IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jingze Dai
August 1, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collections (with SpaceX API, web scraping, and data wrangling)
  - Data Analysis (Exploratory Data Analysis, with SQL)
  - Data Visualization (Visualize Data and Data Results)
  - Machine Learning Prediction
- Summary of all results
  - Successfully getting data from data collections
  - In the data analysis (EDA), successfully getting all desired results
  - Machine learning tools give a good prediction on given data

Introduction

- Project background and context
  - SpaceX is one of the top commercial space company. There is a difference between costs: SpaceX says a falcon 9's landing requires 62 million dollars, while others have a figure: up to 165 million dollars. The first figure is much lower is because, SpaceX reuses its launching and landing facilities. We can determine if
- Problems you want to find answers
  - What are differences between successful and unsuccessful landings? (In details)
  - What are overall trends of successful landing probability on different types of launchings?
  - What are relationships between landing results and different factors?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping
- Perform data wrangling
  - Delete unnecessary columns
  - Adding default values to missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use exploratory data analysis with SQL usages, as well as data visualization

Data Collection

- Describe how data sets were collected.
- Method 1: SpaceX API
  - https://api.spacexdata.com/v4/rockets/
- Method 2: Web scraping
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

- https://github.com/daijingz/Data-Science-Machine-Learning/blob/main/Coursera%20Capstone/Data%20Collection%20API.ipynb
- Part 1: Requesting data
  - spacex_url="https://api.spacexdata.com/v4/launches/past"
  - response = requests.get(spacex_url)
- Part 2: Data Normalization
- Part 3: Solving the missing-value problem
- Part 4: Extract columns by requirements
  - In this project there is an example: extract only “Falcon 9” columns

Data Collection - Scraping

- Part 1: Request the Falcon9 Launch Wiki page from its URL
- Part 2: Extract all column/variable names from the HTML table header
- Part 3: Create a data frame by parsing the launch HTML tables

```html
<tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordinated Universal Time">UTC</a
>)
</th>
<th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Falcon 9 first-stage boosters">Vers
ion,<br/>Booster</a> <sup class="reference" id="cite_ref-booster_11-0"><a href="#cite_note-booster-11">[b]</a></sup>
</th>
<th scope="col">Launch site
</th>
<th scope="col">Payload<sup class="reference" id="cite_ref-Dragon_12-0"><a href="#cite_note-Dragon-12">[c]</a></sup>
</th>
<th scope="col">Payload mass
</th>
<th scope="col">Orbit
</th>
<th scope="col">Customer
</th>
<th scope="col">Launch<br/>outcome
</th>
<th scope="col"><a href="/wiki/Falcon_9_first-stage_landing_tests" title="Falcon 9 first-stage landing tests">Booster<b
r/>landing</a>
</th></tr>
```

Successful Results:

https://github.com/daijingz/Data-Science-Machine-Learning/blob/main/Coursera%20Capstone/Data%20Collection%20API%20with%20Webscraping.ipynb

Data Wrangling

- 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.
- https://github.com/daijingz/Data-Science-Machine-Learning/blob/main/Coursera%20Capstone/Data%20Wrangling.ipynb

EDA with Data Visualization

- There are three kinds of charts (diagrams) used:
  - Scatter Graphs (e.g., flight number AND payload mass)
  - Bar Graph (e.g., successful rate AND orbits)
  - Line Graph (e.g., successful rate AND year)
- https://github.com/daijingz/Data-Science-Machine-Learning/blob/main/Coursera%20Capstone/Exploratory%20Data%20Analysis%20for%20Data%20Visualization.ipynb

EDA with SQL

- Completed tasks:
  - Display the names of the 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 acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Other tasks in the notebook below ->
- https://github.com/daijingz/Data-Science-Machine-Learning/blob/main/Coursera%20Capstone/Exploratory%20Data%20Analysis%20using%20SQL.ipynb

Build an Interactive Map with Folium

- The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.
- TASK 1: Mark all launch sites on a map
- TASK 2: Mark the success/failed launches for each site on the map
- TASK 3: Calculate the distances between a launch site to its proximities
- http://localhost:8888/notebooks/Desktop/Coursera%20IBM%20Data%20Science%20Capstone%20Projects/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- There are 10 visualized plots (diagrams) added, and from these plots, we can have a clearer view of relations between payloads and launch sites.
- https://github.com/daijingz/Data-Science-Machine-Learning/blob/main/Coursera%20Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

- Task 1: data preparation
- Task 2: Data standardization
- Task 3: Training
- Task 4: Testing
- Task 5: Comparison results
- https://github.com/daijingz/Data-Science-Machine-Learning/blob/main/Coursera%20Capstone/Machine%20Learning%20Prediction.ipynb

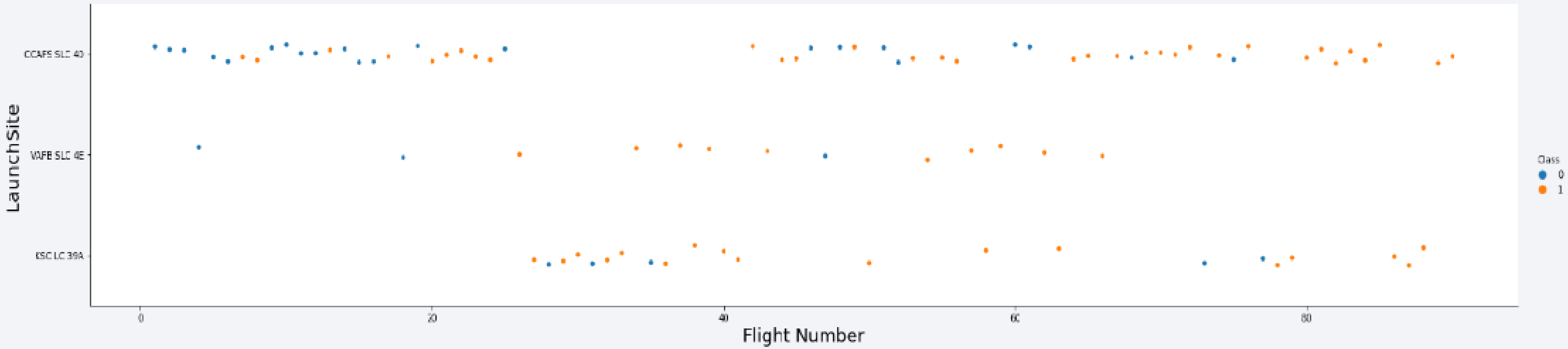
Results

- Exploratory data analysis results (Successful, please see my notebook)
- Interactive analytics demo in screenshots (Successful, please see my notebook)
- Predictive analysis results (Successful, please see my notebook)

Section 2

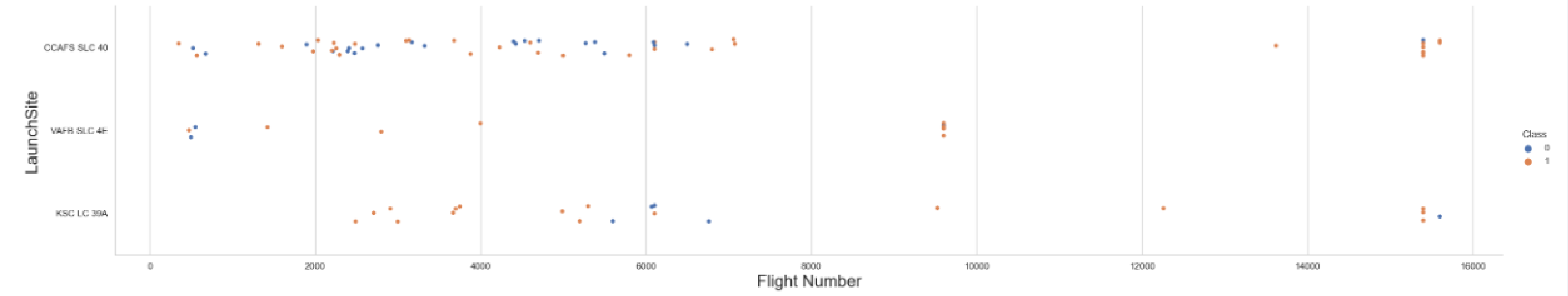
Insights drawn from EDA

Flight Number vs. Launch Site



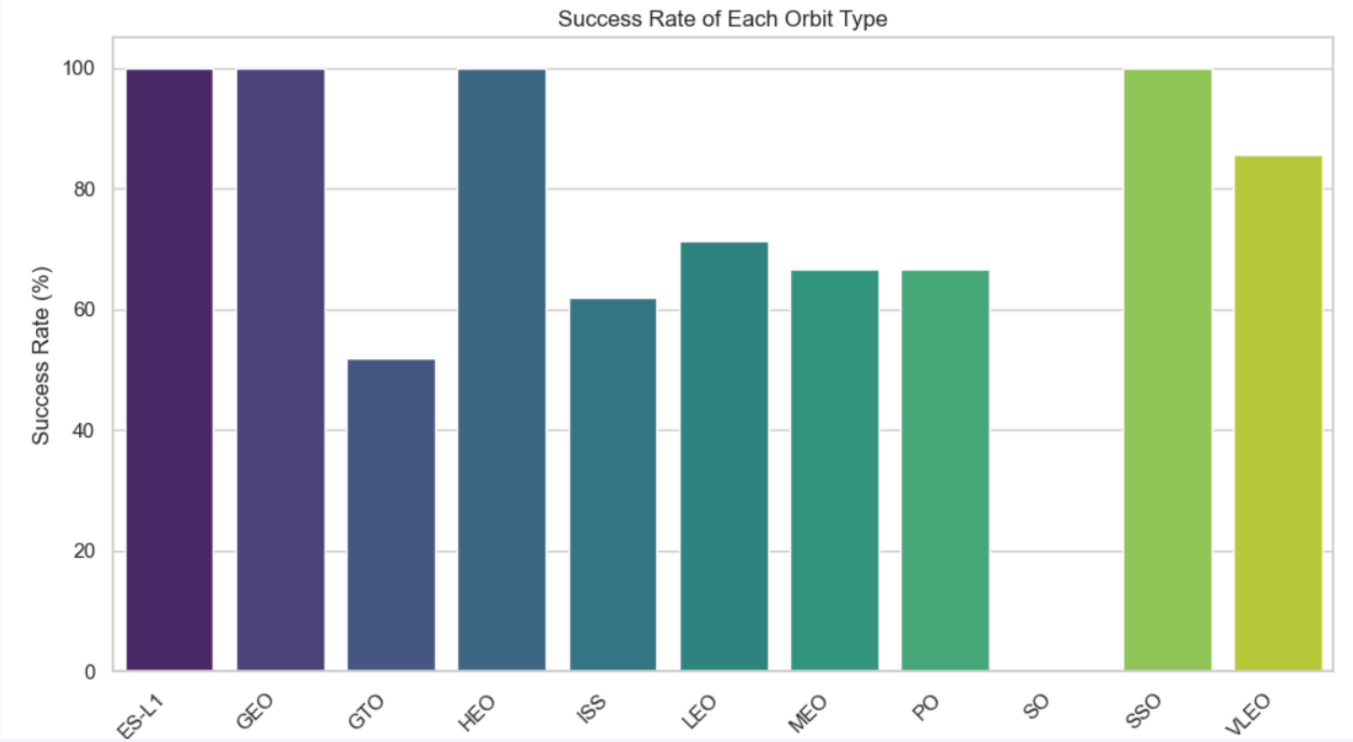
- By observation, we can see that the successful rate is increasing

Payload vs. Launch Site



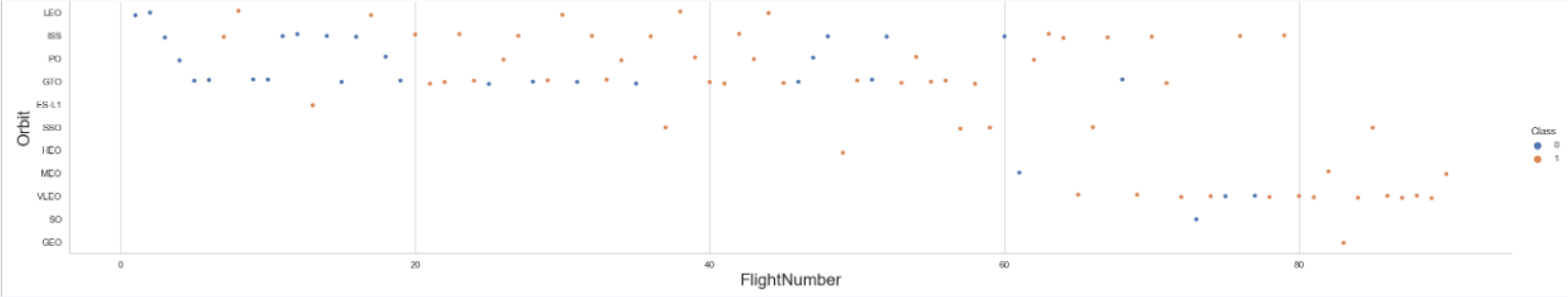
- As we can see, Most launches with payload mass under 7,000 kg are from any launch site (Here, these launches have a high success rate), but launches heavier than 12000 kg happens mostly only at CCAFS SLC 40 and KSC LC 39A

Success Rate vs. Orbit Type



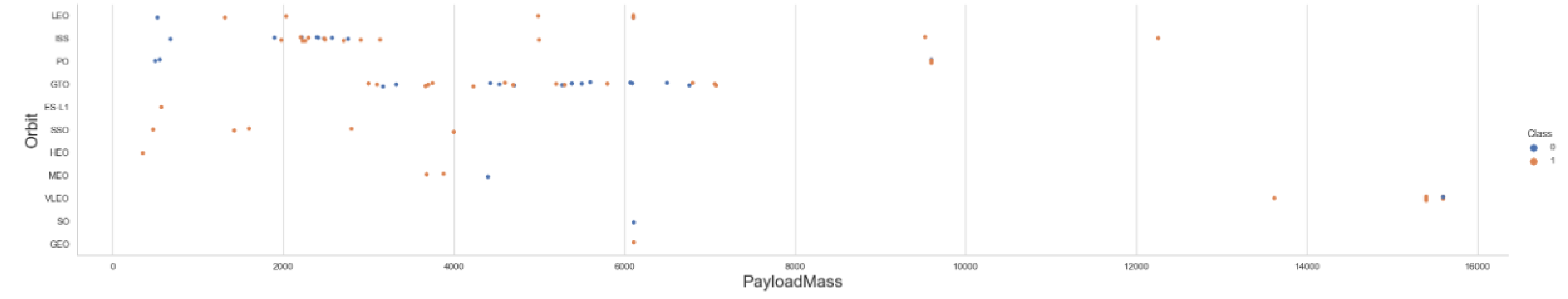
- From this diagram, seems like ES-L1, GEO, HEO, SSO have a perfect success rate.

Flight Number vs. Orbit Type



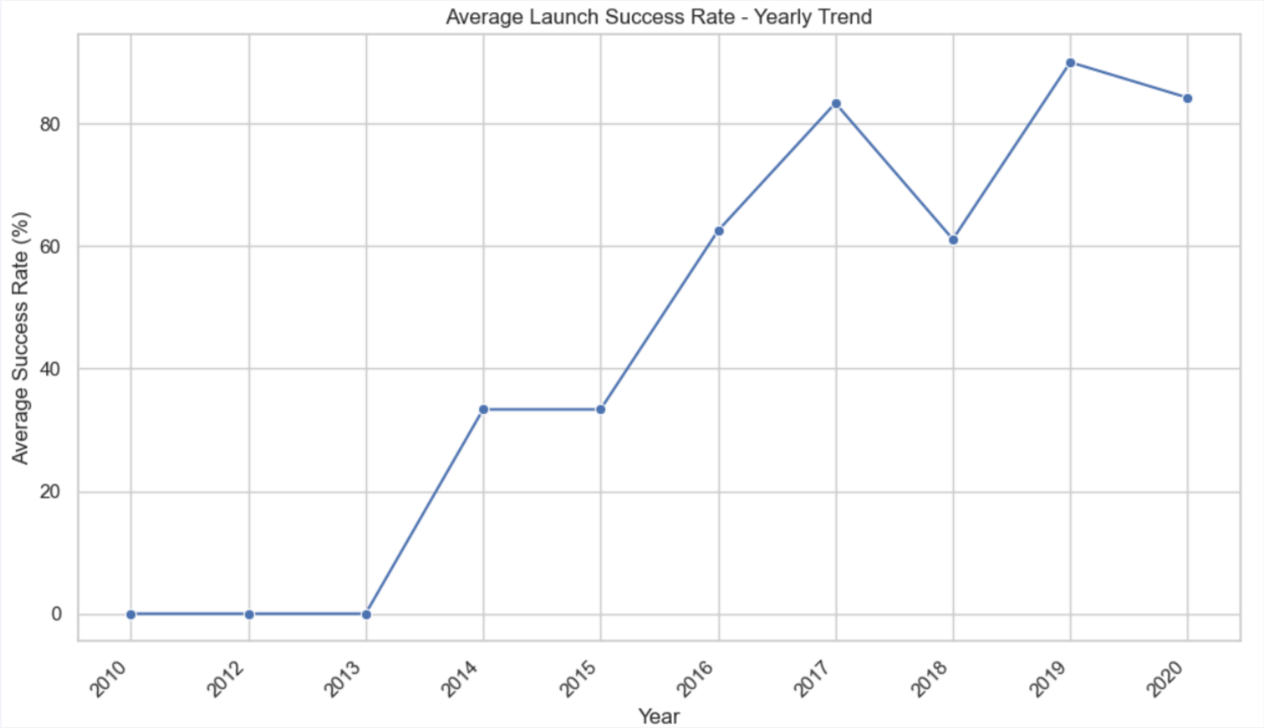
- Success rate increases over and over time

Payload vs. Orbit Type



- In conclusion, most successful landing have a lighter payload mass, so a heavier payload mass has a higher probability of failure

Launch Success Yearly Trend



- The overall success rate is increasing since 2013

All Launch Site Names

- There are four launch site names (shown in the screenshot below, not None)

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Result:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Result:

Total_Payload_Mass
45596.0

Average Payload Mass by F9 v1.1

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

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

First_Successful_Landing
01/08/2018

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
- Result:

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
- Result

Mission_Outcome	Total_Count
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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

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

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

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	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
- Result

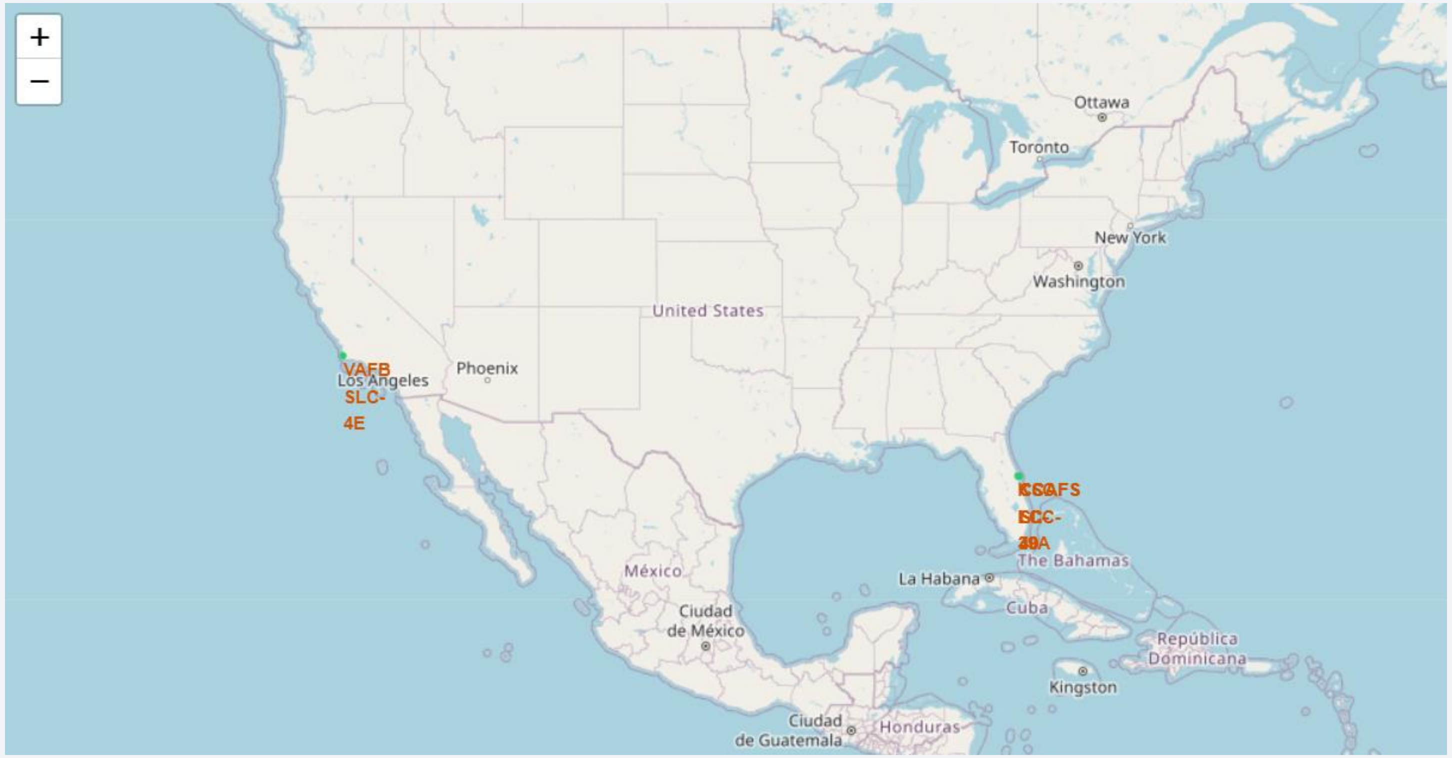
landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

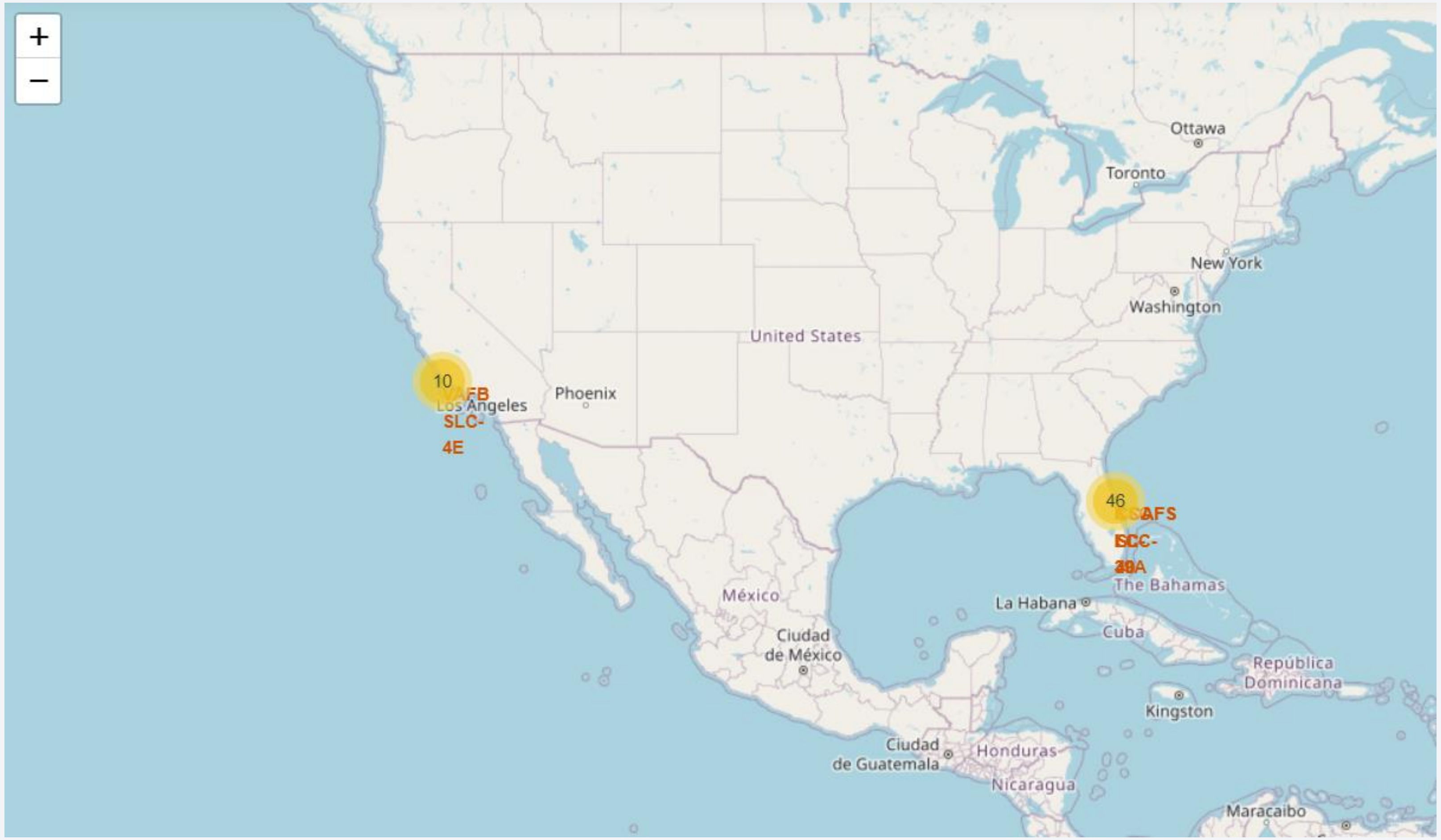
Launch Outcome by Sites

- Two areas of launch sites are near center of roads and railways, and they are both in the south area.



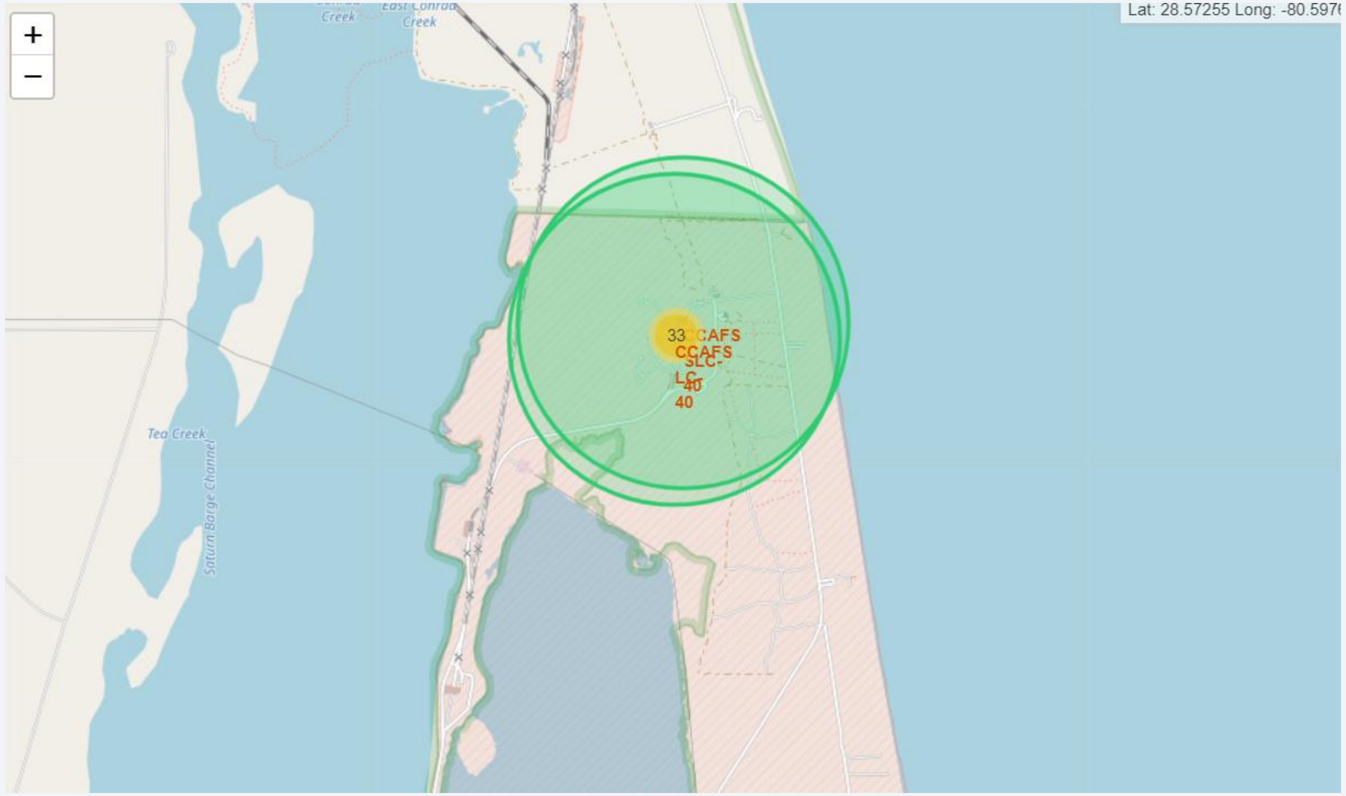
Launch Outcome by Sea

- Green one means successful result and rad one means failure



Logistics and Safety

- KSC LC-39A is logistic and safe, because it is far away from my populated places.

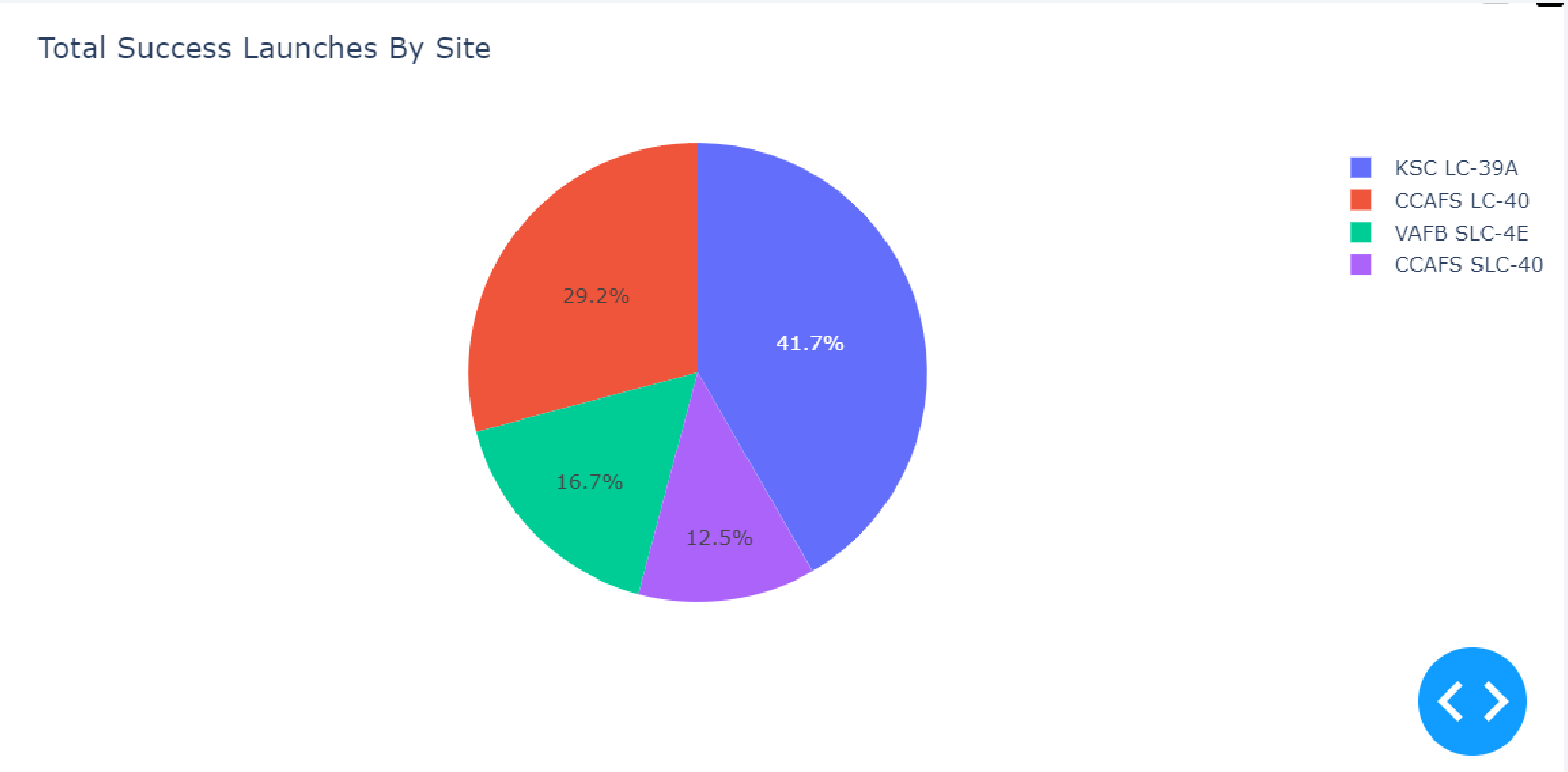


Section 4

Build a Dashboard with Plotly Dash

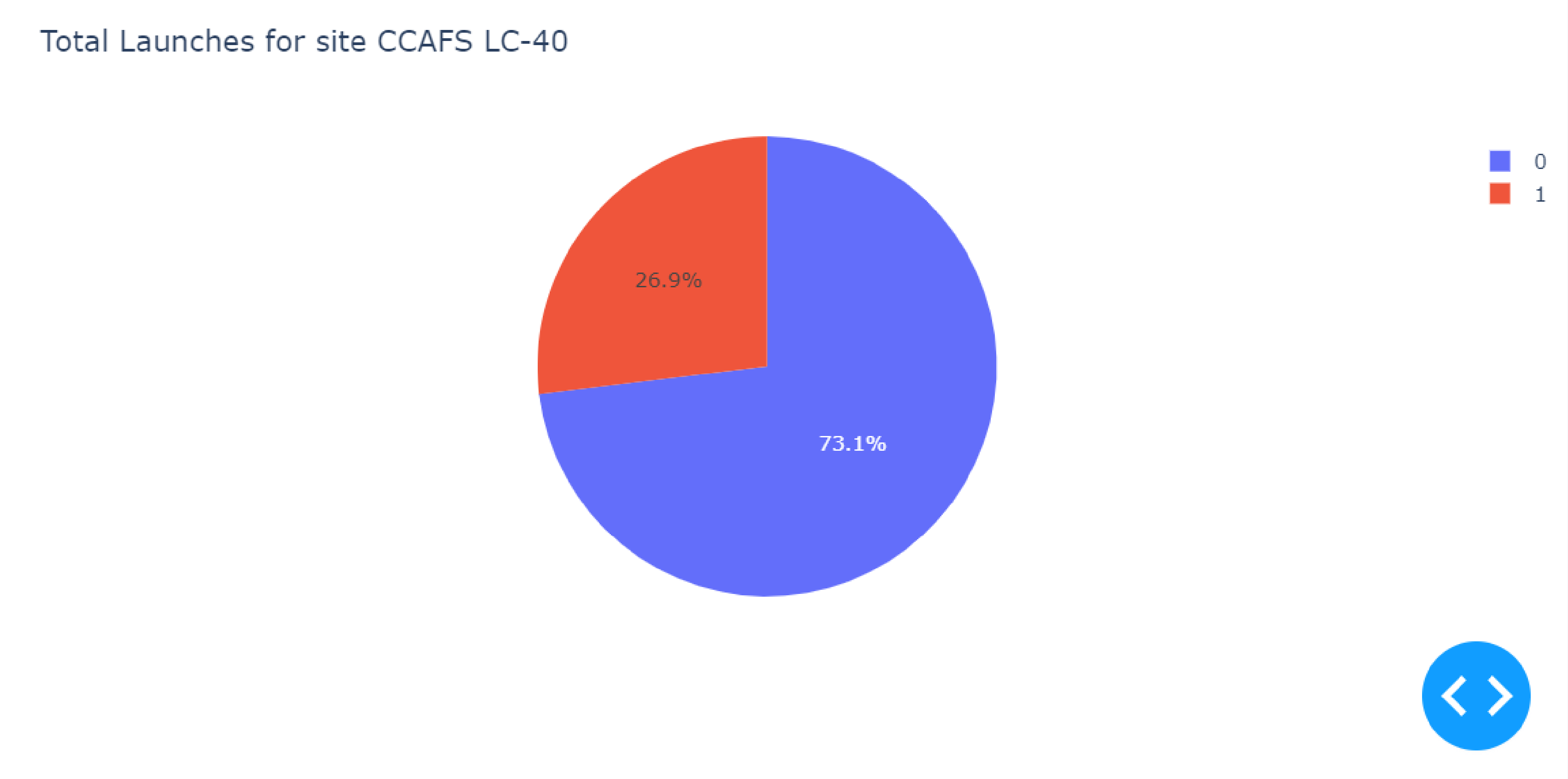
Successful Launches by Site

- KSC LC-39A has the most successful launch, 41.7%



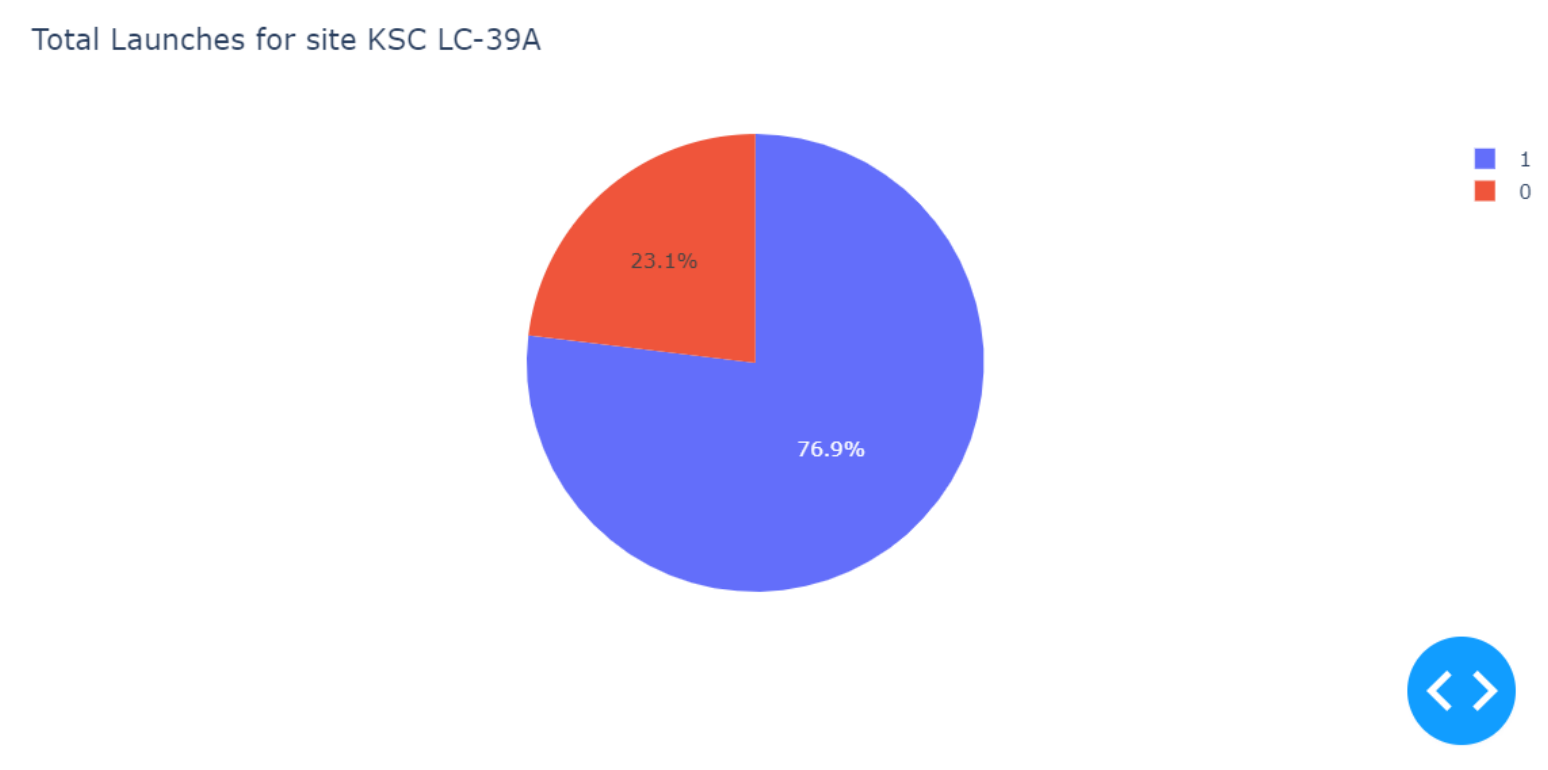
CCAFS LC-40 Launches

- Successful rate is 73.1%



KSC LC-39A Launches

- Successful rate: KSC LC-39A -> 76.9%



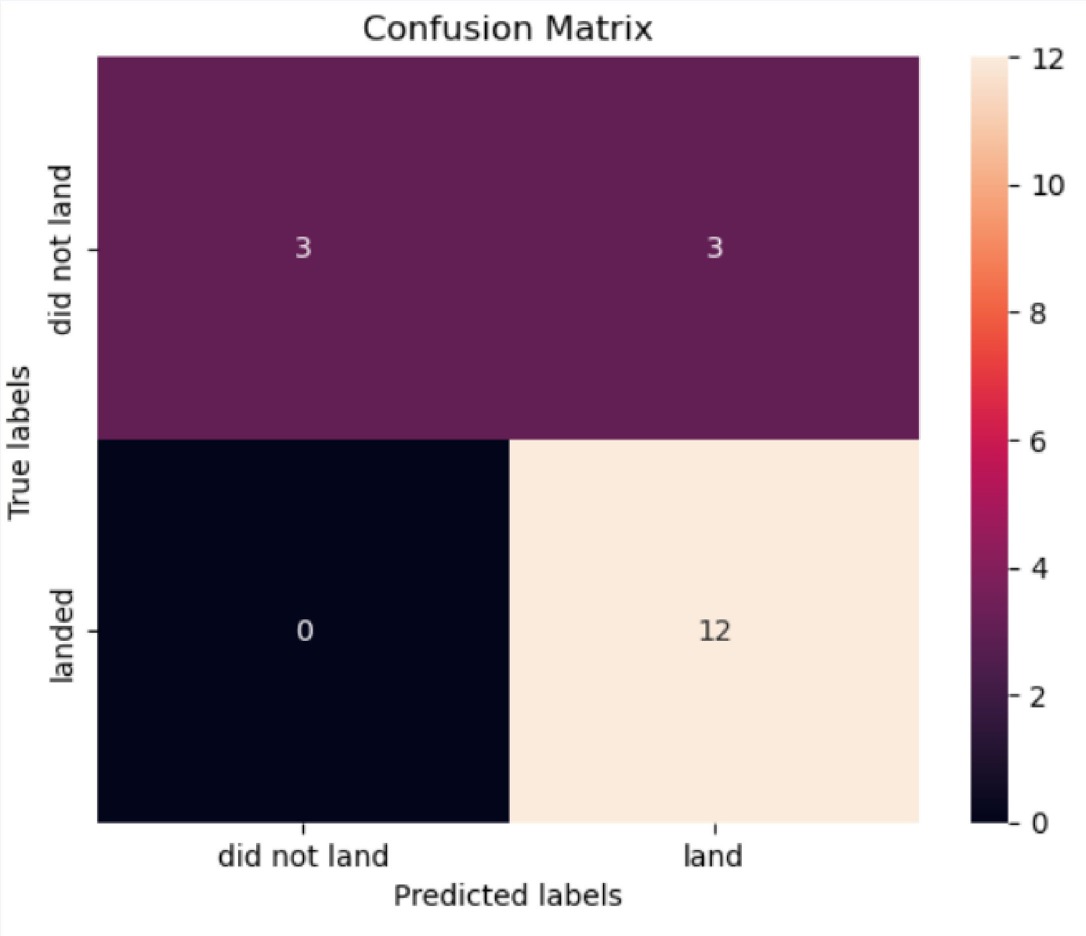
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
  - SVM
  - Logical Regression
  - Decision Tree
  - KNN
- Best Performing Model: Logistic Regression
- Accuracy on Test Data: 0.8333333333333334

Confusion Matrix



- True Positive and False Negative are mostly distributed. So machine learning is accurate.

Conclusions

- Logical regression classifier can be used to predict successful landings and reduce costs (This is the optimal choice)
- KSC LC-39A has the best result
- Launches between 7000 kg and 10000 kg are the most safe.

Appendix

- Some notebooks may have inconsistencies, but all questions I have at least one successful result.
- I use screenshot inside.

Thank you!