IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Guoqing Wang
2/2/2025

Outline

Executive Summary

- **Summary of methodologies**

  The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:

  - Collect data using SpaceX REST API and web scraping techniques
  - Wrangle data to create success/fail outcome variable
  - Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
  - Analyze the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total number of successful and failed outcomes
  - Explore launch site success rates and proximity to geographical markers
  - Visualize the launch sites with the most success and successful payload ranges
  - Build machine learning models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

- **Summary of all results**
  - **Exploratory Data Analysis:**
    - Launch success has been improved over time
    - Effect of payload on the launch success rate is site dependent
    - Orbits SSO has a 100% success rate
  - **Visualization/Analytics:**
    - Most launch sites are near the equator, and all are close to the coast
  - **Predictive Analytics:**
    - All machine learning models performed similarly on the test set with >80% accuracies.

Introduction

Project background and context

- SpaceX, a pioneer in the space sector, aims to make space travel accessible to all. Its achievements include delivering spacecraft to the International Space Station, deploying a satellite network for global internet access, and conducting manned missions. The company keeps costs low—around $62 million per launch—thanks to its innovative reuse of the Falcon 9 rocket's first stage, whereas competitors who cannot reuse this component face costs of $165 million or more per launch. By predicting whether the first stage will successfully land, we can estimate the launch cost. This assessment can be performed using public data combined with machine learning models to forecast if SpaceX—or another company—will be able to reuse the first stage.

Problems

- How do payload mass, launch site, number of flights, and orbits affect first-stage landing success
- What is the rate of successful landings over time
- What is the best predictive model for successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - By SpaceX REST API and web scraping techniques
- Perform data wrangling
  - By filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- Perform exploratory data analysis (EDA)
  - By visualization and SQL
- Perform interactive visual analytics
  - By Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

1. Request and parse the SpaceX launch data using the GET request

2. Extract data using custom functions to make data dictionary

3. Filter the dataframe to only include Falcon 9 launches

4. Dealing with Missing Values

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain
```

We should see that the request was successfull with the 200 status response code

```
response = requests.get(static_json_url)
```

```
# Call getBoosterVersion
getBoosterVersion(data)
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
```

```
# Filter the DataFrame to keep only Falcon 9 launches
data_falcon9 = data_falcon[data_falcon['BoosterVersion'] == 'Falcon 9']

# Verify the filtering
print(data_falcon9['BoosterVersion'].unique())
```

```
# Calculate the mean value of PayloadMass column
payload_mean = data_falcon9['PayloadMass'].mean()

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, payload_mean, inplace=True)
```

GitHub URL : https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX/blob/main/jupyter-labs-spacex-data-collection-api-v2-w.ipynb

Data Collection - Scraping

1. Request the Falcon9 Launch Wiki page from its URL

2. Extract all column/variable names from the HTML table header

3. Process table content using customer functions and generate data dictionary

4. Create a data frame by importing dictionary

```
# use requests.get() method with the provided static_url
# assign the response to a object
response_F9 = requests.get(static_url)
```

```
for th in first_launch_table.find_all('th'):
    name = extract_column_from_header(th)
    if name is not None and len(name) > 0:
        column_names.append(name)
```

```
def landing_status(table_cells):
    """
    This function returns the landing status from the HTML table cell
    Input: the  element of a table data cell extracts extra row
    """
    out=[i for i in table_cells.strings][0]
    return out
```

```
# Booster landing
booster_landing = landing_status(row[8])
launch_dict["Booster landing"].append(booster_landing)
```

```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

GitHub URL : https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX/blob/main/jupyter-labs-webscraping-w.ipynb

Data Wrangling

calculate the number of launches on each site

```
# Apply value_counts() on column LaunchSite
launch_site_counts = df['LaunchSite'].value_counts()
launch_site_counts
```

calculate the number and occurrence of each orbit

```
# Apply value_counts on Orbit column
orbit_counts = df['Orbit'].value_counts()
```

calculate the number and occurrence of mission outcome per orbit type

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
df['landing_class'] = [0 if outcome in bad_outcomes else 1 for outcome in df['Outcome']]
```

create a landing outcome label from Outcome column

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
print(landing_outcomes)
```

GitHub URL : https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX/blob/main/labs-jupyter-spacex-Data%20wrangling-v2-w.ipynb

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis

- View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists
- Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.

GitHub URL : https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX/blob/main/jupyter-labs-eda-dataviz-v2-w.ipynb

EDA with SQL

Display:

- Names of unique launch sites
- 5 records where launch site begins with 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

GitHub URL : https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX/blob/main/EDA%20with%20SQL-w.ipynb

Build an Interactive Map with Folium

1. To highlight all launch sites on a map
   - **folium.Circle()** was used to create a circle for each launch site
   - **folium.Marker()** was used to create a marker for each launch site
2. To mark the success/failed launches for each site on the map
   - **folium.Marker()** was used to mark launching events with label as green/red if success/failed
   - **MarkerCluster()** was used to group launching events from the same site
3. To track the distances between a launch site to its proximities
   - **MousePosition** was used to get the coordinate on the map
   - **folium.PolyLine()** was used to draw a line between the the marker to the launch site, city, railway, highway

GitHub URL : https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX/blob/main/lab-jupyter-launch-site-location-v2-w.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

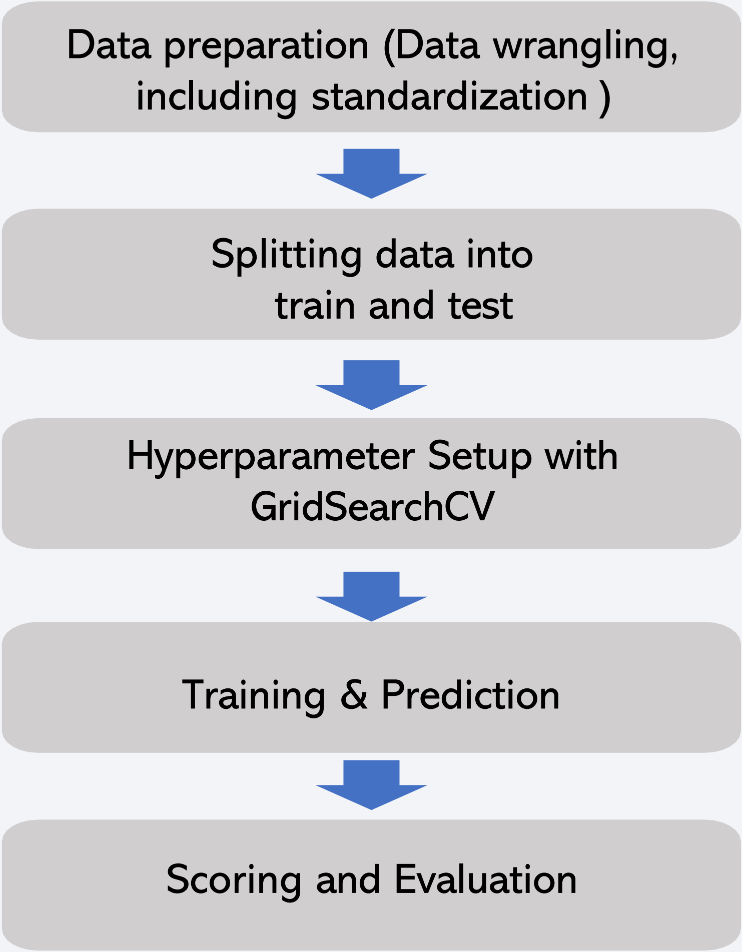
GitHub URL : https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX/blob/main/Build_a_Dashboard_Application_with_Plotly_Dash_w.ipynb

Predictive Analysis (Classification)

Major step:

- **Create** NumPy array from the Class column
- **Standardize** the data with StandardScaler. Fit and transform the data.
- **Split** the data using train_test_split
- **Create** a GridSearchCV object with cv=10 for parameter optimization
- **Apply** GridSearchCV on different algorithms: logistic regression, support vector machine, decision tree , K-Nearest Neighbor
- **Calculate** accuracy on the test data for all models
- **Assess** the confusion matrix for all models
- **Identify** the best model using Accuracy

GitHub URL : https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1-wg.ipynb



Results

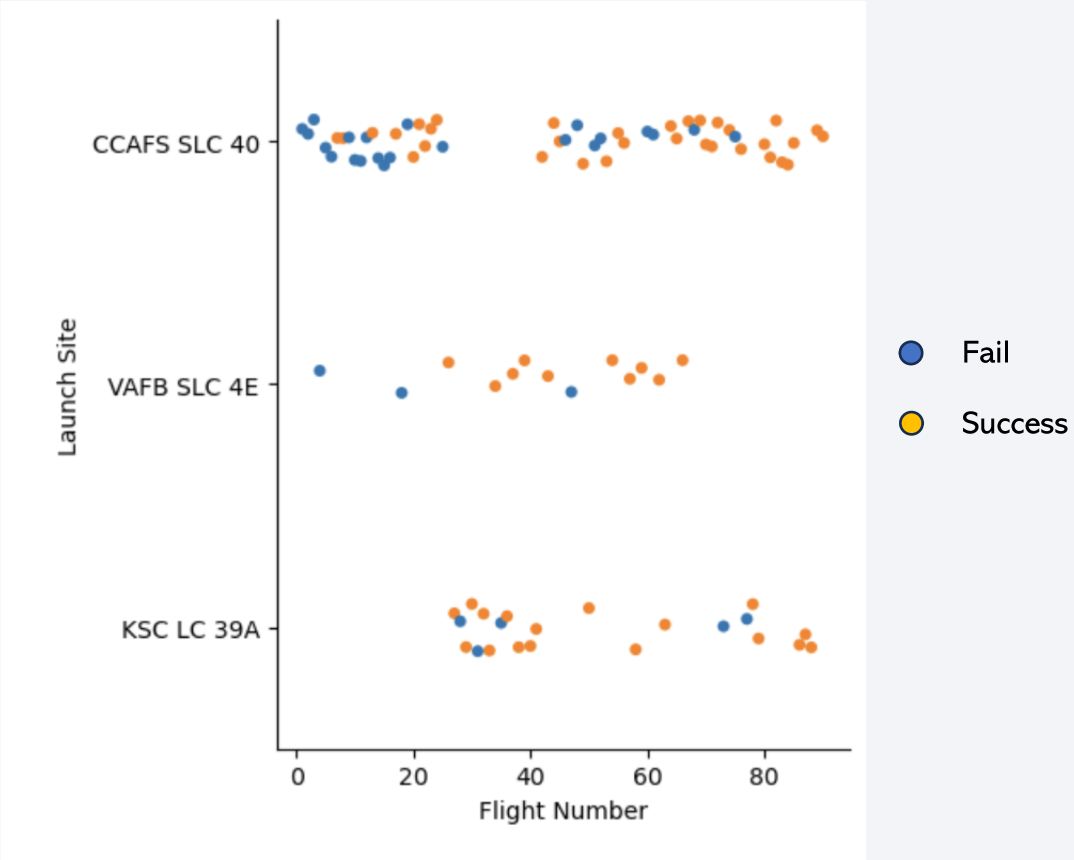
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

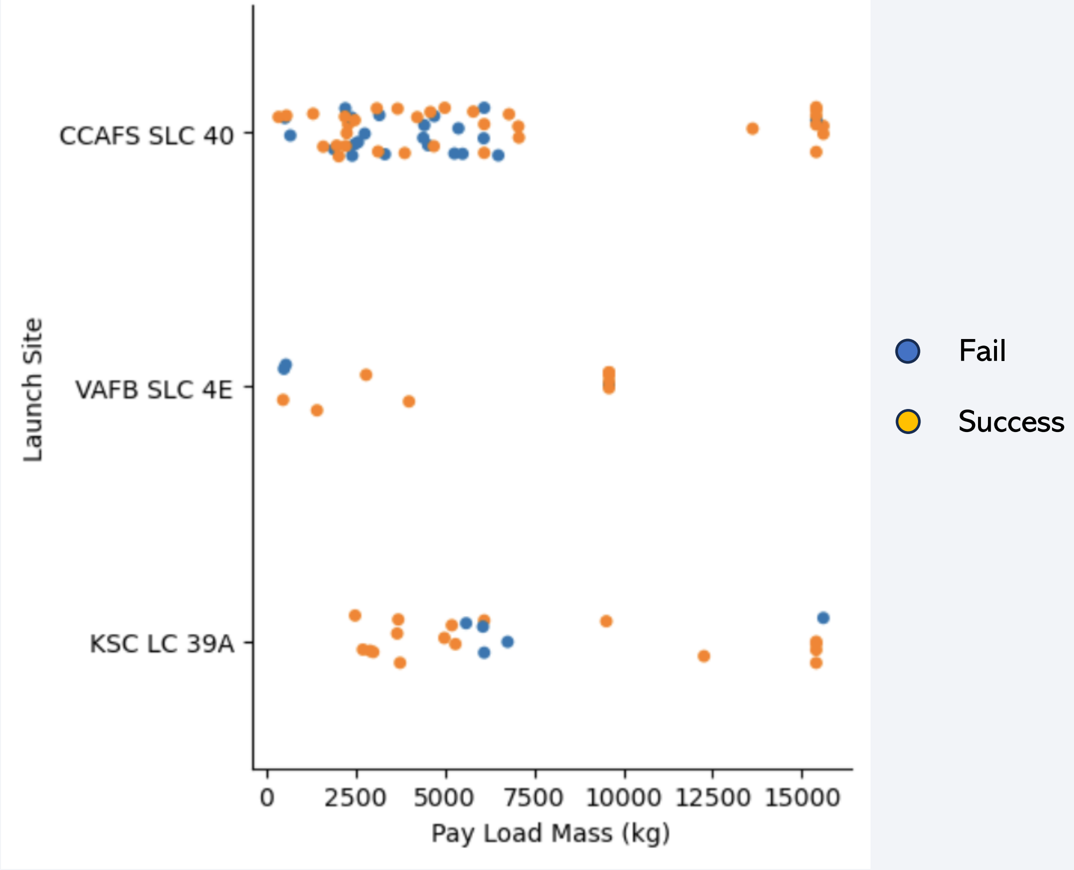
Flight Number *vs.* Launch Site

- Around half of launches were from CCAFS SLC 40 launch site
- Later flights had a higher success rate
- After flight number 80, all launches are successful



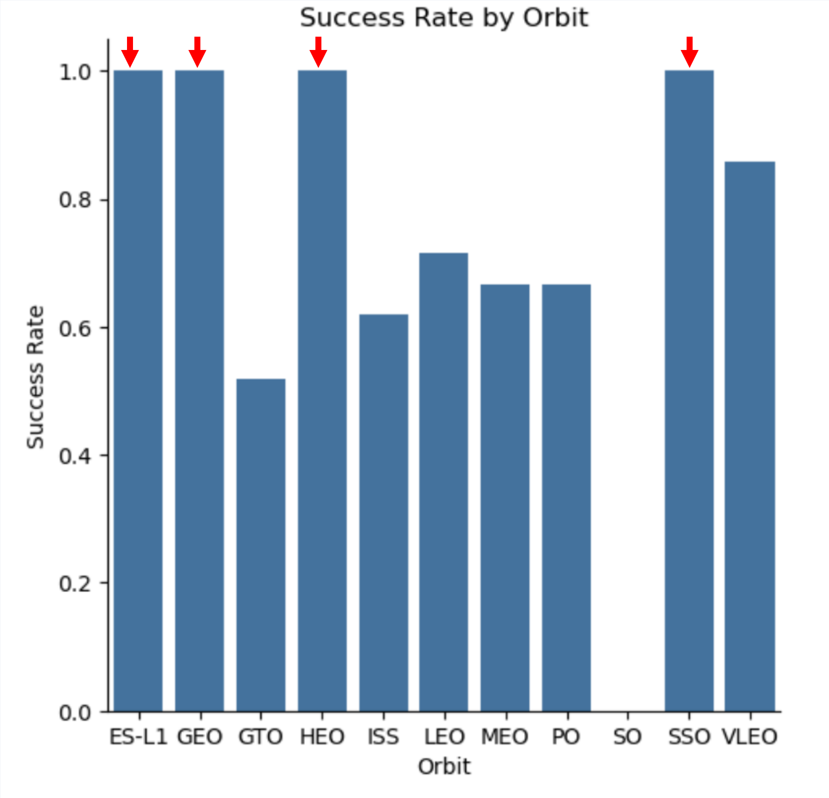
Payload *vs.* Launch Site

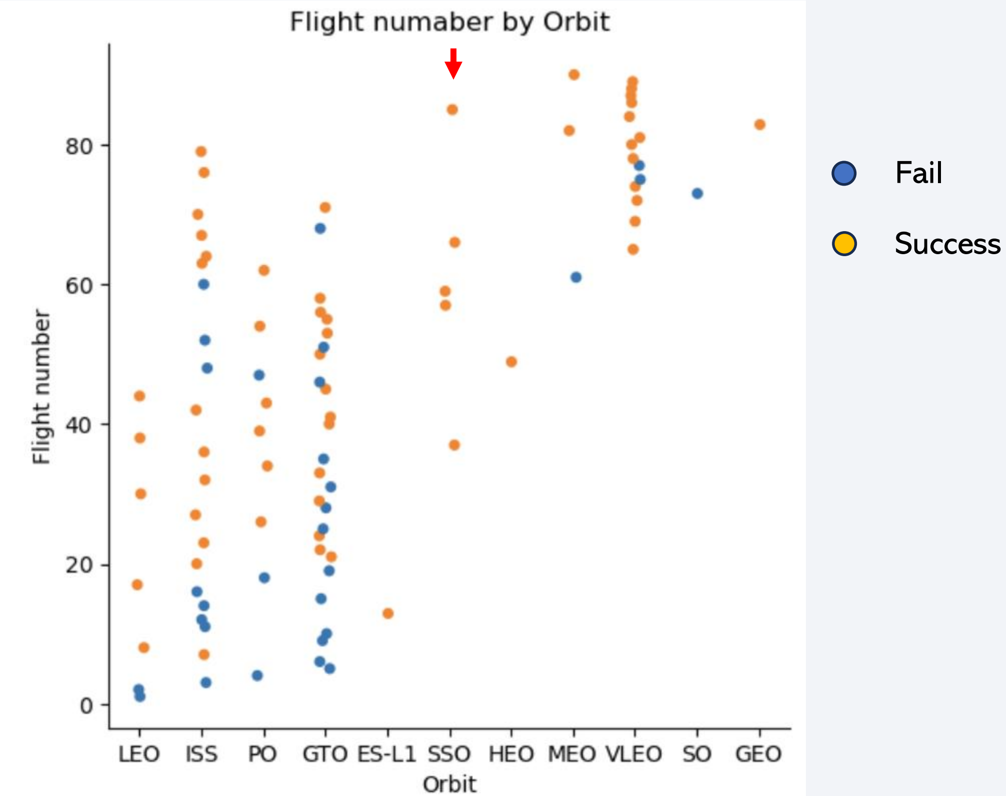
- Most launces with a payload greater than 7,000 kg had successful landing
- KSC LC 39A had a 100% success rate for launches less than 5,500 kg



Success Rate *vs.* Orbit Type

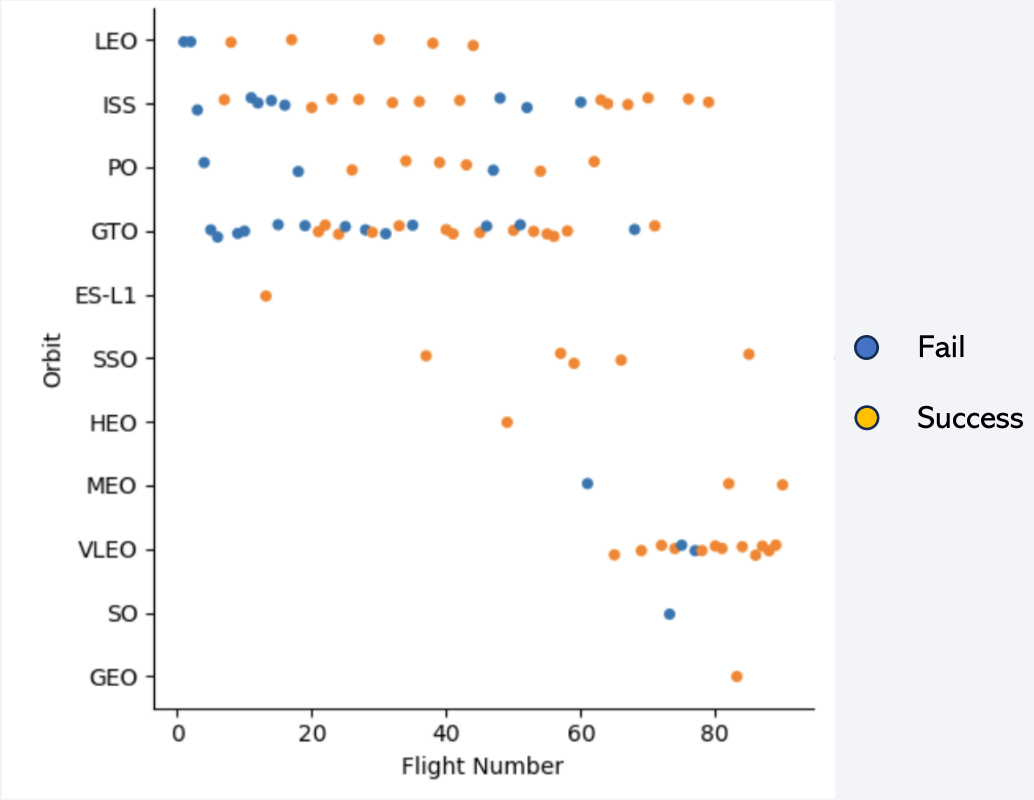
- Although ES-L1, GEO, HEO and SSO had 100% Success Rate (left figure), ES-L1, GEO and HEO only had one flight (right figure). Therefore, the highest success rate of SSO is more trustable





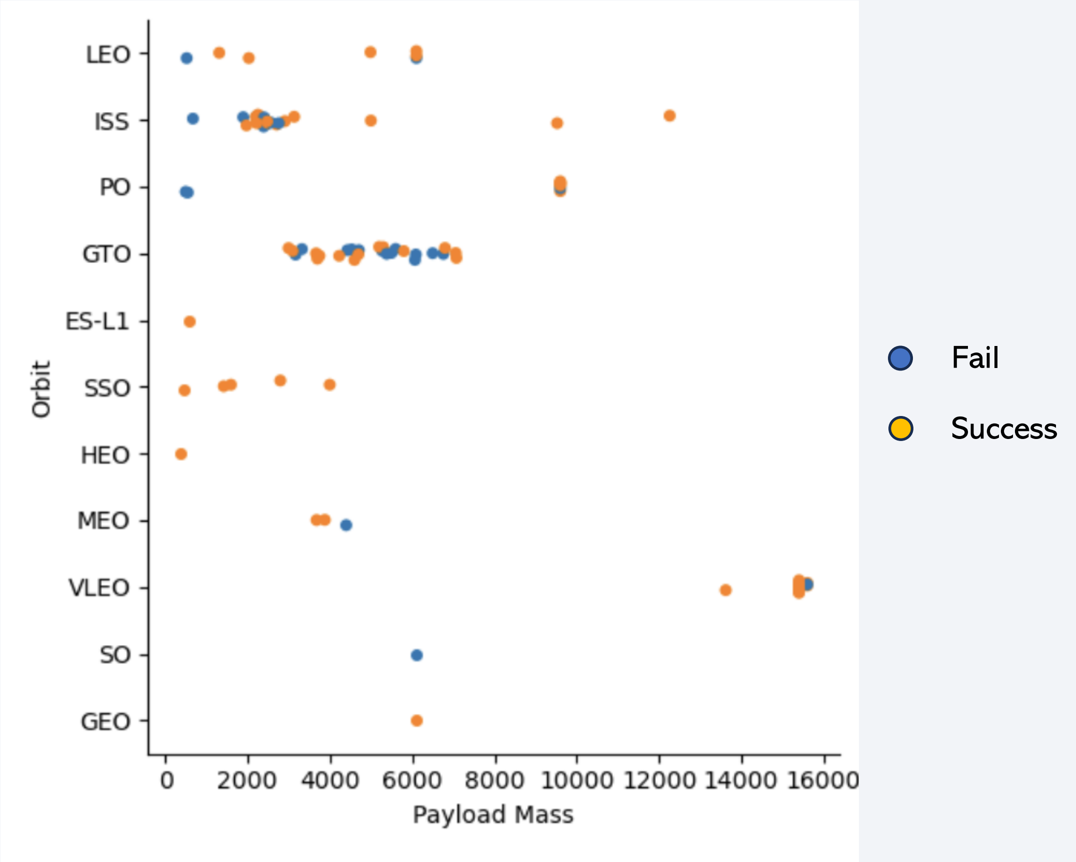
Flight Number *vs.* Orbit Type

- There was a tread that the success rate increased with the number of flights for each orbit
- All 5 flights had successful landing for the SSO orbit
- After 2 flights, all the rest flights had successful landing for the LEO orbit
- The GTO orbit does not follow this trend



Payload *vs.* Orbit Type

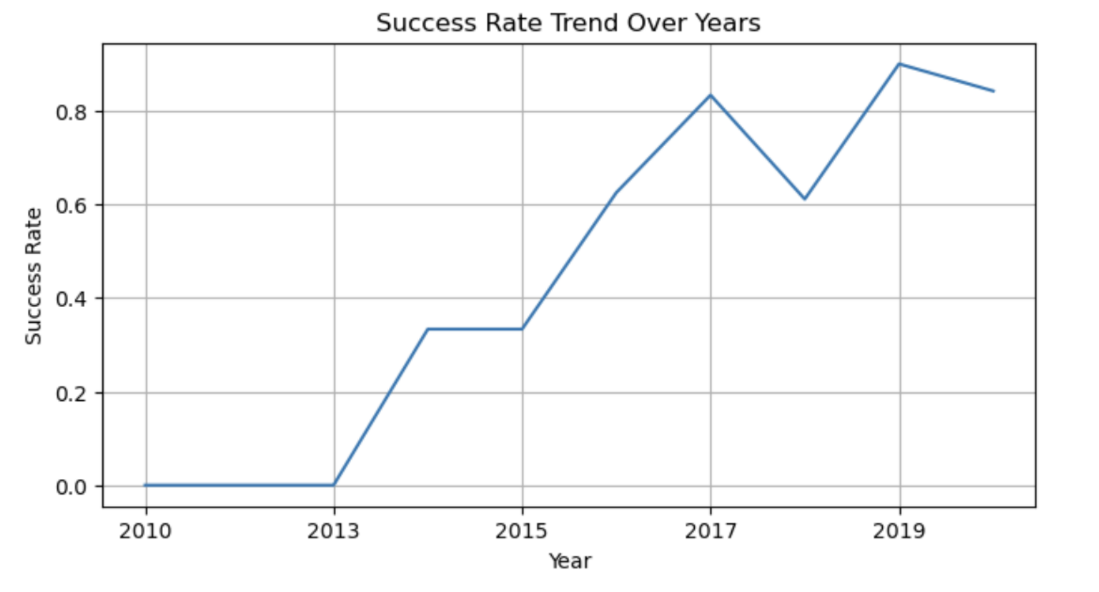
- Heavy payloads are better with LEO, ISS and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

The success rate

- Been improved since 2013 overall
- Dropped between 2017 and 2019
- Been higher than 0.6 since 2016
- Never been higher than 0.9



All Launch Site Names

- There are 4 launch sites in total

```
%sql SELECT DISTINCT "Launch_Site"FROM SPACEXTABLE
```

* sqlite:///my_data1.db
Done.

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

Launch Site Names Begin with 'CCA'

- Two sites' names begin with "CCA" (CCAFS LC-40, CCAFS SLC-40)
- 5 records where launch sites begin with `CCA` are shown below

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- **45,596 kg** (total) payload carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer='NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- **2,535 kg** (average) carried by booster version F9 v1.1

```
%%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_Payload_Mass
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
```

Avg_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

- **The date when the first successful landing outcome in ground pad was achieved is 12/12/2015**

```
%%sql SELECT MIN(Date)
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 4 boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
  AND PAYLOAD_MASS__KG_> 4000
  AND PAYLOAD_MASS__KG_< 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failed Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

```
%%sql SELECT Mission_Outcome, COUNT(*) AS Count
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- There are total 12 distinct booster versions which have carried the maximum payload mass

```
%%sql SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

```
 * sqlite:///my_data1.db
Done.
```

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

In 2015

- Two landing on drone ship failed
- The month, date, booster version, launch site were listed (right table)

```
%%sql SELECT
    substr(Date, 6, 2) AS Month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE substr(Date, 1, 4) = '2015'
  AND Landing_Outcome = 'Failure (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

- **The count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order were listed in right table**

```
%%sql SELECT Landing_Outcome, COUNT(*) AS Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

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

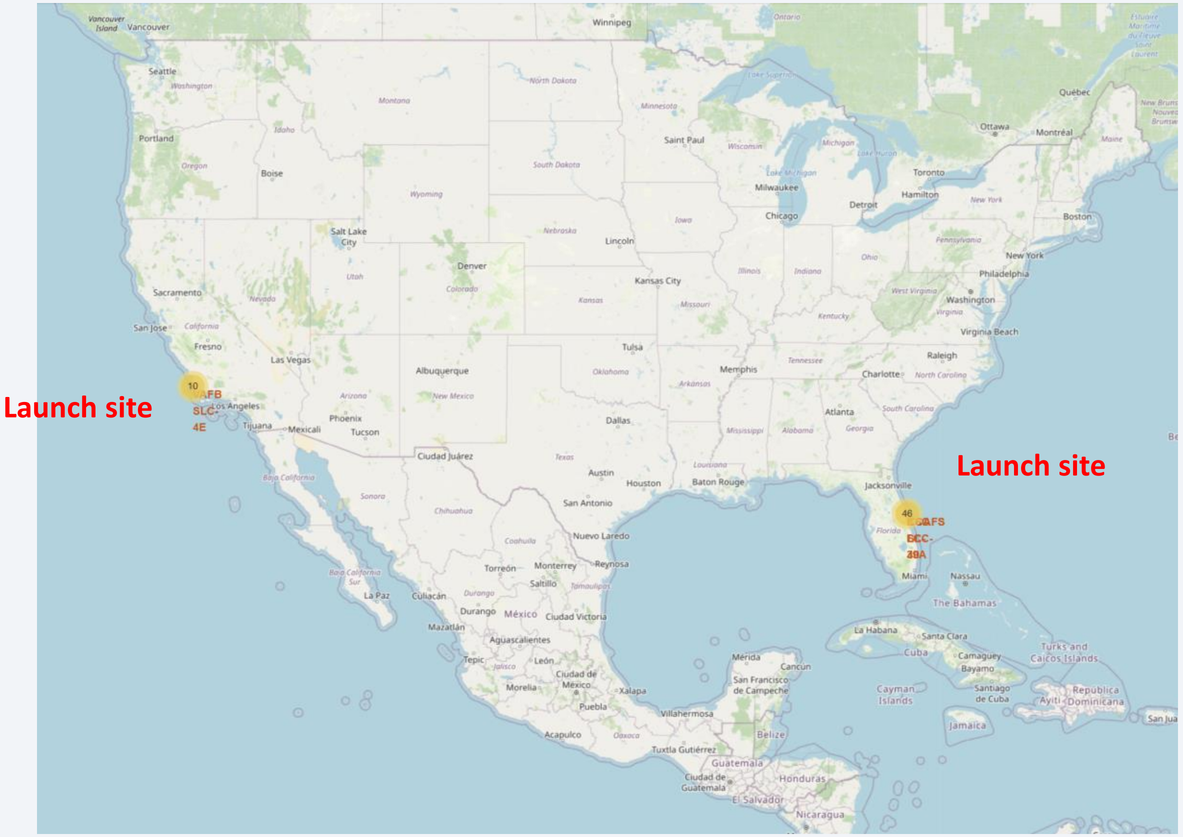
Section 3

Launch Sites Proximities Analysis

Launch Sites Distribution on Map

Observations

- Site VAFB SLC-4E is in California
- Sites CCAFS LC-40, CCAFS SLC-40, KSC LC-39A are in Florida
- All are close to the sea:
  - It can help transportation, launching safety and rocket recycling
- All are near Equator:
  - It can help rockets to get an additional natural boost - due to the rotational speed of earth, which will save cost on fuel and boosters

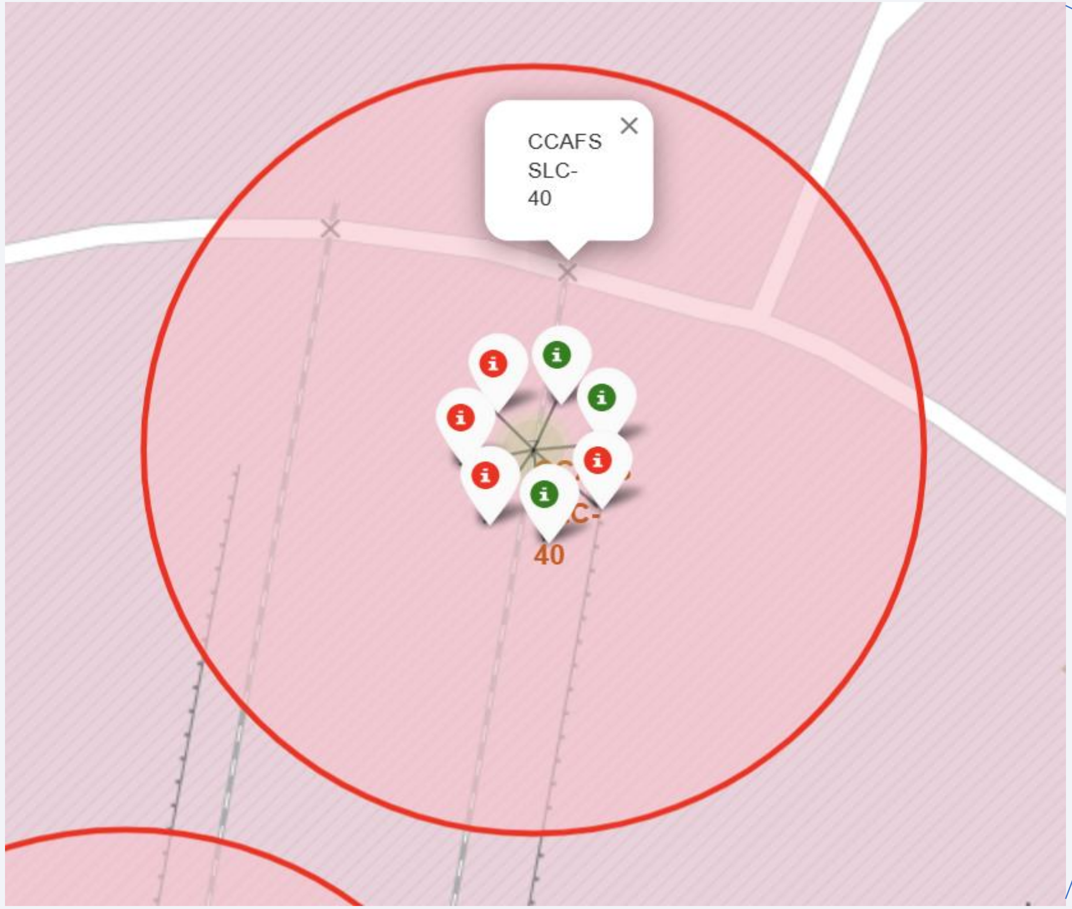


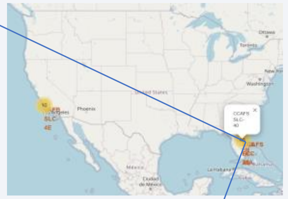
Launch Outcome in Site CCAFS SLC-40

Outcomes:

- Green markers for successful launches
- Red markers for unsuccessful launches
- Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

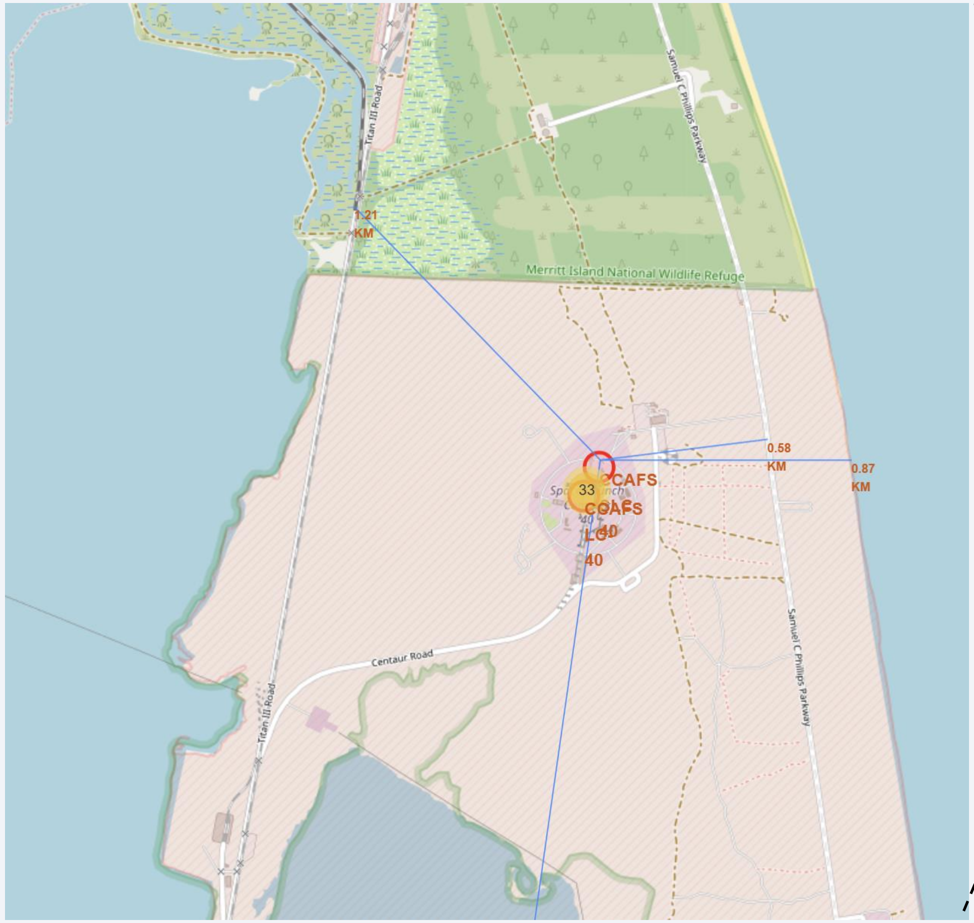
CCAFS SLC-40 is a launching site in Florida





Distance Between Launch Site to Its Proximities

- **Site CCAFS SLC-40**
  - .87 km from nearest coastline
  - 1.21 km from nearest railway
  - 23.23 km from nearest city
  - 19.87 km from nearest highway
- **The location is idea for transportation and safety consideration**

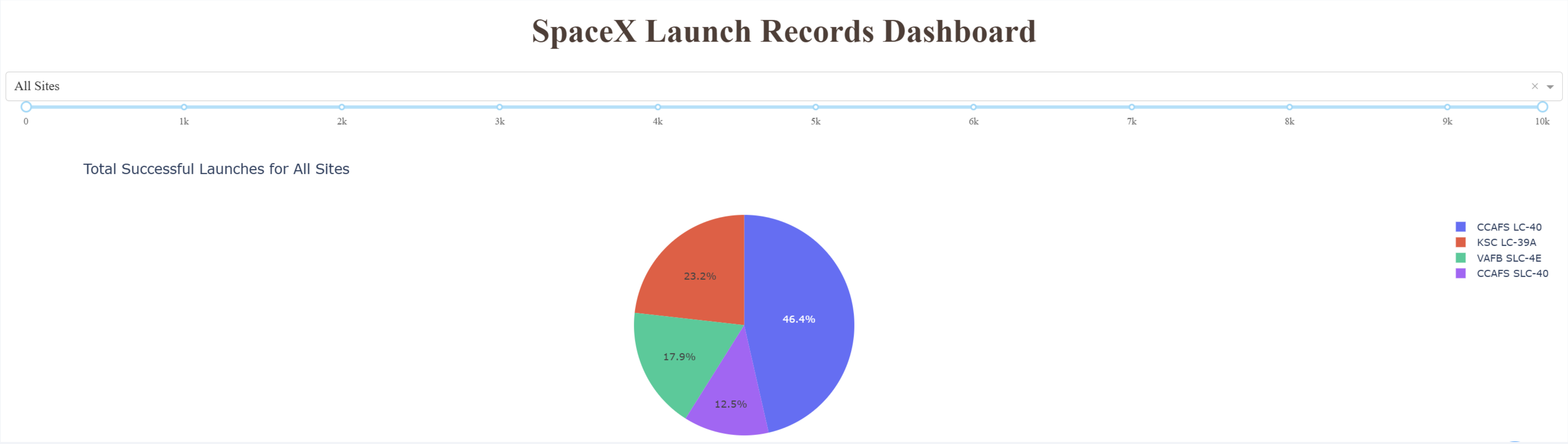


Section 4

Build a Dashboard with Plotly Dash

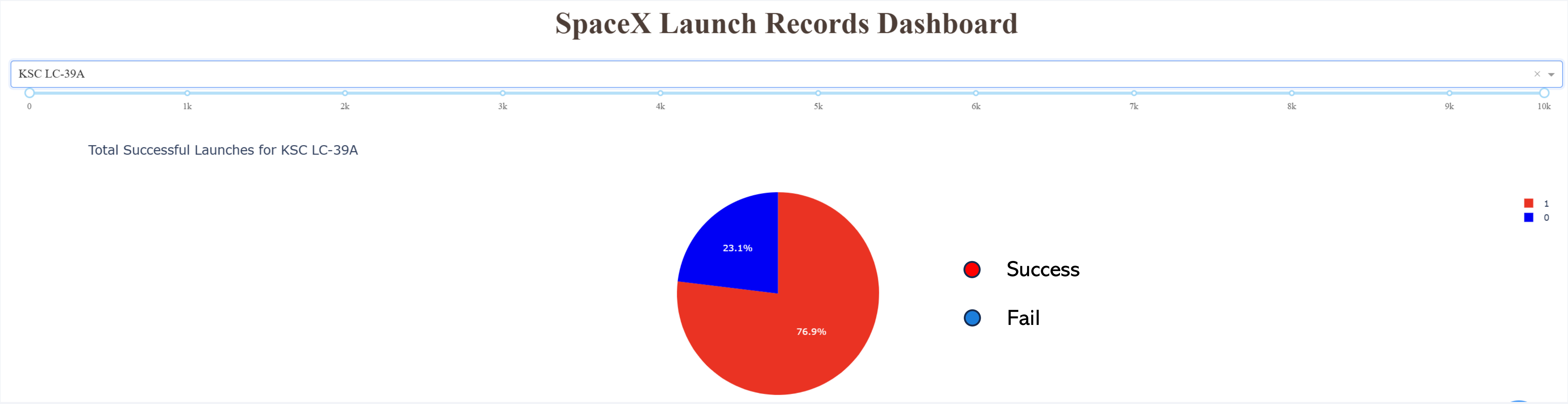
Launch Success by Site

- **CCAPS LC-40** has the **most successful launches** among launch sites (**46.4%**)



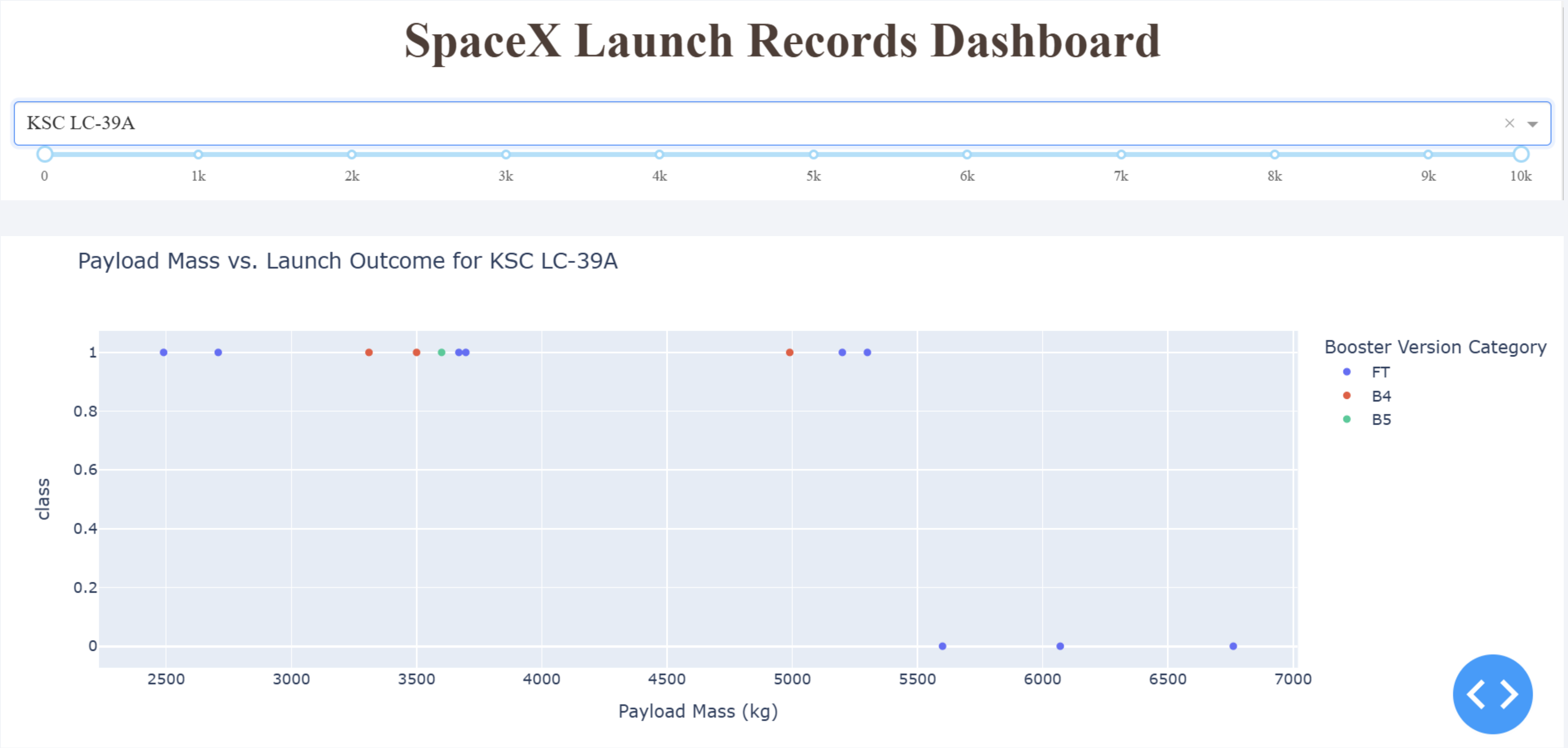
Launch Success (KSC LC-39A)

- **CCAFS KSC LC-39A**has **42.9% success launch ratio, which is the highest among all 4 sites**



Payload Mass and Success

- **It seems KSC LC-39 has higher success when the payload <5500**
  - 1 indicating successful outcome and 0 indicating an unsuccessful outcome

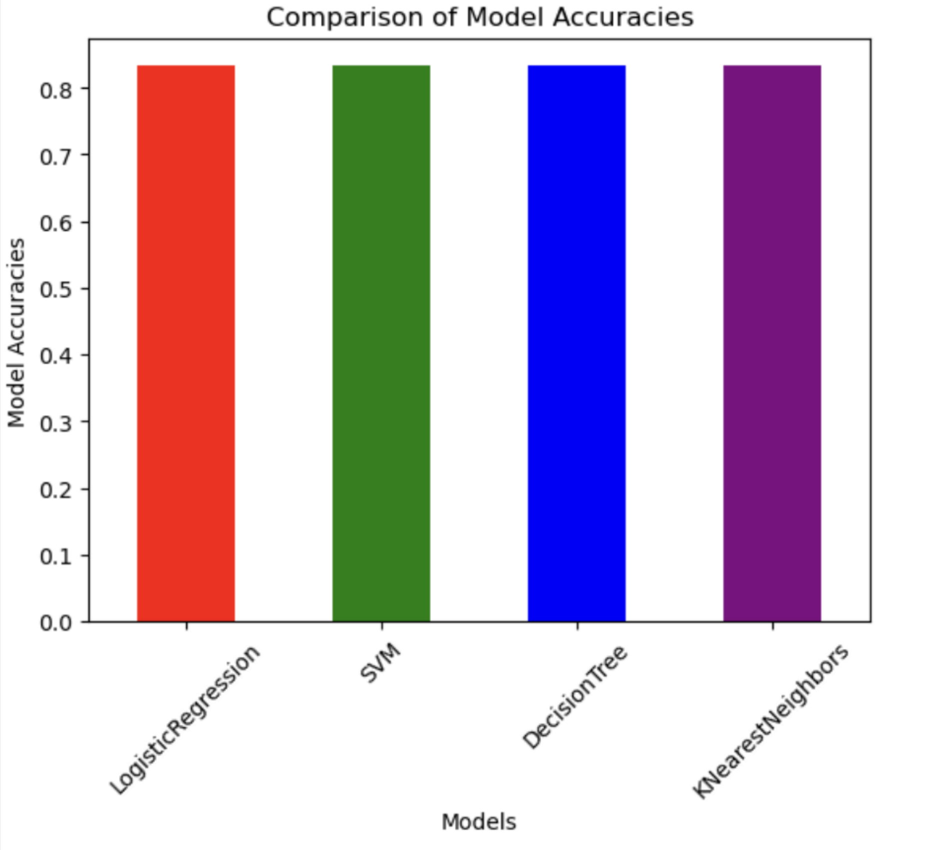


Section 5

Predictive Analysis (Classification)

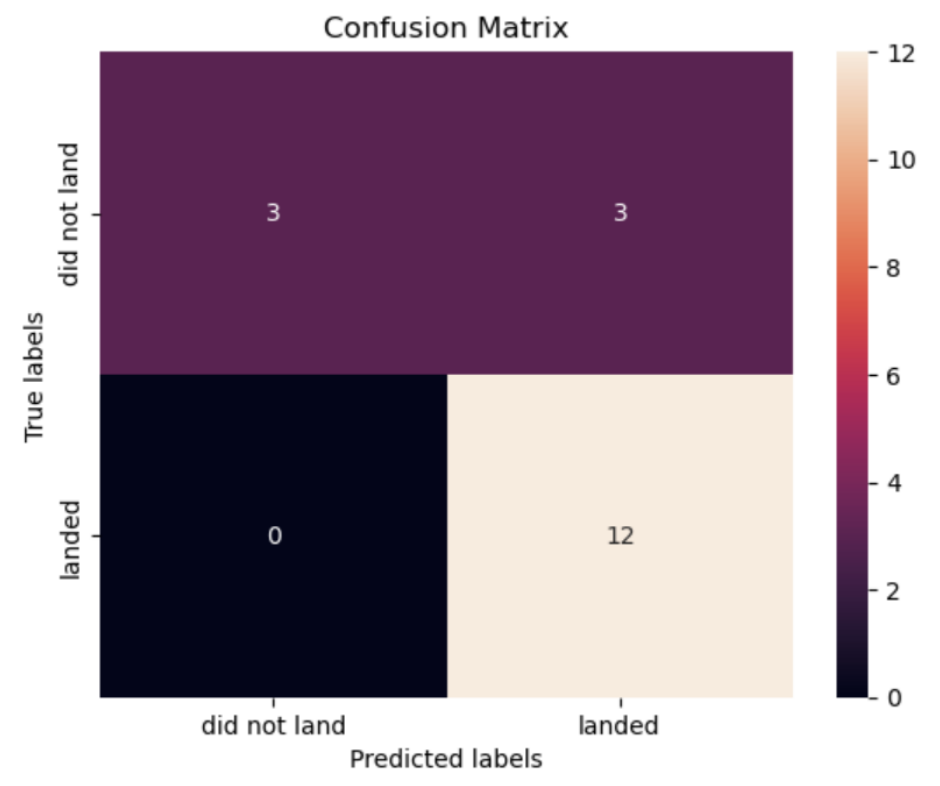
Classification Accuracy

- All 4 models have same accuracies on test data



Confusion Matrix

- **Confusion matrices of Decision Tree on the test data (right figure)**
- **Confusion Matrix Outputs:**
  - 12 True positive
  - 3 True negative
  - 3 False positive
  - 0 False Negative
- **Precision** = TP / (TP + FP)
  - 12 / 15 = .80
- **Recall** = TP / (TP + FN)
  - 12 / 12 = 1
- **F1 Score** = 2 * (Precision * Recall) / (Precision + Recall)
  - 2 * (.8 * 1) / (.8 + 1) = .89
- **Accuracy** = (TP + TN) / (TP + TN + FP + FN) = .833



Conclusions

- The launch success rate has steadily increased over time, indicating continuous advancements in rocket technology
- Various factors, including orbit type, launch site, and payload mass, appear to influence success rates. Expanding the dataset and incorporating more detailed variables will help validate and better understand these patterns
- All four machine learning models achieve over 80% accuracy on the test set, highlighting the promising role of data science in commercial space launches

Appendix

All code and results can be found in the following link

- https://github.com/Notes610041/Applied-Data-Science-Capstone-SapceX

Thank you!