IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Wendy Melyana
19 April 2024

Outline

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

Executive Summary

Summary of methodologies:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
  - SQL
  - Visualization: Pandas and Matplotlib
- Creation of Interactive Visual Analytics and Dashboard
  - Interactive Visuals: Folium (Interactive Map)
  - Interactive Dashboard: Plotly Dash
- Predictive Analysis (Classification)

Summary of all results:

- Exploratory Data Analysis Results
- Interactive Analytics Demos Screenshots
- Predictive Analysis Results

Introduction

Project background and context

With the current commercial space age where different companies are trying to make space travel affordable. We, as a new rocket company Space Y, would like to create a cost-effective rocket.

It is found that one of the most successful rocket company, Space X, manages to launch rockets at a relatively inexpensive cost of 62 millions. Which are much cheaper compared to other companies that are launching it at costs beyond 165 millions.

This cost reductions was because they are able to reuse the first stage of their Falcon 9. Therefore, by being able to predict the successful rate of landing and reusing the first stage, we can then determine the cost required for a launch.

Problems you want to find answers

- Determine the price of each launch
- Whether variables such as payload mass, launch sites, number of flights and orbits effects the successful landing of the first stage
- Finding the best algorithm to classify for each case

Section 1

Methodology

Methodology

Data collection methodology:

- SpaceX REST API
- Web Scrapping table data from Wikipedia

Perform data wrangling

- Filtering of data
- Dealing with missing values
- Usage of One Hot Encoding to prepare the date 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

- Building, fine-tuning and evaluating the classification models to get the best results

Data Collection

The data collected from two sources which was from SpaceX REST API and SpaceX's Wikipedia page. The later was required to assist in creating a complete dataset regarding the successful landing of the rocket itself that was missing in the SpaceX REST API.

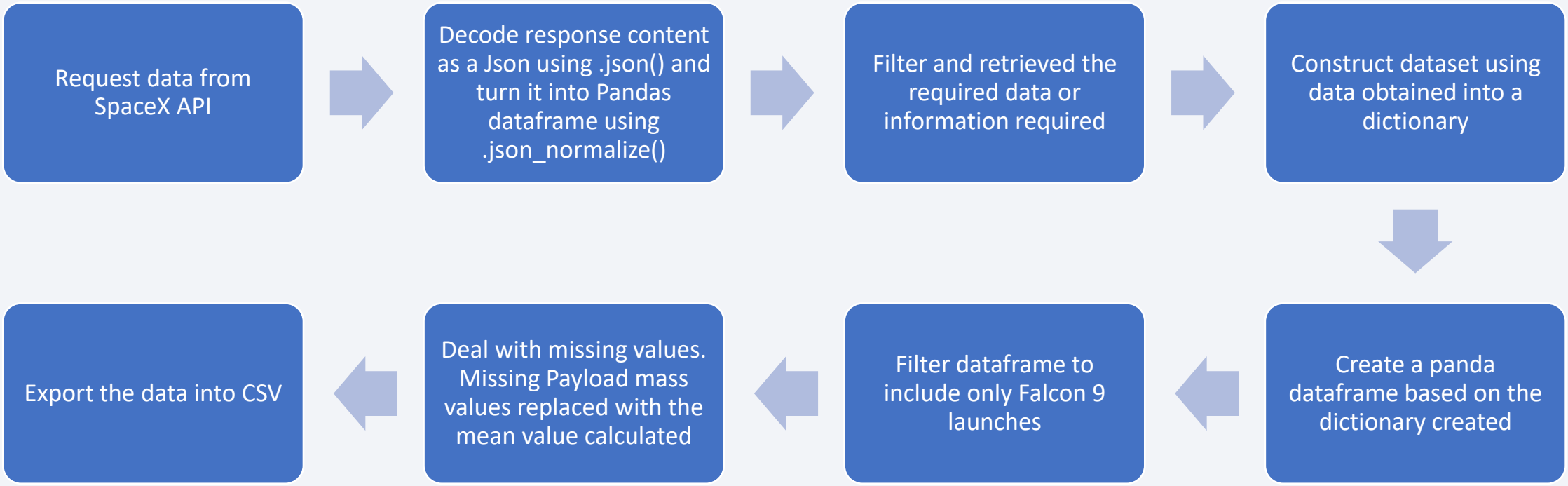
Columns obtain from SpaceX REST API:

- Flight Number
- Date
- Booster Version
- Payload Mass
- Orbit
- Launch Sites
- Outcome
- Flights
- Grid Fins
- Reused
- Legs
- Landing Pad
- Block
- Reused Count
- Serial
- Longitude & Latitude

Columns obtain from SpaceX Wikipedia:

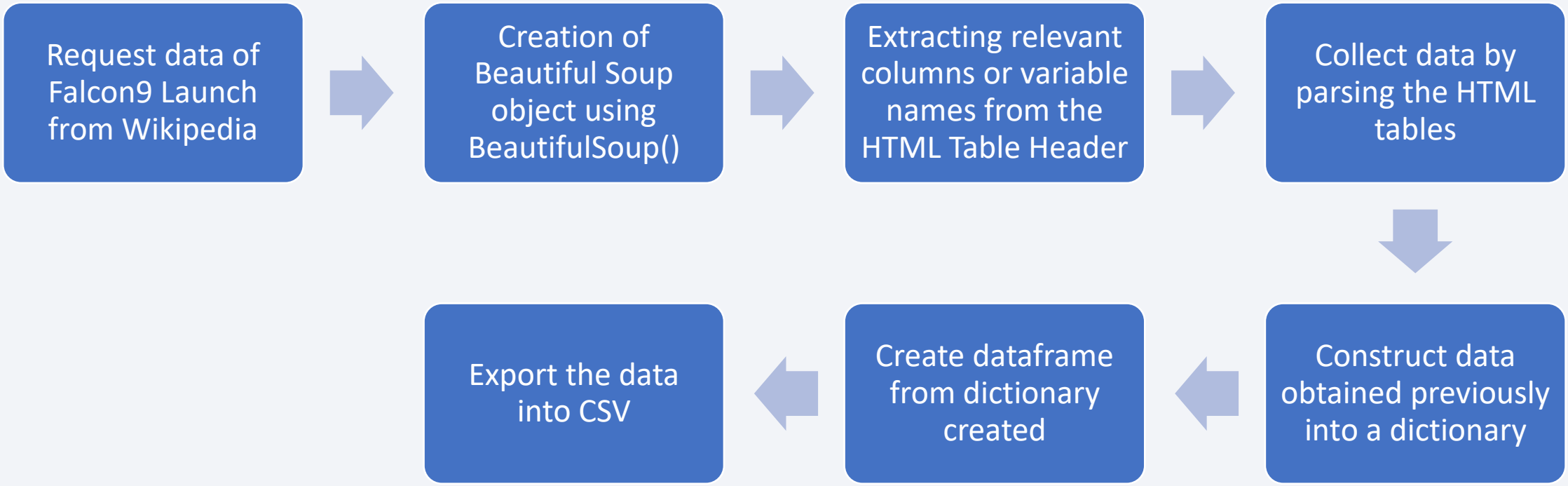
- Flight Number
- Launch Site
- Payload
- Payload Mass
- Orbit
- Customer
- Launch Outcome
- Version Booster
- Booster Landing
- Date
- Time

Data Collection – SpaceX API



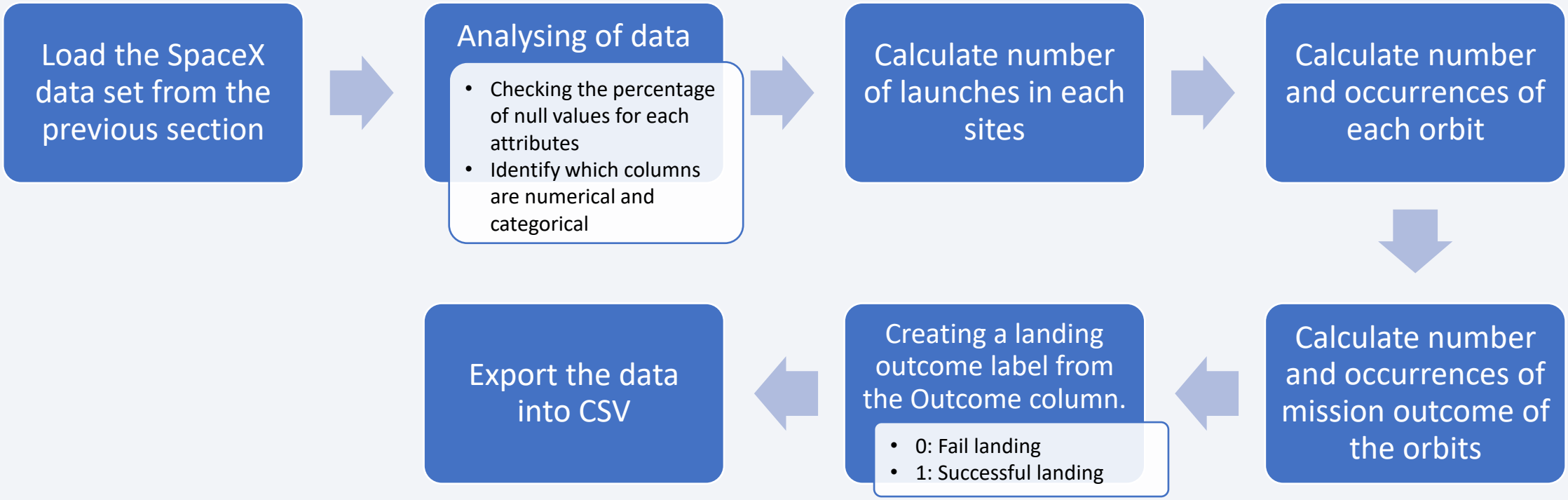
GitHub Link: Data Collection – SpaceX API

Data Collection - Scraping



GitHub Link: Data Collection - Scraping

Data Wrangling



GitHub Link: Data Wrangling

EDA with Data Visualization

Charts Plotted:

- **Scatterplot:**
  - Flight Number vs Launch Site
  - Payload vs Launch Site
  - Flight Number vs Orbit Type
  - Payload vs Orbit Type
- **Bar Chart:**
  - Successful Rate of each Orbit Type
- **Line Chart:**
  - Yearly trend of Successful Launches

Scatterplot was used to show the relationship between variables to see if there exists any relationships between them. If there is, it could then be used for machine learning model.

Bar Chart is useful to show comparisons among the different categories easily. In this case, to show how each categories of different orbit types compares and measured between each other.

Line Chart can easily display visualization of any trends of data over time, in this case by the year.

GitHub Link: EDA with Data Visualization

EDA with SQL

SQL Queries Performed:

- Display the names of 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 the 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 between 4000 to 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster version which have carried the maximum payload mass.
- List the records that displays the month names, failure landing outcome in drone ship, booster version, launch site for the months in year 2005
- Rank the count of landing outcomes between the date 2010- 06-04 to 2017-03-20 in descending order

GitHub Link: EDA with SQL

Build an Interactive Map with Folium

Map Objects Created and Added:

- **Circles:**
  - Launch Sites
- **Markers:**
  - Launch Sites
  - Successful or Failed launches for each Launch Sites
  - Nearest landmarks from CCAFS SLC-40
    - Closest Coastline
    - Closest City
    - Closest Railway
    - Closest Highway
- **Lines:**
  - Nearest landmarks from CCAFS SLC-40
    - Closest Coastline
    - Closest City
    - Closest Railway
    - Closest Highway

Circles added to help us identify the location of the launch site easily by its distinct shape and color against the world map itself.

Colored markers used to assist us to differentiate each launch site and its number of successful and unsuccessful launches, green and red markers, respectively.

Lines helps us to draw a distance line to show how far it is from the nearest landmarks clearly, instead of second guessing which point in the map itself.

GitHub Link: Build an Interactive Map with Folium

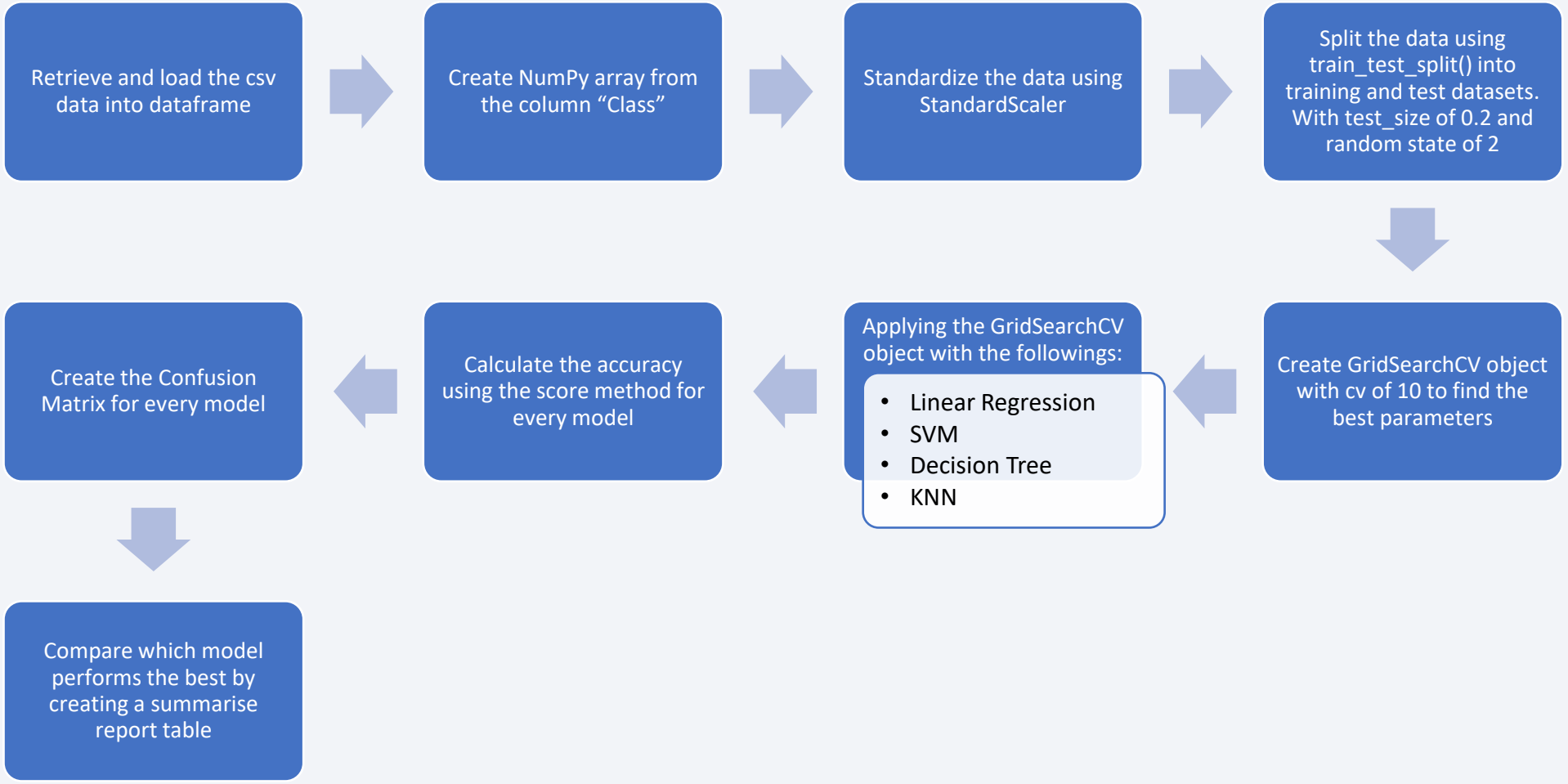
Build a Dashboard with Plotly Dash

Plots and Graph Interactions Added:

- **Launch Sites drop down selection**
  - To allow users to choose between showing the overview of all launch sites or any specific sites
- **Pie Chart**
  - Success rate of each launch sites when compared to each other
  - Success rates of each launch sites when narrowed down
- **Payload Mass range slider**
  - To allow users to adjust the payload mass for each launch
- **Scatter Chart**
  - Payload mass vs Scatter Rate for different Booster Versions

GitHub Link: Dashboard with Plotly Dash

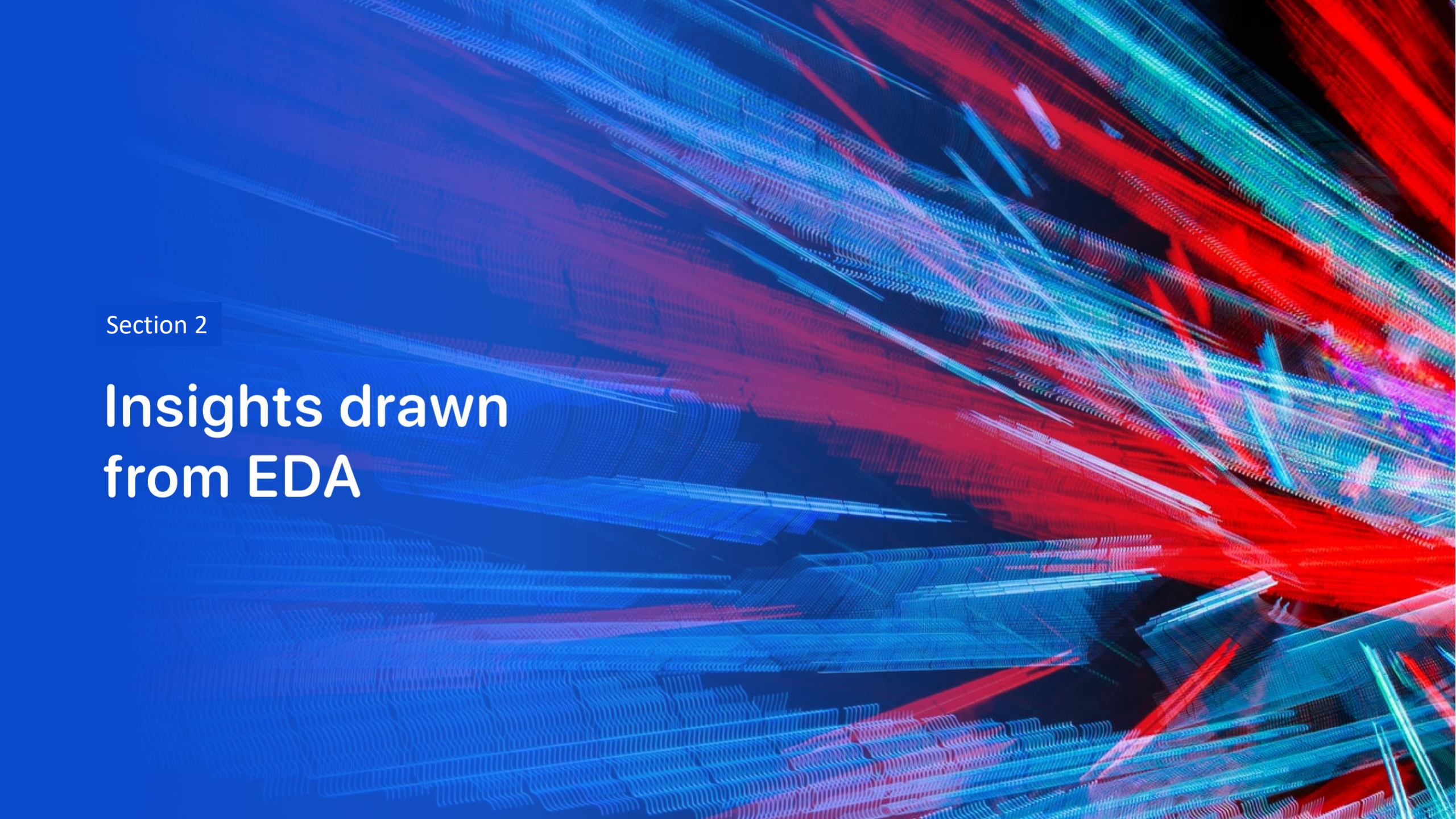
Predictive Analysis (Classification)



[GitHub Link: Predictive Analysis (Classification)]

Results

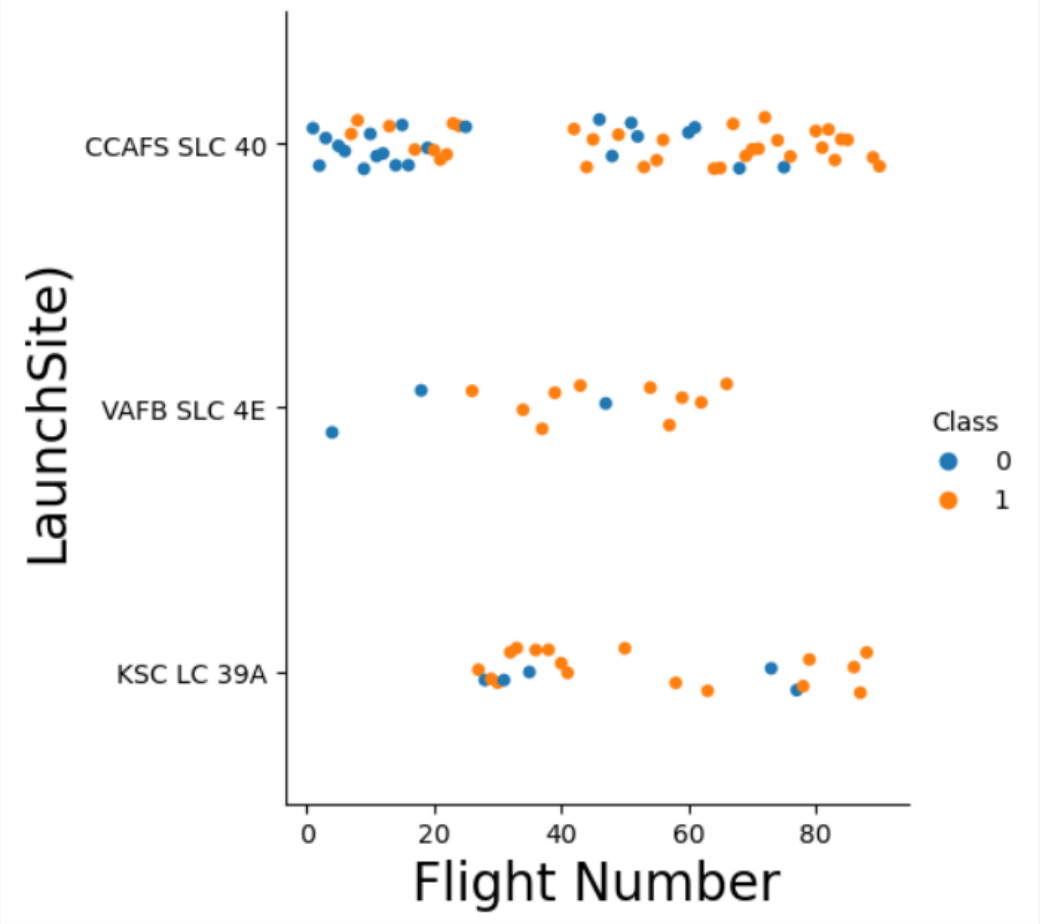
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

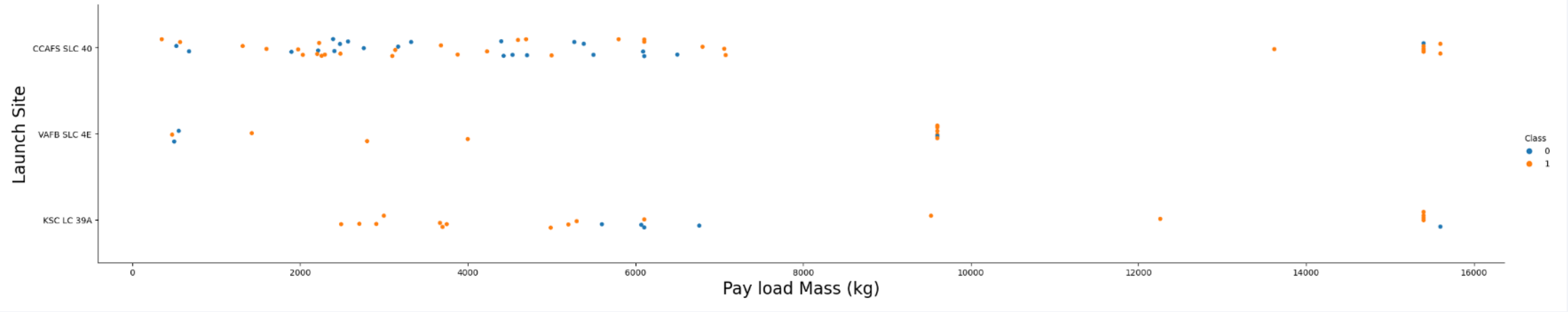
Exploratory Data Analysis with Visualization

Flight Number vs. Launch Site

- Older flights tend to have higher failures while newer flights have higher success
- Most launches done at CCAFS SLC 40
- More successful launches can be found in Launch Site VAFB SLC 4E and KSC LC 39A



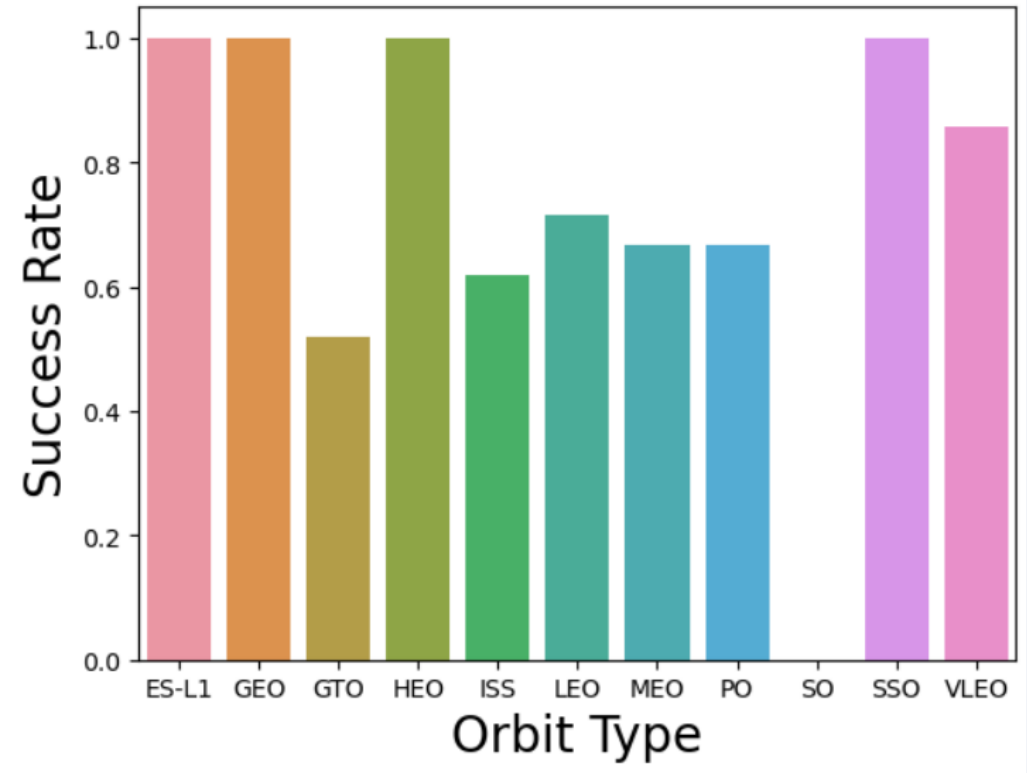
Payload vs. Launch Site



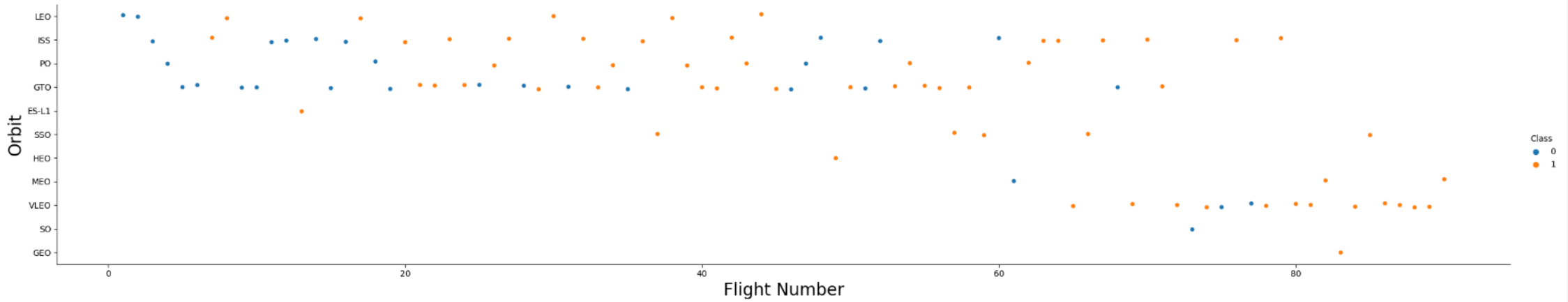
- There are no heavy payload launches at Launch Site VAFB SLC 4E. The max it has done is below 10,000kg
- Relatively there are higher successful launch rates for Payload Mass above the 8,000kg mark compared to those below the 8,000kg mark
- Launch Site KSC LC 39A has a good launching records for Payload Mass below 6,000kg mark

Success Rate vs. Orbit Type

- Orbit Types with 100% Success Rate Record:
  - ES-L1, GEO, HEO and SSO
- Orbit Types with 0% Success Rate Record:
  - SO
- Orbit Types with Success Rate Record from 50% to 85%:
  - GTO, ISS, LEO, MEO, PO, VLEO

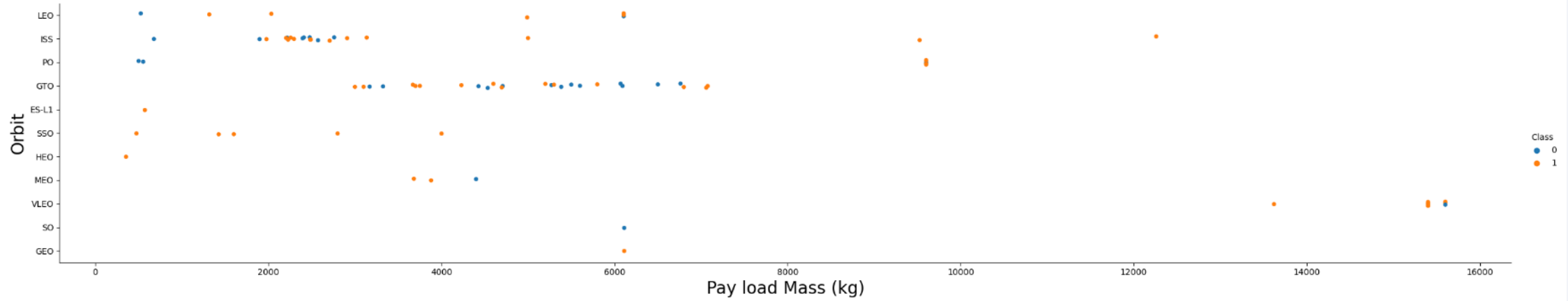


Flight Number vs. Orbit Type



- LEO orbit success rate is correlated to the number of flight
- No correlation can be seen for orbit type GTO with the number of flight

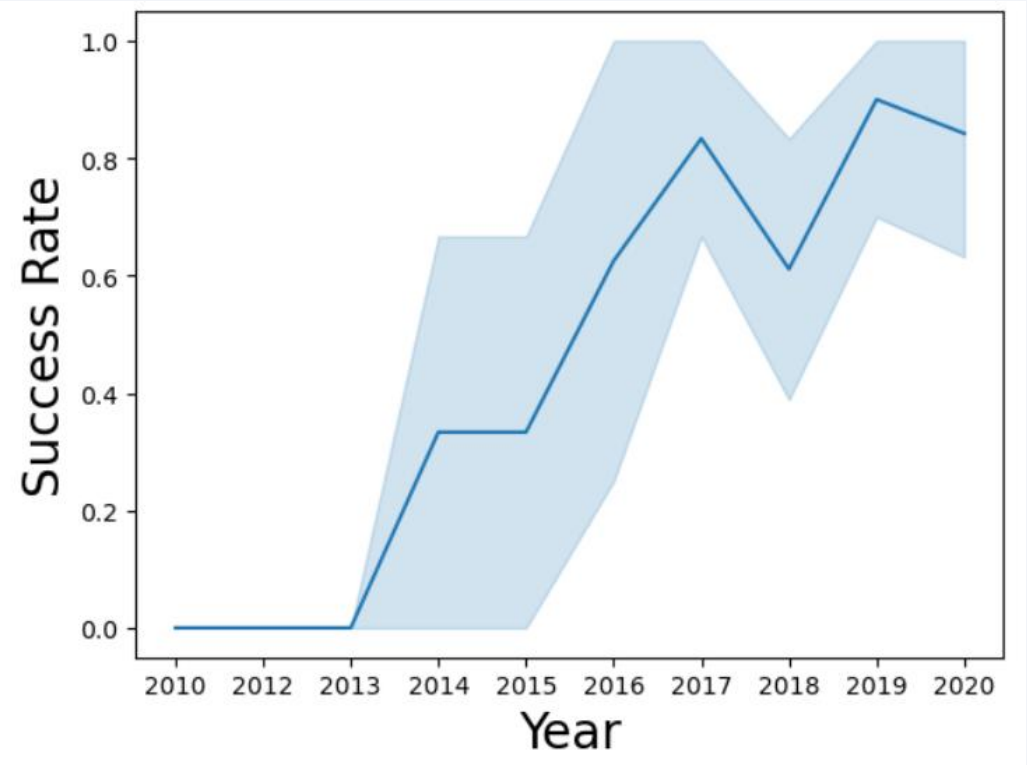
Payload vs. Orbit Type



- For heavier Payload Mass, more successful landing can be found for orbit type LEO, ISS and PO
- Orbit type SSO has a more successful landing for lighter Payload Mass
- Nothing can be derived for orbit type GTO since there is a mixture of successful and failure landing despite its Payload Mass

Launch Success Yearly Trend

- A steady increase in successful launches can be found from 2013
- There is a drastic drop in successful launches seen between the year 2017 and 2018
- Successful launches rate improves after 2018 to 2019 before another slight drop from 2019 to 2020



Exploratory Data Analysis with SQL

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
In [70]: %sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[70]:

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

- Display the names of the unique launch sites in the space mission
- There are 4 unique launch sites

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [18]:
```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5
```

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

Out[18]:

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

- Display 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [32]: %sql SELECT Customer, SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
 * sqlite:///my_data1.db
Done.
```

Out[32]:

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

- Display the total payload mass carried by boosters launched by NASA (CRS)
- The total payload mass is 45,596 kg

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
In [36]: %sql SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1"
 * sqlite:///my_data1.db
Done.
```

Out[36]:

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

- Display average payload mass carried by booster version F9 v1.1
- The average Payload Mass carried by Booster Version F9 v1.1 IS 2,928.4 kg

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [41]:
```
%%sql
SELECT MIN(Date), Landing_Outcome
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE "Success%"
LIMIT 1
```

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

Out[41]:

MIN(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

- List the date when the first successful landing outcome in ground pad was achieved.
- The date is on 22 December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [47]:
```
%%sql
SELECT Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE "%Success%drone ship%" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

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

Out[47]:

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

- List 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

Task 7

List the total number of successful and failure mission outcomes

In [51]:
```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome)
FROM SPACEXTABLE
GROUP BY Mission_Outcome
```

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

Out[51]:

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

- List the total number of successful and failure mission outcomes
- 99 Success count
- 1 Failure count
- 1 Success count with Payload Status being unclear

Boosters Carried Maximum Payload

- List the names of the booster_versions which have carried the maximum payload mass

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [55]:
```
%%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
)
```

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

Out[55]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [60]:
```
%%sql
SELECT Date, substr(Date, 6, 2), Landing_Outcome, Booster_Version
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) = "2015"
AND Landing_Outcome LIKE "%Fail%drone ship%"
```

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

Out[60]:

Date	substr(Date, 6, 2)	Landing_Outcome	Booster_Version
2015-01-10	01	Failure (drone ship)	F9 v1.1 B1012
2015-04-14	04	Failure (drone ship)	F9 v1.1 B1015

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

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

Task 10

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.

In [68]:
```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Count
FROM SPACEXTABLE
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT(Landing_Outcome) DESC
```

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

Out[68]:

Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- 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.

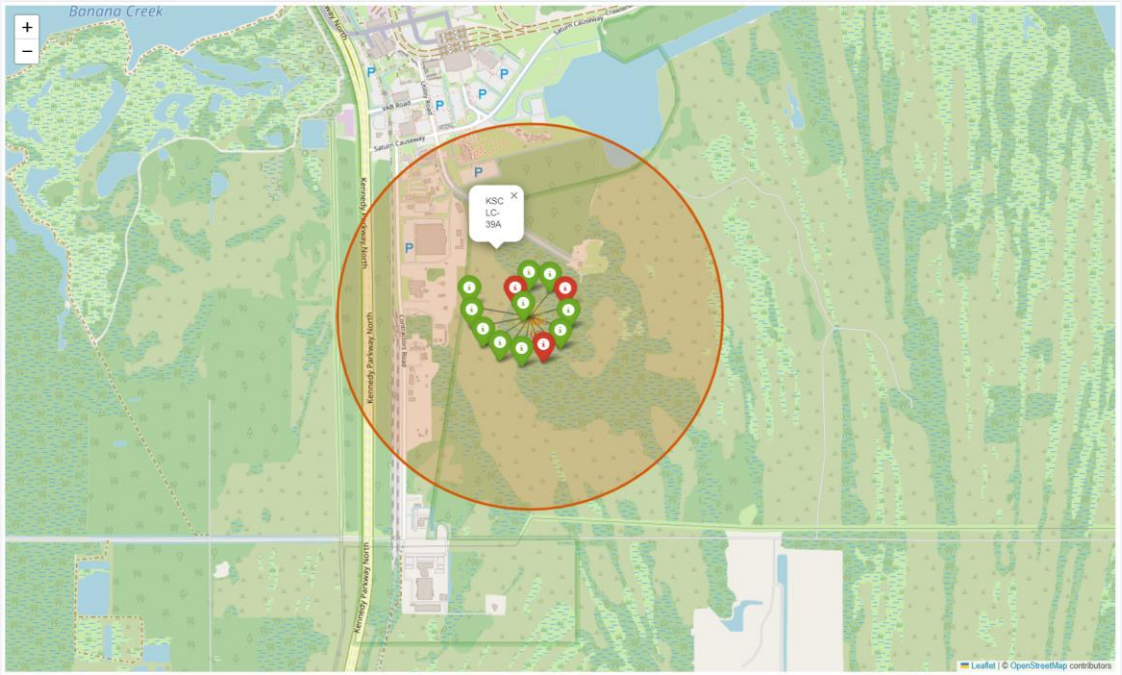
Section 3

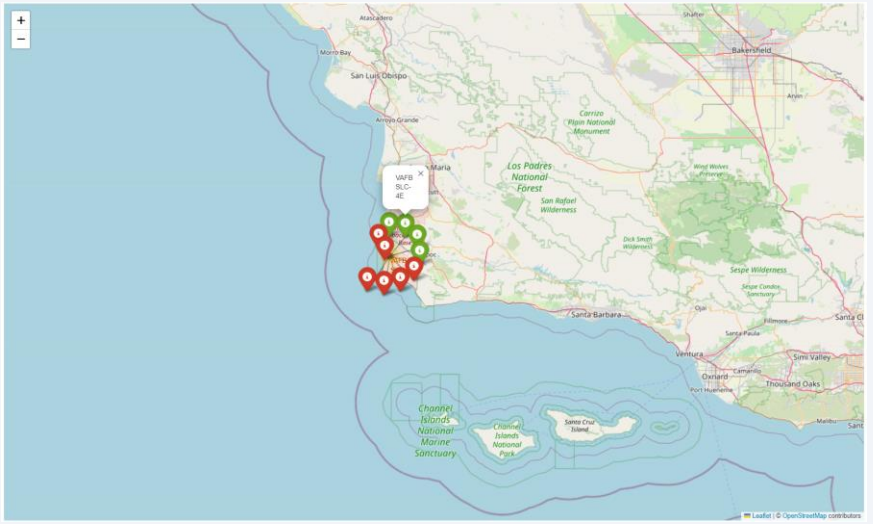
Launch Sites Proximities Analysis

Launch Sites' location markers on the Global Map

- The launch sites are in close to proximity to the equator
- The land tends to move faster the closer it is to the equator
- This will require lesser propellant as compared to launching further from the equator. [1]
- Launch sites are relatively close to the coast. This is to ensure that any fallen debris will fall into the water instead of human civilizations

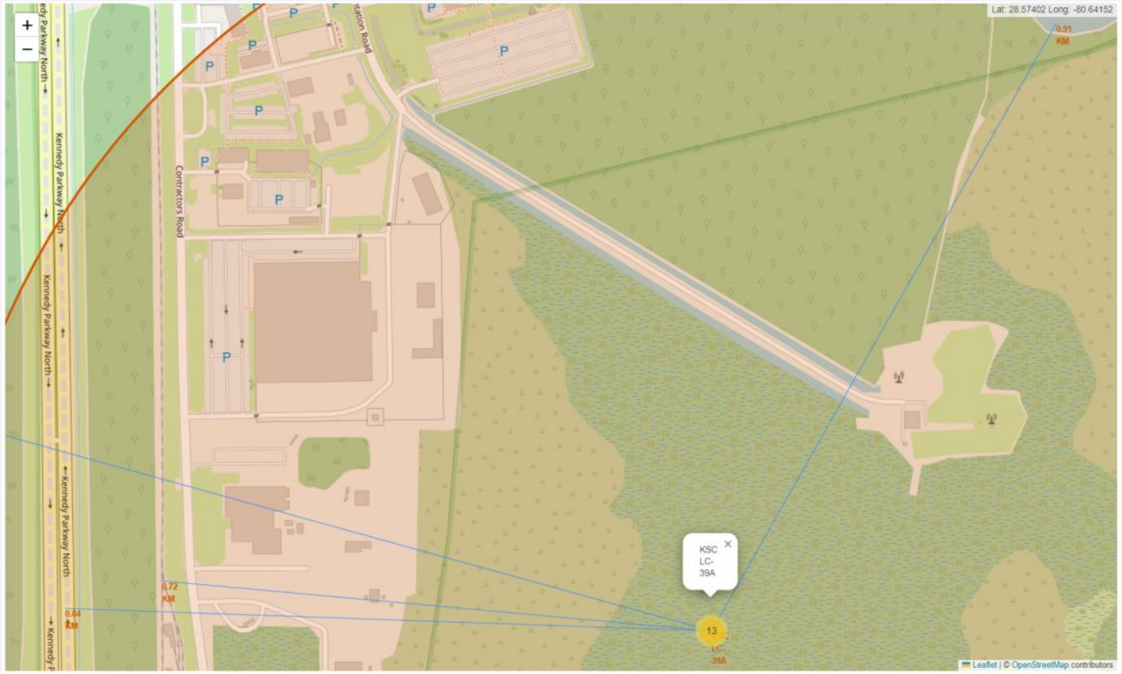
Success Rate of Launches at KSC LC-39A





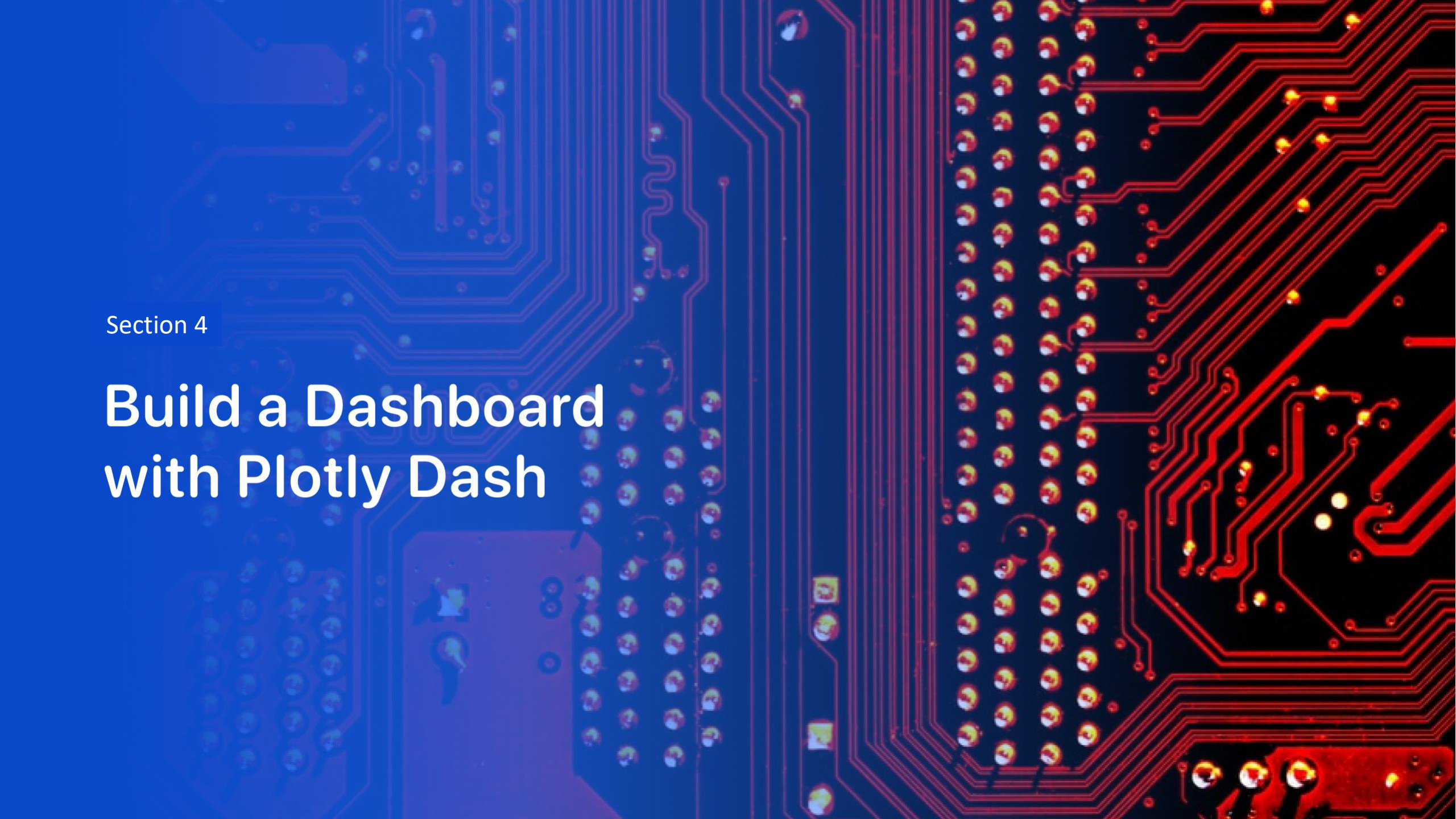
- Launch Site KSC LC-39A has the highest success rate of launches
- Markers:
  - Green: Successful Launches
  - Red: Failed Launches

Site KSC LC-39A closest landmarks proximity

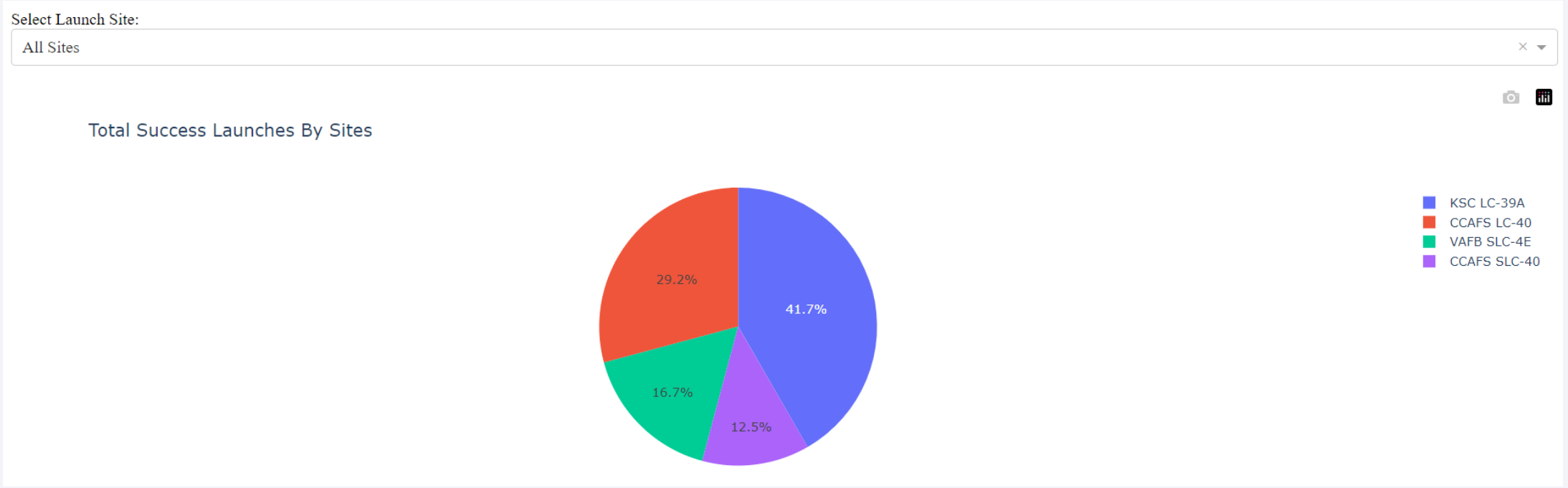


- Closest Landmarks Distances:
  - Coastline: 0.91 km
  - Railway: 0.72 km
  - Highway: 0.84 km

- Launch Site KSC LC-39A closest city is Titusville. 16.28 km away from launch site.

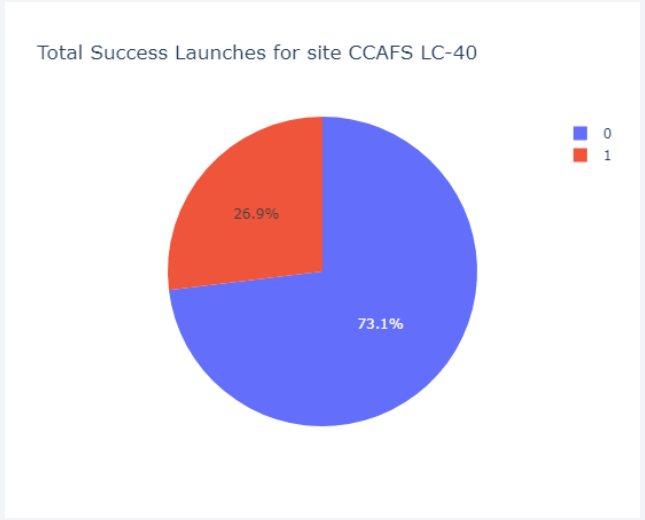

Section 4
Build a Dashboard with Plotly Dash

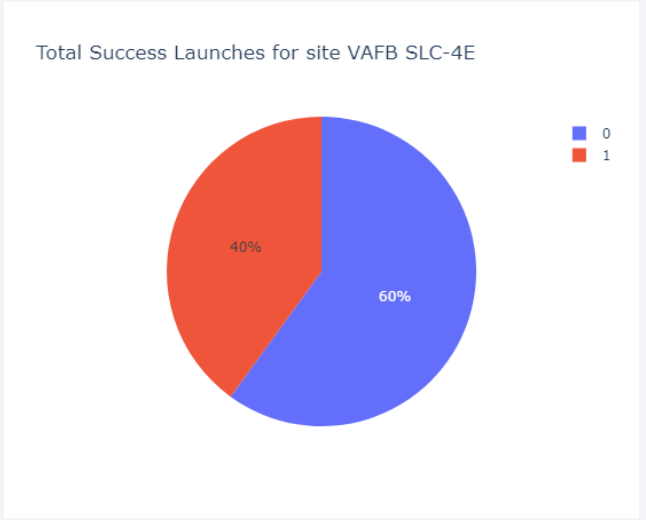
Total Successful Launches for All Sites

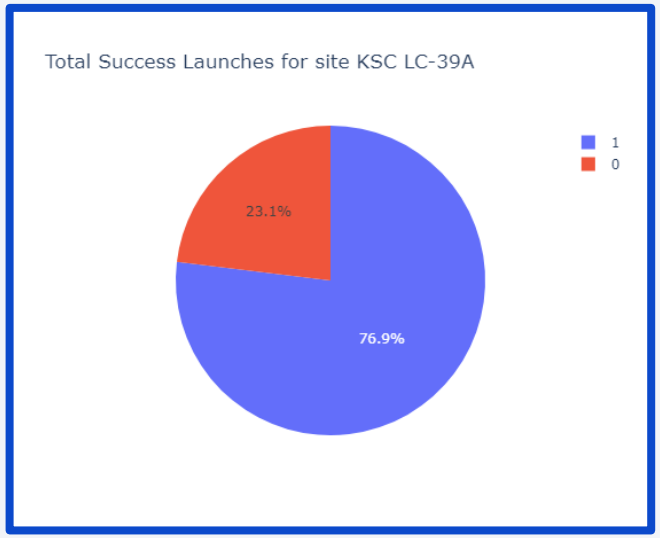


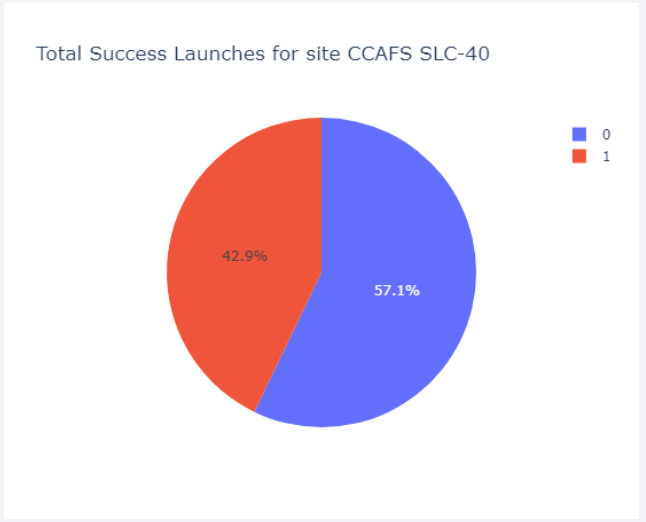
- Highest percentage of successful launches was done at KSC LC 39A

Successful Launches By Each Sites



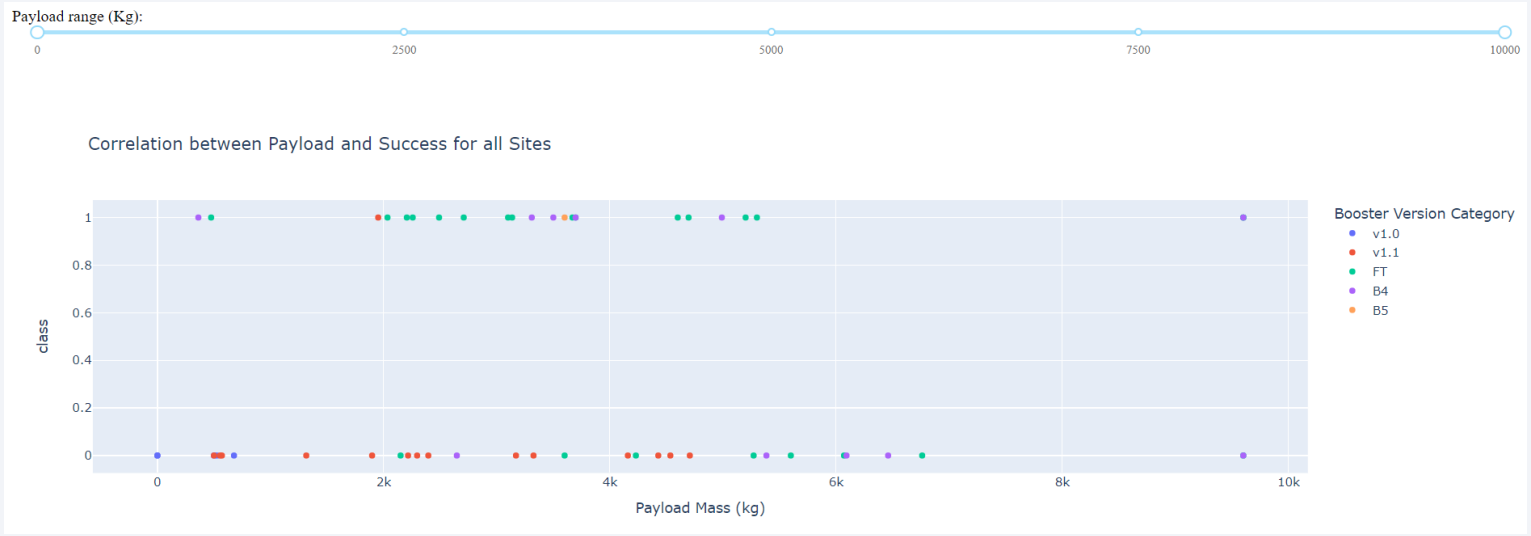


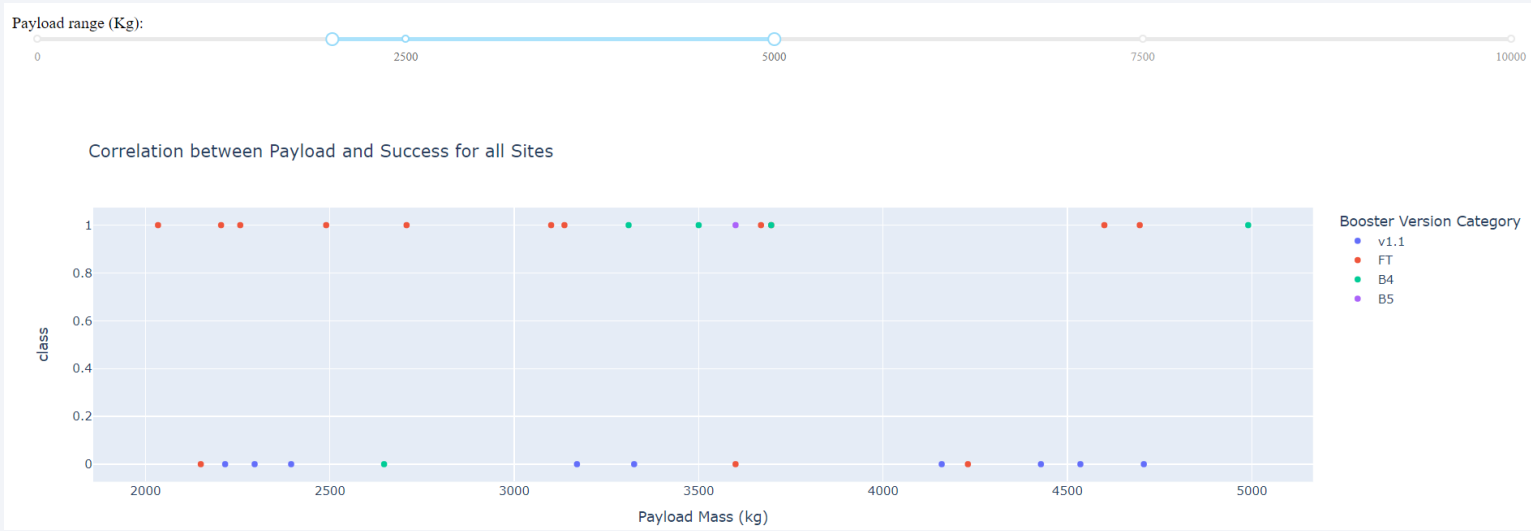




- Launch Site KSC LC-39A has the highest successful launch rate ratio or 76.9%
- Followed by launch site CCAFS LC-40 with a ratio of 73.1%. A difference of 3.8%

Total Successful Launches for All Sites





- There are more successes of launches for payload mass of between 2,000 kg to 5,000 kg
- In that range, booster version category FT has a higher success rate of launches

Section 5

Predictive Analysis (Classification)

Classification Accuracy

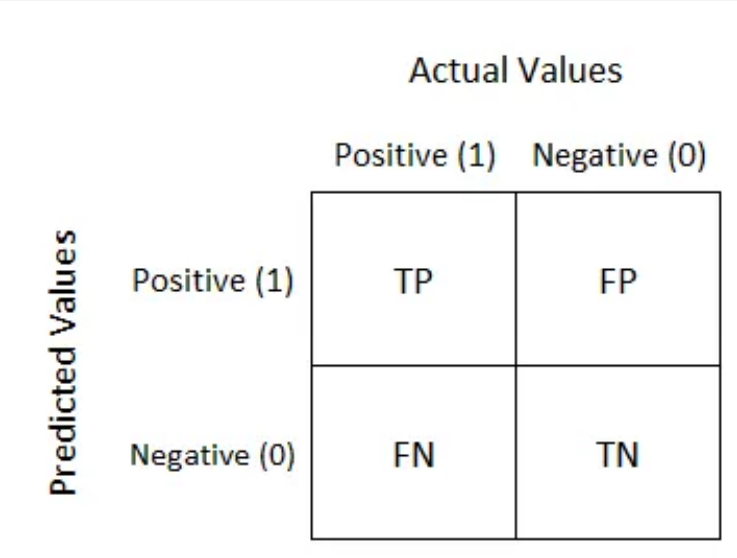
- The Tree Model has the best training score. However, its test score is the lowest amongst the rest.

```
The best training model is:  Tree
The best test model is:  LR
```

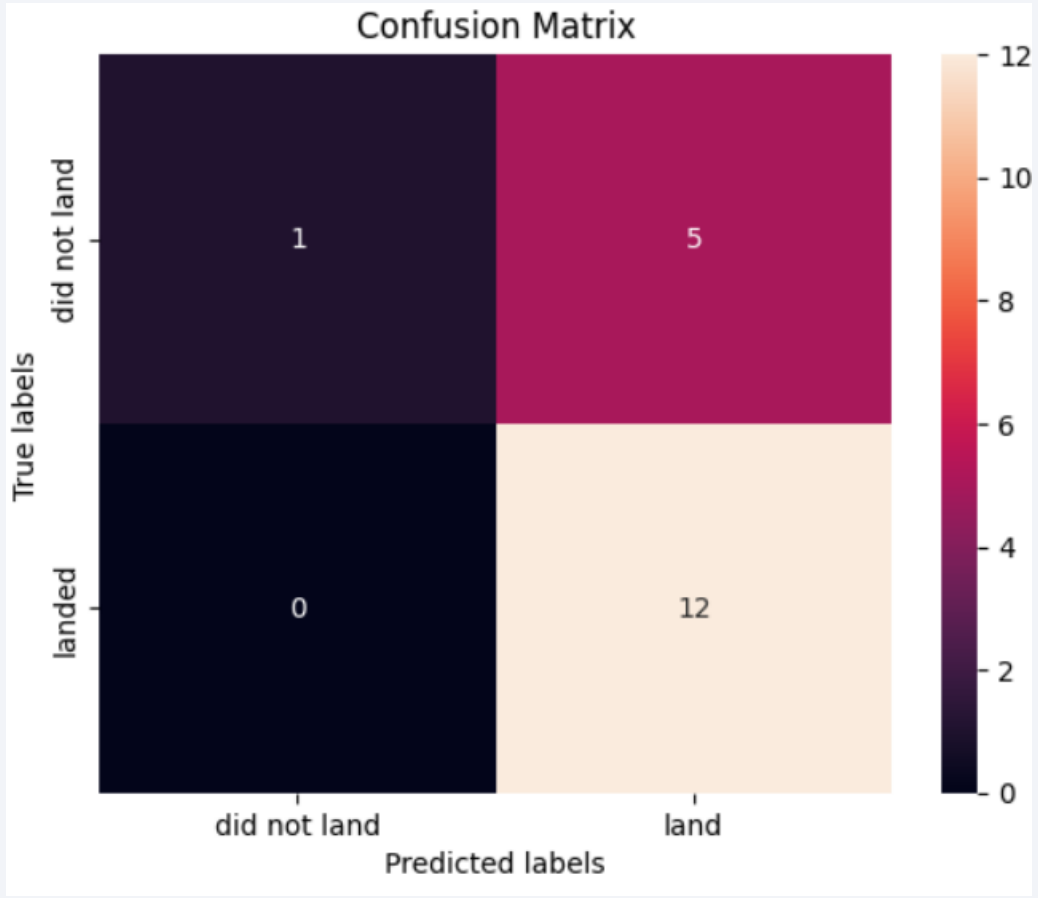
	Model	Training Best Score	Test Score
0	LR	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Tree	0.875000	0.722222
3	KNN	0.848214	0.833333

Confusion Matrix – Tree Model

- Confusion Matrix of the Tree Model
- A high value of 5 of False Positive which is a little bit alarming for a test set of 18 values

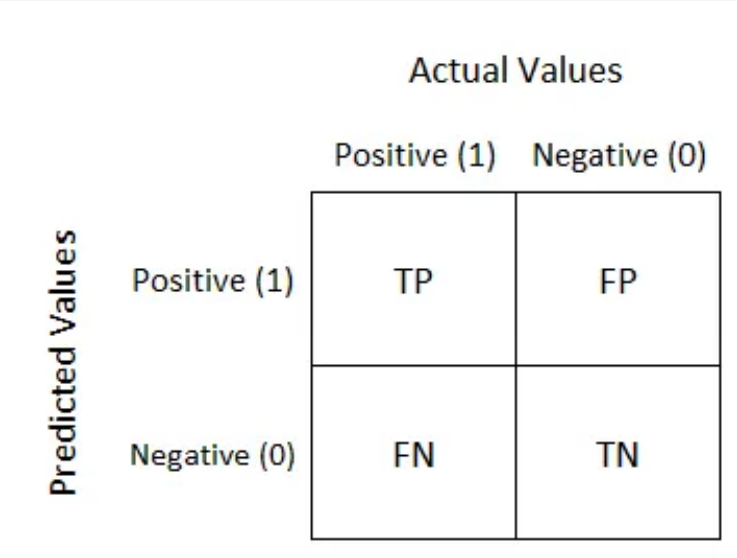


Confusion Matrix Explained [2]

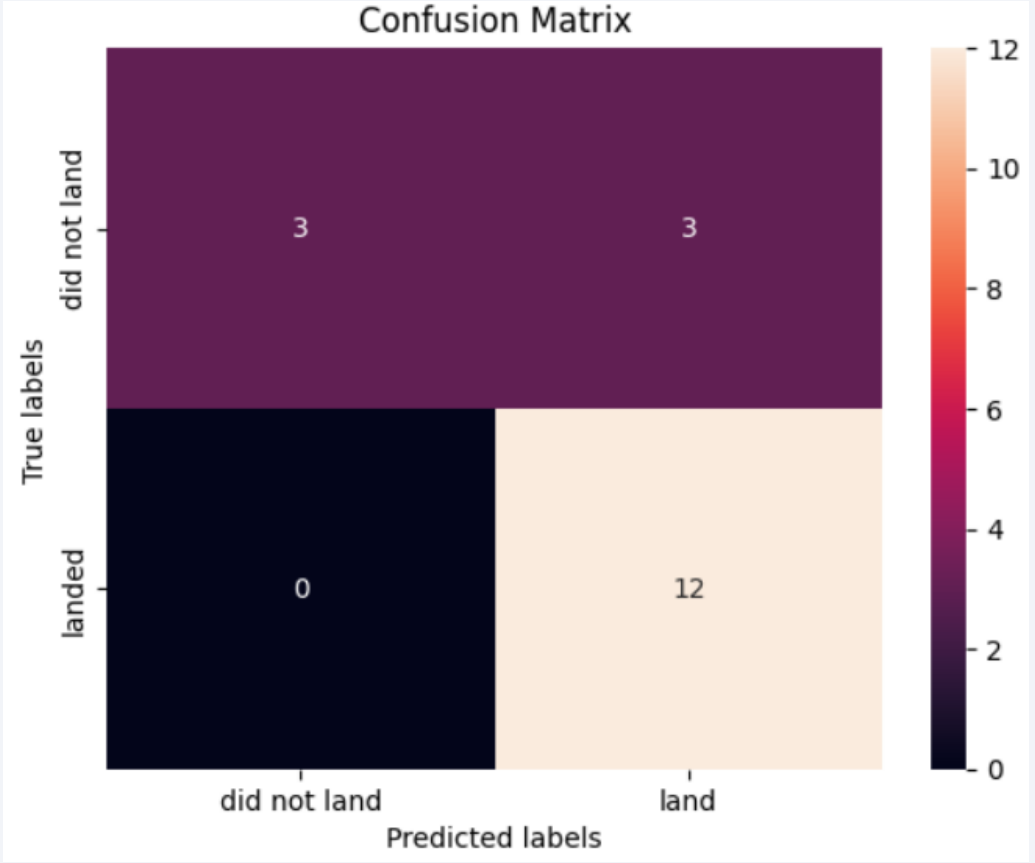


Confusion Matrix – Other Models

- Confusion Matrix of Linear Regression, SVM and KNN
- A value of 3 False Positive which is better than the Tree Model



Confusion Matrix Explained [2]



Conclusions

- Launch site KSC LC 39A has the highest success rate of launches
  - Able to carry a huge range of Payload Mass
  - Very good records or 100% successful launches for Payload Mass below 6,000kg
- Orbit Types ES-L1 GEO, HEO, SSO has a 100% success rate of launches
- Recommendations of Orbit Types based on Payload Mass:
  - Heavy Payload Mass: LEO, ISS and PO
  - Lighter Payload Mass: SSO
- Overall, success rate of launches increases over the year
- Despite the Tree Model having the best training model, it has a high value of False Positive.
- SVM or KNN is a better model for this data set as it has the 2nd highest training score, highest test score and better False Positive value

Appendix

References:

[1] NASA, “Chapter 14: Launch”, 2024. [Online] Available: https://science.nasa.gov/learn/basics-of-space-flight/chapter14-1/. [Accessed: 22 April 2024]

[2] S. Narkhede, “Understanding Confusion Matrix”, 2018. [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62. [Accessed: 22 April 2024]

Special Thanks To:

Coursera

IBM

Instructors

Thank you!