IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pedro Azevedo
10/05/2022

Outline

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

Executive Summary

- Data was collected from public SpaceX API and by scrapping SpaceX Wikipedia page and used Beautiful soup library for it.
- Labels were created for the column ‘Class’ which denotes all the successful landings.
- EDA (Exploratory Data Analysis) using SQL using notebooks Watson Studio and magicSQL
- Explored Queries using group by ,aggregate functions.
- Processed the data and cleared it
- Trained different Machine Learning algorithms to predict a successful landing

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

- Data collection methodology:
  - Trough the SPACEX API and SpaceX Wikipedia page
- Perform data wrangling
  - Processed using BeautifulSoup and Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Created train test sets and performed cross validation on SVMs , KNN, Decision Trees and LogisticRegression

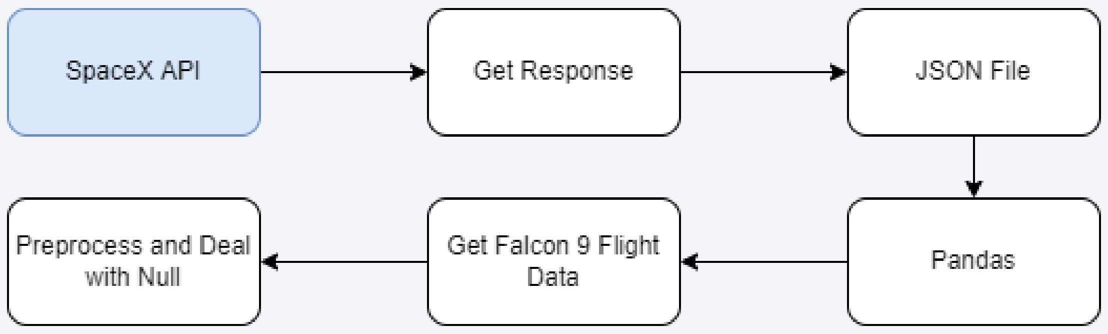
Data Collection

- To have a complete set of data, the following was collected through two distinct methods.
- First the SpaceX API was used to get rocket info booster info and flight history. After the following some Data Scraping was performed on the SpaceX wikipedia page where Falcon 9 launch info was extracted.

Data Collection – SpaceX API

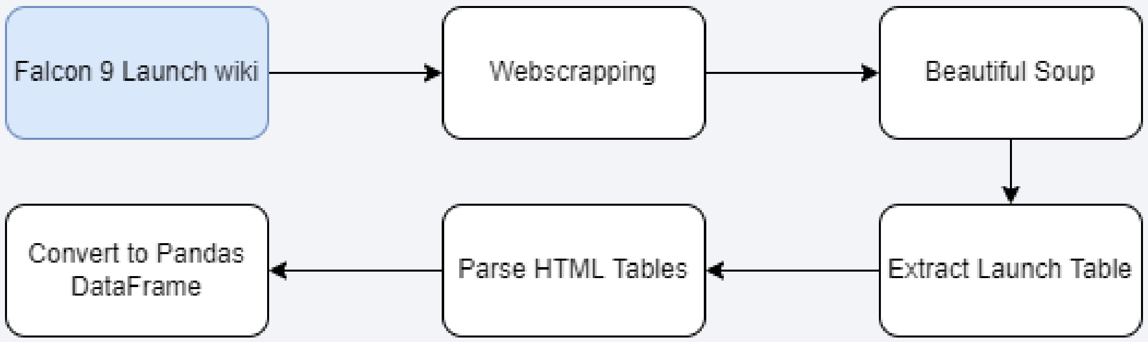
- Data was collected through SpaceX REST calls. Most specifically their *"rockets" "launchpads"* ,*"payloads"* and *"pastlaunches"* endpoints.

- [SpaceX-Rocket-Prediction/Data collection API.ipynb at main · callmesora/SpaceX-Rocket-Prediction (github.com)]()



Data Collection - Scraping

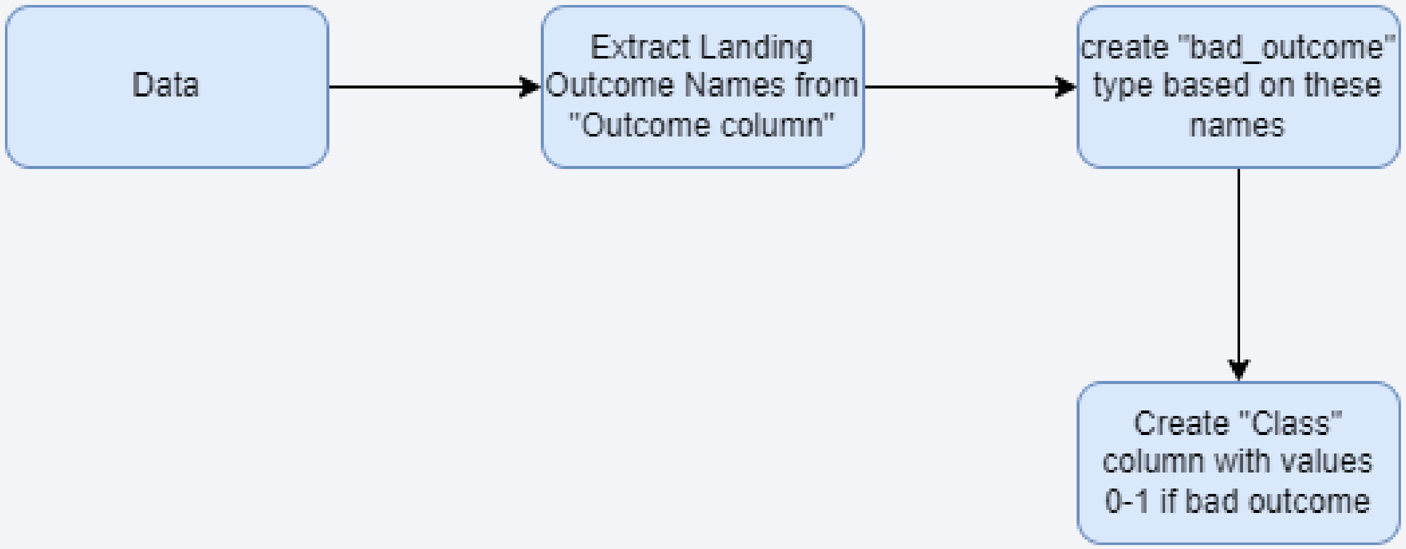
- Extracted Table data from the Falcon 9 Launch data wikipage.
- Parsed this data using BeautifulSoup and converted it to a Pandas DataFrame



- [SpaceX-Rocket-Prediction/Data Collection with Web Scrapping.ipynb at main · callmesora/SpaceX-Rocket-Prediction (github.com)]

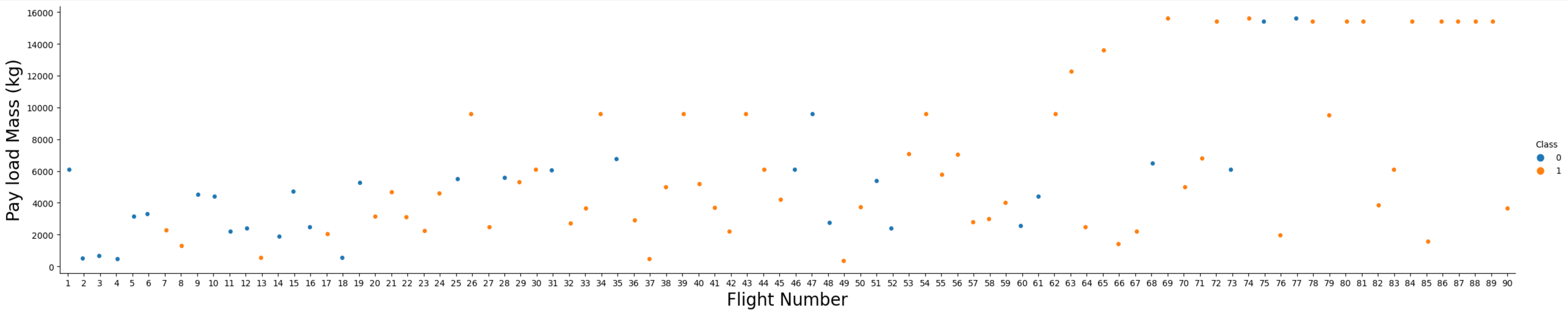
Data Wrangling

- The outcome column values where processed to create a new binary class representing successful landings with values [0,1]
- [SpaceX-Rocket-Prediction/EDA Lab.ipynb at main · callmesora/SpaceX-Rocket-Prediction (github.com)]



EDA with Data Visualization

- We see that as the flight number increases, the first stage is more likely to land successfully.
- [SpaceX-Rocket-Prediction/jupyter-labs-eda-dataviz.ipynb at main · callmesora/SpaceX-Rocket-Prediction (github.com)]



EDA with SQL

- Multiple SQL queries were performed in order to determine the relations between different variables and some statistics properties. Namely the following:
  - Launch Site names
  - All Launch sites beginning with CCA
  - Total Payload mass
  - The average payload carried by the F9 v1.1 booster
  - The first successful landing
  - Successful Drone Ship landings of a given payload
  - MIssion Success Rate
  - Boosters That carried the maximum payload
  - SpaceX-Rocket-Prediction/SQL Lab EDA.ipynb at main · callmesora/SpaceX-Rocket-Prediction (github.com)

Build an Interactive Map with Folium

- A Folium map was created with Circles surrounding the Launch Sites (to better locate them)
- Markers signifying successful/failed launches (to get a sence of the porportion)
- Lines indicating the distance between relevant landmarks such as rails/cities/coast
- SpaceX-Rocket-Prediction/lab_jupyter_launch_site_location.ipynb at main · callmesora/SpaceX-Rocket-Prediction (github.com)

Build a Dashboard with Plotly Dash

- Created plots and graphs comparing the sucessrate in different launch sites and within a given launch site.
- Analysed the success rate given different boosters and a payload range
- [SpaceX-Rocket-Prediction/spacex_dash_app.py at main · callmesora/SpaceX-Rocket-Prediction (github.com)]

Predictive Analysis (Classification)

- 4 Distinct models were created. Given these models all of them were trained using a 10 fold cross validation and a test train split of 80/20.
- Decision Tree, KNN , SVM and Logistic Classification were the models of choice
- SpaceX-Rocket-Prediction/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · callmesora/SpaceX-Rocket-Prediction (github.com)

Results

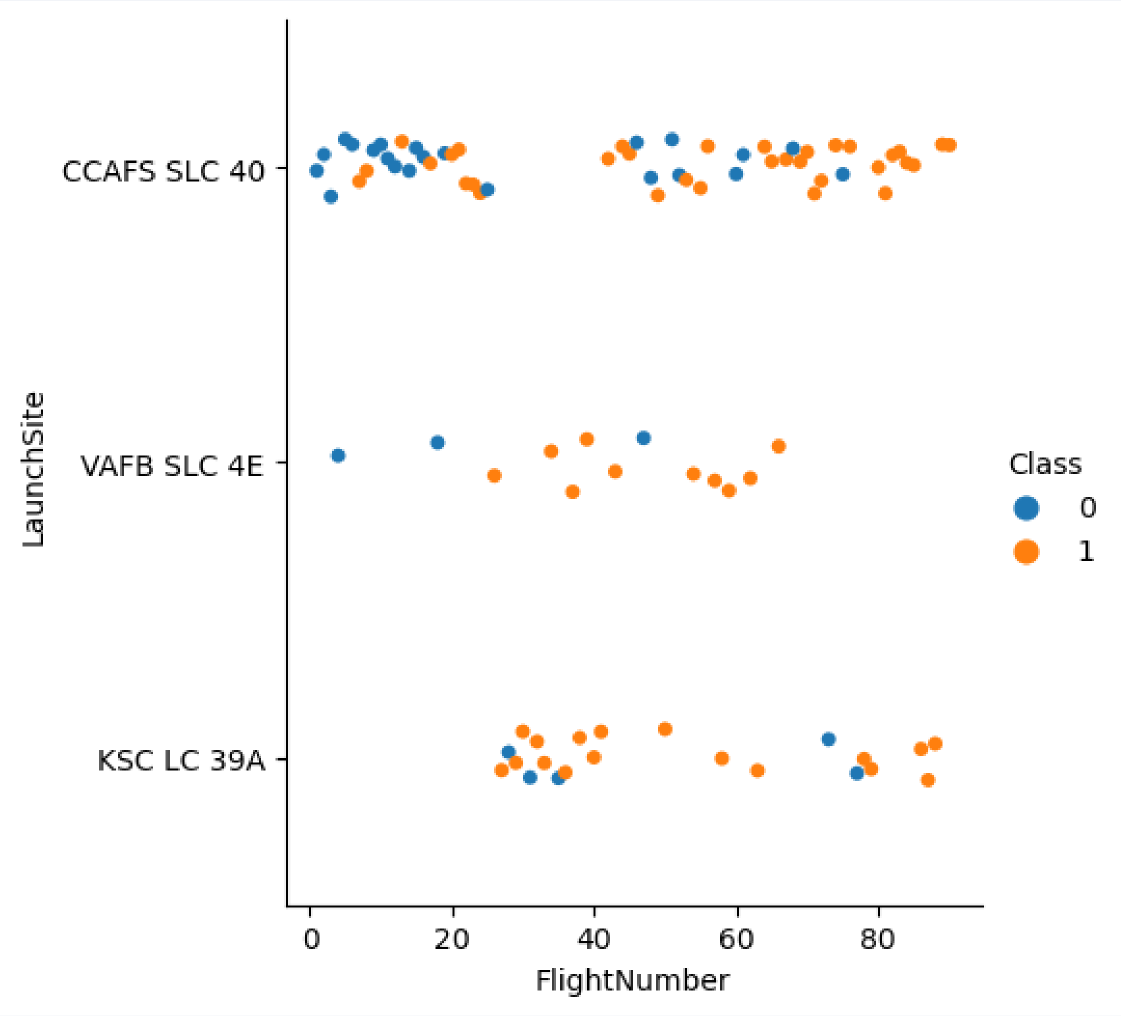
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

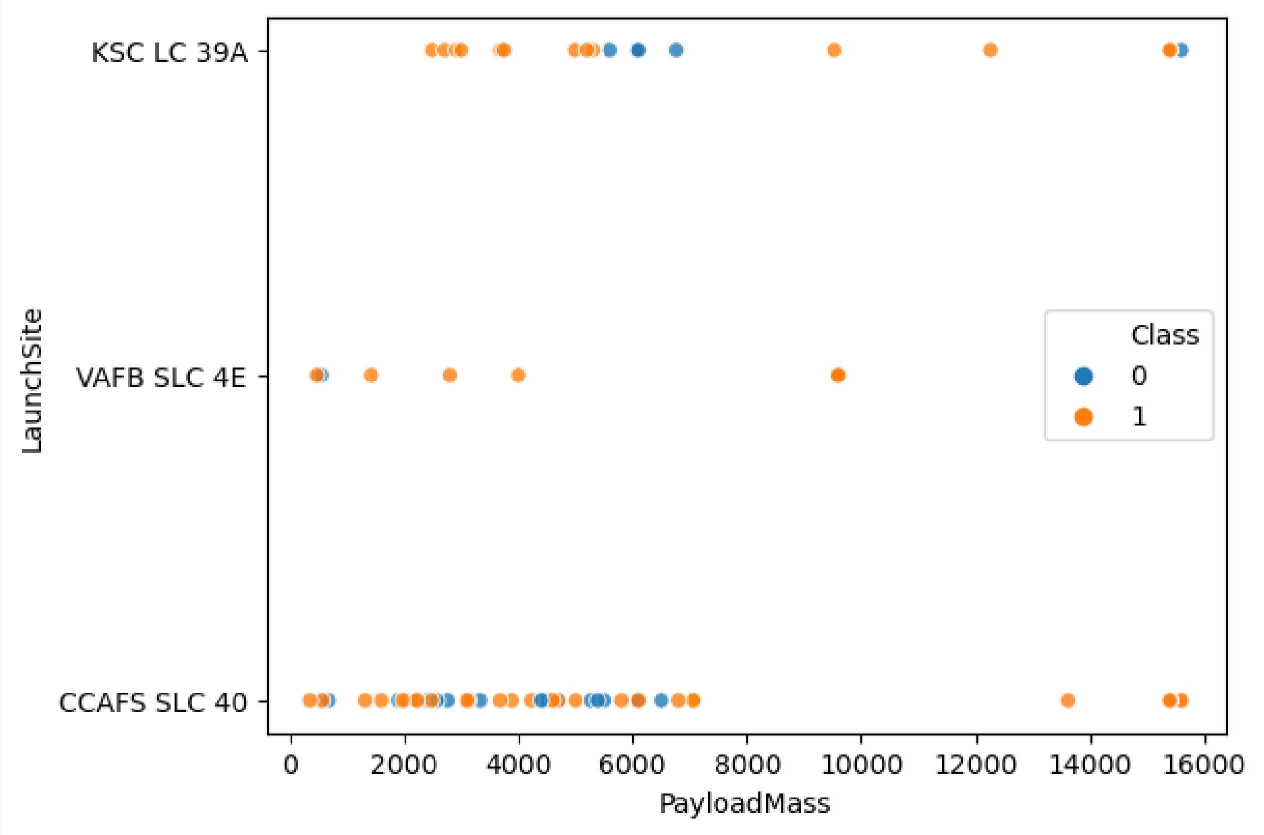
- Early flights (the first 20) were much more likely to fail then the current ones




```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue
sns.catplot(data = df , x = 'FlightNumber', y = 'LaunchSite', hue='Class')  # early flight numbers see
```

Payload vs. Launch Site

- For the VAFB-SLC launchsite there are **no rockets launched for heavypayload mass(greater than 10000).**

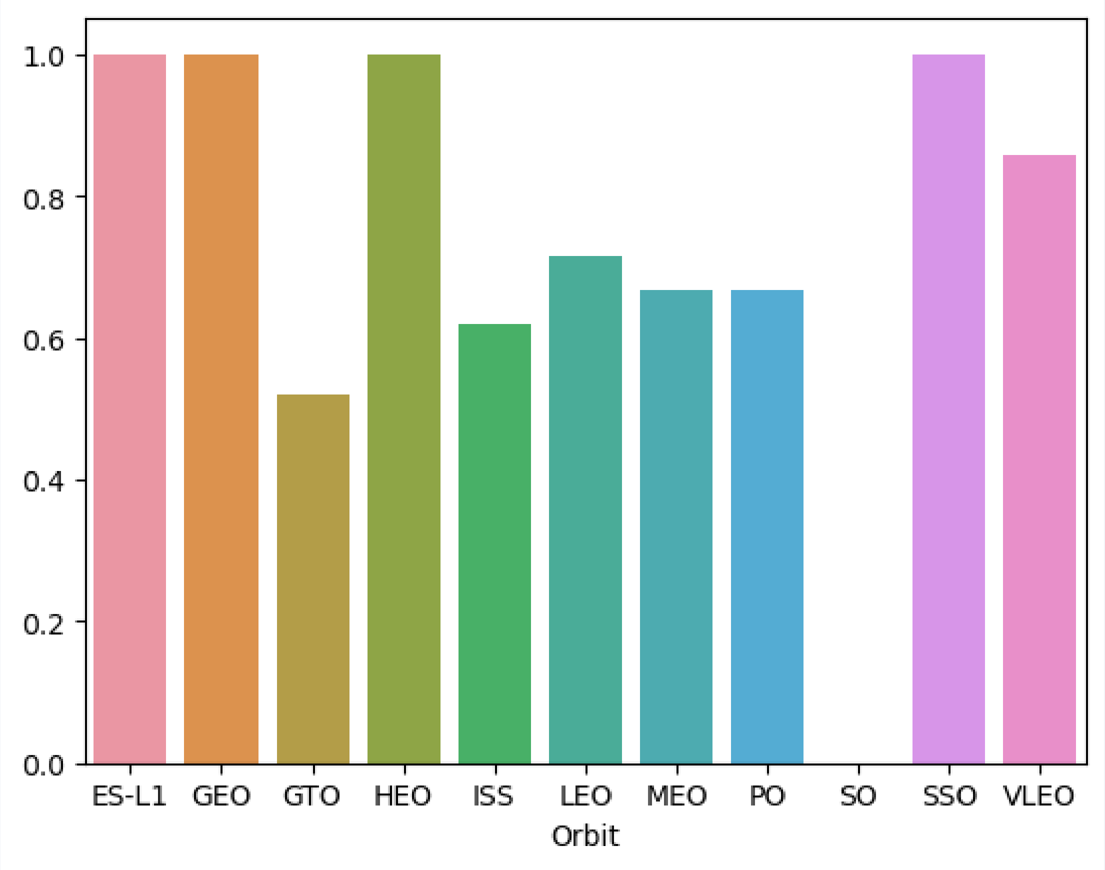



```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, a
sns.scatterplot(data = df , x = 'PayloadMass', y = 'LaunchSite', hue = 'Class', alpha=0.8)
```

Success Rate vs. Orbit Type

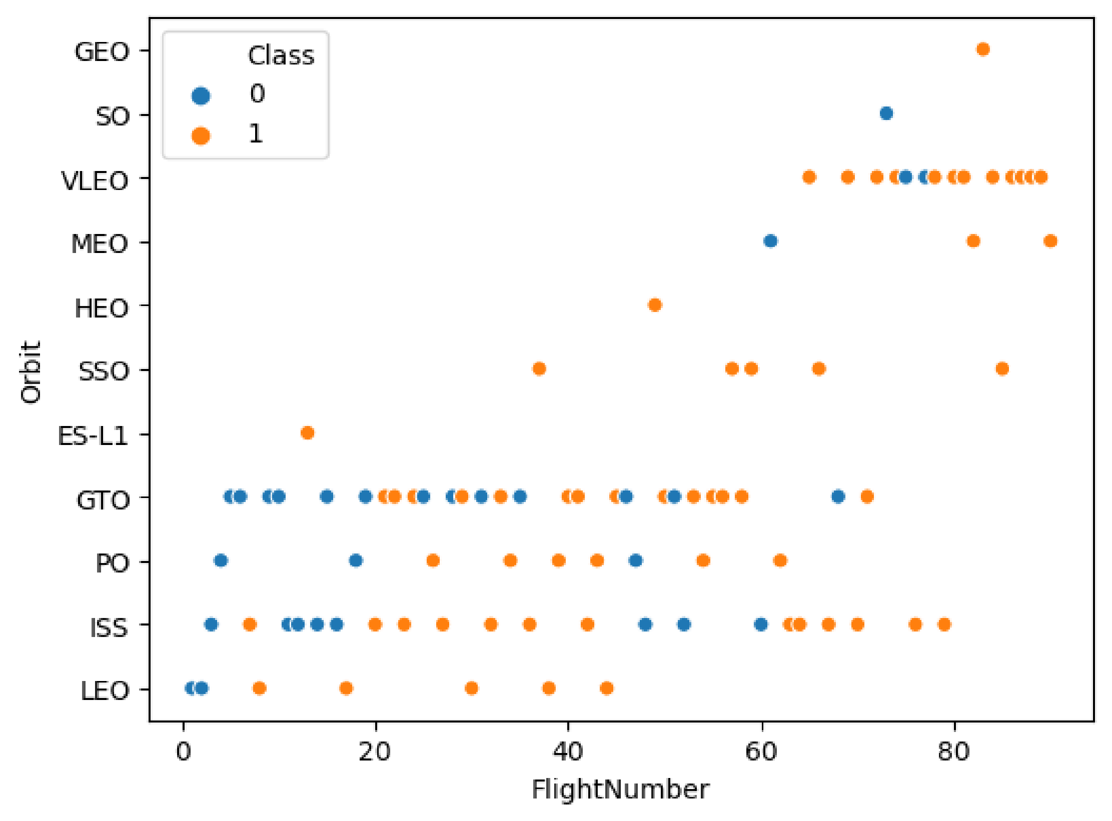
- GTO, ISS are the two lowest success rate orbits
- ES-L1, GEO, HEO, SSO are the ones with the highest success rate

```
# HINT use groupby method on Orbit column and get the mean of Class column
bar_data = df.groupby('Orbit')['Class'].mean()
sns.barplot(bar_data.index,bar_data.values)
```



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.

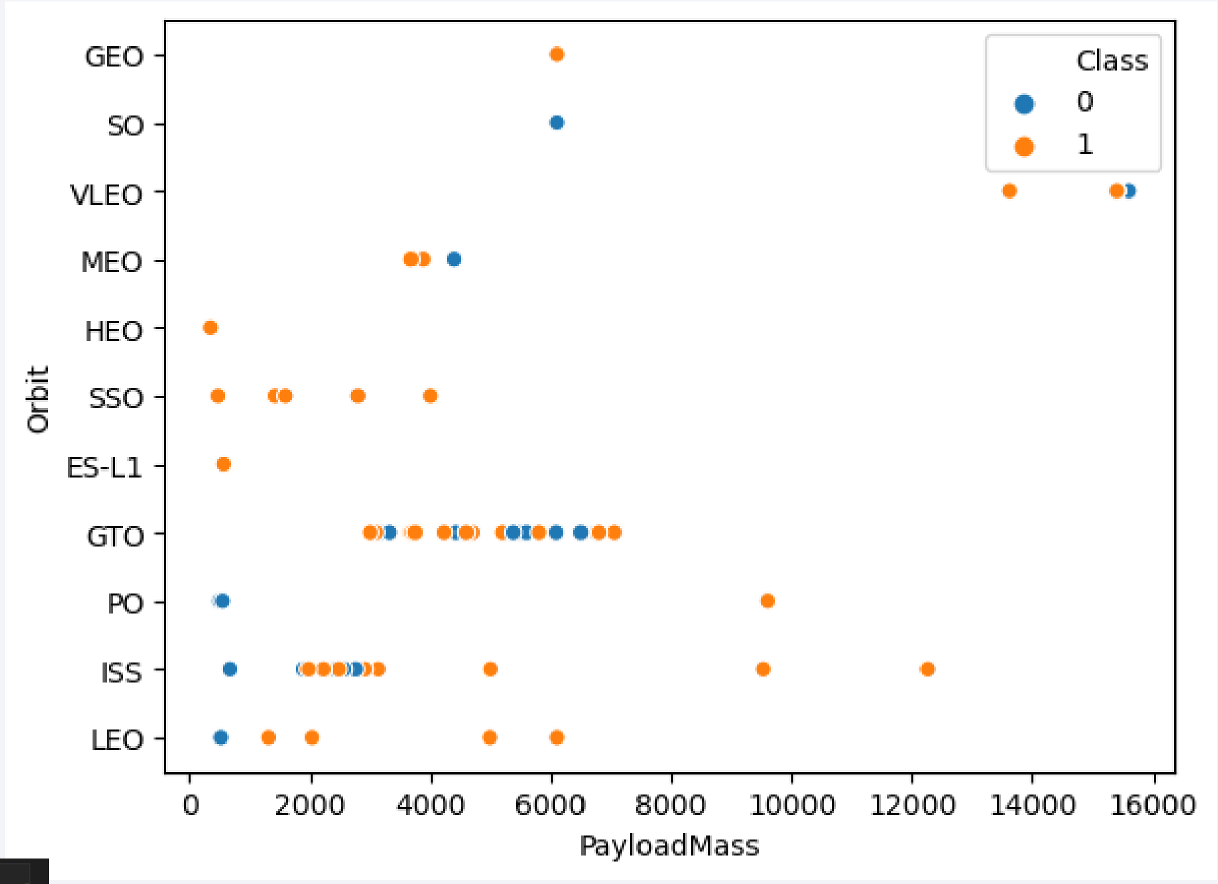



```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Or
sns.scatterplot(data= df , x='FlightNumber', y = 'Orbit', hue = 'Class')
```

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.




```python
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class val
sns.scatterplot(data = df, x = 'PayloadMass', y='Orbit', hue='Class')
```

Python

```
<AxesSubplot:xlabel='PayloadMass', ylabel='Orbit'>
```

Launch Success Yearly Trend

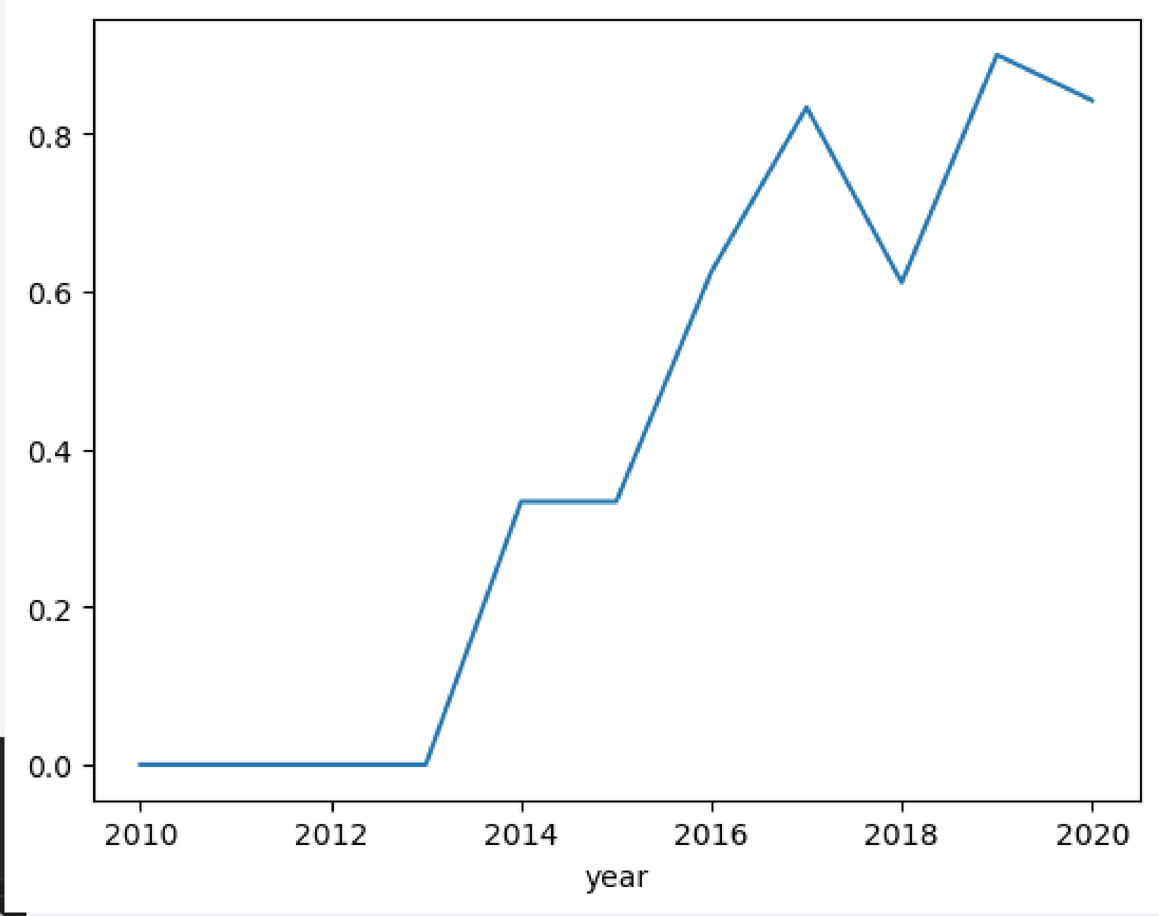
- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
df['year'] = pd.DatetimeIndex(df['Date']).year
year_sucess = df.groupby('year')['Class'].mean()

sns.lineplot(x=year_sucess.index, y= year_sucess.values) #sucess rateincreases

<AxesSubplot:xlabel='year'>
```



All Launch Site Names

```
%%sql
SELECT DISTINCT(launch_site) FROM SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
%%sql
SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5
```

16]

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Summed the total payload

```
%%sql
SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE customer = 'NASA (CRS)'
```

5]

1
22007

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Filtered where the version had 'F9 v1.1' in the name

```
%%sql
SELECT AVG(payload_mass__kg_) FROM SPACEXTBL WHERE booster_version LIKE 'F9 v1.1%'
```

1
3226

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Selected the minimum date where the landing outcome was sucessfull

```
%%sql
SELECT MIN(DATE) FROM SPACEXTBL WHERE landing__outcome LIKE 'Success (ground pad)'
```

```
2017-01-05
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 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
```

booster_version
F9 FT B1022
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%%sql
SELECT COUNT(*) FROM SPACEXTBL GROUP BY mission_outcome
```

1
44
1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Filtered by payload mass with subquery to find out the max payload_mass

```
%%sql
SELECT DISTINCT(booster_version) FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1058.3
F9 B5 B1060.2

2015 Launch Records

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

```
%%sql
SELECT booster_version, launch_site, landing__outcome FROM SPACEXTBL WHERE DATE LIKE '2015%'
```

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1013	CCAFS LC-40	Controlled (ocean)
F9 v1.1 B1014	CCAFS LC-40	No attempt

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(*)  FROM SPACEXTBL  WHERE DATE BETWEEN '2010-06-04' and '2017-03-20'  GROUP BY landing__outcome   ORDER BY COUNT(*)
```
Python

landing__outcome	2
No attempt	7
Failure (drone ship)	2
Success (drone ship)	2
Success (ground pad)	2
Controlled (ocean)	1
Failure (parachute)	1

Section 3

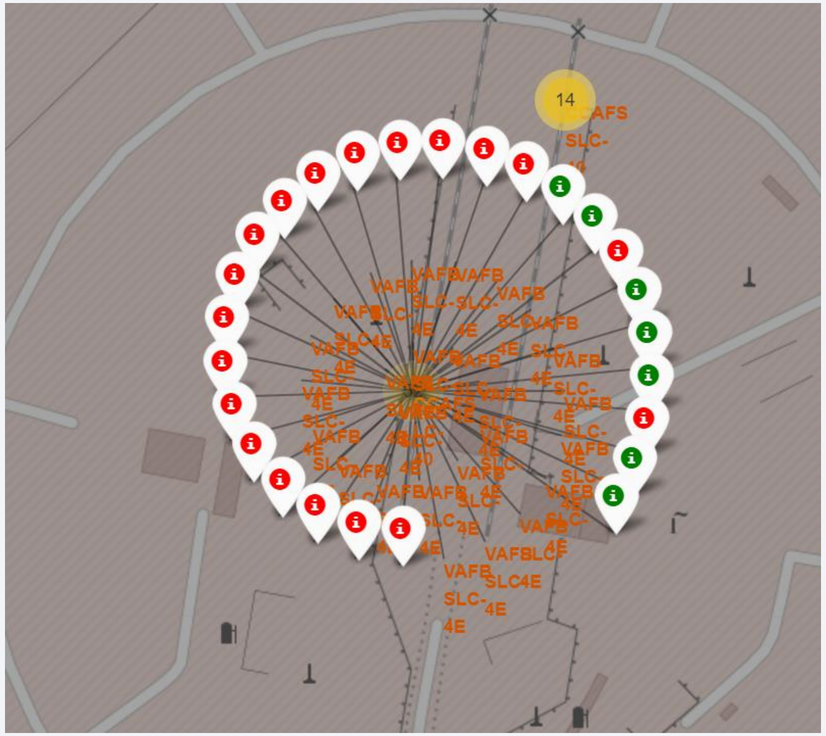
Launch Sites Proximities Analysis

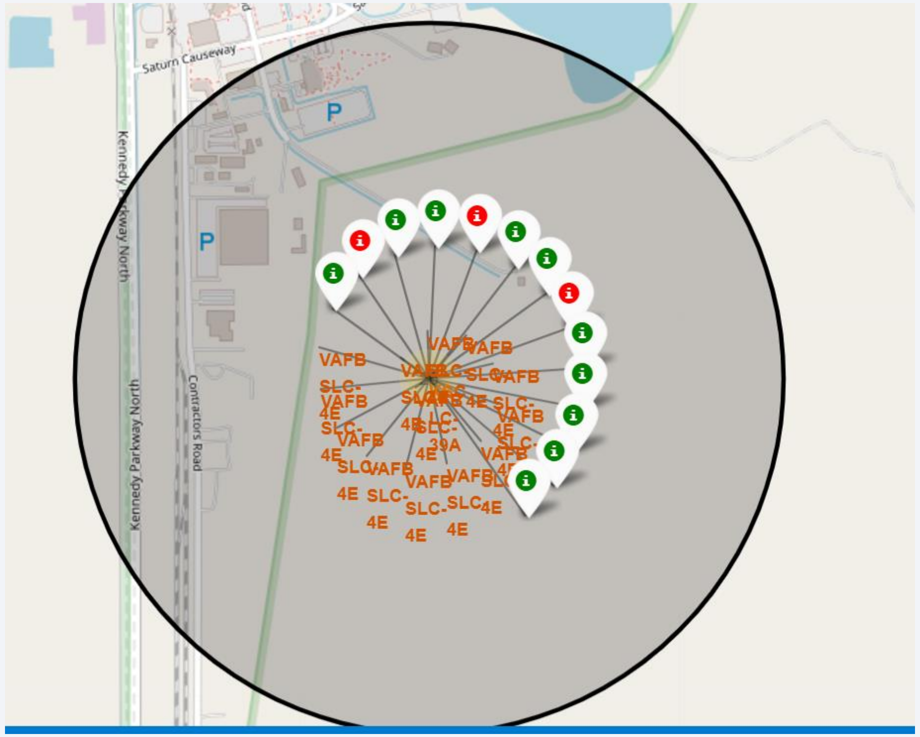
SPACEX Launch Sites in USA

- All the launch site locations are situated with a similar latitude (closer to the equator)

Success Rate by Location

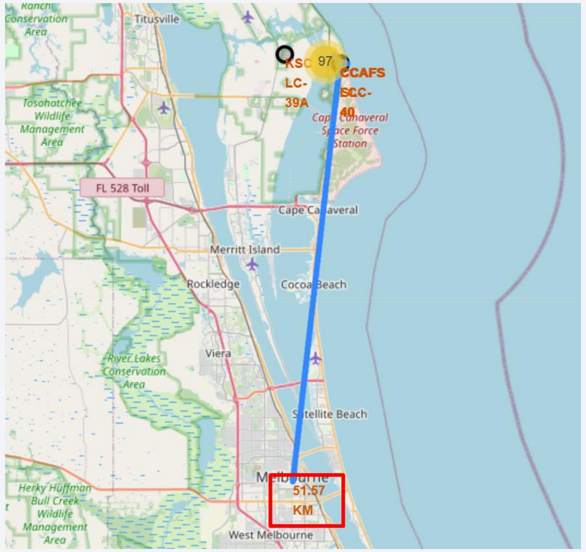
- KSC have a much higher success rate compared to CCAs..
- The success rate varies considerably from location

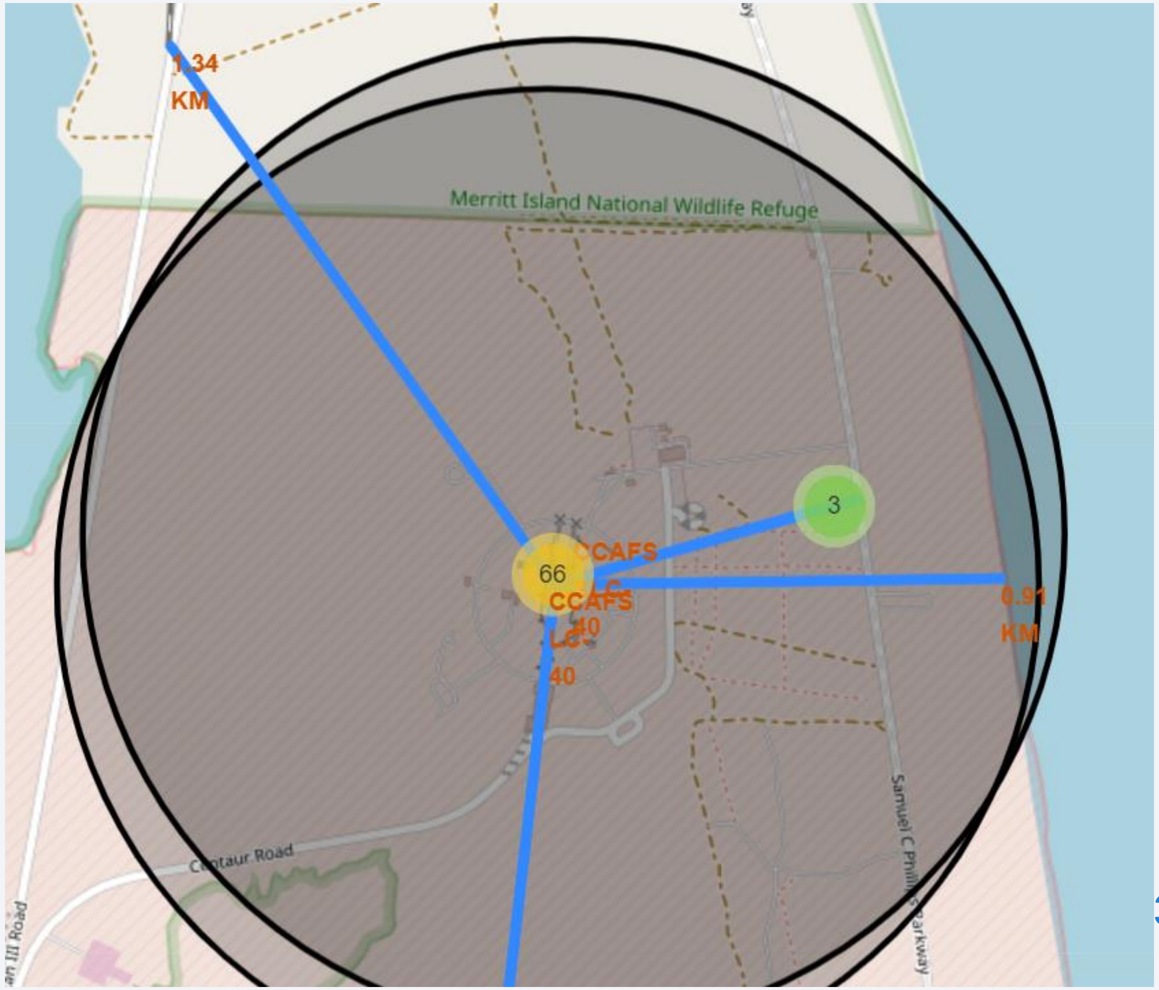




Launch site proximity to cities/coast/rails

- The main launch location is situated:
- 1.34 KM away from the closest rail.
- 6.93 KM away from the coast
- 51.57 KM away from the closest city



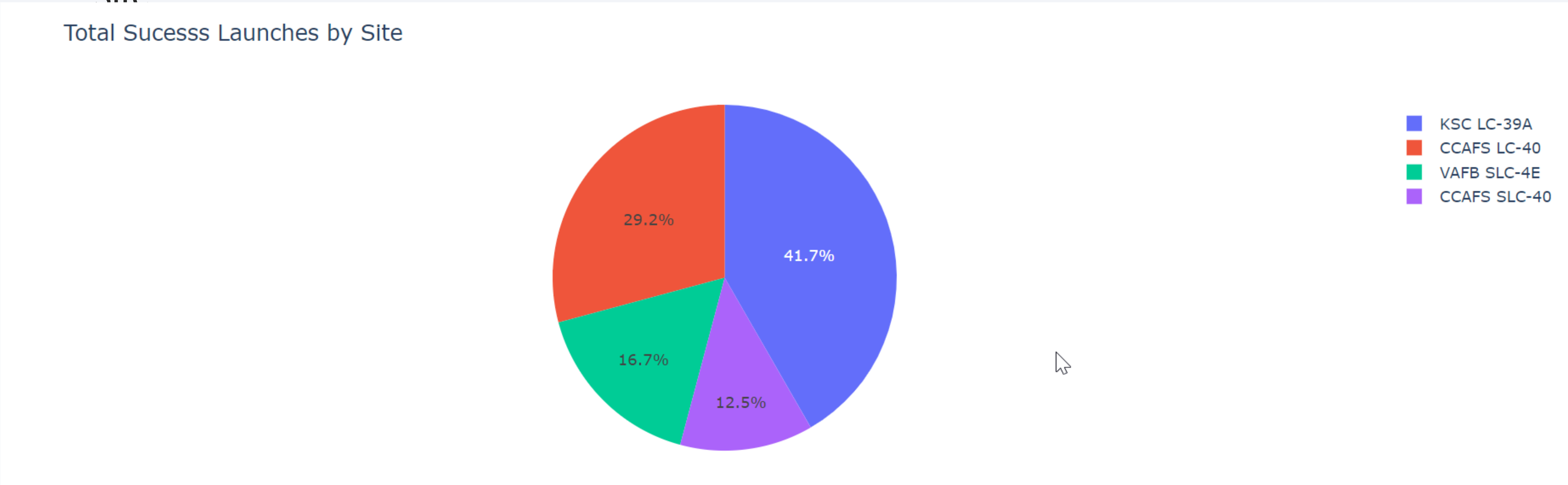


Section 4

Build a Dashboard with Plotly Dash

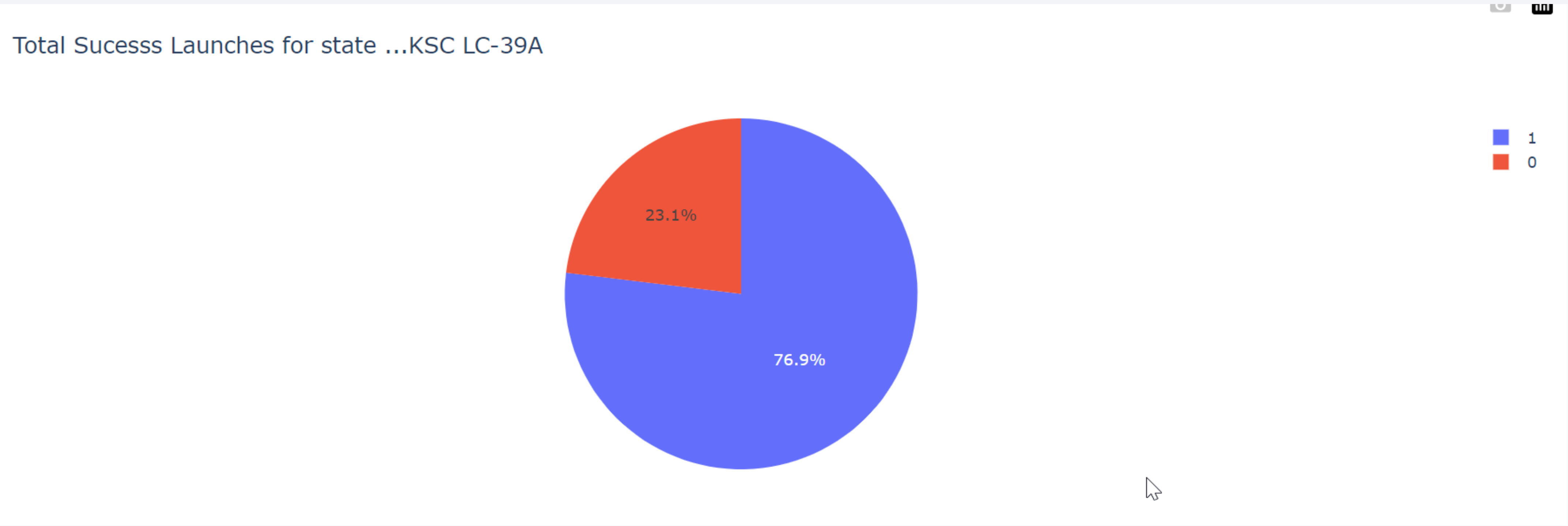
Successful Launches by site

- Vast majority of the successful launches were performed on the KSC-LC site



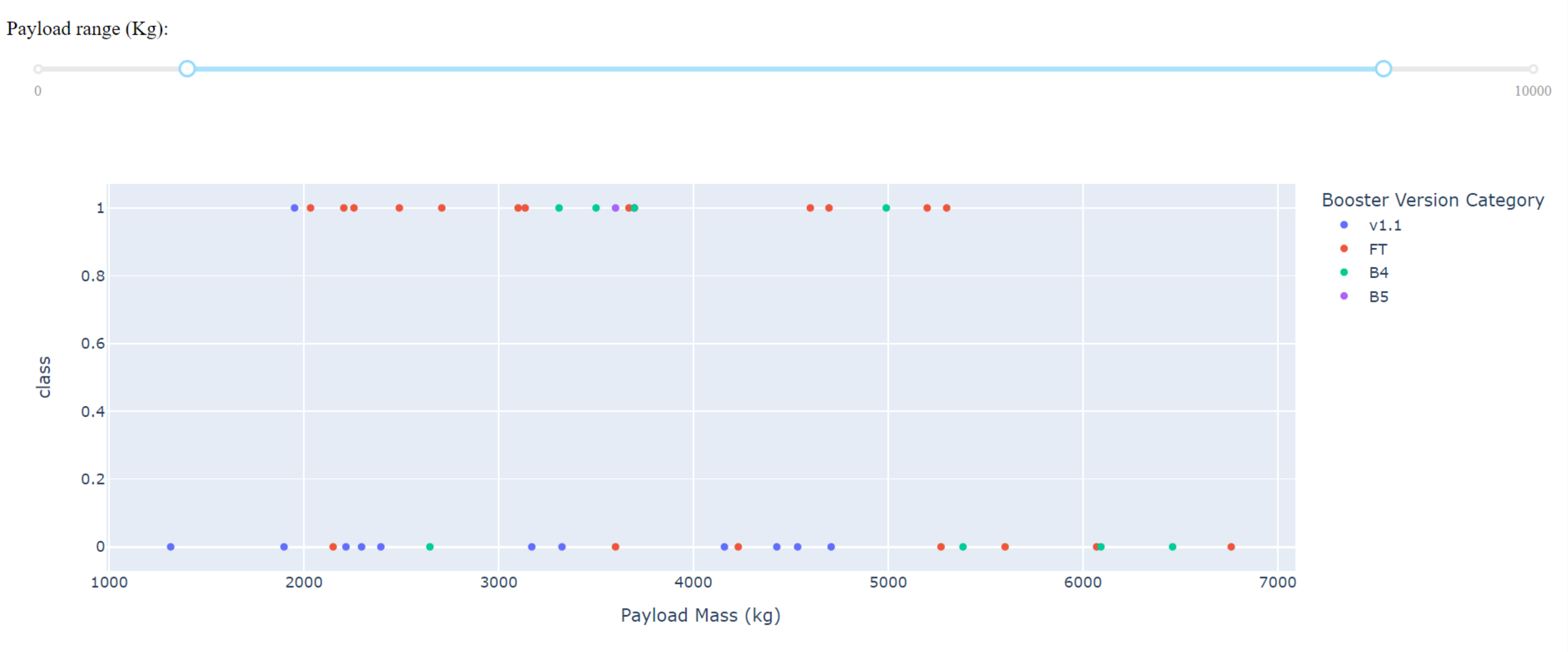
The launch site with the highest success rate

- Over ¾ of the launches in KSC-LC were successful launches



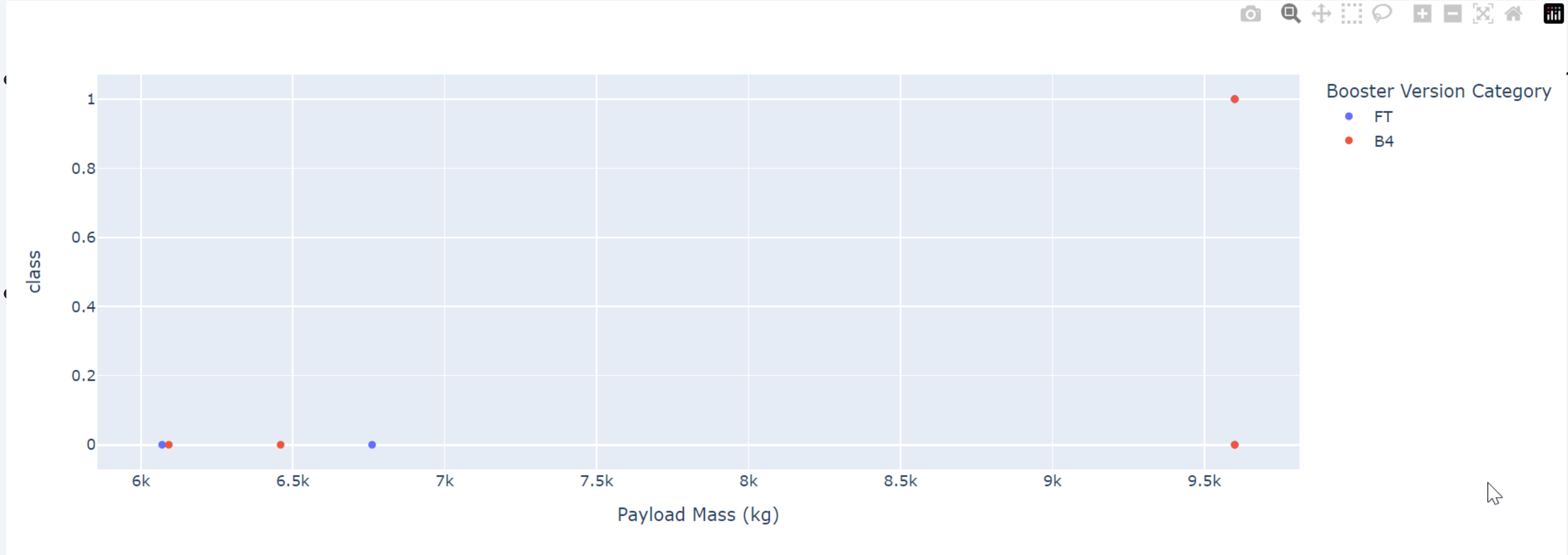
Booster version success rate analysis

- V1.1 Seems to be the booster that fails more



Which boosters are used in higher payloads

- For higher payloads only FT and B4 Boosters seem to be used



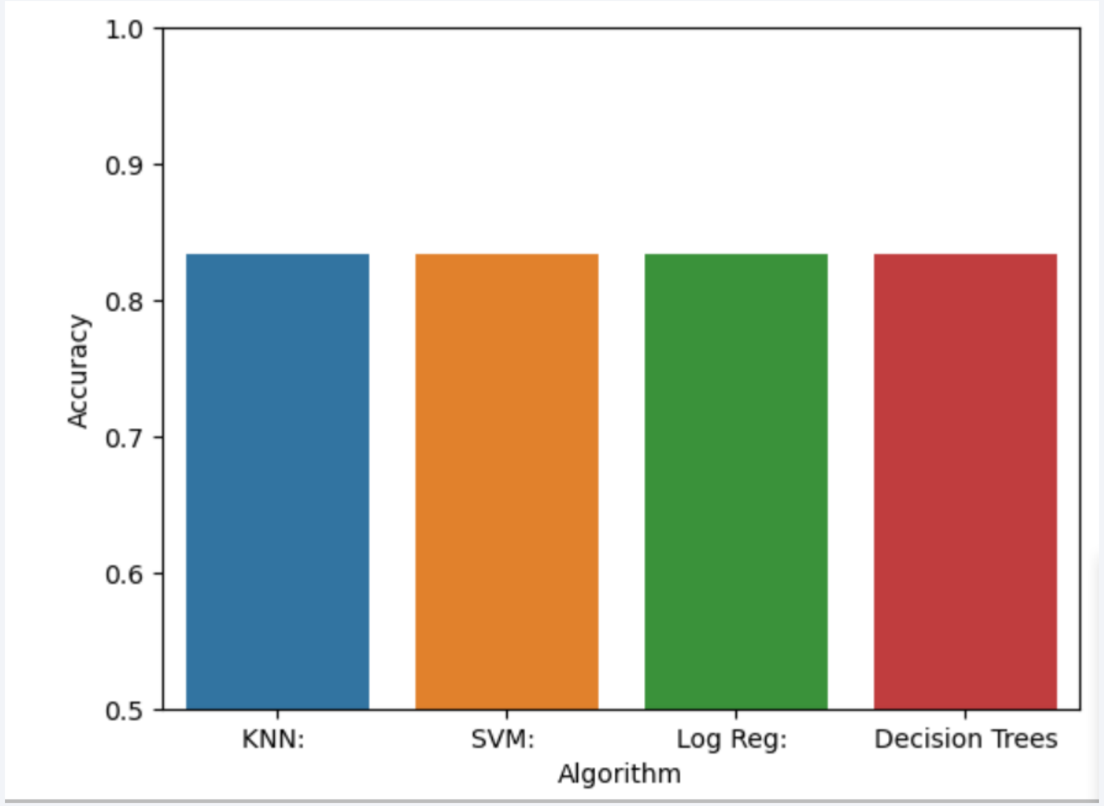
Section 5

Predictive Analysis (Classification)

Classification Accuracy

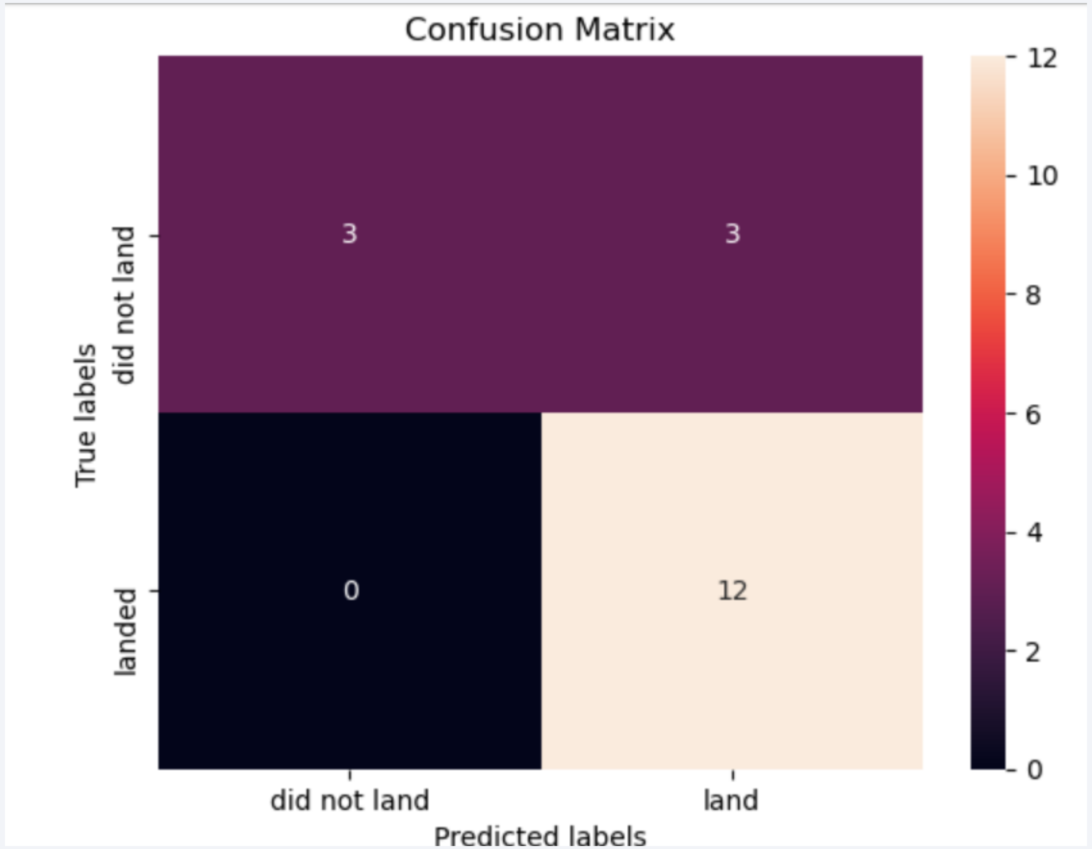
- All the different models perform with quite a similar metrics

	Algorithm	Accuracy
0	KNN:	0.833333
1	SVM:	0.833333
2	Log Reg:	0.833333
3	Decision Trees	0.833333



Confusion Matrix

- 3 Corrects are predicted incorrectly has a successful landing.
- Given that all models perform the same the confusion matrix is identical



Conclusions

- KSC-LC is the highest success rate launch site
- The models can predict a successful landing with up to 83% accuracy
- SpaceX launch sites are located relatively nearby the coast and equator
- Booster V1.1 seems to fail a lot for the time being
- Location site, payload size and booster seem have high importance in predicting the success of the launch

Thank you!