IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ignatius Evans Erlangga
20th of June, 2024

Outline

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

Executive Summary

- I collected data from the public SpaceX API and SpaceX Wikipedia page. I created a 'class' column to classify successful landings. Using SQL, visualization techniques, folium maps, and dashboards, I explored the data. I selected relevant columns as features for analysis. Categorical variables were converted to binary using one-hot encoding. After standardizing the data, I employed GridSearchCV to optimize parameters for machine learning models. Finally, I visualized the accuracy scores of all models.

- Four machine learning models were generated: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. Each achieved an accuracy rate of approximately 83.33%. Interestingly, all models tended to over-predict successful landings. To improve model accuracy and reliability, additional data is necessary for better determination.

Introduction

Background :

The commercial space age is now a reality. SpaceX leads with competitive pricing at $62 million compared to $165 million USD for its rivals. This cost advantage is largely attributed to SpaceX's capability to recover and reuse the first stage of its rockets. Space Y is aiming to rival SpaceX in this competitive market.

(SpaceX)

Task :

Space Y has assigned us the responsibility of developing a machine learning model aimed at forecasting the successful recovery of Stage 1 rockets. This initiative aligns with their goal to enhance operational efficiency and reliability in their space missions, aiming to compete effectively in the commercial space industry, which is increasingly driven by the ability to recover and reuse rocket components.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data collection methodology involved combining data from the SpaceX public API and the SpaceX Wikipedia page.
- Perform data wrangling
  - Data Wrangling was perform to clean and prepare the data, afterwards Successful and Unsuccessful Landings were classified based on specific criteria.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - EDA was conducted using SQL and various visualization techniques to gain insights.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Interactive visual analytics were implemented.
- Perform predictive analysis using classification models
  - Predictive analysis was carried out using classification models, and model parameters were fine-tuned using GridSearchCV.

Data Collection

The data collection process included utilizing API requests from SpaceX's public API and scraping data from a table found in SpaceX's Wikipedia entry. The subsequent slide will illustrate the flowchart depicting the data collection process from the API, followed by another slide detailing the flowchart for data collection through web scraping.

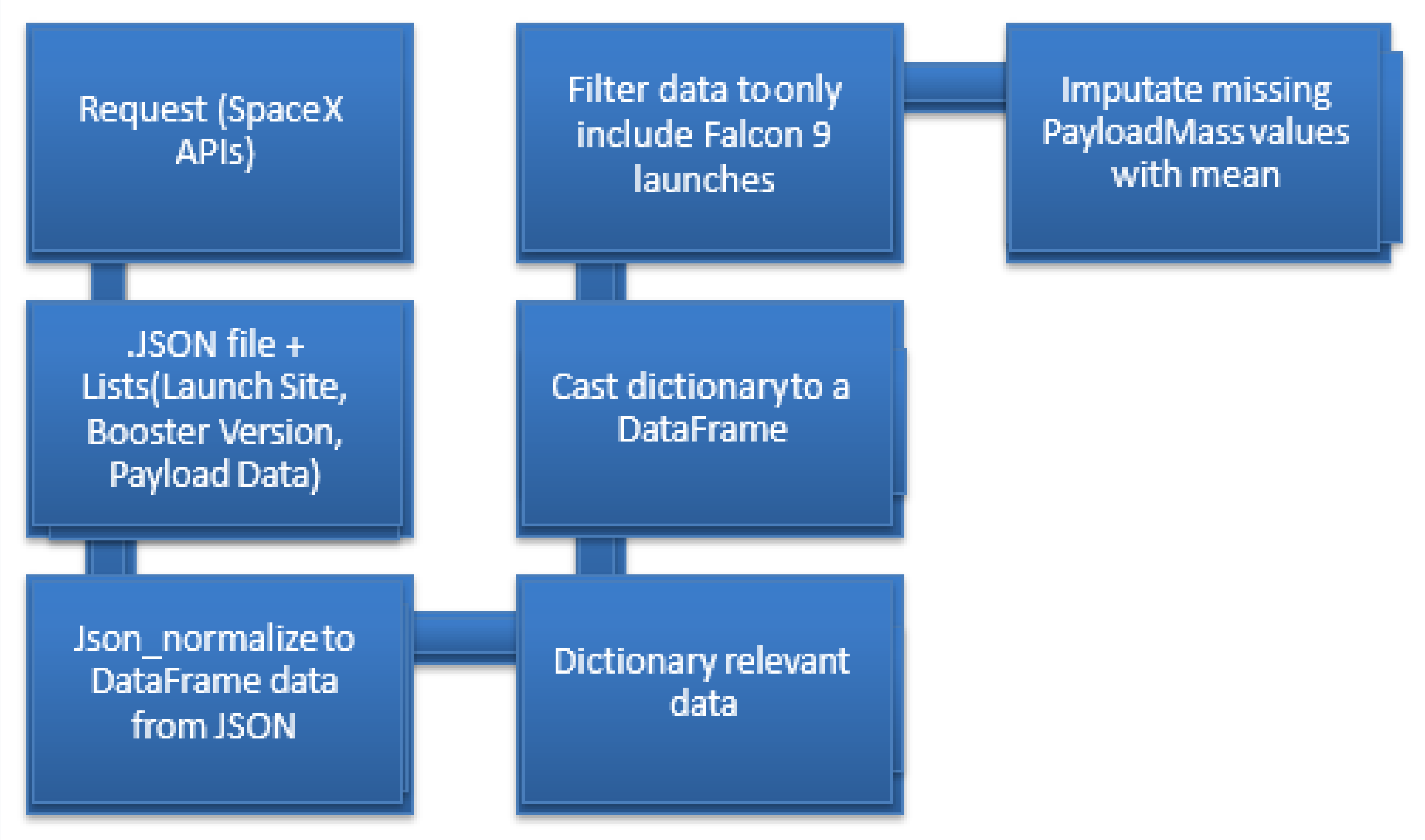
- Space X API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- Wikipedia Webscrape Data Columns:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

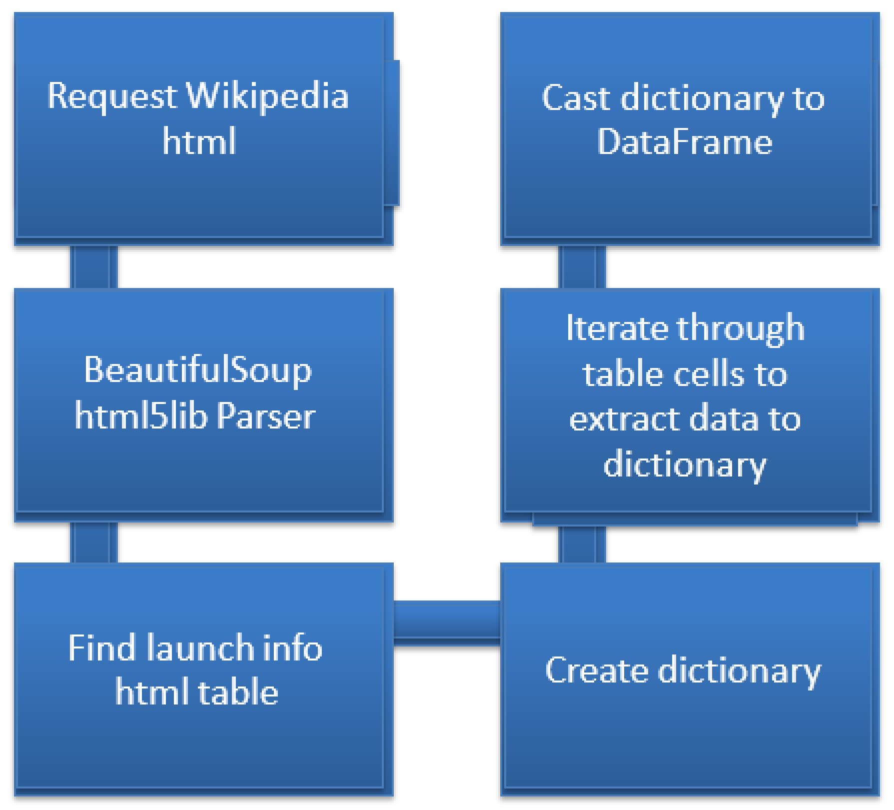
Data Collection – SpaceX API



GitHub URL :

https://github.com/evanserlangga/IBMDSPFP/blob/ae5350a75d04afc0344e178494715908fb1de441/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



GitHub URL :

https://github.com/evanserlangga/IBMDSPFP/blob/ae5350a75d04afc0344e178494715908fb1de441/jupyter-labs-webscraping.ipynb

Data Wrangling

To create a training label with landing outcomes where successful landings are represented as 1 and failures as 0, follow these steps:

1. The 'Outcome' column consists of two components: 'Mission Outcome' and 'Landing Location'.
2. Create a new column called 'class' for the training label. This column will have a value of 1 if the 'Mission Outcome' is True, and 0 otherwise. The value mapping is as follows:
   - True ASDS, True RTLS, and True Ocean are mapped to 1.
   - None None, False ASDS, None ASDS, False Ocean, and False RTLS are mapped to 0.

This approach ensures that successful landings are correctly labeled for training the machine learning model.

GitHub URL :

https://github.com/evanserlangga/IBMDSPFP/blob/ae5350a75d04afc0344e178494715908fb1de441/labs_jupyter_spacex_Data_wrangling.ipynb

EDA with Data Visualization

Exploratory Data Analysis (EDA) was conducted on several variables, including Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. The analysis involved creating various plots to examine the relationships between these variables:

- **Flight Number vs. Payload Mass**
- **Flight Number vs. Launch Site**
- **Payload Mass vs. Launch Site**
- **Orbit vs. Success Rate**
- **Flight Number vs. Orbit**
- **Payload Mass vs. Orbit**
- **Success Yearly Trend**

Scatter plots, line charts, and bar plots were utilized to compare these relationships and determine if significant correlations existed. These insights were crucial for selecting features to train the machine learning model.

GitHub URL :

https://github.com/evanserlangga/IBMDSPFP/blob/ae5350a75d04afc0344e178494715908fb1de441/edadataviz.ipynb

EDA with SQL

GitHub URL : https://github.com/evanserlangga/IBMDSPFP/blob/ae5350a75d04afc0344e178494715908fb1de441/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- SQL queries performed :
- 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. Use a subquery
- 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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

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 we could discover some of the factors by analyzing the existing launch site locations.

Folium maps were used to mark launch sites, indicating both successful and unsuccessful landings, and to demonstrate proximity to key locations such as railways, highways, the coast, and cities. This visualization helps to understand the strategic placement of launch sites and provides insights into the success rates of landings relative to their locations.

GitHub URL :

https://github.com/evanserlangga/IBMDSPFP/blob/ae5350a75d04afc0344e178494715908fb1de441/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

The dashboard includes both a pie chart and a scatter plot for comprehensive data visualization.

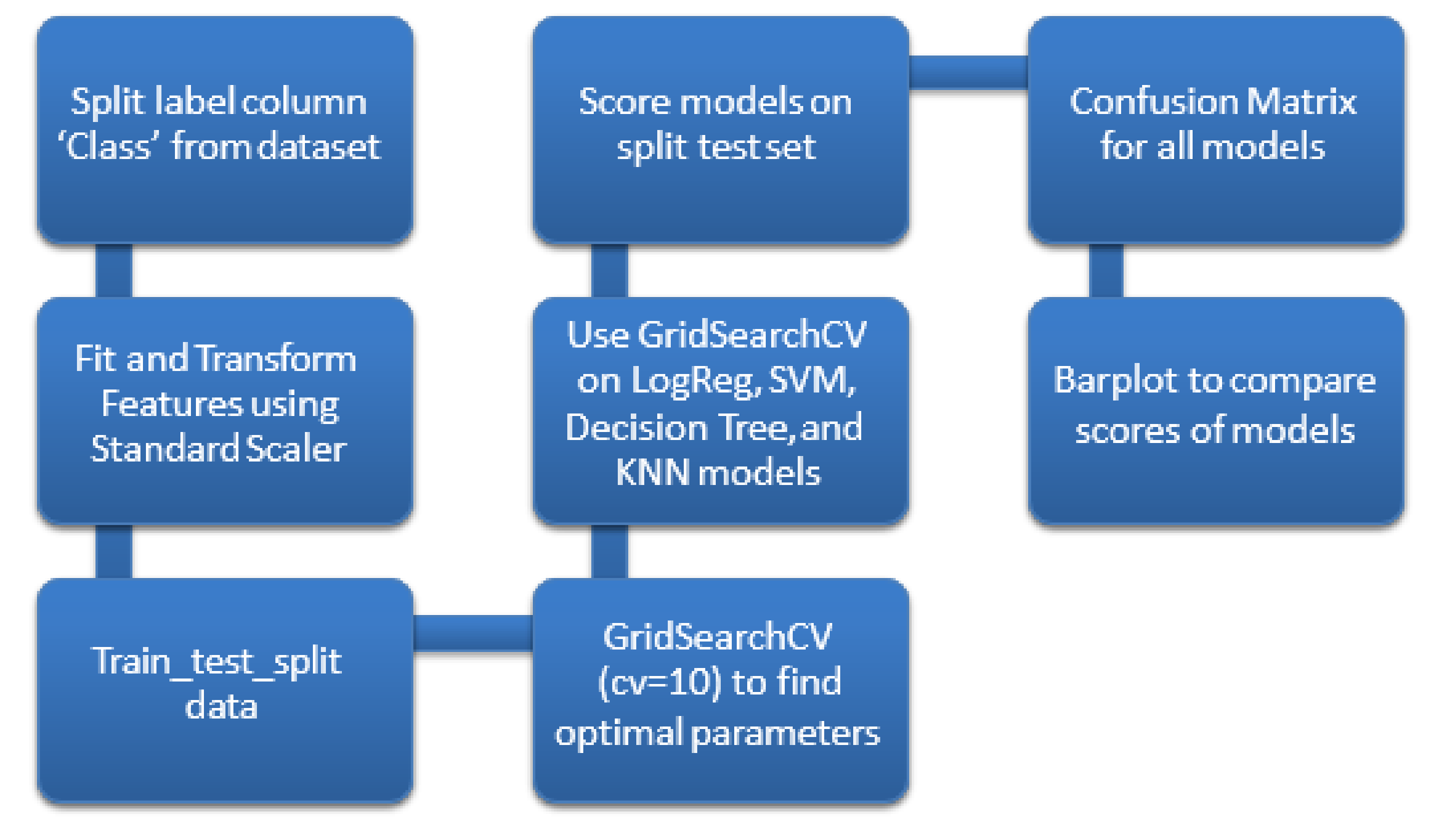
- **Pie Chart**: This chart can display the distribution of successful landings across all launch sites or the success rates of individual launch sites. It effectively visualizes the launch site success rates.
- **Scatter Plot**: This plot allows for interactive inputs, where users can select either all sites or a specific site, and adjust the payload mass using a slider ranging from 0 to 10,000 kg. The scatter plot helps to observe how success rates vary across different launch sites, payload masses, and booster version categories.

These visual tools aid in analyzing the factors influencing successful landings.

GitHub URL :

https://github.com/evanserlangga/IBMDSPFP/blob/289c3bb6d51329c83454dfe144674c1d9fba7a1b/spacex_dash_app.py

Predictive Analysis (Classification)



GitHub URL :

https://github.com/evanserlangga/IBMDSPFP/blob/289c3bb6d51329c83454dfe144674c1d9fba7a1b/SpaceX_Machine_Learning_Prediction_Part_5_jupyterlite.ipynb

Results

In the upcoming slides, I will provide a comprehensive overview of the work completed during this project. First, I will delve into the results of our exploratory data analysis (EDA), which involved a thorough examination of key variables such as Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. This analysis helped to identify patterns and relationships within the data that are crucial for developing accurate predictive models.

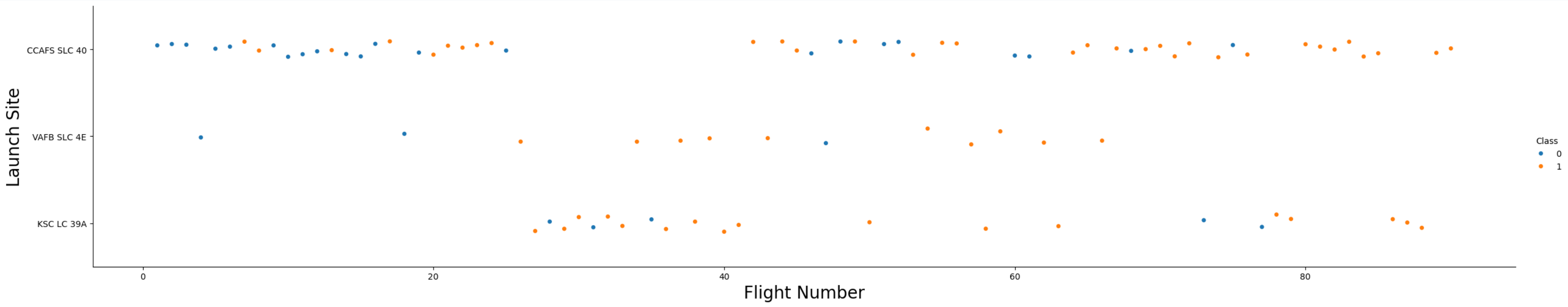
Next, I will present an interactive analytics demo through a series of detailed screenshots. This demo will illustrate how advanced visualization tools, including Folium maps and Plotly Dash, were utilized to create dynamic and insightful visual representations of the data. These visualizations enabled a deeper understanding of the spatial distribution of launch sites, the success rates of landings, and other significant factors influencing mission outcomes.

Finally, I will share the results of our predictive analysis. This section will cover the development and evaluation of several machine learning models designed to predict the success of Stage 1 rocket landings. I will discuss the model selection process, the tuning of hyperparameters using GridSearchCV, and the performance metrics that indicate the accuracy and reliability of the models. This comprehensive presentation aims to demonstrate the effectiveness of the analytical methods and tools employed in this project, ultimately contributing to Space Y's goal of achieving successful rocket recoveries.

Section 2

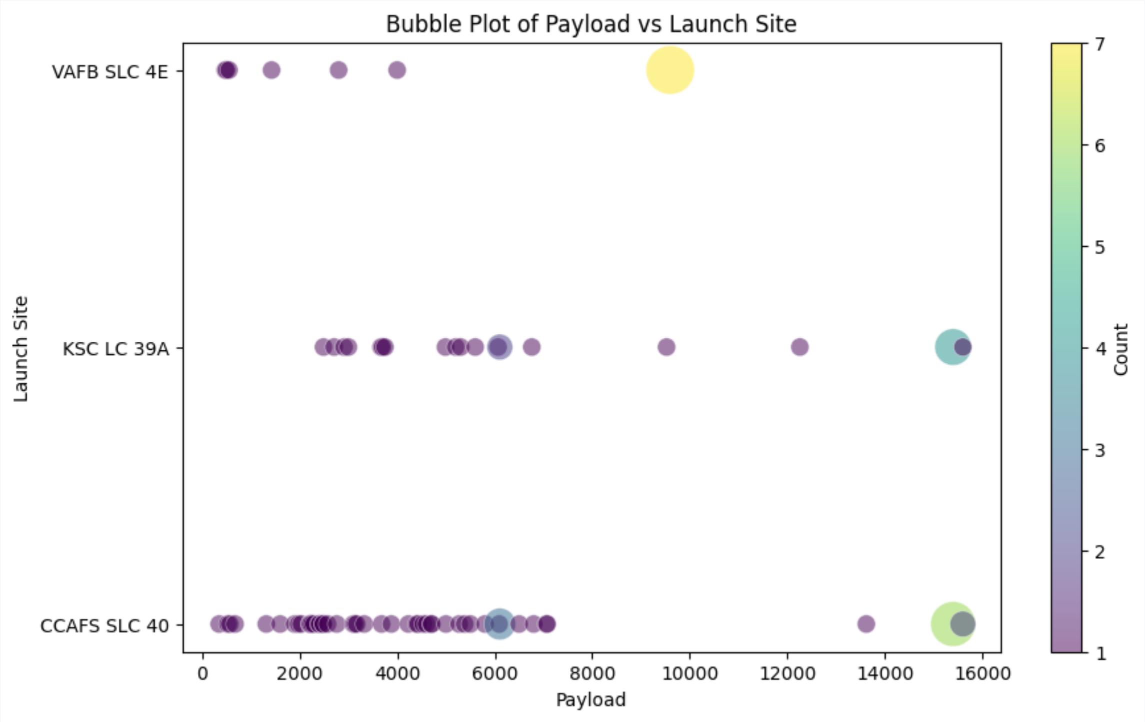
Insights drawn from EDA

Flight Number vs. Launch Site



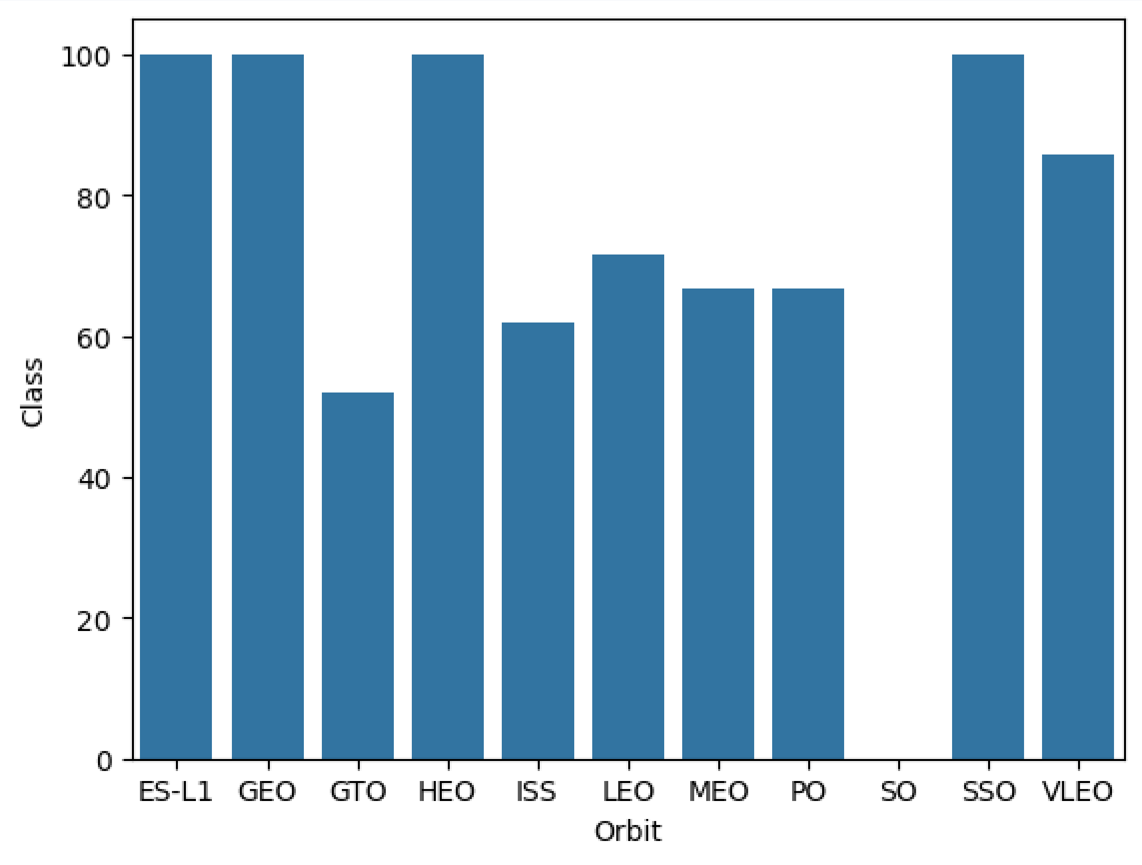
The graphic indicates a clear increase in the success rate of landings over time, as reflected in the Flight Number. Notably, there appears to be a significant breakthrough around the 20th flight, which substantially boosted the success rate. Additionally, Cape Canaveral Air Force Station (CCAFS) stands out as the primary launch site, handling the highest volume of launches.

Payload vs. Launch Site



The data shows that payload masses predominantly range between 0 and 6000 kg. Furthermore, it appears that different launch sites tend to handle varying payload masses, indicating a potential specialization or preference for certain payload ranges at specific locations.

Success Rate vs. Orbit Type

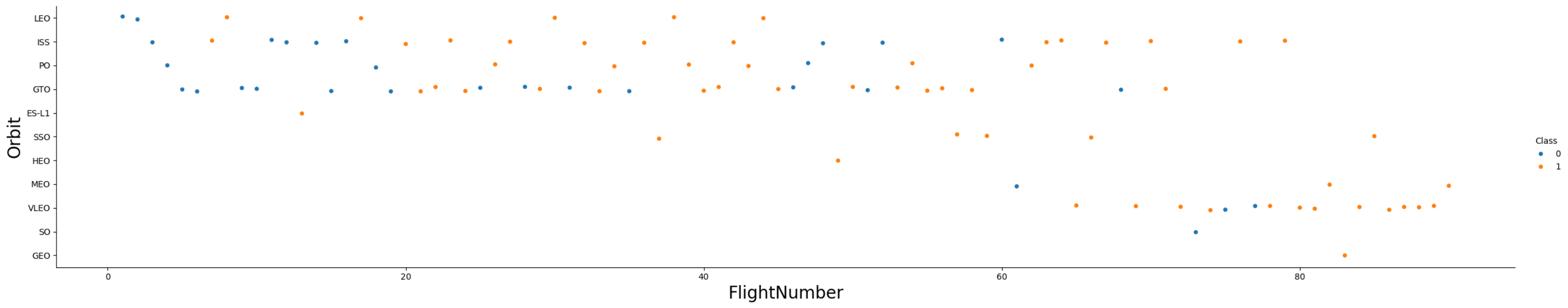


The data analysis reveals the following success rates for different orbits, with sample sizes indicated in parentheses:

- ES-L1 (1), GEO (1), and HEO (1) each have a 100% success rate.
- SSO (5) also maintains a 100% success rate.
- VLEO (14) shows a decent success rate with a significant number of attempts.
- SO (1) has a 0% success rate.
- GTO (27) has approximately a 50% success rate but has the largest sample size.

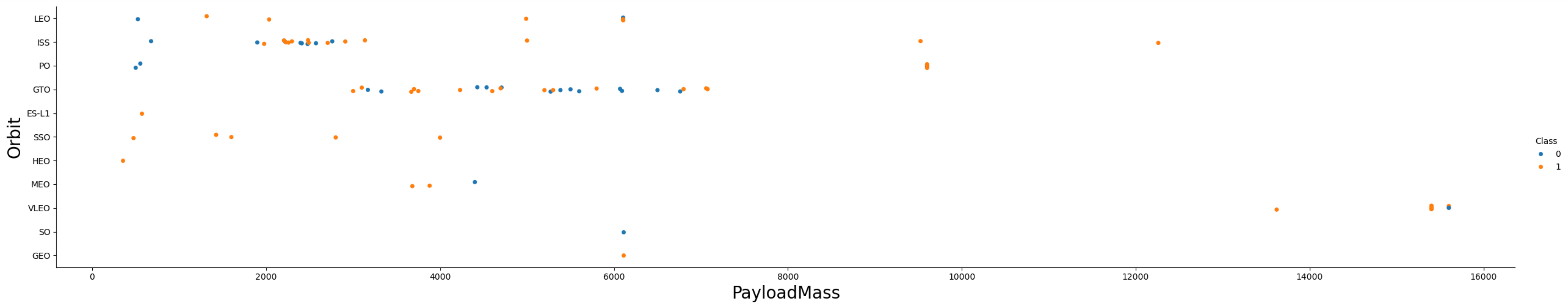
Success Rate Scale with 0 as 0% ;
0.6 as 60% ; 1 as 100%

Flight Number vs. Orbit Type



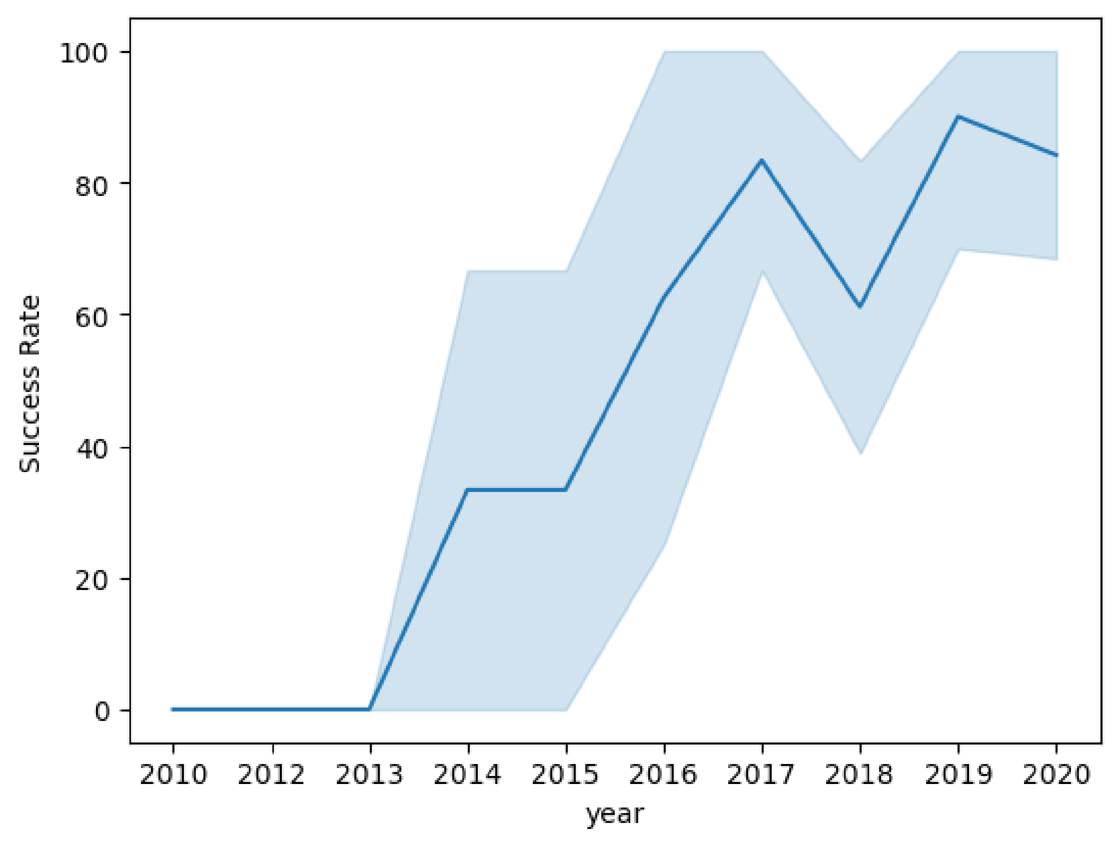
Over the course of Flight Number, there is observable variability in SpaceX's launch orbit preferences, which correlates closely with launch outcomes. Initially focusing on LEO orbits, which saw moderate success, SpaceX later shifted towards VLEO orbits in recent launches. The data suggests that SpaceX tends to perform better in lower orbits or Sun-synchronous orbits, highlighting a strategic preference that aligns with higher success rates in these specific orbital configurations.

Payload vs. Orbit Type



There appears to be a correlation between payload mass and orbit type based on the analysis. Orbits such as LEO and SSO tend to accommodate payloads with relatively lower mass. In contrast, VLEO, which shows high success rates, typically handles payloads at the higher end of the mass spectrum. This observation underscores how payload characteristics, including mass, are tailored to suit specific orbital requirements and operational success in SpaceX's missions.

Launch Success Yearly Trend



The analysis indicates a general trend of increasing success rates over time since 2013, with a notable slight dip observed in 2018. In recent years, success rates have stabilized around 80%, reflecting a consistent improvement in SpaceX's operational reliability and mission outcomes.

95% confidence interval (light blue shading)

All Launch Site Names

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

In [9]:
```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

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

Out[9]:

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

Based on the query and the information provided, it appears that there are three unique launch site names in the database:

1. CCAFS SLC-40 (Cape Canaveral Air Force Station Space Launch Complex 40)
2. KSC LC-39A (Kennedy Space Center Launch Complex 39A)
3. VAFB SLC-4E (Vandenberg Air Force Base Space Launch Complex 4E)

The variations like CCAFS SLC-40 and CCAFS LC-40 likely refer to the same launch site but with data entry errors or variations in naming conventions.

Launch Site Names Begin with 'CCA'

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

In [24]:
```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

Out[24]:

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

Showing the first five entries in database with Launch Site name beginning with CCA.

Total Payload Mass

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

In [11]:
```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

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

Out[11]:

sum
45596

This query calculates the total payload mass in kilograms for missions where NASA was the customer. The use of CRS (Commercial Resupply Services) indicates that these payloads were specifically intended for delivery to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

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

In [12]:
```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

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

Out[12]:

Average
2534.6666666666665

This query computes the average payload mass for launches utilizing the booster version F9 v1.1. The average payload mass associated with F9 1.1 tends to fall on the lower end of our payload mass range.

First Successful Ground Landing Date

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

Hint:Use min function

In [14]:
```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

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

Out[14]:

Date
2010-06-04

This query returns the first successful ground pad landing date.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

This query retrieves the four booster versions that achieved successful drone ship landings with a payload mass ranging between 4000 and 6000 kilograms (non-inclusive).

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [34]:
```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

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

Out[34]:

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

This query provides a count of each mission outcome. SpaceX achieves its mission outcome nearly 99% of the time, indicating that most landing failures are deliberate or intended. Interestingly, there is one launch with an unclear payload status, and unfortunately, one mission resulted in a mid-flight failure.

Boosters Carried Maximum Payload

```
In [45]:
# Execute the first SQL query to get the maximum payload_mass_kg
maxm = %sql SELECT MAX(payload_mass__kg_) FROM SPACEXTBL;

# Extract the maximum value from the result set
maxv = maxm[0][0] if maxm else None

# Execute the second SQL query to fetch booster_version where payload_mass_kg equals maxv
%sql SELECT booster_version FROM SPACEXTBL WHERE payload_mass__kg_ = :maxv;

 * sqlite:///my_data1.db
Done.
 * sqlite:///my_data1.db
Done.
Out[45]:
```

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

This query identifies the booster versions that carried the highest payload masses. These booster versions, all falling under the F9 B5 B10xx.x category, are remarkably similar. This suggests a correlation between the booster version used and the payload mass capacity it can accommodate.

2015 Launch Records

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

In [49]:

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site
from SPACEXDATASET where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

Out[49]:

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

This query retrieves the month, landing outcome, booster version, payload mass (in kilograms), and launch site for launches in 2015 where the Stage 1 failed to land on a drone ship. There were two such occurrences during this period.

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

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

In [61]:

```
%sql select landing__outcome, count(*) as count from SPACEXDATASET
where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing__outcome ORDER BY count Desc
```

```
* ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

Out[61]:

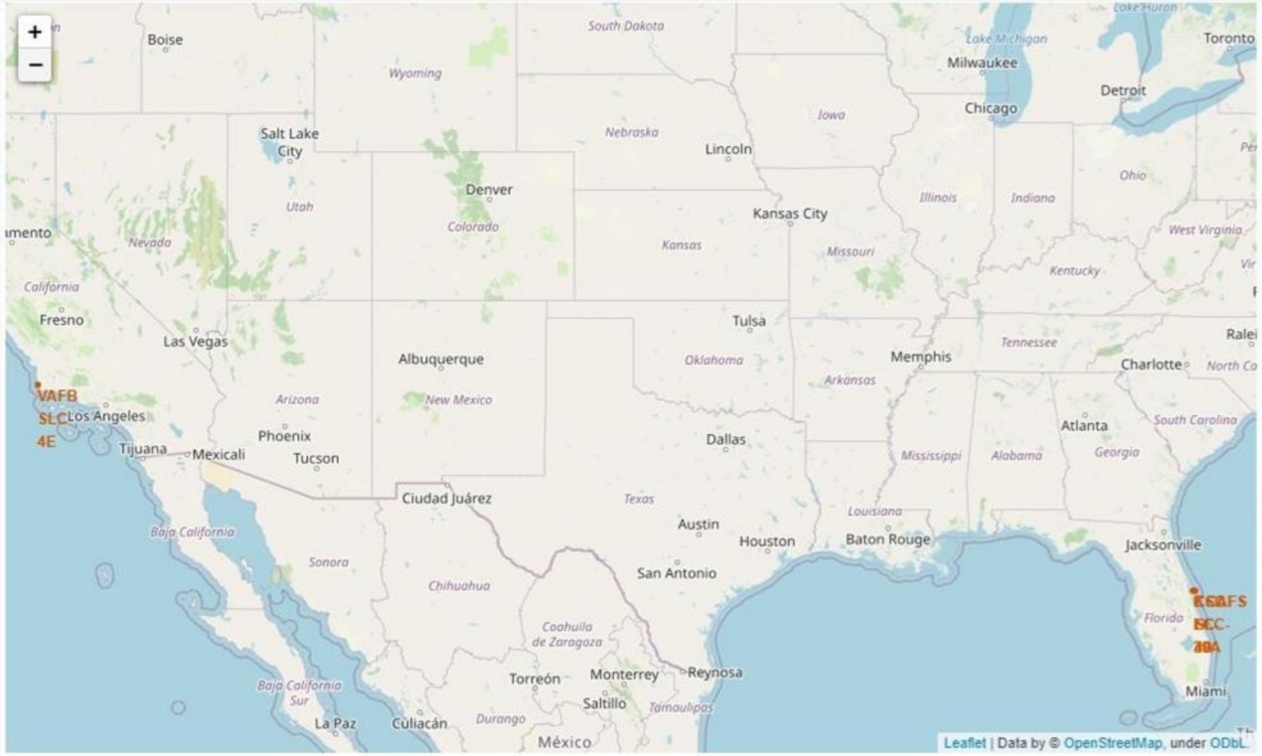
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

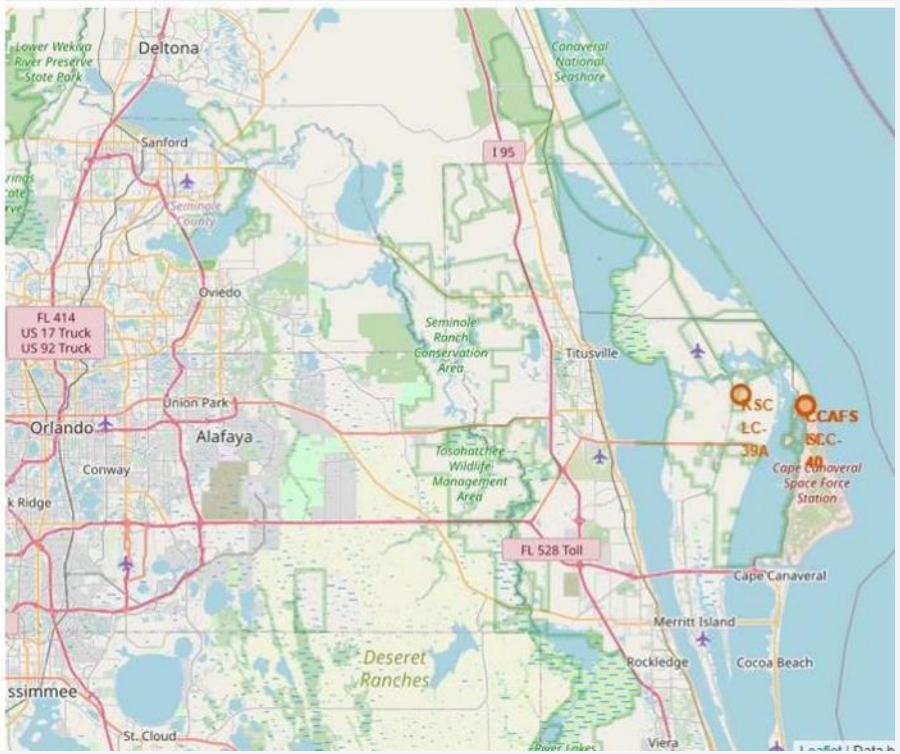
This query provides a list of successful landings between June 4, 2010, and March 20, 2017, inclusive. Successful landings during this period are categorized into two types: drone ship and ground pad landings. In total, there were 8 successful landings documented.

Section 3

Launch Sites Proximities Analysis

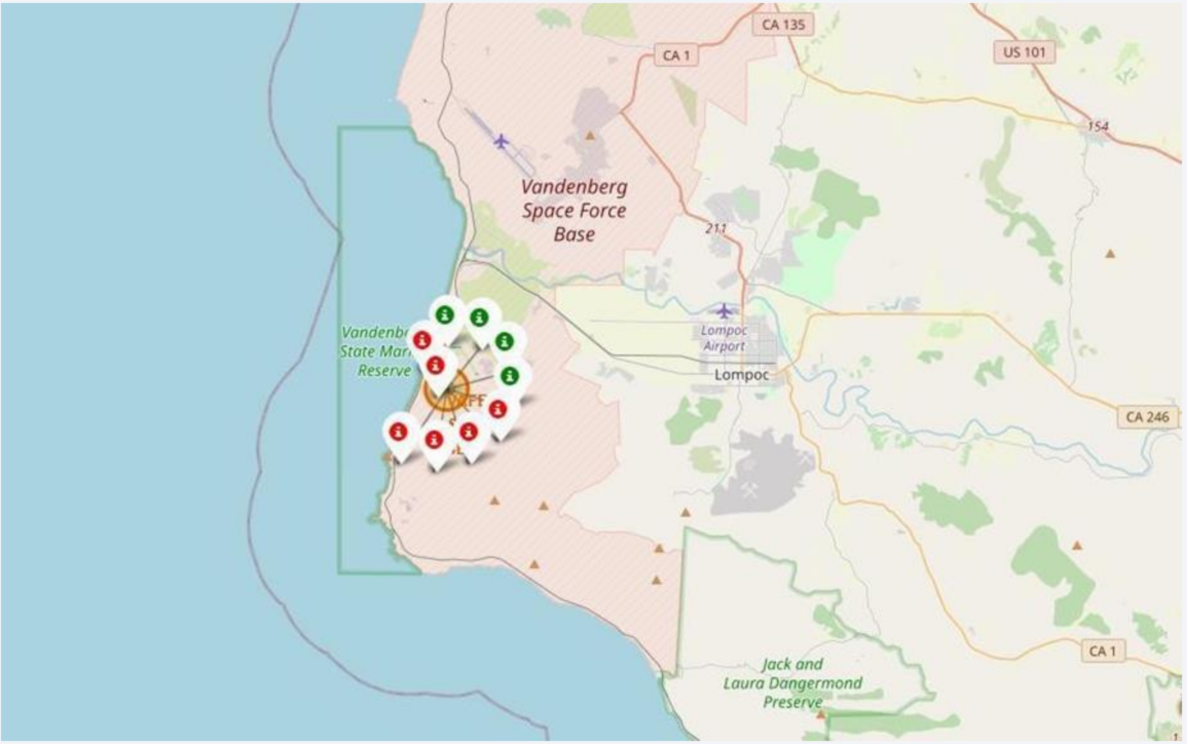
Launch Sites Geography





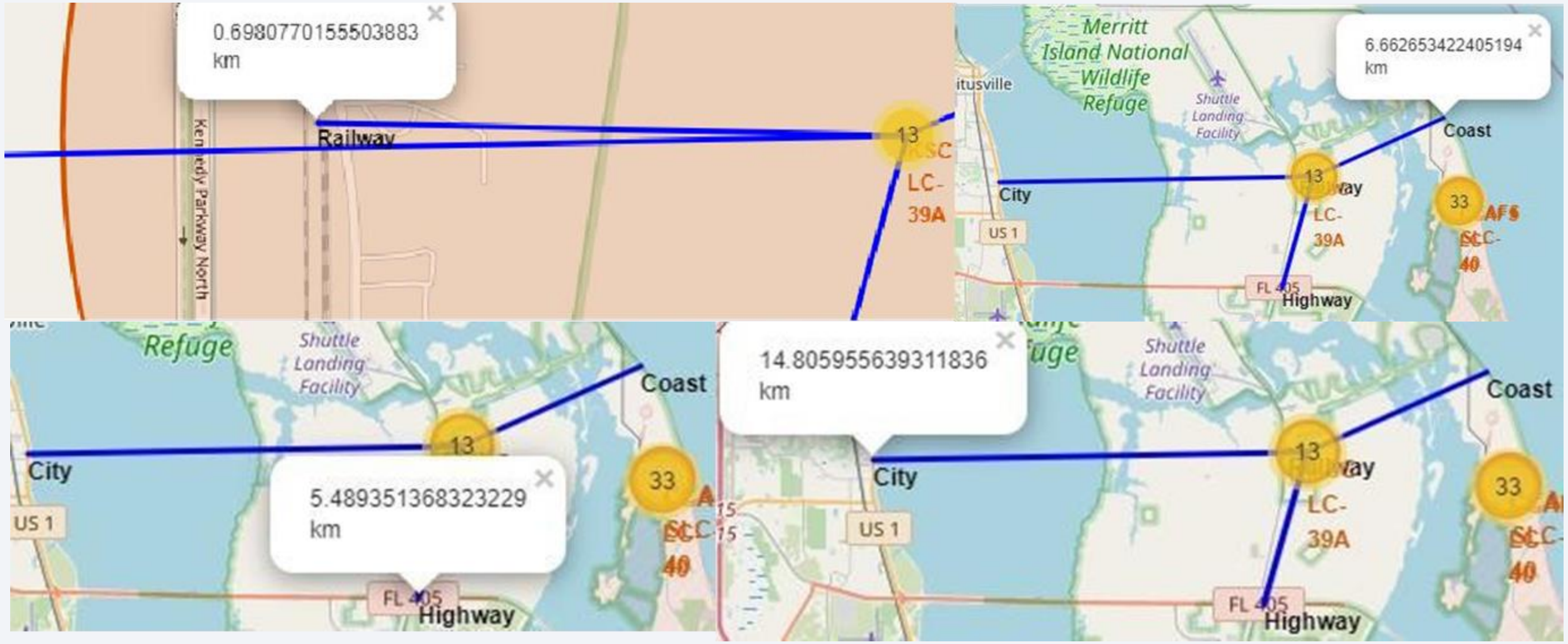
The map on the left illustrates the geographical locations of all launch sites across the United States. Meanwhile, the map on the right focuses specifically on the two launch sites in Florida, highlighting their proximity to each other. Notably, all launch sites are situated in close proximity to the ocean.

Colored-Launch Site Markers



On the Folium map, clusters can be clicked to view individual markers representing successful landings (green icon) and failed landings (red icon). For instance, in the case of VAFB SLC-4E, clicking on the cluster reveals a breakdown of 4 successful landings and 6 failed landings. This interactive feature provides detailed insights into the distribution and outcomes of launches at specific launch sites.

Key Locations Lines

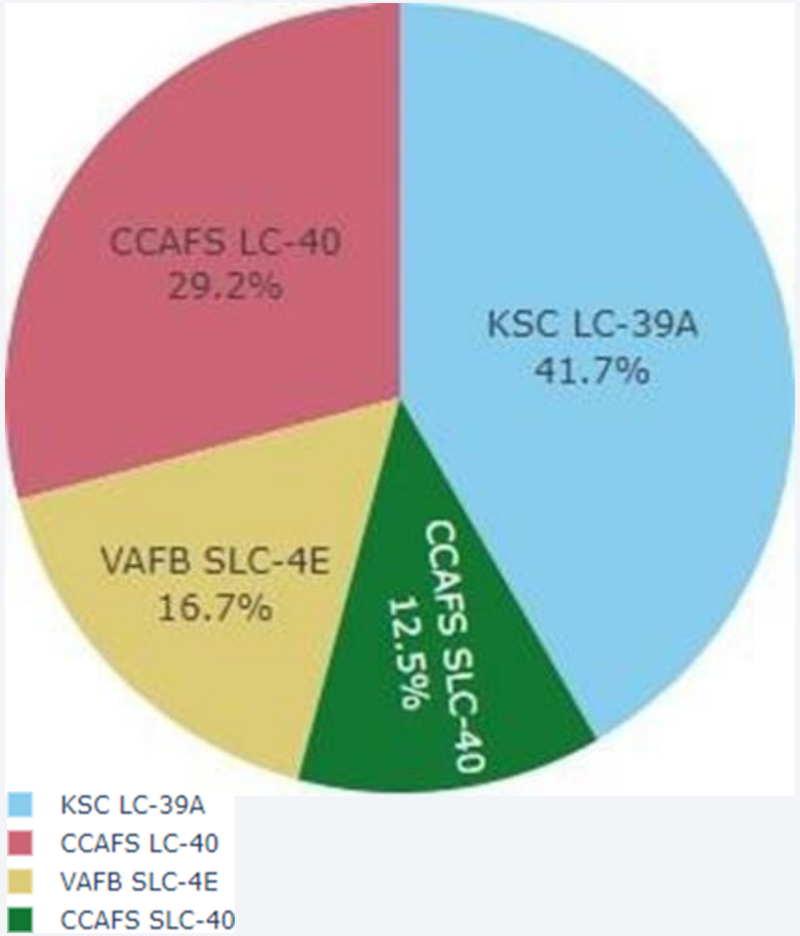


Taking KSC LC-39A as an example, launch sites are positioned in close proximity to railways, facilitating the transportation of large equipment and supplies essential for missions. They are also strategically located near highways, ensuring easy access for personnel and efficient supply transport. Additionally, these sites are situated near coastlines, which provides a safe area for rocket debris in case of launch failures, thereby minimizing risks to densely populated urban areas.

Section 4

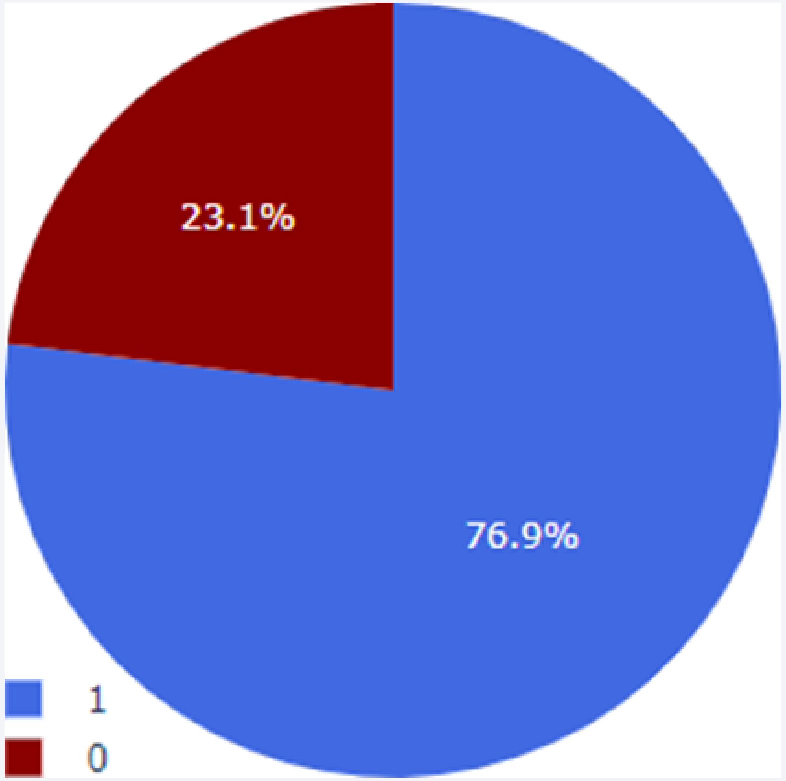
Build a Dashboard with Plotly Dash

Successful Launches Pie Chart



This represents the distribution of successful landings across all launch sites. CCAFS LC-40, which is the former name of CCAFS SLC-40, and KSC have an equal number of successful landings. However, a majority of these successful landings occurred before the name change. VAFB, on the other hand, has the smallest share of successful landings, possibly due to a smaller sample size and the increased challenges associated with launching from the west coast.

Launch Site with the Highest Succession Rate



Blue = Success
Red = Failure

KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

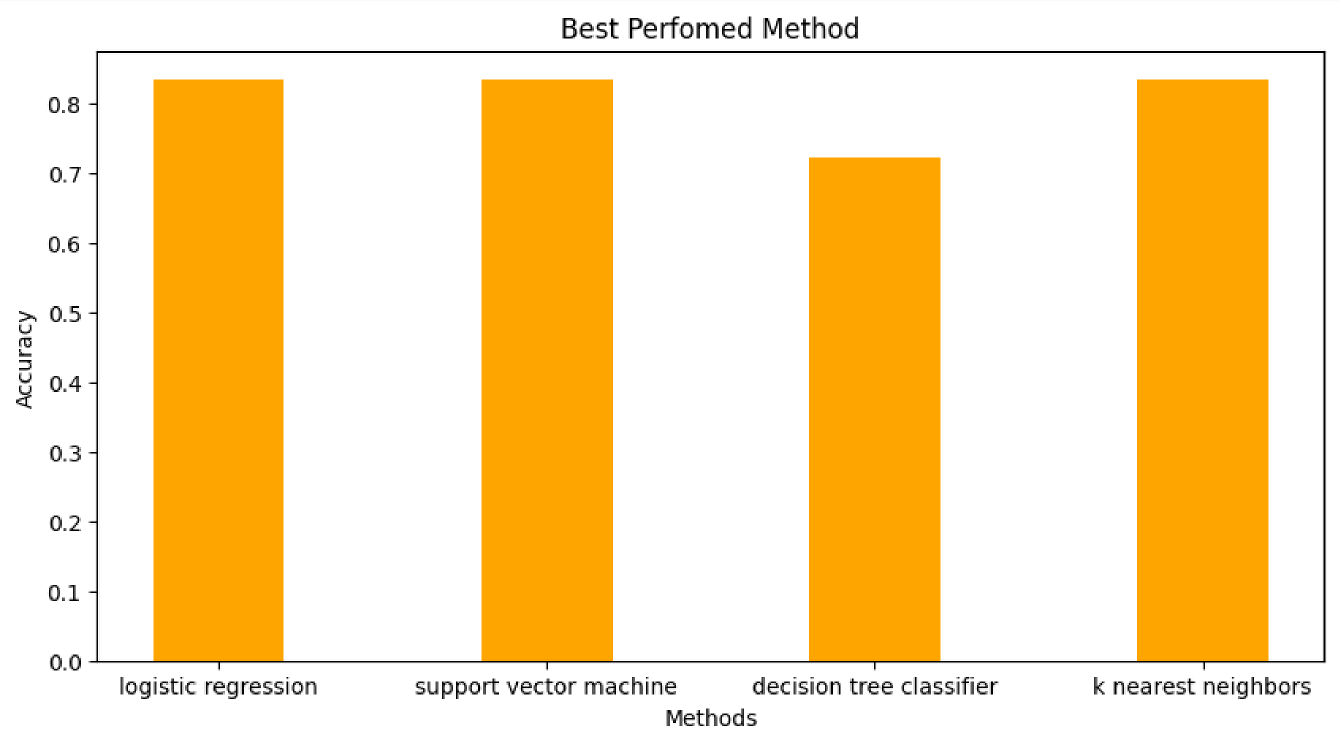
Payload Mass x Success

In the Plotly dashboard, there is a payload range selector that ranges from 0 to 10,000, which is below the maximum payload of 15,600 kg. The 'Class' attribute indicates 1 for successful landings and 0 for failures. The scatter plot also incorporates booster version categories by color and represents the number of launches with point size. Interestingly, within the specific payload range of 0 to 6,000 kg, there are two instances of failed landings despite payloads being recorded as zero kilograms.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



All models achieved a similar accuracy of 83.33% on the test set. However, the test set is quite small, with only 18 samples. This small sample size can result in significant variability in accuracy outcomes, as seen with the Decision Tree Classifier model in repeated tests. To accurately identify the best model, we likely need a larger dataset.

Confusion Matrix

Because all models showed identical performance on the test set, their confusion matrices are also identical. The models correctly predicted 12 successful landings and 3 unsuccessful landings. However, they incorrectly predicted 3 successful landings when the actual result was an unsuccessful landing (false positives). This suggests that our models have a tendency to overpredict successful landings.

Correct predictions are on a diagonal from top left to bottom right.

Conclusions

- Our task was to develop a machine learning model for SpaceY, aiming to compete with SpaceX. The objective was to predict the successful landing of Stage 1 to save approximately $100 million USD.
- We utilized data from a public SpaceX API and web-scraped the SpaceX Wikipedia page. We labeled the data and stored it in a DB2 SQL database, then created a dashboard for visualization.
- The machine learning model we developed achieved an accuracy of 83%. Allon Mask of SpaceY can use this model to predict, with relatively high accuracy, whether a launch will have a successful Stage 1 landing before it takes place, aiding in the decision to proceed with the launch.
- Collecting more data could help in identifying the best machine learning model and further improving accuracy.

Appendix

- GitHub Repository URL :

https://github.com/evanserlangga/IBMDSPFP

- References :

https://en.wikipedia.org/wiki/SpaceX

Thank you!