IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pablo Magallanes-Flores
Feb. 1st, 2023

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

Summary of methodologies

- Data Collection utilizing SpaceX API Web Scraping
- Data Wrangling to ease access and analysis
- Exploratory Data Analysis (EDA) with Visualization tools and SQL
- Creating an interactive map with Folium
- Using Plotly and Dash to create interactive dashboards for EDA
- Classification predictive analytics with Machine Learning

Summary of all results

- Functional Data was collected and organized
- EDA techniques assisted in choosing features for analytical methods
- Machine Learning techniques produced predictive models

Introduction

- Project background and context
    - SpaceX has become the leader in commercial space exploration, the cost to launch its Falcon 9 rocket is 62 million, where competitors average 165 million. They maintain a low cost by having designed their rocket to reuse its first stage and landing it back on Earth. The project aims to predict if the Falcon 9 will have a successful landing of it's first stage, therefore determining the cost of a launch. This information will assist alternate company in creating a competitive bid against SpaceX.
- Questions we want to answer
    - What are the features with the highest contribution in creating a successful or failed stage 1 landing.
    - Using these features, can we predict the outcome of future launches.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- SpaceX API to retrieve relevant data
- Web Scraping from Wikipedia using Beautiful soup

Perform data wrangling

- Taking advantage of pandas to remove unnecessary data and transform categorical data.

Perform exploratory data analysis (EDA) using visualization and SQL

- Utilizing libraries such as Seaborn and Matplotlib

Perform interactive visual analytics using Folium and Plotly Dash

- Allowed us to quickly and intuitively observe trends.

Perform predictive analysis using classification models

- Used ScikitLearn and Cross Validation to build and tune classification models.

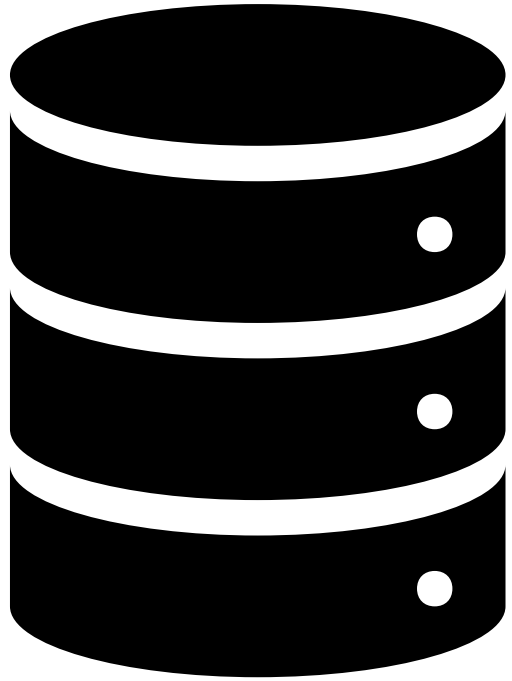
Data Collection

Open-Source Rest API from SpaceX in combination with python:

1. Request json data with GET request
2. Parse and filter the data, resulting in a DataFrame
3. Export the Data to a CSV

Wikipedia with the use of Web Scraping tools in python:

1. Locate and extract the desired data with Beautiful Soup
2. Parse and filter the data in a DataFrame
3. Export the Data to a CSV



Data Collection – SpaceX API

Get a JSON response from the API using a GET request

```
spacex_url="https://api.spacexdata.
response = requests.get(spacex_url)
```

Normalize the JSON into Pandas DataFrame

```
# Use json_normalize meethod to convert th
data = pd.json_normalize(response.json())
```

Use functions to convert data to Dictionary

```
launch_dict = {'FlightNumber': l
'Date': list(data['date']),
'BoosterVersion':BoosterVersio
'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}
```

Create a DataFrame with desired data from dictionary

```
# Create a data from launch_dict
launch_df = pd.DataFrame(launch_dict)
```

Filter and Clean the Data Frame

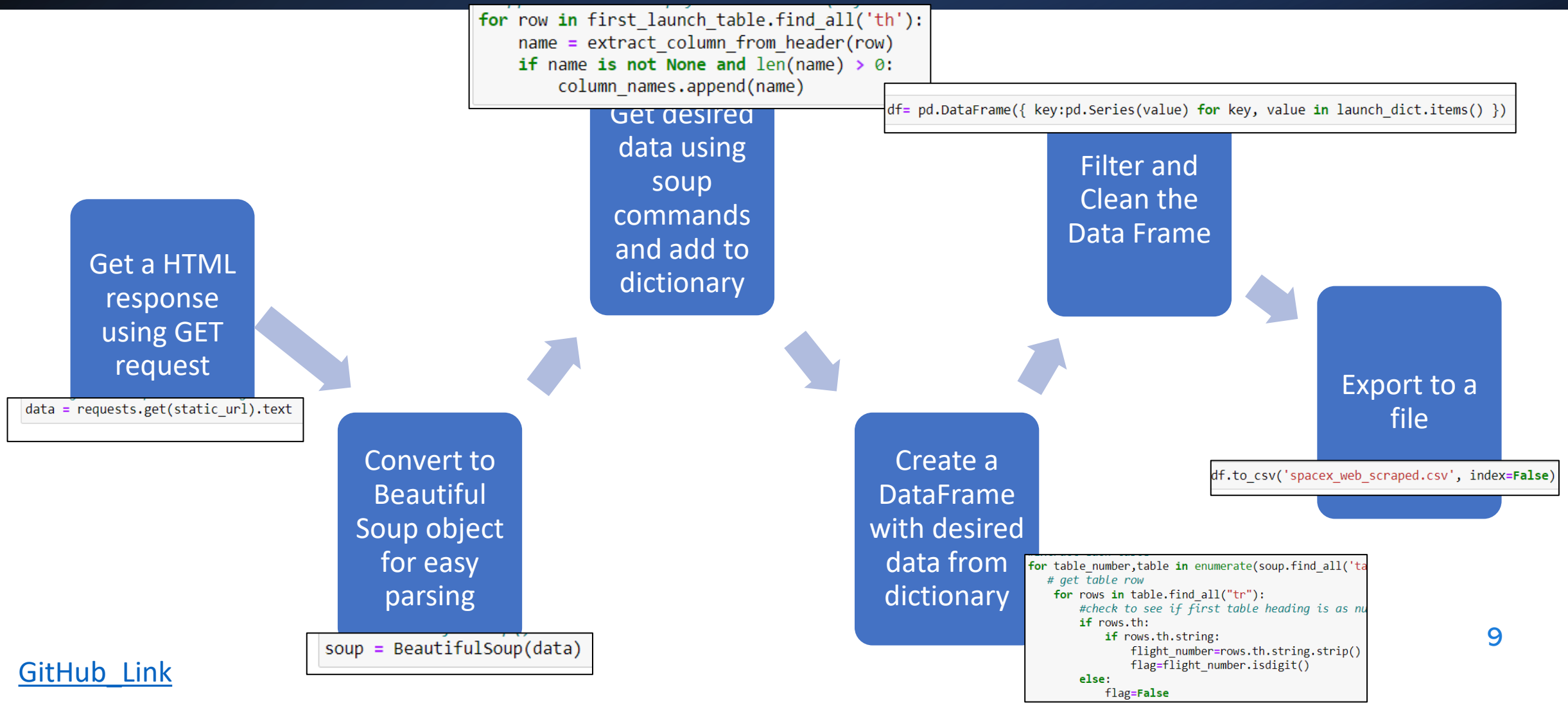
```
# Calculate the mean value of PayloadMass column
payload_mass_mean = data_falcon9["PayloadMass"].mean()
# Replace the np.nan values with its mean value
data_falcon9["PayloadMass"].fillna(value=payload_mass_mean, inplace=True)
data_falcon9.isnull().sum()
```

Export to a file

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

GitHub_Link

Data Collection – Scraping



GitHub_Link

Data Wrangling

Began by importing the data and exploring its data quality features such as null values and data types.

```
FlightNumber        int64
Date               object
BoosterVersion     object
PayloadMass       float64
Orbit              object
LaunchSite         object
Outcome            object
Flights             int64
GridFins             bool
Reused               bool
Legs                 bool
LandingPad         object
Block             float64
ReusedCount         int64
Serial             object
Longitude         float64
Latitude          float64
dtype: object
```

Used pandas to extract summary and categorical data for launch sites, orbit types, and successes/failures

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

```
GTO        27
ISS        21
VLEO       14
PO          9
LEO         7
SSO         5
MEO         3
ES-L1       1
HEO         1
SO          1
GEO         1
```

Appended desired categorical data to our Data Frame

```
landing_class = [0 if outcome in bad_outcomes else 1 for outcome in df['Outcome']]
df['Class']=landing_class
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0

Saved and exported the data to a CSV

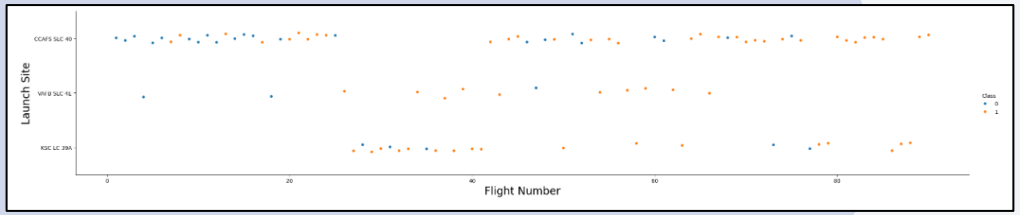
```
df.to_csv("dataset_part_2.csv", index=False)
```


GitHub_Link

EDA with Data Visualization

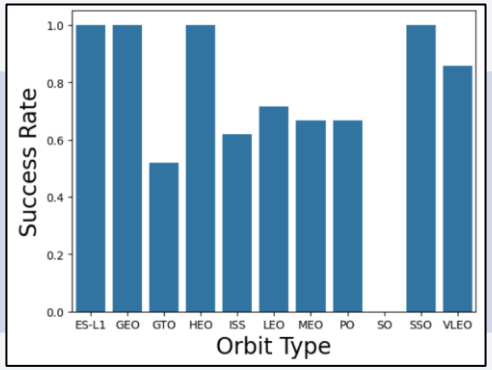
Scatter Plots: Used to show relationships between variables and visualize correlations.

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



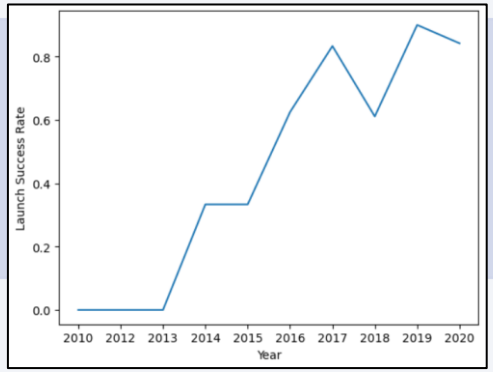
Bar Chart: Used to show relationships between categorical values.

- Success Rate vs. Orbit



Line Graph: Used to show trends for time series data.

- Success Rate vs. Year



GitHub_Link

EDA with SQL

Using sqlalchemy v1.3.9 we performed the following SQL queries:

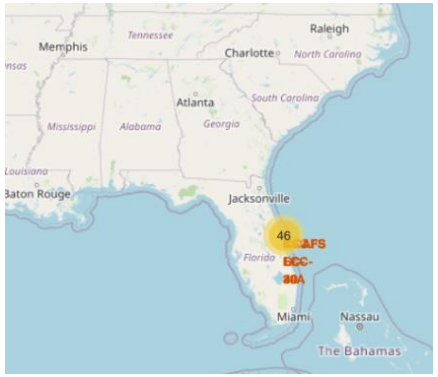
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
- 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

GitHub_Link

Build an Interactive Map with Folium

Created markers, circles, text labels, clusters and lines to visually represent points of interest:

- Added markers with red circles and site name text labels to visually show where the sites are located.
- Created clusters with markers to display different information pertaining to the same coordinates.
- Created colored markers to show successful or failed landings
- Added lines to show the distance from a site location to points of interest, including a coastline, railway and closest metropolitan area.

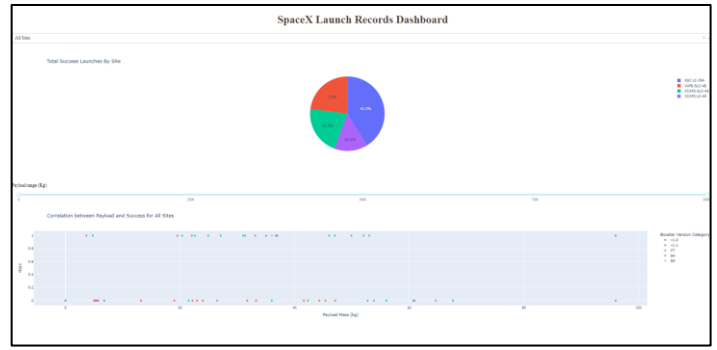


GitHub_Link

Build a Dashboard with Plotly Dash

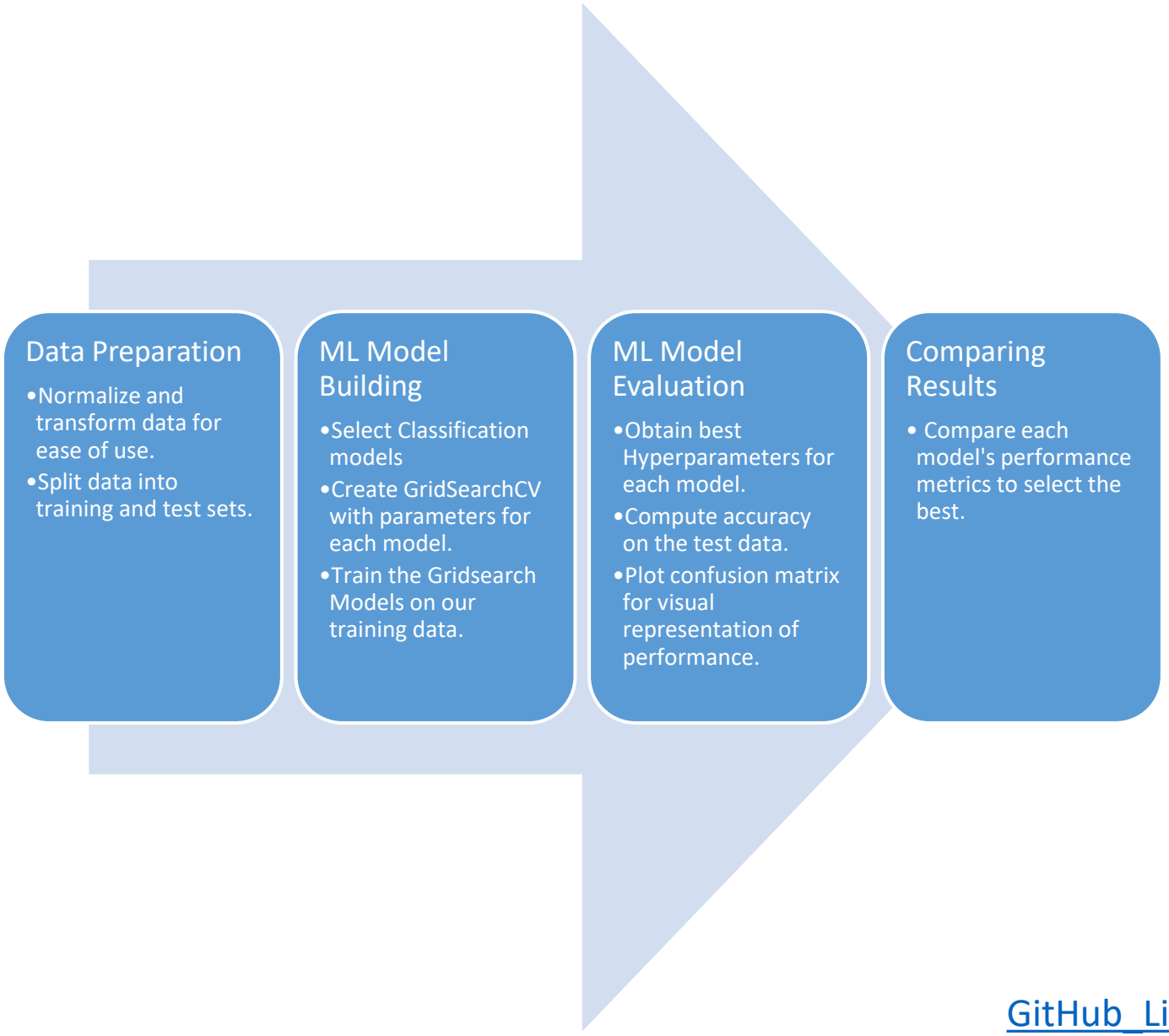
Created a Dashboard with a dropdown and slider. Displaying reactive scatter and pie charts:

- The dropdown allows the user to select if they want to observe visual data for all locations, or a specific location.
- The slider allows the user to narrow their results by specifying payload ranges.
- The Pie Chart displays the percentage of success and failures for each site. If the user wants data on “All” sites, it will display distribution of launches by site.
- The scatter chart displays the correlation between payload mass and landing outcomes.



GitHub Link

Predictive Analysis (Classification)



GitHub_Link

Results

Exploratory data analysis results

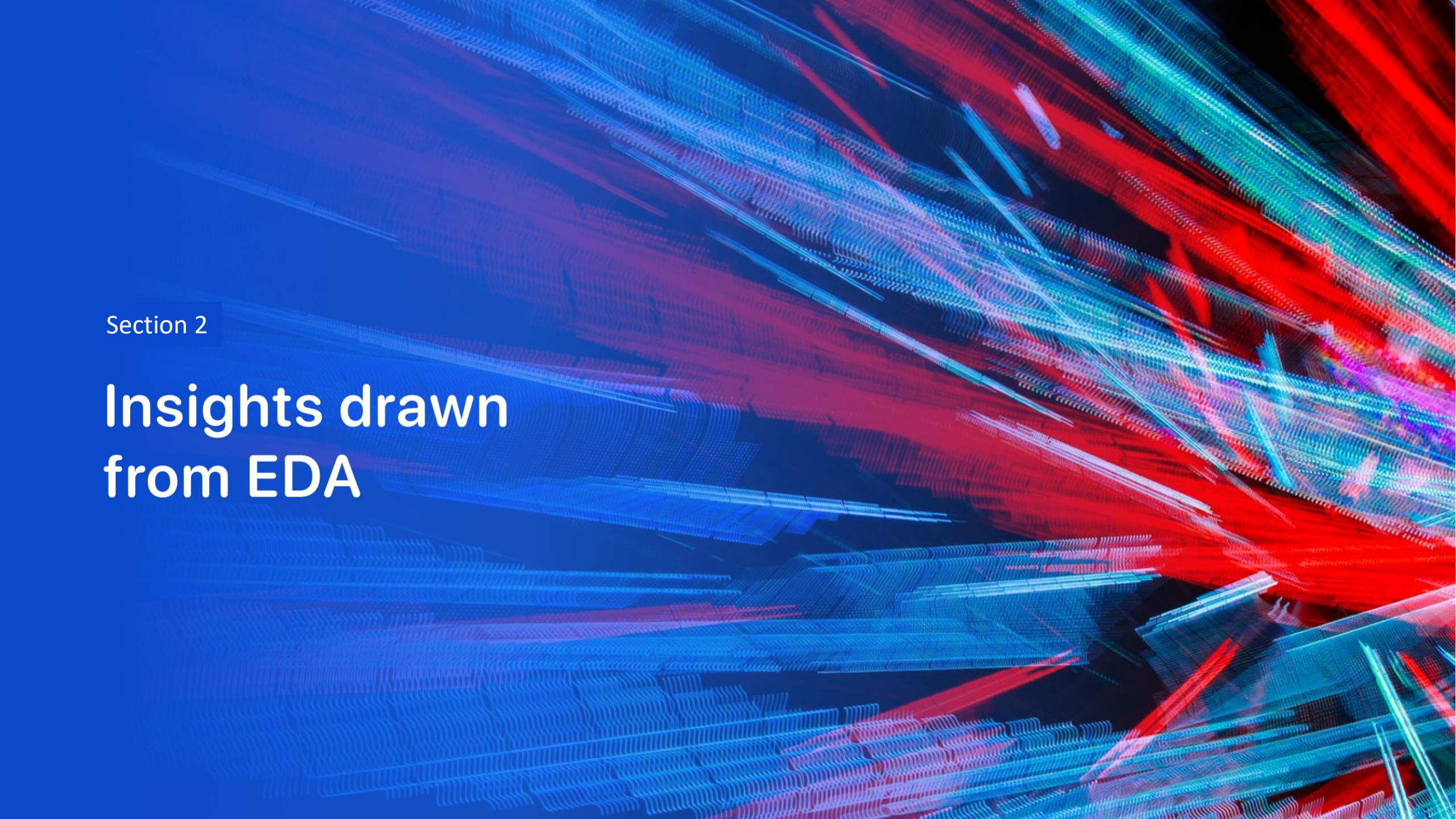
- Location categories is narrowed to 4 sites.
- There are various payloads for each launch.
- Landing success has increased throughout the years.

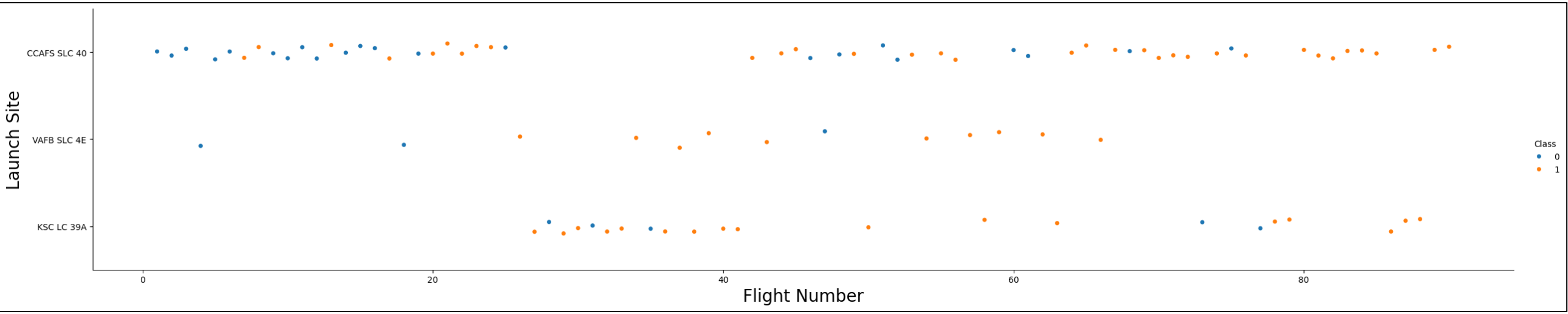
Interactive analytics

- Launches occur near the cost.
- Most launches occur in KSC LC-39A
- Booster v1.1 has the most failures

Predictive analysis results

- Decision Tree Classifier performs best for predicting the outcome of a launch.

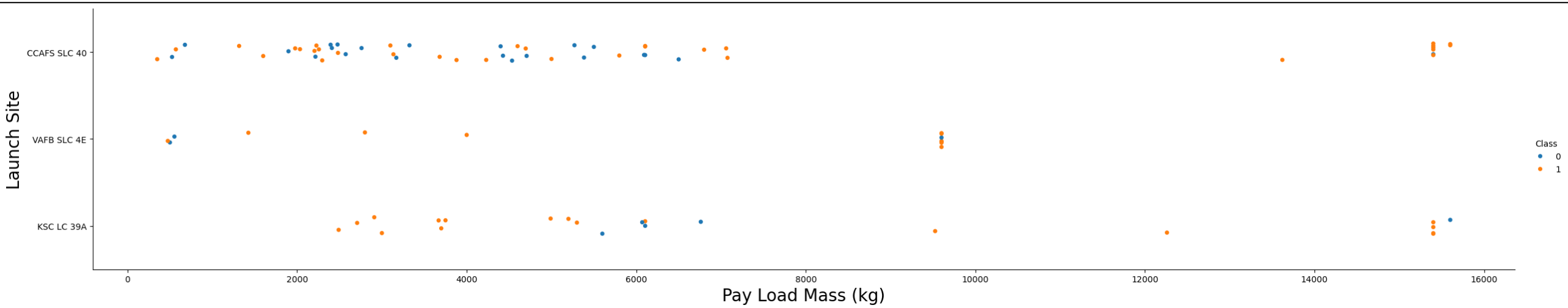

Section 2
Insights drawn from EDA



Observations:

- Most launches occur at CCAFS SLC40
- Early launches failed and recent success rate is 100%

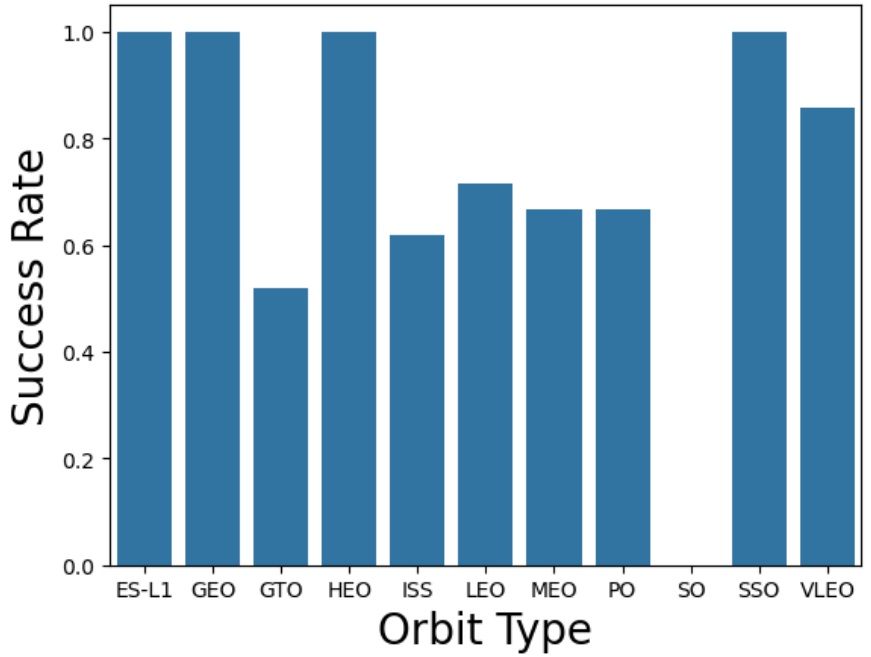
Flight Number vs. Launch Site



Observations:

- Payload and success rate seem to be positively correlated
- KSC LC 39A has a perfect success rate for payloads bellow 5500 kg
- Success Rate for pay loads over 8000 kg are higher then those bellow

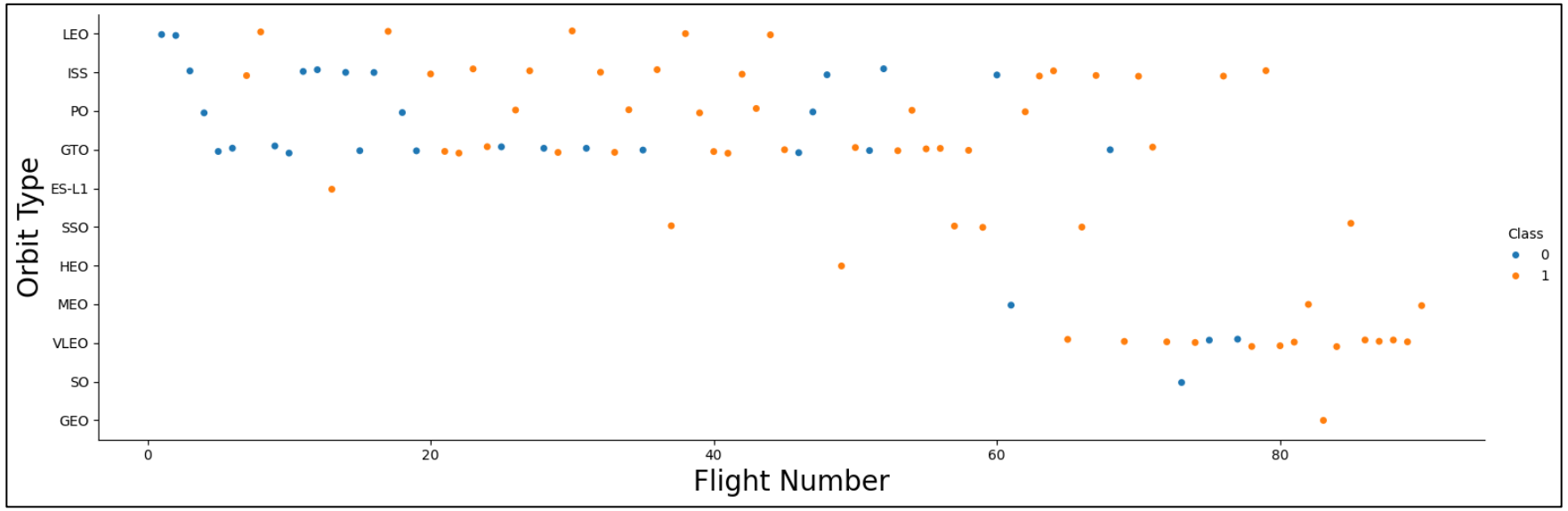
Payload vs. Launch Site



Observations:

- Multiple orbit types have 100% success rate
- Orbit type SO has a 0% success rate

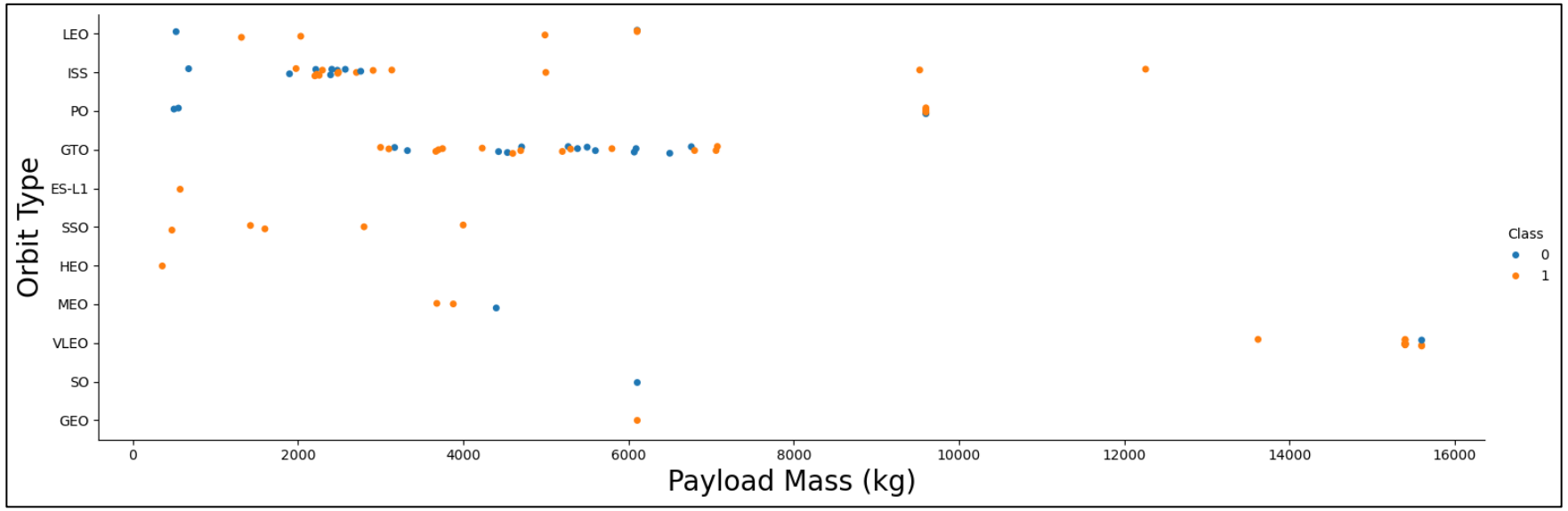
Success Rate vs. Orbit Type



Observations:

- Success Rate improves as over time
- VLEO seems to have become the most common type of launch orbit

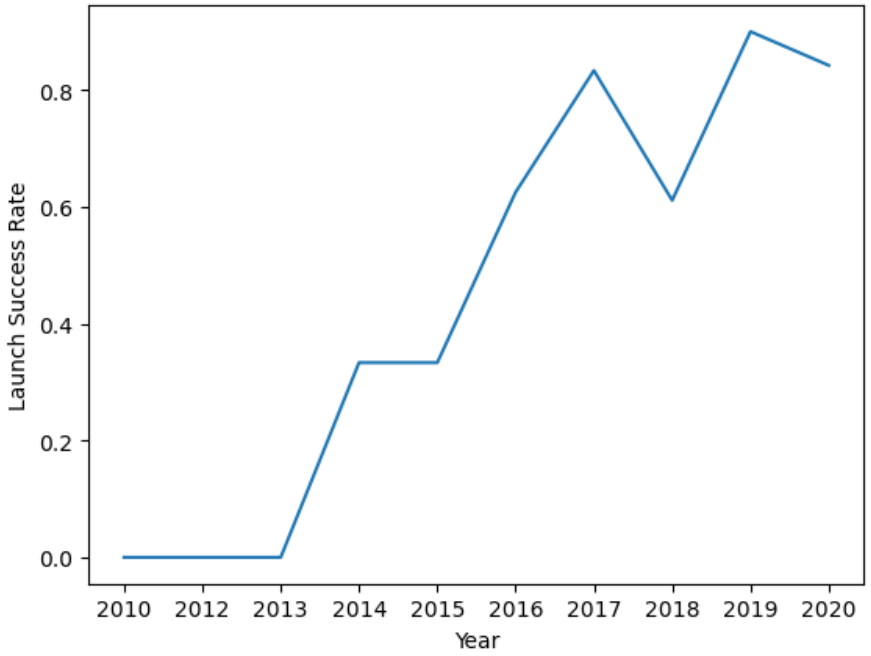
Flight Number vs. Orbit Type



Observations:

- Launches seem to try to keep a payload bellow 8000kg
- VLEO is the only orbit that has attempted payloads above 13000

Payload vs. Orbit Type



Observations:

- Since 2013 success rate has been increasing

Launch Success Yearly Trend

Explanation : Run a query to retrieve all the unique launch sites

```
%%sql
SELECT DISTINCT "Launch_Site"
FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

SQL: All Launch Site Names

Explanation : Run a query to retrieve data on sites who's name begins with "CCA"

```
%%sql
SELECT * FROM SPACEXTBL
WHERE "Launch_Site" LIKE "CCA%"
LIMIT 5
```

```
 * sqlite:///my_data1.db
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

SQL: Launch Site Names Begin with 'CCA'

Explanation : Run a query to calculate the total payload mass

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL
WHERE "Customer" = "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

SUM("PAYLOAD_MASS__KG_")
45596

SQL: Total Payload Mass

Explanation : Run a query to calculate the average payload mass for F9 v1.1

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL
WHERE "Booster_Version" = "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

AVG("PAYLOAD_MASS__KG_")
2928.4

SQL: Average Payload Mass by F9 v1.1

Explanation : Run a query to receive the date of the first successful ground landing

```
%%sql
select min("Date") from SPACEXTBL
where "Landing_Outcome" = "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

min("Date")
2015-12-22

SQL: First Successful Ground Landing Data

Explanation : Run a query to list the boosters which have a success in drone ship and have a payload mass greater then 4000 but less than 6000

```
%%sql
select "Booster_Version" from SPACEXTBL
where "Landing_Outcome" = "Success (drone ship)"
and "PAYLOAD_MASS__KG_" between 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

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

SQL: Successful Drone Ship Landing with Payload Between 4000 and 6000

Explanation : Run a query to list the total number of successful and failure mission outcomes

```
%%sql
select "Mission_outcome" , count("Mission_Outcome")from SPACEXTBL
group by "Mission_outcome";
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	count("Mission_Outcome")
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

SQL: Total Number of Successful and Failure Mission Outcomes

Explanation : Run a query to list the names of booster versions which have carried the maximum payload mass.

```
%%sql
select Booster_Version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
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

SQL: Boosters Carried Maximum Payload

Explanation : Run a query to display the month names, failure landing outcomes in drone ship, booster version, launch site for the months in 2015

```
%%sql
select substr(Date, 6,2) as month, date, Booster_Version, Launch_Site, Landing_Outcome from SPACEXTBL
where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

SQL: 2015 Launch Records

Explanation : Run a query to rank the count of landing outcomes between 2010-06-04 and 2017-03-20

```
%%sql
select Landing_Outcome, count(*) as Outcome_Counts from SPACEXTBL
where date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by Outcome_Counts desc;
```

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

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

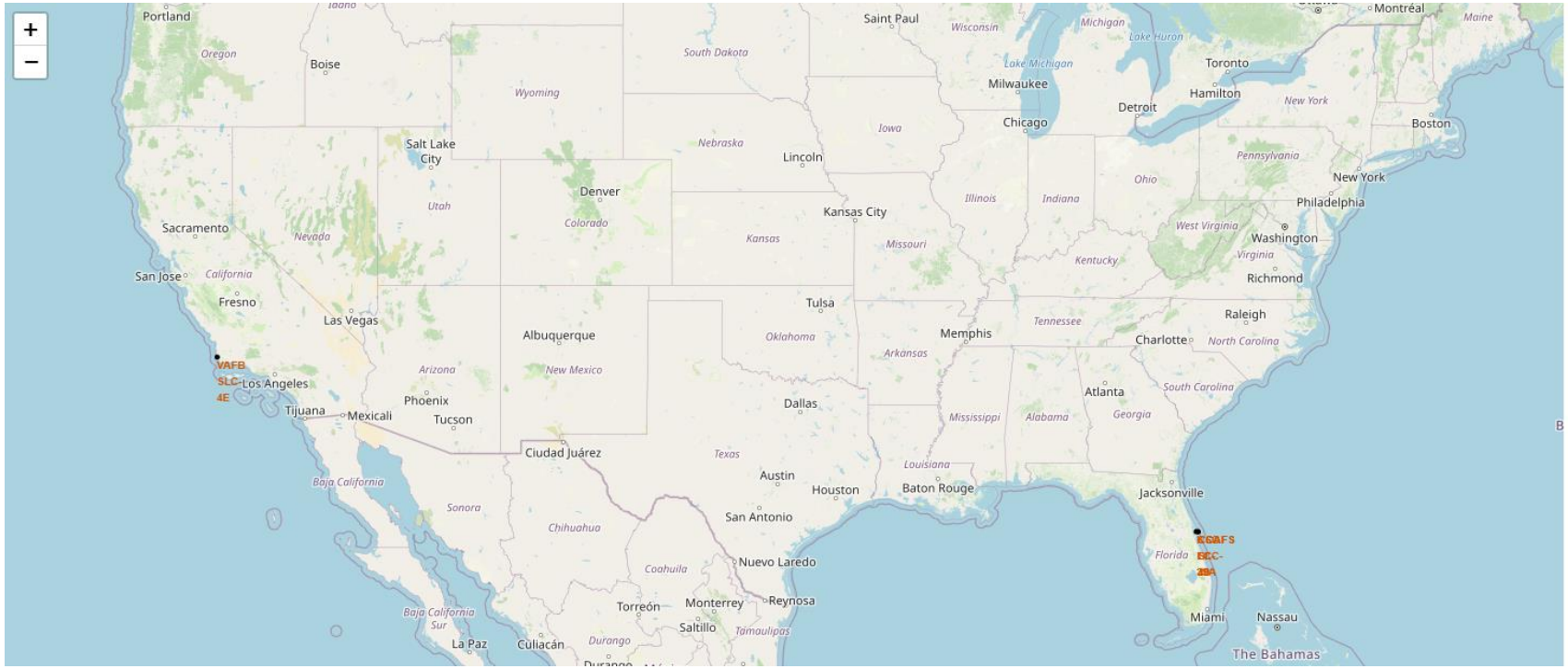
Section 3

Launch Sites Proximities Analysis

Folium: All Launch Sites

Insight:

We observe that all launch sites are located near the ocean.

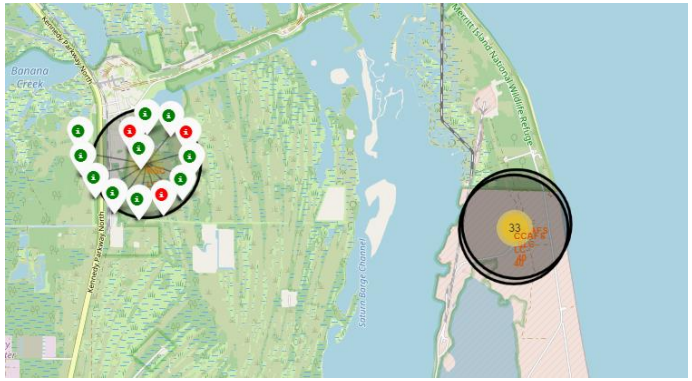


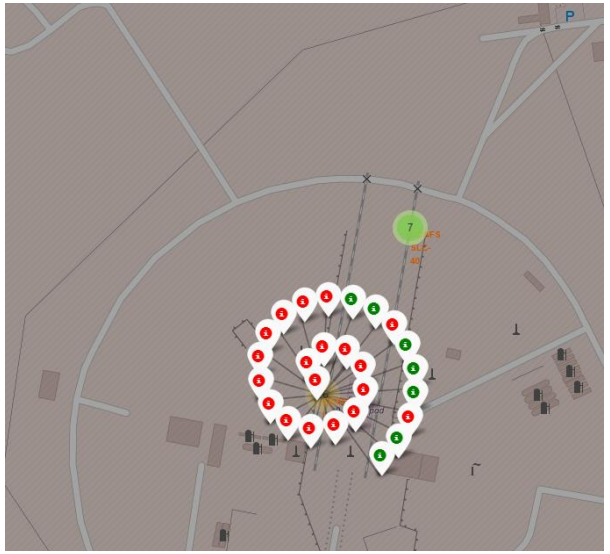
Folium: Launch Outcome markers

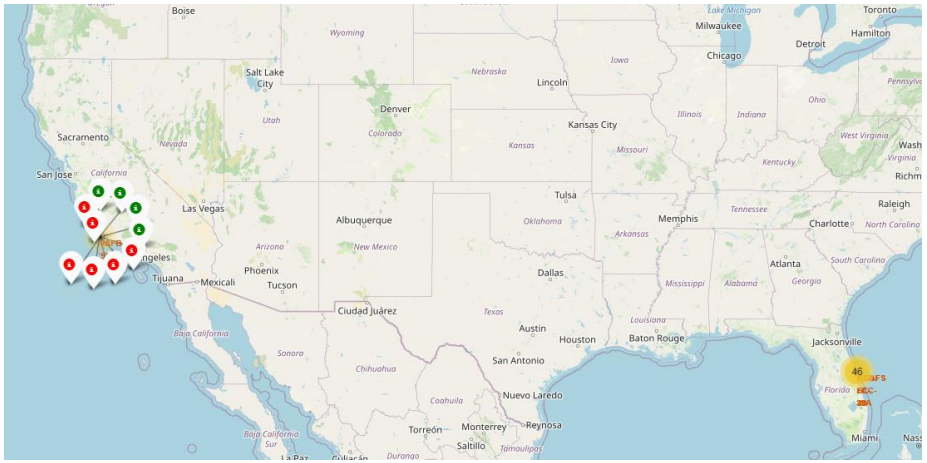
Explanation:

Color labeled markers were added to each site to quickly get a visual representation of the success rates.

- Red == Failure
- Green == Success



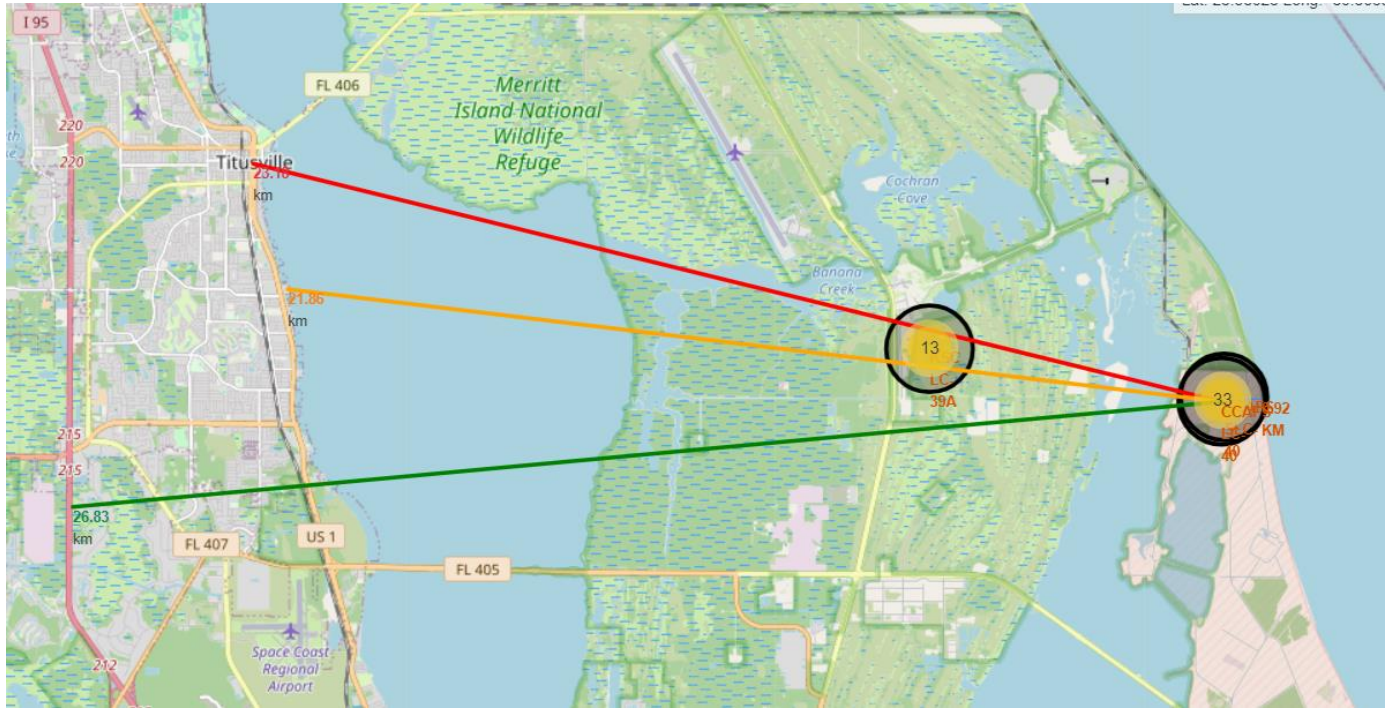


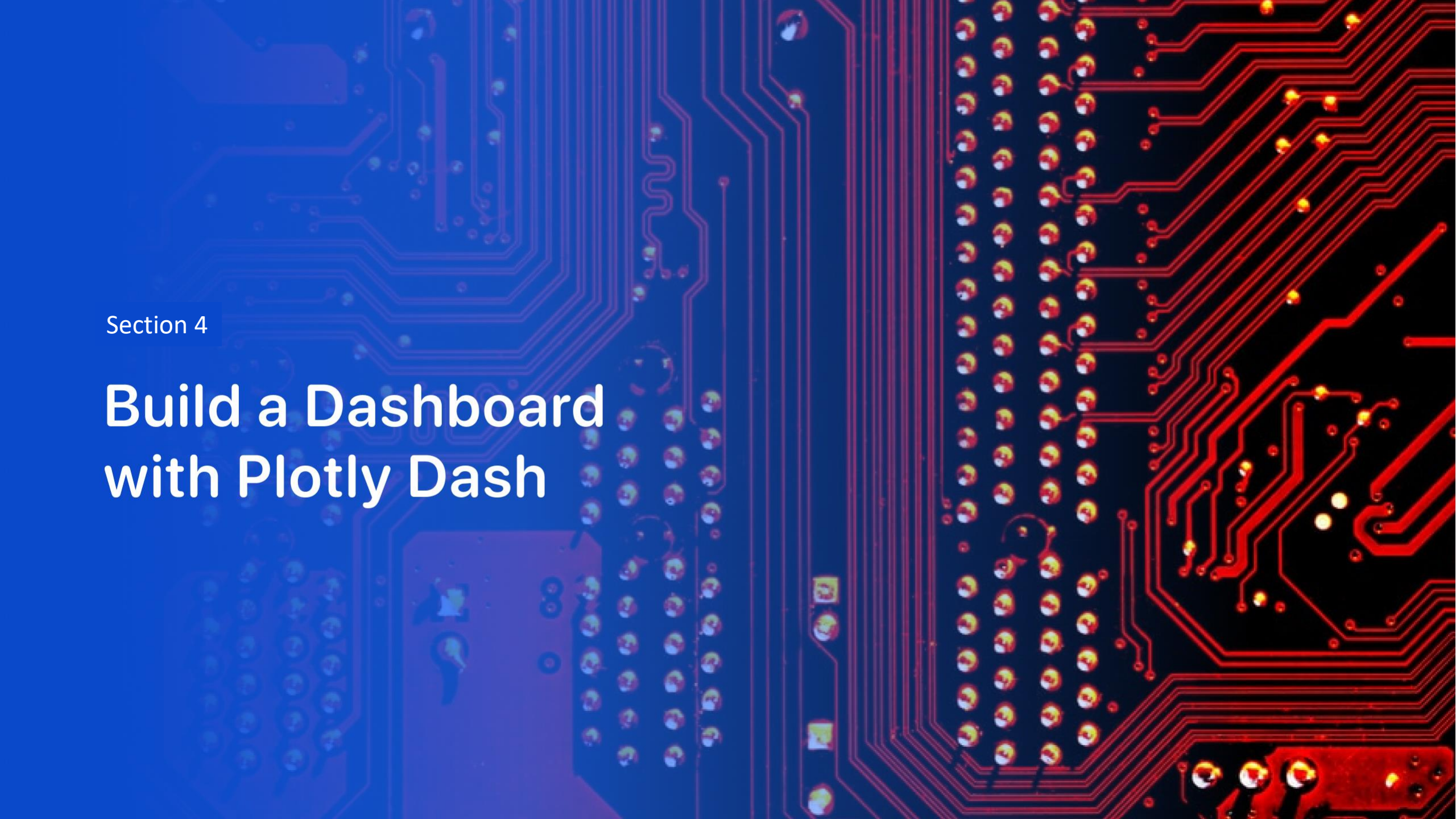


Folium: Distance from Launch site to Points of Interest

Explanation: Added lines and labels to show the distance from CCAFS SLC-40 to specific points of interest:

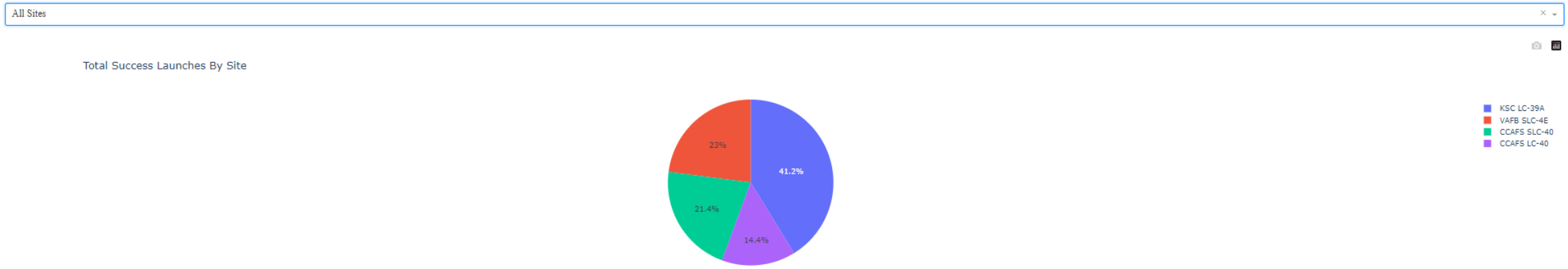
- A Coastline
- A Railway
- A close Metropolitan area




Section 4
Build a Dashboard with Plotly Dash

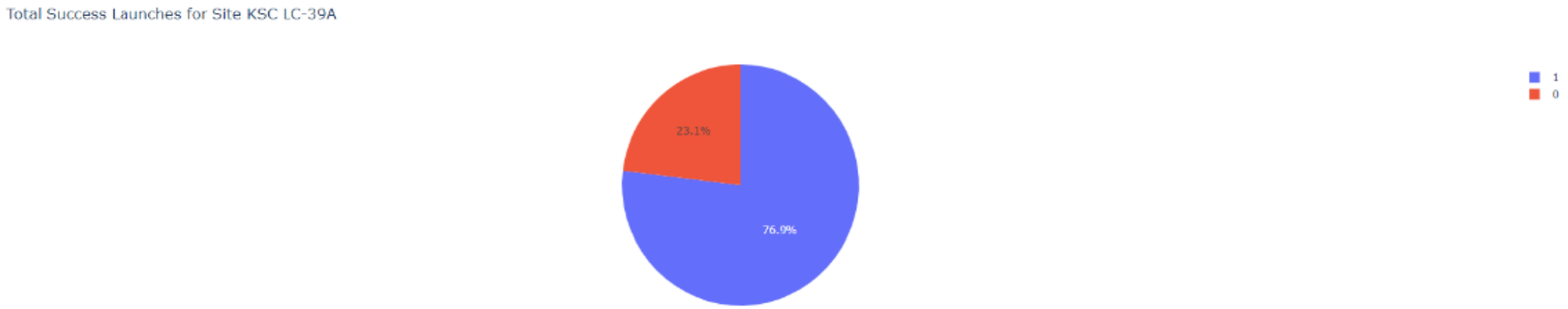
Dash: Launch Success – All Sites

Findings: The pie chart clearly shows that most successful come from KSC LC39A.



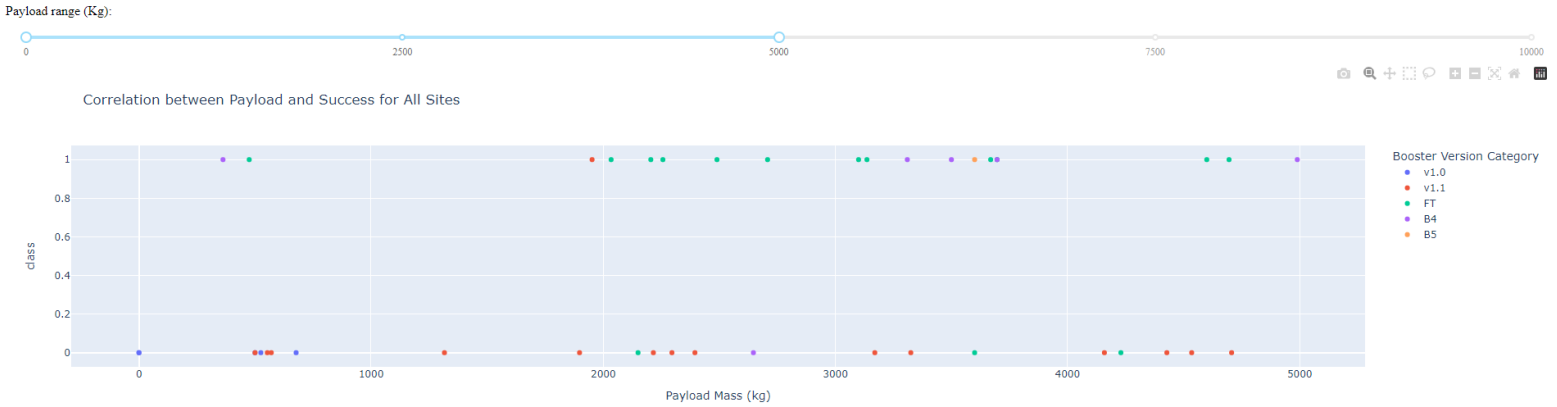
Dash: Launch Site with Highest Success Rate

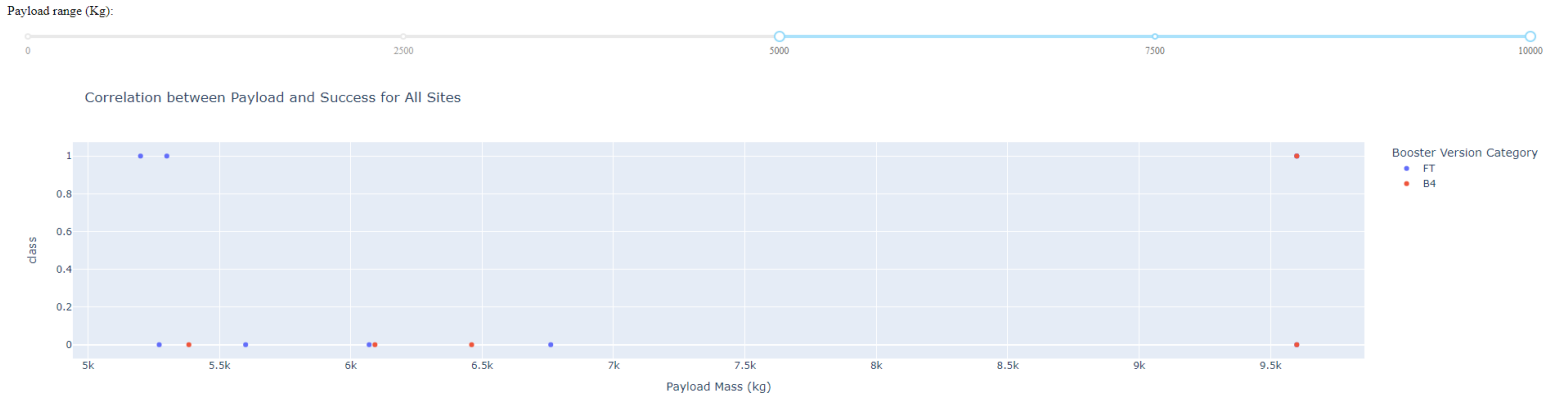
Findings: The pie chart shows that KSC LC-39A has a success rate of 76.9%



Dash: Low and High Payload Mass vs. Outcomes for (All Sites)

Findings: The Plots show that the success rate is higher for payloads bellow 5000kg





Section 5

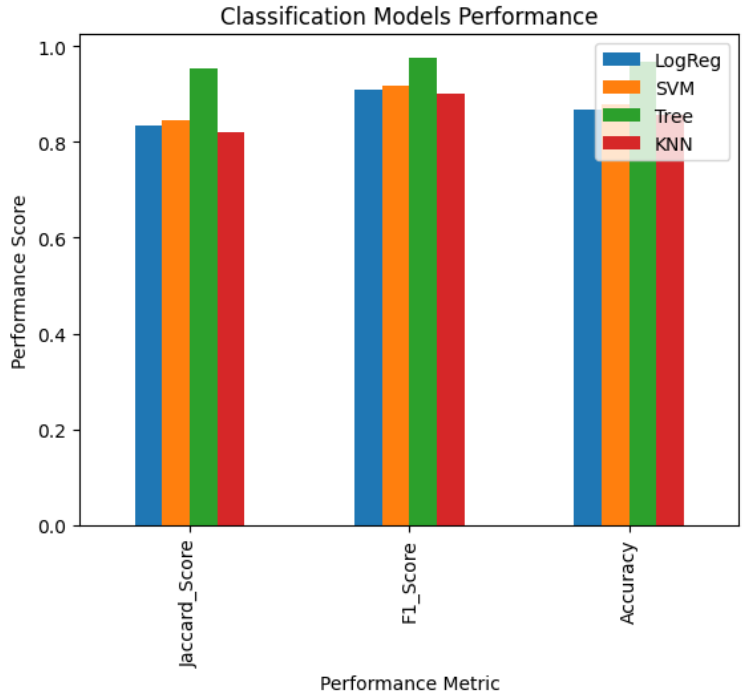
Predictive Analysis (Classification)

ML - Classification Accuracy

Explanations:

- Beyond accuracy, the Jaccard and F1 scores were calculated to help us choose the best model. The scores and plot is shown
- Based on these results we can see that the **Decision Tree** is the best model.

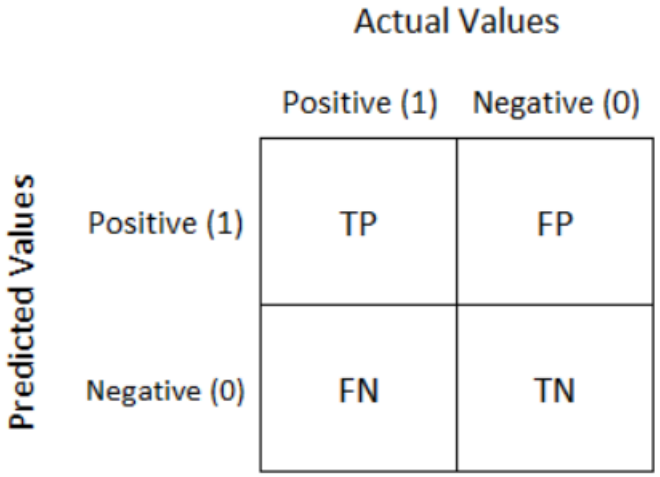
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.952381	0.819444
F1_Score	0.909091	0.916031	0.975610	0.900763
Accuracy	0.866667	0.877778	0.966667	0.855556

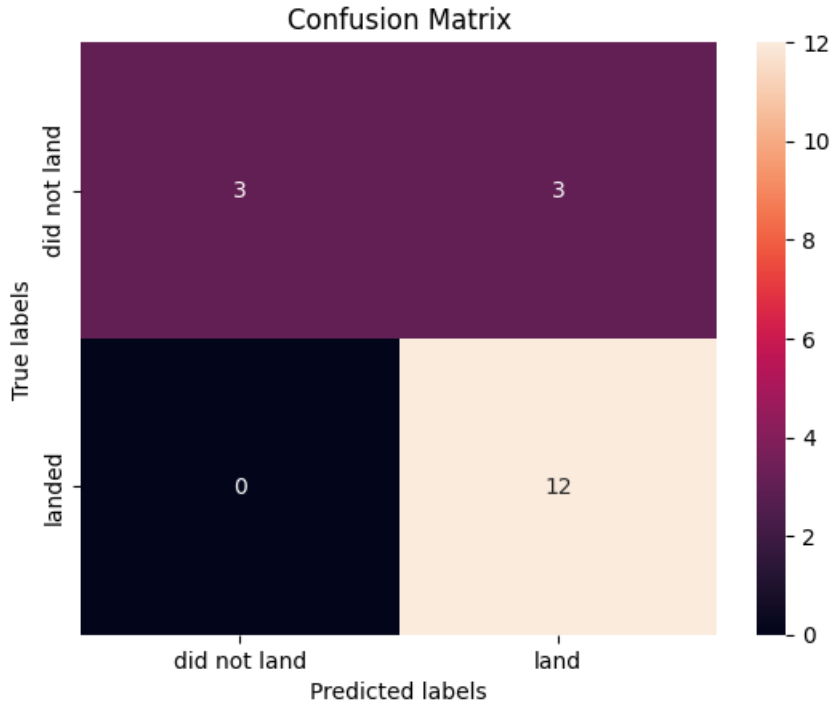


ML – Confusion Matrix Decision Tree

Explanations:

- The confusion matrix reports the number of true positives, false negatives, false positives, and true negatives
- This shows that with Decision Tree performs great in creating no False Negatives, but can produce false positives.





Conclusions

To try and predict the outcome of a launch, a **Decision Tree Classifier** is the best model to use, but we must be aware that it may produce False Positives.

Orbit SO has yet to have a successful launch and landing. ES-L1, GEO, HEO, SSO have 100% success rates

The success rate of landings has increased dramatically since 2013 and currently hovers around 90%.

Lighter payloads have more success than heavier ones.

The site with best success rates is KSC LC039A.

Launch sites are located near the ocean, likely for safety reasons.

The most common type of failure is “drone ship”

Appendix

- Confusion Matrix reference image was provided via the web from:
  - https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Thank you!