IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mukundan Kuthalam
11/10/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Predict whether SpaceX could reuse its boosters using its historic flights
- Useful for a competitor that will have to contend with cheapness of SpaceX flights
- Some interesting conclusions from analysis and experiments:
  - Landings got more successful with experience
  - Correlated with success: the CCAFS LC-40 site, payload masses in the range of about 2000 kg to 3700 kg, and the FT booster
  - Classifier has accuracy of 83.3%
- Conclusion: Mission parameters and launch site can be established in accordance with what has been successful, and we have a reliable classifier to check if the rest of the mission details can lead to reuse

Intro - Background

- SpaceY mission: Ensure the best-quality commercial space travel
- Main competition: SpaceX with its relatively cheap missions
  - Stage 1 booster reuse
- Learning from SpaceX's experience => Learn from their data
- **Can we predict when SpaceX will reuse its boosters by using its launch history to predict when a landing will be successful, thus leading to cheaper flights of our own?**
  - Use all launch data for analysis
  - Focus on Falcon 9 (latest missions) for prediction

Intro – Conclusions in some detail

- C1: Landings were more likely to be successful as time went on, regardless of payload mass or orbit types (53 successes from March 2017 onwards, compared to 8 pre-March 2017)
- C2: Launch sites always kept away from cities but stayed close to coasts, roads, and railways.
- C3: The CCAFS LC-40 site, payload masses in the range of about 2000 kg to 3700 kg, and the FT booster are the most “proven” components of a successful landing (i.e., correlate to more successes).
- C4: There is no clear winner in which machine learning model yields the best prediction but all models yield a test accuracy of about 83.3%

Section 1

Methodology

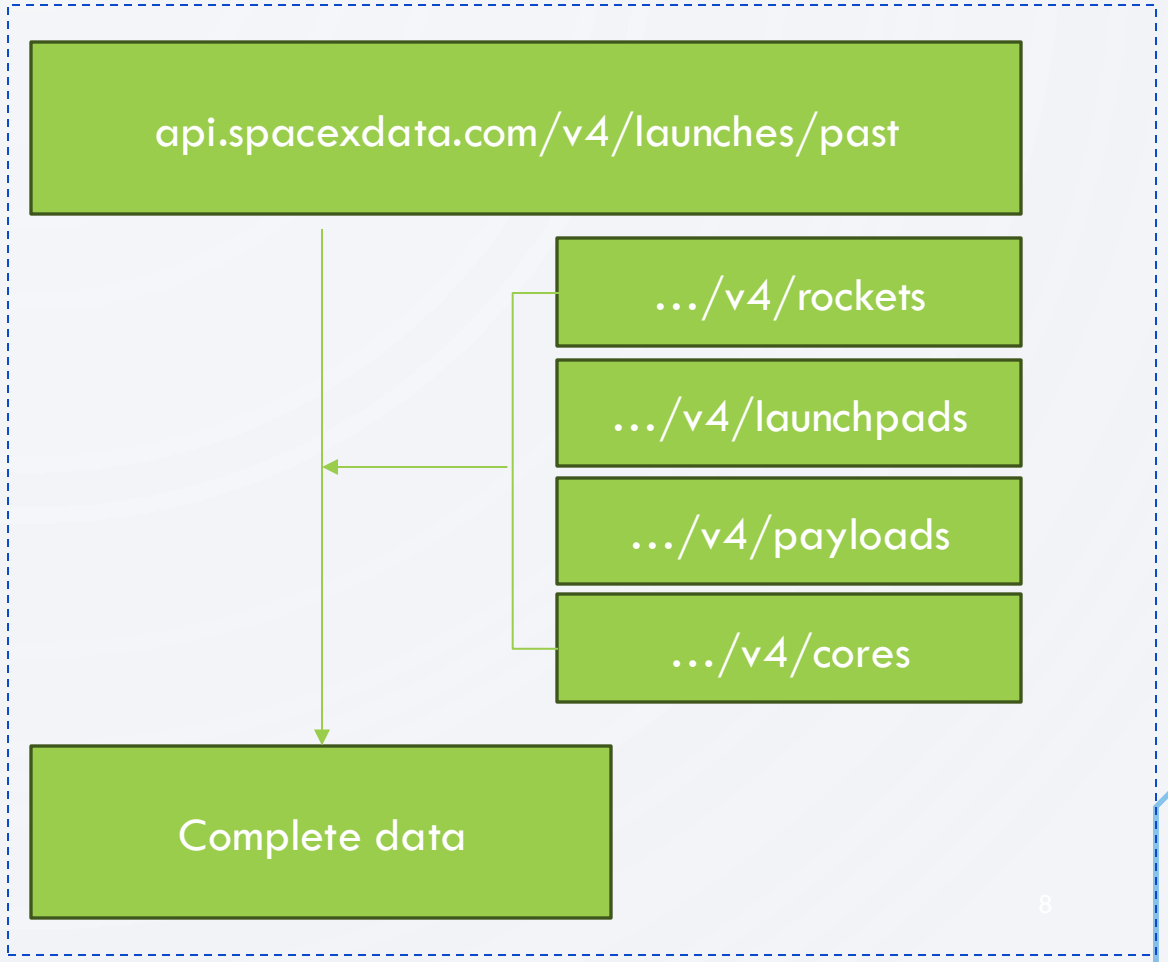
Methodology

Executive Summary

- All data was collected using the r/SpaceX API, including mission dates, booster versions, landing outcomes, and several other details
- Missing payload mass data was filled using the mean and categorial vars were one-hot encoded
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models with 20% of data held out for testing and a final test accuracy of about 83.3%

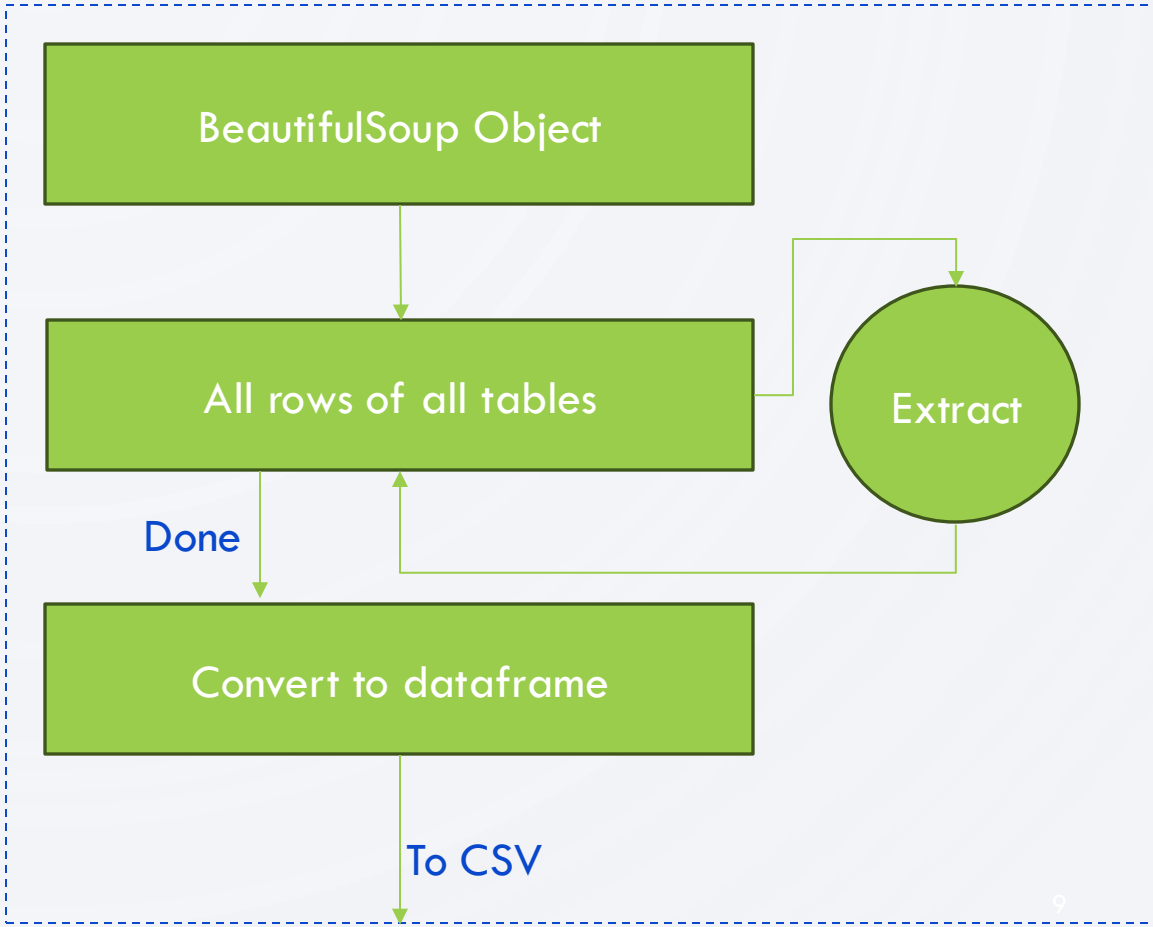
Data Collection – SpaceX API

- API calls to the r/SpaceX group
- Past launches as initial endpoint
  - Remaining info gathered from rockets, launchpads, payload and core endpoints
- Most data from cores endpoint (e.g., existing reuse data, landing outcome)
- Filtered to only Falcon 9 launches for prediction
- GitHub URL: https://github.com/kuthalam/Data_Science_Capstone/blob/main/Capstone_API-Data-Collection.ipynb



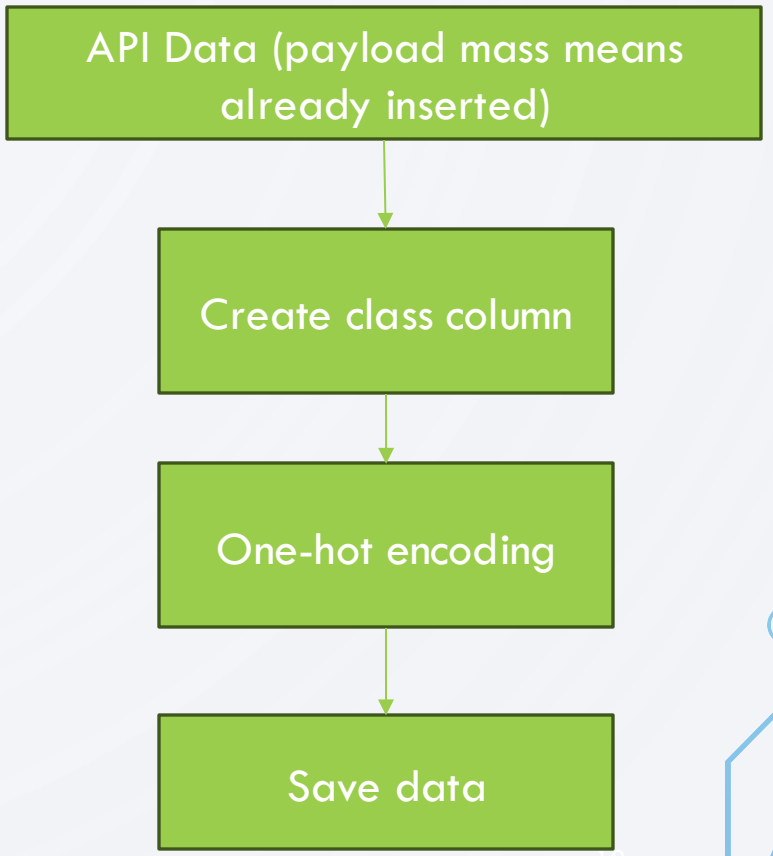
Data Collection - Scraping

- Used BeautifulSoup4 to iterate through tables on Wikipedia
  - Parsed through tables and rows for extra info like whether a landing was attempted
- Raw data from Wikipedia to Python dict to Pandas dataframe to CSV
- Comes up again during analysis
- GitHub URL: https://github.com/kuthalam/Data_Science_Capstone/blob/main/Capstone_Webscraping.ipynb



Data Wrangling

- Processed data was from the API (i.e., filtered to Falcon 9)
- Straightforward: Missing payload mass values were filled with mean
- About 30% of landing pads were missing
- Outcomes were used to create a target value column Class
  - 0 – Bad outcome
  - 1 – Good outcome
  - Note failure to land may be because there was no attempt
  - 66.7% success rate
- Categorial variables were one-hot encoded
- GitHub URL: https://github.com/kuthalam/Data_Science_Capstone/blob/main/Capstone_Data-Wrangling.ipynb



EDA with Data Visualization

- Primary goal with EDA: What features of the Falcon 9 missions correlated with each other and landing success?
- Scatter Plots for Q1: Flight number vs. payload mass, orbit type vs. flight number, orbit vs. payload, launch site vs. payload
- Other plots: Success by launch site, success average by orbit type, success average per year
- GitHub URL: https://github.com/kuthalam/Data_Science_Capstone/blob/main/Capstone_DataViz-EDA.ipynb

EDA with SQL

- Primary goal: With all the launch data, what sort of numerical analysis can we do when we start a deep dive and what useful numbers can we quickly pull without that detailed look?
- Queries just to get used to the data (e.g., unique launch sites, names of boosters in a payload range)
- Useful facts (e.g., number of successful and failed flights, landing outcomes between 2010 and 2017)
- GitHub URL: https://github.com/kuthalam/Data_Science_Capstone/blob/main/Capstone_SQL-EDA.ipynb

Build an Interactive Map with Folium

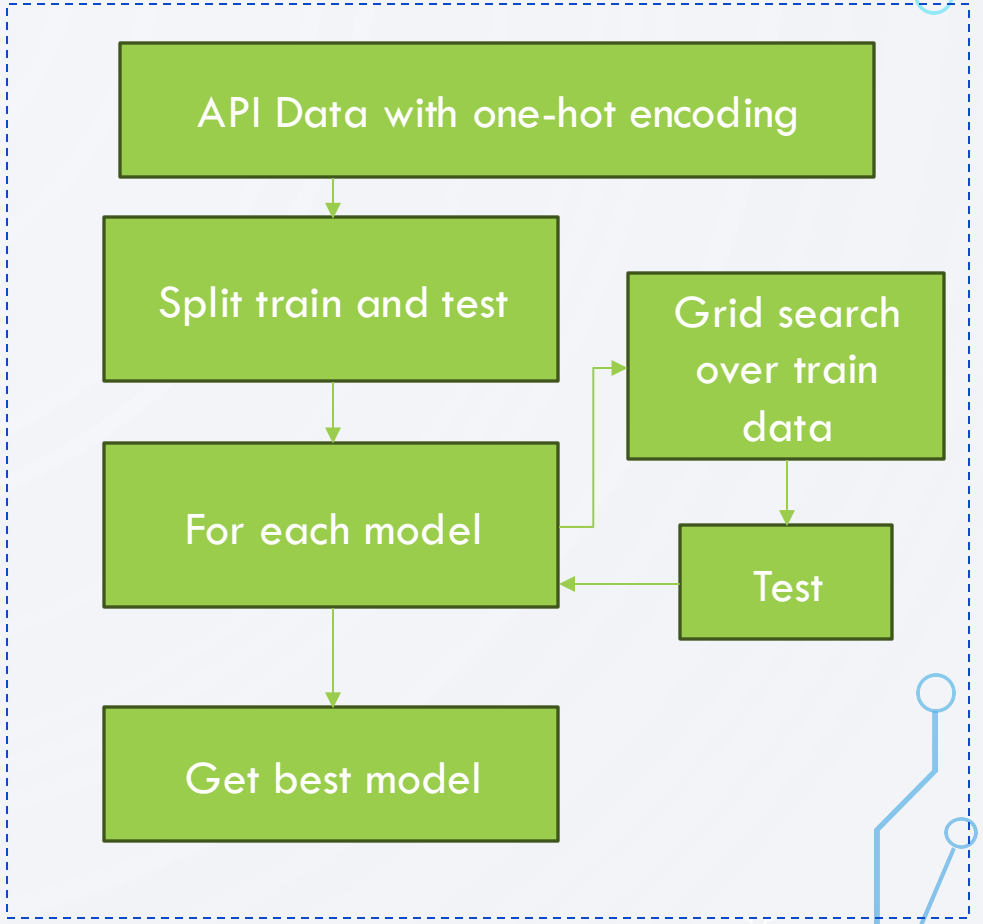
- Primary goal: “what about the successful missions can be gleaned from the geography of the launch sites?”
- Focused on marking sites of successful and failed launches, where the flight sites are, and their distances from notable landmarks
  - Where should SpaceY keep its launch sites?
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- GitHub URL: https://github.com/kuthalam/Data_Science_Capstone/blob/main/Capstone_Folium.ipynb

Build a Dashboard with Plotly Dash

- Primary goal: “What feature values map to the most successes?”
- Pie chart of success vs. failure percentage across launch site
  - Usefulness of landing site as a feature
- Scatter plot of payload mass vs. landing outcome with a hue according to booster used
  - Payload range for successful missions
- GitHub URL: https://github.com/kuthalam/Data_Science_Capstone/blob/main/Capstone_dash_app.py

Predictive Analysis (Classification)

- Features: 'FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial'
  - One-hot encoded: Orbit, LaunchSite, LandingPad, and Serial
- Test set size of 20% with a set random seed
- Classifiers: Logistic Regression, SVM, Decision Tree, and KNN
- Best training and test accuracy: ~87.3% and ~83.3%
- GitHub URL: https://github.com/kuthalam/Data_Science_Capstone/blob/main/Capstone_ML_Prediction.ipynb



Results

EDA insights:

- SpaceX got excellent at successful launches with a jump in 2017
- No one feature is particularly predictive

From the interactive analytics:

- The CCAFS launch site, a payload range of 2000 – 3700 kg, and the FT boosters correlated most with a successful launch
- Landing sites stayed near roads and railways but away from cities

Predictive analytics results:

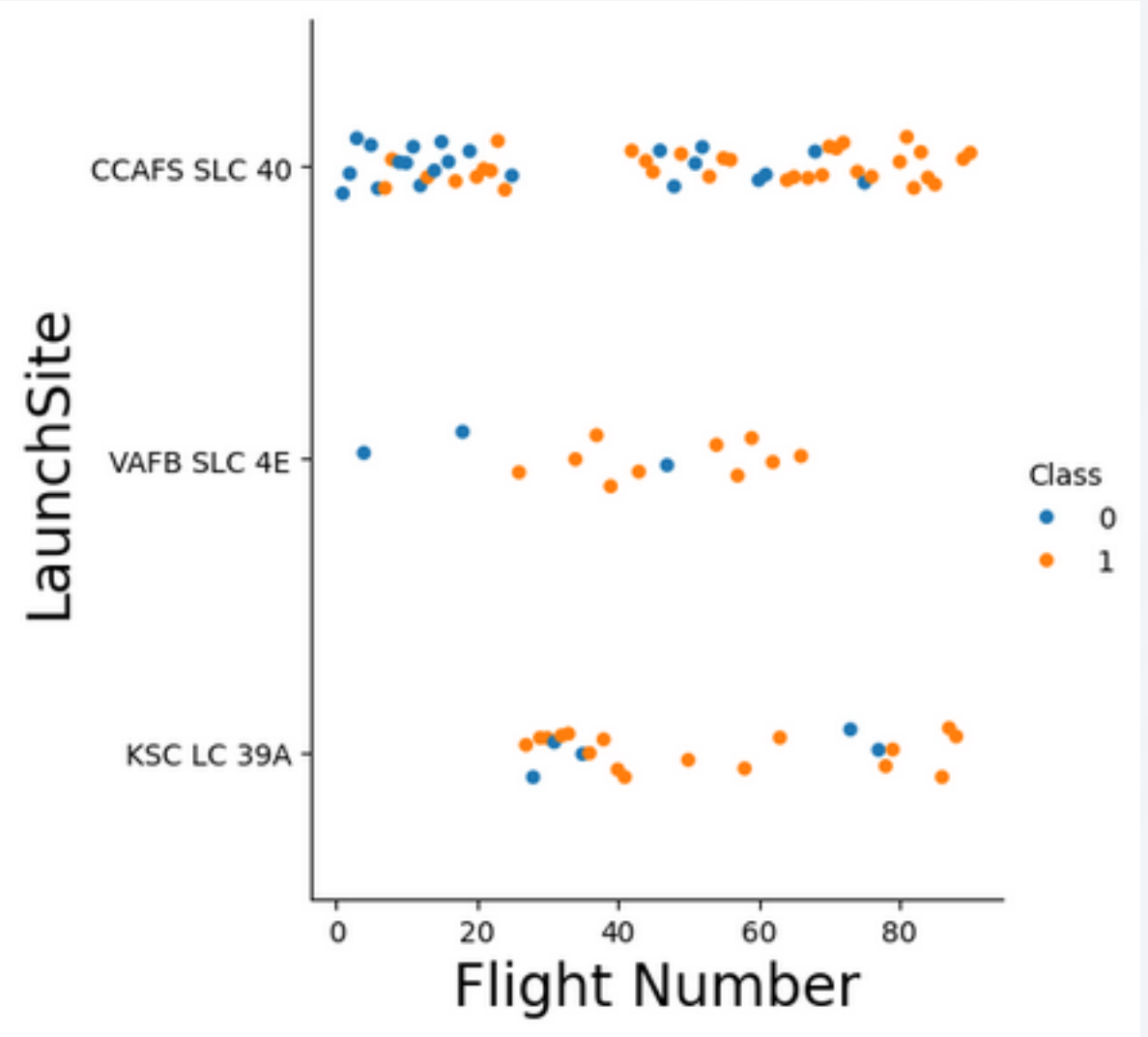
- No one model proved superior as all had a test accuracy of 83.3% though decision trees (DT) had the best training accuracy

Section 2

Insights drawn from EDA

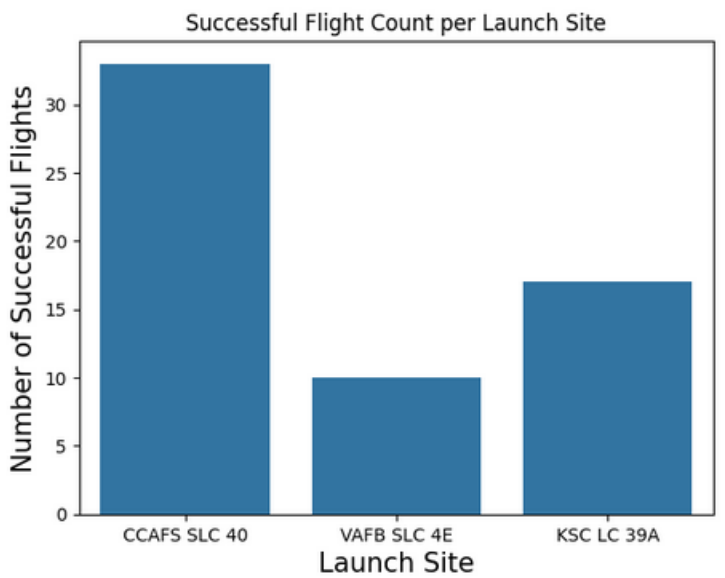
Flight Number vs. Launch Site

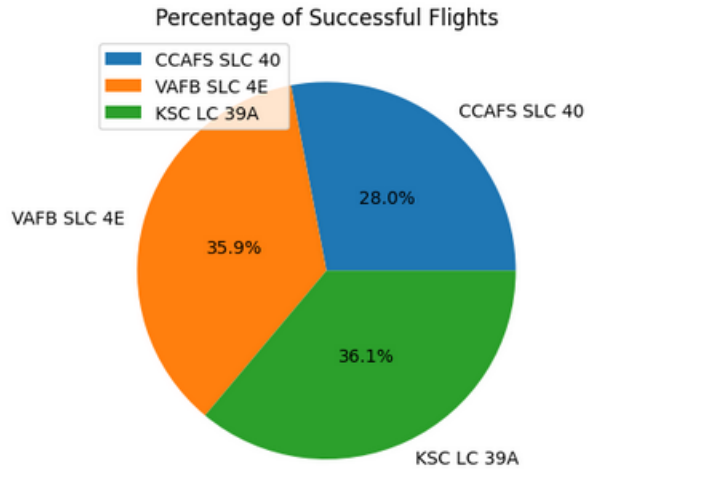
- Lots of flights in Cape Canaveral (CCAFS) along with majority of successes
- Note many successes in VAFB but not as many missions
- Later flight number: more likely to succeed
  - Recall conclusion 1 on how SpaceX got good (slide 5)



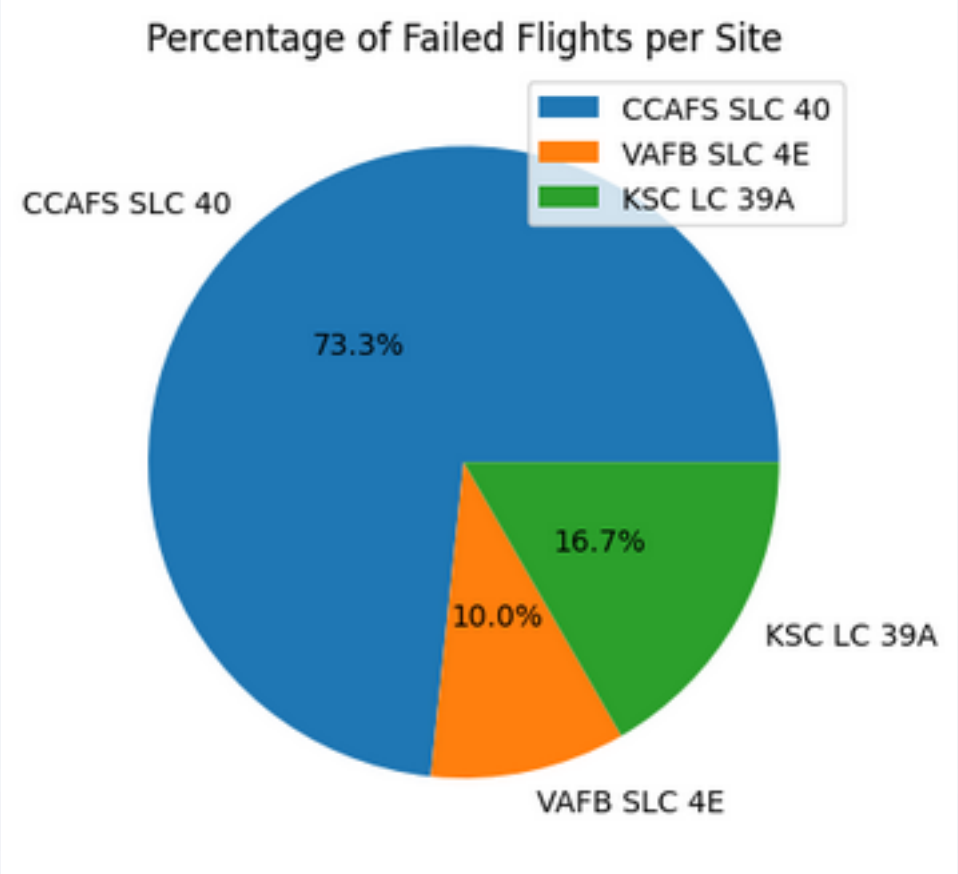
More on how good each launch site was

- More successes in the Floridian locations (CCAFS and KSC) but percentages were about equal
- Lot of failures in CCAFS so hard to say that launch site is a useful predictive feature



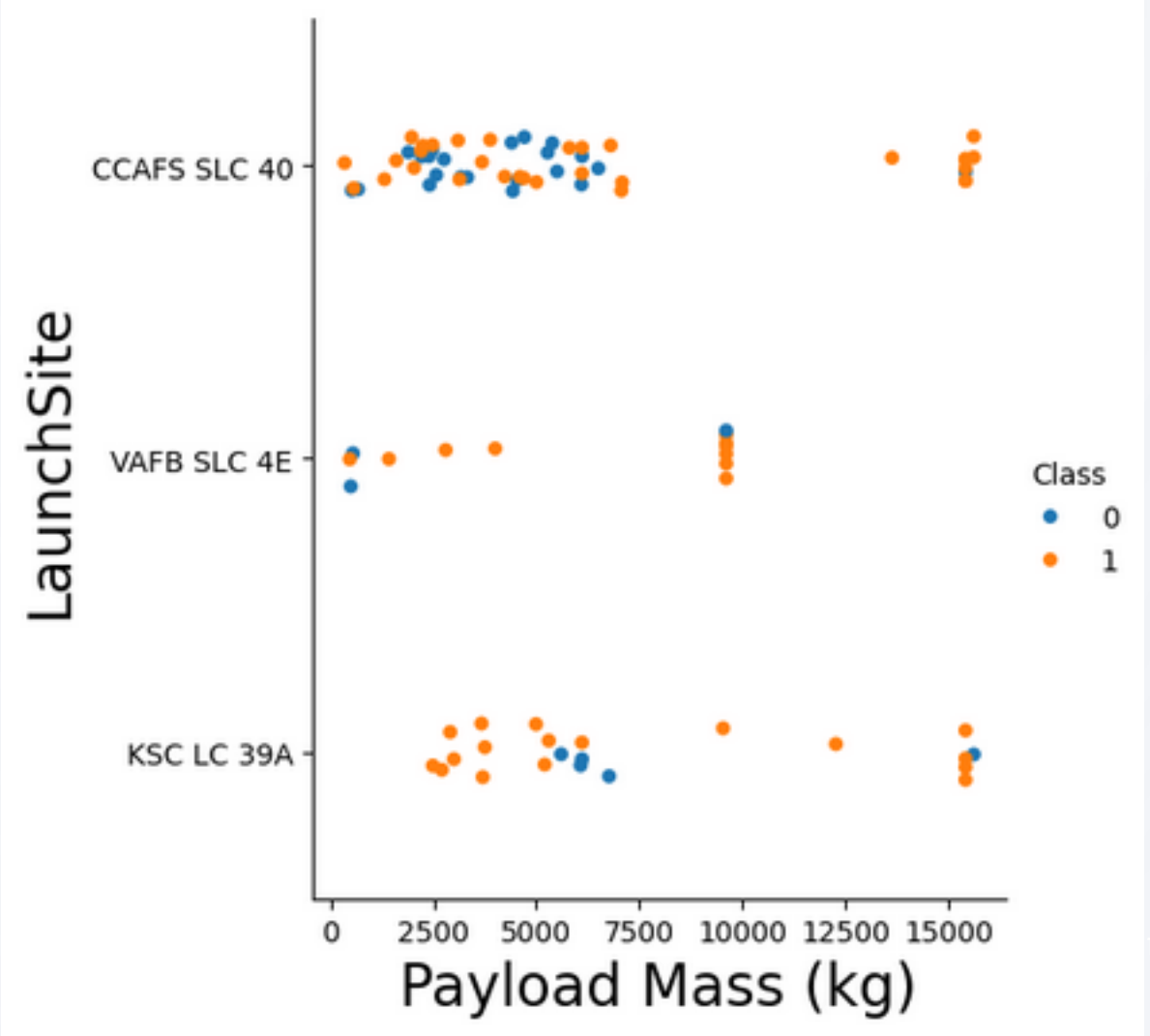


That slide I promised before



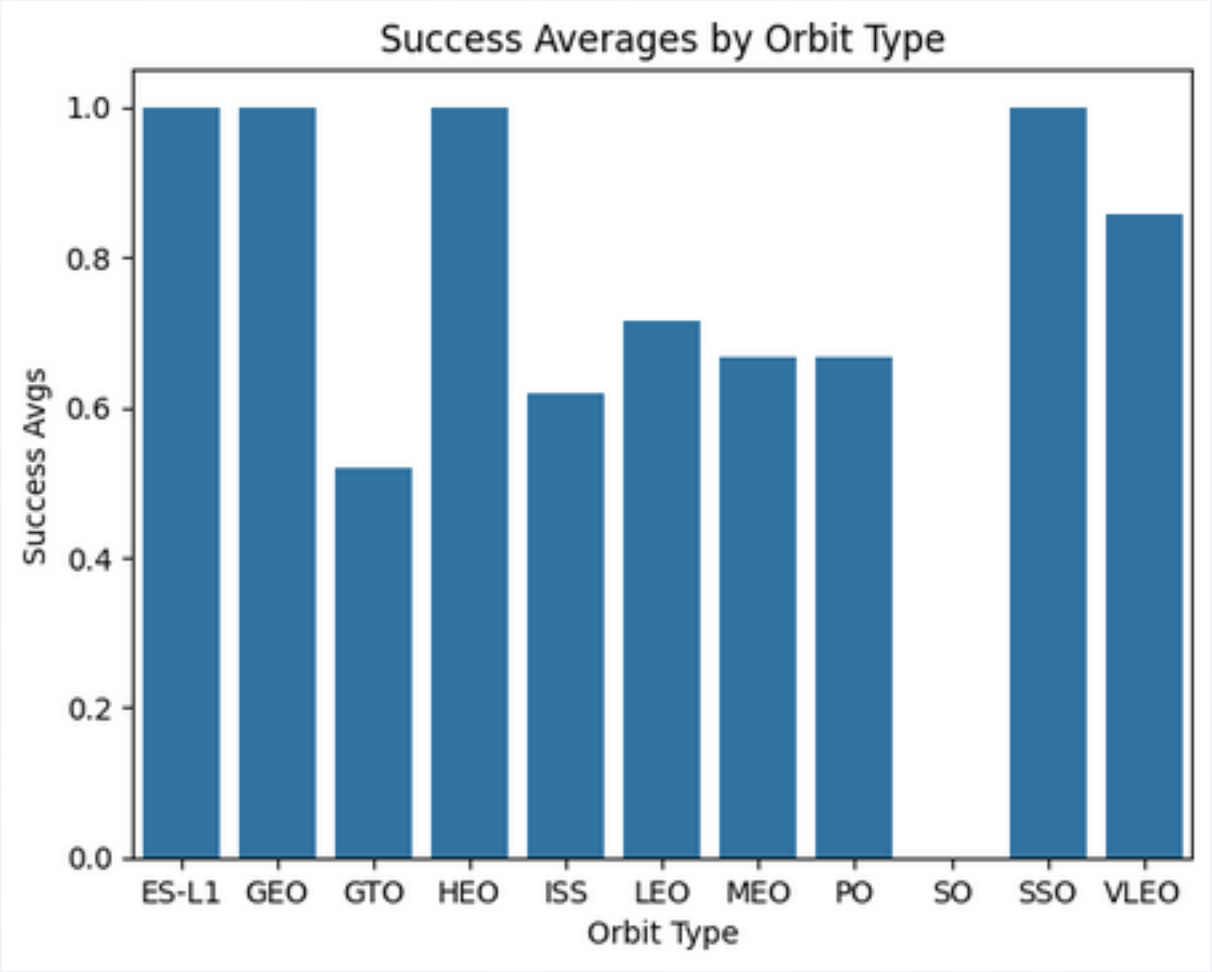
Payload vs. Launch Site

- Past about 9000 kg of payload, landings were almost always successful
- VAFB is mostly tested for 9-10k kg payloads
- Follow-up: is there a correlation between payload mass and resources put into the mission
  - High payload => landing that is more set up for success?



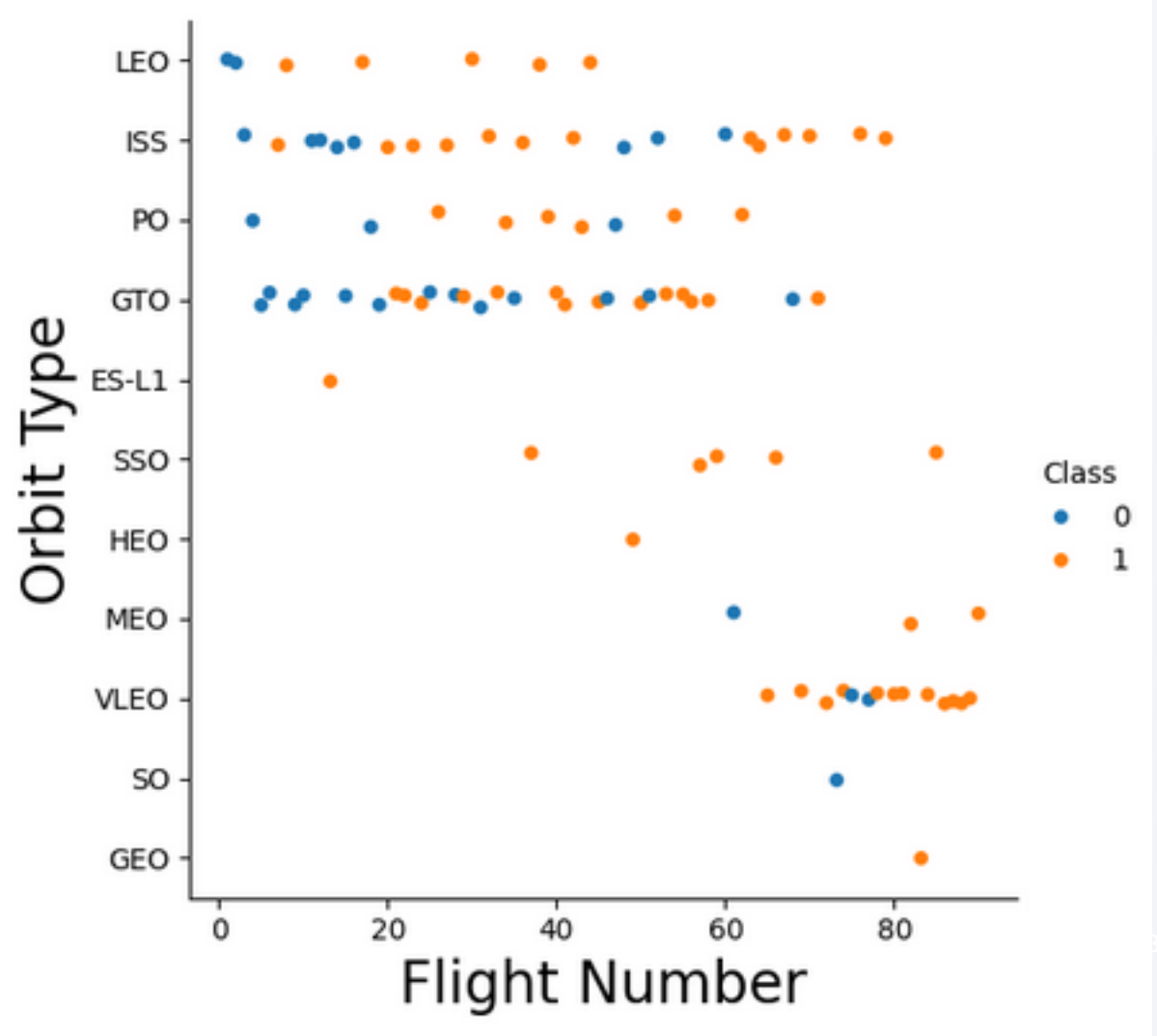
Success Rate vs. Orbit Type

- 4 orbits look to be the most successful: ES-L1, GEO, HEO, and SSO
- Recall our exploration of percentages vs. absolute number of success
- Are these orbits tried and tested?



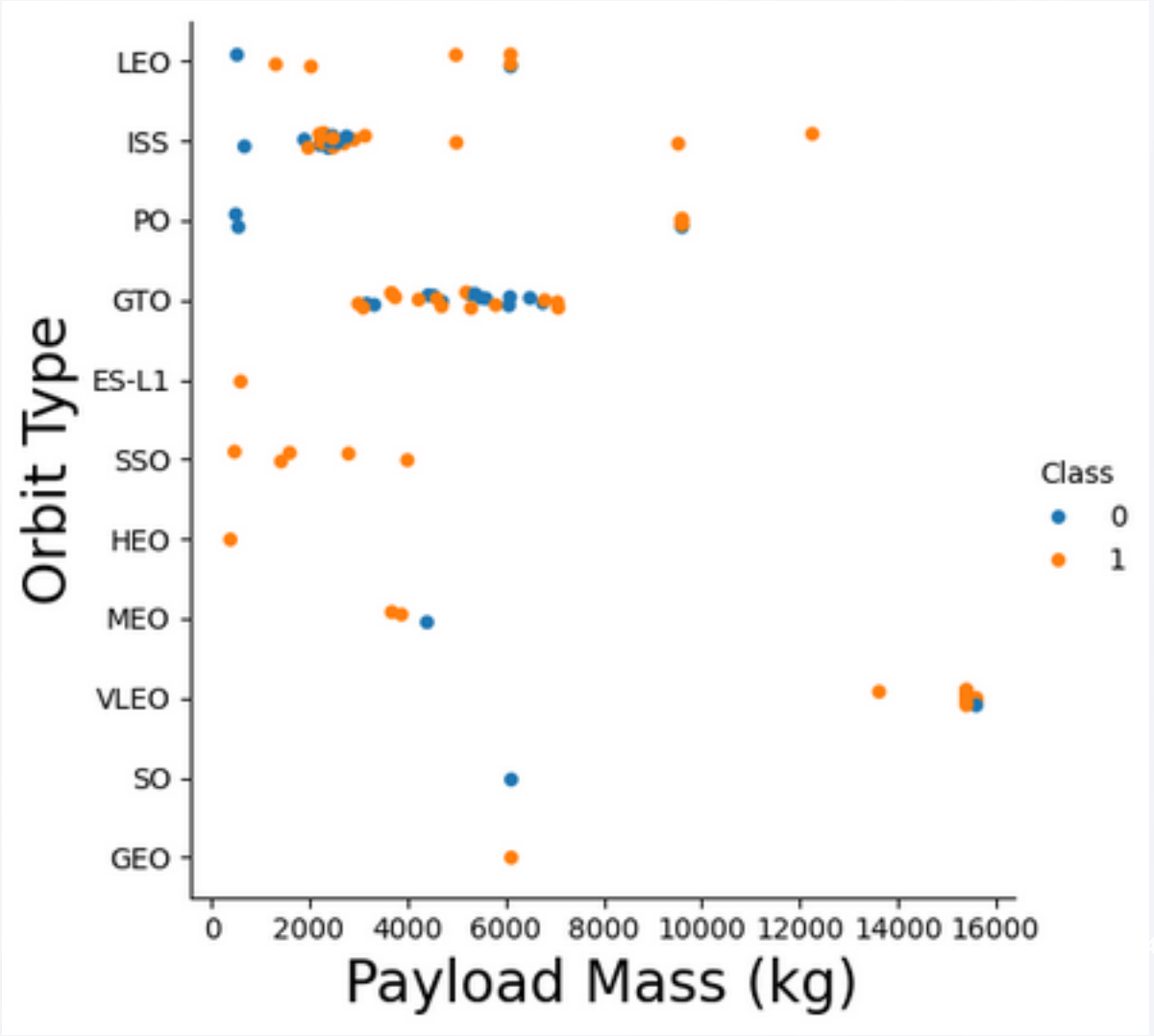
Flight Number vs. Orbit Type

- Now let's look at ES-L1, GEO, SSO, and HEO
- None have more than 5 missions
- VLEO, ISS, and GTO look more promising



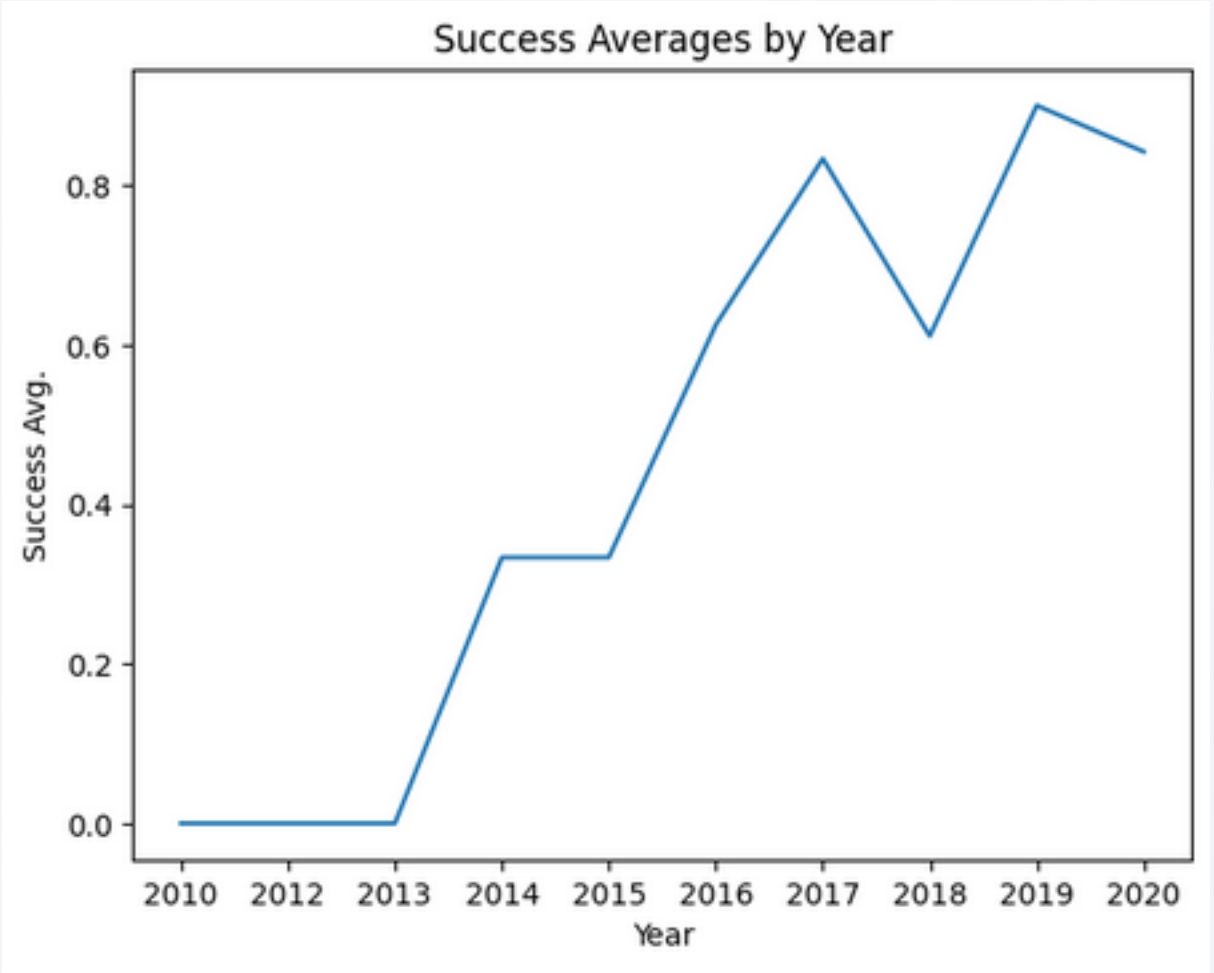
Payload vs. Orbit Type

- Adding payload adds interesting constraints
- Most ISS successes were in the high 2k to mid 3k kg payloads
- GTO is 4k-8k kg
- VLEO seems to almost always have high payloads



Launch Success Yearly Trend

- Really cements conclusion 1: SpaceX got very good at launches over time with 2017 onwards being generally great
- Saw this with flight number vs. launch site and more numbers in the SQL analysis



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

- Speaking of proof of conclusion 1 with numbers, consider this SQL query:
  - SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Date BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC

Landing_Outcome	COUNT(Landing_Outcome)
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

Now see what happens after 2017-03-20

Landing_Outcome	COUNT(Landing_Outcome)
Success	38
No attempt	11
Success (drone ship)	9
Success (ground pad)	6
Failure	3
Controlled (ocean)	2
No attempt	1

Looks quite the increase in successes, would you not say?

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(*) FROM SPACEXTABLE WHERE Landing_Outcome LIKE "Success%"
```

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

COUNT(*)
61

```
%sql SELECT COUNT(*) FROM SPACEXTABLE WHERE Landing_Outcome LIKE "Failure%"
```

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

COUNT(*)
10

- Remaining landings were classified as controlled or uncontrolled, not attempted, or precluded (not sure what the latter means)

Remaining Slides are For Submission Requirements

All Launch Site Names

- Query: SELECT DISTINCT Launch_Site FROM SPACEXTABLE

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

- LC-40 vs. SLC-40: Naming difference that splits pre-Falcon 9 era from the Falcon 9 era [1]

Launch Site Names Begin with 'CCA'

- Query: SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5

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

- See how Landing Outcome gives us detailed info including whether a landing was attempted

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer="NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Looks like NASA is a frequent customer as one may expect

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
[13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version="F9 v1.1"
 * sqlite:///my_data1.db
Done.
[13]: AVG(PAYLOAD_MASS__KG_)
      2928.4
```

- Looks like the F9 v1.1 boosters tended to receive lightweight missions
- Either smaller payloads or lighter (read: more delicate) payloads

First Successful Ground Landing Date

```
[17]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome LIKE "Success%"
 * sqlite:///my_data1.db
Done.
[17]: MIN(Date)
      2015-12-22
```

- SpaceX was established in 2002 and its first launch was in 2008 so this is not the first ever successful launch to be clear [2]

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[15]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome="Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_

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

[15]:

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

- Full query: SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome="Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

Boosters That Carried Maximum Payload

```
[19]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
 * sqlite:///my_data1.db
Done.
```

[19]:

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

Quite a few boosters have been tried and tested

2015 Launch Records

```
[20]: %sql SELECT substr(Date, 6,2), Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = "Failure (drone ship)" AND substr(Date

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

[20]:

substr(Date, 6,2)	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- Full query: SELECT substr(Date, 6,2), Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = "Failure (drone ship)" AND substr(Date,0,5)='2015'

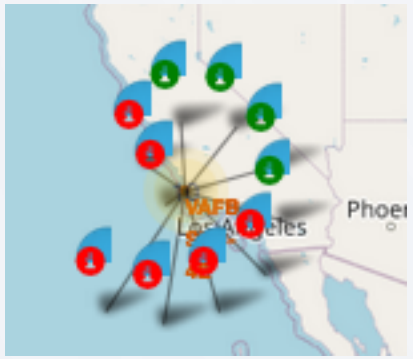
Section 3

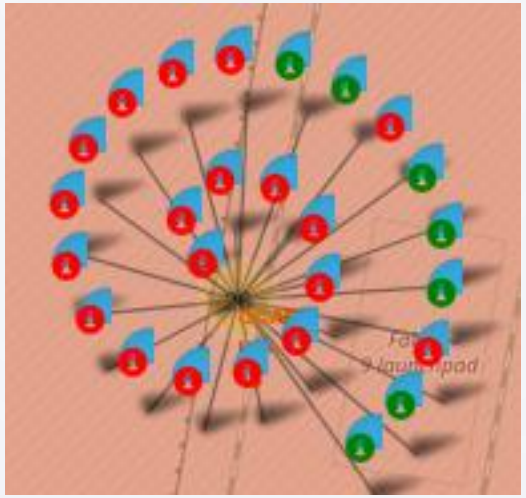
Launch Sites Proximities Analysis

Zoomed out view of all launch sites

- Note how the sites are all coastal, particularly on both sides
- Consistent with previous visuals, there are more sites and likely more data for the East coast

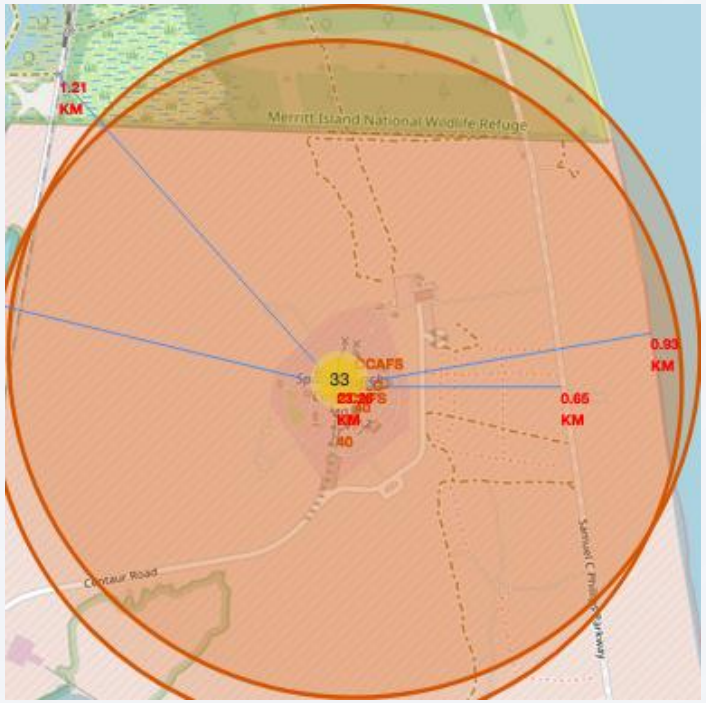
Successful and failed landing launch points





- A few screenshots of points of success and failure
- Notice the middle picture is from the East coast where we had already seen had a lot of failed landings

What is Near a Launch Site?

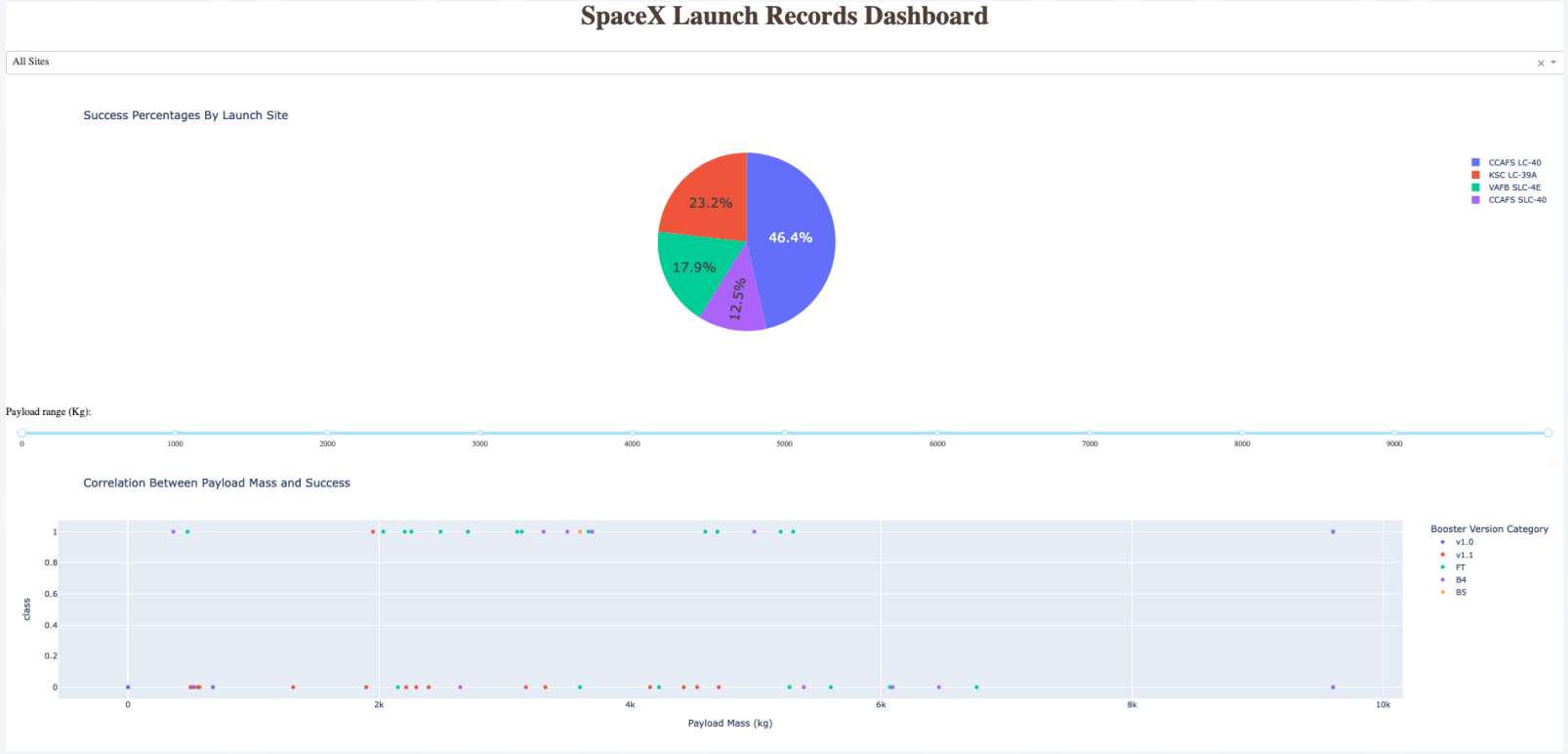


- Notice how the CCAFS launch point is extremely close to the coast, highway, and railway but is 23 km away from the nearest city
- Consistent with conclusion 2 that launch sites are close to coasts, roads, and railways but far from cities (slide 5)

Section 4

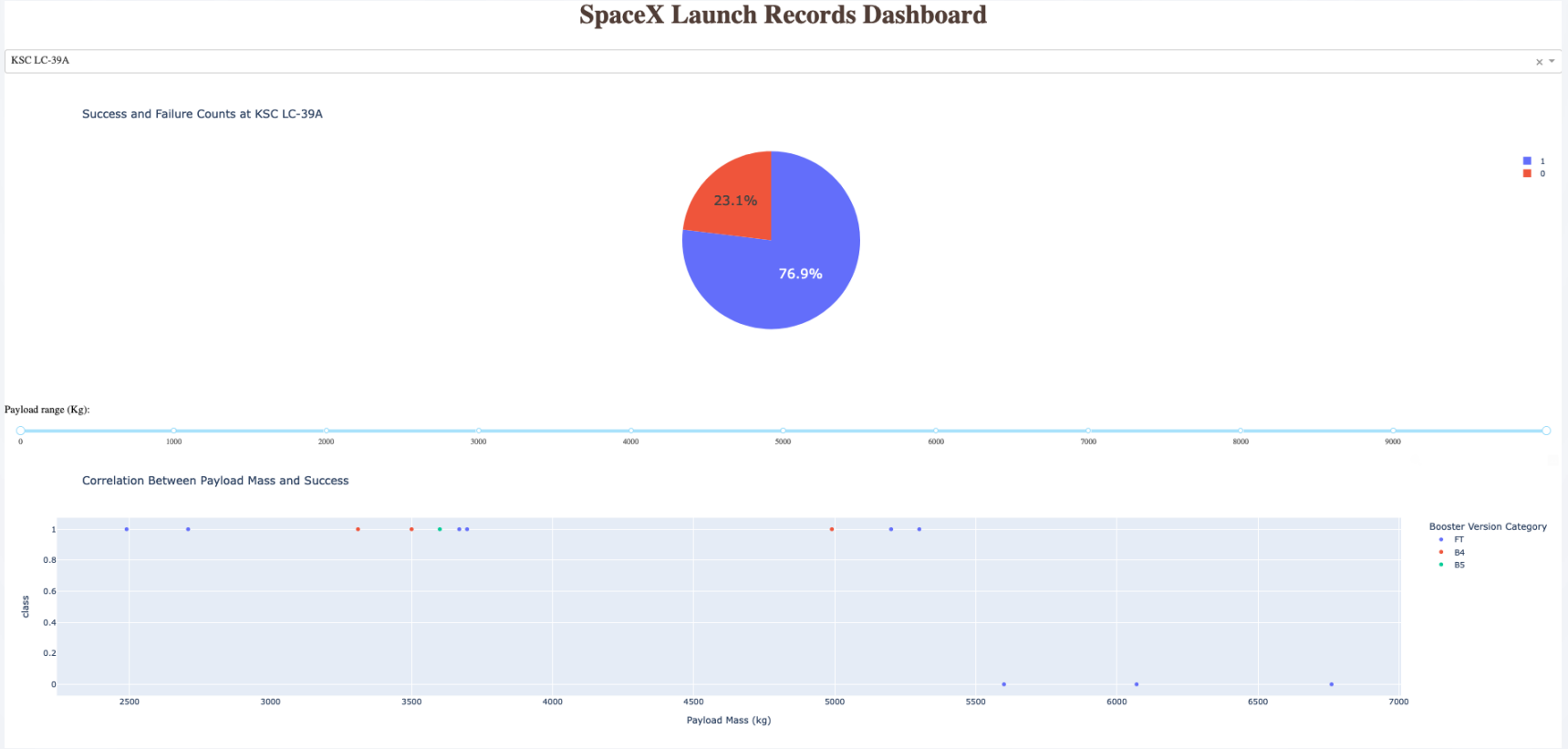
Build a Dashboard with Plotly Dash

Dashboard for all launch sites



- The CCAFS sites combined for 58.9% of successes
- Consistent with conclusion 3: CCAFS sites, payloads between 2k and ~3.7k kg, and the FT booster had most successful flights (green dot => FT booster)

Site with highest success rate: KSC LC-39A



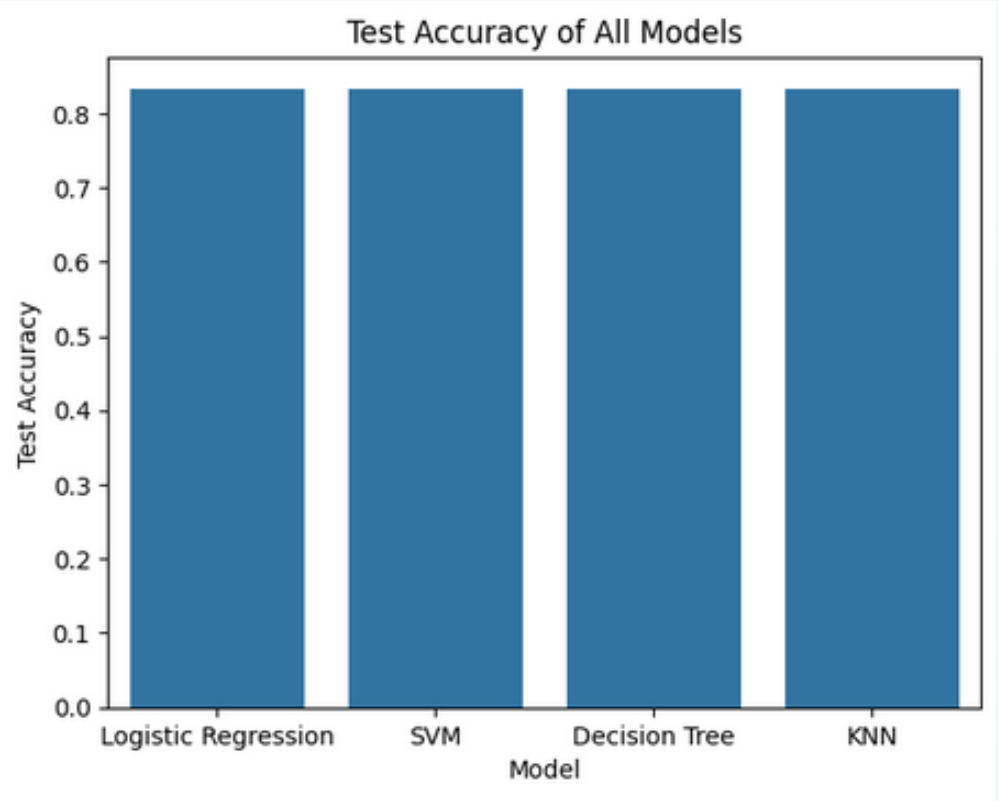
- Highest success rates among all the sites, but again: not the most launches!

Screenshots of the payload chart close up

Section 5

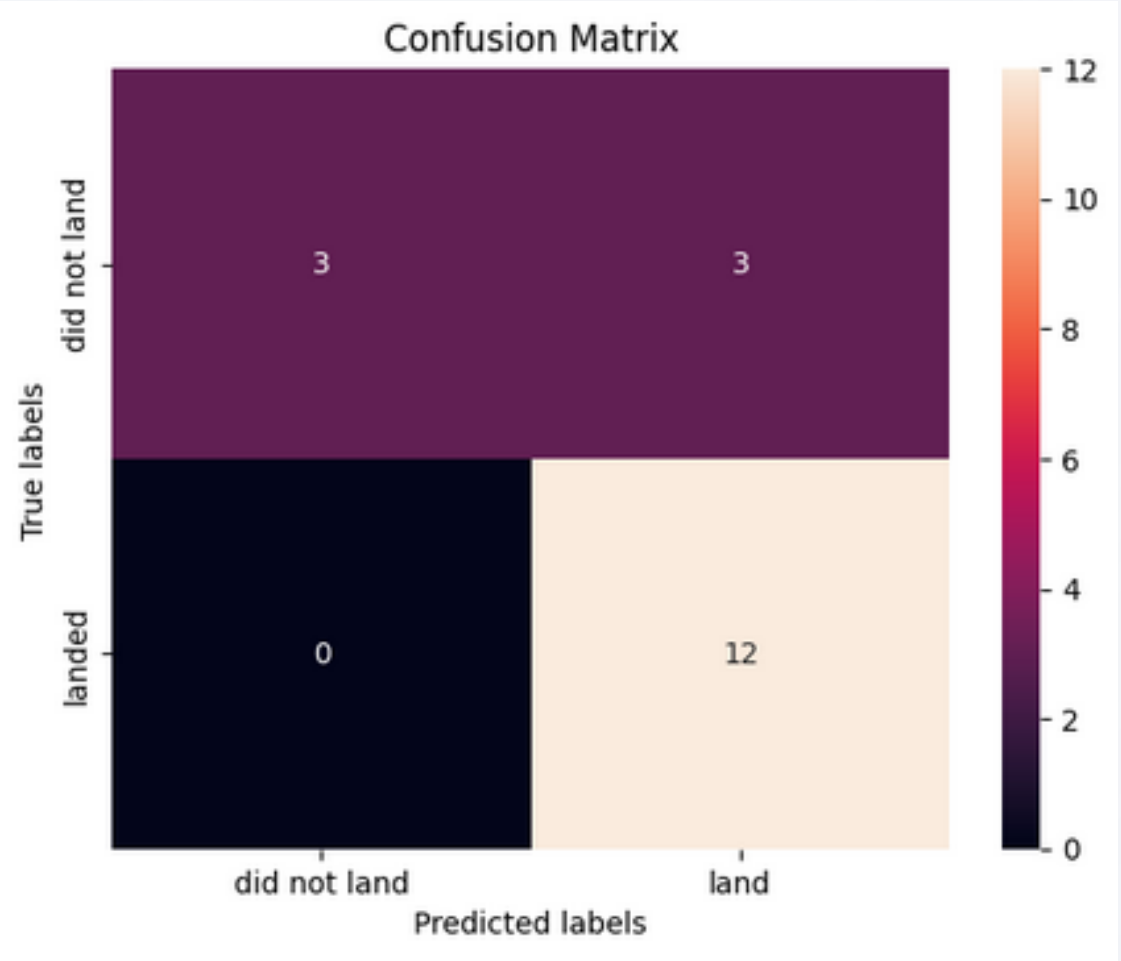
Predictive Analysis (Classification)

Classification Accuracy



- No one model is better than another
  - All are at about 83.3% test accuracy

Confusion Matrix



- All the models performed

Conclusions

- C1: Landings were more likely to be successful as time went on, regardless of payload mass or orbit types (53 successes from March 2017 onwards, compared to 8 pre-March 2017)
- C2: Launch sites always kept away from cities but stayed close to coasts, roads, and railways.
- C3: The CCAFS LC-40 site, payload masses in the range of about 2000 kg to 3700 kg, and the FT booster are the most “proven” components of a successful landing (i.e., correlate to more successes).
- C4: There is no clear winner in which machine learning model yields the best prediction but all models yield a test accuracy of about 83.3%

Thank you!

References

- [1] https://ccspacemuseum.org/facilities/launch-complex-40/
- [2] https://www.britannica.com/topic/SpaceX