IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christophe GRENIER

09/04/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies:
  - Data Collection: SpaceX API, Web Scraping of Falcon 9 Wikipedia
  - Data wrangling and initial exploratory analysis: pandas, sql
  - Data analysis and visualization: Matplotlib, seaborn and plotly graph of data. Creation of a dynamic dashboard through Plotly Dash
  - Data modeling: ML modelization through logistic regression, support vector machine (SVM), decision tree classifiers and k-nearest neighbors (KNN). Each will include a hyperparameter optimization step though GridsearchCV
- In this work, we have come to the following results:
  - Multiple model(s) capable of predicting the success of first stage landing with 83.3% accuracy have been built, despite limited data available. As more data is made available, we expect further improvement especially around lowering number of false positive. These models as well as our exploratory analysis will enable any third party the chance to calculate what the best options to minimize cost of each of their launches are.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- If we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch or make informed choices when using SpaceX.

Section 1

Methodology

Methodology

Executive Summary

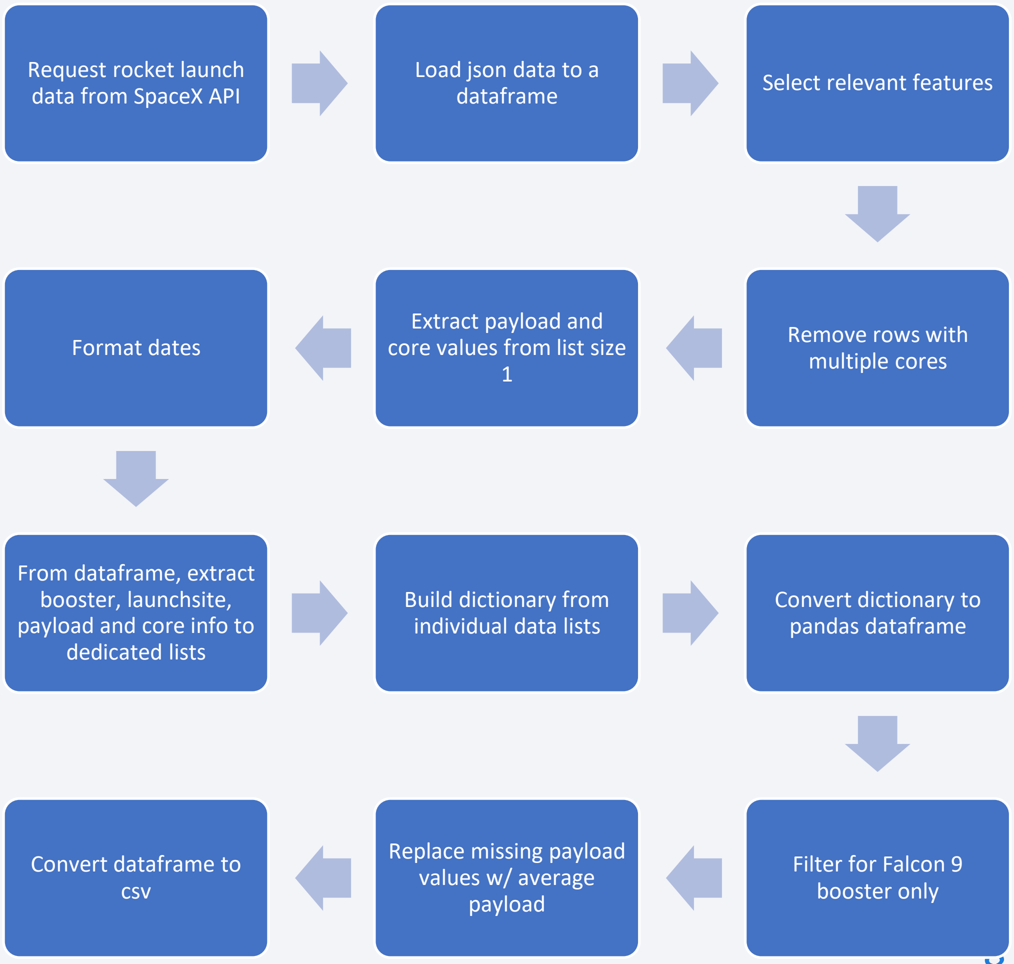
- Data collection methodology:
  - Data was collected by connecting to the SpaceX API and webscraping the Wikipedia page on Falcon 9 launches
- Perform data wrangling
  - Grouping was used to get some insight about distribution of launches across factors. Mission outcomes were one-hot encoded.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - 4 models are evaluated (logistic regression, svm, decision tree and k-nearest neighbors)
  - Each model is optimized through gridsearch on hyperparameters with a 10-fold cross validation and fitting to training data. Precision scores were then calculated using test data

Data Collection

- Data is collected from SpaceX API, then key information is extracted to a dictionary that is thenprocessed into a dataframe
- Data is filtered to look only at Falcon 9 boosters
- To connect information on mission outcomes, we scrape Wikipedia page listing outcomes of Falcon 9 launches using beautiful soup
- Initial look indicate some payload data is missing. Replaced it by average payload value

Data Collection – SpaceX API

- Data collection from SpaceX REST process:



Request rocket launch data from SpaceX API

Load json data to a dataframe

Select relevant features

Remove rows with multiple cores

Extract payload and core values from list size 1

Format dates

From dataframe, extract booster, launchsite, payload and core info to dedicated lists

Build dictionary from individual data lists

Convert dictionary to pandas dataframe

Filter for Falcon 9 booster only

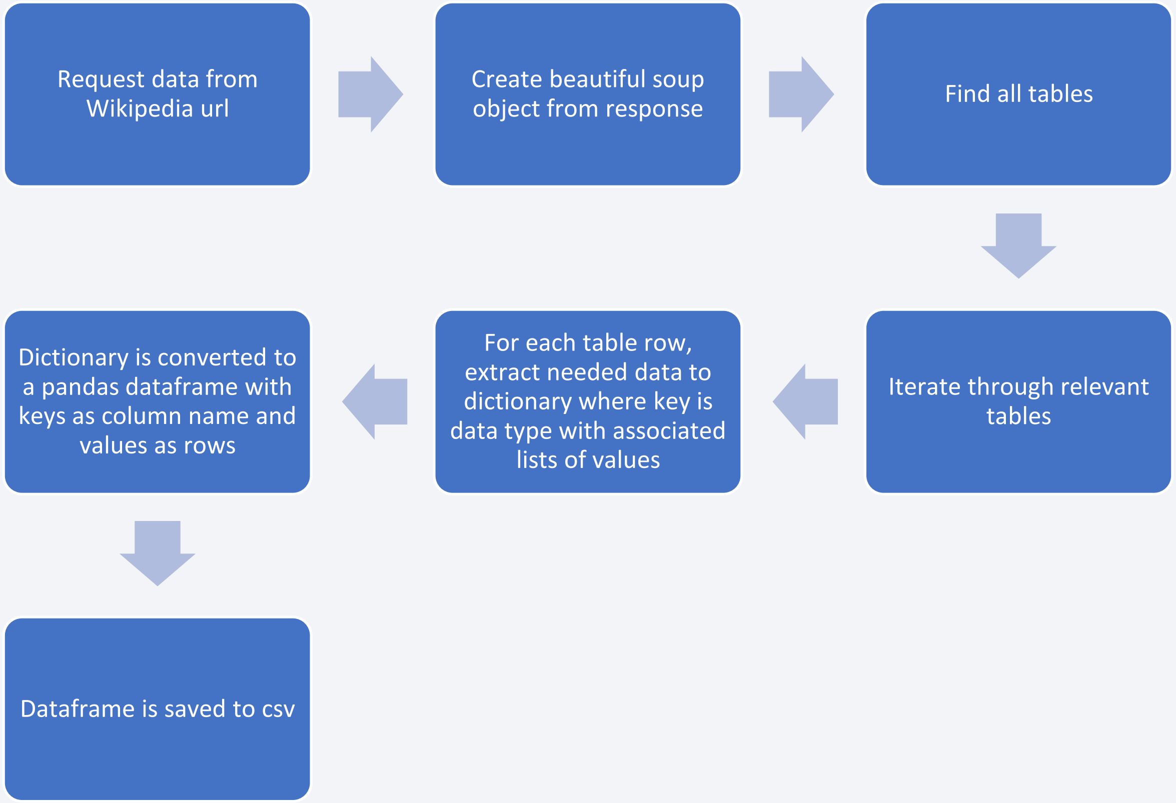
Replace missing payload values w/ average payload

Convert dataframe to csv

- GitHub URL: capstone/Lab 1 spacex-data-collection-api.ipynb at main · cgrenier79/capstone (github.com)

Data Collection - Scraping

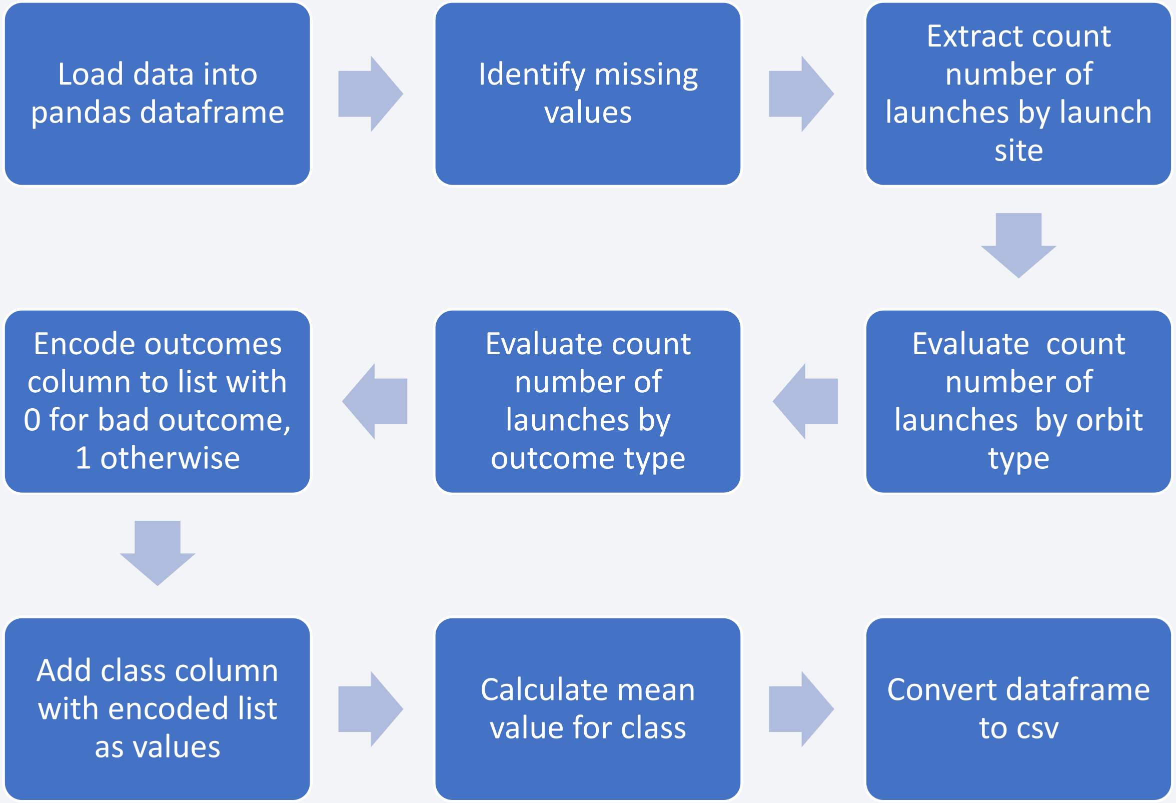
- Web scraping process:


Request data from Wikipedia url
Create beautiful soup object from response
Find all tables
Iterate through relevant tables
For each table row, extract needed data to dictionary where key is data type with associated lists of values
Dictionary is converted to a pandas dataframe with keys as column name and values as rows
Dataframe is saved to csv

- GitHub URL: capstone/lab 1 part 2 webscraping wikipedia falcon 9 launches.ipynb at main · cgrenier79/capstone (github.com)

Data Wrangling

- Data was processed according to present flowchart:
- GitHub URL: [capstone/Lab 2 spacex-Data wrangling.ipynb at main · cgrenier79/capstone (github.com)]



Load data into pandas dataframe

Identify missing values

Extract count number of launches by launch site

Evaluate count number of launches by orbit type

Evaluate count number of launches by outcome type

Encode outcomes column to list with 0 for bad outcome, 1 otherwise

Add class column with encoded list as values

Calculate mean value for class

Convert dataframe to csv

EDA with Data Visualization

- To analyze the data, we plotted the following:
  - Visualize the relationship between Success rate, Payload, Flightnumber through scatterplot
  - Visualize the relationship between Success rate, Flight Number, Launch Site through scatterplot
  - Visualize the relationship between Success rate, Payload Mass and Launch Site through scatterplot
  - Visualize the relationship between Success rate of each Orbit type through barplot
  - Visualize the relationship between Success rate, FlightNumber and Orbit type through scatterplot
  - Visualize the relationship between Success rate Payload Mass and Orbit type through scatterplot
  - Visualize the launch success yearly trend through lineplot

- Add the GitHub URL: capstone/EDA with matplotlib.ipynb at main · cgrenier79/capstone (github.com)

EDA with SQL

- SQL queries performed:
  - %sql SELECT * FROM SPACEXTABLE LIMIT 5
  - %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
  - %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
  - %sql SELECT PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Customer='NASA (CRS)’
  - %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1’
  - %sql SELECT min(Date) FROM SPACEXTABLE WHERE Mission_Outcome='Success’
  - %sql SELECT Booster_Version FROM SPACEXTABLE WHERE ((Landing_Outcome LIKE '%drone ship%') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000))
  - %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
  - %sql SELECT DISTINCT(Booster_Version) FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
  - %sql SELECT substr(Date, 6,2) as 'Month', substr(Date,0,5) as 'Year', substr(Date,9,2) as 'Day', Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE (Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5) = '2015’)
  - %sql SELECT COUNT(Landing_Outcome),Landing_Outcome FROM (SELECT * FROM SPACEXTABLE WHERE ((substr(Date,0,5) || substr(Date,6,2) || substr(Date,9,2)) between '20100604' and '20170320')) GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC

- capstone/EDA analysis SQL data.ipynb at main · cgrenier79/capstone (github.com)

Build an Interactive Map with Folium

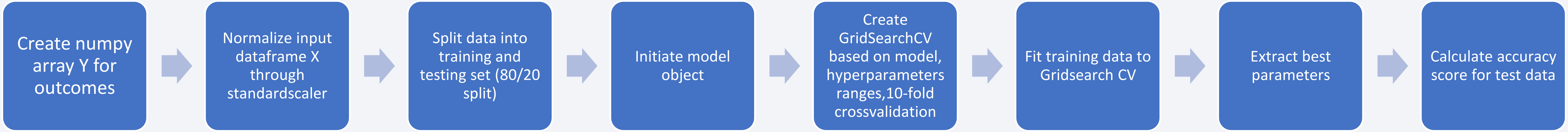
- After creating the Folium map displaying launch sites information the following elements were added:
  - Each site location was marked through circle object with its name as popup label
  - Each launch outcomes at each site were marked through MarkerCluster objects (green colored if launch is successful, red otherwise). These allow us to identify the most successful launch site easily.
  - MousePosition object was added to enable the user to identify position (lat, long) for key locations. This will enable gauging their distance to the launch site. This is key to determine an ideal site.
  - Add Marker and PolyLine objects to show and display the distance to key locations around the map such as nearest city, coastline, highway and railways. Adding these elements and their proximity to a launch site to the map will make the selection of an ideal launch site easy.
- capstone/Lab Folium_launch_site_locations.ipynb at main · cgrenier79/capstone (github.com)

Build a Dashboard with Plotly Dash

- We built a dashboard with Dash with two plots:
  - First plot is a piechart
    - When “All sites” is selected from dropdown the pie chart will show the distribution of launch successes amongst launch sites
    - When a specific launch site is selected, the piechart will show the relative % of launch successes and failures
  - Second plot is a scatterplot
    - When “All sites” is selected from dropdown, the scatterchart will show launch outcome vs payload, colored according to booster type
    - When a specific launch site is selected, the same plot will be shown but limited to selected site
    - A slider is added for the second chart enabling user to select a custom range of payloads
- This dashboard enables us to determine the most promising, and to recommend a range of payload to maximize chance of a successful landing of the first stage
- GitHub URL: https://github.com/cgrenier79/capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- 4 models were evaluated: logistic regression, svm, decision tree and k-nearest neighbors. The following flow was used for each of them:
- GitHub URL: capstone/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · cgrenier79/capstone (github.com)



Create numpy array Y for outcomes

Normalize input dataframe X through standardscaler

Split data into training and testing set (80/20 split)

Initiate model object

Create GridSearchCV based on model, hyperparameters ranges,10-fold crossvalidation

Fit training data to Gridsearch CV

Extract best parameters

Calculate accuracy score for test data

Results

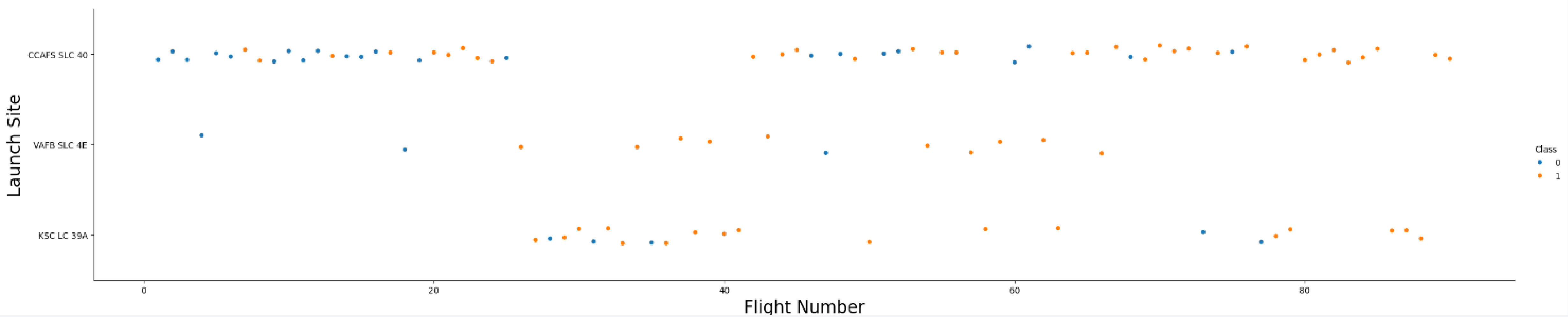
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

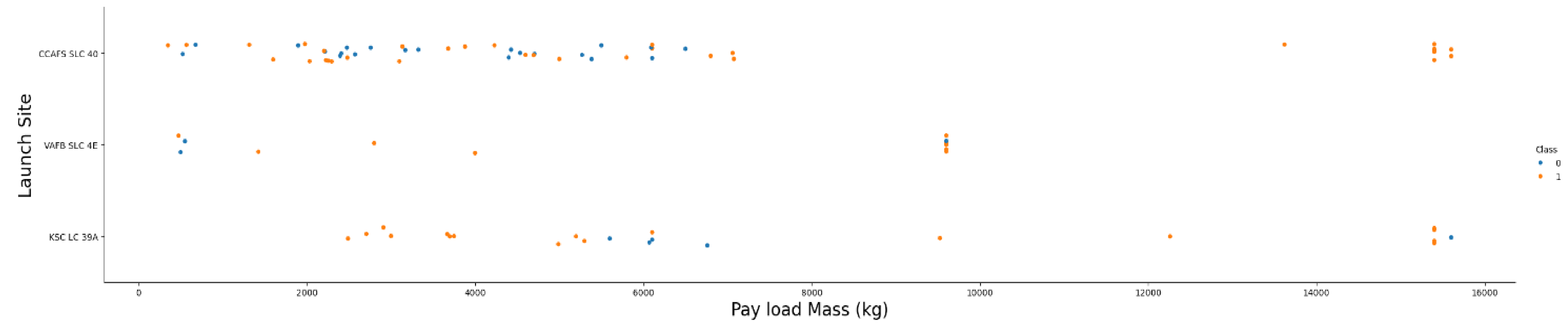
- Scatter plot of Flight Number vs. Launch Site



- Clear trend of success rate improving over time at all sites

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

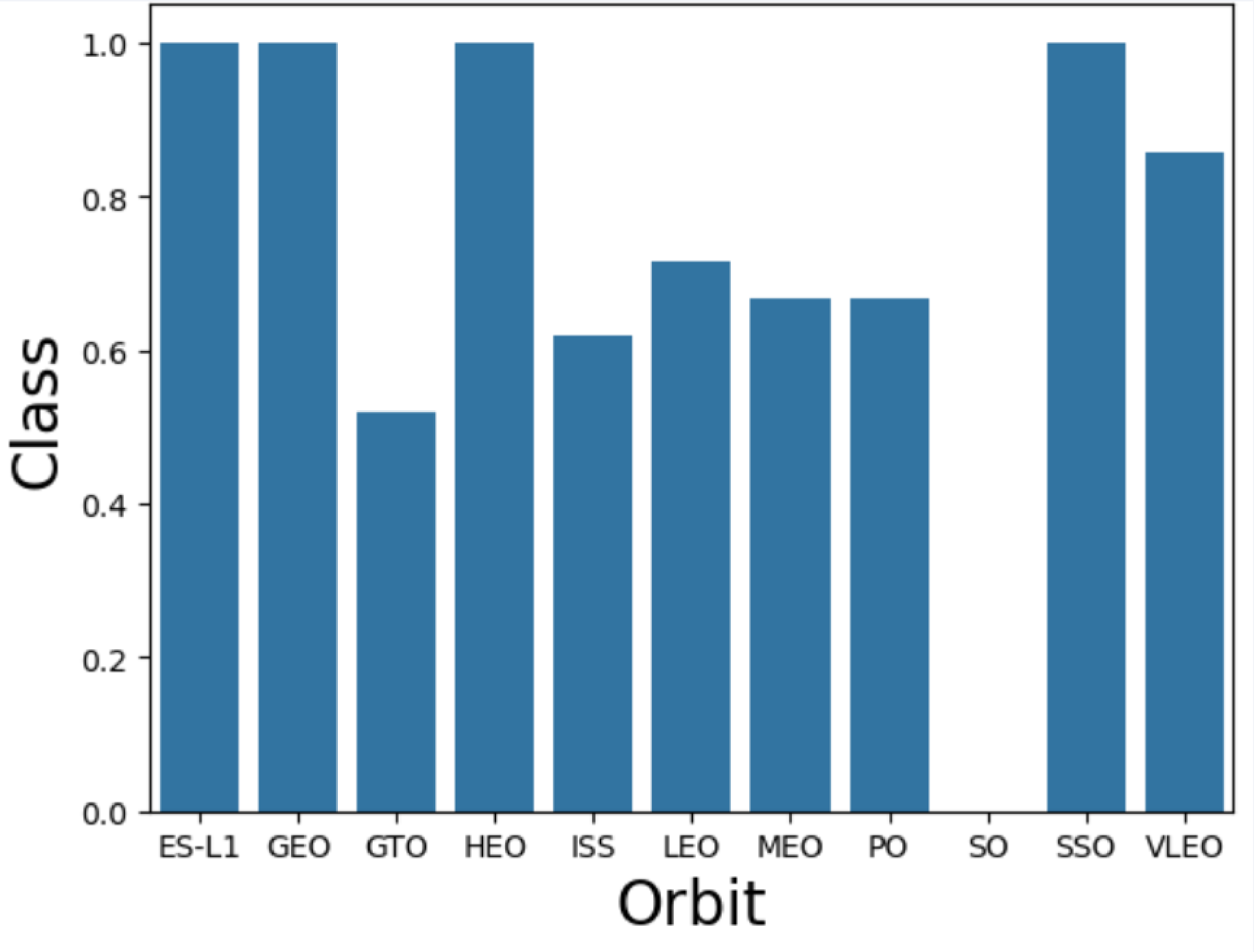


- No heavy payloads (>10000kgs) are launched from VAFB SLC 4E !!

Success Rate vs. Orbit Type

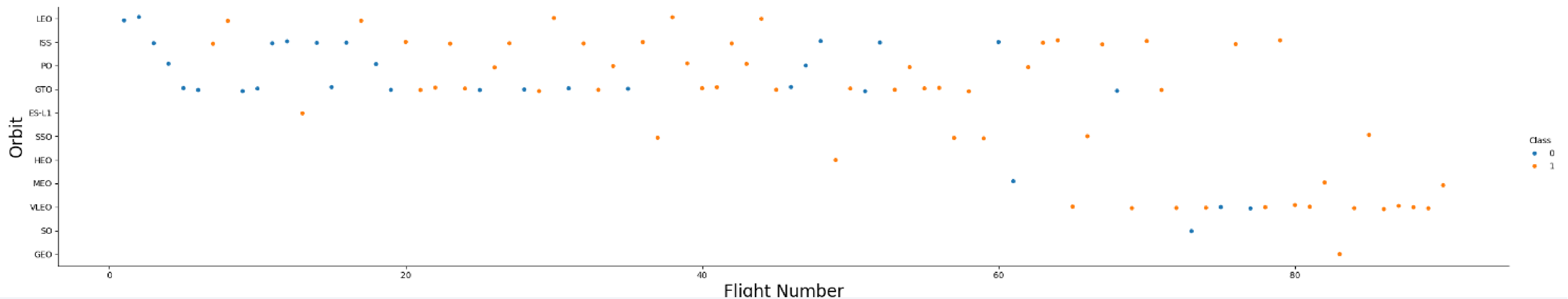
- Bar chart for the success rate of each orbit type

- SSO, HEO, GEO and ES-L1 have perfect records
- VLEO has good success rate as well at 85%
- GTO has only 50% success



Flight Number vs. Orbit Type

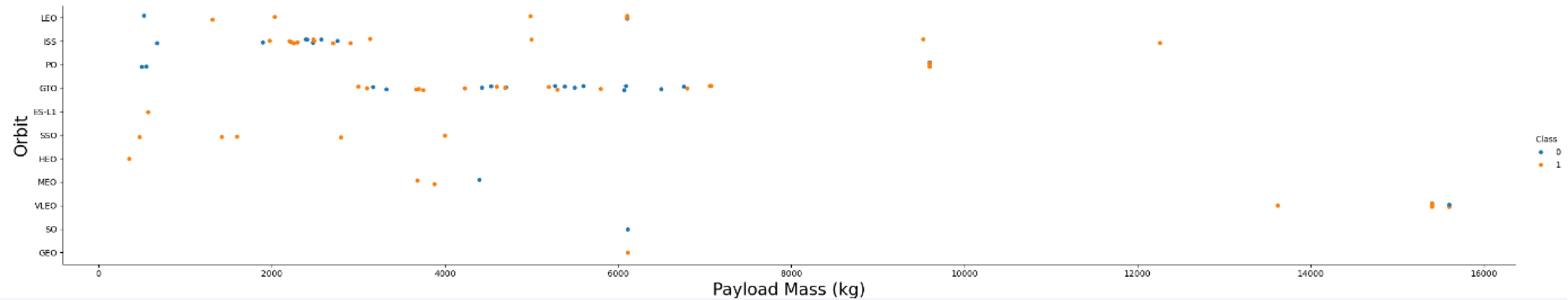
- Scatter point of Flight number vs. Orbit type



- For LEO orbit, the success rate is related to the flight number with later flight all being successful
- For GTO orbit, no clear trend in success rate
- Most recent flight in VLEO orbit have been more numerous than many other orbits with significant successes across latest flights

Payload vs. Orbit Type

- Scatter point of payload vs. orbit type

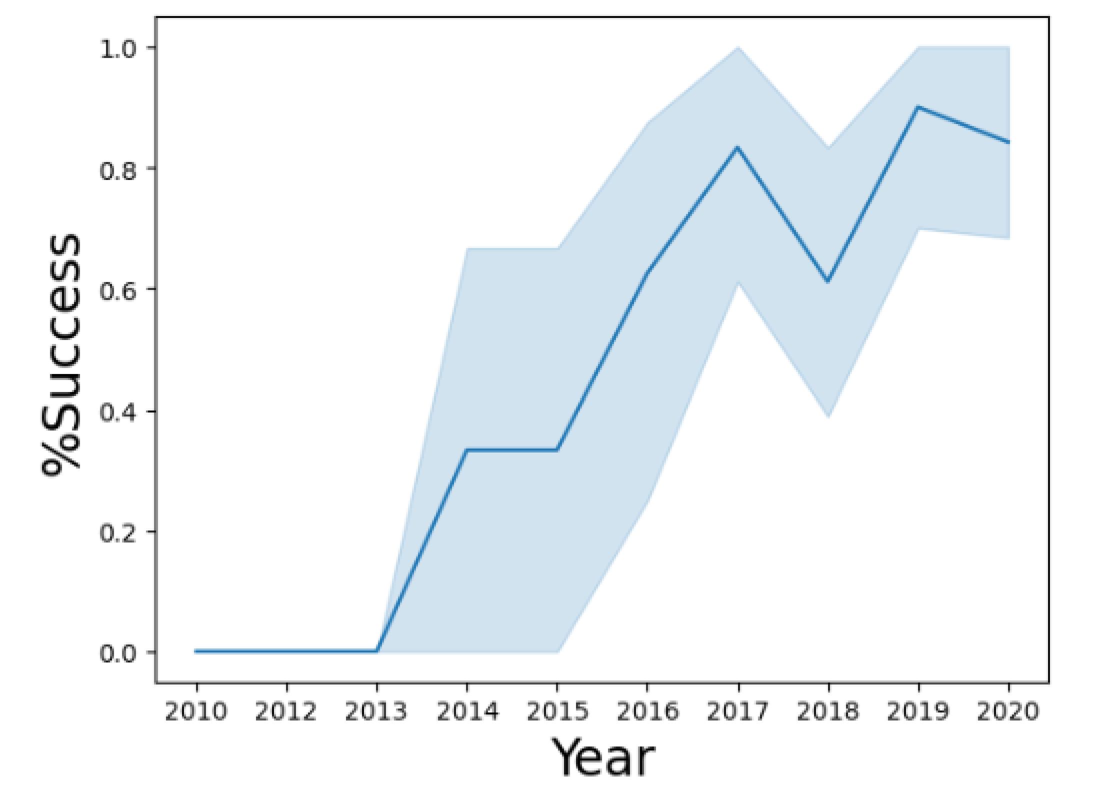


- With heavier payloads higher success for landing for Polar, LEO and ISS orbits
- For GTO orbit, mixed results of success and failure

Launch Success Yearly Trend

- Line chart of yearly average success rate

- Overall success rate improved significantly over time

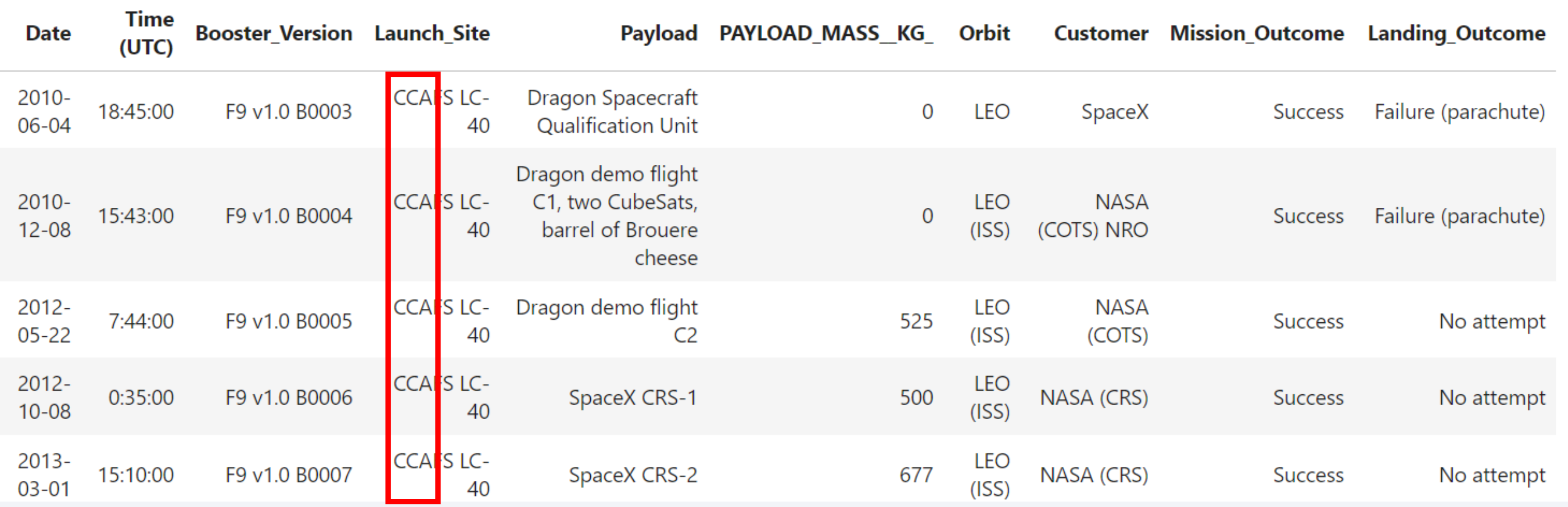


All Launch Site Names

- Names of the unique launch sites: CCAFS LC-40, CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A
- SQL query used:
  - To get unique sites we leverage the DISTINCT function
  - %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
- Results: 4 unique launch sites seen. Two sites are located next to one another (LC-40 and SLC-40)

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`:

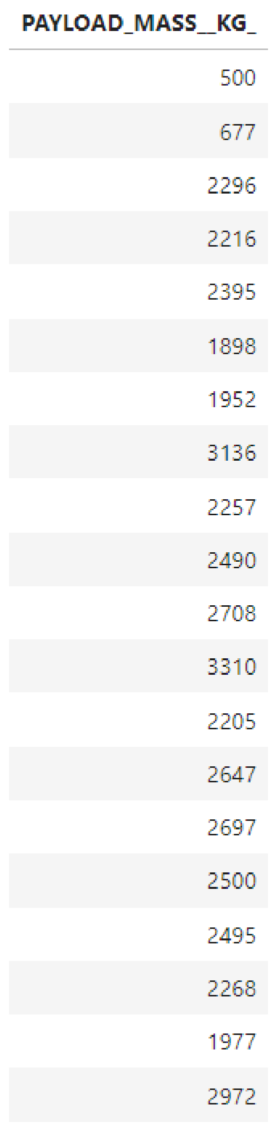


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	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

- SQL query used:
  - Leveraging WHERE….LIKE function for filtering records
  - %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Total Payload Mass

- Total payload carried by boosters from NASA
- Query result used:
  - %sql SELECT PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Customer='NASA (CRS)’
- Results:
  - Minimum payload is 500kgs and maximum payload is 3310kgs



PAYLOAD_MASS__KG_
500
677
2296
2216
2395
1898
1952
3136
2257
2490
2708
3310
2205
2647
2697
2500
2495
2268
1977
2972

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: 2928.4kgs
- Query used:
  - Leveraged AVG function with filtering using WHERE to select proper booster
  - %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'

First Successful Ground Landing Date

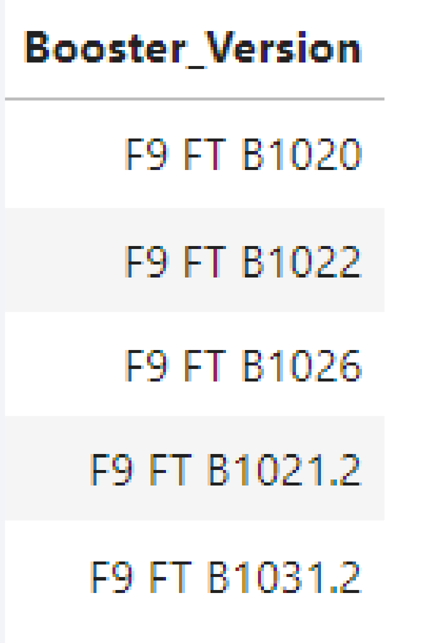
- Dates of the first successful landing outcome on ground pad is 2010-06-04
- Query used:
  - Select minimum function applied to date and records filtered for success of outcome
  - %sql SELECT min(Date) FROM SPACEXTABLE WHERE Mission_Outcome='Success'

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Query used:
  - Used WHERE…..LIKE and WHERE…..BETWEEN to filter based on payload range and landing type
  - %sql SELECT Booster_Version FROM SPACEXTABLE WHERE ((Landing_Outcome LIKE '%drone ship%') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000))

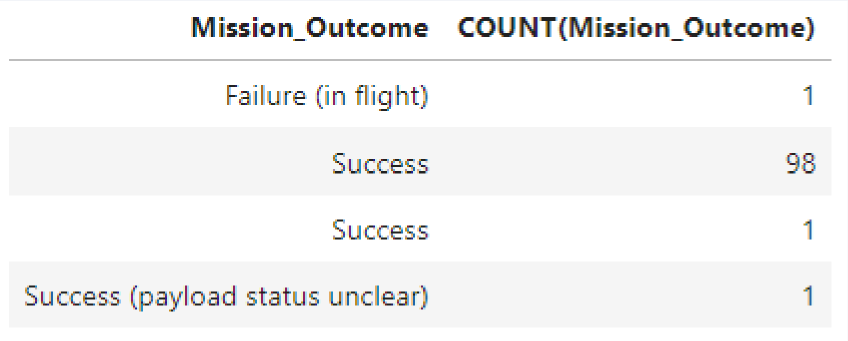
- Results: 5 boosters identified with payload between 4000 and 6000 that landed on drone ships



Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

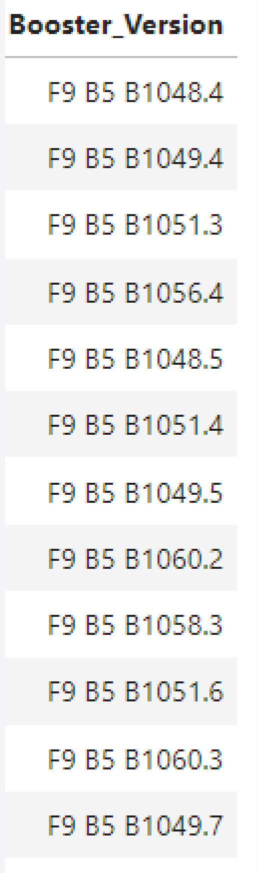
- Calculate the total number of successful and failure mission outcomes
- Query used:
  - Use grouping BY function on mission outcome, and display outcomes and outcomes count
  - %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
- Results:
  - Overall mission is achieved with success rate of > 99%
  - This is separate from landing success!!!!



Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

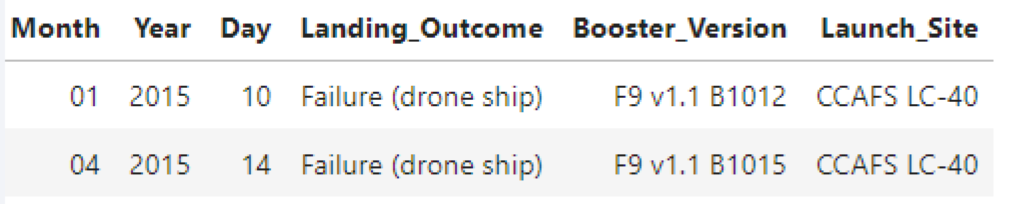
- List the names of the booster which have carried the maximum payload mass
- Query used:
  - Used distinct function on booster type, then leveraged subquery to limit records to only ones that matches the maximum payload
  - %sql SELECT DISTINCT(Booster_Version) FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
- Results:
  - 15 boosters have been used to carry out maximum payload



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

- Query used:
  - Filtered records to find only 2015 launches and where landing on drone ship was a failure
  - Used substr() function on date to select year, month or day
  - %sql SELECT substr(Date, 6,2) as 'Month', substr(Date,0,5) as 'Year', substr(Date,9,2) as 'Day', Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE (Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5) = '2015’)
- Results:
  - 2 records found



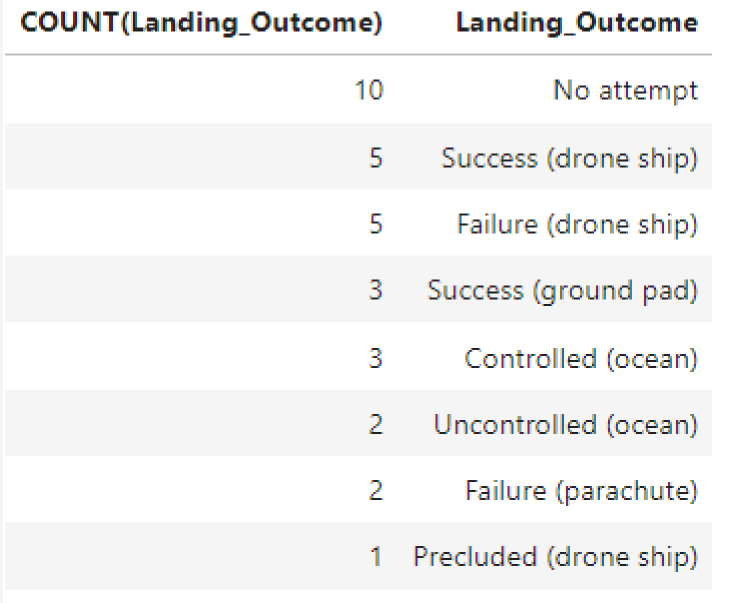
Month	Year	Day	Landing_Outcome	Booster_Version	Launch_Site
01	2015	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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

- Query used:
  - Leverage count function on landing outcomes, concatenated date from SUBSTR() extraction of year, month or date
  - %sql SELECT COUNT(Landing_Outcome),Landing_Outcome FROM (SELECT * FROM SPACEXTABLE WHERE ((substr(Date,0,5) || substr(Date,6,2) || substr(Date,9,2)) between '20100604' and '20170320')) GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC

- Results: most attempts and successes were on drone ships between 20100604 and 201703020



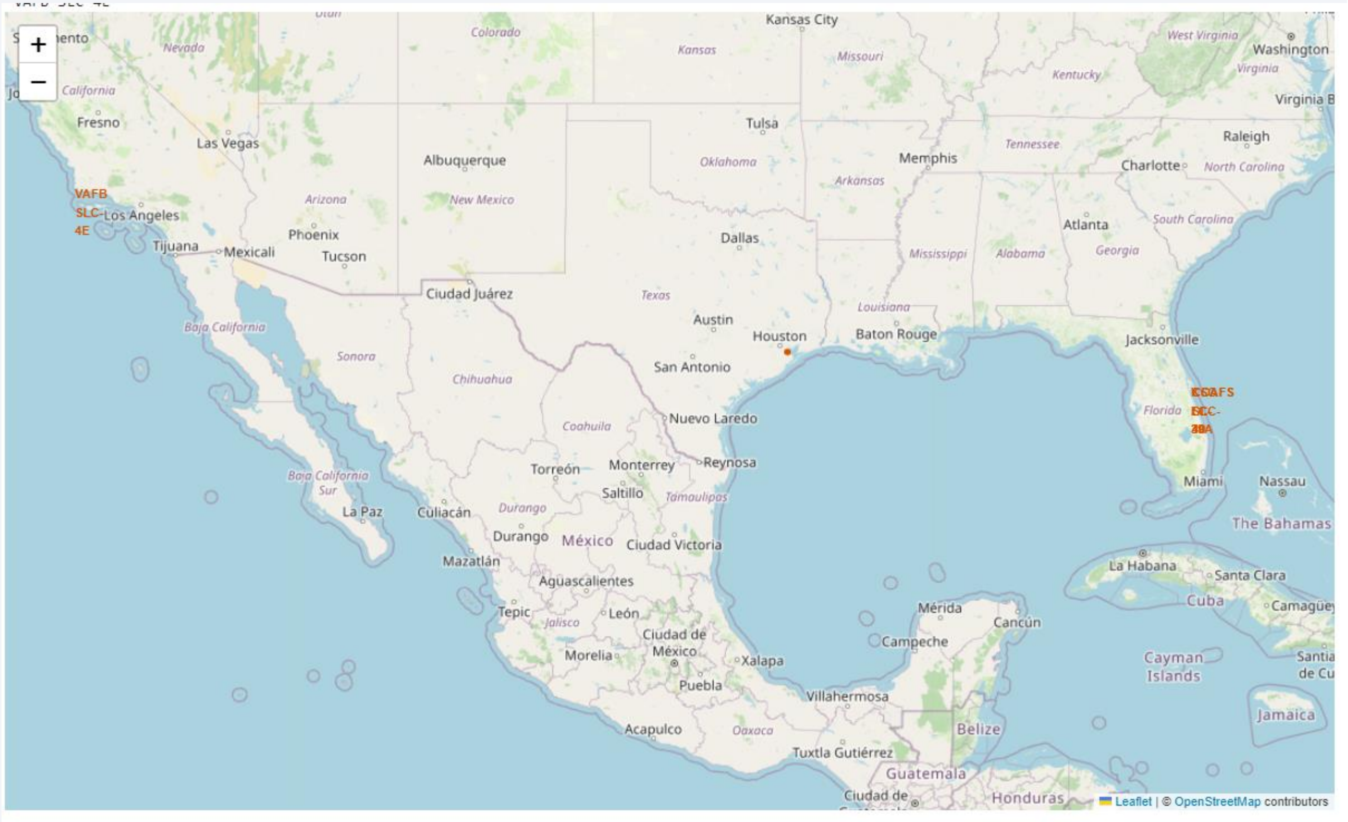
COUNT(Landing_Outcome)	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

Map of SpaceX Launch Sites

- Here we have a terrain map showing all launch sites in red:
  - CCAFS LC-40 (Florida)
  - CCAFS SLC-40 (Florida)
  - KSC LC-39A (Florida)
  - VAFB SLC-4E (California)
- The sites in Florida are too close to each other so labels overlap

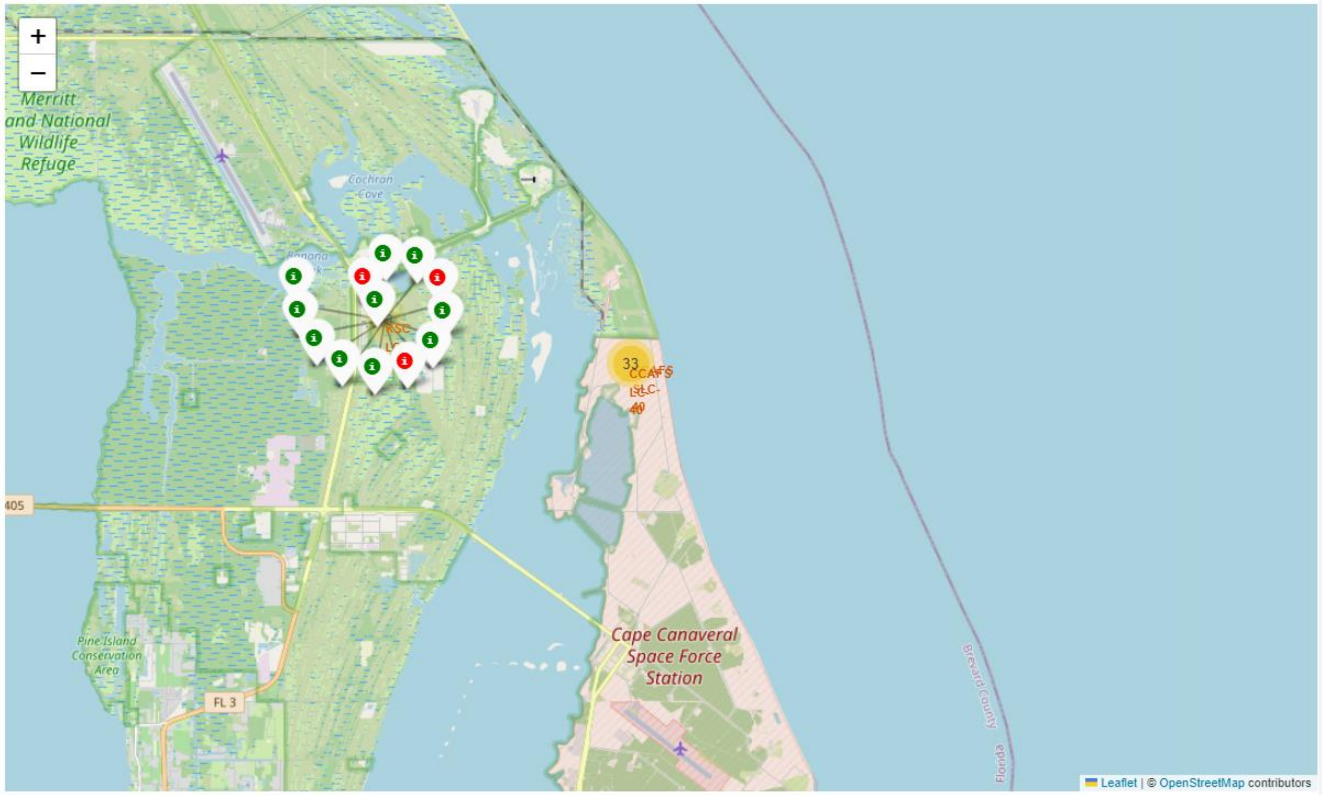


Map w/ Site Markers for Success/Failure for Each Launches

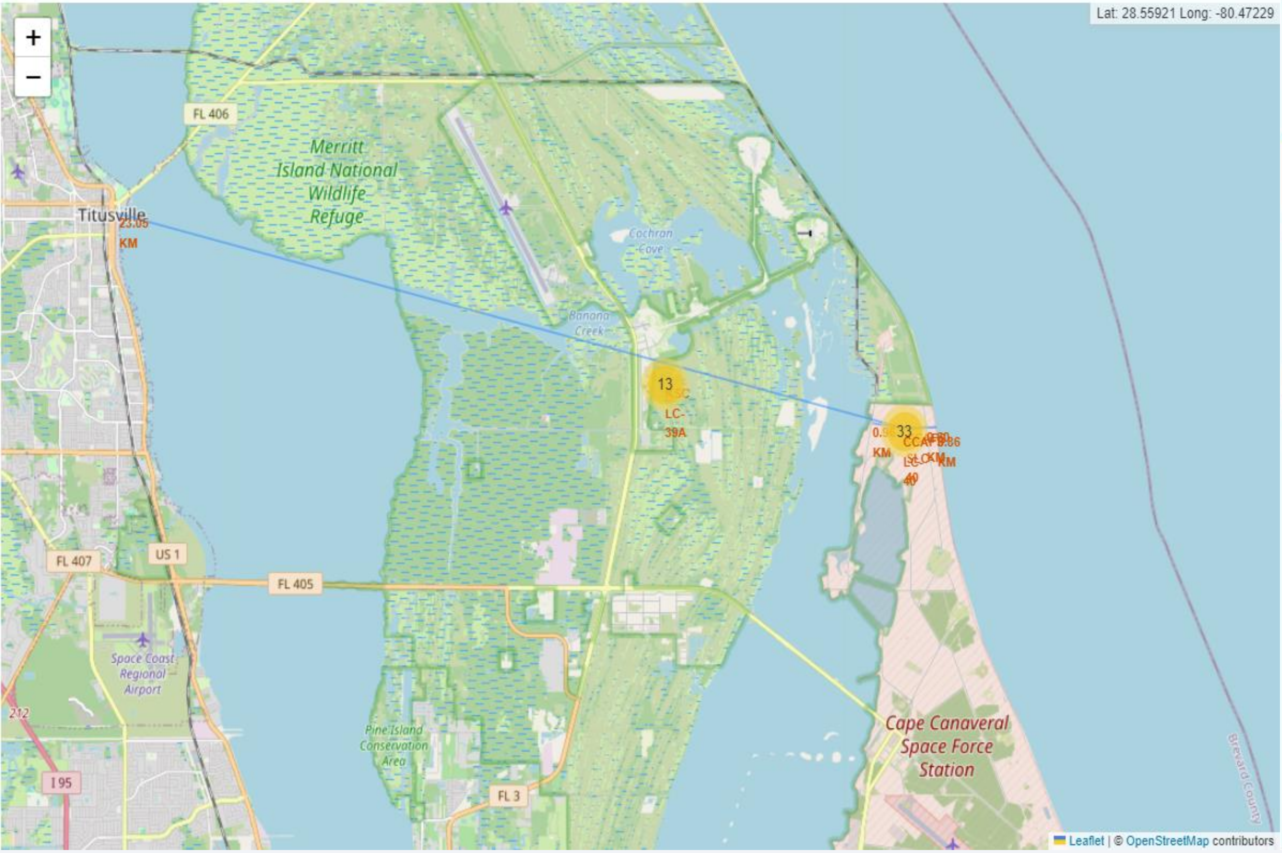
- Clicking on any given sites will provide a set of green (success) and red (Failure) markers

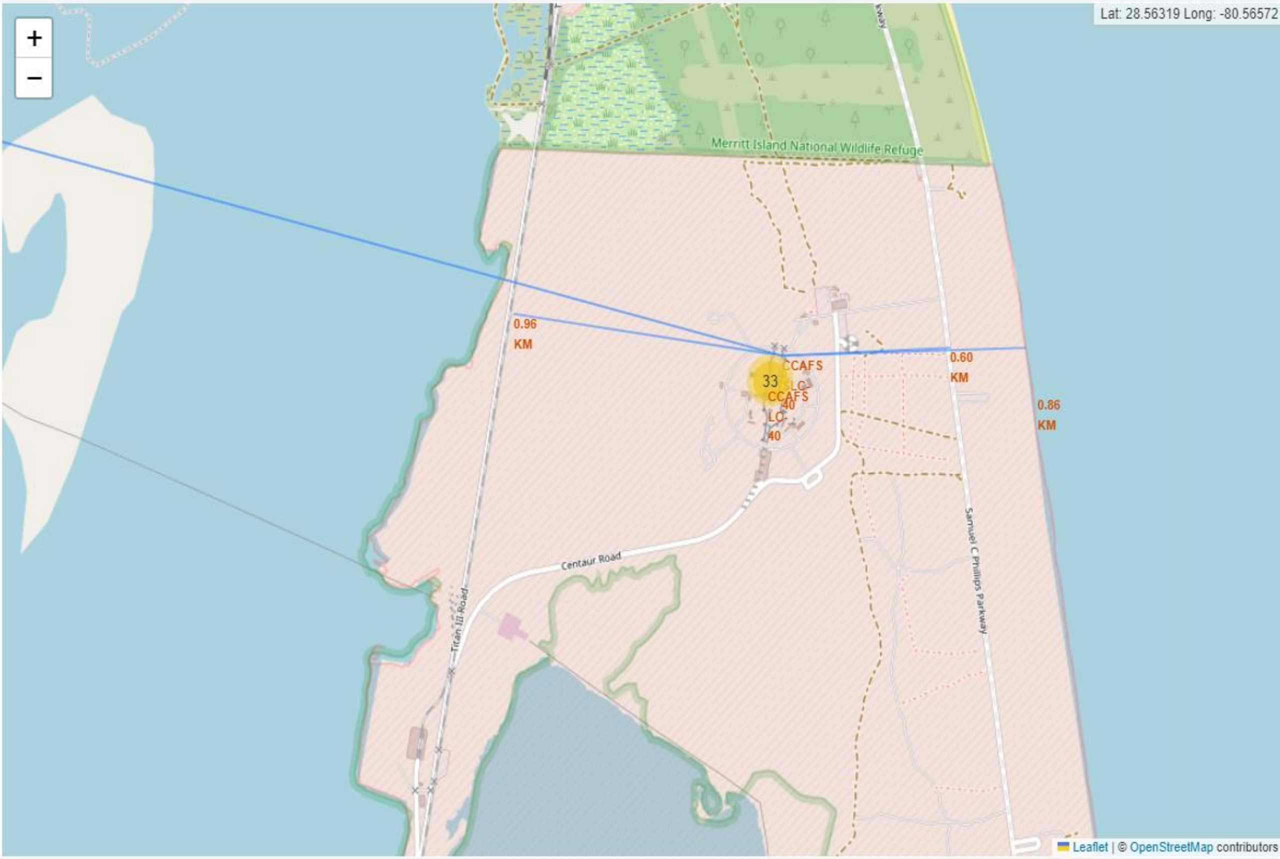
- In screenshot the KSC LC-39A was highlighted. The markers show 13 launches with:
  - 3 failures
  - 10 successes

- The KSC LC-39A has (visually) the most successful launches rate



Map of CCAFS Launch Sites w/ Distances to Key Loacations

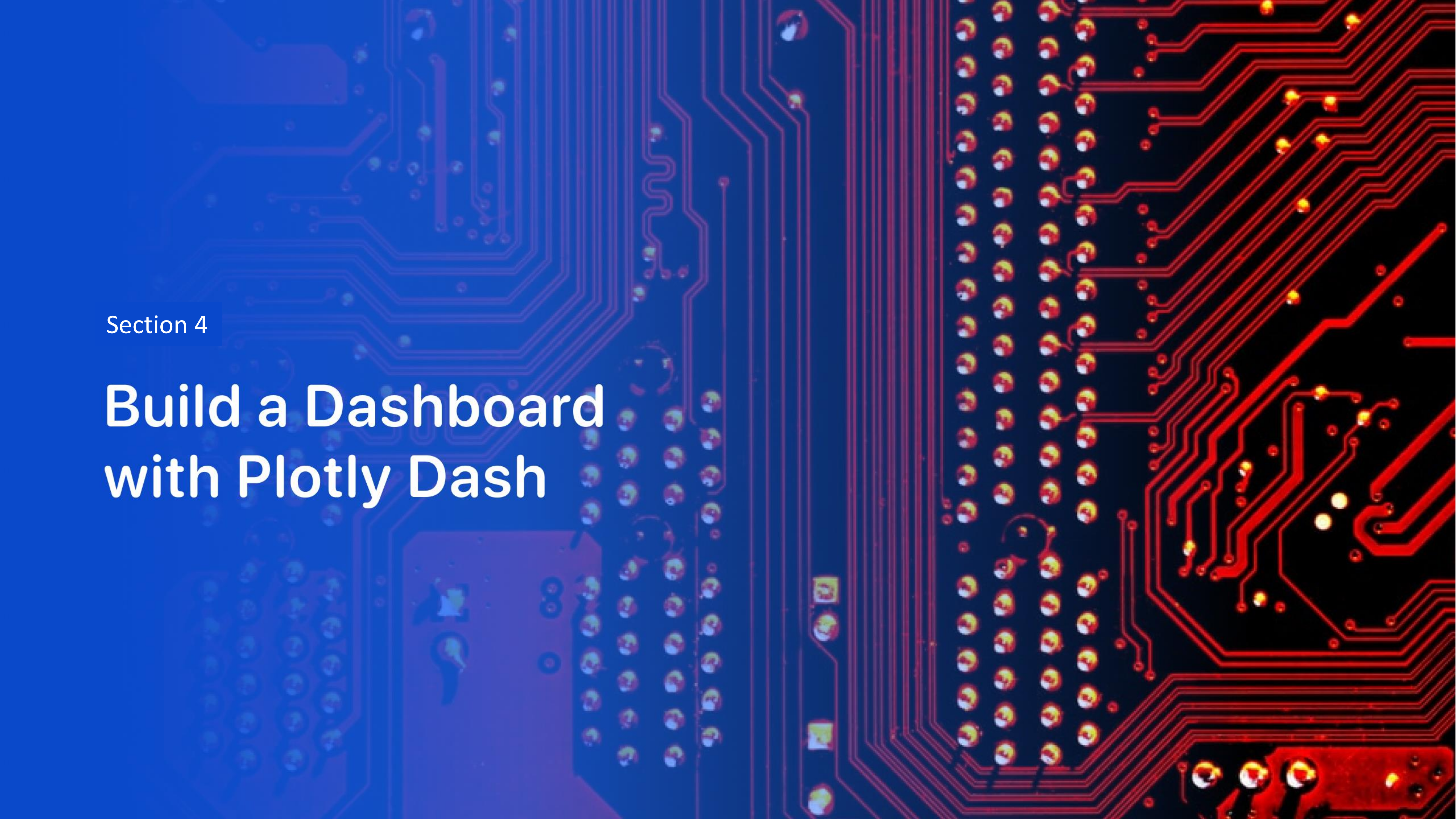

Lat: 28.55921 Long: -80.47229
+
−
Titusville
23.05 KM
FL 406
Merritt Island National Wildlife Refuge
Cochran Cove
Banana Creek
13
LC-39A
33
CCAFS
0.86 KM
LC-40
FL 407
US 1
FL 405
Space Coast Regional Airport
I 95
Pine Island Conservation Area
FL 3
Cape Canaveral Space Force Station
Brevard County
Leaflet | © OpenStreetMap contributors


Lat: 28.56319 Long: -80.56572
+
−
Merritt Island National Wildlife Refuge
0.96 KM
CCAFS
33
LC-40
0.60 KM
0.86 KM
Centaur Road
Titan III Road
Samuel C Phillips Parkway
Leaflet | © OpenStreetMap contributors

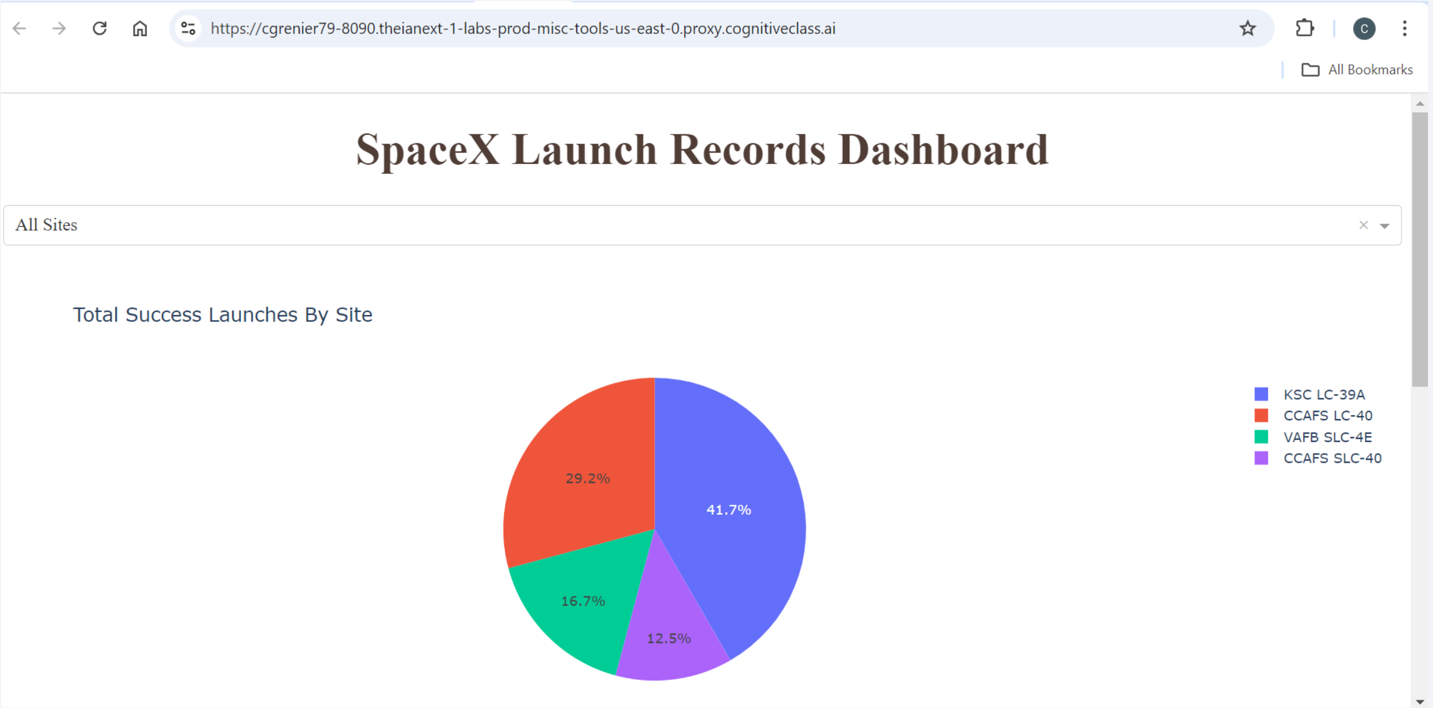
- The drawn line with added distance indicate closest city (Titusville) is ~23km away, closest railway is 0.96km, closest highway is 0.6km and closest coastline is 0.86km

Section 4

Build a Dashboard with Plotly Dash



Total Successful Launches By Site



https://cgrenier79-8090.theianext-1-labs-prod-misc-tools-us-east-0.proxy.cognitiveclass.ai

SpaceX Launch Records Dashboard

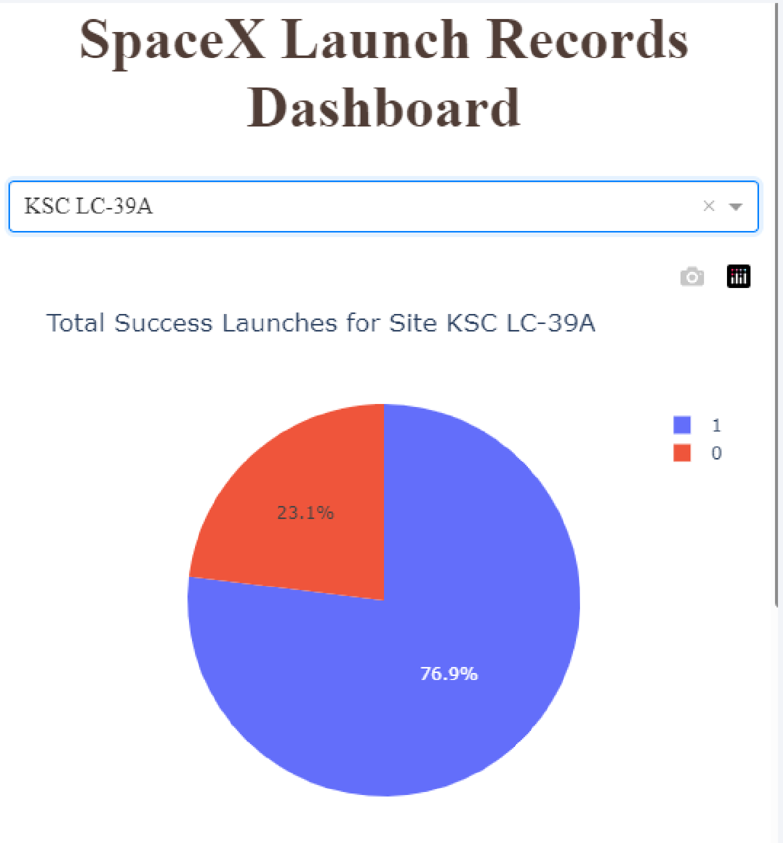
All Sites

Total Success Launches By Site

- All sites selected in drop down to show % of all successes for each sites
- Conclusions: KSC LC-39A provide the highest fraction of all successful landings

Analysis of Success Rate for KSC LC-39A

- Selected KSC LC-39A site which from previous slide had the highest success rate
- The pie chart indicates 76.9% of launches at that site were successful

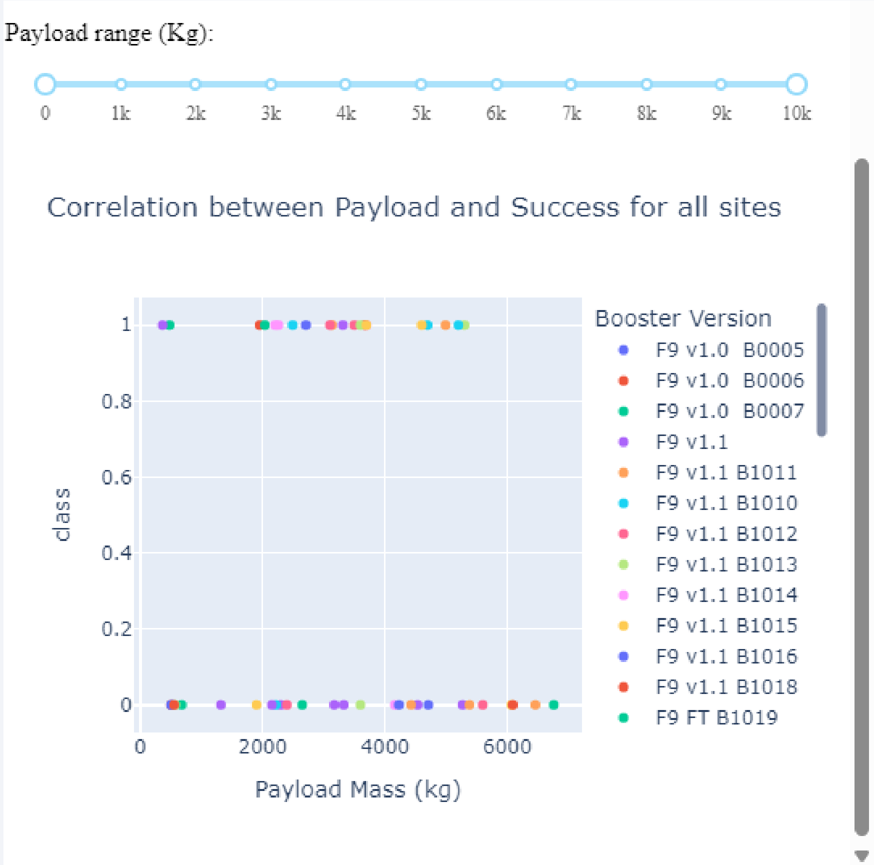


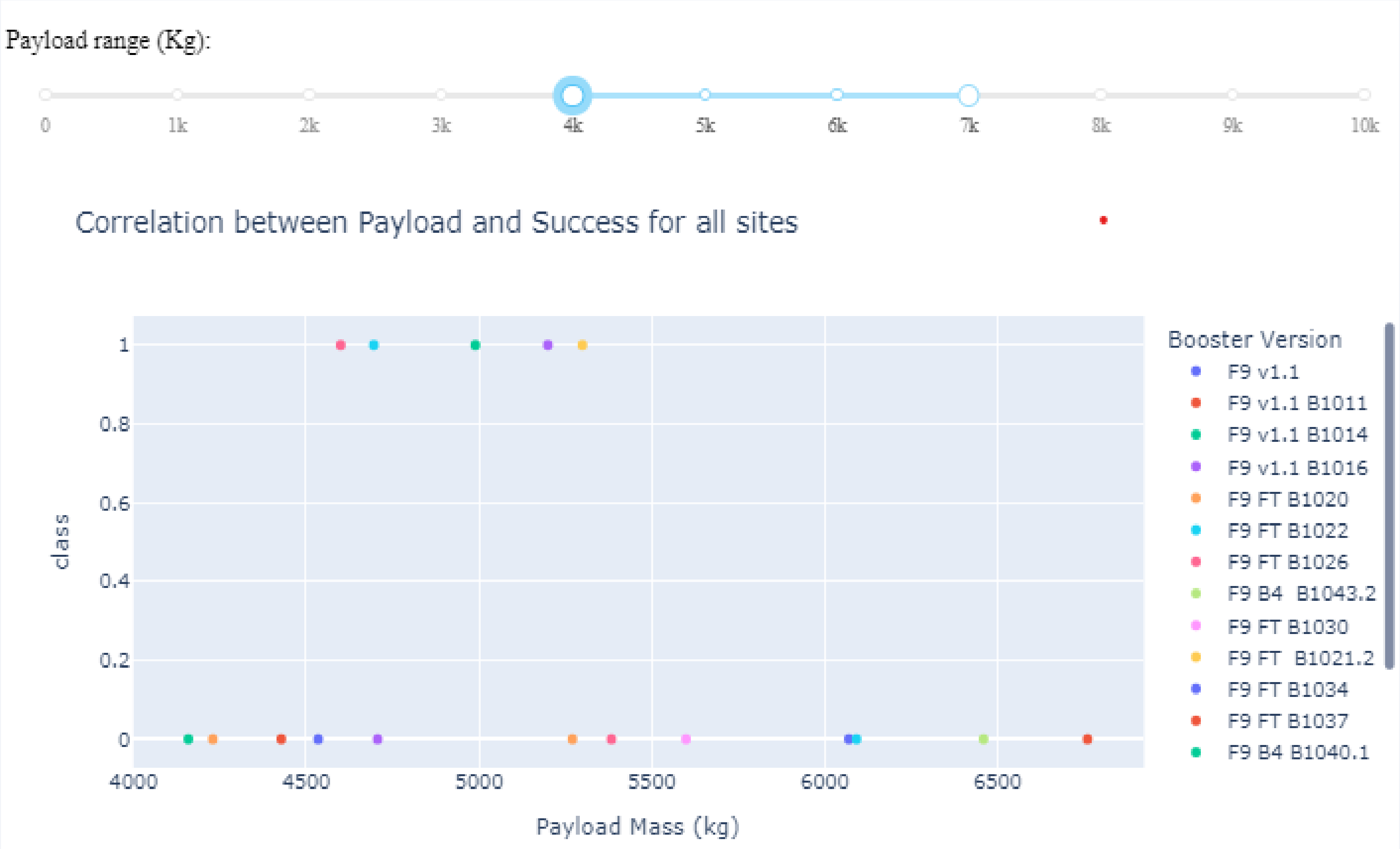
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for Site KSC LC-39A

Correlation between Payload and Success for ALL SITES





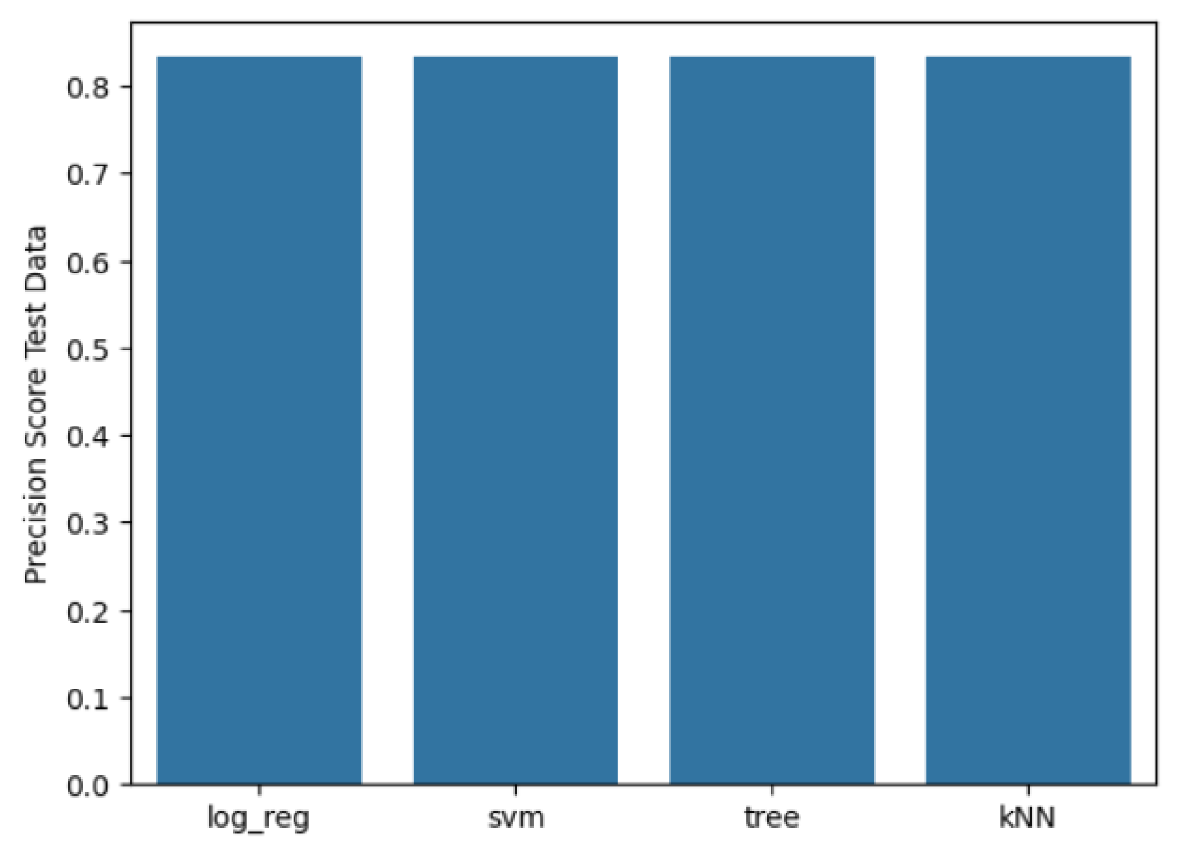
- Clearly, success rates drop drastically above 5500kgs, which is easier to see with the smaller range of payloads

Section 5

Predictive Analysis (Classification)

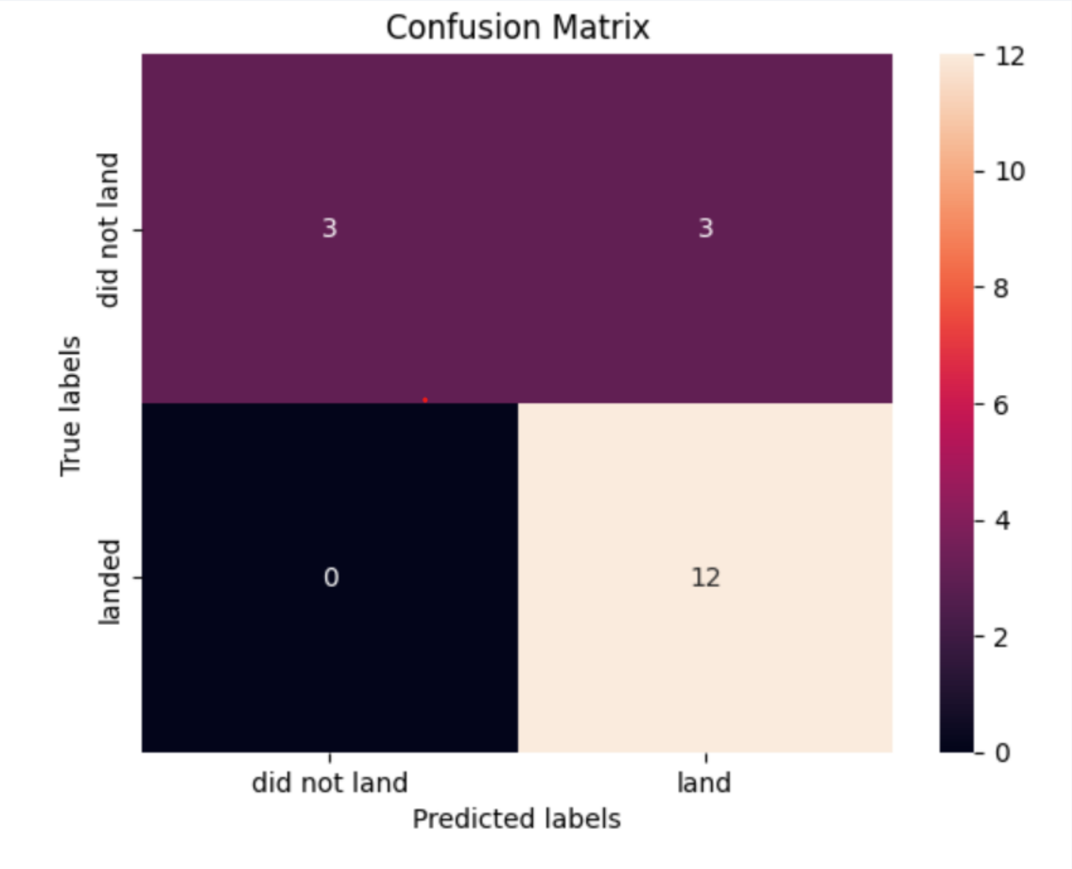
Classification Accuracy

- All models have the same accuracy score after hyperparameters optimization
- This result is likely due to the small size of the dataset
- These models should be re-evaluated as more data becomes available



Confusion Matrix

- The confusion matrix is the same for all 4 model types after hyperparameter optimization:
  - Tree Classifier
  - SVM
  - Logistic Regression
  - kNN neighbors
- 15 correctly classified
- 0 false Negatives
- 3 false Positives



Conclusions

- It appears there is a limit to payload that can be accommodated for successful landing of the first stage
- KSC LC-39A site which from previous slide had the highest success rate
- A model with 83.3% accuracy was built to predict landing outcomes.

Thank you!