IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

S. Kailash Venkat
21st May 2024

Outline

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

Executive Summary

- **Summary of Methodologies:**
  - Data collection via API & web scraping
  - Data wrangling and cleaning
  - Exploratory Data Analysis (EDA) using visualization and SQL
  - Interactive visual analytics using Folium and Plotly Dash
  - Predictive analysis using classification models

Introduction

- **Project Background and Context:**
  - SpaceX launches satellites at 70% lower cost than competitors by reusing rockets.
  - The launched rockets are reused by not disposing them off into the ocean. Instead, the boosters land back safely and are used again.
- **Problem Statement:**
  - Predict the probability of booster landing success based on launch site, payload, orbit, and booster version.

Section 1

Methodology

Methodology

Executive Summary

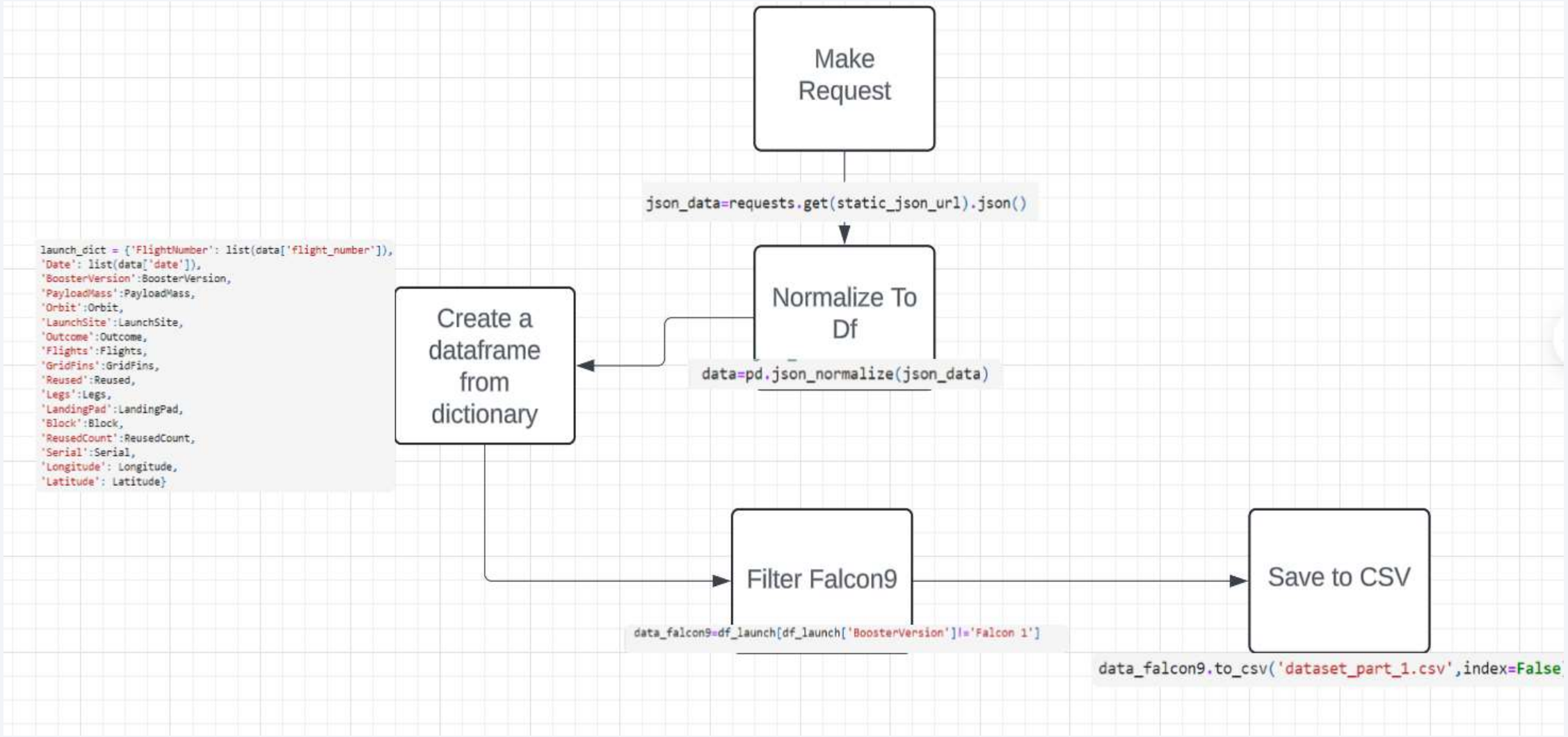
- **Data Collection:**
  - API & Web Scraping
- **Data Wrangling:**
  - Extract, Load, Transform
- **Cleaning Data:**
  - Convert labels to dummy integers
- **EDA:**
  - Visualization and SQL
- **Interactive Analysis:**
  - Folium and Plotly Dash
- **Predictive Analysis:**
  - Machine Learning Models

Data Collection

- **Process**:
  - Make requests to the REST API
  - Normalize JSON to DataFrame
  - Create DataFrame from dictionary
  - Filter for Falcon 9
  - Save to CSV
- **Link to Notebook**:

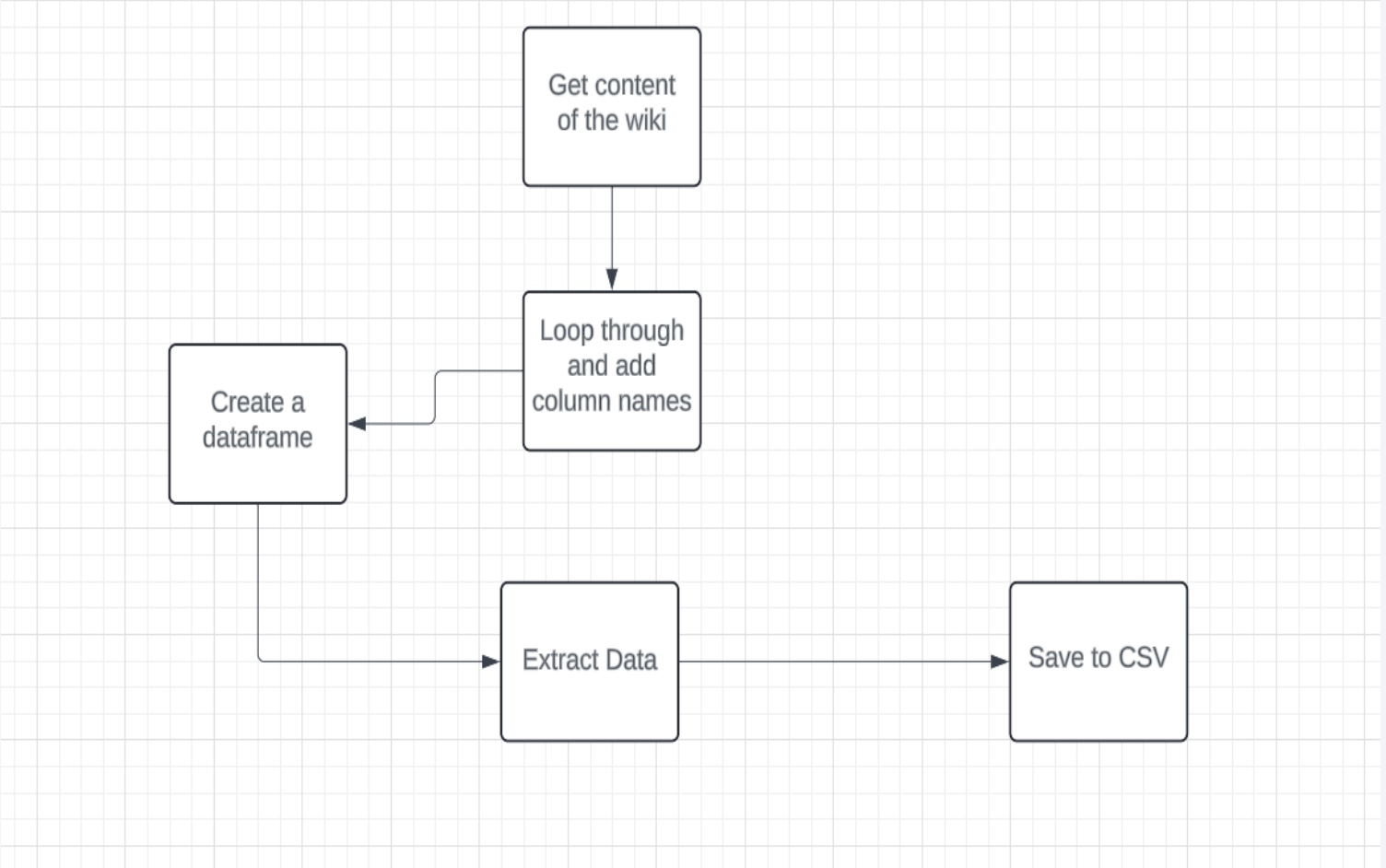
  Data Collection

Data Collection – SpaceX API



Data Collection - Scraping

- **Process**:
  - Get content from Wikipedia
  - Normalize to DataFrame
  - Extract data in a loop
  - Add column names
  - Save to CSV
- **Link to Notebook**:
  - Web Scraping



Data Wrangling

- **Objective:**
- Prepare the dataset for predictive analysis by converting complex mission outcomes into binary training labels.
- **Process:**
- **Mission Outcome Categories:**
  - Different outcomes were recorded for booster landings:
    - **True Ocean**: Successfully landed in a specific ocean region.
    - **False Ocean**: Unsuccessfully landed in a specific ocean region.
    - **True RTLS (Return to Launch Site)**: Successfully landed on a ground pad.
    - **False RTLS**: Unsuccessfully landed on a ground pad.
    - **True ASDS (Autonomous Spaceport Drone Ship)**: Successfully landed on a drone ship.
    - **False ASDS**: Unsuccessfully landed on a drone ship.

Data Wrangling

- **Conversion to Training Labels:**
  - The mission outcomes were simplified into binary labels:
    - **1**: Indicates a successful landing (True Ocean, True RTLS, True ASDS).
    - **0**: Indicates an unsuccessful landing (False Ocean, False RTLS, False ASDS).
- Link :

  Wrangling

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

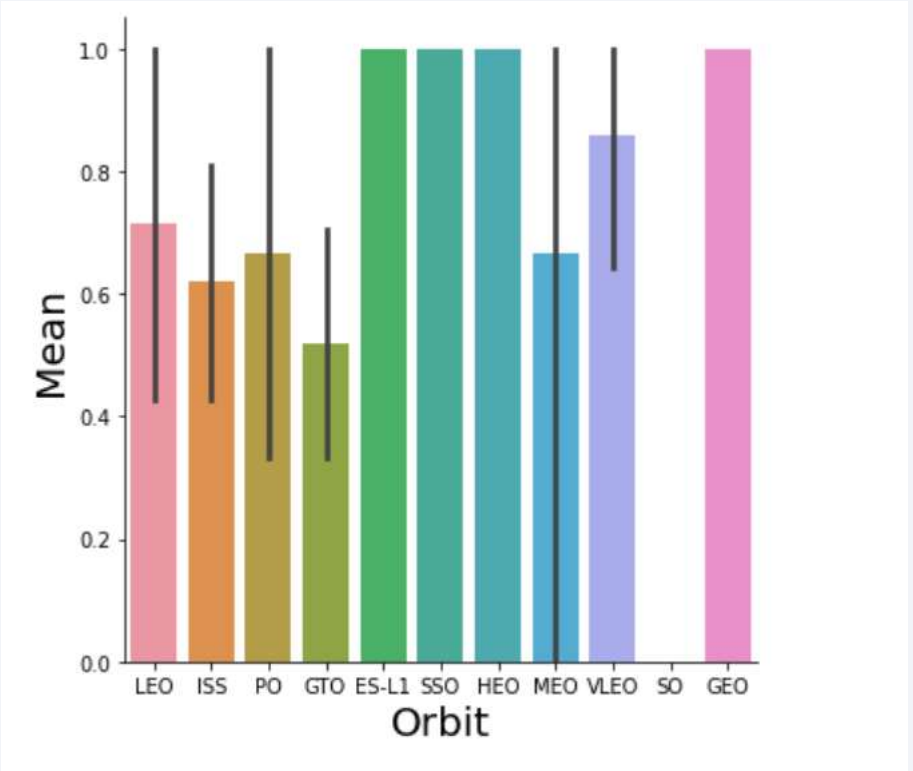
```
landing_class = []
for key,value in df["Outcome"].items():

    if value in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
```

EDA with Data Visualization

- **Summary of Plotted Charts**:
  - Flight number vs. Launch Site
  - Payload vs. Launch Site
  - Success rate vs. Orbit type
  - Flight number vs. Orbit type
  - Payload vs. Orbit type
  - Success rate trend
- **Link to Notebook**:

Data Visualization



Relation between Orbit and success rate. The figure shows that some orbits like the ES-L1 have higher success rate than others.

EDA with SQL

- **SQL Queries Summary**:
  - Unique Sites
  - Max Payload
  - Average Payload
  - First Successful Landing Day
  - Success and Failures Count
  - Boosters with Max Payload
- **Link to Notebook**:

  SQL EDA

Build an Interactive Map with Folium

- **Map Objects**:
  - Markers, circles, lines
- **Purpose**:
  - Visualize launch sites and outcomes
- **Link to Notebook**:

Folium

Build a Dashboard with Plotly Dash

- **Plots and Interactions**:
  - Launch success count by site
  - Payload vs. Launch Outcome with range slider
- **Link to Notebook**:

Dashboard

Predictive Analysis (Classification)

- **Model Development Process**:
  - KNN, Decision Tree, Logistic Regression, Support Vector Machine
- **Best Model**:
  - Decision tree with 0.88 Jaccard_score,0.937 F1_score and 0.91 accuracy
- **Link to Notebook**:

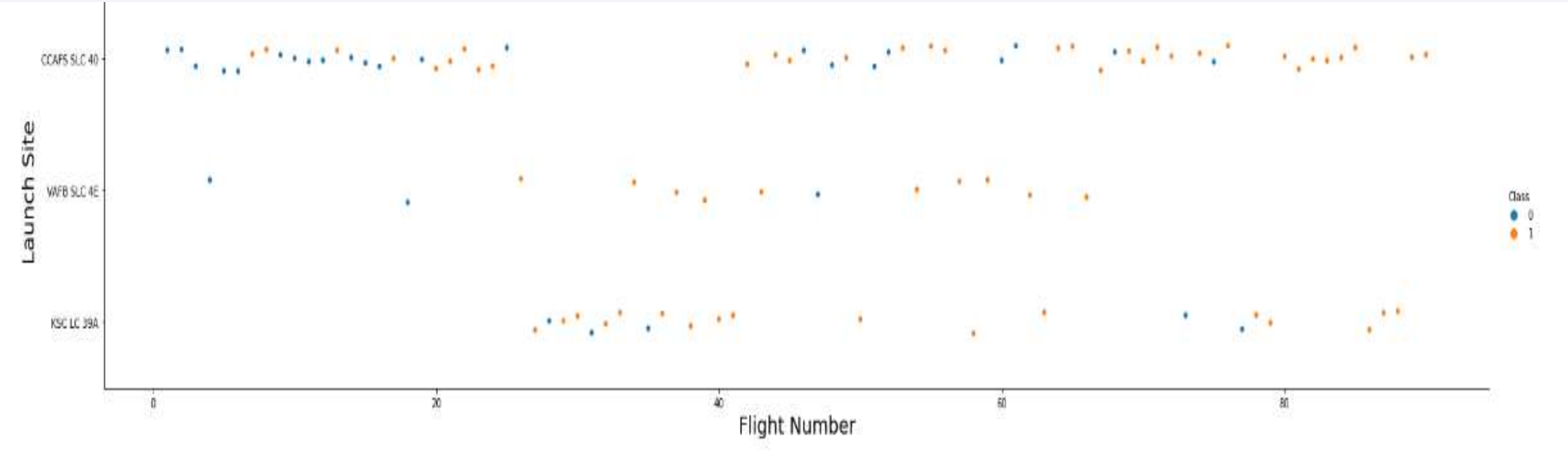
  Modelling

Results

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

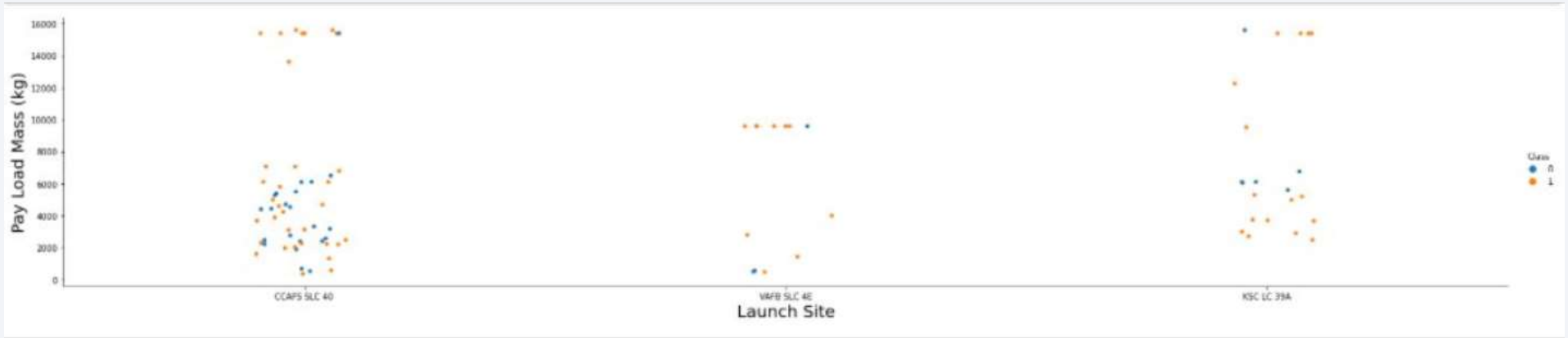

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Most launches are from CCAFS SLC-40 site
- Least launches are from VAFB SLC 4E site
- Most of the flights launched from VAFB SLC 4E site have class 1

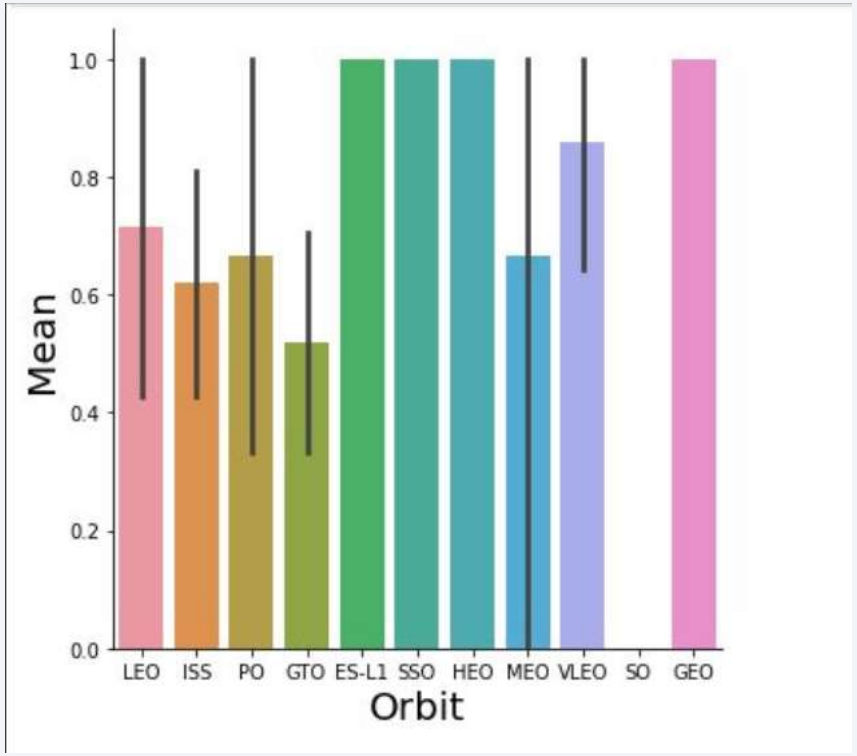
Payload vs. Launch Site



From the Visualization we can concluded that:

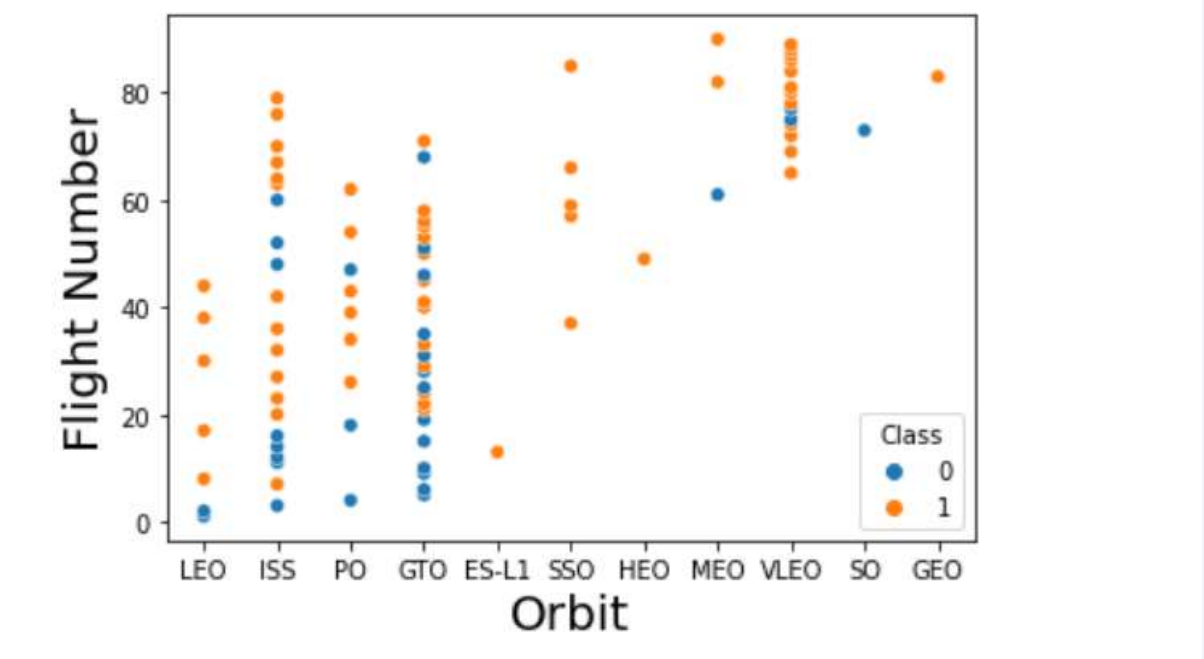
- VAFB SLC 4E has Low Payload launches
- CCAFS SLC 40 has more Higher Payload Launches and Low Payload Lauches .

Success Rate vs. Orbit Type



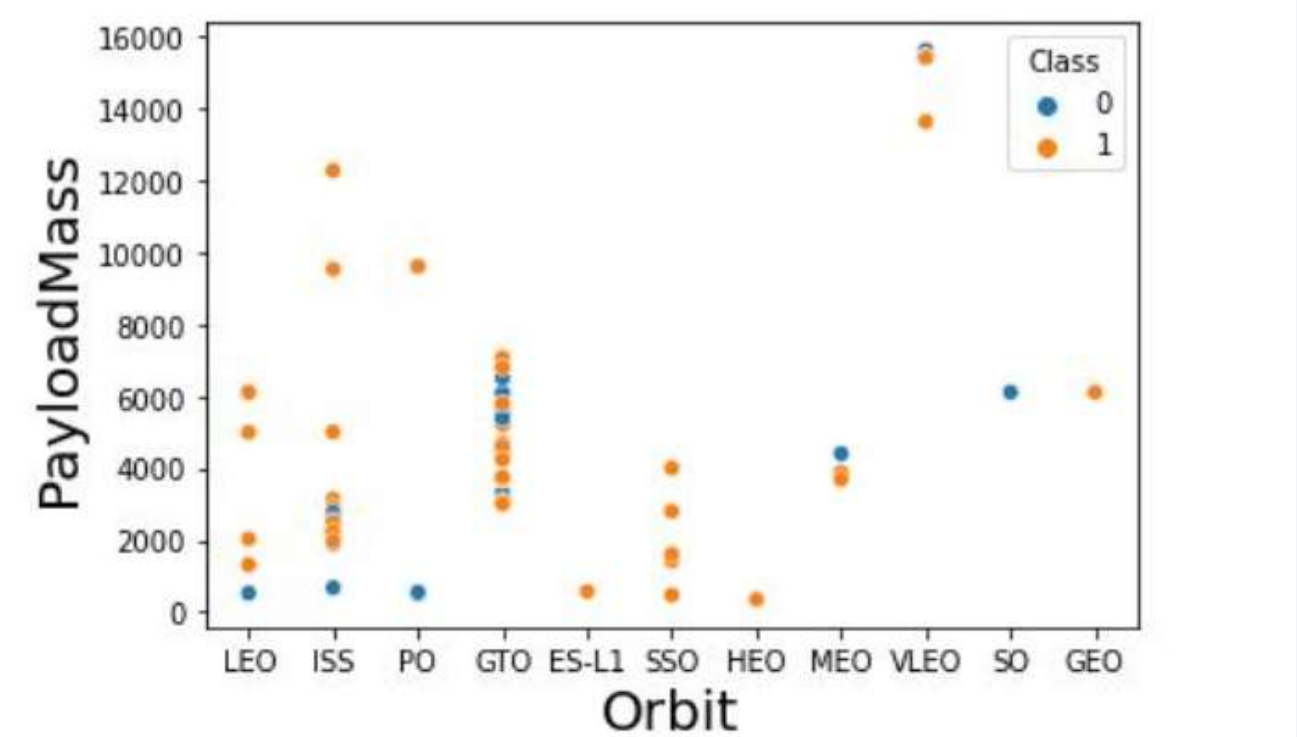
Orbits like ES-L1,SSO,HEO and GEO have a 100% success rate while SO has a really low success rate

Flight Number vs. Orbit Type



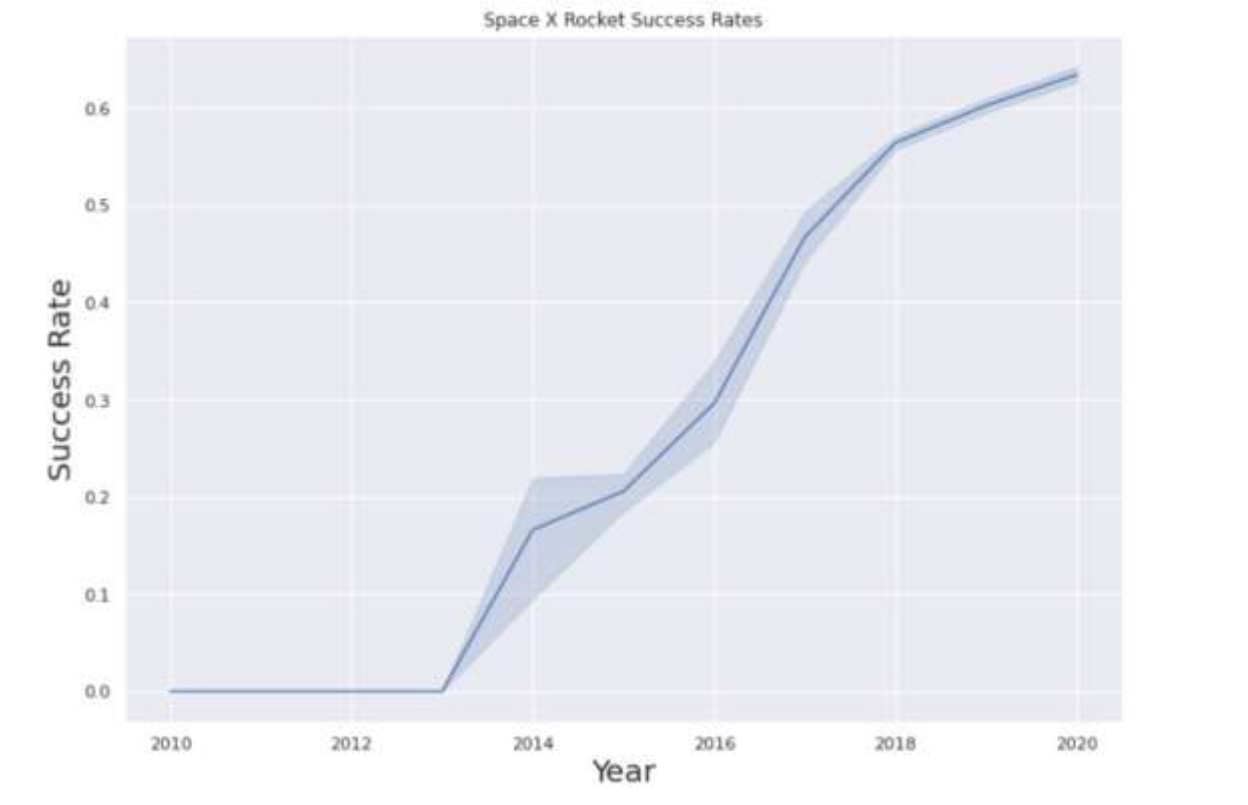
- Most flights are from GTO and ISS
- Least flights are from ES-L1 and SO
- VLEO has a very high success rate
- SO has a very poor success rate

Payload vs. Orbit Type



- High payload launches are usually for VLEO and low payload launches are for HEO,SSO,ES-L1
- GTO has moderate payload weights and has a pretty average success rate

Launch Success Yearly Trend



As the technology of SpaceX gets better every year, we can observe that the success rate of launches increases each year and follows a positive trend.

All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

Out[4]:

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

We can display all the launch sites using the select command in sql and the sites are distinct. There are a total of 4 launch sites.

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

Used select statement to display records. We have used the where and like clauses to filter the records to only display those with launch sites beginning with CAA. Limit 5 is used to retrieve only 5 records that match our criterion.

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.clo
Done.
```

Out[6]:

total_payload_mass
45596

The sum aggregate function is used to calculate the total mass of the payload. This is displayed as total_payload_mass using the as method.

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

The average payload weight is selected using the avg aggregate function. To limit the booster type to F9 v1.1 we use the where clause.

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

We have used the min aggregate function to find the minimum date that satisfies our condition specified by the where clause which makes sure that the launch was a success.

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

WE have used the between method along with the where clause to display all the booster versions who had a successful drone ship landing with payload between 4000 and 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.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

We have displayed the total successful and failure mission outcomes using the group by clause and the aggregate function count(*).

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

We have printed the boosters that carried the maximum payload using nested queries. We first find the max payload using the max aggregate function in the inner query. Then we use the where clause to print the necessary output.

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)

Listed all the failed launch records in 2015 using the select clause in conjunction with the where clause to restrict our search to failures and 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

We have ranked the landing outcomes using the order by method. As the highest value had to be the first record, we have used desc.

Section 3

Launch Sites Proximities Analysis

Folium map showing the launch sites

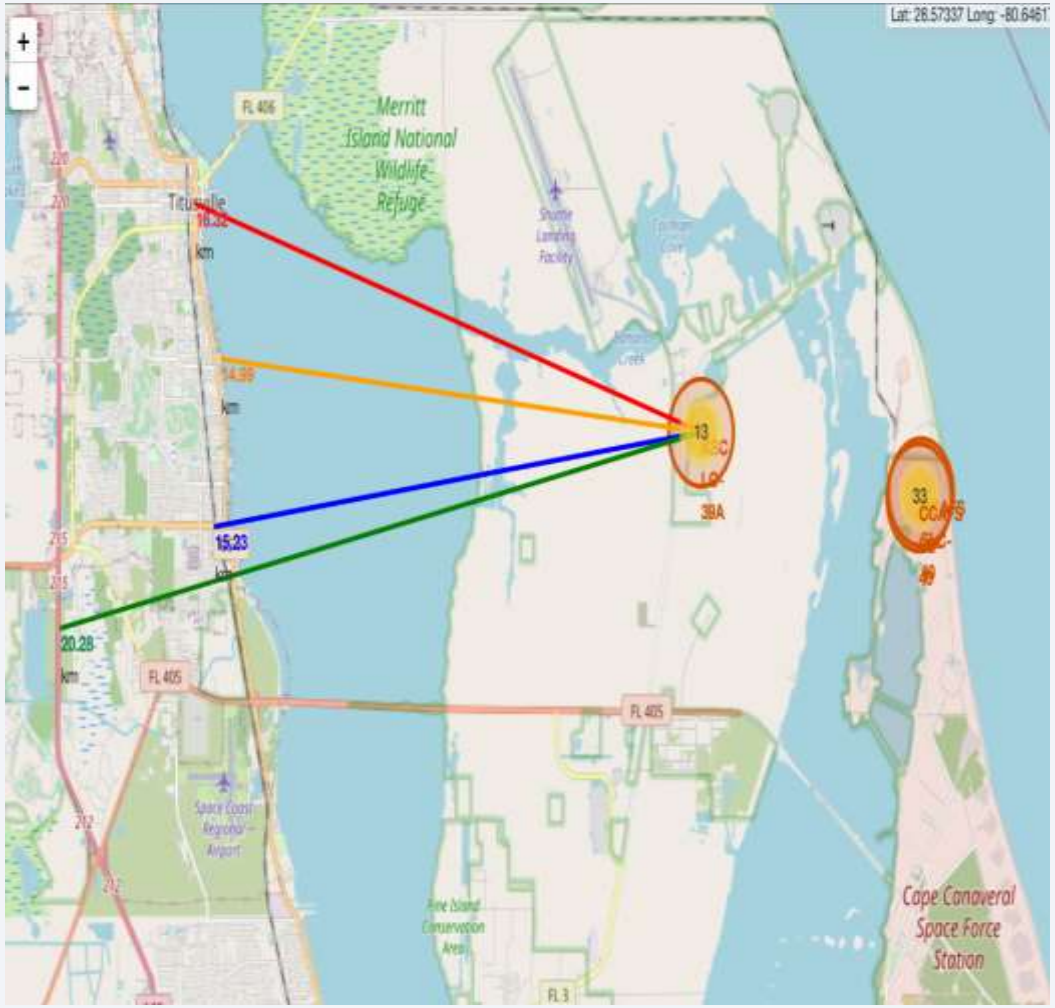
The above map shows all the launch sites. We can notice that all the sites are coastal areas. This is because any failure during launch can be easily handled by piloting the rocket into the sea. This reduces the risks of launch failure. Also, the launch pads are near the equator as it is more efficient to launch from the equator due to the earth's rotation.

Color labelled launch records

The map alongside shows markers in the launch site. These labels should help the user identify which sites have higher success rates.

- The GREEN labels represent successful launches
- The RED labels represent failed launches

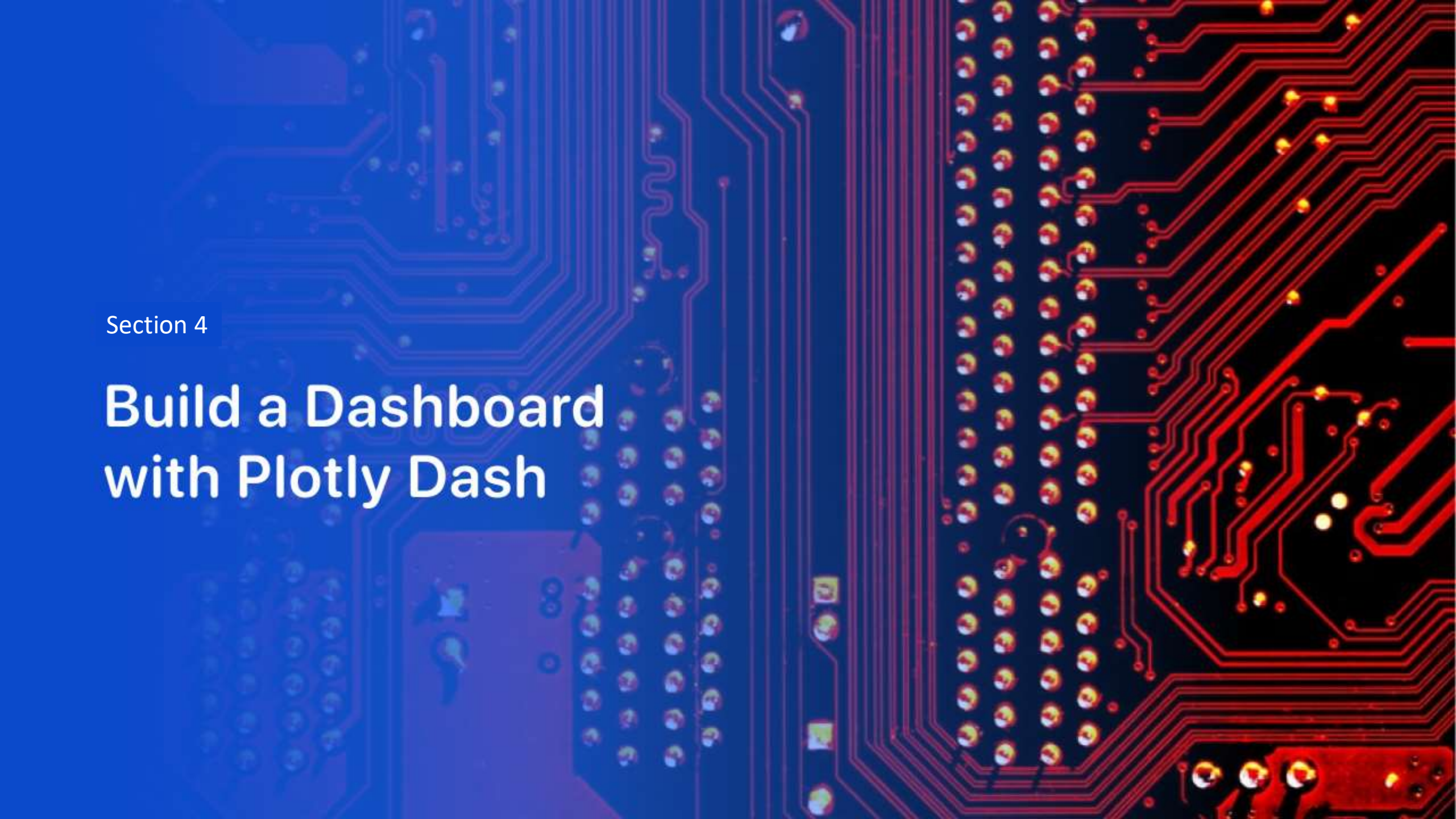
Launch site to its proximities



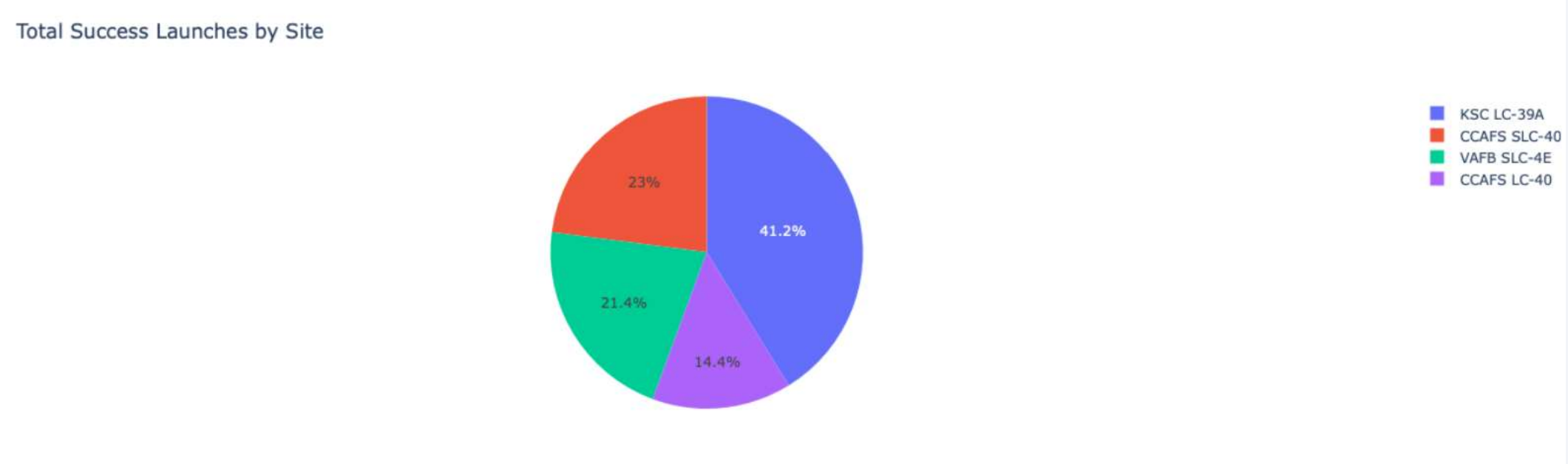
From the analysis of the figure alongside, we can clearly see the important areas close to the launch site KSC LC-39A.

- Close to the railways (16km)
- Close to the highway (20km)
- Close to the coastline (15km)

The site is also close to the city Titusville(16km). Failed launches may cause some damages.

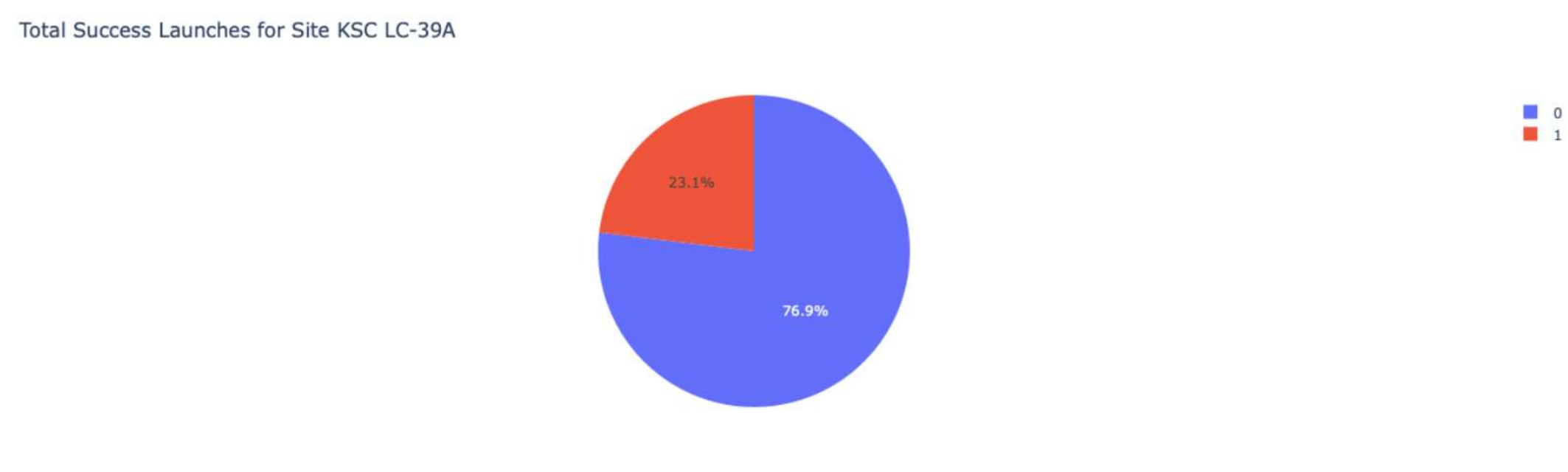

Section 4
Build a Dashboard with Plotly Dash

Launch success count for sites



From the pie chart, we can clearly see that KSC LC-39A has the highest launch success rate out of all the launch sites.

Launch site with the highest launch success ratio



The site KSC LC-39A has the highest ratio of successful launches with a 3:1 success ratio.

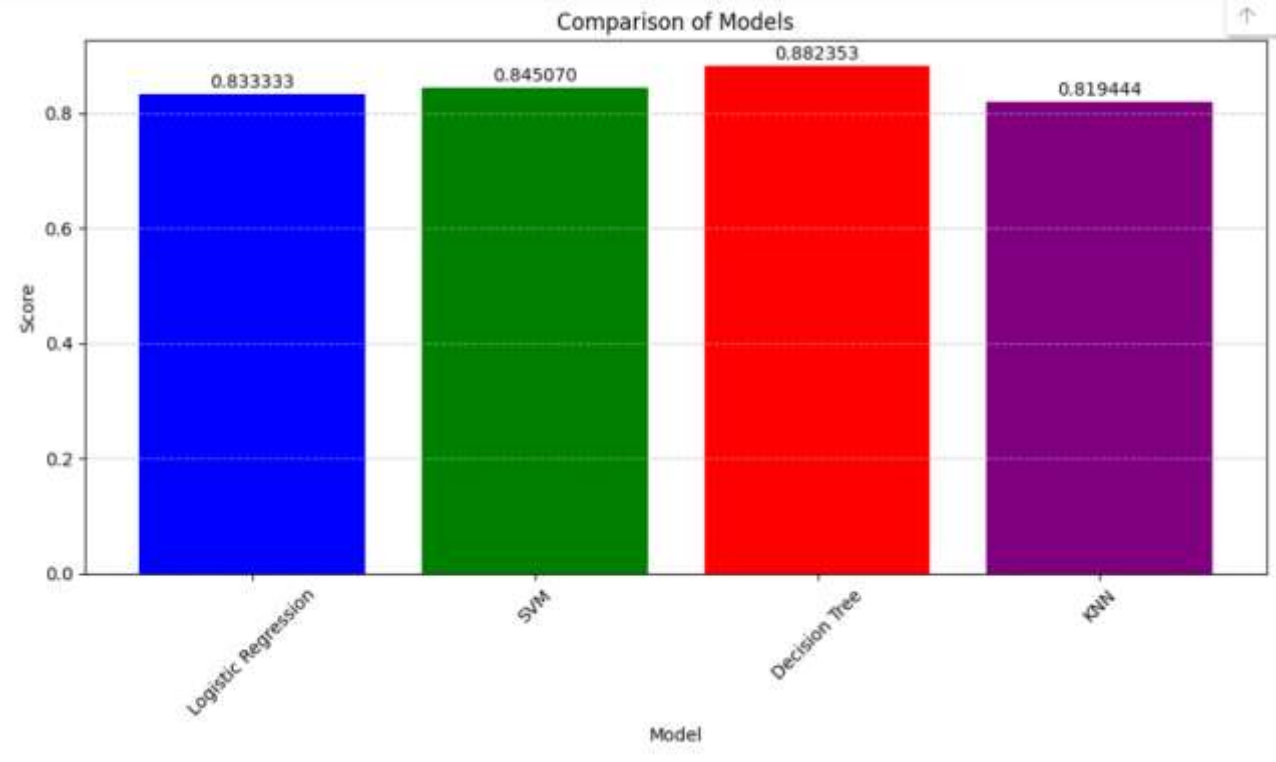
Payload mass vs Launch outcome

From the charts, it can be inferred that payloads between 2000 to 5500 kg have the highest success rates.

Section 5

Predictive Analysis (Classification)

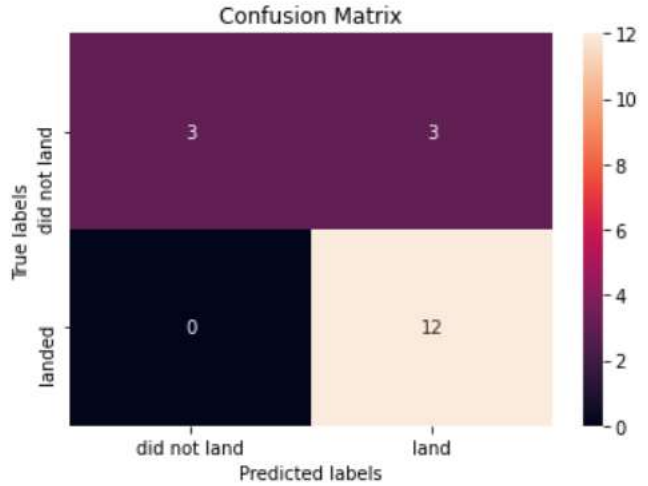
Classification Accuracy



The model with the highest accuracy is the decision tree

Confusion Matrix

```
In [24]: tree_yhat = svm_cv.predict(X_test)
         plot_confusion_matrix(Y_test, tree_yhat)
```



The best model is the decision tree. Given above is its confusion matrix. There are no false positives but the model may give some true negatives.

Conclusions

- Decision Tree Model is the best algorithm for this dataset with close to 90% accuracy.
- 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 as technology advances.
- 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.

Thank you!