IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ashani Kallichurum

2021/11/29

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- EDA with Data Visualization
- EDA with SQL
- Building an Interactive Map using Folium
- Building a Dashboard using Plotly Dash
- Machine Learning(Predictive Analysis)

Summary of all results

We gained insight into the relationship between Weight of the Payload and the Success rate of the Launch, which Booster has the highest success rate, as well as the most accurate Classification Model to be used.

Introduction

Project background and context

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, therefore determining the success rate of the first stage plays and important factor to cost.

Problems you want to find answers

In order to determine the Cost of the Launches, its is relevant to determine the relationship between the various variables, and the success rate of the Launches.

This will assist in selecting the correct conditions to conduct these Launches successfully, whilst maintaining a competitive edge on the Cost.

Section 1

Methodology

Methodology

Executive Summary

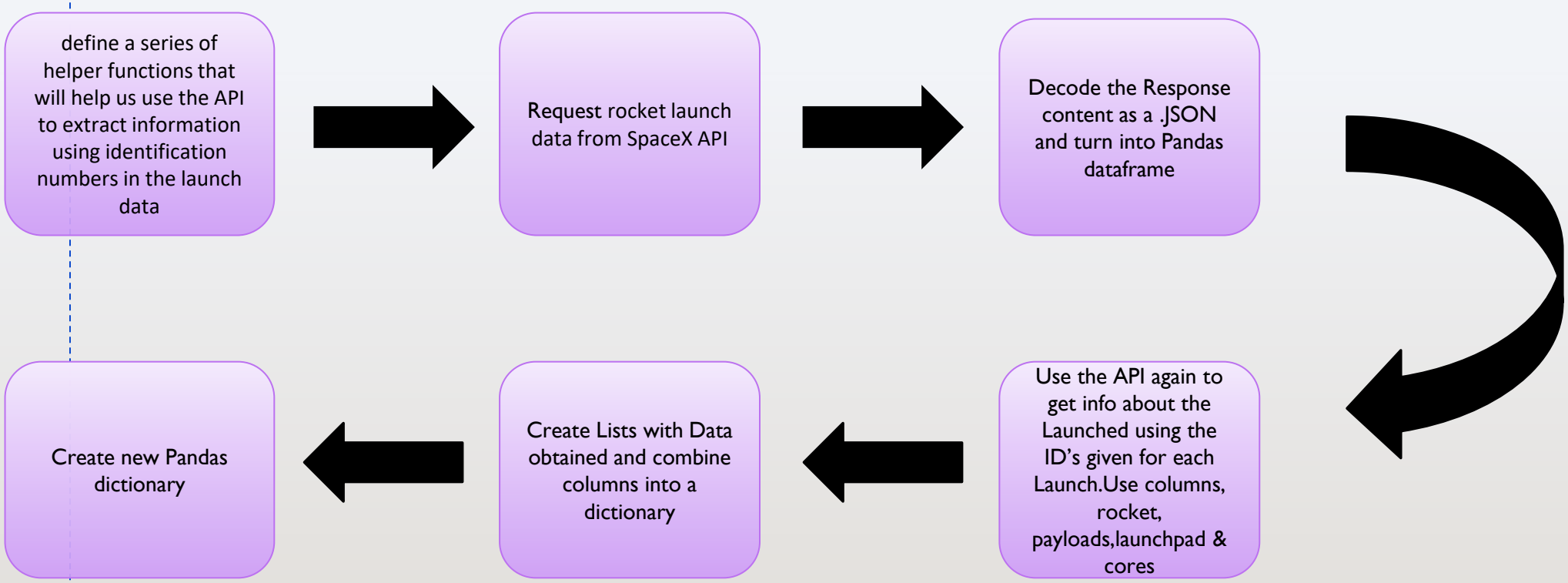
- Data collection methodology:
  - Obtained data using an API, specifically the SpaceX REST API.
  - Webscraped Spacex Wikipedia page
- Perform data wrangling
  - We performed Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models
- Perform 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

Collection of Data:

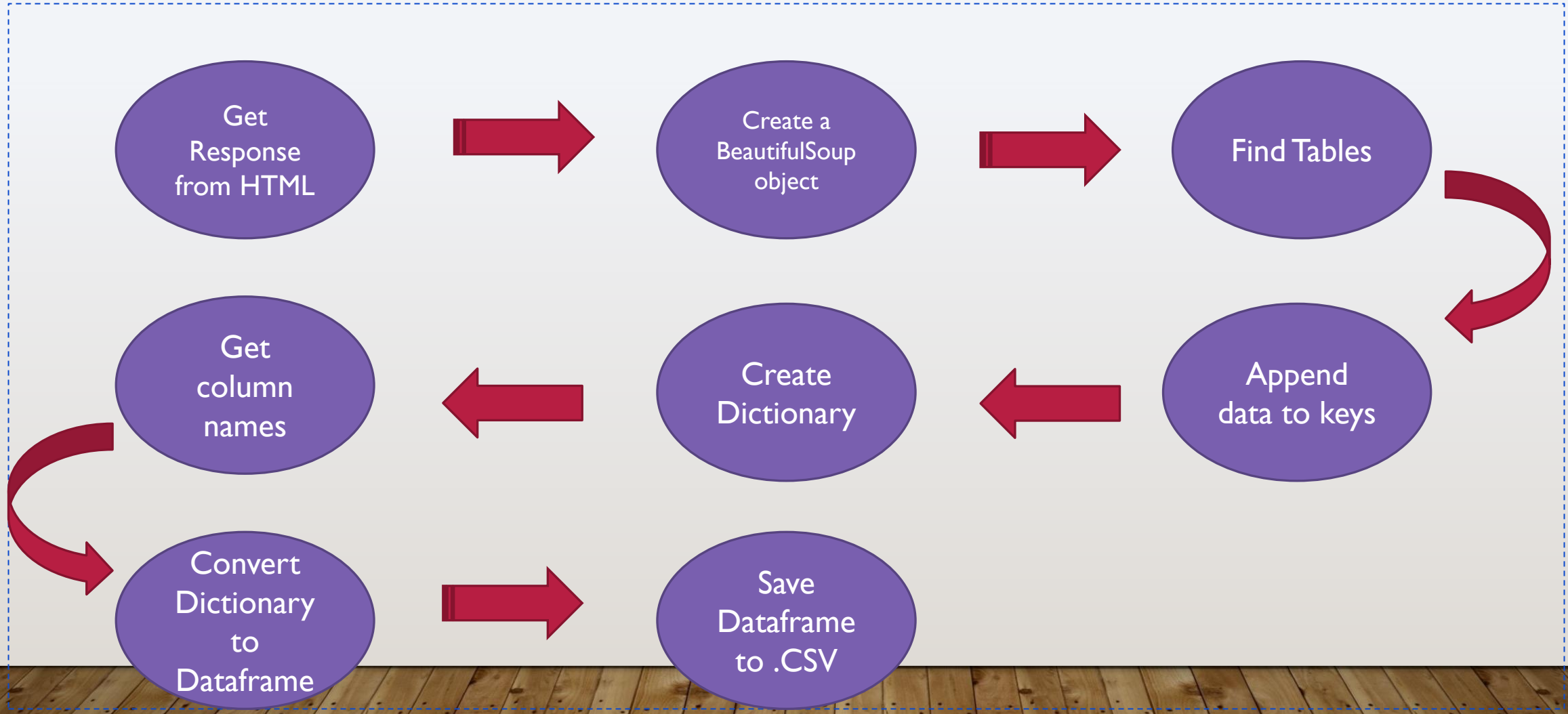
- We used SpaceX launch data that is gathered from an API, specifically the SpaceX REST API
- This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.
- The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.

Data Collection – SpaceX API



Data Collection Github

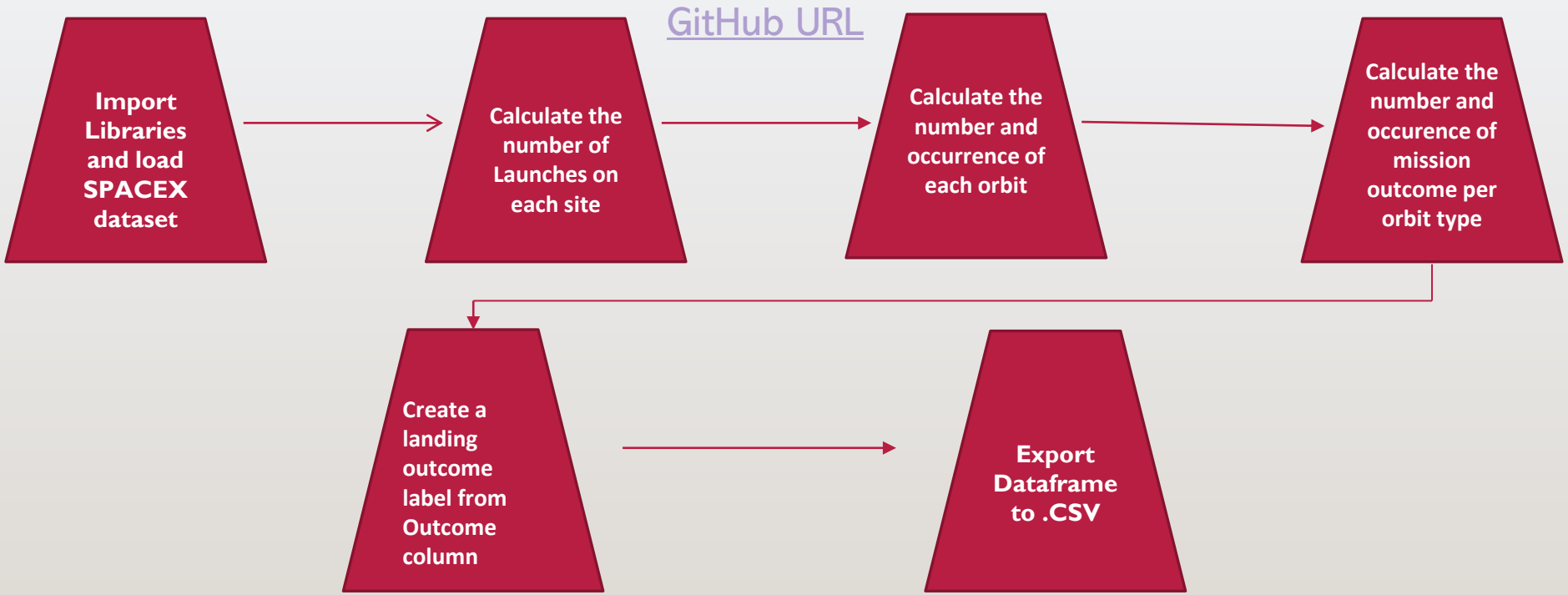
Data Collection - Scraping



GitHub URL

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.
In this lab we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.



EDA with Data Visualization

Bar Chart

A bar chart is used when you want to show a distribution of data points or perform a comparison of metric values across different subgroups of your data. From a bar chart, we can see which groups are highest or most common, and how other groups compare against the others.

We used the Bar Chart to **Visualize the relationship between success rate of each orbit type.**

Scatter Plot

A scatter diagram is used **to show the relationship between two kinds of data**. It could be the relationship between a cause and an effect, between one cause and another, or even between one cause and two others.

We used the Scatter plot to visualize:

Flight Number vs Payload Mass

Flight Number vs Launch Site

Payload vs Launch Site

Flight number vs Orbit type

Payload vs Orbit Type

Line Chart

A line chart is a type of chart used to show information that changes over time.

Line charts are created by plotting a series of several points and connecting them with a straight line. Line charts are used to track changes over short and long periods of time.

We used the Line Chart to **Visualize the launch success yearly trend.**

GitHub URL

EDA with SQL

12

Questions resolved using SQL Queries :

- 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 using a subquery
- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for 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 URL

Build an Interactive Map with Folium

13

- We added a *Circle Marker* around each Launch site, with a label of the name of the Launch site, using the Longitudes and Latitudes coordinates of the Launch sites.
- Using Green and Red *Markers* on the Map in a MarkerCluster(),we assigned the dataframe launch_outcomes, to classes, 0 and 1.
- *Polyline* was drawn on the map to measure distance to landmarks
- *Mouse Position* to get coordinates of a point when you hover with mouse.

GitHub URL

Dashboard with Plotly Dash

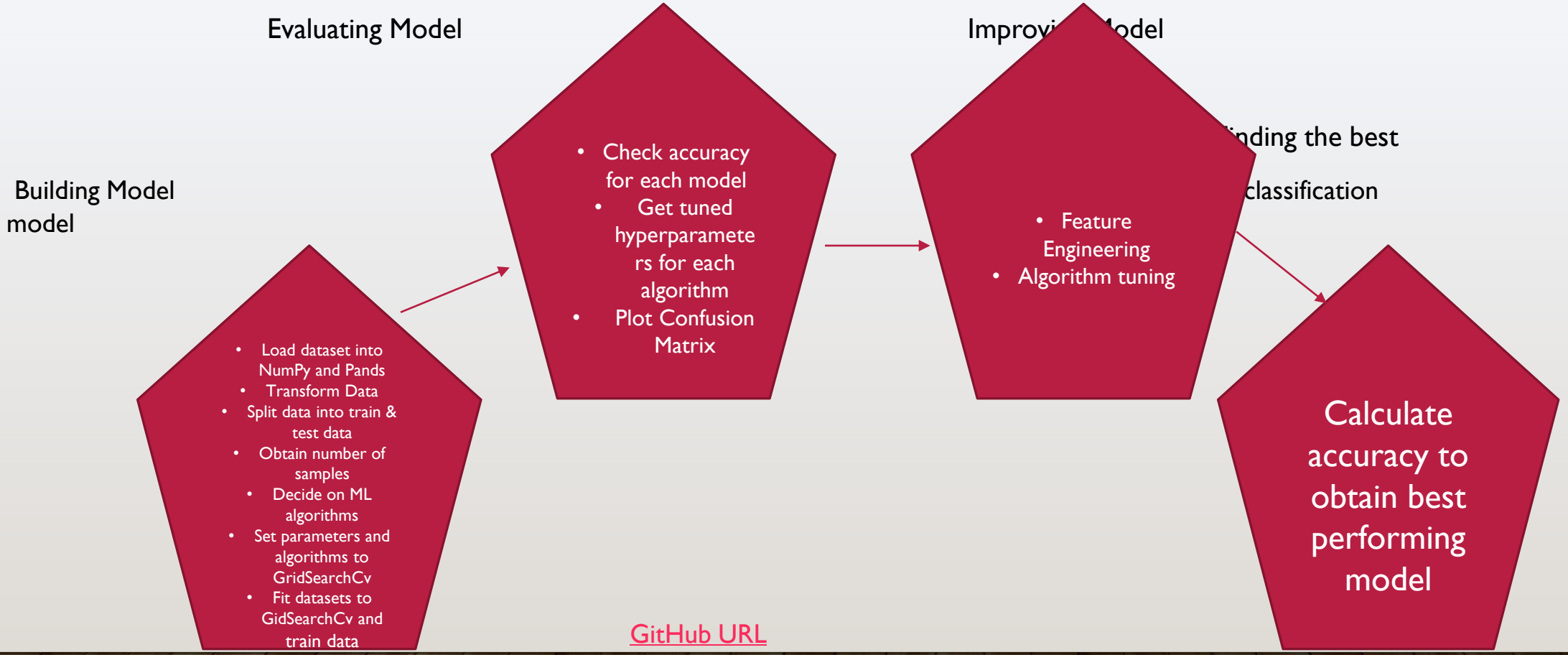
Pie Chart was used to show success rate for a selected Launch site/All Launch sites.

Scatter Plot graph was used to reflect the Launch outcomes against the Booster Payload Mass

GitHub URL to Source Code

Predictive Analysis (Classification)

15



GitHub URL

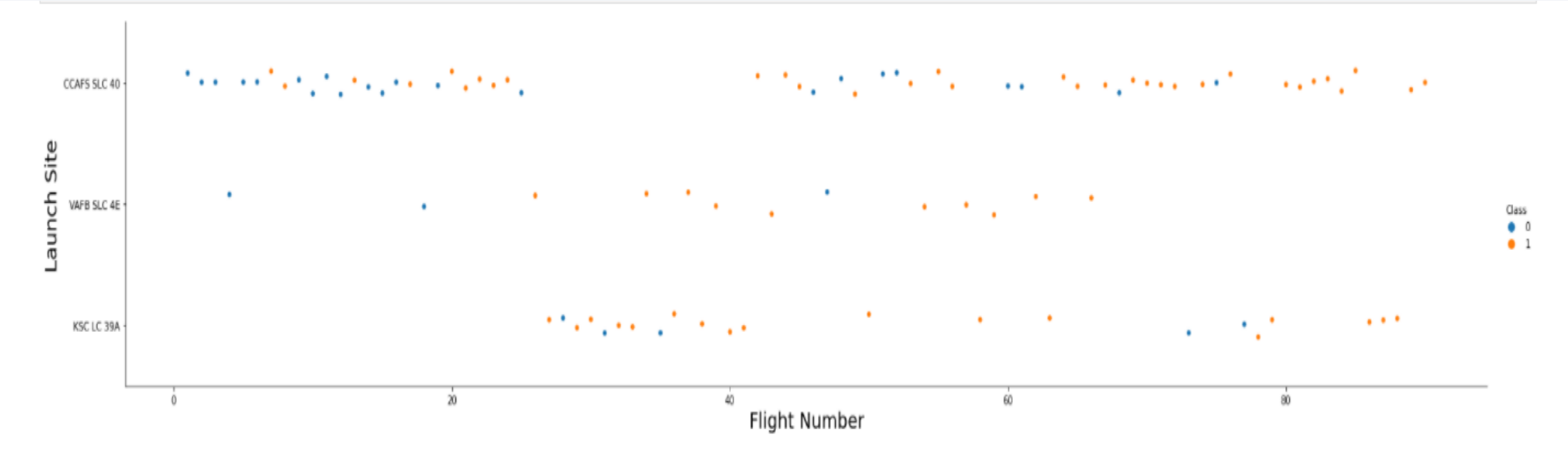
Results

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

Section 2

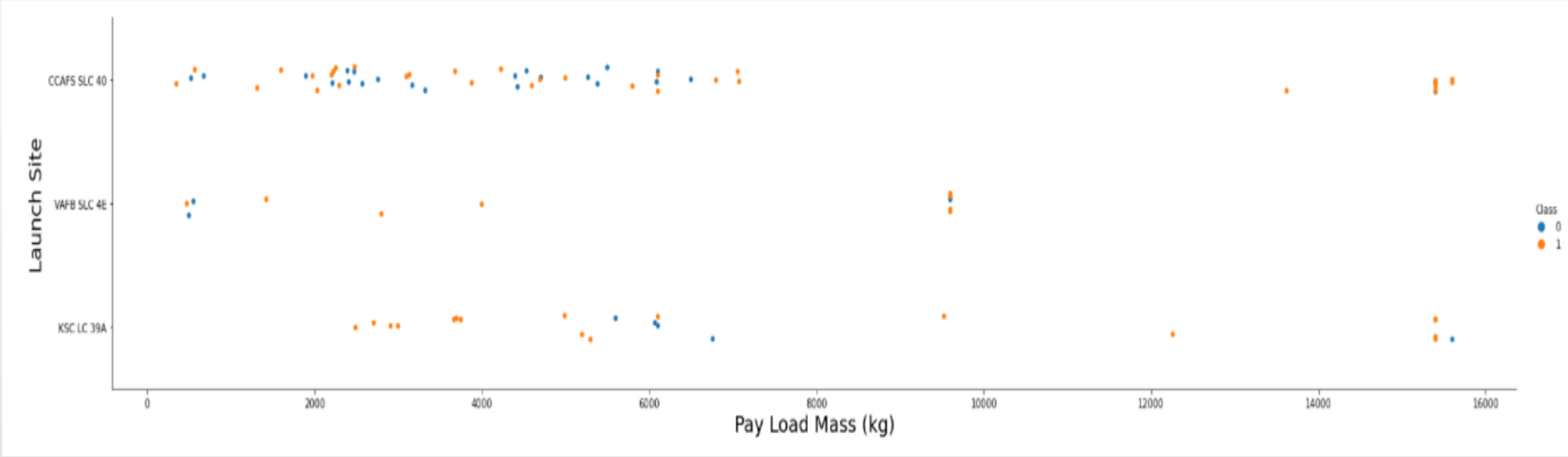
Insights drawn from EDA

Flight Number vs. Launch Site



The results of the graph indicate the Number of successful launches increases with the increase in Number of Launches at a Launch site.

Payload vs. Launch Site

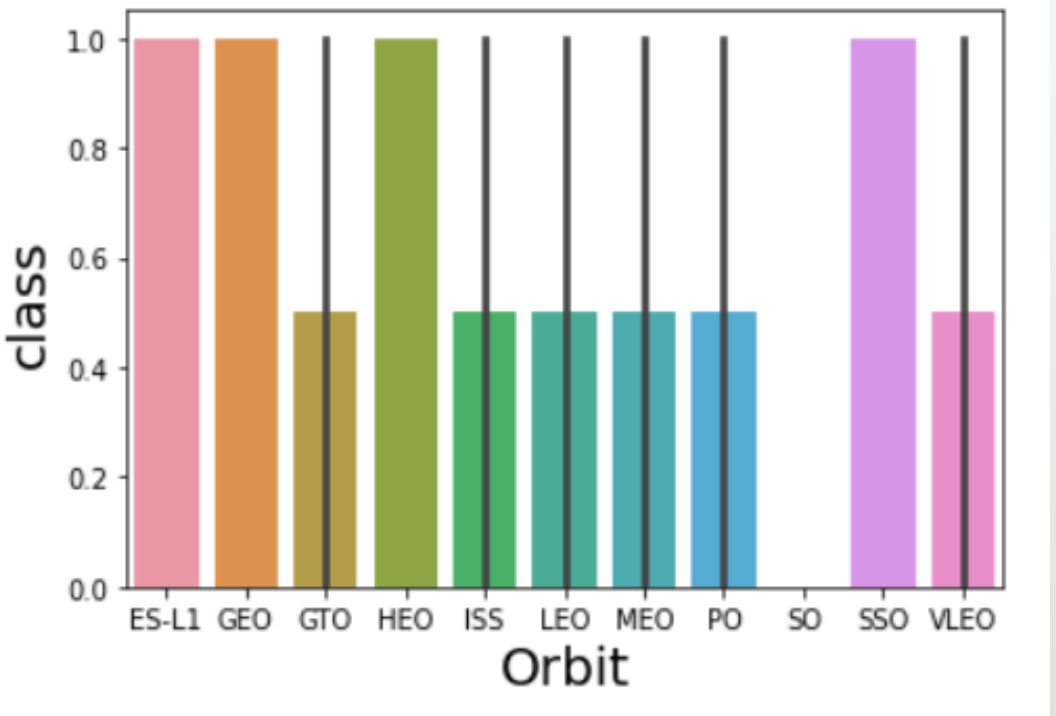


The results of the graph indicate, a higher success rate with a higher Payload Mass.

Success Rate vs. Orbit Type

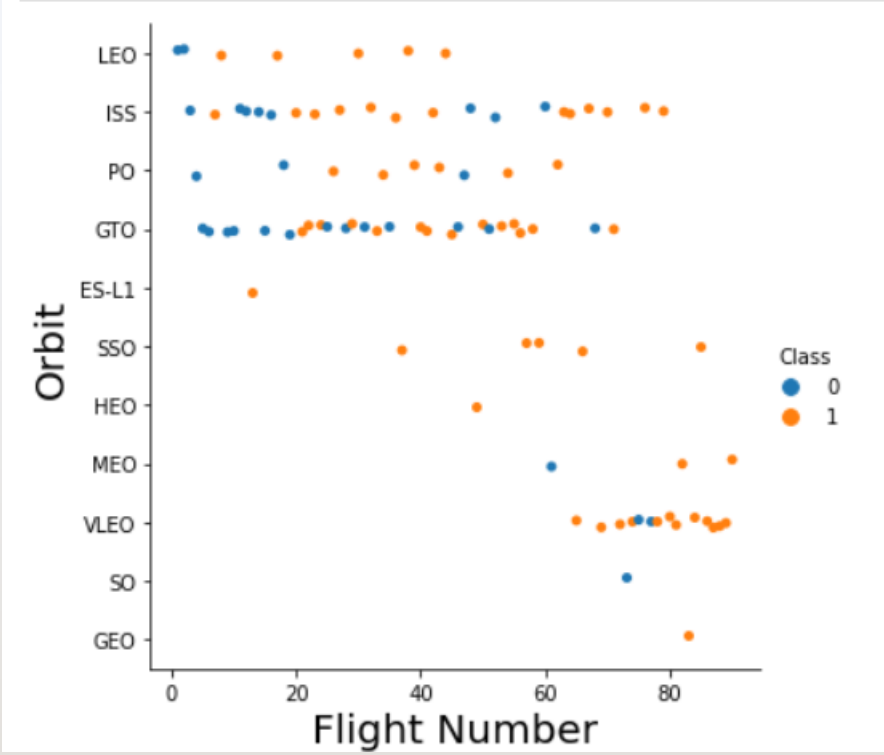
Orbits ES-L1, GEO, HEO and SSO have the best Su

Rate.



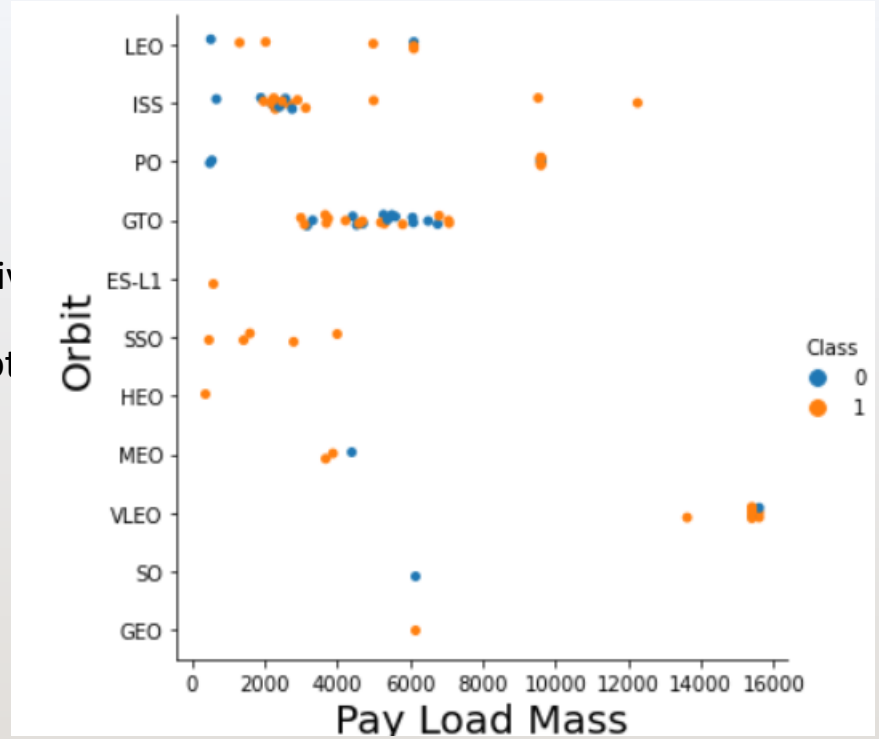
Flight Number vs. Orbit Type

There appears to be a positive relationship between the number of flights and the Success rate.

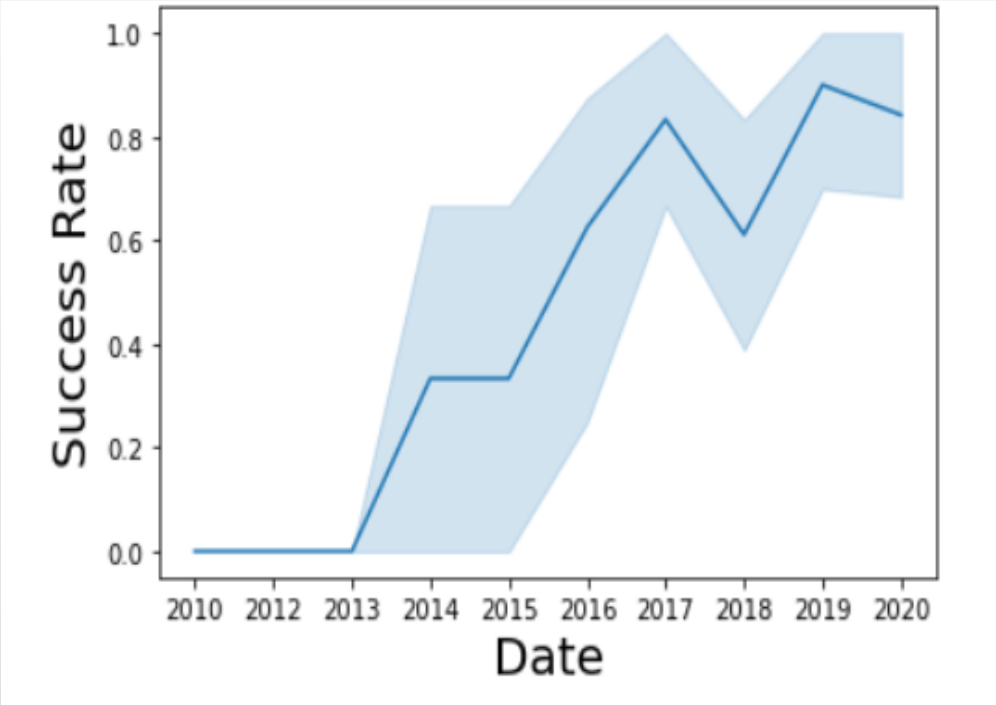


Payload vs. Orbit Type

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



Launch Success Yearly Trend



We observe that the sucess rate since 2013

kept increasing till 2020.

All Launch Site Names

SQL Query

```
%sql select distinct (LAUNCH_SITE) from SPACEX
```

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

The word, distinct, in the query finds only the unique Launch site names from the SPACEX table.

Launch Site Names Begin with 'CCA'

SQL Query

%sql select * from SPACEX where LAUNCH_SITE like 'CCA%' limit 5

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

where and **like** have been used to retrieve Launch site names from the dataset which resemble “CCA”. **limit** was used to retrieve only 5 results.

Total Payload Mass

SQL Query

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEX where CUSTOMER = 'NASA (CRS)'
```

I
45596

Sum summated the Payload Mass column. **Where** is used to filter only results for ‘NASA(CRS)’

Average Payload Mass by F9 v1.1

SQL Query

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEX where BOOSTER_VERSION = 'F9 v1.1'
```

I
2928

Avg calculates the Payload Mass average. **Where** is used to filter the results relevant to ‘F9v1.1’

First Successful Ground Landing Date

SQL Query

```
%sql select min(DATE) from SPACEX where Landing__Outcome = 'Success (ground pad)'
```

I
2015-12-22

Min retrieves the data for the minimum date from the DATE column. **Where** clause filters the data for results for Success(ground pad) from the Landing_Outcome column.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query

```
%sql select BOOSTER_VERSION from SPACEX where Landing__Outcome = 'Success (drone ship)' and
PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

And & the **where** clauses, were used to filter the data to reflect results specific to the requested Payload Mass and Landing Outcome.

Total Number of Successful and Failure Mission Outcomes

30

SQL Query

```
%sql select count(MISSION_OUTCOME) from SPACEX where
MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in
flight)'
```

1
100

Count and **where** clauses were used to filter out the data for the number of relevant Mission Outcomes.

Boosters Carried Maximum Payload

SQL Query

```
%sql select BOOSTER_VERSION from SPACEX where PAYLOAD_MASS__KG_ =
(select max(PAYLOAD_MASS__KG_) from SPACEX)
```

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

Select and **where** clauses, including a subquery, were used to filter the data for the requested results.

2015 Launch Records for Failed landing_outcomes in drone ship

SQL Query

```
%sql select substr(Date, 6, 2) as Month, Landing_Outcome, booster_version, launch_site from spacextbl where
substr(Date, 1, 4) = '2015' and Landing_Outcome like 'Fail%'
```

Month	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

We used **substr(Date, 4, 2)** as month to get the months and **substr(Date,7,4)='2015'** for year

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

SQL Query

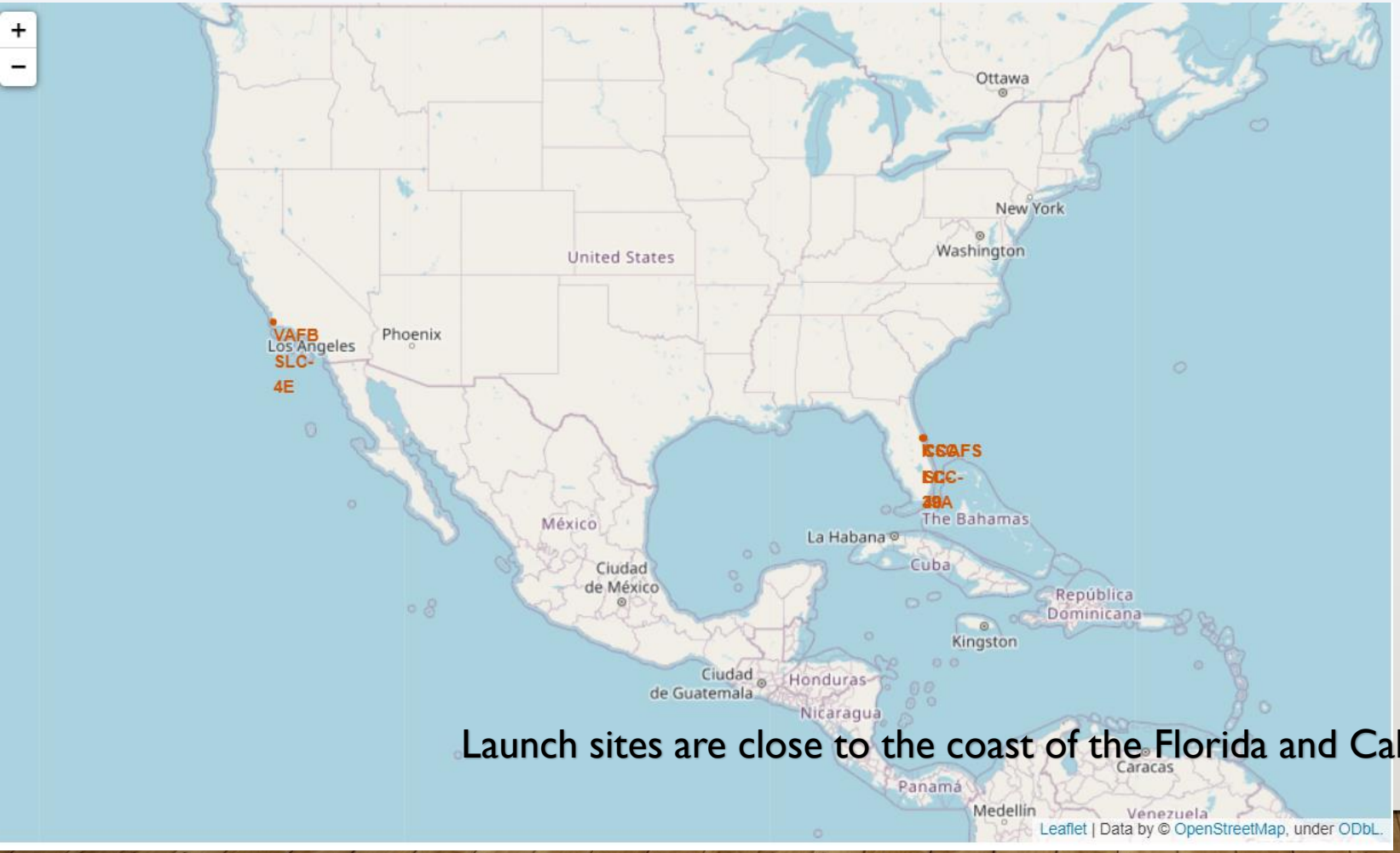
```
%sql SELECT LANDING__OUTCOME,COUNT(LANDING__OUTCOME) FROM SPACEX WHERE DATE
BETWEEN '2010-06-04' AND '2017-03-20' AND LANDING__OUTCOME LIKE 'Success%'GROUP BY
LANDING__OUTCOME
```

landing_outcome	2
Success (drone ship)	5
Success (ground pad)	3

WHERE, LIKE & **Group** clauses were used to filter the number of successful Landings.

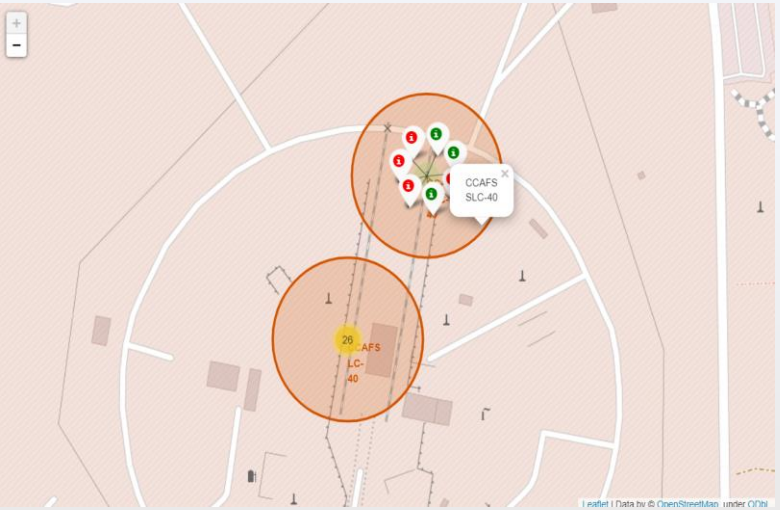

Section 4
Launch Sites
Proximities Analysis

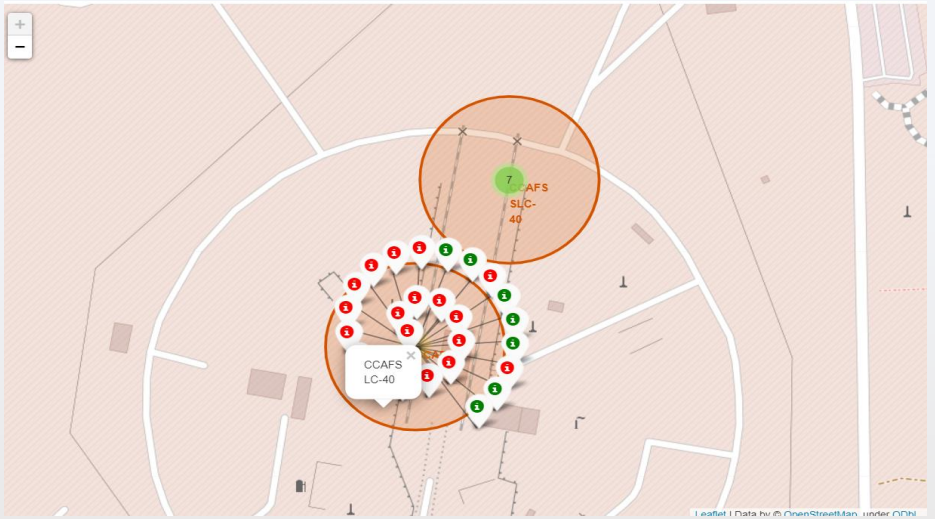
Launch Site Markers

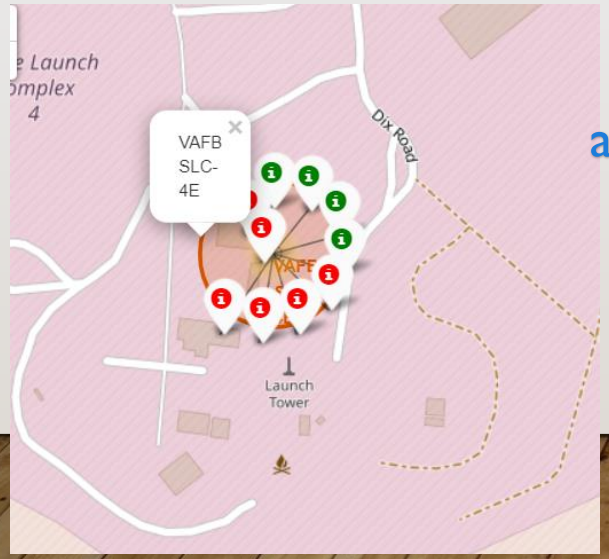


Launch sites are close to the coast of the Florida and California in USA

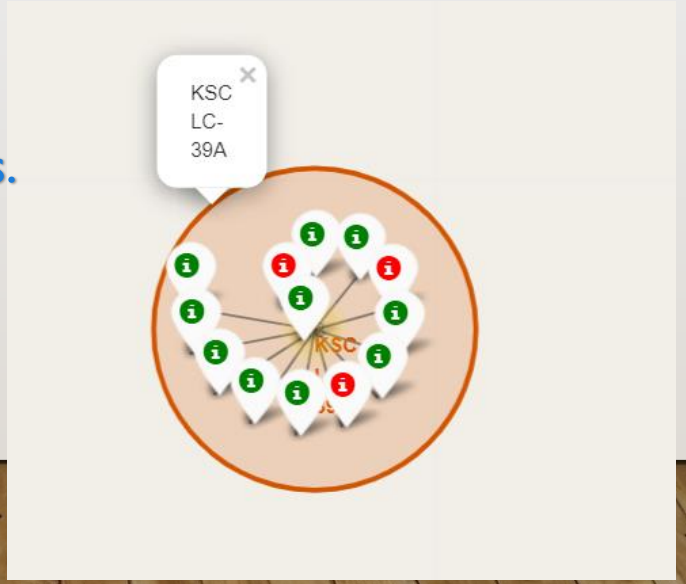
Markers reflecting Success and Failure for Launches





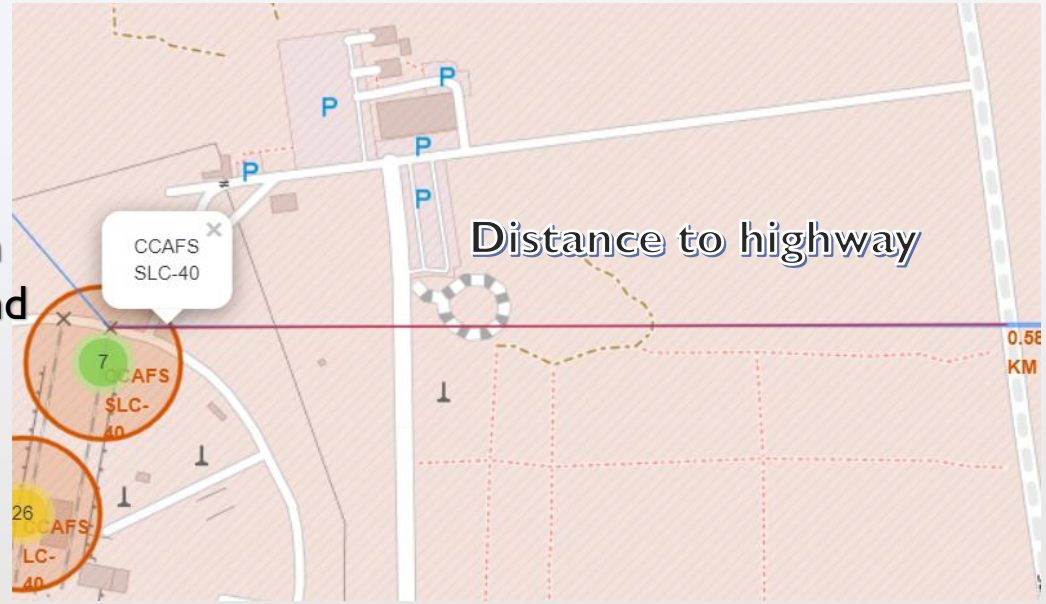


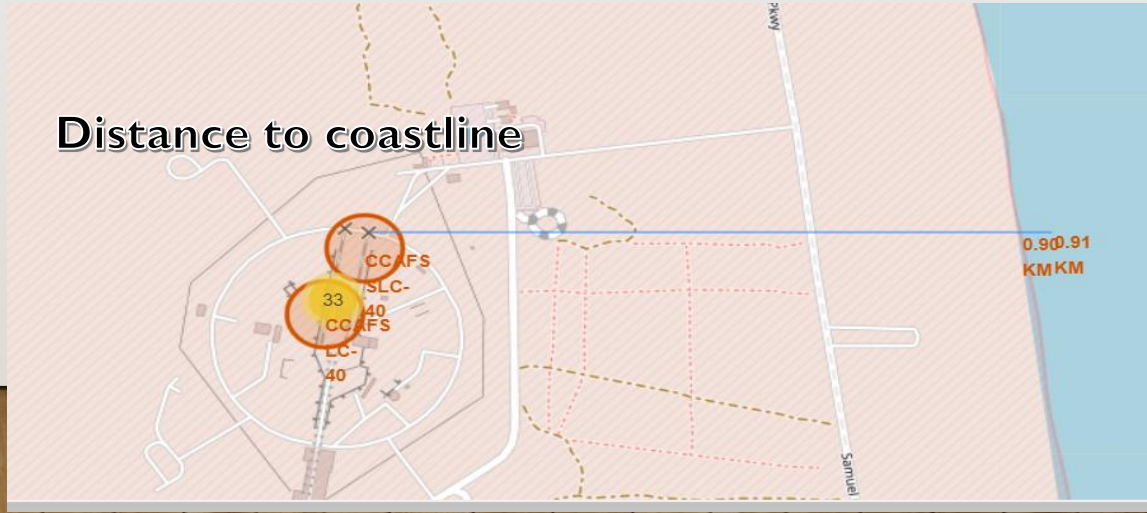
Red markers indicate Failure and green marker reflect success. KSC LC-39A reflects high success



CCAFS SLC-40 Launch site distance to proximities

The launch site is closest in Proximity to the highway and furthest from the railway



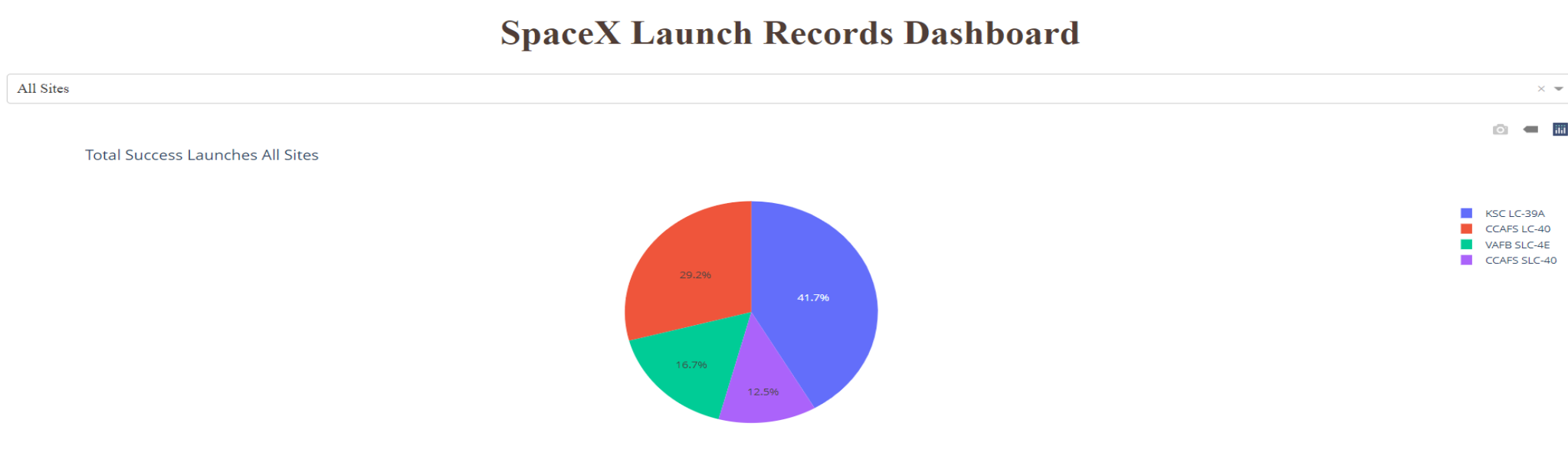


Section 5

Build a Dashboard with Plotly Dash

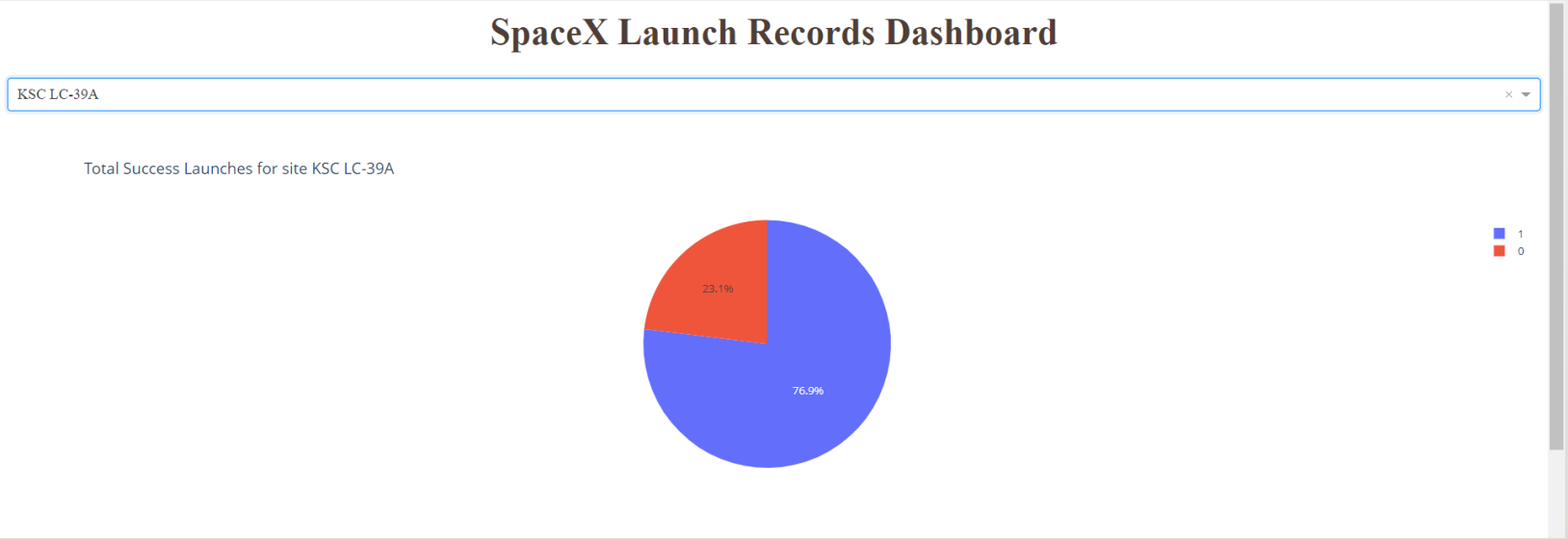
Launch Success Plotly Dashboard

KSC LC-39A had the most amount of successful Launches

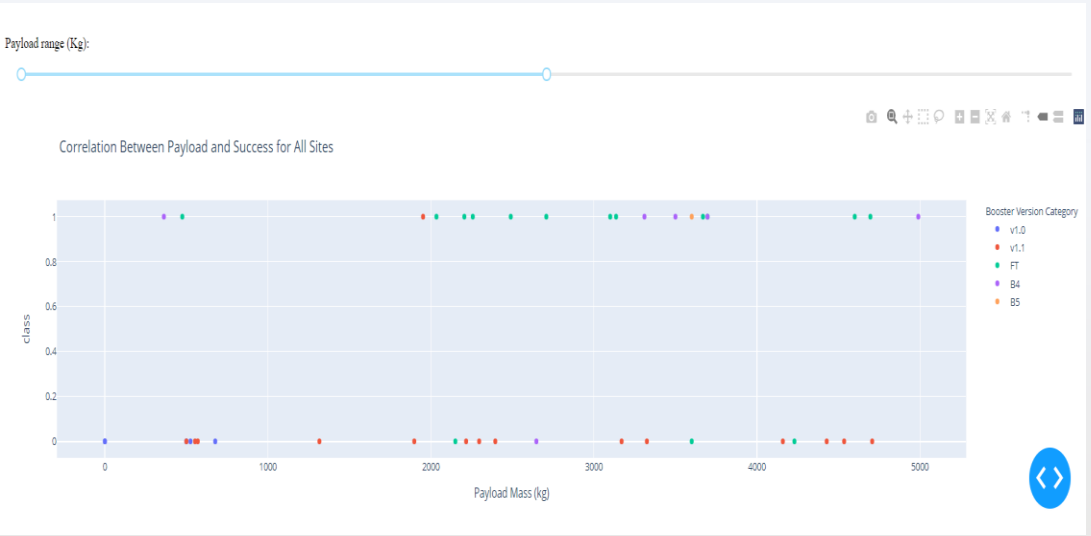


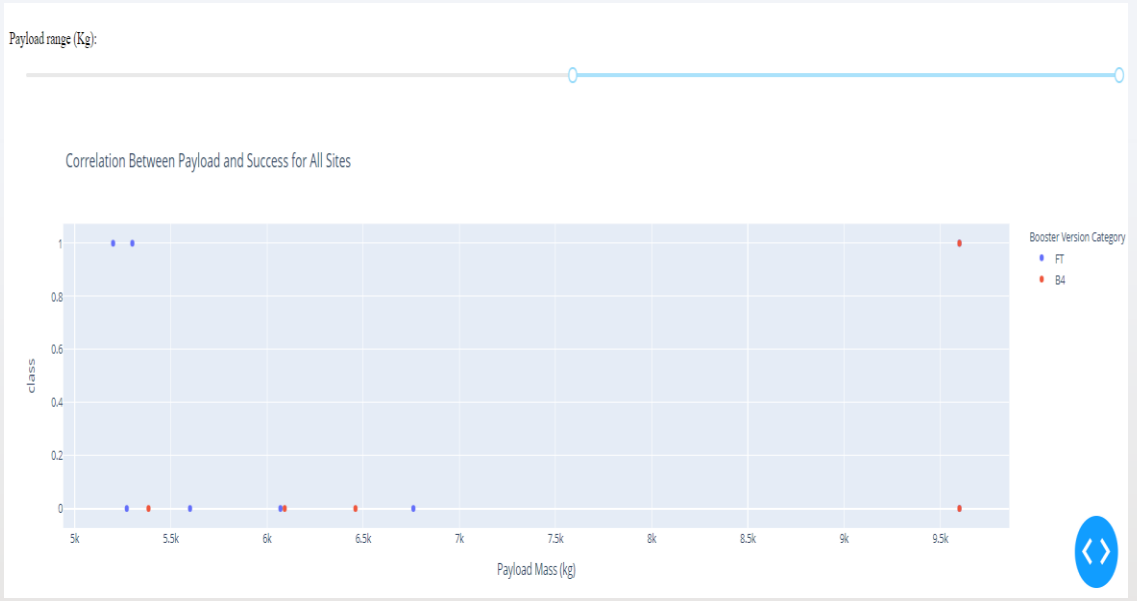
Highest Launch site success ratio

KSC-LC-39A has a success ratio of 76,9%



Payload vs. Launch Outcome scatter plot

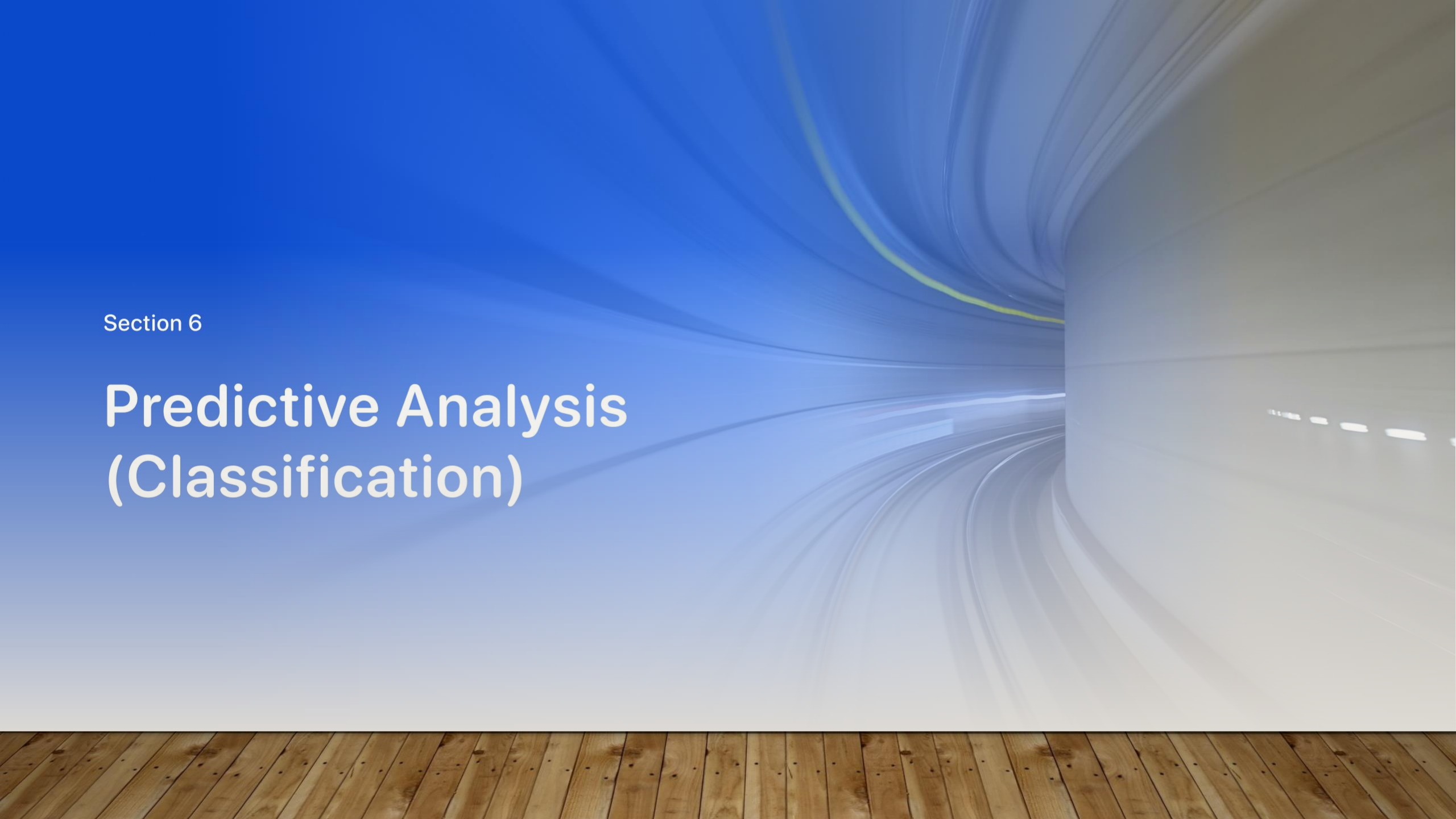




The Success rate for Low weighted Payloads are higher than its is for higher weighted Payloads

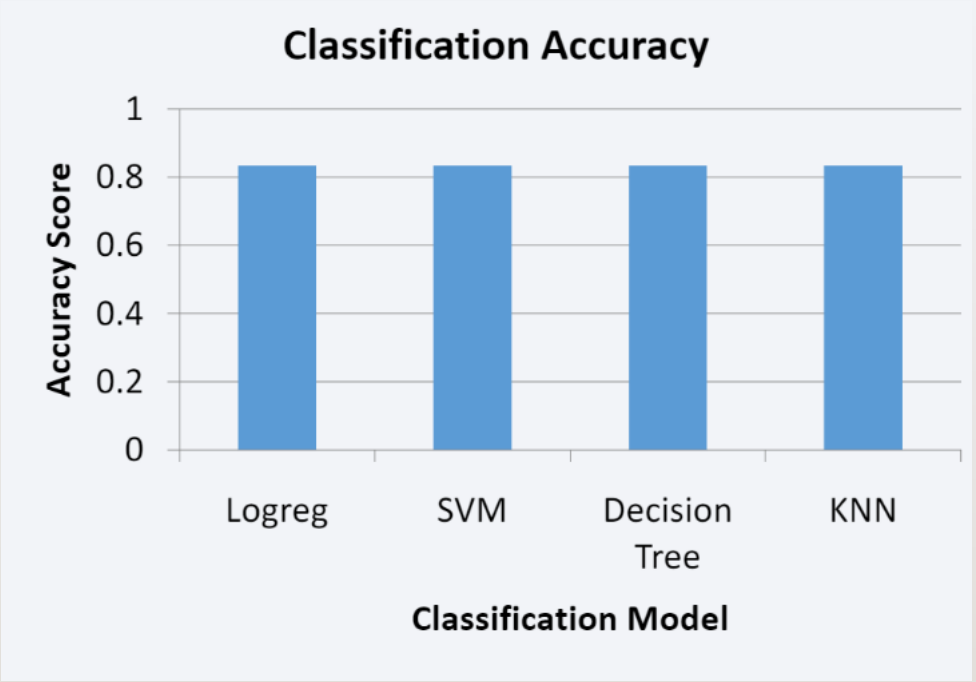
Section 6

Predictive Analysis (Classification)

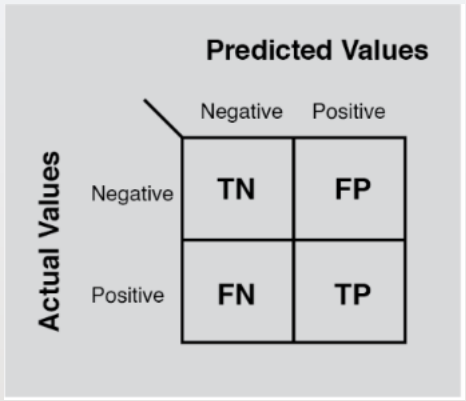


Classification Accuracy

All models had an accuracy of 83.33%

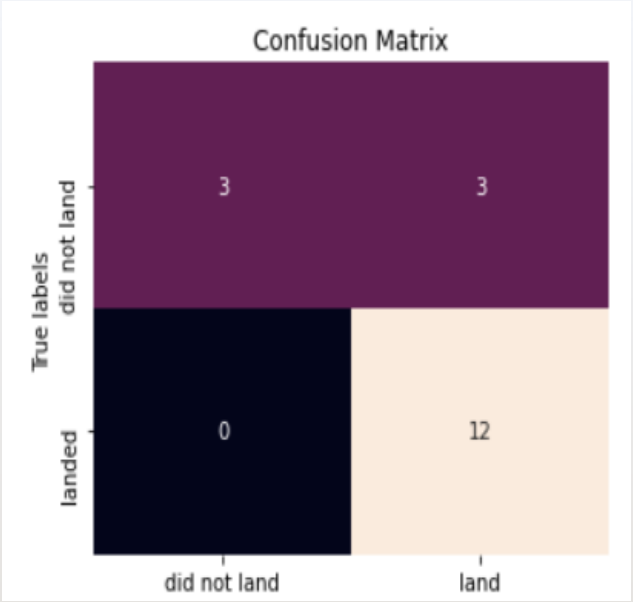


Confusion Matrix



-All models reflected the same outcome for the Confusion Matrix

-The model did accurately predict 12 True positives And 3 False positives



Conclusions

- The Best Algorithm from the performed algorithms is Tree
- KSC LC-39A has the highest success rate
- Low weighted Payloads result in a higher success rate
- The higher the number of flights, the higher the Success rate

Appendix

- https://github.com/ashani84/Coursera_Capstone.git

Thank you!