IBM Developer SKILLS NETWORK

Data Science Capstone Project

TING CHONG NA
29.12.2022

OUTLINE

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

EXECUTIVE SUMMARY

Summary of methodologies

- Data Collection API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Visualization
- Interactive Visual Analytics with Folium
- Interactive Dashboard with Ploty Dash
- Prediction Analysis (Classification)

Summary of all Results

- Exploratory Data Analysis (EDA)
- Interactive Visual Analytics and Dashboard
- Prediction Analysis (Classification)

INTRODUCTION

Project background and context

- SpaceX is the most successful company of the commercial space age, which making space travel affordable for everyone.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars.
- Other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Objective / Problems

- To determine the price of each launch
- To determine if SpaceX will reuse the first stage
- Train a machine learning model, create dashboards and use public information to predict if SpaceX will reuse the first stage

METHODOLOGY

Data Collection	• with SpaceX REST API • with Web Scraping
Data Wrangling	• Dealing with Missing Values
Exploratory Data Analysis (EDA)	• with SQL • with Visualization
Interactive Visual Analytics	• Map - Folium • Dashboard – Plotly Dash
Predictive Analysis	• using Classification Models

DATA COLLECTION

SpaceX REST API

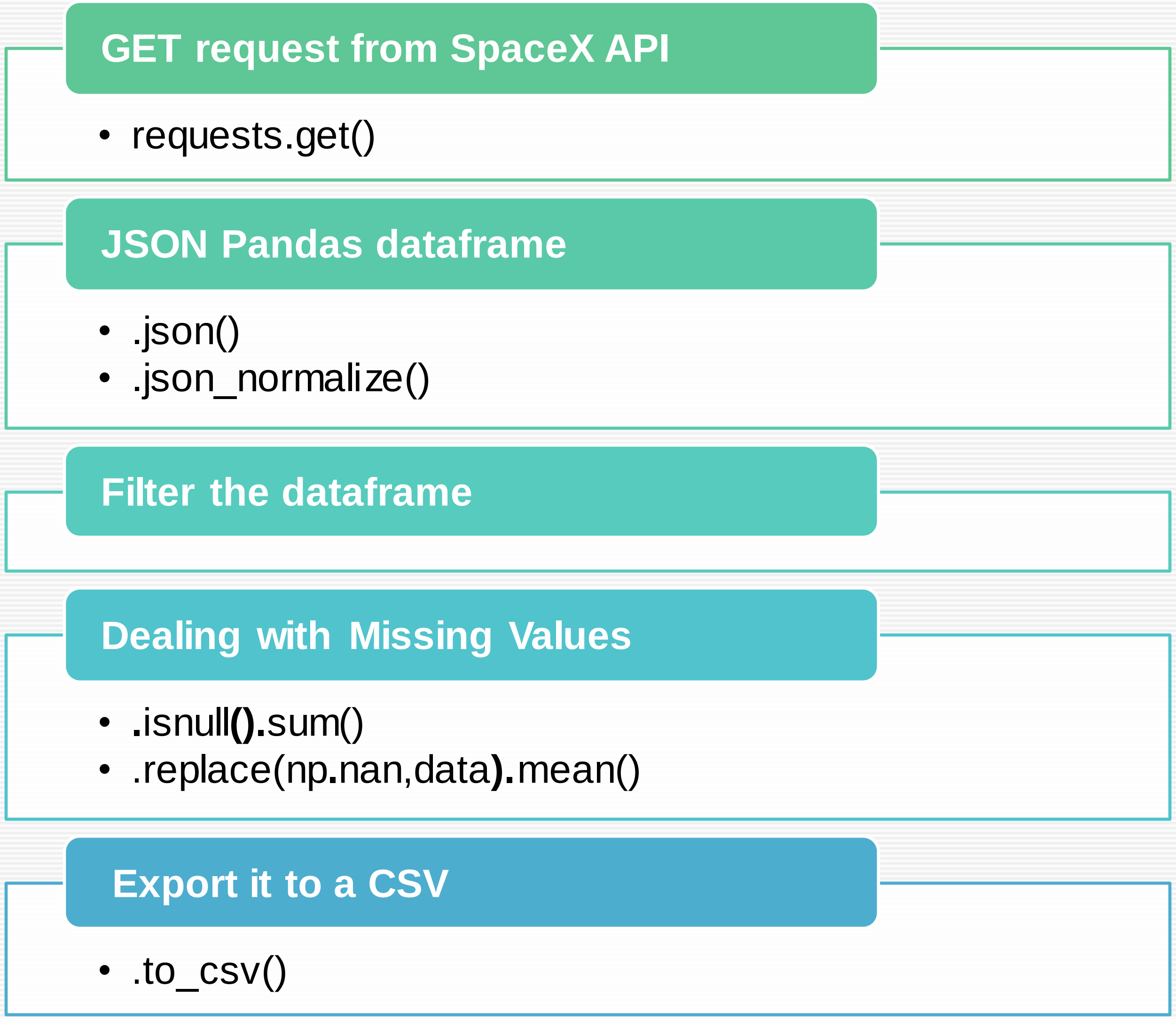
- Request to the SpaceX API
- Clean the requested data

Web Scraping

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

DATA COLLECTION – SPACEX API

GITHUB-DATA COLLECTION API



GET request from SpaceX API

- requests.get()

JSON Pandas dataframe

- .json()
- .json_normalize()

Filter the dataframe

Dealing with Missing Values

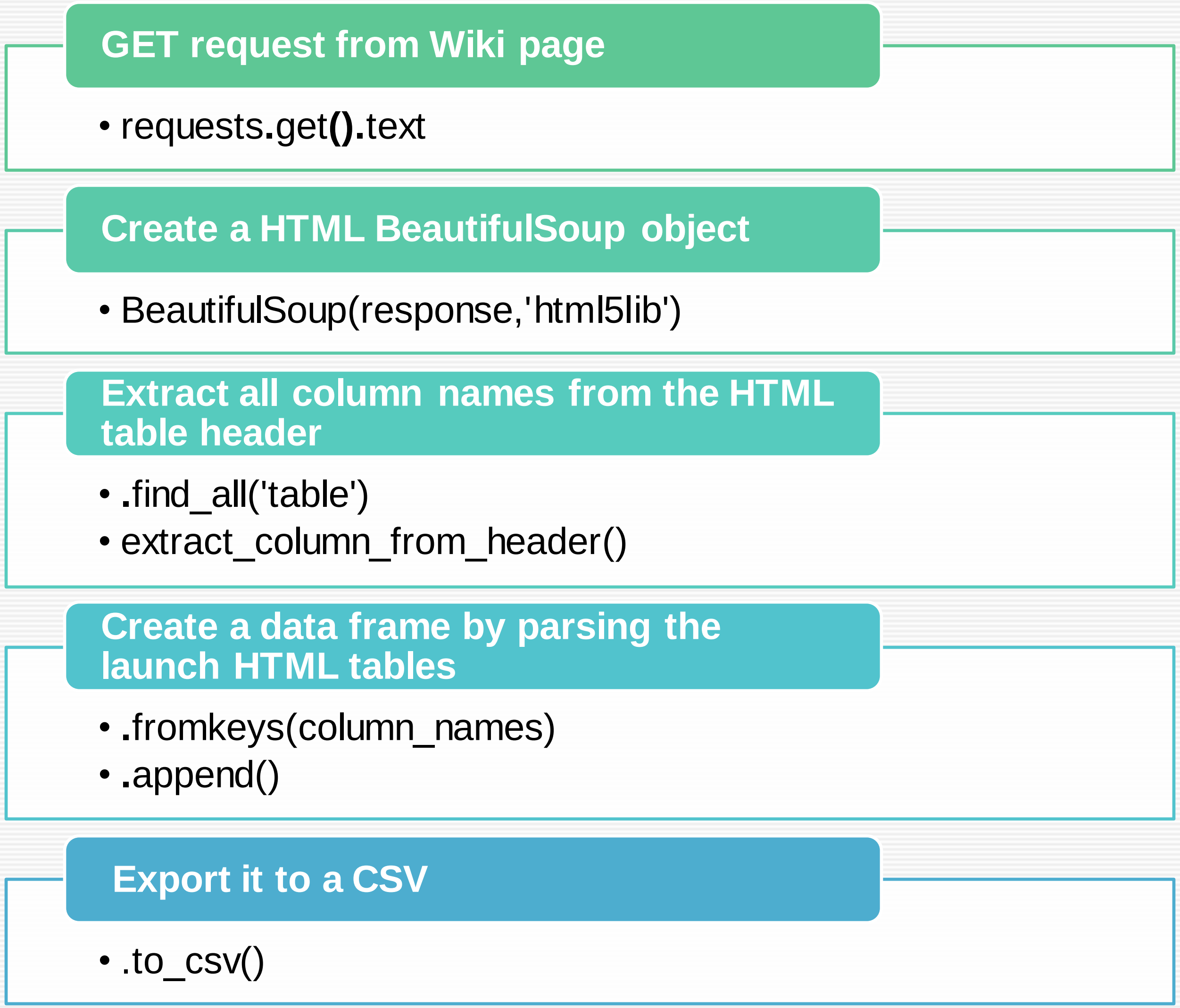
- .isnull().sum()
- .replace(np.nan,data).mean()

Export it to a CSV

- .to_csv()

DATA COLLECTION – WEB SCRAPING

GITHUB-DATA COLLECTION WEB SCRAPING



GET request from Wiki page

- requests.get().text

Create a HTML BeautifulSoup object

- BeautifulSoup(response,'html5lib')

Extract all column names from the HTML table header

- .find_all('table')
- extract_column_from_header()

Create a data frame by parsing the launch HTML tables

- .fromkeys(column_names)
- .append()

Export it to a CSV

- .to_csv()

DATA WRANGLING

GITHUB-DATAWRANGLING EDA

- Perform some Exploratory Data Analysis (EDA) to find some patterns in the data

- Determine what would be the label for training supervised models

- Convert those outcomes into Training Labels

- `1` = booster successfully
- `0` = booster unsuccessful.

Calculate the number of launches on each site

- .value_counts()

Calculate the number and occurrence of each orbit

- .value_counts()

Calculate the number and occurence of mission outcome per orbit type

- .value_counts()

Create a landing outcome label from Outcome column

Export it to a CSV

- .to_csv()

EDA WITH SQL

GITHUB-EDA SQL

DISTINCT()	• Display All Launch Site Names
LIKE 'CCA%' / LIMIT	• Display 5 records where Launch Site Names Begin with 'CCA'
SUM()	• Display Total Payload Mass
AVG()	• Display Average Payload Mass by F9 v1.1
MIN()	• List First Successful Ground Landing Date
AND	• List Successful Drone Ship Landing with Payload between 4000 and 6000
COUNT()	• List Total Number of Successful and Failure Mission Outcomes
SUBQUERY	• List Boosters Carried Maximum Payload
YEAR()	• List 2015 Launch Records
BETWEEN...AND... / DESC	• Rank Landing Outcomes Between 2010-06-04 and 2017-03-20, in descending order

EDA WITH DATA VISUALIZATION

GITHUB-EDA DATA VISUALIZATION

SCATTER POINT CHART

- Show relationship between 2 different numeric variables
- sns.catplot(x,y,hue,data)

BAR CHART

- Best used for categorical data, compare between different groups
- sns.barplot(x,y,hue,data)

LINE CHART

- Track changes over a periods of time
- sns.lineplot(x,y,data)

INTERACTIVE MAP WITH FOLIUM

[GITHUB-INTERACTIVE MAP WITH FOLIUM]

MARKER

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- folium.map.Marker(coordinate, icon=DivIcon (icon_size,icon_anchor, html='<div style="font-size; color;"><b>%s</b></div>' % 'label',))

CIRCLE

- Add a highlighted Circle area with a text label on a specific coordinate
- folium.Circle(coordinate, radius, color, fill=True) .add_child(folium.Popup(...))

POLYLINE

- Draw a line between 2 points
- folium.PolyLine(locations=coordinates, weight=1)

INTERACTIVE DASHBOARD WITH PLOTLY DASH

GITHUB-INTERACTIVE DASHBOARD WITH PLOTLY DASH

DROPDOWN INPUT COMPONENT

- To select different launch sites
- dcc.Dropdown(id, options=[{'label','value'},{'label', 'value'},...], value, placeholder, searchable=True)

PIE CHART

- Add a callback function to render the Pie Chart based on selected site dropdown
- @app.callback(Output(component_id='PieChart', component_property='figure'), Input(component_id='Dropdown', component_property='value'))
- px.pie(data, values, names, title)

RANGE SLIDER

- Add a Range Slider to Select Payload
- dcc.RangeSlider(id, min, max, step, marks, value=[min_value, max_value])

SCATTER PLOT

- Add a callback function to render the Scatter Plot
- Observe how payload may be correlated with mission outcomes for selected sites
- @app.callback(Output(component_id='ScatterPlot', component_property='figure'), [Input(component_id='Dropdown', component_property='value'), Input(component_id='RangeSlider', component_property='value)])
- px.scatter(data, x, y, color, title)

PREDICTIVE ANALYSIS (CLASSIFICATION)

GITHUB-PREDICTIVE ANALYSIS (CLASSIFICATION)

Create a column for the class

- .to_numpy()

Standardize the data

- preprocessing.StandardScaler().fit(X).transform(X)

Split into Training data and Test data

- train_test_split(X,Y,test_size,random_state)

Model

- lr=LogisticRegression()
- svm = SVC()
- tree = DecisionTreeClassifier()
- KNN = KNeighborsClassifier()

Apply GridSearchCV object on Models

- GridSearchCV(estimators,parameters,cv).fit(X_train,Y_train)

Find the Best Parameters and Accuracy on the Validation data

- .best_params_
- .best_score_

Calculate the Accuracy on the Test data

- .score(X_test,Y_test)

Examining the Confusion Matrix

- .predict(X_test)
- plot_confusion_matrix(Y_test,yhat)

RESULTS

Exploratory Data Analysis (EDA)

with SQL

with Visualization

Interactive Analytics Demo with Screenshots

Map – Folium

Dashboard – Plotly Dash

Predictive Analysis

Classification Models

EDA WITH SQL

All Launch Site Names

Launch Site Names Begin with 'CCA'

Total Payload Mass

Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

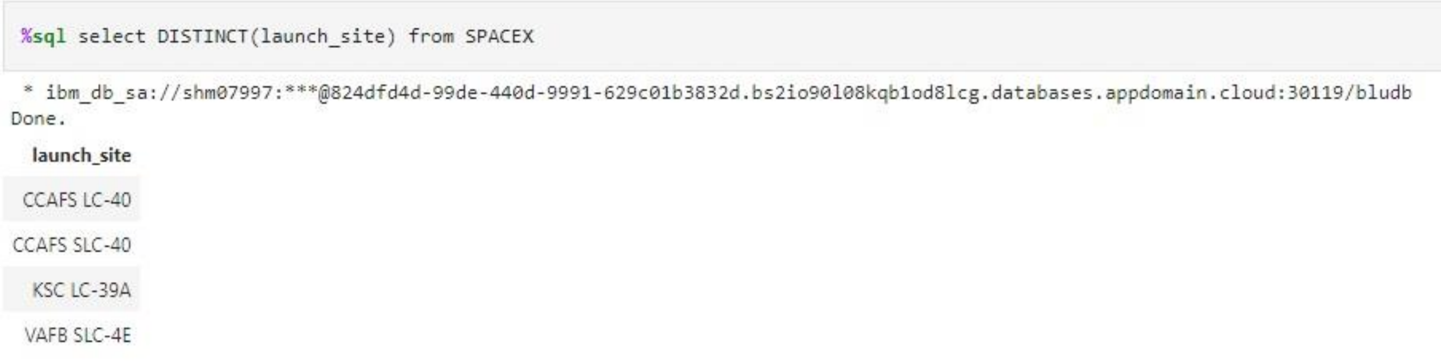
Successful Drone Ship Landing with Payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

2015 Launch Records

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



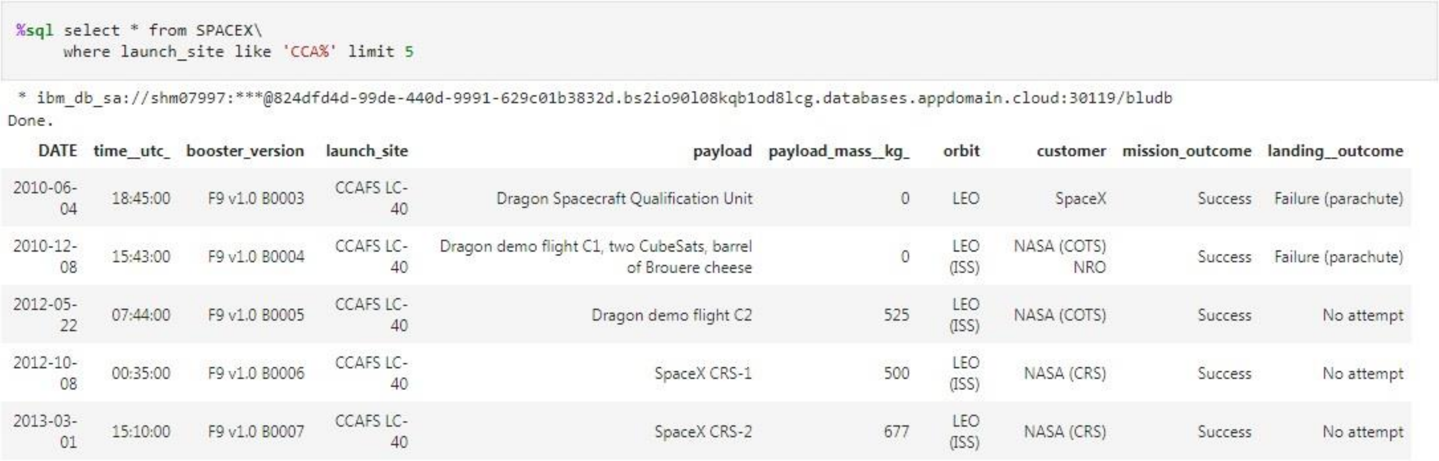
```
%sql select DISTINCT(launch_site) from SPACEX
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

Q: Find the names of the unique launch sites

ANS: select DISTINCT(launch_site) from SPACEX



```
%sql select * from SPACEX\
     where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/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

Q: Find 5 records where launch sites begin with `CCA`

ANS: select * from SPACEX where launch_site like 'CCA%' limit 5

EDA WITH SQL

All Launch Site Names

Launch Site Names Begin with 'CCA'

Total Payload Mass

Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

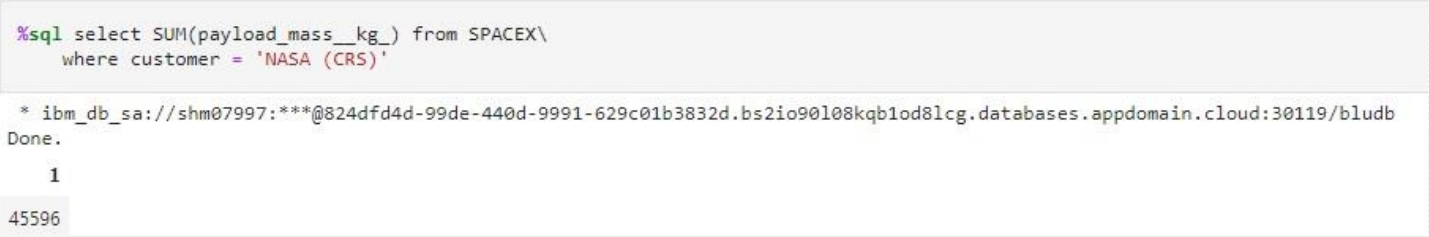
Successful Drone Ship Landing with Payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

2015 Launch Records

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

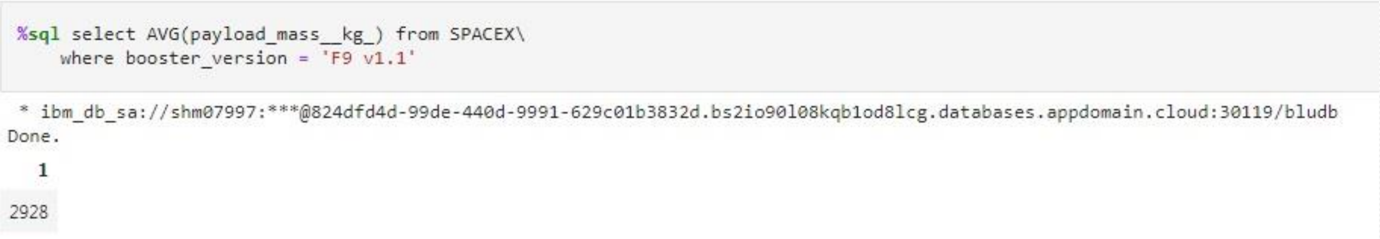


```
%sql select SUM(payload_mass__kg_) from SPACEX\
    where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
    1
45596
```

Q: Calculate the total payload carried by boosters from NASA

ANS: select SUM(payload_mass__kg_) from SPACEX where customer = 'NASA (CRS)'



```
%sql select AVG(payload_mass__kg_) from SPACEX\
    where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
   1
2928
```

Q: Calculate the average payload mass carried by booster version F9 v1.1

ANS: select AVG(payload_mass__kg_) from SPACEX where booster_version = 'F9 v1.1'

EDA WITH SQL

All Launch Site Names

Launch Site Names Begin with 'CCA'

Total Payload Mass

Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

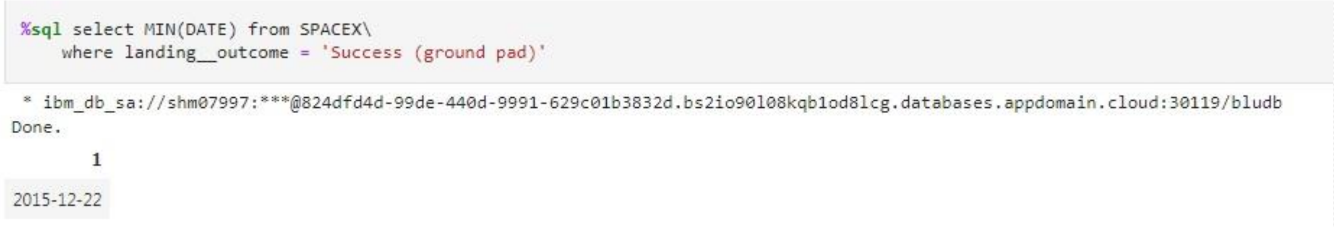
Successful Drone Ship Landing with Payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

2015 Launch Records

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

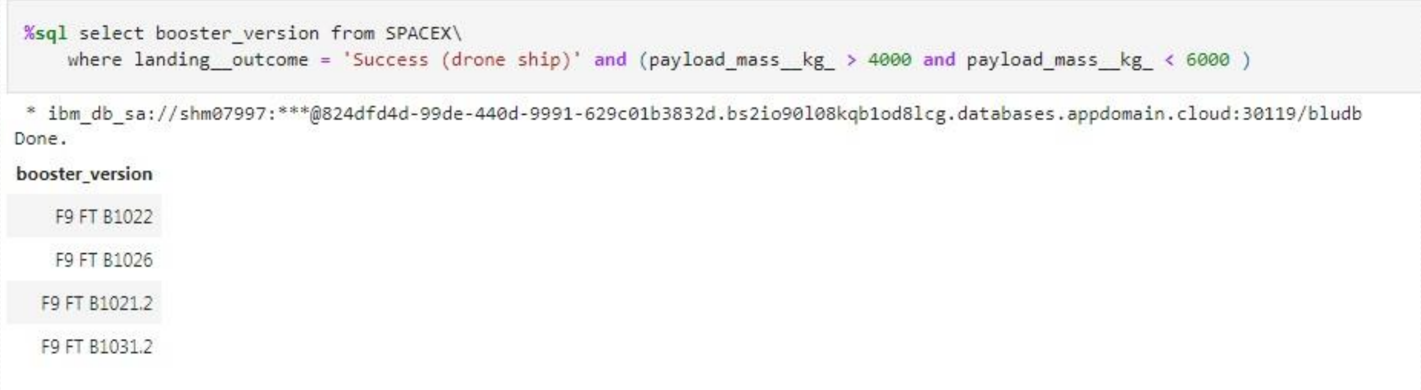


```
%sql select MIN(DATE) from SPACEX\
    where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
         1
2015-12-22
```

Q: Find the dates of the first successful landing outcome on ground pad

ANS: select MIN(DATE) from SPACEX where landing__outcome = 'Success (ground pad)'



```
%sql select booster_version from SPACEX\
    where landing__outcome = 'Success (drone ship)' and (payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000 )
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
booster_version
    F9 FT B1022
    F9 FT B1026
  F9 FT B1021.2
  F9 FT B1031.2
```

Q: List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

ANS: select booster_version from SPACEX where landing__outcome = 'Success (drone ship)' and (payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000)

EDA WITH SQL

All Launch Site Names

Launch Site Names Begin with 'CCA'

Total Payload Mass

Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

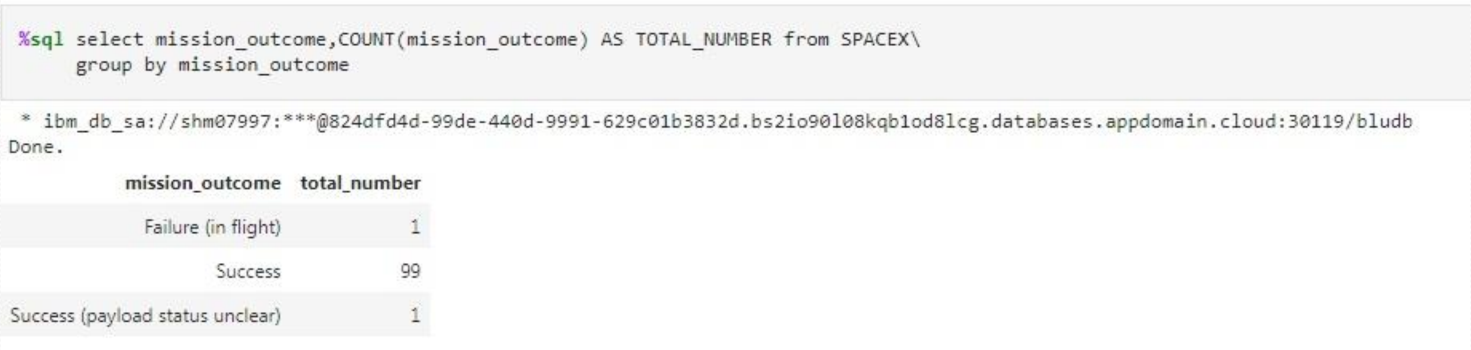
Successful Drone Ship Landing with Payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

2015 Launch Records

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



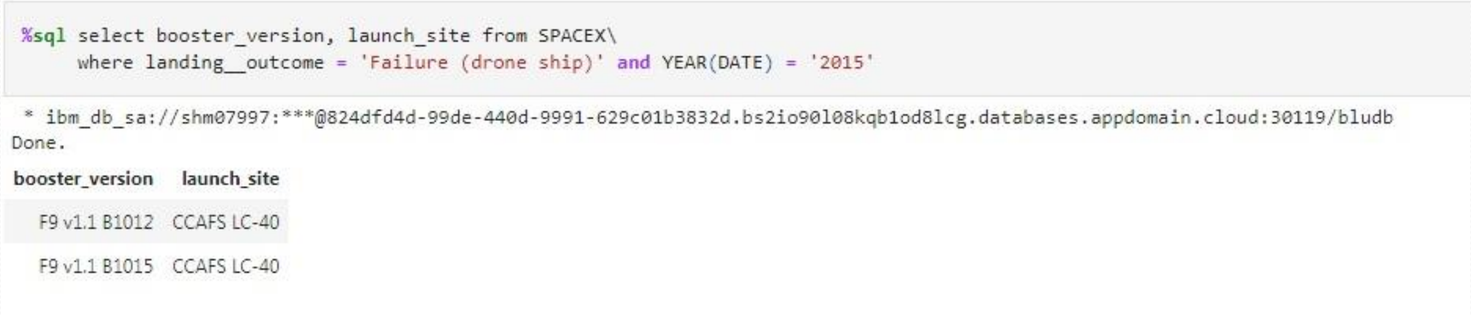
```
%sql select mission_outcome,COUNT(mission_outcome) AS TOTAL_NUMBER from SPACEX\
     group by mission_outcome
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

Q: Calculate the total number of successful and failure mission outcomes

ANS: select mission_outcome,COUNT(mission_outcome) AS TOTAL_NUMBER from SPACEX group by mission_outcome



```
%sql select booster_version, launch_site from SPACEX\
     where landing__outcome = 'Failure (drone ship)' and YEAR(DATE) = '2015'
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Q: List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

ANS: select booster_version, launch_site from SPACEX where landing__outcome = 'Failure (drone ship)' and YEAR(DATE) = '2015'

EDA WITH SQL

All Launch Site Names

Launch Site Names Begin with 'CCA'

Total Payload Mass

Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

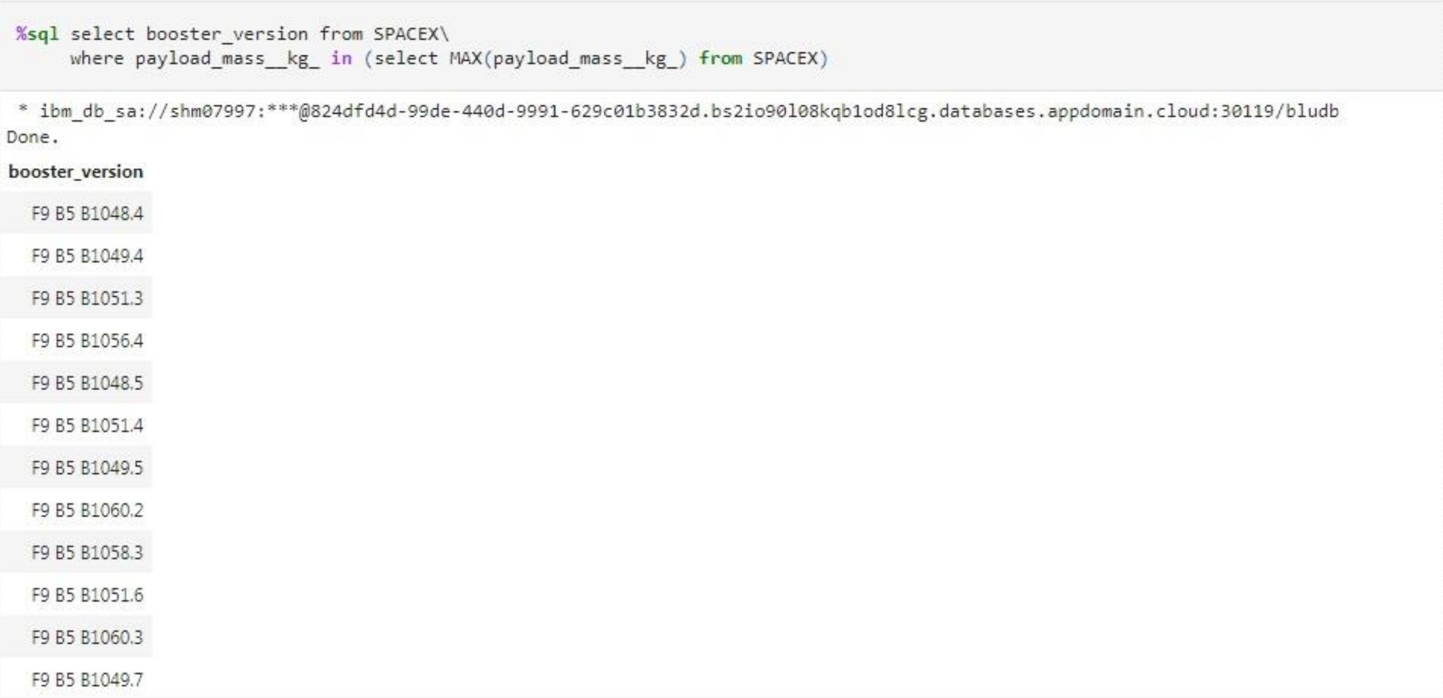
Successful Drone Ship Landing with Payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

2015 Launch Records

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



```
%sql select booster_version from SPACEX\
     where payload_mass__kg_ in (select MAX(payload_mass__kg_) from SPACEX)
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/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

Q: List the names of the booster which have carried the maximum payload mass

ANS: select booster_version from SPACEX where payload_mass__kg_ in (select MAX(payload_mass__kg_) from SPACEX)

EDA WITH SQL

All Launch Site Names

Launch Site Names Begin with 'CCA'

Total Payload Mass

Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

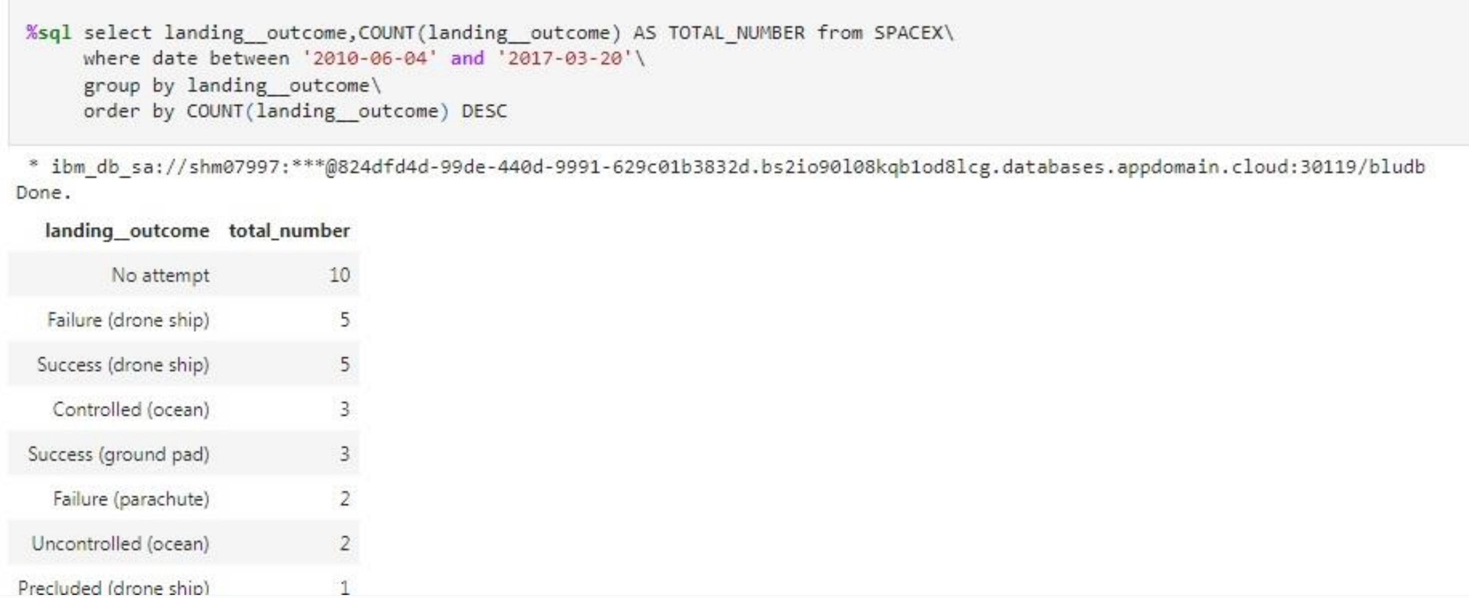
Successful Drone Ship Landing
with Payload between 4000 and 6000

Total Number of Successful
and Failure Mission Outcomes

Boosters Carried Maximum Payload

2015 Launch Records

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



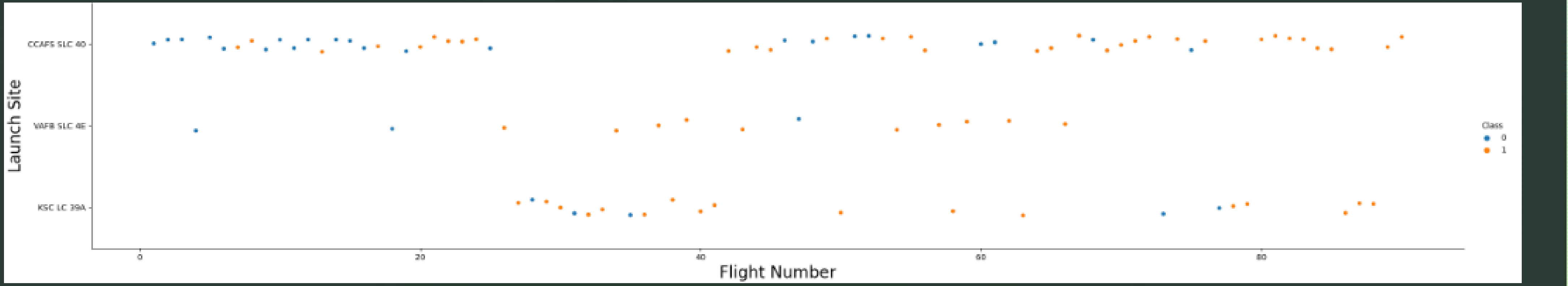
```
%sql select landing__outcome,COUNT(landing__outcome) AS TOTAL_NUMBER from SPACEX\
     where date between '2010-06-04' and '2017-03-20'\
     group by landing__outcome\
     order by COUNT(landing__outcome) DESC
```

```
 * ibm_db_sa://shm07997:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

landing__outcome	total_number
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

Q: Rank 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

ANS: select landing__outcome,COUNT(landing__outcome) AS TOTAL_NUMBER from SPACEX where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by COUNT(landing__outcome) DESC

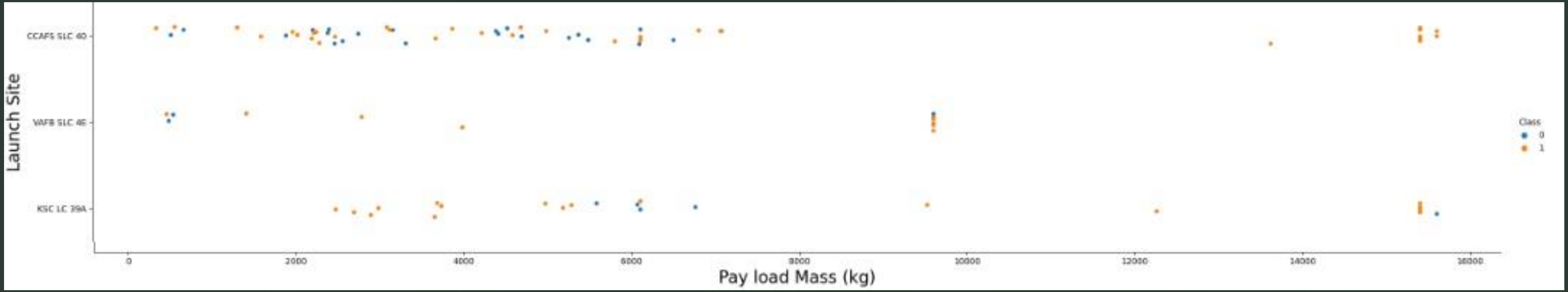


EDA WITH DATA VISUALIZATION

Flight Number vs.
Launch Site
(Scatter Plot)

Explanation:

KSC LC 39A and VAFB SLC 4E have higher success rate (Class 1) , compare to CCAFS LC-40.

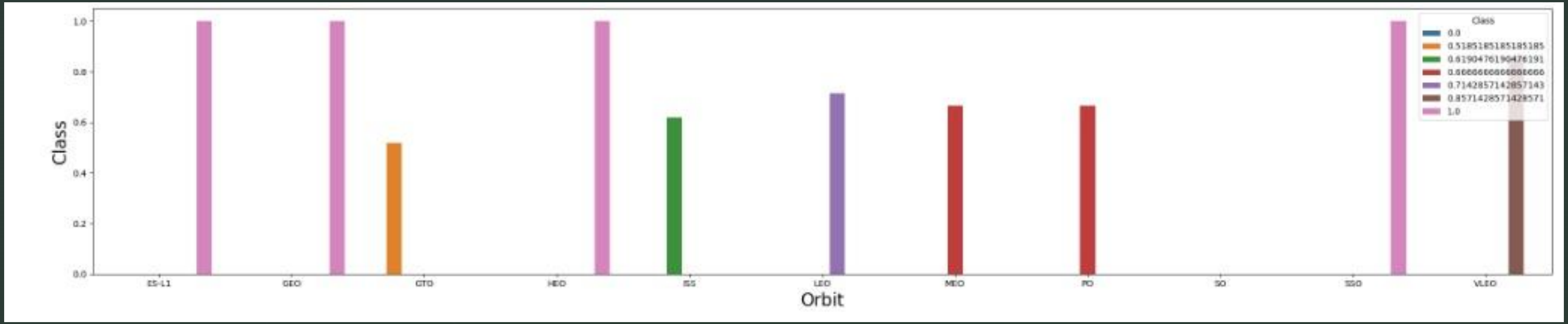


EDA WITH DATA VISUALIZATION

Payload vs. Launch Site (Scatter Plot)

Explanation:

For the VAFB-SLC launch site, there are no rockets launched for heavy Payload mass (greater than 10000).



EDA WITH DATA VISUALIZATION

Success Rate vs.
Orbit Type
(Bar Chart)

Explanation:

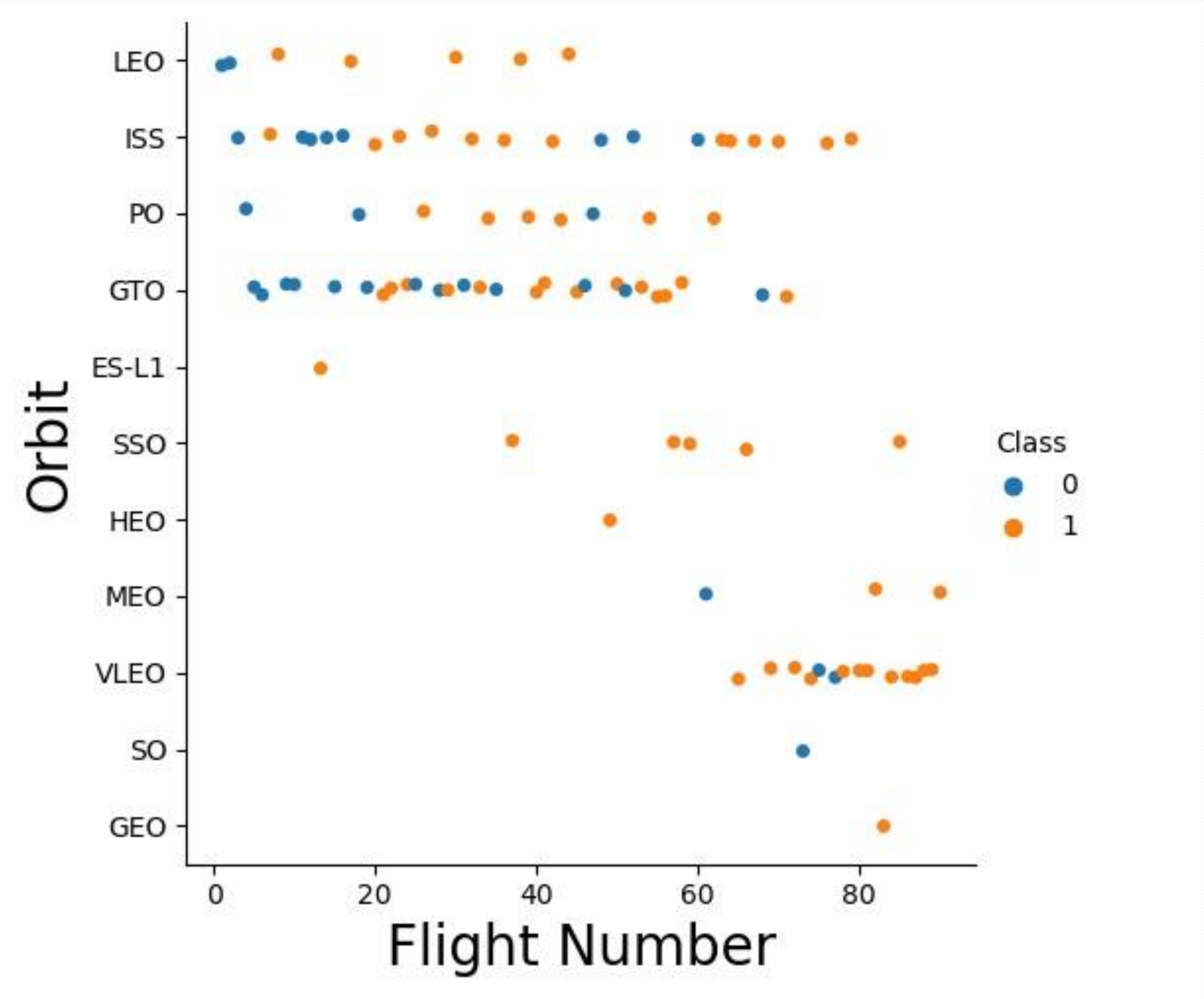
Orbits ES-L1, GEO, HEO, SSO have highest Sucess Rate (100%).
Orbits SO has lowest Sucess Rate (0%).

EDA WITH DATA VISUALIZATION

Flight Number vs. Orbit Type (Scatter Plot)

Explanation:

In the LEO Orbit, the Success appears related to the number of Flights.
There seems to be no relationship between Flight number when in GTO Orbit.

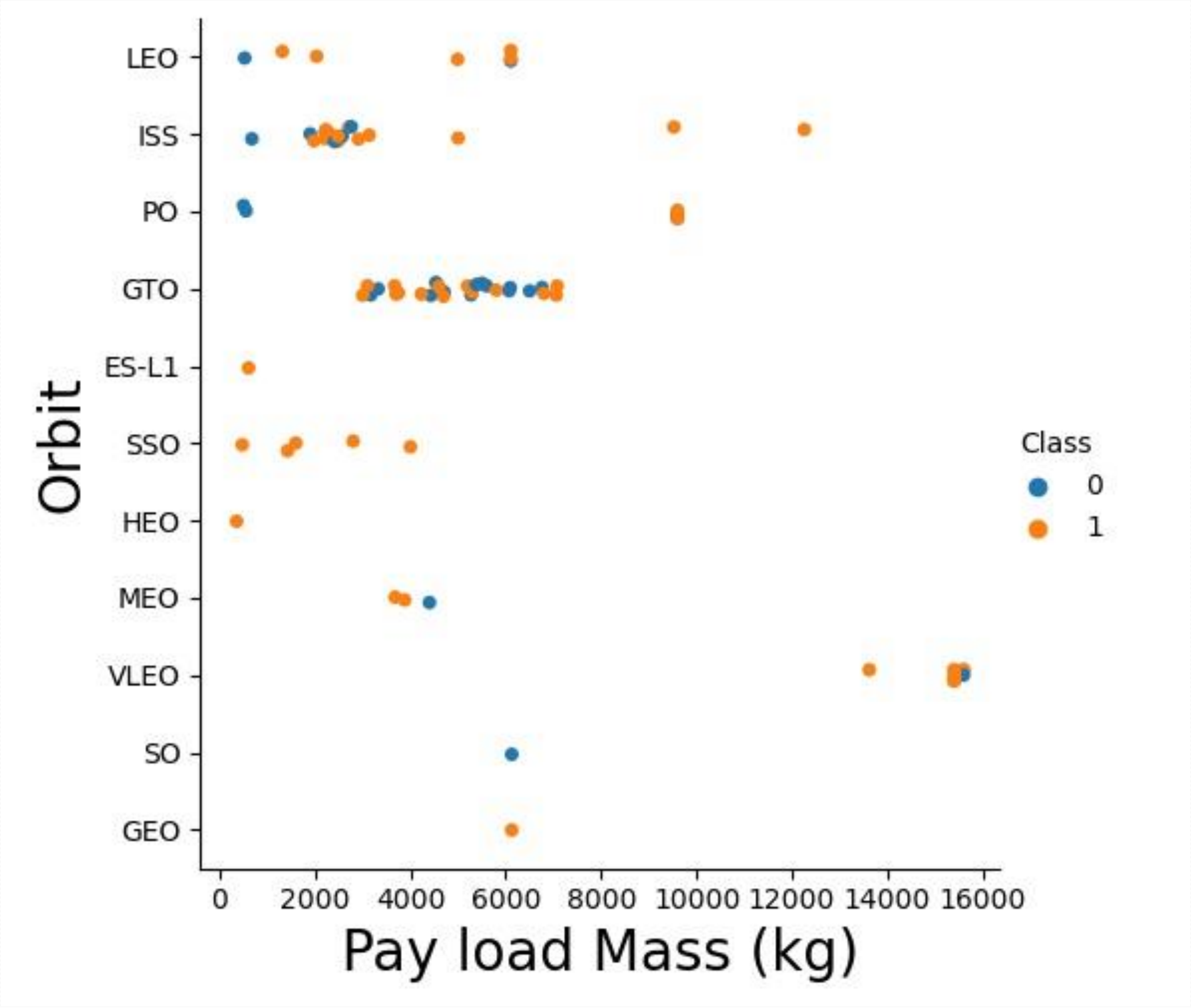


EDA WITH DATA VISUALIZATION

Payload vs. Orbit Type (Scatter Plot)

Explanation:

With heavy Payloads, the successful landing (Class = 1) are more for Orbit Polar, LEO and ISS.

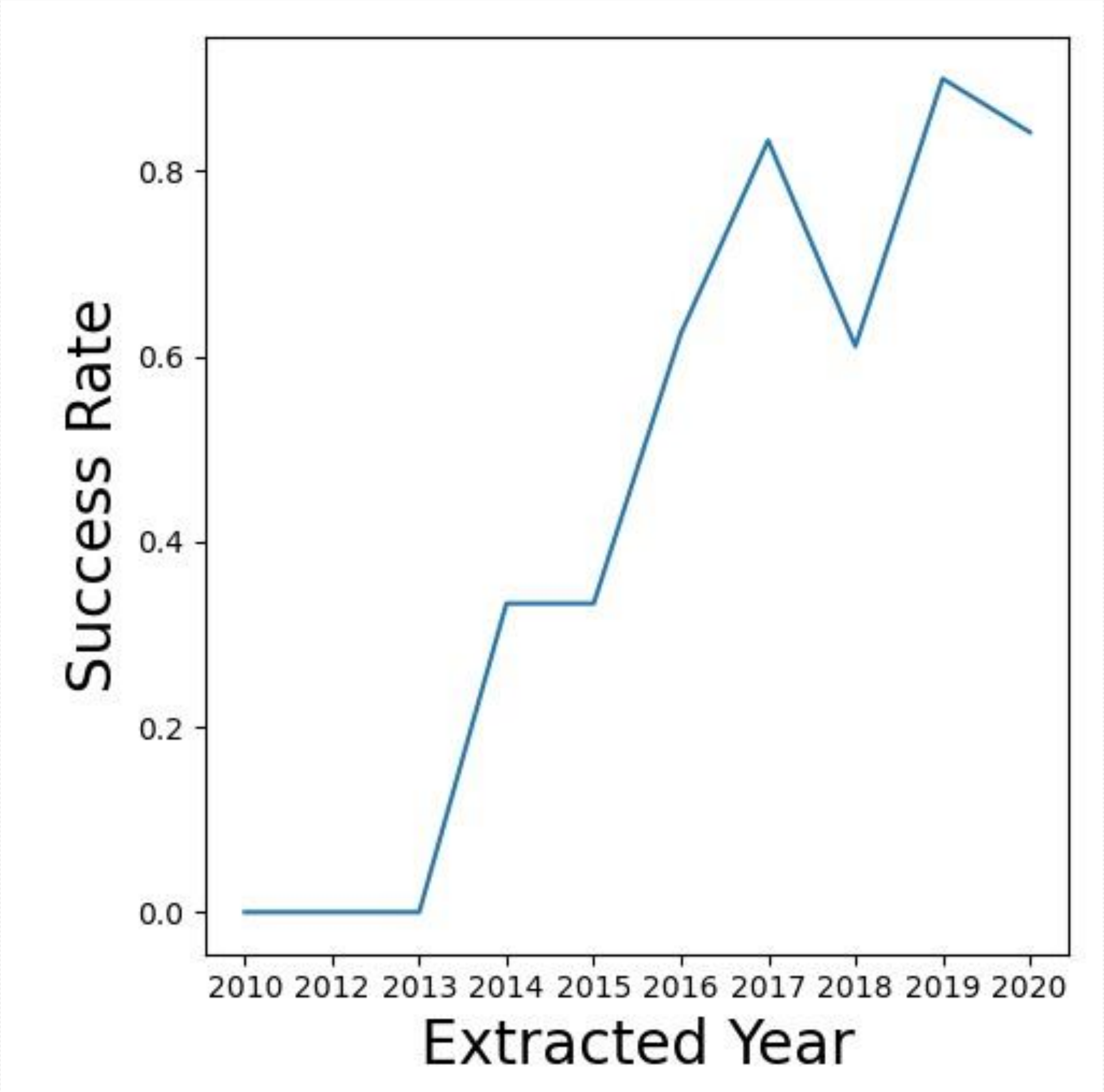


EDA WITH DATA VISUALIZATION

Launch Success Yearly Trend (Line Chart)

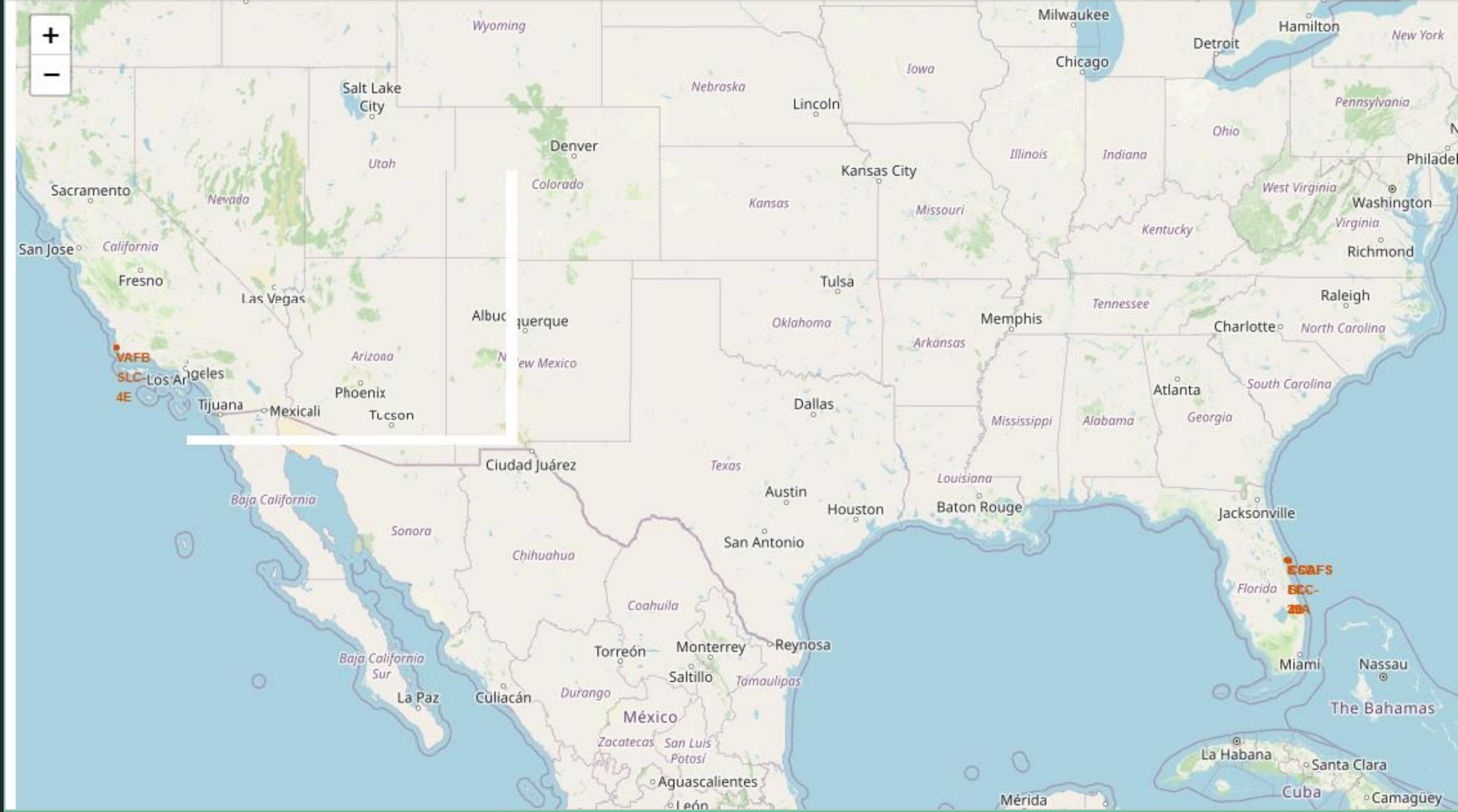
Explanation:

The Sucess Rate since 2013 kept increasing till 2020.



INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

Mark all launch sites' location markers on a global map



- **MARKER**

folium.Circle(coordinate, radius=1000, color='#000000', fill=True) .add_child(folium.Popup(...))

- **CIRCLE**

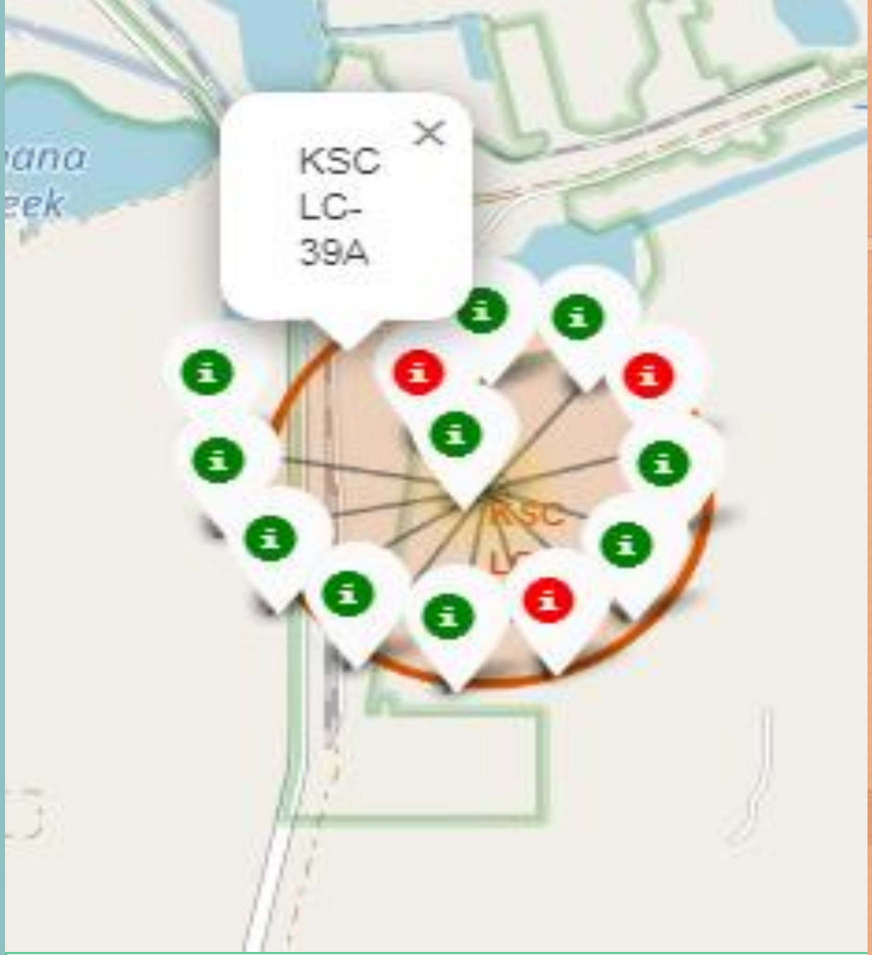
folium.map.Marker(coordinate, icon=DivIcon(icon_size=(20,20),icon_anchor=(0,0), html='<div style="font-size: 12; color:#d35400;"><b>%s</b></div>' % 'label',))

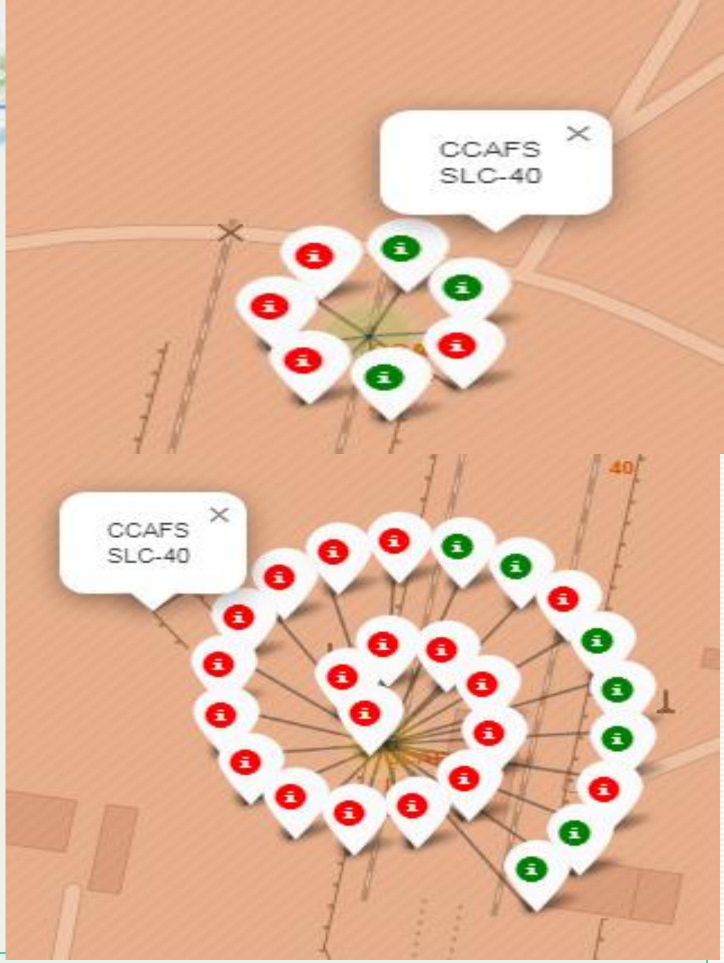
Explanation:

All launch sites are proximity to the Equator line.
All launch sites are very close proximity to the coast.

INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

Mark color-labeled launch outcomes (success / failed launches for each site) on the map





Explanation:

KSC LC 39A launch sites have relatively high success rates (Green Marker = Successful).

- **MARKERCLUSTER**

marker_cluster = MarkerCluster()

- **MARKER**

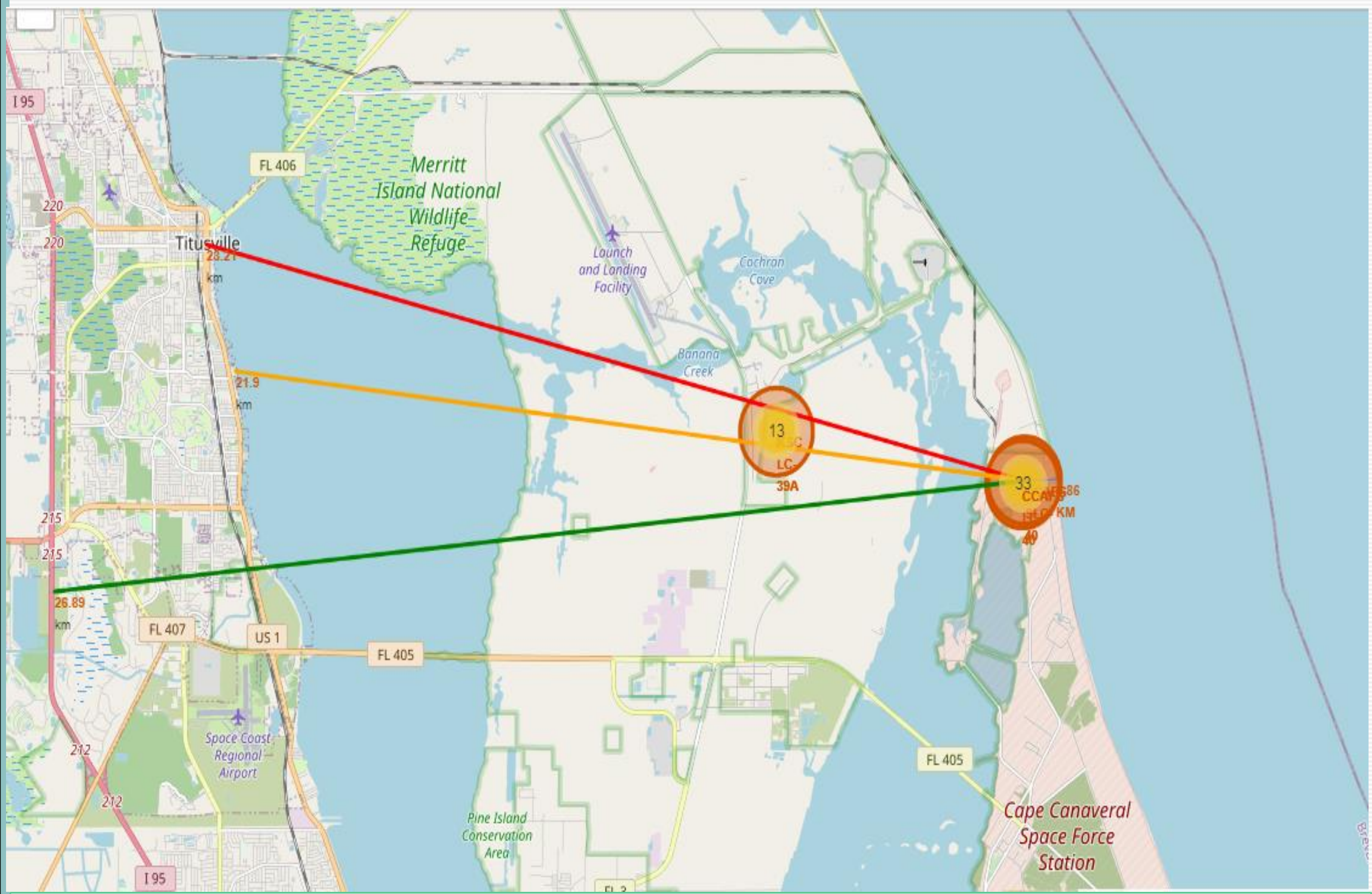
folium.Circle(coordinate, radius=1000, color='#000000', fill=True) .add_child(folium.Popup(...))

INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

Mark selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

Explanation:

For KSC LC 39A launch sites, in close proximity to railways, highways and coastline.



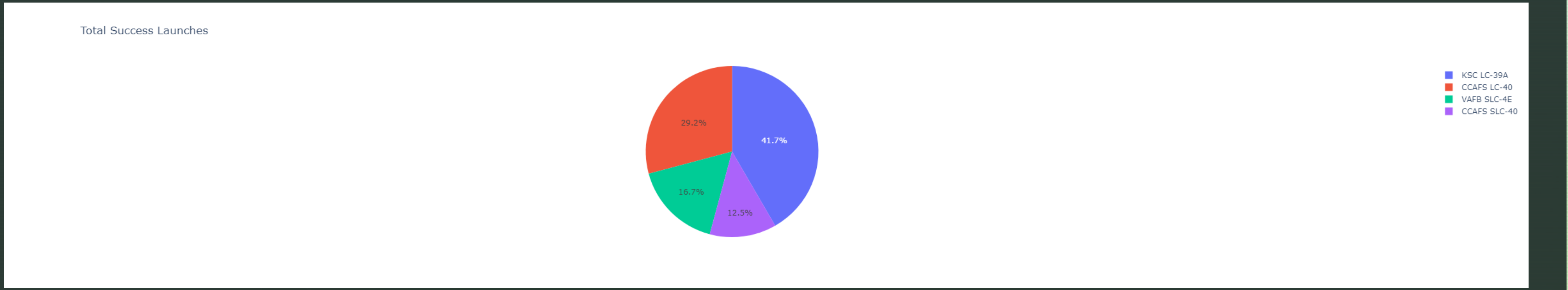
I 95
FL 406
Merritt Island National Wildlife Refuge
Titusville
28.21 km
Launch and Landing Facility
Cochran Cove
Banana Creek
21.9 km
13
LC 39A
33
CCAFS
KM
40
215
26.89 km
FL 407
US 1
FL 405
Space Coast Regional Airport
212
Pine Island Conservation Area
Cape Canaveral Space Force Station

- **MOUSEPOSITION / DISTANCE COASTLINE & RAILWAY & HIGHWAY**

distance_coastline = calculate_distance(launch_site_lat, launch_site_lon, coastline_lat, coastline_lon)

- **POLYLINE**

lines=folium.PolyLine(locations=coordinates, weight=1)

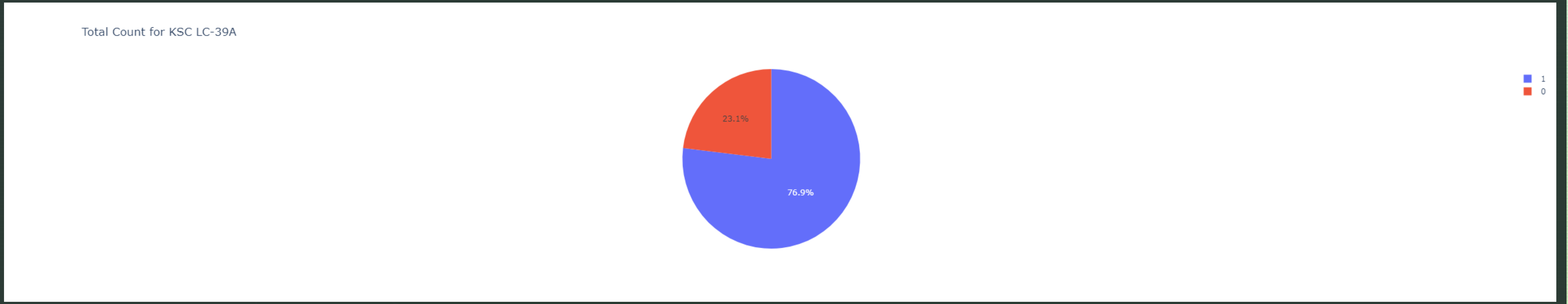


INTERACTIVE DASHBOARD WITH PLOTLY DASH

Launch Success Count for All Sites in Pie Chart

Explanation:

KSC LC 39A has the most successful launches (41.7%) from All Sites.

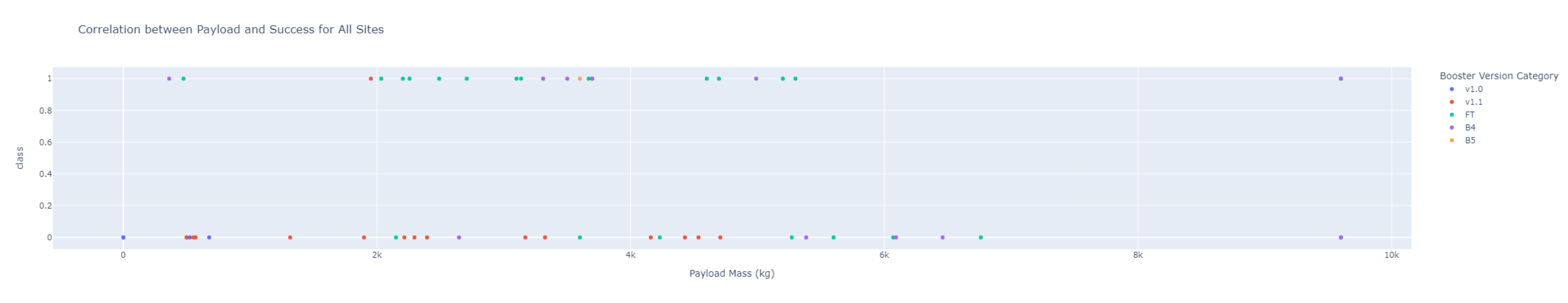


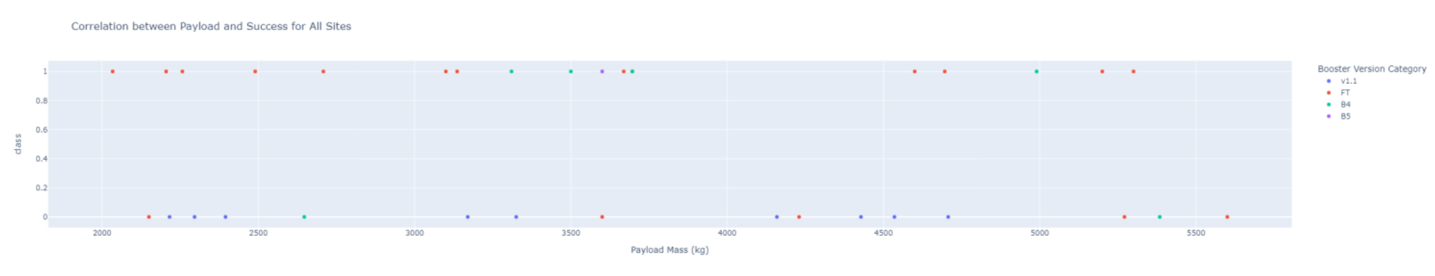
INTERACTIVE DASHBOARD WITH PLOTLY DASH

Pie Chart for the launch site with highest launch success ratio

Explanation:

KSC LC 39A has the highest launch site success ratio (76.9%) for Class 1, while 23.1% for Class 0.





INTERACTIVE DASHBOARD WITH PLOTLY DASH

Payload vs. Launch Outcome Scatter Plot for all sites,
with different payload selected in the Range Slider

Explanation:

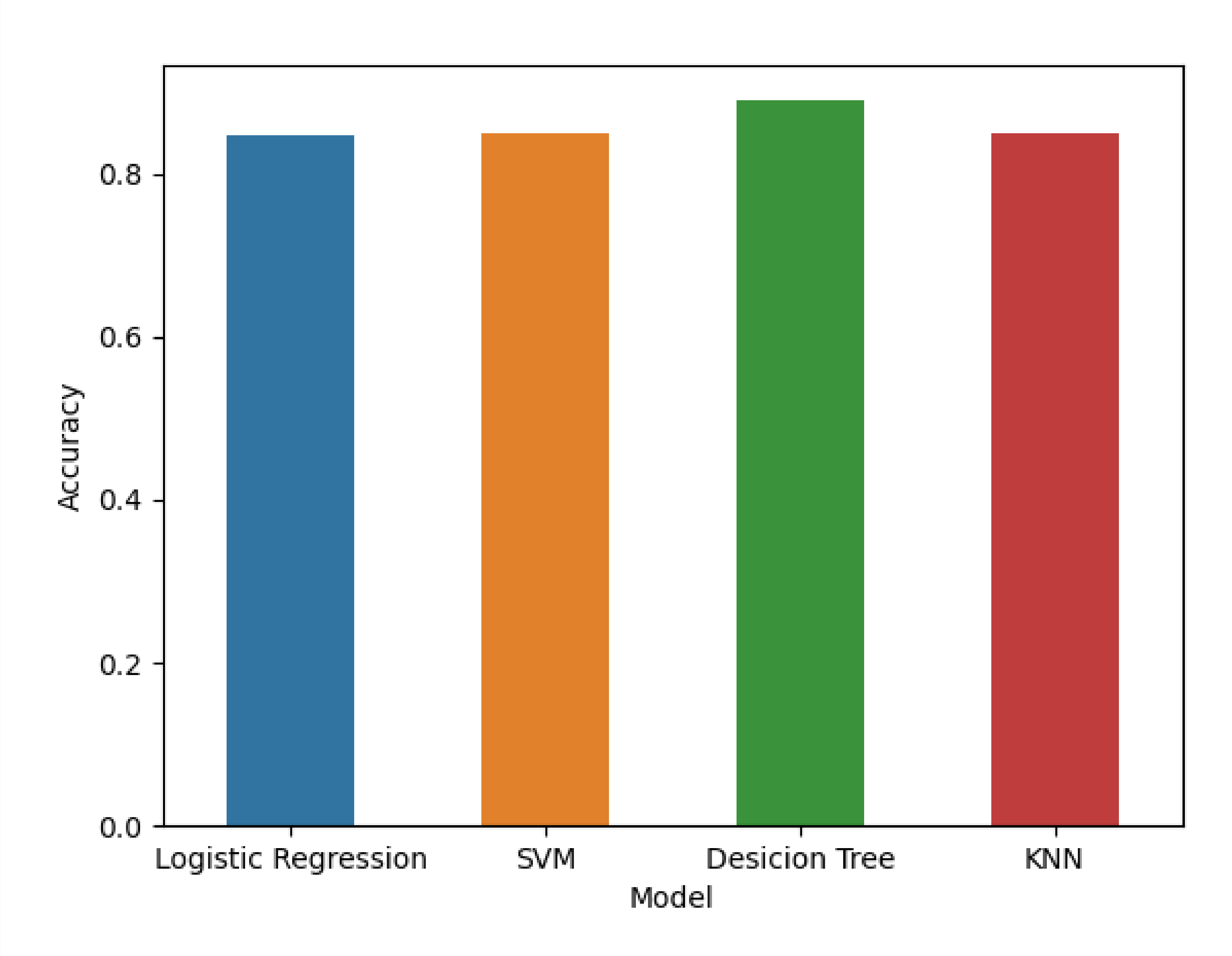
Payload range between 2000kg and 5500kg has highest success rate.

PREDICTION ANALYSIS

Classification Accuracy
(Bar Chart)

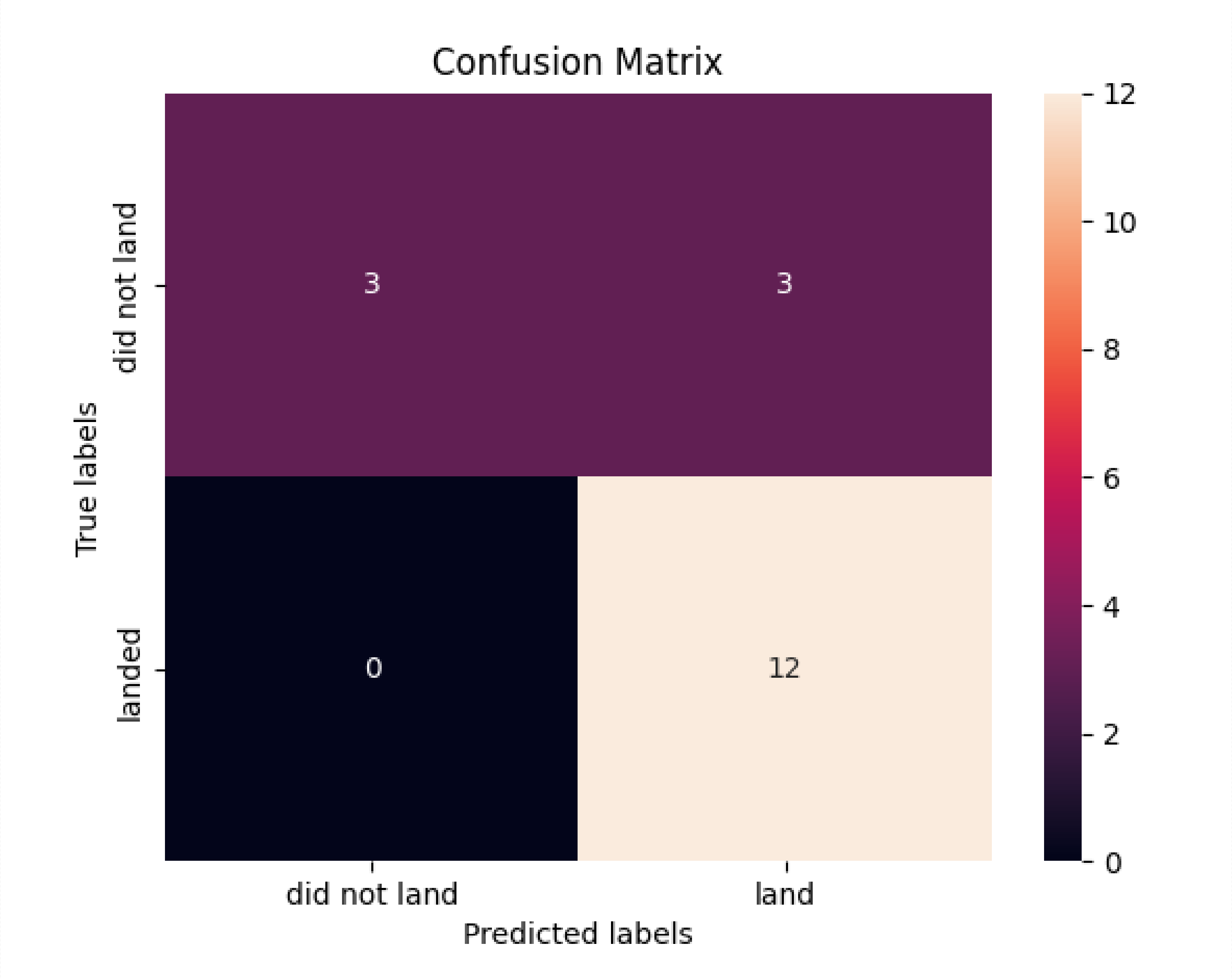
Explanation:

Decision Tree Model performs best. It has the highest classification accuracy.



PREDICTION ANALYSIS

Confusion Matrix



Explanation:

For Decision Tree Model, the major problem is False Positives.

CONCLUSION

KSC LC 39A has the most successful launches from All Sites.

KSC LC 39A launch site is in proximity to the Equator line and very close proximity to the coast.

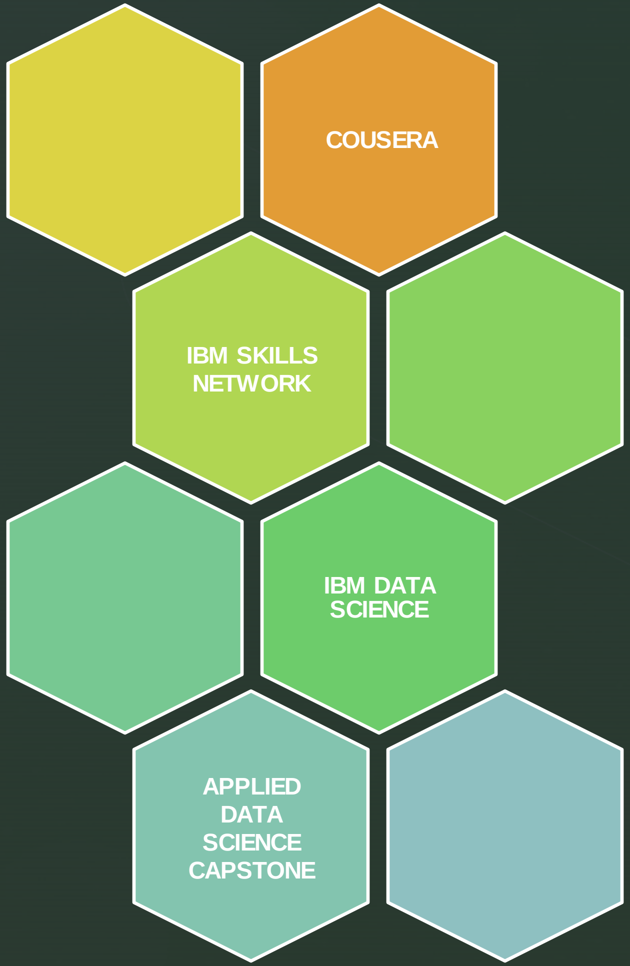
Payload range between 2000kg and 5500kg has highest success rate.

Orbits ES-L1, GEO, HEO, SSO have 100% Sucess Rate.

The Sucess Rate kept increasing over the years.

Decision Tree Model performs best for this dataset.

APPENDIX



COUSERA

IBM SKILLS NETWORK

IBM DATA SCIENCE

APPLIED DATA SCIENCE CAPSTONE