IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Dang Van Hiep>
<02-Feb-2024>

Outline

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

Executive Summary

- Summary of methodologies
  - Collecting the Data
  - Data Wrangling
  - Exploratory Analysis Using SQL
  - EDA with Visualization
  - Data Visualization with Folium
  - Interactive Dashboard with PlotlyDash
  - Machine Learning Prediction (Classification)
- Summary of all results
  - Exploratory Analysis Results
  - Interactive Visualization Results
  - Predictive Analysis Results

Introduction

- Project background and context
  - The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX.
  - The second stage of a rocket helps bring the payload to orbit, but the first stage does most of the work and is much larger.
  - 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.
  - The Space Y company would like to compete with SpaceX founded by Billionaire industrialist EllonMusk.
- Problems you want to find answers
  - If we can determine the first stage of landing, we can determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - From SpaceX REST API
  - With Web Scraping from Wiki page
- Perform data wrangling
  - Dealing with Missing Values, Feature Engineering, Scaling, Dummies Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Sklearn LogisticRegression , SVM, DecisionTreeClassifier , KNeighborsClassifier algorithms
  - GrigSearch parameters tuning, 10 folds Cross Validation

Data Collection

There are 2 way to collect data:

- SpaceX REST API
  - Performed GET request to the SpaceX REST API various endpoints starting with
  - https://api.spacexdata.com/v4/
  - Responses in the form of a list of JSON objects were gathered
  - JSON format was converted into Pandas DataFrameusing the json_normalizefunction
- Web Scraping
  - Performed an HTTP GET request to the Falcon9 Launch HTML Wiki page
  - Used Python BeautifulSouppackage to web scrape HTML tables from response
  - Parsed the data from HTML tables and converted into a Pandas DataFrame

Data Collection – SpaceX API

- GET request to SpaceX RESTAPI

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

response = requests.get(spacex_url)

content = response.json()

data = pd.json_normalize (content)
```

https://github.com/hiepdv/My-submission-Applied-Data-Science-Capstone/blob/042e2c1c8dbd607237683e88fd9035addc18d9cd/Hiepdv-DataCollectionAPI.ipynb

Data Collection - Scraping

- Web Scraping Using Python BeautifulSoup

```
static_url="https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

response = requests.get(static_url).text

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

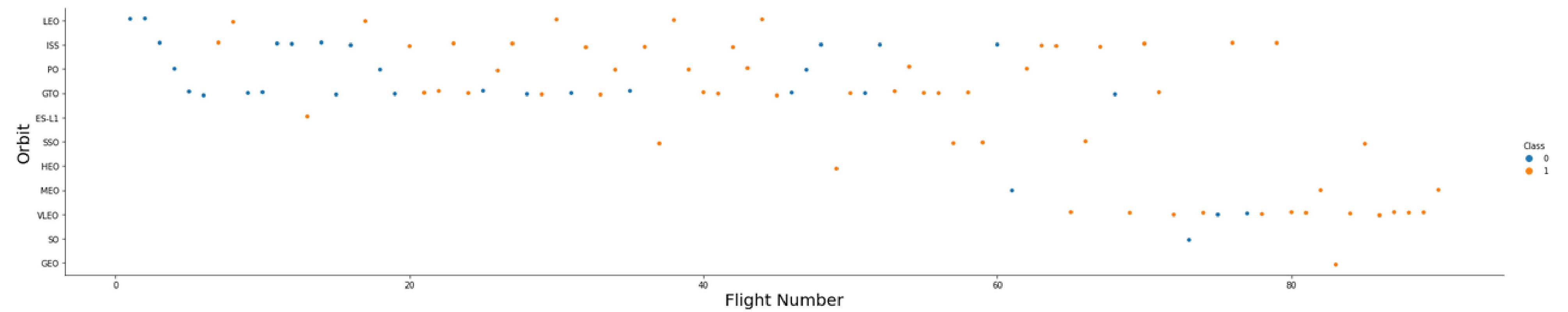
https://github.com/hiepdv/My-submission-Applied-Data-Science-Capstone/blob/042e2c1c8dbd607237683e88fd9035addc18d9cd/Hiepdv-DataCollectionWebscraping.ipynb

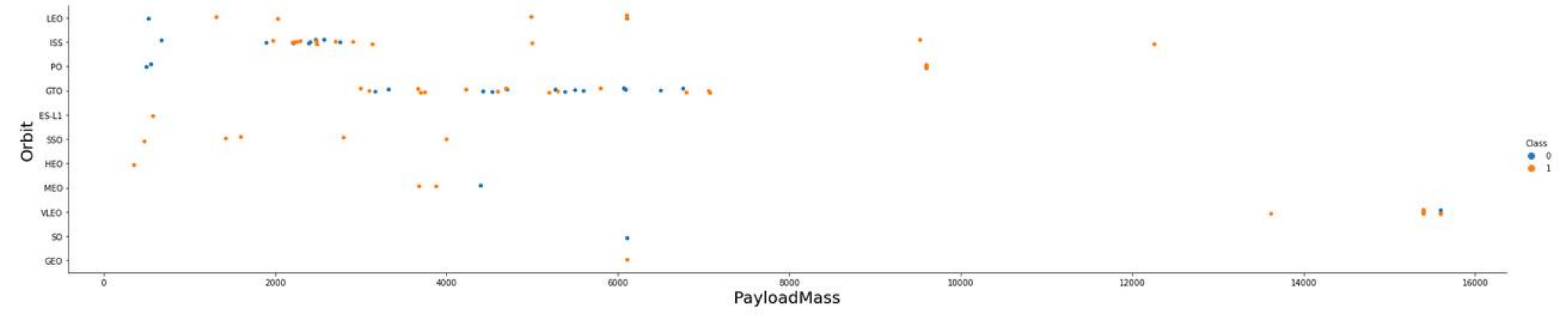
Data Wrangling

- Payload Mass missing values replaced with mean value (SpaceX API code)
- Calculated the percentage of the missing values in each attribute
- Identified which columns are numerical and categorical
- Determined the number of launches on each site
- Determine the number and occurrence of each orbit
- Created a landing outcome label from Outcome column

https://github.com/hiepdv/My-submission-Applied-Data-Science-Capstone/blob/0f06c656917dee12b9ea23e824bca77308310a4a/Hiepdv-DataWrangling.ipynb

EDA with Data Visualization





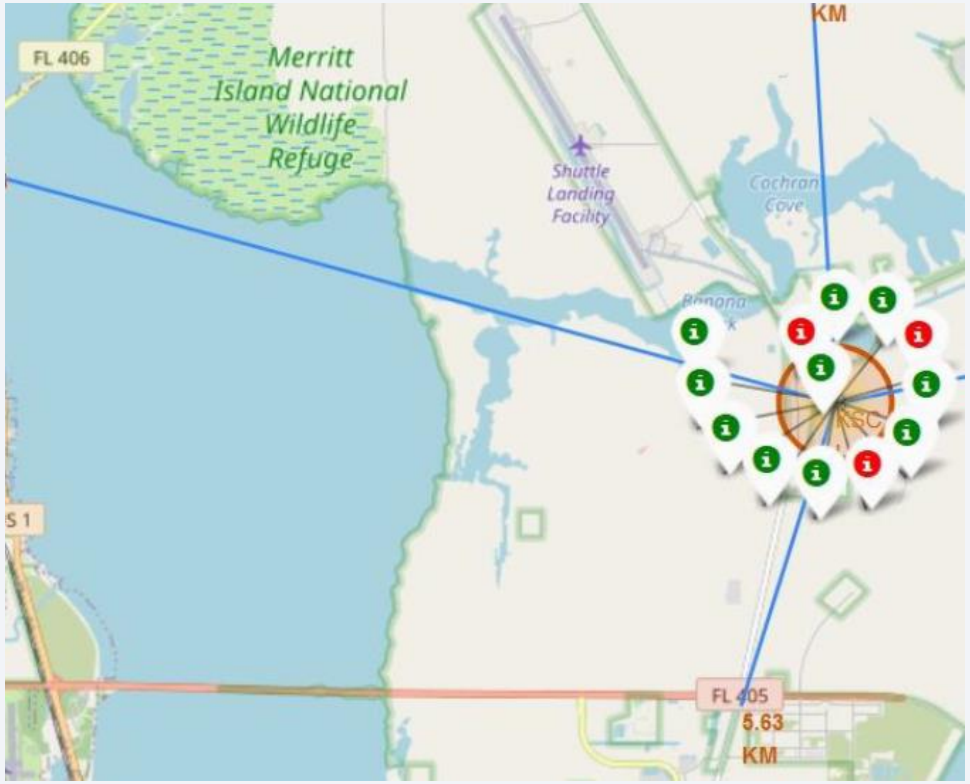
https://github.com/hiepdv/My-submission-Applied-Data-Science-Capstone/blob/0f06c656917dee12b9ea23e824bca77308310a4a/Hiepdv-EDA_DataVisualization.ipynb

EDA with SQL

```
select unique(LAUNCH_SITE) from SPACEXTBL

select *from SPACEXTBL where LAUNCH_SITE like'CCA%' limit(5)

select SUM(payload_mass__kg_)from SPACEXTBLwherecustomer='NASA (CRS)'

select avg(payload_mass__kg_)from SPACEXTBLwhere booster_versionlike 'F9 v1.1'

select min(DATE)from SPACEXTBLwhere Landing_Outcome= 'Success (ground pad)'

select booster_versionfrom SPACEXTBLwhere Landing_Outcome= 'Success (drone ship)'and
payload_mass__kg_ between 4000 and 6000

selectmission_outcome,count(mission_outcome)fromSPACEXTBLgroup bymission_outcome

select booster_versionfrom SPACEXTBLwhere payload_mass__kg_ in (select max(payload_mass__kg_)
from SPACEXTBL)

select Landing_Outcome, booster_version,launch_sitefromSPACEXTBLwhereLanding_Outcome=
'Failure (drone ship)'and EXTRACT(YEAR FROM DATE) =2015

select Landing_Outcome, count(Landing_Outcome) as totalfrom SPACEXTBLwhere DATE between
'2010-06-04'and '2017-03-20'group byLanding_Outcomeorder bytotal DESC
```

Build an Interactive Map with Folium

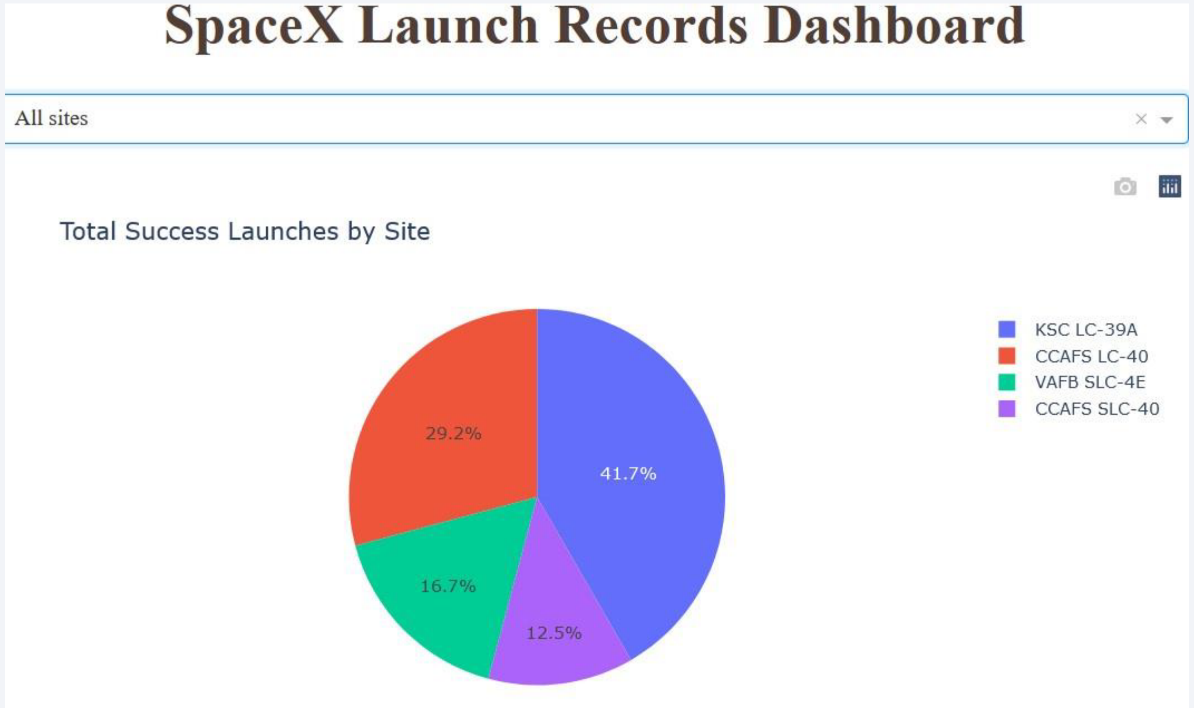
- The launch success rate may depend on many factors such as payload mass, orbit type.
- It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- The goal of geo plots is to analyzing the existing launch site locations, discover the factors involved in finding an optimal location for building a launch site.

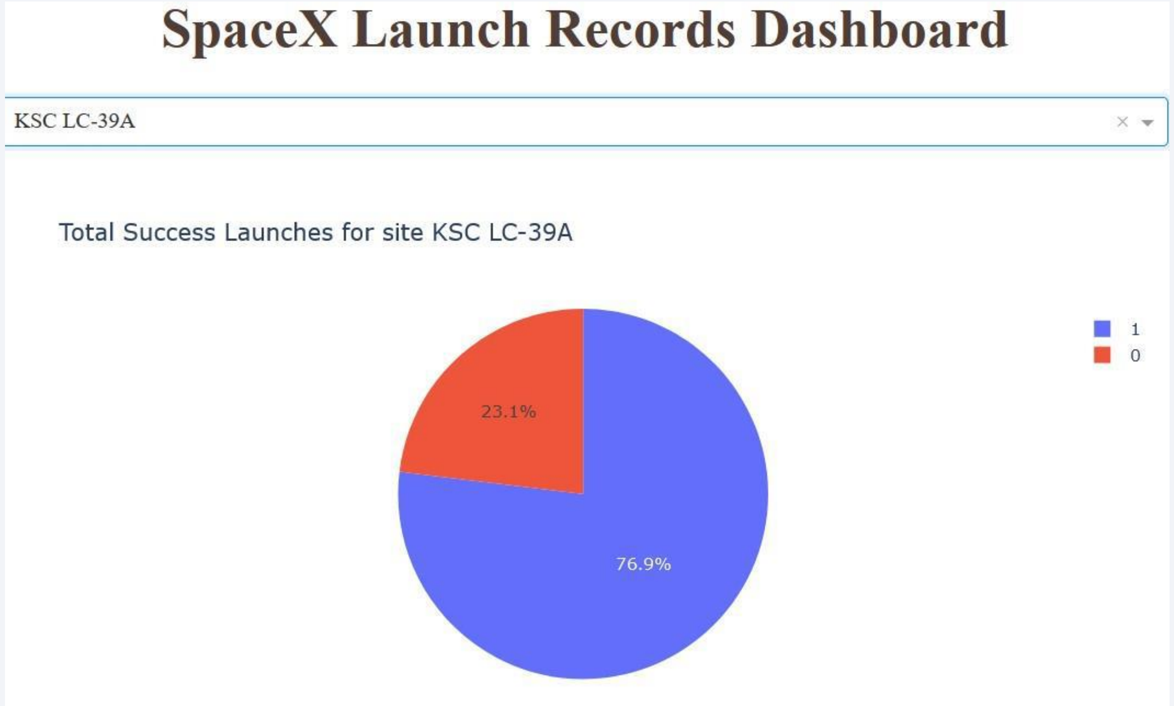


https://github.com/hiepdv/My-submission-Applied-Data-Science-Capstone/blob/0f06c656917dee12b9ea23e824bca77308310a4a/LaunchSite_Location.ipynb

Build a Dashboard with Plotly Dash

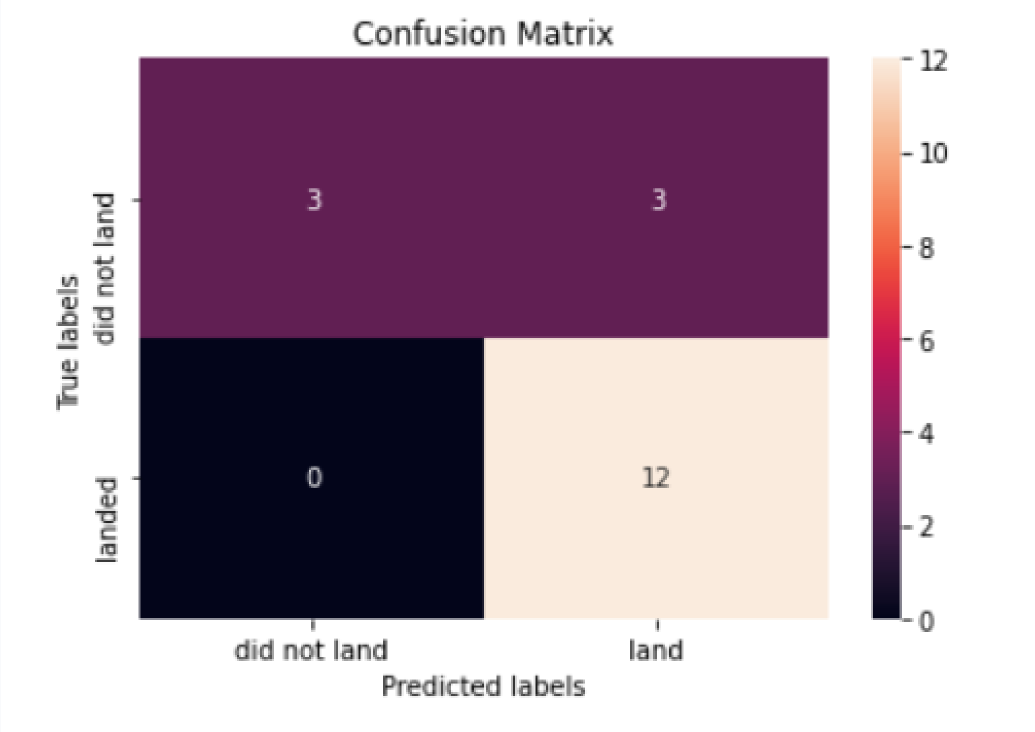
- Interactive visualization of successful launches per site/ all sites





Predictive Analysis (Classification)

- KNN, SVM, DecisionTree, LogisticRegression models with tuned hyperparameters by GridSearchCV were built and evaluated by 10 folds Cross Validation.
- The highest predictive outcome of 83.3% have KNN, SVM and LogisticRegression algoriths



https://github.com/hiepdv/My-submission-Applied-Data-Science-Capstone/blob/b70e93acb021c8f63122cdc03d97288af74d4139/Machine%20Learning%20Prediction.ipynb

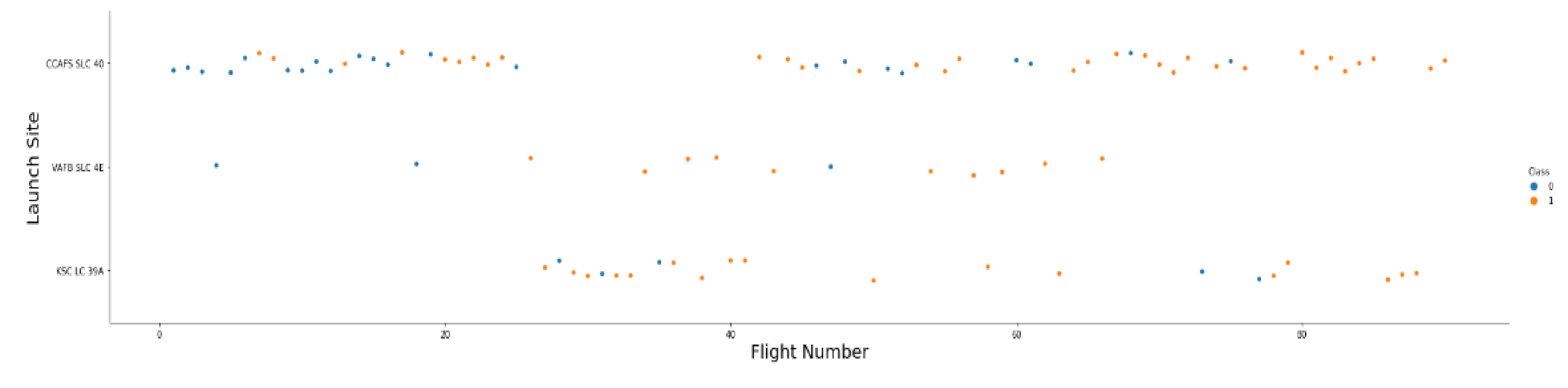
Results

- The most successful rate have ES L1, GEO, HEO, SSO orbits
- Since 2013 successful launches rate increased from 0 to 90%
- For Booster Version FT the optimal payload mass seems to be roughly between 2000 and 4000
- The highest rate of successful launches has KSC LC 38A site
- KNeighbourClassifier , LogisticRegression and SVM
- performed the best on test dataset (83.3% accuracy)

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

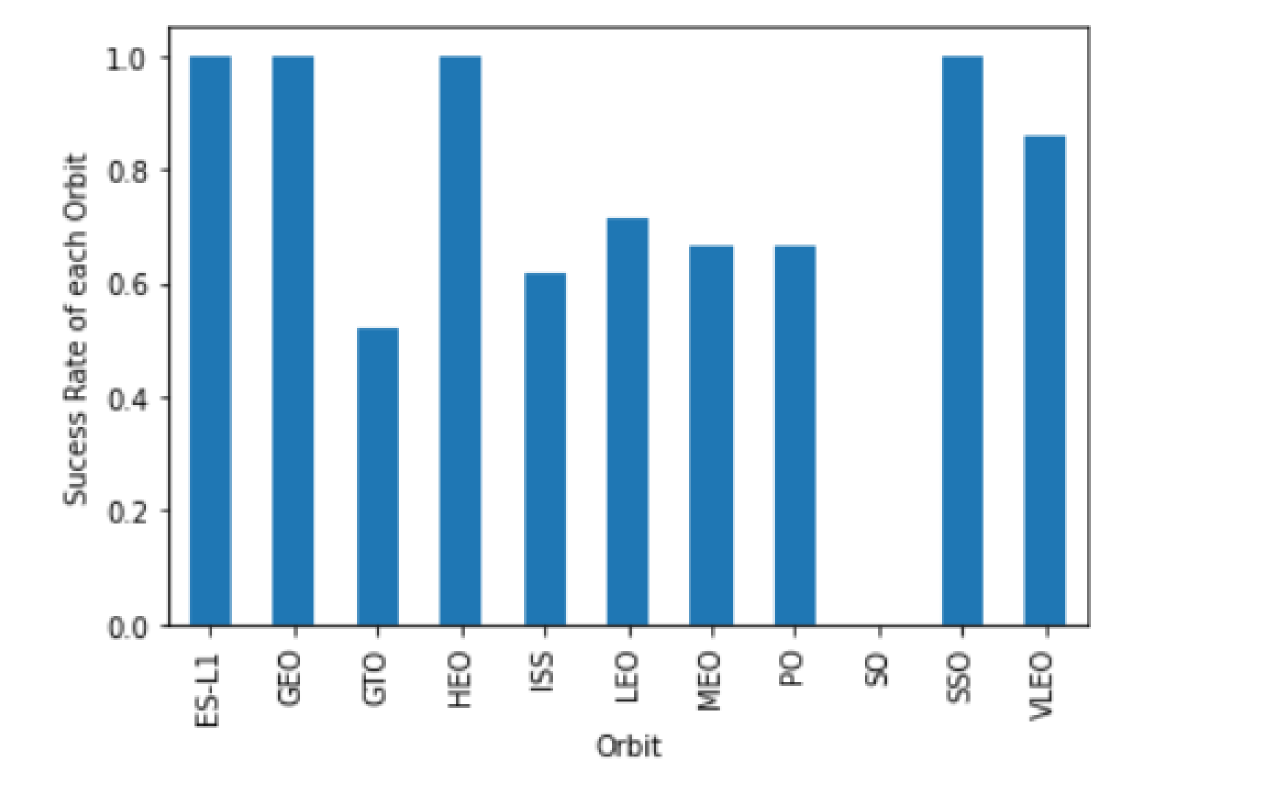


- Main Launche Site is CCAFS SLC 40 site

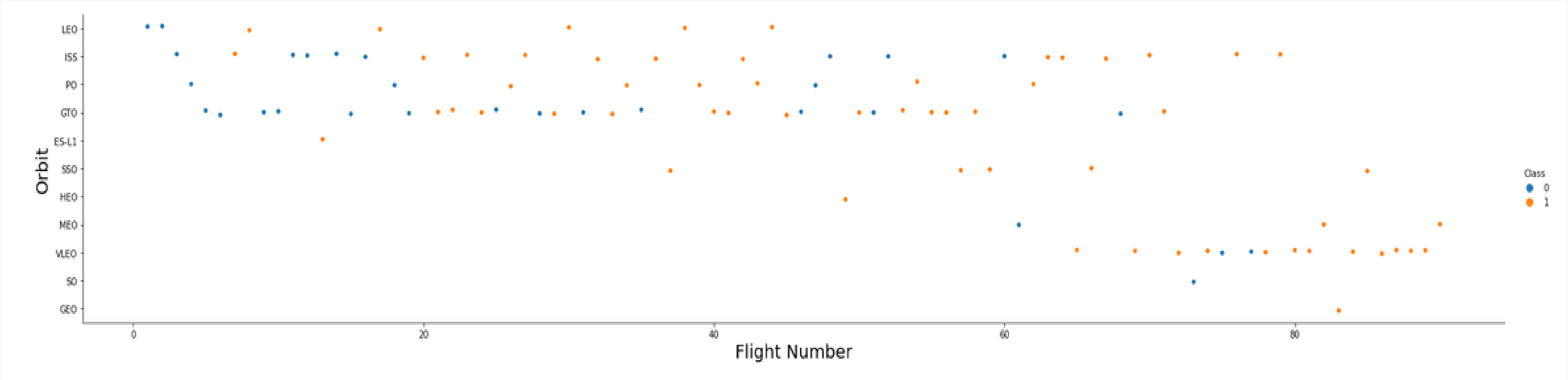
Payload vs. Launch Site

Success Rate vs. Orbit Type

- 4 orbits with high rate of success are ES=L1, GEO, HEO, SSO orbits launches

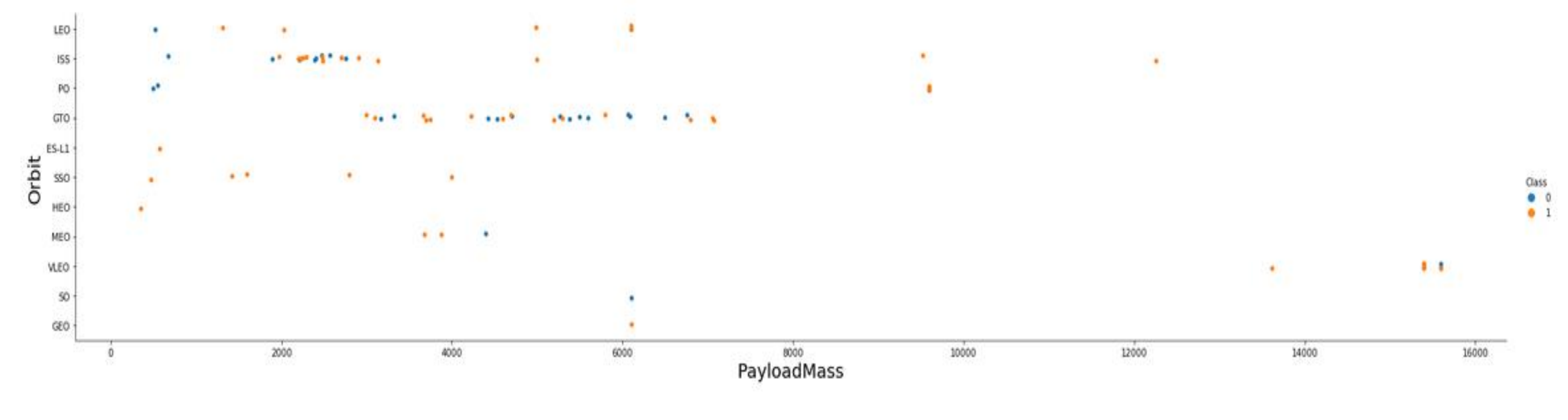


Flight Number vs. Orbit Type



- VLEO orbit gain the highest popularity among all types

Payload vs. Orbit Type



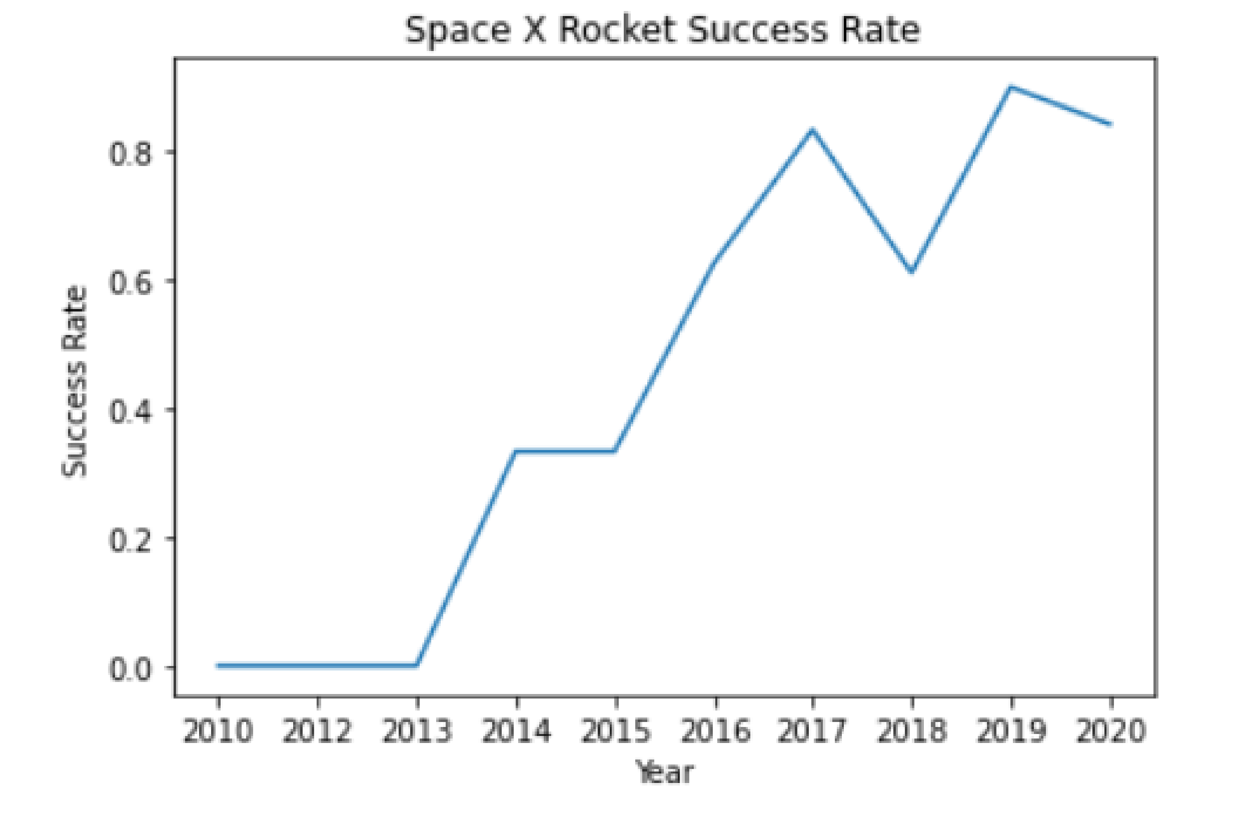
- There are mostly two clusters of payload mass

~1500 3200 (ISS orbit)

~2200 7200 (GTO orbit)

Launch Success Yearly Trend

- From 2013, success rate improved rapidly and kept high rate from 2017



All Launch Site Names

- There are 4 Launch Site

  CCAFS LC-40

  CCAFS SLC-40

  KSC LC-39A

  VAFB SLC-4E

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

Launch Site Names Begin with 'CCA'

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

%sql

SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5

Total Payload Mass

```
%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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

```
%sql SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'
```

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

MIN("DATE")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

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

Done.
```

''''''''

SUCCESS	FAILURE
100	1

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

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

Done.
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

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

2015 Launch Records

```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

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

Done.
```

,,,,,,,,,,,,,,,,

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

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

Done.
```

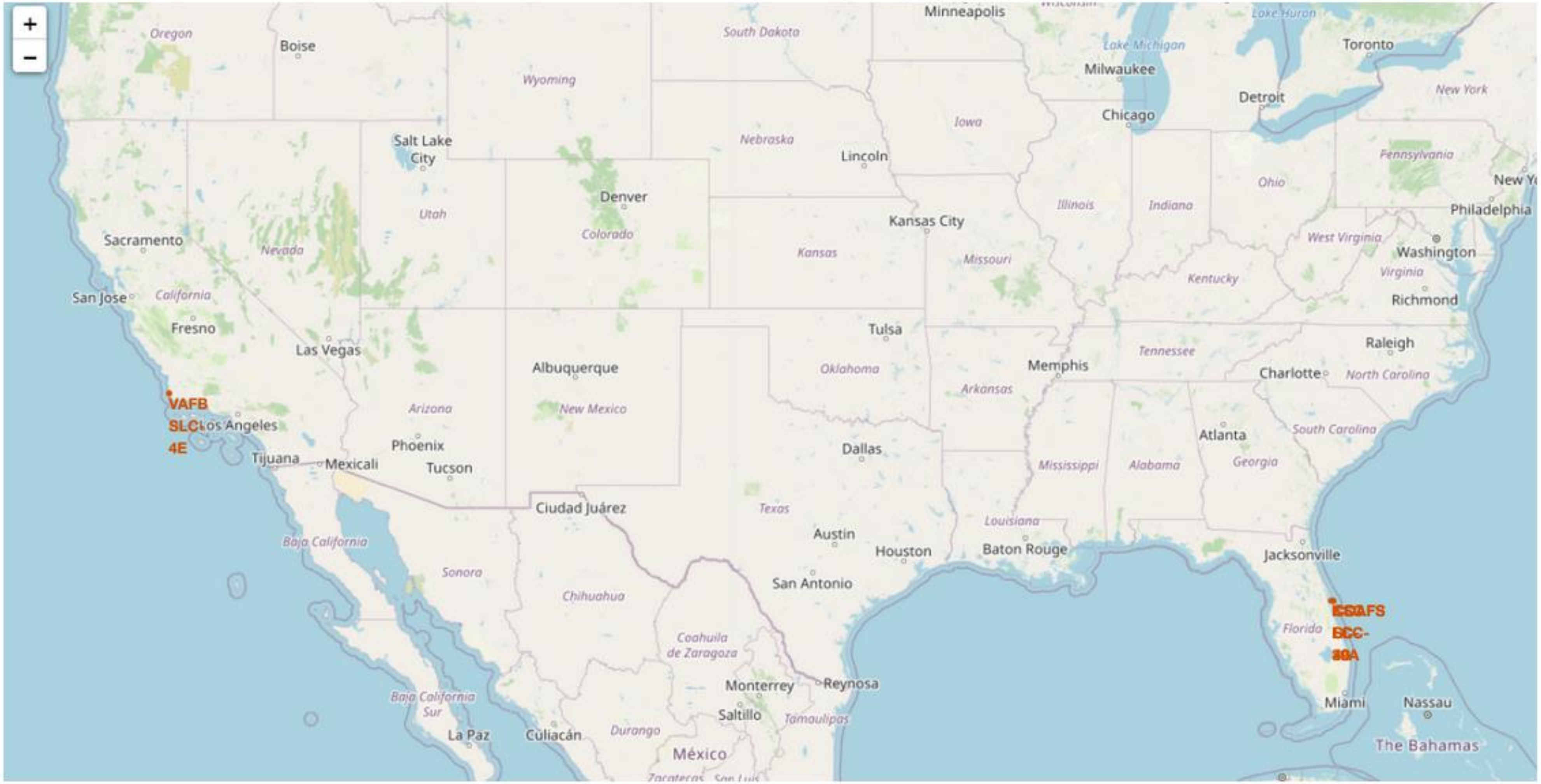
''''''''''''''''''

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

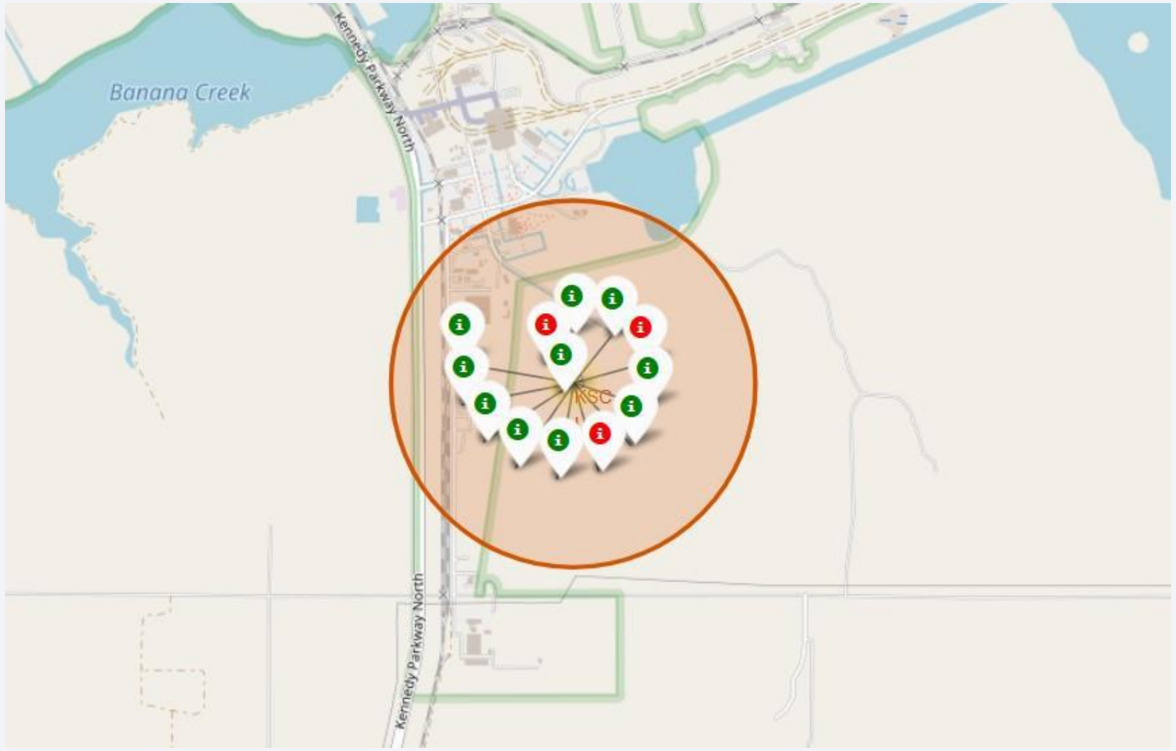
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



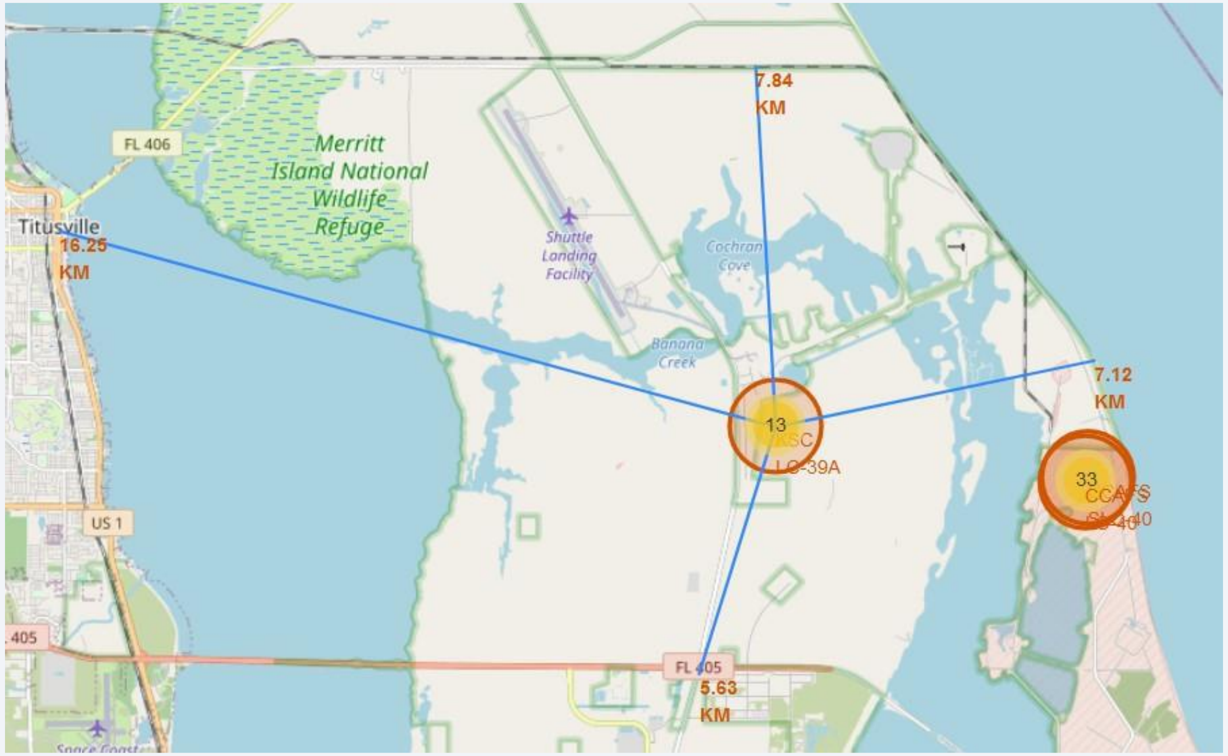
<Folium Map Screenshot 2>

- KSC LC-39A is the most successful site with 10 of 13 successful launches outcomes



<Folium Map Screenshot 3>

- All sites are in a close proximity to coast line and railway (max~7km)

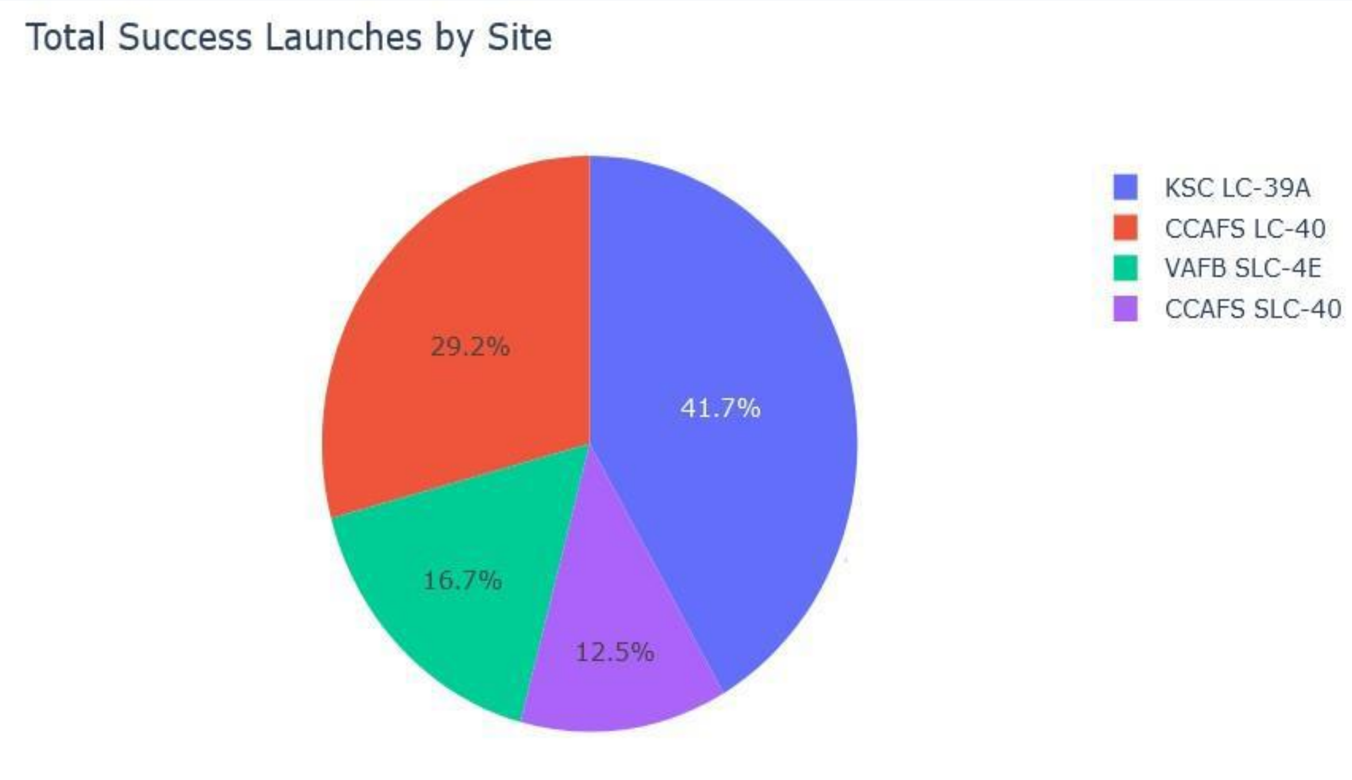


Section 4

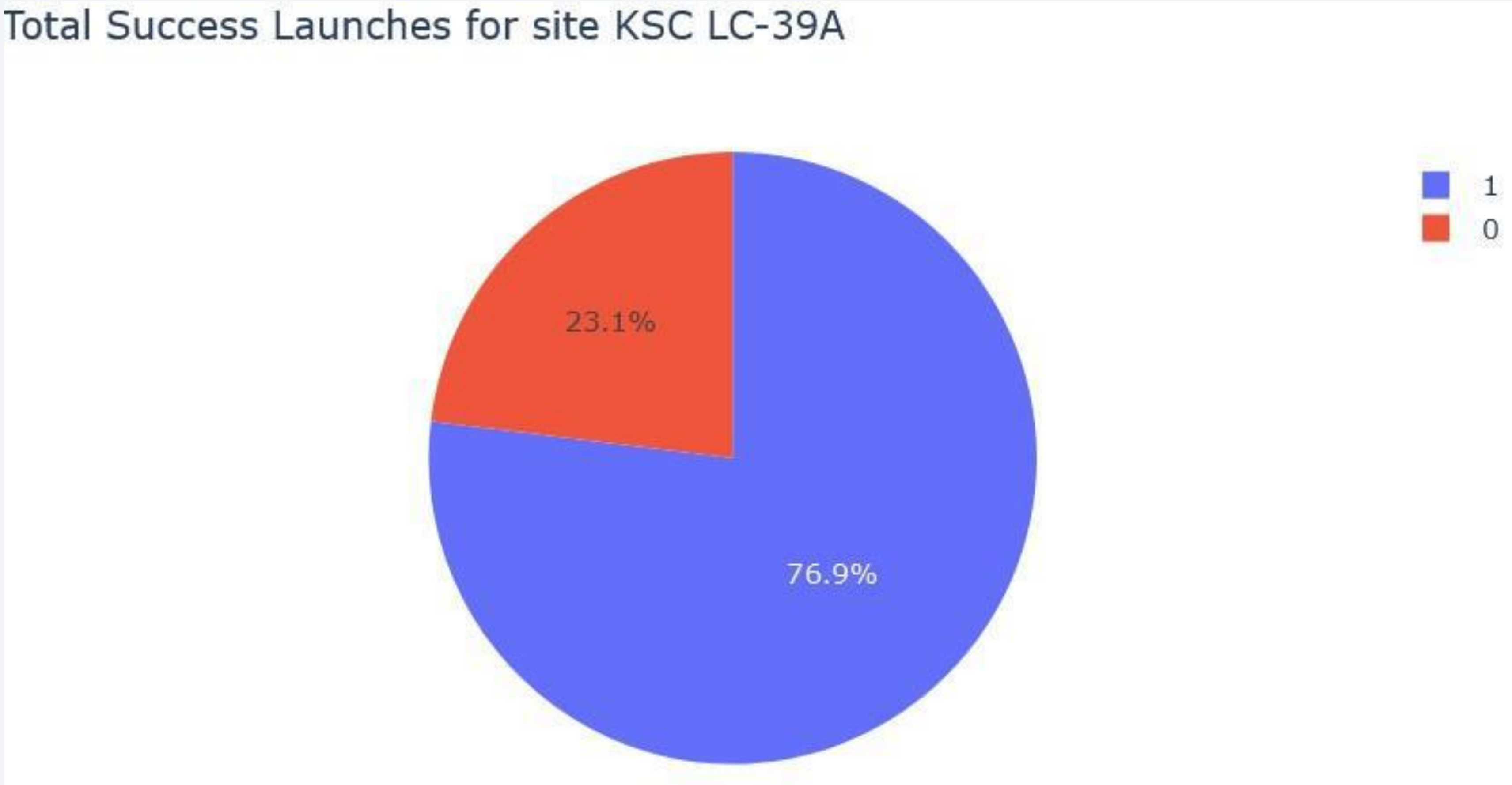
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

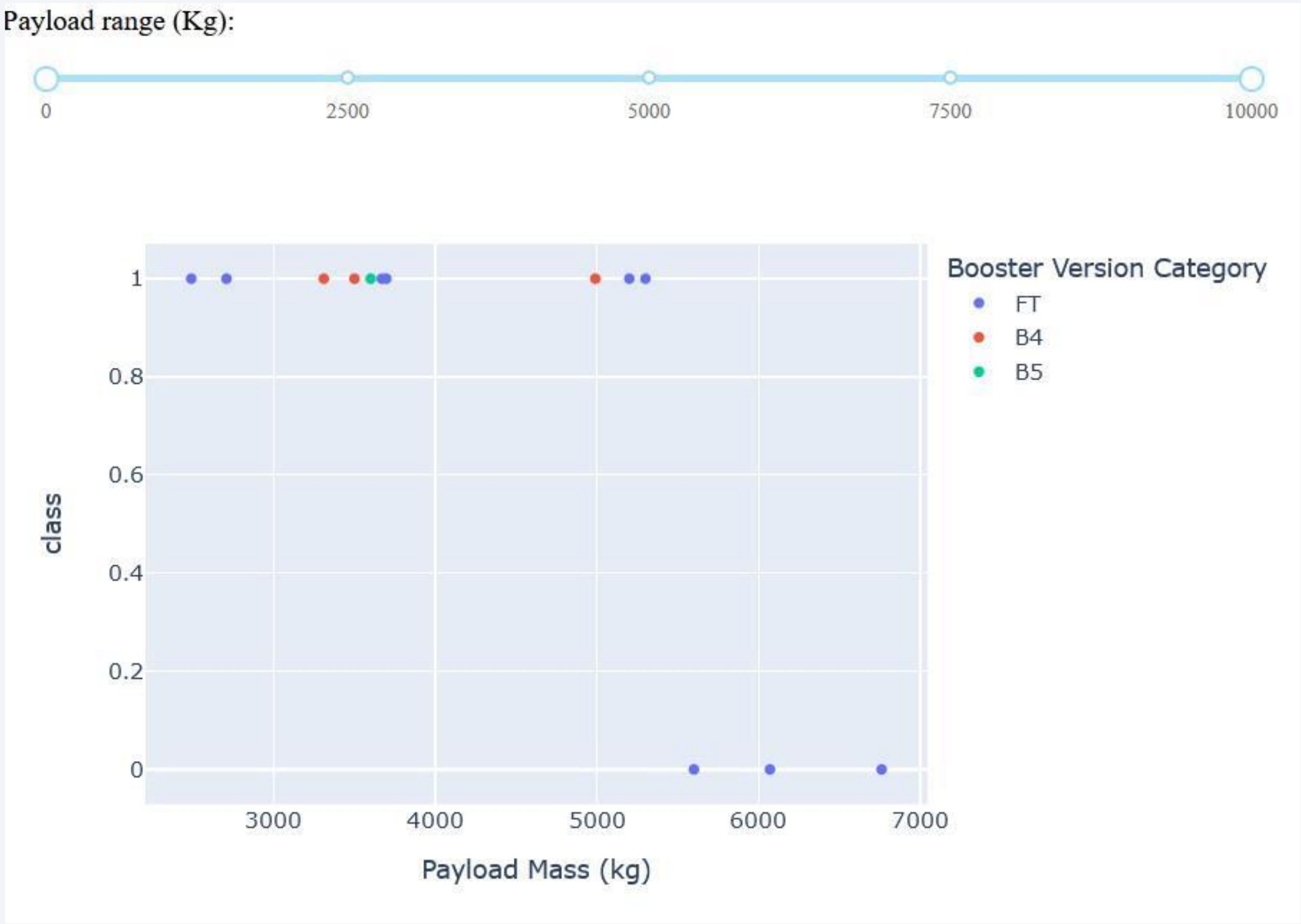
- KSC LC-39A is the higher success rate



<Dashboard Screenshot 2>



<Dashboard Screenshot 3>



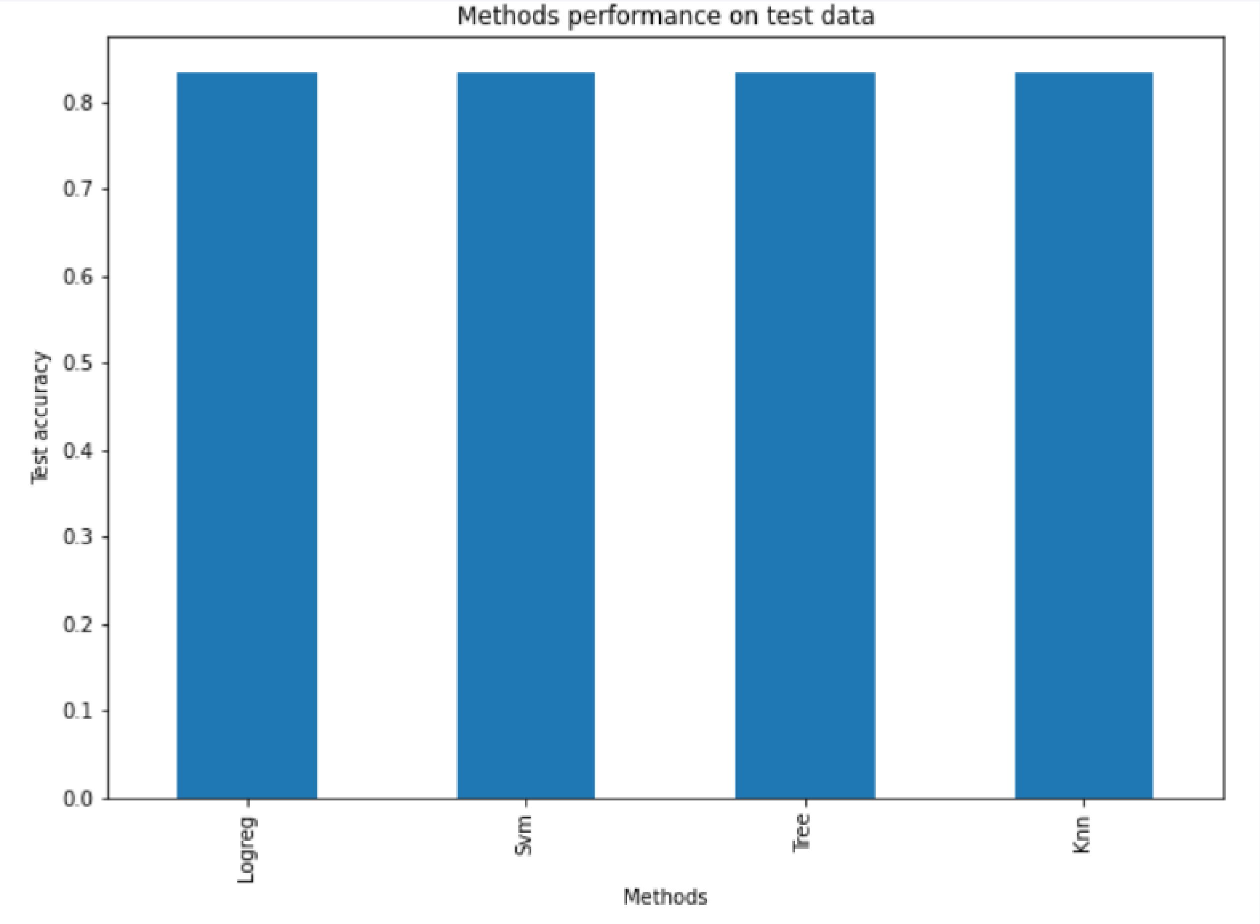
Section 5

Predictive Analysis (Classification)

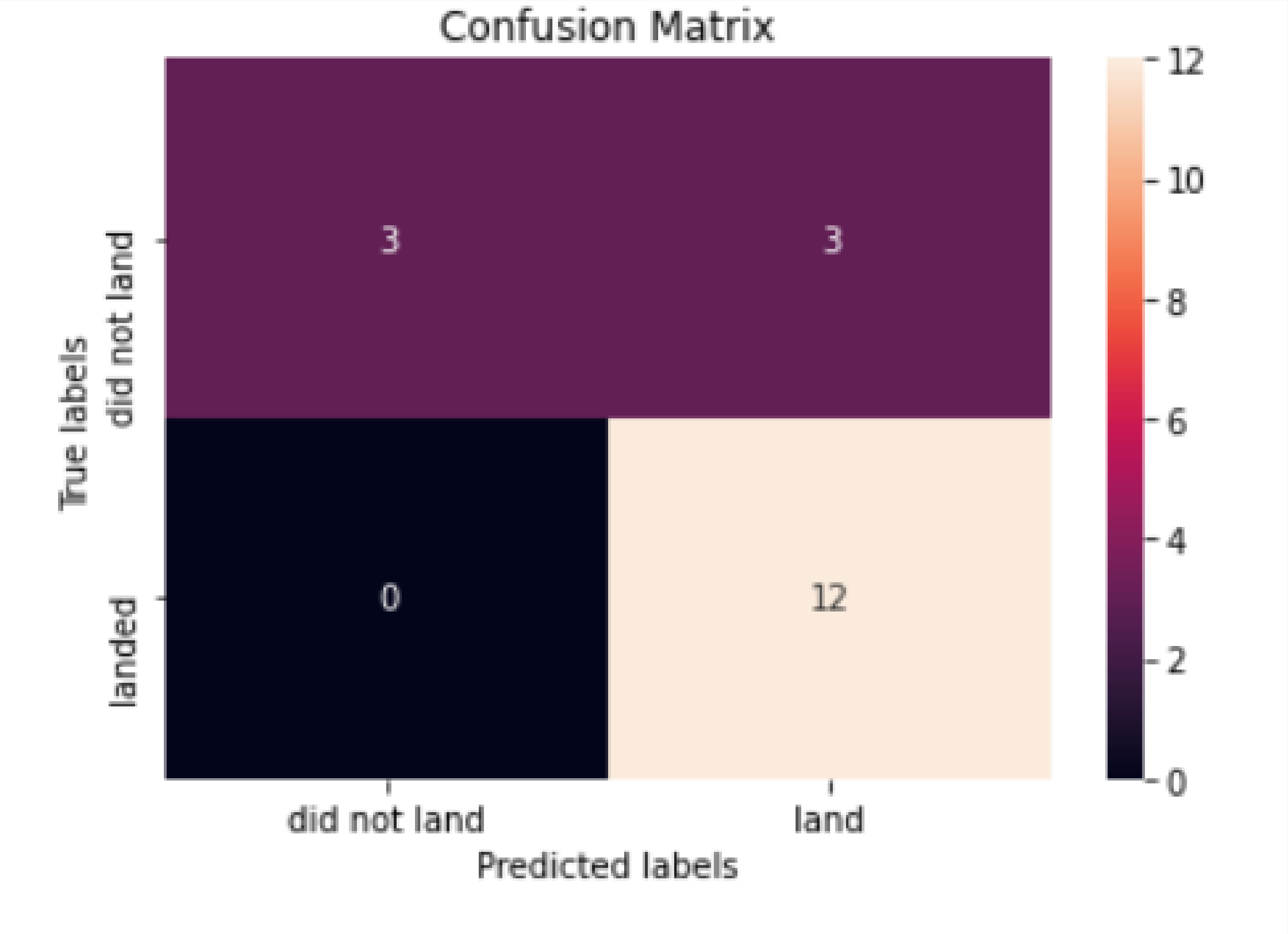
Classification Accuracy

- All models have same accuracy on test dataset

	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.876786	0.833333
Knn	0.848214	0.833333



Confusion Matrix



Conclusions

- The most successful orbit type are ES-L1, GEO, HEO, SSO
- The most successful site is KSC LC-39A (77% success rate)
- Payload Mass lower than 5500 havechances for successful launch
- The best performed Classifier for this project are KNeighborClassifier, SVM, LogisticRegression
- Technologies are constantly developing and from the Launch Success Yearly Trend could be made conclusion that in the future rate of successful launches will continue increasing

Appendix

- GitHub Repo for Capstone project

https://github.com/hiepdv/My-submission-Applied-Data-Science-Capstone

Thank you!