Winning Space Race with Data Science

Sang Kha
2021-11-07

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all 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, much of the savings is because SpaceX can reuse the first stage.
- Problems you want to find answers: determine the cost of a launch, according to the result of the first stage. This information can be used for our company, SpaceY, wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

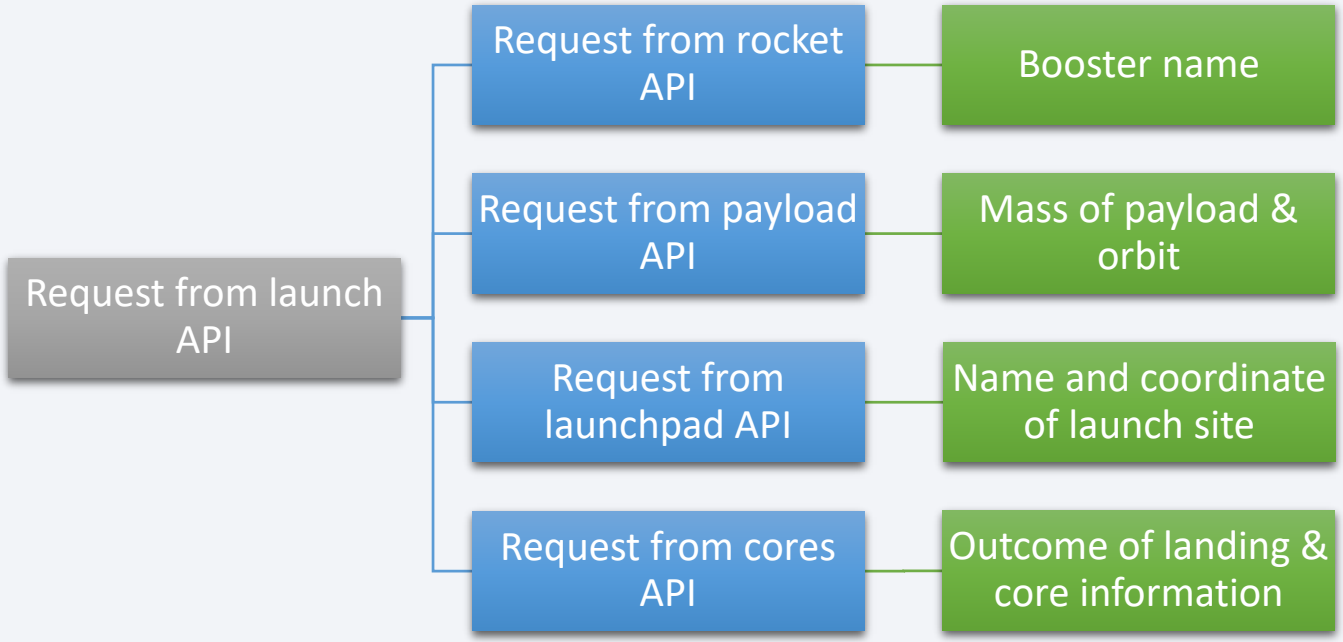
Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

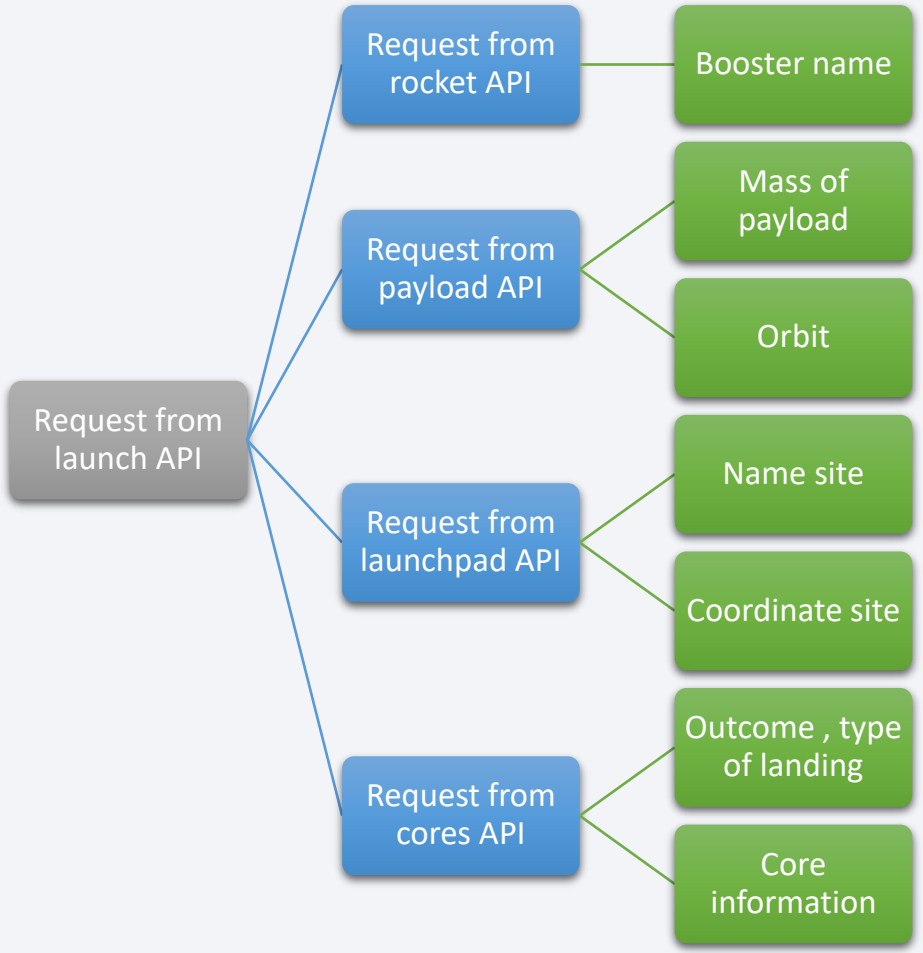
Data Collection

- Data was collected from SpaceX API: https://api.spacexdata.com/v4/, using 'request' and 'BeautifulSoup' library in Python
- Data collection process:



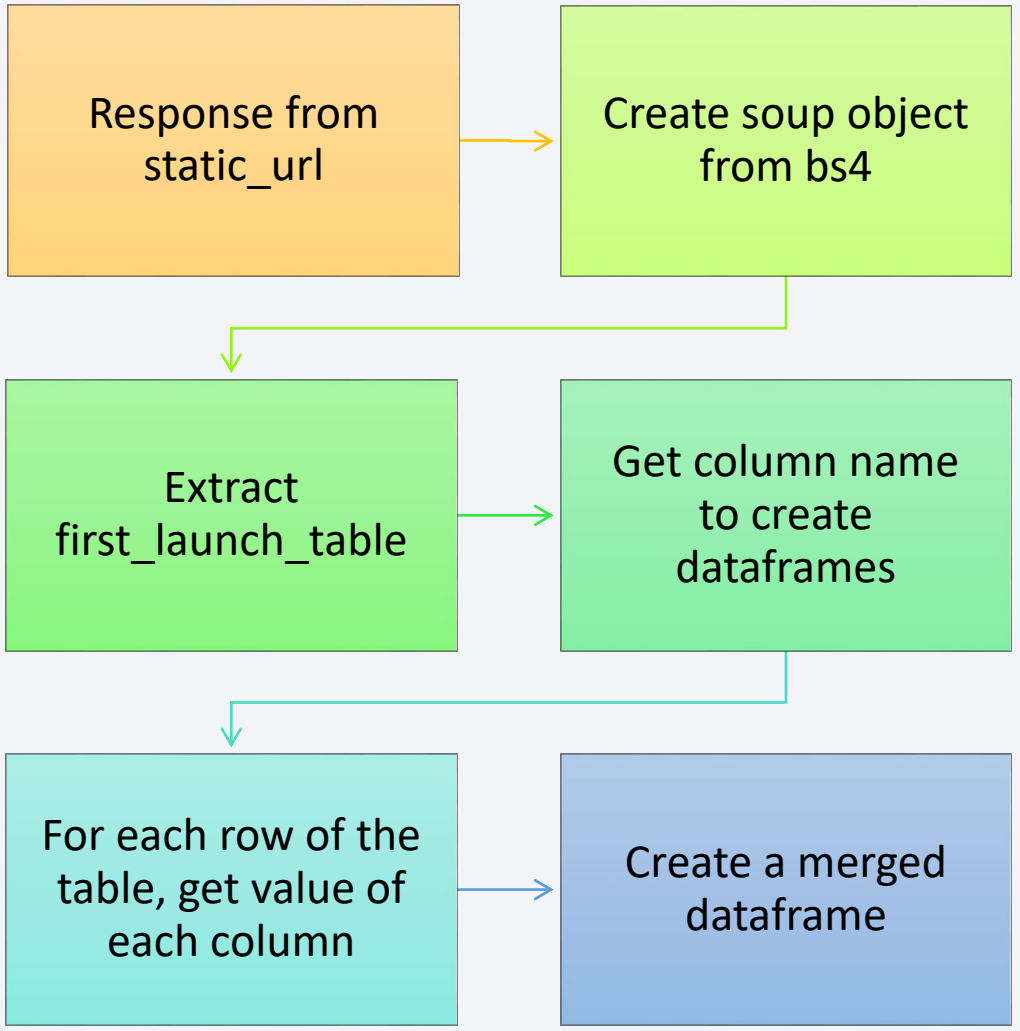
Data Collection – SpaceX API

- First, we request each API and store in different data frames. Then, we merged them into one after create a dictionary (call launch_dict).
- GitHub URL: https://github.com/khasang12-khmt/IBM_Capstone/blob/master/data_collection.ipynb



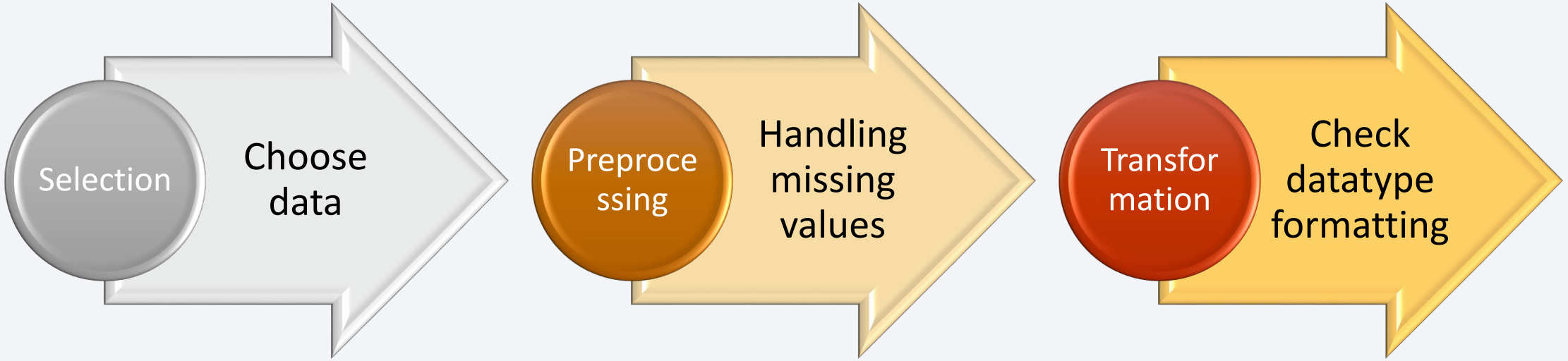
Data Collection - Scraping

- Web scraping process using BeautifulSoup
- GitHub URL: https://github.com/khasang12-khmt/IBM_Capstone/blob/master/data_scraping.ipynb



Data Wrangling

- Describe how data were processed:
- Missing values: find if there is a NA value in each column
- Data type: find if data types for every column are correct
- GitHub URL: https://github.com/khasang12-khmt/IBM_Capstone/blob/master/data_wrangling.ipynb



EDA with Data Visualization

- Type of plots used in the project:

- Scatterplot: to see relationship between two numerical variables
- Line chart to visualize the trend (time event)
- Bar chart to see relationship between one numerical and one categorical variable

- Github URL: https://github.com/khasang12-khmt/IBM_Capstone/blob/master/eda_dataviz.ipynb

EDA with SQL

- SQL queries performed:

- *Display unique launch sites in the space mission*
- *Display 5 records where launch sites begin with the string 'CCA'*
- *Display the total payload mass carried by boosters launched by NASA (CRS)*
- *Display average payload mass carried by booster version F9 v1.1*
- *List the date when the first successful landing outcome in ground pad was achieved.*

EDA with SQL

- SQL queries performed (cont'd):

- List the date when the first successful landing outcome in ground pad was achieved.

- List the total number of successful and failure mission outcomes

- List the names of the booster_versions which have carried the maximum payload mass.

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

- 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

- GitHub URL: https://github.com/khasang12-khmt/IBM_Capstone/blob/master/eda_sql.ipynb

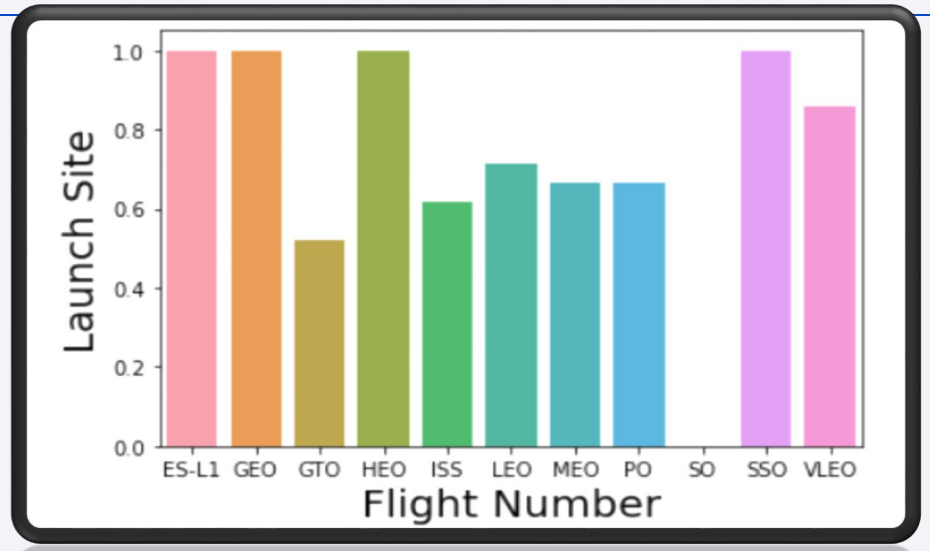
Build an Interactive Map with Folium

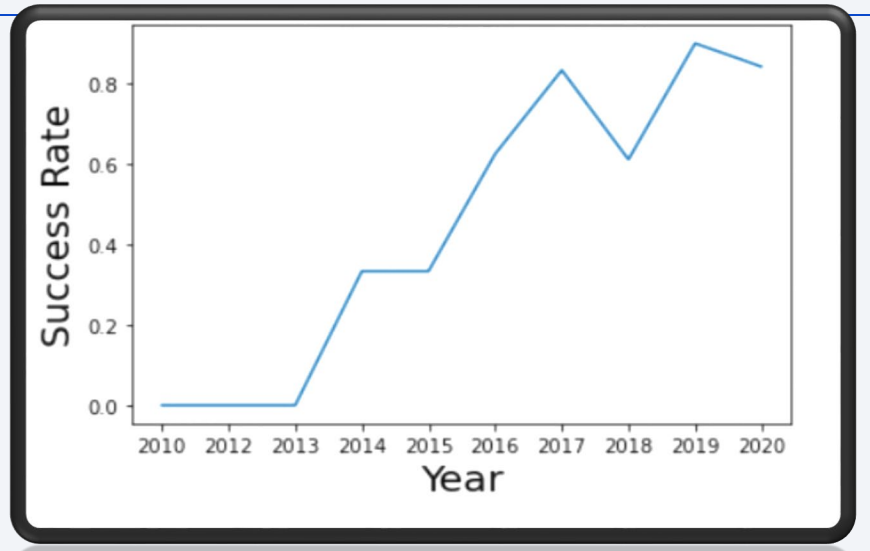
- Map objects added to a folium map:
- Circle: to show the area of launching site
- Marker: to show the label of launching site
- Marker_cluster: to show the information of duplicated launching records (number of launches, successed or failed)
- Mouse_position: get the coordinate (Lat, Long) for a mouse over on the map
- PolyLine: to show the distance between two places
- GitHub URL: https://github.com/khasang12-khmt/IBM_Capstone/blob/master/folium_launch_site_location.ipynb

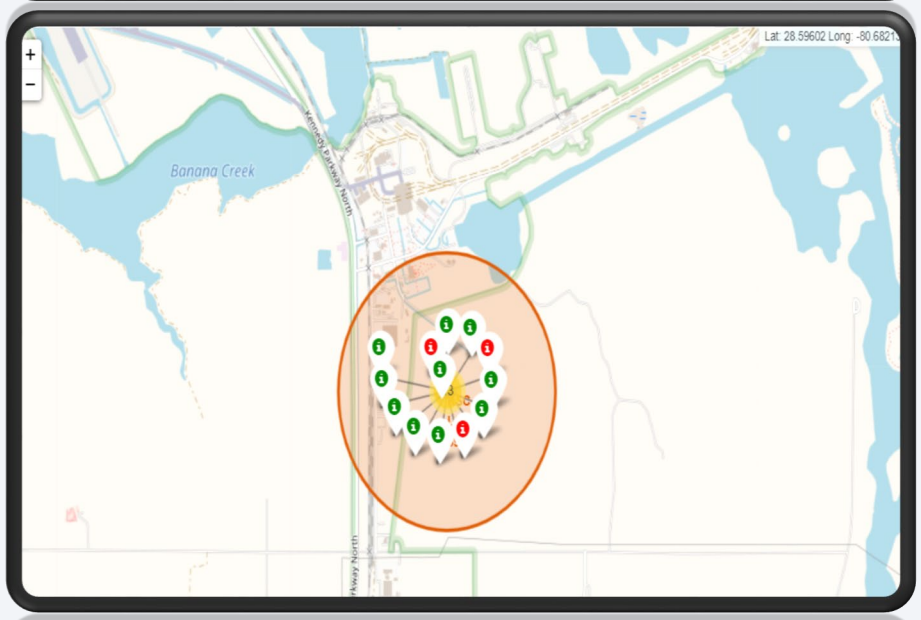
Predictive Analysis (Classification)

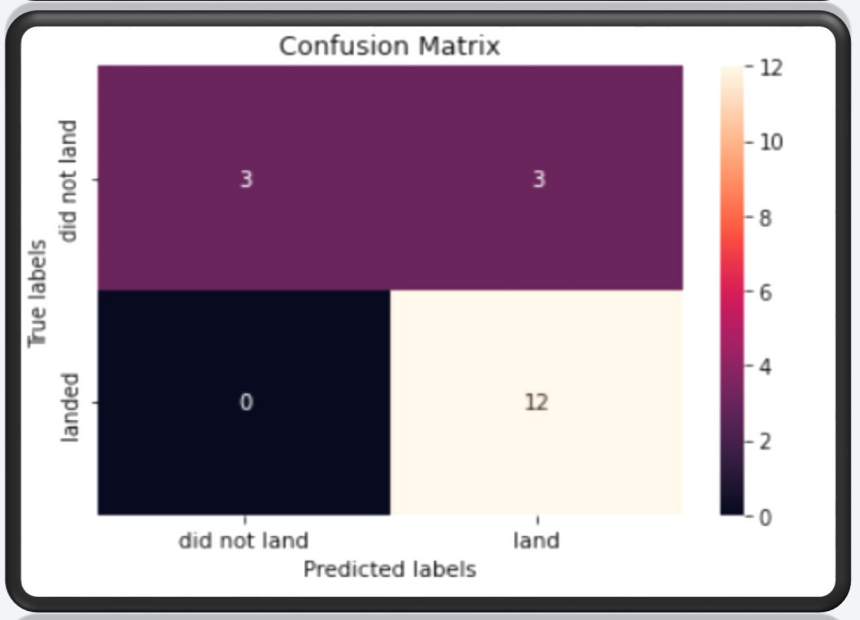
- Model development process is described as below
- GitHub URL: https://github.com/khasang12-khmt/IBM_Capstone/blob/master/Machine%20Learning%20Prediction.ipynb

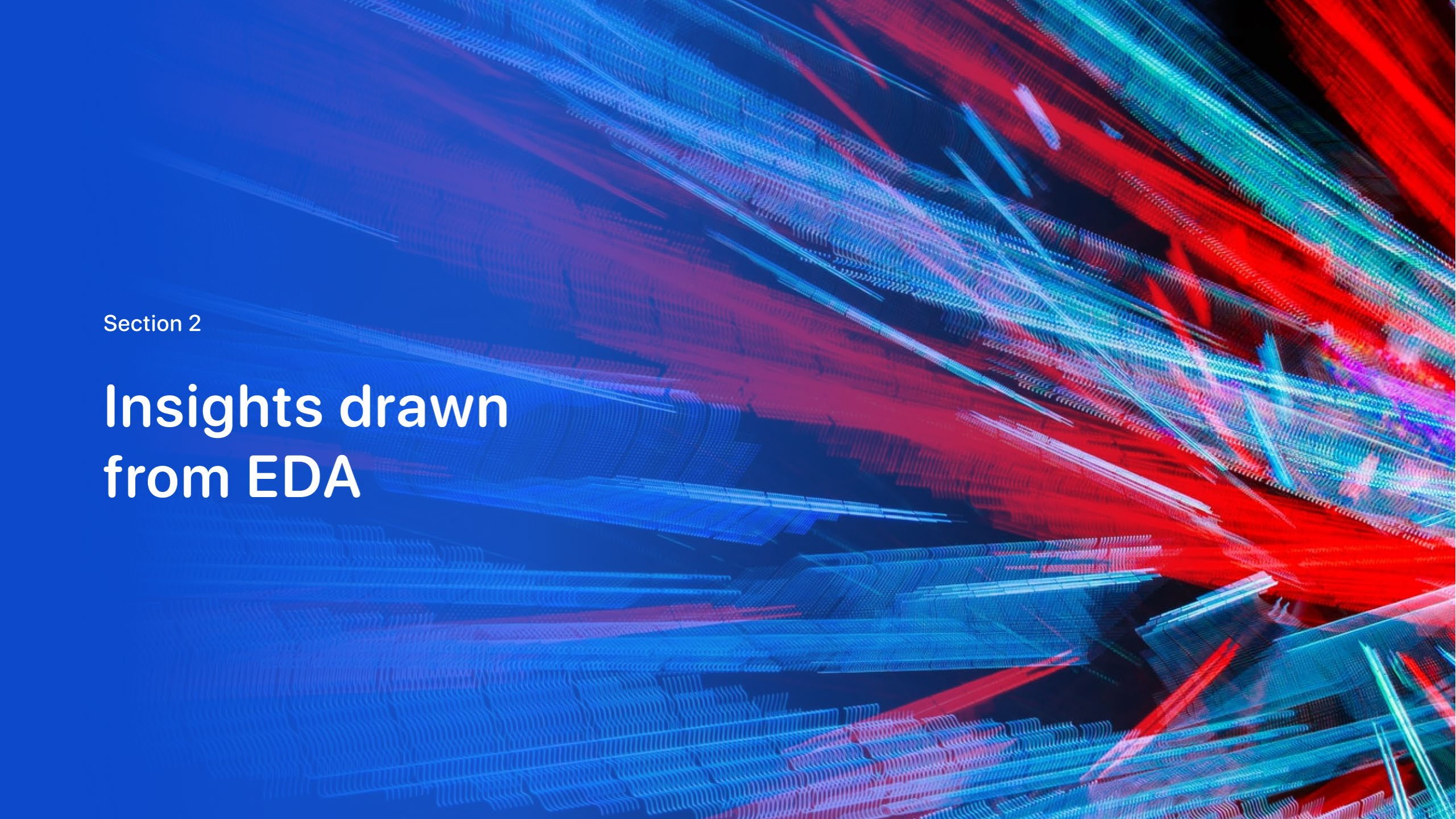
Results


Launch Site
1.0
0.8
0.6
0.4
0.2
0.0
ES-L1 GEO GTO HEO ISS LEO MEO PO SO SSO VLEO
Flight Number

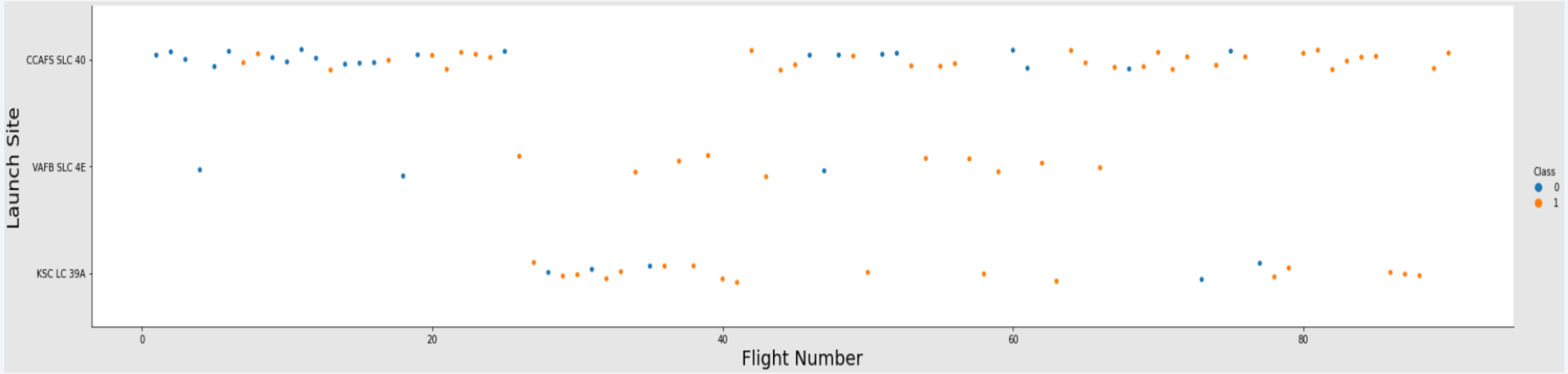

Success Rate
0.8
0.6
0.4
0.2
0.0
2010 2012 2013 2014 2015 2016 2017 2018 2019 2020
Year


Banana Creek


Confusion Matrix
True labels
did not land
landed
3
3
0
12
did not land
land
Predicted labels
12
10
8
6
4
2
0

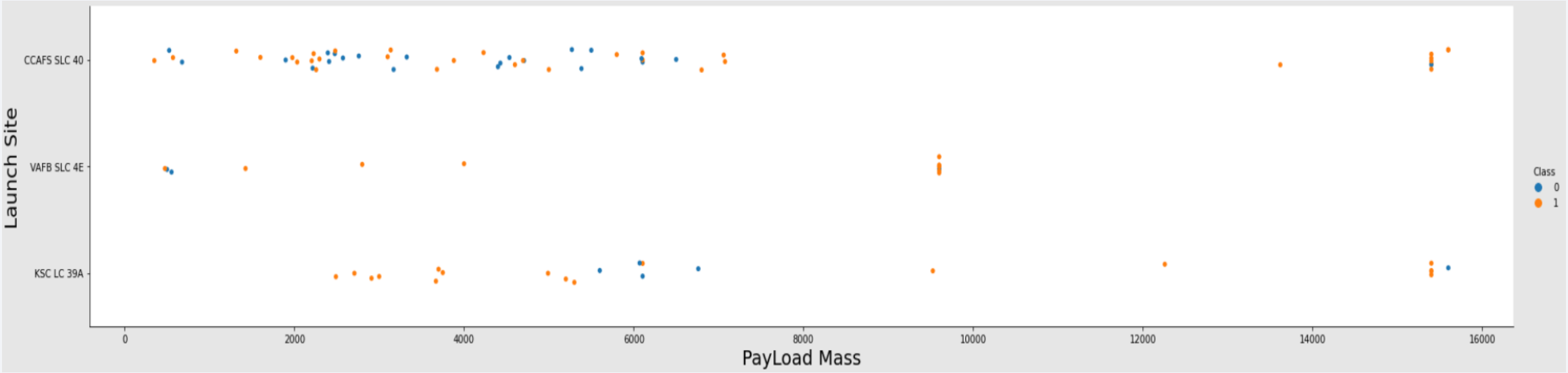

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- Explanation: As the number of flights increases, the first stage is more likely to land successfully. Different launch sites have different success rates.

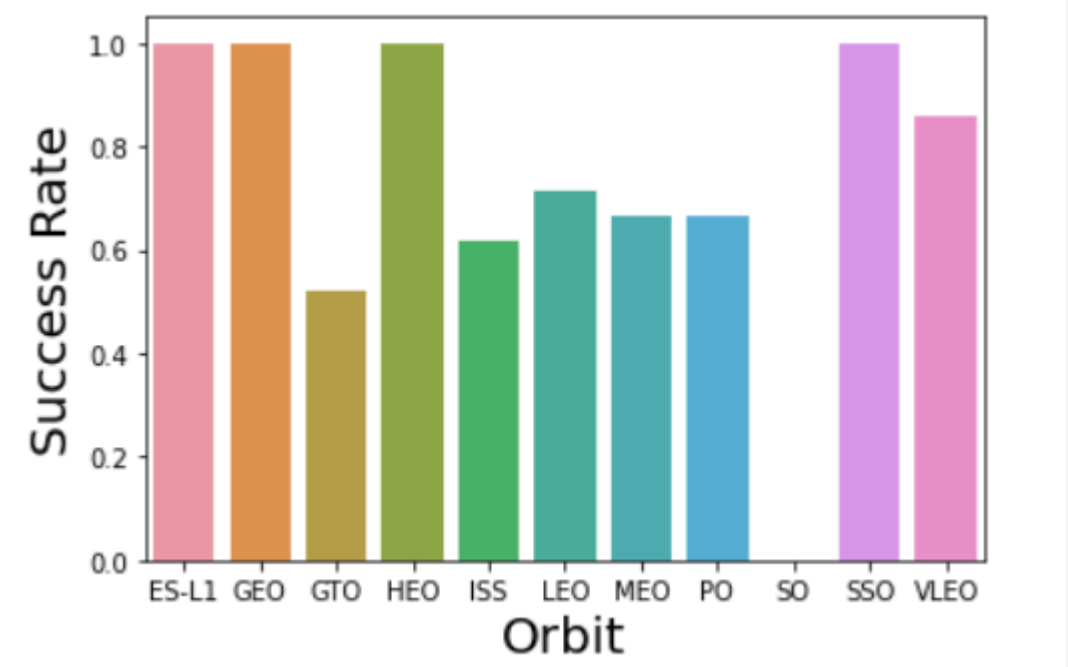
Payload vs. Launch Site



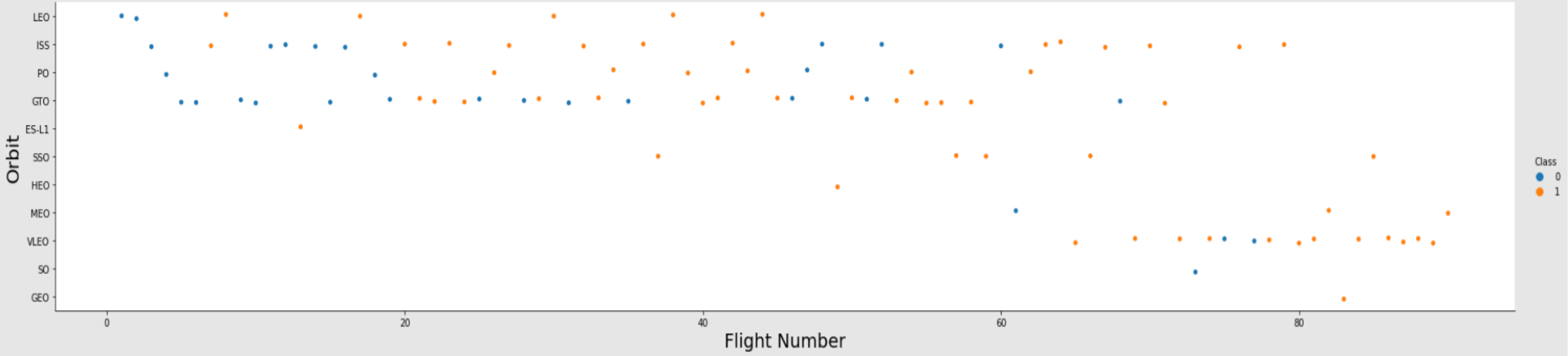
- Explanation: As the mass increases, the first stage is less likely to land successfully. Different launch sites have different success rates.

Success Rate vs. Orbit Type

- Explanations: Different orbits has different success rate. ES-L1, GEO, HEO, and SSO had success rate of 100%. Meanwhile, SO never succeeded.

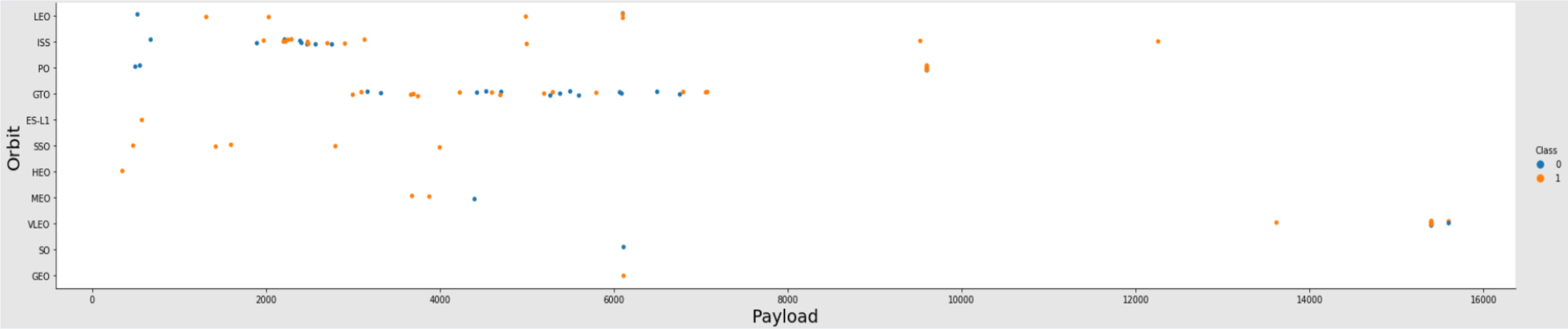


Flight Number vs. Orbit Type



- Explanation: As the number of flights increases, the first stage is more likely to land successfully. Different orbits have different success rates.

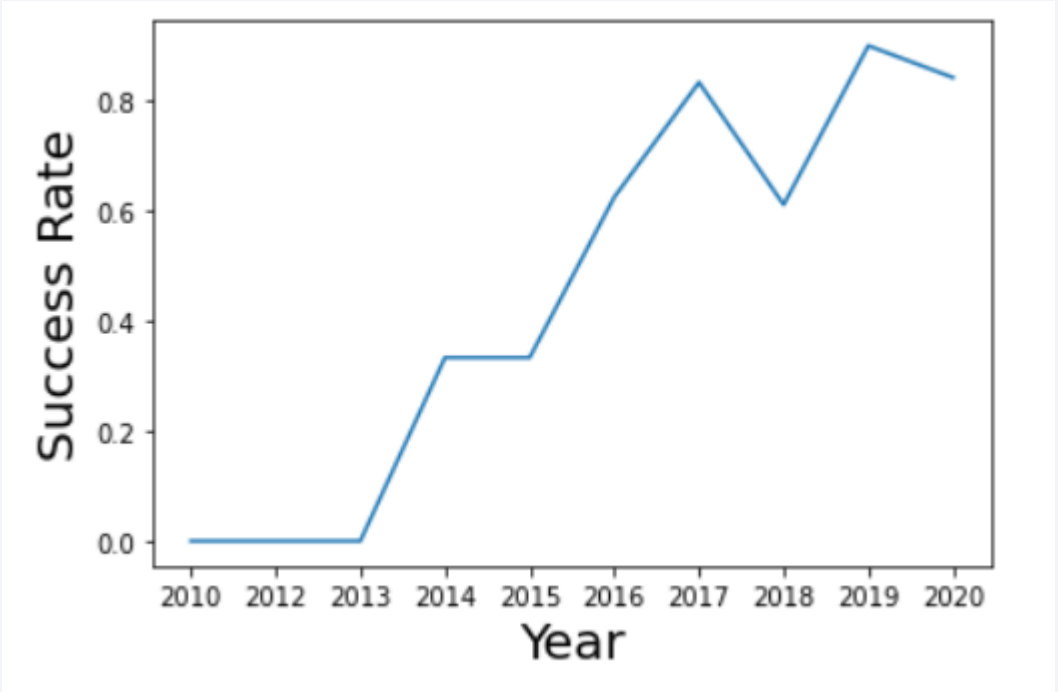
Payload vs. Orbit Type



- Explanations: As the mass increases, the first stage is more likely to failed. ES-LI, SSO, HEO are most successful orbit with small payload.

Launch Success Yearly Trend

- Explanations: From 2013, the success rate increases until 2020, despite some minor fluctuations.



All Launch Site Names

- Unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
- Short explanation: selection unique launch sites from the table

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

Launch Site Names Begin with 'CCA'

- Short explanation: get rows from the table which the first 3 characters equals 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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Short explanation: get sum from ‘payload_mass__kg_’ column where customer equals 'NASA (CRS)'

sum_payload_nasa_crs
45596

Average Payload Mass by F9 v1.1

- Short explanation: get average from ‘payload_mass__kg_’ column where ‘booster_version’ starts by 'F9 v1.1’ string

average_f9v11
2534

First Successful Ground Landing Date

- Short explanation: get min of date from ‘payload_mass__kg_’ column where ‘landing__outcome’ contains 'Success’ string

first_landing_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Short explanation: get booster_version from ‘payload_mass__kg_’ column where ‘payload_mass__kg_’ > 4000 and ‘payload_mass__kg_’ < 6000 and ‘landing__outcome’ like 'Success (drone ship)'

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

Total Number of Successful and Failure Mission Outcomes

- Short explanation: select ‘mission_outcome’, count of duplicated outcome from table.

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

Boosters Carried Maximum Payload

- Short explanation: select ‘booster_version’ and ‘payload_mass__kg_’ where its mass equals max of payload_mass

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

2015 Launch Records

- Short explanation: select ‘landing__outcome’, ’booster_version’ and ‘launch_site’ from table where ‘landing_outcome’ contains ‘failure’ string and year of date is 2015

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

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

- Short explanation: select ‘landing__outcome’ and count its duplicated values, where the date is between 2010-06-04 and 2017-03-20, and ranked from the most counted outcome.

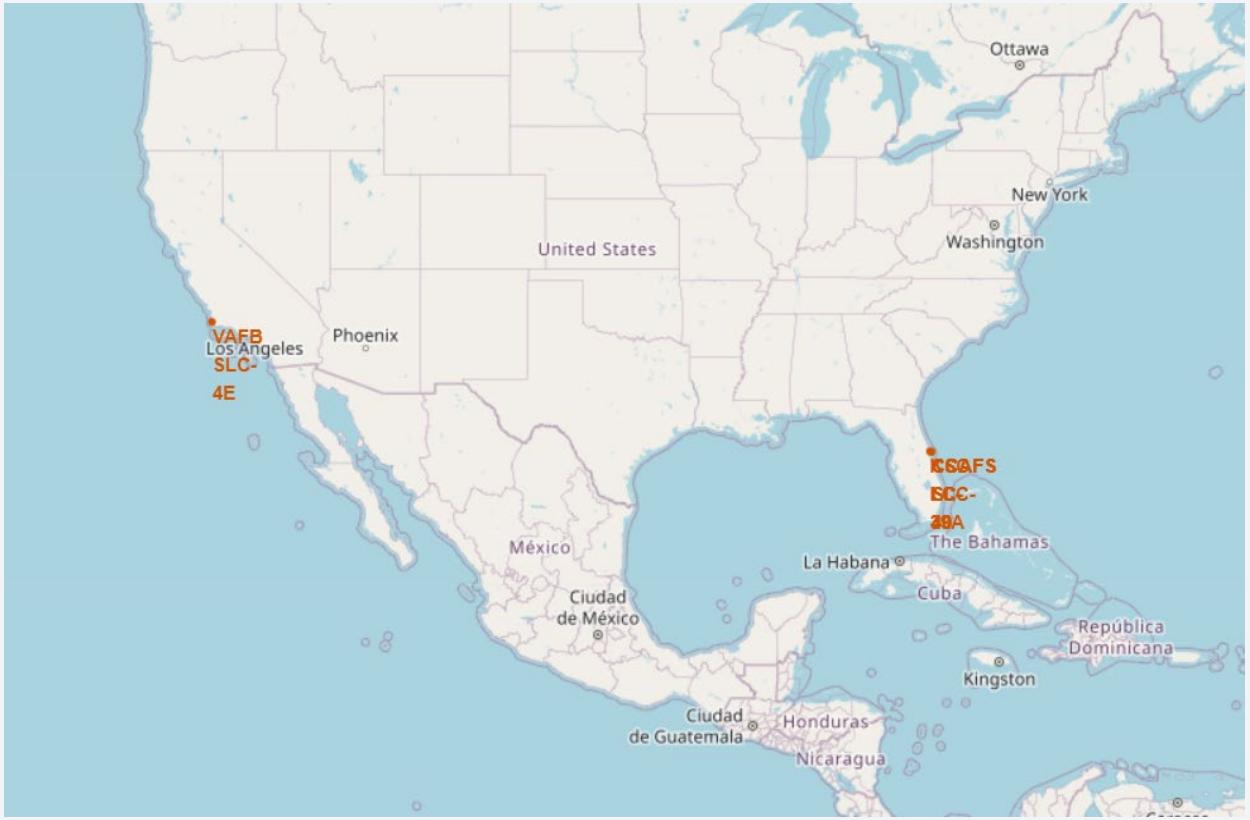
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

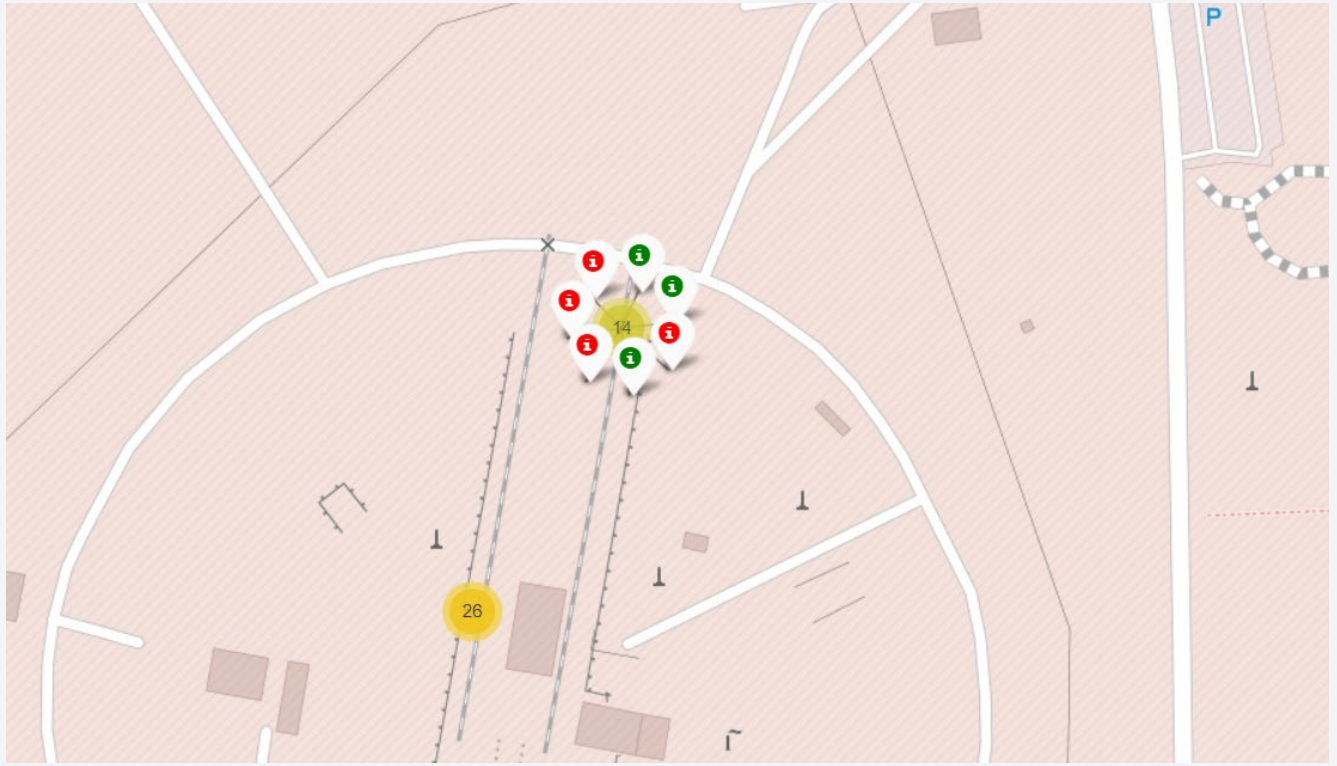
Mark all launch sites on a map

- All launch sites concentrated at two places only



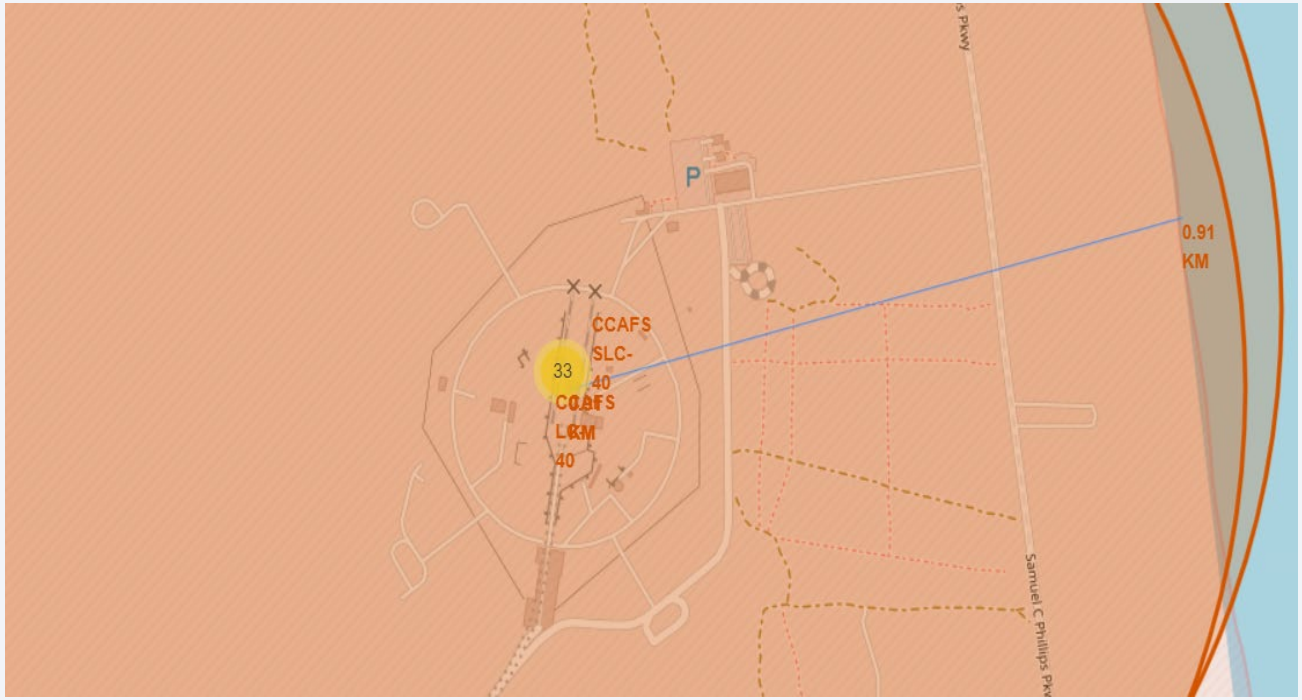
Mark the success/failed launches for each site on the map

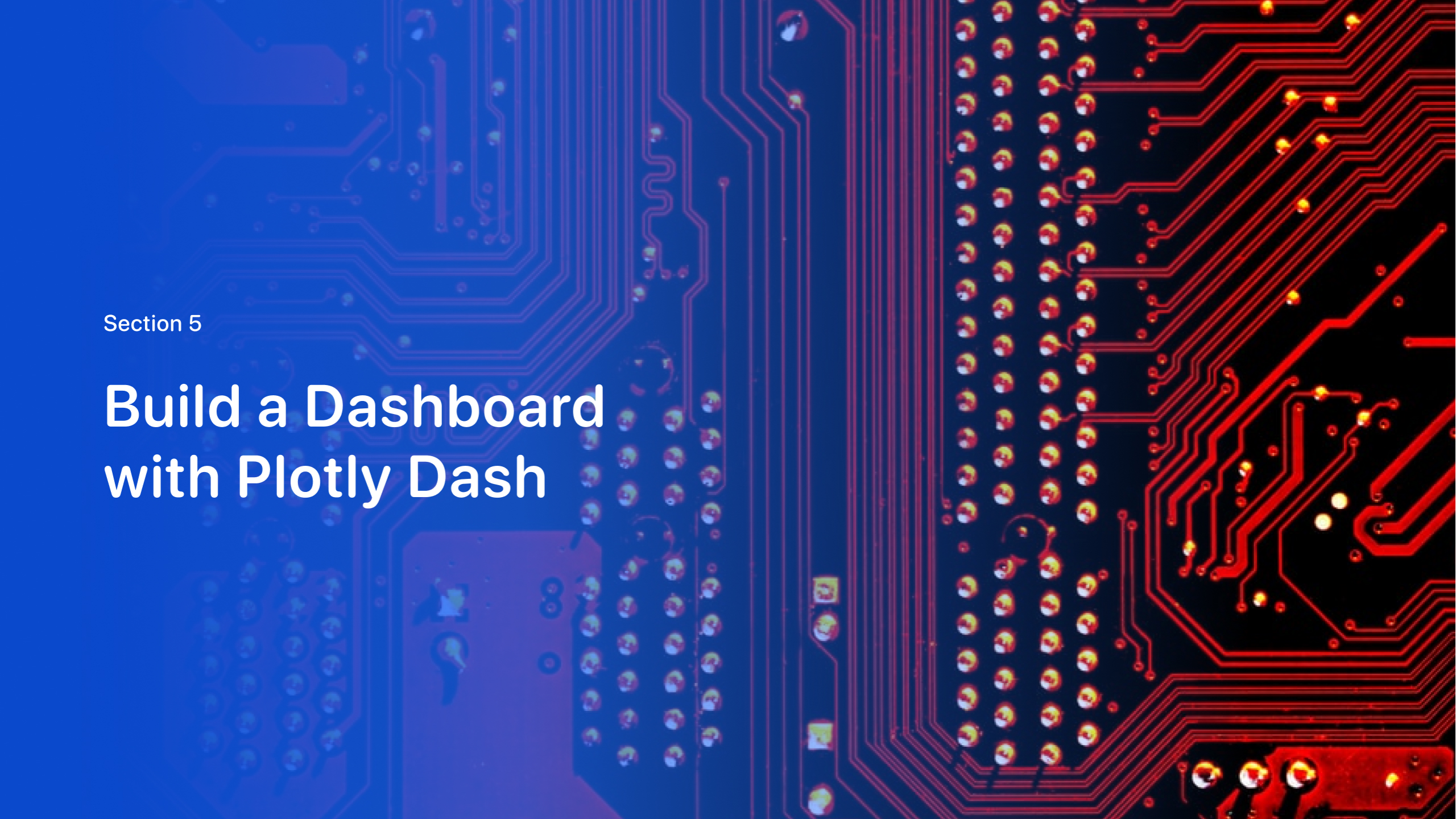
- Green icons represent success launches, while red one represent failed records.
- In the screenshot, this place has 7 launches, and three of them are success.



Calculate the distances between a launch site to its proximities

- The distance from CCAFS LC-40 launch site to nearest coastline is approximately 0.91km





Section 5

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

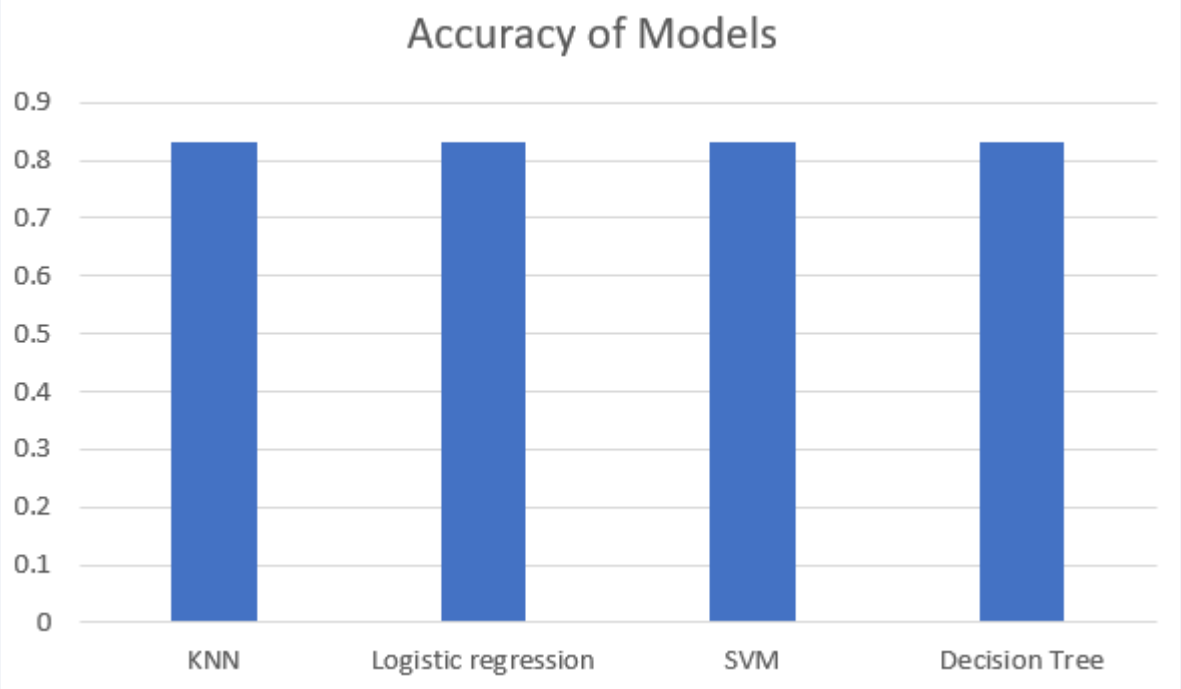
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 6

Predictive Analysis (Classification)

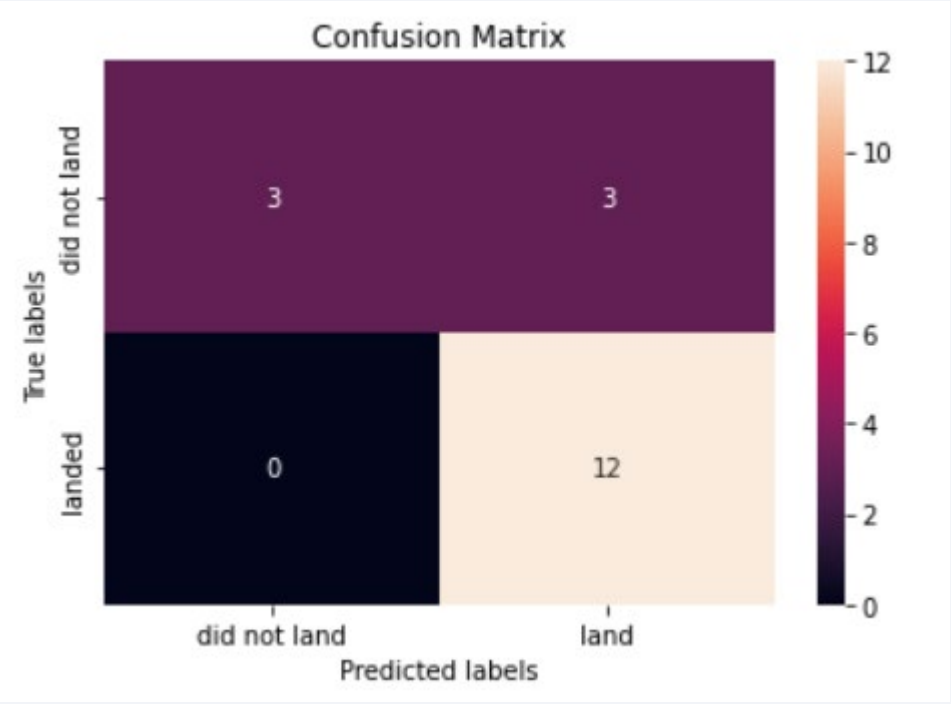
Classification Accuracy

- Practically all these algorithms give the same result



Confusion Matrix

- Using confusion matrix, we can see that the model can distinguish between different classes with high accuracy. The only problem is with false positives (predicted landed but in fact didn't)



Conclusions

- After the project, we can make a model with 83,33% chance that it will predict True. We can utilize the model, in relation with present data, to make a prediction that the first landing will be successful or failed.
- Furthermore, SpaceX can use the model to change their process system (place that take off, payload,…) in order to maximize the positive result.

Appendix

- IBM Data Scientist: Joseph Santarcangelo and his team
- Coursera Discussion Forum
- Stackoverflow
- Python, Pandas, Folium,.. Documentary on Google.

Thank you!