IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cléber Rodrigo de Souza
02/03/2022

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 Plotly Dash
    - 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 Allon Musk.
- Problems you want to find answers

  - Therefore, if we can determine if the first stage will land, 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

- 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 DataFrame using the json_normalize function
- Web Scraping
  - Performed an HTTP GET request to the Falcon9 Launch HTML Wiki page
  - Used Python BeautifulSoup package 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 REST API

https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/jupyter-labs-spacex-data-collection-api.ipynb

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

response = requests.get(spacex_url)

content = response.json()

data = pd.json_normalize(content)
```

Data Collection - Scraping

- Web Scraping Using Python BeautifulSoup package

https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/jupyter-labs-webscraping.ipynb

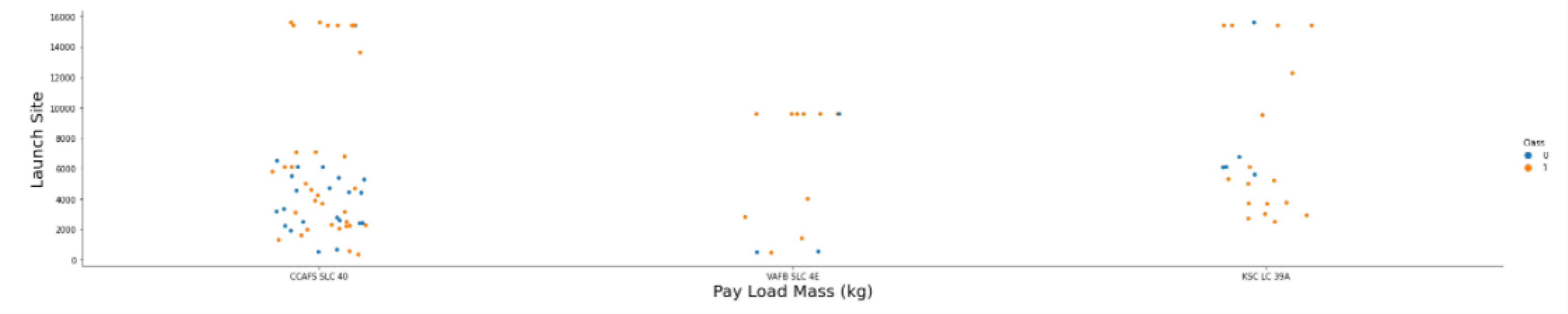
```
static_url = https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

response = requests.get(static_url).text

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

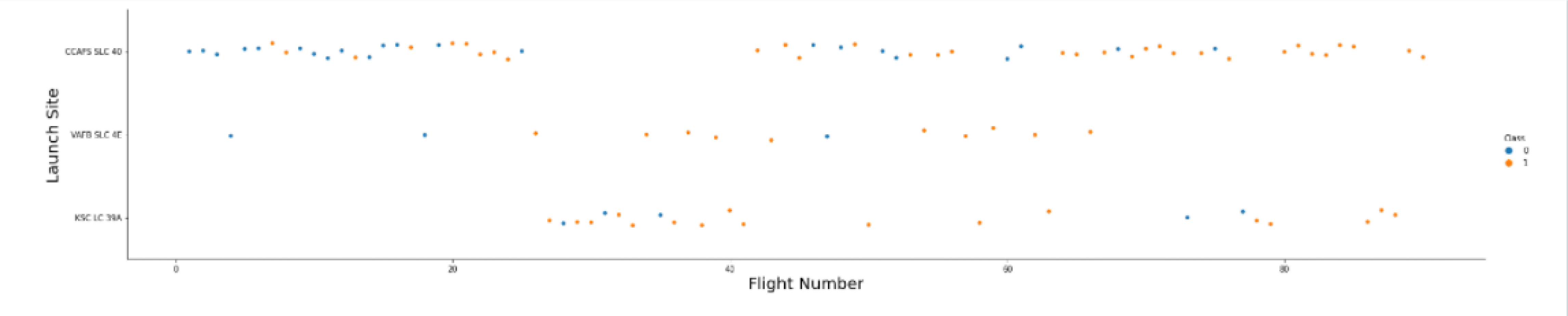
Data Wrangling

- 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/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e1318
0cfb54810a89ae6/labs-jupyter-spacex-Data_wrangling.ipynb

EDA with Data Visualization





https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Display names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- 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 when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 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

https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/jupyter-labs-eda-sql-coursera.ipynb

EDA with SQL (queries)

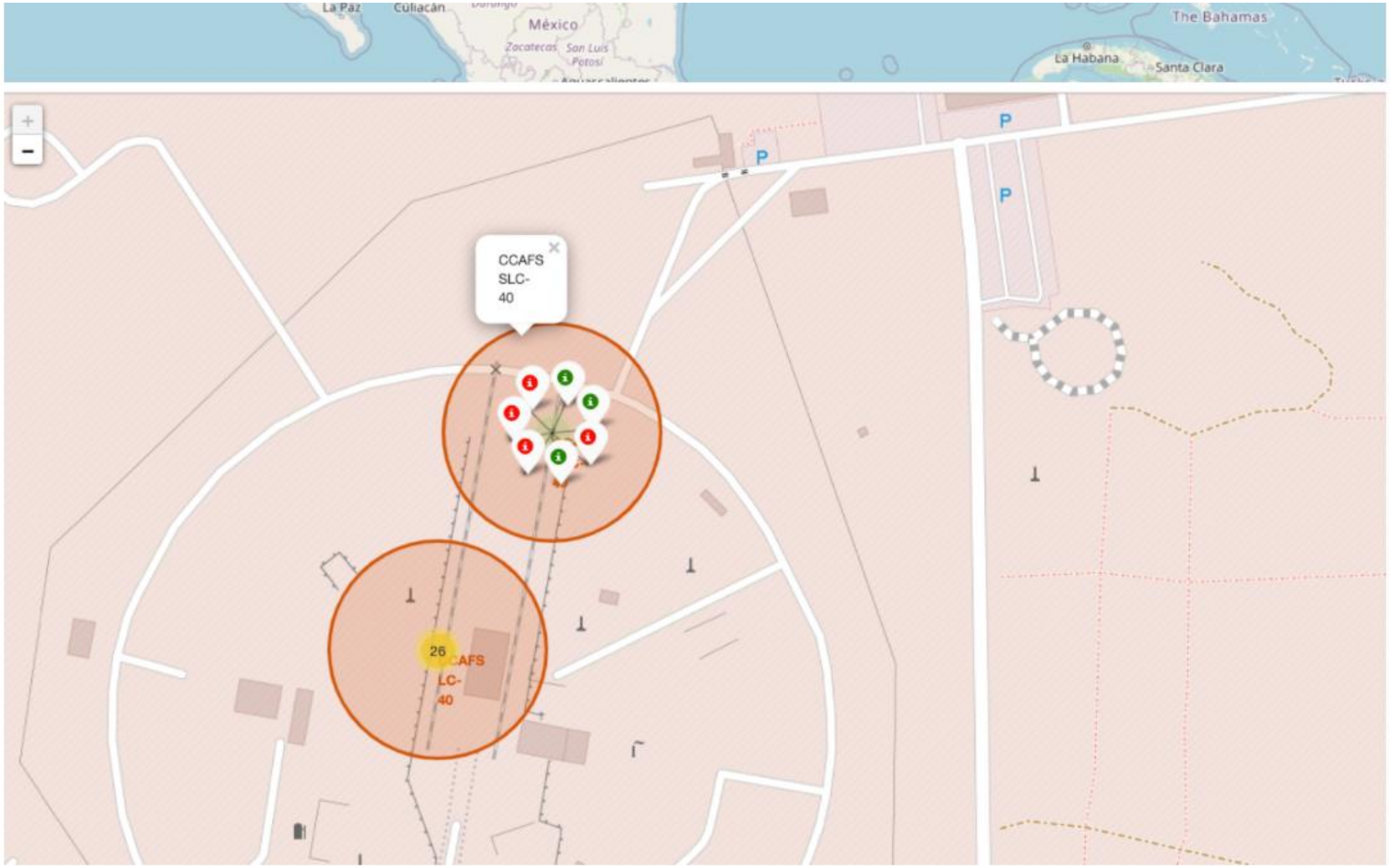
- select unique(LAUNCH_SITE) from SPACEXTBL
- select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit(5)
- select SUM(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)'
- select avg(payload_mass__kg_) from SPACEXTBL where booster_version like 'F9 v1.1'
- select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
- select booster_version from SPACEXTBL where Landing_Outcome= 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
- select mission_outcome, count(mission_outcome) from SPACEXTBL group by mission_outcome
- select booster_version from SPACEXTBL where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXTBL)
- select Landing_Outcome, booster_version, launch_site from SPACEXTBL where Landing_Outcome = 'Failure (drone ship)' and EXTRACT(YEAR FROM DATE) = 2015
- select Landing_Outcome, count(Landing_Outcome) as total from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by total 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.

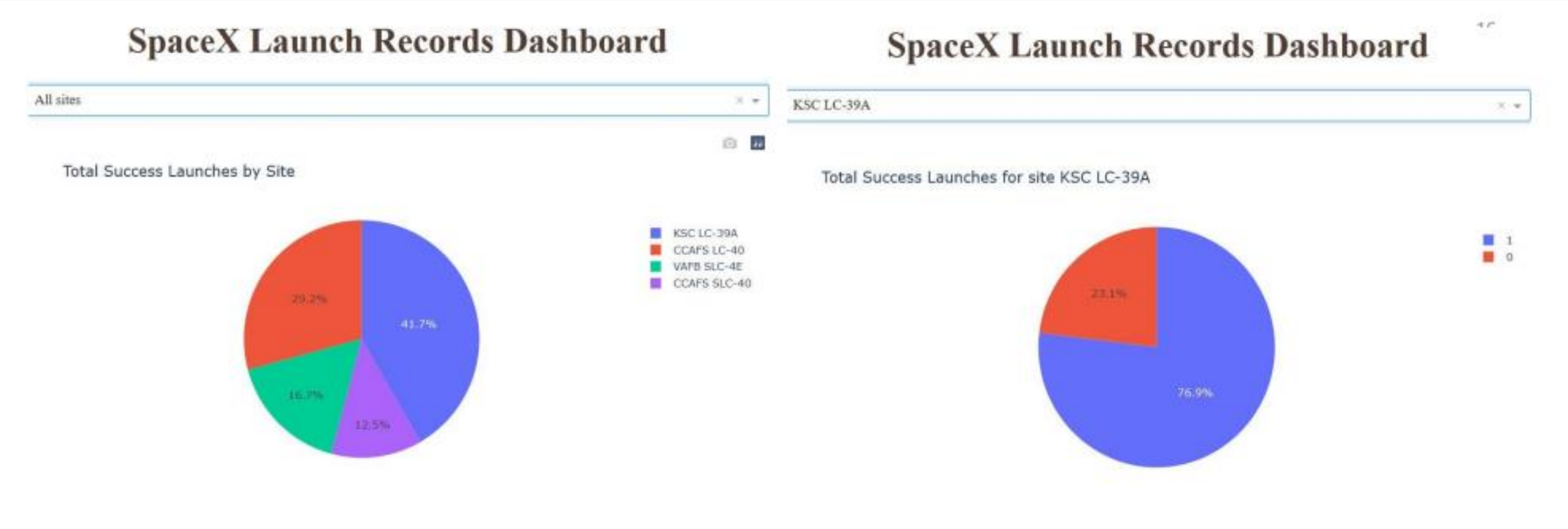
https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e1318
0cfb54810a89ae6/lab_jupyter_launch_site_location.ipynb

Build an Interactive Map with Folium



Build a Dashboard with Plotly Dash

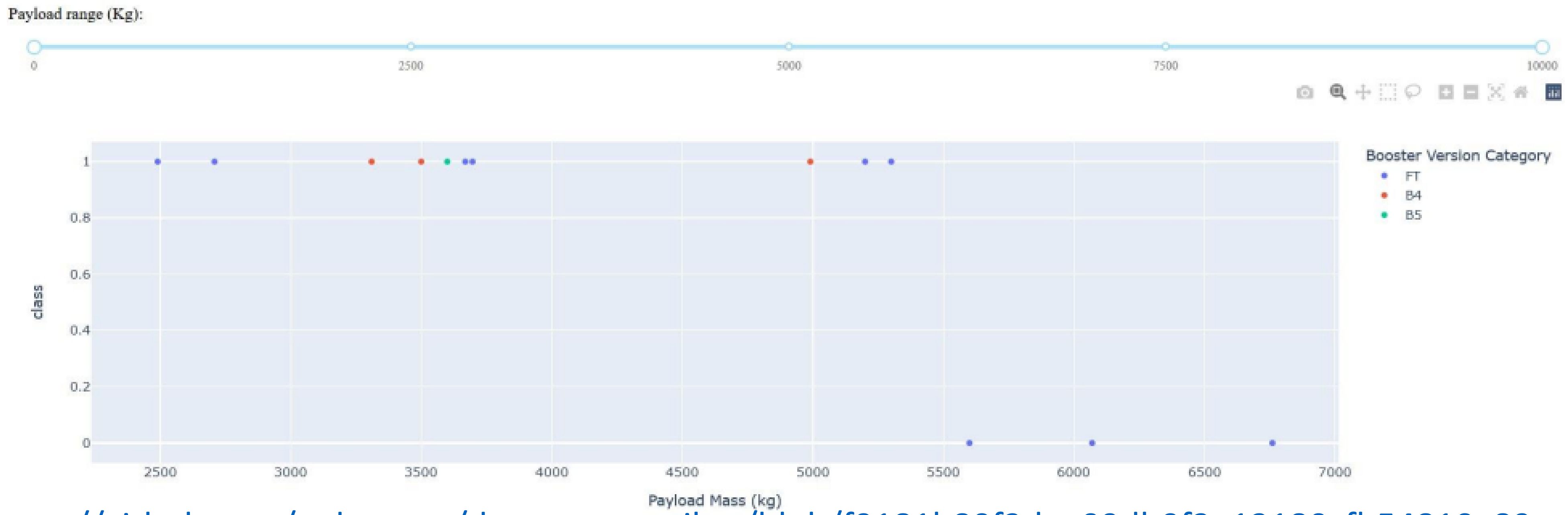
- Interactive visualization of successful launches per site/ all sites



https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/dash_spacex.py

Build a Dashboard with Plotly Dash

- Correlation between payload mass for different Booster Version and successful launch outcome

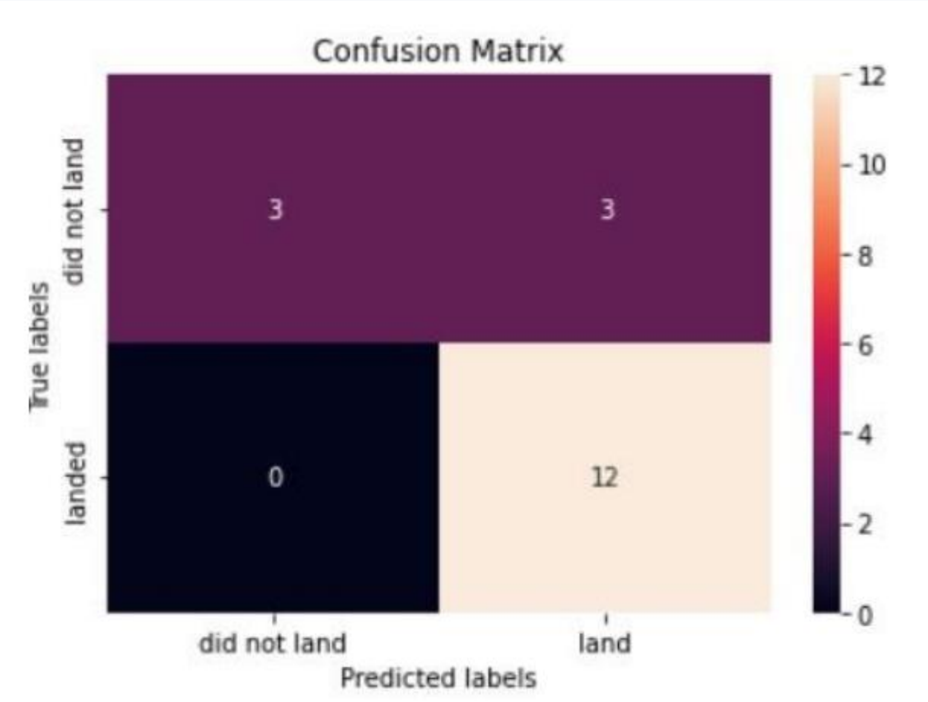


https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/dash_spacex.py

Predictive Analysis (Classification)

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

https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/SpaceX_Machine_Learning_Prediction.ipynb



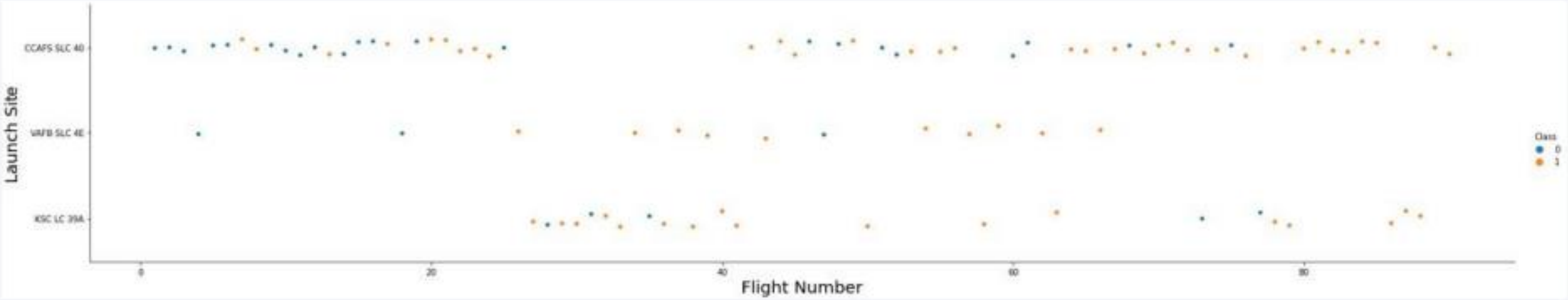
Results

- The most successful rate have ES-L1, GEO, HEO, SSO orbits
- Since 2013 successful launches rate increased from 0 to almost 80-90%
- 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

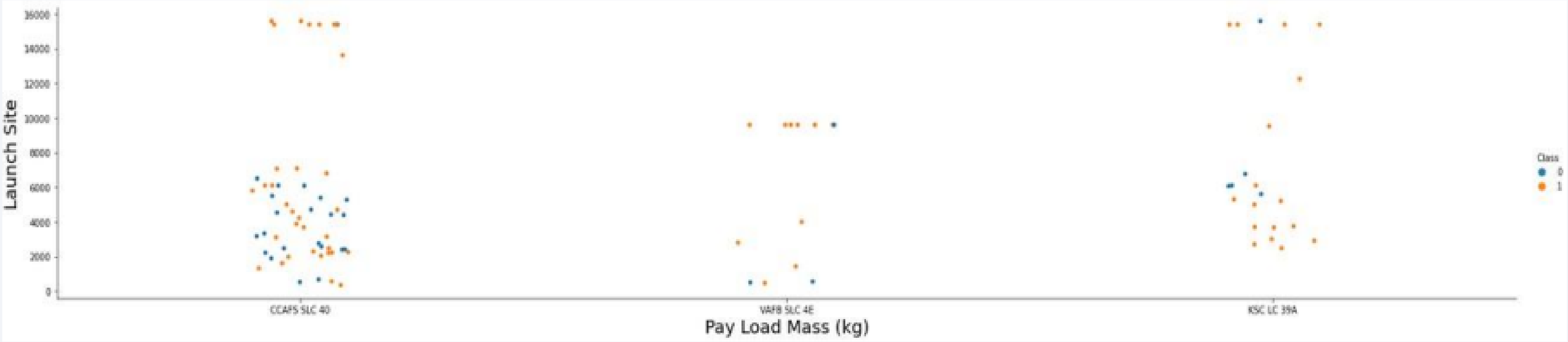
Insights drawn from EDA

Flight Number vs. Launch Site



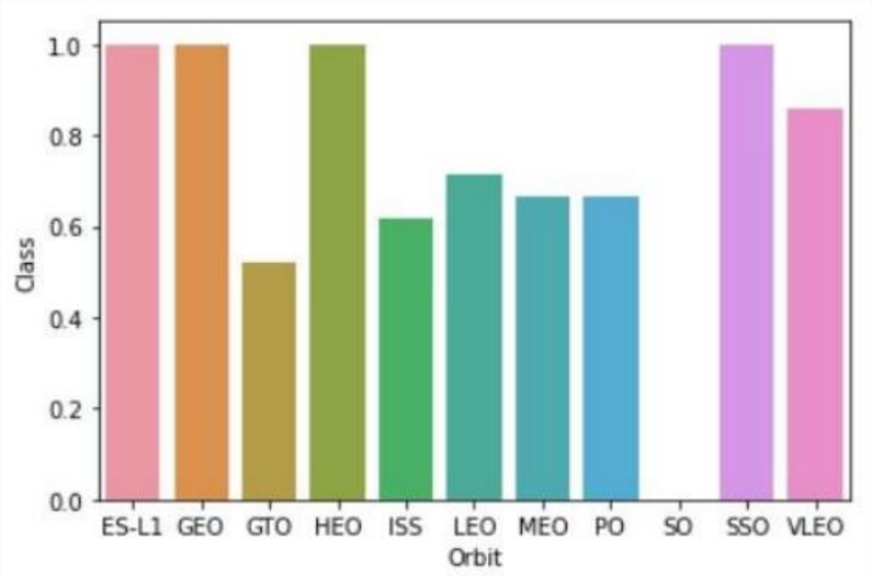
- The majority of launches are made from CCAFS SLC 40 site

Payload vs. Launch Site



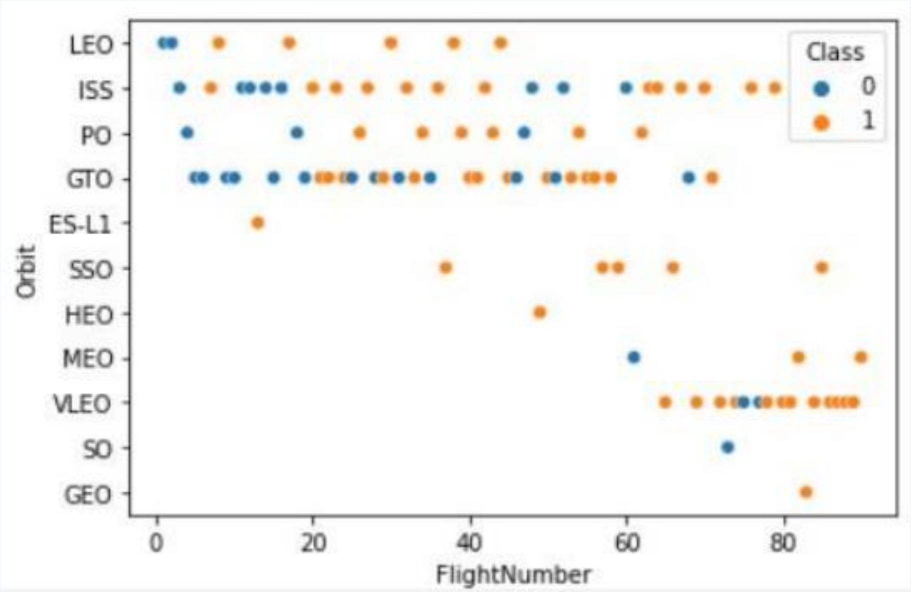
- Almost all launces from CCAFS SLC 40 with high payload were successful.
- The most successful rate of launches seem to be from KSC LC 39A regardless payload.

Success Rate vs. Orbit Type



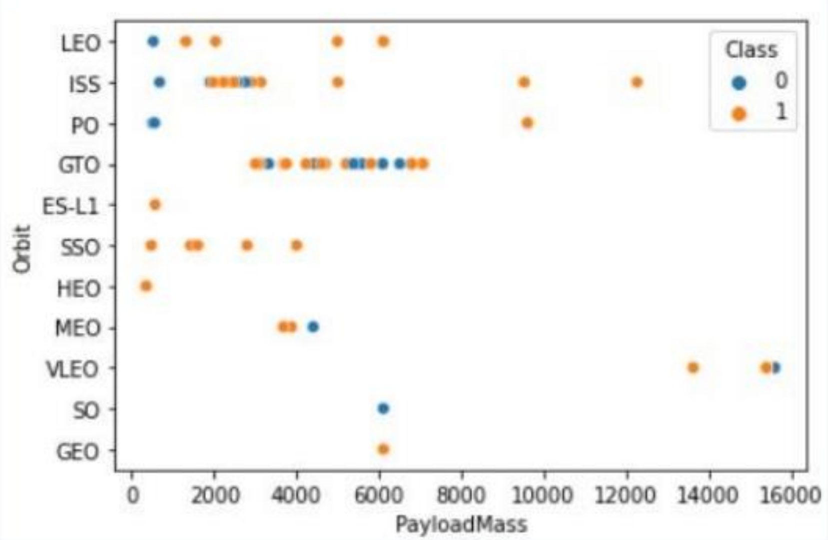
- The highest rate of success have ES=L1, GEO, HEO, SSO orbits launches

Flight Number vs. Orbit Type



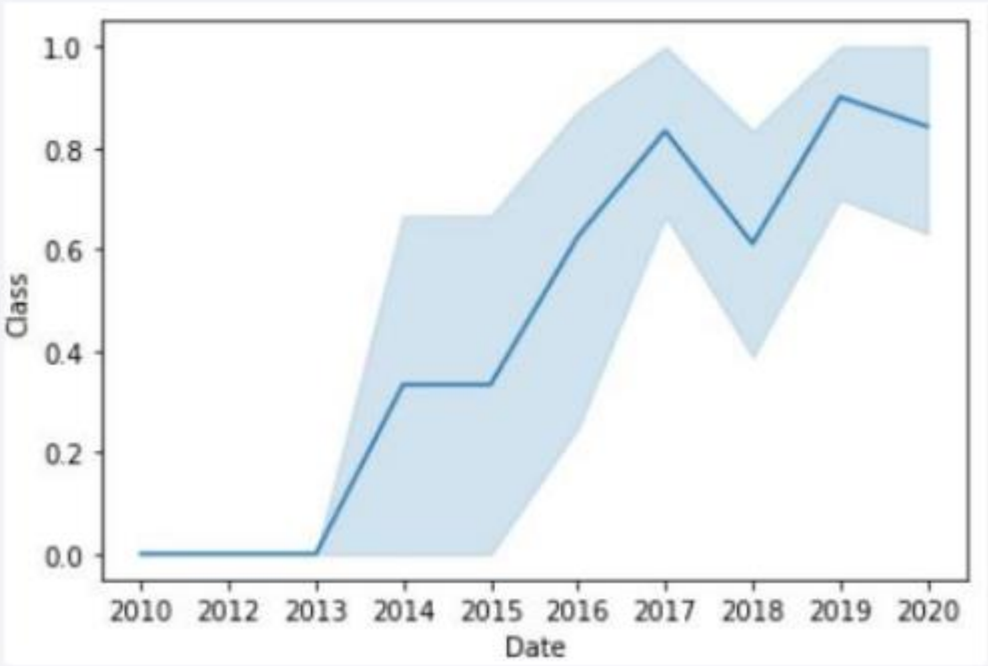
- VLEO orbit gain the highest popularity among all types
- ISS have pretty the regular amount of launches during the whole period

Payload vs. Orbit Type



- There is mostly two clusters of payload mass
- ~1500 – 3200 (ISS orbit)
- ~2200 – 7200 (GTO orbit)

Launch Success Yearly Trend



- Launch success rate was constantly improving since 2013 with an exception during the year 2017 and reach almost 85% to the end of 2019

All Launch Site Names

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

- The Falcon 9 rockets have been launched only from 4 different sites

Launch Site Names Begin with 'CCA'

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

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

Total Payload Mass

```
%%sql
select SUM(payload_mass__kg_)
from SPACEXTBL
where customer='NASA (CRS)'
```

1
45596

Average Payload Mass by F9 v1.1

```
%%sql
select avg(payload_mass__kg_)
from SPACEXTBL
where booster_version like 'F9 v1.1'
```

1
2928

First Successful Ground Landing Date

```
%%sql
select min(DATE)
from SPACEXTBL
where Landing_Outcome = 'Success (ground pad)'
```

2015-12-22

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_ between 4000 and 6000
```

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 mission_outcome, count(mission_outcome)
from SPACEXTBL
group by mission_outcome
```

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

Boosters Carried Maximum Payload

```
%%sql
select booster_version
from SPACEXTBL
where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXTBL)
```

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 Landing_Outcome, booster_version, launch_site
from SPACEXTBL
where Landing_Outcome= 'Failure (drone ship)'
    and EXTRACT(YEAR FROM DATE)=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

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

```
%%sql
select Landing_Outcome, count(Landing_Outcome) as total
from SPACEXTBL
where DATE between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by total DESC
```

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

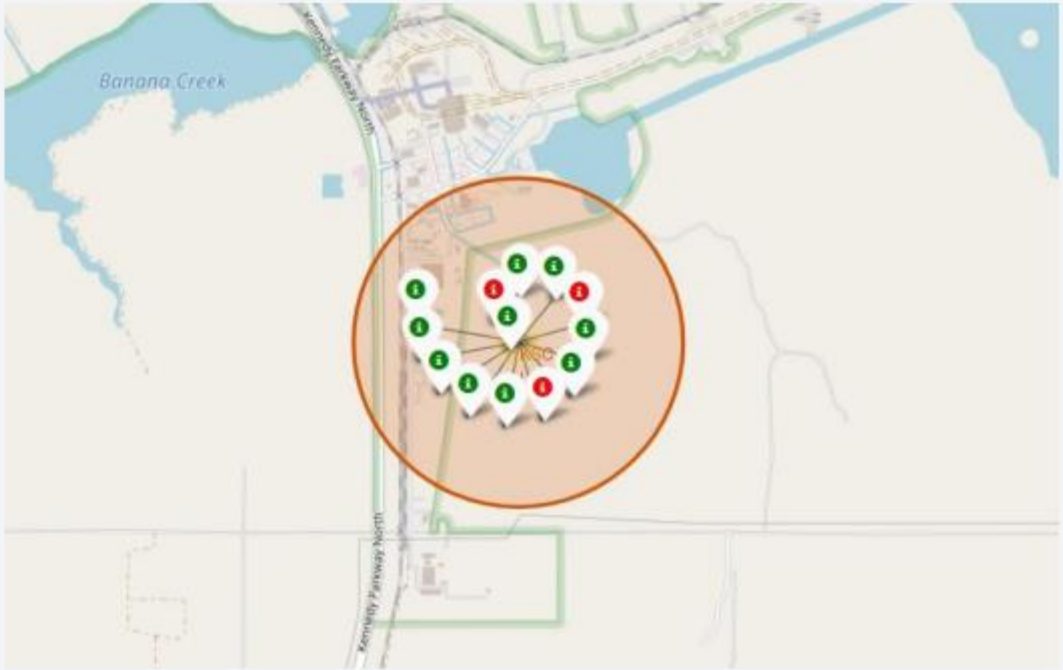
Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

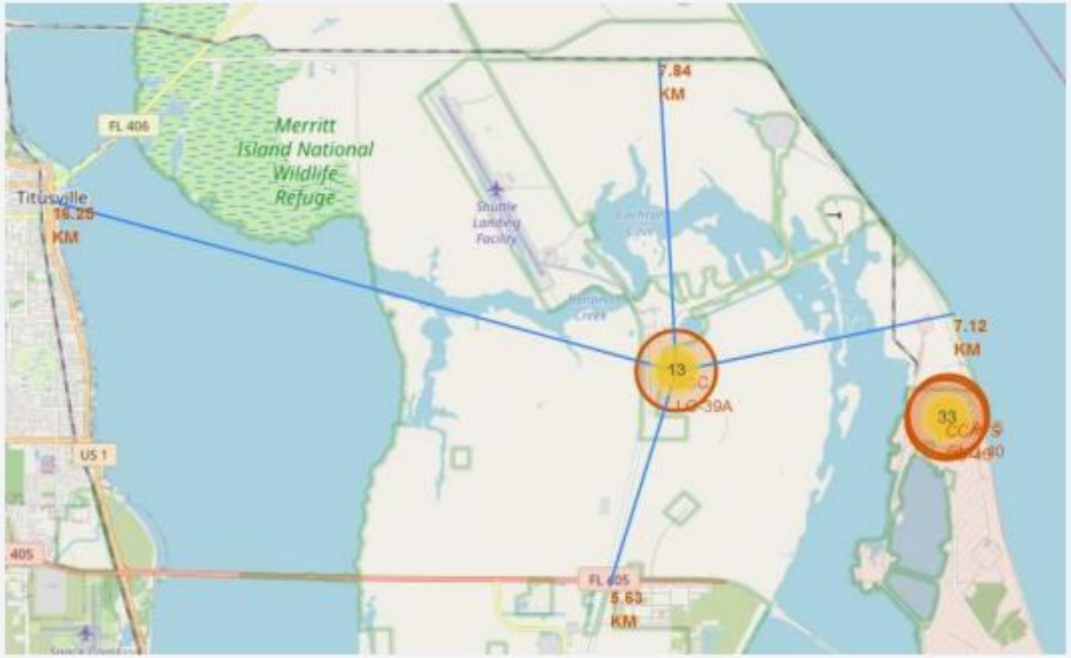
There is 4 launch sites but 3 of them are clustered on East Coast of Florida

<Folium Map Screenshot 2>



KSC LC-39A is the most successful site with 10 of 13 (77%) successful launches outcomes

<Folium Map Screenshot 3>

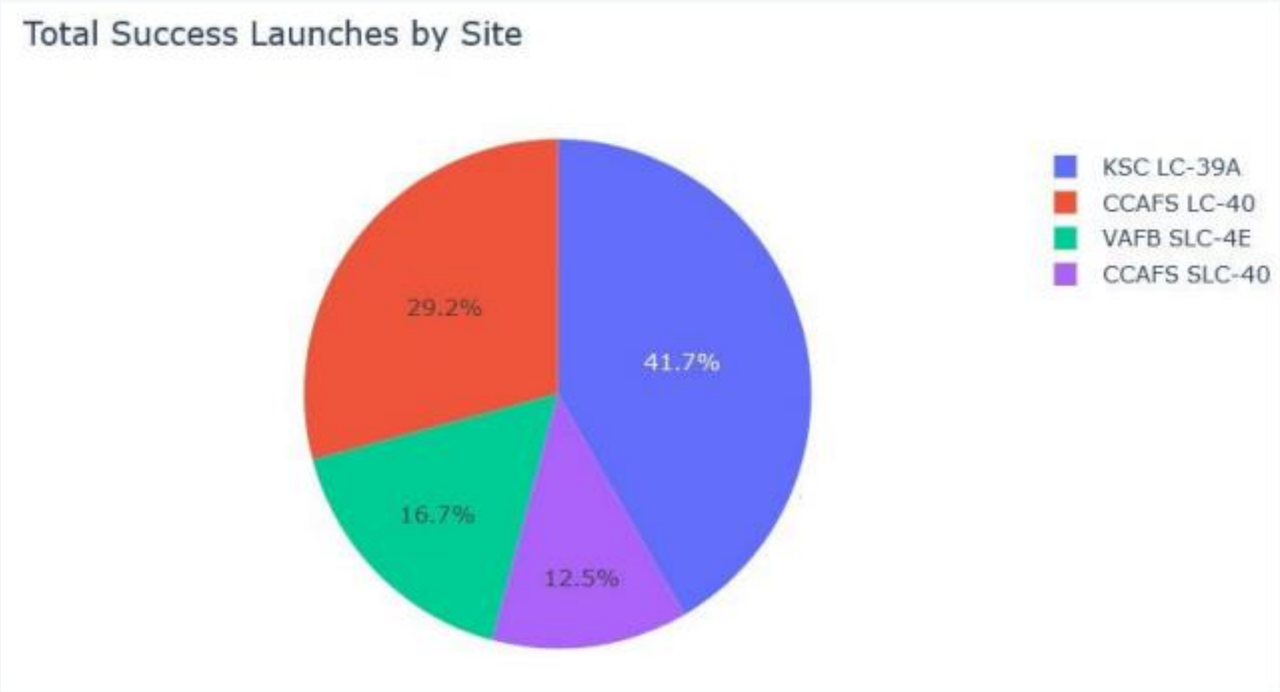


All sites are in a close proximity to coast line and railway (max ~7km)

Section 4

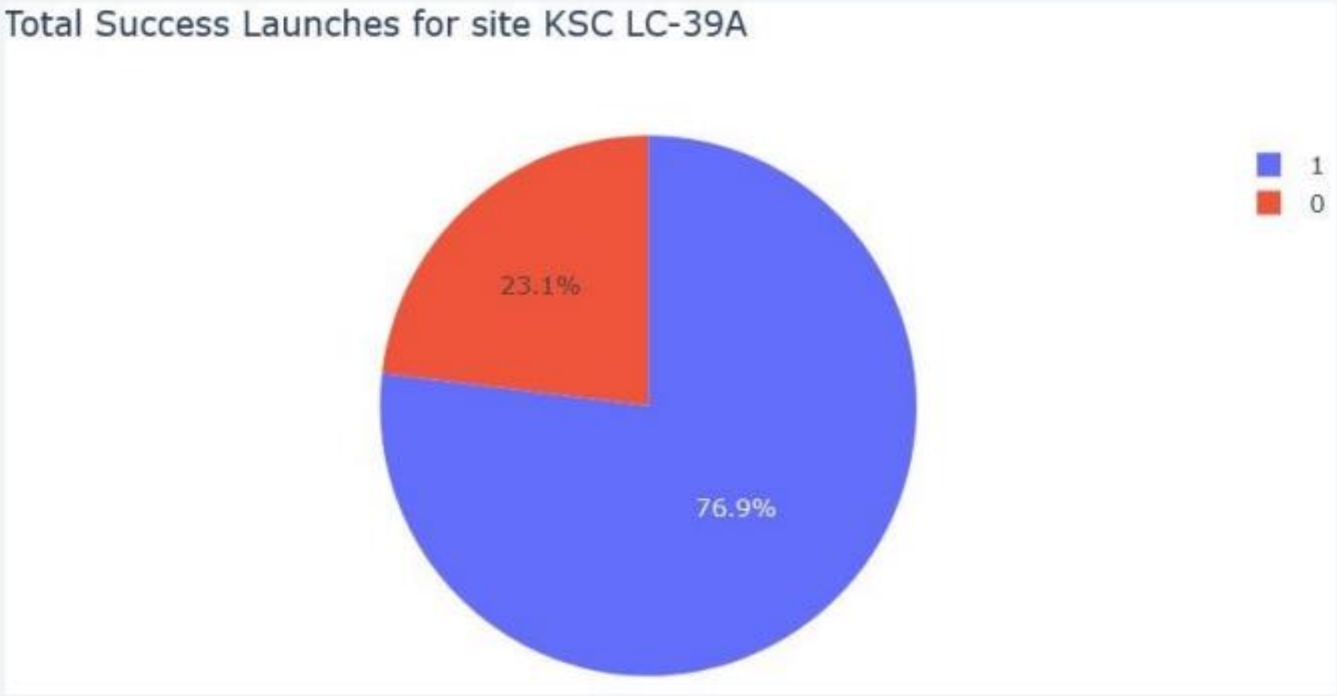
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

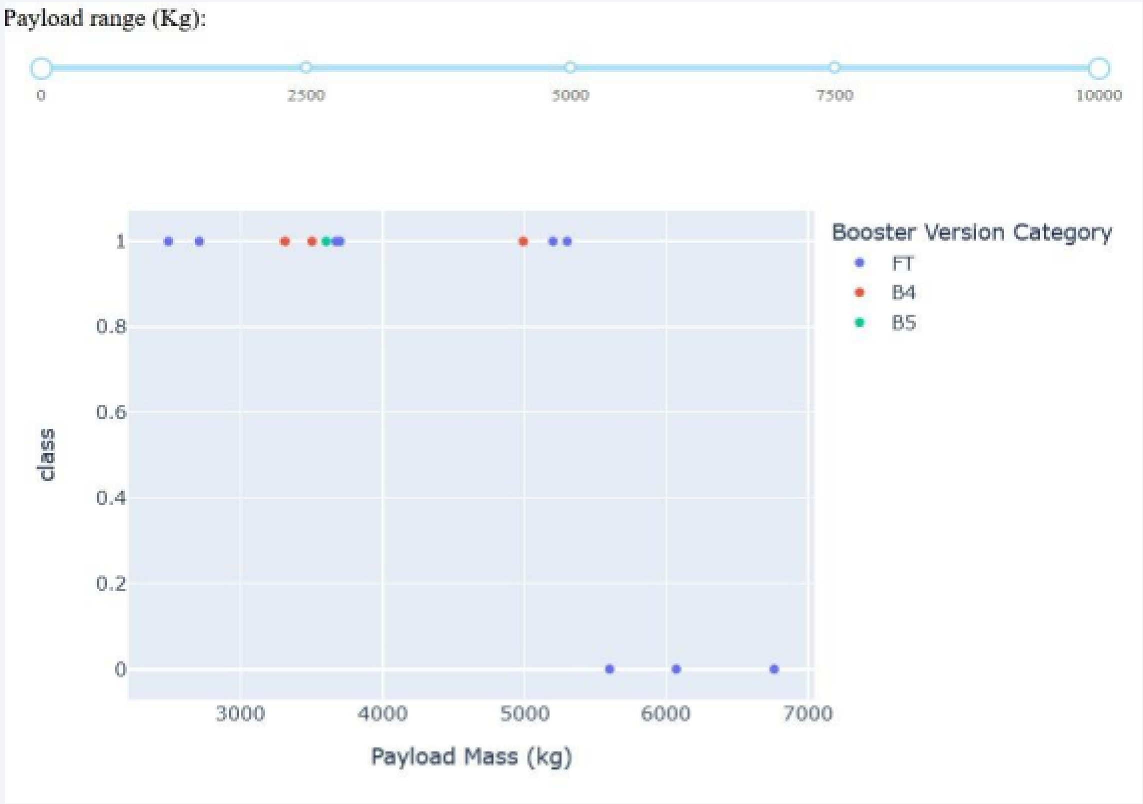


KSC LC-39A has significantly higher success rate

<Dashboard Screenshot 2>



<Dashboard Screenshot 3>



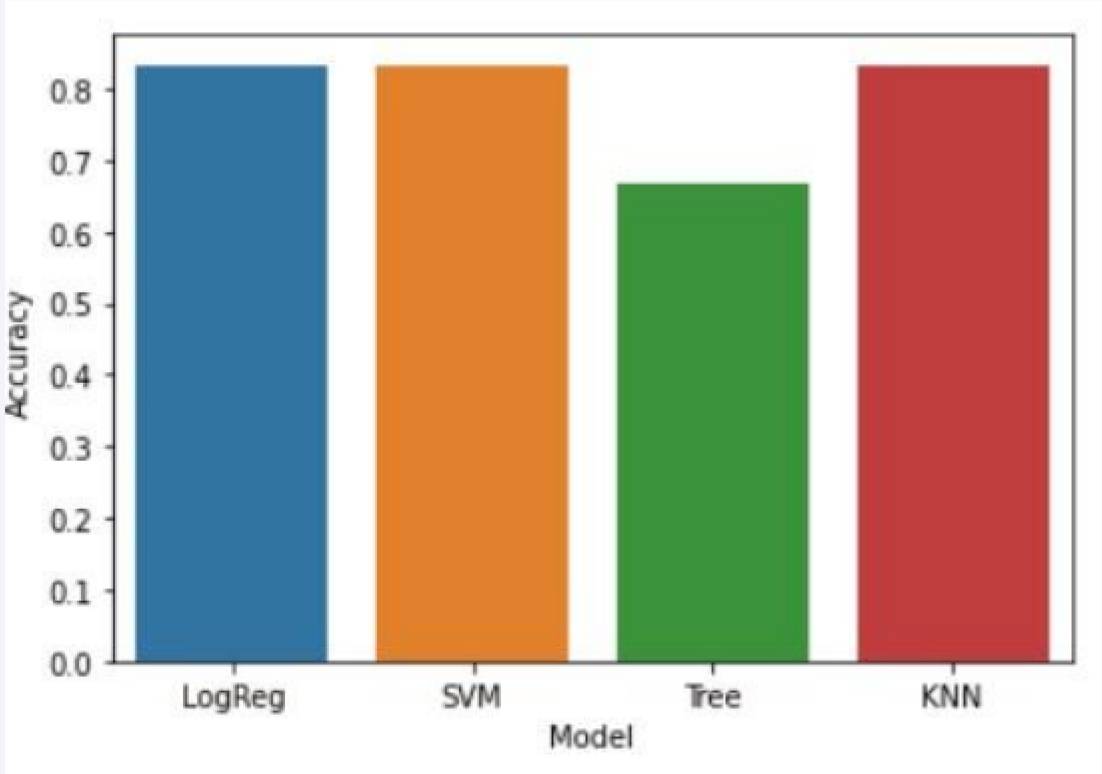
Lower payload mass leads to higher chances for successful launch

Section 5

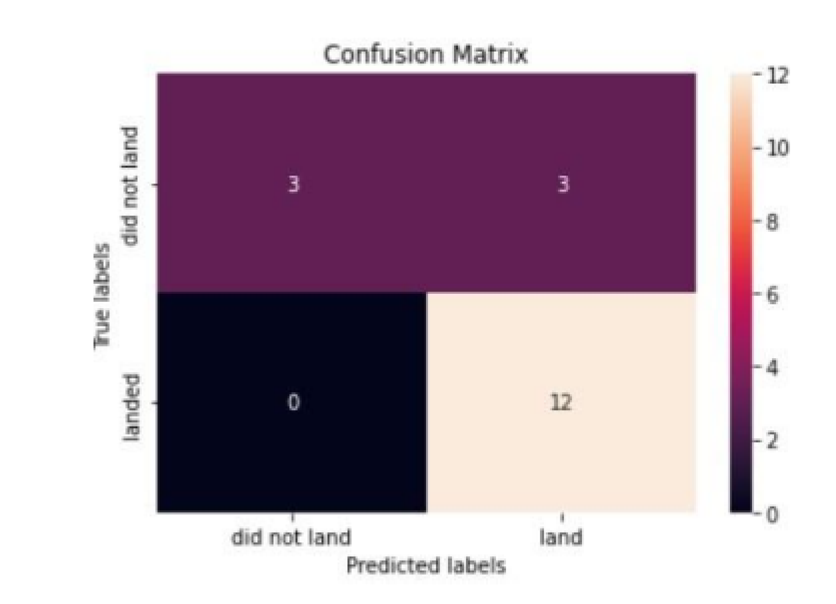
Predictive Analysis (Classification)

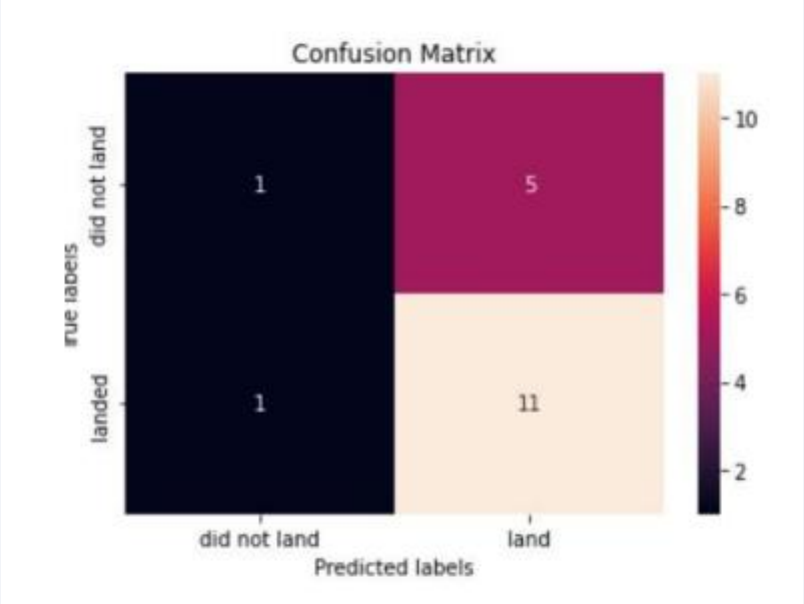
Classification Accuracy

- KNN, SVM, LogisticRegression models have same accuracy on test dataset



Confusion Matrix





- KNN, SVM and Logistic vs. DecissionTree
- KNN, SVM, LogisticRegression algorithms have same predictive accuracy on test dataset
- All this 3 models have Type I Error with 3 False positive outcomes

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 have chances 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/crdesouza/ds_capstone_ibm

- SQL queries

https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/jupyter-labs-eda-sql-coursera.ipynb

- Python Dash app

https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/dash_spacex.py

- Machine Learning Prediction

https://github.com/crdesouza/ds_capstone_ibm/blob/f9181b80f2dea09db0f3e13180cfb54810a89ae6/SpaceX_Machine_Learning_Prediction.ipynb

Thank you!