IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mario Easie
April 13, 2025

Outline

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

Executive Summary

Methodologies Used:

- The project began with **data collection** from relevant sources, gathering datasets necessary for analysis.
- This was followed by **data wrangling**, which involved cleaning, transforming, and preparing the data for further processing.
- An **exploratory data analysis (EDA)** was conducted using data visualization tools to uncover patterns, trends, and insights.
- Additionally, EDA was performed using **SQL** for efficient querying and filtering of data.
- An **interactive map** was built using **Folium**, allowing geographic data to be visualized dynamically.
- A **dashboard** was created with **Plotly Dash**, enabling users to interact with key metrics and visuals.
- Finally, **predictive analysis** was conducted using **classification models** to forecast or categorize future outcomes based on the data.

Results Summary:

- The EDA provided valuable insights into the dataset, which were summarized visually.
- Screenshots of the **interactive analytics tools** demonstrated how users could explore the data in real-time.
- The **predictive analysis** delivered classification results, showcasing the effectiveness of the models built during the project.

Introduction

Project Background and Context:

SpaceX is currently the leading company in the commercial space industry, helping to make space travel more affordable. The company's Falcon 9 rocket launches are listed on its website at approximately $62 million per launch. In contrast, other launch providers charge around $165 million per launch. A major reason for SpaceX's cost advantage is its ability to reuse the first stage of the rocket.

The goal of this project is to determine whether a rocket's first stage will successfully land, based on available data. If we can predict this, we can also estimate the overall cost of a launch. Using publicly available information and machine learning techniques, this project aims to predict whether the first stage of the rocket will be reused.

Questions to be Answered:

- How do factors like payload mass, launch site, number of flights, and orbit type influence the success of the first stage landing?

- Has the success rate of first stage landings improved over time?

- What is the most effective algorithm for binary classification in predicting successful landings?

Section 1

Methodology

Methodology

Data Collection Methodology:

- Retrieved data using the **SpaceX REST API**.
- Performed **web scraping** from Wikipedia to supplement the dataset.

Data Wrangling:

- **Filtered the data** to focus on relevant variables.
- Handled **missing values** to maintain data integrity.
- Applied **One Hot Encoding** to convert categorical variables into binary format for classification.

Exploratory Data Analysis (EDA):

- Conducted EDA using **visualization tools** and **SQL queries** to uncover patterns and relationships in the data.

Interactive Visual Analytics:

- Built **interactive visualizations** using **Folium** for geographic mapping and **Plotly Dash** for dynamic dashboards.

Predictive Analysis Using Classification Models:

- Developed, tuned, and evaluated **classification models** to accurately predict outcomes and ensure optimal performance.

Data Collection

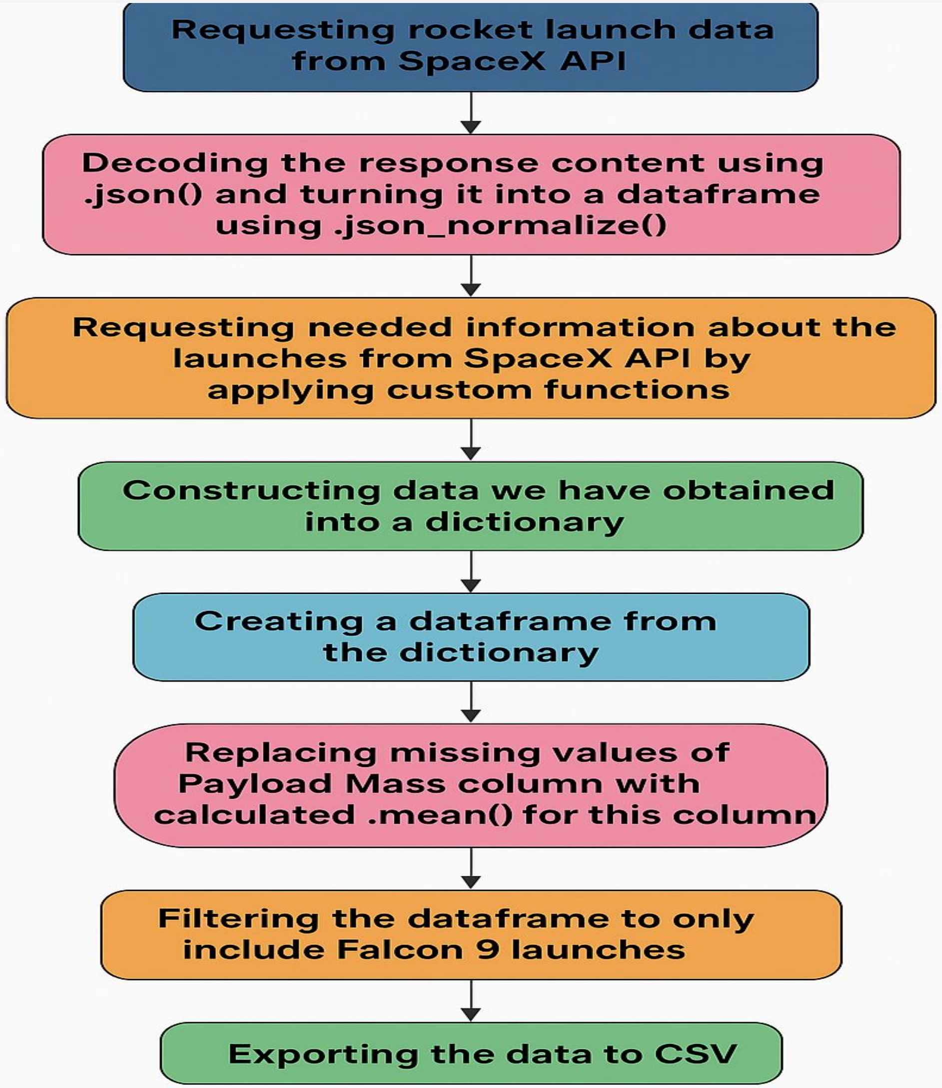
The data collection process involved using a combination of API requests from the SpaceX REST API and web scraping data from a table on SpaceX's Wikipedia page. Both methods were necessary to obtain complete information and support a more detailed analysis of the launches.

Data columns retrieved from the SpaceX REST API include: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

Data columns obtained through web scraping from Wikipedia include: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

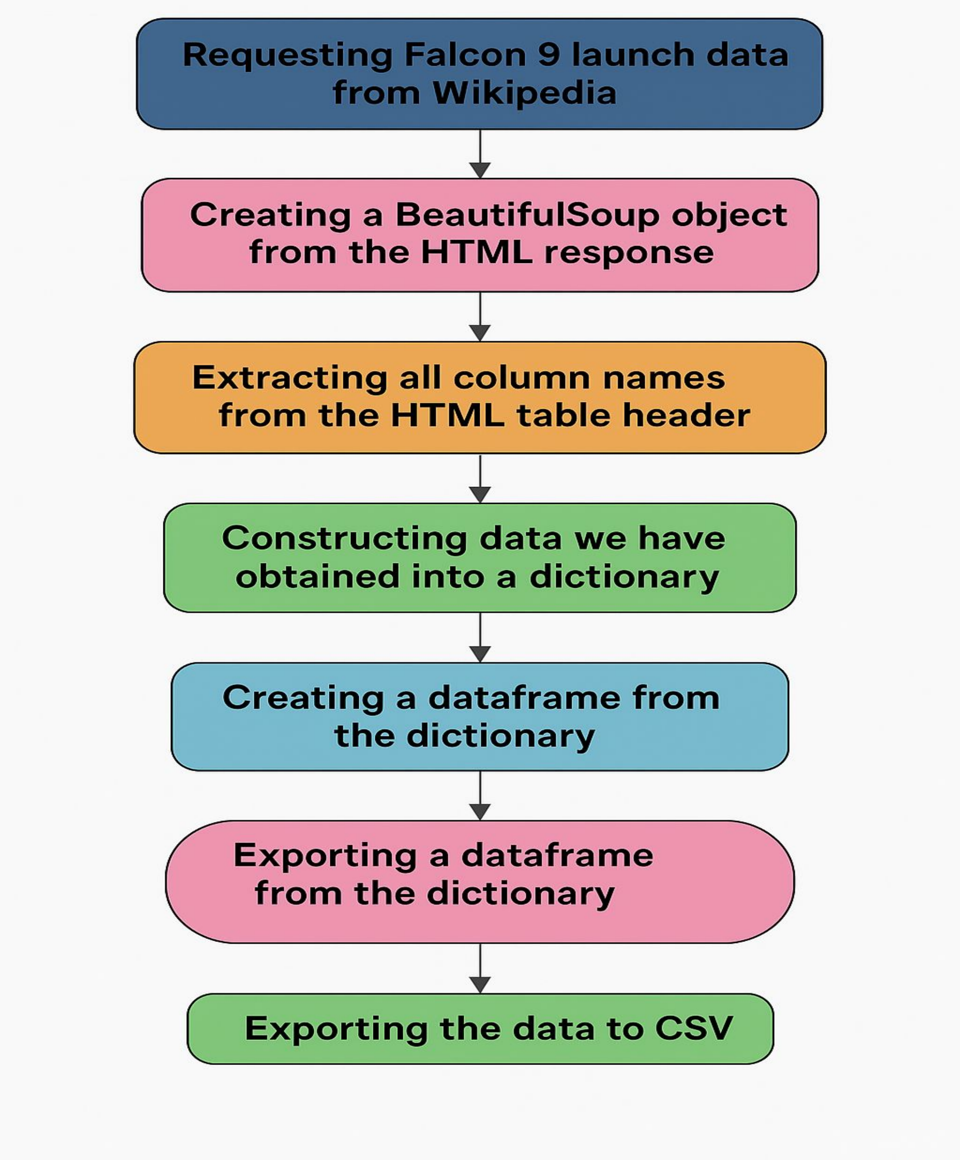
Data Collection – SpaceX API

https://github.com/marioeasie/coursera-labs-mario-easie.git



Data Collection - Scraping

https://github.com/marioeasie/coursera-labs-mario-easie.git



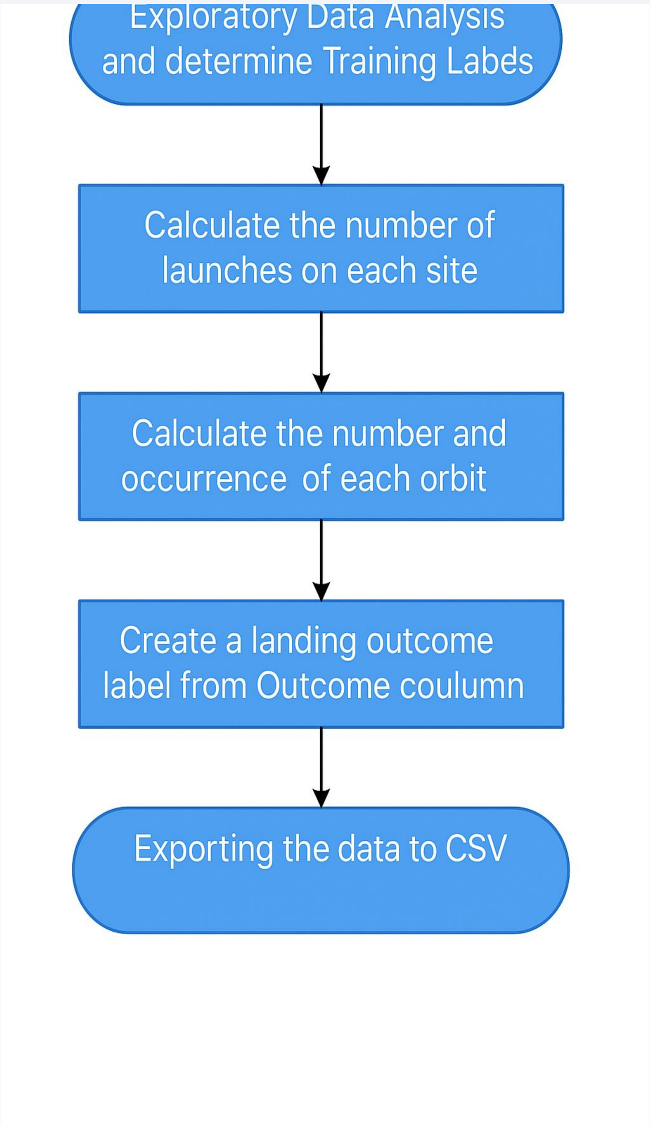
Data Wrangling

The dataset includes various cases where the booster landing was unsuccessful. Some missions attempted landings but failed due to accidents. For example:

- **True Ocean** indicates a successful landing in the ocean, while **False Ocean** indicates a failed ocean landing.
- **True RTLS** means the booster landed successfully on a ground pad; **False RTLS** means it failed to do so.
- **True ASDS** represents a successful landing on a drone ship, whereas **False ASDS** indicates an unsuccessful drone ship landing.

These outcomes are converted into training labels, where **“1”** indicates a successful landing and **“0”** indicates a failure.

https://github.com/marioeasie/coursera-labs-mario-easie.git



EDA with Data Visualization

- In order to adequately show the relationship between variables, scatter plot and bar chart visualization tools were used to explore the data. Scatter plots help show trends that can be used to train machine learning models. Bar charts help detect unique identifiers in categories that are more discrete. The charts used include:

  Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Payload Mass, Orbit Type vs Flight Number, and Orbit Type vs Payload Mass. The success rate per year was also charted to visualize trends.

- https://github.com/marioeasie/coursera-labs-mario-easie.git

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

https://github.com/marioeasie/coursera-labs-mario-easie.git

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.

https://github.com/marioeasie/coursera-labs-mario-easie.git

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:

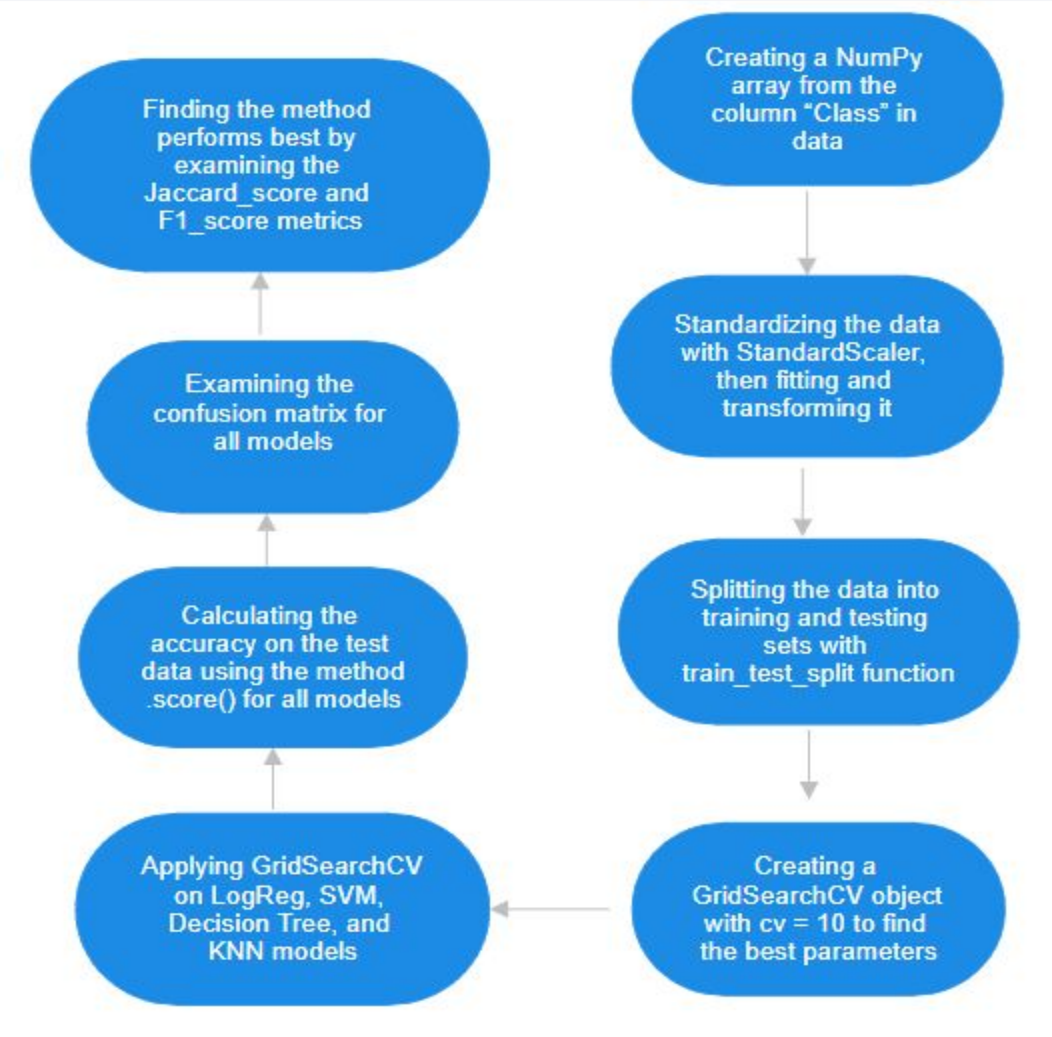
- Added a slider to select Payload range.

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.

https://github.com/marioeasie/coursera-labs-mario-easie.git

Predictive Analysis (Classification)



https://github.com/marioeasie/coursera-labs-mario-easie.git

Results

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

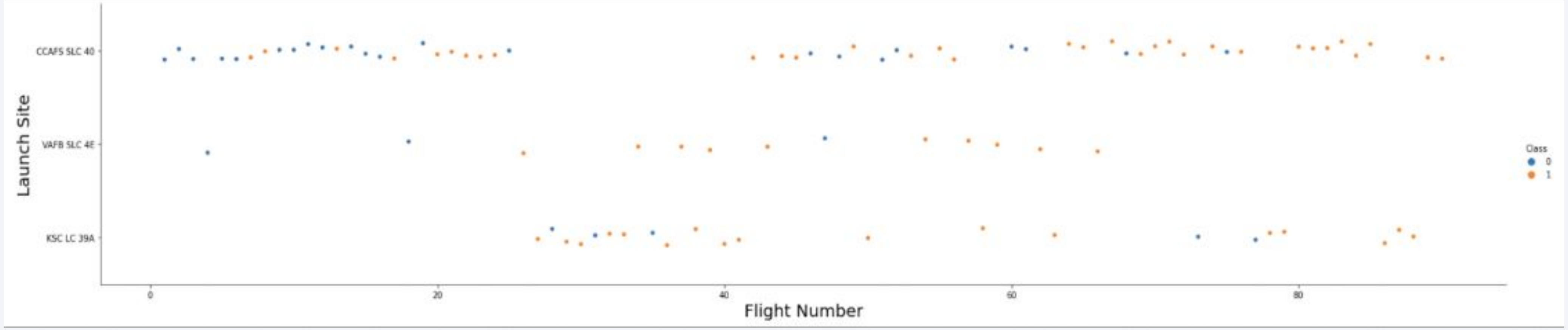
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Explanation:

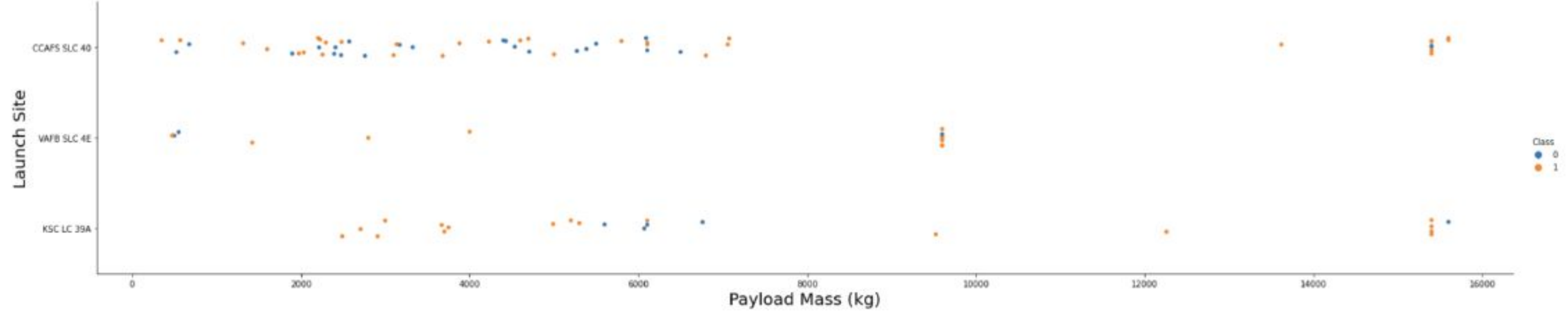
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success



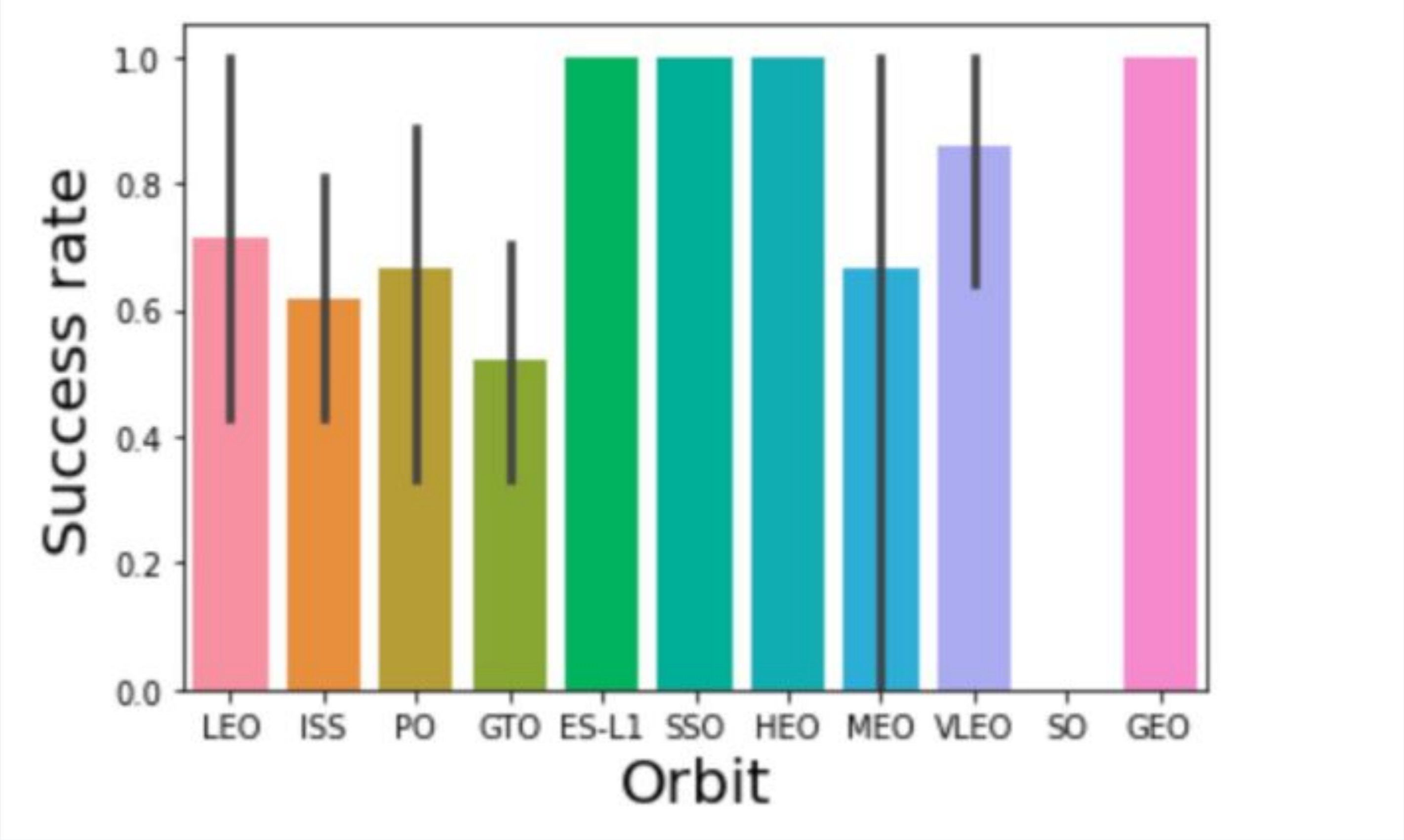
Payload vs. Launch Site

Explanation:

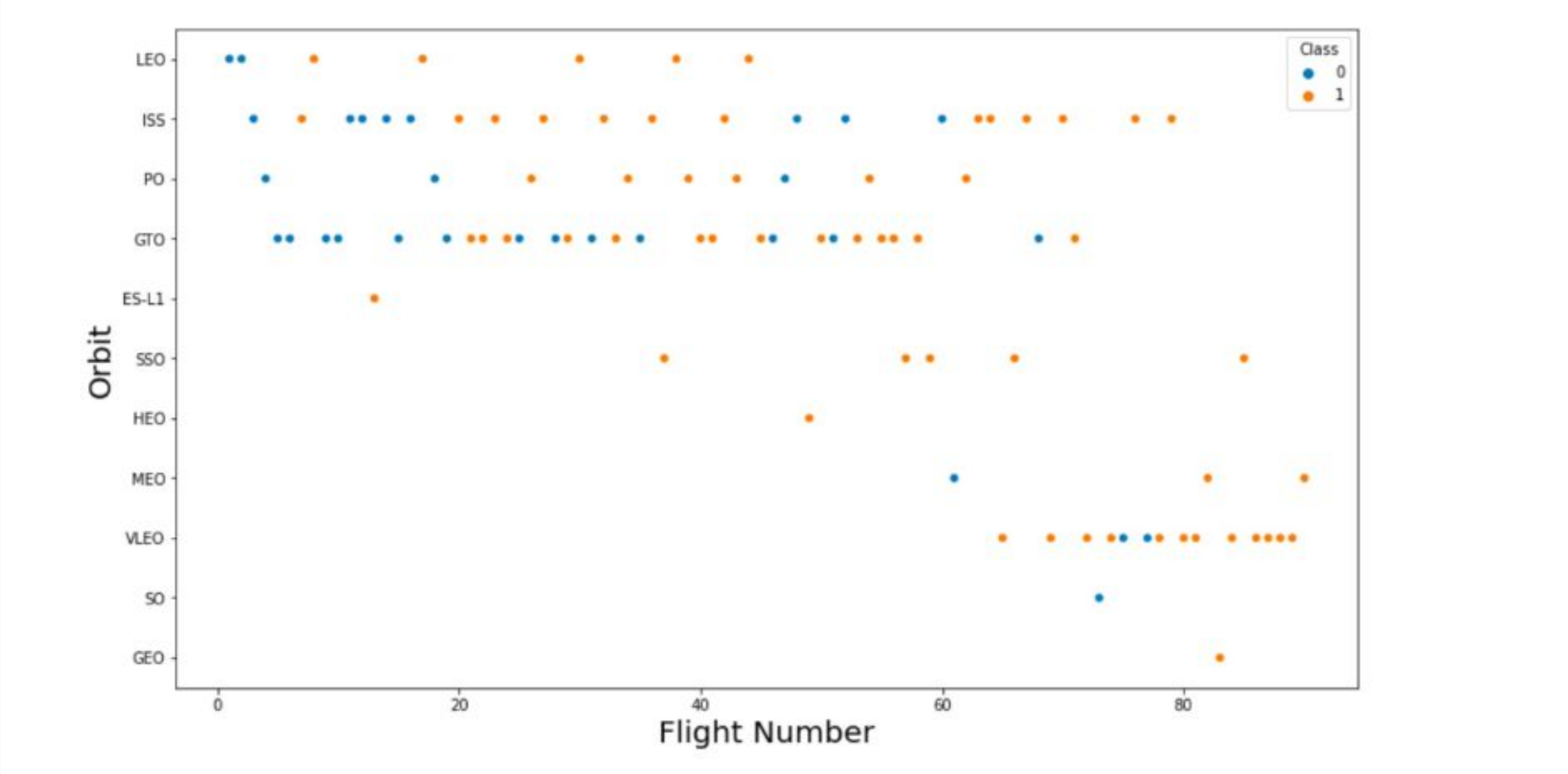
- For every launch site the higher the payload mass, the higher the success
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too



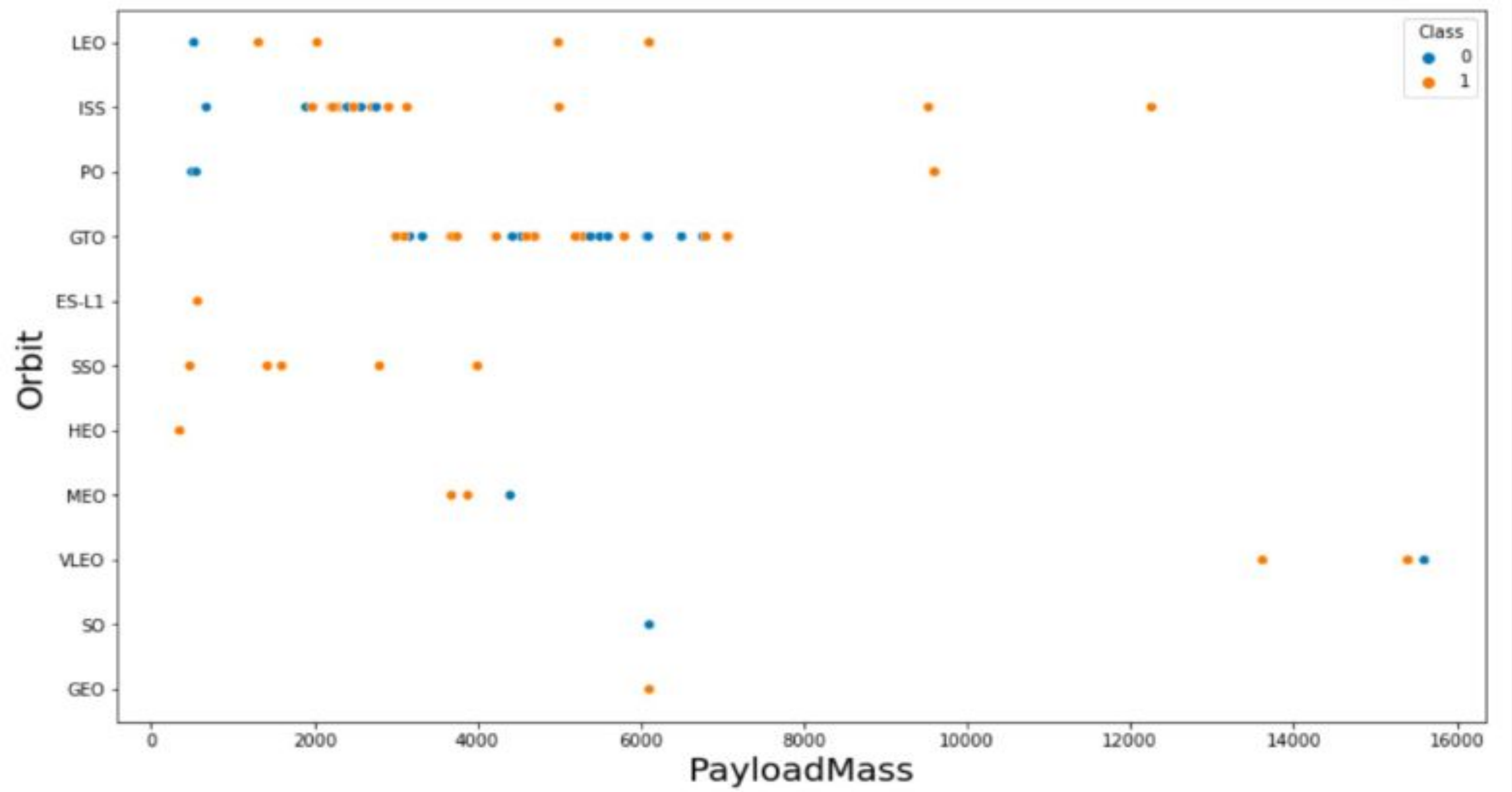
Success Rate vs. Orbit Type



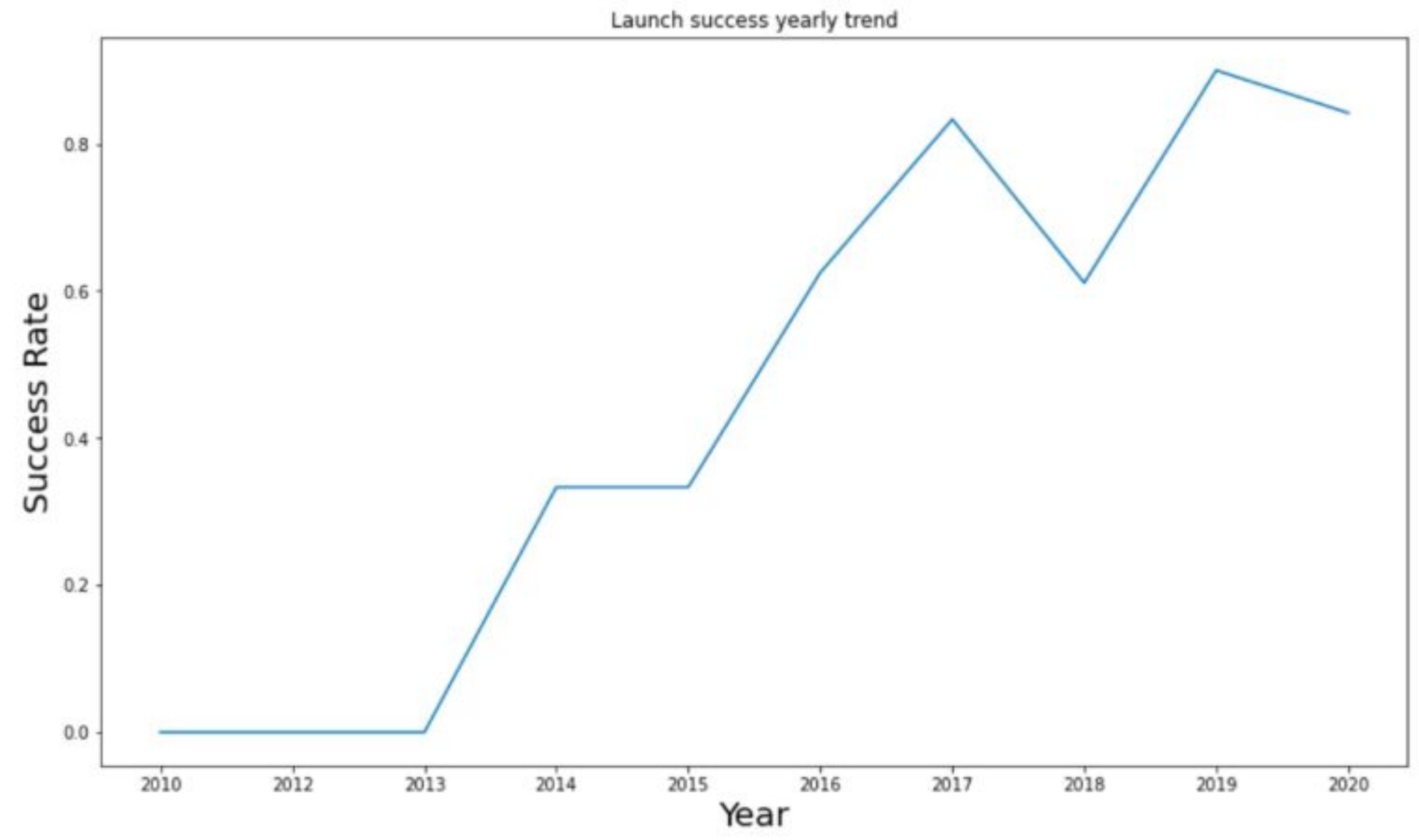
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[4]:
```

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

Explanation: Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

```
In [5]: %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.
```

Out[5]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation: Displaying 5 records where launch sites begin with the string ‘CCA’

Total Payload Mass

```
In [6]: %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.
```

Out[6]:

total_payload_mass
45596

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

Average Payload Mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

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

Out[7]:
```

average_payload_mass
2534

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

First Successful Ground Landing Date

```
In [8]: %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.
Out[8]:
```

first_successful_landing
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

```
In [9]: %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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[9]:
```

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

```
In [10]: %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.
```

Out[10]:

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

Explanation: Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
In [11]: %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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[11]:
```

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 version which have carried the maximum payload mass.

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)

Explanation: Listing the failed landing outcomes in drone ship, their booster versions and launch sites names for the months in year 2015.

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

```
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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

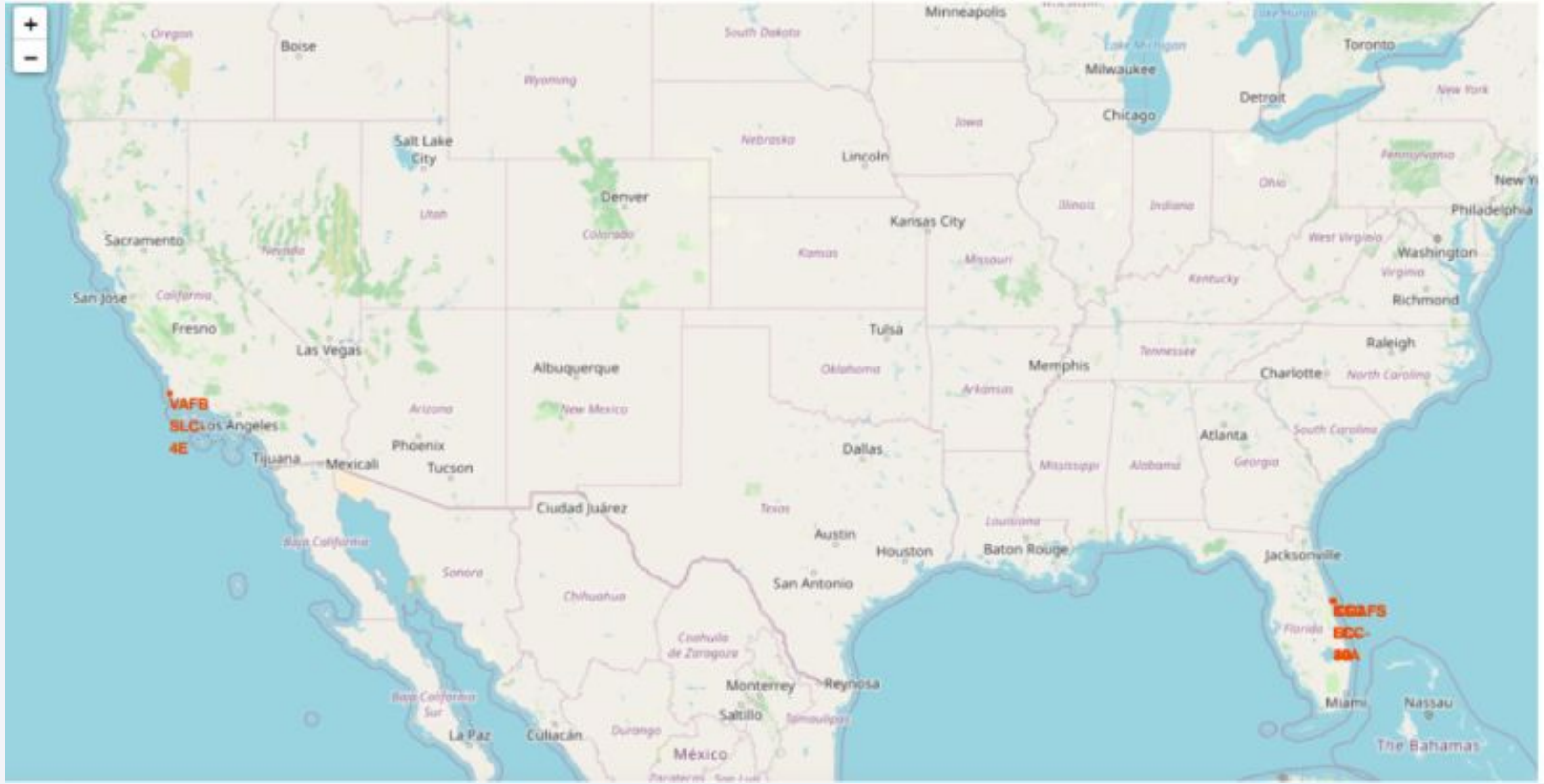
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Explanation: Ranking the count of landing outcomes such as failure drone ship or success ground pad between the dates 2010-06-04 and 2017-03-20 in descending order.

Section 3

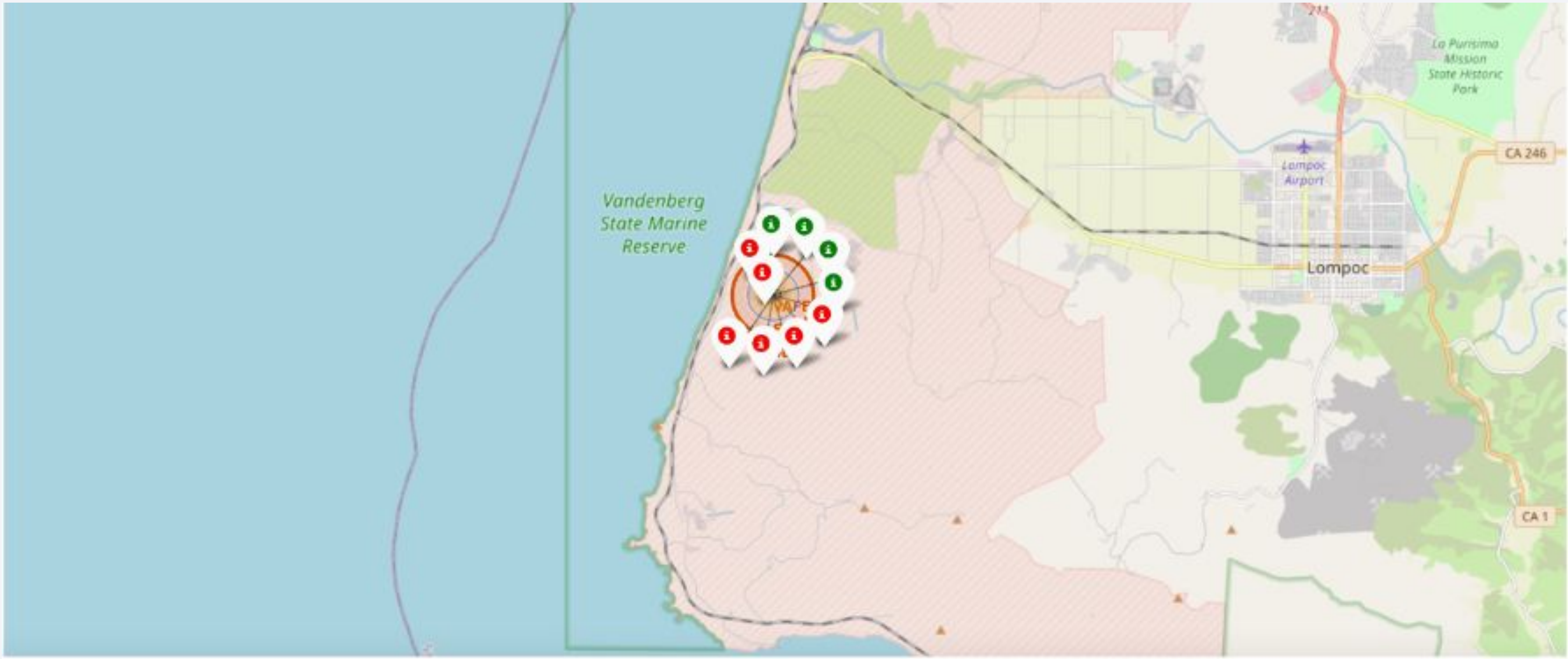
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



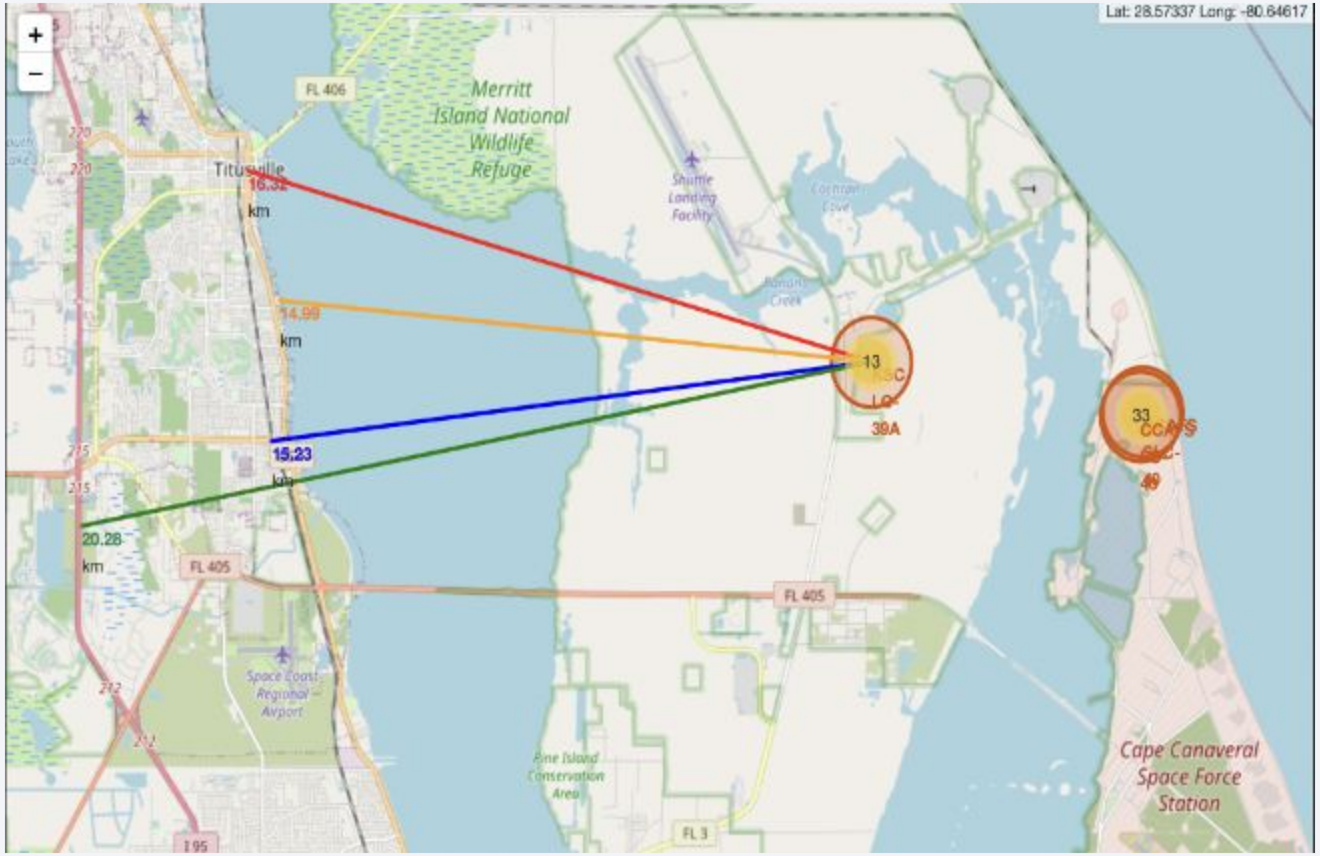
All Launch Sites on Map

<Folium Map Screenshot 2>



The succeeded launches and failed launches for each site on map if we zoom in on one of the launch site, we can see green and red tags. each green tag represent a successful launch while each red tag represents a failed launch.

<Folium Map Screenshot 3>



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)

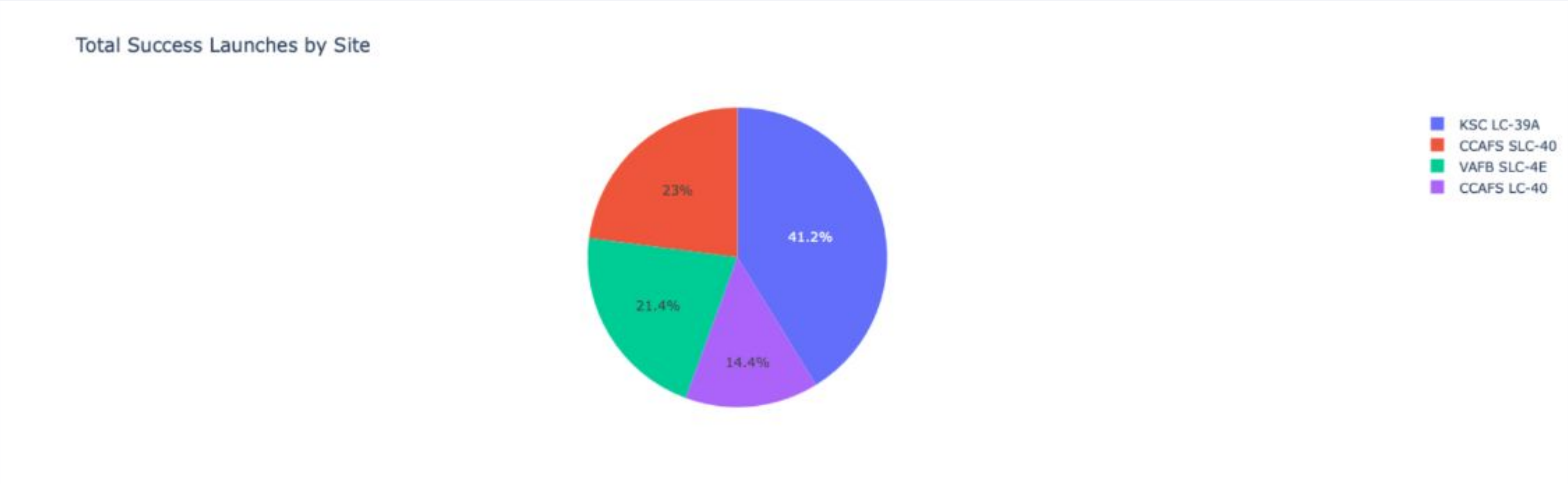
• Also the launch site KSC LC-39A is relatively close to its closest city Titusville (16.32 km).

• Failed rocket with its high speed can cover distances like 15-20 km in a few seconds. It could be potentially dangerous to populated areas

Section 4

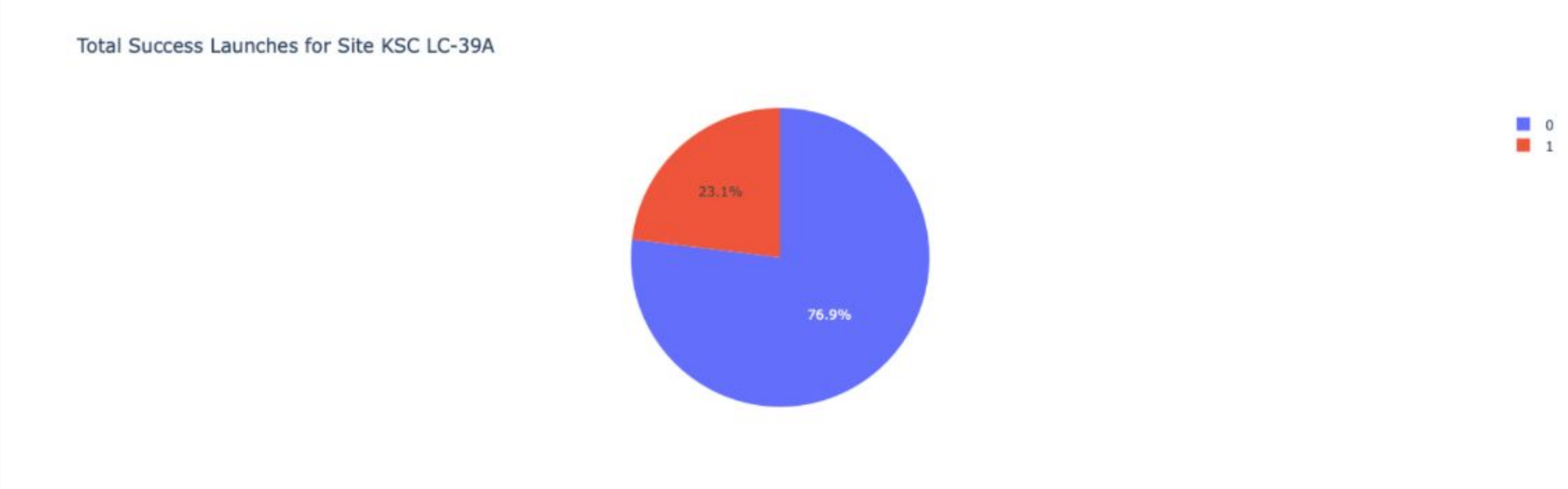
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



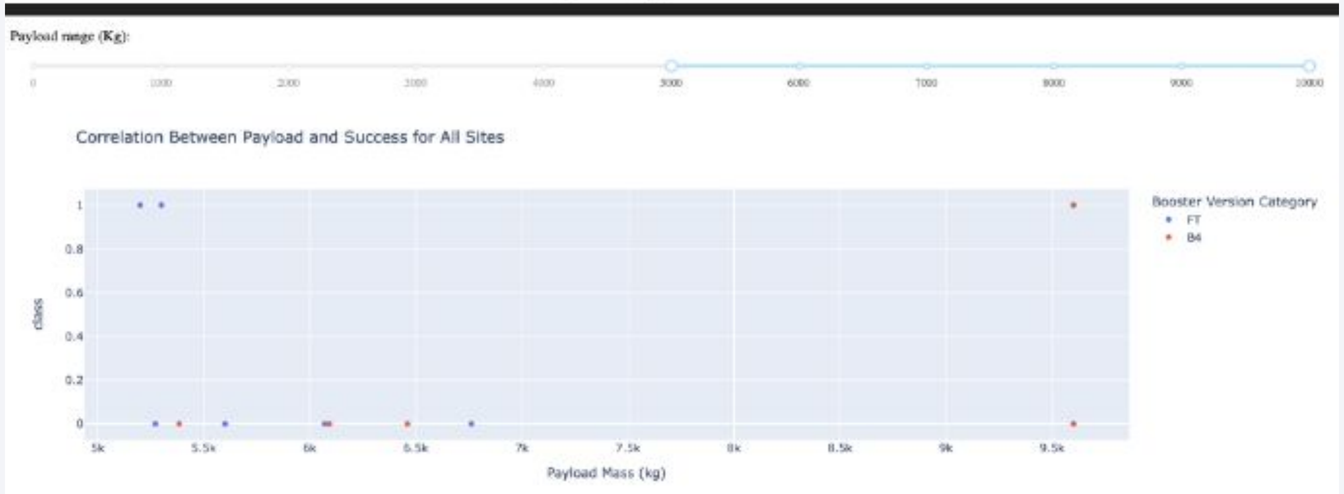
Explanation: The chart clearly shows thats from all the sites, KSC LC-39A has the most successful launches.

<Dashboard Screenshot 2>



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

<Dashboard Screenshot 3>



Explanation The chart shows 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, we can not confirm which method performs best.

• Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.

• The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

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

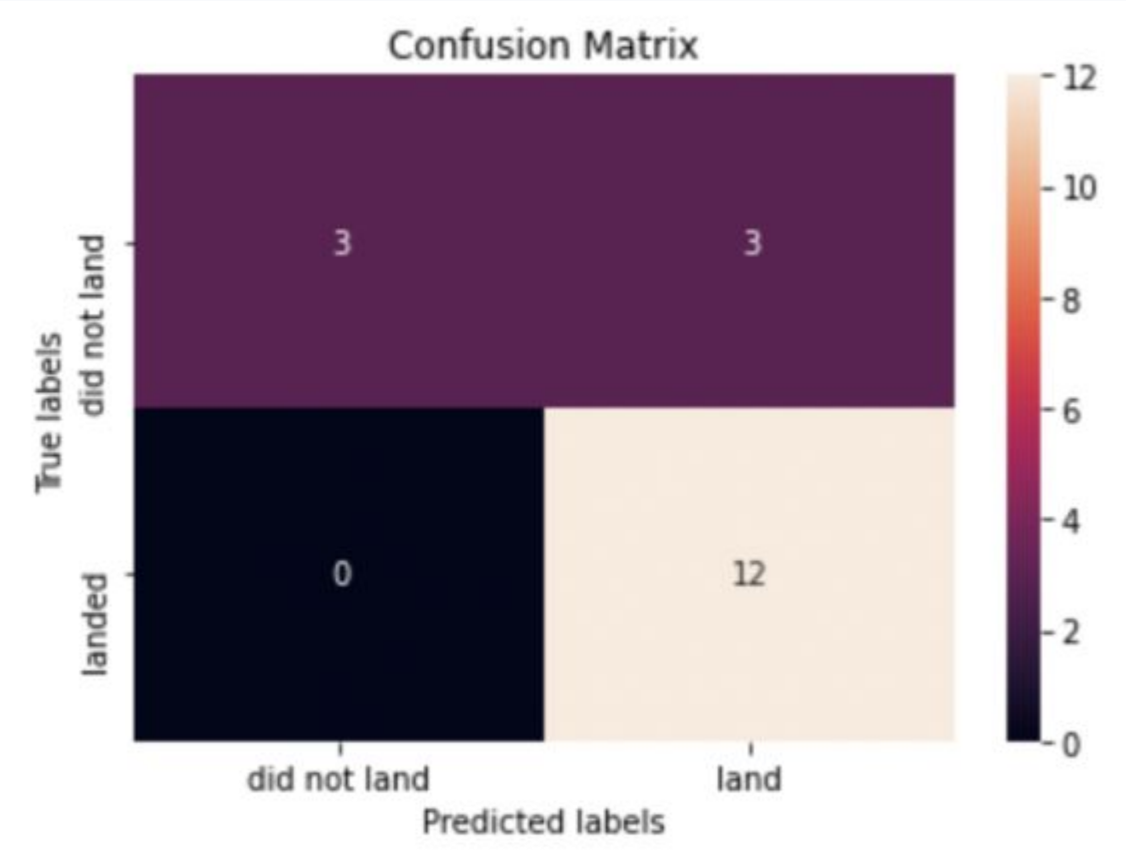
Score and accuracy of the entire data set

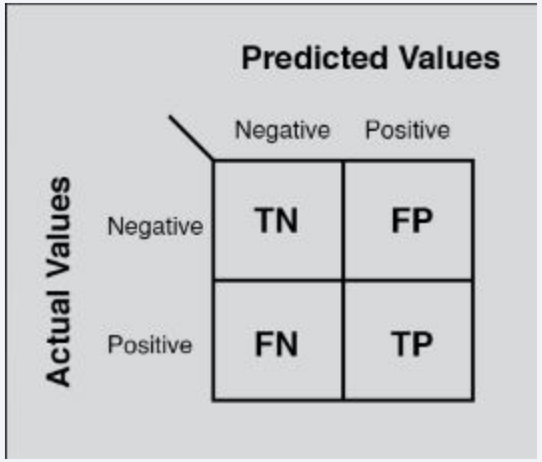
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation:

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





Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate

Appendix

Thanks to: Instructors, Coursera and IBM


Thank you!