IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yejun (Nicole) Tu
May 2023

OUTLINE

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

EXECUTIVE SUMMARY

Methodology

- Data collection included pulling records from the SpaceX REST API and web scraping public Wikipedia HTML data
- Data cleaning and wrangling ensured data was prepped for exploratory analysis and machine learning model
- EDA consisted of data visualization of key parameters, SQL filtering and representation of data, and mapping of launch records using Folium

Results

- All SpaceX launch sites are near the coast and not in high latitude areas
- The most successful launch site is KSC LC-39A
- We can predict landing success using a decision tree ML model with 88.9% accuracy

INTRODUCTION

- The first stage in a launch is typically very expensive and large
- Success in the first stage can be due to a failed landing, or may be the result of sacrificing the launch due to some other parameters (e.g. payload mass, orbit, or customer)
- To be competitive with SpaceX, who has been able to safely cut the cost of first stage expenses, SpaceY needs to be able to make predictions about cost in the first stage

How accurately can we determine if the first stage will land?

If we can predict launch success, we can determine the cost of a launch, and make SpaceY competitive against SpaceX

Section 1

Methodology

METHODOLOGY

- Data collection methodology:
  - Data collection included pulling records from the SpaceX REST API and web scraping public Wikipedia HTML data
- Perform data wrangling
  - Data wrangling ensured missing values were adjusted, and organized and standardized for exploratory analysis and machine learning 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
  - Standardized data were split into train/test sets and fit using four types of classification models, and best parameters were chosen with a grid search technique before assessing model accuracy for the best fit model from each type

Data Collection

Data was collected using an **API** and via **web scraping**

SpaceX REST API was used to source data on the company's rocket launches with focus on **Falcon 9** launches

Additional **Falcon 9** and Falcon Heavy Launch data were **scraped** from **Wikipedia HTML data**

Data Collection – SpaceX API

- SpaceX REST API was used to:

1. GET request to SpaceX REST API
2. Parse launch data
3. Filter data for Falcon 9 launches
4. Clean missing data

Data Collection - Scraping

- Web scraping from Wikipedia extracted additional Falcon 9 and Falcon Heavy Launch Records by:

Extract data as HTML → Parse HTML tables → Convert HTML data to Pandas dataframe

Data Wrangling

- Data wrangling involved **cleaning** and **labeling** data to make it appropriate for modeling, specifically:

Calculating the number of launches per site → Determining the number and occurrence of orbits → Calculating the number and occurrence of mission outcomes for each orbit → Labeling landing outcomes

EDA WITH DATA VISUALIZATION

- Exploratory data analysis partly included data visualization
- Data collected on SpaceX launches were plotted in the following ways:
    - Flight number vs. Launch site (scatter plot)
    - Payload mass vs. Launch site (scatter plot)
    - Success rate vs. Orbit type (bar pot)
    - Flight number vs. Orbit type (scatter plot)
    - Payload mass vs. Orbit type (scatter plot)
    - Yearly Trend of Launch Success (line plot)

EDA WITH SQL

- Data was further explored with SQL, suing the workflow:
  - Display unique launch sites with DISTINCT
  - Display 5 launch sites with ‘CCA’ string using LIKE ‘%CCA%’ LIMIT 5
  - View total payload mass with SUM(PAYLOAD_MASS__KG_) for NASA CRS customers
  - View average payload with AVG(PAYLOAD_MASS__KG_) for booster version LIKE '%F9 v1.1%’
  - See first success date with MIN(Date) and WHERE Landing_Outcome LIKE '%Success%’
  - Viewing boosters within a payload range and in success landing cases with WHERE Landing_Outcome LIKE '%Success%' AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
  - Counting successful and failed missed with COUNT(Mission_Outcome) and GROUP BY Mission_Outcome
  - Listing boosters with max payloads using a subquery, WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
  - Listing cases with date information and case data using date syntax and LIKE
  - Ranking landing outcomes between 2010-06-04 and 2017-03-20 using functions like COUNT, DISTINCT, BETWEEN, ORDER BY

BUILD AN INTERACTIVE MAP WITH FOLIUM

- In a map representing landing sites for launches by SpaceX, markers and other information were added
- All launch sites were marked on a map of North America with a marker and circle as a foundation for representing the coordinate location and vicinity of launch locations
- Next, success and failure launches were marked at each launch sites for each launch record, and were described with a color-coded marker (green for success, red for fail)
- Then, distance between launch sites and proximities of other features
    - A launch site in Florida and its proximity to the nearest coastline was calculated to determine ease of water landing for launches
    - A launch site in a Florida and proximity to the nearest wildlife refuge were calculated to consider risk of launches failing and encroaching on a refuge

BUILD A DASHBOARD WITH PLOTLY DASH

- In a dashboard with Plotly, I represented launch data using several elements:
  - A dropdown menu to filter data by launch site for ease of viewing all or subset data
  - A pie chart of success/failure launches at all site, or at a single site, to represent the proportion of successful and failed cases
  - A slider to select payload range to filter through heavy or light loads
  - A scatter plot showing the correlation between payload and launch success, to consider the relationship between payload mass and chance of success visually

Predictive Analysis (Classification)

- Four model types were tested on standardized data to classify launch success: logistic regression, support vector machine, decision tree, and K nearest neighbor
- Best parameters were determined using a grid search technique
- The final model built of each type was used to assess accuracy, and the best model was chosen based on accuracy score

Create model types with training data

Grid search for best parameters

Evaluate model accuracy

Model selection using test data accuracy

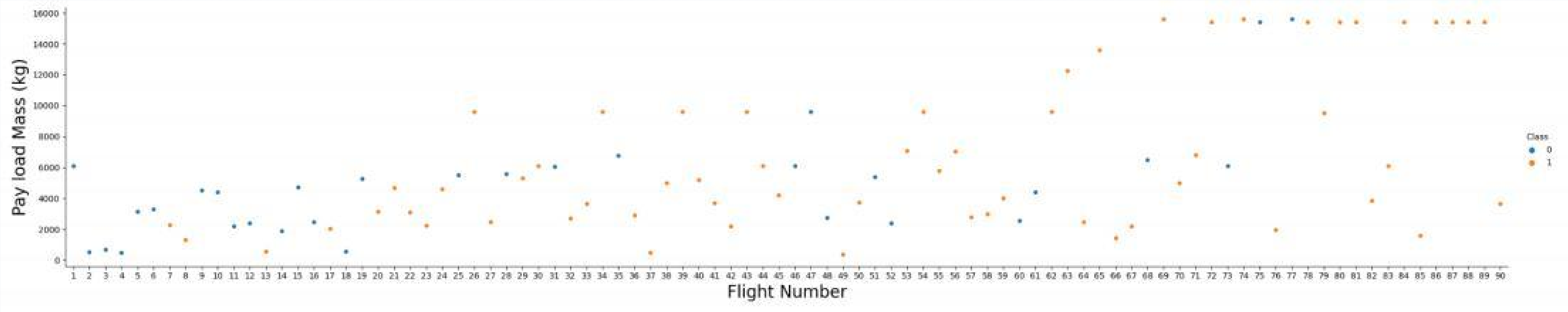
Results

- Exploratory data analysis showed the relationship between launch data parameters, such as payload mass, orbit type, and launch success, as well as geographic relationships
- Interactive analytics provide additional ways of exploring the provided data
- Predictive analysis demonstrated that we can predict launch success with 88.9% accuracy with the given data

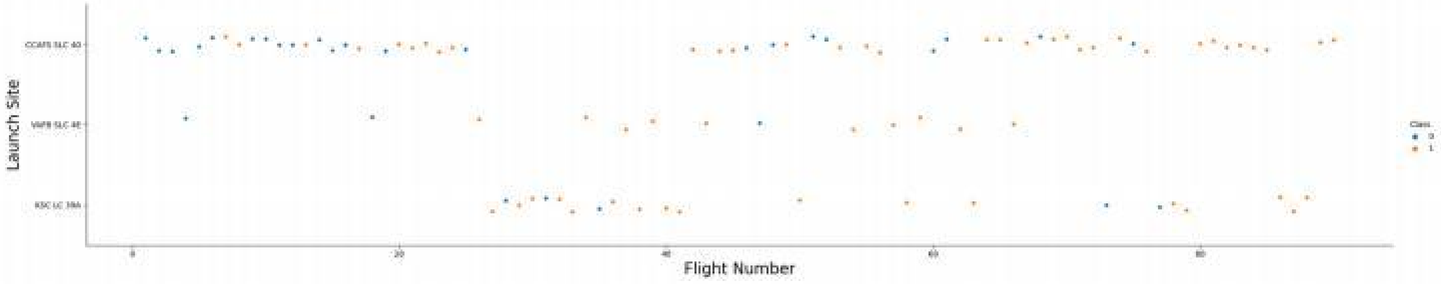
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site




```
### 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()
```



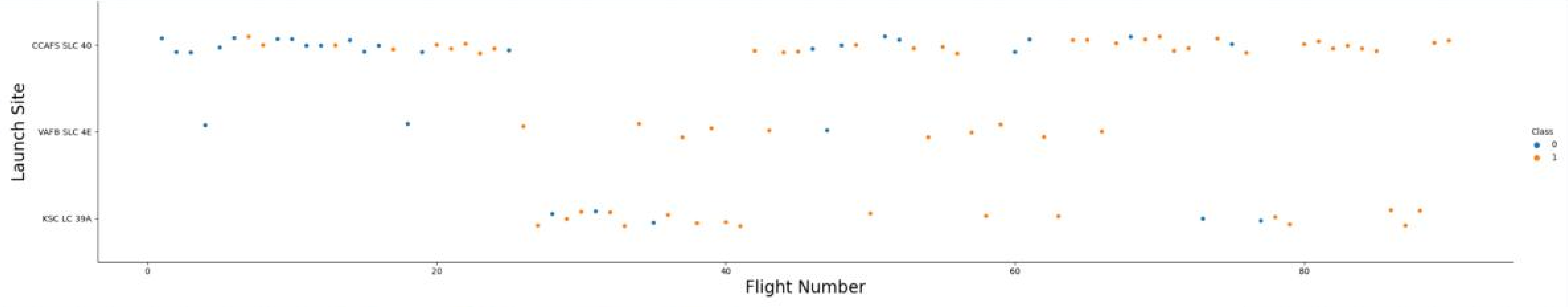
Use the function `catplot` to plot `FlightNumber` vs `LaunchSite`, set the parameter `x` parameter to `FlightNumber` ,set the `y` to `Launch Site` and set the parameter `hue` to `'class'`

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
```

Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

The flight number for launch site CCAFS SLC 40 has abroader range than for other sites.

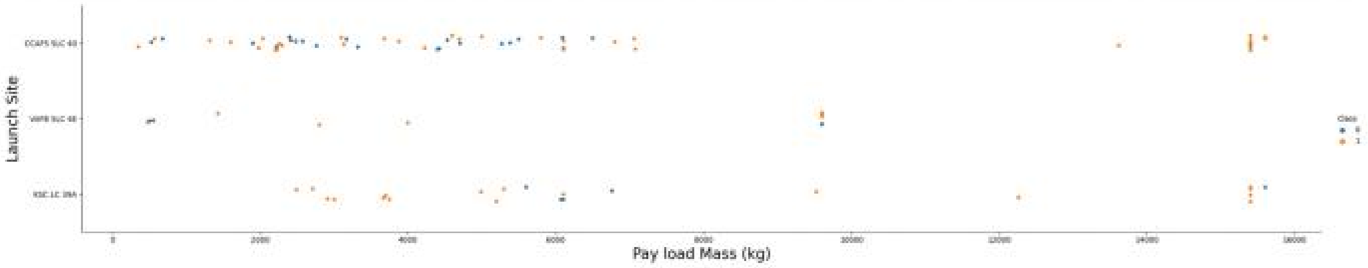
Payload vs. Launch Site




```
### TASK 2: Visualize the relationship between Payload and Launch Site
```

We also want to observe if there is any relationship between launch sites and their payload mass.

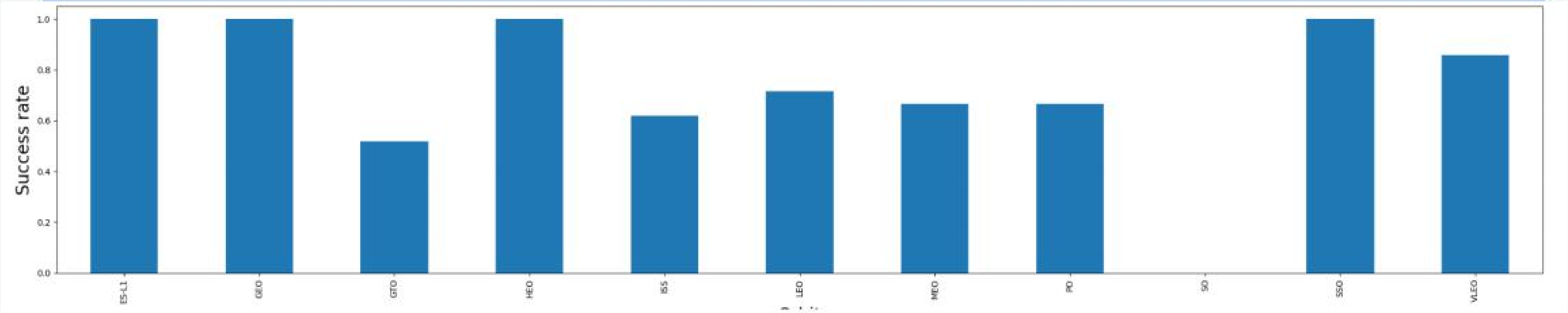
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(x="PayloadMass", y="LaunchSite", hue="Class", data=df, aspect = 5)
plt.ylabel("Launch Site",fontsize=20)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.show()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



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

The launch site VAFB SLC 4E has no heavy (>10000 kg) rocket launches, while the other sites have both heavy and lighter rocket launch loads.

Success Rate vs. Orbit Type

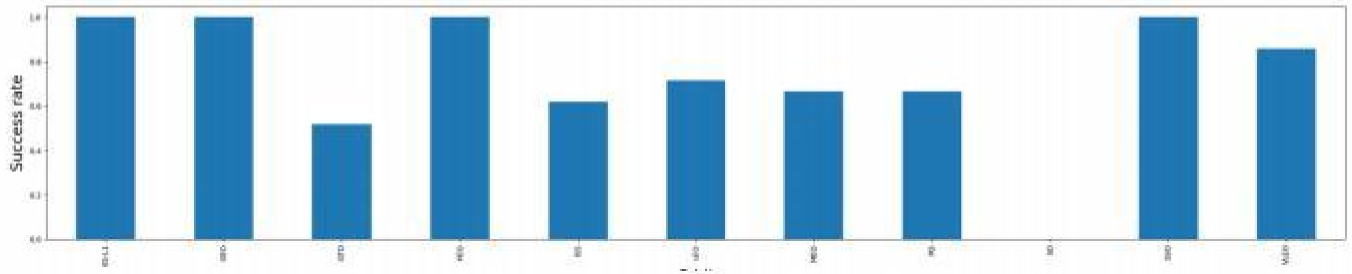


TASK 3: Visualize the relationship between success rate of each orbit type

Next, we want to visually check if there are any relationship between success rate and orbit type.

Let's create a `bar chart` for the sucess rate of each orbit

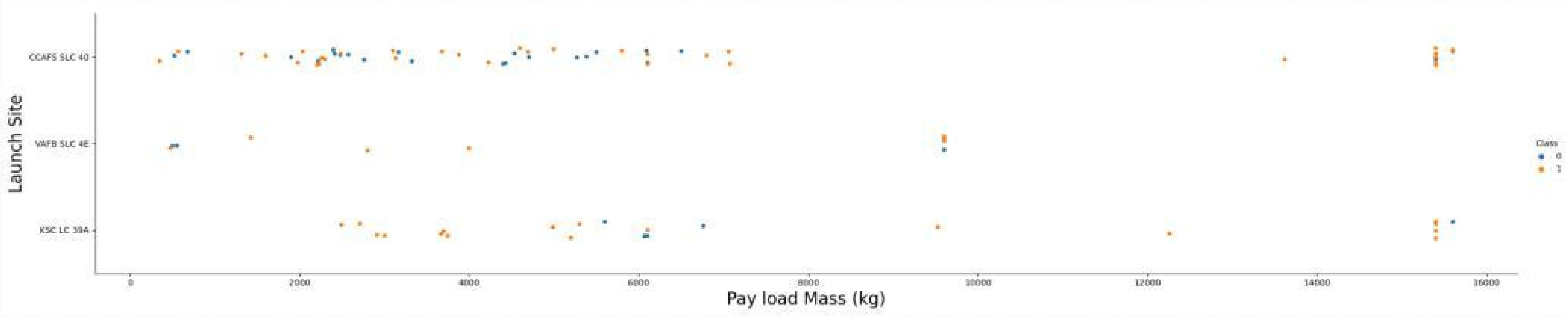
```
# HINT use groupby method on Orbit column and get the mean of Class column
df.groupby("Orbit").mean()['Class'].plot(kind='bar')
plt.ylabel("Success rate",fontsize=20)
plt.xlabel("Orbit",fontsize=20)
plt.show()
```



Analyze the ploted bar chart try to find which orbits have high sucess rate.

The orbits HEO, SSO, GEO, and ES-L1 have the highest success rates for SpaceX Falcon 9 launches, and the SO orbit has the lowest success rate.

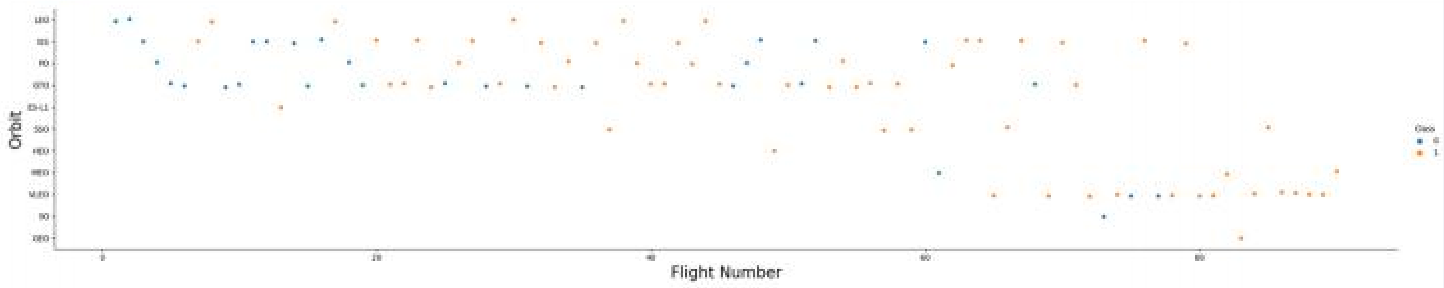
Flight Number vs. Orbit Type



TASK 4: Visualize the relationship between FlightNumber and Orbit type

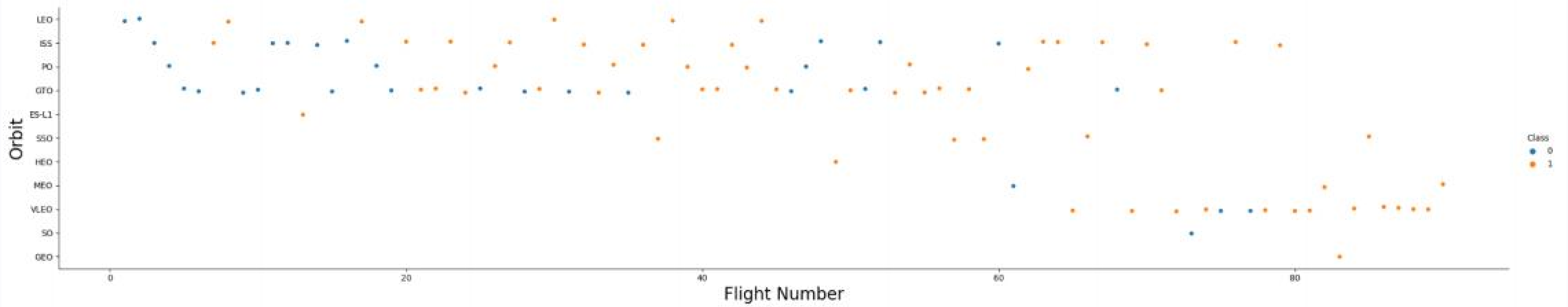
For each orbit, we want to see if there is any relationship between FlightNumber and Orbit type.

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

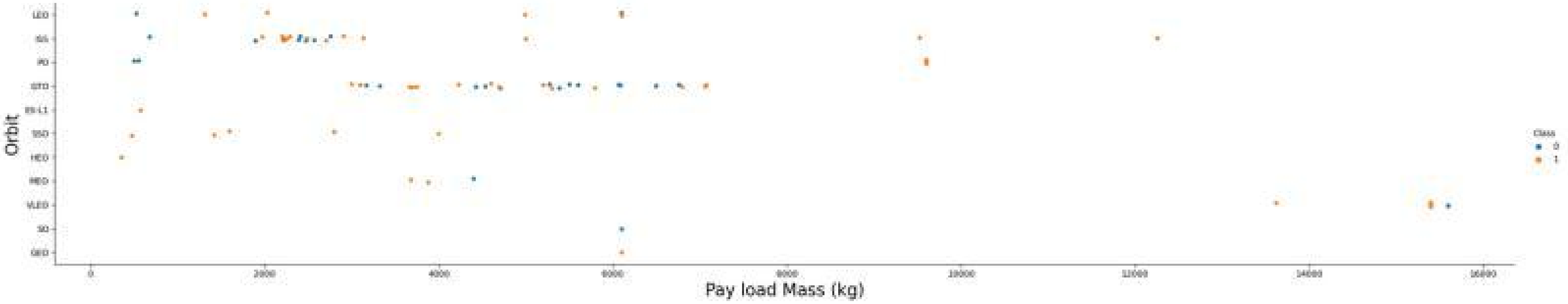


You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit. **There is no relationship with GTO orbit and others and flight number, but there seems to be a relationship between LEO orbit and flight number**

Payload vs. Orbit Type



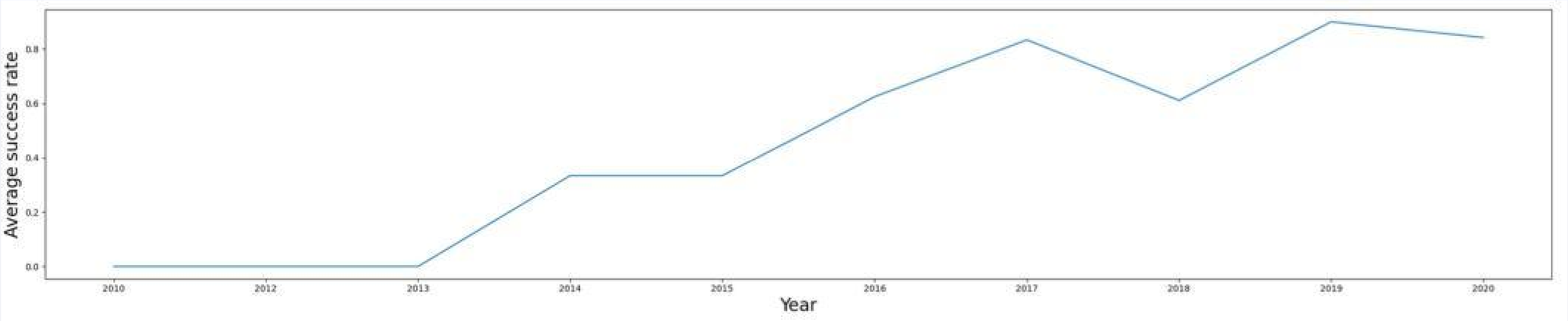

```python
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.ylabel("Orbit",fontsize=20)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

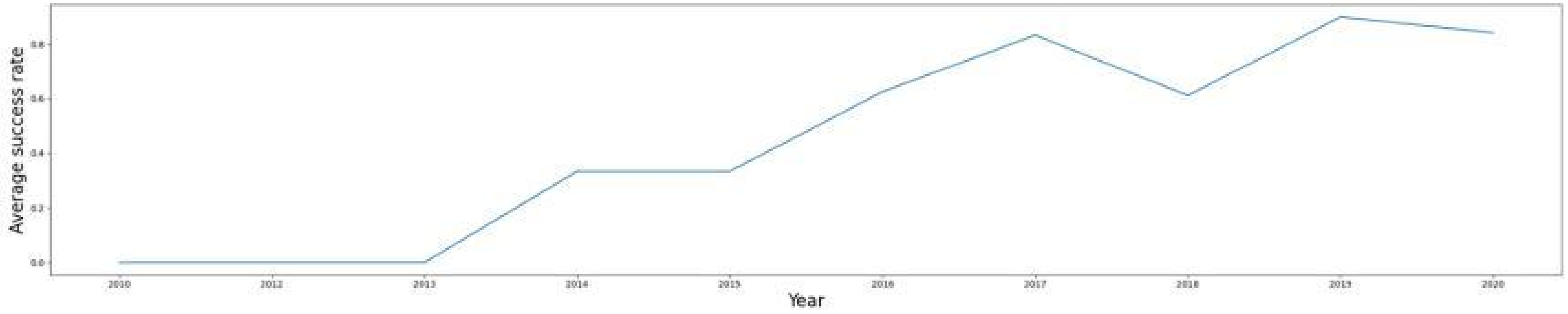
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend




```python
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
df2 = df.groupby(by="Date").mean()
df2.reset_index(inplace=True)

sns.lineplot(data=df2, x="Date", y="Class")
plt.xlabel("Year",fontsize=20)
plt.ylabel("Average success rate",fontsize=20)
plt.show()
```



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

All Launch Site Names

- All launch sites were selected from the SpaceX SQL data table

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

Launch Site Names Begin with 'CCA'

- The first 5 records where launch sites begin with `CCA` were selected using LIKE

```
%%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

- The total payload carried by boosters from NASA was calculated with SUM()

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER LIKE '%CRS%'
```

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

SUM(PAYLOAD_MASS__KG_)
48213.0

Average Payload Mass by F9 v1. 1

- The average payload mass carried by booster version F9 v1. 1 was calculated form the filtered data with AVG

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version like '%F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad were found by filtering data for success cases

```
%%sql
SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%'
```

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

MIN(Date)
01/07/2020

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were found using successive filtering

```
%%sql
SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%' AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
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

- The total number of successful and failure mission outcomes was determined with COUNT

```
%%sql
SELECT COUNT(Mission_Outcome),Mission_Outcome FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

COUNT(Mission_Outcome)	Mission_Outcome
0	None
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass were found using subquery to find the maximum load in each group

```
%%sql
SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTB
```

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

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

2015 Launch Records

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 were found by filtering the data table

```
%%sql

SELECT Date,Landing_Outcome,Booster_Version,Launch_Site,substr(Date, 4, 2) AS Month,substr(Date,7,4) AS Year FROM SPACEXTBL
    WHERE Landing_Outcome LIKE '%Failure%' AND Year='2015'
```

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

Date	Landing_Outcome	Booster_Version	Launch_Site	Month	Year
01/10/2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	10	2015
14/04/2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	04	2015

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 'Date', COUNT(DISTINCT Landing_Outcome) FROM SPACEXTBL WHERE 'Date' BETWEEN '2010-06-04' AND '2017-03-20'
    ORDER BY COUNT(Landing_Outcome) DESC
```

Section 3

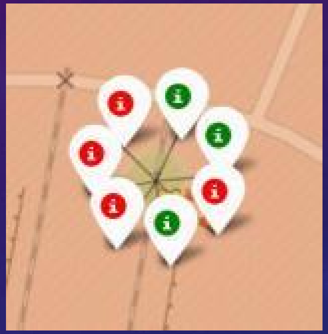
Launch Sites Proximities Analysis

LAUNCH SITES (FOLIUM MAP)

- Locations of the four launch sites in North America. One site is in California, USA, and three sites are in Florida, USA.
- Sites in Florida are closer to the Equator than the California site.
- All sites are close to the respective coastline (either Pacific or Atlantic Ocean) to ensure a water landing is possible

LABELED SUCCESS MARKERS AT LAUNCH SITES

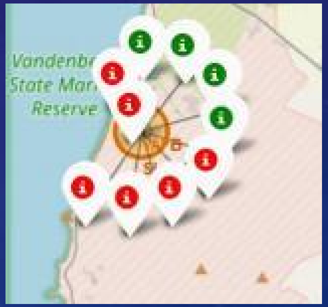
- Successful and failed launch records were shown with a marker at all four sites
- Green markers indicate successful launch cases at a site, red markers indicate a failed launch
- CCAFS SLC-40 has the most failed records, and KSC LC-39A has the most successes, when considering absolute counts



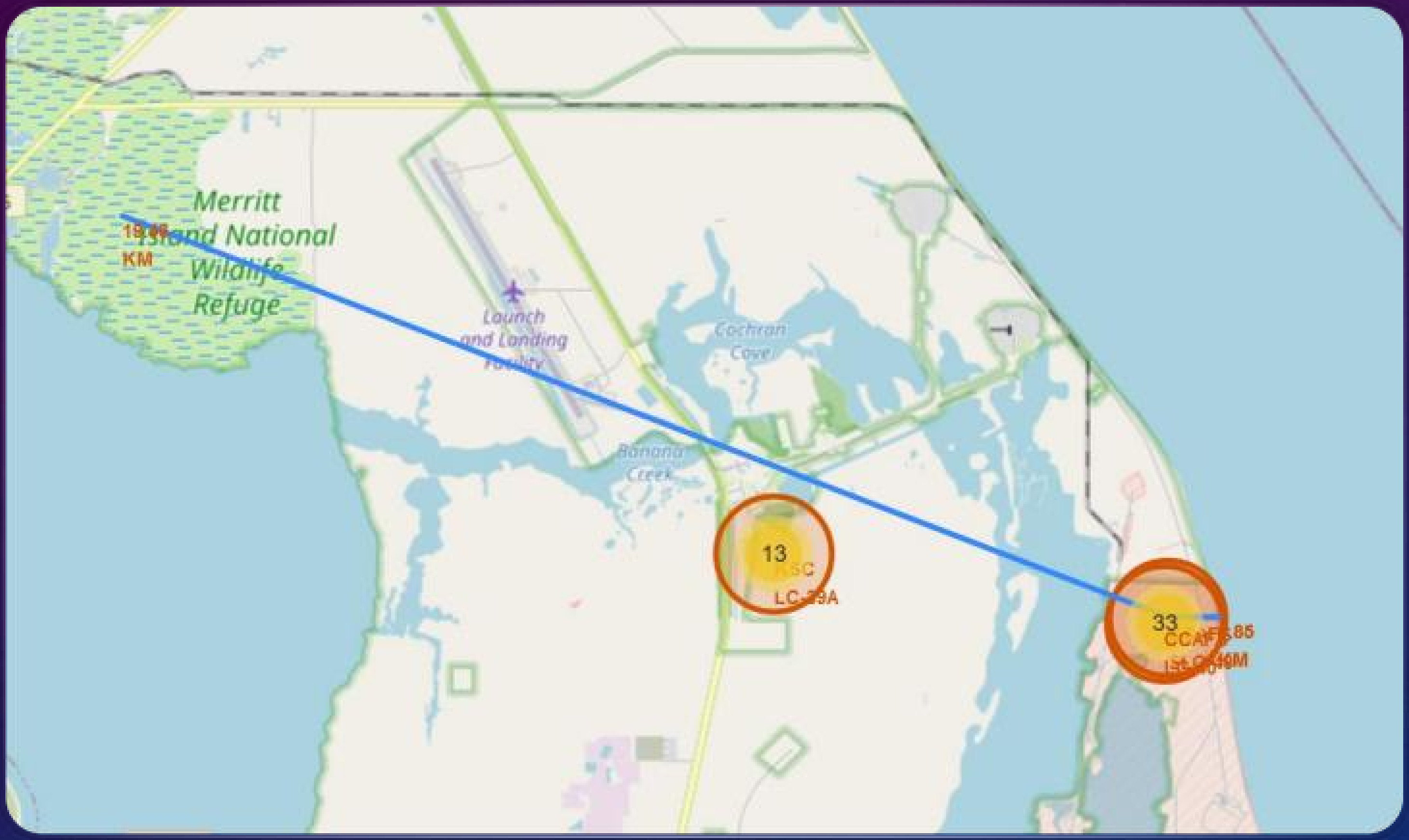
CCAFS LC-40

CCAFS SLC-40

VAFB SLC-4E



KSC LC-39A



CCAFS SLC-40 DISTANCE TO PROXIMITIES

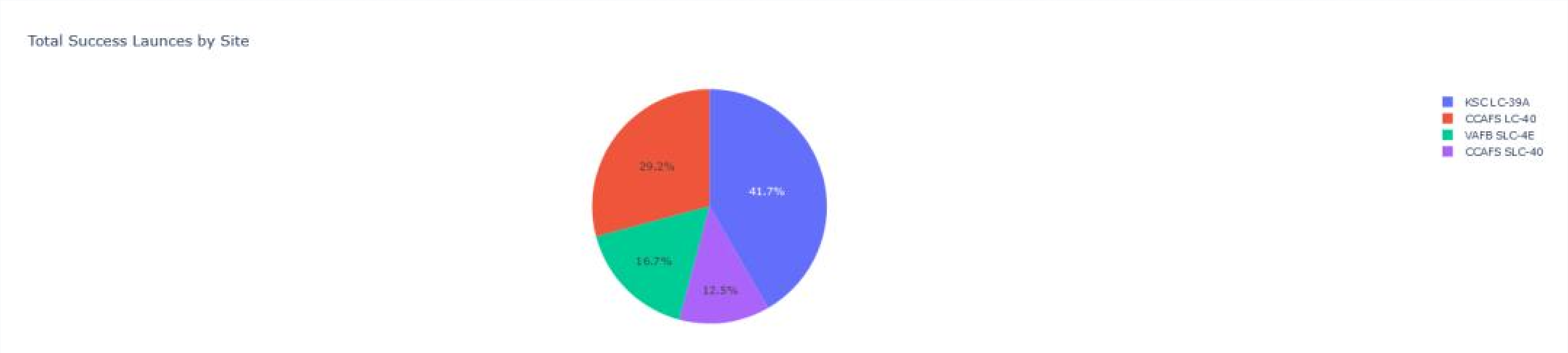
- Two proximate distances are shown from the CCAFS SLC-40 launch site
- The distance to the nearest coastline was 0.85 km, and the distance to the Merritt Island National Wildlife Refuge was 19.5 km
- This launch site is quite close to the coastline, but noticeably farther from the wildlife refuge

Section 4

Build a Dashboard with Plotly Dash

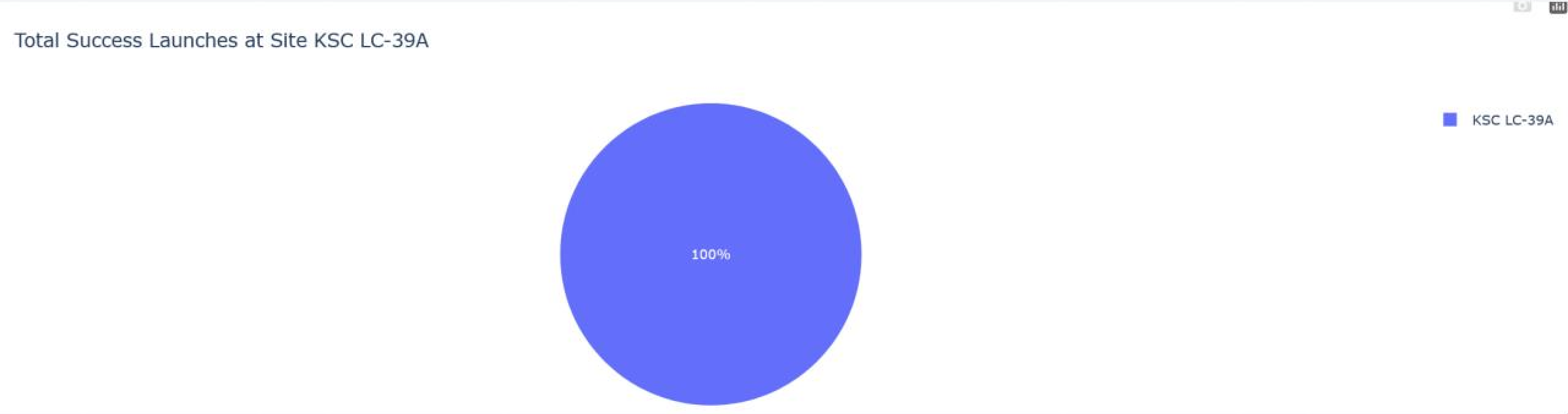
Total Successful Launches

- Total successful launches for all sites are shown in a pie chart
- We can see that site KSC LC-39A has the greatest proportion of successful launches (41.7%) and the CCAFS SLC-40 site has the lowest (12.5%)



Launches at KSC LC-39A Site

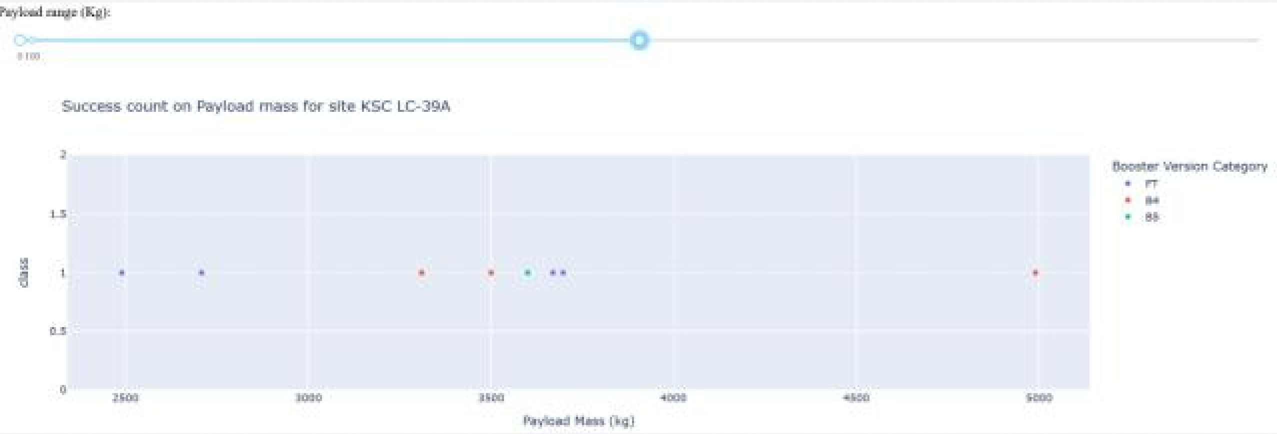
- Here we can see the KSC LC-39A has a high rate of successful launches

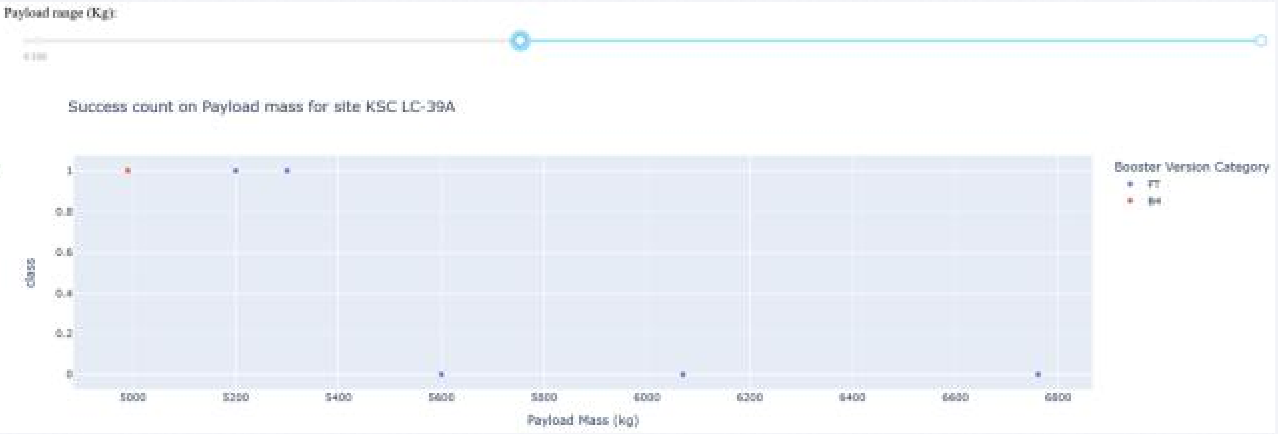


Payload Mass and Launch Success

- Payload mass and launch success outcome is represented in a scatterplot where payload mass can be varied using a slider

- Using two cases, we can see that the lower payload range (mass = 0-5000 kg) shows more successes compared to the higher range (mass = 5000-6800 kg)





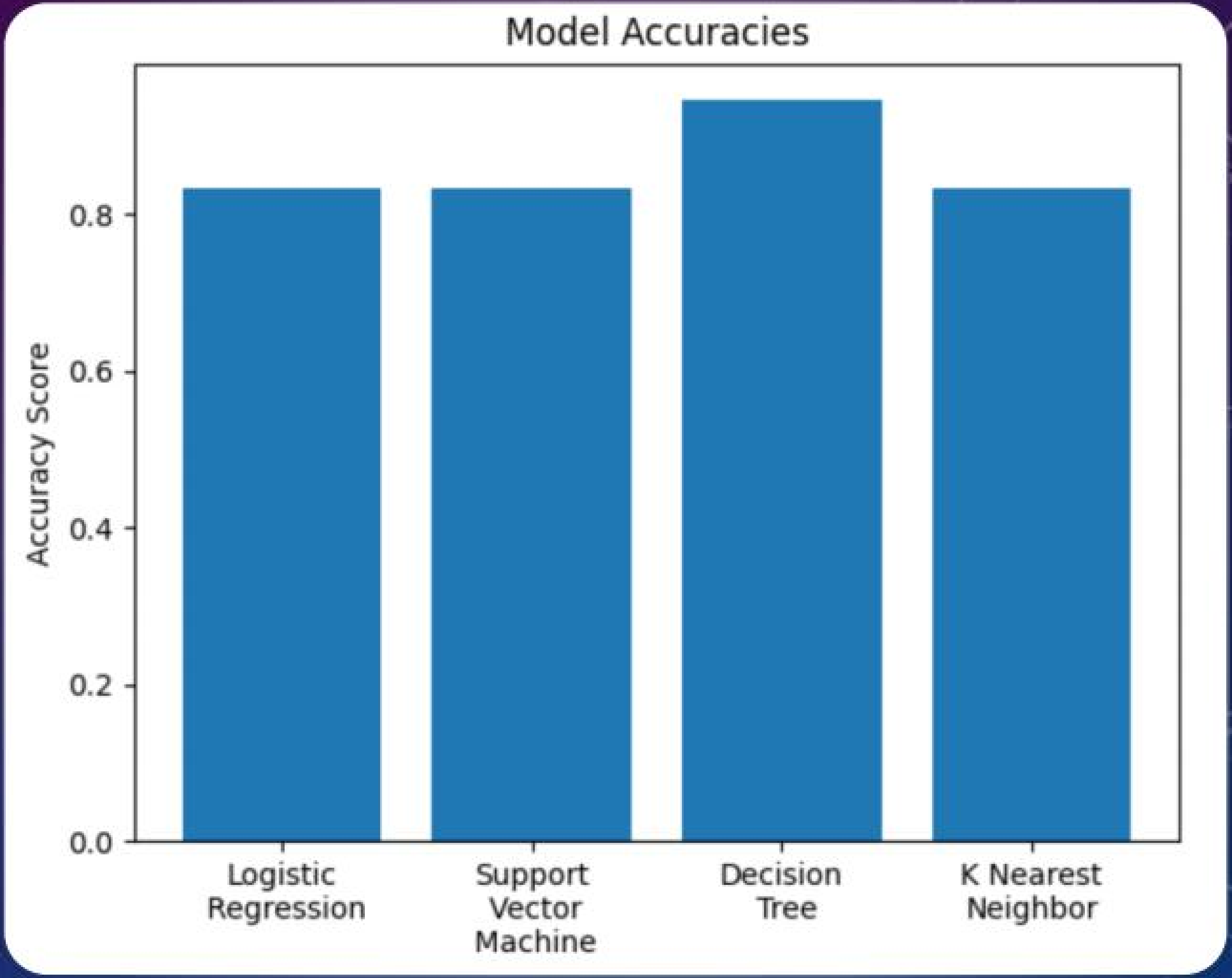
Section 5

Predictive Analysis (Classification)

CLASSIFICATION ACCURACY

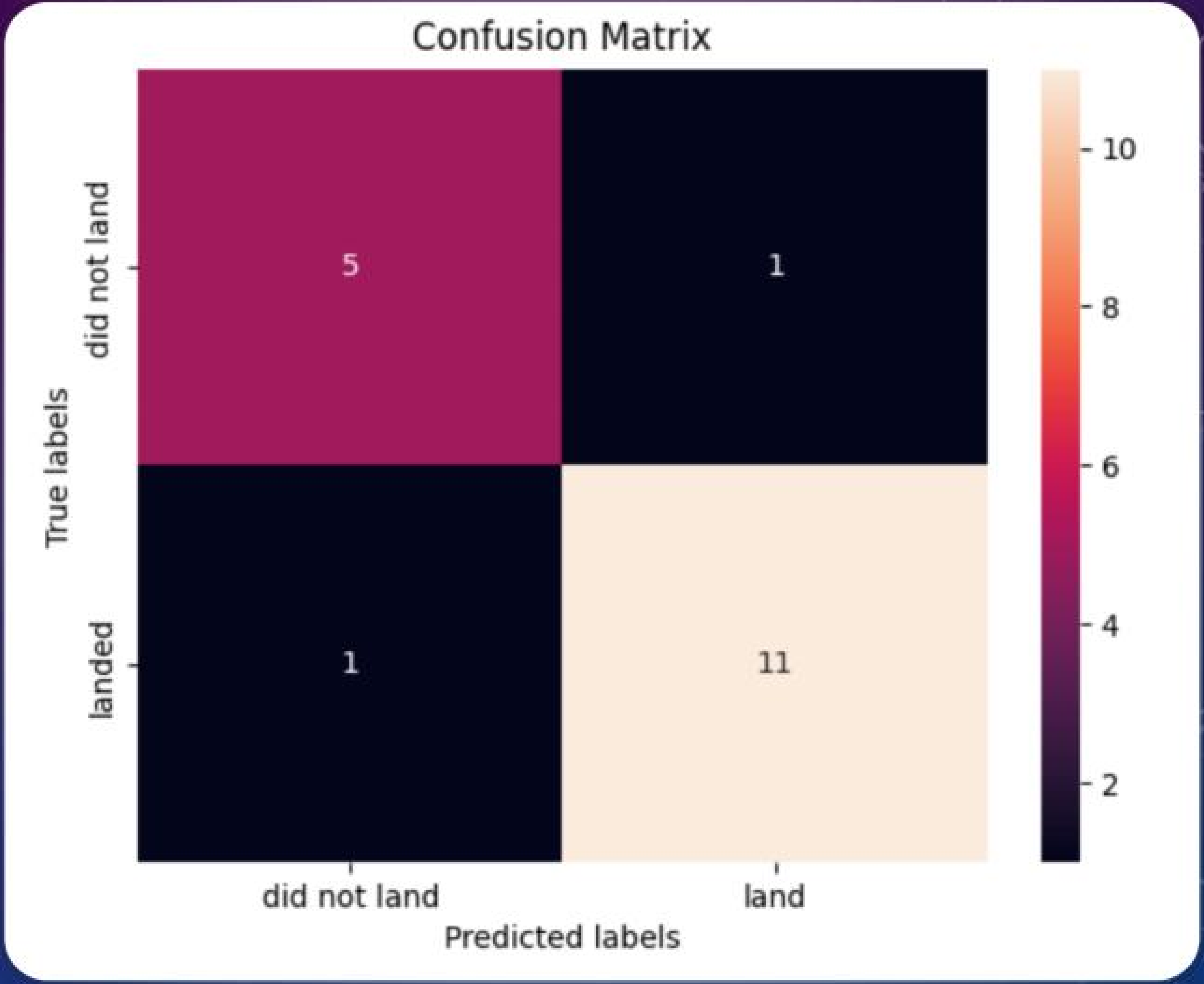
- Several classification models were built were varying accuracy

- The model with the highest accuracy is the Decision Tree model



CONFUSION MATRIX

- In the confusion matrix from the decision tree model, we can see that only one false positive and one false negative case was found in model assessment, suggesting good accuracy
- Scoring this model gave an accuracy measure of 0.889



ADDITIONAL INSIGHT: NEXT STEPS

- To further improve classification, additional modeling approaches may be considered for understanding the relationships between parameters and launch success (or launch cost!)
- Regression analysis maybe useful for uncovering some relationships between the considered data and the cost of a launch, which is the key element for SpaceY's success as a competitor
- SpaceY may also consider a cost analysis for launch site locations

CONCLUSIONS

- Given the data, we can predict the success or failure of a launch by SpaceX with 88.9% accuracy
- With this prediction accuracy, we can continue to make inference about cost expectations for SpaceY in order to be competitive with SpaceX
- SpaceY would likely benefit from position launch sites with similar attributes as SpaceX, for example with a close proximity to a coastline and in a warm location

Thank you!