Winning Space Race with Data Science

Shashank Kukreti
21 March 2022

Outline

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

Executive Summary

- Data collected from the public SpaceX API and SpaceX Wikipedia page. Created Column ‘Class’ which represent the success of landing. Data explored with the help of SQL, Visualized using Plotly, Folium and seaborn and dashboard created with Dash. Cleaned the data by selecting only the relevant columns as Features. Then changed all categorical variable to binary variable. Standardized the data and used GridSearchCV to find best parameters for Machine Learning models. Fit the data into models and then visualized the accuracy.

- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

Project background

- Private players are entering Space race.
- Space X provides the cost of rocket launch as $65 million whereas another companies launches cost $165 million
- SpaceX saves money by recovering the Stage 1 of rocket launch.
- SpaceY wants to compete with SpaceX

Problems

- SpaceY wants us to predict the successful recovery of stage 1 of Rocket Launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected from the SpaceX public API and the SpaceX Wikipedia page
- Perform data wrangling
  - Adding column ‘class’ with value 1 for successful landing and value 0 for rest
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection

.

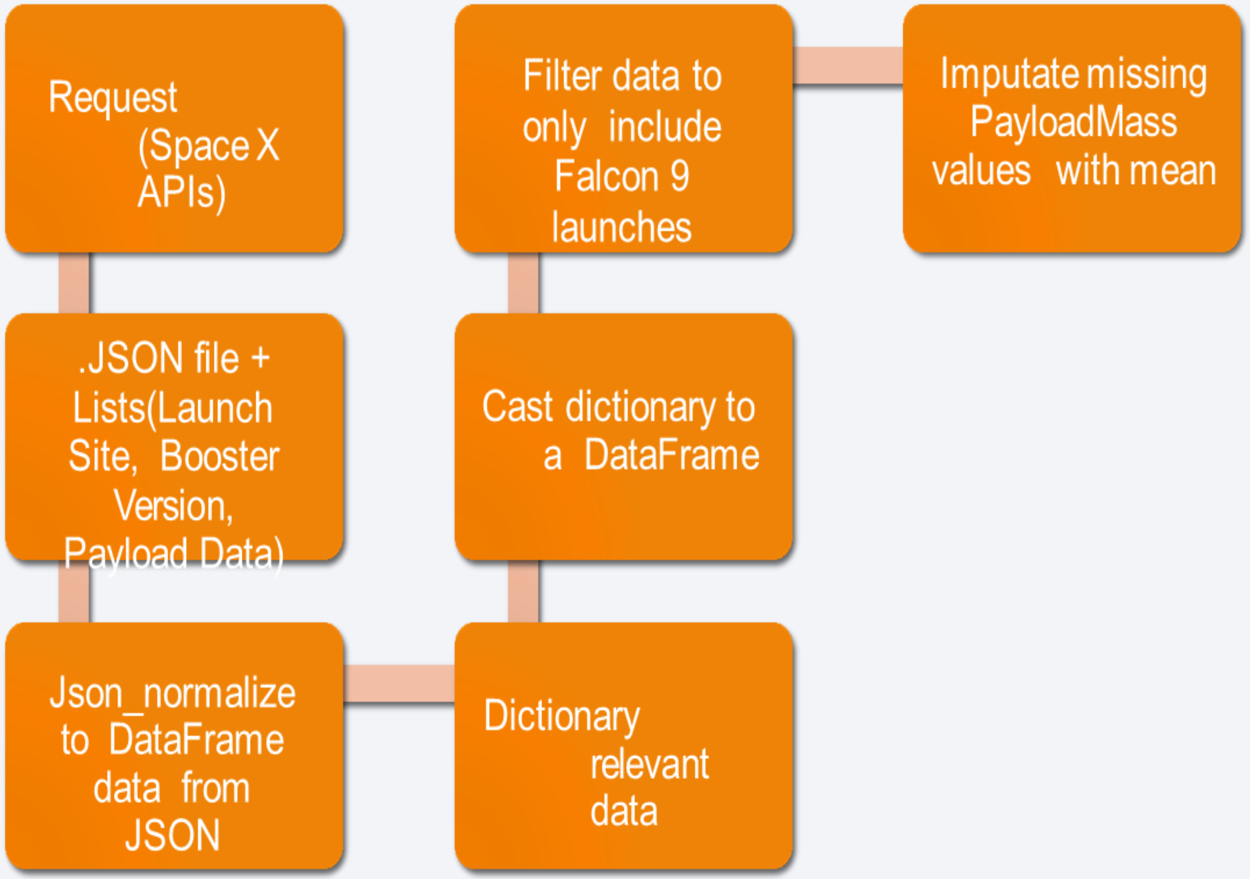
- Data sets collected primary with two process.

1. Making API requests to SpaceX public API
2. Web Scrapping the table in space X Wikipedia page.

Data Collection - SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

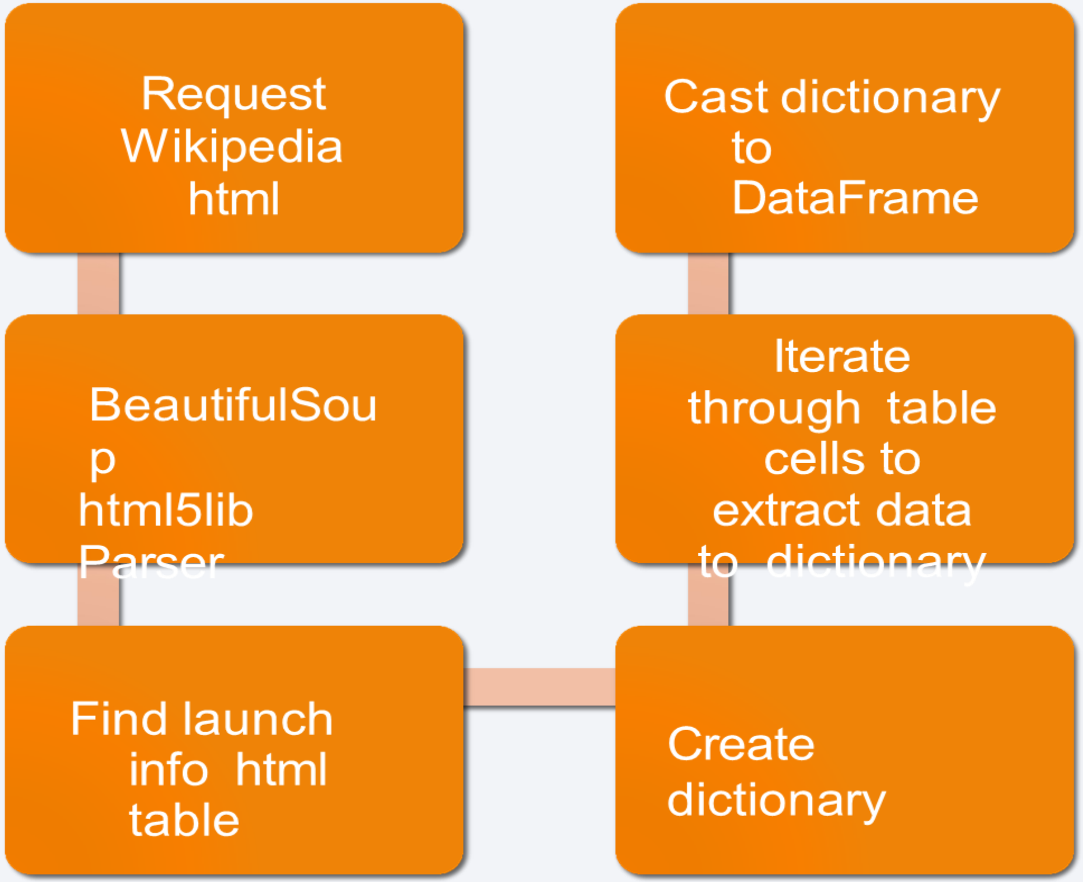
- GitHub url:- https://github.com/Shashank-colab/SpaceX_project/blob/master/spaceXapi.ipynb



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- GitHub url :-[Web scrap]



Data Wrangling

- Create a new column ‘Class’ where successful landing =1 and failure landing = 0
- As outcome column has two components: ‘mission outcome’ , ’landing site’
- If mission outcome is True then column “Class” value is 1 else value is 0
- Mapping
- True ASDS, True RTLS, & True Ocean - set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS - set to -> 0

- GitHub url:-https://github.com/Shashank-colab/SpaceX_project/blob/master/datawrangling.ipynb

EDA with Data Visualization

- EDA performed on various variable Flight Number, Payload mass, Launch site , Orbit, Class and year

Plots used

- Payload mass vs Flight Number, Flight Number vs Launch Site , Payload mass vs Launch site, orbit vs success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
- Scatter plots , line charts and bar plots are used to decide if the relationship exits and independent variables can be used to classify the dependent variable
- GitHub url:- https://github.com/Shashank-colab/SpaceX_project/blob/master/visualization.ipynb

EDA with SQL

- Connect to IBM DB2 database
- select unique launch sites
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

- GitHub url :- https://github.com/Shashank-colab/SpaceX_project/blob/master/sql.ipynb

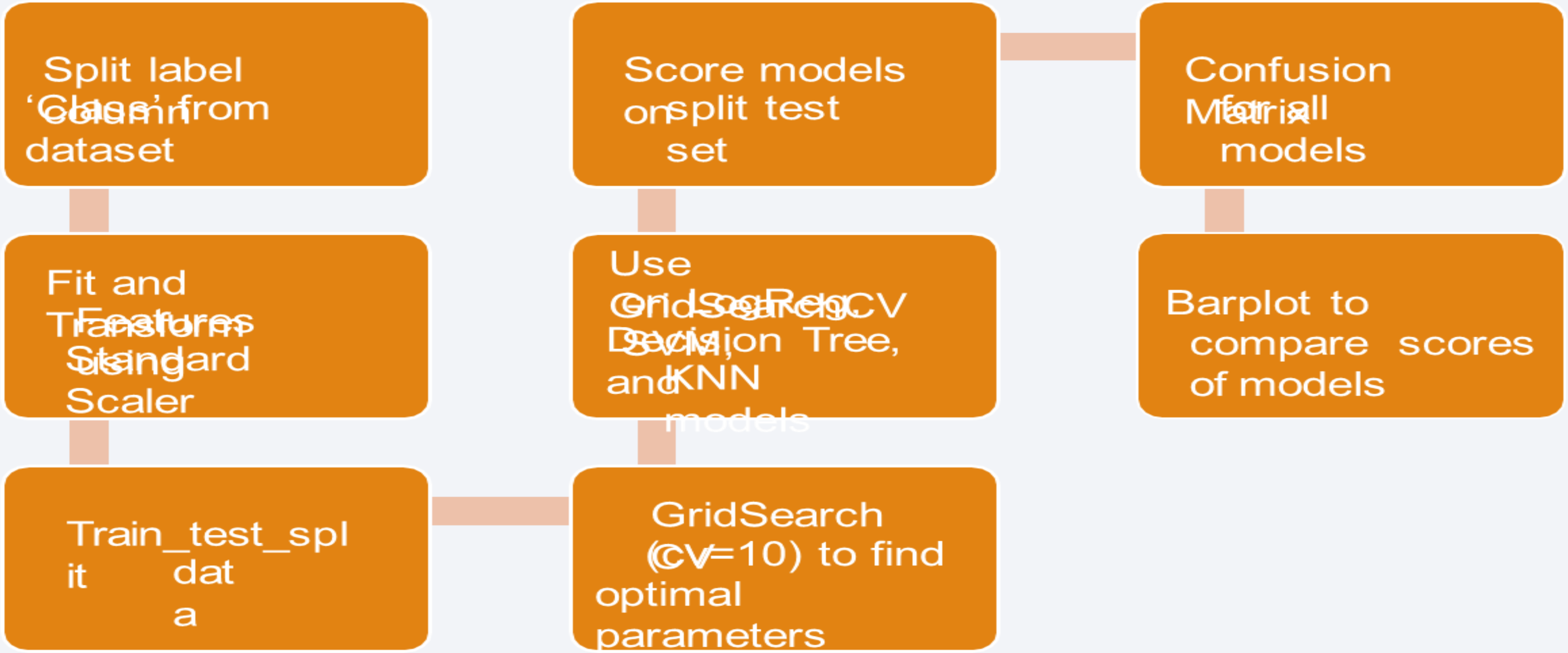
Build an Interactive Map with Folium

- Folium maps marks launch sites ,successful and unsuccessful landing and distance between launch site to key locations
- We mark launch sites to get the clear picture about the geographical benefits of the locations
- Github url :- https://github.com/Shashank-colab/SpaceX_project/blob/master/folium.ipynb

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and Scatterplot
- Explain why you added those plots and interactions
- Pie chart takes the input of launching sites and tell us about the successful landing from that launching site
- Scatter plot takes two input launching site and payload mass it help us to see how success changes across launch sites, payload mass and booster version category
- GitHub :- https://github.com/Shashank-colab/SpaceX_project/blob/master/Dash.py

Predictive Analysis (Classification)



- GitHub url :- https://github.com/Shashank-colab/SpaceX_project/blob/master/ml.ipynb

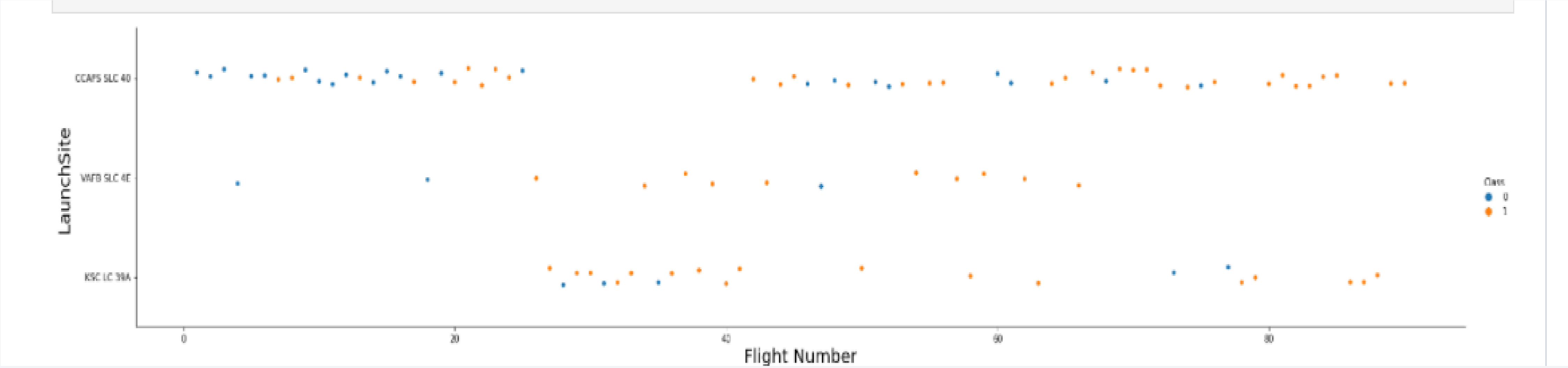
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

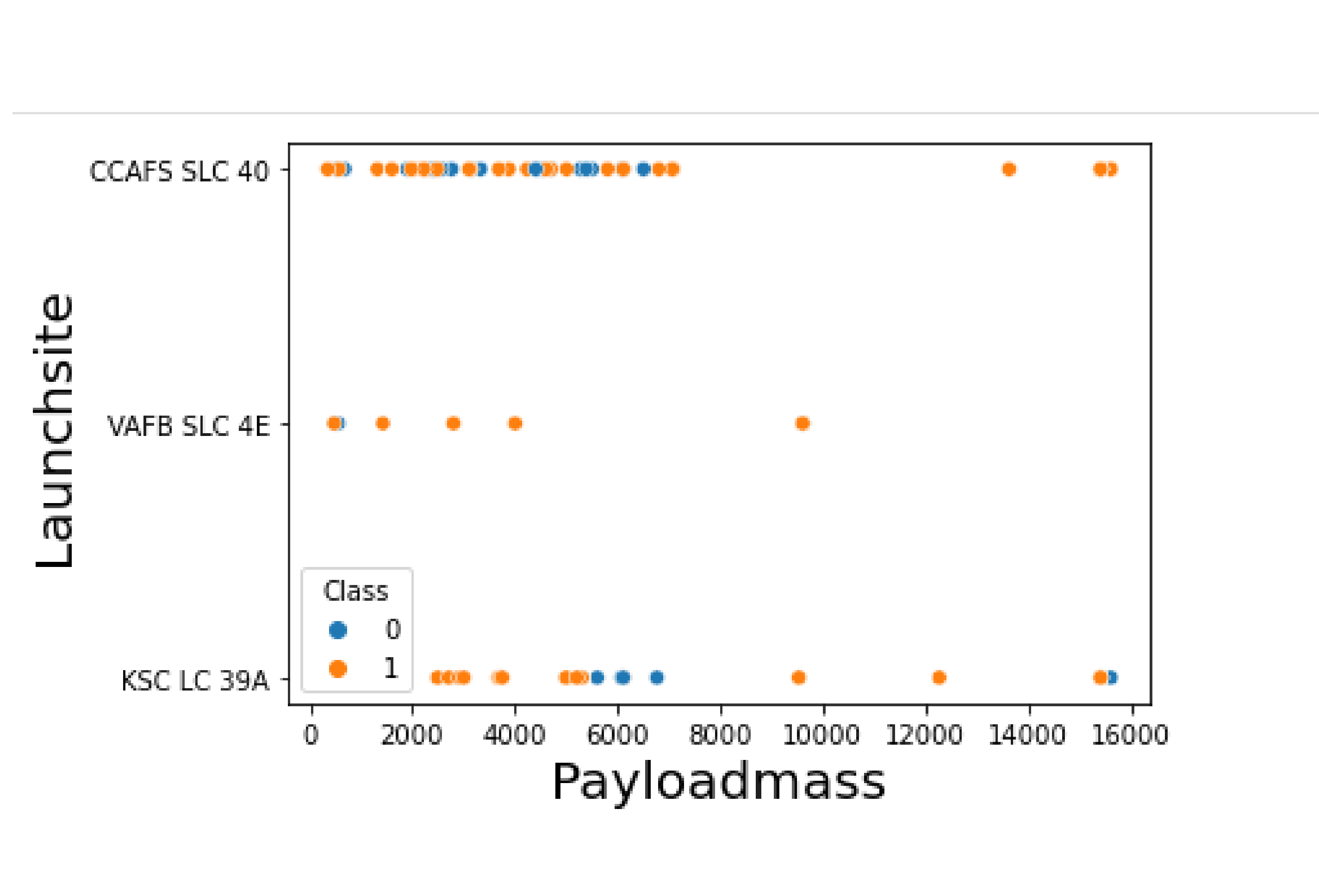
Insights drawn from EDA

Flight Number vs. Launch Site



- Great increase in success rate after 20 launches
- CCAFS is the main launch site

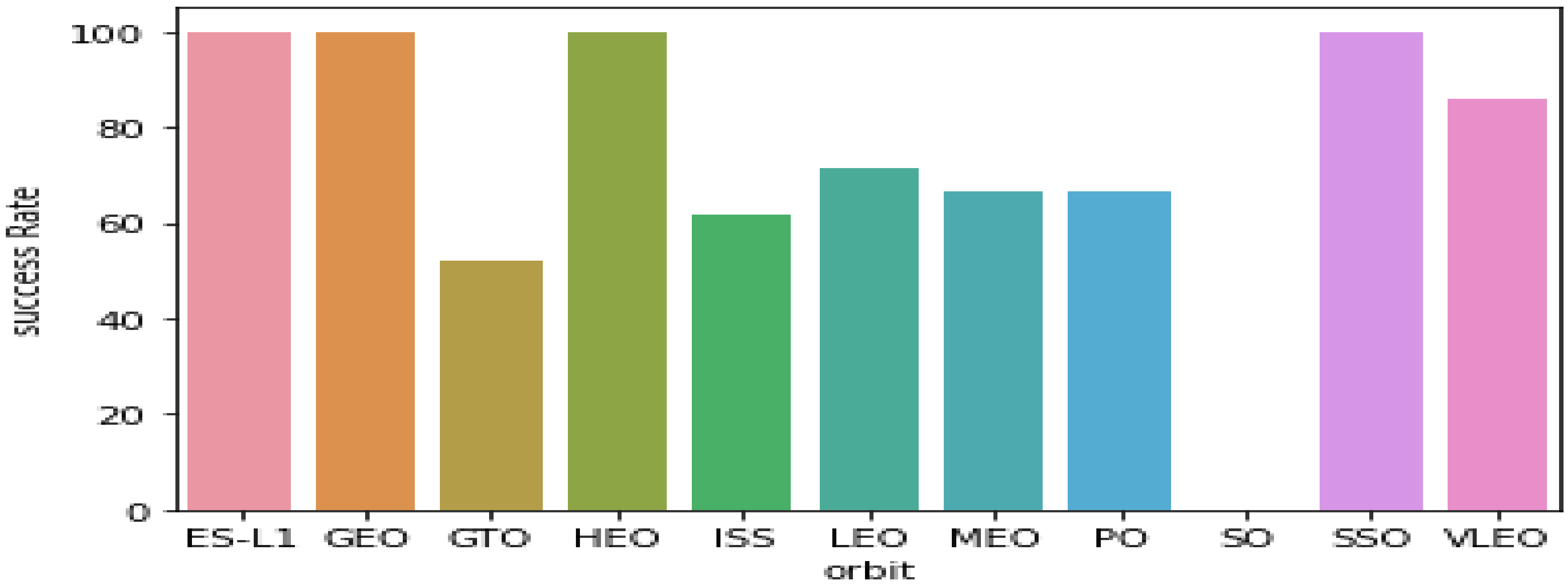
Payload vs. Launch Site



VAFB launch site have a good success rate but less launches

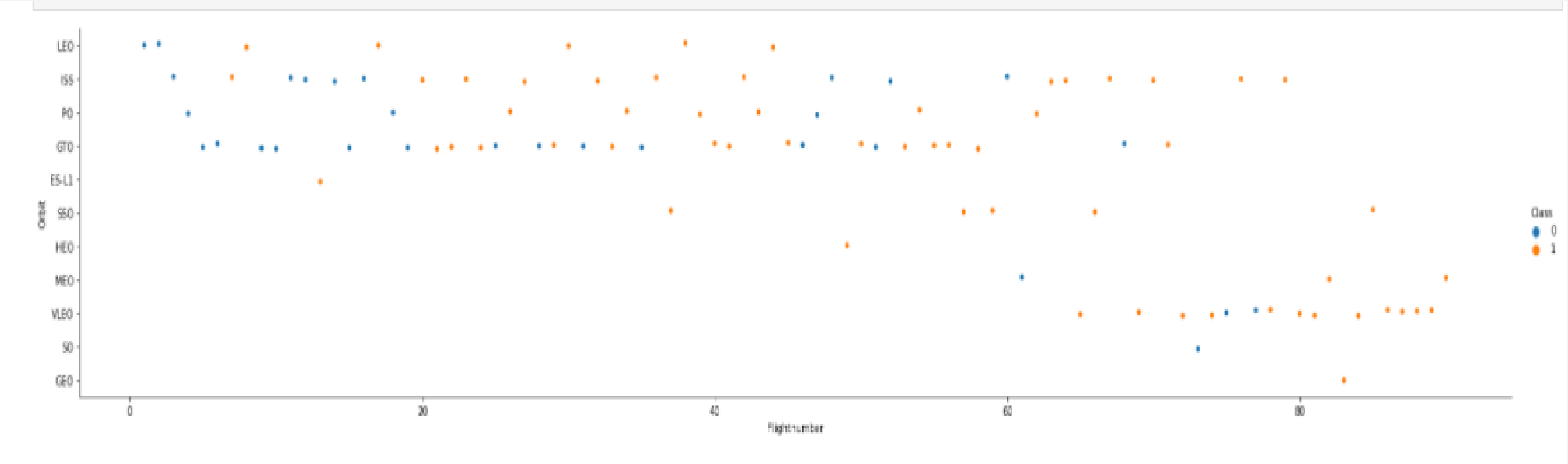
Mostly Payload mass fall in range 0-6000 kg

Success Rate vs. Orbit Type



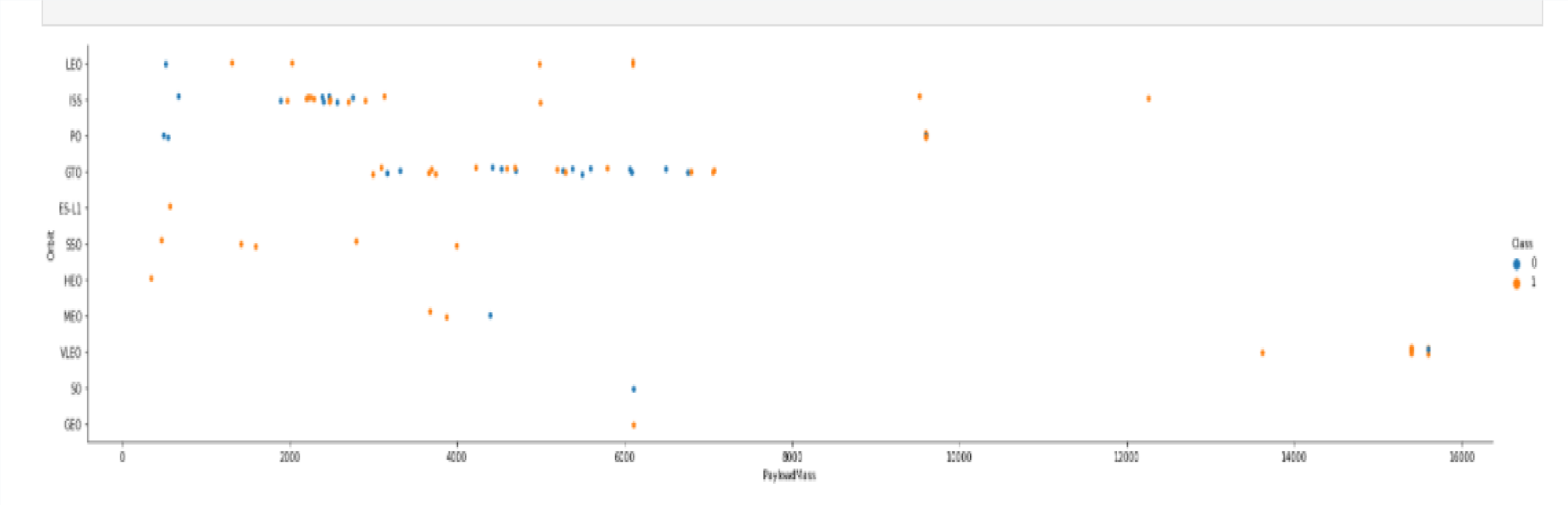
- ES-L1, GEO , HEO ,SSO orbit has 100% success rate
- Where as SO have 0% success rate

Flight Number vs. Orbit Type



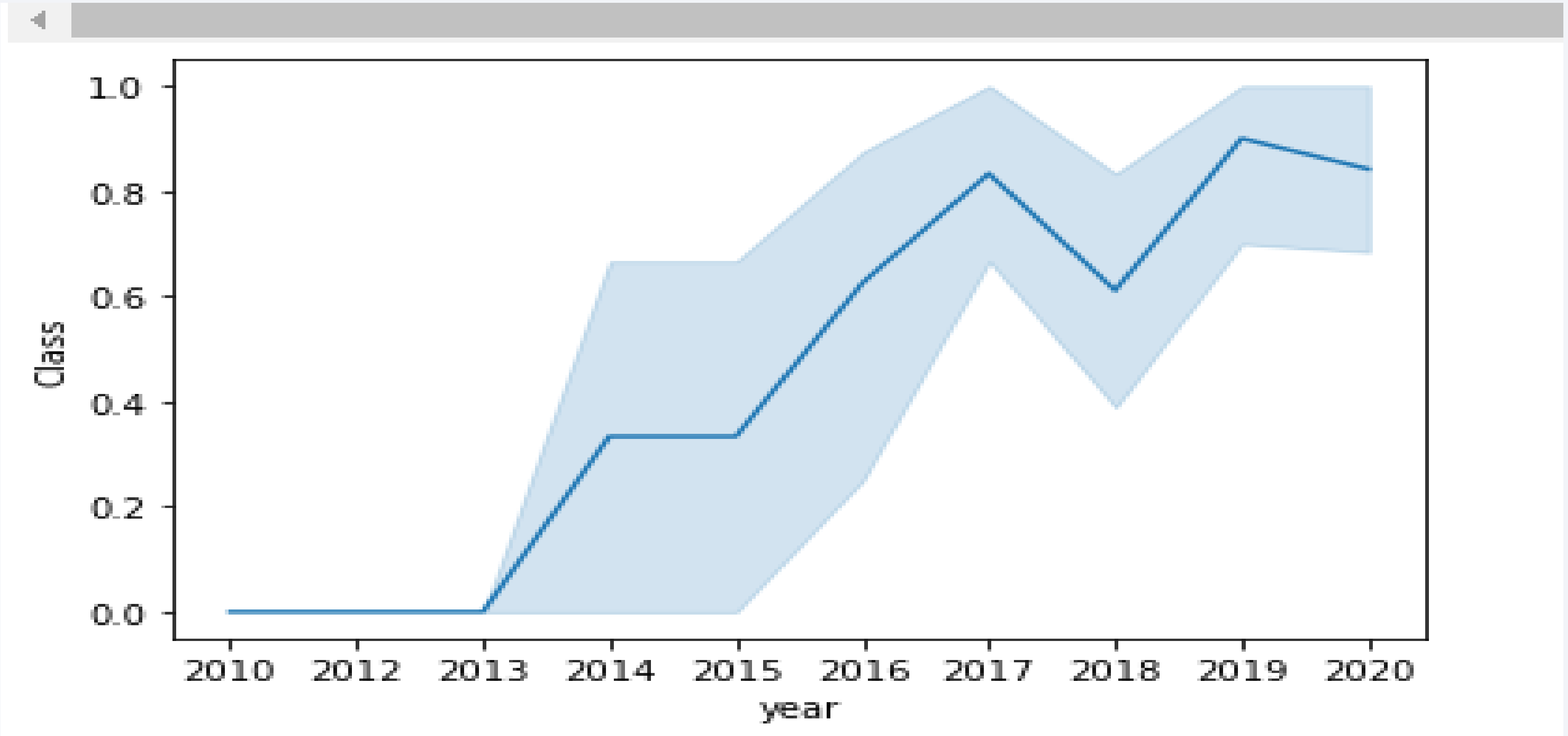
- Launch Orbit Preferences change over flight number
- SpaceX Started with LEO orbits Which Saw Moderate Success

Payload vs. Orbit Type



- LEO and SSO have low payload mass

Launch Success Yearly Trend



- Success is in uptrend since 2013 with slight dip in 2017-2018
- Success rate in recent year is at 80%

All Launch Site Names

In [7]:
```
%sql select distinct(launch_site) from SP
```

```
 * ibm_db_sa://fwx28092:***@764264db-9824
Done.
```

Out[7]:

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

- CCAFS LC-40 and CCAFS SLC -40 are same launch site with data entry error
- There are 3 unique launch sites

Launch Site Names Begin with 'CCA'

In [9]:
```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://fwx28092:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

Out[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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Launch site Names Begin with ‘CCA’

Total Payload Mass from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

In [24]:
```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://fwx28092:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

Out[24]:

1
45596

- Total payload Mass by NASA is 45596 kg

Average Payload Mass by F9 v1.1

Task 4

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

In [28]:
```
%sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://fwx28092:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

Out[28]:

1
2928

- The average Payload mass for booster version F9 v1.1 is on lower end

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

- First ground pad landing was started in 2005

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

- There are 4 booster version that have successful drone ship landings.

Total Number of Successful and Failure Mission Outcomes

In [64]:

```
%sql select mission_outcome,count(mission_outcome) from SPACEXTBL group by mission_outcome
```

```
 * ibm_db_sa://fwx28092:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases
Done.
```

Out[64]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- SpaceX Most flight appears to be successful.
- Out of 101 only 1 flight is unsuccessful.

Boosters Carried Maximum Payload

In [76]:

```
%sql select distinct(booster_version) from :
```

```
 * ibm_db_sa://fwx28092:***@764264db-9824-4b
Done.
```

Out[76]:

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

- The Booster versions are very similar and of F9 B5 B10 variety.

2015 Launch Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- There are two such occurrences where failed Drone Ship Landing Records in 2015.

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

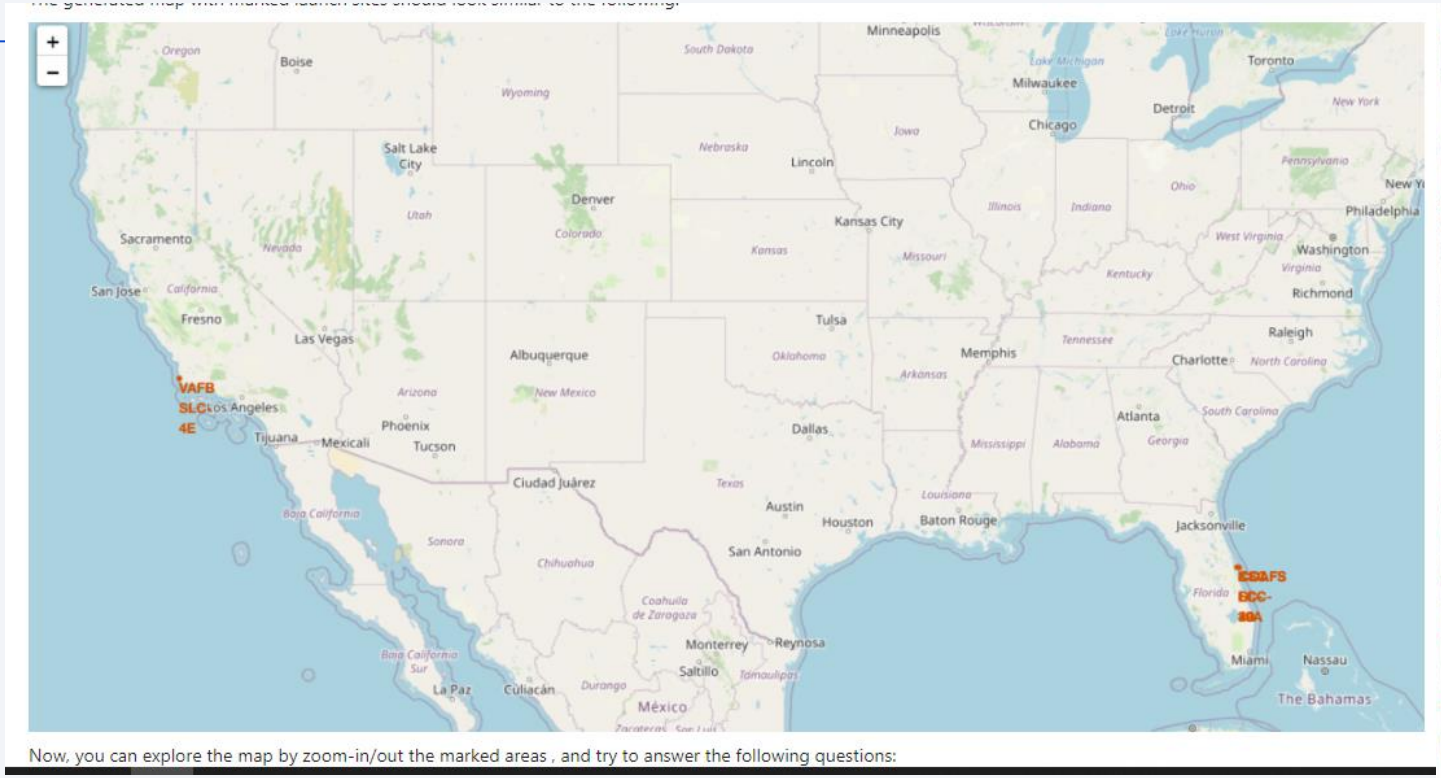
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

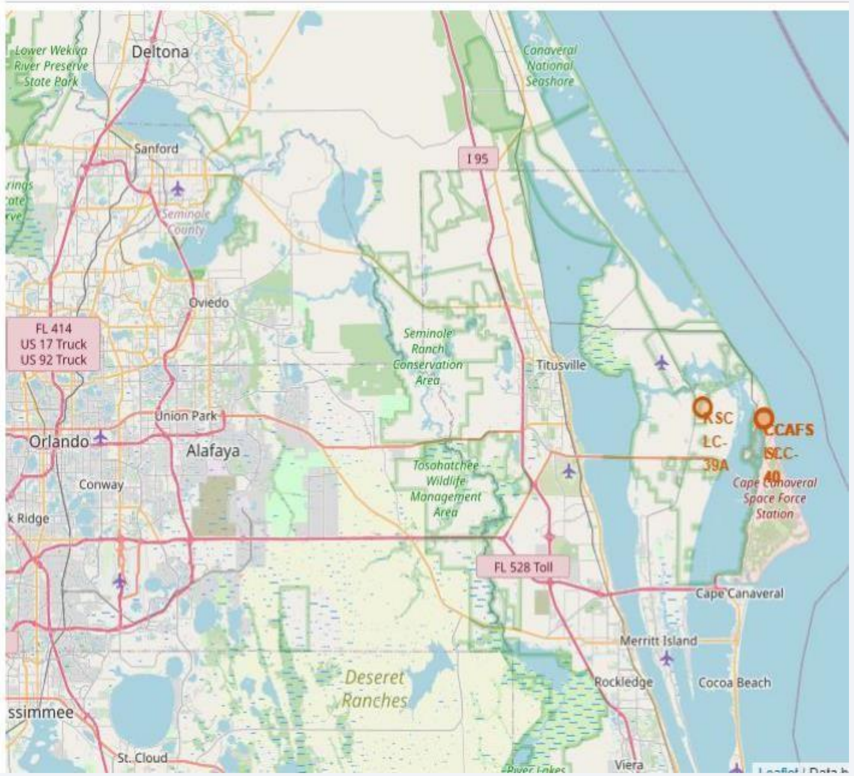
- There are two successful landing outcomes between 2010-06-04 and 2010-03-20

Section 3

Launch Sites Proximities Analysis

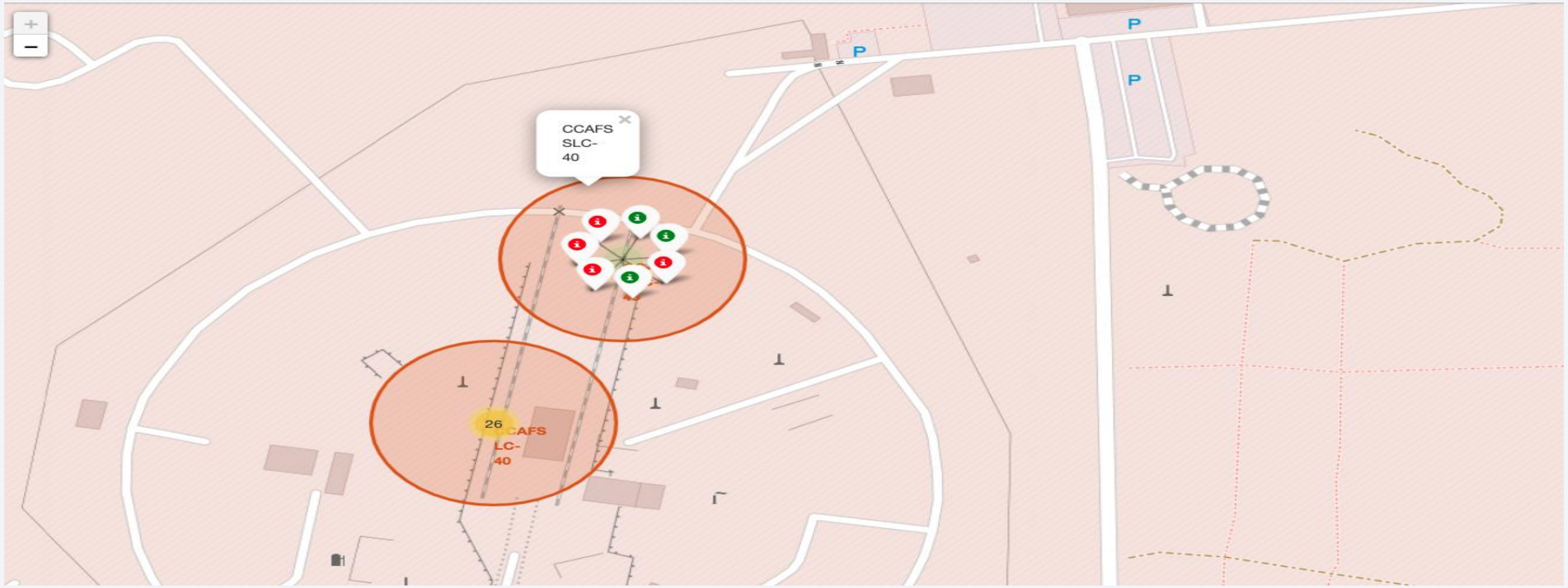
Launch Site Locations





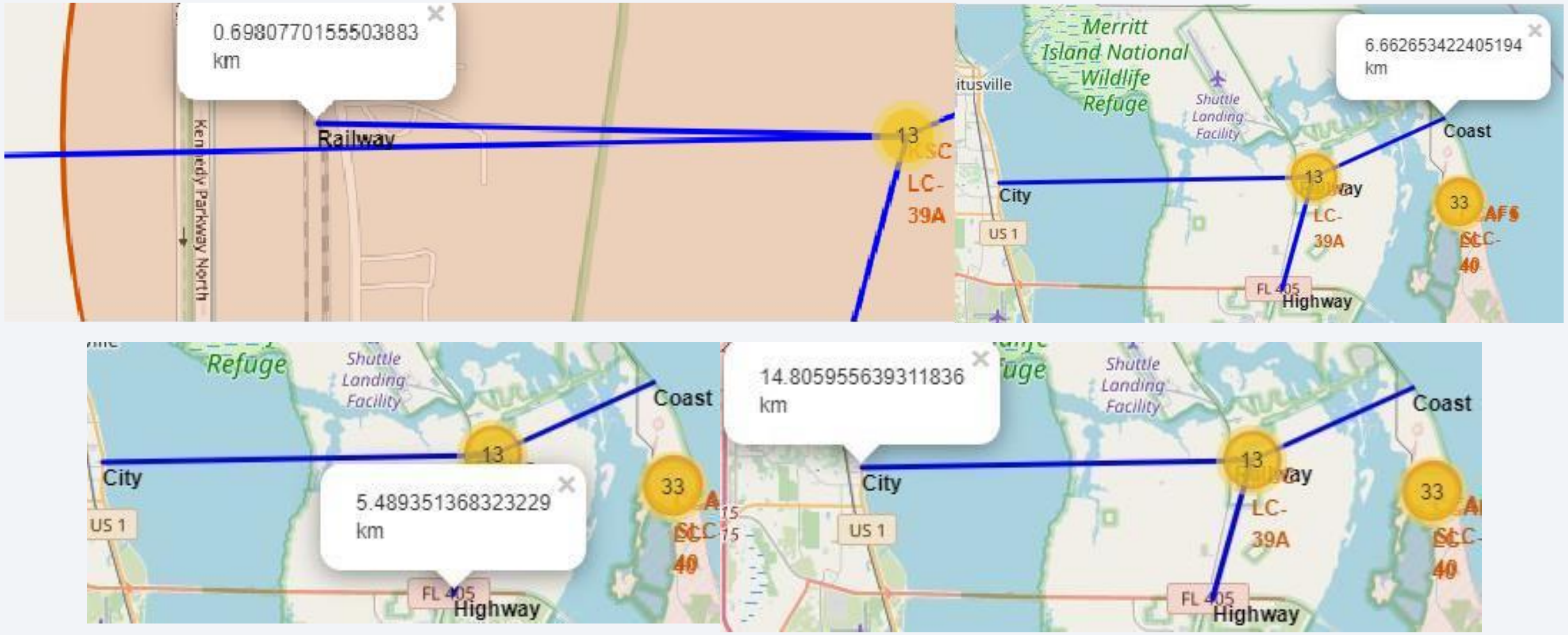
- The left image shows all the launch sites locations in us map
- As two sites are closer to each other so right image shows zoom map

Color labeled Launch



- Clusters on folium map can be clicked and shows successful launch with color and failed launch with red color

Key Locations Proximites

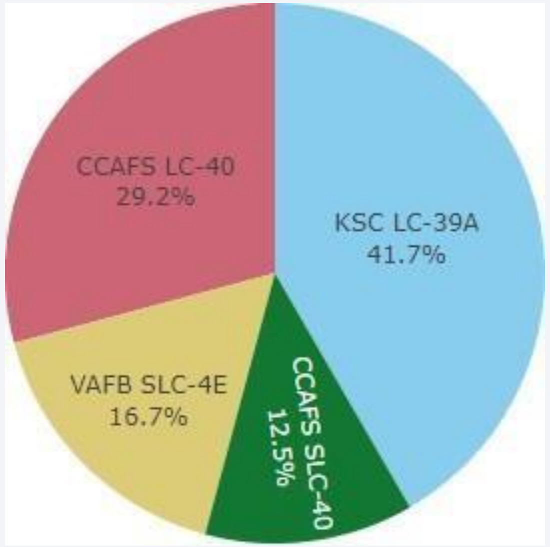


- Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas

Section 4

Build a Dashboard with Plotly Dash

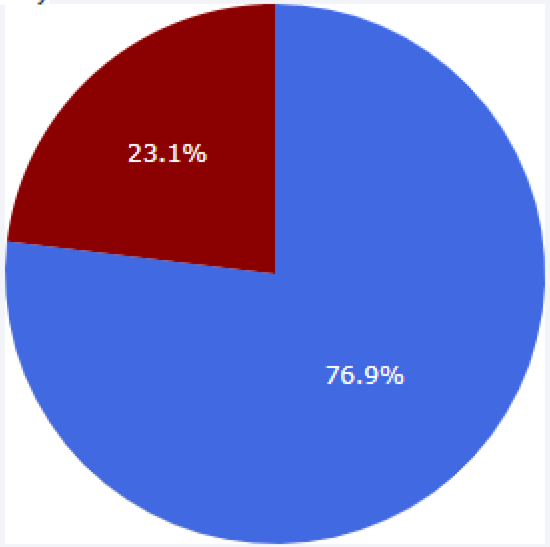
launch success count for all sites



- KSC has the highest count of successful landings.
- VAFB has the smallest share of successful landings

Highest Success Rate Launch Site

KSC LC-39A Success Rate (blue=success)



- KSC LC-39A has the Highest success Rate launch Site

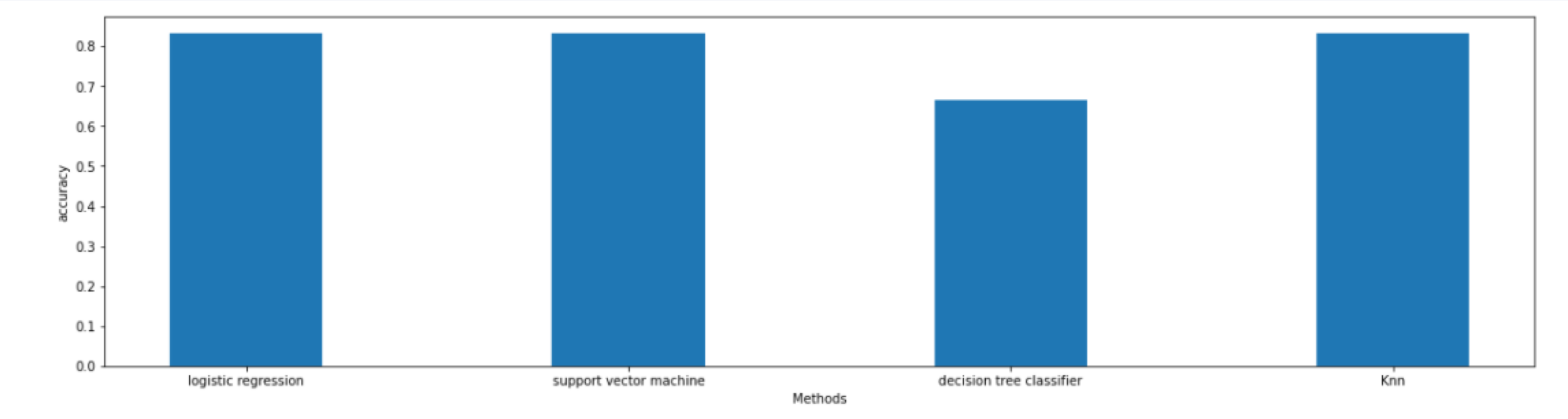
Payload vs. Launch Outcome scatter plot for all sites

- Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg

Section 5

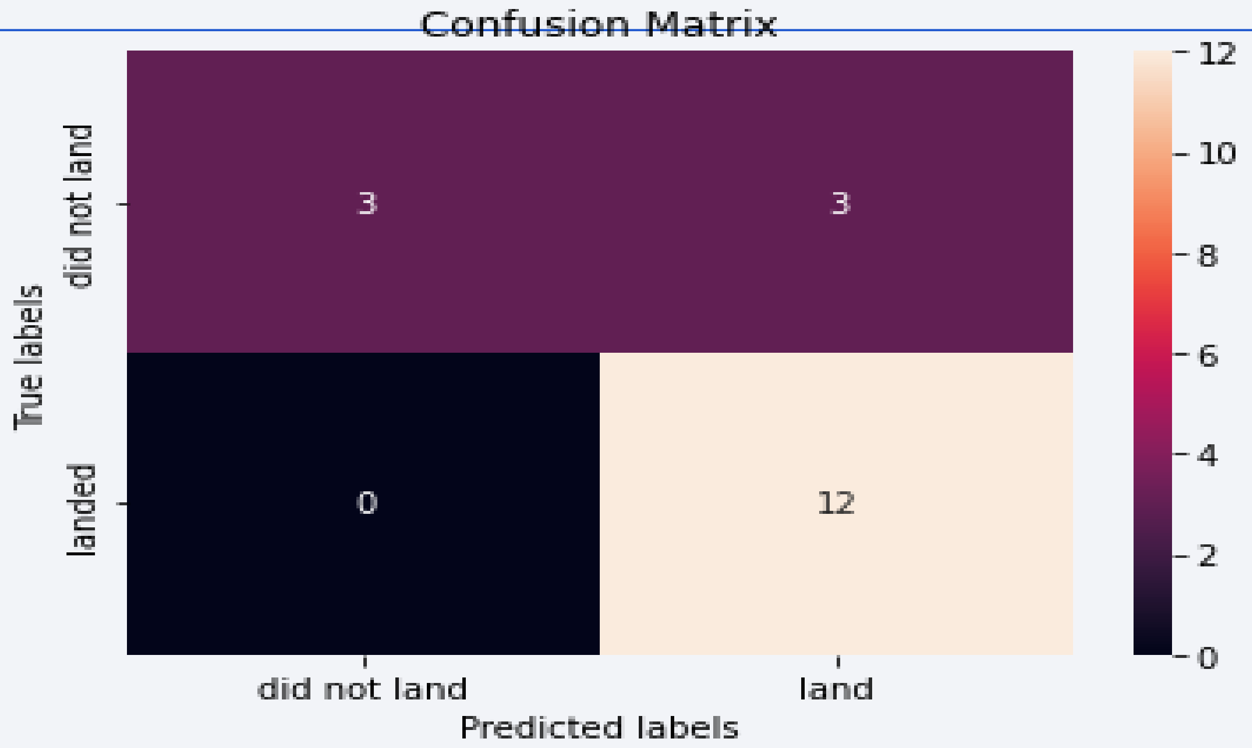
Predictive Analysis (Classification)

Classification Accuracy



- All Models have the same accuracy except decision tree classifier
- The highest accuracy is 83.33%
- We need more data to determine the best model

Confusion Matrix



- Since three models have performed same the confusion matrix is also same.
- Our model predicted 3 unsuccessful land and 12 successful land launch correctly.
- And 3 unsuccessful land incorrectly

Conclusions

- We have develop a machine learning model for SpaceY who wants to bid against the Space X
- The goal is to predict when the Stage 1 will successfully recovered to save $ 100 million
- Collect the Data from SpaceX API and SpaceX Wikipedia page
- Cleaned the data and store it in DB2 data base
- Created a dashboard for visualization
- Created a Machine learning model with 83% accuracy
- SpaceY can use this model to predict the cost for launch and bid against SpaceY
- If possible collect more data to get better machine learning model.

Appendix

- GitHub repository link :-

https://github.com/Shashank-colab/SpaceX_project

- Instructors:
- **Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo**

Thank you!