IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kaushik Chakram

19/06/2025

Outline

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

Executive Summary

(Methodology) Data Collection using

- SpaceX API
- Wikipedia (via Beautiful Soup)

(Methodology) Exploratory Data Analysis (EDA) via

- SQL and Data Visualization using Matplotlib.
- Data wrangling using feature engineering and data cleaning.

(Methodology) Interactive Data Analysis

- Built a Dashboard using plotly and Dash.
- Geospatial Analysis using folium to get insights about SpaceX launch sites.

Results

- Trained different ML Classification models to predict the outcome of a SpaceX Falcon 9 launch.
- Decision Tree Classifier had highest accuracy 87% (after hypertuning parameters) compared to the other models.

Introduction

- **Project Specification**: Develop a ML Classification Model to predict if SpaceX's Falcon 9 will land successfully or not.
- **Project Goal:** Being able to successfully predict the outcome implies we can estimate the cost of a launch.
- **Use Case**: ability to potentially bid for or against the launches of SpaceX i.e. a rival company say SpaceY can also be wary of the potential costs /success rates for a similar private rocket company.
- **Insights**: What are those essential features in the data that we can use to predict the launch outcome successfully.
- **Model of Choice**: We compare 4 different ML models and use parameter hypertuning to select a model that best predicts the success rate of the launches.

Section 1

Methodology

Methodology

Executive Summary

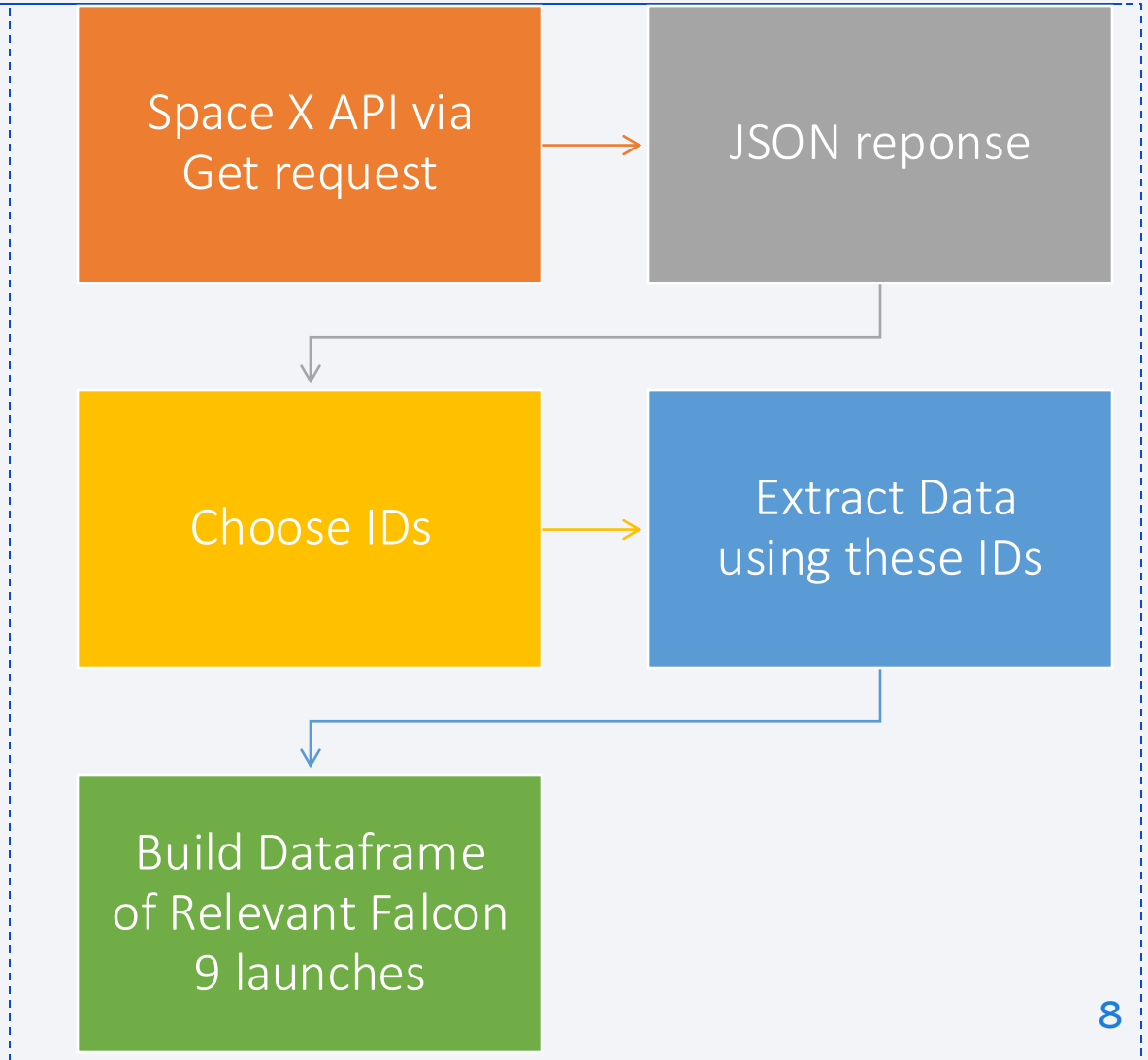
- Data collection methodology:
  - Data collected using SpaceX API and web scraped Wiki using Beautiful Soup.
  - Focused on Falcon 9 launches and stored data as a Dataframe.
- Perform data wrangling
  - Here we cleaned the data by treating numerical and categorical data separately.
  - Treated missing values and recast the data such that it can be used in a ML model.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models.

Data Collection

- Data was collected using two methods:
  - SpaceX API: Using API Request and JSON response
    - Using Get requests
    - Turn JSON response into pandas dataframe via pd.json_normalize()
- Web Scraped: a Wikipedia page using Beautiful Soup HTML page:
  - Use Beautiful Soup find_all method to get remaining data about Falcon 9 rocket from a Wikipedia page.

Data Collection – SpaceX API

- Use SpaceX API to extract some attributes and records of the Falcon 9 data
- Use some helper functions to extract the useful information from the API based on their IDs.
- Names of IDs used to extract data via helper functions: BoosterVersion, LaunchSite, PayloadData, and CoreData.
- Link to code here

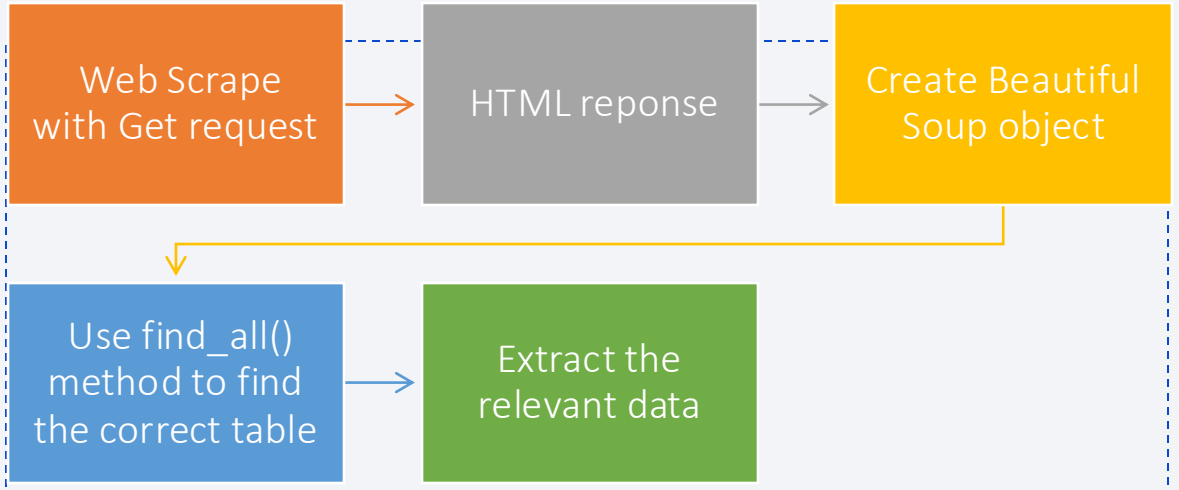


Data Collection - Scraping

- Web Scraped this Wikipedia page

Containing information of Falcon 9 rocket launches.

- Used Beautiful Soup's HTML parser to parse the data
- Use the find_all() method to find the appropriate table with the relevant information.
- Extract the columns with the data.
- Link to code here

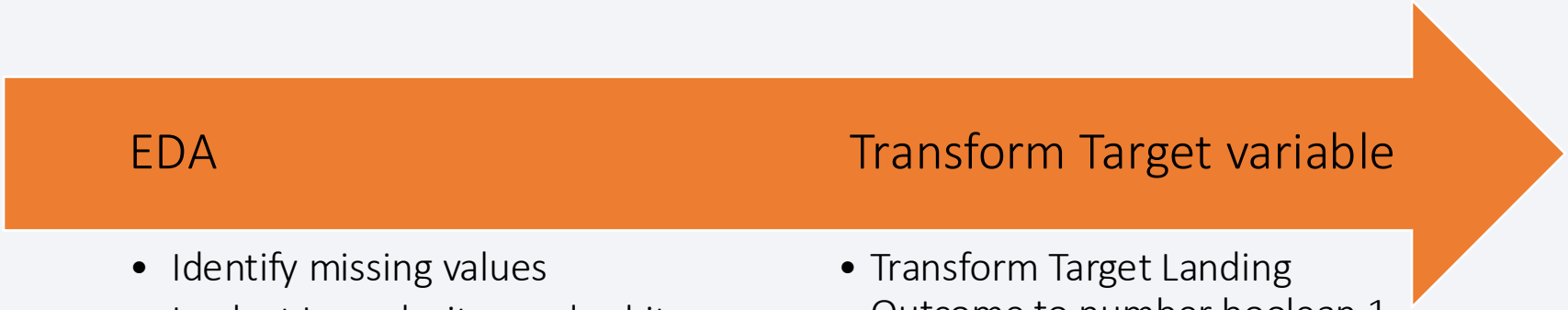


Columns we Extract:
Flight No., Launch Site, Payload, Payload mass, Orbit, Customer, Launch Outcome, Version Booster, Booster landing, Date, and Time.

Data Wrangling

- Start by exploring the data for missing values and identify the data types present.
- Explore the launch sites used and the orbit types of the satellites launched using value_counts() method.
- We found that most rockets are launched from CCAFS SLC 40 site (55 launches) and the most popular orbit type was geosynchronous orbit at 35,786 KM (GTO orbit.)
- Setup data for ML analysis (feature engineering) on the outcome column to recast into (numerical) boolean i.e 1 or 0 into a new column called class.

- Link to code here



EDA with Data Visualization

- Here we want to investigate what are the features that are well correlated with the target variable landing outcome or Class column (i.e. success rate).

- Plotted the following (grouped by class or Success rate):
  - Flight number vs. Launch SIte (Catplot)
  - Payload Mass (kg) vs. launch site (scatter plot)
  - Flight number vs. Orbit type (scatter plot)
  - Payload Mass (kg) vs. Orbit Type (scatter plot)

- Also plotted:
  - Orbit type vs. Success rate (bar plot)
  - Year vs. Success Rate (Line plot)

Link to code here

EDA with SQL

- The following SQL queries were used to get the following information:

1. Names of unique launch sites.
2. Five instances of launch sites that contain the string 'CCA'.
3. Total Payload Mass carried by NASA (CRS) Boosters.
4. Average Payload mass carried by D9 v1.1 booster.
5. The date when the first landing outcome on the ground pad was achieved successfully.
6. Booster names that had success in drone ship and contains a payload mass between 4000 and 6000 Kg.
7. Total number of missions outcomes that were successful and failures.
8. Booster version names that have carried the Maximum amount of payload mass.
9. Records containing month names, failed landing outcomes in a drone ship, booster versions, launch site in year 2015.
10. Count of landing outcomes ranked between the dates of 2010-Jun-04 and 2017-Mar-20, in descending order.

Link to Code here

Build an Interactive Map with Folium

- Mainly three tasks were performed for geospatial analysis using folium.
- Created a map object centered (and marked a circle) at NASA Johnson Space Center (NJSC) at Houston, TX. Also, marked a circle at all launch sites in the data. In addition, all sites contained a label marker to display the site's name.
- Used a Marker cluster object to mark the success/failure of each launch site on the map indicated by the color red: failure and green: success.
- Calculated the distance from the launch sites to other locations such as: railways, highways, coastline, and cities using a function calculate_distance() and mouse position to get the coordinates of locations.
- Lastly, use the calculated distances to draw a line between the coordinates to show on the map. This is done using the PolyLine object.
- Link to code here.

Build a Dashboard with Plotly Dash

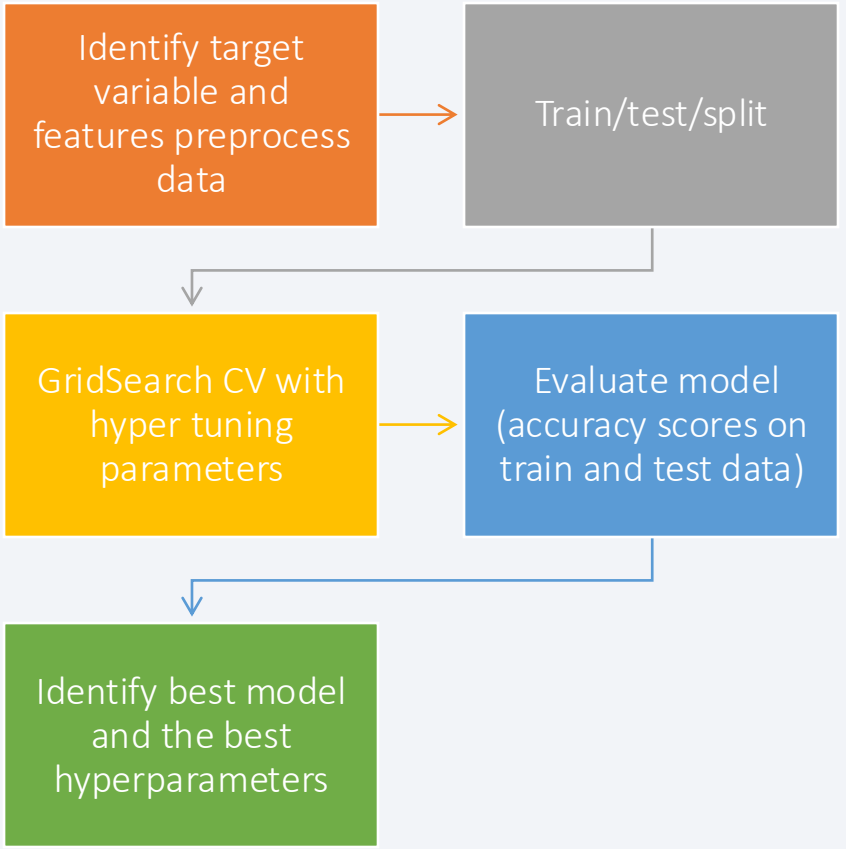
- Built a dashboard for an interactive way to visualize the data.
- These specific interactive plots are study the best/worst: launch site, booster version and the correlation between payload mass and launch outcome.
- Interactions are brought in via a Dropdown menu where we can choose to look at all the launch sites or each one separately. Once the choice is made it displays:
  - Pie chart of the total share of successful landings for each launch site.
  - Pie chart that displays the success or failure rate of that particular launch site.
- Another interaction is brought via a Range Slider for the payload mass (kg).
- This is used to show the scatter plot of payload mass vs. Success/failure for each booster versions. Payload mass can be controlled using the slider.
- Link to code here.

Predictive Analysis (Classification)

- Identify target variable class column i.e. success/failure
- Preprocess i.e. transform features appropriately.
- Train (80%) test (20%) split data.
- Setup GridSearch CV (with 5 folds) and hyper tuning parameters to fit for four ML Classification models:
  - Logistic regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K- Nearest Neighbors (KNN)

- Evaluate the model using accuracy scores on train data then test data along with the confusion matrix.

- Use this to identify the Best model along with the best hyper parameter choices for that model.

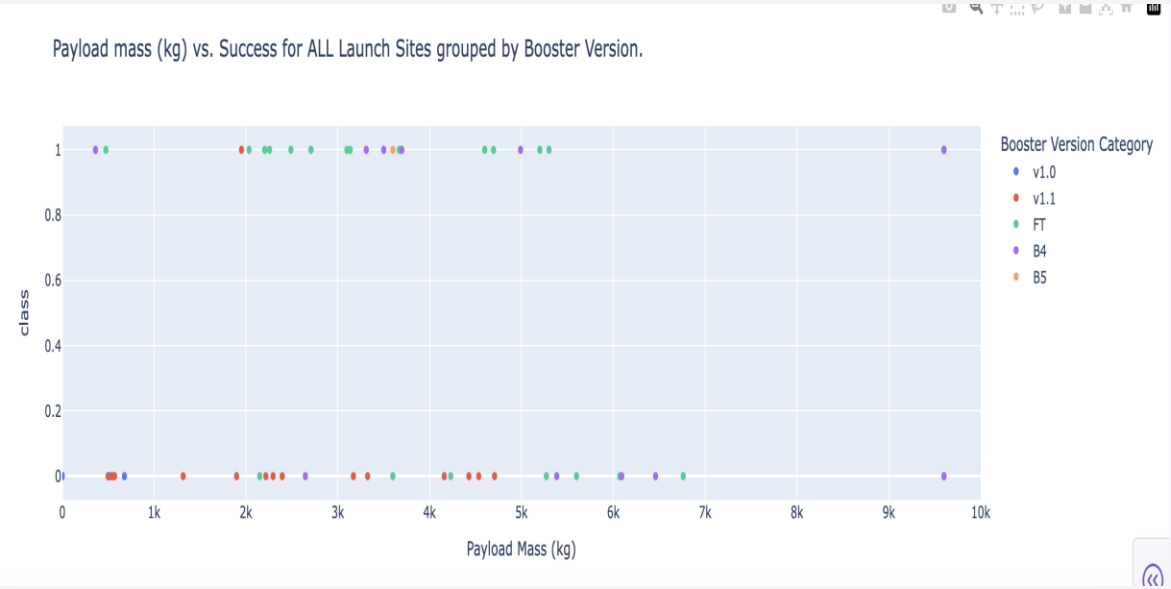


Link to code here.

Results

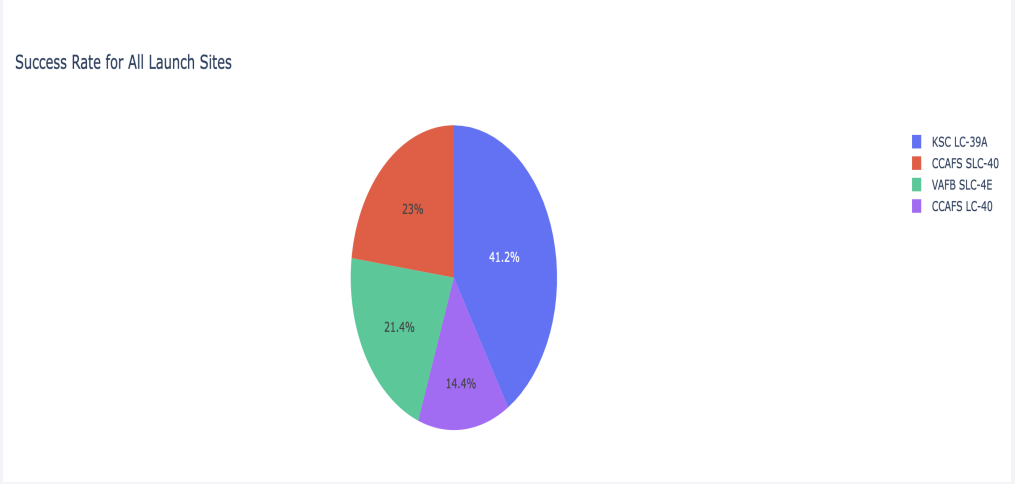
EDA via SQL:

- Found that Total Payload mass taken by NASA (CRS) boosters: 48213 kg.
- Found that avg. Payload mass taken by F9 v 1.1 boosters: 2534.66 kg.
- 22-Dec-2015 was the first date when a landing outcome on a ground pad was achieved.
- Total of 61 successful and 10 failure mission counts.
- Between the dates 04-Jun-2010 and 20-Mar-2017 most landing outcomes was via a drone ship, while least were using a parachute.



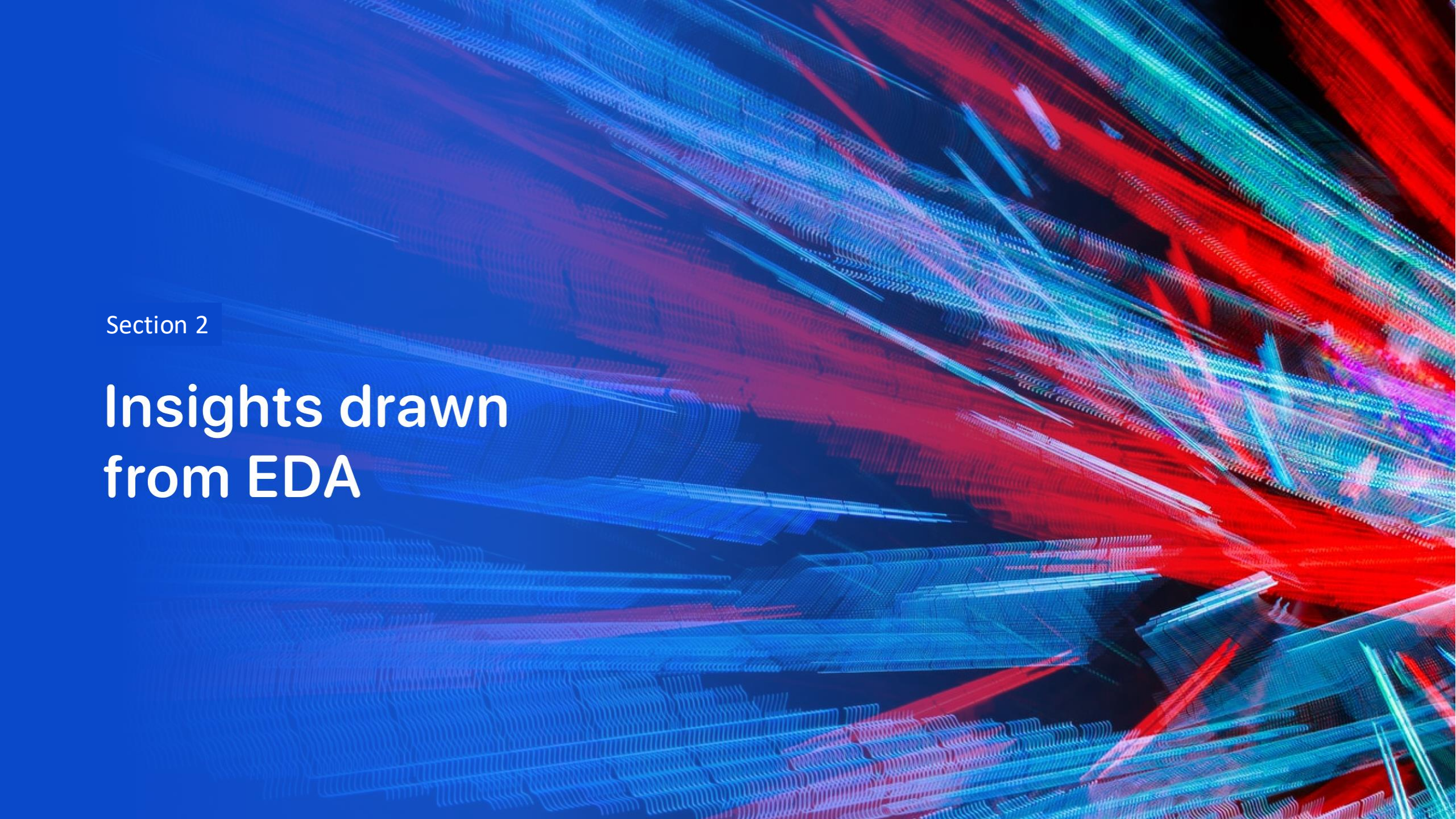
EDA via Data visualization:

- Success rates for launches are different for each launch sites.
- KSC LC-39A is most successful and CCAFS SLC-40 is least successful launch site.
- Across launch sites if payload mass was more than 8000 kg, then those tend to be successful.
- Top 4 successful orbit types are: ES-L1, GEO, HEO, SSO.
- VLEO orbit has more success with more flight numbers whereas, it is much more difficult to tell for GTO orbit type.
- GTO orbit type has both success and failures with a payload mass in the range of 2000 kg –8000 kg.
- Success rate for launches have sharply increased since 2013.



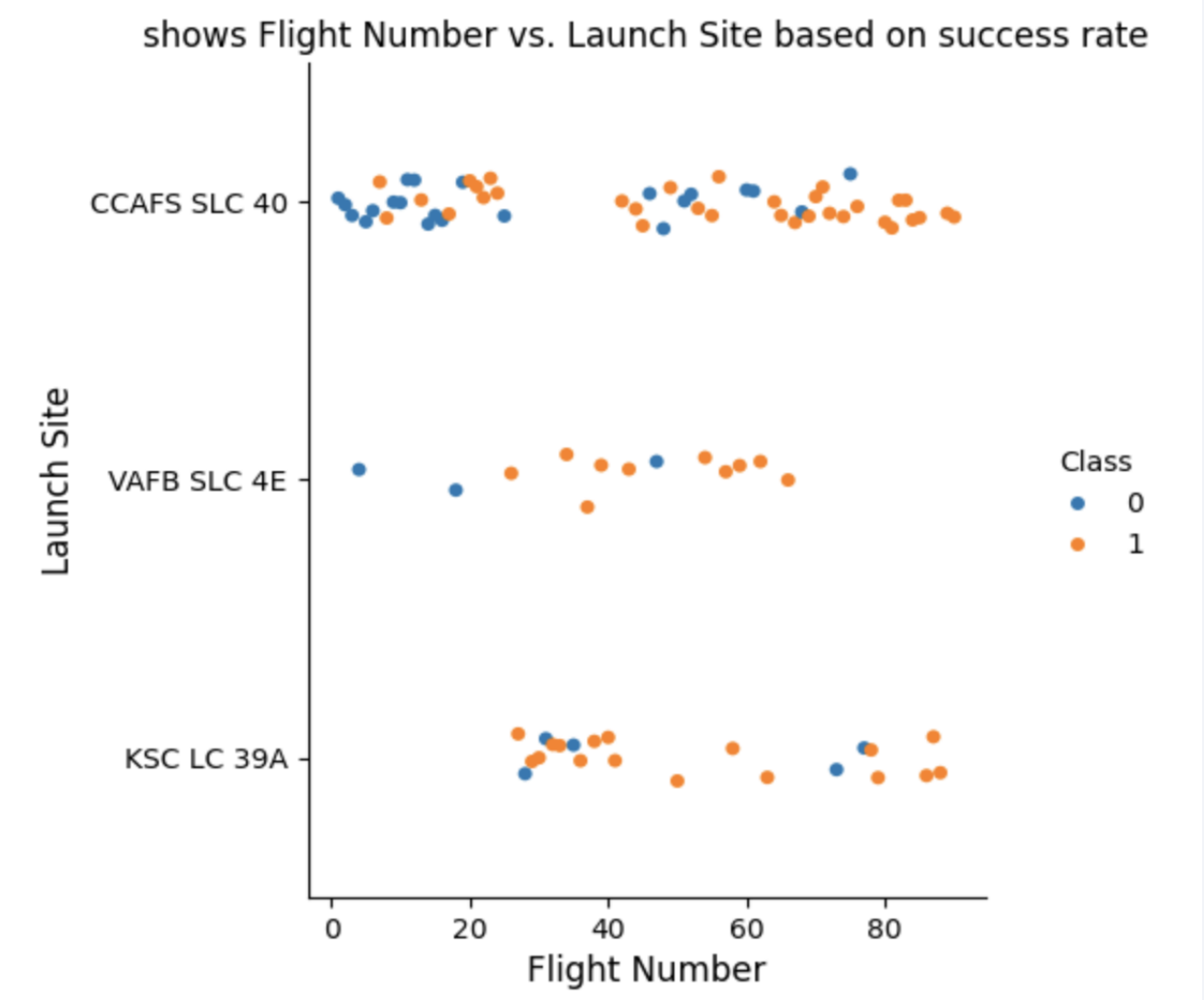
Results (contd.)

- The predictive analysis portion yielded the following results:
  - The best ML classification model out of those tested was a Decision Tree model with the following hyperparameters:
    - Criterion: 'Entropy' max_features: 'sqrt'
    - Max_depth: 2 min_samples_split: 2
    - Splitter: 'random' min_samples_leaf: 4
    - Out of all the other models tested Decision Tree had the most accuracy of 0.875 (or 87.5%).


Section 2
Insights drawn from EDA

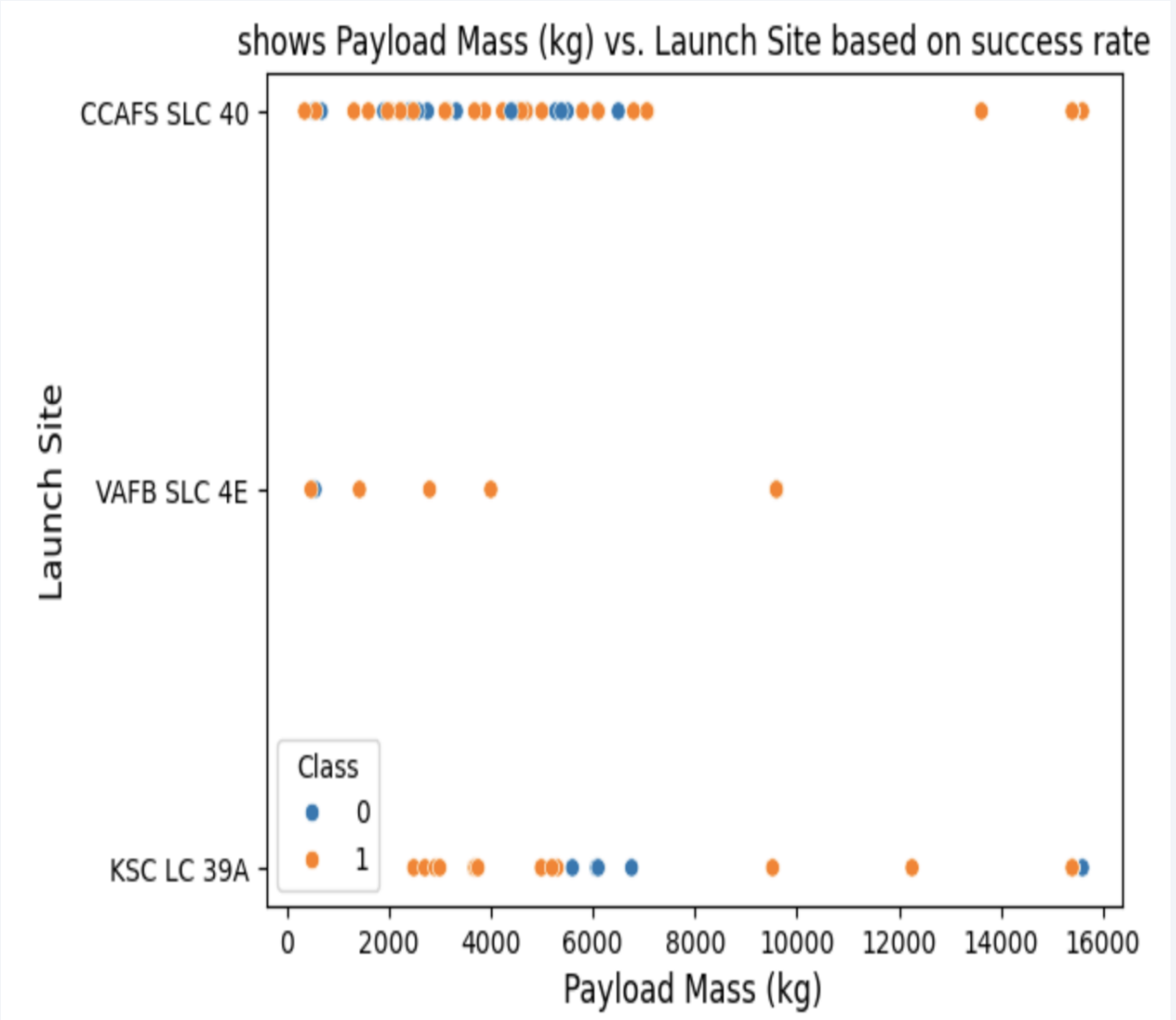
Flight Number vs. Launch Site

- Most successful landings are in KSC LC 39A with more flights launched than VAFB SLC 4E site.
- VAFB SLC 4E has highly successful landings, but with least number of flights launched.
- CCAFS SLC 40 has launched lots of flights, and the success has grown when the number of flights are high.



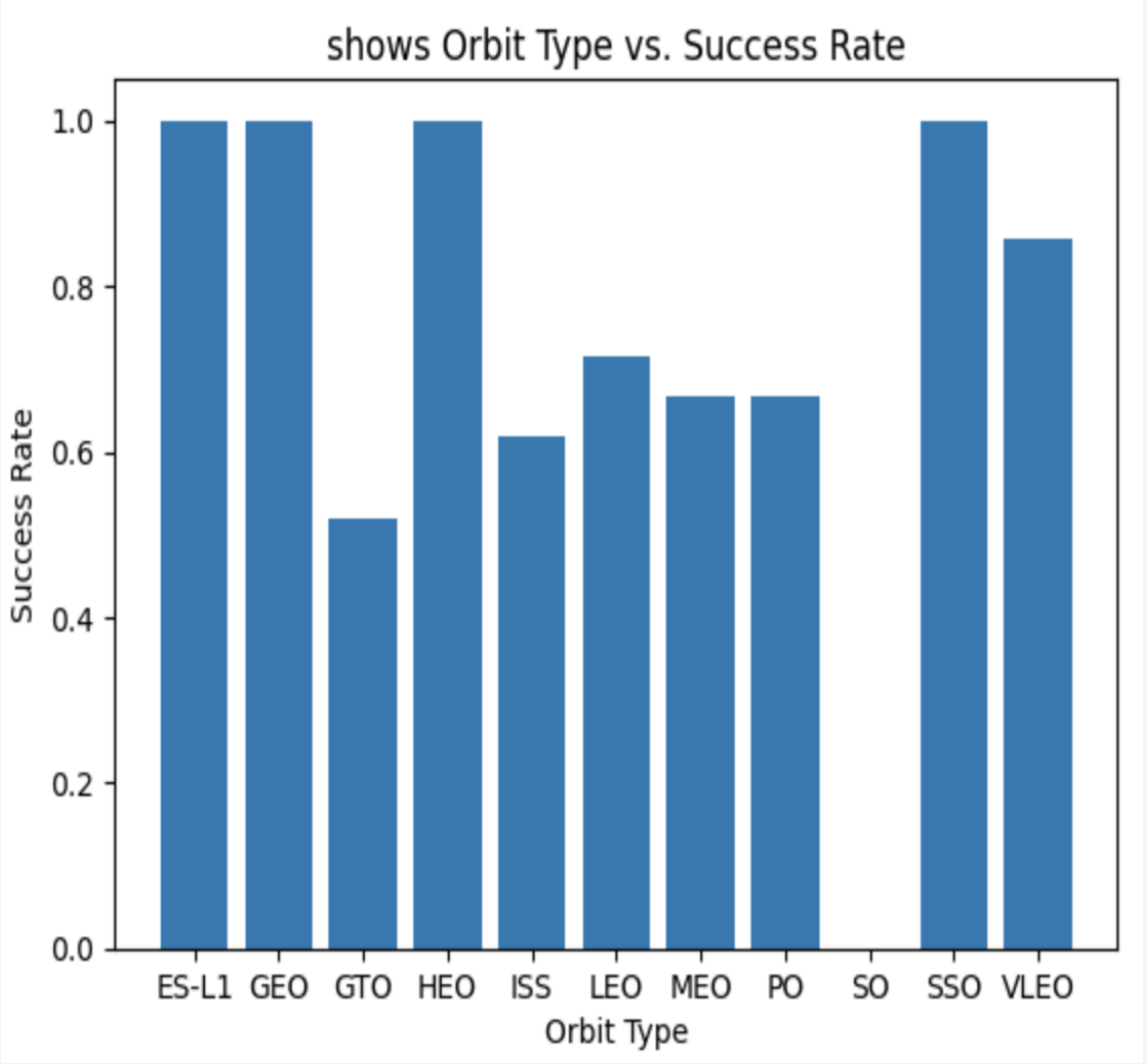
Payload vs. Launch Site

- VAFB SLC 4E site has more success as the payload mass increases, but again least number of flights.
- Most Flights of CCAFS SLC 40 has a payload mass in the range between 0 kg to 8000 kg and it has launched most number of flights.
- KSC LC 39A mostly has successful missions with payload less than 6000 kg, but cannot state an exact trend.



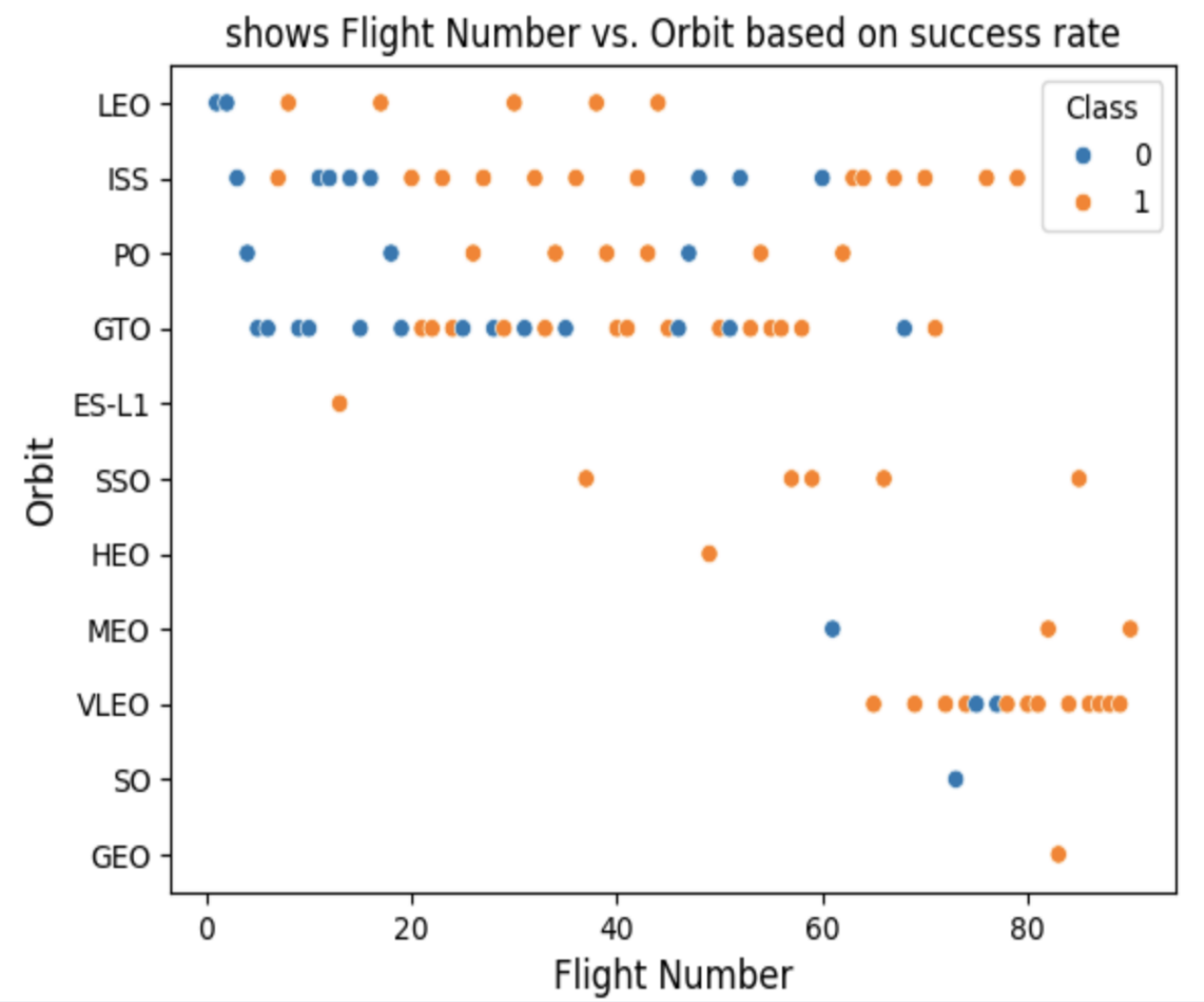
Success Rate vs. Orbit Type

- Top four orbit types with high success rate:
  - ES-L1, GEO, HEO, SSO.
- SO orbit type is the least successful.
- Caveat: ES-L1,GEO, SO, and HEO contains only one data point.
- This will be very apparent in the plot on the next slide.



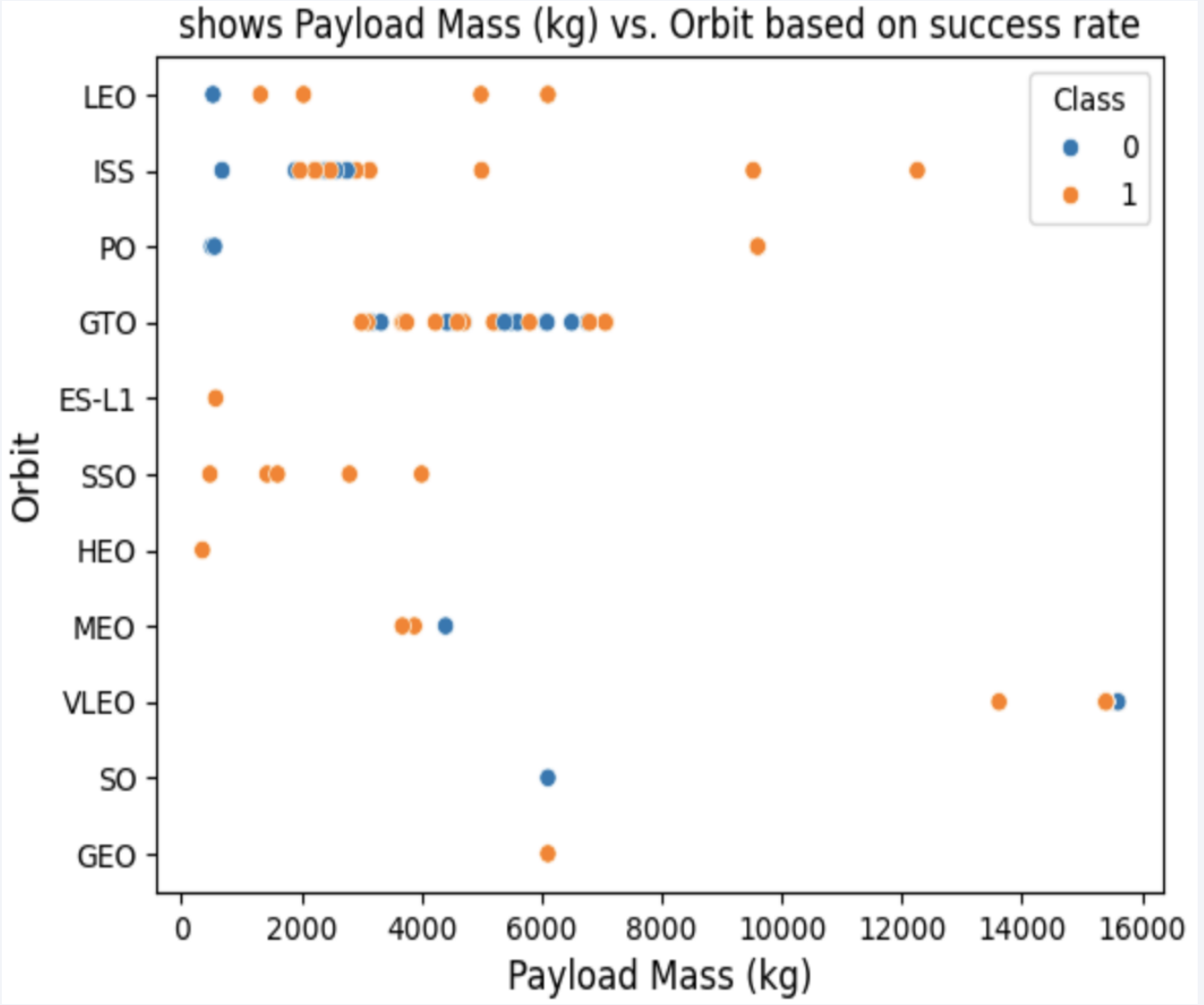
Flight Number vs. Orbit Type

- GTO contains the most number of missions and still has no real trend with the number of flight launches.
- LEO, SSO, VLEO have more successes mostly when the flight numbers are more than 10.
- Notice that the caveat mentioned in the previous slide can be seen here, because ES-L1,SO, HEO, and GEO.
- ISS orbit has the most number of well distributed flight numbers with mix of both success and failures.



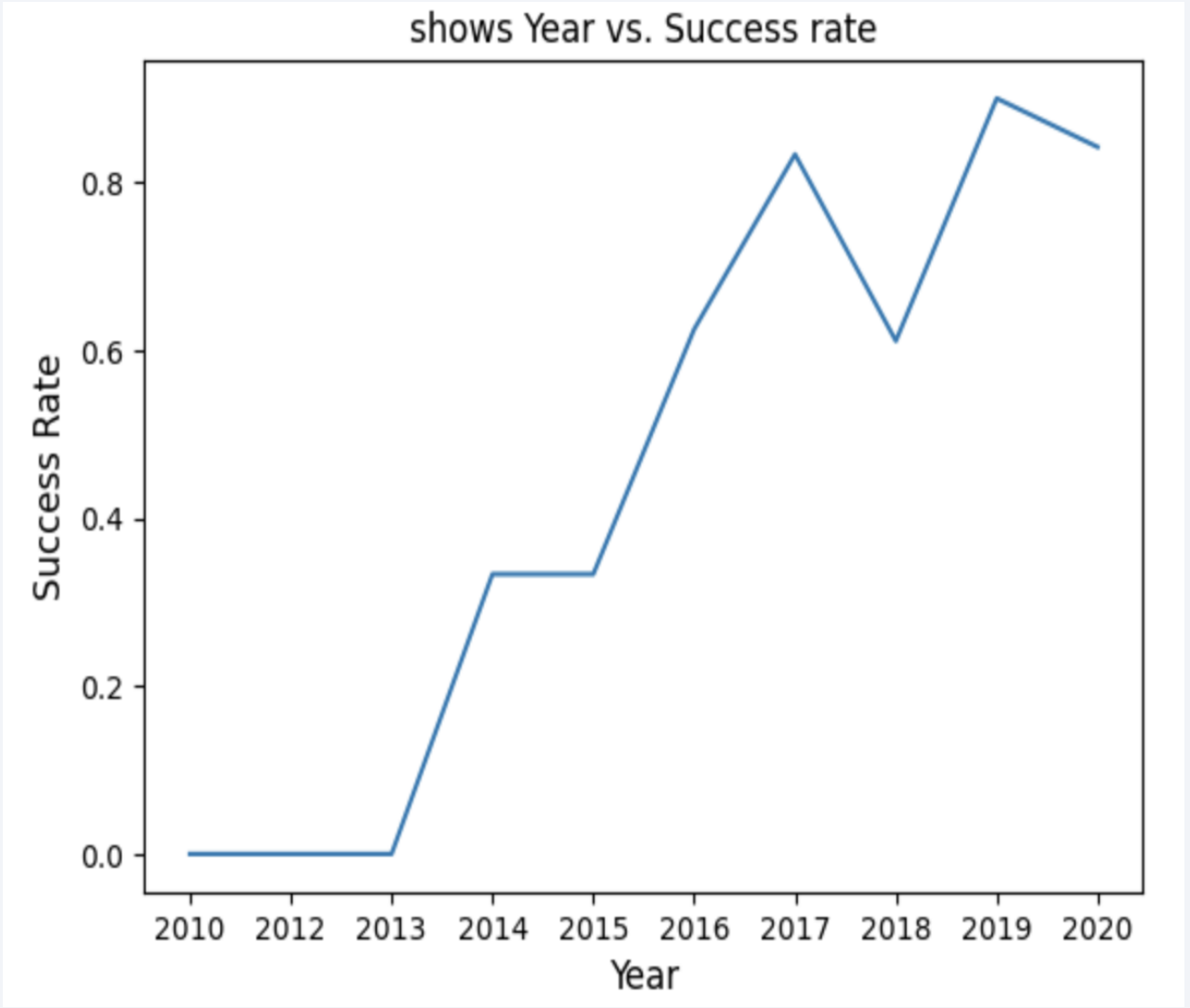
Payload vs. Orbit Type

- GTO orbit type contains no real trend, because it has both success/failures for its payload mass range.
- For LEO, ISS, and PO if the payload mass is high then the mission is also successful.
- VLEO is the only orbit type where the payload mass is the highest with more than 1200 kg.



Launch Success Yearly Trend

- 2018 is the only exception where the success rate decreases.
- News online (link) suggests that in 2018 this occurred during the Zuma mission where there was a failure with the payload adapter.
- However, there is an overall sharp increase in the success rate over post 2013.



All Launch Site Names

- The SQL query used for this is as follows:
  - `%sql select distinct "Launch_Site" from SPACEXTABLE;`

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

- There are four distinct launch sites in the dataset.

Launch Site Names Begin with 'CCA'

- Query to get 5 records where launch sites begin with `CCA`:

```
%sql select * from SPACEXTABLE where Lower("Launch_Site") like
'%cca%' limit 5;
```

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

Total Payload Mass

- The SQL query used for this is as follows:

```
%%sql select SUM("PAYLOAD_MASS__KG_") as "total payload mass kg" from SPACEXTABLE where Customer
like '%CRS%';
```

- The result of this query is 48213 kg.

total payload mass kg
48213

Average Payload Mass by F9 v1.1

- The SQL query used for this is as follows:

```
%%sql select AVG("PAYLOAD_MASS__KG_") as "avg_payload_mass_by_f9_v1_1" from SPACEXTABLE
where "Booster_Version" like '%F9 v1.1%';
```

- The result of this query is 2534.66 kg.

avg_payload_mass_by_f9_v1_1
2534.6666666666665

First Successful Ground Landing Date

- The SQL query used for this is as follows:

```
%sql select min(Date) from SPACEXTABLE where "Landing_Outcome" like '%Success (ground pad)%';
```

- The result of this query is: 22- Dec-2015.

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The SQL query used for this is as follows:

```
%%sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome"
like '%Success (drone ship)%'
    and "PAYLOAD_MASS__KG_" between 4000 and 6000;
```

- The result of this query is:
- The result of this query shows successful done ship missions

Where payload mass range is between 4000 and 6000.

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

Total Number of Successful and Failure Mission Outcomes

- The SQL query used for this is as follows:

```
%%sql
/*
This makes more of sense because space x did not have just one mission failure.
You get only one failure if you go by "Mission_Outcome"
*/
select (select count("Landing_Outcome") from SPACEXTABLE where "Landing_Outcome" like
'%Success%')
    as "Success_Count",(select count("Landing_Outcome") from SPACEXTABLE
    where "Landing_Outcome" like 'Failure%') as "Failure_Count";
```

- The result of this query is as follows:

Success_Count	Failure_Count
61	10

Boosters Carried Maximum Payload

- The SQL query used for this is as follows:

```
%%sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (
select max("PAYLOAD_MASS__KG_") from SPACEXTABLE);
```

- The result of this query is shown on the right:
- The result shows all Booster versions where it carries 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

- The SQL query used for this is as follows:

```
%%sql
select substr(Date,6,2) as month, "Landing_Outcome", "Booster_Version",
"Launch_Site"
    from SPACEXTABLE where substr(Date,0,5) = '2015' and "Landing_Outcome"
like '%Failure (drone ship)%';
```

- The result of this query is as follows:
- The months of the failed (drone ship) landings in 2015 are January and April at CCAFSLC-40 site.

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

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

- The SQL query used for this is as follows:

```
%%sql select "Landing_Outcome", count("Landing_Outcome") as "Landing_Count" from SPACEXTABLE where Date
between '2010-06-04' and '2017-03-20' group by "Landing_outcome" order by "Landing_Count" desc;
```

- The result of this query on right:

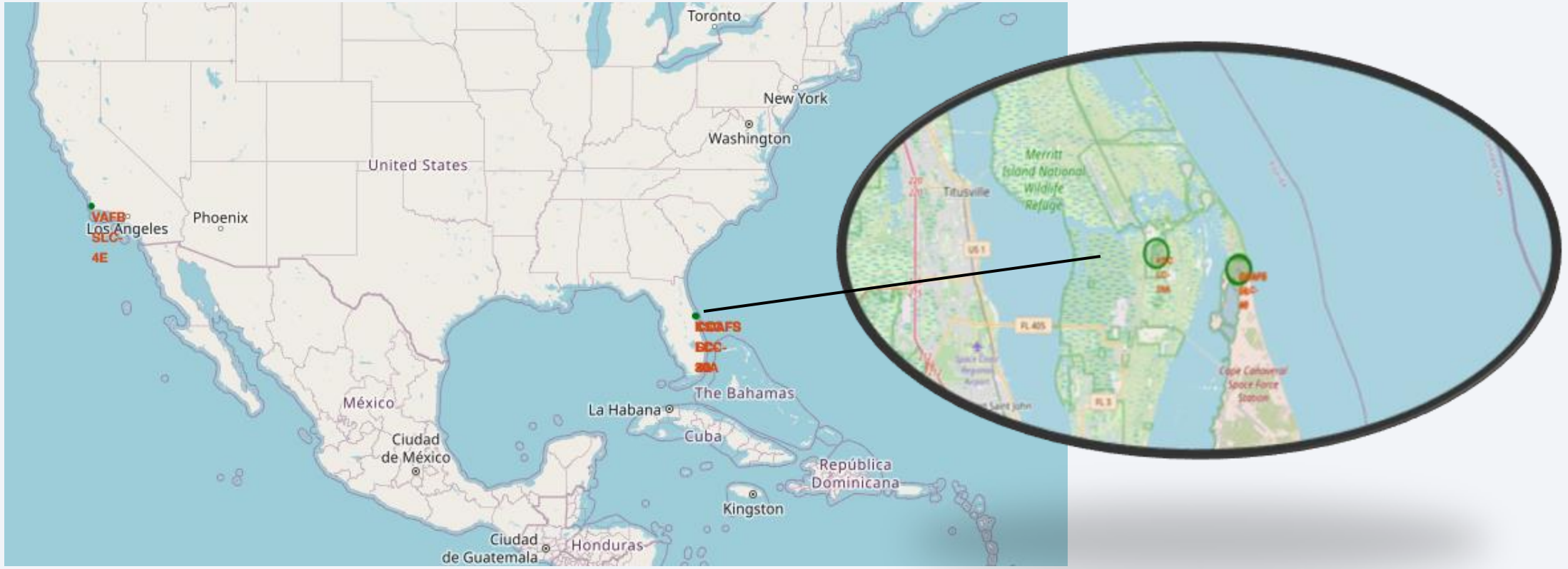
- This query counts the landing outcomes that contain either failed drone ship mission or success in ground pad between this time frame: 2010-06-04-2-17-03-20.
- The result shows that the top outcome has missions where it did not contain an attempt to land
- The second lowest count is for launches that contains failed landings with parchutes.
- The lowest count is for missions with precluded drone ships.

Landing_Outcome	Landing_Count
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

Section 3

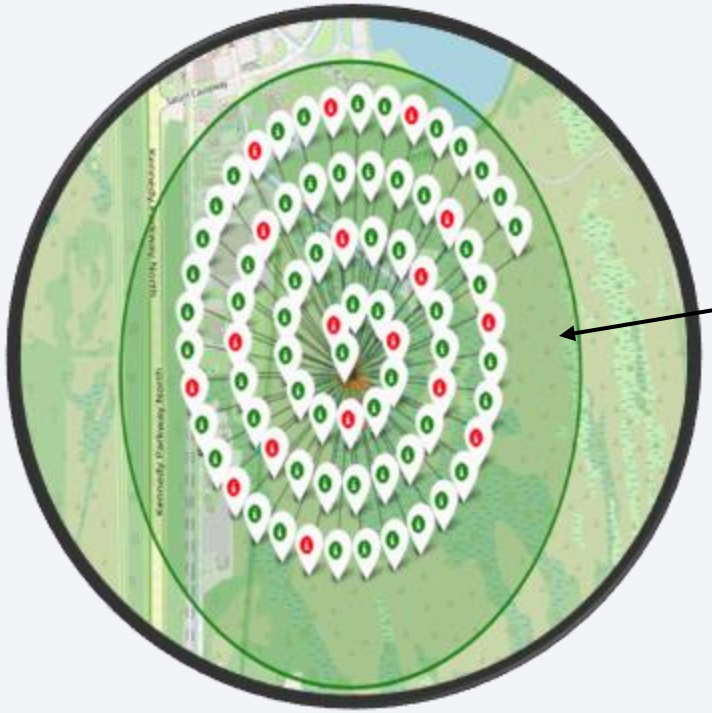
Launch Sites Proximities Analysis

Find Launch Sites on the Map



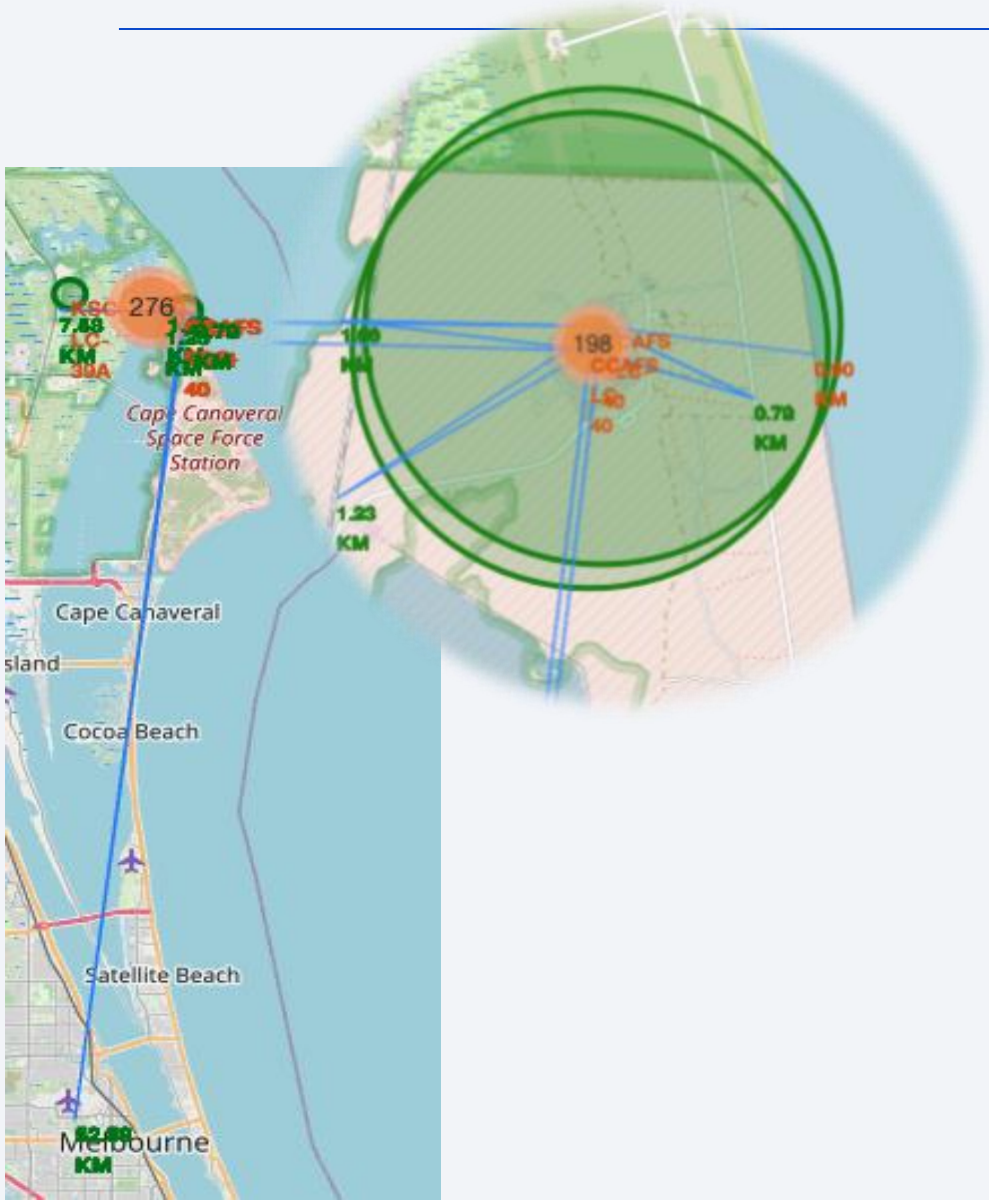
- Added a circle and marked all launch site locations on the map.
- For example the above picture on left shows the broad picture whereas on the right we see a zoomed in version of the launch sites on the eastern side of the united states

Clustering Launch Outcomes

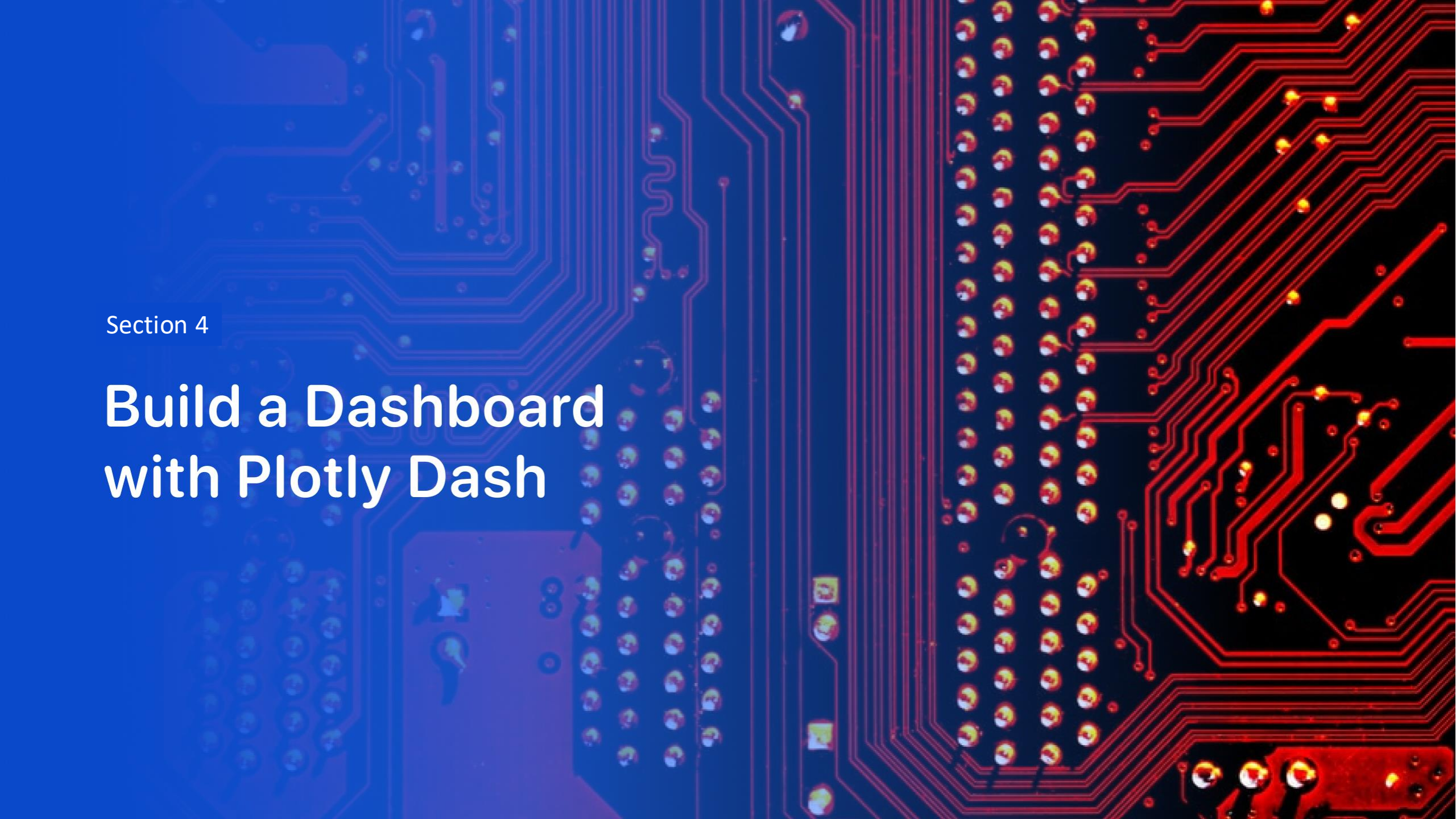


- We clustered based on launch outcome for each launch site and marked them on the map. Success shown in green icon, failure: red icon
- For example KSC-LC-39-A is shown on the right with its success/fail launches marked on the left (zoomed version).

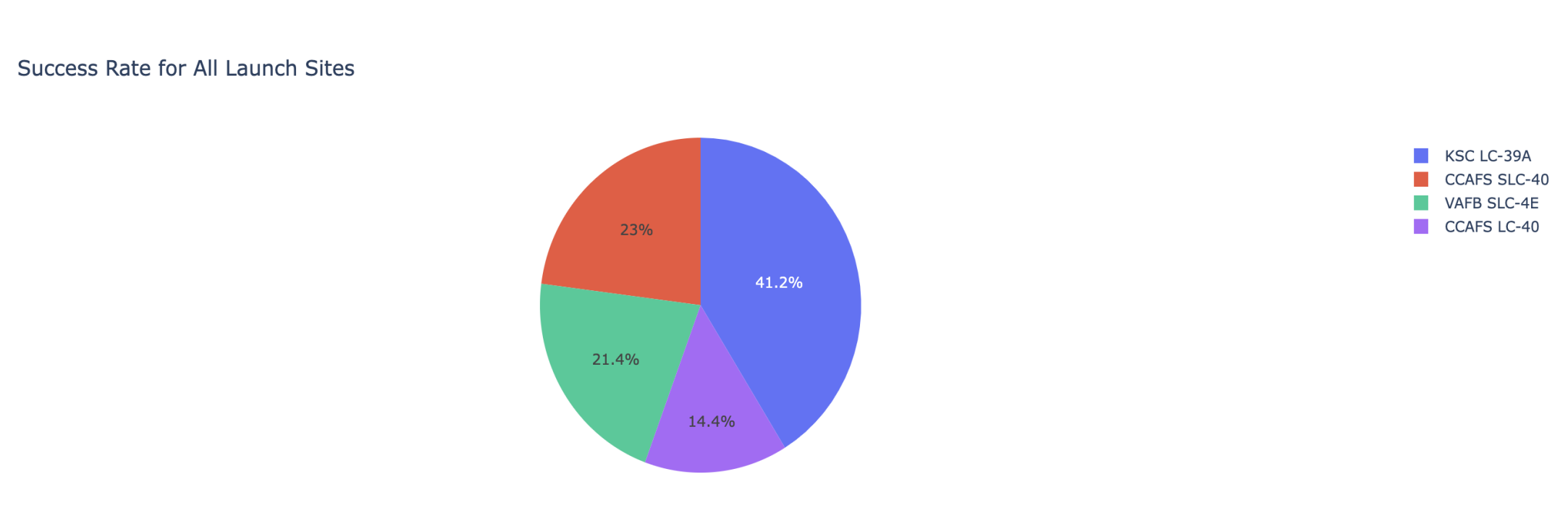
Geospatial Proximity Analysis



- We focus on CCAFS LC-40 and CCAFS SLC-40 on Easter side of US since they are located fairly close to each other.
- We find that it is pretty close to railways about 1.23 km next to a place called Titan III Rd.
- Nearest highway is about 0.70 km called Samuel Phillips Parkway.
- Fairly close to the coastline 1.0 km on the western side and 0.90 km on the eastern side.
- The nearest Big city Melbourne FL,US is pretty far away about 52.99 km.
- Makes sense that the launch sites are away from cities and close to coast in the event of mission failure.

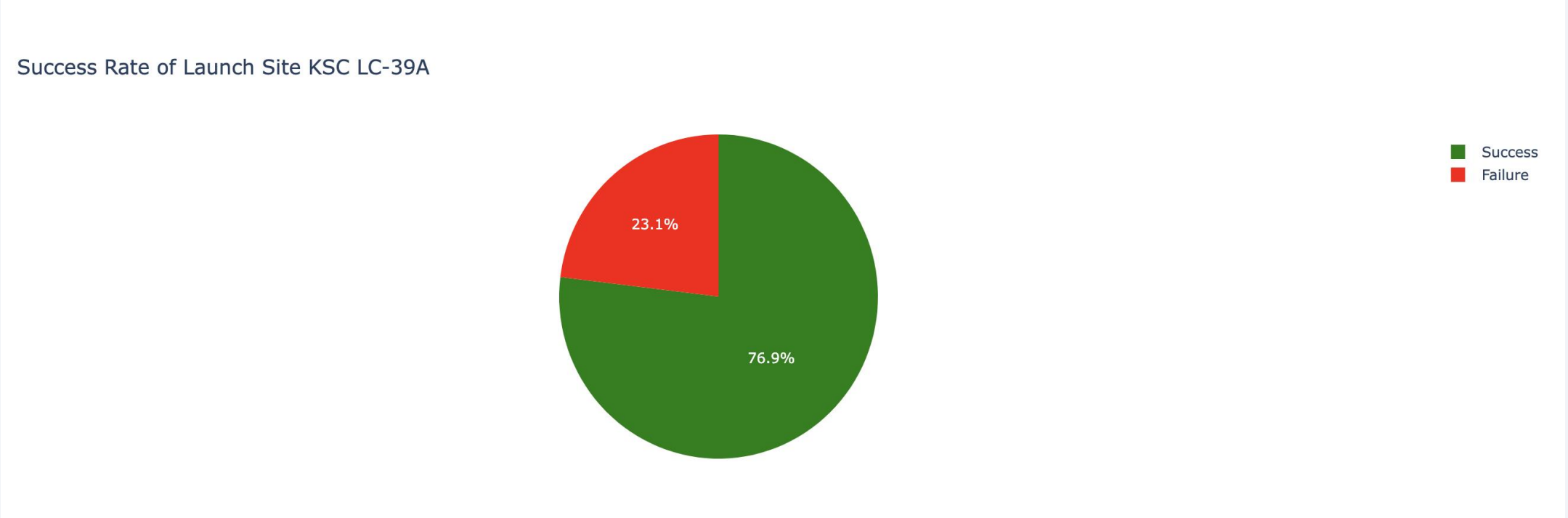

Section 4
Build a Dashboard with Plotly Dash

Dashboard: Total Success Rate Pie Chart



- After selecting "All Sites" from the dropdown menu, we get the pie chart representing the total success rate for each of the four launch sites.
- The result matches what we got in our EDA visualization.
- Most successful site is KSC LC-39A and least successful is CCAFS LC-40.

Dashboard: Most Successful launch site Pie chart



- The most successful site is KSC LC-39A where success is 76.9% and 23.1% failures.

Dashboard: Payload Mass vs. Outcome group by Booster Version

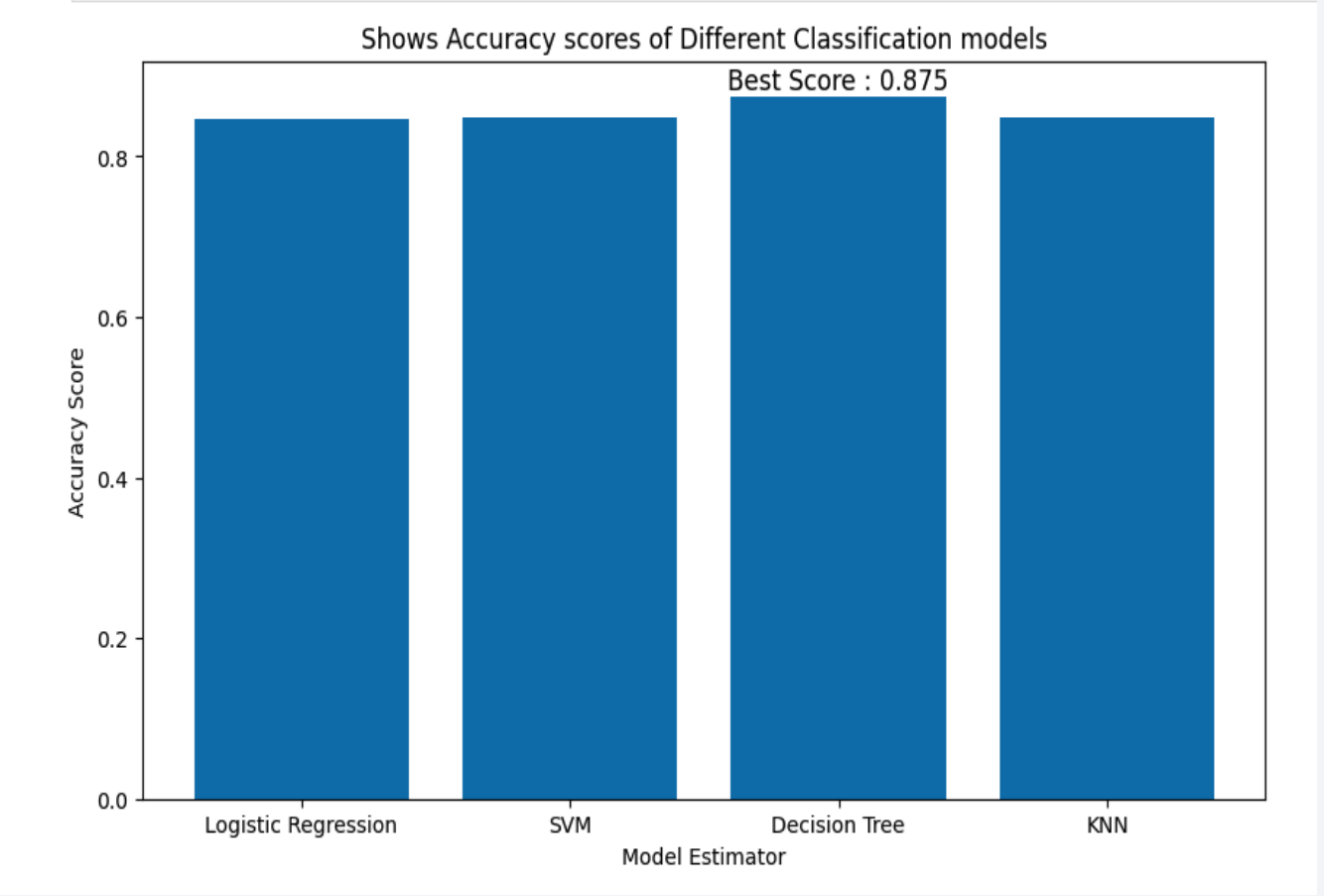
- Using the payload slider we use a meaningful range for this variable and we can see the success for every booster version that was used
- B5 has only one successful mission when payload mass is at 3600 kg.
- Booster versions FT and B4 have bit more successful missions but it does not have a real trend with payload mass
- Version 1.1. has mostly failed missions in this payload mass range.

Section 5

Predictive Analysis (Classification)

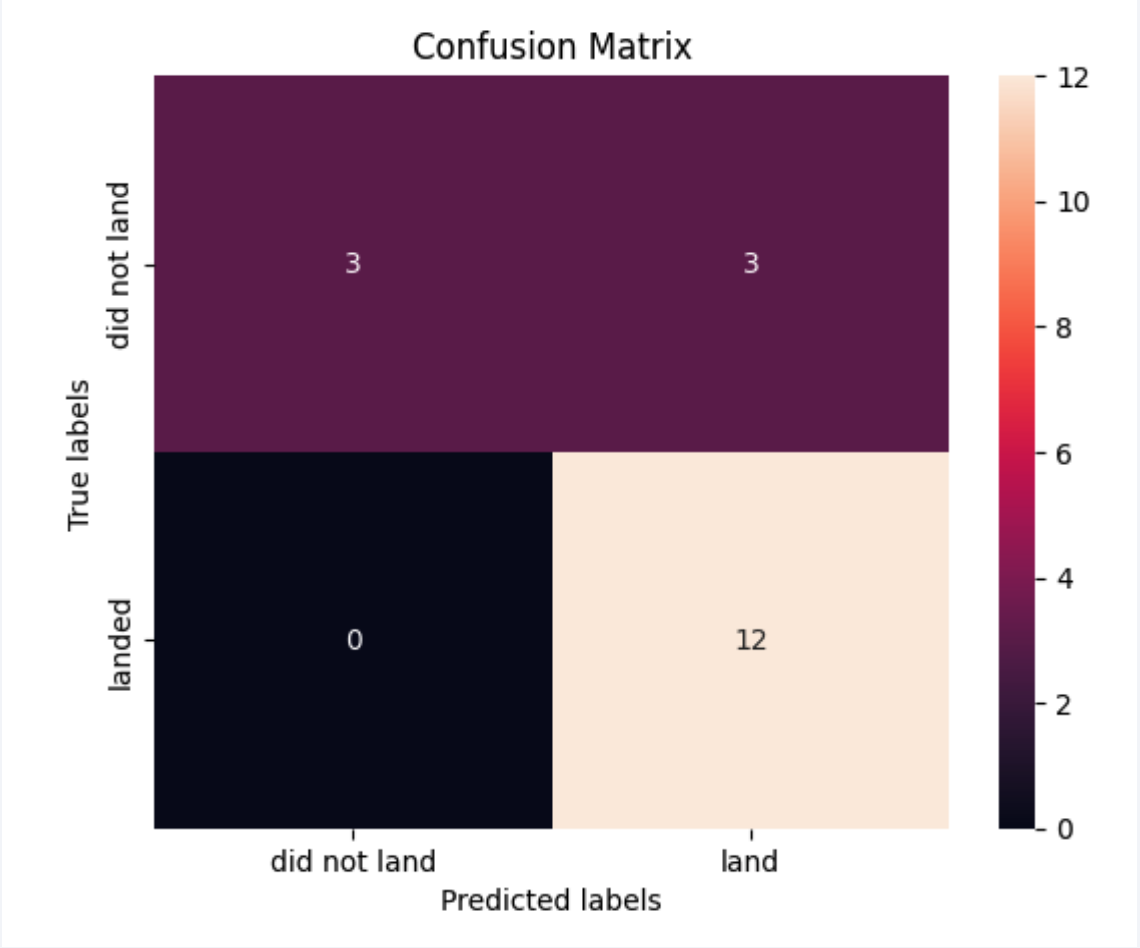
Classification Accuracy

- Compared to all the ML classification models tested:
  - Decision Tree Classifier was the best model with an accuracy score of 0.875.
- The other models have an accuracy score of approx. 0.84.



Confusion Matrix

- Interestingly All four models produced the same confusion matrix.
- The model has trouble with top right entries i.e. the rockets did not land actually, but the model classifies it as though it did land.

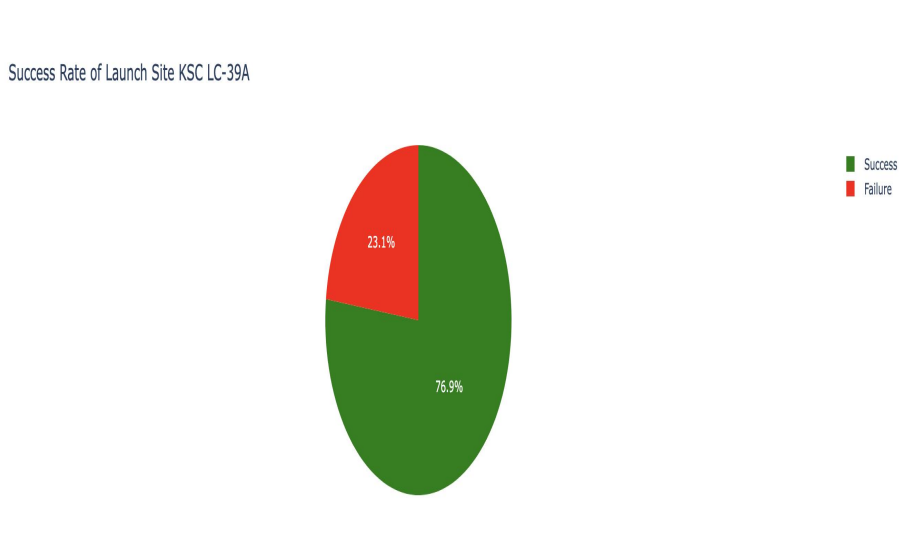


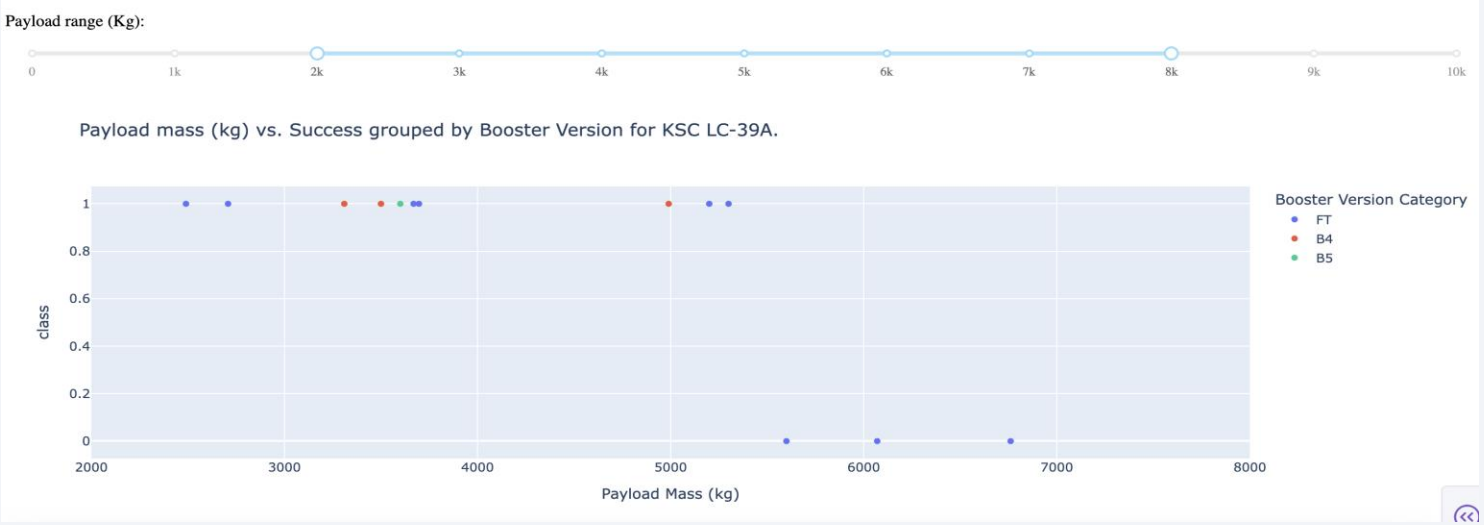
Conclusions

- In this project we investigate SpaceX rocket launch data to predict launch outcome.
- Started with SpaceX API and web scrape with BeautifulSoup to get data.
- Then investigated with EDA data visualization and SQL. Along with interactive and geospatial analysis with plotly dash and folium, respectively.
- Finally we tested four different ML classification models with hyperparameter tuning with gridsearch cv.
- This analysis suggests that Decision tree calssifier was the best model compared to the other models with an accuracy score of 0.875.
- In context this implies that this model can predict the launch outcome (success/fail) of the mission with an accuracy of 87.5%

Appendix

- Confirm KSC LC-39A is successful with another scatter plot





- The plot on the right also confirms the pie chart on the left, because the scatter plot shows that regardless of the booster version for KSC LC -39A site mostly the missions where successful.

Thank you!