DATA SCIENCE CAPSTONE PROJECT

Calvin Rosinbaum

6/10/2025

OUTLINE

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

Executive Summary

Summary of methodologies

- Data Collection via API, Web Scraping
- Exploratory Data Analysis (EDA) with Data Visualization
- EDA with SQL
- Interactive Map with Folium
- Dashboards with Plotly Dash
- Predictive Analysis

Summary of all results

- Exploratory Data Analysis results
- Interactive maps and dashboard
- Predictive results

INTRODUCTION

Project background and context

The aim of this project is to predict if the Falcon 9 first stage will successfully land. SpaceX says on its website that the Falcon 9 rocket launch cost 62 million dollars. Other providers cost upward of 165 million dollars each. The difference is explained by the fact that SpaceX can reuse the first stage. By determining if the stage will land, we can determine the cost of a launch.

Problems you want to find answers

- ➢ What are the main characteristics of a successful or failed landing?
- ➢ What are the effects each relationship of the rocket variables has in relation to the success or failure of a landing?
- ➢ What are the conditions which will allow SpaceX to achieve the best landing success rate?

Section 1: Methodology

Executive Summary

- **Data collection methodology:**
    - SpaceX REST API
    - Web Scrapping
- **Perform data wrangling:**
    - Dropping unnecessary columns
    - OneHot Encoding for classification models
- **exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models:**
    - How to build, tune, evaluate classification models

Data Collection

Datasets are collected from Rest SpaceX API and web-scrapping Wikipedia

❖ The information obtained by the API are rocket, launches, payload information

SpaceX Rest API call → API returns JSON file → Make Dataframe from JSON → Clean Data and export it

The information obtained by the webs-crapping of Wikipedia are launches, landing, payload information

❖ URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Get HTML response from Wikipedia → Extract data with BeautifulSoup → Make Dataframe → Export Data

DATA COLLECTION - SPACEX API

1. Response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
```

2. Convert Response to JSON File

```
data = response.json()
data = pd.json_normalize(data)
```

3. Transform data

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
getBoosterVersion(data)
```

4. Create dictionary with data

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

5. Create Dataframe

```
data = pd.DataFrame.from_dict(launch_dict)
```

6. Filter Dataframe

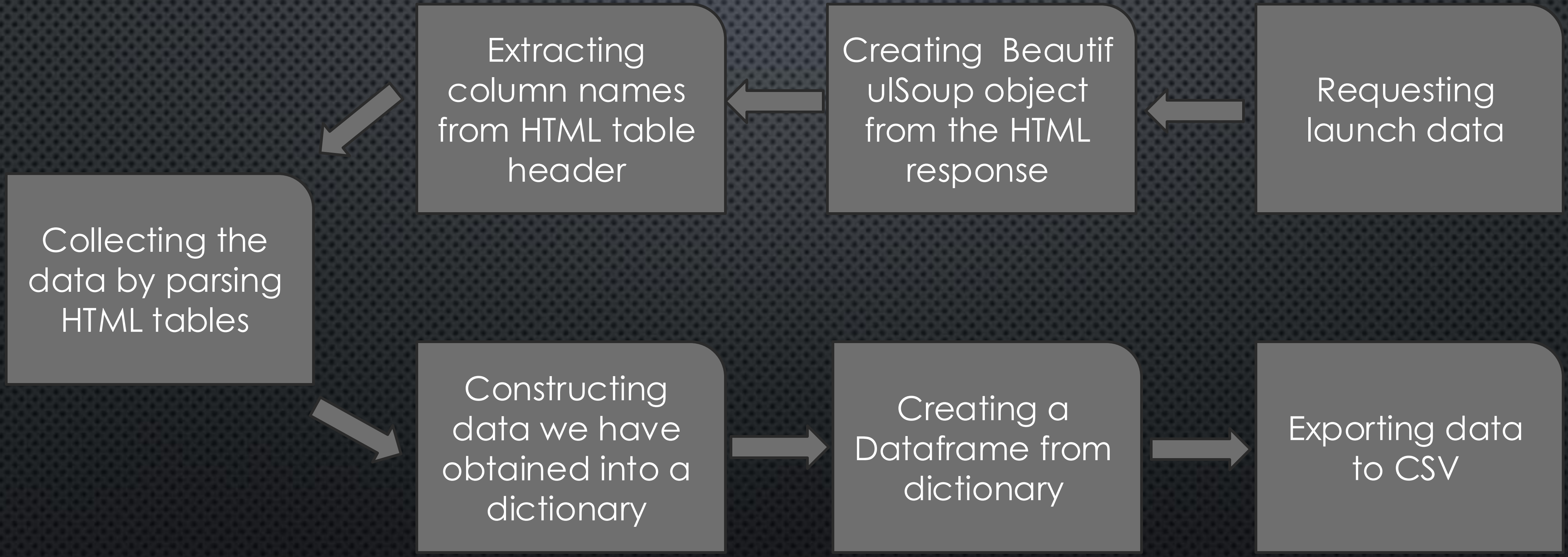
```
data_falcon9 = data[data['BoosterVersion']!='Falcon 1']
```

7. Export file

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Link

DATA COLLECTION – WEB SCRAPPING



DATA WRANGLING

- There are several cases where the booster did not land successfully
    - True Ocean, True RTLS, True ASDS mean mission success
    - False Ocean, False RTLS, False ASDS means the mission failed
- Transform string variables into categorical variables, where 1 means the mission has been successful, and 0 means the mission failed

[Link]

Exploratory Data Analysis and determining Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number of mission outcomes per orbit type and how many

Create a landing outcome label from Outcome column

Exporting data to CSV

EDA – DATA VISUALIZATION

<u>Charts Plotted:</u>

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs Orbit Type

Link

- Scatter plots of success rate yearly trends show variable relationships
  - If relationship exists => could be used in machine learning model
- Bar charts show comparisons among categories
  - The goal:
    - Show the relationship between specific categories being compared and measured
- Line charts show trends in data over time

EDA - SQL

WE PERFORMED SQL QUERIES TO GATHER AND UNDERSTAND DATA FROM DATASET

DISPLAY:

- NAMES OF THE UNIQUE LAUNCH SITES IN THE SPACE MISSION
- 5 RECORDS WHERE LAUNCH SITES BEGIN WITH THE STRING 'CCA'
- TOTAL PAYLOAD MASS CARRIED BY BOOSTERS LAUNCHED BY NASA (CRS)
- AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9 V1.1
- DATE OF FIRST SUCCESSFUL LANDING OUTCOME USING THE GROUND PAD
- LIST NAMES OF THE BOOSTERS WHICH HAVE SUCCESS PAYLOAD MASS GREATER THAN 4000 BUT LESS THAN 6000
- LIST TOTAL # OF SUCCESSFUL AND FAILED MISSION OUTCOMES
- LIST NAMES OF THE BOOSTER_VERSIONS WHICH HAVE CARRIED MAXIMUM PAYLOAD
- LIST RECORDS WHICH WILL DISPLAY THE MONTH NAMES, LANDING_OUUTCOMES IN DRONE SHIP, BOOSTER VERSIONS, LAUNCH_SITE FOR THE MONTHS IN YEAR 2015
- RANK THE COUNT OF SUCCESSFUL LANDIING_OUTCOMES BETWEEN THE DATE 06-04-2010 - 03-20-2017 IN DESCENDING ORDER

Link

INTERACTIVE MAP BUILDING WITH FOLIUM

Markers of all Launch Sites using Latitude and Longitude:

- Added Marker with Circle, Popup and Text Labels of NASA Johnson Space Center
- Added Markers with Circle, Popup and Text Labels of all Launch Sites

Colored Markers of the launch outcomes for each Launch Site:

- Markers of success (Green) and failed (Red) launches using Marker Cluster to identify high success rates launch areas

Distances between a Launch Site to its proximities:

- Lines to show distances between the Launch Site and proximities to railways, highway, coastline and cities

Link

Dashboard Building – Plotly Dash

- Dropdown has pie chart, range slider and scatter plot variables
  - allows users to choose one or all launch sites
- Pie chart shows the total success and total failure for launch site chosen
- Scatter chart shows the relationship between two variables
  - For Example: Success vs Payload Mass

Link

PREDICTIVE ANALYSIS (CLASSIFICATION)

Data preparation:

- Load dataset
- Normalize data
- Split data into training and test sets

Model preparation:

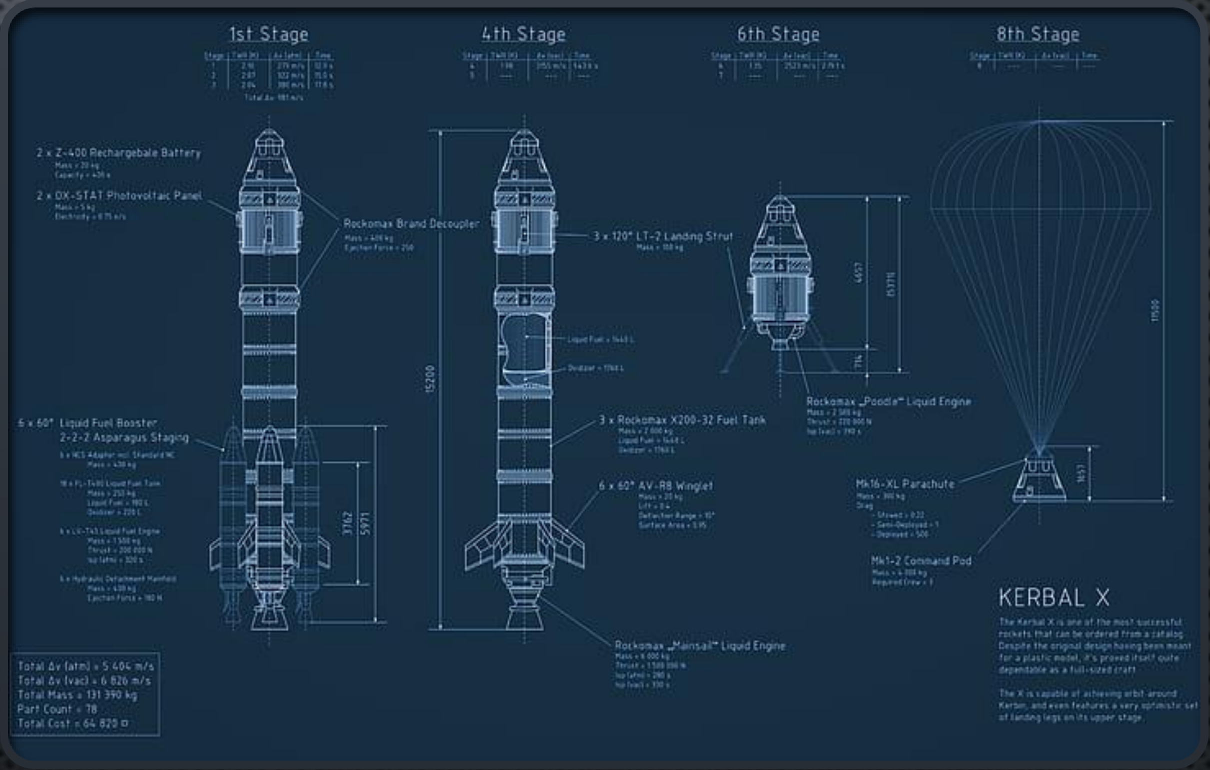
- Selection of machine learning algorithms
- Set parameters each algorithm to GridSearchCV
- Training GridSearchModel models via training dataset

Model evaluation:

- Get best hyperparameters for each model
- Compute accuracy for models with test dataset
- Plot Confusion Matrix

Model comparison:

- Comparison of models according to accuracy
  - o Model with the best accuracy will be chosen

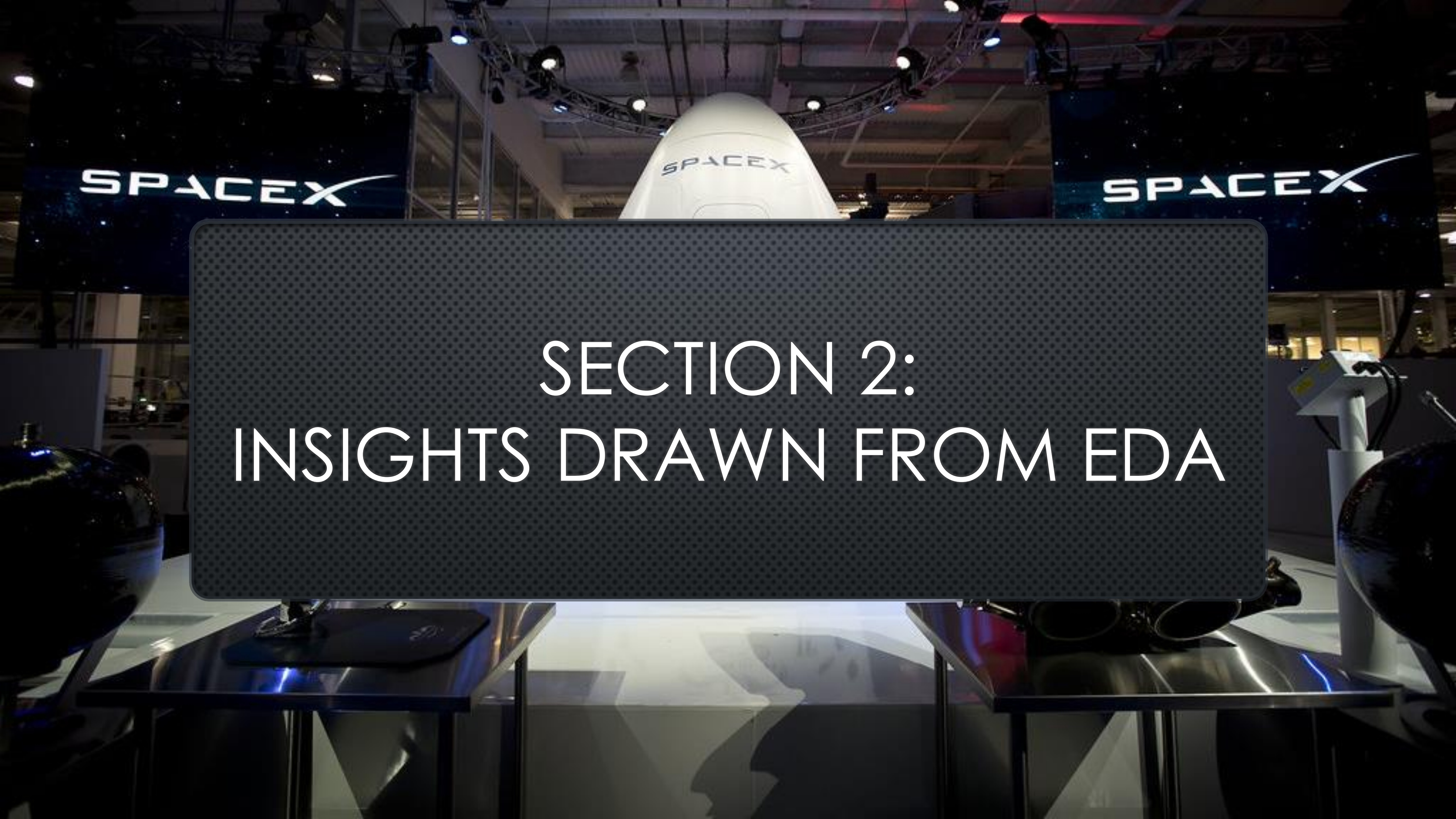


Link

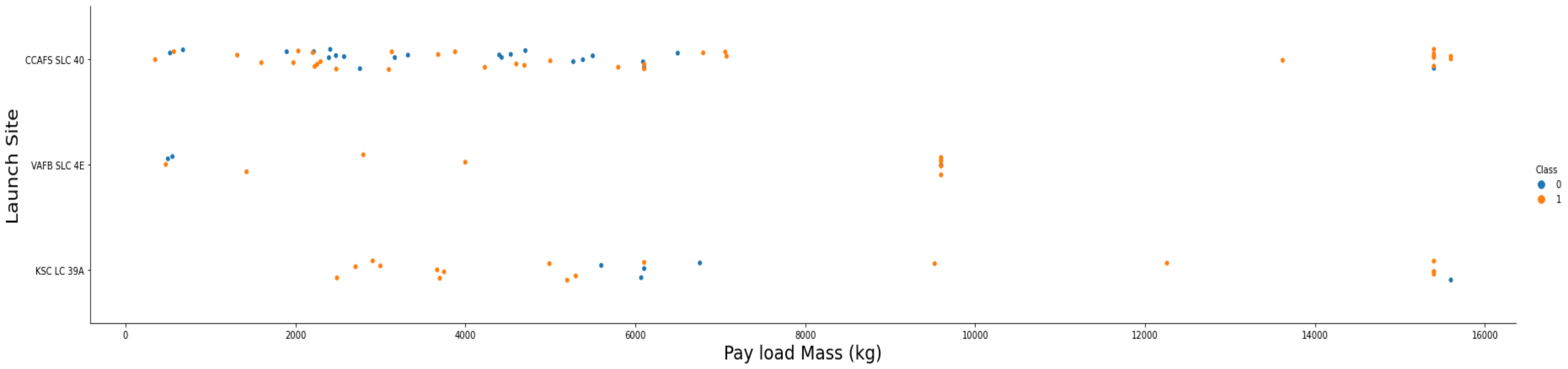
Results

- Exploratory data analysis
- Interactive analytics
- Predictive analysis

SECTION 2: INSIGHTS DRAWN FROM EDA



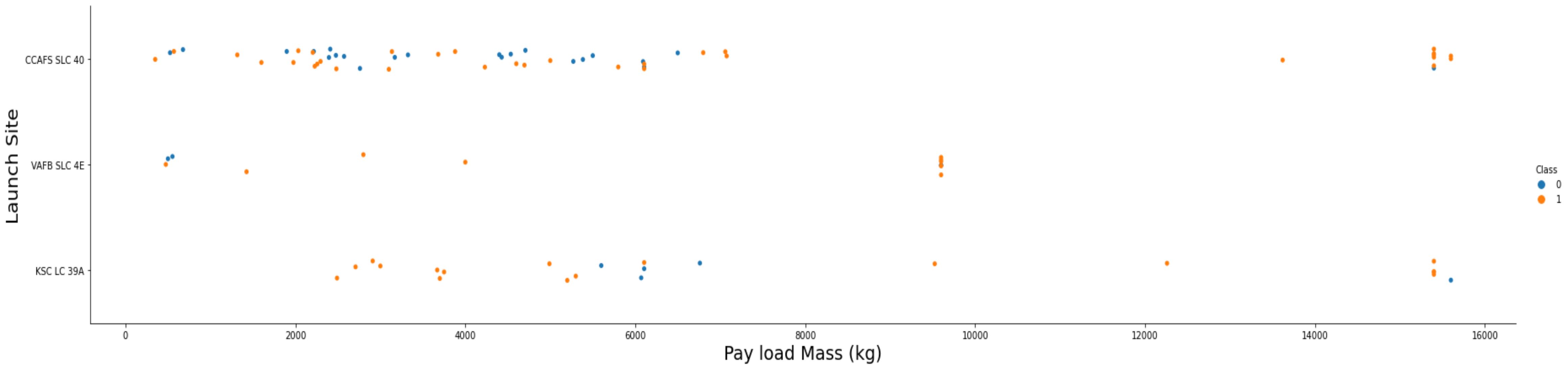
FLIGHT NUMBER VS. LAUNCH SITE



<u>EXPLANATION:</u>

- ✓ EARLIEST FLIGHTS FAILED WHILE THE LATEST FLIGHTS SUCCEEDED
- ✓ CCAFS SLC 40 LAUNCH SITE HAS ABOUT HALF OF ALL LAUNCHES
- ✓ VAFB SLC 4E AND KSC LC 39A HAVE HIGHER SUCCESS RATES
- ✓ ASSUMPTION: EACH NEW LAUNCH HAS A HIGHER RATE OF SUCCESS

PAYLOAD VS. LAUNCH SITE



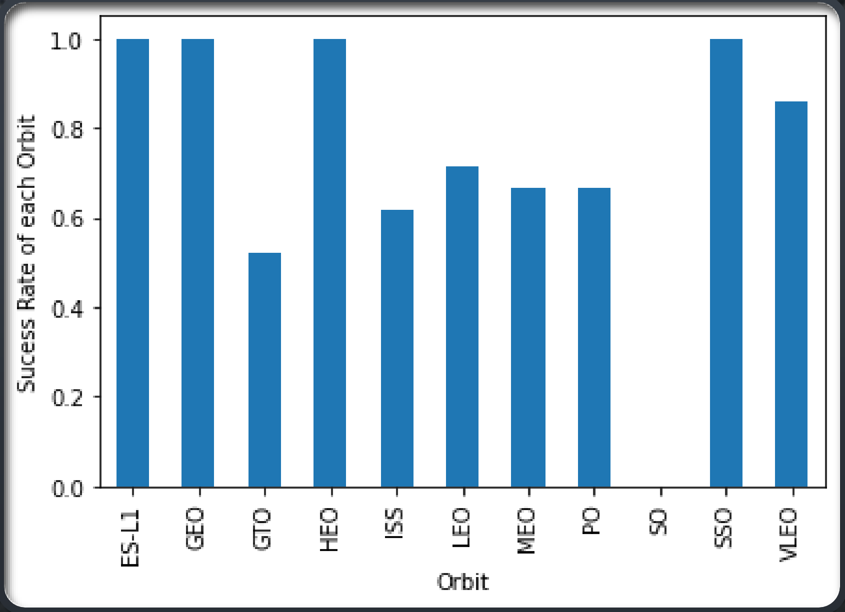
<u>Explanation:</u>

- ✓ For every launch site the higher the payload mass => higher the success rate
- ✓ Most launches with payload mass over 7000kg => success
- ✓ KSC LC 39A => 100% success rate for payload mass under 550kg

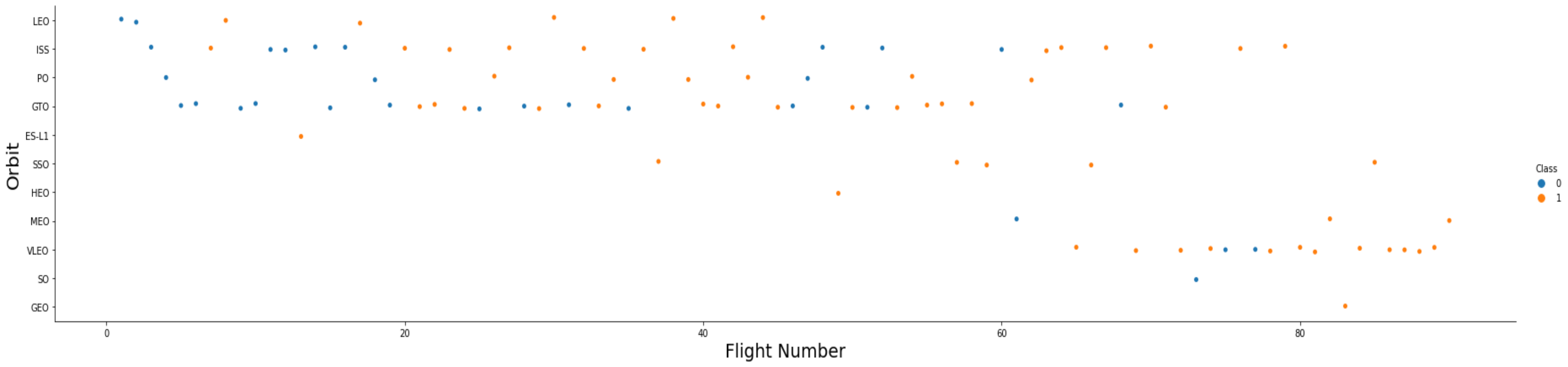
SUCCESS RATE VS. ORBIT TYPE

EXPLANATION:

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



FLIGHT NUMBER VS. ORBIT TYPE

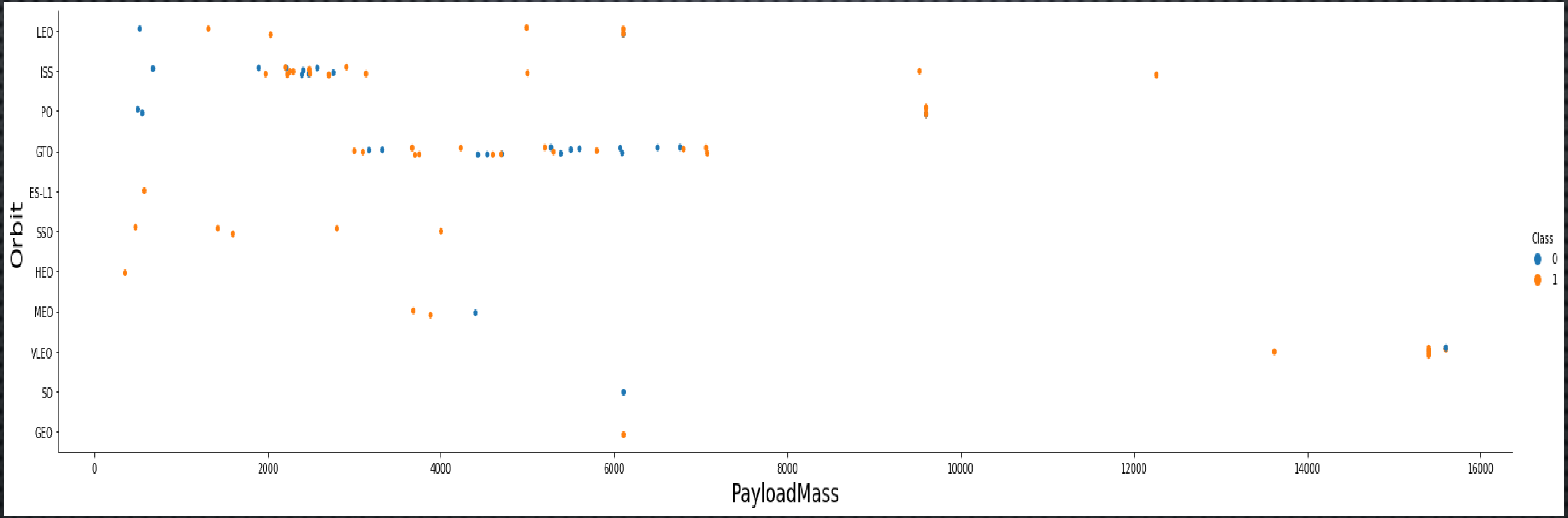


Notice success rate increases with the number of flights for the LEO orbit.

For some orbits like GTO, there is no relation between the success rate and the number of flights.

Can suppose that high success rates of some orbits like SSO or HEO are due to the knowledge of former launches for other orbits.

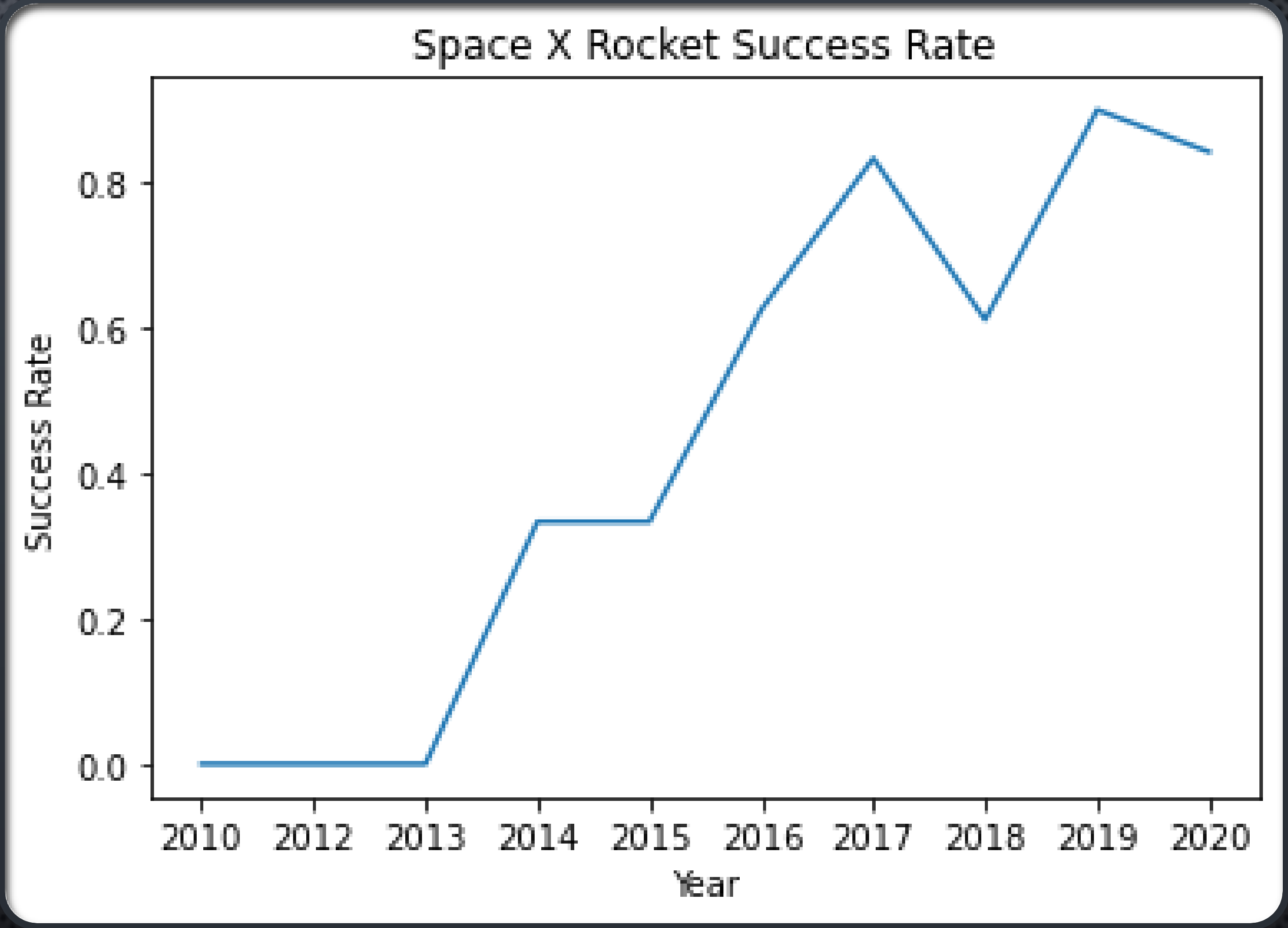
PAYLOAD MASS VS. ORBIT TYPE



- ✓ Heavy payloads have a negative relationship on GTO orbits, and are positive on GTO and Polar LEO (ISS) orbits

LAUNCH SUCCESS RATE – YEARLY TREND

- ✓ Success rate since 2013 increased until 2020



EDA WITH SQL – LAUNCH SITE NAMES

Code:

```
SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
```

Results:

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

- ❖ Use of the DISTINCT gets rid of duplicates

SITE NAMES WITH 'CCA' IN THE NAME

- The "WHERE" clause followed by "LIKE", filters launch sites that contain the substring 'CCA', and "LIMIT 5" shows only 5 records

Code:

```
SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

Results:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

TOTAL PAYLOAD MASS

- ✓ THIS QUERY RETURNS THE SUM OF ALL PAYLOAD MASSES "WHERE" THE CUSTOMER IS NASA (CRS)

Code:

```
SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
```

Results:

SUM("PAYLOAD_MASS__KG_")
45596

AVERAGE PAYLOAD MASS BY F9 V1.1

- ❖ QUERY RETURNS AVERAGE OF ALL PAYLOAD MASSES WHERE THE BOOSTER VERSION CONTAINS THE SUBSTRING F9 v1.1

Code:

```
SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

Results:

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

FIRST SUCCESSFUL LANDING DATE

Code:

```
SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'
```

Results:

MIN("DATE")
01-05-2017

- Query selects the oldest successful landing
- The "WHERE" clause filters datasets to keep records of landing success.
  - The "MIN" function selects the record with the oldest date

PAYLOAD MASS BETWEEN 4000 AND 6000

Code:

```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

Results:

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

- Query produces the booster versions where landing = success and payload mass = between 4000kg-6000kg
- "WHERE" and "AND" clauses filter data

BOOSTERS THAT CARRIED MAX PAYLOAD

Code:

```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

- Subquery to filter data by returning only the heaviest payload mass with "MAX" function
- Main query uses subquery results and returns unique boosters using "SELECT" and "DISTINCT" with the heaviest payload

Results:

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

Code:

```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

- ❖ List the failed landing outcomes for the drone ship, booster versions, and launch site names for the months in the year 2015

Results:

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

RANK LANDING OUTCOMES BETWEEN 06-04-2010 AND 03-20-2017

Code:

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

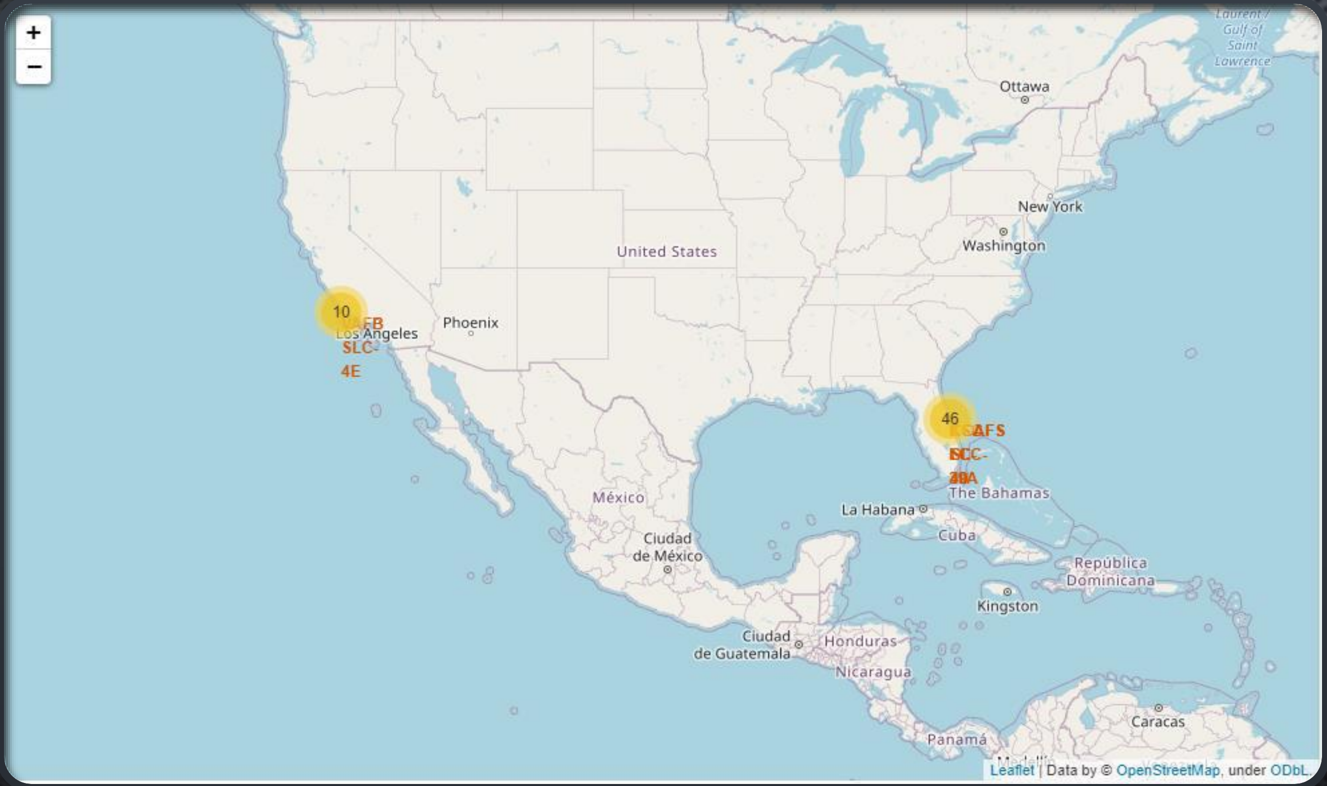
Results:

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

- Ranking the count of landing outcomes:
  - Failure or Success between the dates 06-04-2010 and 03-20-2017 in descending order

SECTION 4: LAUNCH SITE PROXIMITY ANALYSIS

FOLIUM MAP - GROUND STATIONS

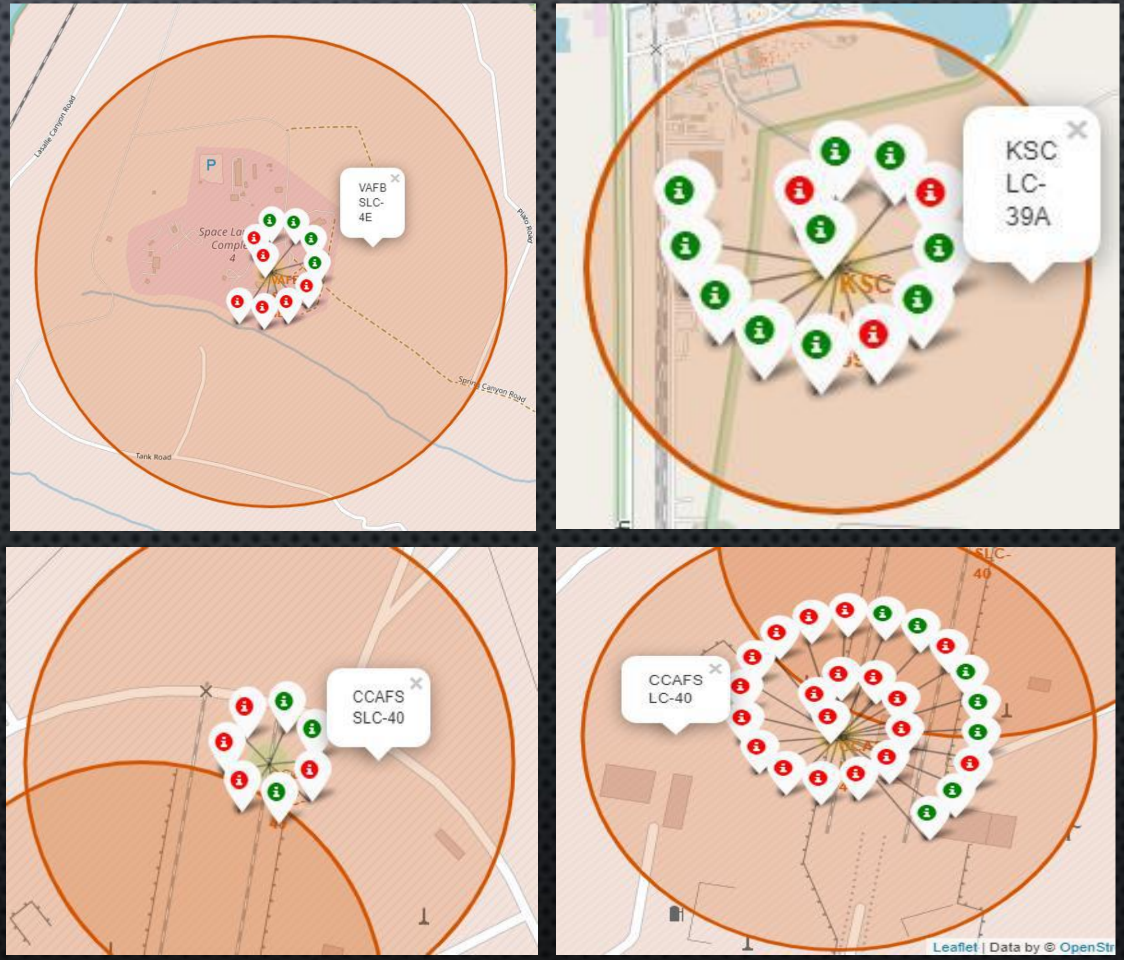


❖ SpaceX launch sites are located on the coasts of the United States

FOLIUM MAP - COLOR LABELED MARKERS

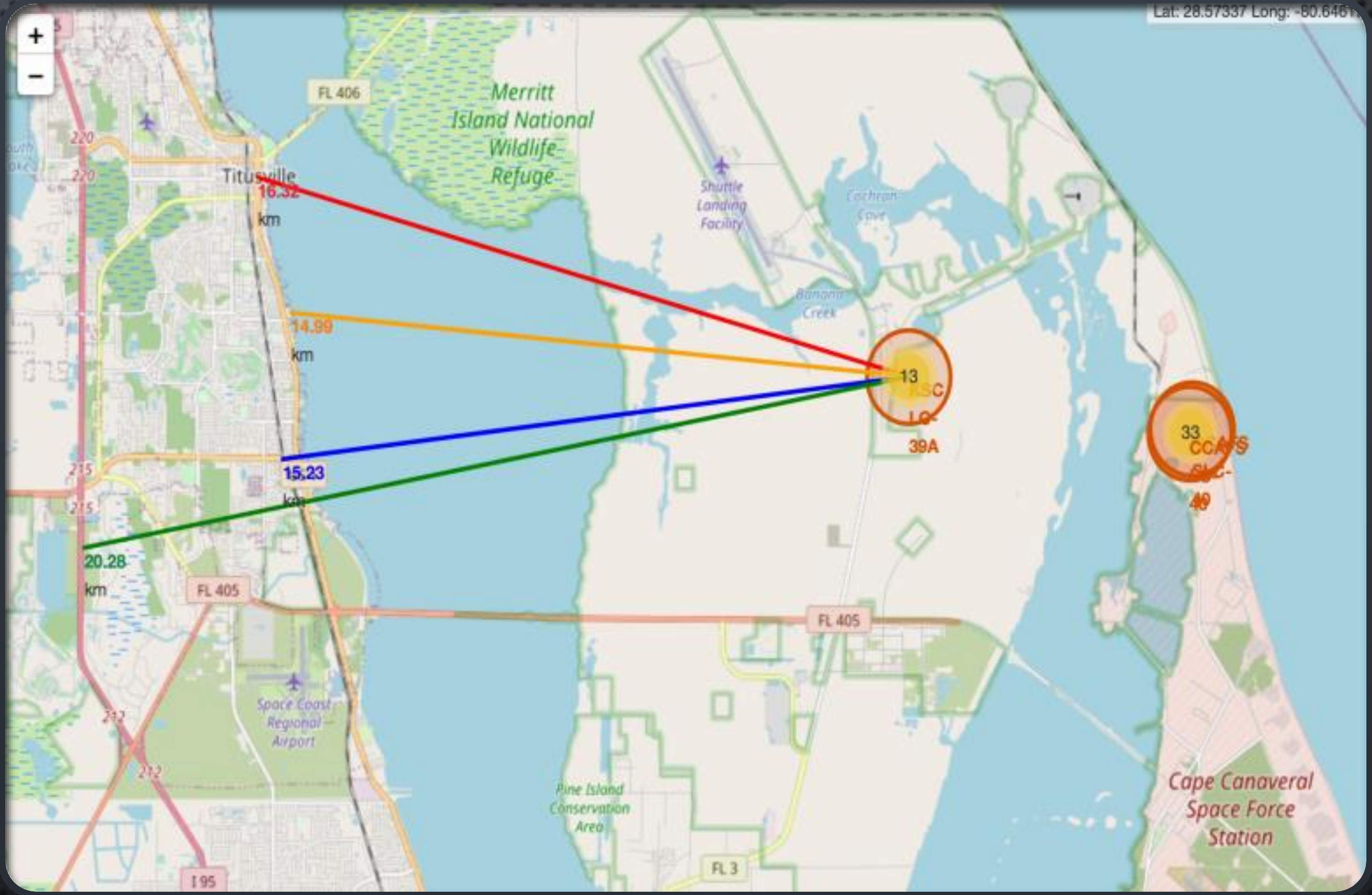
<u>EXPLANATION:</u>

- SHOULD BE ABLE TO EASILY IDENTIFY WHICH LAUNCH SITES HAVE RELATIVELY HIGH SUCCESS RATES
  - GREEN = SUCCESS
  - RED = FAILED
- LAUNCH SITE KSC LC-39A HAS A VERY HIGH SUCCESS RATE



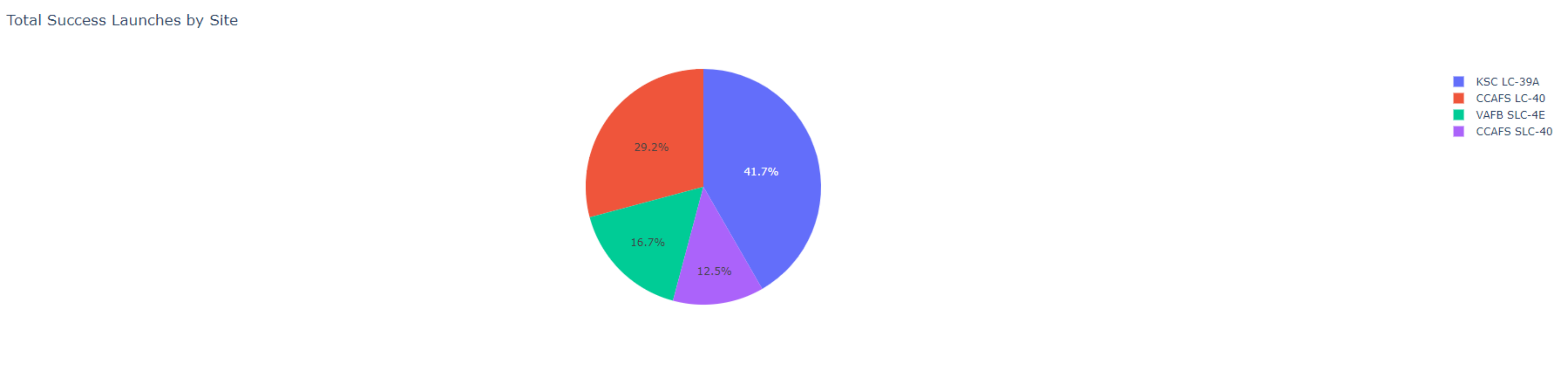
FOLIUM MAP - DISTANCES BETWEEN KSC LC-39A AND IT'S PROXIMITIES

- VIA VISUAL ANALYSIS OF THE LAUNCH SITE KSC LC-39A:
  A. RELATIVE CLOSE TO RAILWAY (15.23 KM)
  B. RELATIVE CLOSE TO HIGHWAY (20.28 KM)
  C. RELATIVE CLOSE TO COASTLINE (14.99 KM)
  D. LAUNCH SITE KSC LC-39A SEMI-CLOSE TO THE CITY TITUSVILLE (16.32 KM)



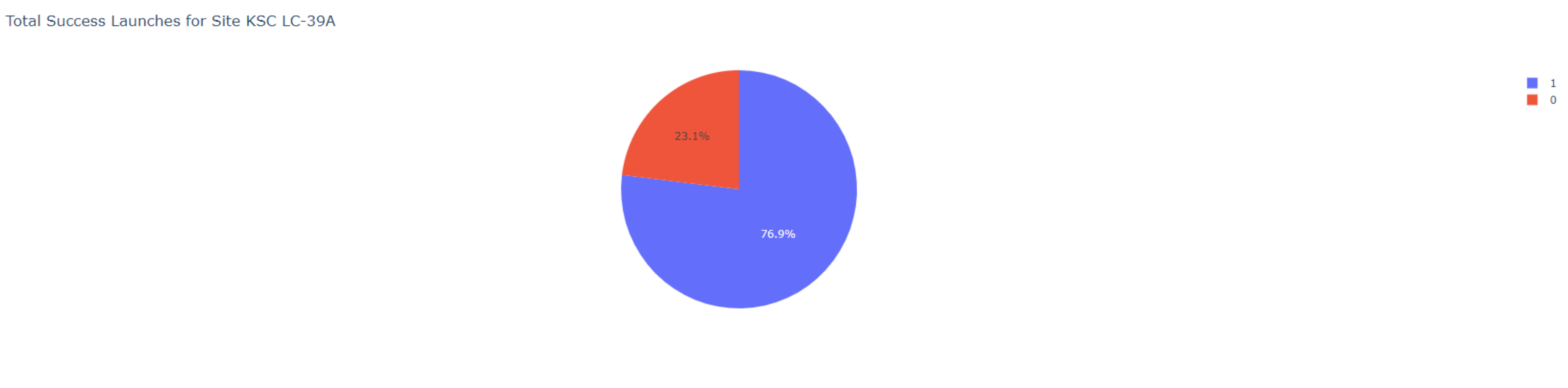
SECTION 5: BUILD A DASHBOARD WITH PLOTLY

DASHBOARD - TOTAL SUCCESS BY SITE



- WE SEE THAT KSC LC 39A =>
  BEST SUCCESS RATE OF LAUNCHES

DASHBOARD - TOTAL SUCCESS LAUNCHES FOR SITE KSC LC 39A



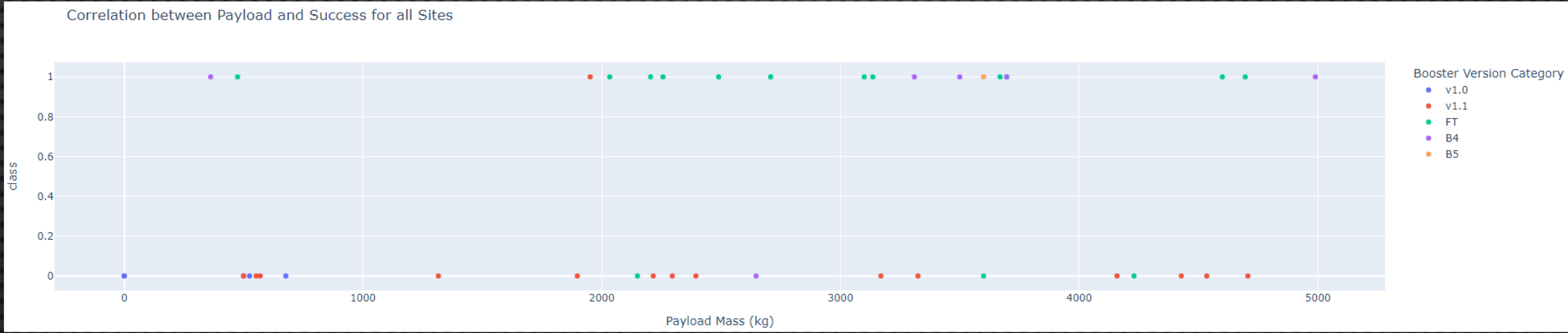
KSC LC 39A:

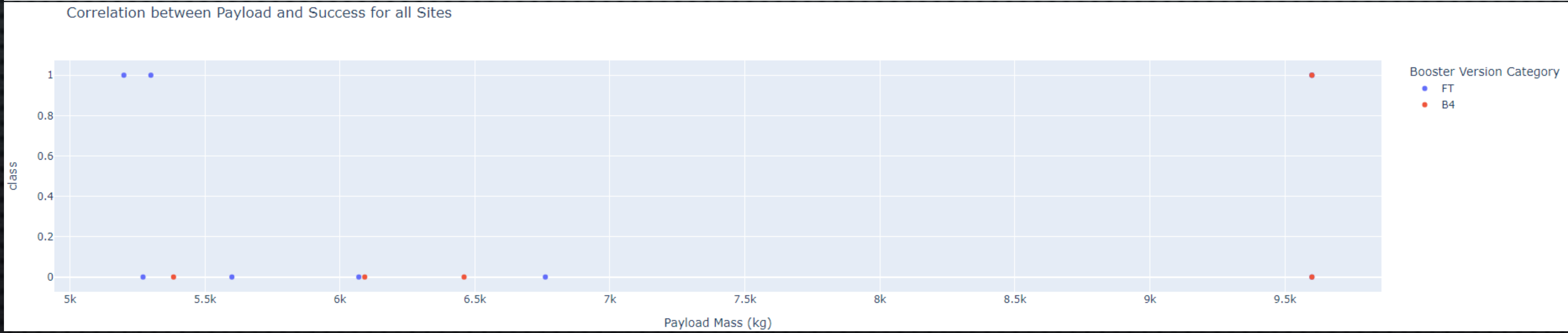
76.9% SUCCESS RATE

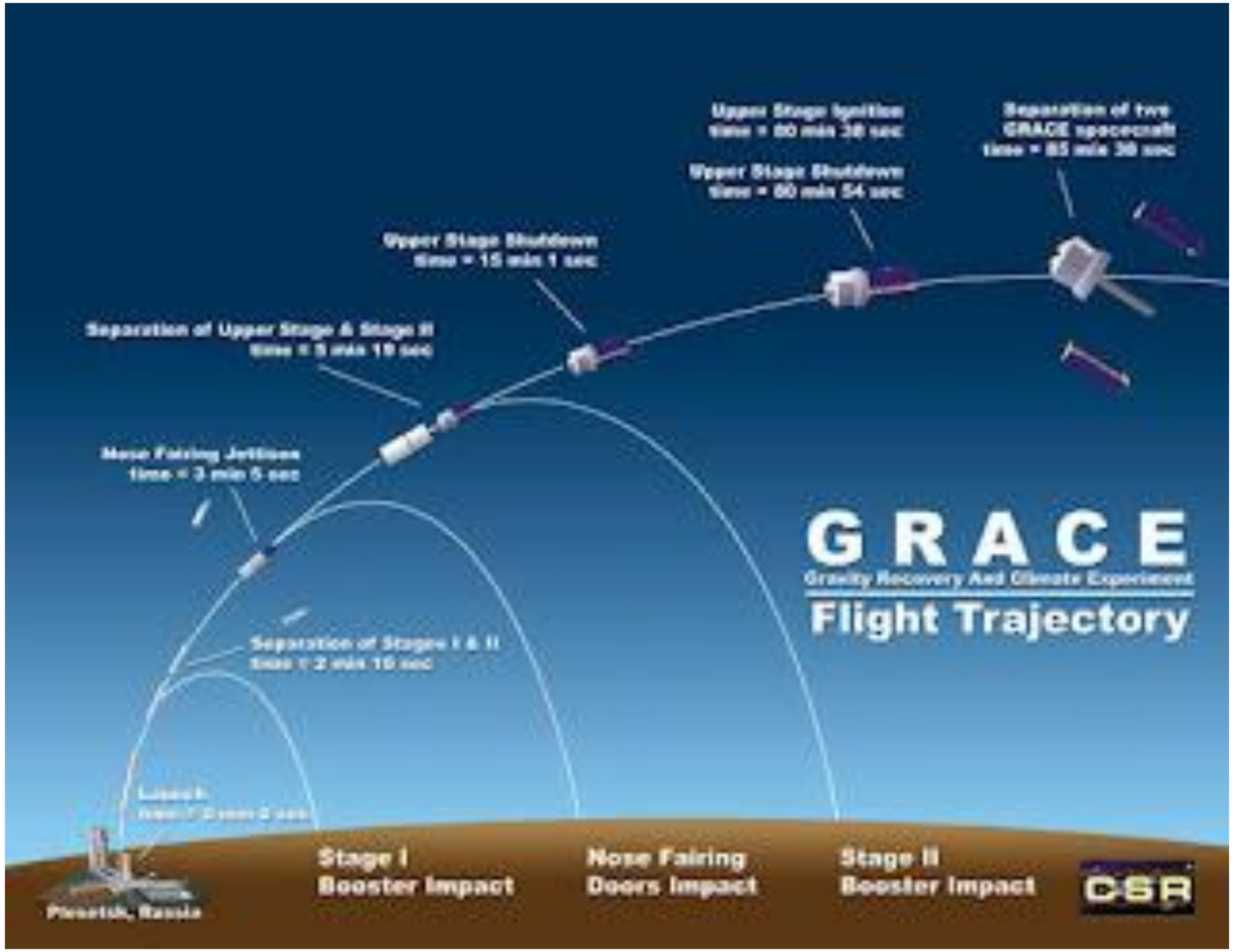
23.1% FAILURE RATE

DASHBOARD - PAYLOAD MASS VS OUTCOME FOR ALL SITES WITH DIFFERENT PAYLOAD MASS SELECTED

- ✓ Low payloads => better success rate than the heavy







SECTION 6 – PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY

Explanation:

- Via scores of the Test Set:
  - Cannot confirm which method performs best
- Same Test Set scores:
  - May be due to the small test size (18). Tested all methods based on the whole Dataset.
- Scores of whole Dataset:
  - Confirm the best model is the "Decision Tree Model". Model has higher scores and highest accuracy.

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

Scores and Accuracy of the Test Set

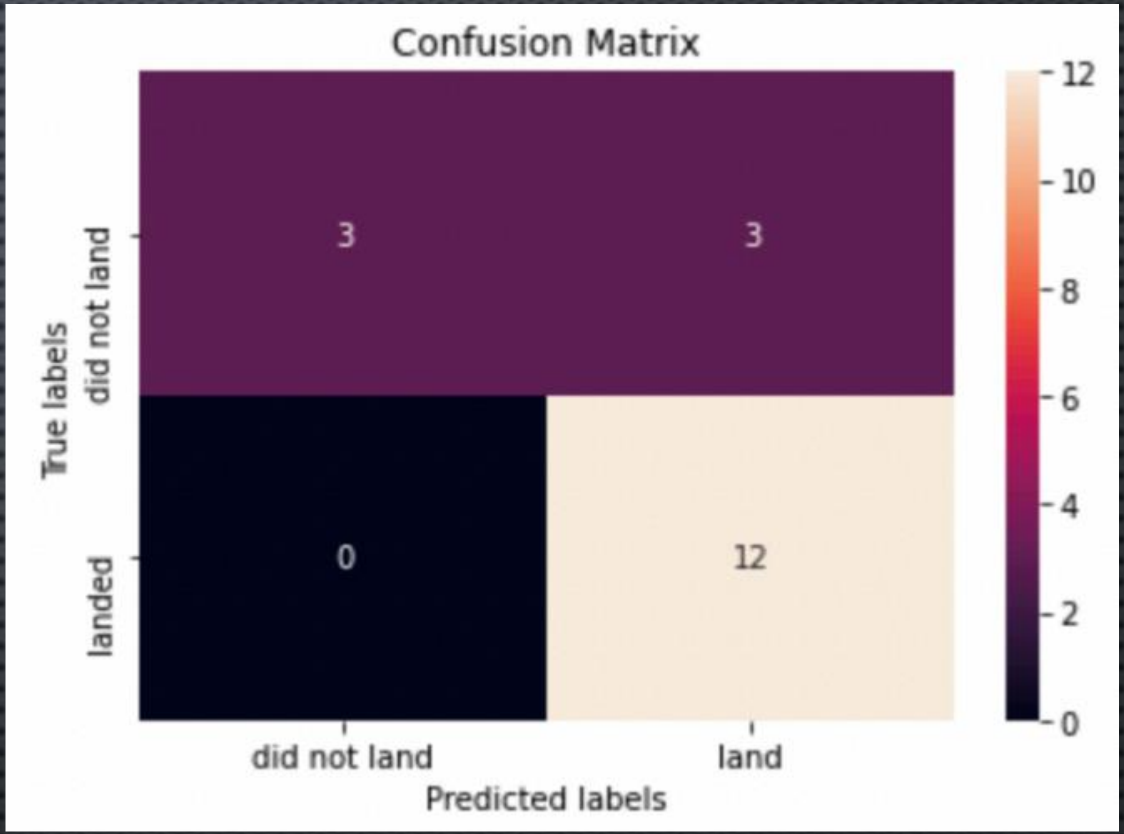
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

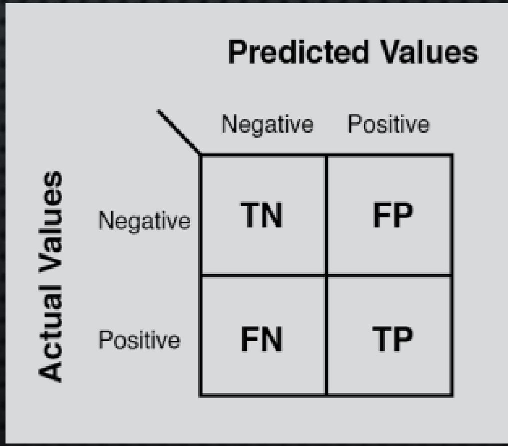
CONFUSION MATRIX

Findings:

- Logistic regression can distinguish between the different classes
- See a major problem is false positives

✓ **All matrices were the same**





Conclusion

- ✓ Success of a mission can be explained by several factors, such as the launch site, the orbit type, and especially the number of previous launches. We can assume that there has been a gain in knowledge between launches that allowed to go from failure to success.
- ✓ Orbits with the best success rates: GEO, HEO, SSO, ES L1
- ✓ Depending on the orbits, the payload mass can be an independent variable to take into account when trying to find success. Though some orbit flights require a lighter or heavier payload, generally low weighted payloads perform better than the heavy ones.
- ✓ With the current data, we cannot explain why some launch sites are better than others (KSC LC 39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data by utilizing other resources such as NASA launches in the past for more data to bring into our test and training sets.
- ✓ The "Decision Tree Algorithm" was chosen as the best model because of its high train accuracy (even if the test accuracy between all the models used is identical).

THANK YOU!!!