IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohamed Jassim
16/06/2023

Outline

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

Executive Summary

Summary of methodologies

These project follows these steps:

- Data collection
- Data wrangling
- Exploratory data analysis
- Interactive visual analytics
- Predictive Analysis

Summary of Results

This project produced the following outputs and visulaization

- Exploratory Data Analysis (EDA) results
- Geospatial Analytics
- Interactive dashboard
- Predictive analysis of classification models

Introduction

- SpaceX launches Falcon 9 rockets at a cost of around $62m. This is considerably cheaper than other providers (which usually cost upwards of $165m), and much of the savings are because SpaceX can land, and then re-use the first stage of the rocket.
- If we can make predictions on whether the first stage will land, we can determine the cost of a launch, and use this information to assess whether or not an alternate company should bid and SpaceX for a rocket launch.
- This project will ultimately predict if the Space X Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected via SpaceX’s API at “https://api.spacexdata.com”
- Perform data wrangling
    - After parsing the API HTML output data, the information was compiled into a pandas DataFrame for further analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Various classification methods were used to determine the best approach for estimating the success rate. The methods included logistic regression, SVM, KNN, and decision tree.

Data Collection

- Data was collected via the SpaceX API as well as by webscraping an html table of data from a Wikipedia page on SpaceX launch history.

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- GitHub URL of the completed SpaceX API calls notebook https://github.com/Jassim41/course/blob/f1fa8823e640f366b10f8b53103500398722ac09/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Using BeautifulSoup, html web data was collected from a wiki table, cleaned, and converted into a panda DataFramefor further analysis.

Html Table

[hide] Flight No.	Date and time (UTC)	Version, Booster	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule,[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 ♺ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 ♺ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)

DataFrame

```
df.head()
```

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX\n	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	\nNASA (COTS)\nNRO\n\n	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)\n	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)\n	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)\n	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Github URL for web scraping:
https://github.com/Jassim41/course/blob/f1fa8823e640f366b10f8b53103500398722ac09/jupyter-labs-webscraping.ipynb

Data Wrangling

- Using the pandas DataFrame, the data was cleaned of null values. The original data contained additional outcome parameters such as landing zone type and whether or not the mission attempted to land at all. In order to apply this information to a classification scheme, the landing outcome definitions were compiled into a simple binary result: 0 for failure and 1

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	F
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

GitHub URL of your completed data wrangling: [Link]

EDA with Data Visualization

- Data charts were used to explore the data:
- Scatter plots of Flight Number vs. Payload Mass and Launch Site and Orbit Type
- By coloring the markers by Class (success or fail), we could see if any trends occurred as more flights were conducted.
- Bar chart of success rate for each Orbit type.
- Scatter plot to show relationship between Orbit type and Payload.
- Line chart showing success rate by year

Github notebook completed link:
https://github.com/Jassim41/course/blob/f1fa8823e640f366b10f8b53103500398722ac09/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Some of the SQL Queries performed on the database include:
- Determined the unique Launch Site names
- Displayed 5 records from launch sites with “CAA” in the name
- Determined the total payload mass and avg. payload for specific booster type.
- Determined the total number of successes and failures
- Determined which booster types carried the maximum payload.

Github link for completed notebook:
https://github.com/Jassim41/course/blob/f1fa8823e640f366b10f8b53103500398722ac09/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium was used to explore information related to the launch sites.
- The locations of the launch sites was marked and labeled on a map.
- One location is in California and the other three are in Florida.
- Additional marker clusters were added to showcase the various successful and failed landing outcomes at each site. By zooming in and clicking on each site, you can see more detail as a result of the clusters.
- Polylines were used to show distances between the sites and nearby locations, such as a

Github link for completed notebook:
https://github.com/Jassim41/course/blob/f1fa8823e640f366b10f8b53103500398722ac09/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Using dash, we built a dashboard that is accessible via a web browser when run.
- In the dashboard were a number of items:
- A drop down list allowing the user to select one of the launch sites or select all of them.
- A bar charts showing the success rate of each site or all the sites, depending on the dropdown list selection.
- A slider allowing the user to select different ranges of payload values.
- A scatter plot showing the landing outcome for the launch sites.
- The scatter plot either showed all the data or specific site data for a given payload

Github link for completed notebook:
https://github.com/Jassim41/course/blob/f1fa8823e640f366b10f8b53103500398722ac09/space_dash.py

Predictive Analysis (Classification)

- Four different classification schemes were used on the data.
    1. Logistic Regression
    2. SVM
    3. KNN
    4. Decision Tree
- Each method was optimized via a GridSearchCV function that iterates through a list of parameters to find which combination performed the best.
- A confusion matrix was the plotted for each method when used on the test data to see how well each method performed

Github link for completed notebook:
https://github.com/Jassim41/course/blob/f1fa8823e640f366b10f8b53103500398722ac09/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

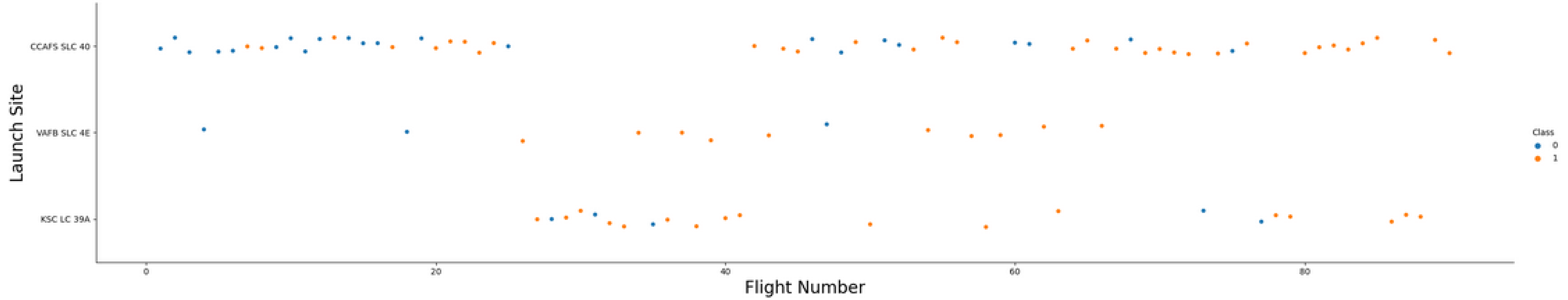
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

6]:

```
### 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()
```



- Show the screenshot of the scatter plot with explanations

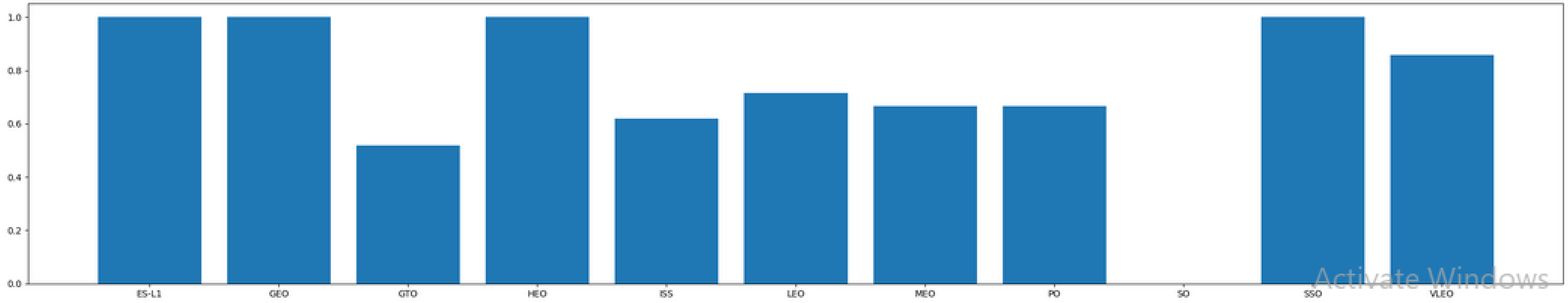
Payload vs. Launch Site

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

- 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).

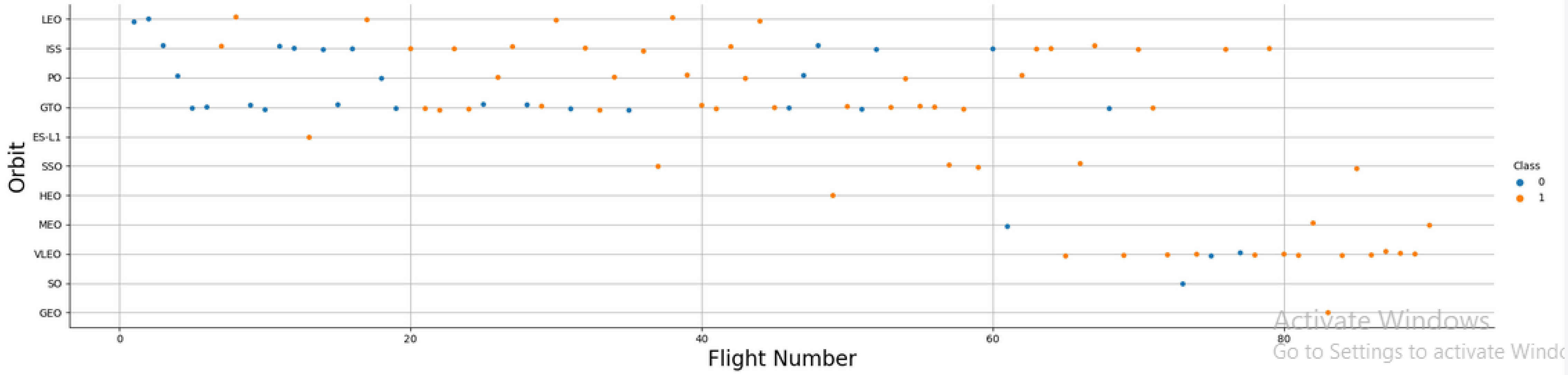
Success Rate vs. Orbit Type

```
# HINT use groupby method on Orbit column and get the mean of Class column
# Group the data by Orbit and calculate the mean of Class column

bardata = df.groupby(['Orbit']).mean()['Class']
x = bardata.keys()
h = bardata.values
plt.bar(x=x, height=h)
plt.show()
```



Flight Number vs. Orbit Type

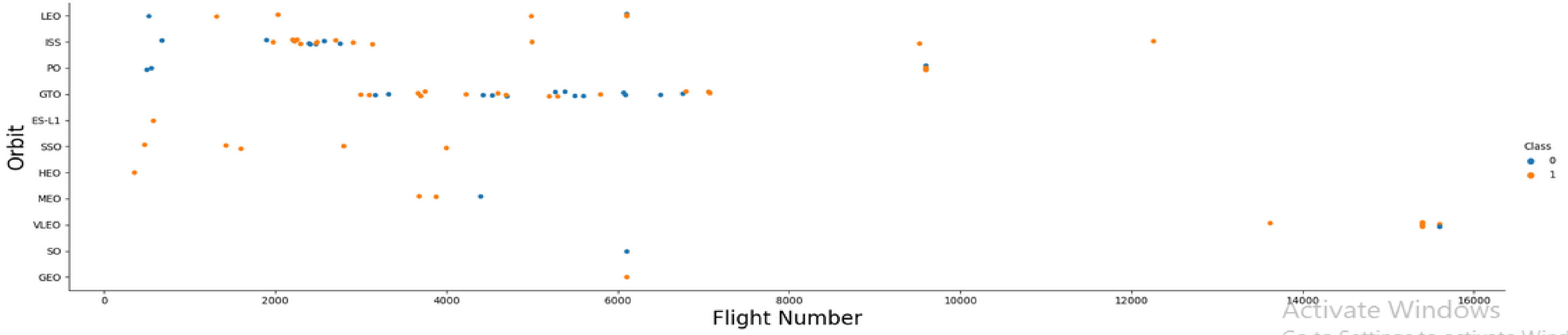
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the cla
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect=4)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.grid()
plt.show()
```



- 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

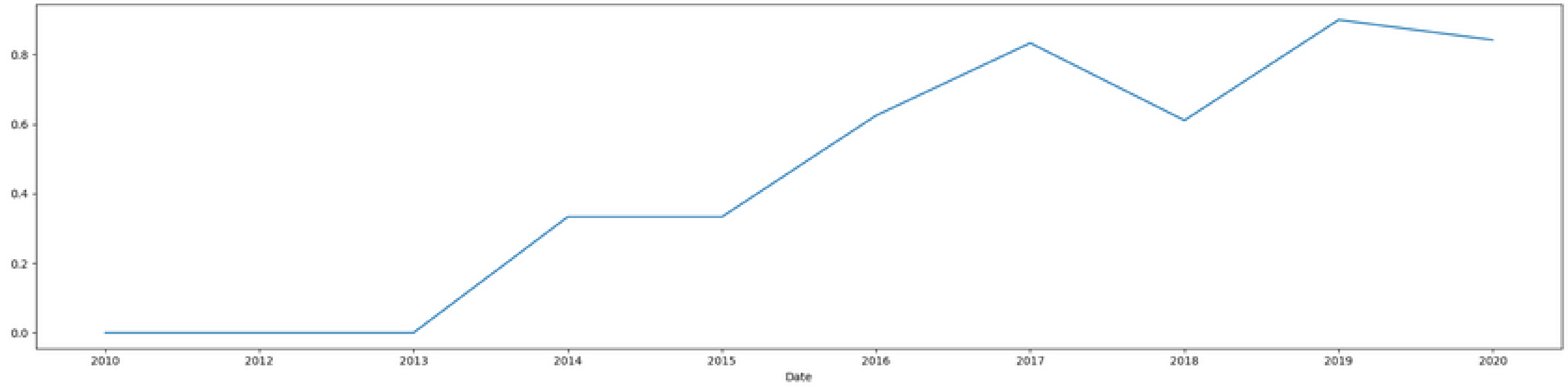
```
# Plot a scatter point chart with x axis to be payload and y axis to be the Orbit, and hue to be the class va
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect=4)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)

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

```
18]:
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
linedata = df.groupby(['Date']).mean()['Class']
x = linedata.keys()
y = linedata.values
sns.lineplot(x=x, y=y)
plt.show()
```



The success rate since 2013 kept increasing till 2020

All Launch Site Names

Task 1

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

- Display the names of the unique launch sites such as CCAFS LC-40,VAFB SCL-4E,KSC LC-39A, CCAFS SLC-40

Launch Site Names Begin with 'CCA'

```
In [9]: %sql Select * from spacextbl where "Launch_Site" like 'CCA%' limit 5
```

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

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Ou
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (para
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 (para
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 a
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No a
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No a

- The list of 5 record with the site names that are begin with the ‘CCA’

Total Payload Mass

Task 3

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

```
[11]: %sql select sum("PAYLOAD_MASS__KG_") as sum_of_payload from  spacextbl where "Customer" ="NASA (CRS)"
```

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

[11]:

sum_of_payload
45596.0

- The sum of the payload mass carried by boosters launched by NASA (CRS) is 45596.0

Average Payload Mass by F9 v1.1

Task 4

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

```
#sql select * from spacextbl
%sql select avg("PAYLOAD_MASS__KG_") as avg_of_payload from  spacextbl where "Booster_Version"= "F9 v1.1"
```

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

avg_of_payload
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2928.4

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
%sql select max("Date") as first_sucessful_landing from spacextbl where "Landing_Outcome"="Success (ground pad)"
```

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

first_sucessful_landing
22/12/2015

- The dates of the first successful landing outcome on ground pad is 22/12/2015.

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 "Booster_Version" from spacextbl where "Landing_Outcome"="Success (drone ship)" and "PAYLOAD_MASS__KG
```

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

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

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are: F9 FTB1022, F9 FTB1026, F9 FTB1021.2, F9 FTB1031.2.

Total Number of Successful and Failure Mission Outcomes

```
[6]: %sql select "Mission_Outcome",count("Mission_Outcome") as Total from spacextbl group by "Mission_Outcome"

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

[6]:

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

- The total number of successful and failure mission outcomes is 100 success and 1 failure(in flight).

Boosters Carried Maximum Payload

[15]:

```
%sql select "Booster_Version" from spacextbl where "PAYLOAD_MASS__KG_" =( select max("PAYLOAD_MASS__KG_") from sp
```

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

t[15]:

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

```
]: %sql select  substr("Date", 4, 2) as month ,"Booster_Version","Landing_Outcome","Launch_Site" from spacextbl wher
```

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

month	Booster_Version	Landing_Outcome	Launch_Site
10	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

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

```
%%sql
select "landing_outcome", count("landing_outcome") as "Count" from SPACEXTBL
where "landing_outcome" like "%Success%" and "Date" between strftime('%d/%m/%Y', "2010-06-04") and strftime('%d/%
group by "landing_outcome"
order by "Count" Desc;
```

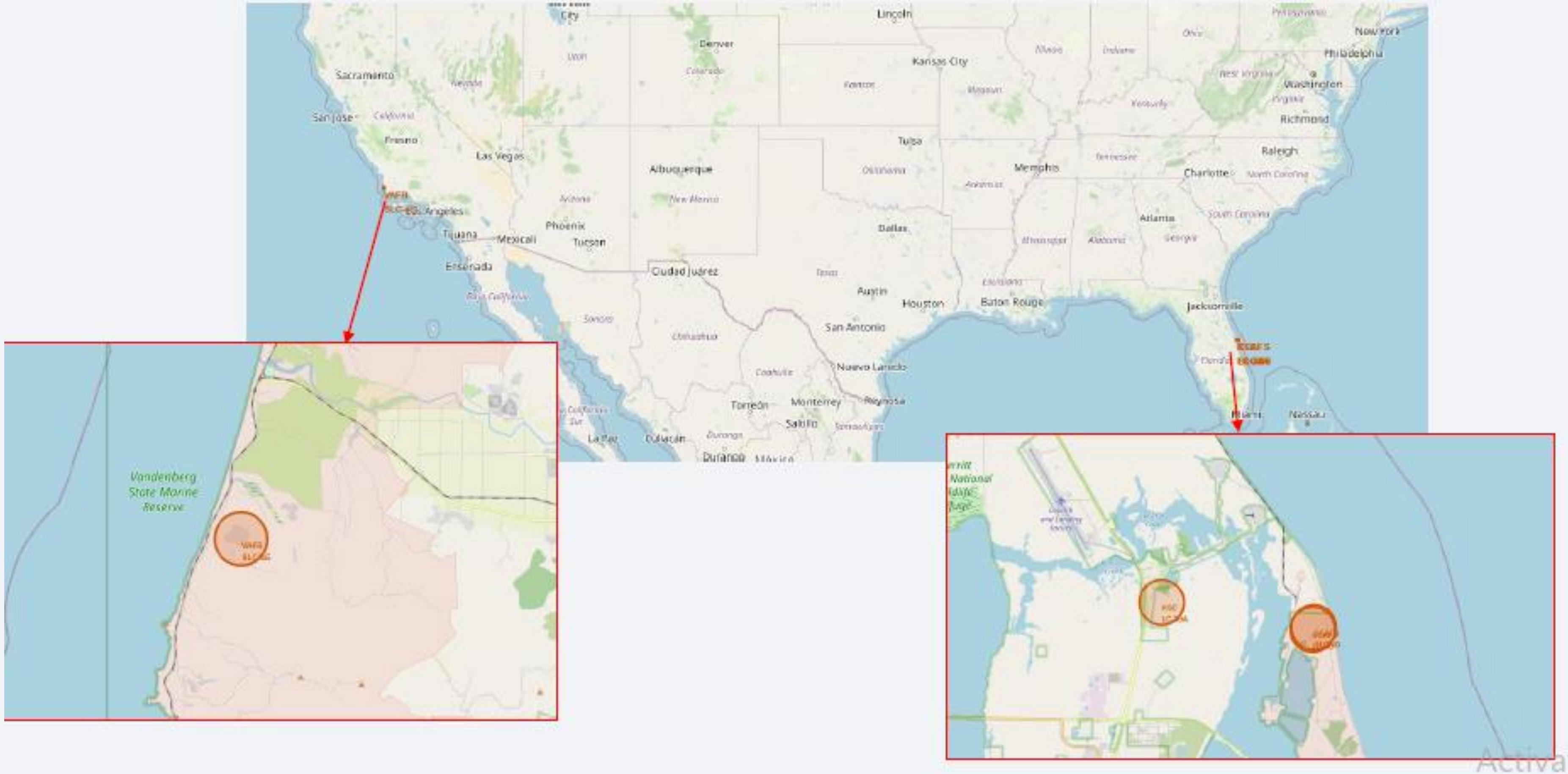
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	7

Section 3

Launch Sites Proximities Analysis

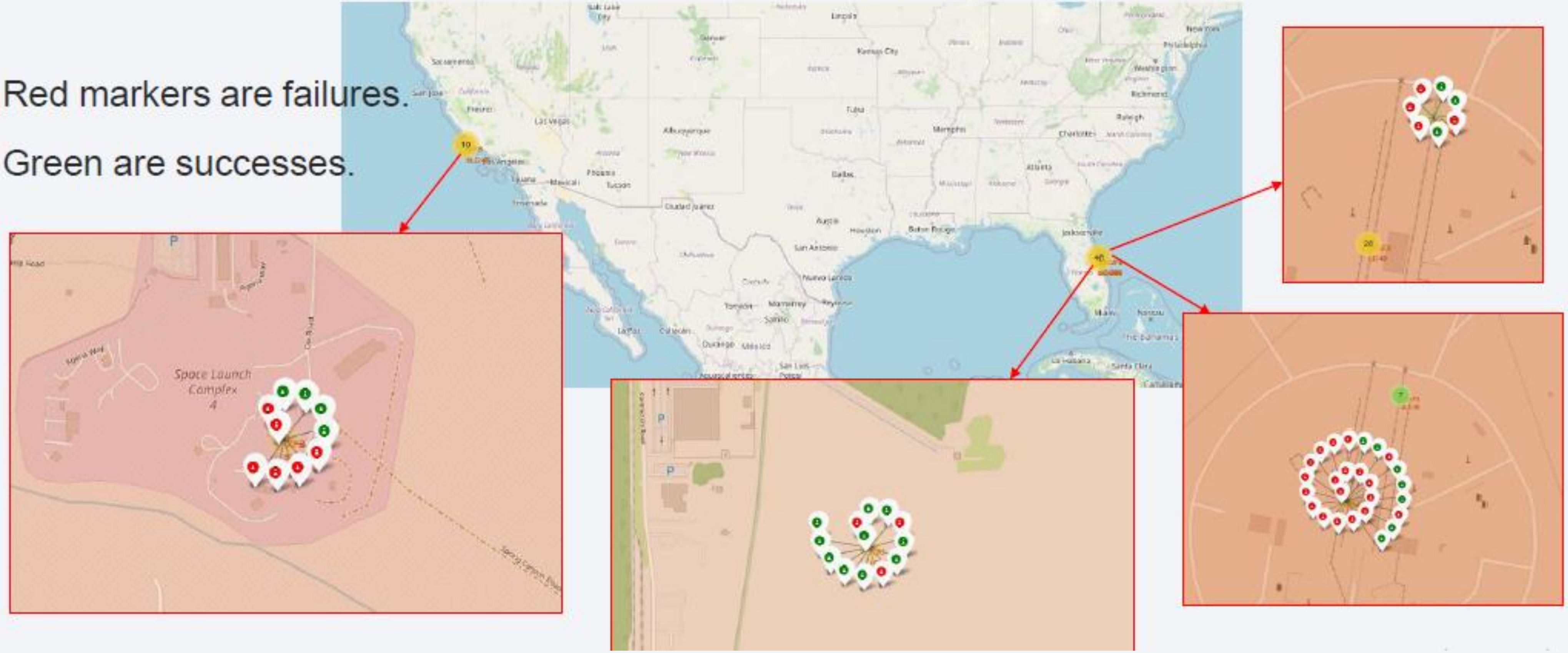
FOLIUM MARKERS SHOWING THE LAUNCH SITE LOCATIONS



Folium Marker clusters showing the success and failures in each sites

Red markers are failures.

Green are successes.



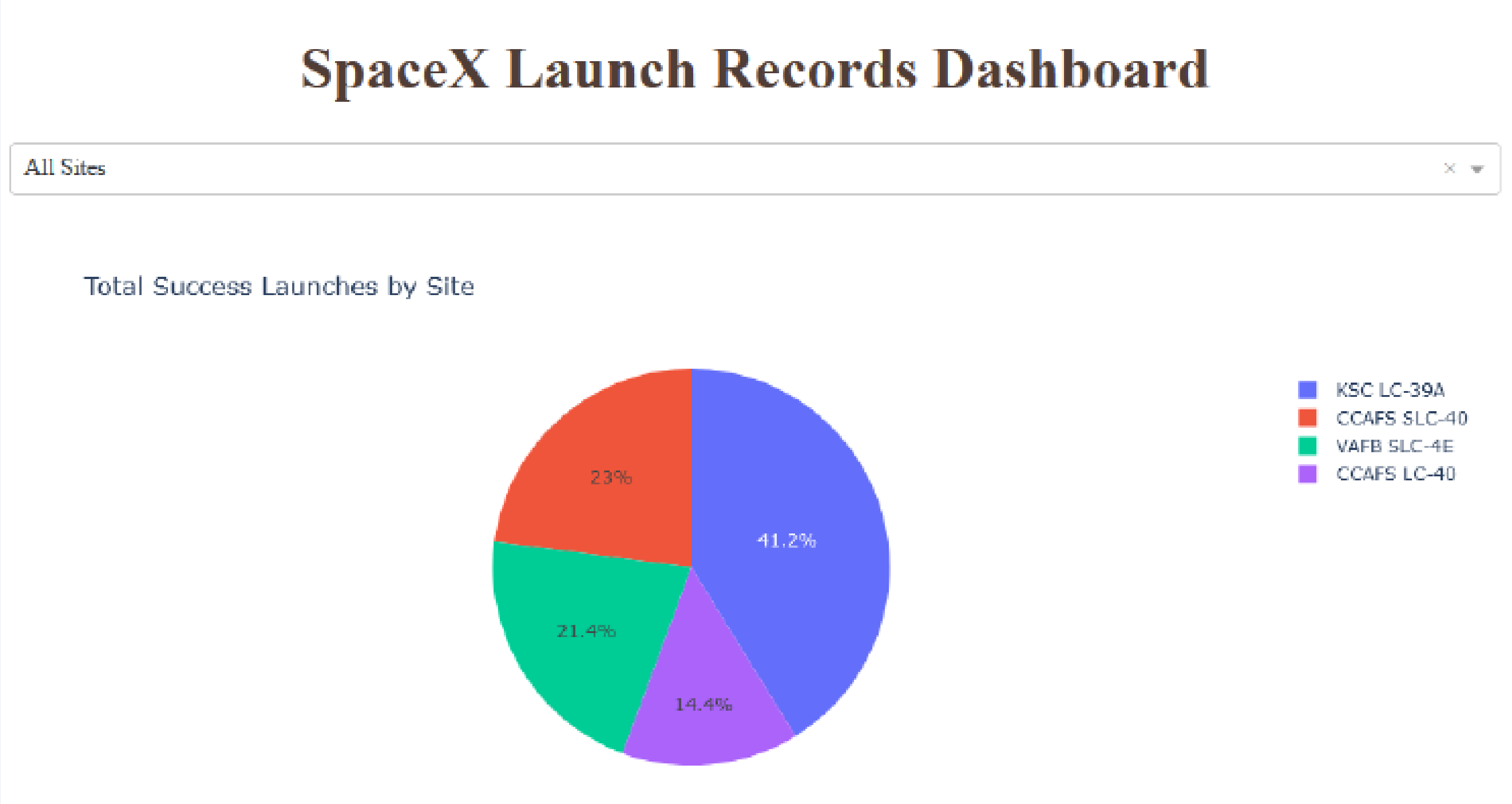
<Folium Map Screenshot 3>

- The distance is shown as 0.51 km

Section 4

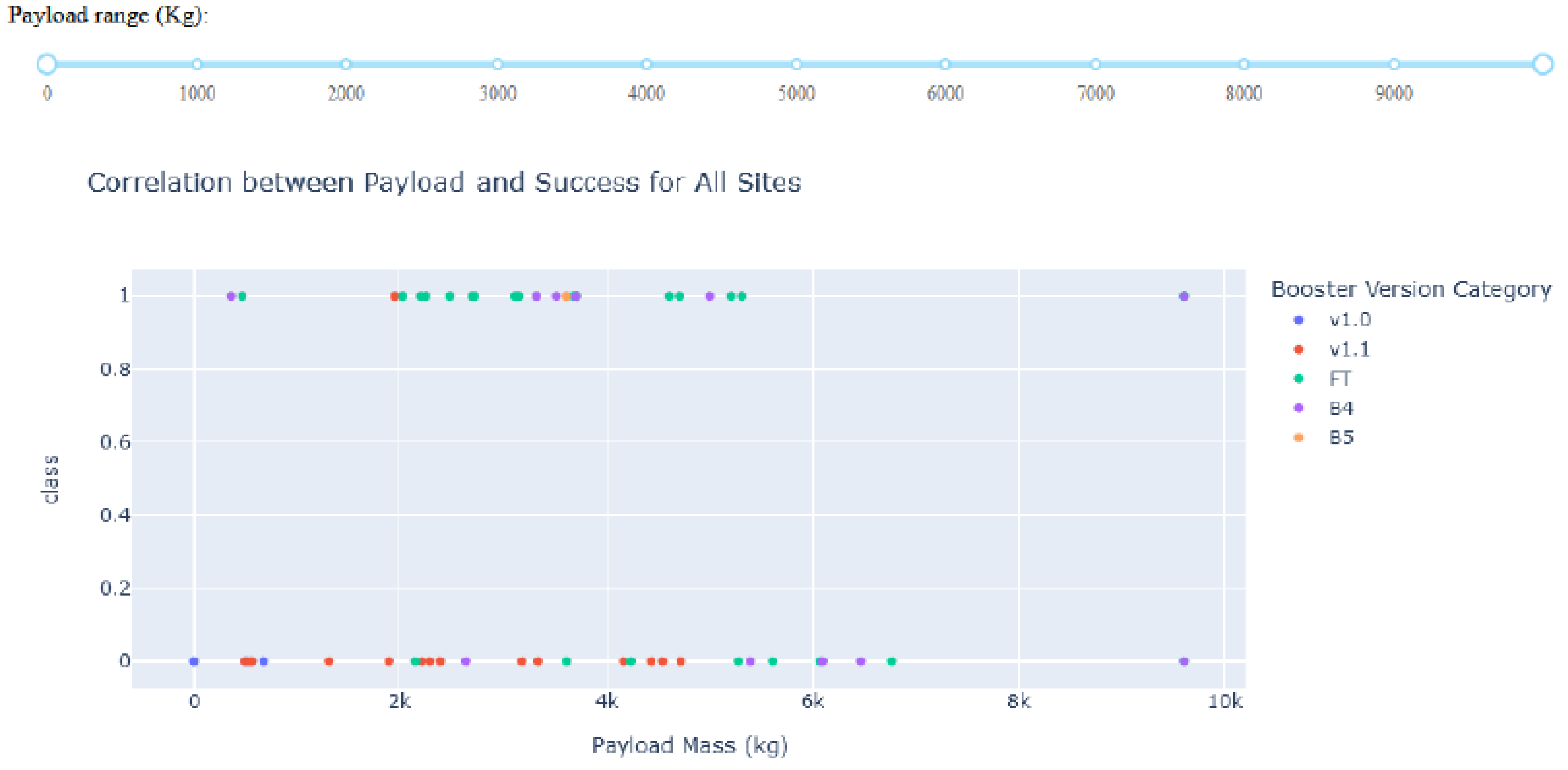
Build a Dashboard with Plotly Dash

SPACEX LAUNCH RECORDS DASHBOARD



- KSC LC-39A has the highest rate of successful launches.

Correlation between payload and success for all sites



- Both successful and failed launches were recorded across the range of tested payloads. The most successful booster version appears to be in the FT category.

Dashboard

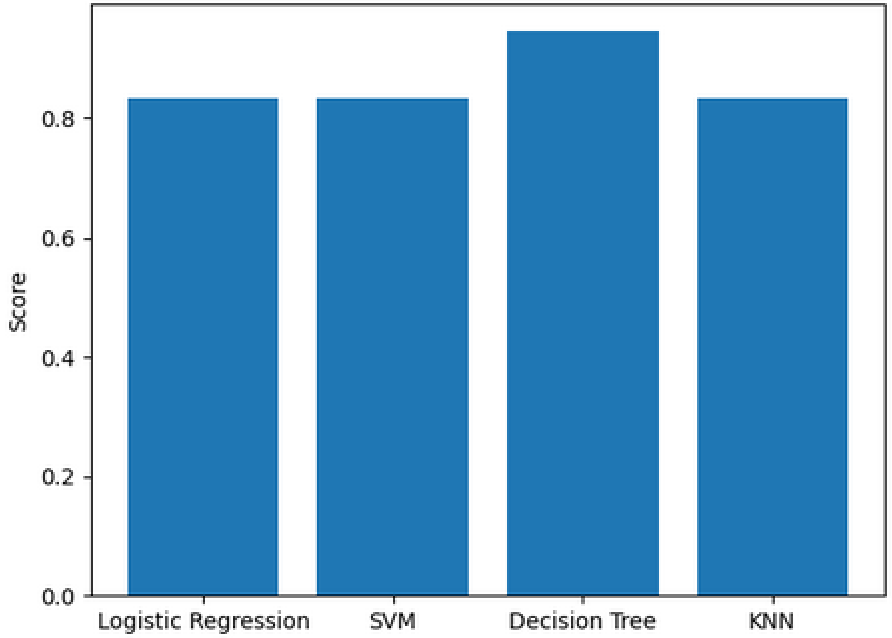
Payload range (Kg):
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Correlation between Payload and Success for All Sites
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Section 5

Predictive Analysis (Classification)

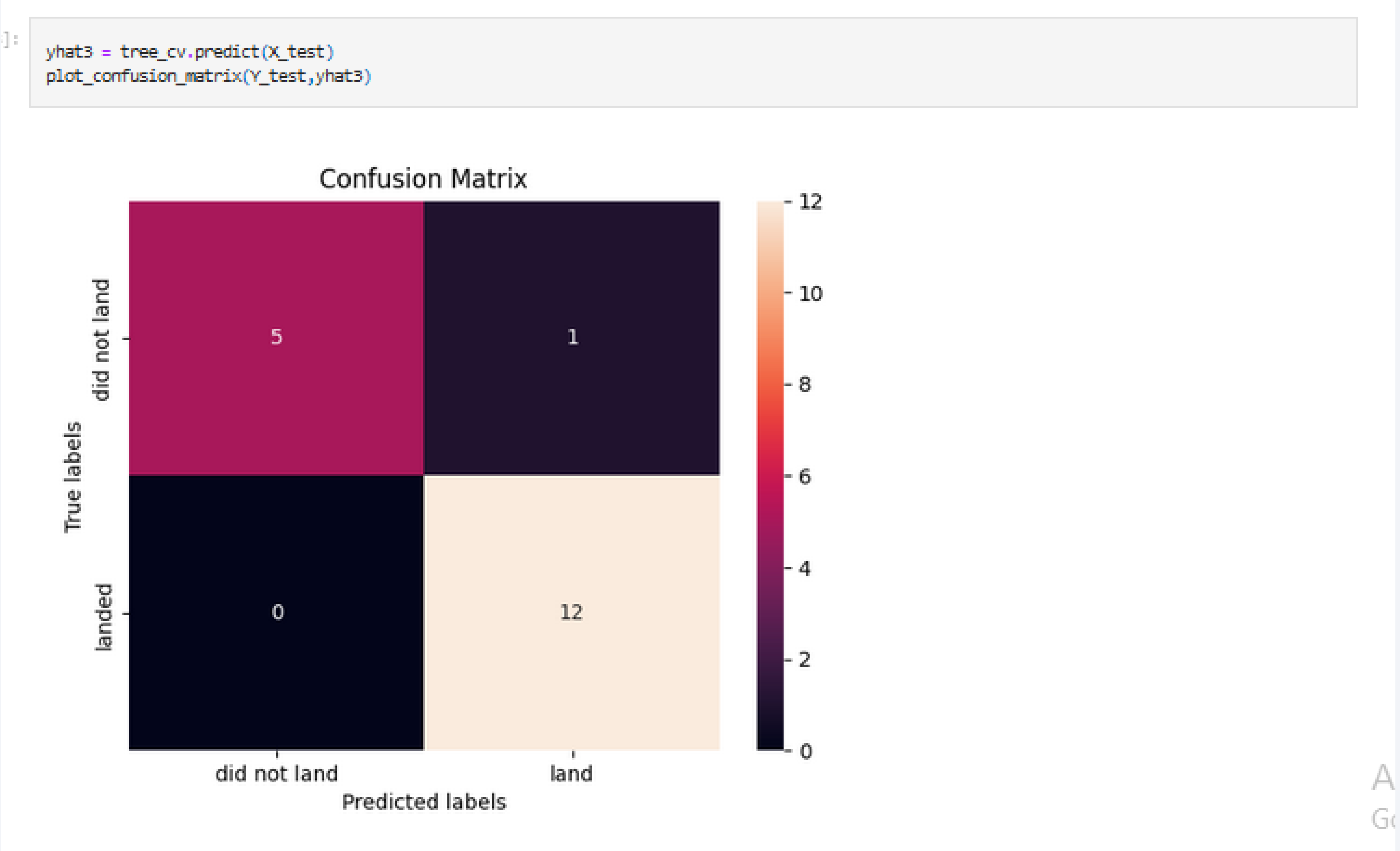
Classification Accuracy

```
In [99]:
scorelist = [logreg_score, svm_score, tree_score, knn_score]
methodlist = ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN']

plt.bar(x=methodlist, height=scorelist)
plt.ylabel('Score')
plt.show()
```



- The Logistic, KNN, and SVM models all resulted in an accuracy score of 0.833. The Decision Tree performed the best with an accuracy score of 0.9444 when used on the test data.

Confusion Matrix



- The confusion matrix for the decision tree model shows why it performed the best. For the 6 failed landings, only 1 record was falsely labeled as successful.
- For the successful landings, all 12 records were correctly labeled.

Conclusions

- We have shown that publicly available data can be assessed to draw meaningful conclusions. Using the data from SpaceX, we have learned about the various types of boosters, timelines, launch sites, landing outcomes, and payloads for various flights.
- Using a few machine learning models, we were able to assess some parsed data related to whether a particular rocket stage would land or not.
- Logistic, SVM, KNN, and Decision Tree classifiers were used on the data.
- It was found that the decision tree model gave the best prediction for the landing outcome based on the data at hand.

Thank you!