IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nadine Forster
04.12.2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
- Collecting the Data
- Data Wrangling
- Exploratory Analysis Using SQL
- Exploratory Analysis Using Pandas and Matplotlib
- Interactive Visual Analytics with Folium
- Interactive Dashboard with Ploty Dash
- Predictive Analysis (Classification)

- Summary of all results
- Exploratory Data Analysis results
- Predictive results

Project background and context

SpaceX is a rocket company that has made launching satellites much cheaper. They charge $62 million for a Falcon 9 rocket launch, while other companies charge over $165 million. The big reason for this lower cost is that SpaceX can reuse the first stage of their rockets, saving about 70% of the cost. If we can predict whether the first stage will land and be reused, we can figure out how much a launch will cost. Using public data and machine learning, we will try to predict if SpaceX will reuse the first stage of their rockets

Problems I want to find answers

Factors such as the launch site, payload mass, orbit type, landing pad location, and booster version influence the success of the first stage landing, and understanding their impact is important.

Identifying the best algorithm for predicting whether a booster will successfully land (binary classification) is a key goal.

Introduction

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

Data collection through API requests from SpaceX REST API & Web Scraping from SpaceX's Wiki page

Perform data wrangling

Data was processed through extracting load & perform

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Data Collection

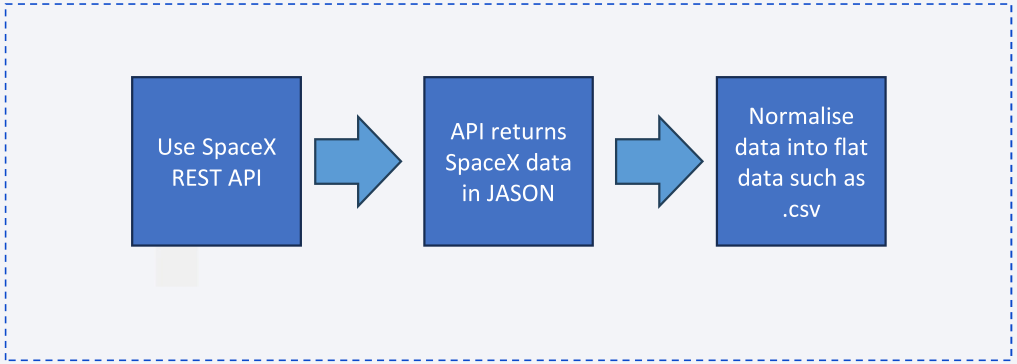
Used REST API to extract data in form of JSON

Transformed to data frame

Used inbuilt Python Pandas method for normalization

Data Collection – SpaceX API

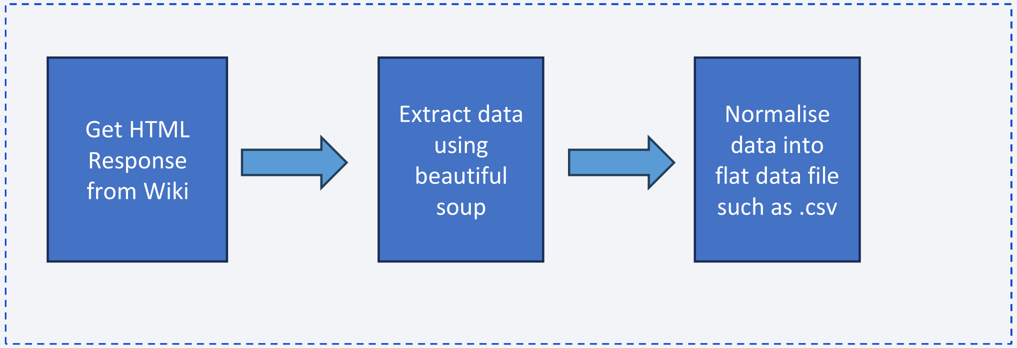
Flowchart of data collection with SpaceX REST calls



GitHub URL of the completed SpaceX API calls notebook: https://github.com/NadineForster/Capstone/blob/master/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping

Flowchart of web scraping process

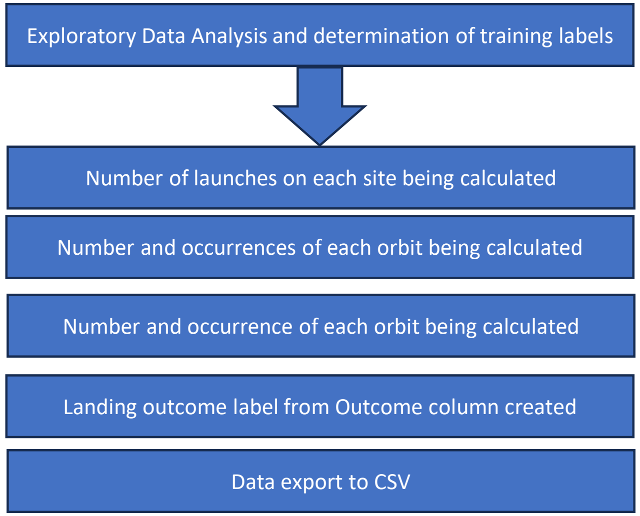


GitHub URL of the completed web scraping notebook:

https://github.com/NadineForster/Capstone/blob/master/jupyter-labs-webscraping.ipynb

Data Wrangling

Flowcharts data wrangling process using key phrases and flowcharts



GitHub URL of completed data wrangling related notebooks:
https://github.com/NadineForster/Capstone/blob/master/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

Charts that were plotted

Flight Number vs Payload Mass

Flight Number vs Launch Site

Payload Mass vs Launch Site

Orbit Type vs Success Rate

Flight Number vs Orbit Type

Payload Mass vs Orbit Type

Success Rate Yearly Trend

Reasons why we used those charts

Scatter plots reveal relationships between variables, helping to determine their suitability for machine learning models.

Bar charts compare discrete categories to highlight differences in their associated values.

Line charts depict trends over time, making them ideal for time series analysis.

GitHub URL of completed EDA with data visualization notebook:
https://github.com/NadineForster/Capstone/blob/master/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- SQL queries performed
- Unique Launch Sites
- Filter Launch Sites
- Total NASA (CRS) Payload
- Average Payload for F9 v1.1
- First Ground Pad Success
- Boosters with Specific Payload
- Mission Outcomes
- Max Payload Booster
- Drone Ship Failures
- Landing Outcome Rankings

GitHub URL of completed EDA with SQL notebook:
https://github.com/NadineForster/Capstone/blob/master/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The following map objects were used

Use of for NASA marker with a circle, popup, and text label

Johnson Space Center using its latitude and longitude as the starting location for all launch sites to show their locations and how close they are to the equator and coasts.

Use of coloured markers

Green markers for successful launches and red markers for failed launches, grouped in a Marker Cluster to show which sites have more successes

Use of coloured lines

To show distances from Launch Site KSC LC-39A (as an example) to nearby features like railways, highways, coastlines, and the closest city

GitHub URL of completed interactive map with Folium map: https://github.com/NadineForster/Capstone/blob/master/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

Added a dropdown list to enable Launch Site selection.

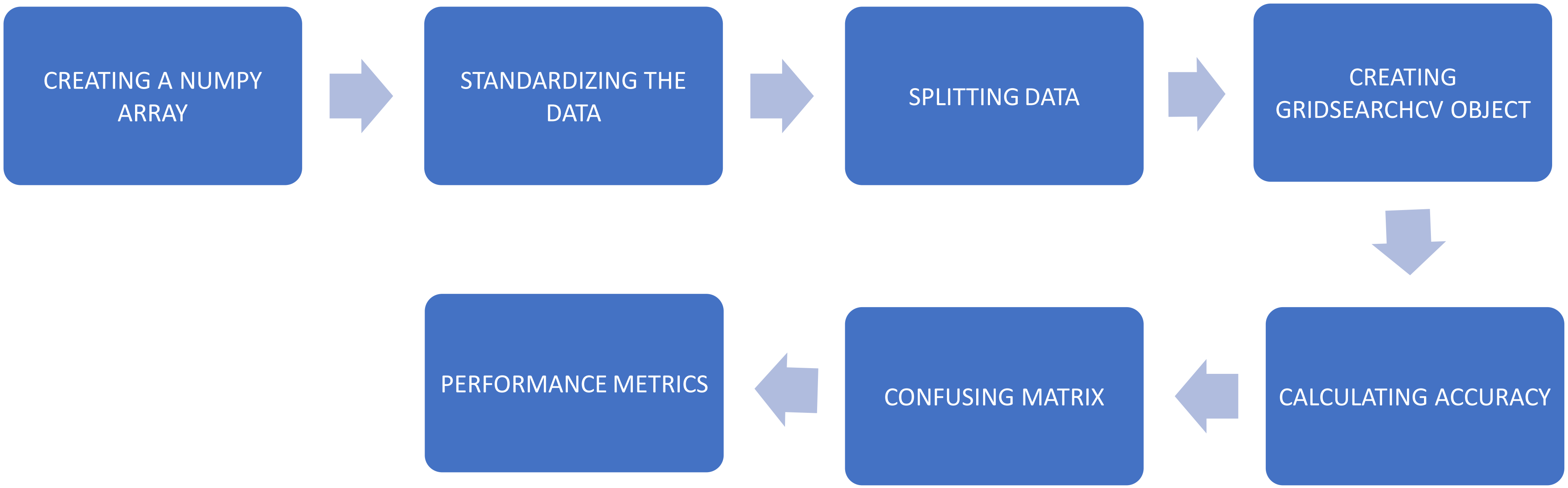
Added a pie chart to show the total successful launches count for all sites

- Added a slider to select Payload range.

- Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub URL of completed Plotly Dash lab:
https://github.com/NadineForster/Capstone/blob/master/main.py

Predictive Analysis (Classification)



GitHub URL of completed predictive analysis lab:
https://github.com/NadineForster/Capstone/blob/master/jupyter-labs-spacex-data-collection-api-v2.ipynb

Results

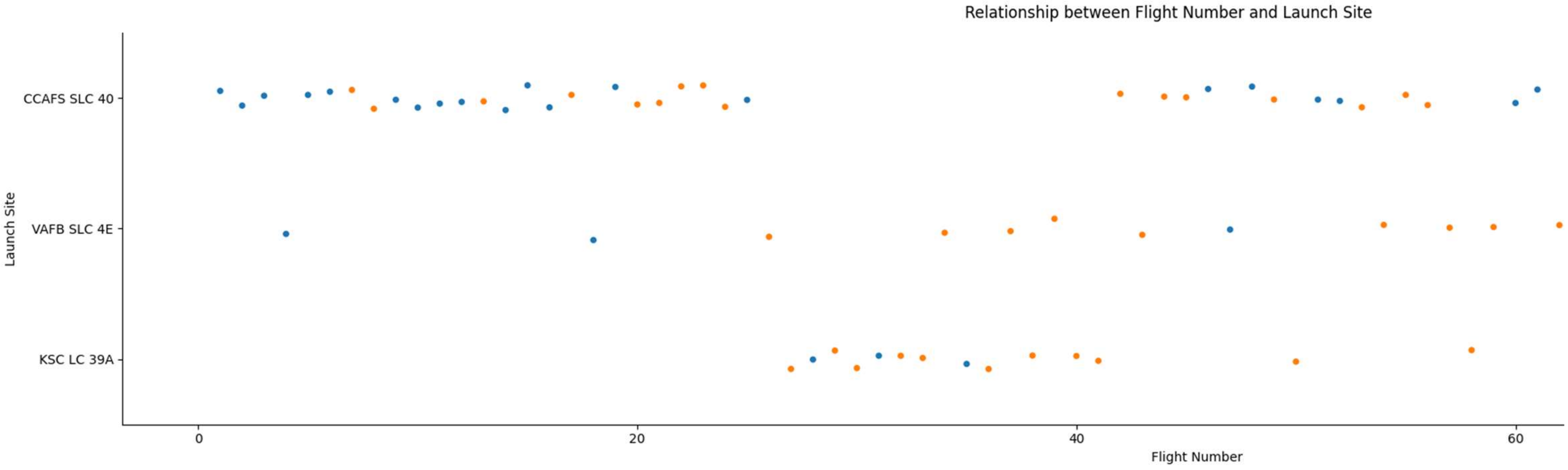
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

Section 2

Insights drawn from EDA

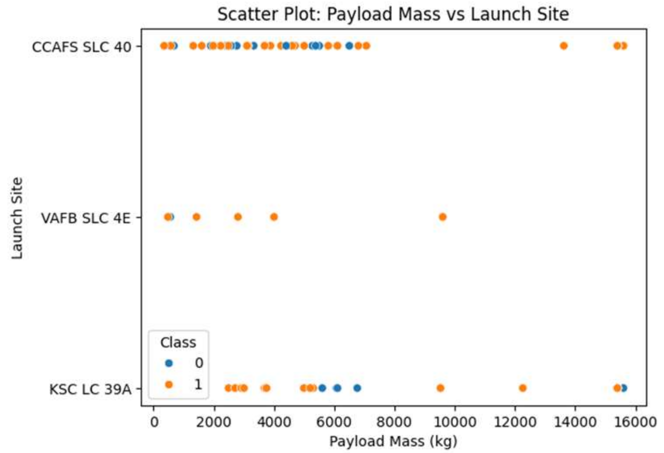


Flight Number vs. Launch Site

- Earliest flights all failed whereas the latest flights succeeded.
- VADB SLC 4E and KSC LC 39A have higher success rates

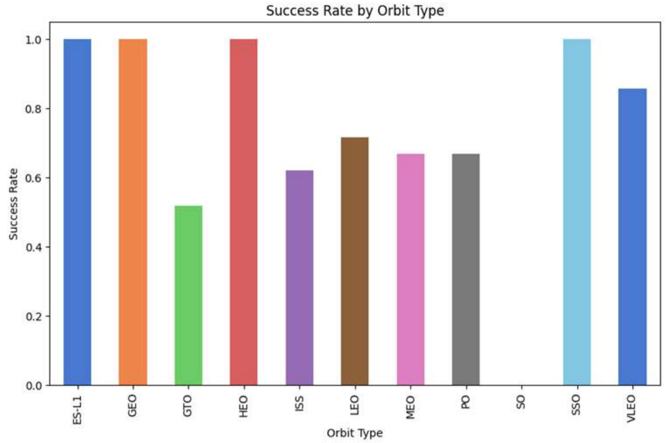
Payload vs. Launch Site

- Almost all the launches with payload mass over 7k kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 6k kg.



Success Rate vs. Orbit Type

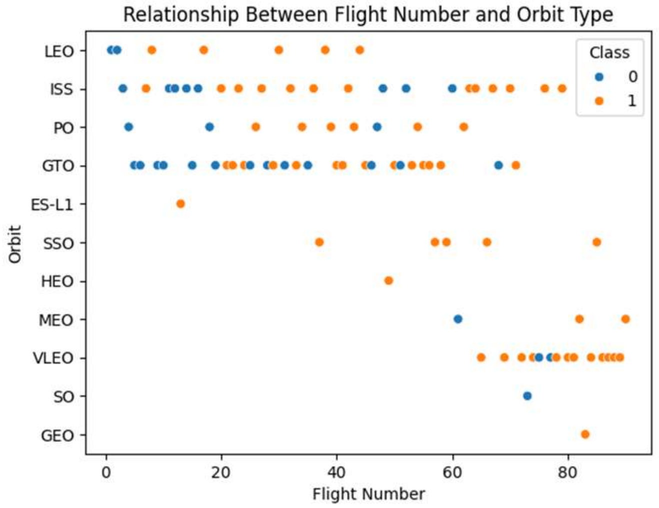
- ES-L1, GEO, HEO, SSO – these are orbits with 100 % success rate
- SO – this is one with 0% success rate
- GTO, ISS, LEO, MEO,PO – these are orbits with 50-85% success rate



Flight Number vs. Orbit Type

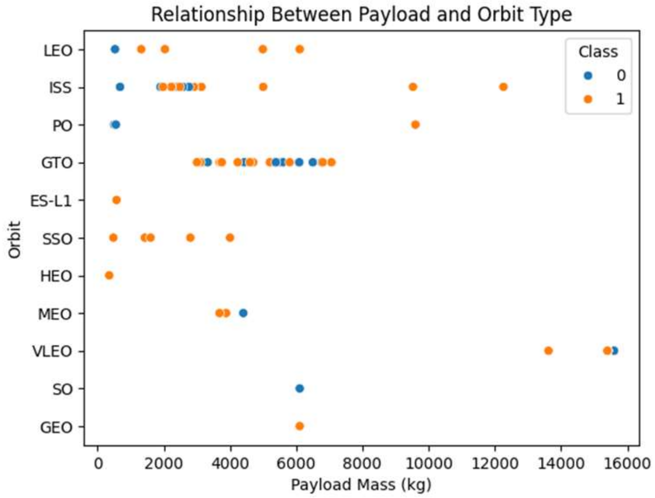
- The success rate in the LEO orbit appears to be related to the number of flights

- There does not appear to be any correlation between flight number and success rate in the GTO orbit



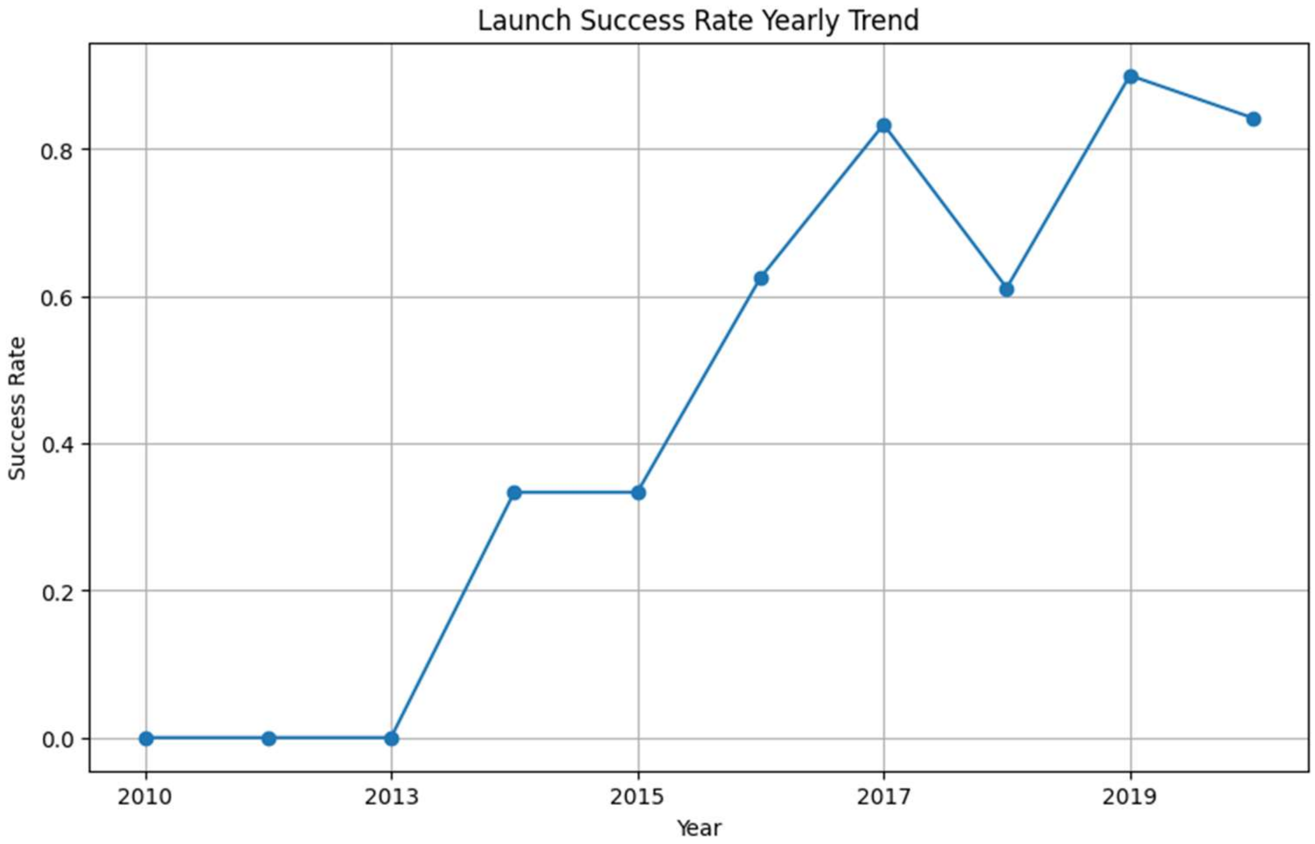
Payload vs. Orbit Type

- Heavy payloads reduce the likelihood of success in GTO orbits
- However, they have a positive effect on Polar LEO (ISS) orbits.



There is an increase of the success rate since 2013 and a rise until 2019.

Launch Success Yearly Trend



In [10]:
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

Out[10]:

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

These are the names of the unique launch site in the space mission

All Launch Site Names

Here are five records where the strings begins with CCA.

Launch Site Names Begin with 'CCA'

```
In [11]: # Use the %sql magic command to execute the query
         %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;

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

Out[11]:

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

```
In [12]:  # Execute the SQL query using %sql magic
          %sql SELECT SUM(payload_mass__kg_) FROM SPACEXTABLE WHERE customer = 'NASA (CRS)'

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

Out[12]:

SUM(payload_mass__kg_)
45596

Here is the total payload carried by boosters launched by NASA (CSR)

Average Payload Mass by F9 v1.1

```
In [14]: %sql SELECT AVG(payload_mass__kg_) FROM SPACEXTABLE WHERE booster_version = 'F9 v1.1'

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

Out[14]:

AVG(payload_mass__kg_)
2928.4

Here is the average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

Here is the date when the first successful landing outcome in ground pad took place.

```
In [15]: %sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'

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

Out[15]:

MIN(DATE)
2015-12-22

In [17]:
```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "Payload_Mass__kg_" BETI
```

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

Out[17]:

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

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

Successful Drone Ship Landing with Payload between 4000 and 6000

In [33]:
```
%sql SELECT Mission_Outcome,COUNT(*) AS numbers FROM SPACEXTABLE GROUP BY Mission_Outcome
```
```
 * sqlite:///my_data1.db
Done.
```
Out[33]:

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

These are the total number of successful and failure mission outcomes

Total Number of Successful and Failure Mission Outcomes

```
In [20]: %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLI
```

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

Out[20]:

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

Here are the names of the booster which have carried the maximum payload mass listed

Boosters Carried Maximum Payload

In [21]:
```
%sql SELECT CASE substr(Date, 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March' WHEN '04' THE
```
```
 * sqlite:///my_data1.db
Done.
```
Out[21]:

Month_Name	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

Here are the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

2015 Launch Records

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

In [22]:
```
%sql SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' (
```

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

Out[22]:

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

Here is the 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

Section 3

Launch Sites Proximities Analysis

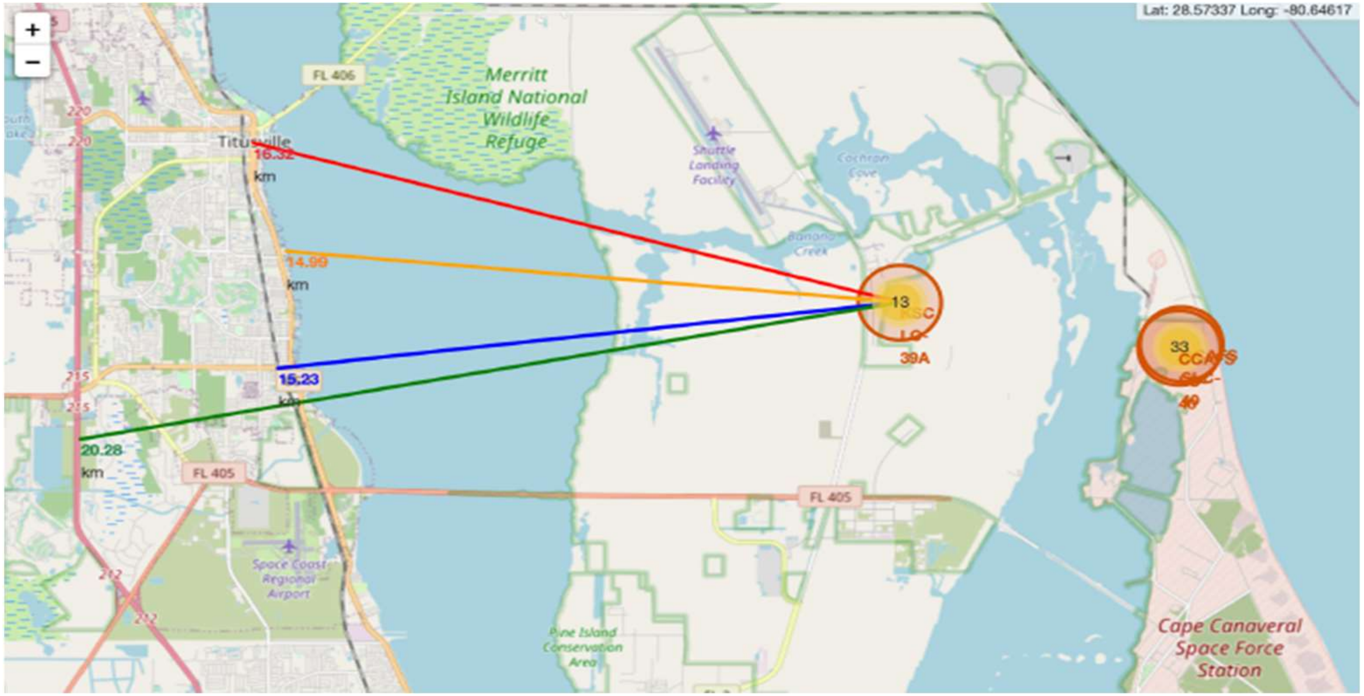
All launch site's location markers on a global map

On the map shown are launch site's location markers

Colour-labeled launch records on the map

On the map shown are successful launches with a green marker and failed launches with a red marker

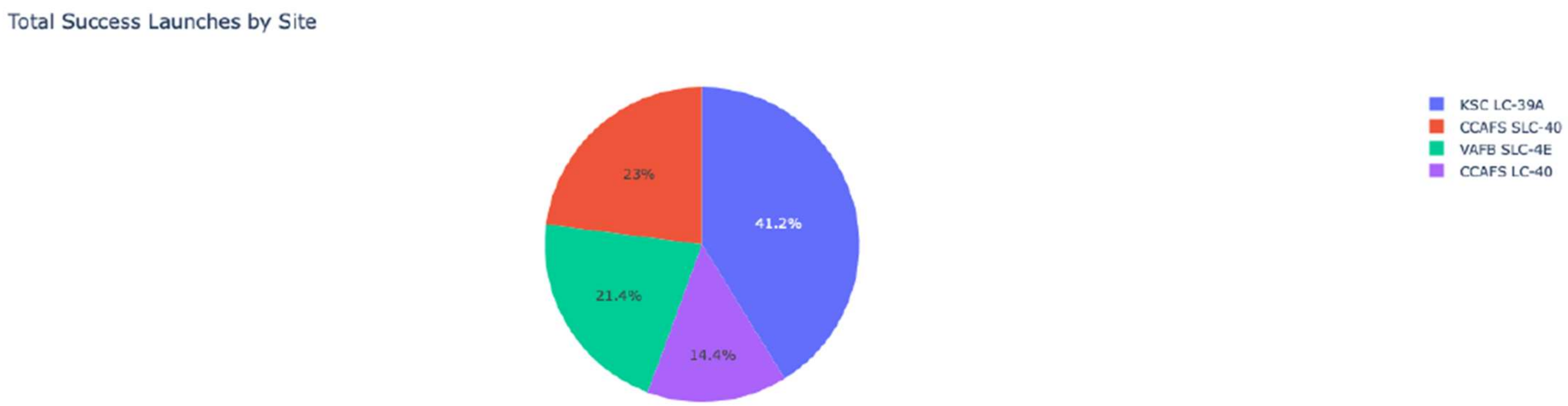
Distance from the launch site KSC LC-39A to its proximities



On the map shown is with coloured line markers distances from Launch Site KSC LC-39A to nearby features such as railways, highways, coastlines, and the closest city

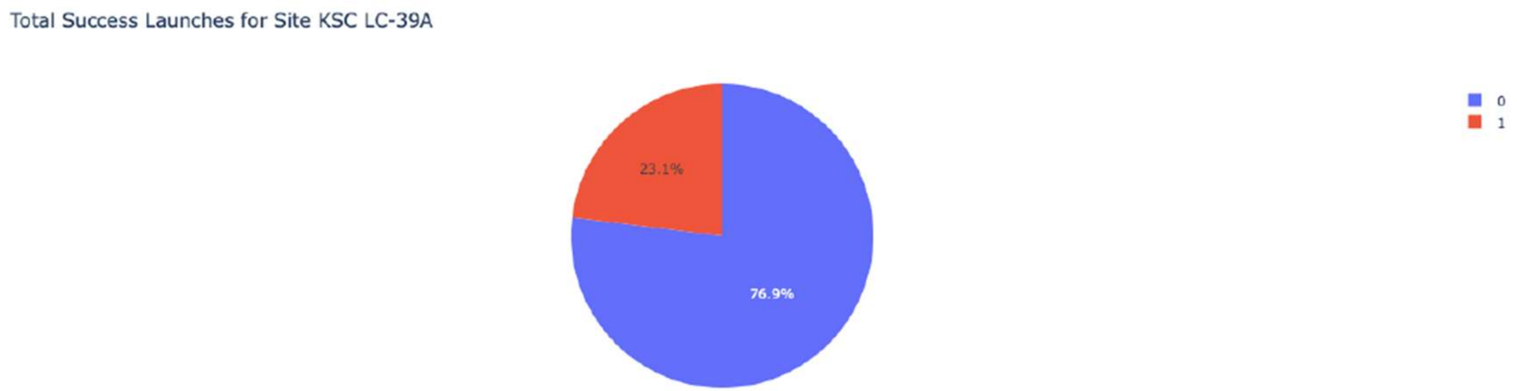
Section 4

Build a Dashboard with Plotly Dash



Shown here is a pie chart that shows that KSC LC-39A has most of the successful launches

Launch success count for all sites



Here you can see a pie chart that shows that KSC LC-39A has the highest launch success rate at 76.9%

Launch site with highest launch success ratio

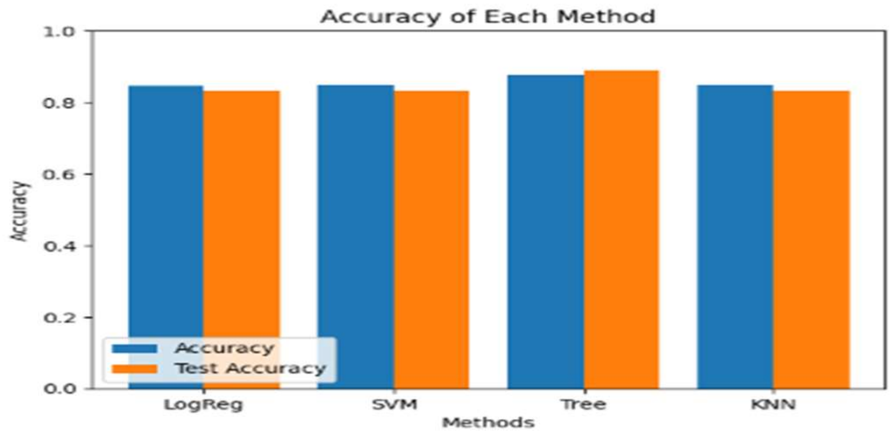
Payload Mass vs. Launch Outcome for all sites

Charts state payloads between 200 and 5500 kg as the highest success rate

Section 5

Predictive Analysis (Classification)

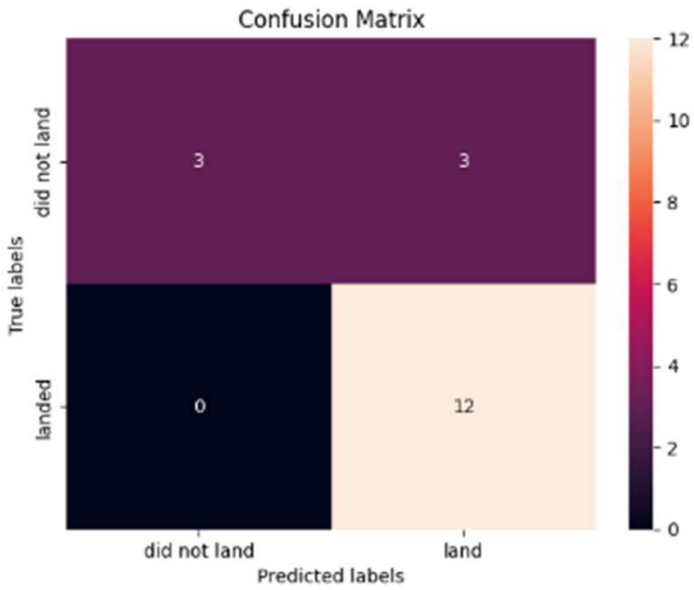
Classification Accuracy



Visualisation of the built model accuracy for all built classification models in a bar chart

Model	Accuracy
LogReg	0.84643
SVM	0.84821
Tree	0.87679
KNN	0.84821

Here we can see that the tree has the best classification accuracy



Looking at Confusion Matrix we can see that the major problem is false positives

Confusion Matrix

Conclusions

- The Decision Tree Model works best for this dataset
- Launches with lighter payloads perform better than those with heavier payloads
- Most launch sites are near the Equator and all are close to the coast
- Launch success rates have improved over time.
- KSC LC-39A has the highest success rate among all launch sites
- Orbits like ES-L1, GEO, HEO, and SSO have a 100% success rate


Thank you!