IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yousef Essam Aziz
18-09-2021

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

- **Project background and context**

- SpaceX is an American aerospace manufacturer, space transportation services and communications company headquartered in Hawthorne, California. SpaceX was founded in 2002 by Elon Musk with the goal of reducing space transportation costs to enable the colonization of Mars. SpaceX manufactures the Falcon 9 and Falcon Heavy launch vehicles, several rocket engines, Dragon cargo, crew spacecraft and Starlink communications satellites.

- The company 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 SpaceX can reuse the irst stage. Therefore, if we can determine if the irst stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- **Problems you want to find answers**

  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - Does the rate of successful landings increase over the years?
  - What is the best algorithm that can be used for binary __classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a 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
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

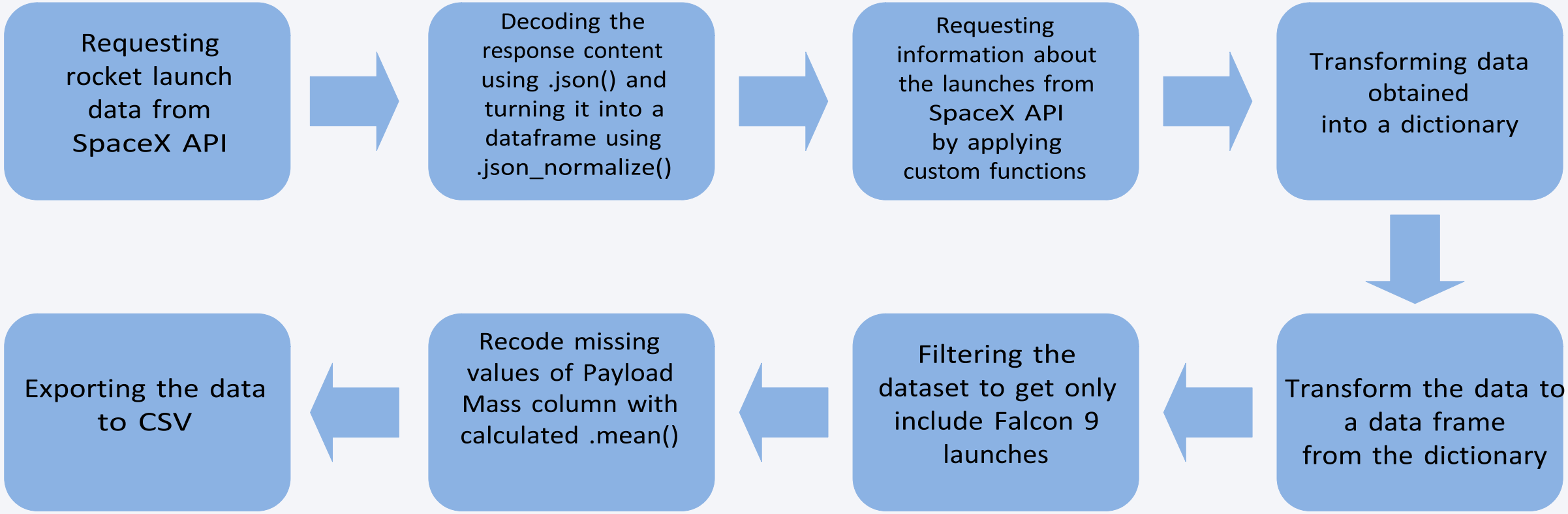
Data Columns are obtained by using SpaceX REST API:

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

Data Columns are obtained by using Wikipedia Web Scraping:

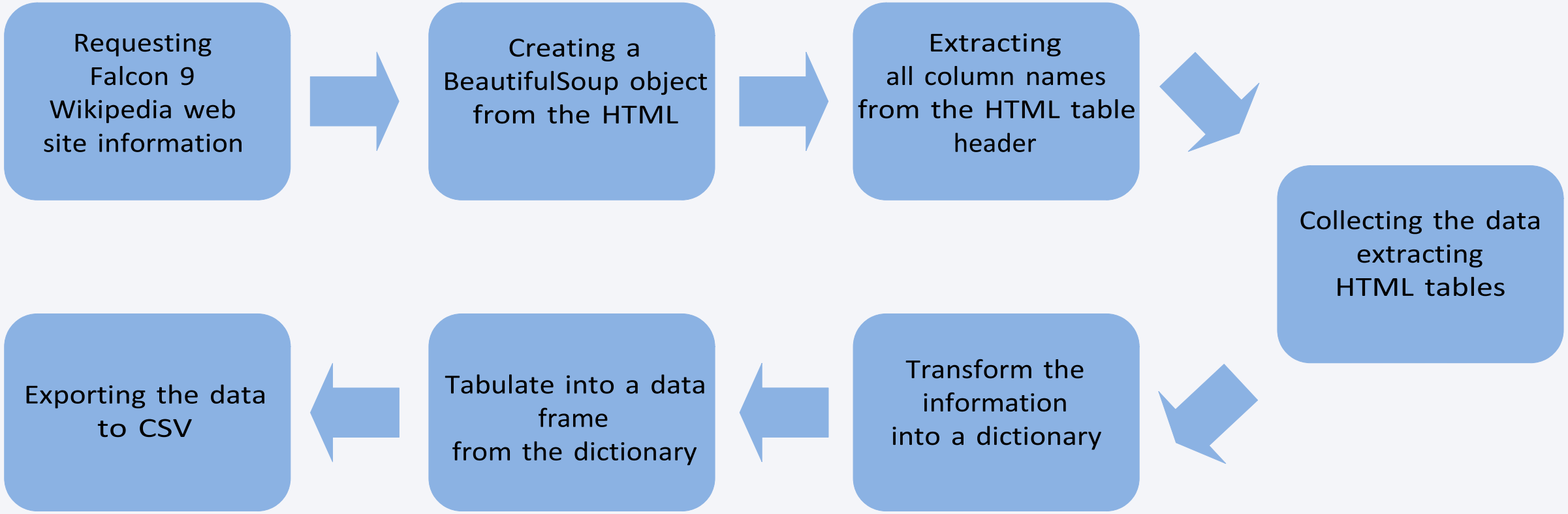
- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Github URL: Data Collection AP

Data Collection - Scraping



Github URL: Data Collection with Web Scraping

Data Wrangling

Do an exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

Github URL: Data Wrangling

EDA with Data Visualization

Charts 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 and
- Success Rate Yearly Trend

Scatter plots show the relationship between both variables. If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The objective 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 URL: EDA with Data Visualization

EDA with SQL

SQL queries in Exploratory Data Analysis:

- Query of names of the unique launch sites in the space mission
- Query of 5 records where launch sites begin with the string ‘CCA'
- Query of total payload mass carried by boosters launched by NASA (CRS)
- Query of average payload mass carried by booster version F9 v1.1
- List of the date when the first successful landing outcome in ground pad was achieved
- List of the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List of the total number of successful and failure mission outcomes
- List of the names of the booster versions which have carried the maximum payload mass
- List of the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Finally, Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the date 2010-0604 and 2017-03-20 in descending order

Github URL: EDA with SQL

Build an Interactive Map with Folium module

To Mark of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates.
- Repeat the action showing their geographical locations and proximity to Equator and coasts.

Select a colour Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

Github URL: Interactive Visualization with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

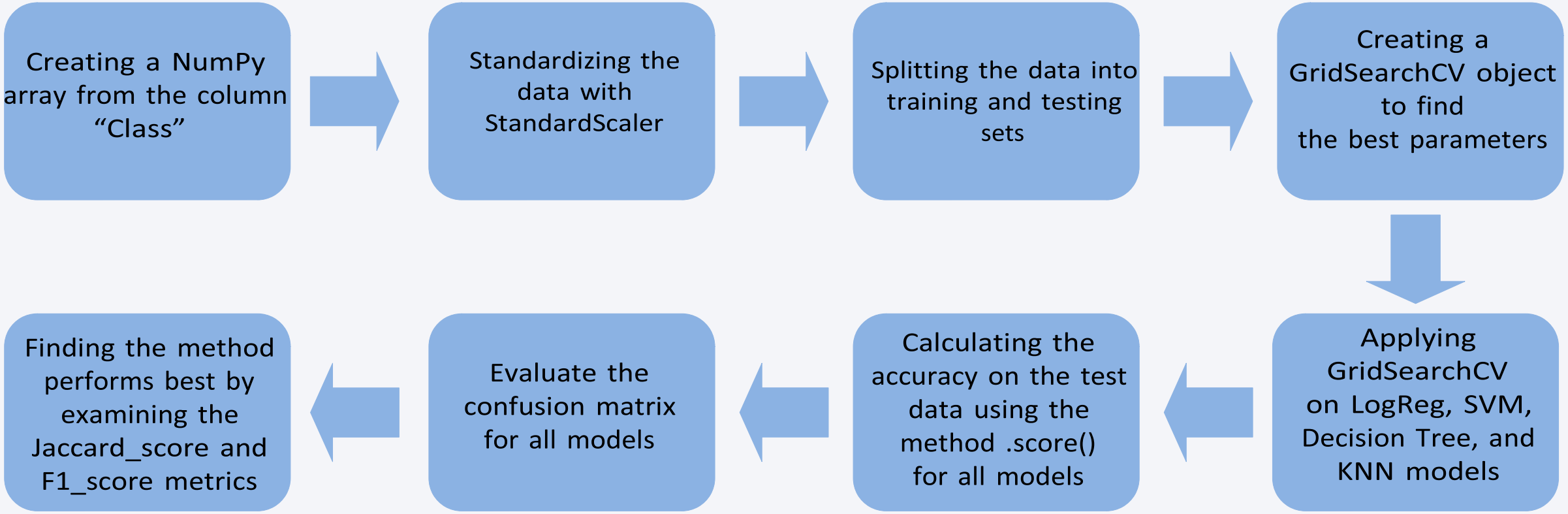
- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

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

Github URL: SpaceX Dash app (code)

Predictive Analysis (Classification)



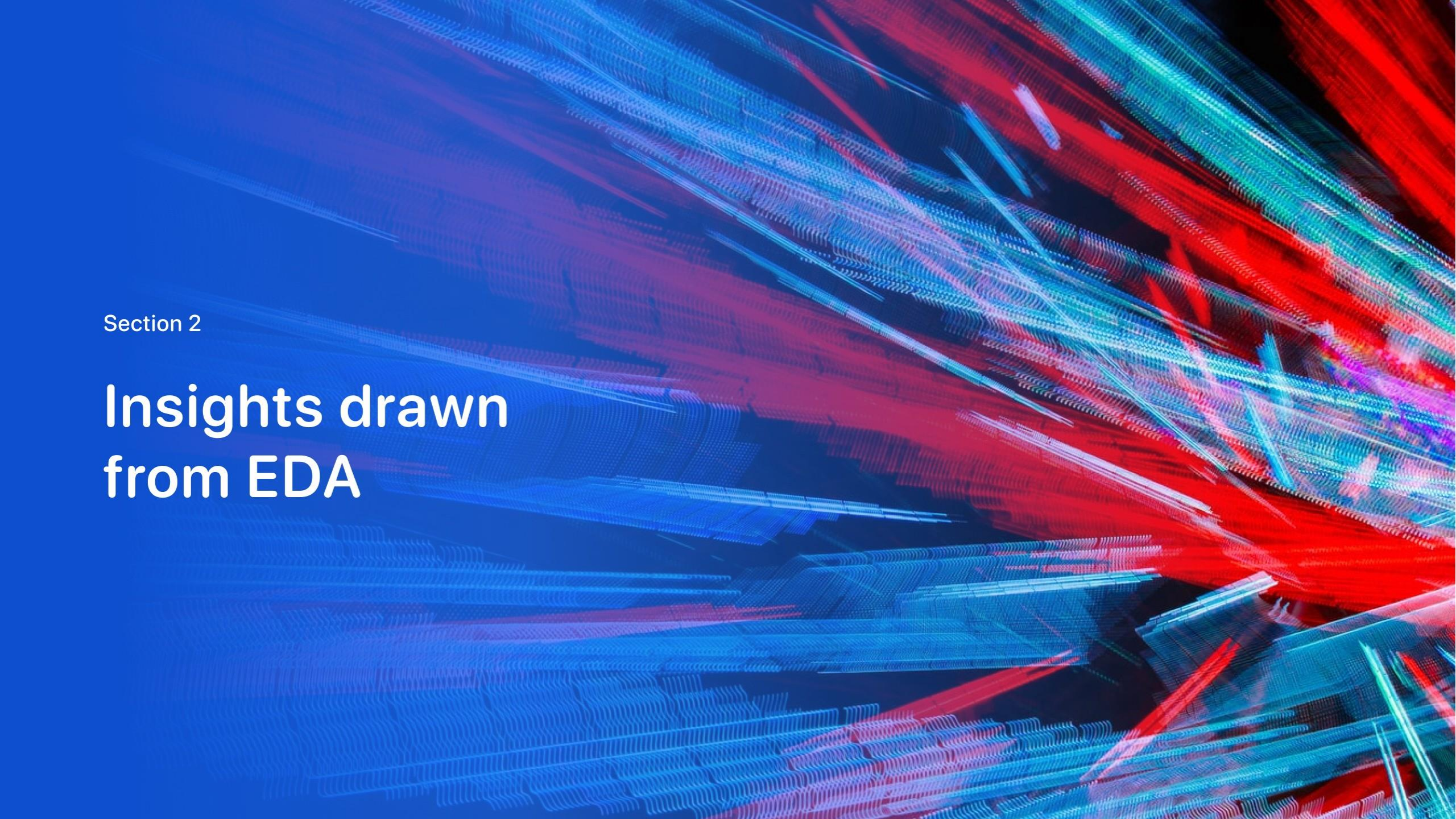
Github URL: Machine Learning Prediction

Results

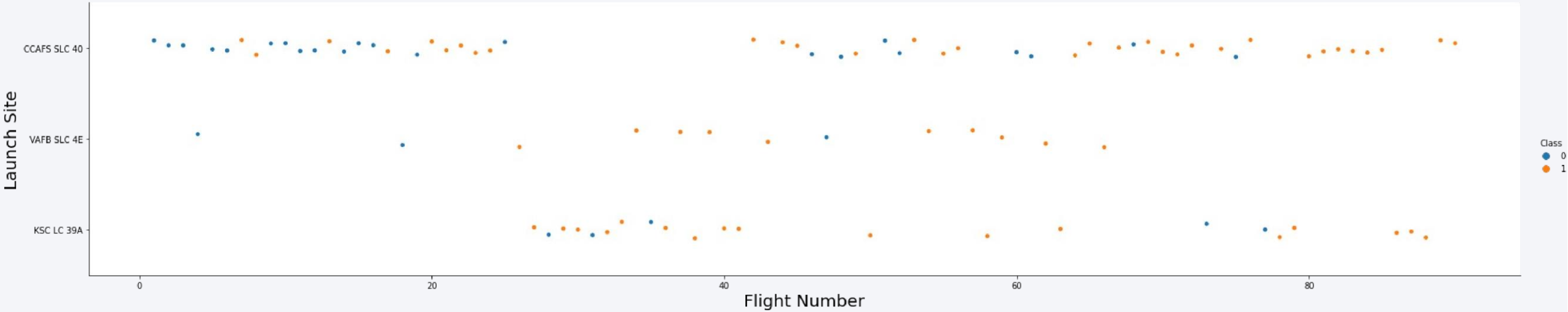
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



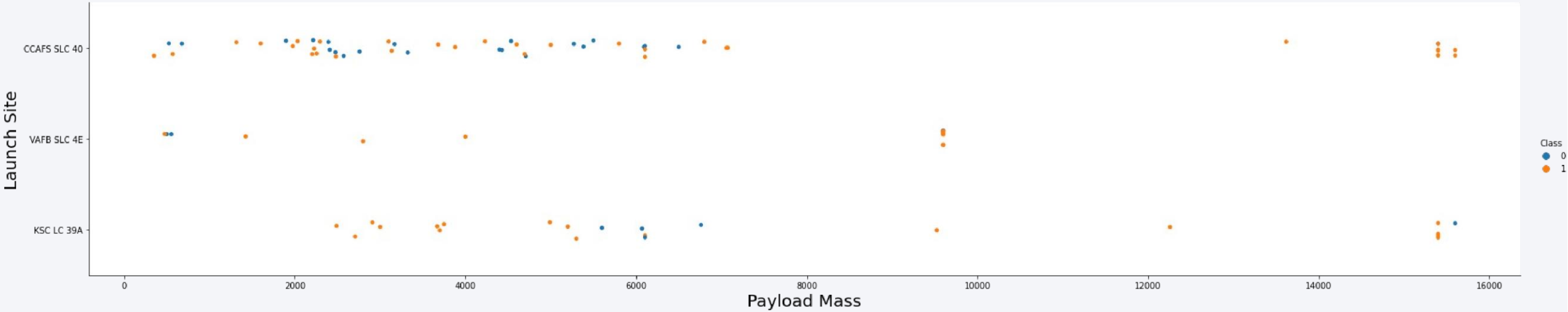
Flight Number vs. Launch Site



Explanation:

- The earliest lights all failed while the latest lights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

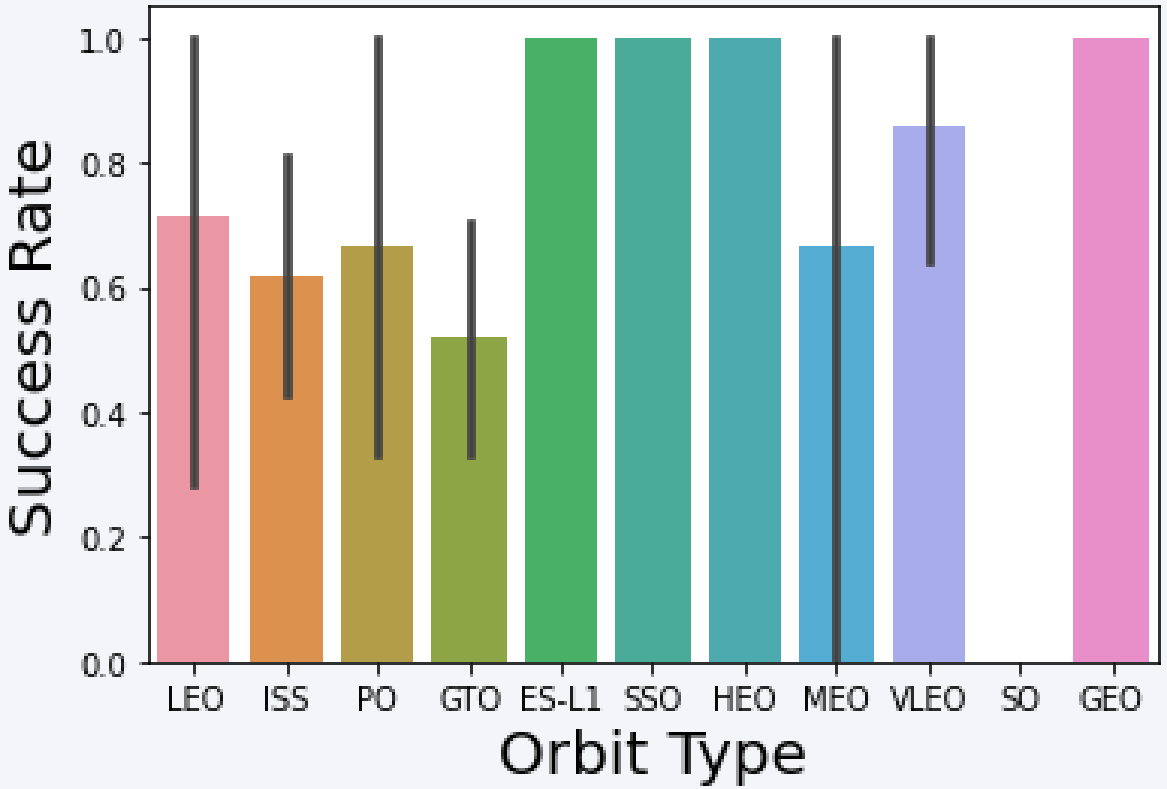
Payload vs. Launch Site



Explanation:

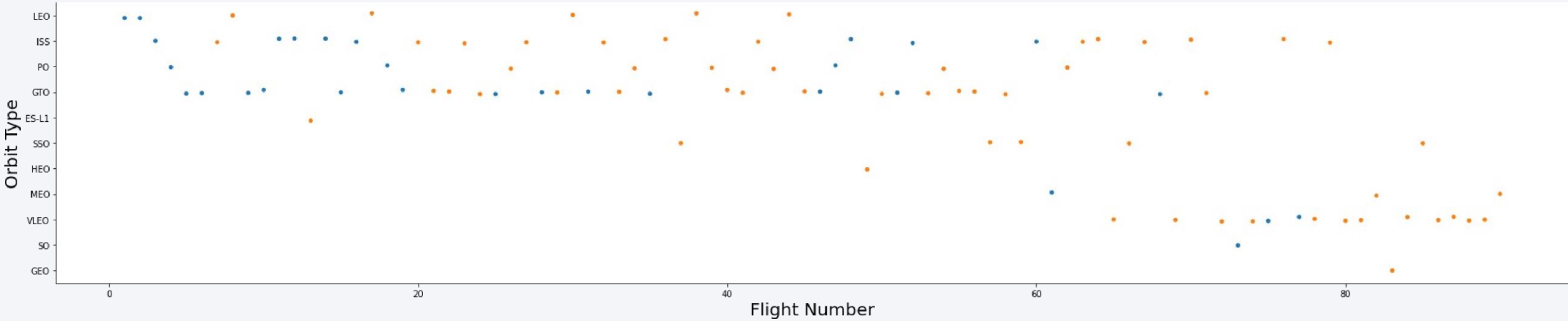
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

Success Rate vs. Orbit Type



Explanation:

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO



Explanation:

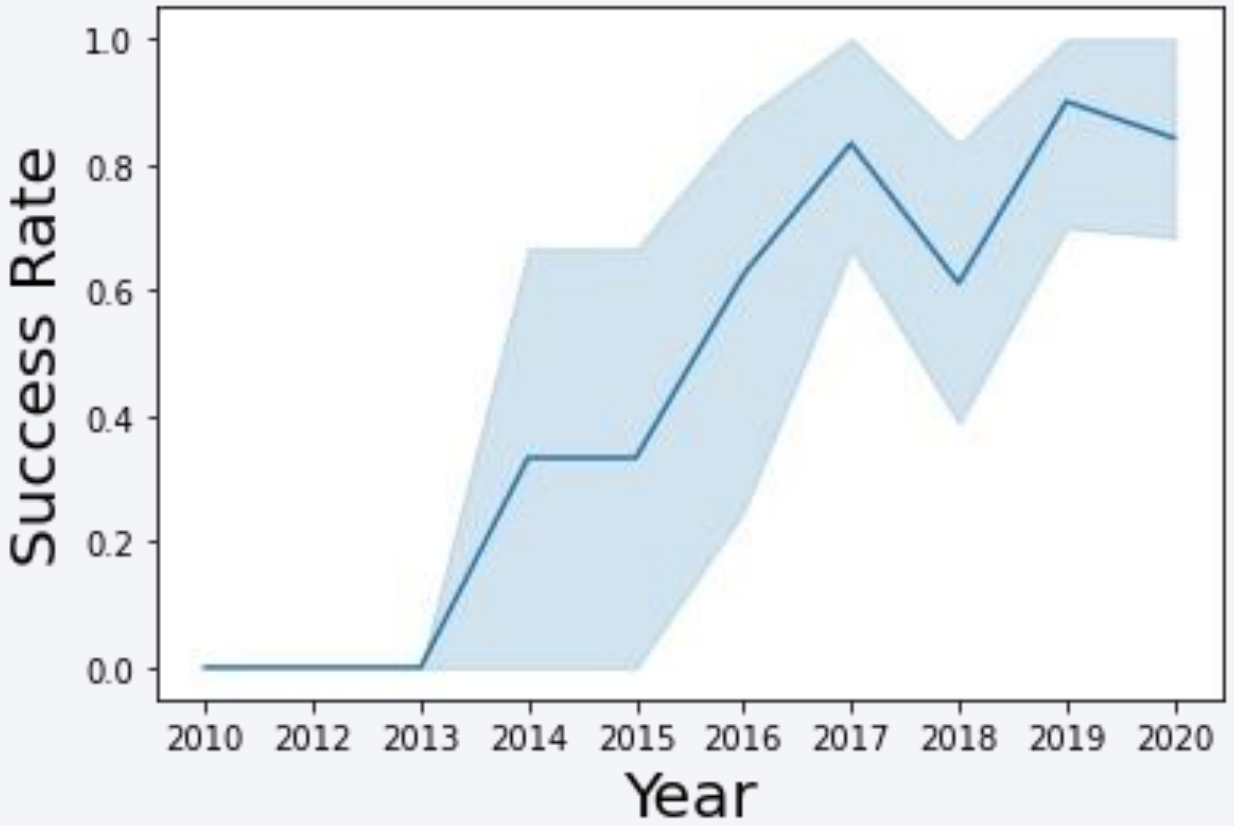
- In the LEO orbit the Success is related to the number of lights; on the other hand, there seems to be no relationship between light number when in GTO orbit.

Payload vs. Orbit Type

Explanation:

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



Explanation:

- The success rate since 2013 kept increasing till 2020.

All Launch Site Names

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

```
In [3]:  %%sql
         SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[3]:

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

Explanation:

- Showing the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

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

In [4]:
```
%%sql
SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

Out[4]:

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

Explanation:

- Showing 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

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

In [5]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass by NASA (CRS)" FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

Out[5]:

Total Payload Mass by NASA (CRS)
45596

Explanation:

- Showing the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

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

```
In [6]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload Mass by Booster Version F9 v1.1" FROM SPACEXTBL WHERE BOOS
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

Out[6]:

Average Payload Mass by Booster Version F9 v1.1
2928

Explanation:

- Displaying average payload mass carried by booster version F9 v1.1.

First Successful Ground Landing Date

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

Hint:Use min function

In [7]:
```
%%sql
SELECT MIN(DATE) AS "First Succesful Landing Outcome in Ground Pad" FROM SPACEXTBL WHERE Landing_Outcome = 'Succe
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

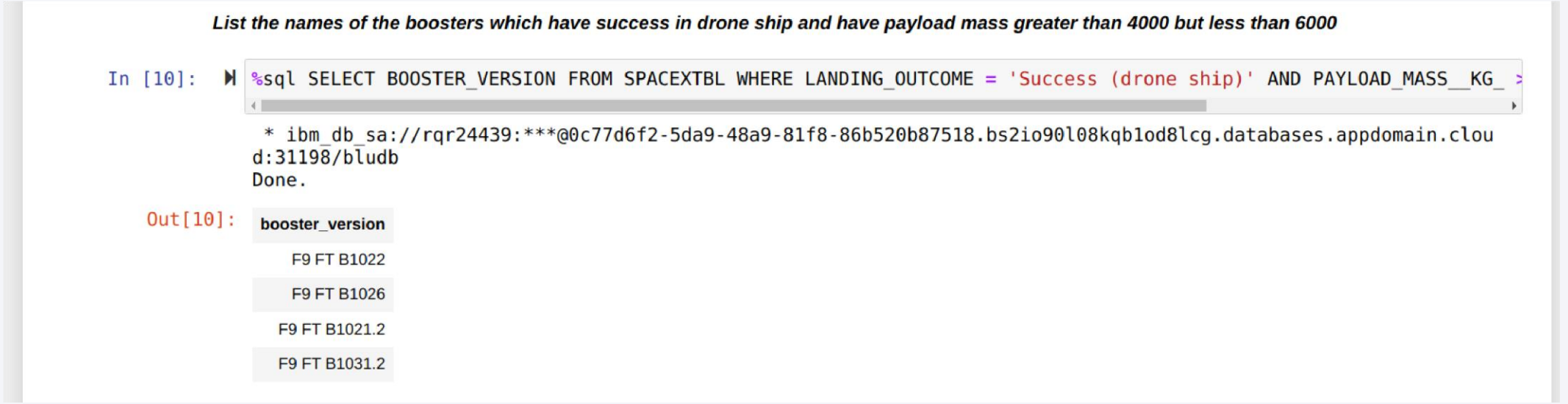
Out[7]:

First Succesful Landing Outcome in Ground Pad
2015-12-22

Explanation:

- Getting the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000



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

In [10]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ >
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

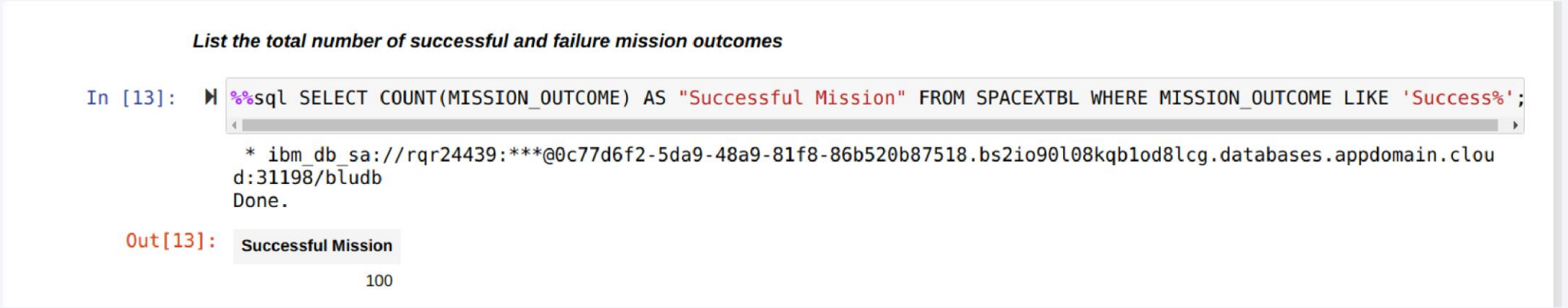
Out[10]:

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

Explanation:

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

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

In [13]:
```
%%sql SELECT COUNT(MISSION_OUTCOME) AS "Successful Mission" FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%';
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

Out[13]:

Successful Mission
100

Explanation:

- Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

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

In [14]:
```
%%sql
SELECT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEXTBL WHERE PAYLOAD_
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

Out[14]:

Booster Versions which 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

Explanation:

- Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

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

In [20]:
```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND Landing_Outcome = 'Failure (drone
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

Out[20]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Explanation:

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

In [27]:
```
%sql SELECT Landing_Outcome AS "Landing Outcome", COUNT(Landing_Outcome) AS "Success Count" FROM SPACEXTBL WHERE
```

```
 * ibm_db_sa://rqr24439:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.clou
d:31198/bludb
Done.
```

Out[27]:

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

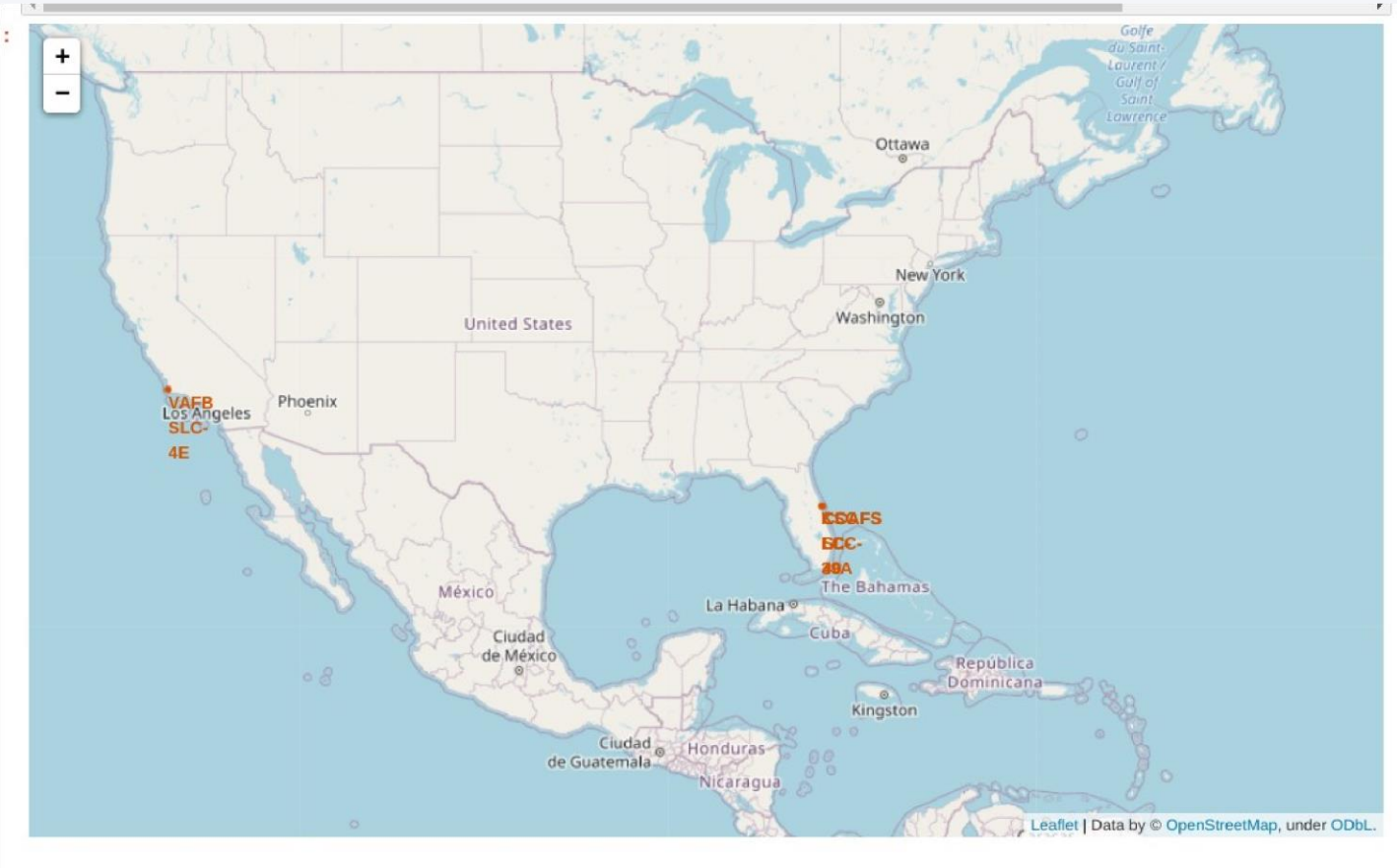
Explanation:

- Ranking 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 4

Launch Sites Proximities Analysis

All launch sites location markers on a global map

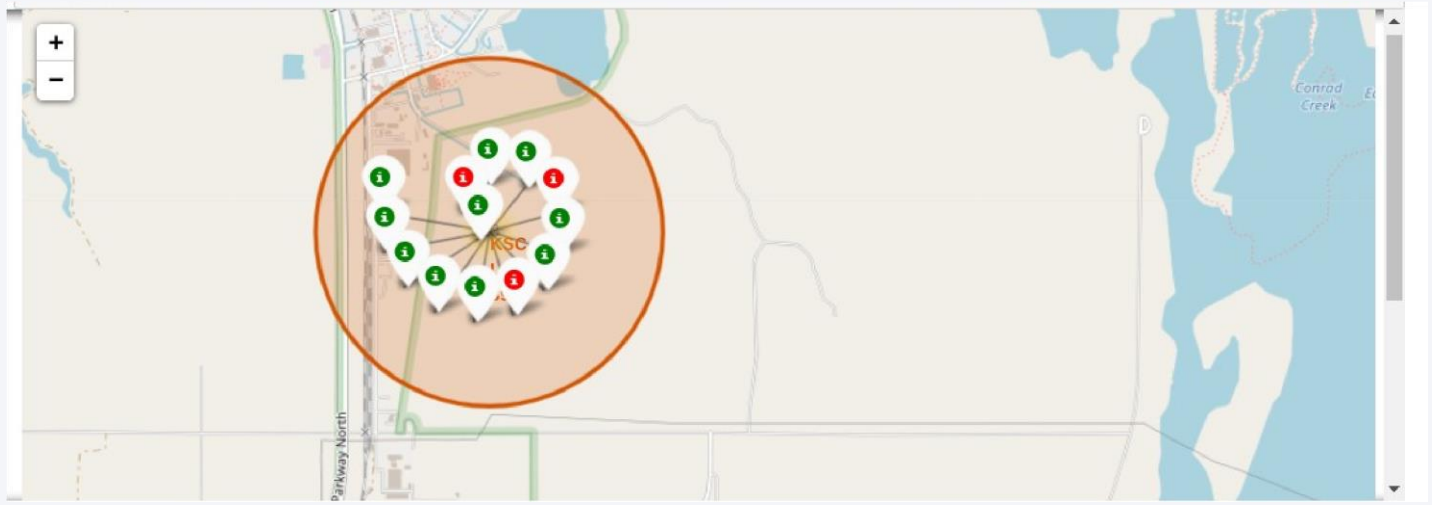


Explanation:

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. So, 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, because of the inertia. This speed will help the rocket keep up a good enough speed to stay in orbit.

- All launch sites are beside the coast, while launching rockets towards the ocean it helps to reduce the risk of having any problems or in the worst scenario exploding near people.

Colour-labeled launch records on the map



Explanation:

- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch

- Launch Site KSC LC-39A has a very high Success Rate.

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

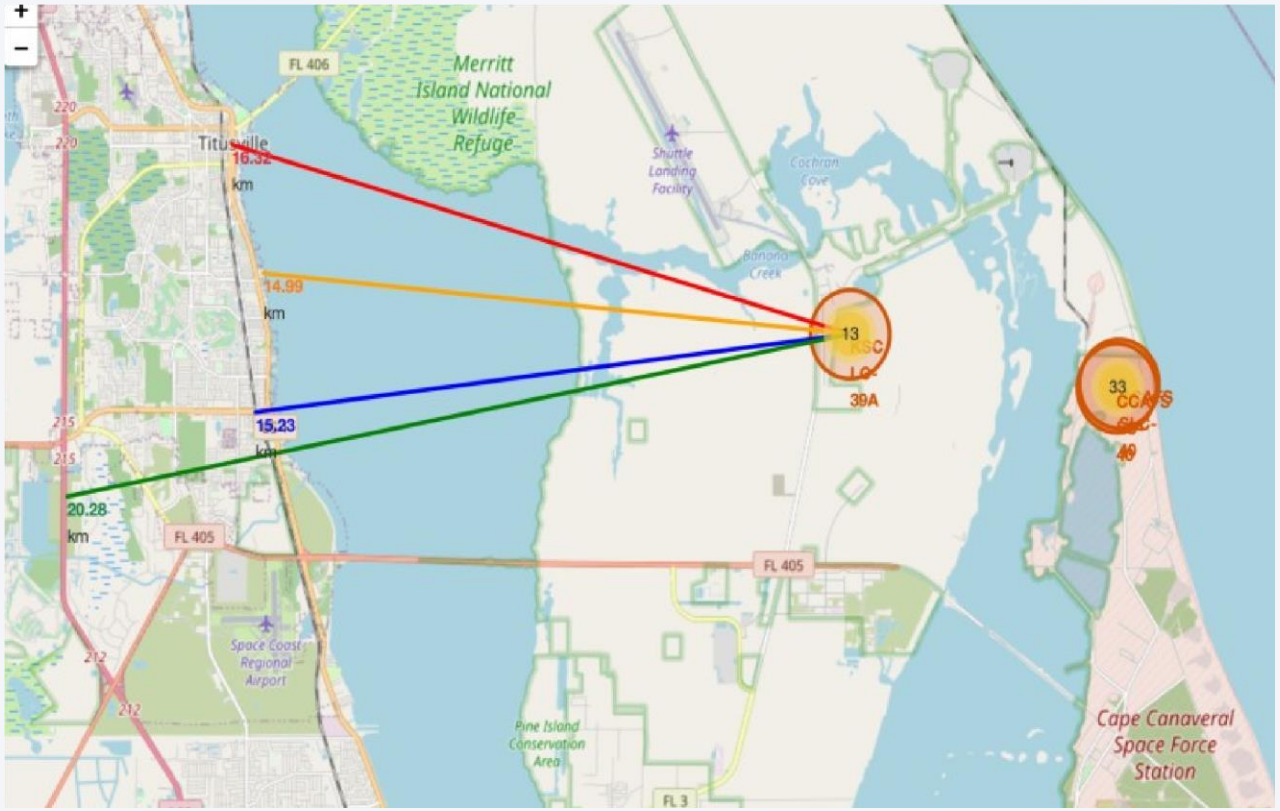
Explanation:

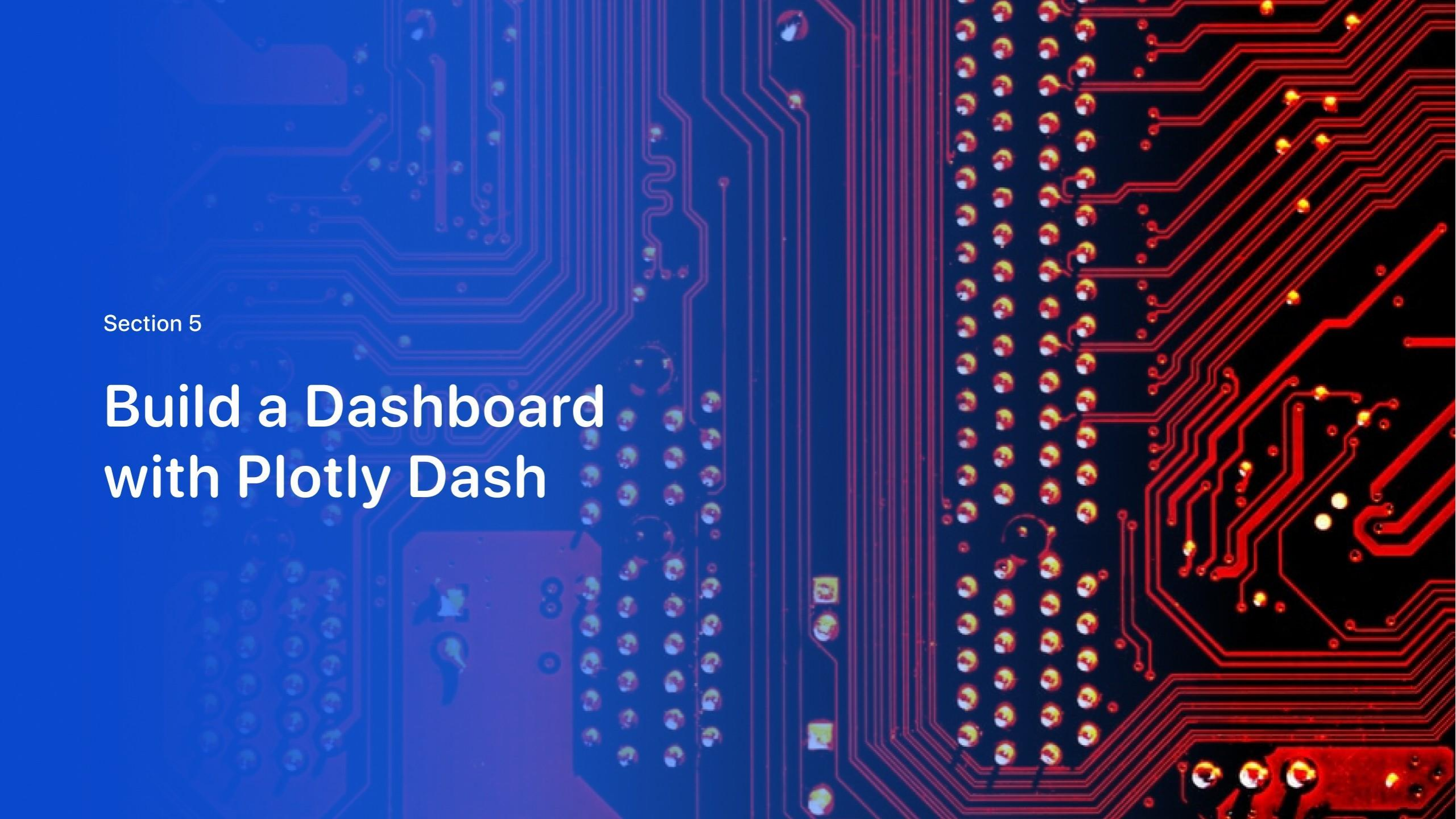
- From the visual analysis of the launch site KSC LC-39A we can see that it is:

  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)

- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to where people live.

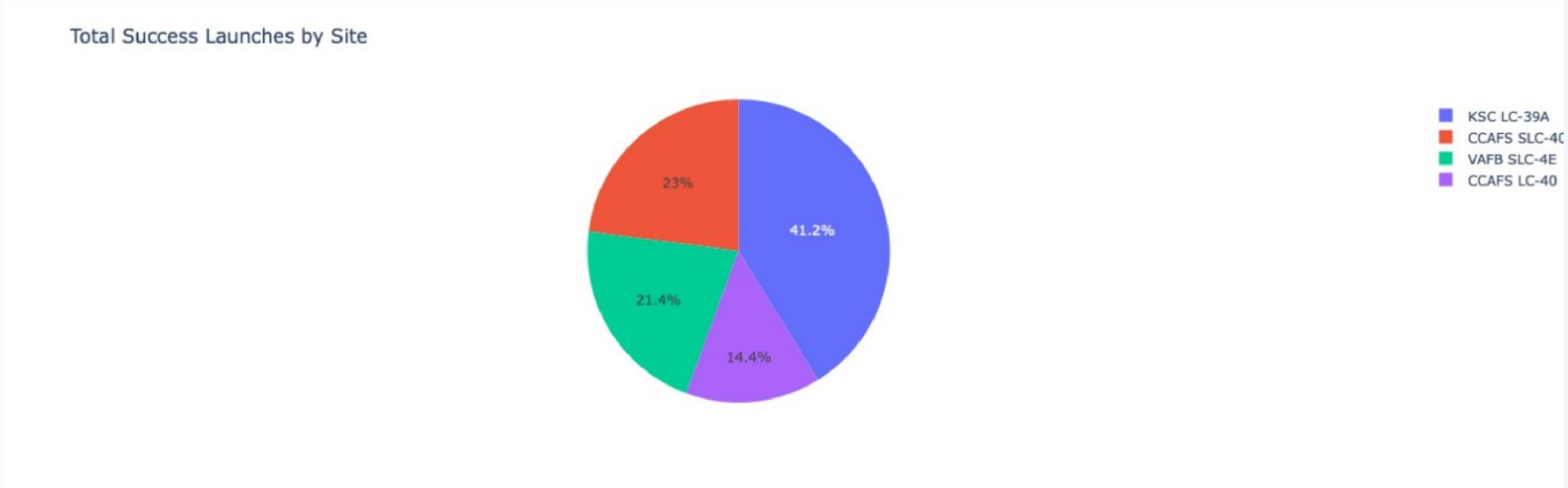




Section 5

Build a Dashboard with Plotly Dash

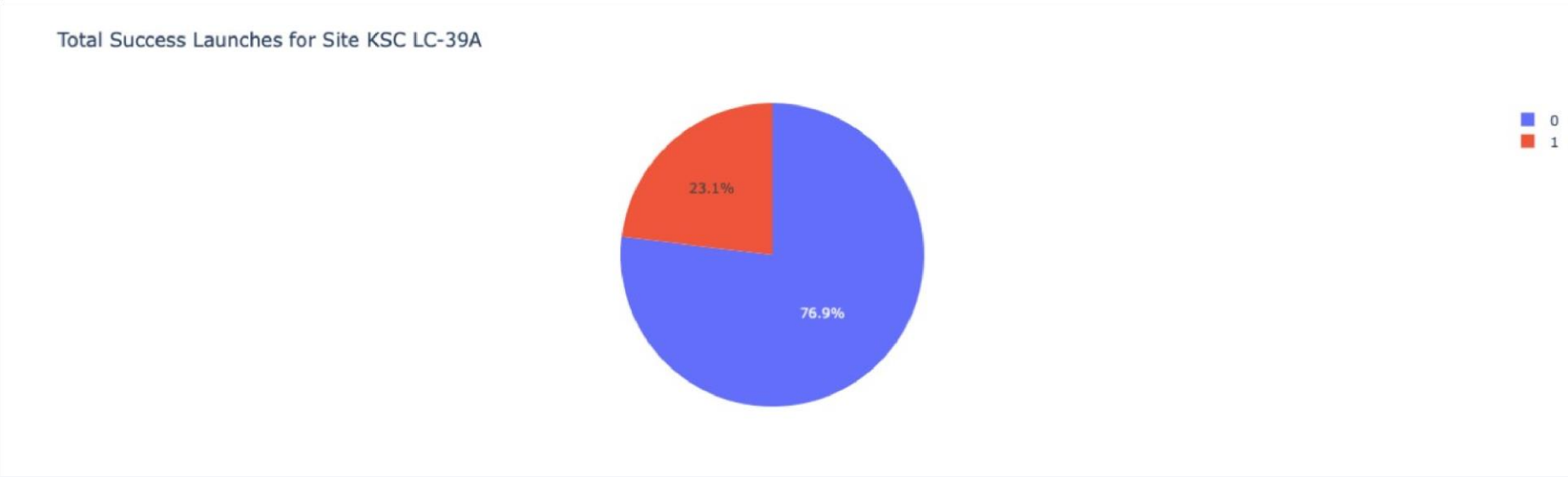
Launch success count for all sites



Explanation:

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch site with highest launch success ratio



Explanation:

- KSC LC-39A has the highest launch success rate (77%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites

Explanation:

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

Scores and Accuracy of the Test Set

- Based on the scores, we are not able to confirm which method performs best.

- Same Test Set scores, one possible explanation could be due to the small test sample size (18 samples). Therefore, we made a comparison among all methods based on the whole Dataset.

- The scores of the whole Dataset confirm that the best model is the **Decision Tree Model** This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

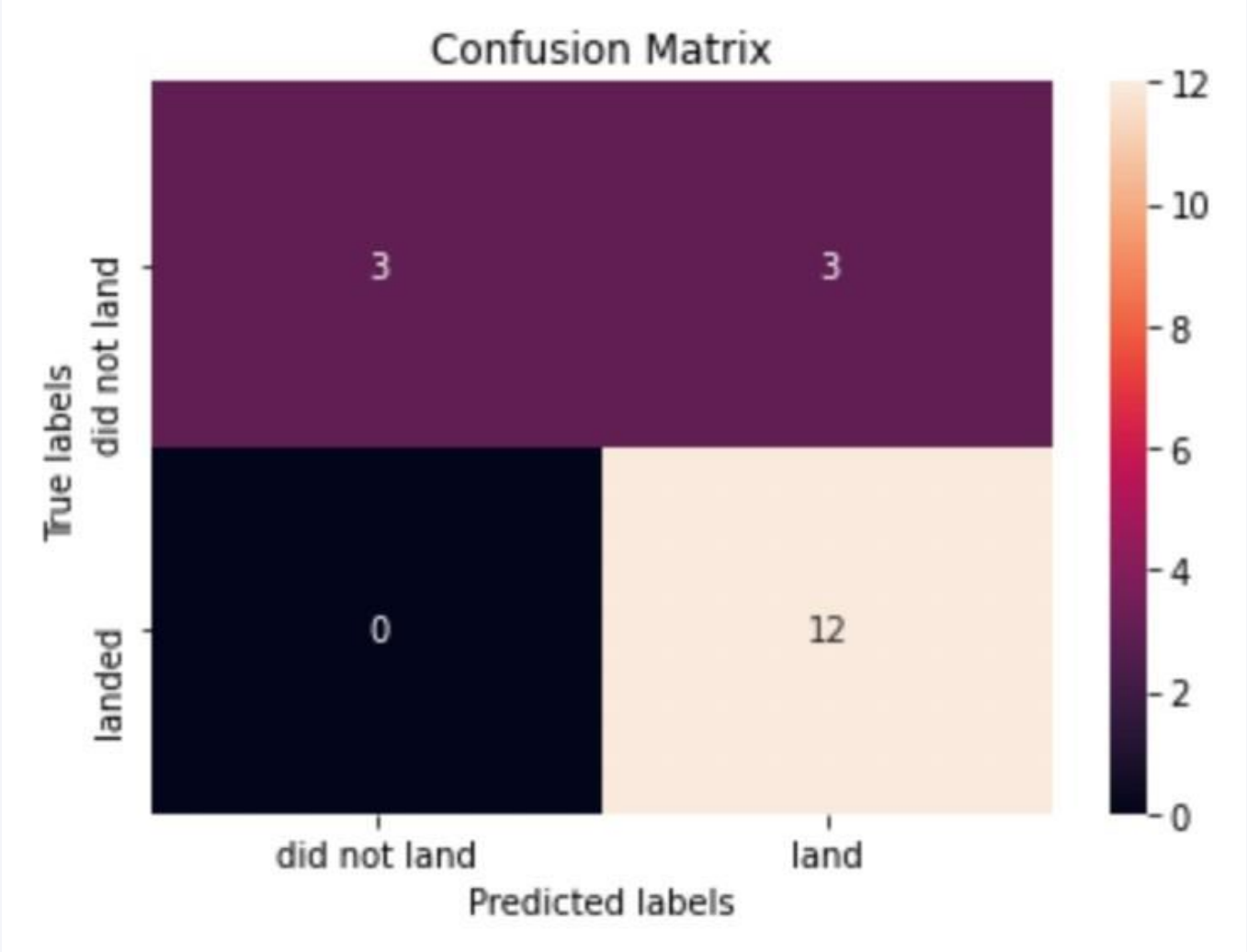
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation:

- As the confusion matrix shows, the logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

Decision Tree Model is the best algorithm for this dataset.

- Launches with a low payload mass usually show better results than launches with a larger payload mass.
- Most of launch sites are closer to the Equator line and all the sites are just beside the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have100% success rate.

Appendix

Thanks!

Thank you!