IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lihan Tu (Vicky)
2/2/2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with API
  - Data Collection with Web Scraping
  - Data Wrangling
  - EDA with SQL
  - EDA with Features Engineering
  - Visual Mapping Analytics with Folium
  - Plotly Dash Dashboard
  - Machine Learning Predictive Analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context
  - We will predict if the Falcon 9 first stage will land successfully. 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. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.

- Problems you want to find answers
  - Find out how variables like payload mass, orbit, booster version, or launch site affect the landing outcome.
  - Which machine learning model is the best one to predict landing outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection using SpaceX API
  - Data collection vis Web Scraping on Wikipedia
- Perform data wrangling
  - Filtering out necessary data
  - Dealing with Missing Data
  - One hot converting to convert categorical data into binary format
  - Determine Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, evaluate logistic regression, SVM, Decision Tree, and K Nearest Neighbors classification models via Scikit-Learn

Data Collection

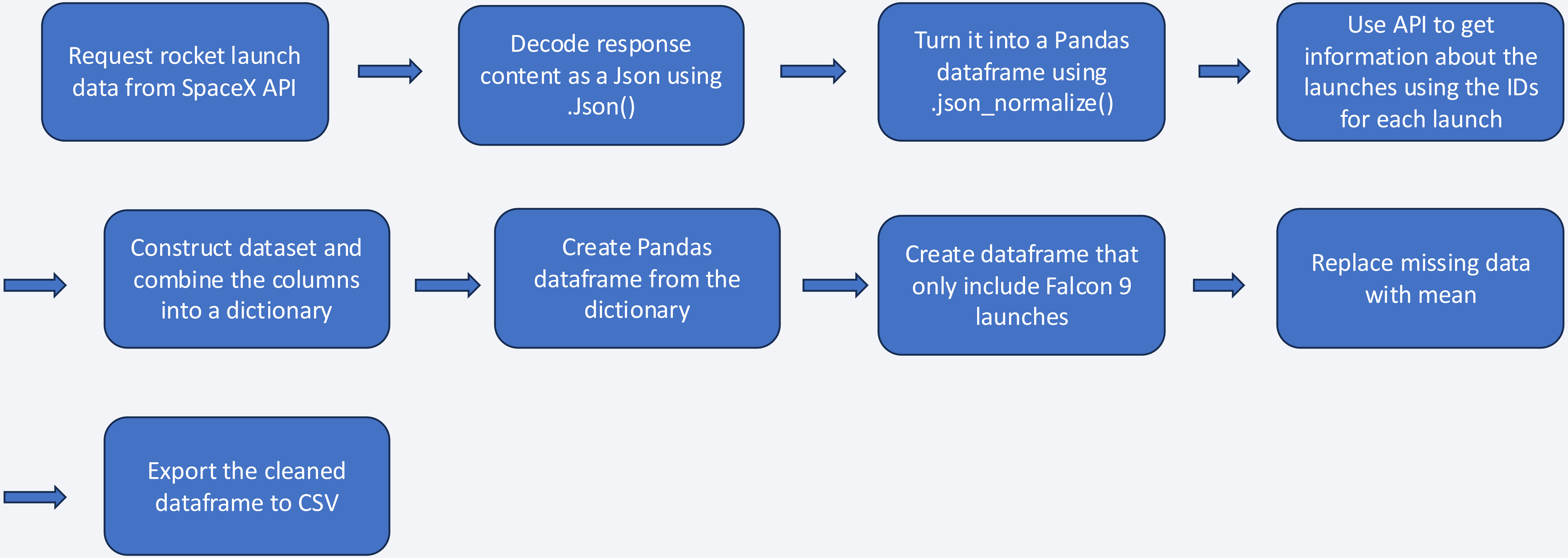
- Data was collected from two sources to ensure we have a full dataset:
  - SpaceX API
  - Web Scraping on Wikipedia
- The following data columns collected from SpaceX API will be used in this project:

BoosterVersion ,PayloadMass , Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- The following data columns scrapped from Wikipedia will be used in this project:

Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Request rocket launch data from SpaceX API

Decode response content as a Json using .Json()

Turn it into a Pandas dataframe using .json_normalize()

Use API to get information about the launches using the IDs for each launch

Construct dataset and combine the columns into a dictionary

Create Pandas dataframe from the dictionary

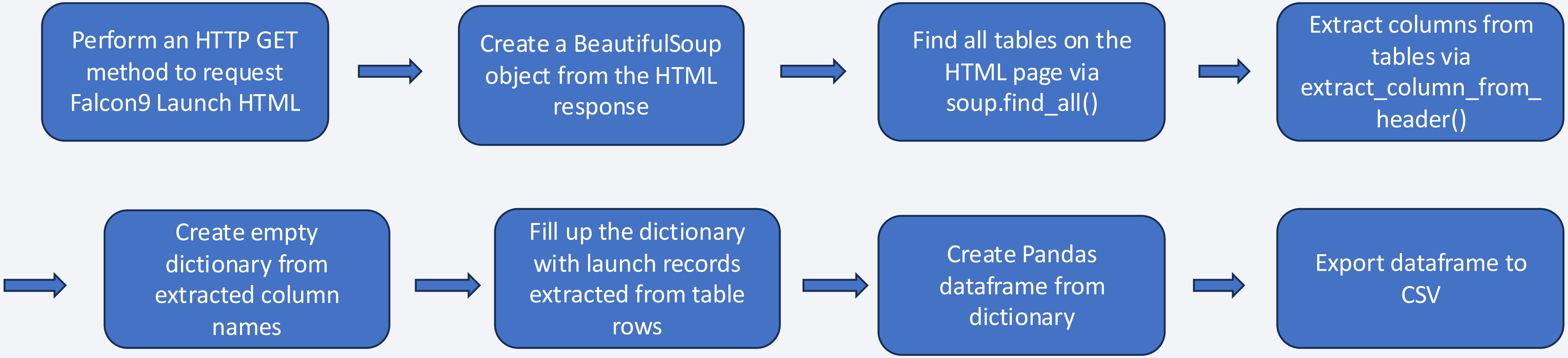
Create dataframe that only include Falcon 9 launches

Replace missing data with mean

Export the cleaned dataframe to CSV

Github URL: [Github/Data Collection via SpaceX API]()

Data Collection – Web Scraping



Perform an HTTP GET method to request Falcon9 Launch HTML

Create a BeautifulSoup object from the HTML response

Find all tables on the HTML page via soup.find_all()

Extract columns from tables via extract_column_from_header()

Create empty dictionary from extracted column names

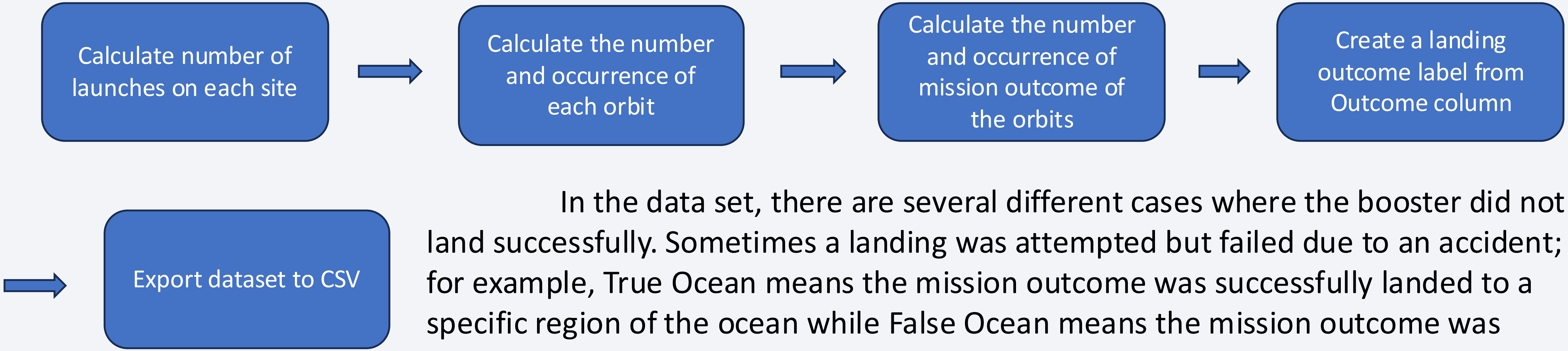
Fill up the dictionary with launch records extracted from table rows

Create Pandas dataframe from dictionary

Export dataframe to CSV

Github URL: Github/Data Collection with Web Scraping

Data Wrangling



Calculate number of launches on each site → Calculate the number and occurrence of each orbit → Calculate the number and occurrence of mission outcome of the orbits → Create a landing outcome label from Outcome column → Export dataset to CSV

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

In this lab we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

Github URL: Github/Data Wrangling

EDA with Data Visualization

1. **Visualized the relationship between Flight Number and Launch Site via scatter plot**
2. **Visualized the relationship between Payload Mass and Launch Site via scatter plot**
3. **Visualized the relationship between success rate of each orbit type via bar plot**
4. **Visualize the relationship between FlightNumber and Orbit type via scatter plot**
5. **Visualize the relationship between Payload Mass and Orbit type via scatter plot**
6. **Visualize the launch success yearly trend via line plot**

Github URL: Github/EDA Data Visualization

EDA with SQL

Github URL: Github/EDA with SQL

1. **SQL query to display the names of the unique launch sites:** "select Distinct Launch_Site from SPACEXTABLE”
2. **Display 5 records where launch sites begin with the string 'CCA’:** “select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5”
3. **Display the total payload mass carried by boosters launched by NASA (CRS):** “select Customer, sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)’”
4. **Display average payload mass carried by booster version F9 v1.1:** “select Booster_Version, avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1’”
5. **List the date when the first succesful landing outcome in ground pad was achieved: “**select Landing_Outcome, min(Date) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)’ “
6. **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:** “select Landing_Outcome, PAYLOAD_MASS__KG_, Booster_Version from SPACEXTABLE where Landing_Outcome like 'Success (drone ship)' and PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_< 6000”
7. **List the total number of successful and failure mission outcomes:** “select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome”
8. **List the names of the booster_versions which have carried the maximum payload mass:** “select DISTINCT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)”
9. **List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015: “**select substr(Date, 6,2) as 'Month', substr(Date,0,5) as 'Year', Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome like 'Failure (drone ship)’ “
10. **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.:** “select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(Landing_Outcome) desc”

Build an Interactive Map with Folium

- 1. Add the Folium circle and marker for each SpaceX launch site, based on their coordinates from the dataset, on the Folium Map to indicate their locations on a world map
- 2. Create a new column in the dataset called “marker_color” that will be green if class = 1 and red if class = 2
- Create markers to mark all the successes and failures for each launch site using marker_color
- Add mouse_position on the map to display and find display of any points of interests
- Find the coordinates of the railways, highways, coastline, and city closest to a launch site.
- Use a Folium marker to mark the distances of all the above proximities to the launch site
- Use Folium PolyLine to draw lines between each of the proximities and the launch site
- Explain why you added those objects

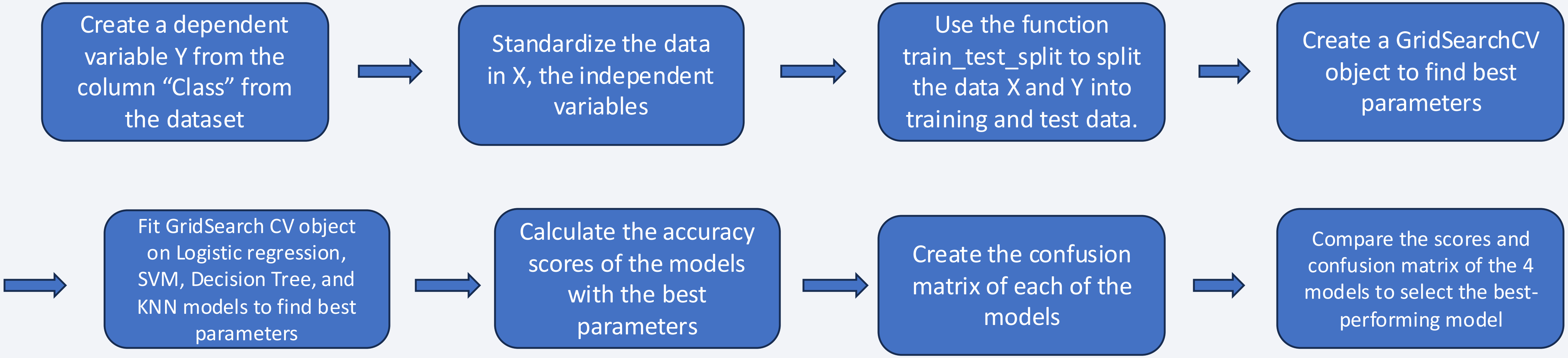
Github URL: Github/Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- Add a dropdown list to enable Launch Site selection to be used on charts that are to be created later to display information relevant to selected site
- Add a pie chart to show the total successful launches count for all sites. If a specific launch site was selected, show the Success vs. Failed counts for the site
- Add a slider to select payload range to be used on the later chart to display information to selected payload range
- Add a scatter chart to show the correlation between payload and launch success

Github URL: Github/SpaceX Plotly Dash

Predictive Analysis (Classification)



Create a dependent variable Y from the column “Class” from the dataset

Standardize the data in X, the independent variables

Use the function train_test_split to split the data X and Y into training and test data.

Create a GridSearchCV object to find best parameters

Fit GridSearch CV object on Logistic regression, SVM, Decision Tree, and KNN models to find best parameters

Calculate the accuracy scores of the models with the best parameters

Create the confusion matrix of each of the models

Compare the scores and confusion matrix of the 4 models to select the best-performing model

Github URL: Github/Prediction Analysis

Results

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

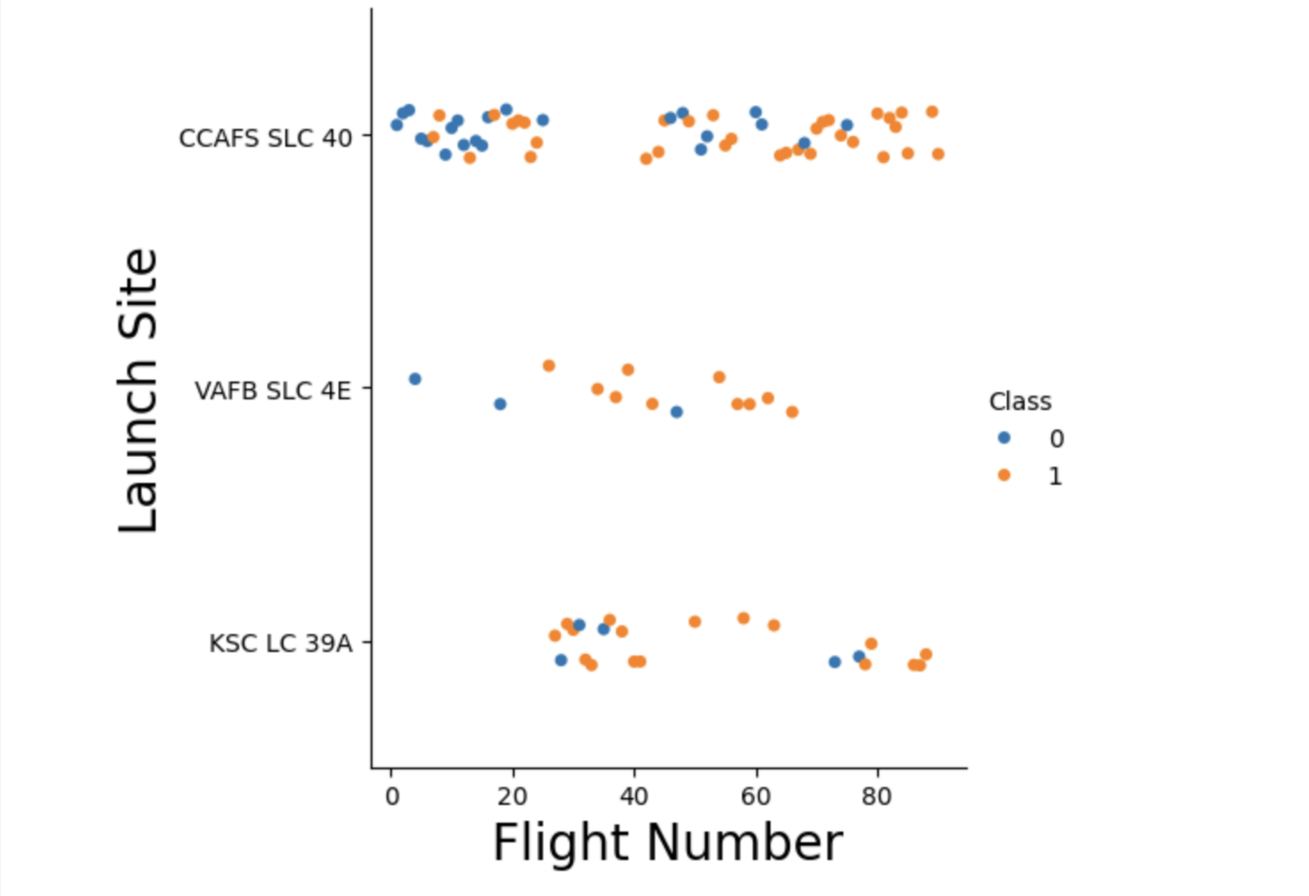
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

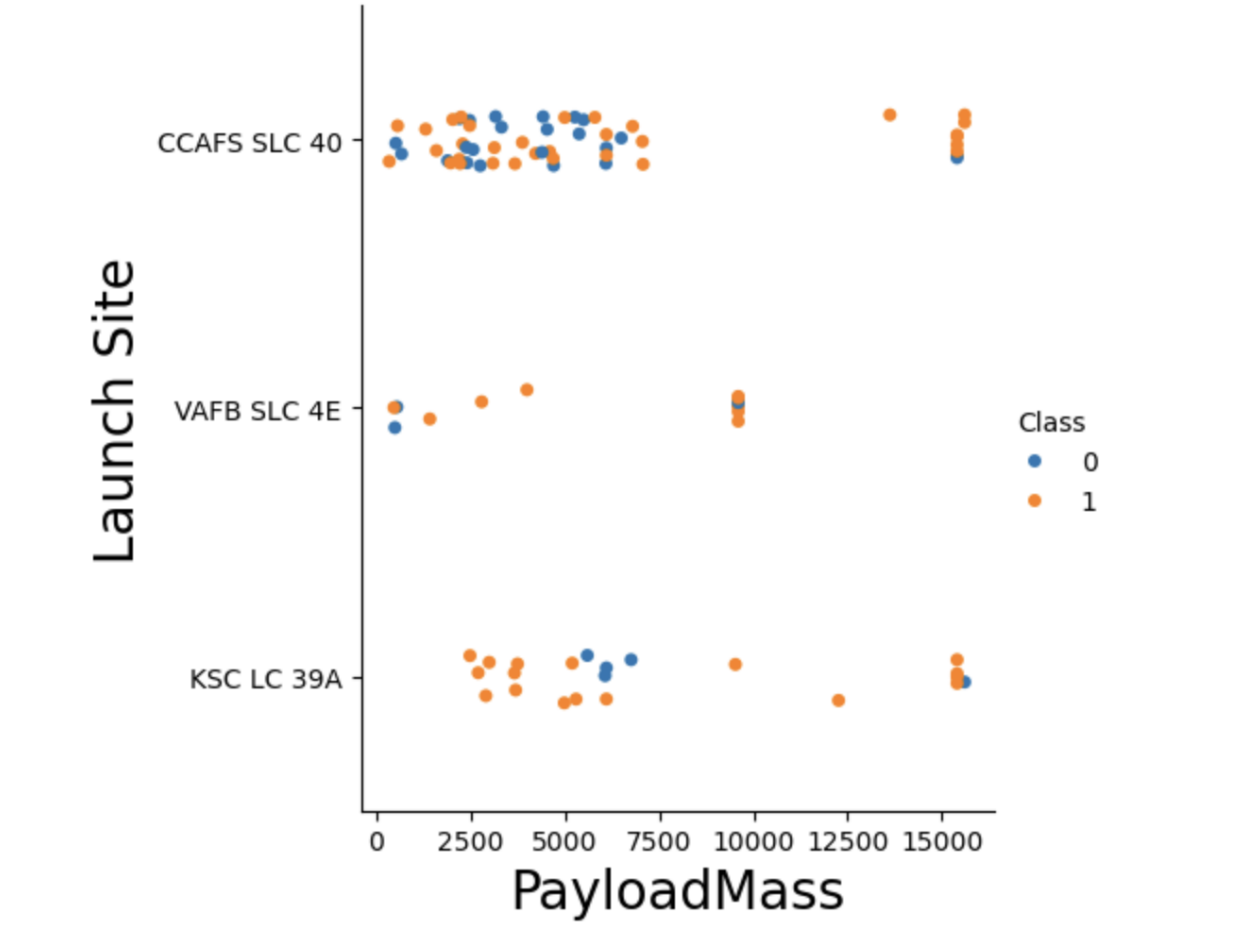
- Launches at site VAFB SLC 4E show the trend of higher number of successes as number of flights increases

- Launches at site KSC LC 39A all have number of flights over 20.



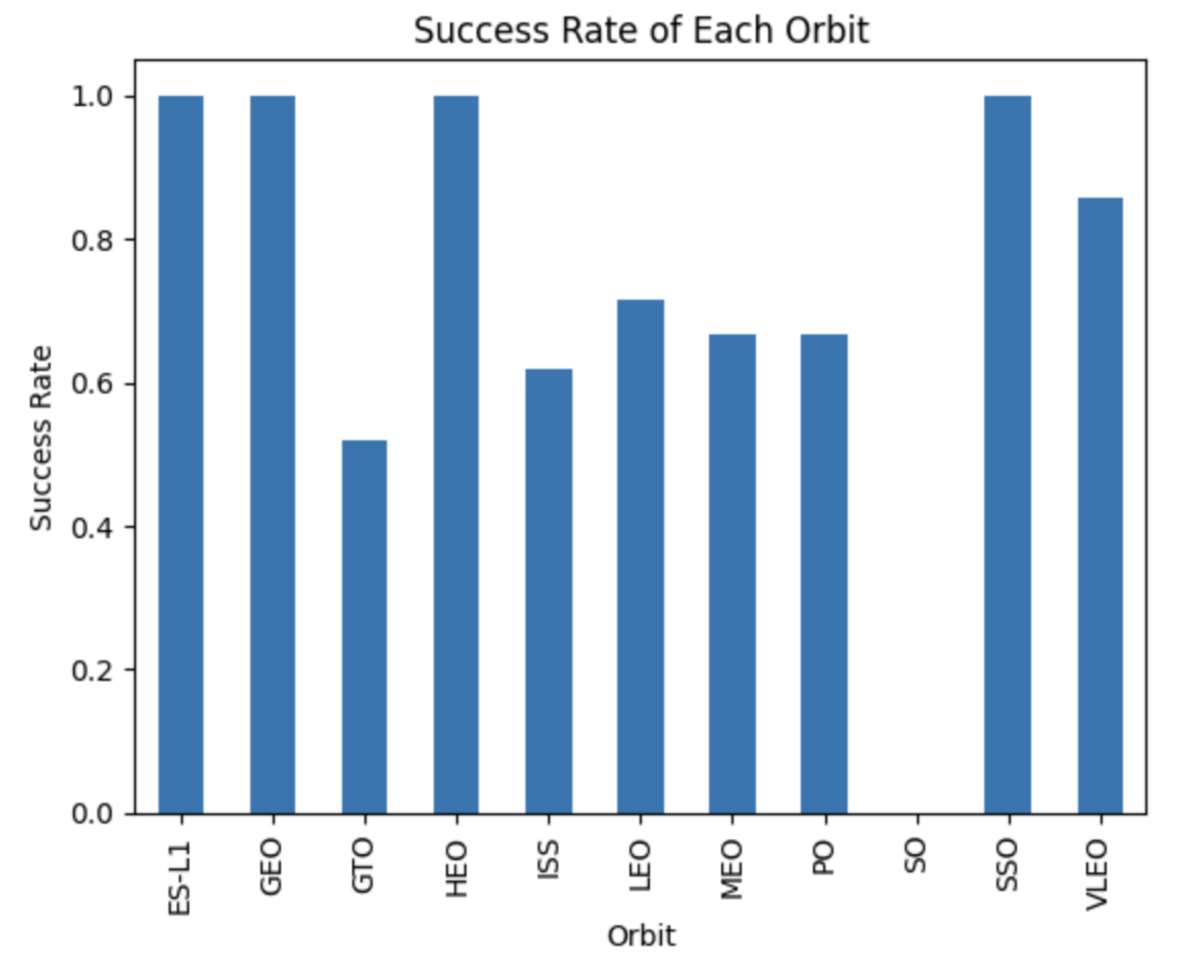
Payload vs. Launch Site

- For site CCAFS SLC 40, there is no payload mass between 7,500 and 12,500. There is also no relationship between Payload Mass and launch successes for this site.
- For site KSC LC 39A, all Payload Mass are above 2,500.
- For both sites VAFB SLC 4E and KSC LC 39A, there are more successes than failures.



Success Rate vs. Orbit Type

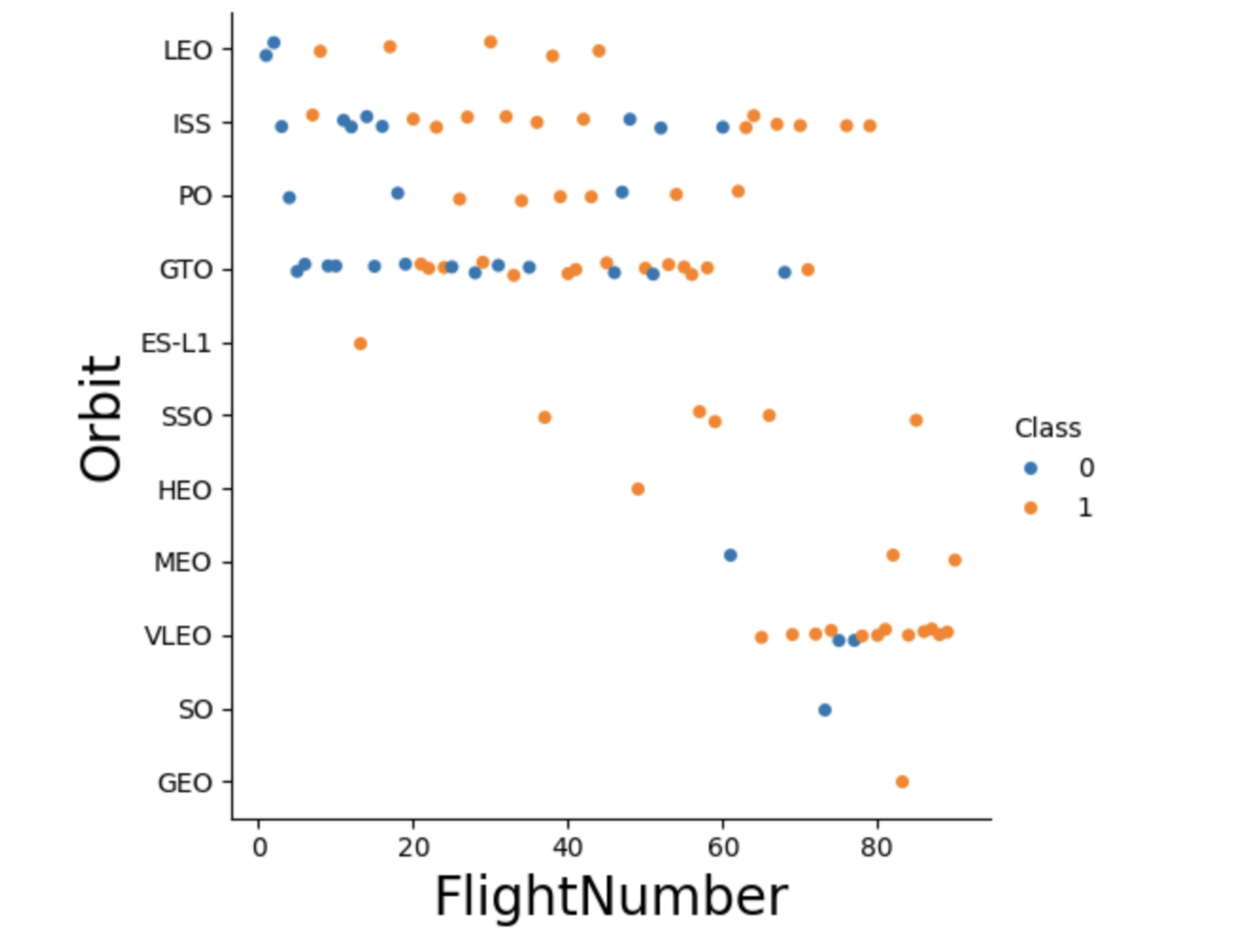
- The success rates are the highest for the orbits ES-L1, GEO, HEO and SSO, all at 1.
- However GEO data is unreliable due to it only has one record and therefore could be an outlier.
- Orbit SO has the lowest success rate of 0. However later chart shows that SO only has 1 launch so it can be considered an outlier.
- GTO will be the orbit with the lowest success rate after taking out SO.



Flight Number vs. Orbit Type

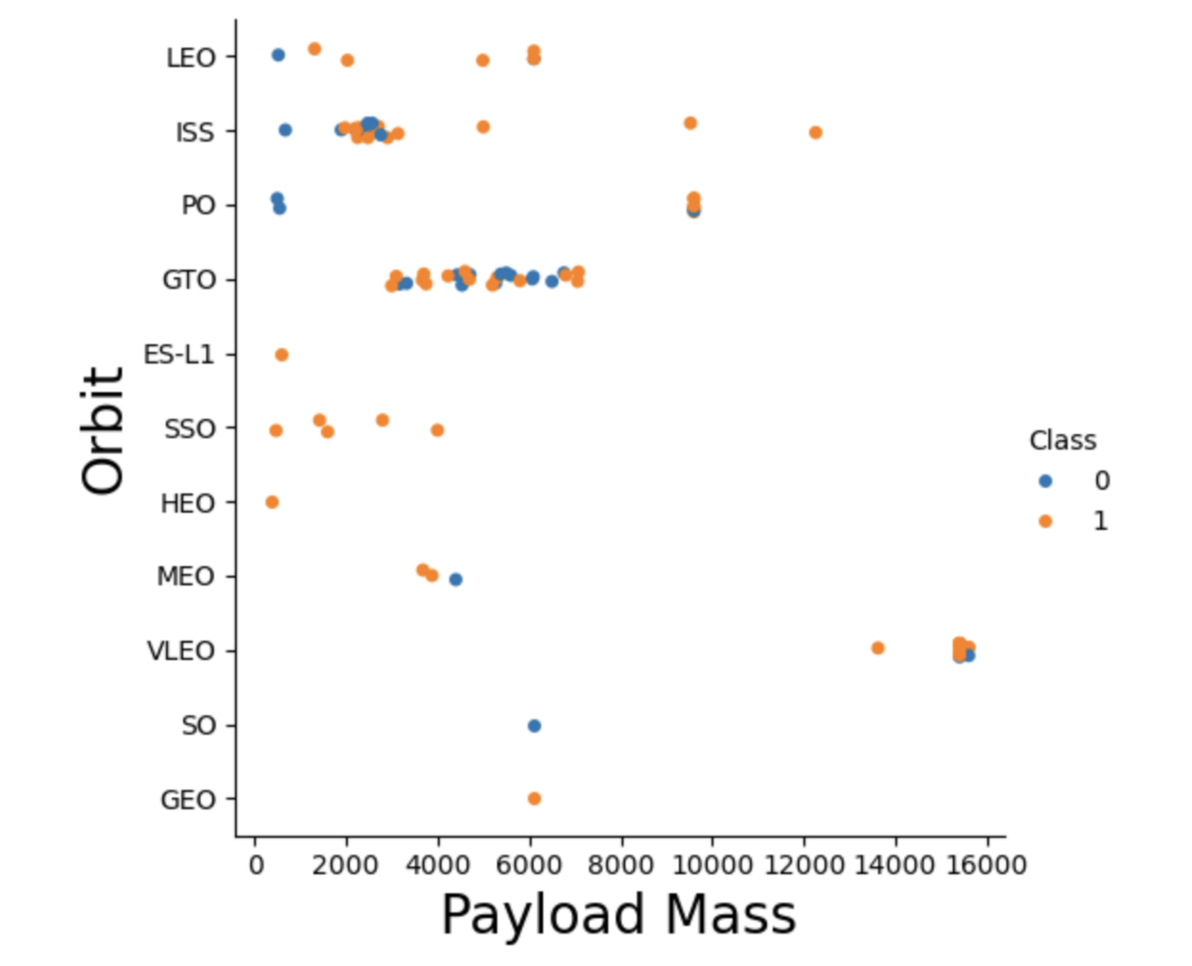
- SSO and HEO orbits have launches with flight number above 40. VLEO and SO orbits have launches with flight number above 60. GEO orbit has launches with flight number above 80. These orbits all have more successes than failures.

- LEO, ISS, PO, GTO, and ES-L1 have launches with flights numbers from 0 to 80. There is no significant relationship in number of successes and flight number for these orbits.



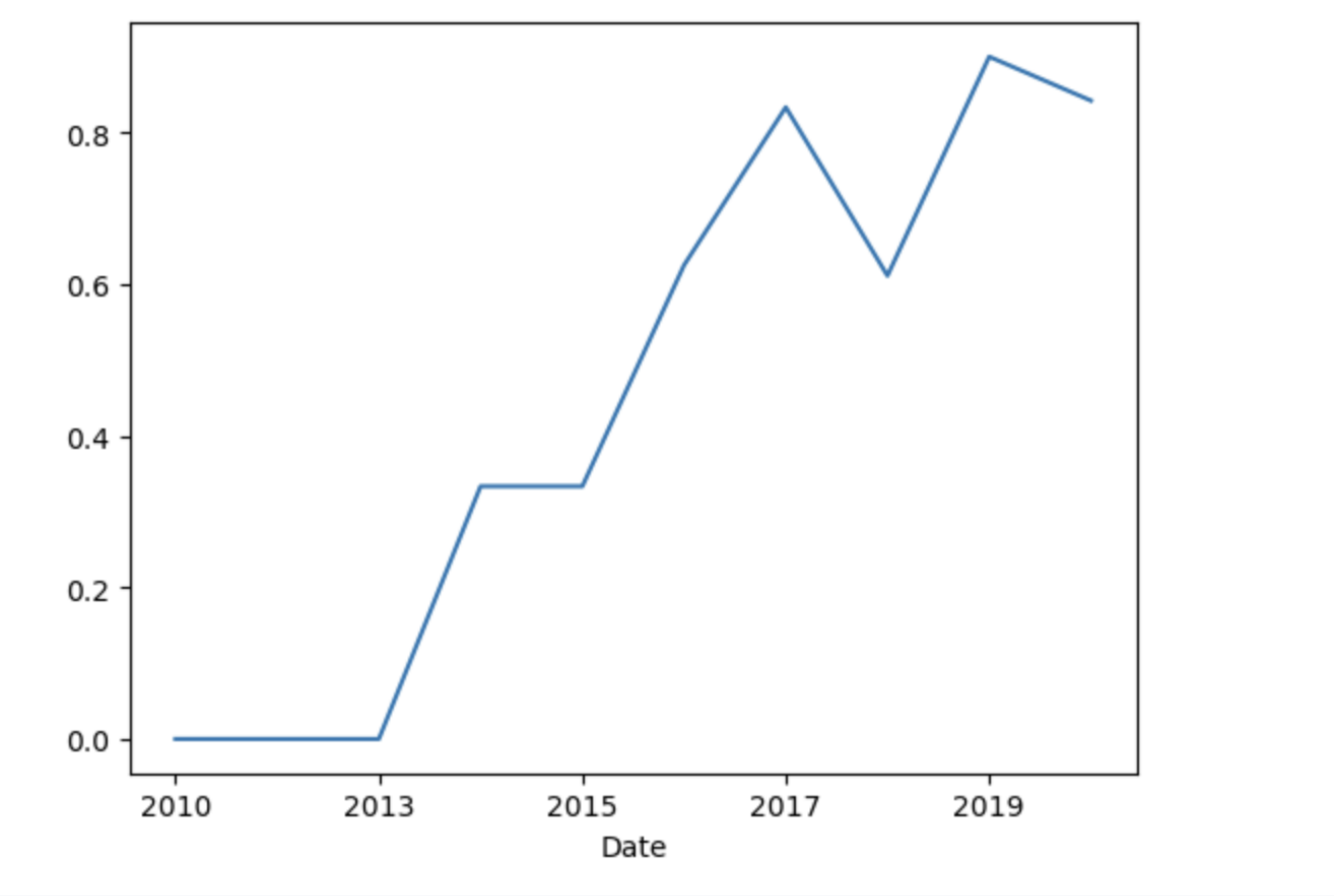
Payload vs. Orbit Type

- LEO, ISS, PO all have payload mass from 0 to 12,000 with more success than failure at higher payload mass
- GTO has payload mass from 2000 and 8000 and no relationship in success and payload mass
- ES-L1, SSO, HEO, and MEO have payload masses from 0 to 4000 with mostly successes.
- VLEO has payload mass above 12,000
- SO and GEO has payload mass of 6000 and only 1 launch at each orbit. Success/failure trend at these orbits cannot be determined without more data.

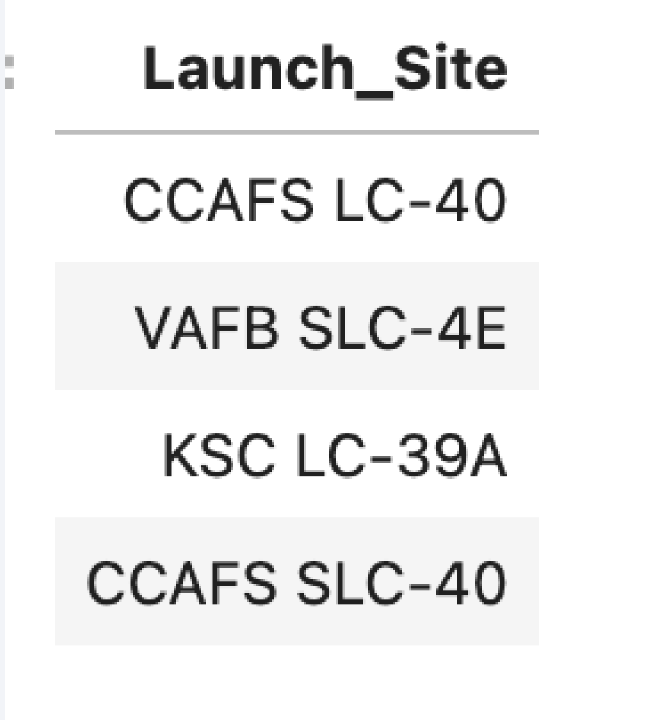


Launch Success Yearly Trend

- Yearly average success rate has been increasing at a drastic rate ever since 2013
- The success rate dropped in 2018 but rose significantly again in 2019.



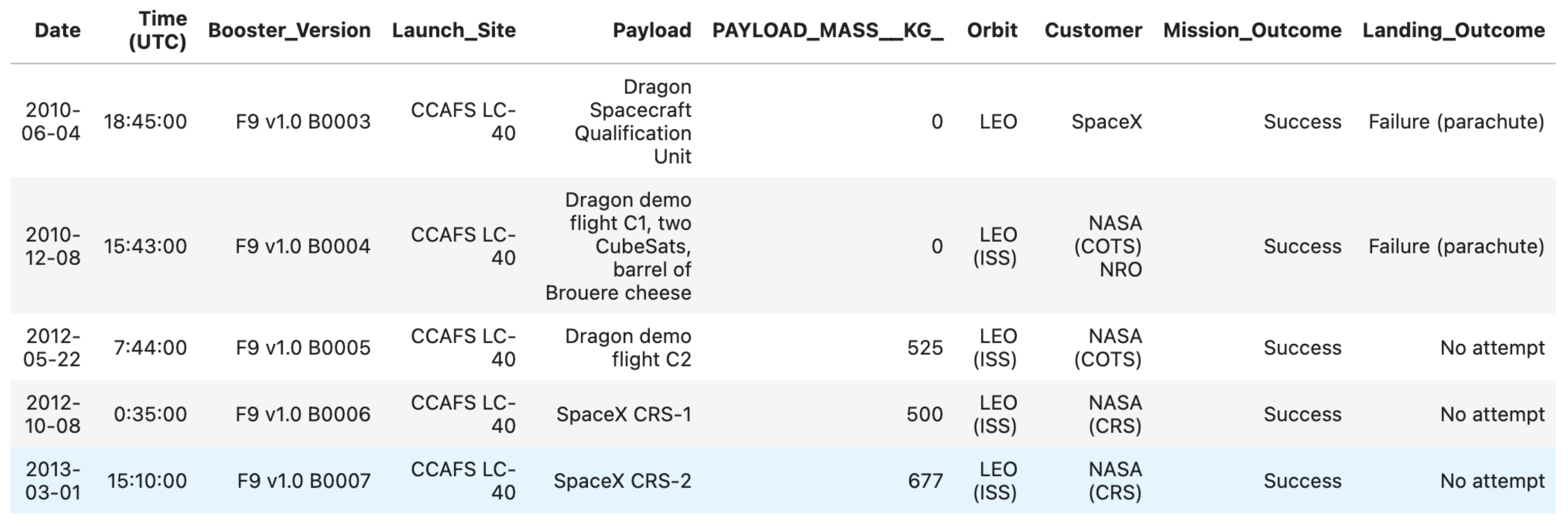
All Launch Site Names



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

- There are 4 launch sites. The SQL query result shown above are their names.

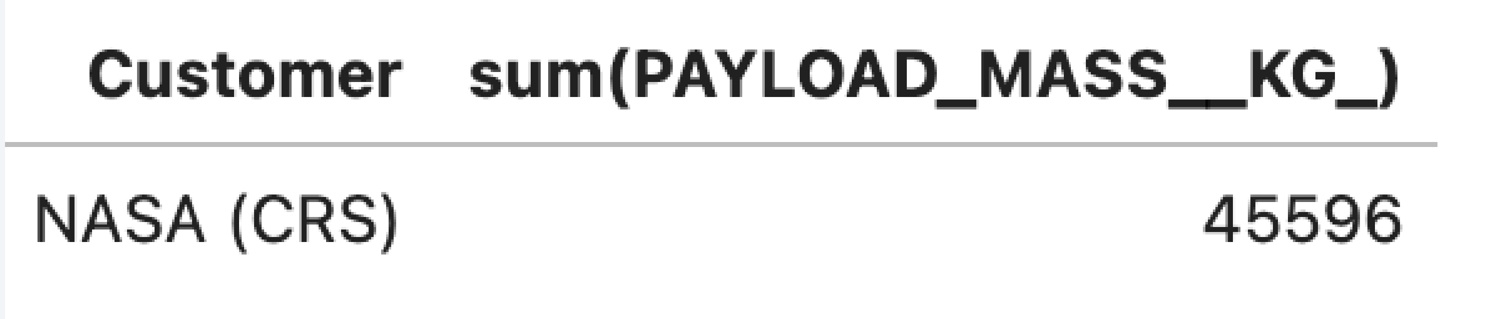
Launch Site Names Begin with 'CCA'



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

- Above is the SQL query result for the top 5 records of launch sites with name that begins with ‘CCA’.

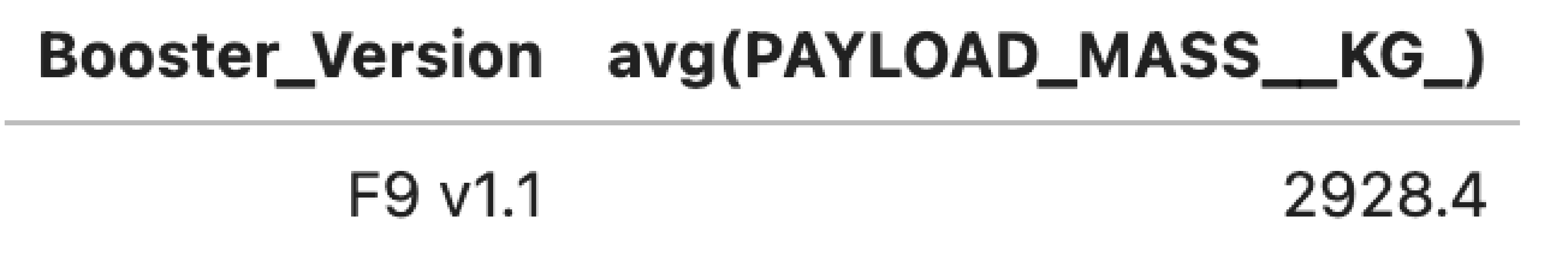
Total Payload Mass



Customer	sum(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

- The above SQL query calculated the total payload mass for customer, NASA, as 45,596 KG.

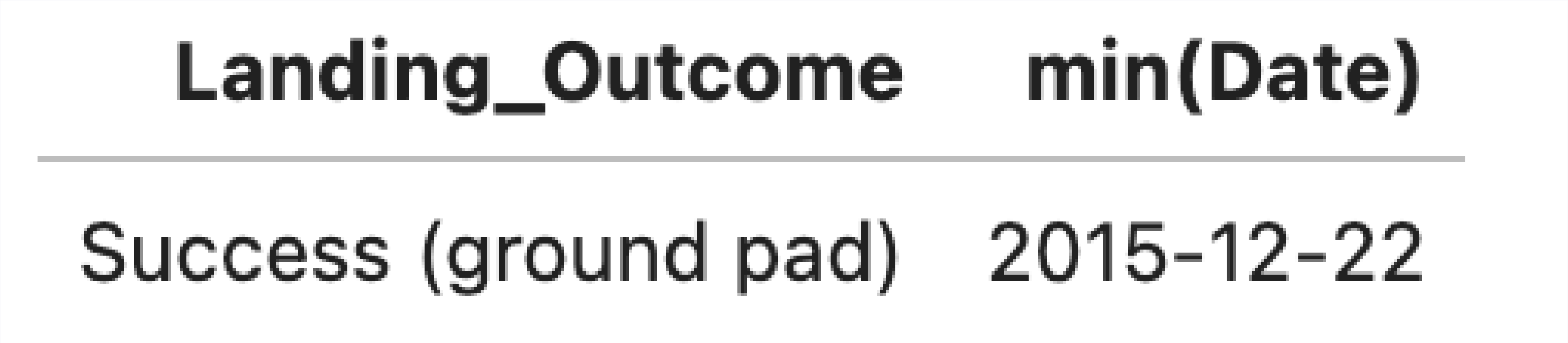
Average Payload Mass by F9 v1.1



Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

- The above SQL query calculated the average payload mass carried by booster version F9 v1.1 as 2928.4 KG.

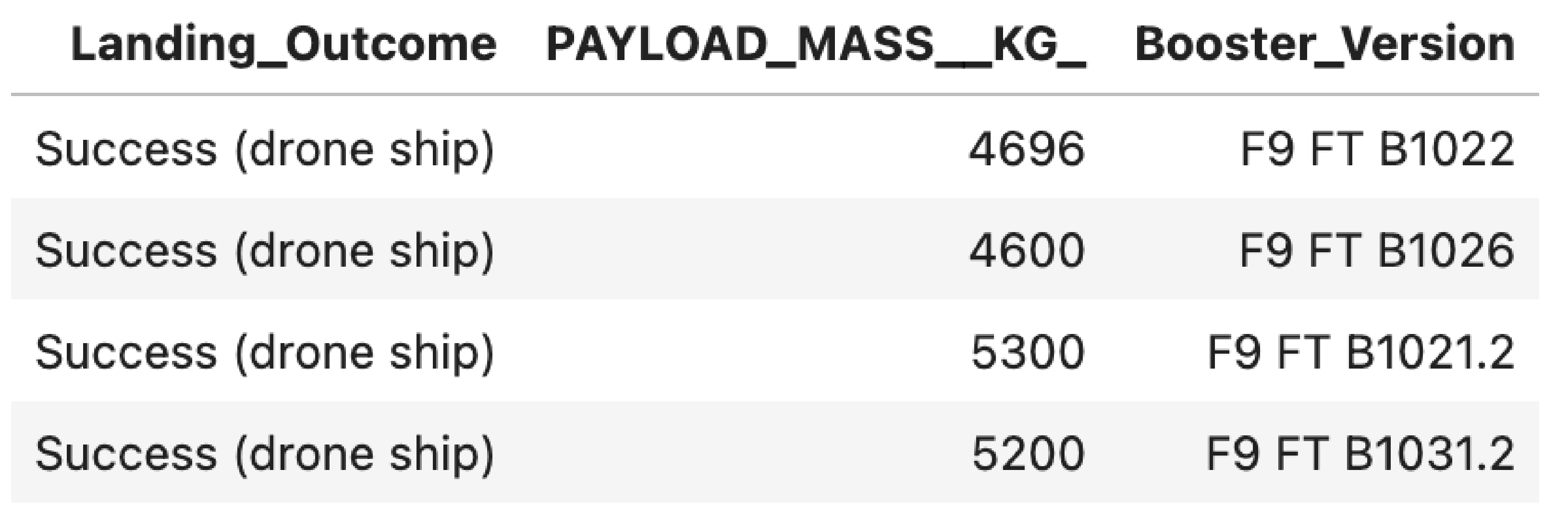
First Successful Ground Landing Date



Landing_Outcome	min(Date)
Success (ground pad)	2015-12-22

- The SQL query above is the date of the first successful ground pad landing is Dec 22nd, 2015.

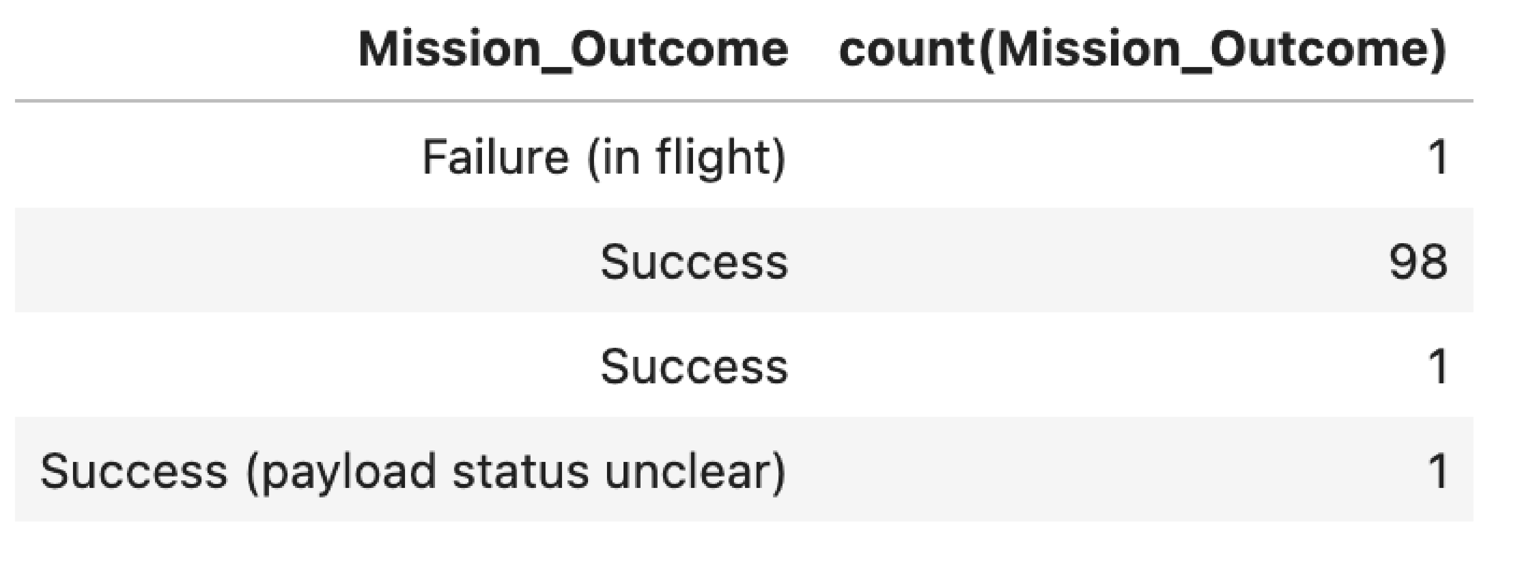
Successful Drone Ship Landing with Payload between 4000 and 6000



Landing_Outcome	PAYLOAD_MASS__KG_	Booster_Version
Success (drone ship)	4696	F9 FT B1022
Success (drone ship)	4600	F9 FT B1026
Success (drone ship)	5300	F9 FT B1021.2
Success (drone ship)	5200	F9 FT B1031.2

- SQL query retrieved the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 as F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

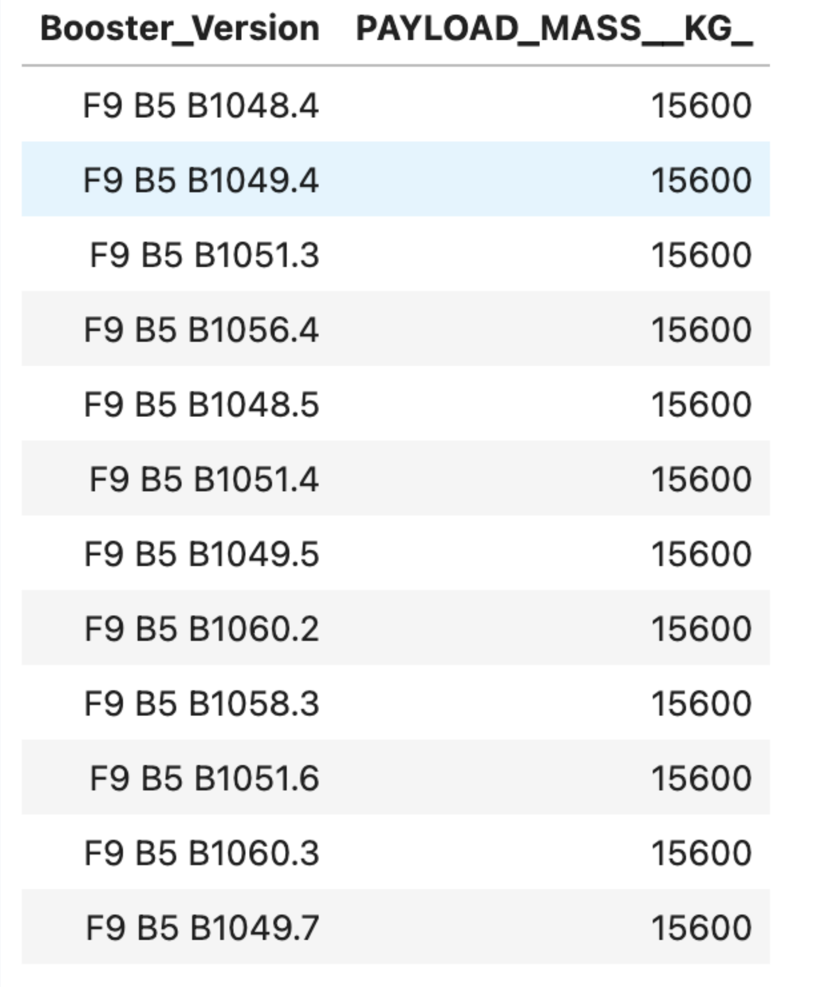
Total Number of Successful and Failure Mission Outcomes



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

- The above table retrieved through SQL query shows the total number of successful and failure mission outcomes.

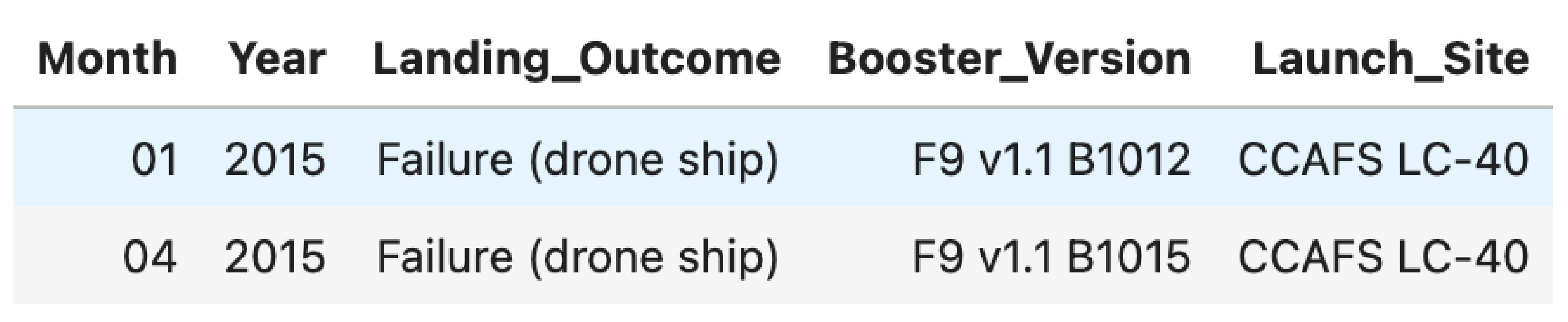
Boosters Carried Maximum Payload



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

- To the left is a list of Booster Versions that carried the maximum Payload Mass of 15,600 KG, retrieved through SQL query.

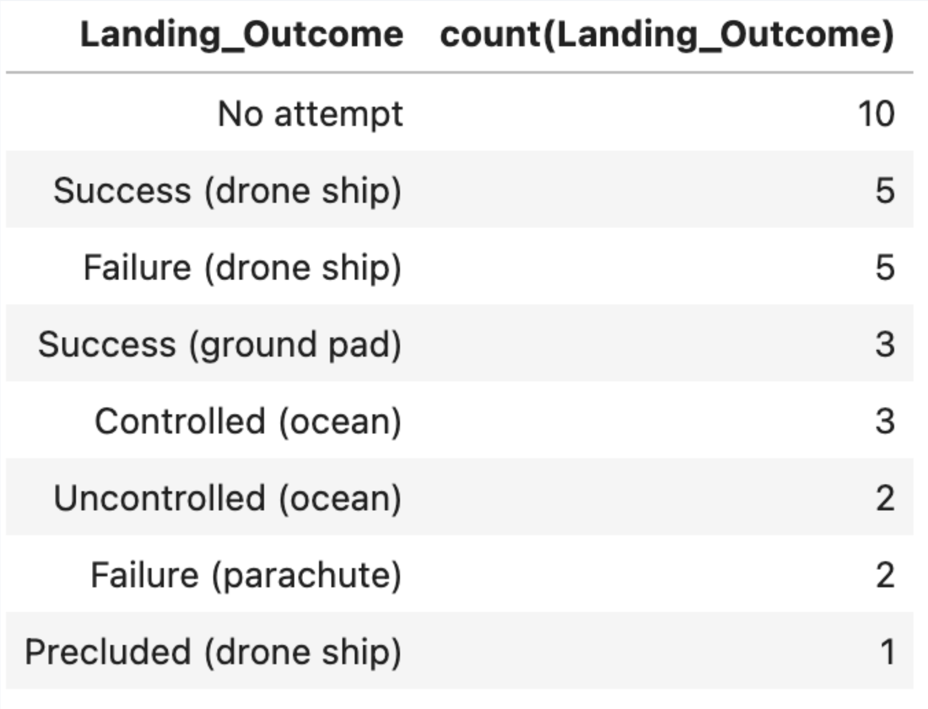
2015 Launch Records



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

- Above are the records that show the months, failed landing outcomes in drone ship, their booster versions, and launch site names in year 2015, retrieved through SQL query.

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



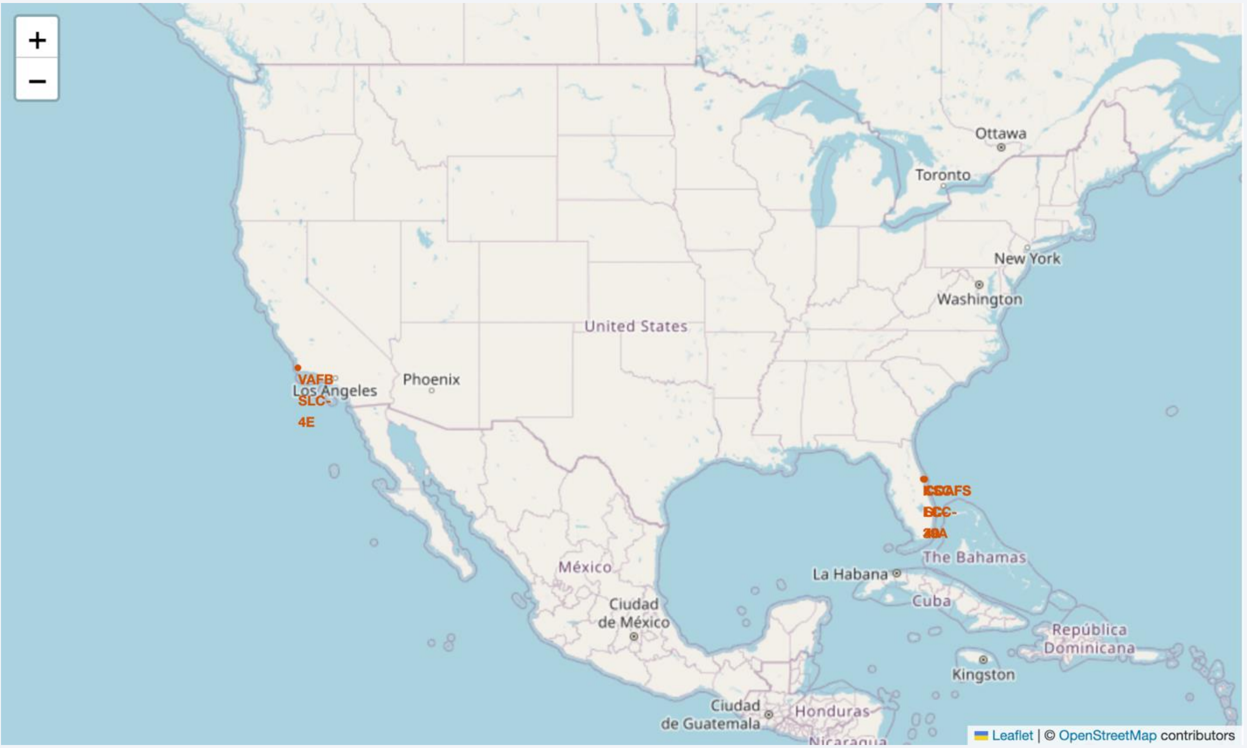
Landing_Outcome	count(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Above is the counts 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, retrieved through SQL query.

Section 3

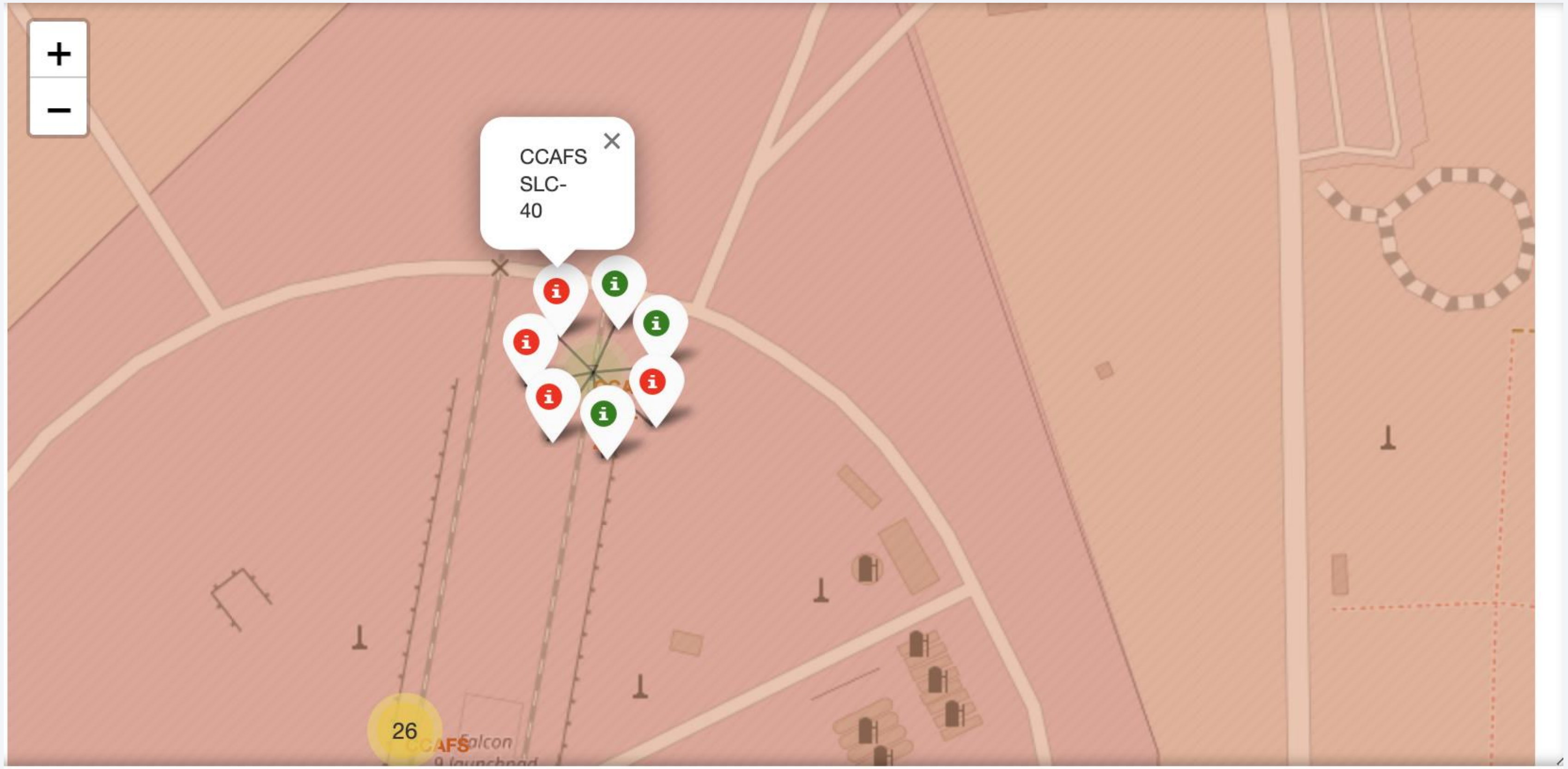
Launch Sites Proximities Analysis

Launch Sites Locations



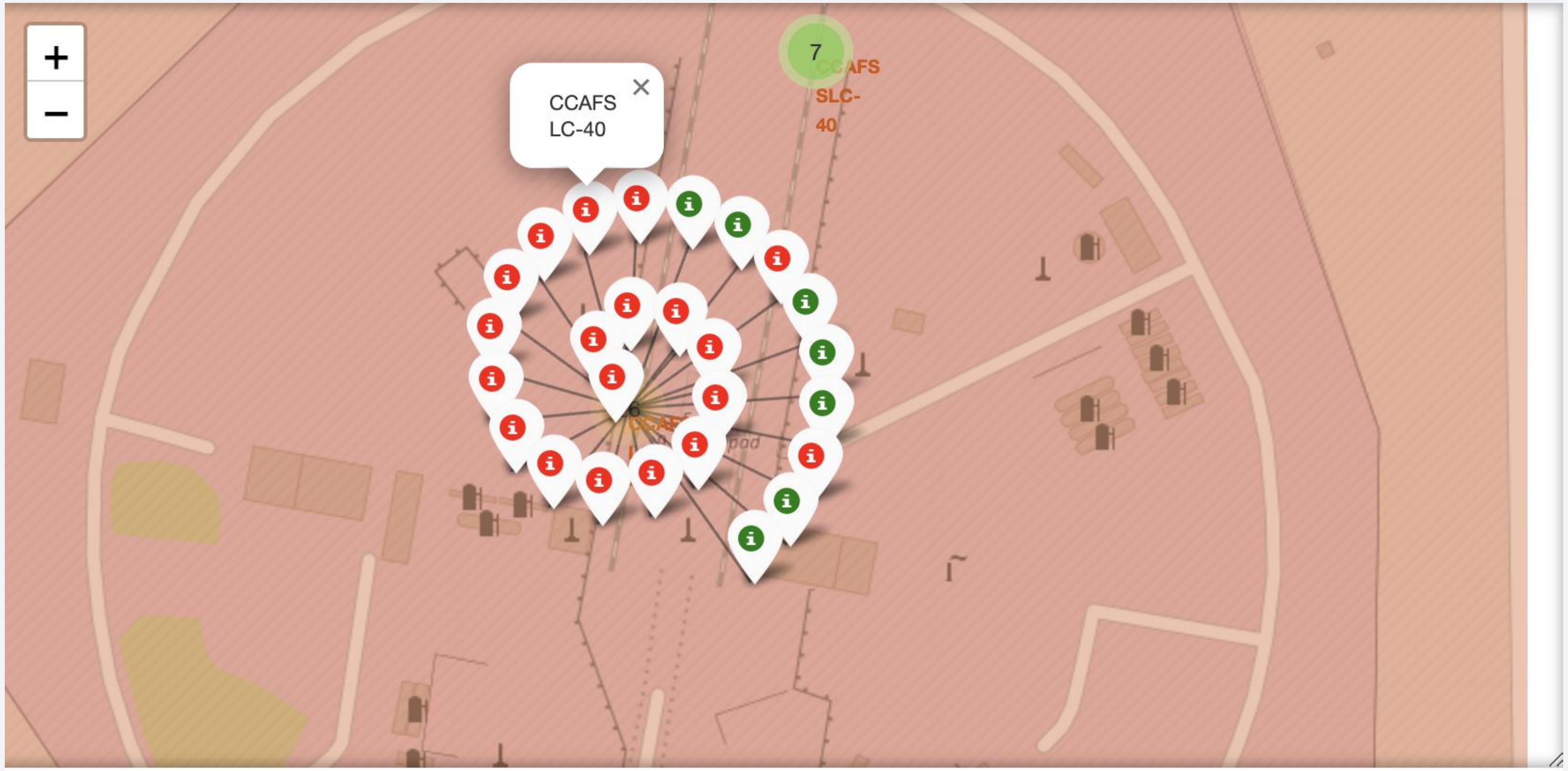
- Above image is United States on the global map. The spots marked in red are all the SpaceX launch sites.
- The launch sites are all on the edge of the United States with close proximity to the sea, either on the West Coast (California) or the East Coast (Florida).
- The launch sites are also all relatively close to the Equator line.

CCAFS SLC – 40 Launch Outcomes



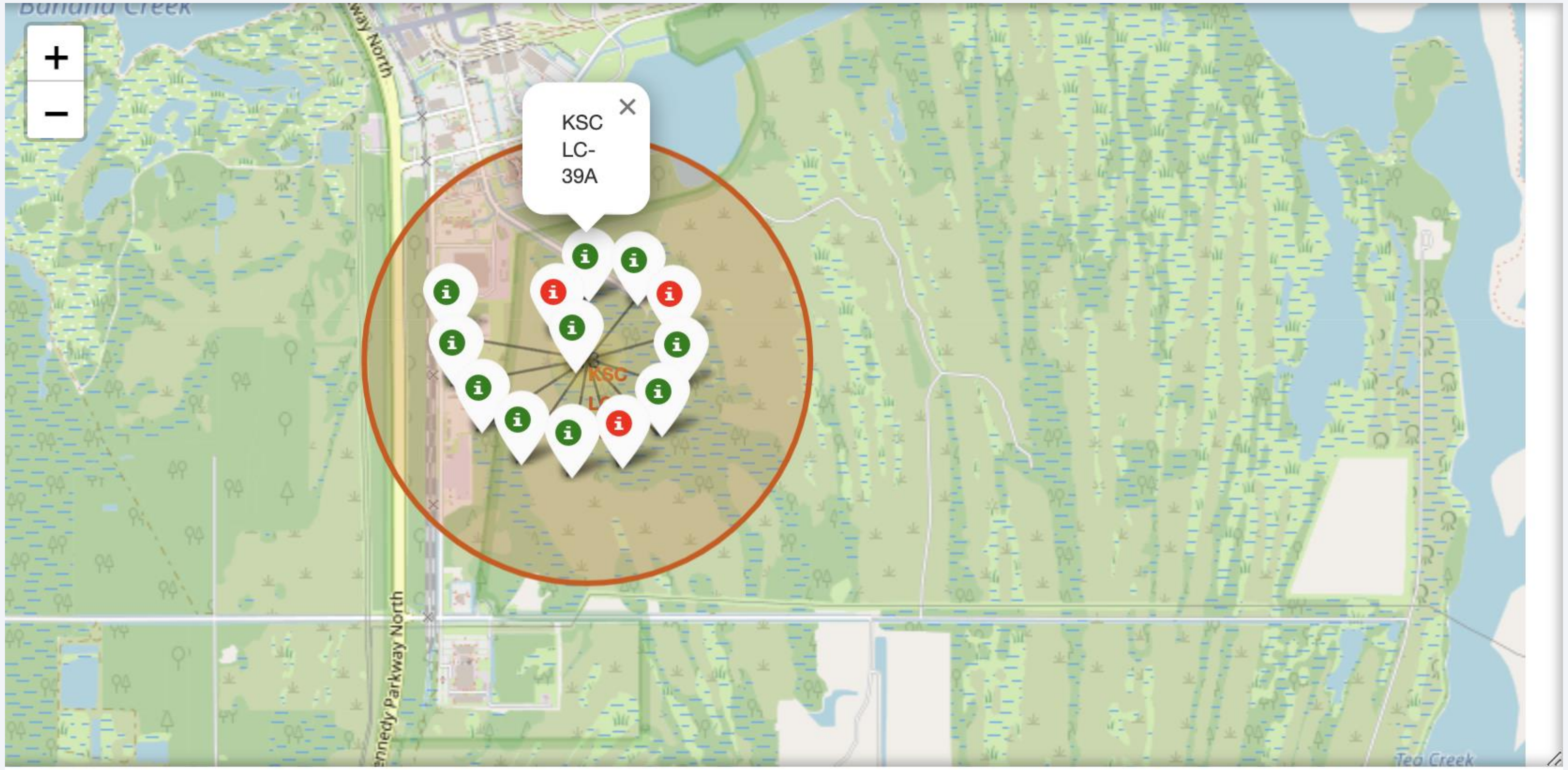
- Image above shows the launch outcomes for site CCAFS SLC – 40.
- There are 7 launches at this site. Green markers mark the successes and red markers mark the failures.

CCAFS LC – 40 Launch Outcomes



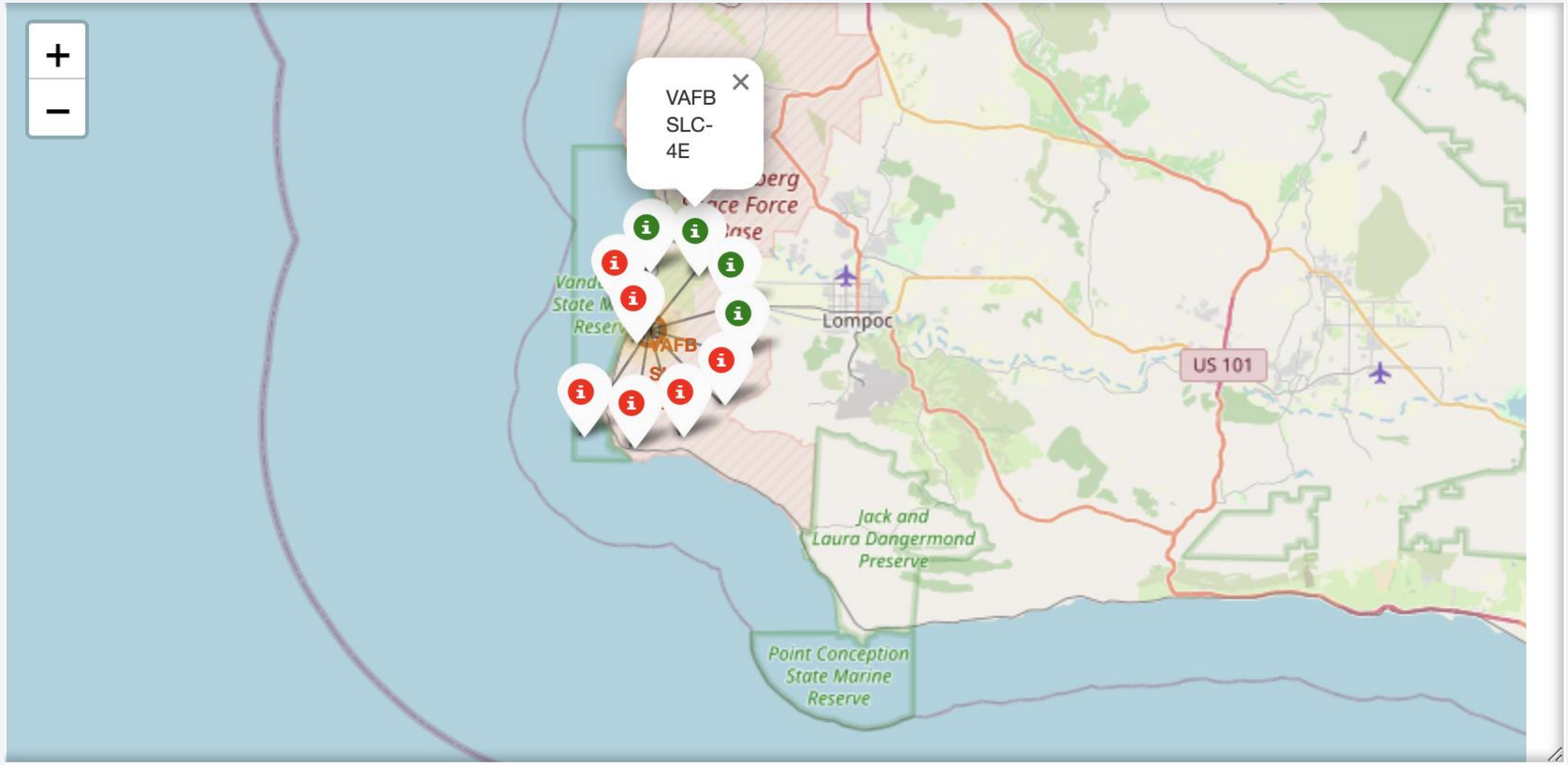
- Image above shows the launch outcomes for site CCAFS LC – 40.
- There are 26 launches at this site. Green markers mark the successes and red markers mark the failures.

KSC LC – 39A Launch Outcomes



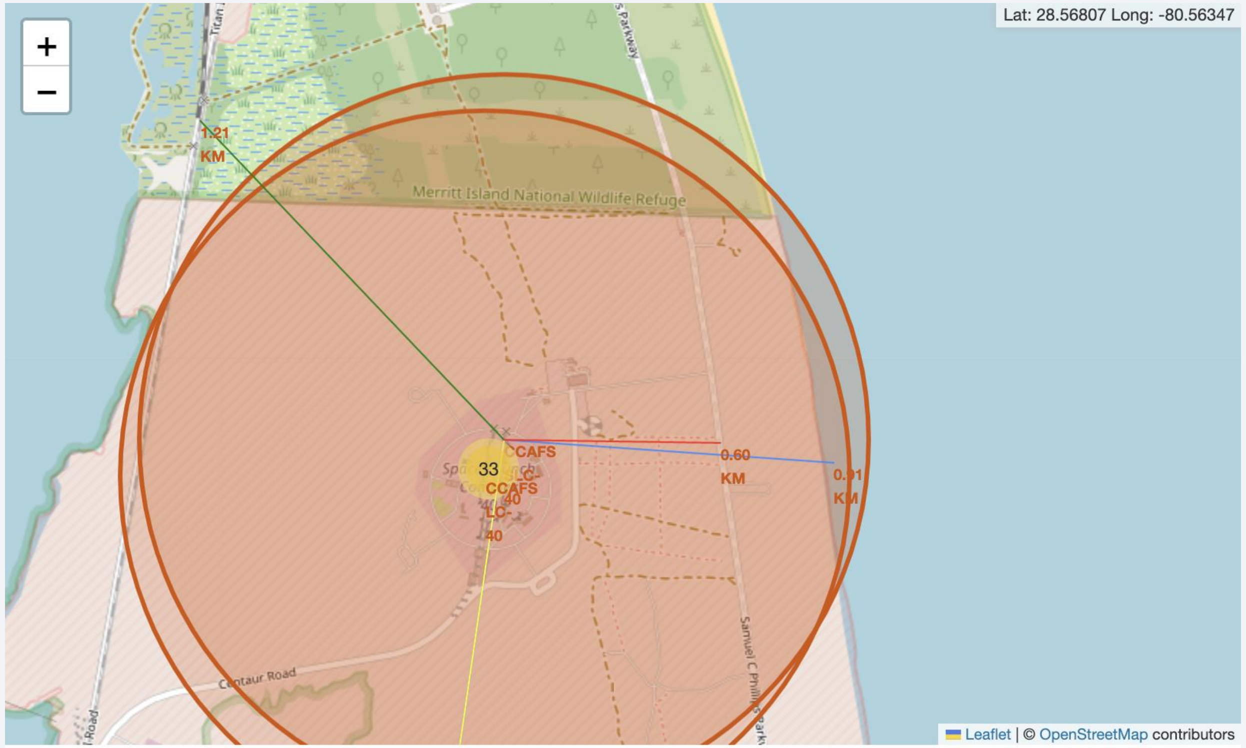
- Image above shows the launch outcomes for site KSC LC – 39A.
- There are 13 launches at this site. Green markers mark the successes and red markers mark the failures.

VAFB SLC – 4E Launch Outcomes



- Image above shows the launch outcomes for site VAFB SLC – 4E.
- There are 10 launches at this site. Green markers mark the successes and red markers mark the failures.

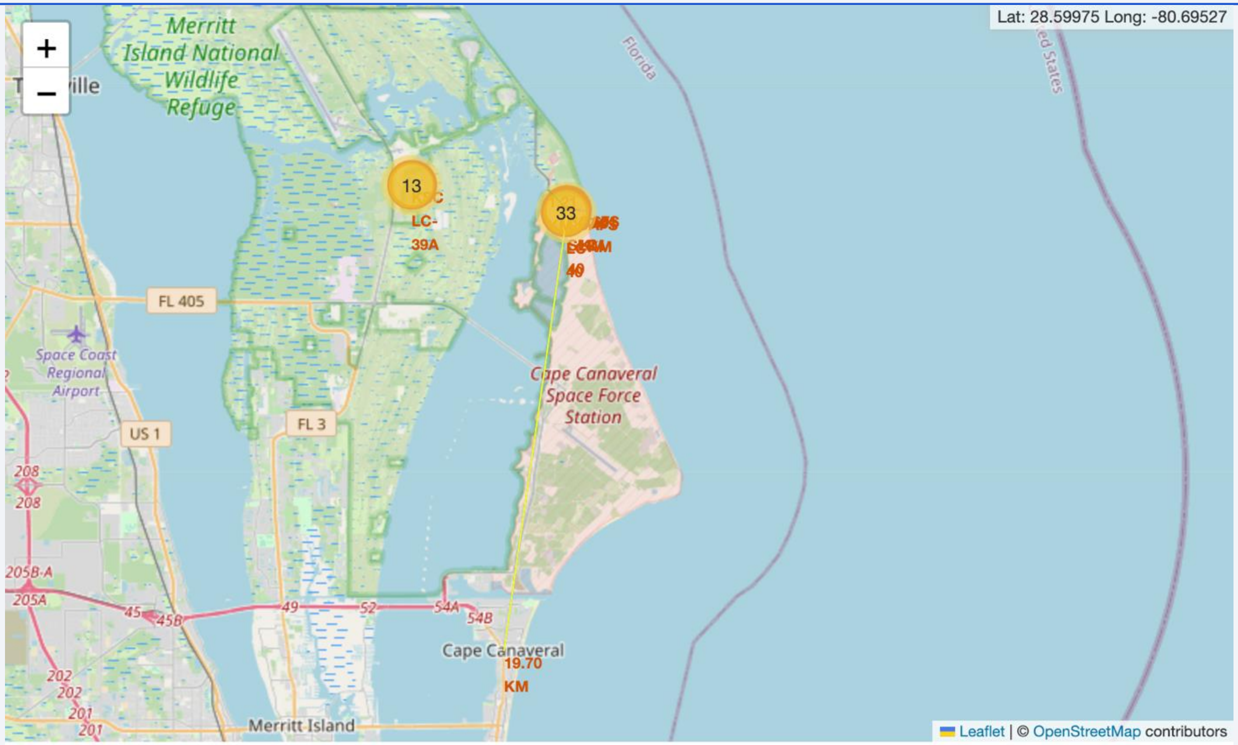
Proximity to Railroad, Highway, or Coastline


Lat: 28.56807 Long: -80.56347
+
−
1.21 KM
Merritt Island National Wildlife Refuge
33
CCAFS LC-40
CCAFS LC-40
0.60 KM
0.91 KM
Centaur Road
Samuel C Phillips Parkway
Leaflet | © OpenStreetMap contributors

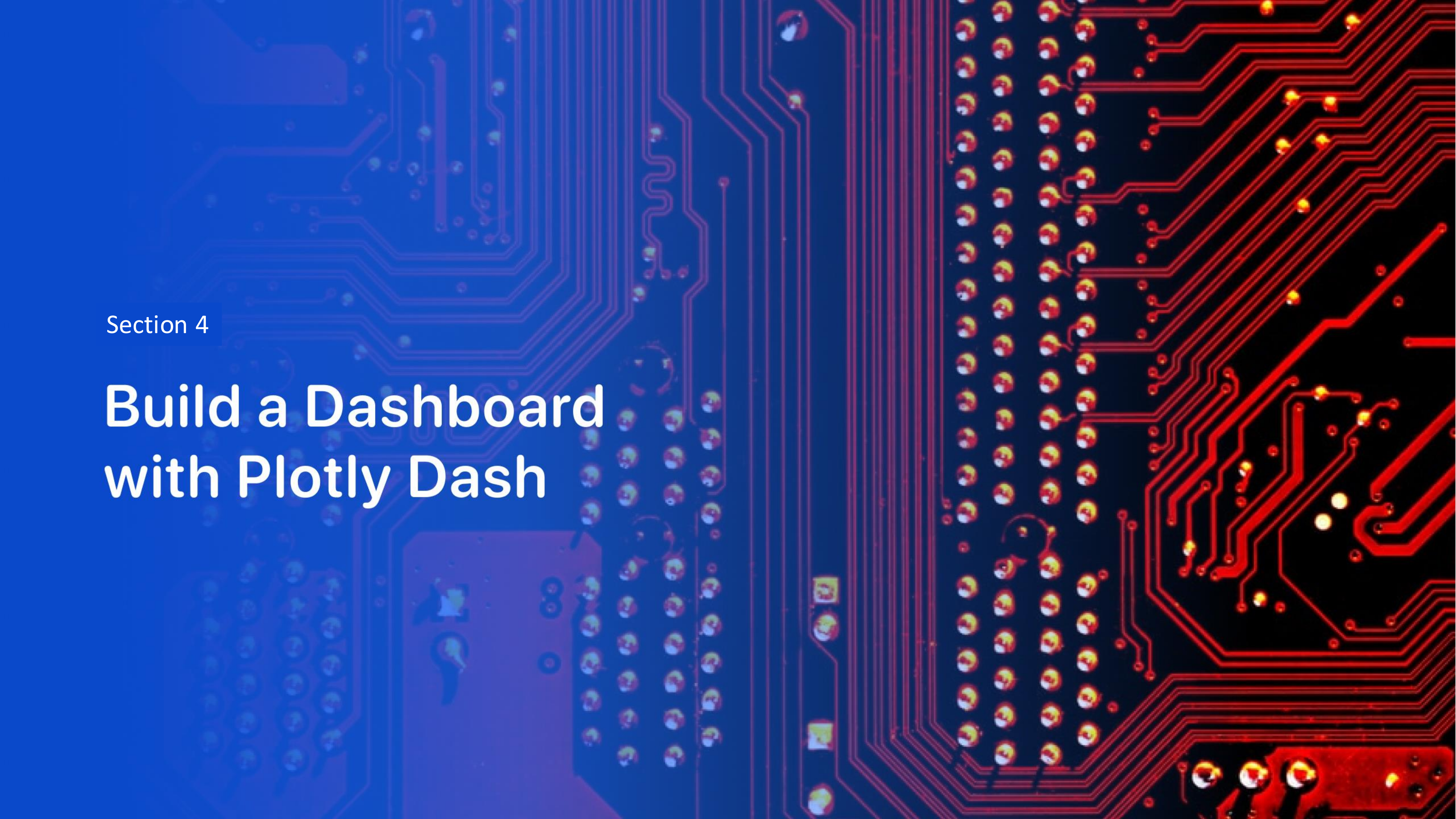
- Chart above shows the distance from CCAFS SLC – 40 to the closest railroad, highway, or coastline:
  - Proximity to coastline is marked in a blue line with a distance of 0.91km
  - Proximity to railroad is marked in a green line with a distance of 1.21km
  - Proximity to highway is marked in a red line with a distance of 0.60km

- CCAFS SLC – 40 is very, very close to coastline, railroad, or the highway.

Proximity to City


Lat: 28.59975 Long: -80.69527
+
−
Merritt Island National Wildlife Refuge
13
LC-39A
33
Florida
FL 405
Space Coast Regional Airport
US 1
FL 3
Cape Canaveral Space Force Station
Cape Canaveral
19.70 KM
Merritt Island
Leaflet | © OpenStreetMap contributors

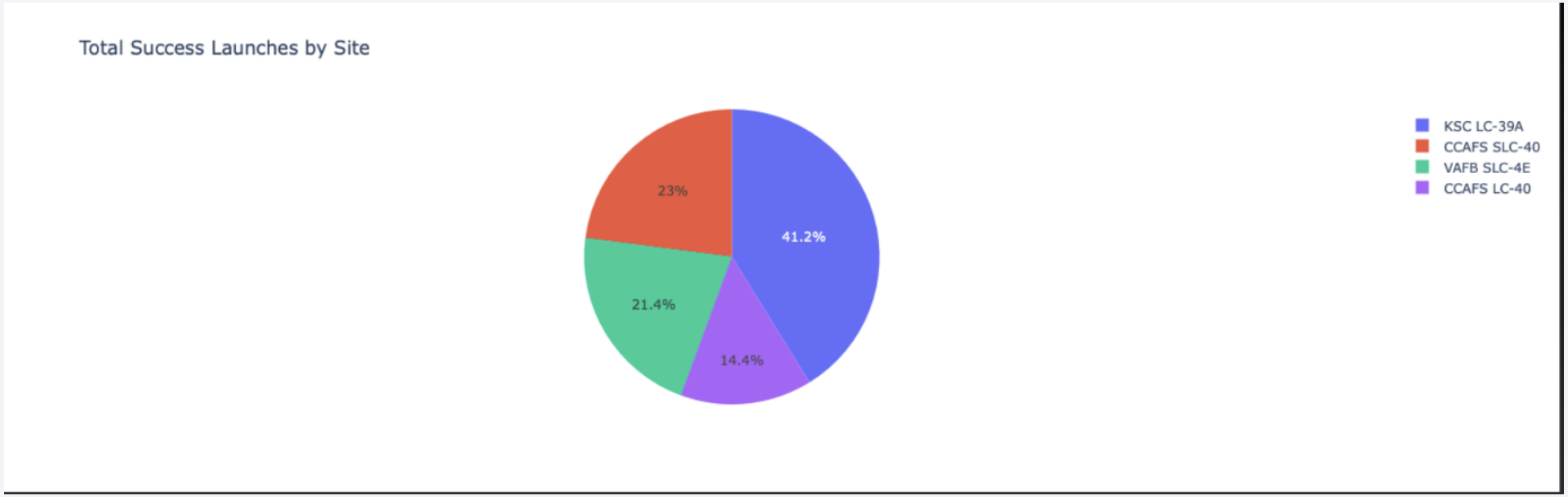
- Chart above shows the distance from CCAFS SLC – 40 to the closest city, Cape Canaveral. The line is marked in yellow. The distance is 19.70km.
- CCAFS SLC – 40 is quite a distance from its closest city, therefore launches from this site are at low risk of impacting the lives of people living in Cape Canaveral.



Section 4

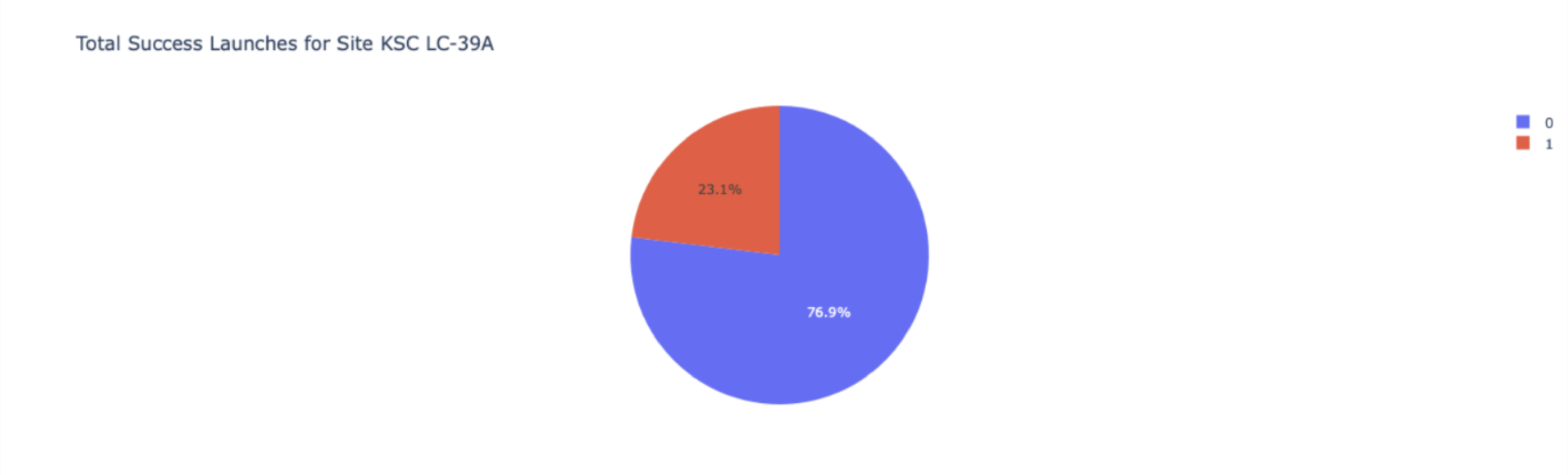
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- The chart above shows that the site with the most success launches is KSC LC-39A.

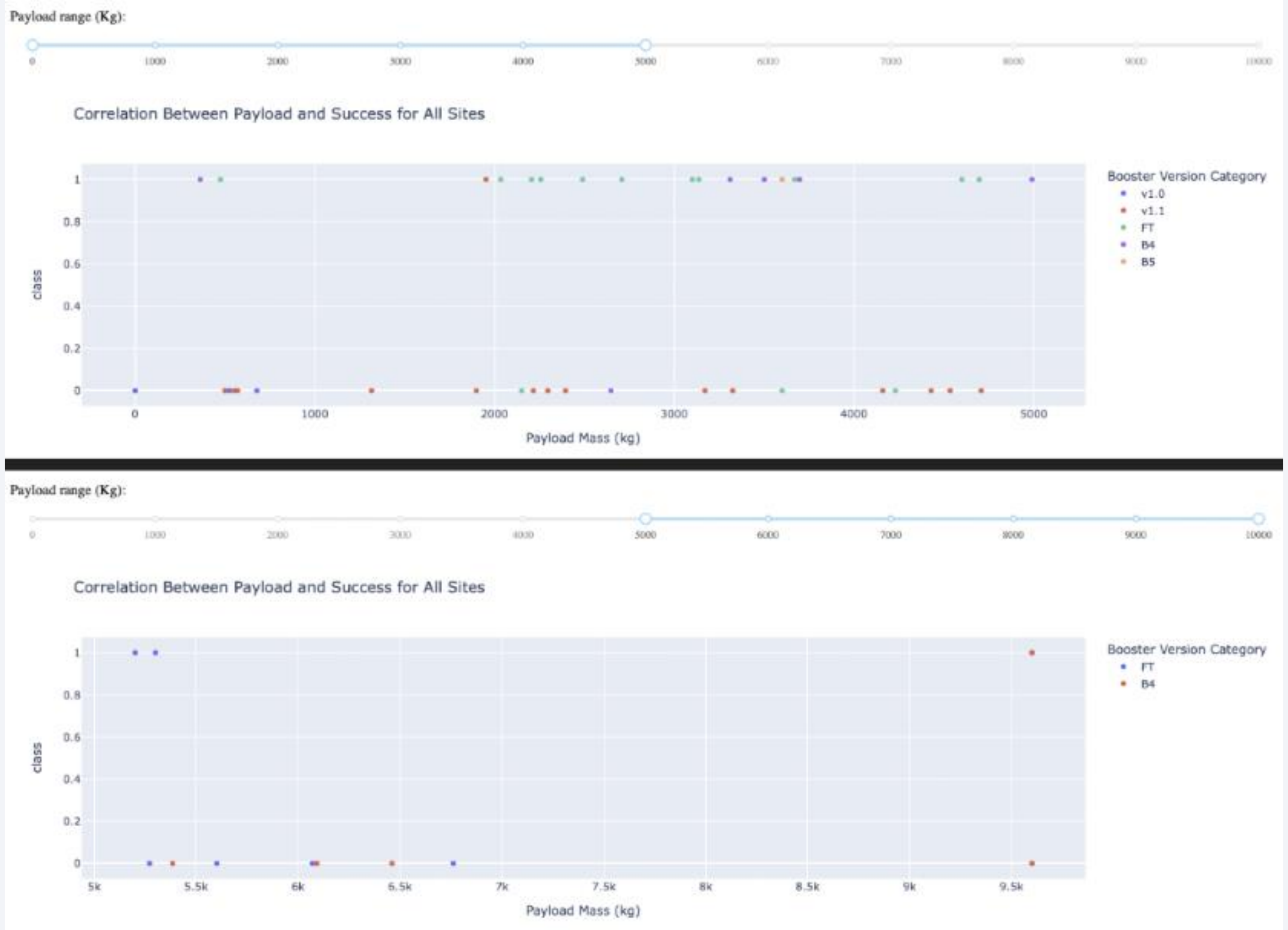
Success Launches Ratio KSC LC-39A



- KSC LC-39A, the site with the highest total success launches, has a success ratio of 23.1%.

Payload Mass vs Launch Outcome

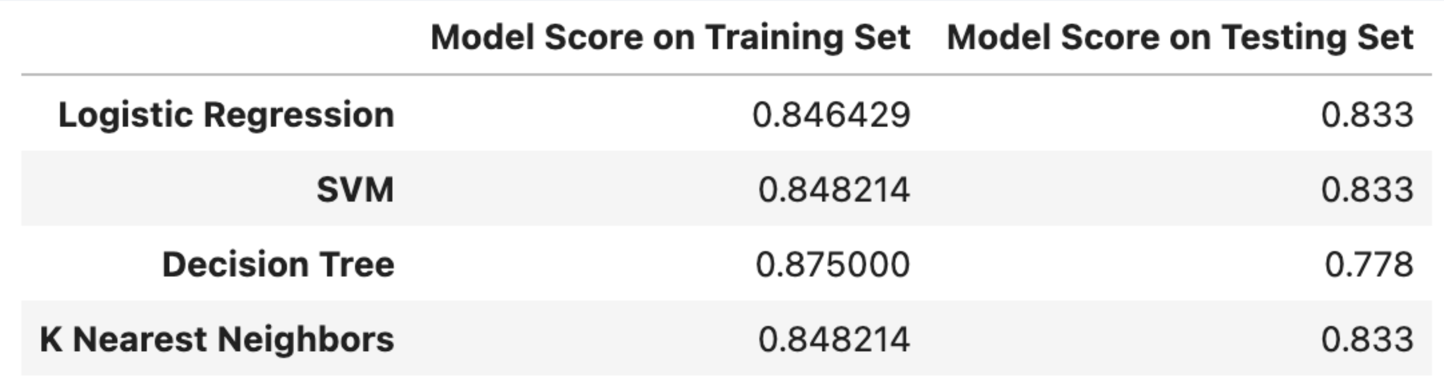
- The most successes happen when payload range is between 2,000 and 4,000.
- Booster version category FT (in green in the chart) has the highest number of successes.



Section 5

Predictive Analysis (Classification)

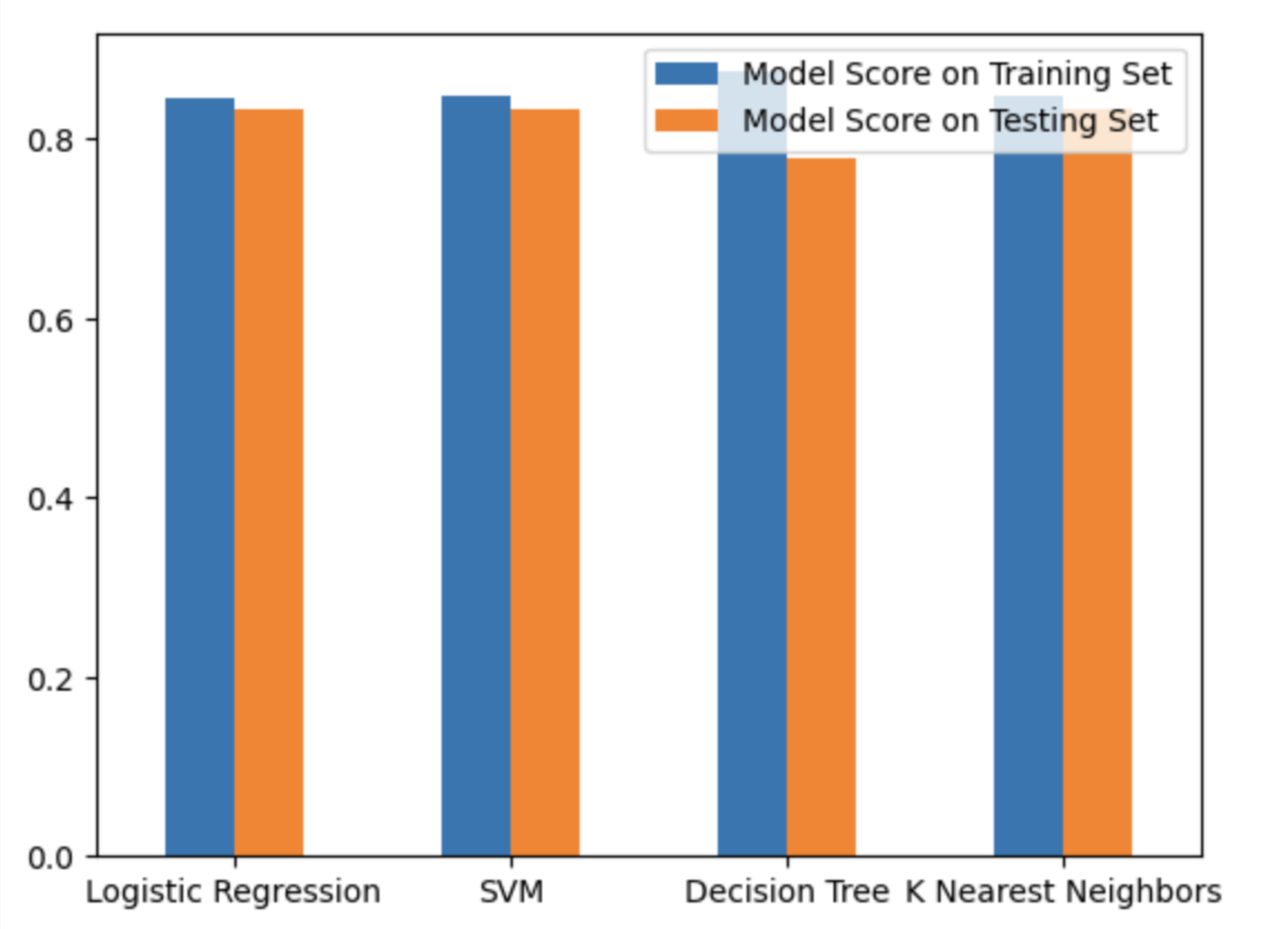
Classification Accuracy – Score Table



	Model Score on Training Set	Model Score on Testing Set
Logistic Regression	0.846429	0.833
SVM	0.848214	0.833
Decision Tree	0.875000	0.778
K Nearest Neighbors	0.848214	0.833

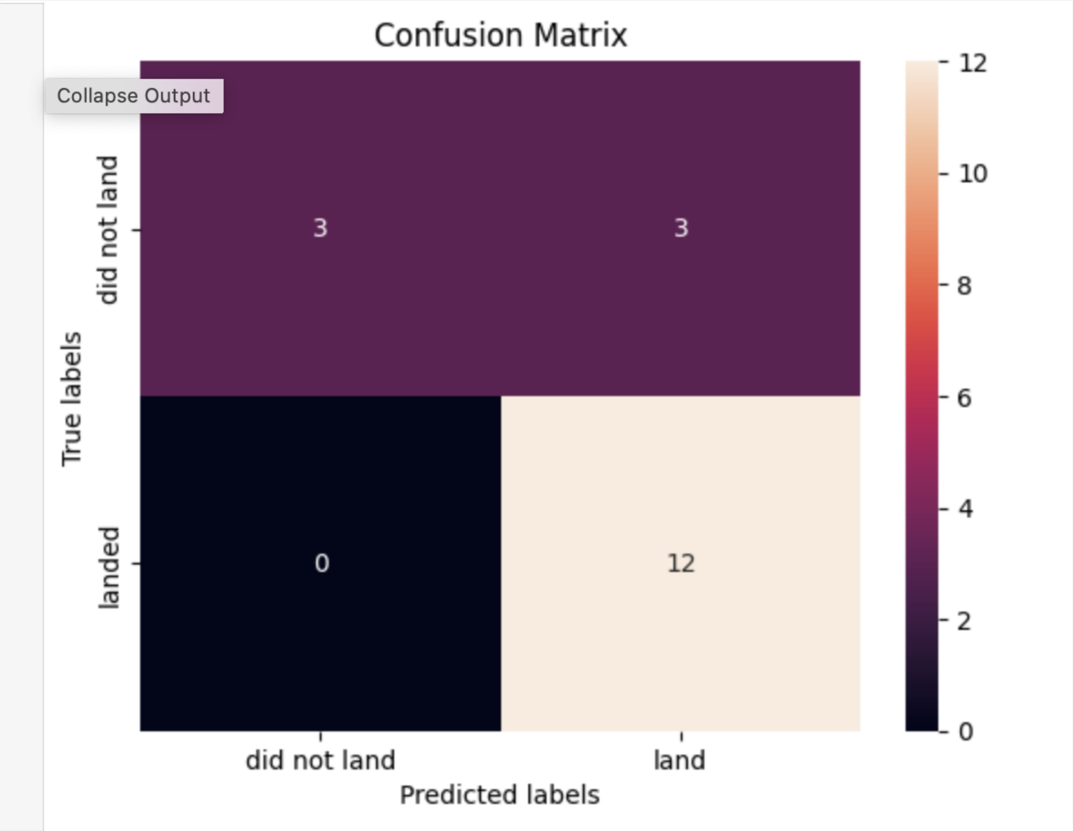
- Decision Tree has the highest training model score of 0.875 but low testing score of 0.778 indicating over-training. Therefore it is not the best model.
- The rest of the three all have the same testing score of 0.833.
- SVM and K Nearest Neighbors both have the same training score of 0.848214, higher than that of Logistic regression.

Classification Accuracy – Chart



- The accuracy score chart shows similar trend as the score table.
- Decision tree model is over-trained.
- SVM and K Nearest Neighbors have both have slightest higher training score than logistic regression. Selecting both of them as the best-performing models.

Confusion Matrix



- SVN and K Nearest Neighbors models not only have the same accuracy scores, but also have the same confusion matrix.
- There are 12 out of 12 (100%) true positives and 3 out of 6 (50%) false positives.

Conclusions

- Number of launch successes increases as number of flights increases.
- Orbits ES-L1, HEO, and SSO have the highest success rates.
- Yearly average success rate has been rising significantly since 2013.
- All the launch sites are located near the coastlines and close to the equator line
- Site KSC LC-39A has the most success launches.
- Lower payload mass have more success launches than higher payload mass.
- The best machine learning models to make predictions on launch success are SVM and K Nearest Neighbors.

Appendix

This project is conducted under well-written guidance from the IBM Applied Data Science Capstone course on Coursera Platform.

For reference, the link to the course is: IBM Applied Data Science Capstone

Thank you!