IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mikhail Petushok
4/28/2023

Outline

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

Executive Summary

In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.

- 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 has emerged as the top-performing company in the commercial space industry by revolutionizing space travel and making it more accessible. To promote its services, the company has advertised its Falcon 9 rocket launches on its website, offering them at a comparatively lower cost of $62 million, as opposed to other providers who charge upwards of $165 million per launch. This remarkable cost reduction is primarily attributed to SpaceX's ability to reuse the first stage of their rockets. Therefore, if we can accurately predict the first stage's successful landing, we can estimate the total cost of a launch. Utilizing publicly available information and machine learning models, we will endeavor to forecast whether or not SpaceX will be able to reuse the first stage.

- **Problems we want to find answers**

In what ways do factors such as payload mass, launch location, number of flights, and orbits impact the first stage landing success?

Is there a rising trend in the success rate of first stage landings over time?

Which binary classification algorithm would be most suitable for this scenario?

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

To conduct a comprehensive analysis of launches, we employed a combination of SpaceX REST API requests and web scraping to extract data from a table on SpaceX's Wikipedia page. We utilized both of these data collection methods to ensure we had a complete dataset for a more thorough analysis.

- The columns were obtained through the SpaceX REST API:

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

- The columns were acquired through web scraping Wikipedia:

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

[Data Collection]

Data Collection – SpaceX API

- The utilized API, located at https://api.spacexdata.com/v4/rockets/, supplies information regarding various rocket launches conducted by SpaceX. To focus solely on Falcon 9 launches, we filtered the data accordingly. Any incomplete data was replaced with the mean value of its respective column. The final dataset consisted of 90 rows or instances and 17 columns or features. The initial rows of the dataset are illustrated in the image below:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

SpaceX API calls

Data Collection - Scraping

- The data was obtained through web scraping from the following URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922. This website exclusively presents information regarding Falcon 9 launches. The resulting dataset comprised 121 rows or instances and 11 columns or features. The initial rows of the dataset are depicted in the image below:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

[Web scraping]

Data Wrangling

Within the dataset, there are numerous instances where the booster did not achieve a successful landing. In some cases, a landing was attempted but ultimately failed due to an accident. For example, the designation "True Ocean" indicates a successful landing in a specific region of the ocean, while "False Ocean" signifies an unsuccessful landing in a designated oceanic location. Similarly, "True RTLS" and "False RTLS" signify successful and unsuccessful landings, respectively, on a ground pad. "True ASDS" and "False ASDS" indicate successful and unsuccessful landings, respectively, on a drone ship. We transformed these outcomes into Training Labels, where "1" represented a successful booster landing and "0" represented an unsuccessful landing.

[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 depict the correlation between variables, and if a correlation is present, they can be utilized in a machine learning model. Bar charts compare discrete categories to illustrate the relationship between the compared categories and a measured value. Line charts demonstrate the patterns in data over time, specifically in a time series.

[Data Visualization]()

EDA with SQL

- Performed SQL queries
  - Showing a list of unique launch sites in space missions.
  - Displaying 5 records where the launch sites begin with the string 'CCA.'
  - Showing the total payload mass of boosters launched by NASA (CRS).
  - Displaying the average payload mass of booster version F9 v1.1.
  - Listing the date of the first successful landing outcome on a ground pad.
  - Listing the names of boosters that successfully landed on a drone ship and carried a payload mass greater than 4000 but less than 6000.
  - Listing the total number of successful and failed mission outcomes.
  - Listing the names of booster versions that have carried the maximum payload mass.
  - Listing the failed landing outcomes on a drone ship, along with their booster versions and launch site names, for the months in the year 2015.
  - Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates of 2010-06-04 and 2017-03-20 in descending order.

EDA with SQL

Build an Interactive Map with Folium

- Markers of all Launch Sites:
  - A Marker with a Circle, Popup Label, and Text Label was added to the NASA Johnson Space Center, using its latitude and longitude coordinates as the starting point to display its location.
  - Markers with Circles, Popup Labels, and Text Labels were added to all Launch Sites, using their latitude and longitude coordinates to display their geographic locations and their proximity to the Equator and coasts.
- Colored Markers of the launch outcomes for each Launch Site:
  - Markers have been added in different colors, green for success and red for failure, using a Marker Cluster technique. This is done to determine which launch sites have comparatively higher success rates.
- Distances between a Launch Site to its proximities:
  - Colored lines were added to display the distances between the launch site KSC LC-39A (as an example) and nearby features such as railways, highways, coastlines, and the closest city.

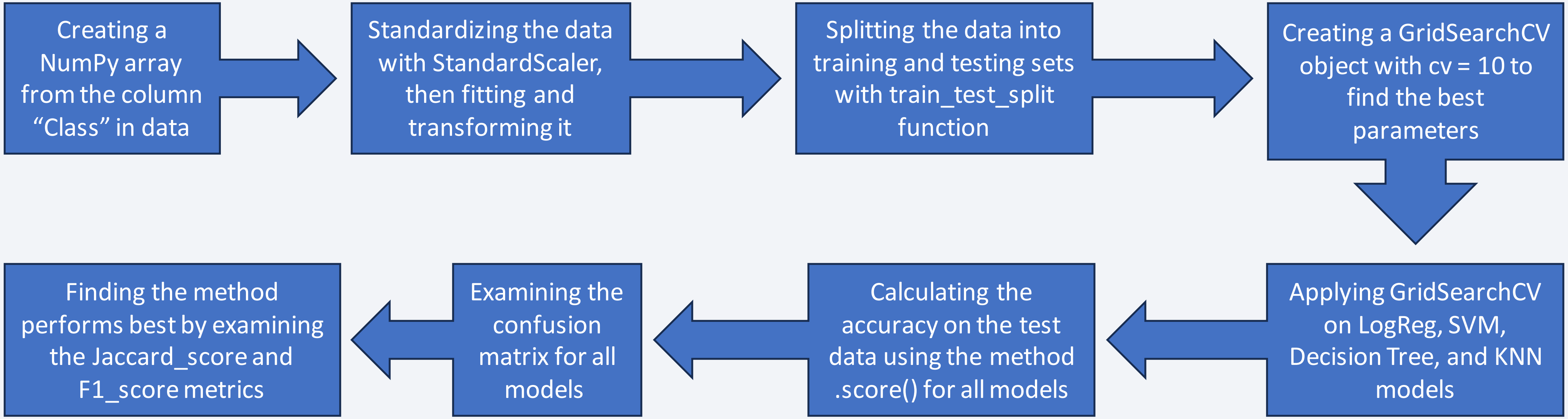
[Map with Folium]()

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.

Plotly Dash

Predictive Analysis (Classification)



- git

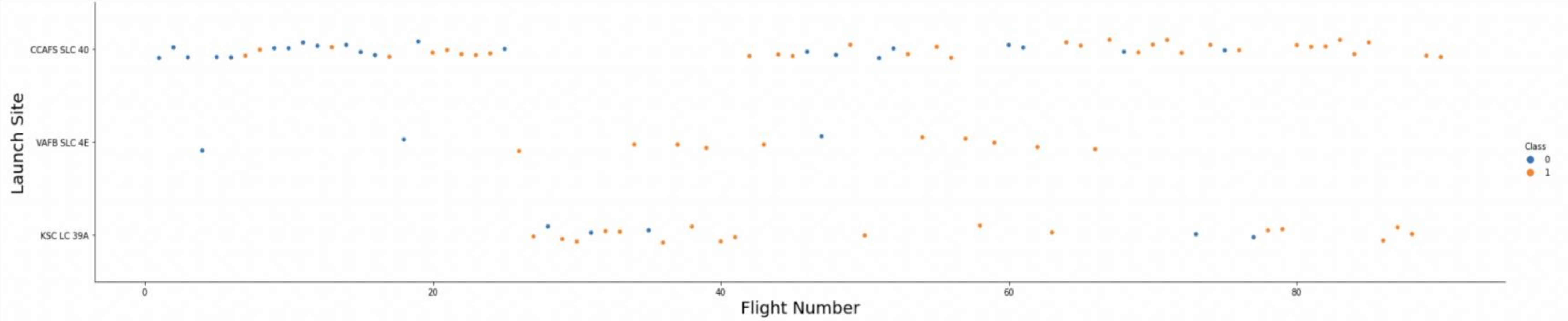
Results

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

Section 2

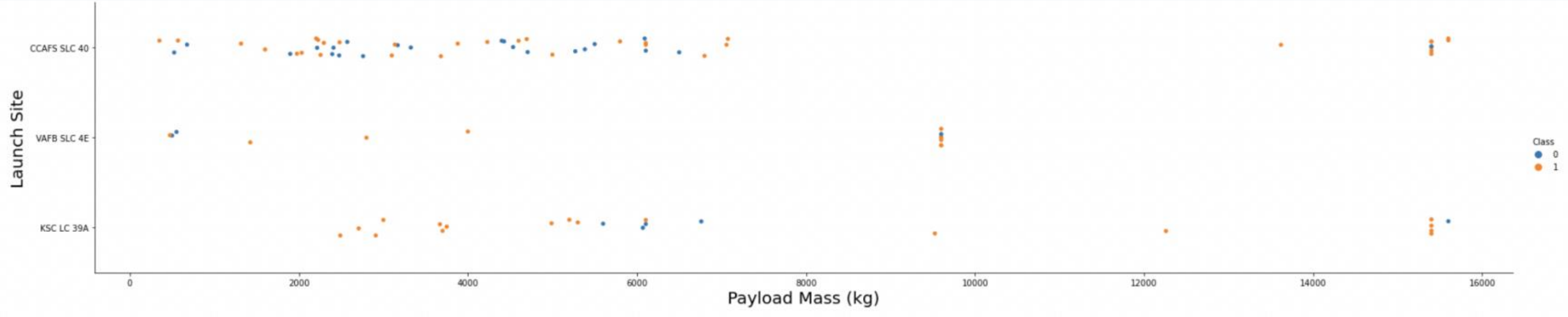
Insights drawn from EDA

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.

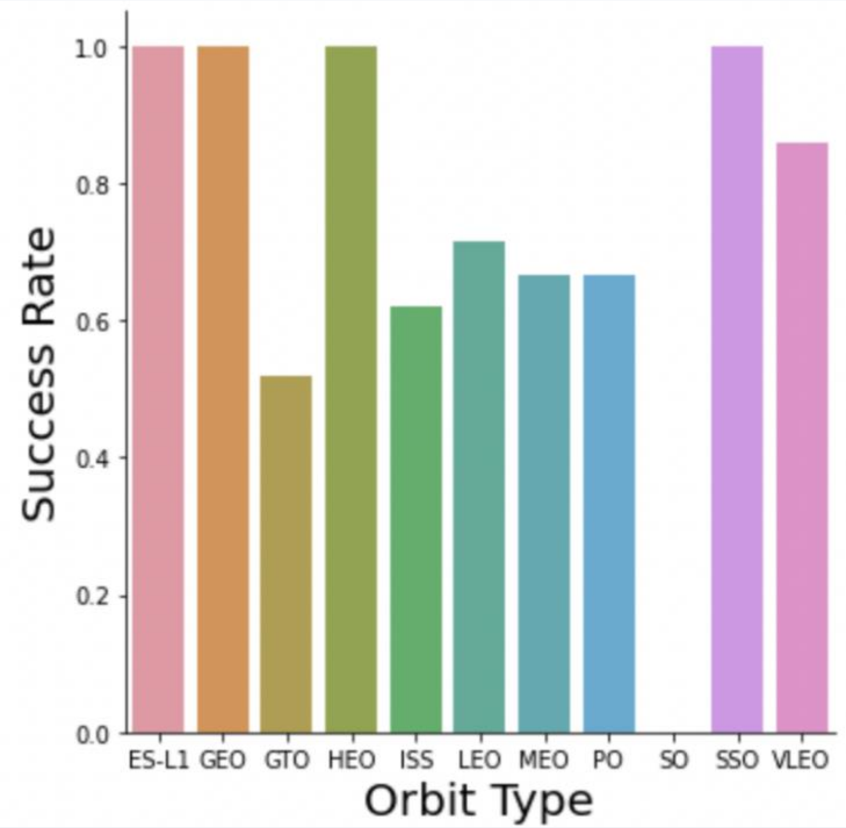
Success Rate vs. Orbit Type

- Explanation:
  - Orbits with 100% success rate:

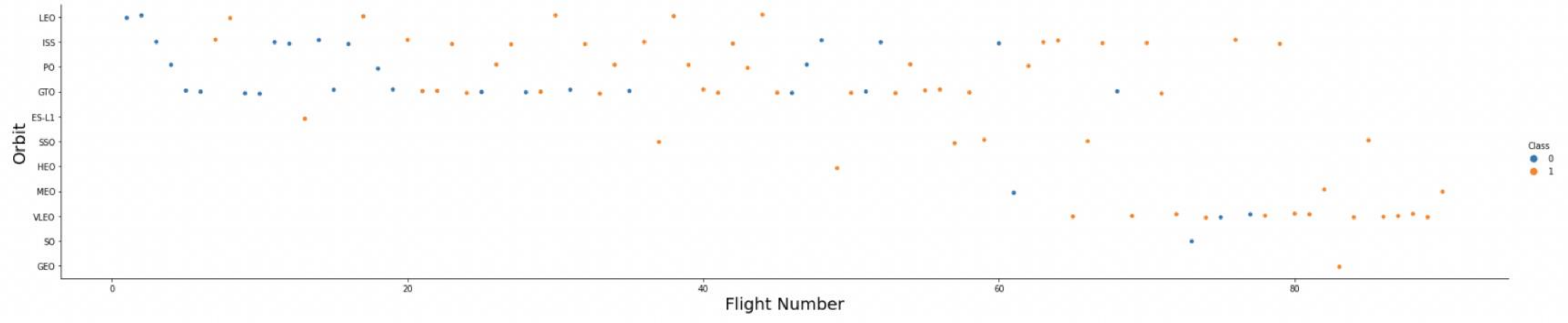
    - ES-L1, GEO, HEO, SSO
  - Orbits with 0% success rate:

    - SO
  - Orbits with success rate between 50% and 85%:

    - GTO, ISS, LEO, MEO, PO

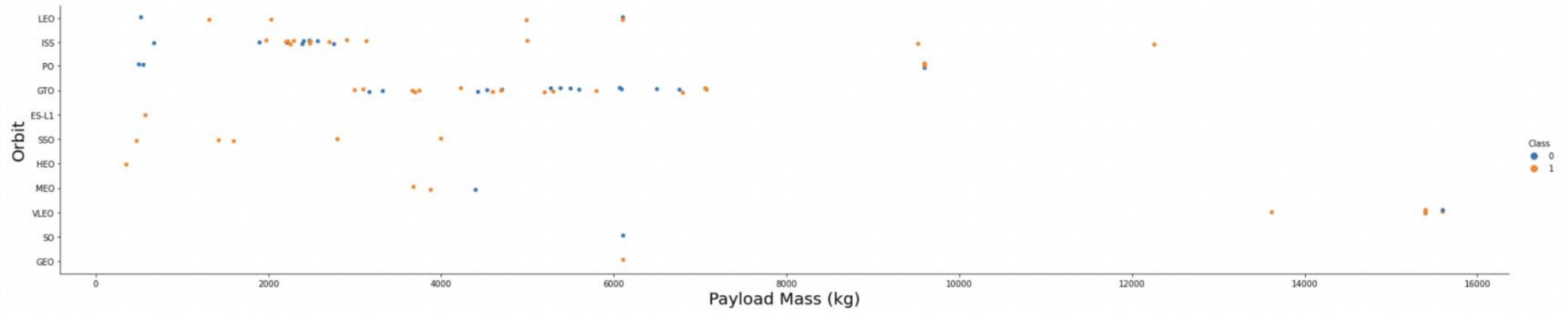


Flight Number vs. Orbit Type



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

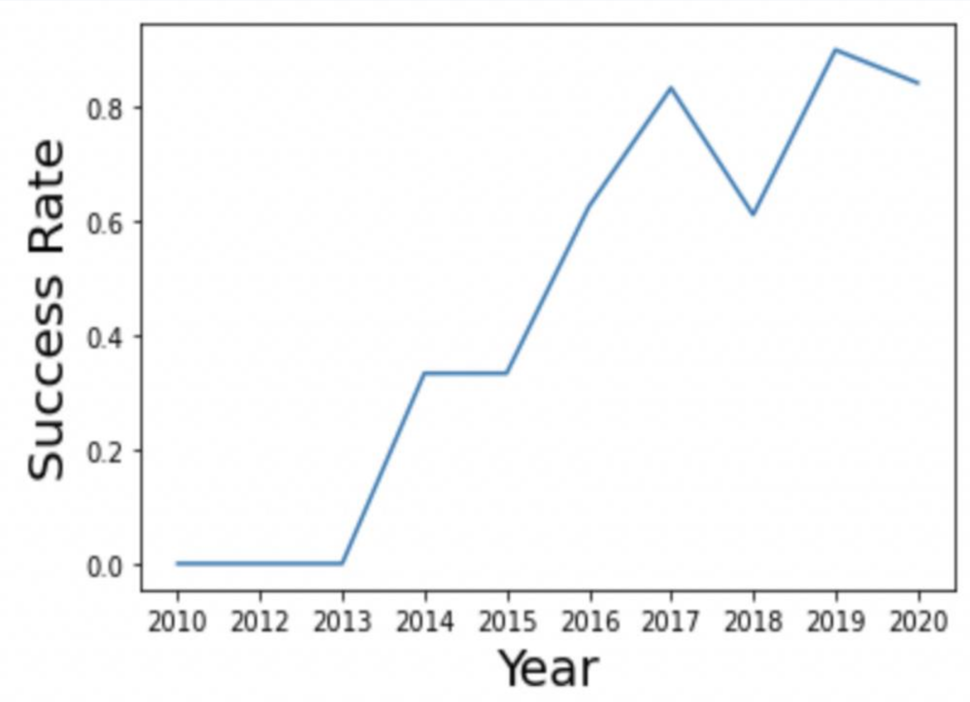
Payload vs. Orbit Type



- 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 since 2013 kept increasing till 2020.



All Launch Site Names

[8]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

[8]:

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

- Explanation:
  - Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

```
[11]: %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

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

[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Explanation:
  - Displaying 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)

```
[13]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'

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

[13]:

SUM(PAYLOAD_MASS__KG_)
45596

- Explanation:
  - Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

```
[15]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 V1.1%'

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

[15]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

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

First Successful Ground Landing Date

```
[13]: %sql SELECT min(Date) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success (ground pad)'

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

[13]:

min(Date)
01-05-2017

- Explanation:
  - Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[14]: %sql SELECT Booster_Version FROM SPACEXTBL where "Landing _Outcome" LIKE '%drone ship%' and PAYLOAD_MASS__KG_ between 4000 and 6000
 * sqlite:///my_data1.db
Done.
```

[14]:

Booster_Version
F9 FT B1020
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

```
[28]: %sql SELECT Mission_Outcome, COUNT(*) as num_outcomes FROM SPACEXTBL GROUP BY Mission_Outcome;

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

[28]:

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

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

Boosters Carried Maximum Payload

```
[31]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) from SPACEXTBL);

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

[31]:

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

2015 Launch Records

```
[18]: %sql SELECT substr(Date, 4, 2) AS month_name, date, "Landing _Outcome", Booster_Version, launch_Site FROM SPACEXTBL where "Landing _Outcome" Like '%Failure%' and substr(Date, 7, 4) = '2015';
 * sqlite:///my_data1.db
Done.
```

[18]:

month_name	Date	Landing _Outcome	Booster_Version	Launch_Site
01	10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	14-04-2015	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

```
[19]: %sql SELECT "Landing _Outcome", COUNT(*) as num_successes FROM SPACEXTBL where Date between '04-06-2010' AND '20-03-2017' GROUP BY "Landing _Outcome" ORDER BY num_successes DESC;

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

[19]:

Landing _Outcome	num_successes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

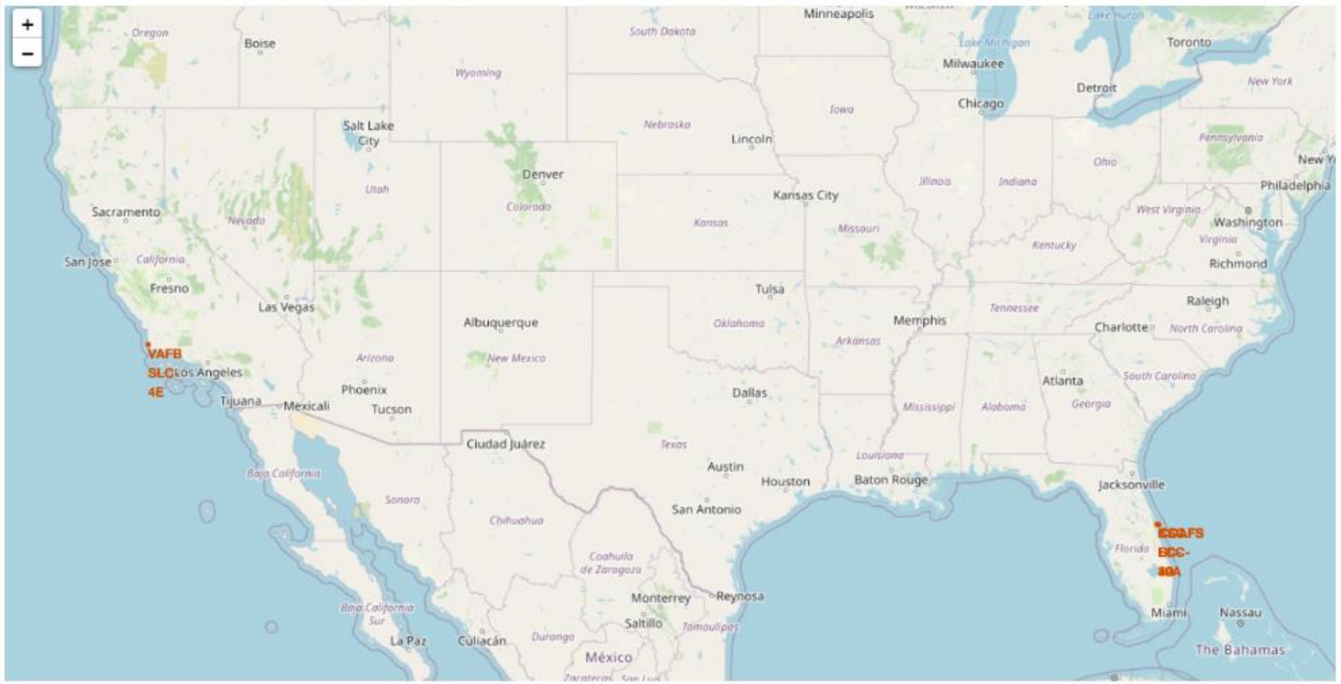
- Explanation:
    - Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

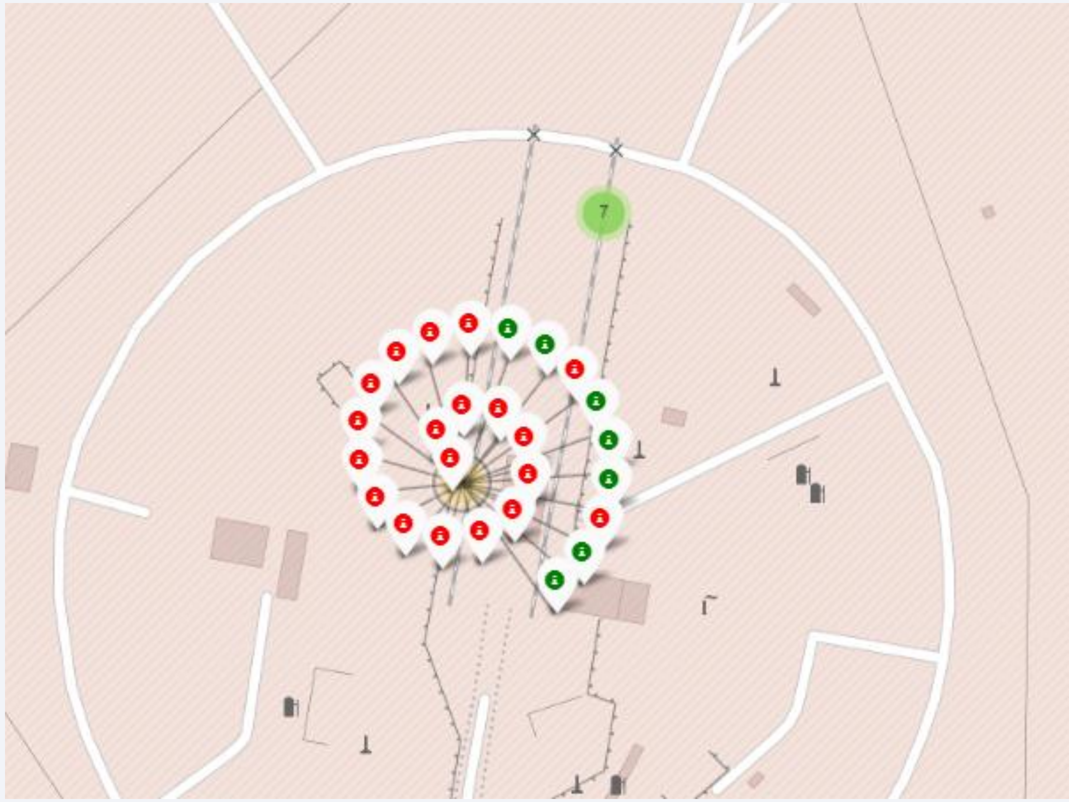
All launch sites' location markers on a global map

- Explanation:
  - Launch sites are typically located close to the Equator, where the land is moving faster than any other place on the Earth's surface. At the Equator, objects on the Earth's surface are already moving at a speed of 1670 km/hour. Therefore, when a spacecraft is launched from the Equator, it enters space while maintaining the same speed at which it was moving before launch due to inertia. This high speed is advantageous for the spacecraft to maintain its orbit around the Earth.
  - The launch sites are located very near to the coast, which reduces the risk of debris from the rocket falling or exploding near populated areas.



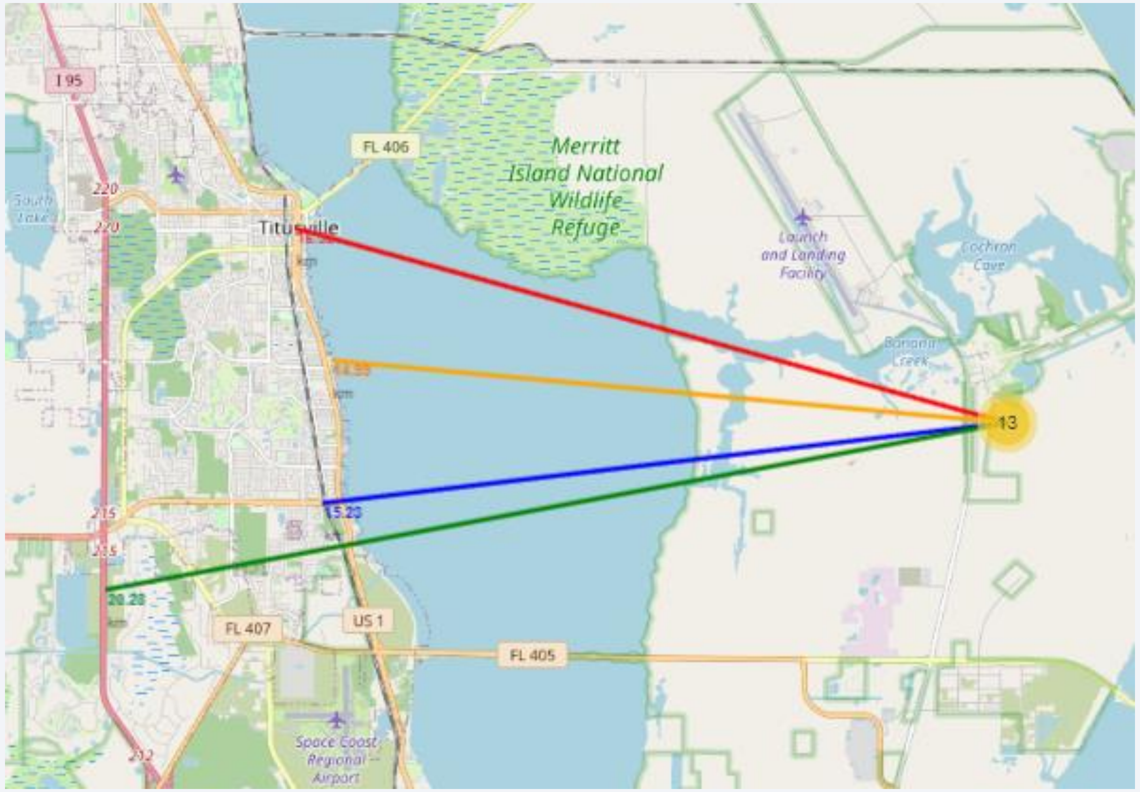
Color-labeled launch records on the map

- Explanation:
  - The markers with color-coded labels allow for easy identification of launch sites with relatively high success rates.
    - Green marker indicates successful launches
    - Red marker indicates failed launches.
  - Among these launch sites, KSC LC-39A stands out with an exceptionally high success rate.



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

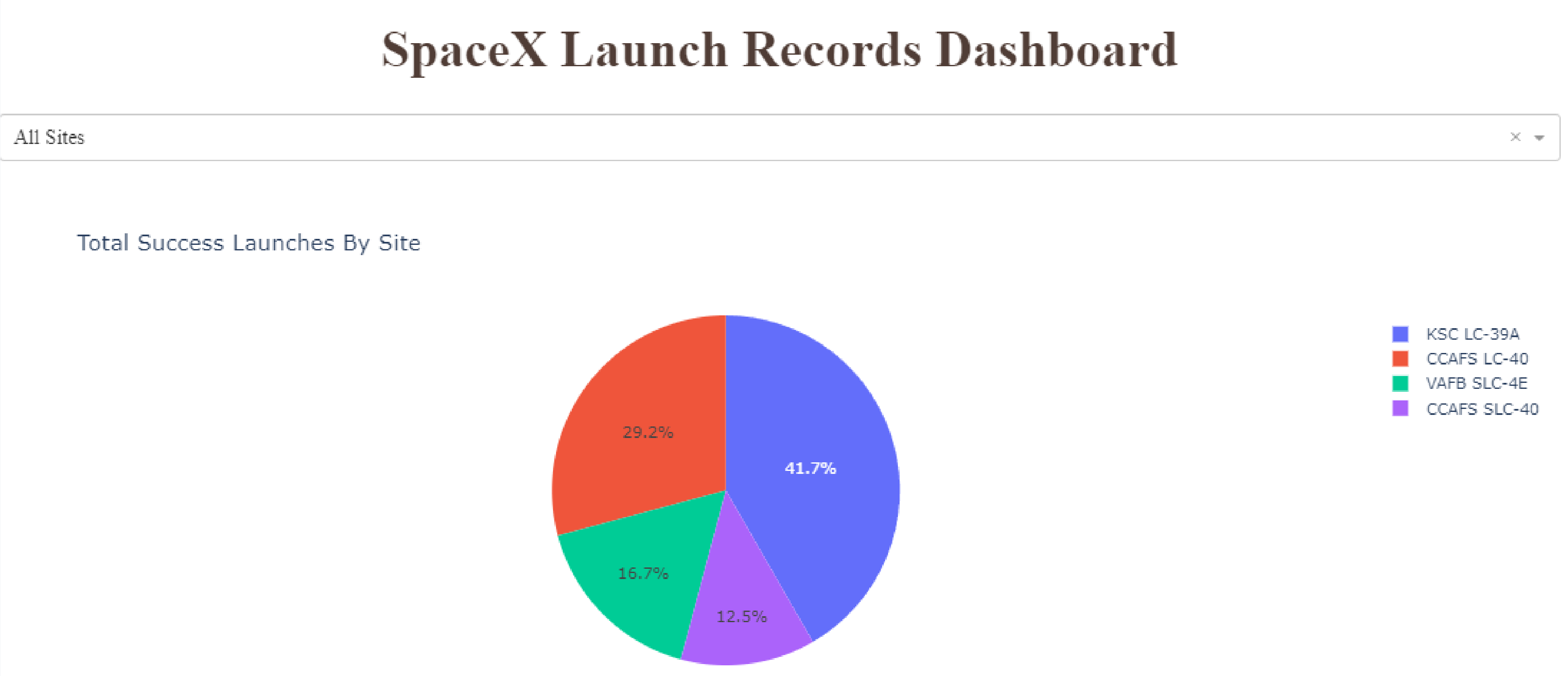
- Explanation:
  - The visual examination of the KSC LC-39A launch site reveals its proximity to various features:
    - railway (15.23 km)
    - highway (20.28 km)
    - coastline (14.99 km)
  - Additionally, the launch site is relatively close to the nearby city of Titusville (16.32 km).
  - Considering the high speed of a failed rocket, it can travel a distance of 15-20 km within a few seconds, which poses a potential risk to densely populated areas.



Section 4

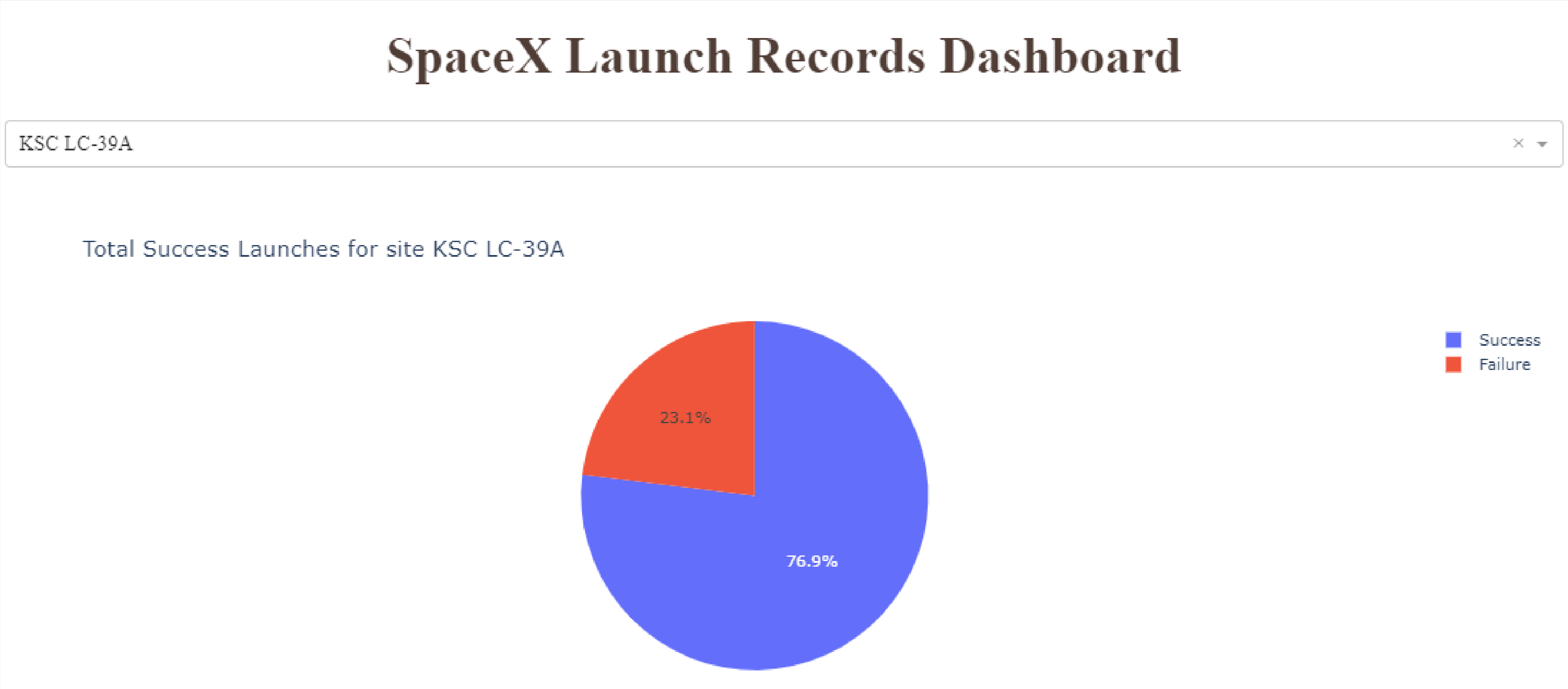
Build a Dashboard with Plotly Dash

Launch success count for all sites



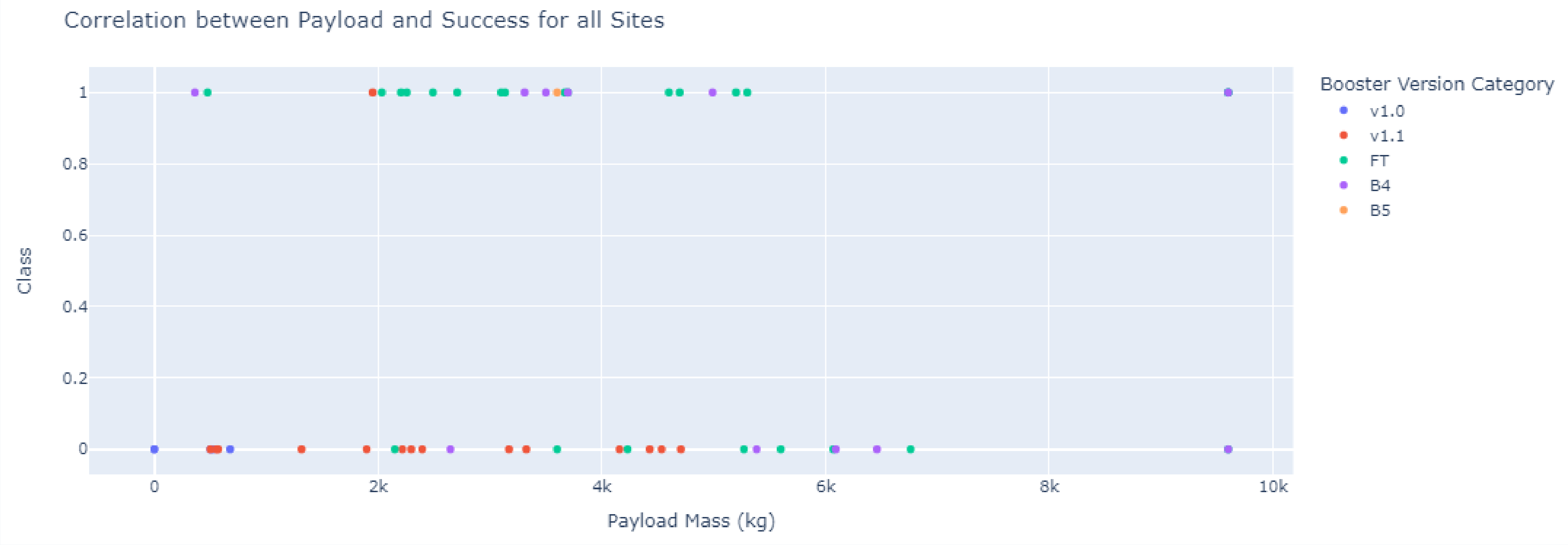
- Explanation:
  - The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch site with highest launch success ratio



- Explanation:
  - KSC LC-39A has the highest launch success rate (76.9%).

Payload Mass vs. Launch Outcome for all sites



- Explanation:
  - The charts show that payloads between 1900 and 4000 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

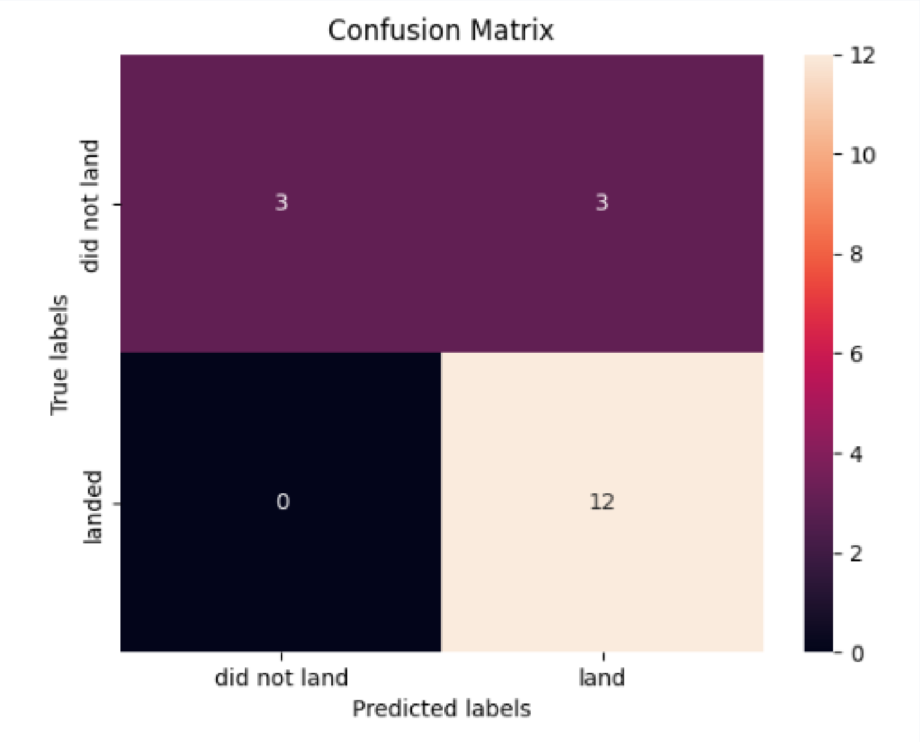
- Explanation:
  - The Test Set scores do not provide sufficient evidence to determine which method performs the best.
  - Since the test sample size was small (only 18 samples), it is possible that the Same Test Set scores were obtained by chance. To address this issue, we evaluated all methods using the entire dataset.
  - Based on the scores obtained from the entire dataset, it is evident that the Decision Tree Model outperforms the other models. This model not only exhibits higher scores but also the highest level of accuracy.

Accuracy of the Data Sets

	logReg	SVM	Tree	KNN
Test Data Set	0.83333	0.83333	0.83333	0.83333
Entire Data Set	0.84643	0.84821	0.90179	0.84821

Confusion Matrix

- Explanation:
  - Upon analyzing the confusion matrix, it becomes evident that logistic regression is capable of distinguishing between the various classes. However, the major issue lies with the occurrence of false positives.



Conclusions

- The optimal algorithm for this dataset is the Decision Tree Model.
- Launches with lower payload masses yield superior outcomes compared to those with larger payloads.
- The majority of launch sites are located near the Equator and are situated very close to the coastline.
- Over time, the success rate of launches has improved.
- Among all the launch sites, KSC LC-39A has the highest success rate.
- Orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.

Appendix

Special Thanks to:

Coursera

IBM

Thank you!