Winning Space Race with Data Science

NIGAY WARIS

November 19th 2022

Outline

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

Executive Summary

- For this project, we used a variety of methodologies like :
  - Websrapping with BeautifulSoup ;
  - SQL Queries to extract key findings ;
  - Folium to analyze data in a more interactive way;
  - Machine Learning algorithms to predict if a land will be successful or not.
- Summary of all results :
  - The CCAFS SLC 40 launch site had the most launch (55) ;
  - But it was the less successful launch site (60% vs 77% for KSC LC-39A)

Introduction

- SpaceX proved to all of us that it was possible to build rocket for a much SMALLER amount of money than the NASA was doing. Moreover, they showed the world that it was possible to improve every technological aspect with less budget.
- With this project, we want to answers some problems :
  - Will a competitive company be able to have a better successful launch rate ?
  - Where the launch site should be ?
  - Is Payload a dominant factor ?

Section 1

Methodology

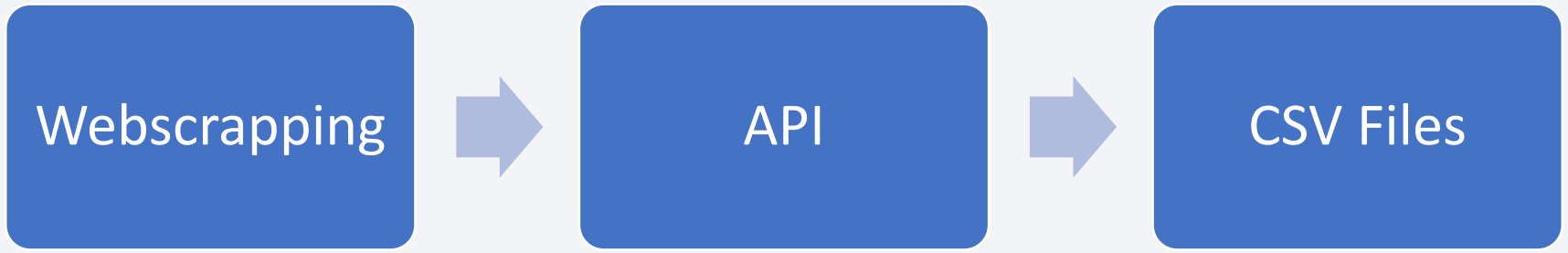
Methodology

Executive Summary

- Data collection methodology:
  - Web scraping using BeautifulSoup
- Perform data wrangling
  - It was processed by dropping non usefull columns,
  - By indexing data, and replacing NuLL values with meaningfull data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - With SVM, Decision Tree, Clusters etc.

Data Collection

- The data was collected by using web scrapping and loading CSV files.



From the `rocket` column we would like to learn the booster name.

```
# Takes the dataset and uses the rocket column to call the API and append the data to the list
def getBoosterVersion(data):
    for x in data['rocket']:
       if x:
        response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
        BoosterVersion.append(response['name'])
```

Data Collection – SpaceX API

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

- On the right, you can see a sample of the outcome from the API Request.

- GitHub URL: https://github.com/NigayWaris/IBM-Final-Project/blob/main/test.ipynb


```
b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"links":{"pat
tps://images2.imgbox.com/5b/02/QcxHUb5V_o.png"},"reddit":{"campaign":null,"launch":null,"media":nul
bcast":"https://www.youtube.com/watch?v=0a_00nJ_Y88","youtube_id":"0a_00nJ_Y88","article":"https://
ipedia":"https://en.wikipedia.org/wiki/DemoSat"},"static_fire_date_utc":"2006-03-17T00:00:00.000Z",
95eda69955f709d1eb","success":false,"failures":[{"time":33,"altitude":null,"reason":"merlin engine
rew":[],"ships":[],"capsules":[],"payloads":["5eb0e4b5b6c3bb0006eeb1e1"],"launchpad":"5e9e4502f509
2:30:00.000Z","date_unix":1143239400,"date_local":"2006-03-25T10:30:00+12:00","date_precision":"hou
1,"gridfins":false,"legs":false,"reused":false,"landing_attempt":false,"landing_success":null,"land
rary_id":null,"id":"5eb87cd9ffd86e000604b32a"},{"fairings":{"reused":false,"recovery_attempt":false
imgbox.com/f9/4a/ZboXReNb_o.png","large":"https://images2.imgbox.com/80/a2/bkWotCIS_o.png"},"reddit
{"small":[],"original":[]},"presskit":null,"webcast":"https://www.youtube.com/watch?v=Lk4zQ2wP-Nc",
lcon-1-rocket-fails-reach-orbit.html","wikipedia":"https://en.wikipedia.org/wiki/DemoSat"},"static_
ocket":"5e9d0d95eda69955f709d1eb","success":false,"failures":[{"time":301,"altitude":289,"reason":"
ccessful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine
stage","crew":[],"ships":[],"capsules":[],"payloads":["5eb0e4b6b6c3bb0006eeb1e2"],"launchpad":"5e9e
03-21T01:10:00.000Z","date_unix":1174439400,"date_local":"2007-03-21T13:10:00+12:00","date_precisio
ight":1,"gridfins":false,"legs":false,"reused":false,"landing_attempt":false,"landing_success":null
ch_library_id":null,"id":"5eb87cdaffd86e000604b32b"},{"fairings":{"reused":false,"recovery_attempt"
ages2.imgbox.com/6c/cb/na1tzhHs_o.png","large":"https://images2.imgbox.com/4a/80/k1oAkY0k_o.png"},"
kr":{"small":[],"original":[]},"presskit":null,"webcast":"https://www.youtube.com/watch?v=v0w9p3U8
02/11/falcon-1-flight-3-mission-summary","wikipedia":"https://en.wikipedia.org/wiki/Trailblazer_(sa
false,"window":0,"rocket":"5e9d0d95eda69955f709d1eb","success":false,"failures":[{"time":140,"altit
and stage 2"}],"details":"Residual stage 1 thrust led to collision between stage 1 and stage 2","cr
```


Now let's start requesting rocket launch data from SpaceX API with the following URL:

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

Data Collection - Scraping

- In this part, we used BeautifulSoup to webscrape the Wikipedia page.
- The objective was to extract Falcon 9 launch records.
- GitHub URL: https://github.com/NigayWaris/IBM-Final-Project/blob/main/2%20-%20Webscraping.ipynb

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

Data Wrangling

- In this part, we performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- We calculated the number of launches on each site. This is where we discovered that CCAFS SLC 40 was the site with the most launches (55).
- GitHub URL: https://github.com/NigayWaris/IBM-Final-Project/blob/main/EDA%20-%20Final%20Project.ipynb

EDA with Data Visualization

- In this part, we predicted if the Falcon 9 first stage will land successfully. 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 due to the fact that SpaceX can reuse the first stage.
- We used “catplot” and “bar chart” to visualize the data
- GitHub URL: https://github.com/NigayWaris/IBM-Final-Project/blob/main/EDA%20Matplotlib.ipynb

EDA with SQL

- Summary of the SQL queries performed :
  - SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
  - SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
  - SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.0%';
  - Other form of “min”, “avg” and “distinct” queries.

- GitHub URL: https://github.com/NigayWaris/IBM-Final-Project/blob/main/EDA%20SQL%20-%20IBM%20Final%20Project.ipynb

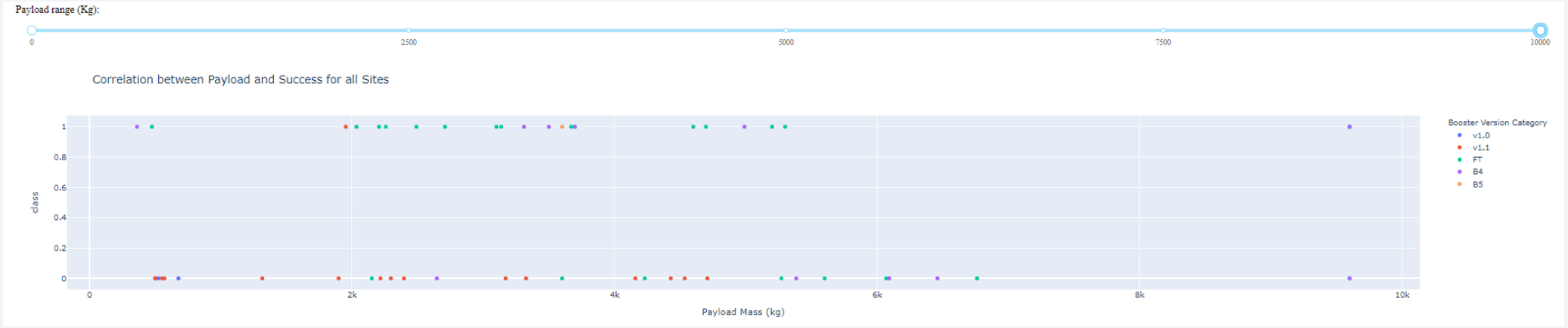
Build an Interactive Map with Folium

- In this part, we used the longitude and latitude to show a map were the launch site are. We added objects like circles, markers and so on, using this kind of code :
  - circle = folium.Circle(nasa_coordinate, radius=1000, color='#d35400', fill=**True**).add_child(folium.Popup('NASA Johnson Space Center'))
- We added those objects in order to have more precise and meaningful insights on the data where rockets were launched.
- GitHub URL: https://github.com/NigayWaris/IBM-Final-Project/blob/main/Week%203%20-%20Interactive%20Visual%20Analytics%20(Folium%20lab).ipynb

Build a Dashboard with Plotly Dash

- We used scatter plot to build an interactive Dashboard, allowing us to see the correlation between Payload and the success rate for all of the launch sites :

- GitHub URL: https://github.com/NigayWaris/IBM-Final-Project/blob/main/PlotlyDash.py

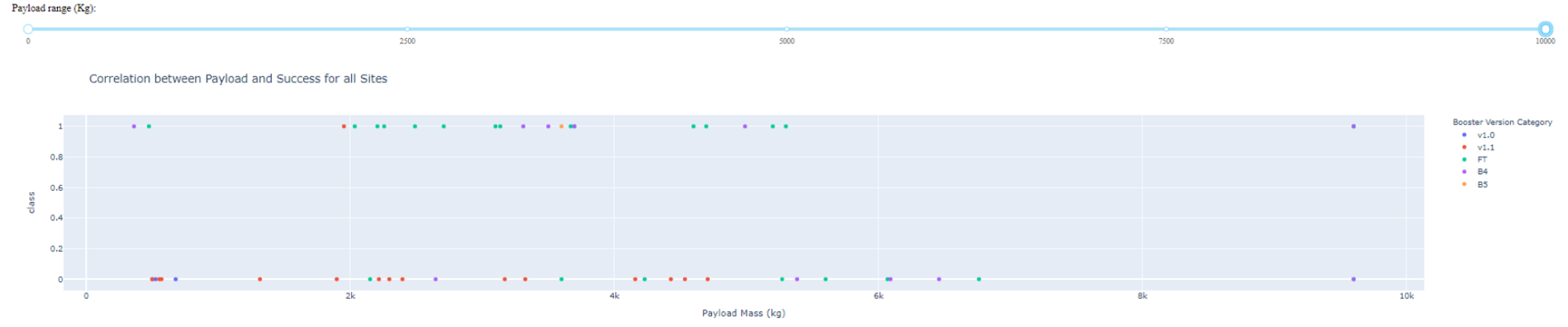


Predictive Analysis (Classification)

- In this part, we imported seaborn, sklearn libraries to be able to build the ML process.
- Here are the steps we conducted :
  - Create a column for the class;
  - Standardized the data;
  - Split into training data set and test data set;
  - Used SVM, Decision Tree etc.
- GitHub URL: https://github.com/NigayWaris/IBM-Final-Project/blob/main/Week%204%20-%20ML%20Prediction.ipynb

Results

- Interactive analytics demo in screenshots



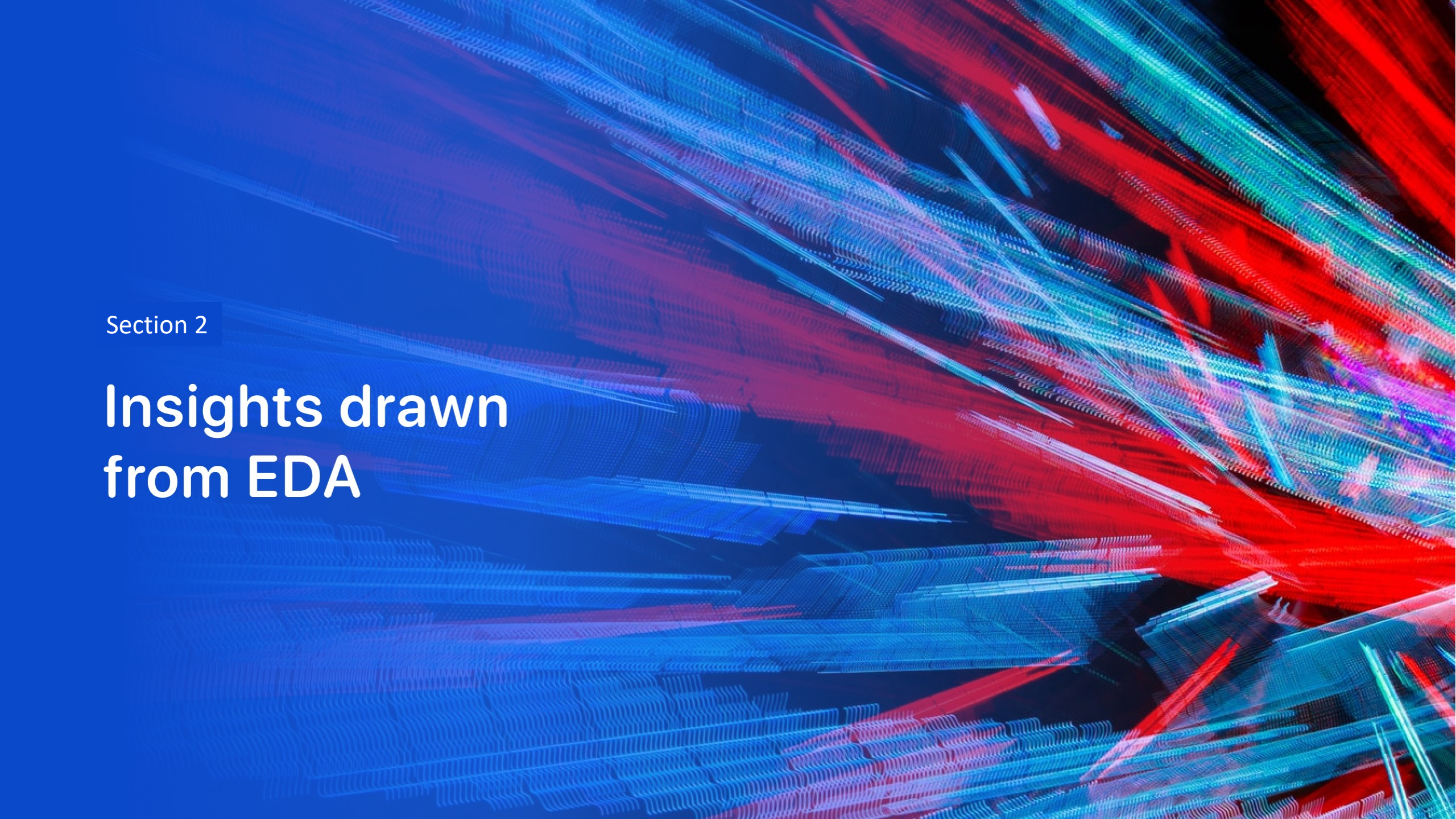
- Predictive analysis results

TASK 12

Find the method performs best:

```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.6666666666666666
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

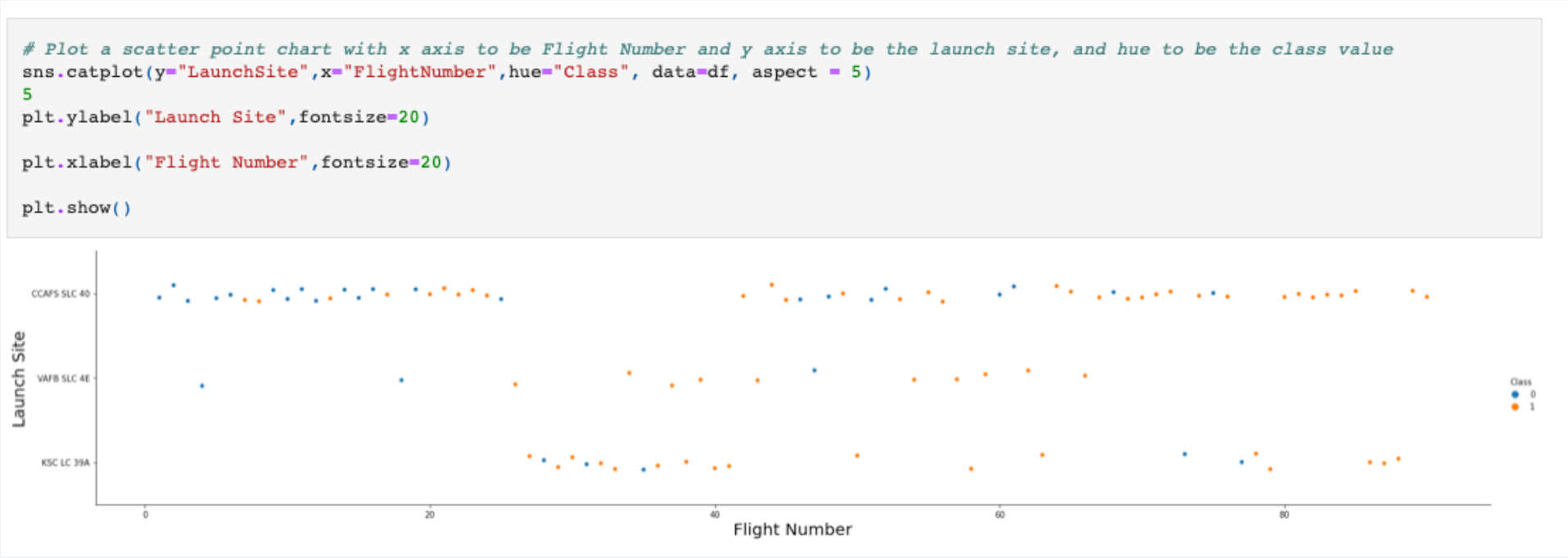

Section 2
Insights drawn from EDA

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)
5
plt.ylabel("Launch Site",fontsize=20)

plt.xlabel("Flight Number",fontsize=20)

plt.show()
```

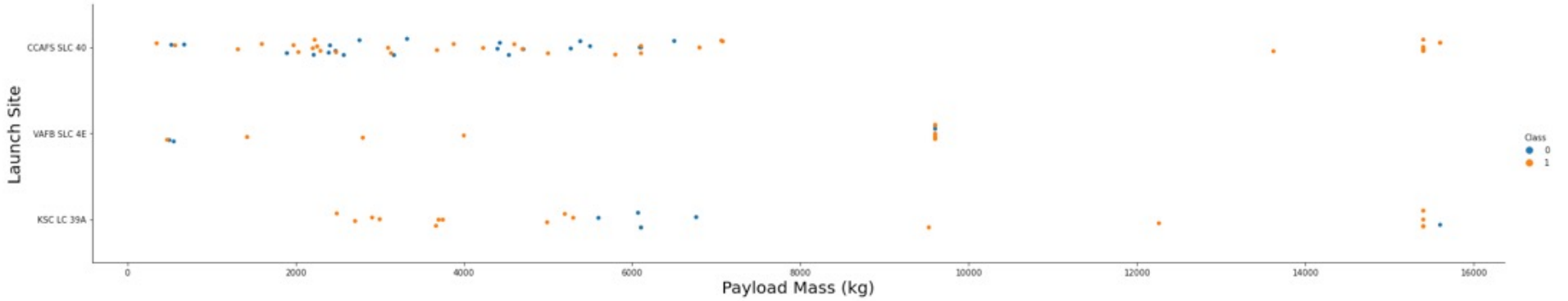


- This scatter plot shows that there is a lot more launch at the CCAFS SLC 40 site.
- It also shows that flights between 25 and 40 were pratically all conducted in KSC LC 39A site.

Payload vs. 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(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

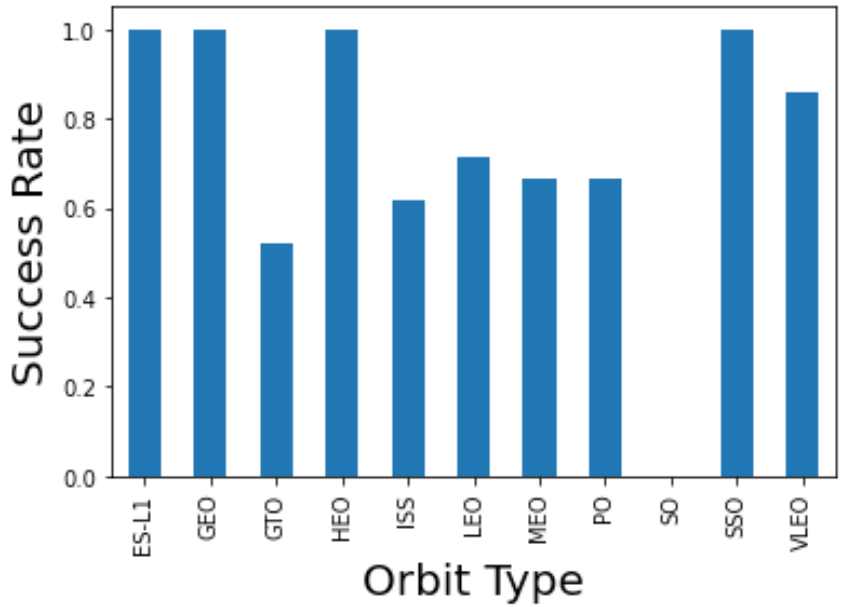


- Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- The least successful Orbit Type is GTO.
- In order to find these results we grouped by the Orbit column.

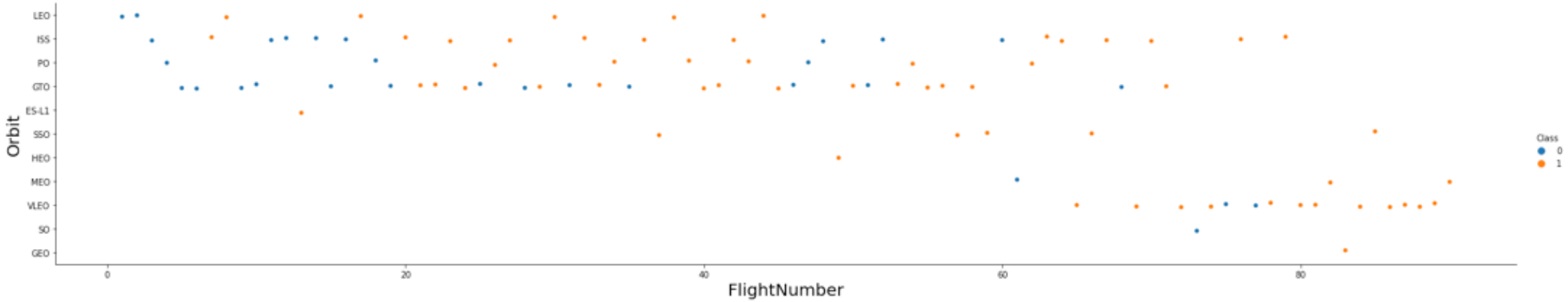
```
# HINT use groupby method on Orbit column and get the mean of Class column
df.groupby("Orbit").mean()['Class'].plot(kind='bar')
plt.xlabel("Orbit Type",fontsize=20)
plt.ylabel("Success Rate",fontsize=20)
plt.show()
```



Flight Number vs. 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.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

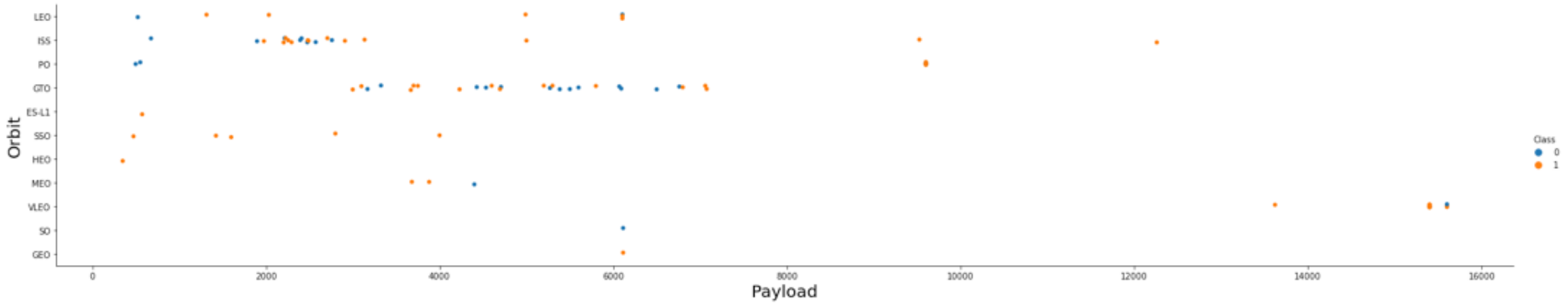


- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

Similarly, we can plot the Payload vs. Orbit scatter point charts to reveal the relationship between Payload and Orbit type

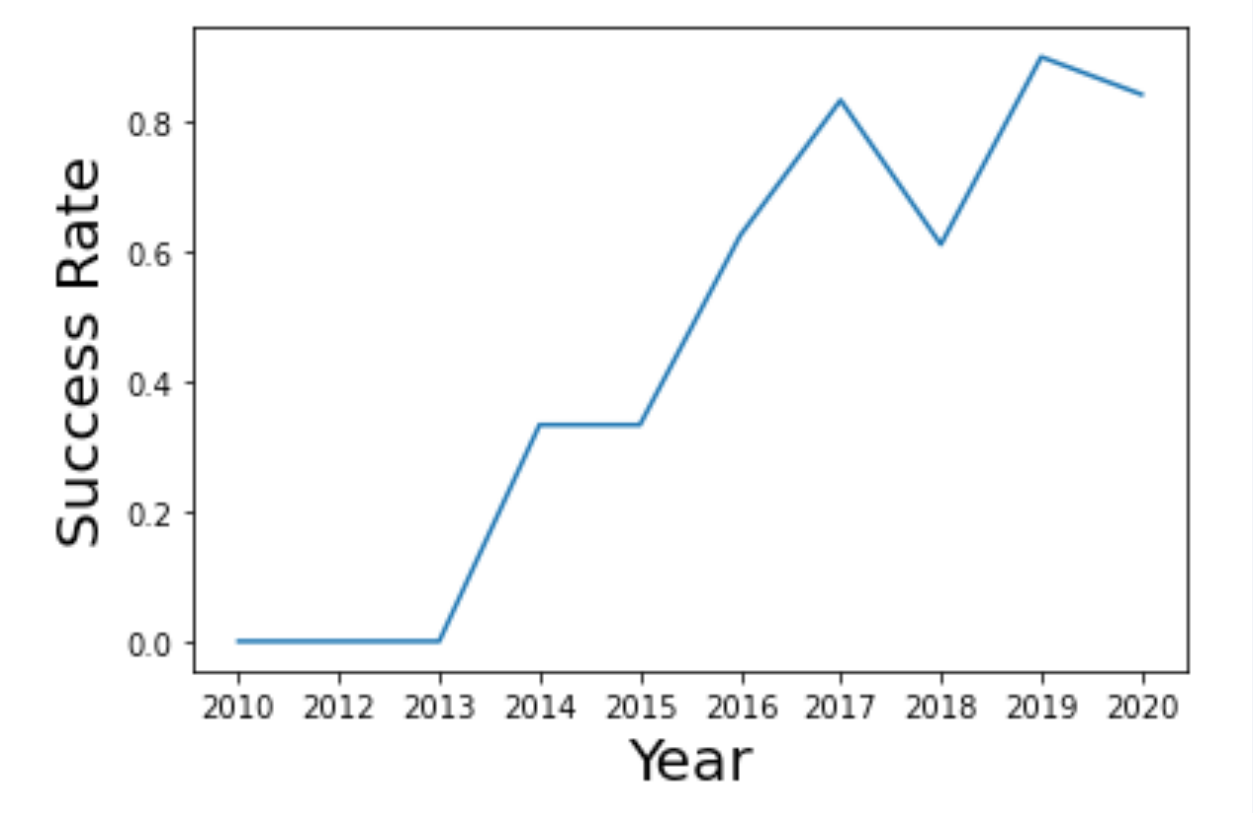
```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



- 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

- You can observe that the sucess rate since 2013 kept increasing till 2020.



All Launch Site Names

- Find the names of the unique launch sites
- **%sql** SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
- The query use the DISTINCT function

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- **%sql** SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
- We used the function like on the column LAUNCH_SITE.

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Simple SUM query on the Payload column.

Task 3

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

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

```
 * ibm_db_sa://cdm26718:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Simple AVG function on Payload with a WHERE and LIKE function to filter.

Task 4

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.0%';
```

```
 * ibm_db_sa://cdm26718:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

1
340

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Simple MIN(Date) function with the filter Success.

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

Hint:Use min function

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing__Outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://cdm26718:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

1
2015-12-22

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
- Simple WHERE query with a AND & < operator.

Task 6

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 4000 < PAYLO
```

```
 * ibm_db_sa://cdm26718:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- We used an ALIAS to rename the column. Simple COUNT() function.

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://cdm26718:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Introduction of a subquery with a MAX() function here.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_
```

```
 * ibm_db_sa://cdm26718:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Same as before.

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE Landing__Outcome = 'Failure
```

```
 * ibm_db_sa://cdm26718:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Task 10

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 LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

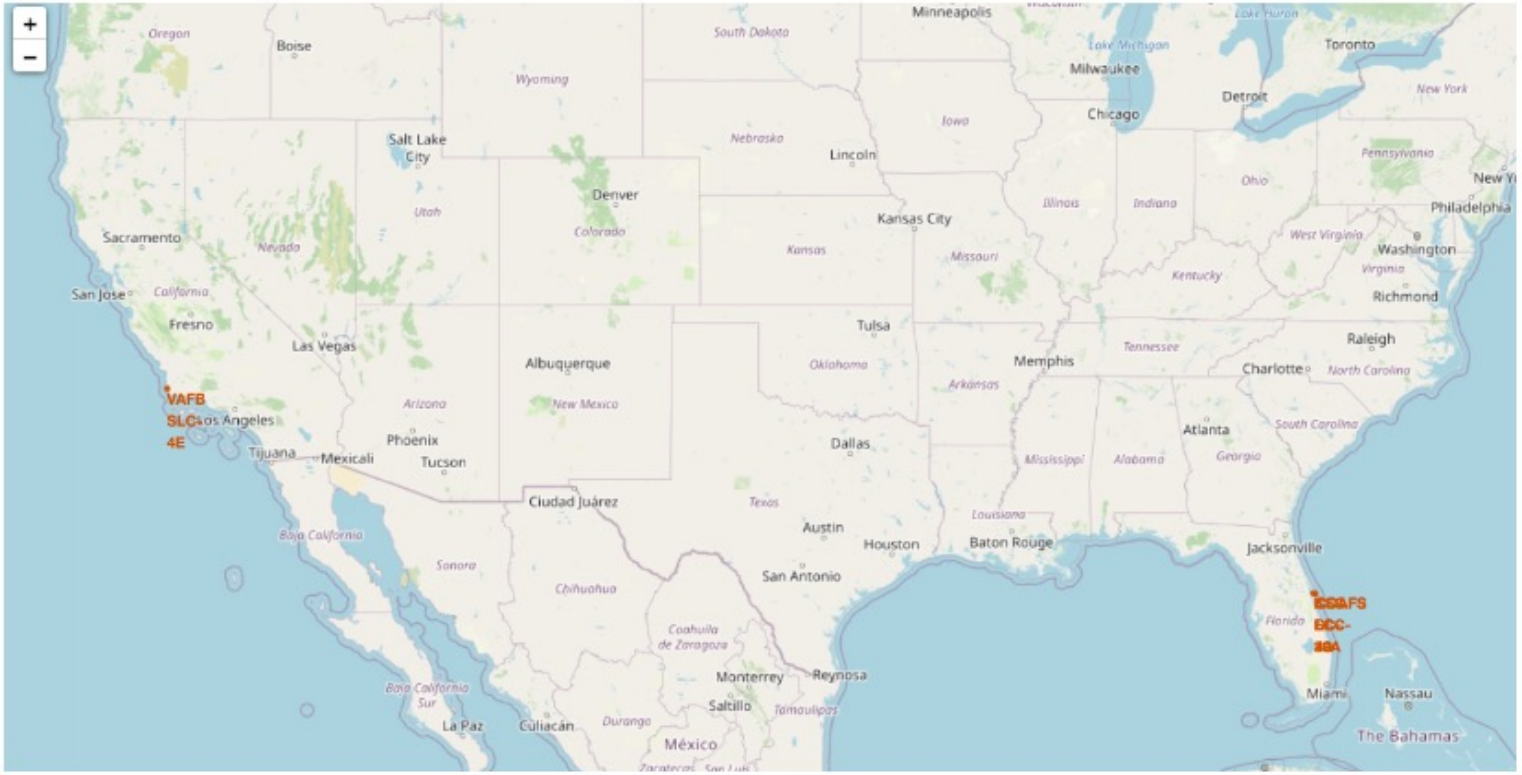
```
 * ibm_db_sa://cdm26718:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

landing__outcome	total_number
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 3
Launch Sites
Proximities Analysis

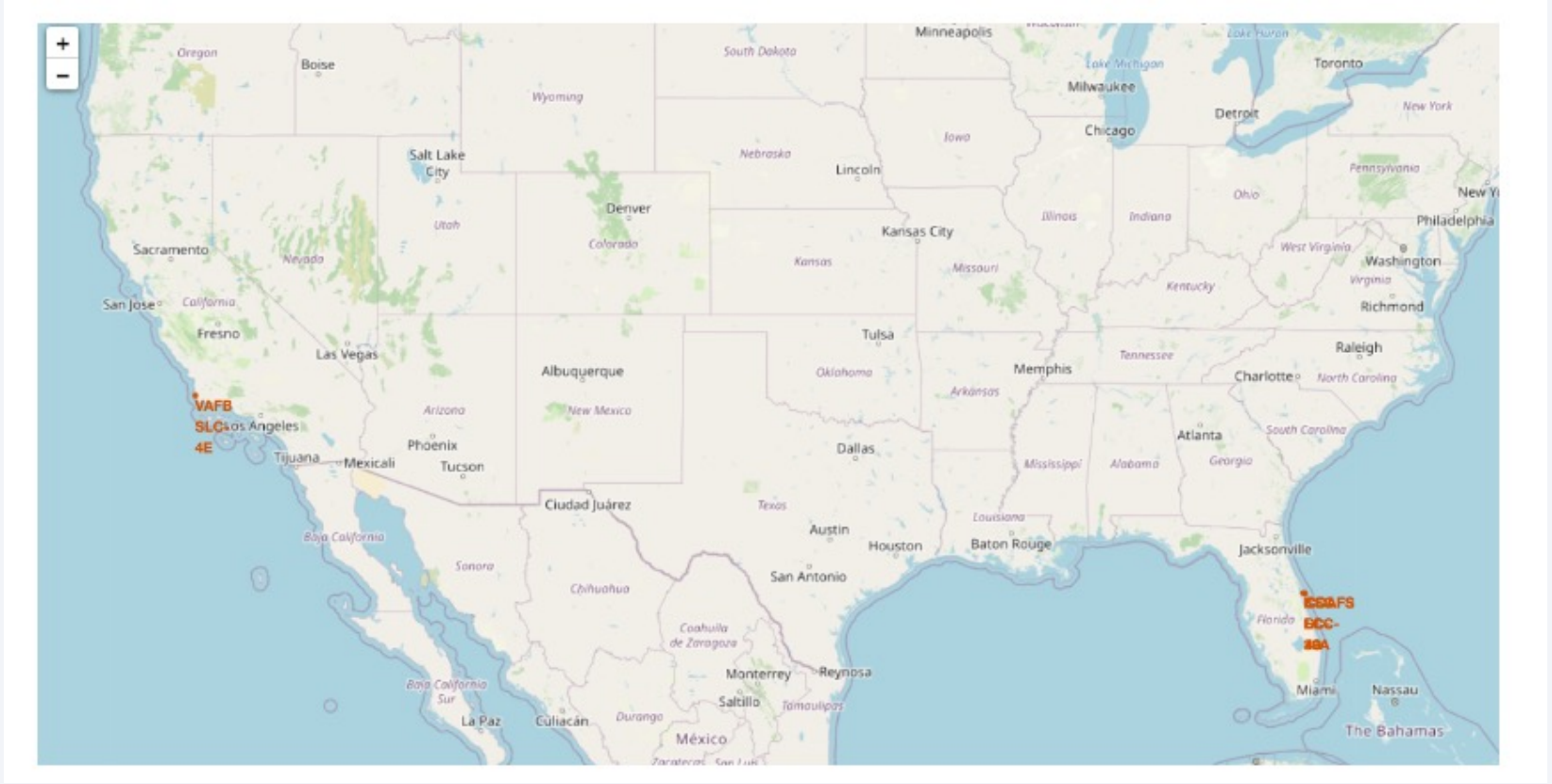
Geo localization of Launch Sites

All the sites are close to the Equator line. Why ? Less distance.



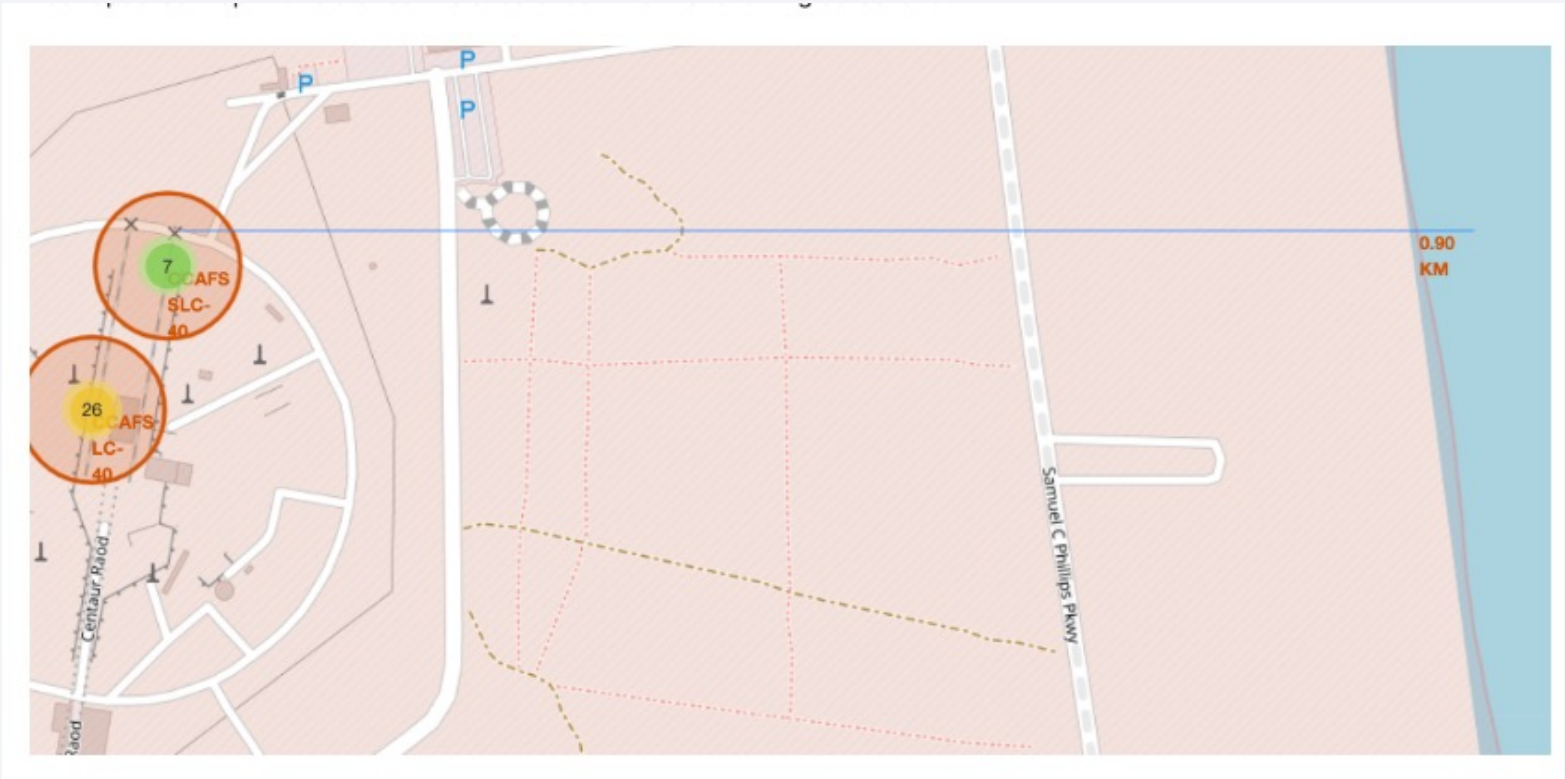
Number of launches per site

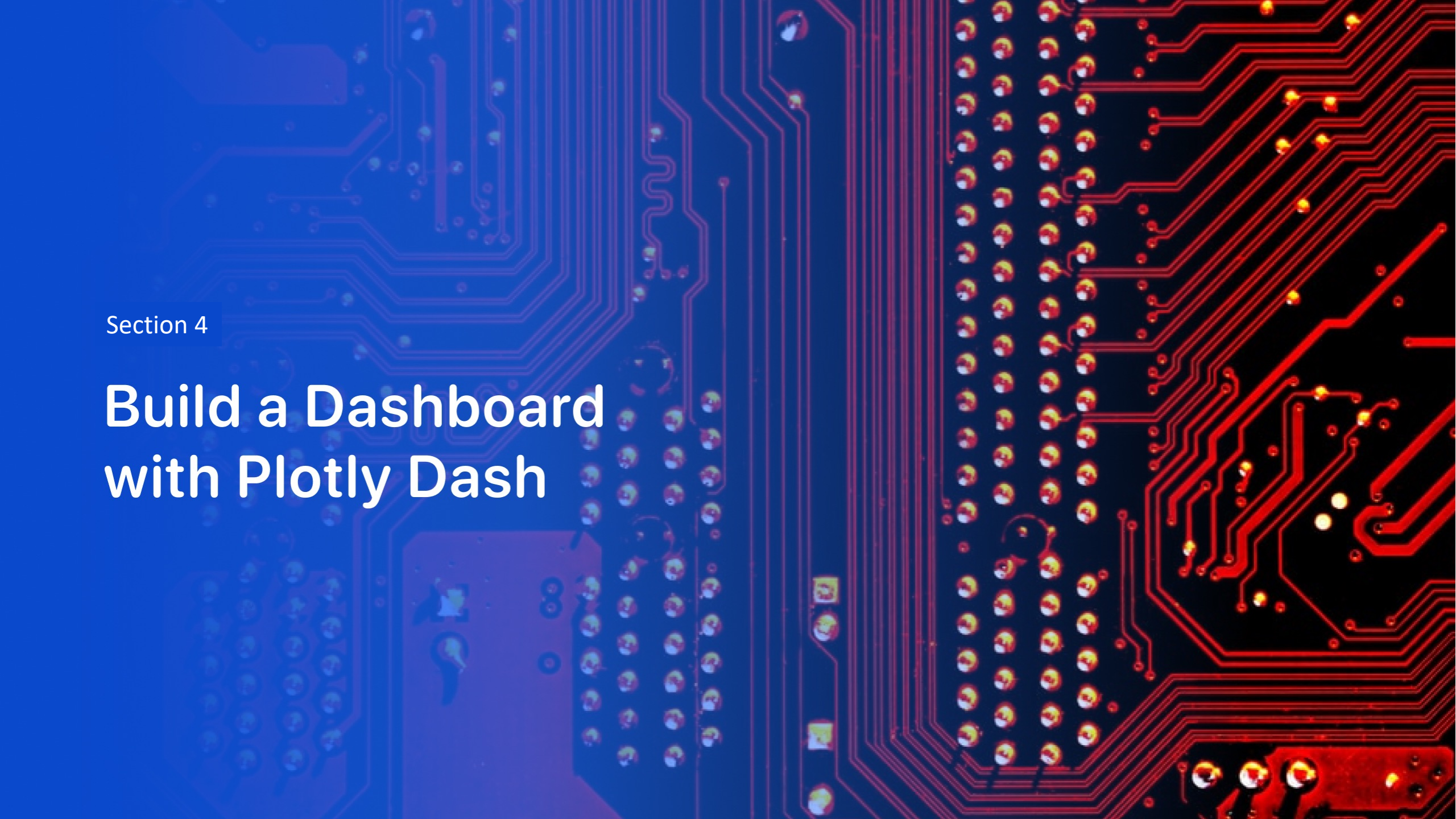
- ¾ of launch where made in Florida.



Distance from Coastline

- Close to the coastline to make it safer in case of incident.



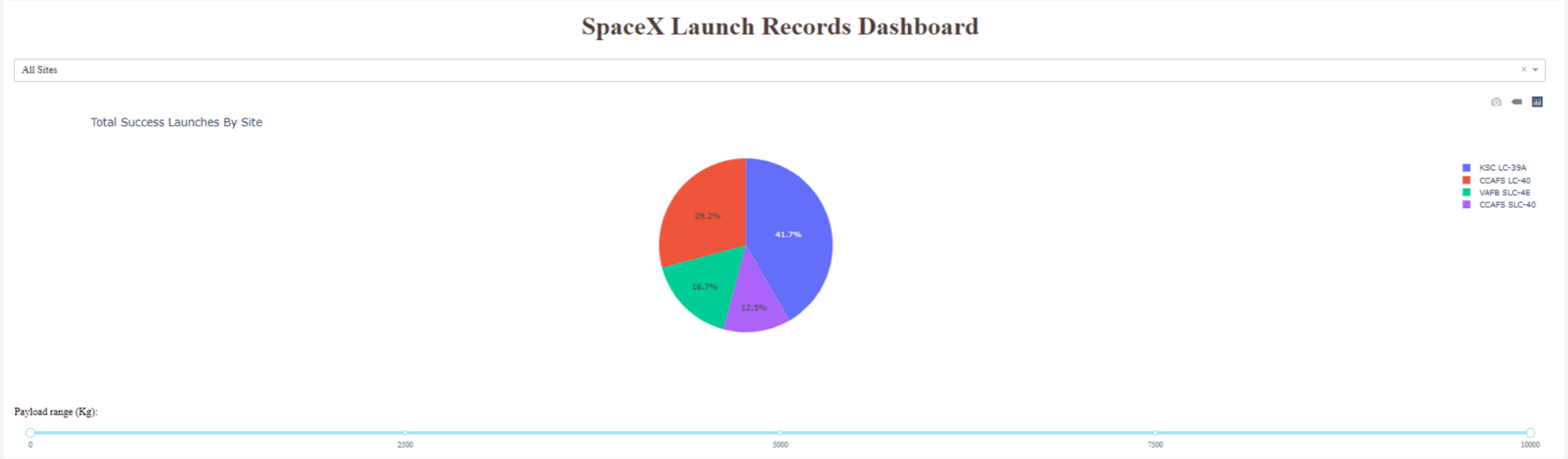


Section 4

Build a Dashboard with Plotly Dash

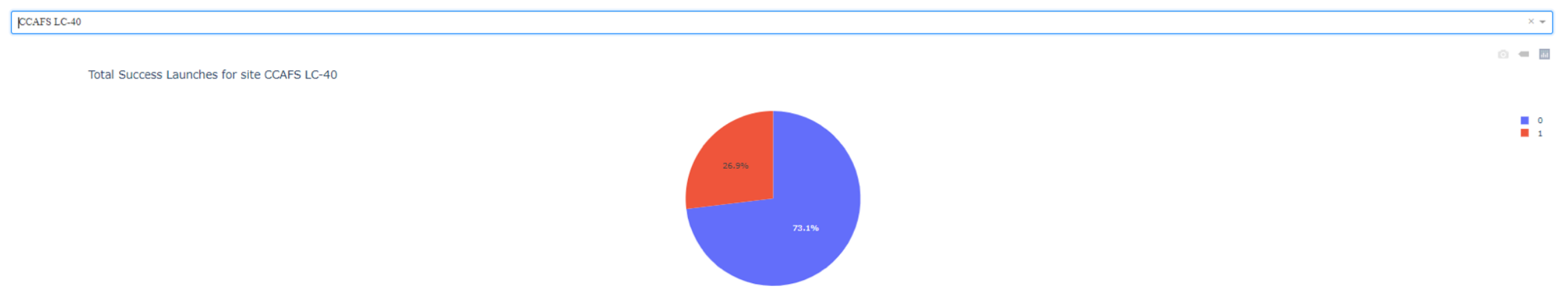
Successful launches per sites

- KSC LC-39A is, by far, the most successful launch site



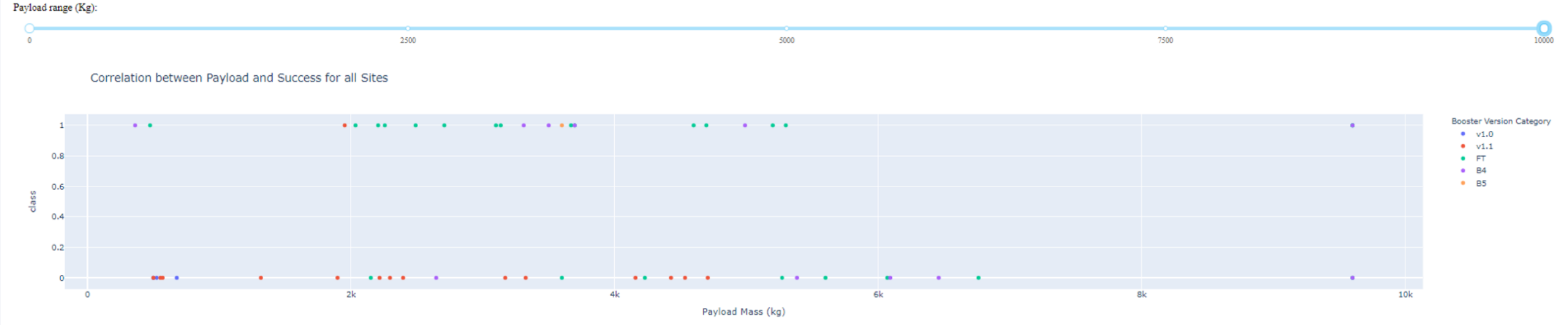
CCAFS LC-40 Success launches

- 73,1% of success.



Payload vs Outcome

- The bigger is the mass the more successful the launch is.

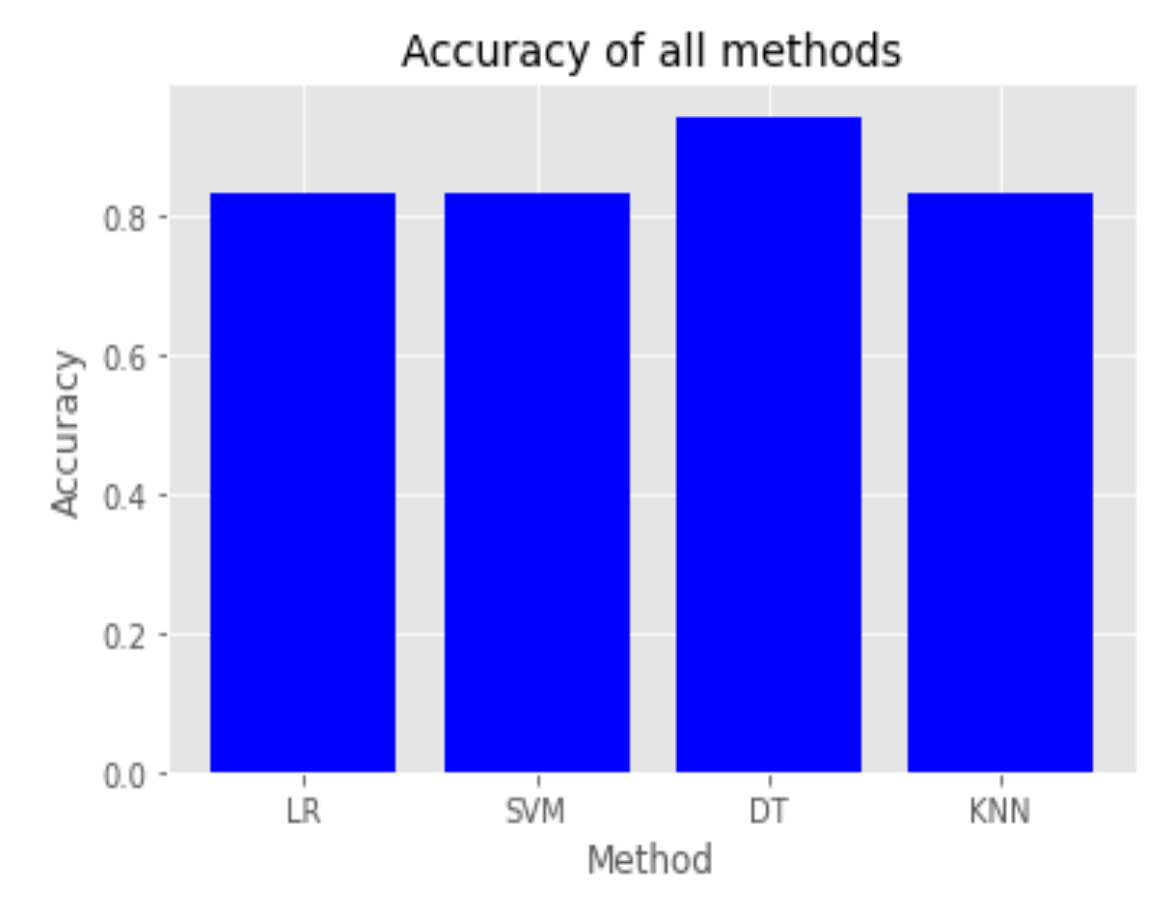


Section 5

Predictive Analysis (Classification)

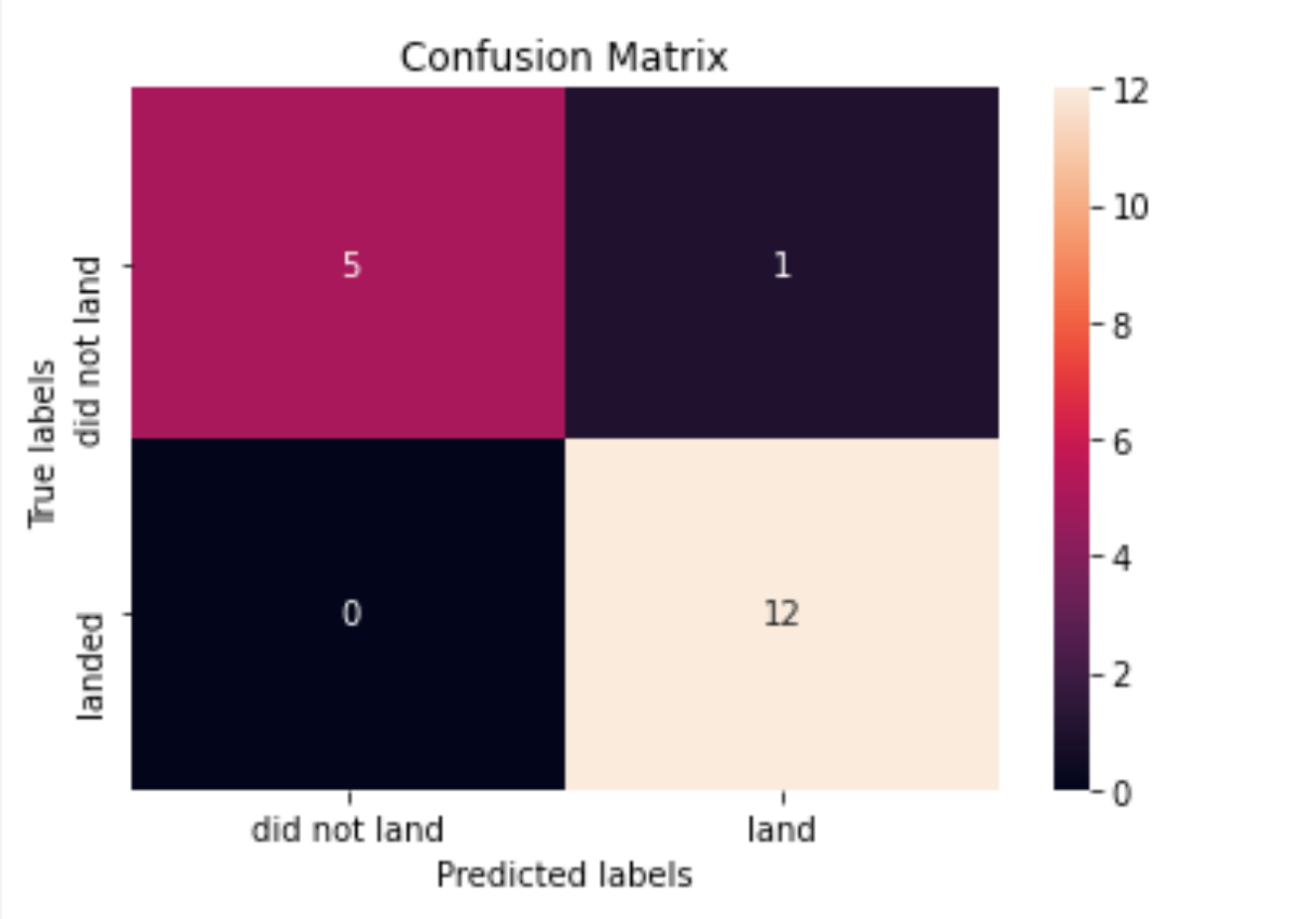
Classification Accuracy

- Decision Tree is the most accurate model in this case.



Confusion Matrix

- There is only one missing point.
- All the other model had 3 missing points.



Conclusions

- The heavier the Payload the better.
- The closer you are to the Equator line the more successful your launches will be.
- Don’t launch your rocket on the GTO Orbit.

Thank you!