

Winning Space Race with Data Science

SZ
Oct 5, 2023



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 the most successful company of the commercial space age, making space travel affordable. 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 first stage. Therefore, if we can determine if the first 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.

Questions to be answered

- 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

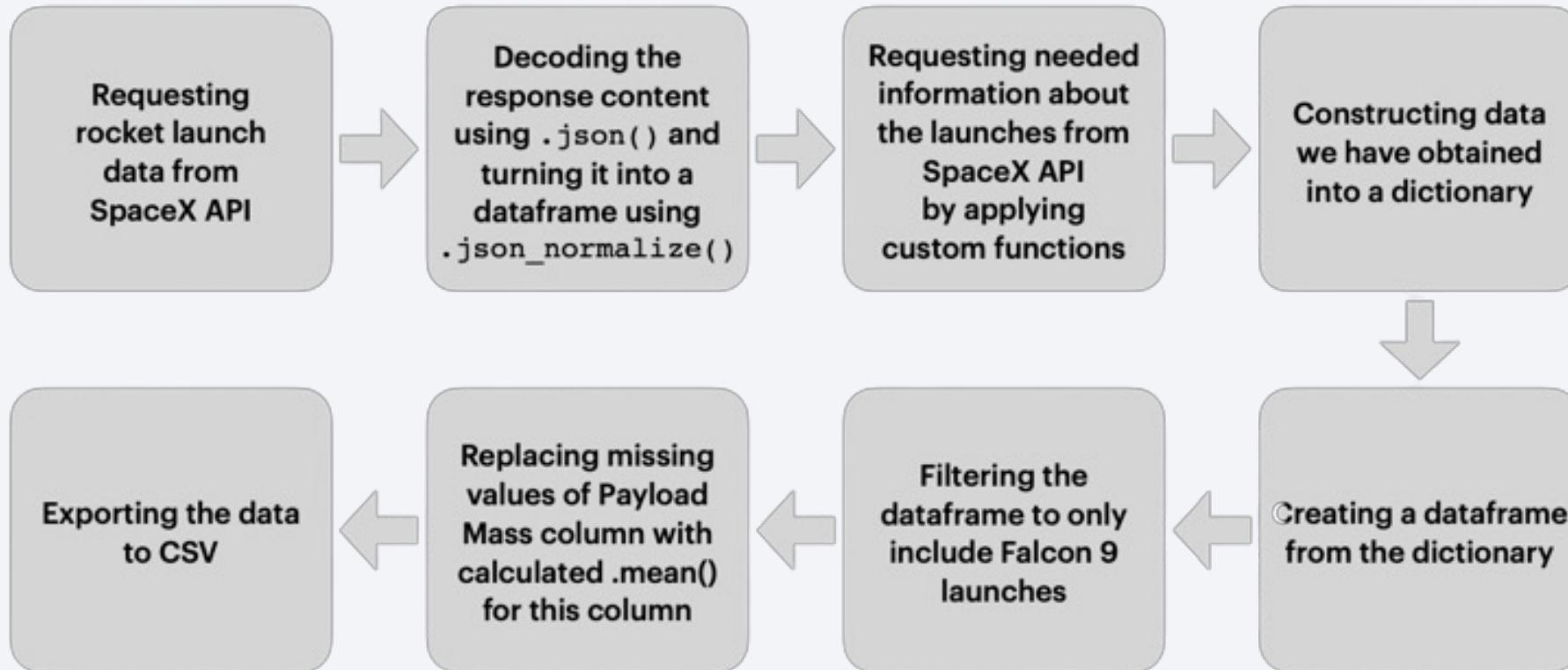
Data Collection

The 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.

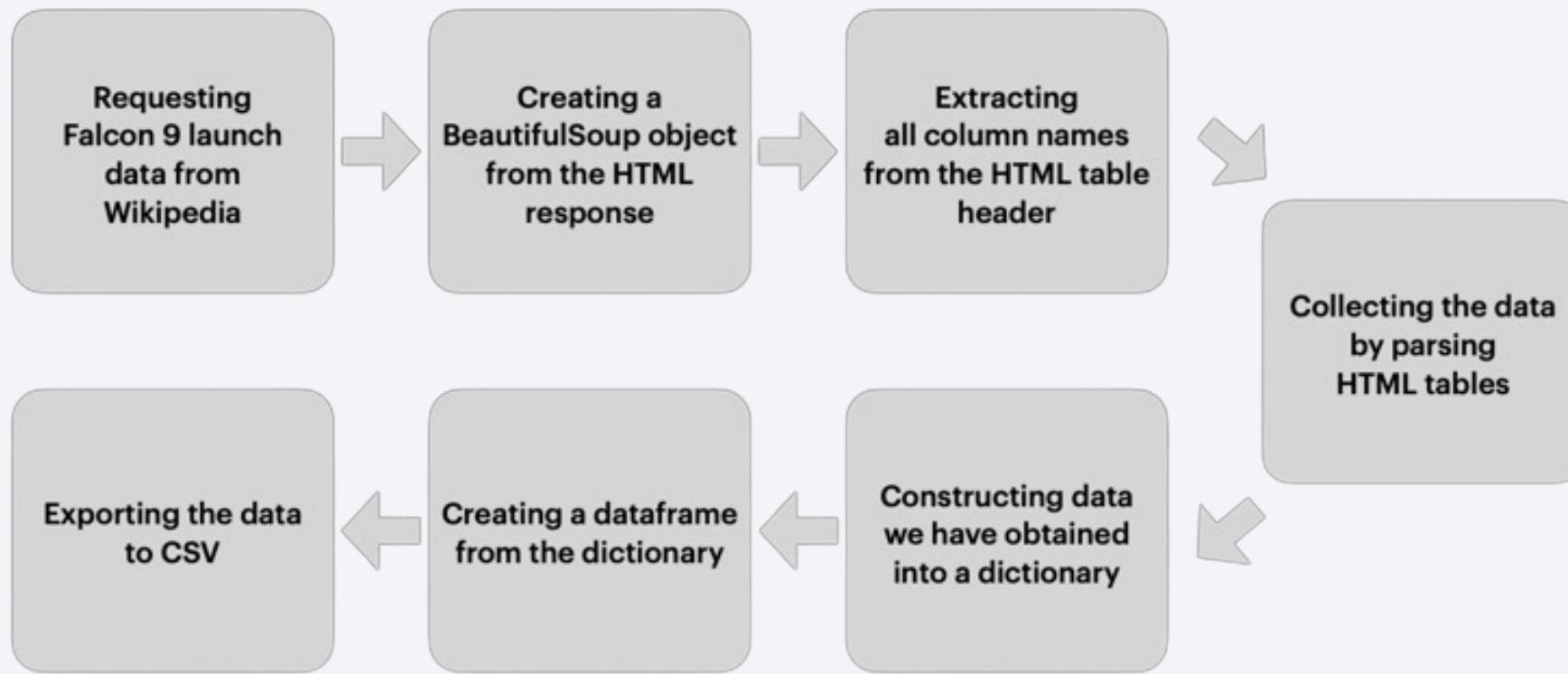
- Data Columns are obtained by using SpaceX REST API:
 - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
 - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



[GitHub URL: Data Collection API](#)

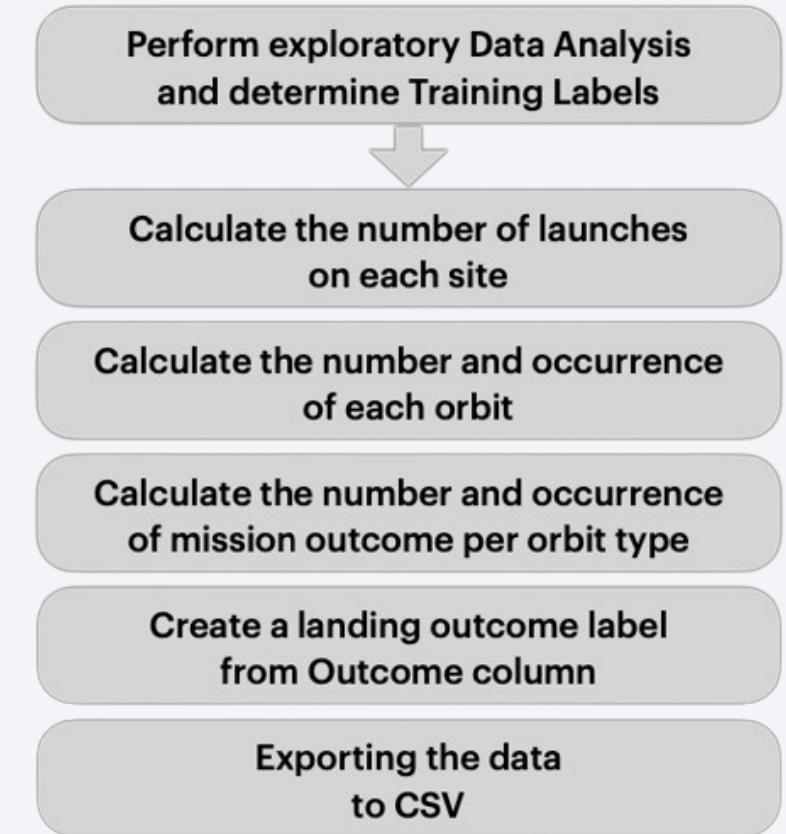
Data Collection - Scraping



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 while 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.



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).

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 first 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 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

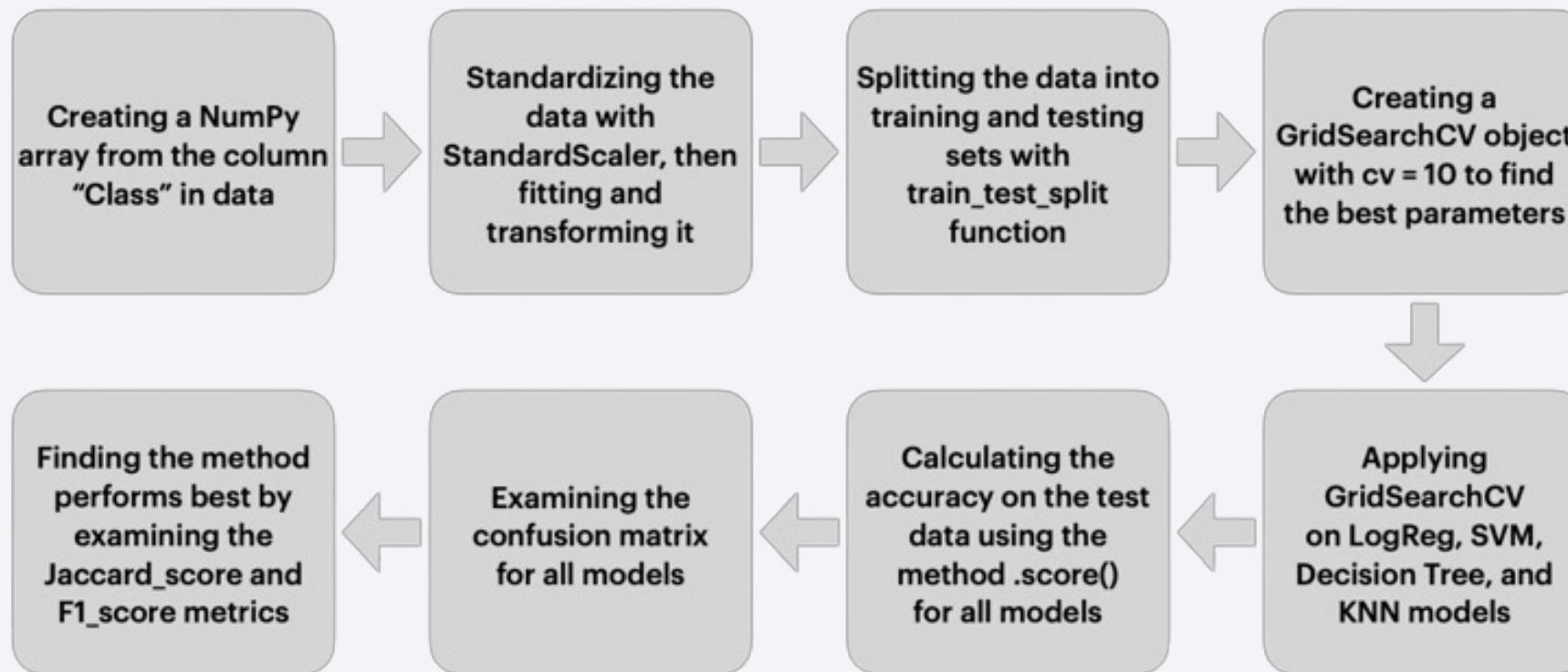
Build an Interactive Map with Folium

- **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.
- **Colored Markers of the launch outcomes for each Launch Site:**
 - Added colored 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 colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

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.

Predictive Analysis (Classification)



Results

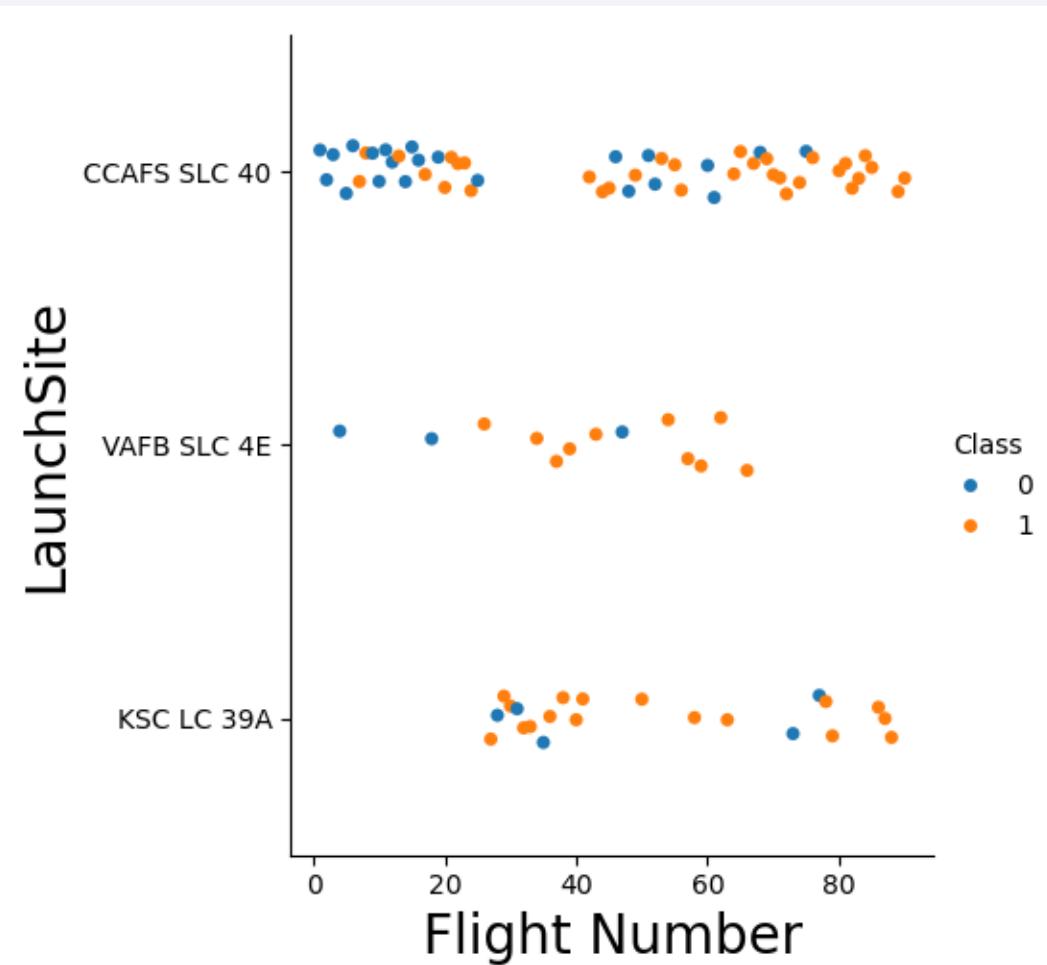
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a three-dimensional space or a network of data points. The overall effect is futuristic and dynamic.

Section 2

Insights drawn from EDA

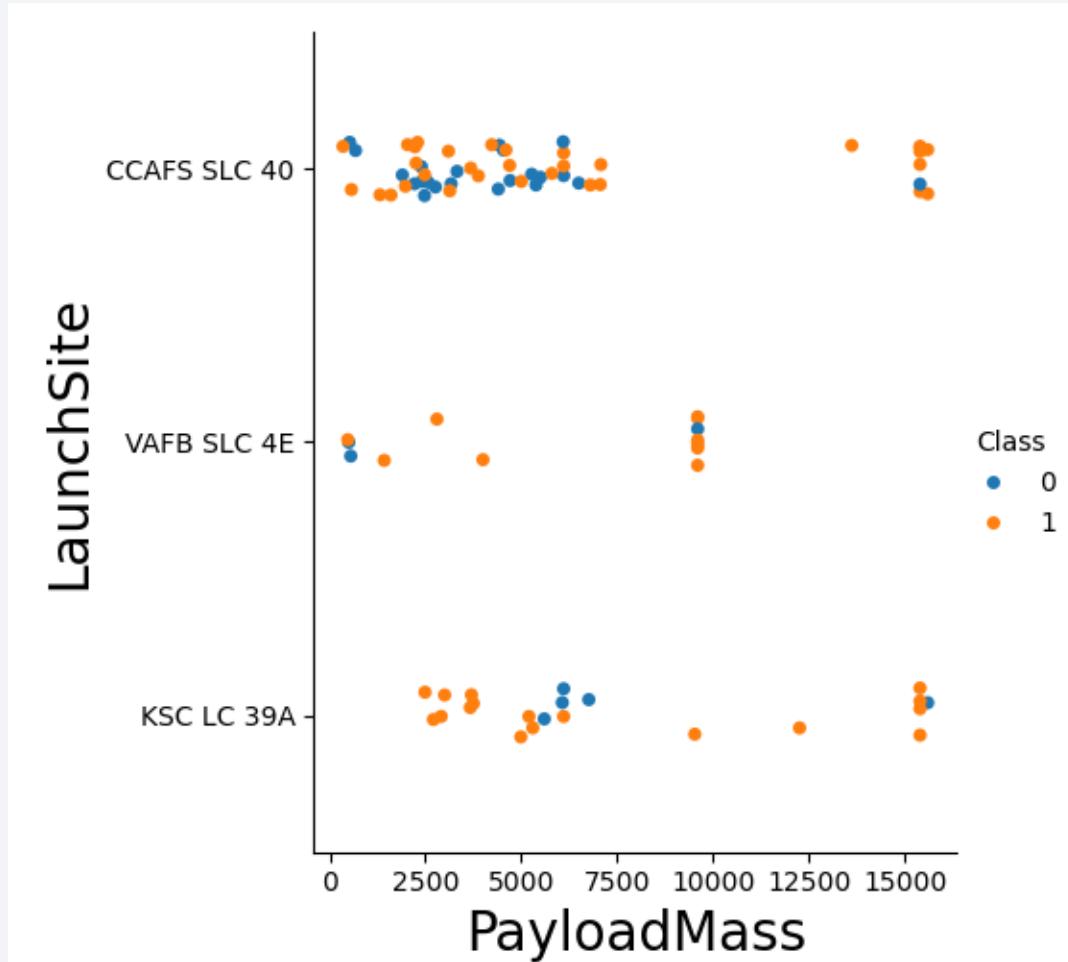
Flight Number vs. Launch Site



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.

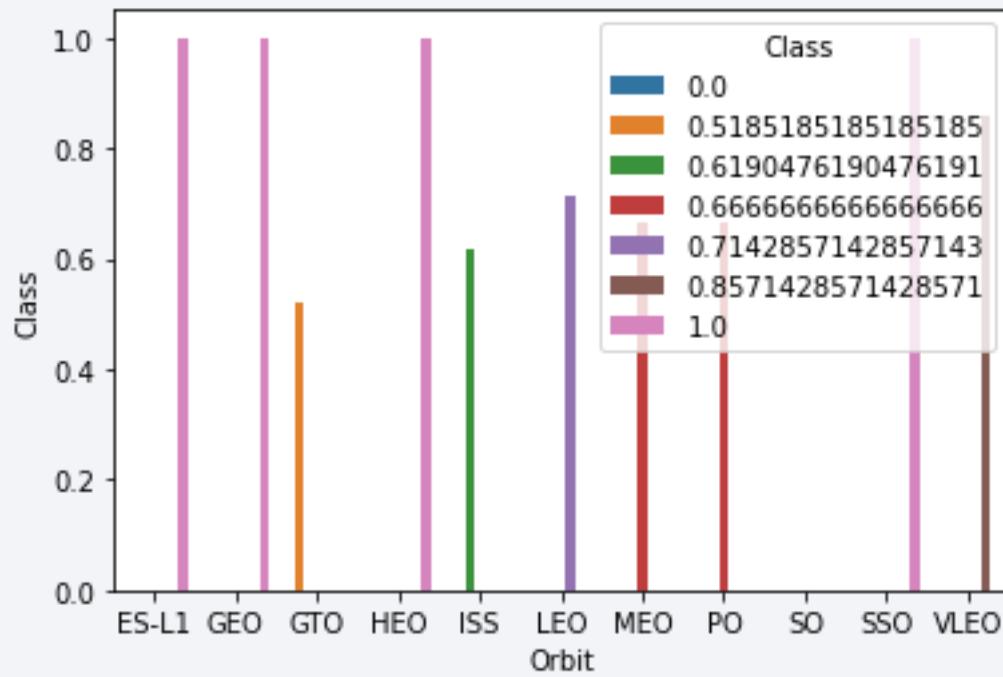
Payload vs. Launch Site



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.

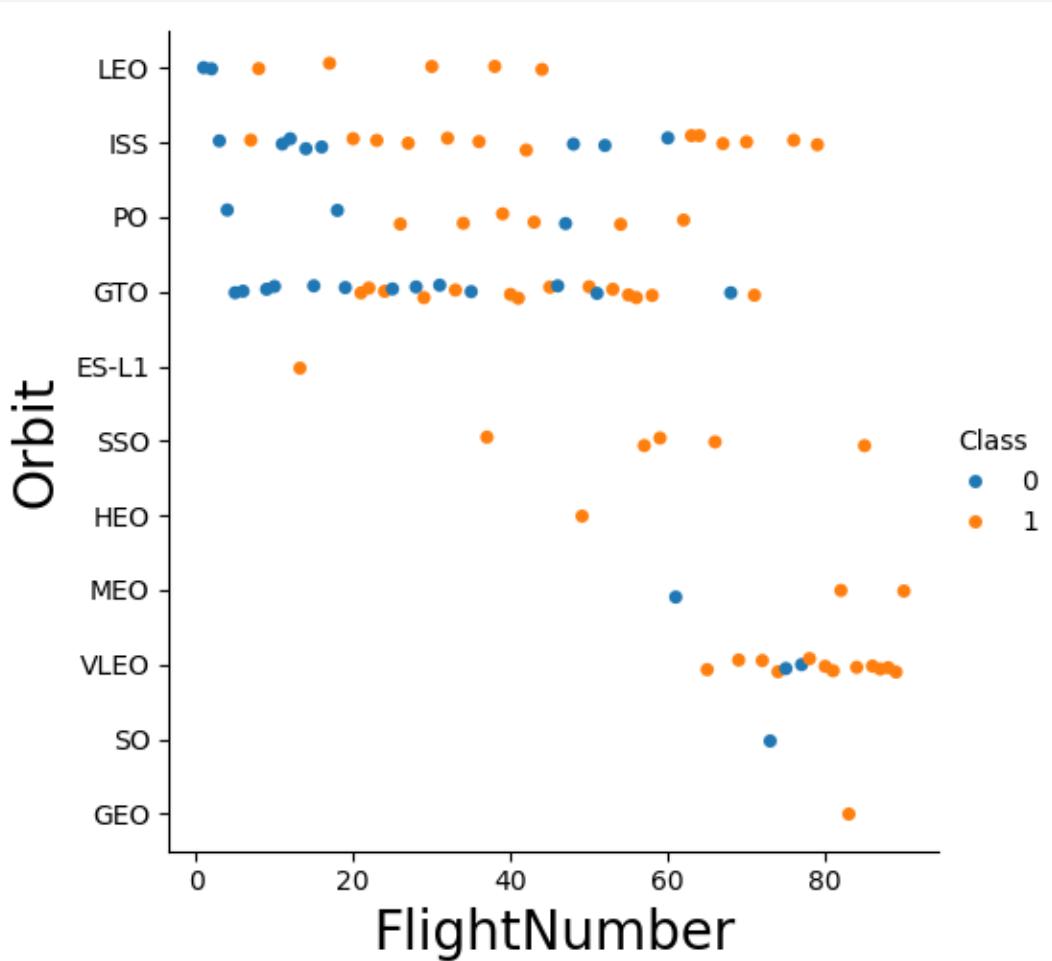
Success Rate vs. Orbit Type



Explanation:

- 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

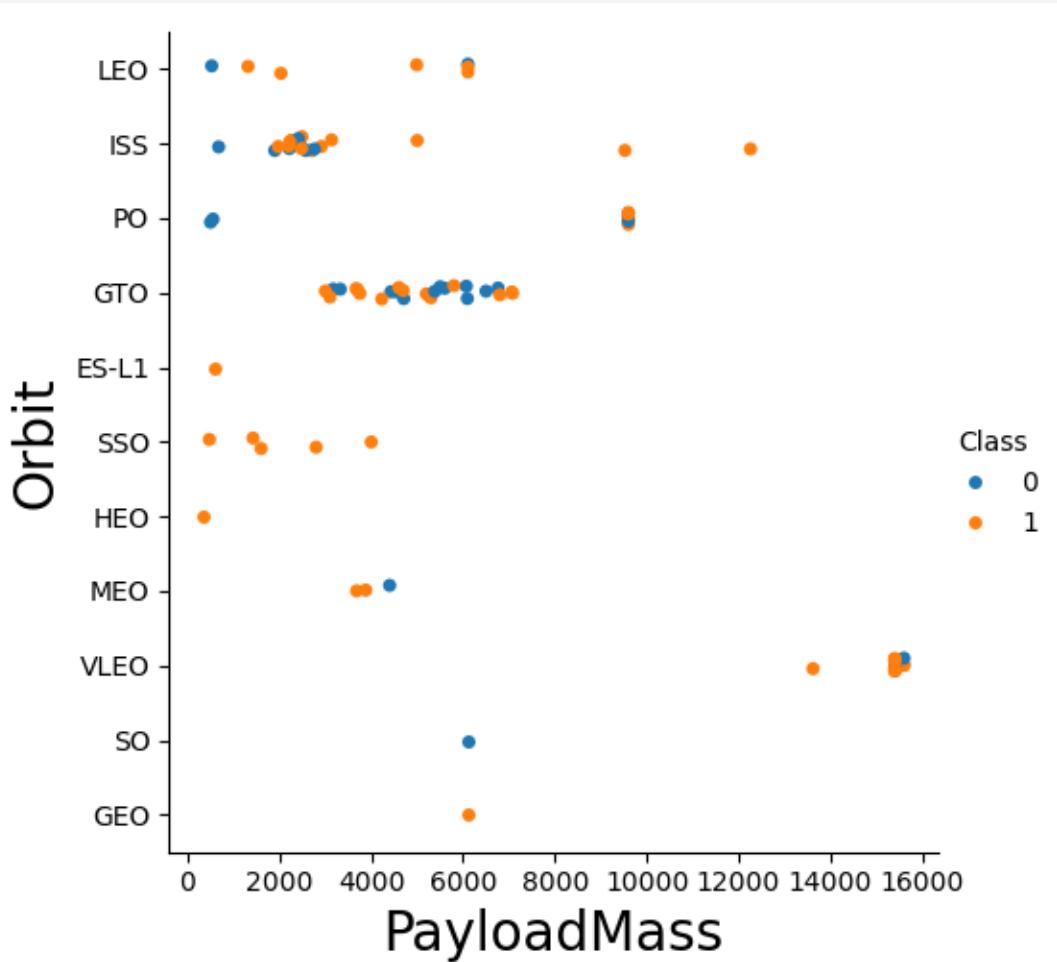
Flight Number vs. Orbit Type



Explanation:

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

Payload vs. Orbit Type



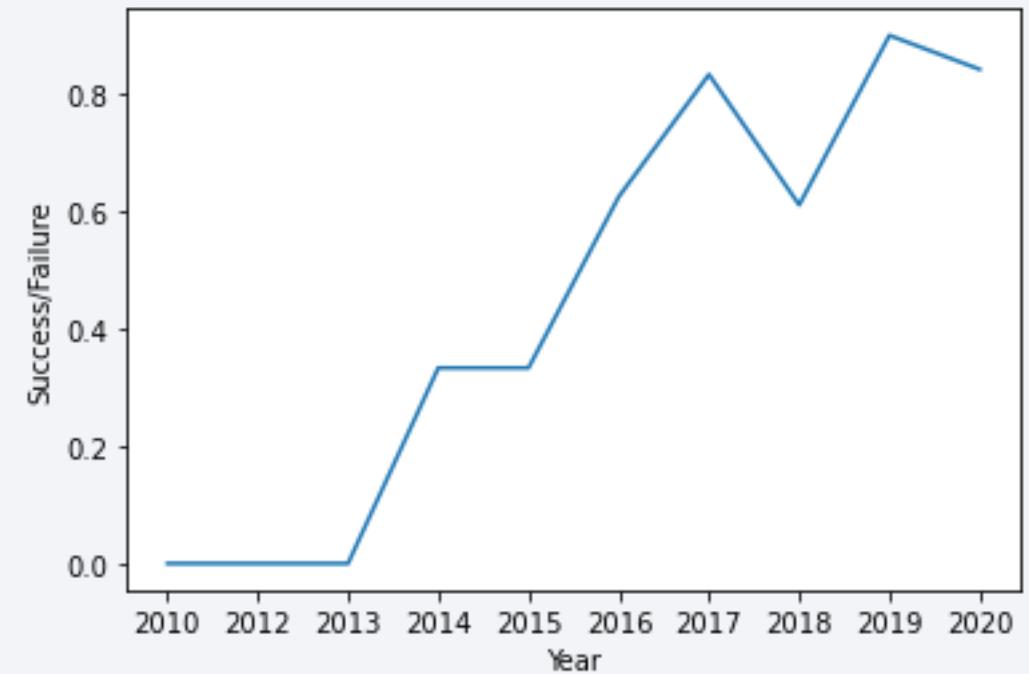
Explanation:

- 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 increased from 2013 until 2020



All Launch Site Names

Explanation:

- The names of the unique launch sites

```
In [55]: sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;  
* sqlite:///my_data1.db  
Done.  
Out[55]: Launch_Site  
_____  
CCAFS LC-40  
CCAFS SLC-40  
KSC LC-39A  
VAFB SLC-4E
```

Launch Site Names Begin with 'CCA'

Explanation:

- The 5 records where launch sites begin with `CCA`

In [56]:	sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;								
Out [56]:	* sqlite:///my_data1.db Done.								
	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcom
	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
	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)
	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Explanation:

- Total payload carried by boosters from NASA

```
In [57]: sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
```

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

```
Out[57]: TOTAL_PAYLOAD
```

```
111268
```

Average Payload Mass by F9 v1.1

Explanation:

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

```
In [58]: sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';  
* sqlite:///my_data1.db  
Done.  
Out[58]: AVG_PAYLOAD  
2928.4
```

First Successful Ground Landing Date

Explanation

- The date of the first successful landing outcome on ground pad: December 22, 2015

```
In [59]: sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';

* sqlite://my_data1.db
Done.

Out[59]: FIRST_SUCCESS_GP
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

Explanation:

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

Booster_Version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Explanation:

- The total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

Explanation:

- The names of the booster which have carried the maximum payload mass

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

2015 Launch Records

Explanation:

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

1	mission_outcome	booster_version	launch_site
1	Success	F9 v1.1 B1012	CCAFS LC-40
2	Success	F9 v1.1 B1013	CCAFS LC-40
3	Success	F9 v1.1 B1014	CCAFS LC-40
4	Success	F9 v1.1 B1015	CCAFS LC-40
4	Success	F9 v1.1 B1016	CCAFS LC-40
6	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

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

Explanation:

- Ranking of 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

Out [44]:	landing_outcome
	No attempt
	Success (ground pad)
	Success (drone ship)
	Success (drone ship)
	Success (ground pad)
	Failure (drone ship)
	Success (drone ship)
	Success (drone ship)
	Success (drone ship)
	Failure (drone ship)
	Failure (drone ship)
	Success (ground pad)
	Precluded (drone ship)
	No attempt
	Failure (drone ship)
	No attempt
	Controlled (ocean)
	Failure (drone ship)
	Uncontrolled (ocean)
	No attempt
	No attempt
	Controlled (ocean)
	Controlled (ocean)
	No attempt
	No attempt
	Uncontrolled (ocean)
	No attempt
	No attempt
	No attempt
	Failure (parachute)
	Failure (parachute)

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against a dark blue-black void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States appears. In the upper right, the green and yellow glow of the aurora borealis is visible. The atmosphere of the Earth is thin and hazy, appearing as a light blue band near the horizon.

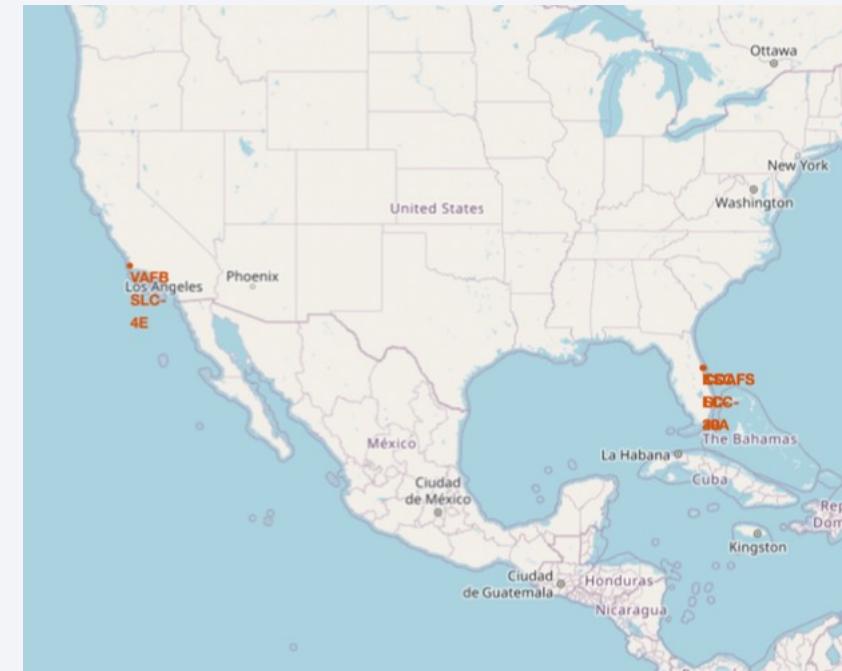
Section 3

Launch Sites Proximities Analysis

Location of all launch sites on a global map

Explanation:

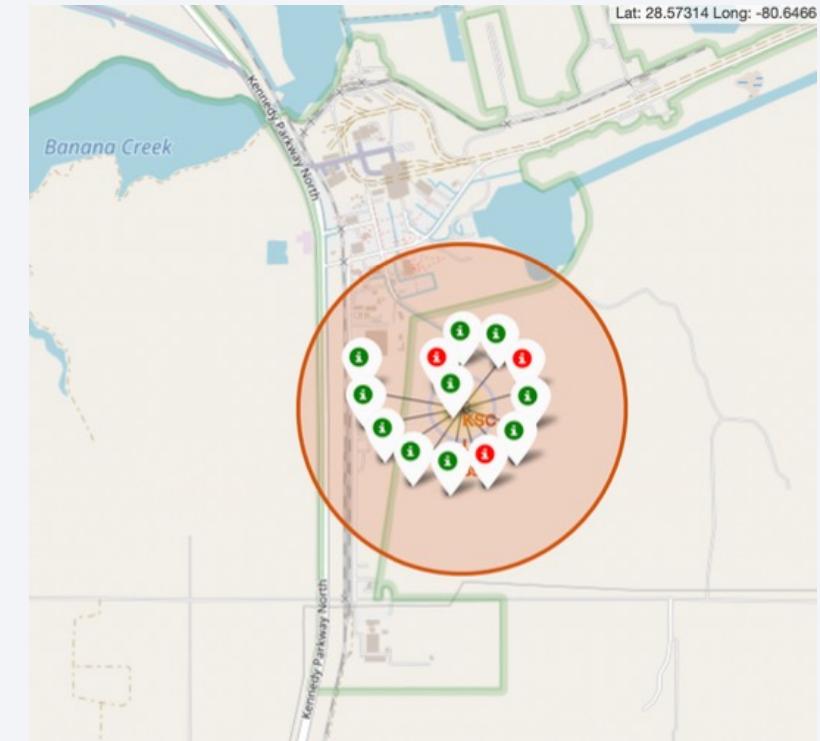
- Most of Launch sites are in proximity to the Equator. The land is moving
- All launch sites are also very close to the coast.
- Launching rockets towards the ocean minimizes the risk of explosive debris reaching populated areas.



Color-labeled launch records on the map

Explanation:

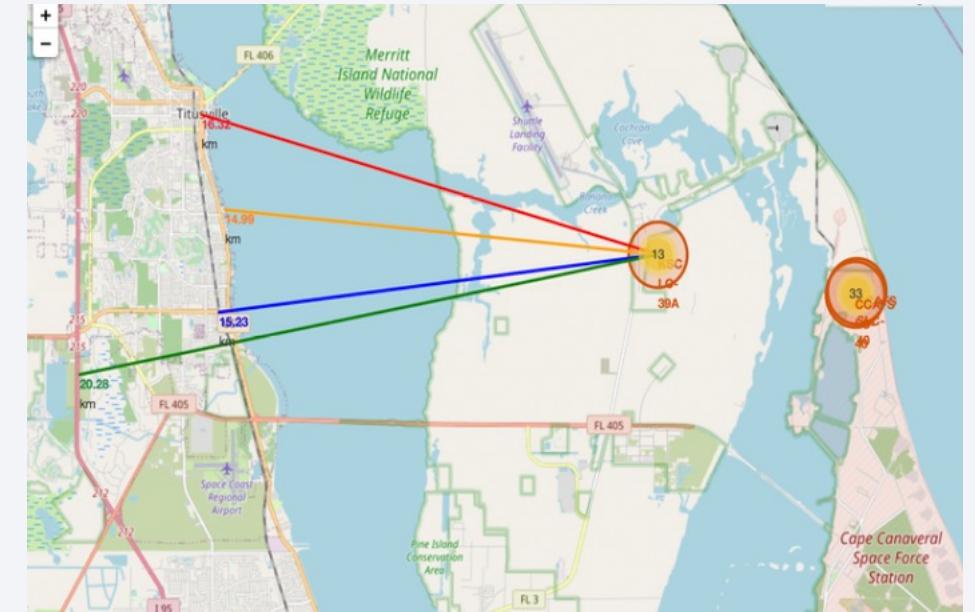
- From the color-labeled markers, we should be able to easily identify which launch sites have relatively high success rates.
 - **Green** Marker = Successful Launch
 - **Red** Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.



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

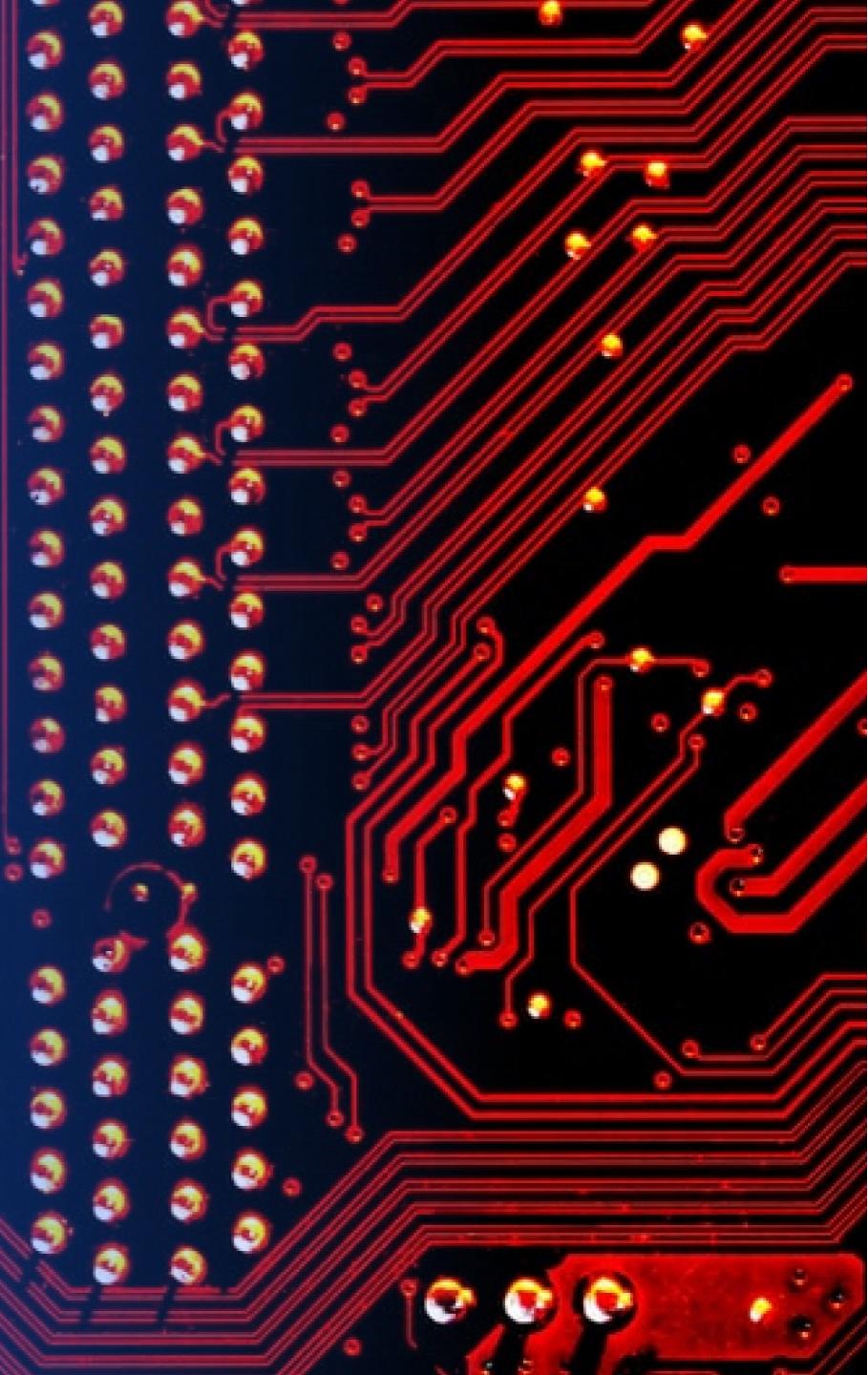
Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
 - relatively close to railway (15.23 km)
 - relatively close to highway (20.28 km)
 - relatively close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relatively 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 populated areas



Section 4

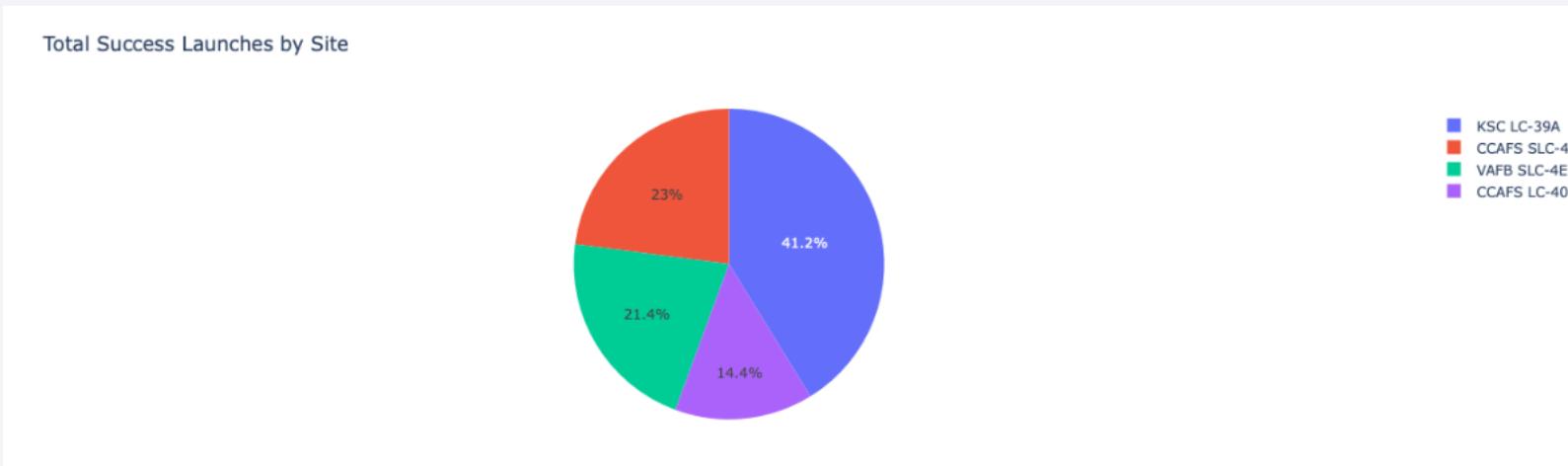
Build a Dashboard with Plotly Dash



Launch success count for all sites

Explanation:

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

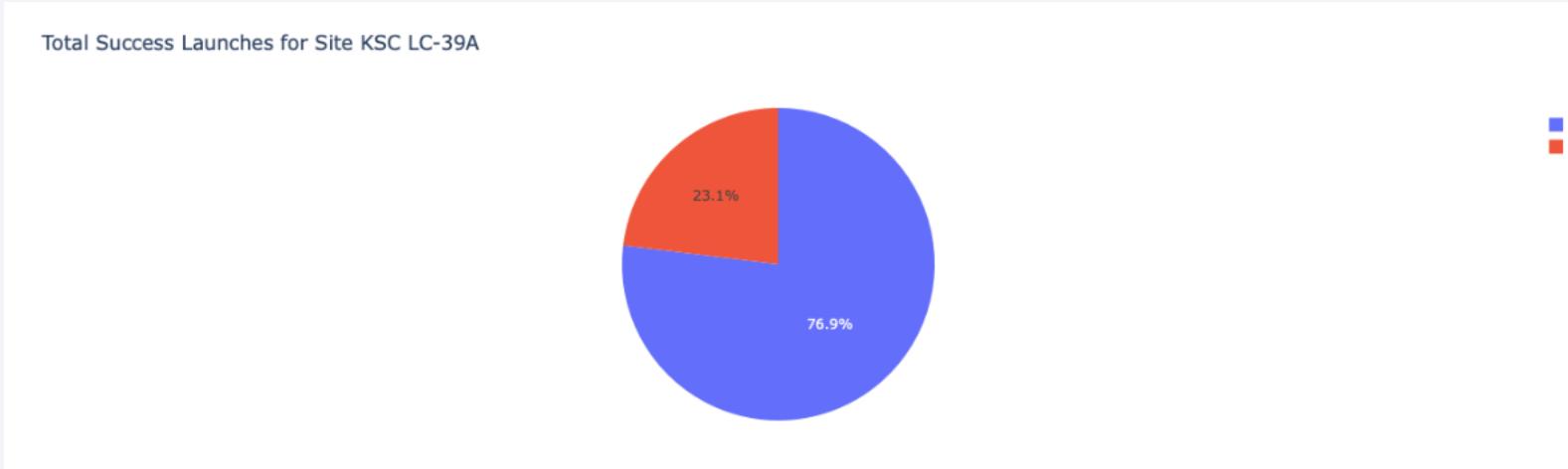


Launch site with highest launch success ratio

Explanation:

KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and

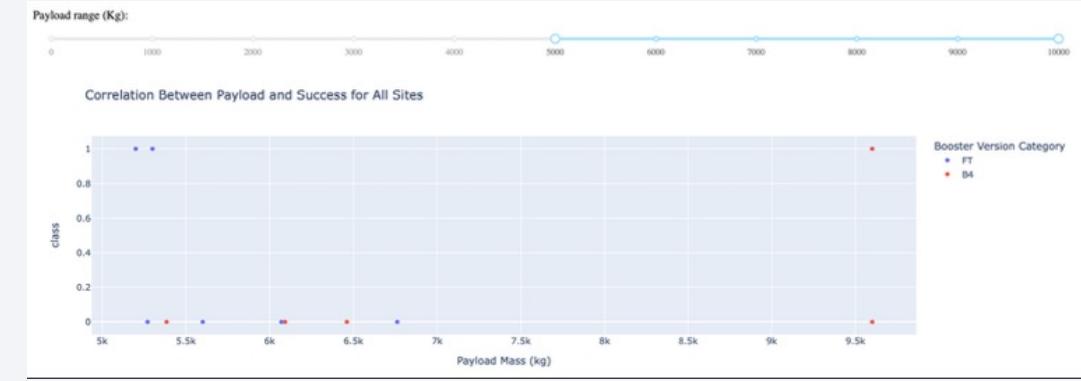
- only 3 failed landings.



Payload Mass vs. Launch Outcome for all sites

Explanation:

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



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- Based on the scores of the Test Set, we cannot confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested 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

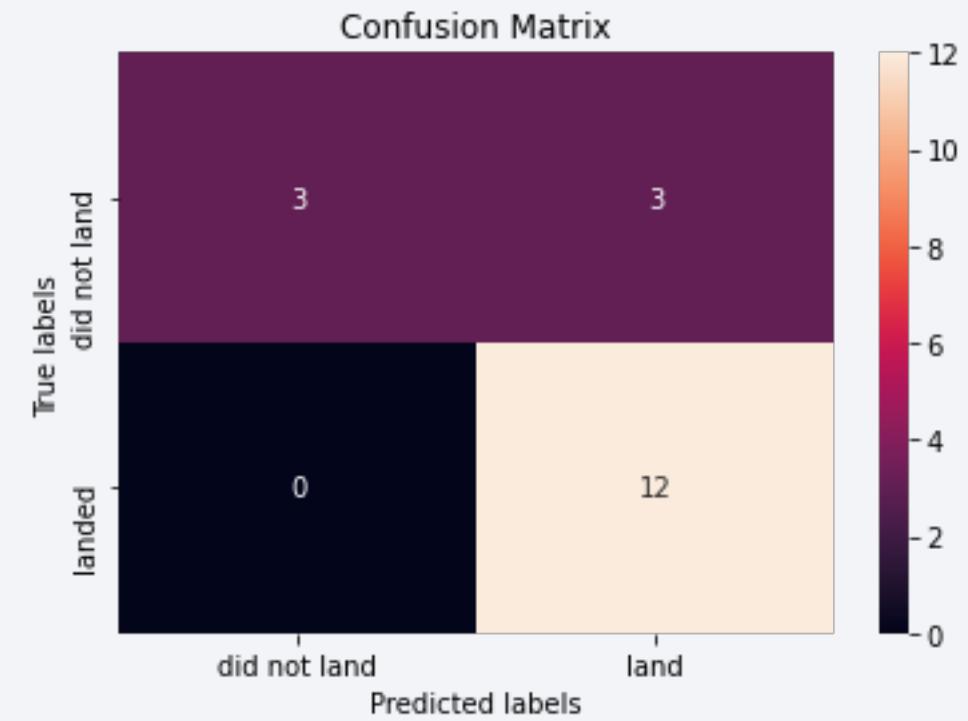
Scores and Accuracy of the Dataset

	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:

- We see that logistic regression can distinguish between the different classes and that the major problem is false positives.



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.
- 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

- See [Github repository](#) for Python code snippets, SQL queries, charts, Notebook outputs, or data sets that used during this project

Thank you!

