Winning Space Race with Data Science

Greg Sasso
04/14/2023

Outline

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

Executive Summary

Methods!

- Collected data on Spacex launches by accessing the Spacex api and scraping the Falcon 9 Wikipedia page
- Classified landing outcomes as successful or not successful
- Perform exploratory data analysis using visualization and SQL
- Create interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Results!

- Launches have gotten more successful over time
- Launch sites are all close to the coast
- ES-L1, GEO, and SSO are the most successful Orbit types
- Classification models generally predict which launches will land successfully

Introduction

- Spacex has been launching rockets for almost 15 years. Understanding it's successes and failures is crucial improving consumer space travel.
- This projects seeks to answer several questions:
  - What makes a rocket launch successful?
  - Where are successful launches located?
  - What tools can we use to make sure more launches are successful in the future?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected data on Spacex launches by accessing the Spacex ap and scraping the Falcon 9 Wikipedia page
- Perform data wrangling
  - Classified landing outcomes as successful or not successful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used Scikit-learn to choose optimal classification models

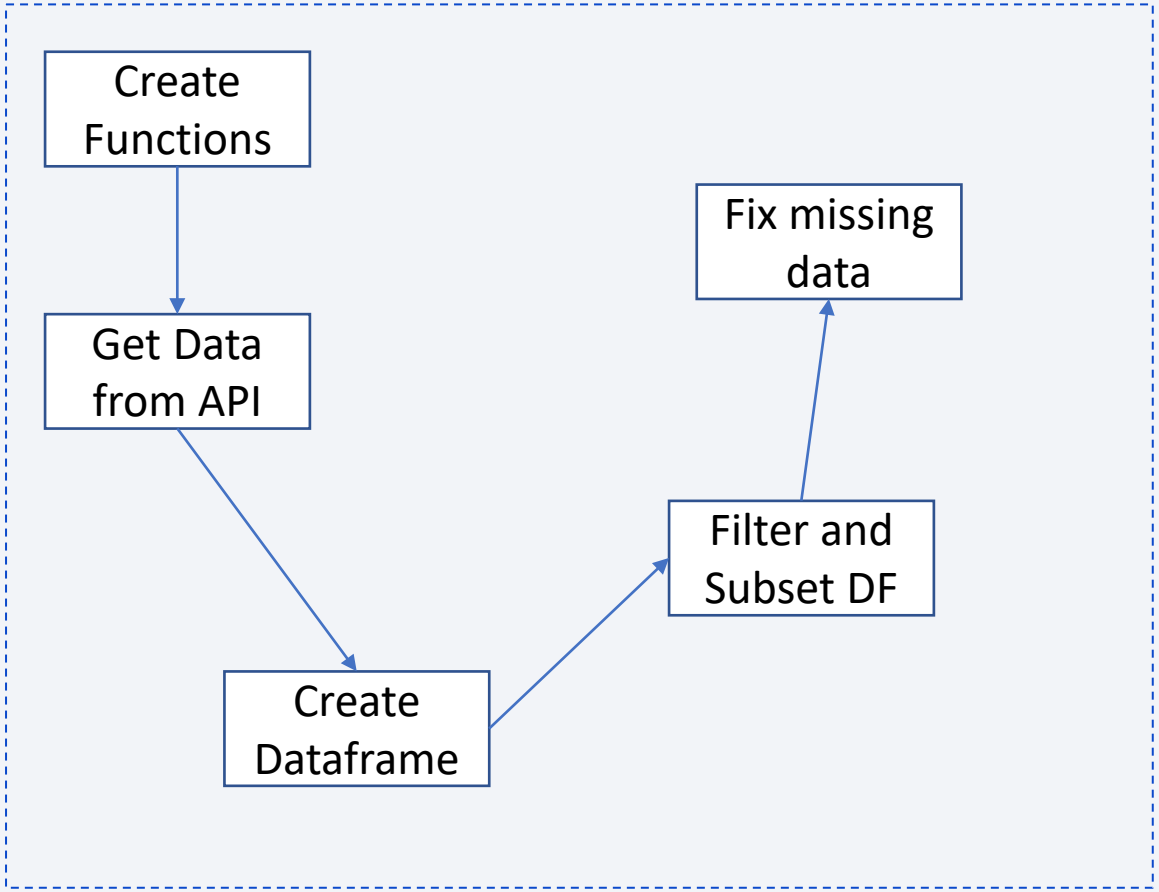
Data Collection

- I worked with two data sets
  - Data collected from the Spacex API.
    - Accessed multiple CSVs and merged them together to create the core Spacex data frame
  - Data from the Falcon 9 Wikipedia page
    - Used beautiful soup to scrape Falcon 9 data from tables on the Wikipedia page.

Data Collection – SpaceX API

1. Create functions to get specific data
2. Call data from Spacex URL
3. Transform called data to Dataframe
4. Separate out date specific to Falcon 9
5. Fix missing data by using average values

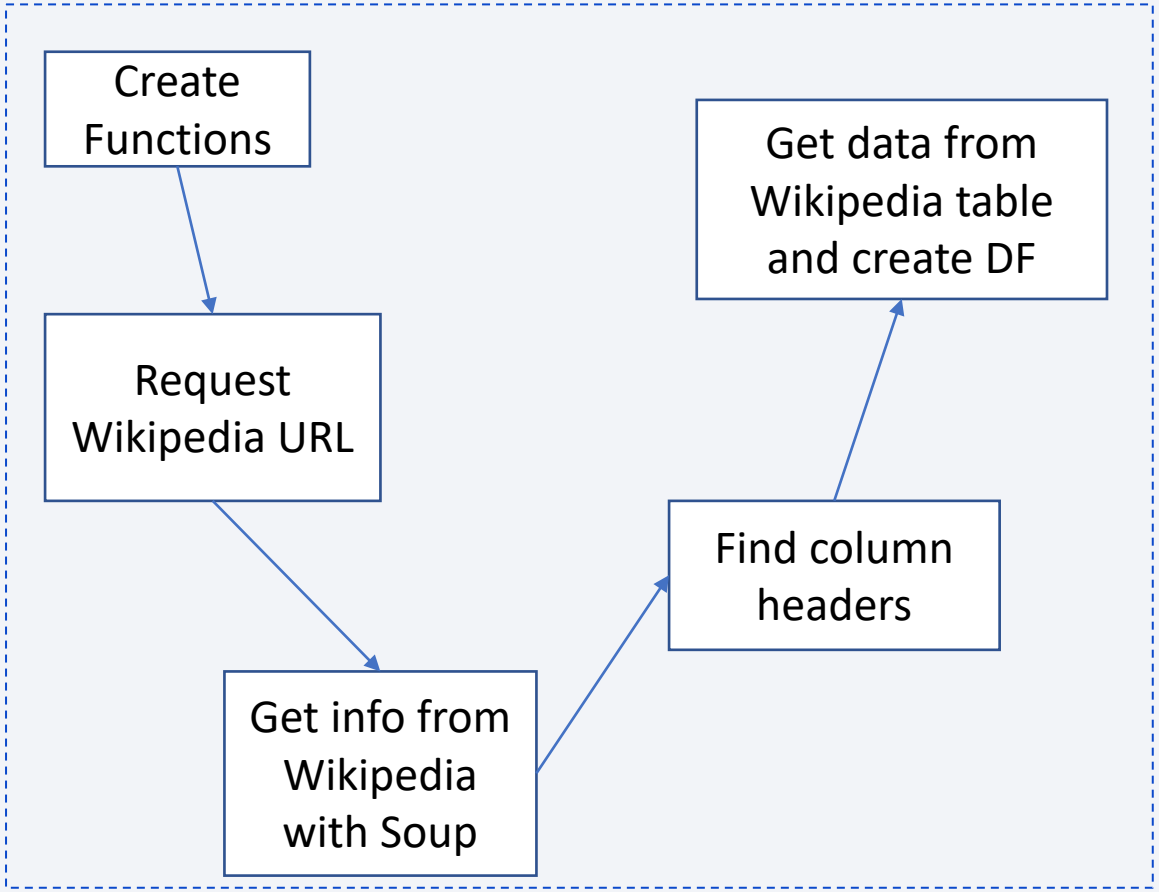
- Code: https://github.com/gregsasso/Final-Project/blob/main/Data%20Collection.ipynb



Data Collection - Scraping

1. Define functions to get specific data from the Wikipedia page table
2. Request Wikipedia url
3. Use BeautifulSoup to read information from Wikipedia page
4. Create column headers from Wikipedia table headers
5. Extract data from Table on Falcon 9 launches to create data frame

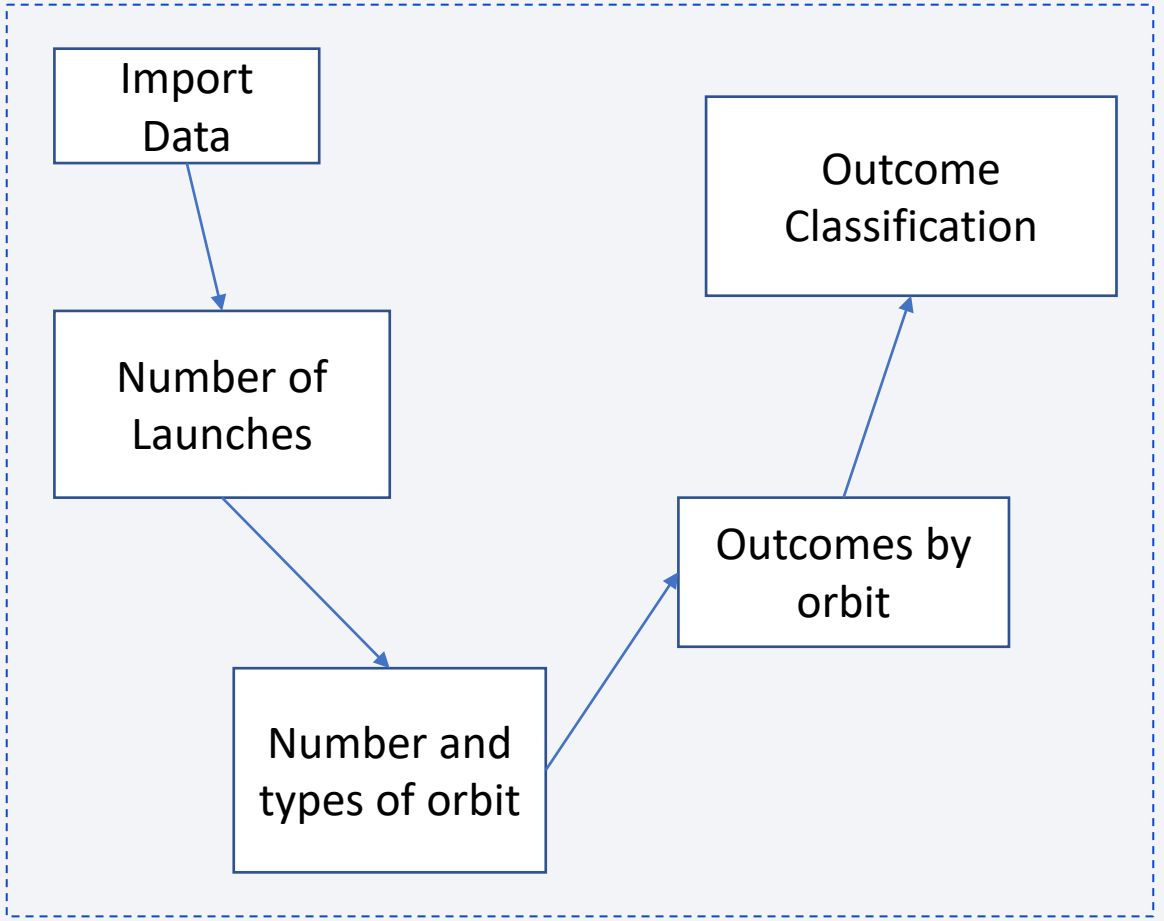
- Code: https://github.com/gregsasso/Final-Project/blob/main/Web%20Scraping.ipynb



Data Wrangling

1. Import data from csv
2. Check number of launches by site
3. Check number of each type of orbit
4. Check outcomes by orbit
5. Classify outcomes as either good or bad

- Code: https://github.com/gregsasso/Final-Project/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- Used seaborn scatter plots to visualize multiple two variable relationships grouped by success type (Flight number vs launch site, Payload vs launch site, Payload vs Orbit type)
  - These allows us to quickly see if there are any relationships between the two variables
- Used a bar chart to see the success types of various orbits
  - Bar chart allow us to compare multiple categories of a single variable quickly
- Used a line chart to track success rates over time.
  - Line chart is ideal for check time trends
- Code: https://github.com/gregsasso/Final-Project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries used for exploratory analysis:
  - Queried the distinct launch site names
  - Displayed 5 launches where the launch site began with CCA
  - Found the combined payload for all launches
  - Found the average payload for Falcon 9 V1.1
  - Found the date of the first successful landing
  - Displayed boosters with moderate payloads that had successful drone landings
  - Calculated total number of success and failures
  - Found boosters that carried the highest payload
  - Found the failed landings in 2015
  - Made a list in ascending order of all landing outcomes
- Code: https://github.com/gregsasso/Final-Project/blob/main/SQL.ipynb

Build an Interactive Map with Folium

- Added markers for all launch sites
- Added circles to all launch sites to indicate number of lunches per site
- Further added market clusters to indicate successful and unsuccessful launches per site
- Added lines to important nearby infrastructure such as highways, cities, train lines, and the coast
- Added all these objects help visualize important determinants of location and which locations had more successful launches
- Code: https://github.com/gregsasso/Final-Project/blob/main/lab_jupyter_launch_site_location.ipynb

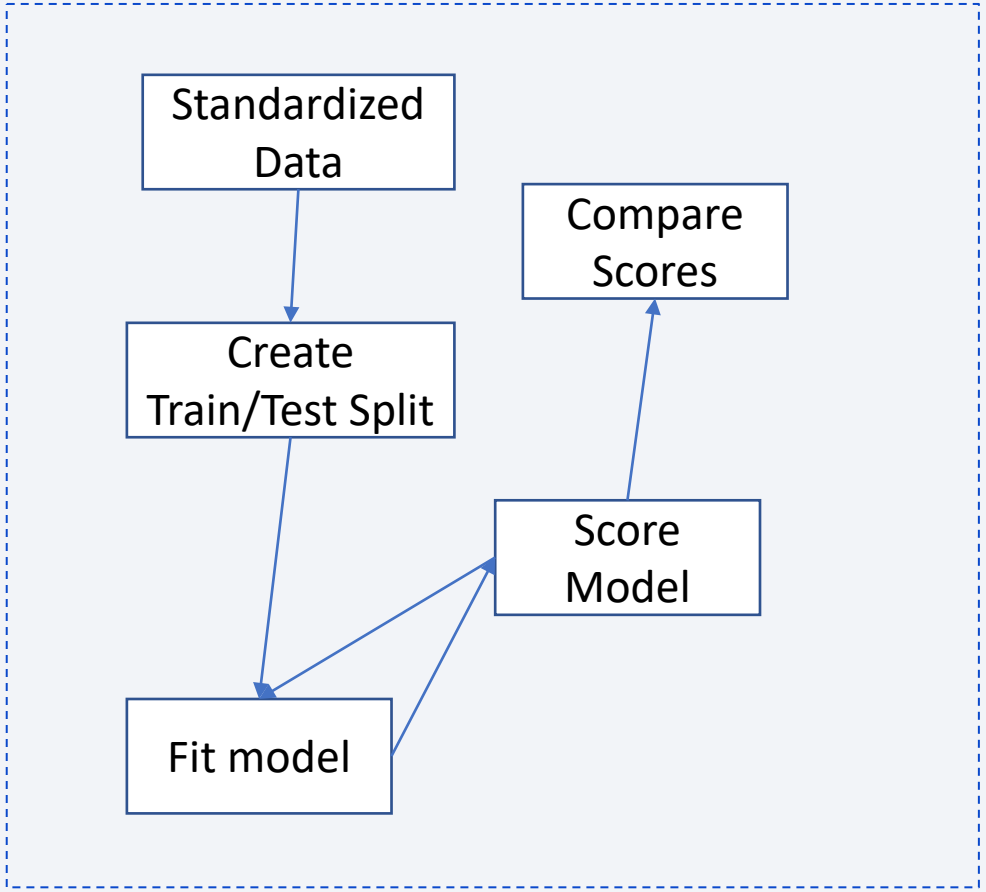
Build a Dashboard with Plotly Dash

- Created interactive piecharts to show successful launches broken down by launch site. If a specific launch site is selected, the pie chart shows the breakdown of successes vs failures
  - This chart allows sites to be compared by success rate at a quick glance
- Created an interactive scatter plot showing the breakdown of successful launches by payload and launch site.
  - This chart allows quick comparison about which payloads are most successful and a quick comparison between which sites launch which payloads.
- Code: https://github.com/gregsasso/Final-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

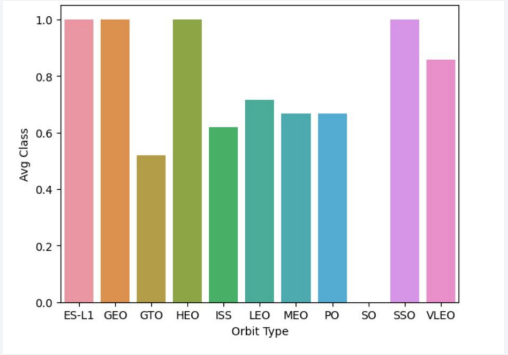
1. First step is to standardize the data and split the data in to training and test samples
2. Then the following steps to fit and score each model type (Logistic, SVM, Decision Tree, KNN):
    1. Do a grid search to find optimal parameters
    2. Estimate model
    3. Find the score (predictive accuracy) of the model
3. Compare scores across models to find the optimal classification model.

- Code: https://github.com/gregsasso/Final-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



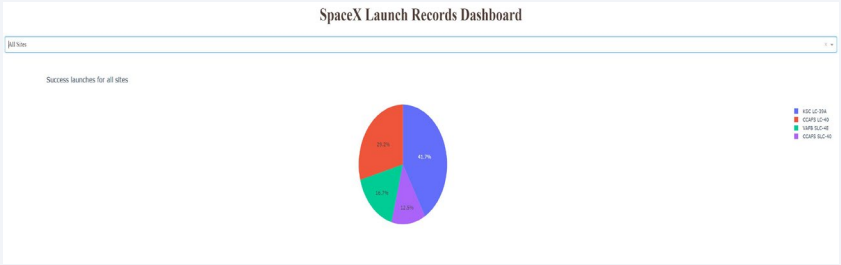
Results

- Exploratory data shows correlations between different launch characteristics
  - Example of orbit type and launch success:



  - SQL queries used to find specific groupings such as all successful launches for medium payload boosters

- Used interactive dashboard to summarize successes and payloads:



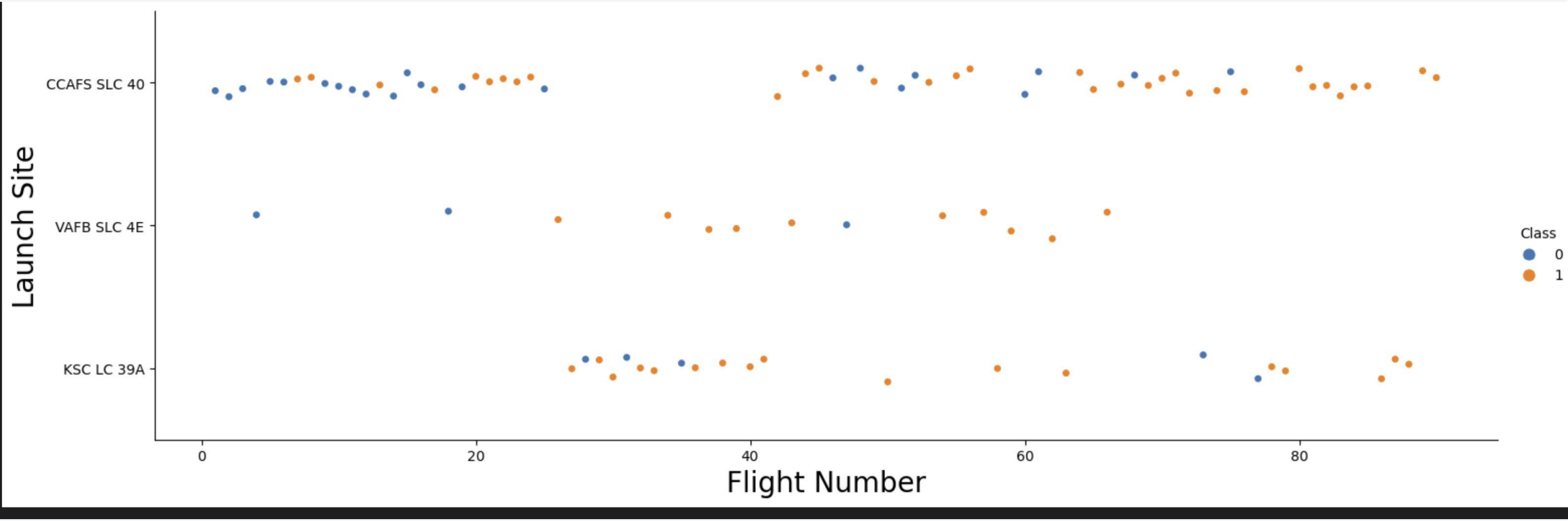
- Used machine learning classification algorithms

to predict which launches would be successful and found KNN is the best algorithm for this task

Section 2

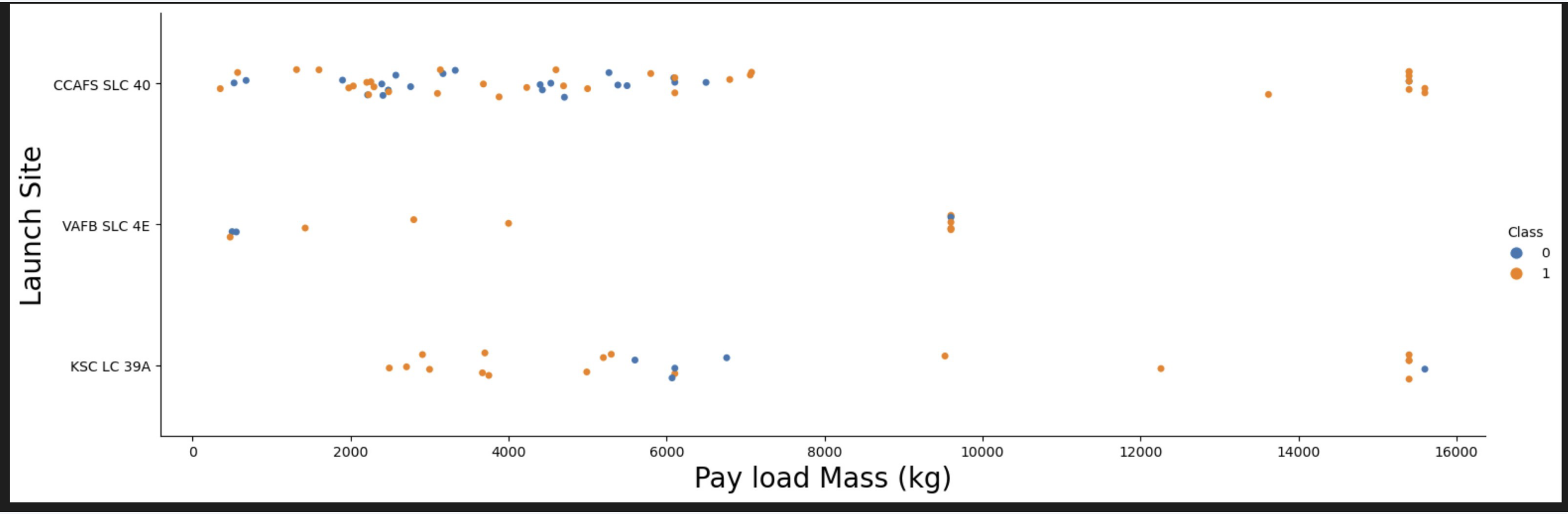
Insights drawn from EDA

Flight Number vs. Launch Site



- We can see that CCAFS has by far the most launches. KSC was used predominantly for launches between 23 and 41 and then a few after. VAFB has few launches.

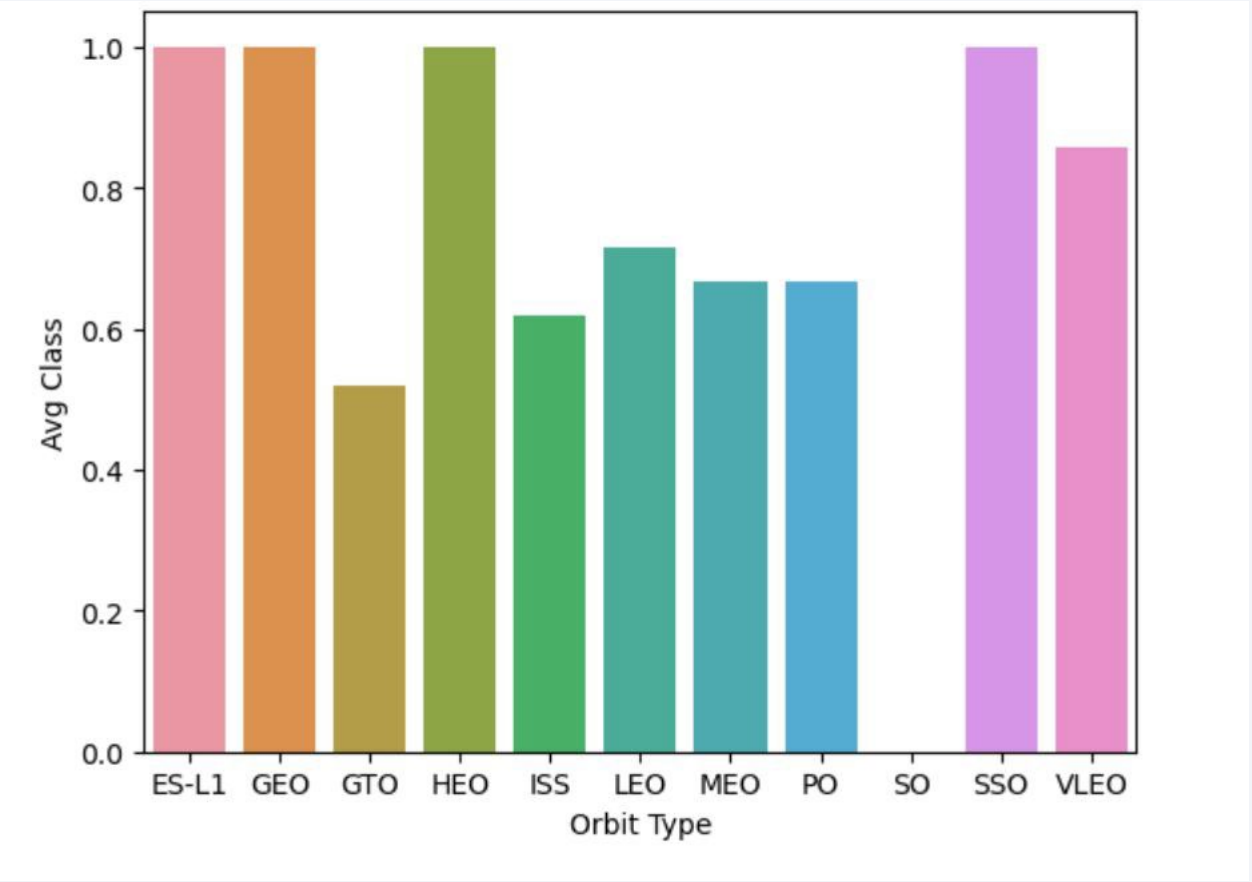
Payload vs. Launch Site



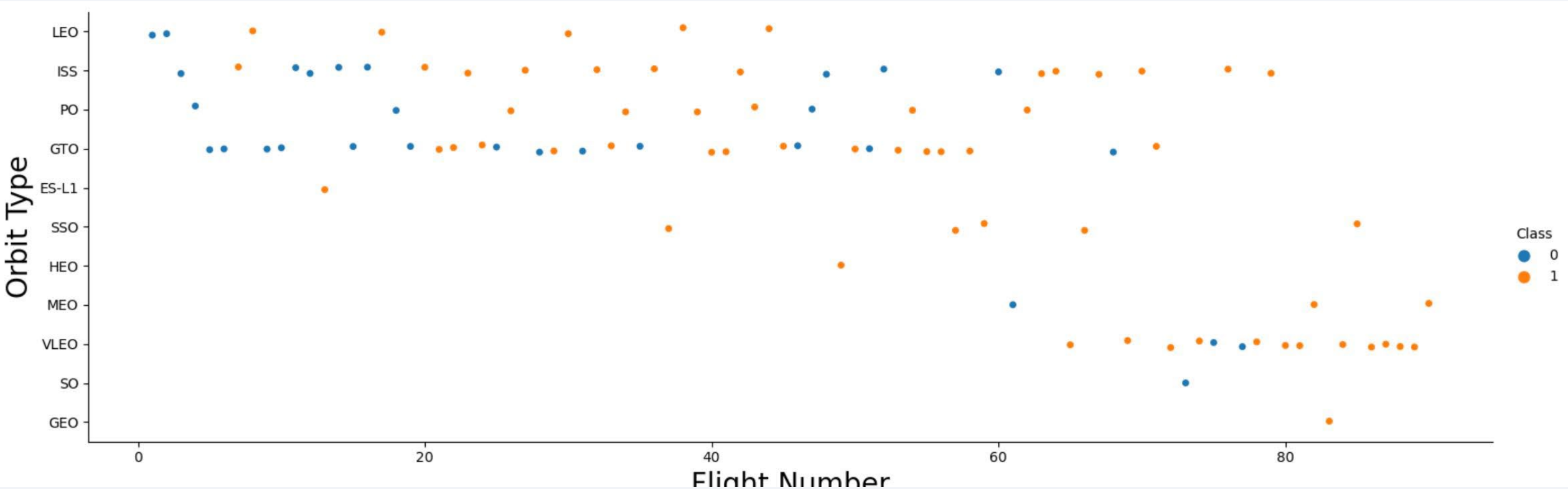
- The VAFB site only handles low and moderate payloads. CCAFS has the greatest payload variety. Both CCAFS and KSC handle the heaviest payloads.

Success Rate vs. Orbit Type

- This chart shows the success rate (“Class” on the y axis) by various Orbit types
- ES-L1, GEO, and SSO are the most successful, having perfect success rates.
- SO is the worst, having no successes.

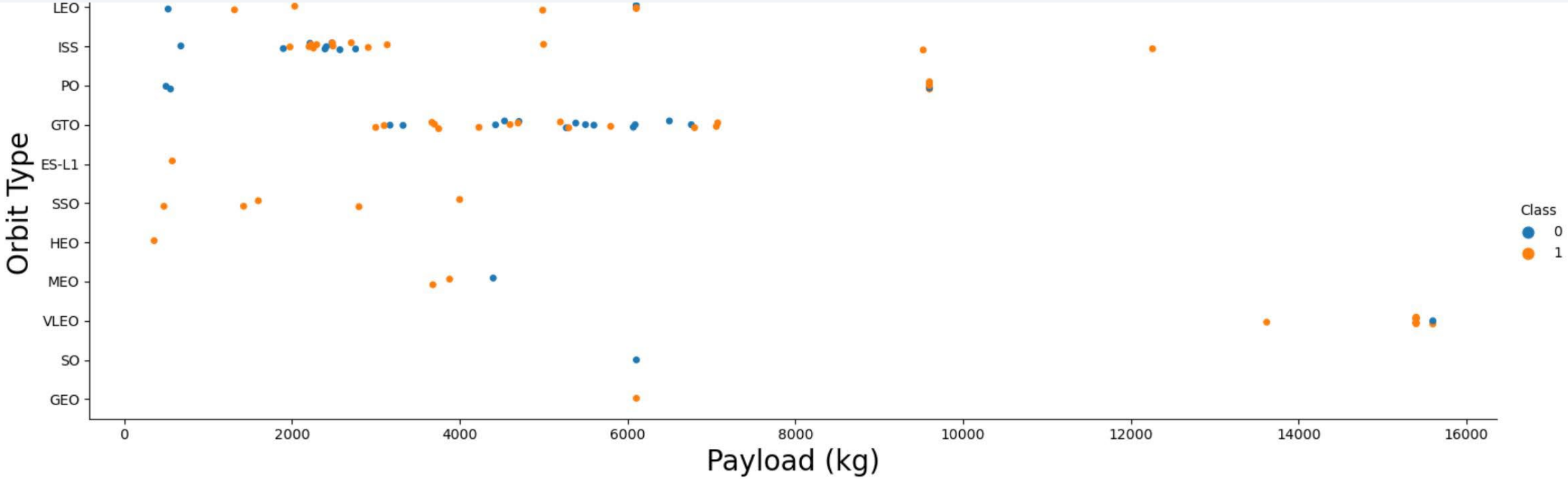


Flight Number vs. Orbit Type



- The scatter plot shows flight number on the x axis and orbit type on the y axis.
- Note that type of orbits has changed over time. LEO, ISS, PO, and GTO were common early. VLEO has be come much more common for later flights.

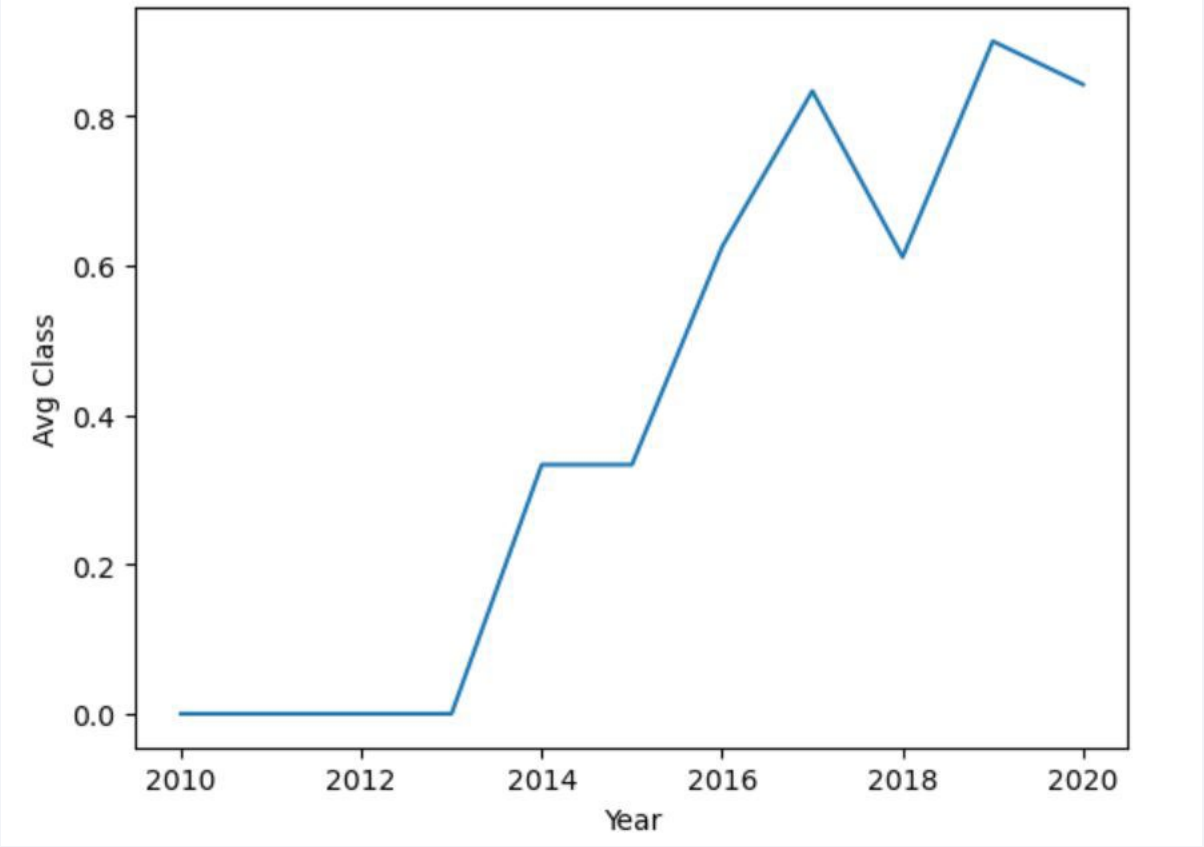
Payload vs. Orbit Type



- With payload on the x axis and orbit type on the y axis, we can see that low payloads are concentrated with the ISS orbit type. Moderate payloads are concentrated in the GTO orbit type, and heavy payloads are concentrated in the VLEO orbit type.

Launch Success Yearly Trend

- There is a clear trend of more successful launches over time with 0 in 2010 and 80% in 2020.



All Launch Site Names

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

- Using a sql query with the DISTINCT operator, we were able to find a list of all Spacex launch sites

Launch Site Names Begin with 'CCA'

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

- Looked for the first five records where launch site began with ‘CCA’, and copied all information for those records.

Total Payload Mass

- The total Payload from all booster is 619967. To find this, I used the Sum function on the payload_mass_kg_ column.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534
- I used the AVG function on payload_mass_kg and WHERE Booster_version LIKE 'F9 v1.1%' to only count booster F9 v1.1.

First Successful Ground Landing Date

The first successful ground landing was 2010-06-04

- I used the MIN function on DATE where mission_outcome LIKE 'Success' to find the first successful landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

- I used a query to only select boosters WHERE landing_outcome LIKE 'Success (drone%' AND payload_mass_kg_ BETWEEN 4000 AND 6000. This limited the results to only the ones sought.

Total Number of Successful and Failure Mission Outcomes

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- I found the total number of successes and failures by using COUNT and grouping by mission_outcome.

Boosters Carried Maximum Payload

- Used a subquery to select only boosters where payload = MAX(payload_mass_kg_).

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

2015 Launch Records

	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Used a query to only select year 2015 and only recorders where landing_outcome had ‘Failure (drone ship)’

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

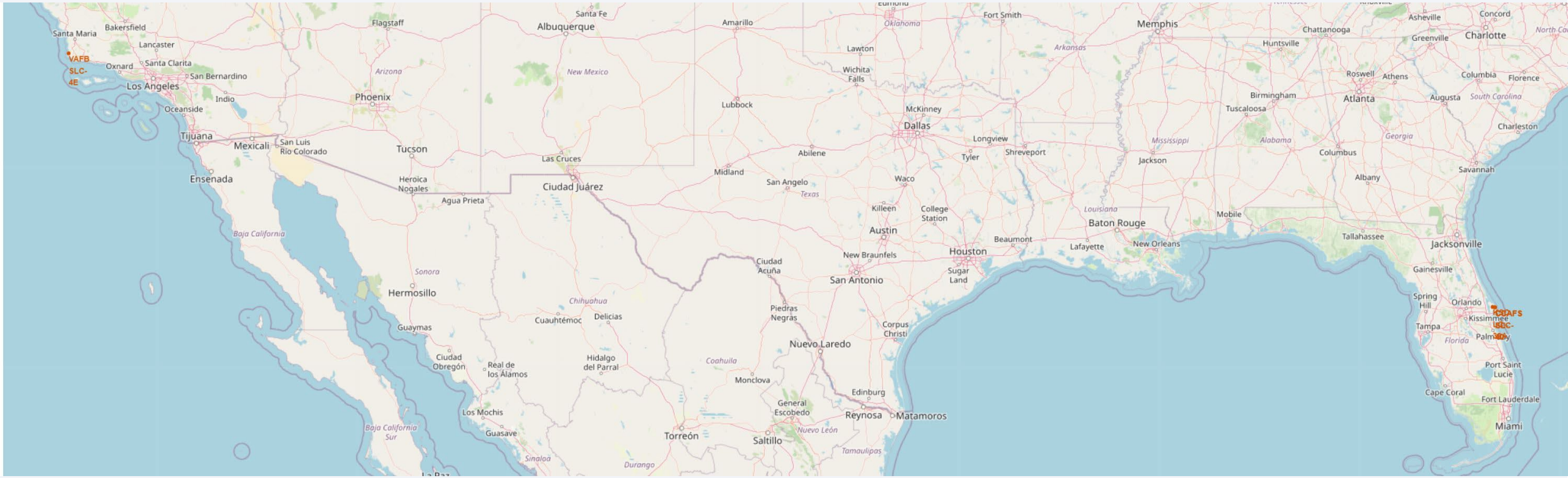
Selected the count of landing_outcomes and grouped by landing_outcome and then ordered them in ascending order.

1	**landing_outcome**
1	Precluded (drone ship)
2	Failure (parachute)
2	Uncontrolled (ocean)
3	Failure
5	Controlled (ocean)
5	Failure (drone ship)
9	Success (ground pad)
14	Success (drone ship)
22	No attempt
38	Success

Section 3

Launch Sites Proximities Analysis

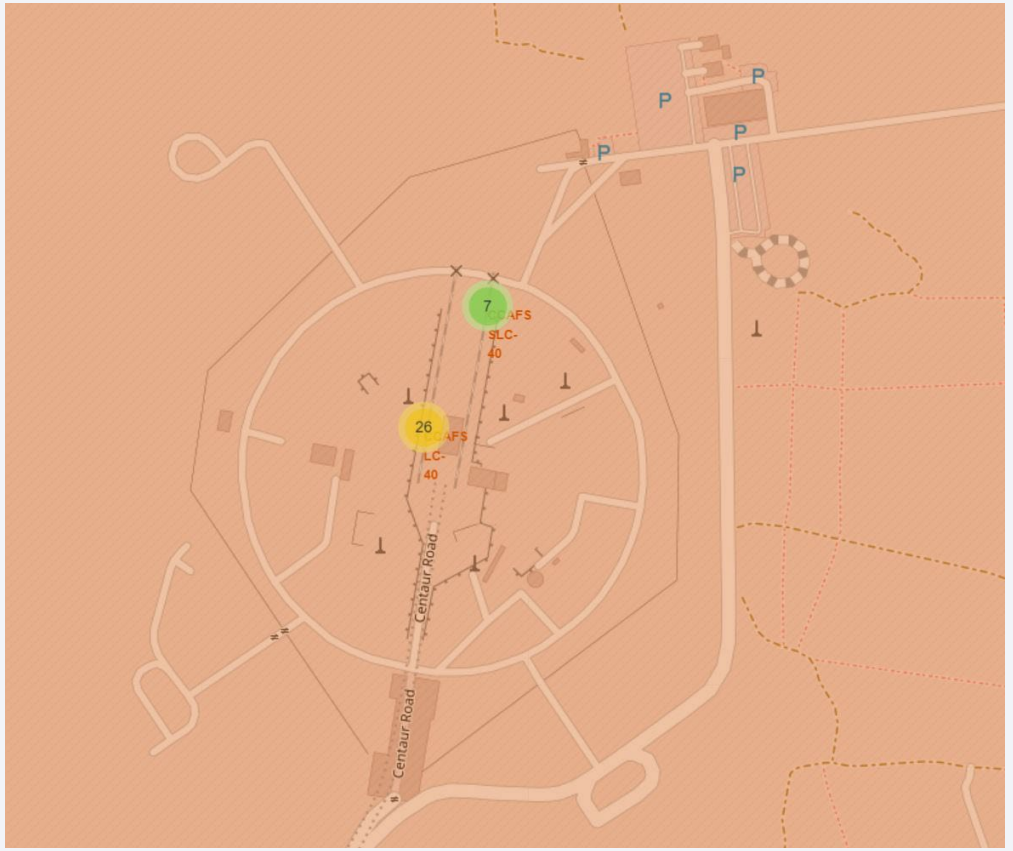
Map of Launch Sites



- Here we can see that there is one launch site (VAFB SLC-4E) in California west of LA and the other three launch sites are on the east coast of Floriday.

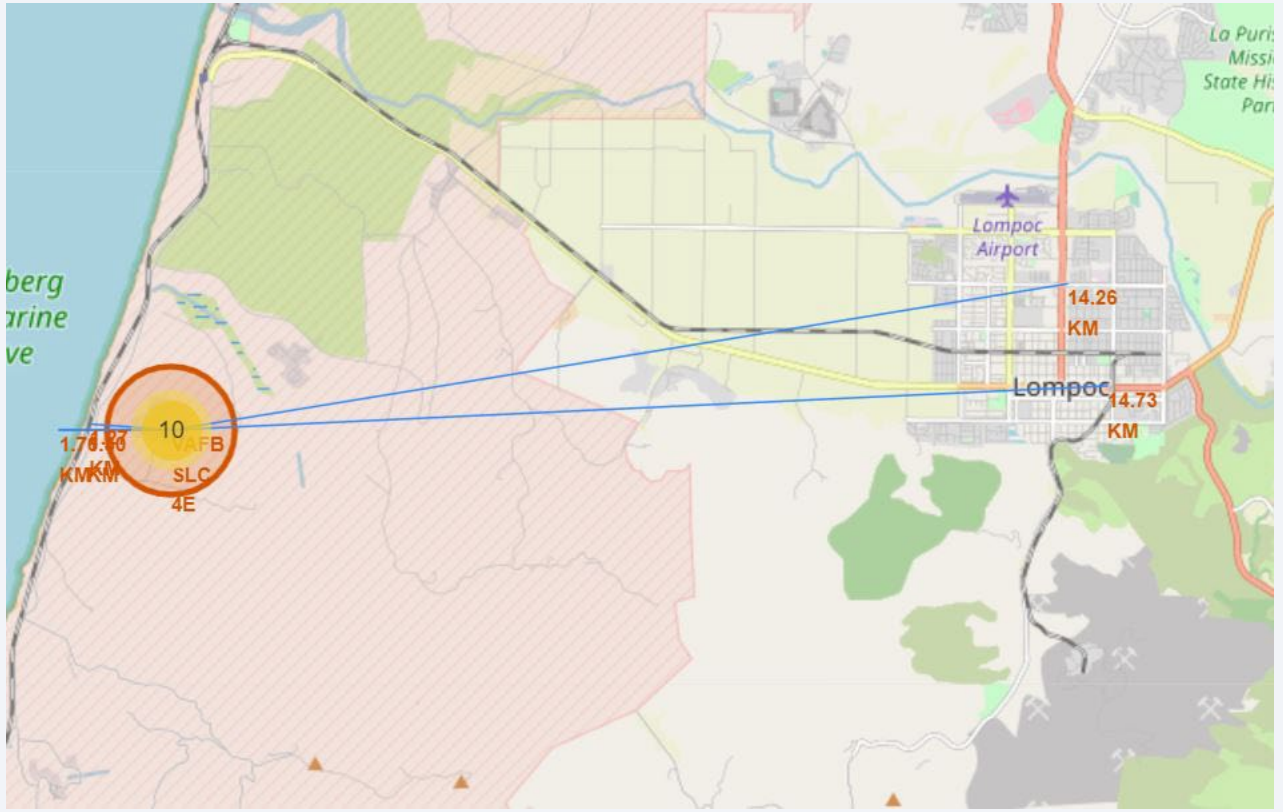
Successful Launches

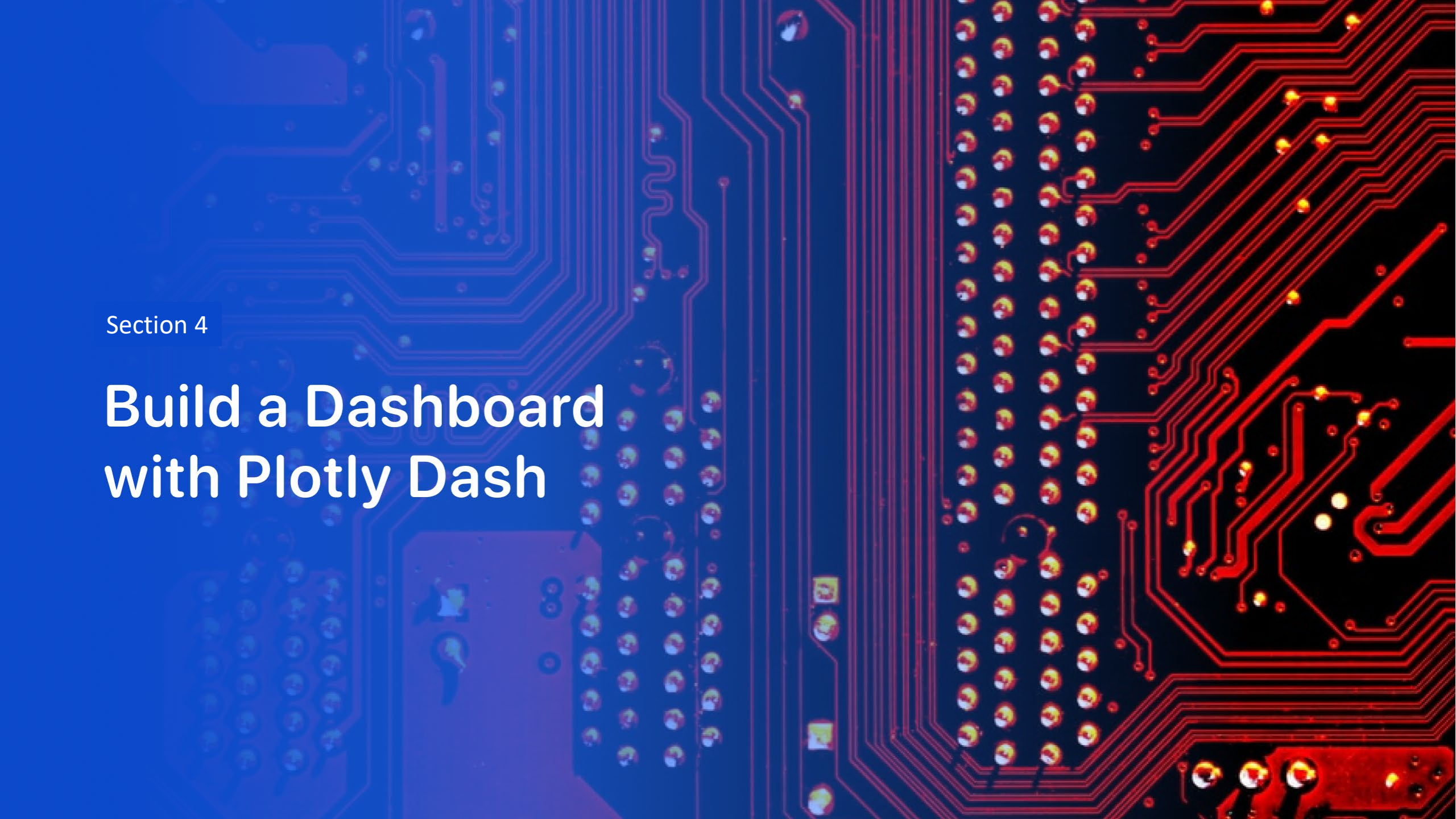
- This map shows successful and unsucesful launches for two of the Florida launches sites (CCAFS LC-40 and CCAFS SLC-40)



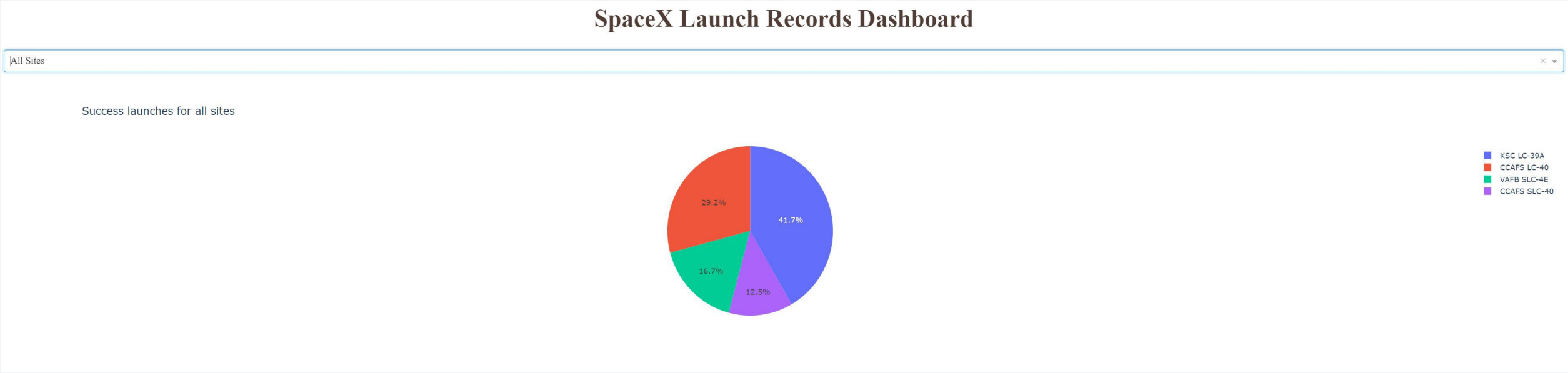
Launch Site Characteristics

- This map shows that site VAFB SLC-4E is very close both to the coast and to a train line.
- It is moderately close (under 15 kilometers) to a small city (Lompoc) and to a major highway.



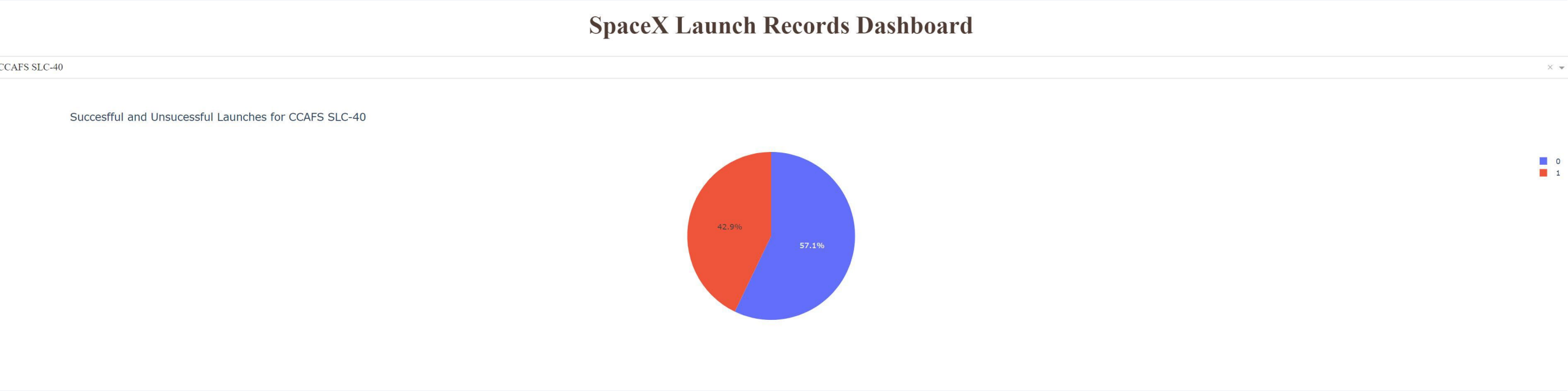

Section 4
Build a Dashboard with Plotly Dash

Successes by Launch Site



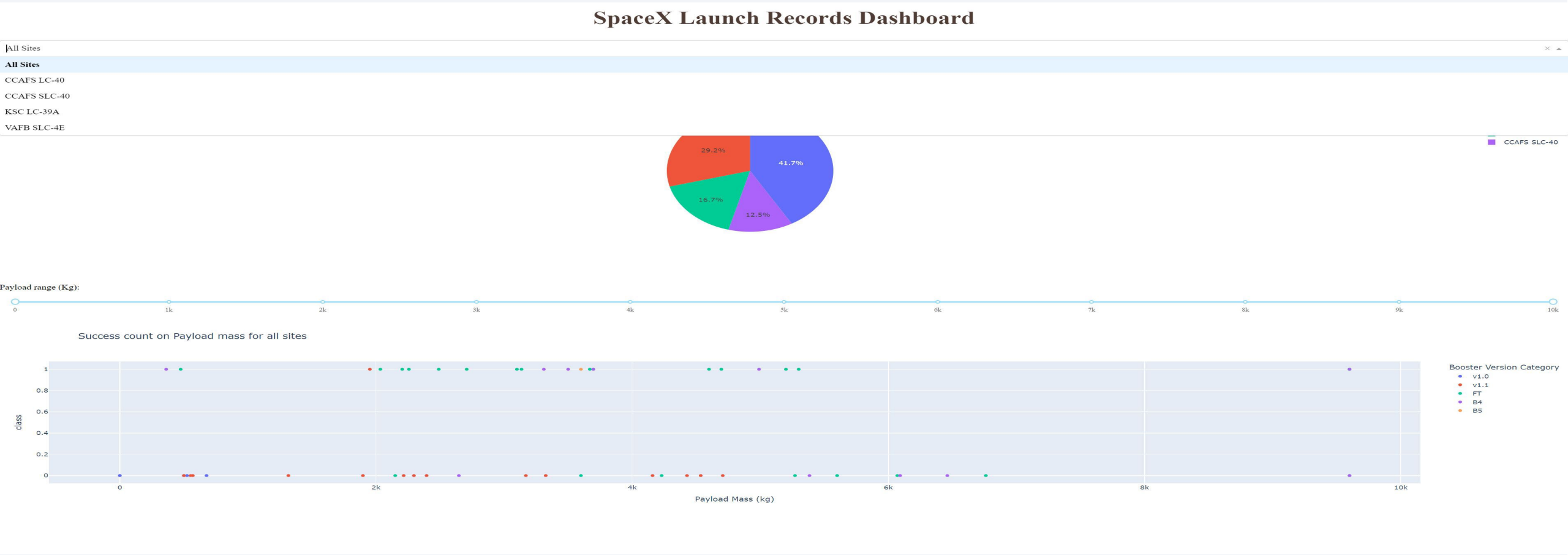
- This piechart shows which launch sites had the greatest share of successful launches. Site KSC LC-39A has the most successful launches of any launch site.

Launch site CCAFS SLC-4



This piechart shows that Site CCAFS SLC-40 had a more unsuccessful launches (57.1 %) than successful launches (42.9%).

Booster success rates.



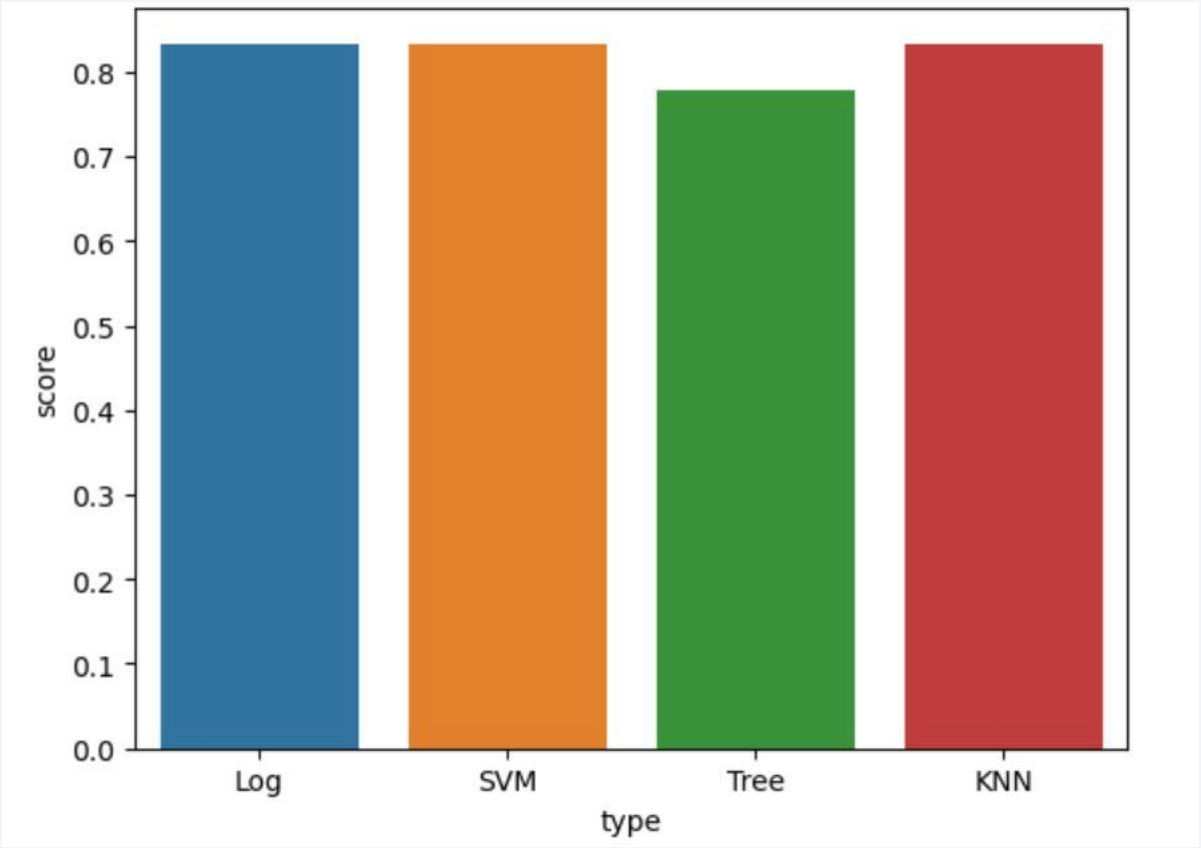
The bottom scatter plot shows the various successes and failures for different booster versions.

Section 5

Predictive Analysis (Classification)

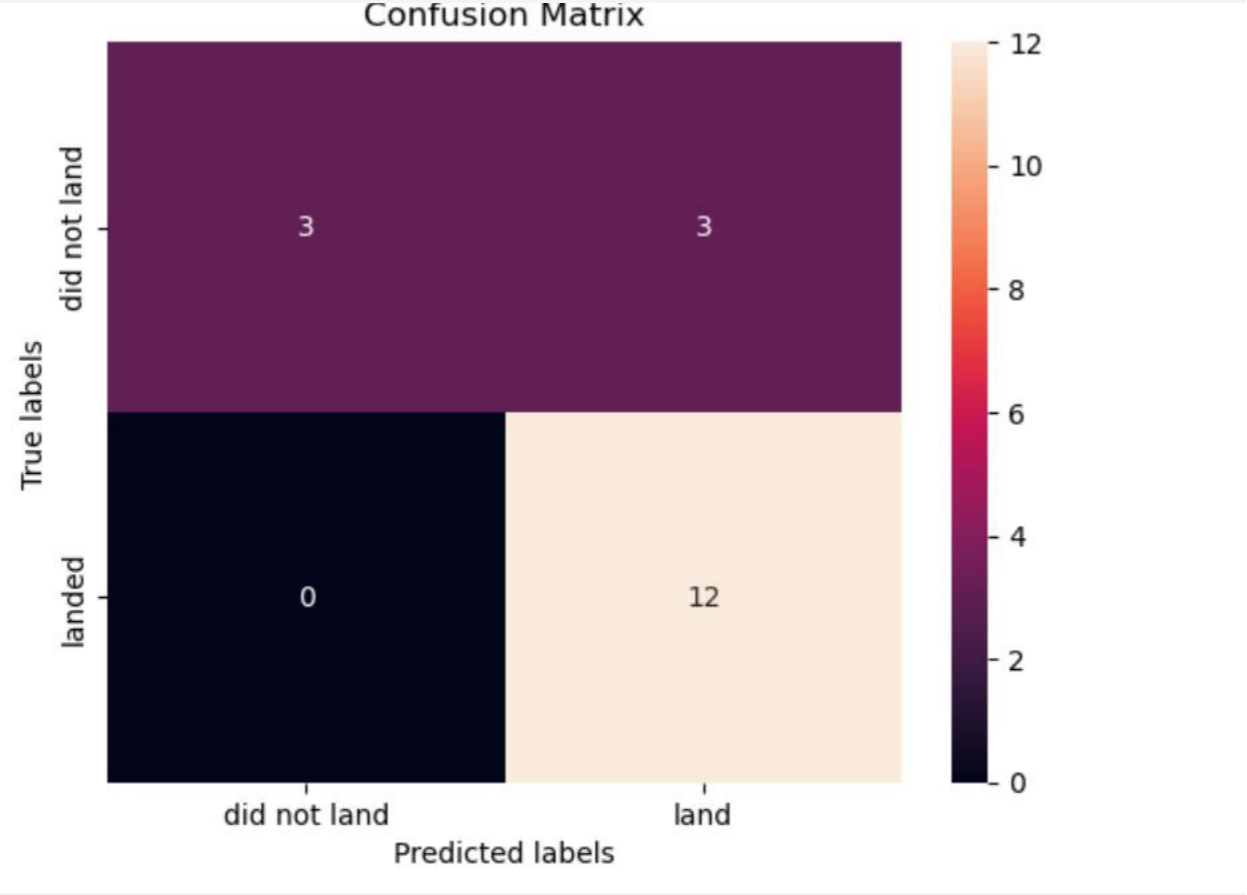
Classification Accuracy

- Log, SVM, and KNN all performed similarly. Decision Tree was the worst performing model.



Confusion Matrix

- This is the confusion matrix for the KNN classification model. It predicts launches that landed correctly. It incorrectly predicts some launches that did not land would land.



Conclusions

- Spacex has had a number of launches over the past decade+
- The success of these launches has improved over time
- In this report, we have shown characteristics that correlate with successful landings
- Hopefully by showing which characteristics contribute to successful launches, we can improve the proportion of successful launches even more.

Appendix

- Python notebooks used for this analysis: https://github.com/gregsasso/Final-Project
- Sample SQL query: %sql SELECT COUNT(*), landing_outcome FROM spacex GROUP BY landing_outcome ORDER BY COUNT(landing_outcome)
- Sample python code:

```
df_year = df.groupby(df['Year'], as_index=False).mean()

#df_orbit

sns.lineplot(x="Year", y="Class", data=df_year)

plt.xlabel("Year", fontsize=10)

plt.ylabel("Avg Class", fontsize=10)

plt.show()
```

Thank you!