IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Manish
06/02/2024

Outline

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

Executive Summary

- Summary of methodologies
    - Data Collection
    - Data Wrangling
    - EDA with SQL
    - EDA with Data Visualization
    - Building an interactive map with Folium
    - Building a Dashboard with Plotly Dash
    - Predictive Analysis (Classification)
- Summary of all results
    - EDA Result
    - Interactive analytics demo in screenshots
    - Predictive analysis results

Introduction

Project background and context

The most prosperous business of the commercial space era, SpaceX has reduced the cost of space travel. On its website, the firm promotes Falcon 9 rocket flights, which start at 62 million dollars; in comparison, other suppliers charge up to 165 million dollars per launch; a large portion of the cost savings are attributable to SpaceX's ability to reuse the first stage. Thus, we can calculate the cost of a launch if we can ascertain if the first stage will land. We are going to make a prediction about whether SpaceX will reuse the first stage based on publicly available data and learning models.

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
  - Web Scraping from Wikipedia using Beautiful Soup
- 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 LR, SVN, Decision Tree and KNN models to ensure the best results.

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping 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.

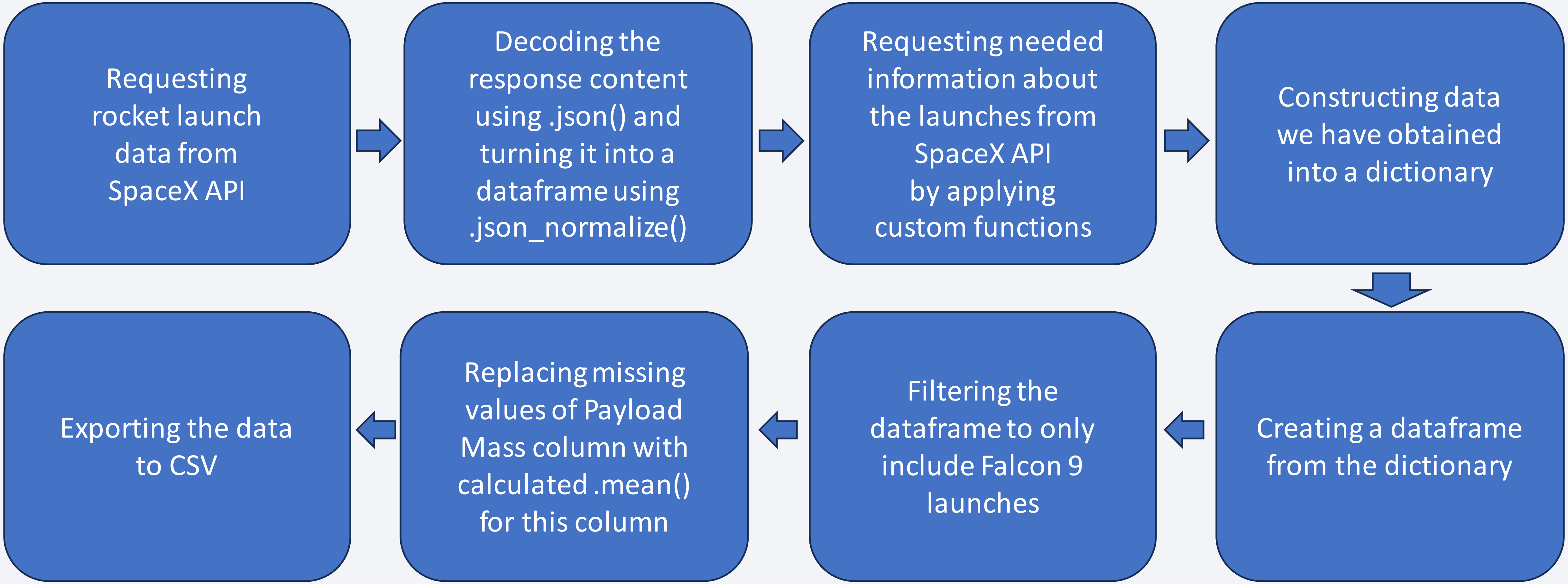
Data Columns are obtained by using SpaceX REST API:

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:

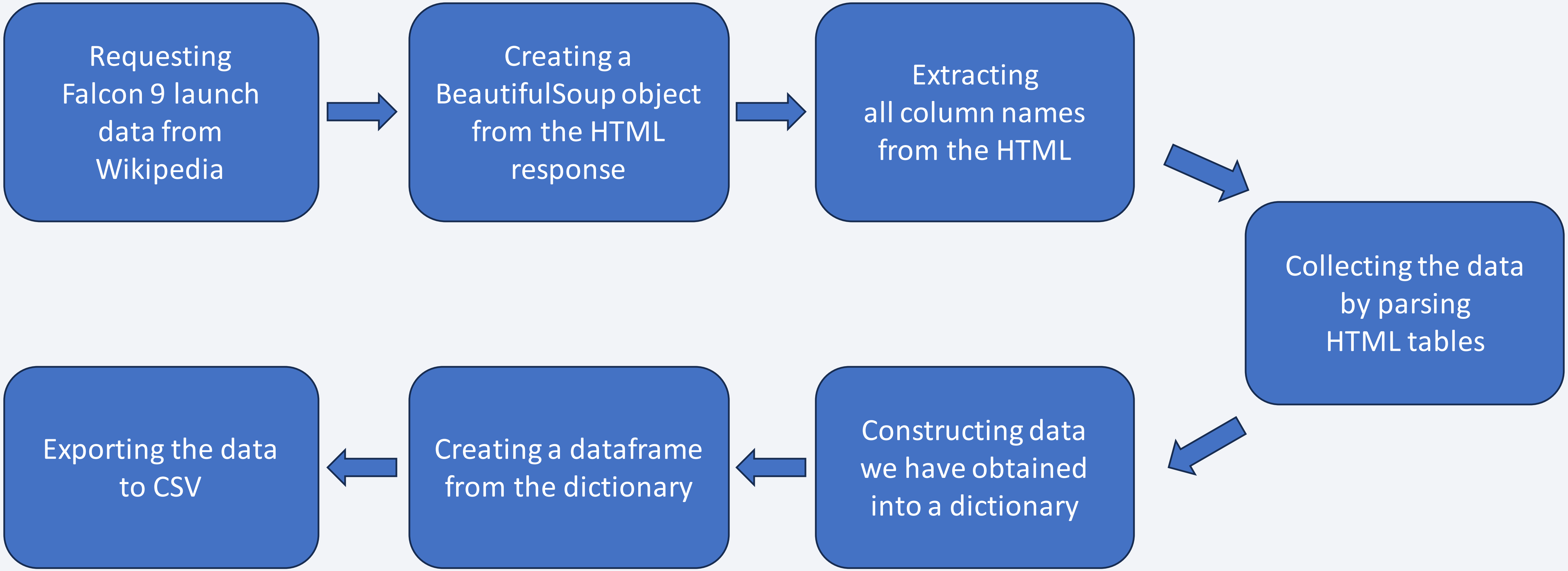
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



[GitHub URL: Data Collection using API]

Data Collection - Scraping

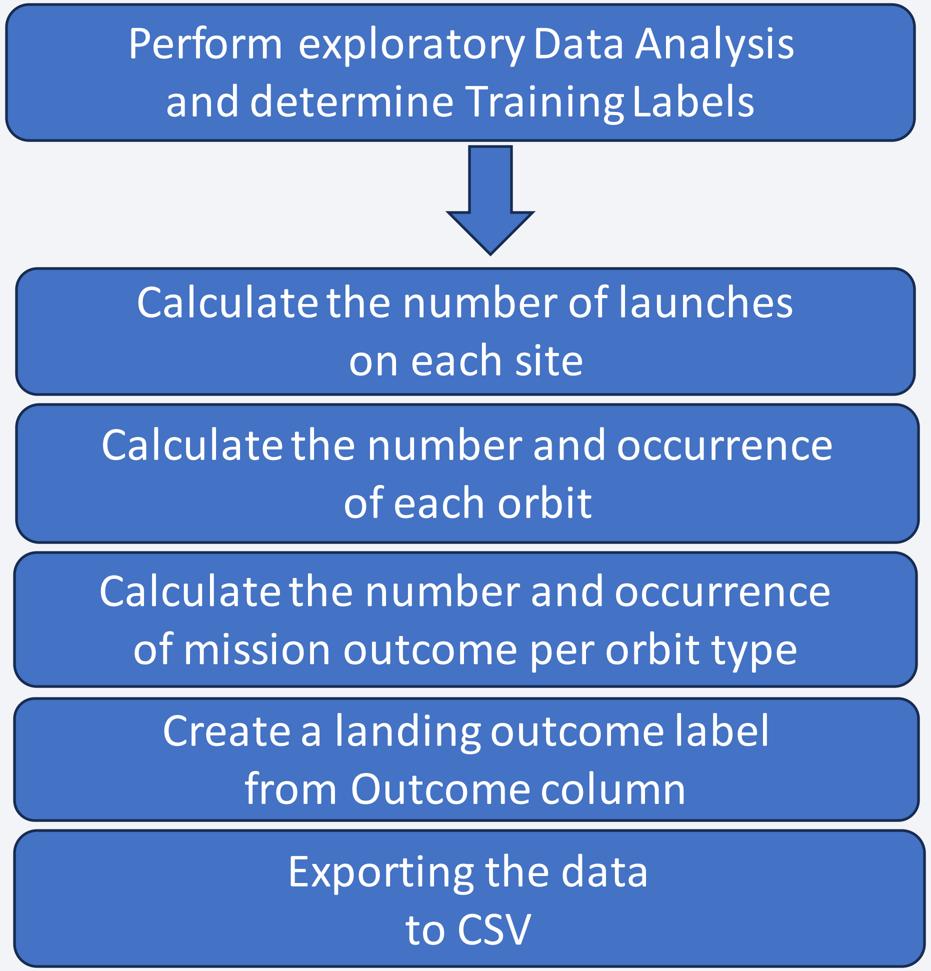


[GitHub URL: Data Collection using Web Scraping]

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground Pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.



[GitHub URL: Data Wrangling]

EDA with Data Visualization

Charts 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 and Success Rate Yearly Trend

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

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 URL: EDA with Data Visualization]

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the irst successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- 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

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured 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 coloured 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 Visual Analytics 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:

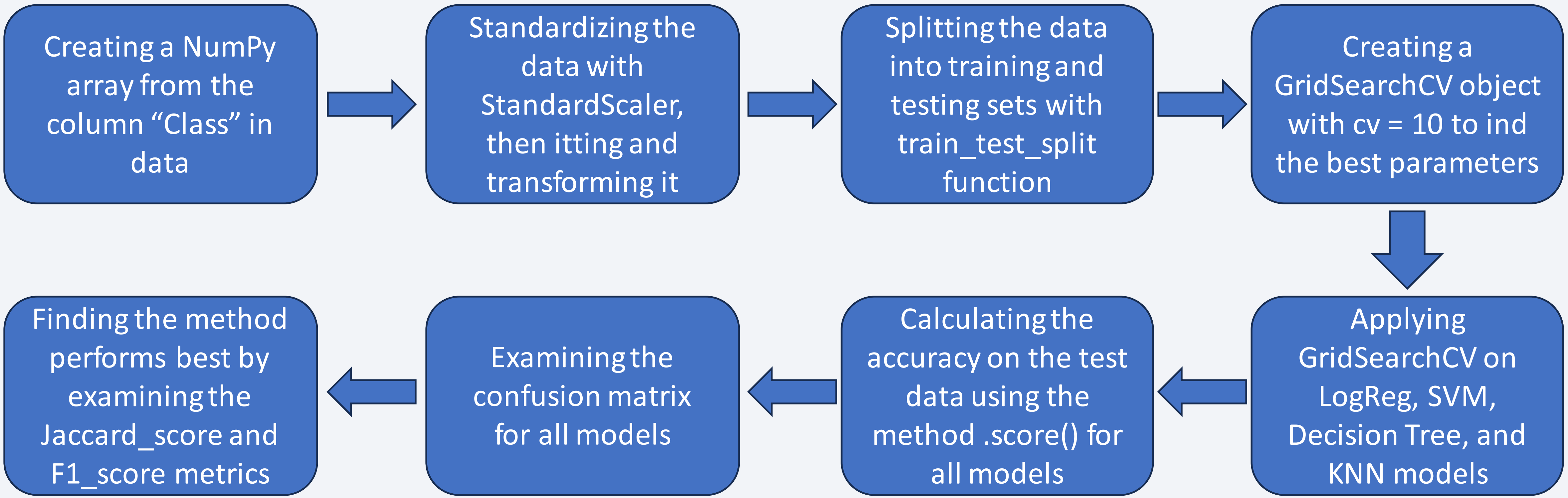
- Added a slider to select Payload range.

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: Interactive Dashboard with Plotly Dash]

Predictive Analysis (Classification)



GitHub URL: Predictive Classification using Machine Learning

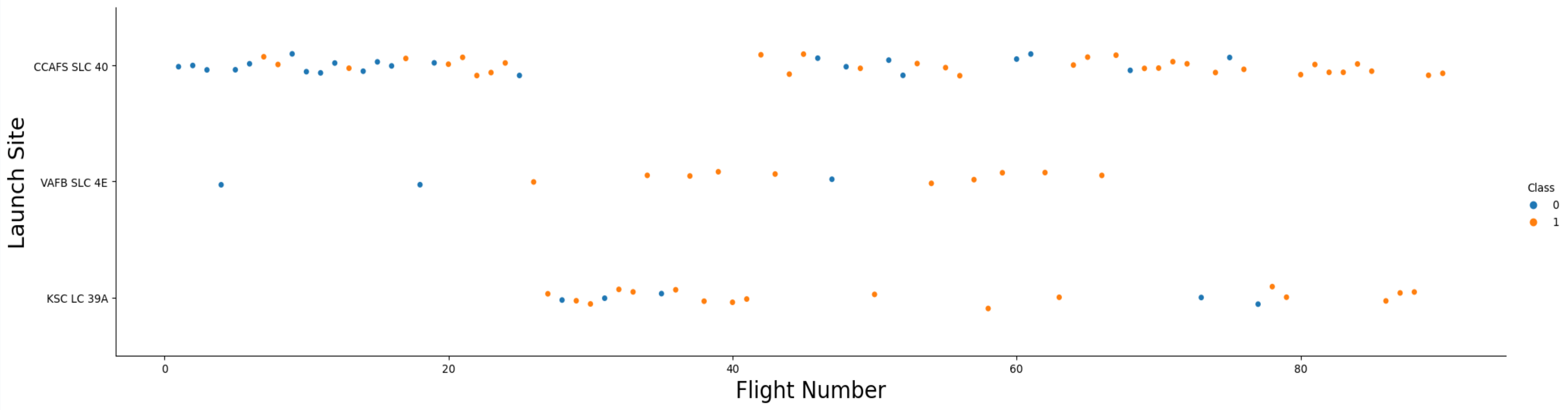
Results

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

Section 2

Insights drawn from EDA

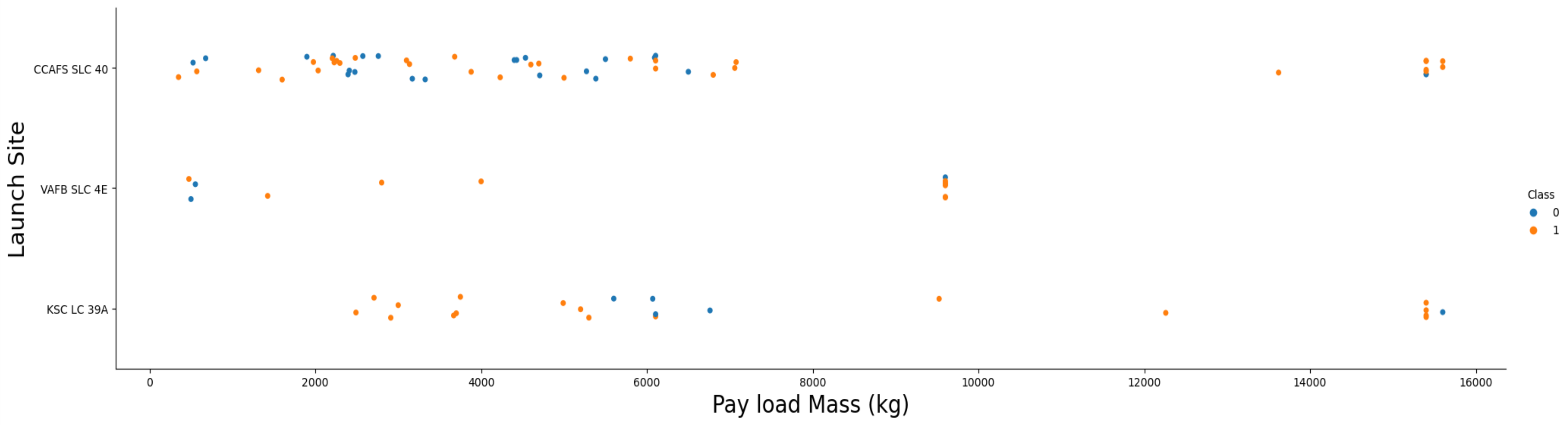
Flight Number vs. Launch Site



Explanation:

- The earliest flights all failed while the latest flights 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



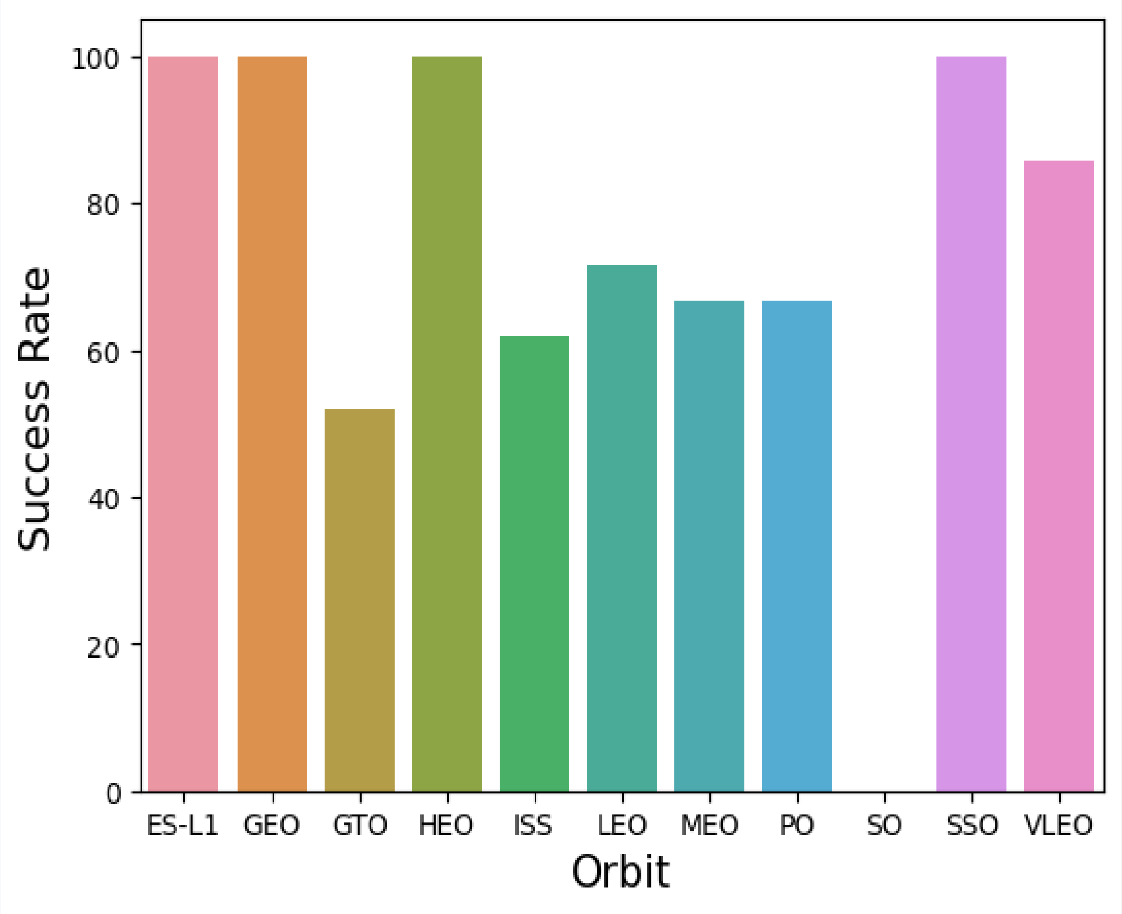
Explanation:

- 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.

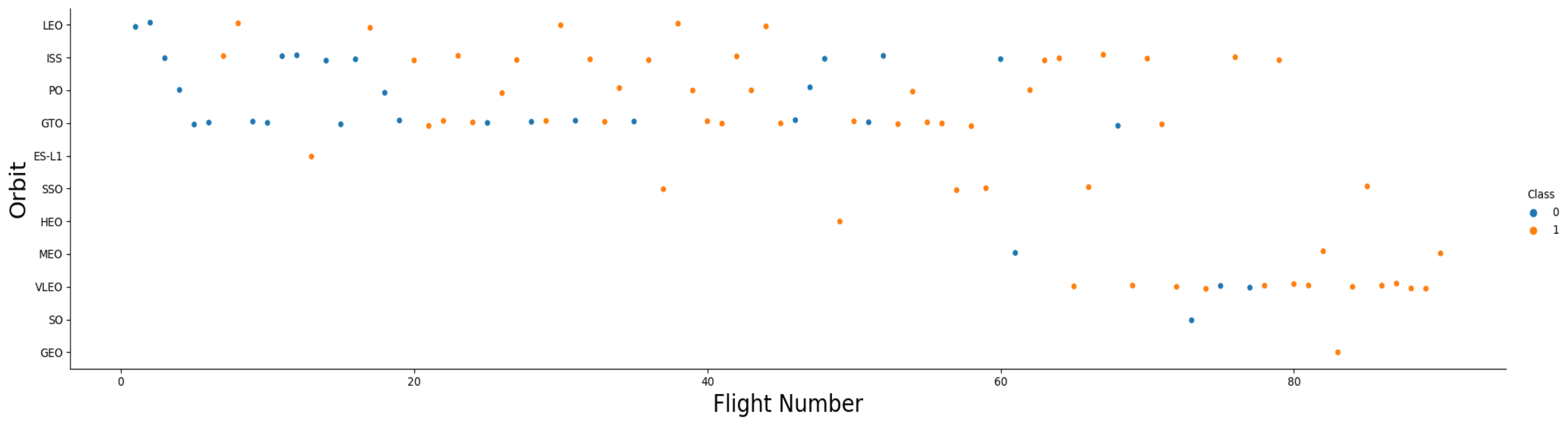
Success Rate vs. Orbit Type

Explanation:

- Orbit ES-L1, GEO, HEO and SSO have 100% success rate.
- Orbit SO has 0% success.
- Others orbit such as GTO, ISS, LEO, MEO, PO and VLEO have success rate between 50% to 85%.



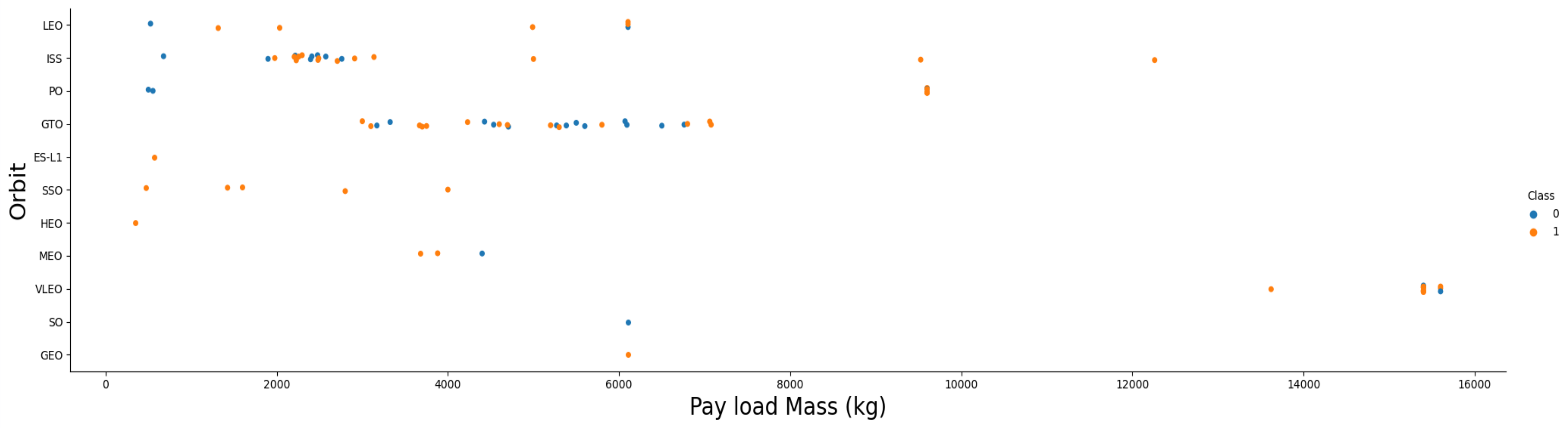
Flight Number vs. Orbit Type



Explanation:

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

Payload vs. Orbit Type



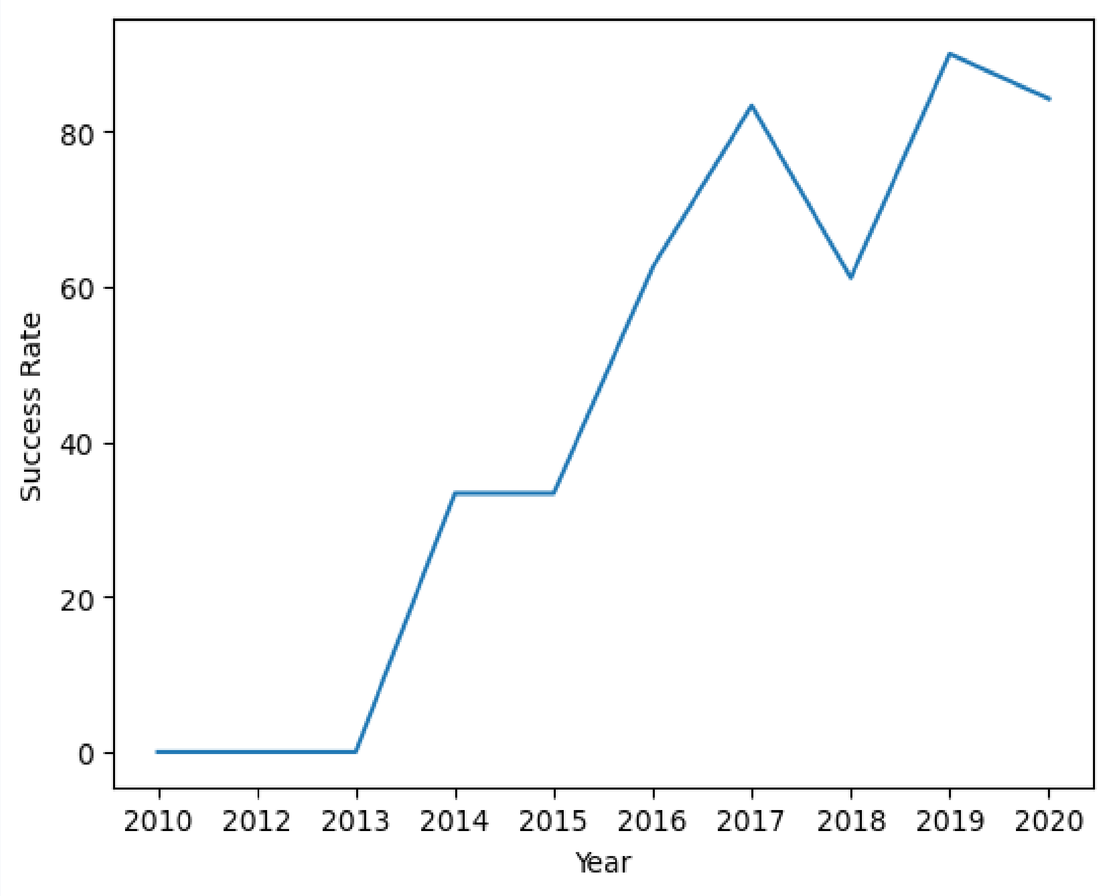
Explanation:

- Heavy payload has positive influence on Polar, LEO and ISS orbit.
- For GTO orbit, we cannot distinguish between positive and negative influence.

Launch Success Yearly Trend

Explanation:

- The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

```
In [15]: %sql select distinct Launch_Site from SPACEXTABLE

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

Out[15]:

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

Explanation:

- There are 4 distinct launch site i.e. CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

Launch Site Names Begin with 'CCA'

```
In [16]: %sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5;
 * sqlite:///my_data1.db
Done.
```

Out[16]:

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

Explanation:

- First 5 launch site whose name begin with "CCA" is CCAFS LC-40 and mission outcome for all 5 launch is successful.

Total Payload Mass

```
In [23]: %sql select sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass_by_NASA_CRS from SPACEXTABLE where Customer = "NASA (CRS)"

 * sqlite:///my_data1.db
Done.

Out[23]:
```

Total_Payload_Mass_by_NASA_CRS
45596

Explanation:

- Total payload mass by "NASA (CRS)" is 45596 kg.

Average Payload Mass by F9 v1.1

```
In [25]: sql select avg(PAYLOAD_MASS__KG_) as "Avg_Payload_Mass_by_F9_V1.1" from SPACEXTABLE where Booster_Version = "F9 v1.1

 * sqlite:///my_data1.db
Done.

Out[25]:
```

Avg_Payload_Mass_by_F9_V1.1
2928.4

Explanation:

- Average payload mass for booster version "F9 v1.1" is 2928.4 kg.

First Successful Ground Landing Date

```
In [30]: %sql select min("Date") as "Date when the first succesful landing outcome in ground pad was acheived" from SPACEXTA

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

Out[30]:

Date when the first succesful landing outcome in ground pad was acheived
2015-12-22

Explanation:

- The first succesful landing outcome in ground pad was achieved on 22 December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [32]: %sql select Booster_Version from SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and (PAYLOAD_MASS__KG_

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

Out[32]:

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

Explanation:

- Booster version F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2 have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of Successful and Failure Mission Outcomes

```
In [35]: %sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome;

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

Out[35]:

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

Explanation:

- There are 100 successful mission out of 101 mission and 1 is failed in flight.

Boosters Carried Maximum Payload

```
In [38]: %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTA

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

Out[38]:

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

Explanation:

- There are 12 booster version which carried maximum payload mass.

2015 Launch Records

```
In [40]: %sql select substr("Date", 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where sub

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

Out[40]:

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	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

```
In [43]: %sql select Landing_Outcome, count(Landing_Outcome) as No_of_Landing_Outcome, Rank() over(order by count(Landing_Ou
```

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

Out[43]:

Landing_Outcome	No_of_Landing_Outcome	Rank
No attempt	9	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

Explanation:

- No attempt has 9 landing outcome which is highest between the date 2010-06-04 and 2017-03-20.
- Success and Failure in drone ship has 5 landing outcome each between the date 2010-06-04 and 2017-03-20.
- Precluded has 1 landing outcome which is lowest between the date 2010-06-04 and 2017-03-20.

Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on global map

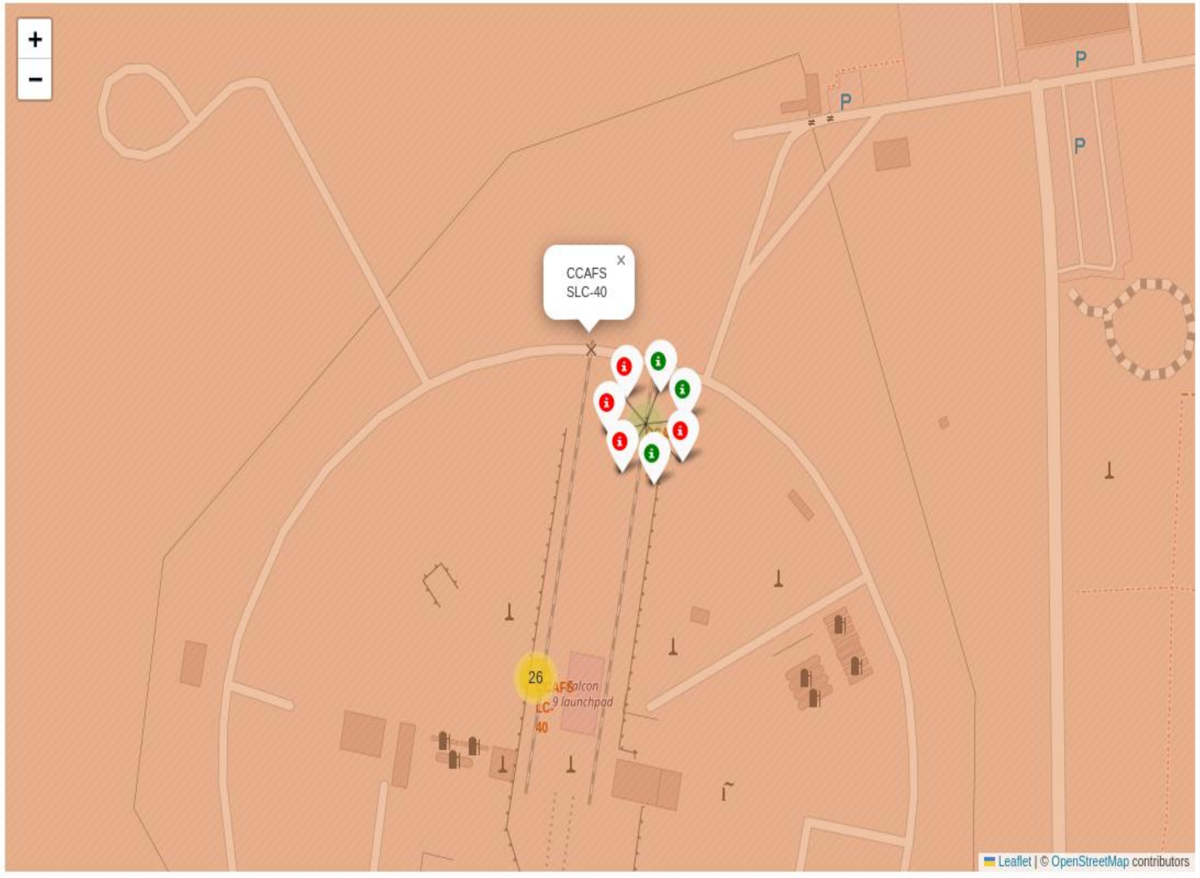
Explanation:

- 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. Anything on the surface of the Earth at the equator is already moving at 1670 km/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 because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimise the risk of having any debris dropping or exploding near people.

Colored-label launch site on map

Explanation:

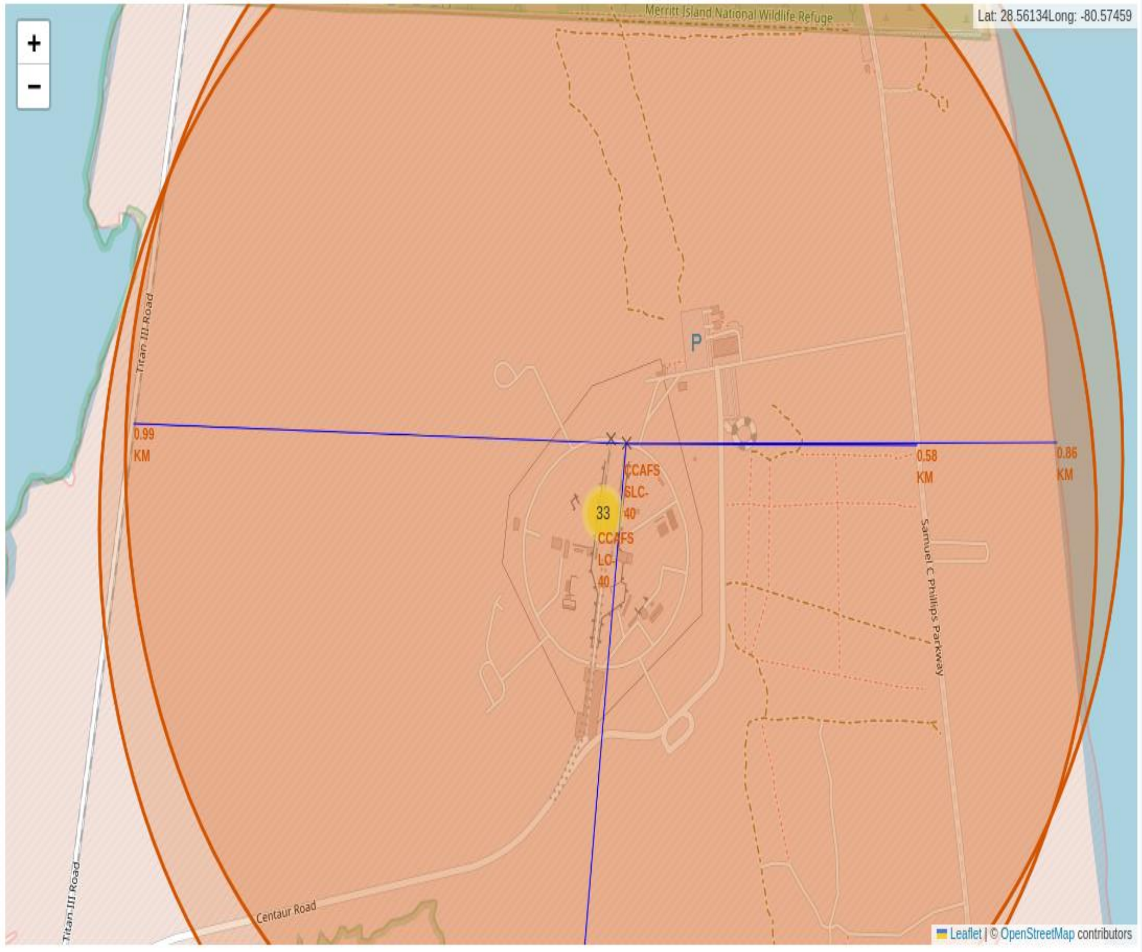
- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.



Distance from launch site CCAFS SLC-40 to its proximity

Explanation:

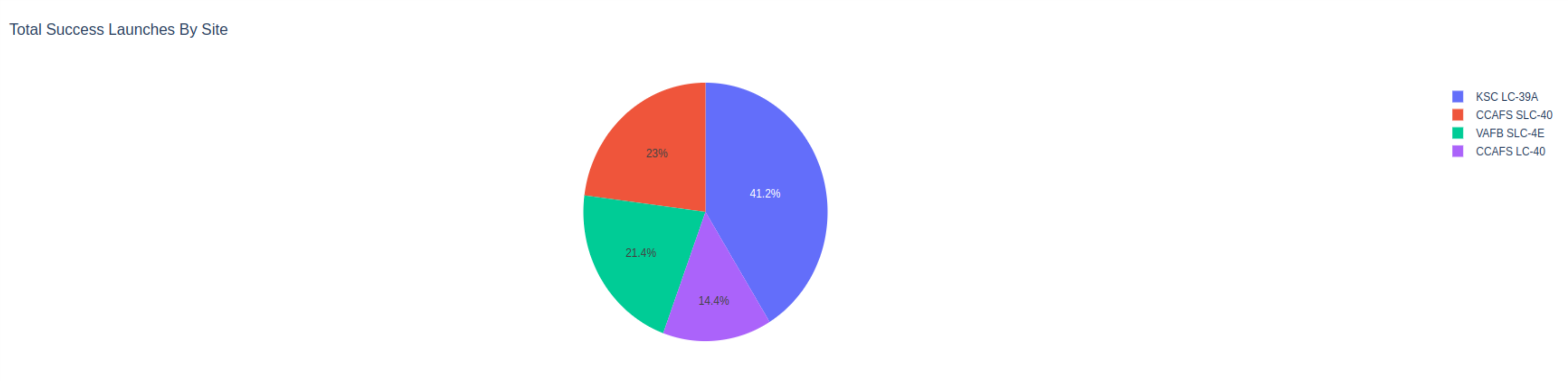
- From the visual analysis of the launch site CCAFS SLC-40 we can clearly see that it is:
  - relative close to railway (0.99 km)
  - relative close to highway (0.58 km)
  - relative close to coastline (0.86 km)
- Also the launch site CCAFS SLC-40 is relative close to its closest city Melbourn (50.94 km).



Section 4

Build a Dashboard with Plotly Dash

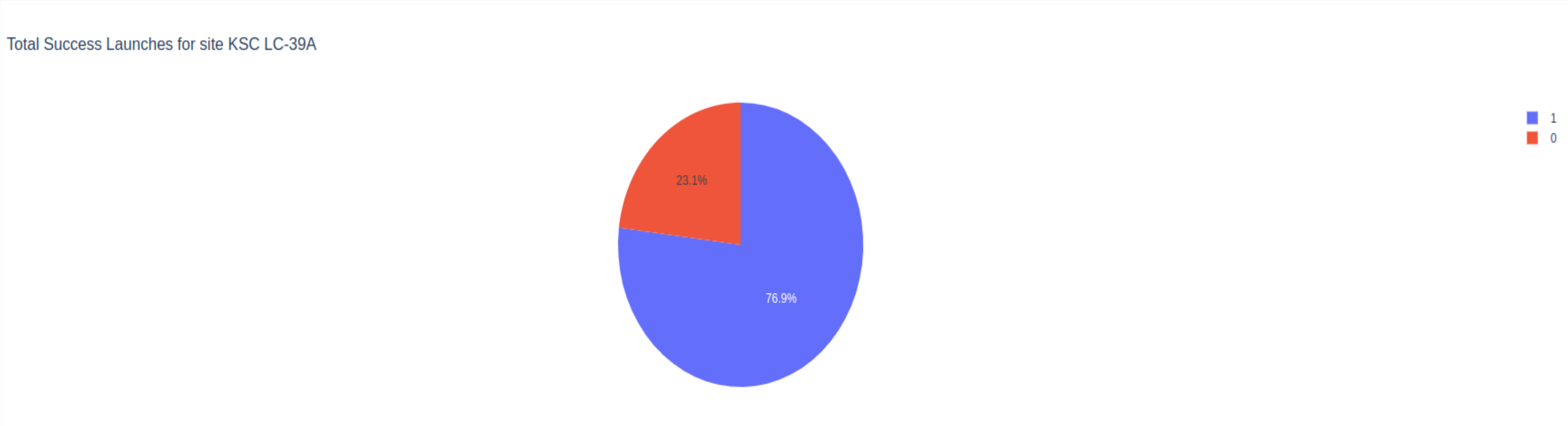
Launch Success count for all sites



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



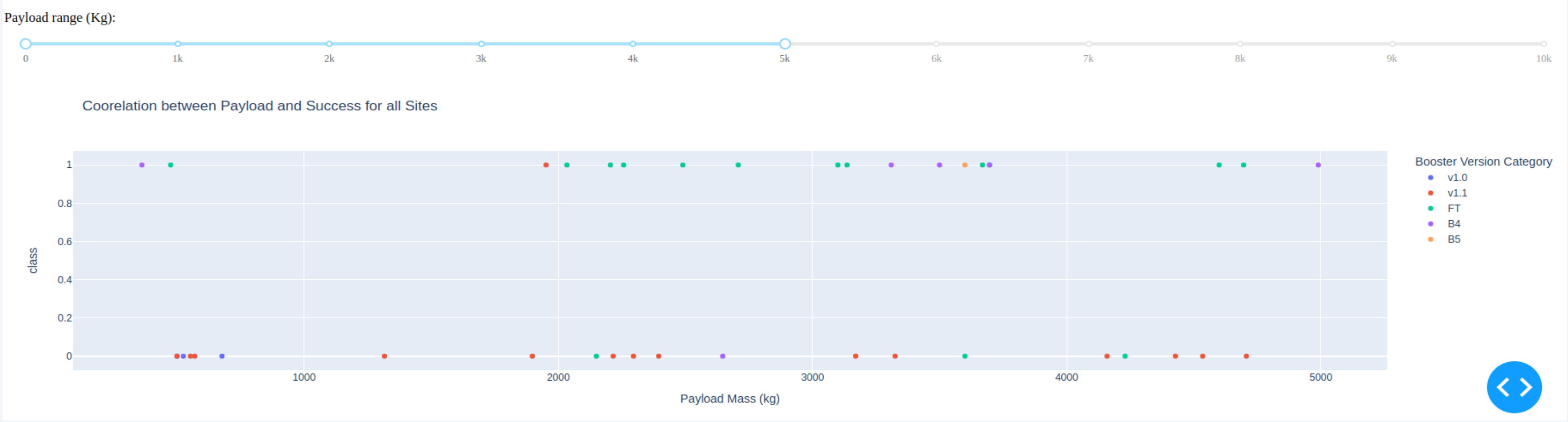
Explanation:

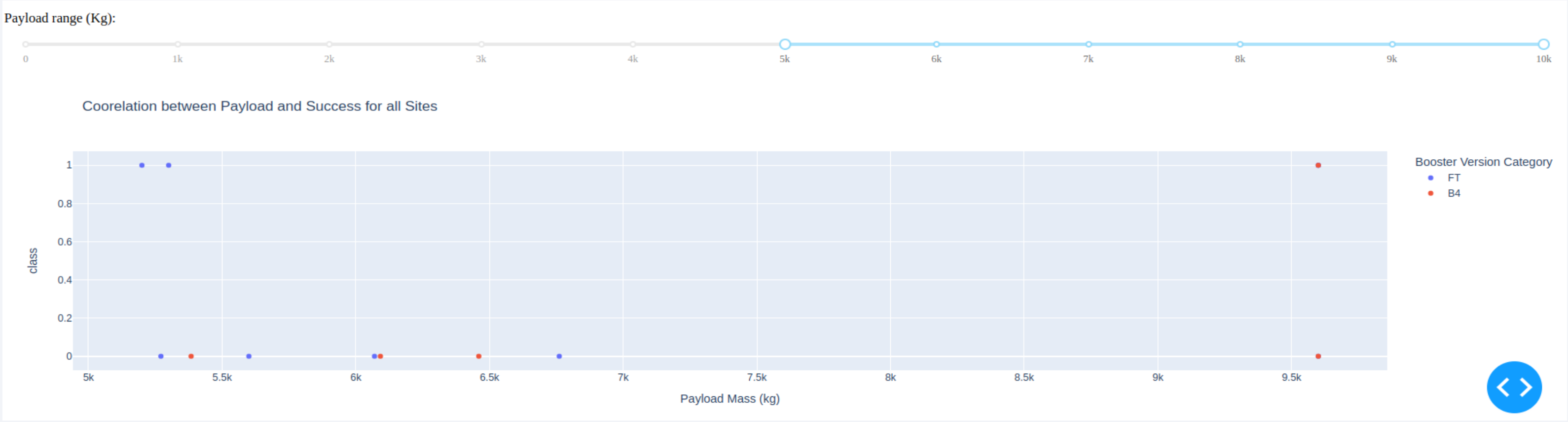
- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all site

Explanation:

- The charts shows that payloads between between 0 to 5000 kg have highest success rate.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- Based on the scores of the Test Set, we can say that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.
- The scores of the whole Dataset are almost same for all model.

Score and Accuracy on test set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.846154	0.800000
F1_Score	0.888889	0.888889	0.916667	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

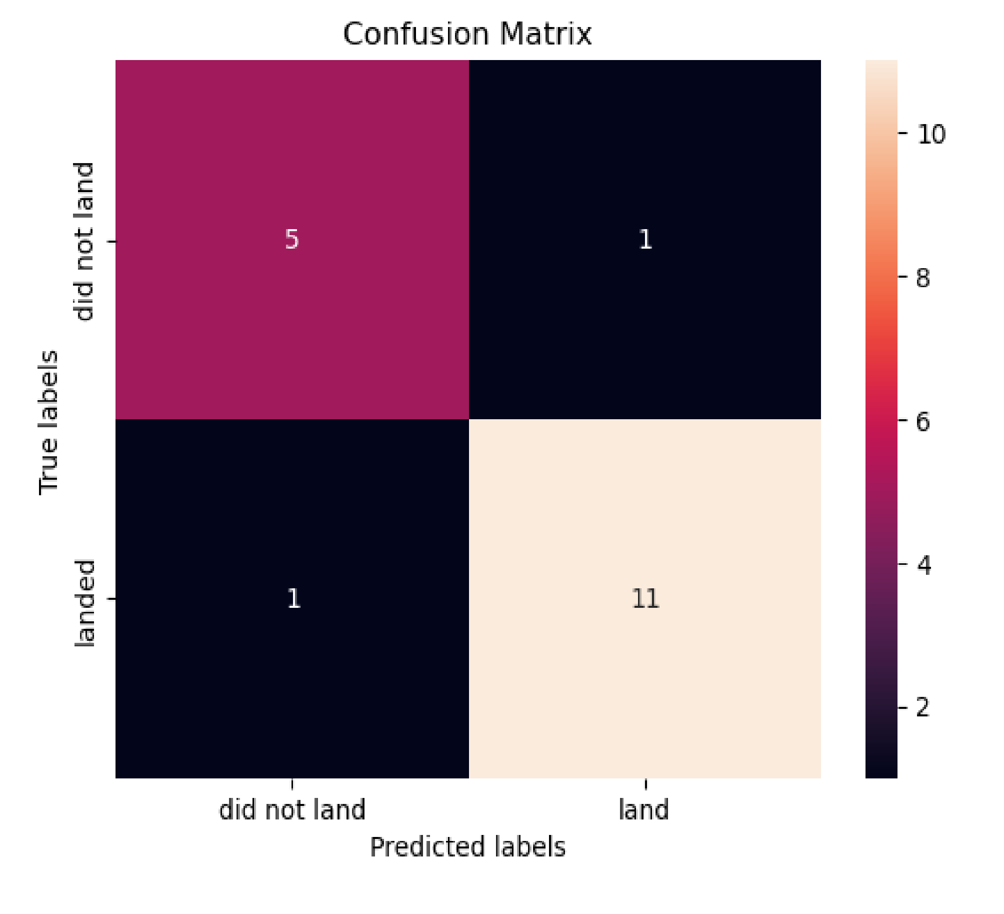
Score and Accuracy on whole dataset

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.843750	0.819444
F1_Score	0.909091	0.916031	0.915254	0.900763
Accuracy	0.866667	0.877778	0.888889	0.855556

Confusion Matrix

Explanation:

- Confusion Matrix for Decision Tree model.
- From the confusion matrix, we can say that False Positive is 1 and False Negative is also 1.



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years except 2018.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- SQL Query for listing the date when the first succesful landing outcome in ground pad was acheived.
  - %sql select min("Date") as "Date when the first succesful landing outcome in ground pad was acheived" from SPACEXTABLE where Landing_Outcome = "Success (ground pad)";
- SQL Query for listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - %sql select Booster_Version from SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000);
- SQL Query for listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - %sql select substr("Date", 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where substr("Date",0,5) = "2015" and Landing_Outcome = "Failure (drone ship)";
- SQL Quer for 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.
  - %sql select Landing_Outcome, count(Landing_Outcome) as No_of_Landing_Outcome, Rank() over(order by count(Landing_Outcome) desc) Rank from SPACEXTABLE where "Date" between "2010-06-04" and "2017-02-20" group by Landing_Outcome;

Thank you!