Winning Space Race with Data Science

Elizabeth Medlin

April 12, 2024

Outline

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

Executive Summary

- This Presentation will identify the predictive factors of a successful Falcon 9 launch.
- The first section will describe the methodology used to collect and clean the relevant data, from API scraping to choosing the proper algorithms to predict launch outcomes.
- The second section will display exploratory visualizations, providing context for the collected data.
- The final section will relate the best algorithm for predictive analysis, and its assessment of the most important factors in First Stage recovery.

Introduction

- Since the mid-20th century scientists, public organizations, and business professionals have sought to increase the economic viability of commercial space travel.
- Major innovations have occurred since the Space Shuttle days; the Falcon 9 rocket system from SpaceX purports to re-use the First Stage of launching system.
- According to SpaceX, First Stage reuse decreases the launch cost from $165 million to $62 million—a 62% drop in cost.
- To realize these tremendous cost savings, a competitor to SpaceX must identify the most important factors leading to a successful launch—including recovery of the First Stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web Scraping
  - Open Source Database Collection
- Data wrangling
  - Pandas Dataframe Processing
- Visualization and SQL:
- Advanced Visualization with Folium and Plotly Dash:
- Predictive analysis using classification models
  - How to build, tune, evaluate classification models

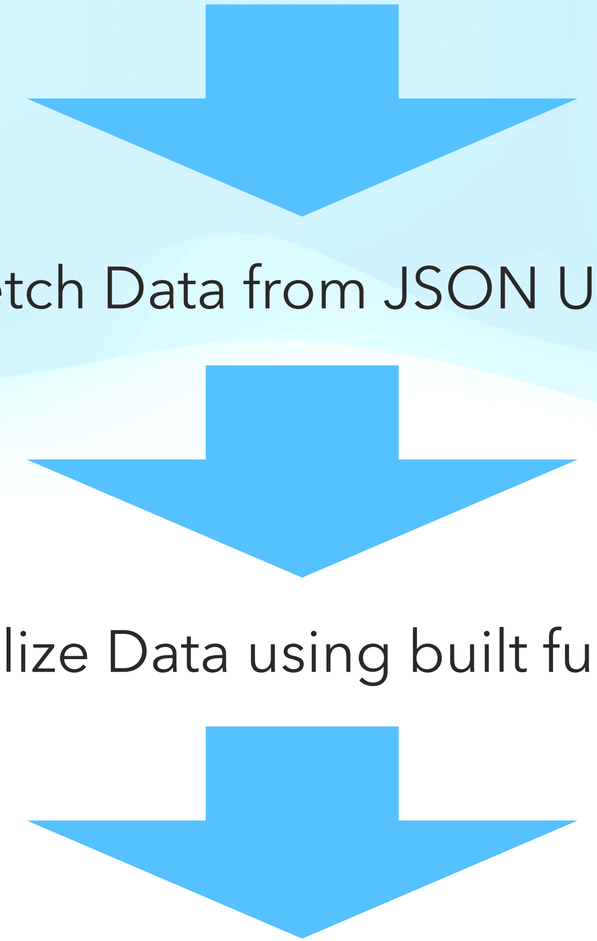
Data Collection

- Data was processed using Web Scraping and API for SpaceX datasets.
- The API data came through relatively clean, and was processed into a Pandas dataframe with the edition of a binary Launch Outcome column.
- Web-scraped data from Wikipedia required Beautiful Soup and an HTML parser, as well as several built functions, to be normalized and converted to dataframe form.

Data Collection – SpaceX API

- I defined the functions to call data cores from the SpaceX URL. With the data imported, I normalized it and converted it to a pandas dataframe.

- https://github.com/felonious-twunk/IBM-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- After importing the data from a static URL, I converted it to BeautifulSoup using an HTML parser. From there, the data was sent to a dataframe for visualization and machine learning.

- https://github.com/felonious-twunk/IBM-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Null values were replaced with category mean. Then, I extracted columns and data type with the purpose of converting launch outcomes to a binary integer form.
- https://github.com/felonious-twunk/IBM-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Replace Null Values with category mean

Get column data types and value counts

Create binary Outcome column for ML models

Yeehaw

EDA with Data Visualization

- Scatterplots plotted several variables against each other:
  - Flight # and Launch Site
  - Payload and Launch Site
  - Flight # and Orbit Type
  - Payload and Orbit Type

- A bar plot measured the success rate within each orbit type.

- A line plot marked the improving success rate year over year.

- https://github.com/felonious-twunk/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

The following SQL queries were performed:

- Names of unique launch sites
- Launch Sites beginning with CCA
- Total Mass Carried by NASA-launched Boosters
- Average Payload Mass of Falcon 9 v.1.1
- First successful ground pad landing
- Successful drone ship boosters with payload >4000 and <6000
- Total successful and failed mission outcomes
- Booster versions that have carried maximum payload mass
- Selected columns of all launches in 2015
- Ranked count of landing outcomes in 2016 and 2017

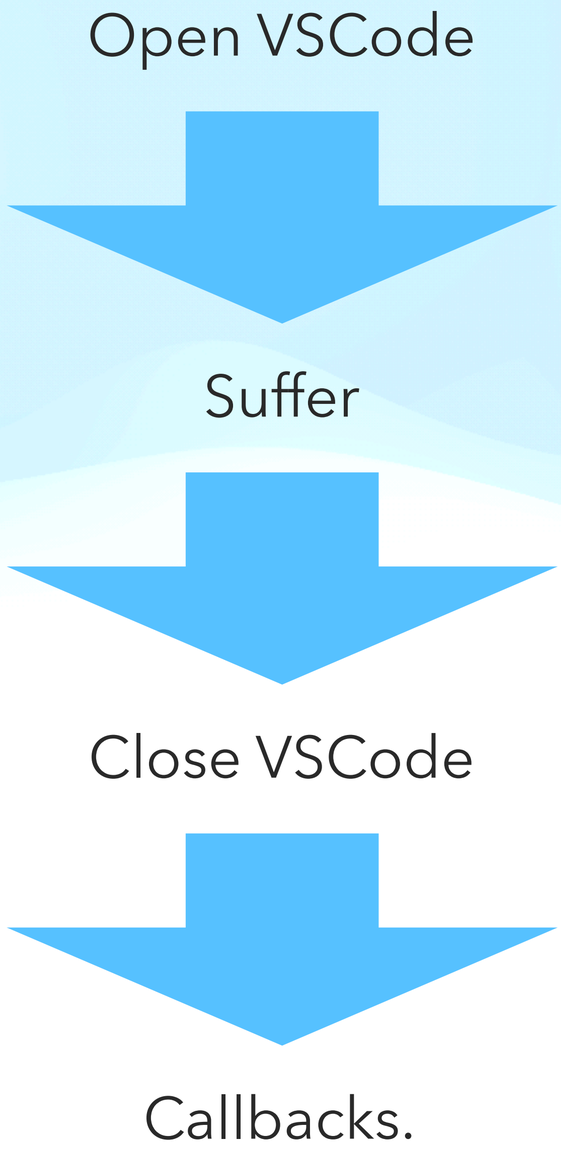
- https://github.com/felonious-twunk/IBM-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- A Folium map was used to mark launch sites and cluster launches that occurred at each site. Distance to the ocean and nearest city were included in the map.
- Distance between the launch sites and ocean is helpful for understanding the process of First Stage recovery. Most failed landings occur over the ocean to minimize risk to the surrounding population.
- https://github.com/felonious-twunk/IBM-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

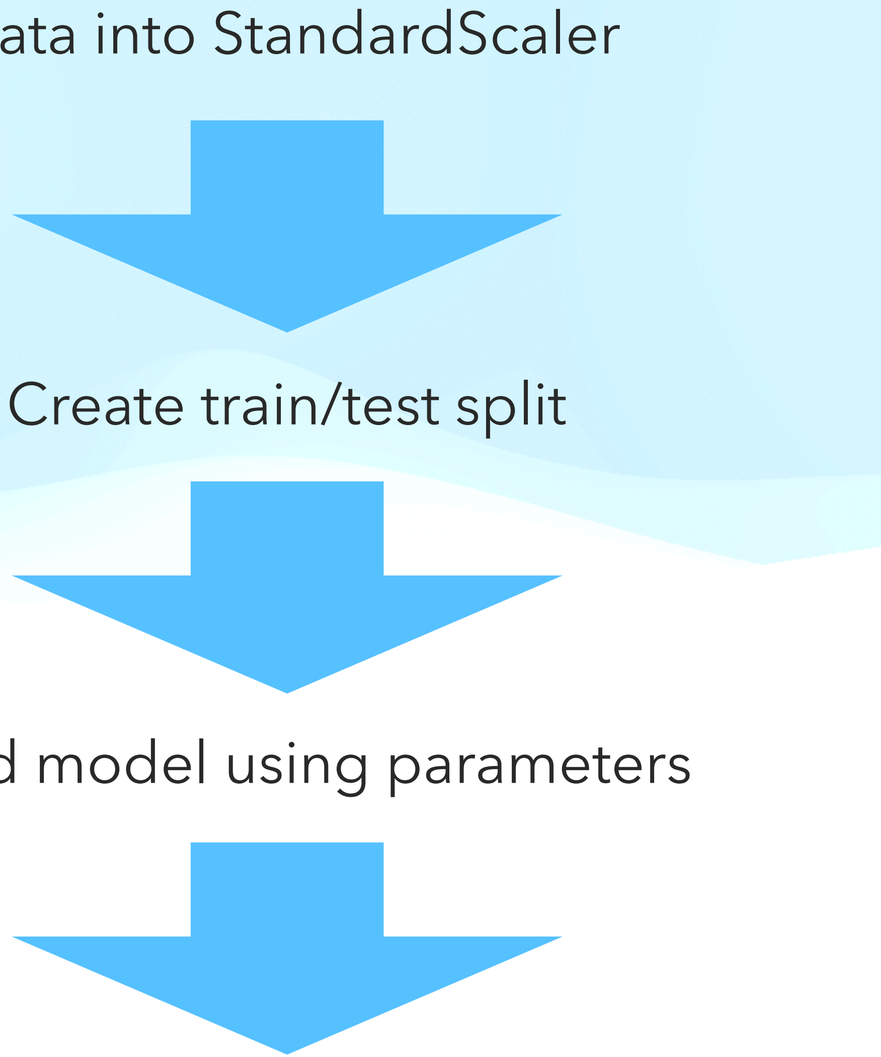
Build a Dashboard with Plotly Dash

- I hate dash!
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose



Predictive Analysis (Classification)

- I put the data through a standard scaler, then tested it on four algorithms: SVN, K-Nearest Neighbor, Decision Tree, and Logistic Regression.
- The models were calibrated with a train-test split of 0.2 and random count of 2.
- https://github.com/felonious-twunk/IBM-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

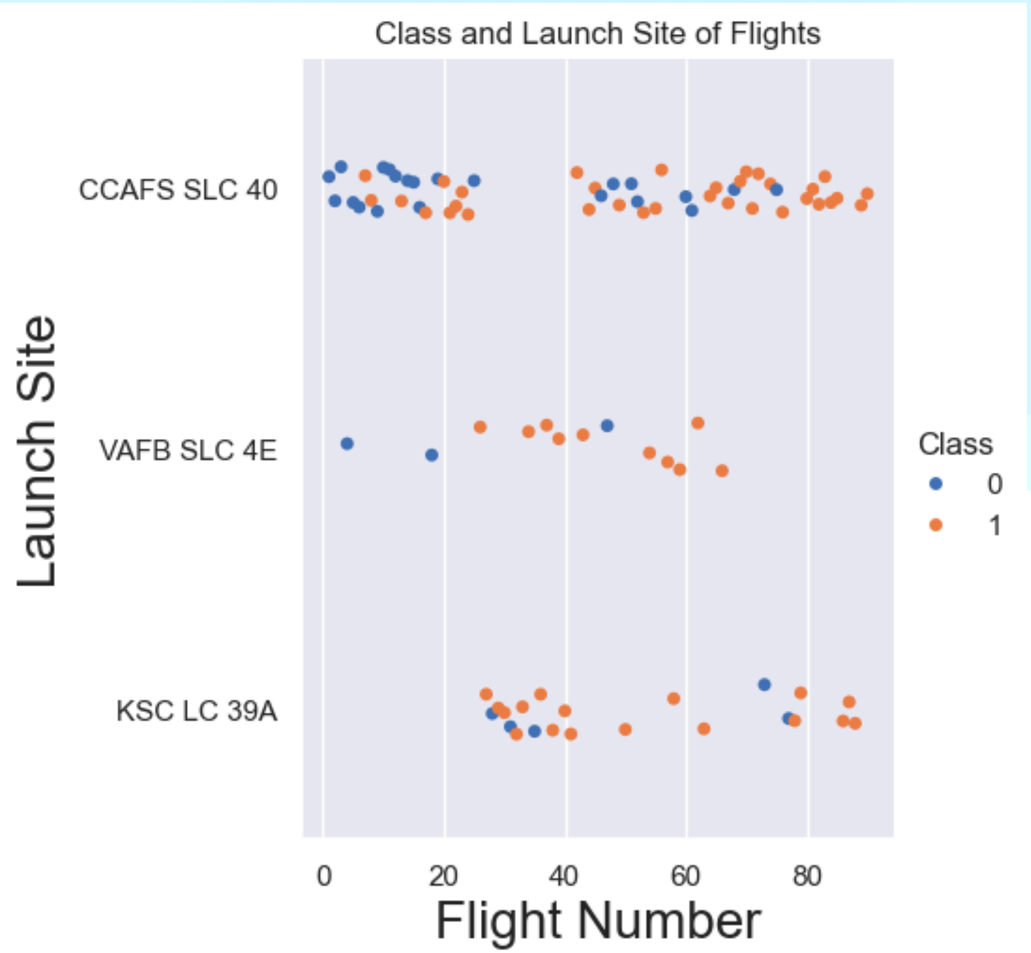
- Later launches have a higher rate of success, and larger booster rockets have facilitated an increase in average payload mass over this time period.
- The dash slides will include charts of the exploratory findings.
- A Decision Tree Algorithm has shown the highest accuracy in predicting the success or failure of a given launch.

Section 2

Insights drawn from EDA

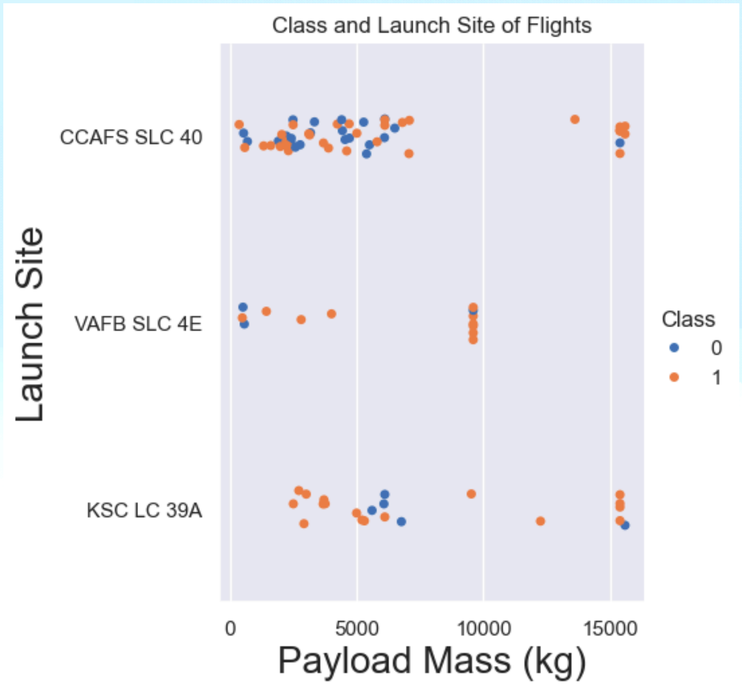
Flight Number vs. Launch Site

- There is no correlation between flight number and launch site
- The scatter distribution foreshadows a correlation between later launches and successful outcomes



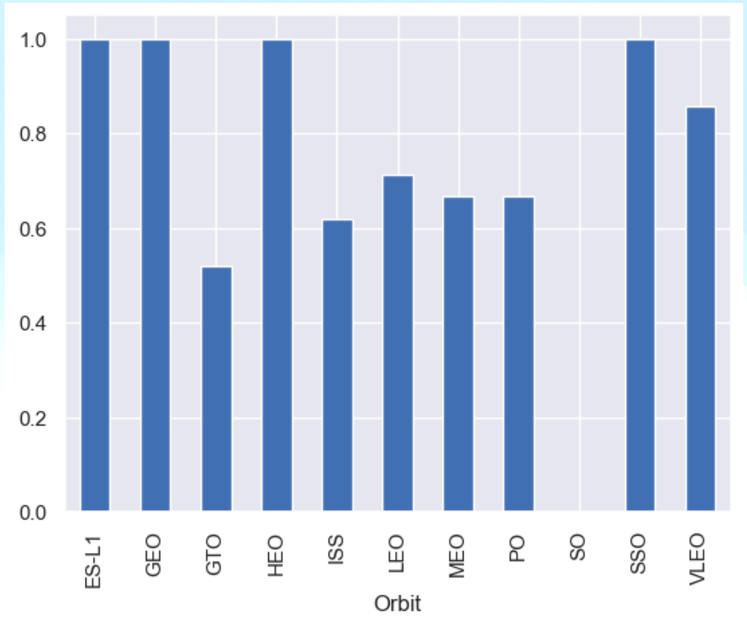
Payload vs. Launch Site

- Nearly all payloads around the 1000kg mark were launched from VAFB in Southern California, rather than Florida.
- These launches were predominantly successful.



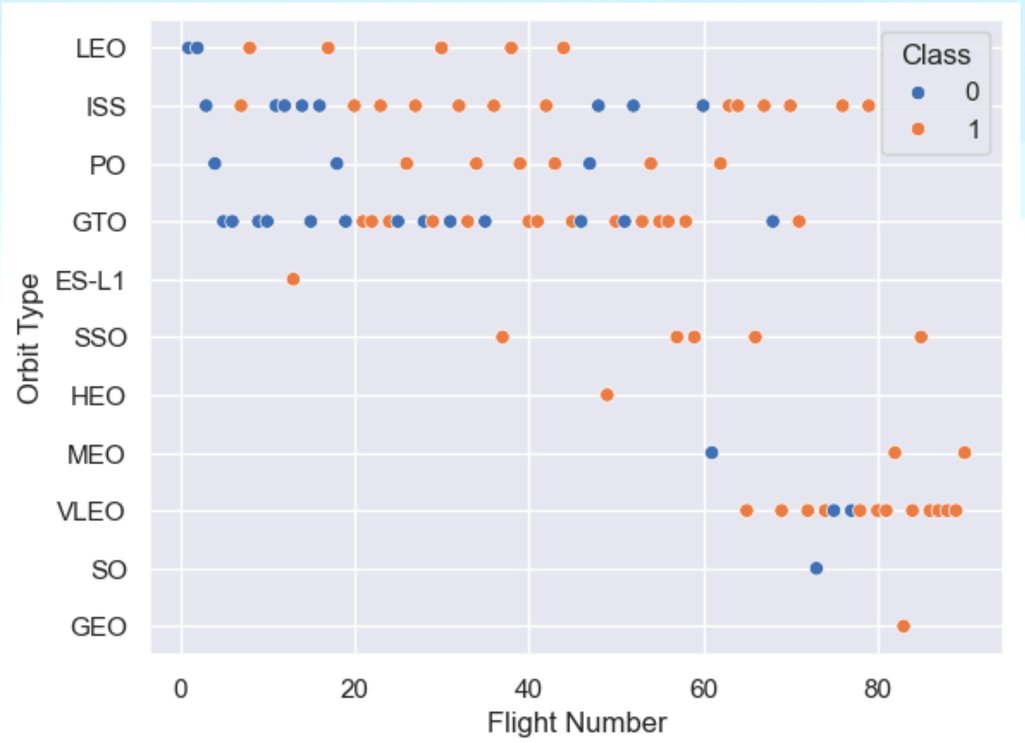
Success Rate vs. Orbit Type

- Among lower orbits, SSO has the highest success rate, followed by VLEO.
- In higher ranges, ES-L1, GEO, and HEO have a 100% success rate. GTO and the intermediate orbits have a lower success rate, though no explaining variable is immediately evident.



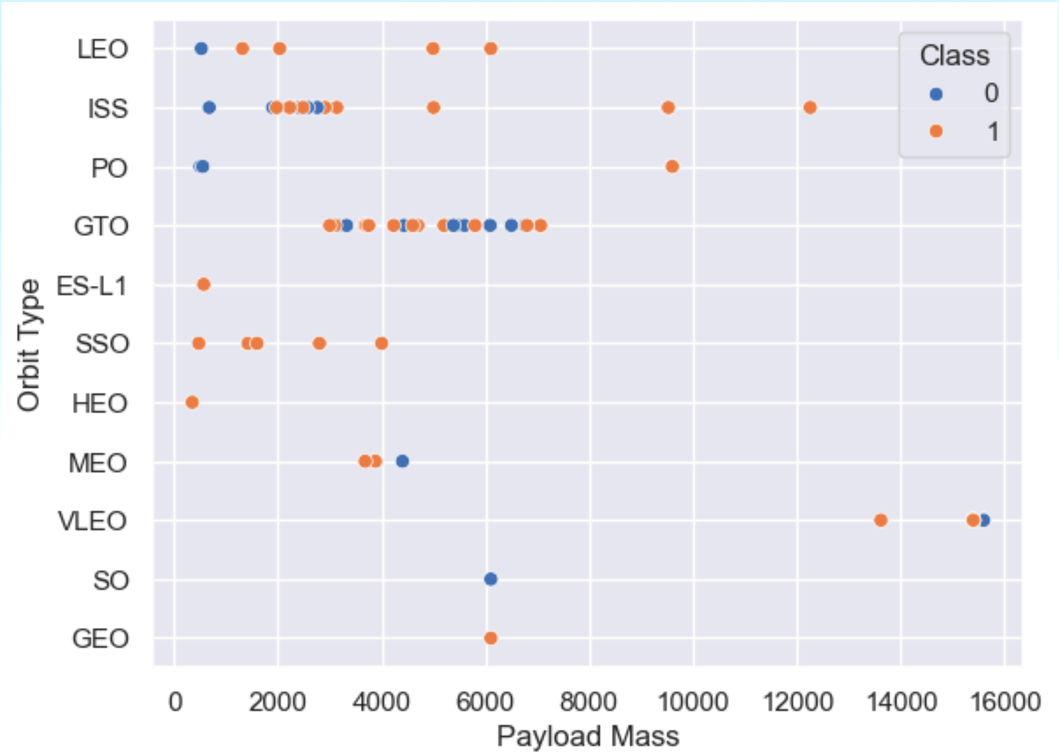
Flight Number vs. Orbit Type

- When referenced against the previous bar chart, low sample sizes in some orbits complicate the drawing of conclusions.
- HEO, ES-L1, GEO, and SO each have one launch
- The top 3 Orbit types (GTO, ISS, and VLEO) return more nuanced results, with VLEO showing the highest success rate of the three.



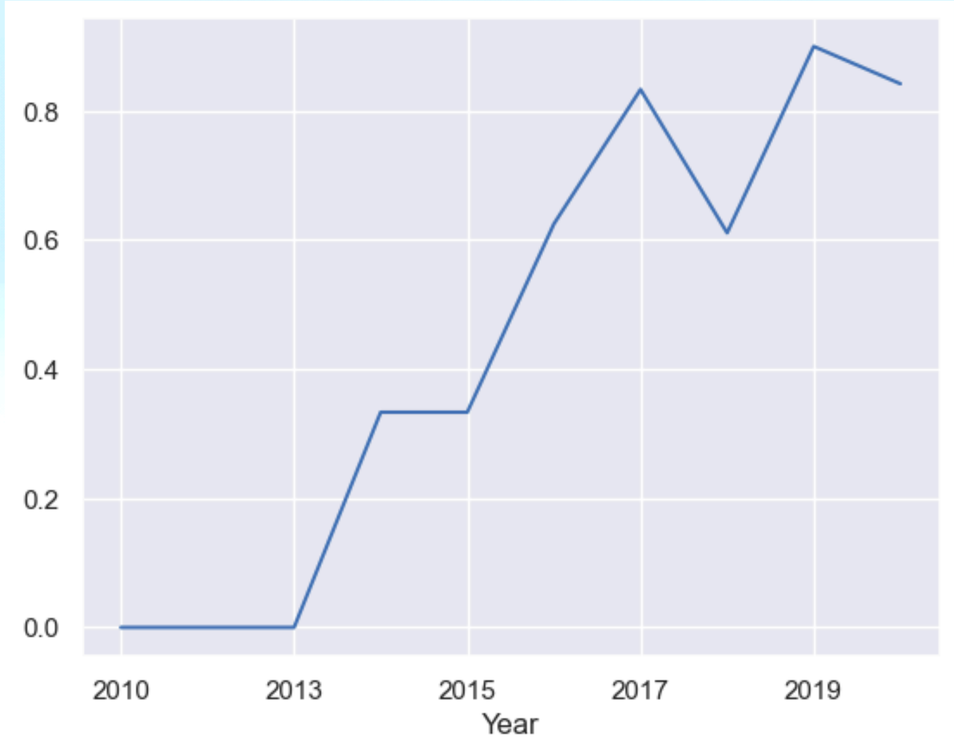
Payload vs. Orbit Type

- The ISS Orbit Flights show a cluster around 2000kg, likely due to routine calculations for International Space Station missions.
- VLEO supports the highest payloads, with a notable outlier to the much-higher ISS orbit at 12000 kg.



Launch Success Yearly Trend

- Between 2014 and 2020, engineers at SpaceX were able to learn from data and experience to markedly increase the success rate of its missions.
- 2018 saw a large uptick in number of flights, including the Spaceman Roadster stunt. This could explain the dip in launch outcomes, as some flights were primarily publicity vehicles.



All Launch Site Names

- The Falcon 9 launches from 4 sites:
  - One in Florida:
    - CCAFS LC-40
    - CCAFS SLC-40
    - KSC LC-39A
    - VAFB SLC-4E

```
%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

Launch Site Names Begin with 'CCA'

- SpaceX launches from two sites at Cape Canaveral Space Force Station (prev. Air Force Base).
- Most flights launch from LC-40 rather than the SLC.

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE "CCA%" LIMI
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MA
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	

Total Payload Mass

- The Falcon 9 has carried a total payload of 45596 kg using boosters supplied by NASA.
- The cluster of low-payload ISS orbit flights suggest that this total payload was distributed across many routine missions.

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Custom
 * sqlite:///my_data1.db
Done.
SUM("PAYLOAD_MASS__KG_")
45596
```

Average Payload Mass by F9 v1.1

- The average payload carried by Falcon 9 v1.1, a model in use during 2013 and 2014, is 2928.4 kg.
- This query result does not include launches utilizing boosters.

Task 4

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booste
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- The first successful Ground Landing occurred on December 22, 2015.
- Five years of observed launches led the engineering team to this moment, and the successes picked up from there.

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" IS
 * sqlite:///my_data1.db
Done.
```

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000 kg

- Four boosters carry payloads in the 4000 to 6000 kg range:
  - F9 FT B1022
  - F9 FT B1026F9
  - FT B1021.2
  - F9 FT B1031.2

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Not every unrecovered first stage denotes an unsuccessful mission.
- Of 101 missions, nearly all have been recorded as successful.
- Controlled crash-landings are common in the space flight industry.

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL ORDER
```

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

Mission_Outcome	QTY
Success	101

Boosters Carried Maximum Payload

- Larger boosters are required to carry the maximum payload mass of 15600 kg.
- All flights with a payload mass over 10000 kg have occurred since 2019. As the engineers learn, bigger projects may be tackled.

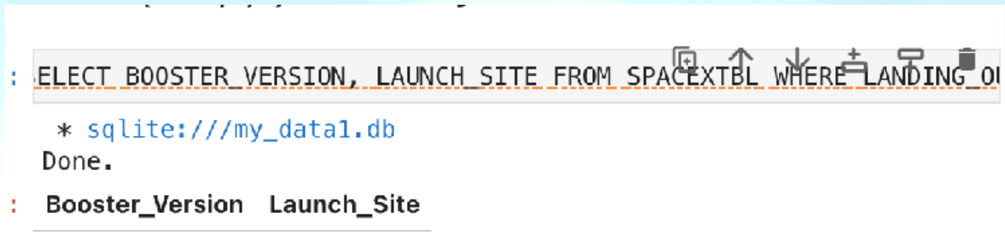
```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD
```

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

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

- In 2015, two launches failed to land on the drone-ship.
  - The missions launched from CCAFS LC-40.
  - Their booster rockets were B1012 and B1015, respectively.
  - Both were bound for the ISS in Low Earth Orbit.



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

- No attempt understandably leads the pack, as early SpaceX flights did not set a goal of landing the first stage.
- Most early attempts at landing occurred over the ocean, minimizing risk in the event of failure.
- Few ground landings were attempted between 2015 and 2017, but these were largely successful.

```
sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	QTY
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

Section 3

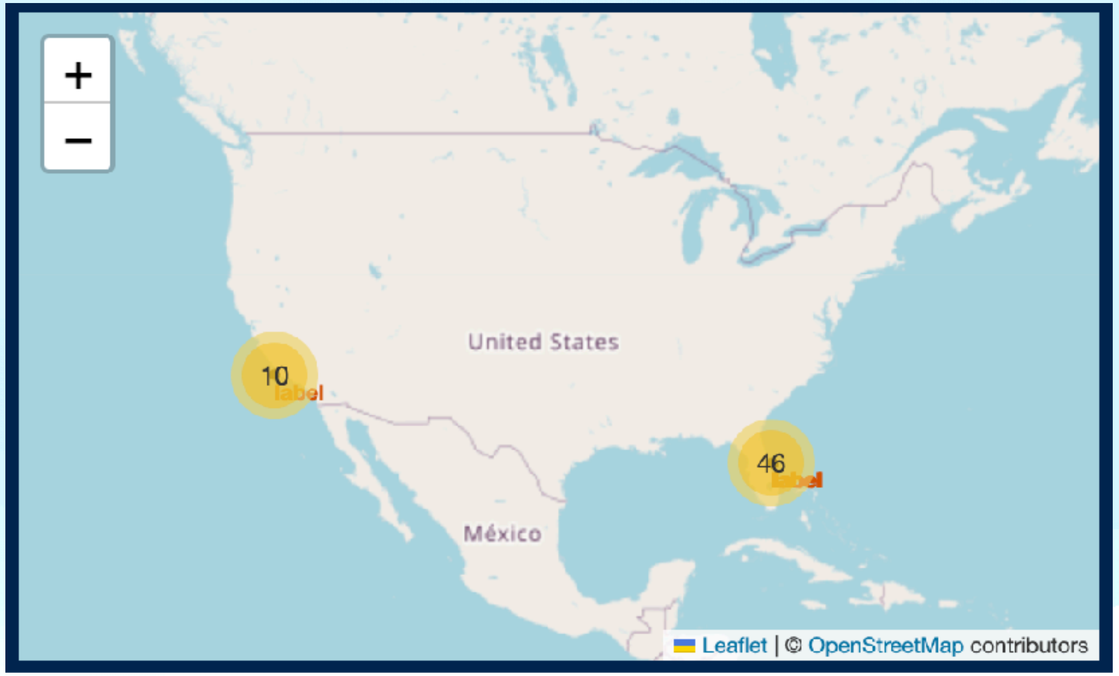
Launch Sites Proximities Analysis

Map of Falcon 9 Launch Sites

- The Falcon 9 launches from 4 sites:
  - One in Florida:
    - CCAFS LC-40
    - CCAFS SLC-40
    - KSC LC-39A
  - And one in California:
    - VAFB SLC-4E

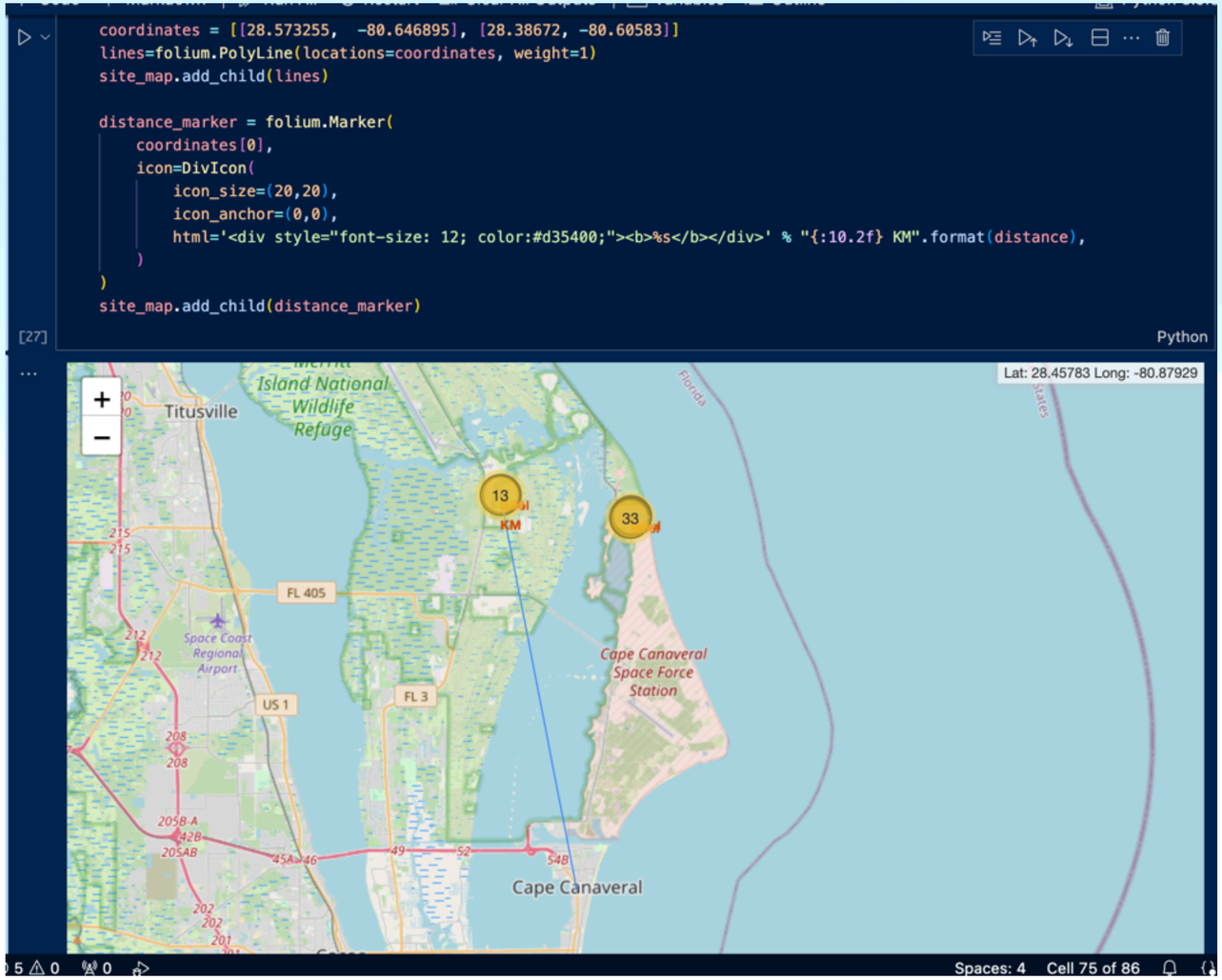
Launch Location Numbers

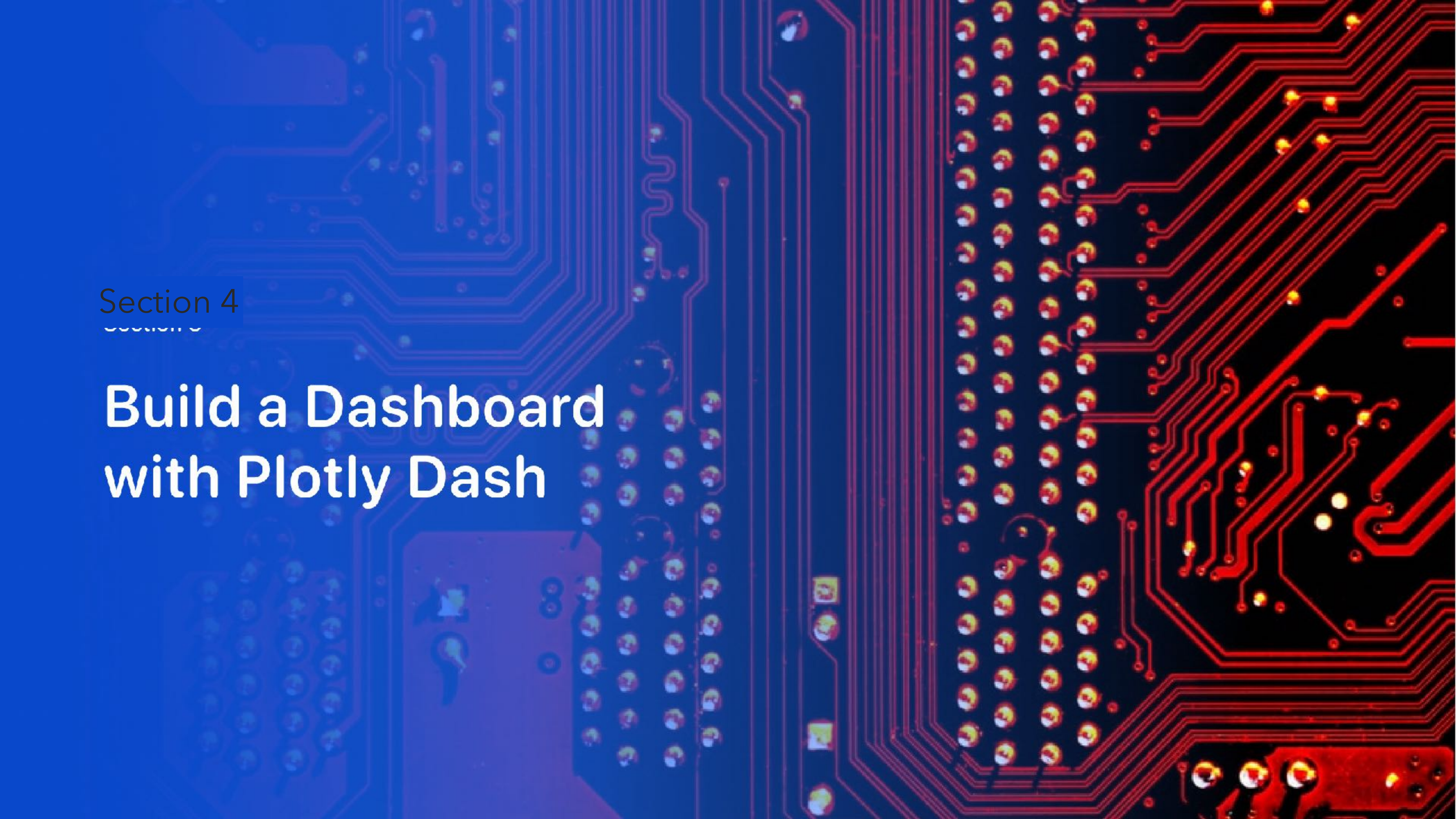
- Florida Launch sites are slightly overrepresented in the data, with 33 of 56 launches occurring in the CCAFS launch sites.
- KSC LC-39A holds another 13 launches; VAFS in California holds the other 10.
- Cape Canaveral hosts significant rocket-launch infrastructure and benefits from proximity to the equator, at around 28 degrees N longitude.



Distance to City – Cape Canaveral

- The Cape Canaveral launch sites benefit from close proximity to the water. This allows SpaceX to test both ground pad and ocean pad landings.
- Planned recovery failures are safer in the water, rather than land.
- Short driving distance to Cape Canaveral.

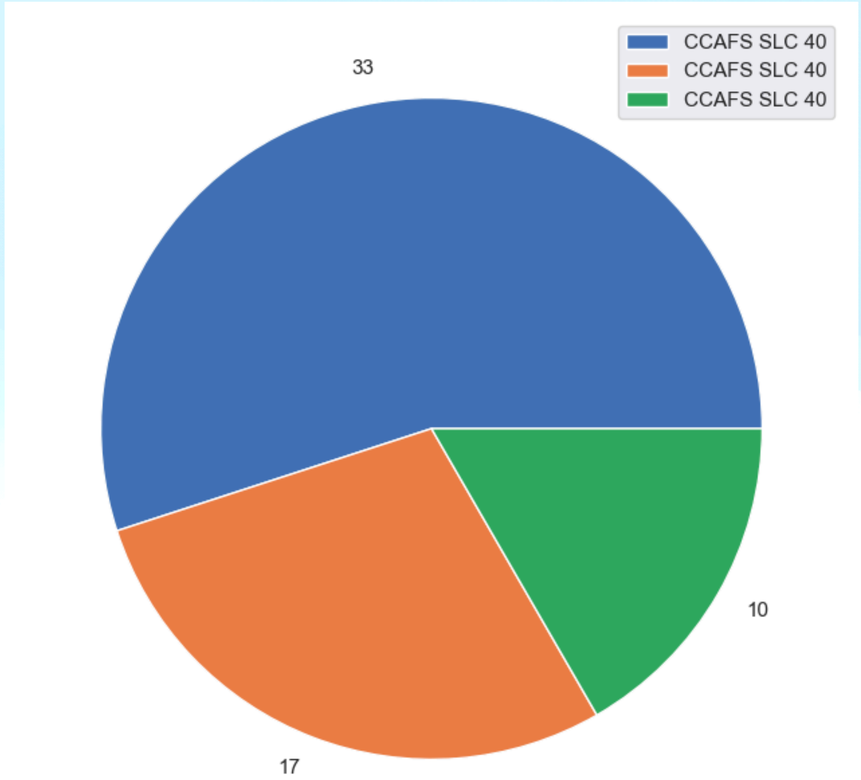



Section 4
Build a Dashboard
with Plotly Dash

Launch Success Charts

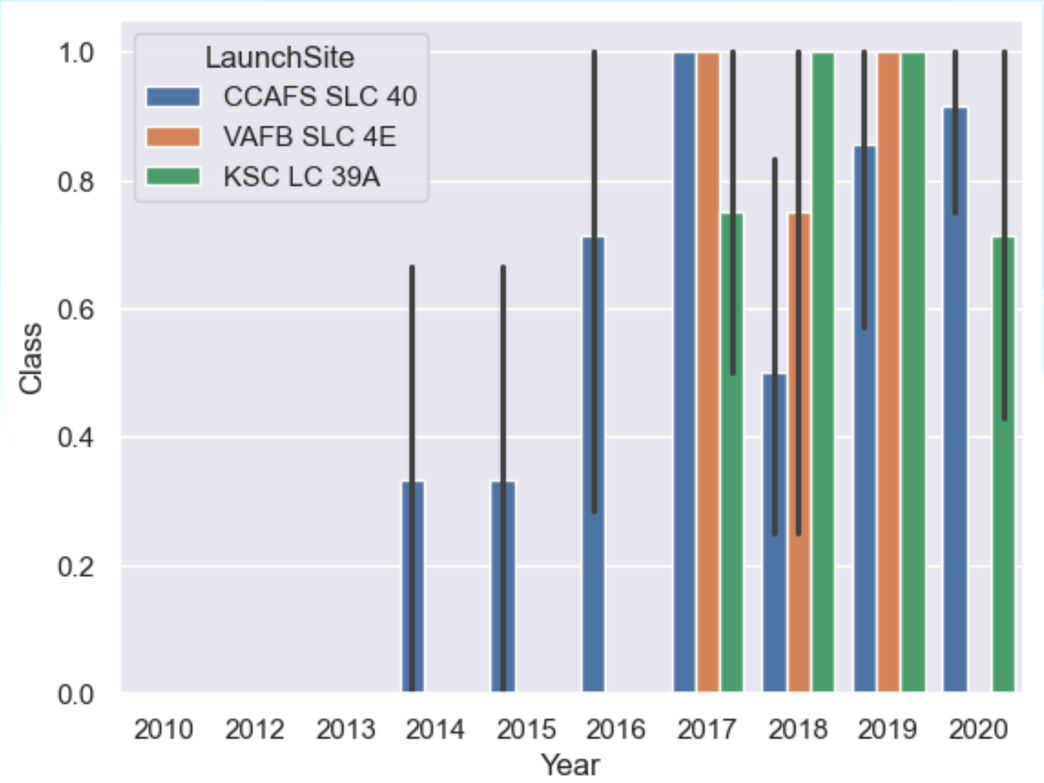
- Show the screenshot of launch success count for all sites, in a piechart

- Explain the important elements and findings on the screenshot



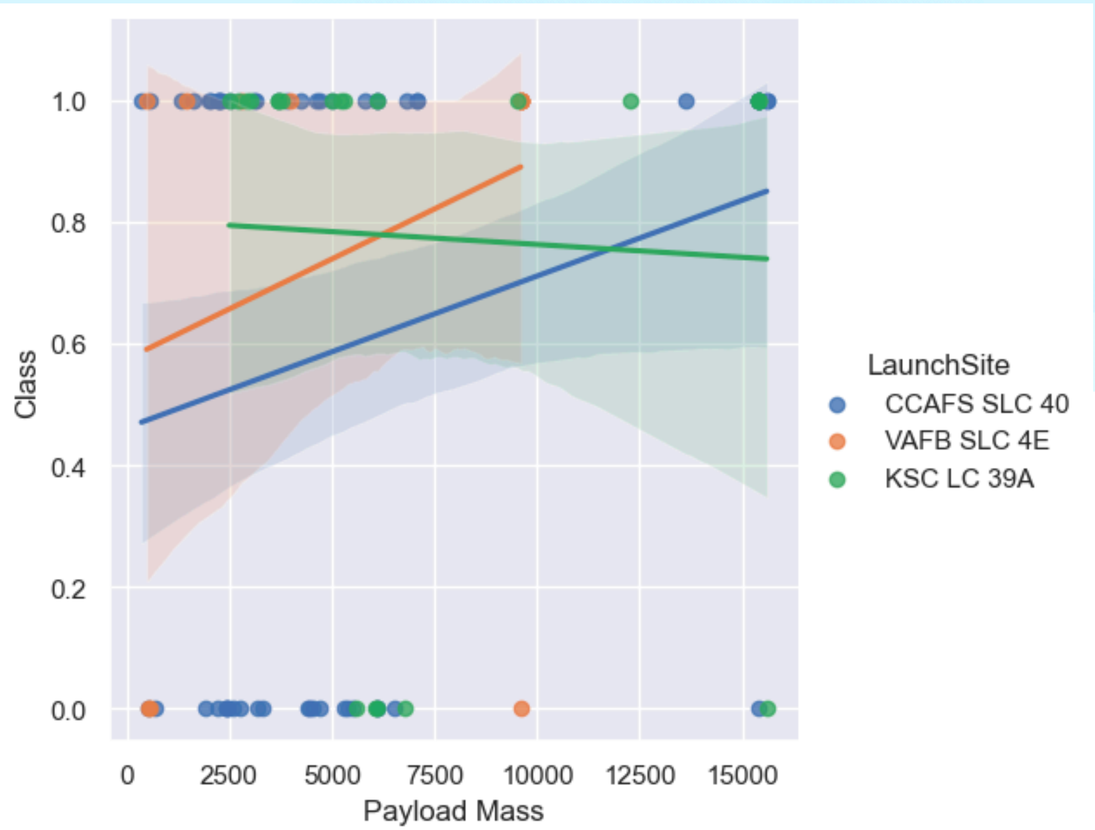
Success Rates of Each Launch Sites

- The success rate of CCAFS SLC 40 shows positive trend year-over-year.
- Smaller numbers of launches exaggerate variations in the data for the other two launch sites.



Launch Outcome and Payload

- This chart shows the launch outcome variance by payload mass, with additional parameters for launch site.
- Higher payload mass indicates more advanced boosters, which have a higher success rate on average.

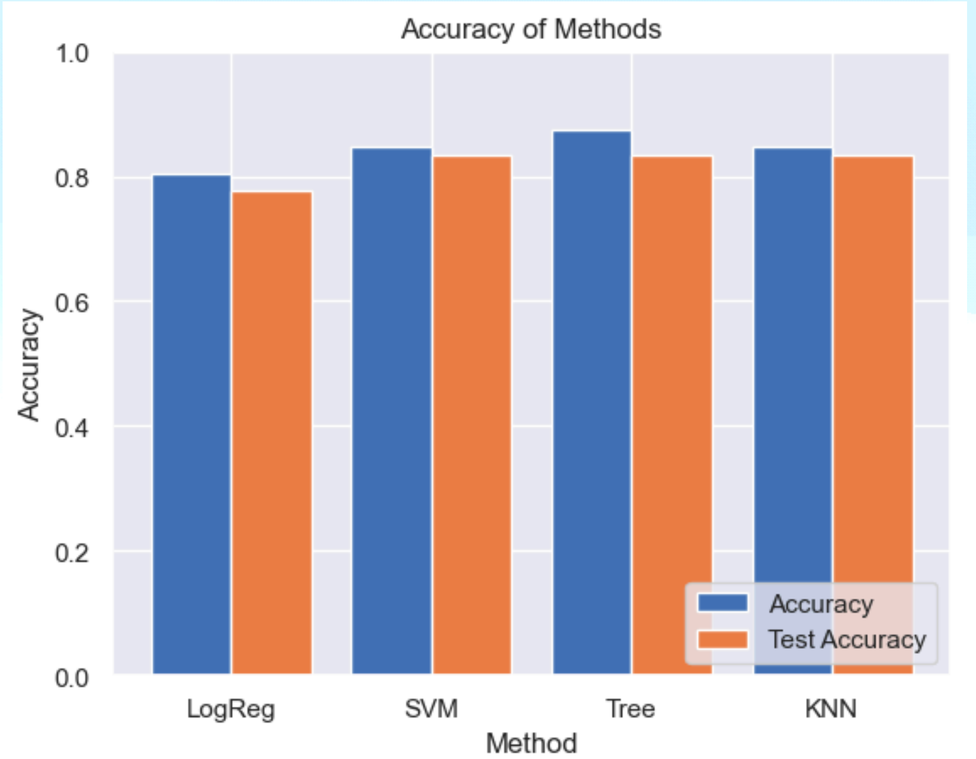


Section 5

Predictive Analysis (Classification)

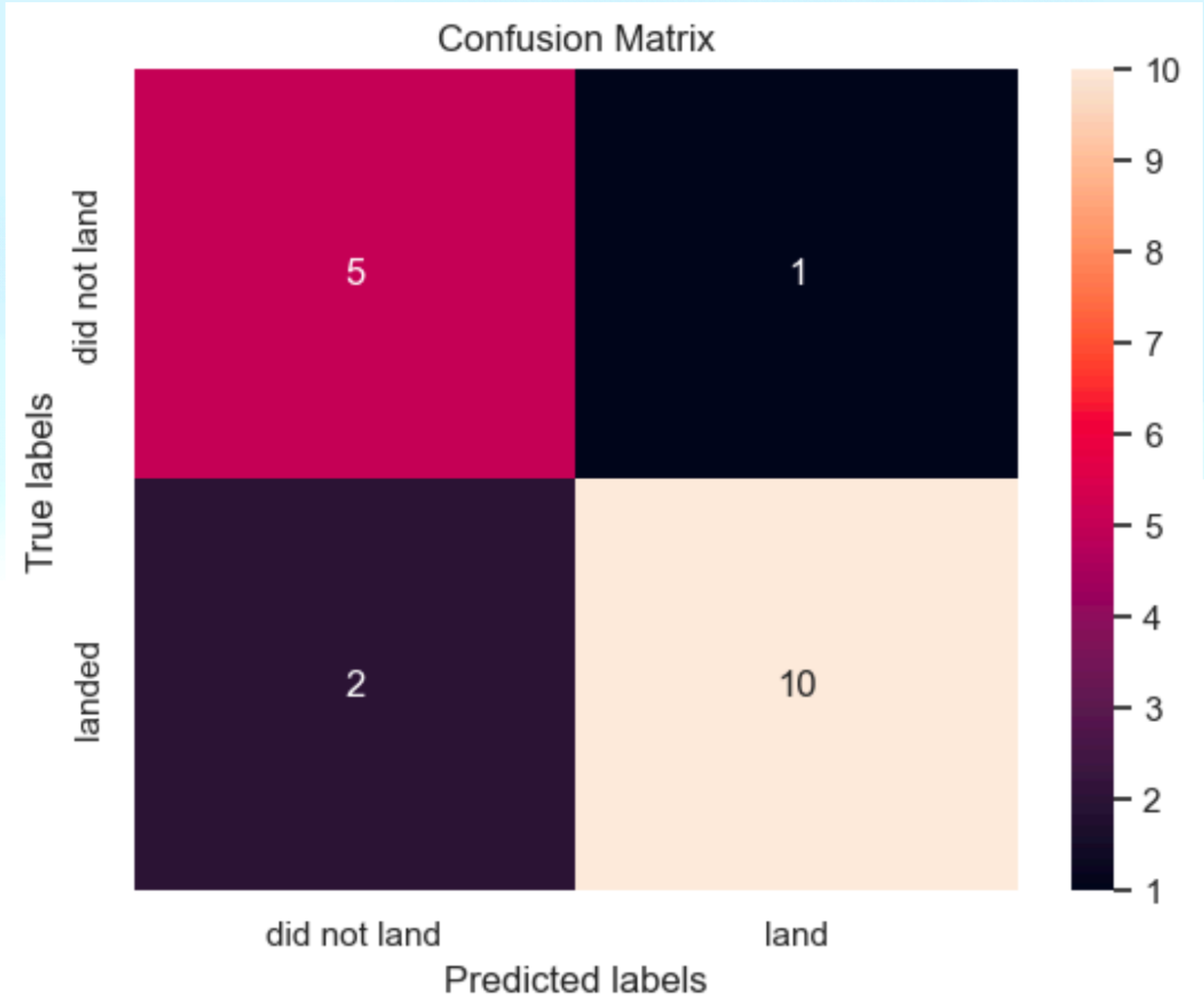
Classification Accuracy

- The bar chart on the right displays the machine learning algorithms trained to test the data.
- Through multiple random states and test sizes, the Decision Tree Algorithm showed a small but persistent advantage in predicting Launch Outcomes.
- It currently holds an 83% prediction success rate.



Confusion Matrix

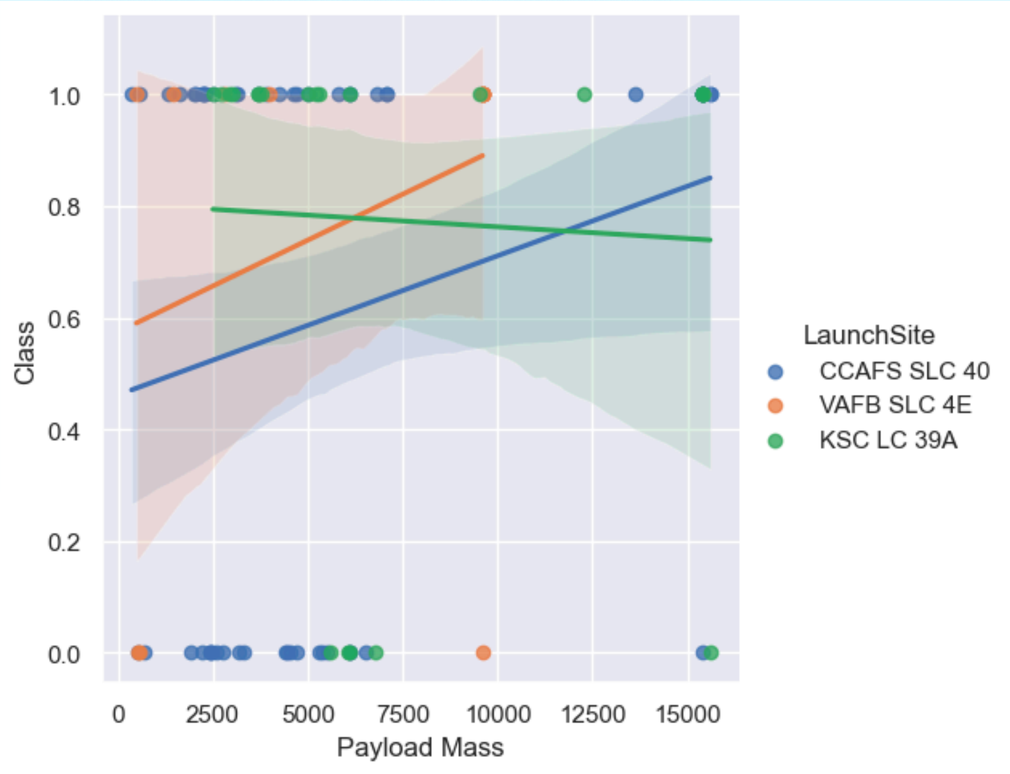
- The Decision Tree algorithm returned one false positive and two false negatives.
- Other classification algorithms returned a higher number of false positives.

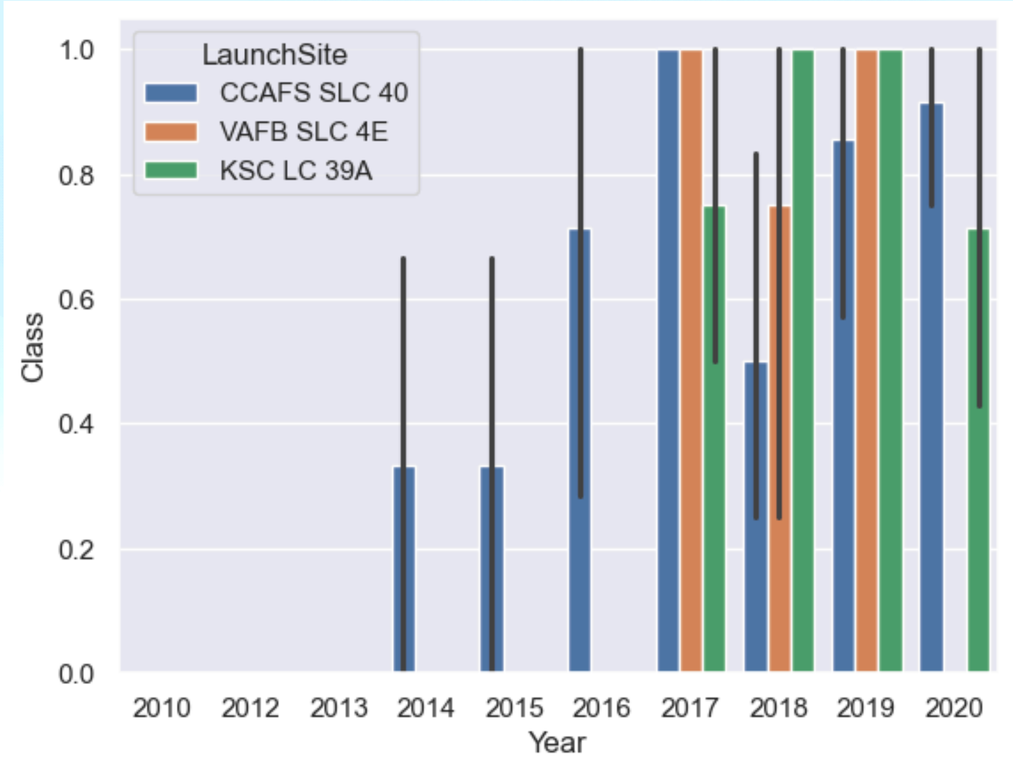


Conclusions

- The primary variable affecting the Launch success rate is Time. This comes as no surprise, given the ability of engineers to learn from previous launches.
- The Florida launch sites appears to have a lower success rate, but closer examination of the data shows a small sample size for the California site. Further data is needed.
- The Decision Tree algorithm holds an 83% success rate in identifying Launch outcomes, but still returns the occasional false positive. Future researchers should continue to test it on upcoming datasets with different randomization parameters.

Appendix





- Fig. 1: Bigger is better, or bigger is later is better?
- Fig. 2: Cape Canaveral leads in launch count…especially in the tricky early years.

Thank you!