Winning Space Race with Data Science

Seoyoung Cho
05/17/2023

Outline

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

Executive Summary

- Perform Data collection, wrangling to understand and prepare the necessary data.
- Use EDA to analyze the data
- Visualize the data analysis using Folium, and Plotly
- Calculate the performance of machine learning model prediction

Introduction

- A newly built “Space Y” company wants to compete with the “Space X” company founded by Allon Musk. The task is to determine the price of each launch, gather information, create a dashboard for the team, and predict whether the Space X will reuse the first stage of rocket launch using machine learning algorithms.

Section 1

Methodology

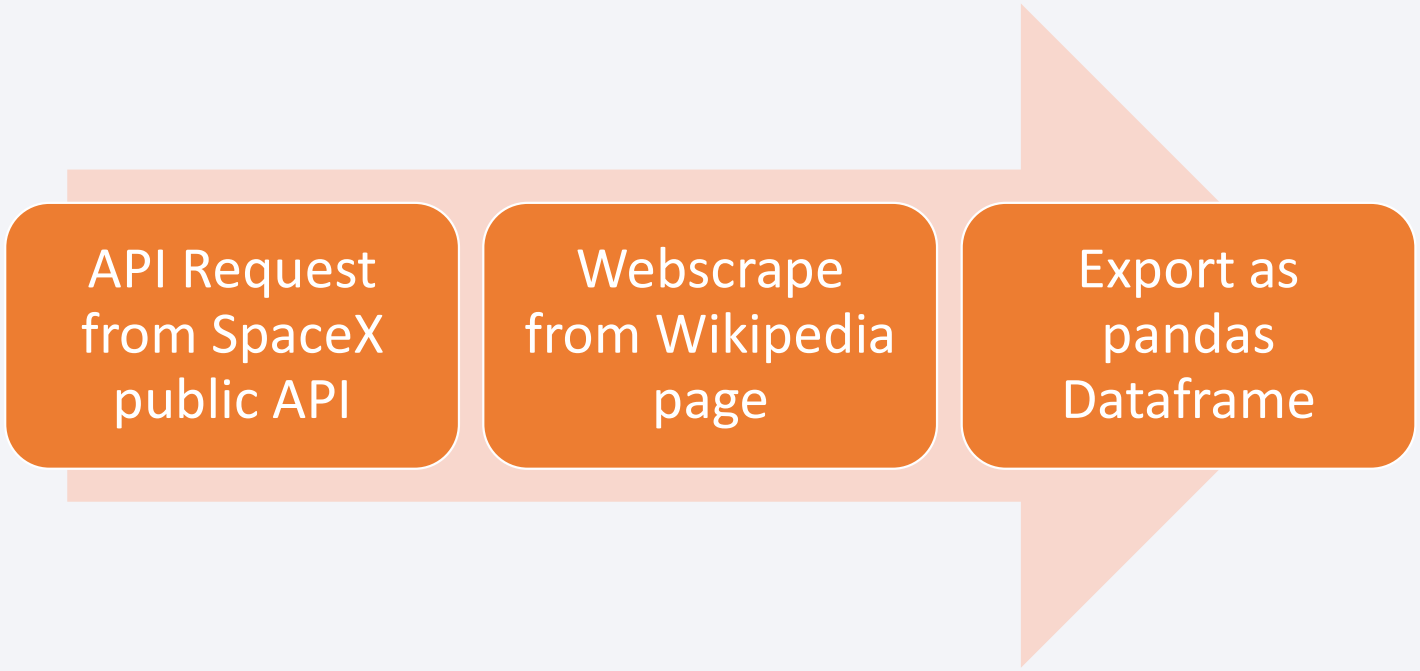
Methodology

Executive Summary

- Data collection methodology:
  - Data collected from SpaceX public API and Wikipedia page.
- Perform data wrangling
  - Data are standardized, hot encoded when necessary
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used GridSearchCV to compare 4 models (Logistic Regression, SVM, Decision Tree, KNN)

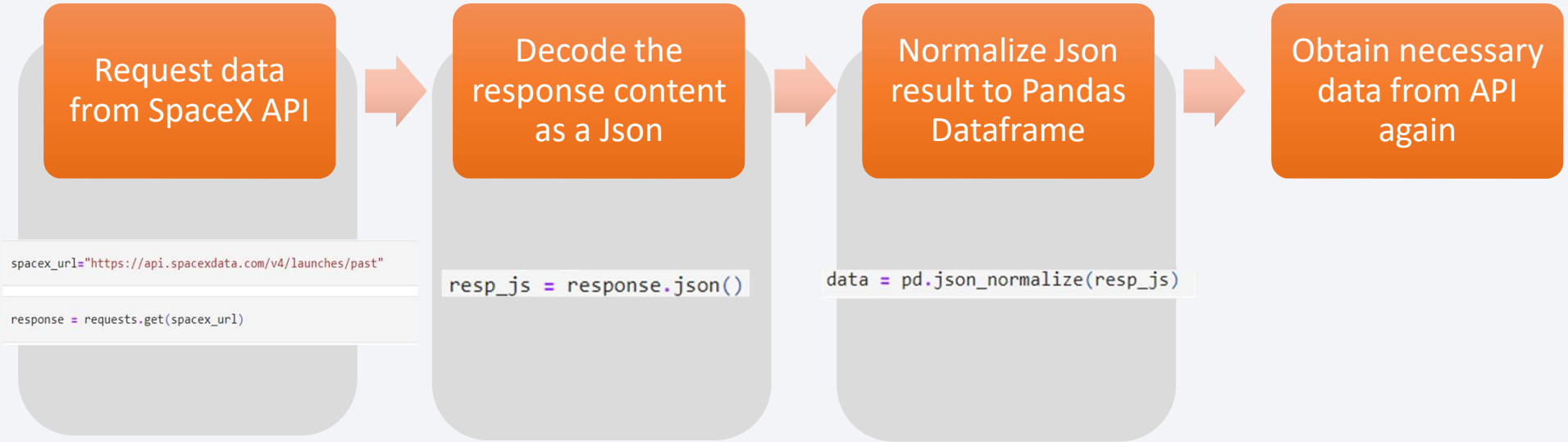
Data Collection

- Data sets collected from SpaceX public API request, SpaceX Wikipedia web scrape.



Data Collection – SpaceX API

- https://github.com/scho249/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API.ipynb



Data Collection - Scraping

- https://github.com/scho249/Applied-Data-Science-Capstone/blob/master/Data_Collection_with_Webscraping_lab.ipynb

```
# Use BeautifulSoup() to create a BeautifulSoup object
soup = BeautifulSoup(response.content, "html.parser")
```

```
df=pd.DataFrame(launch_dict)
```


```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

```
for th in first_launch_table.find_all('th'):
    col = extract_column_from_header(th)
    if col and len(col) > 0:
        column_names.append(col)
```

Data Wrangling

- Data processed by performing EDA and determining training labels:

1. Figure out the shape and size of the dataframe (df.shape())
2. Determine the object types of each column in the dataframe (df.dtypes())
3. Find out the number of occurrences of major variables such as Orbit, LaunchSite, and Outcome using value_counts()
4. Determine the success rate by converting the outcomes in numeric form. (fail = 0, success = 1)

- https://github.com/scho249/Applied-Data-Science-Capstone/blob/master/Data_Wrangling_lab.ipynb

EDA with Data Visualization

- Scatter point chart used to visualize the relationships between:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Flight Number and Orbit type
  - Payload and Orbit type
- Bar graph to find which orbits have high success rate
- Line chart to get the average launch success trend

https://github.com/scho249/Applied-Data-Science-Capstone/blob/master/EDA_with_Visualization_lab.ipynb

EDA with SQL

Some queries include:

- Using bullet point format, summarize the SQL queries you performed
- Getting the Unique launch site
- 5 records of launch sites beginning with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by F9 v1.1
- Date of the first successful landing in ground pad
- https://github.com/scho249/Applied-Data-Science-Capstone/blob/master/EDA_Sql_lab.ipynb

Build an Interactive Map with Folium

- Circles and markers added to mark all launch sites. Markers are color coded based on the launch outcome (green: success, red: fail)
- Lines added to show a distance between the launch site and the closest coastline, highway, railway, and city.

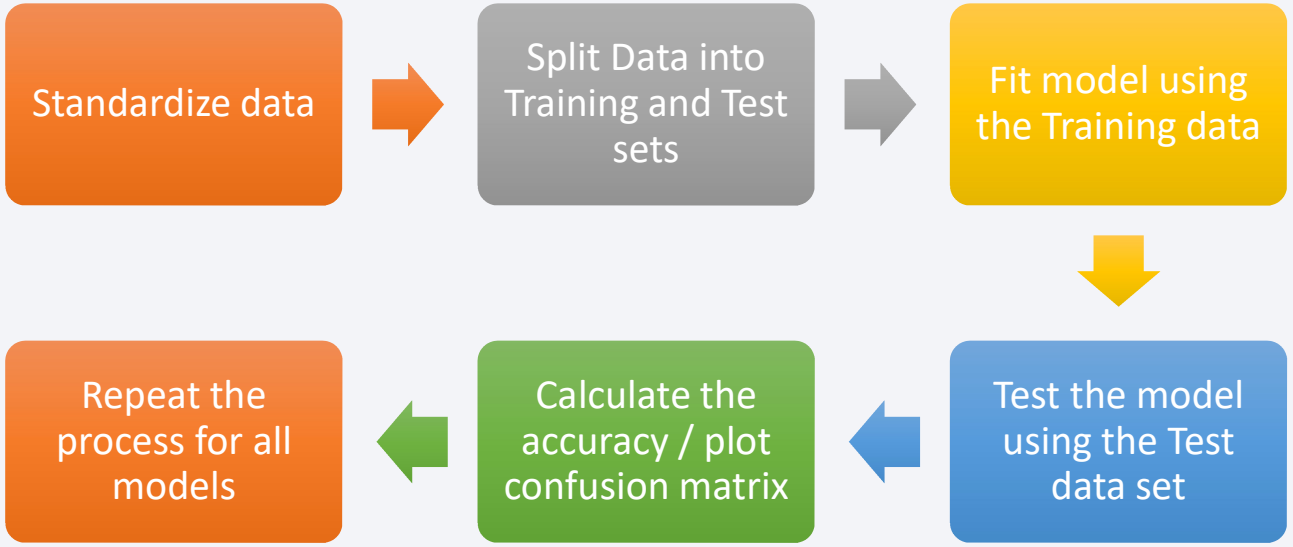
https://github.com/scho249/Applied-Data-Science-Capstone/blob/master/Interactive_Visual_Analytics_with_Folium_lab.ipynb

Build a Dashboard with Plotly Dash

- Pie charts used to visualize the successful launches of Space X for all and each launch sites
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

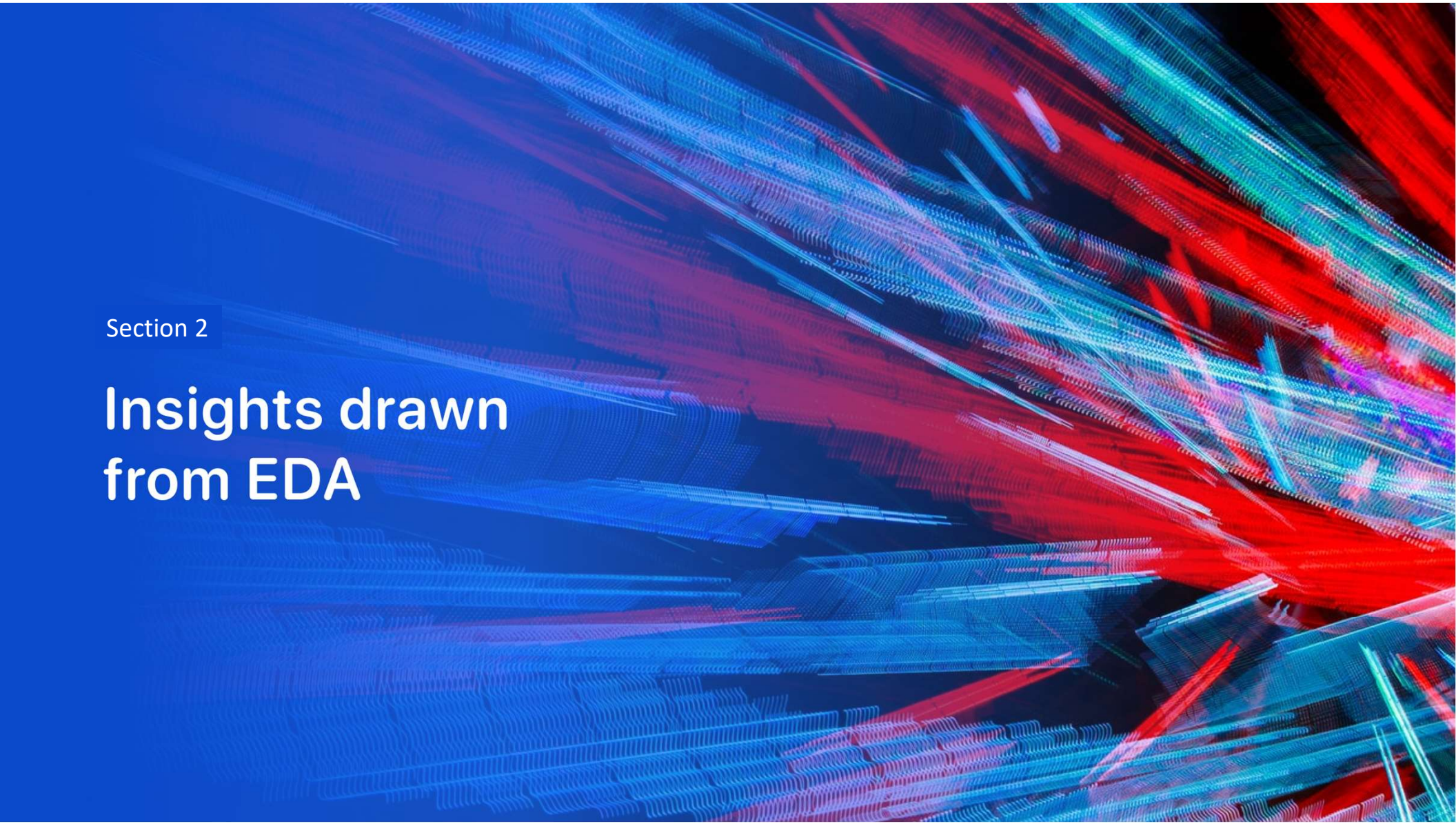
Predictive Analysis (Classification)

- Built Logistic Regression, SVM, Decision Tree, KNN models to compare the accuracy of prediction
- https://github.com/scho249/Applied-Data-Science-Capstone/blob/master/Machine_Learning_Model_lab.ipynb



Results

- General trend of success rate has increased over time. Current rate is about 80%.
- Launch sites are located near coast line. KSC LC-39A site has the highest success rate.
- The Decision Tree model had the highest accuracy of predicting the landing outcome

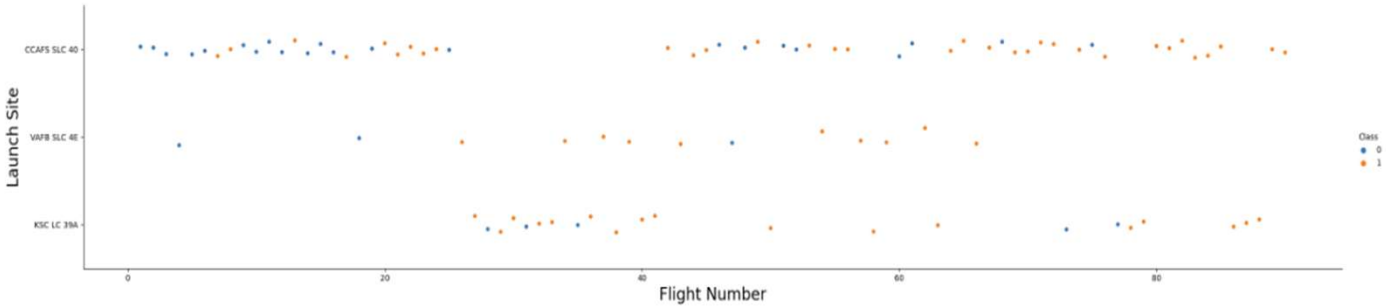


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

```
]:  # 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
    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()
```

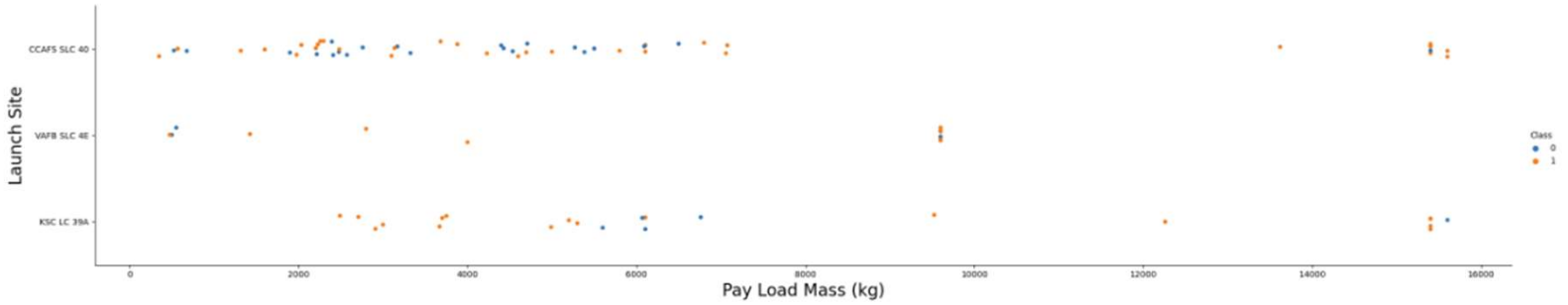


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

- In general, as the flight number increases, the success rate also increases.
- For the site CCAFS SLC 40, clearly the higher flight number resulted in the higher success rate.

Payload vs. Launch Site

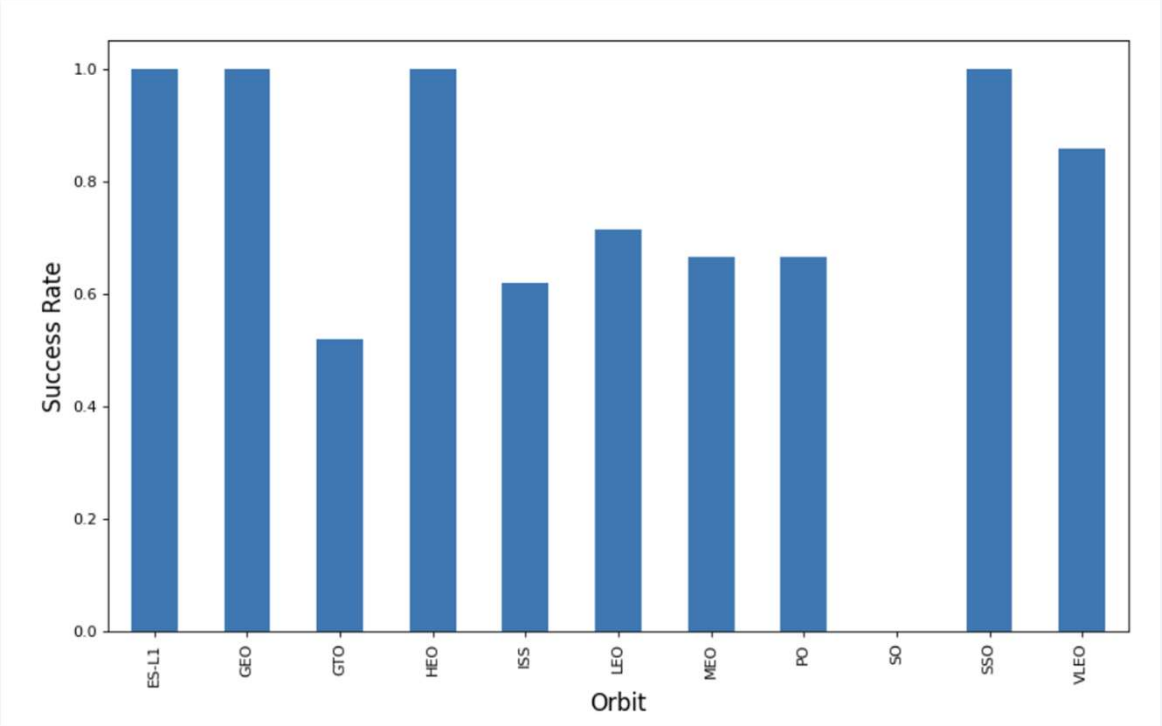
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()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



- For the Launch Site VAFB SLC 4E, there are no rockets launched with over 10000kg of payload mass. In general, more rockets with pay load of under 8000kg are launched for all sites.

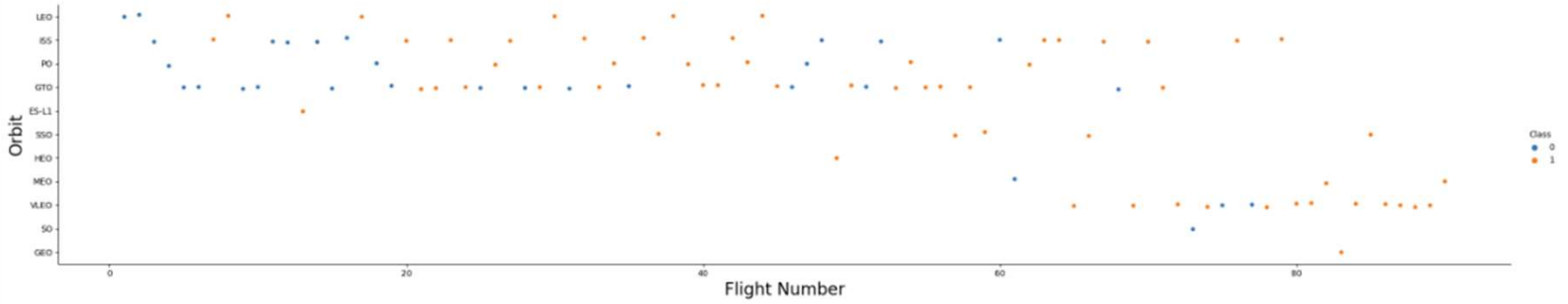
Success Rate vs. Orbit Type

- Orbit type ES-L1, GEO, HEO, SSO have the highest success rate of 1.0.
- Orbit type with the lowest success rate is SO at about rate of 0.



Flight Number vs. Orbit Type

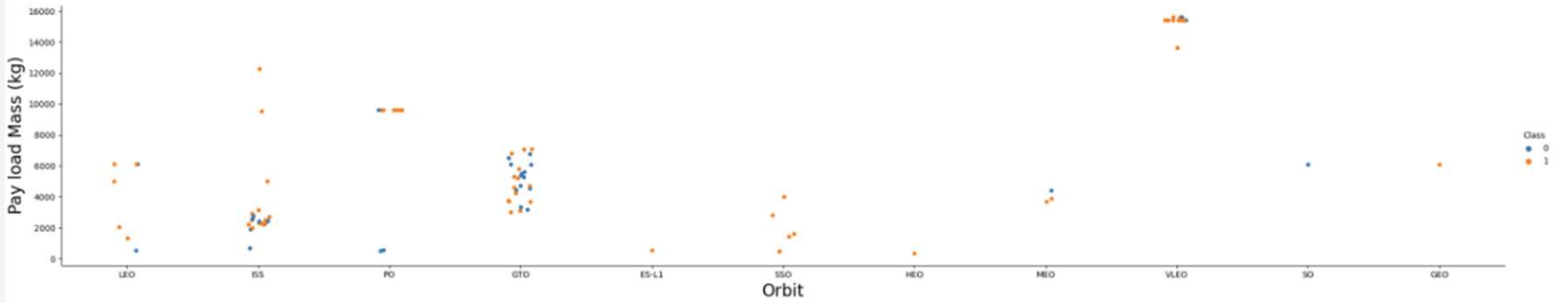
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



- In general, the success rate increases as the flight number increases for all orbit type.
- Orbit type SSO has the highest success rate.
- Orbit type LEO only performed with lower Flight Number and orbit type VLEO performed only with higher Flight Number

Payload vs. Orbit Type

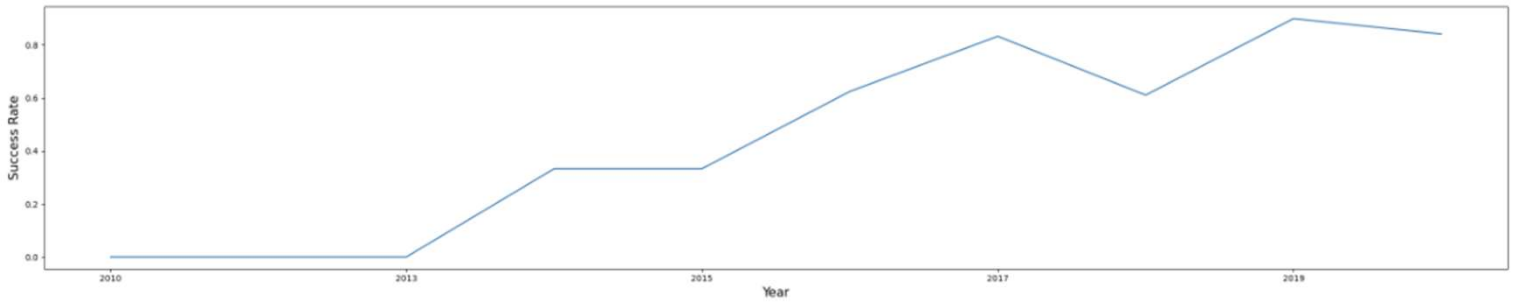
```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="PayloadMass", x="Orbit", hue="Class", data=df, aspect = 5)
plt.xlabel("Orbit",fontsize=20)
plt.ylabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```



- For each orbit type, it seems like only the certain range of pay load mass is included. (EX: GTO has payload Mass ranging from 3000 – 7000kg, VLEO ranging from 14000 – 16000kg)

Launch Success Yearly Trend

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
df.groupby(['Date']).mean()['Class'].plot(kind='line')
plt.xlabel("Year",fontsize=15)
plt.ylabel("Success Rate",fontsize=15)
plt.show()
```



- In general, the success rate is increasing. The highest success rate is in 2019.
- Although the trend is increasing in general, the performance dropped in 2018 in comparison to the success rate in 2017 and 2019.

All Launch Site Names

- Using the DISTINCT command, extracted all the unique launch_Site values from the SPACEXTBL

```
%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'

- Use ‘LIKE’ and the wildcard ‘%’ to perform pattern matching which allows to find all launch site beginning with ‘CCA’
- Use ‘LIMIT’ to only present 5 records.

```
%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_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
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 attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- ‘SUM’ function to add all the Payload Mass where customer is ‘NASA (CRS)’.
- The total payload mass calculated is 45596kg

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

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

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- AVG() function to calculate the mean payload mass of Booster version ‘F9 v1.1’
- The average payload mass is 2928.4kg

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

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

AVG_Payload_Mass
2928.4

First Successful Ground Landing Date

- MIN() function gives the earliest date where the Landing Outcome is ‘Success (ground pad)’
- The first successful date was 01-5-2017

```
%sql SELECT MIN(DATE) as First_successful FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)'
```

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

First_successful
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' AND (PAYLOAD
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- Booster versions F9 FT B1022, F9 FT B1026, F9 FT B1021.2 F9 FT B10312.2had successful Drone Ship landing with Payload between 4000 and 6000kg.

Total Number of Successful and Failure Mission Outcomes

- Use COUNT() and GROUP BY to count the number of outcomes by the Mission Outcome. Total of 100 successful outcome, and 1 failure.

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

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

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

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT PAYLOAD_MASS__KG_ FROM SPACE
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

```
%sql SELECT substr(Date, 4, 2) as Month, Booster_Version, Launch_Site, "Landing _Outcome" FROM SPACEXTBL WHERE "Landing _Out
```

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

Month	Booster_Version	Launch_Site	Landing _Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- ‘Success’ had the highest number of count 20, followed by ‘Success (drone ship)’ with 8 counts and ‘Success (ground pad)’ with 6 counts

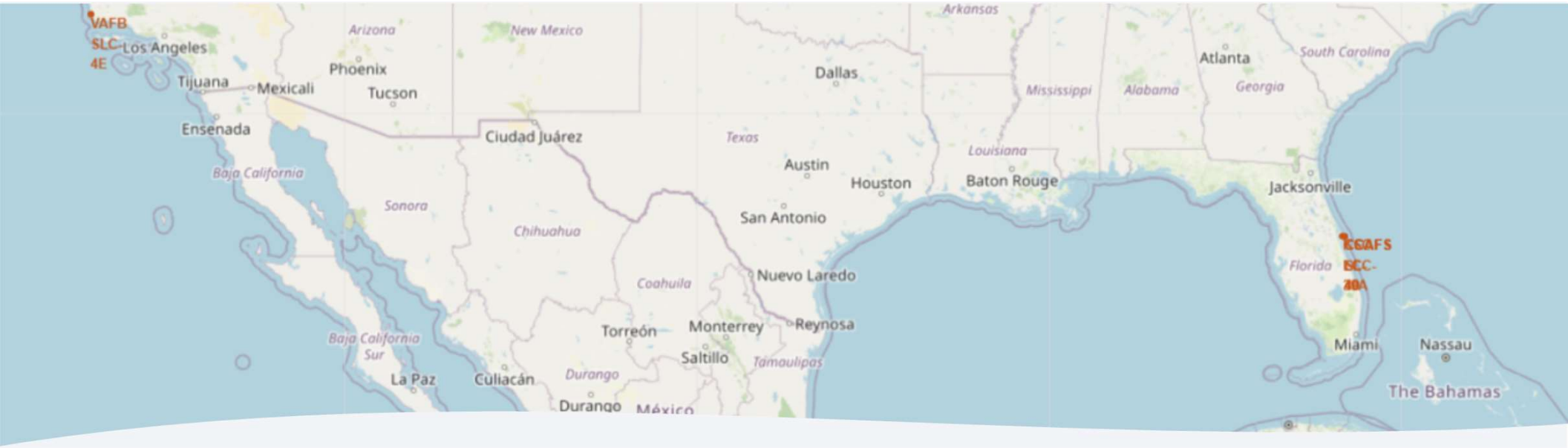
```
%sql SELECT "Landing _Outcome", COUNT(*) as "Count" FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "%Success%" AND Date Betwee
```

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

Landing _Outcome	Count
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

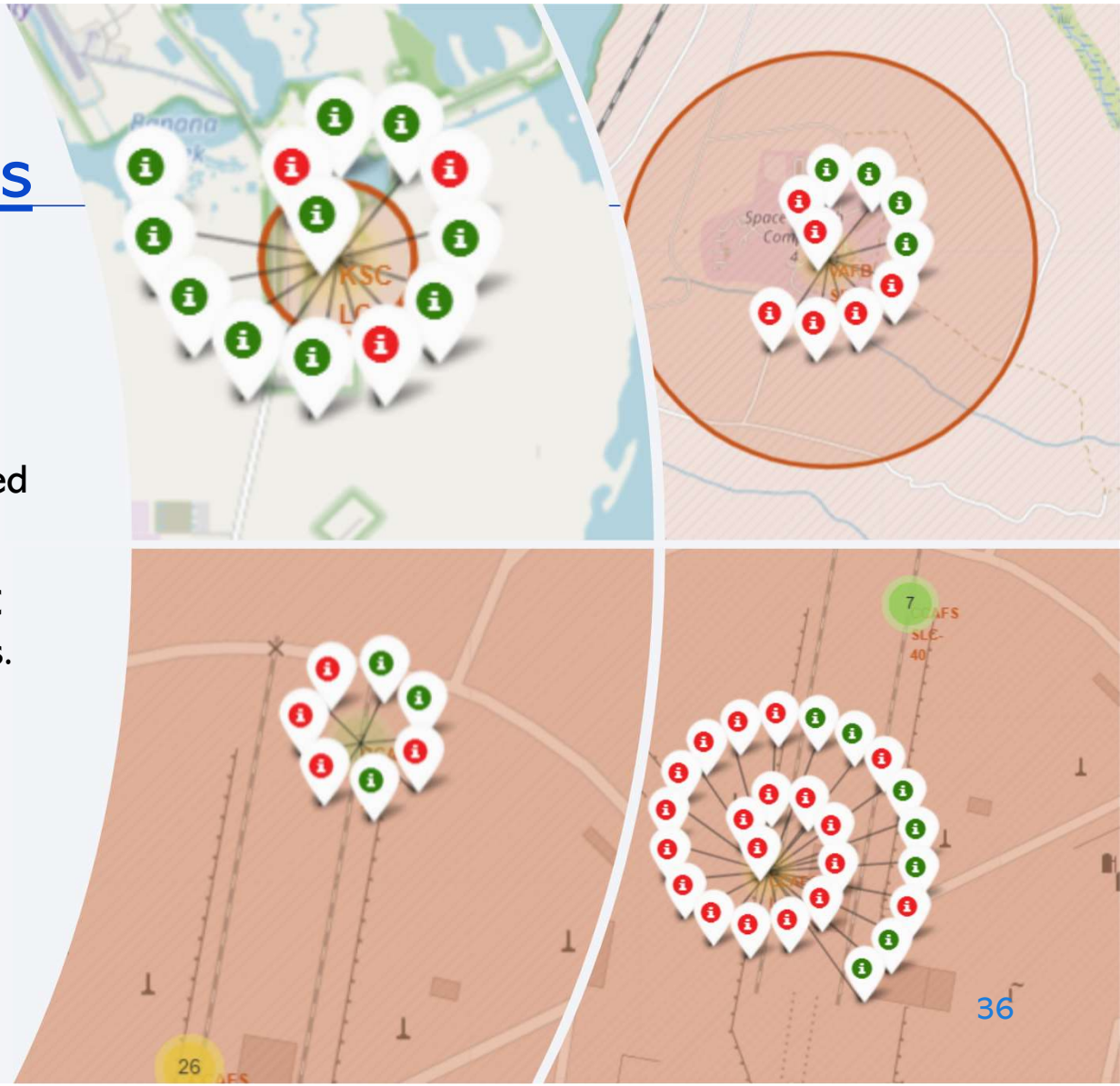


Launch Sites marked on Global Map

- Circles and Markers placed based on the coordinates of each launch site.
- All launch sites are placed very close to the coastal line and above the equator line.
- Launch Site VAFB SLC-4E is in the West U.S (in California) and the remaining sites are in the South-East U.S (Florida)

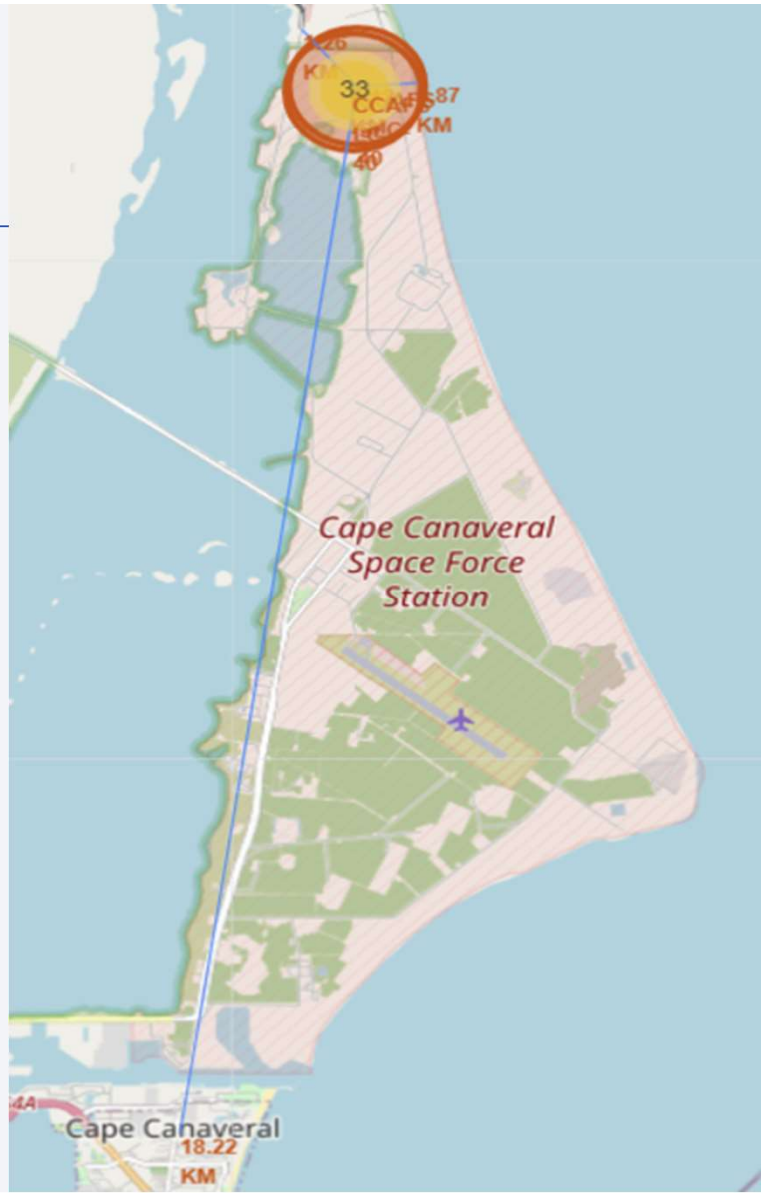
Success/Failed Launches for each Launch Site

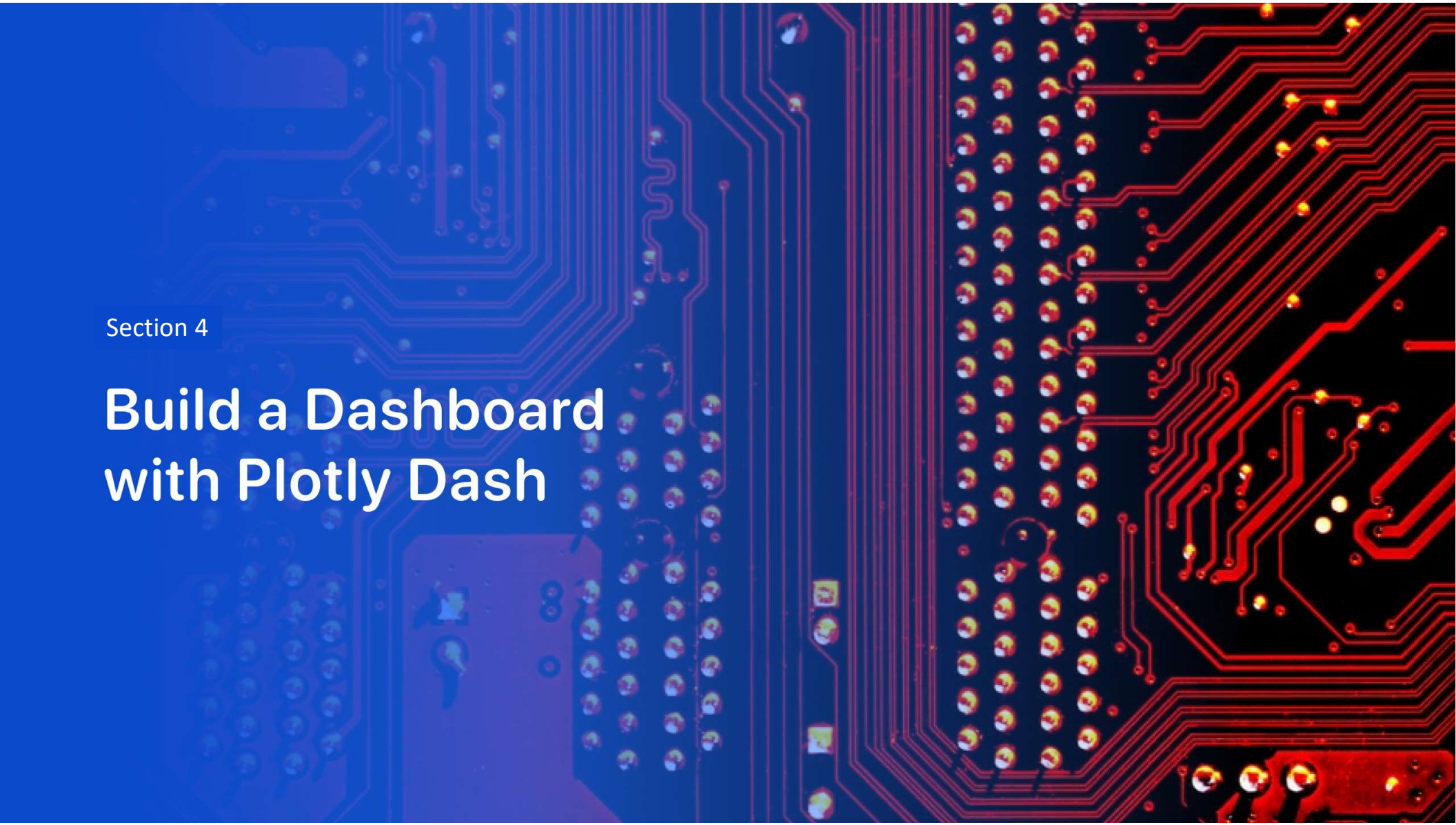
- Green markers indicate Successful launches and Red markers indicate Failed launches
- We can see that the success rate of KSC LC-39A is higher that other launch sites.



Distance to Proximities

- For CCAFS SLC-40, the distance from the launch site to:
  - Coastal line is 0.87km
  - Highway is 1.26km
  - Closest City (Cape Canaveral) is 18.22km



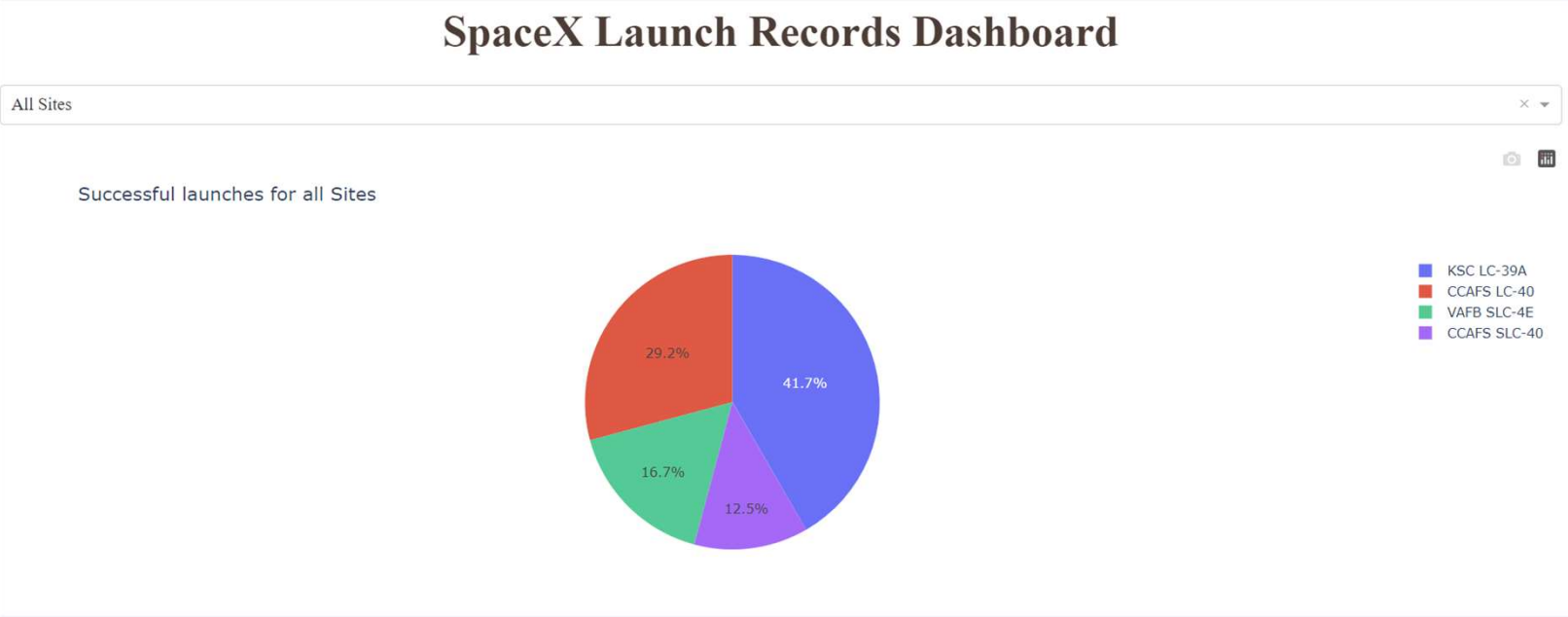


Section 4

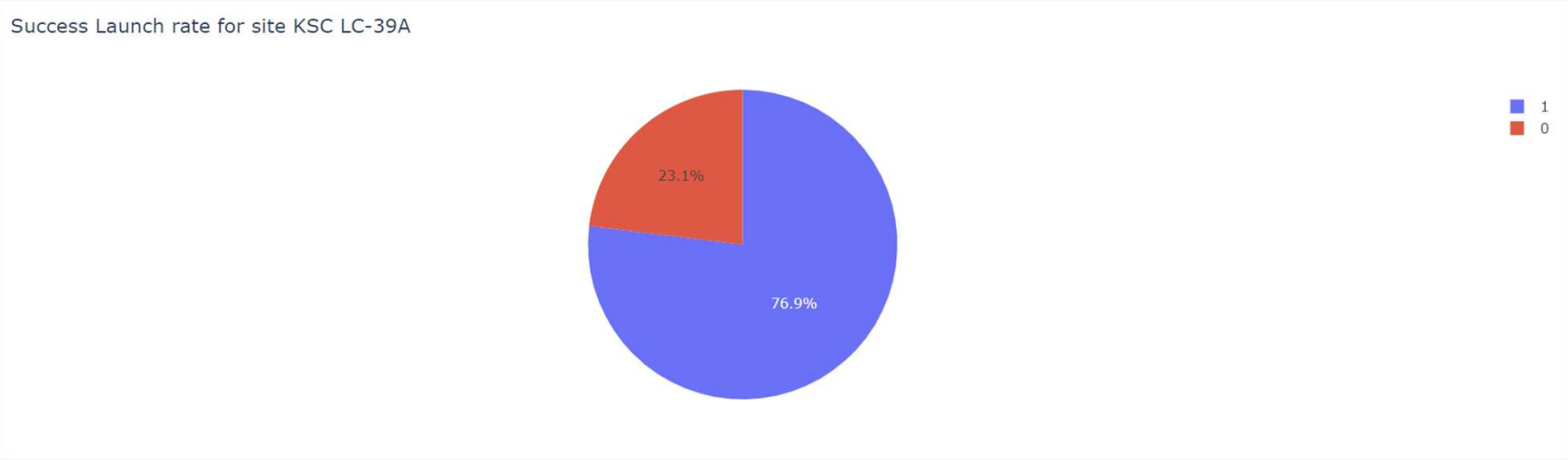
Build a Dashboard with Plotly Dash

Successful Launch rates for all Launch Sites

- KSC LC-39A shown to have the highest success rate with 41.7% rate.
- The result is consistent with the result from the Folium map (slide p. 36)



<Dashboard Screenshot 2>



- KSC LC-39A has the highest success launch ratio. The pie chart shows the success rate is 76.9%.

<Dashboard Screenshot 3>

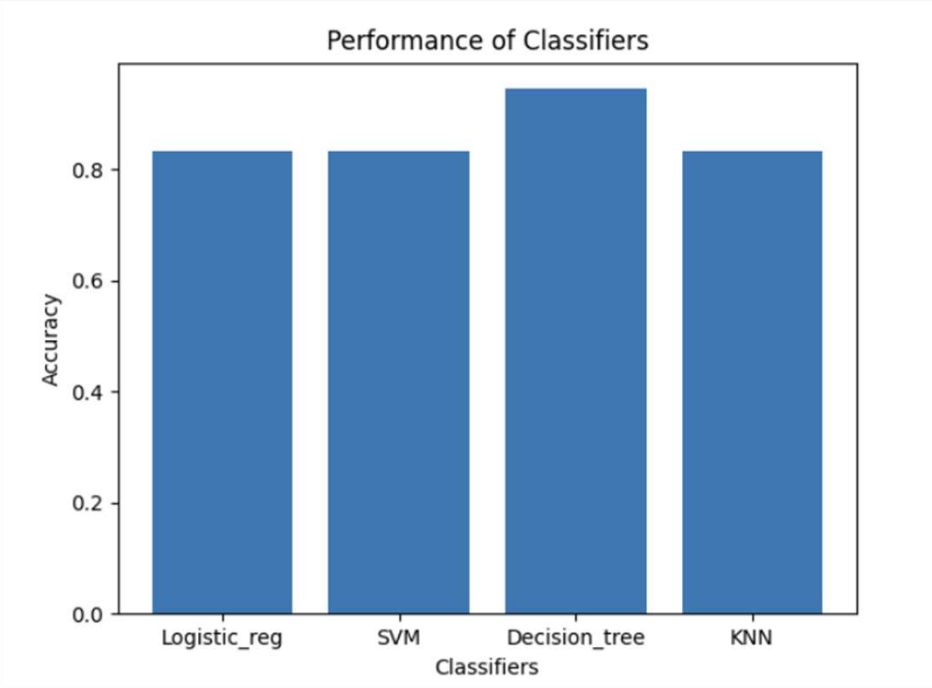
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

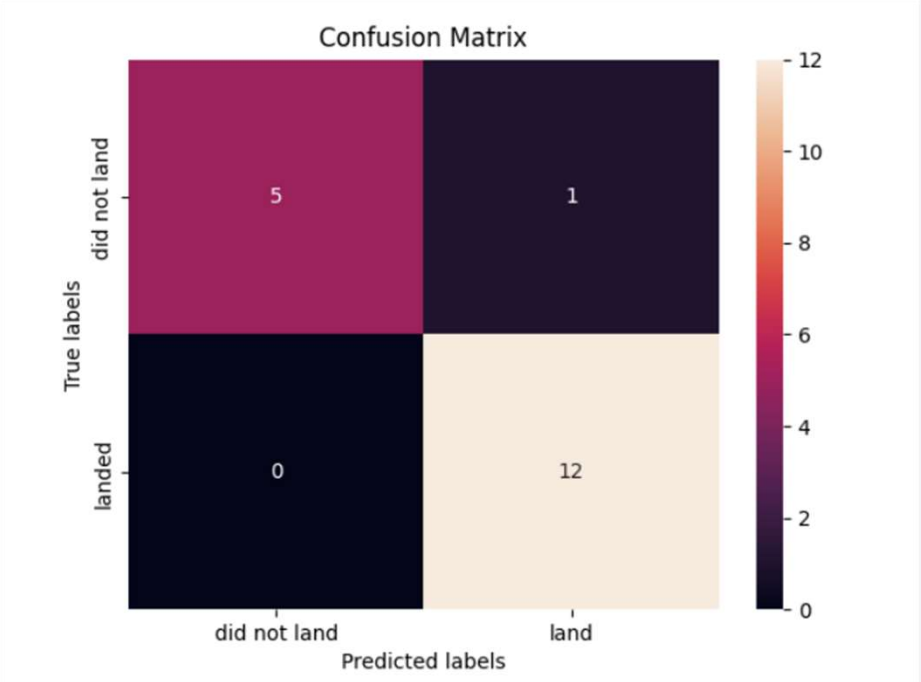
Classification Accuracy

- Based on the bar chart, a classification model with the highest accuracy is Decision Tree classifier with accuracy at about 0.944. Other classification models have the same accuracy of 0.833.



Confusion Matrix

- Confusion Matrix of Decision Tree classifier.
- Assuming landing = Positive and Did not land = Negative:
  - TP = 12, TN = 5, FP = 1, FN = 0
- Thus:
  - Accuracy = (12+5)/(12+5+1) = 0.944
  - Precision = (12)/(12+1) = 0.923
  - Recall = (12)/(12+0) = 1



Conclusions

- More rockets with pay load of under 8000kg are launched for all sites.
- The first successful ground landing date was 01-5-2017
- All launch sites are placed very close to the coastal line and above the equator line.
- KSC LC-39A shown to have the highest success rate with 41.7% rate.
- Decision Tree classifier had the highest accuracy over other models with accuracy at about 0.944.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!