IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Raghavendran S
01/06/2023

Outline

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

Executive Summary

- **Summary of methodologies**

Data was collected by web scraping. Exploratory Data Analysis(EDA) was performed on the data to gain insights and different Machine Learning models were applied.

- **Summary of Results**

EDA helped us to determine the important features that can be used for prediction and the best ML model was used for prediction.

Introduction

- The objective is to determine if Space Y can compete with SpaceX.
- Our goal was to determine if Space X will reuse their first stage or not. By finding landing outcome of first stage we can find the cost of the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API("https://api.spacexdata.com/v4/rockets/)
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - A landing outcome label was created from outcome column as 0 or 1 for failure and success
- 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 was split into training and test sets and the optimal parameters were chosen using GridSearch

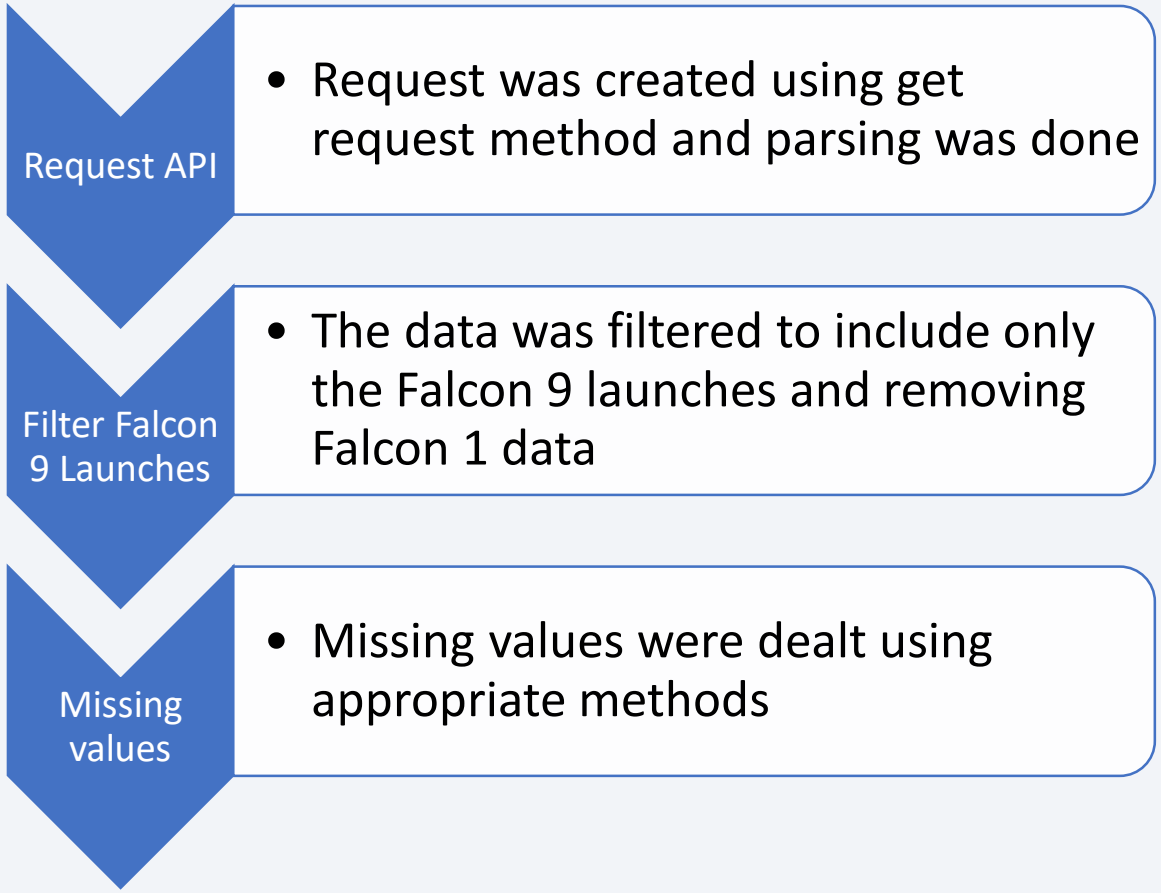
Data Collection

- Data was collection from Space X API("https://api.spacexdata.com/v4/rockets/”) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- Space X has a public API that can be used for Data Collection

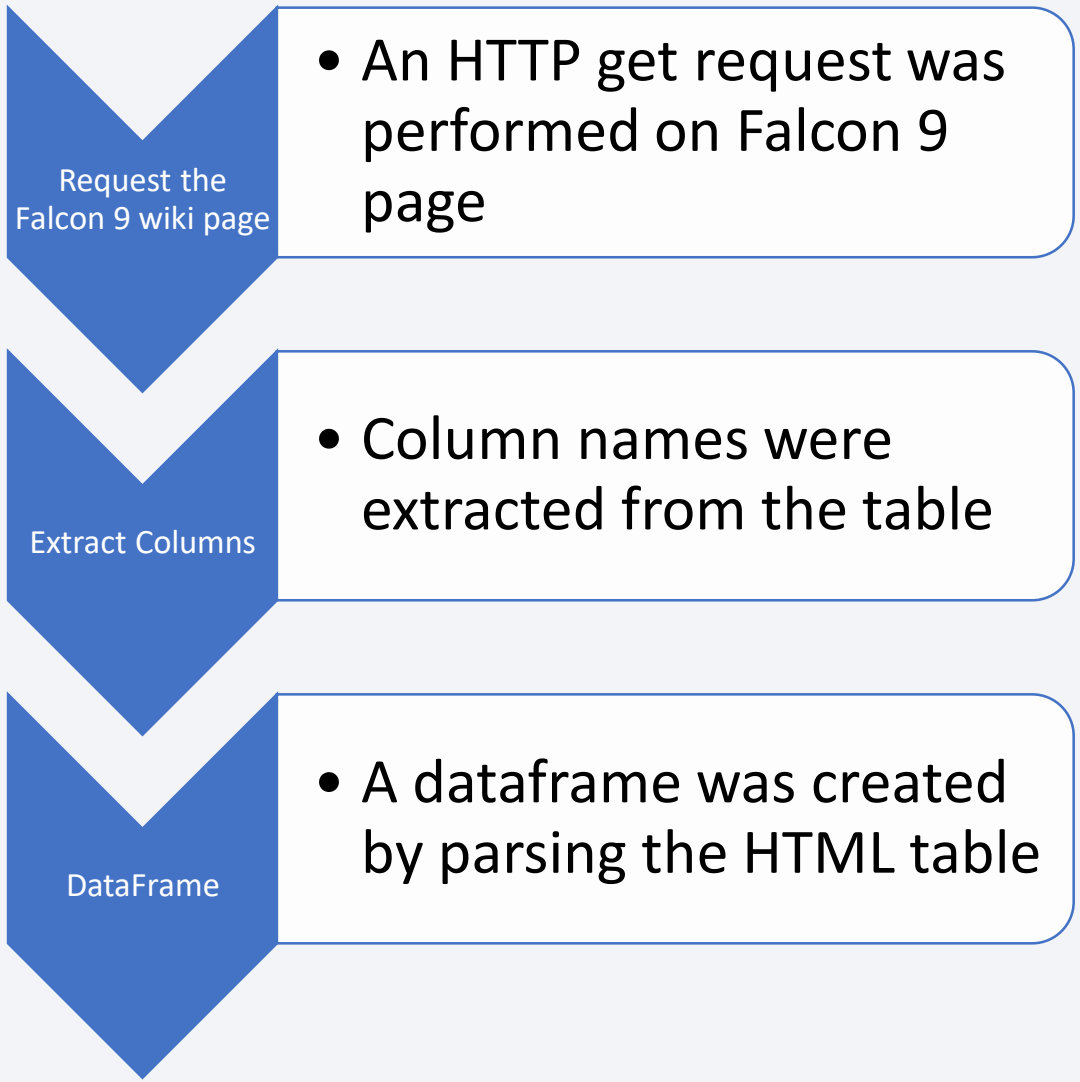
- Code :https://github.com/ragha1097/FirstRepo/blob/main/M1L1(Data%20Collection%20and%20Wrangling%20basics).ipynb



Data Collection - Scraping

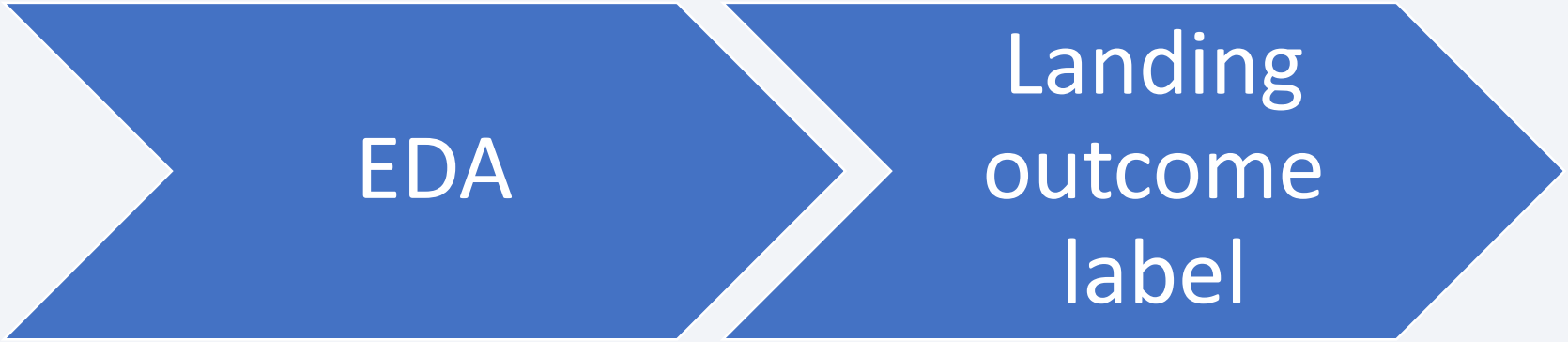
- Web Scraping was done to obtain data from Wikipedia

- Code https://github.com/ragha1097/FirstRepo/blob/main/M1L2(Web%20Scraping).ipynb



Data Wrangling

- EDA was performed on the data initially.
- A label outcome was created based (0 for failure and 1 for success)



https://github.com/ragha1097/FirstRepo/blob/main/M1L3(Data%20Wrangling).ipynb

EDA with Data Visualization

- Scatter plot and Bar plots were used to determine relationships between various columns and also the their relationship with the landing outcome
- https://github.com/ragha1097/FirstRepo/blob/main/M2L2(EDA%20With%20Visualization).ipynb

EDA with SQL

- Determine unique launch names
- Determine 5 records beginning with ‘CCA’
- Determine total payloadmass carried by boosters launched by NASA(CRS)
- Determine avg payloadmass carried by booster version F9 v1.1
- Determine the date of first successful landing outcome in groundpad
- Determine the names of unique boosters satisfying a certain criteria
- Determine the total number of successful and failure missions
- Names of booster versions which carried max payloadmass
- List launch records for the year 2015
- Rank the count of successful missions in a certain period

https://github.com/ragha1097/FirstRepo/blob/main/M2L1(EDA%20with%20SQL).ipynb

Build an interactive map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers to indicate launch sites
- Circles indicate highlights areas
- Lines indicate distance between 2 coordinates
- Marker clusters indicate group of events

Build a dashboard with PlotlyDash

- 2 charts were used for visualization-Pie chart and Scatter Chart
- This dashboard application contains components such as dropdown list and a range slider to interact with the above charts.
- This dashboard can be used to
    - Which site has the highest launches?
    - Which site has the highest launch success rate?
    - Which payload range(s) has the highest launch success rate?
    - Which payload range(s) has the lowest launch success rate?
    - Which booster version has the highest launch success rate?

Predictive Analysis (Classification)

- Data was split into train and test sets and 4 ML models were tested and the accuracy and confusion matrix of each model was determined

NumPy Array
- Numpy array was created for the column class

Standardize the data
- Standard scaling was performed on the data

Train Test Split
- Data was split into train and test sets with a test size of 0.2

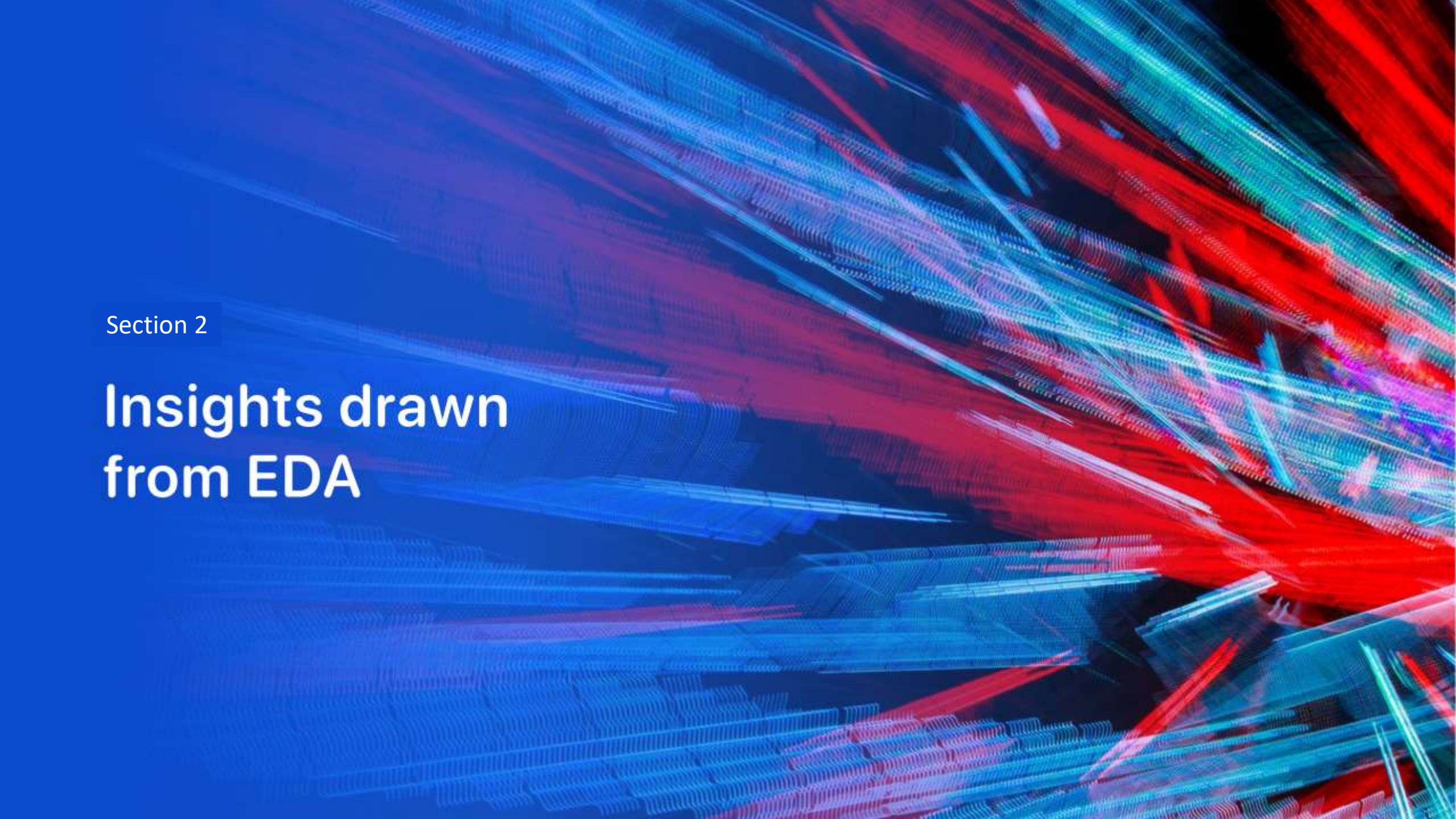
ML models
- Linear regression,SVM,KNN and Decision tree models were applied

Accuracy
- Accuracy was calculated and confusion matrix was printed and the results were compared

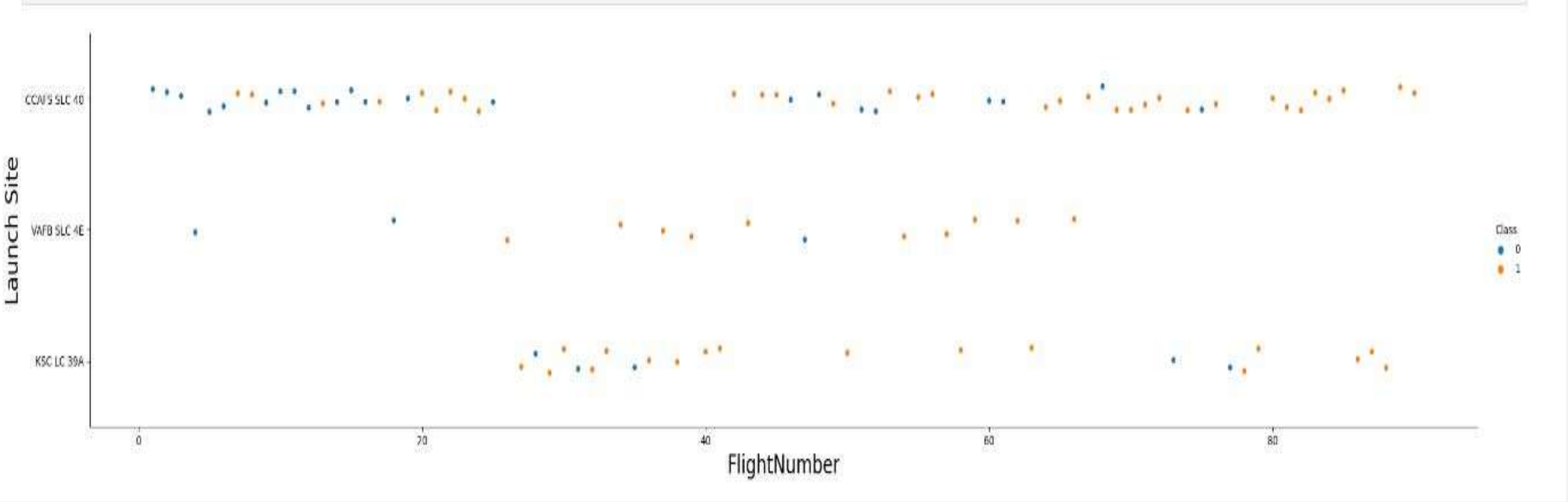
https://github.com/ragha1097/FirstRepo/blob/main/M4L1(ML).ipynb

Results

- Space X uses different launch sites
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).
- For Leo orbit success appears to be related to number of flights
- Landing outcomes increased from the year 2013

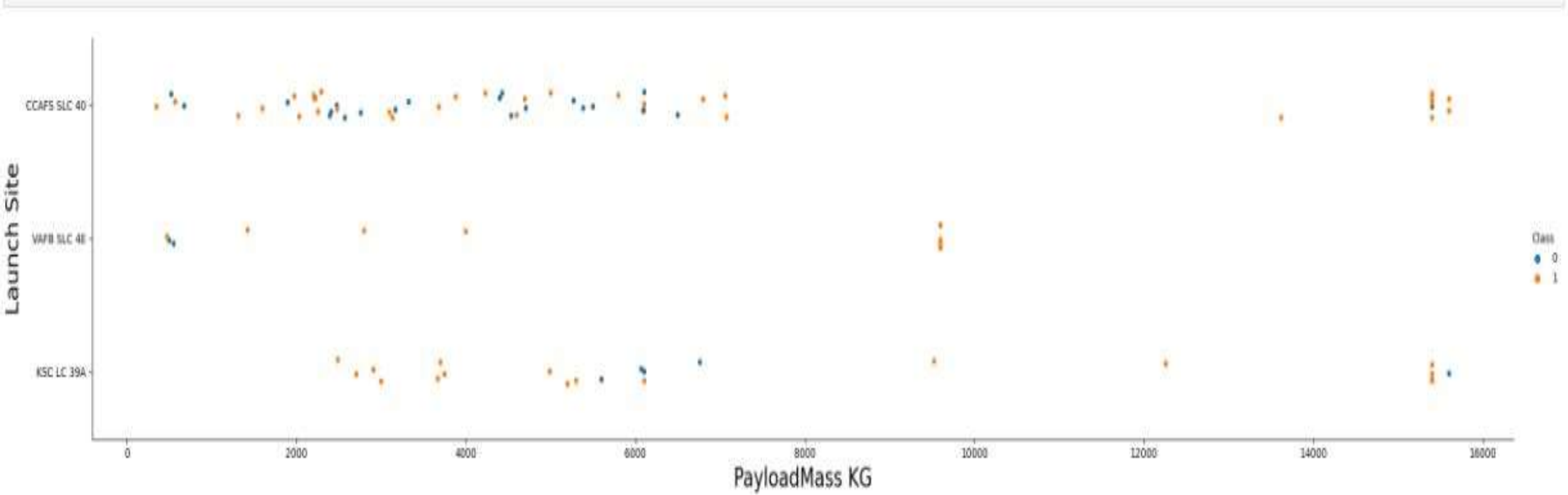

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



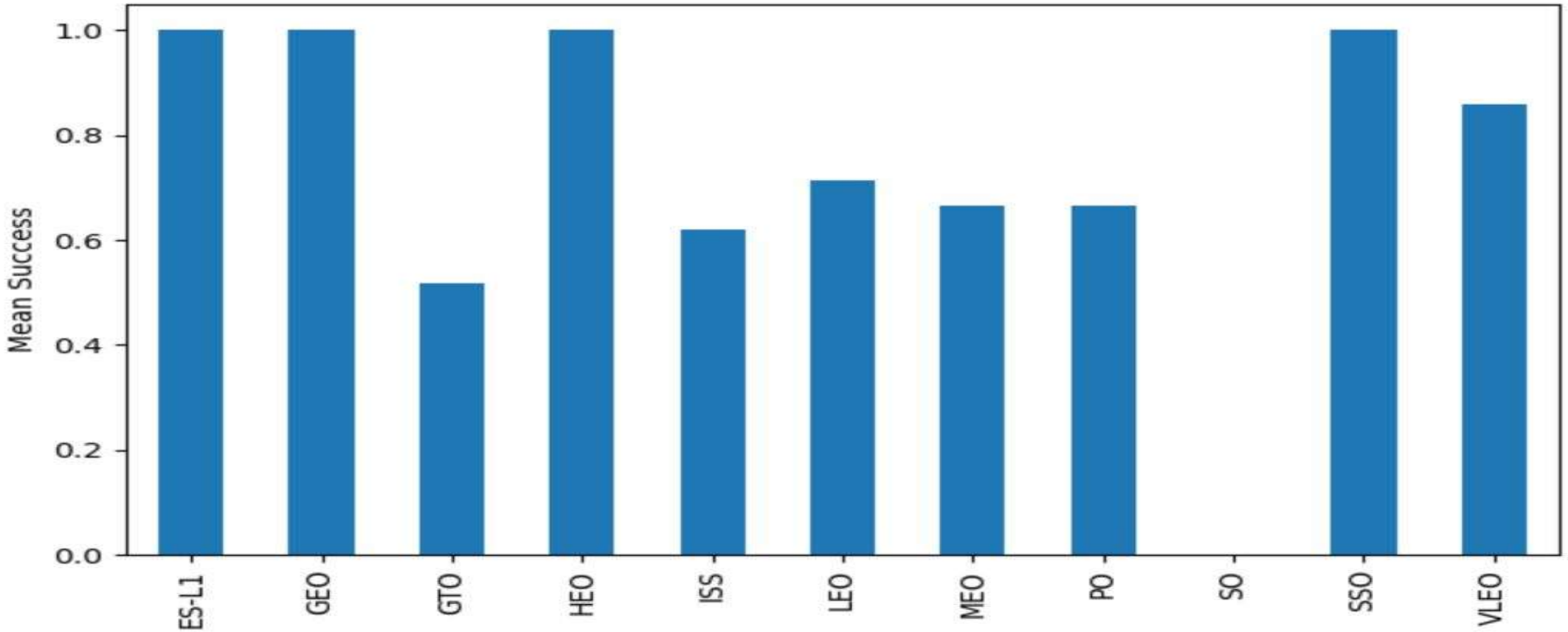
- Success of launch improves with increase in FlightNumber
- Different Launch Sites have different success rate.CCAFS LC-40 has 60% success rate while the remaining two have 77%

Payload vs. Launch Site



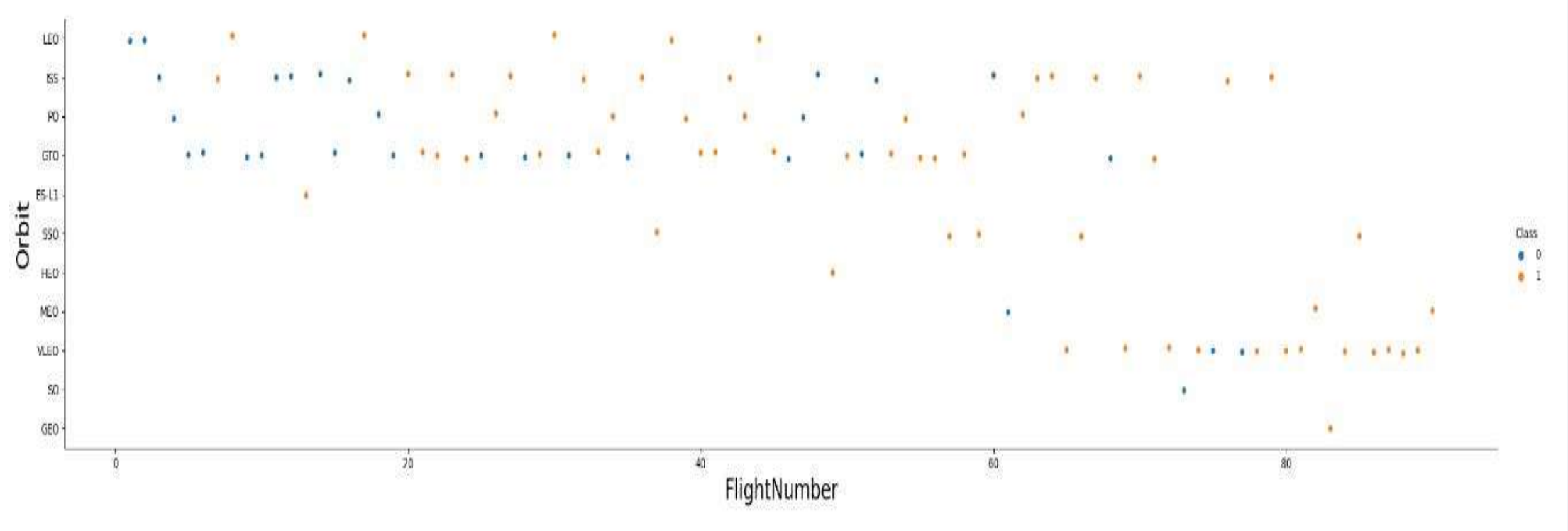
- For VAFB-SLC launch site there are no rockets launched above a payload of 1000kg
- High Success rate for Heavy PayloadMass

Success Rate vs. Orbit Type



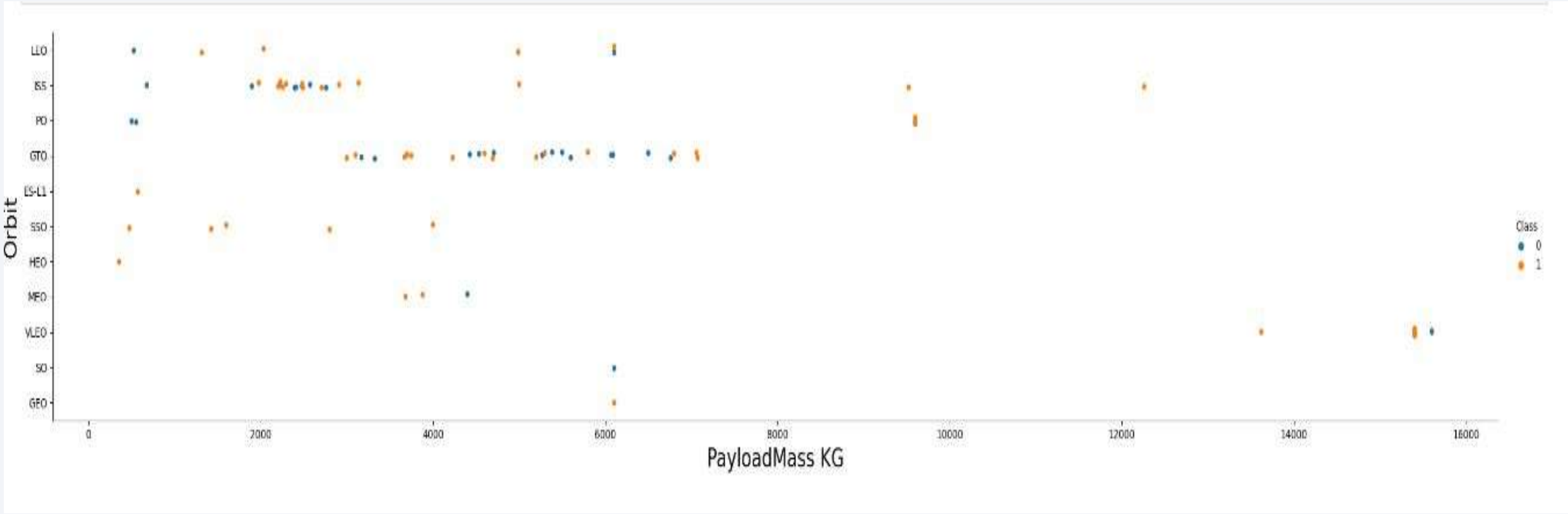
ES-L1,GEO,HEO and SSO have high success rates.

Flight Number vs. Orbit Type



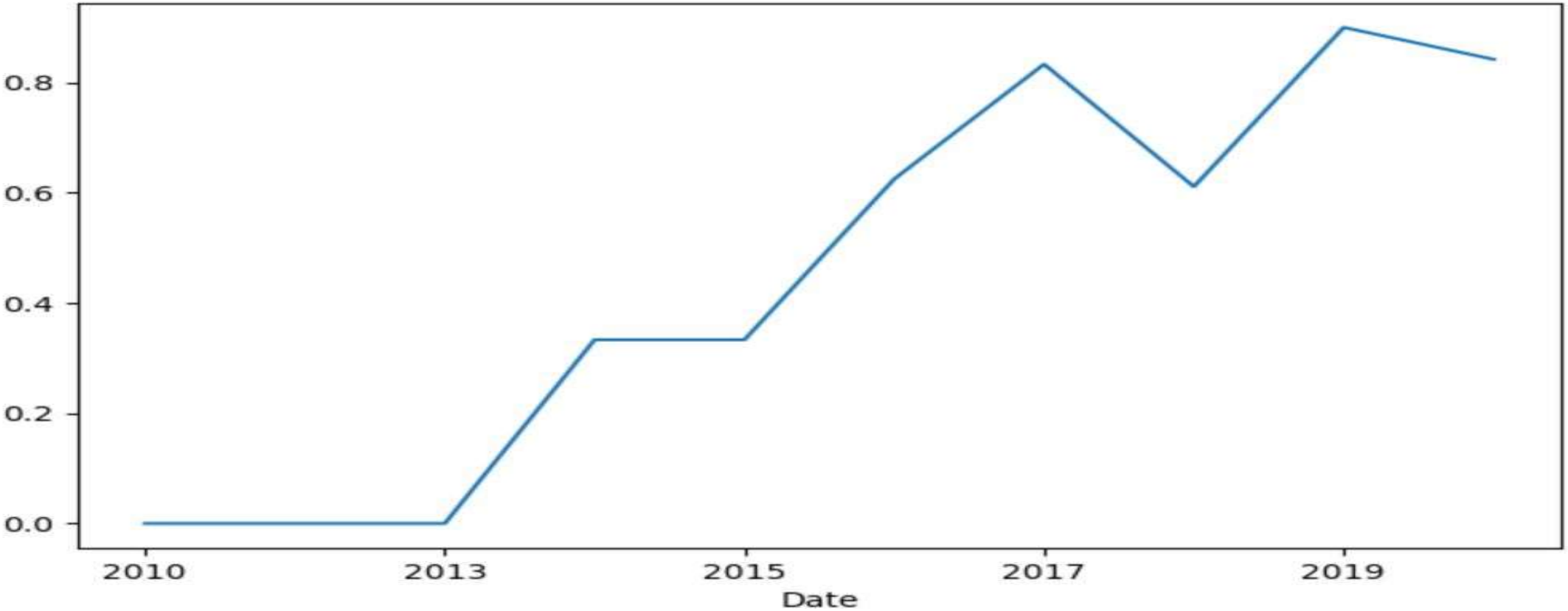
- For LEO orbit success rate increases with Flight Number
- For GTO orbit there is no clear relationship between success and Flight Number

Payload vs. Orbit Type



- With heavy payloads successful landing rates are more for Polar, LEO and ISS

Launch Success Yearly Trend



- Success rate is on the rise from the year 2013

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
]: 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

- DISTINCT keyword is used to obtain the names of all unique launch sites
- There are 4 unique launch sites as shown above

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- The above query is used to obtain 5 records where launch site begins with ‘CCA’

Total Payload Mass

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)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596.0

- Agg function SUM() is used to find total payload mass carried by boosters launched by NASA(CRS)
- Total PayloadMass carried by boosters launched by NASA(CRS) is 45596 Kg

Average Payload Mass by F9 v1.1

Task 4

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

```
%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

- Aggregate Function AVG() is used to obtain average of values
- Avg payloadmass where booster version = F9 v1.1 is 2928.4 Kg

First Successful Ground Landing Date

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

Hint:Use min function

```
%sql SELECT MIN(Date) as DATE from SPACEXTBL WHERE LANDING_OUTCOME ='Success (ground pad)';
```

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

DATE
22/12/2015

- First successful landing outcome in ground pad was achieved on 22/12/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

[89]:
```
%sql SELECT Booster_Version from SPACEXTBL WHERE LANDING_OUTCOME ='Success (drone ship)' AND PAYLOAD_MASS__
```

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

[89]:

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

- 4 Booster versions were successful in drone ship launch and have a payload mass between 4000 and 6000 Kg

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
.6]: %sql SELECT MISSION_OUTCOME,COUNT(MISSION_OUTCOME) from SPACEXTBL GROUP BY MISSION_OUTCOME;
 * sqlite:///my_data1.db
Done.
```

.6]:

Mission_Outcome	COUNT(MISSION_OUTCOME)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- High Success rate

Boosters Carried Maximum Payload

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

```
[99]: %sql select booster_version from SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTBL) ORDER BY BOOSTER_VERSION;
 * sqlite:///my_data1.db
Done.
```

[99]:

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

- A sub-query is used to find max payloadmass.
- Booster versions with maximum payload mass are then displayed in order of Booster Version
- 12 Booster Versions can be seen in the output above

2015 Launch Records

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

[104]:
```
%sql SELECT substr(Date,4,2) as Month,substr(Date,7,4) as Year,date,Booster_version,Landing_outcome,launch_site from SPACEXTBL WHERE SUBSTR(Date,7,4)='
```

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

[104]:

Month	Year	Date	Booster_Version	Landing_Outcome	Launch_Site
10	2015	01/10/2015	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	2015	14/04/2015	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

- There are 2 records for the year 2015

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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
[106]: %sql SELECT LANDING_OUTCOME,COUNT(LANDING_OUTCOME) as C from SPACEXTBL WHERE Date between '04-06-2010' AND '20-03-2017' GROUP BY LANDING_OUTCOME ORDER
```

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

[106]:

Landing_Outcome	C
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

- Landing outcome and their counts are displayed in descending order of count

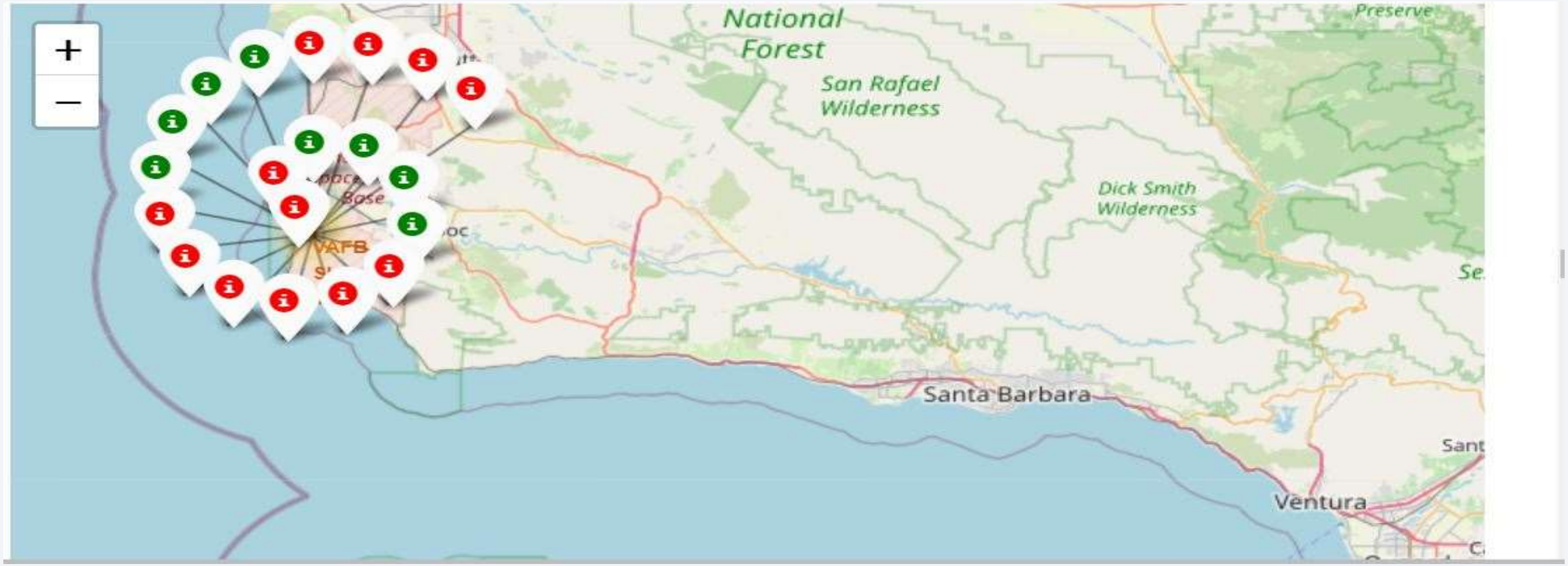
Section 3

Launch Sites Proximities Analysis

Launch Sites

- Launch Sites are safer near the sea

Launch Outcomes for site VAFB SLC-4E



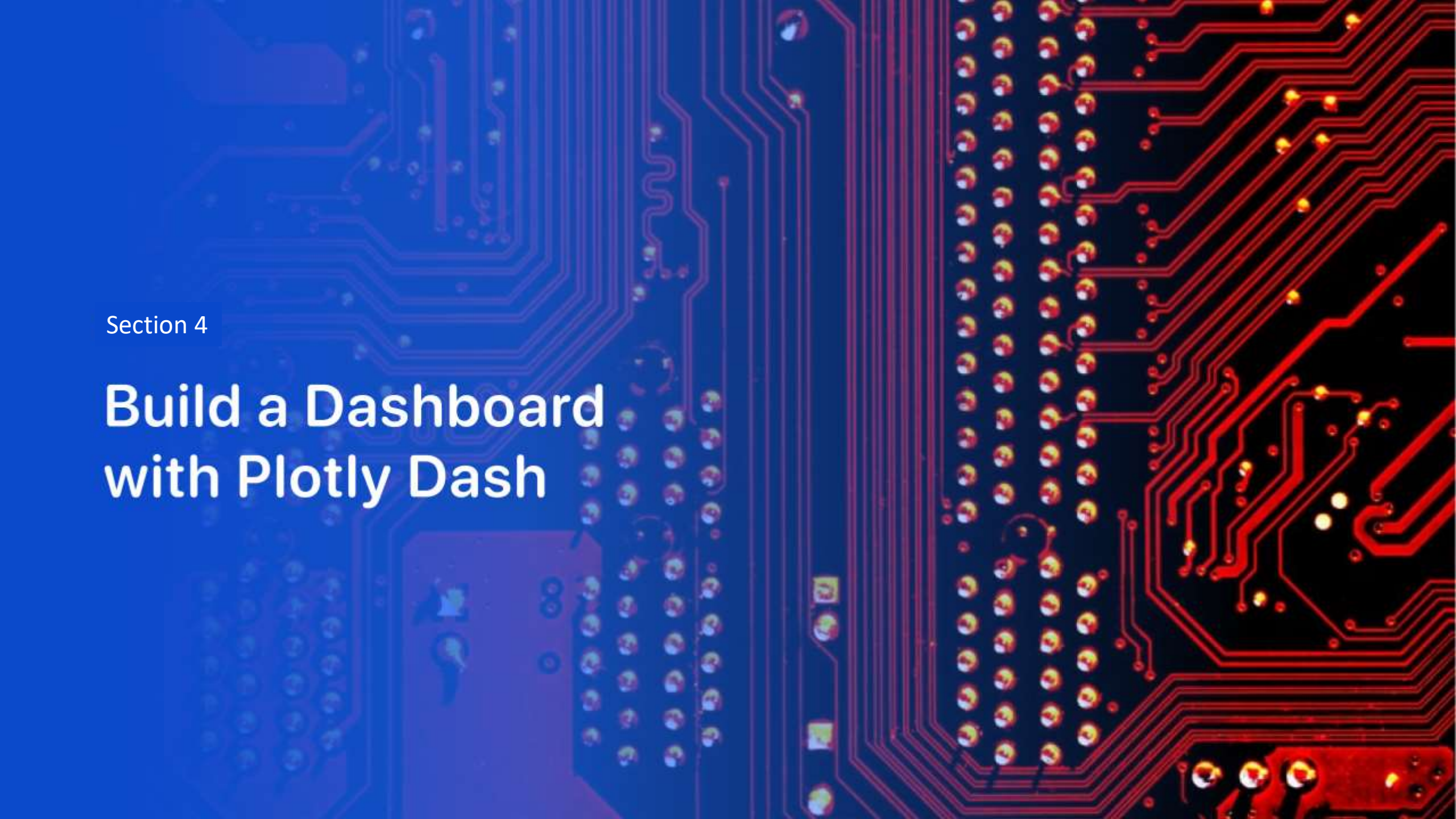
- Green markers indicate success
- Red Markers indicate failure

CCAFS SLC-40 Distance to coastline

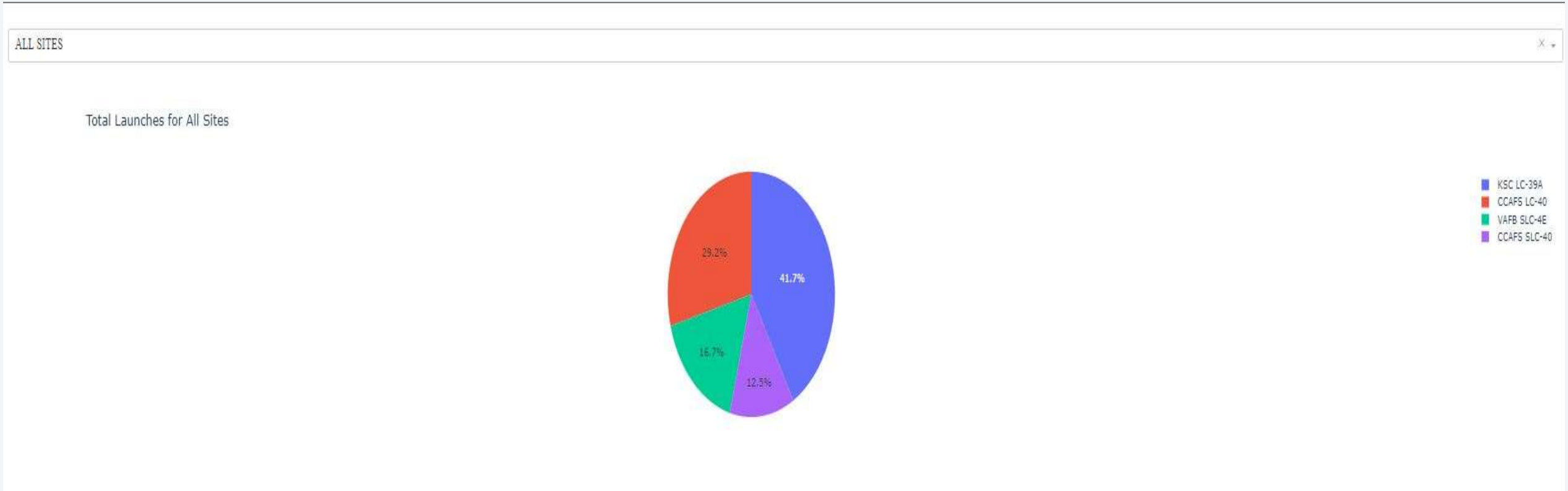
- Distance is indicated by blue line

Section 4

Build a Dashboard with Plotly Dash

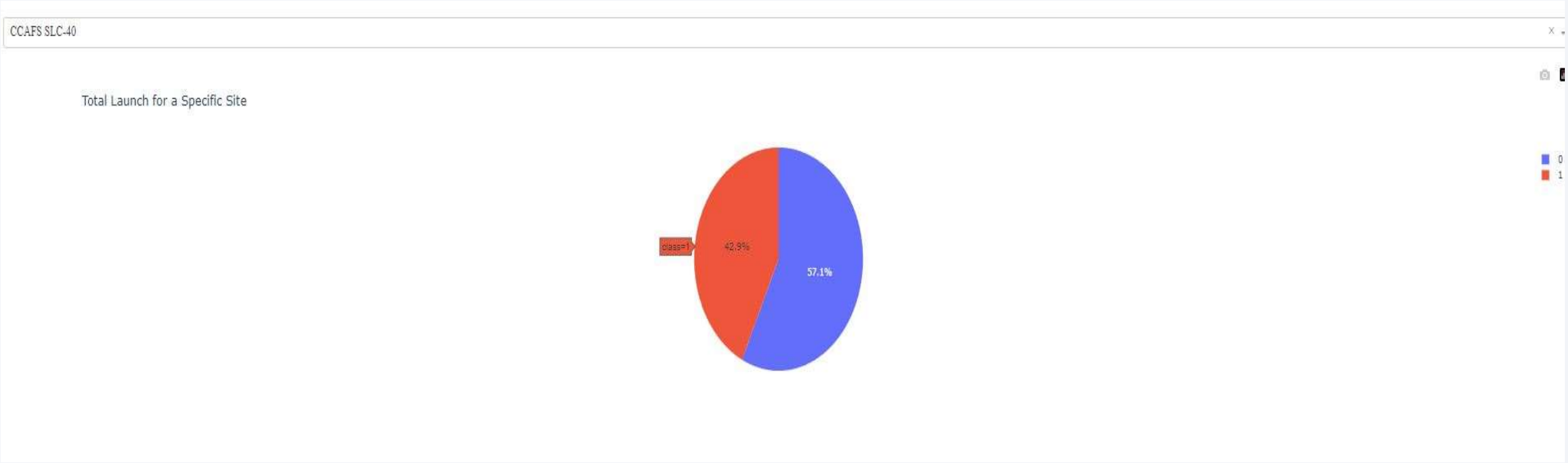


Successful Launches by Site



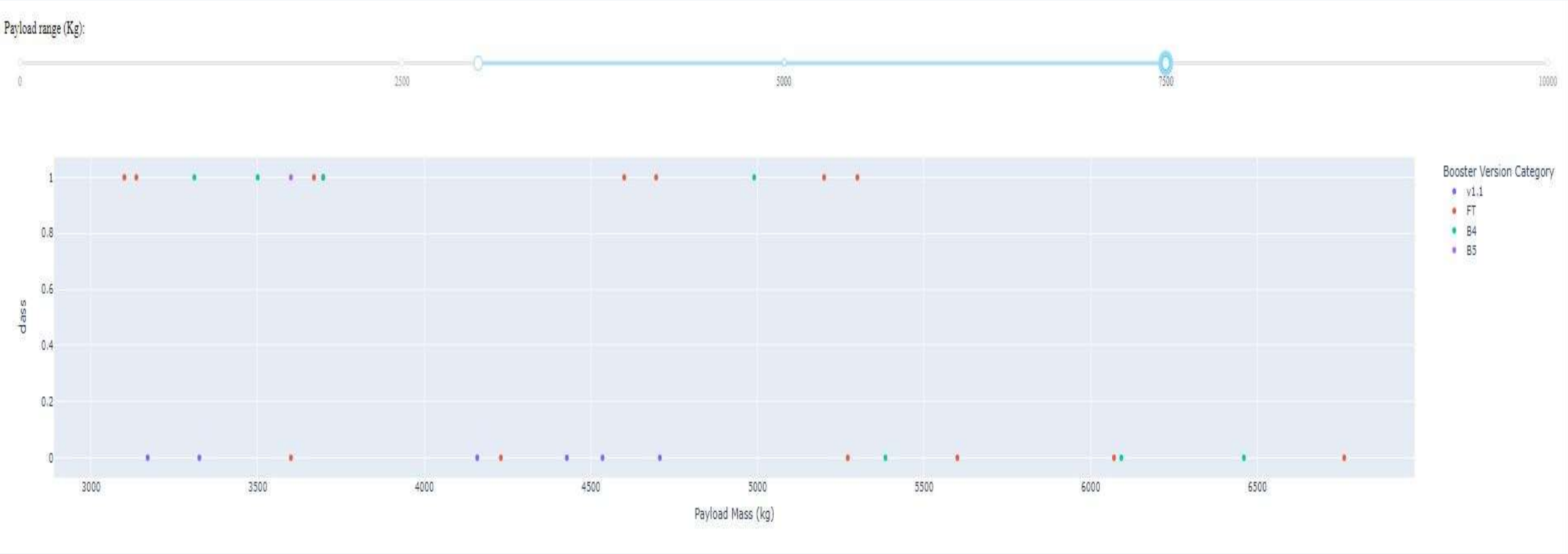
- Most of the successful launches are from KSC LC-39A(41.7%)

CCAFS SLC-40



- Only 12.5 % of successful launches were from CCAFS SLC-40 (from the previous slide)
- More than 50% of launches have failed from this site which explains the very low contribution.

Payload vs Outcome



- Payloads lesser than 5500 kg and FT boosters have a high success percentage

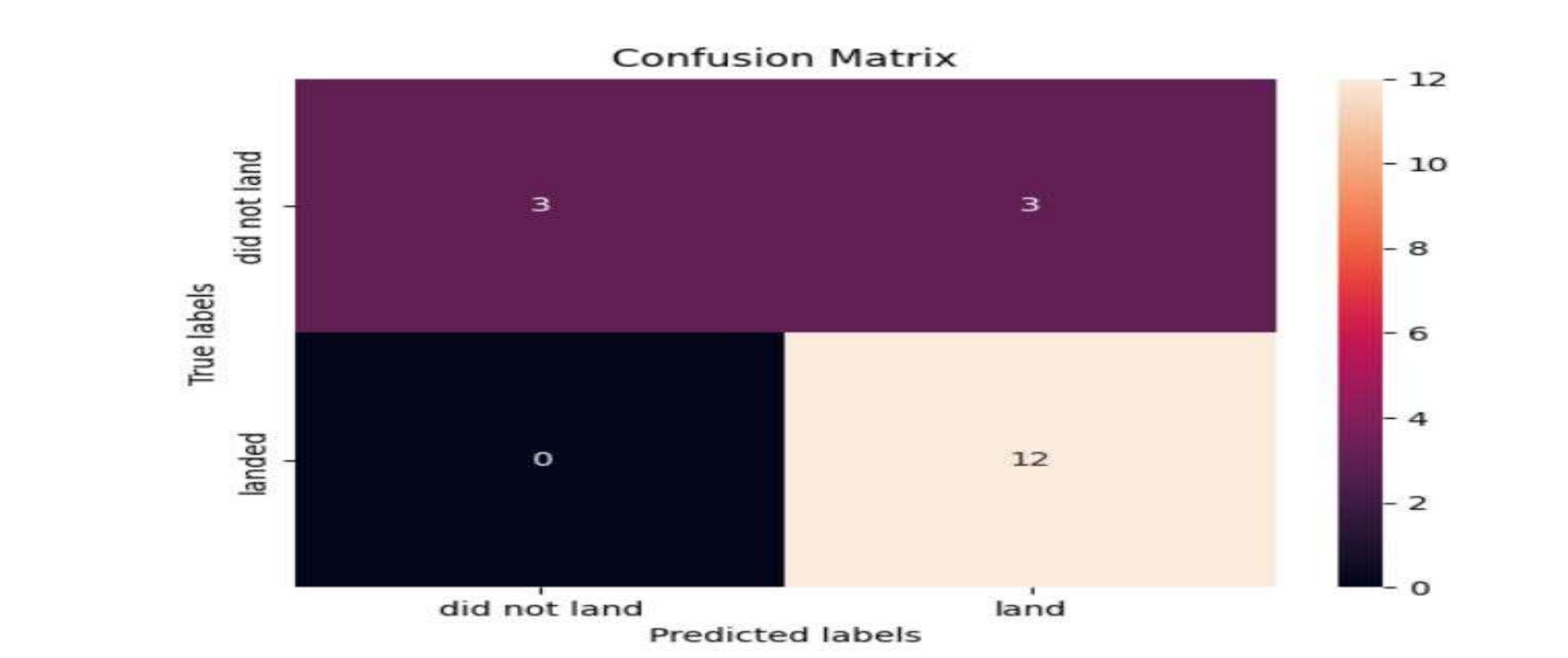
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistic regression, SVM and KNN have the same test accuracy of 83.3333%
- SVM has the highest train accuracy of 88.8%

Confusion Matrix-SVM



- Confusion matrix has a very good true positive prediction

Conclusions

- KSC LC-39 has a very high launch success rate
- Launches below 5500Kg have a high success rate
- Launch sites near ocean are often safe
- CCAFS SLC-40 has a very low success rate
- There has been a steady increase in success rate from the year 2013
- SVM model can be used to predict successful landings

Appendix

- The increase in success rate from 2013 can also be attributed to the advancements in technology.

Thank you!