Winning Space Race with Data Science

Bruno Amaral
15-12-2023

Outline

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

Executive Summary

Summary of methodologies
-Data collection
-Data wrangling
-EDA with data visualization
-EDA with SQL
-Exploring and preparing data
-Interactive visual analysis with Folium
-Building interactive HTML dashboard with Dash Pllotly
-Predictive analysis(classification methods)

Summary of all results

-EDA for best launching predictions.

-Machine Learning Prediction shows the best model for predict if the 1st launch will land successfully,mostly in the oceans.

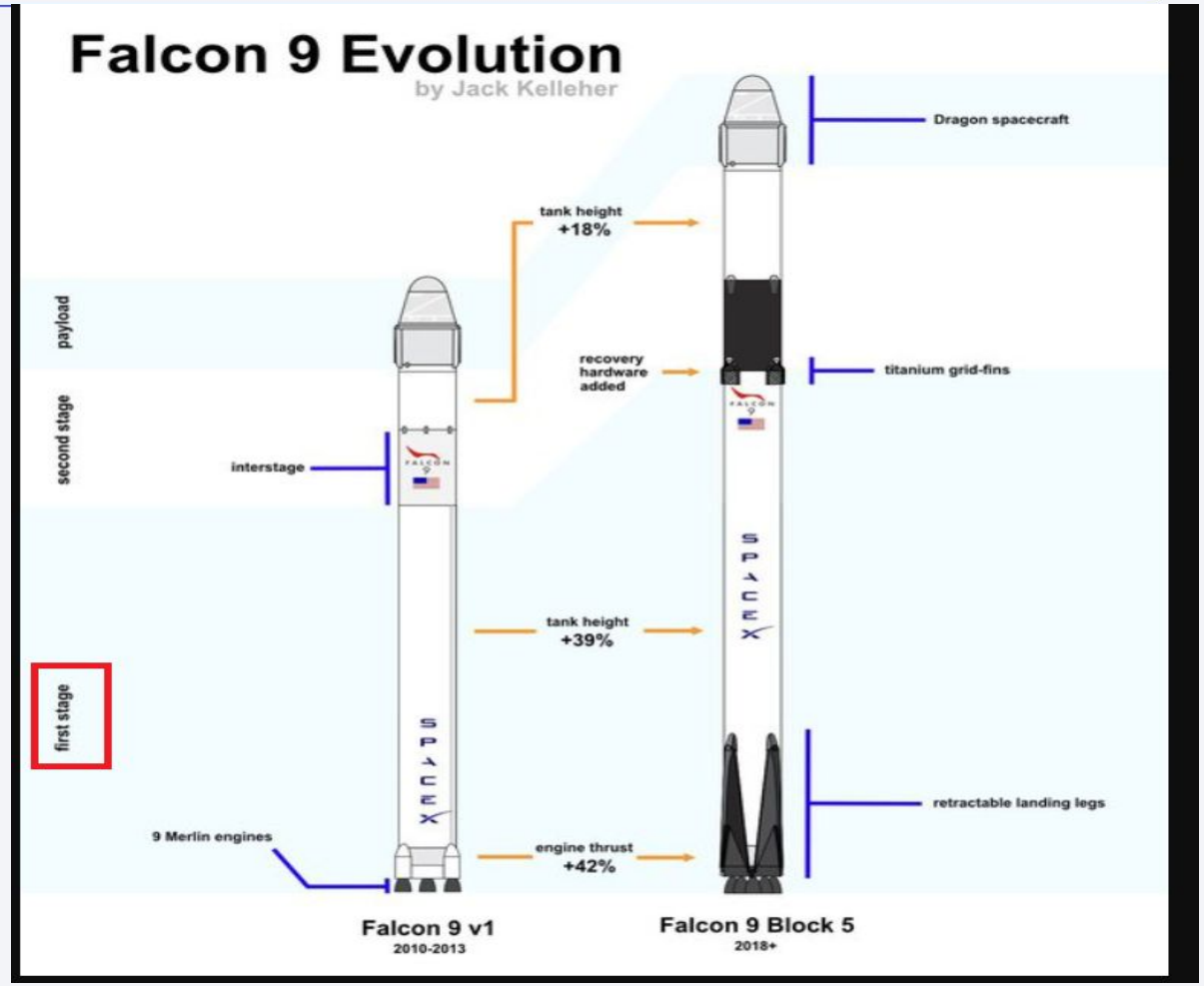
Introduction

- **Project background and context**

-Space X can save a lot of money by reusing the first stage of each launch,so
our job is to adapt and study their data so we can use it in a rival company named space Y.

- **Problems you want to find answers**

-Check the best launch site for successful launches and predict if the first stage will land correctly so it could be reused ,therefore saving big money for Space Y



Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  -Space X API

  -Web scraping from wikipedia

- Perform data wrangling

  -Cleaning the data ,eliminate null values and less important columns,add classification column

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

-Collect the data,normalize it ,split in train and test data to be evaluate by 4 methods o classification (LOG Regression,SVM,KNN and Decision Tree),check which one has the best accuracy

Data Collection

- **Describe how data sets were collected.**

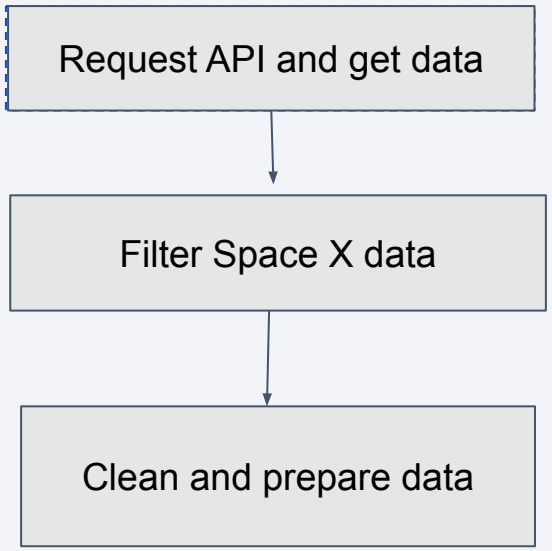
  - Data sets were collected via Wikipedia with BeautifulSoup and Space x API,using python programming language with pandas library.

Data Collection – SpaceX API

-The data was request from Spacex X API,from it we can get all data necessary for the project.

-Github link:

https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab1%20Data%20Collection%20API%20(Application%20Programming%20Interface).ipynb

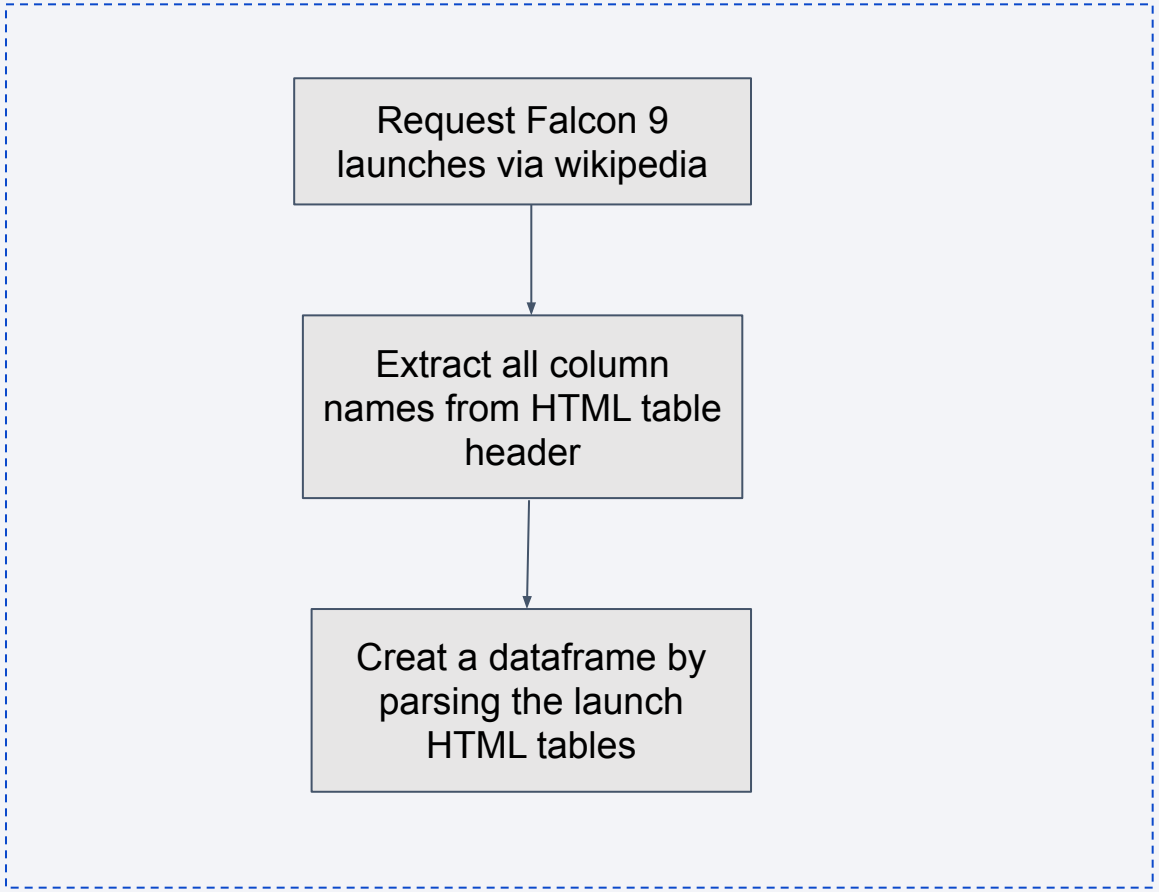


Data Collection - Scraping

- Data was collected via wikipedia with BeautifulSoup:https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922..Then extract Falcon 9 launch records in HTML table, parsing the data and convert to pandas dataframe.

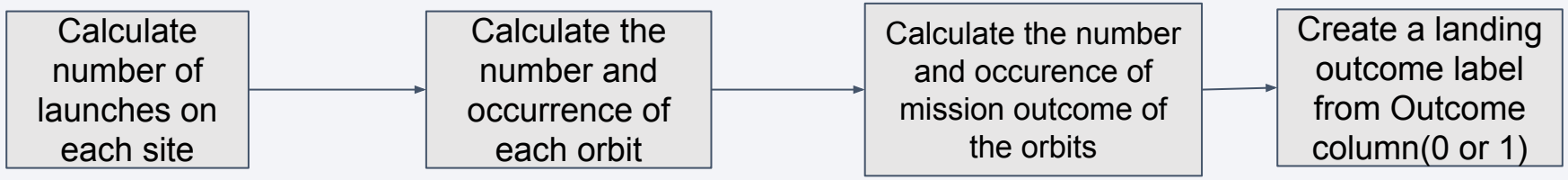
-Github link:

https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab2%20Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

-Convert the successful launches in 1 and 0 for unsuccessful and add to our dataframe for further analysis.

- 

- Github link::

  https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab3%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Charts :

  The scatter plot was used because of the relationship between 2 variables (0 and 1 for class),the bar plot was used between orbit group and successful rate(fixed value per orbit) and with variables changing ver time we can use a line plot

  For payload mass and Flight number for each site was plotted a scatter plot,for orbit type vs successful rate was plotted a bar plot and for successful rate per date was plotted a line plot.

- Github Link:

  https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab5%20Graphs%20Exploring%20and%20Preparing%20Data.ipynb

EDA with SQL

- SQL queries :
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
  - 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.
- Github Link:

  https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab4%20EDA(Exploratory%20data%20analysis)%20with%20SQL.ipynb

Build an Interactive Map with Folium

- folium.Map:Creattion of a map with center location to be NASA Johnson Space Center at Houston, Texas.
- folium.Circle:highlighted circle area with a text label on a specific coordinate
- folium.Marker:to give more information of certain points

Github link:

Images-https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab6%20images.png

Lab-https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab6%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

Charts:

- Pie Chart
- Scatter Chart

Pie chart is useful because we can visualize by parts and colours the percentage os successful launches by launch site,is more interactive,easier and faster to analyze.

SCatter charts are good to analyze large data with numerical variables in this case 0 and 1 .

- Github Link: https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab7%20spacex_dash.py

Predictive Analysis (Classification)

- First we standardize the data then split in train and test data to be evaluate by 4 methods o classification (LOG Regression,SVM,KNN and Decision Tree),check which one has the best accuracy

- 

- Github link:

  https://github.com/Brunoamaral0/Coursera-capstone-project-data-science/blob/main/Lab8%20Machine%20Learning%20Prediction.ipynb

Results

Exploratory data analysis results

-CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.(in flight number x lunch site y graph)

- VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

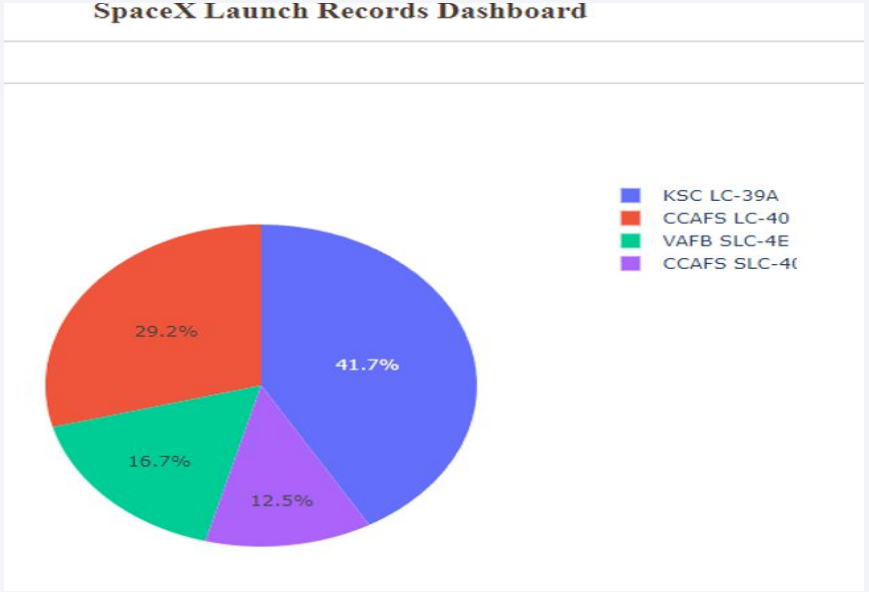
-Orbits type with best successful rate are ES-L1,GEO,HEO,SSO

-Success rate since 2013 kept increase, heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISSing till 2020

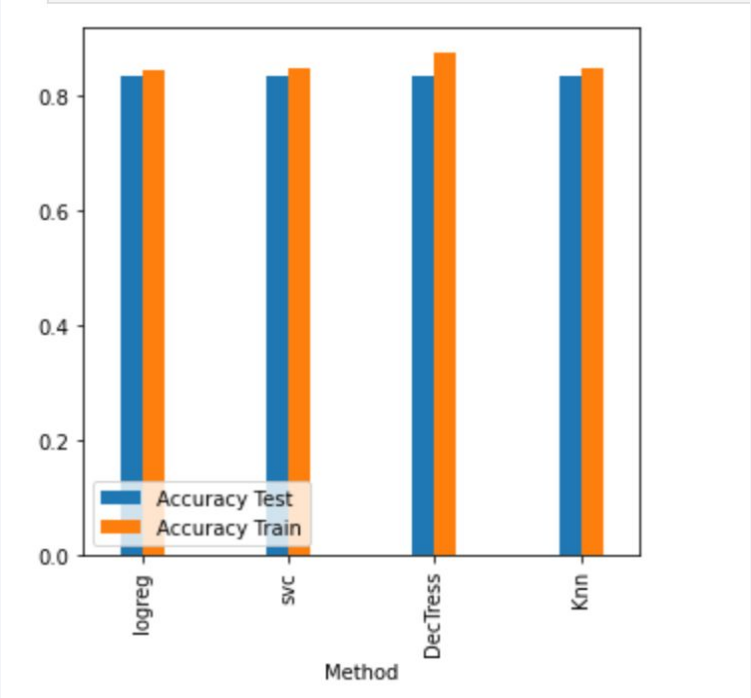
-Success rate of launches since 2013 kept increasing till 2020

Dash results (KSC and CCAFS have highest success launch rate)

Note:dataframe used in dahs is different from the one used EDA

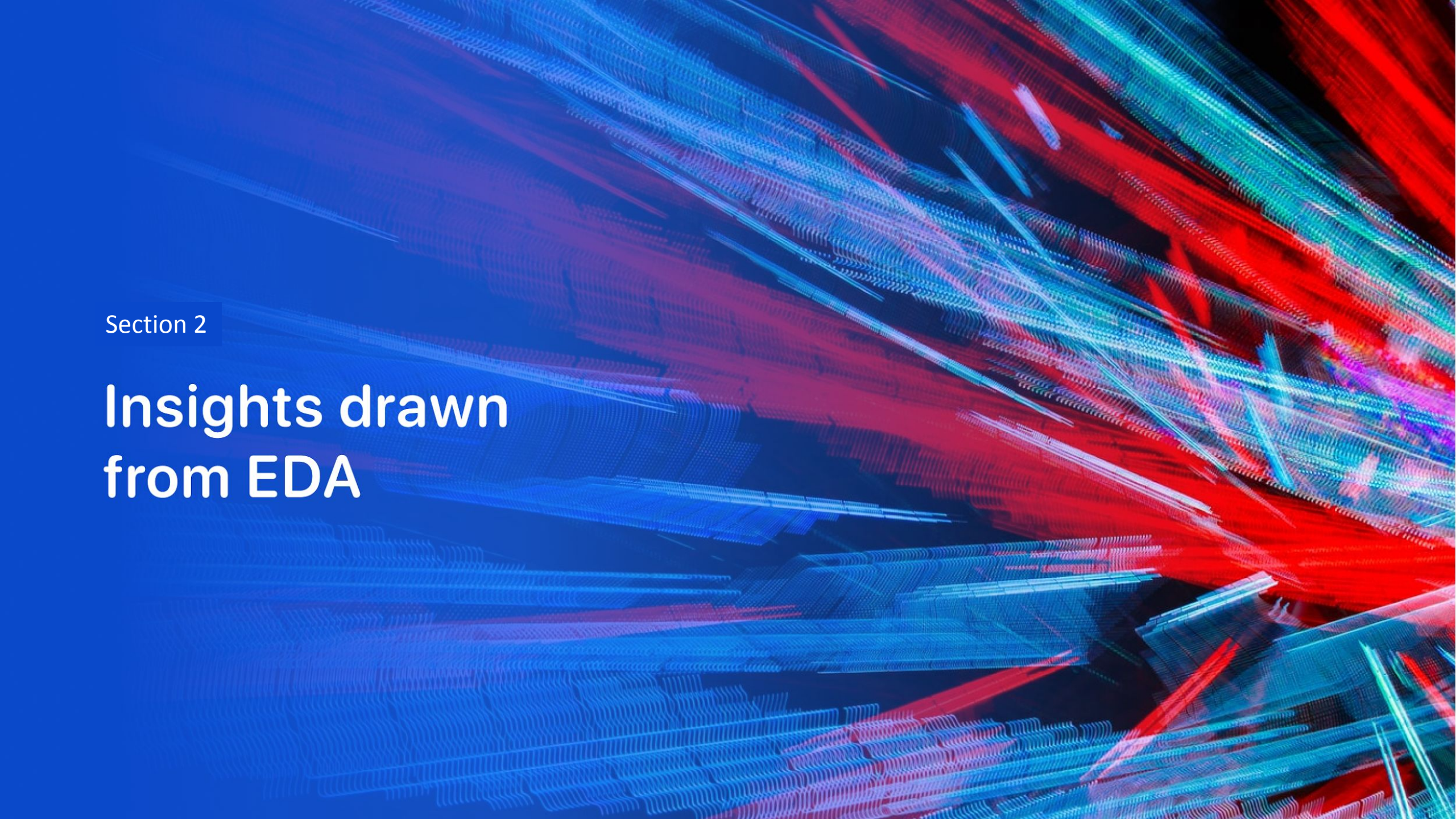


Predictive analysis results

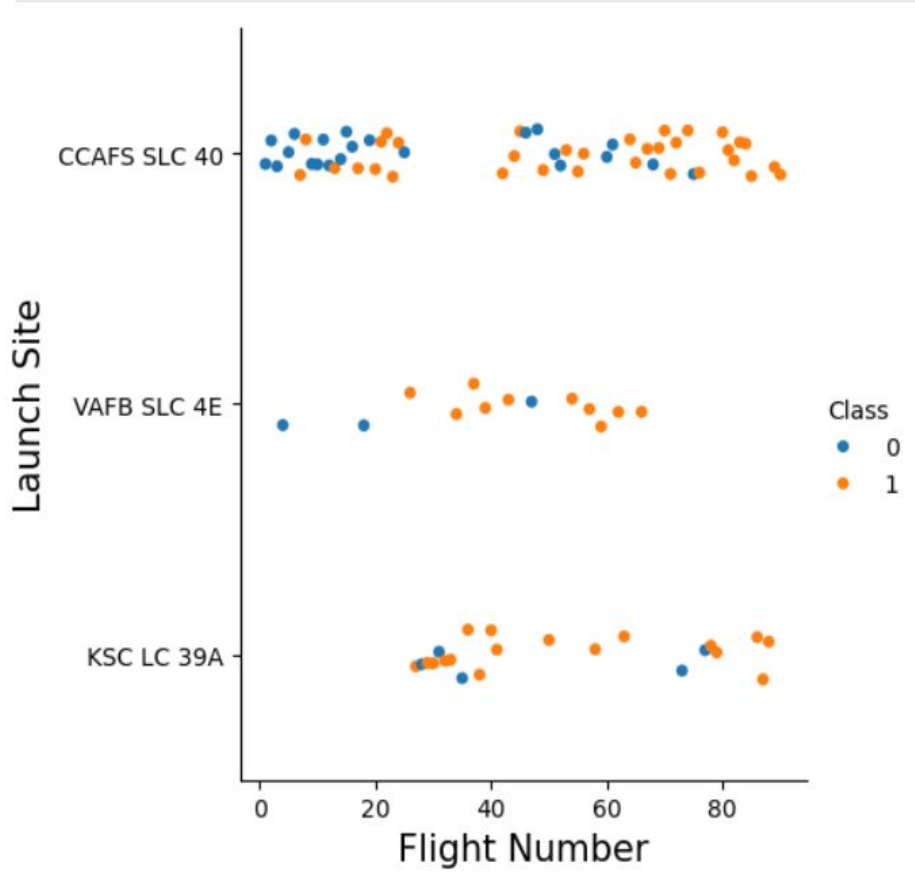


Section 2

Insights drawn from EDA



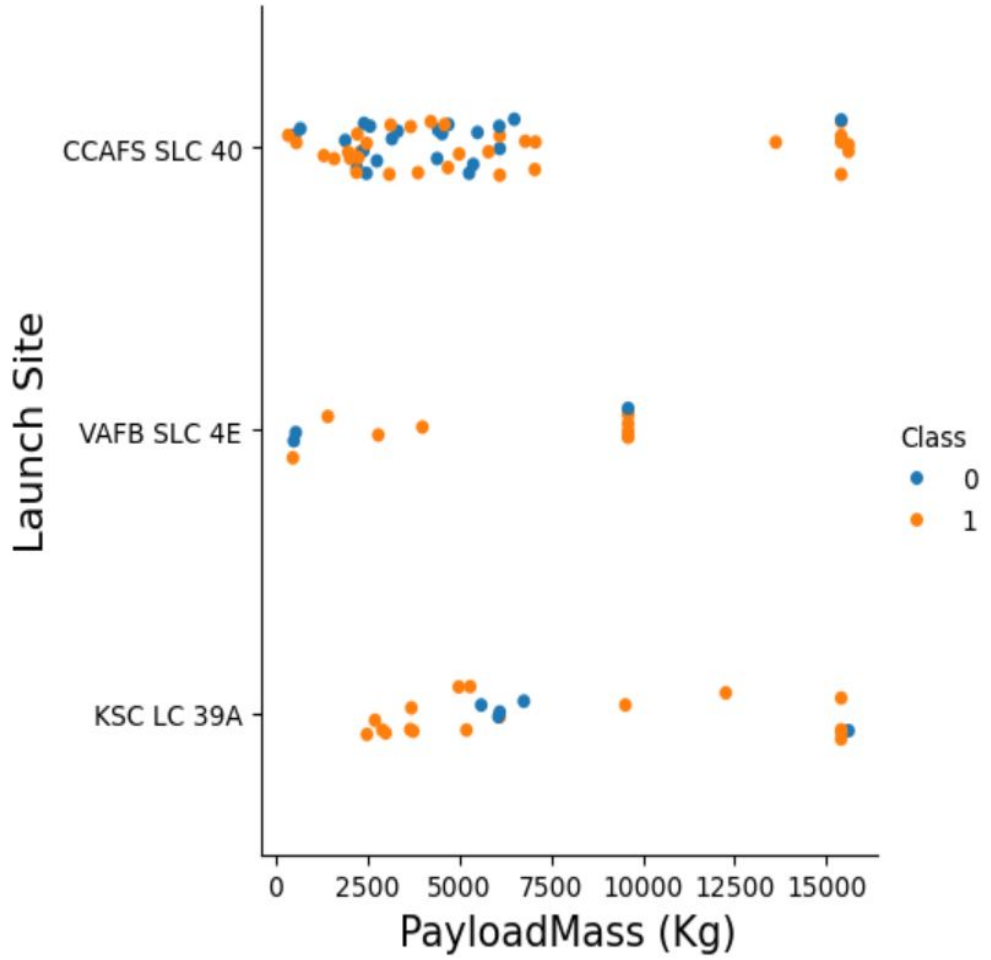
Flight Number vs. Launch Site



- In the graphs we can verify that the launch sites with highest rate of success landings are KSC and VAFB with 15 successful and 5 unsuccessful ,rate of 15/20=75% for KSC and 10 successful and 3 unsuccessful for VAFB rate =77%.

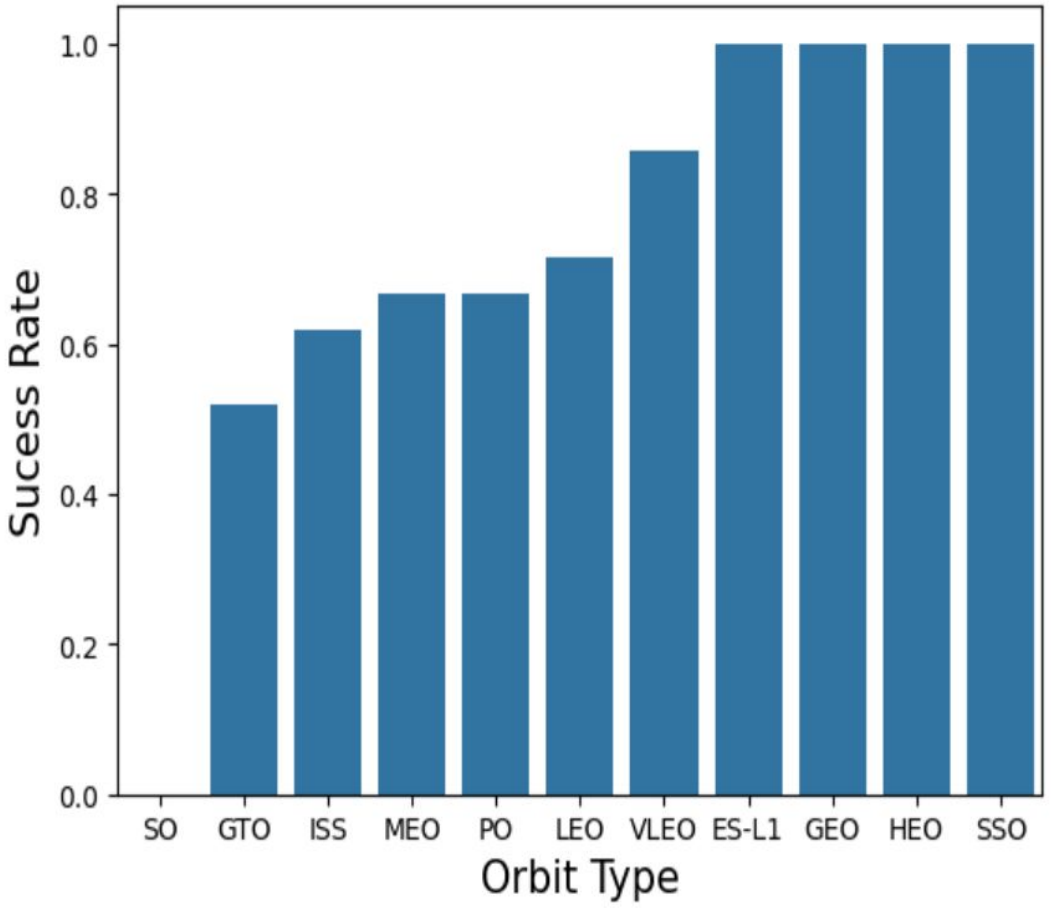
- CCAFS has only 60 % with 33 successful and 22 unsuccessful rate =60%.

Payload vs. Launch Site



- CCAFS has good success rate with payloads higher than 10 tons with 5 successful and 1 unsuccessful rate =83,3% .
- VAFB has no record for payloads above 10 tons while under 10 tons the results are inconclusive due the lack of launches.
- KSC has all launches successful for payloads under or equal 5 tons,while between 5 tons and 7,5 tons the launches were all unsuccessful,above 10 tons there were only one successful launch to 5 successful ,but due to lack of launches we can't say that if we use payload with more than 10 tons the launch has high rate for success.

Success Rate vs. Orbit Type

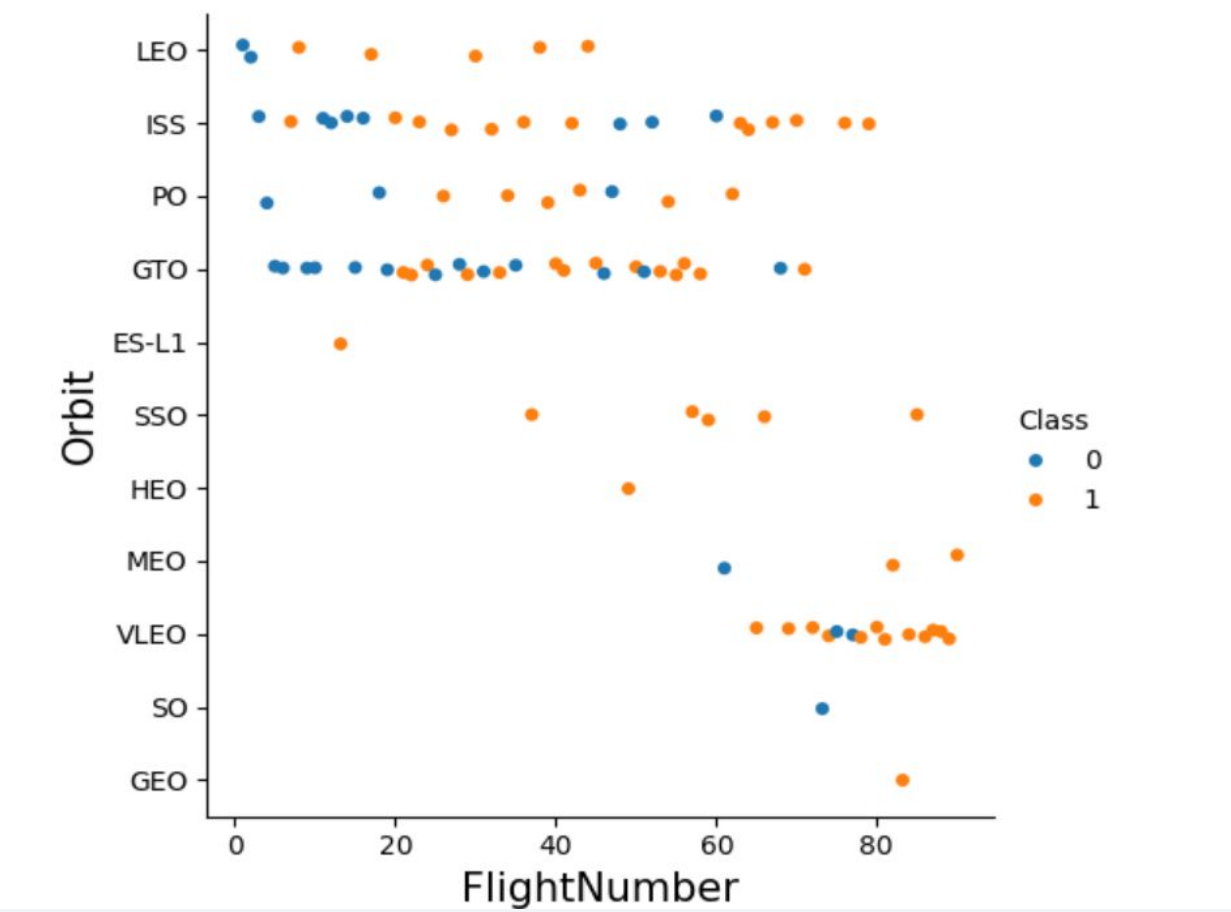


- The orbits with highest success rate were ES-L1,GEO,HEO,SSO .Is important to emphasize that ES-L1,GEO,HEO only made one launch so very inconclusive, while SSO did 5 launches all succeed.

- VLEO still the best orbit for prediction next launch due to high rate,85% and 14 launches

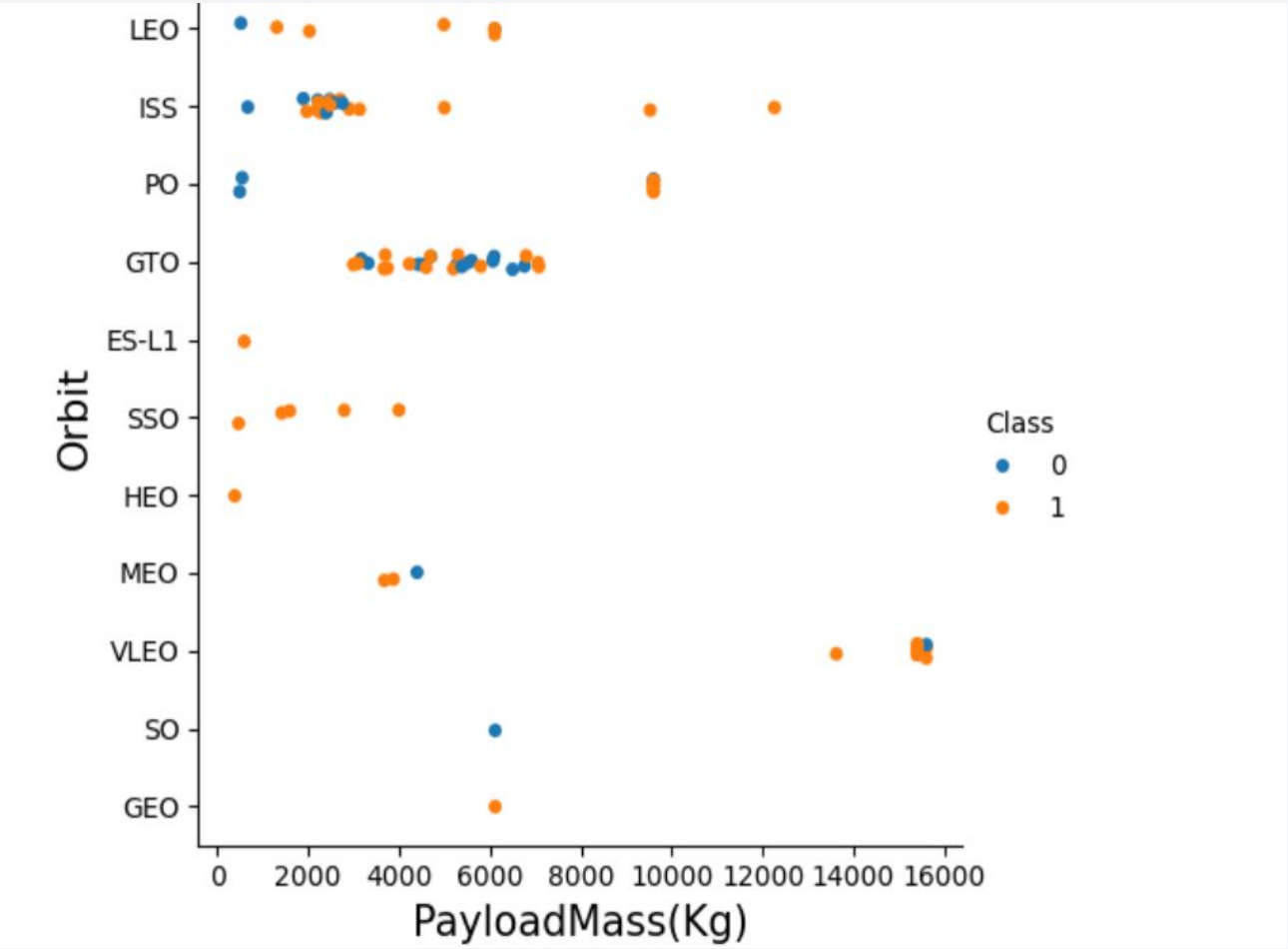
- All the other orbits are inconclusive very few launches.

Flight Number vs. Orbit Type



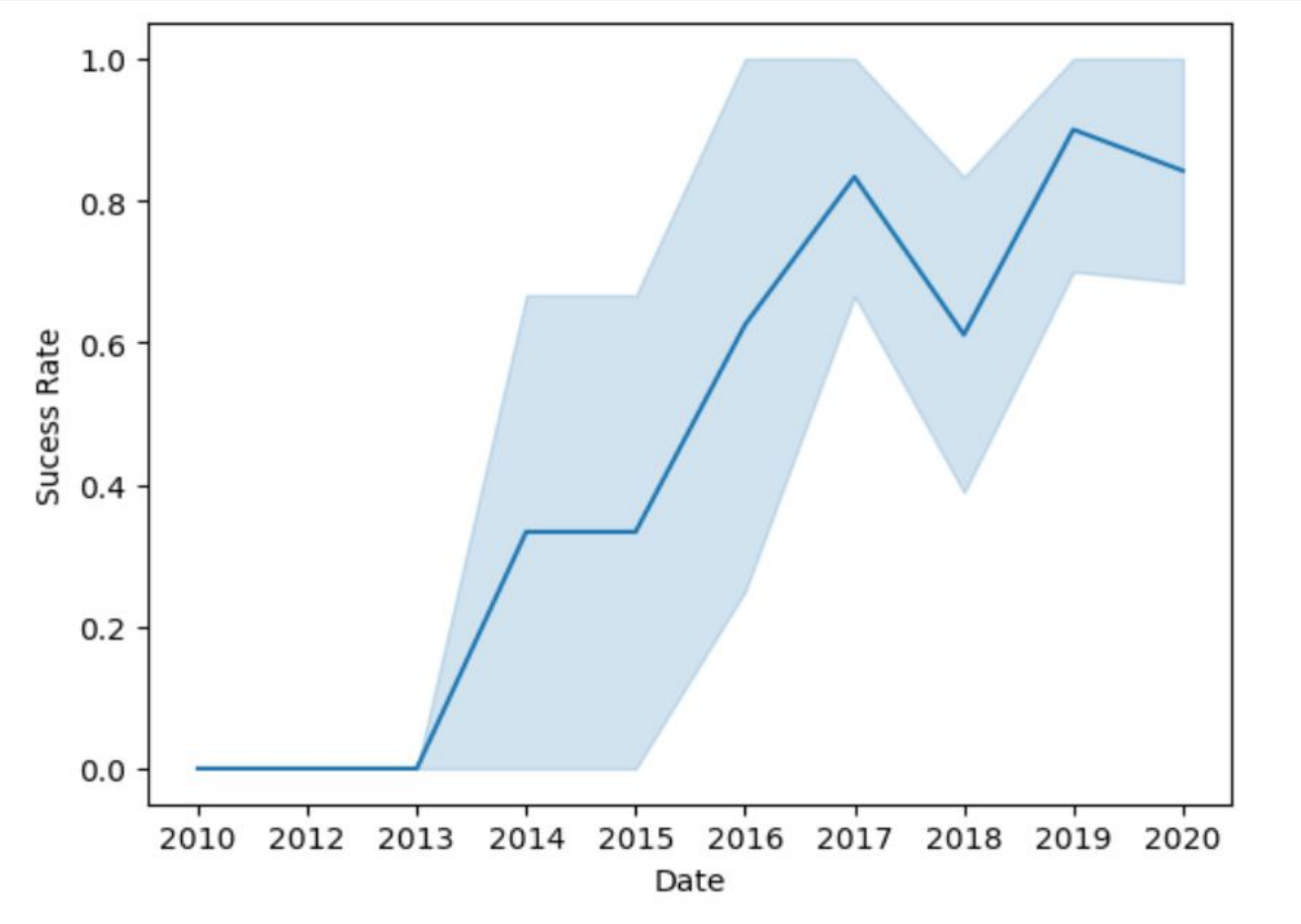
- SSO did 5 launches all succeed,100% success rate.
- VLEO is the best orbit for prediction next launch due to high rate,85% and 14 launches
- No launches were made with SO orbit

Payload vs. Orbit Type



- SSO orbit had payloads mass all under or equal 4 tons .
- VLEO orbit had all launches higher than 12 tons and less than 16 tons
- GTO orbit many launches between payload masses of 2 and 8 tons

Launch Success Yearly Trend



- Success rate has grown over time since 2013.
- In the period of 2016 and 2018 ,success decreased by 20 %.

All Launch Site Names

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

- The mission has 4 launch sites in our dataframe.

Launch Site Names Begin with 'CCA'

9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Exemple with 5 launch sites started by “CCA” in our dataframe and their labels.

Total Payload Mass

10]:
```
%sql SELECT sum(PAYLOAD_MASS__KG_) as sum from SPACEXTABLE where Customer like '%NASA (CRS)%'
```

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

10]:

sum
48213

row 76 contain [NASA (CRS), Kacific 1] ,if is only NASA (CRS) the code must be

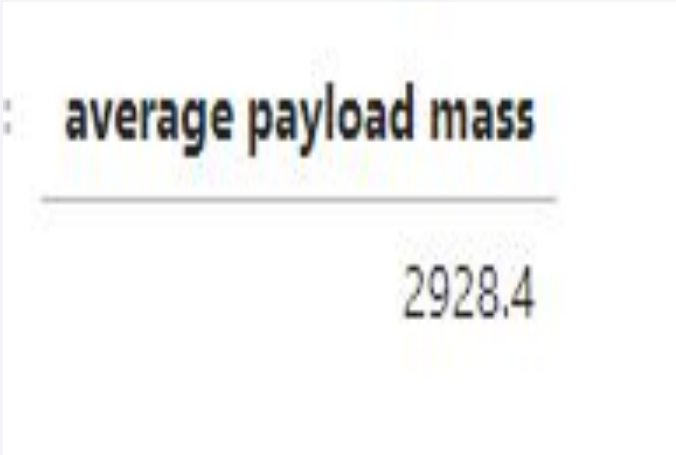
11]:
```
%sql SELECT sum(PAYLOAD_MASS__KG_) as sum from SPACEXTABLE where Customer ='NASA (CRS)'
```

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

11]:

sum
45596

Average Payload Mass by F9 v1.1



- Average payload mass by F9 v1.1 under 3 tons.
- Average of 2928 kg

First Successful Ground Landing Date

min (Date)
2015-12-22

- Date where first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Only 4 boosters landing in drone ship successful with weights between 4 and 6 tons.

Total Number of Successful and Failure Mission Outcomes

outcomes
1
98
1
1

- 98 successful outcomes
- 3 outcomes were not 100 % successful

Boosters Carried Maximum Payload

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

- All boosters recorded that carry the maximum payload mass

2015 Launch Records

Month Name	Booster_Version	Landing_Outcome	Launch_Site
1	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
4	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

- Failure landing outcomes in drone ship in the year 2015 in months of january and april for boosters B1012 and B1015.

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

number	Landing_Outcome
9	Success (ground pad)
5	Failure (drone ship)

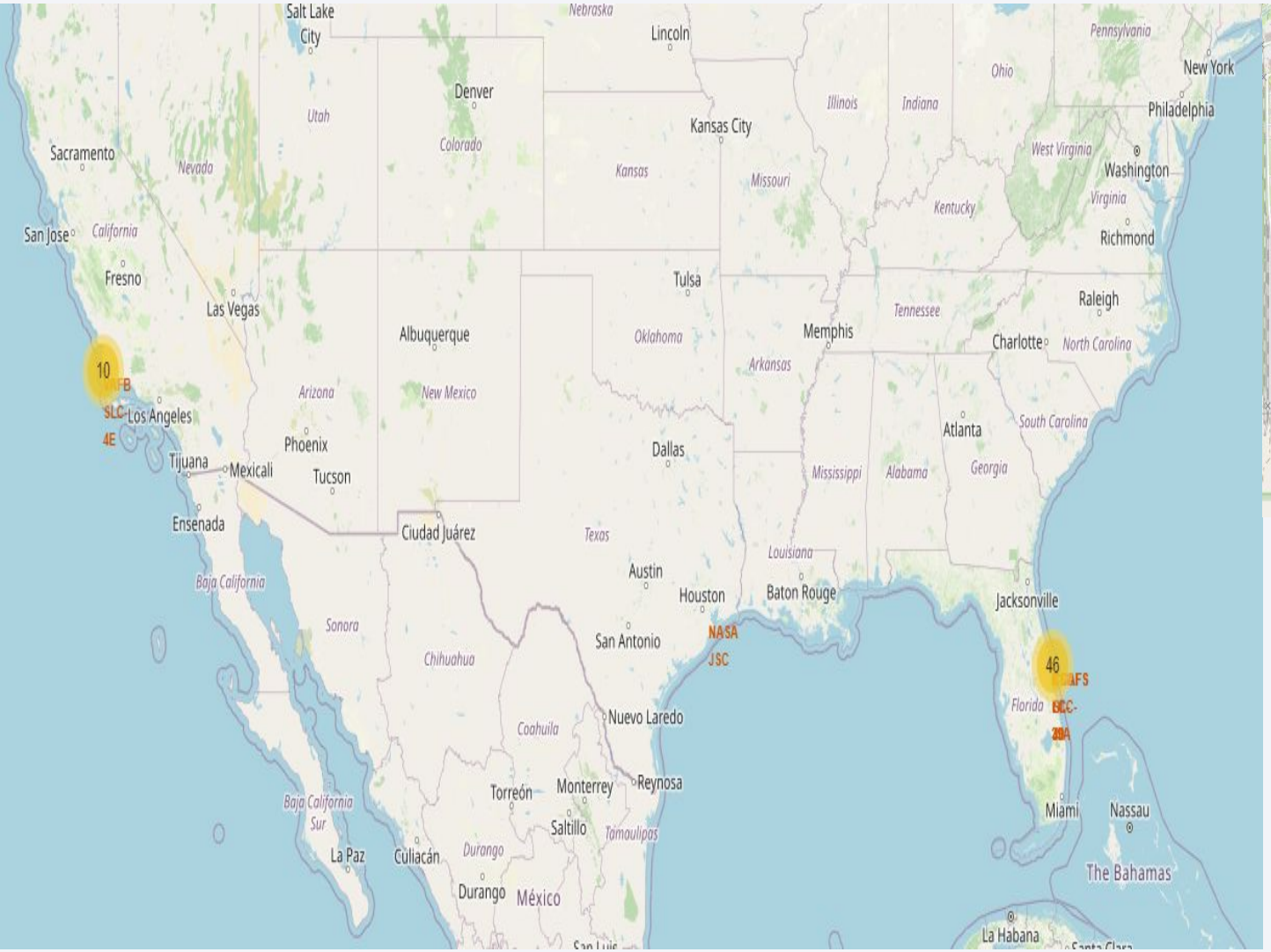
- 9 launches successfully landing on ground
- 5 failures in landing via drone ship

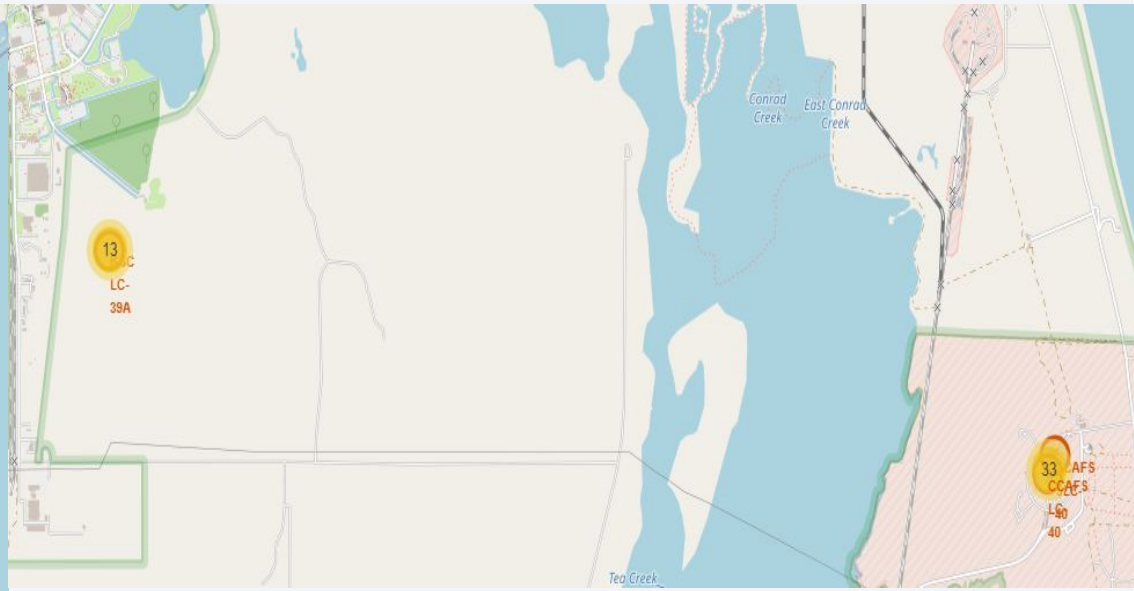

Section 3
Launch Sites
Proximities Analysis

Launch sites

- All launch sites made on the coast.
- Far from cities.

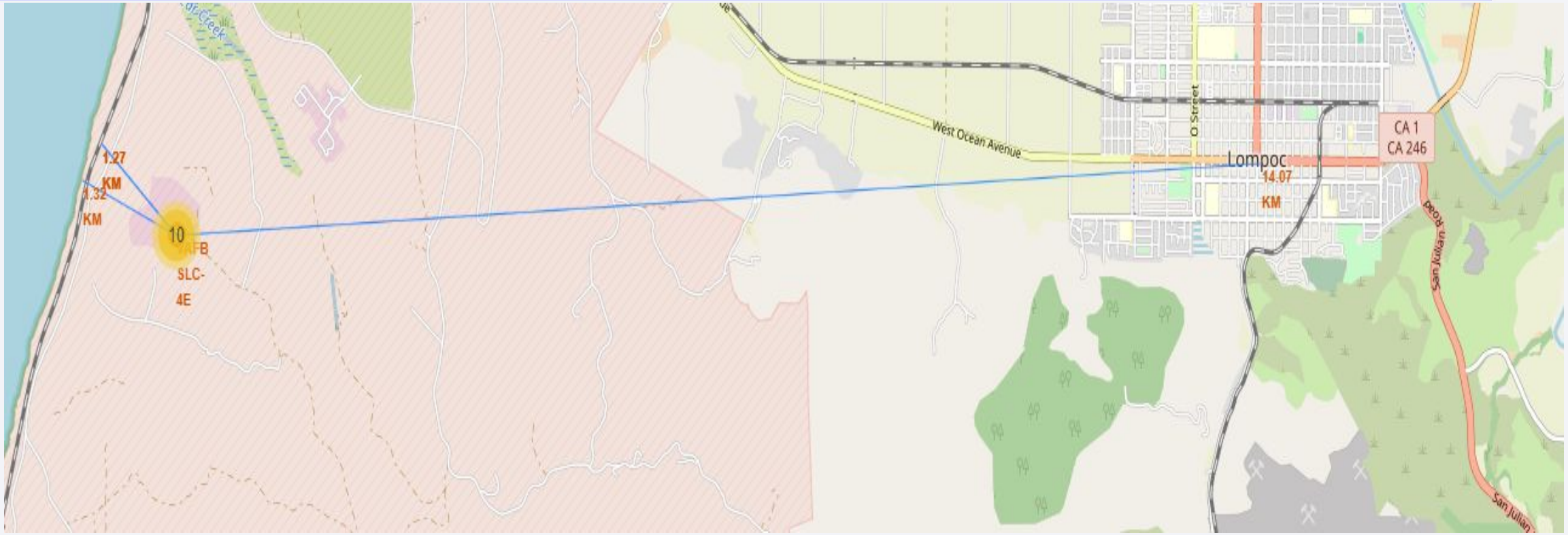
Number of launches per site



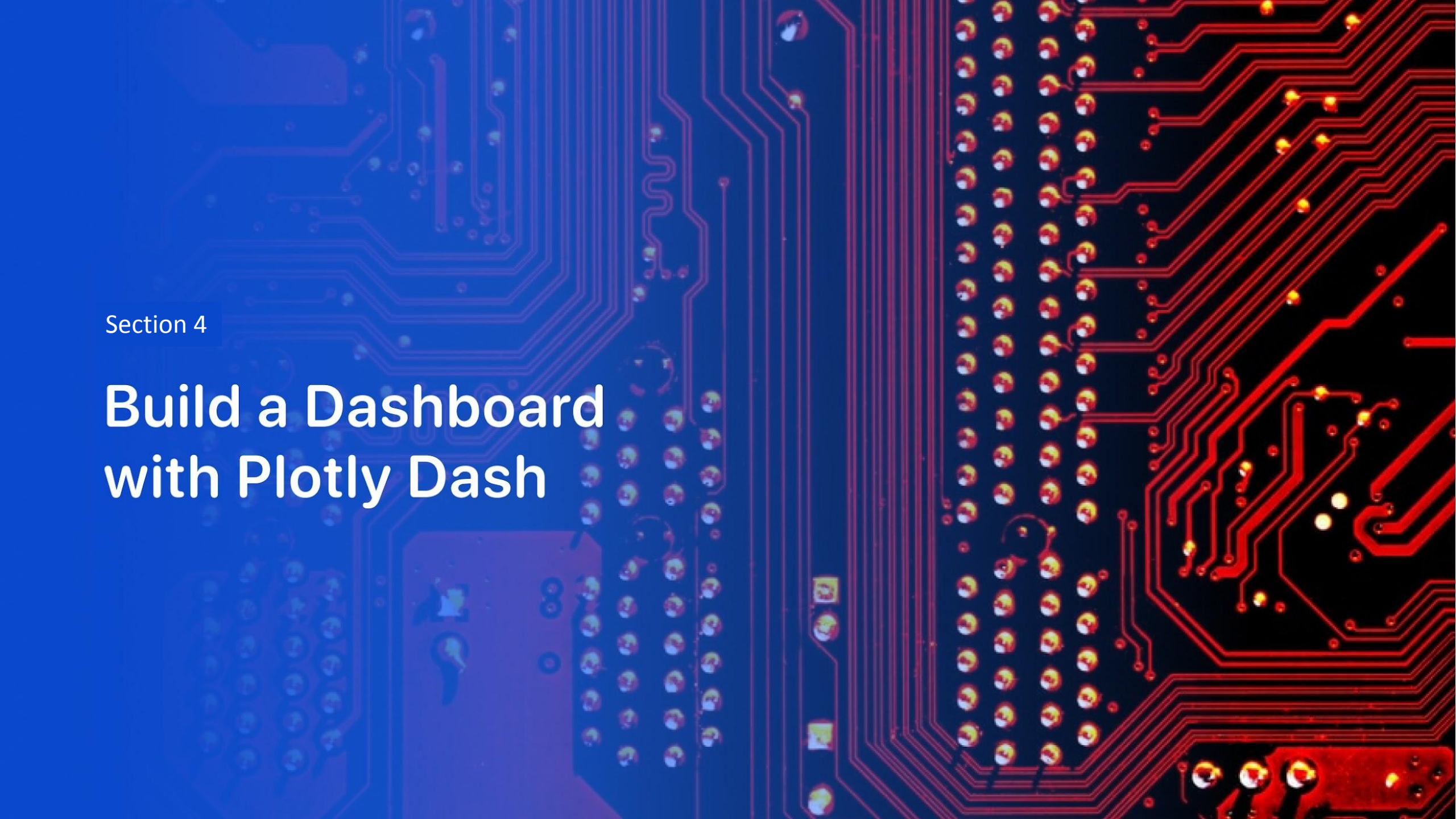


- on the left we can check the number of launches on each site, 10 launches west side and 46 east side.
- on the right we can see more detail the exact location of the 46 launches.

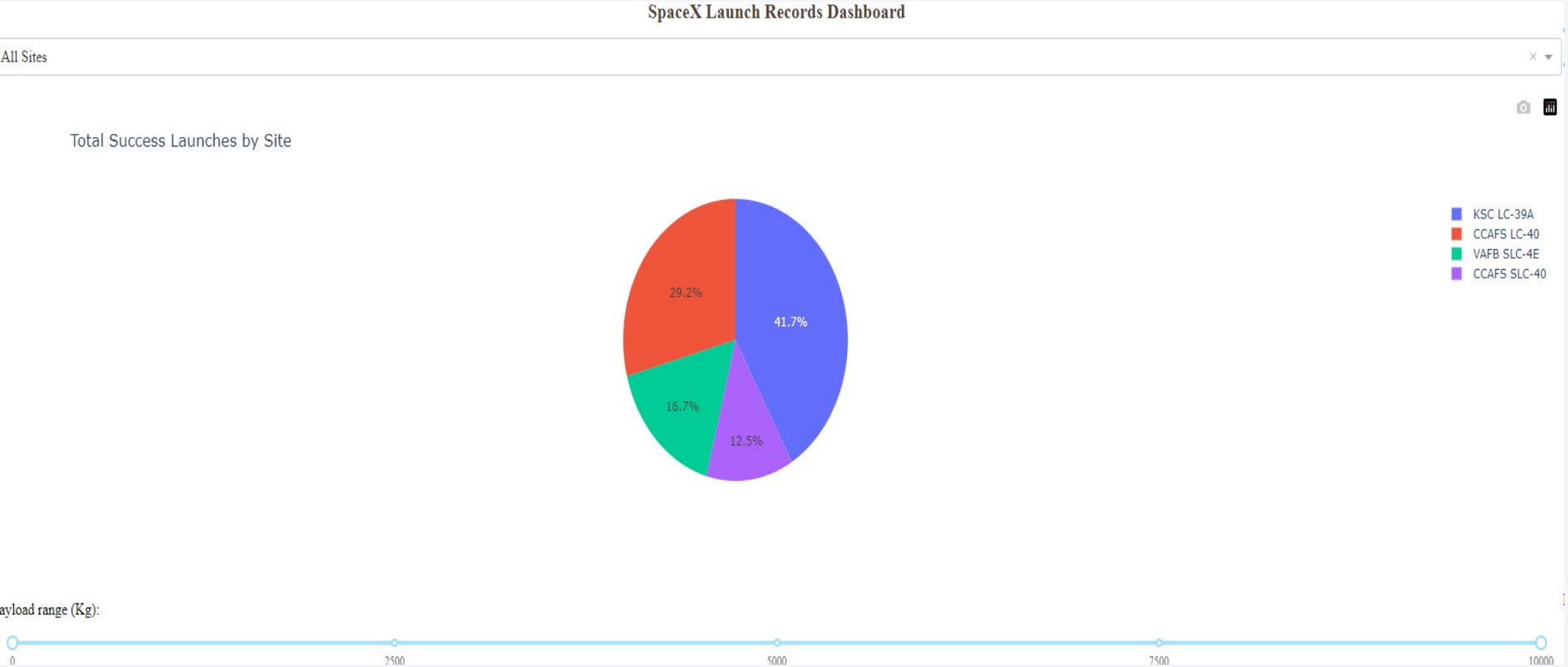
Nearby distances (Highway,Railway,Coastline)



- Nearby distances in the launches sites from the west site map location
- 1 km and 270m from the nearest railway,1km and 320m from the nearest coastline and 14 km and 70 m from nearest highway


Section 4
Build a Dashboard with Plotly Dash

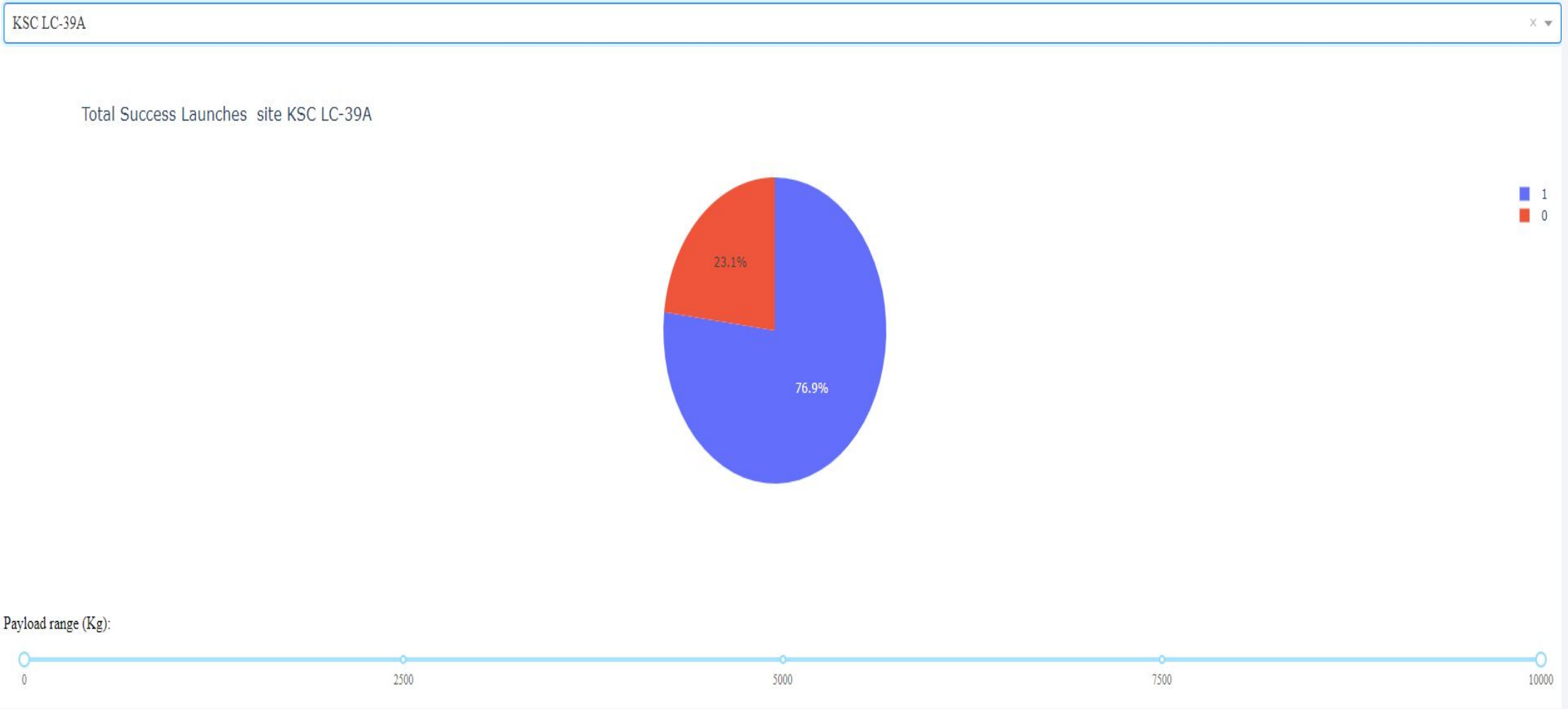
Successful launches per site



Rank success per launch site.

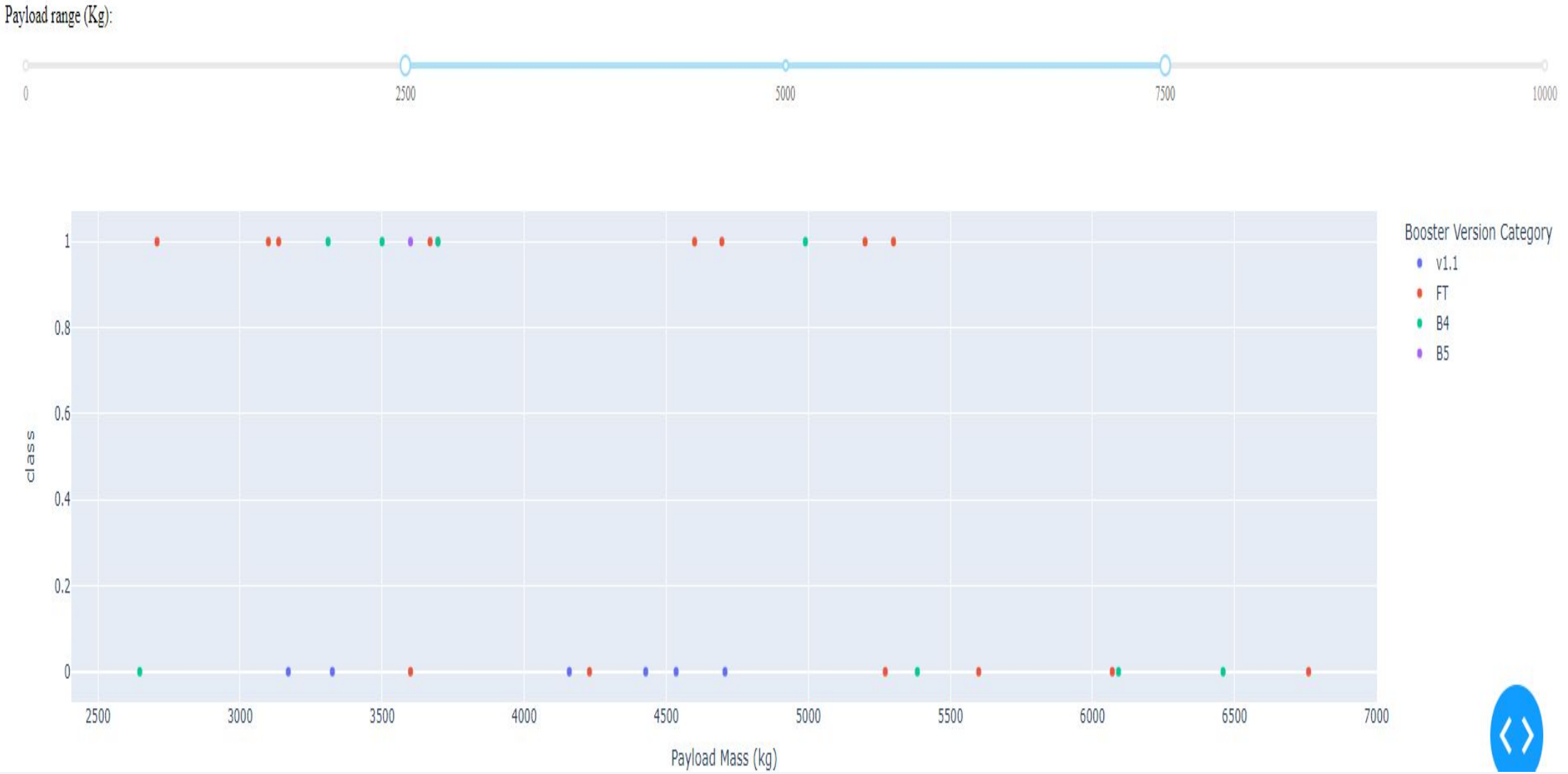
1- KSC LC-39A (41.4%)
2-CCAFS LC -40(29.2%)
3-VAFB SLC -4E(16.7%)
4-CCAFS SLC -40(12.5%)

Launch site with highest sucess



- Success :77 %
- Unsuccess : 23.1%

Payload vs Launch Outcome



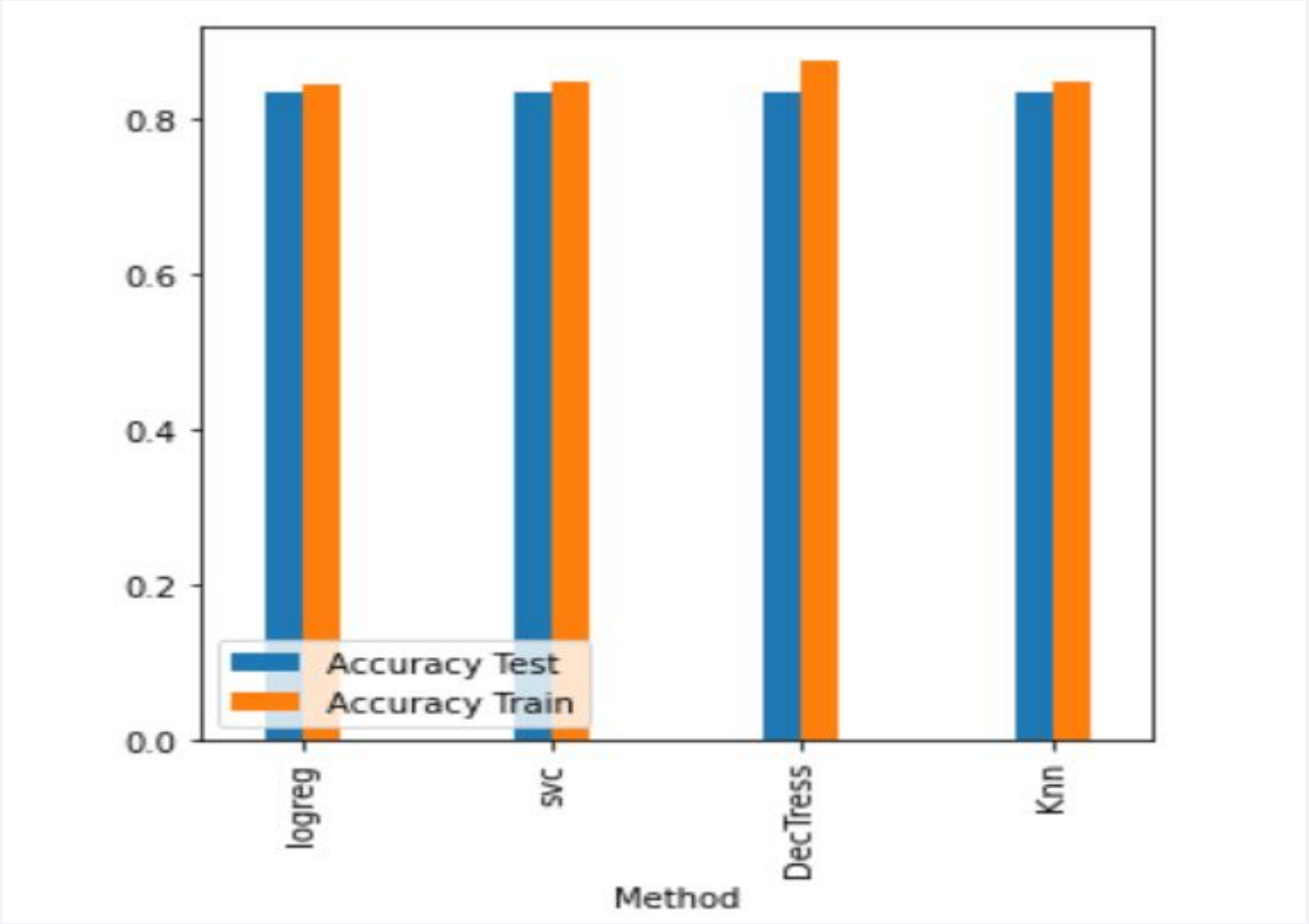
Payload mass: 2.5 tons to 7.5 tons.

Highest success with FT boosters version ,followed by B4.

Section 5

Predictive Analysis (Classification)

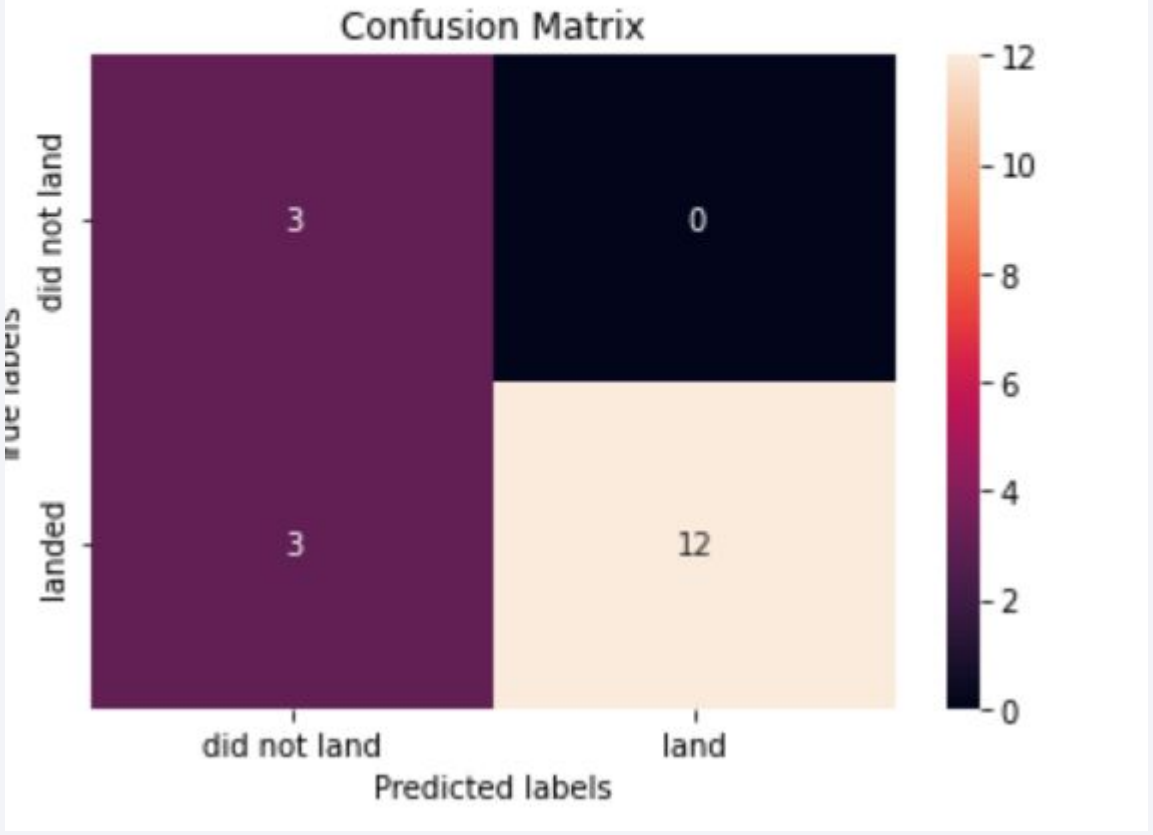
Classification Accuracy



- WE can verify that the best method for training model is the Decision tree method while at same time is the worst for testing model.The other 3 models are very similar in both test and train data.

Confusion Matrix

```
yhat=tree_cv.predict(X_test)
plot_confusion_matrix(yhat,Y_test)
```



- 12 true positives and 3 false negatives
- 3 false positives
- 3 launches bad predicted in total of 18

Conclusions

- After analyze all the work done from the data collected,we can say that the best launch site was KSC LC 39A nearby coastlines ,relative to the weight of the payloads ,under 5 tons and above 10 tons the success of landing and reuse the 1st stage is higher,success rate has grown over time since 2013 due the quality of the spaceships and the quality and treatment of the data,we can use the 4 machine learning prediction methods due to their proximity of accuracy ,but decision tree one has slightly higher accuracy in train test,so must be better for future launches to our company Space Y so we can save a lot of money in rocket launches field.

Appendix

- In the folium maps lab the images are attach in png file,because when uploaded on github they don't appear in the jupyter notebook file.
- In the dash plotly dash the images are attach in png file ,when can only upload the code of the html dashboard on github
- From slide 24 to 33 the queries were obtained via SQL programming language.

Thank you!