IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeny Rachel Samuel
4/12/2024

Outline

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

Executive Summary

- In this Capstone Project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms
- The main steps in this project include:
  - Data collection, wrangling and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- Our graphs show that some features of the rocket launches have correlation with the outcome of the launches ie, success or failure

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants a bid against SpaceX for a rocket launch.
- Many unsuccessful landings are actually intentional. Sometimes, SpaceX chooses to execute a controlled landing in the ocean.
- The primary question we're addressing is this: Given a set of features regarding the Falcon 9 rocket launch, including its payload mass, orbit type, and other factors, will the first stage of the rocket successfully land?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping
- 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
  - Pandas and Nump
  - SQL

Methodology

Executive Summary

- Data Visualization
  - Matplotlib and Seaborn
  - Folium
  - Dash
- Perform predictive analysis using classification models
  - Logistic Regression
  - Support Vector Machine(SVM)
  - Decision Tree
  - K-Nearest Neighbour(KNN)

Data Collection

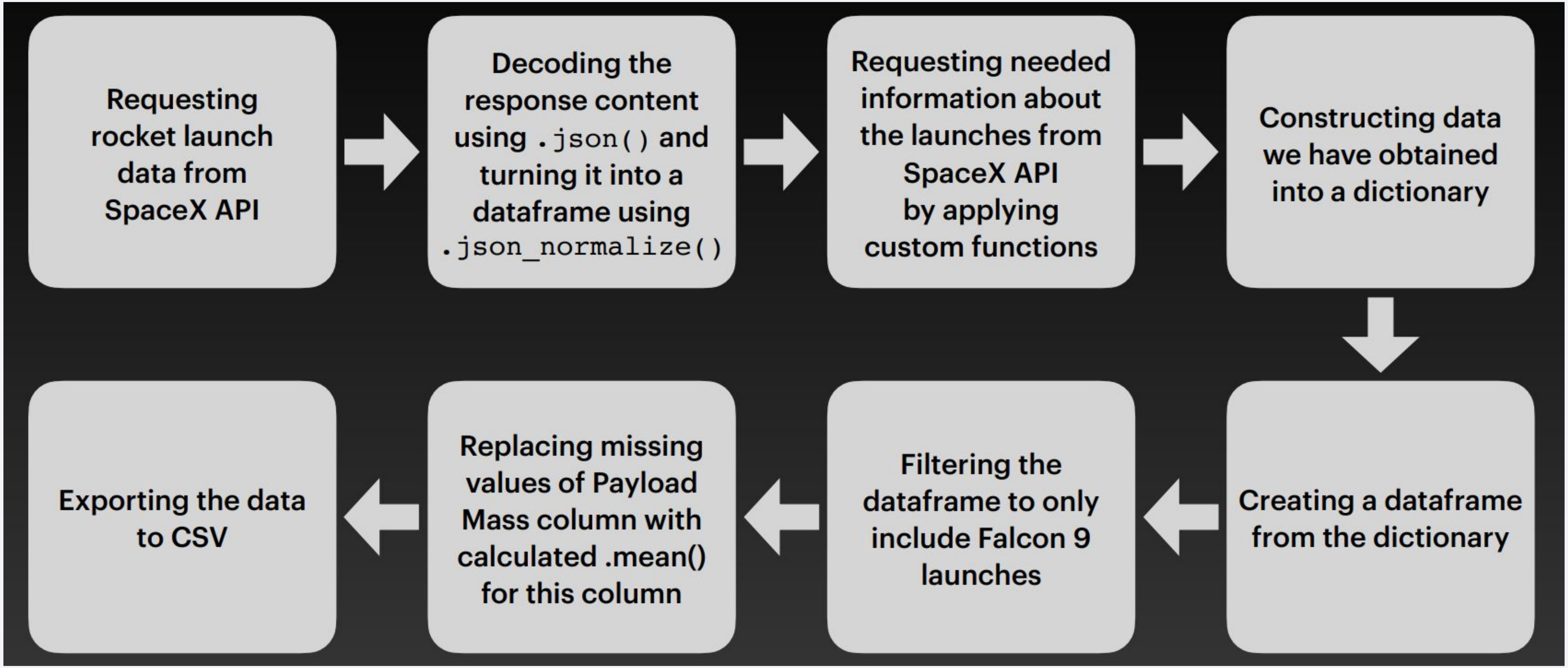
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

- 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 are obtained by using SpaceX REST API:
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

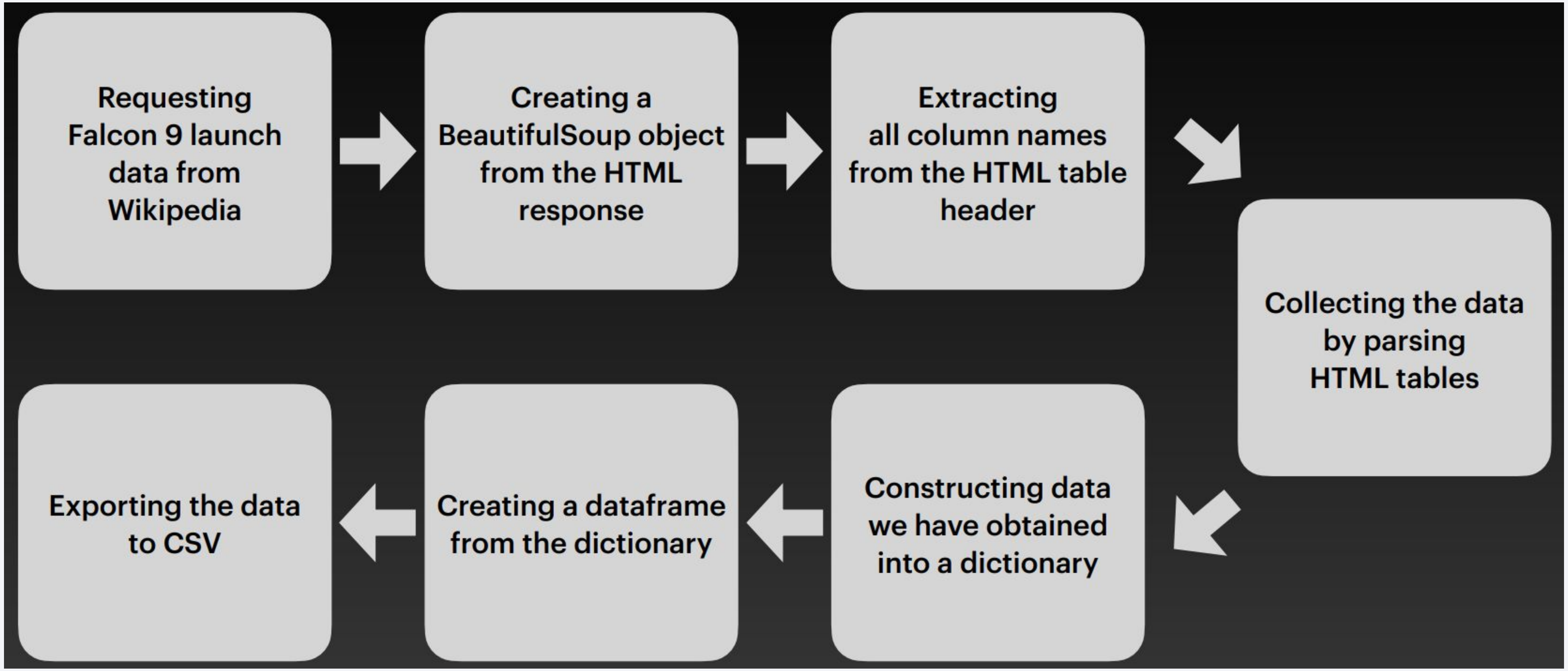
- Data Columns are obtained by using Wikipedia Web Scraping:
    - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub Link For Data Collection SpaceX API:
[SpaceX API]

Data Collection - Scraping



GitHub Link For Data Collection Scraping: [Web Scraping]

Data Wrangling

- In the dataset, there are several different scenarios where the booster did not land successfully. Sometimes, a landing was attempted but failed due to an accident. For instance, "True Ocean" indicates that the mission outcome was successfully landed in a specific region of the ocean, while "False Ocean" means the mission outcome was unsuccessfully landed in a specific region of the ocean. "True RTLS" signifies that the mission outcome was successfully landed on a ground pad, whereas "False RTLS" denotes an unsuccessful landing on a ground pad. Similarly, "True ASDS" indicates a successful landing on a drone ship, while "False ASDS" indicates an unsuccessful landing on a drone ship.
- Our main task is to convert these outcomes into training labels: "1" signifies the booster successfully landed, while "0" indicates it was unsuccessful.

GitHub Link: [Data Wrangling]

EDA with Data Visualization

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

GitHub Link: EDA with Data Visualization

EDA with SQL

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

Build an Interactive Map with Folium

- Markers have been added for all Launch Sites. For instance
    - Marker with a Circle, Popup Label, and Text Label of NASA Johnson Space Center has been included, utilizing its latitude and longitude coordinates as the starting location.
    - Markers with Circles, Popup Labels, and Text Labels for all Launch Sites have been added to display their geographical positions and proximity to the Equator and coastlines.
- Colored Markers representing the launch outcomes have been incorporated for each Launch Site.
    - Green for successful launches
    - Red for failed ones.
    - Marker Cluster has been used to identify Launch Sites with relatively high success rates
- Furthermore, colored Lines have been added to illustrate the distances between Launch Site KSC LC-39A (as an example) and its proximities, such as Railway, Highway, Coastline, and the Closest City.

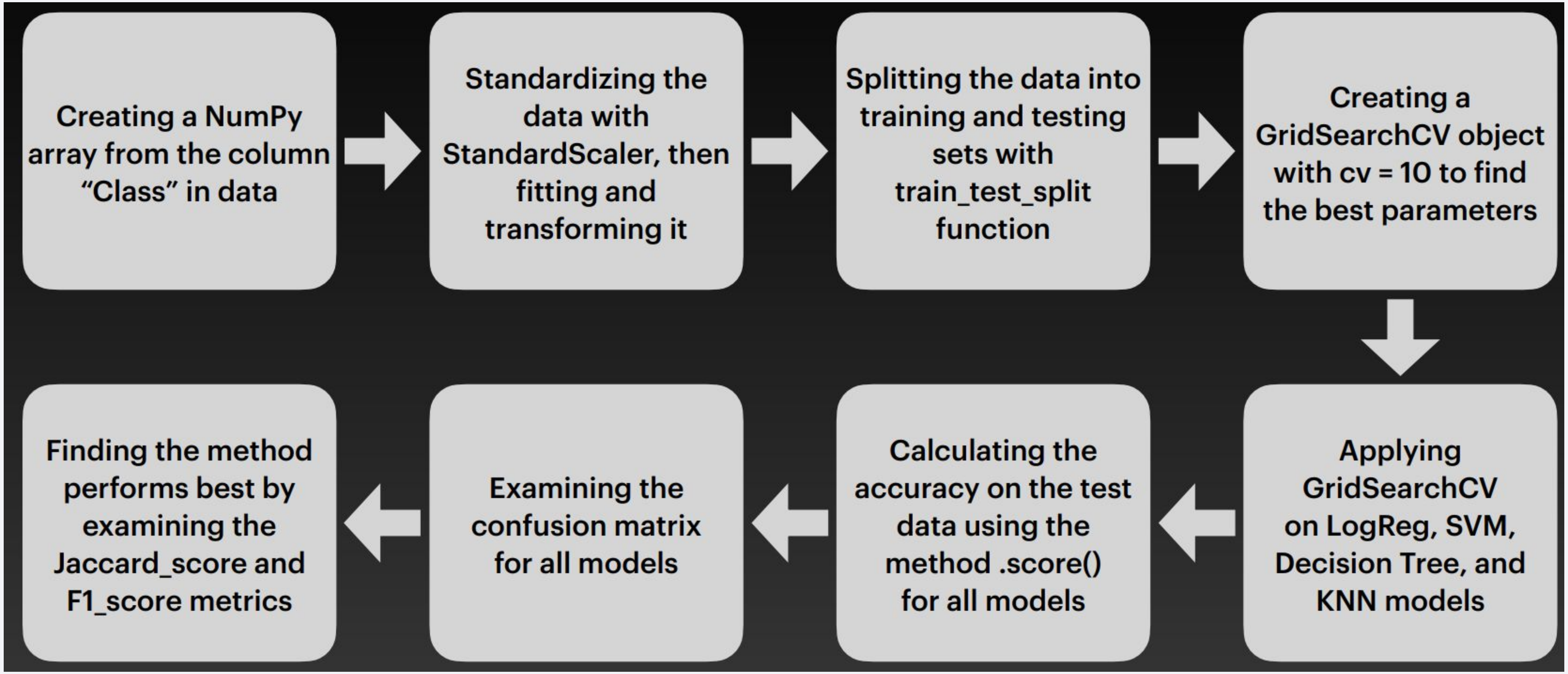
GitHub Link: [Interactive Map with Folium]

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success

GitHub Link: [Dashboard with Plotly Dash]

Predictive Analysis (Classification)



GitHub Link: [Predictive Analysis]()

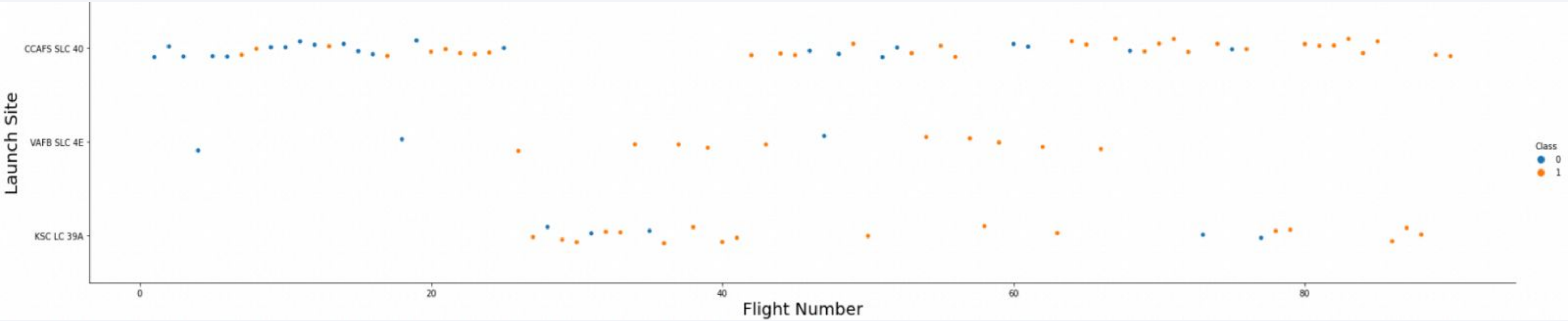
Results

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

Section 2

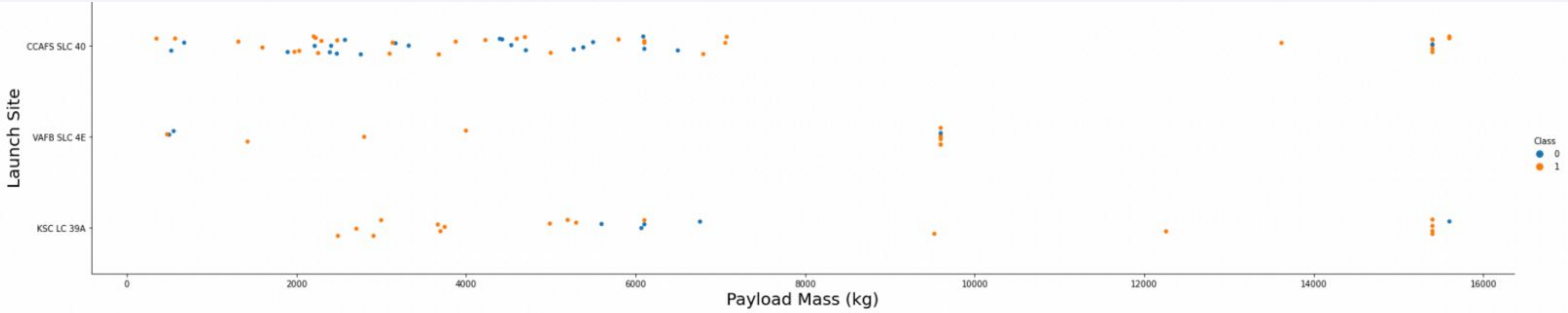
Insights drawn from EDA

Flight Number vs. Launch Site



- Explanation:
- The earliest flights all failed while the latest flights all succeeded.
- It can be assumed that each new launch has a higher rate of success

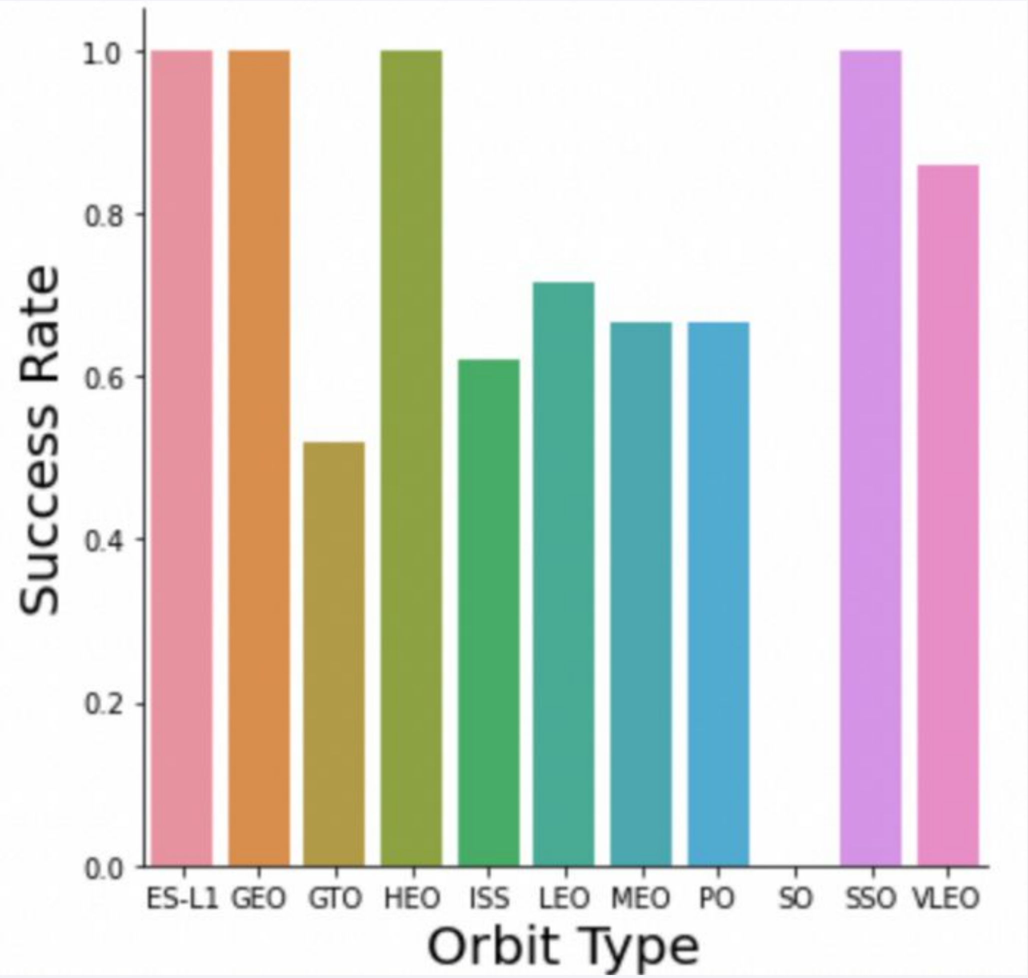
Payload vs. Launch Site



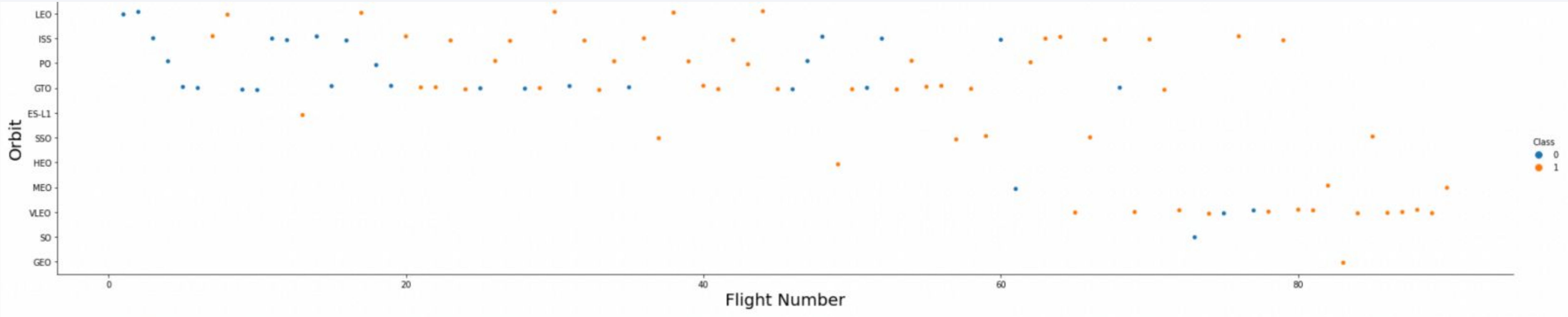
- Explanation:
  - For every launch site the higher the payload mass, the higher the success rate.
  - Most of the launches with payload mass over 7000 kg were successful.
  - KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

Success Rate vs. Orbit Type

- Explanation:
  - Orbits with 100% success rate:
    - ES-L1, GEO, HEO, SSO
  - Orbits with 0% success rate:
    - SO
  - Orbits with success rate between 50% and 85%:
    - GTO, ISS, LEO, MEO, P

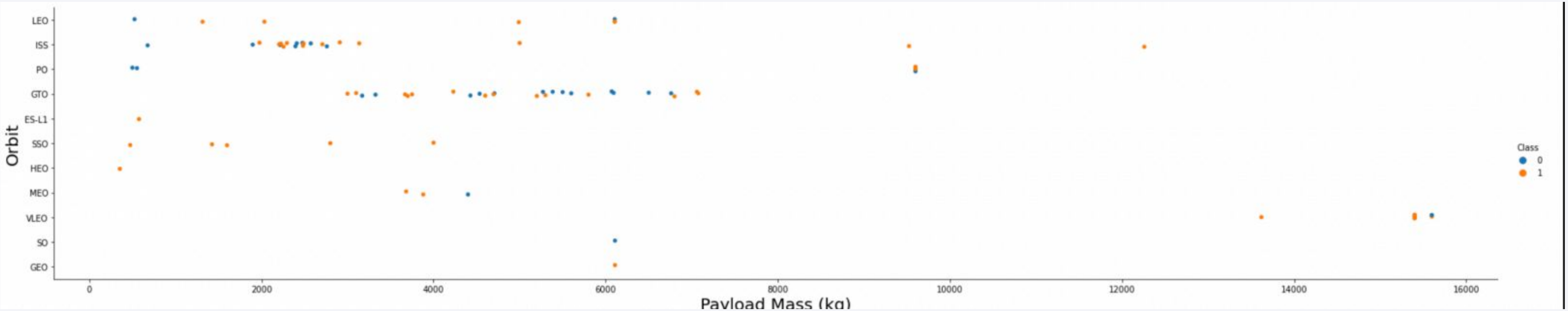


Flight Number vs. Orbit Type



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

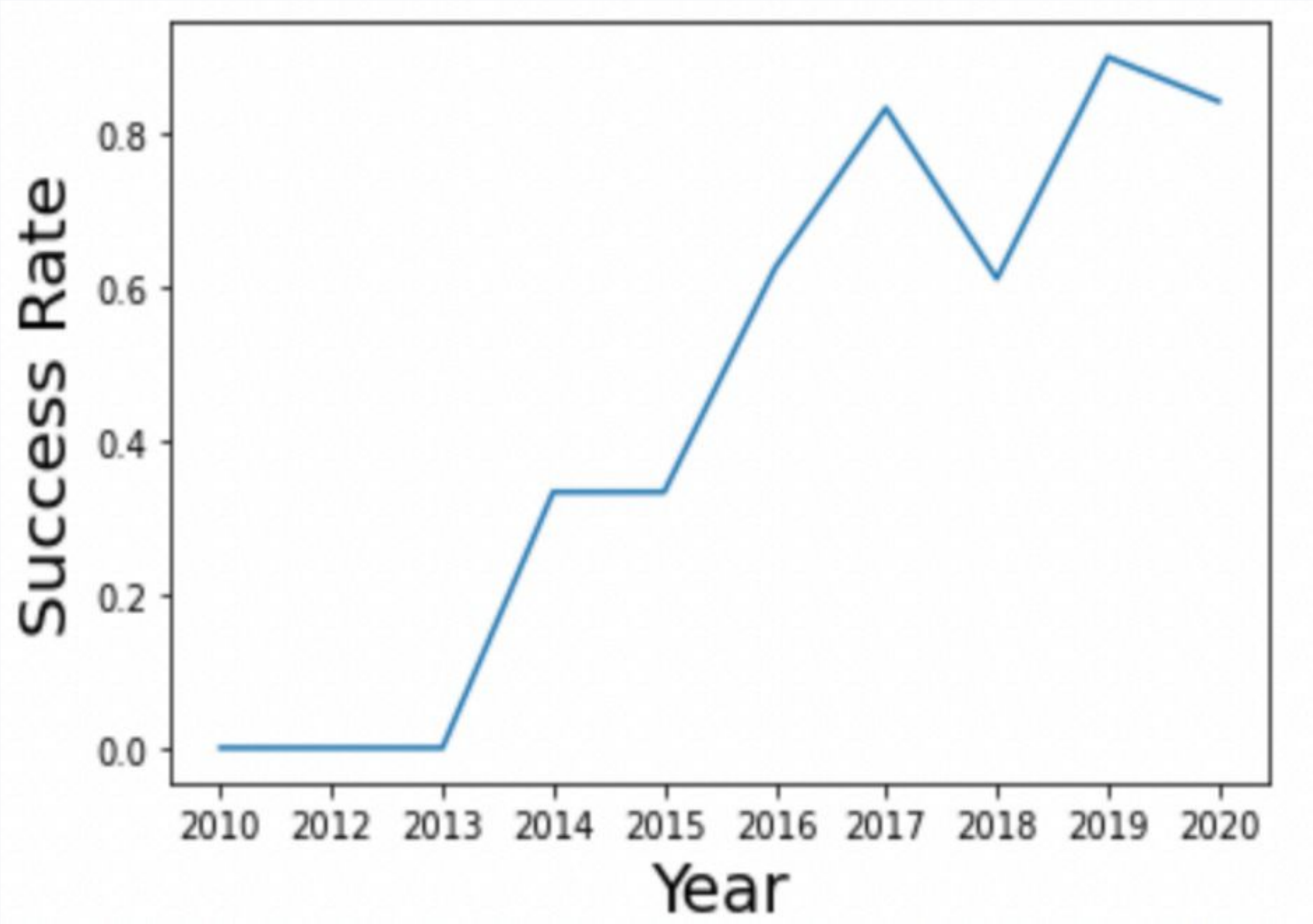
Payload vs. Orbit Type



- Explanation: •
    - Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits

Launch Success Yearly Trend

- Explanation:
  - The success rate since 2013 kept increasing till 2020



All Launch Site Names

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

In [16]:
```
%%sql
SELECT DISTINCT launch_site FROM SPACEXTBL;
```

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

Out[16]:

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

Launch Site Names Begin with 'CCA'

In [17]:

```
%%sql
SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;
```

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

Out[17]:

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

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

Total Payload Mass

In [18]:

```
%%sql
SELECT sum(payload_mass__kg_) AS "Total payload mass (NASA (CRS))" FROM SPACEXTBL WHERE customer = 'NASA (CRS)';
```

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

Out[18]:

Total payload mass (NASA (CRS))
45596

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

Average Payload Mass by F9 v1.1

In [19]:
```
%%sql
SELECT AVG(payload_mass__kg_) AS "Average payload mass (booster version F9 v1.1)" FROM SPACEXTBL WHERE booster_version LIKE
```

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

Out[19]:

Average payload mass (booster version F9 v1.1)
2534.6666666666665

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

First Successful Ground Landing Date

```
In [21]:
%%sql
SELECT min(DATE) AS "First successful landing outcome in ground pad" FROM SPACEXTBL WHERE landing_outcome = 'Success (grounc
```

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

Out[21]:

First successful landing outcome in ground pad
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

In [24]:
```
%%sql
SELECT booster_version FROM SPACEXTBL WHERE landing_outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND
```

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

Out[24]:

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

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

In [25]:

```
%%sql
SELECT 'Success' AS "Outcome", count(*) AS "Count" FROM SPACEXTBL WHERE landing_outcome LIKE 'Success%'
UNION ALL
SELECT 'Failure' AS "Outcome", count(*) AS "Count" FROM SPACEXTBL WHERE landing_outcome NOT LIKE 'Success%'
UNION ALL
SELECT '(All)' AS "Outcome", count(*) AS "Count" FROM SPACEXTBL;
```

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

Out[25]:

Outcome	Count
Success	61
Failure	40
(All)	101

Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

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

```
In [26]: %%sql
SELECT DISTINCT booster_version
FROM SPACEXTBL
WHERE payload_mass__kg_ = (
    SELECT max(payload_mass__kg_)
    FROM SPACEXTBL
)
```

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

Out[26]:

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

```
In [12]: %%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.
```

Out[12]:

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)

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

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 [38]:

```
%%sql

SELECT landing_outcome, COUNT(*) AS "Count"
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' and '2017-03-20'
GROUP BY landing_outcome
ORDER BY Count DESC
;
```

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

Out[38]:

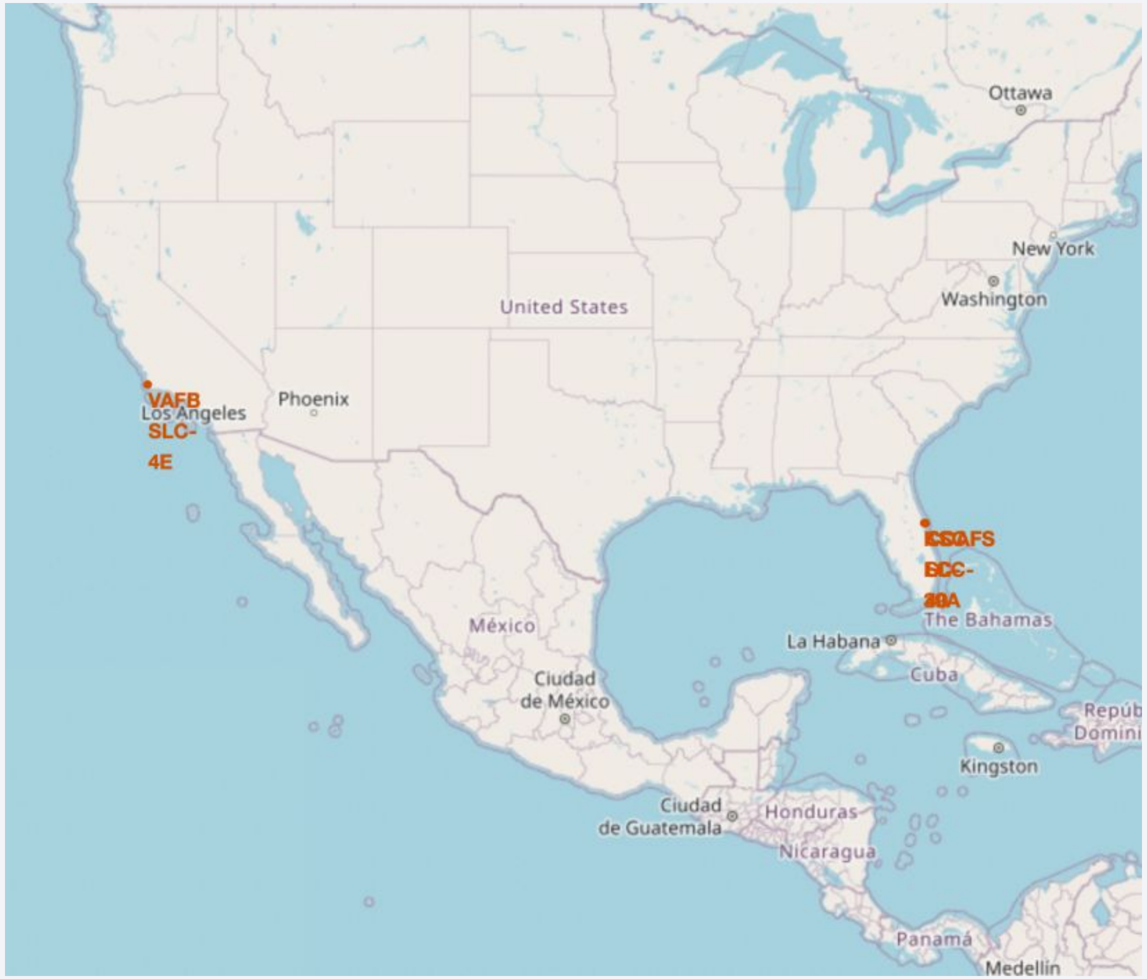
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

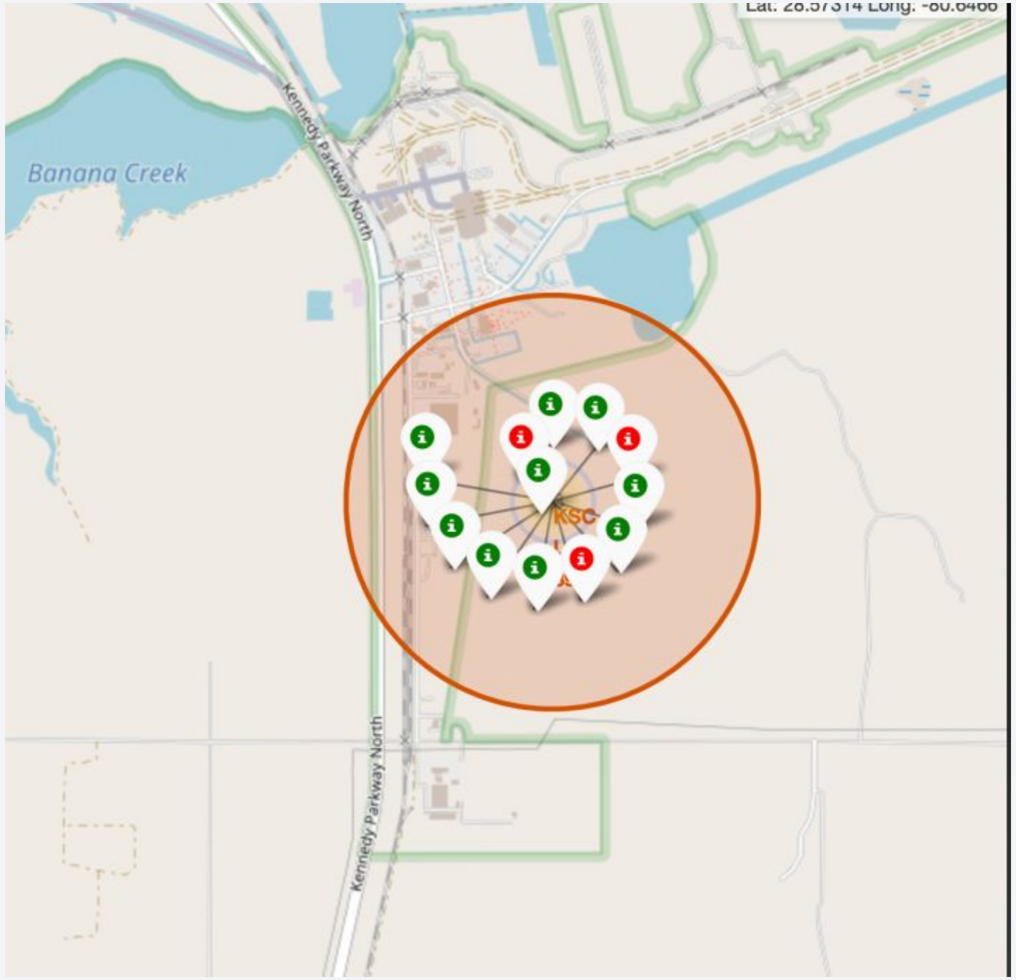
All launch sites' location markers on a map

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people



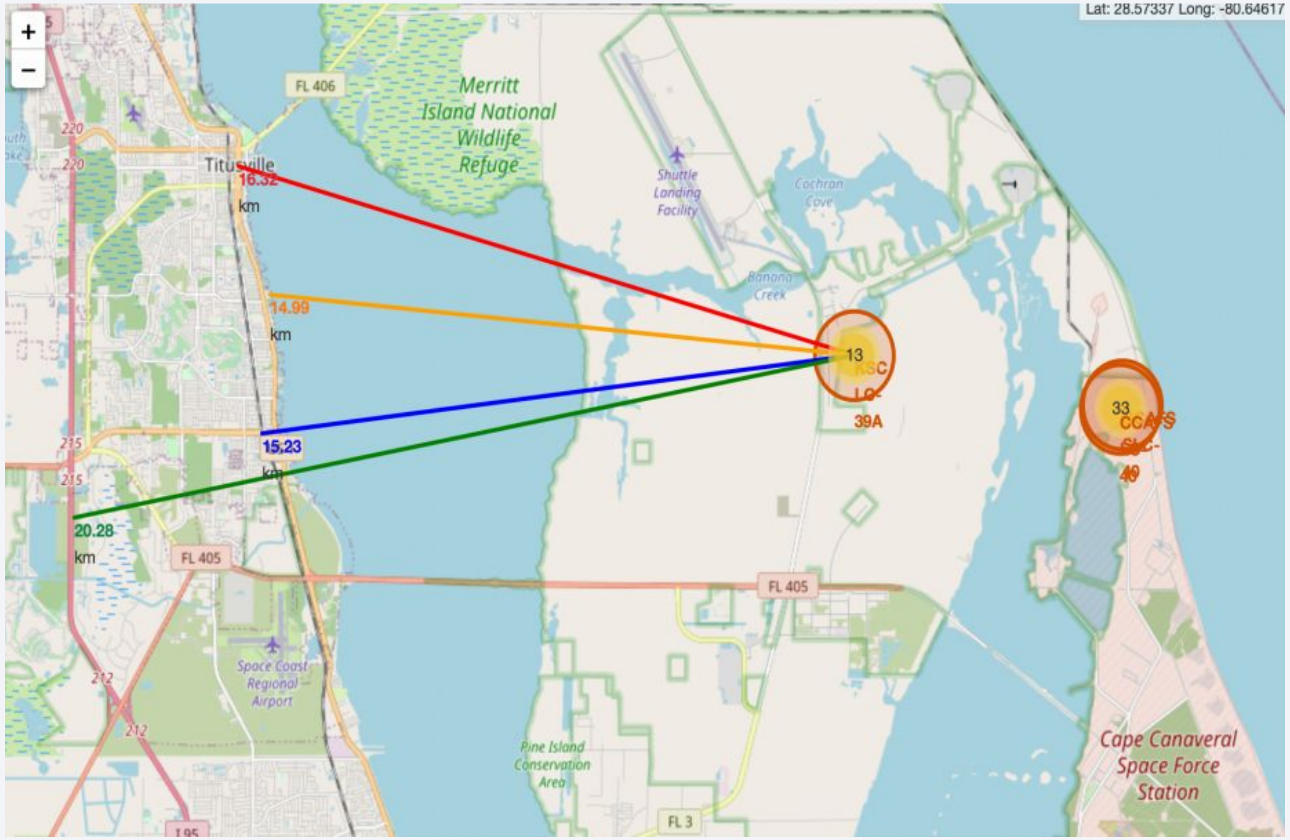
Colour-labeled launch records on the map

- 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



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

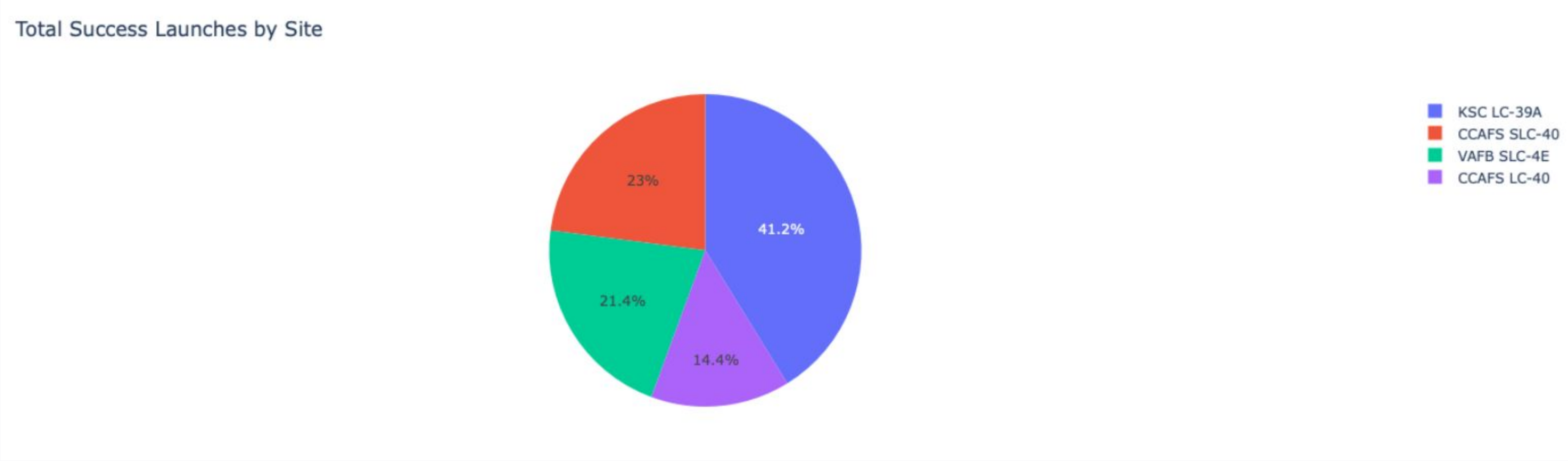
- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km



Section 4

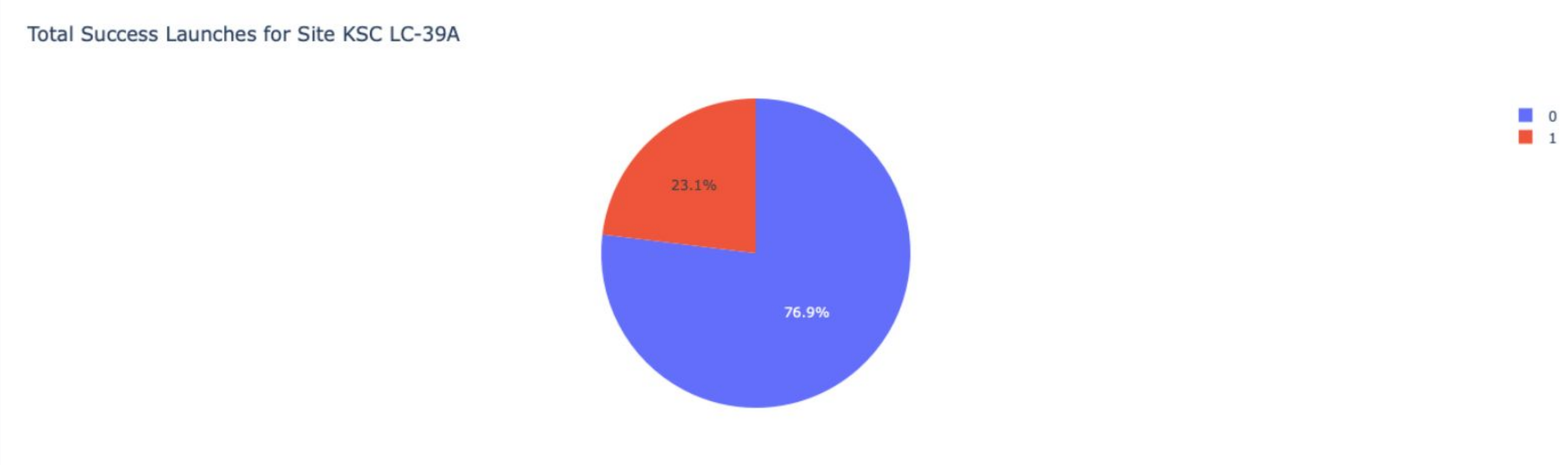
Build a Dashboard with Plotly Dash

Launch success count for all sites



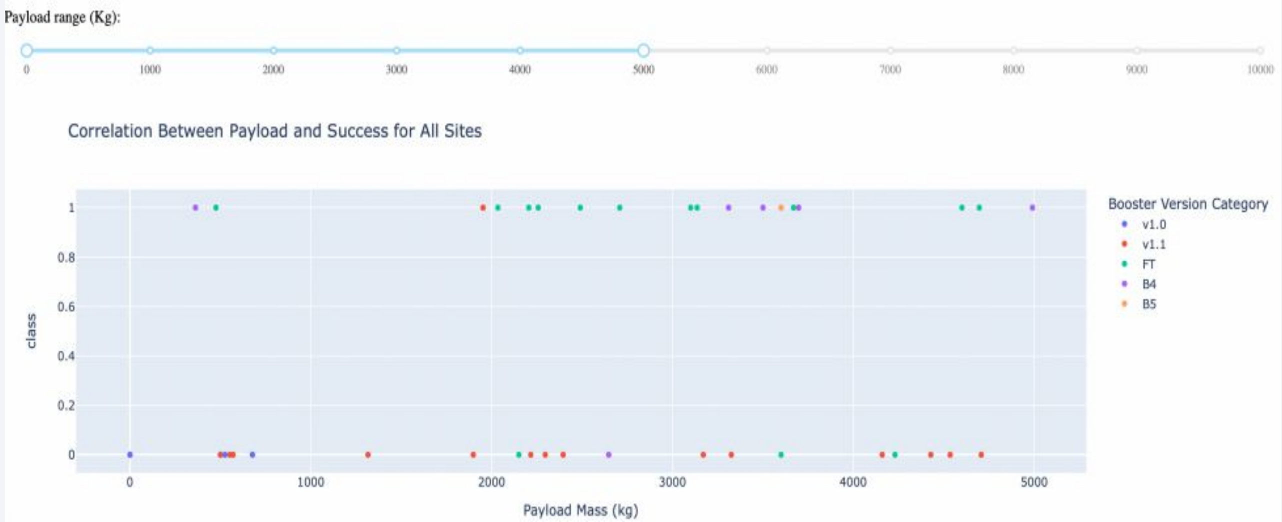
The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

Launch site with highest launch success ratio



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



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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- 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

Scores and Accuracy of the Test Set

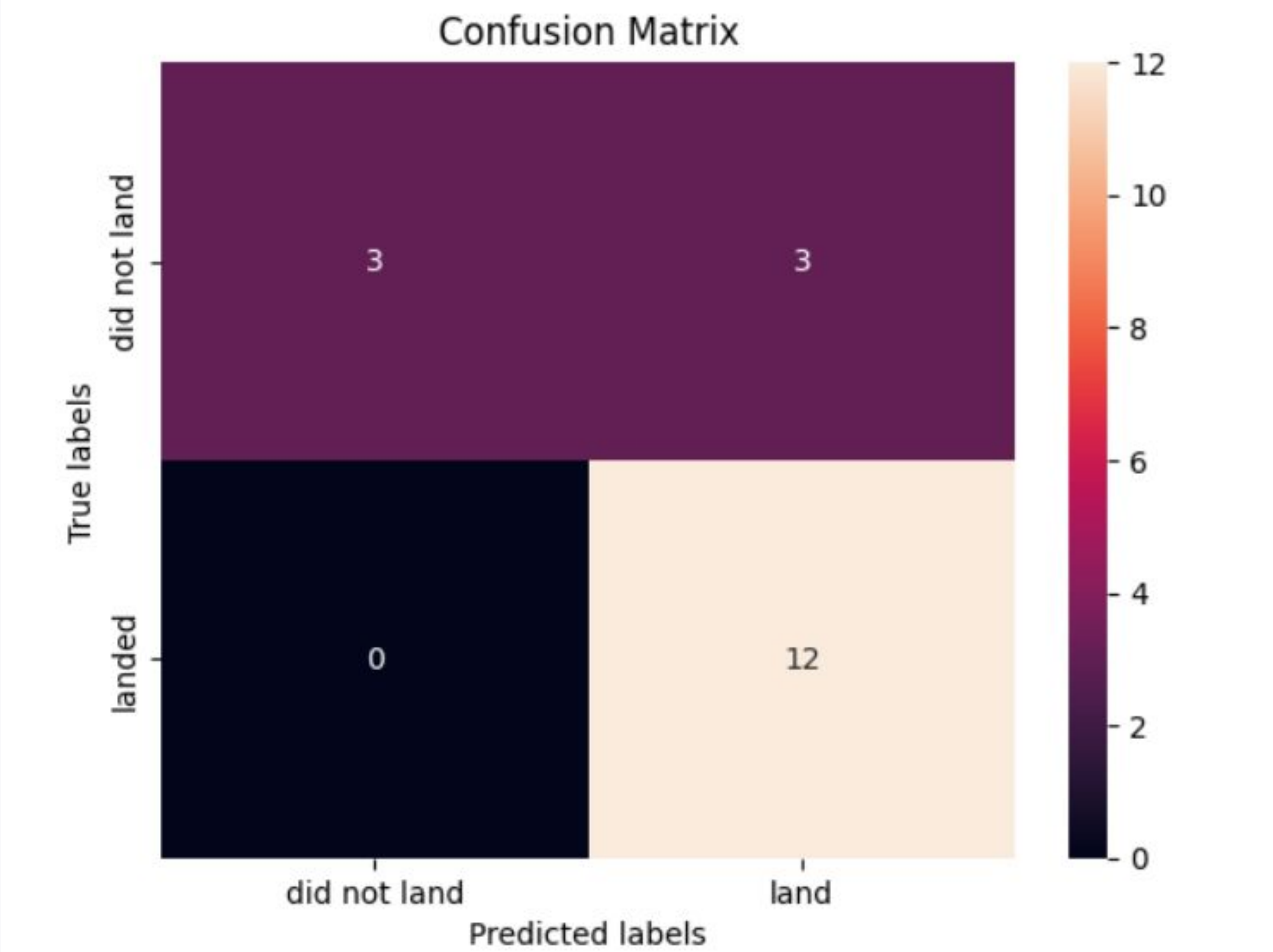
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Entire Data 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

Confusion Matrix

- 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

Based on the characteristics of your dataset, a Decision Tree model seems like a suitable choice because it can effectively capture the relationships between the variables:

- Payload Mass: It seems that launches with lower payload mass tend to perform better, suggesting a potential split point for the decision tree.
- Geographical Factors: The proximity to the Equator line and the coast could be important features for the model to consider when predicting launch success.
- Temporal Trends: The increasing success rate of launches over the years indicates a temporal aspect that a decision tree can incorporate into its splits.
- Launch Site Success Rates*: Knowing that KSC LC-39A has the highest success rate can be a significant node in the decision tree, guiding predictions based on the chosen launch site.
- Orbit Types*: Orbits with a 100% success rate could also be a feature in the decision tree, influencing predictions based on the intended orbit.

All of these characteristics suggest that a Decision Tree model can effectively handle the complexity and variety of factors present in your dataset.

Appendix

Special Thanks to: Instructors Coursera IBM

Thank you!