Winning Space Race with Data Science

Eric Agyei
06/24/2022

Outline

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

Executive Summary

- Summary of methodologies

The main aim of this project was to use the best data science tools and practices to predict whether the SpaceX Falcon 9 first stage will land successfully through selected features. This is to aid us to reuse the first stage so as to cut costs in the future.

- In order to do this, we would be going through the following:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis I
  - Interactive data visualization
  - Machine learning prediction
- It was revealed at the end of the project that some features have a correlation with the outcome as to being successful or failure whiles others had no correlation. It finally, concluded that decision tree was the best machine learning algorithm to be used to predict if the first stage will land successfully.

Introduction

- Project background and context

The commercial space age is here and companies are making space travel affordable for everyone. The most successful is SpaceX and one reason for this is the relatively inexpensive of the rocket launches. Whiles other providers cost upwards of 165 million dollars each, SpaceX advertises its Falcon 9 rocket launches on its website with a cost of 62 million dollars. Much of the savings is because SpaceX can reuse the first stage.

This therefore brings us to the problem, that is, to determine whether the first stage will land so as to determine the cost of a launch.

For us to solve this problem we would be analyzing features about Falcon 9 rocket launch such as its payload mass, orbit type, launch site, amongst others using the best data science tools.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through SpaceX API and Web Scraping with Python Beautiful Soup
- Perform data wrangling
  - Data wrangling was done with Pandas, Numpy and SQL to filter the data and deal with missing values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - These classification models included Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbors

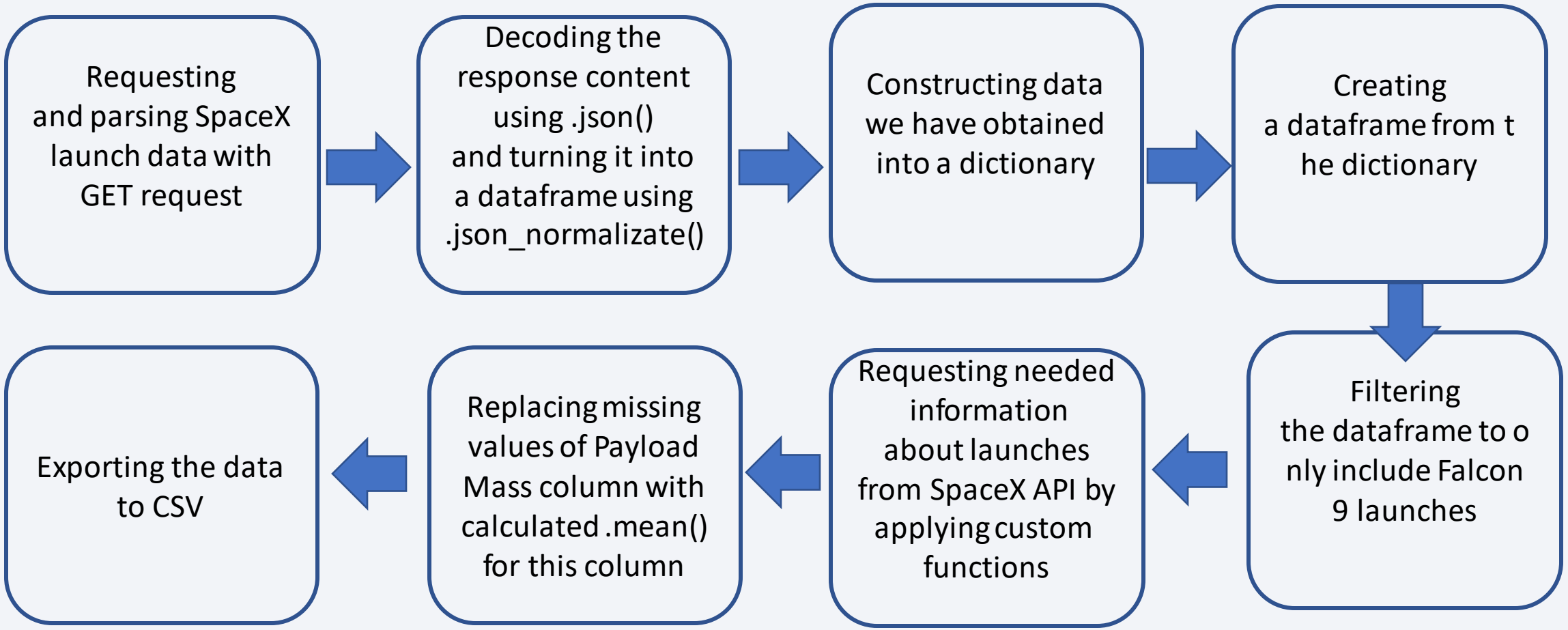
Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. Both data collection methods were used so as to get complete information about the launches.

Data columns such as Flight Number, Date, Booster Version, Payloads Mass, Orbit, Launch Site, Outcome, Flights, Serial, Longitude and Latitude were obtained using SpaceX REST API.

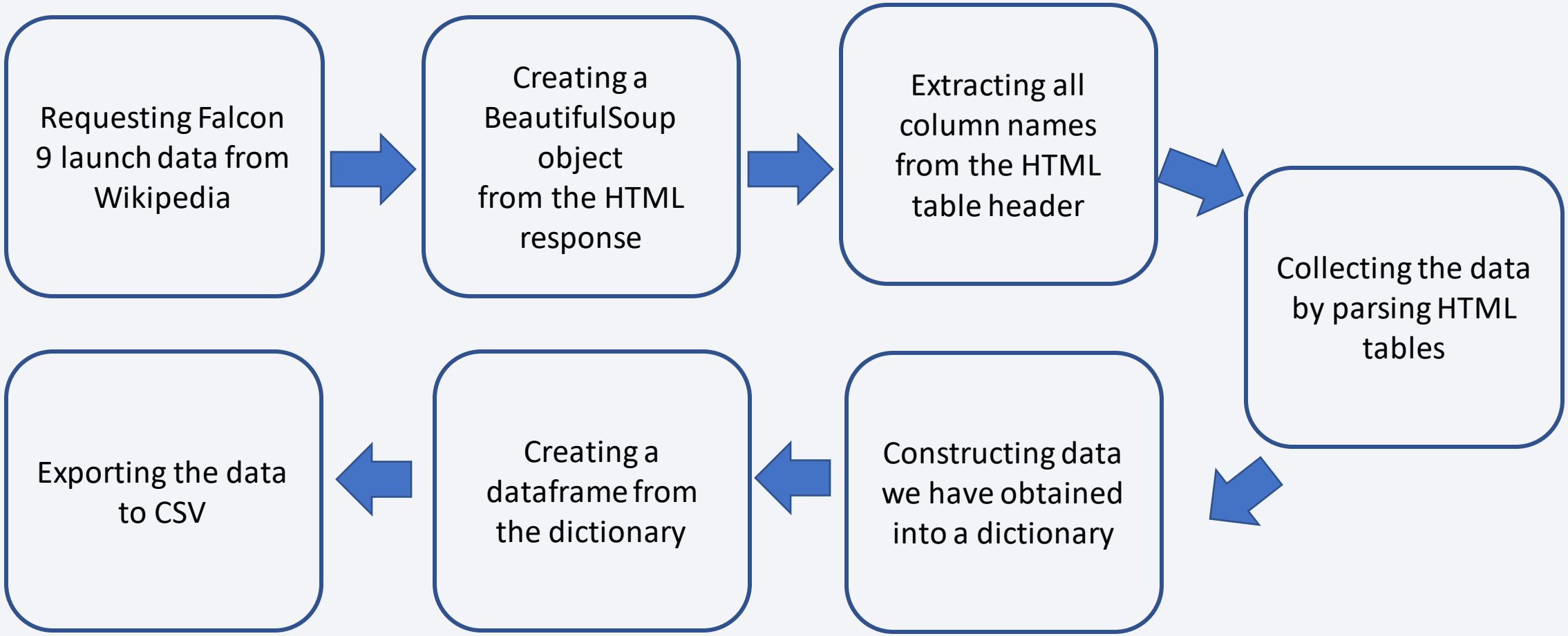
Web Scraping was also used to obtain same information from Wikipedia

Data Collection – SpaceX API



GitHub Link: Data Collection - SpaceX API

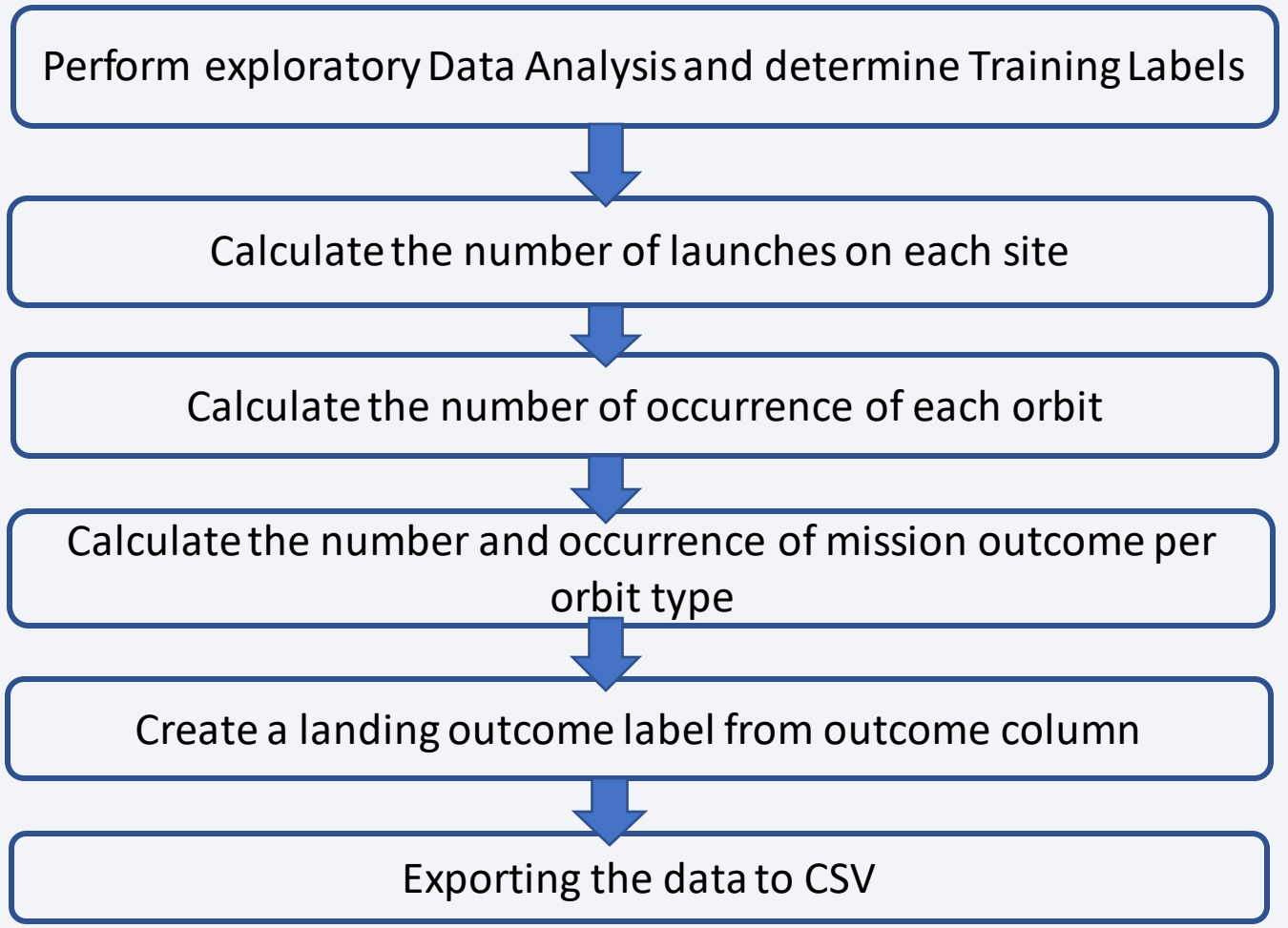
Data Collection - Scraping



GitHub Link: Data Collection - Scraping

Data Wrangling

The data set consisted of different cases where the booster sometimes did not land successfully or attempted but failed due to accident. The data was therefore converted into training and testing binary labels "1" for a successful outcome and "0" for failure or unsuccessful outcome.



GitHub Link: Data Wrangling

EDA with Data Visualization

The following charts were plotted:

- Flight Number vs Payload
- 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
- Success Rate Yearly Trend

Scatter plots show the relationship between variables which is very useful in machine learning model.

Bar charts show comparisons among discrete categories. This helps to show the relationship between specific features being compared.

Line charts show trends in data over time.

GitHub: EDA with Data Visualization

EDA with SQL

- SQL queries performed:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2005
  - Ranking the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

GitHub: EDA with SQL

Build an Interactive Map with Folium

- Markers of all Launch Sites:
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to the coasts.
- Coloured Markers of the launch outcomes for each Launch Site:
  - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
  - Added coloured Lines to show distances between the Launch Sites to Coastline and Railway.

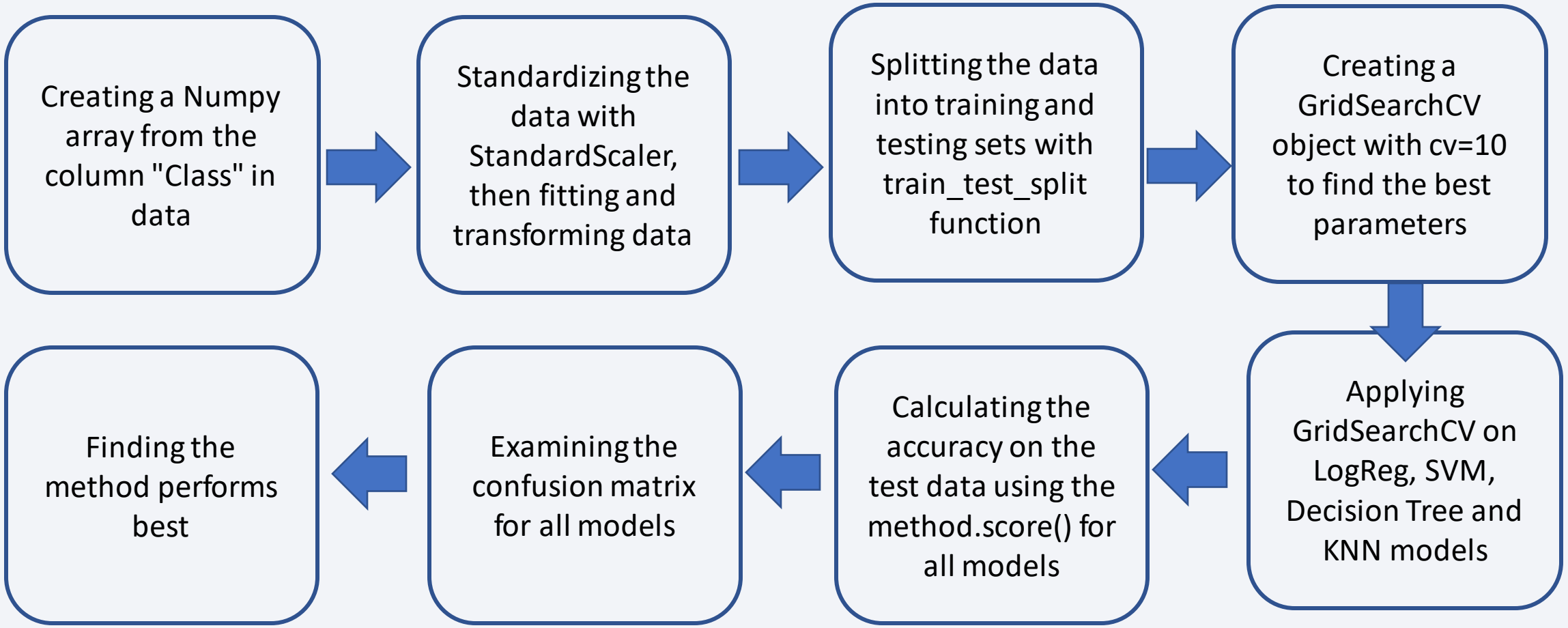
GitHub Link: Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches:
  - Added a pie chart to show the total successful launches count for all sites and the Success vs Failed counts for the site, if a specific Launch Site was selected.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs Success Rate for the different Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub Link: Dashboard with Plotly Dash

Predictive Analysis (Classification)



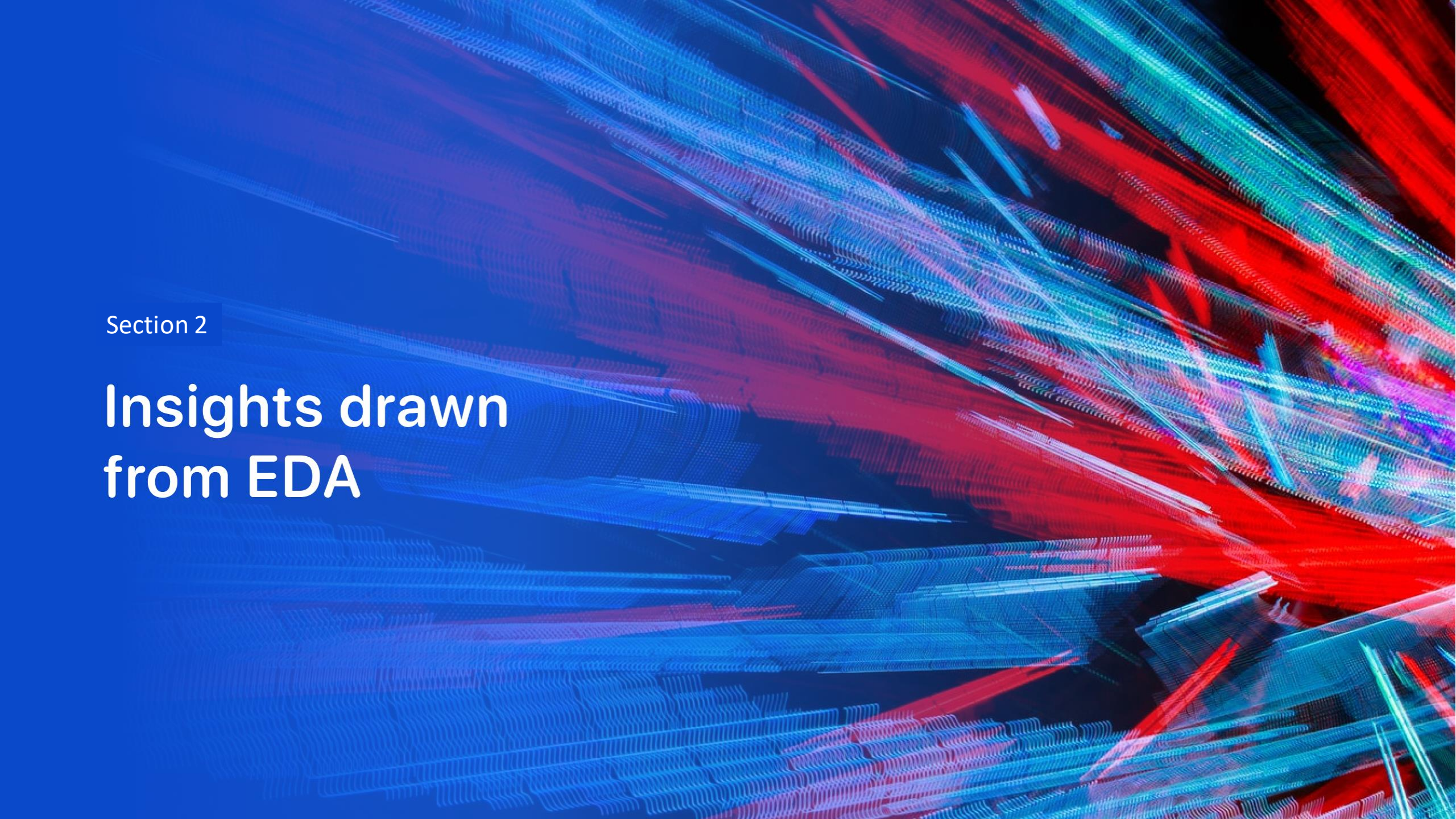
GitHub Link: Predictive Analysis

Results

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

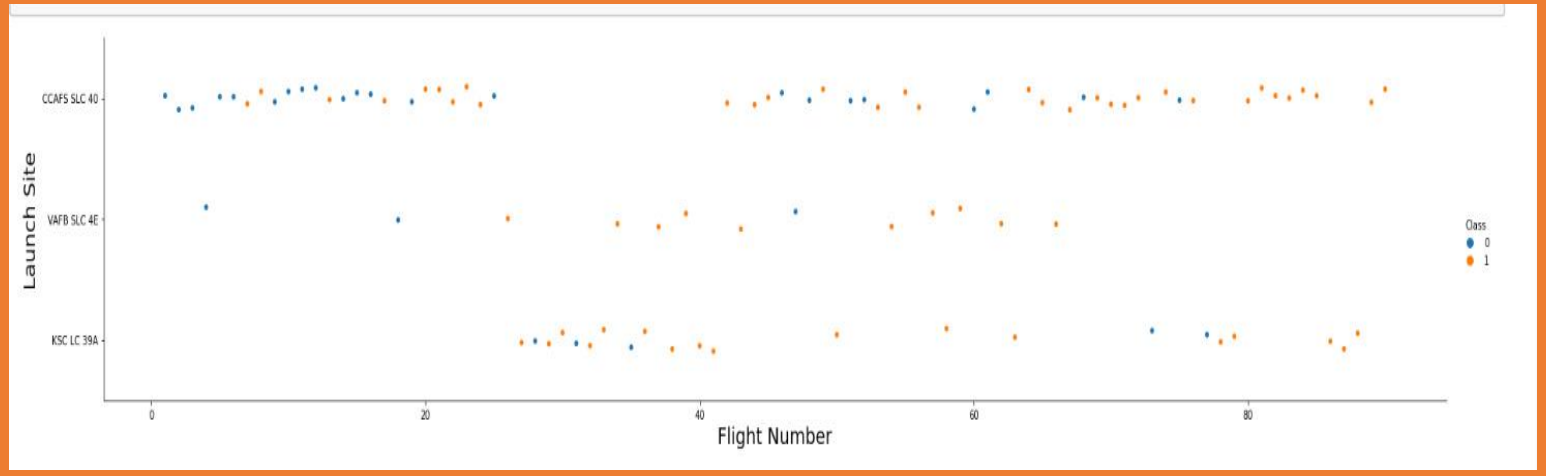
Section 2

Insights drawn from EDA



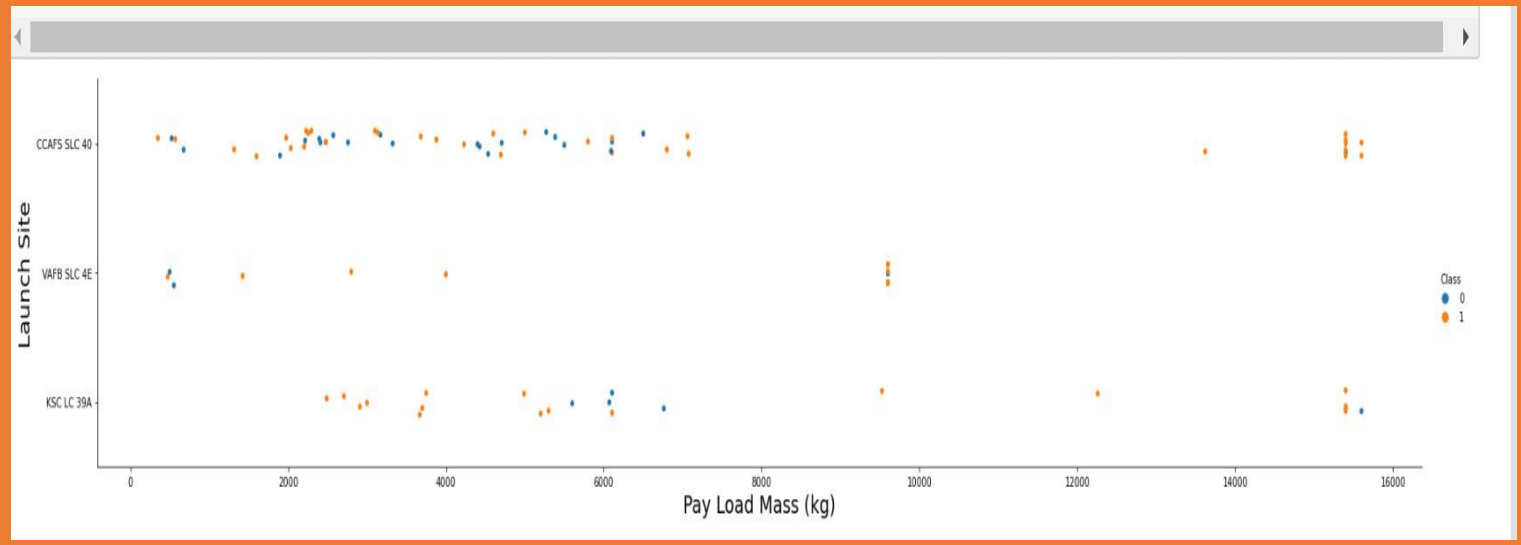
Flight Number vs. Launch Site

- The earlier flights all failed while the later all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.



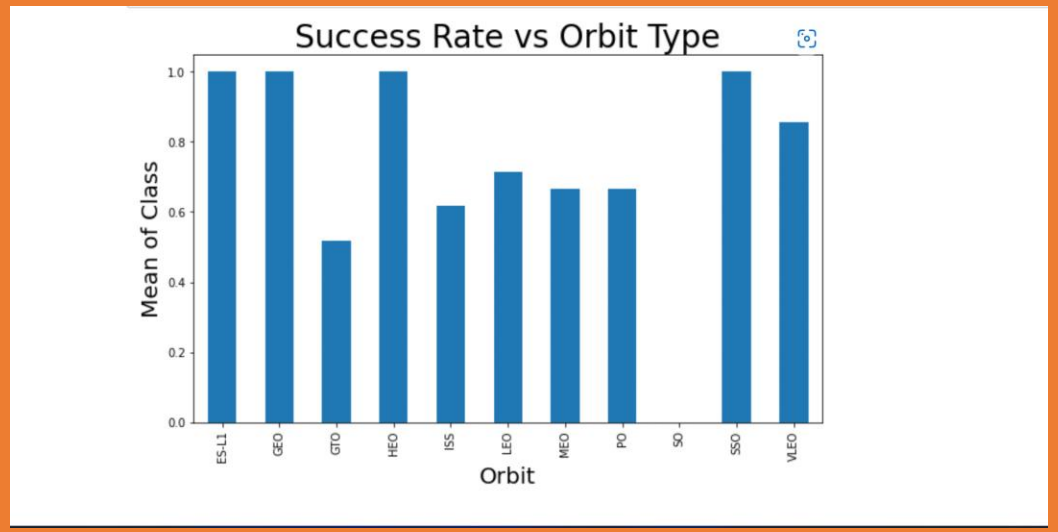
Payload vs. Launch Site

- There are no rockets launched for heavy payload mass greater than 10000 for the VAFB-SLC launch site.



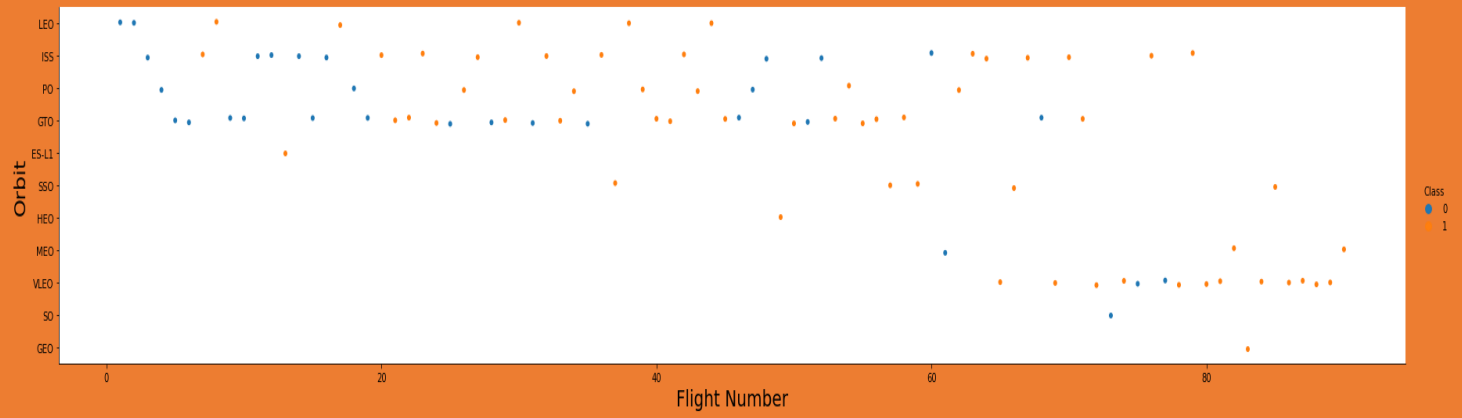
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO orbits had 100% success rate
- SO orbit had 0% success rate.
- GTO, ISS, LEO, MEO, PO orbits had between 50% and 85% success rates.



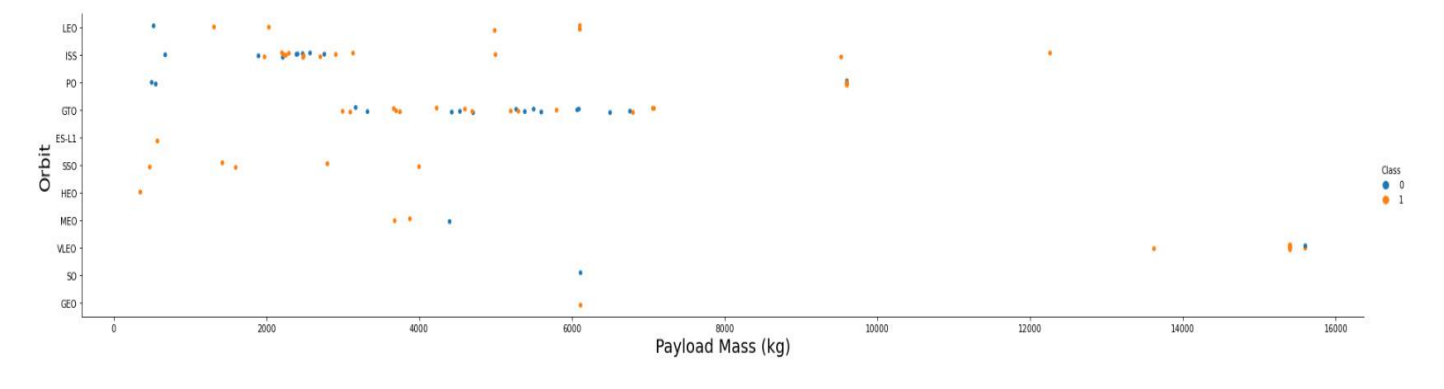
Flight Number vs. Orbit Type

- LEO orbit the Success appears related to the number of flights, on the other hand, there seems to be no relationship between flight number when in GTO orbit.



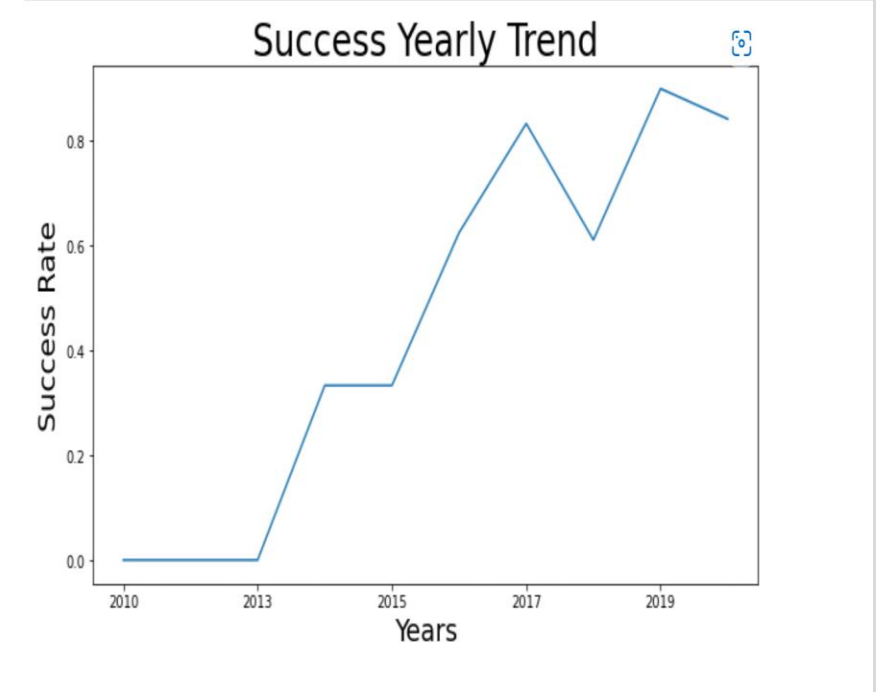
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing are both there.



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2020.



All Launch Site Names

- Displaying the names of the unique launch sites in the space mission.

```
In [7]: %sql Select distinct Launch_Site from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[7]:

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

Launch Site Names Begin with 'CCA'

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

In [8]:
```
%%sql
Select * from SPACEXTBL
where Launch_Site like 'CCA%'
limit 5
```

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

Out[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Displaying the total payload mass carried by booster launched by NASA (CRS)

In [9]:
```
%%sql
Select customer, sum(payload_mass__kg_) as "Total Payload Mass"
from (Select Customer, payload_mass__kg_ from SPACEXTBL where Customer like 'NASA (CRS)')
Group by Customer
```

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

Out[9]:

Customer	Total Payload Mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

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

In [10]:
```
%%sql
Select Booster_Version, avg(payload_mass__kg_) as "Average Payload Mass"
from (Select Booster_Version, payload_mass__kg_ from SPACEXTBL where Booster_Version like 'F9 v1.1')
Group by Booster_Version
```

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

Out[10]:

Booster_Version	Average Payload Mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- Listing the date when the first successful landing outcome in ground pad was achieved.

In [43]:
```
%%sql
select min(substr(Date, 7, 4)||"-"||substr(Date, 4, 2)||"-"||substr(Date, 1, 2)) as "First Success"
from SPACEXTBL
where "Landing _Outcome" = "Success (ground pad)"
```

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

Out[43]:

First Success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [44]:
```
%%sql
Select Booster_Version, payload_mass__kg_, "landing _outcome"
from SPACEXTBL
where 4000 < payload_mass__kg_ and payload_mass__kg_ < 6000 and "landing _outcome" = "Success (drone ship)"
```

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

Out[44]:

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

Total Number of Successful and Failure Mission Outcomes

- Listing the total number of successful and failure mission outcomes.

In [58]:
```
%%sql
Select mission_outcome, count(mission_outcome) as "Total"
from SPACEXTBL
Group by trim(mission_outcome)
```

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

Out[58]:

Mission_Outcome	Total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass.

In [59]:
```
%%sql
Select distinct booster_version, payload_mass__kg_
from SPACEXTBL
where payload_mass__kg_ = (Select max(payload_mass__kg_) from SPACEXTBL)
```

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

Out[59]:

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

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

In [71]:
```
%%sql
Select substr(date, 4, 2) as "Months", "landing _outcome", booster_version, launch_site
from SPACEXTBL
where "landing _outcome" = "Failure (drone ship)" and substr(Date, 7, 4) = "2015"
```

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

Out[71]:

Months	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

In [20]:
```
%%sql
Select "landing _outcome", count(*) as count_outcomes
from SPACEXTBL
where substr(Date, 7, 4)||"-"||substr(Date, 4, 2)||"-"||substr(Date, 1, 2) between "2010-06-04" and "2017-03-20"
group by "landing _outcome"
order by count_outcomes desc
```

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

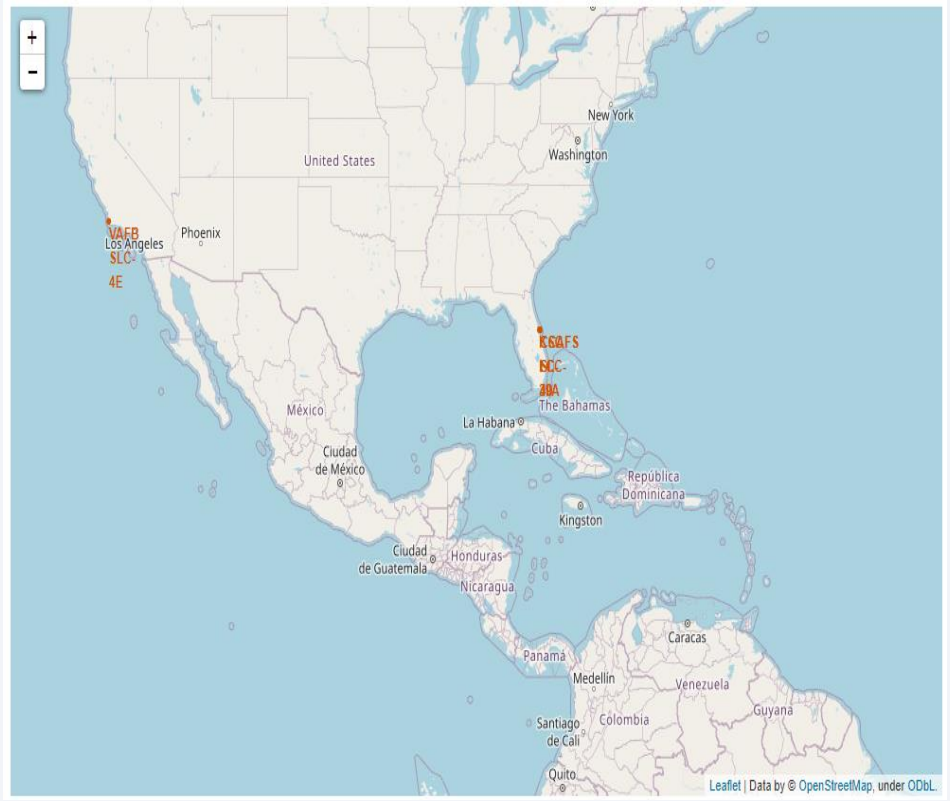
Out[20]:

Landing _Outcome	count_outcomes
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


Section 3
Launch Sites
Proximities Analysis

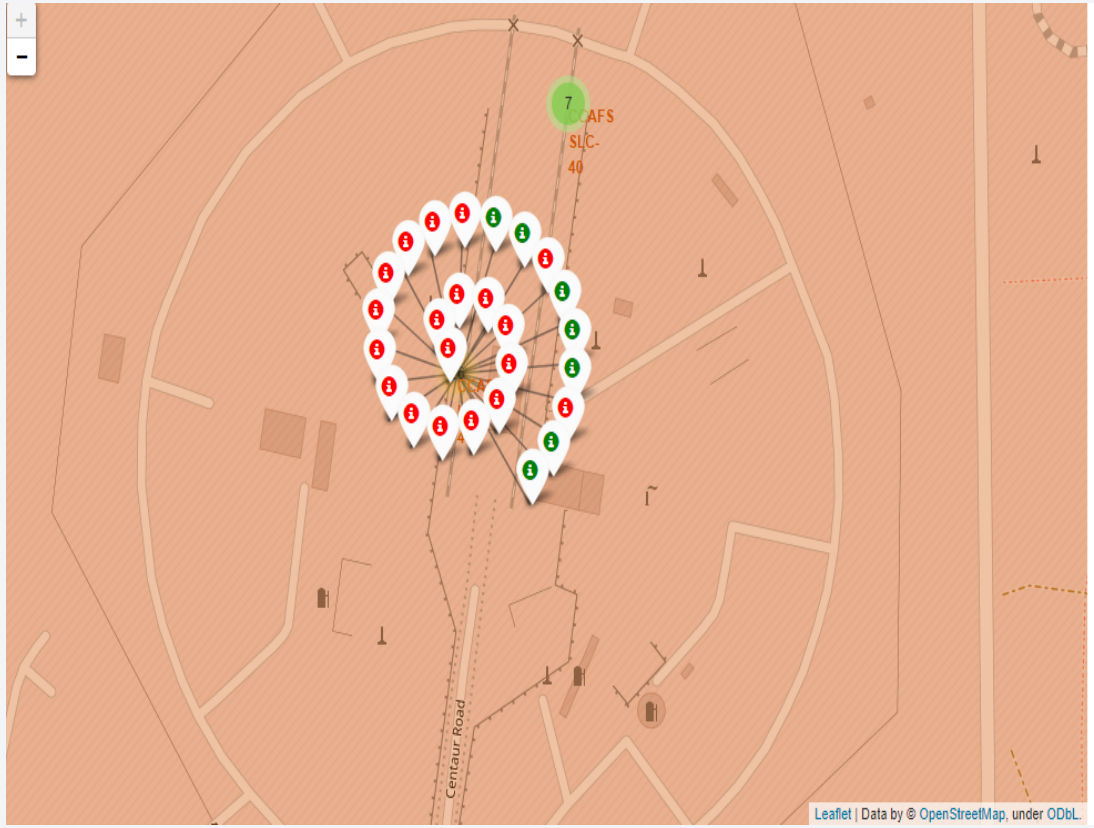
All Launch Sites

- Most of launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. If a ship is launched from the Equator it goes up into space, and it is also moves around the Earth at the same speed it was moving before launch. This speed helps spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast. This minimizes the risk of having any debris dropping or exploding near the population.



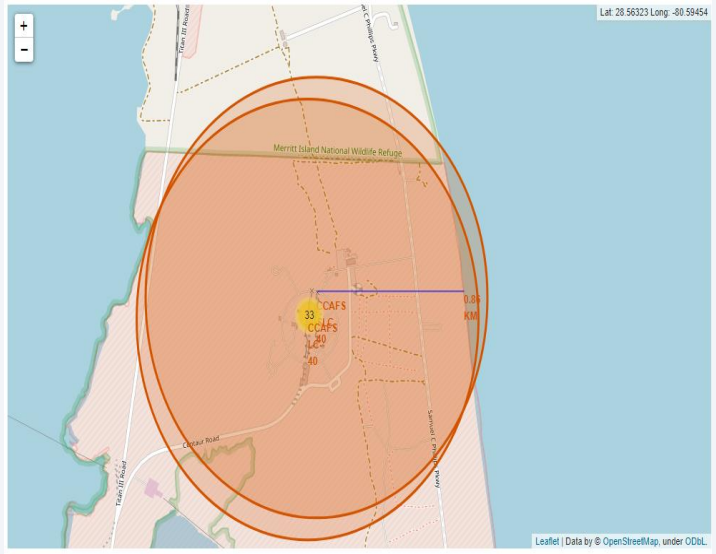
Colour-Labeled Launch Outcomes

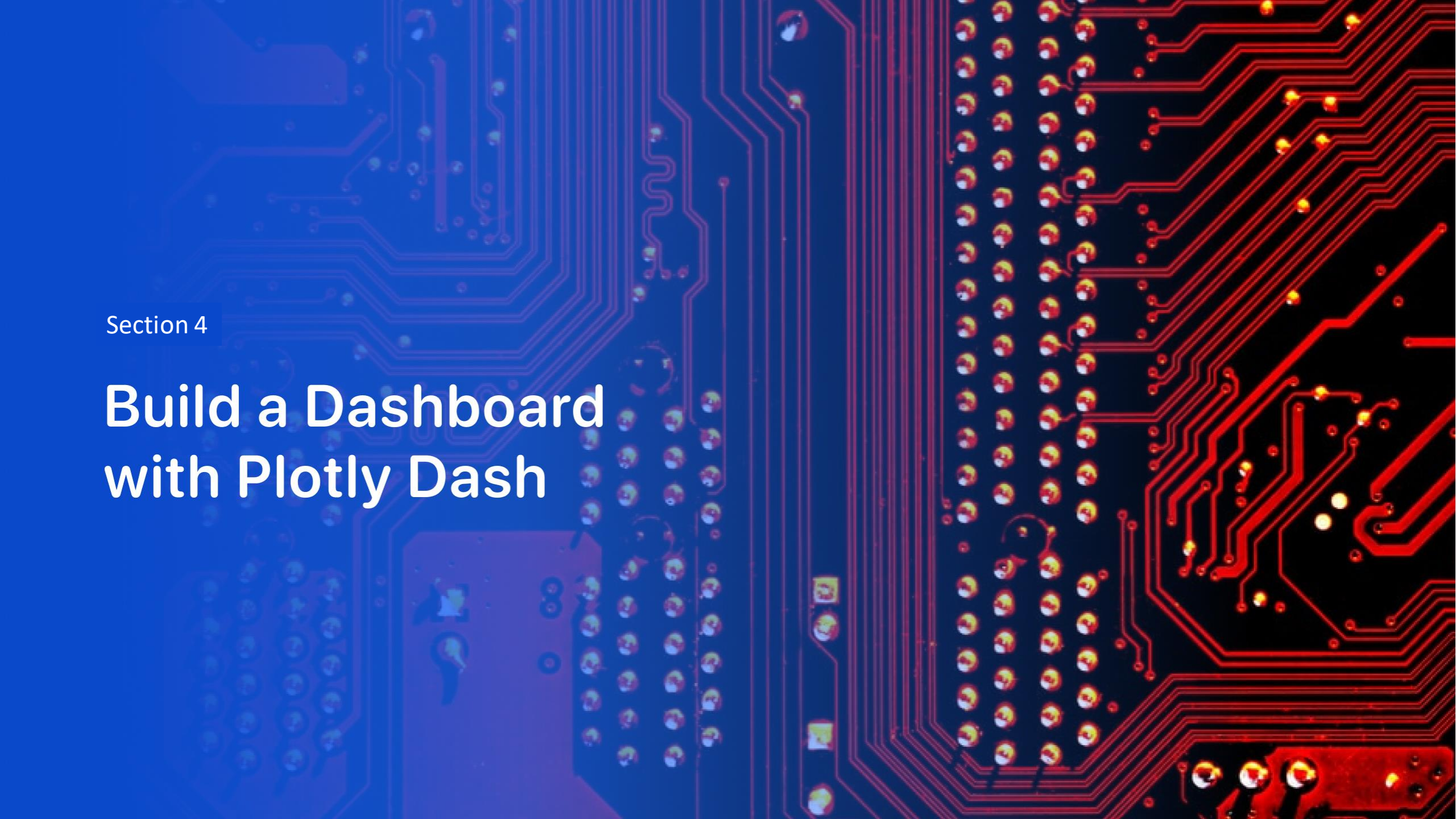
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
- Green Marker = Successful Launch
- Red Marker = Failed Launch



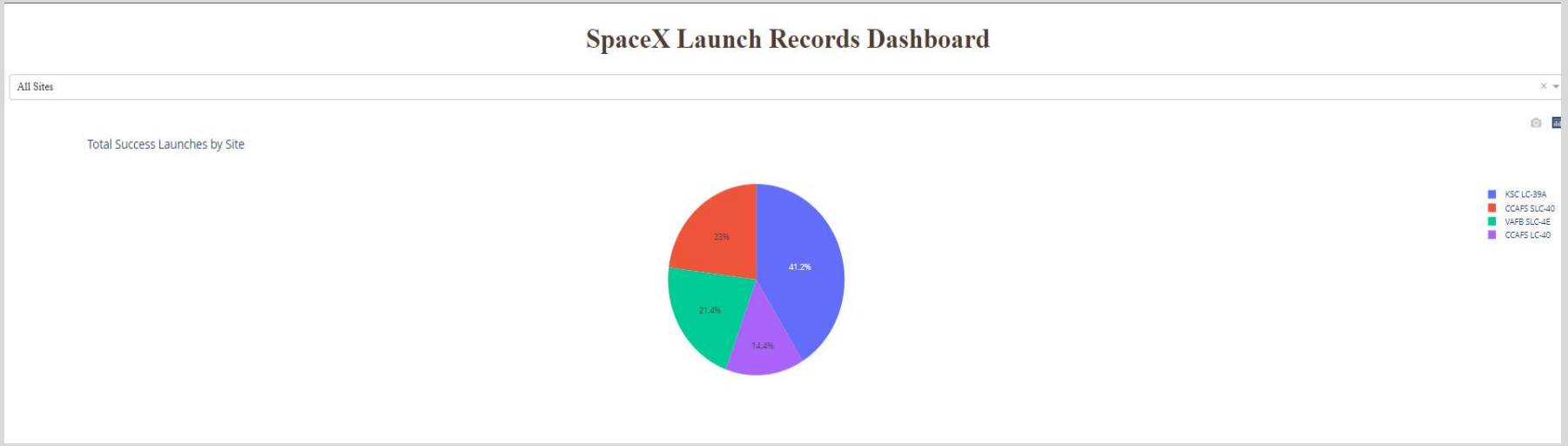
Distance A Launch Site

- The distance between a launch site, CCAFS LC-40 and the nearest coastline.



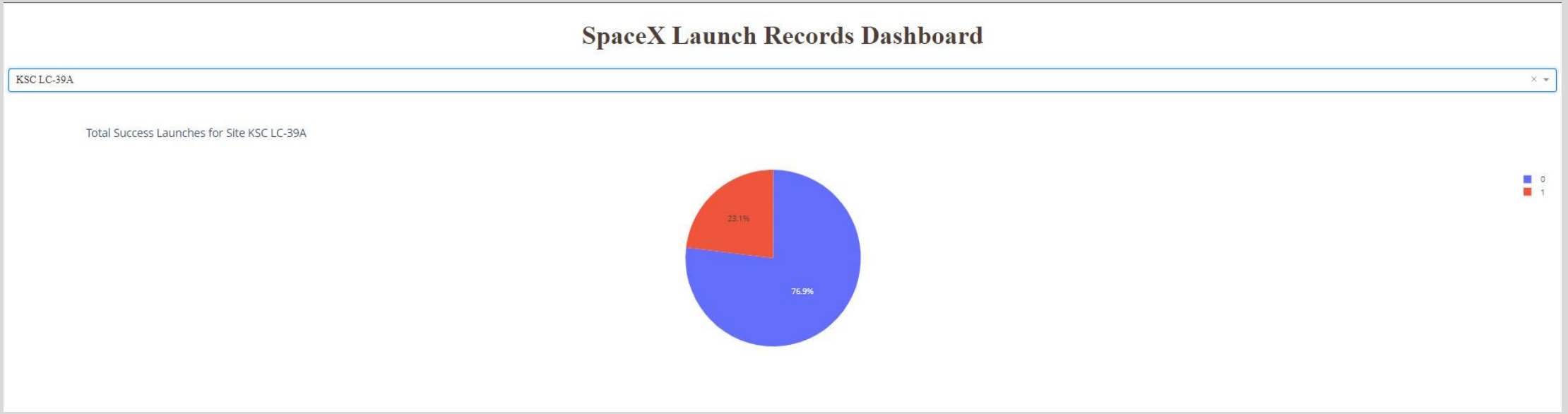

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



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

Launch Site With Highest Launch Success Ratio



- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only failed landings.

Payload Mass vs. Launch Outcome for All Sites

- Payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Based on the accuracy scores of the Test Set, we cannot conform which method performed best.
- However, the best scores of on the Training Set showed that Decision Tree has the best score.

Accuracy scores on Test Sets

Out[37]:

	Accuracy Scores
Logistic regresssion	0.833333
SVM	0.833333
KNN	0.833333
Decision tree	0.666667

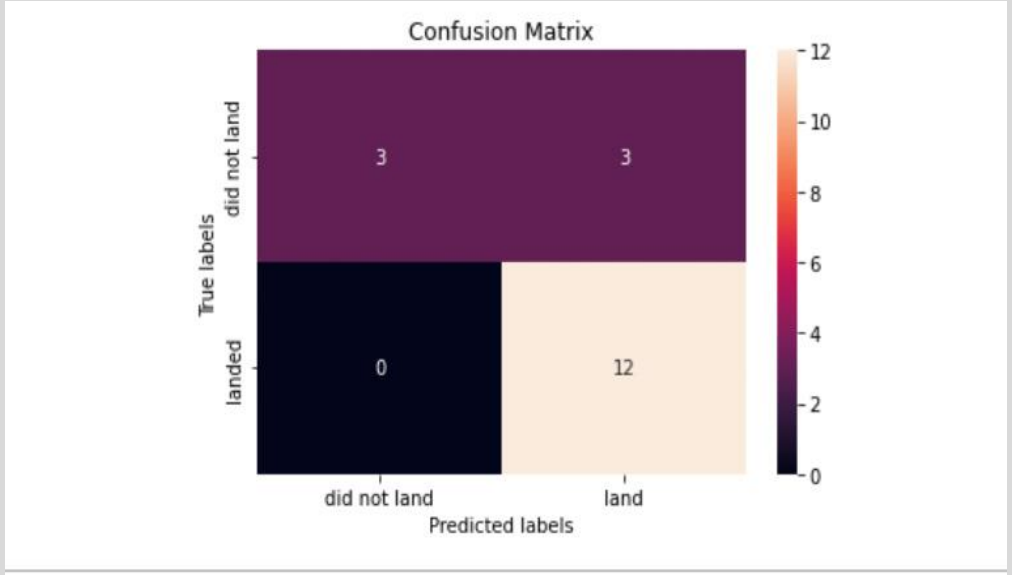
Accuracy scores on Training Sets

Out[38]:

	Best Scores
Decision tree	0.875000
KNN	0.848214
SVM	0.848214
Logistic regresssion	0.846429

Confusion Matrix

- Examining the confusion matrix, we see that almost all the models could distinguish between the different classes. The major problems is false positives.



Conclusions

- In conclusion we could agree with the following
  - Decision Tree Model is the best algorithm for this dataset.
  - Launches with a low payload mass show better results than launches with a larger payload mass.
  - Most launch sites are in proximity to the Equator line and all the sites are very close in proximity to the coast.
  - The success rate of launches increases over the years.
  - KSC LC-39A has the highest success rate of the launches from all the sites.
  - Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

Special Thanks to:

The IBM instructors

Coursera

Colleagues of IBM Data Science Professional Course

Thank you!