IBM Developer
SKILLS NETWORK

Data Science Capstone Project

Lakshmi Priya Ganesan

03-04-2022

Outline

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
- ✓ Interactive analytics demo in screenshots
- ✓ Modelling and Predictive analysis

Introduction

➢ Project background

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost as high as 165 million dollars each. Based on the available dataset and machine learning models, we are going to predict if SpaceX will reuse the first stage and if the Falcon 9 first stage will land successfully.

➢ Problems you want to find answers

- Whether the Falcon 9 first stage will land successfully
- Does the rate of successful landings increase over the years?
- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

Methodology

Methodology

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 for 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

- Build , Tune and Evaluate the model for best results

Data Collection

Data collection

Using REST API and Web Scrapping to get complete information

REST API

Request and parse the SpaceX launch dataset using GET request

Filter Falcon 9 Data

Deal missing values

Wikipedia Web Scrapping

Using Beautiful Soup extract Falcon 9 Launch records

Parsing HTML data to Pandas Data frame

Data Collection – SpaceX API

1 → 2 → 3

1. GET Requests

- Extract Booster version , Launch site coordinates , Payload and core information

2. Filter

- Extract only Falcon 9 data

3. Deal Missing values

- Replace NULL with Payload Mean values

[Data_Collection_SpaceX_API]

Data Collection - Scraping

1

2

3

Beautiful Soup

Using Beautiful soup object extract web data as HTML tables

Extract

- Extract variables and columns from HTML Tables

Create Data Frame

Create pandas data frame with launch records

[Data_Collection_WebScrapping]

Data Wrangling

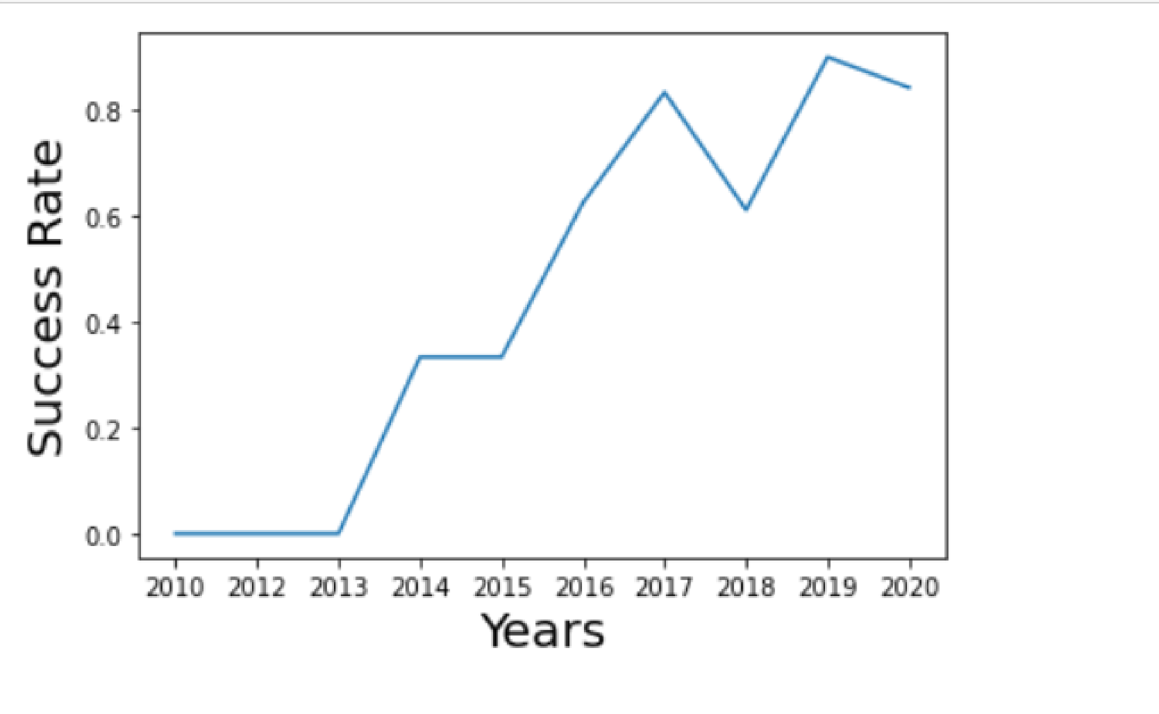
1. Perform exploratory Data analysis and determine training labels
2. Calculate number of launches on each site
3. Calculate number and occurrences of each orbit
4. Calculate number and occurrences of mission outcomes for each orbit type
5. Create landing outcome label for Outcome column
6. Export data to CSV

Data Wrangling

EDA with Data Visualization

Plots for visualization

- Scatter Plot :
  - ✓ Flight number vs Launch Site
  - ✓ Payload and Launch Site
  - ✓ Flight number vs Orbit Type
  - ✓ Payload vs Orbit Type
- Bar Chart :
  - ✓ Success Rate vs Orbit Type
- Line Graph :
  - ✓ Success rates vs Years



EDA_Data_Visualisation

EDA with SQL

- Extract distinct launch sites data

```
%sql select distinct launch_site from SPACEXDATA
```

- Average Payload Mass carried by Booster version F9 V1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATA where booster_version like '%F9 v1.1%'
```

- Total success and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATA group by mission_outcome;
```

[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

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

Plotly_Dash

Predictive Analysis (Classification)

1. Creating a NumPy array from the column "Class" in data
2. Standardizing the data with Standard Scaler, fit and transform
3. Splitting the data into training and testing sets with train_test_split function
4. Creating a GridSearchCV object with cv = 10 to find the best parameters
5. Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
6. Calculating the accuracy on the test data using the method .score() for all models
7. Examining the confusion matrix for all models
8. Finding the method performs best by examining the Jaccard_score and F1_score metrics

[Predictive Analysis]

Results

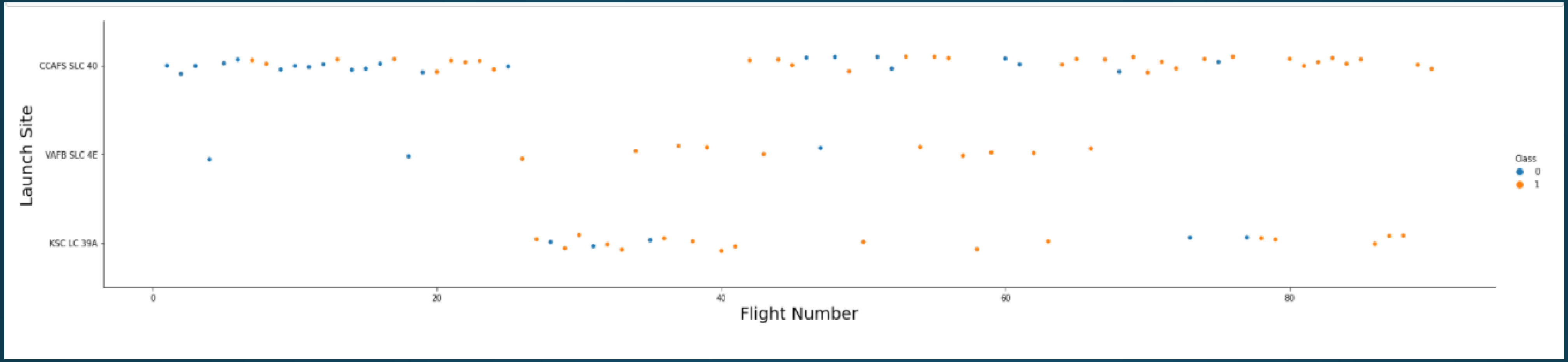
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

EDA with Visualization and SQL

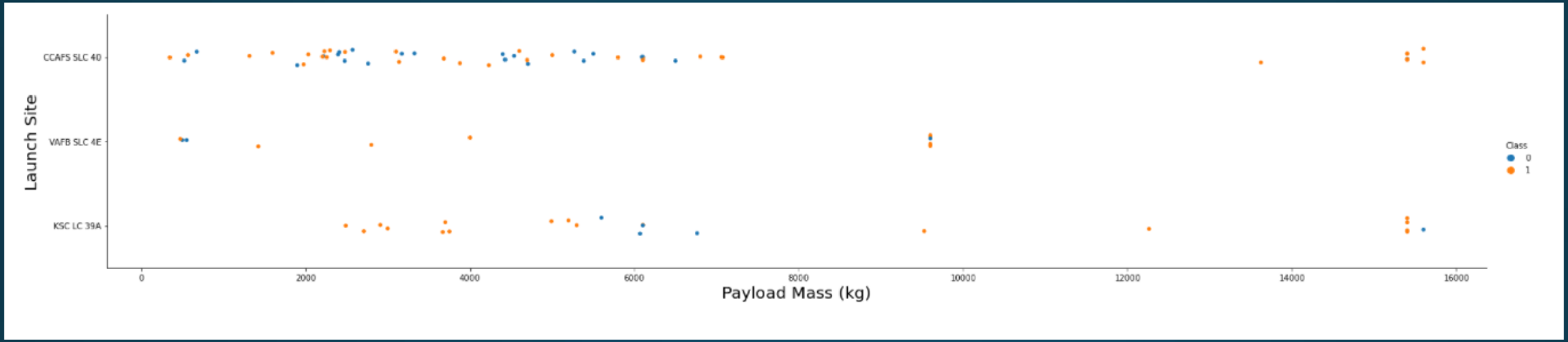
Flight Number vs. Launch Site



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.

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:

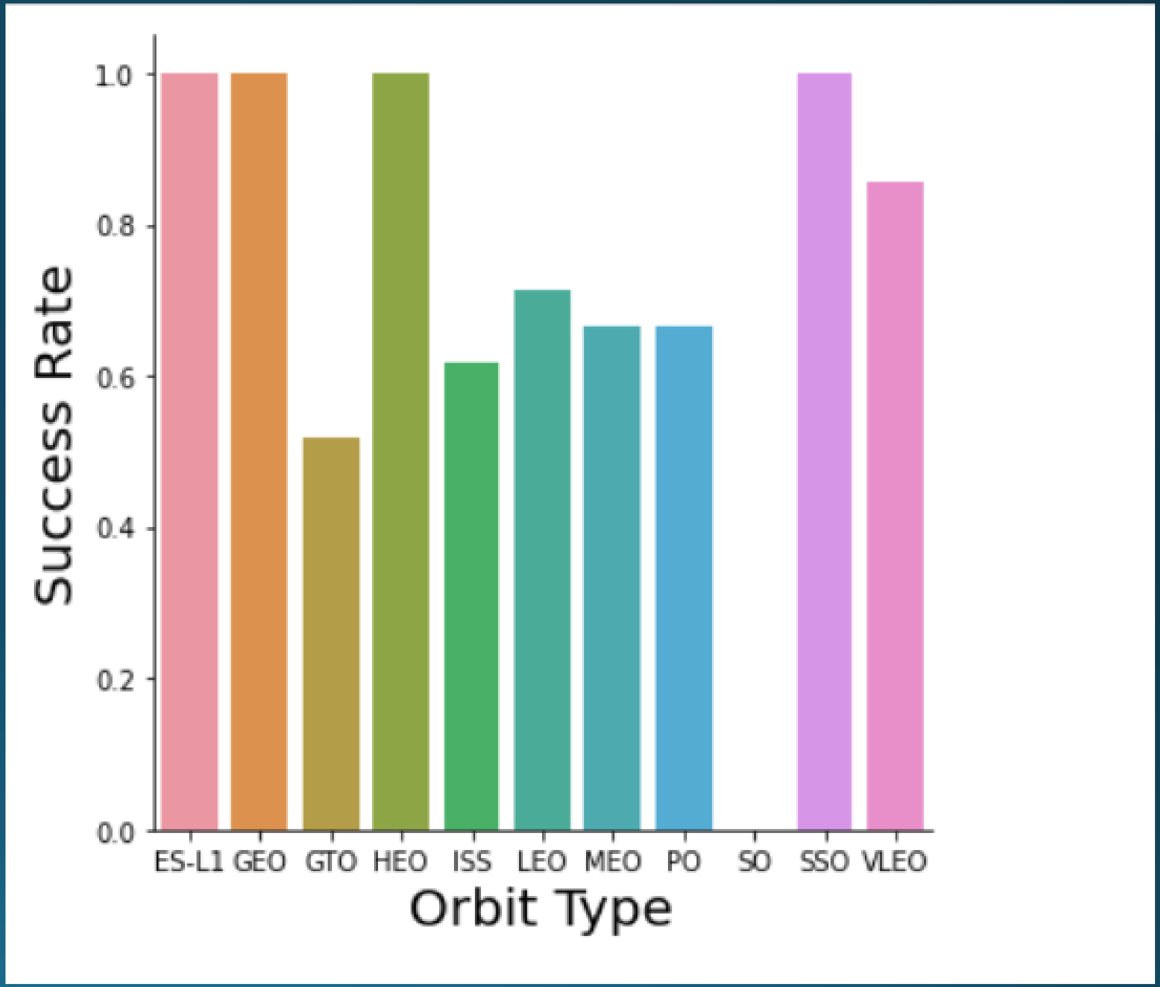
ES-L1, GEO, HEO, SSO

✓Orbits with 0% success rate:

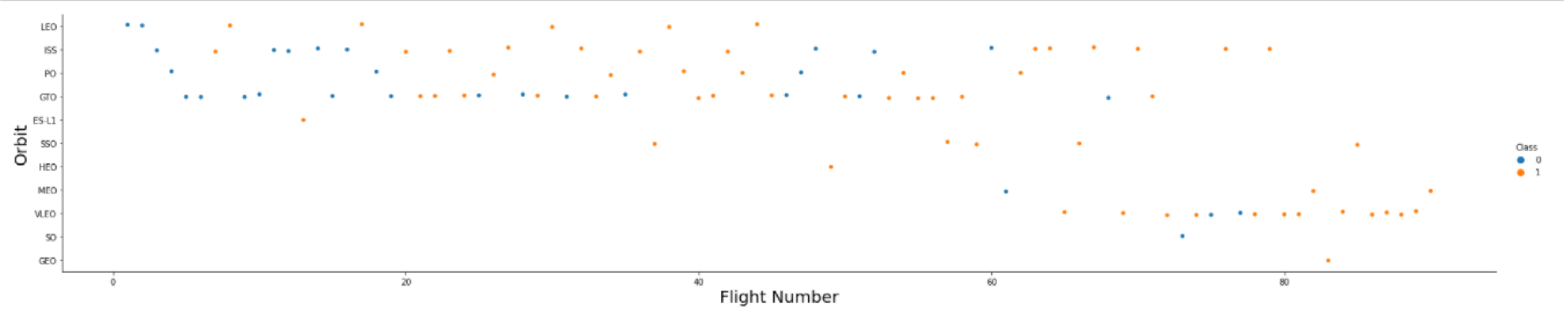
SO

✓Orbits with success rate between 50% and 85%:

GTO, ISS, LEO, MEO, PO



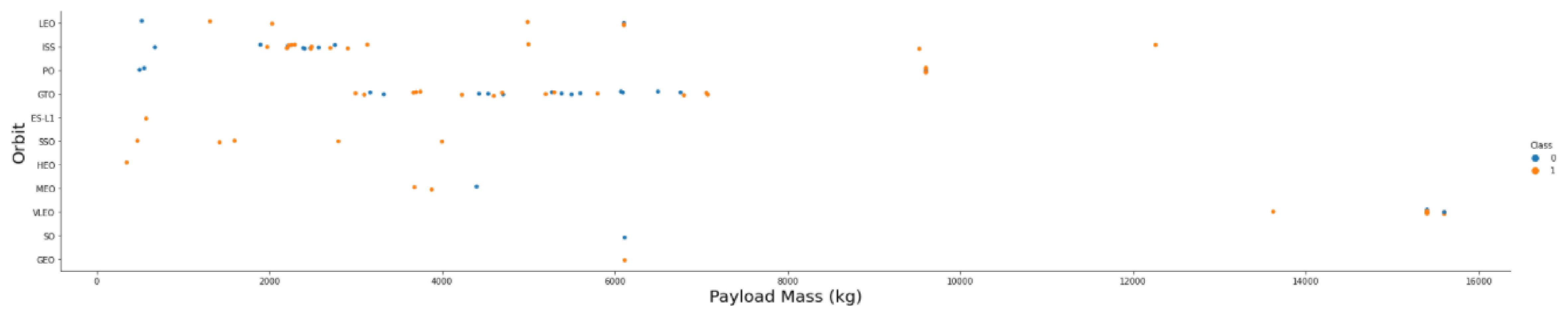
Flight Number vs. Orbit Type



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 GEO orbit

Payload vs. Orbit Type



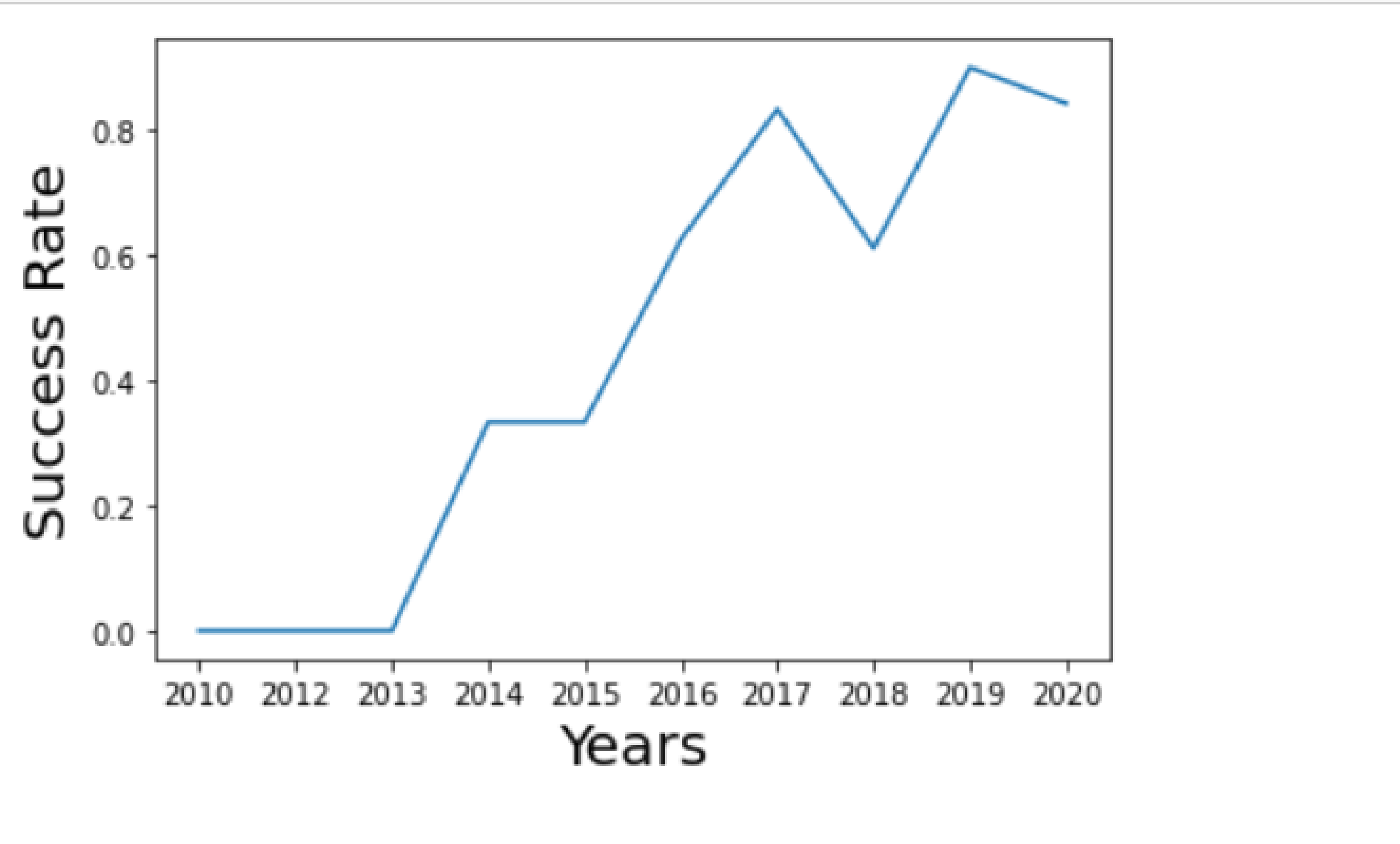
Explanation:

✓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

```
%sql select distinct launch_site from SPACEXDATA;
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

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'

```
%sql select * from SPACEXDATA where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
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	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

```
%sql Select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATA where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
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 average_payload_mass from SPACEXDATA where booster_version like '%F9 v1.1%'
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

average_payload_mass
2534

Explanation:

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

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXDATA where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

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

```
%sql select booster_version from SPACEXDATA where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 a
nd 6000
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
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_number from SPACEXDATA group by mission_outcome;
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

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

```
%sql Select booster_version from SPACEXDATA where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXDATA)
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
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

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATA
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

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_outcomes from SPACEXDATA
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://mjl30932:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

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 date 2010-06-04 and 2017-03-20 in descending order

Interactive map with Folium

All launch sites' location markers on a global map

Explanation:

✓All launch sites are in very close proximity to the coast, as it minimises the risk of having any debris dropping or exploding near people

Coloured Launch records on the map

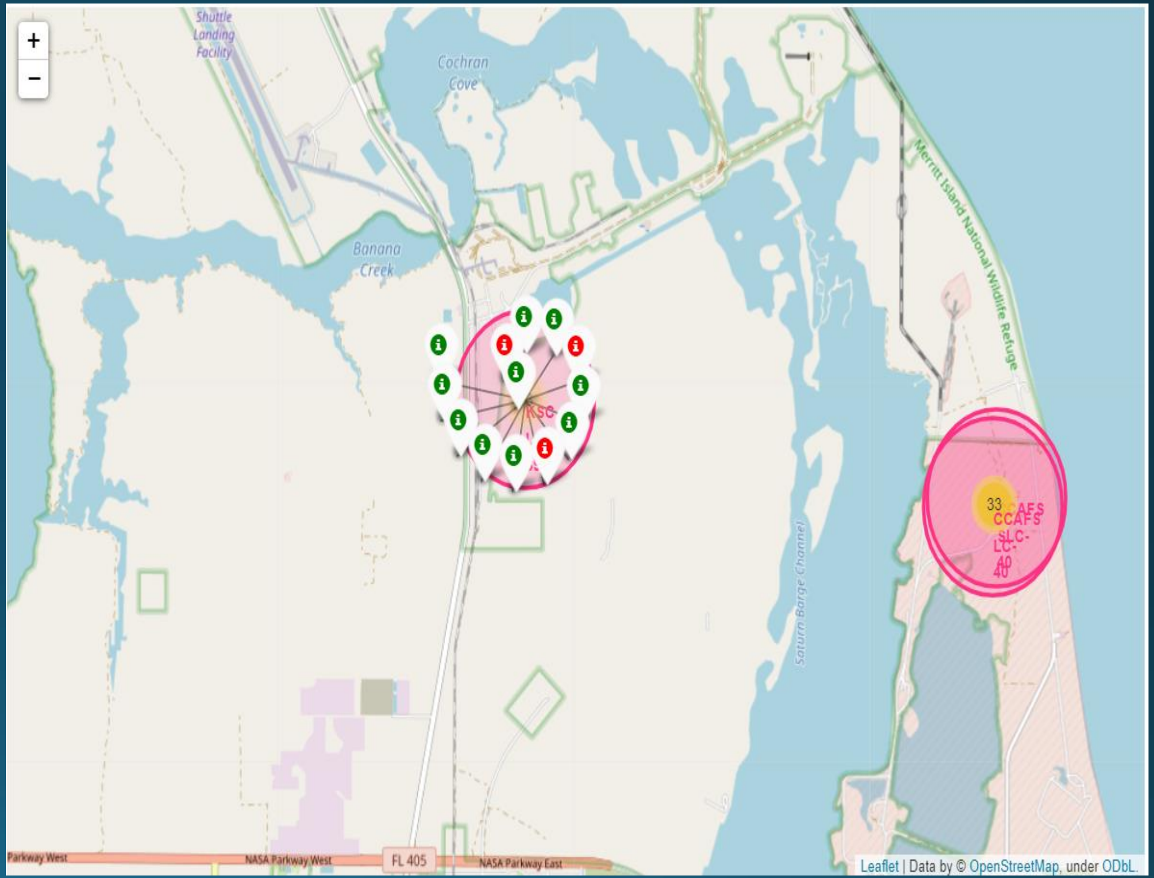
Explanation:

✓From the coloured markers we can easily identify which launch sites have relatively high success rates.

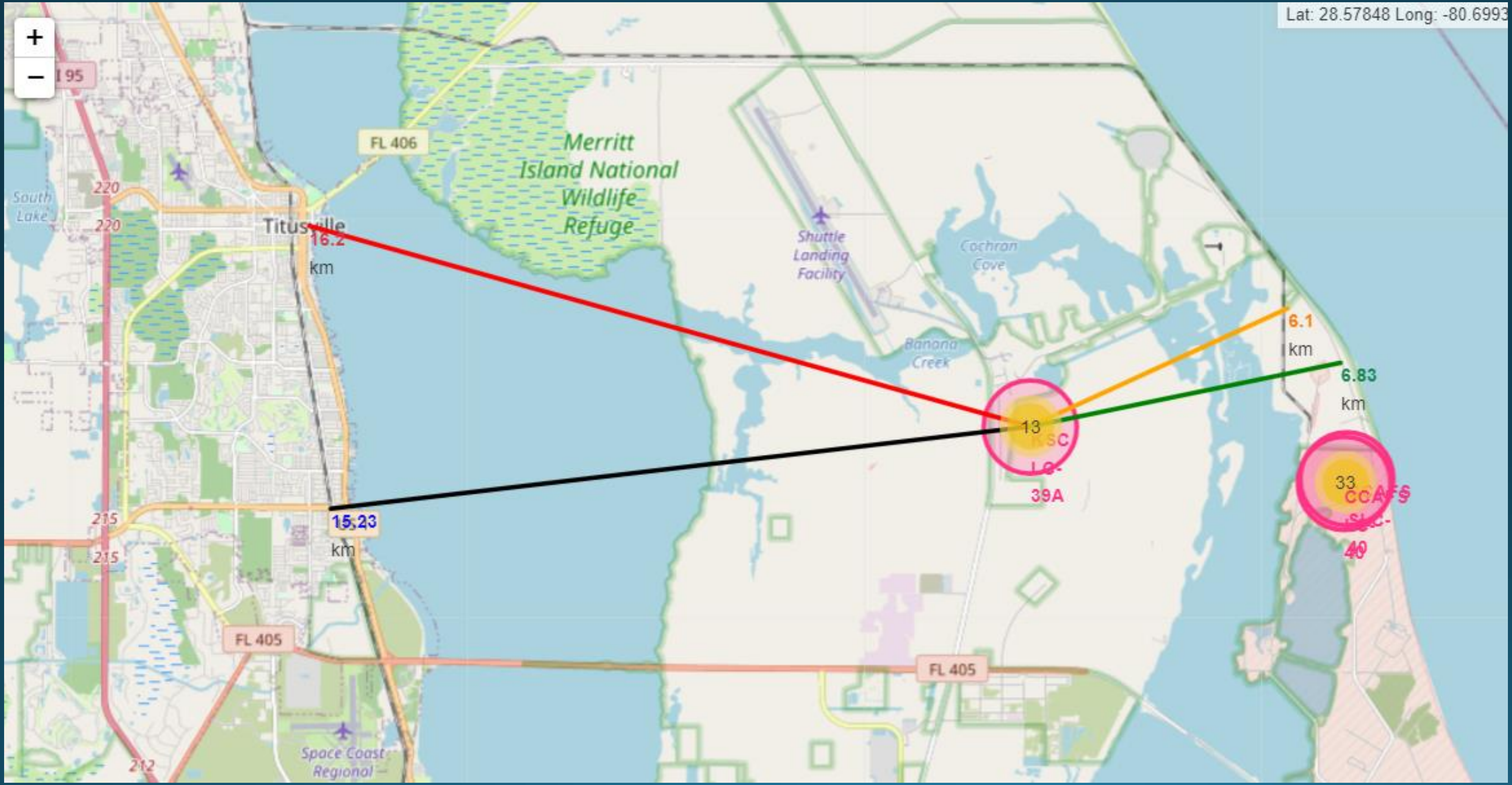
✓Green Marker = Successful Launch

✓Red Marker = Failed Launch

✓Inference : Launch Site KSC LC-39A has a very high Success Rate

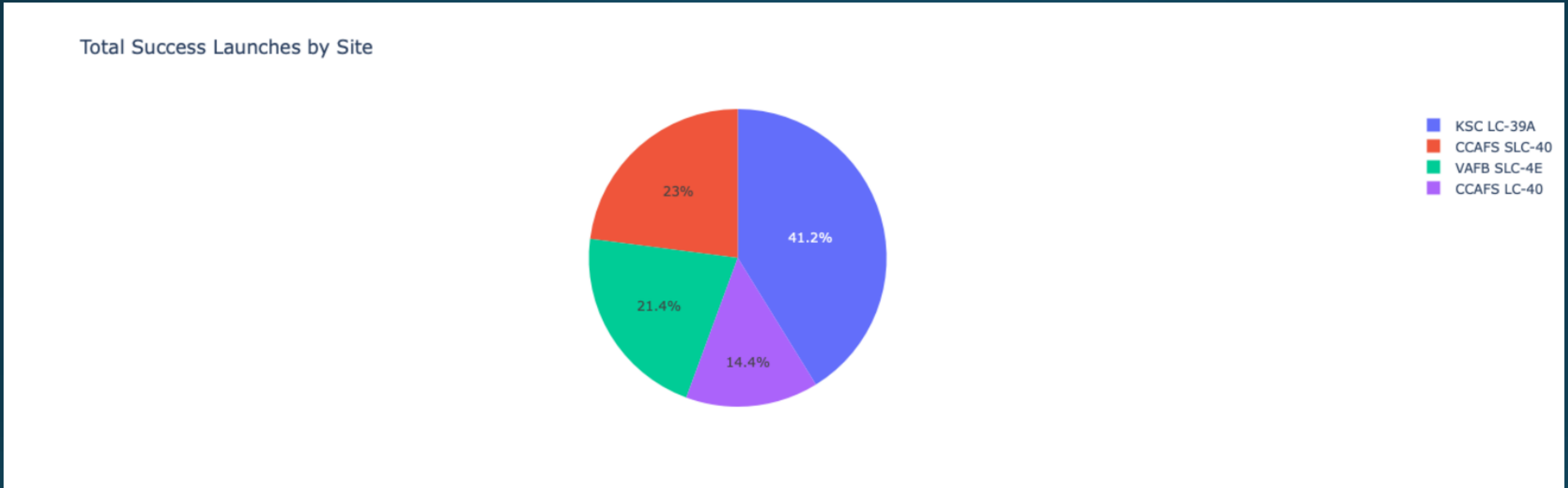


Distance from the launch site KSC LC-39A to its proximities



Build a Dashboard with Plotly Dash

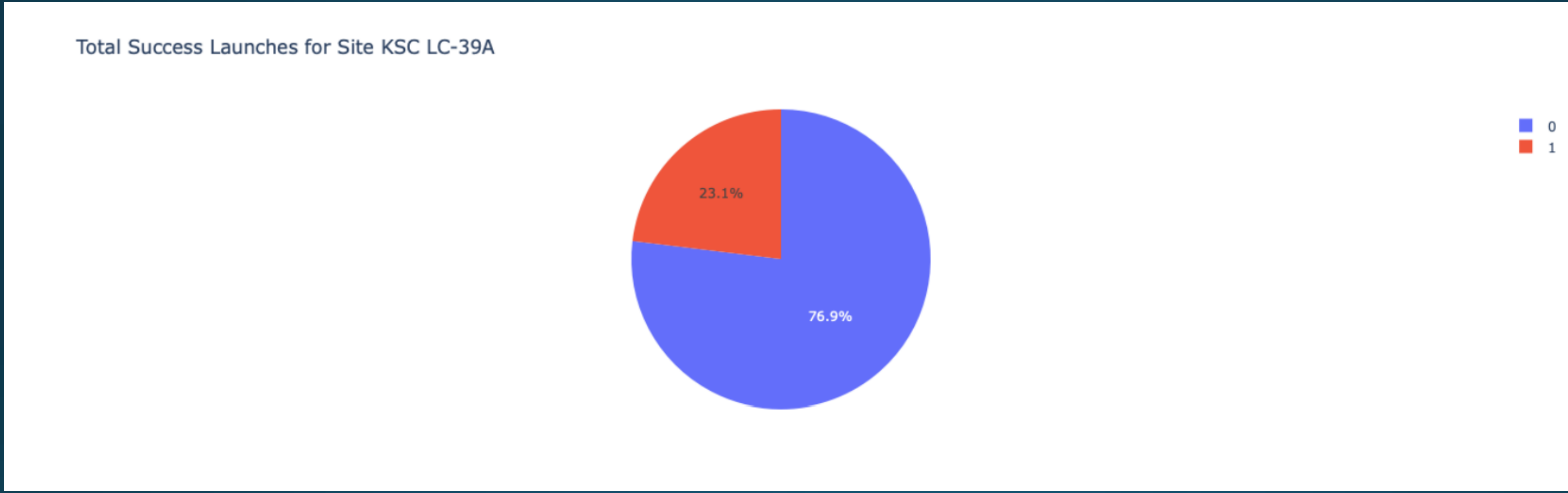
Launch success count for all sites



Explanation:

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

Launch site with highest launch success ratio



Explanation:

✓KSC LC-39A has the highest launch success rate (76.9%).

Payload Mass vs. Launch Outcome for all sites

Explanation:

✓Highest Success rate is when payloads are not heavy .

Predictive analysis (Classification)

Classification Accuracy

Explanation:

✓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 Test Dataset

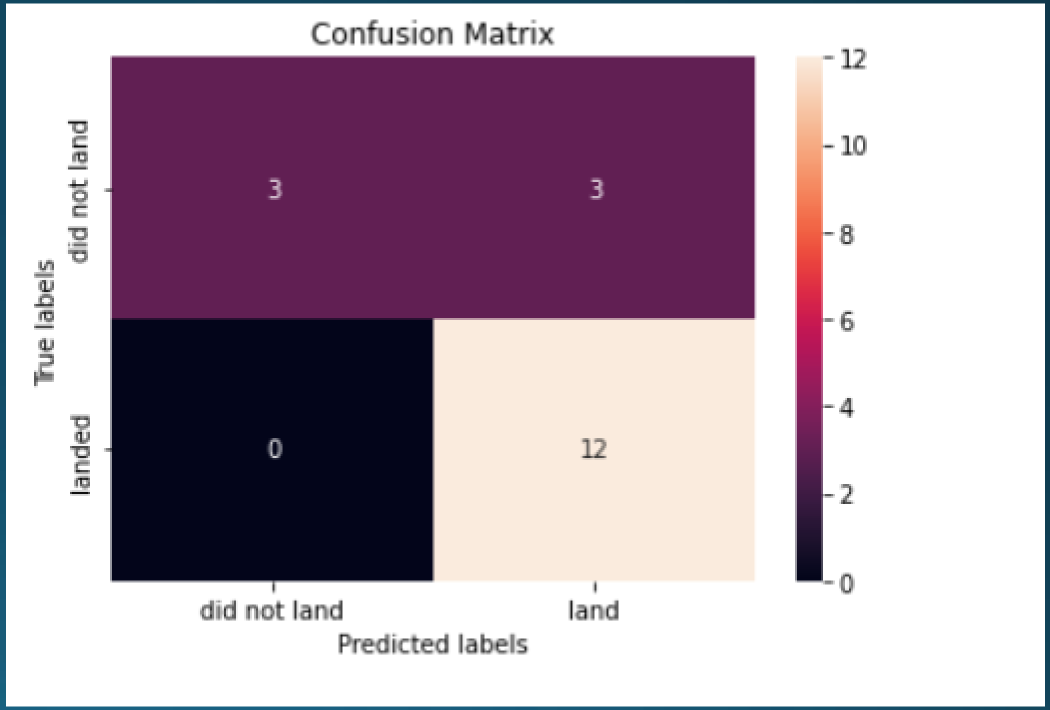
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and accuracy of Entire Dataset

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.840580	0.819444
F1_Score	0.909091	0.916031	0.913386	0.900763
Accuracy	0.866667	0.877778	0.877778	0.855556

Confusion Matrix

Examining the confusion matrix , we can infer that True labels that did not land was predicted incorrectly as landed by the model .



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- When Payload mass is not heavy then it is most likely to be a successful launch .
- The success rate of launches increases over the years.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate

Special Thanks

IBM
Coursera
Instructors