IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jennifer Mueller
10 February 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection and Wrangling
  - Exploratory Data Analysis
  - Folium Interactive Map Exploration
  - Interactive Dashboarding via Plotly
  - Predictive Analysis via Classification Algorithms

- Summary of all results
  - Exploratory Data Analysis
  - Interactive Dashboarding
  - Machine Learning, Predictive Analysis

Introduction

Space Y is a new player in the rocket launching space, joining such notable companies as SpaceX, Rocket Labs, and Virgin Galactic. Space X is the most cost-competitive of these companies, due in large part to its ability to reuse their rockets’ first stage.

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of $62 million, compared to other providers whose cost can run upward of $165 million.
As a new company in this space, if we can determine the ability of the first stage to land, the cost of a launch can be predicted.

In this analysis, we consider:
- The relationship of variables on the success of SpaceX to land its first stage
- The environmental conditions necessary for a successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Pandas dataframes were created from information collected directly via SpaceX API (https://api.spacexdata.com/v4/) and webscraping on the Falcon 9 Wikipedia page.
- Perform data wrangling
  - Landing Outcomes were normalized into a Class column where 0 was an unsuccessful first stage landing and 1 was successful
  - The success rate for each launch site was calculated
  - Missing data was replaced with the mean of available results
  - One Hot Encoding was used to convert categorical values to binary values
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Multiple classification algorithms were trained, including Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors
  - Datasets were split with a 80% train and 20% test size

Data Collection – SpaceX API

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

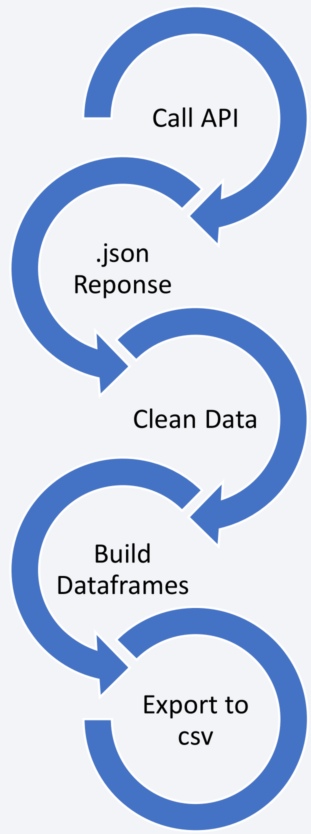
```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
data=pd.json_normalize(response.json())
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc']]
```

2. Create a primary source dataframe from a customized launch dictionary for further analysis

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
new_df = pd.DataFrame(launch_dict)
```

3. Filter dataframe for Falcon9 launches

```
data_falcon9 = new_df[(new_df.BoosterVersion=='Falcon 9')]
```



Github link

Data Collection - Scraping

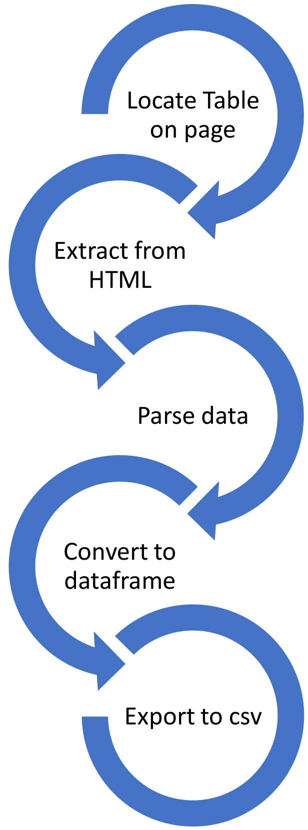
1. Request page from its URL

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
data = requests.get(static_url).text
soup = BeautifulSoup(data,'html.parser')
```

2. Extract column/variable names from header

```
html_tables = soup.findAll('table')
html_tables
first_launch_table = html_tables[2]
for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if (name != None and len(name) > 0):
        column_names.append(name)
print(column_names)
```

```
['Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome']
```



[Github Webscraping]

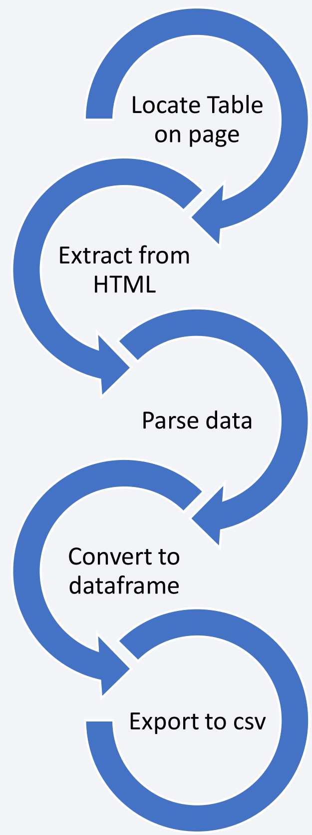
Data Collection – Scraping (con’t)

3. Create dataframe by parsing tables

```
launch_dict= dict.fromkeys(column_names)
```

Sample for Flight No. feature

```
for rows in table.find_all("tr"):
    #check to see if first table heading is as number corresponding to launch a number
    if rows.th:
        if rows.th.string:
            flight_number=rows.th.string.strip()
            flag=flight_number.isdigit()
    else:
        flag=False
    #get table element
    row=rows.find_all('td')
    #if it is number save cells in a dictonary
    if flag:
        extracted_row += 1
        # Flight Number value
        # TODO: Append the flight_number into launch_dict with key `Flight No.` - CHECK
        launch_dict['Flight No.'].append(flight_number)
        #print(flight_number)
        datatimelist=date_time(row[0])
```



[Github Webscraping]

Data Wrangling

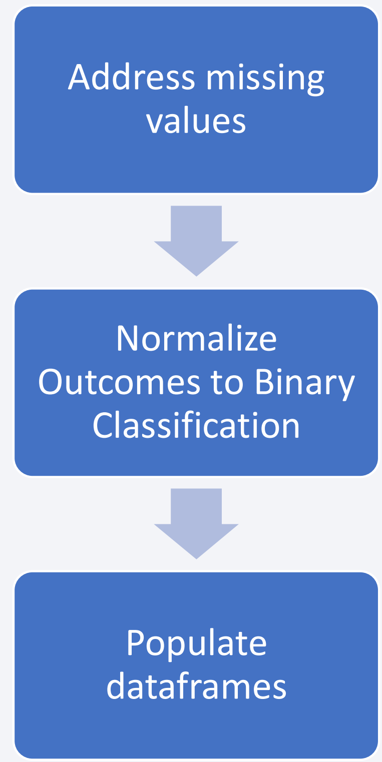
- API Data – missing values for Payload Mass (kg) were replaced by the mean

```
data_falcon9.isnull().sum()

FlightNumber     0
Date             0
BoosterVersion   0
PayloadMass      5
```

```
print ("The Payload Mass mean is: ",data_falcon9[['PayloadMass']].mean())

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].fillna(6123.547647)
```

- Webscraping – fill in parsed launch records

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



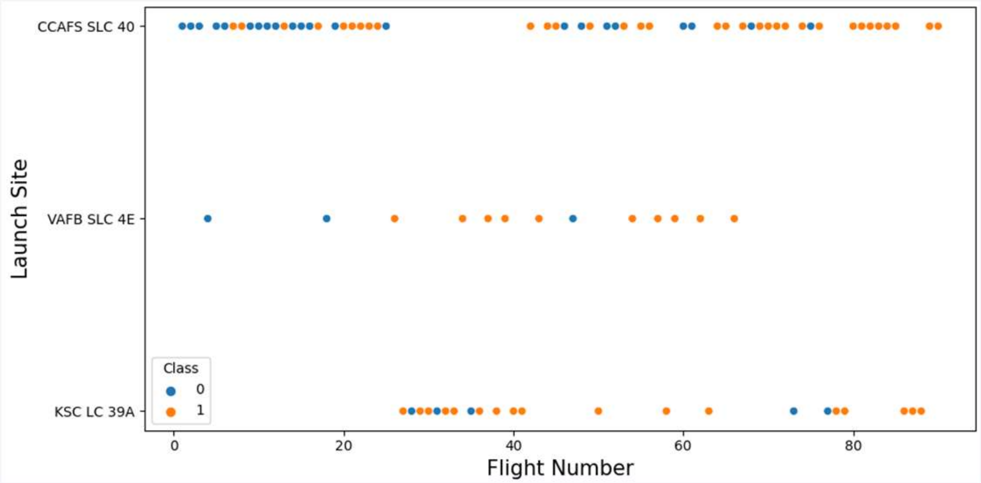
[Github API Wrangling](#) [Github Webscraping](#)

EDA with Data Visualization

Flight number represents continuous launches. A scatter plot identified that as flight number increases, the first stage is more likely to land successfully; it seems the more massive the payload, the less likely the first stage will return. Additional plots were created to drill into those details

Name	Type	Summary Results
Launch Site to Flight No (fig 1)	Catplot	Different launch sites display different success rates. Plot displays as a series to easily show success rates over time
Launch Site to Payload	Scatterplot	Displays as a trend series for success rate of launches based on payload, except at the very highest amounts where there is a higher degree of success
Class to Orbit	Bar	Success rate for each orbit type is displayed side by side for easy comparison

Fig 1



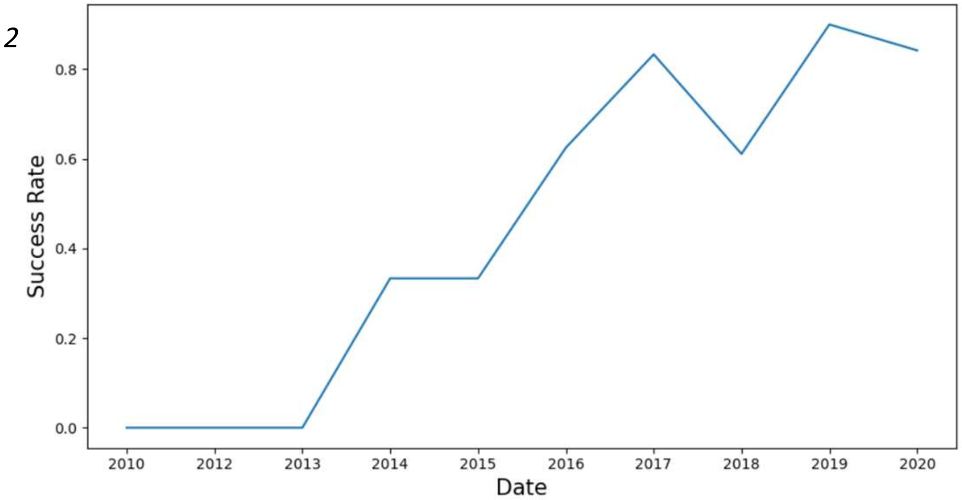
[EDA Notebook]

EDA with Data Visualization (con’t)

Flight number represents continuous launches. A scatter plot identified that as flight number increases, the first stage is more likely to land successfully; it seems the more massive the payload, the less likely the first stage will return. Additional plots were created to drill into those details

Name	Type	Summary Results
Orbit to Flight No	Scatterplot	Displays as a trend series for success rate of launches based on orbit type. Except for LEO orbit, there is no clear relationship demonstrated a higher degree of success
Orbit to Payload	Scatterplot	Displays as a trend series for success rate. With heavy payloads the successful landing or positive landing rate is more frequent for Polar, LEO and ISS orbits
Success Rate by Year (Fig 2)	Line	Over time, the success rate continues to climb

Fig 2



[EDA Notebook]()

EDA with SQL

Multiple queries were run to gather additional information about the datset:

- 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 succesful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- 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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

SQL notebook

Build an Interactive Map with Folium

Interactive maps were produced that marked all launch sites, including successful and unsuccessful launches, and distance calculated to important proximities (railways, highways, cities)

- Latitude and longitude were used to create Circle Markers highlighting the locations of the launch sites
- Launch outcomes were displayed as marker clusters around launch sites to display successful/unsuccessful launches by site
- Distance lines were drawn from launch sites to key proximity areas.
    - Launch sites are generally close to railways, but further away from highways and cities
    - Launch sites are coastal but somewhat north of the equator

Github link

Build a Dashboard with Plotly Dash

An interactive dashboard was built for users to explore payload and and successful launch metrics by site

- Drop down created to allow users to select an individual site for review
- Pie Chart was displayed based on site illustrating the percentage of success/failed launches
- A slider was added allowing the users to control the size of the payload mass when reviewing results
- Scatter graph added to illustrate correlation between payload and launch success

Github link

Predictive Analysis (Classification)

Model Development
Dataset transformed to be read as a numpy array
Data split into train and test sets (80/20)
Determined use of 4 different classification algorithms would be tried: Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbor
Parameters set to GridSearchCV
Fit data to GridSearch CV and trained the dataset

Model Evaluation
For each algorithm, Scikit-learn .score method was used to evaluate accuracy of the test set

Model Improvement was accomplished through feature engineering and algorithm tuning and found the best classification model

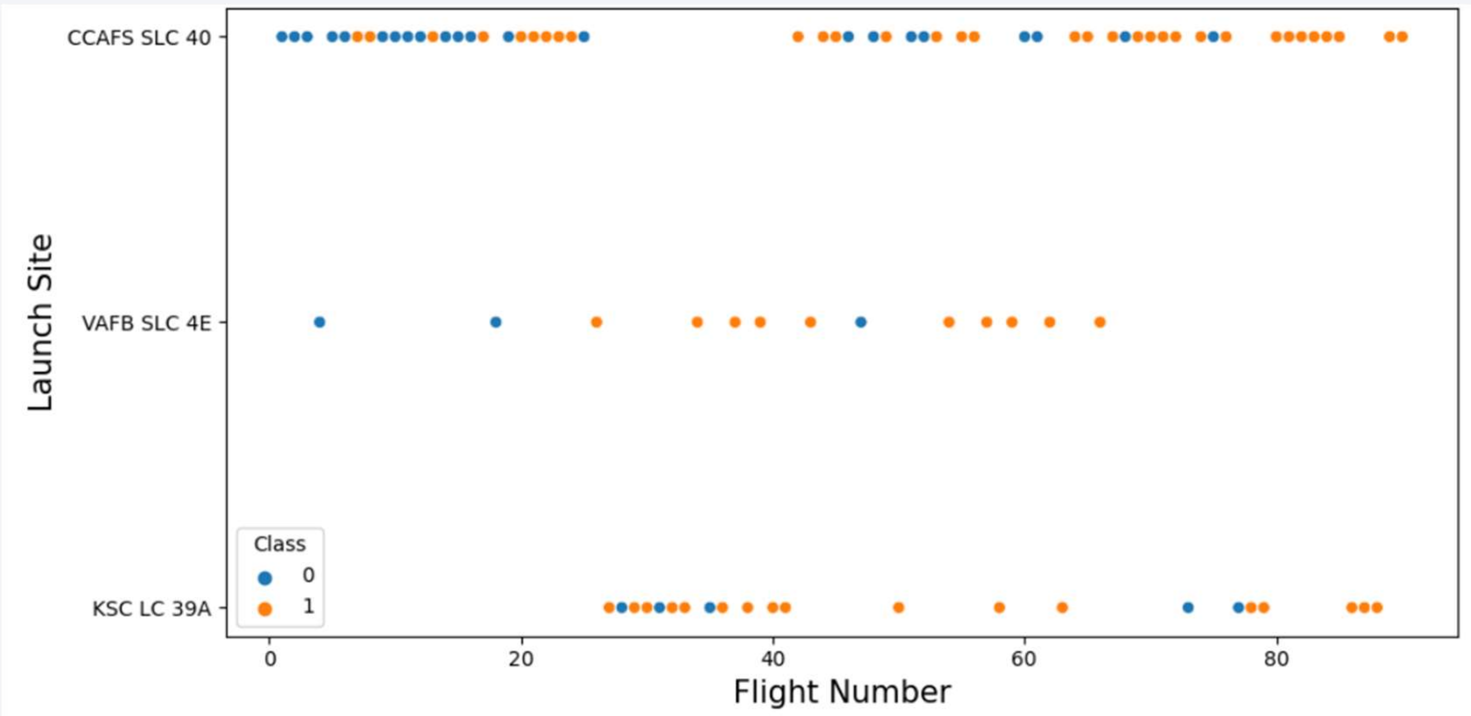
Results

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

Section 2

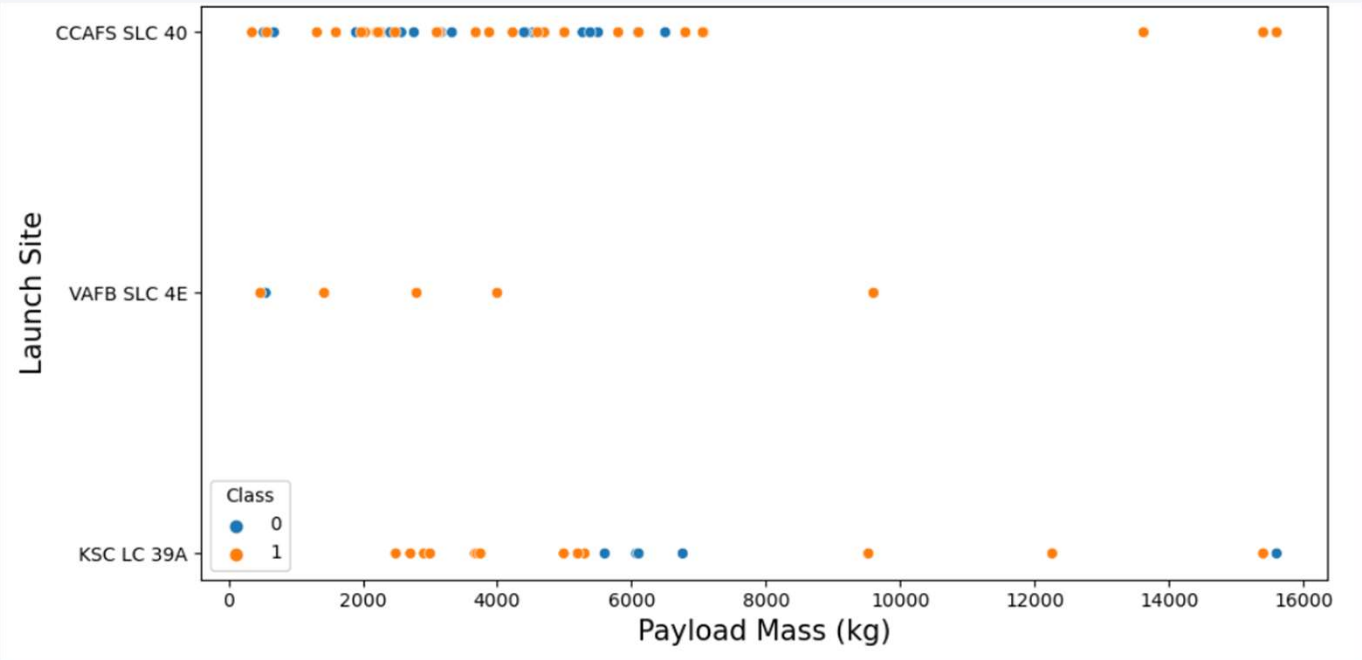
Insights drawn from EDA

Flight Number vs. Launch Site



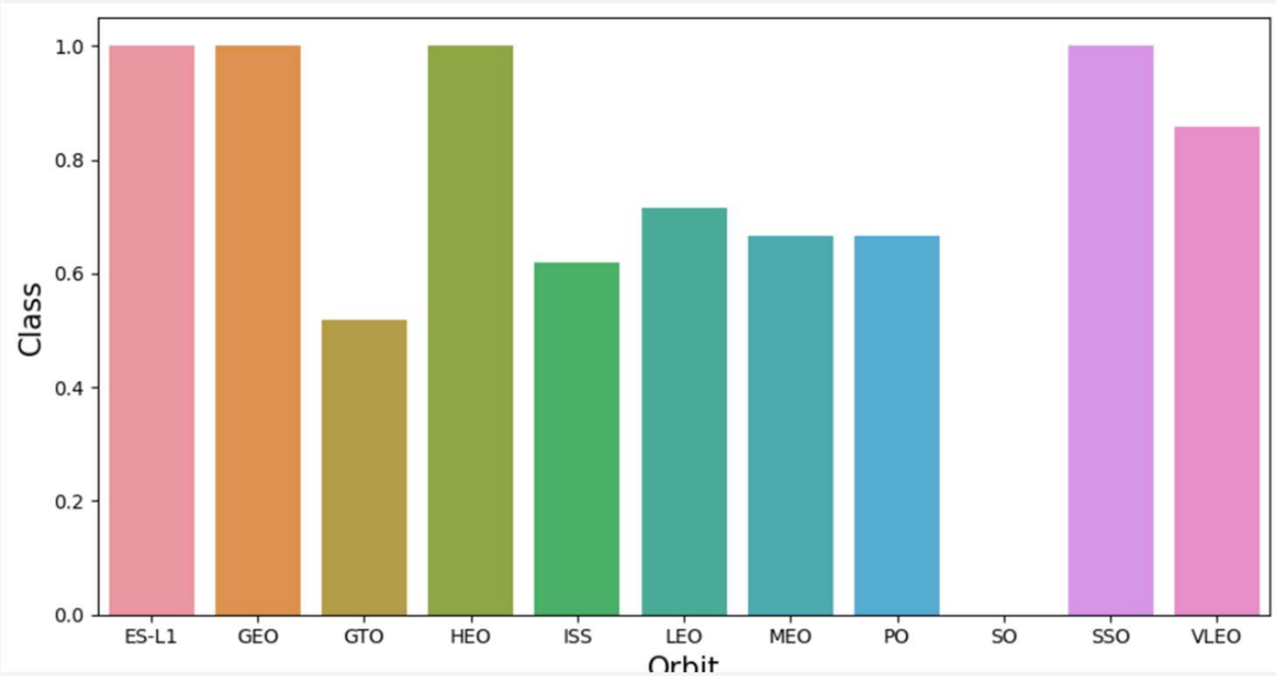
Different launch sites have different success rates. Over time, there are more successes than failures

Payload vs. Launch Site



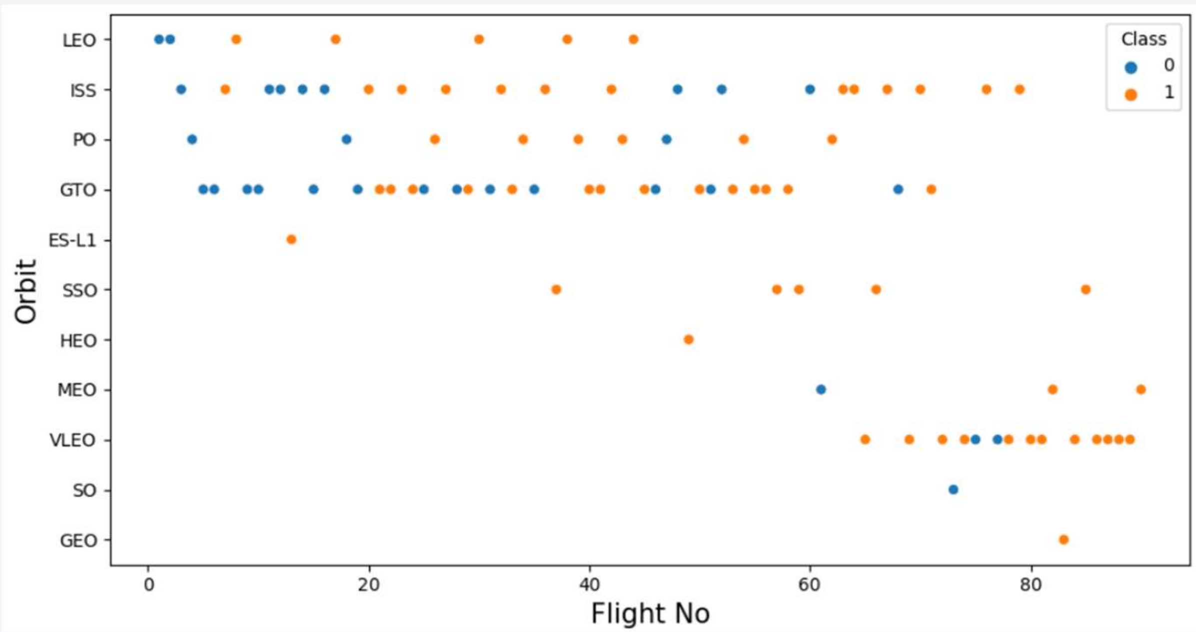
For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000).
There is no clear indicator that payload mass is correlated with a successful landing outcome

Success Rate vs. Orbit Type



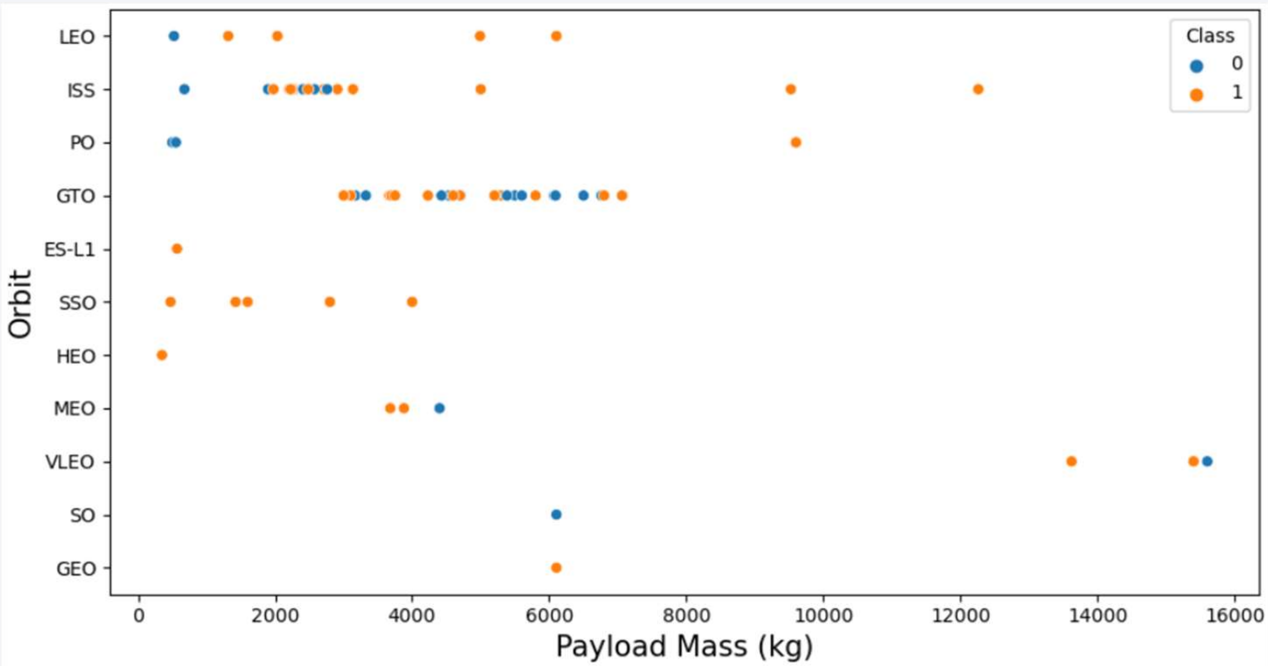
Success rates are mixed for orbit type, with ES-L1, GEO, HEO, and SSO having the highest degrees of success

Flight Number vs. Orbit Type



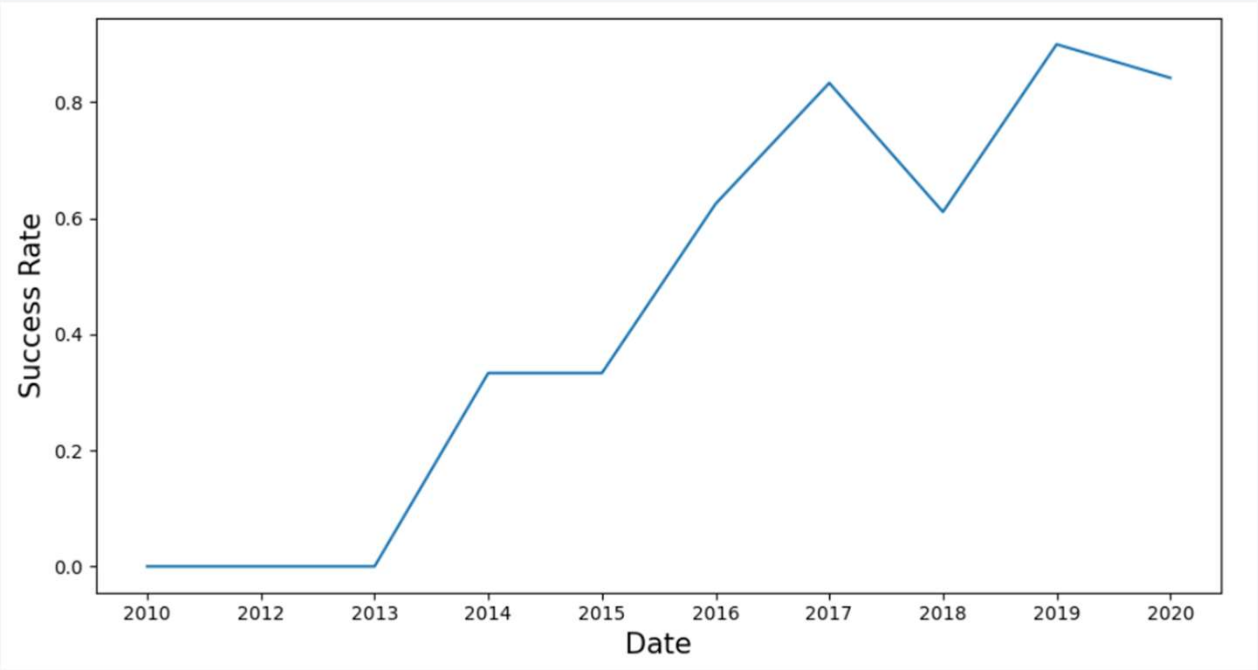
The LEO orbit success appears related to the number of flights. However, for other orbit types there is no strong relationship demonstrated

Payload vs. Orbit Type



With heavy payloads, the successful landing rate is higher for Polar, LEO and ISS. However, GTO is less clear and results more mixed.

Launch Success Yearly Trend



The success rate has been steeply increasing from 2013 through 2020

All Launch Site Names

```
%%sql
select distinct LAUNCH_SITE from spacextbl
```

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

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

- There are 4 distinct launch site location names

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * FROM spacextbl
   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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- This list displays 5 representative records of launch sites beginning with “CCA”

Total Payload Mass

```
%%sql
SELECT sum(payload_mass__kg_) as Total_Payload_Mass_for_NASACRS from spacextbl
    WHERE customer = 'NASA (CRS)'
```

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

Total_Payload_Mass_for_NASACRS
45596

- The total payload carried for NASA (CRS) is 45,596

Average Payload Mass by F9 v1.1

```
%%sql
SELECT avg(payload_mass__kg_) as Avg_Payload_Mass_for_F9 from spacextbl
    WHERE booster_version = 'F9 v1.1'
```

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

Avg_Payload_Mass_for_F9
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2,928.40

First Successful Ground Landing Date

```
%%sql
SELECT min(Date) as First_Successful_Ground_Pad_Landing from spacextbl
    WHERE "Landing _Outcome" = "Success (ground pad)"
```

Date	Landing _Outcome
22-12-2015	Success (ground pad)

- The date of the first successful ground pad landing is 22 Dec 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ from spacextbl
    WHERE "Landing _Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6001
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

These are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) as Total_Count from spacextbl
    GROUP BY MISSION_OUTCOME
```

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

Mission_Outcome	Total_Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Here are the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%%sql
SELECT  Booster_Version, PAYLOAD_MASS__KG_ from spacextbl
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM spacextbl)
```

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

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

Here are the names of the boosters which have carried the maximum payload mass

2015 Launch Records

```
%%sql
select substr(Date,4,2) as Month, Booster_Version, Launch_Site, ("Landing _Outcome") from spacextbl
where ("Landing _Outcome") = "Failure (drone ship)" and substr(Date,7,4) = "2015"
```

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

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

- Here is a list of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%%sql
select Date, ("Landing _Outcome"), count("Landing _Outcome") as TotalCounts from spacextbl
group by ("Landing _Outcome")
order by TotalCounts desc
```

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

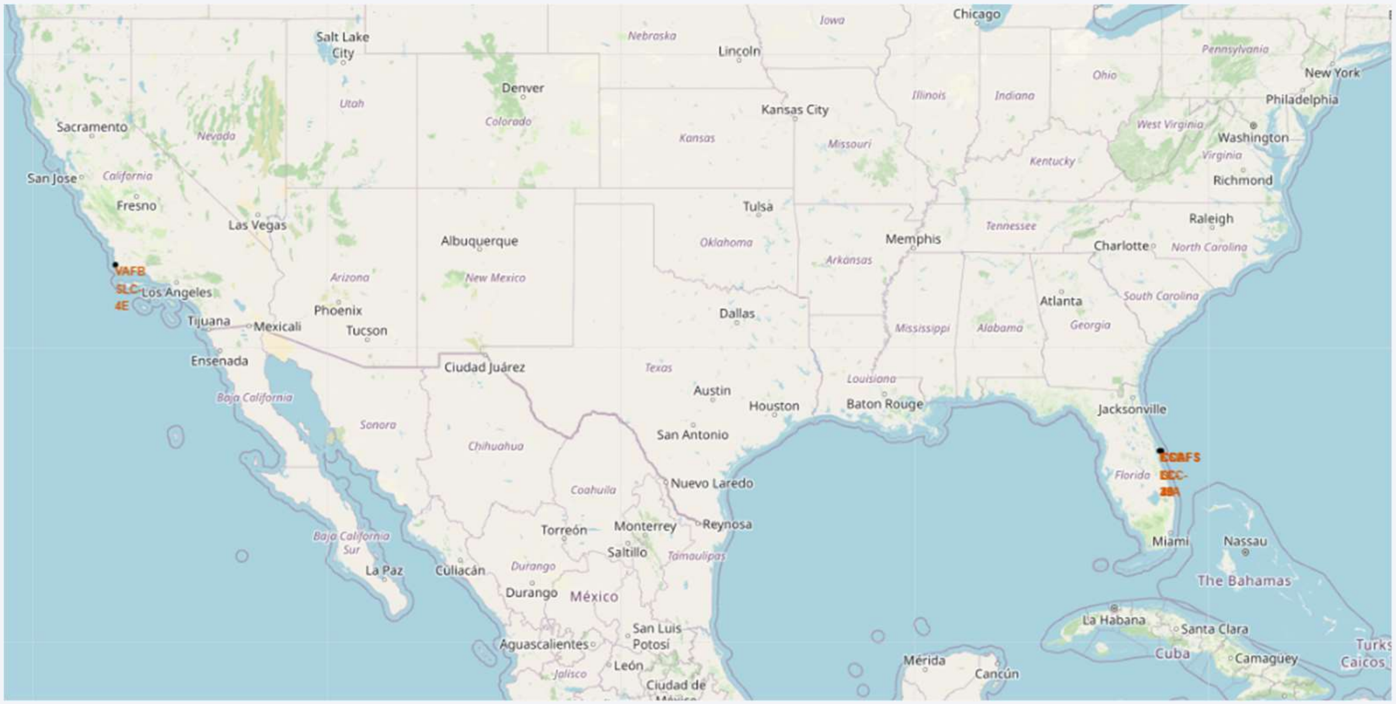
Date	Landing _Outcome	TotalCounts
22-07-2018	Success	38
22-05-2012	No attempt	21
08-04-2016	Success (drone ship)	14
22-12-2015	Success (ground pad)	9
10-01-2015	Failure (drone ship)	5
18-04-2014	Controlled (ocean)	5
05-12-2018	Failure	3
29-09-2013	Uncontrolled (ocean)	2
04-06-2010	Failure (parachute)	2
28-06-2015	Precluded (drone ship)	1
06-08-2019	No attempt	1

Here is a rank of 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

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

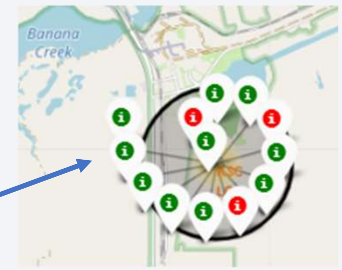


All launch site locations are coastal and north of the equator by a sizeable margin

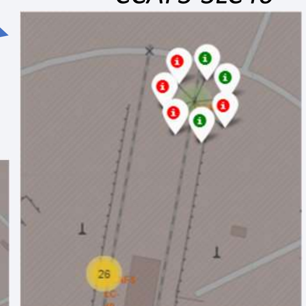
Launch outcomes by site

VAFB-SLC4E

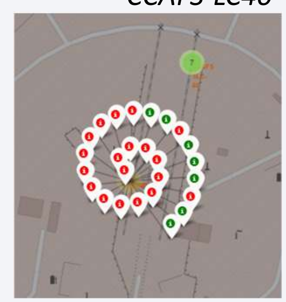
KSC-LC



CCAFS-SLC40



CCAFS-LC40

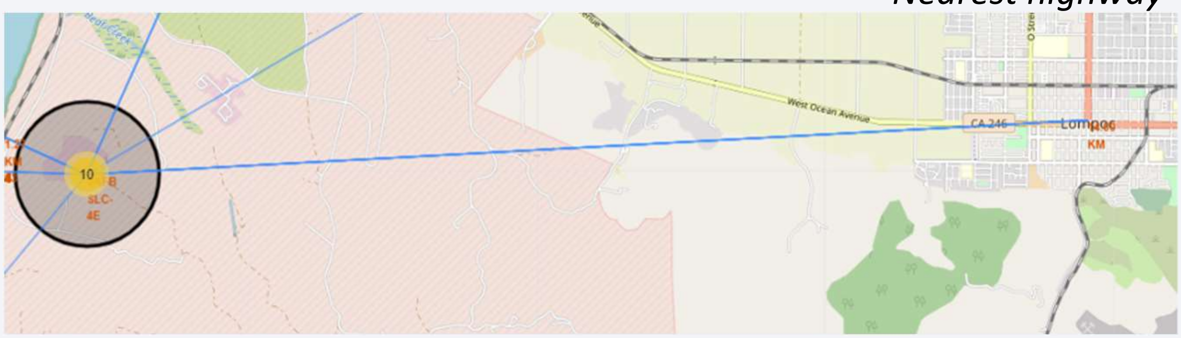


- Red outcome markers are failed launches
- Green outcome markers are successful launches

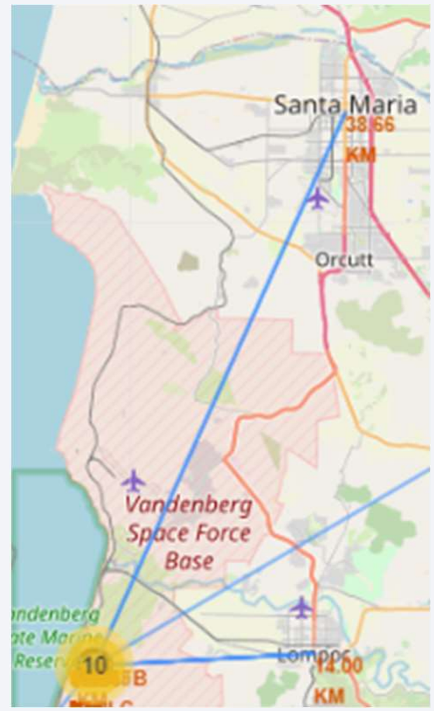
Launchsite distance to landmarks

Nearest railway and coastline

Nearest highway



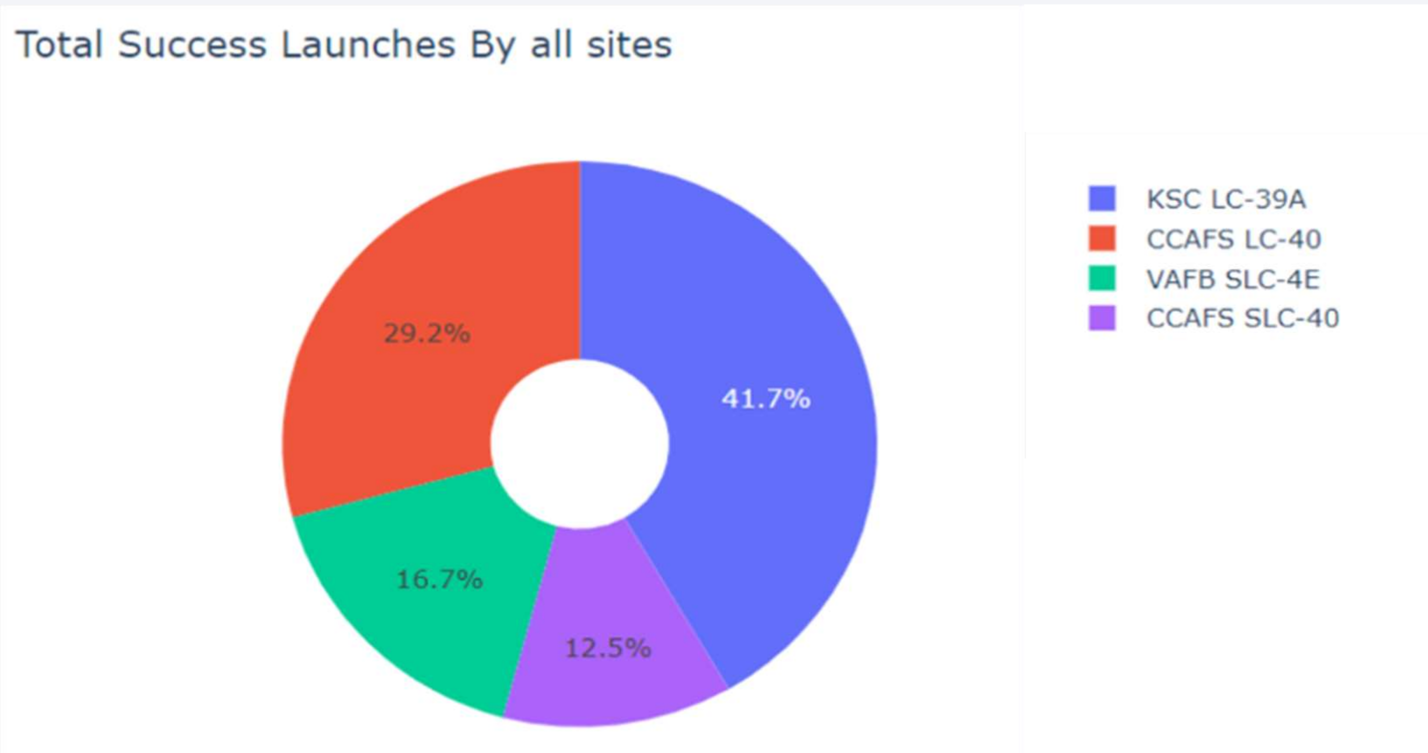
Nearest city



Section 4

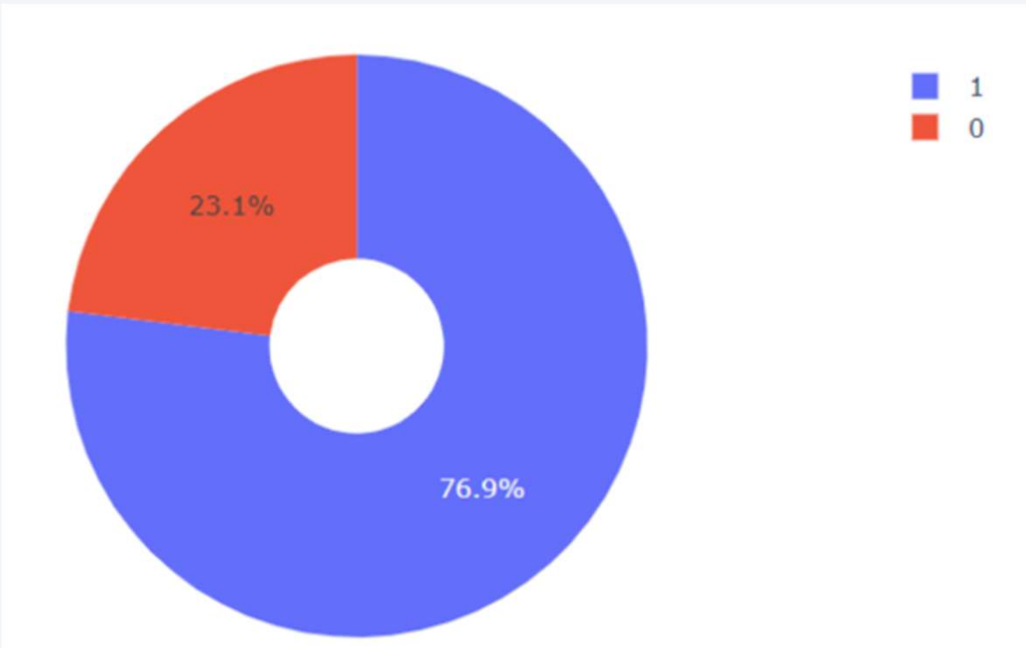
Build a Dashboard with Plotly Dash

Successful landing outcomes by site



KSC LC-39A had the highest number of successful launches compared to other sites

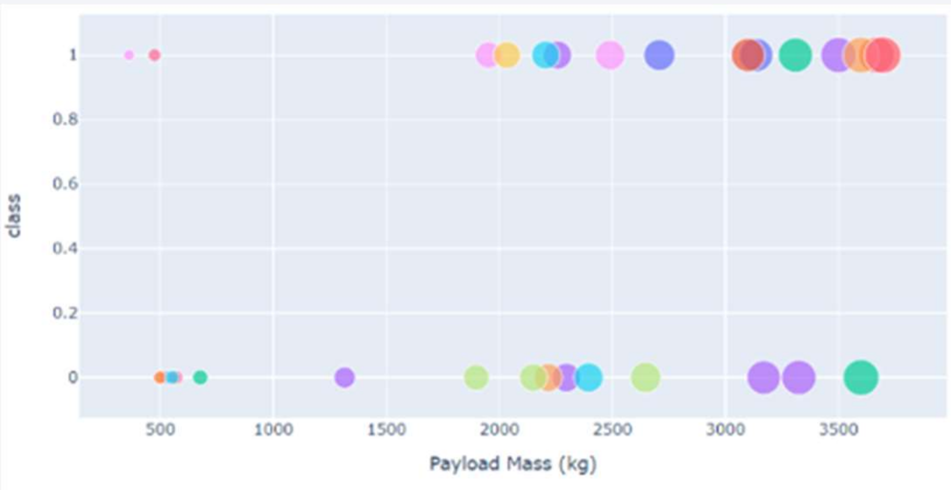
KSC LC-39A launch outcomes



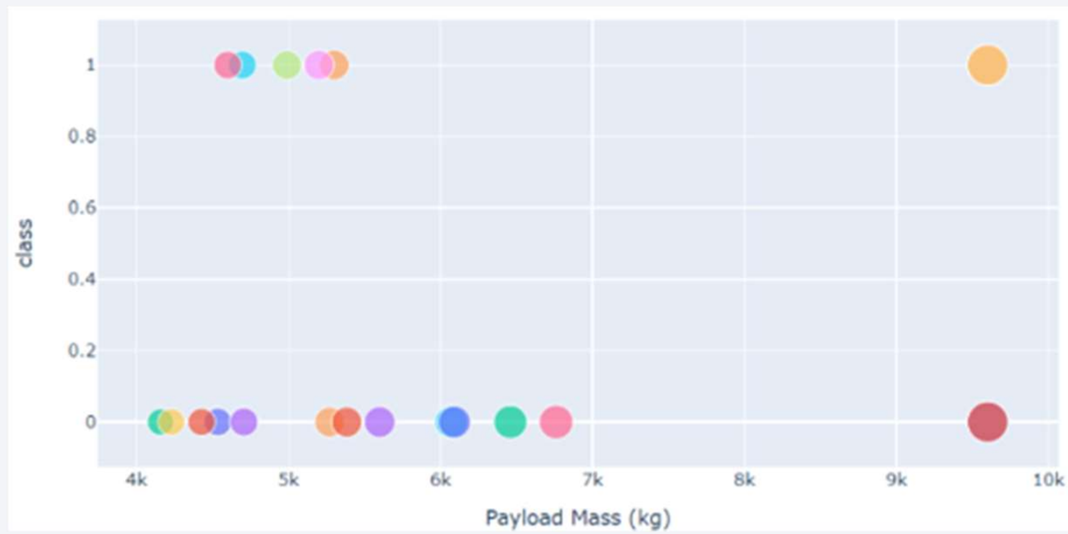
- KSC experienced a nearly 77% success ratio for all launches

Payload to Launch Outcomes for all sites

Payload at 4,000 kg



Payload at 10,000 kg

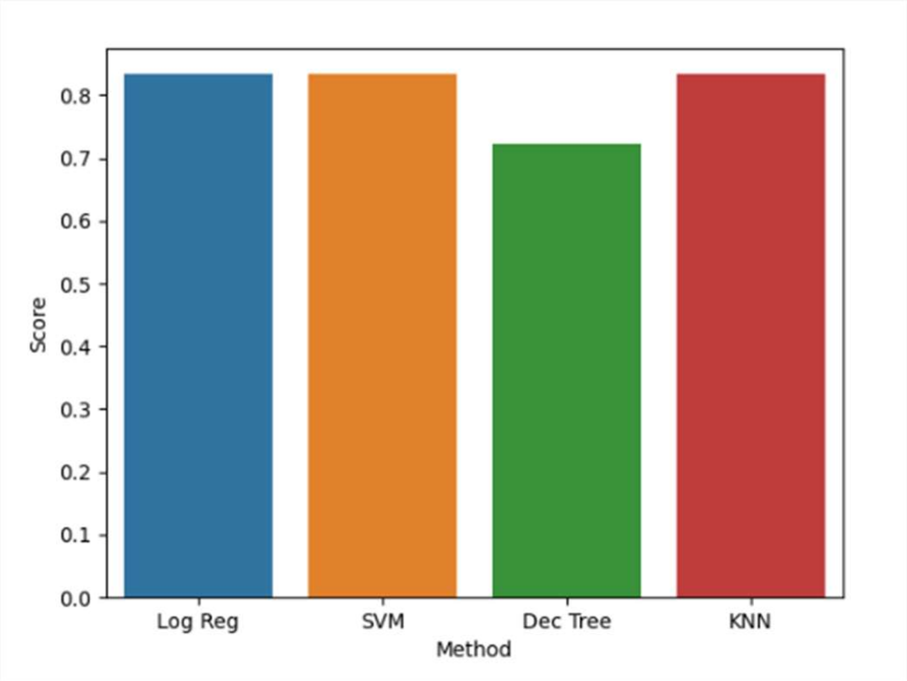


- Success rate is somewhat mixed for various payload weights with no clear differentiator

Section 5

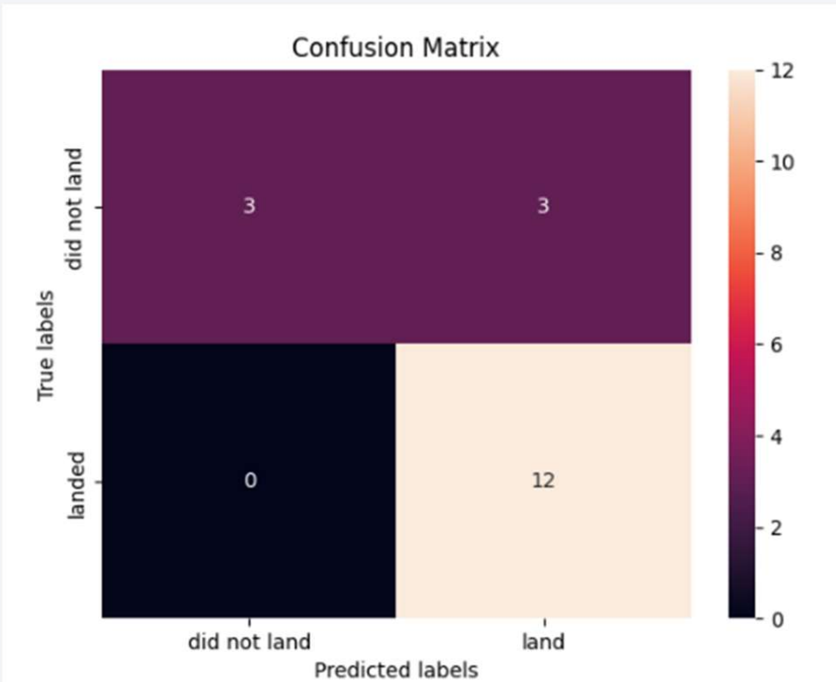
Predictive Analysis (Classification)

Classification Accuracy



All models except Decision Tree have the same accuracy score

Confusion Matrix



Confusion matrix is identical for all models, showing that it can distinguish between classes but yields a high degree of false positives

Conclusions

- The more flights that originate form a site, the greater the overall success rate of a reusable first stage
- Low level orbits have a higher degree of success
- Payload does not have a material impact on successful reusability of a first stage
- All machine learning algorithms except Decision Tree yield the same level of accuracy


Thank you!