IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Linh T. HOANG
April 27, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting the data via making a get request to the SpaceX API
  - Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
  - Data wrangling on the collected data
  - Exploratory Data Analysis on the SpaceX dataset using SQL
  - Exploratory Data Analysis and Feature Engineering with Pandas & Matplotlib
  - SpaceX Launch Sites Locations Analysis with Folium
  - Building an interactive dashboard with Plotly
  - Machine Learning Prediction on the SpaceX dataset
- Summary of results: the best prediction model achieved an accuracy of 83%

Introduction

- 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.
- Therefore 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.
- In this capstone, we predict whether the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

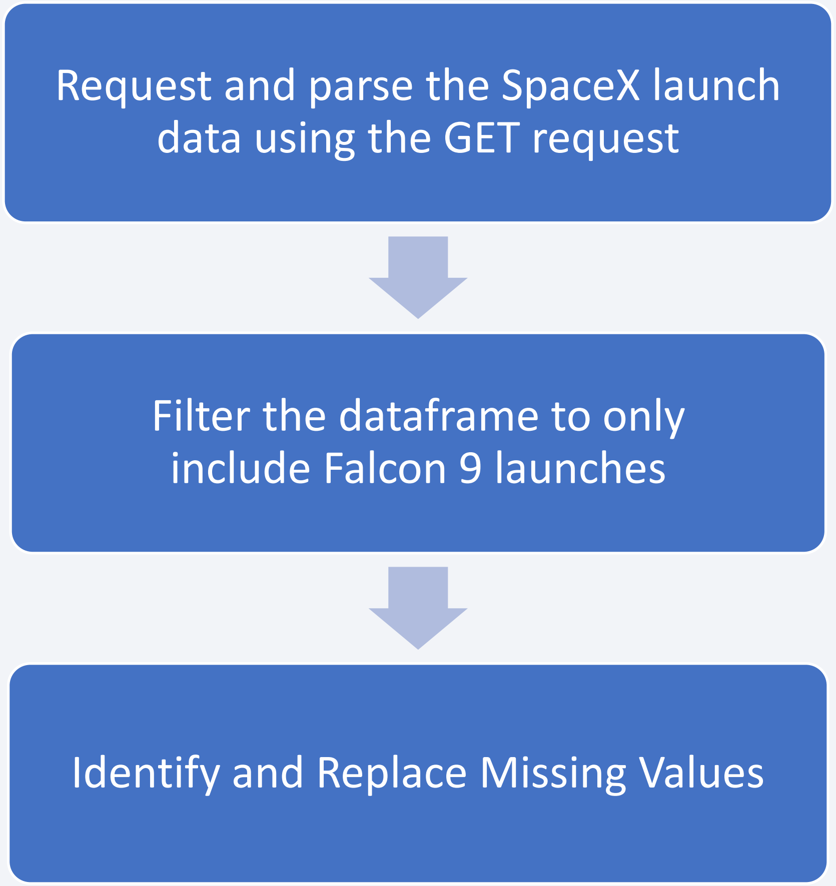
Executive Summary

- Data collection methodology:
  - Data was collected via the SpaceX API and webscraping from Wikipedia
- Perform data wrangling
  - Identify and handle missing values, apply one-hot encoding on some data columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build Logistic Regression, Support Vector Machine (SVM), Decision Tree, and KNN Classifiers
  - Fine-tune hyper-parameters via GridSearchCV

Data Collection – SpaceX API

- Figure: data collection process with SpaceX REST calls

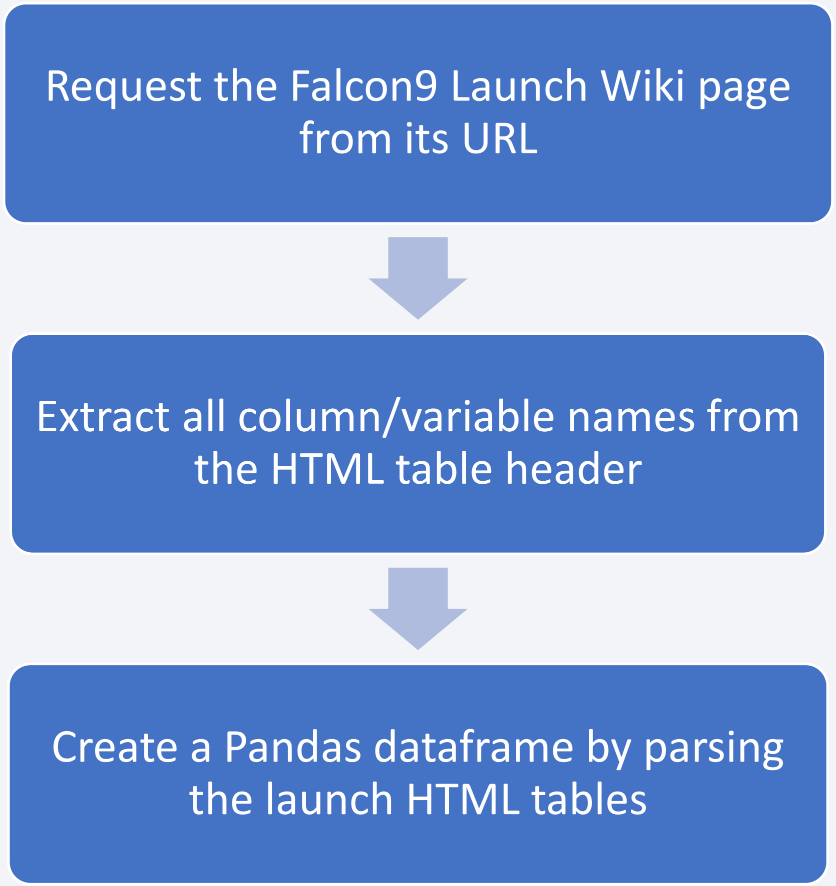
- GitHub URL to the notebook: https://github.com/linhhbk/SpaceX-Falcon9/blob/main/W1A_Data_Collection_API_Lab.ipynb



Data Collection – Web Scraping

- Figure: data collection process with web scraping

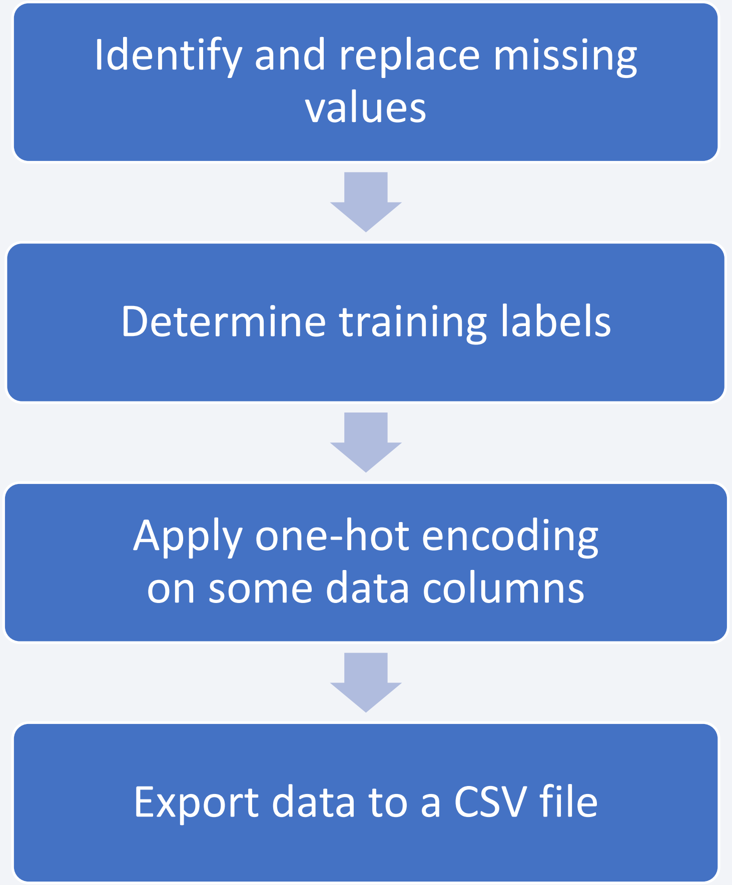
- GitHub URL to the web scraping notebook: https://github.com/linhhbk/SpaceX-Falcon9/blob/main/W1A_Data_Collection_Webscraping.ipynb



Data Wrangling

- Figure: data wrangling process

- GitHub URL to the data wrangling notebook: https://github.com/linhhbk/SpaceX-Falcon9/blob/main/W1B_Data_Wrangling.ipynb



EDA with Data Visualization

- The following charts were plotted to visualize the relationship between corresponding attributes:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Success Rate of each Orbit type
  - Flight Number and Orbit type
  - Payload and Orbit type
  - Yearly trend of the launch success rate
- GitHub URL of to the EDA with data visualization notebook: https://github.com/linhhbk/SpaceX-Falcon9/blob/main/W2B_Exploratory_Data_Analysis_with_Pandas.ipynb

EDA with SQL

- Summary of the SQL queries performed:
  - SELECT * FROM * WHERE *
  - LIMIT, DISTINCT, COUNT
  - LIKE, GROUP BY, ORDER BY
  - AVG, MAX, MIN, … and Implicit JOIN
  - Subquey, etc.
- GitHub URL to the EDA with SQL notebook: https://github.com/linhhbk/SpaceX-Falcon9/blob/main/W2A_Exploratory_Data_Analysis_with_SQL.ipynb

Build an Interactive Map with Folium

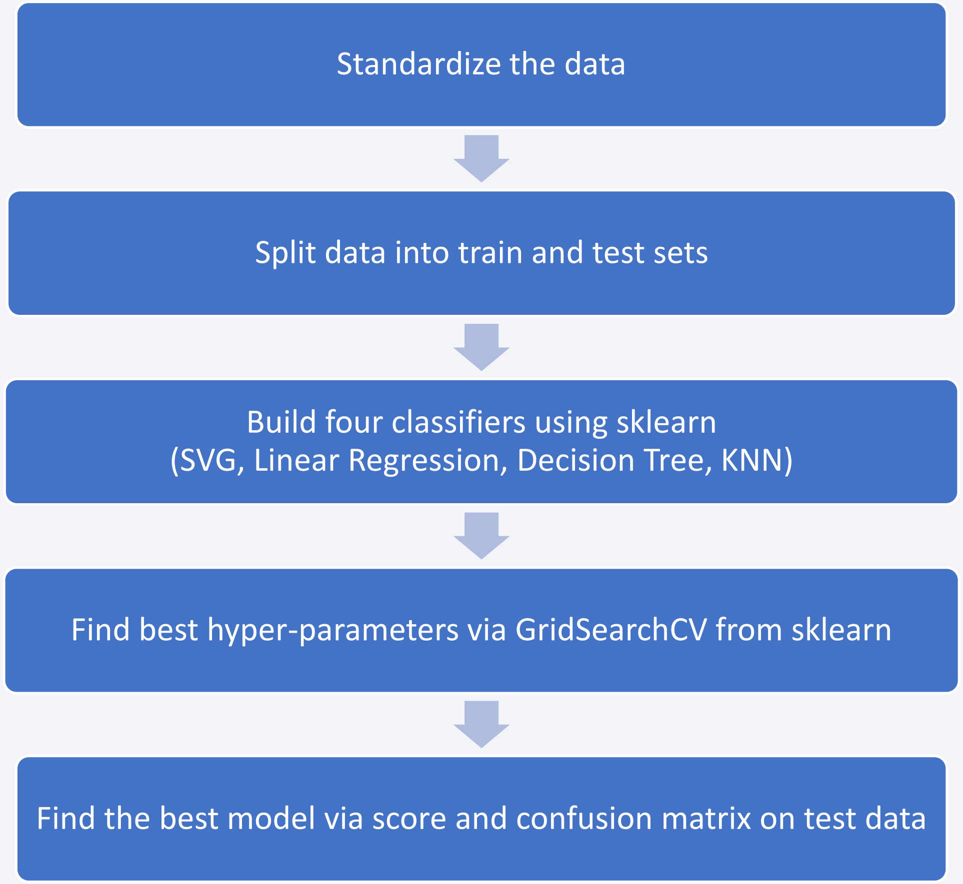
- Map objects created and added to a folium map:
  markers, marker cluster, circles, lines, etc.
- These objects were added to analyze the existing launch site locations of SpaceX.
  The launch success rate may depend on the location and proximities of a launch site.
- GitHub URL to the interactive map with Folium notebook:
  https://github.com/linhhbk/SpaceX-Falcon9/blob/main/W3A_Interactive_Visual_Analytics_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- Summary of plots/graphs and interactions added to the dashboard:
  - A dropdown list to enable Launch Site selection
  - A pie chart to show the total successful launches count
  - A slider to select payload range
  - A scatter chart to show the correlation between payload and launch success
- Those plots and interactions are for interactive visual analytics on the SpaceX dataset.
- GitHub URL to the Plotly Dash lab: https://github.com/linhhbk/SpaceX-Falcon9/blob/main/W3B_SpaceX_Dash_App.py

Predictive Analysis (Classification)

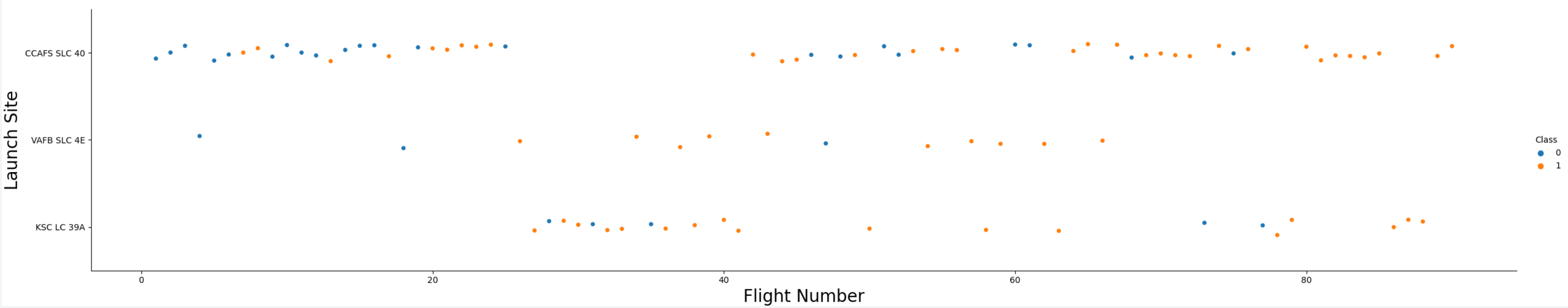
- Summarize how you built, evaluated, improved, and found the best performing classification model
- Figure: model development process
- GitHub URL to the predictive analysis lab: https://github.com/linhhbk/SpaceX-Falcon9/blob/main/W4_SpaceX_Machine_Learning_Prediction.ipynb



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

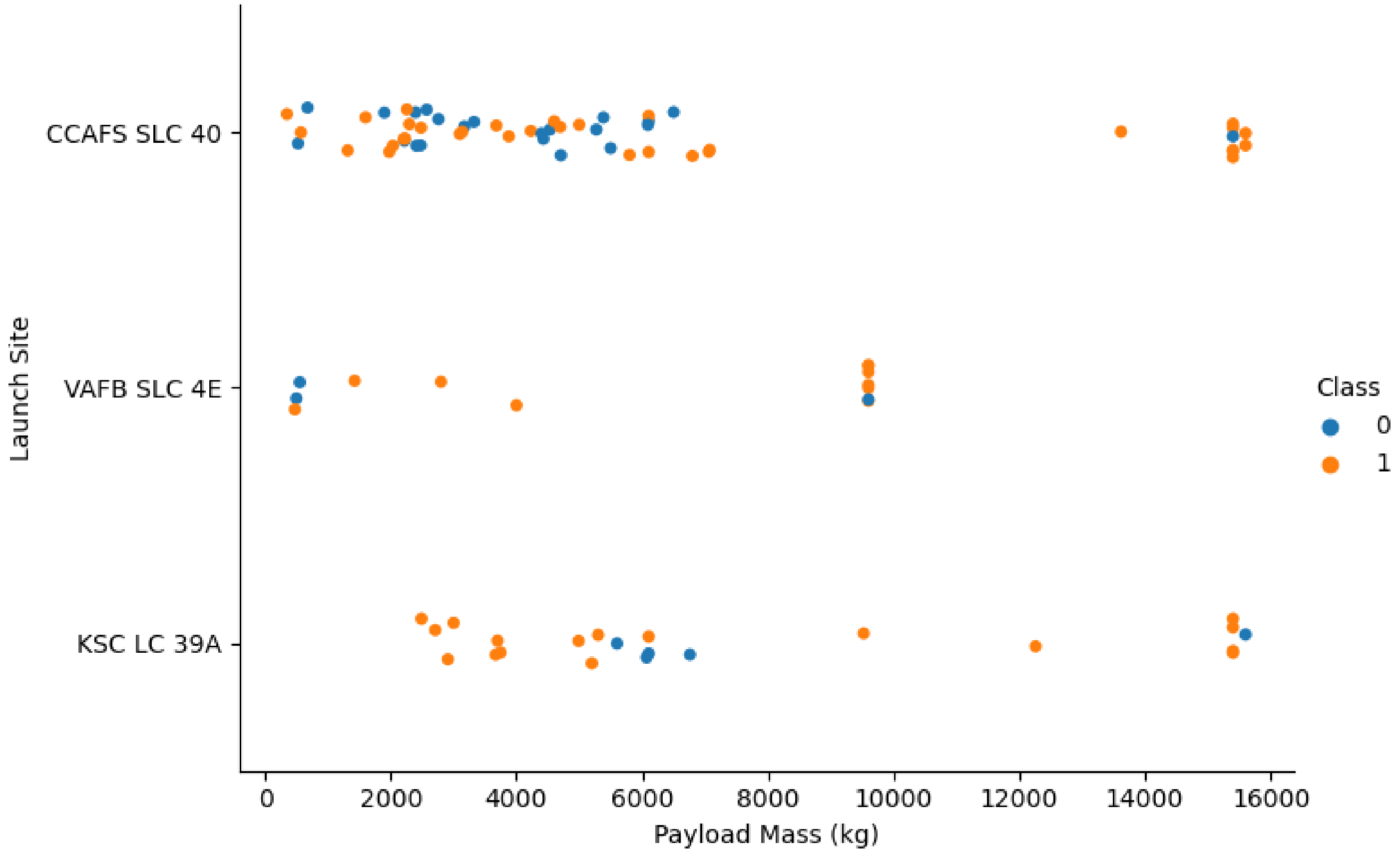


- Figure: a scatter plot of Flight Number vs. Launch Site
- Explanations:
  - The launch success rate increases over time (i.e., when the flight number increases)
  - The launches are not even between launch sites, with CCAFS SLC 40 being the site with most launches

Payload vs. Launch Site

- Figure: a scatter plot of Payload vs. Launch Site
- Explanations:

  For VAFB-SLC 4E launchsite, there are no rockets launched for heavypayload mass (greater than 10000)

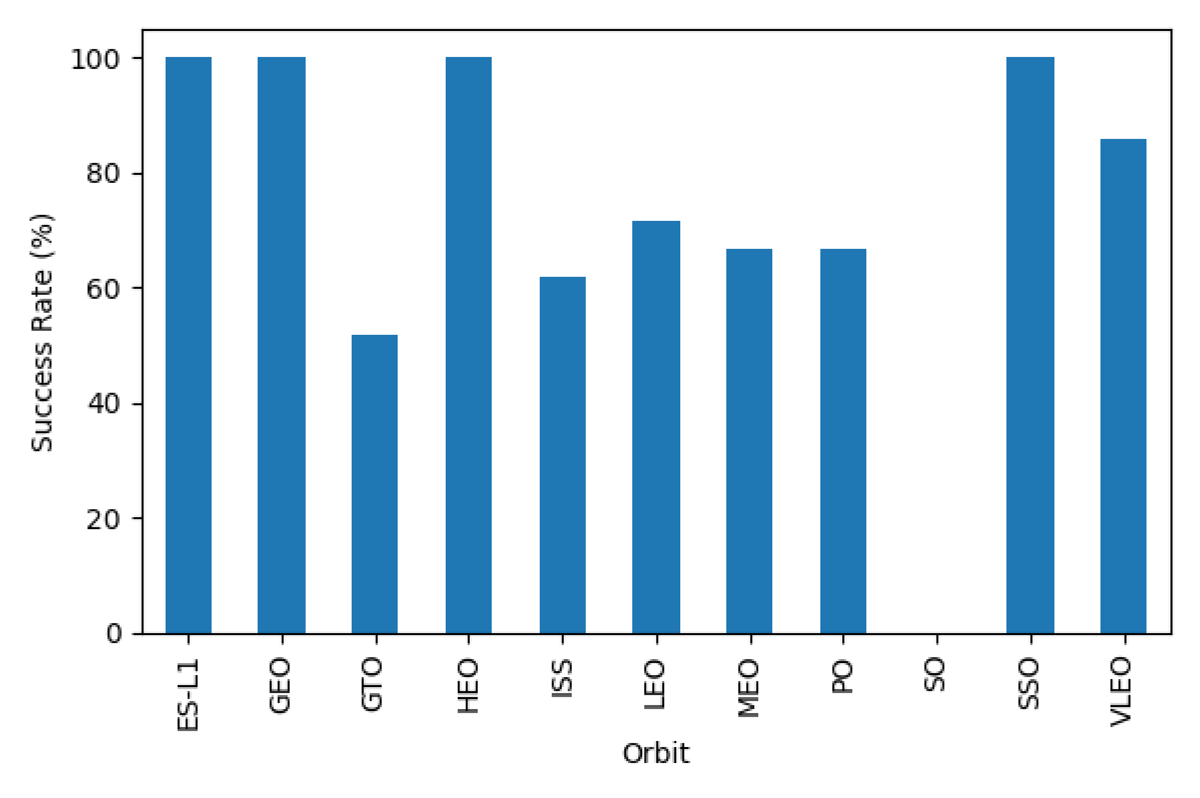


Success Rate vs. Orbit Type

- Figure: a bar chart for the success rate of each orbit type

- Explanations:

  The orbits with high success rate are: ES-L1, GEO, HEO, and SSO



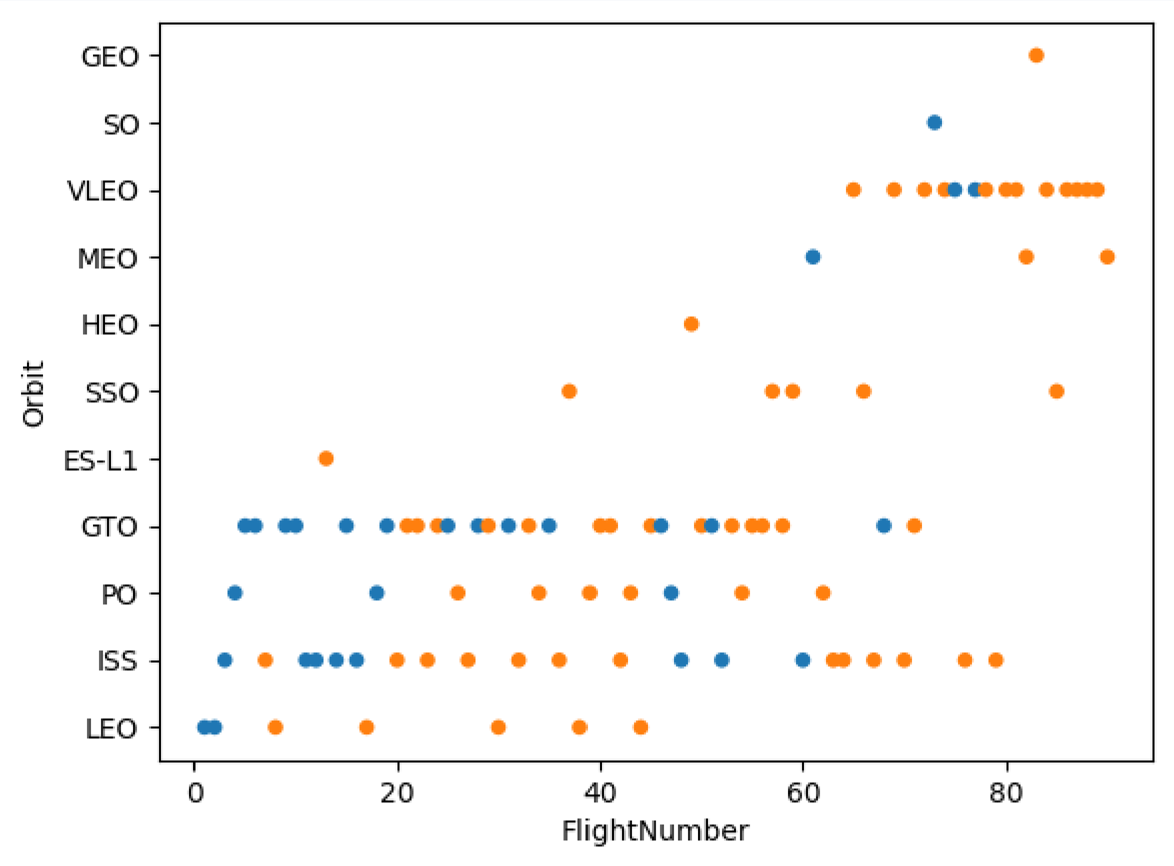
Flight Number vs. Orbit Type

- Figure: a scatter point of Flight number vs. Orbit type

- Explanations:

  In the LEO orbit, the success appears related to the number of flights;

  On the other hand, there seems to be no relationship between flight number when in GTO orbit.

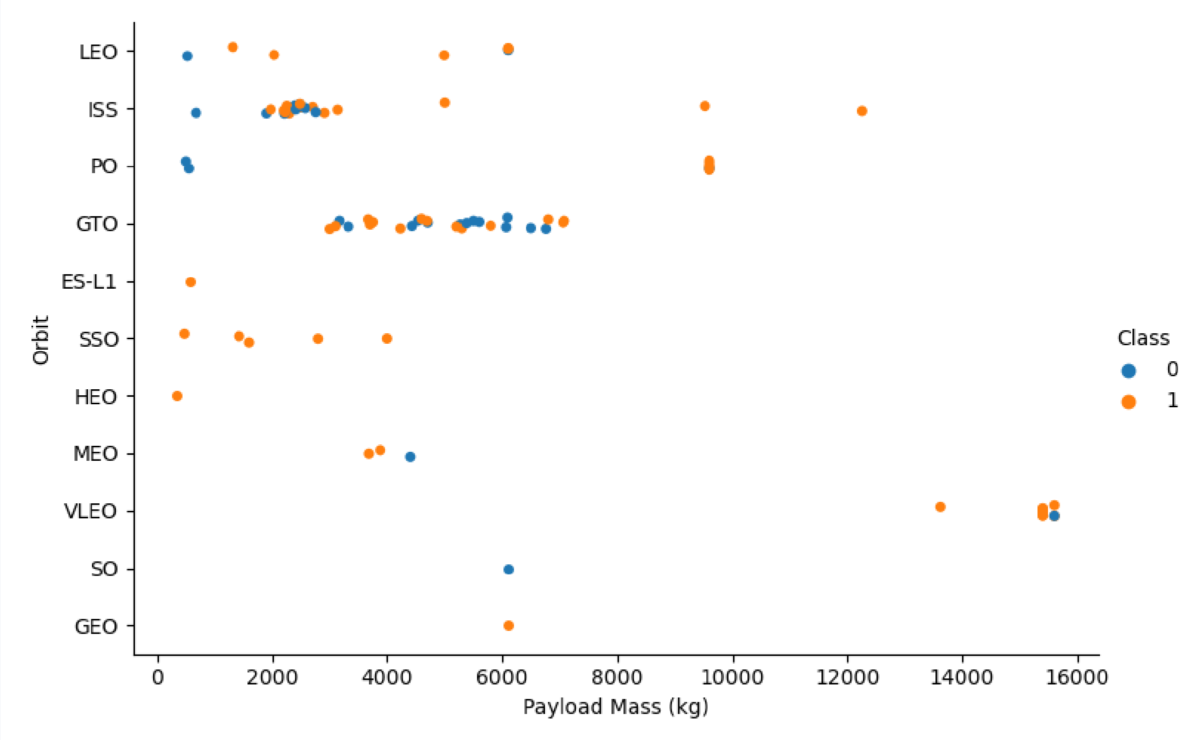


Payload vs. Orbit Type

- Figure: a scatter point of Payload vs. Orbit type
- Explanations:

  With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS.

  However for GTO, we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.

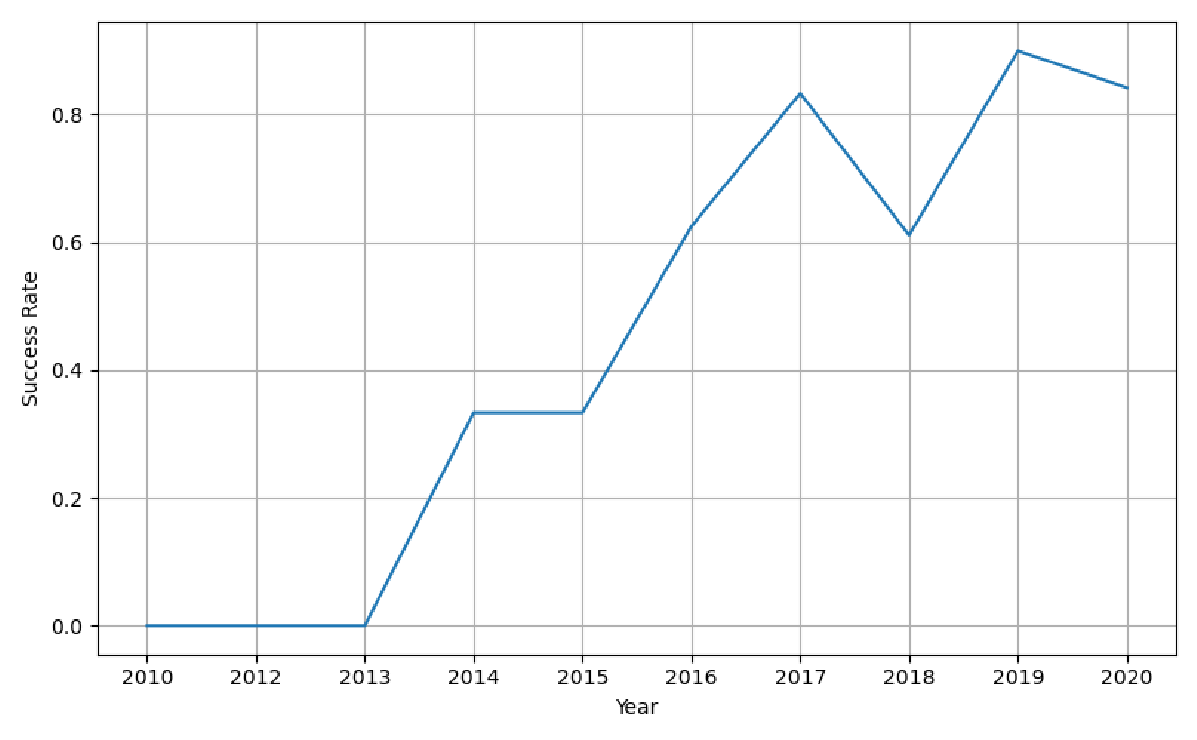


Launch Success Yearly Trend

- Figure: a line chart of yearly average success rate

- Explanations:

  The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Figure: query result to find the names of the unique launch sites
- Explanation: there are four launch sites in total

Task 1

Display the names of the unique launch sites in the space mission

```
In [4]: %sql select distinct(launch_site) from spacex;

 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

Figure: query result when finding 5 records where launch sites begin with `CCA`

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [9]:
```
%%sql
select *
from spacex
where launch_site like'CCA%'
limit 5;
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[9]:

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

- Figure: query result when calculating the total payload carried by boosters from NASA
- Explanation: The total payload is 45,596 Kg

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [11]:

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEX
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[11]:

1
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Explanation: The average payload mass is 2,534 Kg

Task 4

Display average payload mass carried by booster version F9 v1.1

In [16]:

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEX
WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[16]:

1
2534

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Explanation: The first successful landing was on December 22, 2015

Task 5

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

Hint:Use min function

In [17]:

```
%%sql
SELECT MIN(DATE)
FROM SPACEX
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[17]:

1
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
- Explanation: There are five boosters that satisfy the conditions.

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [18]:

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEX
WHERE LANDING__OUTCOME='Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[18]:

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
- Explanation: There are 99 successful missions and only 2 failed ones

Task 7

List the total number of successful and failure mission outcomes

In [24]:

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL_NUMBER_OF_LAUNCHES
FROM SPACEX
GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[24]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Explanation: There are 12 boosters that have carried the maximum payload mass

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [27]:

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEX
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[27]:

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
- Explanation: There are two launches that satisfy the conditions

Task 9

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

In [30]:

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, DATE
FROM SPACEX
WHERE LANDING__OUTCOME = 'Failure (drone ship)'
AND YEAR(DATE)=2015;
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[30]:

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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
- Explanation: In the given time period, landing on drone ships is the most.

Task 10

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

In [37]:

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT_OF_LAUNCHES
FROM SPACEX
WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY COUNT_OF_LAUNCHES DESC;
```

```
 * ibm_db_sa://qwq44400:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[37]:

landing__outcome	count_of_launches
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

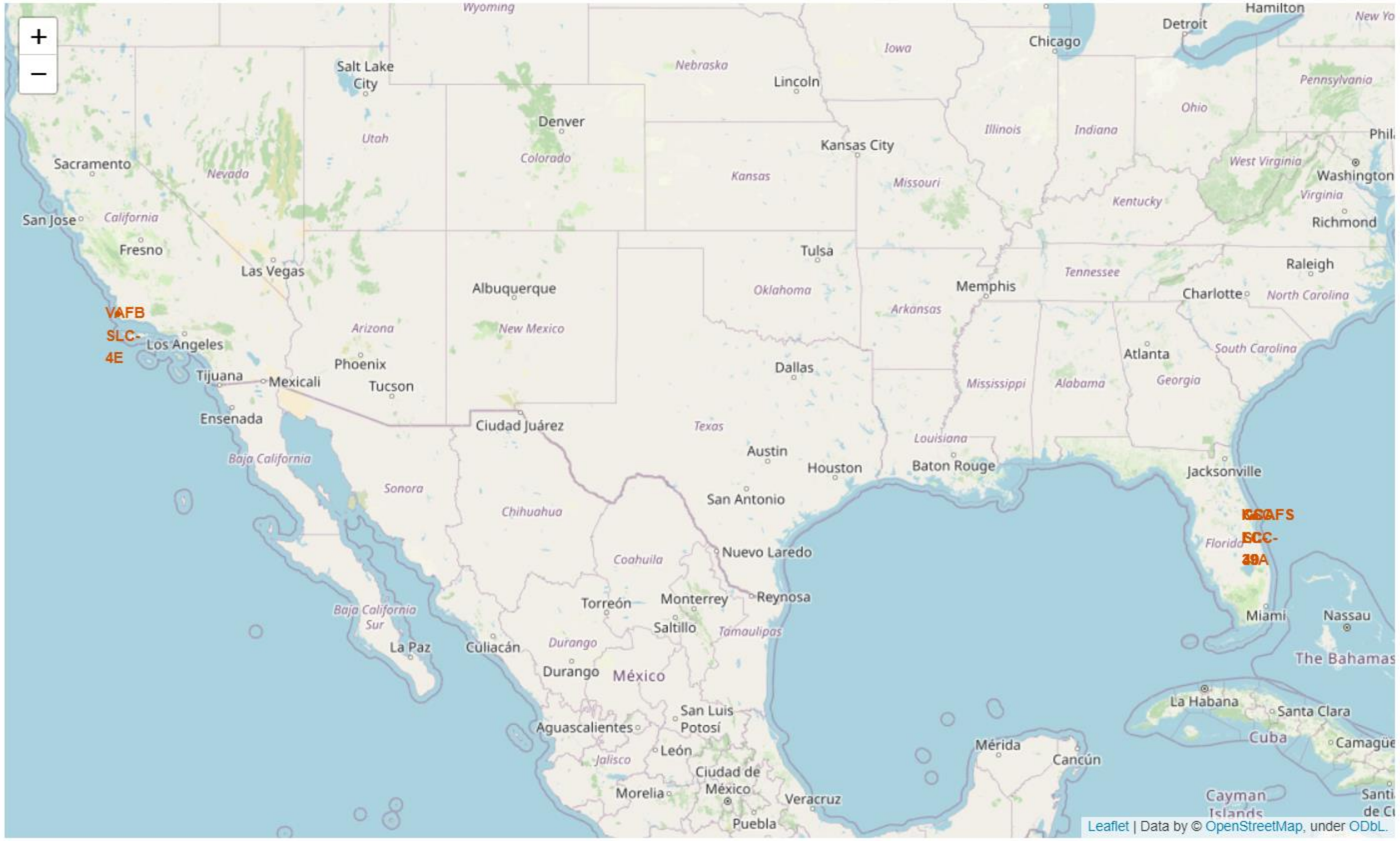
Launch Sites Proximities Analysis

Locations of all launch sites on a global map

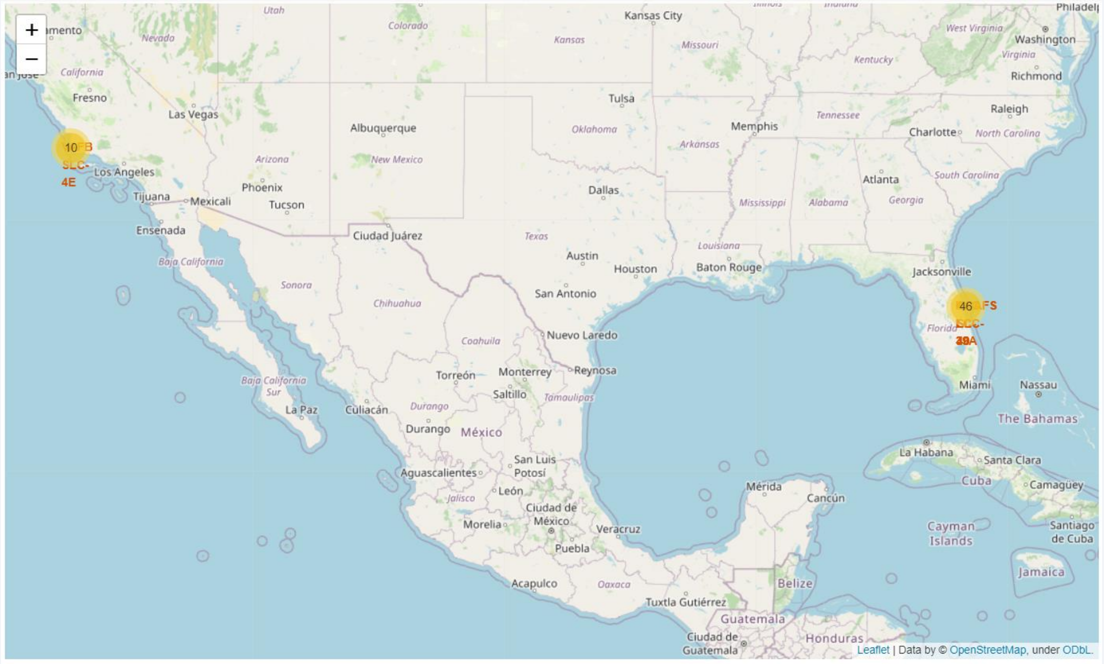
- Figure: locations of all launch sites on a global map, plotted using Folium
- Findings:

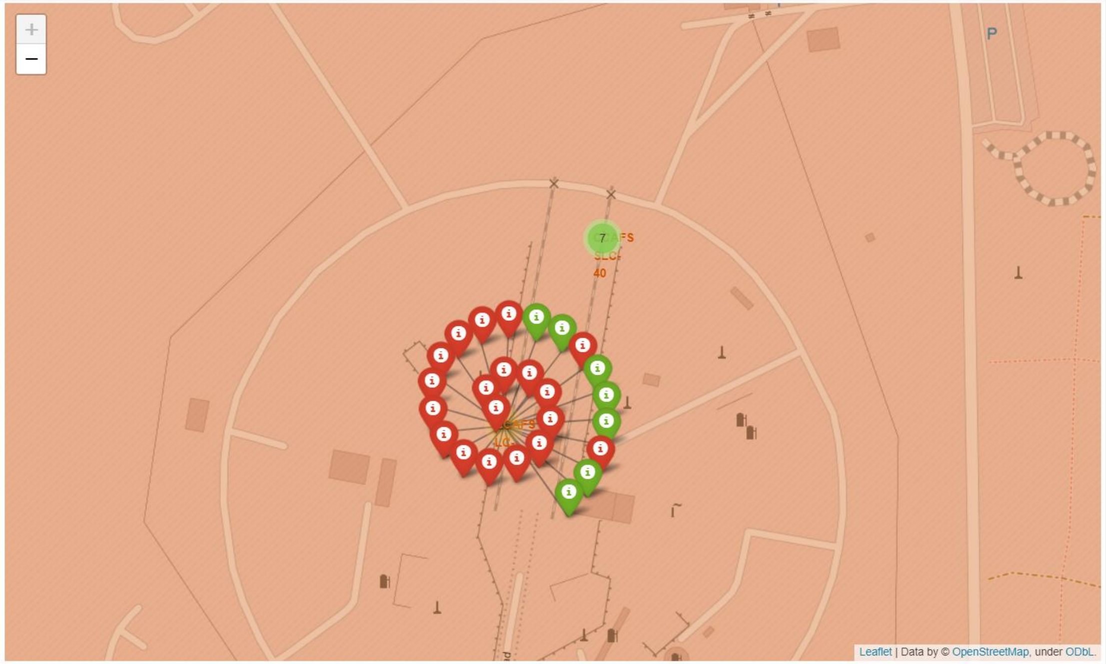
All launch sites are in proximity to the Equator line. Rockets launched from these sites get an additional natural boost that helps save on fuel and boosters.

All launch site sare in very close proximity to the coast to minimize damage in the event of an accident



Marking the success/failed launches on the map





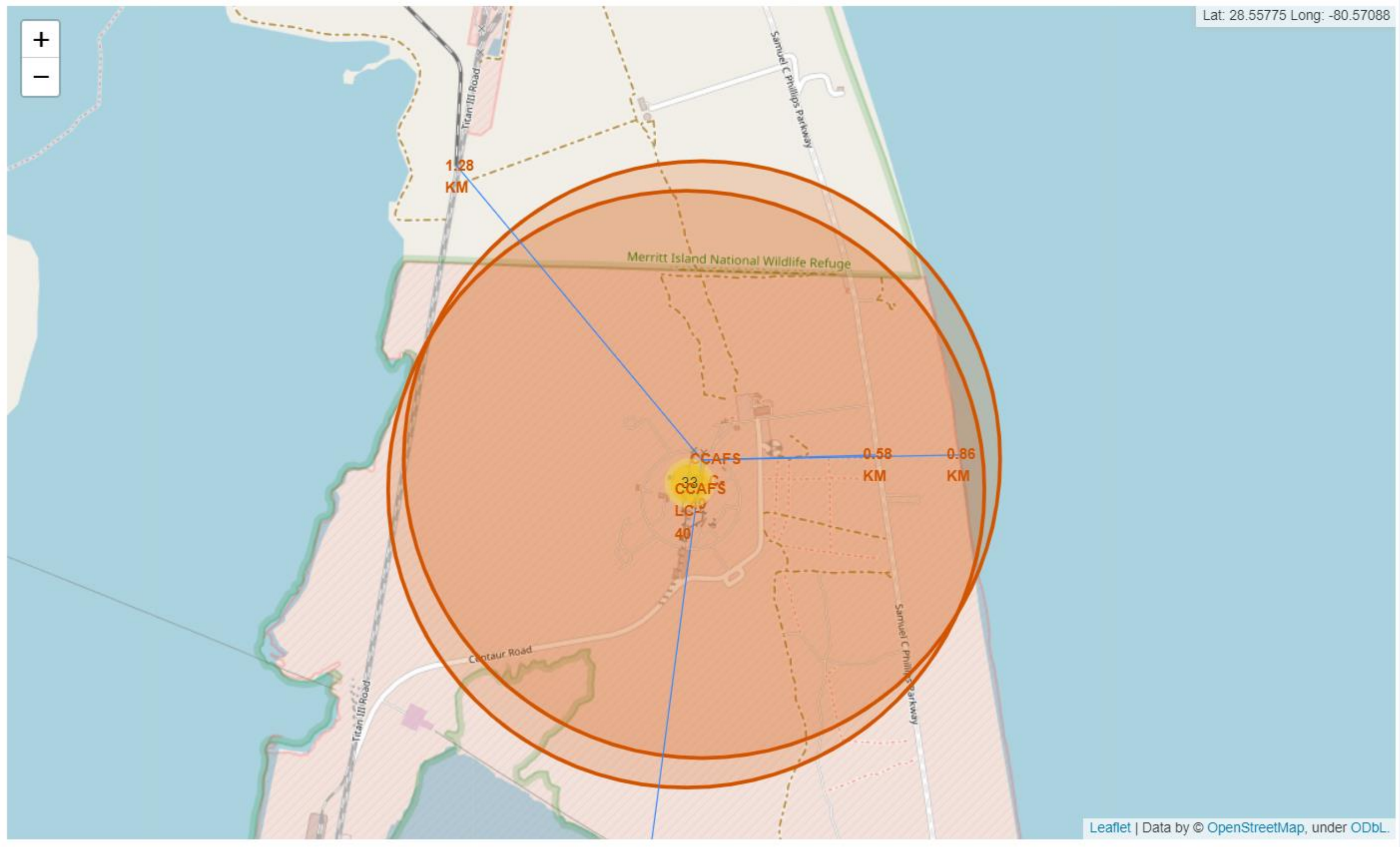
- Figure: color-labeled launch outcomes for each site on the map
- Explanations: Green and red markers denote success and failed launches, respectively. From this map, we can identify which launch sites have relatively high success rates.

Proximities of Launch Sites

- Figure: Launch sites' proximities (railway, highway, coastline, etc.) with distance calculated and displayed
- Explanations:

  Launch sites are in close proximity to railways and highways to facilitate transportation.

  Launch sites are in close proximity to coastline keep certain distance away from cities to minimize damage in the event of an accident.



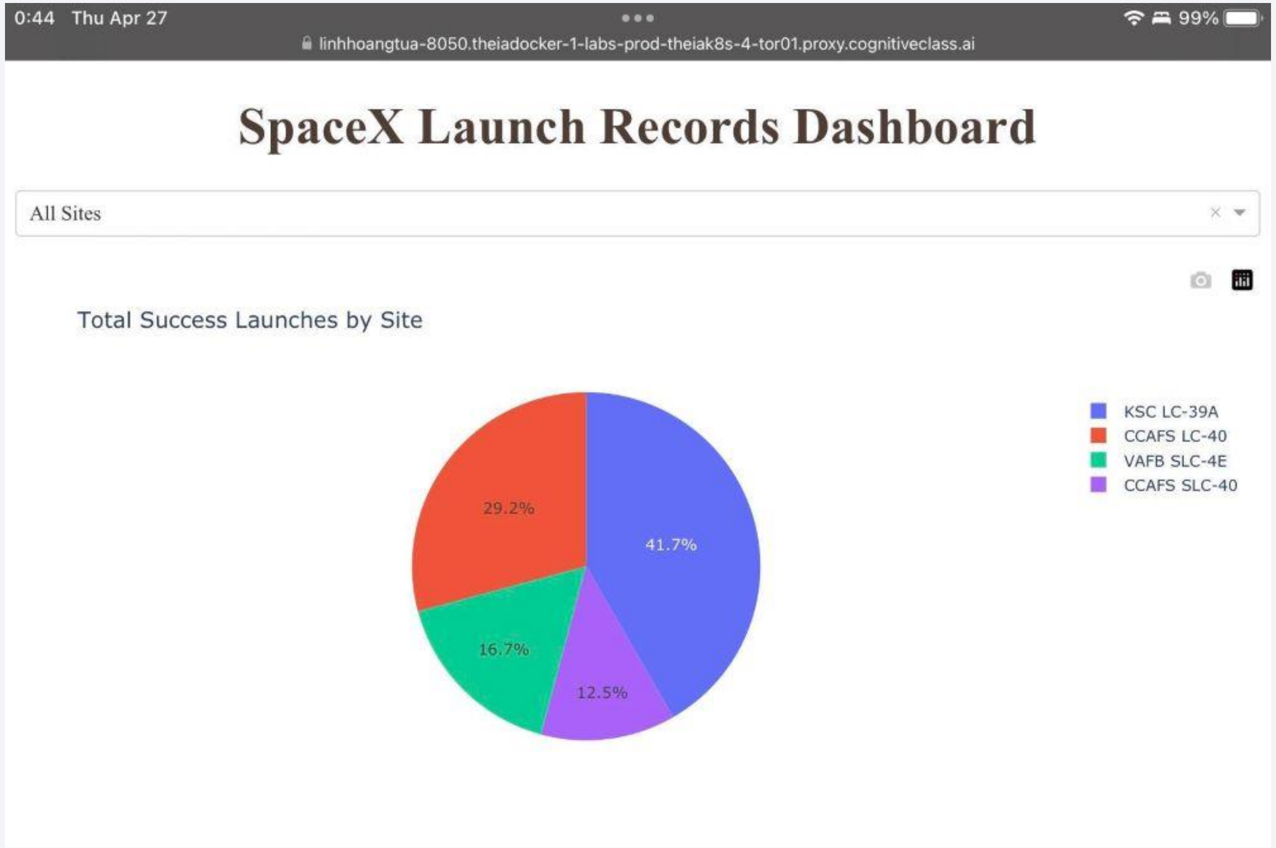
Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

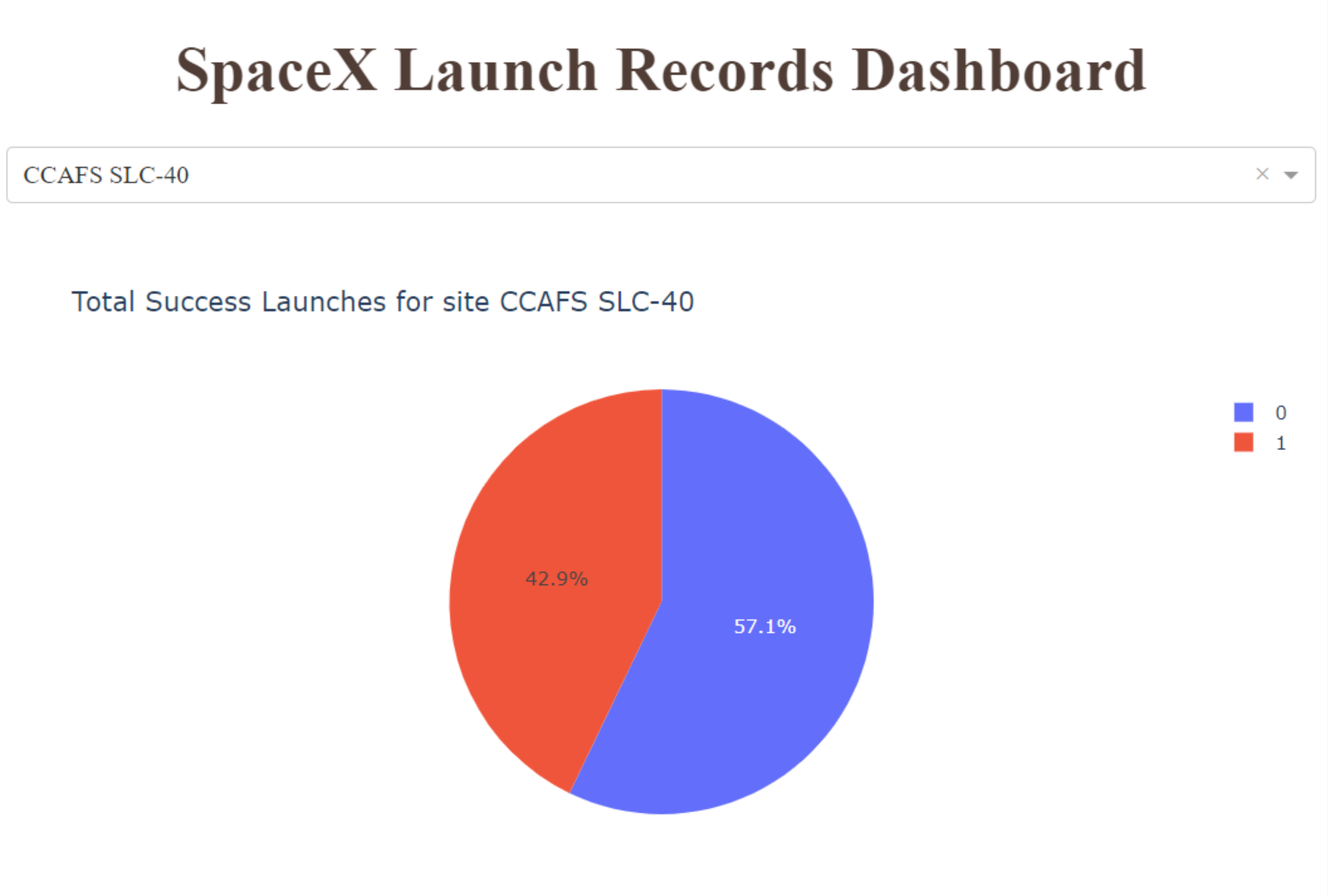
- Figure: a screenshot of launch success count for all sites
- Explanations:

  KSC LC-39A is the launch site with the most successful launches



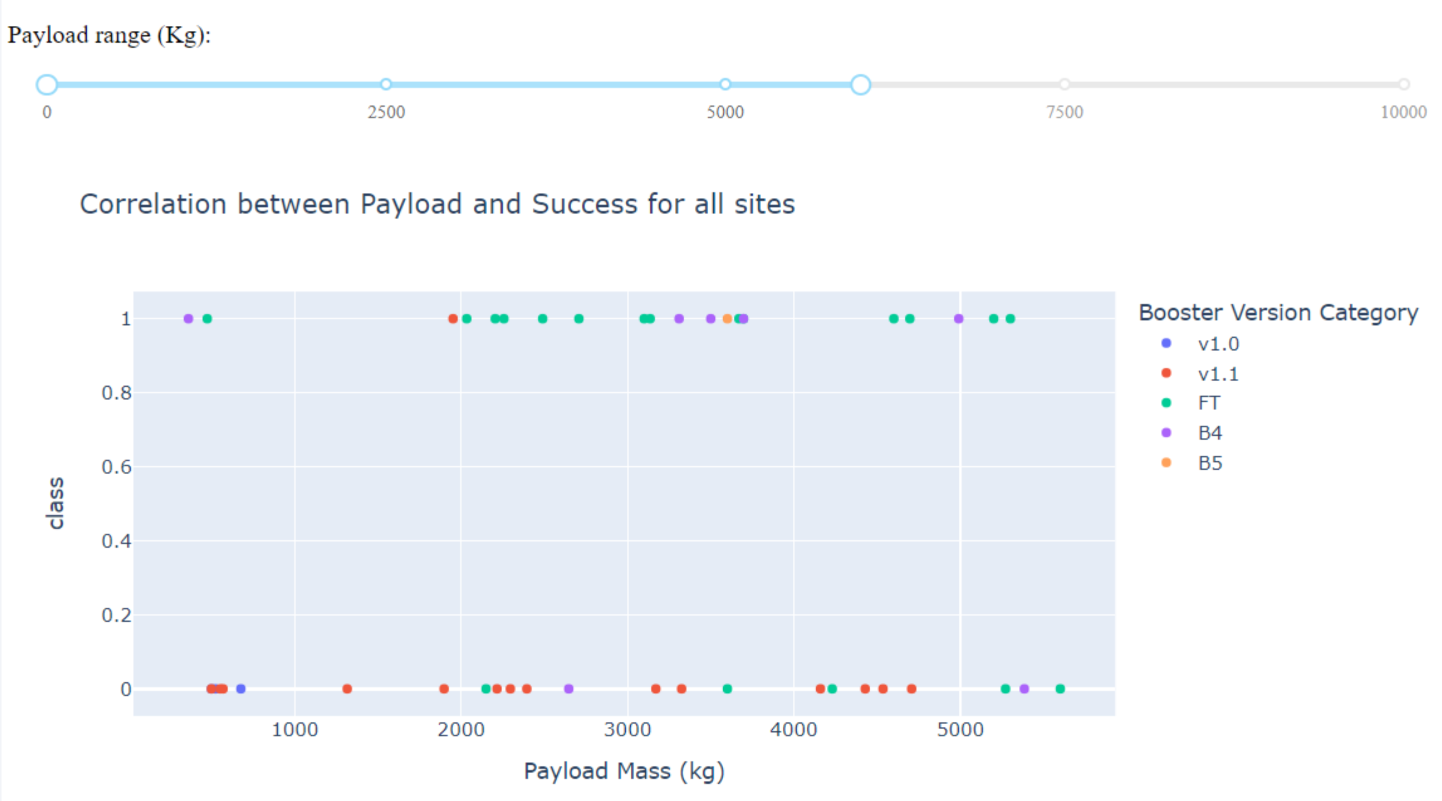
Total Success Launches for site CCAFS SLC-40

- Figure: Total success launches of the site with highest launch success ratio, CCAFS SLC-40
- Explanatioins: CCAFS SLC-40 is the site with the highest launch success ratio of 42.9%



Payload vs. Launch Outcome Scatter Plot for All Sites

- Figure: Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



- Explanations:

  Payload range from 500 to 6,000 Kg and booster version FT have the largest success rate.

Section 5

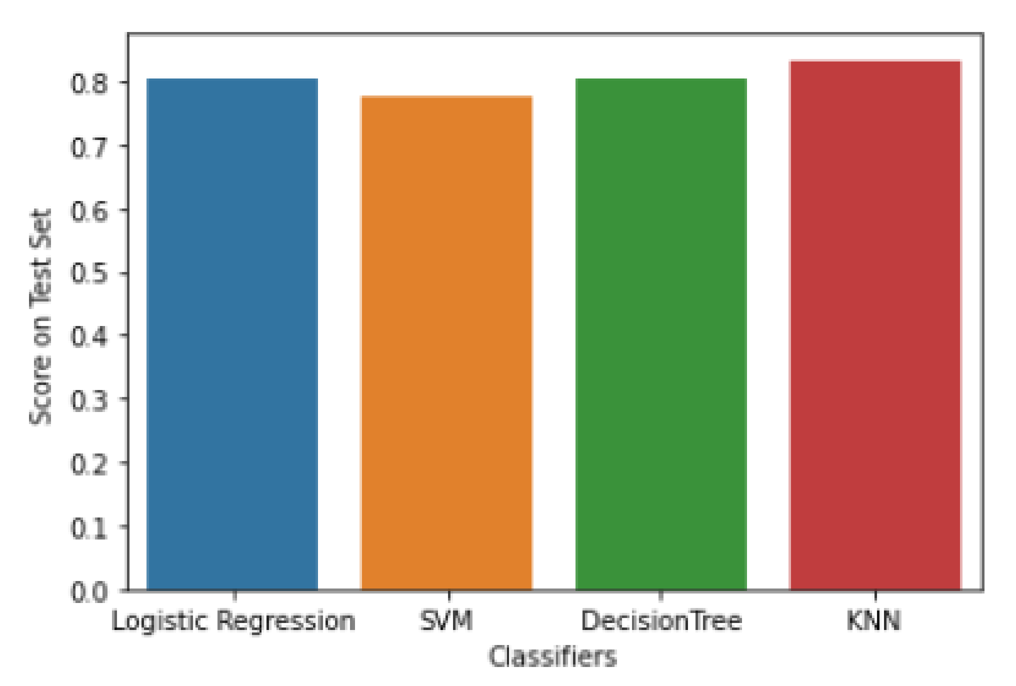
Predictive Analysis (Classification)

Classification Accuracy

- Figure: model accuracy for all built classification models

- Finding: KNN model has the highest classification accuracy

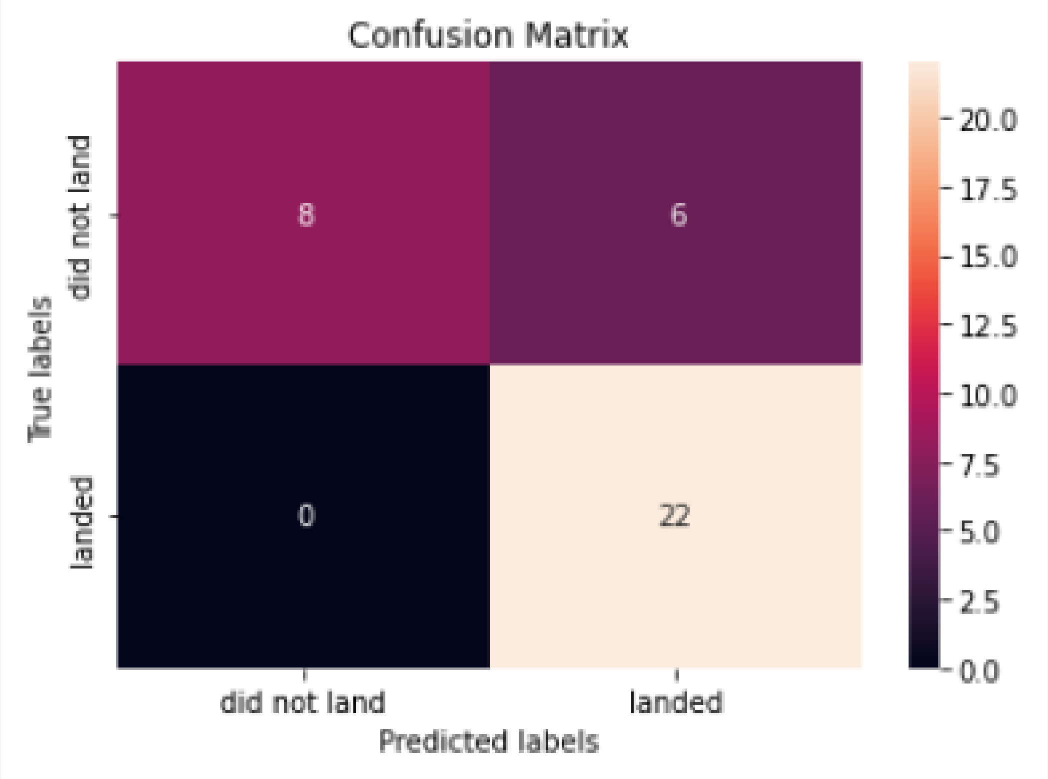
- NOTE: the results are with a quite small dataset with 98 data points, 40% of them are used for testing



Confusion Matrix

- Figure: the confusion matrix of the best performing model, KNN Classifier

- Finding: the KNN Classifier correctly recognized all successful landings



Conclusions

In this capstone project:

- Data was collected via the SpaceX API and webscraping from Wikipedia
- Data wrangling was conducted on the collected data
- Exploratory data analysis (EDA) was conducted using visualization and SQL
- Interactive visual analytics was conducted using Folium and Plotly Dash
- Predictive analysis was conducted using machine learning classification models
    - The best prediction model achived an accuracy of 83%

Appendix

- GitHub Link to all Labs: https://github.com/linhhbk/SpaceX-Falcon9

Thank you!