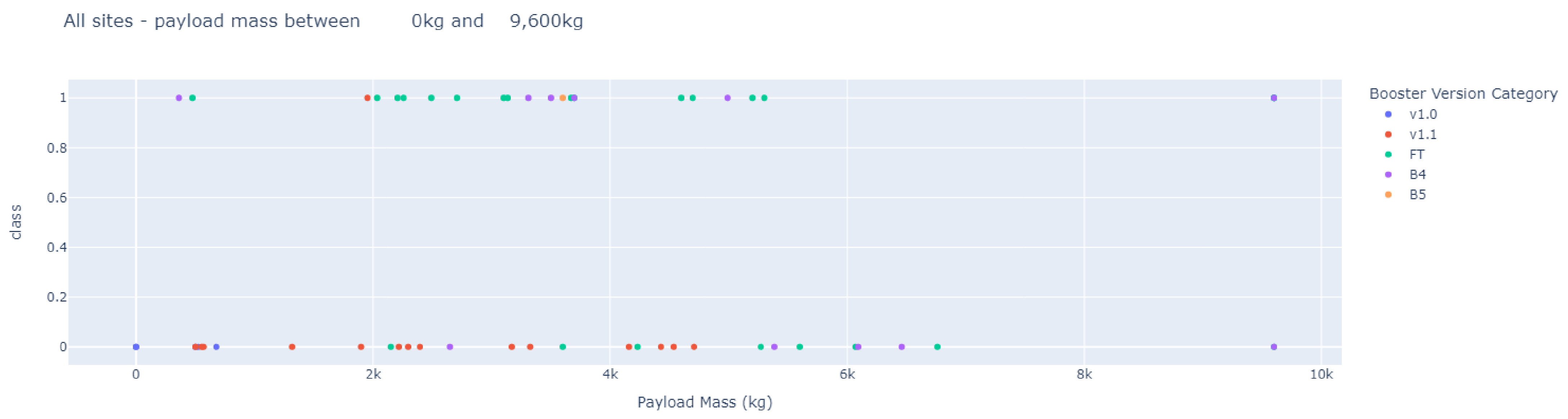
Plotly Visualization Results

The observation is made in that the location that has the most success launches is KSC LC-39A.



Plotly Visualization Results Contin.

The scatter plot shows that the booster version and payload mass are good indicators of successful launches.



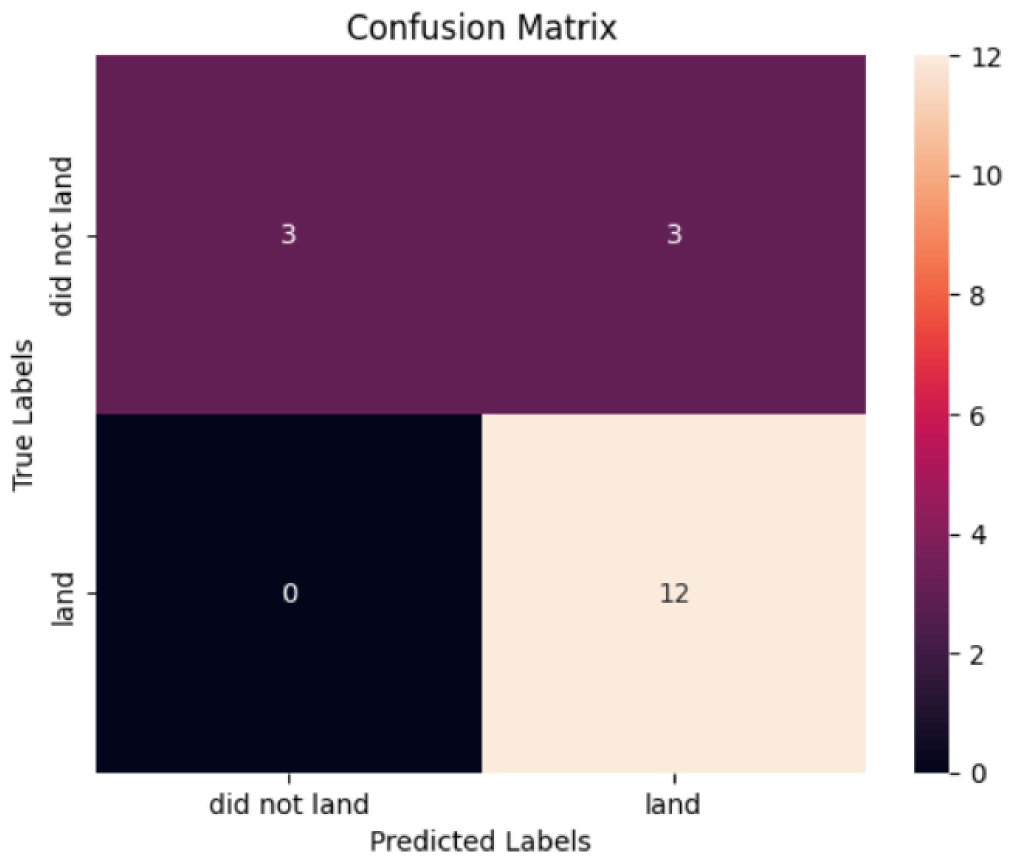
Predictive Analysis

- The computed accuracy for the models is as follows:

1. Logistic Regression: 83.3%
2. Support Vector Machine: 83.3%
3. Decision Tree Classifier: 83.3%
4. K-Nearest Neighbors (K-NN): 84.8%

- The K-NN classification model performed the best for this data. Although all four compared about the same. *The training and testing sizes used in this are relatively small.

Confusion Matrix of K-NN



- The testing size is 18 samples.
- 12/12 Were correctly classified as "land" when labeled as "land"
- 3/6 Were correctly classified as "did not land" when labeled as "did not land"
- 3/6 Were incorrectly labeled as "land" when labeled as "did not land" - False Positives

CONCLUSION

- The K-Nearest Neighbor model performed the best on the hold out testing set.
- Launches made at the KSC LC-39A location perform the best by having high greatest amount of successful launches.
- The launch sites indicated from the dataset are located close to the ocean.
- Although most of mission outcomes are successful, the rate of success continued to improve over time.

APPENDIX

- Additional modeling, feature engineering, model selection, and hyperparameter tuning can be employed to improve classification accuracy.

Thanks!