IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Paul Chambiras
11th February 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Insights from Exploratory Data Analysis
- Launch Sites Proximities Analysis
- Build a Dashboard with Plotly Dash
- Predictive Analysis (Classification)
- Conclusion

Executive Summary

Summary of Methodologies utilised:

- Data Collection using API
- Data Collection using Web Scrapping
- Data Wrangling techniques
- Exploratory Data Analysis (EDA) with SQL
- EDA and Visualization
- Interactive Visual Analytics using Folium
- Dashboard using Plotly Dash
- Machine Learning Prediction Techniques

Summary of all Results Calculated:

- EDA results and outcomes
- Interactive analytics via screenshots
- Predictive Analysis outcomes

Introduction

Project Background

The commercial space age is upon us, and corporations are making space travel somewhat affordable.

- Virgin Galactic is providing suborbital spaceflights.
- Rocket Lab is a small satellite provider.
- Blue Origin manufactures sub-orbital and orbital reusable rockets.
- The most successful is SpaceX.

SpaceX's accomplishments to-date include the following:

- Sending spacecraft to the International Space Station.
- Starlink is a satellite internet constellation providing satellite Internet access.
- Sending human-crewed missions to Space.

One reason SpaceX can do this is that rocket launches have become relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of USD 67 million; other providers cost upward of USD 170 million. The savings occur such that SpaceX can reuse the first-stage rocket. Therefore, if the first stage lands, we can determine the cost of a launch.

This project will attempt to determine if SpaceX can reuse the first-stage rocket based on public information and machine learning models.

Introduction (cont.)

Challenges that Require Solutions

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first-stage rocket landing?
- Does the rate of successful first-stage landings increase over time?
- What is the most suitable algorithm that can be used for binary classification for our project?

Section 1

Methodology

Methodology

Data Collection Methodology:

- Using SpaceX Rest API (https://api.spacexdata.com/v4/rockets/), and
- Web Scraping techniques from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches/)

Data Wrangling:

- Data was summarised and analysed using Python to deliver reliable data, filtering the data, remediating missing values, and using encoding techniques to prepare the data for binary classification

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 was normalized and ML models were trained/ tested using parameters

Data Collection

The data collection process involved a combination of data gathering techniques such as:

1. API requests from SpaceX REST API, and
2. Web Scraping data from a table represented in SpaceX's Wikipedia user entry.

Both data collection methods were used in order to obtain complete space rocket information regarding the launches for a thorough detailed analysis.

Data Columns obtained by using the SpaceX REST API technique included the following:

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

Data Columns obtained by using the Wikipedia Webscraping technique included the following:

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

Data Collection – SpaceX API

- SpaceX data was sourced by using a public API.
- The data was collected using json and pandas libraries via a Jupyter notebook.
- **GitHub source code repo**:

https://github.com/pchambiras/IBM-Applied-Data-Science-Capstone

Data Collection - Web scraping

- Data obtained through Wikipedia and analysed via html extraction using web scraping techniques and parsed into a dataframe for python command manipulation.

- **GitHub source code repo**:

https://github.com/pchambiras/IBM-Applied-Data-Science-Capstone

Data Wrangling

Exploratory Data Analysis was performed on the datasets, analysing launches per site, orbit occurrences and outcome of all missions, given this a landing outcome was created for labelling.

GitHub source code repo:

https://github.com/pchambiras/IBM-Applied-Data-Science-Capstone

EDA with Data Visualization

- Scatterplots, Histograms, and bar plots were the plotting techniques used. These options were simple yet effective to use and explore data that was required to be analysed.
- Paired variables that were analysed were as follows:

  Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

- Scatter plots show the relationship between variables, and if a relationship exists, they can be used in machine learning modelling.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series).
- **GitHub source code repo:**

  https://github.com/pchambiras/IBM-Applied-Data-Science-Capstone

EDA with SQL

- The data set was loaded into an IBM DB2 Database
- Data queries were performed in this DB using SQL Python integration
- Data queries provided a more robust and value-added understanding of the dataset within the DB.
- The queried information investigated the following:

  launch site names, overall mission outcomes, various pay load sizes for customers, rocket booster versions, and overall landing outcomes.

- **GitHub source code repo for more detailed information, code and results**:

  https://github.com/pchambiras/IBM-Applied-Data-Science-Capstone

Build an Interactive Map with Folium

- The interactive Folium maps assisted with marking the launch sites, where the successful and unsuccessful landings occurred, and a proximity visualisation technique to view key map location data: i.e., Railway, Highway, Coast, and City.
- Folium maps facilitated to understand why certain launch sites were chosen and why.
- They also provided visualising of successful landings that were relative to the location.
- Visual markers easily indicated launch sites, with circles that highlight launch areas, and clustering that was applied to event groups per coordinates.

- **GitHub source code repo for more detailed information, code and results**:

  https://github.com/pchambiras/IBM-Applied-Data-Science-Capstone

Build a Dashboard with Plotly Dash

- Graphs and plots based on number of launches per site and payload range were obtained to analyse and identify the relation of payloads and launch sites.
- This information provided a useful indication as to the ideal launch sites per rockets dependent on their payload size.

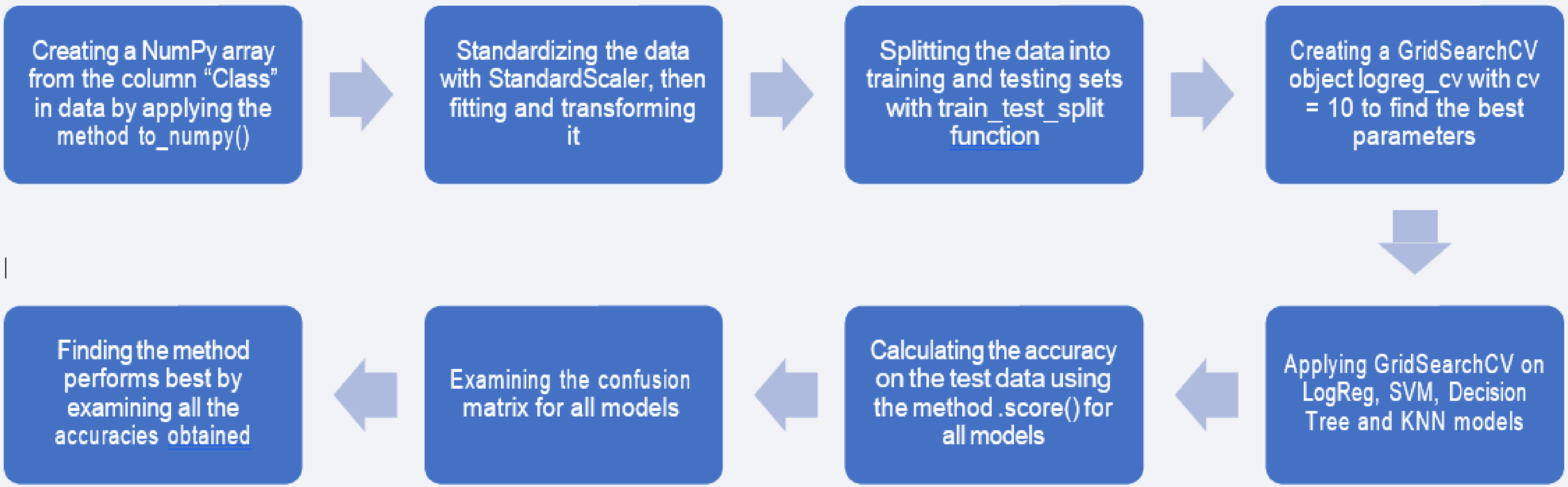
- **GitHub source code repo for more detailed information, code and results**:

  https://github.com/pchambiras/IBM-Applied-Data-Science-Capstone

Predictive Analysis (Classification)

- Four classification models were used for predictive analysis:
  1. Logistic Regression,
  2. Support Vector Machine,
  3. Decision Tree and
  4. K Nearest Neighbours, using the next strategy

Predictive Analysis (Classification) cont.



- **GitHub source code repo for more detailed information, code and results**:

  https://github.com/pchambiras/IBM-Applied-Data-Science-Capstone

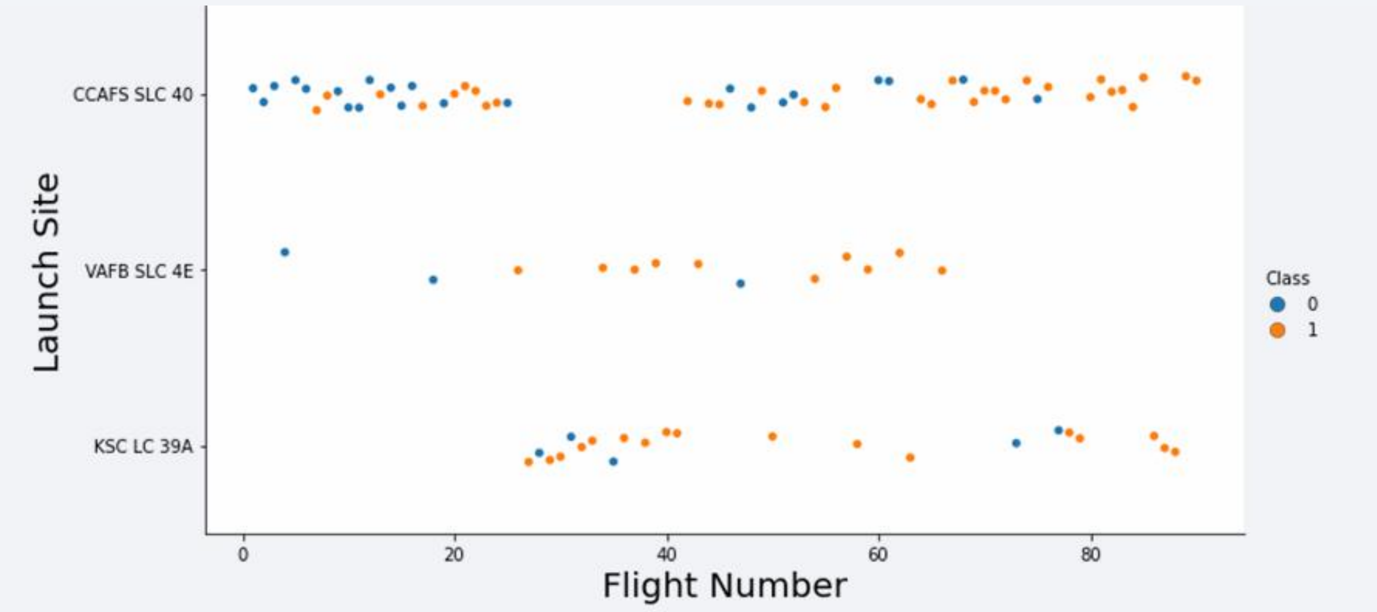
Results

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

Section 2

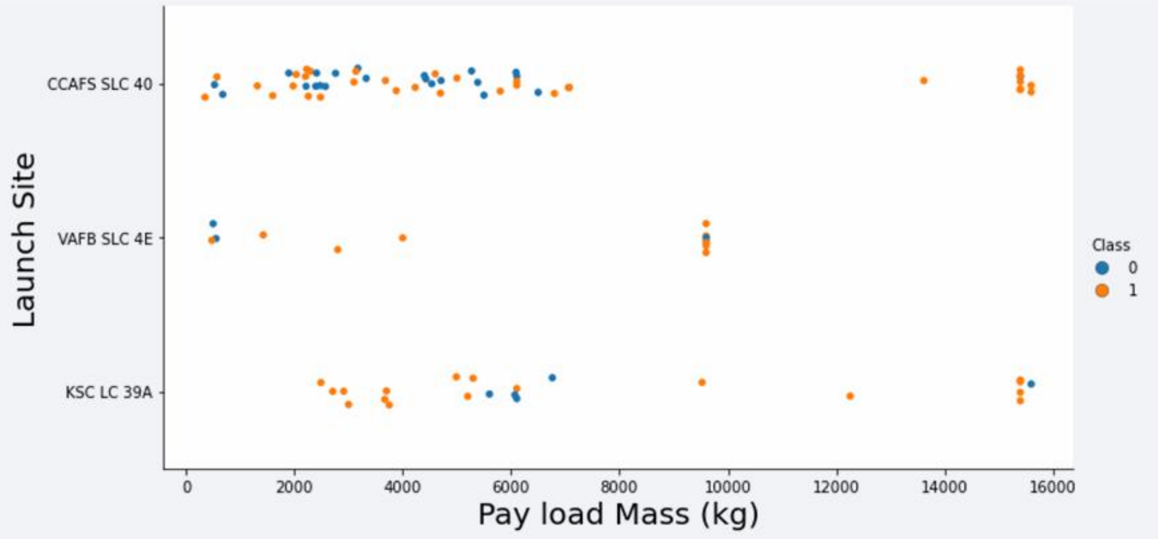
Insights drawn from EDA

Flight Number vs. Launch Site



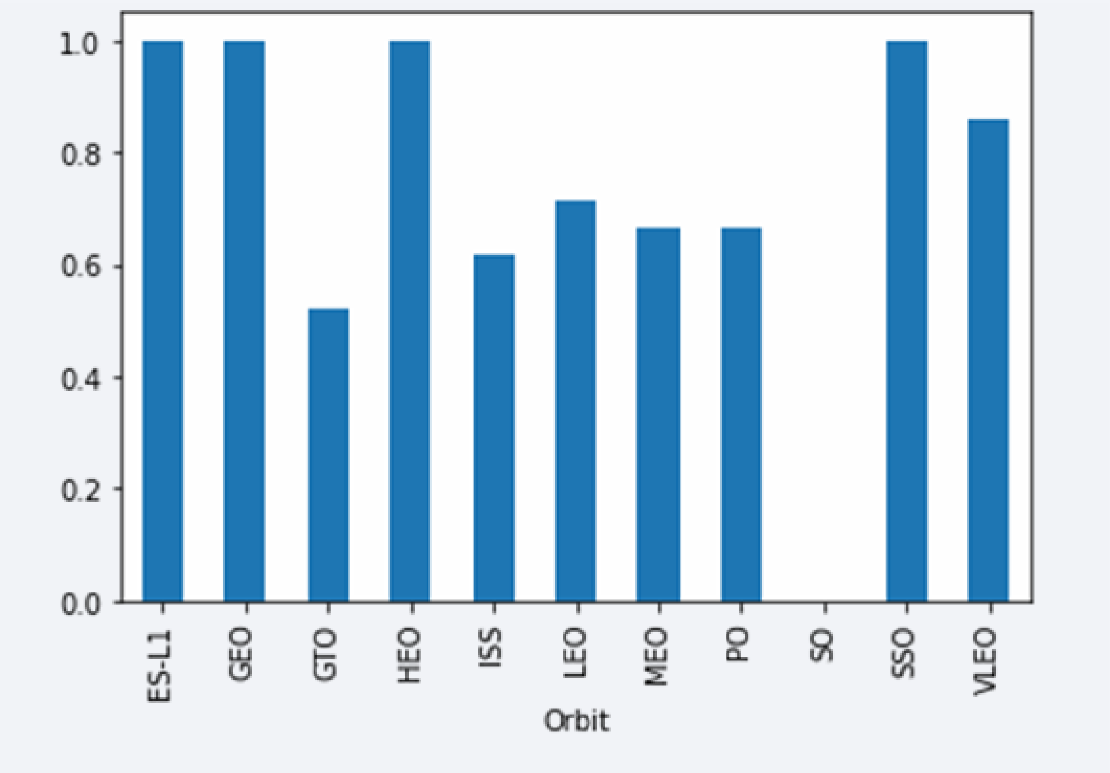
1. All the earliest flights failed while the latest flights were successful
2. CCAFS SLC 40 launch site comprised almost half of all launches in total
3. Both VAFB SLC 4E and KSC LC 39A had higher success rates than CCAFS SLC 40

Payload vs. Launch Site



1. Launch site - VAFB-SLC 4E launch site - no rockets launched from this site with a payload mass >10 000 kg (or 10 tonnes)
2. Majority of launches - with a payload mass over 7000 kg (or 7 tonnes) - were successful
3. Launch site - KSC LC 39A - has a 100% success launch rate for payload mass under 5500 kg (or 5.5 tonnes)

Success Rate vs. Orbit Type



Orbits with 100% success rate:

- ES-L1, GEO, HEO and SSO

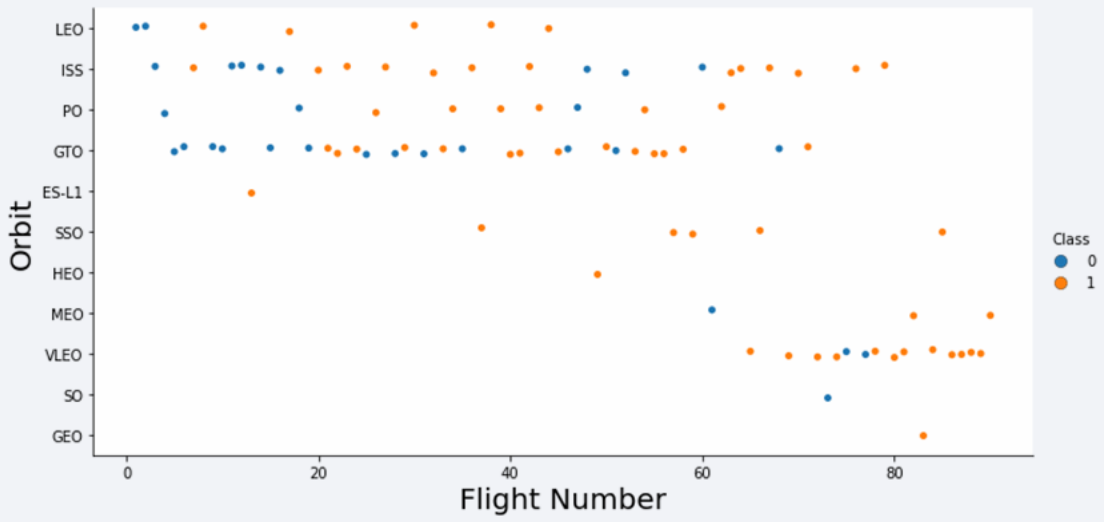
Orbits with 0% success rate:

- SO

Orbits with success rate between 50% and 85%:

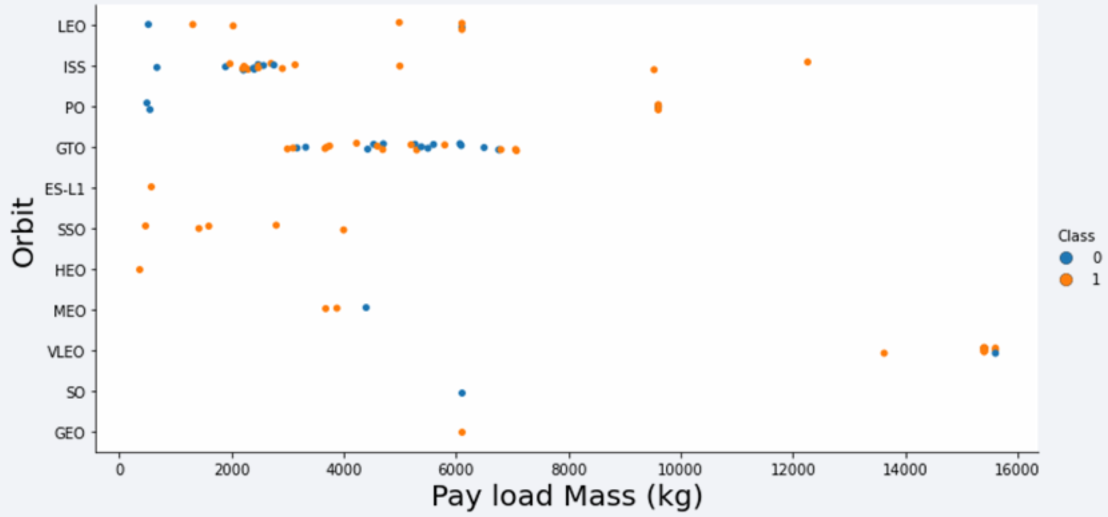
- GTO, ISS, LEO, MEO and PO

Flight Number vs. Orbit Type



1. For the LEO orbit, its success appears to be related to the number of flights,
2. Alternatively, there appears to be no relationship between flight number when in GTO orbit.

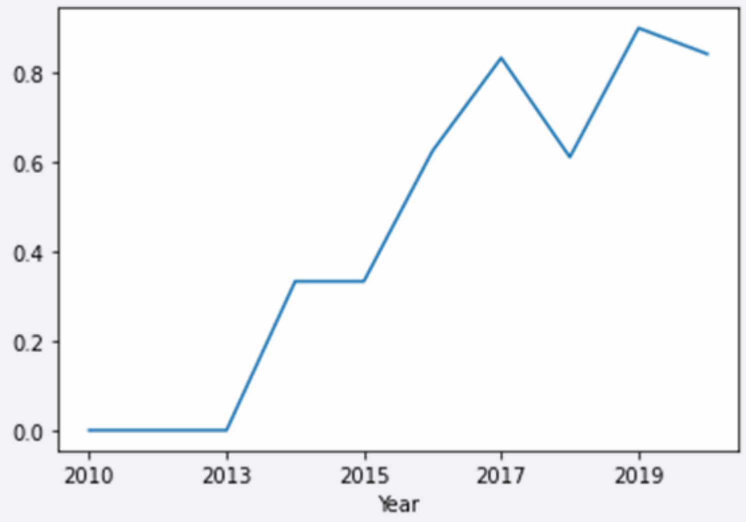
Payload vs. Orbit Type



Regarding heavy payloads, the successful or positive landing rate was higher for PO, LEO and ISS.

For GTO, there are both positive and negative landing rates (or an unsuccessful mission) due to the data present.

Launch Success Yearly Trend



1. Since 2013, the launch success rate was positive and was increasing until 2017
2. However between 2017 and 2018, the success rate dropped approximately 20%.

All Launch Site Names

```
%sql select distinct(launch_site) from SPACEXDATASET
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.data
bases.appdomain.cloud:30119/bludb
Done.
```

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

The SQL statement discovered 4 different launch sites

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

Displayed 5 records where the launch site names begin with 'CCA'.

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
45596

1. Displayed the total payload mass carried by boosters launched by NASA (CRS).
2. The result was: 45,596 kg.

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXDATASET where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
2928

The average payload mass carried by booster version F9 v1.1 was: 2,928 kg

First Successful Ground Landing Date

```
%sql select min(DATE) from SPACEXDATASET where landing__outcome = 'Success (ground pad)'

 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
2015-12-22

The date when the first successful landing outcome occurred in ground pad was achieved on: December 22, 2015

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_>'4000' and payload_mass__kg_<'600
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

The names of the boosters which have success in drone ship and have payload mass greater than 4,000 but less than 6,000 are:

F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(mission_outcome) from SPACEXDATASET where mission_outcome = 'Success' or mission_outcome like 'Failure%'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
100

Answer is 100, both successful and failed mission outcomes

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
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

Using a subquery, there were 12 different booster versions which have carried the maximum payload mass

2015 Launch Records

```
%sql select month(date) as month, landing__outcome, booster_version, launch_site from SPACEXDATASET \
where landing__outcome = 'Failure (drone ship)' and year(date) = '2015'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

The results showed one failed landing outcome in January, another in April, and on the same launch site - CCAFS LC-40

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

```
%sql select landing__outcome, count(landing__outcome) as "Total Count" from SPACEXDATASET \
where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by count(landing__outcome) desc
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

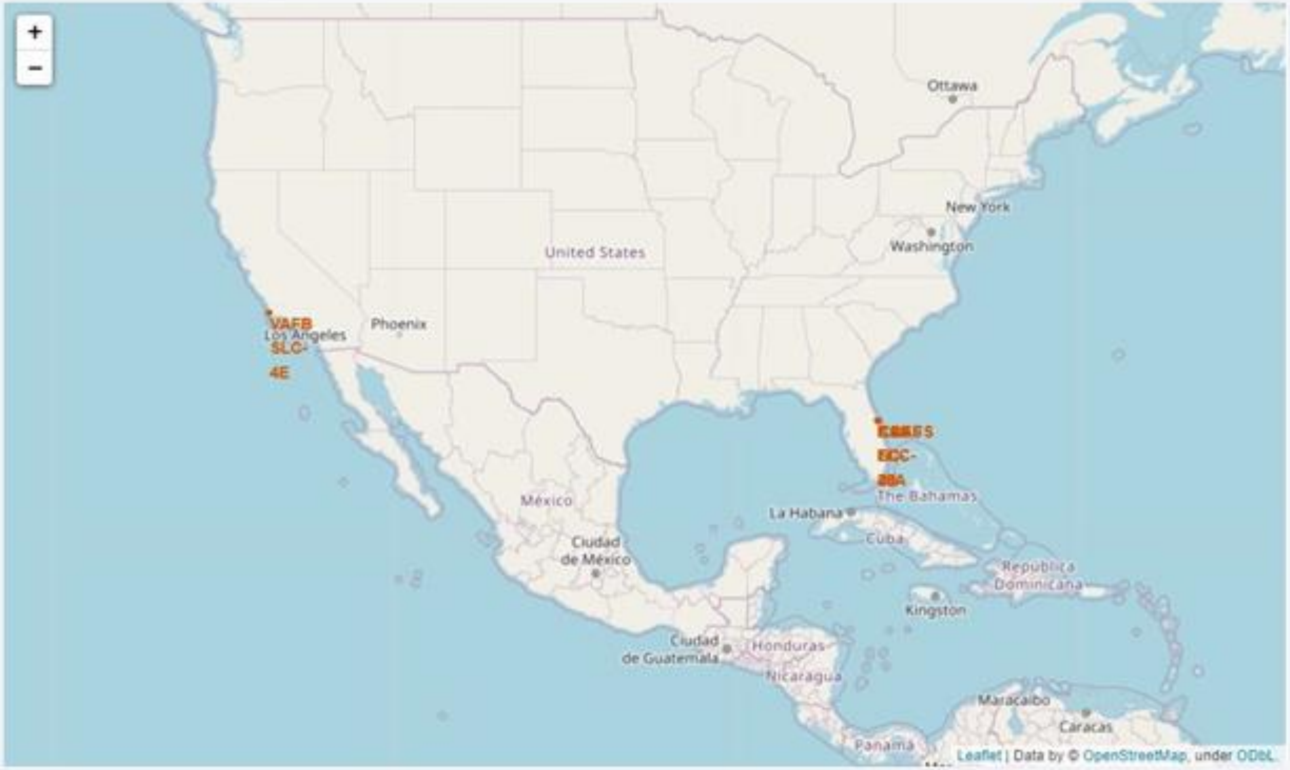
landing__outcome	Total Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Between 2010-06-04 and 2017-03-20, there were 31 launches

Section 3

Launch Sites Proximities Analysis

Launch site's locations markers on a global map



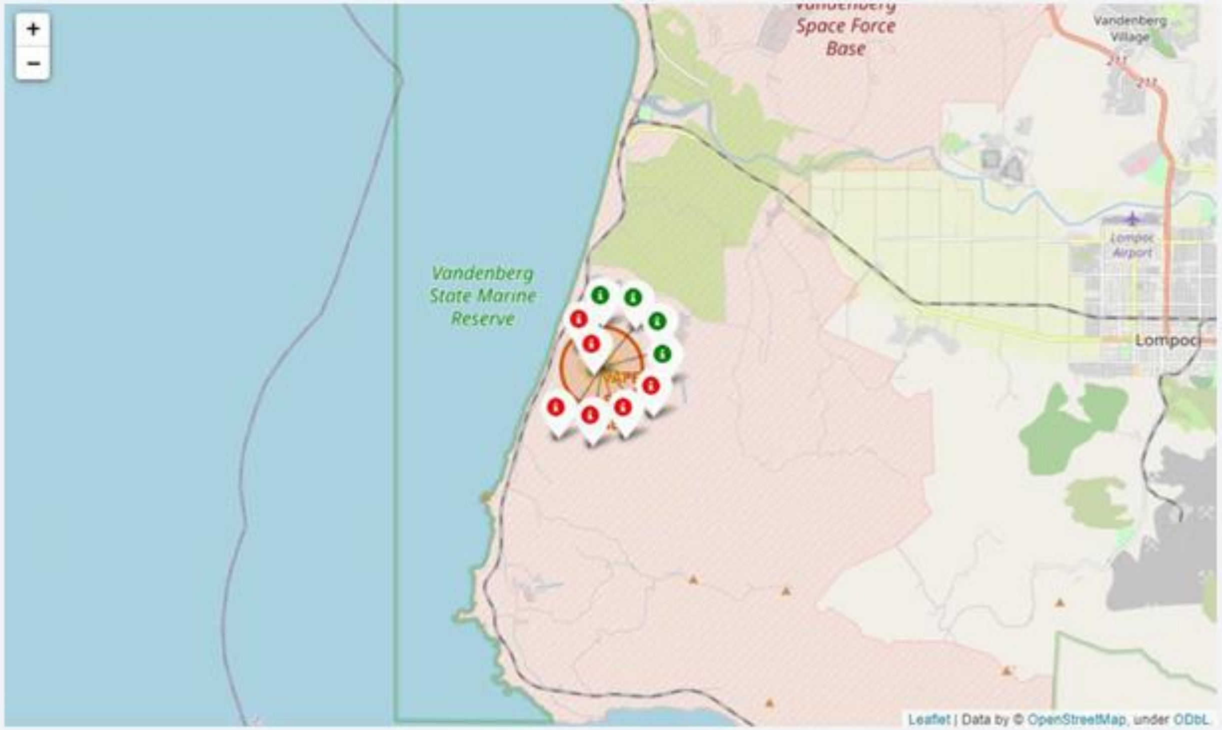
- **Are all launch sites within proximity to the Earth’s Equator line?**

Most of them are, because anything on the surface of the Earth at the equator is already moving at 1670 kilometres per hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is due to inertia. This speed will help the spacecraft maintain enough speed to stay in orbit.

- **Are all launch sites in close proximity to the coast?**

Most of them are, because launching a rocket from the east coast gives an additional boost to the rocket, due to the rotational speed of the Earth. Also, these rockets travel eastward, so if anything goes wrong during their ascent, the debris would essentially fall into an ocean's waters, far away from populated areas.

Success and failed launches for each site on the map

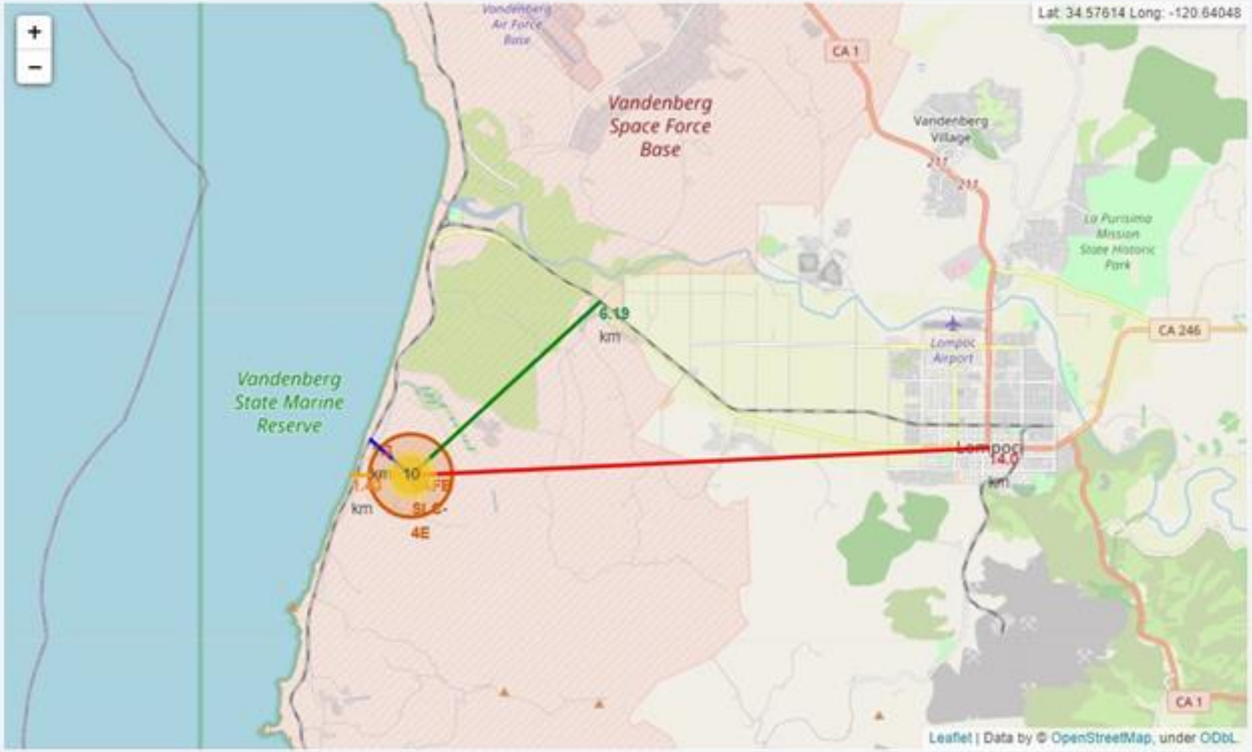


For a successful launch (class = 1), a **green** marker was used, and if the launch failed, it was coloured in **red**.

Regarding launch site - VAFB SLC-4E – there were 10 launches in total:

- **6** were unsuccessful
- **4** were successful

Distance from VAFB SLC-4E launch site to its proximities



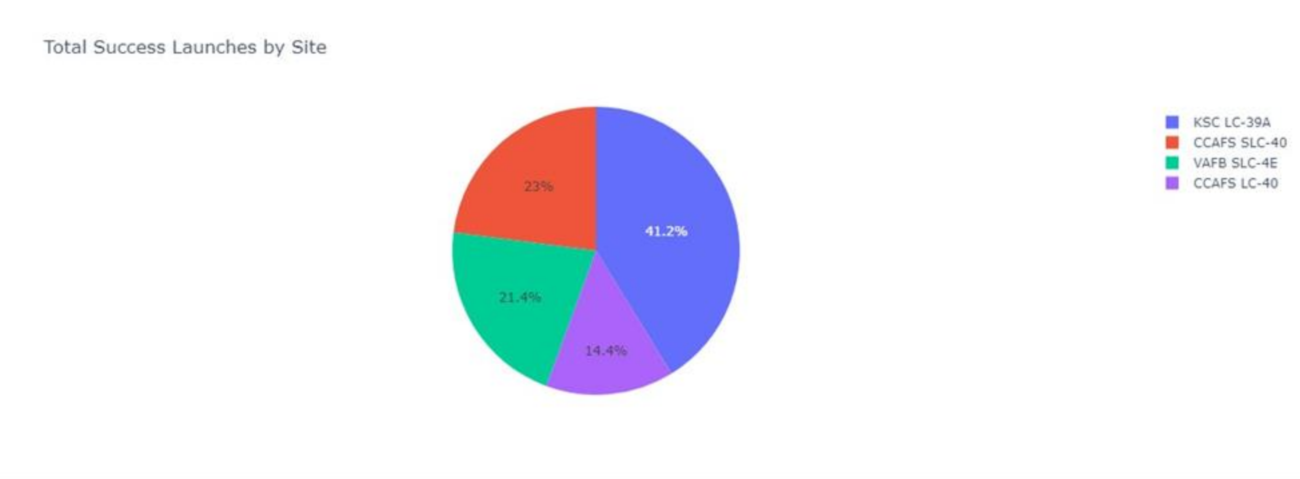
For launch site: **VAFB SLC-4E**

1. It is relatively close to a railway line - 1.3 km
2. It is close to a highway - 6.19 km
3. It is relatively close to the coastline - 1.43 km
4. It is quite close to its nearest city of Lompoc – 14 km.

Section 4

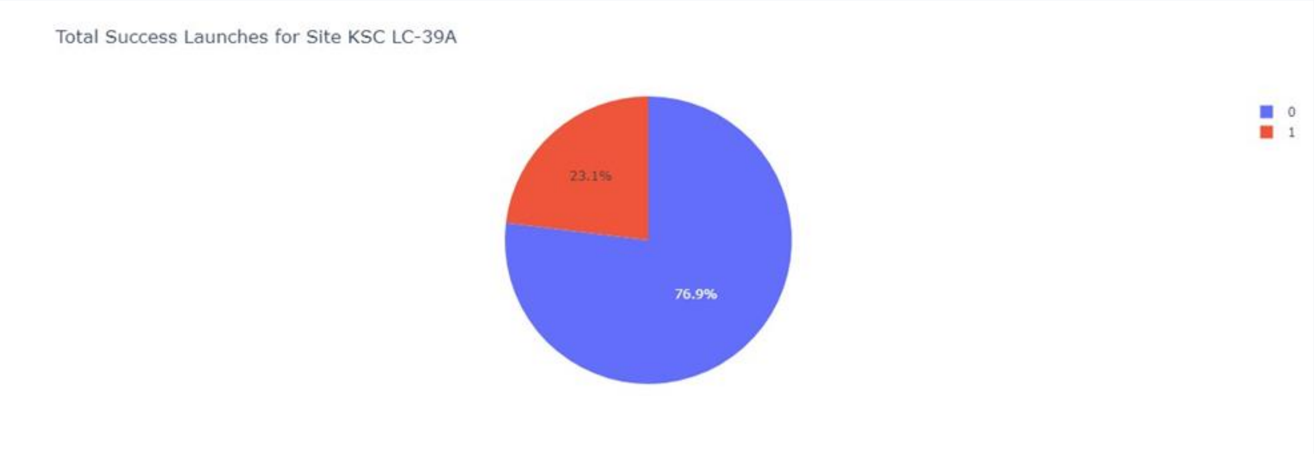
Build a Dashboard with Plotly Dash

Total Success Launches by Site



1. KSC LC-39A is the launch site with the highest success with launches
2. CCAFS LC-40 is the launch site with the lowest success with launches

Launch Site with highest launch success ratio



KSC LC-39A has the highest launch success rate with 76.9%

1. 10 were successful,
2. 3 were unsuccessful

Payload Mass vs Launch Outcome for all sites

1. We can see from the charts that payloads between 2000 and 5500 kg have the highest success rate.

2. Booster version FT has the highest success rate between the payloads range (between 2000 and 5500 kg).

Section 5

Predictive Analysis (Classification)

Classification Accuracy

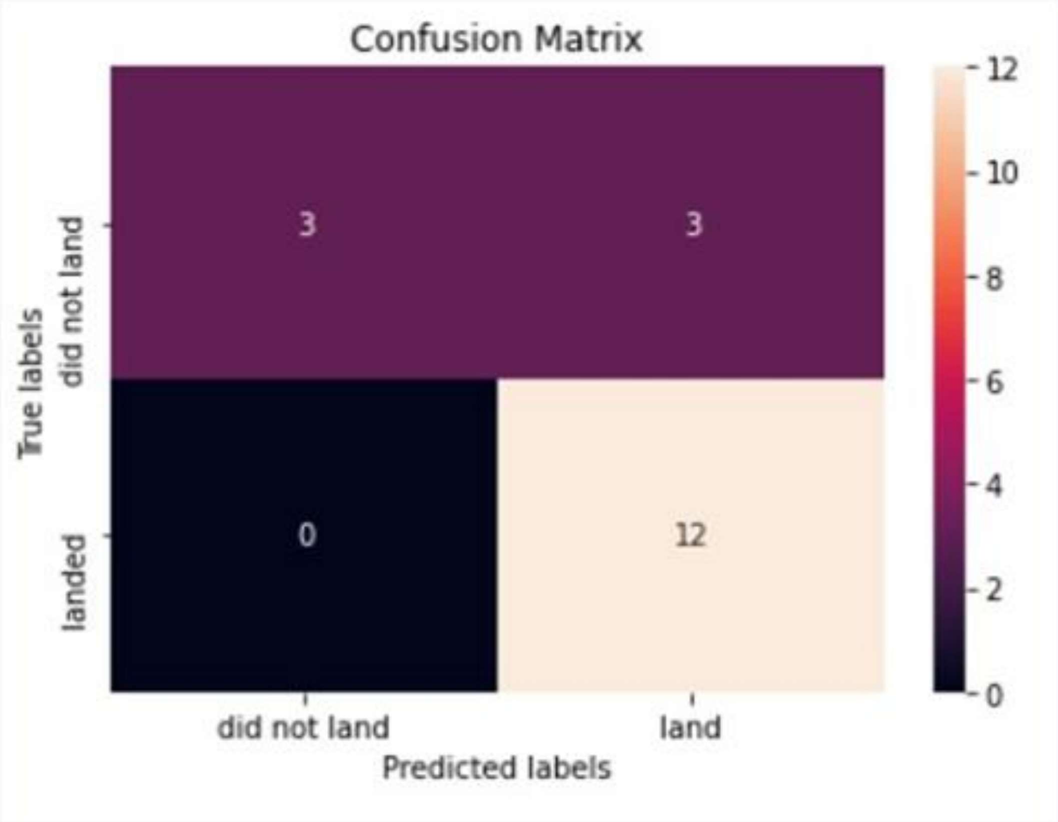
```
#The results are practically the same.

print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.8333333333333334
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

- Based on the scores for the test set, we can not confirm which method has performed the best because the results are almost identical.
- This is because the dataset is small and has lesser values (18 samples in all).

Confusion Matrix



Since the results are almost identical, the confusion matrix is similar viewing for all methods tested.

Conclusions

1. Orbits ES-L1, GEO, HEO and SSO all experienced a success rate of 100%
2. Launches with a lower payload mass had better outcomes than launches with a larger payload mass
3. The success rate of launches increased over time
4. Majority of the launch sites were in proximity to the Equator line, and all the launch sites are in close proximity to the coastline
5. KSC LC-39A has the highest launch success rate of all the sites
6. The accuracy results are almost identical, therefore we can not conclude which ML modelling method (i.e. - Logistics Regression, Support Vector Machine, Decision tree or KNN) performed the best

Thank you!