IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Subhajit Mondal

31-07-2025

Outline

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

Executive Summary

Summary of methodologies:

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches at a cost of $62 million, significantly lower than competitors who charge upwards of $165 million.

This cost advantage is largely due to SpaceX's ability to reuse the first-stage rocket.

If we can predict whether the first stage will successfully land, we can estimate the true cost of a launch.

This insight is valuable for alternate providers looking to compete with SpaceX in the commercial launch market.

In this project, we build a machine learning pipeline to predict first-stage landing success using historical launch data.

- Problems you want to find answers
- 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

- Data collection methodology:
  - Using SpaceX REST API
  - Using Web Scraping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
  - 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

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

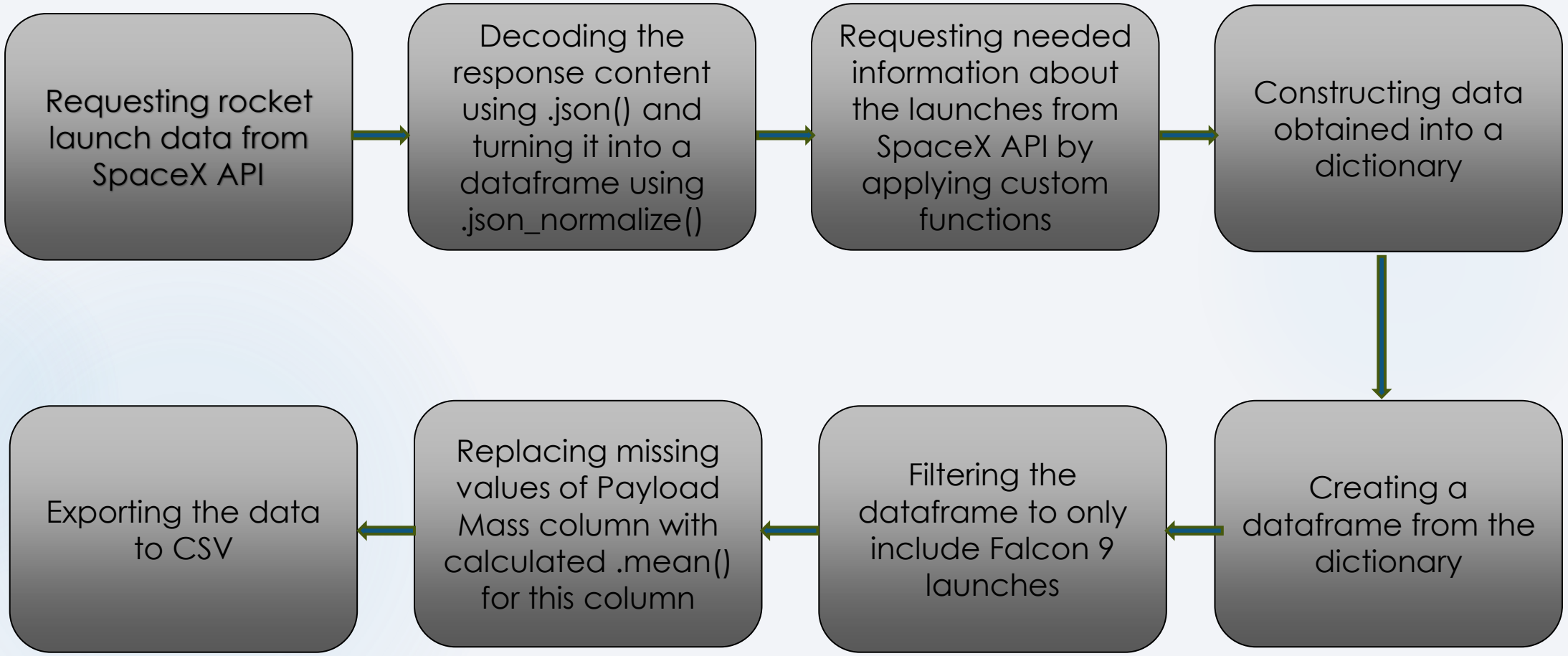
- Data Columns obtained using SpaceX REST API:

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

- Data Columns obtained using Wikipedia Web Scraping:

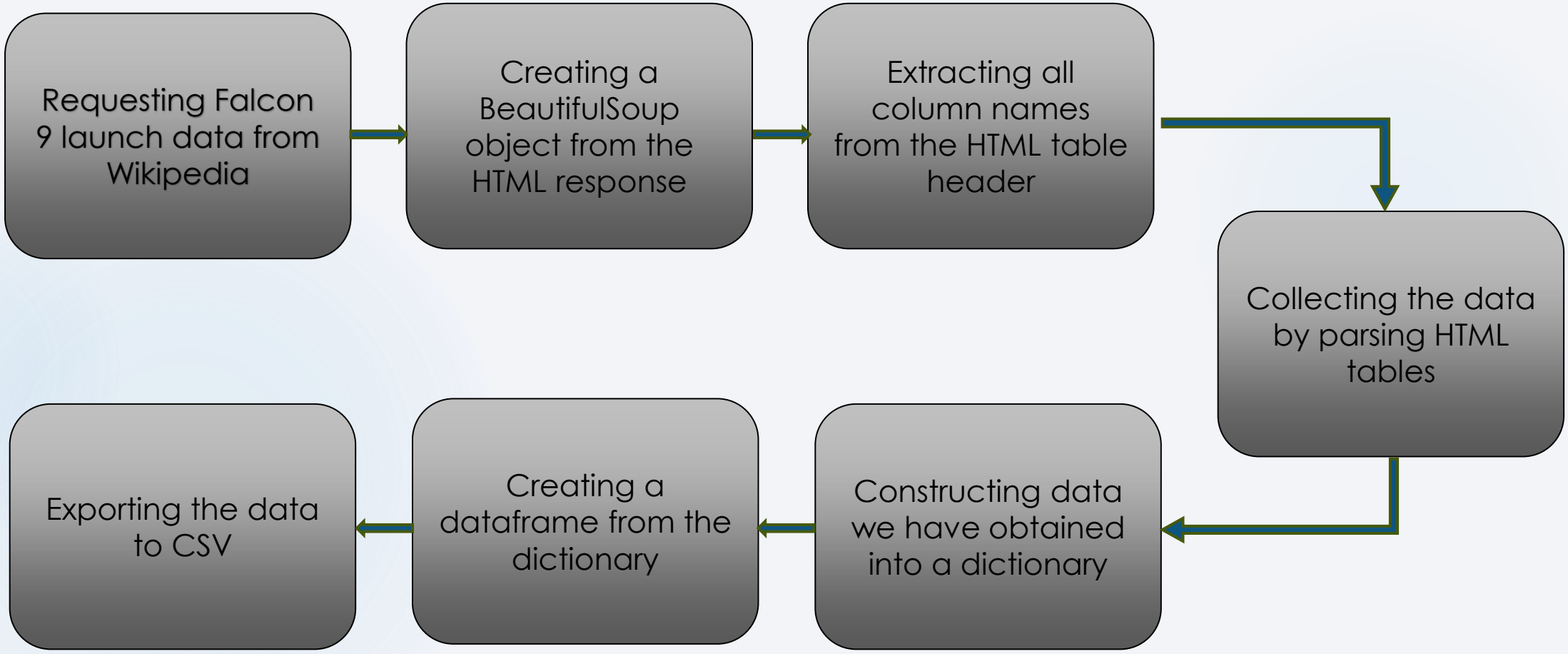
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL: Data Collection with API

Data Collection - Scraping



GitHub URL: Data Collection with Web Scraping

Data Wrangling

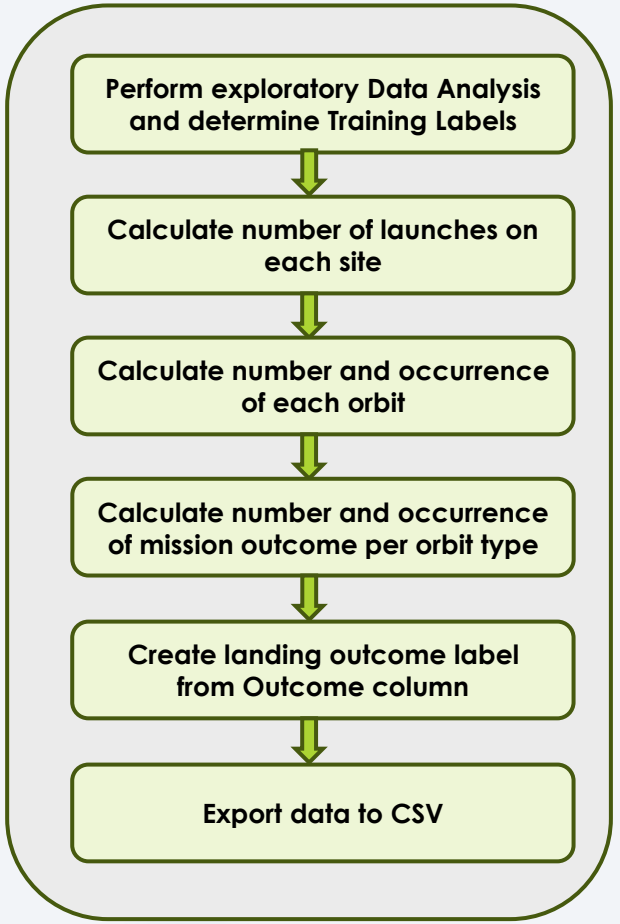
In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example:

- True Ocean means the mission outcome was successfully landed to a specific region of the ocean
- False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean
- True RTLS means the mission outcome was successfully landed to a ground pad
- False RTLS means the mission outcome was unsuccessfully landed to a ground pad
- True ASDS means the mission outcome was successfully landed on a drone ship
- False ASDS means the mission outcome was unsuccessfully landed on a drone ship

We mainly convert those outcomes into Training Labels:

- "1" means the booster successfully landed
- "0" means it was unsuccessful

Steps in Data Analysis



GitHub URL: Data Wrangling

EDA with Data Visualization

Charts were plotted

1. Flight Number vs. Payload Mass
2. Flight Number vs. Launch Site
3. Payload Mass vs. Launch Site
4. Orbit Type vs. Success Rate
5. Flight Number vs. Orbit Type
6. Payload Mass vs. Orbit Type
7. Success Rate Yearly Trend

Chart Types Explained

1. Scatter plots show the relationship between variables. If a relationship exists, they could be used in a machine learning model.
2. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
3. Line charts show trends in data over time (time series).

GitHub URL: EDA with Data Visualization

EDA with SQL

SQL queries performed

1. Displaying the names of the unique launch sites in the space mission
2. Displaying 5 records where launch sites begin with the string 'CCA'
3. Displaying the total payload mass carried by boosters launched by NASA (CRS)
4. Displaying average payload mass carried by booster version F9 v1.1
5. Listing the date when the first successful landing outcome in ground pad was achieved
6. Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. Listing the total number of successful and failure mission outcomes
8. Listing the names of the booster versions which have carried the maximum payload mass
9. Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
10. 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

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

Markers of all Launch Sites:

Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location

Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts

Coloured Markers of the launch outcomes for each Launch Site:

Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates

Distances between a Launch Site to its proximities:

Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

GitHub URL: Interactive Visual Analytics 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:

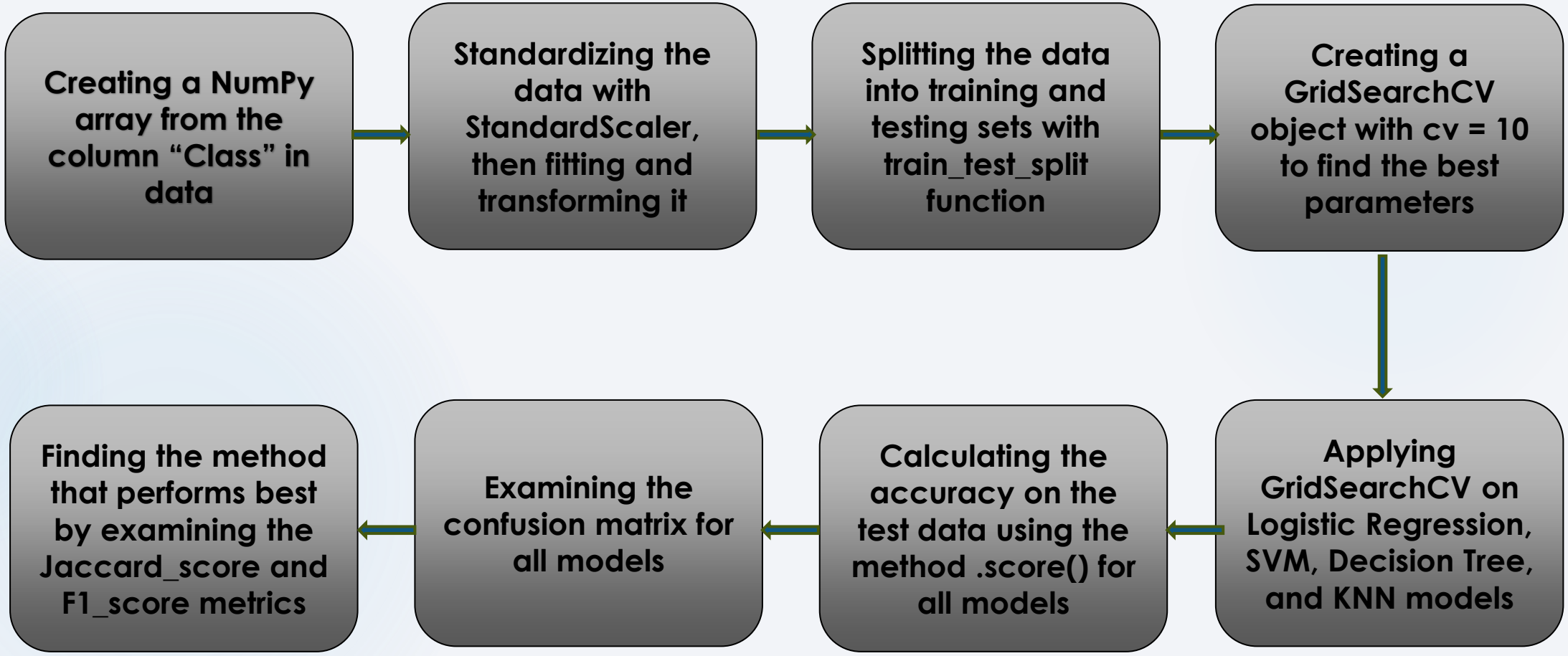
- Added a slider to select Payload range

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
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

GitHub URL: SpaceX Dash App

Predictive Analysis (Classification)



GitHub URL: Machine Learning Prediction

Results

Exploratory data analysis results

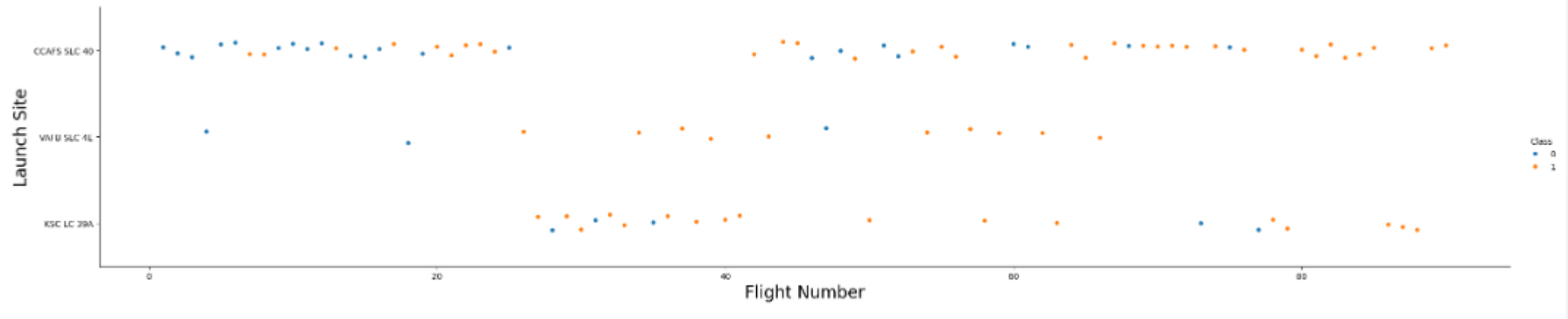
Interactive analytics demo in screenshots

Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



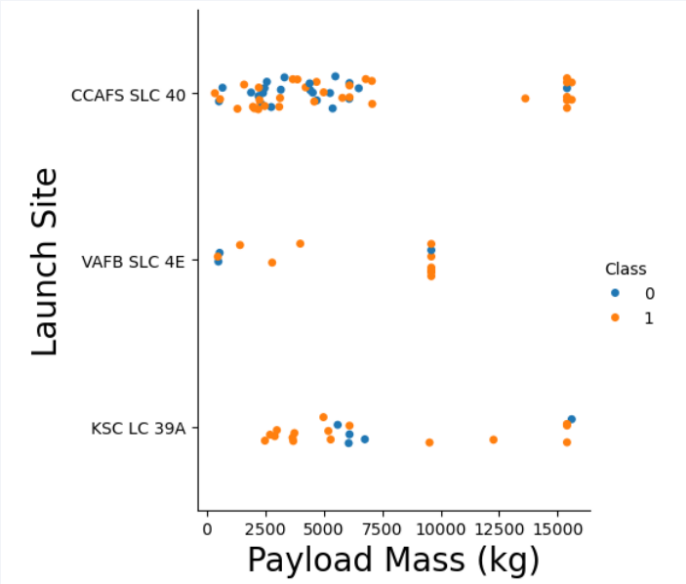
Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about 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

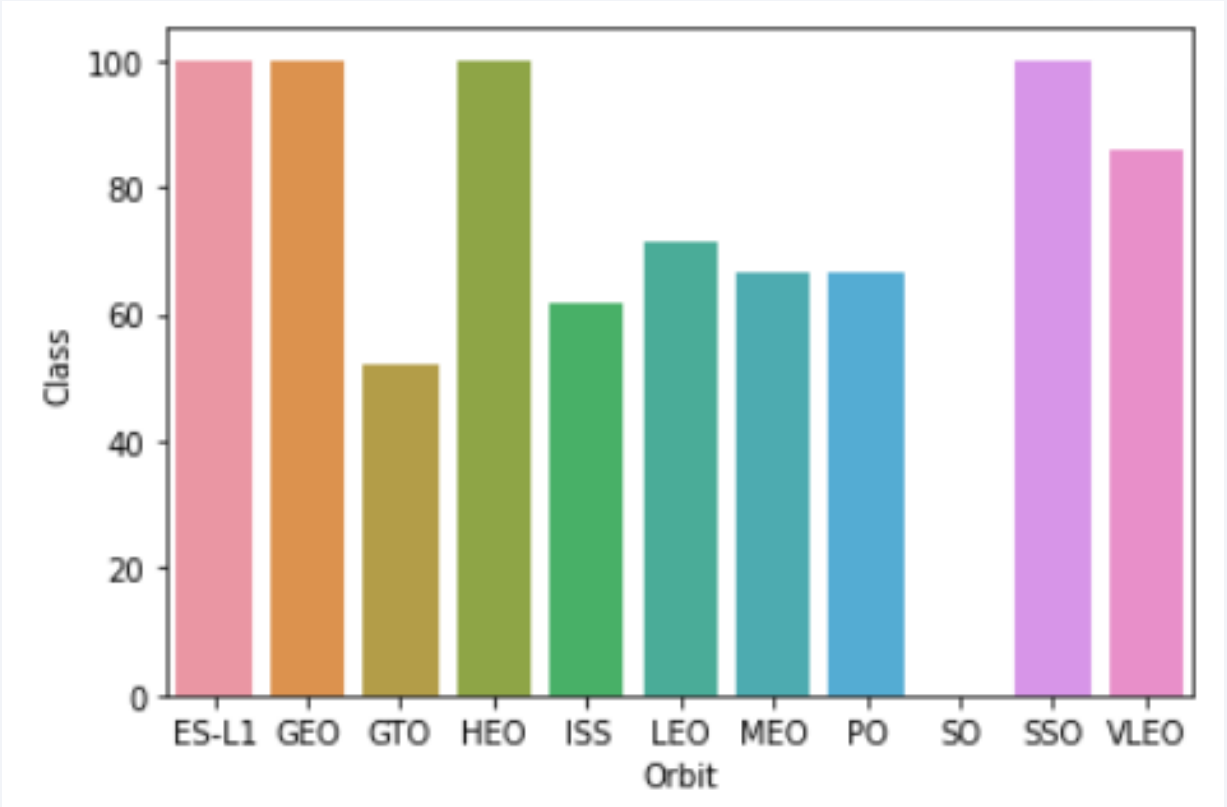
Explanation:

- 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.

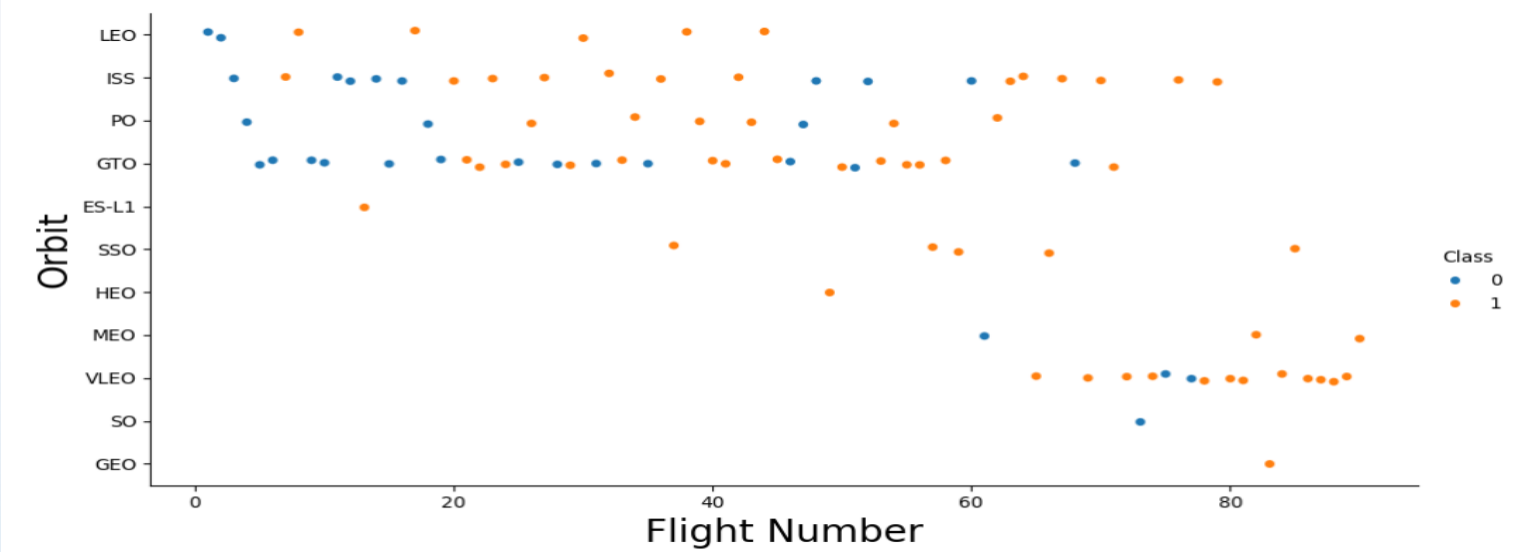


Success Rate vs. Orbit Type

- 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



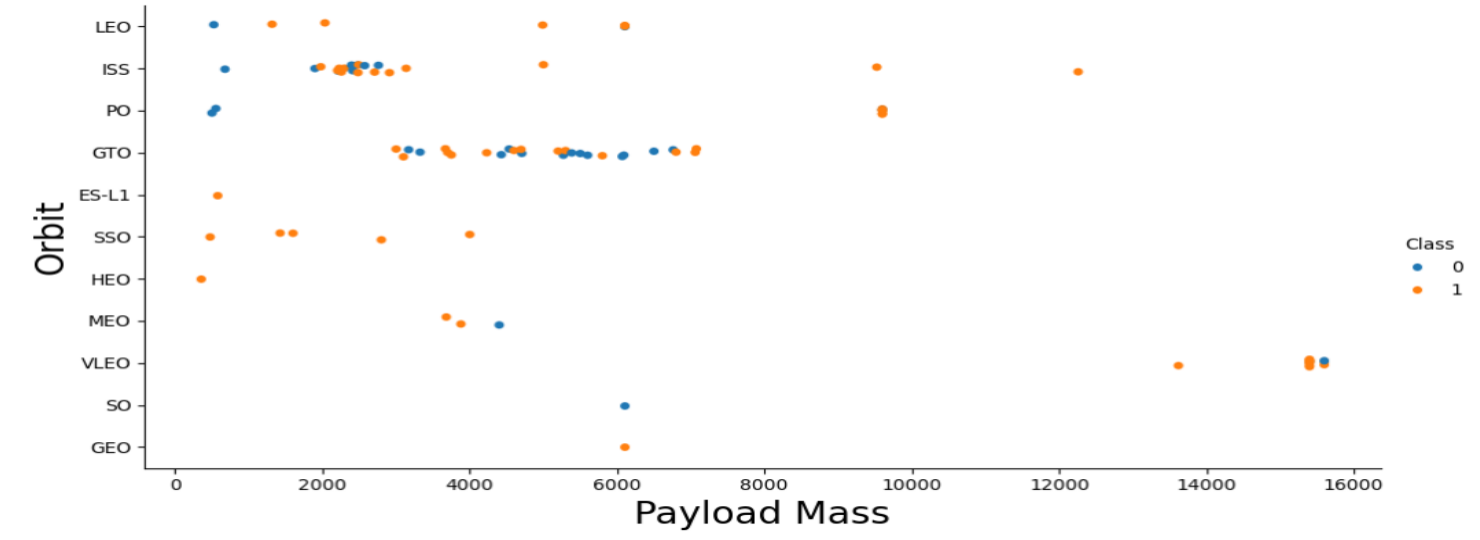
Flight Number vs. Orbit Type



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



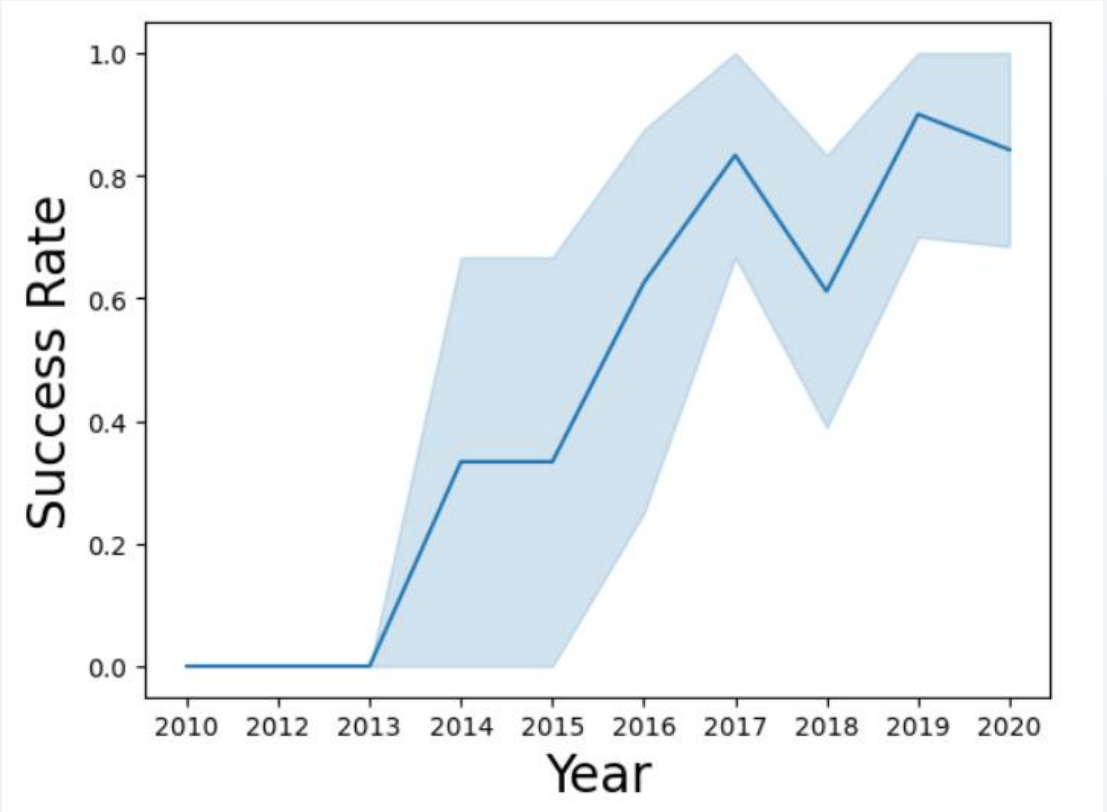
Explanation:

- Heavy payloads have a negative influence on GTO orbits
- And a positive influence 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


```
%sql SELECT DISTINCT Launch_Site from SPACEXTABLE;
```

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

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'


```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

Explanation:

- Displaying 5 records where launch sites begin with the string ‘CCA'

Total Payload Mass


```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" Like 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
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 AVG_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 V1.1%';
```

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

AVG_Payload_Mass
2534.6666666666665

Explanation:

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

First Successful Ground Landing Date


```
%sql Select MIN(Date) from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)'
```

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

MIN(Date)
2015-12-22

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


```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" BETWEE
```

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

Booster_Version
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


```
%sql SELECT Mission_Outcome, Count(*) AS TOTAL_NUMBERS FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

Mission_Outcome	TOTAL_NUMBERS
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


```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

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

2015 Launch Records

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:

- 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


```
%sql SELECT Landing_Outcome, Count(*) as Count_Outcome FROM SPACEXTABLE WHERE Date between '2010-06-04' and '2017-03-20' GR(
```

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

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

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 ascending order

Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on a global map

Explanation:

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to coast, while launching rockets towards ocean it minimises risk of having any debris dropping or exploding near people.

Colour-labeled launch records on the map

Explanation:

- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

  Green Marker = Successful Launch

  Red Marker = Failed Launch

- Launch Site KSC LC-39A has a very high Success Rate.

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

Explanation:

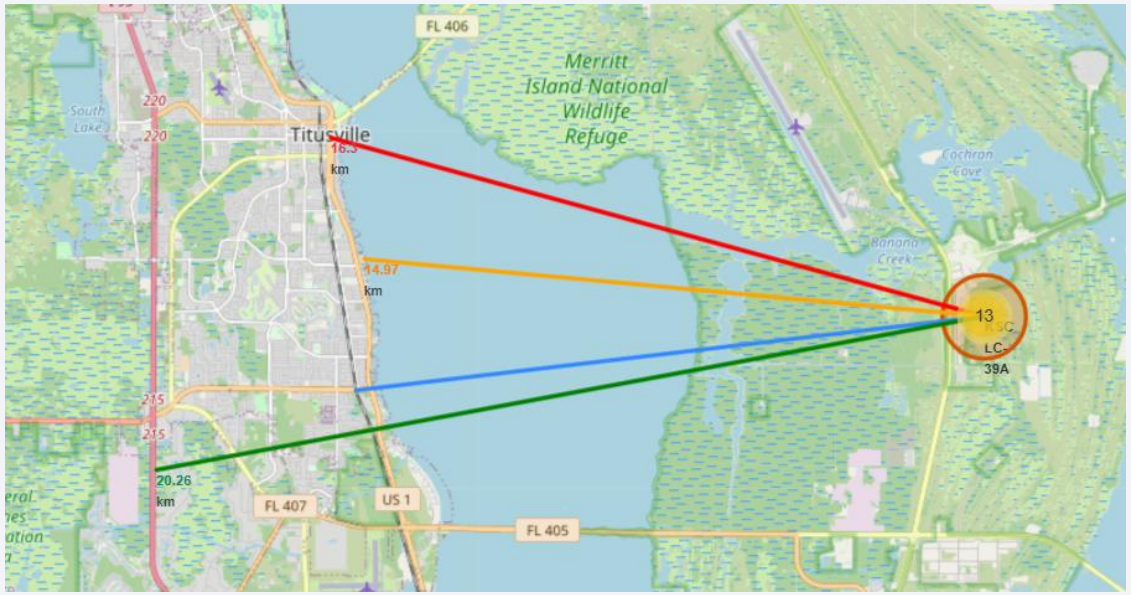
- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

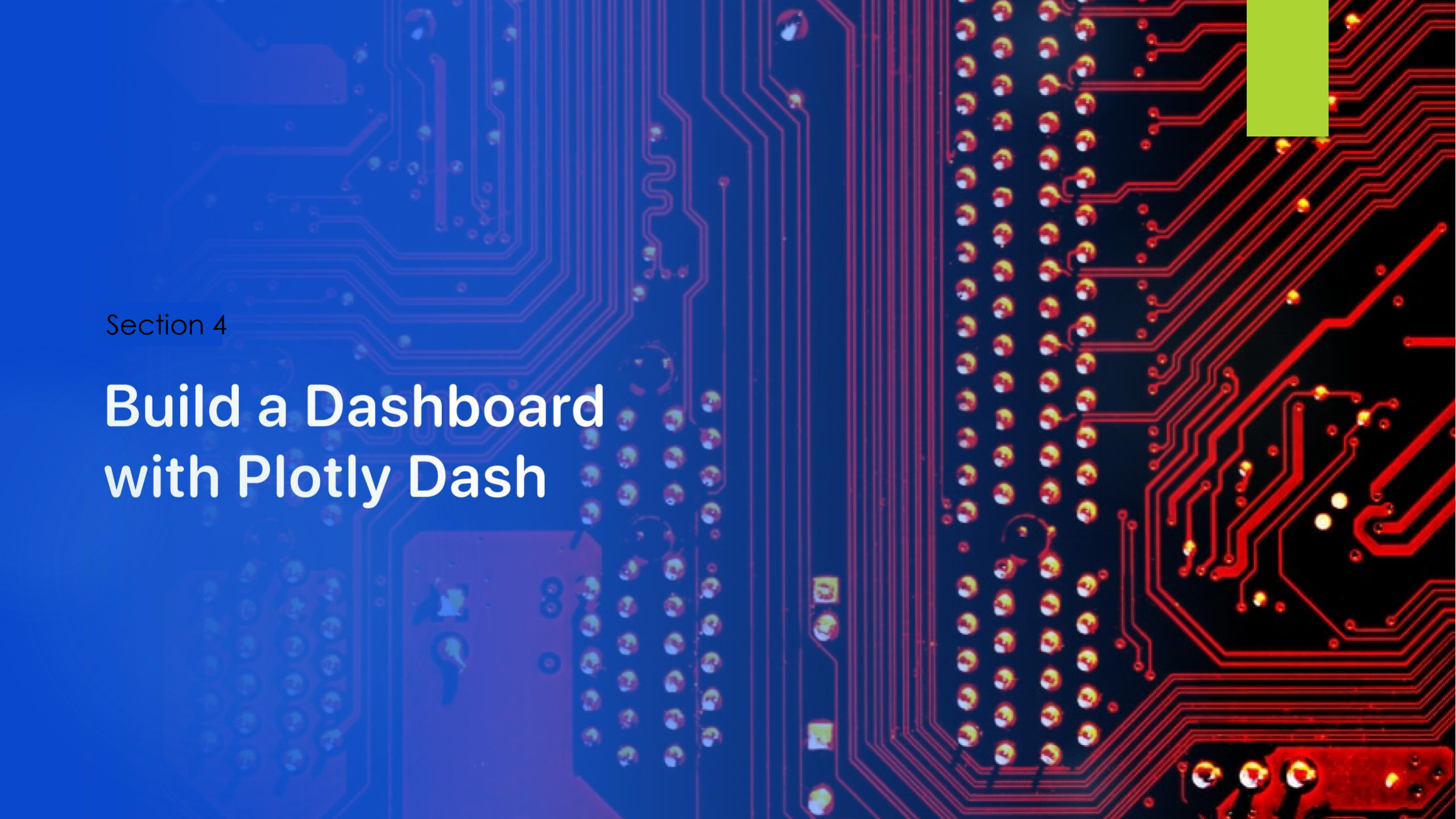
  relative close to railway (15.23 km)

  relative close to highway (20.28 km)

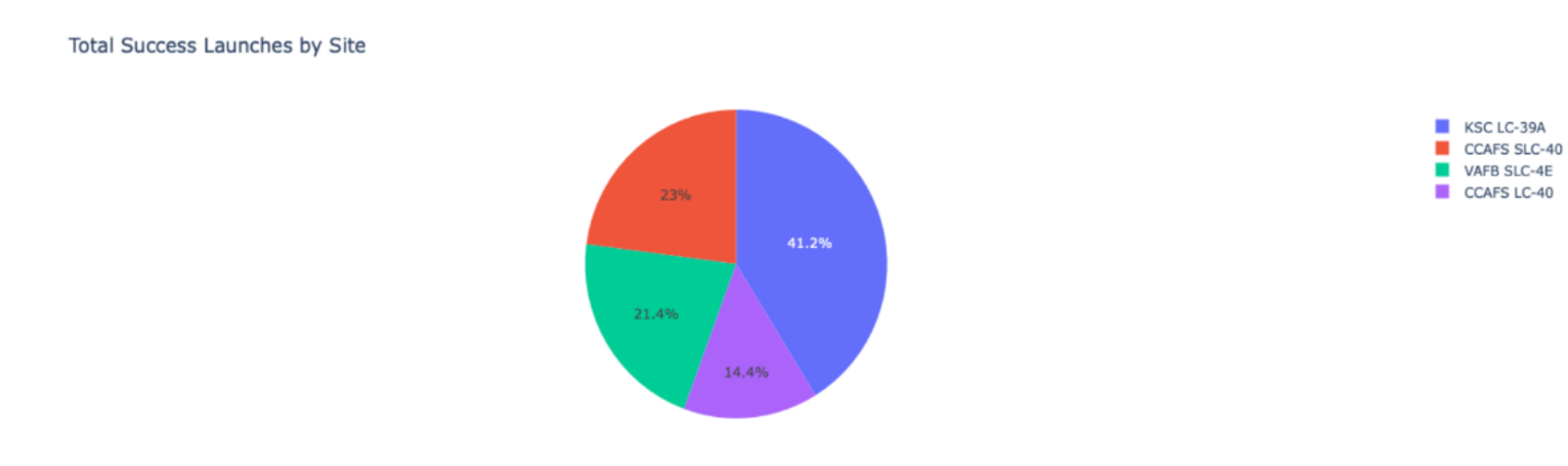
  relative close to coastline (14.99 km)

- 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

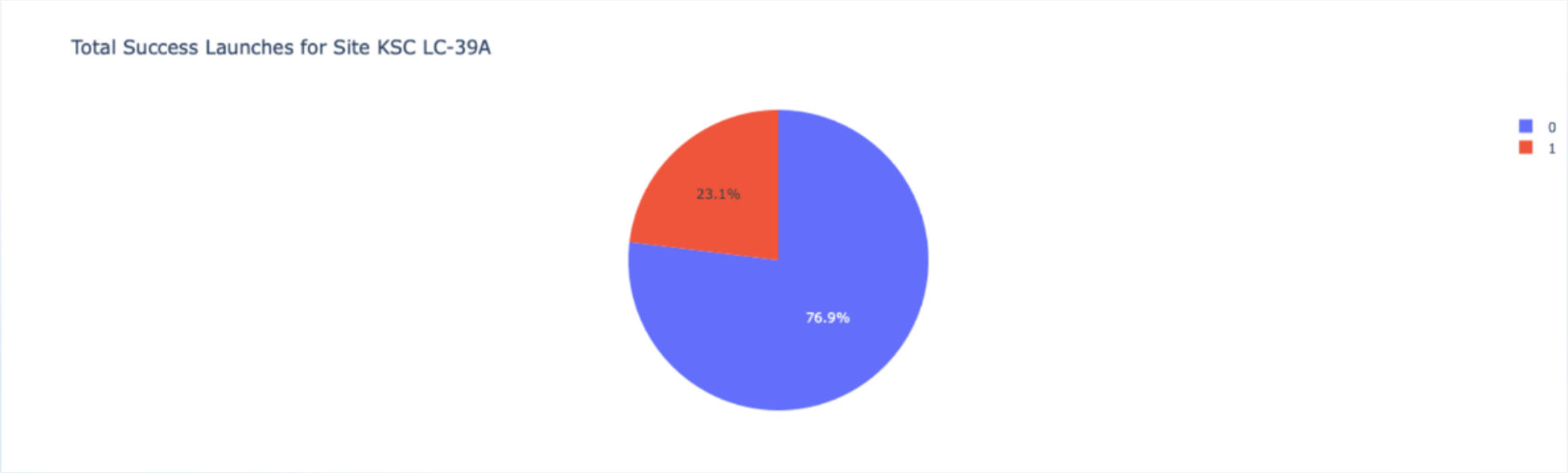
Launch success count for all sites



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

Launch site with highest launch success ratio



Explanation:

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

Payload Mass vs. Launch Outcome for all sites

Explanation:

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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

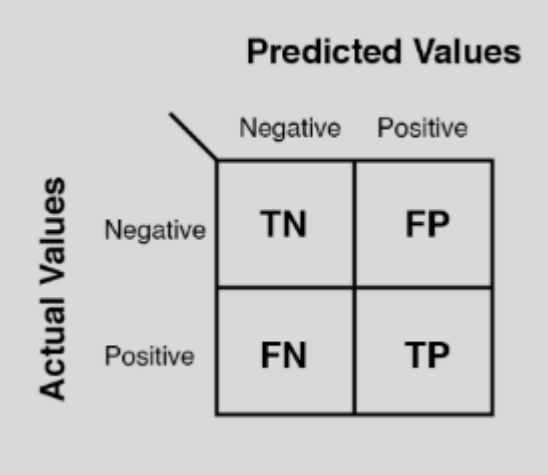
- Based on the scores of the Test Set, Logistic Regression, SVM, and KNN models show identical performance across all metrics.
- This uniformity may be due to the small test sample size (18 samples), which limits the ability to distinguish between models.
- Therefore, we evaluated all models using the entire dataset for a more reliable comparison.
- The results indicate that Logistic Regression, SVM, and KNN outperform the Decision Tree model in terms of accuracy, precision, recall, and F1 score.
- Despite its interpretability, the Decision Tree model has the lowest scores among the tested models.

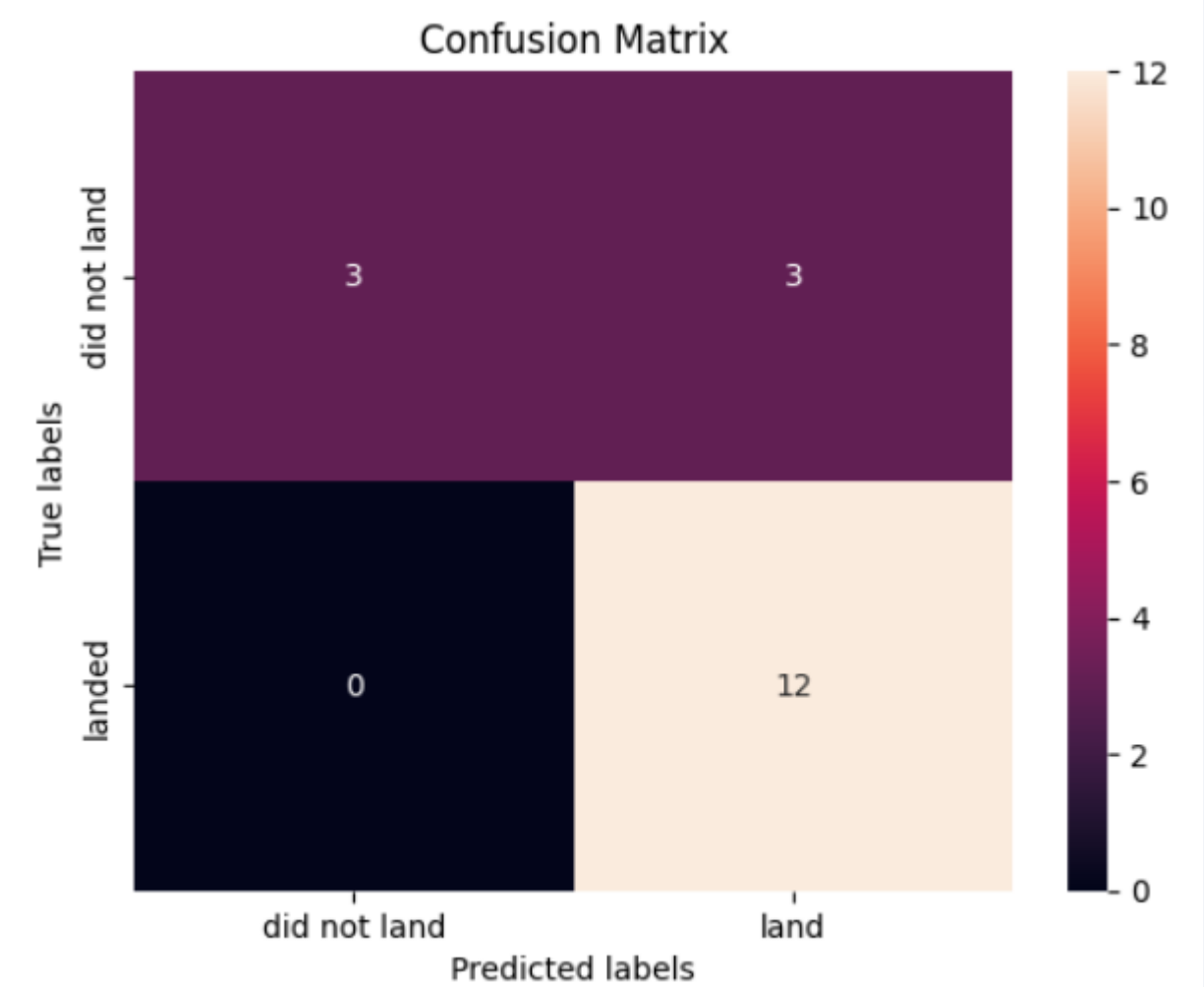
	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.833333	0.866667	0.833333	0.814815
1	SVM	0.833333	0.866667	0.833333	0.814815
2	KNN	0.833333	0.866667	0.833333	0.814815
3	Decision Tree	0.777778	0.773810	0.777778	0.764103

Confusion Matrix

Explanation:

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes, we can see that the major problem is false positives.





Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most launch sites are in proximity to the Equator line, and all 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

Special Thanks to

IBM

Coursera

IBM Skills Network

Instructors

IBM Skills Network Team, Dr. Pooja, Abhishek Gagneja, Romeo Kienzler, Joseph Santarcangelo, Polong Lin, Alex Aklson, Rav Ahuja, Saishruthi Swaminathan, Hima Vasudevan, Azim Hirjani, Aije Egwaikhide, Yan Luo, Svetlana Levitan, Jeff Grossman

Thank you!