IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andy Bernabe
12/19/2023

Outline

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

Executive Summary

Data Methodology

- Business understanding
- Analytic approach
- Data requirements
- Data collection
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment
- Feedback

Introduction

- We will predict if the Falcon 9 first stage will land successfully. 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 if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

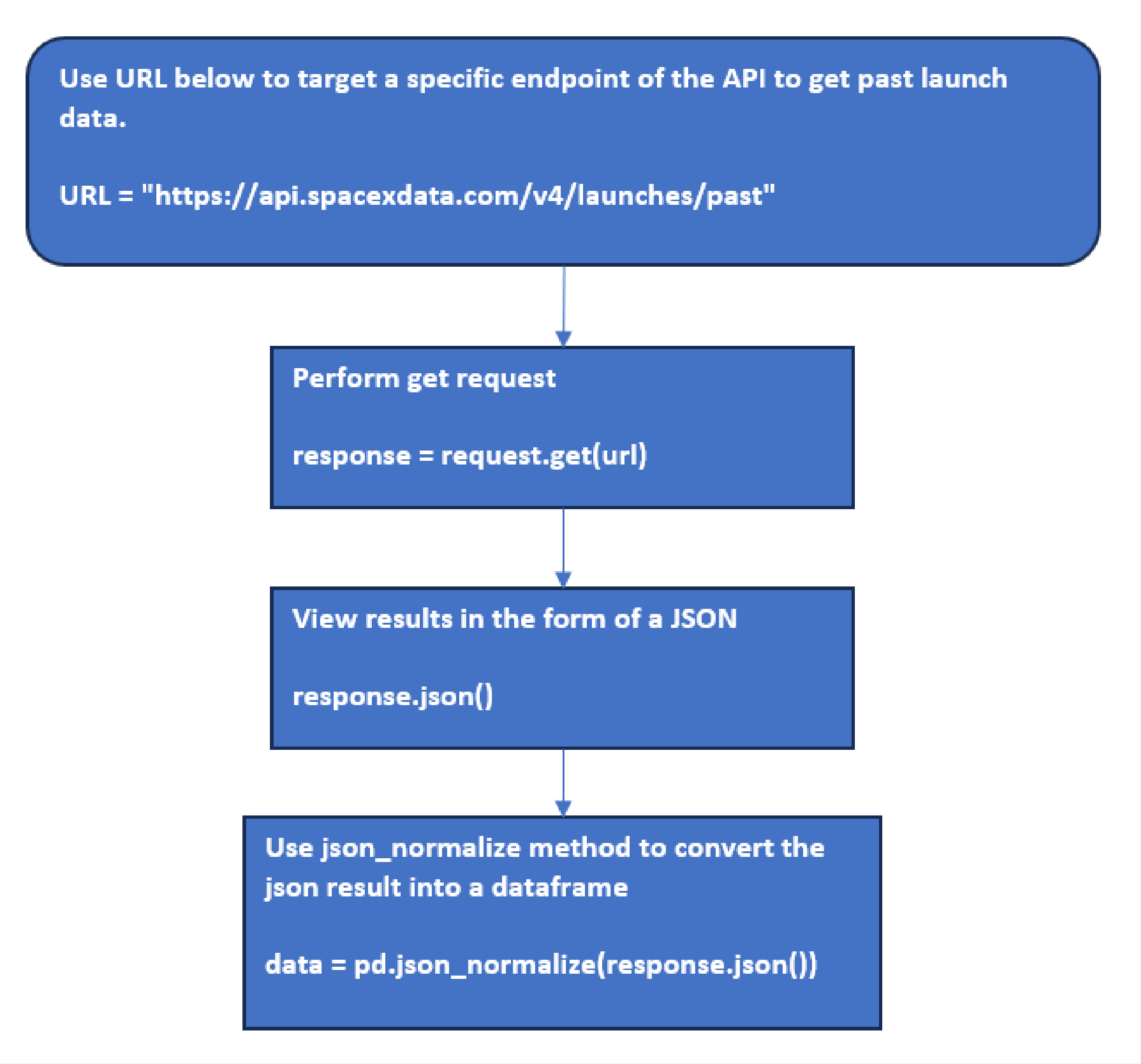
Executive Summary

- Data collection methodology:
  - SpaceX launch data was collected via SpaceX REST API. This API gave data about launches, including information about rocket used, payload delivered, launch specifications, landing specifications, and landing outcome
- Perform data wrangling
  - Data obtained via the SpaceX REST API was converted from html table to data frame and only included data related to Falcon 9.
  - Landing outcome label was created from Outcome column.
- 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

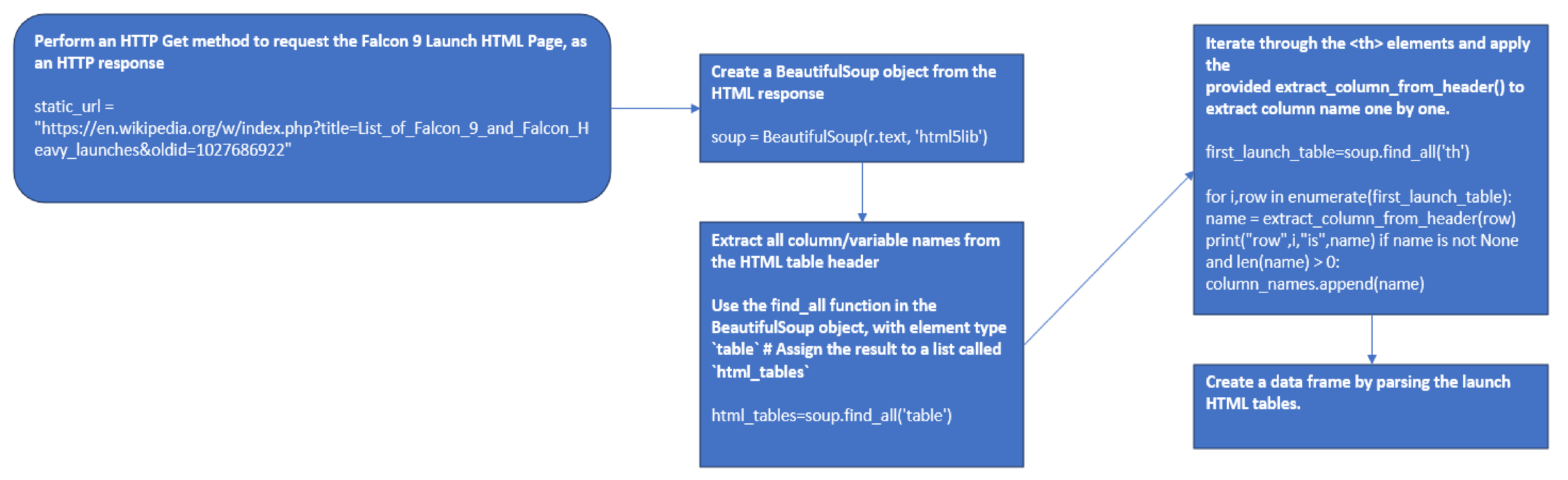
- Data was collected with SpaceX REST API and web scraping Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.

Data Collection – SpaceX API



https://github.com/andy1646/Capstone/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/andy1646/Capstone/blob/main/2_jupyter-labs-webscraping.ipynb

Data Wrangling

Wrangling data using API Process

Convert JSON to a dataframe by using the json_normalize function.

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

Take a subset of the dataframe keeping only the features needed and the flight number, and date_utc.

```
data = data[['rocket', 'payloads',
'launchpad', 'cores', 'flight_number',
'date_utc']]
```

Remove rows with multiple cores because those are falcon rockets with 2 extra rocket boosters and rows that have multiple payloads in a single rocket.

```
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]
```

Since payloads and cores are lists of size 1 we will also extract the single value in the list and replace the feature.

```
data['cores'] = data['cores'].map(lambda x :
x[0])
data['payloads'] =
data['payloads'].map(lambda x : x[0])
```

Convert the date_utc to a datetime datatype and then extracting the date leaving the time.

```
data['date'] =
pd.to_datetime(data['date_utc']).dt.date
```

Using the date we will restrict the dates of the launches.

```
data = data[data['date'] <=
datetime.date(2020, 11, 13)]
```

Creates lists of global variables which will be used to create a new dataframe. Also apply the functions below to get correct column name data

- getBoosterVersion(data)
- getLaunchSite(data)
- getPayloadData(data)
- getCoreData(data)

Filter the dataframe to only include Falcon 9 launches

```
data_falcon9 = df.loc[df['BoosterVersion'] !=
'Falcon 1']
```

Reset the FlgihtNumber column

```
data_falcon9.loc[:,'FlightNumber'] =
list(range(1, data_falcon9.shape[0]+1))
```

Deal with missing values for Payload Mass

```
data_falcon9.isnull().sum()
```

Calculate the mean for Payload Mass and replace null values with it

```
payloadMassAverage =
data_falcon9["PayloadMass"].mean()

data_falcon9['PayloadMass'].fillna(value=
payloadMassAverage, inplace=True)
```

https://github.com/andy1646/Capstone/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb

EDA with Data Visualization

- The following charts below were created to visualize relationship between them
  - Flight Number vs. Launch Site (Scatter Chart)
  - Payload vs. Launch Site (Scatter Chart)
  - Success Rate vs. Orbit Type (Bar Chart)
  - Flight Number vs. Orbit Type (Scatter Chart)
  - Payload vs. Orbit Type (Scatter Chart)
- The following chart below was created to visualize trend
  - Launch Success Yearly Trend (Line Chart)

https://github.com/andy1646/Capstone/blob/main/5_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Performed the following queries below
    1. %sql select distinct "Launch_Site" from SPACEXTABLE
        - Display the names of the unique launch sites in the space mission
    2. %sql select * from SPACEXTABLE where "Launch_Site" like "%CCA%" limit 5
        - Display 5 records where launch sites begin with the string 'CCA'
    3. %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" = "NASA (CRS)“
        - Display the total payload mass carried by boosters launched by NASA (CRS)
    4. %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" = "F9 v1.1“
        - Display average payload mass carried by booster version F9 v1.1
    5. %sql select min(Date) from SPACEXTABLE where "Mission_Outcome" = "Success“
        - List the date when the first succesful landing outcome in ground pad was achieved
    6. %sql select distinct "Booster_Version" from SPACEXTABLE where "Mission_Outcome" = "Success" and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
        - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    7. %sql select Mission_Outcome, count(Mission_Outcome) as Total_Number from SPACEXTABLE group by Mission_Outcome
        - List the total number of successful and failure mission outcomes

EDA with SQL

- Performed the following queries below

8. %sql select Booster_Version from (select Booster_Version, max(PAYLOAD_MASS__KG_) from SPACEXTABLE group by Booster_Version

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

9. %sql select substr(Date, 6,2) as Month, Booster_Version, Launch_Site from SPACEXTABLE where "Landing_Outcome" = "Failure”

    - 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

10. %sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE where Date >= '2010-06-04' and Date <= '2017-03-20’

    - 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

https://github.com/andy1646/Capstone/blob/main/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- The following map objects below were created and added to the folium map
  1. Map – to generate map
  2. Circle - to add a highlighted circle area with a text label on a specific coordinate
  3. Marker – to add various objects such as markers, circles, lines, text to map
  4. Polyline – to draw a line between launch site to selected point

https://github.com/andy1646/Capstone/blob/main/6_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

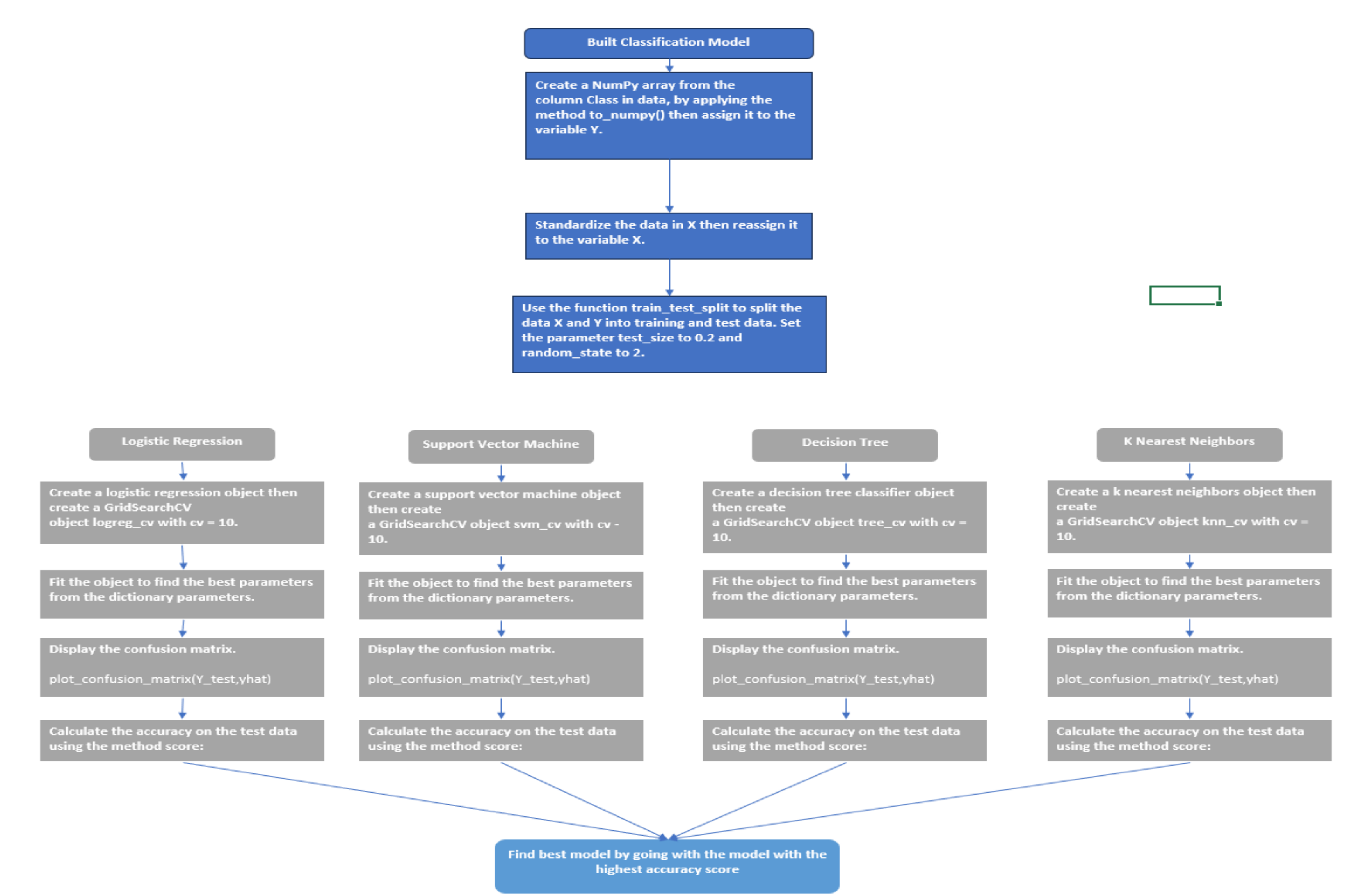
Added the following interactions

1. Launch Site Drop-down Input Component
   - To render success-pie-chart based on selected site dropdown
2. Range Slider to Select Payload
   - To render the success-payload-scatter-chart scatter plot

Added the following graphs

3. Pie Chart – To visualize total success launch by site or by selected site
4. Scatter Chart – To visualize relationship between payload mass, launch outcome, and booster version

Predictive Analysis (Classification)

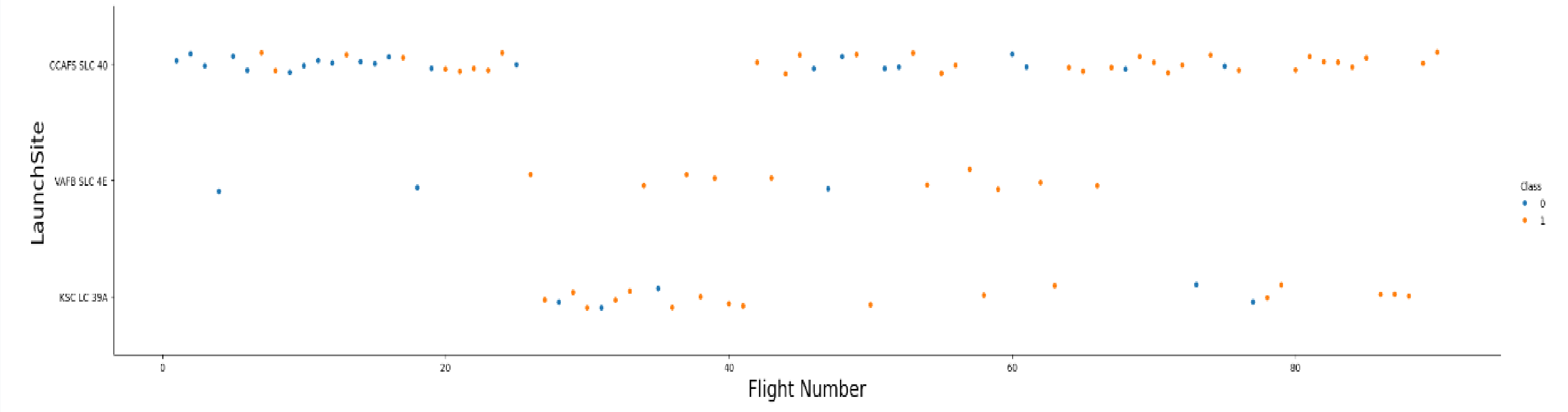


https://github.com/andy1646/Capstone/blob/main/7_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Section 2

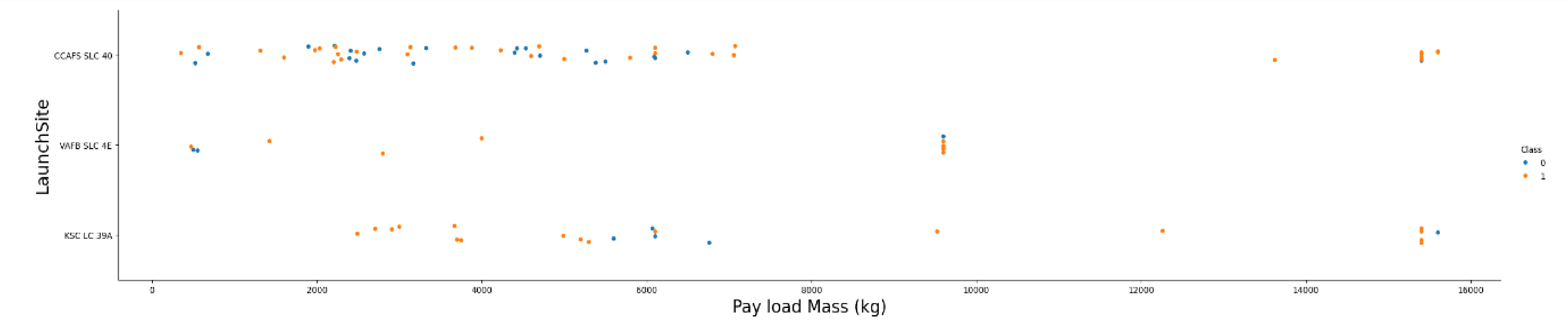
Insights drawn from EDA

Flight Number vs. Launch Site



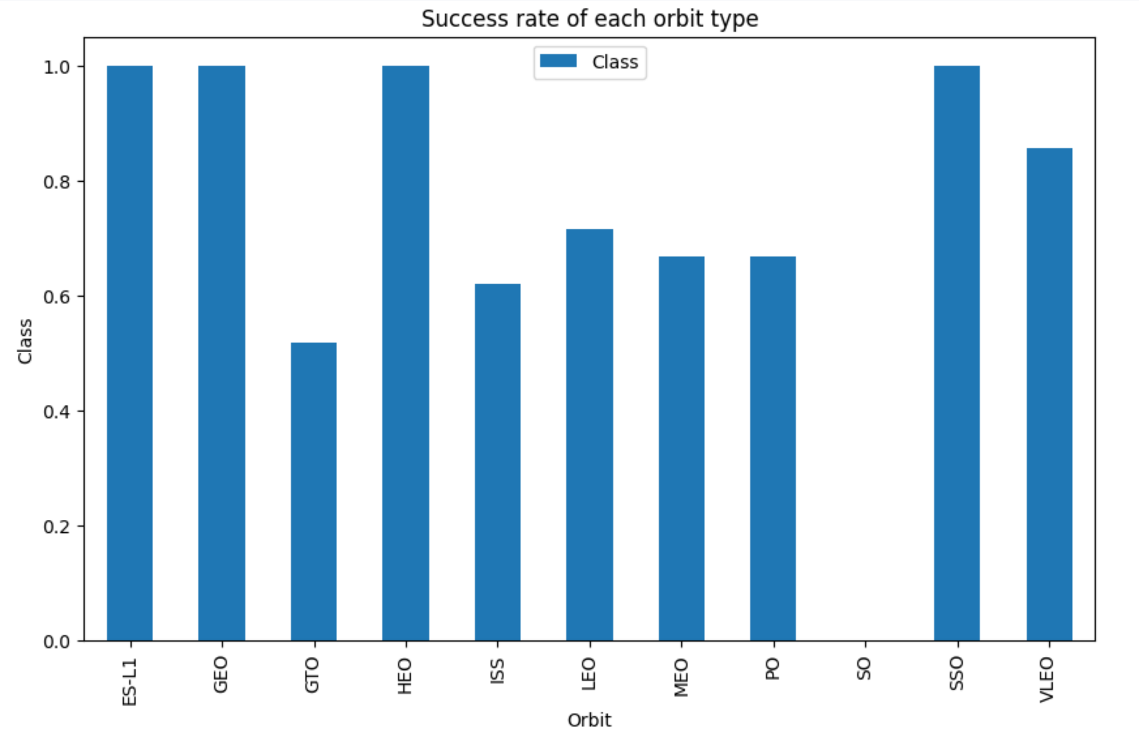
Based on visuals above, launch success increases as the flight number increases.

Payload vs. Launch Site



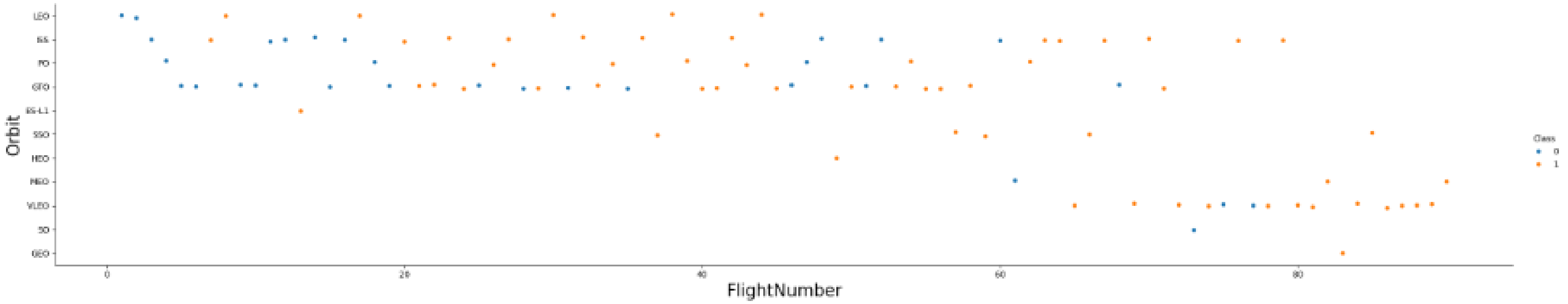
Based on visuals above, launch success
increases as the pay load mass increases.

Success Rate vs. Orbit Type



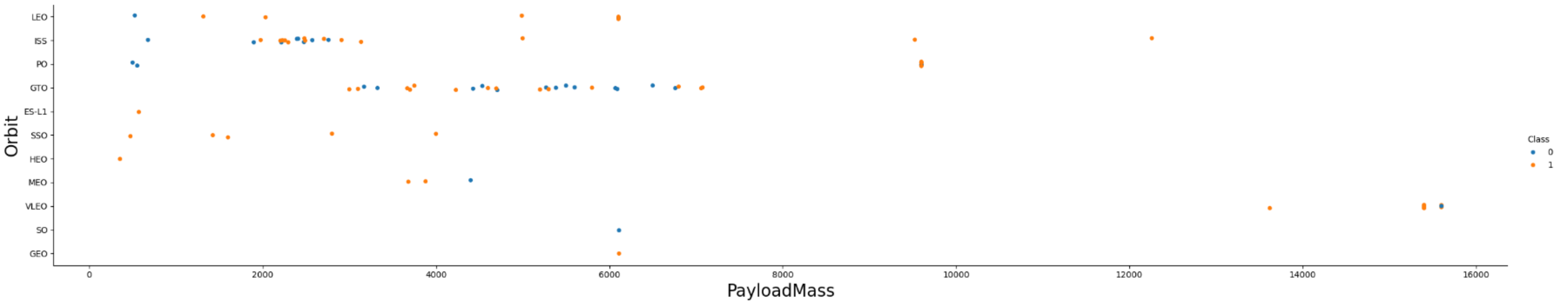
Based on visuals above, orbit GTO has the worse success rate while ES-L1, GEO, HEO and SSO have the best.

Flight Number vs. Orbit Type



You should see that in the 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.

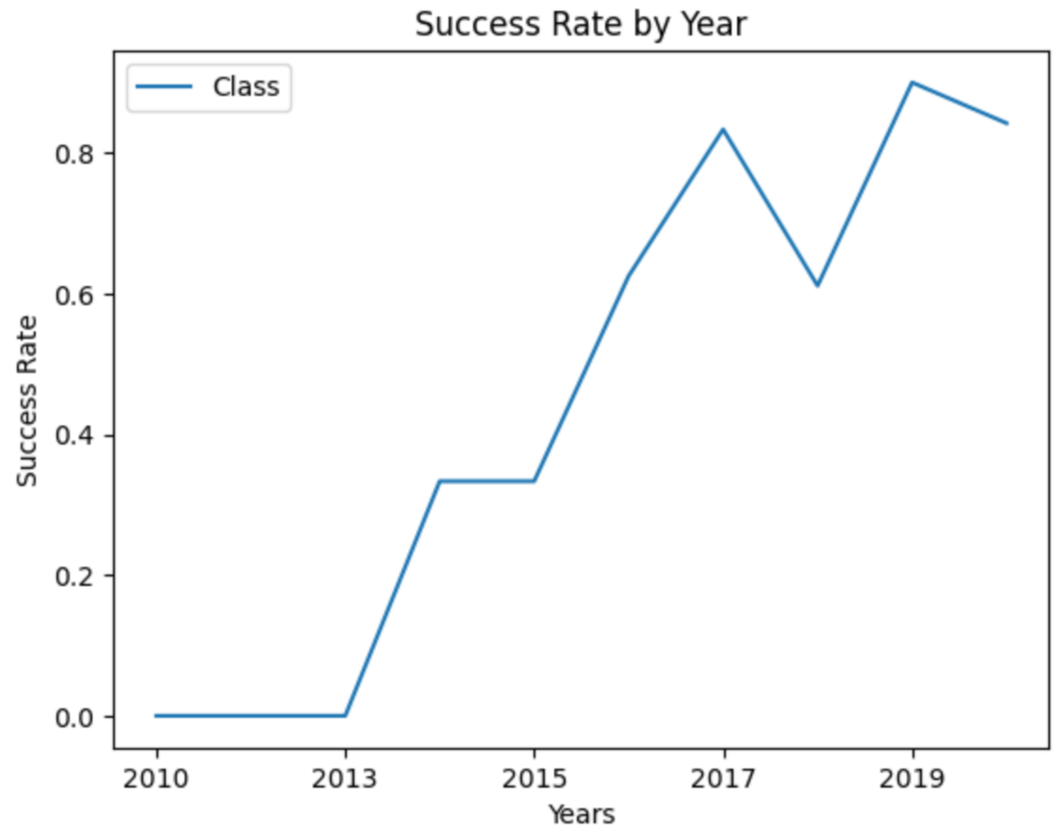
Payload vs. Orbit Type



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(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

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

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

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

```
%sql select * from SPACEXTABLE where "Launch_Site" like "%CCA%" limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcon
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachut
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 (parachut
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Total Payload Mass

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

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" = "NASA (CRS)"
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" = "F9 v1.1"
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql select min(Date) from SPACEXTABLE where "Mission_Outcome" = "Success"
```

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

min(Date)
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select distinct "Booster_Version" from SPACEXTABLE where "Mission_Outcome" = "Success" and PAYLOAD_MASS__KG_ > 4000
```

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

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) as Total_Number from SPACEXTABLE group by Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

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

```
%sql select Booster_Version from (select Booster_Version, max(PAYLOAD_MASS__KG_) from SPACEXTABLE group by Booster_Versio
```

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

Booster_Version
F9 B4 B1039.2
F9 B4 B1040.2
F9 B4 B1041.2
F9 B4 B1043.2
F9 B4 B1039.1
F9 B4 B1040.1
F9 B4 B1041.1
F9 B4 B1042.1
F9 B4 B1043.1
F9 B4 B1044
F9 B4 B1045.1
F9 B4 B1045.2
F9 B5 B1046.1
F9 B5 B1046.2
F9 B5 B1046.3
F9 B5 B1046.4
F9 B5 B1047.2
F9 B5 B1047.3
F9 B5 B1048.2
F9 B5 B1048.3
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.2
F9 B5 B1049.3
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.6
F9 B5 B1049.7

2015 Launch Records

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.

```
%sql select substr(Date, 6,2) as Month, Booster_Version, Launch_Site from SPACEXTABLE where "Landing_Outcome" = "Failure
```

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

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

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

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.

```
%sql select Landing_Outcome, count(Landing_Outcome) from SPACEXTABLE where Date >= '2010-06-04' and Date <= '2017-03-20'
```

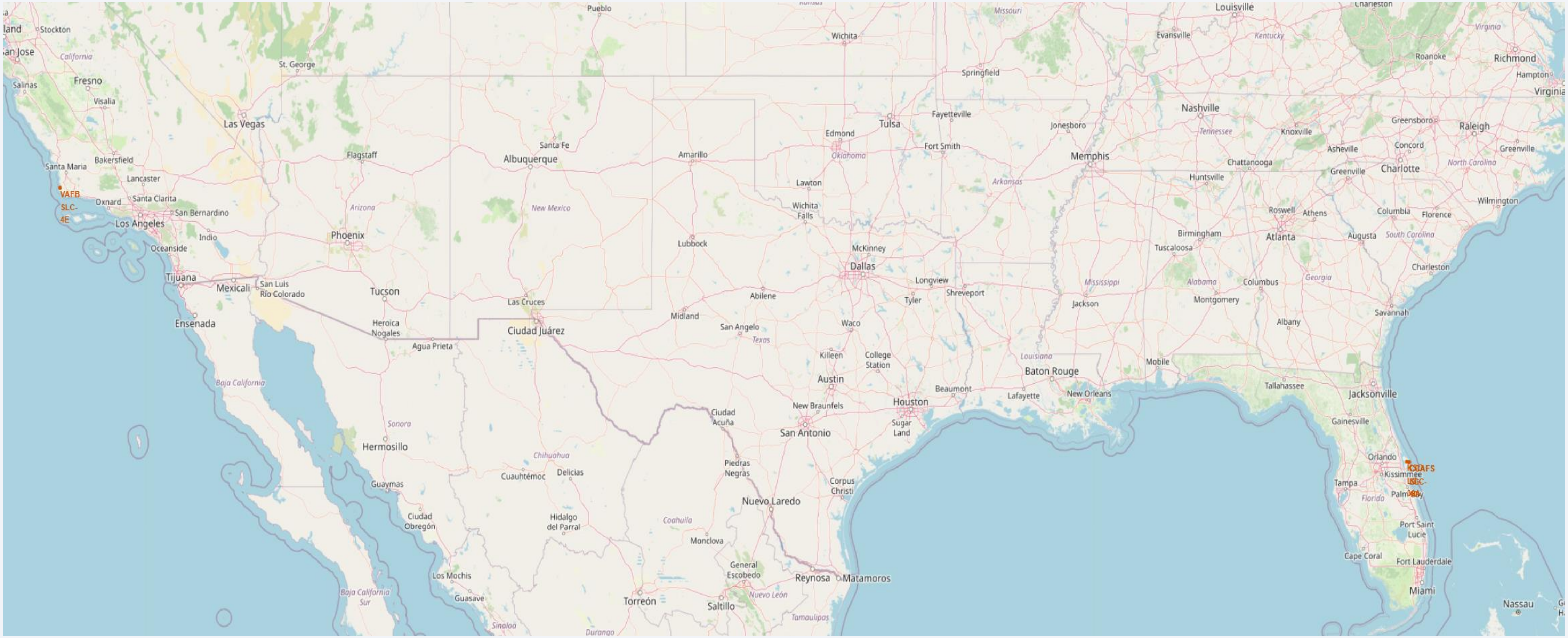
```
* sqlite:///my_data1.db
Done.
```

Landing_Outcome	count(Landing_Outcome)
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

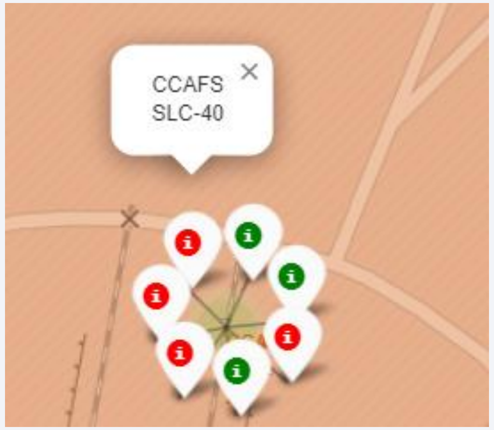
Launch Sites



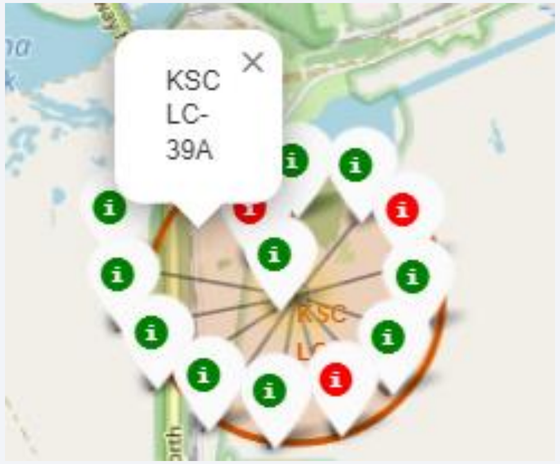
Three launch sites are in central Florida and one launch site is in southern California.

Launch Outcomes by Site

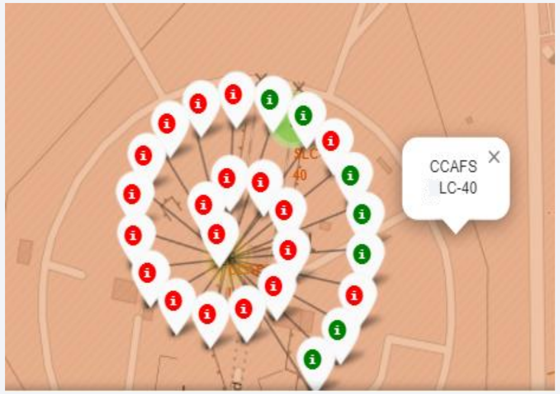
10 launches with 40% success rate



7 launches with 43% success rate

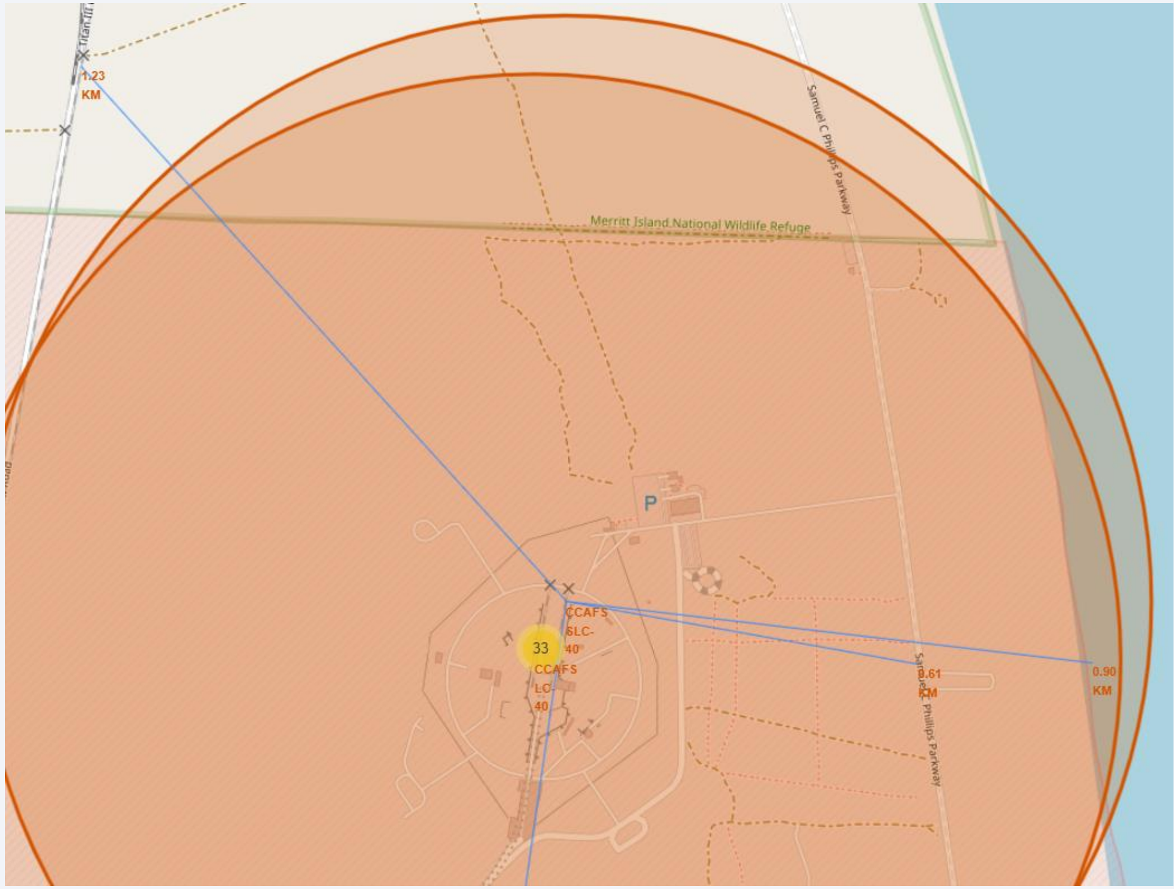


13 launches with 77% success rate



26 launches with 27% success rate

Launch Site to its Proximities



Launch Site CCAFS SLC-40 is
0.61 KM from nearest highway
0.90 KM from nearest coastline
1.23KM from nearest railway
18.22 from nearest city

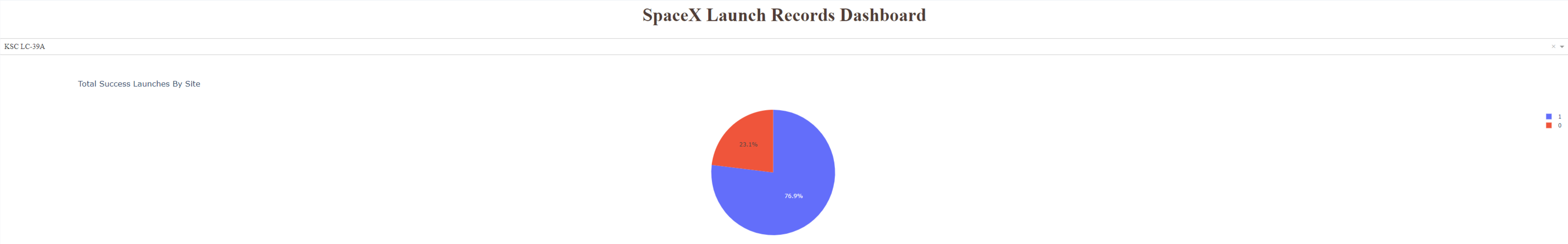
Section 4

Build a Dashboard with Plotly Dash

Launch Success by Site

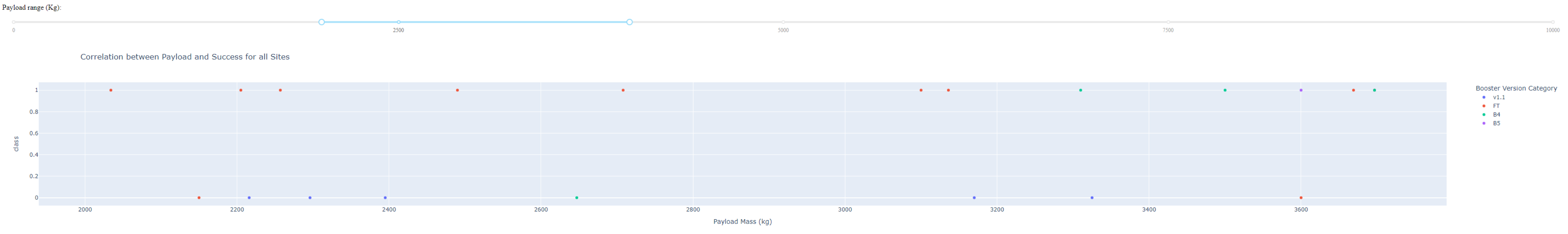
Almost half of the successful launches come from site KSC LC-39A while the least come from site CCAFS SLC-40.

Launch Site with highest launch success ratio



KSC LC-39A has the highest launch success ratio with 76.9% success rate

Payload vs Launch Outcome



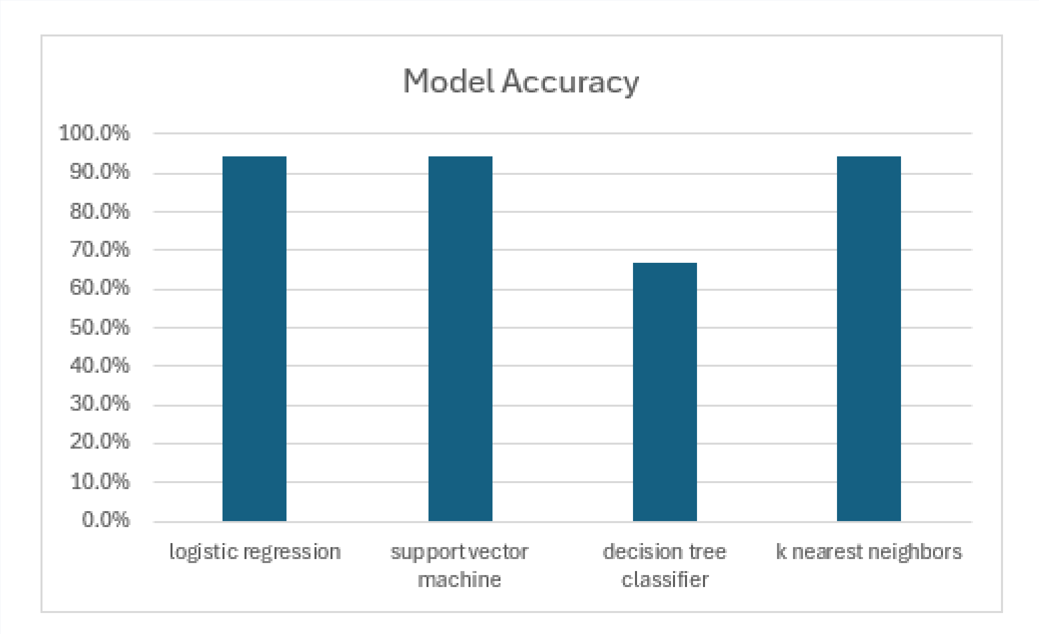
Almost half the launches are booster version FT and have the highest success rate.

Payload range between 2K and 4K have the highest success rate.

Section 5

Predictive Analysis (Classification)

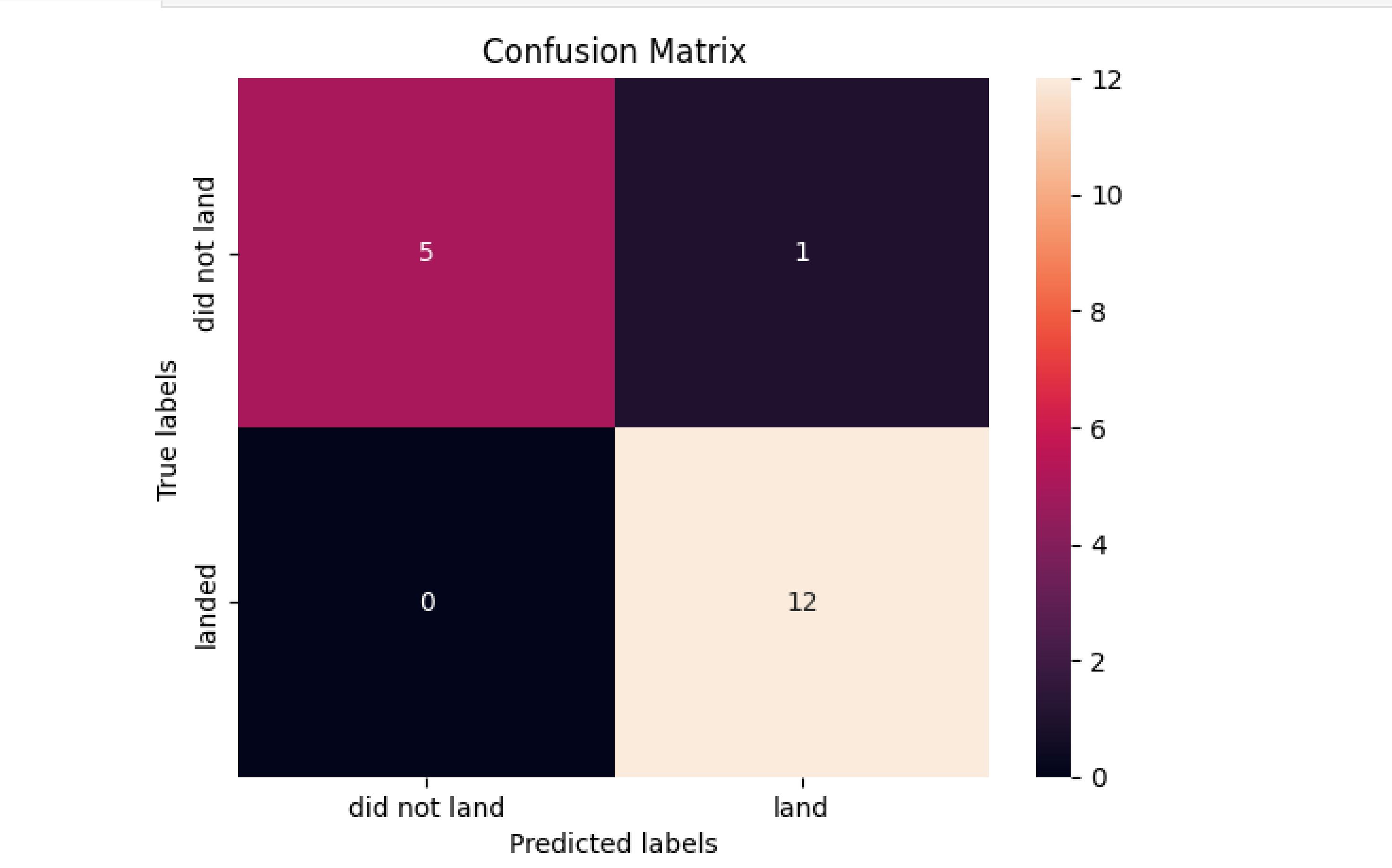
Classification Accuracy



The following models below had the highest accuracy and had the same accuracy.

Logistic Regression
Support Vector Machine
K Nearest Neighbors

Confusion Matrix



Conclusions

- Since 2013, success rate has continued to increase.
- Orbit GTO has the worse success rate while ES-L1, GEO, HEO and SSO have the best.
- KSC LC-39A has the highest launch success ratio with 76.9% success rate.
- Almost half of the successful launches come from site KSC LC-39A while the least come from site CCAFS SLC-40.
- Almost half the launches are booster version FT and have the highest success rate.
- Payload range between 2K and 4K have the highest success rate.
- The following models below had the highest accuracy 94% and had the same accuracy. We can use any of the models to determine if a launch will be successful.
  - Logistic Regression
  - Support Vector Machine
  - K Nearest Neighbors

Thank you!