IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Delmon Arous
31.07.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - Predictive Analysis (Classification) using Machine Learning
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive Analytics in screenshots
  - Predictive Analysis result

Introduction

- Project background and context

  Space X 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 Space X 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 space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers

  - What factors determine if the rocket will launch and land successfully?
  - The interaction amongst various features that determine the success rate of a successful launch and landing
  - What operating conditions needs to be in place to ensure a launch and successful landing program

Section 1

Methodology

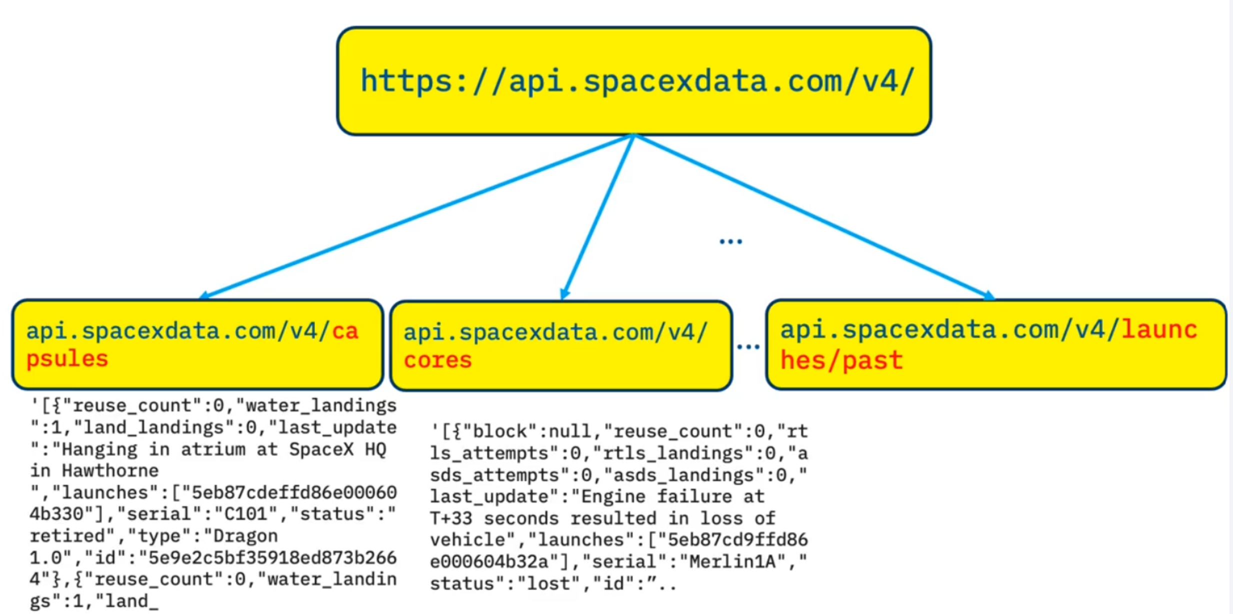
Methodology

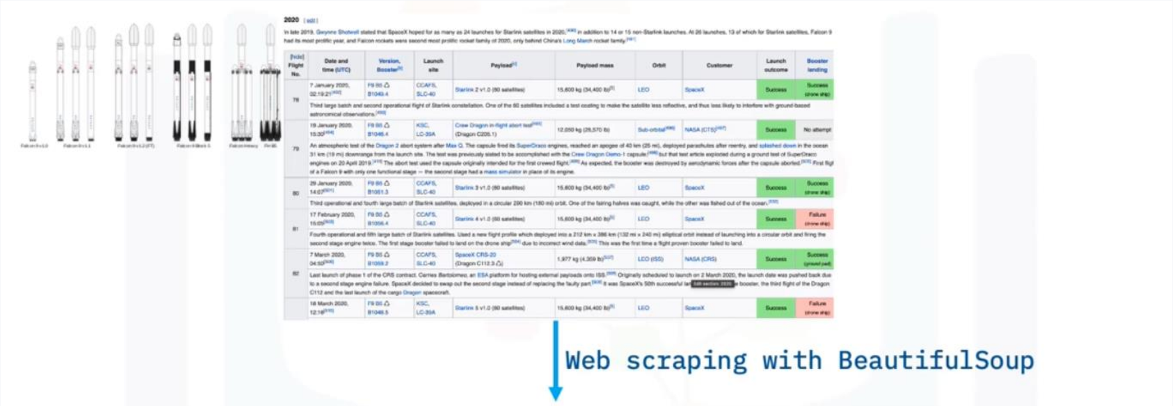
Executive Summary

- Data collection methodology:
  - The SpaceX launch data was gathered from the SpaceX REST API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) (BeautifulSoup)
- Perform data wrangling
  - Collected data was transformed into a clean dataset to make the dataset viable and meaningful for analysis (Pandas)
- 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 a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully (incl. preprocessing, standardizing data and dividing data into training and testing datasets for model training and evaluation)

Data Collection

- The SpaceX REST API (https://api.spacexdata.com/v4/rockets/) yield data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome
- Goal: use this data to predict whether SpaceX will attempt to land a rocket or not
- Web scraping Falcon 9 Launch records (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) with BeautifulSoup

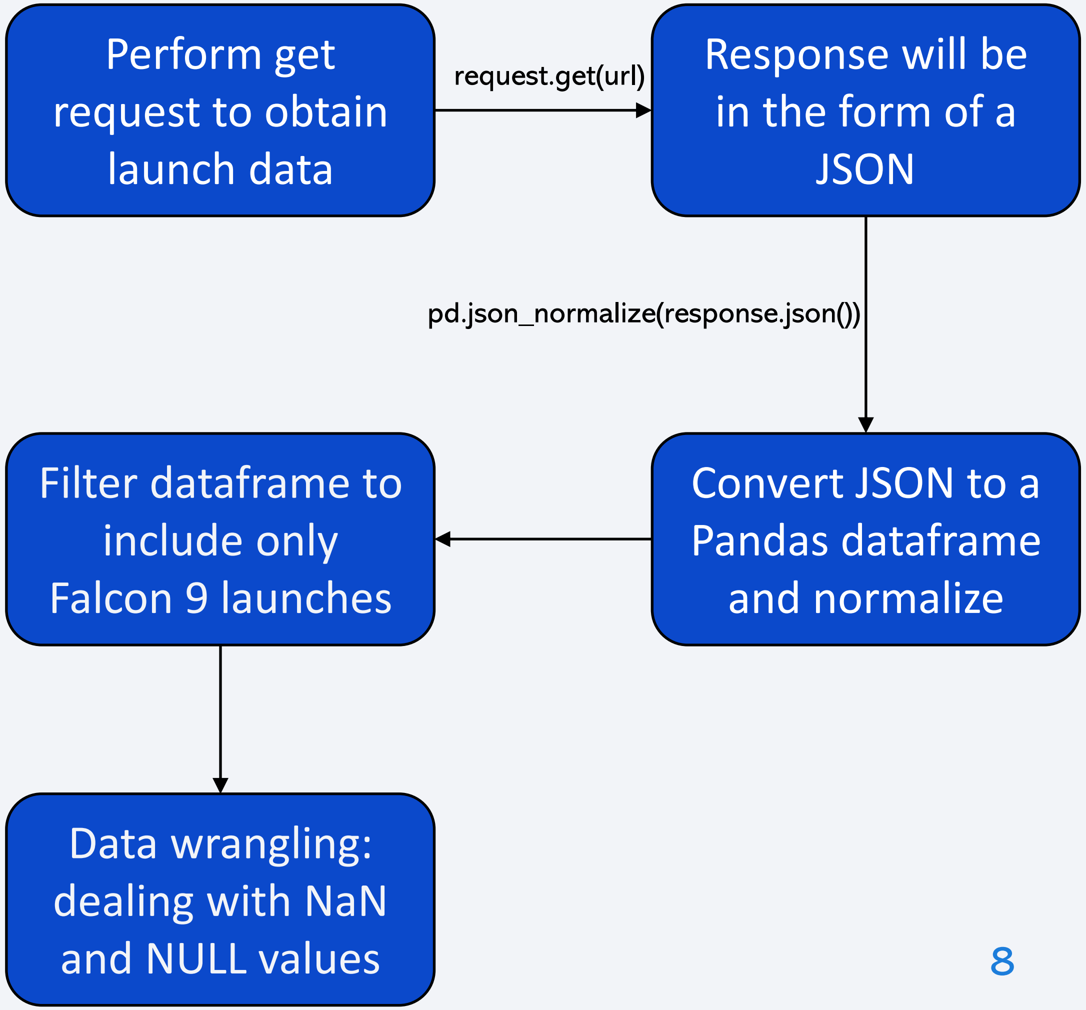




	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data Collection – SpaceX API

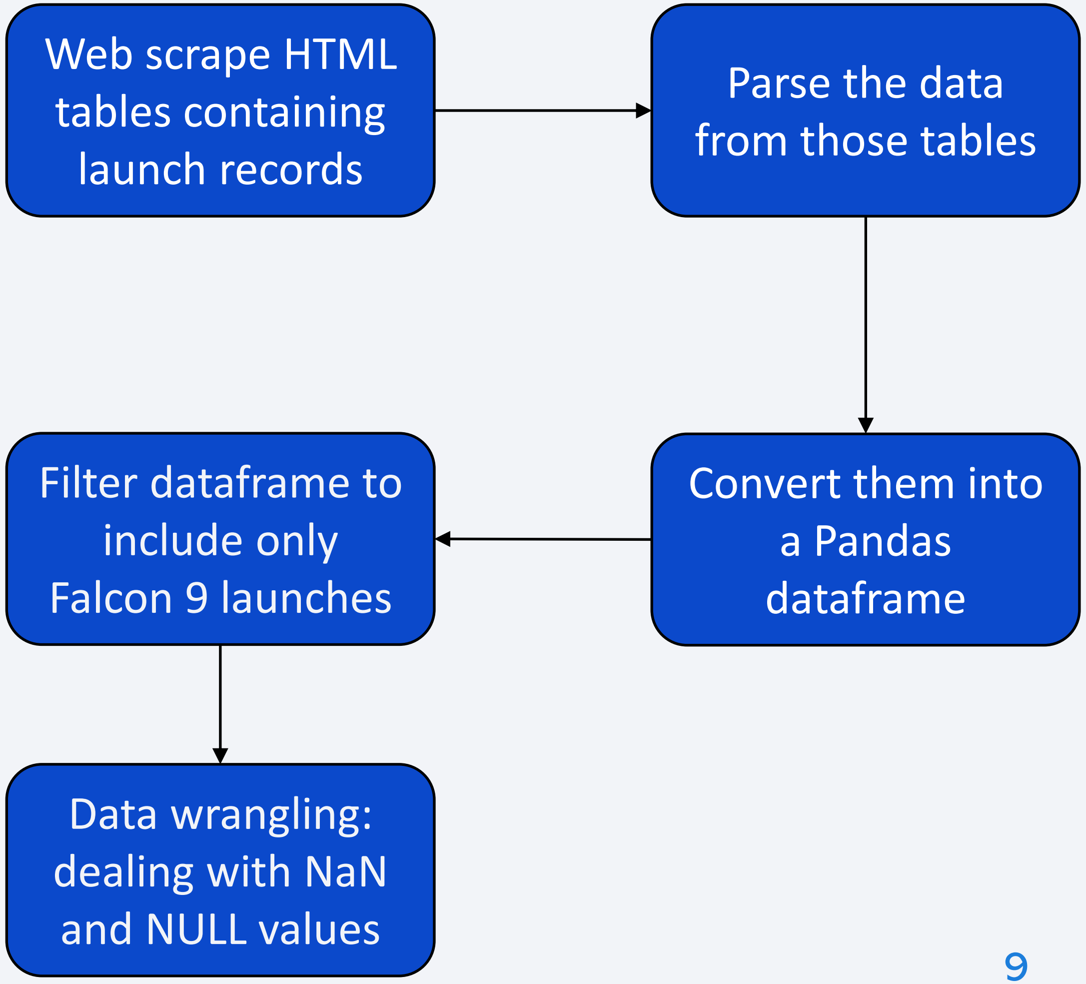
- SpaceX offers a public API from where data can be obtained and then used
- GitHub URL: https://github.com/tehm0n/Coursera/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Module%201%20-%20Introduction/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

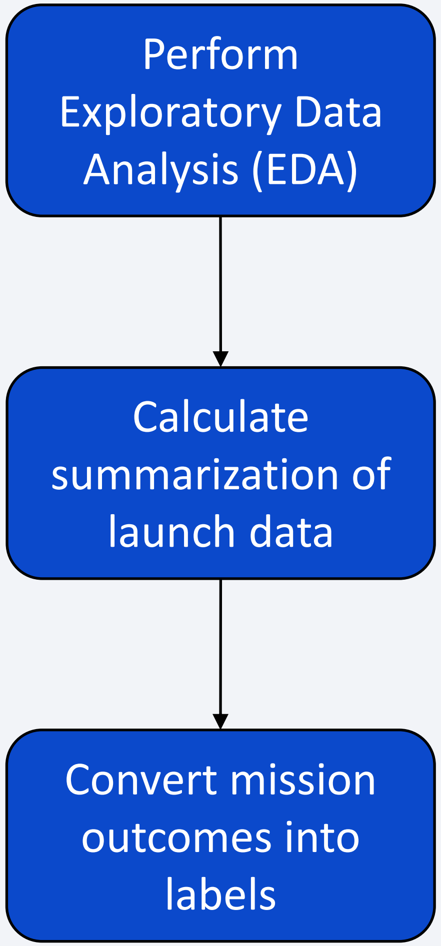
- Python BeautifulSoup package to web scrape some HTML tables from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) that contain valuable Falcon 9 launch records

- GitHub URL: https://github.com/tehm0n/Coursera/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Module%201%20-%20Introduction/jupyter-labs-webscraping.ipynb



Data Wrangling

- Perform some Exploratory Data Analysis (EDA); dealing with missing values
- Summarize launches on each site, occurrences of each orbit and occurrences of mission outcome per orbit type
- Convert mission outcomes into training labels (0/1)
- GitHub URL: https://github.com/tehm0n/Coursera/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Module%201%20-%20Introduction/labs-jupyter-spacex-data_wrangling_jupyterlite.ipynb



EDA with SQL

- Perform some Exploratory Data Analysis using a database with SQL
- The following SQL queries were 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 successful 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 through 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 between the date 2010-06-04 and 2017-03-20, in descending order
- GitHub URL: https://github.com/tehm0n/Coursera/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with Data Visualization

- Perform Exploratory Data Analysis (EDA) and Feature Engineering to predict if the Falcon 9 first stage will land successfully
- Employing Pandas and Matplotlib to explore data and to produce scatterplots, line plots and bar plots to visualize the relationship between pair of features:
  - Scatter plot: Flight number vs Launch site, Payload mass vs Launch site, Flight number vs Orbit type, Payload mass vs Orbit type
  - Bar plot: Orbit type vs Success rate
  - Line plot: Yearly trend of launch success
- GitHub URL: https://github.com/tehm0n/Coursera/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Module%202%20-%20Exploratory%20Data%20Analysis%20(EDA)/jupyter-labs-eda-dataviz.ipynb

Build an Interactive Map with Folium

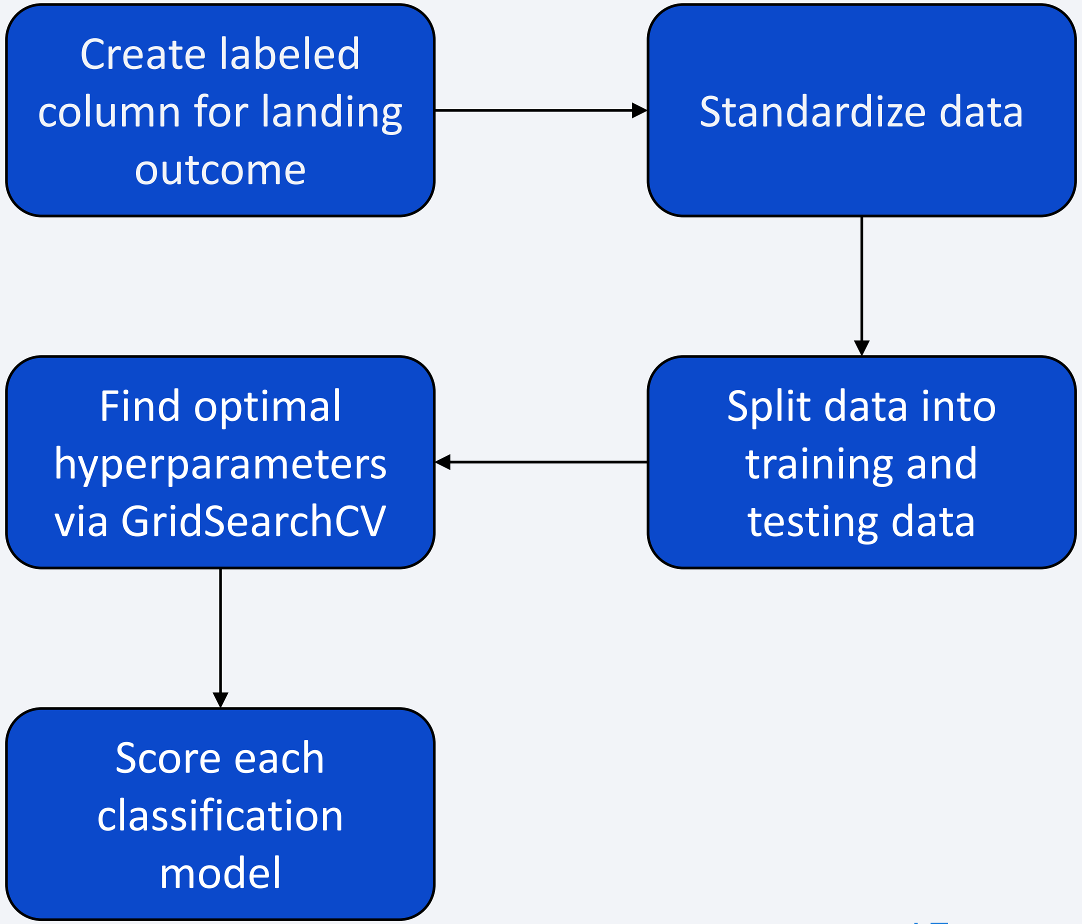
- Analyze launch site locations and proximities with Folium:
  - mark the launch site locations and their close proximities on an interactive map
  - explore the map with created markers and discover any underlying patterns
  - by that, explain how to choose an optimal launch site
- Tasks:
  - mark all launch sites on a Folium map by adding markers (points) and circles (highlighted areas) around a coordinate points
  - mark the success/failed launches for each site on the map by adding color-labeled markers in marker clusters
  - calculate the distances between a launch site to its proximities by adding lines (indicating distances between two coordinate points)
- GitHub URL: https://github.com/tehm0n/Coursera/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Build a dashboard application with the Python Plotly Dash package
- Contains input components such as a dropdown list (for launch site selection) and a range slider (for payload range selection) to interact with:
  - a pie chart showing the total launches for each launch site
  - a scatter point chart showing the correlation between payload mass and launch success for different booster version
- Can be employed to find insights from the SpaceX dataset more easily than with static graphs
- GitHub URL: https://github.com/tehm0n/Coursera/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Module%203%20-%20Interactive%20Visual%20Analytics%20and%20Dashboard/spacex_dash_app.py

Predictive Analysis (Classification)

- Build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully. This includes:
  - Preprocessing, allowing to create a column for landing outcome labels and standardize data (NumPy, Pandas, Scikit-Learn)
  - Split into training data and testing data (Scikit-Learn)
  - Classification model training by performing a Grid Search, allowing to find the hyperparameters that yields most optimal performance for the model algorithm (Scikit-Learn)
  - Using the best hyperparameter values, find the classification model with the best accuracy using the testing data (Scikit-Learn)
- Classification models tested: logistic regression, support vector machines, decision tree classifier and K-nearest neighbors
- GitHub URL: https://github.com/tehm0n/Coursera/blob/main/Course%2010%20-%20Applied%20Data%20Science%20Capstone/Module%204%20-%20Predictive%20Analysis%20(Classification)/SpaceX_Machine_Learning_Prediction_Part_5.ipynb



Results

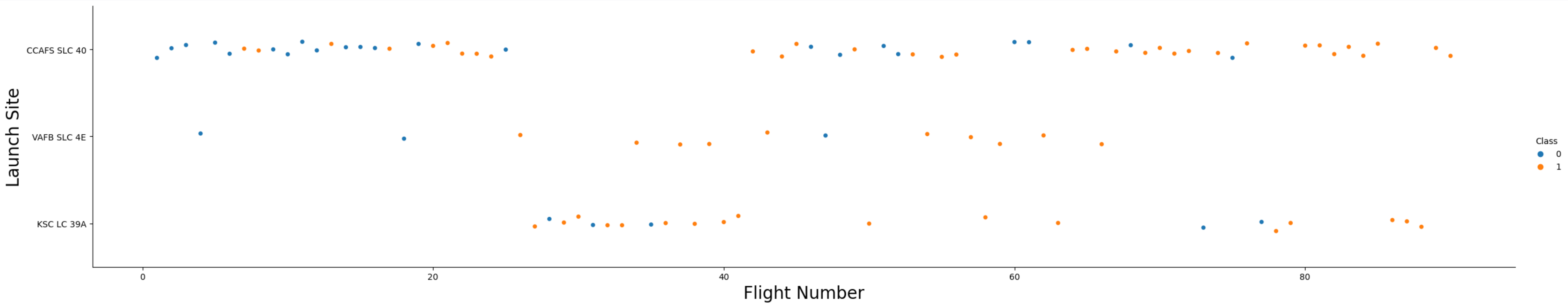
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

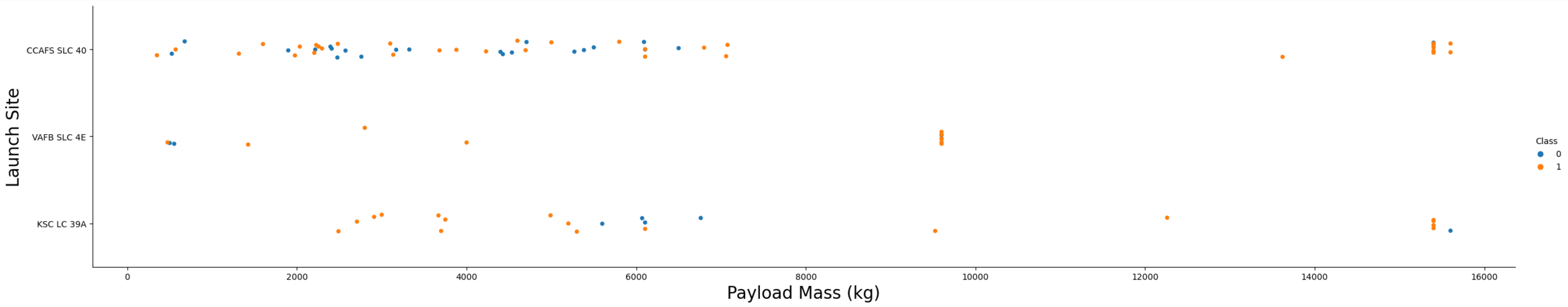
Flight Number vs. Launch Site

- The larger the flight amount at a launch site, the greater the success rate at a launch site



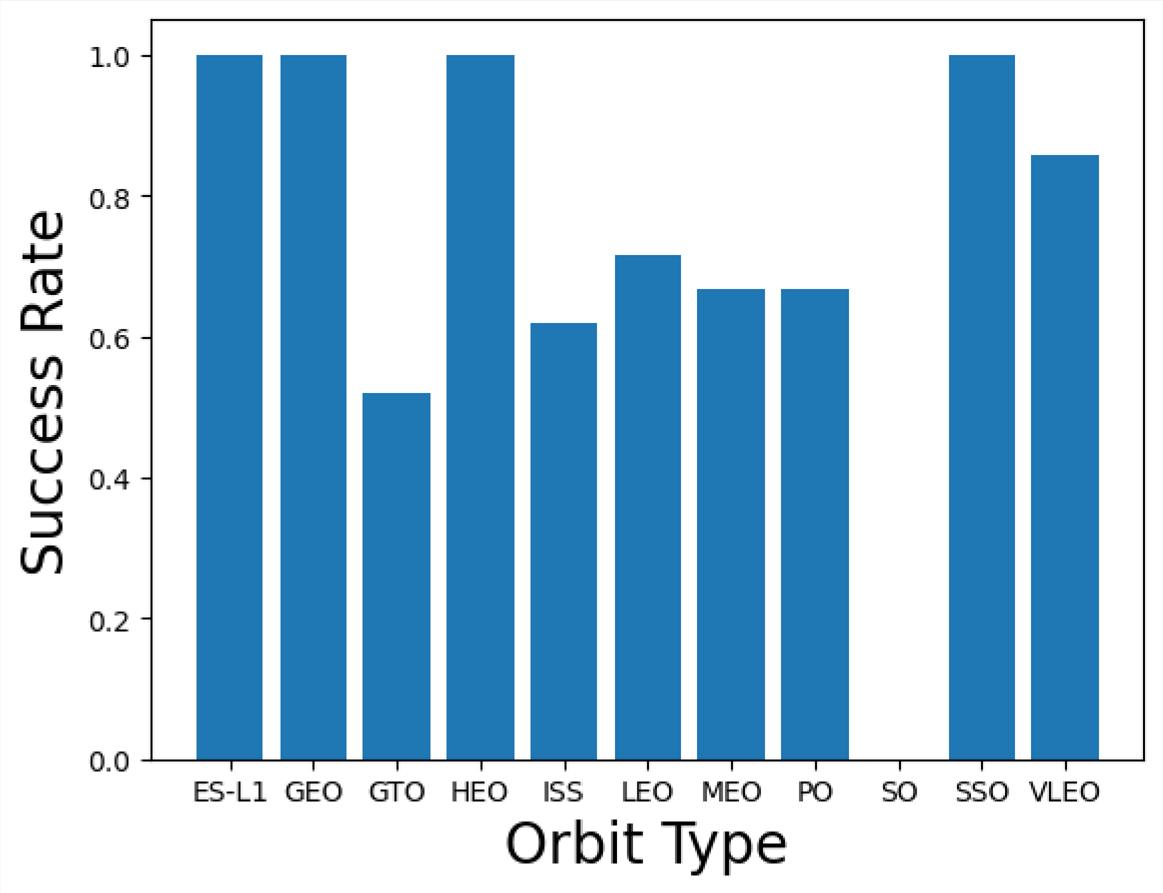
Payload vs. Launch Site

- For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000)
- The greater the payload mass for launch site CCAFS SLC 40, the higher the success rate for the rocket



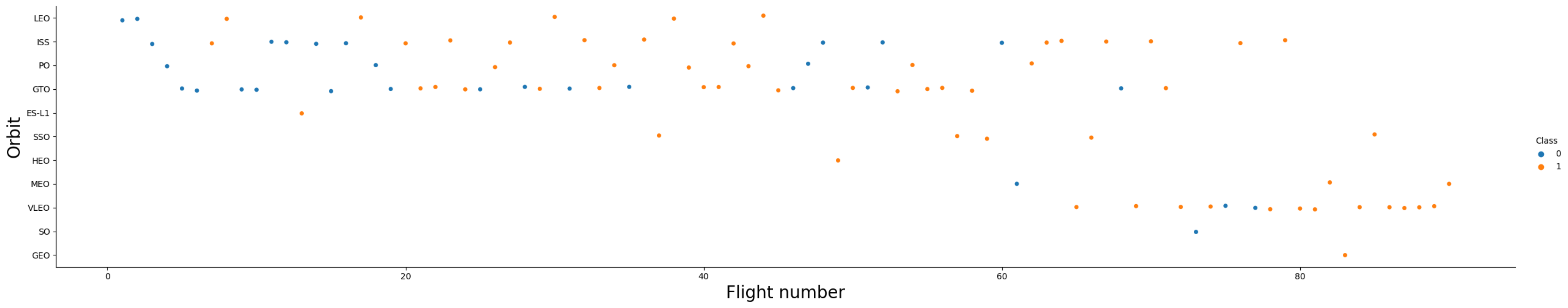
Success Rate vs. Orbit Type

- The orbits ES-L1, GEO, HEO, SSO and VLEO have the highest success rates



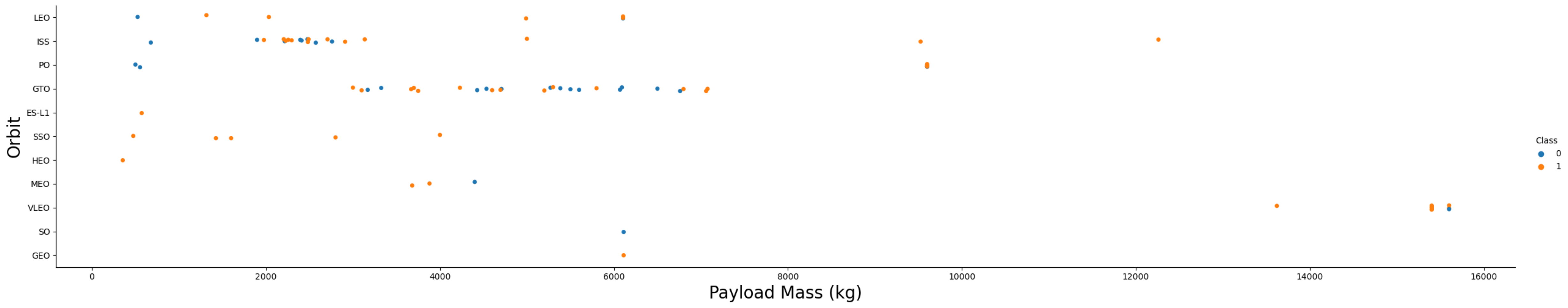
Flight Number vs. Orbit Type

- For the LEO orbit, the success appears related to the number of flights (higher success for increasing flight number)
- For the GTO orbit, there is no relationship between the orbit and flight number
- Generally, the success rate increases with increasing flight number for each orbit



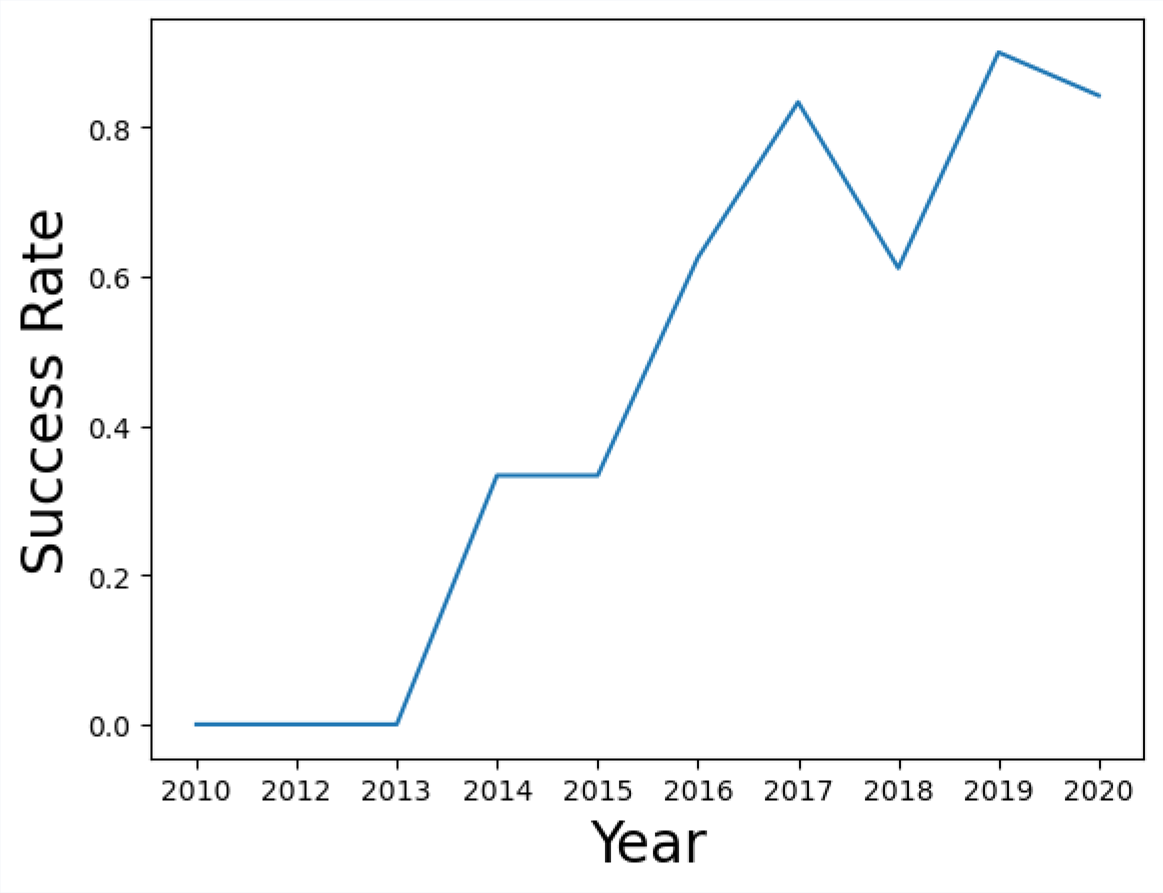
Payload vs. Orbit Type

- With heavy payloads, the successful landing rate is more for the Polar, LEO and ISS orbit
- For the GTO orbit, cannot distinguish relationship with payload as successful and unsuccessful landing rate are prominent



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

- DISTINCT function is employed to display the unique names of the launch sites in the space mission: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

Task 1

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

In [8]:
```
%%sql
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with 'CCA'
- The combination of WHERE clause and the wildcard LIKE "%CCA%" was used to find relevant launch sites

Task 2

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

In [9]:

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

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

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No atte
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No atte
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

- Calculate the total payload carried by boosters from NASA: 45596.0 kg
- The SUM() function was used to calculate total payload mass, while the WHERE clause was used to filter for boosters launched by NASA (CRS)

Task 3

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

In [10]:

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass by NASA (CRS) (kg)"
FROM SPACEXTABLE
WHERE Customer = "NASA (CRS)";
```

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

Out[10]:

Total Payload Mass by NASA (CRS) (kg)
45596.0

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1: 2534.67 kg
- The AVG() function was used to calculate average payload mass, while the combination of WHERE clause and the wildcard LIKE “F9 v1.1%” was used to filter for rockets with booster version F9 v1.1

Task 4

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

In [11]:

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload Mass Carried by Booster Version F9 v1.1 (kg)"
FROM SPACEXTABLE
WHERE Booster_Version
LIKE "F9 v1.1%";
```

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

Out[11]:

Average Payload Mass Carried by Booster Version F9 v1.1 (kg)
2534.6666666666665

First Successful Ground Landing Date

- Find the date of the first successful landing outcome on ground pad: 22/12/2015
- The MAX() function and the WHERE clause is used to find the earliest successful landing outcome on ground pad

Task 5

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

Hint:Use min function

In [12]:

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

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

Out[12]:

MAX(Date)
22/12/2015

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: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2
- Employed a WHERE clause to filter which booster have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

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 [13]:
```
%%sql
SELECT DISTINCT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (drone ship)" AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

Out[13]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696.0
F9 FT B1026	4600.0
F9 FT B1021.2	5300.0
F9 FT B1031.2	5200.0

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes: Failure in flight (1), Success (99), Success with payload status unclear (1)
- The COUNT() function and the GROUP BY clause is used to calculate the total number of mission outcomes

Task 7

List the total number of successful and failure mission outcomes

In [14]:
```
%%sql
SELECT Mission_Outcome, COUNT(*) AS "Count"
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

Out[14]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- A subquery in the WHERE clause and the MAX() function is employed to find the relevant boosters

Task 8

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

In [15]:

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

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

Out[15]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Used a combinations of the WHERE and AND clause to filter for failed landing outcomes in drone ship for year 2015

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [16]:

```
%%sql
SELECT SUBSTR(Date,4,2) AS Month, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = "Failure (drone ship)"
AND SUBSTR(Date,7,4) = "2015";
```

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

Out[16]:

Month	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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
- The COUNT() function is used to count the different landing outcomes, the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2017-03-20, the GROUP BY clause to classify landing outcomes and the ORDER BY clause to order the grouped landing outcomes in descending order

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 [17]:

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

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

Out[17]:

Landing_Outcome	Count
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

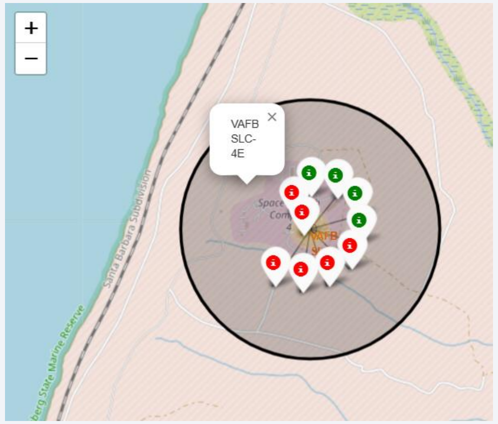
All launch sites marked on a Folium map

- All SpaceX launch sites are located near the coast in the Unites States; 3 sites in Florida and 1 site in California
- Markers are (longitude, latitude) coordinate points and circles are highlighted areas around each coordinate point

Successful/failed launches for each site

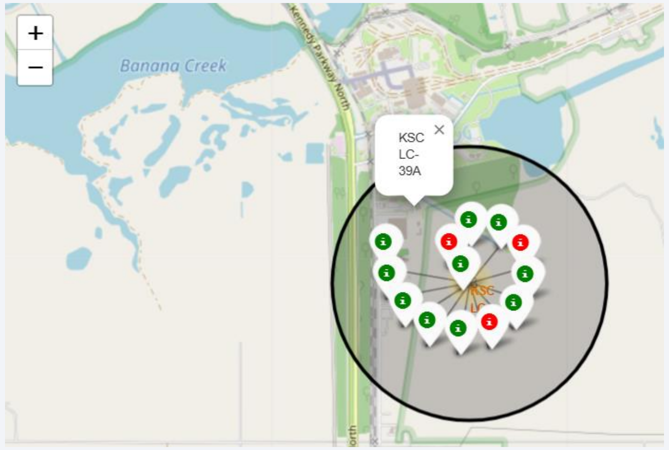
- Success and failed launches for each launch site on the Folium map is displayed using color-labeled markers in marker clusters
- Green markers represent successful launches, red markers represents failed launches

Site VAFB SLC-4E

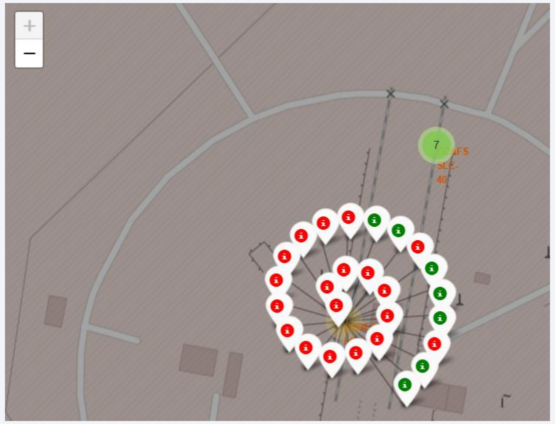


Launch site in California

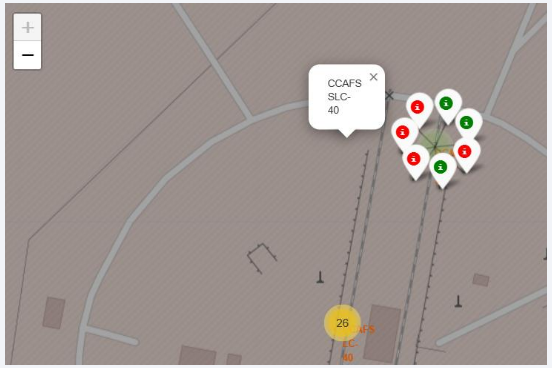
Site KSC LC-39A



Site CCAFS LC-40



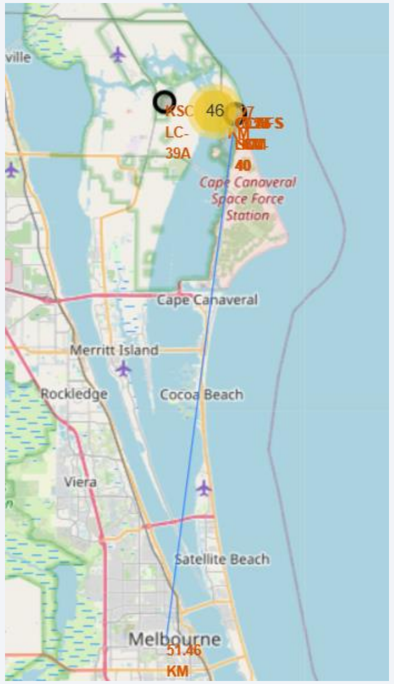
Site CCAFS SLC-40



Launch sites in Florida

Proximity calculations from launch site

- Estimate and display distances between a selected launch site (CCAFS SLC-40) to its proximities by adding lines (indicating distances between two coordinate points)
- Distance between the CCAFS SLC-40 site to its nearest proximities: coastline (0.86 km), railway (NASA Railroad; 1.27 km), highway (Samuel C Philips Pkwy; 0.58 km), city (Melbourne; 51.46 km)

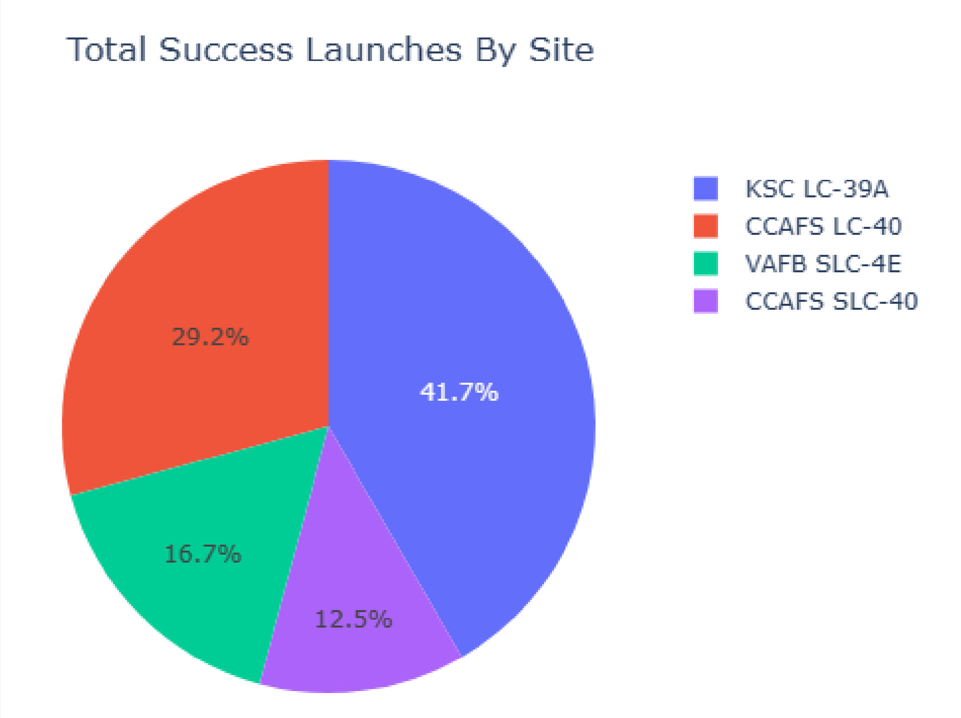


Section 4

Build a Dashboard with Plotly Dash

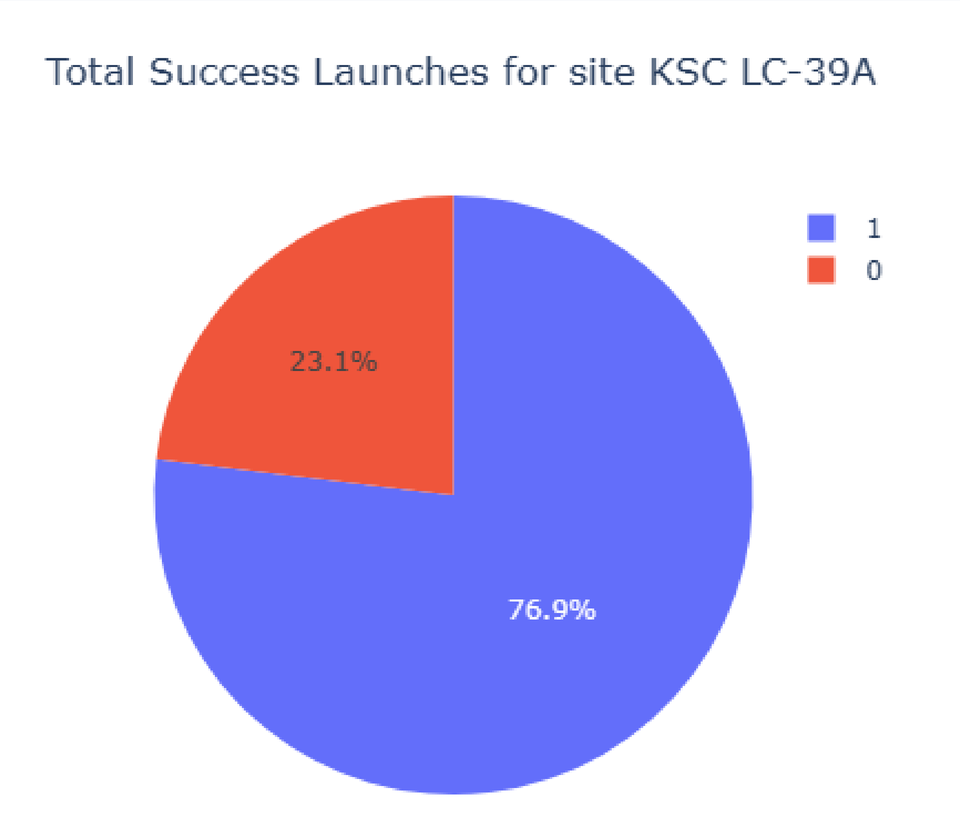
Launch Success Count for All Sites

- Pie chart of total successful launches by site
- The KSC LC-39A has the highest launch success ratio of all the sites



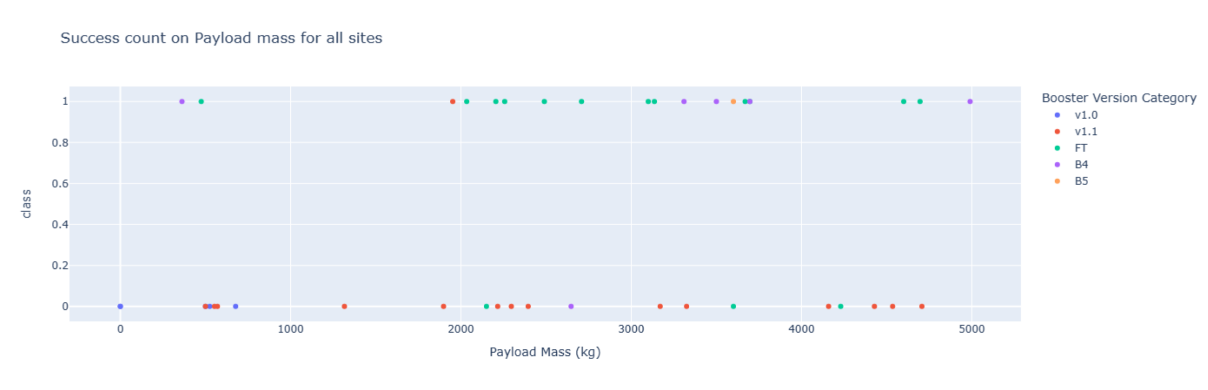
Total Successful launches for Launch Site KSC LC-39A

- Pie chart for the launch site (KSC LC-39A) with highest launch success ratio
- The KSC LC-39A site achieved a 76.9% success rate (blue) while getting a 23.1% failure rate (red)

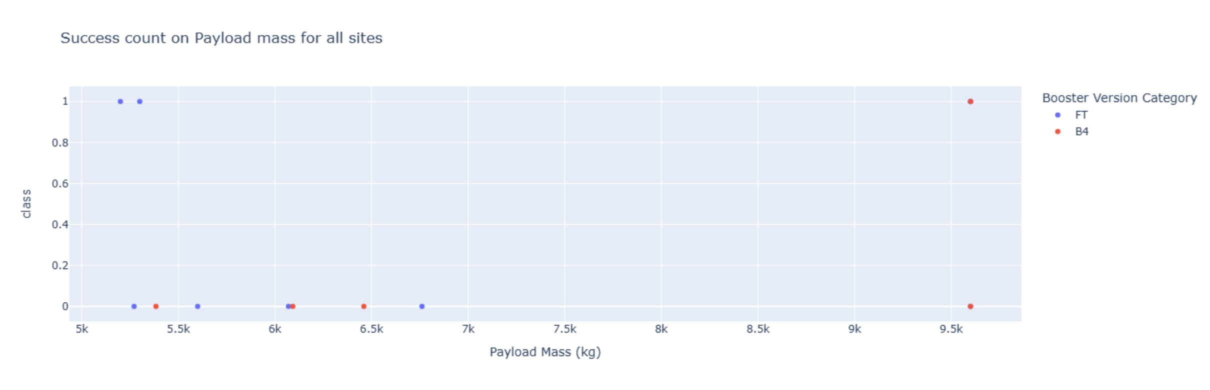


Payload Mass vs Launch Outcome for all Launch Sites

- Payload (kg) vs. Launch Outcome (0/1) scatter plot for all launch sites, with different payload selected in the range slider
- For all booster versions, the success rate with light-weighted payloads is higher than with heavy-weighted payloads
- For low payload mass; booster version FT has the highest successful launch count, while booster version v1.1 has the most unsuccessful count
- For high payload mass; booster version FT has the highest successful launch count, while both booster version FT and B4 has equally unsuccessful counts



Payload range: 0 – 5000 kg (low payload mass)



Payload range: 5000 – 10000 kg (high payload mass)

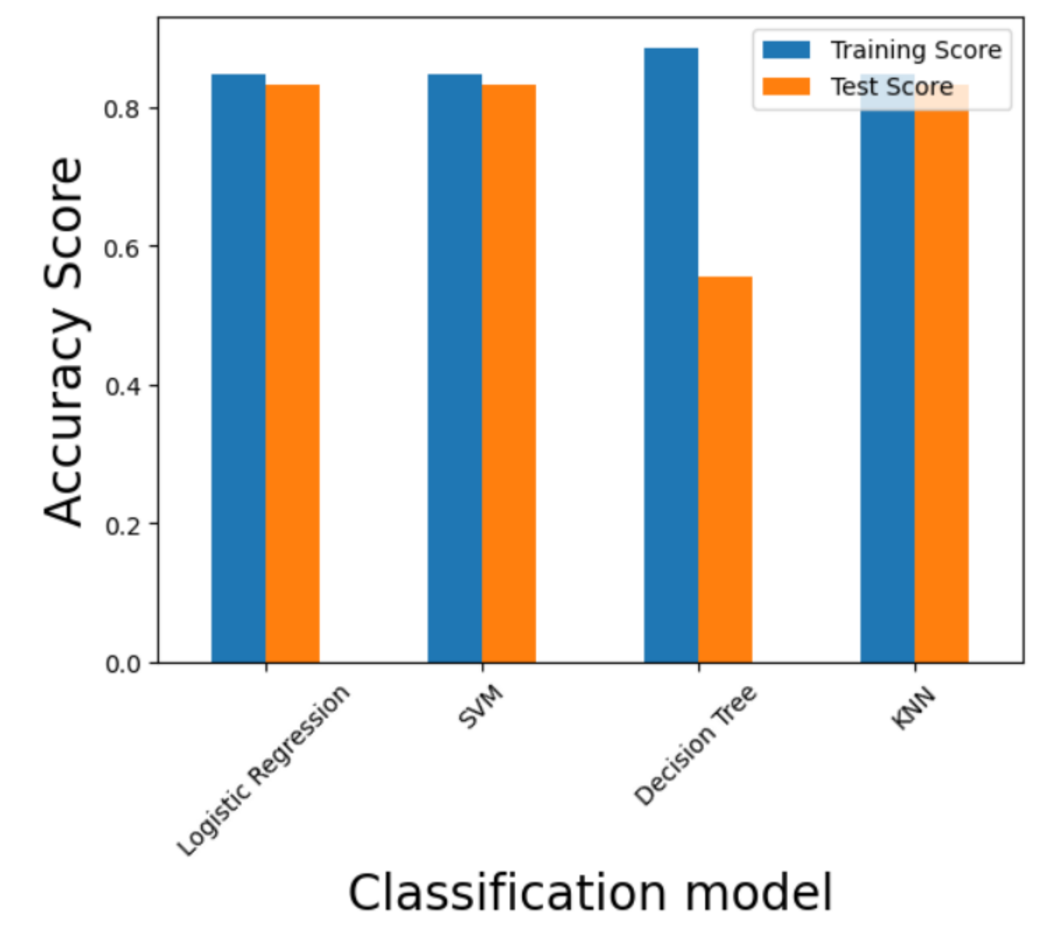
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classifier yields the best accuracy score for the training dataset, but worst accuracy score for the testing dataset
- Logistic Regression, SVM and KNN yields the best accuracy score for the testing dataset

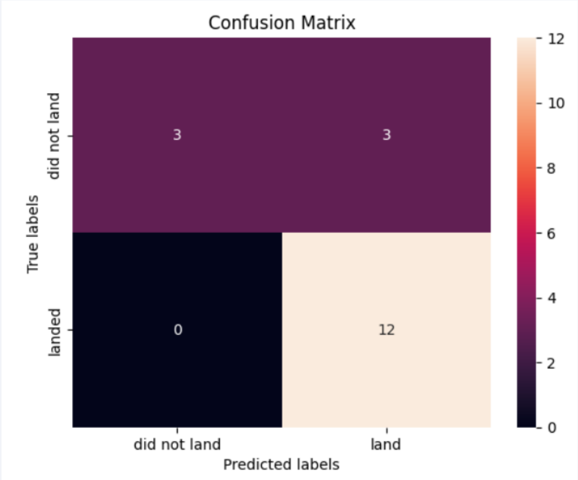
	Method	Training Score	Test Score
0	Logistic Regression	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Decision Tree	0.885714	0.555556
3	KNN	0.848214	0.833333



Confusion Matrix

- The confusion matrix for Logistic Regression, SVM and KNN shows that the methods can distinguish between the different classes. The major problem is the false positives for the test dataset, .i.e. unsuccessful landing marked as successful landing by the classifier

- The confusion matrix for the Decision Tree shows that the classifier struggles with false positives and false negatives for the test dataset



Logistic Regression

SVM

KNN

Decision Tree

Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site
- Launch success rate started to increase in 2013 till 2020
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate
- KSC LC-39A had the most successful launches of any sites
- The Decision tree classifier is the best machine learning algorithm for this task

Thank you!