IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jihad Akbar
11 March, 2023

Outline

IBM Developer
SKILLS NETWORK

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

Executive Summary

- SpaceY is a new company offering commercial rocket launches and they intend to compete with SpaceX.
- SpaceX offers launch services at a starting price of $62 million, which includes some fuel reserved for landing the first stage rocket booster so that it can be reused.
- Based on public statements made by SpaceX, it is estimated that the cost to build the first stage Falcon 9 booster is upwards of $15 million, not including R&D cost recoupment or profit margin.
- This report was able to use models based on mission parameters such as payload mass and desired orbit to predict the successful landing of the first stage rocket booster with an 83.3% accuracy level.
- SpaceY plans to use these predictions as a way to make more informed bids against SpaceX by using them as a proxy for the cost of a launch.

Introduction: Background

- As part of the Applied Data Science Capstone course, I have created this report where I am playing the role of a data scientist working for SpaceY, a new rocket company.
- The data science findings and models presented in this report will assist SpaceY in making more informed bids against SpaceX for a rocket launch.

Introduction: Business Problem

- SpaceX promotes Falcon 9 rocket launches for a price of $62 million if the first stage can be reused.
- The cost to build the first stage is estimated to be over $15 million, excluding R&D cost recoupment or profit margin.
- However, there are instances where SpaceX will forego the first stage based on factors such as payload, orbit, and customer requirements.
- This report's objective is to predict the probability of the first stage rocket successfully landing as an indicator of the cost of a launch.

Section 1

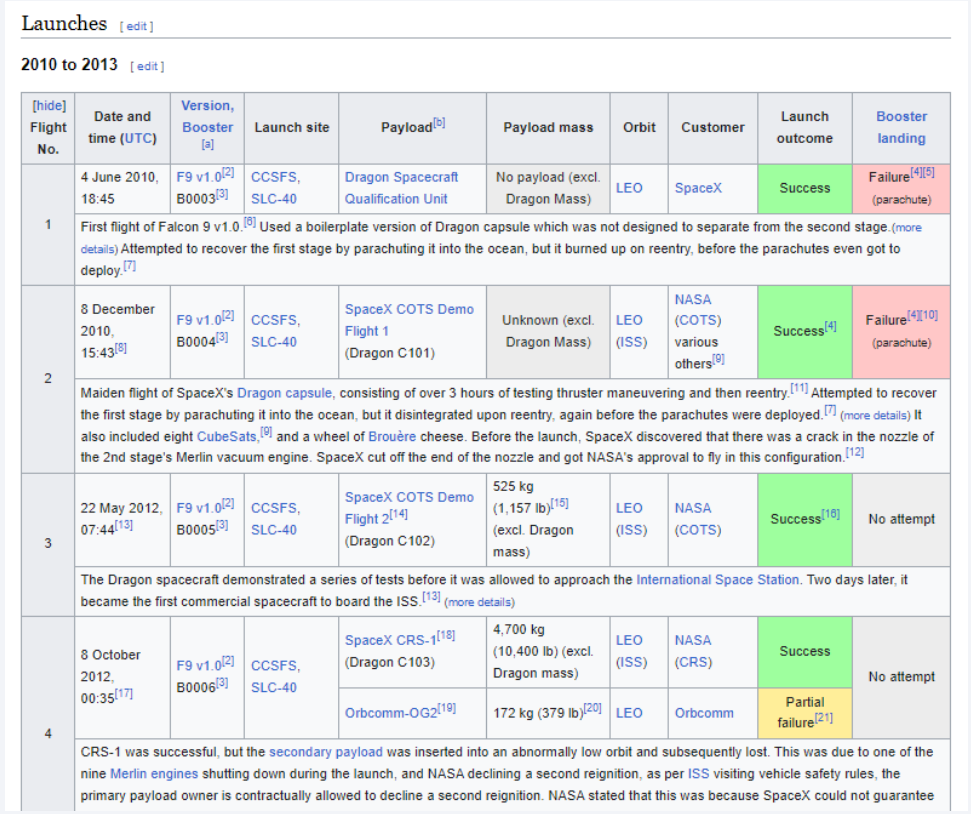
Methodology

Methodology

This report utilized the following data science methodology:

1. Data collection
2. Data wrangling
3. Exploratory data analysis
4. Data visualization
5. Model development
6. Reporting results to stakeholders

Data Collection – SpaceX API and Scraping



Launches [edit]

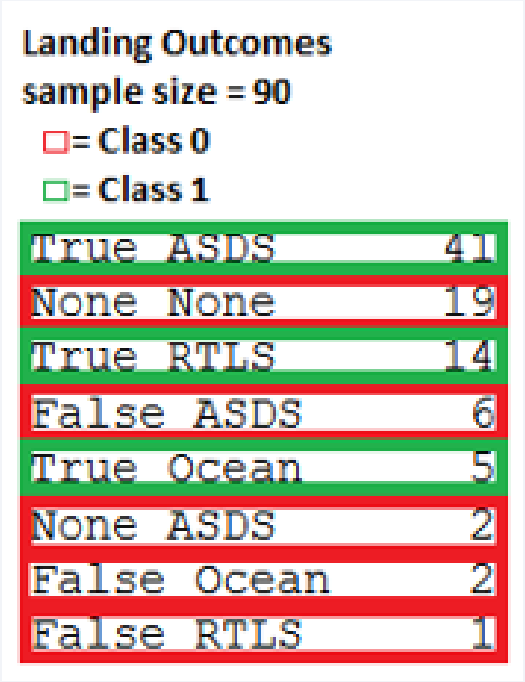
2010 to 2013 [edit]

[hide] Flight No.	Date and time (UTC)	Version, Booster [a]	Launch site	Payload[b]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[2] B0003[3]	CCSFS, SLC-40	Dragon Spacecraft Qualification Unit	No payload (excl. Dragon Mass)	LEO	SpaceX	Success	Failure[4][5] (parachute)
	First flight of Falcon 9 v1.0.[6] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage. (more details) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even got to deploy.[7]								
2	8 December 2010, 15:43[8]	F9 v1.0[2] B0004[3]	CCSFS, SLC-40	SpaceX COTS Demo Flight 1 (Dragon C101)	Unknown (excl. Dragon Mass)	LEO (ISS)	NASA (COTS) various others[9]	Success[4]	Failure[4][10] (parachute)
	Maiden flight of SpaceX's Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and then reentry.[11] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, again before the parachutes were deployed.[7] (more details) It also included eight CubeSats,[9] and a wheel of Brouère cheese. Before the launch, SpaceX discovered that there was a crack in the nozzle of the 2nd stage's Merlin vacuum engine. SpaceX cut off the end of the nozzle and got NASA's approval to fly in this configuration.[12]								
3	22 May 2012, 07:44[13]	F9 v1.0[2] B0005[3]	CCSFS, SLC-40	SpaceX COTS Demo Flight 2[14] (Dragon C102)	525 kg (1,157 lb)[15] (excl. Dragon mass)	LEO (ISS)	NASA (COTS)	Success[16]	No attempt
	The Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[13] (more details)								
4	8 October 2012, 00:35[17]	F9 v1.0[2] B0006[3]	CCSFS, SLC-40	SpaceX CRS-1[18] (Dragon C103)	4,700 kg (10,400 lb) (excl. Dragon mass)	LEO (ISS)	NASA (CRS)	Success	No attempt
				Orbcomm-OG2[19]	172 kg (379 lb)[20]	LEO	Orbcomm	Partial failure[21]	
	CRS-1 was successful, but the secondary payload was inserted into an abnormally low orbit and subsequently lost. This was due to one of the nine Merlin engines shutting down during the launch, and NASA declining a second reignition, as per ISS visiting vehicle safety rules, the primary payload owner is contractually allowed to decline a second reignition. NASA stated that this was because SpaceX could not guarantee								

How the initial launch data appeared on the first page of Wikipedia before it was extracted through web scraping

- API
  - Obtained past launch information for SpaceX from an Open Source REST API
  - Utilized a GET request to retrieve and analyze the SpaceX launch data
  - Narrowed down the dataset to contain only Falcon 9 launches
  - Substituted the absent payload mass values from confidential missions with the mean.
- Web Scraping
  - Obtained the launch data from the Wikipedia page 'List of Falcon 9 and Falcon Heavy Launches'
  - Accessed the Falcon 9 Launch Wiki page by using its Wikipedia URL
  - Extracted the names of all the columns/variables from the header of the HTML table
  - Analyzed the table and transformed it into a Pandas data frame.

Data Wrangling

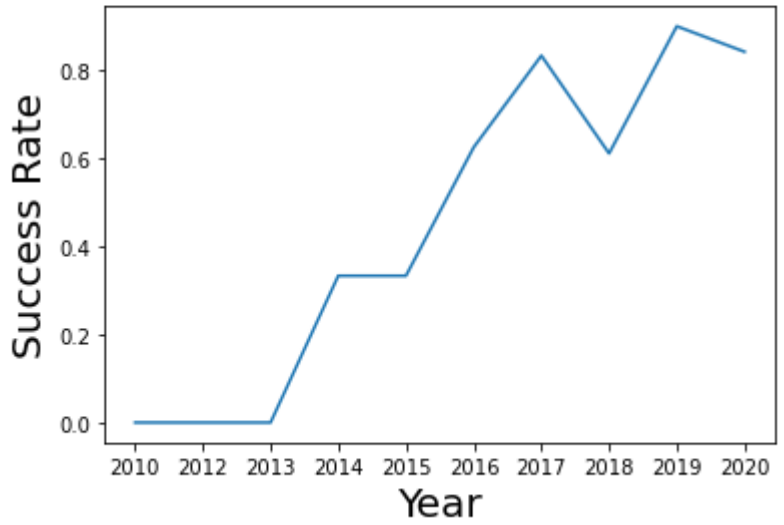


- The data was analyzed to determine the appropriate label for supervised model training.
  - Calculated the number of launches on each site, occurrence of each orbit, and occurrence of mission outcome per orbit type

- Created a landing outcome training label from 'Outcome' column
  - Training label: 'Class'
  - Class = 0: first stage booster did not land successfully
    - None None: not attempted
    - None ASDS: unable to be attempted due to launch failure
    - False ASDS: drone ship landing failed
    - False Ocean: ocean landing failed
    - False RTLS: ground pad landing failed
  - Class = 1: first stage booster landed successfully
    - True ASDS: drone ship landing succeeded
    - True RTLS: ground pad landing succeeded
    - True Ocean: ocean landing succeeded

EDA with SQL and Data Visualization

```
In [11]: # Plot a line chart with x axis to be the extracted
         df1=pd.DataFrame(Extract_year(df['Date']),columns =[
         df1['Class']=df['Class']
         sns.lineplot(data=df1, x=np.unique(Extract_year(df['
         plt.xlabel("Year", fontsize=20)
         plt.ylabel("Success Rate", fontsize=20)
         plt.show()
```



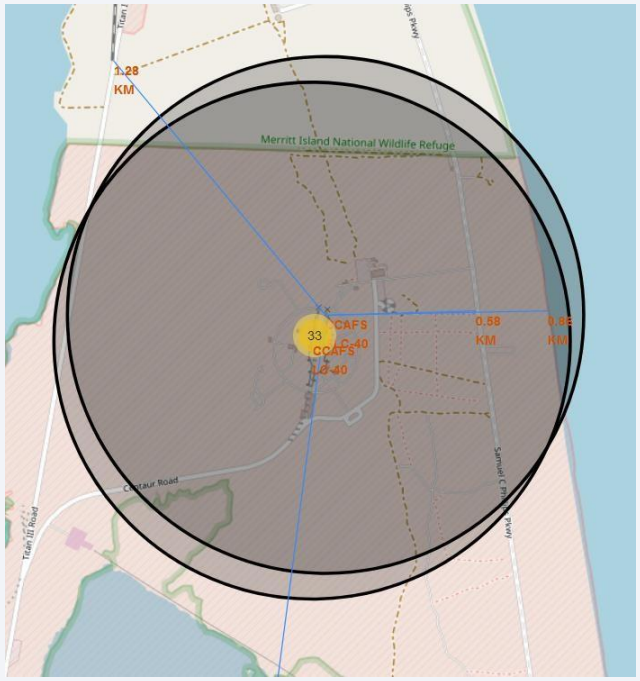
screenshot of Year vs. Success rate plot

- EDA with SQL
  - Loaded data into an IBM DB2 instance
  - Ran SQL queries to display and list information about:
    - Launch sites
    - Payload masses
    - Booster versions
    - Mission outcomes
    - Booster landings

- EDA with visualization
  - Read the dataset into a Pandas dataframe
  - Used Matplotlib and Seaborn visualization libraries to plot:
    - FlightNumber x PayloadMass †
    - FlightNumber x LaunchSite †
    - Payload x LaunchSite †
    - Orbit type x Success rate
    - FlightNumber x Orbit type †
    - Payload x Orbit type †
    - Year x Success rate

† = with Class overlayed (1st stage booster landing outcome)

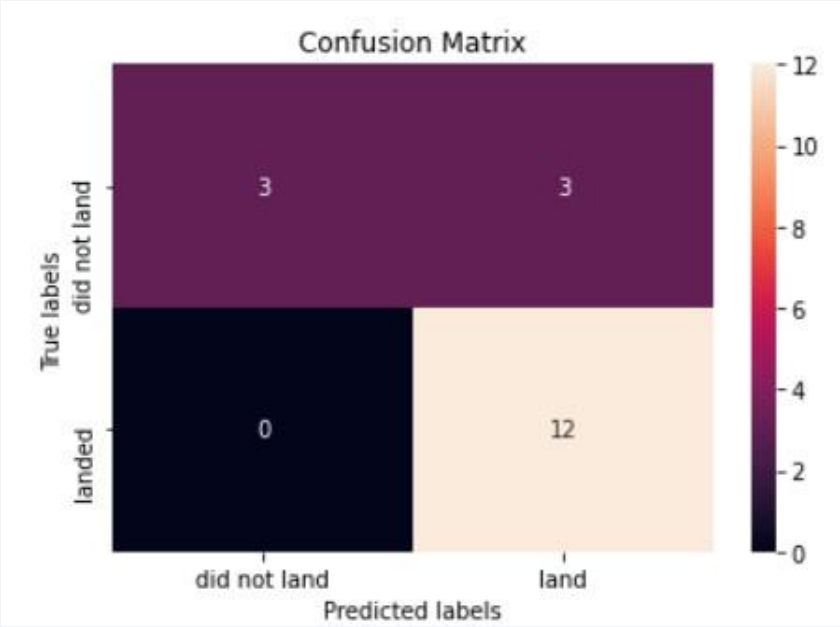
Build an Interactive Map with Folium and Dash



Screenshot of interactive Folium map showing proximity from CCAFS SLC-40 launch site to nearby railway, highway, and coastline

- Data Visualization with Folium
  - Used Python interactive mapping library called Folium
  - Marked all launch sites on a map
  - Marked the successful/failed launches for each site on map
  - Calculated the distances between a launch site to its proximities:
    - Railways
    - Highways
    - Coastlines
    - Cities
- Launch Records Dashboard using Dash
  - Used Python interactive dashboarding library called Plotly Dash to enable stakeholders to explore and manipulate data in an interactive and real-time way
  - Pie chart showing success rate
    - Color coded by launch site
  - Scatter chart showing payload mass vs. landing outcome
    - Color coded by booster version
    - With range slider for limiting payload amount
  - Drop-down menu to choose between all sites and individual launch sites

Predictive Analysis (Classification)



Confusion matrix of logistic regression model, showing 15 correct predictions and 3 false positives

- Imported libraries and defined function to create confusion matrix
  - Pandas
  - Numpy
  - Matplotlib
  - Seaborn
  - Sklearn

- Loaded the dataframe created during data collection
- Created a column for our training label 'Class' created during data wrangling
- Standardized the data
- Split the data into training data and test data

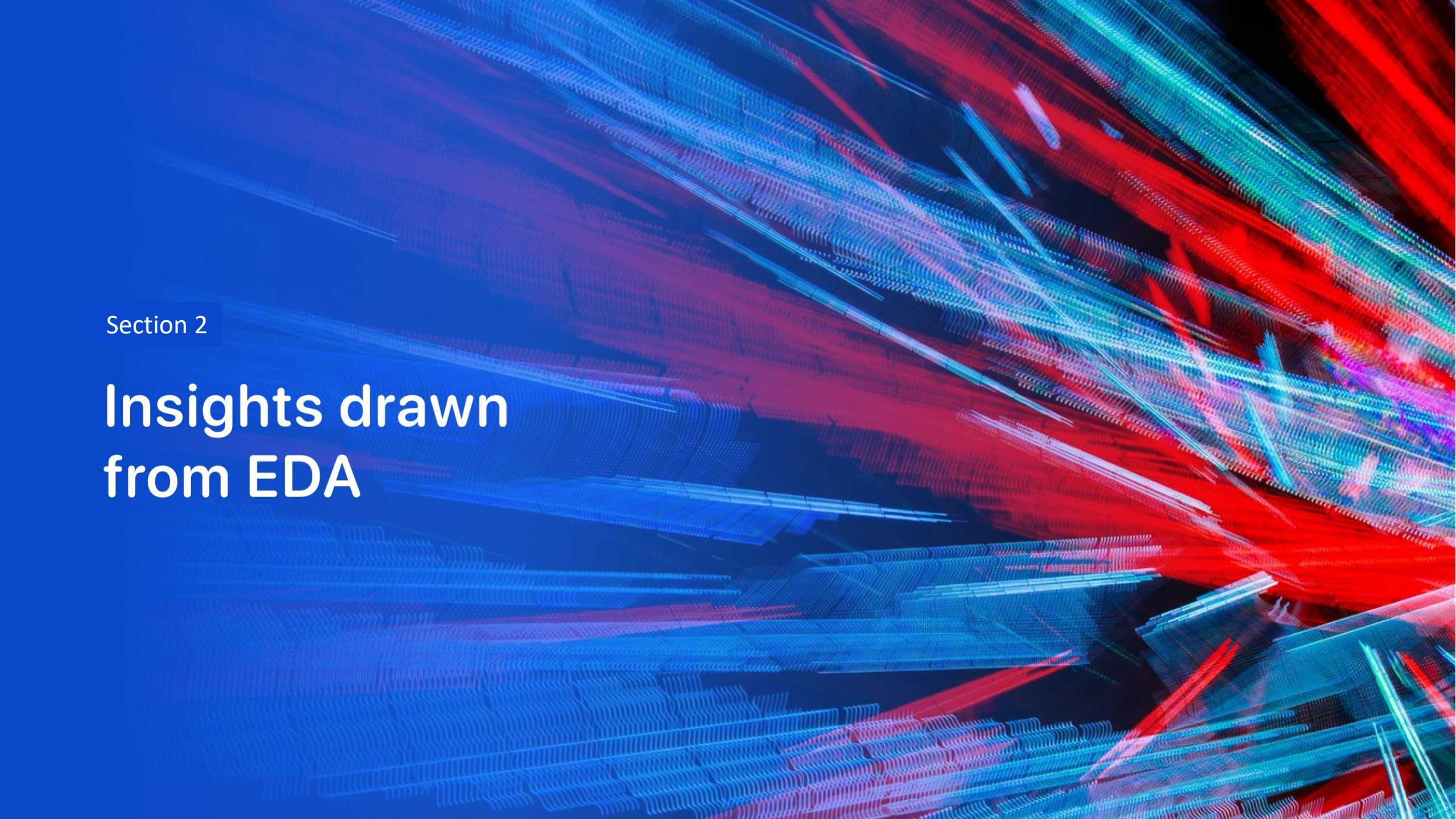
- Fit the training data to various model types:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors Classifier

- Used a cross-validated grid-search over a variety of hyperparameters to select the best ones for each model
  - Enabled by Scikit-learn library function GridSearchCV

- Evaluated accuracy of each model using test data to select the best model

Results

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


Section 2
Insights drawn from EDA

Results: EDA with SQL

The team at SpaceY had some very specific questions to answer with SQL:

- What launch sites has SpaceX used?
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

- Examine launch site and date records where launch sites begin with the string 'CCA', do they overlap?
  - Last launch from CCAFS LC-40 was 2016-08-14
  - First launch from CCAFS SLC-40 was 2017-12-15
  - Wikipedia confirms Cape Canaveral Space Launch Complex 40 was renamed in 2017

- Display the total payload mass carried by boosters launched by NASA (CRS)
  - 45,596 KG, total

- Display average payload mass carried by booster version F9 v1.1
  - 340 KG, average

- List the date when the first successful landing outcome in ground pad was achieved.
  - 2015-12-22, more than 5 years after the first Falcon 9 launch on 2010-06-04

Results: EDA with SQL (continued)

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

- F9 FT B1021.1
- F9 FT B1023.1
- F9 FT B1029.2
- F9 FT B1038.1
- F9 B4 B1042.1
- F9 B4 B1045.1
- F9 B5 B1046.1

Rank the count of landing outcomes between the date 2010-06- 04 and 2017-03-20, in descending order

- 10 - No attempt5 - Failure (drone ship)
- 5 - Success (drone ship)
- 3 - Controlled (ocean)
- 3 - Success (ground pad)
- 2 - Failure (parachute)
- 2 - Uncontrolled (ocean)
- 1 - Precluded (drone ship)

List the names of the booster versions which have carried the maximum payload mass.

- F9 B5 B1048.4
- F9 B5 B1048.5
- F9 B5 B1049.4
- F9 B5 B1049.5
- F9 B5 B1049.7
- F9 B5 B1051.3
- F9 B5 B1051.4
- F9 B5 B1051.6
- F9 B5 B1056.4
- F9 B5 B1058.3
- F9 B5 B1060.2
- F9 B5 B1060.3

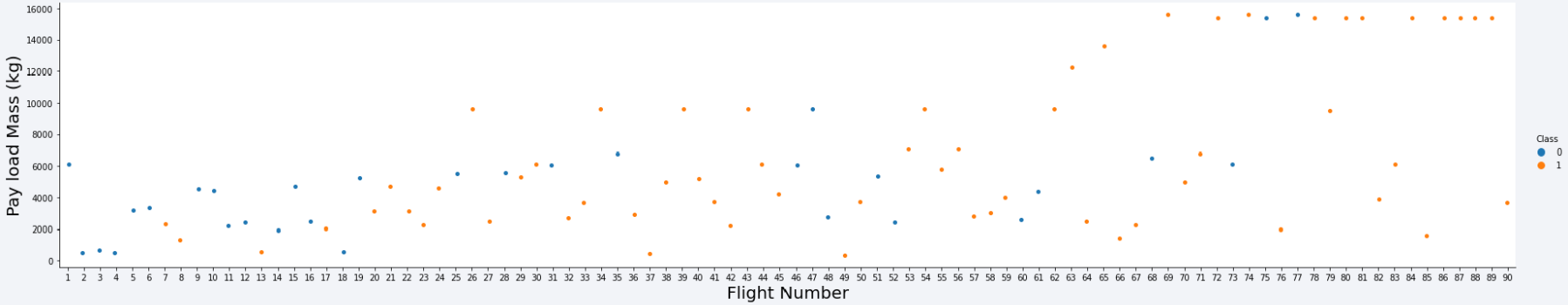
List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Failure (drone ship), F9 v1.1 B1012, CCAFS LC-40
- Failure (drone ship), F9 v1.1 B1015, CCAFS LC-40

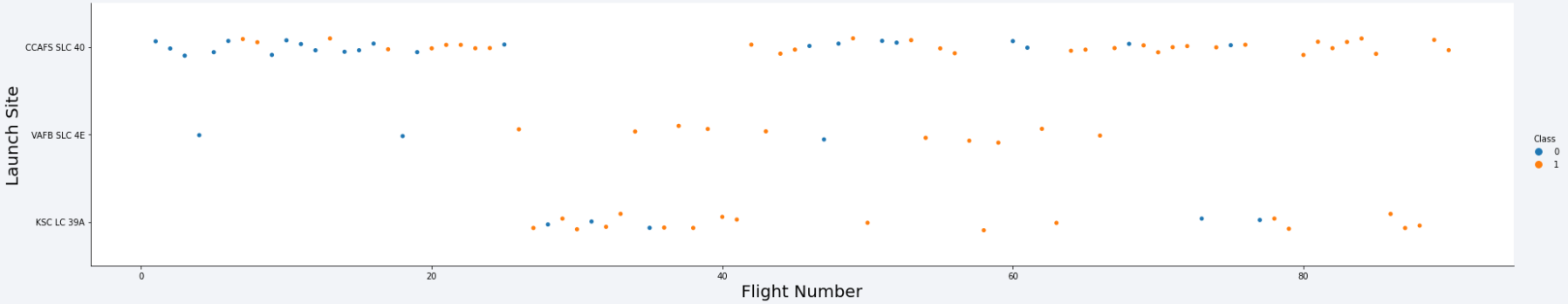
List the total number of successful and failure mission outcomes

- 1 - Failure (in flight)
- 99 - Success
- 1 - Success (payload status unclear)

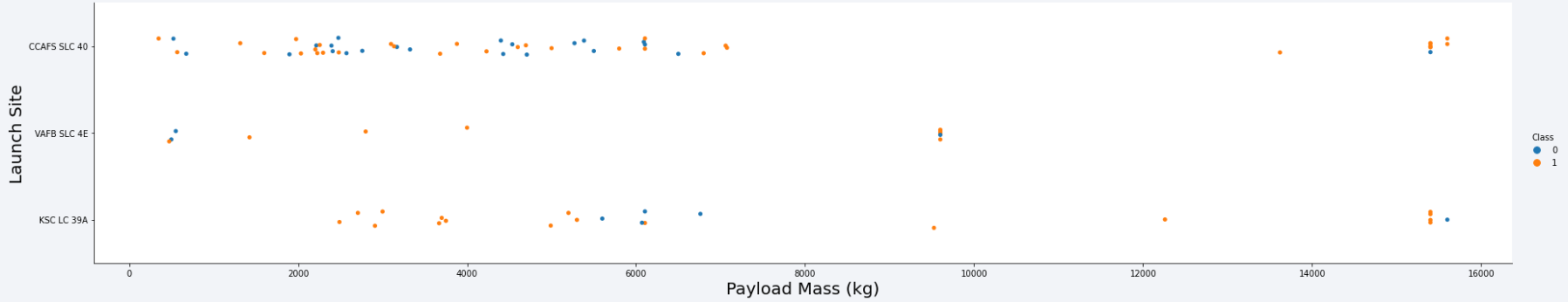
Results: EDA with Visualization



FlightNumber x PayloadMass, 1st stage landing success positively correlated with continuous launch attempts, while negatively correlated with payload mass

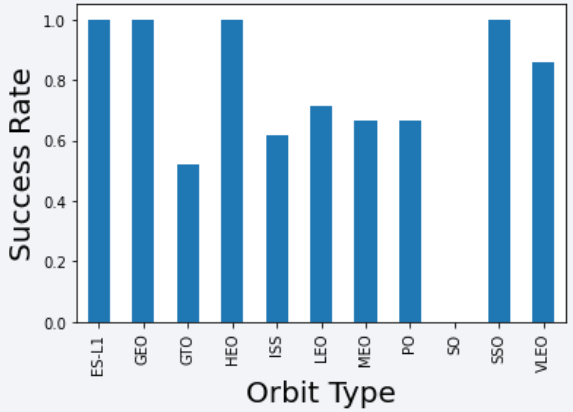


FlightNumber x LaunchSite, CCAFS SLC 40 appears to have been where most of the early 1st stage landing failures took place

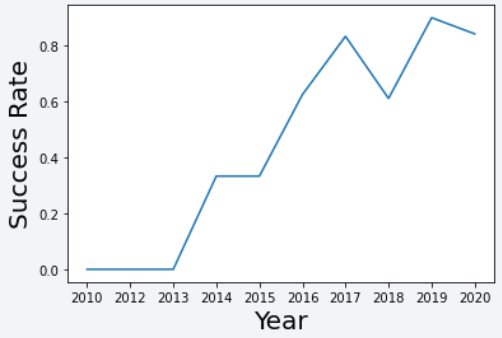


PayloadMass x LaunchSite, CCAFS SLC 40 and KSC LC 39A appear to be favored for heavier payloads

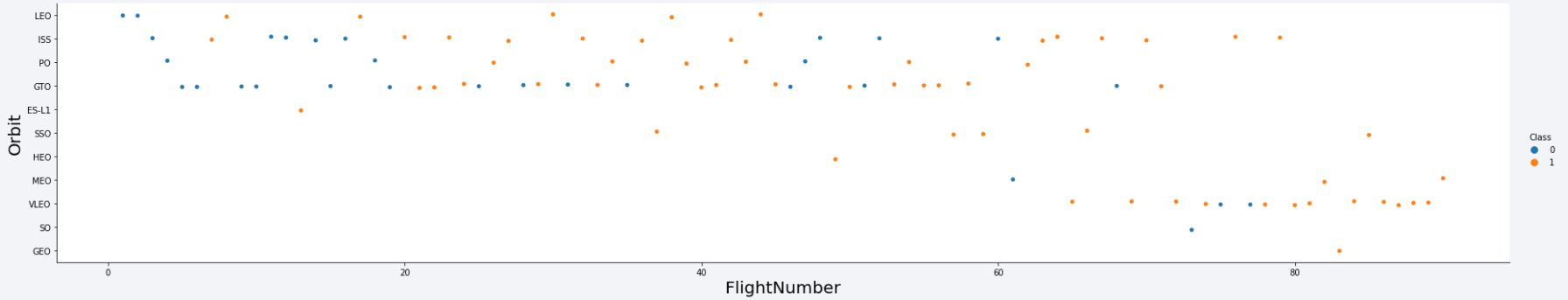
Results: EDA with Visualization (continued)



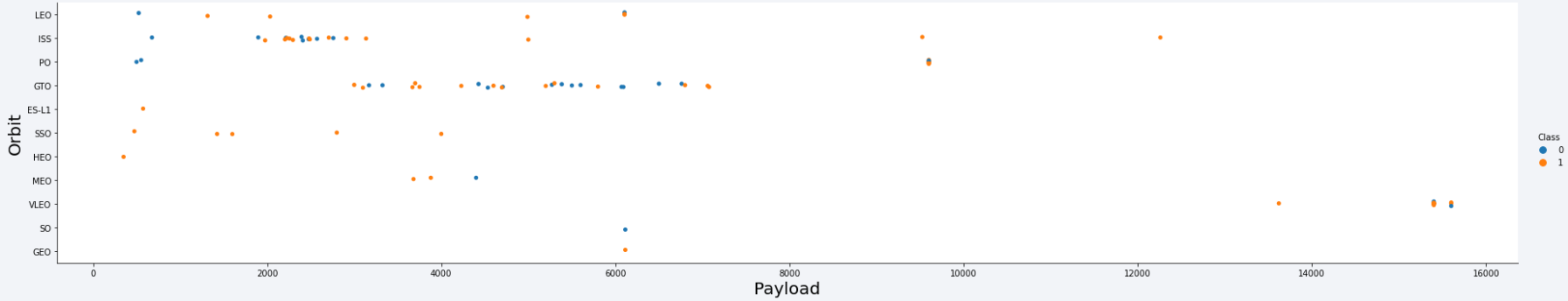
Orbit type x Success rate, All orbit types except 'SO' have had successful 1st stage landings



Year x Success rate, success rate trending positively on a yearly basis since 2013



FlightNumber x Orbit type, flight number positively correlated with 1st stage recovery for all orbit types

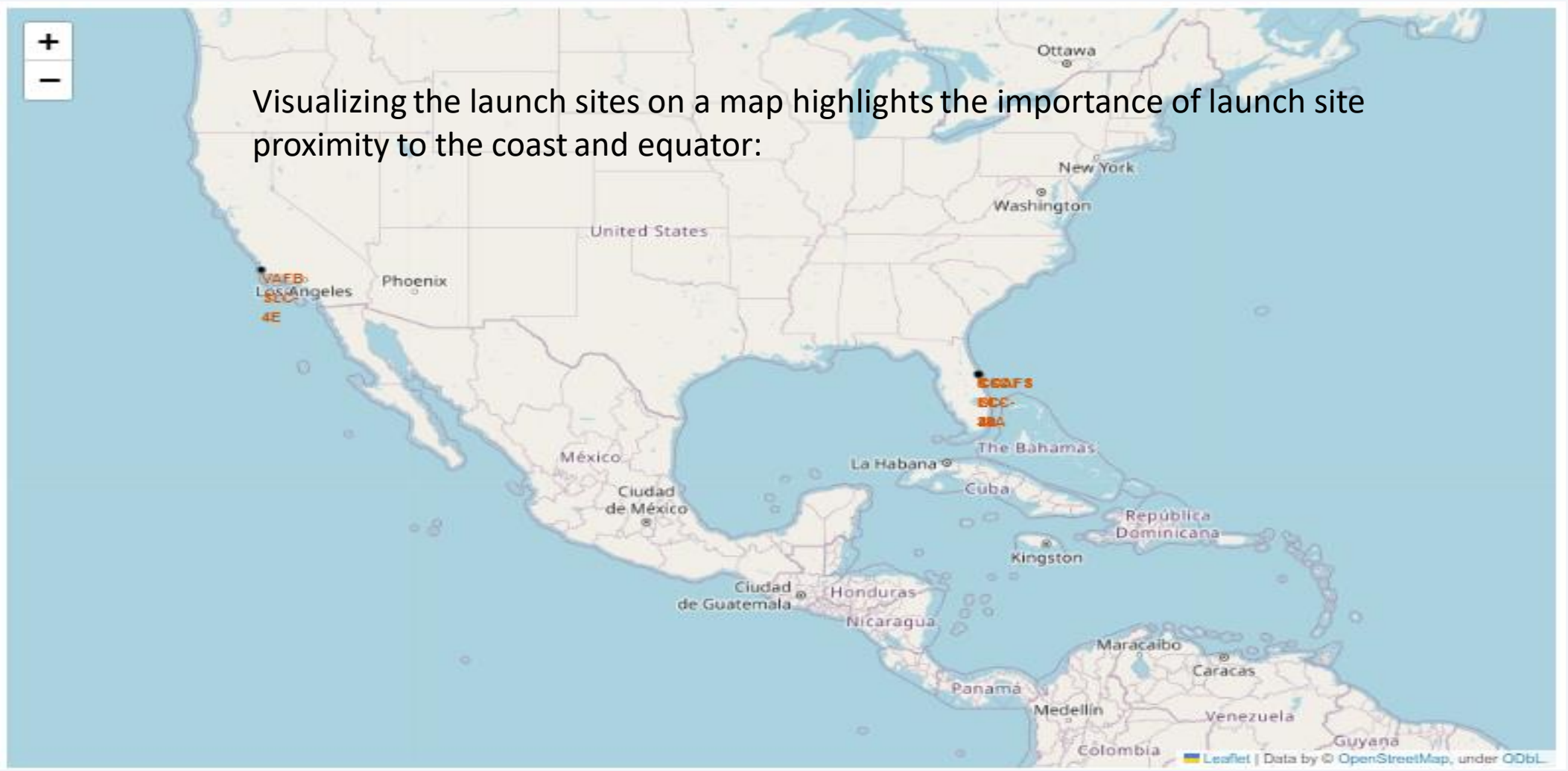


PayloadMass x Orbit type, heavier payloads have a negative influence on GTO orbits and positive influence on ISS orbits

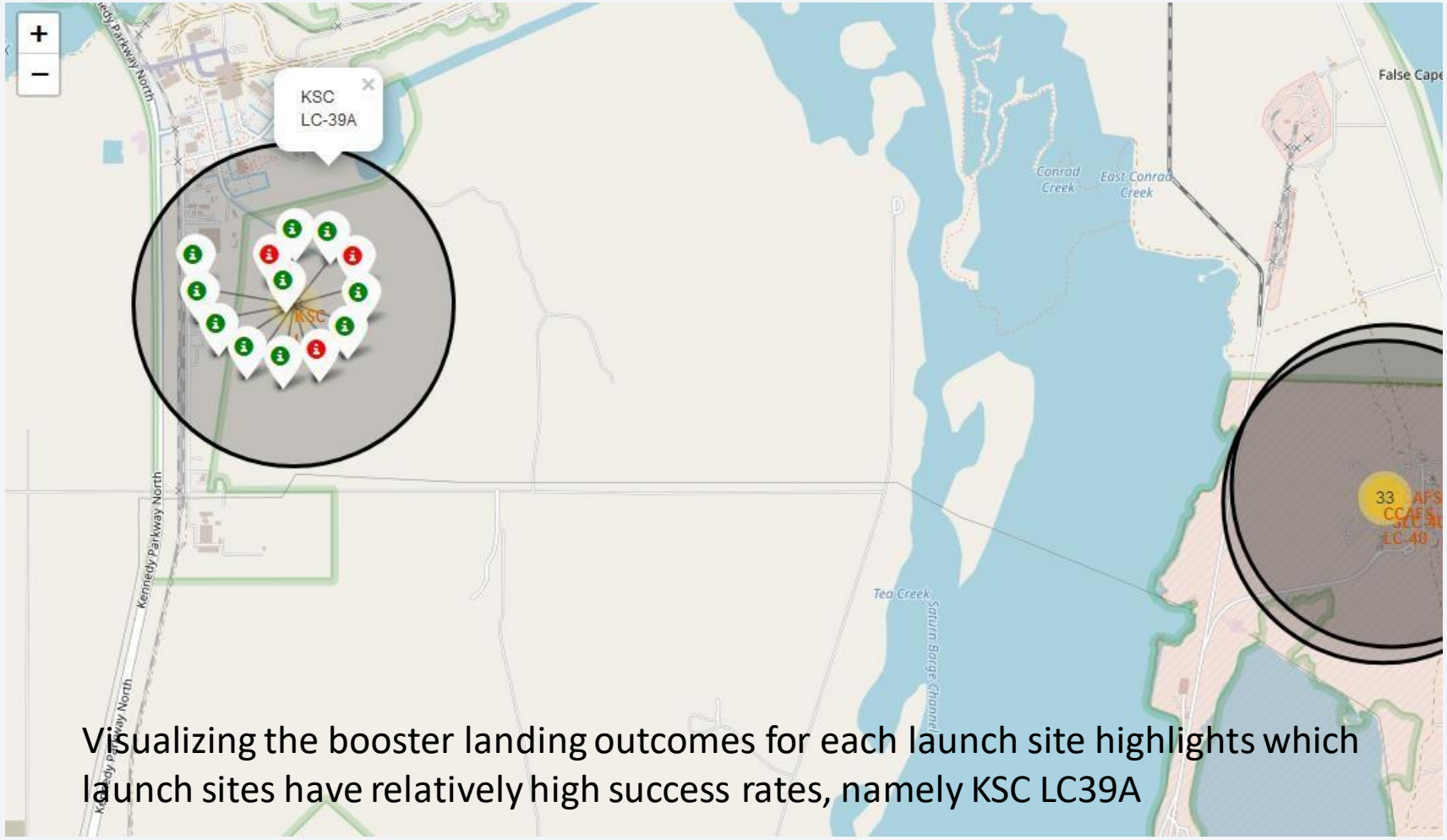
Section 3

Launch Sites Proximities Analysis

Results: Data Visualization with Folium

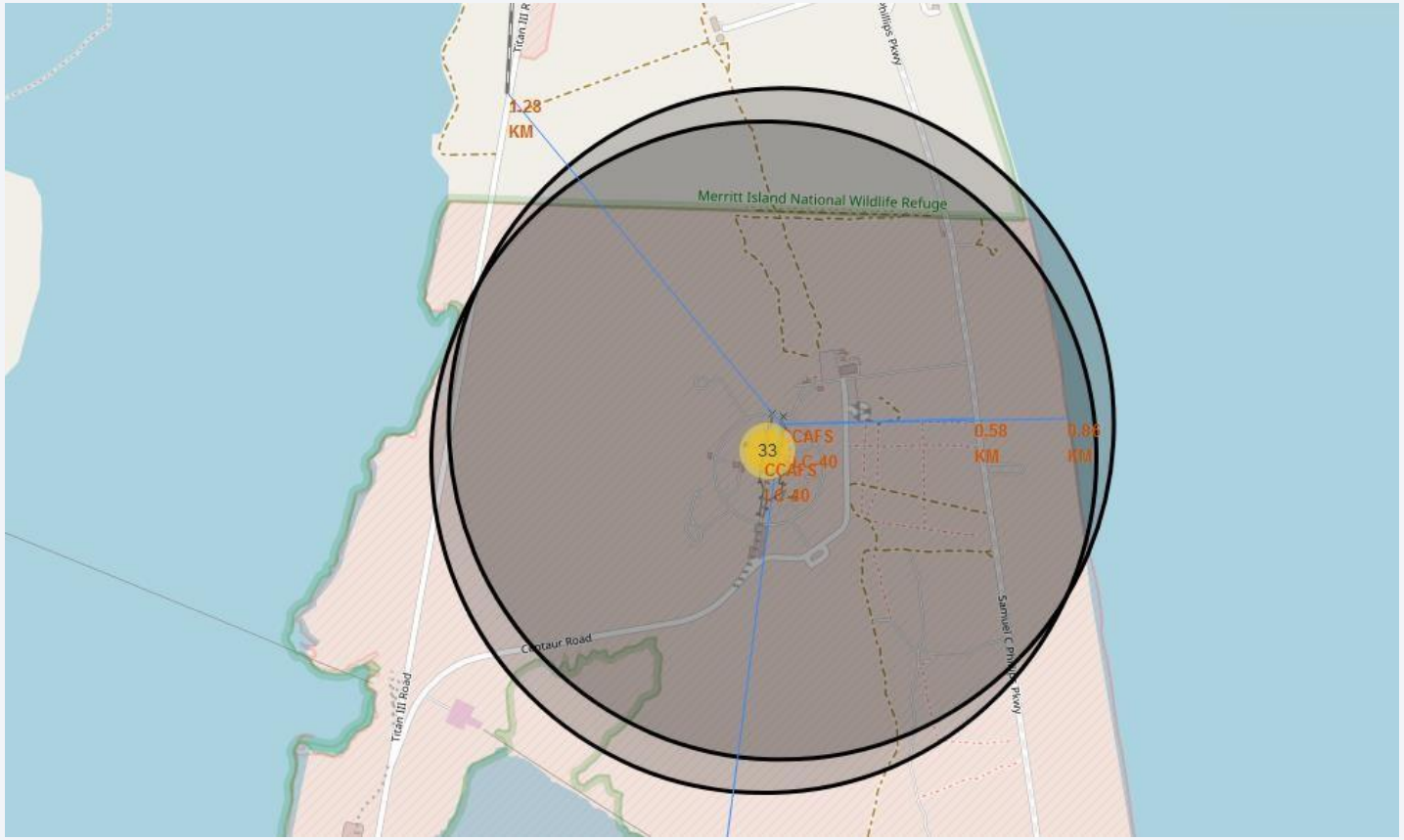


Results: Data Visualization with Folium (continued)

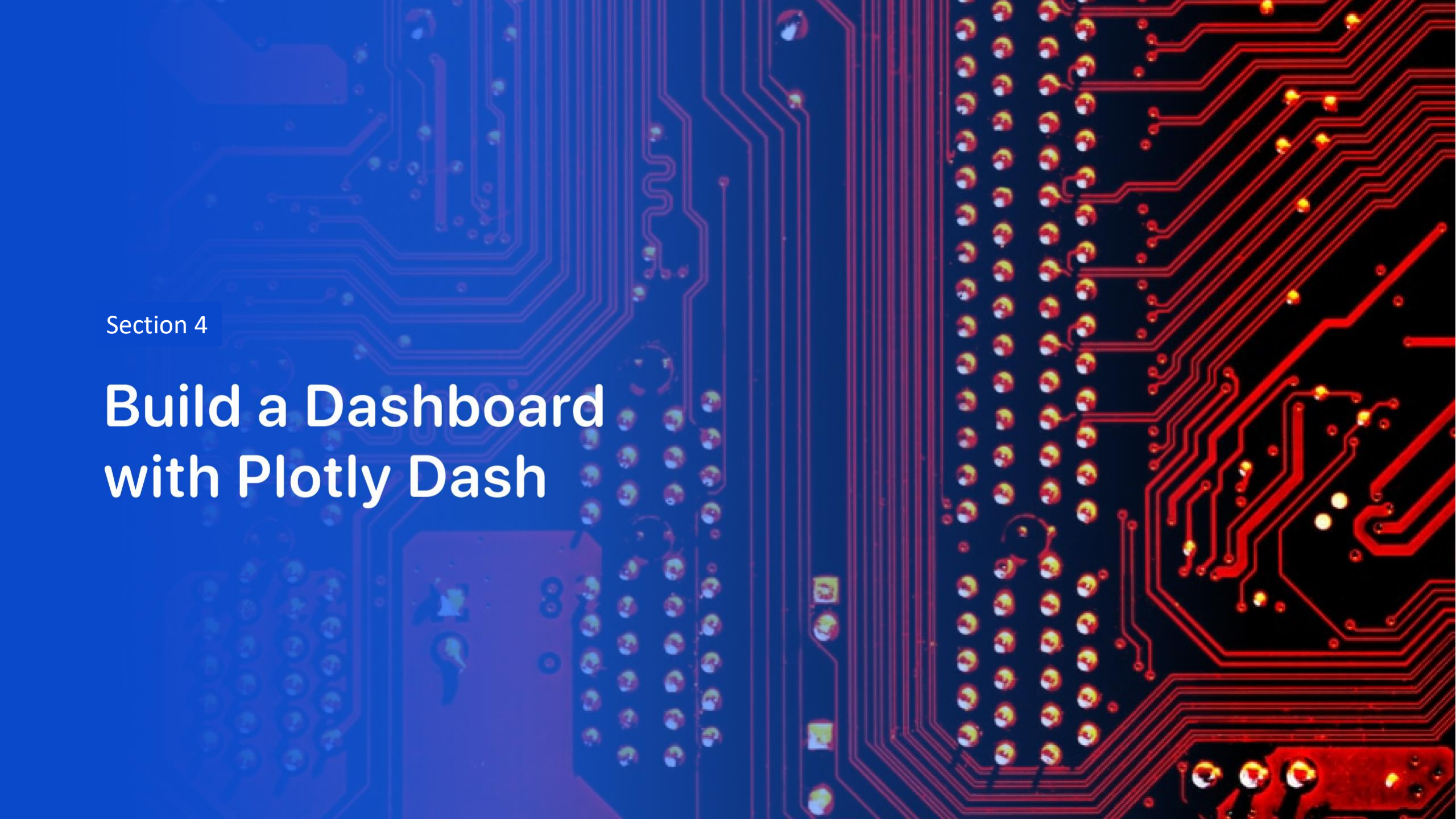


Visualizing the booster landing outcomes for each launch site highlights which launch sites have relatively high success rates, namely KSC LC39A

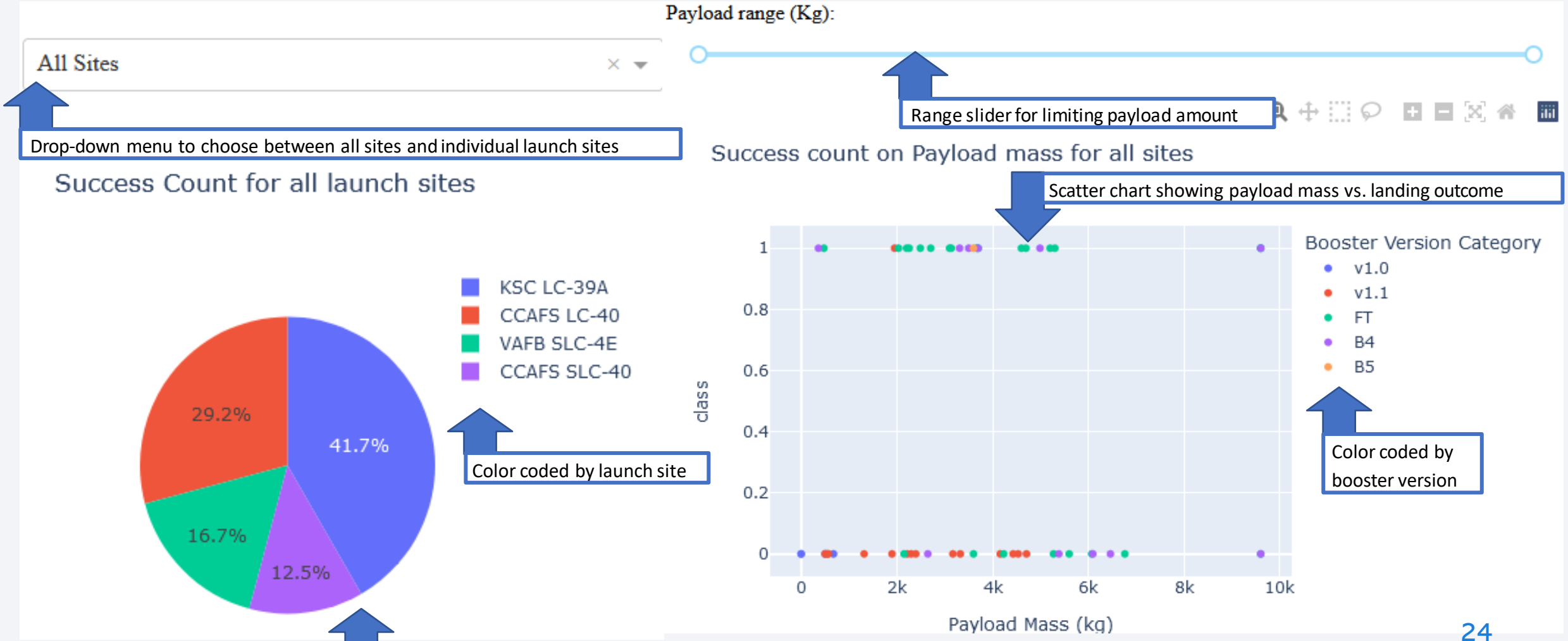
Results: Data Visualization with Folium (continued)



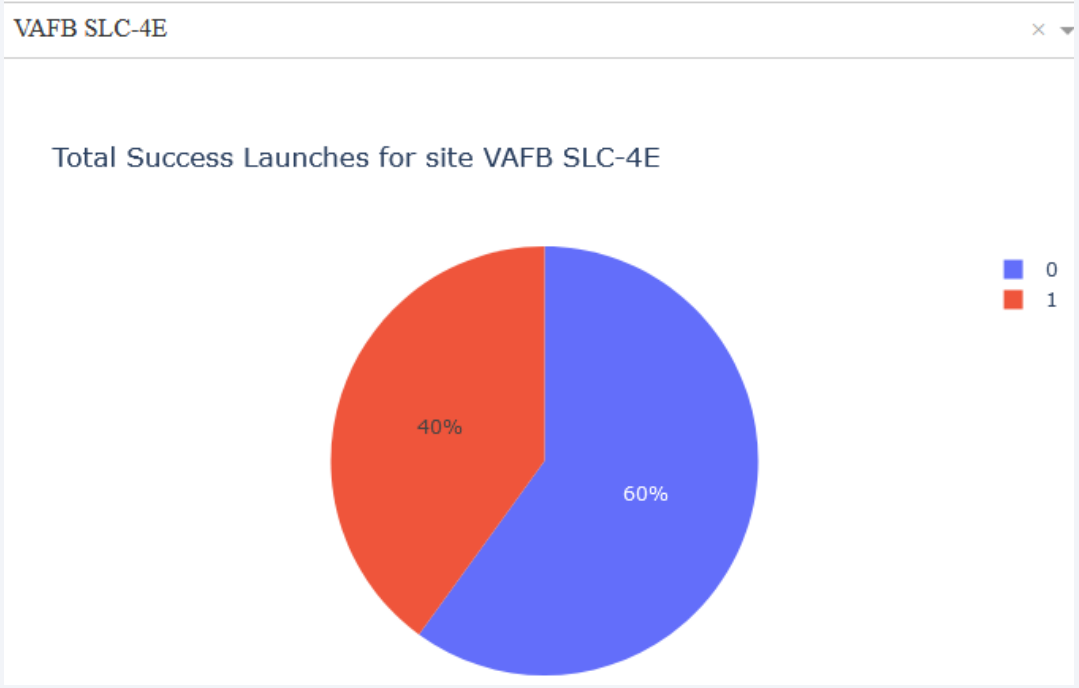
- Visualizing the railway, highway, coastline, and city proximities for each launch site allows us to see how close each is, for example, proximities for CCAFS SLC-40:
  - railway: 1.28 km
    - transporting heavy cargo
  - highway: 0.58 km
    - transporting personel and equipment
  - coastline: 0.86 km
    - optionality to abort launch and attempt water landing
    - minimizing risk from falling debris
  - city: 51.43 km
    - minimizing danger to population dense areas.


Section 4
Build a Dashboard with Plotly Dash

Results: Launch Records Dashboard



Results: Launch Records Dashboard (continued)



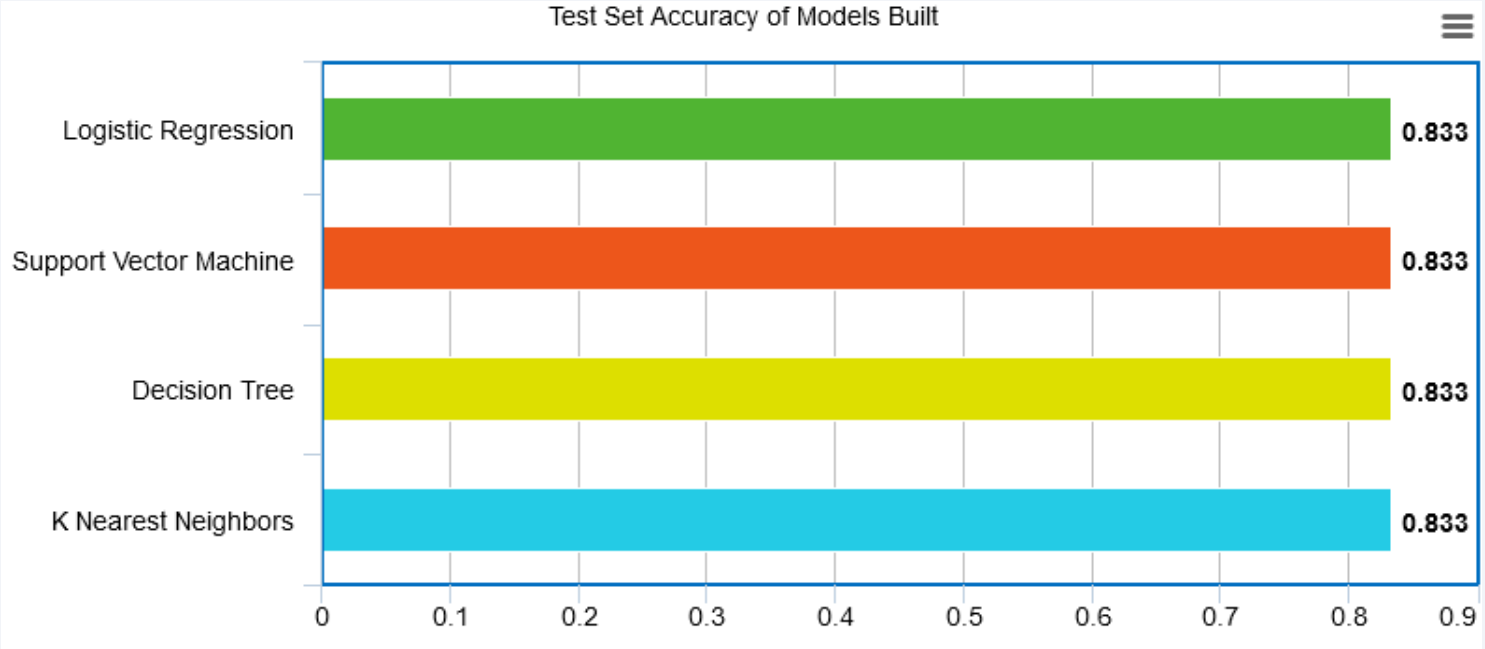
Example dashboard view:
Booster landing sucess rate for VAFB SLC-4E

- Explore the dashboard yourself:
  - Enabling stakeholders to explore and manipulate the data in an interactive and real-time way
- Dashboard observations:
  - FAFB SLC-4E had the heaviest successful booster landing success
  - KSC LC-39A has the highest booster landing success rate
  - Payloads < 5,300 kg had the highest booster landing success rate
  - Payloads > 5,300 kg had the lowest booster landing success rate

Section 5

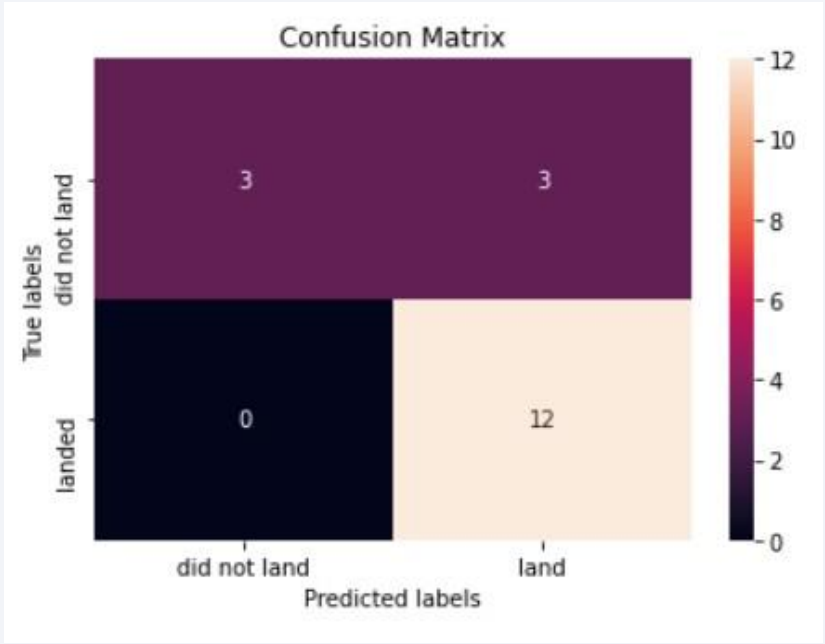
Predictive Analysis (Classification)

Classification Accuracy



Each of the four models built came back with the same accuracy score, 83.33%

Confusion Matrix



- The confusion matrices of the best performing models (4-way-tie) are the same
- The major problem is false positives as evidenced by the models incorrectly predicting the 1st stage booster to land in 3 out of 18 samples in the test set

Conclusions

- Using the models from this report SpaceY can predict when SpaceX will successfully land the 1st stage booster with 83.3% accuracy
- SpaceX public statements indicate the 1st stage booster costs upwards of $15 million to build
- This will enable SpaceY to make more informed bids against SpaceX, since they will have a good idea when to expect the SpaceX bid to include the cost of a sacrificed 1st stage booster
- With a list price of $62 million per launch, sacrificing the $15+ million 1st stage, would put the SpaceX bid at upwards of $77 million

- Biggest opportunities going forward to make even more informed bids:
  - Freeze the best performing combination of model and hyperparameters and re-fit using the whole dataset instead of just the training data
    - Potentially better than using only part of the data to fit the model, but you would no longer be able to measure the accuracy of the resulting model
  - Incorporate additional launch data to the dataset and model as it becomes available
  - Subdivide the current model into two models
    - Predict if SpaceX will ATTEMPT to land the 1st stage
    - Predict if SpaceX will SUCCEED in their attempt
  - Create a related model that predicts if SpaceX will launch using a previously-flown 1st stage booster
    - Would enable SpaceY to take into account when the SpaceX bid would likely include a discount

Appendix

- Notebooks to recreate dataset, analysis, and models:
  - Data Collection API.ipynb
  - Data Collection with Web Scraping.ipynb
  - EDA.ipynb
  - EDA with SQL.ipynb
  - EDA with Visualization.ipynb
  - Data Visualization with Folium.ipynb
  - spacex_dash_app.py
  - Machine Learning Prediction.ipynb
- Acknowledgments
  - Thank you to Joseph Santarcangelo at IBM for creating the course and materials
- References
  - https://aviationweek.com/defense-space/space/podcast-interview-spacexs-elon-musk
    - Interview with Elon Musk where he discloses the 1st stage booster to cost upwards of $15 million
  - https://datascience.stackexchange.com/a/33050
    - Explanation of why you would rebuild your model using the full dataset
  - https://www.spacex.com/vehicles/falcon-9/
    - Source of SpaceX's advertised $62 million launch price

Thank you!