IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andrzej Szcz
March 24, 2024

Outline

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

Executive Summary

- Summary of methodologies:
    - Data Collection
    - Data Wrangling
    - EDA with Visualisation
    - EDA with SQL
    - Interactive maps (Folium)
    - Dashboard (Plotly, Dash)
    - Predictive analysis
- Summary of all results
    - Preliminary analysis
    - Maps and dashboards
    - Results of predictions

Introduction

- Project background and context

  Our goal is to predict succesfull land of Falcon 9 in first stage. Space X claims that Falcon 9 cost of is 62 milion dollars, however they assume that SpaceX can reuse the first stagege. Therefore they are cheaper than other providers. So based determination if the first stage will land, we can determine cost of a launch. This information is valuable for other comapnies competing wit SpaceX

- Problems you want to find answers:
  - Conditions of successful landing
  - Outcome dependent onf different variables with succes rate

Section 1

Methodology

Methodology

Executive Summary

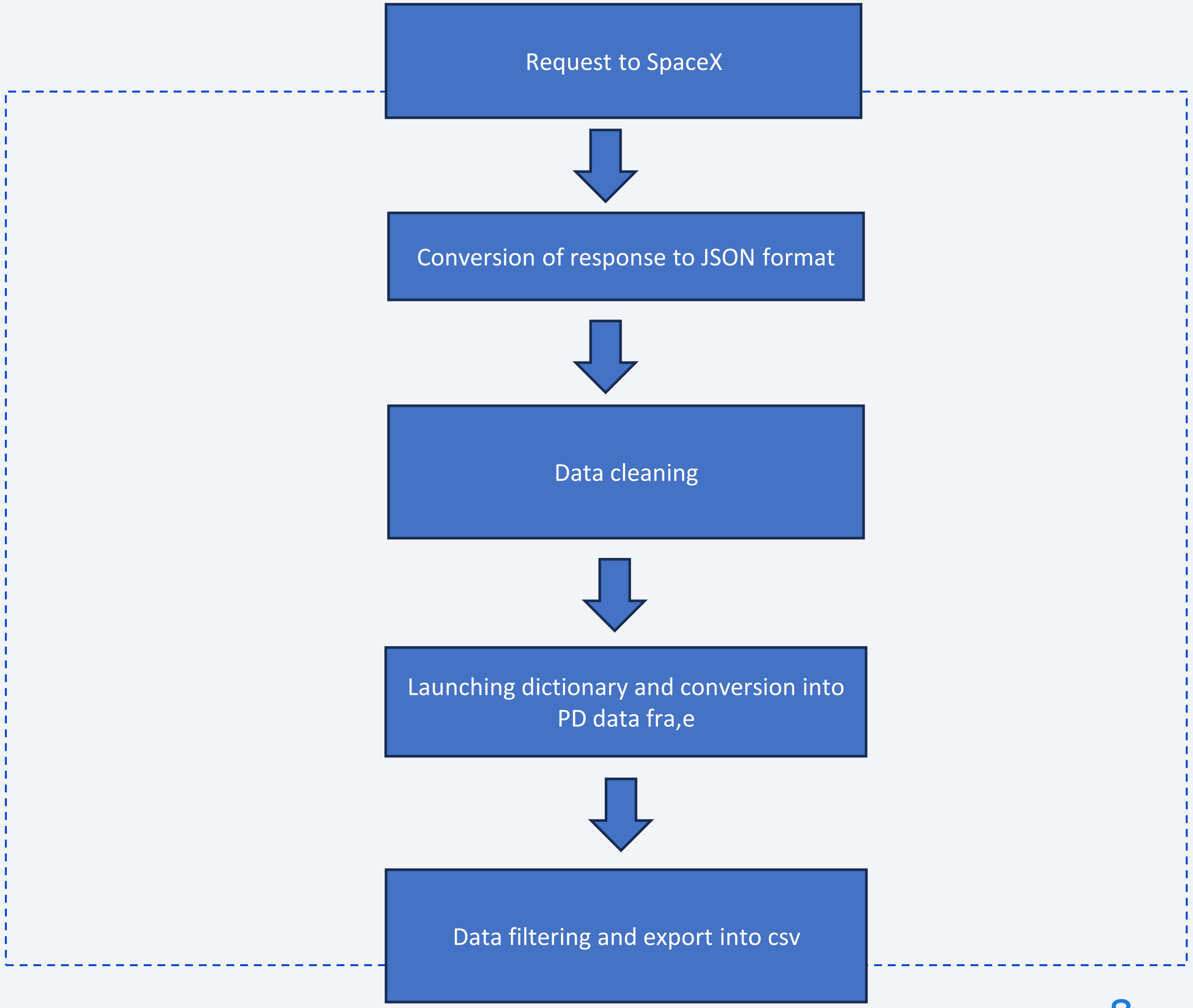
- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping (Wikipedia)
- Perform data wrangling
  - Data cleaned from irrelevant columns and transformed (one hot encoding for ML)
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Scatter, Bar and Line plots to find patterns of data
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data collected through SpaceX Rest Api
- Scope of data includes data about launches, rocket used, payload, landing specifications and landing outcome
- Webscrapping from Wikipedia

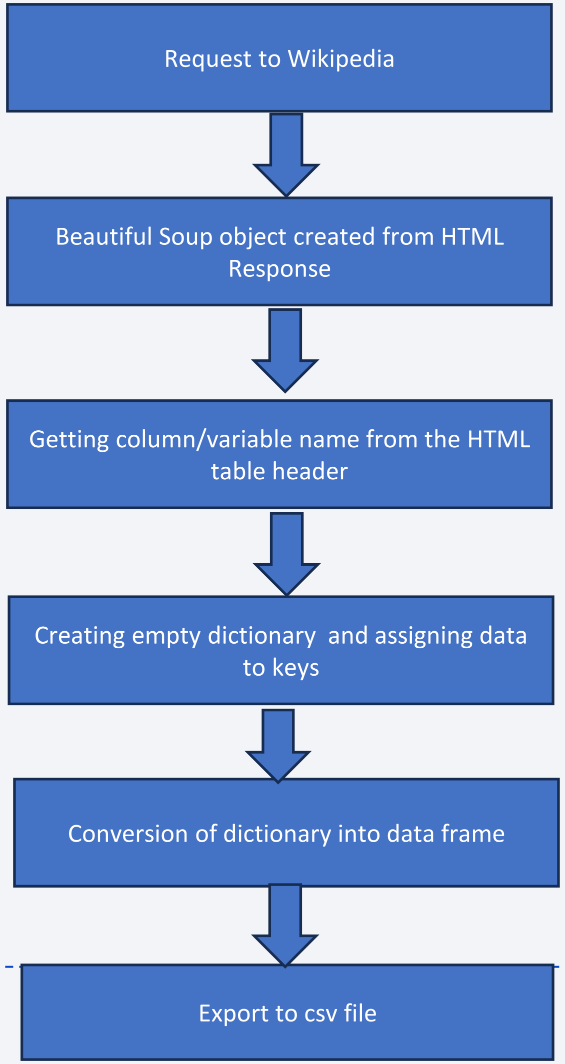
Data Collection – SpaceX API

- Request to SpaceX API and check the data format.
- Clean the data
- Convert the data into csv format
- https://github.com/andszcz/IBM_Data_Science_Final_Project/blob/main/Lab_1_Collecting_the_data.ipynb



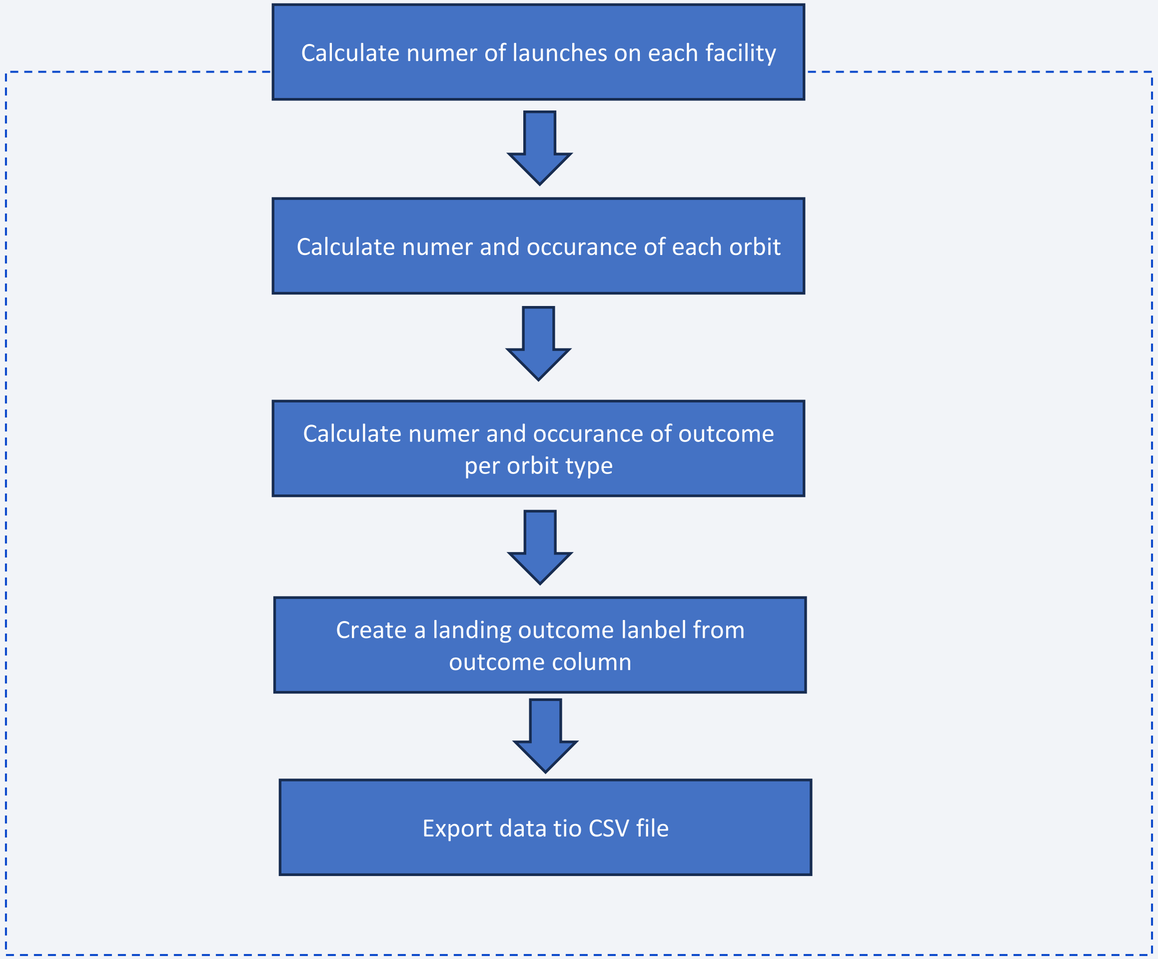
Data Collection - Scraping

- With BeautifulSoup webscrap data fom wikipedia page
- Find tables and get column names.
- Create dictionary, assign data and convert it into dataframe.
- Export data to csv file.
- https://github.com/andszcz/IBM_Data_Science_Final_Project/blob/main/Webscraping.ipynb



Data Wrangling

- The aim of data wrangling stage is to find paterns in data.
- The data set contains several different cases where the booster land was succesfull or unsuccesfull (False).
- Eg. Oceans mean landing in ocean, RTLS on a ground pad ASDS on a droneship
- You need to present your data wrangling process using key phrases and flowcharts.
- True means success False means failure.
- Outcomes are converted into Training Labels where 1 is success and 0 is failure.
- https://github.com/andszcz/IBM_Data_Science_Final_Project/blob/main/Lab_2_Data_Wrangling.ipynb



EDA with Data Visualization

- Scatter Plots – to show how different features are correlated.
- Bar Graph – to compare different groups of date (we would like to present which orbits have the highest success rate).
- Line Graph – to present and analyse trends and to make predictions based on it.

- https://github.com/andszcz/IBM_Data_Science_Final_Project/blob/main/EDA_with_data_vis.ipynb

EDA with SQL

- Distinct of LAUNCH_SITE to display the names of the unique launch sites in the space mission.
- Display of 5 records where launch sites begin with the string ,KSC
- 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 where the succesful landing outcome in drone ship was acheived.
- List the names of the boosters which have success in ground pad 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.
- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- 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,
- https://github.com/andszcz/IBM_Data_Science_Final_Project/blob/main/EDA_with_SQL.ipynb

Build an Interactive Map with Folium

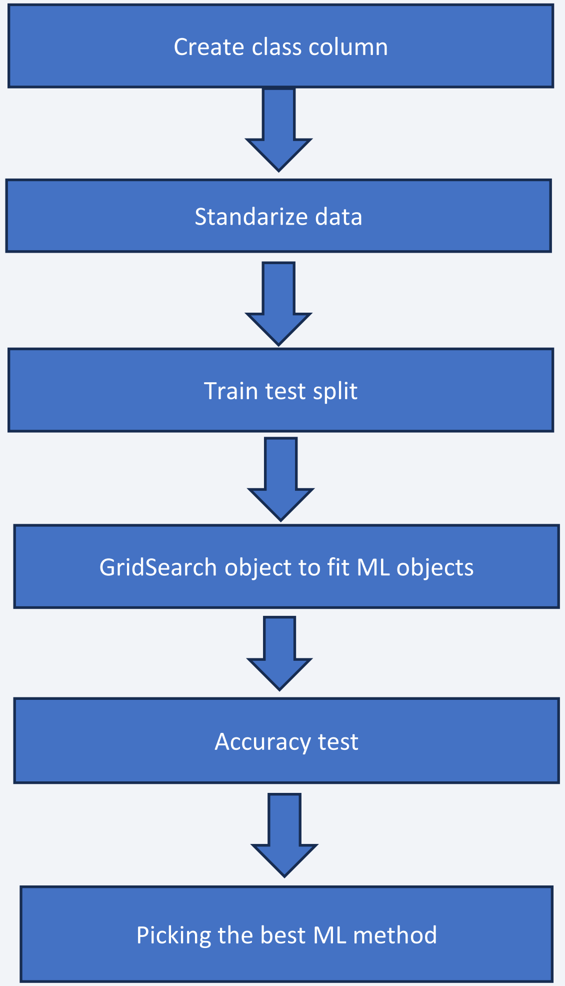
- We added following objects to a folium map:
  - Polylines – to connect SpaceX launch sites with nearest landmarks like railways, cities and coastlines.
  - Circles – to highlight area of launch sites.
  - Marker cluster – red represents failure and greene represents succes
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- Pie chart to show:
  - succes rate of all launch sites,
  - Proportion of fails and successes of given launch site
- Scatter Plot to:
  - show correlation between Misssion Ooutcome and Payload Mass (Kg) for different Booster Versions
  - to filter Payload Mass by a weight range using the slider

Predictive Analysis (Classification)

- Model development:
    - Preprocessing and data standarization
    - Train test split
    - Optimizing parameters with Grid Search
    - Model training
- Model evaluation:
    - Assessment of accuracy
    - Getting best hyperparameters
    - Plotting confusion matrix
- Finding the best model
    - Picking the model with the best accuracy score



- https://github.com/andszcz/IBM_Data_Science_Final_Project/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

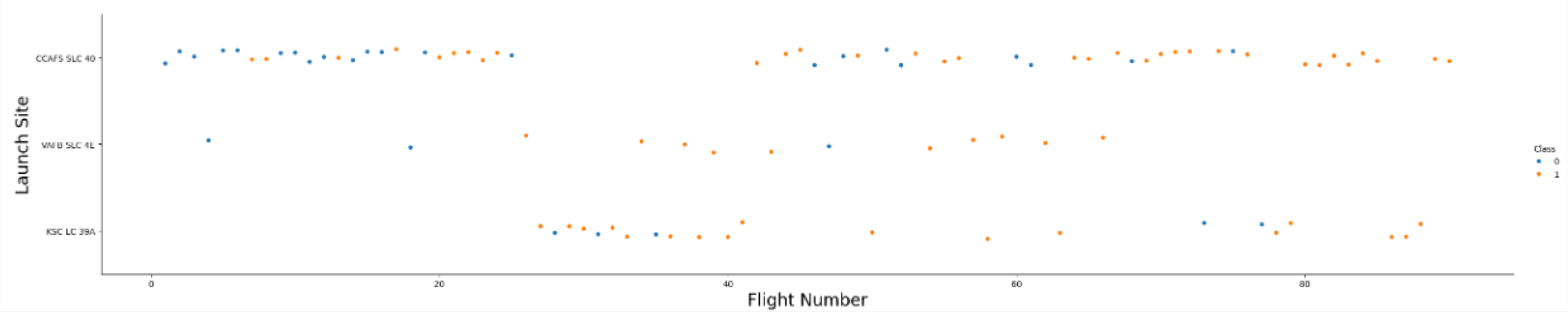
Results

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

Section 2

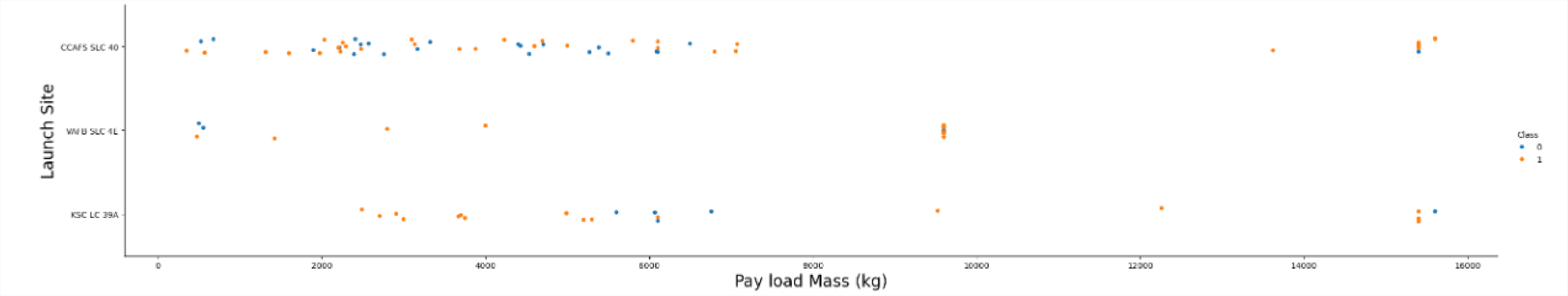
Insights drawn from EDA

Flight Number vs. Launch Site



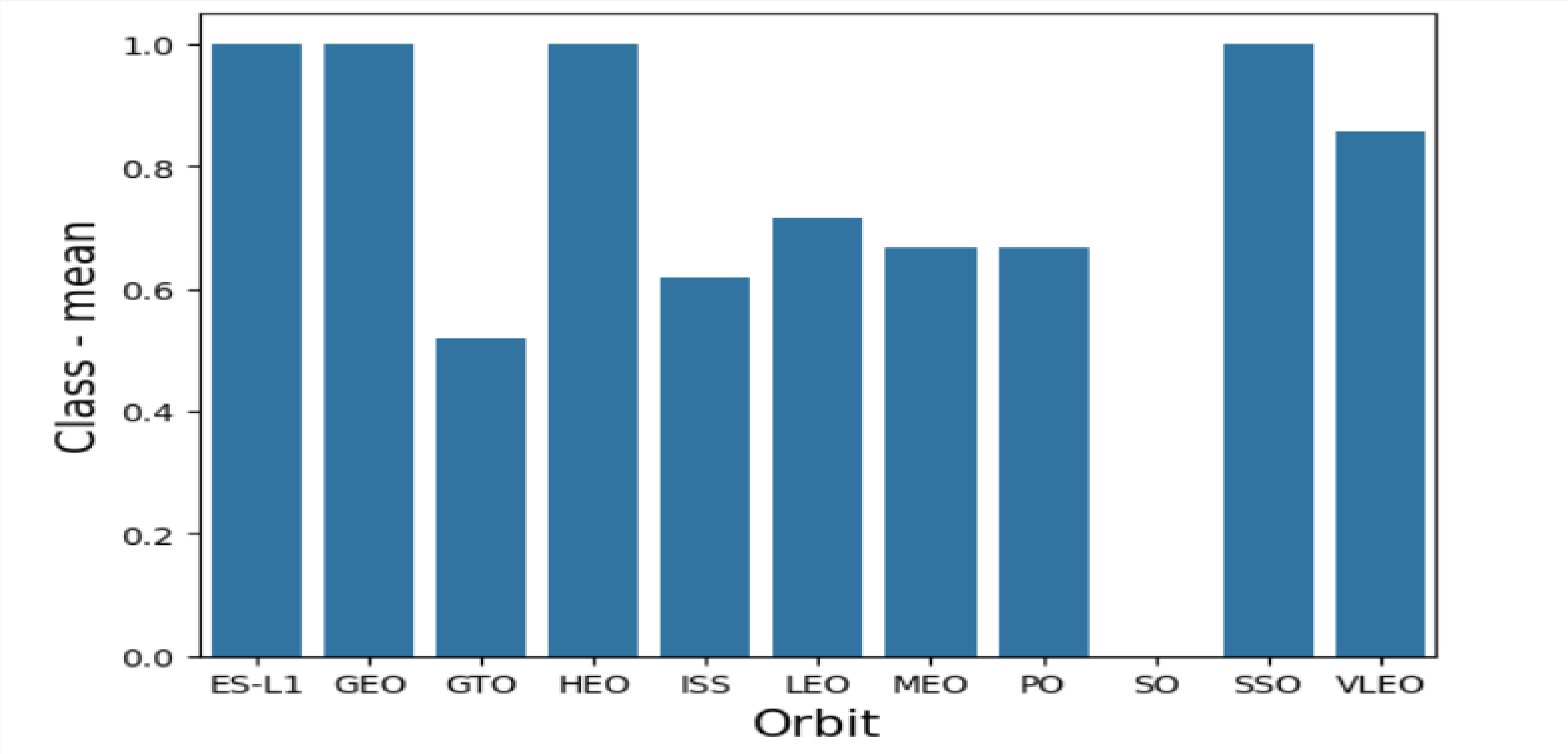
- Launch sites with higher number of flights are more succesfull.
- The most successful launch site is CCAFS SLC 40

Payload vs. Launch Site



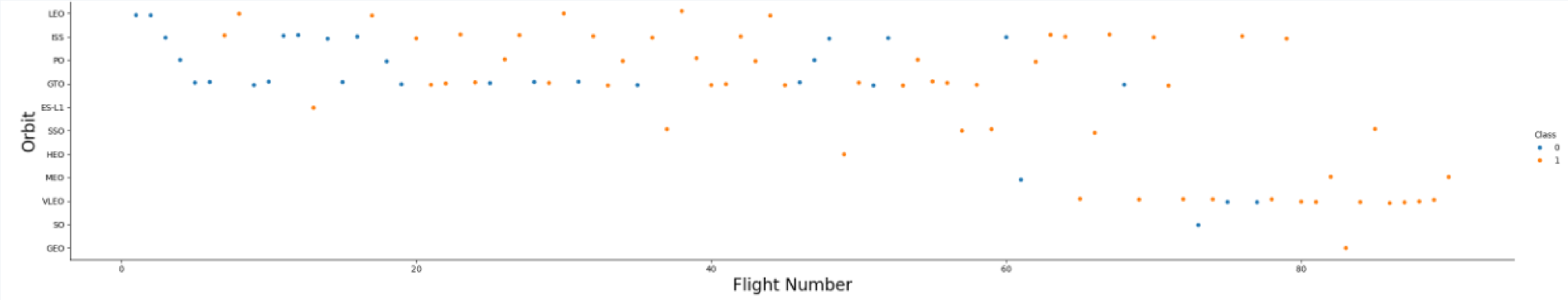
- It is not clear if successful launches depends on Pay load Mass

Success Rate vs. Orbit Type



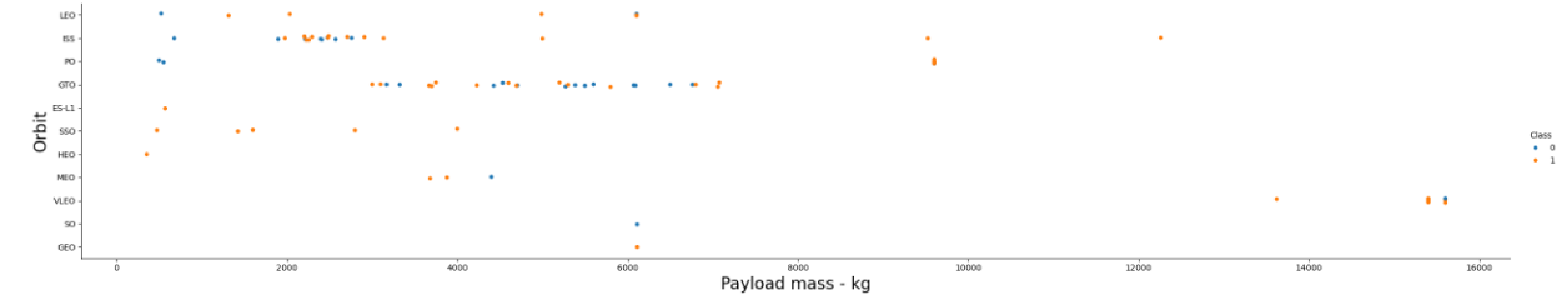
- Orbits ES-L1, GEO, HEO, SSO have 100% succes rate.
- There is no successful landing for orbit SO.
- All other orbits have success rate between 50-90%

Flight Number vs. Orbit Type



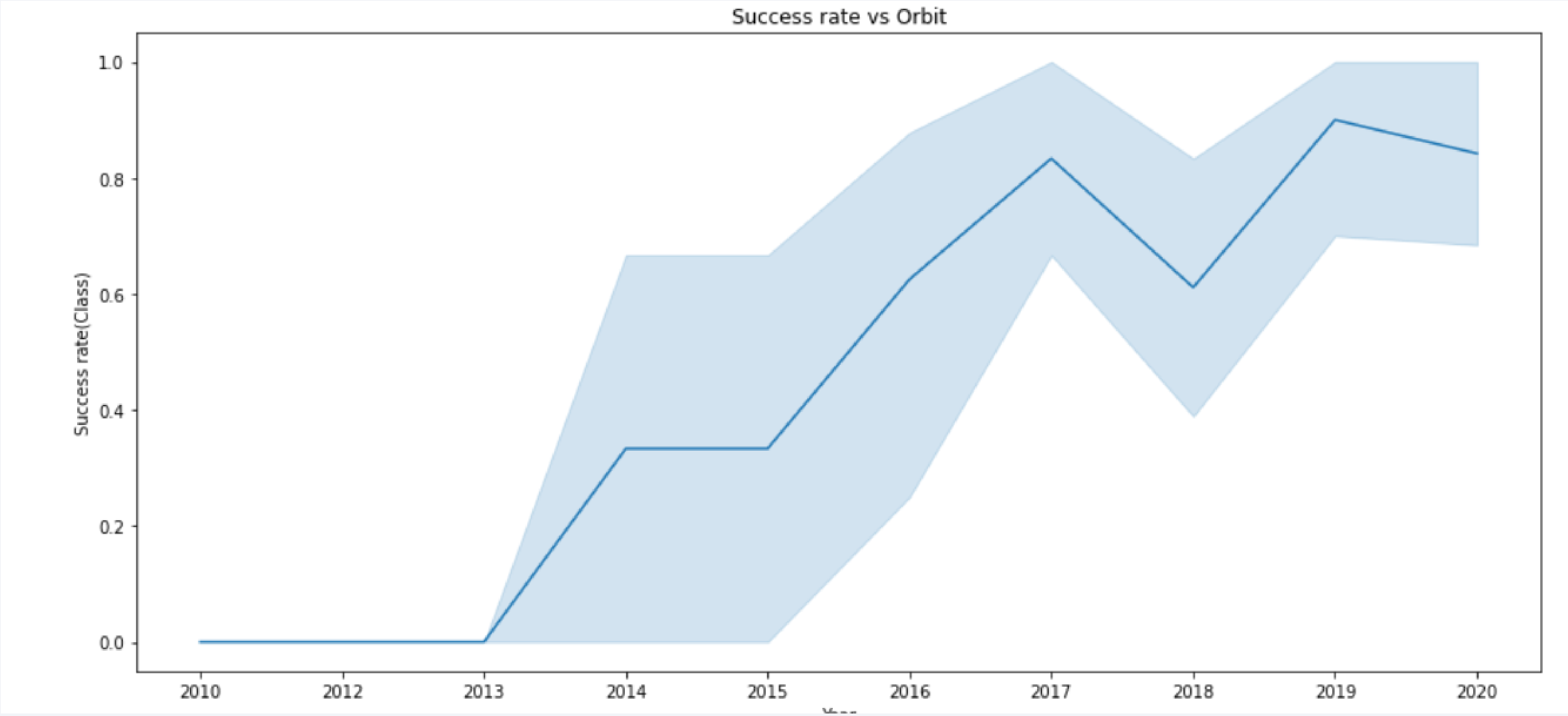
- For Orbits: LEO, VLEO, ISS, it seems that success rate increases with flight nuber.
- It is not so obvious in case of GTO orbit.

Payload vs. Orbit Type



- In most cases payload mass is up to 8000 kg
- Success rate seems to increase with payload mass for ISS and Leo.
- It is not so clear in case of GTO (both, positive and negative landings are present).

Launch Success Yearly Trend



- Succes rate significantly increased from 2013 to 2020.

All Launch Site Names

```
In [10]: %sql select distinct LAUNCH_SITE from SPACEXTBL

 * sqlite:///my_data1.db
Done.

Out[10]:
```

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

- We retrieved unique names of launch site with sql distinct statement.

Launch Site Names Begin with 'KSC'

Task 2

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

In [16]:
```
%sql select * from SPACEXTBL WHERE LAUNCH_SITE LIKE 'KSC%' LIMIT 5
```

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

Out[16]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- To find records where launch sites’ names start with \`KSC\` we use filter: WHERE LAUNCH_SITE LIKE 'KSC%’
- To show exactly 5 records we use: LIMIT 5

Total Payload Mass

In [19]:
```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

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

Out[19]:

SUM(PAYLOAD_MASS__KG_)
45596

- To calculate total payload we use sql sum function.
- To limit outcomes to NASA's boosters we use filter: WHERE CUSTOMER = 'NASA (CRS)'.

Average Payload Mass by F9 v1.1

```
In [21]: %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Booster_Version = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.

Out[21]:
```

AVG(PAYLOAD_MASS__KG_)
2928.4

- To calculate total payload we use sql AVG function.
- To limit outcomes to NASA’s boosters we use filter: WHERE Booster_Version = 'F9 v1.1'

First Successful Ground Landing Date

```
In [22]: %sql select MIN(Date) from SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)'

 * sqlite:///my_data1.db
Done.

Out[22]:
```

MIN(Date)
2016-04-08

- To find the dates of the first successful landing we use MIN function.
- To limit outcome to drone ship we use filter: WHERE Landing_Outcome = 'Success (drone ship)'

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [26]: %sql select distinct Booster_Version from SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
         and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

          * sqlite:///my_data1.db
         Done.

Out[26]:
```

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1

- To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 we use distinct Booster_Version and filter results with:

  WHERE Landing_Outcome = 'Success (ground pad)’

  and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

Total Number of Successful and Failure Mission Outcomes

In [27]:
```
%sql select Mission_Outcome, COUNT(*) from SPACEXTBL GROUP BY Mission_Outcome
```

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

Out[27]:

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

- To calculate the total number of successful and failure mission outcomes, we need to count Mission_Outcome with GROUP BY Mission_Outcome

Boosters Carried Maximum Payload

```
In [29]: %sql select distinct Booster_Version from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

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

Out[29]:

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

- We query distinct BOOSTER_VERSION from SPACEXTBL
- We create subquery to find maximum PAYLOAD_MASS_KG_ and use it as a filter.

2015 Launch Records

```
In [33]: %sql select substr(Date,6,2) month, substr(Date,9,2) date, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL
WHERE substr(Date,0,5) ='2017' AND  Landing_Outcome= 'Success (ground pad)'

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

Out[33]:

month	date	Landing_Outcome	Booster_Version	Launch_Site
02	19	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	01	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	03	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	14	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	07	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	15	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- We use usbstr function do display the month names, dates and years of succesful landing_outcomes.

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

In [36]:
```
%sql select Landing_Outcome, COUNT(*) from SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC
```

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

Out[36]:

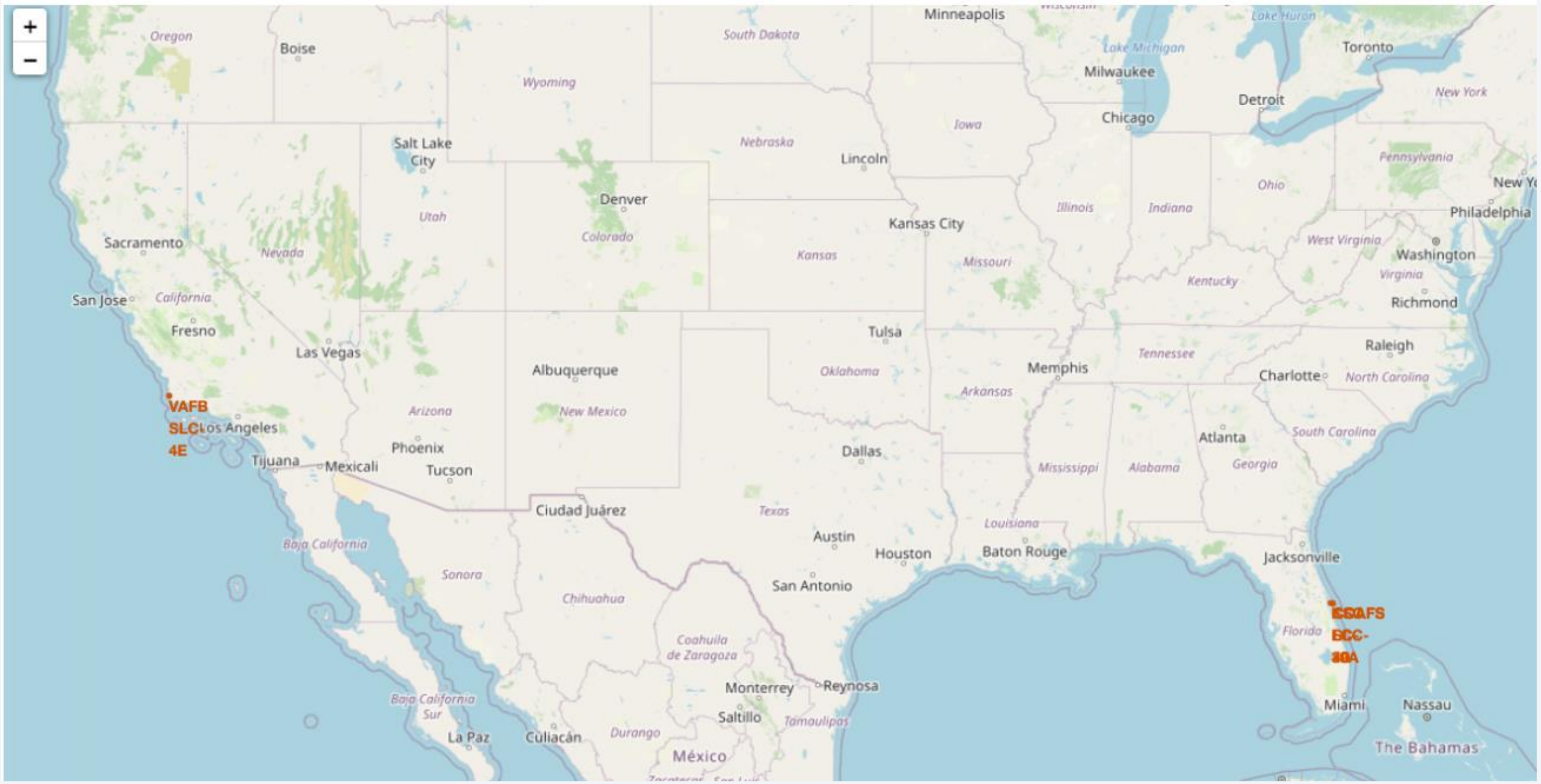
Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- We filter respective dates in where clause
- We use count and group by to find total number of landing outcomes
- We use ORDER DESC to rank values

Section 3

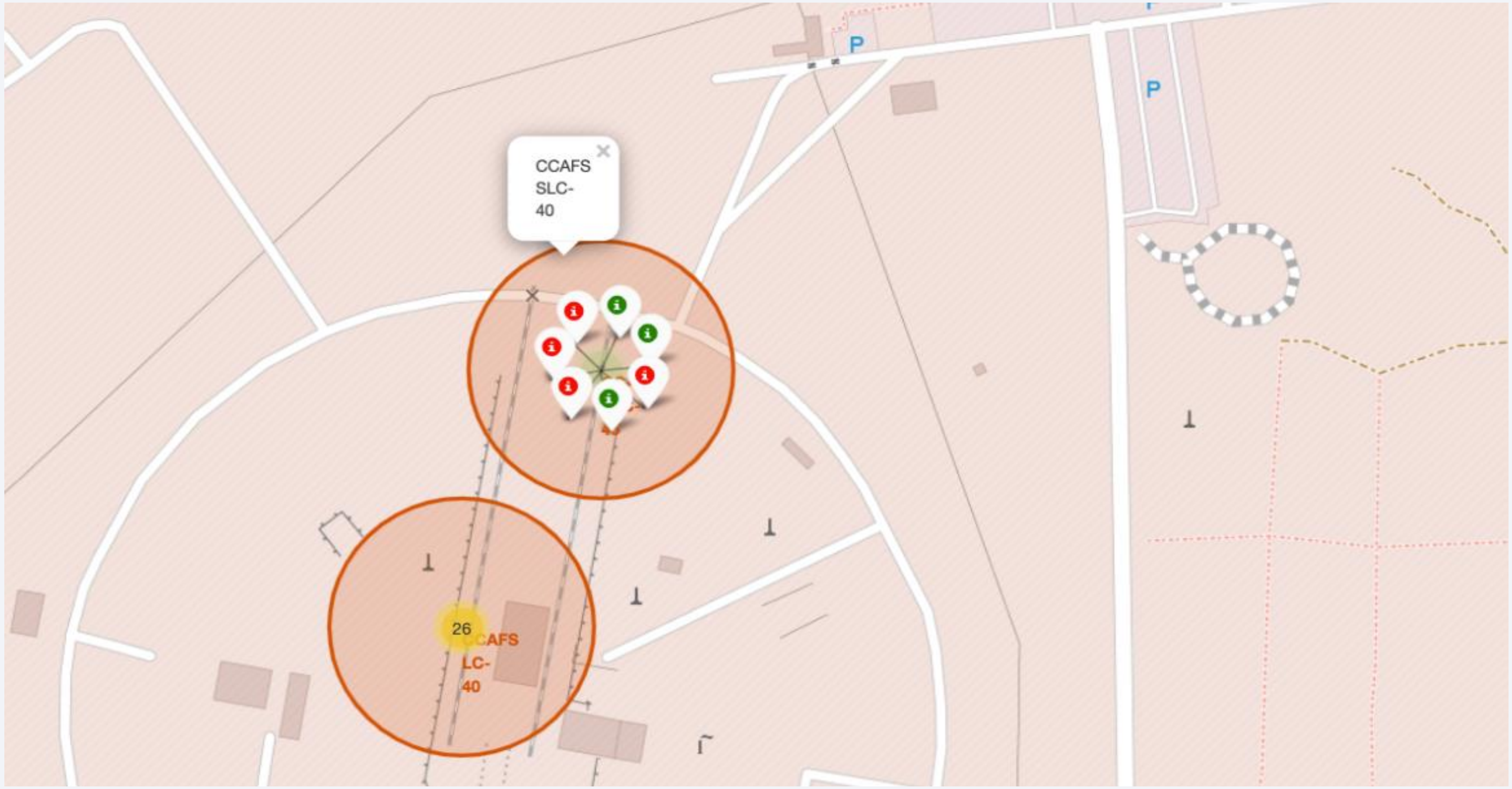
Launch Sites Proximities Analysis

Launch sites locations



- All launch sites are located on the USAs seecost, in Florida and California.

Markers of success/failed launches



- Green and red labels are signs of succes/failed landing.
- It helps us in assessment of success rate of each lunch site.

Distance to proximities

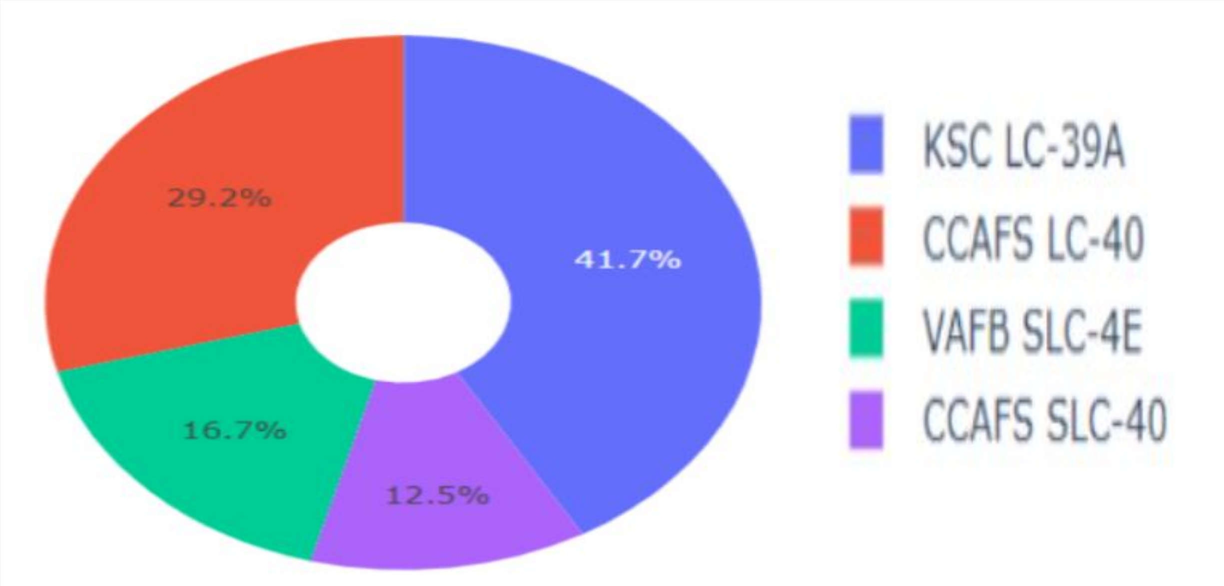
The launch site is in:

- close distance to the seecoast (0.9 km)
- 29 km away from the highway
- 78 km away from Orlando

Section 4

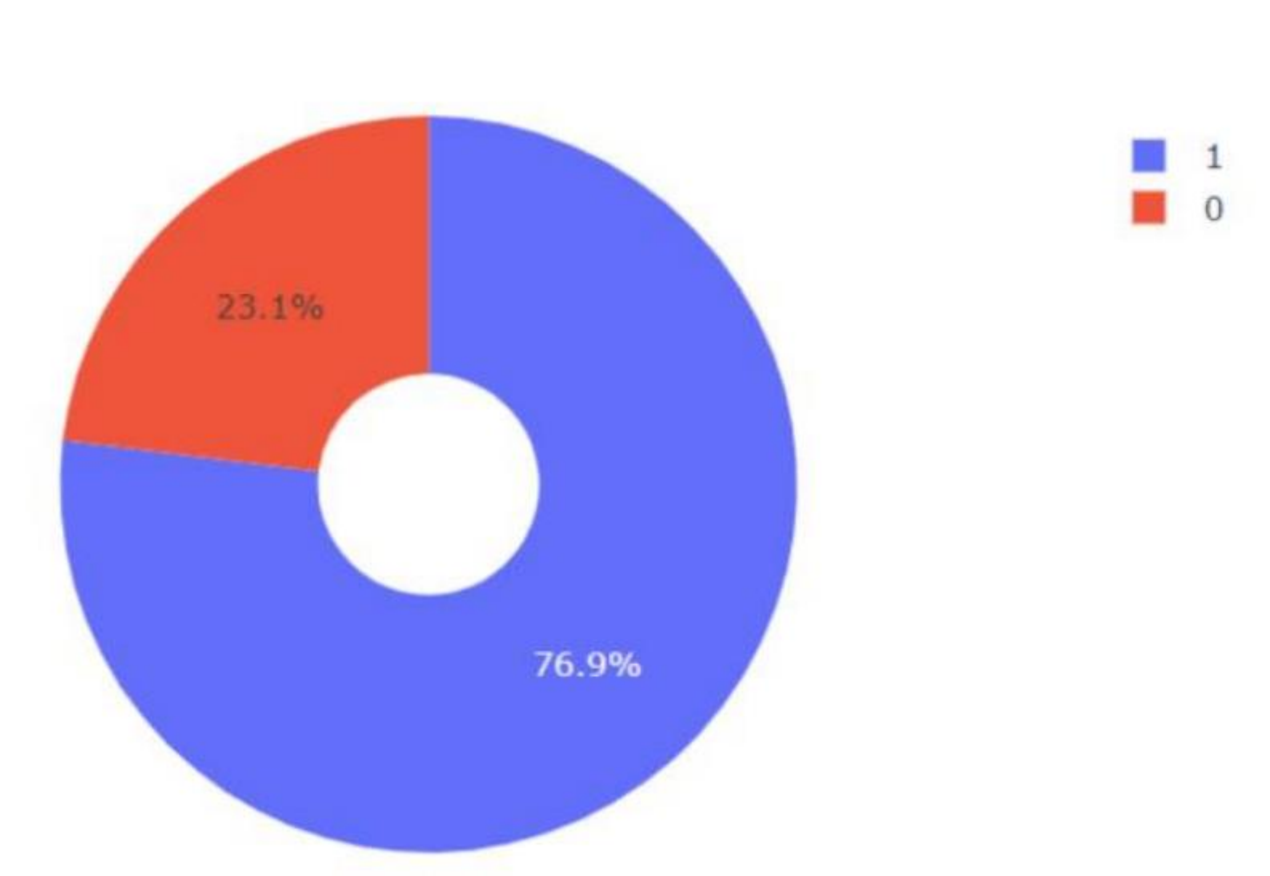
Build a Dashboard with Plotly Dash

Launch sites – success rate



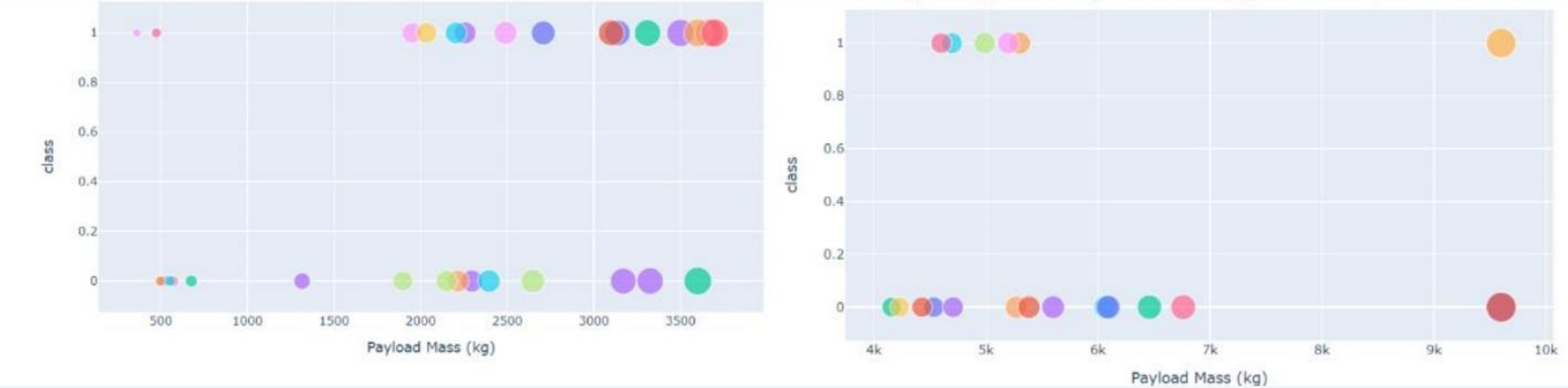
- KSC LC-39A has the highest success rate

Launch site with the highest success ratio



- 76.9 % of KSC LC-39A launches ended with success while 23,1% ended with failure

Success rate vs Payload Mass

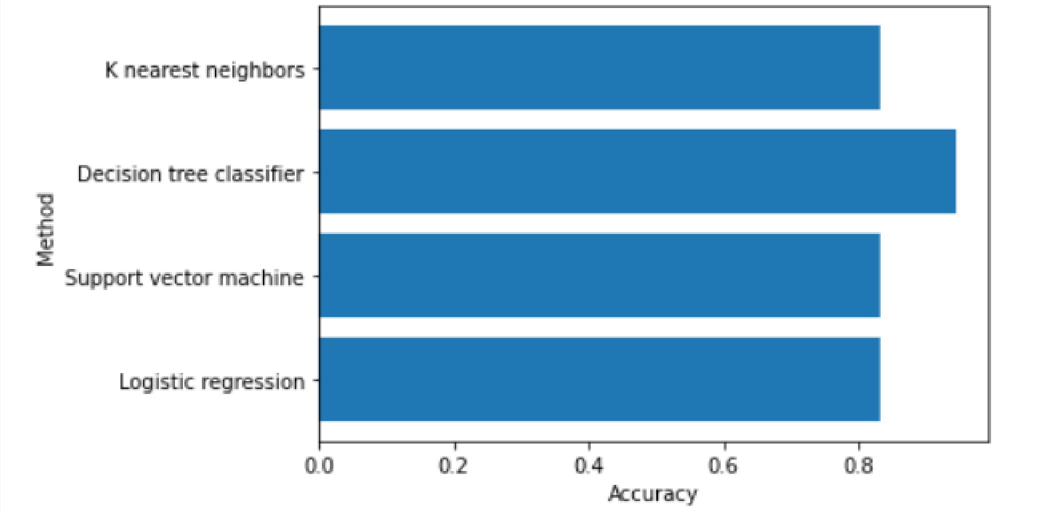


- Low weighted payloads have higher success rate

Section 5

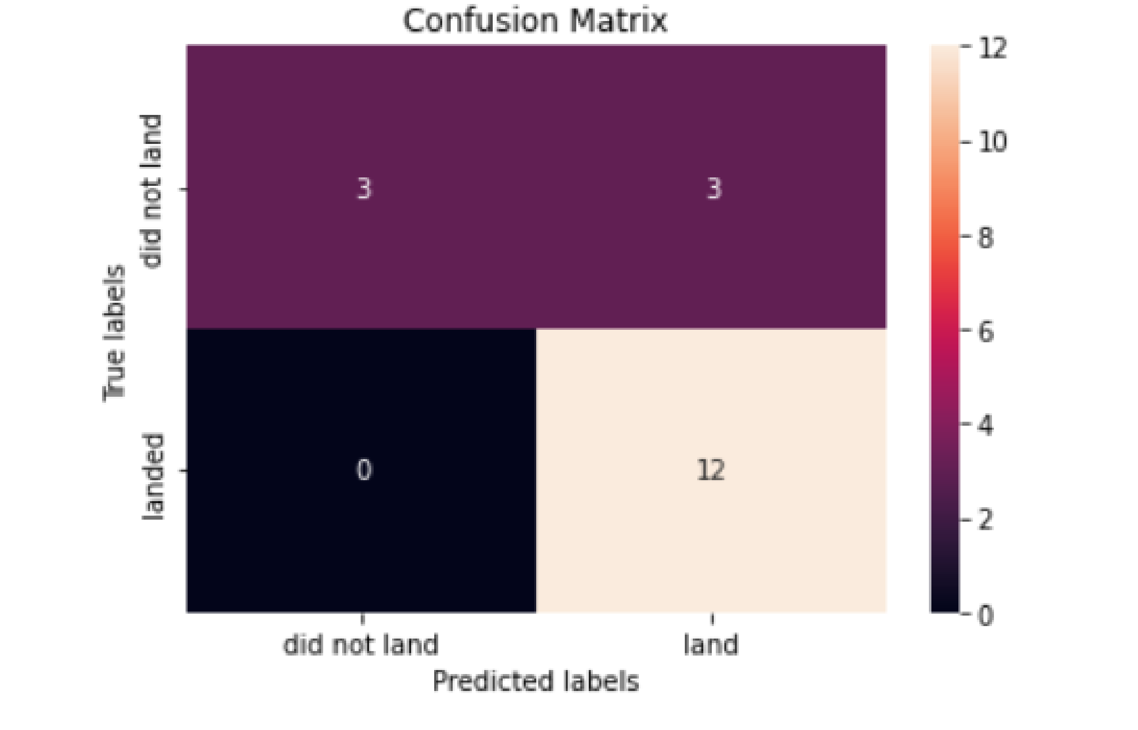
Predictive Analysis (Classification)

Classification Accuracy



- The highest classification accuracy has Decision tree classifier

Confusion Matrix



- The model:
    - Correctly predicted12 successful landings (true positive),
    - Wrongly predicted 3 successful landings (false positive),
    - Correctly predicted 3 unssuccesful landings (true negative),

Conclusions

- There was observed positive correlation between numer of flights and success rate (succes rate significantly increased from 2013 to 2020).
- The most successful launch site is CCAFS SLC 40.
- Orbits ES-L1, GEO, HEO, SSO have 100% succes rate
- The best predictive model is the Decision Tree Classifier.

Appendix

- GitHub Repositiory: https://github.com/andszcz/IBM_Data_Science_Final_Project

Thank you!