IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Keerthi Vijay
27-12-2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- The object of this project is to apply a complete data-driven approach to predict the favorable outcome to winning the space race with SpaceX landing data
- Data has been collected and cleaned by data wrangling and performed exploratory analysis using visualization and SQL. Next, performed interactive visual analytics using Folium and Plotly Dash. Finally predictive analysis using classification model has been performed
- The results are summarized with exploratory data analysis, interactive demo and predictive analysis.

Introduction

- Project background and context
  - SpaceX 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 SpaceX can reuse the first stage. Therefore, this project is done to determine if the first stage will land, then can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - Will SpaceX land successfully in first stage?
  - What are the parameters that influence the landing outcome
  - How to use the insight for the SpaceX rocket launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using the SpaceX API, rocket launch data has been collected. Response content using the GET request has been decoded as Json.
- Perform data wrangling
  - Replaced the missing values with mean in the PayloadMass and calculated the number of launches on each site, number and occurrence of each orbit and mission outcome per orbit type.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split the dataset into training and testing data with test size 0.2. Created GridSeach for different classification models and calculated the accuracy.

Data Collection

- Describe how data sets were collected.
  - Request and parse the SpaceX launch data using the GET request and created a dataframe. Some of the data still contained id. Next used APIs to get the real value for those ids. Finally, a dataframe is created with required data.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- SpaceX REST calls using key phrases and flowcharts

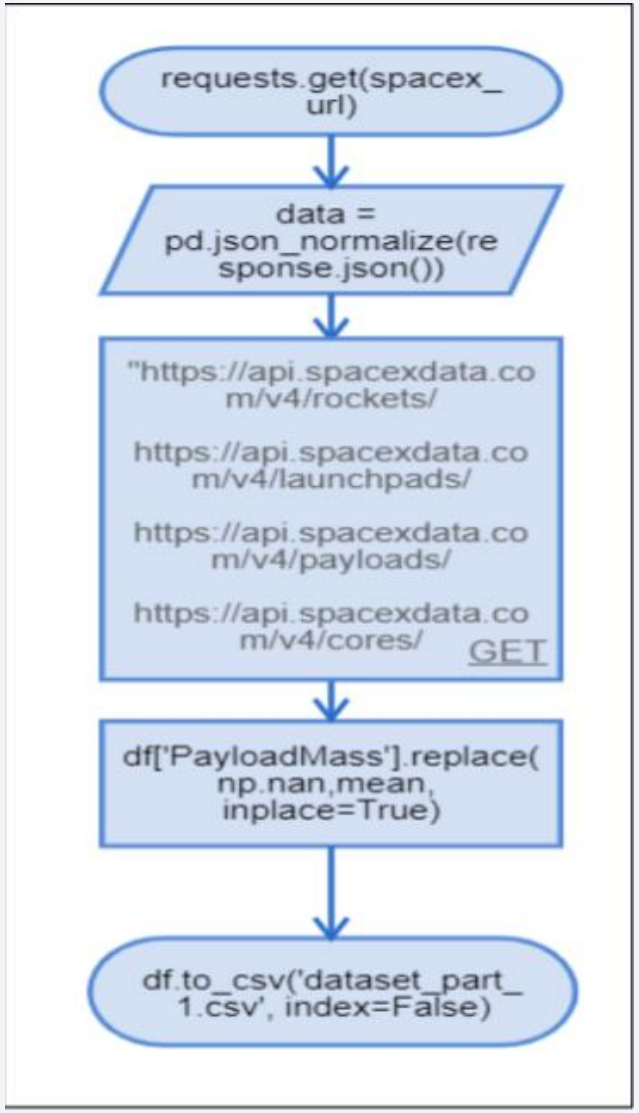
- GitHub URL
  https://github.com/kvi24/IBM-Cousera/blob/master/Final%20Course%20-%20Week%201.ipynb

Request and parse the SpaceX launch data using the GET request

Decode the response content as a JSON and turn it into Pandas dataframe using json_normalize

Use the API to get the information about the launches using the IDs given for each launch.

The mean and the replace() function to replace np.nan values in the data with the calculated mean



Data Collection - Scraping

- Web scraping process

- GitHub URL
  https://github.com/kvi24/IBM-Cousera/blob/master/Data%20Collection%20with%20web%20scrapping.ipynb

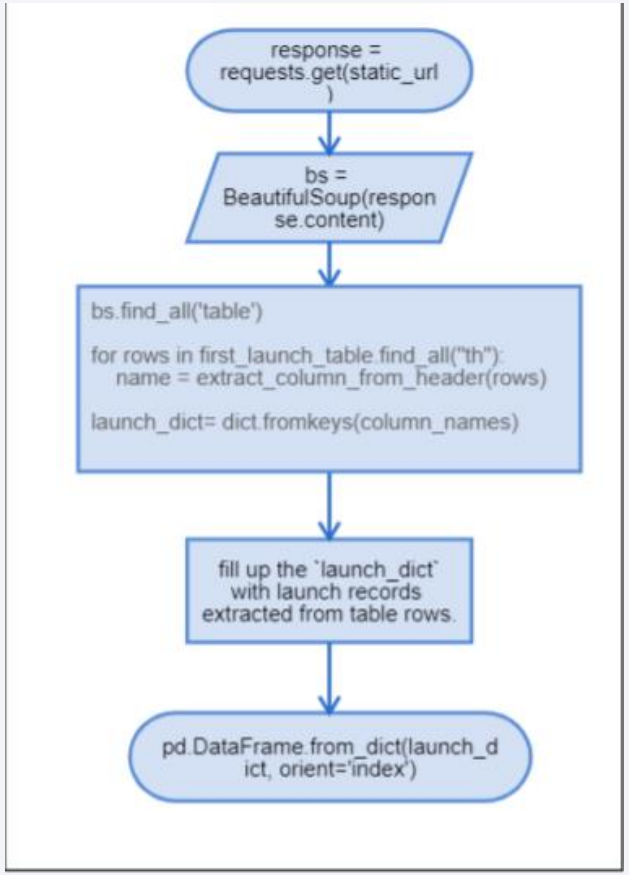
Extract Falcon 9 launch records HTML table from wiki

Parse the table and convert it into Pandas dataframe

Request the Falcon 9Launch Wiki page from URL

Extract all variable names from the HTML table header

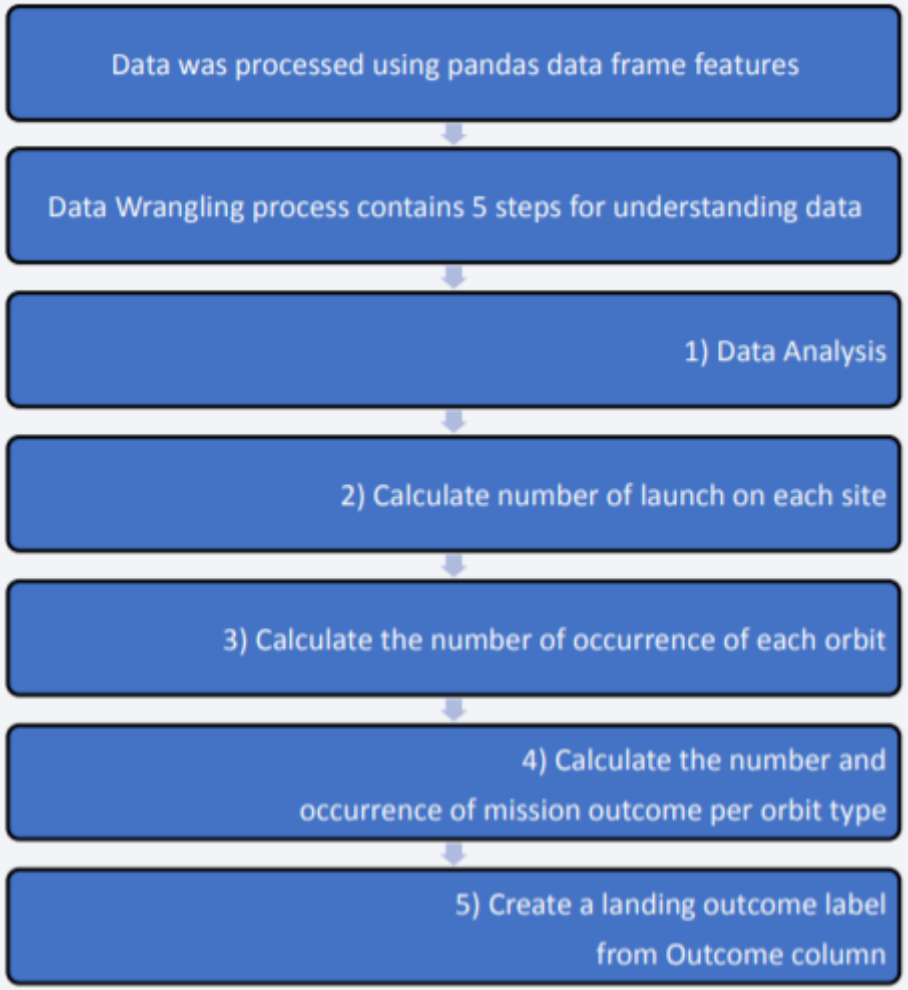
Create a dataframe by parsing the launch HTML tables

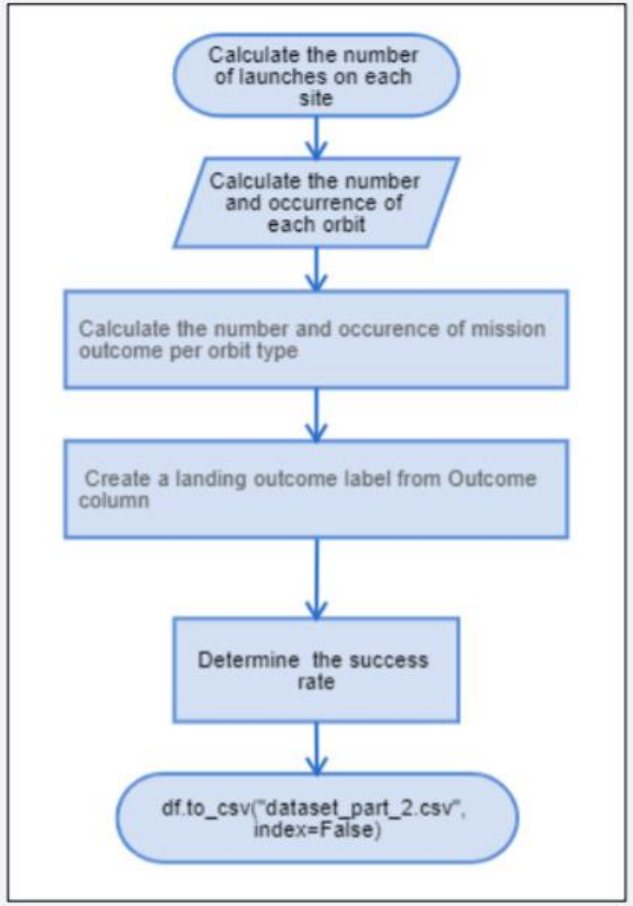


Data Wrangling

- Data wrangling process

- GitHub URL
  - https://github.com/kvi24/IBM-Cousera/blob/master/Final%20Course%20-%20Week%201.ipynb
  - https://github.com/kvi24/IBM-Cousera/blob/master/EDA.ipynb





EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Scatter plot helps to visualize the data – show the data pattern and identify the correlation between the variables
  - Bar chart makes it easy to compare sets of data between different groups
  - Line graph shows the data variables and trends very clearly to make predictions
- GitHub URL
  - https://github.com/kvi24/IBM-Cousera/blob/master/EDA%20Week%202%20-%20Part2.ipynb

EDA with SQL

- SQL queries performed

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass. Use a subquery
9. List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010 06 04 and 2017 0 3 20, in descending order

- GitHub URL
  - https://github.com/kvi24/IBM-Cousera/blob/master/EDA%20Week%202%20-%20Part2.ipynb

Build an Interactive Map with Folium

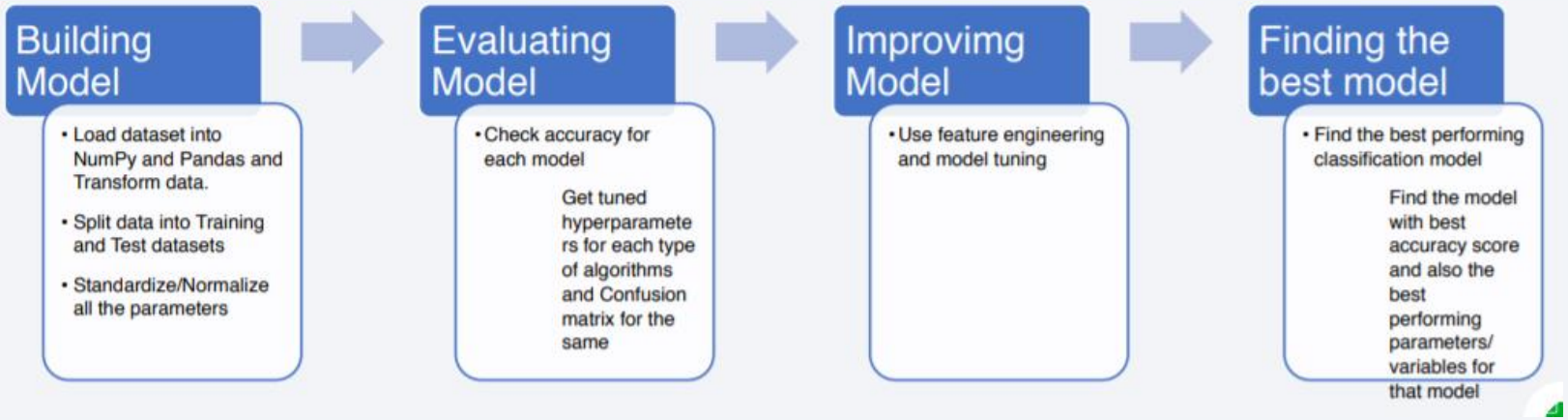
- Folium map summary
  - Marked all launch sites visually on a map with coordinates
  - Marked the success/failure launches for each site
  - Calculated the distance between various landmarks
- Explain why you added those objects
  - Helps to answer the questions easily:
    - Are launch sites near railways, highways, coastlines and cities?
- GitHub URL
  - https://github.com/kvi24/IBM-Cousera/blob/master/Week%203%20-%20Interactive%20Visual%20Analytics%20with%20Folium%20.ipynb

Build a Dashboard with Plotly Dash

- Dashboard Summary
  - Plotly is used to interact with the data visualizations
- Explain why you added those plots and interactions
  - Pie Chart to show number of launches from each launch site as well as number of success/failure launches from each sites
  - Scatter Plot to show the relationship between the success of a launch and Payload for different versions of boosters
- GitHub URL
  - https://github.com/kvi24/IBM-Cousera/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Classification model Summary



- GitHub URL
  - https://github.com/kvi24/IBM-Cousera/blob/master/Machine%20Learning%20Prediction%20Lab.ipynb

Results

Exploratory data analysis results

landing__outcome	COUNT
Success (drone ship)	2
Success (ground pad)	2
Failure (drone ship)	1
No attempt	1

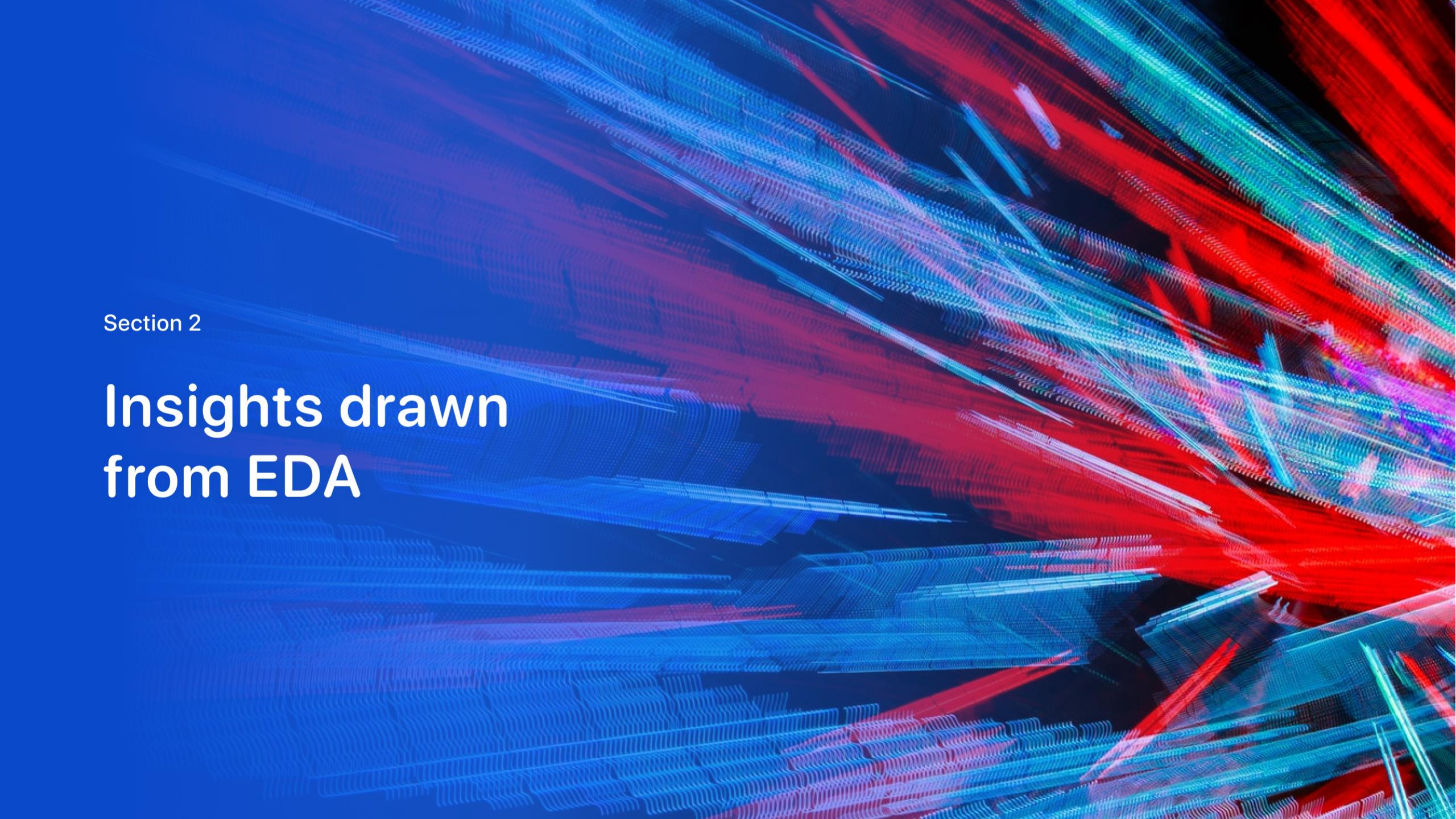
Interactive analytics demo in screenshots

Predictive analysis results

Model	Accuracy	Accuracy using Score
Logistic Regression	84.64%	83.34%
SVM	84.82%	83.34%
Decision Tree	87.68%	88.89%
KNN	84.82%	83.34%

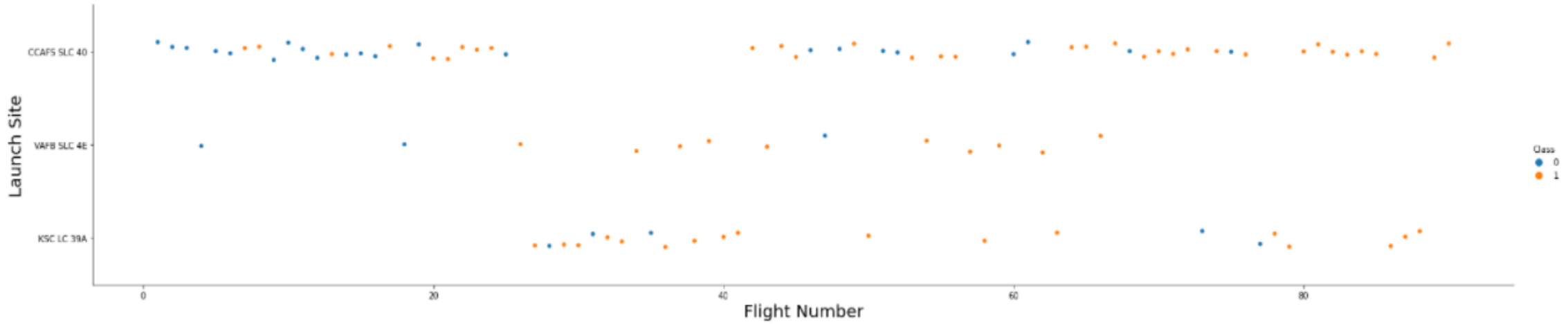
Section 2

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()
```

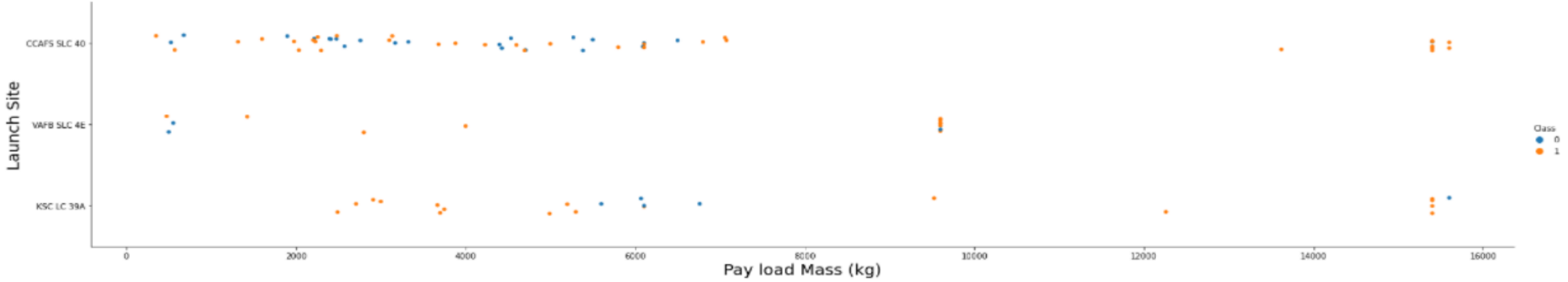


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Payload vs. Launch Site

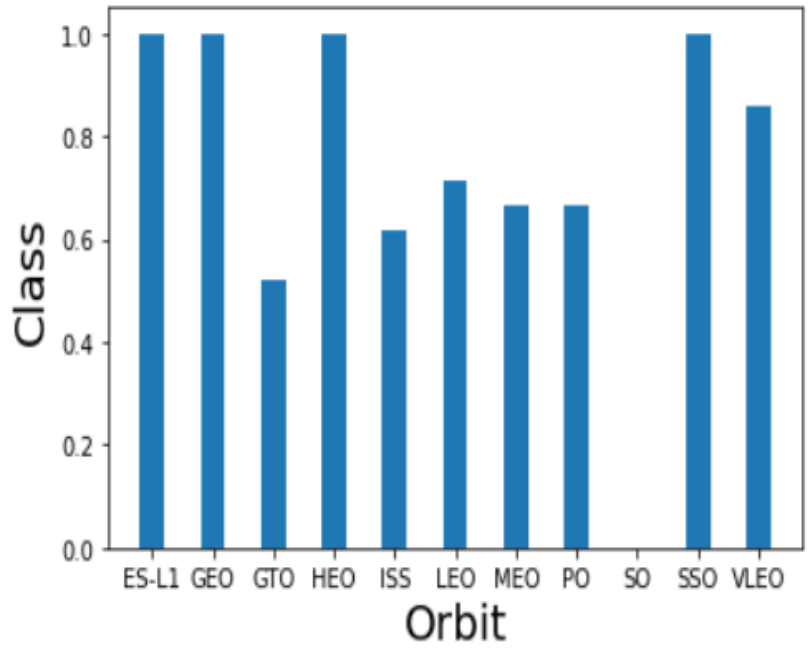
```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class valu
e
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()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



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).

Success Rate vs. Orbit Type

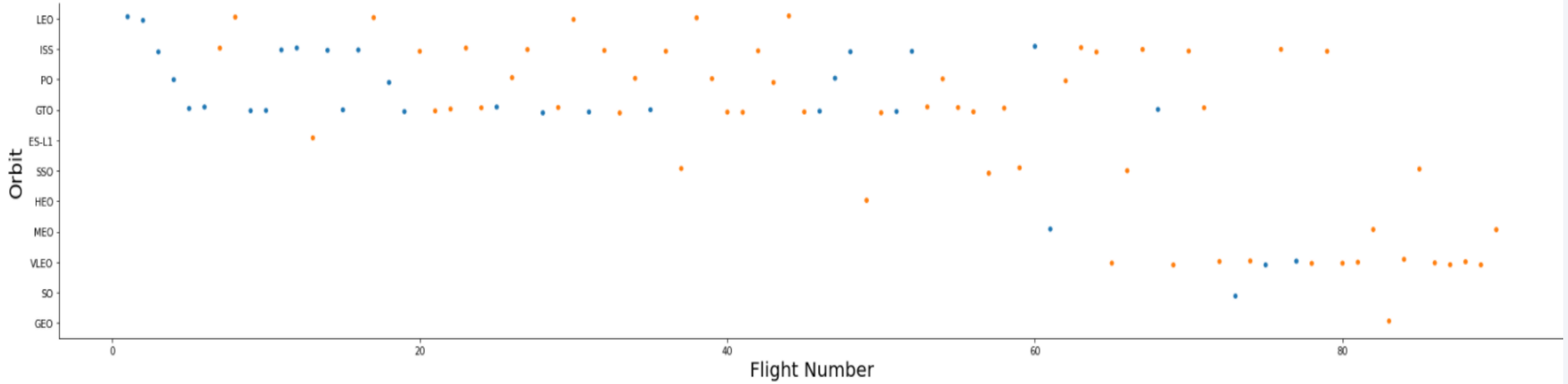
```
# HINT use groupby method on Orbit column and get the mean of Class column
#df[['Orbit','Class']].groupby('Orbit').mean().sort_values('Class')
plt.bar(np.unique(df["Orbit"]),df.groupby(['Orbit']).mean()['Class'],width = 0.4)
plt.xlabel("Orbit", fontsize = 20)
plt.ylabel("Class", fontsize = 20)
plt.show()
```



Analyze the ploted bar chart try to find which orbits have high sucess rate. Orbits SSO, ES-L1, GEO, HEO and VLEO are the ones with 80% or more success rate

Flight Number vs. Orbit Type

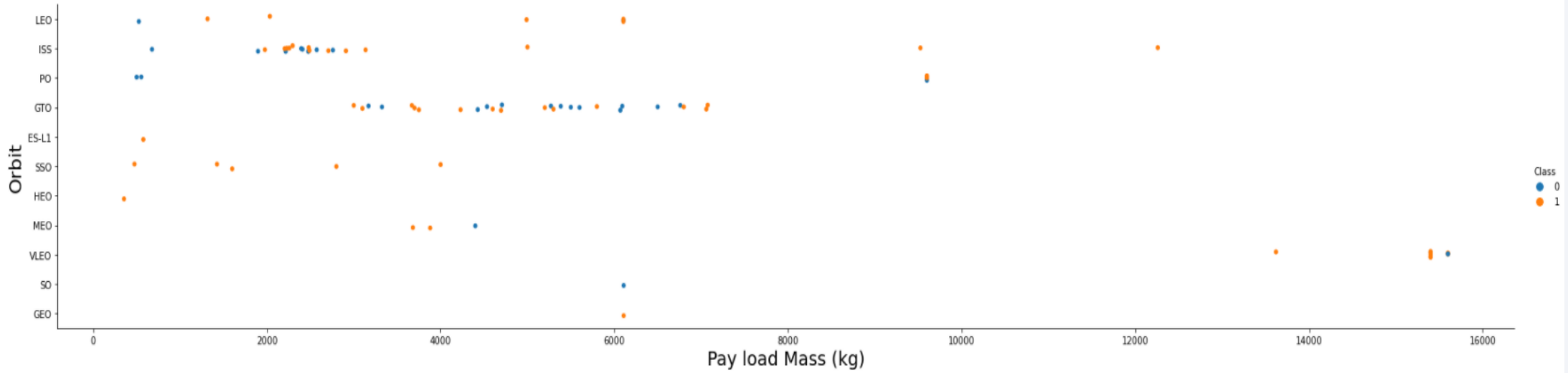
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



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.

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="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



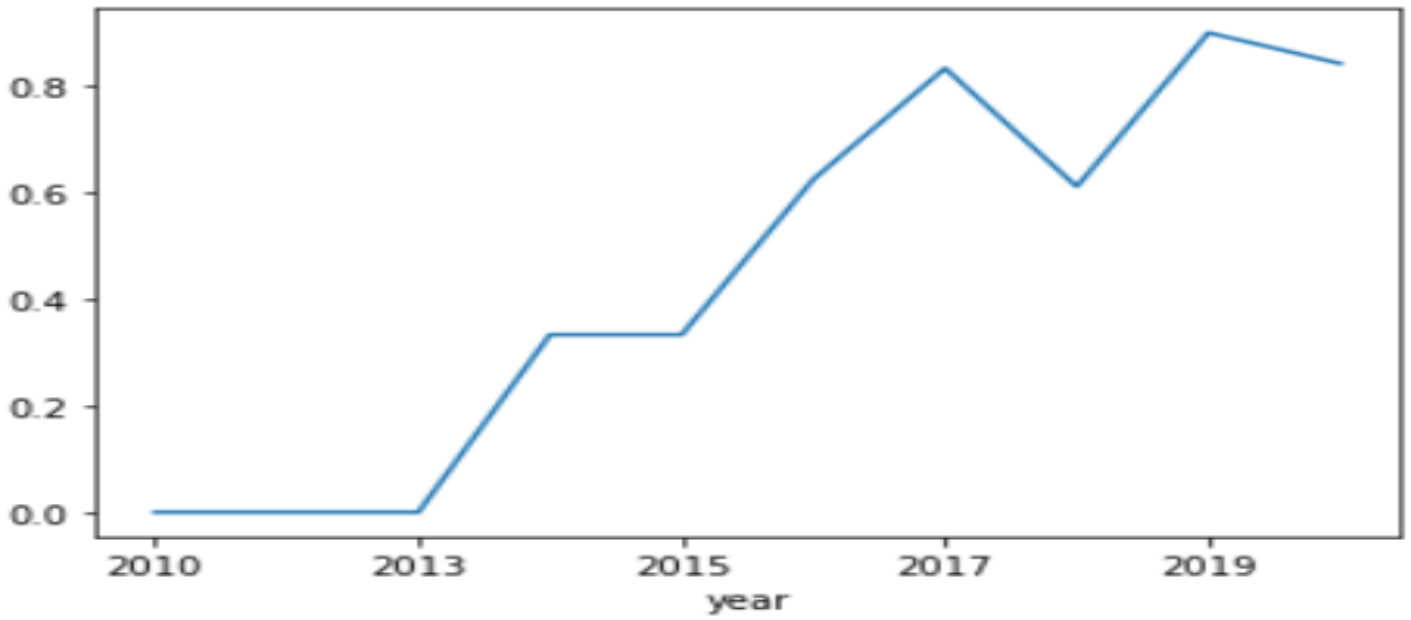
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

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

```
# Plot a line chart with x axis to be the extracted year and y axis
df1 = pd.DataFrame(Extract_year(df['Date']),columns = ['year'])
df1['Class'] = df['Class']
df2 =df1.groupby('year')['Class'].mean()
df2.plot(kind='line', y='Success Launch')
```

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



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

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'

```
SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%';
```

DATE	Time (UTC)	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD	PAYLOAD_MASS__KG_	ORBIT	CUS
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	Space
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA

Total Payload Mass

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

45596 Kg

Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9
v1.1%';
```

2534 Kg

First Successful Ground Landing Date

```
SELECT MIN(DATE) FROM SPACEXTBL WHERE ("Landing _Outcome") = 'Success (ground pad)';
```

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT BOOSTER_VERSION, "Landing _Outcome", PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE ("Landing _Outcome") = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEEN
4000 AND 6000);
```

BOOSTER_VERSION	Landing _Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

```
SELECT COUNT("Landing _Outcome") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE
'Success%';
```

61

```
SELECT COUNT("Landing _Outcome") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE
'Failure%';
```

10

Boosters Carried Maximum Payload

```
SELECT DISTINCT(BOOSTER_VERSION), PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE
PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

2015 Launch Records

```
SELECT "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE, YEAR(DATE) AS Year FROM
SPACEXTBL WHERE YEAR(DATE) = '2015' AND "Landing _Outcome" = 'Failure (drone ship)'
```

Landing _Outcome	BOOSTER_VERSION	LAUNCH_SITE	YEAR
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015

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

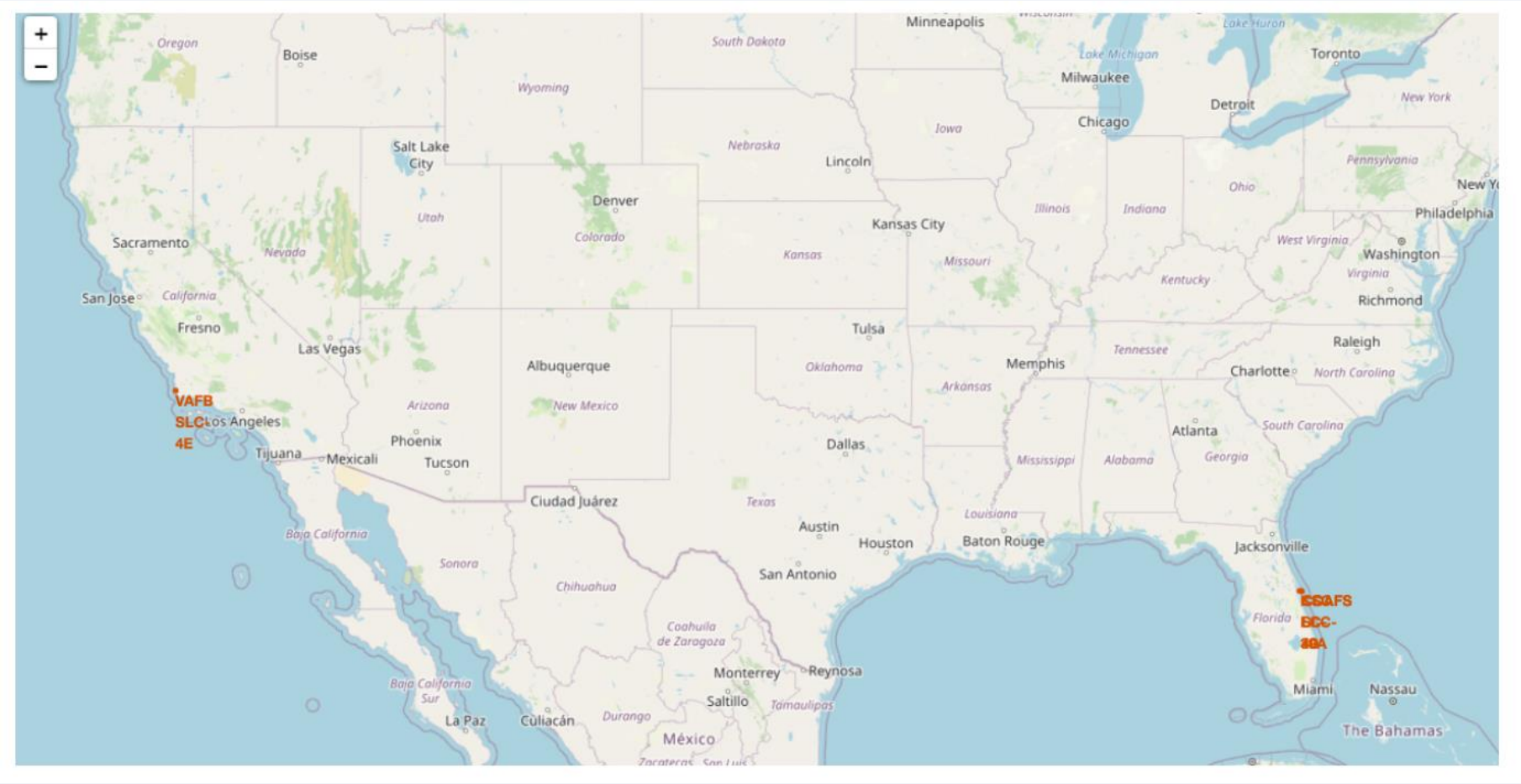
```
SELECT "Landing _Outcome", COUNT("Landing _Outcome") AS Count FROM SPACEXTBL WHERE
DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing _Outcome" ORDER BY Count
DESC;
```

Landing _Outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

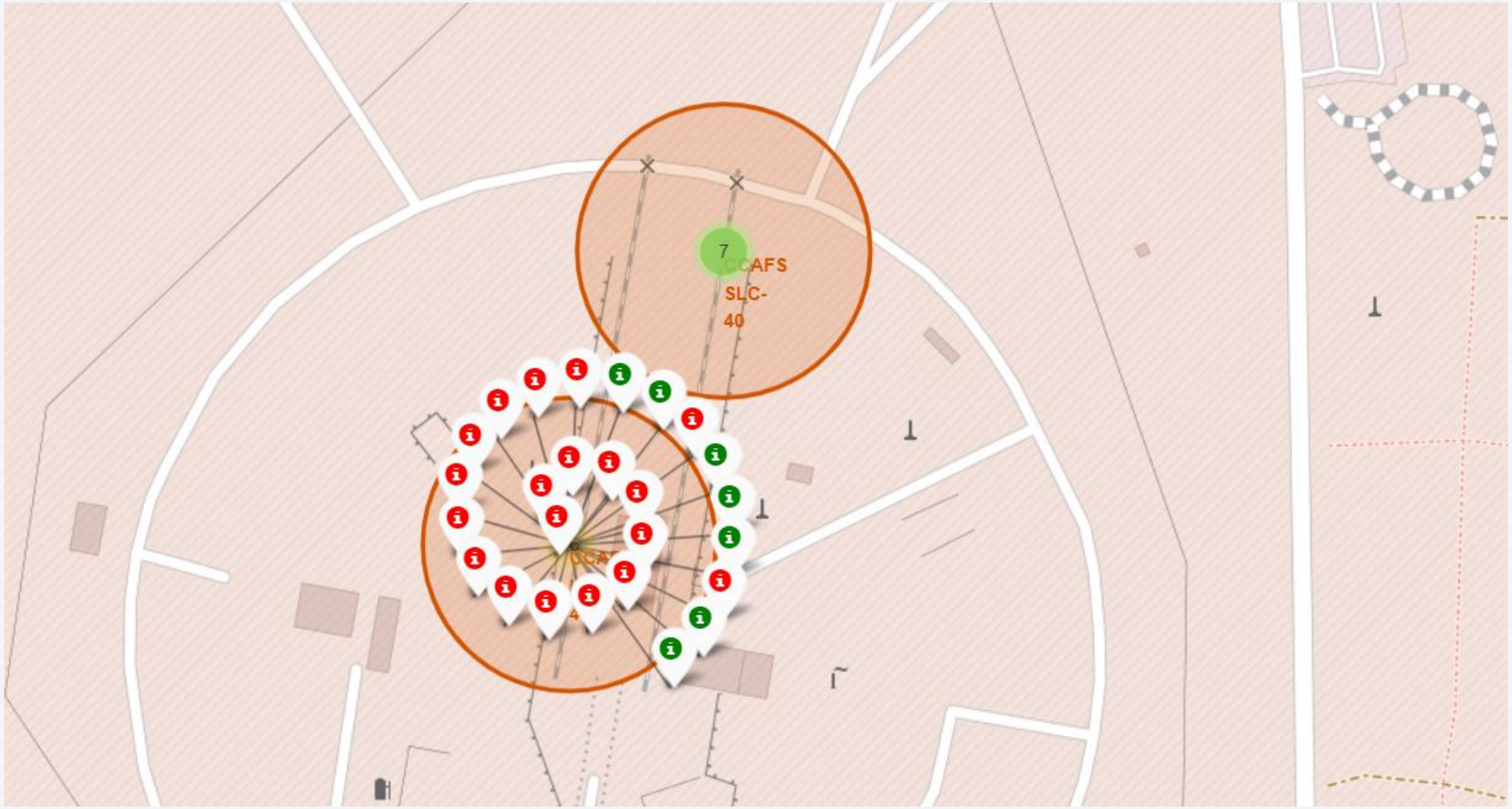
Launch Sites Proximities Analysis

Launch Sites on Map



SpaceX launch sites are in Florida – USA east coast and California – USA west coast

Launch Outcomes

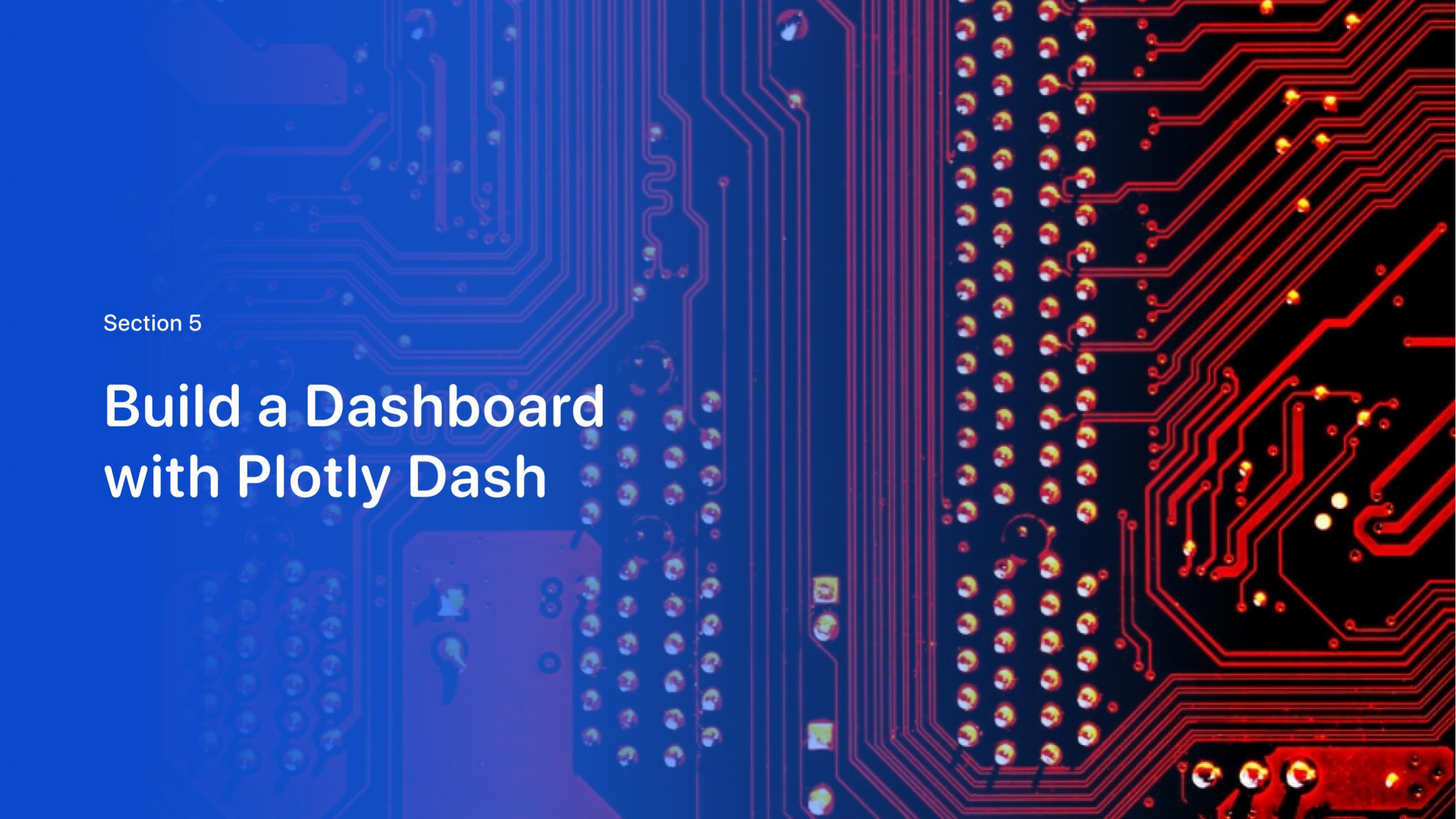


Green colored are successful launches and red colored are the failed launches

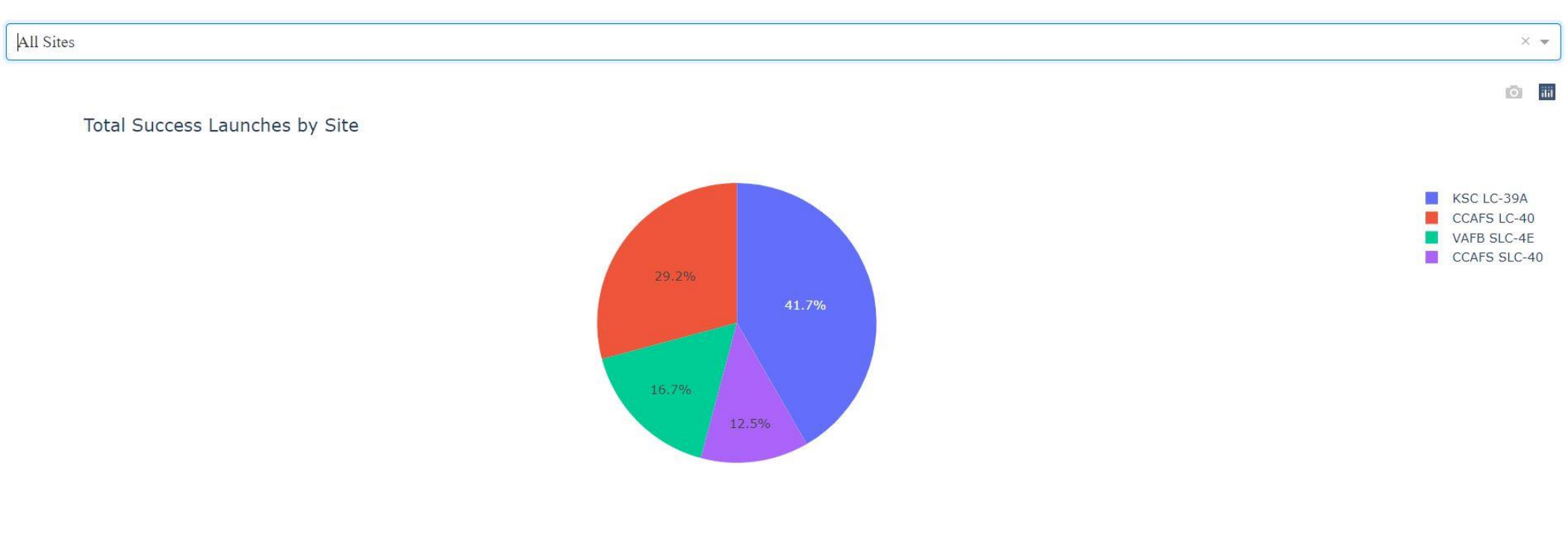
Launch site proximities

Section 5

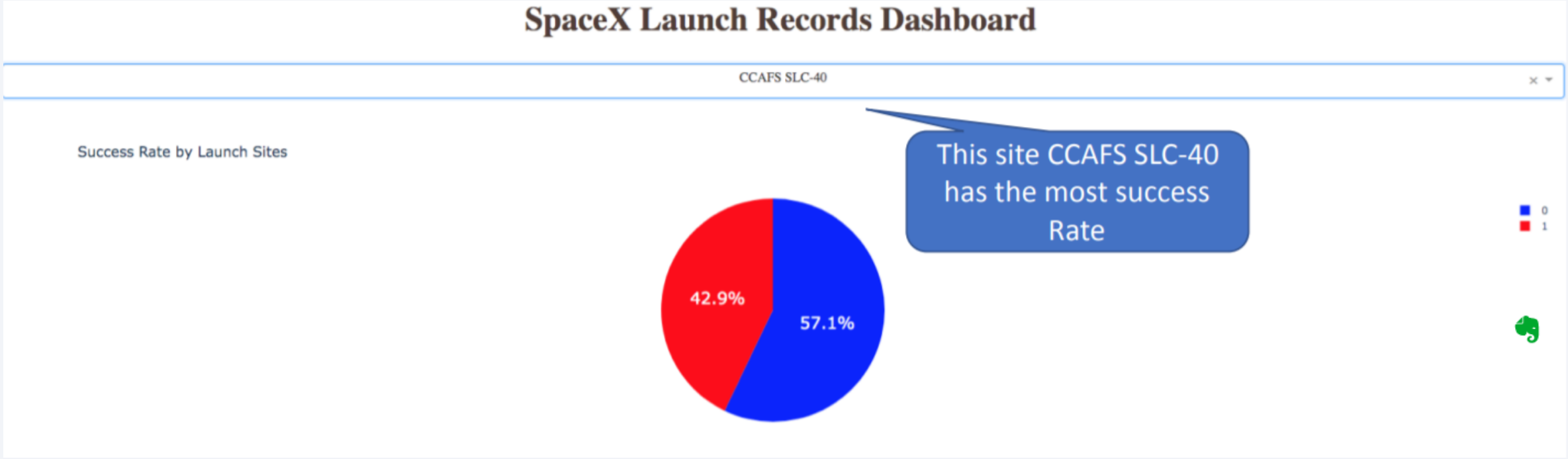
Build a Dashboard with Plotly Dash



Success rate by Launch sites



Most successful launch site



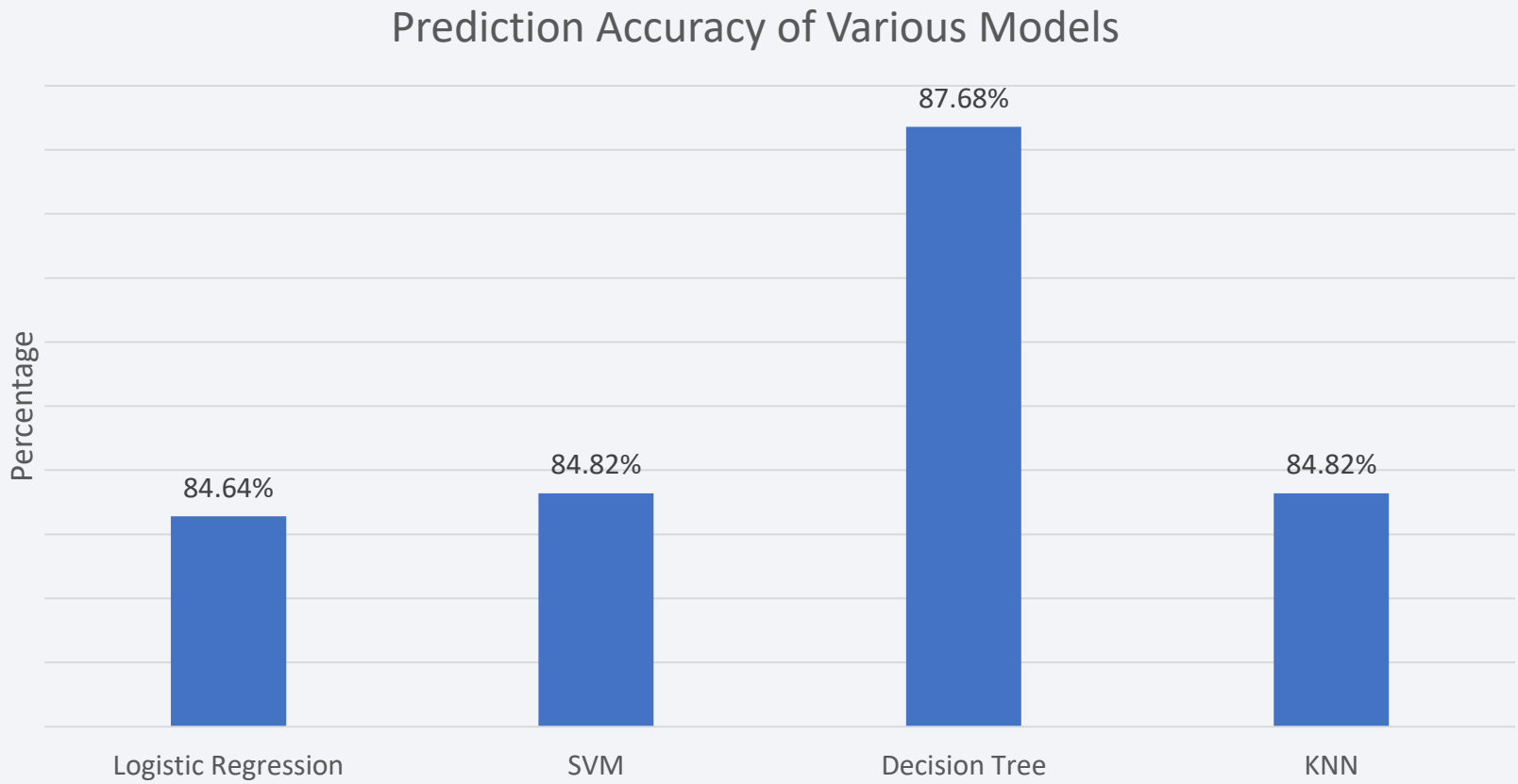
Payload vs. Launch Outcome Scatter Plot

Lower Payload launches are more successful

Section 6

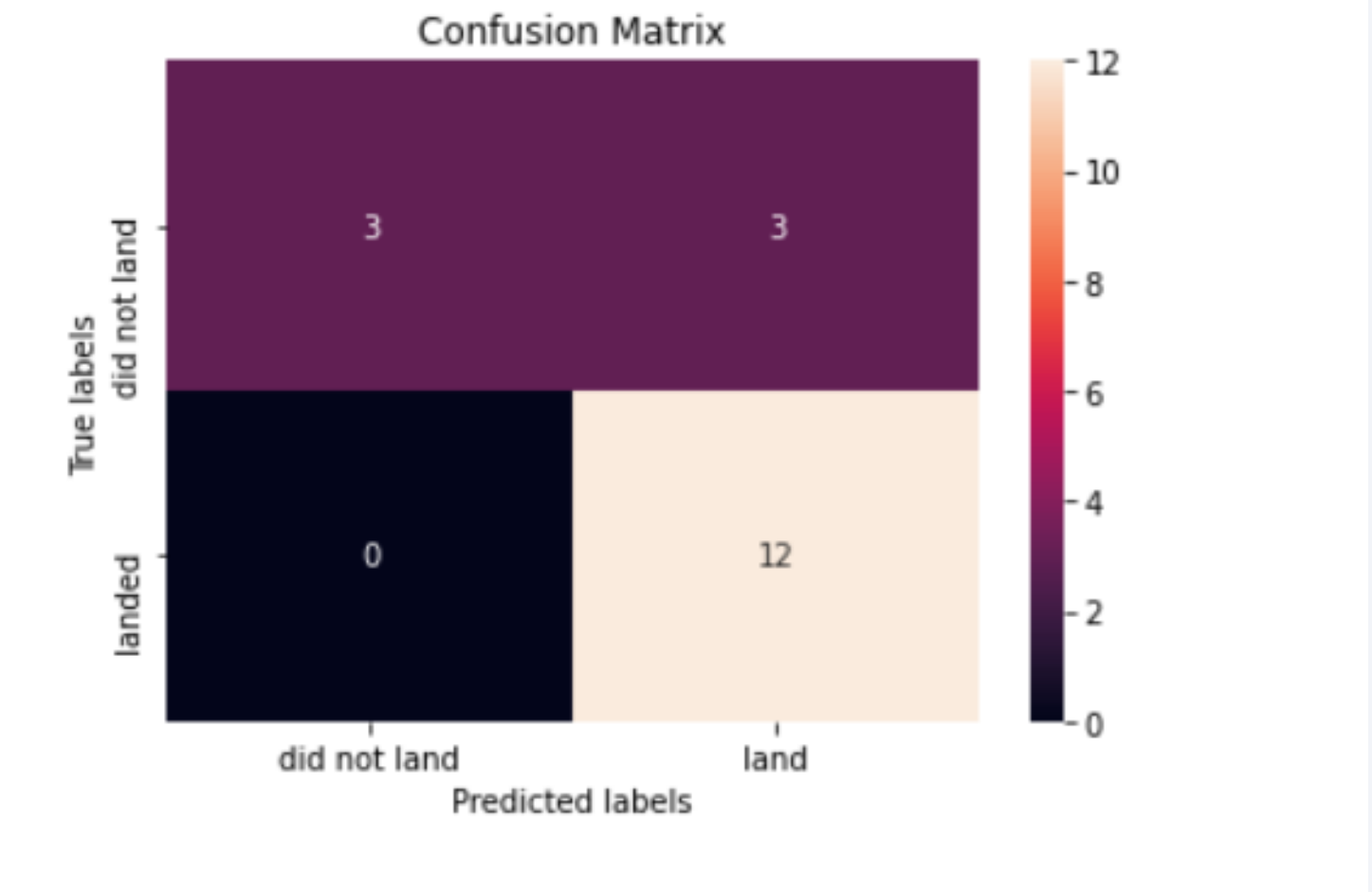
Predictive Analysis (Classification)

Classification Accuracy



Decision Tree model got high accuracy

Confusion Matrix



Decision tree confusion matrix

Conclusions

- Decision Tree Model is the best classifier model with high classification accuracy
- The lower payload launches have higher success rate than heavier payloads
- Site KSC LC-39A has the most successful launches from all sites
- F9 Booster versions v1.0, v1.1, FT, B4, B5 have the highest launch success rates
- The SpaceX launches have been continuously getting better from year 2013 to 2020 based on data so they have the best chances for perfecting their launches in future

Thank you!