IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alysia P.
July 16, 2025

Outline

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

Executive Summary

- Summary of Methodology
  - Data Collection using API
  - Data Collection with web scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction

Introduction

- In this capstone I have taken the role of a Data Scientist working for SpaceX to:
  - Determine if the first launch will be reused
  - Determine the cost of each launch
  - Gather information and create dashboards

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected via a SpaceX API and web scraping Wikipedia
- Perform data wrangling
  - Clean the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Determine and utilize the best Machine Learning model

Data Collection

- The data was collected via API and web scraping

Enter the URL for analysis

Request and parse SpaceX Launch data using the GET request

Decode the response from JSON to Pandas

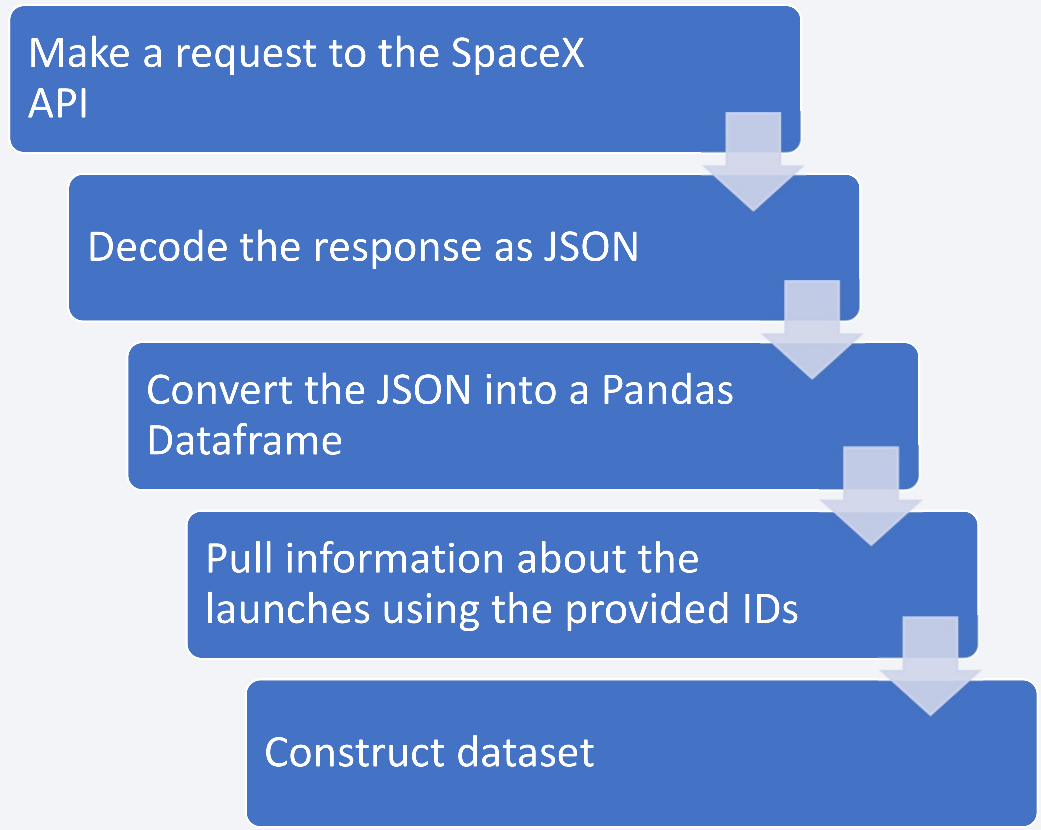
Use API to get information from the launches

Filter DF to include only launches by Falcon9 and replace null values

Data Collection – SpaceX API

- Data was collected using API followed the steps included. Below is the link for the completed notebook

- https://github.com/apetite-cat/IBM_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data was collected using web scraping followed the steps included. Below is the link for the completed notebook

- https://github.com/apetite-cat/IBM_Data_Science_Capstone/blob/main/jupyter-labs-webscraping.ipynb

Launch Wikipedia page for Falcon9 from its URL

From the HTML, create a beautiful soup object

Extract column names

Create an empty dictionary from the list of column names with keys

Fill in launch records and convert to CSV

Data Wrangling

- Data Wrangling was completed following the steps included. Below is the link for the completed notebook

- https://github.com/apetite-cat/IBM_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

COUNT the number of launches per site

COUNT the number of each orbit

COUNT the number and occurrence of the outcome of the mission by orbit type

Use Outcome column to create a label

EDA with Data Visualization

- Graphical representations of the data included:
  - Scatter plot
    - Flight Number vs Payload Mass
    - Flight Number vs Launch Sites
    - Payload vs Launch sites
    - Flight Number vs Orbit Type
    - Payload vs Orbit Type
  - Bar Chart
    - Depicting the success rate of each orbit
  - Line Graph
    - Depicting the success rate by date

- GitHub URL: https://github.com/apetite-cat/IBM_Data_Science_Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- SQL queries performed
  - Display names of the unique launch sites
  - Display records where launch site begins with 'CCA
  - Calculate the total payload mass carried by boosters launched
  - Display the payload mass carried by booster version F9 v1.1
  - List the first successful landing outcome date
  - List names of boosters with success in drone ship and have a payload mass between 4,000 and 6,000
  - Display the total number of successful and failed outcomes
  - Using a subquery, show the names of the boosters that have carried the max payload
  - List failed landings in drone ships, booster versions, and launch site names in 2015
  - Rank the count of outcomes

- GitHub URL:https://github.com/apetite-cat/IBM_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- The interactive map included Folium Markers to show the SpaceX Launch sites, in addition to their nearest notable landmarks (highways, cities, etc)
- Polylines were used to connect the launch sites to their nearest land marks
  - Green indicates success
  - Red indicates failure
- GitHub URL: https://github.com/apetite-cat/IBM_Data_Science_Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Plotly was utilized to create scatter charts and pie graphs depicting the launch records of Spacex
- These visualizations aide in the understanding of influential factors pertaining to the success rate of each site
  - Factors include payload mass and booster versions

Predictive Analysis (Classification)

- Machine Learning was used to predicitive analytics. Scikit-Learn was the tool applied.

1. Create columns
2. Standardize the data
3. Split the data in training and test data
4. Create a GridSearchCV object
5. Fit different ML objects
6. Calculate the accuracy of the test data for ML methods
7. Choose the ML method with the best accuracy
8. Compare the predictions with the real labels

- GitHub URL: https://github.com/apetite-cat/IBM_Data_Science_Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

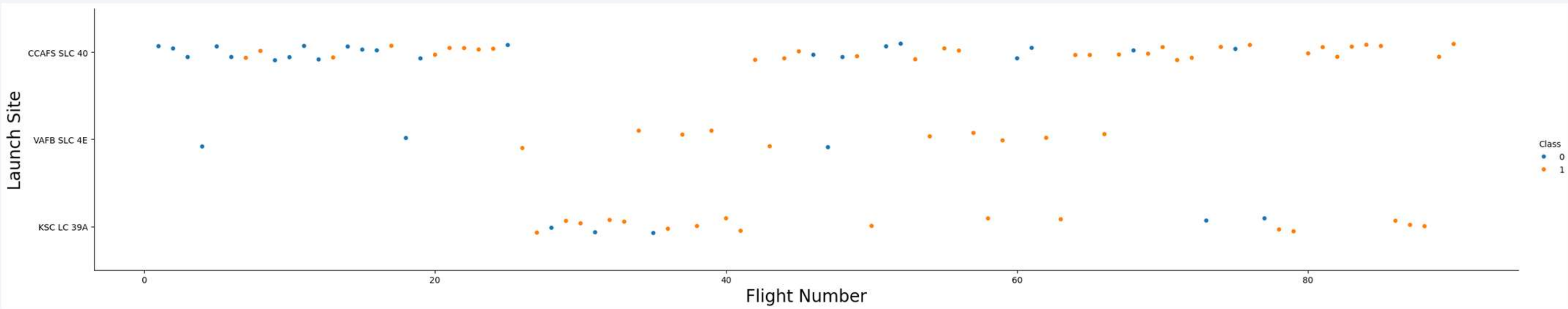
Results

- Exploratory data analysis results indicate that there may be a correlation between the landing outcome and the flight number. There is also an increase in successful landing outcomes since 2015
- Observations include the proximity of the launch sites to highways, railways, and large bodies of water. This could be for convenience of transportation of equipment and landing rockets in the water
- Predictive analysis results indicate the success of landing the rockets may be predicted with 83.33% accuracy

Section 2

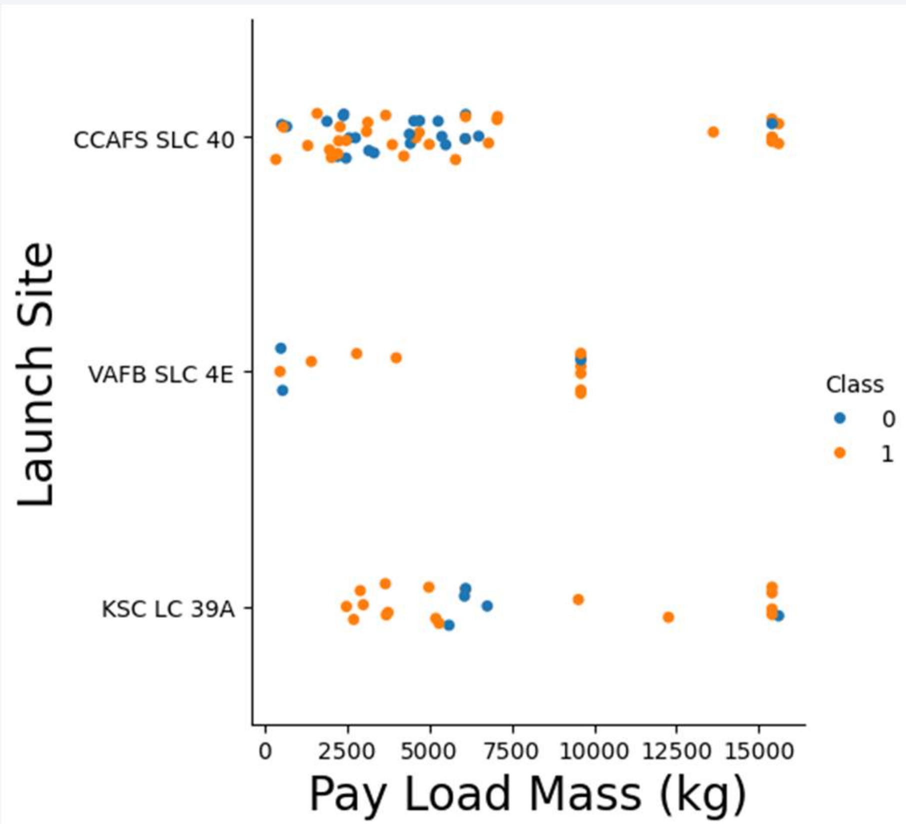
Insights drawn from EDA

Flight Number vs. Launch Site



- There were more successful landings as the flight number increased
- The most launches came from launch site CCAFS SLC 40

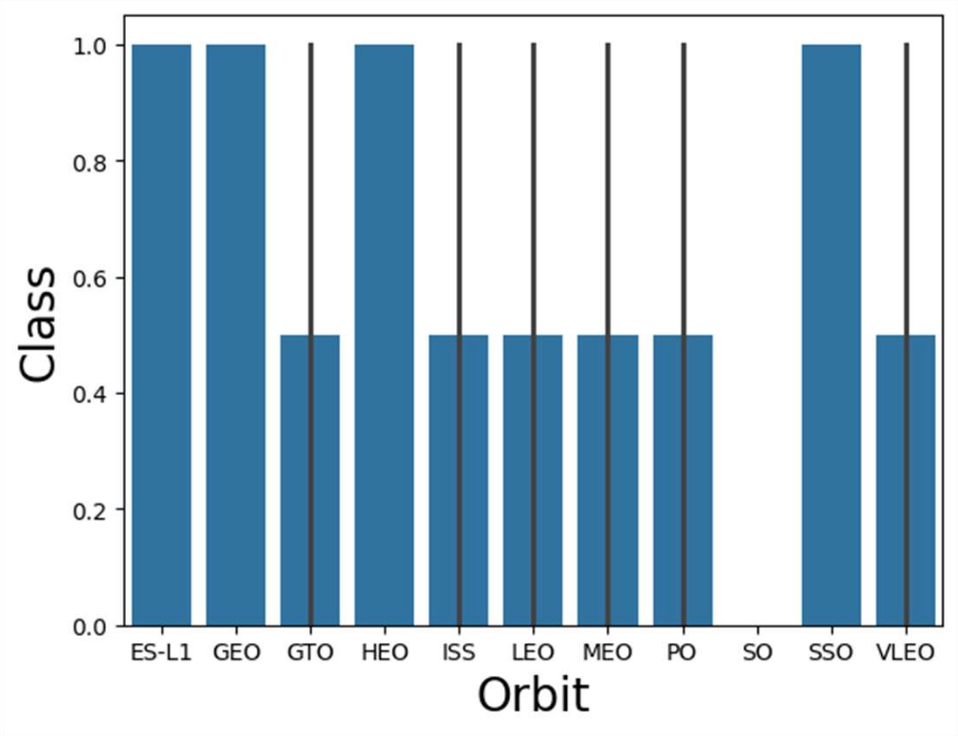
Payload vs. Launch Site



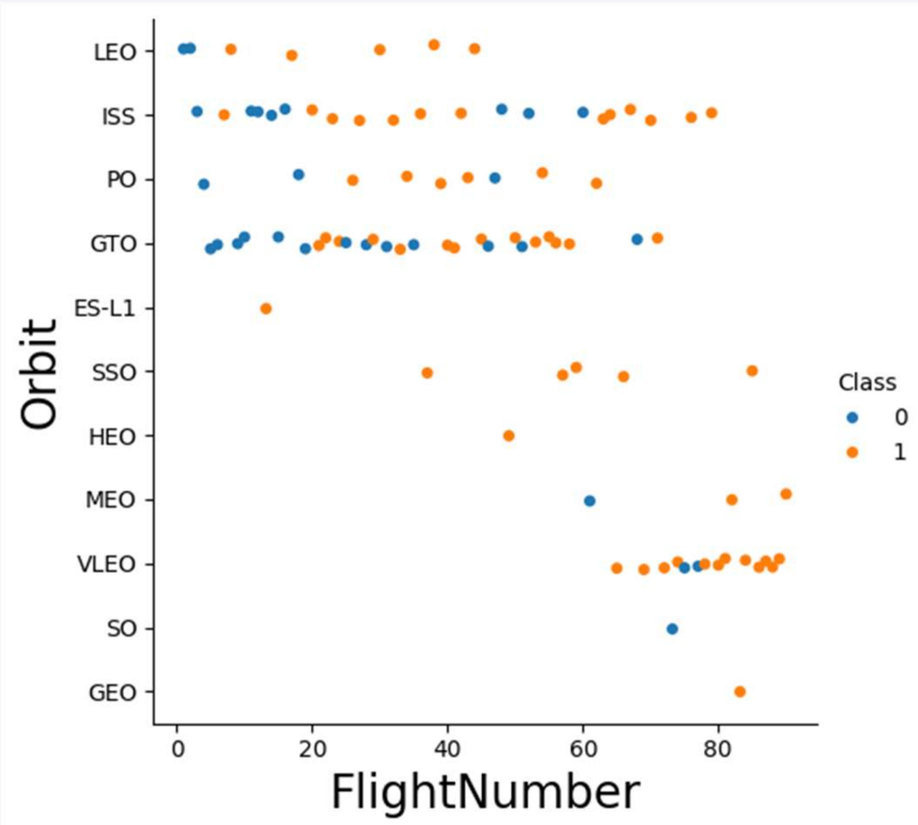
- It is notable that Launch Site VAFB SLC 4E did not have any launches with a payload mass above 10,000kg

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO all had the highest success rates



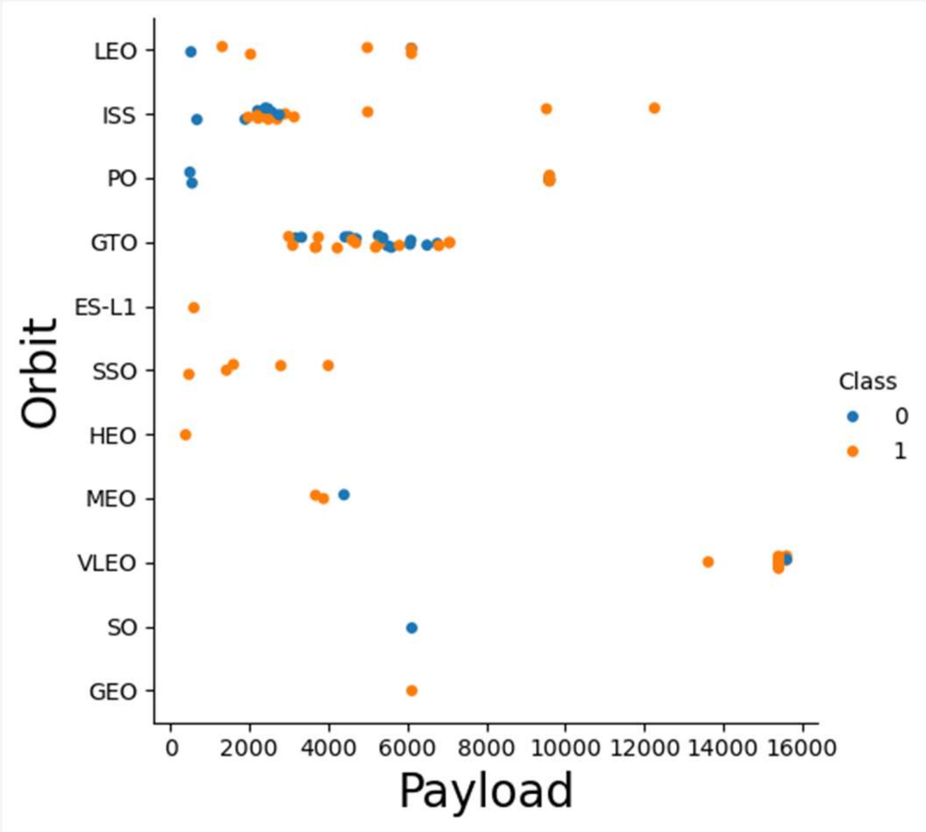
Flight Number vs. Orbit Type



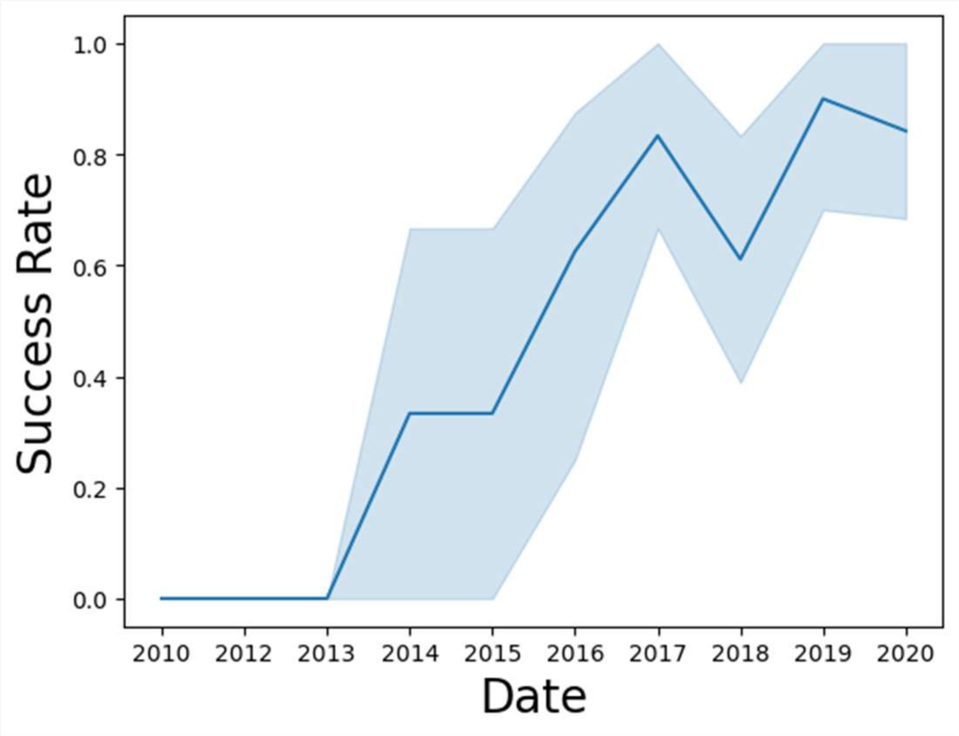
- The LEO orbit success seems to be related to the number of flights
- The GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

- Polar, LEO and ISS appear to have a more positive landing rate for landing with heavy payloads
- GTO is difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend



- Since 2013, there is a significant increase in success rate

All Launch Site Names

- Within the data, the following four sites were included with rocket landing attempts:
    - CCAFS LC-40
    - CCAFS SLC-40
    - KSC LC-39A
    - VAFB SLC-4E

```
%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

Launch Site Names Begin with 'CCA'

- These are 5 records where launch sites begin with CCA
- This data includes information from organizations outside of SpaceX ([Customer])

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

- This is the calculated total payload carried by boosters from NASA

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

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- This is the calculated average payload mass carried by booster version F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was 2015-12-22

```
%sql SELECT MIN(DATE) FROM SPACEXTBL 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

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

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

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_  >4000 AND
```

```
 * 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

- The total number of successful and failure mission outcomes is shown below

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(MISSION_OUTCOME) FROM SPACEXTBL WHERE MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failur
```

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

COUNT(MISSION_OUTCOME)
99

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass
  - 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

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

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

```
 * 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

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

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

```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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.

```
%sql select * from SPACEXTBL where Landing_Outcome = 'Success (ground pad)' and (DATE between '2010-06-04' and '2017-03-20'
```

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

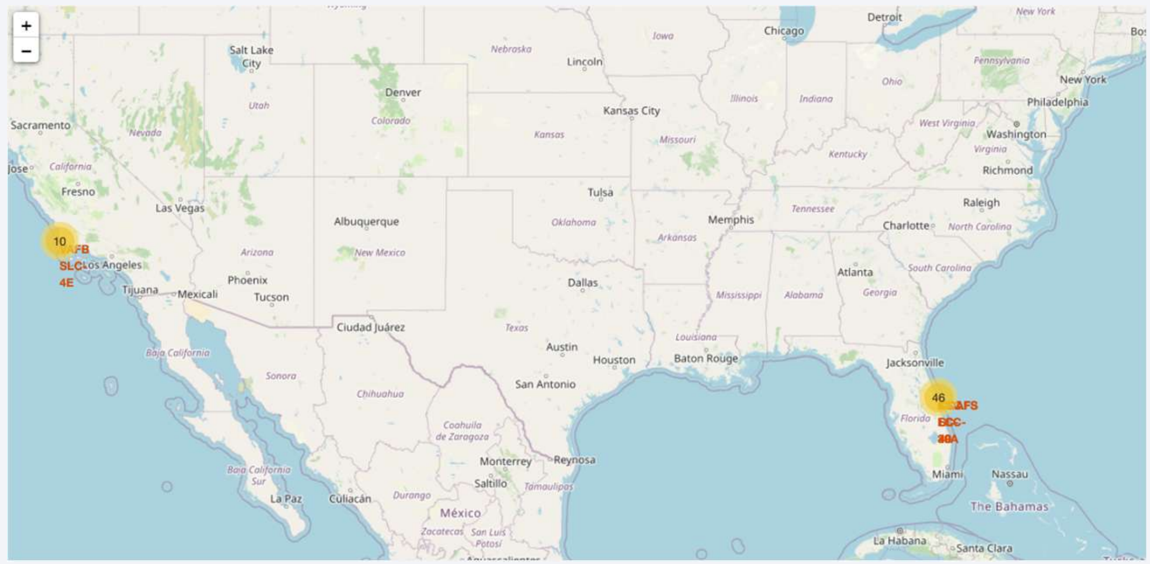
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

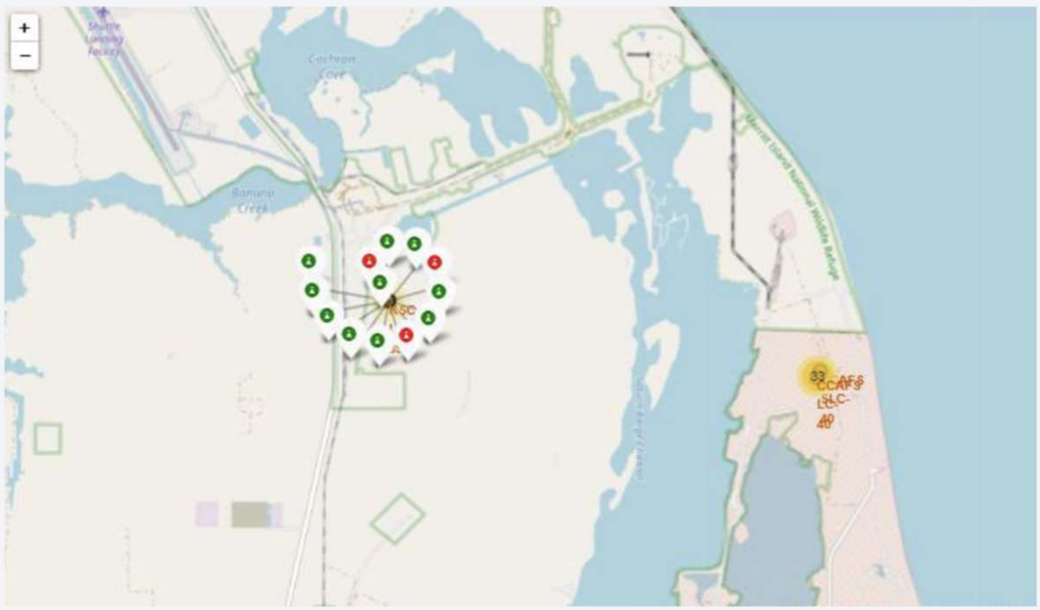
Lauch Site Locations

- Launch Sites are ideally placed near coastlines
- This makes the testing of water landing possible



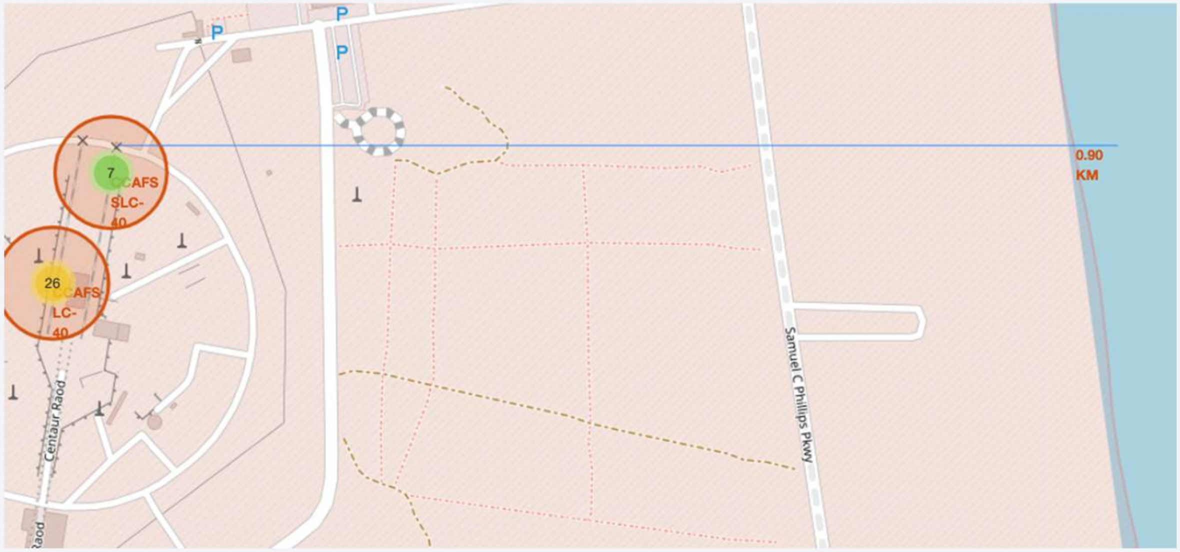
Success Rate of Rocket Launches

- Successful launches are indicated with a green marker while failed launches are red



Coastline Proximity

- This launch site is .9km from the coastline

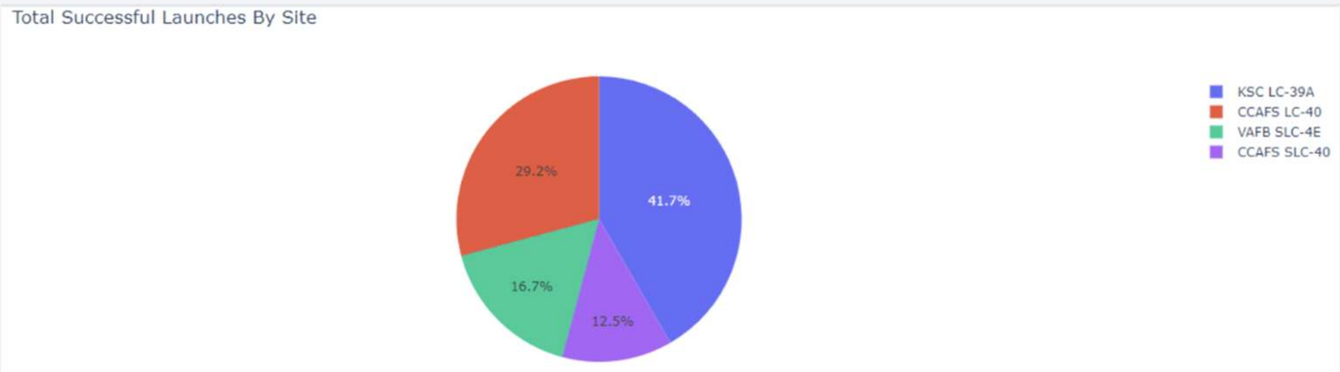


Section 4

Build a Dashboard with Plotly Dash

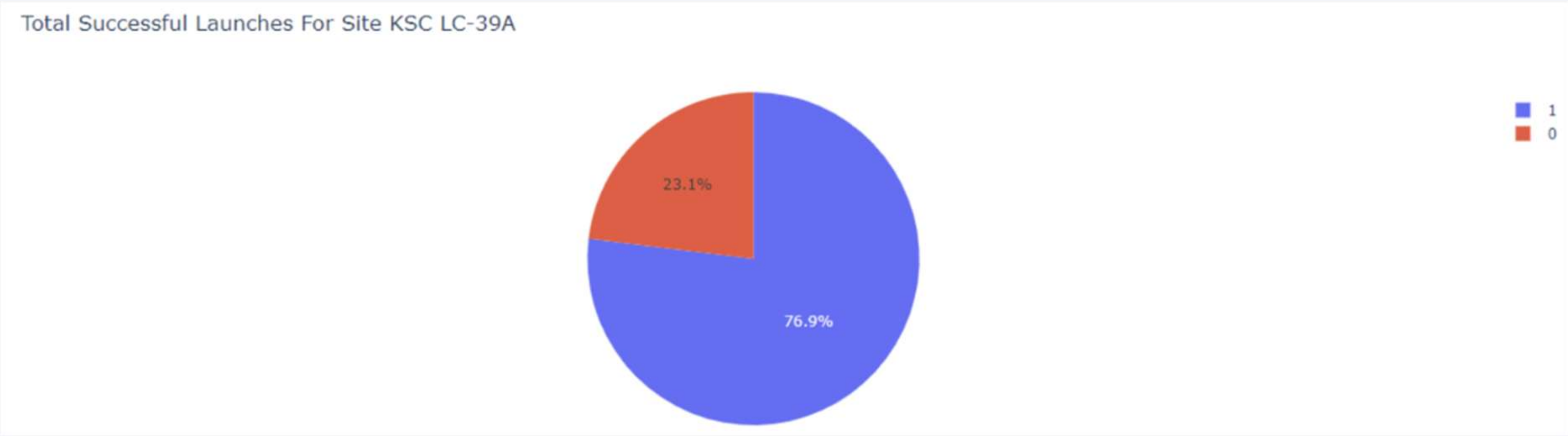
Successful Launches by Site

- KSC LC-39A had the most success, while CCAFS SLC-40 had the least



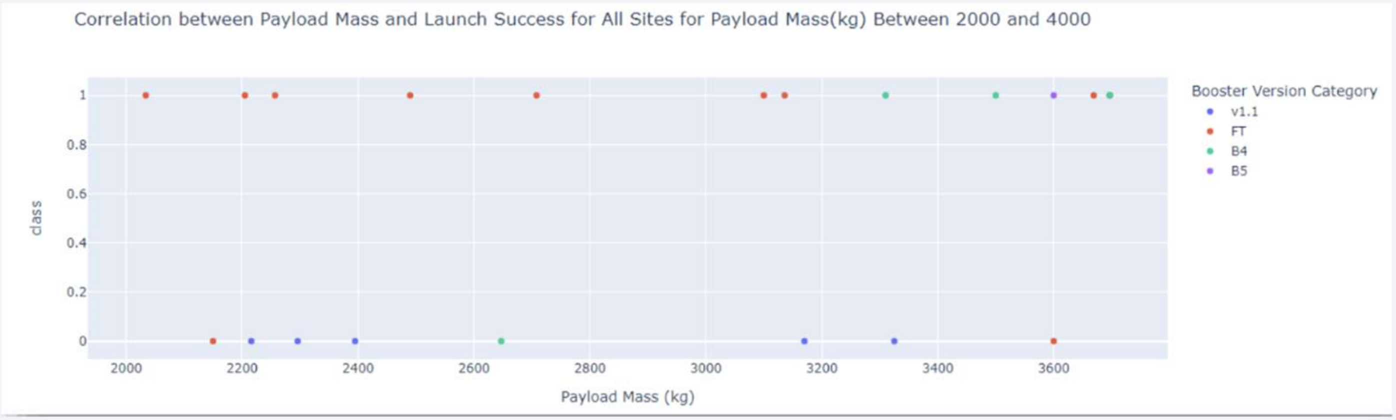
KSC LC-39A Successful Launches

- This site had a success rate of 76.9%



Payload vs. Launch Outcome

- V1.1 appears to have no successful outcomes
- FT and B5 had significantly more success
- B4 is more successful with a higher payload mass



Section 5

Predictive Analysis (Classification)

Classification Accuracy

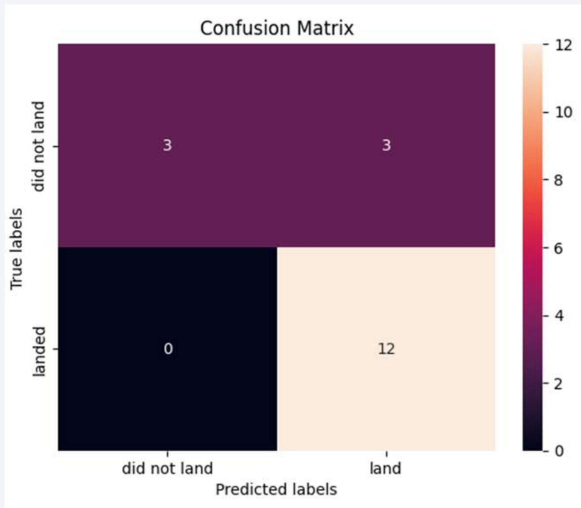
- It appears that all methods have a success rate of 83.33%

```
accuracy = [svm_cv_score, logreg_score, knn_cv_score, tree_cv_score]
accuracy = [i*100 for i in accuracy]

method = ['Support Vector Machine', 'Logistic Regression', 'K Nearest Neighbor', 'Decision Tree']
models = {'ML Method' :method, 'Accuracy Score (%)':accuracy}

ML_df = pd.DataFrame(models)
ML_df
```

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbor	83.333333
3	Decision Tree	83.333333

Confusion Matrix

- This model only failed to accurately predict 3 labels



Conclusions

- The location of the site is directly linked to its success
    - Maintaining close proximity to large bodies of water, highways, railways
    - Distance from large cities
- KSC LC-39A had the highest success rate
- Rocket Landings success has increased dramatically since 2013


Thank you!