IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Li Yanling
10 June 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection via SpaceX REST API and scraping of Wiki page
  - Data wrangling and exploratory data analysis using Python and SQL via sqlite
  - Interactive visualization on a web browser with Plotly & Dash
  - Geospatial visualization with Folium
  - Classification prediction models tested with logistic regression, support vector machine, decision tree, and k-nearest neighbour
- Summary of all results
  - Launch Site KSC LC-39A has the highest rate of successful launches (76.9%).
  - The success rate of launches generally rose over the years; flights before 2014 had 0% success rate while flights in 2019 and 2020 have about 80% success rate.
  - All 4 prediction models were able to predict successful launches with 83% accuracy based on the training data. 83 columns were used as independent variables.

Introduction

- Project background and context:
  - Space X 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 Space X can reuse the first stage.
  - By studying the historical data of past launches, we can analyse the factors affecting the launch outcome and estimate the success rate of future launches.
  - This can help SpaceX to improve their launch success rates in future launches and understand their average cost per successful launch.
  - The analysis can also help SpaceX's competitors to determine the cost and success rate required to bid against SpaceX.
- Questions to be answered:
  - Which parameters affect the success rate of a launch?
  - How has the success rate changed over the years?
  - Which prediction model can be used to predict the success rate of a launch before it is performed?

Section 1

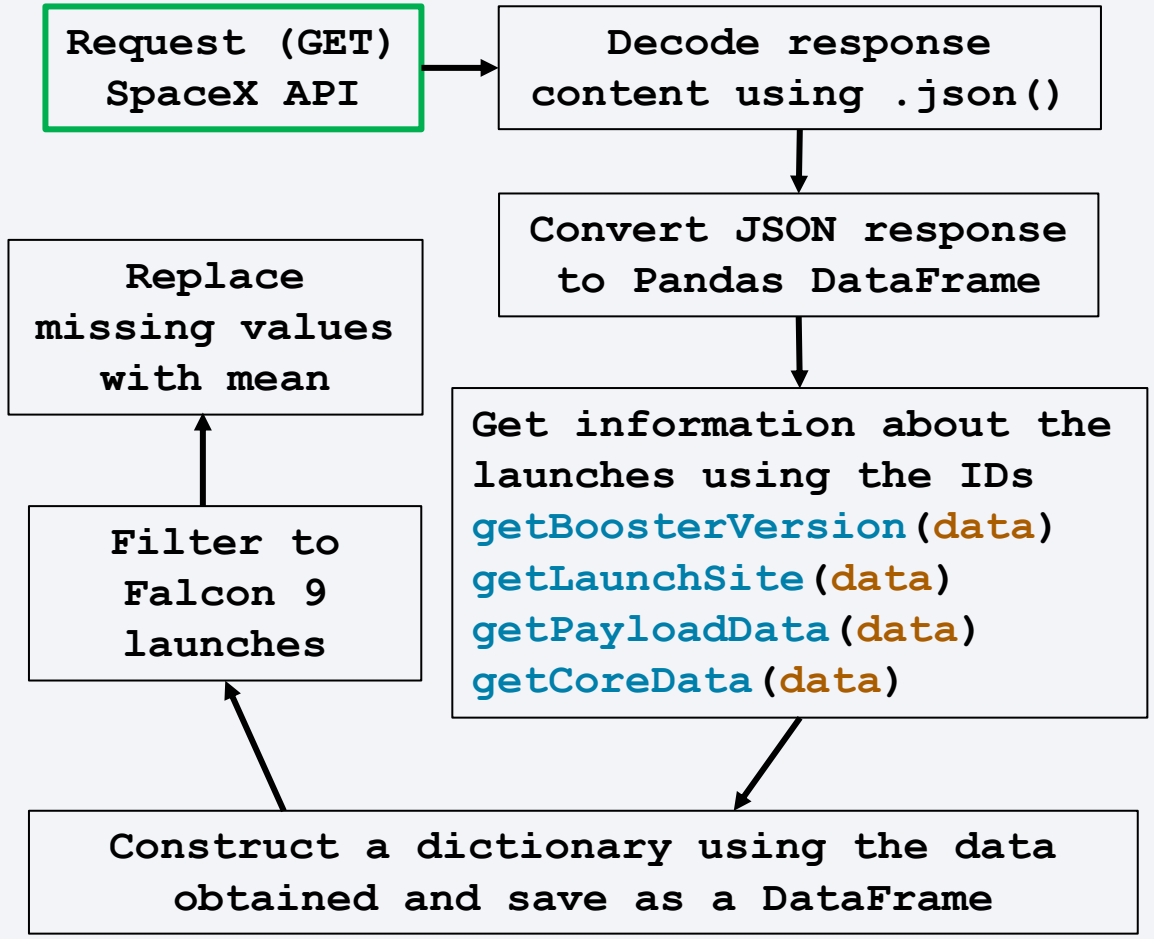
Methodology

Methodology

- Data collection methodology:
    1. SpaceX REST API
    2. Scraping the Falcon 9 Launch Wiki Page
- Perform data wrangling
    - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Build, tune, evaluate classification models

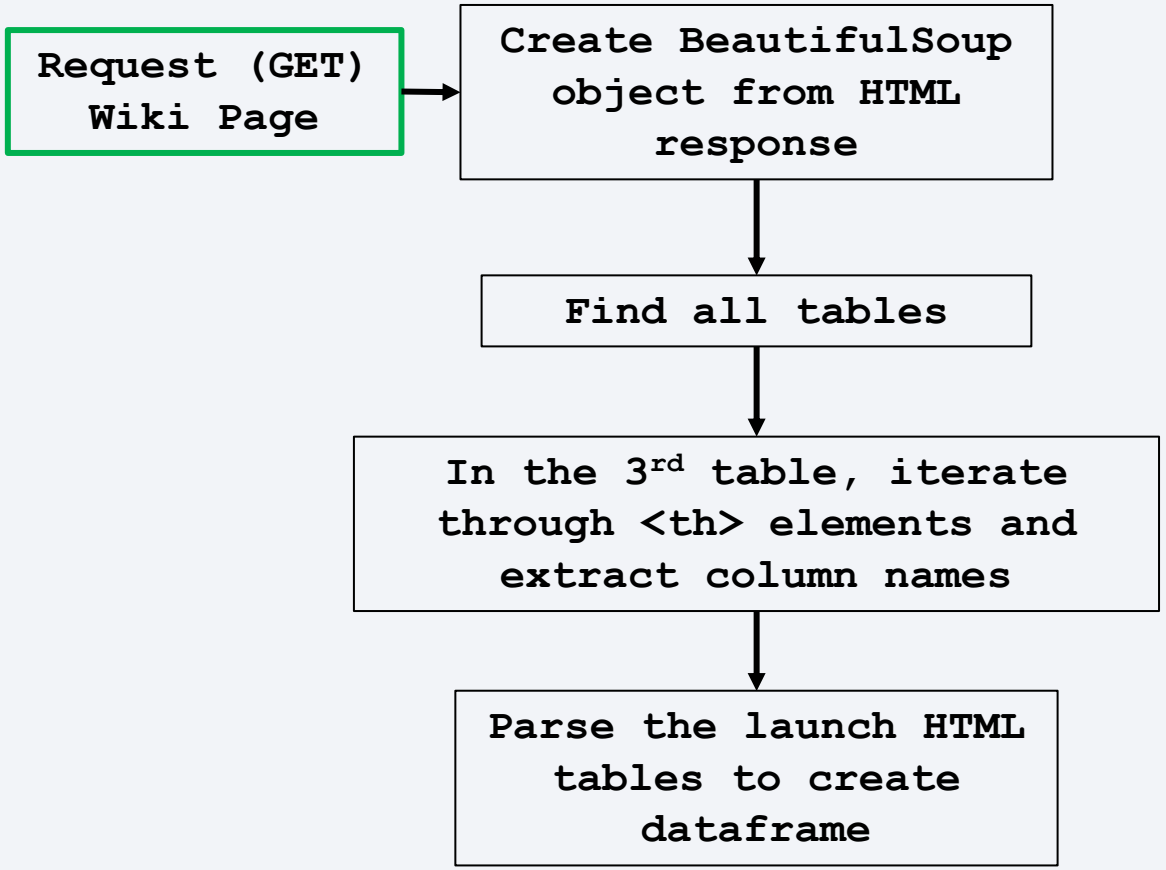
Data Collection – SpaceX API

- The final output includes:
  - *FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude*
- GitHub URL:
  - https://github.com/yanling-yl/IBMDataScienceLab/blob/main/Capstone_Project/jupyter-labs-spacex-data-collection-api.ipynb



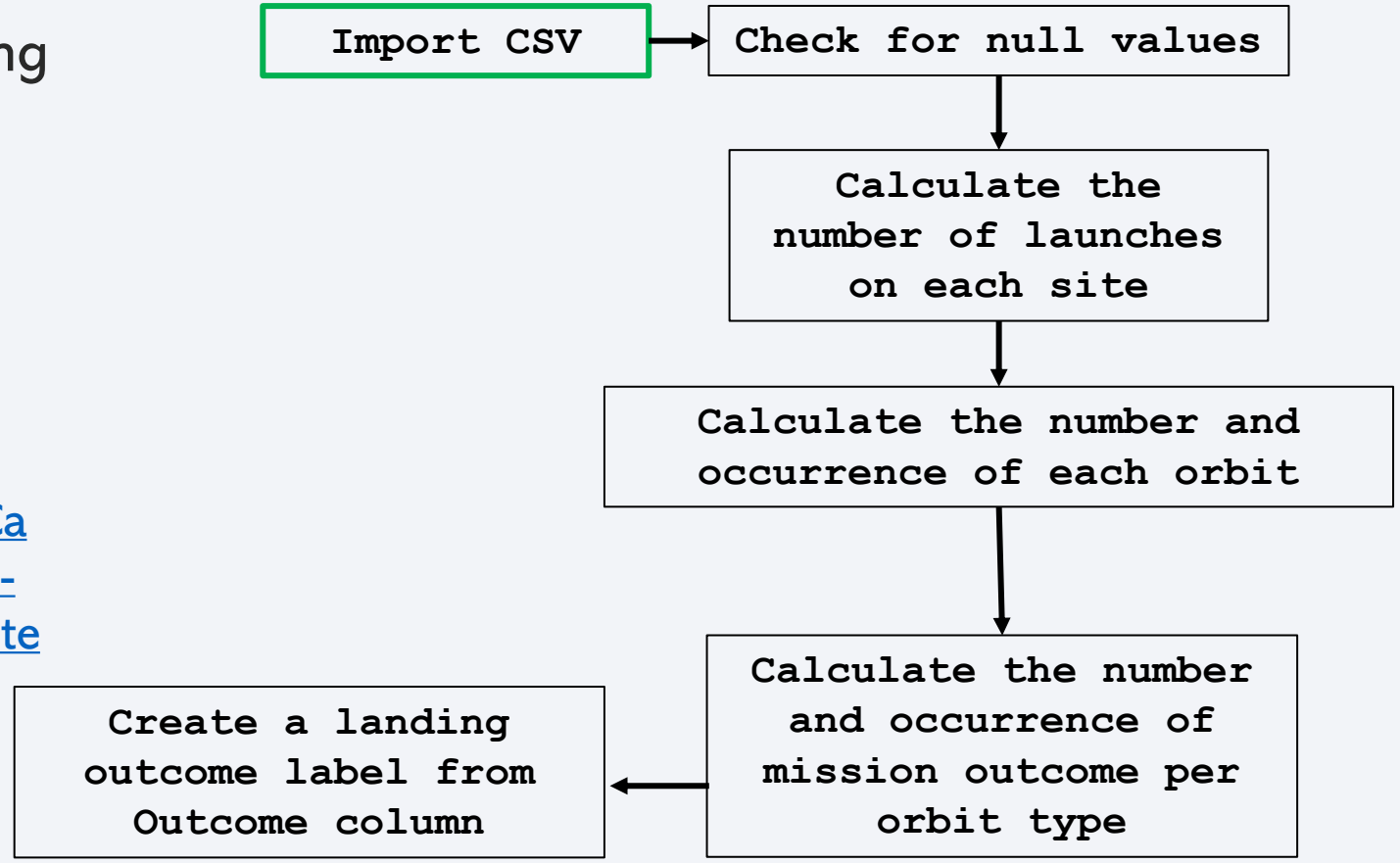
Data Collection – Scraping Wiki

- The final output includes:
  - *Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Flight No., Version Booster, Booster landing, Date, Time*
- GitHub URL:
  - https://github.com/yanling-yl/IBMDataScienceLab/blob/main/Capstone_Project/jupyter-labs-webscraping.ipynb



Data Wrangling

- The final output is a table showing the number of successful/failed launches for each orbit type and each launch site.
- GitHub URL:
  - https://github.com/yanling-yl/IBMDataScienceLab/blob/main/Capstone_Project/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite%20(1).ipynb



Data Wrangling

Final Output

```
df_pivot = pd.pivot_table(df, values='BoosterVersion', index='LaunchSite', columns=['Orbit', 'Class'],
                          aggfunc='count')

df_pivot
```

Orbit	ES-L1	GEO		GTO	HEO		ISS		LEO		MEO		PO	SO	SSO		VLEO
Class	1	1	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1
LaunchSite																	
CCAFS SLC 40	1.0	1.0	10.0	8.0	1.0	8.0	8.0	2.0	3.0	1.0	2.0	NaN	NaN	NaN	1.0	1.0	8.0
KSC LC 39A	NaN	NaN	3.0	6.0	NaN	NaN	5.0	NaN	2.0	NaN	NaN	NaN	NaN	1.0	NaN	1.0	4.0
VAFB SLC 4E	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0	6.0	NaN	4.0	NaN	NaN

EDA with Data Visualization

- Charts plotted for data exploration:
    1. Flight Number vs. Launch Site:
    2. Launch Site vs. Payload Mass
    3. Orbit Type vs. Class (Success Rate)
    4. Flight Number vs. Orbit Type
    5. Payload Mass vs. Orbit Type
    6. Date vs. Class (Success Rate)
- GitHub URL:
    - *https://github.com/yanling-yl/IBMDataScienceLab/blob/main/Capstone_Project/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb*

EDA with SQL

- Exploratory SQL Queries for a basic understanding of the data
  - Unique launch sites, Sites starting with “CCA”, Payload for boosters launched by NASA (CRS)
  - Average payload mass for booster version F9 v1.1, Date of the first successful landing in ground pad
- Aggregations to statistically understand the historical data
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcome
  - List the names of the booster_versions which have carried the maximum payload mass.
  - 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.
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- GitHub URL:
  - https://github.com/yanling-yl/IBMDataScienceLab/blob/main/Capstone_Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map objects were used to visualize launch sites on a map:
  - Markers, circles (aggregations), and lines to a specific location (highway).
  - Maps are interactive and can be zoomed in to have close-up views of specific launch sites.
- GitHub URL:
  - https://github.com/yanling-yl/IBMDataScienceLab/blob/main/Capstone_Project/lab_jupyter_launch_site_location.jupyterlite.ipynb

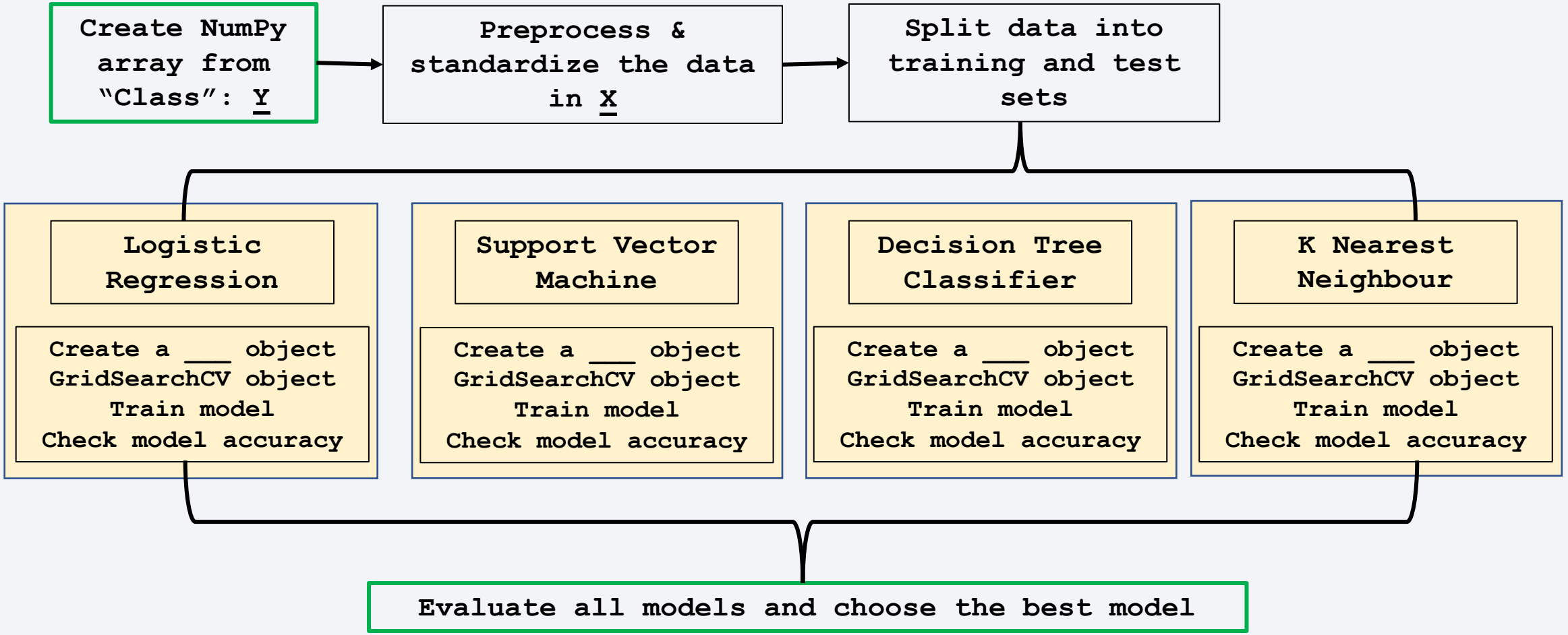
Build a Dashboard with Plotly Dash

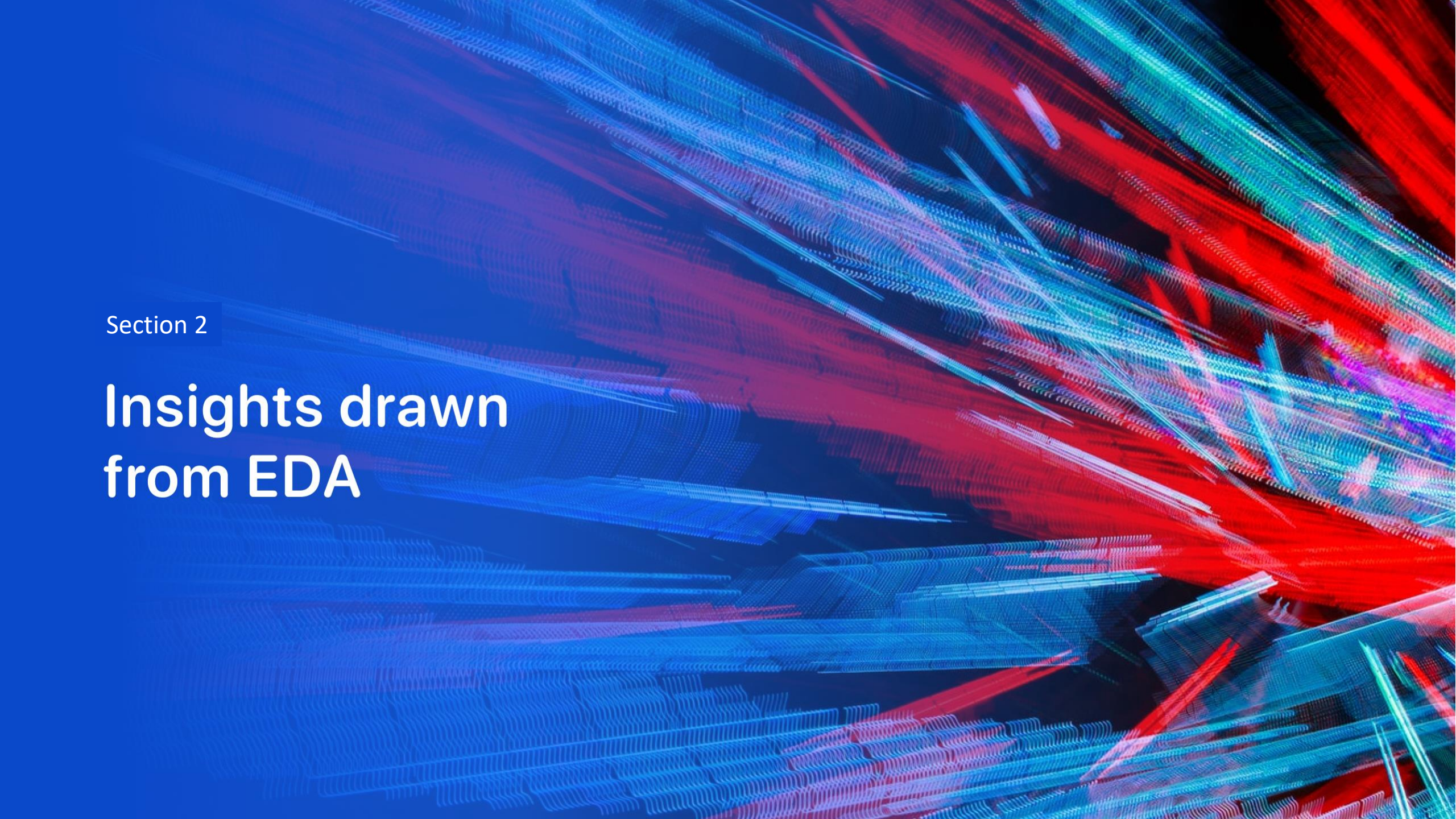
- The interactive charts can help to quickly toggle the visualizations of the launches, successes, and payloads between different launch sites:
  - Success Count for all launch sites and each launch site
  - Payload Mass by Class, including a slider bar to select the payload range
- GitHub URL:
  - https://github.com/yanling-yl/IBMDataScienceLab/blob/main/Capstone_Project/spacex_dash_app.py

Predictive Analysis

- Data was standardized, transformed into NumPy arrays, and split into test/trainsets
- 4 models were tested and the optimal hyperparameters are found using GridSearchCV
  - Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbour
- The flowchart and summary are presented in the next chart
- GitHub URL:
  - https://github.com/yanling-yl/IBMDataScienceLab/blob/main/Capstone_Project/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Predictive Analysis Flow Chart

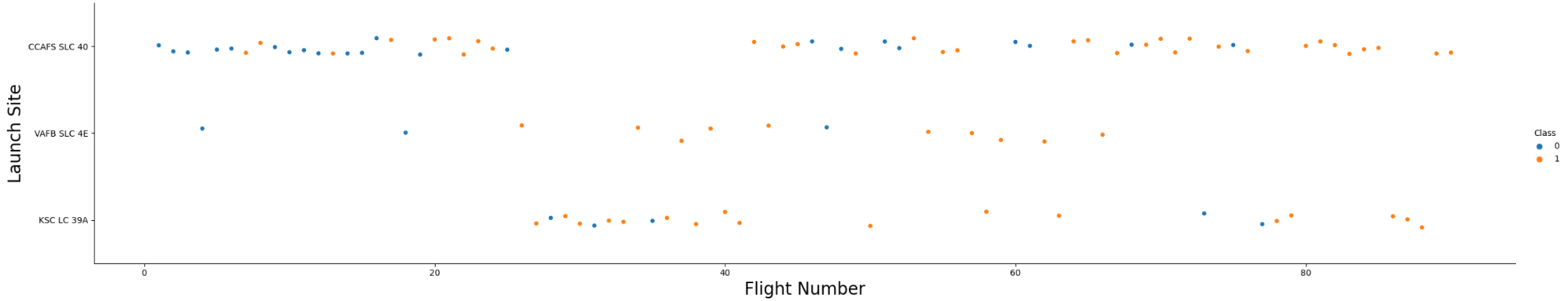



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Flight Number vs. Launch Site
  - CCAFS SLC 40 has the highest number of flights, especially for earlier flights

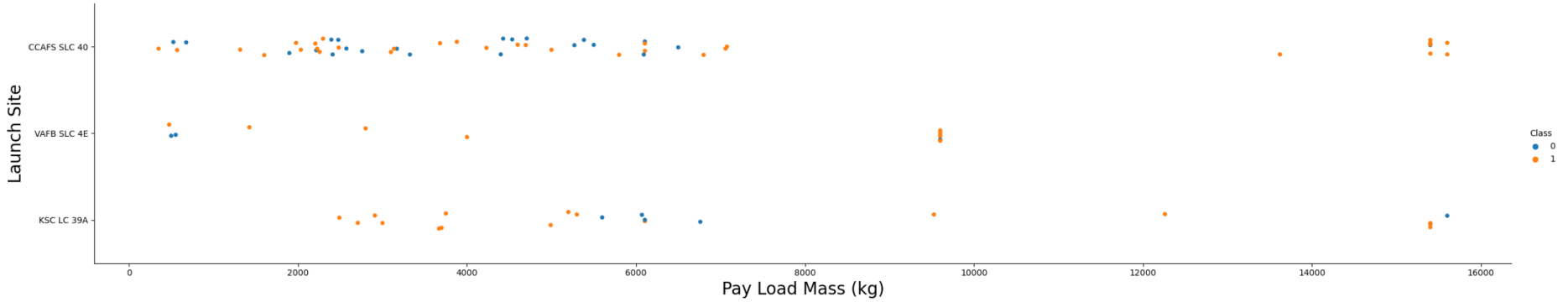
```
### TASK 1: Visualize the relationship between Flight Number and Launch Site
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



Payload vs. Launch Site

- Payload vs. Launch Site
  - Most flights have a payload of less than 8000kg
  - VAFB SLC 4E does not have flights with payload >10000kg

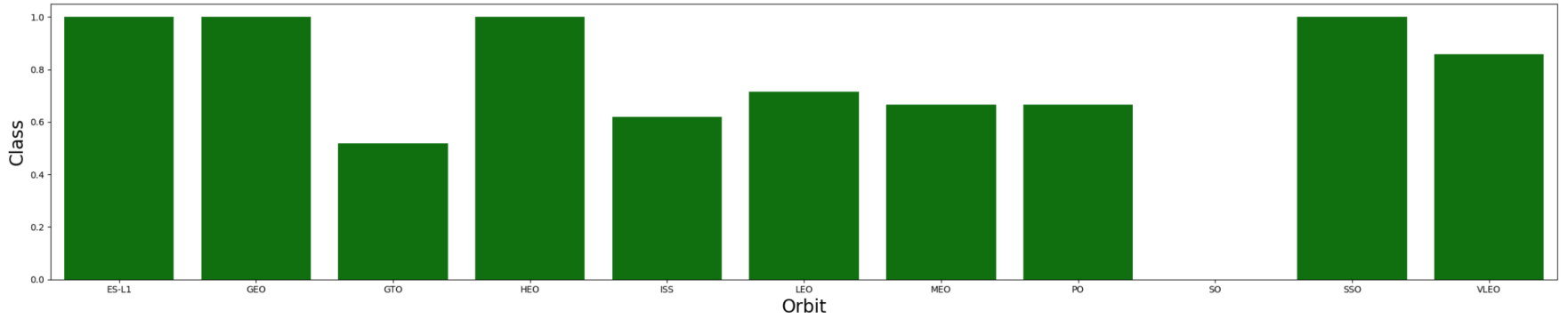
```
### TASK 2: Visualize the relationship between Payload and Launch Site
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



Success Rate vs. Orbit Type

- Success Rate vs. Orbit Type
  - Flights with ES-L1, GEO, HEO, and SSO has 100% success rate. GTO flights have the lowest success rate.

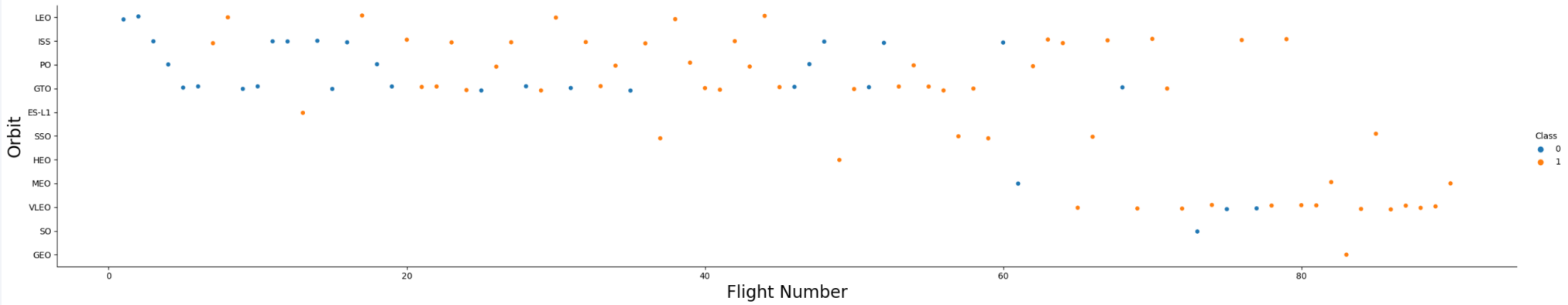
```
### TASK  3: Visualize the relationship between success rate of each orbit type
class_mean_by_orbit_df = df[['Orbit', 'Class']].groupby('Orbit').mean().reset_index()
sns.barplot(data=class_mean_by_orbit_df, x="Orbit", y="Class", color = 'green')
plt.xlabel("Orbit",fontsize=20)
plt.ylabel("Class",fontsize=20)
plt.show()
```



Flight Number vs. Orbit Type

- Flight Number vs. Orbit Type
  - Orbit types like LEO, ISS, PO and GTO are commonly used for the earlier launches.
  - VLEO is a frequently used orbit type for later flights.

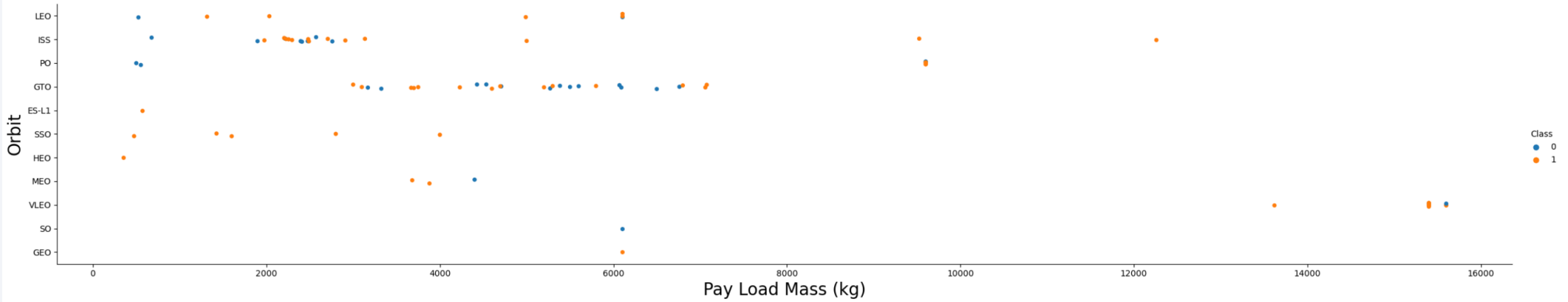
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the c
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Payload vs. Orbit Type

- Payload Mass vs. Orbit Type
  - The flights with the highest payload mass are VLEO orbit flights

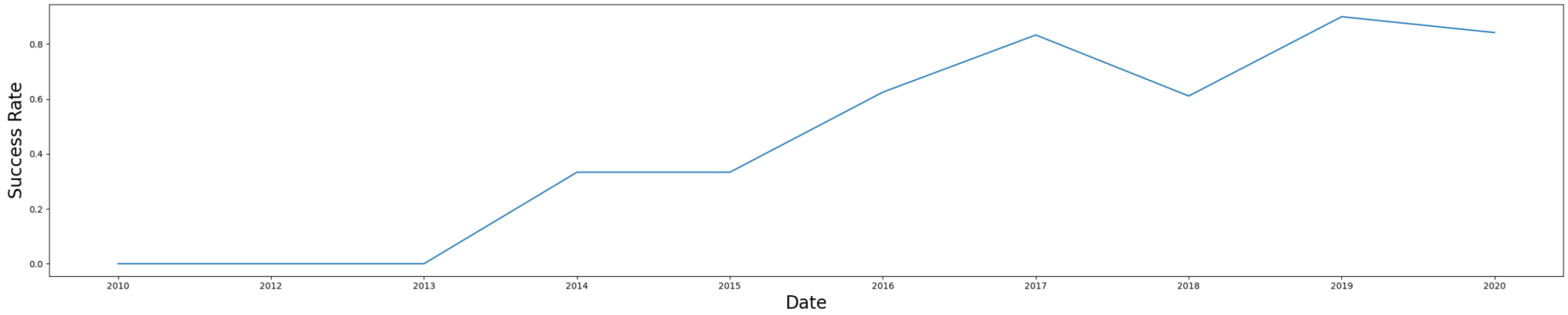
```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



Launch Success Yearly Trend

- Launch Success Yearly Trend
  - The success rate of launches rose over the years, with a dip in 2018.
  - Flights before 2014 had 0% success rate while flights in 2019 and 2020 have about 80% success rate.

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
class_mean_by_year_df = df[['Date', 'Class']].groupby('Date').mean().reset_index()
sns.lineplot(data=class_mean_by_year_df, x="Date", y="Class")
plt.xlabel("Date",fontsize=20)
plt.ylabel("Success Rate",fontsize=20)
plt.show()
```



All Launch Site Names

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

```
%sql select distinct(Launch_Site) from SPACEXTBL;
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

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;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No att
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No att
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - Total payload mass across all launches add up to 45596kg.

```
%sql select sum(PAYLOAD_MASS__KG_) as TotalPayloadMass from SPACEXTBL where Customer = "NASA (CRS)";
```

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

TotalPayloadMass
45596.0

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - Average payload mass carried by booster version F9 v1.1 is 2928.4kg

```
%sql select avg(PAYLOAD_MASS__KG_) as TotalPayloadMass from SPACEXTBL where Booster_Version = "F9 v1.1";
```

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

TotalPayloadMass
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - The date is 22 Dec 2015

```
%%sql
select Date, Landing_Outcome
from SPACEXTBL
where Landing_Outcome like "Success%ground%"
order by Date DESC
limit 1;
```

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

Date	Landing_Outcome
22/12/2015	Success (ground pad)

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
  - The booster versions are: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ > 4000
and PAYLOAD_MASS__KG_ < 6000
and Landing_Outcome like "%Success%drone%";
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696.0
F9 FT B1026	4600.0
F9 FT B1021.2	5300.0
F9 FT B1031.2	5200.0

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome;
```

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

Mission_Outcome	count(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
  - The max load is 15600kg.

```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) as maxmass from SPACEXTBL);
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - There are two failed launches in 2015, both from CCAFS LC-40.

```
%%sql
select substr(Date,7,4) as Year, substr(Date, 4, 2) as Month, Booster_Version, Launch_Site
from SPACEXTBL
where Landing_Outcome like "%Failure%drone%"
and Date like "%2015";
```

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

Year	Month	Booster_Version	Launch_Site
2015	10	F9 v1.1 B1012	CCAFS LC-40
2015	04	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
select Landing_Outcome, count(*) as SuccessCount
from SPACEXTBL
where Landing_Outcome like "Success%"
group by Landing_Outcome
order by SuccessCount DESC
;
```

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

Landing_Outcome	SuccessCount
Success	38
Success (drone ship)	14
Success (ground pad)	9

Section 3

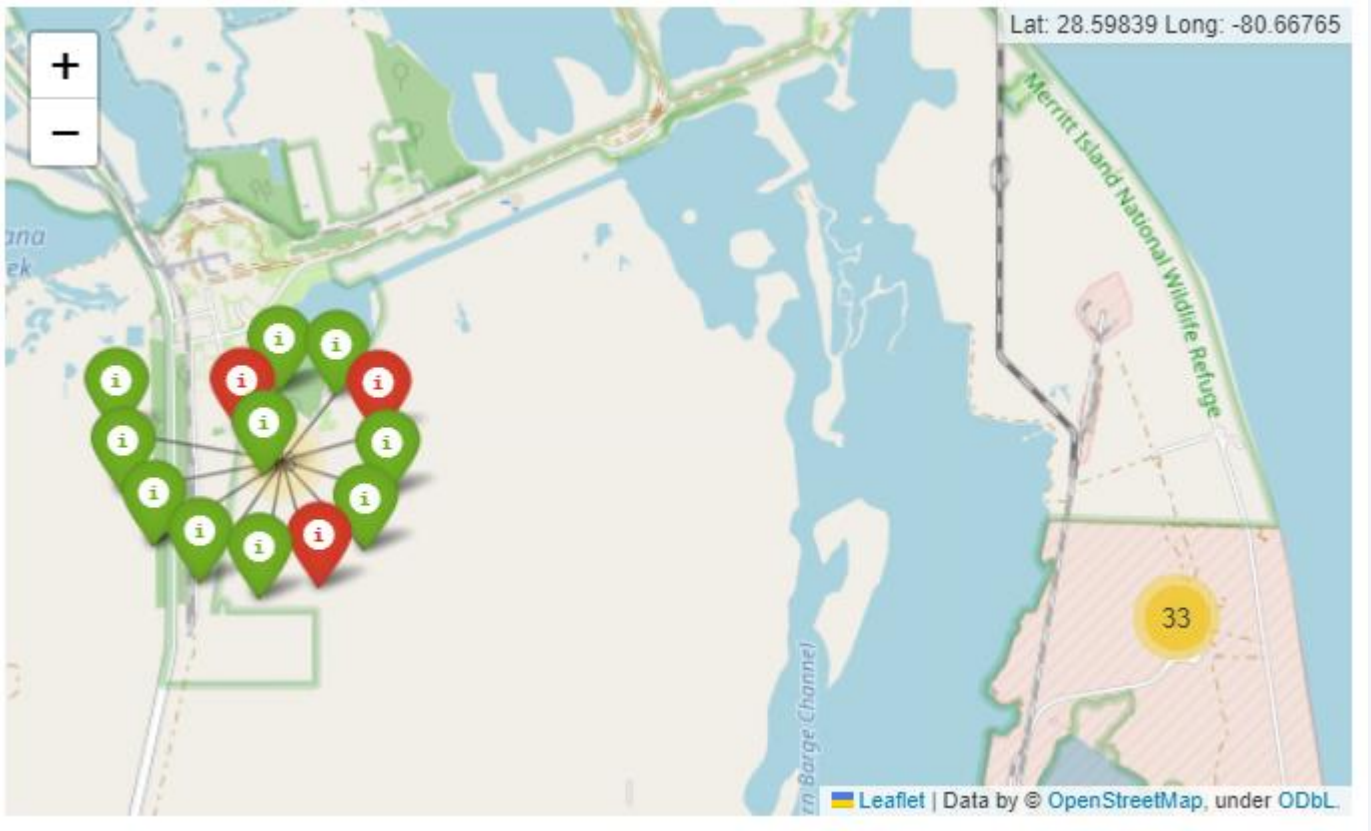
Launch Sites Proximities Analysis

Distribution of launch sites in US

- 10 sites are in the west coast with the same geographical locations.
- 46 sites are in the east with 3 unique geographical locations.

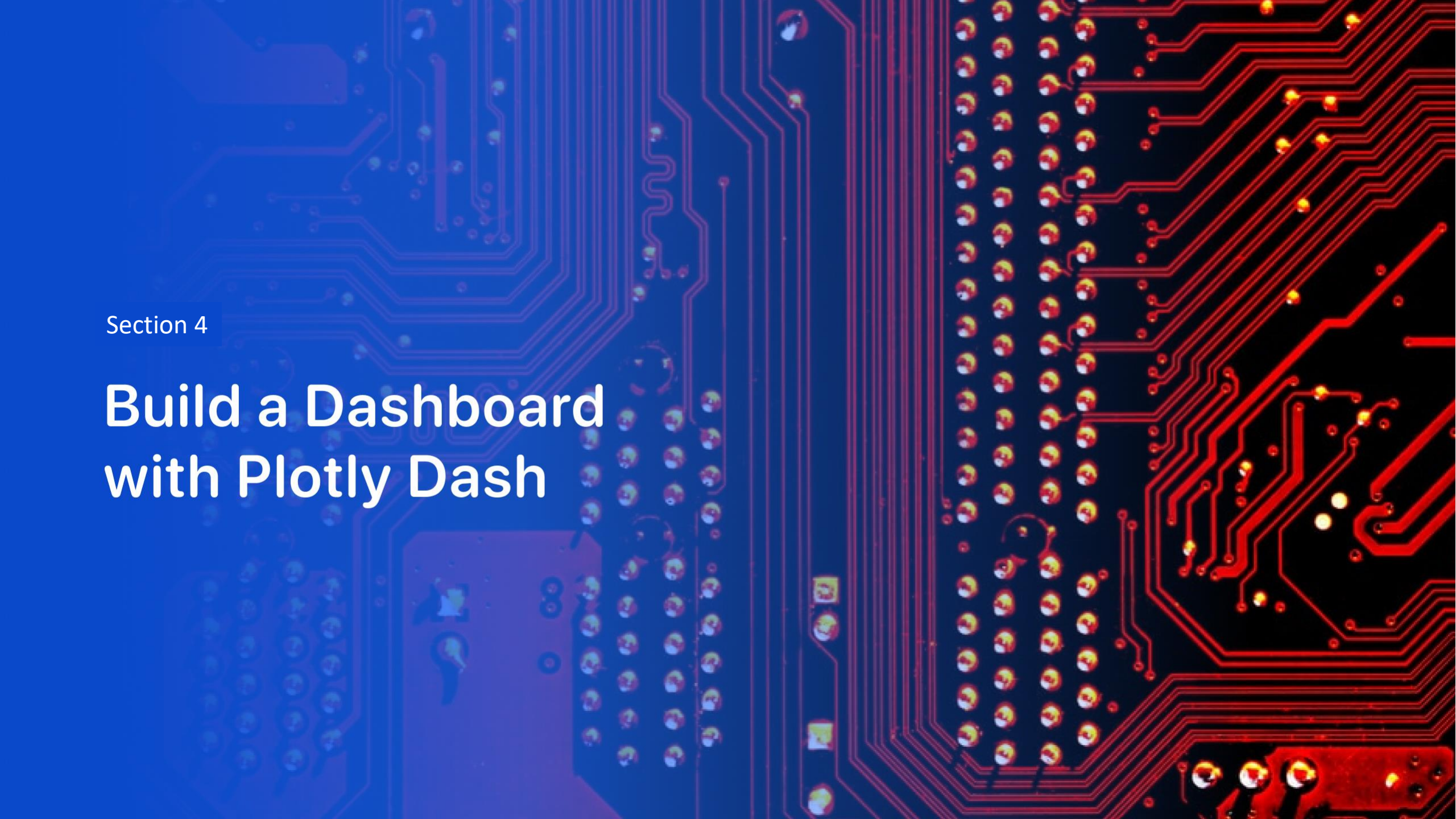
Distribution of launch sites in US East Coast

- A close-up view of the east coast launch sites



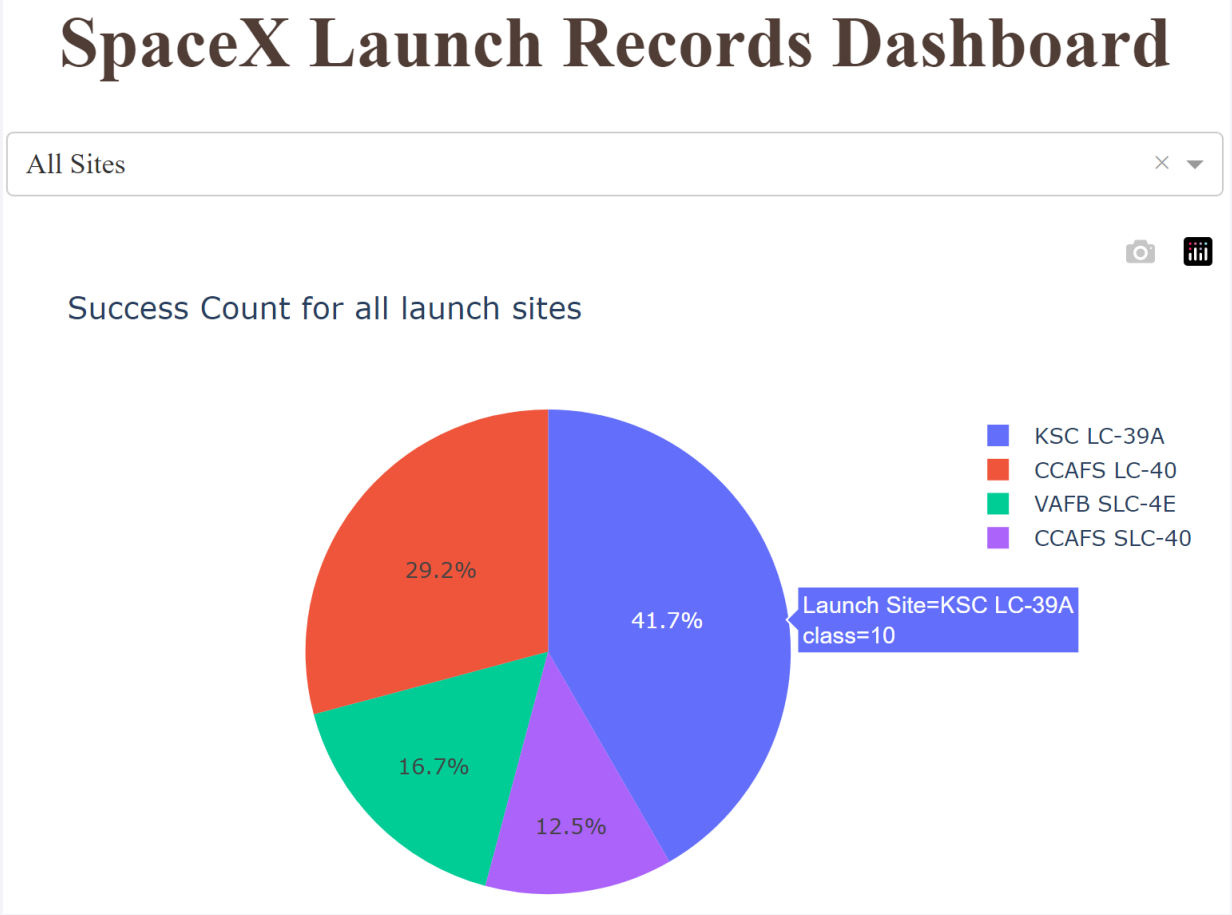
Launch site's proximity to another point

- These 33 sites are 0.58km away from the marked point on the nearest highway.


Section 4
Build a Dashboard with Plotly Dash

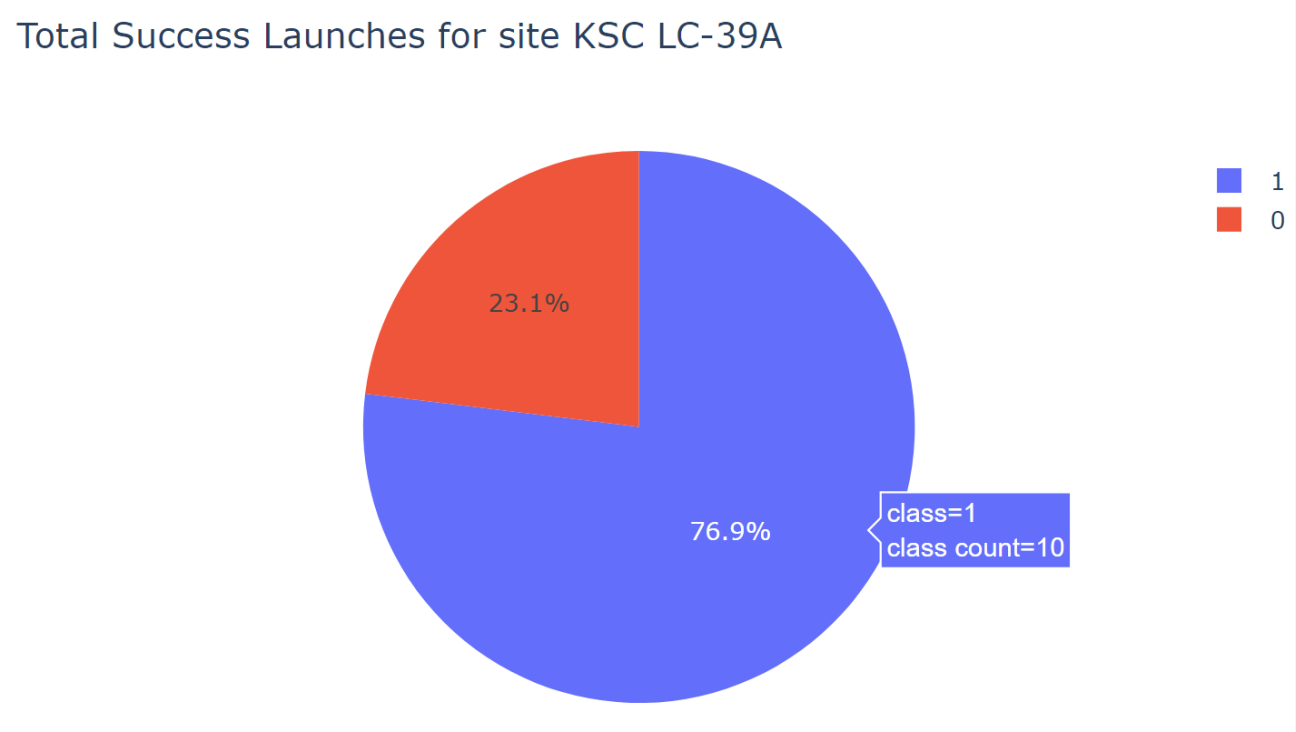
Success Count for All Launch Sites

- Hovering on each slice of the pie chart shows the launch site name and the “class” value is the number of successful launches from that site.
- The purple slice (KSC LC-30A) has the highest number of successful launches.



Successful vs Failed Launches for KSC LC-39A

- KSC LC-39A has the highest rate of successful launches (76.9%).



Payload Mass by Class for All Launches

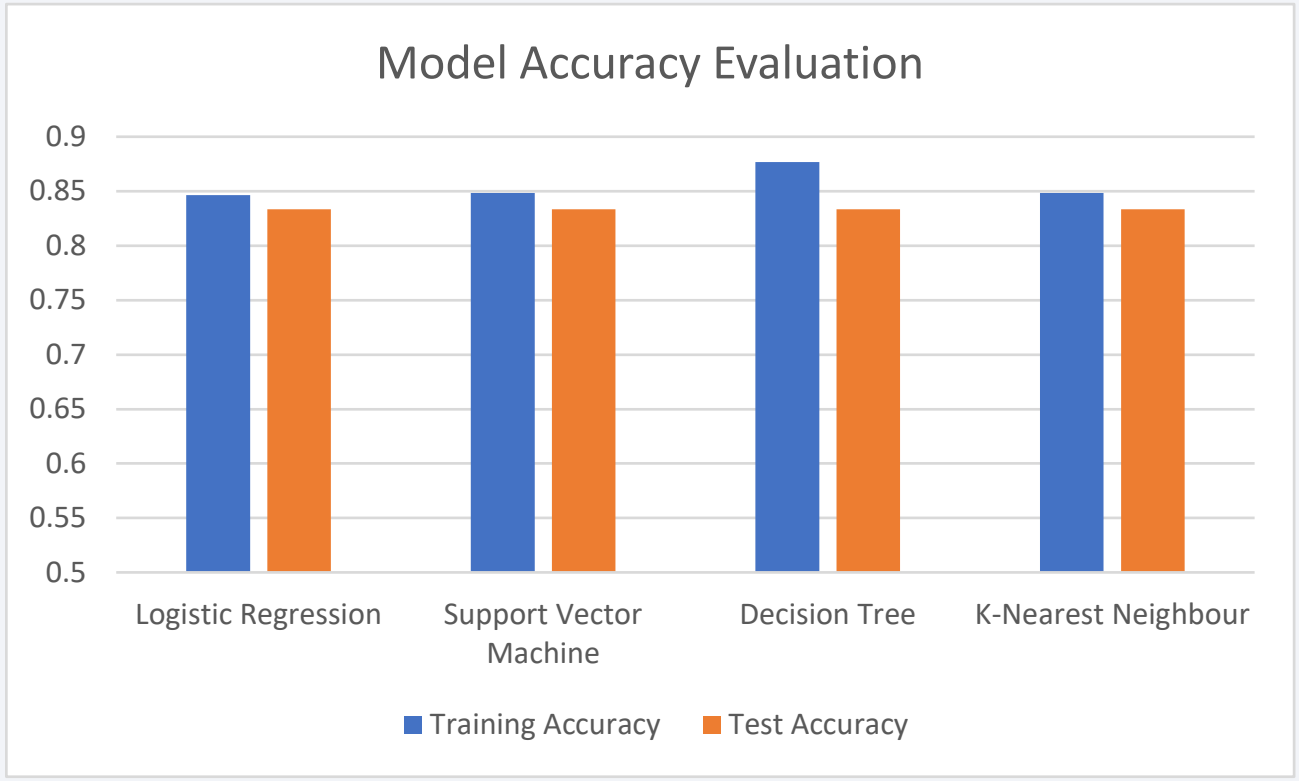
- Most of the successful launches have a payload mass of 2000-4000kg
- Hovering on each point will show the booster version, payload mass, and class

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All the models have the same test accuracy, although the decision tree model had the highest training accuracy

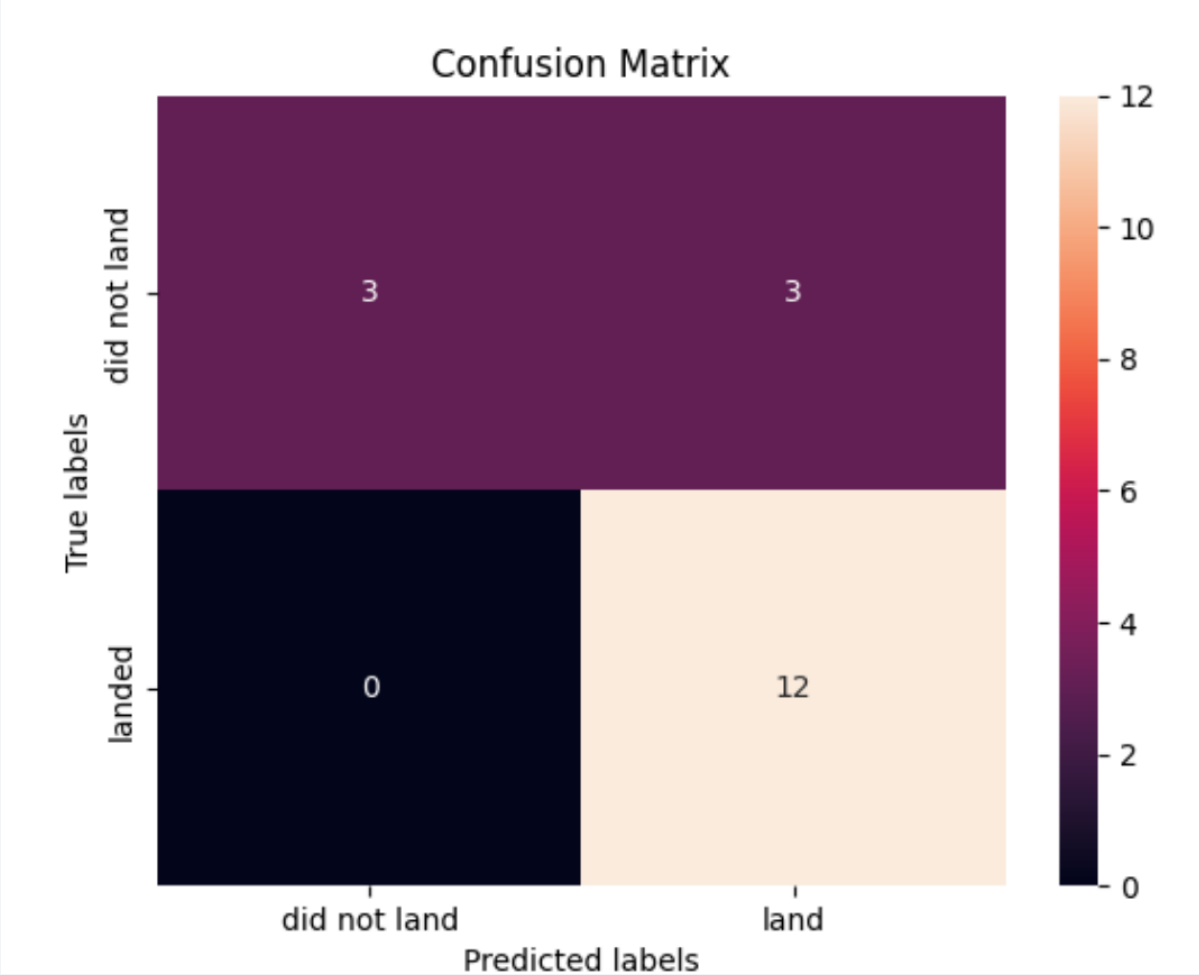


	LR	SVM	Tree	KNN
0	0.846429	0.848214	0.876786	0.848214
1	0.833333	0.833333	0.833333	0.833333

Confusion Matrix (Decision Tree Model)

- True positives: 12
- True negatives: 3
- False positives: 3
- False negatives: 0

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- Performed initial data exploration and created interactive visualizations based on the launch success rates, launch sites, payload mass, year of launch etc.
- Created geographical visualizations of launch sites and successful/failed launches
- Launch Site KSC LC-39A has the highest rate of successful launches (76.9%).
- The success rate of launches generally rose over the years; flights before 2014 had 0% success rate while flights in 2019 and 2020 have about 80% success rate.
- All 4 prediction models were able to predict successful launches with 83% accuracy based on the training data. 83 columns were used as independent variables.

Appendix

- Git Repo: https://github.com/yanling-yl/IBMDataScienceLab/tree/main/Capstone_Project

Number of launches for each launch site

```
# Apply value_counts() on column LaunchSite
df.LaunchSite.value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Number of launches for each orbit type

```
# Apply value_counts on Or
df.Orbit.value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

Thank you!