IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matthew Hill
11/02/2024

Outline

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

Executive Summary

The project involved collecting data from the public SpaceX API and SpaceX Wikipedia page to analyze SpaceX launches and landings. A 'class' column was created to classify successful landings. The data was then explored using SQL queries, various visualizations, folium maps, and dashboards. Relevant columns were selected as features for machine learning models. All categorical variables were converted to binary using one-hot encoding, which likely increased the number of columns significantly. The data was then standardized to prepare it for machine learning algorithms. GridSearchCV was employed to find the optimal parameters for the machine learning models. Finally, the accuracy scores of all models were visualized to compare their performance. This comprehensive approach combined data collection, preprocessing, exploratory data analysis, feature engineering, and machine learning to gain insights into SpaceX's launch and landing outcomes.

Four machine learning models were developed: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. Each model yielded similar results, achieving an accuracy rate of approximately 83.33%. However, all models tended to over-predict successful landings. To enhance model performance and improve accuracy, it is evident that additional data is required for better training and evaluation. This suggests that incorporating more comprehensive datasets could lead to more reliable predictions regarding landing outcomes.

Introduction

Background

- Falcon 9 Rocket launched with a cost of 62 million dollars due to the ability to reuse the first stage, while competitors cost upward of 165 million each. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

- Problems you want to find answers
  - Predict if the Falcon 9 first stage will land successsfully
  - Rate of successful landings
  - Best predictive model for successful landings

Section 1

Methodology

Methodology

- Data collection methodology:
  - Combination of SpaceX public API and Spacex Wikipedia profile
- Perform data wrangling
  - True landings classified as successful and unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridsearchCV to tune models

Data Collection

The data collection process for this project utilized two primary sources: the SpaceX public API and SpaceX's Wikipedia entry. This dual approach allowed for a comprehensive dataset to be compiled.

From the SpaceX public API, we extracted a rich set of data columns including:

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

To complement this, we employed web scraping techniques to gather additional information from a table in SpaceX's Wikipedia entry. This yielded the following columns:

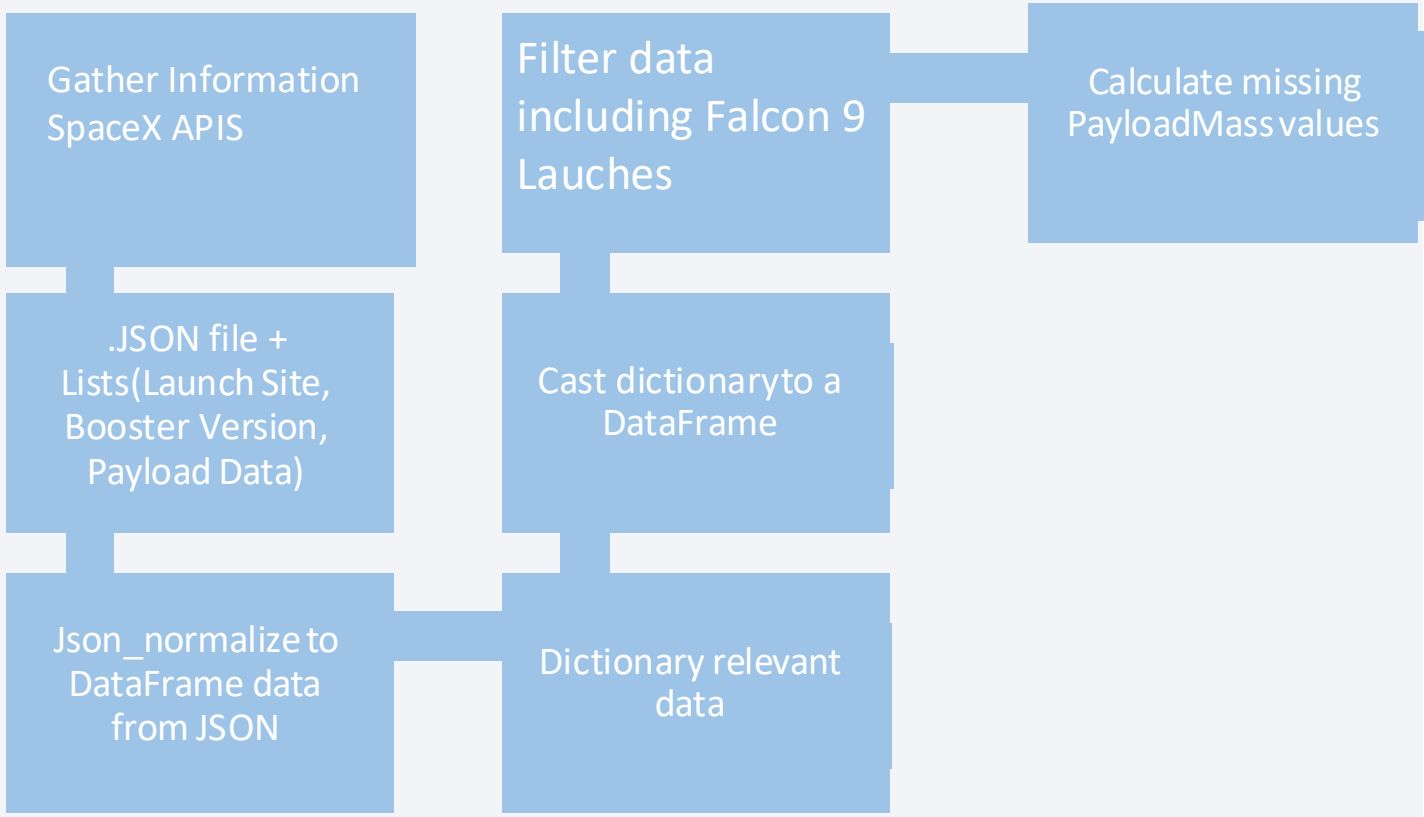
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

The combination of these two data sources provided a comprehensive overview of SpaceX's launches and related information. The subsequent slides in the presentation will illustrate the flowcharts for both the API data collection process and the web scraping methodology, offering a visual representation of how this data was obtained and structured for analysis.

Data Collection – SpaceX API

- GitHub URL:

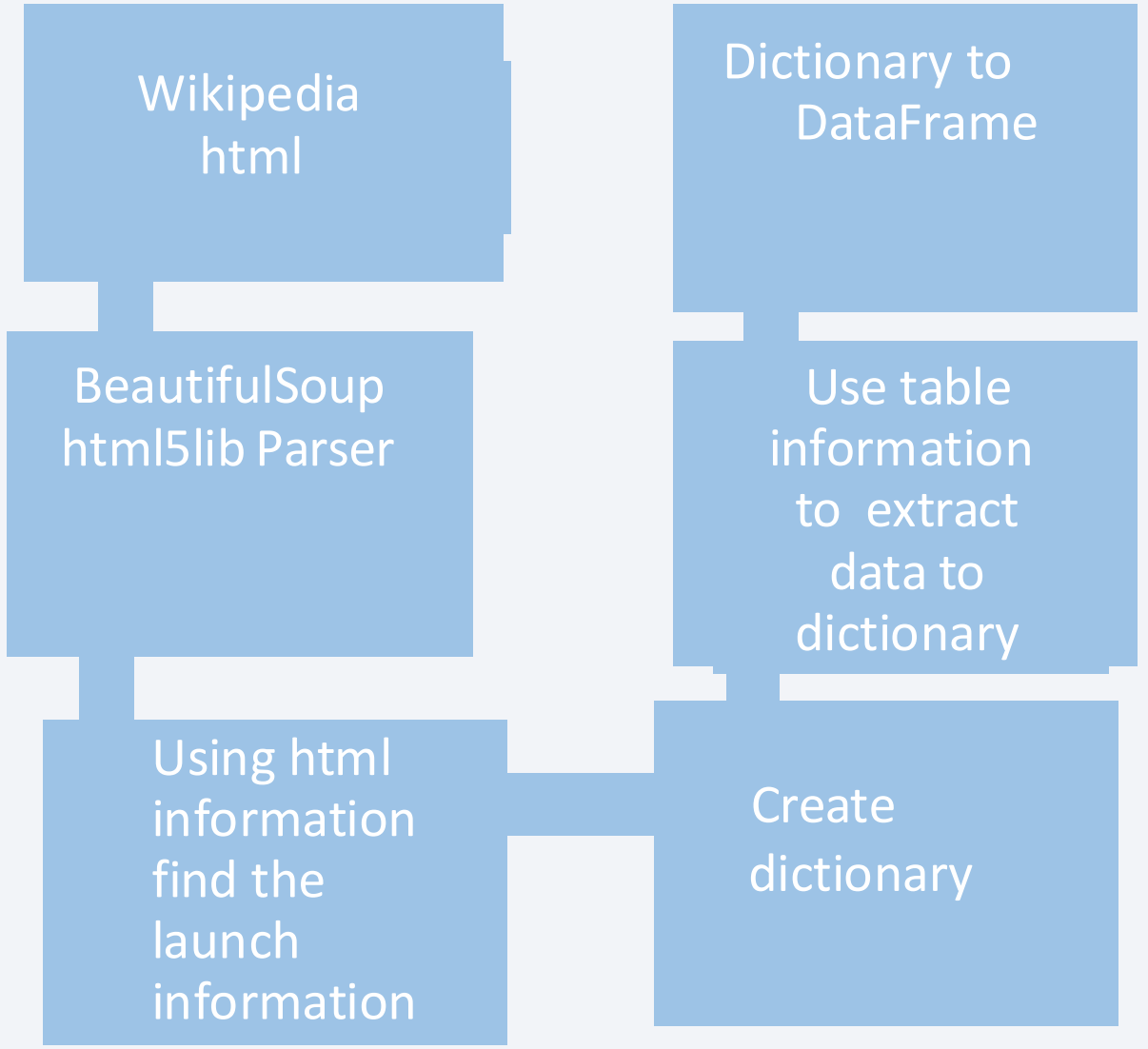
https://github.com/Mhill93/IBM-Data-Science-Certification/blob/main/Capstone%20Project/Data%20Collection%20Api(8).ipynb



Data Collection – Webscraping

GitHub URL:

https://github.com/Mhill93/IBM-Data-Science-Certification/blob/main/Capstone%20Project/Capstone%20Webscraping.ipynb



Data Wrangling

To create a binary classification label for landing outcomes, we'll generate a new column called 'class' in our dataset. This column will be derived from the 'Outcome' column, which contains information about both the mission outcome and landing location. The mapping for the 'class' column will be as follows:

- Successful landings (class = 1) will include:
- True ASDS (Autonomous Spaceport Drone Ship)
- True RTLS (Return to Launch Site)
- True Ocean

Unsuccessful landings (class = 0) will include:

- None None
- False ASDS
- None ASDS
- False Ocean
- False RTLS

This binary classification will allow us to distinguish between successful and unsuccessful landing attempts, providing a clear target variable for our machine learning models. The 'class' column will serve as our training label, with 1 representing a successful landing and 0 representing a failure or unsuccessful attempt.

https://github.com/Mhill93/IBM-Data-Science-Certification/blob/main/Capstone%20Project/Capstone%20Webscraping.ipynb

EDA with Data Visualization

In this project, we conducted an extensive Exploratory Data Analysis (EDA) focusing on key variables including Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. To visualize and analyze the relationships between these variables, we employed a variety of plots and charts. These included scatter plots comparing Flight Number against Payload Mass, and Flight Number against Launch Site. We also created bar plots to examine the relationship between Payload Mass and Launch Site, as well as to visualize the Success Rate across different Orbits. Line charts were utilized to track the Flight Number trends for various Orbits and to display the yearly trend of successful launches. Additionally, we explored the correlation between Payload and Orbit using appropriate visualizations. The primary goal of this comprehensive EDA was to identify and understand potential relationships between variables, which would inform our decision-making process when selecting features for training our machine learning models. By using a combination of scatter plots, line charts, and bar plots, we were able to gain valuable insights into the data patterns and interdependencies, setting a strong foundation for the subsequent modeling phase.

- GitHub URL: https://github.com/Mhill93/IBM-Data-Science-Certification/blob/main/Capstone%20Project/EDA%20with%20Visualization.ipynb

EDA with SQL

The dataset was successfully imported into an IBM DB2 Database to facilitate efficient data management and analysis. Leveraging the SQL Python integration, we executed a series of queries to gain deeper insights into the dataset's structure and content. These queries were strategically designed to extract valuable information about various aspects of SpaceX launches. Specifically, we investigated details such as launch site names, mission outcomes, payload sizes for different customers, booster versions utilized, and landing outcomes. This comprehensive querying approach allowed us to develop a more nuanced understanding of the dataset, providing a solid foundation for subsequent analysis and modeling efforts. By exploring these key variables, we were able to identify patterns and relationships that could be crucial for predicting launch success and optimizing future missions.

- Add the GitHub URL: https://github.com/Mhill93/IBM-Data-Science-Certification/blob/main/Capstone%20Project/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Folium maps were utilized to create interactive visualizations that provide comprehensive insights into SpaceX's launch and landing operations. These maps feature markers for various key elements, including launch sites, successful and unsuccessful landing locations, and proximity to critical infrastructure and geographical features. Specifically, the maps highlight the relative positions of railways, highways, coastlines, and nearby cities in relation to the launch and landing sites. This visual representation offers a clear understanding of the strategic placement of launch sites, taking into account factors such as accessibility, safety, and logistical considerations. Furthermore, the maps effectively illustrate the distribution of successful landings across different geographical locations, allowing for analysis of potential correlations between landing success rates and specific site characteristics or environmental factors. This visualization technique enhances our ability to interpret the complex spatial relationships involved in SpaceX's operations and success rates.

- GitHub URL: https://github.com/Mhill93/IBM-Data-Science-Certification/blob/main/Capstone%20Project/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- The dashboard incorporates two key visualizations: a pie chart and a scatter plot, each offering unique insights into SpaceX launch data. The pie chart provides a flexible view of landing success rates, allowing users to toggle between an overall distribution across all launch sites and individual site-specific success rates. This feature enables a quick comparison of performance between different launch locations. The scatter plot, on the other hand, offers a more detailed analysis by accepting two user inputs: the choice of all sites or a specific launch site, and a payload mass range selectable via a slider from 0 to 10,000 kg. While the pie chart effectively illustrates the success rates of launch sites, the scatter plot delves deeper, revealing potential correlations between landing success and factors such as launch site, payload mass, and booster version category. Together, these interactive visualizations provide a comprehensive tool for exploring the various factors influencing the success of SpaceX launches and landings.

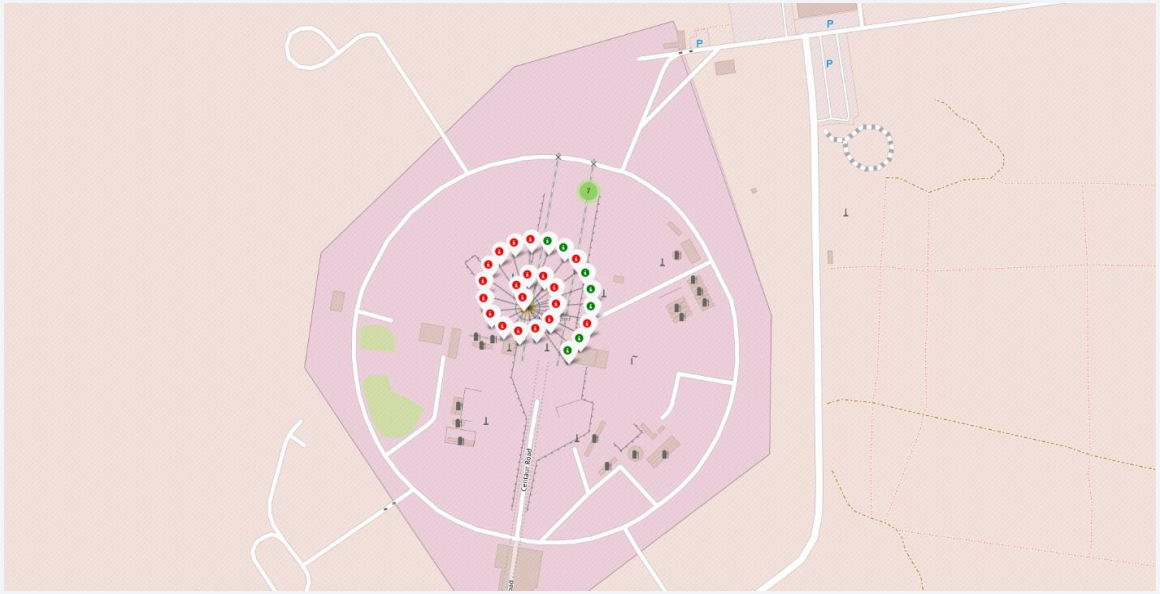
Predictive Analysis (Classification)

- GitHub URL: https://github.com/Mhill93/IBM-Data-Science-Certification/blob/main/Capstone%20Project/Machine%20Learning%20Prediction%20lab.ipynb

Results

Exploratory data analysis results

- Launches have improved over time
- KSC LC-39A has the highest success rate of launching zones

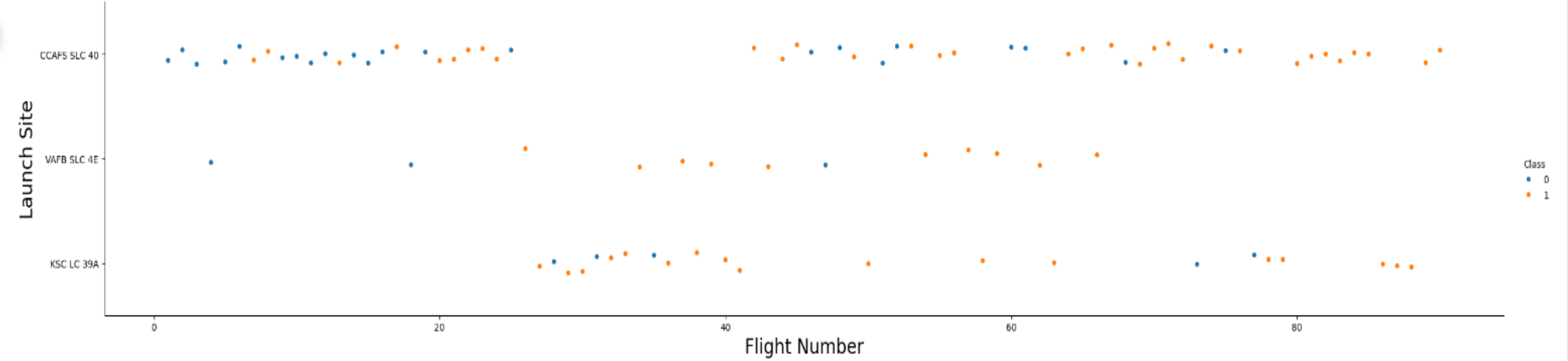


- Predictive analysis results
  - Decision Tree Model is the best model to use

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

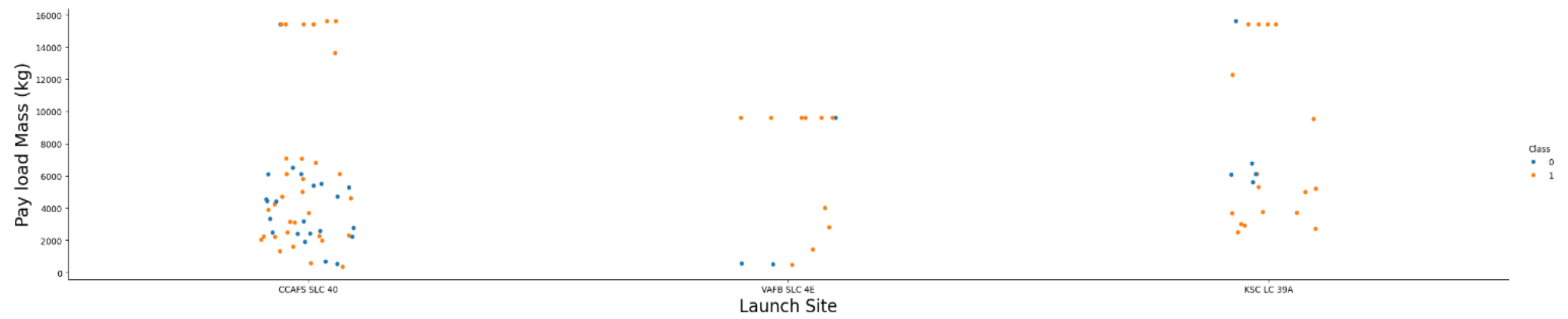


Blue – Fail & Orange – Success

Later Flights saw more success at a higher rate

CCAFS house the most lauches.

Payload vs. Launch Site



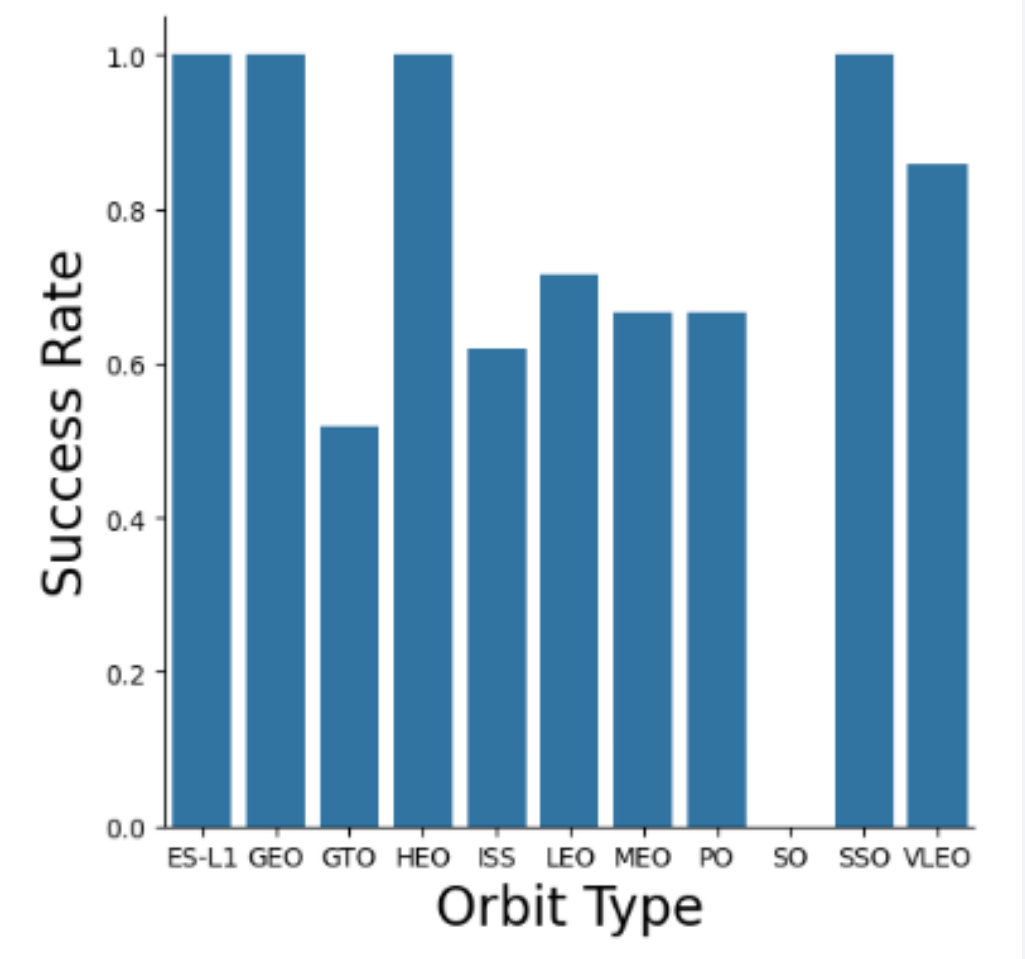
It seems that the higher the payload, the more successful the launches were, especially above 7,000Kg

KSC LC 39A had successful launches, with the payload being under 5,000 Kg

Success Rate vs. Orbit Type

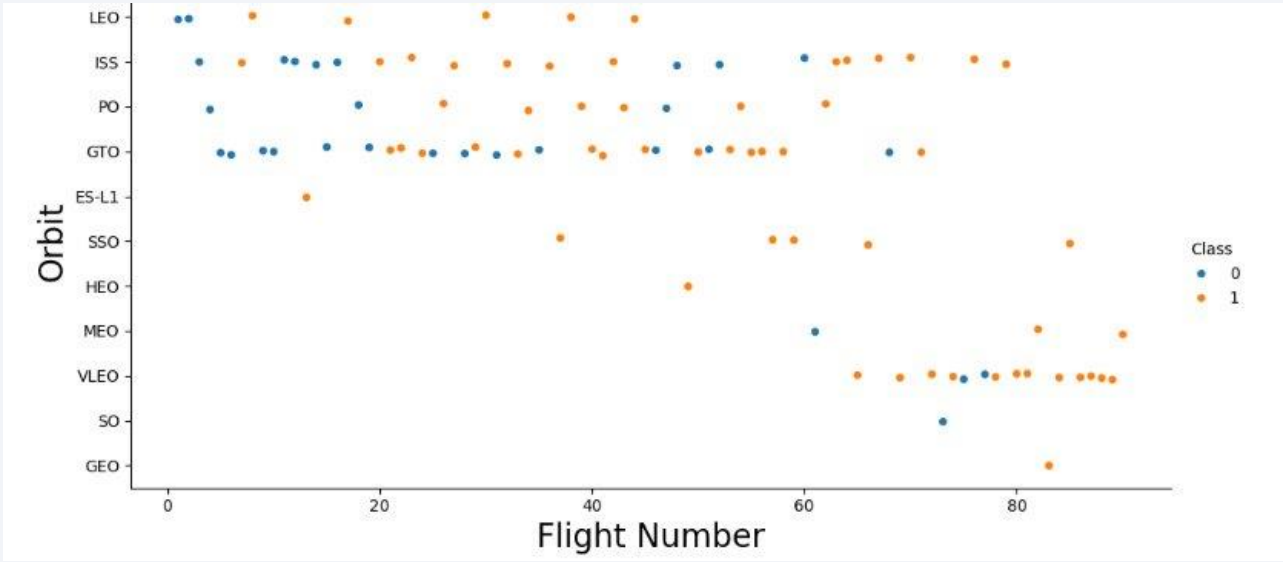
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

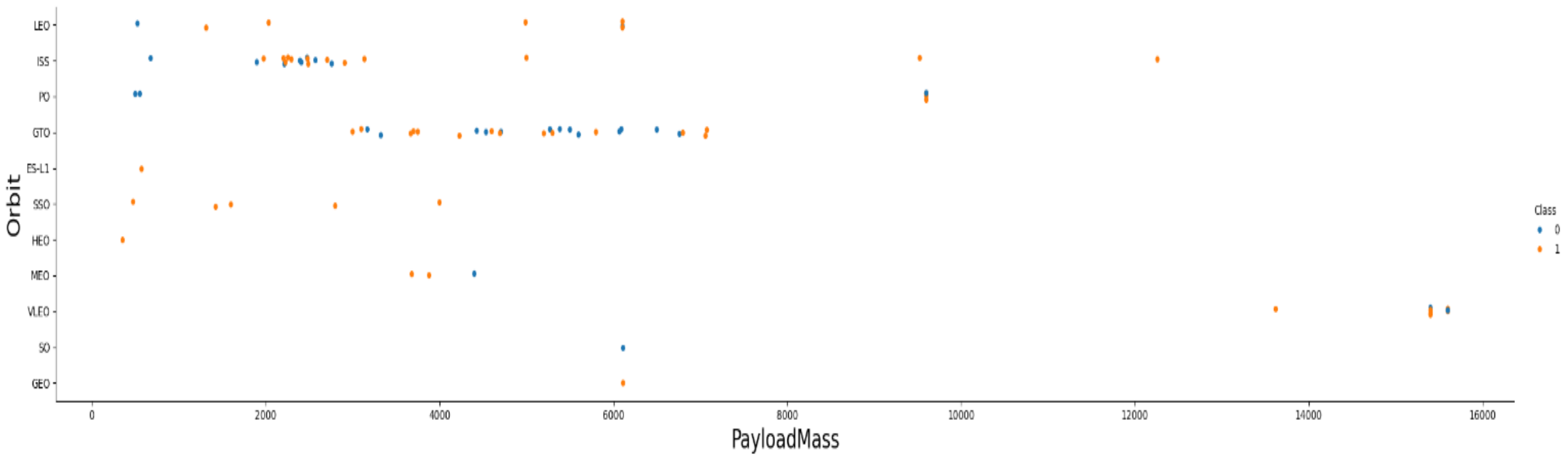


Flight Number vs. Orbit Type

- Success rate increased for the most part with the number of flights for each orbit especially for the LEO orbit. GTO orbit does not follow this trend however



Payload vs. Orbit Type



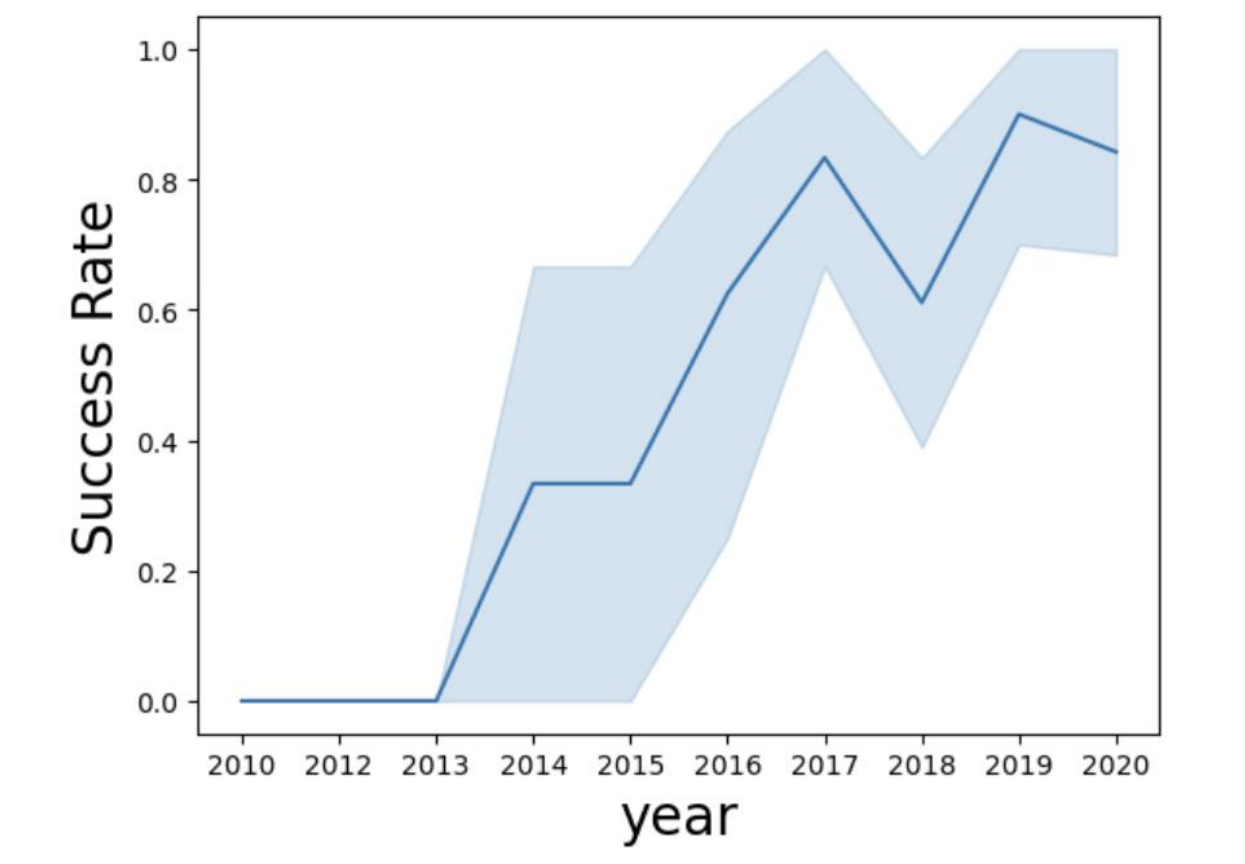
- Heavy payloads seem better with LEO, ISS and PO orbits.
- GTO has mixed results

Launch Success Yearly Trend

Success rate increased from 2013-2017 and 2018-2017

The rate decreased 2017-2018 as well as 2019-2020

In 2020 a success rate of about 80%



All Launch Site Names

The unique names are as follows:

- CCAFS SLC-40
- CCAFS LC-40
- CCAFSSLC-40
- KSC LC-39A
- VAFB SLC-4E
- Similar names may indicate similar sites with a rename. Others may suggest a different site all together.

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

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

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

Launch Site Names Begin with 'CCA'

- Used Query to find 5 records where launch sites 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

Total Payload Mass

- Calculated the total payload carried by boosters from NASA using the query and got an answer of 45596

```
%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

Average Payload Mass by F9 v1.1

- The calculation determines the mean payload mass for missions that utilized the Falcon 9 v1.1 booster variant.
- The typical payload mass carried by the F9 v1.1 falls towards the lower end of our payload capacity spectrum.

```
%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.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

average_payload_mass
2534

First Successful Ground Landing Date

- December 12, 2022 is the date of the first successful landing ground pad

```
%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

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query was refined using the WHERE clause to focus on boosters that achieved successful landings on drone ships. We further narrowed the results by applying an AND condition to identify landings that were not only successful but also involved payloads with a mass between 4,000 and 6,000 units (presumably kilograms or pounds). This combination of criteria allowed us to analyze a specific subset of missions that met both the landing location and payload mass requirements.

```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Used the mission outcome filter to find the results.
- 99% success rate, with one payload being unclear

```
%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

Boosters Carried Maximum Payload

- Each booster can be correlated to the amount that it can haul along with the payload

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.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

2015 Launch Records

- Two Failed landing in 2015. January and April with similar booster versions of F9 v1. B11012 and 1015

```
%%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)

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

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

```
In [13]: %%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.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

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

Section 3

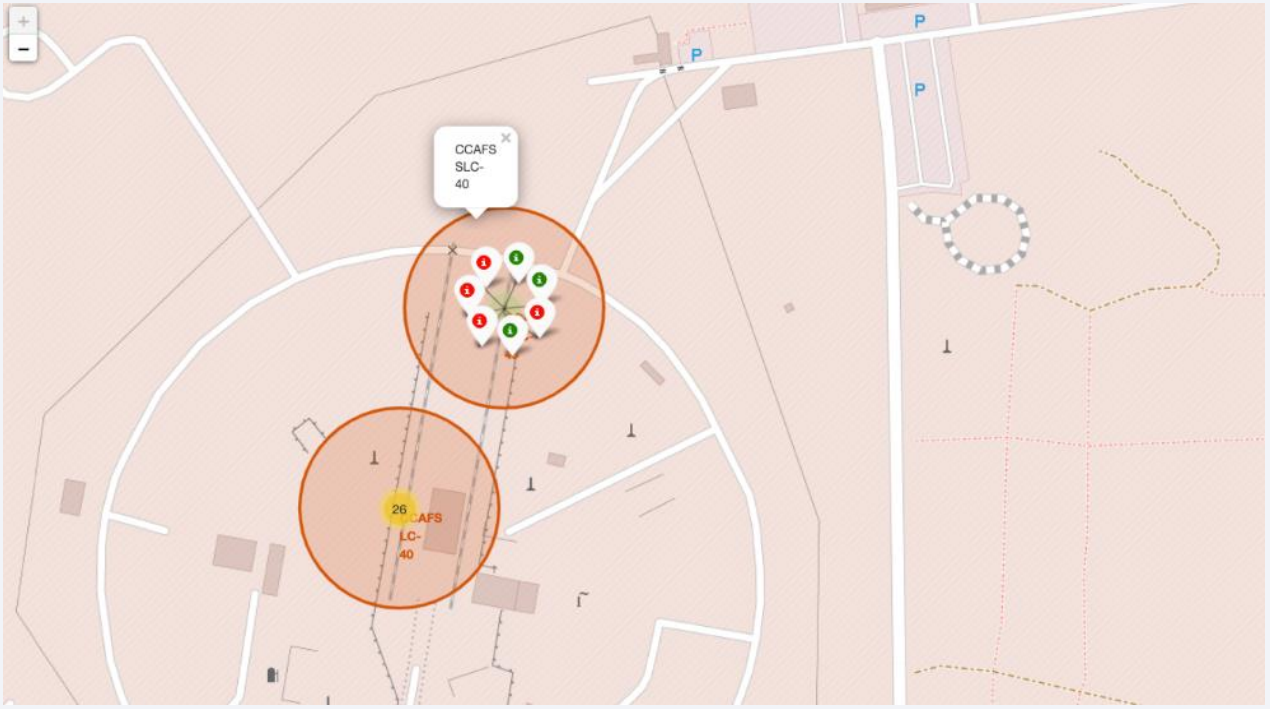
Launch Sites Proximities Analysis

Launch Locations

- Located on the opposite side of the country and as close to the equator as possible is the launch site. These sites will use rockets and the Earth's rotation to launch.

Launch Results

- Launch site CCAFS SLC-40 has a 43% launch rate with 3 of the 7 being successful. Green is successful while red is not.



Proximities

There seems to be a distance of at least 20KM from populated areas for safety use. Being as close to the water as possible. Still close enough to transportation hubs for goods and tool, but far enough not to damage them.

```
print("Railway Distance", railway_distance)
```

```
Railway Distance 21.961465676043673
```

```
print("Highway Distance", highway_distance)
```

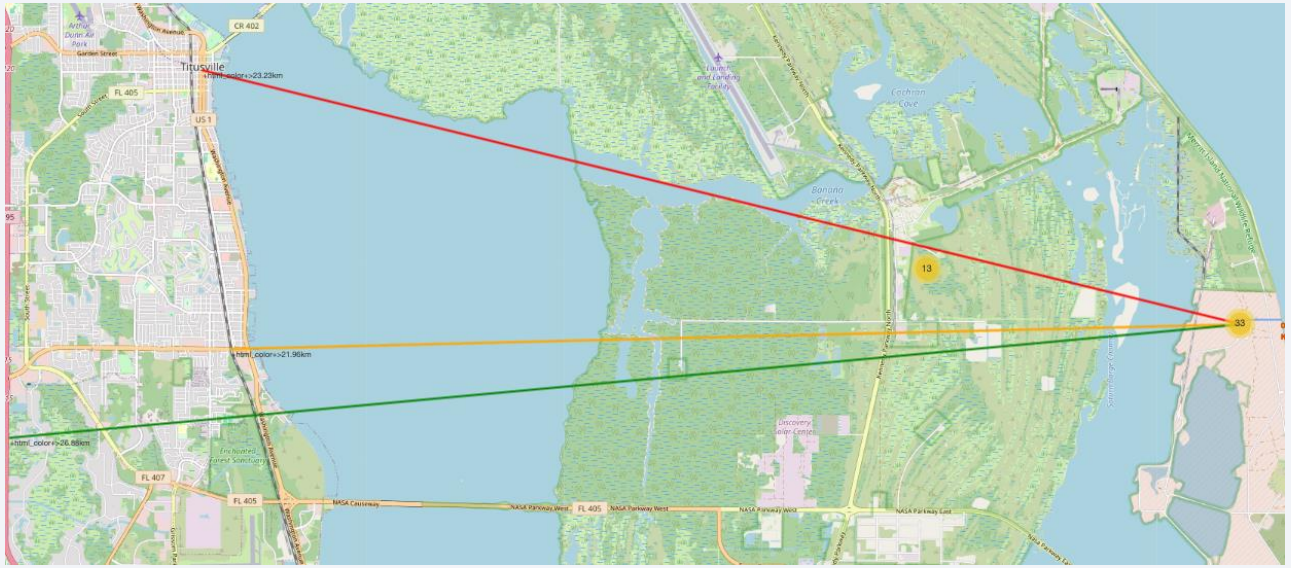
```
Highway Distance 26.88038569681492
```

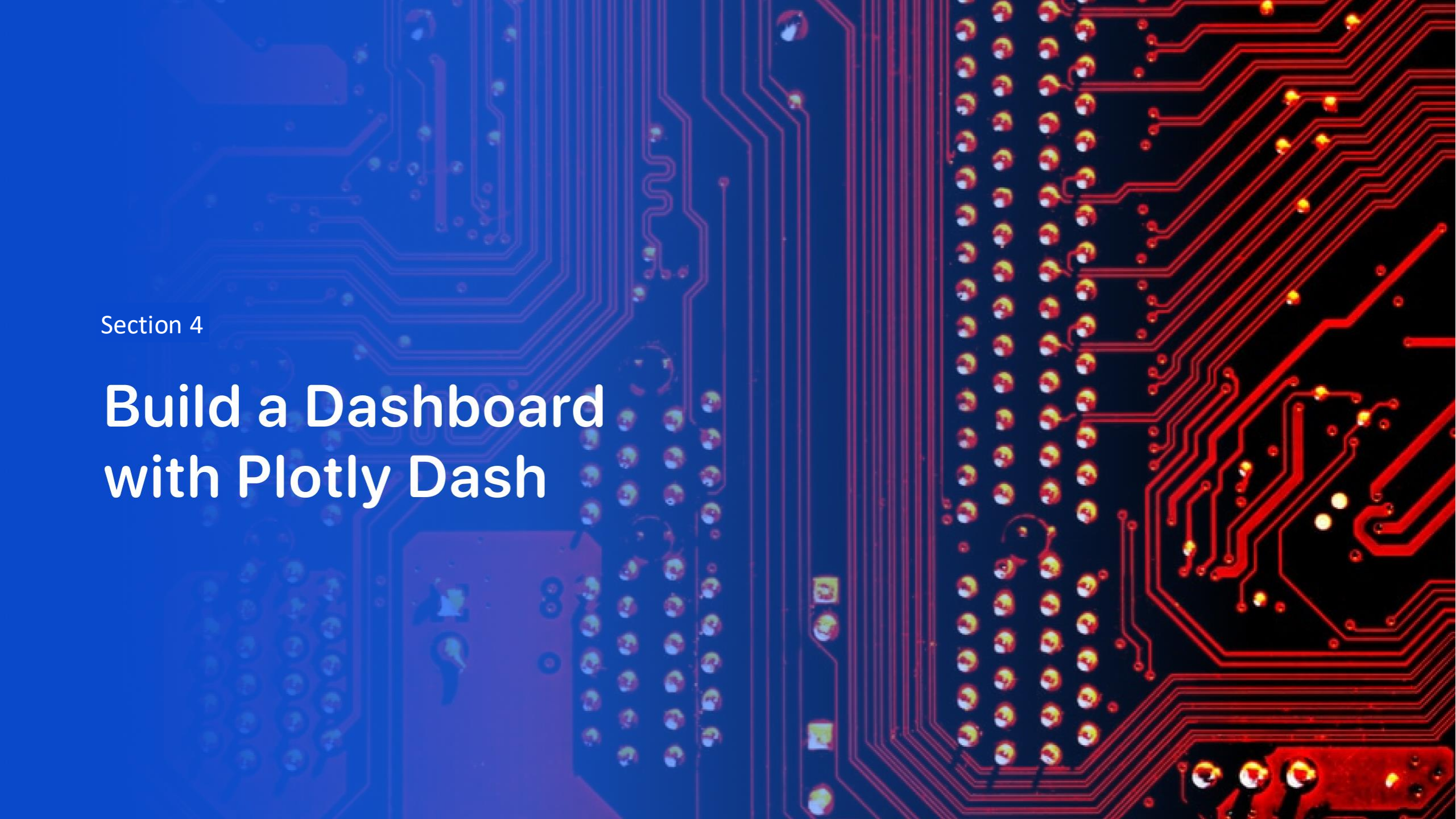
```
print("Coastline Distance", distance_coastline)
```

```
Coastline Distance 0.8627671182499878
```

```
print("City Distance", city_distance)
```

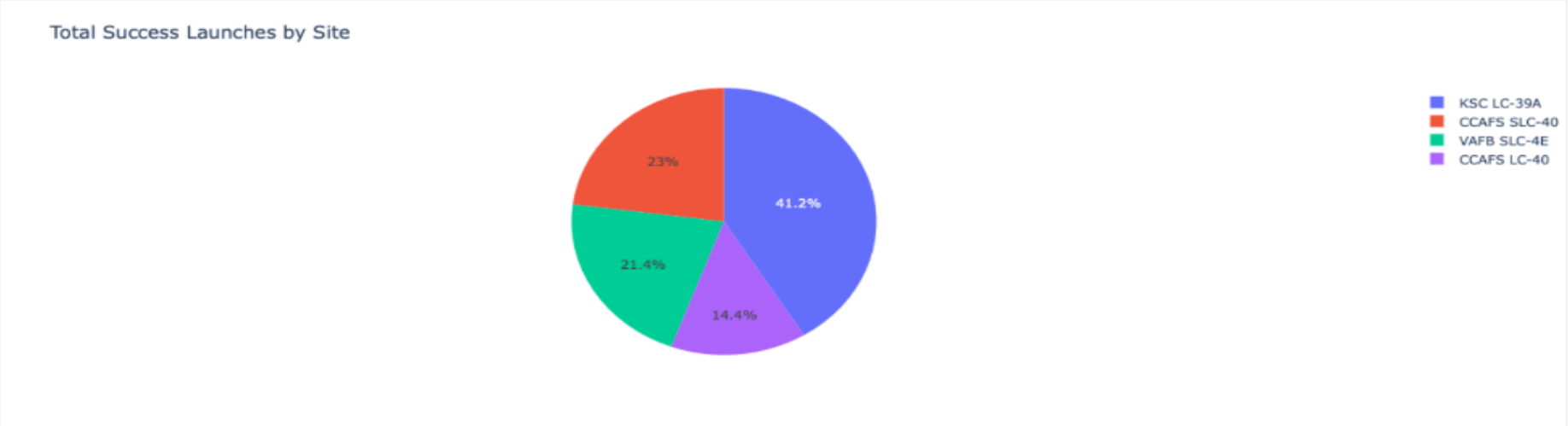
```
City Distance 23.234752126023245
```




Section 4
Build a Dashboard with Plotly Dash

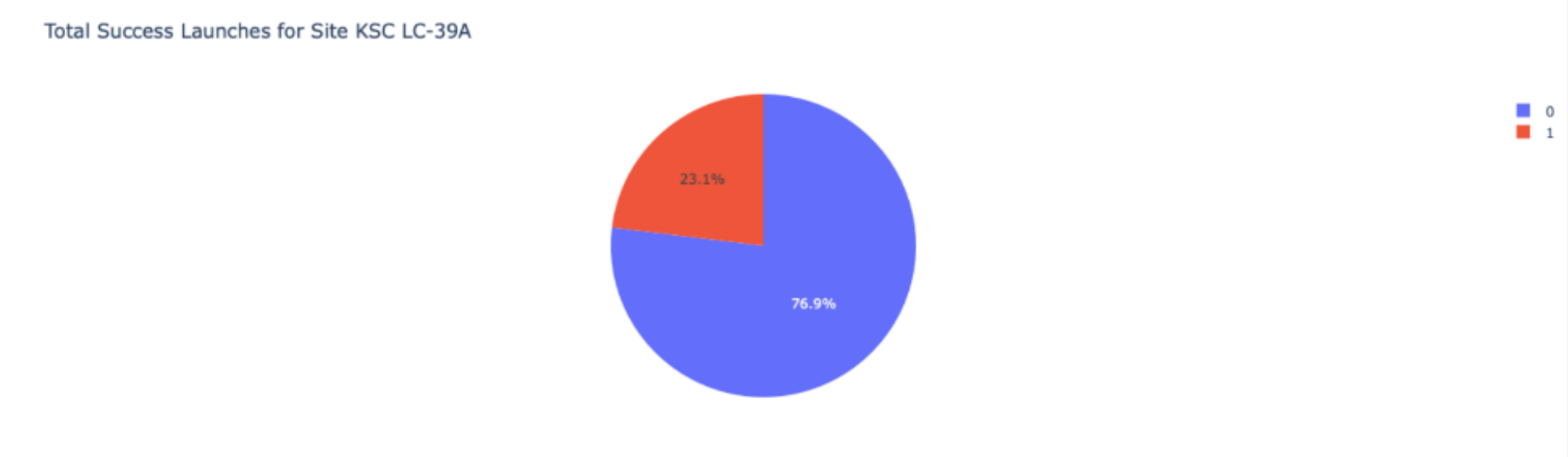
Success by Site

- The distribution of successful landings across all launch sites reveals interesting patterns. Notably, CCAFS LC-40 and CCAFS SLC-40 refer to the same location, with the former being the older designation. When combined, the successful landings at CCAFS and KSC are equal in number. However, it's important to note that a significant portion of these successful landings occurred before the name change from LC-40 to SLC-40. In contrast, VAFB (Vandenberg Air Force Base) shows the lowest proportion of successful landings among the sites. This discrepancy could be attributed to two factors: a smaller sample size of launches from VAFB, and the increased challenges associated with launching from the West Coast. These geographical and operational differences may contribute to the varying success rates observed across the different launch sites.



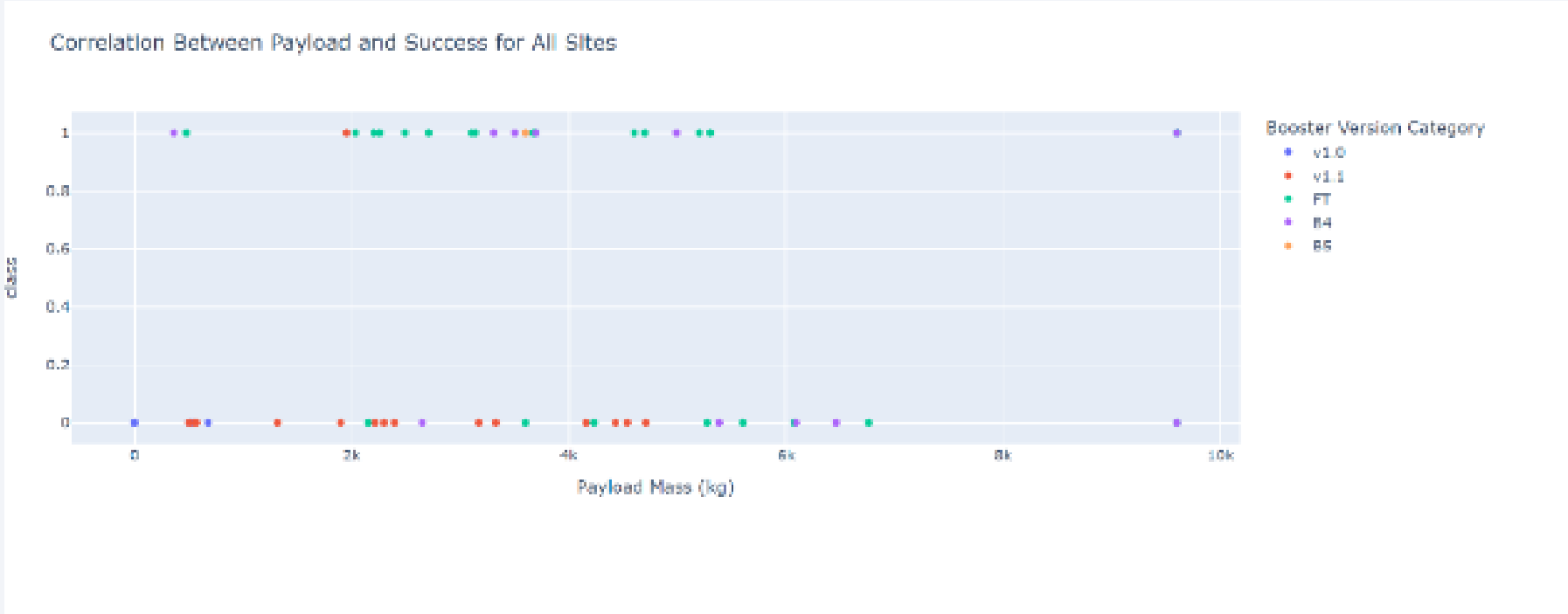
Greatest Launch Rate

- KSC LC-39A has the greatest success rate at 76.9%, only 3 failed while 10 were succesful



Payload Mass and Success Rate

- Payloads betwenn 2,000 kg and 5000 kg have the hightest success rate

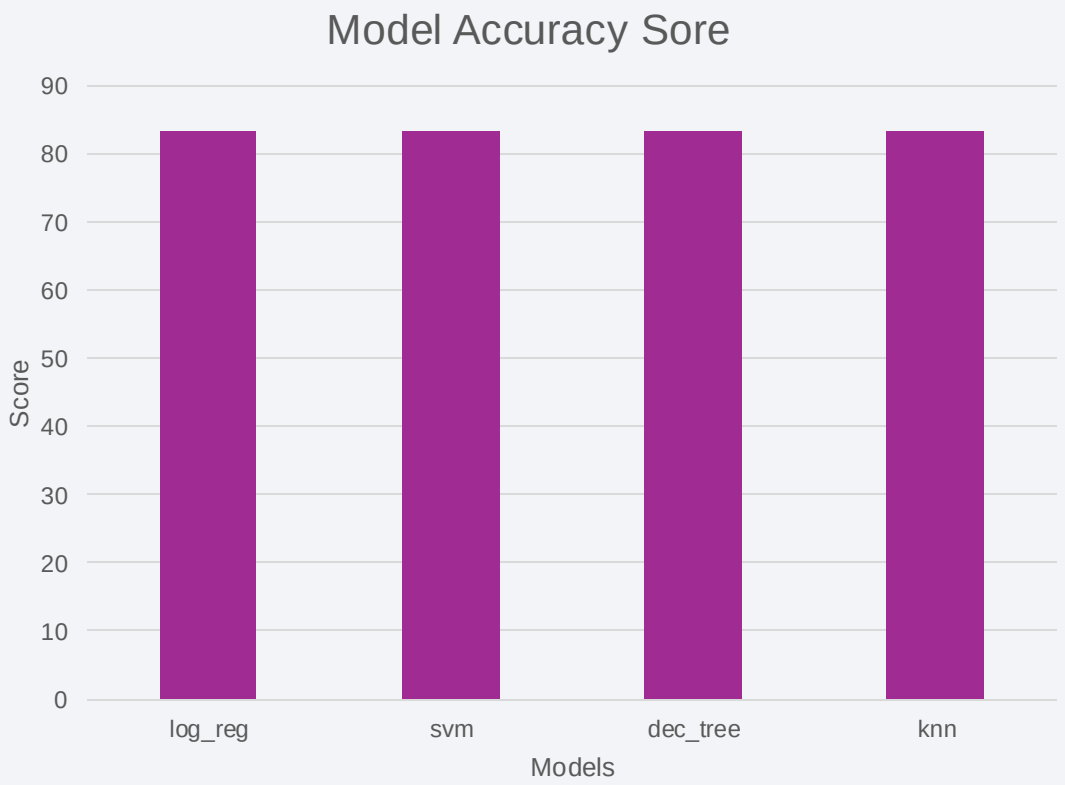


Section 5

Predictive Analysis (Classification)

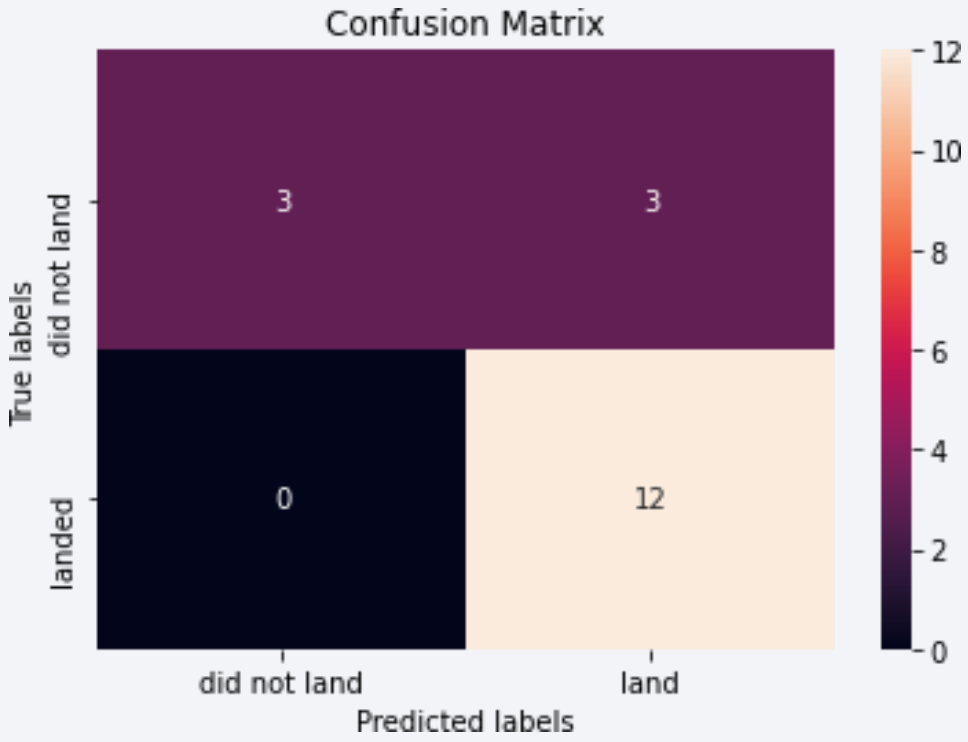
Classification Accuracy

All four machine learning models demonstrated remarkably similar performance on the test set, each achieving an accuracy of 83.33%. However, it's crucial to highlight that the test set was relatively small, consisting of only 18 samples. This limited sample size can lead to significant variability in accuracy results, particularly evident in the Decision Tree Classifier model, which showed fluctuations in performance across multiple runs. The small test set makes it challenging to definitively determine which model is truly superior. To obtain more reliable and stable results, and to make a more informed decision about the best-performing model, it would be beneficial to acquire and incorporate additional data into our analysis.



Confusion Matrix

The performance of all models on the test set was consistent, resulting in identical confusion matrices across the board. The models accurately predicted 12 successful landings when the actual outcome was indeed successful. Additionally, they correctly identified 3 unsuccessful landings when the true label indicated failure. However, there were 3 instances where the models incorrectly predicted successful landings for cases that were actually unsuccessful, resulting in false positives. This pattern suggests that our models have a tendency to overestimate the occurrence of successful landings, indicating a slight bias towards positive outcomes in their predictions.



Conclusions

- The Decision tree classifier is the best machine learning algorithm for this task.
- KSC LC-39A had the most successful launches of any sites.
- Sites are placed in areas for safety reason and to use the Earth's spin for an advantage.
- The larger the payload at a launch site, the greater the success rate at a launch site.
- A larger Data Set would help the predicitive analytics results

Appendix

- Thank you to Coursera and IBM!

Thank you!