IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Terry Chen
2022/3/22

Outline

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

Executive Summary

- With data wrangling, we can first filter out the data we don't need and do some feature engineer before we start to explore it.

- With EDA and Interactive analytics, we can get basic correlations between variables, better understanding of data and easier for choosing features.

- Predictive analysis results, most of the models have similar results, since the dataset is too small.

Introduction

- SpaceX's accomplishments include: Sending spacecraft 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.

- **What features determine the first stage will success (land) or not?**

Section 1

Methodology

Methodology

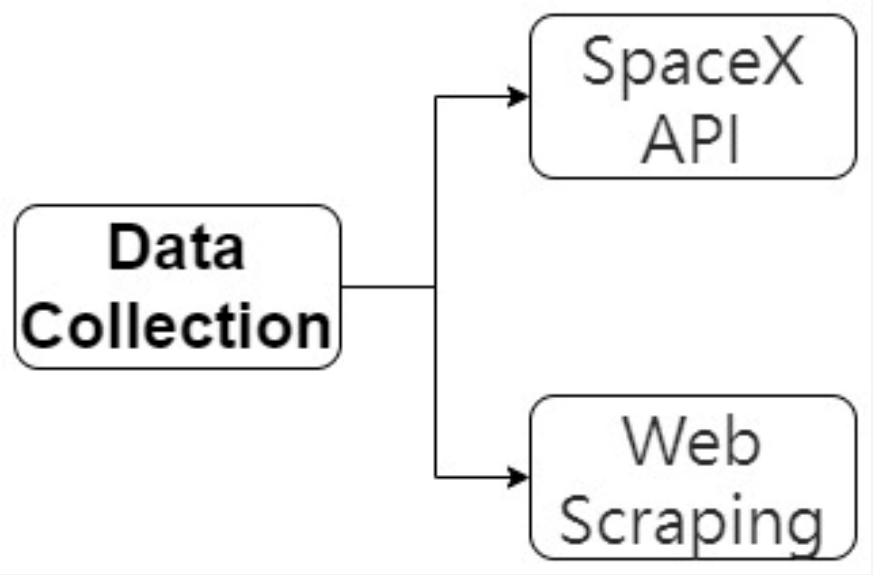
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

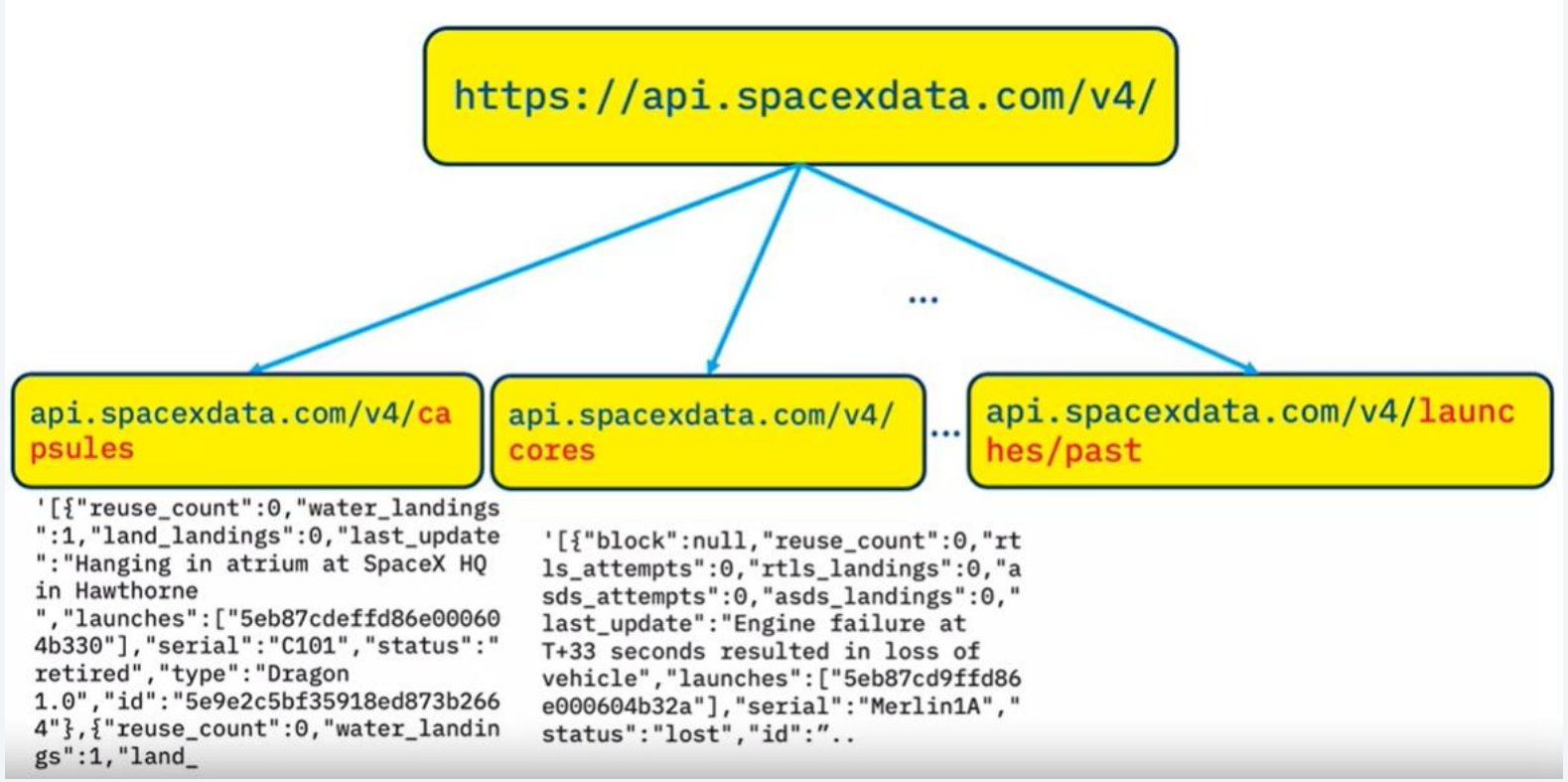
Data Collection

- We collect data through:

1. SpaceX API
2. Web Scraping



Data Collection – SpaceX API



- The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/, with different end points, for example: /capsules and /cores.
- We will be working with the endpoint api.spacexdata.com/v4/launches/past.

View on GitHub

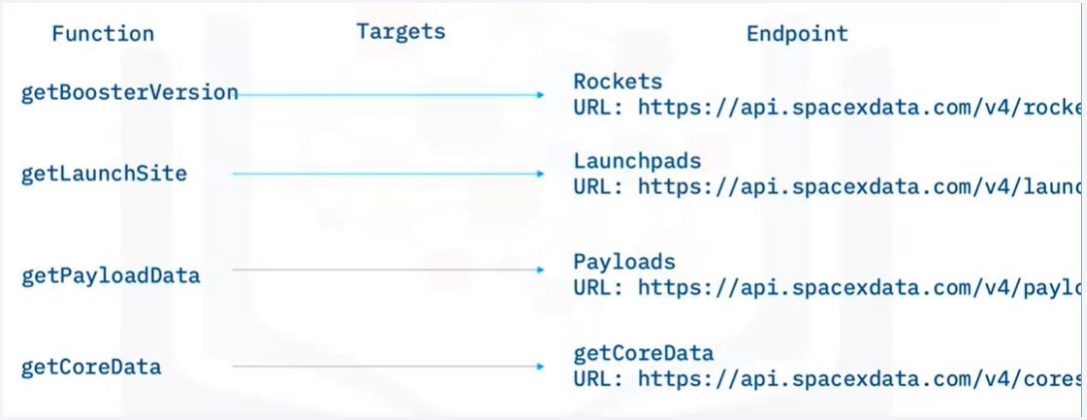
Data Collection - Scraping

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

- Using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records from Wiki pages.

View on GitHub

Data Wrangling



- Wrangling data using API
  - Using the API again targeting another endpoint to gather specific data for each ID number.
- Sampling data
  - Filtering/Sampling the data to remove Falcon 1 launches.
- Dealing with Nulls
  - Replacing the NULL values inside the PayloadMass column by mean.

View on GitHub

EDA with Data Visualization

1. Scatter plot
   - For visualizing correlation between variables.
2. Bar plot
   - For visualizing success rate by category.
3. Line plot
   - For visualizing success rate by time.

View on GitHub

EDA with SQL

- Using SQL queries to:
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster_versions which have carried the maximum payload mass.
  - Rank the count of landing outcomes.
  - Etc.

View on GitHub

Build an Interactive Map with Folium

- Using Interactive Map with Folium to:
  - Add each site's location on a map using site's latitude and longitude coordinates.
  - Mark the success/failed launches for each site on the map.

View on GitHub

Build a Dashboard with Plotly Dash

1. Pie chart
   - Visualizing the success rate.
2. Scatter plot
   - Visualizing the correlation between Payload Mass and success, with different Booster Version Category.
3. Interaction with Payload Mass and Launch Site
   - For easier to view results under different conditions.

View on GitHub

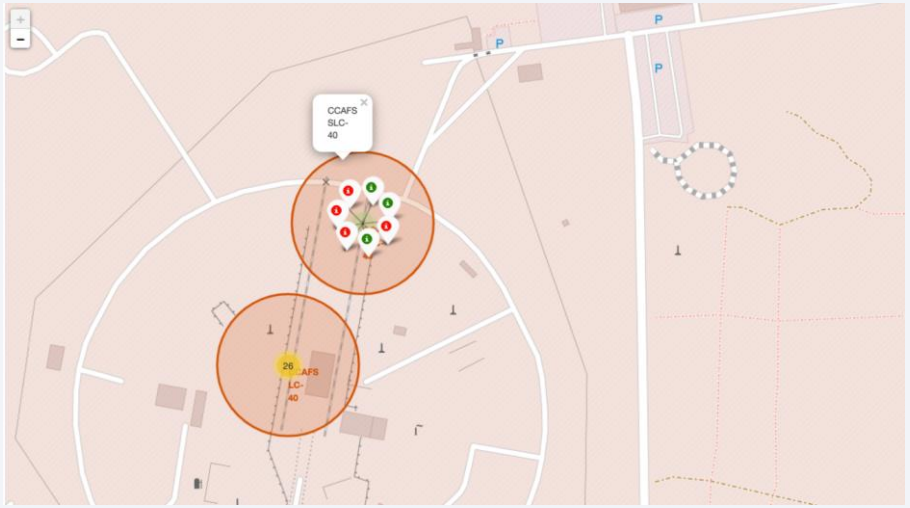
Predictive Analysis (Classification)

- Using Logistic regression, KNN, Decision tree and SVM.
- Doing GridsearchCV for hyperparameters tuning.
- Evaluating performance with accuracy.
- Viewing test set result by plotting confusion matrix.

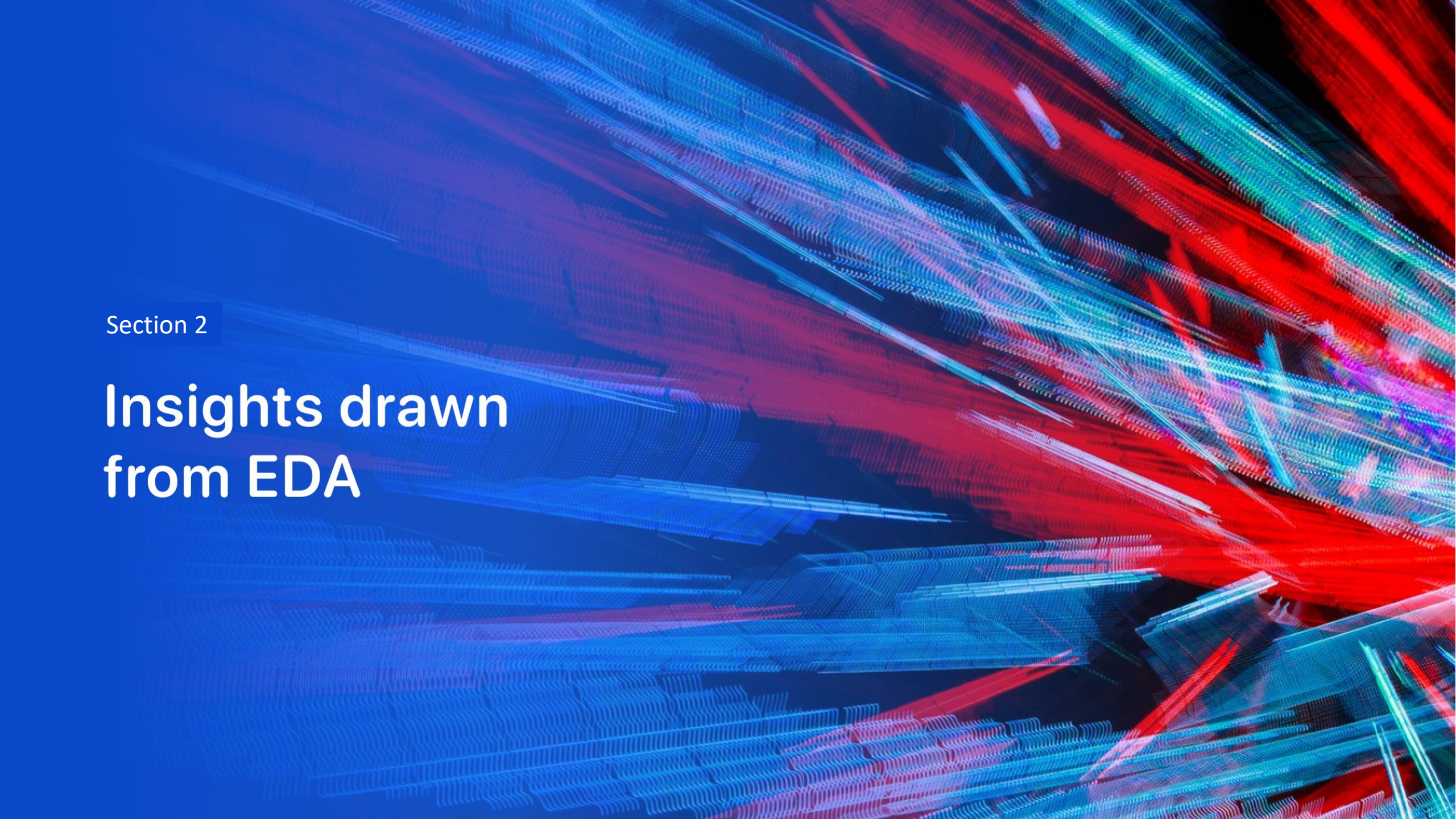
View on GitHub

Results

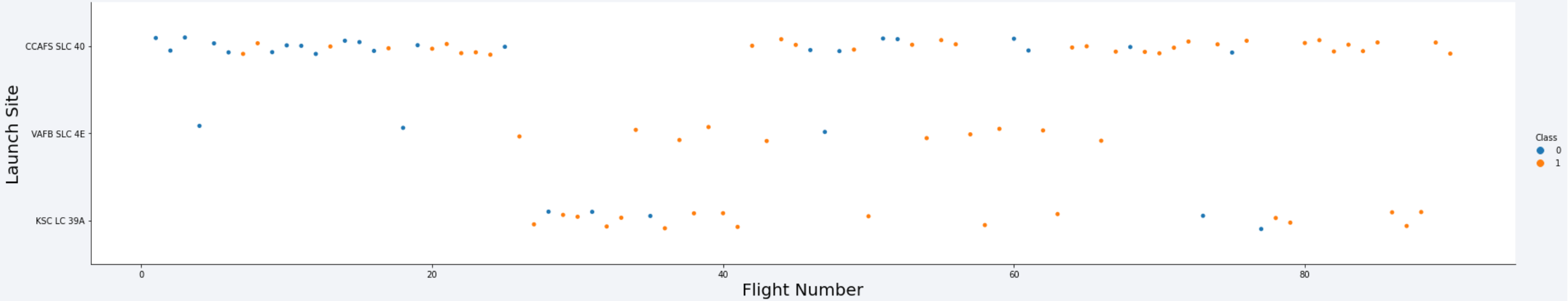
- EDA results:
  - Basic correlations between variables, better understanding for choosing features.
- Interactive analytics demo in screenshots:



- Predictive analysis results:
  - Most of the models have similar results, since the dataset is too small.

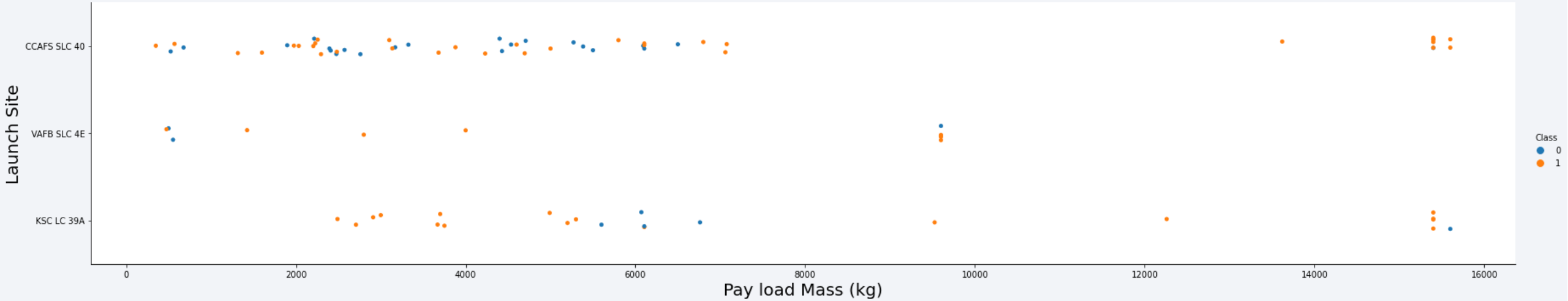

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Observing Flight Number Vs. Launch Site scatter point chart you will find for most of the events occur at the CCAFS-SLC-40 launchsite after Flight Number greater than 40.

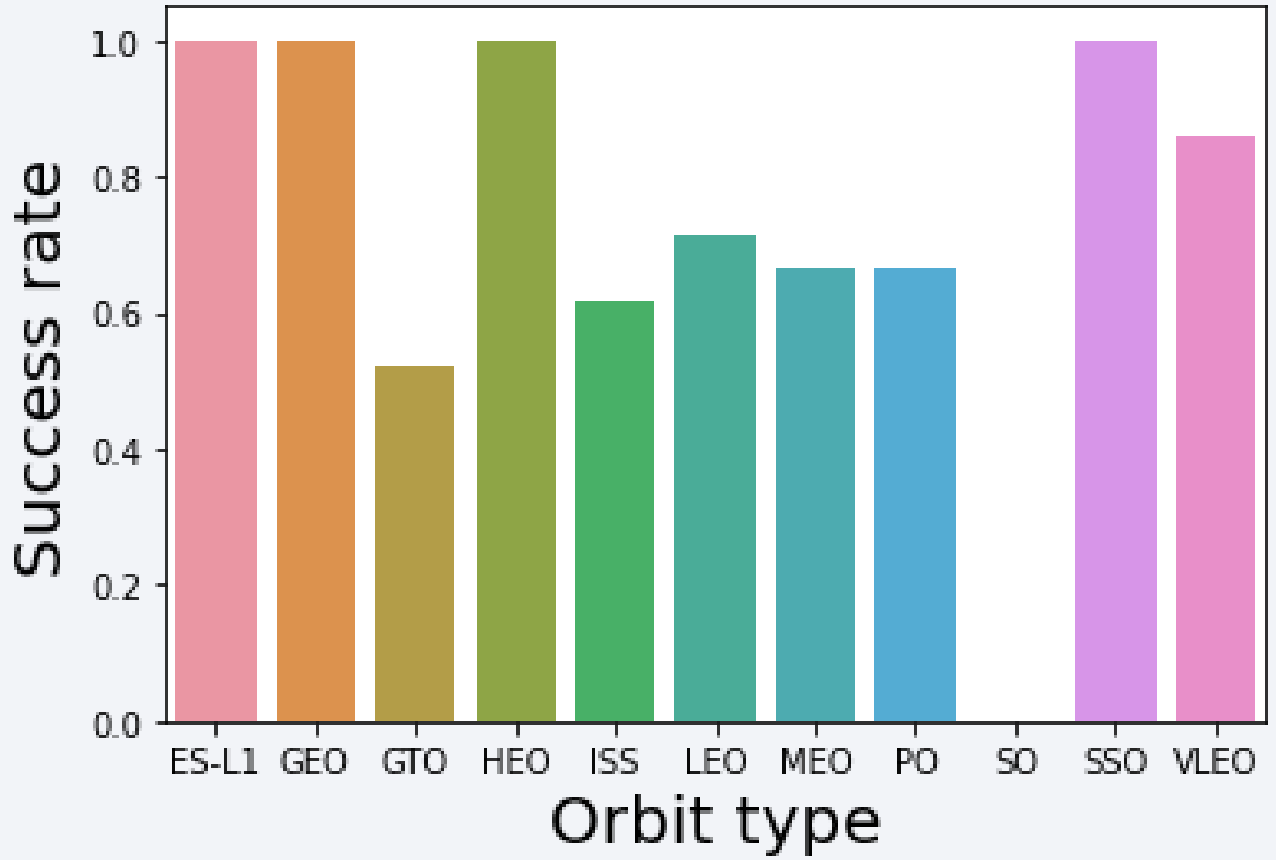
Payload vs. Launch Site



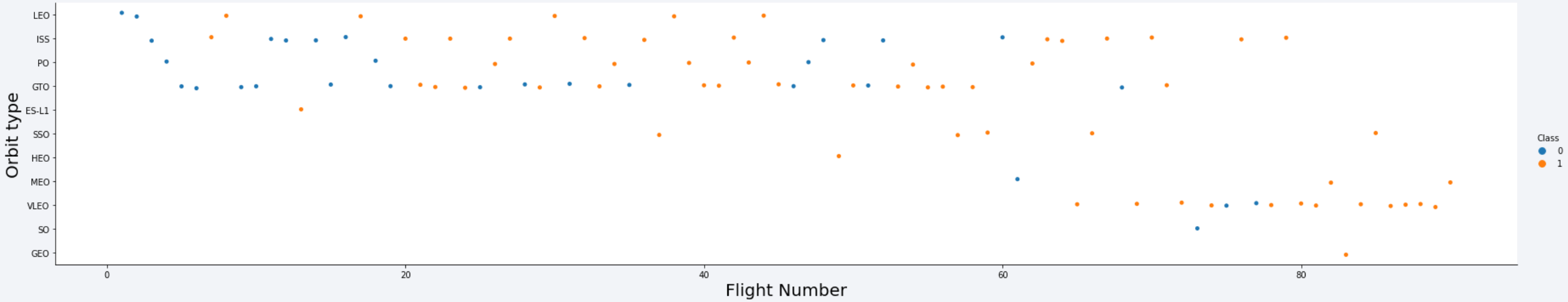
- Observing Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- SO has the success rate of 0%.
- ES-L1, GEO, HEO and SSO have the success rate of 100%.

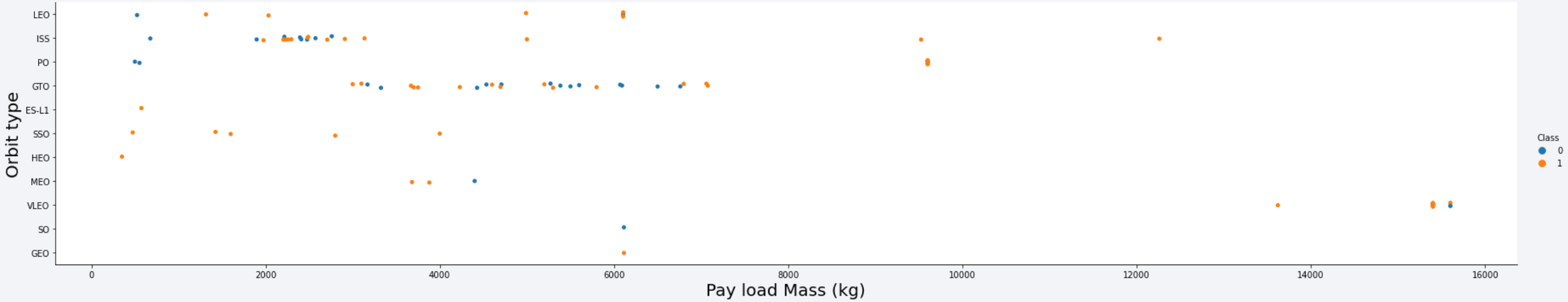


Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

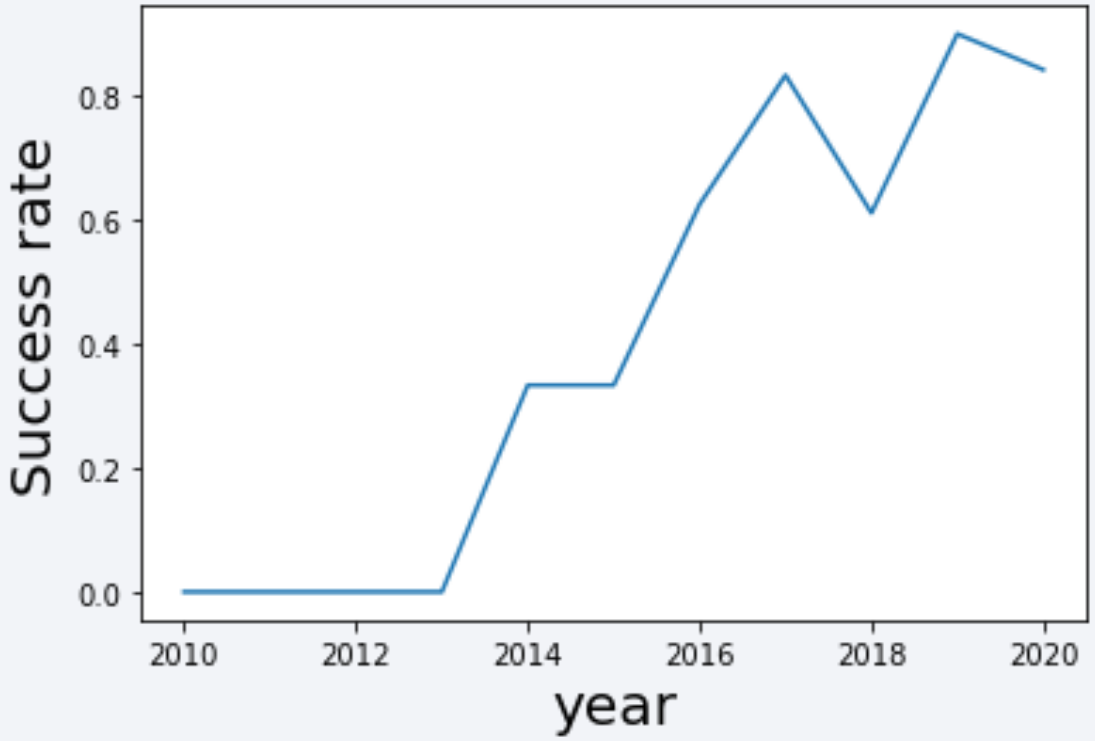
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- The sucess rate since 2013 kept increasing till 2020.
- A drop in 2018.



All Launch Site Names

- There are 4 unique names for launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

- Perhap the CCAFS LC-40 is typo of CCAFS SLC-40, or the old name of it.

Launch Site Names Begin with 'CCA'

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Landing outcomes were all fail with the first five records.

Payload Mass Calculation

- Total Payload Mass
  - Total payload carried by boosters from NASA is 45596.
- Average Payload Mass by F9 v1.1
  - Average payload mass carried by booster version F9 v1.1 is 2928.

First Successful Ground Landing Date

- The date first successful landing outcome on ground pad is 2015-12-22.
- 5 years after the first attempt at 2010-04-06.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:
  - Failure (in flight): 1
  - Success: 99
  - Success (payload status unclear): 1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass with 15600 kg:
  - 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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - F9 v1.1 B1012, CCAFS LC-40
  - F9 v1.1 B1015, CCAFS LC-40

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

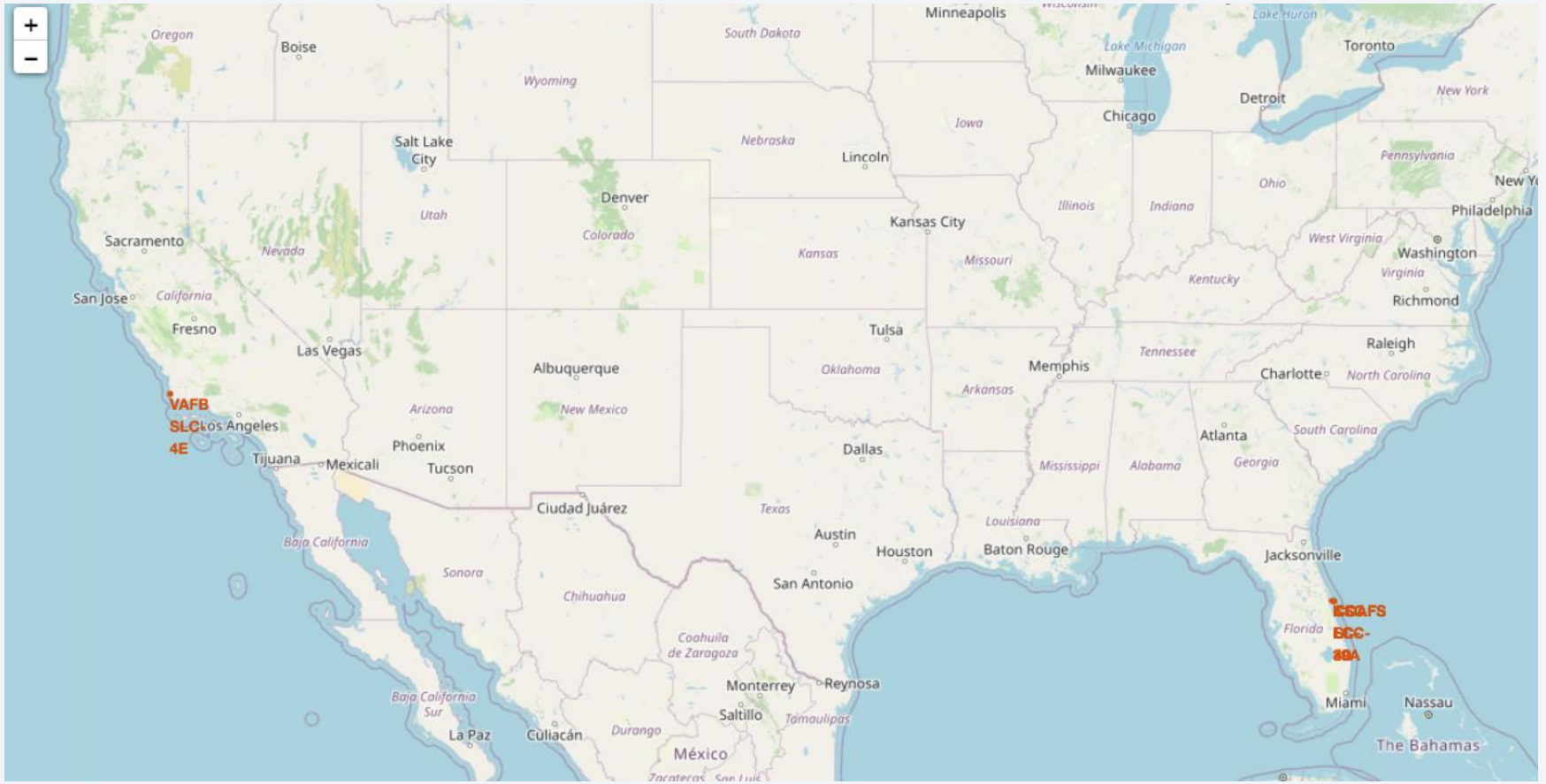
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing _Outcome	event_count	RANK
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	4
Success (ground pad)	3	4
Failure (parachute)	2	6
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	8

Section 3

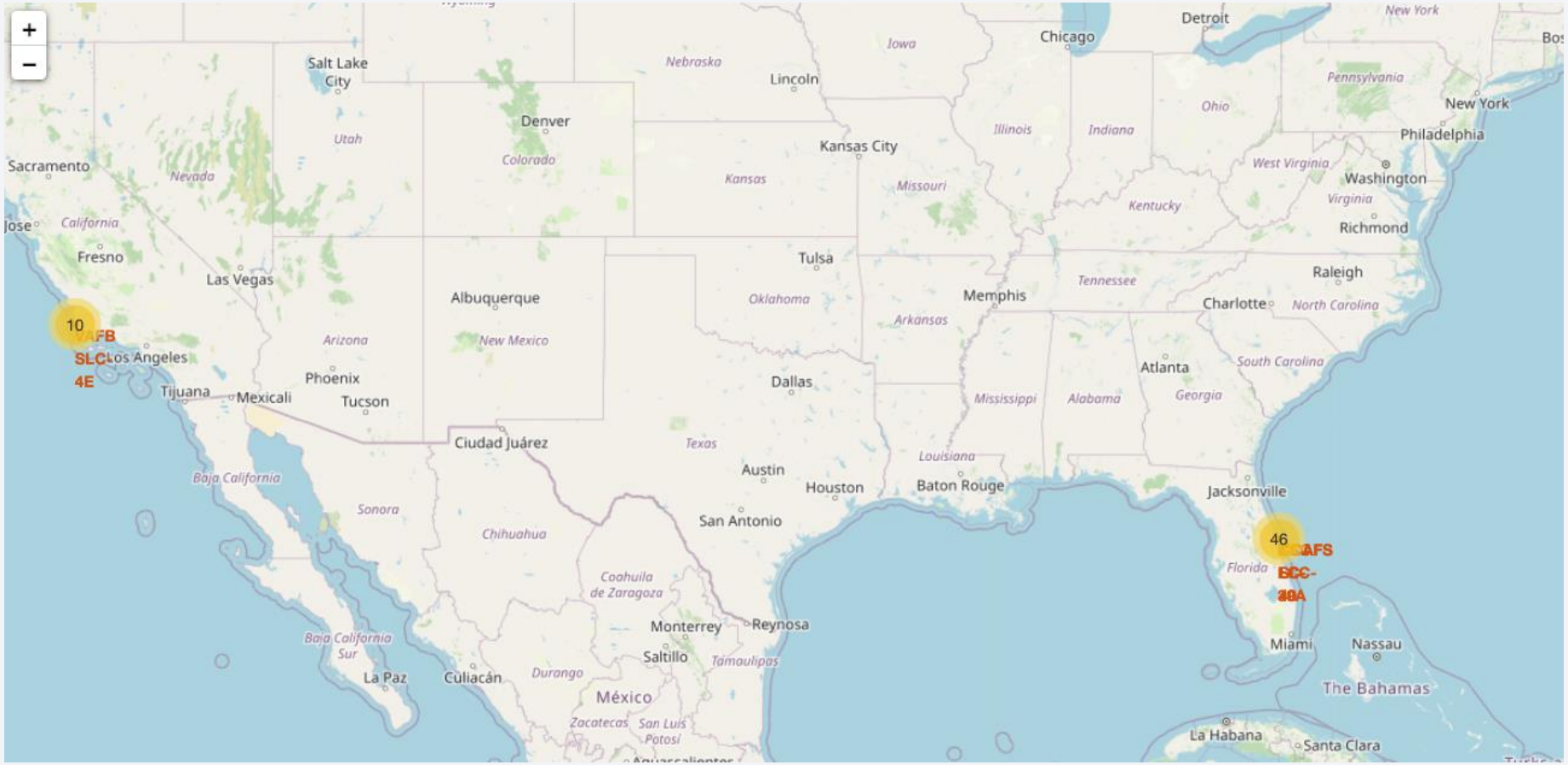
Launch Sites Proximities Analysis

Marking all launch sites

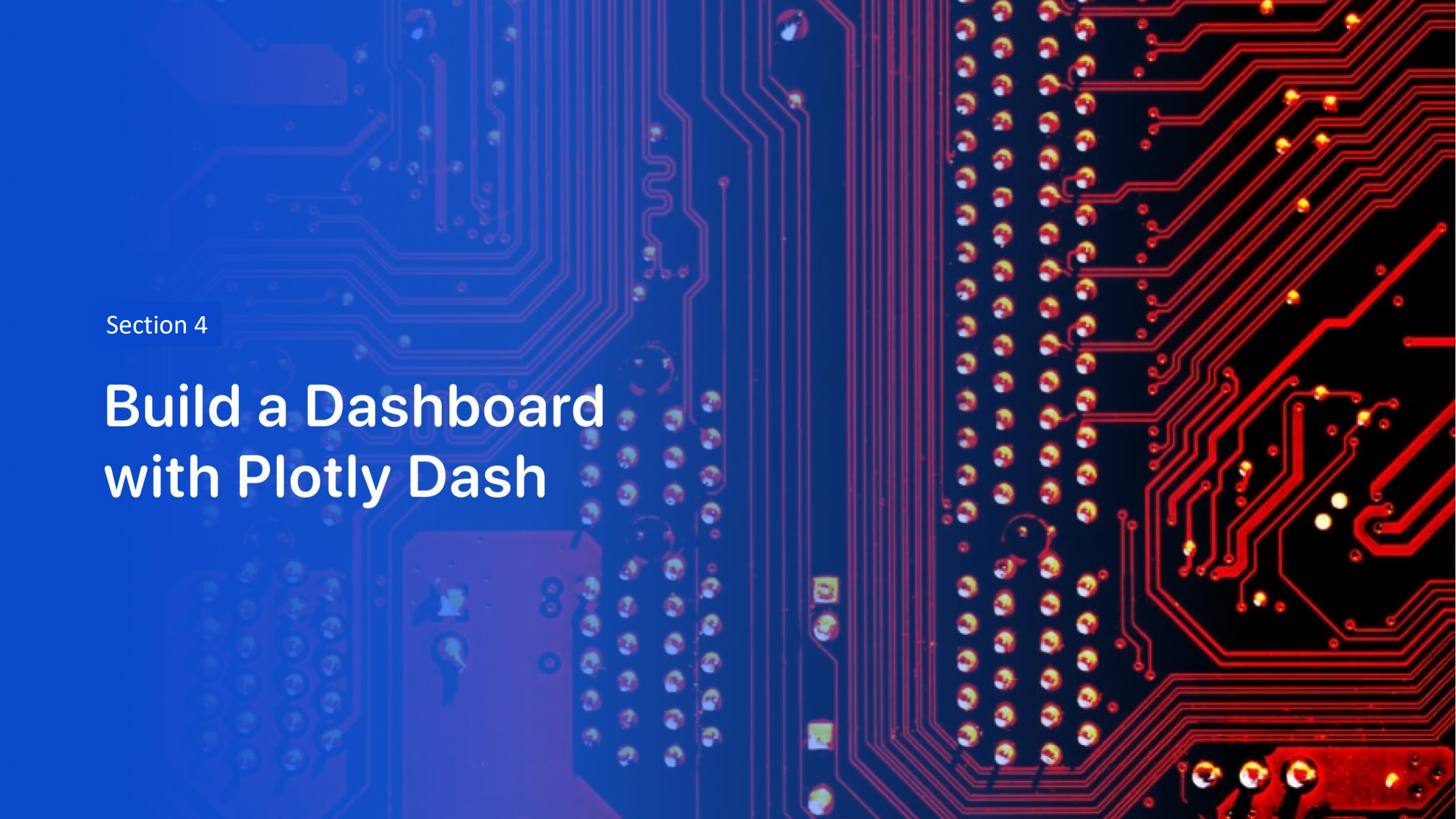


- All launch sites in very close proximity to the coast.

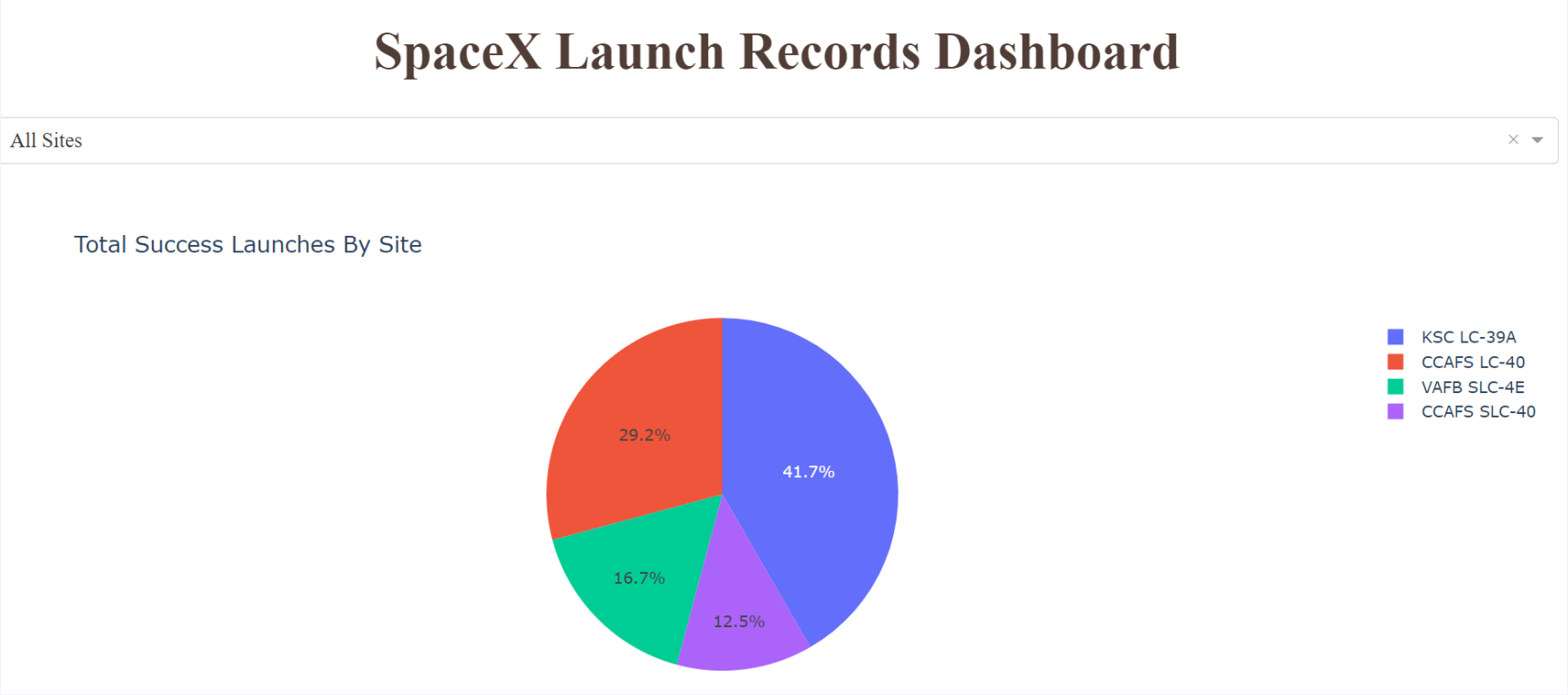
Marking the success/failed launches for each site



- Most of the launches occur at Florida.

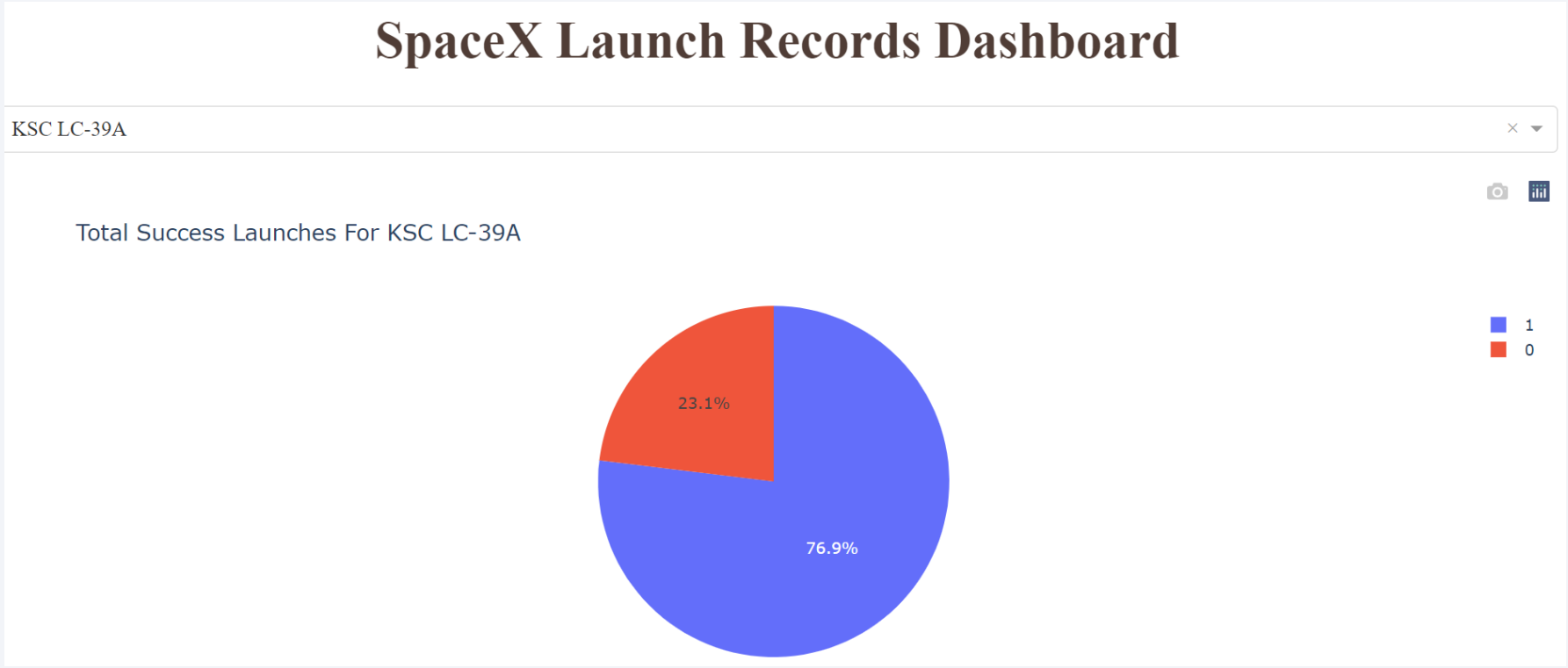

Section 4
Build a Dashboard with Plotly Dash

Pie chart for all sites



- KSC-LC-39A has the most successful launch (41.7%).
- Combining CCAFS-LC-40 and CCAFS-SLC-40 have the same successful launch.

Pie chart for KSC-LC-39A



- KSC-LC-39A has 76.9% success rate.

Scatter plot for KSC-LC-39A

- All of the booster version are FT for Payload greater than 5000.
- All of the outcome fails for Payload greater than 5500, success for Payload less than 5500.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Model	Train Accuracy	Test Accuracy
KNN	0.848	0.833
SVM	0.848	0.833
Decision tree	0.889	0.778
Logistic regression	0.846	0.833

Confusion Matrix

- Both KNN, Logistic regression and SVM got the same confusion matrix.
- All of the models suffer from FP.

Conclusions

- Decision tree is overfitted since high train accuracy with relative low test accuracy.
- All of the model got similar results, since , since the dataset is too small.
- In such small dataset, how the training set and test set split highly affect the final result.

Thank you!