IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yuta A
2022-04-10

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting data using SpaceX API and Web Scraping
  - Data wrangling and Exploratory Data Analysis (EDA)
  - Machine learning prediction fitting various classification models
- Summary of all results
  - Key factors for excellent success rate may be grid fin, launch site and payload mass.
  - We can predict the outcome of each launch by using a decision tree classifier model.

Introduction

- The first stage is the most important factor for rocket launches
  - The first stage is quite large and more expensive than the second stage
  - Space X saves the cost reusing the first stage compared to other rocket providers
- If determine whether each first stage will land successfully, we could save much cost
- Therefore, what I should do is as follows:
  - To determine the key factor for success and the price of each launch;
  - To predict the outcome of Space X launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Extracted data by web scraping and using SpaceX API
- Perform data wrangling
  - Removed unnecessary data, dealt with missing values and created a column used in analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into training and test data, and calculated a score of each model to determine the best fit model

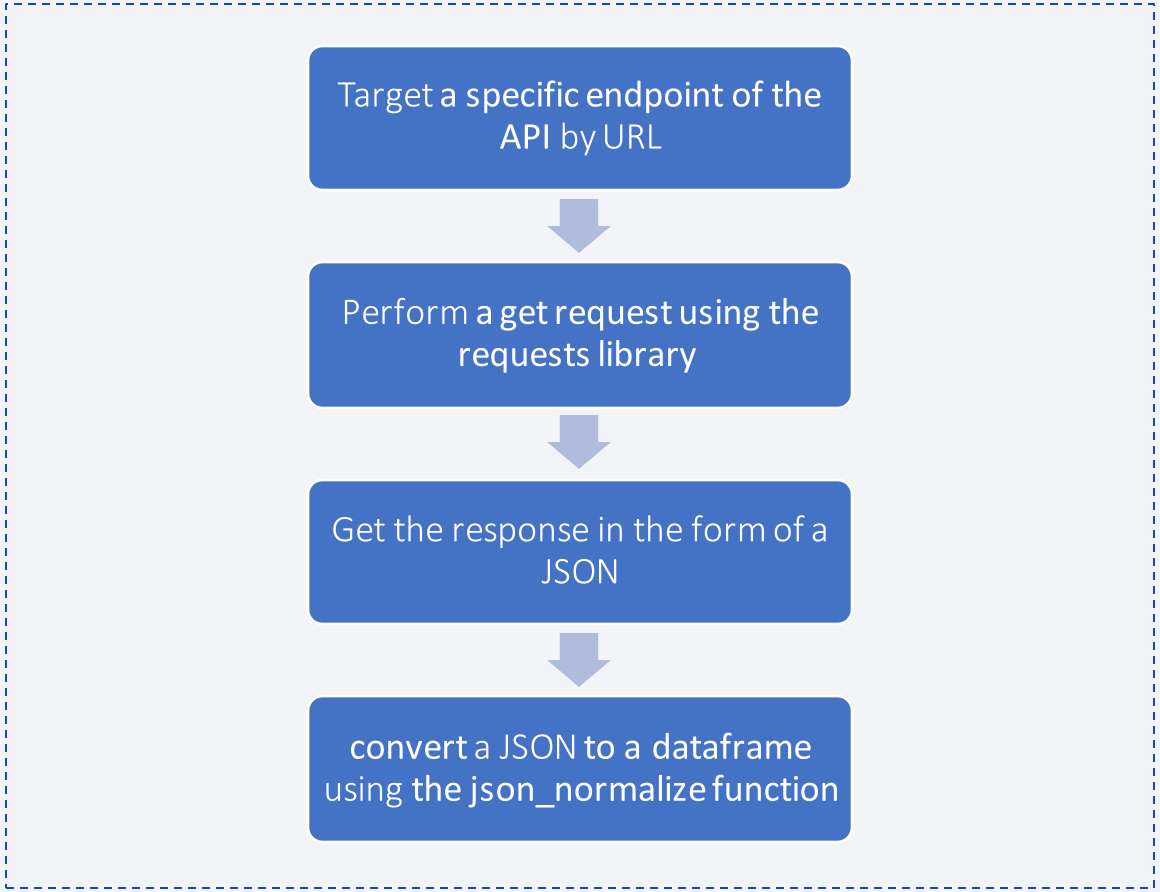
Data Collection

- Extracted a Falcon 9 data using two methods and sources:
  - Made a get request to the SpaceX API (https://api.spacexdata.com/v4/launches/past);
  - Web Scraping from Wikipedia ("List of Falcon 9 and Falcon Heavy launches").

SpaceX API	Web Scraping
1. Request data from SpaceX API using ".get()" 2. Convert the json result into a dataframe 3. (after that, data wrangling)	1. Request a HTML page using HTTP GET method 2. Create a BeautifulSoup object 3. Collect data from HTML table using BeautifulSoup object

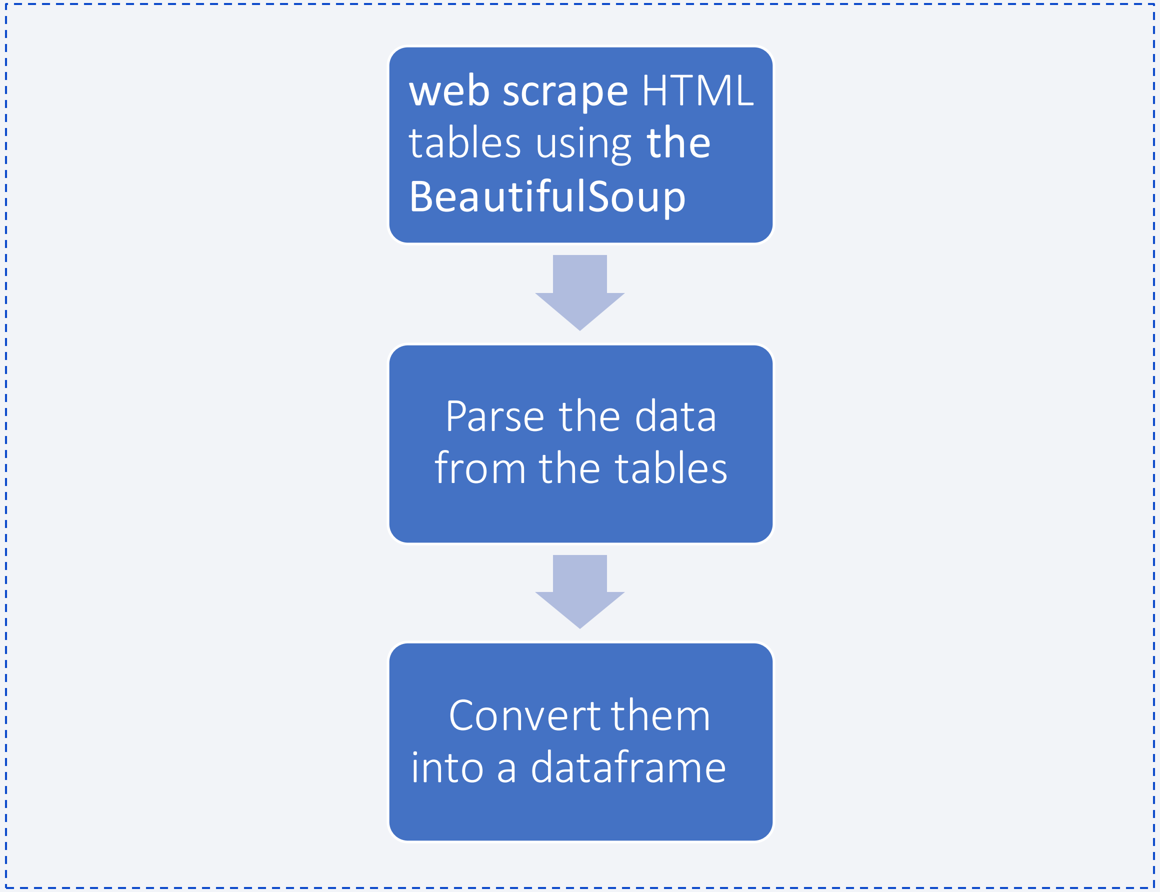
Data Collection – SpaceX API

- Used SpaceX REST API to obtain past launch data
- converted JSON data getting as a response to a flat table to utilize it in the following process
- The detail is as shown in the flow chart to the right
- GitHub URL: https://github.com/Yuta555/applied_data_science_capstone/blob/main/Data%20Collection%20API.ipynb



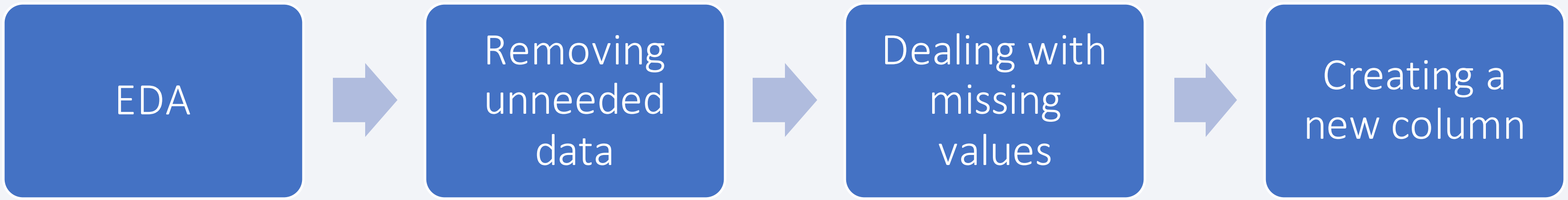
Data Collection - Scraping

- Web scraped to obtain past launch data from Wikipedia
- Converted the data to a flat table to utilize it in the following process
- The detail is as shown in the flow chart to the right
- GitHub URL: https://github.com/Yuta555/applied_data_science_capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

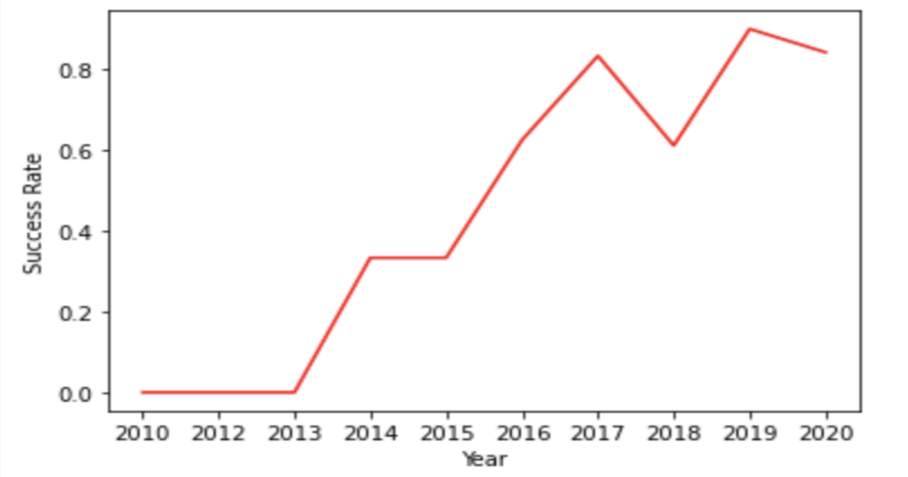
- Initially did fundamental Exploratory Data Analysis (EDA) to understand the Data
- Then cleaned the data to make it easy to manipulate in the post-process
  - Specifically, removed unnecessary data, replaced missing values with something like mean values, and created a new column to make it easier to analyze



- GitHub URL: https://github.com/Yuta555/applied_data_science_capstone/blob/main/EDA.ipynb

EDA with Data Visualization

- Plotted scatter point charts, bar charts and line charts to see if there was any relationship between pair of features. Examples follow:
  - A scatter point chart with x axis to be Flight Number and y axis to Payload Mass;
  - A bar chart depicting success rate of each orbit type;
  - A line chart showing the launch success yearly trend.



- GitHub URL: https://github.com/Yuta555/applied_data_science_capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Display the names of the unique launch sites in the space mission;
  - Display 5 records where launch sites begin with the string 'CCA';
  - Display the total payload mass carried by boosters launched by NASA (CRS);
  - Display average payload mass carried by booster version F9 v1.1;
  - List the date when the first successful landing outcome in ground pad was achieved;
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000;
  - List the total number of successful and failure mission outcomes;
  - List the names of the booster_versions which have carried the maximum payload mass;
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015;
  - Rank 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
- GitHub URL: https://github.com/Yuta555/applied_data_science_capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Added Markers, circles, marker clusters and lines to Folium maps
  - Markers indicated the place where launch sites are.
  - Circles were used to highlight area where launch sites are.
  - Marker clusters simplified a map containing many markers having the same coordinate.
  - Lines were used to indicate the distance between two coordinates.

- GitHub
  URL: https://github.com/Yuta555/applied_data_science_capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Added a launch site drop-down and a range slider to select payload
  - The launch site drop-down can be used to filter the data by selected launch site
  - The range slider makes us select different payload range
- Added a pie chart and a scatter point chart
  - The pie chart is about the percentage of success outcomes based on selected site dropdown
  - The scatter point chart indicates how payload may be correlated with mission outcomes for selected site(s)
- GitHub URL: https://github.com/Yuta555/applied_data_science_capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Fit several classification models as follows: logistic regression; support vector machine; decision tree classifier; and k nearest neighbors
- Split the data into training and testing data to check and enhance the accuracy of the models
- implemented following process on each model

Created an object of each classification model → determined the best parameters using a GridSearchCV function with training datase → Calculated the auccuracy score using testing dataset

- GitHub URL: https://github.com/Yuta555/applied_data_science_capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

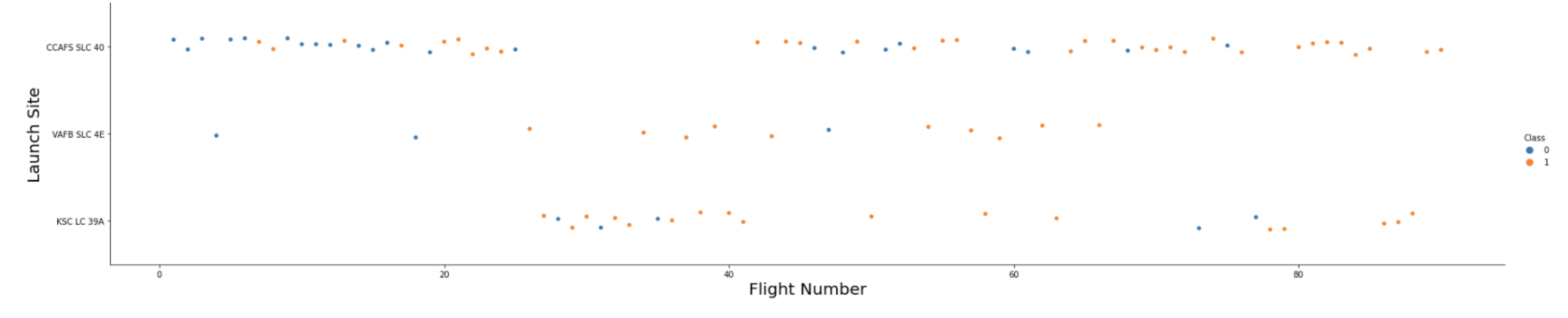
Results

- Summary of the results is as below:
  - Success rate has increased gradually since 2013;
  - Success rate seems to be correlated with grid fin, launch site and payload mass.
  - The most accurate classification model is decision tree classifier in this case.

Section 2

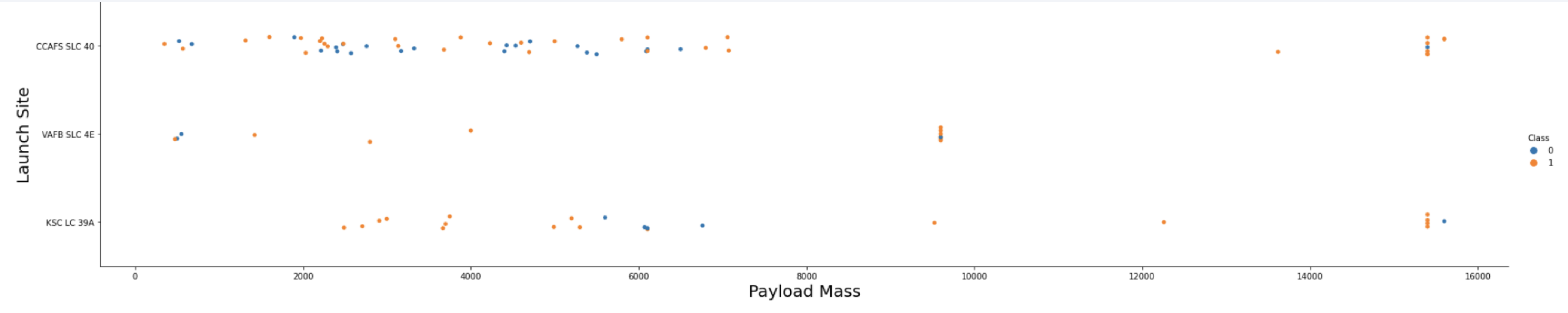
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 had the largest number of flights and improved the success rate nowadays.
- The outcomes of VAFB SLC 4E and KSC LC 39A also seem to get better.

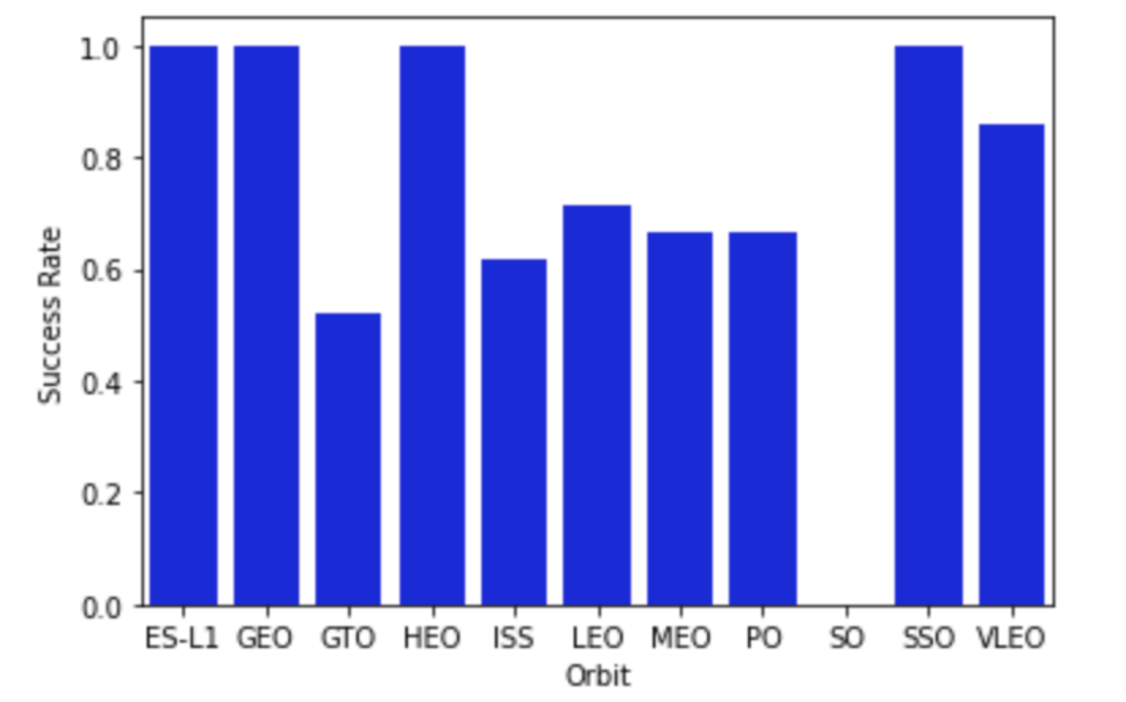
Payload vs. Launch Site



- Payloads under 5,000kg on KSC LC 39A had excellent success rate.
- Payloads over 9,000kg also had high success rate.
- According to the plot above, payloads over 10,000kg may be feasible only on CCAFS SLC 40 and KSC LC 39A.

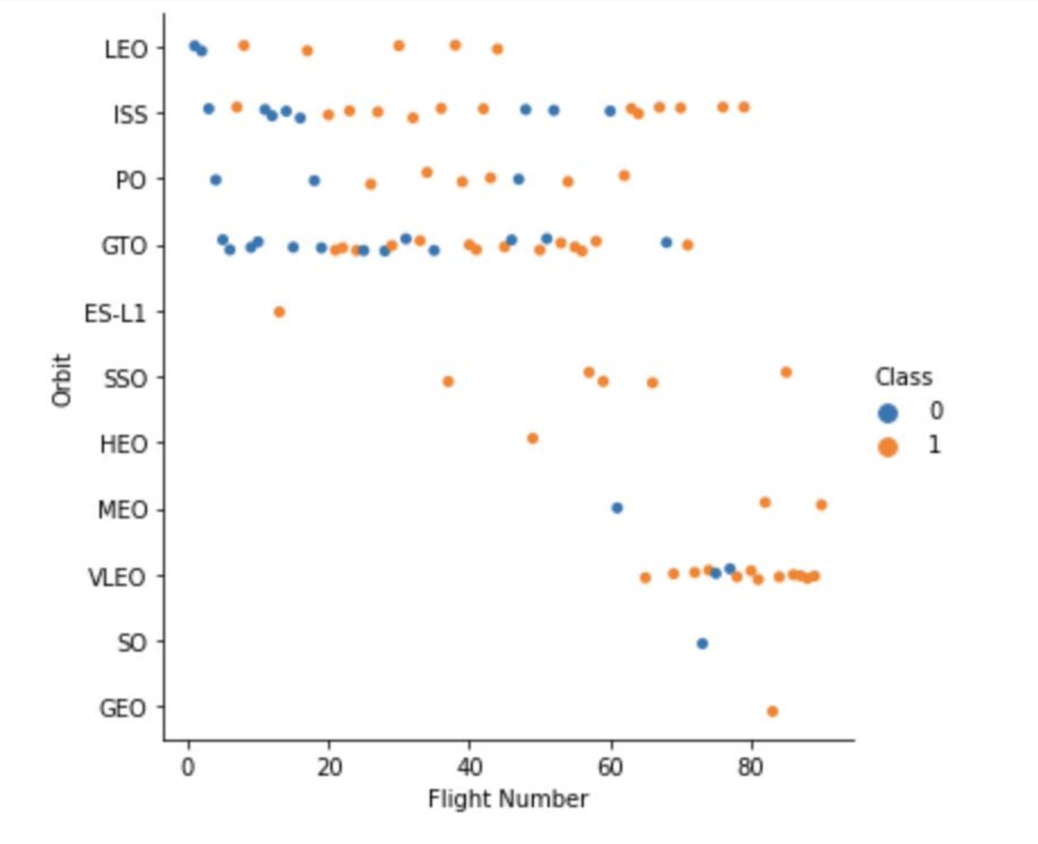
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO and VLEO had a high success rate (above 80%)
- GTO, ISS, LEO, MEO, PO and SO had a low success rate (under 80%)



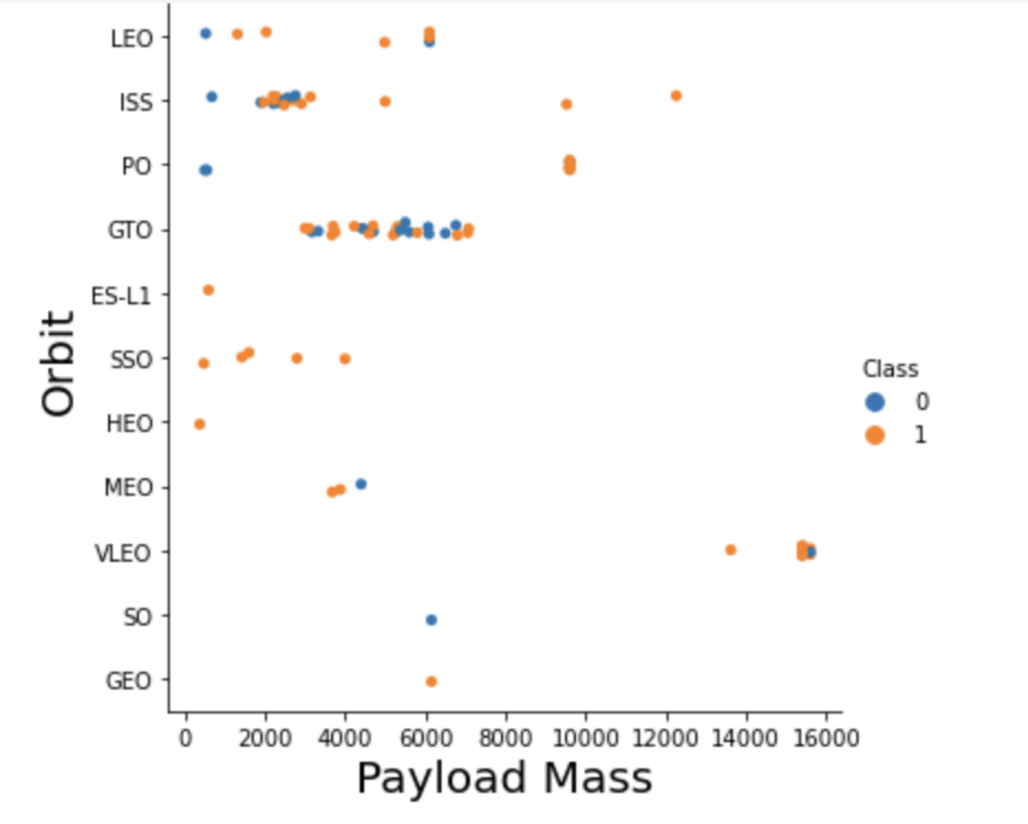
Flight Number vs. Orbit Type

- SSO had 100% success rate whereas it had 5 flights.
- The outcomes of LEO seems to have improved.
- GTO may be difficult to succeed because it had so many flights but the success rate had not improved.



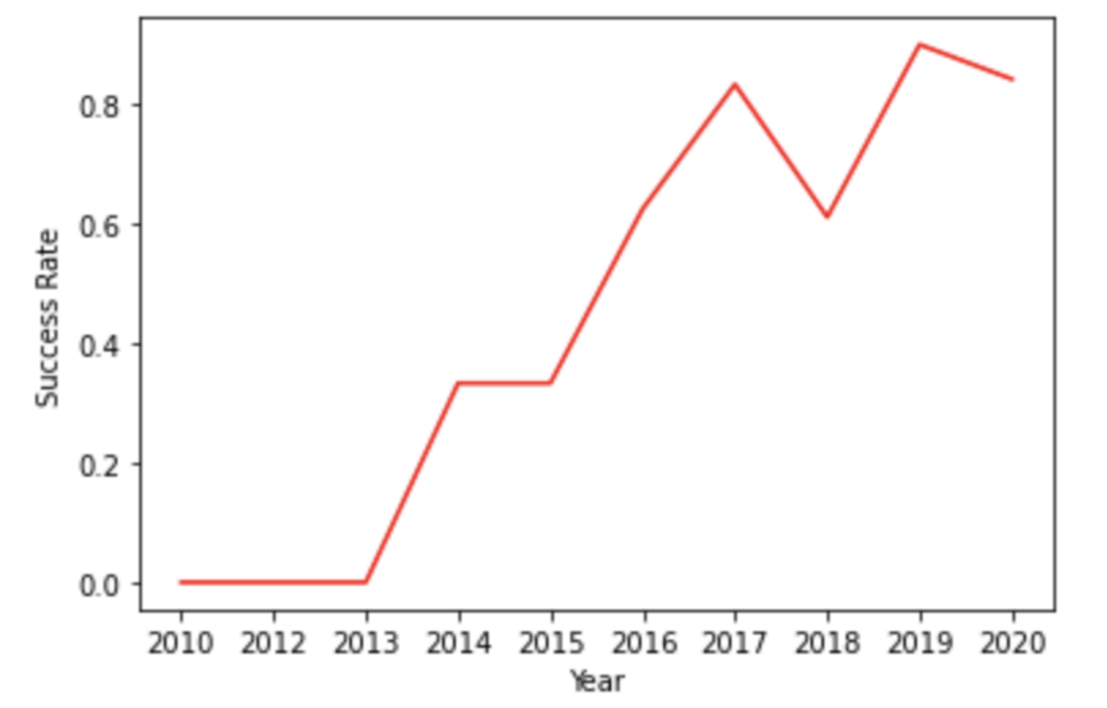
Payload vs. Orbit Type

- All SSO's flight succeeded and the payloads were under 5,000kg.
- When payloads of ISS and PO were over 9,000kg, those success rates were quite high, while total success rate of the orbits were not high.



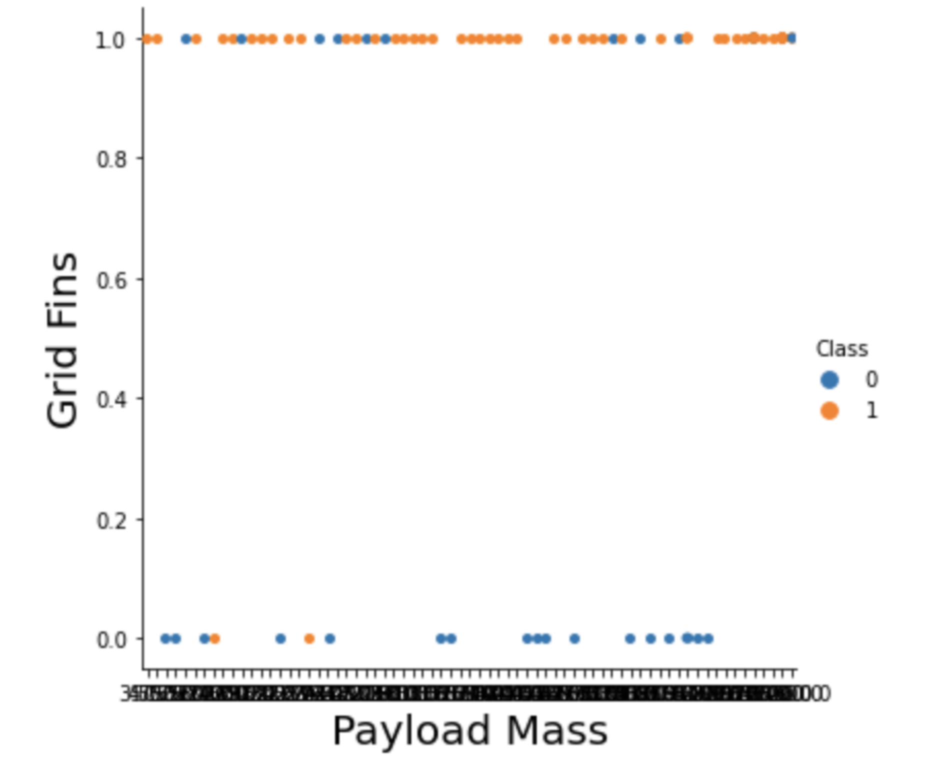
Launch Success Yearly Trend

- We can see that the success rate has increased dramatically since 2013.



(Additional) Grid Fins vs. Payload Mass

- Rockets which had a grid fin had much higher success rate than those which had no grid fin.



All Launch Site Names

- Displayed the names of the unique launch sites in the space mission
- As a result, found 4 launch sites as follows:

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

Launch Site Names Begin with 'CCA'

- Displayed 5 records where launch sites begin with the string 'CCA'
- As a result, got 5 success launch data as follows:

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

Total Payload Mass

- Displayed the total payload mass carried by boosters launched by NASA (CRS);
- Summed payload mass data whose customer data is "NASA (CRS)."

total_payload
45596

Average Payload Mass by F9 v1.1

- Displayed average payload mass carried by booster version F9 v1.1
- The result was as shown below:

average_payload
2928

First Successful Ground Landing Date

- Listed the date when the first successful landing outcome, grouped by landing outcome
- We can see that the date when the first successful landing outcome in ground pad was achieved was "2015-12-22."

Landing _Outcome	min_date
Controlled (ocean)	2014-04-18
Failure	2018-12-05
Failure (drone ship)	2015-01-10
Failure (parachute)	2010-06-04
No attempt	2012-05-22
Success	2018-07-22
Success (drone ship)	2016-04-08
Success (ground pad)	2015-12-22
Uncontrolled (ocean)	2013-09-29

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed the names of the boosters which have success in drone ship and have payloaed mass greater than 4000 but less than 6000
- As a result, 4 booster versions were shown as below:

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

Total Number of Successful and Failure Mission Outcomes

- Listed the total number of successful and failure mission outcomes as below:

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

Boosters Carried Maximum Payload

- Listed the names of the booster_versions which have carried the maximum payload mass.

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

- Listed the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- As a result, only 2 rows were listed as below:

Landing _Outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Ranked 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.

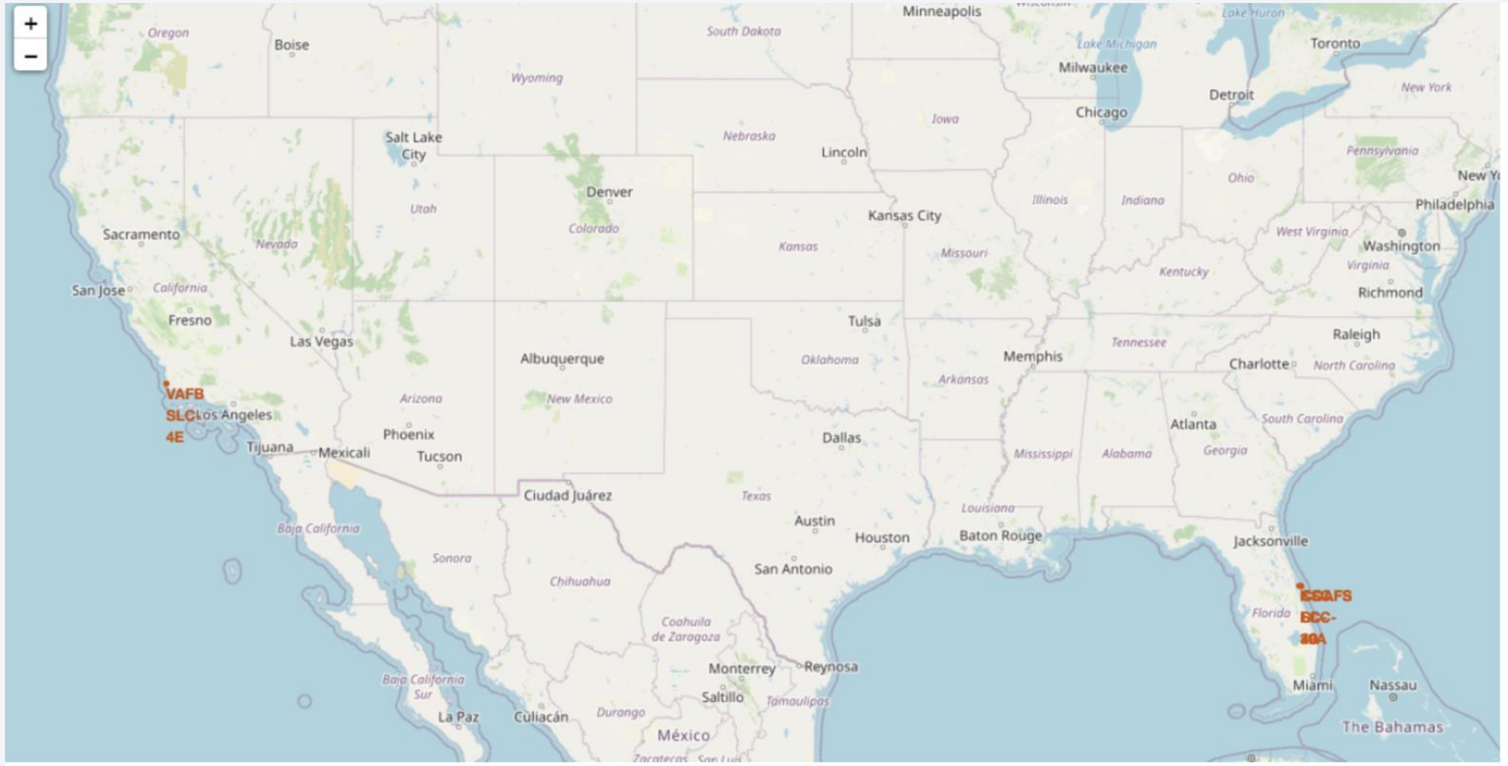
Landing _Outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

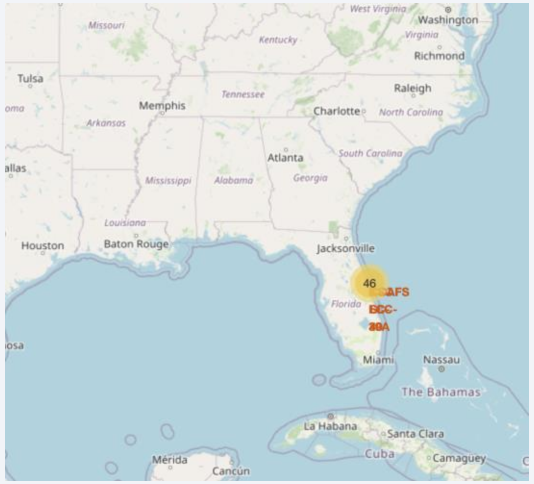
All launch sites

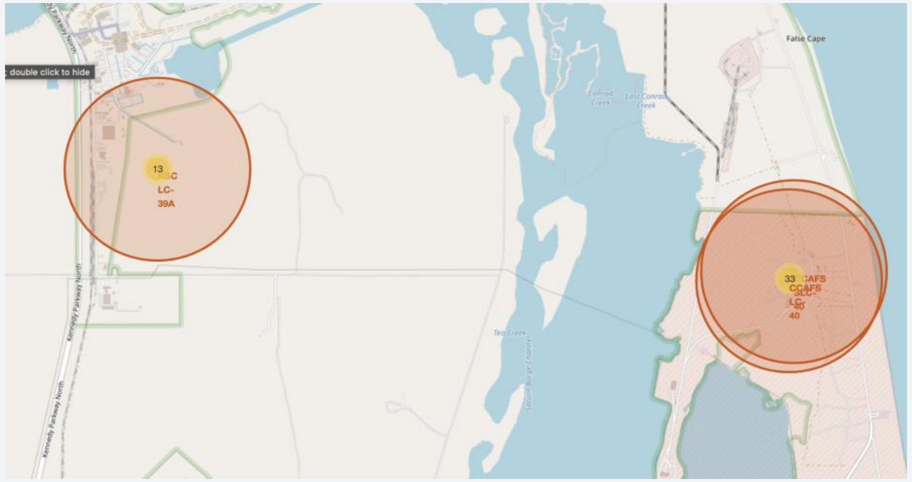
- All launch sites are located beside the sea as below:

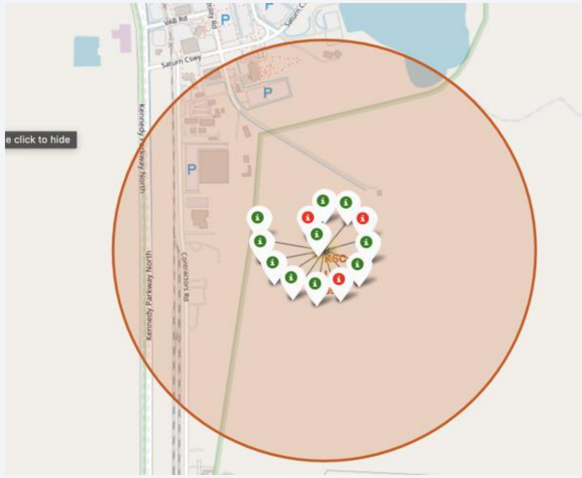


Outcomes of launches for each site

- We can see which sites have a high success rate by checking color-labeled markers.
- Example of KSC LC-39A is as below:

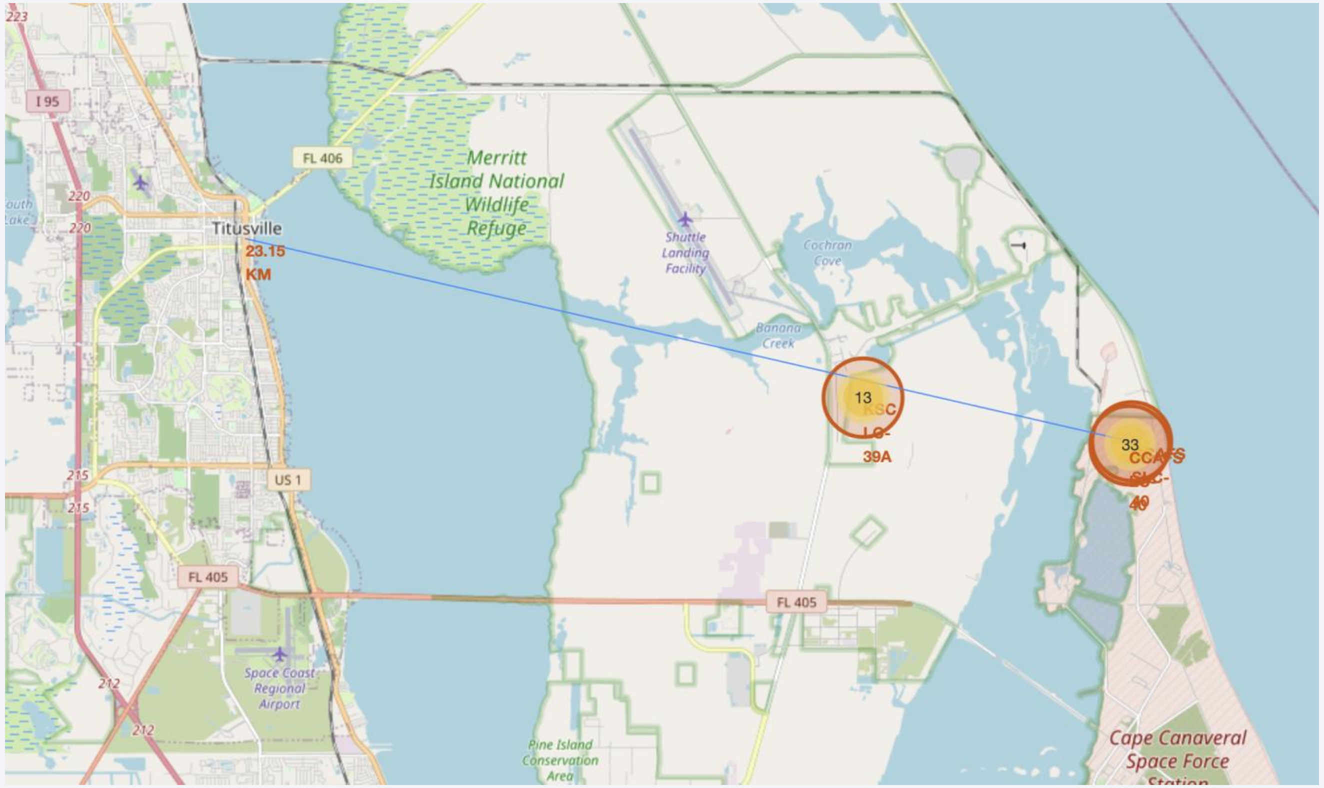






Distance between launch sites and nearest city

- CCAFS SLC-40 may be the safest site because it is more than 20km away from the nearest city, Titusville.

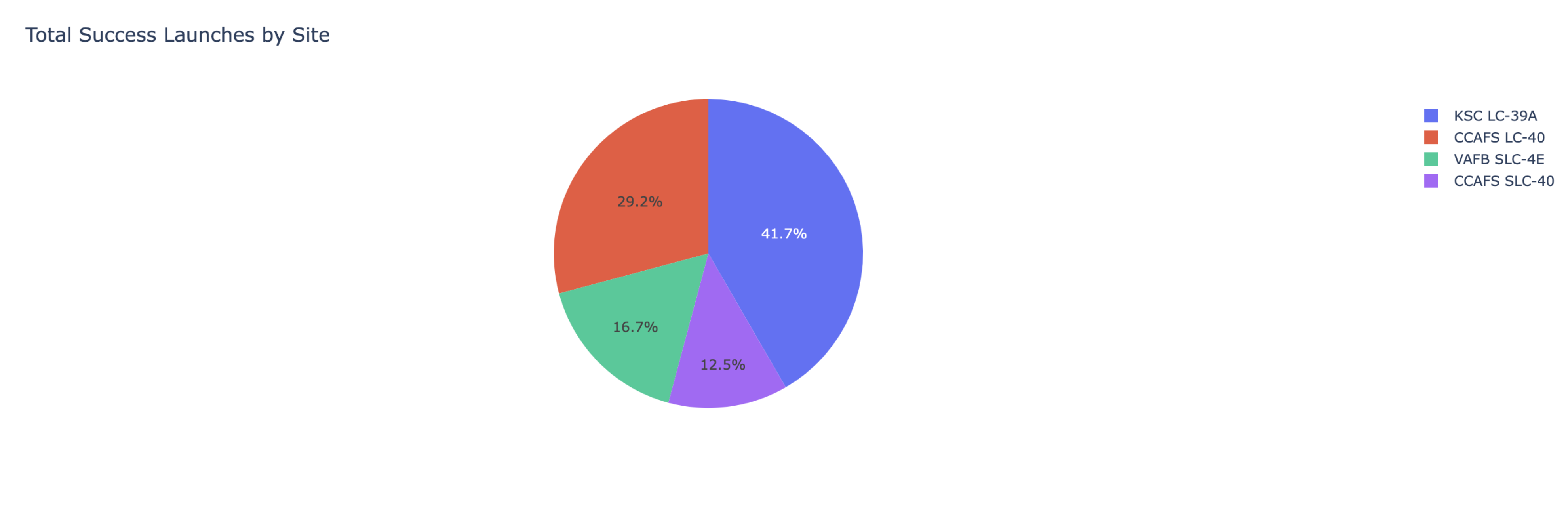


Section 4

Build a Dashboard with Plotly Dash

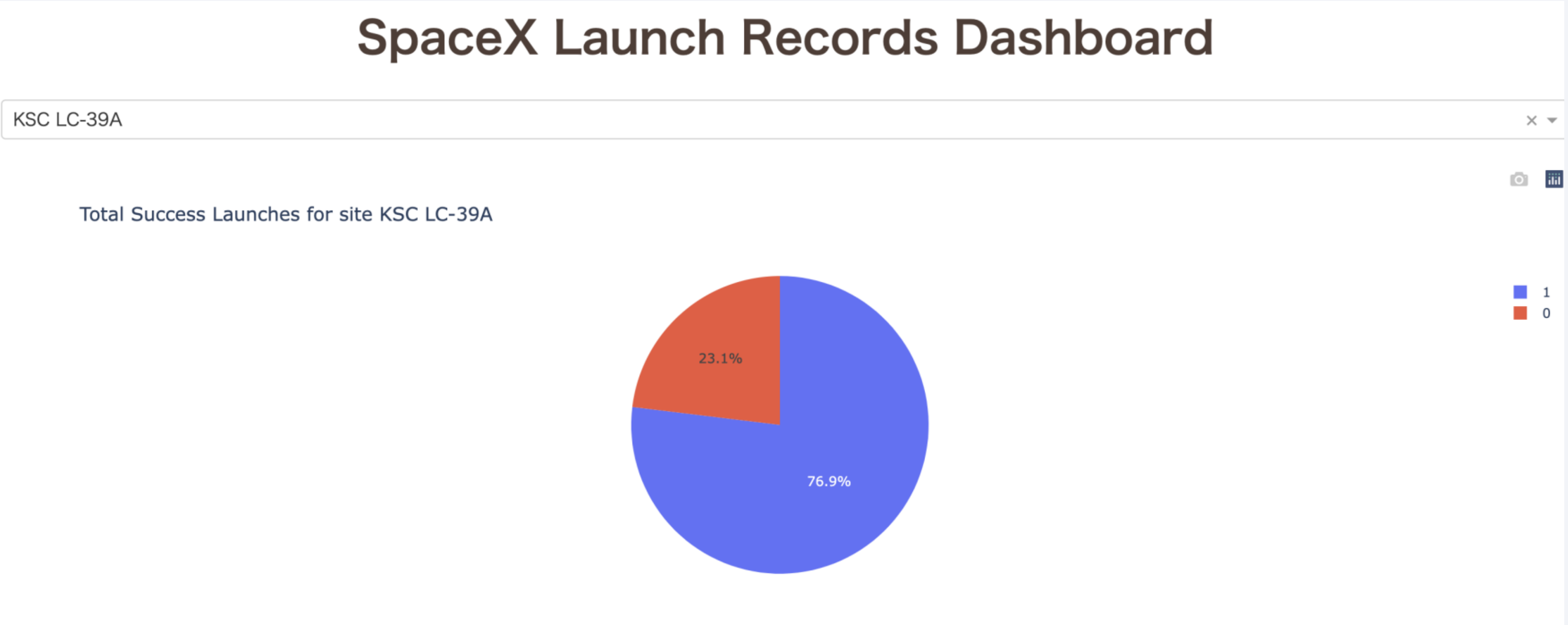
Success Launches by Site

- KSC LC-39A accounted for the largest percentage, followed by CCAFS LC-40.



Launch Success rate for KSC LC-39A

- KSC LC-39A had the highest success rate, which is 76.9%.



Successful Payload and Booster version

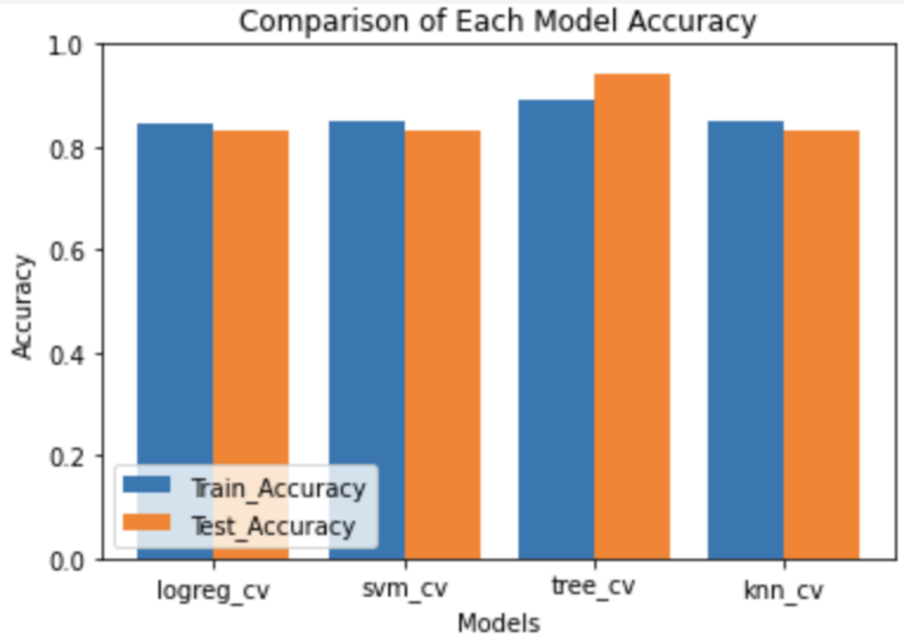
- According to scatter plot, FT boosters whose payloads are between 2,000kg and 6,000kg seem to be most successful.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

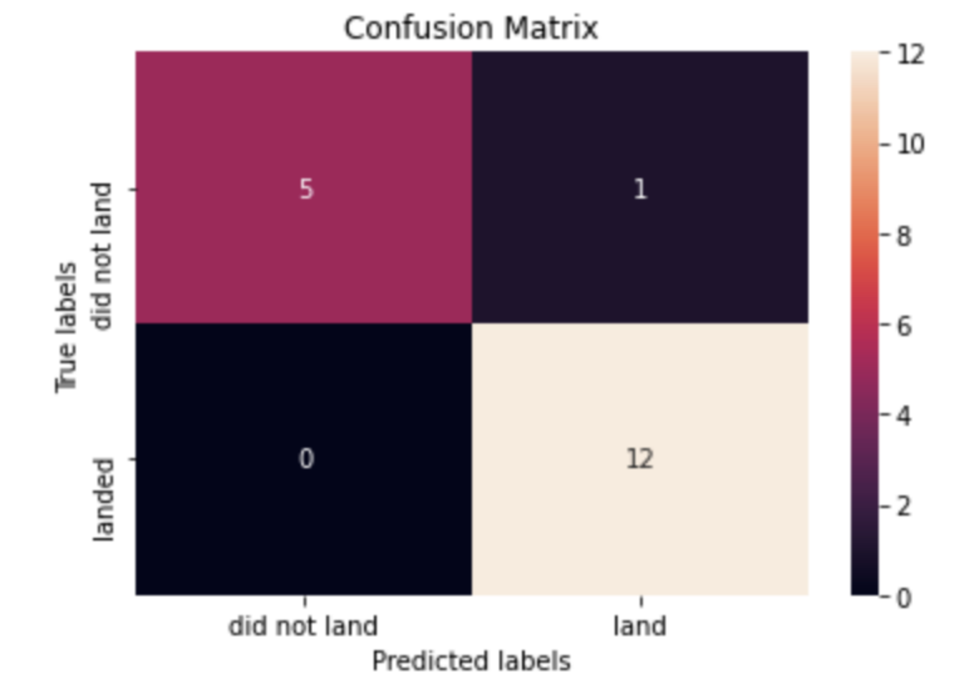
- Fitted 4 classification models and compare the accuracy.
- Decision tree classifier model has the highest accuracy score in both training and testing data, especially the score by testing data is over 90%.



	Train_Accuracy	Test_Accuracy
logreg_cv	0.846429	0.833333
svm_cv	0.848214	0.833333
tree_cv	0.891071	0.944444
knn_cv	0.848214	0.833333

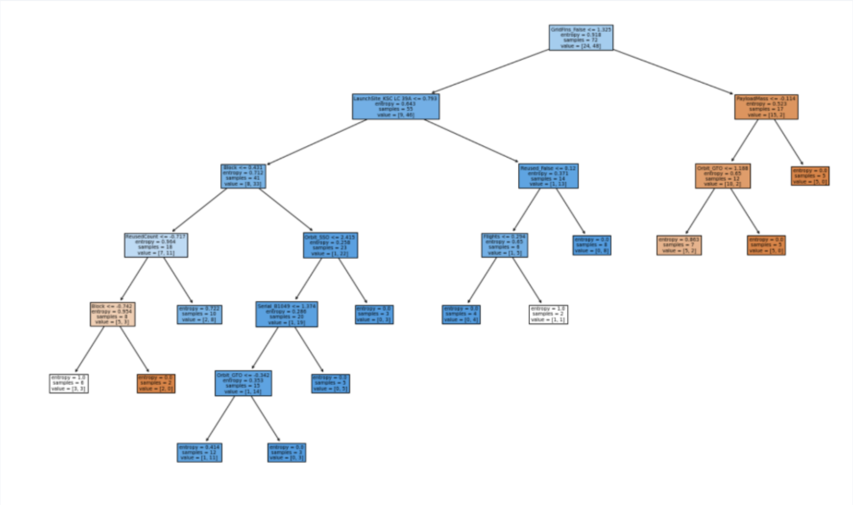
Confusion Matrix

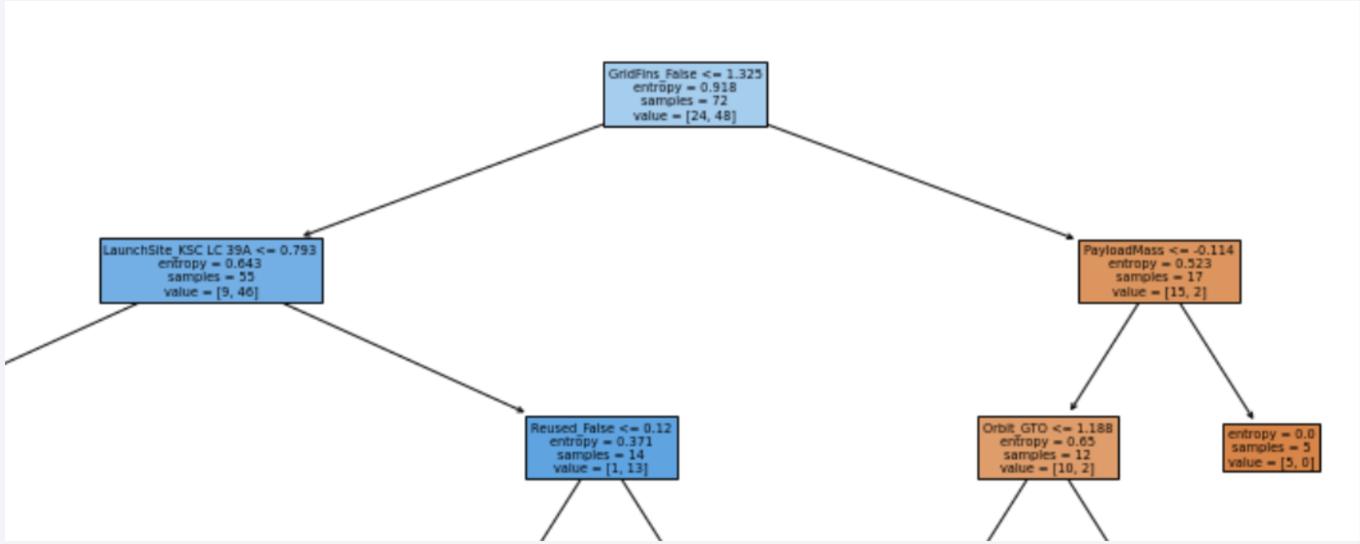
- The confusion matrix of decision tree classifier model is as below.
- Almost all data are true positive or true negative.



Tree Visualization

- Visualized the best-estimated decision tree model to understand the prediction precisely and determine which factor is most effective to the success rate.
- We can see that the most effective factor is grid fins, followed by launch sites and payload mass.





Conclusions

- SpaceX has been improving the success rate gradually.
- I believe that grid fin is the most essential factor for excellent success rate.
- Launch site and payload mass are also effective factors to achieve successful outcomes.
- We can predict the outcome of each launch by using a decision tree classifier model.

Thank you!