IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohd Zaed Fahmi Mohamed Yussof
November 11, 2024

Outline

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

Executive Summary

- Summary of methodologies:
    1. Data collection
    2. Data wrangling
    3. Exploratory data analysis (EDA) using visualization and SQL
    4. Building an interactive visual analytics with Folium and Plotly Dash
    5. Predictive analysis with classification models
- Summary of all results
    1. Exploratory data analysis (EDA) results
    2. Interactive analytics demo
    3. Predictive analysis results

Introduction

- Project background and context

  SpaceX is the leading company in the commercial space industry, making space travel more affordable. The company offers Falcon 9 rocket launches on its website for 62 million dollars, while other providers charge over 165 million dollars. Much of the cost savings come from SpaceX being able to reuse the first stage of the rocket. By predicting wether the first stage will land successfully, we can estimate the cost of a launch. Using publicly available data and machine learning models, we will predict if SpaceX can reuse the first stage.

- Problems you want to find answers

  1. How do factors like payload mass, launch site, number of flights, and orbit types impact the success of the first stage landing?
  2. Has the success rate of landings improved over the years?
  3. Which algorithm is the most effective for binary classification in 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
  - How Building, tuning and evaluation of classification models to ensure the best results

Data Collection

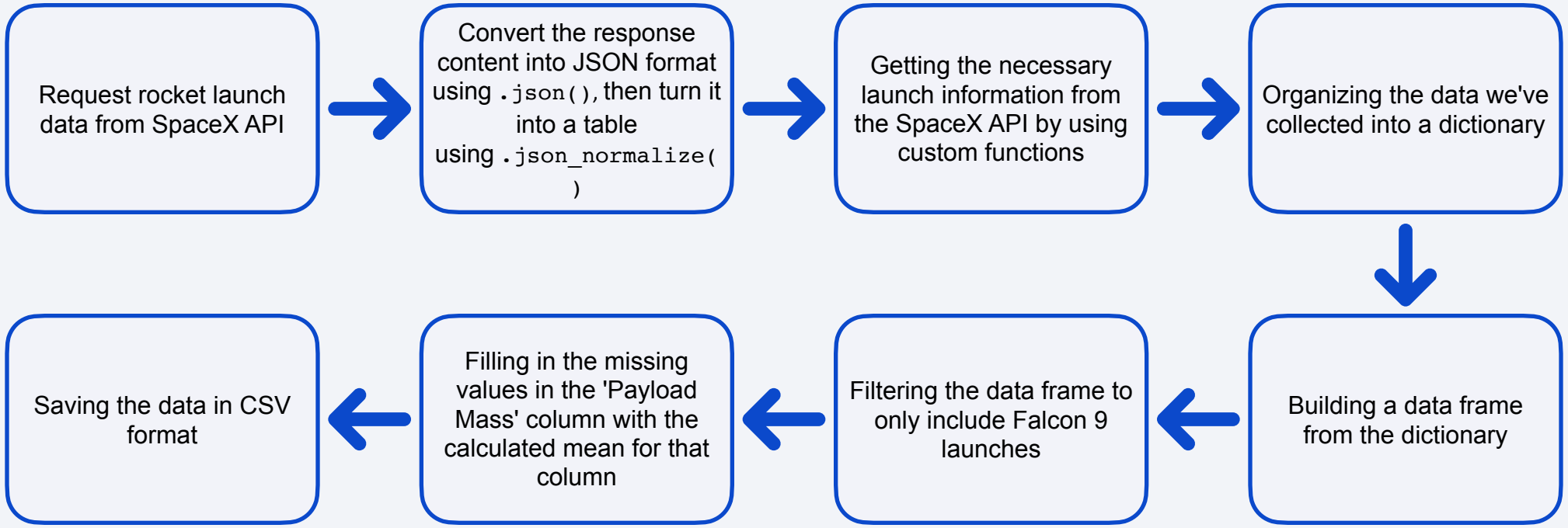
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia page.
- We had to use both of these data collection methods in order to get complete information about the launches for more detailed analysis.
- Data columns are obtained by using SpaceX REST API:

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

- Data columns are obtained by using Wikipedia Web Scraping:

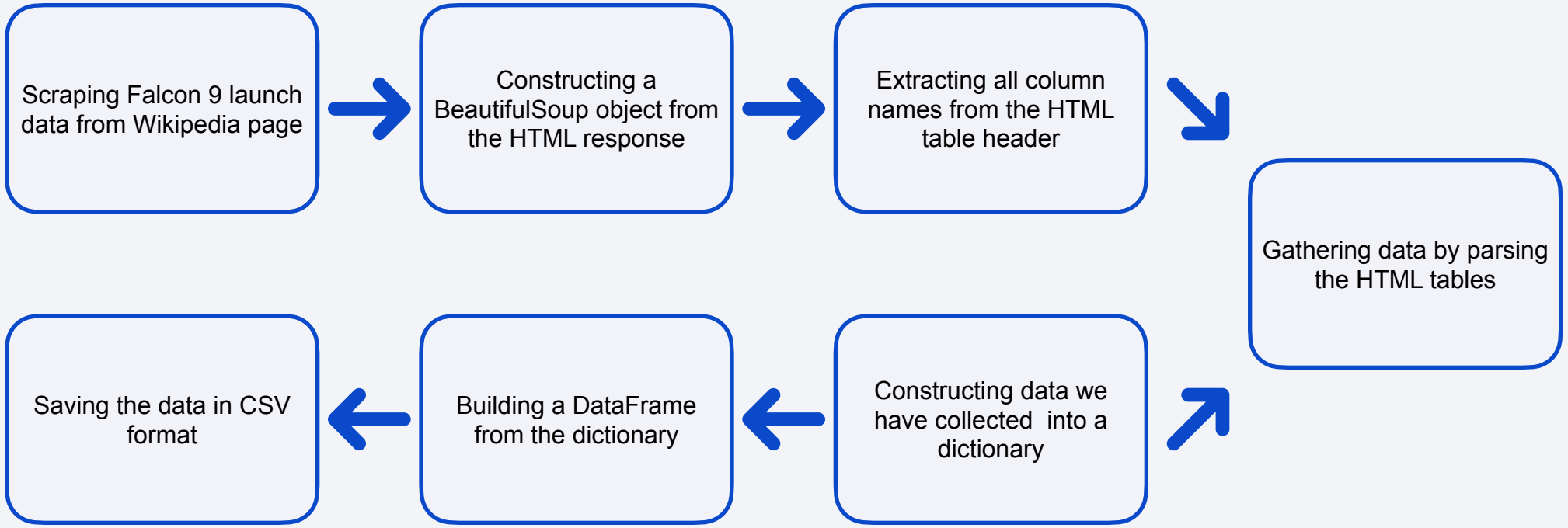
  Fight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb

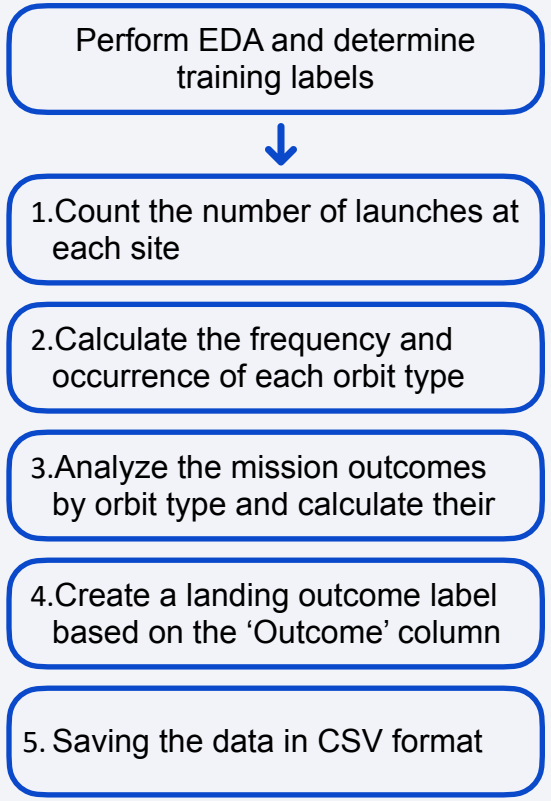
Data Collection - Scraping



https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- The dataset includes various cases where the booster did not land successfully
- Sometimes, a landing was attempted but failed due to accidents
- ‘True Ocean’ indicates a mission outcome where the booster successfully landed in a specific ocean region
- ‘False Ocean’ indicates a mission outcome where the booster unsuccessfully landed in a specific ocean region
- ‘True RTLS’ means the booster successfully landed on a ground pad
- ‘False RTLS’ means the booster unsuccessfully landed on a ground pad
- ‘True ASDS’ means the booster successfully landed on a drone ship
- ‘False ASDS’ means the booster unsuccessfully landed on a drone ship
- These outcomes are mainly converted into training labels, with ‘1' representing a successful landing and ‘0’ representing an unsuccessful landing



https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- Charts plotted include:
  - 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
  - Success Rate Yearly Trend
- Scatter plots illustrate relationships between variables. If a relationship is present, it could be useful for machine learning models
- Bar chart compare discrete categories, showing relationships between the categories and measured value
- Line charts display trends in data over time

https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

Performed SQL queries:

- Displayed the names of the unique launch sites in the space mission
- Retrieved 5 records where launch sites begin with the string ‘CCA’
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed the average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful ground pad landing outcome was achieved
- Listed the names of boosters that successfully landed on a drone ship with payload mass between 4,000 and 6,000
- Displayed the total number of successful and failed mission outcomes
- Listed the booster versions that carried the maximum payload mass
- Retrieved failed landing outcomes on drone ships, along with their booster versions and launch site names, for the months in 2015
- Ranked landing outcomes (e.g., Failure on drone ship or Success on ground pad) by count between the dates 2010-06-04 and 2017-03-20 in descending order

https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers for all Launch Sites:
  - Added markers with circles, popup labels, and text labels for the NASA Johnson Space Center, using its latitude and longitude as the starting point
  - Placed markers with circles, popup labels, and text labels for all launch sites based on their coordinates, highlighting their geographical locations and proximity to the equator and coastlines
- Colored Markers for launch outcomes for each Launch Site:
  - Added colored markers indicating launch outcomes: **green** for successful launches and **red** for failed ones, using a marker cluster to visualize which sites have higher 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

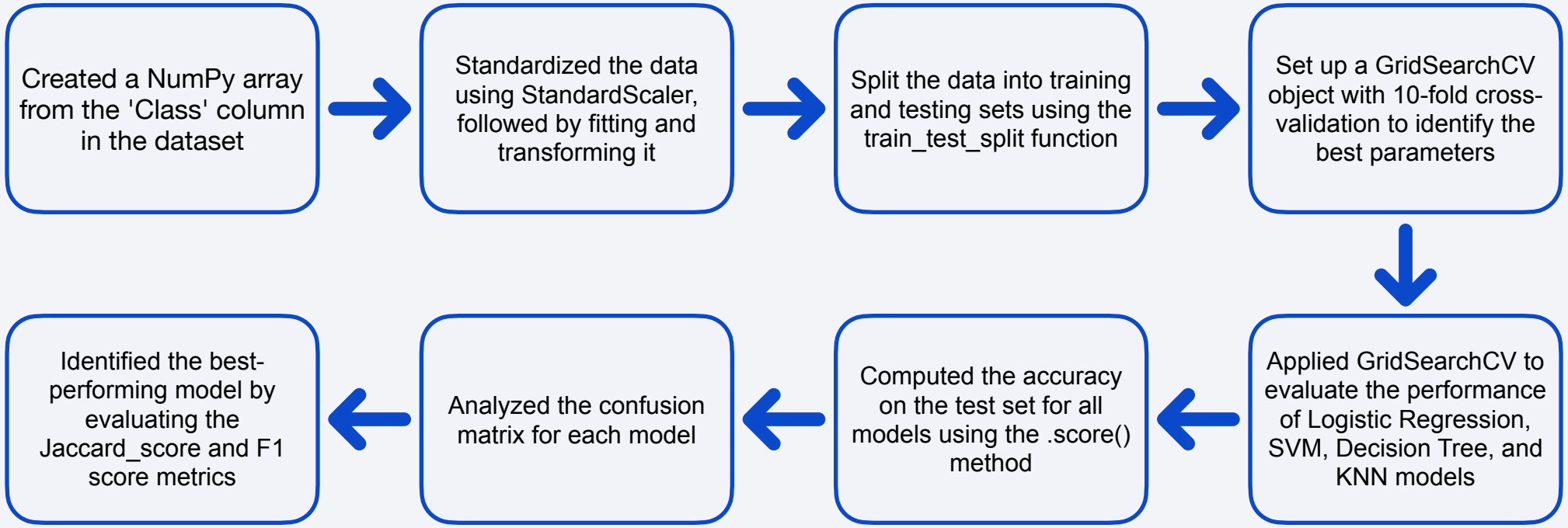
https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown:
  - Added a dropdown menu to allow selection of a specific launch site
- Pie Chart for Successful Launches:
  - Included a pie chart displaying the total count of successful launches for all sites, and for individual sites, showing the ratio of successful vs. failed launches when a specific site is selected
- Payload Mass Range Slider:
  - Added a slider to adjust and select a range for payload mass
- Scatter Plot of Payload Mass vs. Success Rate by Booster Version:
  - Added a scatter plot illustrating the relationship between payload mass and launch success rates across different booster versions

https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

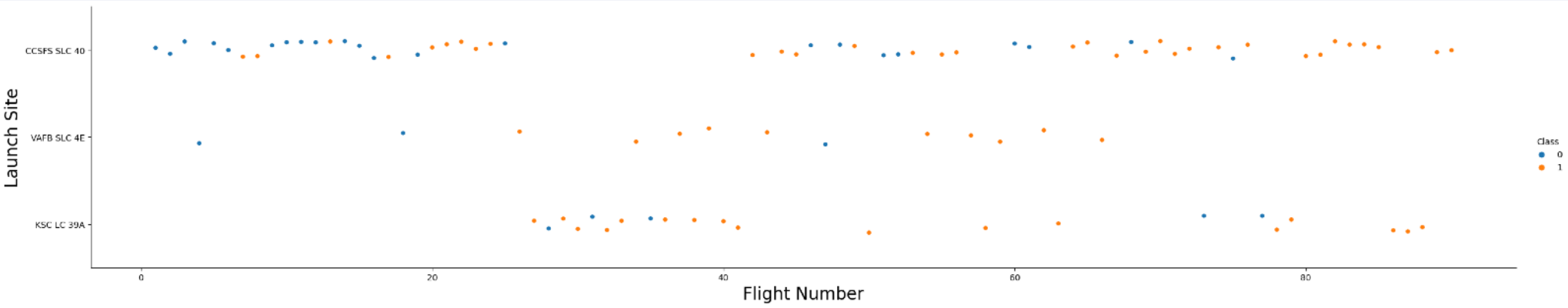
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



Explanation:

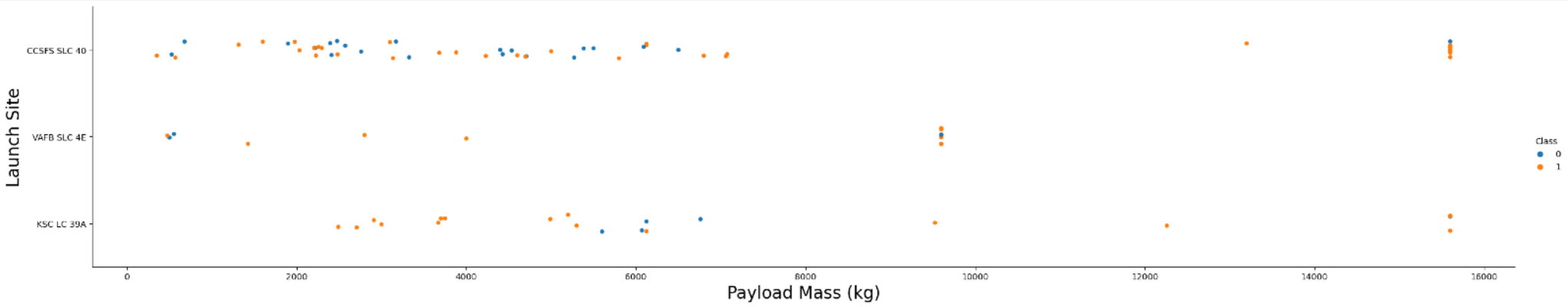
1.The earliest flights all failed while the latest flights all succeeded

2.The CCAFS SLC 40 launch site has about a half of all launches

3.VAFB SLC 4E and KSC LC 39A have higher success rates

4.It can be assumed that each new launch has a higher rate of success

Payload vs. Launch Site



Explanation:

1.For every launch site the higher the payload mass, the higher the success rate

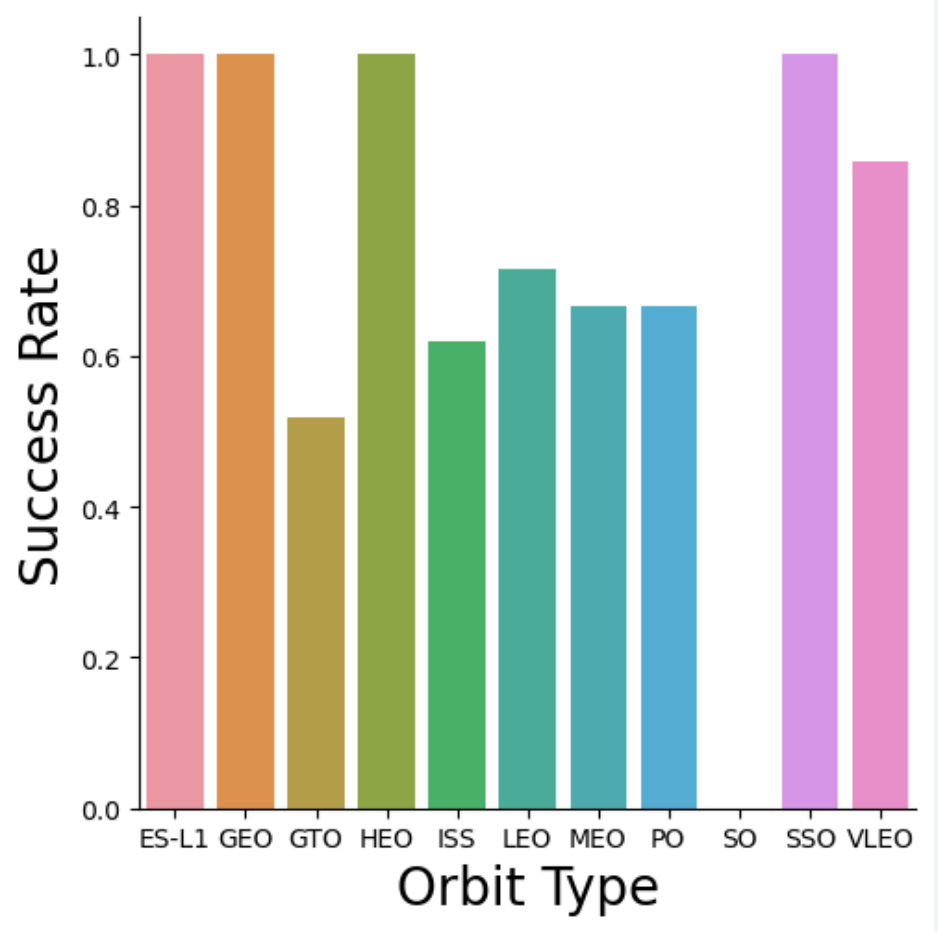
2.Most of the launches with payload mass over 7,000 kg were successful

3.KSC LC 39A has a 100% success rate for payload mass under 5,500 kg

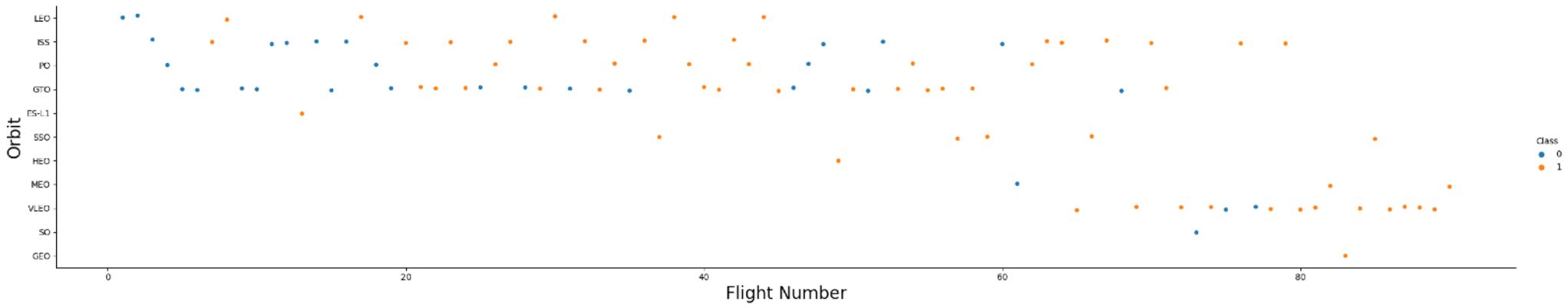
Success Rate vs. Orbit Type

Explanation:

1. Orbit with 100% success rate are: ES-L1, GEO, HEO, and SEO
2. Orbits with 0% success rate are: SO
3. Orbits with success rate between 50% and 85%: GTO, ISSUES, LEO, MEO, and PO



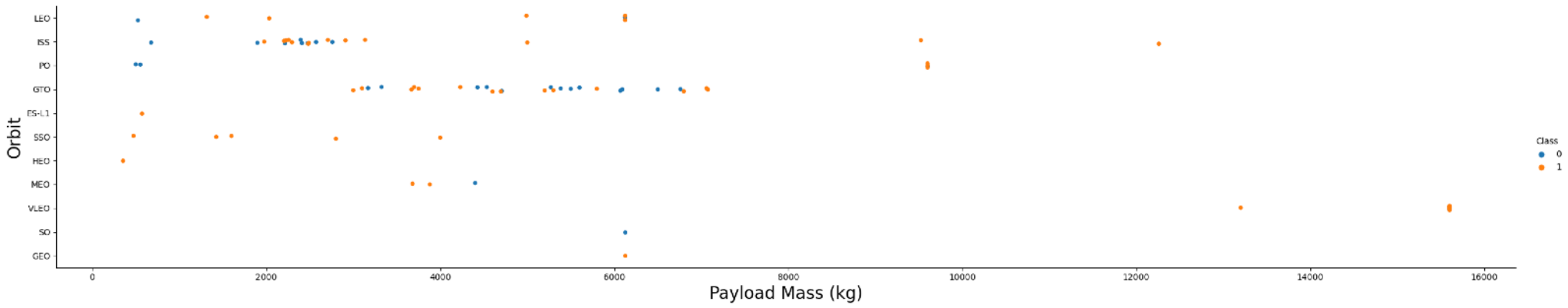
Flight Number vs. Orbit Type



Explanation:

1. The LEO orbit the Success appears related to the number of flights
2. On the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



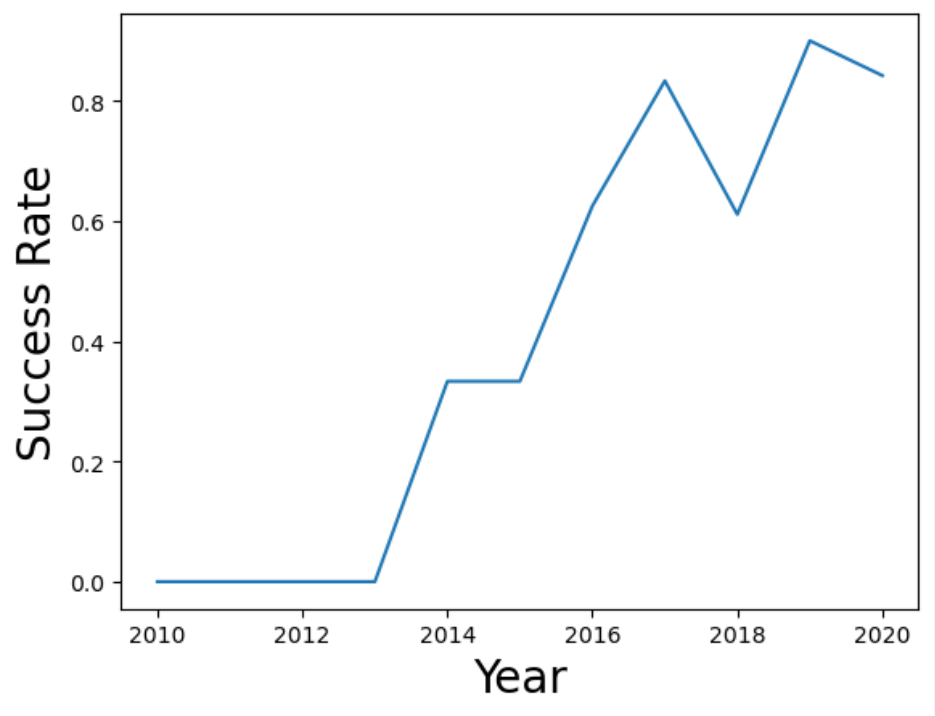
Explanation:

1. Observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits

Launch Success Yearly Trend

Explanation:

1. Observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

Task 1: Display the names of the unique launch sites in the space mission

In [10]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

Task 2: Display 5 records where launch sites begin with the string 'CCA'

In [11]:
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

Out[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Task 3: Display the total payload mass carried by boosters launched by NASA (CRS)

In [12]:
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

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

Out[12]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Task 4: Display average payload mass carried by booster version F9 v1.1

In [13]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

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

Out[13]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Task 5: List the date when the first succesful landing outcome in ground pad was acheived

In [14]:
```
%sql select min(date) as first_successful_landing from SPACEXTBL where landing_outcome = 'Success (ground pad)';
```

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

Out[14]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [15]:
```
1 ersion from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Out[15]:

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

Total Number of Successful and Failure Mission Outcomes

Task 7: List the total number of successful and failure mission outcomes

In [16]:
```
1 %sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
```

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

Out[16]:

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

Boosters Carried Maximum Payload

Task 8: List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [17]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXT
```

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

Out[17]:

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

Task 9: List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

In [18]:
```
%sql SELECT CASE strftime('%m', Date) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March'
```

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

Out[18]:

month_name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Task 10: Rank 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 [19]:
```
%sql SELECT Landing_Outcome, COUNT(*) as outcome_count FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-
```

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

Out[19]:

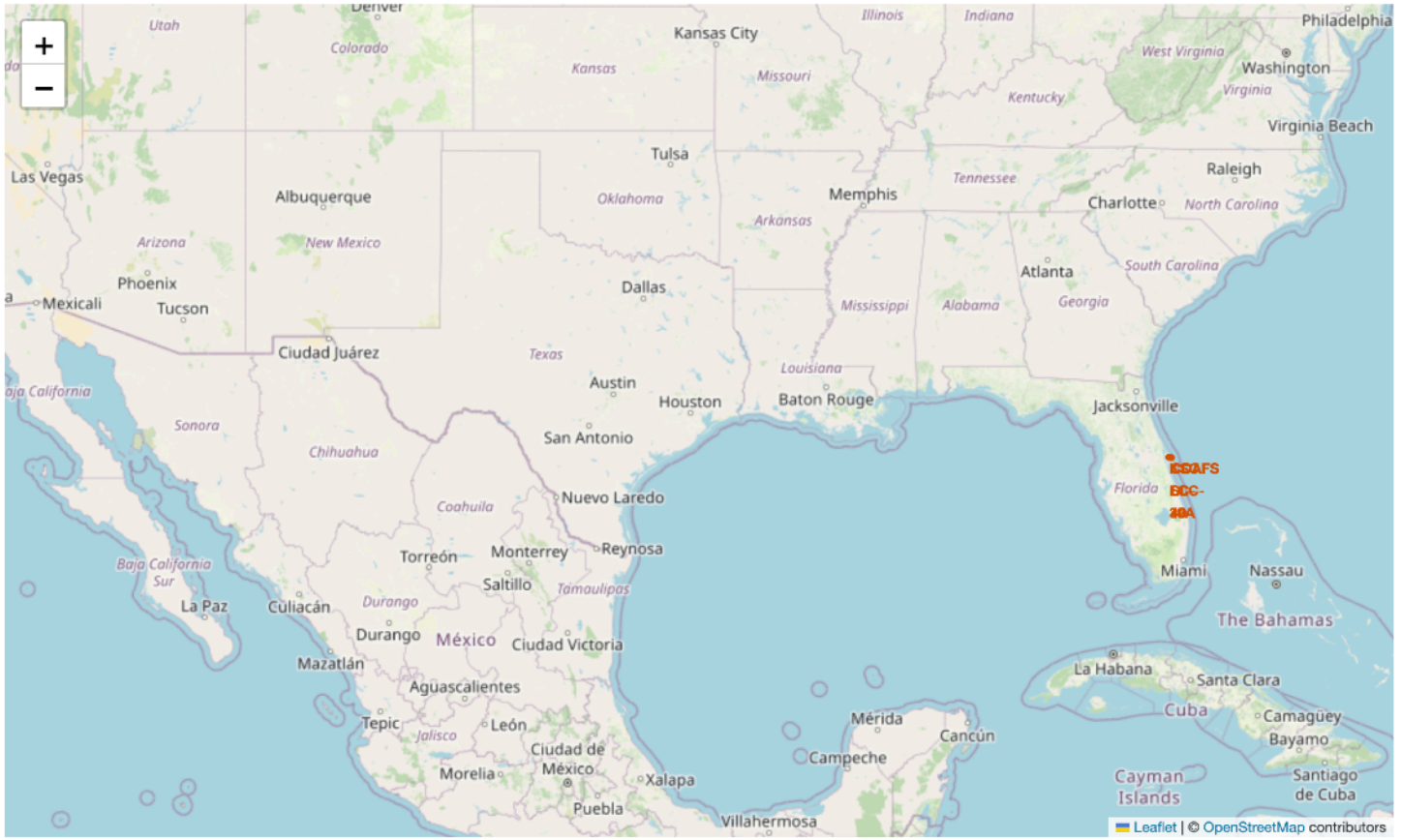
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Site Location Markers

Out[8]:



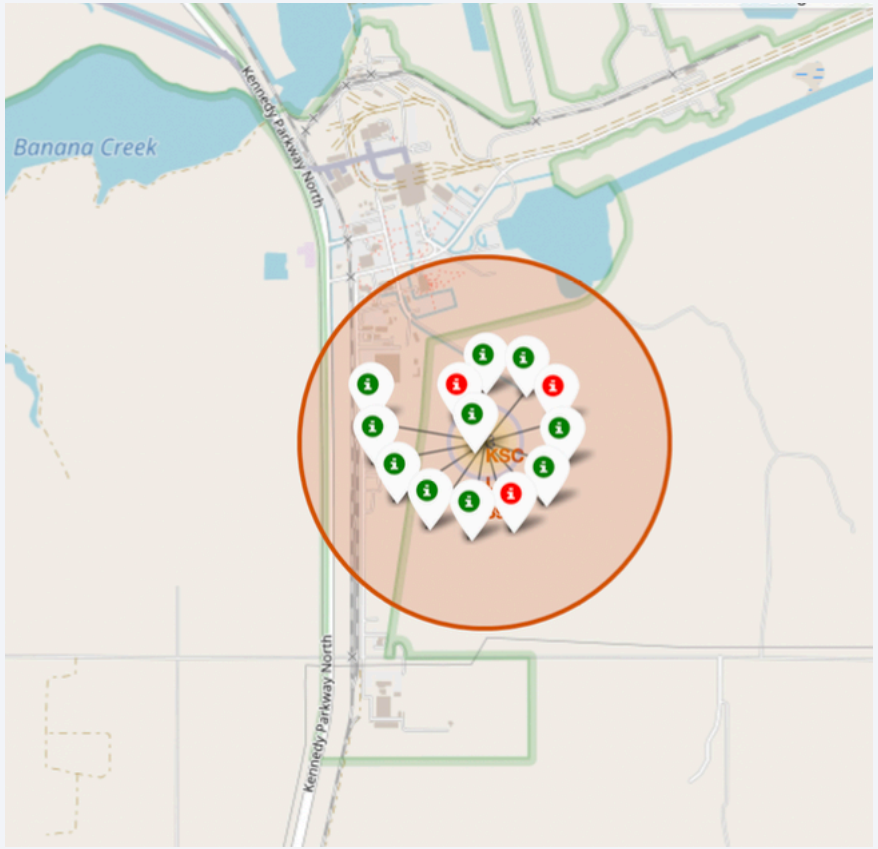
Explanation:

1. Most of Launch sites considered in this project are in proximity to the Equator line. Launch sites are made at the closest point possible to Equator line, because anything on the surface of the Earth at the equator is already moving at the maximum speed (1670 kilometers per hour). For example launching from the equator makes the spacecraft move almost 500 km/hour faster once it is launched compared half way to north pole.
2. All launch sites considered in this project are in very close proximity to the coast While starting rockets towards the ocean we minimise the risk of having any debris dropping or exploding near people.

Color Labelled Launch Records

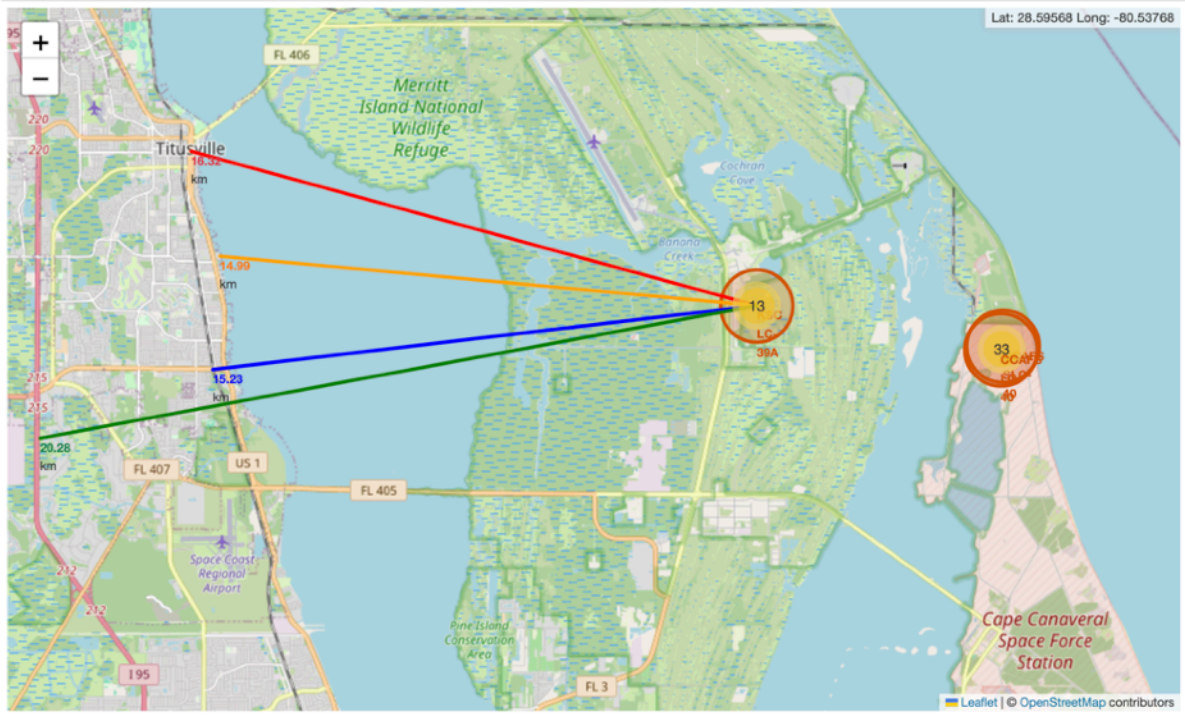
Explanation:

1. With the color markers we can identify which launch sites have relatively high success rates.
2. Green = Successful launch
3. Red = Failed launch



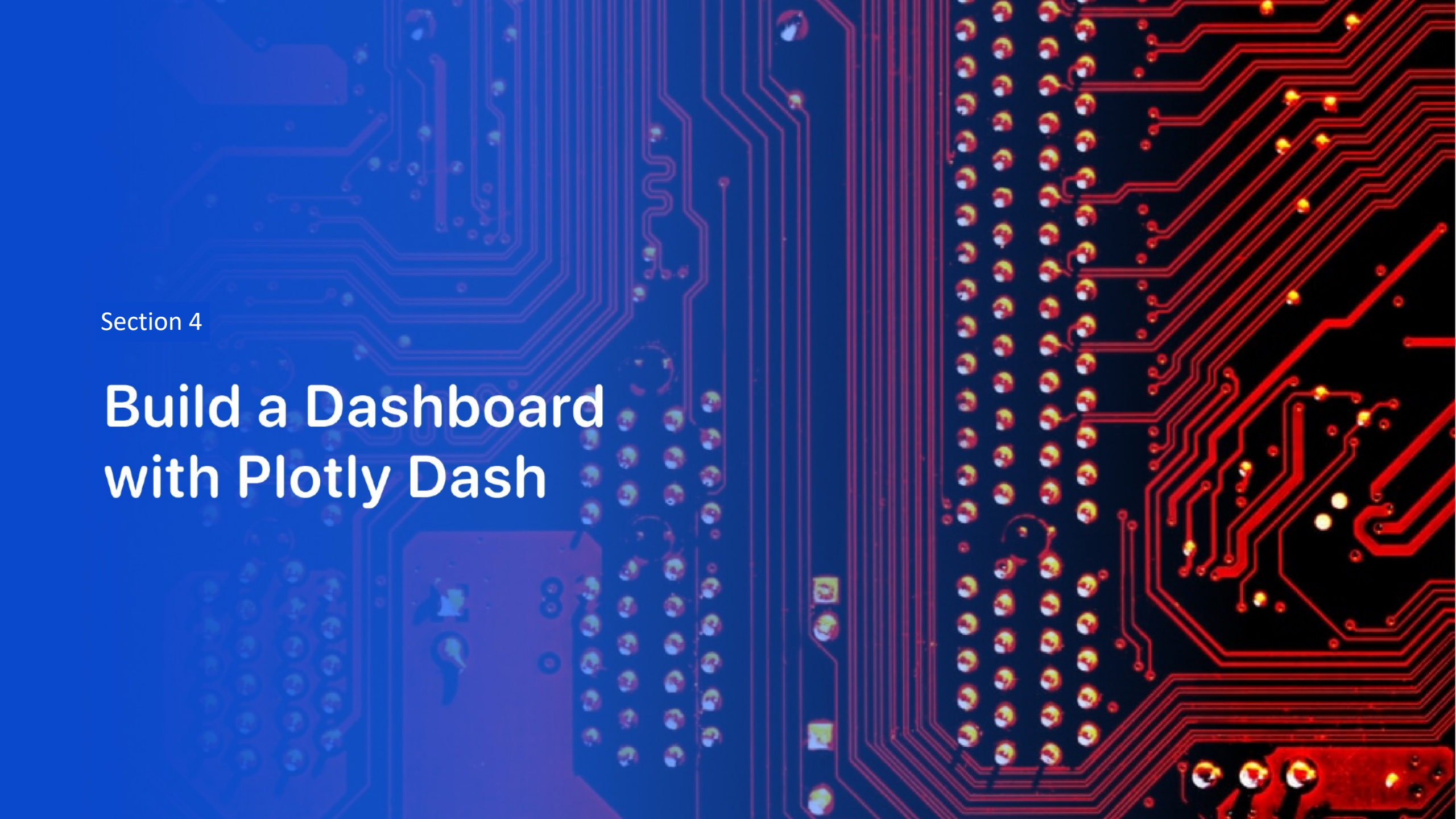
Distance from the Launch Sites

Out[19]:



Explanation:

1. From the visual analysis of the launch site KSC LC-39A we can clearly see that it is a. relative close to railway (15.23 km) b. relative close to highway (20.28 km) c. relative close to coastline (14.99 km)
2. Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km)
3. 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

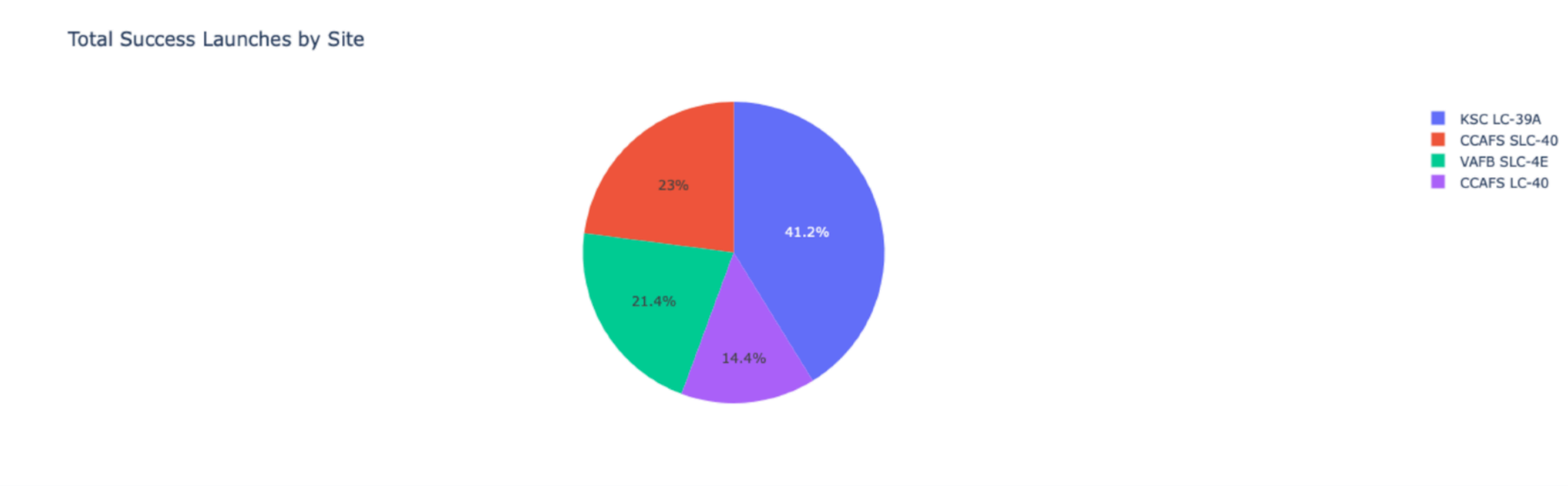
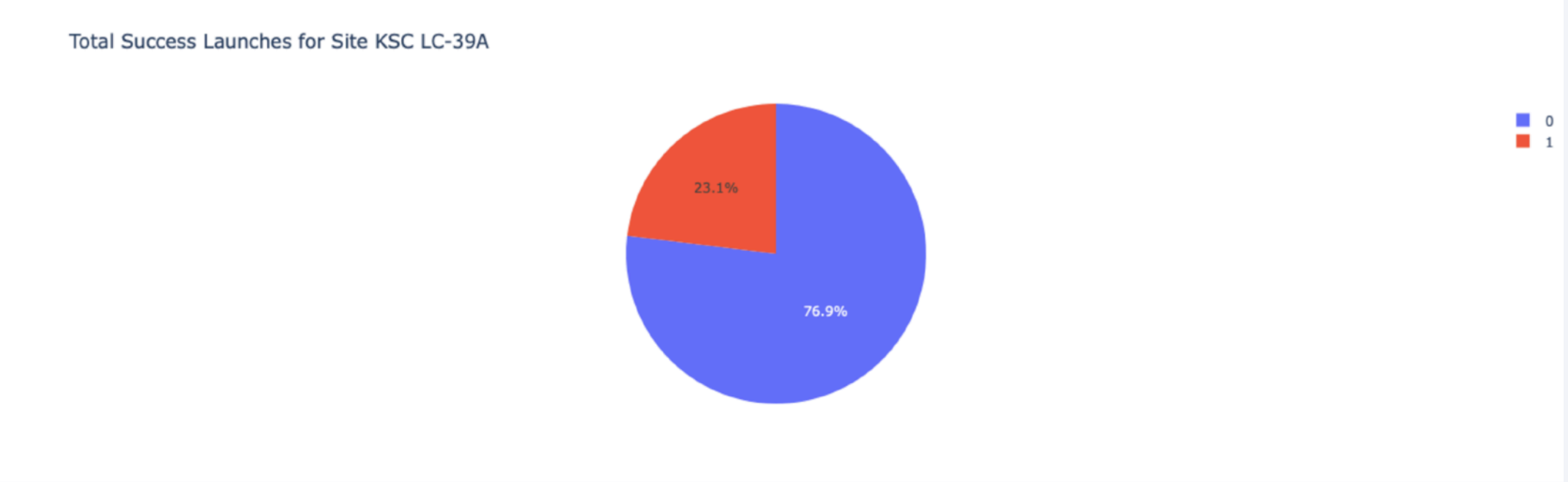


Chart shows that from all the sites, KSC LC-39A has the most successful launches

<Dashboard Screenshot 2>



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

<Dashboard Screenshot 3>

Charts shows that payload between 2,000 kg and 5,500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- From the test set scores alone, we cannot definitively determine which method performs the best.
- The identical test set scores could be a result of the small sample size (18 samples). To address this, we evaluated all methods using the entire dataset.
- The results from the full dataset indicate that the Decision Tree model is the best-performing. It not only achieved higher scores overall but also had the highest accuracy.

Out[30]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.733333	0.923077	0.923077	0.705882
F1_Score	0.846154	0.960000	0.960000	0.827586
Accuracy	0.777778	0.944444	0.833333	0.722222

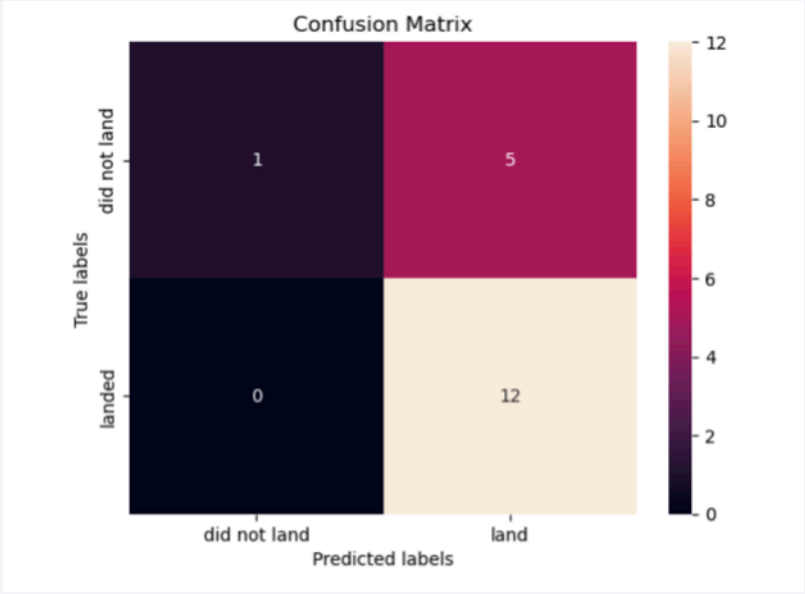
Test Set

Out[31]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.921875	0.906250	0.760000	0.740741
F1_Score	0.959350	0.950820	0.863636	0.851064
Accuracy	0.944444	0.933333	0.800000	0.766667

Whole Data Set

Confusion Matrix



By analyzing the confusion matrix, we observe that logistic regression is capable of distinguishing between the different classes. However, the primary issue lies in the number of false positives.

Conclusions

- The Decision Tree model is the most effective algorithm for this dataset.
  - Launches with lower payload mass tend to have better outcomes compared to those with larger payloads.
  - Most launch sites are near the Equator, and all are located very close to the coast.
  - The success rate of launches has improved steadily over the years.
  - KSC LC-39A boasts the highest success rate among all launch sites.
  - Orbits such as ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.
-

Appendix

https://github.com/zaedyussof88/IBM-Applied-Data-Science-Capstone

Thank you!