Winning Space Race with Data Science

<Chia-Ying Tsao>
<2021/10/06>

Outline

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

Executive Summary

- Summary of methodologies
  - To complete the goal, we apply machine learning models such as SVM, decision tree, KNN and LogisticRegression.
- Summary of all results
  - From the plots we get by EDA, we found the variables that have positive or negative relationship with the success rate.

Introduction

Project background

The project is about predicting if the Falcon9 first stage landing will success.

SpaceX is a company that launches advanced rockets.

Since every launch cost a lot, it could be helpful if we know whether the first stage will success to determine the cost.

Problems we want to find answers

- What are the key factors of the landing success?
- Which model has the highest prediction accuracy?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - HTTP requests / SpaceX APIl
- Perform data wrangling
  - Describe how data was processed
- 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

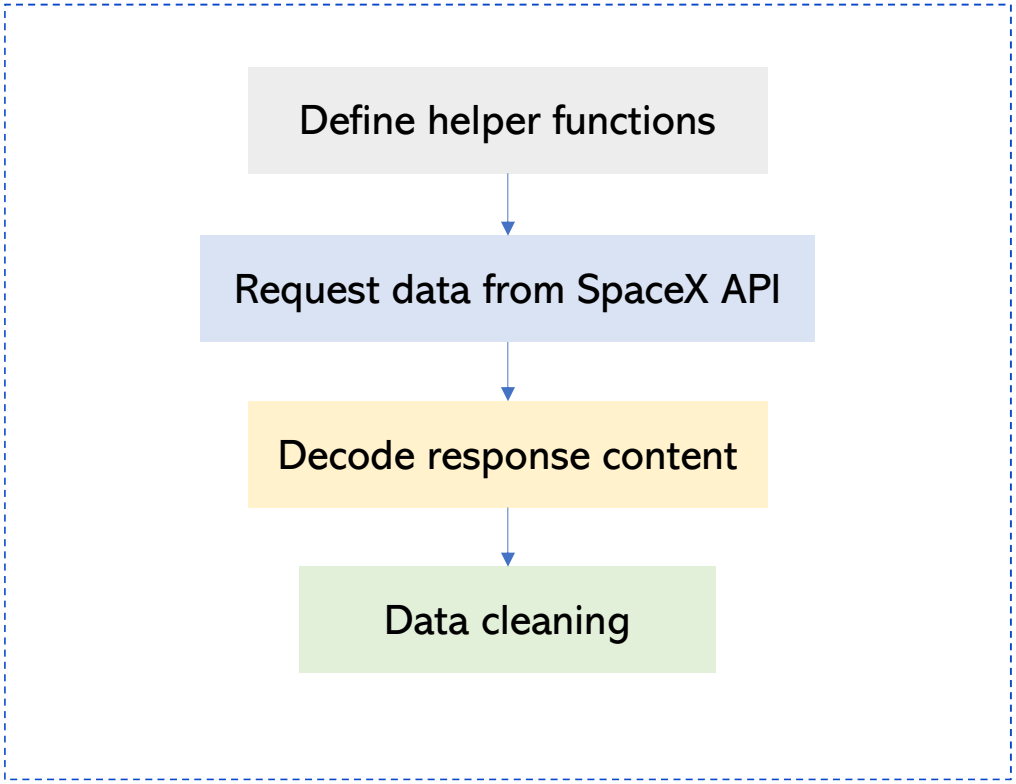
Data Collection

- There are two ways we collect the SpaceX data
  - API
  - Web scraping

Data Collection – SpaceX API

- The data was collected from SpaceX API by python request library
- Data content including booster version, launch site, payload and other information
- GitHub URL

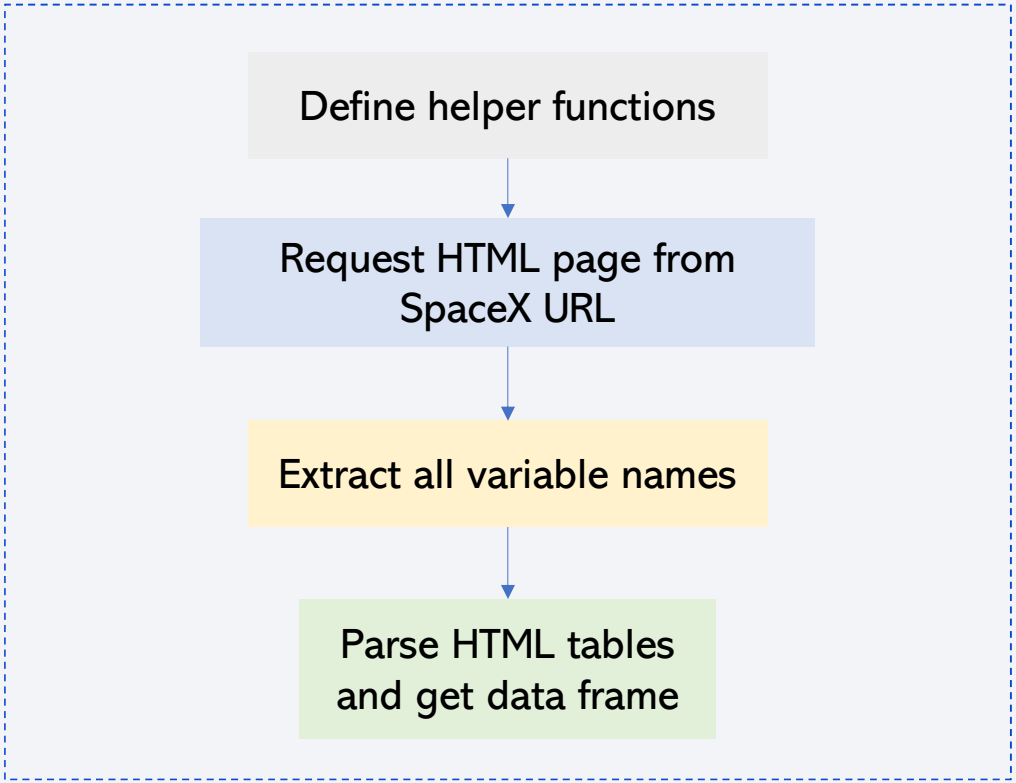
  [https://github.com/elvish7/Coursera-applied-data-science-capstone/blob/main/Data%20Collection%20API%20Lab.ipynb]



Data Collection - Scraping

- Scrape the HTML content by python package beautifulsoup
- Parse the HTML tree structure
- GitHub URL

  [https://github.com/elvish7/Coursera-applied-data-science-capstone/blob/main/Data%20Collection%20API%20Lab.ipynb]



Data Wrangling

- Prepare the data and transform data format
- GitHub URL

[https://github.com/elvish7/Coursera-applied-data-science-capstone/blob/main/Data%20wrangling.ipynb]

Deal with missing values	Data exploration	Derivative features
calculate the percentage of the missing values in each attribute	calculate numbers of launch or occurrence of each label	create “landing_outcome” to label whether the outcome is successful

EDA with Data Visualization

- Chart Types
  - Scatter Plot – visualize whether two variables has observable relationship
  - Bar Plot – compare the success rate of different launch sites
  - Line Plot – show the yearly change
  - Map – locate the launch sites and other landmark on the map
- GitHub URL

  https://github.com/elvish7/Coursera-applied-data-science-capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- SQL queries
  - SELECT * FROM
  - WHERE (condition)
  - DISTICT
  - GROUPBY / ORDERBY
- GitHub URL

  https://github.com/elvish7/Coursera-applied-data-science-capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Summarize map objects
  - Circle - add a highlighted circle area with a text label on a launch site's coordinate
  - Marker – create an icon to show the location name
  - Line – indicate the distance between launch site and other location
- GitHub URL

  https://github.com/elvish7/Coursera-applied-data-science-capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

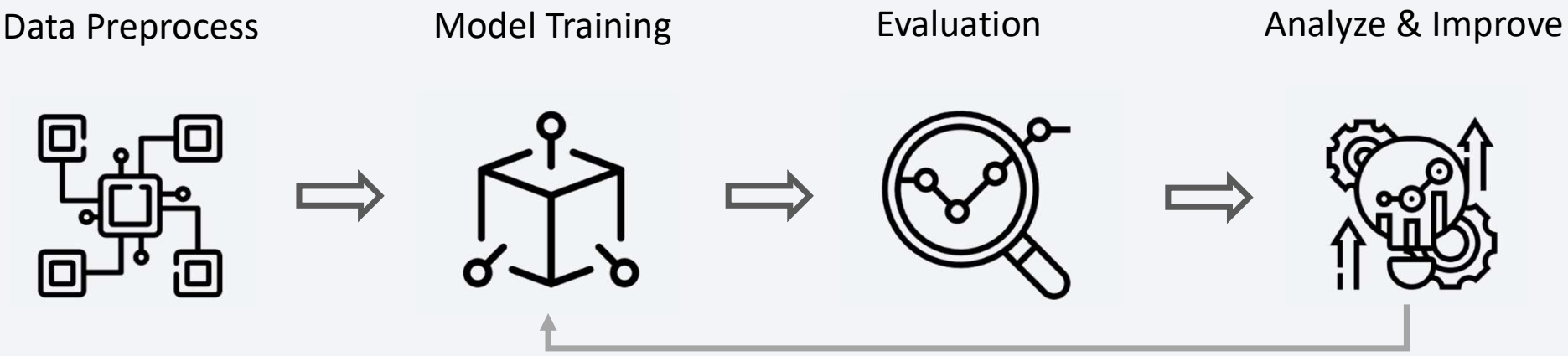
Build a Dashboard with Plotly Dash

- Plot/graphs and interactions on the dashboard
  - Pie chart – show the success rate of each launch site
  - Scatter Plot – show the correlation between payload and outcome
  - Drop down / range slider
- GitHub URL

  https://github.com/elvish7/Coursera-applied-data-science-capstone/blob/main/spacex_dash_app.py

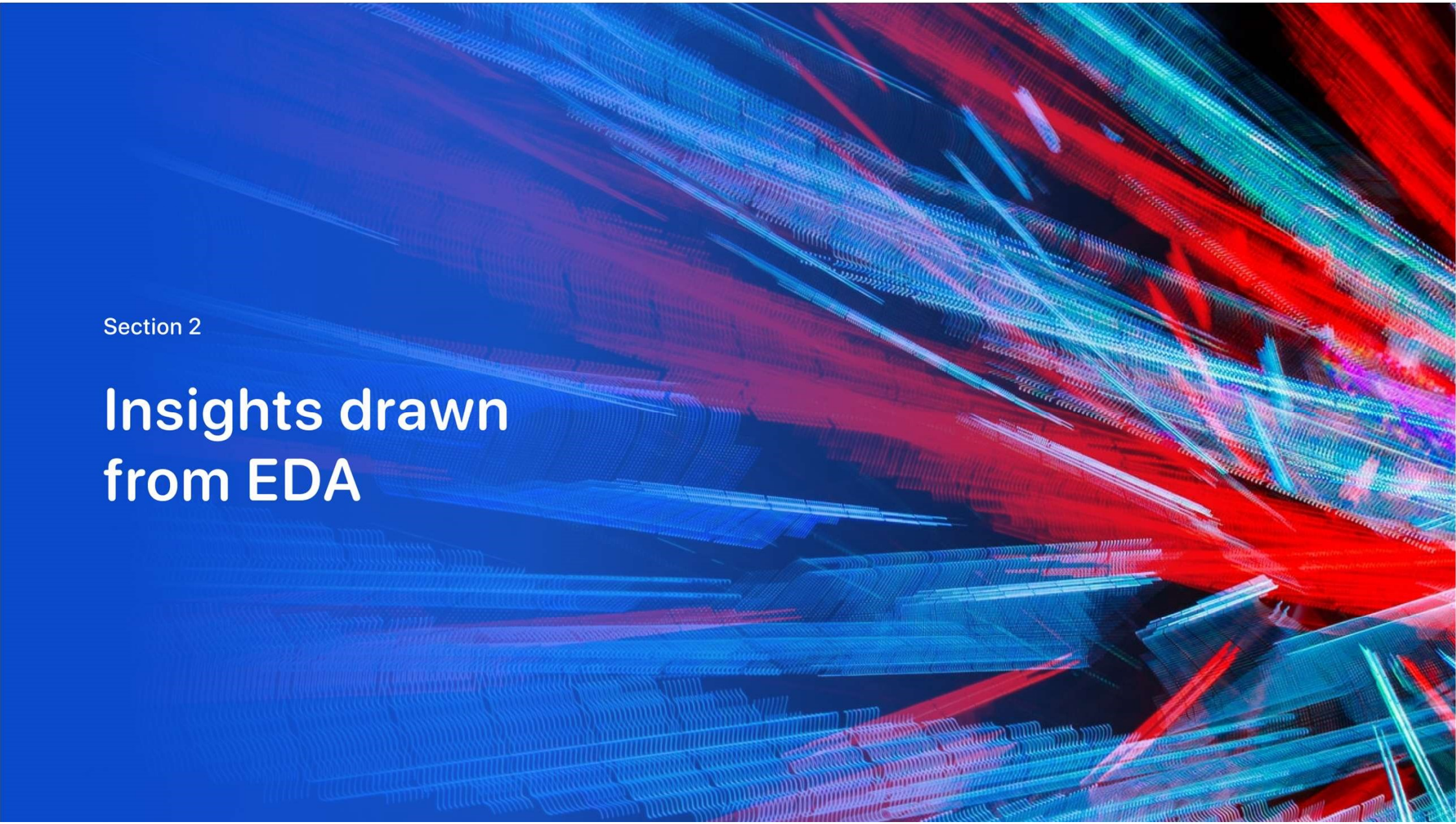
Predictive Analysis (Classification)

- Apply classification models
  - KNN / logistic regression / decision tree / SVM
- GitHub URL https://github.com/elvish7/Coursera-applied-data-science-capstone/blob/main/Machine%20Learning%20Prediction.ipynb



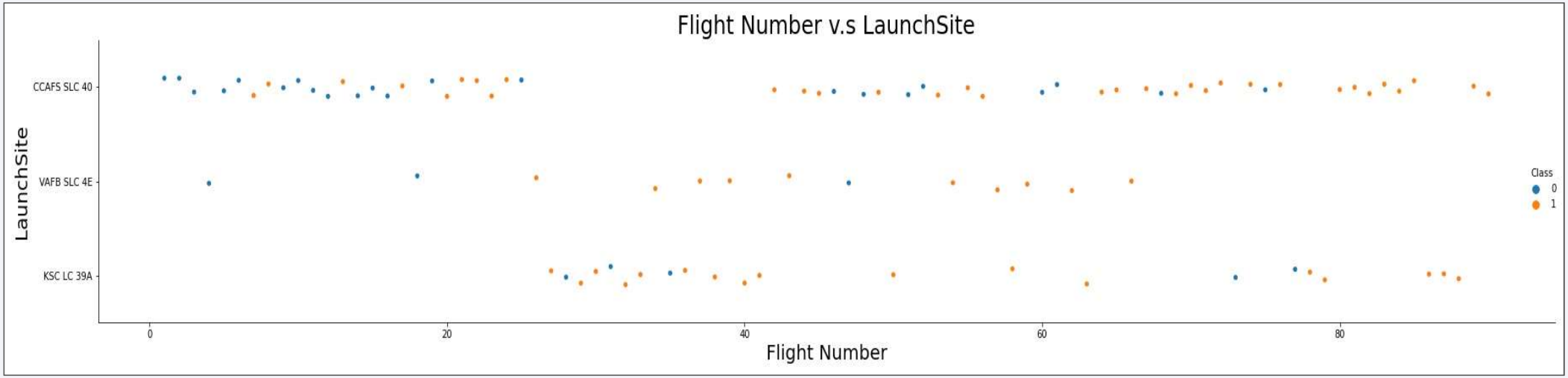
Results

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


Section 2
Insights drawn
from EDA

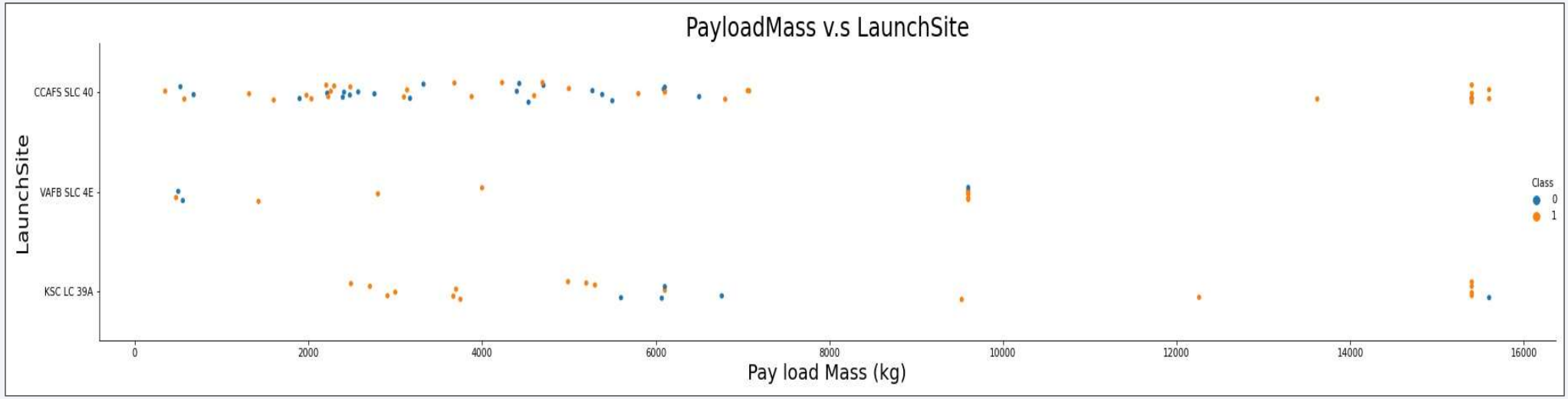
Flight Number vs. Launch Site

- For launch site CCAFS LC-40, it seems to be more likely to success when flight number is high
- KSC LC-39A and VAFB SLC 4E have relatively high flight number
- Most of the launch is success for launch site KSC LC-39A



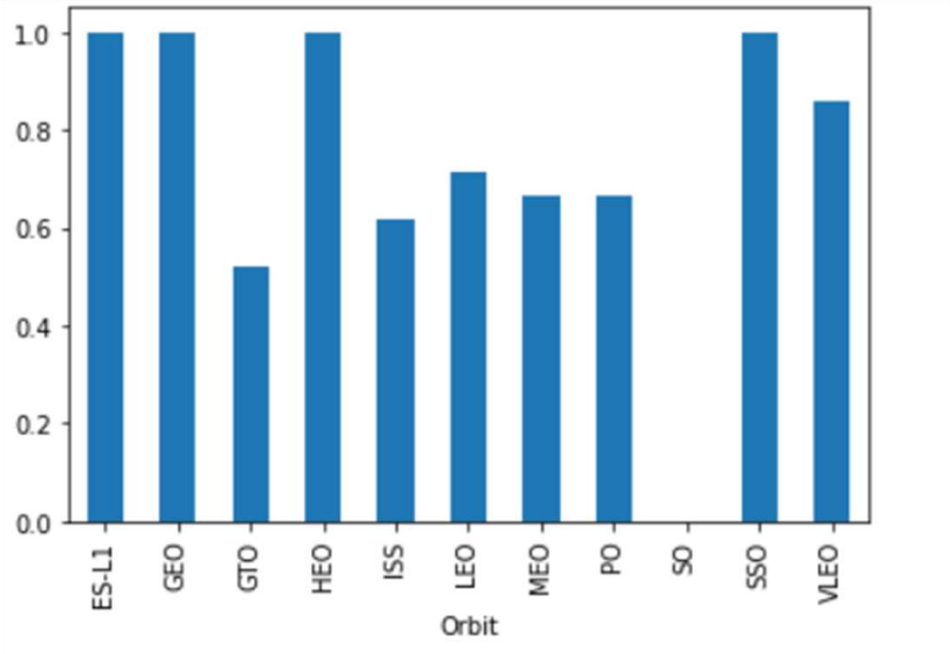
Payload vs. Launch Site

- Most of the launches with high payload mass have successful outcome
- Launch site CCAFS LC-40 has the average lowest payload mass



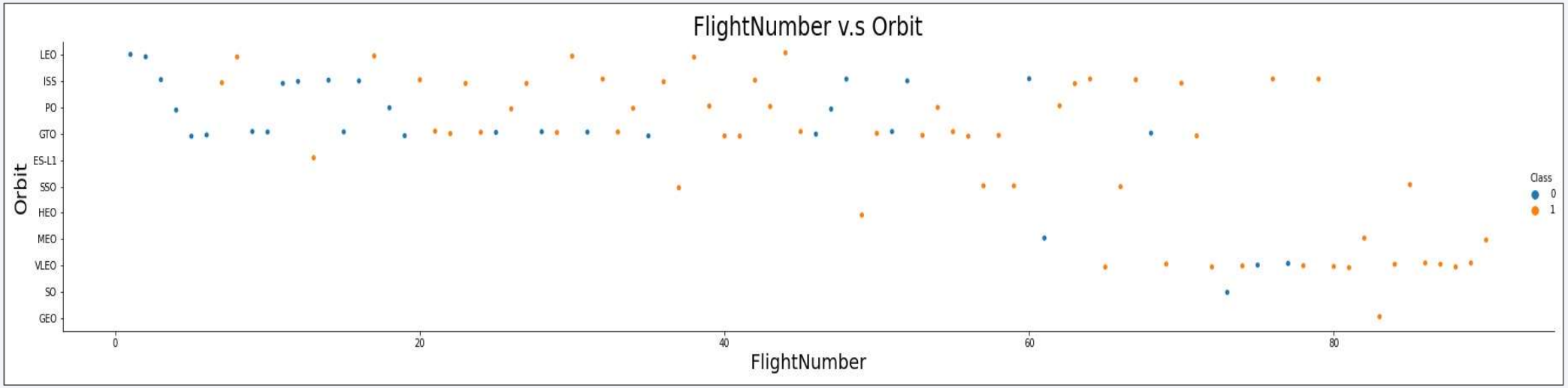
Success Rate vs. Orbit Type

- Orbit type ES-L1, GEO, HEO and SSO has the highest success rate (> 90%)
- Orbit type GTO has approximately 50% for success
- Orbit type SO has the lowest success rate (maybe due to lack of data)



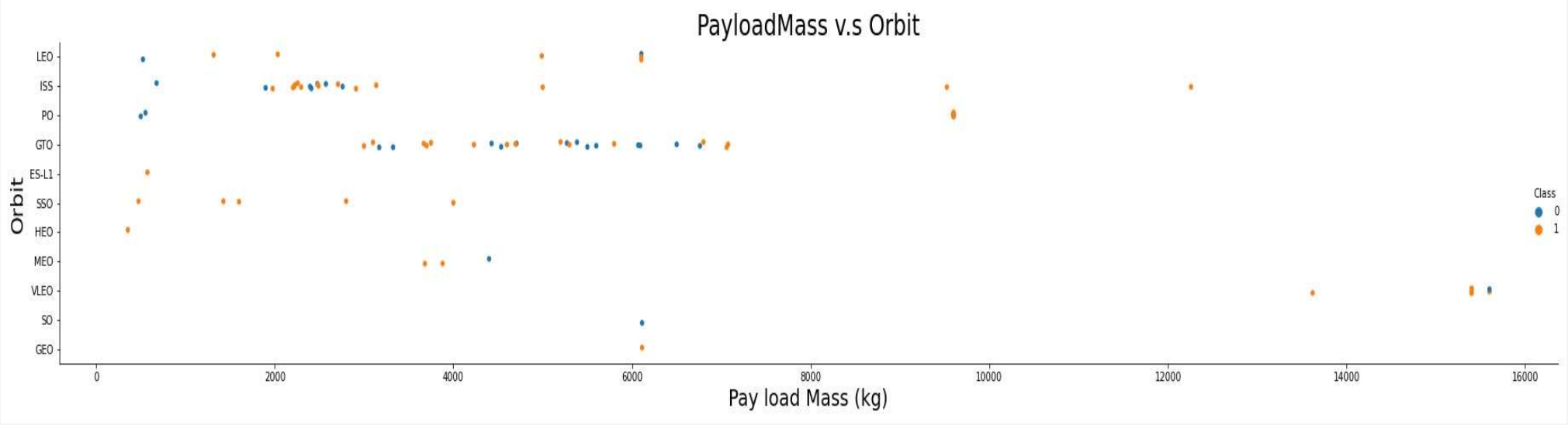
Flight Number vs. Orbit Type

- Some orbit types such as VLEO and MEO have more flight number
- As the flight number increases, the outcome is more likely to be success for every orbit types



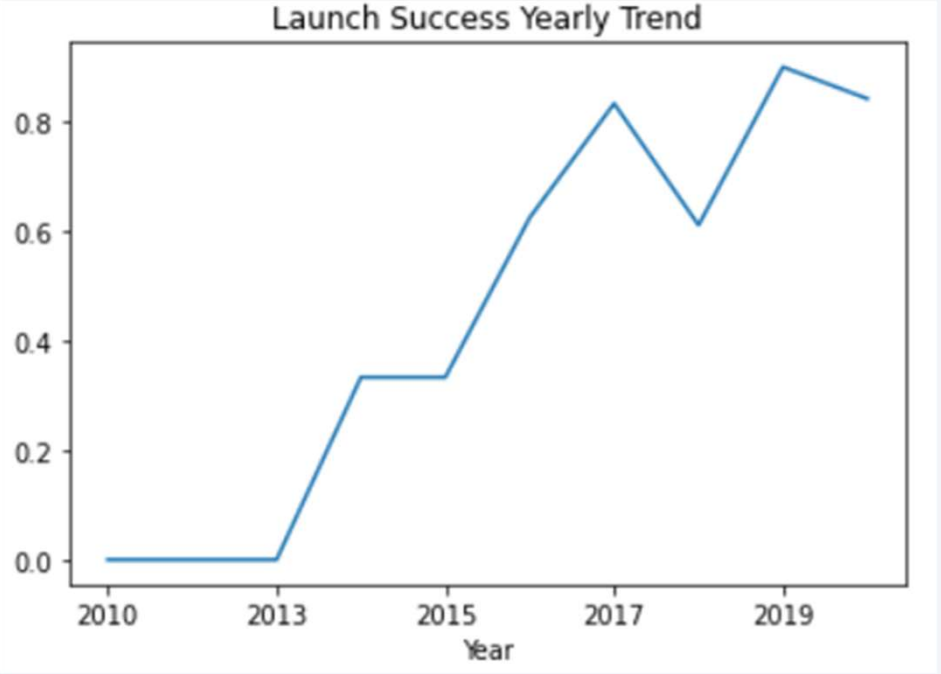
Payload vs. Orbit Type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits



Launch Success Yearly Trend

- success rate since 2013 kept increasing till 2020



All Launch Site Names

- Unique launch sites

-> There are 4 launch sites

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`

-> 2 launch sites start with ‘CCA’

launch_site
CCAFS LC-40
CCAFS SLC-40

Total Payload Mass

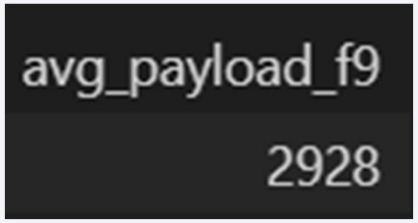
- Calculate the total payload carried by boosters from NASA

-> Most of the payload > 2000

payload_mass__kg_	customer
500	NASA (CRS)
677	NASA (CRS)
2296	NASA (CRS)
2216	NASA (CRS)
2395	NASA (CRS)
1898	NASA (CRS)
1952	NASA (CRS)
3136	NASA (CRS)
2257	NASA (CRS)
2490	NASA (CRS)
2708	NASA (CRS)
3310	NASA (CRS)
2205	NASA (CRS)
2647	NASA (CRS)
2697	NASA (CRS)
2500	NASA (CRS)
2495	NASA (CRS)
2268	NASA (CRS)
1977	NASA (CRS)
2972	NASA (CRS)

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1



the average payload mass carried by booster version F9 v1.1 = 2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

-> the first successful landing outcome happened at 2010-06-04

first_success
2010-06-04

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

-> 4 data fits the condition

launch_site	booster_version	payload_mass__kg_	landing__outcome
CCAFS LC-40	F9 FT B1022	4696	Success (drone ship)
CCAFS LC-40	F9 FT B1026	4600	Success (drone ship)
KSC LC-39A	F9 FT B1021.2	5300	Success (drone ship)
KSC LC-39A	F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

-> there are 100 successful outcomes and only 1 failure

sucess	failure
100	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

-> list part of the data since the limit of space

-> the maximum payload mass mostly > 2000

booster_version	max_payload
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092
F9 B4 B1045.1	362
F9 B4 B1045.2	2697
F9 B5 B1046.1	3600
F9 B5 B1046.2	5800
F9 B5 B1046.3	4000

2015 Launch Records

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

-> launch site CCAFS LC-40 has 2 events of failed landing_outcomes in drone ship

launch_site	booster_version	landing__outcome
CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
CCAFS LC-40	F9 v1.1 B1015	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, in descending order

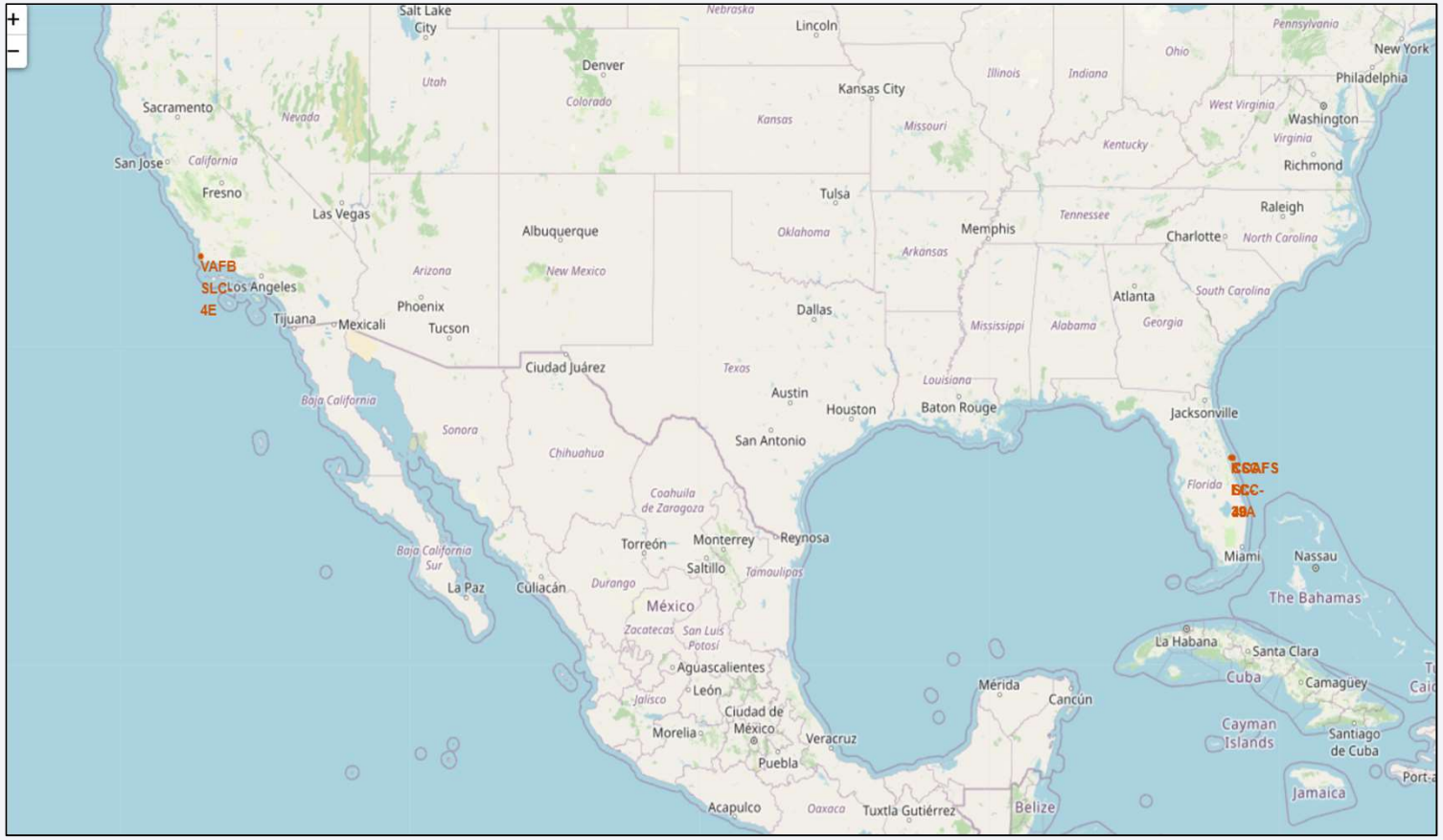
- >the landing outcome that has the highest

  occurrence is No attempt

landing_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 4
Launch Sites
Proximities Analysis

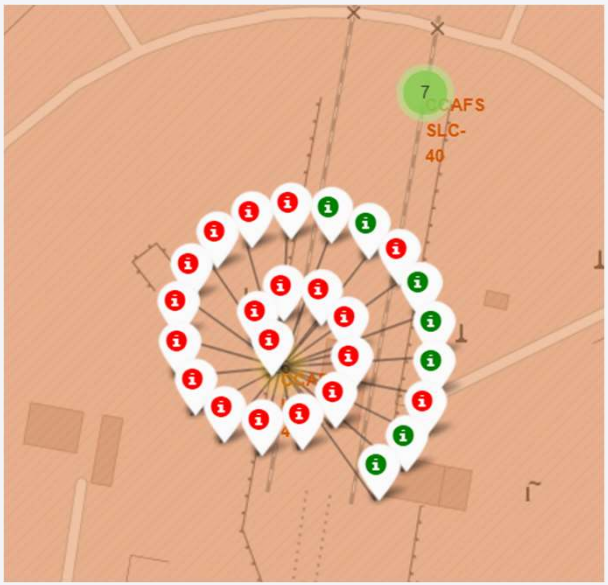
<SpaceX launch sites' location>

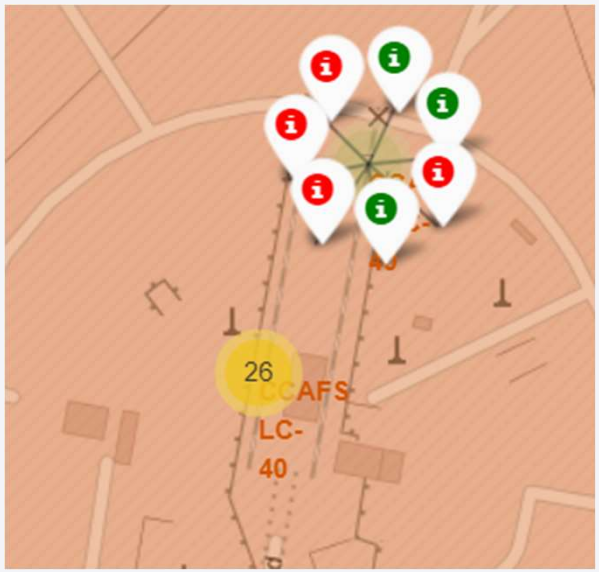


- All launch sites are near to the sea
- Besides VAFB SLC 4E, rest of the 3 launch sites are close

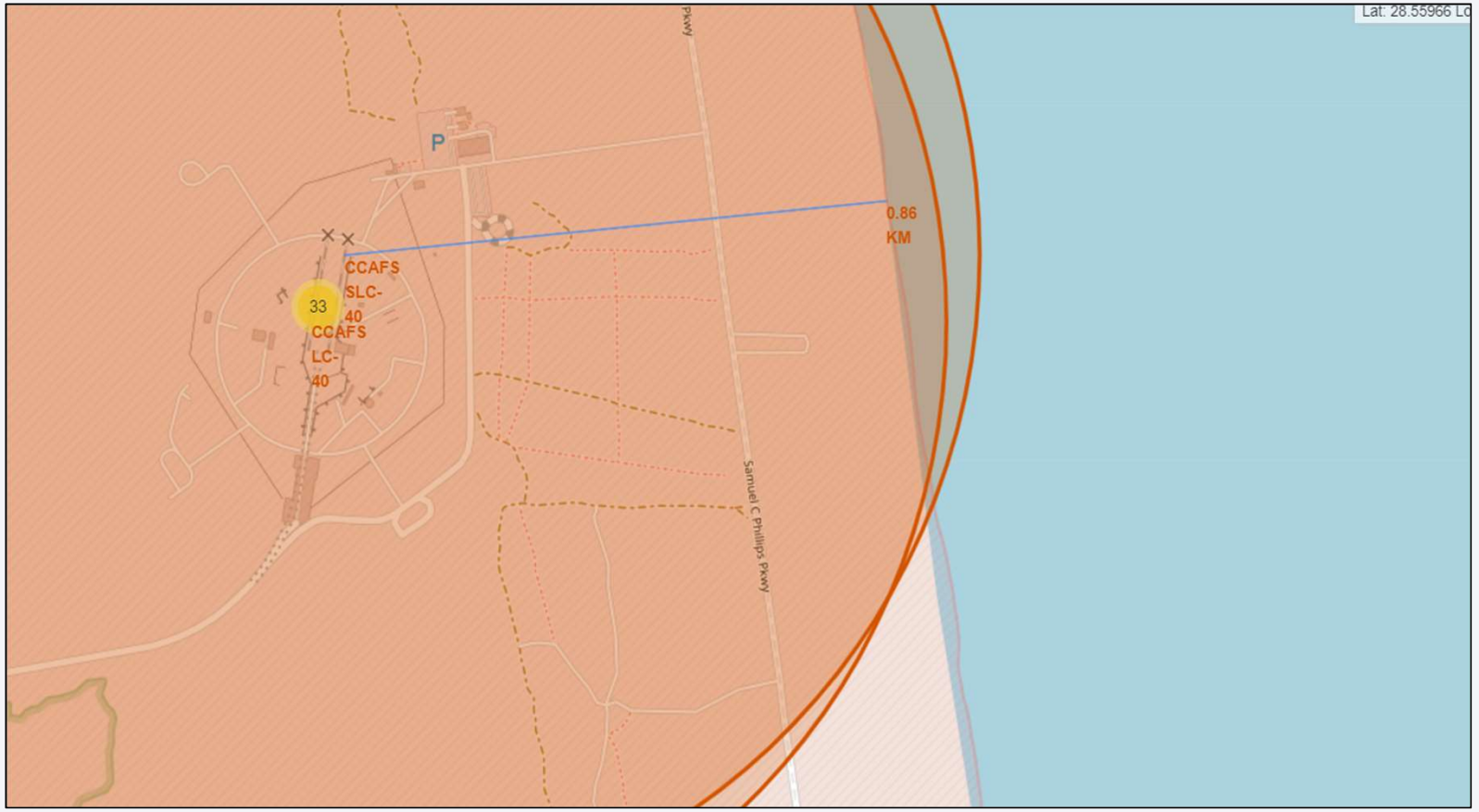
<Folium Map Screenshot 2>

- CCAFS SLC-40 has more failure than success outcomes
- It seems like CCAFS LC-40 has higher success rate

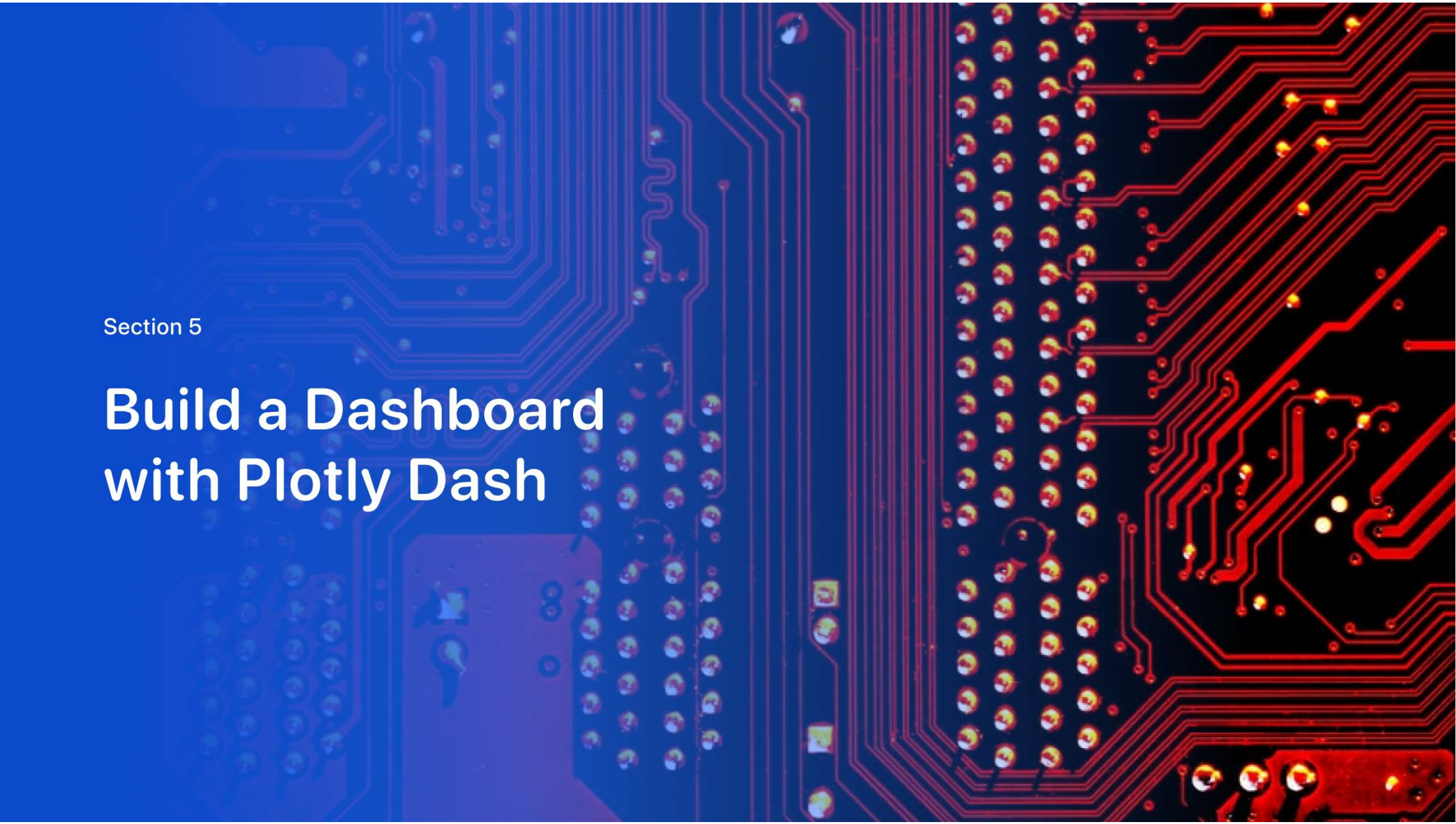




<Folium Map Screenshot 3>

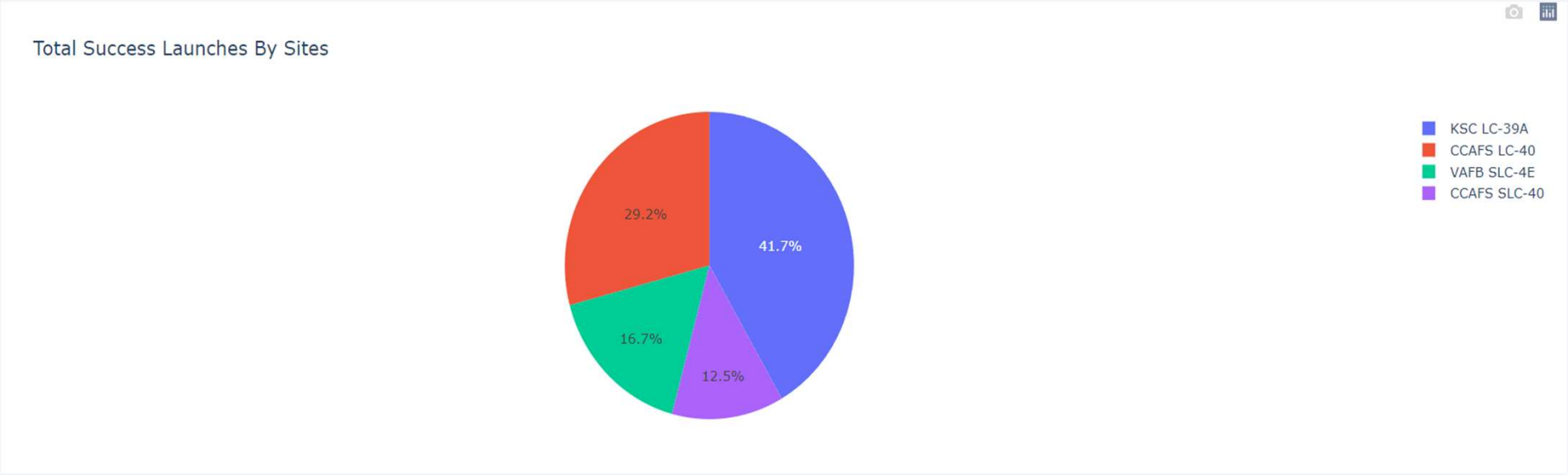


- CCAFS SLC-40 is approximately 0.86km away from the closest coastline


Section 5
Build a Dashboard with Plotly Dash

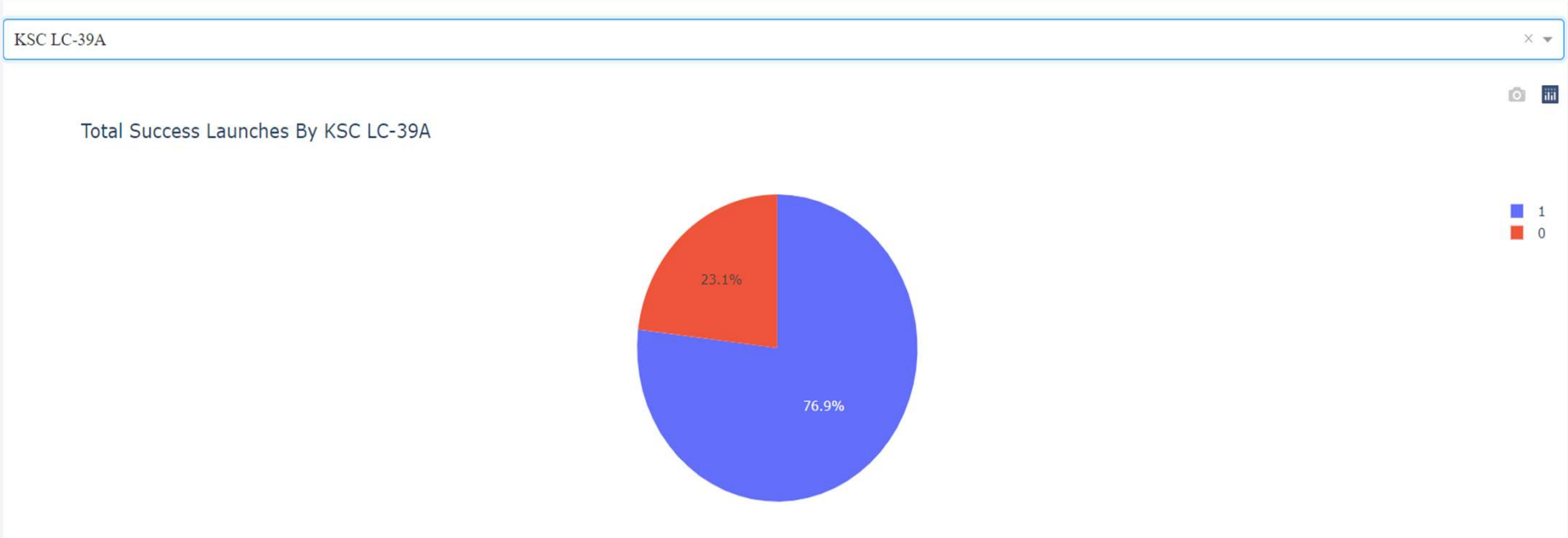
<Total Success Launches By Sites>

- The rank of the launch success
  - KSC LC 39A (class=10) > CCAFS LC-40 (class=7)
    > VAFB SLC-4E (class=4) > CCAFS SLC-40 (class=3)



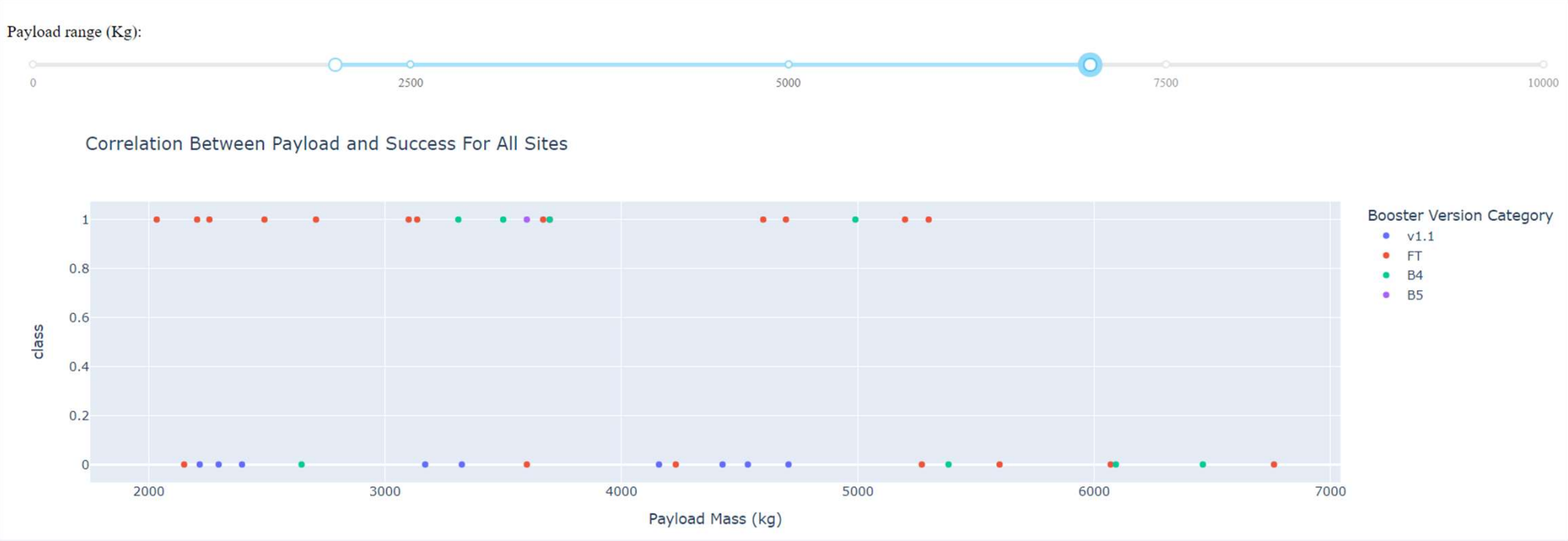
<Launch success ratio of KSC LC-39A>

- The launch site KSC LC 39A has the highest launch success ratio



<Correlation Between Payload & Success>

- Booster version FT has the most success when payload range between 2700-6900

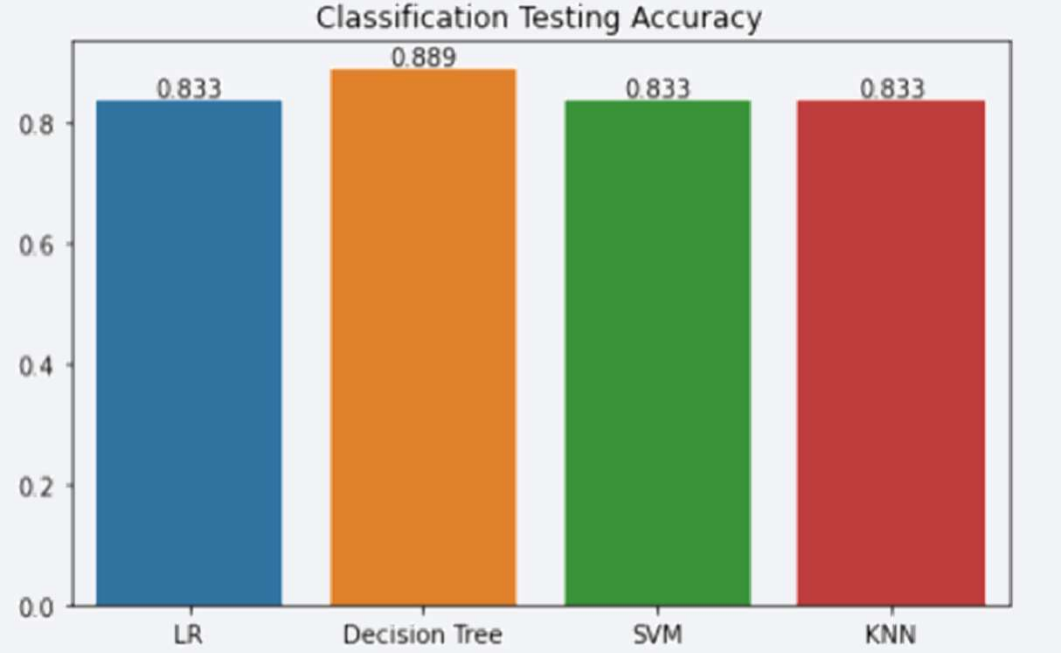


Section 6

Predictive Analysis (Classification)

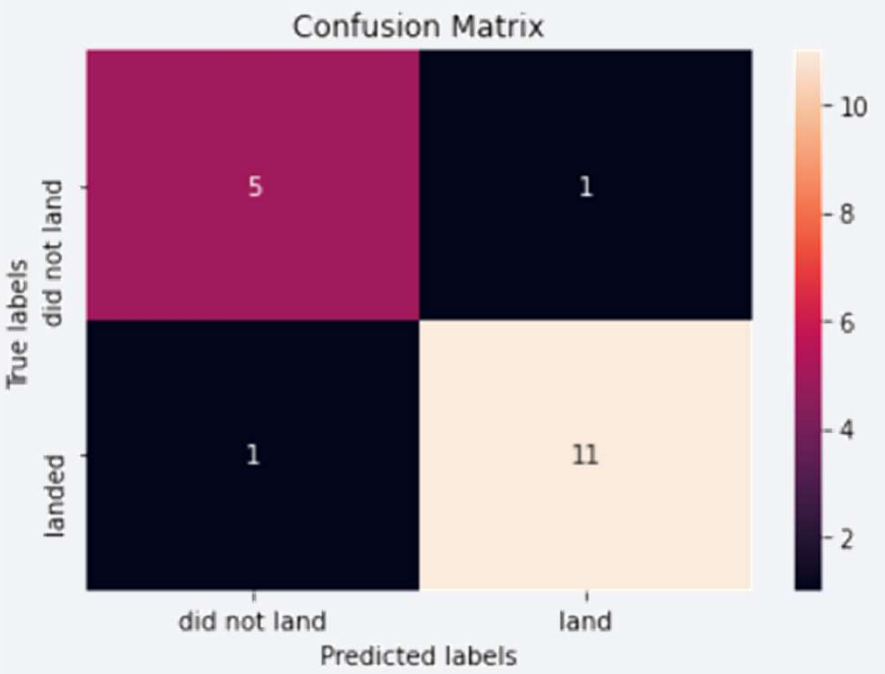
Classification Accuracy

- Decision Tree model has the highest accuracy on testing dataset
- The performance of the models is overall similar



Confusion Matrix

- The higher the diagonal values of the confusion matrix the better
- The diagonal values in the plot of decision tree model are relatively high -> more accurate



Conclusions

- Some variables such as flight number and payload has positive relationship with the landing outcome
- Each launch site or booster version correspond to a observable pattern/feature
- Overall the success rate increase by time
- Basic classification model could have good performance on landing outcome success prediction

Appendix

- All the content of this projects is at my github
- https://github.com/elvish7/Coursera-applied-data-science-capstone

Thank you!