IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Olivia Leniuk
2/26/2022

Outline

Executive Summary

Methodology

Data was collected from the SpaceX API as well as a Space X Wikipedia article on Falcon 9 launches. We modified the data by creating a new column to designate the launch as either successful or unsuccessful, and analyzed our data using visualizations from Matplotlib and seaborn, as well as SQL queries. Further, we used Folium to create interactive maps depicting the various launch sites included in our data and created an interactive dashboard using Plotly Dash. Lastly, several different predictive analysis models we utilized in order to obtain predictions on whether a launch will be successful or not.

Results

Visualization tools were utilized to show relationships between launch success and key variables such as orbit, payload mass, and flight number.

After analyzing data, classification models were trained and tested for predicting successful launches – the decision tree model resulted in the highest accuracy score at 94%.

Introduction

One of the primary reasons Space X has become a front runner in the commercial space field is the cost efficiency resultant of reusing first stage boosters. Space X advertises a Falcon 9 launch at $62 million, whereas competitors' cost is upwards of $165 million. If prospective company Space Y can predict the likelihood of a successful first stage booster landing, it can estimate the cost of a launch.

Questions:

- How do factors such as payload mass, launch site, orbit, etc. affect the landing outcome
- How accurately can we predict a successful vs. unsuccessful landing outcome

Methodology

Executive Summary

- Data collection methodology:
  - Data was sourced from the SpaceX REST API as well as a Wikipedia article on Falcon 9 Launches
- Perform data wrangling
  - As there are several different categories for landings outcomes, we add a column to our data frame to categorize each launch attempt as either successful (1) or unsuccessful (0)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Import desired classification algorithms from sklearn to build models, tune models to best parameter using GridSearchCV, and evaluate accuracy for the best score

Section 1

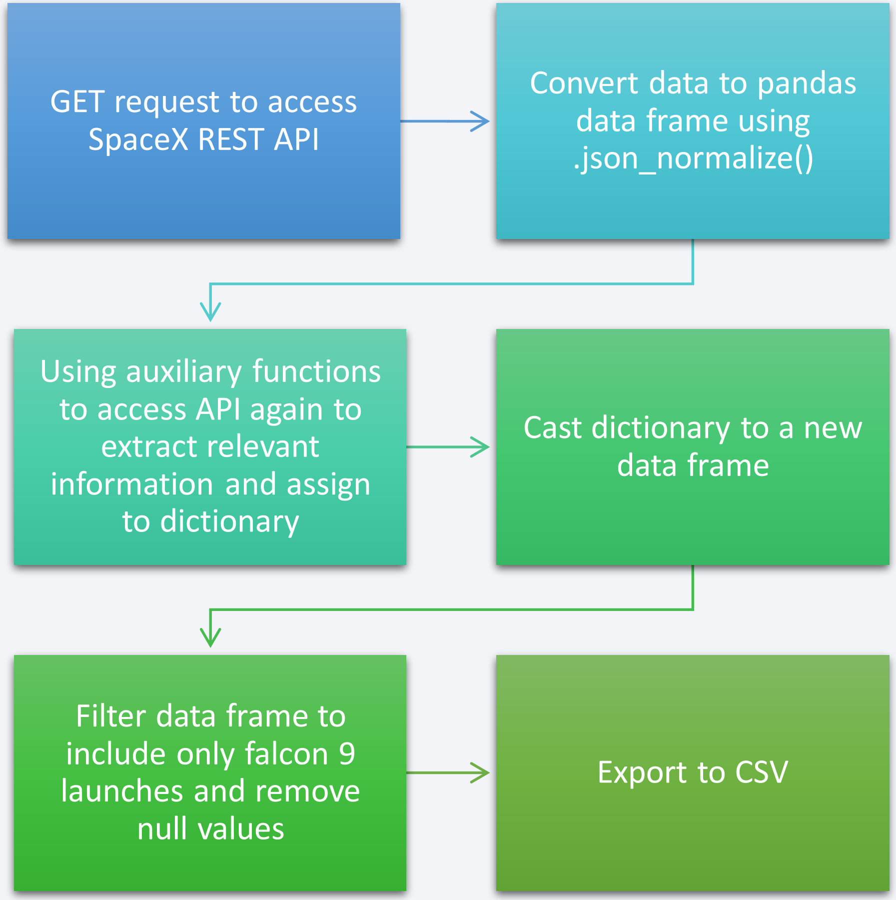
Methodology

Data Collection – SpaceX API

Collected the following data on Falcon 9 launches vis SpaceX API

- Booster name
- Payload mass
- Launch site location (along with longitude and latitude)
- Landing outcome
- Whether grid fins were used, legs were used, core was reused, which landing pad was used
- Core block
- No. of times core has been reused
- Serial of core

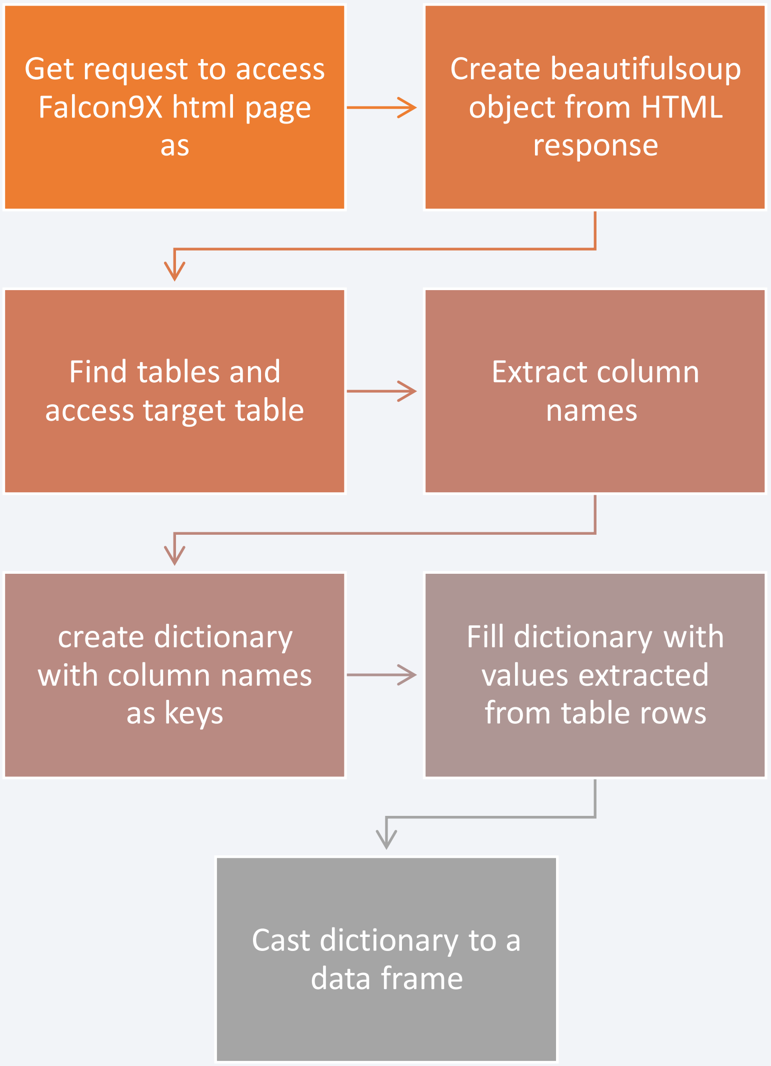
Github URL



Data Collection - Scraping

Collected the following data via web scraping of the Wikipedia article on [Falcon 9 launch history]

- Flight number
- Launch site
- Payload
- Payload mass
- Orbit
- Customer
- Launch outcome
- Version booster
- Booster landing outcome
- Date



[Github URL]

Data Wrangling

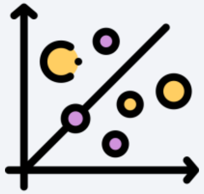
In its original format, the data contained 8 different versions of booster landing outcomes. For instance, True Ocean indicates a successful landing to a designated area of an ocean, while False ASDS indicates an unsuccessful landing to a drone ship. For simplicity towards our purpose, we recategorized each landing as either a successful (1) or unsuccessful (0) launch and created a new column in our data frame to represent its respective class.

1. Obtain list of all distinct landing outcomes
2. Create set distinguishing the unsuccessful launches from successes
3. Create new column to update data frame
4. Create a landing outcome label

[Github URL]

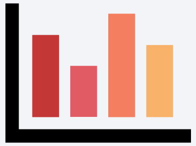
EDA with Data Visualization

The following charts were utilized for visual analysis:



Scatter plots – *ideal for determining if a relationship exists between two variables*

- Flight No. vs Payload Mass
- Flight No. vs Launch Site
- Payload Mass vs Launch Site
- Flight No. vs Orbit
- Payload Mass vs Orbit



Bar Charts – *ideal for comparing categorical data*

- Success Rate by Orbit

Line Graph – *ideal to depict trends over time*

- Average Success Rate by Year

[Github URL]

EDA with SQL

SQL queries used for exploratory data analysis:

- Display unique launch site locations
- Display 5 records for launch sites beginning with "CCA"
- Display total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- Return date for first successful landing outcome was achieved
- List names of boosters which have had success in drone ship landing and payload between 4000 and 6000
- List total number of successful and unsuccessful mission outcomes

Github URL

Build an Interactive Map with Folium

Using Folium I created several interactive maps displaying the various launch sites and notable features relevant to our research.

- Circular markers on launch sites
- Marker clusters to display launch outcomes
- Lines to depict distance between launch sites and proximities

Github URL

Build a Dashboard with Plotly Dash

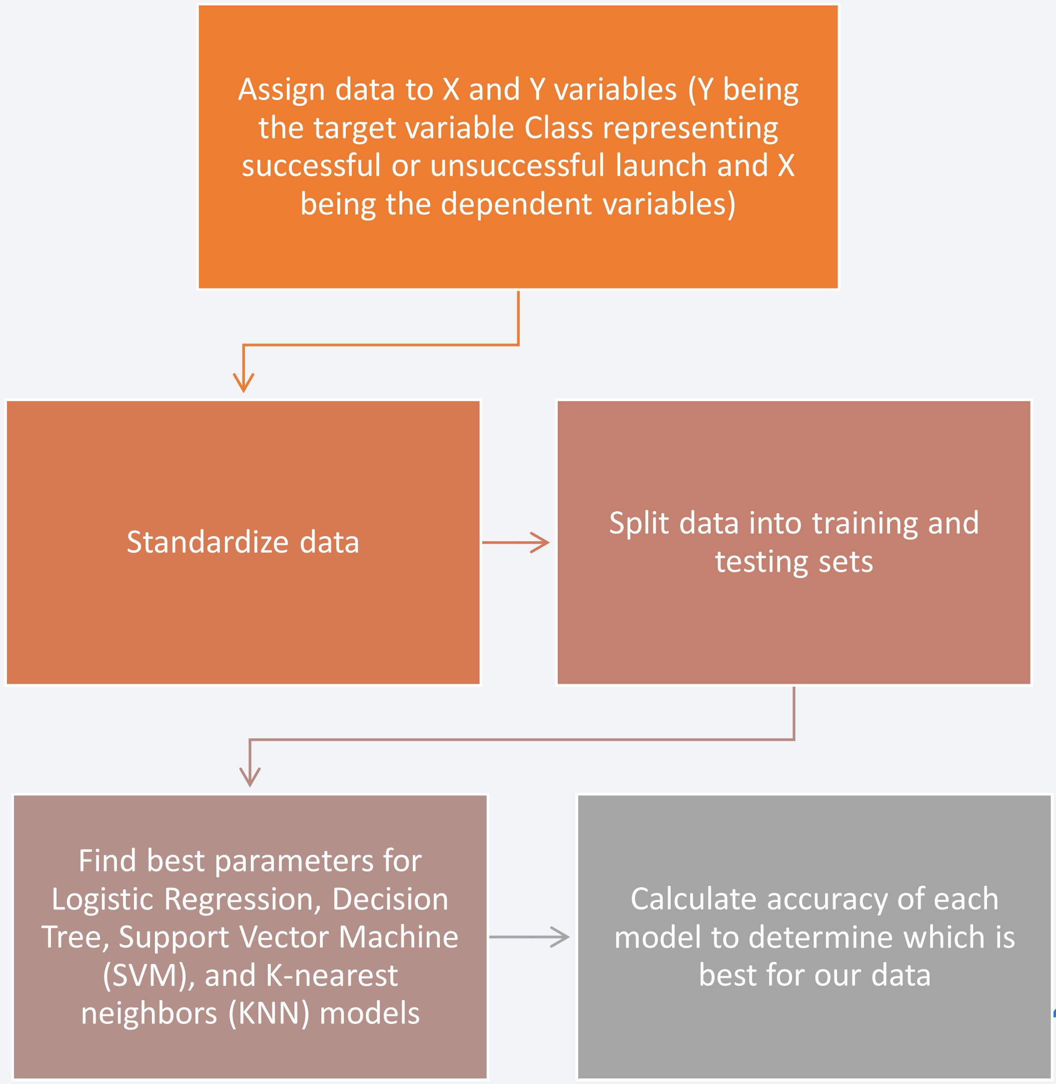
- Using Plotly Dash I created an interactive dashboard including a pie chart depicting success rate as well as a scatterplot showing success rate for each booster version. A dropdown menu allows selection of different launch sites (all data from all sites combined), and interactive payload slider allows specific payload range for scatter plot.
- These tools allow for greater insight into our data in order to seek out significant relationships that might contribute to launch success.

Github URL

Predictive Analysis (Classification)

- Once data has been analyzed and insights gathered, several Machine Learning algorithms are applied for predictive analysis that we can then be used to predict whether a launch will be successful
- Algorithms used:
  - Logistic Regression
  - Decision Tree
  - Support Vector Machine (SVM)
  - K-Nearest Neighbors (KNN)

Github URL

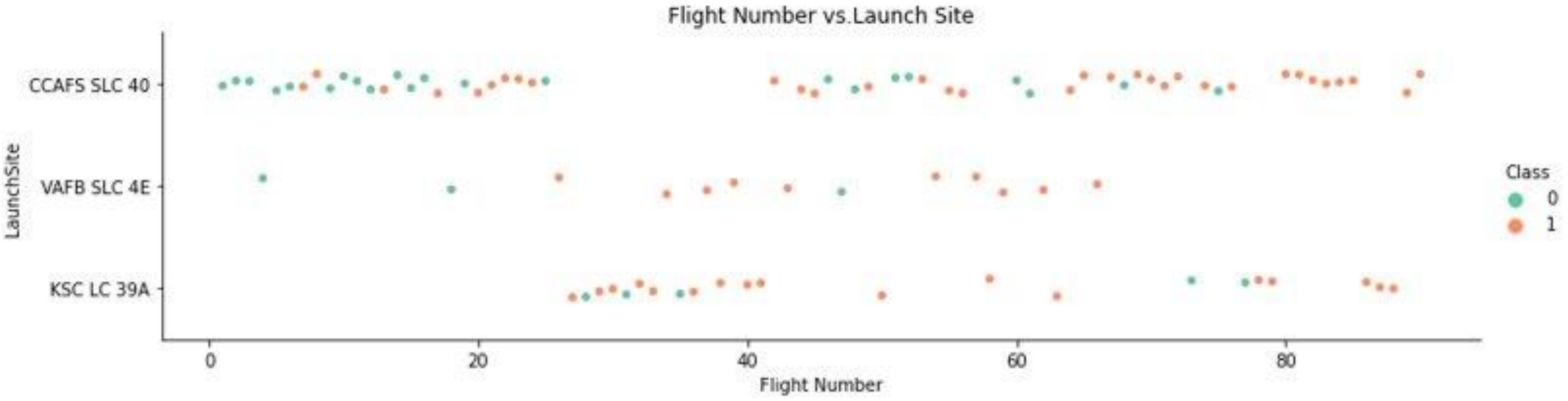


Results

- Exploratory data analysis looking into possible relationships between launch success and several key variables
- Notable relationship between success rate and launch number
- Overall success rate has steadily increased over time, most recently around 80% successful in 2020
- Predictive analysis via Decision tree had a 94% accuracy score suggesting good confidence for future launch predictions

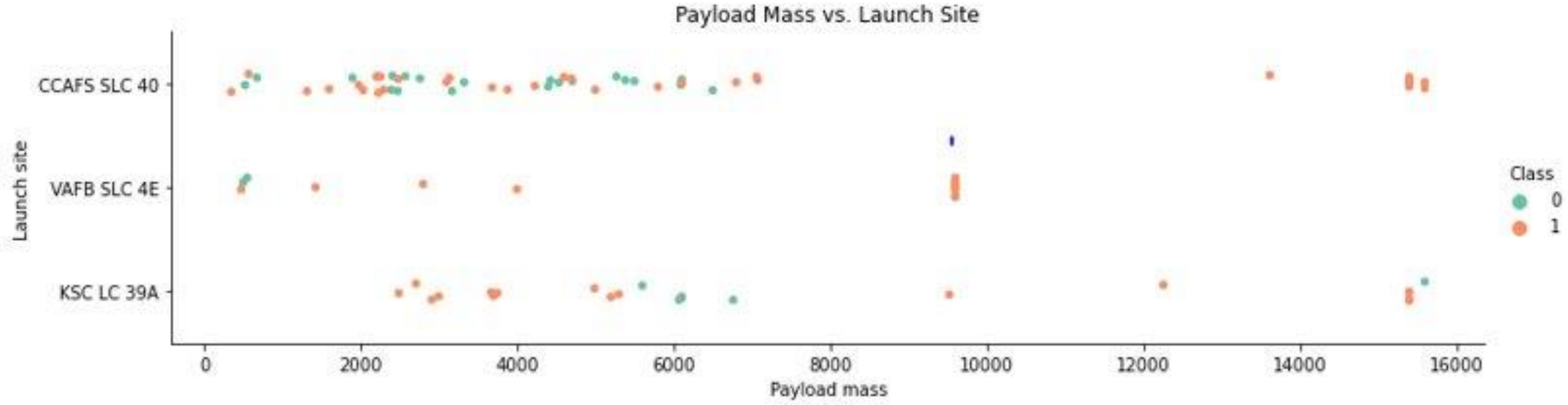
Section 2

Insights drawn from EDA



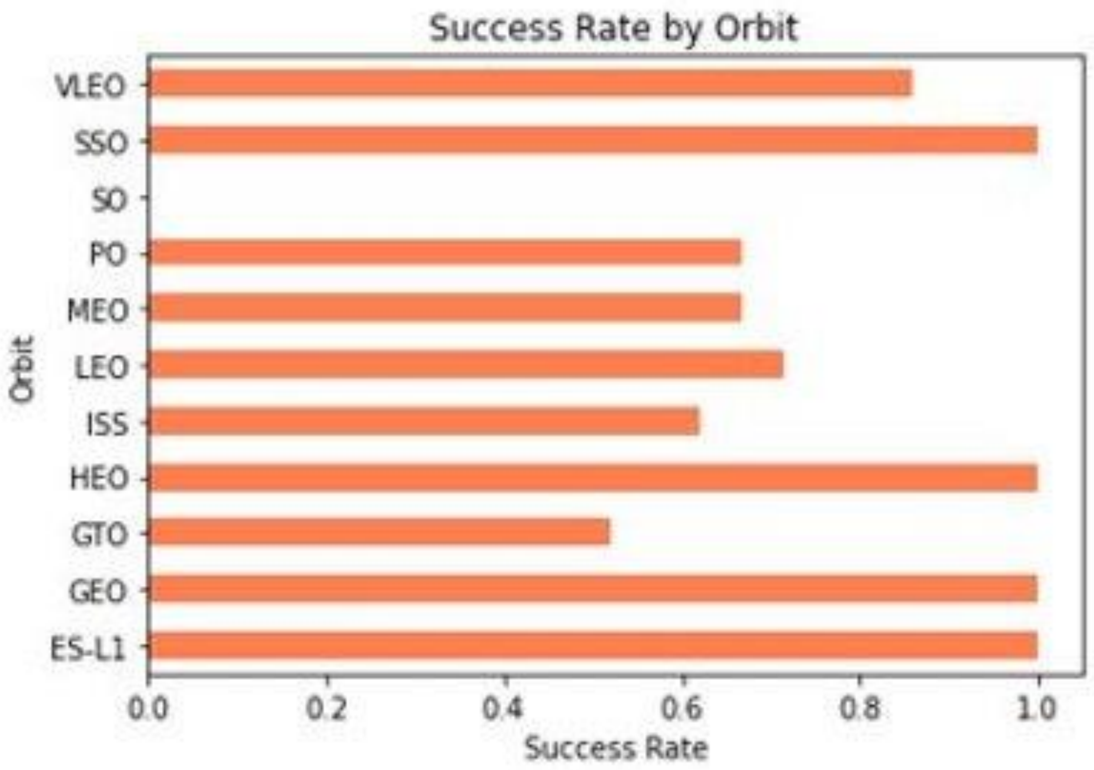
Flight Number vs. Launch Site

- Generally, more success in later flight numbers – indicates with more attempts there are more favorable outcomes



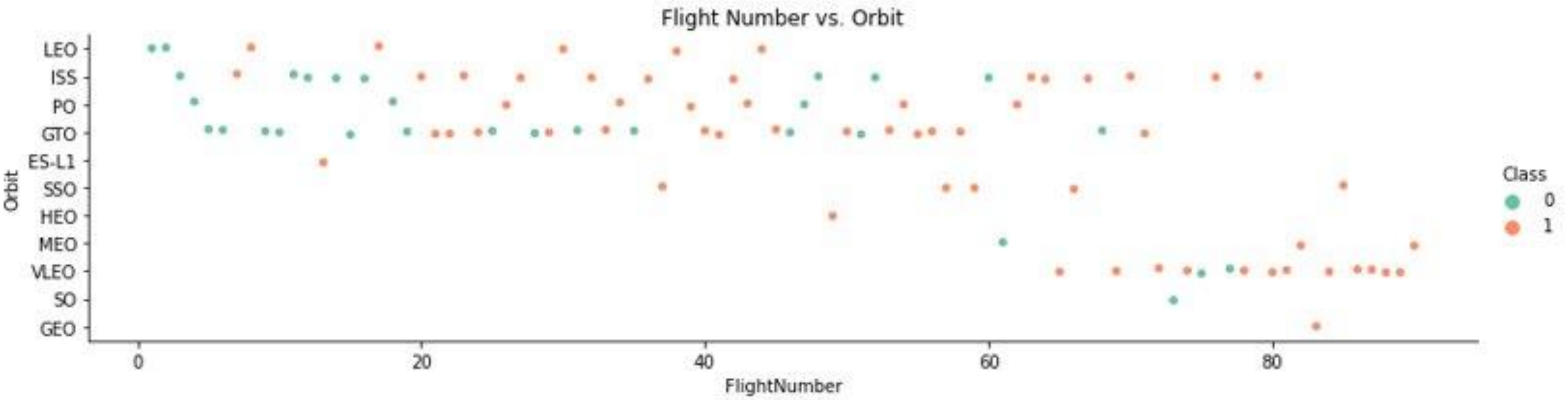
Payload vs. Launch Site

- CCAFS (Cape Canaveral) and VAFB (Vandenberg) appear to have more successful landings with greater payloads, whereas the opposite seems to be true with KSC (Kennedy Space Center)



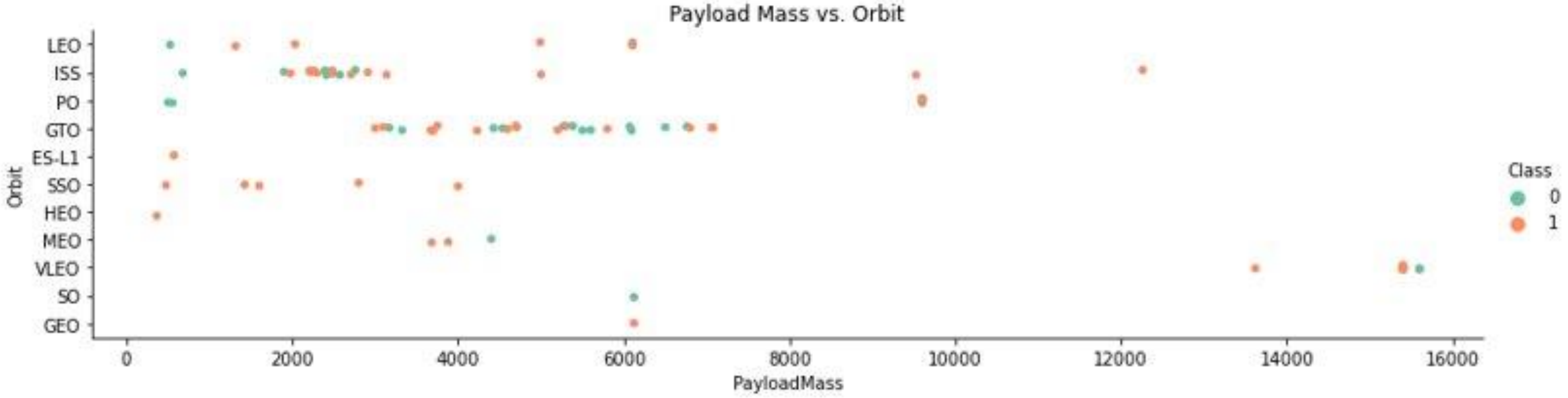
Success Rate vs. Orbit Type

- SSO, HEO, GEO, and ES-L1 all have a perfect success rate
- Further analysis shows this is likely on account of the fewest launches
- VLEO has the next highest success rate with considerable launch history



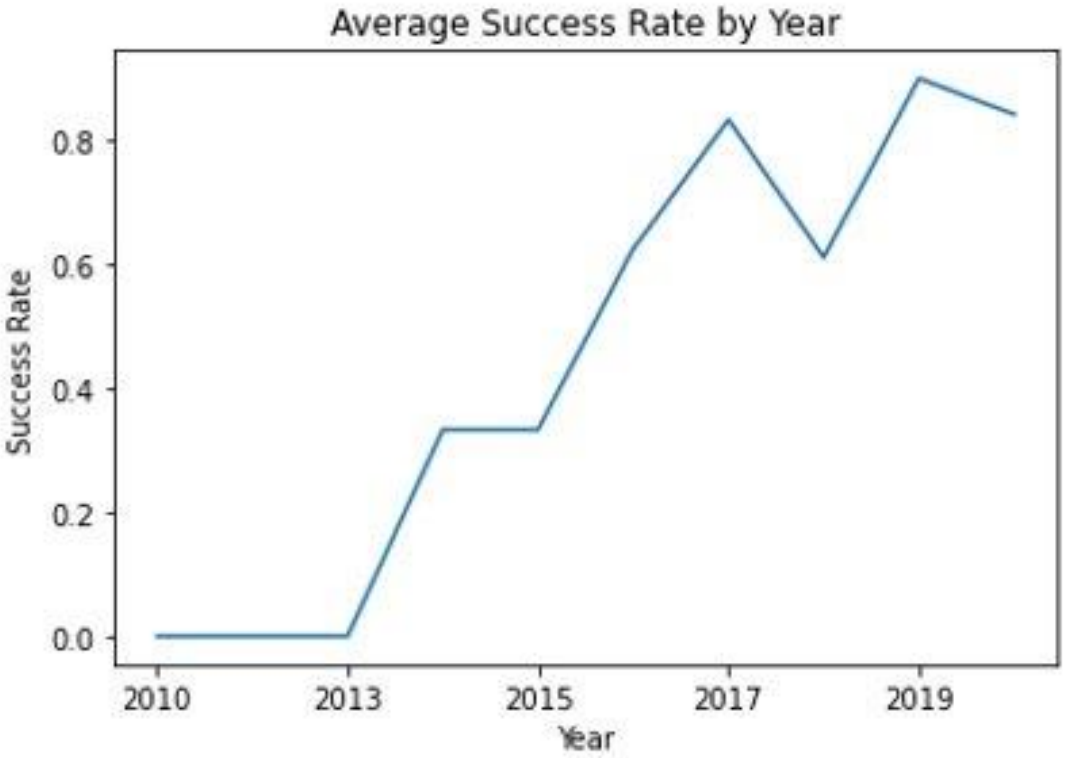
Flight Number vs. Orbit Type

- Successful launches tend to increase with higher flight numbers



Payload vs. Orbit Type

- For the majority of launches payload mass remains below 8000kg
- Heavier launches seem to have a positive influence on LEO, ISS, and PO orbits



Launch Success Yearly Trend

- Average success has increased through 2019 except for a slight decline in 2017

All Launch Site Names

- Three different launch sites:
  - CCAFS SLC-40 refers to Cape Canaveral Space Launch Complex (Florida)
  - KSC LC-40 refers to the Kennedy Space Center Launch Complex (Florida)
  - VAFB SLC-40 refers to the Vandenberg Space Launch Complex (California)

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

- Used **select DISTINCT** query to return unique values in the launch_site column

Note: Originally four distinct launch sites due to multiple records listed under the former name CCAFS LC-40 (Launch Center instead of Space Launch Center) – used UPDATE query to update label

Launch Site Names Begin with 'CCA'

- Used **like** statement to search for records beginning with ‘CCA’ in the launch_site column, and **limit 5** to display only 5 results

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 SLC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS SLC-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 SLC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS SLC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS SLC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass Carried by NASA

- This query returned the total payload mass carried by boosters launched by NASA
- Used **sum** statement to get the total of all values in the payload_mass column, and **where** statement to limit calculations to where the customer column contains NASA(CRS)

total_payload
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
- Used function **avg** to calculate the average of values in the payload_mass column, along with **where** statement to limit calculations to where booster column contains F9 v1.1

avg_payload
2928

First Successful Ground Landing Date

- Data of first successful landing to a ground pad was achieved
- Used **select** statement to obtain the Date column, **where** statement to filter out records where landing_outcome does not match Success (ground pad), and **order by date** to sort by earliest to latest date, and **limit 1** so only the first date is listed

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Used **select** statement to obtain relevant columns and **where** statement to filter only records where landing_outcome column contained ‘Success (drone ship) and payload_mass column contained values between 4000 and 6000

booster_version	landing__outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes
- Used **count** and **group by** statements to obtain a column listing the count for each distinct mission outcome

mission_outcome	**2**
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass
- Used **select** statement to obtain booster_version and payload_mass column and a **where** statement with subquery to specify only records containing the maximum payload mass

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015
- Used select statement to obtain relevant columns, and **where** statement to specify only records where landing_outcome is ‘Failure (drone_ship)’ and year is 2015

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS SLC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS SLC-40	Failure (drone ship)

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

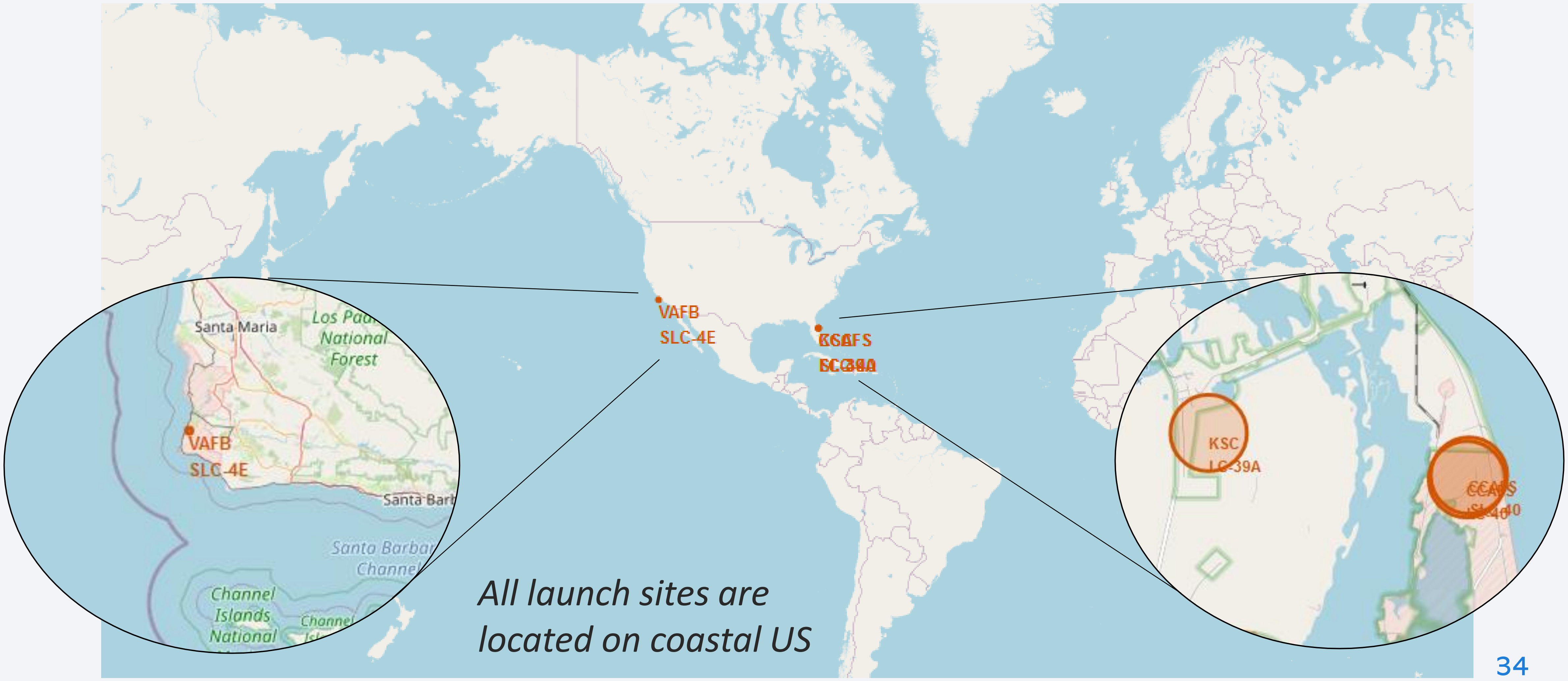
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20
- Used **select DISTINCT**, **count**, and **group by** statements to obtain counts for each distinct landing outcome, and **order by** to list from highest to lowest

landing__outcome	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

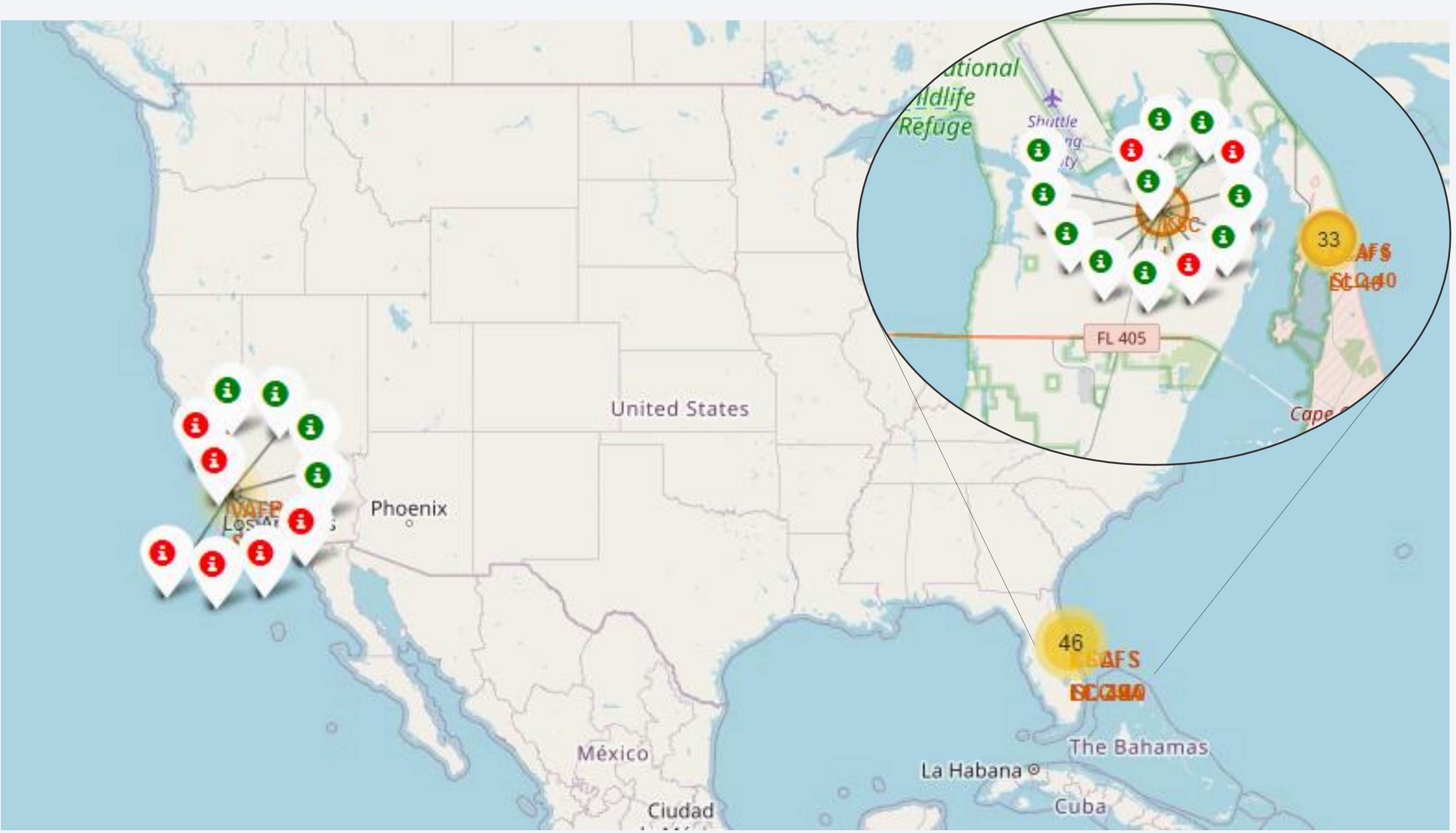
Launch Sites Proximities Analysis

Launch Site Locations



All launch sites are located on coastal US

Launch Outcome Visualization

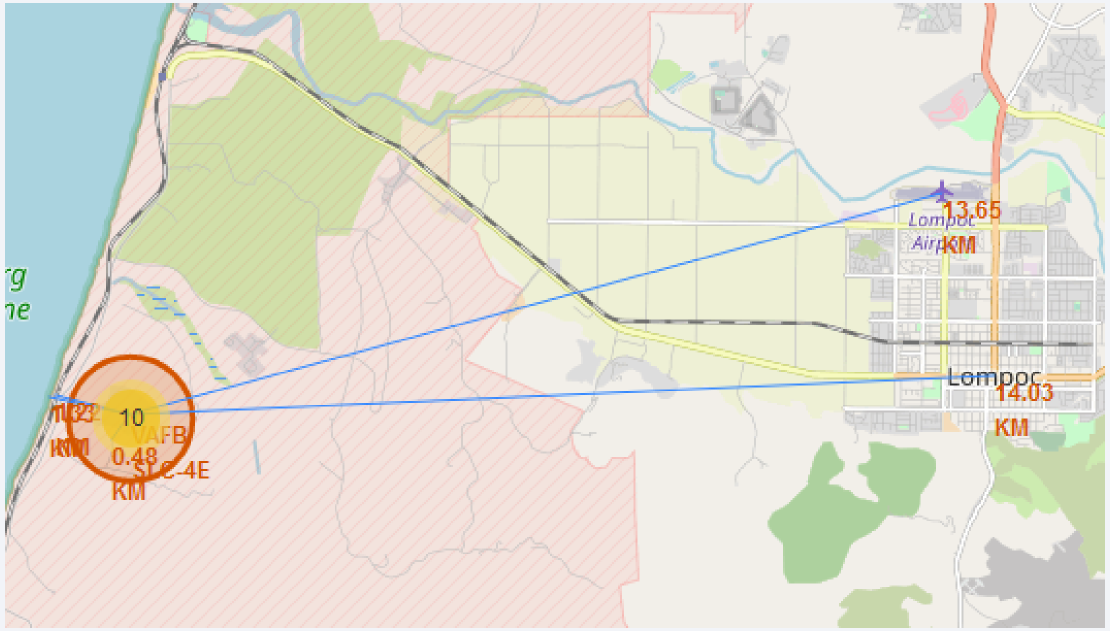


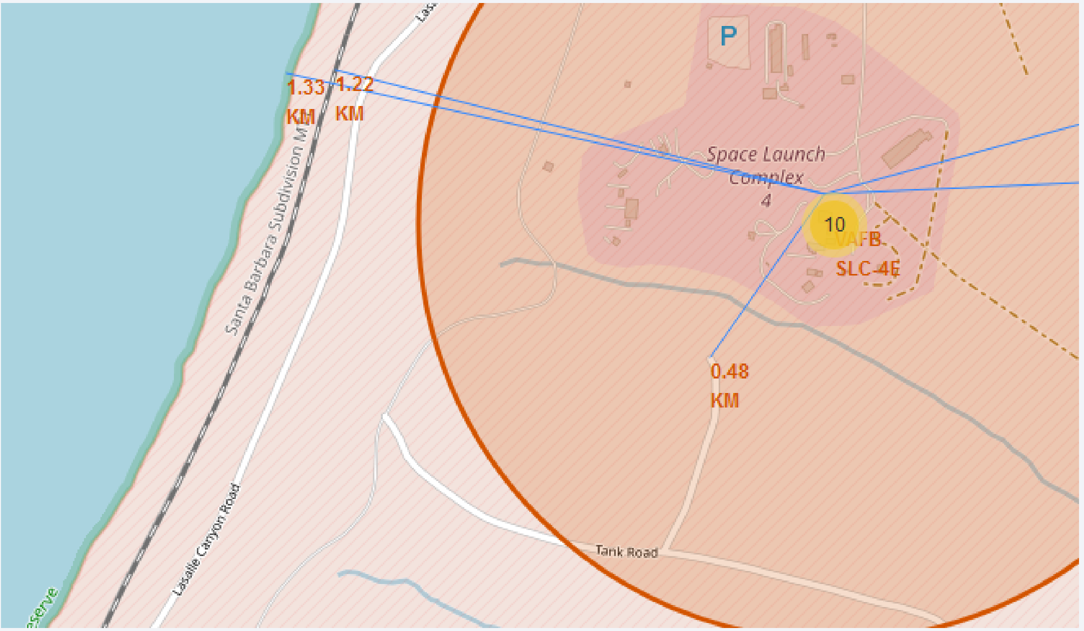
Circular markers depict number of launches in the area – clicking on marker shows specifics out outcome

Green markers indicate successful launch outcomes

Red markers indicate unsuccessful outcomes

Proximities Visualization





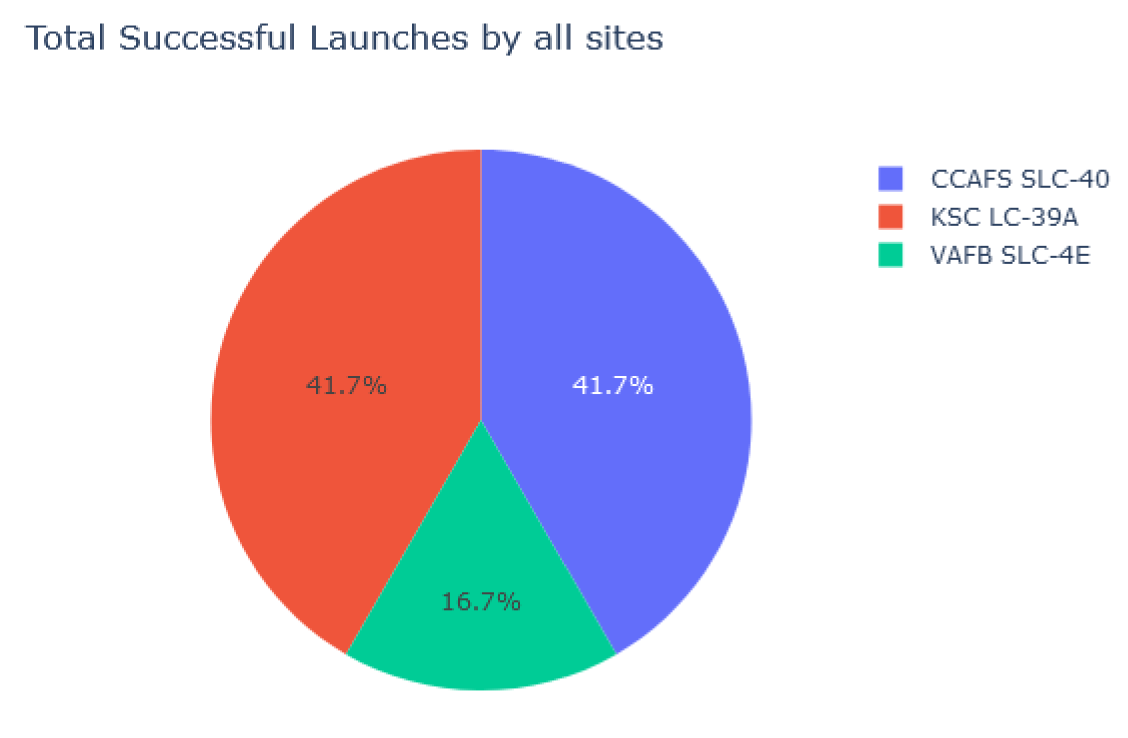
- Distance lines drawn to indicate how far launch site is from proximities including coastline, nearest city, airport, railroad, and roads

Section 4

Build a Dashboard with Plotly Dash

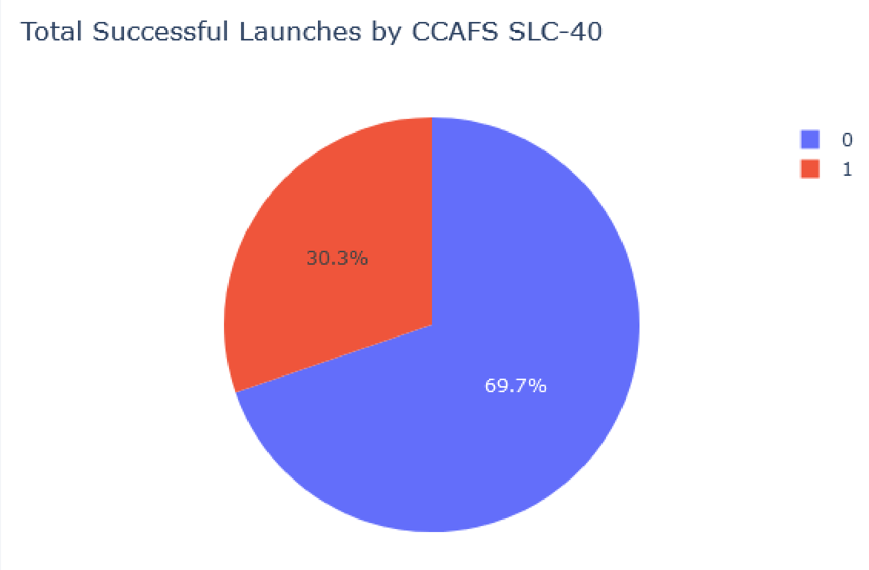
Launch Success by Location

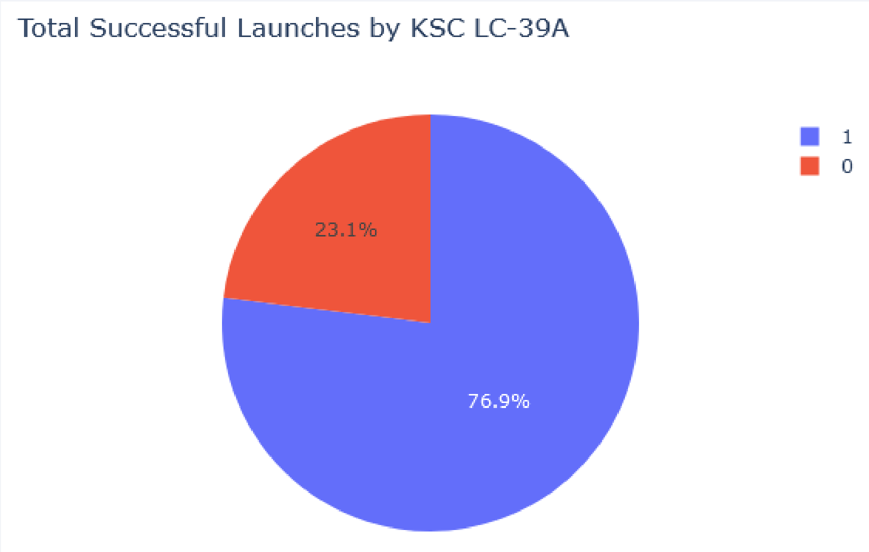
- This pie chart indicates that Kennedy Space Center and Cape Canaveral have an equal share of successful launches at 41.7%, and Vandenberg with the lowest at 16.7%
- *Note: Launch site CCAFS SLC-40 was previously CCAFS LC-40 – original data had some entries under previous name making it appear as if there were 4 separate launch sites. Data was updated to new title accordingly*



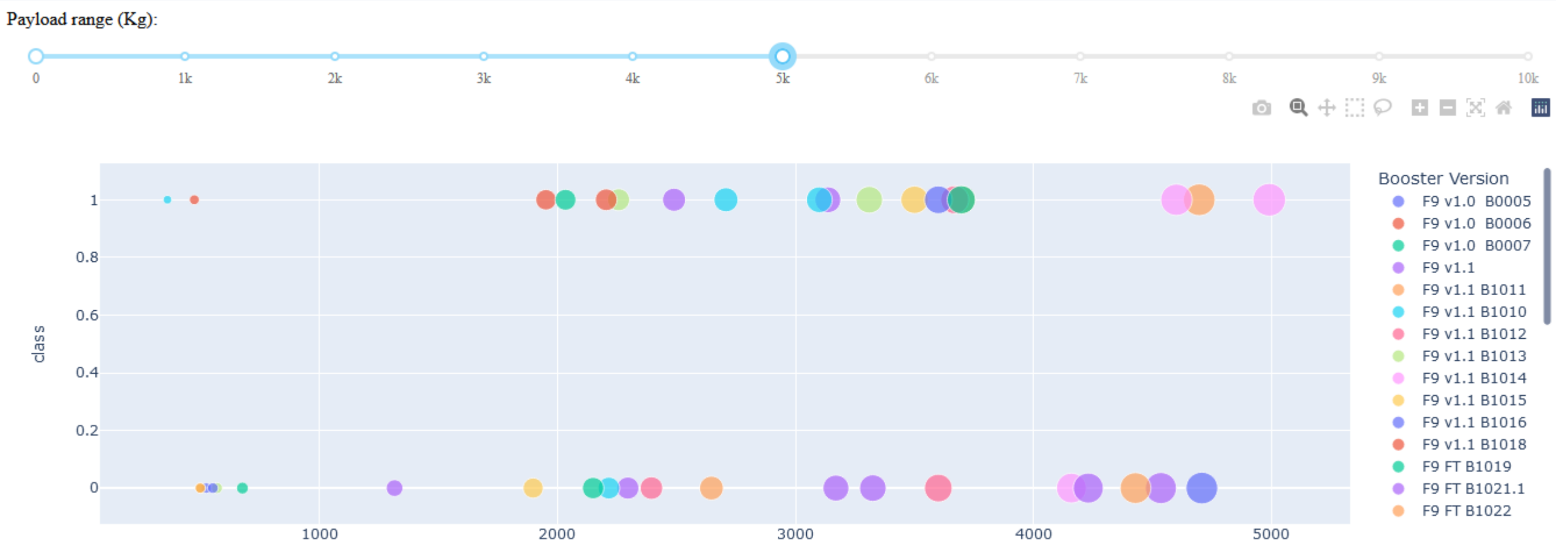
KSC vs CCAFS Launch Success

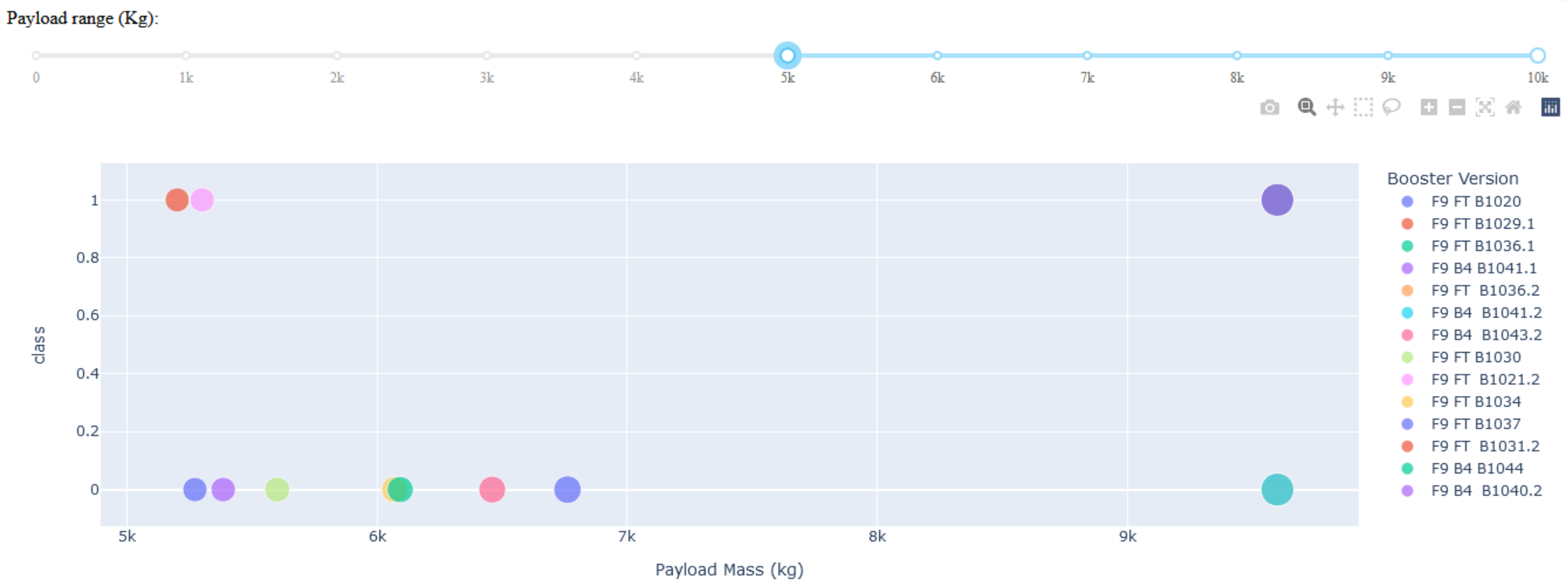
While CCAFS and KSC both share 41.7% of successful launches, KSC has a higher ratio of successful to unsuccessful launches – this is on account of CCAFS having more total launch attempts





Heavy vs Light Payload Mass





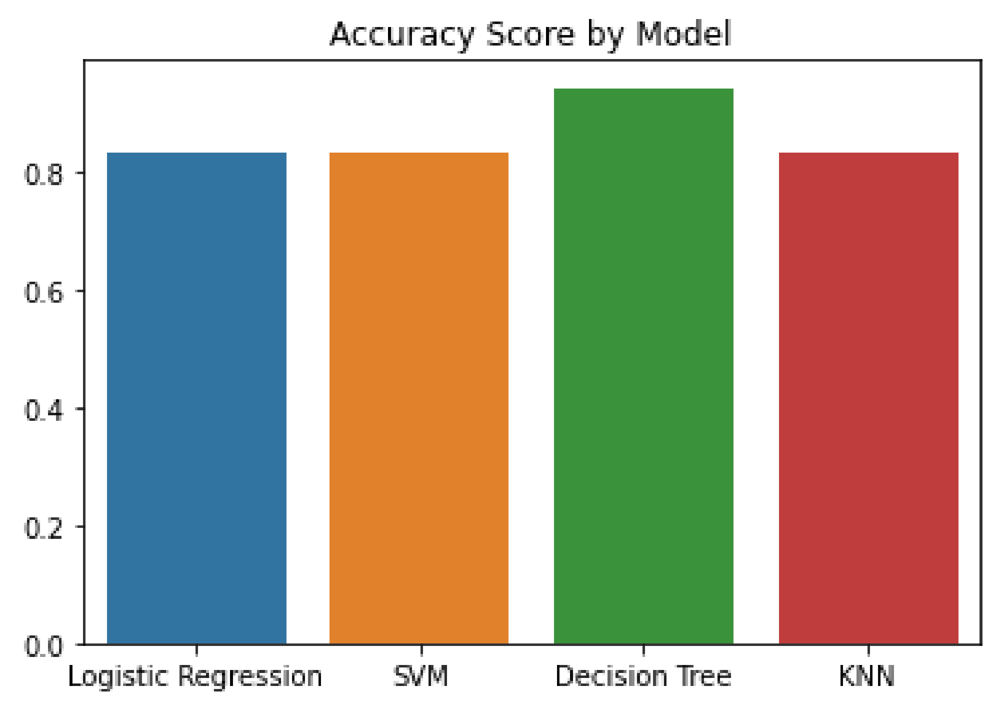
- The majority of all launches have payload mass under 7000kg
- There are no launches with payload mass between 7000-9000kg
- Only two boosters have carried loads over 9000kg, one with successful landing and one unsuccssesful

Section 5

Predictive Analysis (Classification)

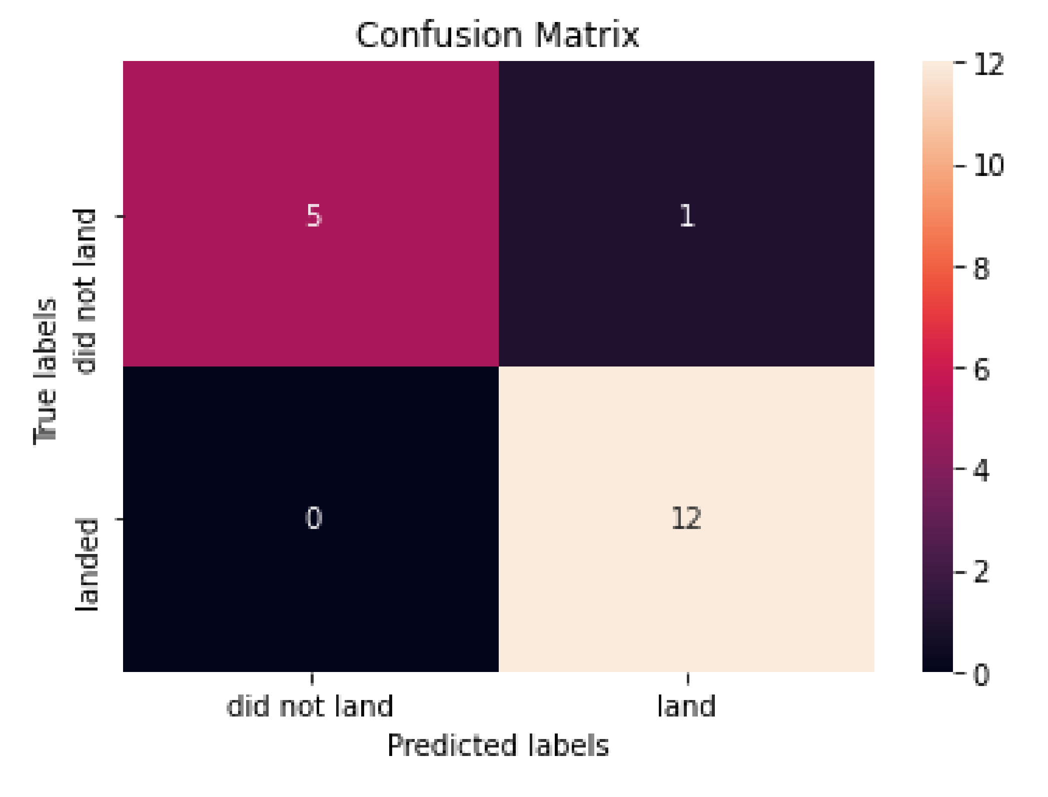
Classification Accuracy

Analysis showed the same accuracy of 0.83 across all models except the decision tree, which had the highest score of 0.94



Confusion Matrix

- Confusion matrix for decision tree
- Shows the decision tree model correctly predicted a successful landing 12 out of 12 times, and only incorrectly predicted an unsuccessful landing in 1 out of 6 launches.



Conclusions

- Notable relationships between success rate and increased flight number
- This is exemplified by the steady increase in average success rate over time shown on slide (19)
- While launch outcome varied from site to site, the over all mission outcome shows only 1 launch mission was unsuccessful – this is because most unsuccessful landings are planned
- Decision tree model gave 94% accuracy – more testing and training data would be useful however this indicates model should make considerably accurate predictions for future launches

Appendix

- [Github Capstone Project URL]
- [SpaceX Falcon 9 Wikipedia article]

Thank you!