IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Utkarsh Goyal
7/4/23

Outline

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

Executive Summary

- As the SpaceX wants to make rockets more affordable and environmentally friendly, we are looking into our rockets launches more successful
- I was able to predict with a 94% accuracy when a rocket will be successfully launched to achieve this goal.
- Using LightGBM, we can make those predictions and find the optimal features to save database costs build a lean model.

Introduction: Background

- In this capstone for the IBM Applied Data Science course, I will try to work as a data scientist.
- I will try to conditions where rockets safely land for SpaceX.

Introduction: Business Problem

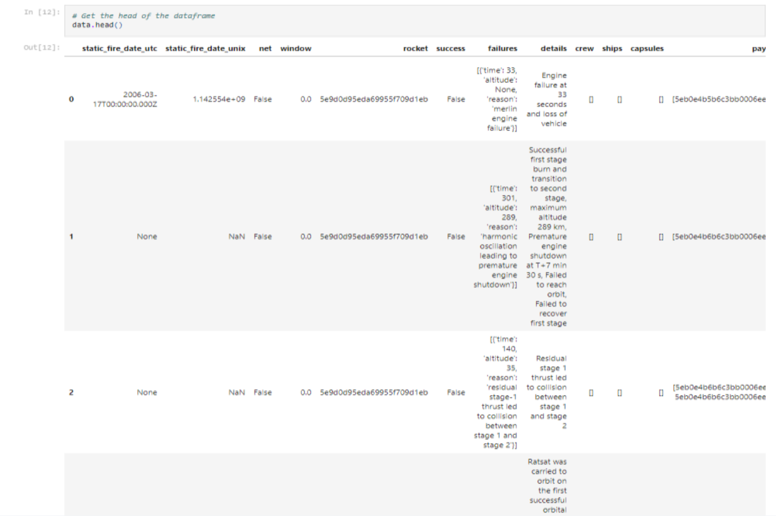
- Predict if the Falcon 9 first stage will land successfully
- SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore, if we can determine if the first stage will land then we save a lot money.

Section 1

Methodology

Data Collection

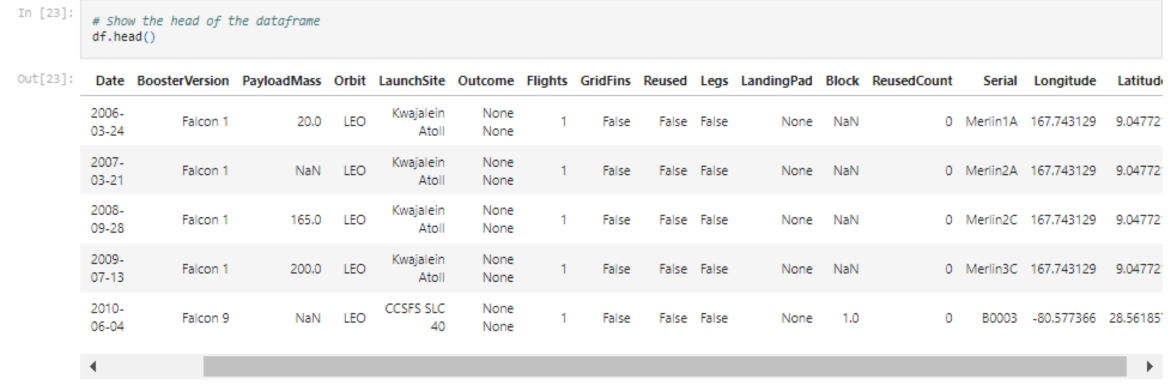
- Acquired historical launch data from [Open-Source REST API](#) for SpaceX
- Requested and parsed the SpaceX launch data using the GET request
- Clean the data to become more readable



In [12]:
```python
# Get the head of the dataframe
data.head()
```

Out[12]:

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details	crew	ships	capsules	pay
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	Engine failure at 33 seconds and loss of vehicle	[]	[]	[]	[5eb0e4b5b6c3bb0006ee
1	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 301, 'altitude': 289, 'reason': 'harmonic oscillation leading to premature engine shutdown'}]	Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage	[]	[]	[]	[5eb0e4b6b6c3bb0006ee
2	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 140, 'altitude': 35, 'reason': 'residual stage-1 thrust led to collision between stage 1 and stage 2'}]	Residual stage 1 thrust led to collision between stage 1 and stage 2	[]	[]	[]	[5eb0e4b6b6c3bb0006ee 5eb0e4b6b6c3bb0006ee
								Ratsat was carried to orbit on the first successful orbital				



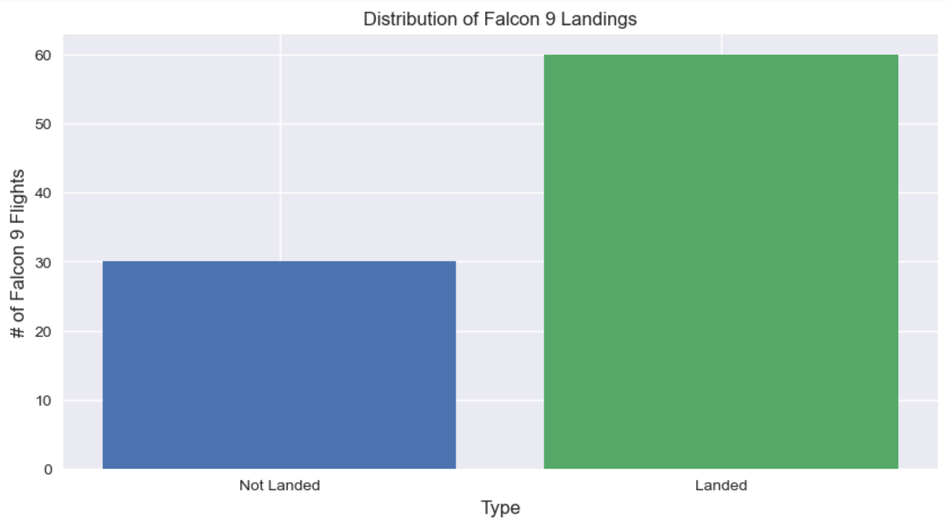
In [23]:
```python
# Show the head of the dataframe
df.head()
```

Out[23]:

Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitud
2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.04772
2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.04772
2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.04772
2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.04772
2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.56185

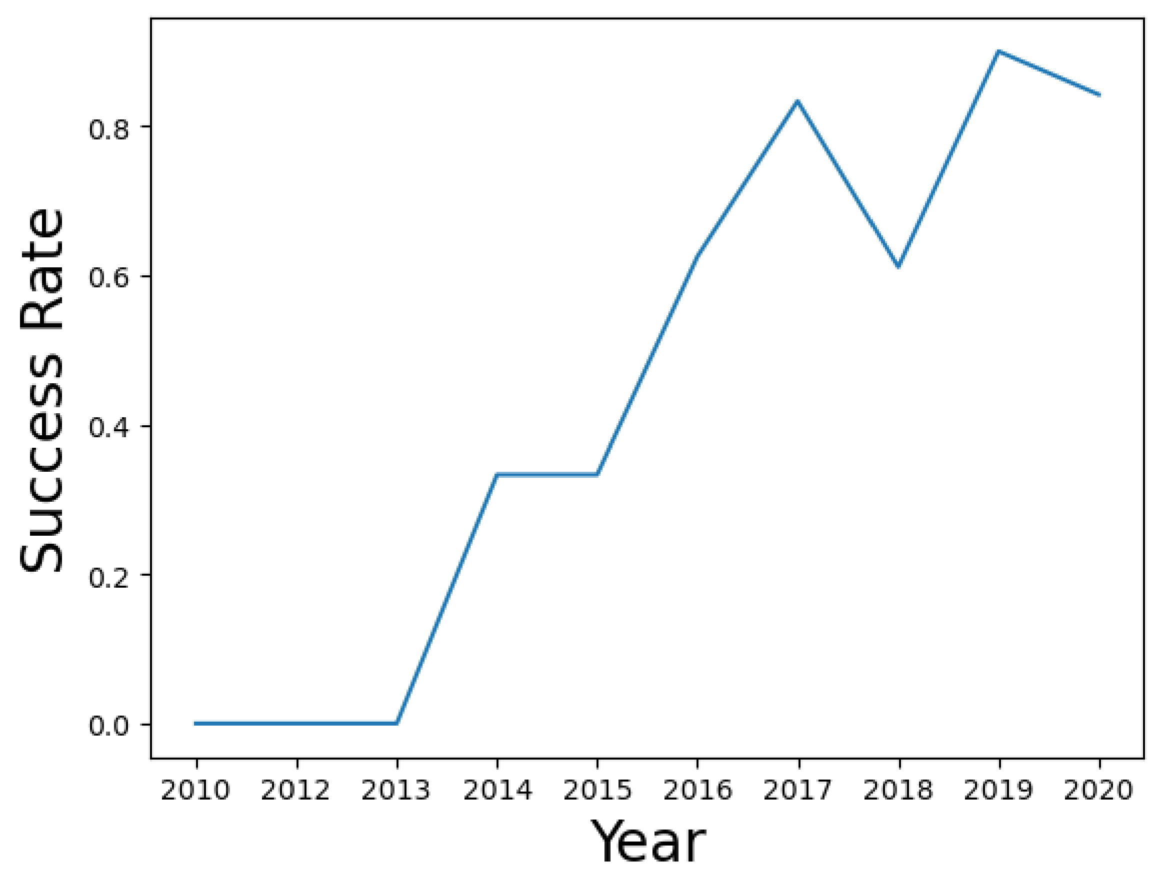
Data Wrangling

- Filtered the data to only include Falcon 9 launches
- Replaced missing payload mass values from classified missions with mean
- Created a landing outcome training label, ‘Class’, from 'Outcome' column
  - Class = 0; first stage booster did not land
    - None None; not attempted
    - None ASDS; unable to be attempted due to launch failure
    - False ASDS; drone ship landing failed
    - False Ocean; ocean landing failed
    - False RTLS; ground pad landing failed
  - Class = 1; first stage booster landed
    - True ASDS; drone ship landing succeeded
    - True RTLS; ground pad landing succeeded
    - True Ocean; ocean landing succeeded



EDA with Data Visualization

- FlightNumber vs PayloadMass
- FlightNumber vs LaunchSite
- Payload vs LaunchSite
- Orbit type vs Success rate
- FlightNumber vs Orbit type
- Payload vs Orbit type
- Success rate over time

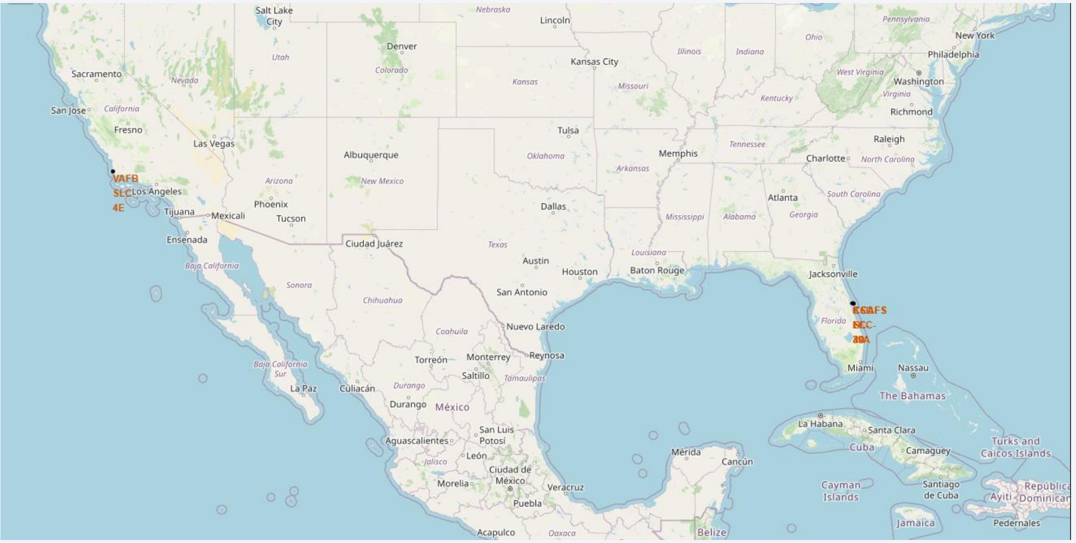


EDA with SQL

- Ran SQL queries to understand:
  - Launch sites
  - Payload masses
  - Booster versions
  - Mission outcomes
  - Booster landings

Build an Interactive Map with Folium

- Marked all launch sites on a map
- Marked the successful/failed launches for each site on map
- Calculated the distances between a launch site to its proximities
  - Railways
  - Highways
  - Coastlines
  - Cities



Build a Dashboard with Plotly Dash

- Pie chart showing success rate by launch site
- Scatter chart showing payload mass vs. landing outcome by booster version with range slider for limiting payload amount
- Drop-down menu to choose between all sites and individual launch sites

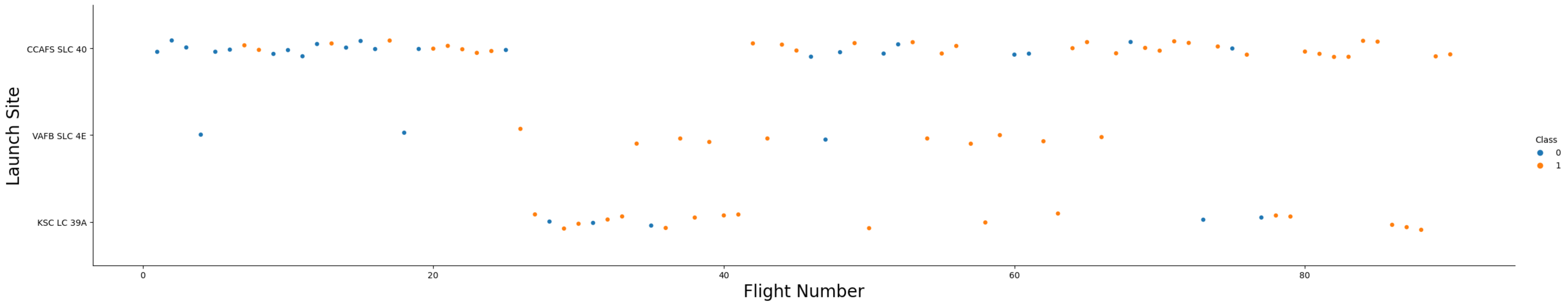
Predictive Analysis (Classification)

- Split into training data and test data
- Fit the training data to various model types
  - Logistic Regression
  - SVM
  - Decision Tree Classifier
  - KNN Classifier
  - Random Forrest Classifier
  - Xgboost Classifier
  - LightGB Classifier
  - ANN
- Used a GridSearchCV to select the best hyperparameters for each model
- Evaluated accuracy of each model using test data to select the best model

Section 2

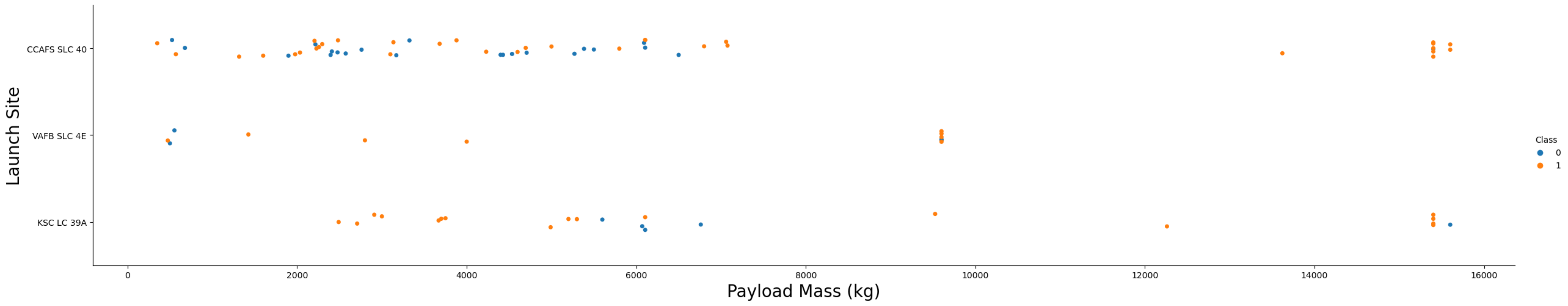
Insights drawn from EDA

Flight Number vs. Launch Site



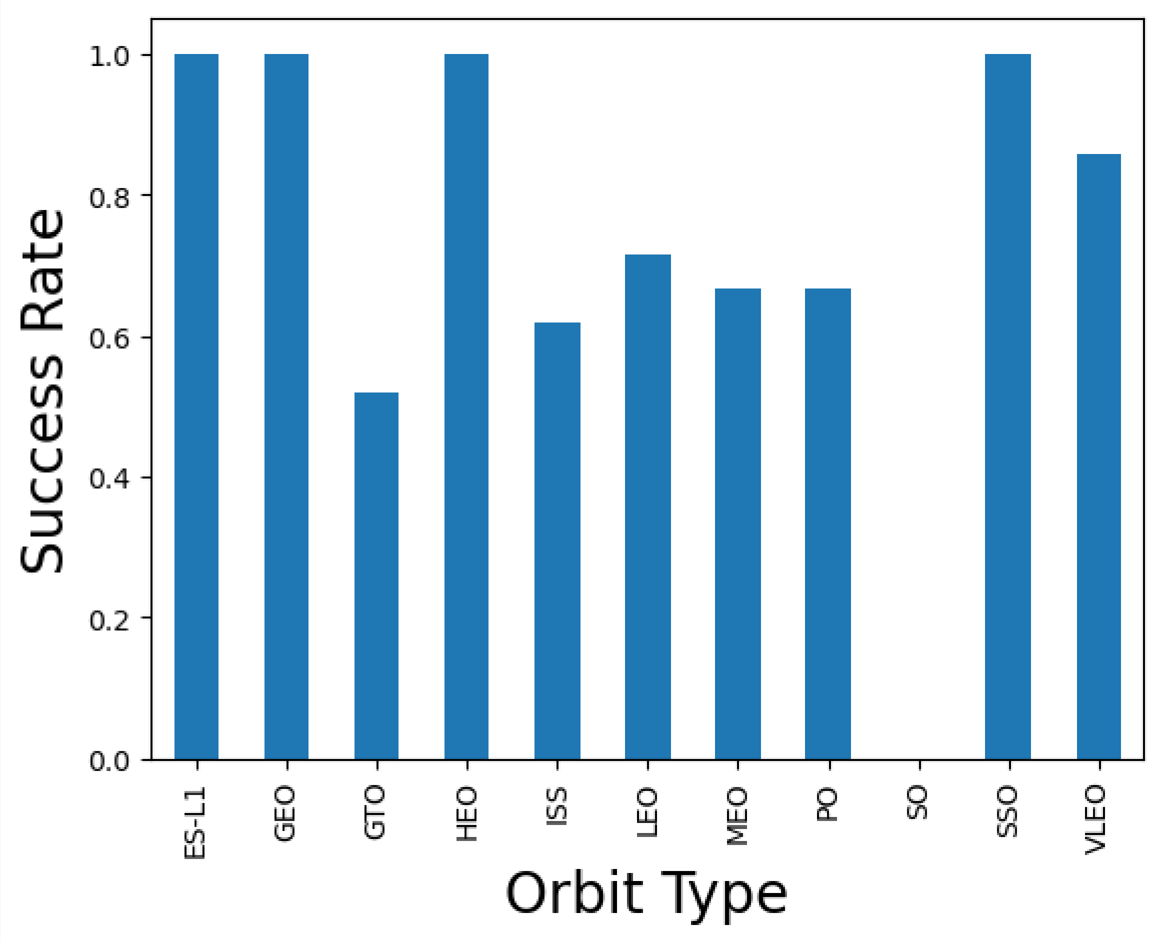
CCAFS SCL 40 has a lower success rate compared to the other two launch sites.

Payload vs. Launch Site



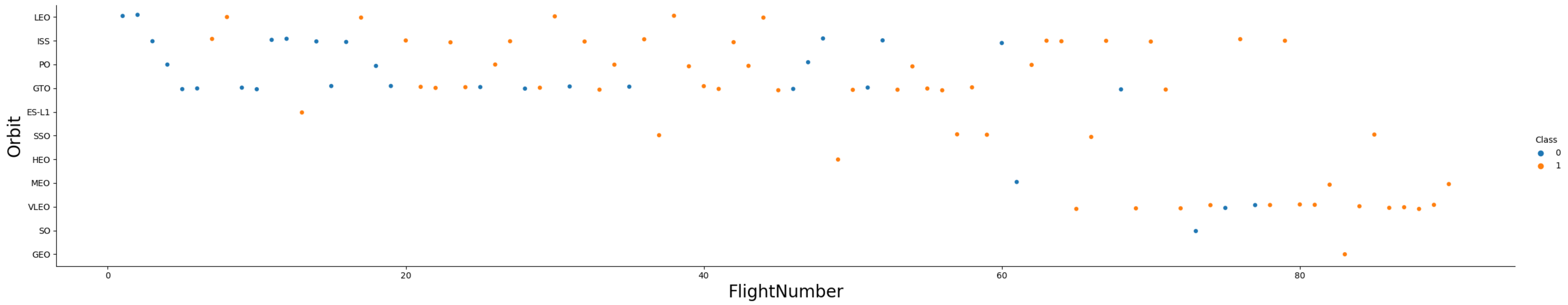
CCAFS SLC 40 and KSC LC 39A appear to have a higher success rate for heavier payloads, but this also means increased costs.

Success Rate vs. Orbit Type



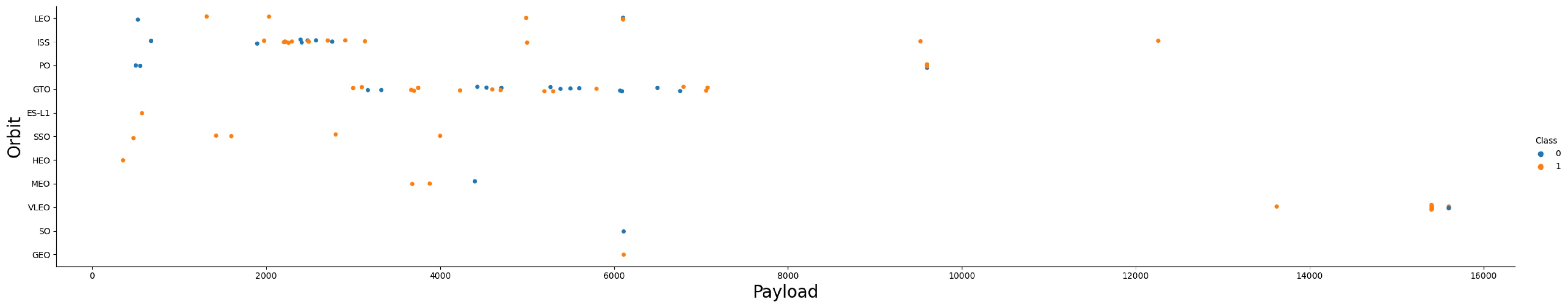
SO has not had a successful flight. The GTO has the lowest success rate, after SO.

Flight Number vs. Orbit Type



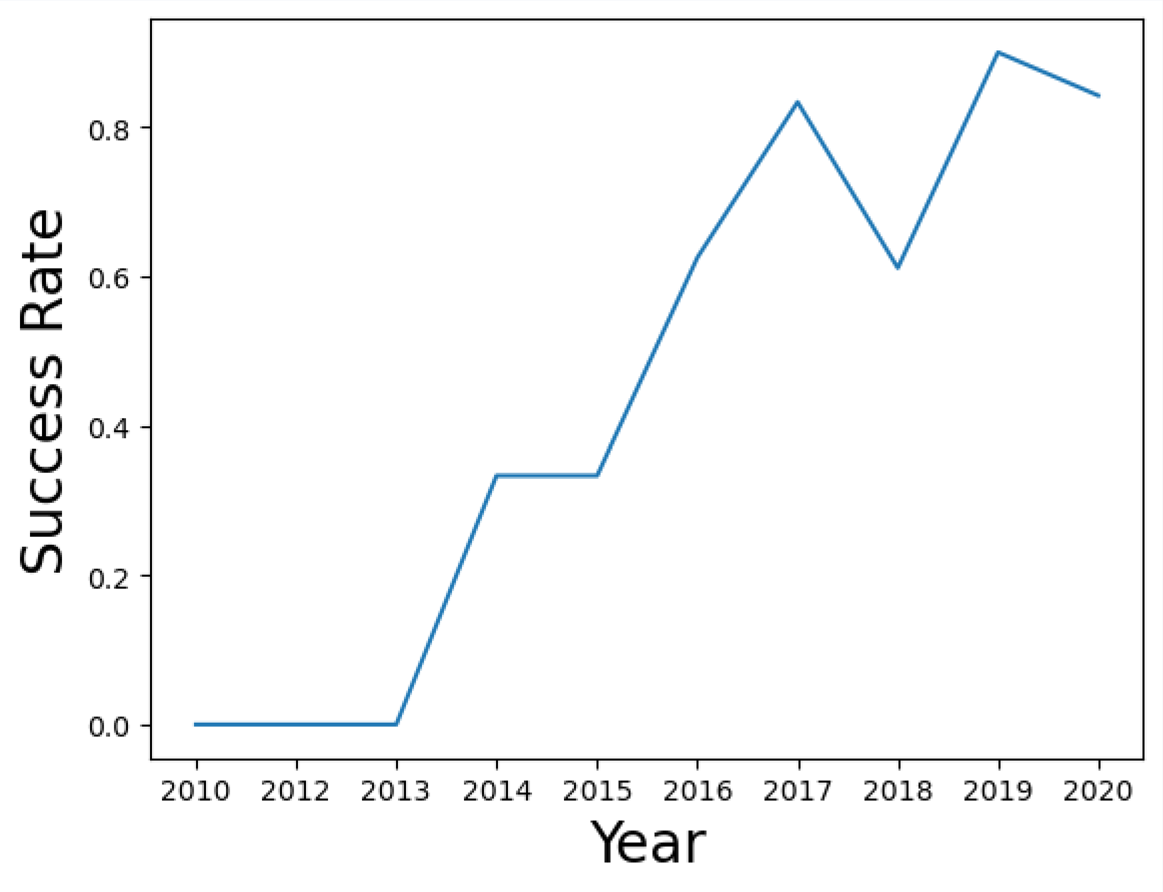
We have more successes in later flight numbers.

Payload vs. Orbit Type



Lighter payloads are recommended for GTO orbits, however heavier payloads are better for VLEO and ISS.

Launch Success Yearly Trend



We learnt from our mistakes and improved as years went on. Now we have 80% success rates.

All Launch Site Names

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

1
45596

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

1
340

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing__Outcome = 'Success (ground pad)';
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

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
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 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

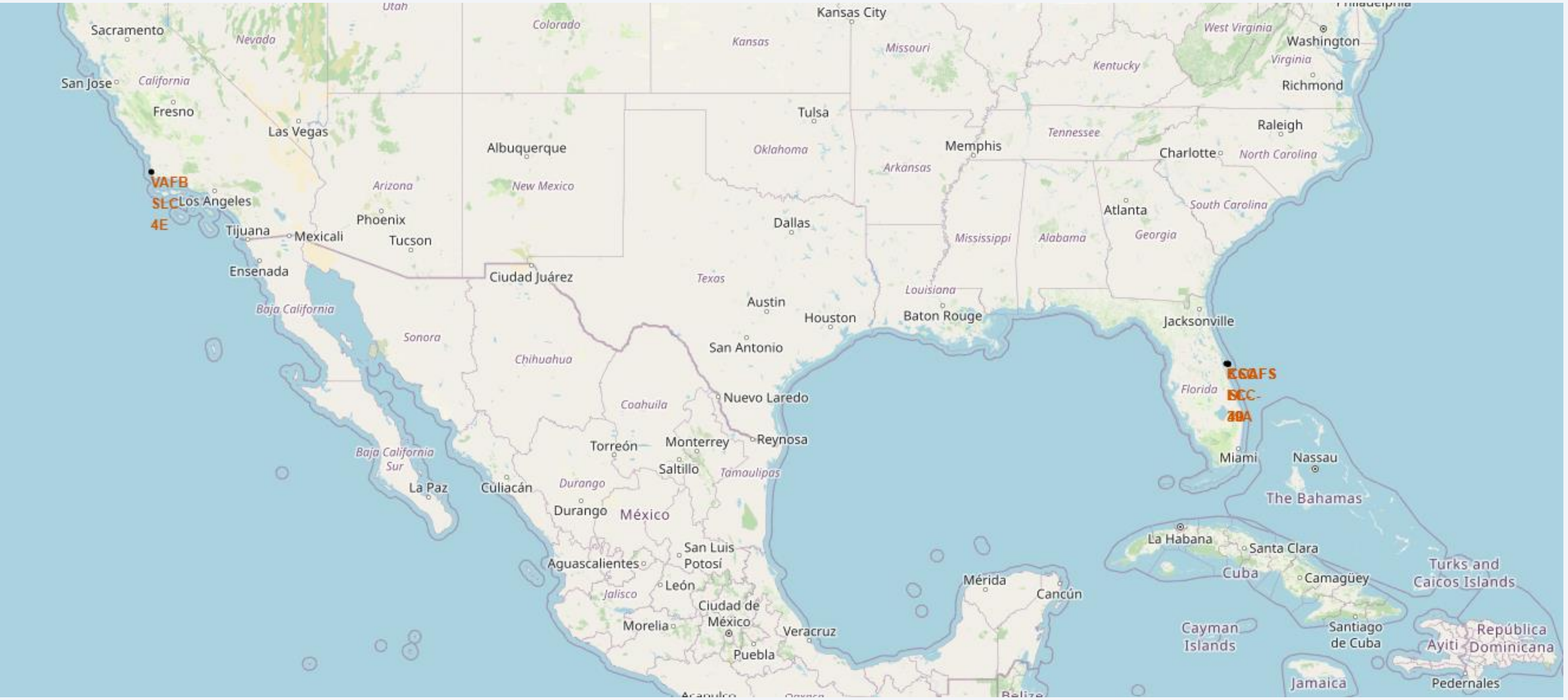
```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

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

Section 3

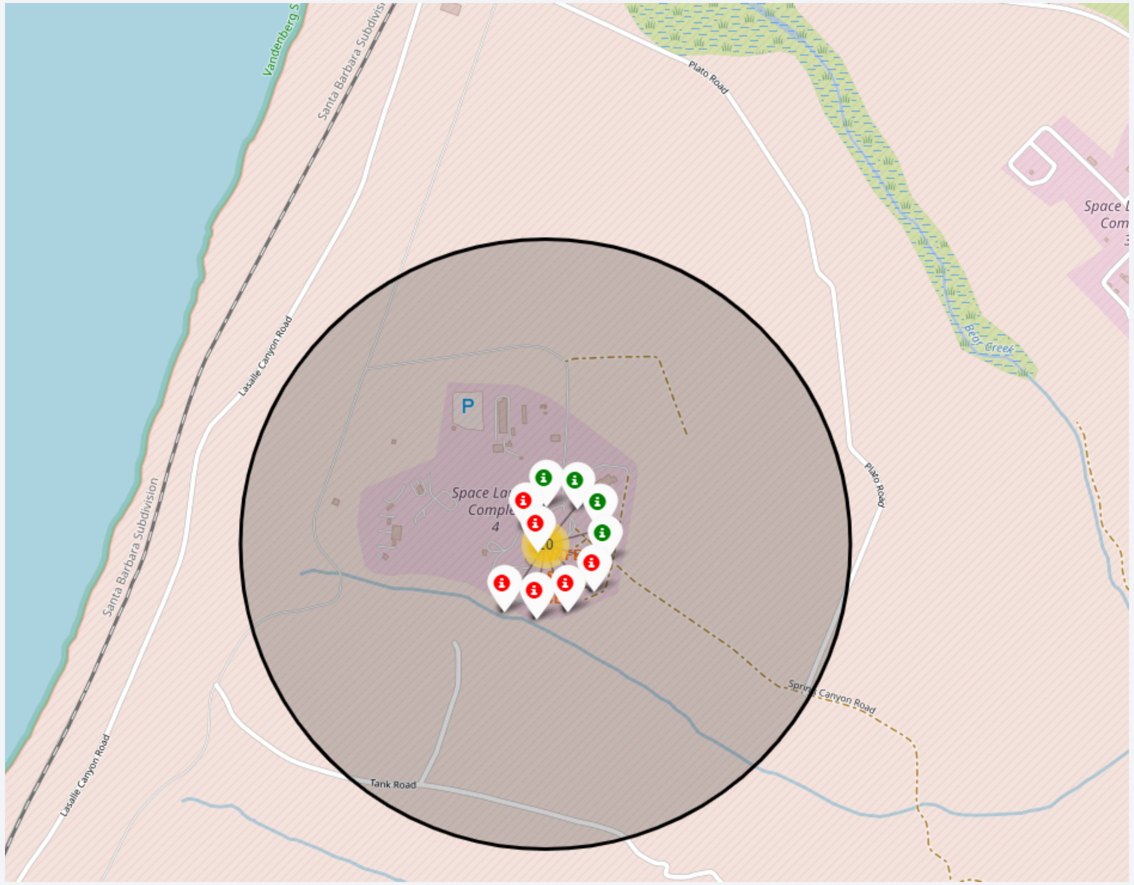
Launch Sites Proximities Analysis

All Launch Sites



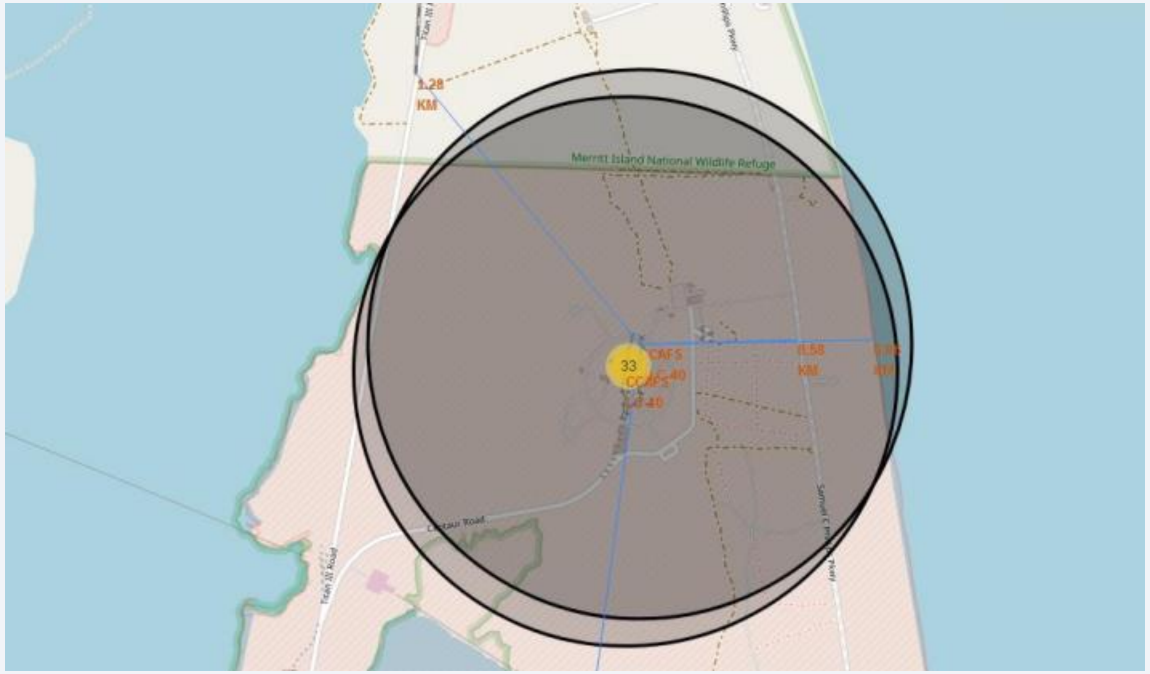
- ¾ launch sites are in Florida
- They are all near the equator

Success/Failures in each locations



- Looking at VAFB SLC-4E
- There are more failures at this location

Launch site proximities

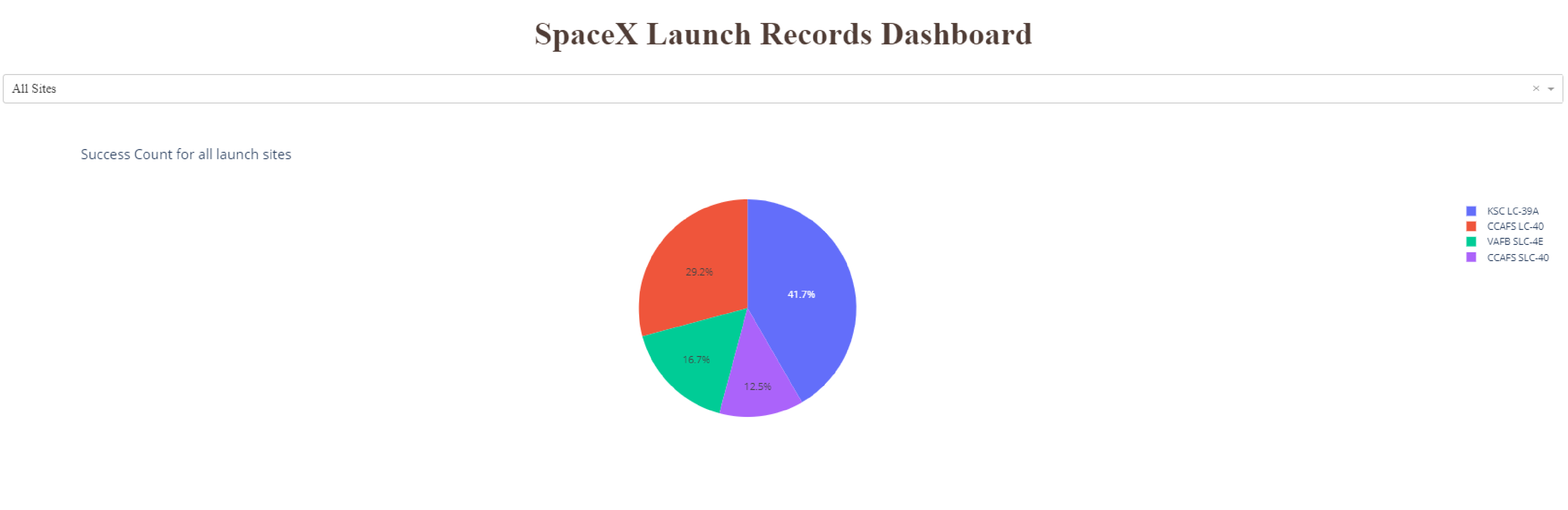


- Looking at CCAFS SLC-40
- It is very far from the city, ~50km to minimize danger

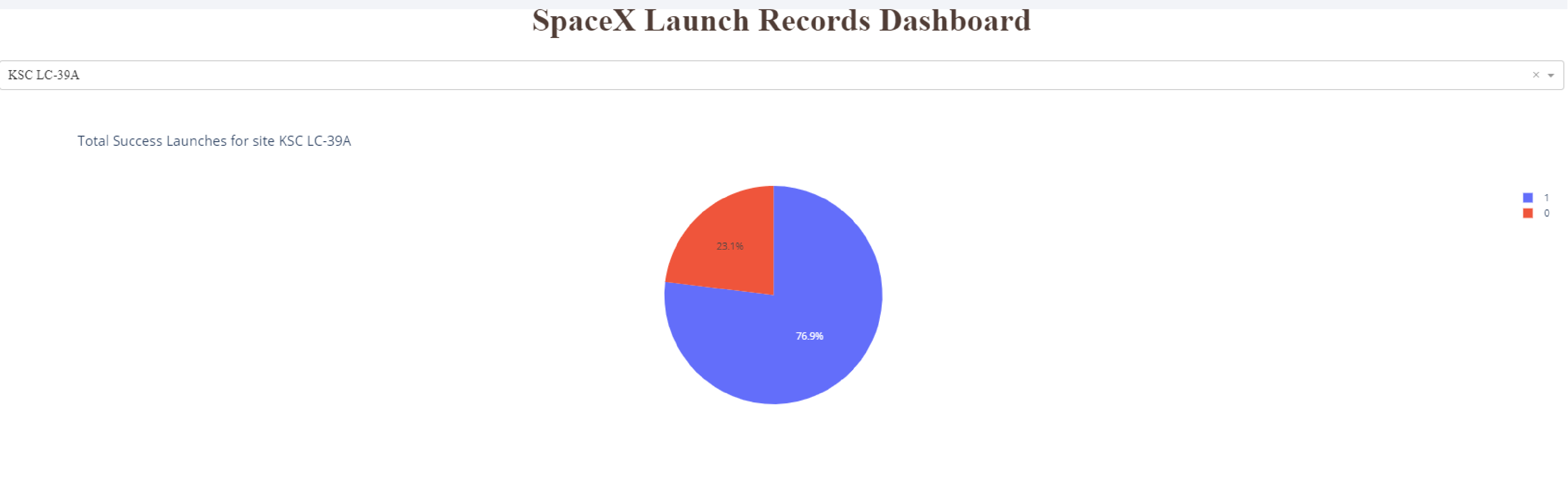
Section 4

Build a Dashboard with Plotly Dash

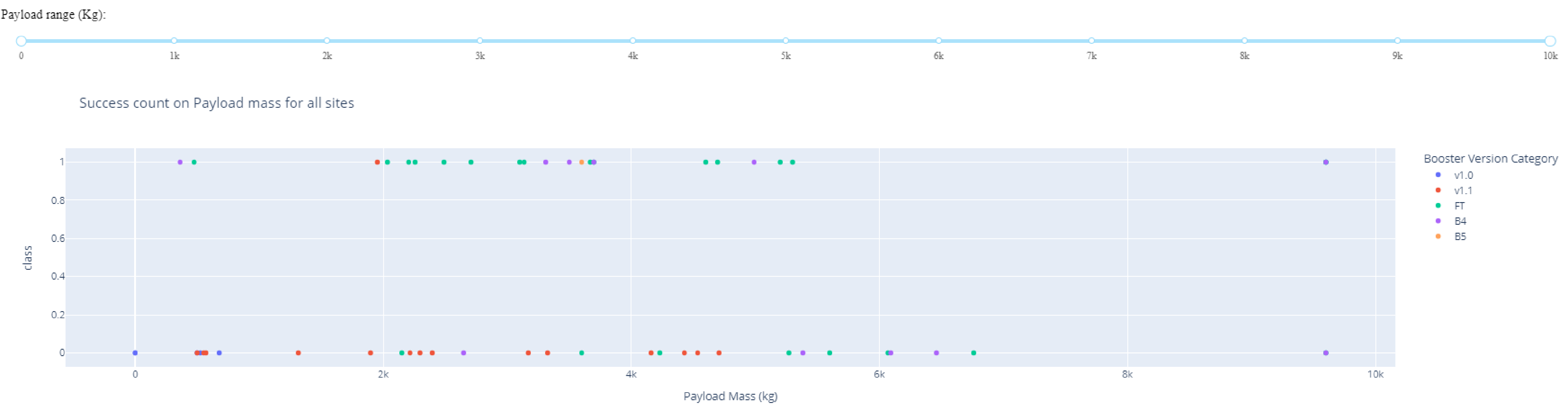
Success Rate across all launch sites



KSC LC-39A success rates



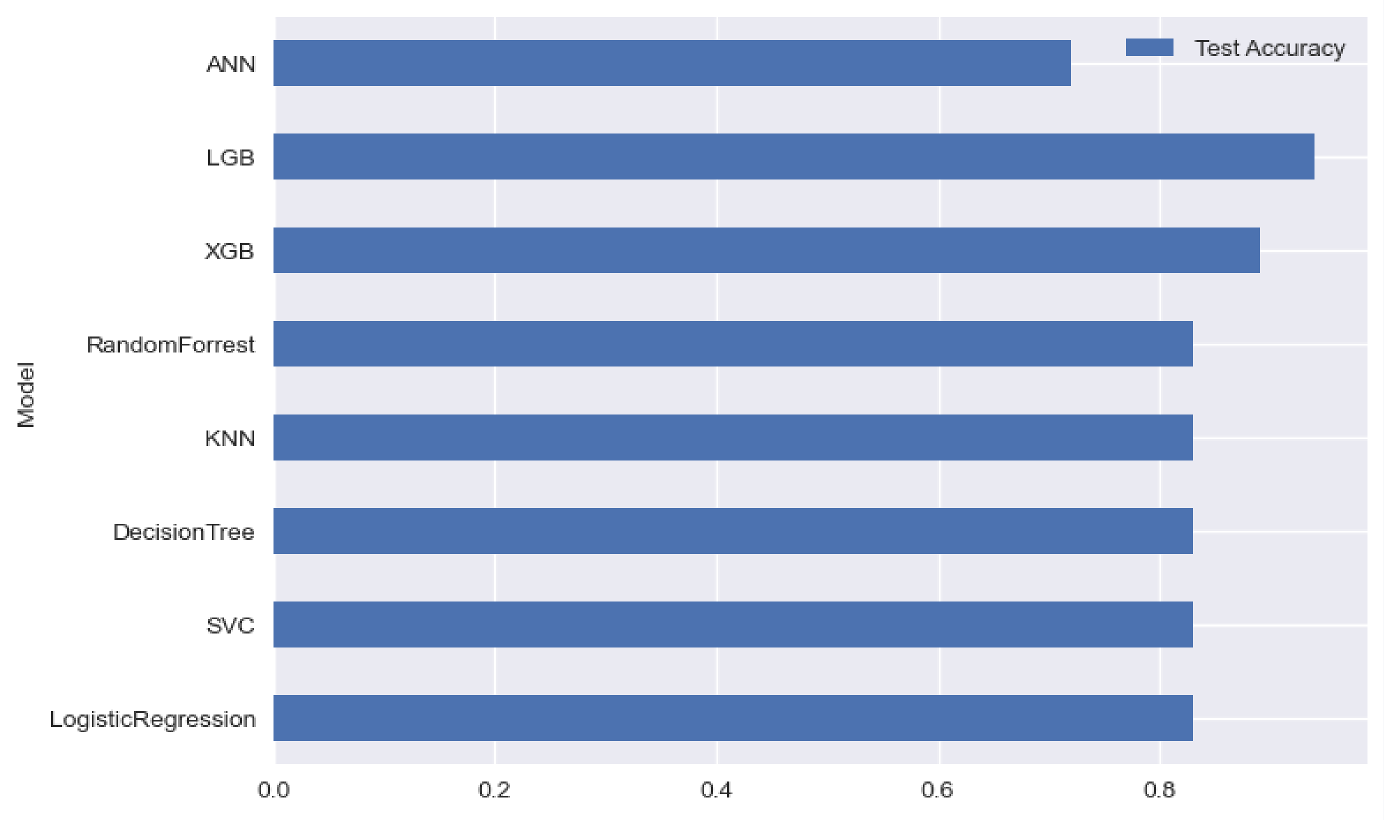
Payload vs. Launch Outcome



Section 5

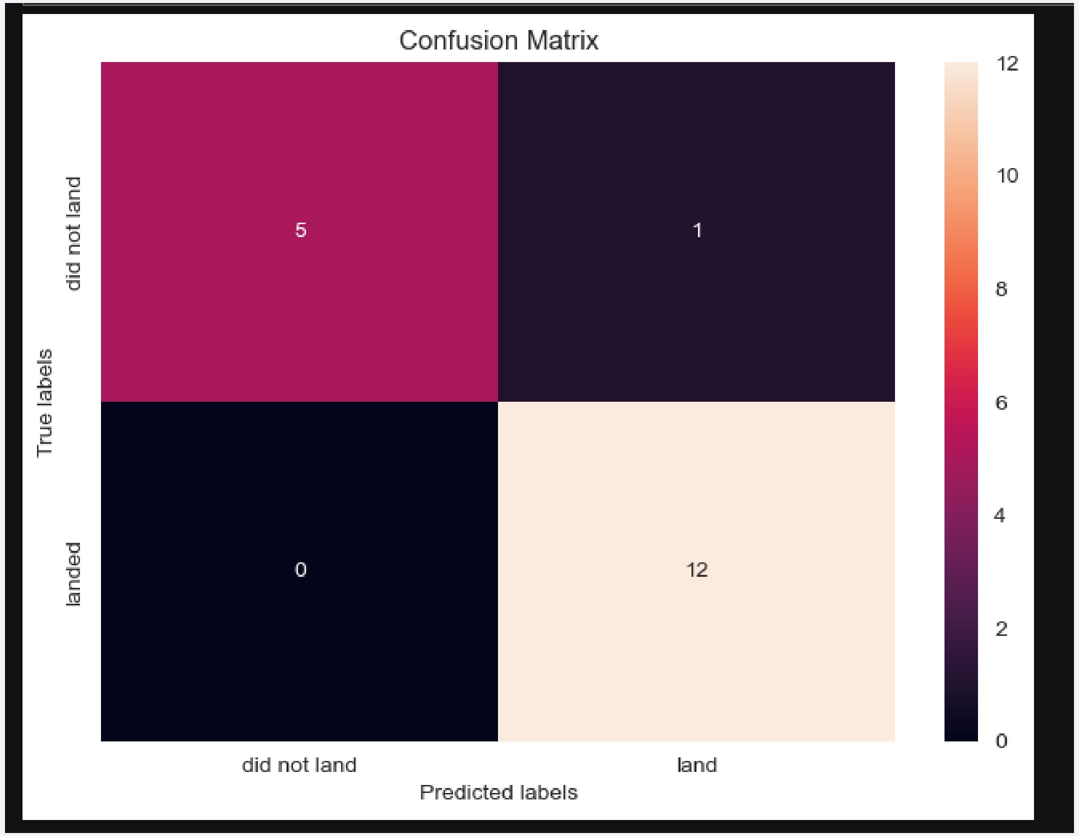
Predictive Analysis (Classification)

Classification Accuracy



- Most Models have 83% test accuracy
- LightGBM has the highest test accuracy, 94%
- Neural Net has the lowest, 72%

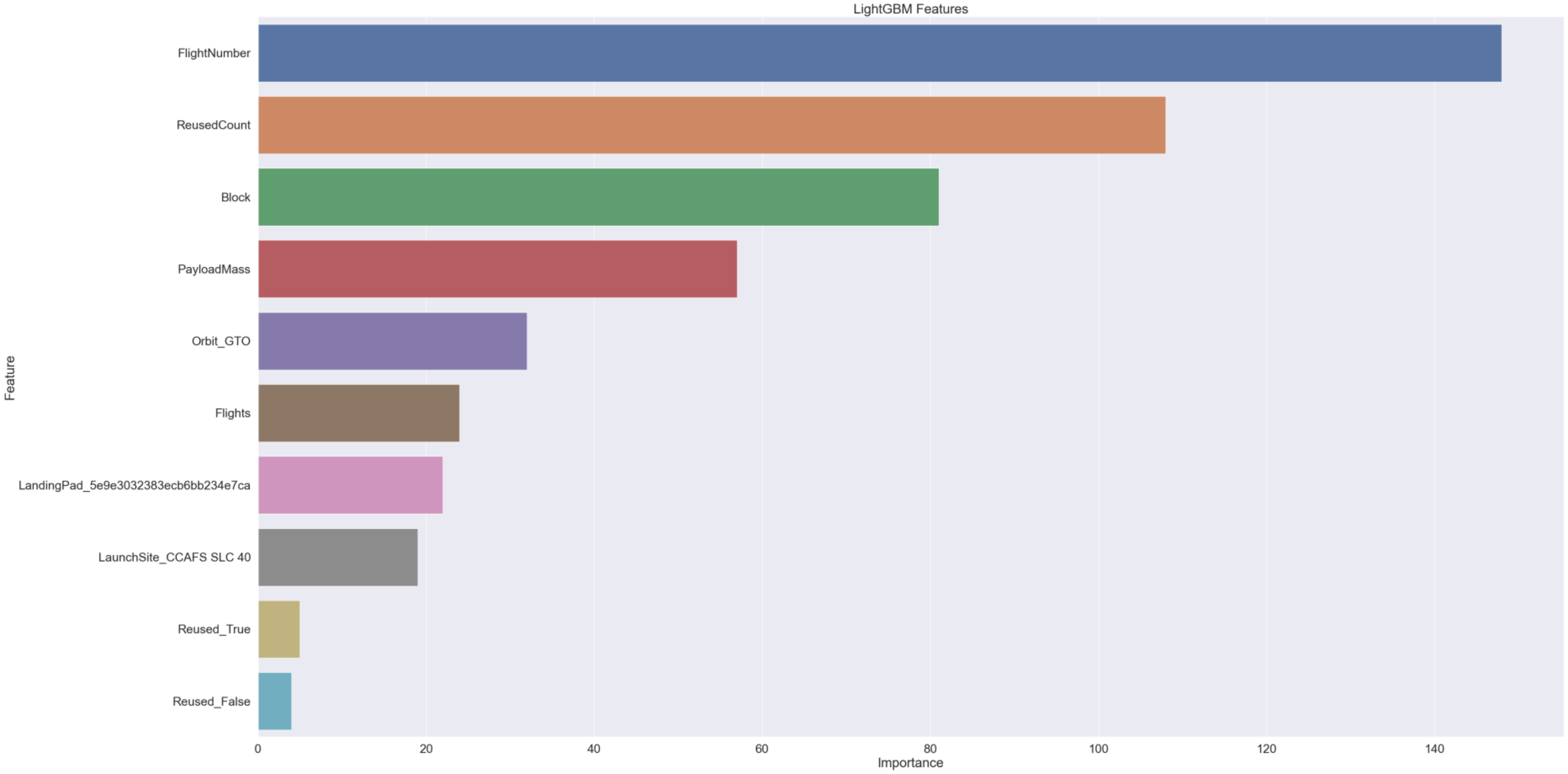
LGB Confusion Matrix



Classification Report:

	precision	recall	f1-score	support
0	1.00	0.83	0.91	6
1	0.92	1.00	0.96	12
accuracy			0.94	18
macro avg	0.96	0.92	0.93	18
weighted avg	0.95	0.94	0.94	18

Feature Importance



Conclusions

- We can predict with a 94% accuracy when a rocket will land
- We can save our costs by only launching optimal rockets with optimal features such as payloads at launch sites to the correct orbit
- We can improve our model by having more flights with the optimal features to refine and tune our model until we make the best rockets
- We can save on data storage by only maintaining the important attributes, when we have more data

Thank you!