IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ELBECHNA Zakaria
16/09/2022
https://github.com/ZakariaElbe1

Outline

• Executive Summary	Slide : 2
• Introduction	Slide : 2
• Methodology	Slide : 5
• Results	Slide : 17
• Conclusion	Slide : 45
• Appendix	Slide : 46

Executive Summary

Summary of methodologies

- Collecting data from public SpaceX API and SpaceX Wikipedia page by web scrapping at first stage;
- Transforming a Categorical variable to be a dependent variable to classify the success and failure landing;
- Exploratory data analysis using SQL, pandas, visualization, folium maps, and dashboard.
- Gathering relevant columns to be used as features. Changed all categorical variables to binary using one hot encoding.
- Standardizing data and using GridSearchCV to find best parameters for machine learning models. Visualize accuracy score of all models.

Summary of all results

- Using four machine learning algorithms for classification: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors;
- Using the best parameters founded by GridSearshCV ,All models produced similar results with accuracy rate equal to 83.33% with similar precision, recall and F1_score.
- All models over predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

Background

- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Space X vs Space Y

Problem:

Space Y want to predict successful Stage 1 recovery;

=>For that , the company ask us to build a machine Learning classification to predict successful stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combining data from SpaceX public API and SpaceX Wikipedia page web scrapping
- Perform data wrangling
  - Classifying landings as success and failure landing
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuning models using GridSearchCV to find the best parameters for every model chosen

Data Collection

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

The features extracted using the methods below are:

- ***Data Columns founded by Space X API :***

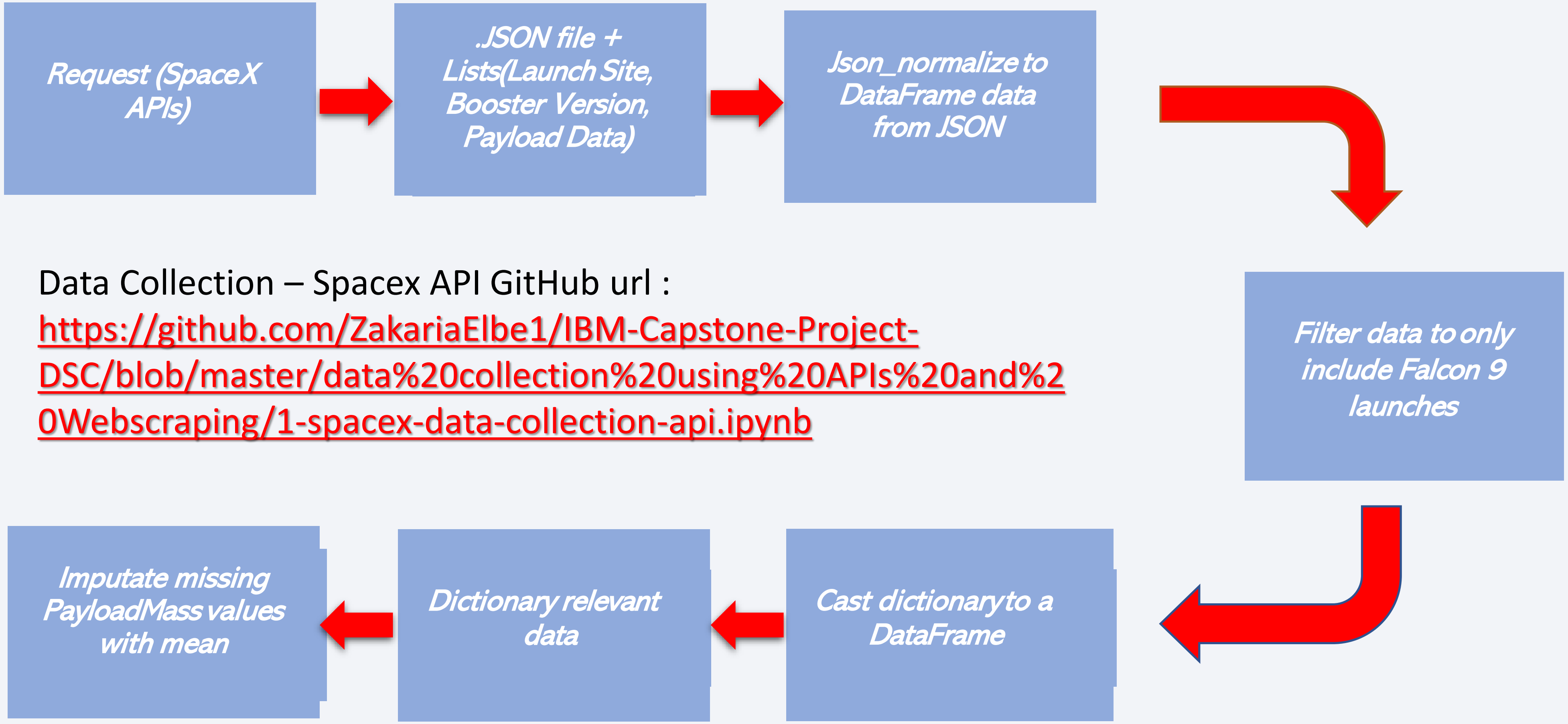
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,

Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- ***Data Columns founded by web scraping the Wikipedia Sapce X page:***

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – Spacex API



Request (SpaceX APIs)

.JSON file + Lists(Launch Site, Booster Version, Payload Data)

Json_normalize to DataFrame data from JSON

Filter data to only include Falcon 9 launches

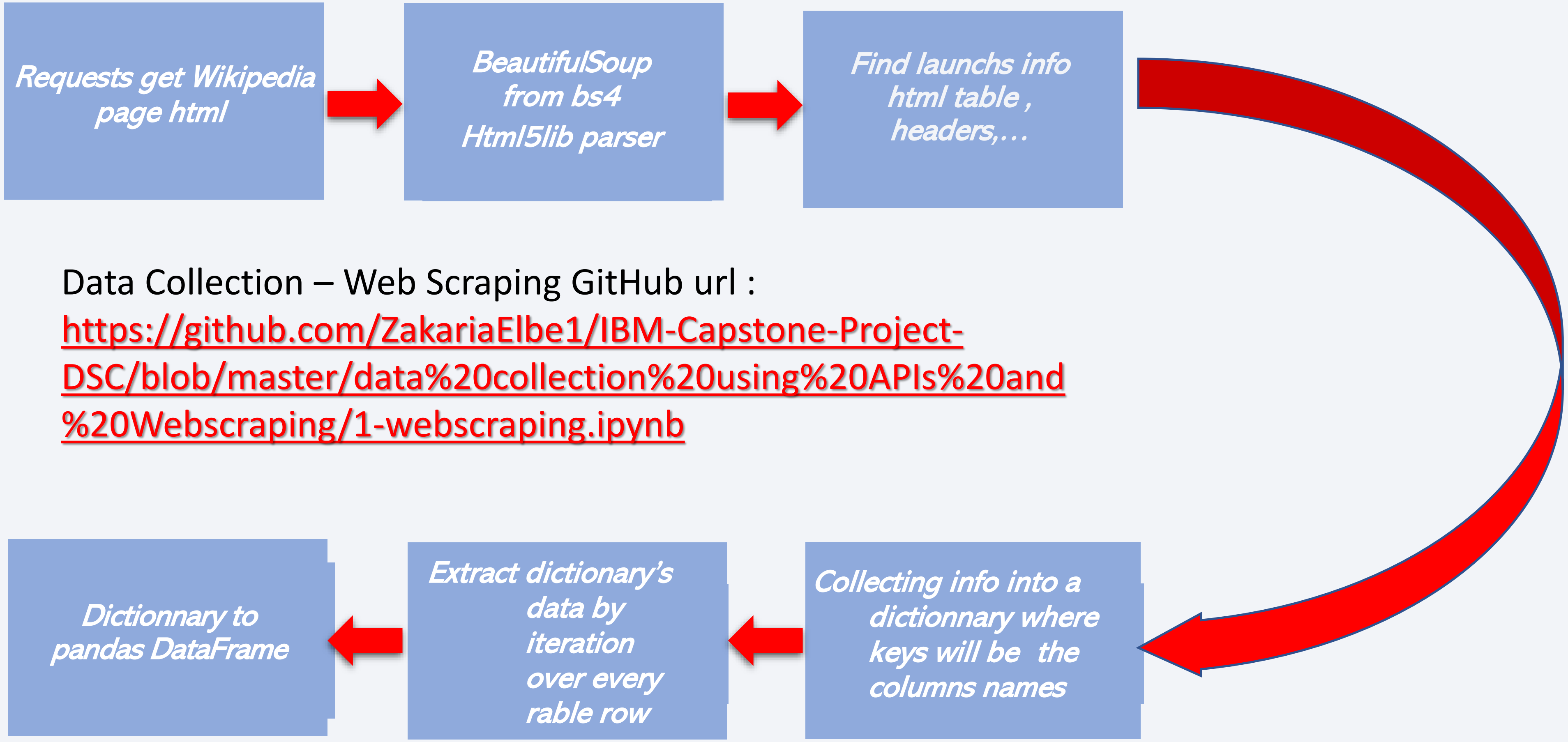
Cast dictionary to a DataFrame

Dictionary relevant data

Imputate missing PayloadMass values with mean

Data Collection – Spacex API GitHub url :
https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/blob/master/data%20collection%20using%20APIs%20and%20Webscraping/1-spacex-data-collection-api.ipynb

Data Collection – Web Scraping



Requests get Wikipedia page html

BeautifulSoup from bs4

Html5lib parser

Find launchs info html table , headers,…

Collecting info into a dictionnary where keys will be the columns names

Extract dictionary's data by iteration over every rable row

Dictionnary to pandas DataFrame

Data Collection – Web Scraping GitHub url :
https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/blob/master/data%20collection%20using%20APIs%20and%20Webscraping/1-webscraping.ipynb

Data Wrangling

- Data transformation : transforming landing outcomes feature from categorical to numeric data; successful = 1 and failure = 0.
- New label column ‘class’ was created (it’s our dependent variable or target) with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise. Value Mapping:
  - True ASDS, True RTLS, & True Ocean – set to value 1
  - None None, False ASDS, None ASDS, False Ocean, False RTLS – set to value 0

Data Wrangling GitHub url :
https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/blob/master/EDA%20of%20Falcon9%20data/2-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year. The main goal from using this process is to define the relationship between features.
- Plots Used:

- Flight Number vs. Payload Mass,
- Flight Number vs. Launch Site,
- Payload Mass vs. Launch Site,
- Orbit vs. Success Rate,
- Flight Number vs. Orbit,
- Payload vs Orbit
- Success Yearly Trend

EDA with data visualization GitHub url :
https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/blob/master/EDA%20of%20Falcon9%20data/2-eda-dataviz.ipynb

- Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model

EDA with SQL

Using SQL for Exploratory Data Analysis , we were able to:

- Understand the Spacex DataSet
- Load the dataset into the corresponding table in a Db2 database
- Execute SQL queries for better understanding our data, such as : informations between features (launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes)

EDA with SQL Github url :

https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/blob/master/EDA%20of%20Falcon9%20data/2-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

The launch success rate depend on the location and proximities of a launch site, for that , we used an interactive map to mark launch sites on map with success rate for each site. For that many objects was created and added to a map using the folium library, such as :

- **Markers** : used to mark a specific site or pinning it geographically using geographic coordinates, it's a way to interpret coordinates to visualize closeness between launch sites
- **Circles** : used to mark a site area as circle by giving a radius and color where the circle's center is a marker already created;
- **Markers cluster** : used for clustering markers on each site by grouping its by a feature, in our situation ,success/failed launch was the feature to split markers by
- **Line** : shows connection between site and a specific chosen location, but before that, we must calculate distance between the two points

Interactive map with Folium Github url :

https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/blob/master/Launch%20sites%20locations%20and%20Dashboard%20for%20analysis/lab_jupyter_launch_site_location.ipynb

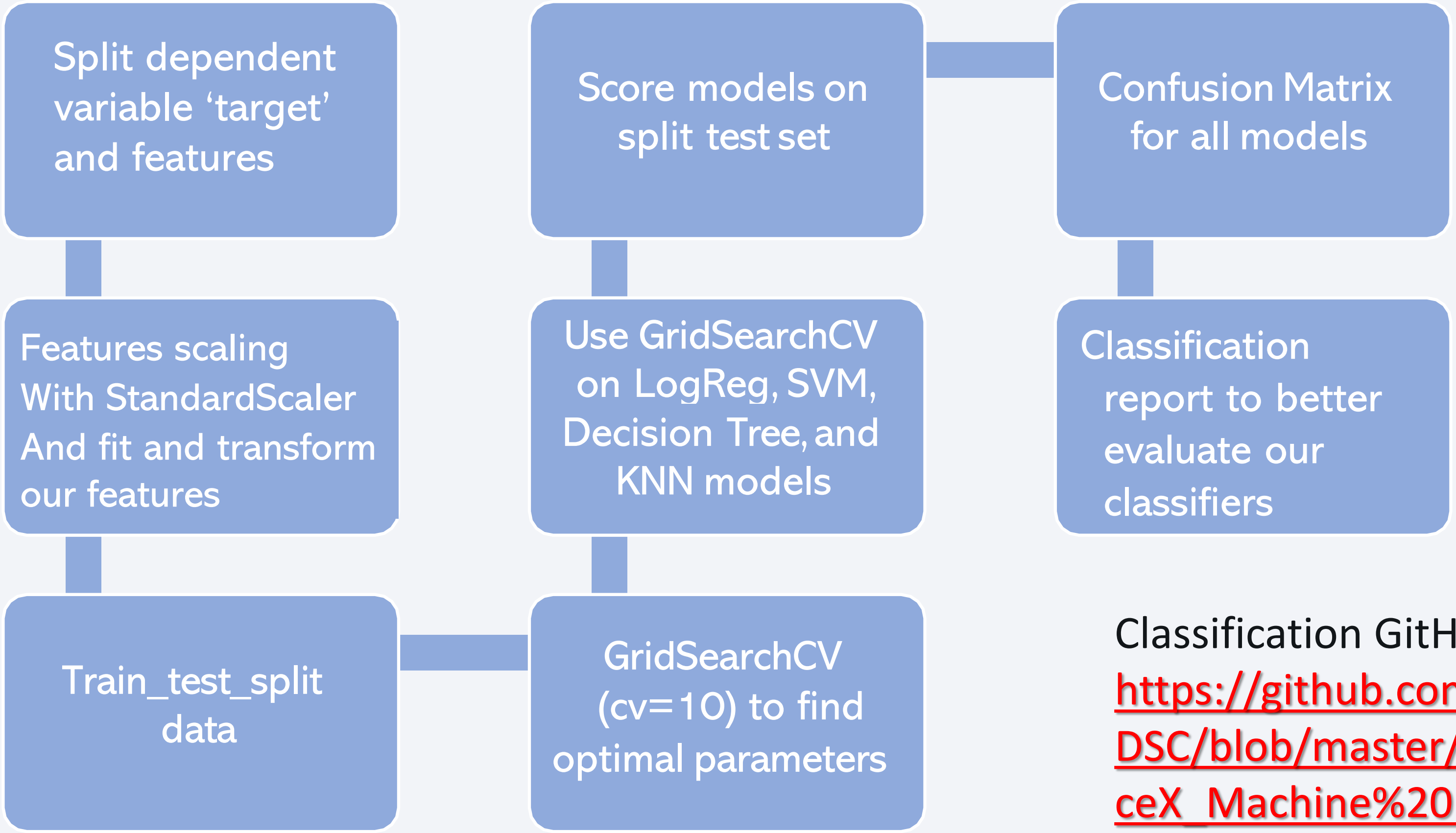
Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot by launch site, for creating our dashboard , we used “dash” library
- Pie chart created using “plotly” lib with express method, it can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot created using “plotly” lib with express method,it takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

My dashboard code - Github url :

https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/blob/master/Launch%20sites%20locations%20and%20Dashboard%20for%20analysis/my%20dashbord%20for%20the%20capstone%20ibm%20ds%20project.ipynb

Predictive Analysis (Classification)



Split dependent variable 'target' and features

Features scaling With StandardScaler And fit and transform our features

Train_test_split data

GridSearchCV (cv=10) to find optimal parameters

Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Score models on split test set

Confusion Matrix for all models

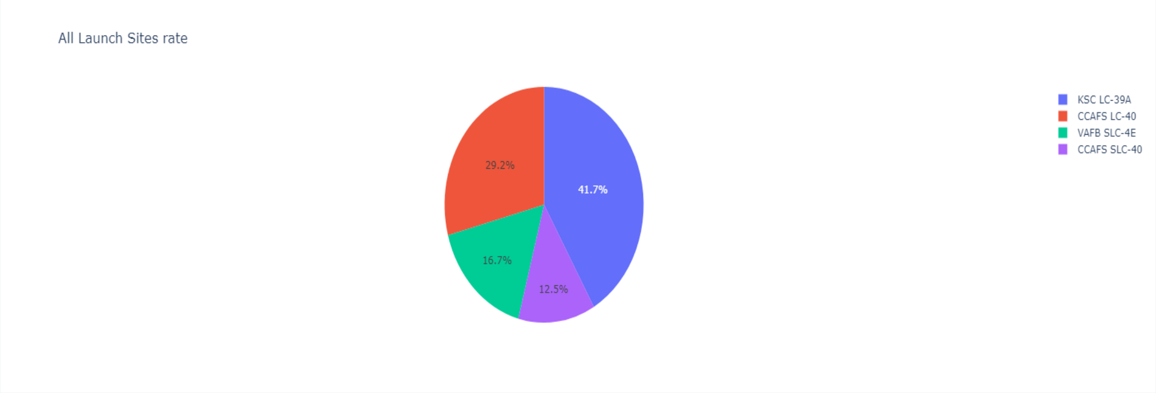
Classification report to better evaluate our classifiers

Classification GitHub url :
https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/blob/master/machine%20learning%20prediction/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

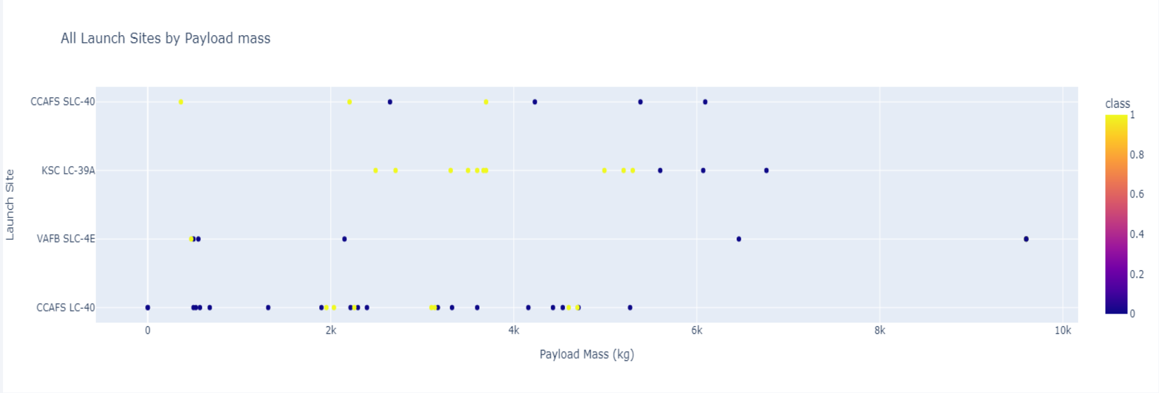
Results

SpaceX Launch Records Dashboard

Pie Chart for All sites



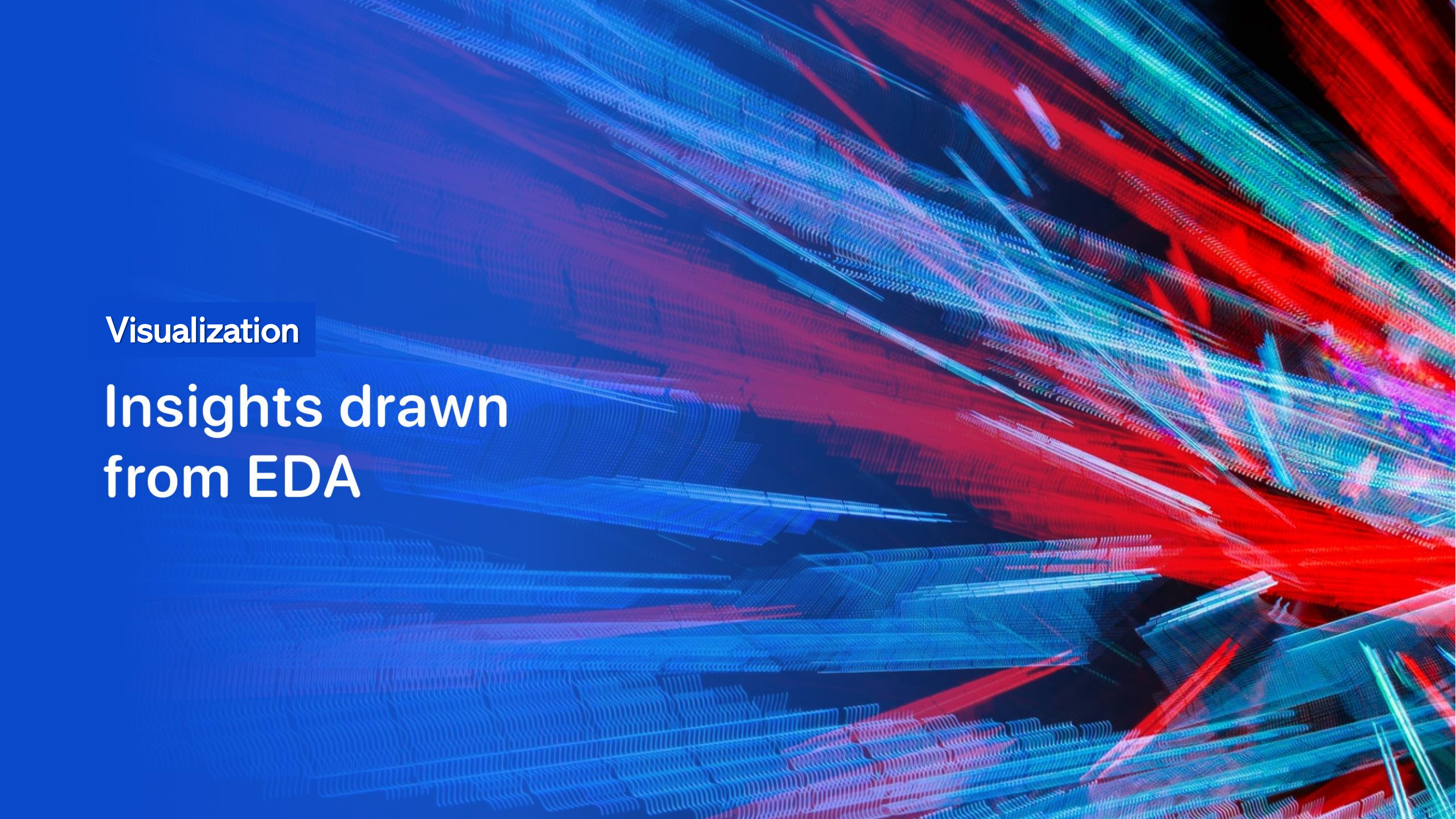
Scatter plot launch sites vs Payload mass



The graphs below belongs to our created dashbord, it shows the success rate for each launch site and the mass of the payload launched from each site.

EDA : Using The EDA process, we did some data manupilation ,engineering ,transformation and visualisation. We transormed data from categorical to numerical using one hot enconding for example , we visualize features to understand relationship between its.

Classification : Our models have the same producted accuracy with a rate of 83,34%

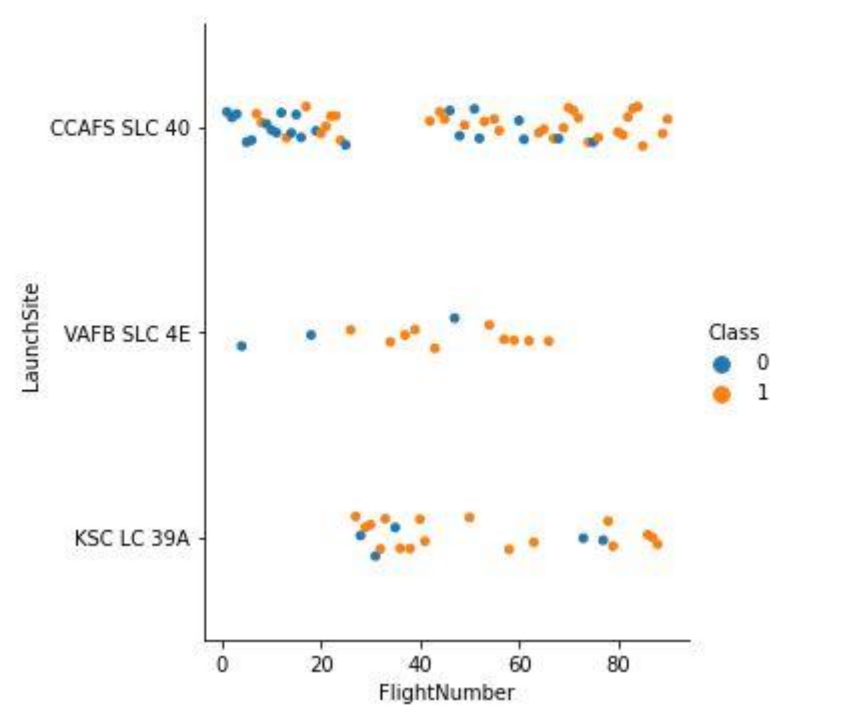


Visualization

Insights drawn from EDA

Flight Number vs. Launch Site

Flight Number vs. Launch Site



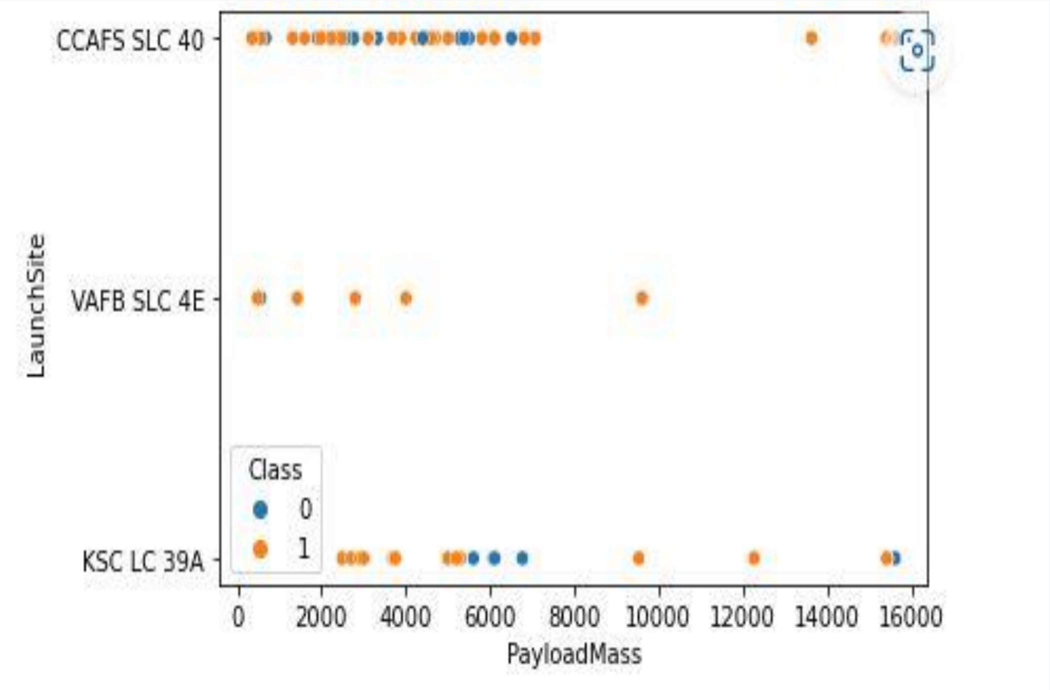
● Failure ● Success

Graph explanations :

- Graph shows that the CCAFS has more volume of flights, it appears to be the main launch site;
- After the 20th flight the success rate increase for every lauch sites, and KSC site appears to be start launching;
- The VAFB has the best success rate against small volume of flights

Payload vs. Launch Site

Scatter plot of Payload vs. Launch Site



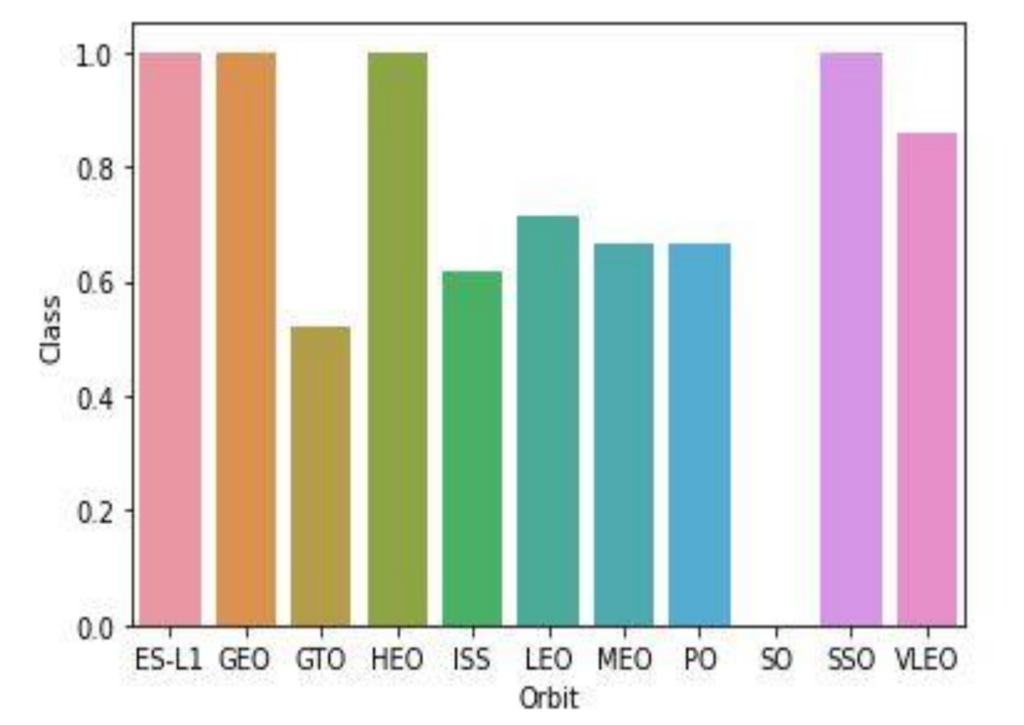
● Failure ● Success

Graph explanations :

- The majority of payloads launched from each site have a masse under 8000kg
- The payloads with a mass bigger than 8000 kg appears to have a good success rate

Success Rate vs. Orbit Type

Bar chart for the success rate of each orbit type

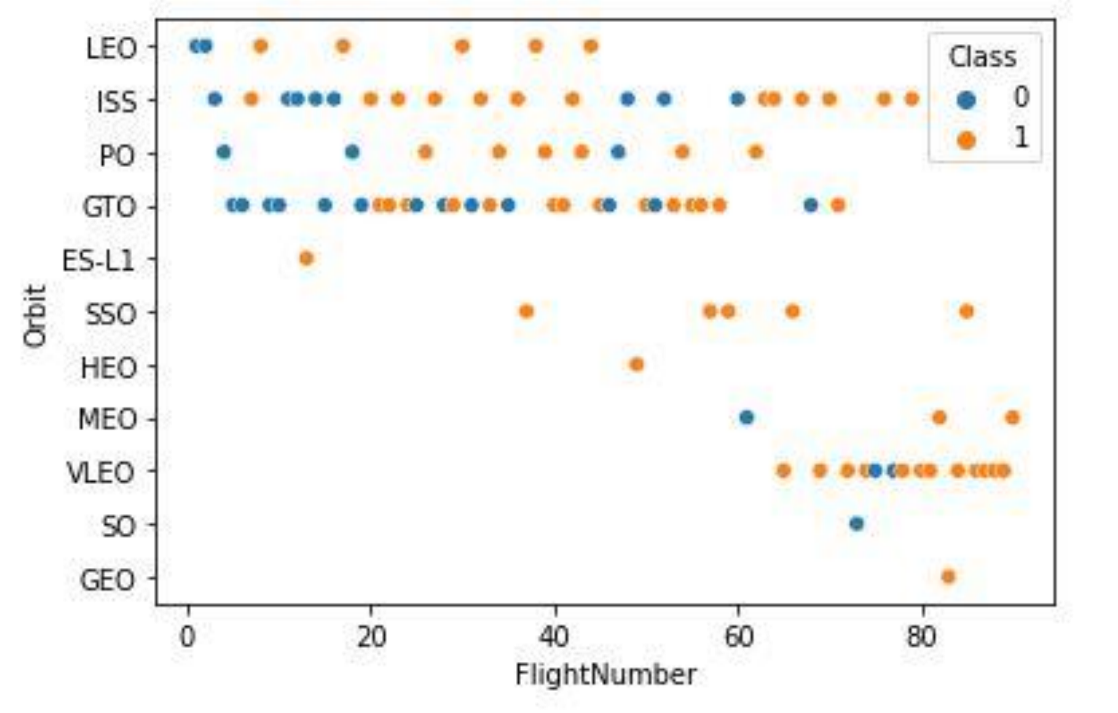


Graph explanations (sample sizes in parenthesis) :

- ES-L1 (1), GEO (1), HEO (1) have 100% success rate
- SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

Flight Number vs. Orbit Type

Scatter point of Flight number vs. Orbit type



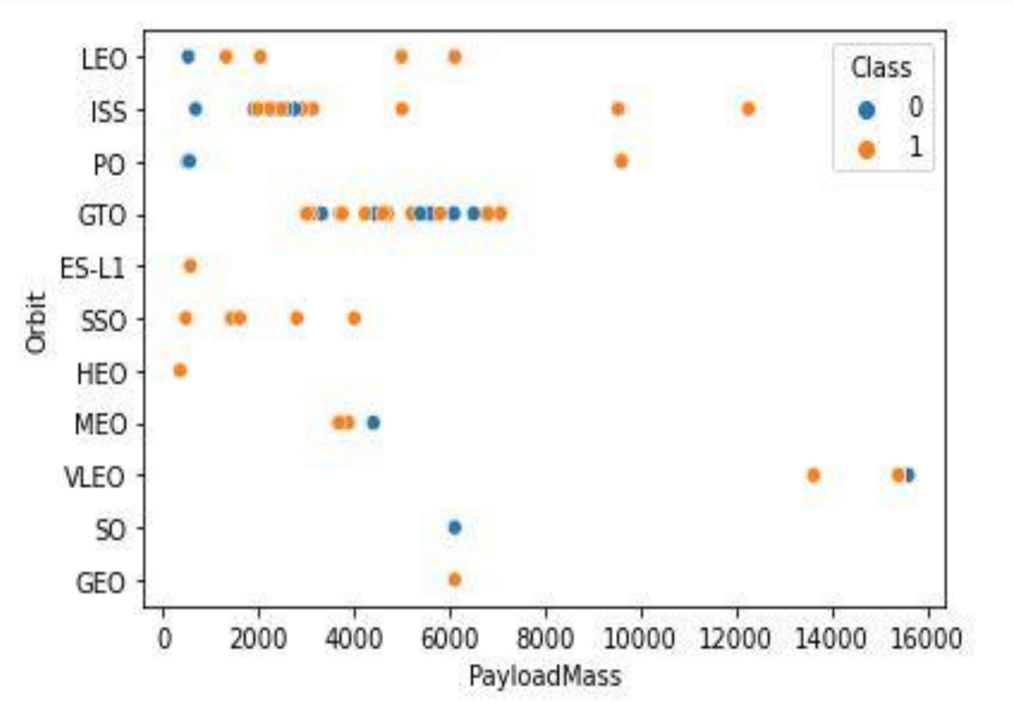
● Failure ● Success

Graph explanations :

- After 50 flight, the success rate increase on different Orbit type other than LEO
- ISS and GTO have big number of flights comparing to other Orbit type
- Flights on VLEO Orbit starts from the 60th flight
- Flights on PO stopped in the 60th and stopped on 45th on LEO Orbit
- in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type

Scatter point of payload vs. orbit type



● Failure ● Success

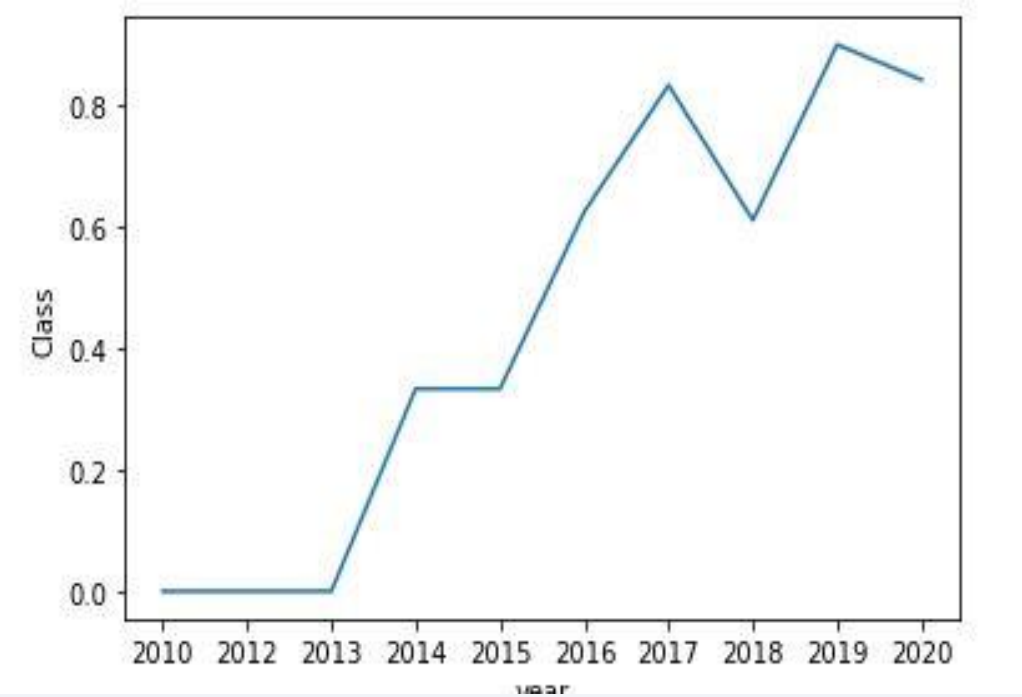
Graph explanations :

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here

Launch Success Yearly Trend

Line chart of yearly average success rate



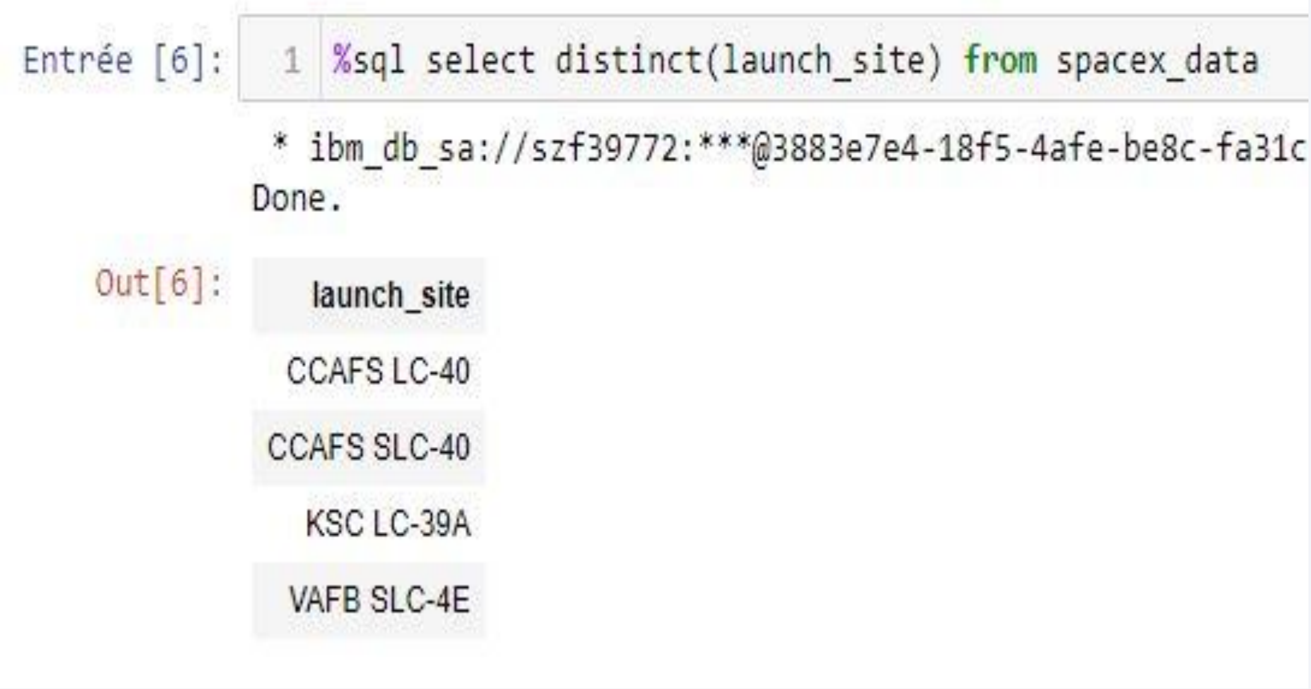
Graph explanations :

the sucess rate since 2013 kept increasing till 2020 with a breakdown in 2018

The biggest success rate was in 2019

All Launch Site Names

Query of names of the unique launch sites from database



Entrée [6]:

```
%sql select distinct(launch_site) from spacex_data
```

```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c
Done.
```

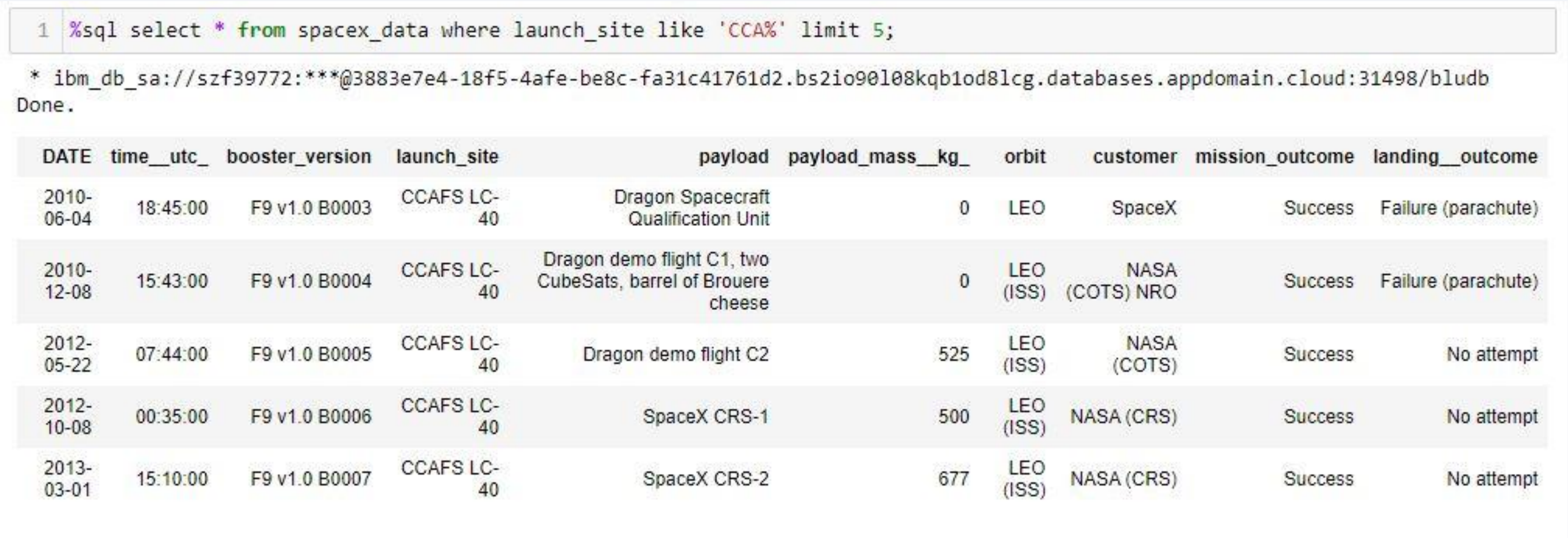
Out[6]:

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

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same
launch site with data entry errors.

Launch Site Names Begin with 'CCA'

Query of 5 records where launch sites begin with `CCA`



```
%sql select * from spacex_data where launch_site like 'CCA%' limit 5;
```

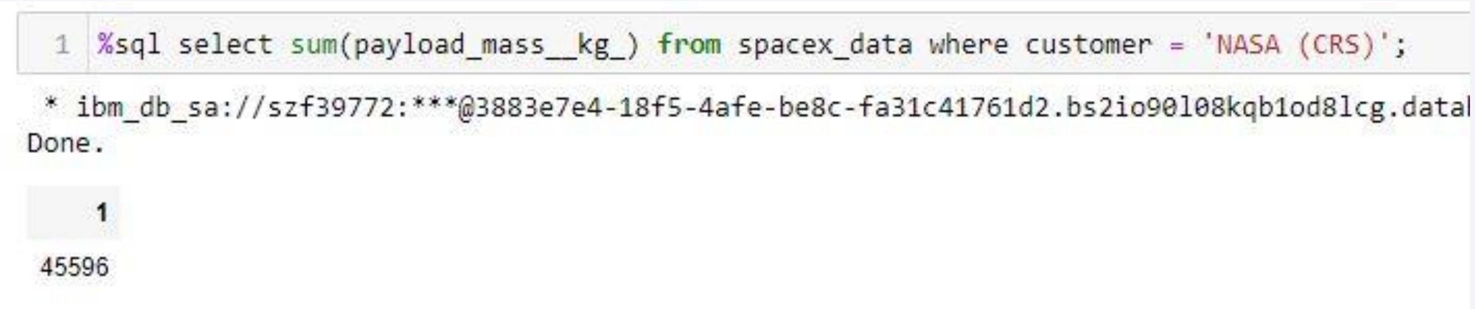
```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

All the 5 records shows a success mission outcome, from CCA launch site

Total Payload Mass

Query of total payload carried by boosters from NASA



```
%sql select sum(payload_mass__kg_) from spacex_data where customer = 'NASA (CRS)';
```

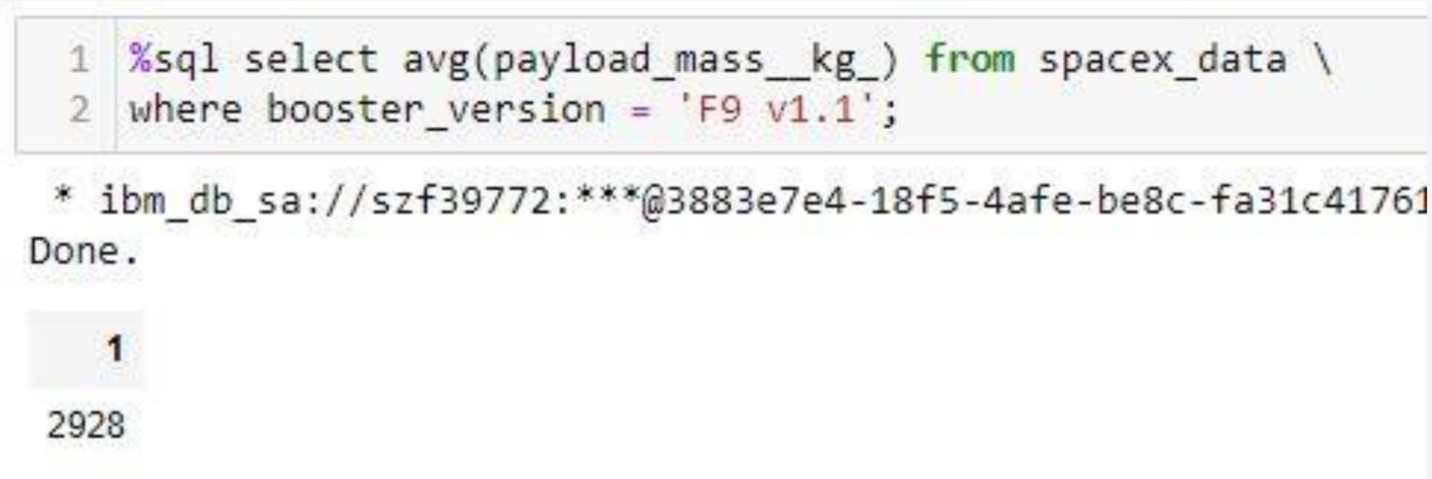
```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.datal
Done.

    1
45596
```

This query sums the total payload mass in kg where NASA was the customer.

Average Payload Mass by F9 v1.1

Query of average payload mass carried by booster version F9 v1.1



```
%sql select avg(payload_mass__kg_) from spacex_data \
where booster_version = 'F9 v1.1';
```

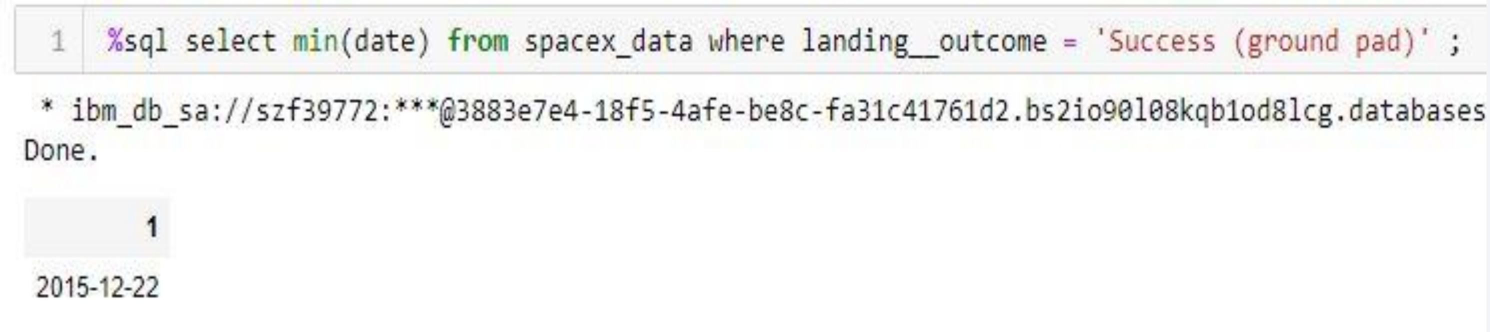
```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761
Done.

   1
2928
```

The payload mass for the booster version F9 v1.1 seems to have a light mass comparing to others booster versions

First Successful Ground Landing Date

Query of the dates of the first successful landing outcome on ground pad



```
%sql select min(date) from spacex_data where landing__outcome = 'Success (ground pad)' ;
```

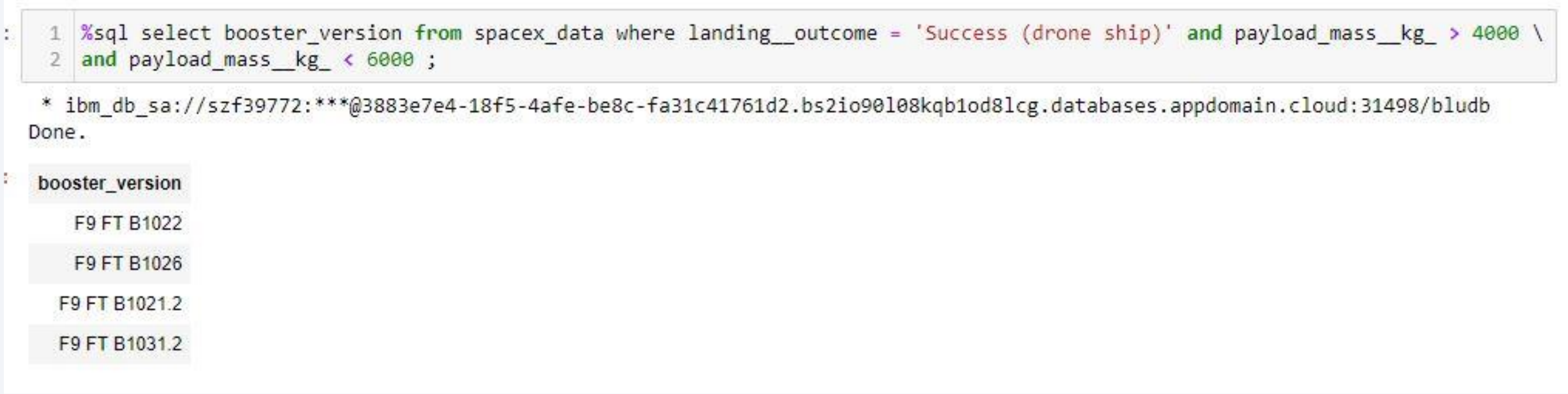
```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases
Done.
```

1
2015-12-22

The first success ground pad landing wasn’t until the end of the year 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Query of List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



```
%sql select booster_version from spacex_data where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 \
and payload_mass__kg_ < 6000 ;
```

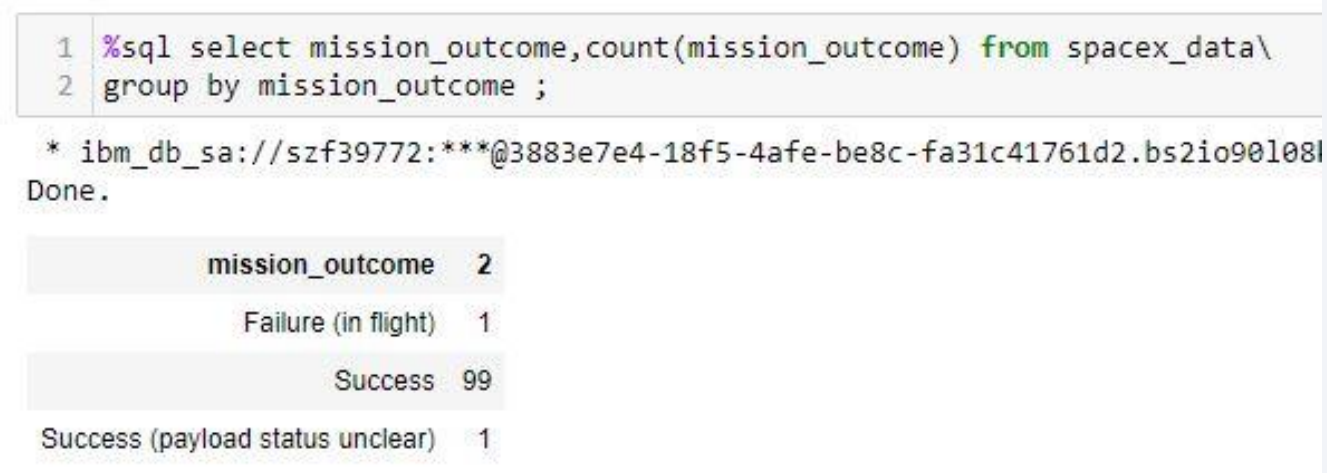
```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

The query shows the four success boosters landed on drone ship with a payload mass between 4000 and 6000 kg excluded

Total Number of Successful and Failure Mission Outcomes

Query of the total number of successful and failure mission outcomes



```
%sql select mission_outcome,count(mission_outcome) from spacex_data\
group by mission_outcome ;
```

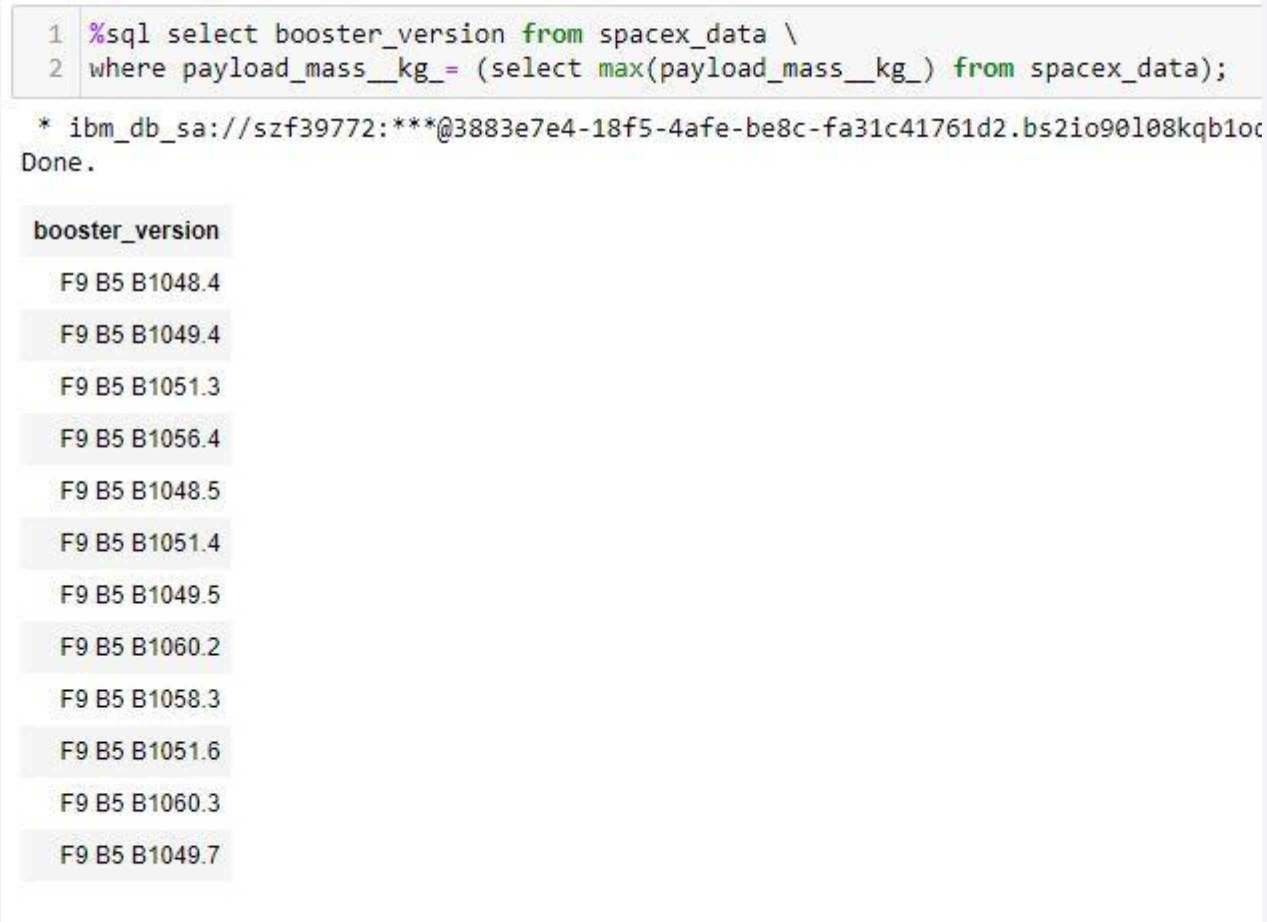
```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08
Done.
```

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

The space X rate of success mission outcome is 98.02%

Boosters Carried Maximum Payload

Query of List the names of the booster which have carried the maximum payload mass



```
%sql select booster_version from spacex_data \
where payload_mass__kg_= (select max(payload_mass__kg_) from spacex_data);
```

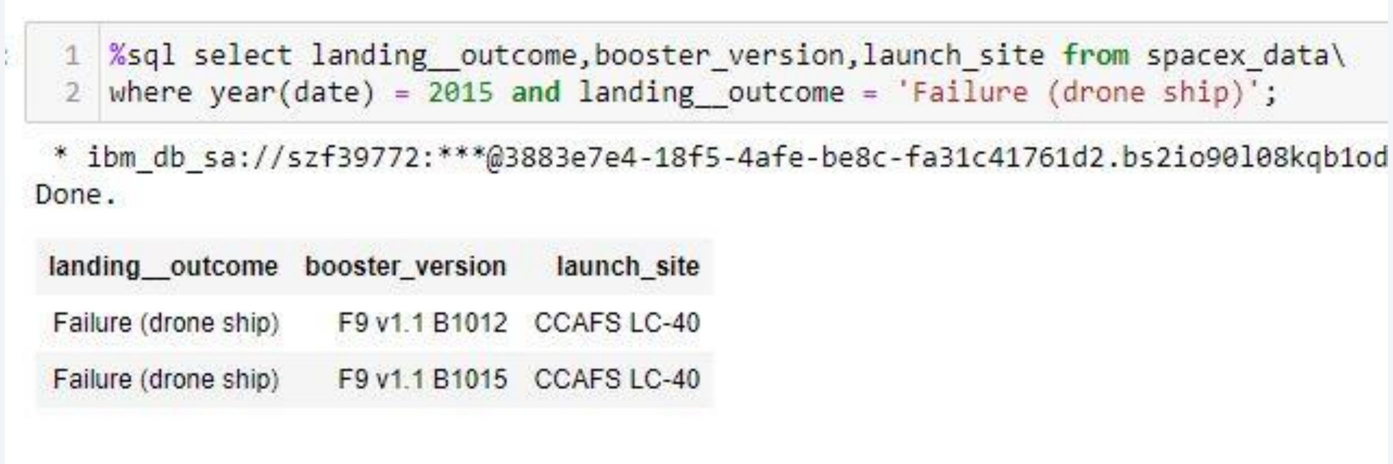
```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1oc
Done.
```

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

The booster version F9 B5 B10** appears to be the heavy ones, it shows that is a correlation between payload mass and the booster version

2015 Launch Records

Query of List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015



```
%sql select landing__outcome,booster_version,launch_site from spacex_data\
where year(date) = 2015 and landing__outcome = 'Failure (drone ship)';
```

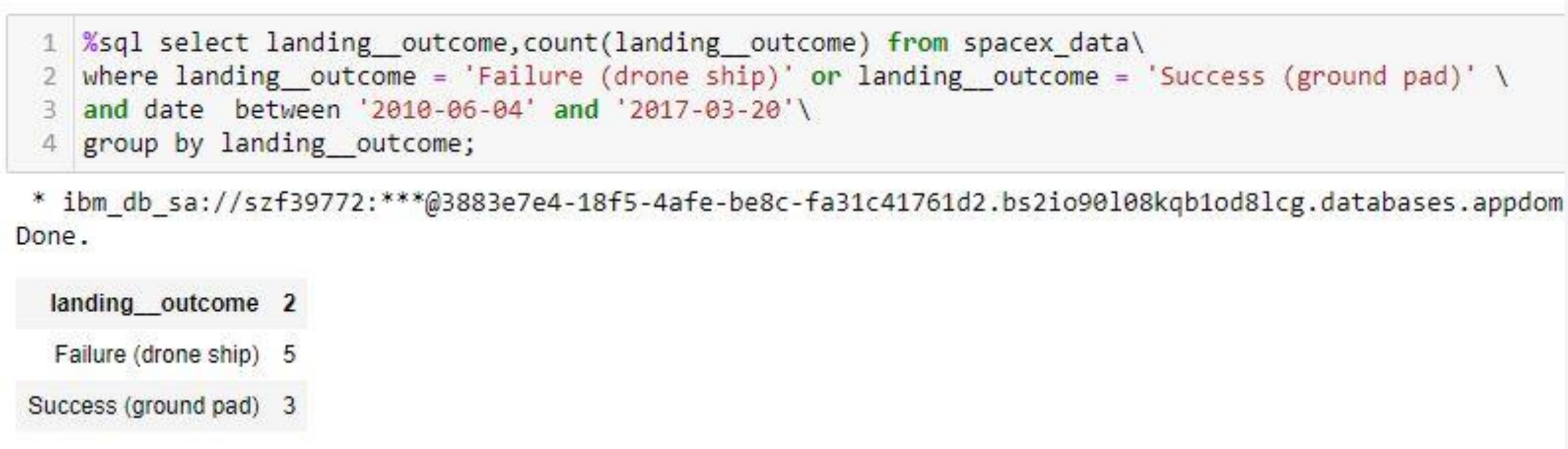
```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od
Done.
```

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

The query result shows that there is a two failed landing on drone ship both launched from the same site which is CCAFS site

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

Query of 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



```
%sql select landing__outcome,count(landing__outcome) from spacex_data\
where landing__outcome = 'Failure (drone ship)' or landing__outcome = 'Success (ground pad)' \
and date  between '2010-06-04' and '2017-03-20'\
group by landing__outcome;
```

```
 * ibm_db_sa://szf39772:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdom
Done.
```

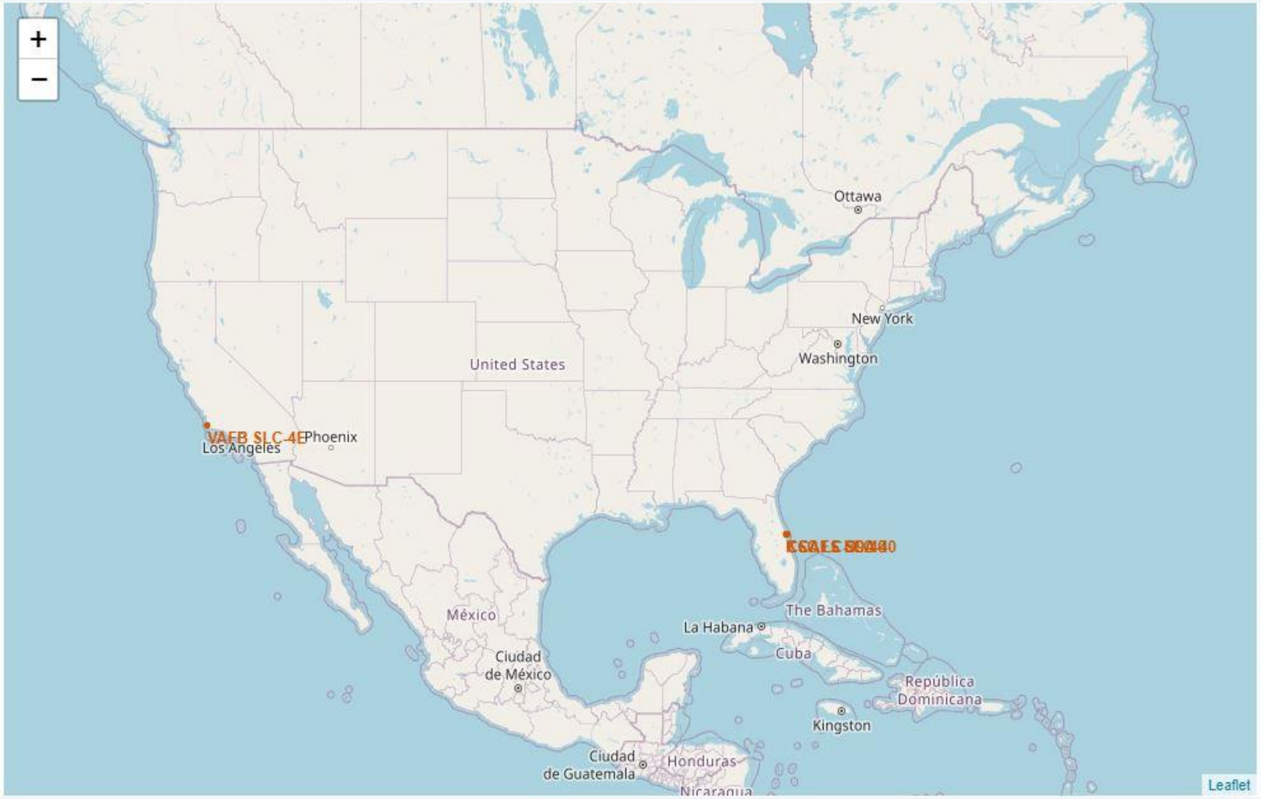
landing__outcome	2
Failure (drone ship)	5
Success (ground pad)	3

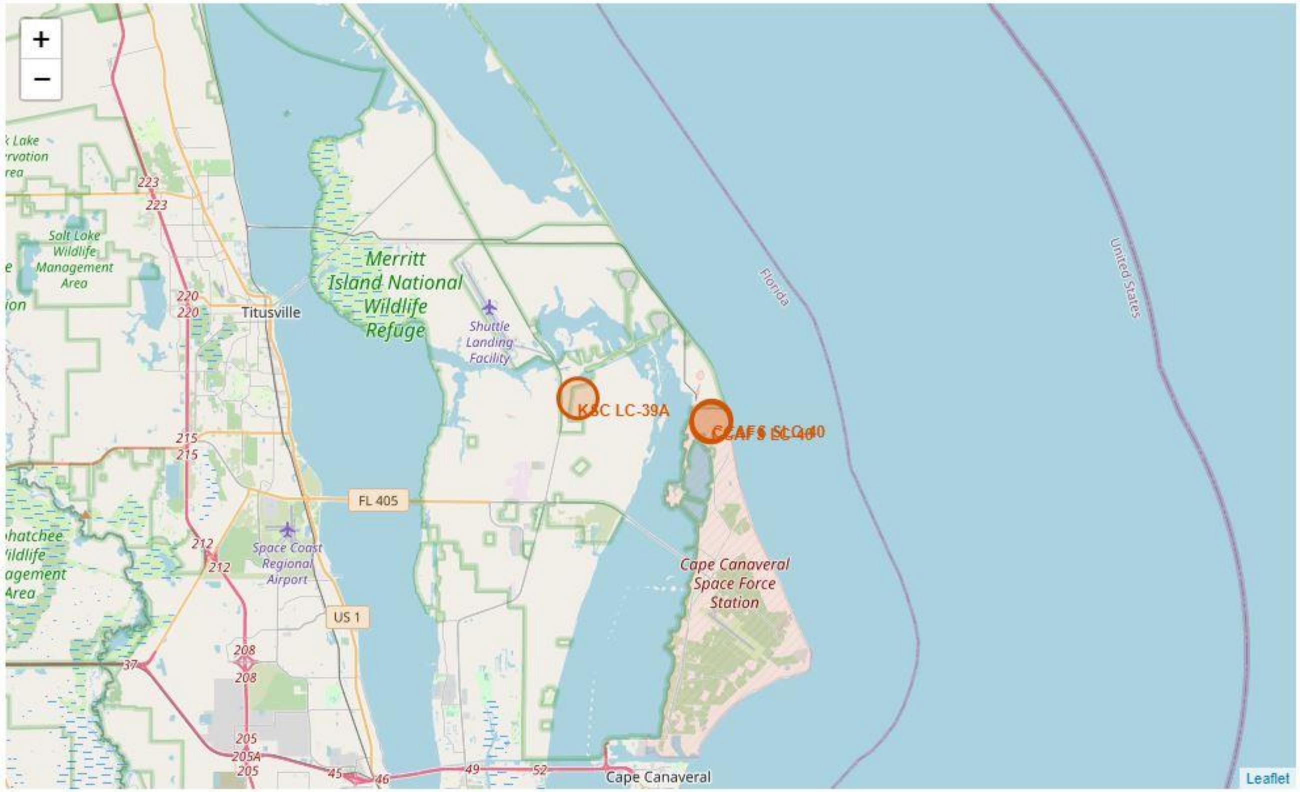
The query result shows a count of 8 of failure drone ship landing (5) and success ground pad landing(3) between the date 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

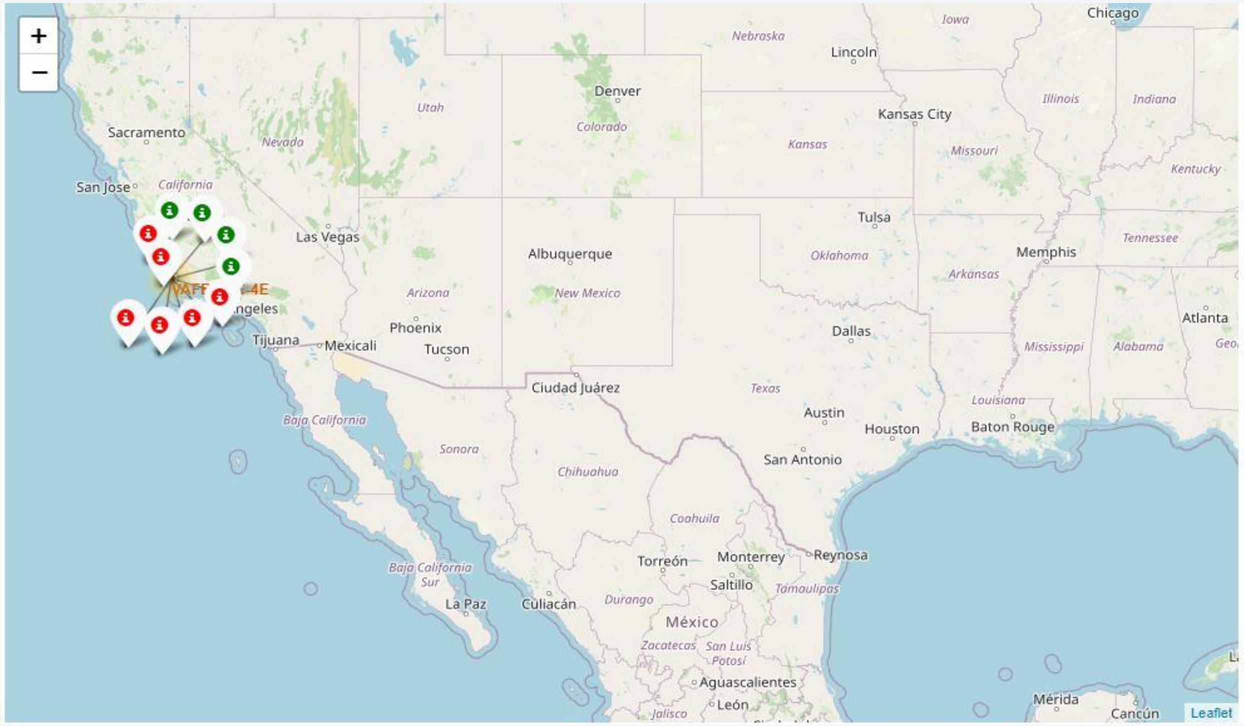
All launch sites locations on map

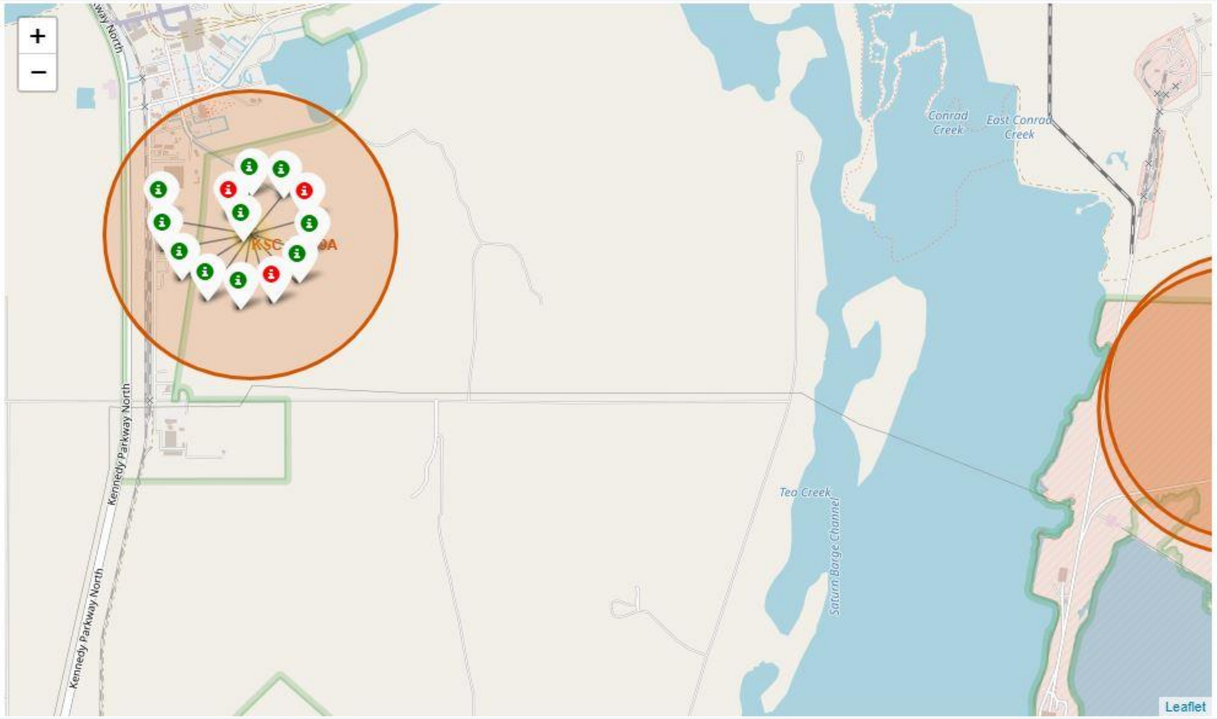



+
−
Salt Lake Wildlife Management Area
Titusville
Merritt Island National Wildlife Refuge
Shuttle Landing Facility
KSC LC-39A
CCAFS SLC-40
FL 405
Space Coast Regional Airport
US 1
Cape Canaveral Space Force Station
Cape Canaveral
Florida
United States
Leaflet

- ○ The right map produced shows that all sites are near to the coast;
- ○ The left map shows launch sites close to Florida coast
- ○ Launch sites are placed on north of the equatorial line

Labeled launch sites by success/failed launches





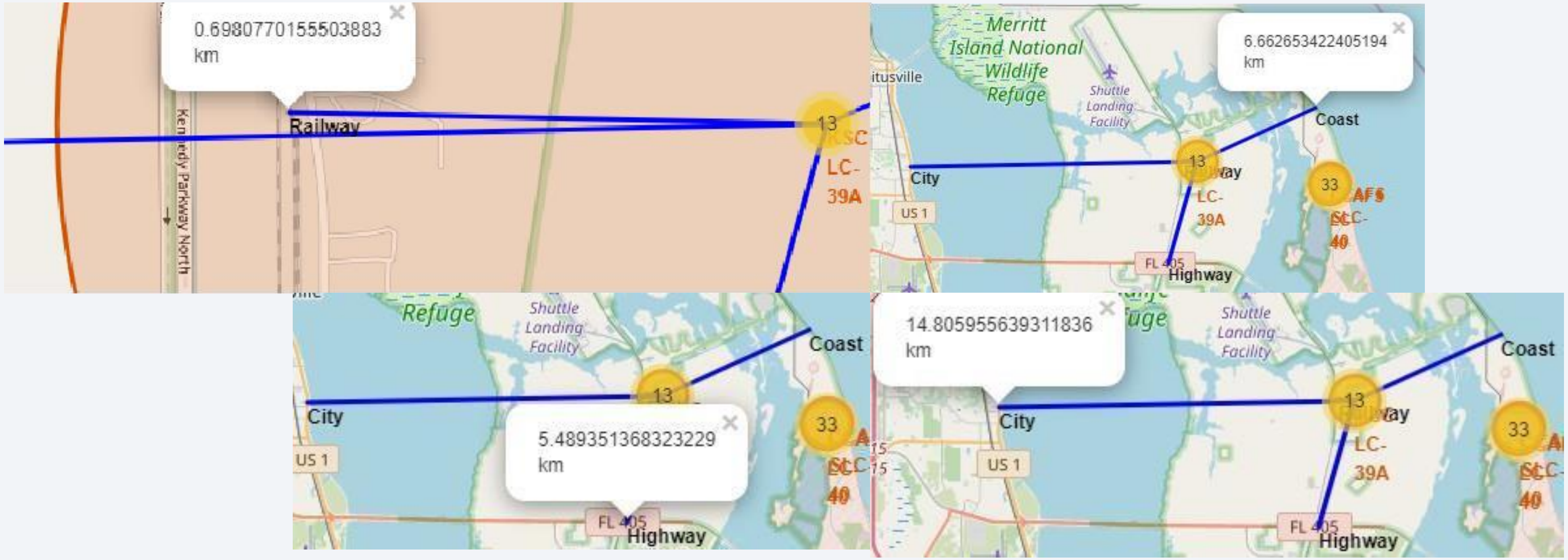
🔴 Failure 🟢 Success

In this example on the left VAFB SLC-4E shows 4 successful landings and 6 failed landings using markers cluster by folium

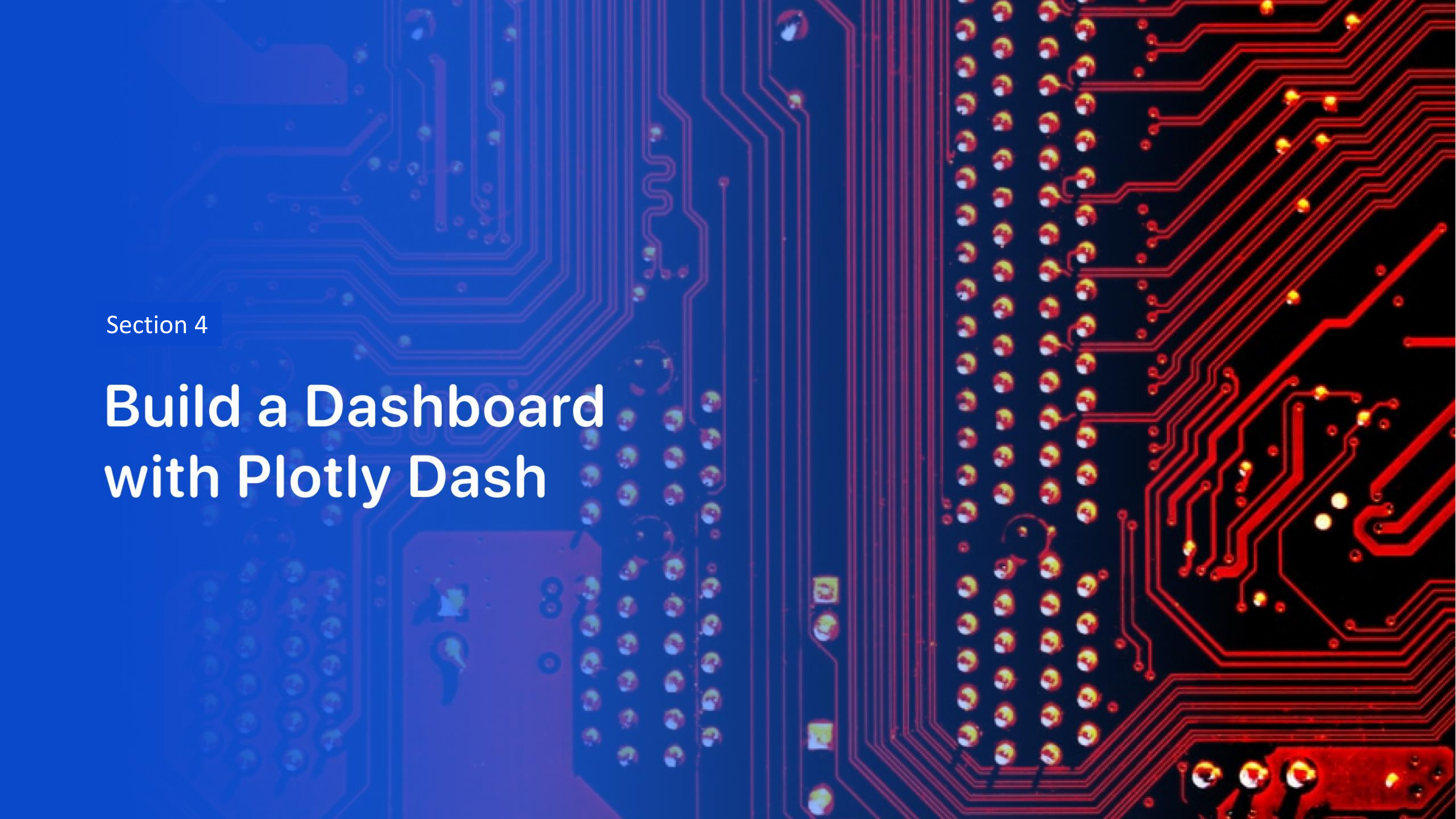
On the right KSC LC-39A shows 10 successful landings and 3 failed landings,

We can conclude that the launch sites on the east have more records and best success landing rate

Locations proximities



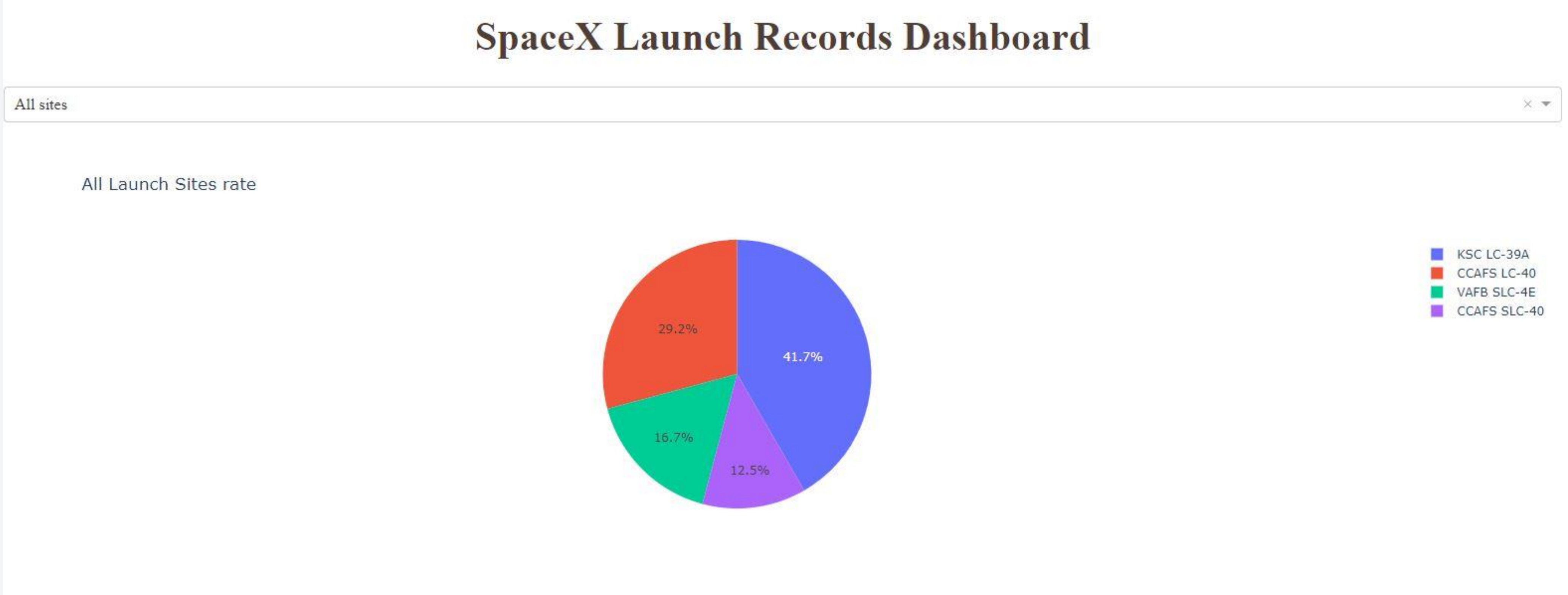
Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid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



Section 4

Build a Dashboard with Plotly Dash

Dashboard : Success rate for all sites

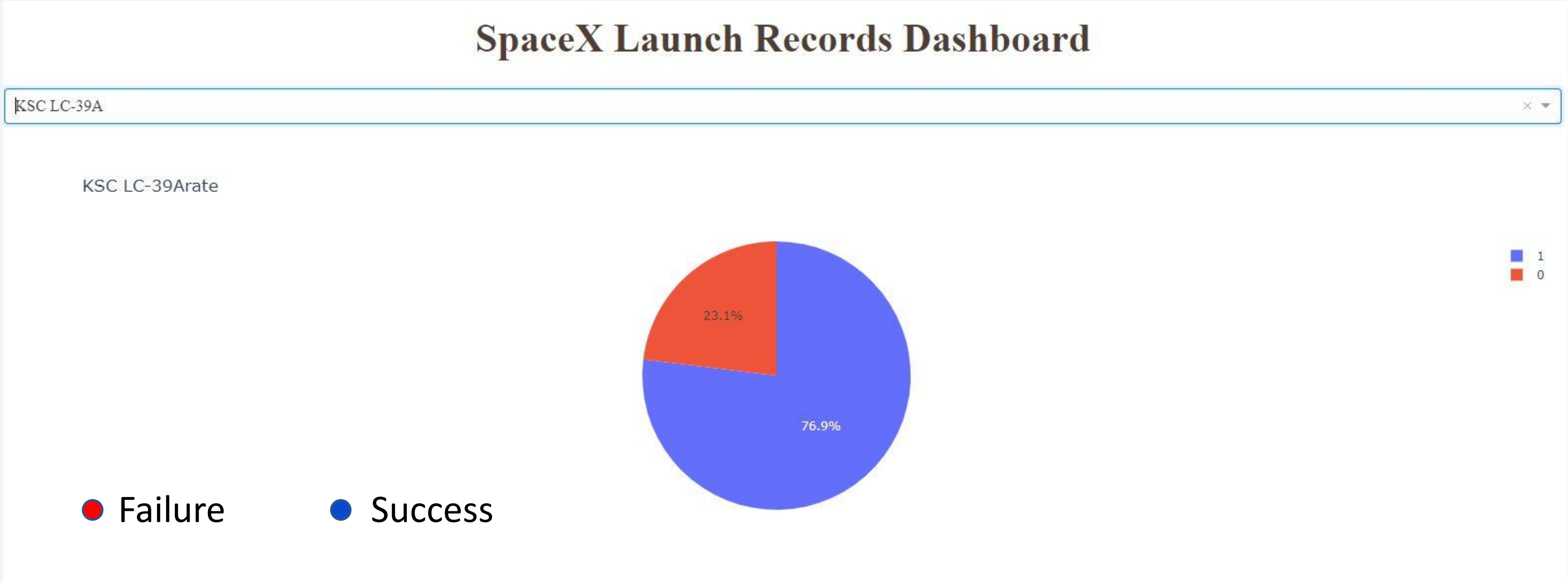


SpaceX Launch Records Dashboard

All sites

The piechart from our dashboard shows that KSC LC-39A site have the best landing success rate, it confirms the result of labeled map locations on the previous section where we concluded that the launch sites located on the east coast have a good success rate comparing to VAFB site located close to the west coast

Dashboard: Success ratio of KSC Site



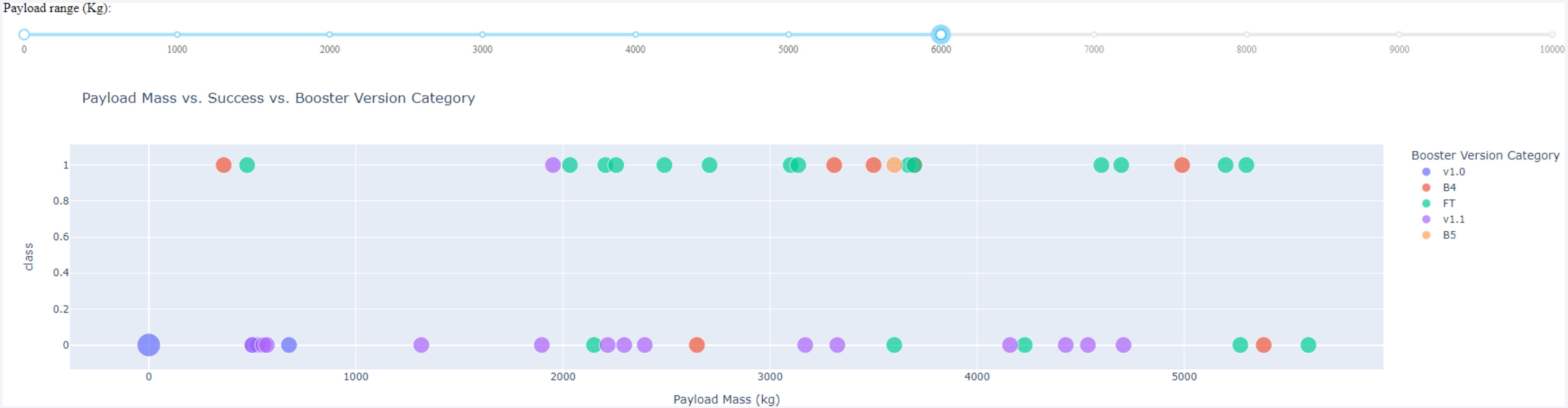
SpaceX Launch Records Dashboard

KSC LC-39A

● Failure ● Success

The pie chart of the launch site with the best success landing ratio with more than 75% success landing

Payload Mass vs. Success vs. Booster Version Category

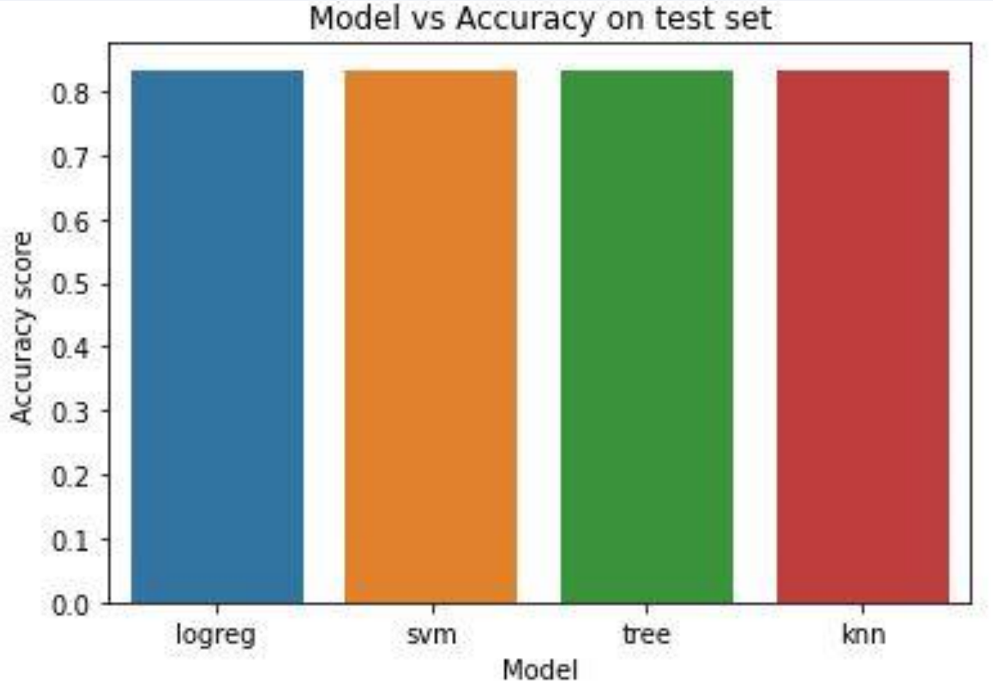


Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range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

Section 6

Predictive Analysis (Classification)

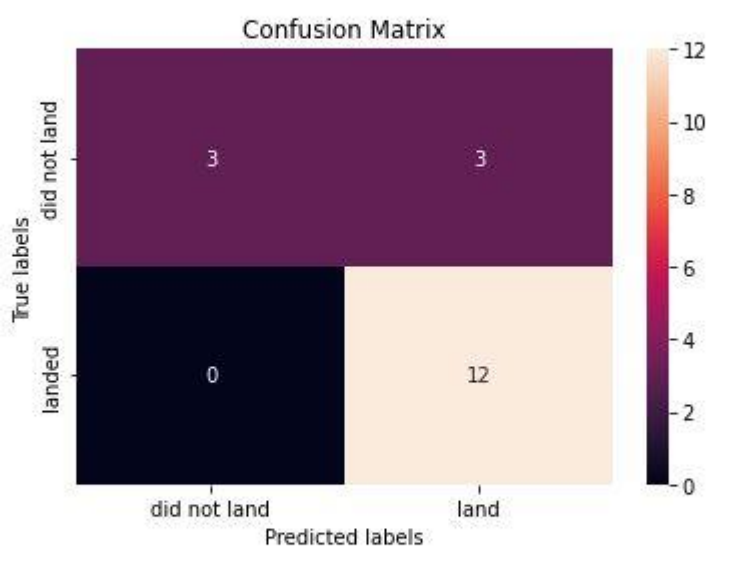
Classification Accuracy



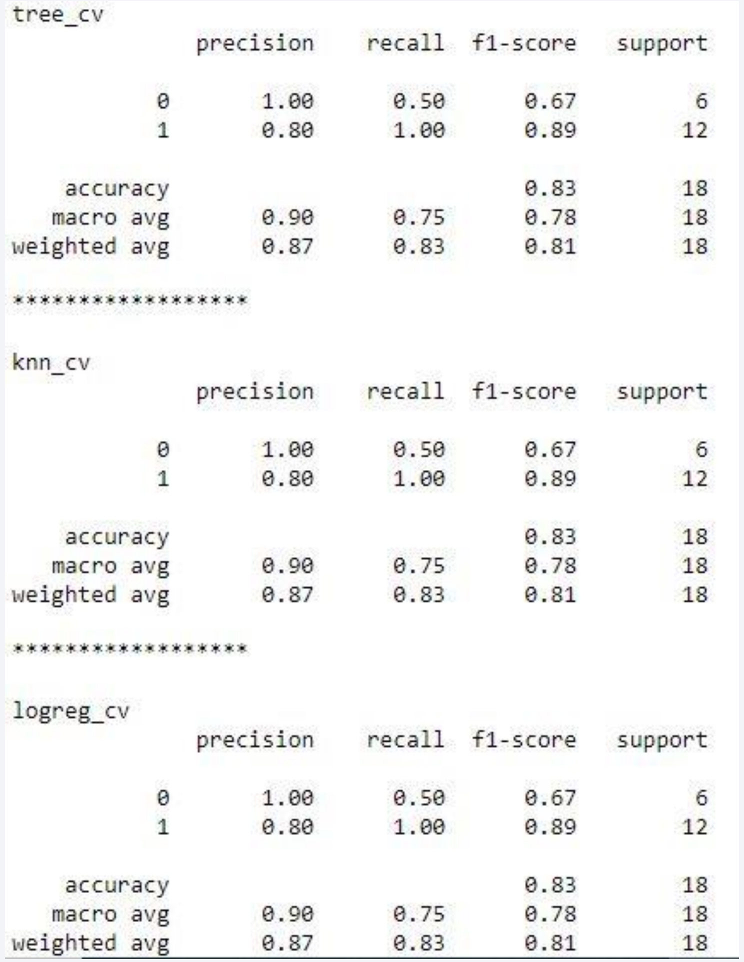
All models had the same accuracy on the test set at 83.33% accuracy. We could explain that by the small test size at only sample size of 18.
It should be noted that the Decision tree model with more than one run produced multiple accuracy scores, it depends on quality of split that the model did.
We likely need more data to determine the best model.

Confusion Matrix

Since the four models had the same accuracy, precision, recall and F1_score on the test set , we displayed only one confusion matrix



The 4 models had missclassify 3 failed landing as a success landings and they classified the rest successfuly



```
tree_cv
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         6
           1       0.80      1.00      0.89        12

    accuracy                           0.83        18
   macro avg       0.90      0.75      0.78        18
weighted avg       0.87      0.83      0.81        18

******************

knn_cv
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         6
           1       0.80      1.00      0.89        12

    accuracy                           0.83        18
   macro avg       0.90      0.75      0.78        18
weighted avg       0.87      0.83      0.81        18

******************

logreg_cv
              precision    recall  f1-score   support

           0       1.00      0.50      0.67         6
           1       0.80      1.00      0.89        12

    accuracy                           0.83        18
   macro avg       0.90      0.75      0.78        18
weighted avg       0.87      0.83      0.81        18
```

Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

Appendix

All the files used in this presentation could be found by checking this Github url :

https://github.com/ZakariaElbe1/IBM-Capstone-Project-DSC/tree/master

And finally i want to thank all the instrutors, coursera platform and all of you guys

Thank you!