



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hasan Hammad
09, September, 2023



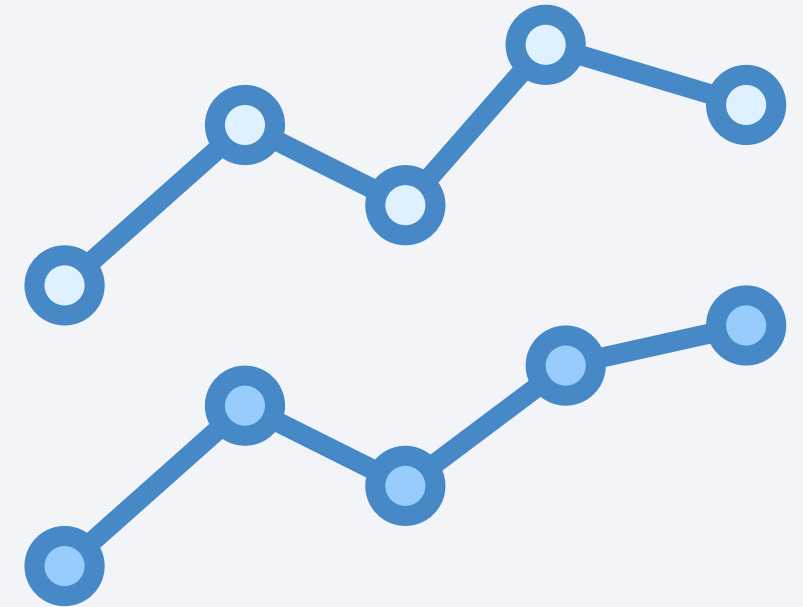
Outline

- 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)



Executive Summary

- Summary of all results
 - ✓ Exploratory Data Analysis results
 - ✓ Interactive analytics demo in screenshots
 - ✓ Predictive analysis results



Introduction

- Project background and context

SpaceX stands out as the preeminent achiever in the era of commercial space exploration, playing a pivotal role in democratizing space travel. On its website, the company showcases **Falcon 9 rocket** launches, priced at a competitive **\$62 million per launch**, a stark contrast to other providers whose charges exceed a staggering **\$165 million for each mission**. A significant portion of this cost disparity stems from SpaceX's pioneering ability to recover and reuse the first stage of its rockets. Consequently, the ability to ascertain the successful landing of **the first stage** serves as a key determinant for **launch cost estimation**. Leveraging publicly available data and advanced **machine learning algorithms**, we endeavor to forecast the likelihood of **SpaceX reusing the first stage in future missions**.

Introduction

- 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:
 - Describe how data was collected
 - Using SpaceX Rest API
 - Using Web Scrapping from Wikipedia
- Perform data wrangling
 - Describe how data was processed
 - Filtering the data
 - Dealing with missing values
 - Using One Hot Encoding to prepare the data to a binary classification



Methodology

Executive Summary

- 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](#) encompassed a dual approach, utilizing both [API requests](#) from [SpaceX's REST API](#) and [web scraping techniques](#) to extract data from a table within SpaceX's [Wikipedia](#) entry. This combined methodology was essential to obtain comprehensive information about their launches, enabling a more thorough and detailed analysis.

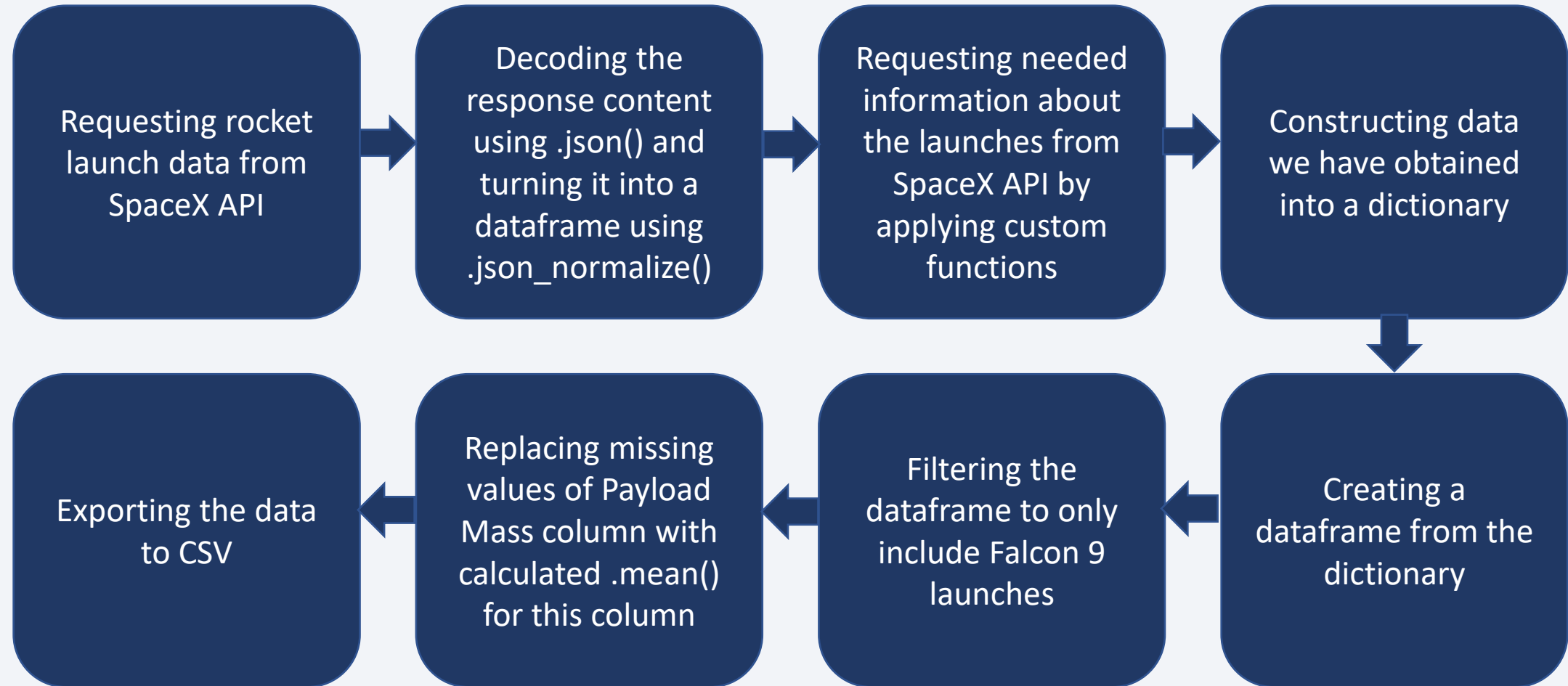
➤ Data Columns are obtained by using SpaceX REST API:

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

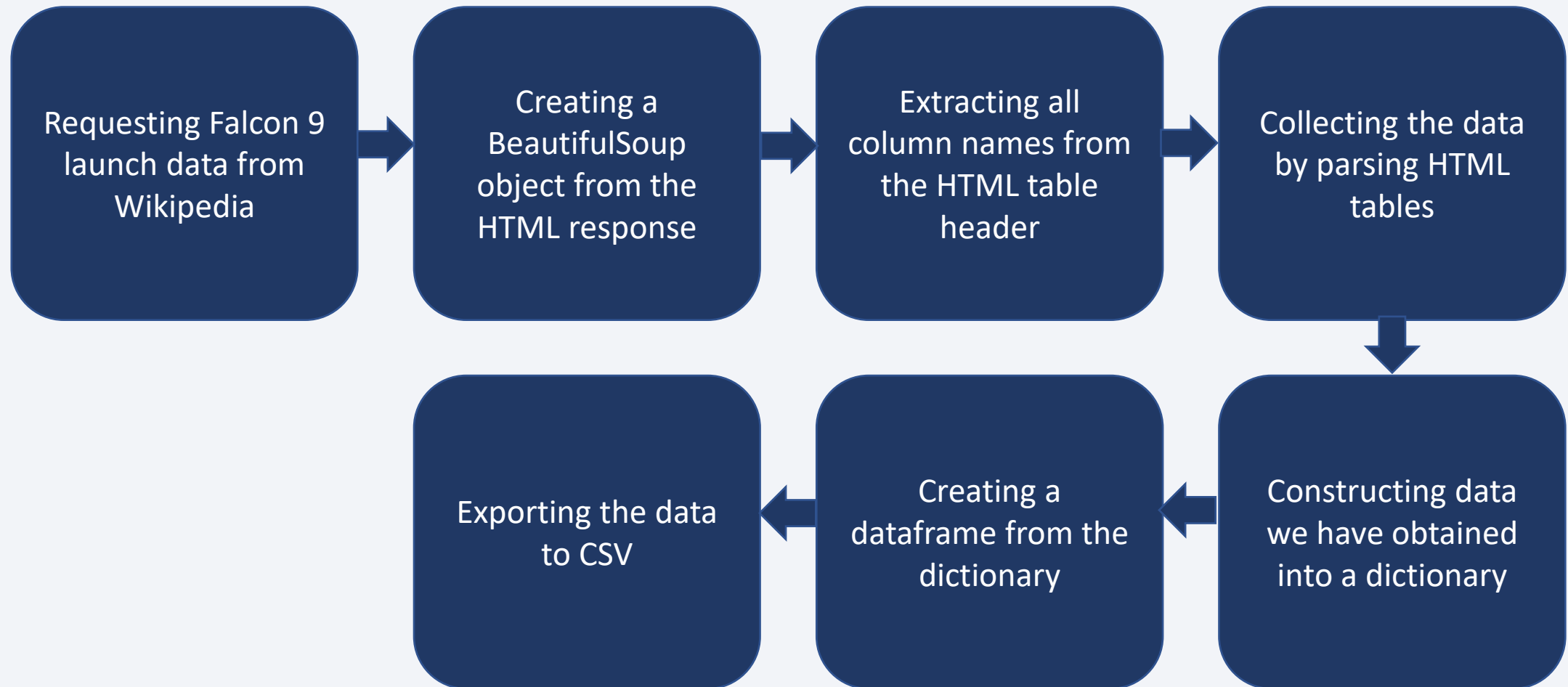
➤ Data Columns are obtained by using Wikipedia Web Scraping:

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

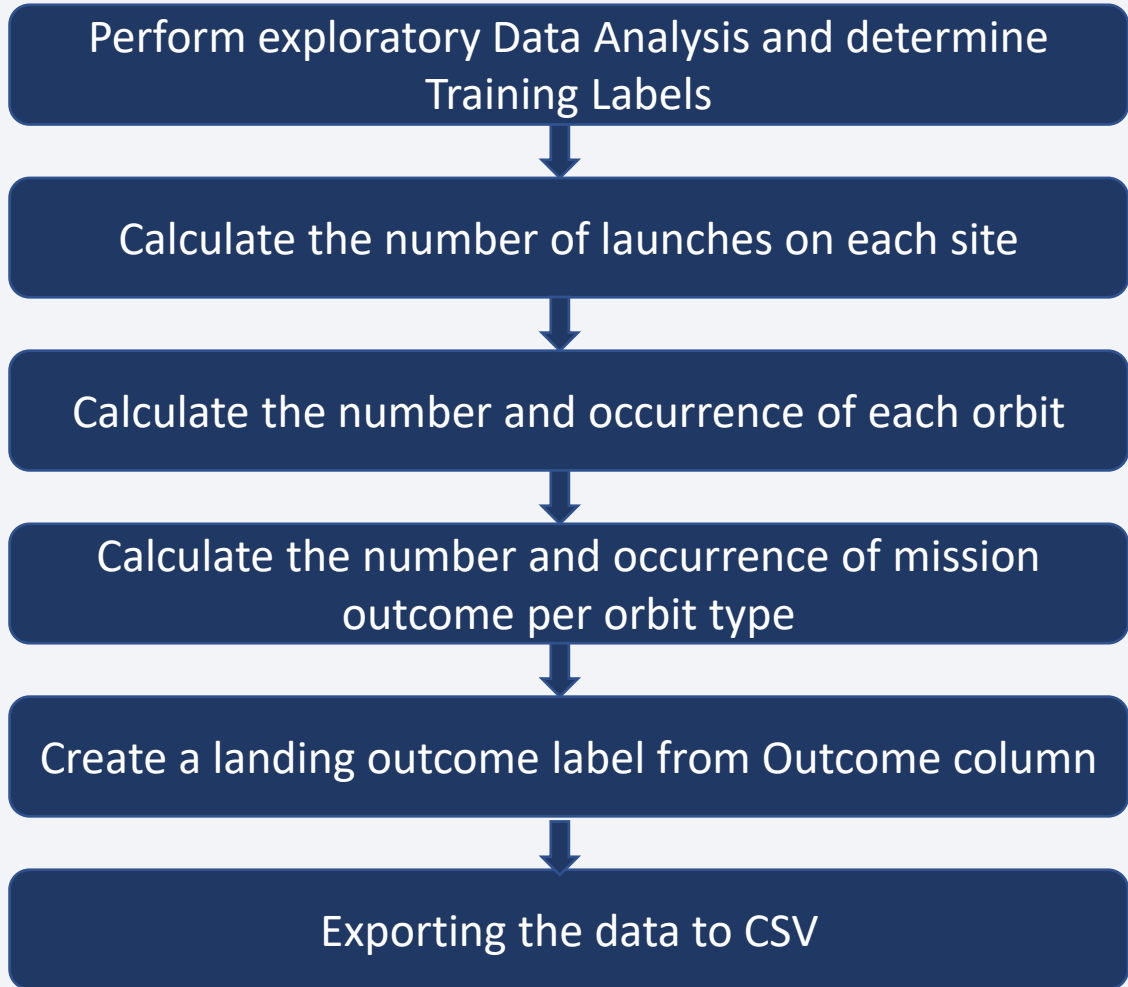
Data Collection – SpaceX API [GitHub URL: Data Collection API](#)



Data Collection - Scraping [GitHub URL: Web Scraping Notebook](#)



Data Wrangling [GitHub URL: Data Wrangling](#)



In the dataset, there are various scenarios where the booster's landing was **not successful**. These instances can be **categorized** based on different criteria. For instance, "**True Ocean**" indicates a **successful landing** in a specific oceanic region, while "**False Ocean**" signifies an **unsuccessful landing** in the ocean. Similarly, "**True RTLS**" denotes a **successful landing** on a ground pad, whereas "**False RTLS**" indicates an **unsuccessful ground pad landing**. Lastly, "**True ASDS**" represents a **successful landing** on a drone ship, while "**False ASDS**" corresponds to an **unsuccessful drone ship landing**.

To streamline these outcomes for analysis, we have translated them into training labels. A label of "**1**" signifies a **successful landing**, while "**0**" denotes an **unsuccessful landing**.

We generated a series of charts to visually represent the data:

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, Success Rate Yearly Trend

- **Scatter plots** within these charts help **uncover potential relationships between variables**, which could be invaluable in constructing machine learning models.
- **Bar charts** were also employed to facilitate **comparisons between discrete categories**. These charts aim to elucidate the connection between specific categories being compared and the corresponding measured values.
- Furthermore, **line charts** were utilized to track data trends over time, providing a time-series perspective on the information.

Conducted SQL queries encompassing various aspects of the space mission data:

- Retrieving the **names of unique launch sites** involved in the space mission.
- Displaying **five records** where launch sites commence with the string '**CCA**.'
- **Calculating the total payload mass** transported by boosters launched under NASA's CRS program.
- Determining the **average payload mass** carried by boosters of the F9 v1.1 version.
- Listing the **date when the initial successful ground pad landing** outcome was achieved.
- Enumerating the **names** of boosters that achieved success on drone ships, with payload masses ranging from **4000 to 6000**.
- Providing a tally of **the total number** of **successful** and **failed** mission **outcomes**.
- Identifying the booster versions responsible for transporting **the maximum payload mass**.
- Listing the **unsuccessful** landing outcomes on drone ships, including the booster versions and launch site names, specifically for the months within the year **2015**.
- Ranking the count of landing outcomes **Failure or Success** between the dates **2010-06-04 and 2017-03-15** **20** in descending order.

Build an Interactive Map with Folium [GitHub URL: Folium Map](#)

Summary of Map Markers and Distances Visualization:

- **Markers for All Launch Sites:** We incorporated markers on the map to represent various launch sites. Specifically, we added markers for the [NASA Johnson Space Center](#) with [circle indicators](#), [popup labels](#), and [text labels](#), using its latitude and longitude coordinates as the starting point. Additionally, we included markers for [all other launch sites](#), each marked with [circles](#), [popup labels](#), and [text labels](#), highlighting their geographical positions in relation to the Equator and nearby coastlines.
- **Colored Launch Outcome Markers:** To provide insights into launch outcomes, we introduced [colored markers](#). Successful launches are denoted by [green markers](#), while unsuccessful ones are [marked in red](#). The use of Marker Clusters aids in identifying launch sites with higher success rates based on the clustering of green markers.
- **Distance Visualization:** We enhanced the visualization by adding [colored lines](#) that display [distances](#) between the [KSC LC-39A](#) launch site (as an example) and its proximity points, such as [railways](#), [highways](#), [coastlines](#), and the [nearest city](#). These lines help viewers understand the spatial relationships between the launch site and its surroundings.

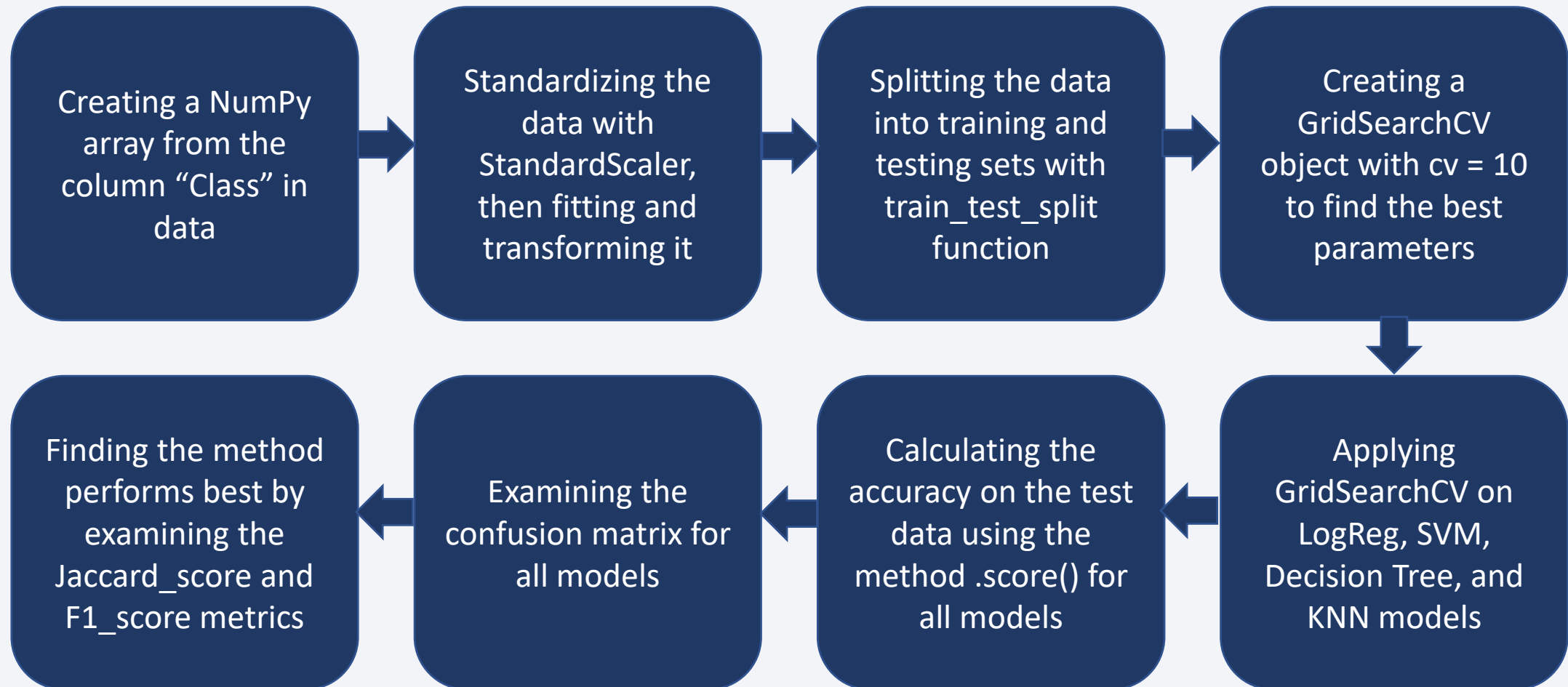
Build a Dashboard with Plotly Dash [GitHub URL: Dashboard](#)

Summary of the Dashboard

- **Launch Sites Dropdown List:** Implemented a convenient dropdown list feature, allowing users to select their preferred **launch site effortlessly**.
- **Pie Chart for Launch Success:** Introduced an informative **pie chart** to visualize the overall count of **successful launches across all sites**. Moreover, if users opt for a specific launch site, the chart **dynamically displays** the success versus failure counts for that particular site, providing valuable insights at a glance.
- **Payload Mass Range Slider:** Incorporated a **user-friendly** slider control that enables users to specify the desired **payload mass range**, enhancing the precision of data exploration.
- **Scatter Chart for Payload vs. Success Rate:** Introduced a **scatter chart** to depict the relationship between **payload mass** and **launch success rate** across different booster versions. This visualization aids in identifying potential correlations and patterns within the data.

Predictive Analysis (Classification)

[GitHub URL: Classification](#)



Results

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

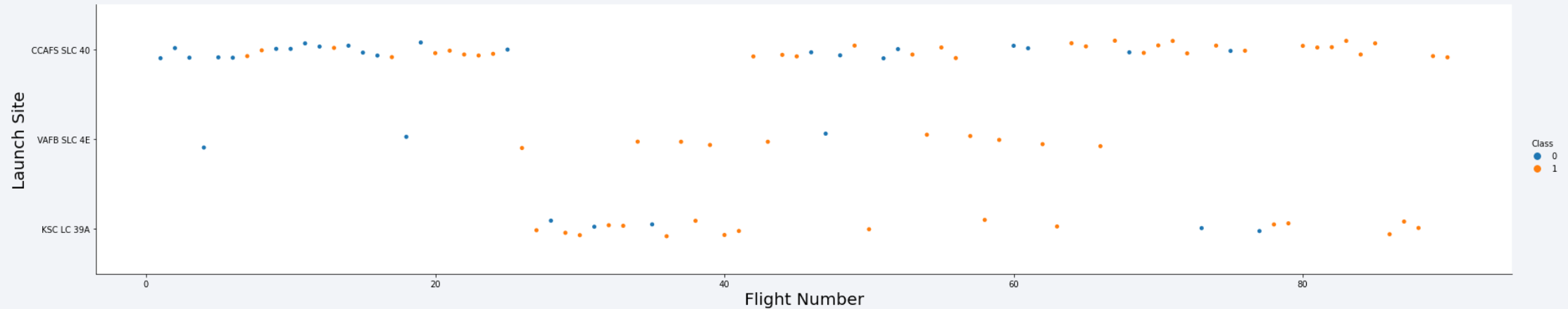


The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

Insights drawn from EDA

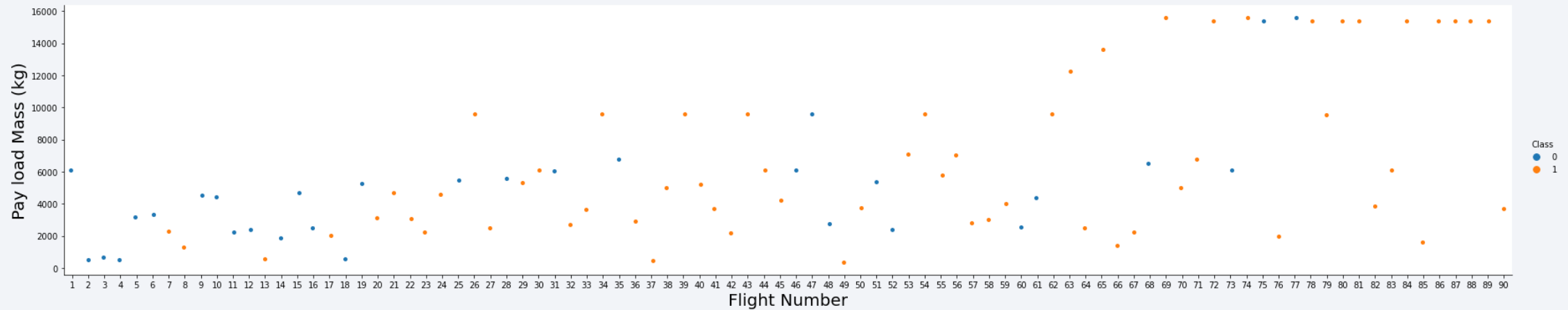
Flight Number vs. Launch Site



Explanation:

- The initial flights in the dataset experienced a series of **failures**, while the most recent flights achieved **success**. This suggests an improvement in mission outcomes over time.
- The CCAFS SLC 40 launch site accounts for approximately half of all launches, indicating its **significance** in the **space mission program**.
- Two specific launch sites, **VAFB SLC 4E** and **KSC LC 39A**, stand out with notably higher success rates compared to others. This underscores their effectiveness and reliability.
- There's a plausible **assumption** that each new launch in the dataset is associated with a **higher likelihood of success**, given the historical progression from **failures to successes**.

Payload vs. Launch Site



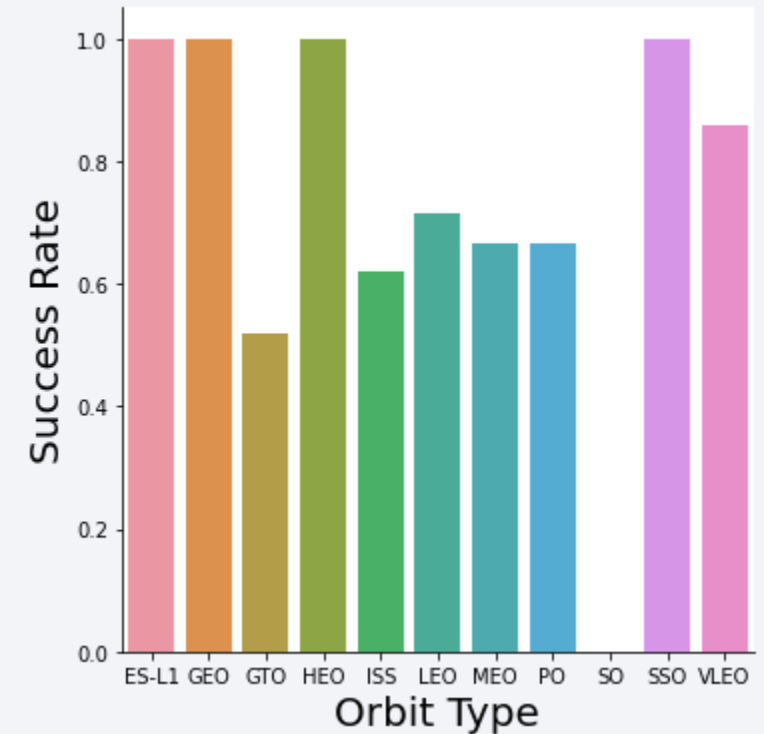
Explanation:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 8000 kg were successful.

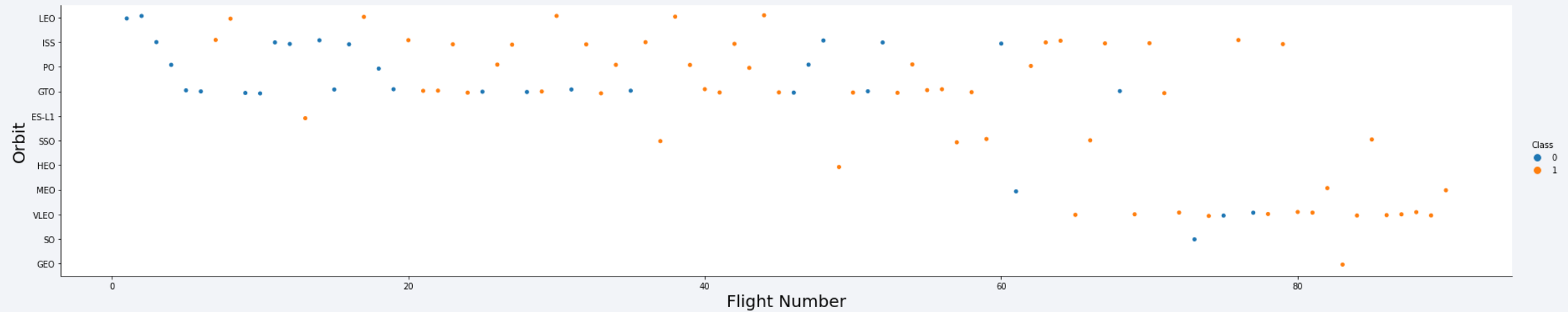
Success Rate vs. Orbit Type

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, VLEO



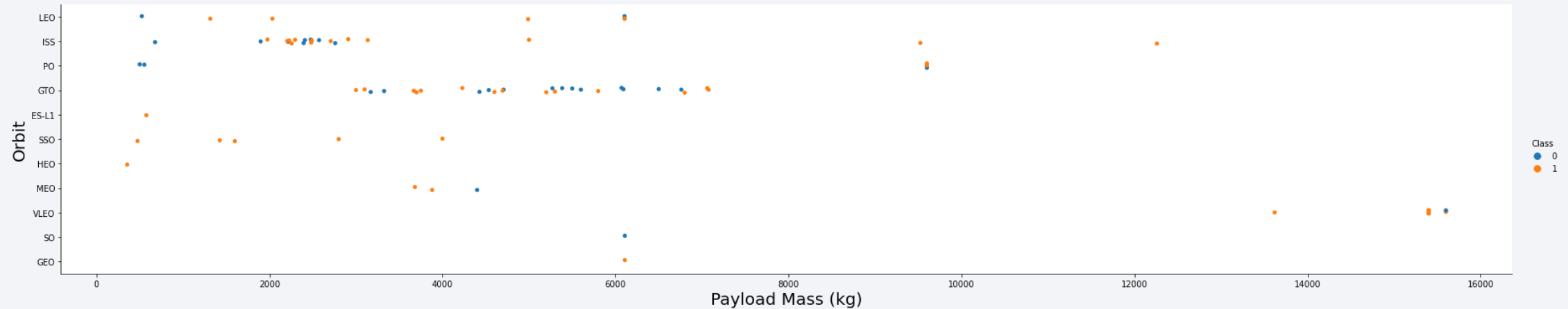
Flight Number vs. Orbit Type



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 and orbit type when the orbit is GTO orbit.

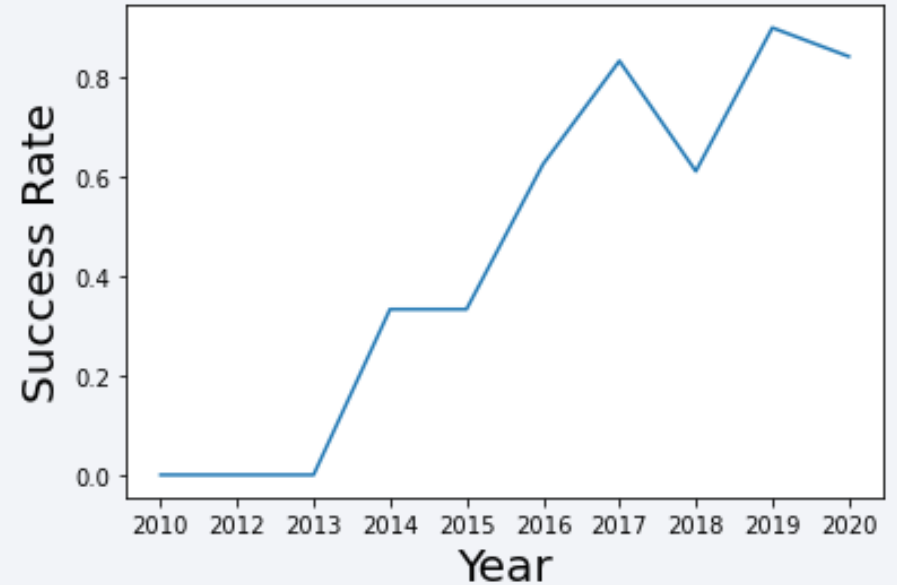
Payload vs. Orbit Type



Explanation:

- Payload mass seems to correlate with orbit LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend



Explanation:

- Success generally **increases** over time since **2013** with a slight dip in **2018**. Success in recent years is at around 80%.

All Launch Site Names

```
%sql select distinct launch_site from SPACEXDATASET;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

launch_site

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

Explanation:

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

Launch Site Names Begin with 'CCA'

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

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/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

Explanation:

- Displaying 5 records where launch sites begin with “CCA”

Total Payload Mass

```
%sql select sum(payload_mass_kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

total_payload_mass

45596

Explanation:

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

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

average_payload_mass

2534

Explanation:

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

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

first_successful_landing

2015-12-22

Explanation:

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

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_ betw  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

booster_version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Explanation:

- Displaying 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

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb  
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

- Displaying the **total number** of **successful** and **failure** mission outcomes.

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASE
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/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

Explanation:

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

2015 Launch Records

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

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:

- Displaying 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

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

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

landing__outcome	count_outcomes
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

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.

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

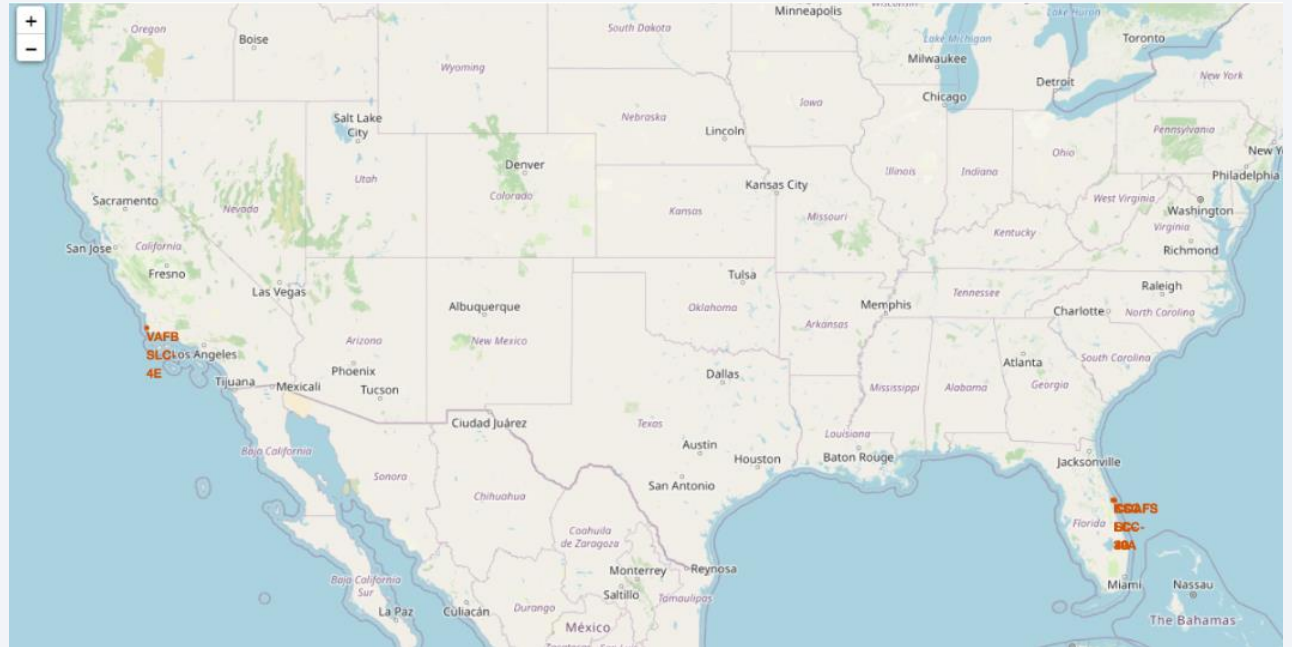
Section 3

Launch Sites Proximities Analysis

Launch Site Locations

Explanation:

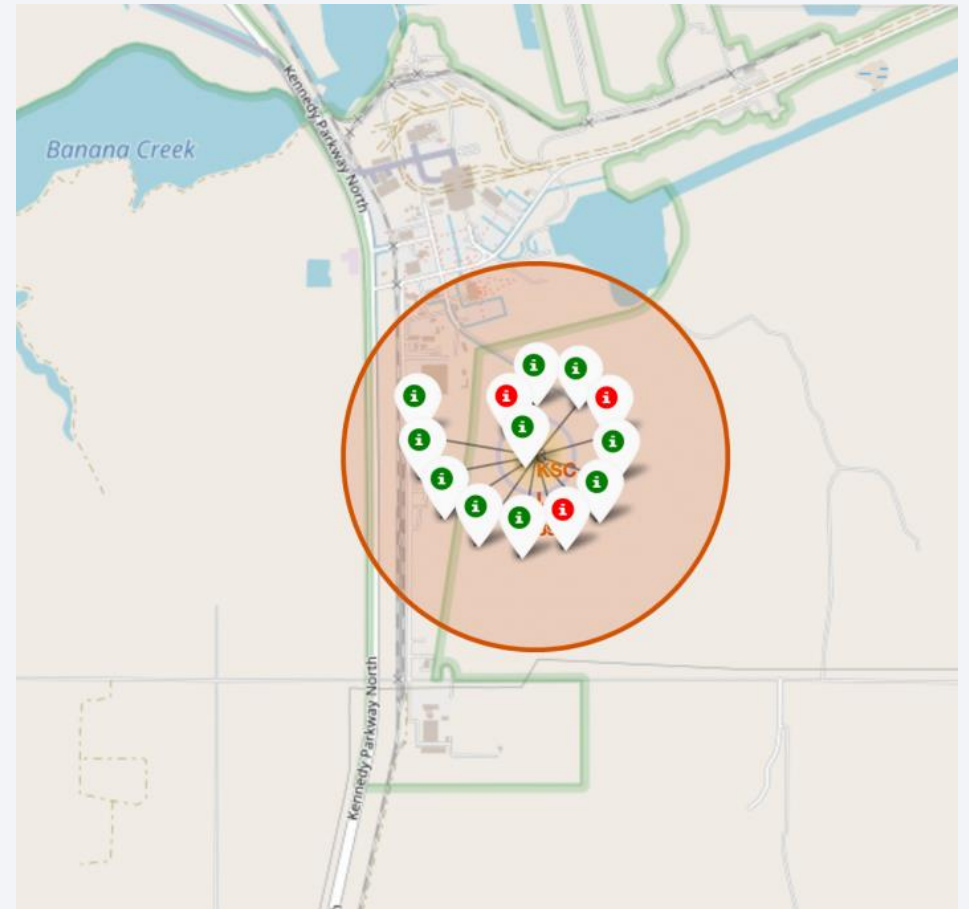
- Most launch sites are situated **near the Equator** due to the region's faster **rotational speed**. When a spacecraft is launched from this location, it not only ascends into space but also retains its **pre-launch velocity**. This velocity is crucial for maintaining the necessary speed to stay in orbit.
- Furthermore, all launch sites are strategically located in **close proximity to coastlines**. Launching rockets towards the **ocean** serves to **minimize the potential danger** of **debris falling** or **exploding** near populated areas.



Color-labeled Launch Markers

Explanation:

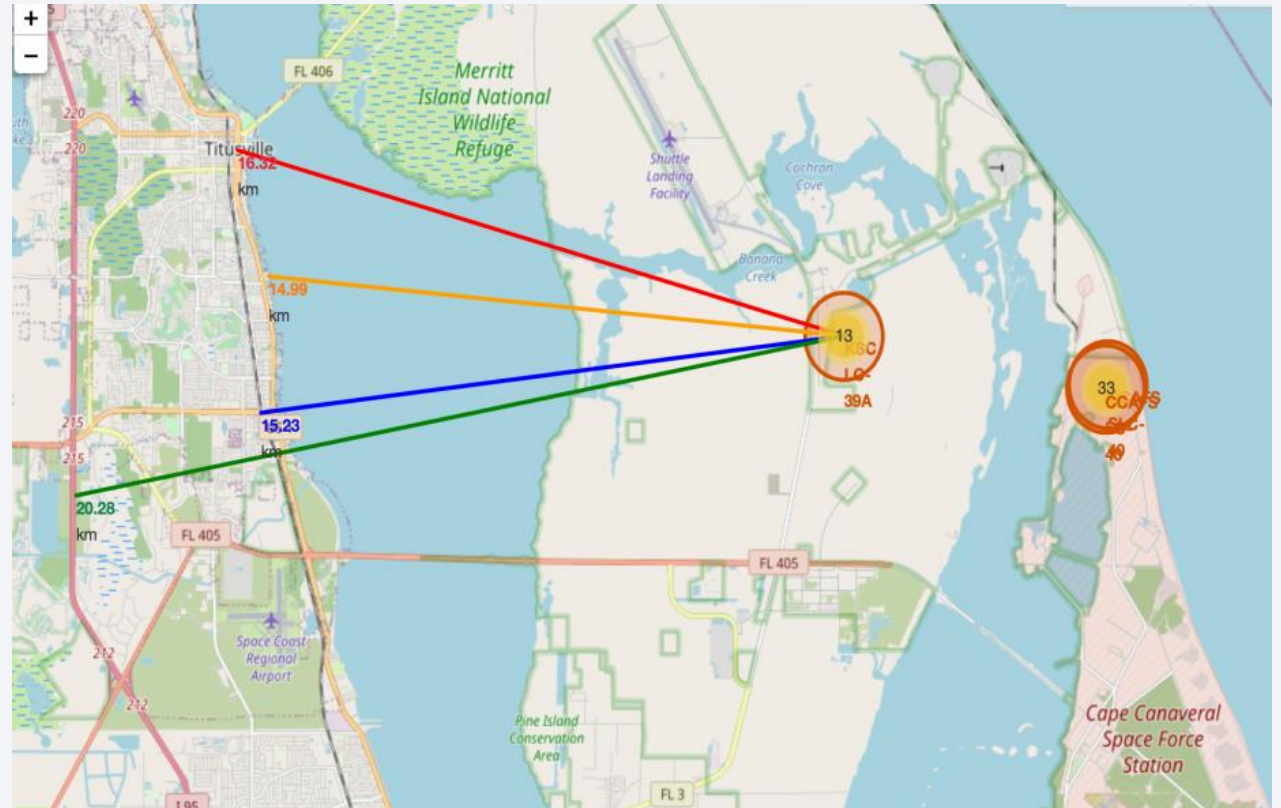
- By referencing **color-coded markers**, we can readily discern **launch sites** with notably high success rates. A **green** marker signifies a **successful launch**, while a **red** marker indicates a **failed** one.
- Launch Site KSC LC-39A boasts an exceptionally impressive track record of **successful launches**.



Location Proximities

Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly 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 populated areas.





Section 4

Build a Dashboard with Plotly Dash

Launch Success Count For All Sites

Total Success Launches by Site



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

Total Success Launches for Site KSC LC-39A



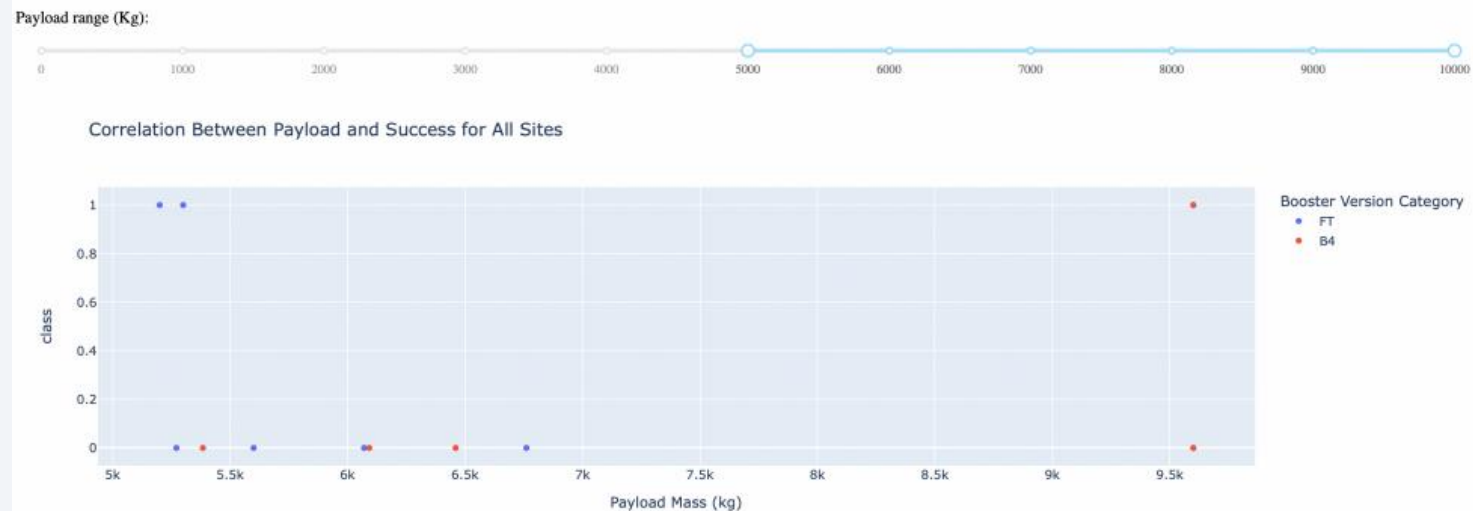
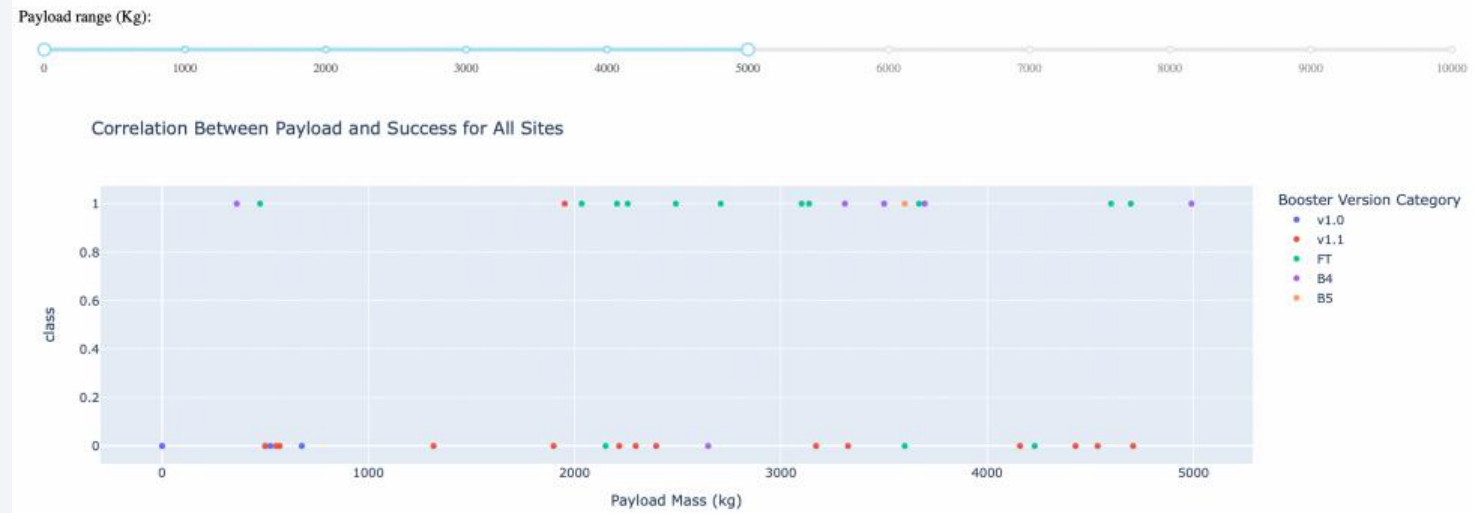
Explanation:

- 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

	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 Test 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

Scores and Accuracy of the Entire Data Set

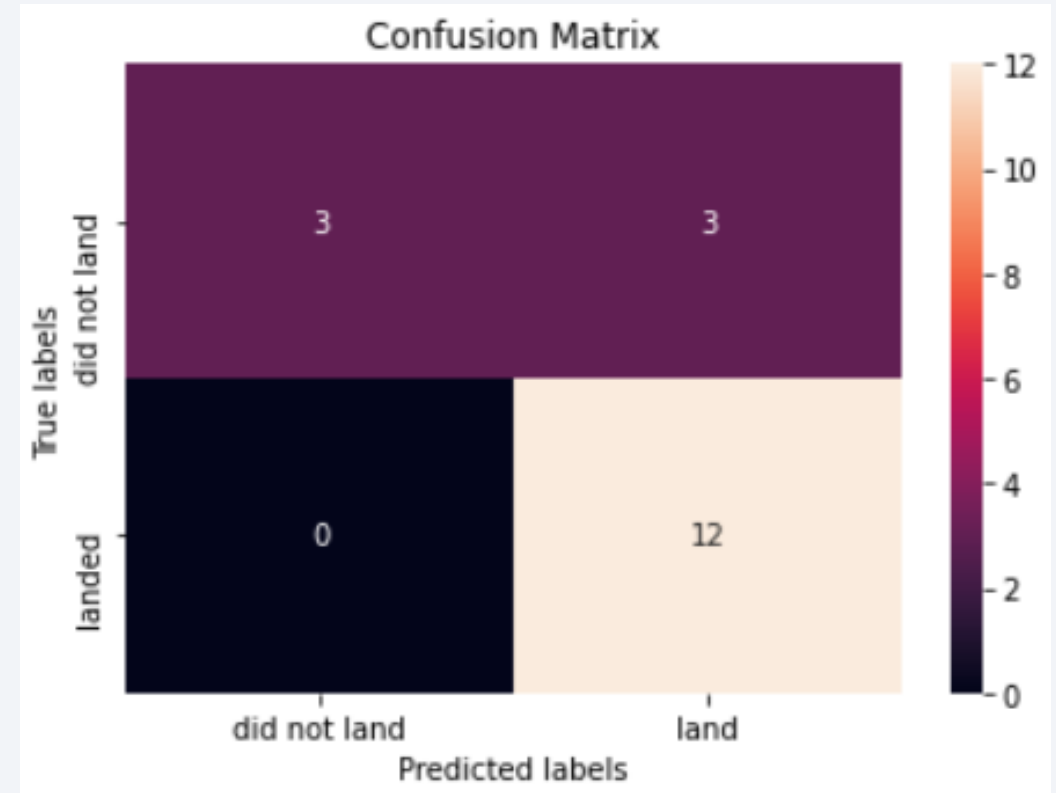
Explanation:

- Based on the scores of the **Test Set**, we **can not** 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 the **highest accuracy**.

Confusion Matrix

Explanation:

- Examining the [confusion matrix](#), we see that [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**.
- 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

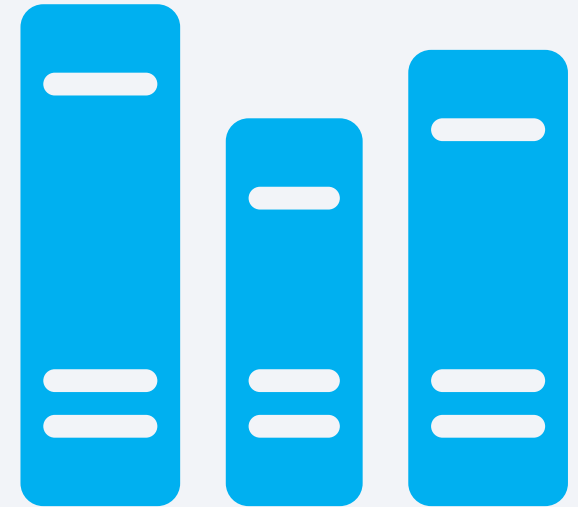
- GitHub Repository URL

[hasanhammad/Applied-Data-Science-Capstone:](#)
[The final task of this capstone project \(github.com\)](#)

- Special Thanks to All Instructors:

[Applied Data Science Capstone - IBM - Course Info | Coursera](#)

- SpaceX data
- Wikipedia



Thank you!

