IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hadis Abarghouyi
4/18/2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; while other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. In this project we focus on the prediction of success landing of first stage.

The used methodology of this project cosists of below steps:

1. Data Collection via API and webscraping
2. Applying EDA methods with SQL and Data Visualization
3. Providing Interactive MAPS with Folium and Dash-Plotly.
4. Do predictive analysis and find the best models between (Decision Tree, KNN, LogisticRegression, SVM)

Results summary:

- Payload mass, Flight number, Booster Version and Orbit are the affective features on the success rate.
- VLEO has complete success rate for flight number more than 80. HEO has 100% success rate.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- For the payload less than 5000, the KSC LC 39A has 100% success rate.
- KSC LC-39A is the site with the best success rate.
- ES-L1, GEO, HEO and SSO have the most success rate.
- Among all models, Decision Tree had the best accuracy (88.88%) and matrix confusion.

Introduction

- Project background and context

we will predict 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 because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we 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

- What are the main characteristics of a successful or failed landing ?
- What are the effects of each relationship of the rocket features on the success or failure of a landing ?
- What are the conditions which will allow SpaceX to achieve the best landing success rate ?

Section 1

Methodology

Methodology

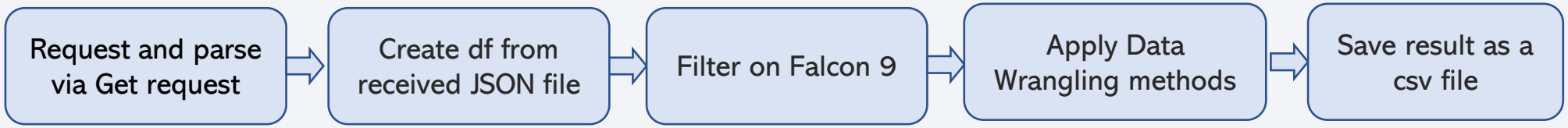
Executive Summary

- Data collection methodology:
  - Request to the SpaceX API
  - Web Scrapting to collect data from 'https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches '.
- Data wrangling
  - Using python for preprocessing steps such as , checking missed values, dropping unnecessary columns, applying one-hot encoding for categorical data and normalize data.
- Exploratory data analysis (EDA)
  - EDA has been done via SQL and data visualization methods.
- Interactive visual analytics
  - Folium and Plotly Dash has been used to perform interactive analysis
- Perform predictive analysis using classification models
  - Python sklearn library has been used to check different classification methods and tune their parameters with GridSearchCV method. Their accuracy and confusion matrix has been used to evaluate the result.

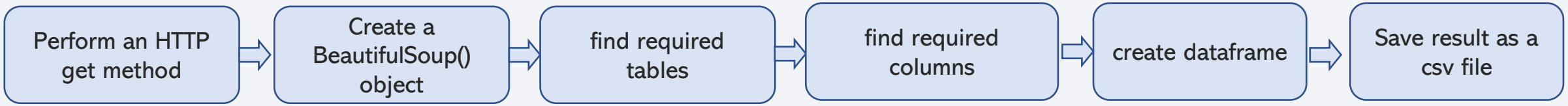
Data Collection

Data has been collected via below two method

1. Request to the SpaceX API



2. Web Scraping to collect data from 'https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches '.



Data Collection – SpaceX API

Request and parse via Get request

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

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

Create df from received JSON file

```
response.json()
data=pd.json_normalize(response.json())
```

Filter on Falcon 9

```
data_falcon9=df[df['BoosterVersion']=='Falcon 9']
```

Apply Data Wrangling methods

```
data_falcon9.isnull().sum()
```

```
# Calculate the mean value of PayloadMass column
PayloadMass_mean=data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass']=data_falcon9['PayloadMass'].replace(np.nan, PayloadMass_mean)
data_falcon9['PayloadMass'].isnull().sum()
```

Save result as a csv file

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Github link for SpaceX API

Data Collection - Scraping

Perform an HTTP get method

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

Create a BeautifulSoup() object

```
soup=BeautifulSoup(response.text,'html.parser')
```

find required tables

```
html_tables=soup.find_all('table')
```

find required columns

```
['Flight No.',
 'Date and time ( )',
 'Launch site',
 'Payload',
 'Payload mass',
 'Orbit',
 'Customer',
 'Launch outcome']
```

create dataframe

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

Save result as a csv file

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

Github link for Scraping

Data Wrangling

Github link for Data Wrangling

- In the dataset, there are several cases where the booster did not land successfully.
  True Ocean, True RTLS, True ASDS means the mission has been successful.
  False Ocean, False RTLS, False ASDS means the mission was a failure.
- We need to transform string variables into categorical variables where 1 and 0 means the mission has been successful and a failure, respectively.

```
df.to_csv("dataset_part_2.csv", index=False
```

```
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40     55
KSC LC 39A       22
VAFB SLC 4E      13
Name: LaunchSite, dtype: int64
```

```
# landing_outcomes = values on Outcome column
landing_outcomes=df['Outcome'].value_counts()
```

Check nulls → Number of launches per site → Number and occurrence of each orbit → Number and occurrence of mission outcome per orbit type → Create a landing outcome label → Save result as a csv file

```
df.isnull().sum()/df.shape[0]*100
```

```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

```
landing_class=~df['Outcome'].isin(bad_outcomes)
#landing_class=landing_class.multiply(landing_cla
landing_class=list(map(int,landing_class))#=landi
```

EDA with Data Visualization

Github link for EDA Visualization

The scatter plots used to find the relation between two variables and their effects on output.

- ✓ As the flight number or Payload increases , the first stage is more likely to land successfully.
- ✓ Different launch sites have different success rates.
  - CAFS LC-40: 60%
  - KSC LC-39A or VAFB SLC 4E :70%
- ✓ There is no relation between some orbits such as GTO and the number of flights. In large flight numbers the VLEO orbit has about 86% success rate.
- ✓ With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

Bar plot is used to find the relation between a numeric (Class) and a categorical (Orbit) data.

- ✓ ES-L1 , GEO and HEO have the highest success rate.

Line plot is used to check time series trend.

- ✓ The success rate since 2013 kept increasing till 2020

Finally, we used the EDA results for feature engineering.

EDA with SQL

Github link for EDA SQL

SQL commandas are used to find below records:

- Find unique Launch_Site lists.
- 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 successful landing outcome in ground pad was achieved.
- 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

Github link for Interactive Map with Folium

Folium map is used to perform more interactive visual analytics in the attached link and below details.

Using different objects on MAP gives us better understanding of data.

Details:

- a red circle at NASA Johnson Space Center's coordinate with a popup label/icon showing its name (folium.circle , folium.map.Maker)
- Add a circle per site to clarify the site's locations. (folium.circle , folium.map.Maker)
- Mark the success/failed launches for each site on MAP with Green/Red color (MarkerCluster())
- MarkerCluster is a good choice when many markers having the same coordinate
- Lines are used to show the distance between a site and any railway, highway, coastline, etc. (folium.PolyLines)

Build a Dashboard with Plotly Dash

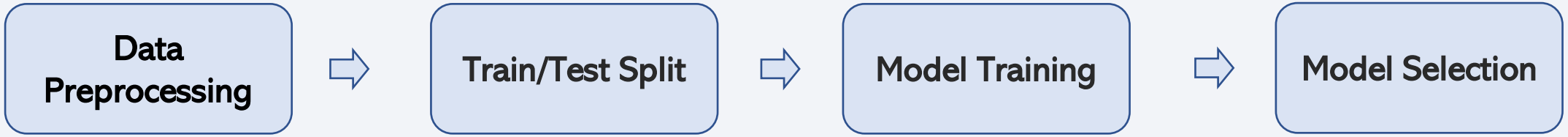
Github link for Plotly Dash

Some charts consist of pie chart, rangeslider, scatter plot and dropdown components have been added to an interactive Plotly dashboard for user investigations.

- A dropdown helps user to filter on specific site or compare the success rate of different launch sites.
- A pie chart has been selected to compare success rate of filtered sites.
- A rangeslider helps user to select desired payload weights
- A scatter plot shows the relation between payload and success rate per different booster version.

Predictive Analysis (Classification)

Github link for Prediction



Change data type
Normalization

Logistic Regression with accuracy: 83.33%
SVM model with accuracy: 83.33%
Decision Tree with accuracy: 88.88%
KNN with accuracy: 83.33%

Decision Tree with accuracy: 88.88%

Results

Exploratory data analysis results

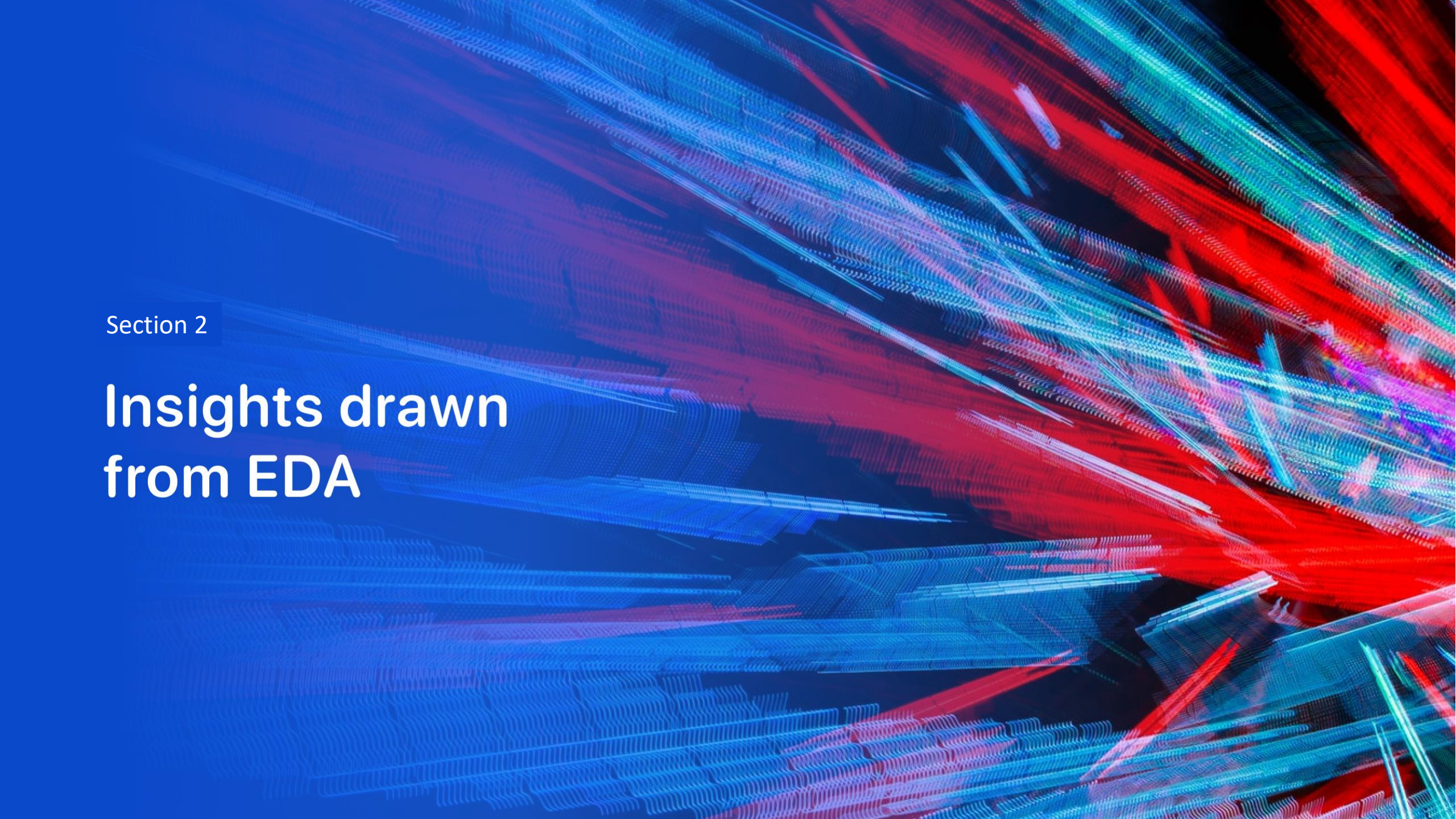
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- The more flight number, the more success rate.
- for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- ES-L1, GEO, HEO and SSO have the most success rate.
- The success rate since 2013 kept increasing till 2020
- the first successful landing outcome on ground pad is 22/12/2015.

Interactive analytics demo in screenshots

- All launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast.
- The most outcomes related to the CCAFC – LC 40.

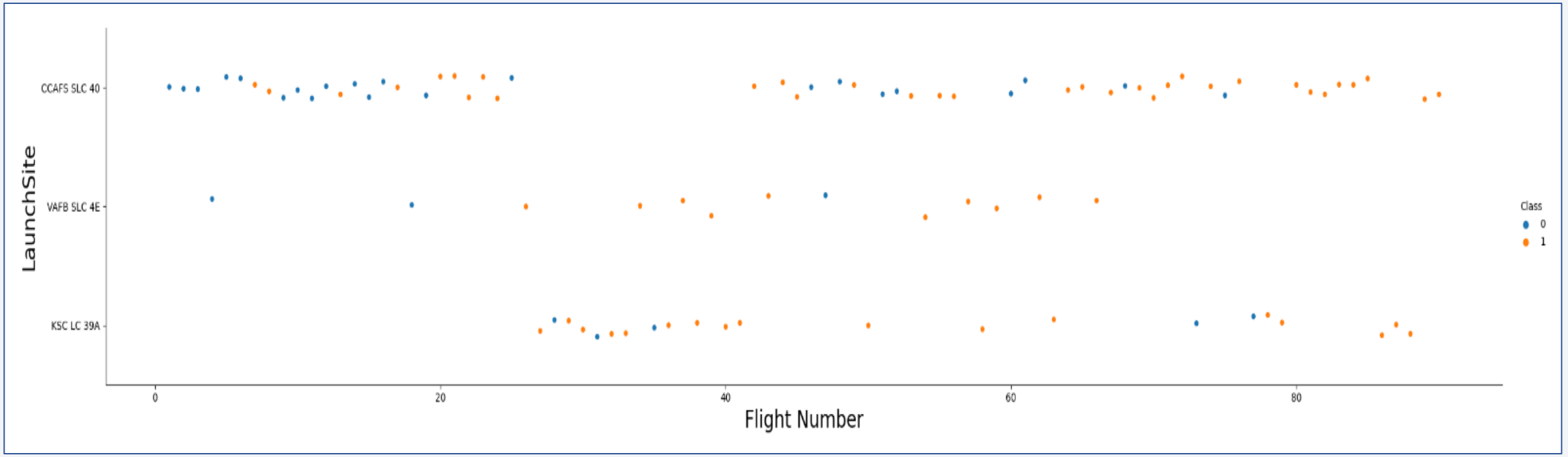
Predictive analysis results

- Checking the accuracy of different classification methods clarifies the Decision Tree is the best choice for prediction with 88.88% accuracy .


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

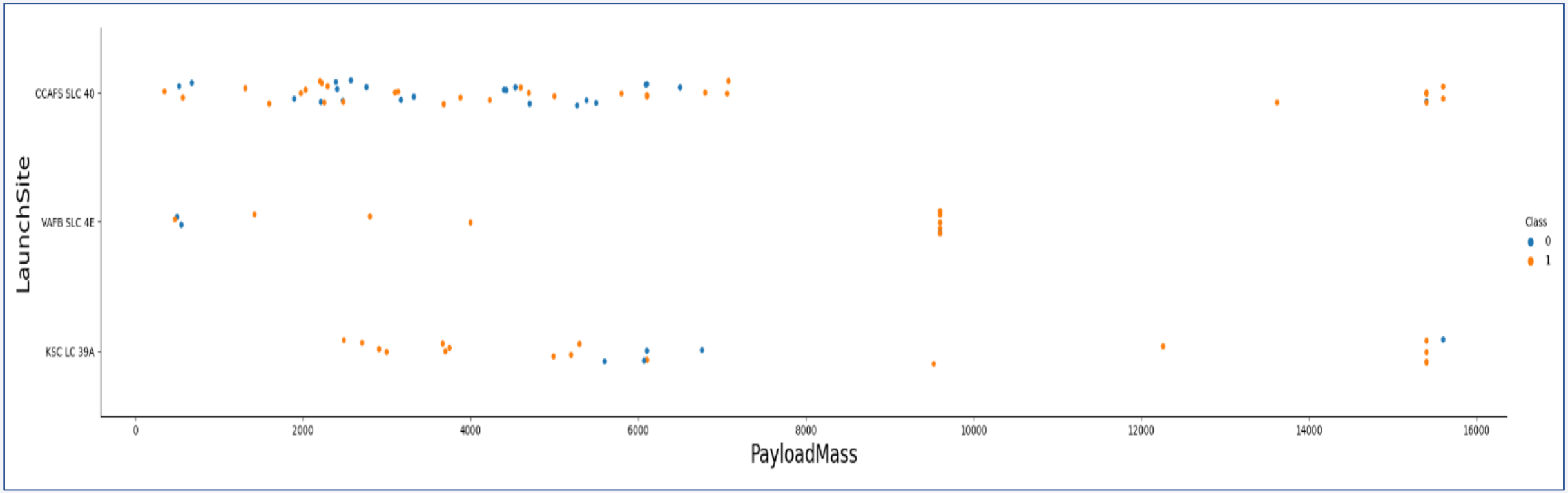
- CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- There were 100% success rate for flight numbers more than 80.
- The more flight number, the more success rate.



Payload vs. Launch Site

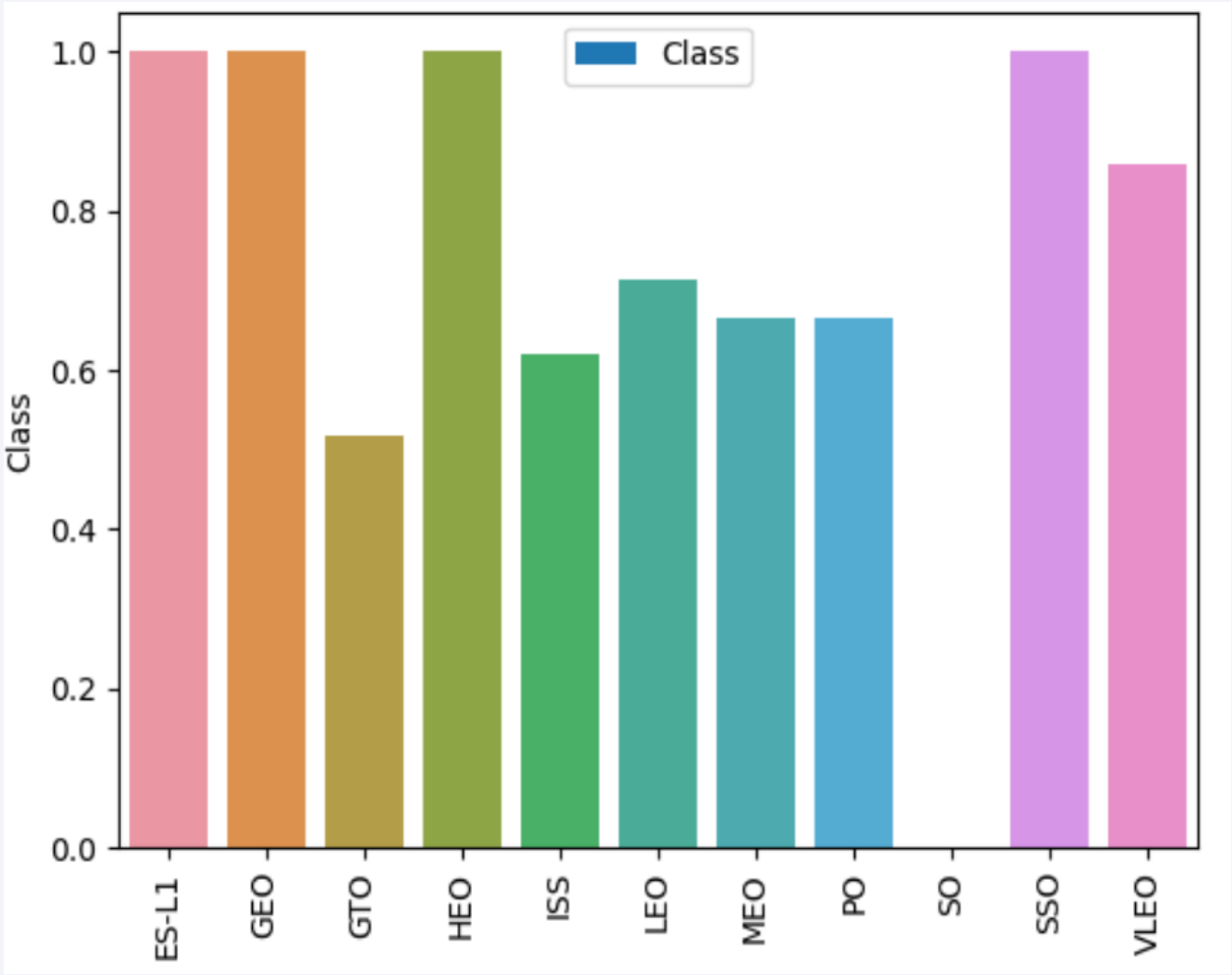
For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

For the payload less than 5000, the KSC LC 39A has 100% success rate.



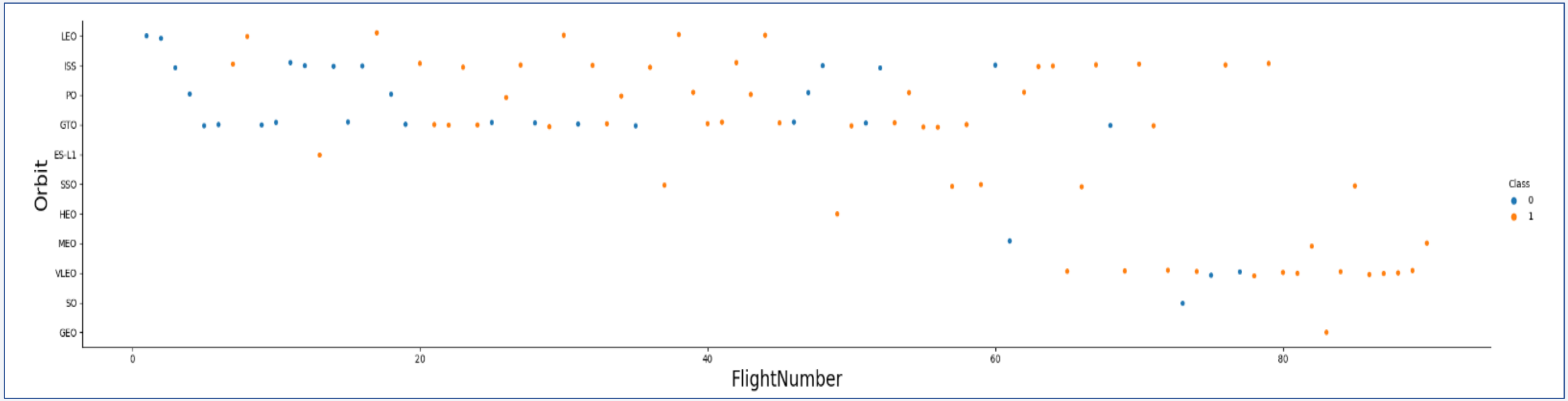
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have the most success rate.
- SO doesn't have any success.



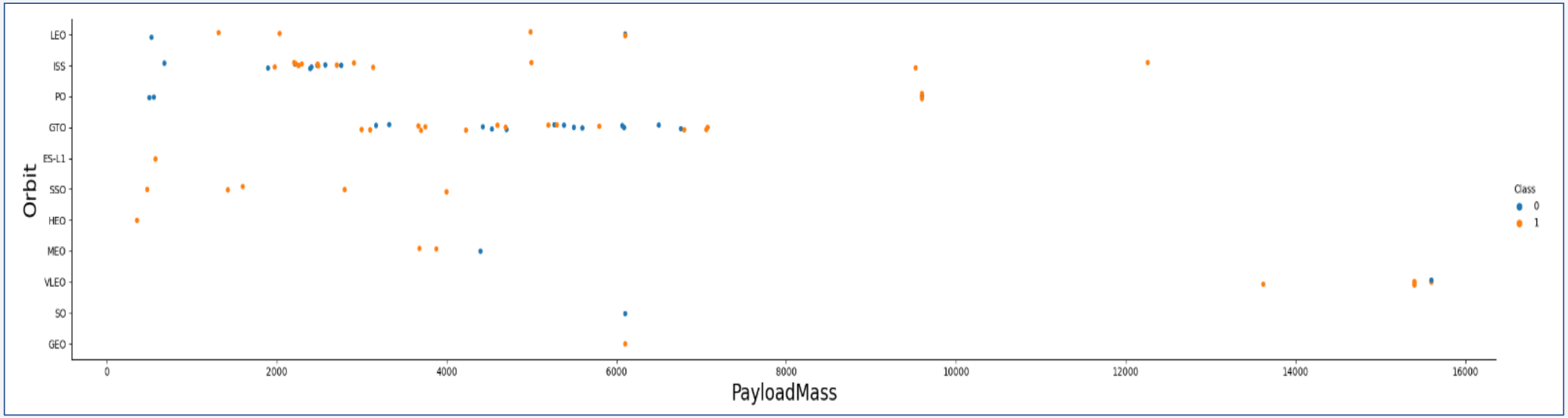
Flight Number vs. Orbit Type

- The LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit.
- VLEO has complete success rate for flight number more than 80.
- HEO has 100% success rate.



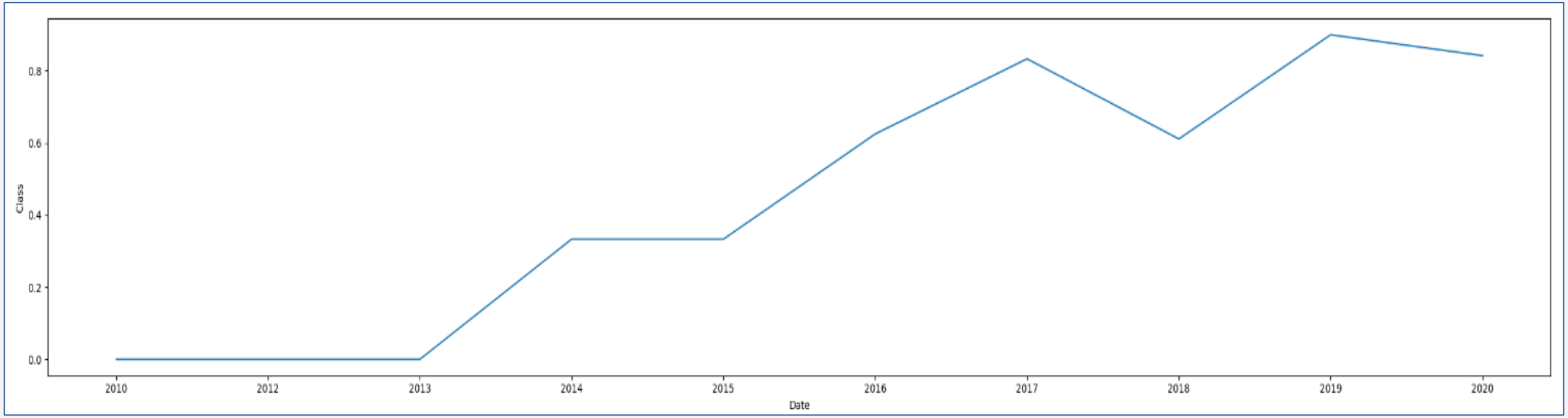
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- SSO, ES-L1, HEO have 100% success rate.
- For ISS the more the payload, the more success rate.



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020



All Launch Site Names

SQL query is used to find the name of launch-site as following:

```
%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

Launch Site Names Begin with 'CCA'

SQL used to find 5 records where launch sites begin with `CCA`:

```
%%sql
select Launch_Site from SPACEXTBL
where Launch_Site like 'CCA%' limit 5
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

The total payload carried by boosters from NASA is:

```
%%sql
select sum(PAYLOAD_MASS__KG_) from SPACEXTBL
where Customer='NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
%%sql
select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL
where Booster_Version='F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad is:

```
%%sql
select Date,substr(Date,1,2) As Day,substr(Date,4,2) As Month,substr(Date,7,4) As Year, "Landing _Outcome" from SPACEXTBL
where "Landing _Outcome" like "Success%"
order by Year, Month, Day limit 1
```

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

Date	Day	Month	Year	Landing _Outcome
22-12-2015	22	12	2015	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

```
%%sql
select Payload,Booster_Version,"Landing _Outcome",PAYLOAD_MASS__KG_ from SPACEXTBL
where "Landing _Outcome"= 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_<6000
```

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

Payload	Booster_Version	Landing _Outcome	PAYLOAD_MASS__KG_
JCSAT-14	F9 FT B1022	Success (drone ship)	4696
JCSAT-16	F9 FT B1026	Success (drone ship)	4600
SES-10	F9 FT B1021.2	Success (drone ship)	5300
SES-11 / EchoStar 105	F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes is:

```
%%sql
select substr(Mission_Outcome,1,7),count(Mission_Outcome) from SPACEXTBL
group by substr(Mission_Outcome,1,7)
```

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

substr(Mission_Outcome,1,7)	count(Mission_Outcome)
Failure	1
Success	100

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass are as following table:

```
%%sql
select "Booster_Version" , "PAYLOAD_MASS__KG_"from SPACEXTBL
where "PAYLOAD_MASS__KG_" in (select MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

```
 * 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

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are:

```
%%sql
select Date, substr(Date, 4, 2) as month , substr(Date,7,4) as year, "Landing _Outcome",Booster_Version,Launch_Site from SPACEXTBL
where year='2015' and "Landing _Outcome"='Failure (drone ship)'
```

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

Date	month	year	Landing _Outcome	Booster_Version	Launch_Site
10-01-2015	01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
14-04-2015	04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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 are as follows:

```
%%sql
SELECT "Landing _Outcome" ,count(*) FROM SPACEXTBL
where CAST(substr(Date,7,4) AS INTEGER) BETWEEN 2011 and 2016
or (CAST(substr(Date,4,2) AS INTEGER)>05 and CAST(substr(Date,7,4) AS INTEGER)=2010 )
or (CAST(substr(Date,4,2) AS INTEGER)=06 and CAST(substr(Date,7,4) AS INTEGER)=2010 and (CAST(substr(Date,1,2) AS INTEGER)>04))
or (CAST(substr(Date,4,2) AS INTEGER)<03 and CAST(substr(Date,7,4) AS INTEGER)=2017 )
or (CAST(substr(Date,4,2) AS INTEGER)=03 and CAST(substr(Date,7,4) AS INTEGER)=2017 and (CAST(substr(Date,1,2) AS INTEGER)<20))
Group by "Landing _Outcome" having "Landing _Outcome" like "%Success%"
order by Count(*) DESC
```

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

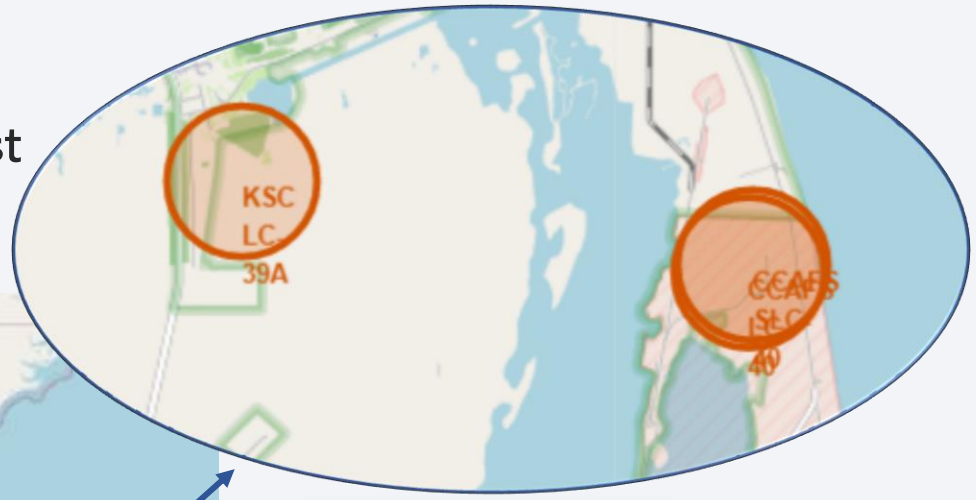
Landing _Outcome	count(*)
Success (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

Launch site locations

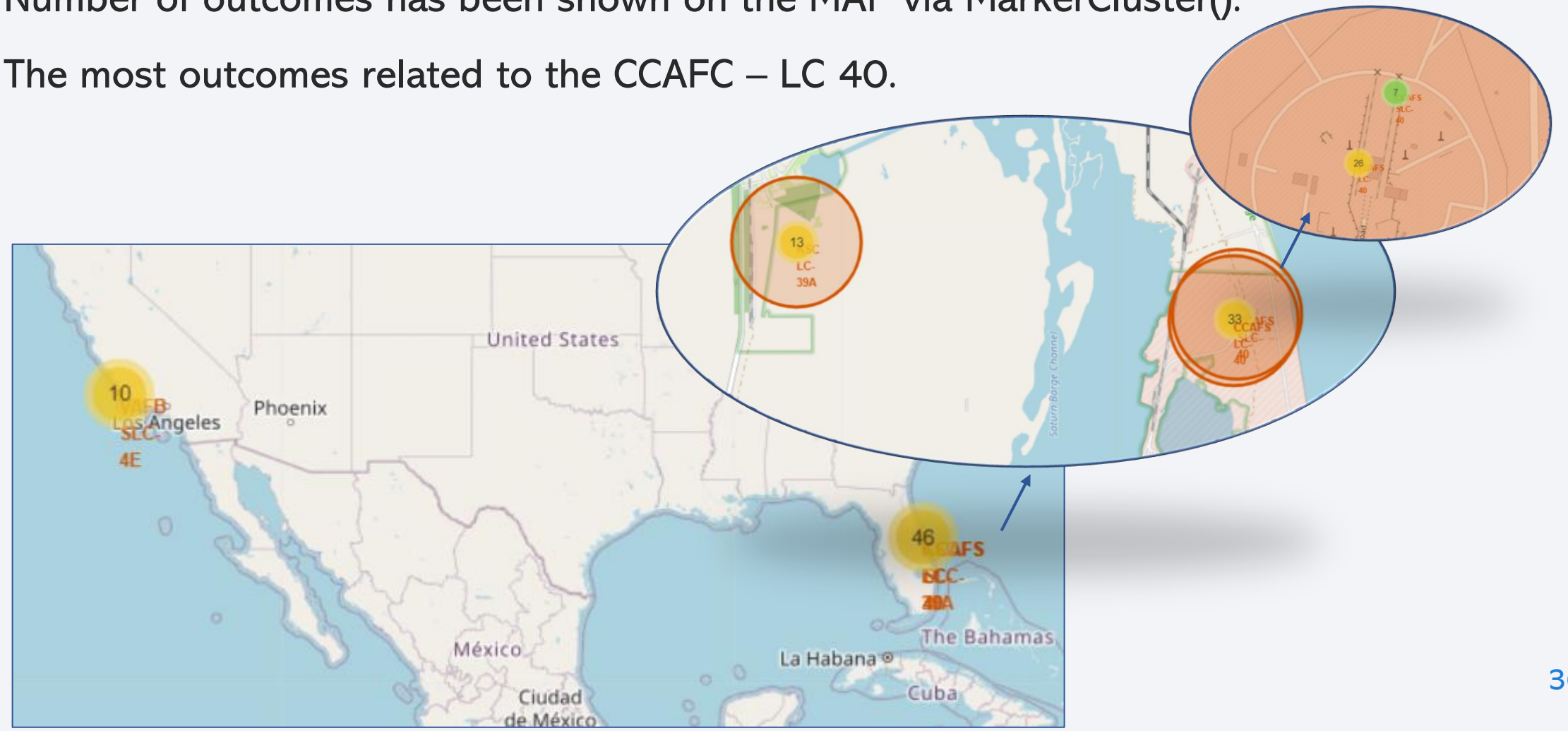
- All launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast.
- One site is in the West and three others are in the east and near eachother.



Cluster of outcomes on MAP

Number of outcomes has been shown on the MAP via MarkerCluster().

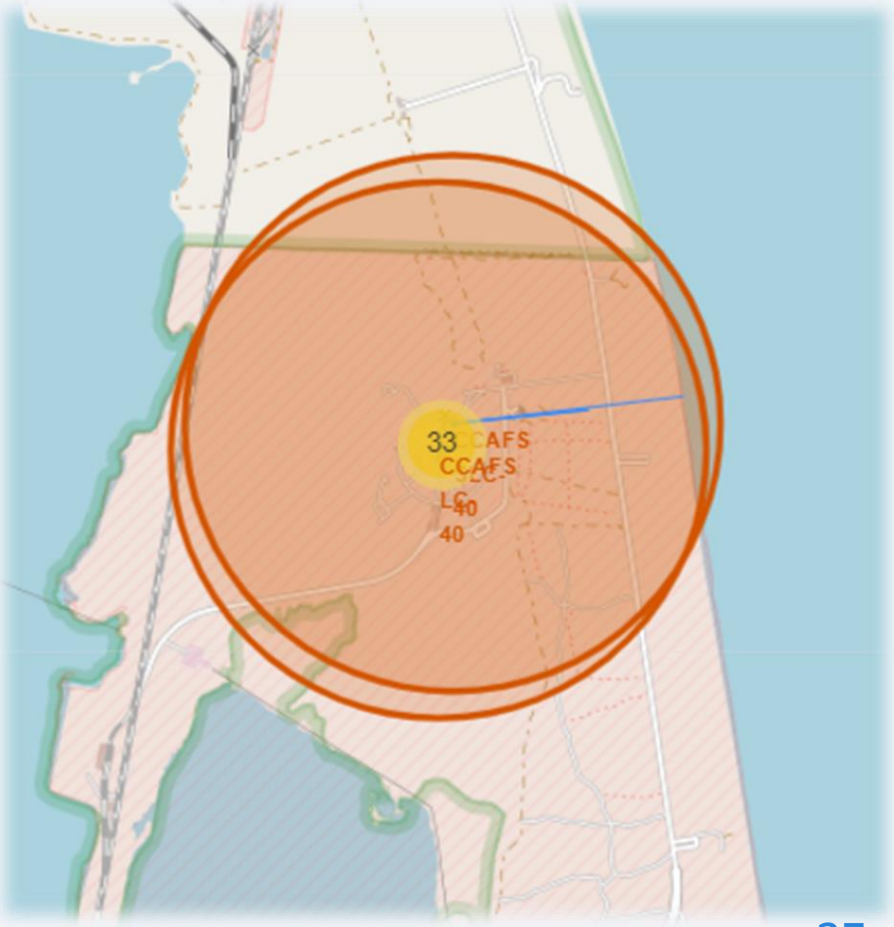
The most outcomes related to the CCAFC – LC 40.

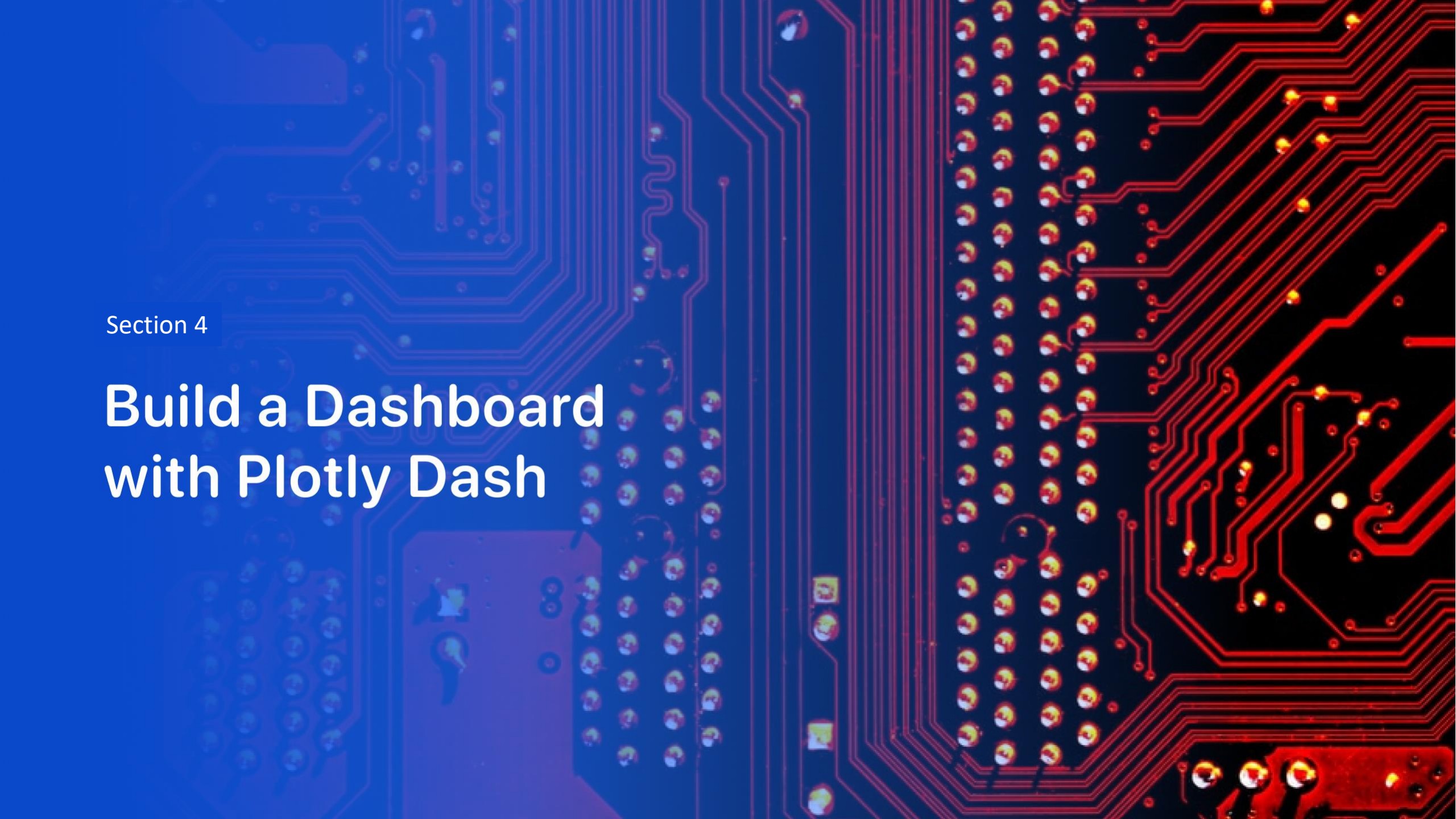


Site distance to coastline

The distance of different sites to different objects in MAP is calculatable.

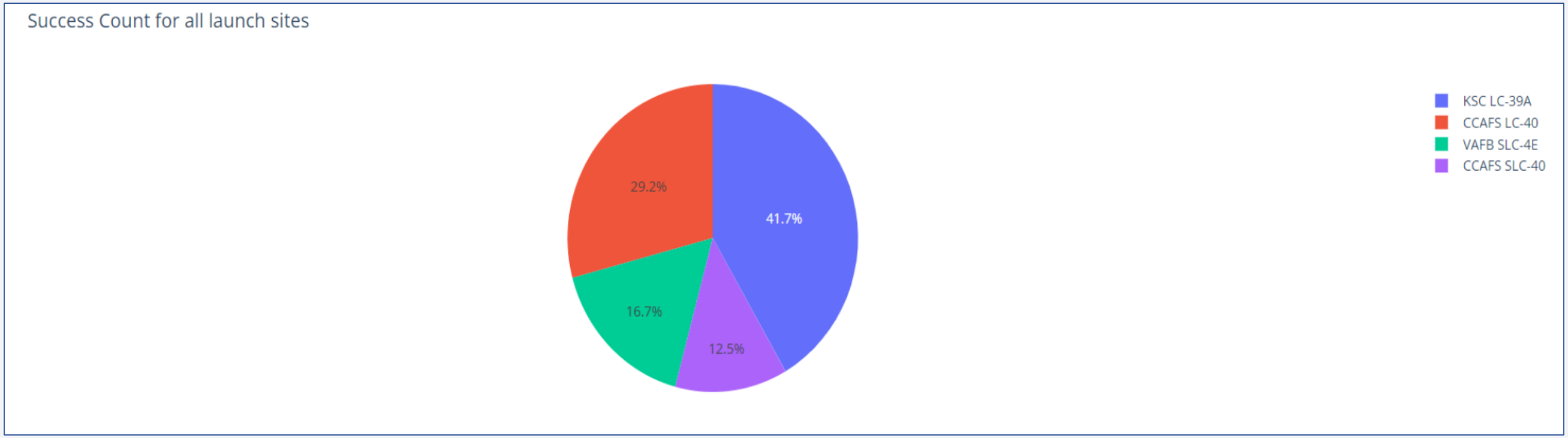
For example, the distance of site to coastline is shown with a polyline and is 0.857 km.




Section 4
Build a Dashboard with Plotly Dash

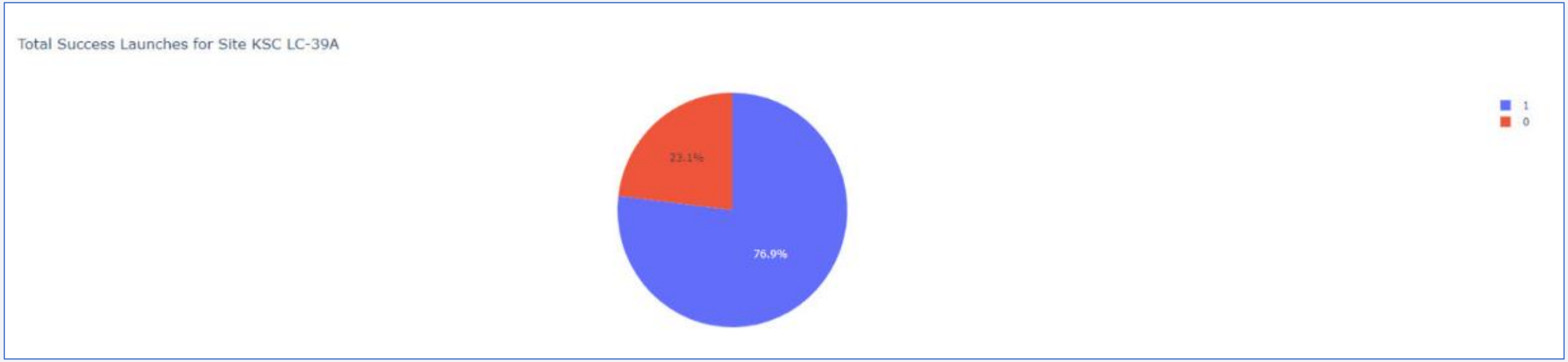
Success rate per launch - site

As shown in the pie-chart, KSC LC-39A and CCAFS SLC-40 have the most and least success rates with 41.7% and 12.5% success rate, respectively.



The launch site with highest success rate

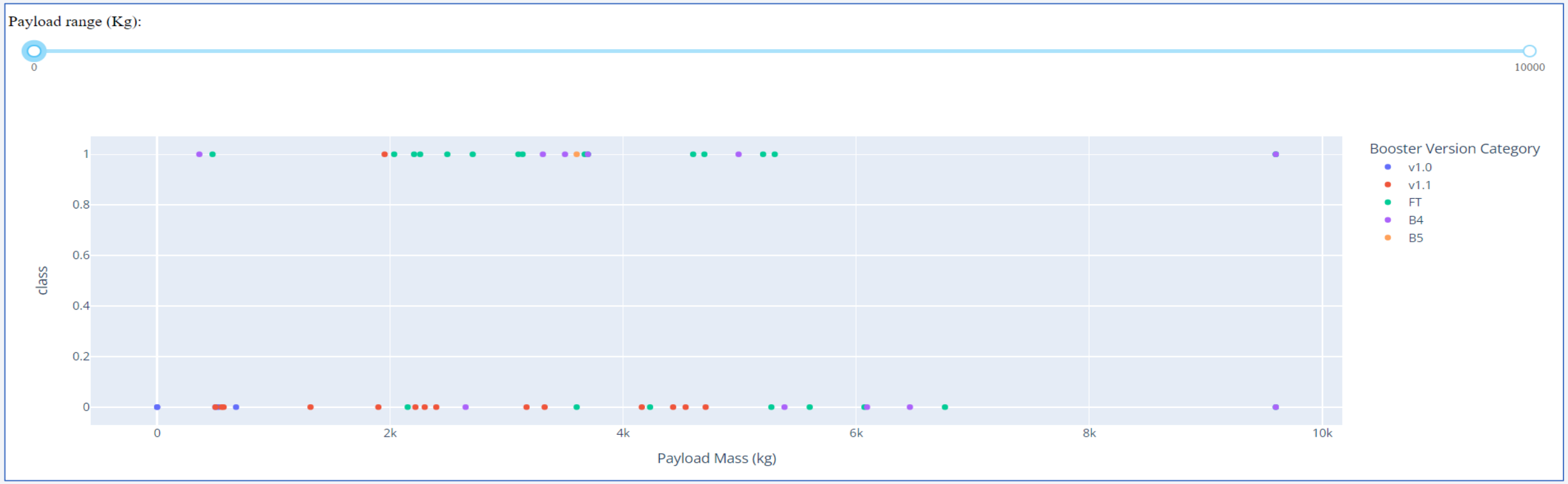
- KSC LC-39A is the site with the best success rate among different sites.
- It has 76.9% success rate and 23.1% failure rate.



Payload vs Launch Outcome

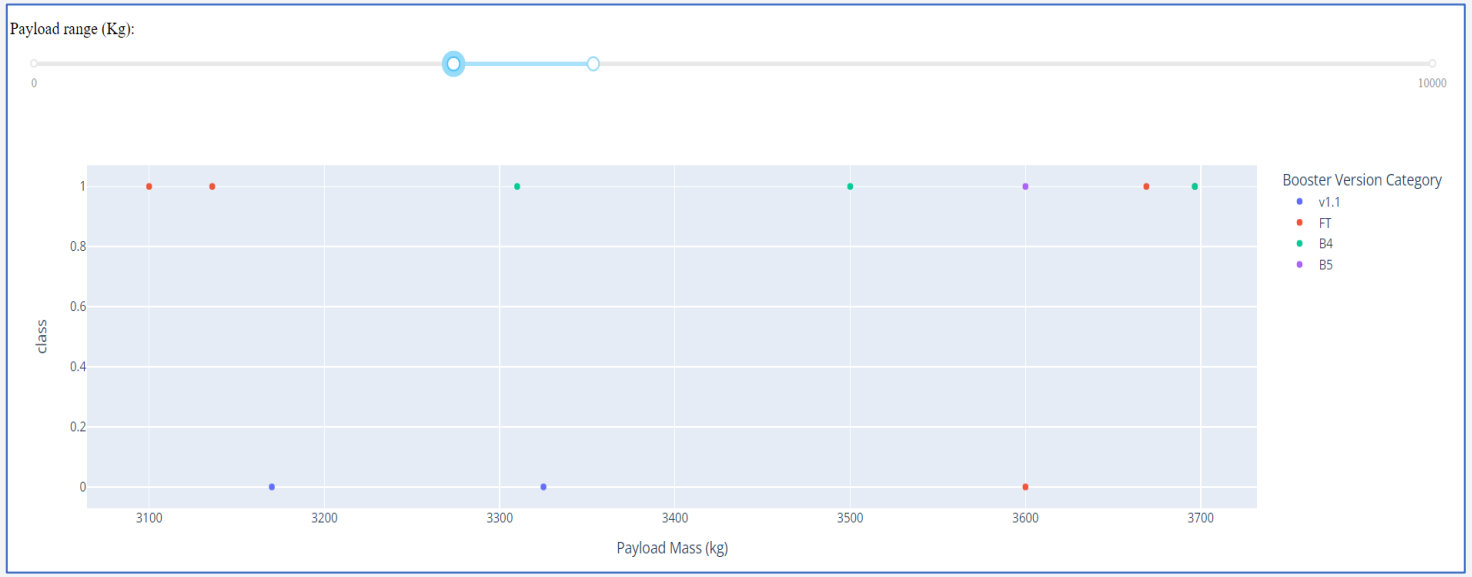
The plot clarifies the affect of payload and booster versions on the outcome.

As shown FT booster version has the most success rate among different booster versions.

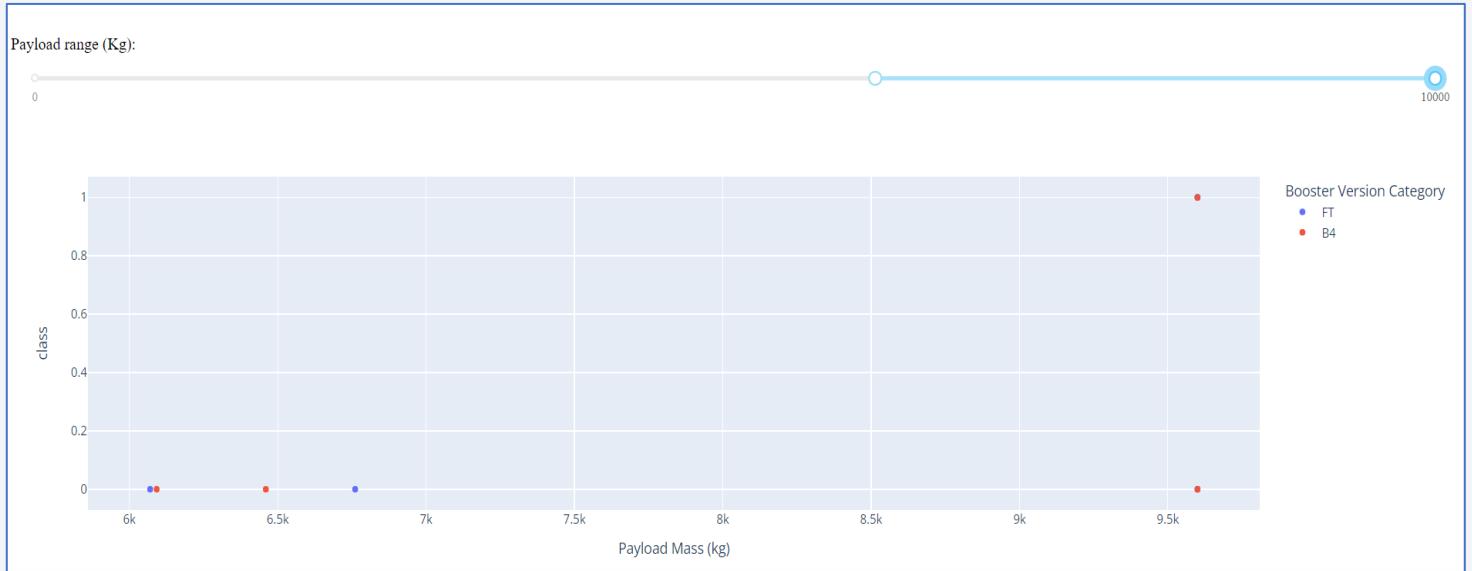


Payload vs Launch Outcome

The most success rate occurs among payload 3.1kg-3.7kg.



The least success rate occurs among the payload 6kg-9.5kg.

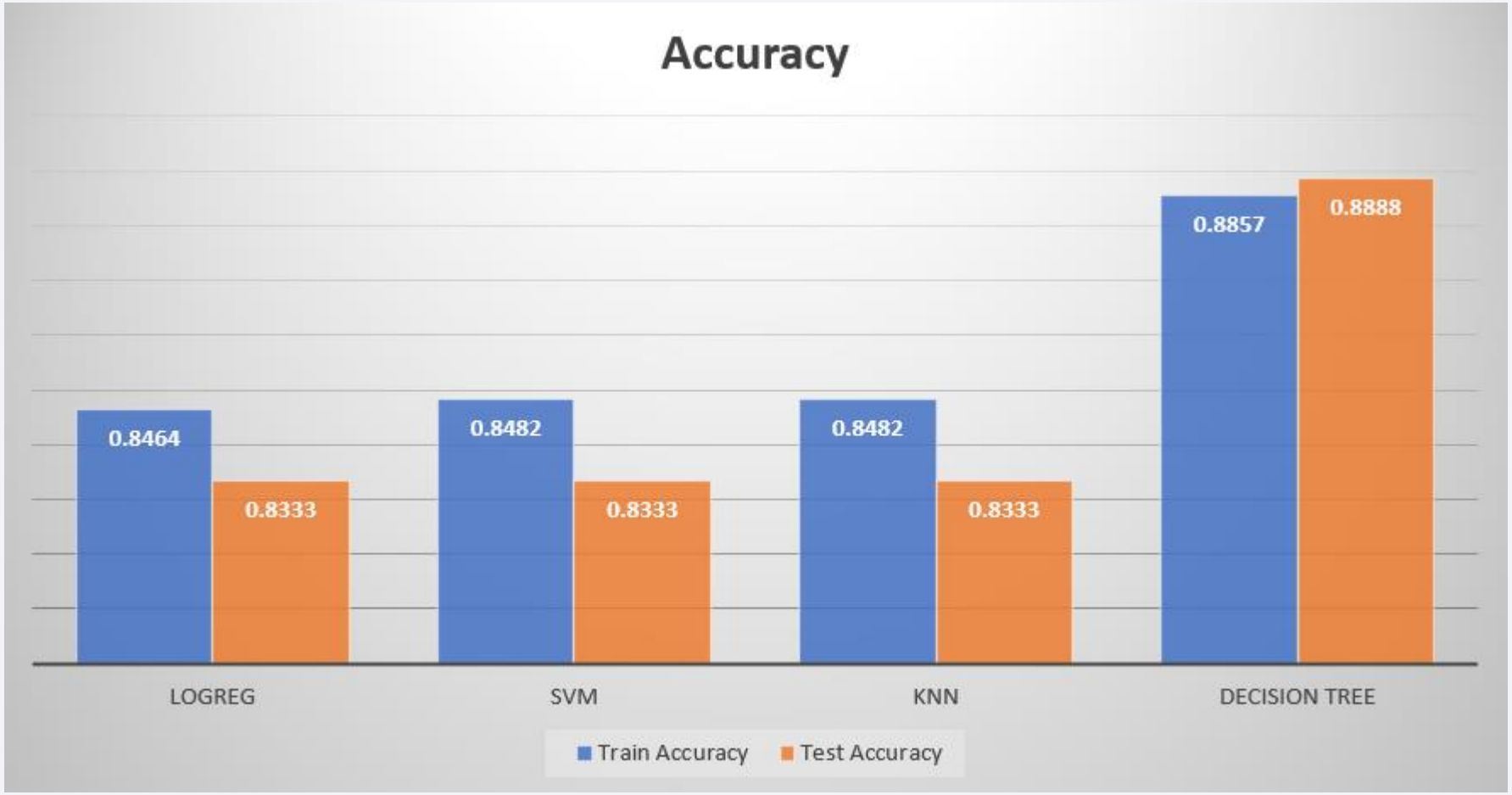


Section 5

Predictive Analysis (Classification)

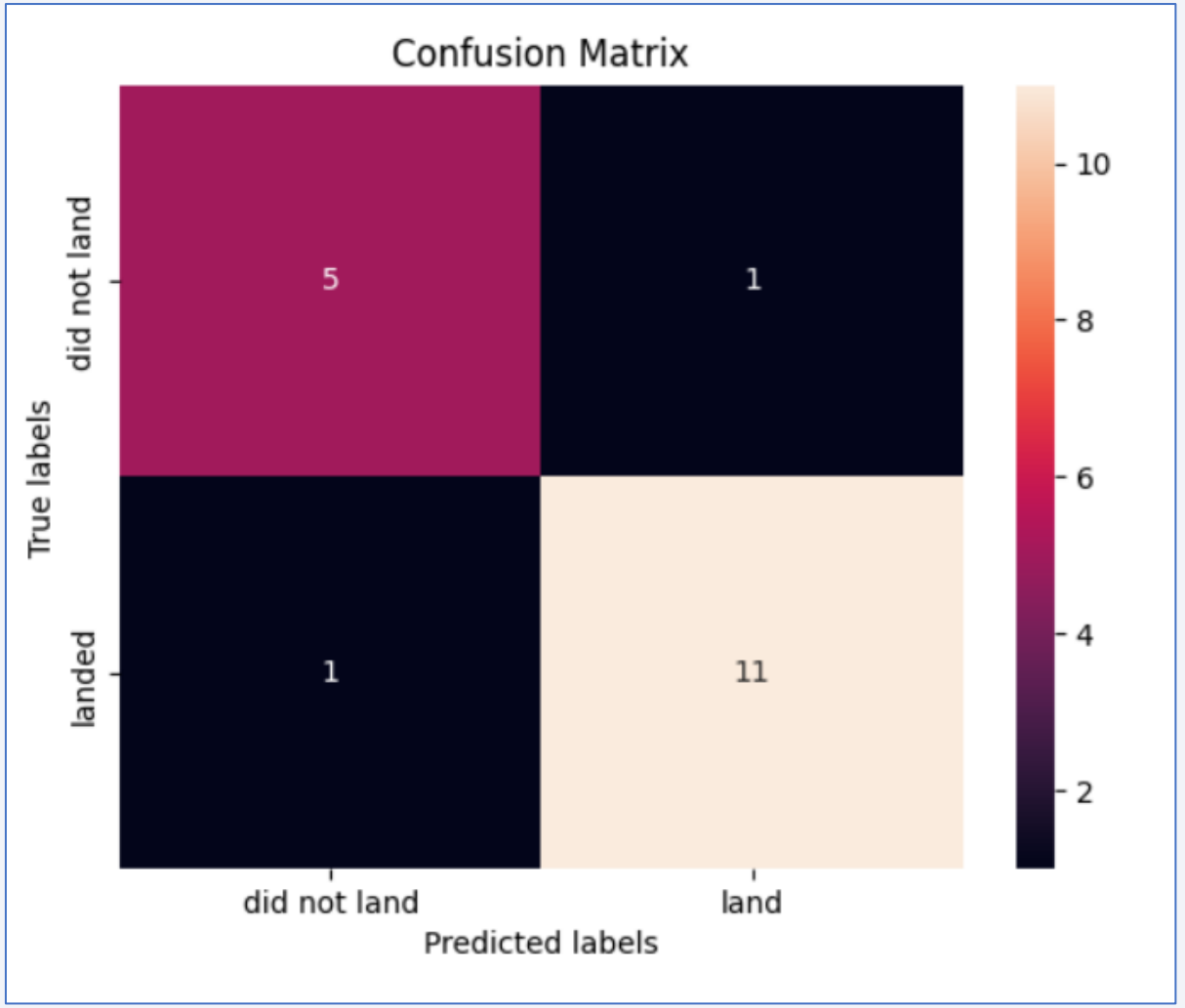
Classification Accuracy

Checking the accuracy of different classification methods clarifies the Decision Tree is the best choice for prediction.



Confusion Matrix

Among different Confusion Matrix, the decision tree confusion matrix has the best False Positive value.



Conclusions

Based on feature engineering, the most impacted features on predicting a successful landing are Payload mass, Flight number, Booster Version and Orbit.

With having these feature and using the DecisionTree model with 88.88% accuracy, we can predict the success or failure landing.

These results are also obtained from the analysis:

- VLEO has complete success rate for flight number more than 80. HEO has 100% success rate.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- For the payload less than 5000, the KSC LC 39A has 100% success rate.
- KSC LC-39A is the site with the best success rate.
- ES-L1, GEO, HEO and SSO have the most success rate.

Appendix

To find the whole project scripts refer to :

https://github.com/HadisAB/Applied-Data-Science-Capstone/tree/main

Thank you!