Winning Space Race with Data Science

Sagadiyev Radmir
November 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection API, Web Scraping
  - Exploratory Data Analysis (EDA). Data Visualization
  - EDA with SQL
  - Interactive Map with Folium
  - Dashboards with Plotly Dash
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive maps and dashboard
  - Predictive results

Introduction

- Project background and context

  The project is about identifying probability of Falcon 9 successful landing.

  The cost of SpaceX Falcon 9 rocket launch is 62 million dollars. While rivals launch cost is about 165 million dollars each, because SpaceX can reuse the first stage. By determining if the stage will land, we can determine the cost of a launch.

- Problems you want to find answers

  Identifying parameters that influence success of landing.

  Identifying conditions which will allow SpaceX to achieve the best landing success rate.

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection – SpaceX API

Calls	Code cell
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)
Decode the response content as a Json using .json()	data = pd.json_normalize(response.json())
Apply getBoosterVersion function method to get the booster version	getBoosterVersion(data)
Create a Pandas data frame from the dictionary launch_dict.	df = pd.DataFrame(launch_dict)
Filter the dataframe to only include Falcon 9 launches	df['BoosterVersion']!='Falcon 1' data_falcon9 = df[(df['BoosterVersion']!='Falcon 1')]

Data Collection - Scraping

Calls	Code cell
Scraping data with the following URL	spacex_url="https://api.spacexdata.com/v4/launches/past"
Using requests.get() method with the provided static_url Assigning the response to a object	page=requests.get(static_url) page.status_code
Using BeautifulSoup() to create a BeautifulSoup object from a response text content	soup = BeautifulSoup(page.content)
Using the find_all function in the BeautifulSoup object, with element type `table`	html_tables = soup.find_all('table')
Creating an empty dictionary with keys from the extracted column names	launch_dict= dict.fromkeys(column_names) ……
Creating a dataframe using launch_dict	df=pd.DataFrame(launch_dict)

Data Wrangling

Calls	Code cell
Identify and calculate the percentage of the missing values in each attribute	df.isnull().sum()/df.count()*100
Identify which columns are numerical and categorical	df.dtypes
Use the method value_counts() on the column LaunchSite to determine the number of launches on each site:	df['LaunchSite'].value_counts()
Use the method .value_counts() to determine the number and occurrence of each orbit in the column Orbit	df['Orbit'].value_counts()
Use the method .value_counts() on the column Outcome to determine the number of landing_outcomes.Then assign it to a variable landing_outcomes.	landing_outcomes = df['Outcome'].value_counts()

Using the Outcome, create a list where the element is zero if the corresponding row in Outcome is in the set bad_outcome; otherwise, it's one. Then assign it to the variable landing_class:	landing_class = []
	for key,value in df["Outcome"].items():
	  if value in bad_outcomes:
	  landing_class.append(0)
	  else:
	  landing_class.append(1)

EDA with Data Visualization

Scatter Graphs	Bar Graph	Line Graph
Flight Number vs. Payload Mass Flight Number vs. Launch Site Payload vs. Launch Site Orbit vs. Flight Number Payload vs. Orbit Type Orbit vs. Payload Mass	Success rate vs. Orbit	Line Graph
Scatter plot shows correlation between variables.	Bar graph shows relationship between numeric and categoric variables.	Line graphs shows trends.

EDA with SQL

Calls	Code cell
load the SQL extension and establish a connection with the database	!pip install sqlalchemy==1.3.9, !pip install ibm_db_sa, !pip install ipython-sql, import sqlite3
Display the names of the unique launch sites in the space mission	select distinct Launch_Site from df
Display 5 records where launch sites begin with the string 'CCA'	select * from df where Launch_Site like 'CCA%' limit 10
Display the total payload mass carried by boosters launched by NASA (CRS)	select sum(PAYLOAD_MASS__KG_) as total from df where customer = 'NASA (CRS)'
Display average payload mass carried by booster version F9 v1.1	select avg(PAYLOAD_MASS__KG_) as avg from df where Booster_Version = 'F9 v1.1'
List the date when the first successful landing outcome in ground pad was acheived.	select min(Date) from df where [Landing _Outcome] = 'Success (drone ship)' or [Landing _Outcome] = 'Success' or [Landing _Outcome] = 'Success (ground pad)'
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery	select Booster_Version from df where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from df)
List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015	select [Landing _Outcome], Booster_Version, Launch_Site from df where [Landing _Outcome] = 'Failure (drone ship)' and Date like '%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	select [Landing _Outcome], count(*), row_number() over(order by count(*) desc) from df group by [Landing _Outcome] order by count(*) desc

Build an Interactive Map with Folium

Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

- Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas
- Red circle at NASA Johnson Space Center's coordinate with label showing its name(folium.Circle, folium.map.Marker).
- Green Markers to show successful landings. Red - unsuccessful landing.
- line between launch site to railway, highway, coastway, city to reveal distance folium.PolyLine.
- These objects are created in order to understand better the problem and the data. Launch sites, successful and unsuccessful landings can be seen.

Build a Dashboard with Plotly Dash

Summarize what plots/graphs and interactions you have added to a dashboard

Calls	Component
Dropdown choosing the launch site or all launch sites	dash_core_components.Dropdown
Pie chart shows help to understand success/failure for the launch site	plotly.express.pie
Rangeslider allows to select a payload mass in a fixed range	dash_core_components.RangeSlider
Scatter chart shows relationship between two variables - Success vs Payload Mass	plotly.express.scatter

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model

Data preparation

Load dataset

Normalize data

Split data into training and test sets.

Model preparation

Selection of machine learning algorithms

Set parameters for each algorithm to GridSearchCV

Training GridSearchModel models with training dataset

Model evaluation

Get best hyperparameters for each type of model

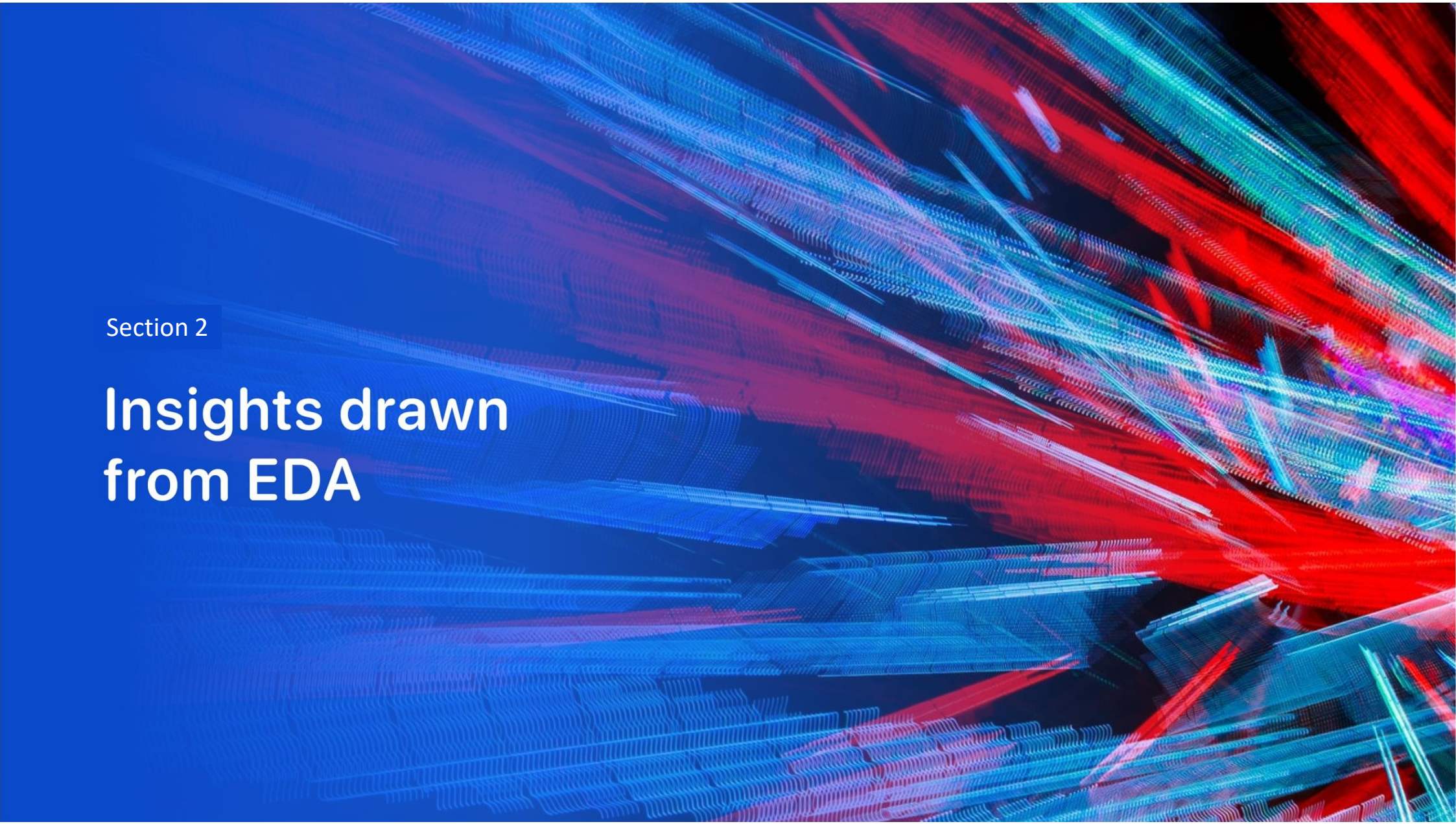
Compute accuracy for each model with test dataset

Plot Confusion Matrix

Model comparison

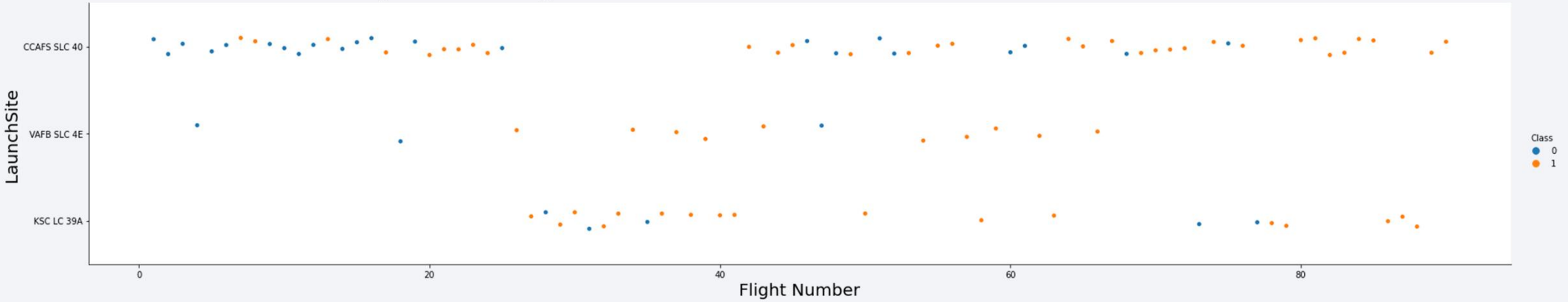
Comparison of models according to their accuracy

The model with the best accuracy will be chosen (see Notebook for result)


Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site

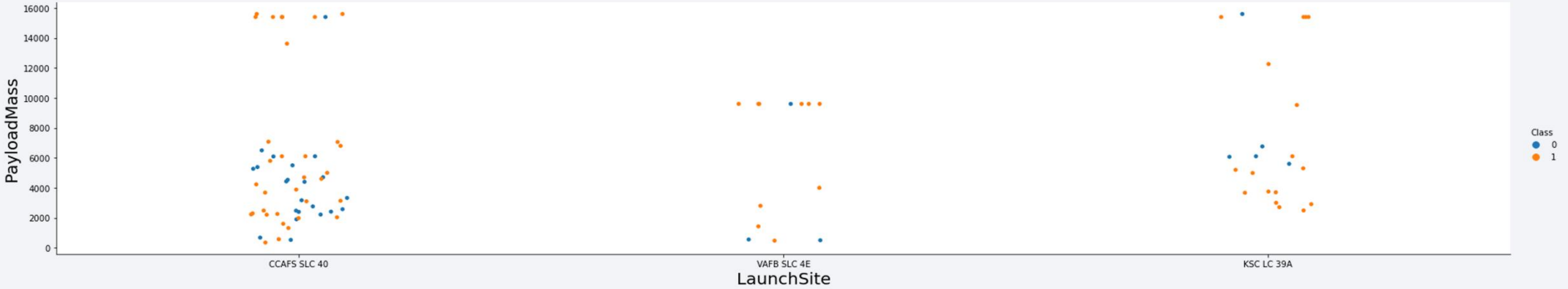
- Show a scatter plot of Flight Number vs. Launch Site



- First 20 – 30 launches took place on CCAFS LC-40. Success rate was far from ideal. But after 60 launches statistics is much better.
- Success rate of VAFB SLC-4E looks promising, but much lower quantity comparing to CCAFS LC-40

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

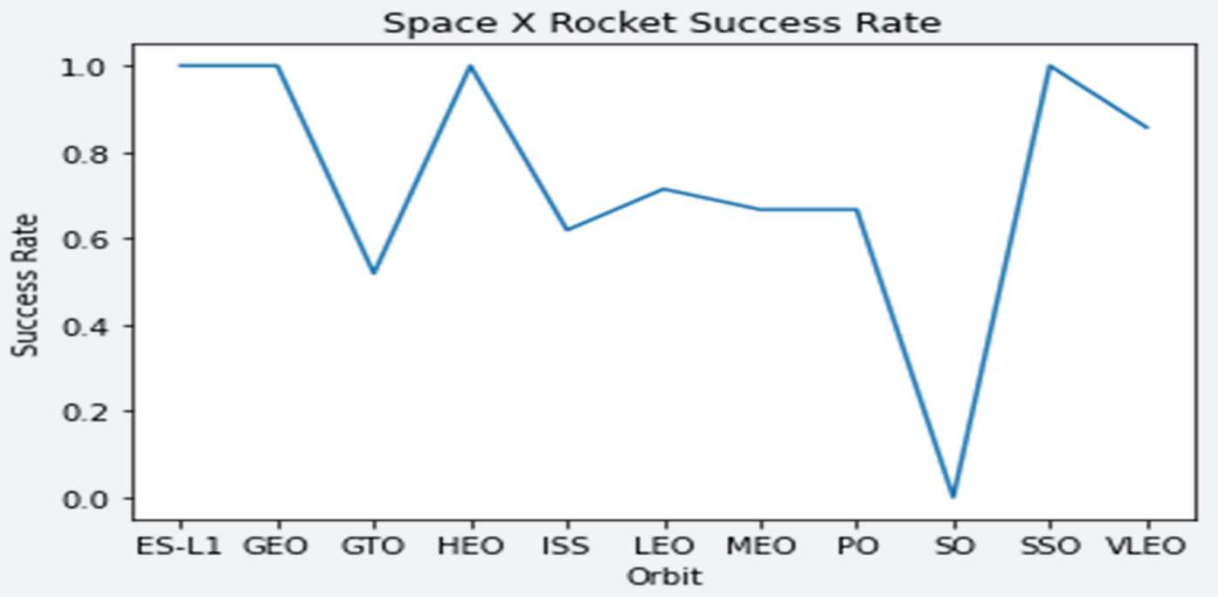


- It seems like heavy payload mass should be better launched CCAFS LC-40 SITE

Launch_Si
CCAFS LC-4
VAFB SLC-4
KSC LC-39A
CCAFS SLC

Success Rate vs. Orbit Type

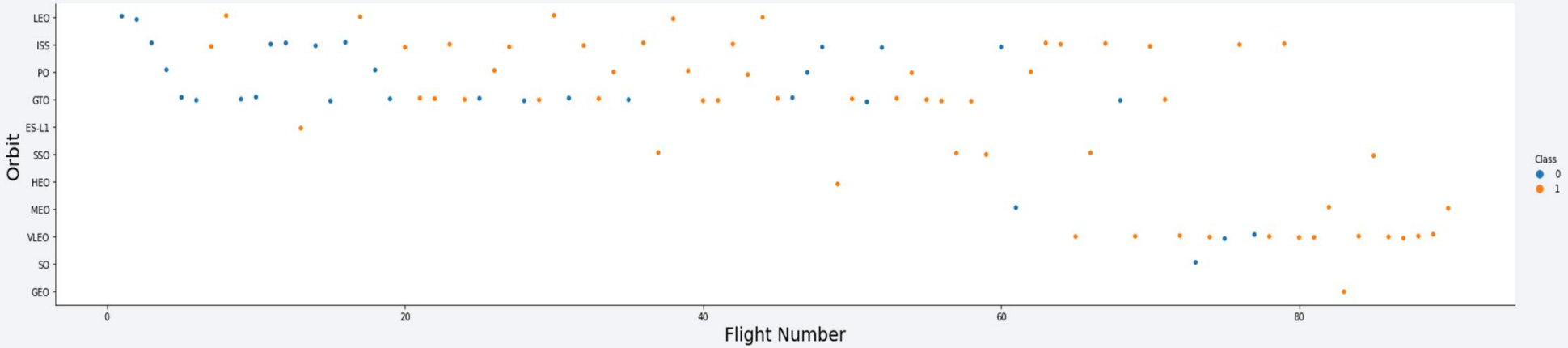
- Show a bar chart for the success rate of each orbit type



- The lowest success rate is for - SO orbit type
- Several orbit type has success rate close to 100%

Flight Number vs. Orbit Type

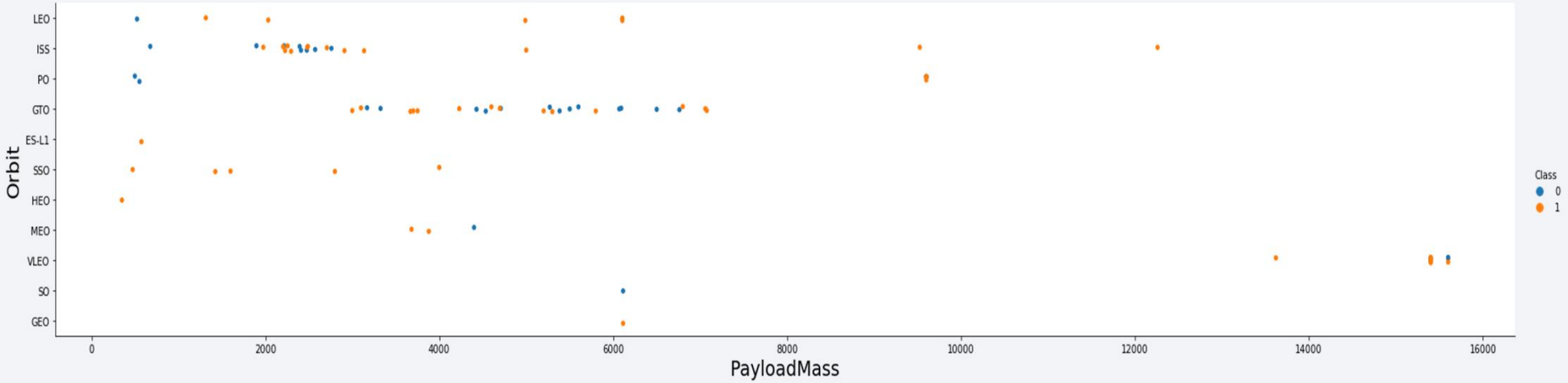
- Show a scatter point of Flight number vs. Orbit type



Success rate of LEO orbit increases with the number of flights.
Success rate of VLEO, SSO or HEO looks good, but number of flights is not significant.

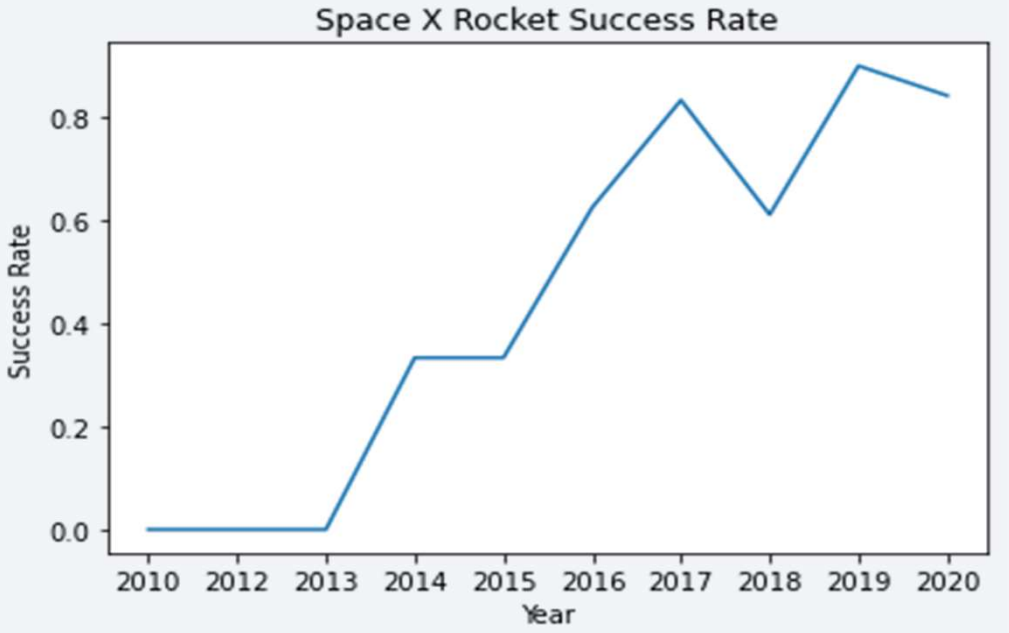
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type



Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- Increase of success rate

All Launch Site Names

- Names of the unique launch sites

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

- Present your query result with a short explanation here

```
%%sql

select distinct Launch_Site from df
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

index	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	08-12-2010	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)
2	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Present your query result with a short explanation here

```
%%sql

select * from df where Launch_Site like 'CCA%' limit 5
```

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

total
45596

- Present your query result with a short explanation here

```
%%sql

select sum(PAYLOAD_MASS__KG_) as total from df where customer = 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

avg
2928.4

- Present your query result with a short explanation here

```
%%sql

select avg(PAYLOAD_MASS__KG_) as avg from df where Booster_Version = 'F9 v1.1'
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

min(Date)
01-05-2017

- Present your query result with a short explanation here

```
%%sql

select min(Date) from df where [Landing _Outcome] = 'Success (drone ship)' or [Landing _Outcome] =
'Success' or [Landing _Outcome] = 'Success (ground pad)'
```

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

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

- Present your query result with a short explanation here

```
%%sql

select Booster_Version from df where [Landing _Outcome] = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000  and PAYLOAD_MASS__KG_ < 6000
```

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

- Present your query result with a short explanation here

```
%%sql

select mission_outcome, count(*) from df group by [Mission_Outcome]
```

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

- Present your query result with a short explanation here

```
%%sql

select Booster_Version from df where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from df)
```

2015 Launch Records

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

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

- Present your query result with a short explanation here

```
%%sql

select [Landing _Outcome], Booster_Version, Launch_Site from df where [Landing _Outcome] = 'Failure (drone ship)'
and Date like '%2015'
```

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

- 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

Landing _Outcome	count(*)	row_number() over(order by count(*) desc)
Success	38	1
No attempt	21	2
Success (drone ship)	14	3
Success (ground pad)	9	4
Failure (drone ship)	5	5
Controlled (ocean)	5	6
Failure	3	7
Uncontrolled (ocean)	2	8
Failure (parachute)	2	9
Precluded (drone ship)	1	10
No attempt	1	11

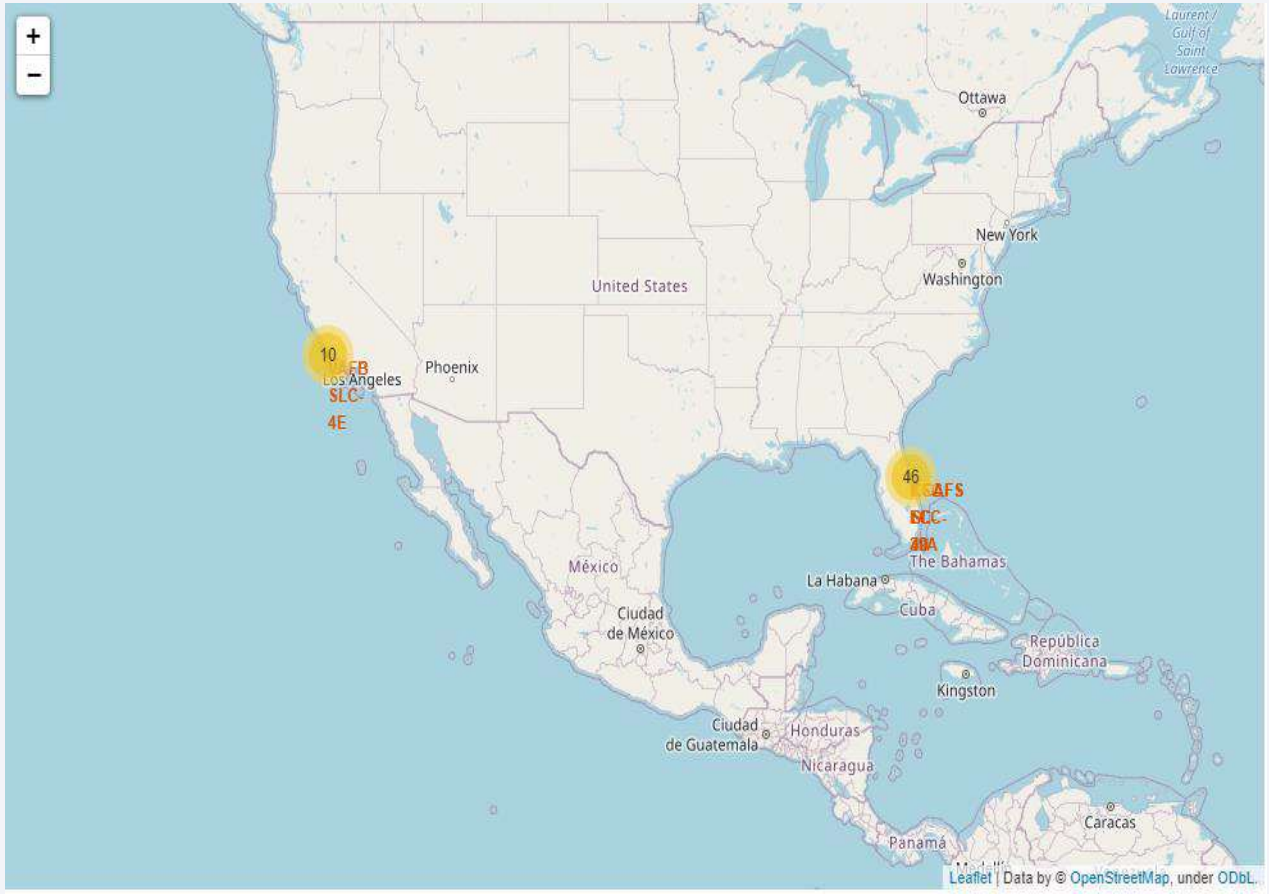
- Present your query result with a short explanation here

```
%%sql

select [Landing _Outcome], count(*), row_number() over(order by count(*) desc) from df group by [Landing _Outcome] order by count(*) desc
```

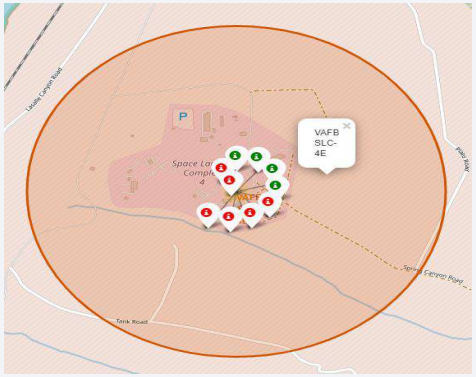

Section 3
Launch Sites
Proximities Analysis

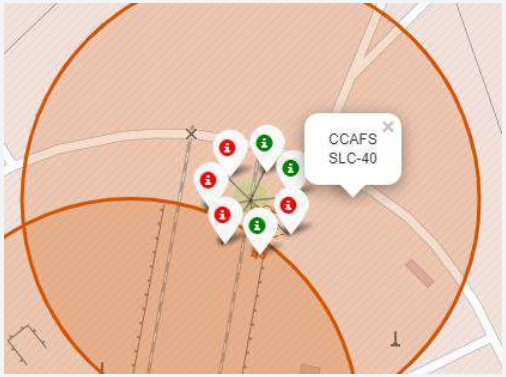
Folium Map. Ground stations

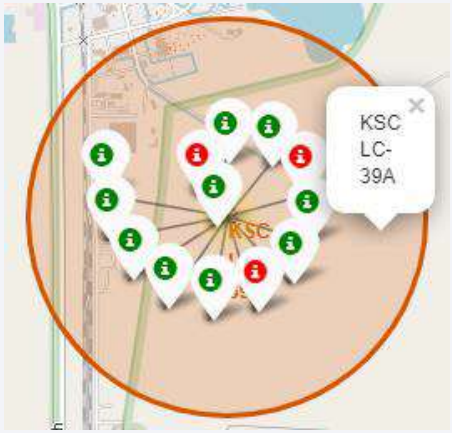


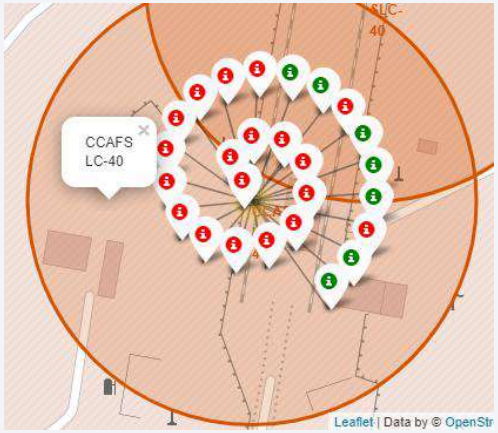
Locations of Space X launch sites

Labeled maps

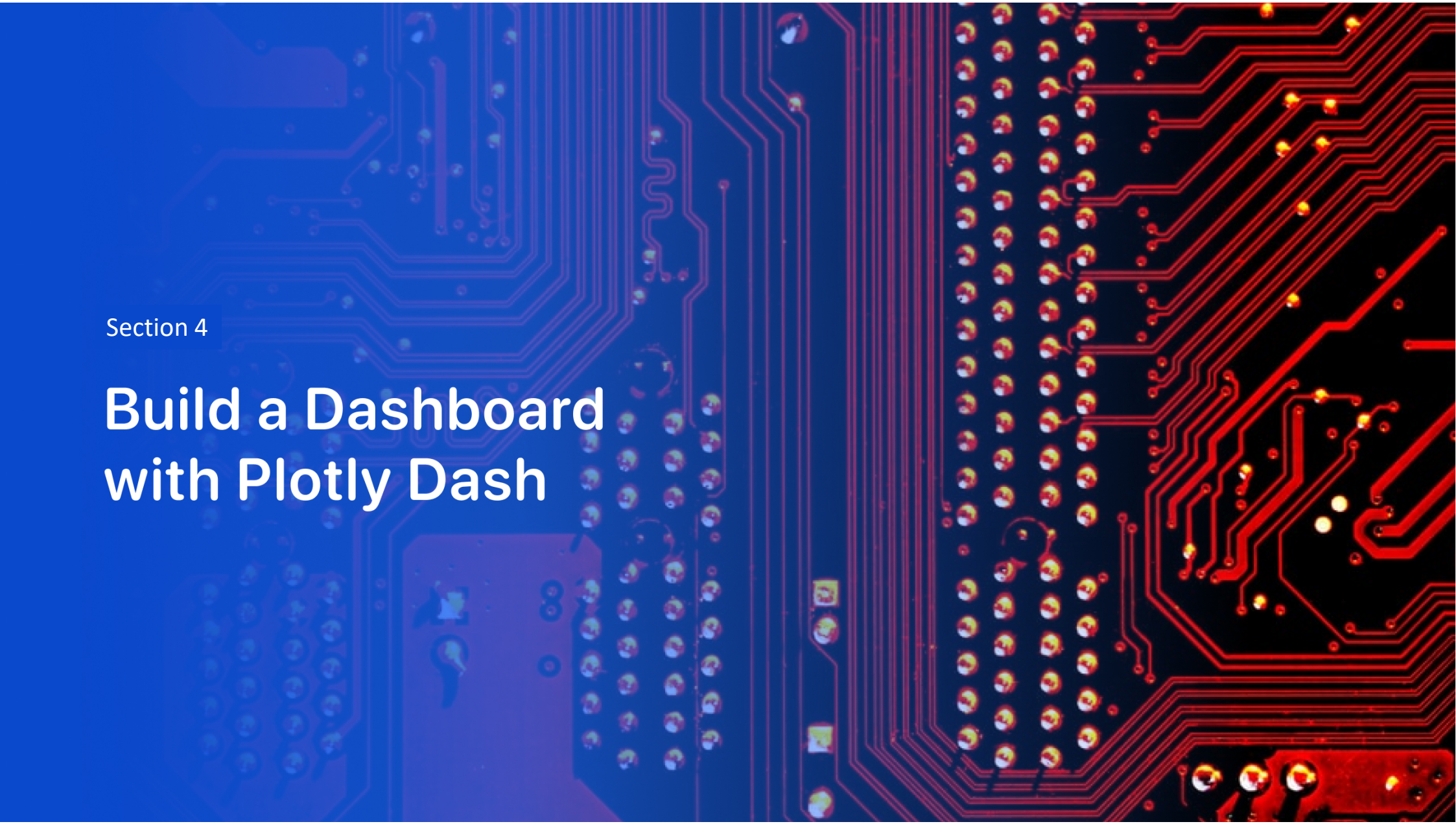




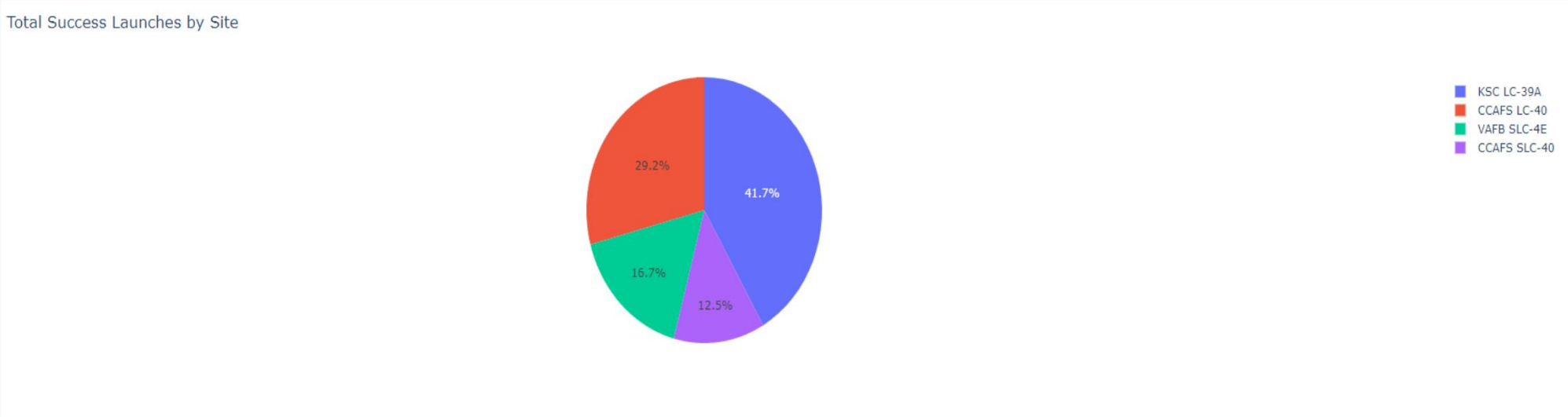




KSC LC-39A has a higher launch success rate.


Section 4
Build a Dashboard with Plotly Dash

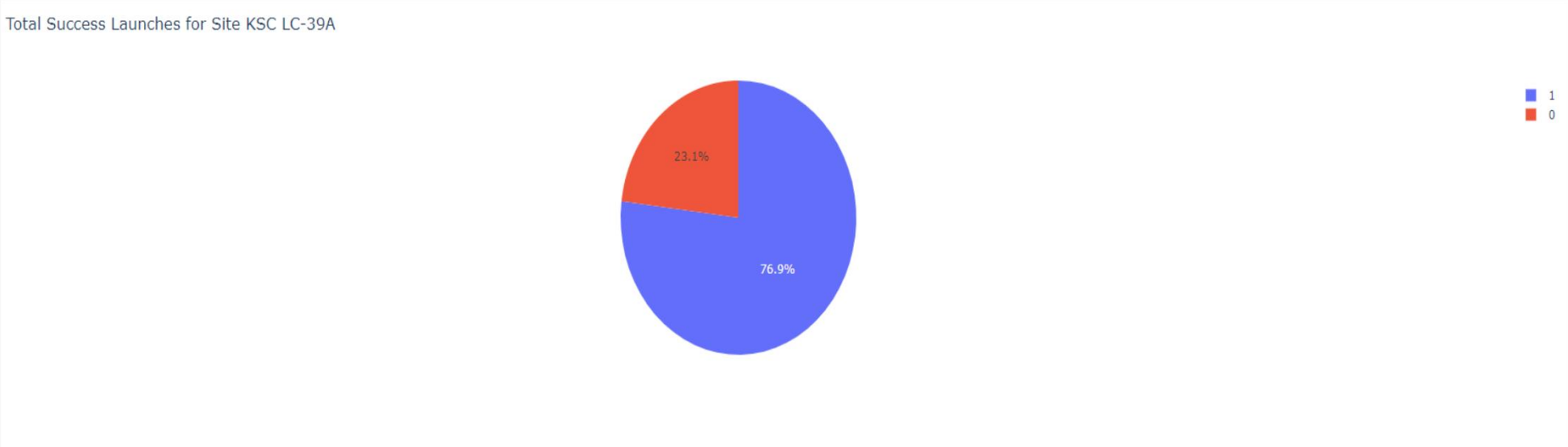
Total success rates



KSC LC-39A - the most successful Site

CCAFS LC-40 – following most successful Site

Total success launches for KSC LC-39A

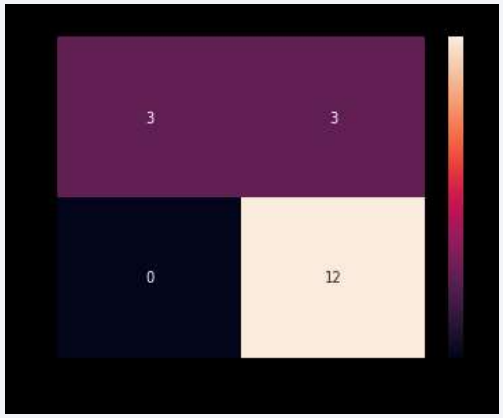


Section 5

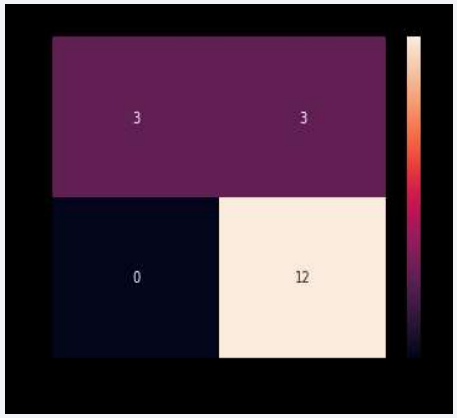
Predictive Analysis (Classification)

Confusion Matrix

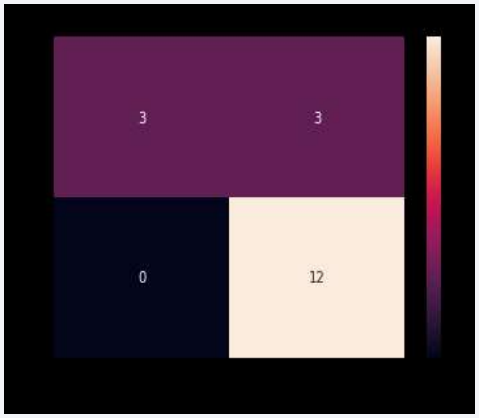
Logistic regression



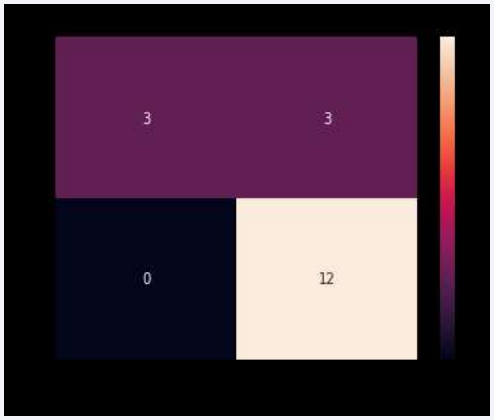
Decision Tree



kNN



SVM



As the test accuracy are all equal, the confusion matrices are also identical.

Conclusions

- Orbits with the best performance have been identified, namely: GEO, HEO, SSO, ES-L1.
- It was revealed that lower payload mass have more success rate than large payload mass . However, the percentage of success depending on the weight is different in different territories.. Some orbits require a light or heavy payload.
- A launcher with best performance settings has been identified - KSC LC-39A.

Thank you!