Winning Space Race with Data Science

Nanjun Chen
08.28.2023

Outline

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

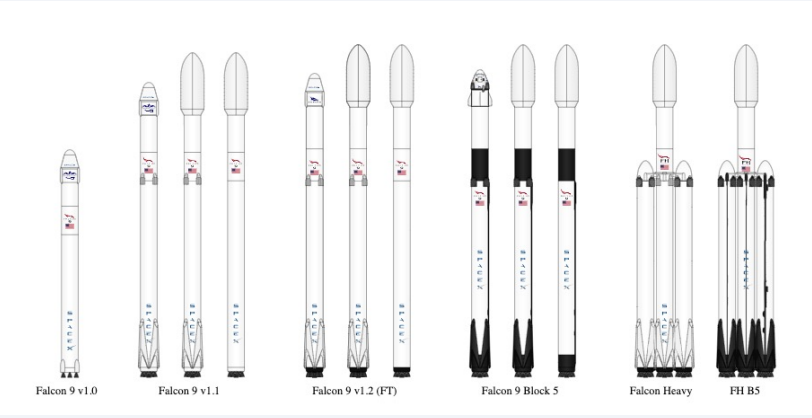
Executive Summary

- Summary of methodologies
  - Data Collection: SpaceX API, Web-scrapping
  - Data Wrangling
  - Exploratory Data Analysis (EDA): visualization, SQL
  - Interactive folium map for launch sites
  - Plotly dashboard
  - Predictive analysis with classification techniques: logistic regression, SVM, decision tree, KNN
- Summary of all results
  - EDA: Visualization of correlation among features of interest (independent variables)
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context
  - 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
  - Savings on the cost is strongly related to whether SpaceX can reuse the first stage. If we can determine if the first stage will land, we can determine the cost of a launch.
  - This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - Will the first stage land successfully?
- Objective
  - Determine the price of each launch;
  - Predict if SpaceX will reuse the first stage - meaning successful landing of first stage.

Source: Universe Today



Section 1

Methodology

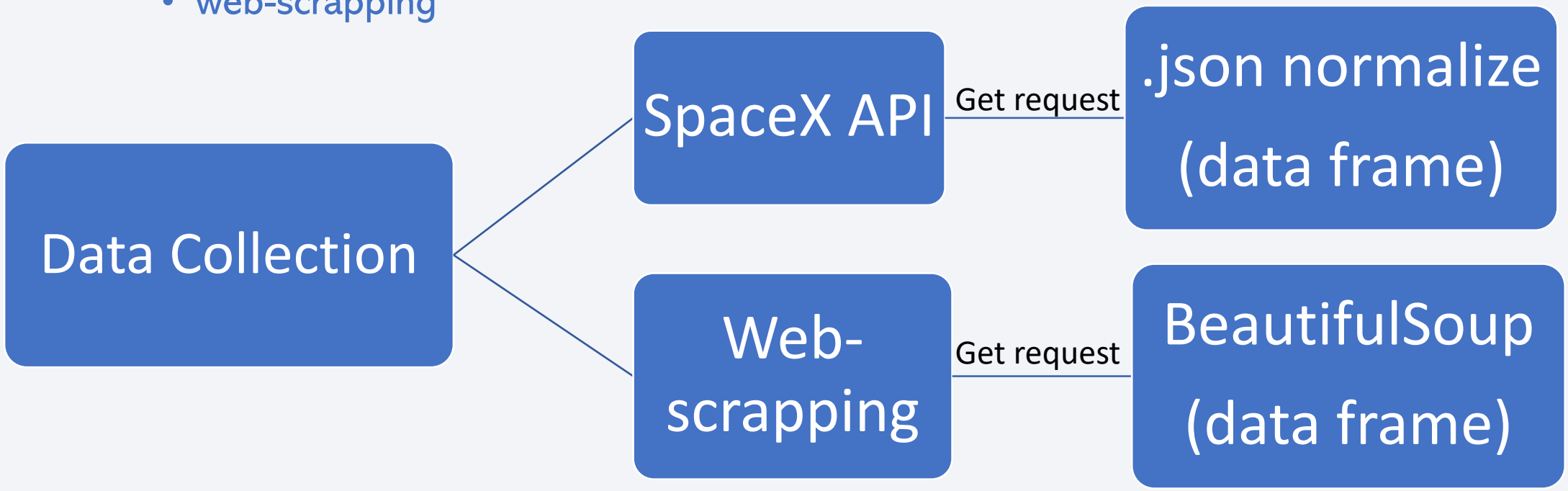
Methodology

Executive Summary

- Data collection methodology:
  - Request to the SpaceX API and convert response content to dataframe
  - Extract Falcon 9 launch record from webpage and parse html tables to dataframe
- Perform data wrangling
  - Clean the requested data, convert it to more readable type, filter it to data of interest (Falcon 9 launches), deal with missing values
  - Find some patterns in the date and determine the label for training supervised models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize data and split into training and testing datasets
  - Tune hyper-parameters for different supervised models with Grid Search method
  - Evaluate the accuracy score and find the model and parameters with best performances

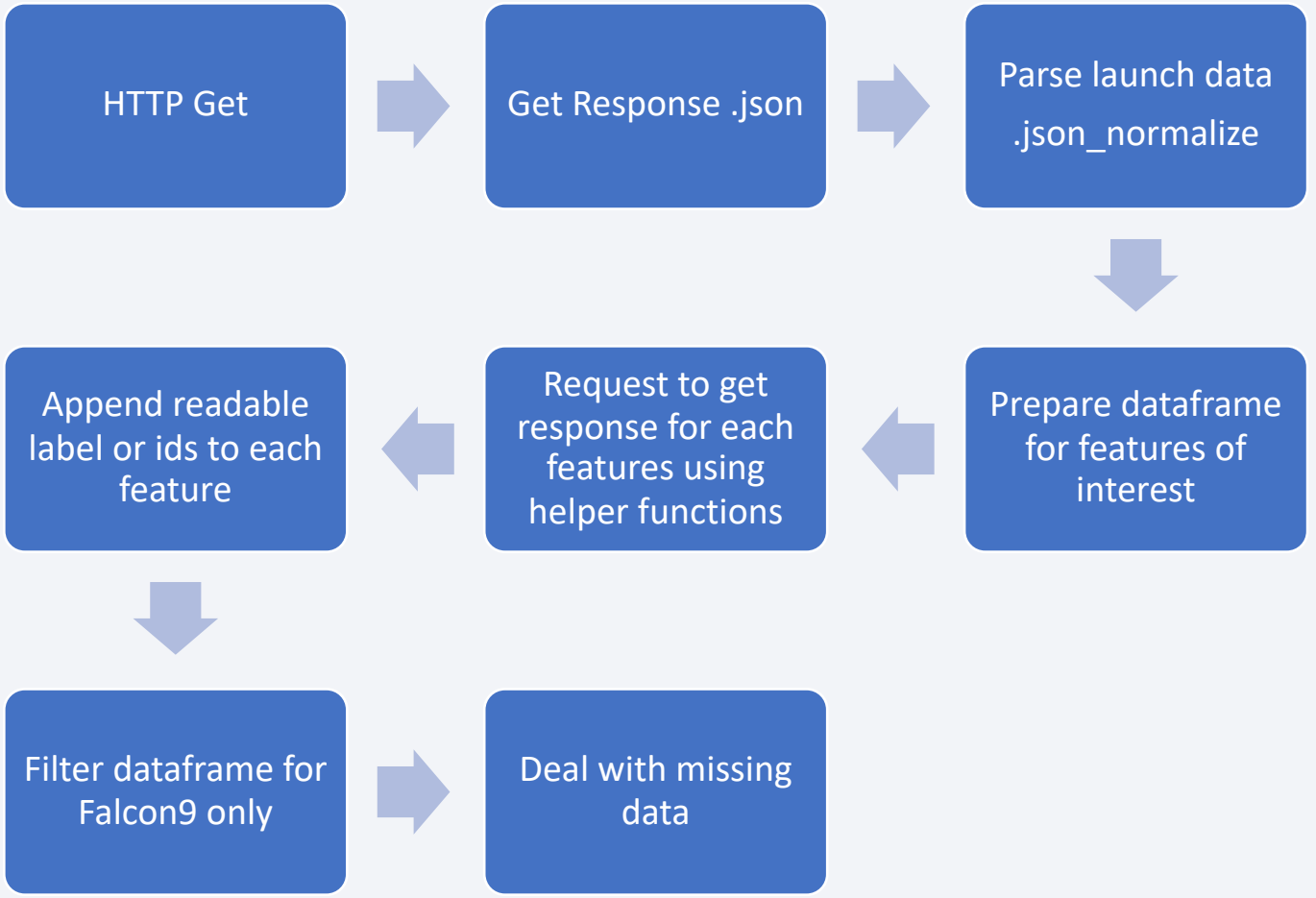
Data Collection

- Datasets were collection through
  - requesting to SpaceX API and
  - web-scrapping



Data Collection – SpaceX API

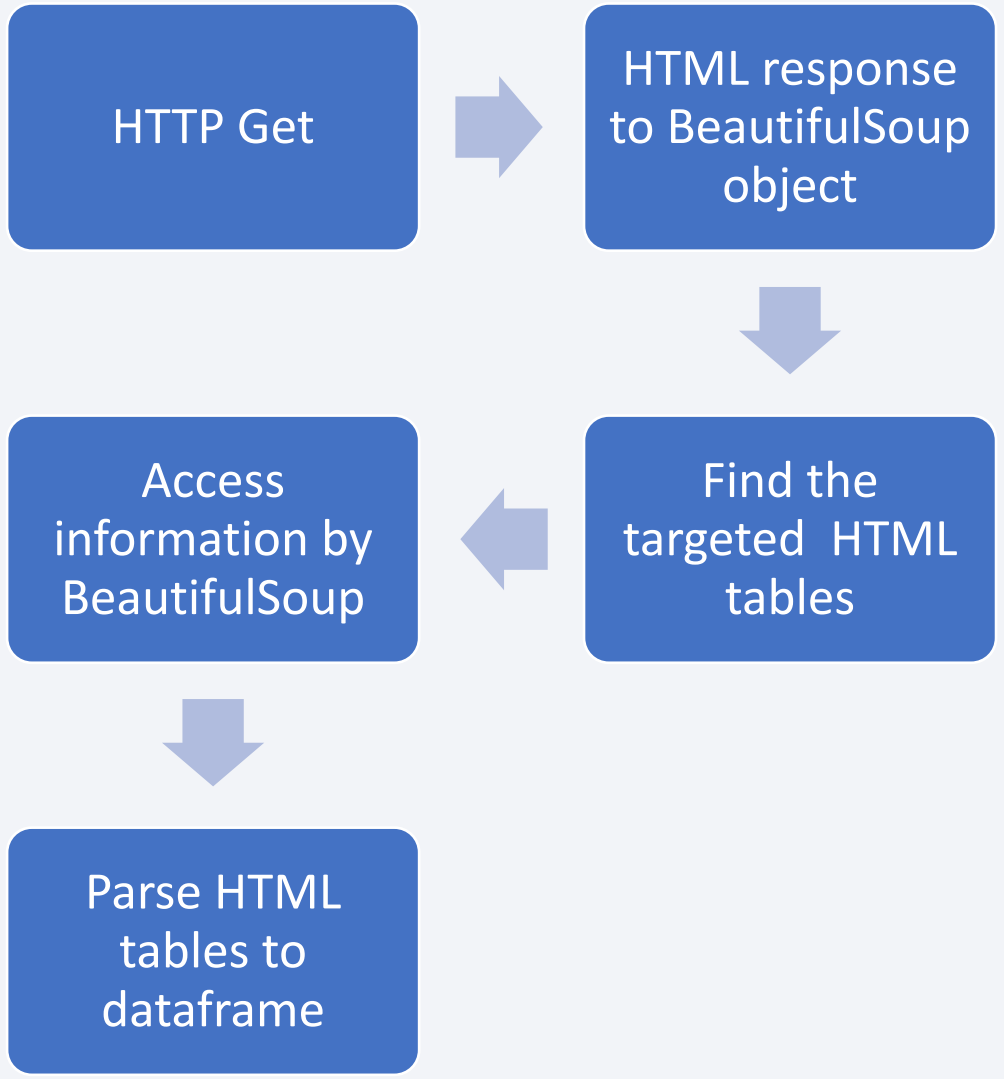
- Request and parse the SpaceX launch data using GET request
- Filter the dataframe to only include Falcon 9 launches
- Minor data wrangling for missing values



GitHub URL: https://github.com/nanjunchen/Rocket-Launch-Data-Science-Project/blob/main/spacex-data-collection-api.ipynb

Data Collection - Scraping

- Request the Falcon9 launch wiki page from its URL
- Extract all columns/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables



GitHub URL:
https://github.com/nanjunchen/Rocket-Launch-Data-Science-Project/blob/main/webscraping.ipynb

Data Wrangling

- Calculate the number of launches on each **launch site**
- Calculate the number of occurrence of each **orbit**
- Calculate the number and occurrence of **mission outcome** of the orbits
- Create a **landing outcome** label from Outcome column and assign to new column

Launch site

LaunchSite	
CCSFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	13

Orbit

Orbit	
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
MEO	3
ES-L1	1
HEO	1
SO	1
GEO	1

Mission outcome

0	True ASDS
1	None None
2	True RTLS
3	False ASDS
4	True Ocean
5	False Ocean
6	None ASDS
7	False RTLS

Landing classification

	Class
0	0
1	0
2	0
3	0
4	0
5	0

GitHub URL: https://github.com/nanjunchen/Rocket-Launch-Data-Science-Project/blob/main/spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plot of Flight Number vs. Payload Mass with overlay of outcome
  - Show flight number and payload mass correlation as well as importance to outcome
- Scatter plot of Flight Number vs. Launch Site with overlay of outcome
  - Understand the correlation of flight number and launch site for successful launch
- Scatter plot of Payload Mass vs. Launch Site with overlay of outcome
  - Show how payload mass is correlated with launch site for successful launch
- Bar plot of landing outcome for different orbit type
  - Demonstrate the relationship between success rate of each orbit type
- Scatter plot of Flight Number vs. Orbit type with overlay of outcome
  - See if there is a relationship between flight number and orbit types for successful launch
- Scatter plot of Payload Mass vs. Orbit type with overlay of outcome
  - See how payload mass and orbit types influence launch outcome
- Line chart of launch success rate as a function of year
  - Show yearly trend of successful launch rate

GitHub URL: https://github.com/nanjunchen/Rocket-Launch-Data-Science-Project/blob/main/eda-dataviz.ipynb

EDA with SQL

- %sql select Launch_Site from SPACEXTABLE group by Launch_Site
- %sql select Launch_Site from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
- %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer =='NASA (CRS)'
- %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version=='F9 v1.1'
- %sql select min(Date) from SPACEXTABLE where Landing_Outcome =='Success (ground pad)'
- %sql select Booster_Version from SPACEXTABLE where Landing_Outcome == 'Success (drone ship)' and PAYLOAD_MASS__KG_>=4000 and PAYLOAD_MASS__KG_<6000
- %sql select COUNT(Mission_Outcome) from SPACEXTABLE
- %sql select Booster_Version from (select max(PAYLOAD_MASS__KG_), Booster_Version from SPACEXTABLE) as Boost
- %%sql select substr(Date,2,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE
  where Landing_Outcome ='Failure (drone ship)' and substr(Date,1,4)='2015'
- %%sql select Date, Landing_Outcome, COUNT(Landing_Outcome) as countL from SPACEXTABLE
  where Date <='2017-03-20' and Date > '2010-06-04'
  group by Landing_Outcome Order by countL DESC

GitHub URL: https://github.com/nanjunchen/Rocket-Launch-Data-Science-Project/blob/main/EDA-with-SQL.ipynb

Build an Interactive Map with Folium

- Circle and marker for NASA Johnson Space Center, TX
  - To initiate the map and label the the location
- Circles and markers for launch sites
  - To show the specific locations of the launch sites
- Marker cluster for launch sites with color differentiated by launch outcome
  - Identify success rate for each launch site
- Marker and lines for distance between launch sites to its proximities, e.g., coastline, railway, highway, etc.
  - Explore and analyze how its proximities affect the launch

GitHub URL: https://github.com/nanjunchen/Rocket-Launch-Data-Science-Project/blob/main/folium_launch_site_location.ipynb

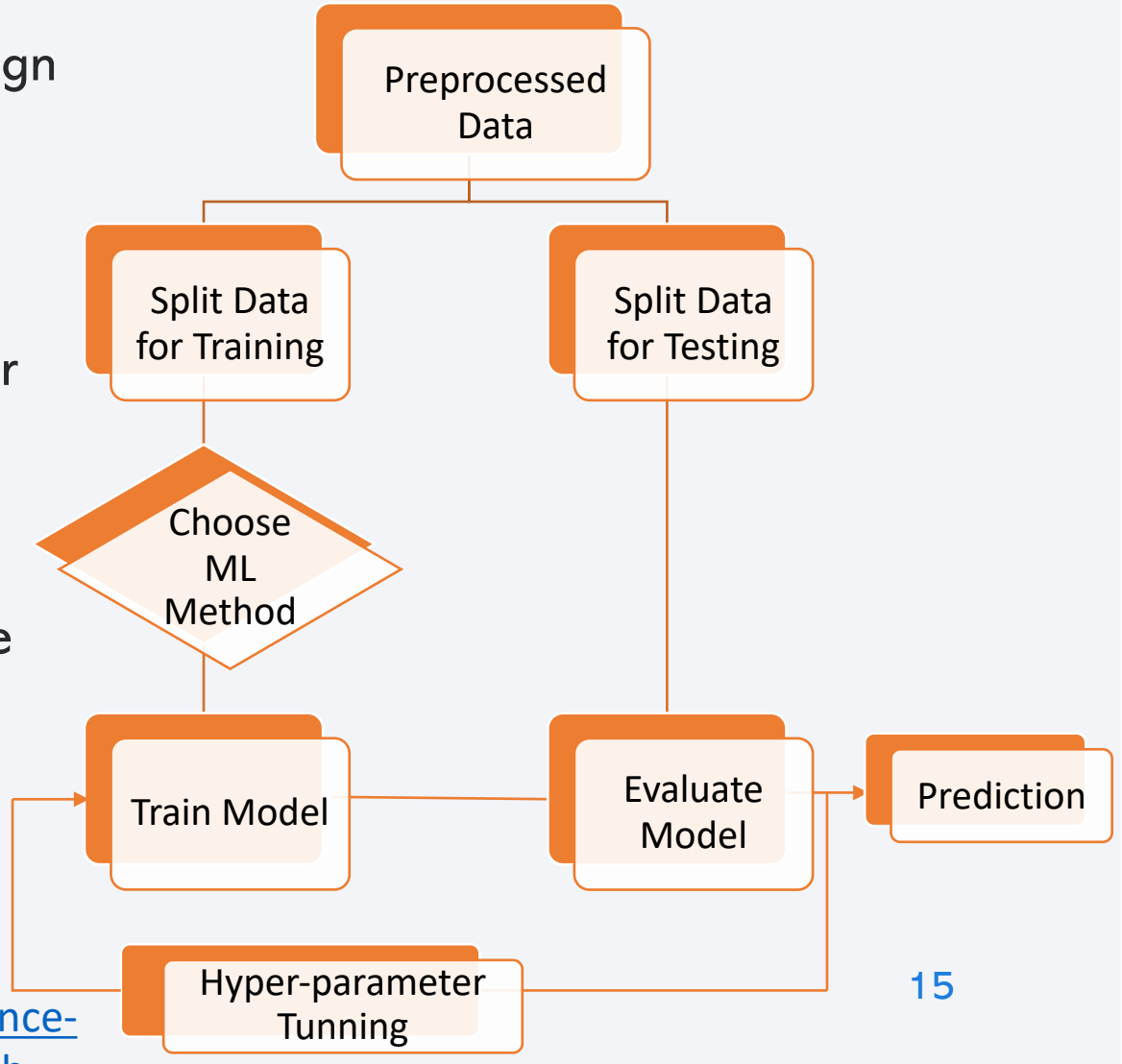
Build a Dashboard with Plotly Dash

- Dropdown list for launch site selection
- Payload mass range slider to select the range of payload mass for data exploration
- Pie chart is used to show successful launches upon all sites or selected launch site
- Scatter chart is to present correlation between payload mass and launch success

GitHub URL: https://github.com/nanjunchen/Rocket-Launch-Data-Science-Project/blob/main/folium_launch_site_location.ipynb

Predictive Analysis (Classification)

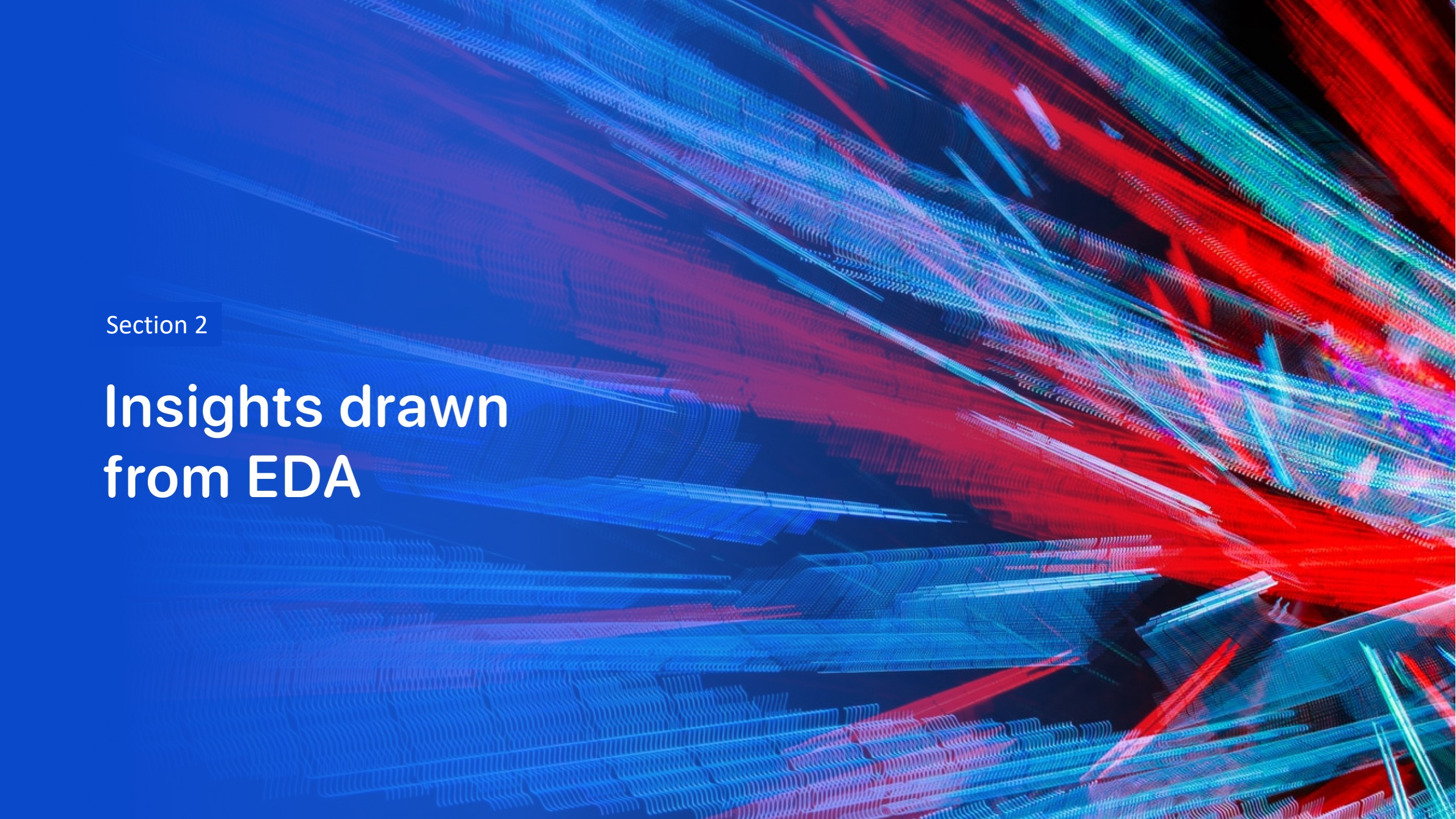
- Create a Numpy array from column “Class” and assign it to dependent variable Y
- Standardize data from independent variables to X
- Split X and Y to training and test data
- Create classification models with multiple parameter sets with Grid Search method
- Train models with training data and find the best parameters for such model
- Evaluate the model with test data by accuracy score and confusion matrix
- Compare the performance by different models with different parameters
- Determine the best performing classification model



GitHub URL: https://github.com/nanjunchen/Rocket-Launch-Data-Science-Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

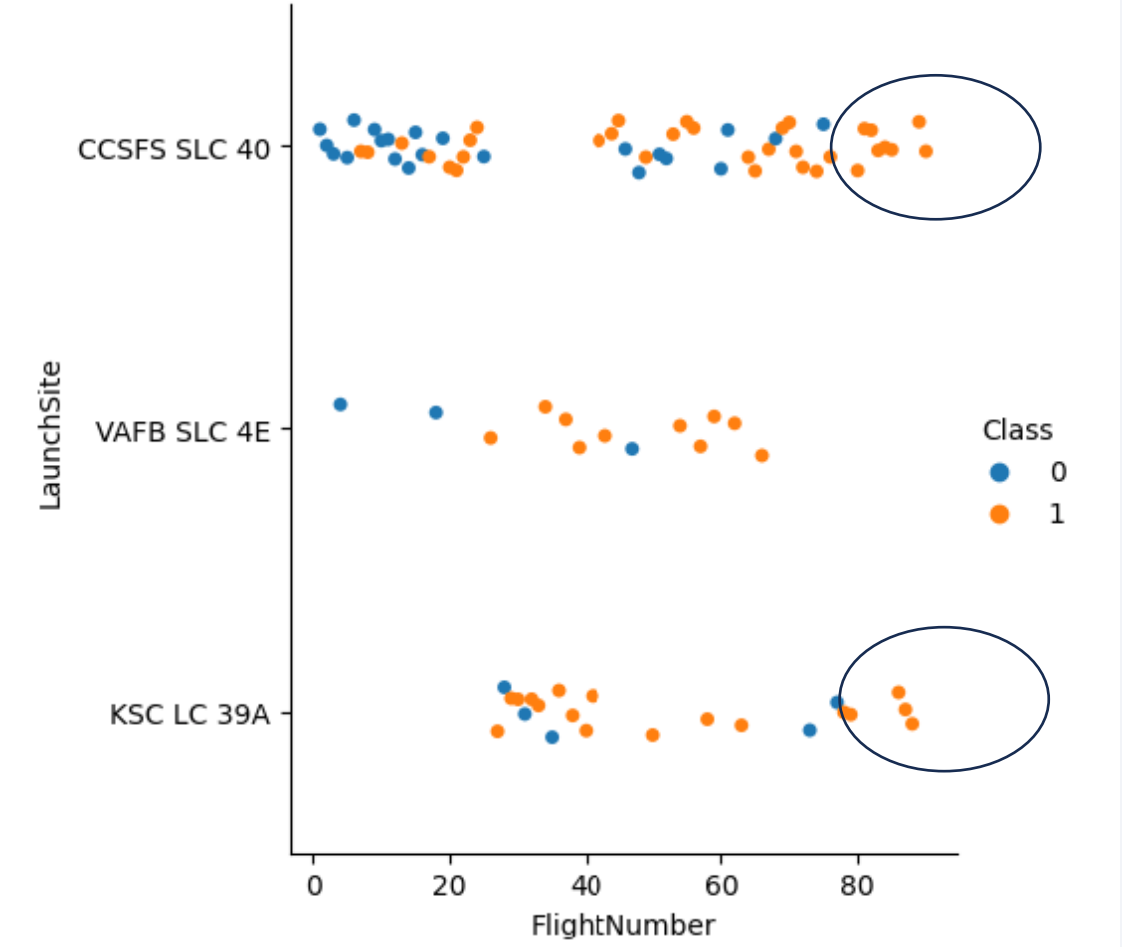
Results

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


Section 2
Insights drawn from EDA

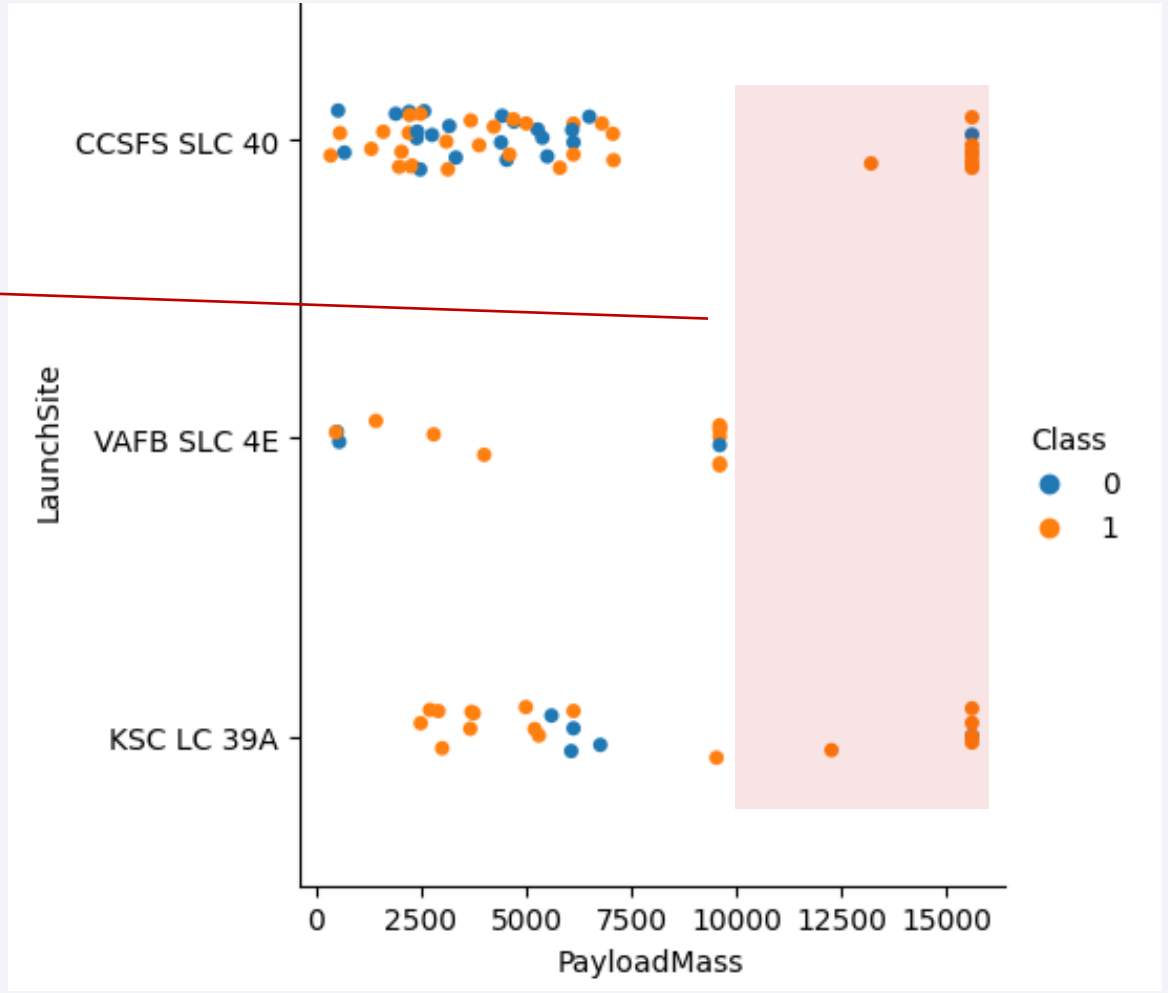
Flight Number vs. Launch Site

- At CCSFS SLC 40 launch site, for Flight Number larger than 80, there exists a high success rate
- VAFB SLC 4E has lower flight number with higher success rate
- KSC LC 39A doesn't have flight number lower than 20
- However, the correlation between flight number and launch sites is not prominent



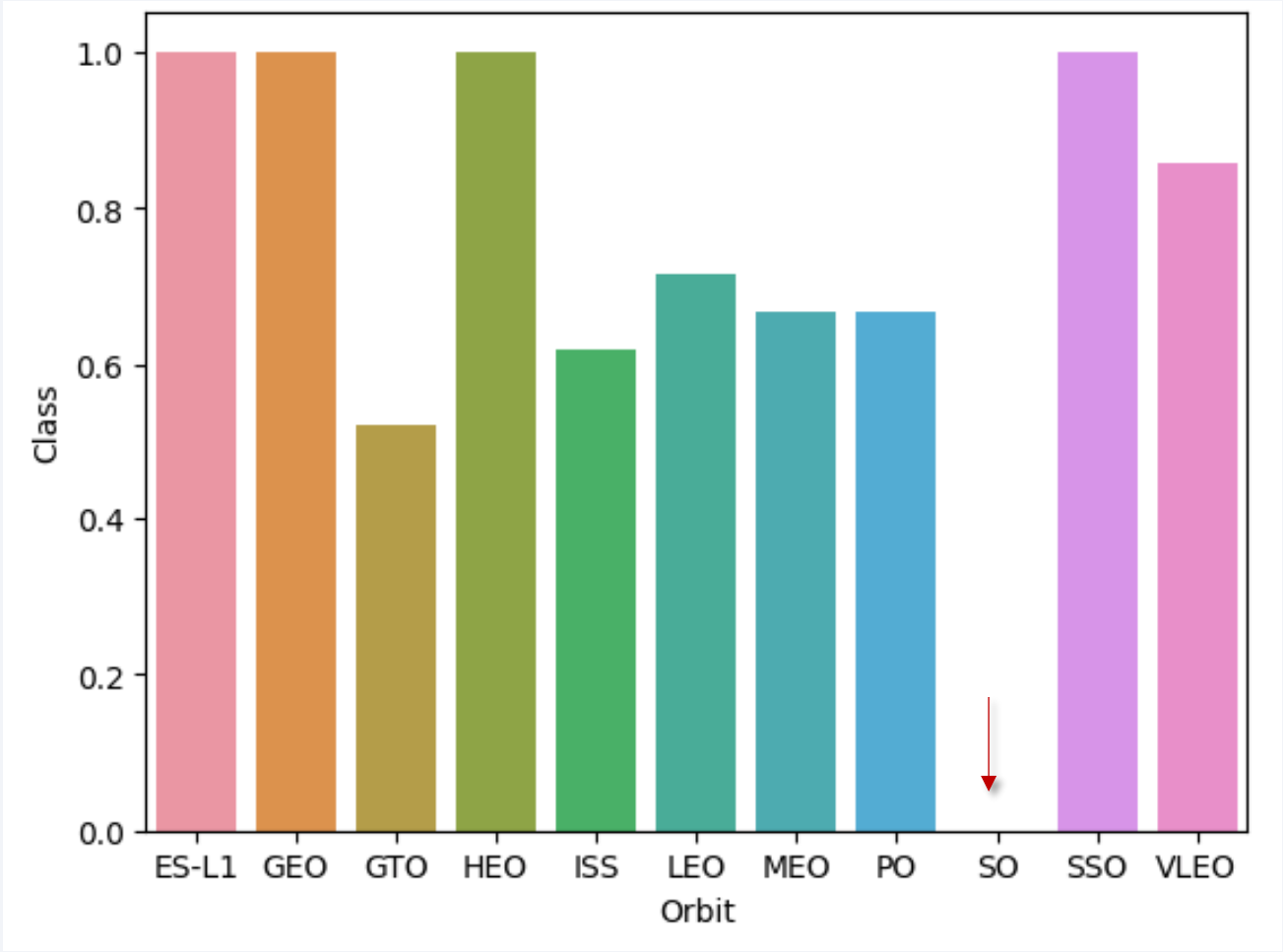
Payload vs. Launch Site

- For heavy payload mass (greater than 10000), CCSFS SLC 40 and KSC LC 39A show a high success rate
- VAFB SLC 4E does not carry heavy payload mass
- Success rate at lower payload mass is not distinguishable among sites



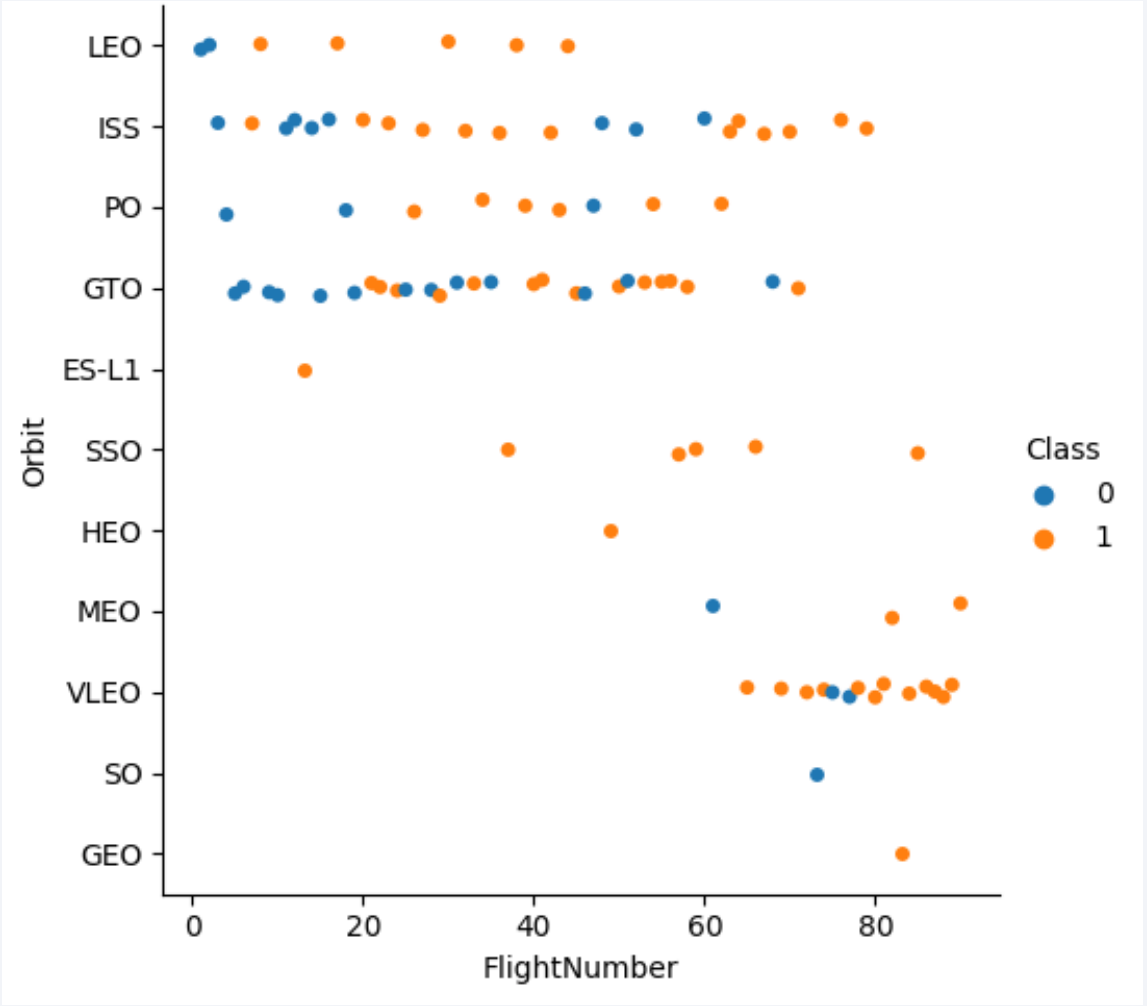
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO all present 100% successful mission rate
- When in SO orbit, all landing fail



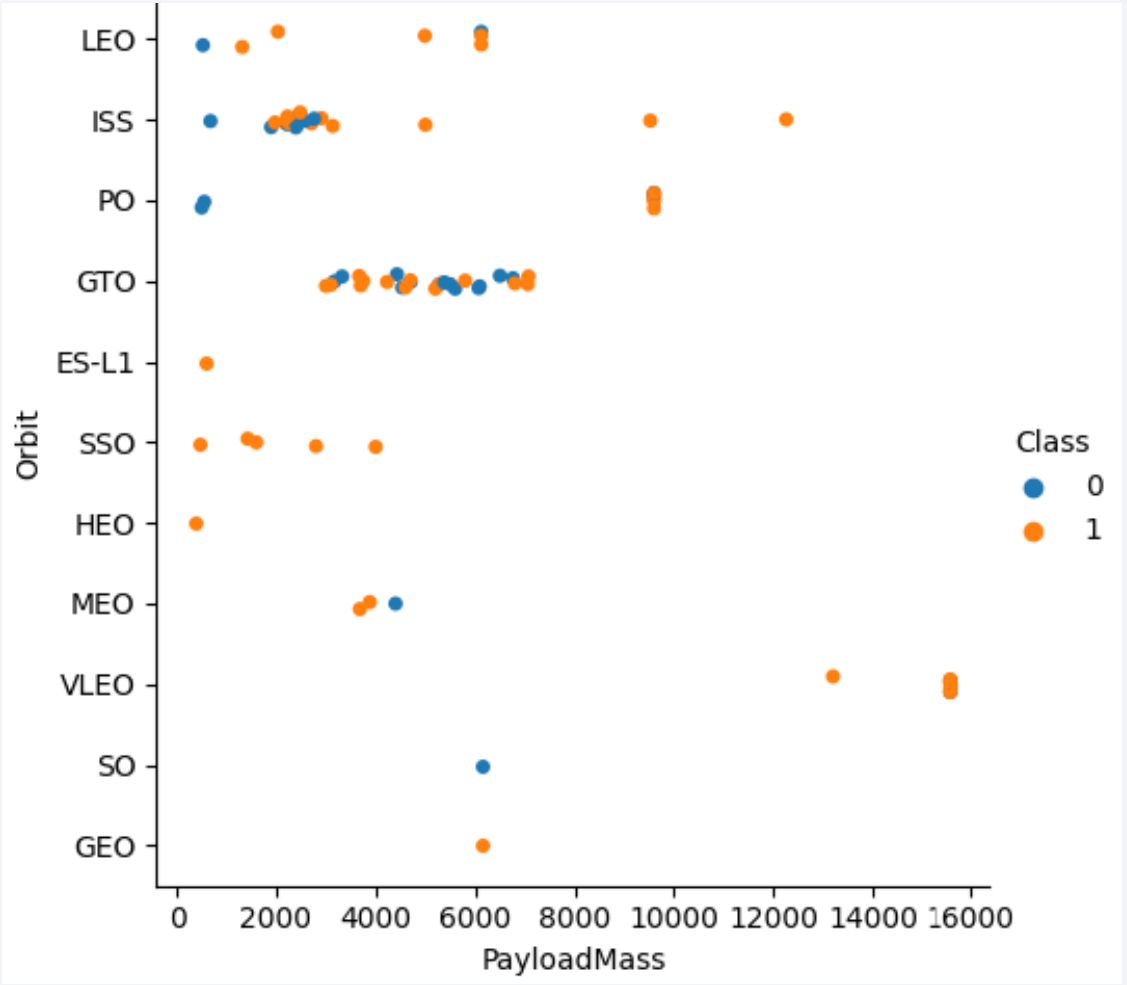
Flight Number vs. Orbit Type

- For LEO orbit, success rate appears related to the number of flights
- Launches in ES-L1, HEO, and GEO orbits are all successful no matter how much number of flights
- No relationship has been found between flight number and GTO orbit



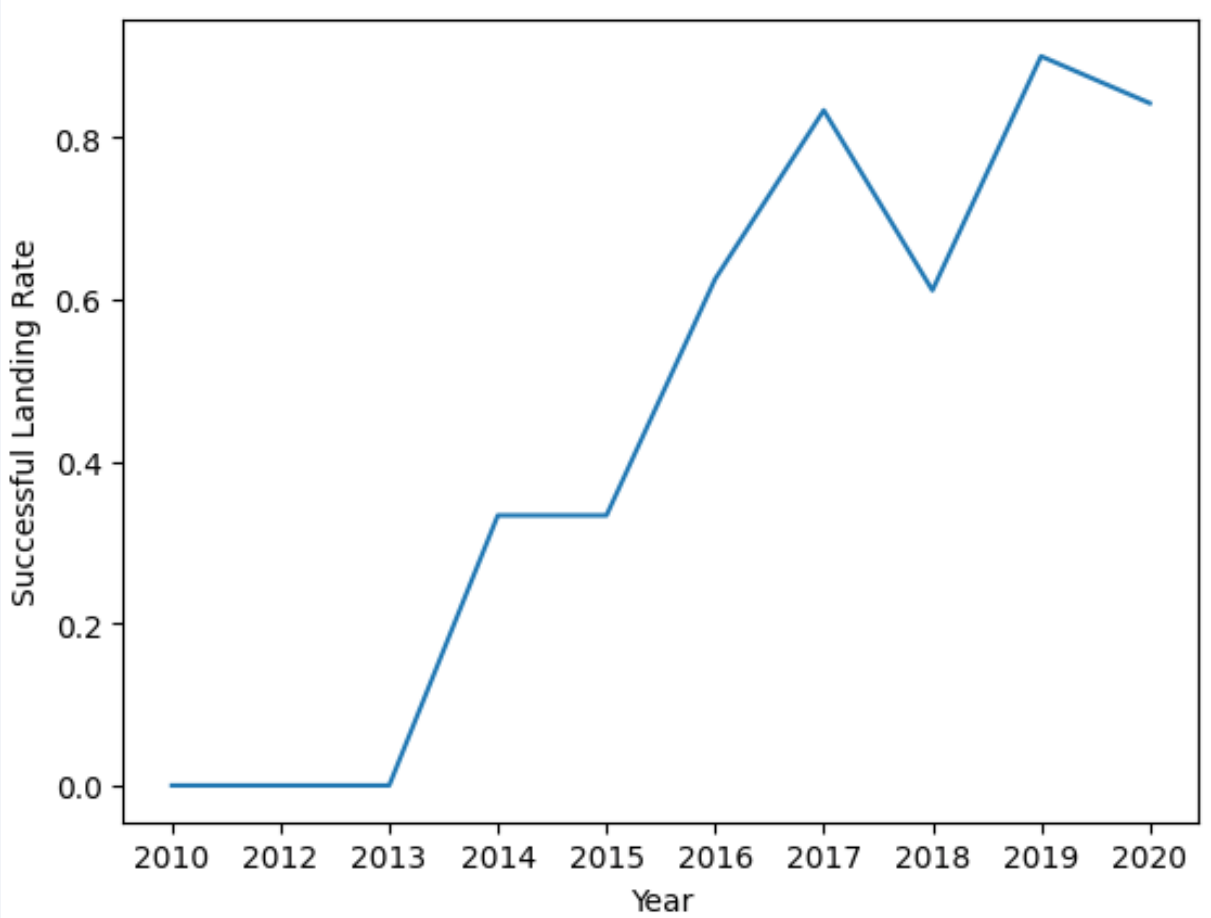
Payload vs. Orbit Type

- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO, and ISS
- When in GTO, no outstanding influence of payload mass has been found on successful/positive landing



Launch Success Yearly Trend

- The success rate kept increasing since 2013
- The rate dropped in 2018 as compared with 2017



All Launch Site Names

- Find the names of the unique launch sites
- Launch site column was queried from the SPACEXTABLE by grouping the unique launch site name
- Four unique launch sites are found

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- From launch site column, we set a WHERE condition for launch site name started with ‘CCA’ and limit to only 5 of them in the returned list

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Where condition is set to when customer is NADA (CRS), the returned payload mass carried by boosters launched by NASA are added together
- The total payload carried by booster from NASA is 45596 kg

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Where condition is set to when booster version is F9 v1.1, the mean value is calculated for returned payload mass carried by this booster
- The average payload mass is 2928.4 kg carried by F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Use min() function on the Date column and set WHERE condition to when landing outcome is 'Success (ground pad)'
- The first successful landing on ground pad was dated on Dec. 22nd, 2015

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- WHERE condition is used to narrow down to a certain range of payload mass and when landing outcome is “Success (drone ship)”, and the booster versions are returned which that satisfy the conditions.
- Four booster versions satisfy the condition

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Use COUNT() function on the mission outcome column to count the number of rows in this column
- There are 101 mission outcomes

COUNT(Mission_Outcome)
101

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Subquery is used to first return the maximum payload mass with its associated booster version and define it as a new table “Boost”. From that, we return the booster version as follow:

Booster_Version
F9 B5 B1048.4

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Get month and year by substr() function and filter data by WHERE conditions for “Failure (drone ship)” and year 2015
- Two records are returned for same launch site but different booster version

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

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

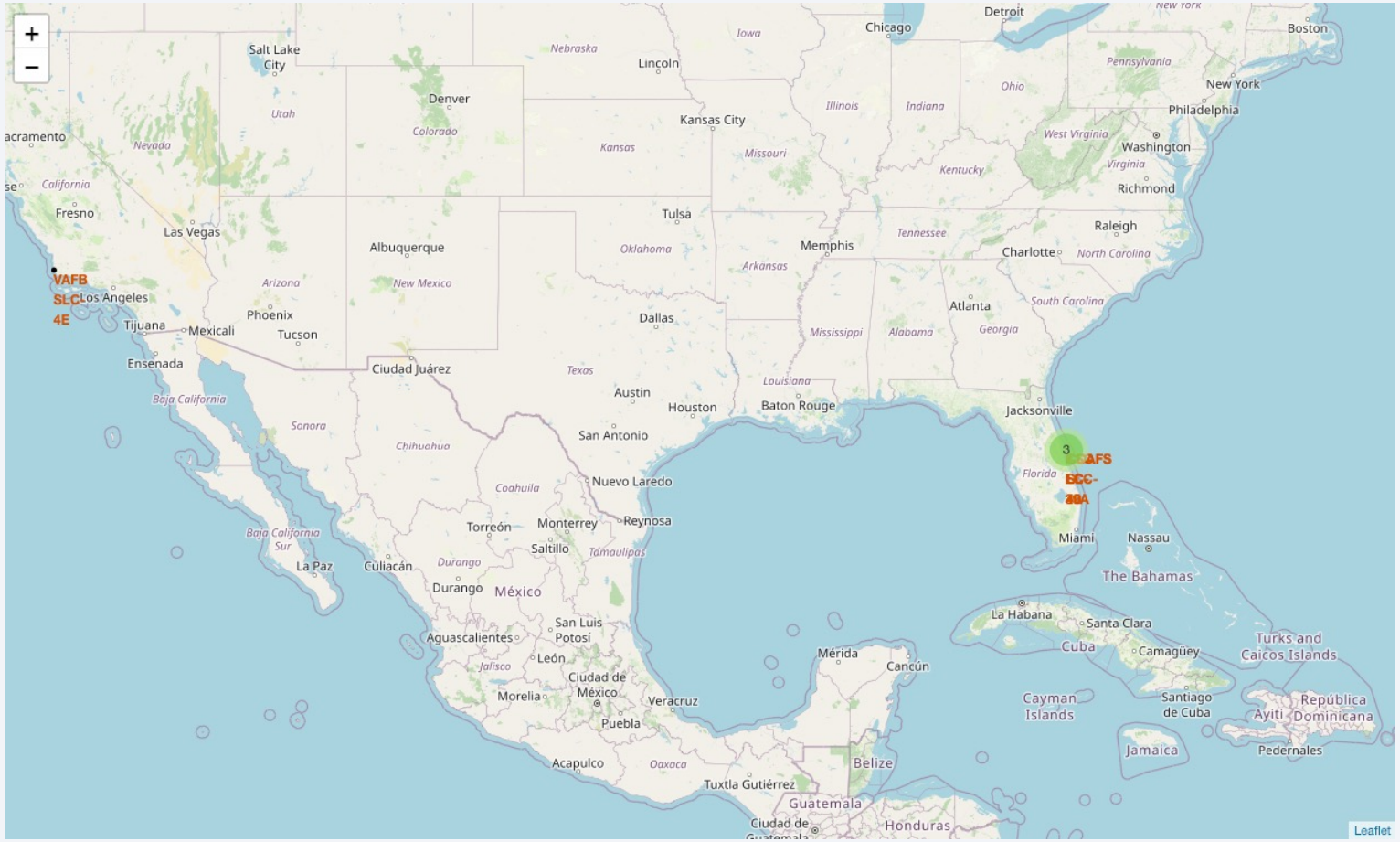
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- The highest rank of landing outcome is “No attempt”

Date	Landing_Outcome	countL
2012-05-22	No attempt	10
2015-12-22	Success (ground pad)	5
2016-08-04	Success (drone ship)	5
2015-10-01	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2015-06-28	Precluded (drone ship)	1
2010-08-12	Failure (parachute)	1


Section 3
Launch Sites
Proximities Analysis

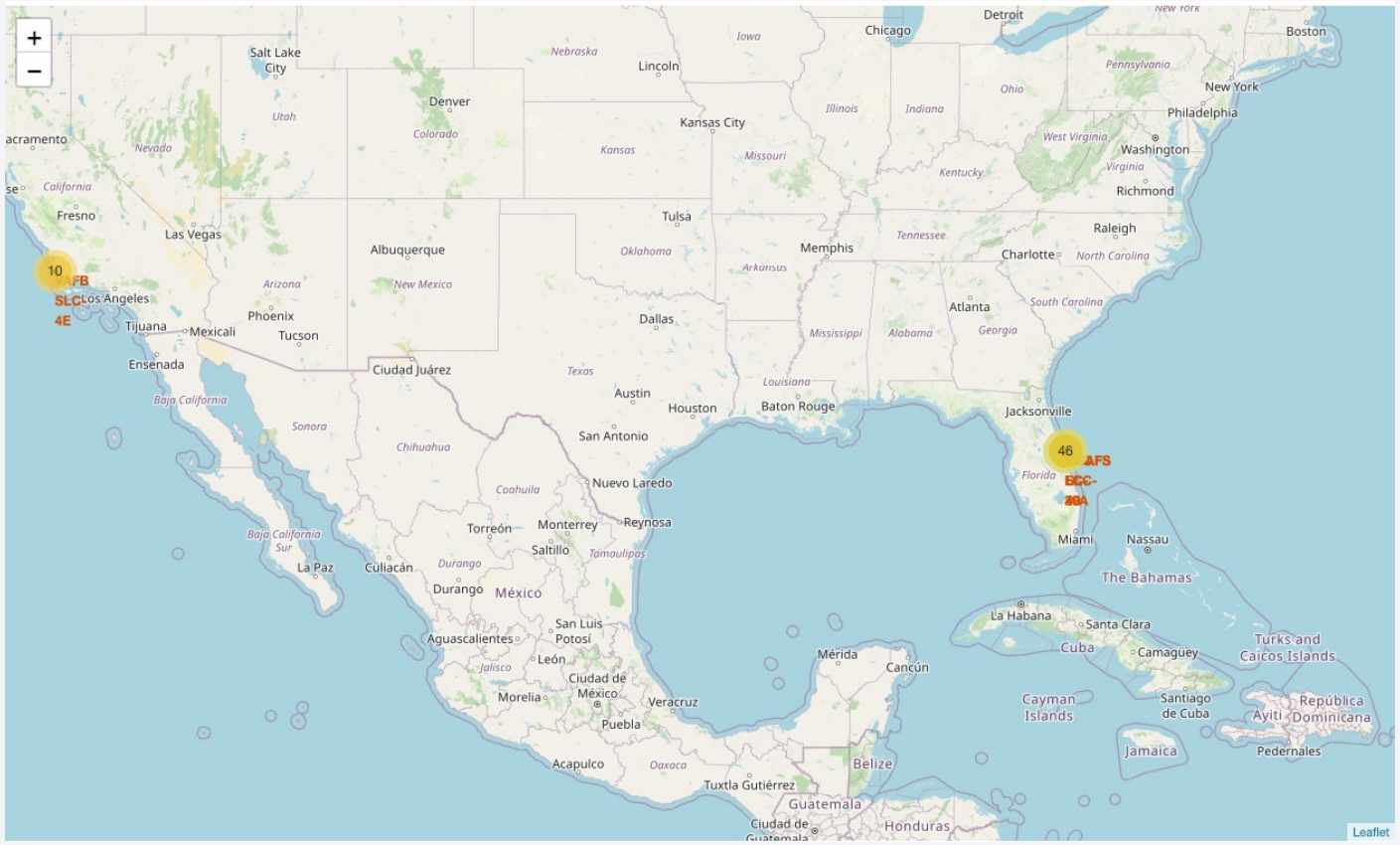
All Launch Sites on Map

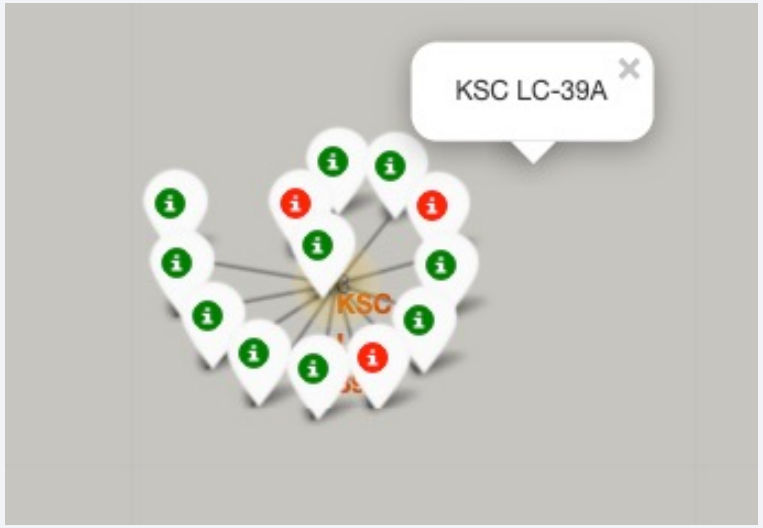
- Four launch sites are marked with one in California and other three in Florida



Success/failed Launch Outcomes for Each Site

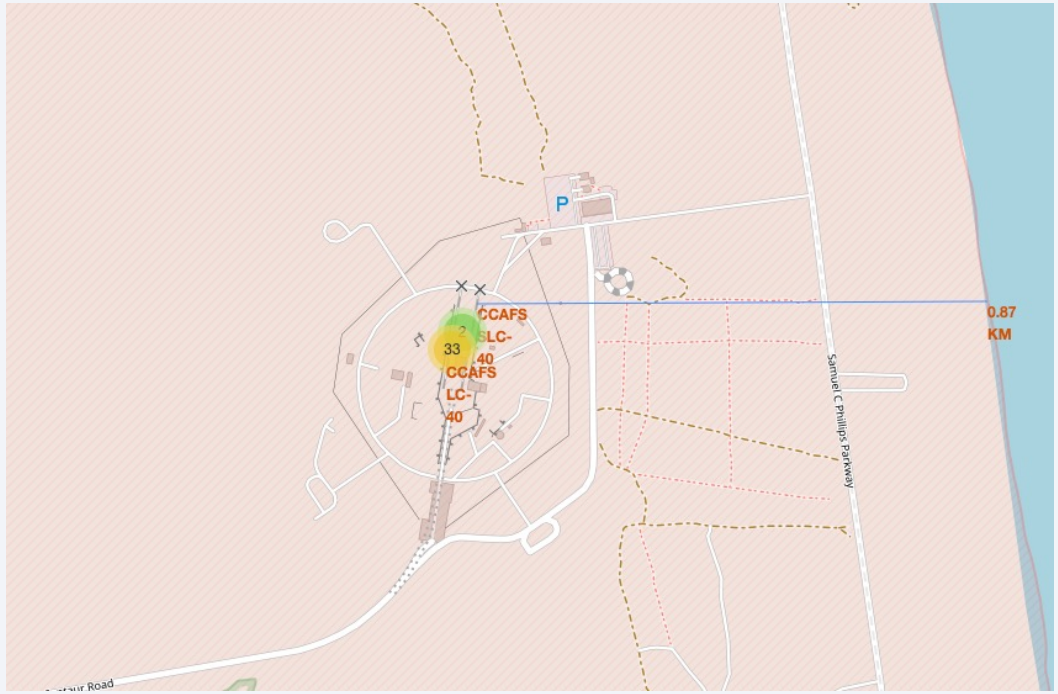
- Success/failed launches are included in Marker Cluster
- Ten successful outcomes and three failures are marked for KSC LC-39A site, which has highest successful rate among sites

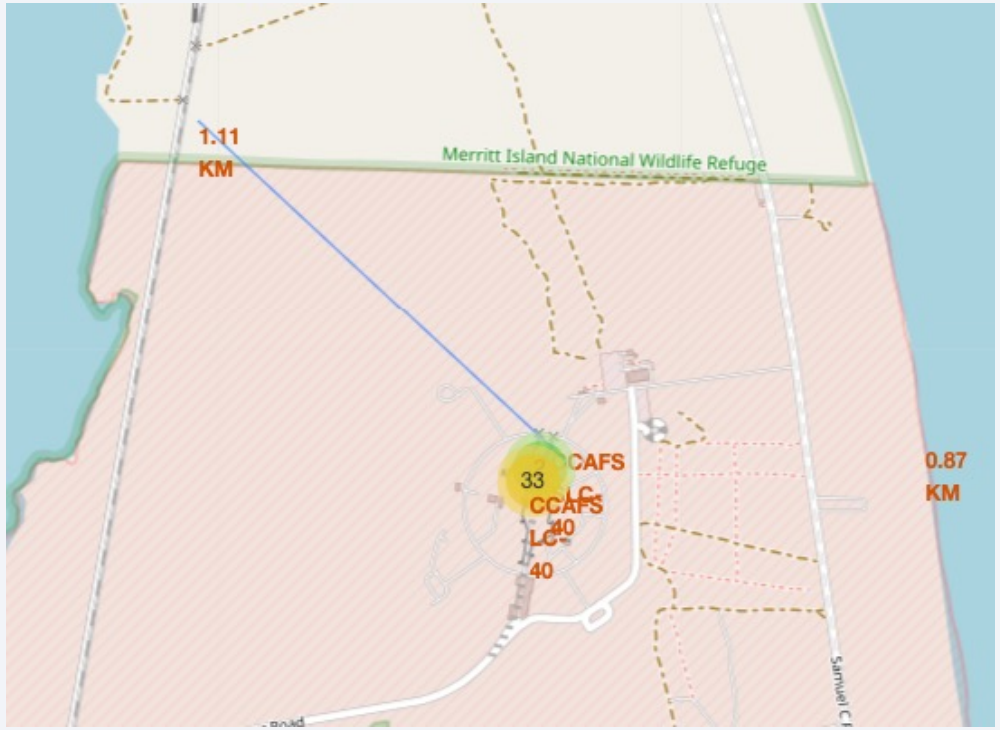


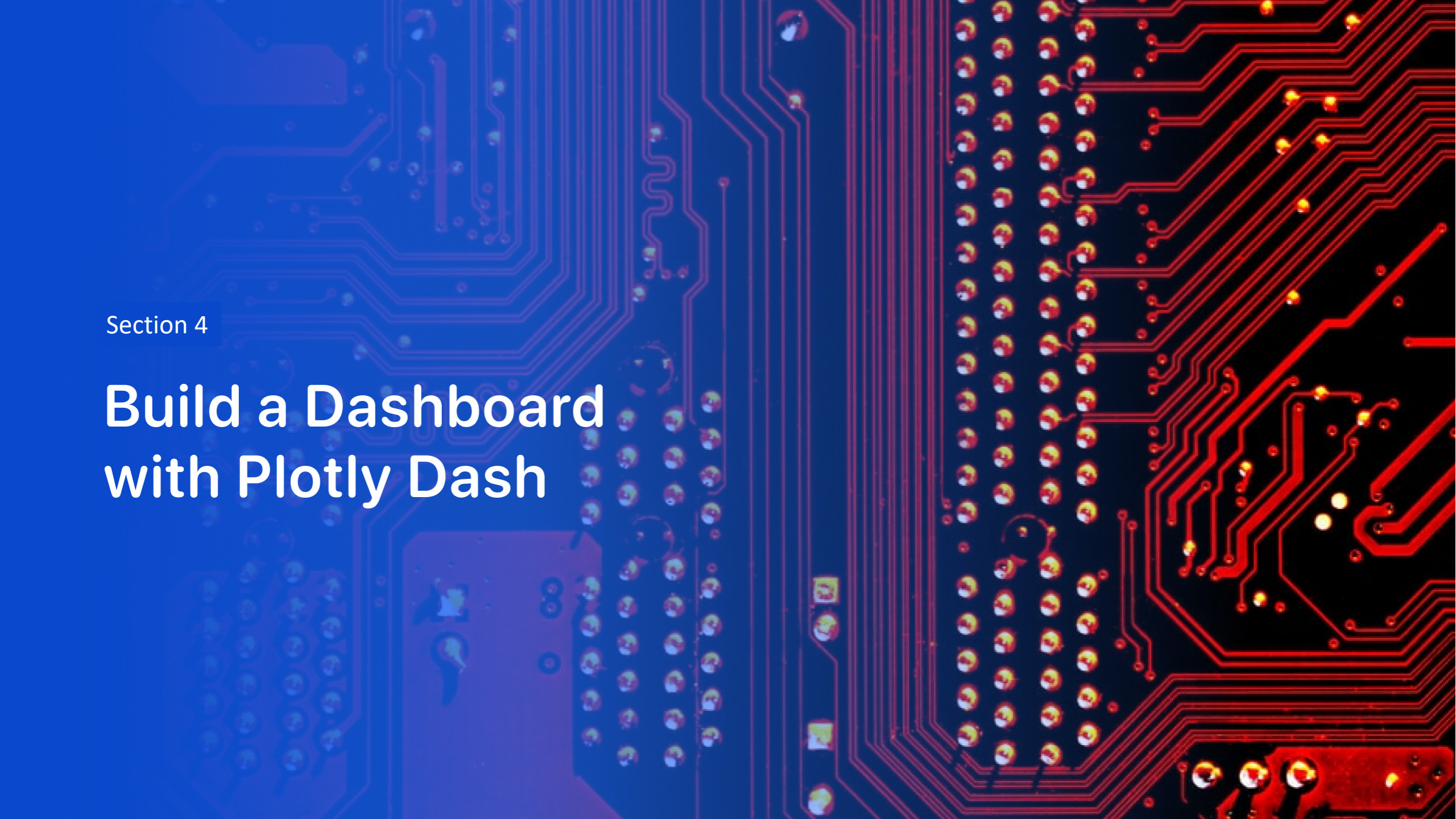


Proximities of Launch Sites

- Launch sites are closer to coastline than railway, highway and city centers

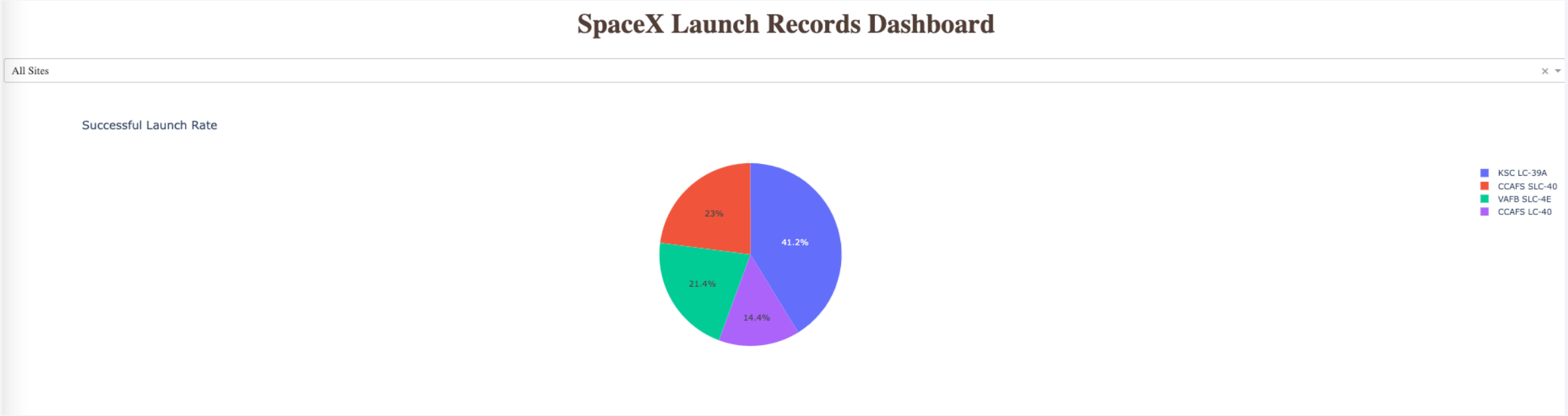





Section 4
Build a Dashboard with Plotly Dash

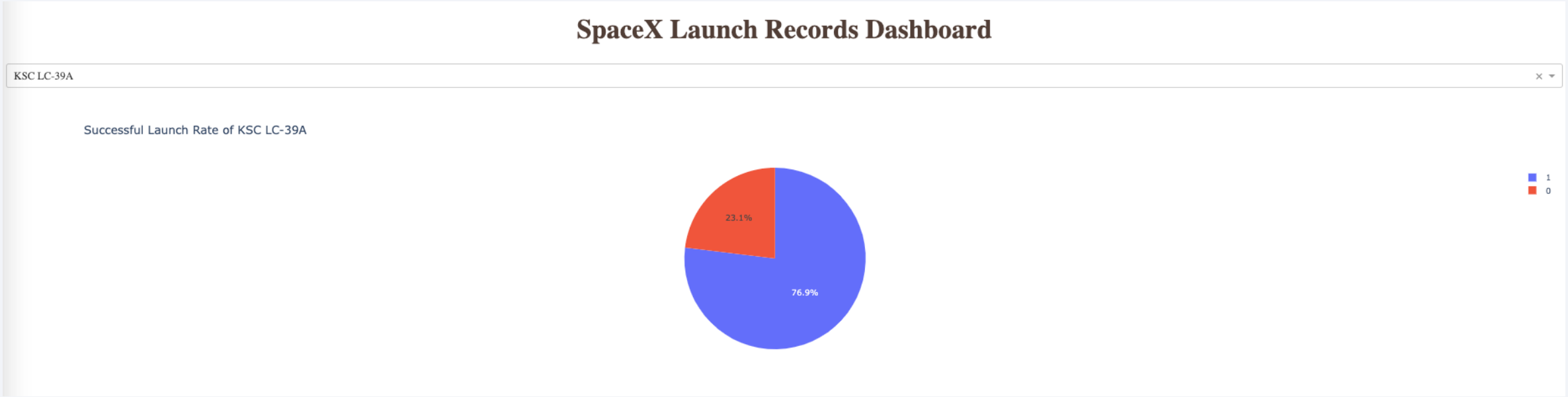
Successful Launch Rate at All Sites

- KSC LC-39A show highest positive landing rate of 41.2% while CCAFS LC-40 presents the lowest with 14.4%
- The other two sites demonstrate a close rate ~ 20%



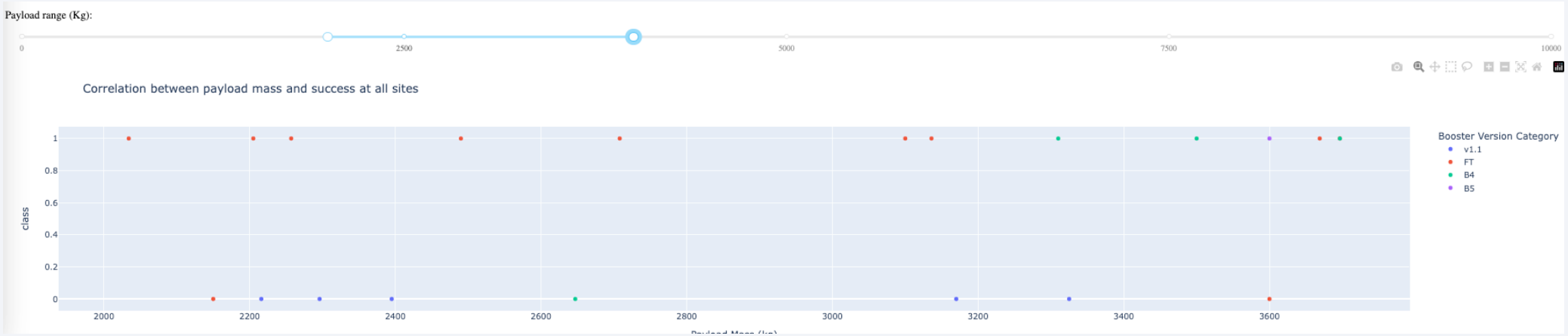
Successful/Failure Launch Rate of KSC LC-39A

- 76.9% of the mission outcomes are successful at KSC LC-39, while 23.1% are failed



Launch Outcome for a Certain Payload Range

- For payload range in 2000~3600 kg, booster FT has the highest successful rate

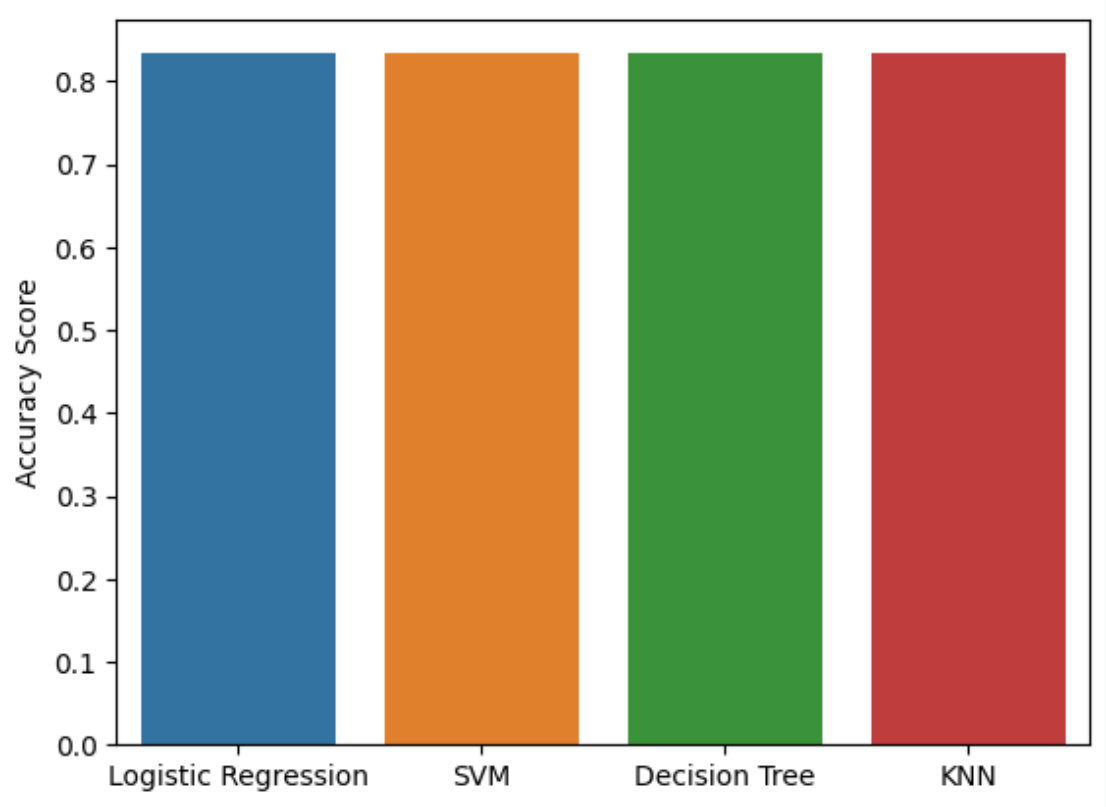


Section 5

Predictive Analysis (Classification)

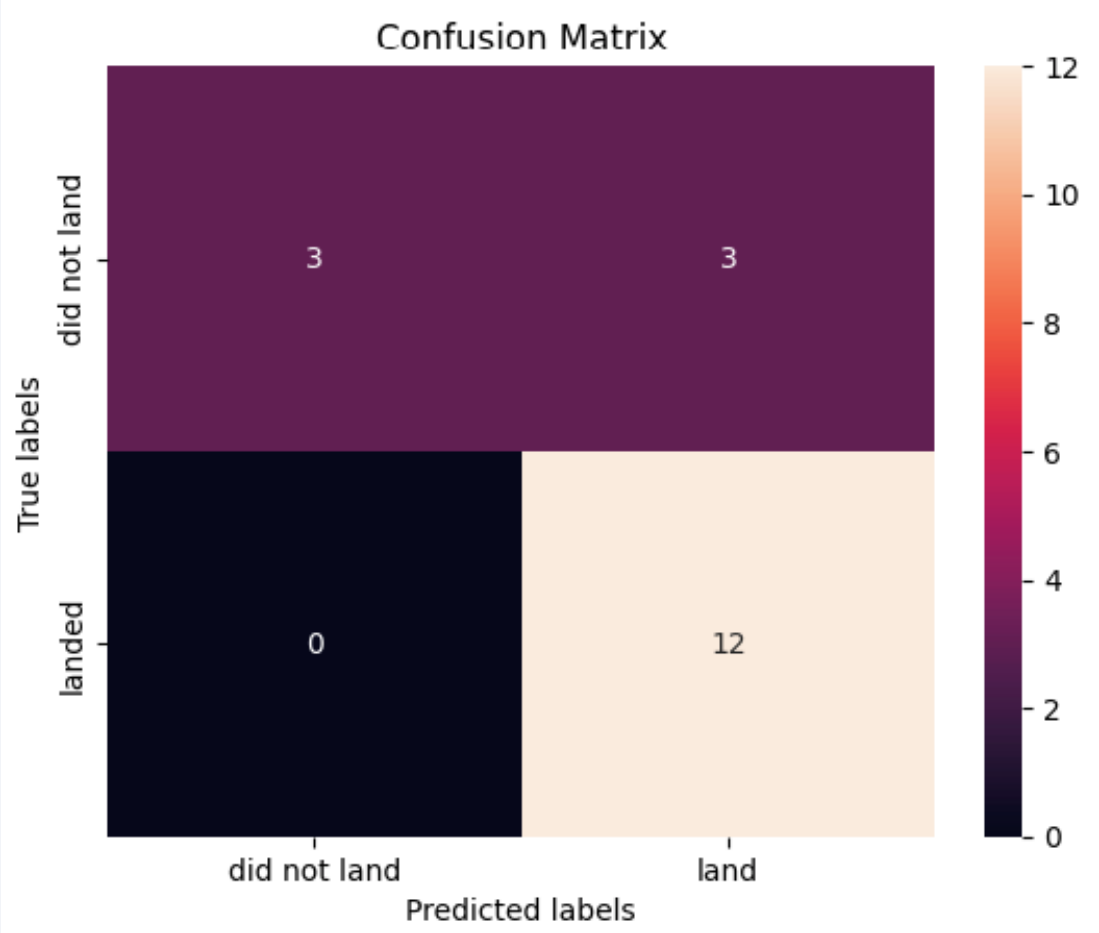
Classification Accuracy

- Accuracy score are practically identical for chosen classification methods
- This is because the dataset is small (only18 samples) and has lesser values



Confusion Matrix

- It has successfully predicted 12 cases with successful landing while 3 cases for failed landing
- There are three inaccurate prediction for those didn't land successfully



Conclusions

- The role of Flight Number, Launch Site, Orbit Type, Payload Mass, Booster Version, etc. played in landing outcomes are extensively explored
- Number of flights has weak correlation with launch site in general to affect the landing
- Orbit type play an important role in determining whether first stage will land
- Successful/Failure landings at each launch site marked on map give a clearly geographic location and the launch site are tend to be located closer to coastline
- Multiple classification techniques are used to train data and evaluate with test data
- Methods chosen show same performance on the prediction attributed to limited test data used for evaluation

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!