IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Spence Cox
7/19/2023

Outline

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

Executive Summary

Methodologies

- Webscraping
- Data Collection API
- Explanatory Data Analysis with SQL
- Machine Learning Models
- Dashboard Creation

Results Summary

- Webscraping and data querying is crucial to understanding the SpaceX company
- Many factors affect the outcome of Falcon 9 launches
- Using a Support Vector Machine model, we can accurately predict the outcome of upcoming SpaceX launches (0.889% accuracy)
- By using the methodology described in this presentation, we can predict whether SpaceX will reuse their rocket. This is of great help to our competing brand, SpaceY

Introduction

Space Y

- New company competing with SpaceX by Allon Musk
- We must determine the price of each launch
- Determine if SpaceX will reuse their rocket
- Train Machine learning model to predict if SpaceX will reuse their first stage rocket

Section 1

Methodology

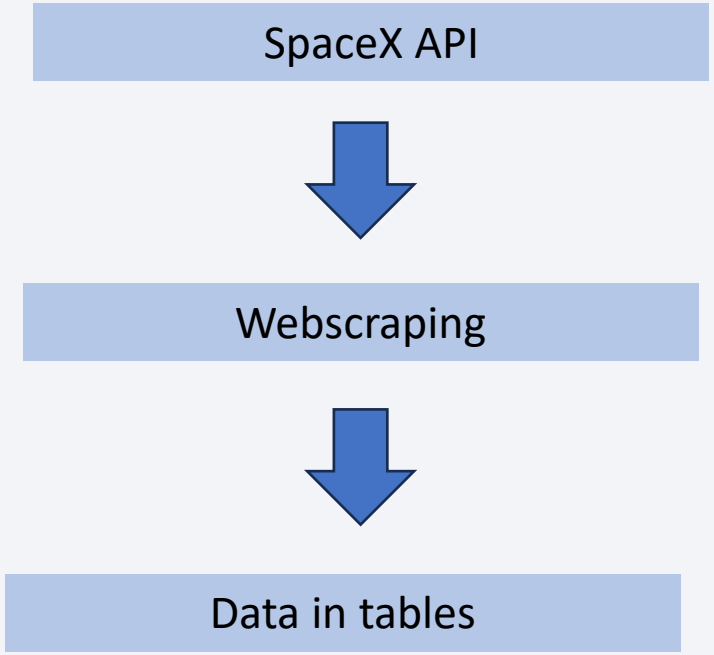
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through a SpaceX API and with webscraping a Wikipedia article
- Perform data wrangling
  - Python code was used to organize the data into a Pandas dataframe
- 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

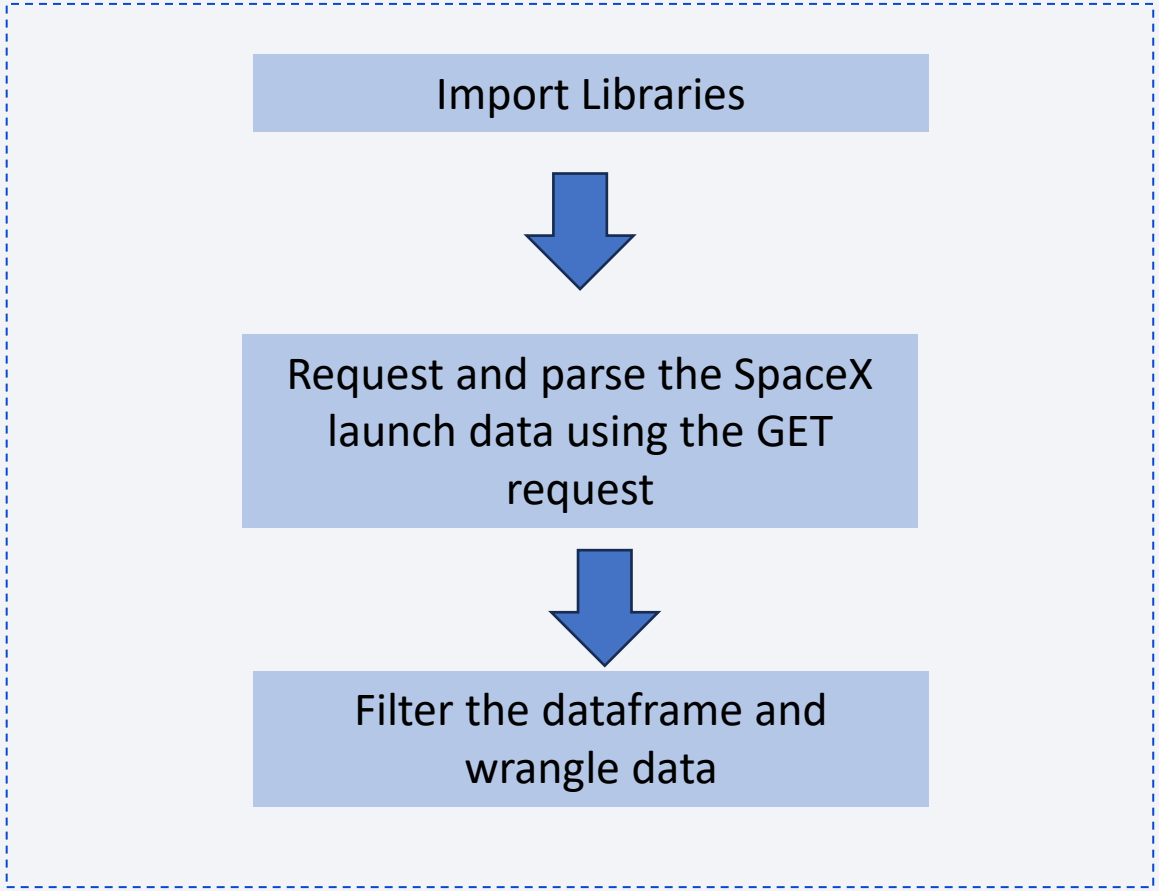
Data Collection

- Data was collected through the **SpaceX Data API** and through **webscraping Wikipedia articles**
- The Requests and Beautiful Soup libraries were used extensively
- These methods brought the data into a table for analysis



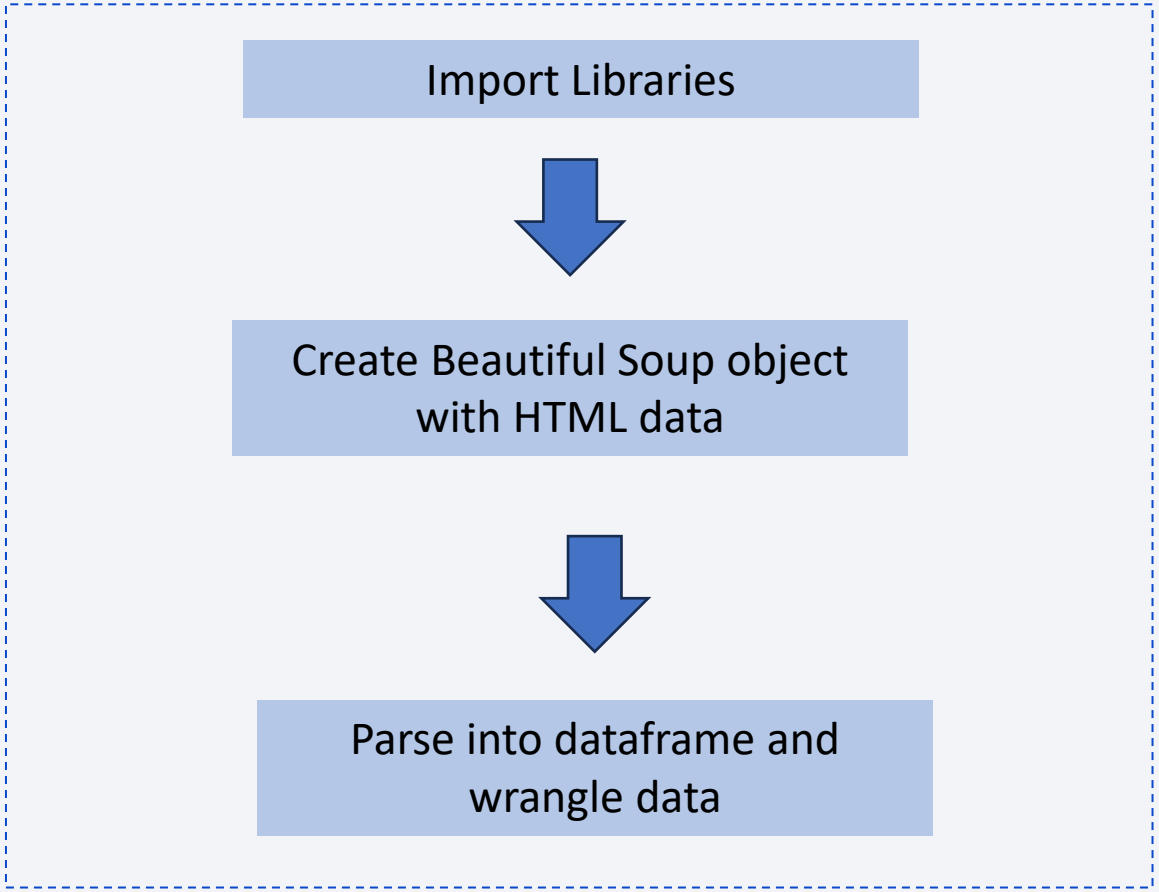
Data Collection – SpaceX API

- Get requests were used to parse launch data from the API
- Data Collection Notebook (github)



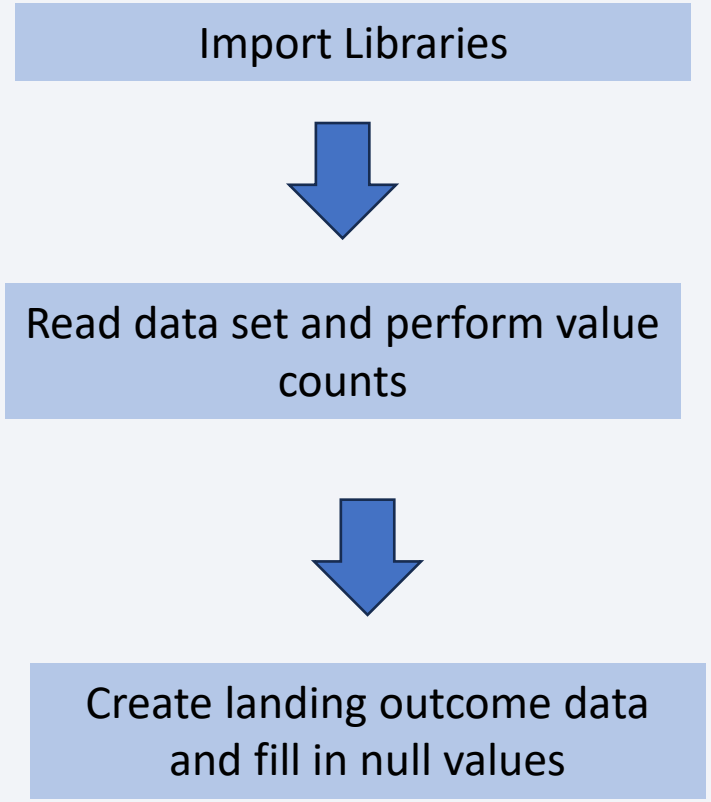
Data Collection - Scraping

- **Request the Falcon9 Launch Wiki page from its URL**
- **Extract all column/variable names from the HTML table header**
- **Create a data frame by parsing the launch HTML tables**
- Data Scraping Notebook (github)



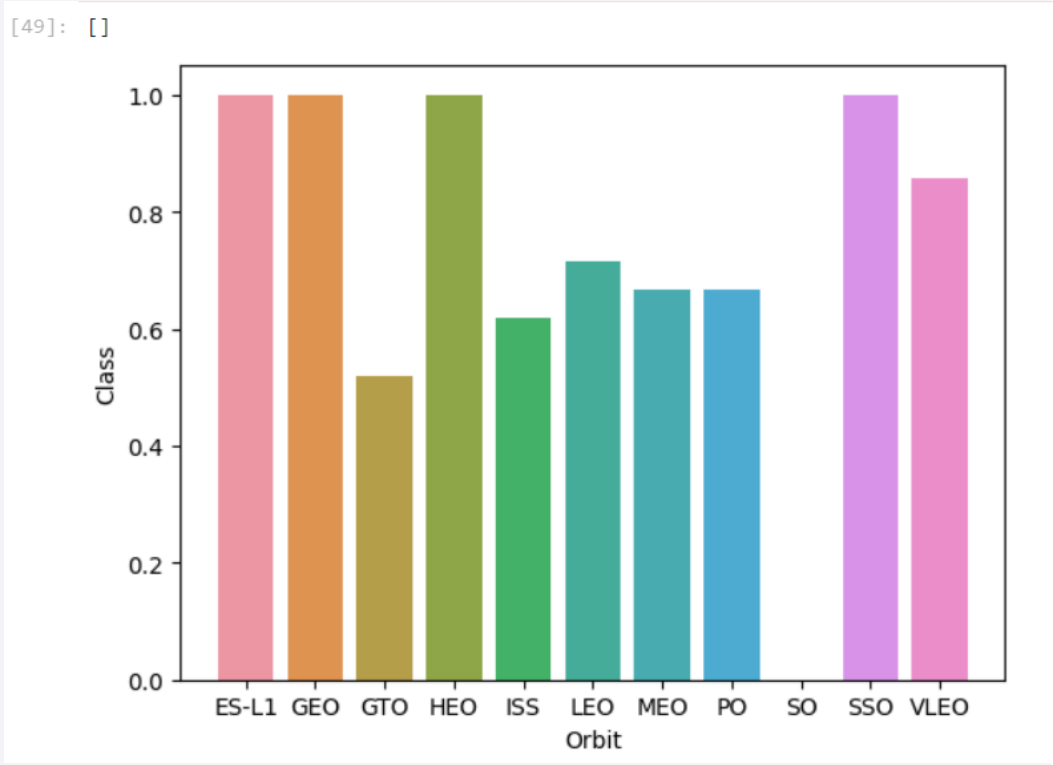
Data Wrangling

- Analyze data from collection
- Calculate number and occurrence of orbits and create landing outcome column
- Save data
- Data Wrangling Notebook (github)



EDA with Data Visualization

- A variety of graphs were used for data visualization
- This included bar graphs, scatterplots, and scatterpoint graphs
- Each graph had a specific use case and explained the data in a specific way.
- To the right is a bar graph showing success rate for different orbit destinations
- [Github notebook for visualization]

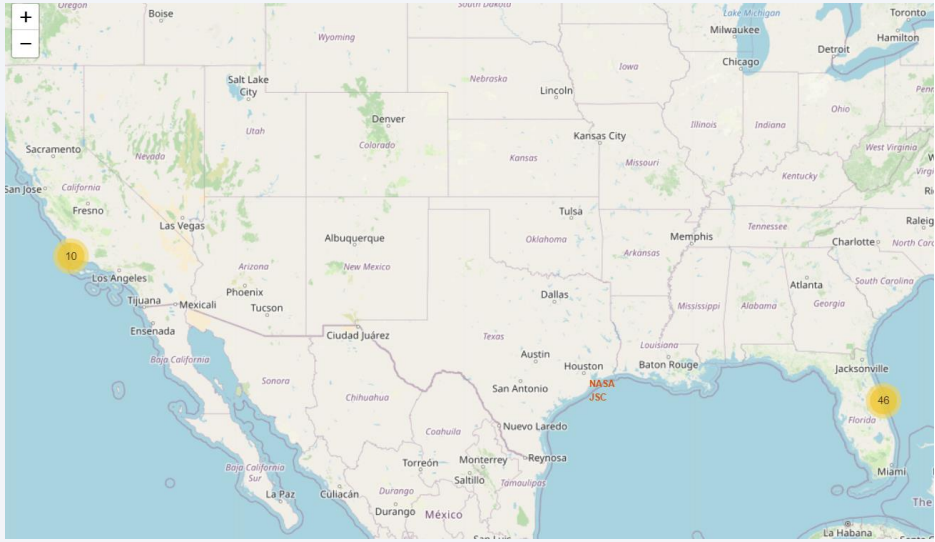


EDA with SQL

- Github link
- Many queries were used to understand the data, including
  - Max payload
  - Outcomes of 2015 missions
  - Value counts
  - Mission outcomes

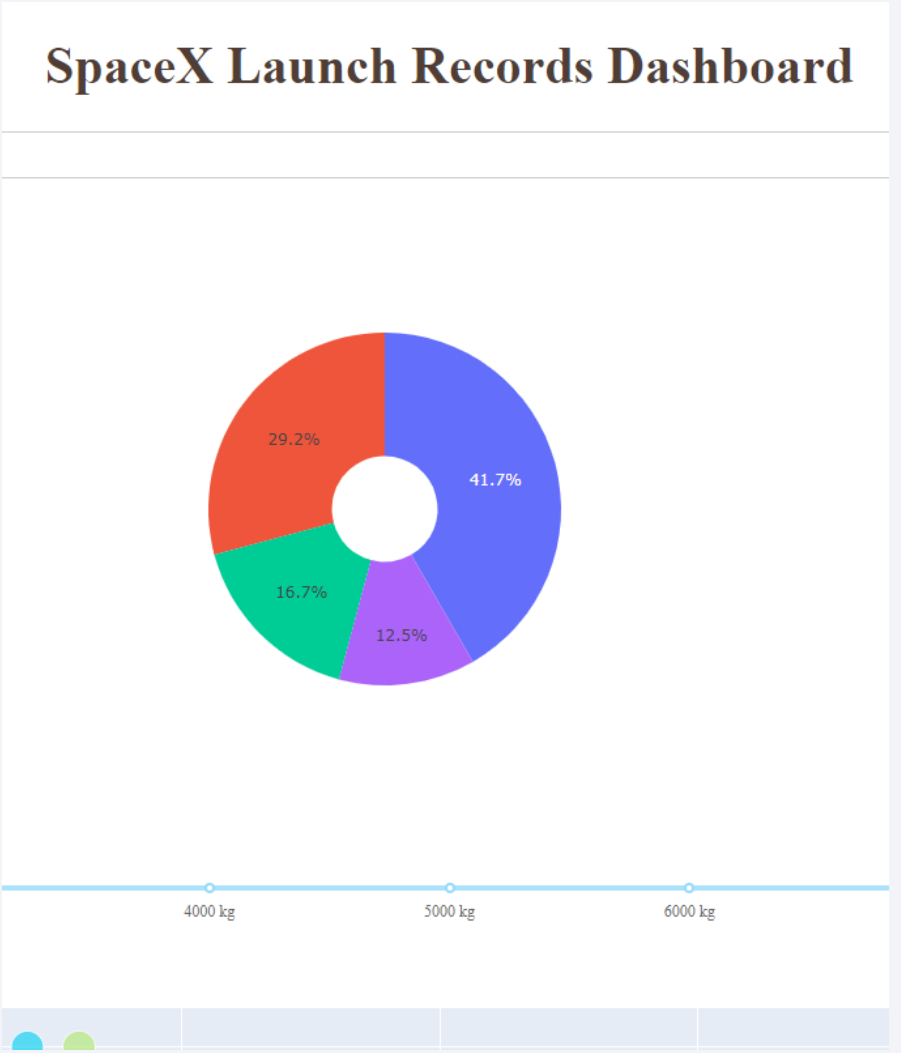
Build an Interactive Map with Folium

- A map point was added for each distinct launch location (four total). Also, each successful and failure launch was added as a map cluster to the map locations to show which locations had the highest success rate.
- Lastly, lines and distances to the nearest landmarks, like highways, coastlines, and railroads were added to show the differences between different launch locations
- Github link



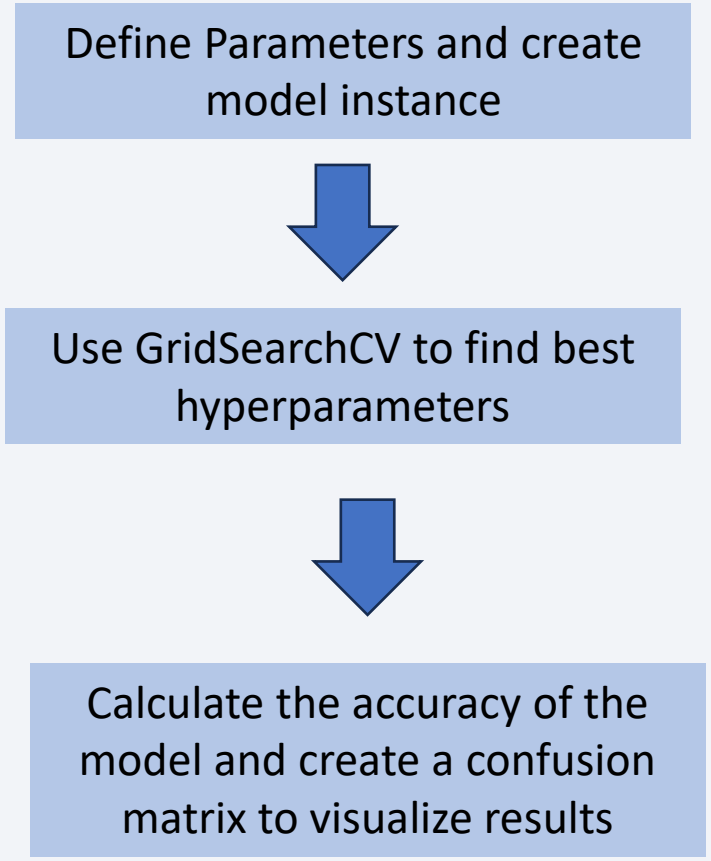
Build a Dashboard with Plotly Dash

- The dashboard contains dynamic pie charts and scatterplots that detail the success rates of launch sites and payload masses
- These plots are important because they allow insight to be drawn about what factors most affect the success of a launch site
- Github Link



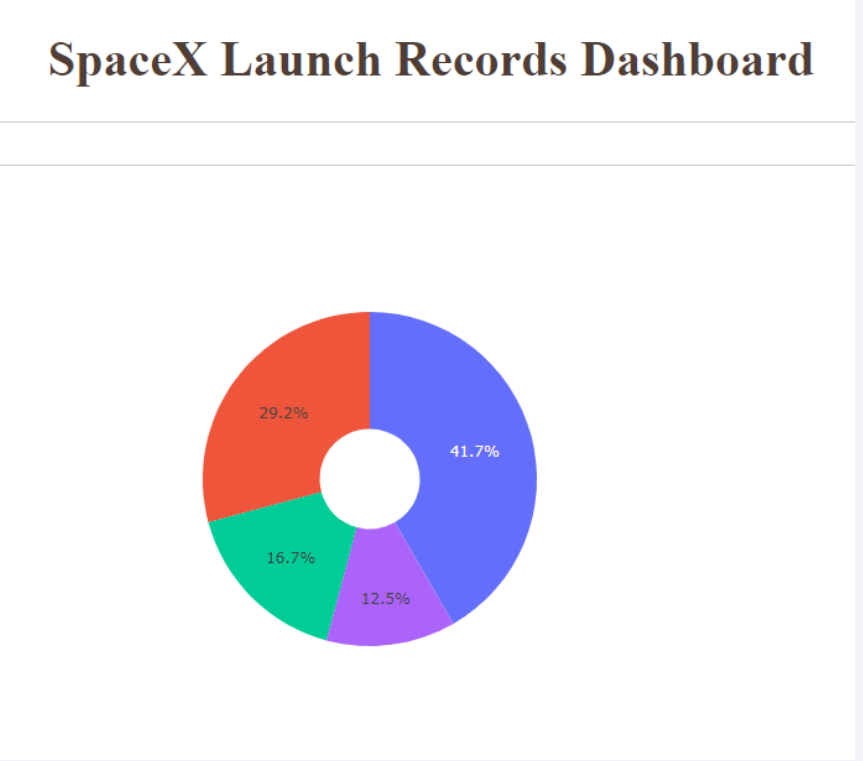
Predictive Analysis (Classification)

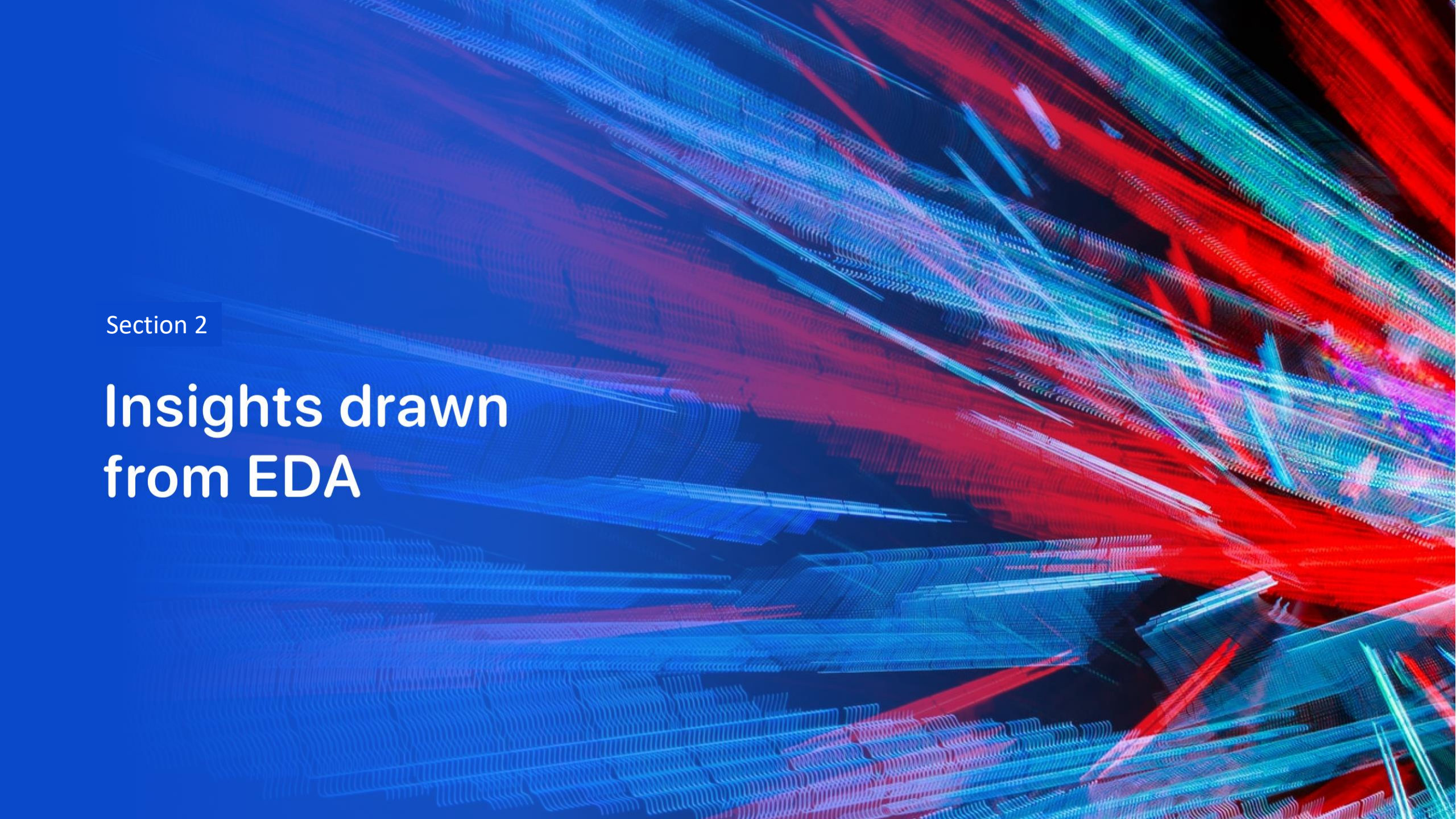
- To find the best classification machine learning model, I trained and tested 4 different models: Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Classification Tree
- The Support Vector Machine model consistently performed the best, with the highest testing accuracy of 0.889.
- Close seconds were the K-Nearest Neighbors model and Logistic Regression Model
- Github Link



Results

- Our exploratory data analysis revealed that there are four distinct launch sites, all with different attributes
- There are also different payload masses and booster rockets used
- To the right is a screenshot of the dashboard
- The predictive analysis showed that a support vector machine model is the best for predicting the outcome of SpaceX launches

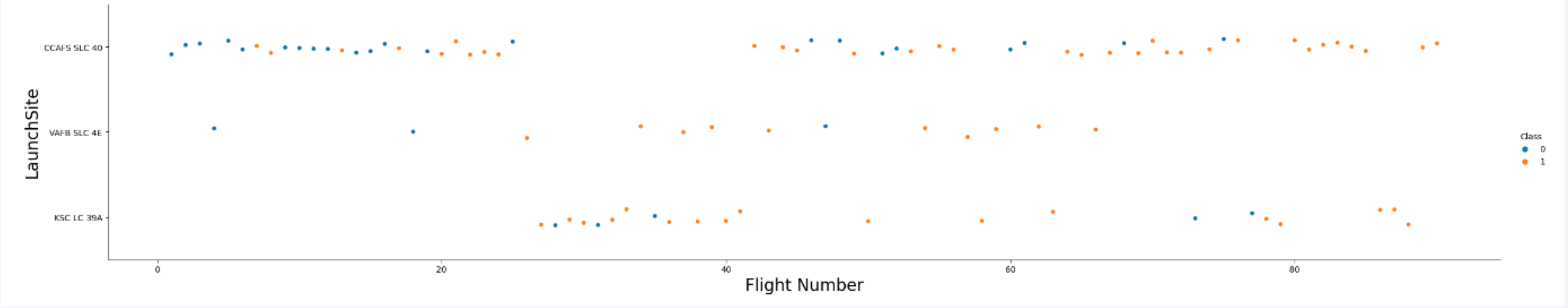



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Scatterplot of Flight Number vs. Launch Site

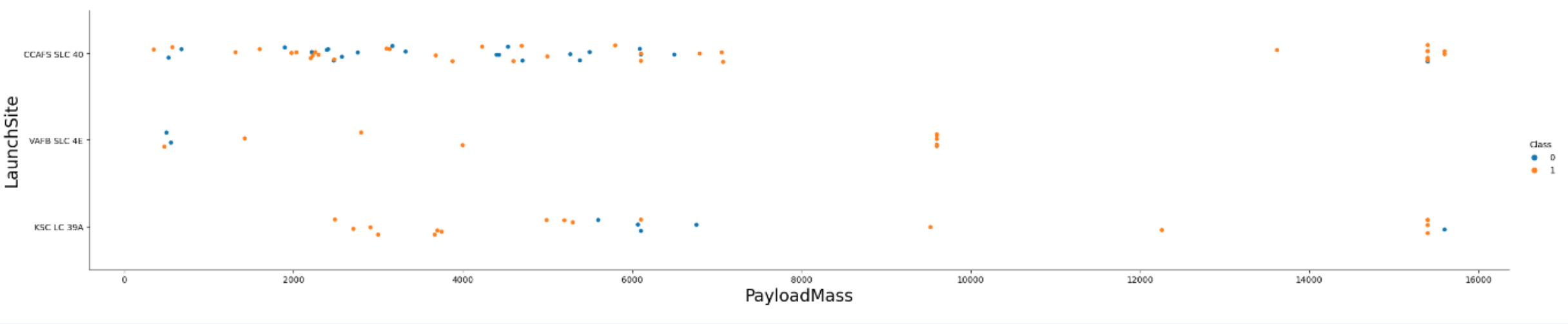
- This shows which flight numbers took off from which launch sites



Payload vs. Launch Site

- Scatterplot of Payload vs. Launch Site

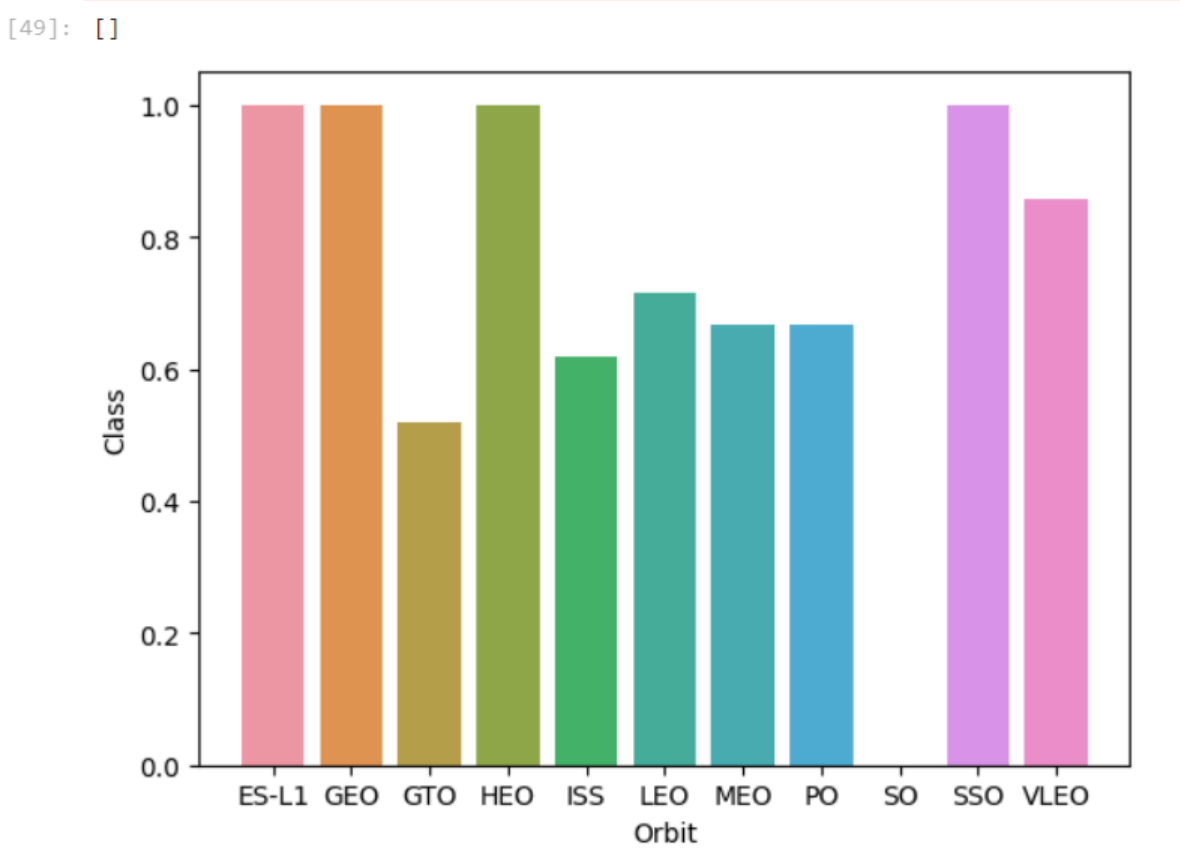
- This shows the relationship between launchsite and payload mass, where we can see that CCA launch site has mostly high payload launches.



Success Rate vs. Orbit Type

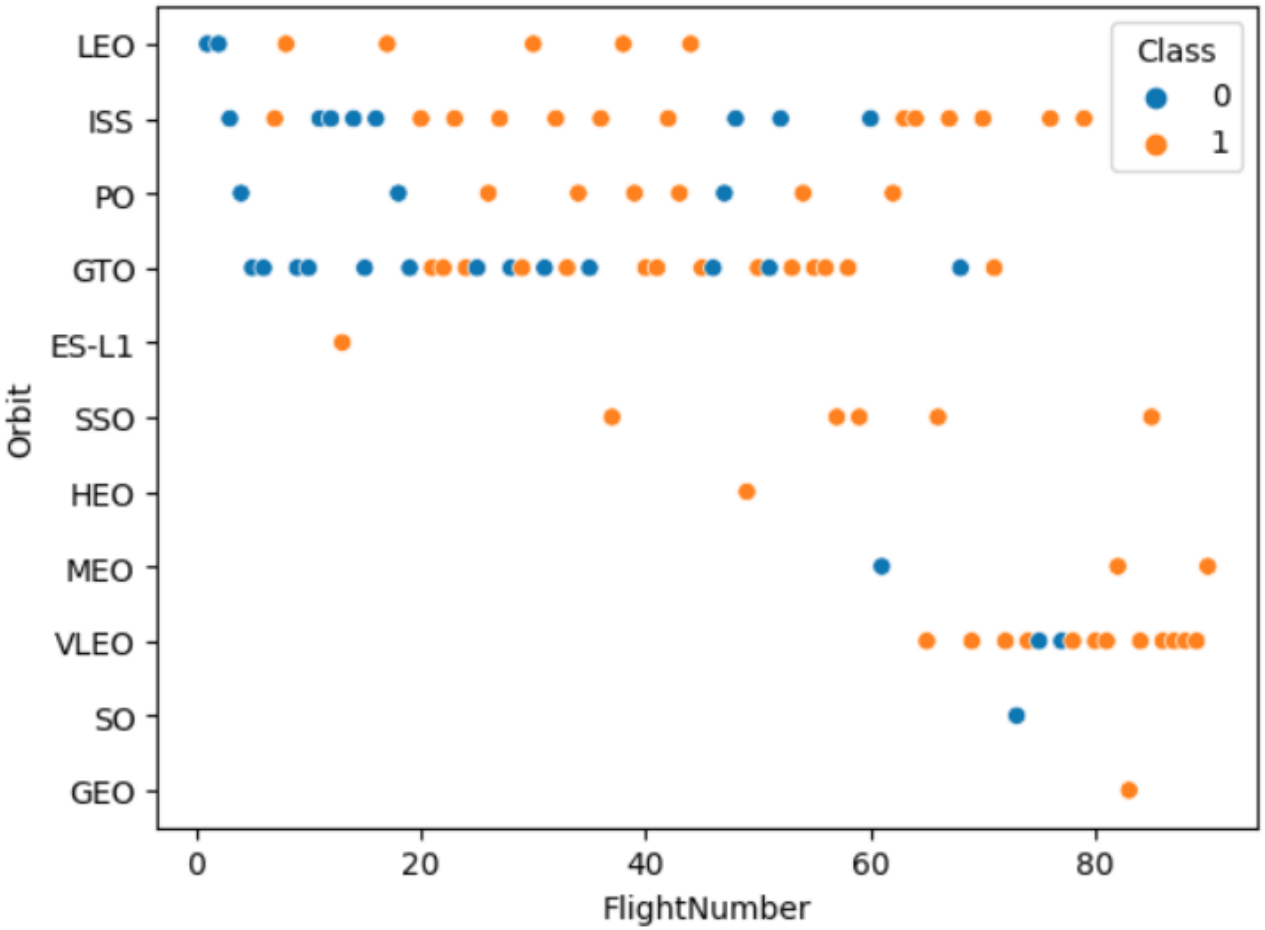
- bar chart for the success rate of each orbit type

- Here we can see that some orbits have higher success rates, like GEO and VLEO, and some have lower, such as GTO.



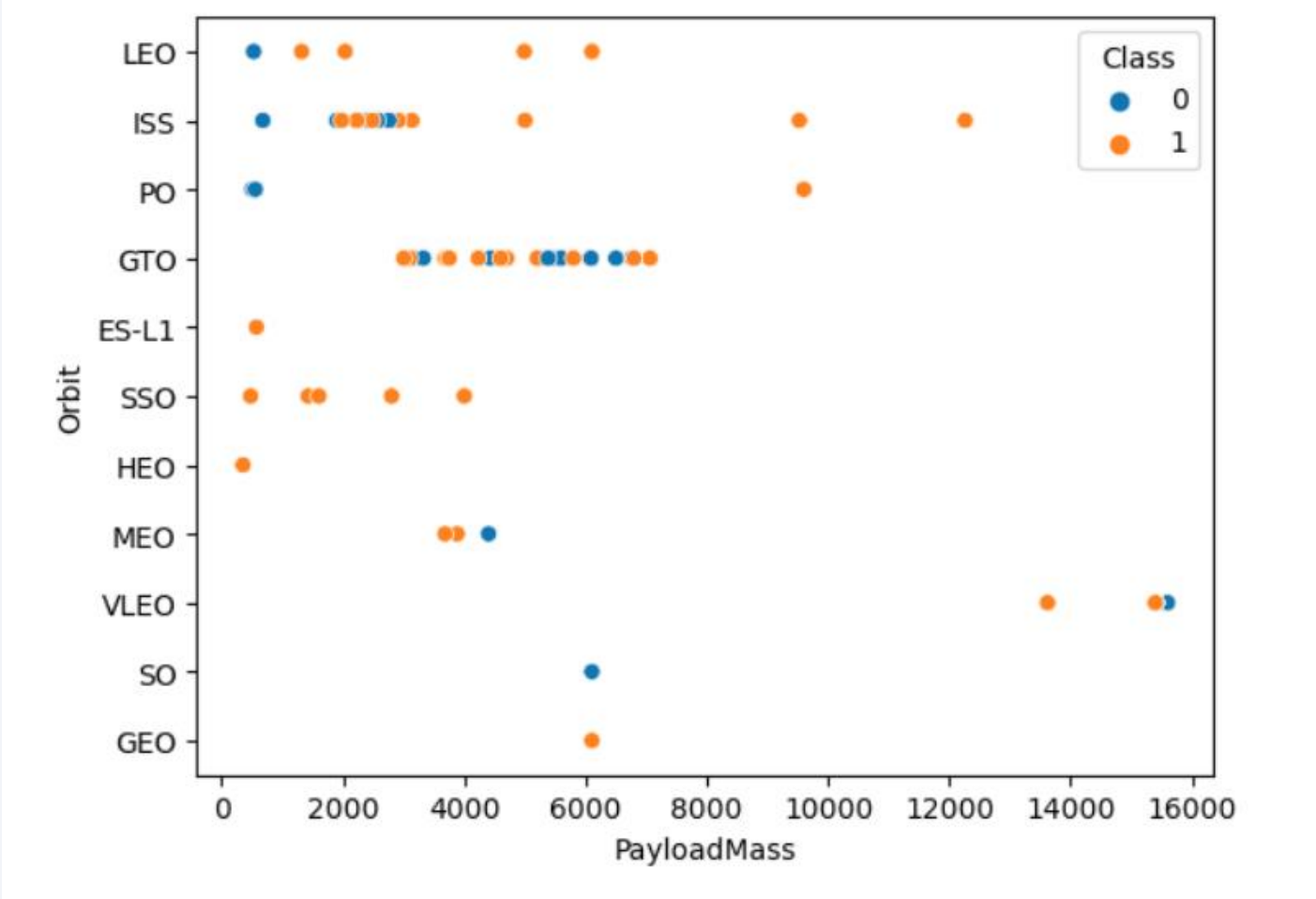
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type, where hue is success
- There have been a lot of attempts at GTO orbit, and a lot of them have failed. Lately, more VLEO flights have been conducted.



Payload vs. Orbit Type

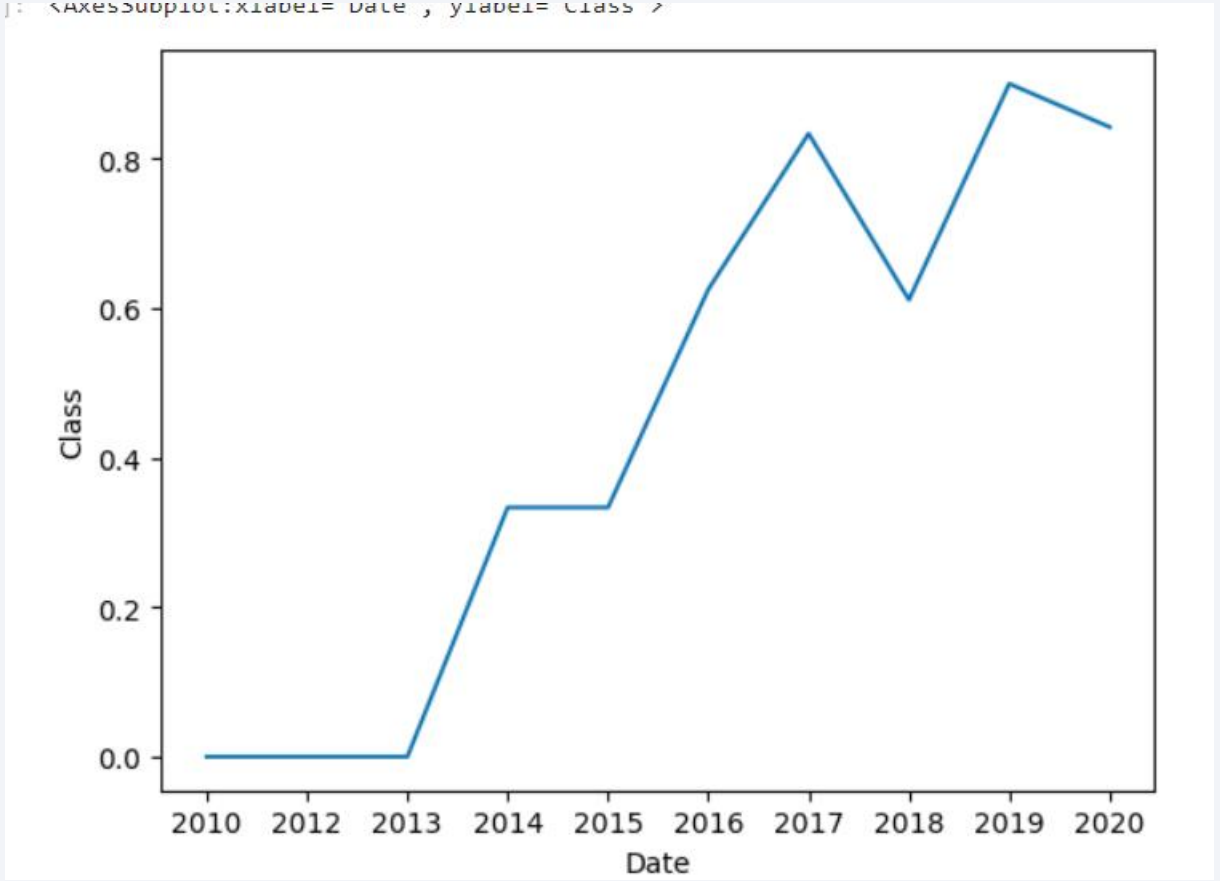
- scatter point of payload vs. orbit type and hue as success
- ISS orbits have a very tight range of mass, while GTO is more spread out. Each orbit has different mass specs.



Launch Success Yearly Trend

- line chart of yearly average success rate

- The average success rate increases over time, meaning SpaceX is getting better at reusing its rockets.

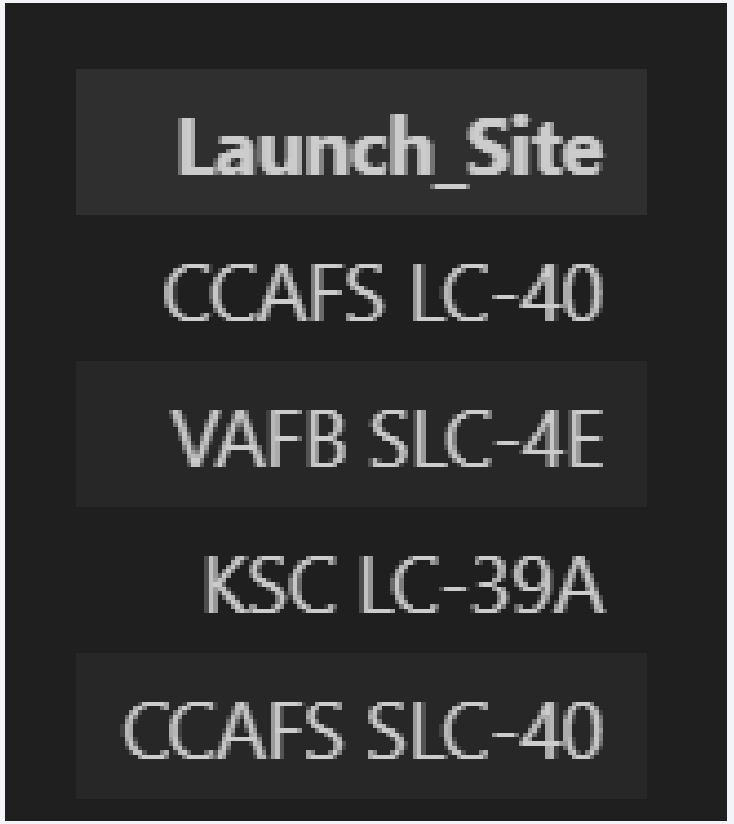


All Launch Site Names

Query:

- select distinct "Launch_Site" from SPACEXTBL

- This displays all the launch sites for the Falcon 9



Launch Site Names Begin with 'CCA'

- Query:

SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

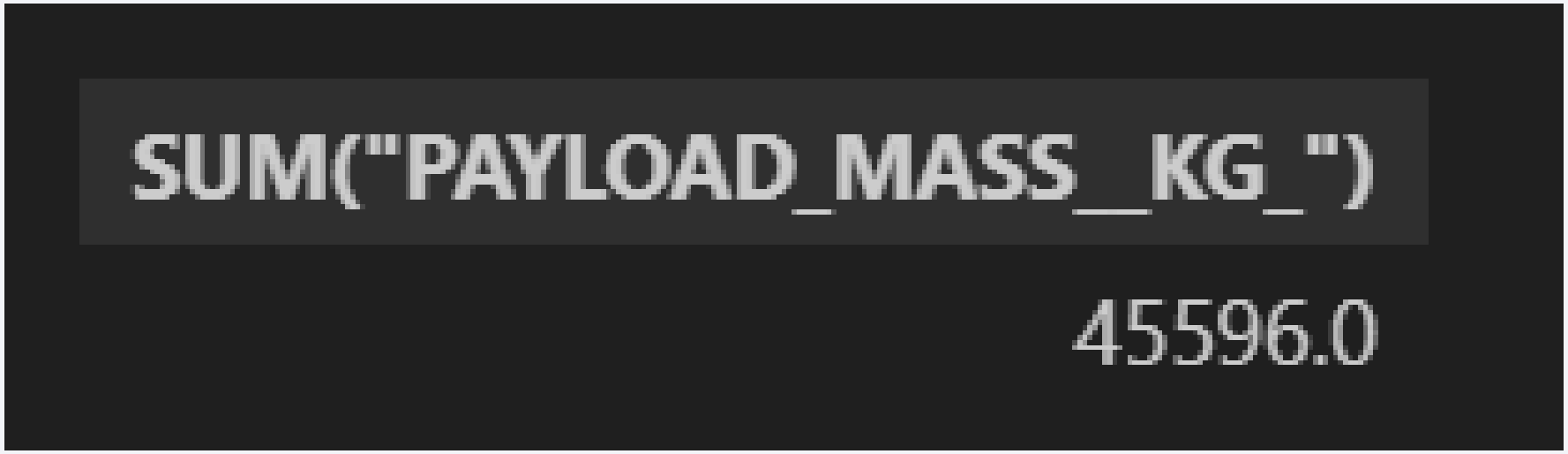
- The earliest five records where the launch site starts with CCA is below. Very interesting.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE Customer = "NASA (CRS)";

- Here is the total mass carried by the boosters, but only the ones launched by NASA.



Average Payload Mass by F9 v1.1

- SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass
- FROM SPACEXTBL
- WHERE Booster_Version LIKE 'F9 v1.1%';
- The average payload mass is shown here in kg

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- SELECT MIN(Date) AS First_Successful_Landing_Date
- FROM SPACEXTBL
- WHERE Landing_Outcome = 'Success (ground pad)';
- Here is the first successful landing outcome yay!

First_Successful_Landing_Date
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT Booster_Version
- FROM SPACEXTBL
- WHERE Landing_Outcome = 'Success (drone ship)'
- AND PAYLOAD_MASS__KG_ > 4000
- AND PAYLOAD_MASS__KG_ < 6000;
- These are the names of the boosters that fit into the specifications shown in the title.

[13]:

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

Total Number of Successful and Failure Mission Outcomes

- SELECT Mission_Outcome, COUNT(*) AS Total_Count
- FROM SPACEXTBL
- GROUP BY Mission_Outcome;
- Here is a list of the total successful missions and failure missions.

Mission_Outcome	Total_Count
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- SELECT Booster_Version
- FROM SPACEXTBL
- WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
- These are the booster versions which have carried the max payload mass, there are a lot of them.

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

- SELECT
- SUBSTR(Date, 4, 2) AS Month,
- Landing_Outcome,
- Booster_Version,
- Launch_Site
- FROM SPACEXTBL
- WHERE SUBSTR(Date, 7, 4) = '2015'
- AND Landing_Outcome LIKE '%Failure%'
- AND Landing_Outcome LIKE '%drone ship%';
- A lot of data is shown about two missions in 2015 which failed to land on the drone ship.

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
SELECT
    Landing_Outcome,
    COUNT(*) AS Landing_Outcome_Count
FROM SPACEXTBL
WHERE DATE(substr(Date, 7, 4) || '-' || substr(Date, 1, 2) || '-' || substr(Date, 4, 2))
    BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Landing_Outcome_Count DESC;
```

- These are the value counts for the different landing types for the date period specified.

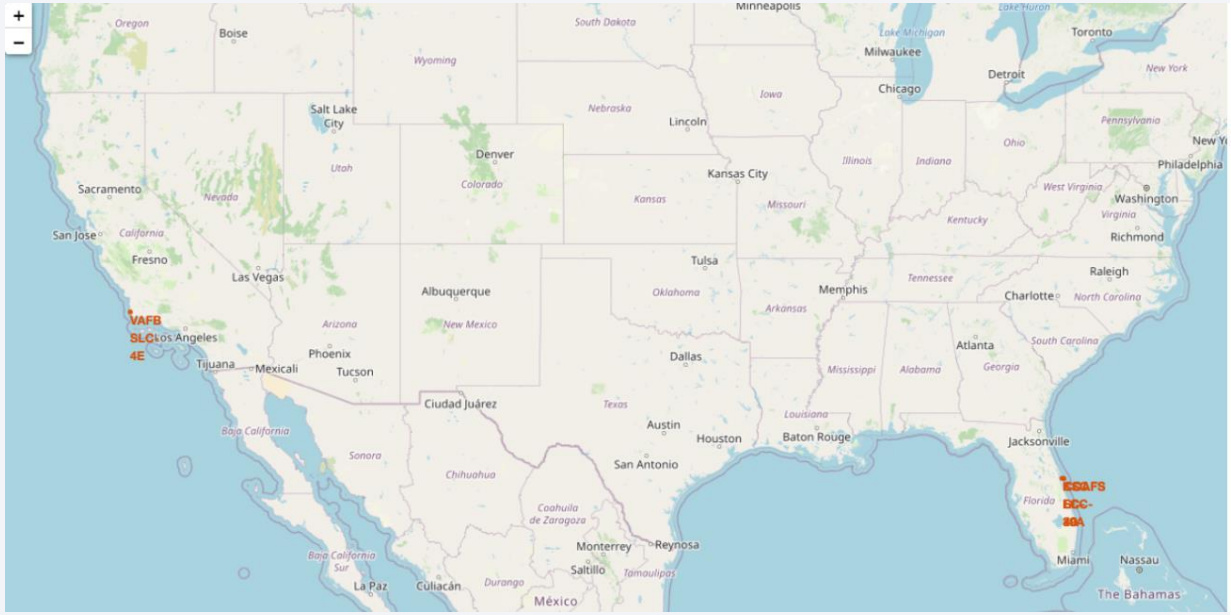
Landing_Outcome	Landing_Outcome_Count
No attempt	7
Success (drone ship)	2
Failure (parachute)	2
Failure (drone ship)	2
Controlled (ocean)	1

Section 3

Launch Sites Proximities Analysis

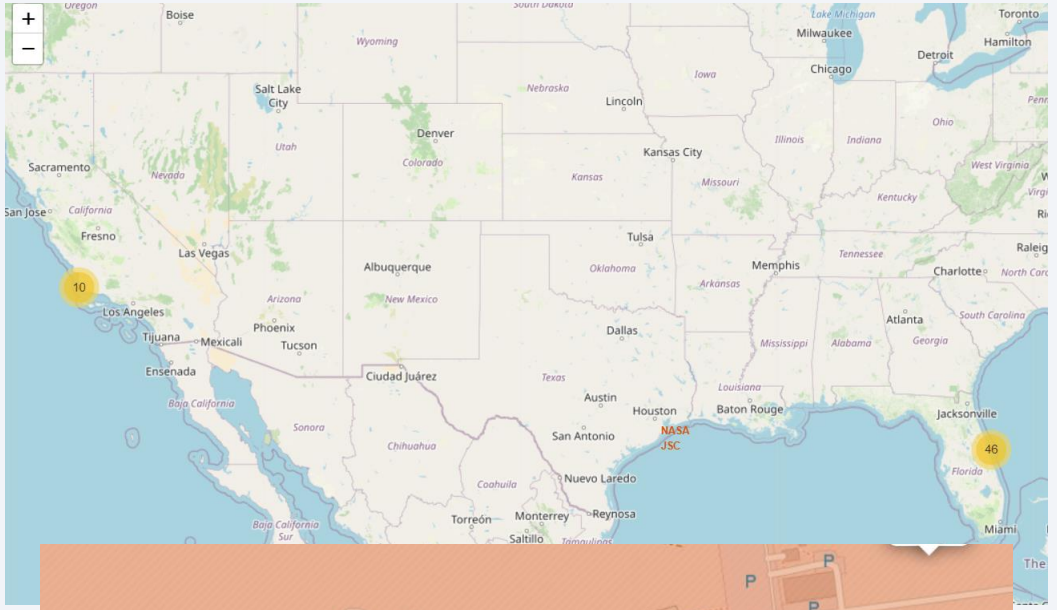
Launch Location Circles

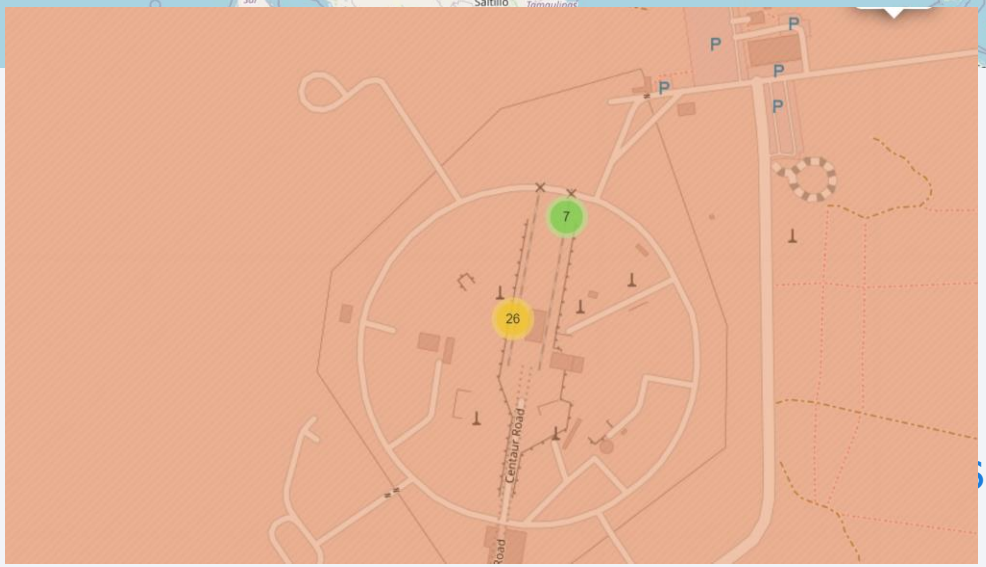
- This map circles all the launch sites for Falcon 9 rockets by SpaceX.
- Some are in California and some are in Florida



Launch Success at Different Locations

- This clip shows that many more launches occurred in Florida than California
- The lower screenshot shows the distribution between successes and failures at the CCAFS SLC launch site.

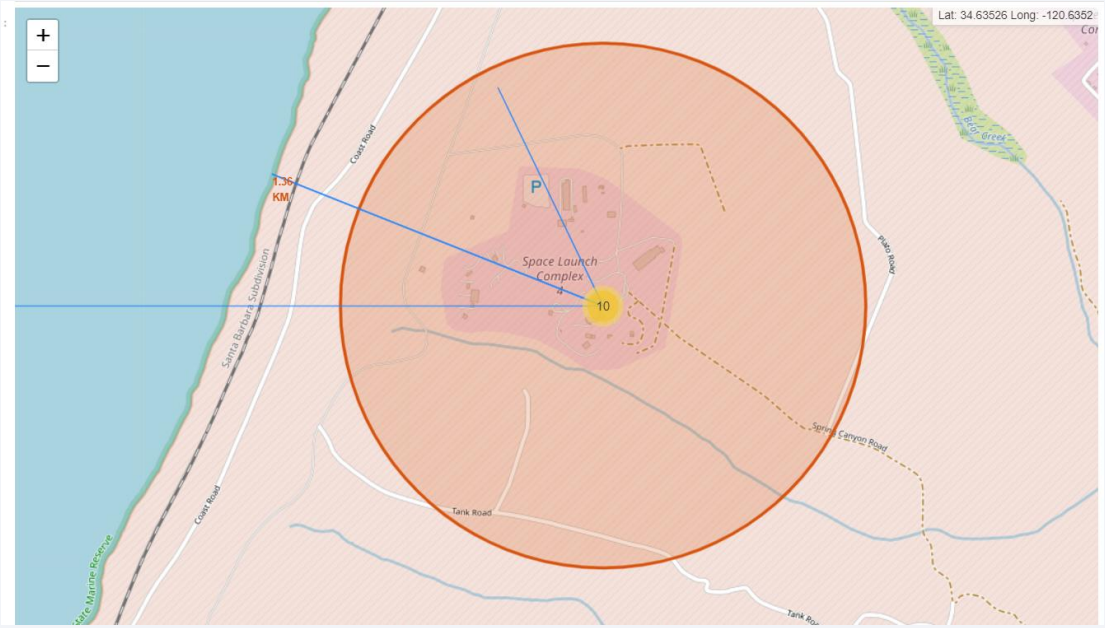


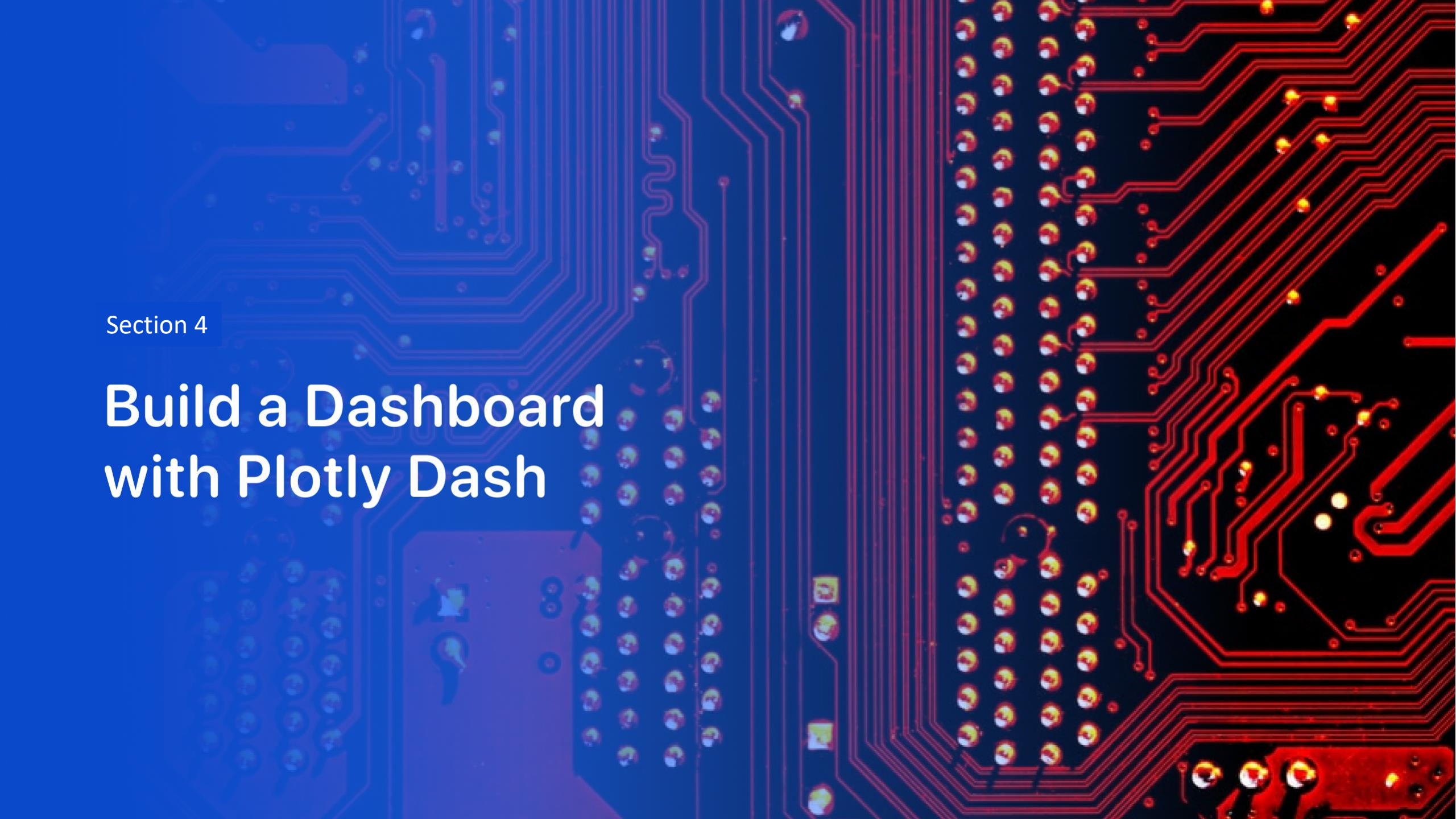


Significant Locations Distances

- Highlighted are the distances to the nearest highway, coastline, and railroad.

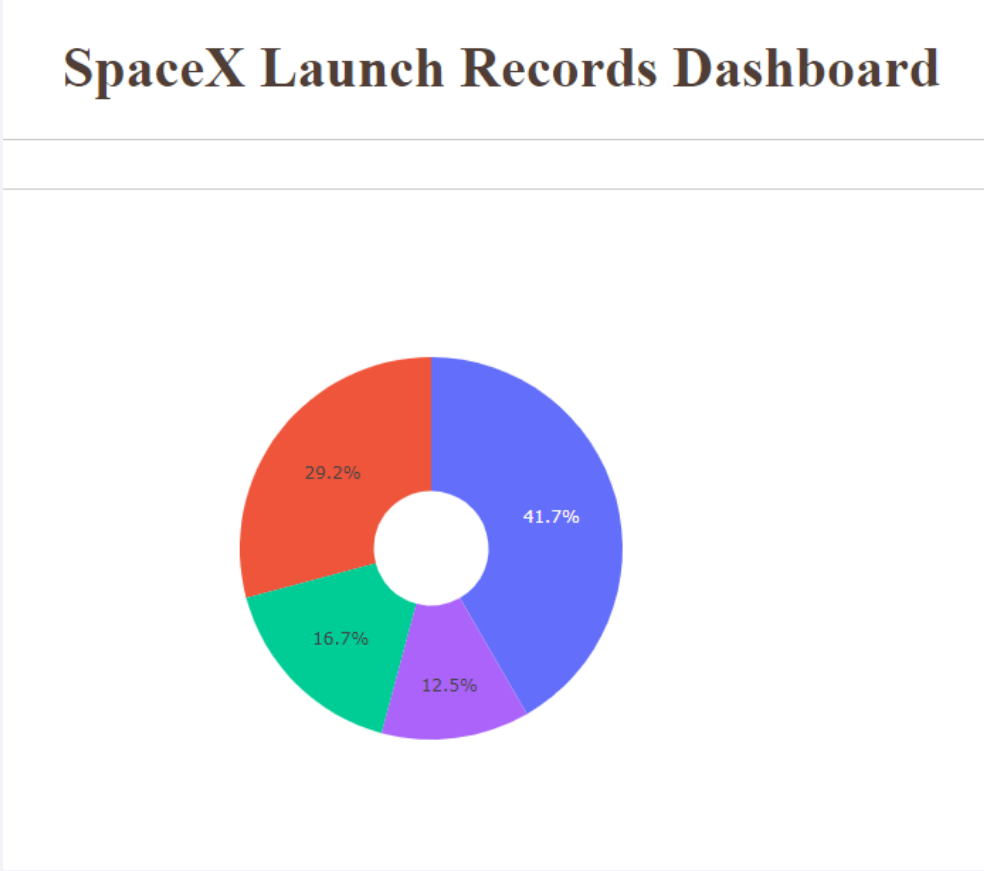
- These are significant local landmarks for this launch site and are important to track.
- Other sites may have different distances to these attractions.
- Sites seem to put distance between the launch pad and cities.




Section 4
Build a Dashboard with Plotly Dash

Proportion of Successful Launches by Site

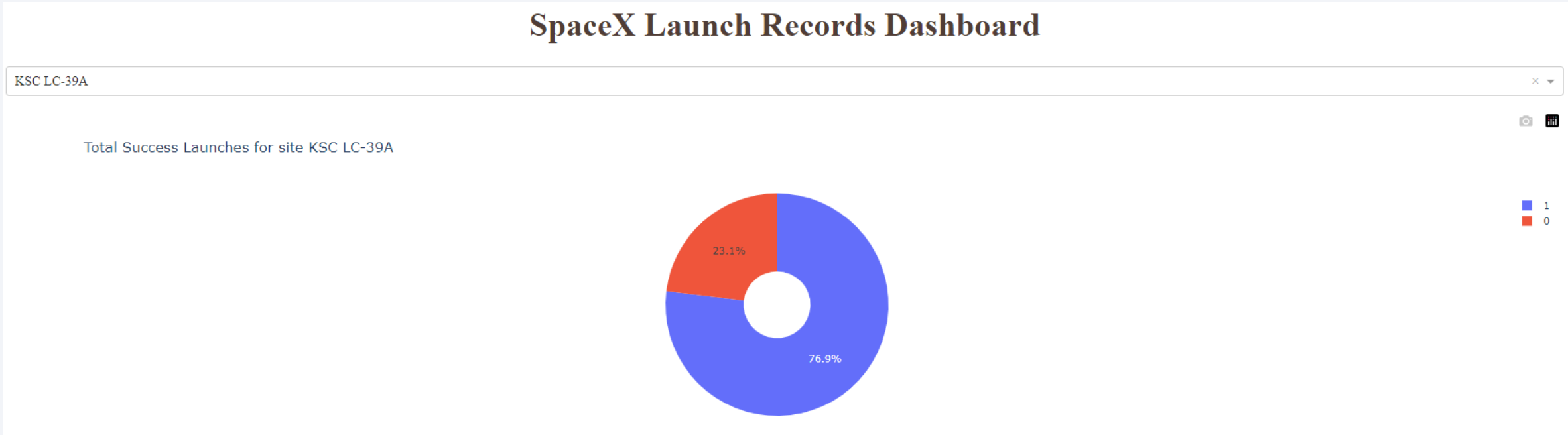
- This pie chart shows the proportion of all successful launches that came from each launch site
- As seen, KSC LC-39A has the largest proportion of successful launches at 41.7%
- CCAFS LC-40 comes in second at 29.2% of successful launches



Most Successful Launch Site: KSC LC-39A

- This pie chart shows the launch site with highest launch success ratio

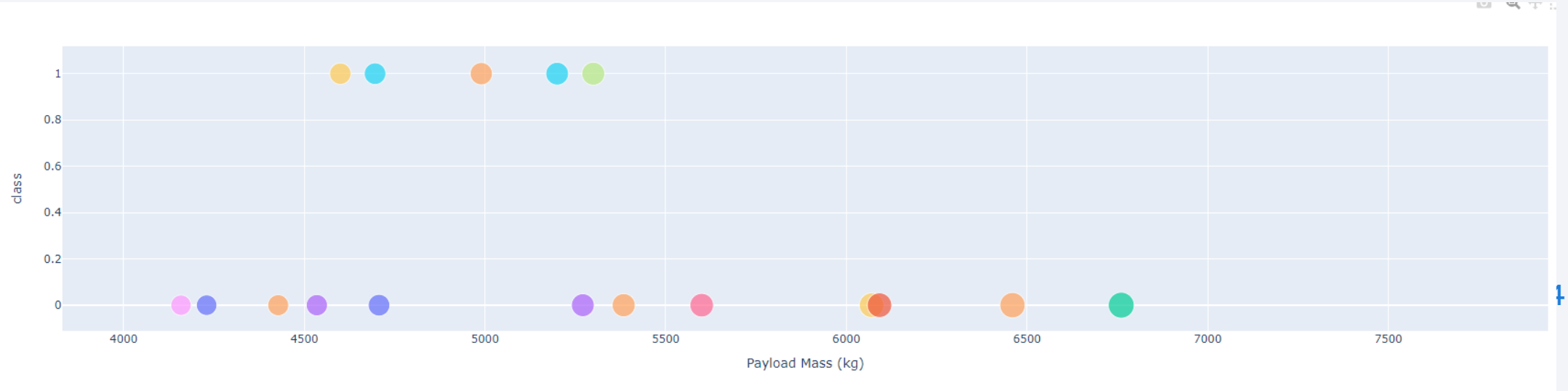
- Over 75% of launches from KSC LC-39A were successful, making it the best launch site that SpaceX has used.



Scatterplot of High Mass Payloads

- Success rate scatterplot for launches with high payload masses
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

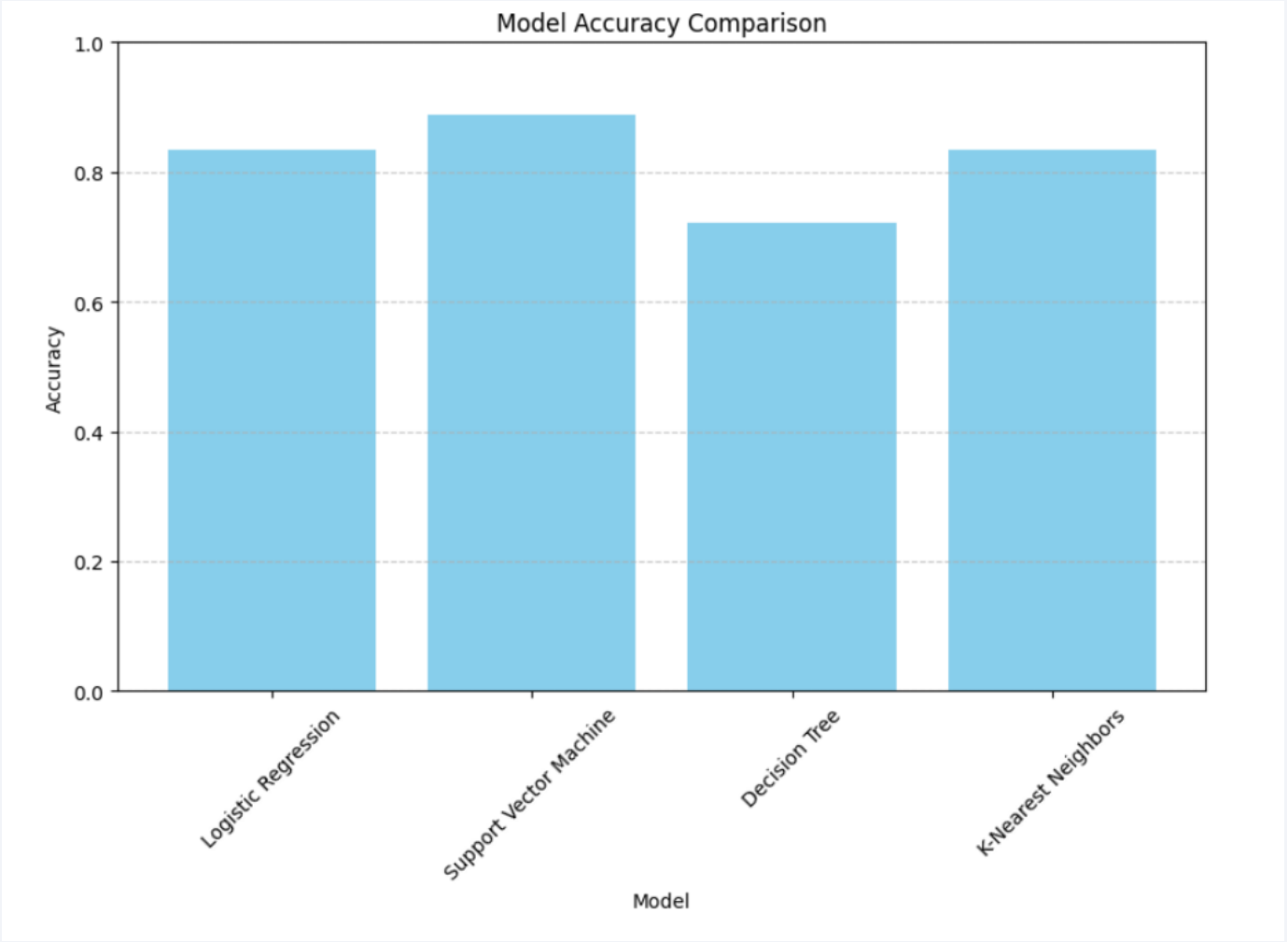


Section 5

Predictive Analysis (Classification)

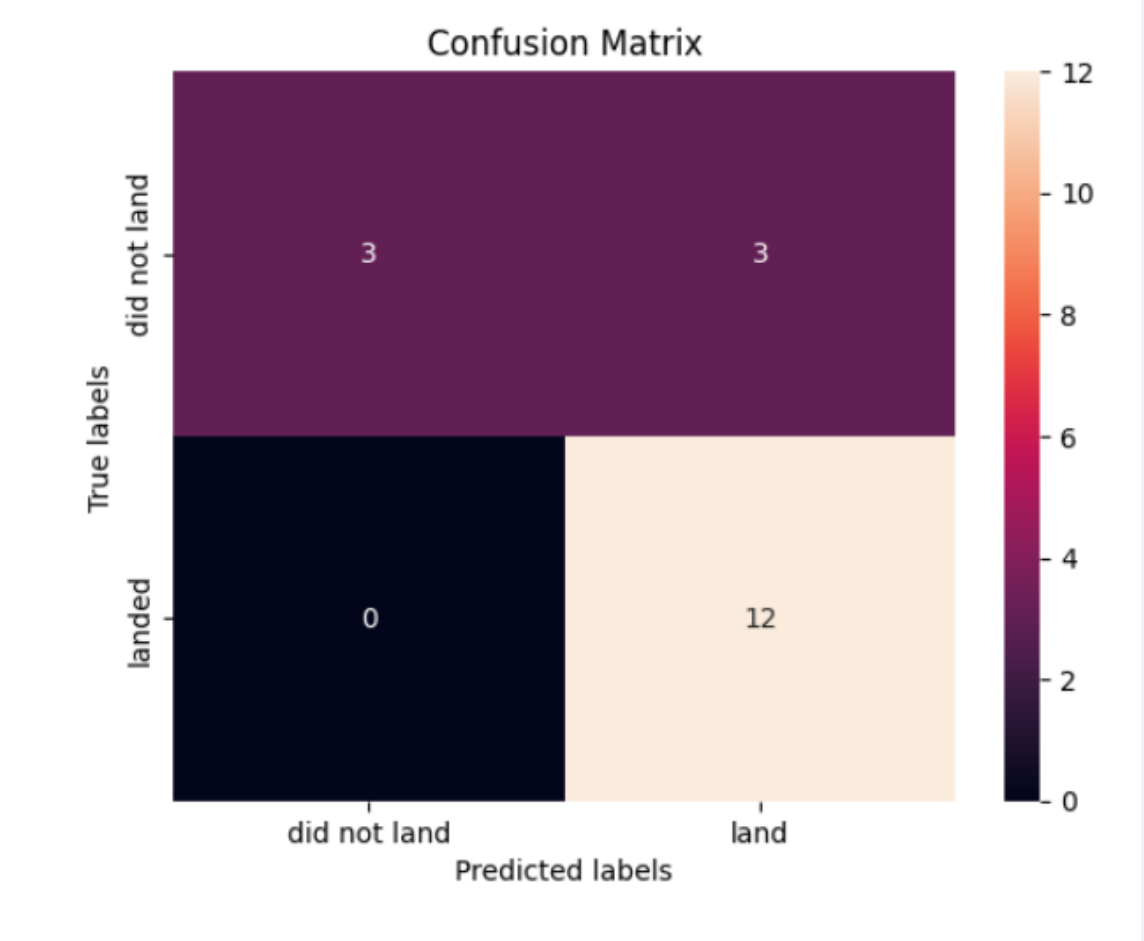
Classification Accuracy

- Model Accuracy visualization for classification models
- As seen in the graph, the Support Vector Machine model has the highest accuracy, making it the best performing model.



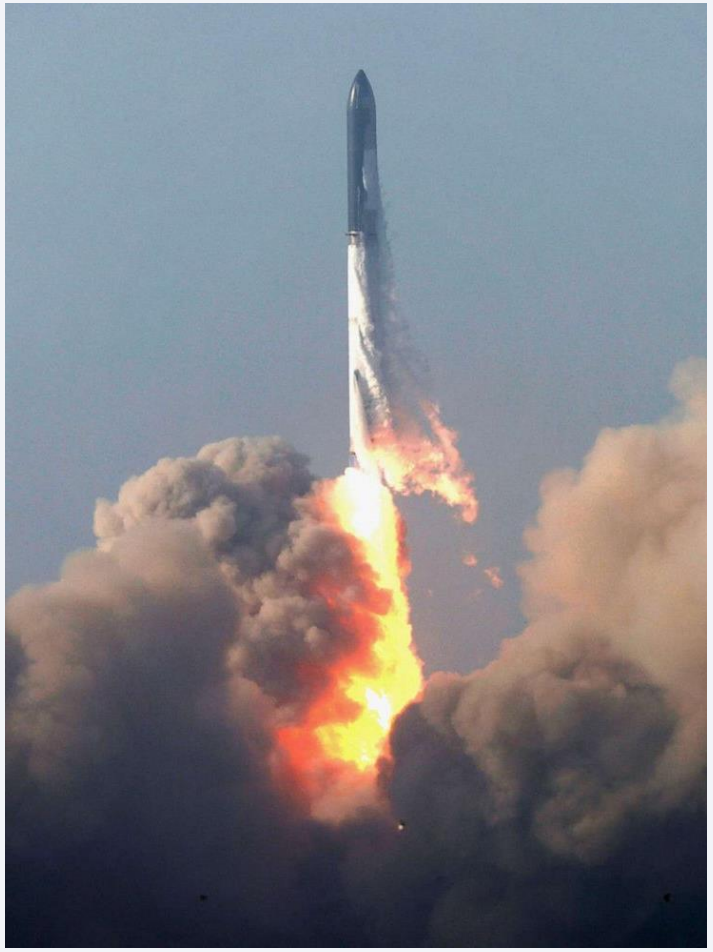
Confusion Matrix for SVM Model

- Shown is a confusion matrix for the best performing model, the Support Vector Machine Model.
- It had a 0.889 accuracy score, making it the best performing classification model to predict SpaceX launch success



Conclusions

- Webscraping and data querying is crucial to understanding the SpaceX company
- Many factors affect the outcome of Falcon 9 launches
- Using a Support Vector Machine model we can accurately predict the outcome of upcoming SpaceX launches (0.889% accuracy)
- By using the methodology described in this presentation, we can predict whether SpaceX will reuse their rocket. This is of great help to our competing brand, SpaceY



Appendix

Example Code for Making a ML Model

```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}

KNN = KNeighborsClassifier()
gscv = GridSearchCV(KNN, parameters, scoring = 'accuracy', cv = 10)
knn_cv = gscv.fit(X_train, Y_train)
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

[13]:

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

Result of an SQL query

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0

Data table from webscraping and API requests.

Thank you!