Data Science Capstone project

Ahmet Bolat

02.09.2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Analyzing cost of a launch of a competitive company called SpaceX
- Determining the price of each launch.
- Determining if SpaceX will reuse the first stage.

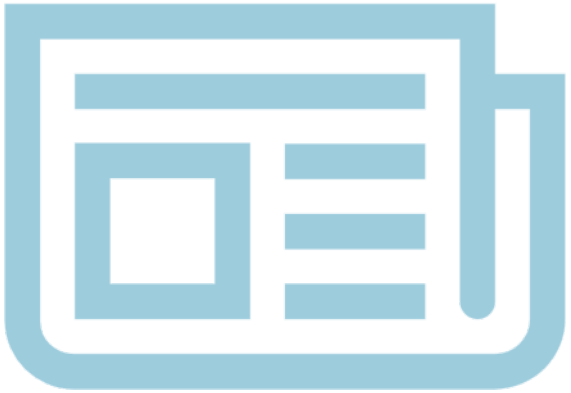
- With respect to data used, machine learning methods such as KNN, SVM and Decision Tree are found equal best-fitters to determine whether SpaceX will land successfully or not.

Introduction

- Project background and context
- Using machine learning models, dashboard of plotly dash, SQL queries, Python notebooks and public information whether SpaceX will land successfully and reuse the first stage.

- Problems you want to find answers
- If we know first stage land, we can determine the cost of the launch

Methodology



- Data collection methodology:
  - Data was collected from the SpaceX REST API.
  - (https://api.spacexdata.com/v4/.)
- Perform data wrangling

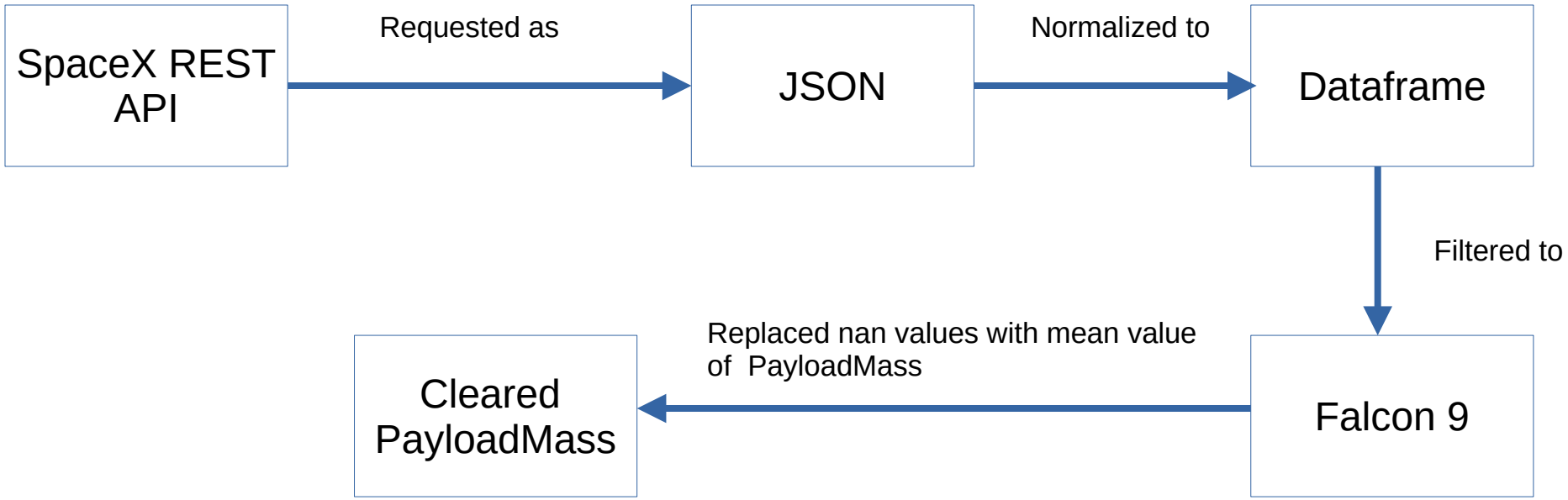
  Normalizing JSON data and making available for dataframe

  . Sampling Data

  dealing with Nulls
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Methodology

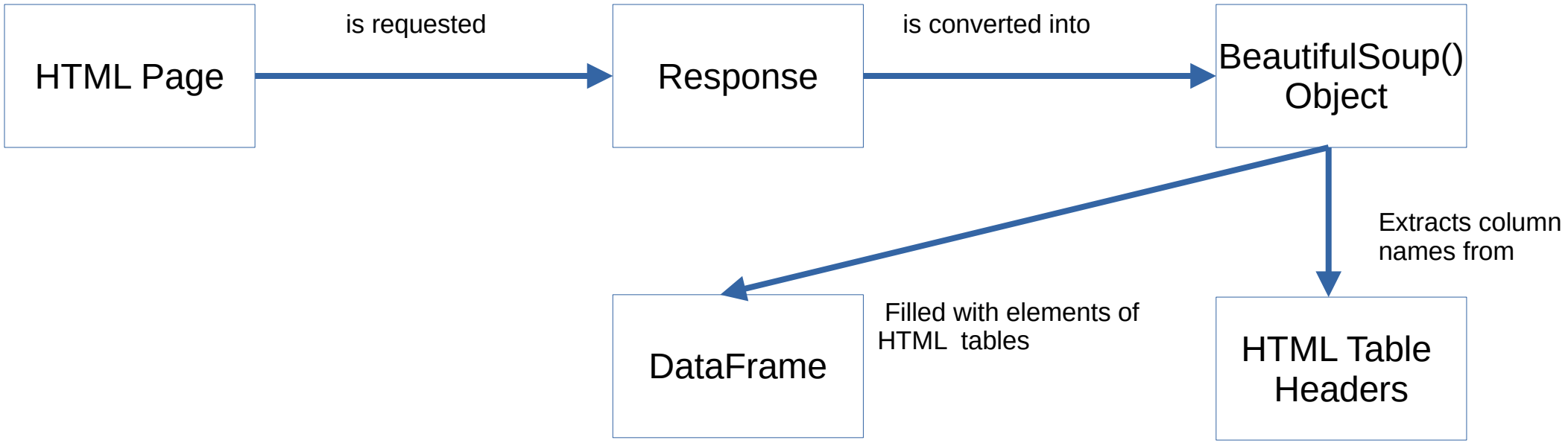
Data Collection - SpaceX API



Please see the notebook:

https://github.com/bolatah/applied-data-science-capstone/blob/master/data_collection_API%20.ipynb

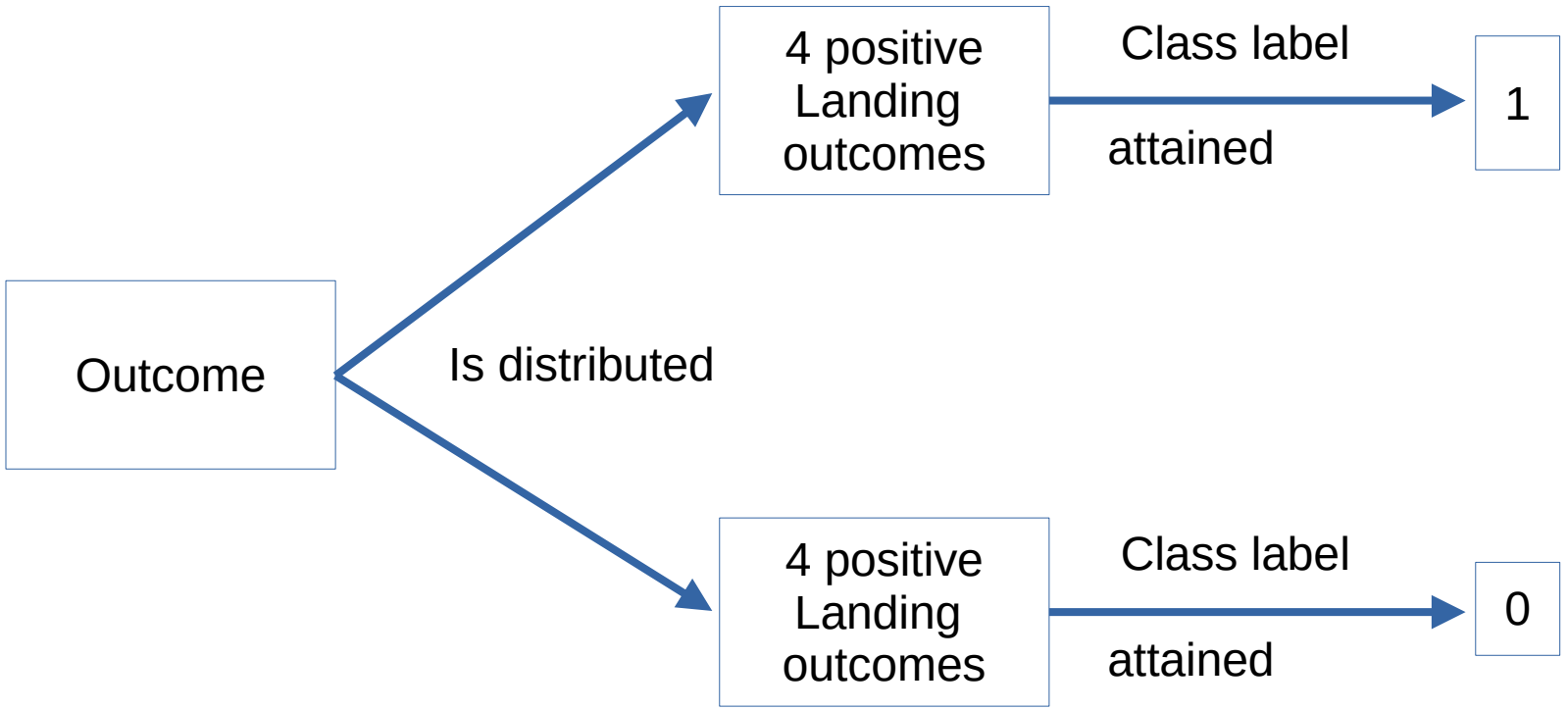
Data Collection – Web Scraping



Please see the notebook:

https://github.com/bolatah/applied-data-science-capstone/blob/master/data_collection_web_scrapping.ipynb

Data Wrangling



Please see the notebook:

https://github.com/bolatah/applied-data-science-capstone/blob/master/data_wrangling_spacex.ipynb

EDA with data visualization

- Scatter point chart:
  - In order to visualize the relationship between the variables better
- Bar chart:
  - It is suitable esp. for values of categorical variables.
- Line chart:

  for visualizing trends esp. yearly trends.

https://github.com/bolatah/applied-data-science-capstone/blob/master/eda_dataviz.ipynb

EDA with SQL – Queries

- DISTINCT
- *
- SELECT, FROM, WHERE
- AVG()
- MIN(), MAX()
- AND, OR
- MONTHNAME()
- LIKE
- ORDER BY
- DESC, ASC

https://github.com/bolatah/applied-data-science-capstone/blob/master/eda_sql.ipynb

Build a Dashboard with Plotly Dash

- Pie chart:
  - To show the successful launches count for all sites together and unique sites.
- Scatter chart:
  - To show the correlation between payload and launch success.
- Dropdown:
  - To enable Launch Site selection
- RangeSlider:
  - To select payload range

https://github.com/bolatah/applied-data-science-capstone/blob/master/spacex_dash_app.py

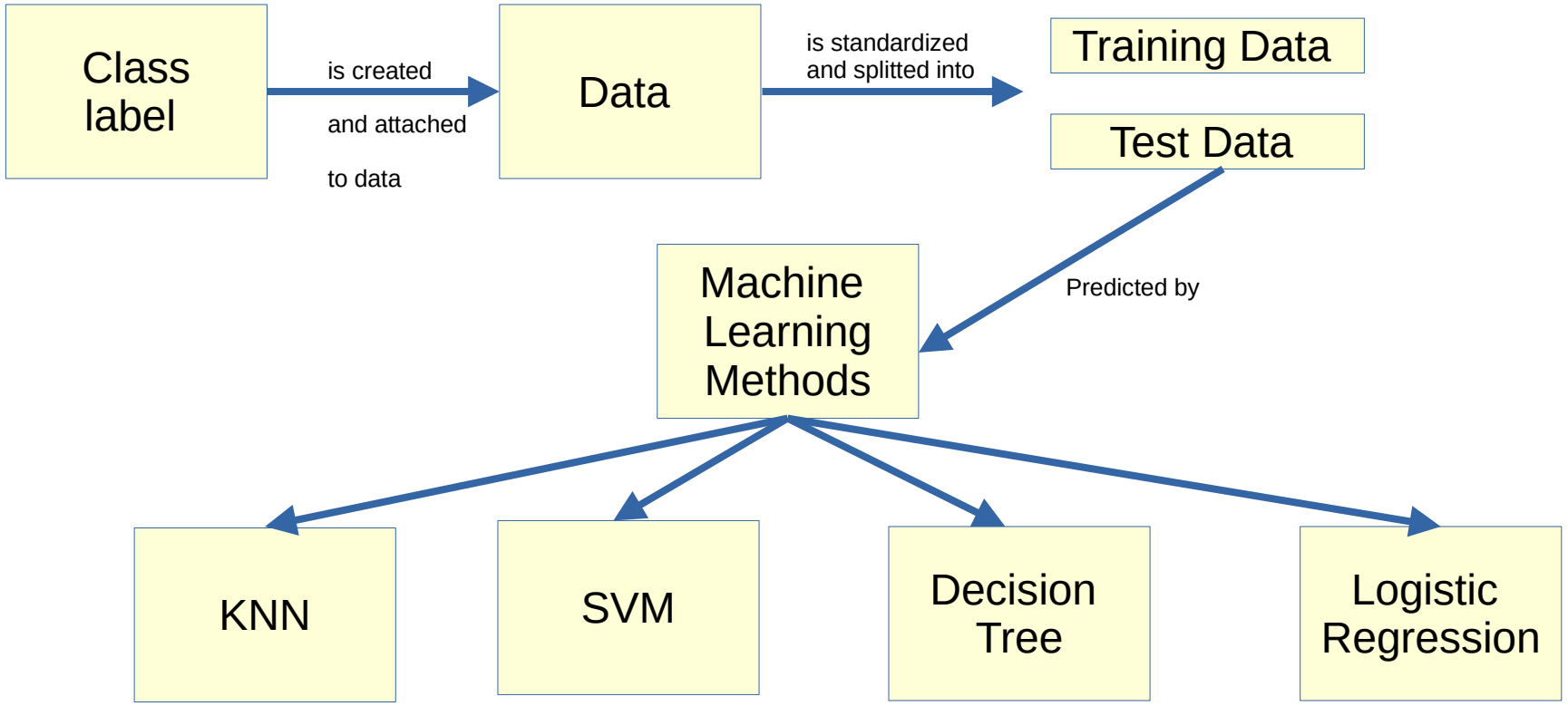
Build an interactive map with Folium

Folium Map objects used in project:

- Marker: To add a marker on the coordinate specified
- Circle: To add a circle on the coordinate specified
- PopUp: To add a popup label showing the name of the Launch Site
- CircleMarker: To add a marker in shape of circle.
- MarketCluster: To add many markers which have the same coordinates
- MousePosition: To get the coordinate for a mouse over a point on the map
- DivIcon: To add an icon as a text label on the coordinate

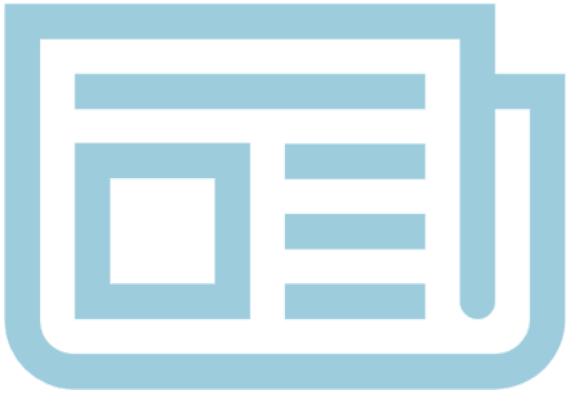
https://github.com/bolatah/applied-data-science-capstone/blob/master/folium_launch_site.ipynb

Predictive analysis (Classification)



https://github.com/bolatah/applied-data-science-capstone/blob/master/machine_learning_prediction.ipynb

Results



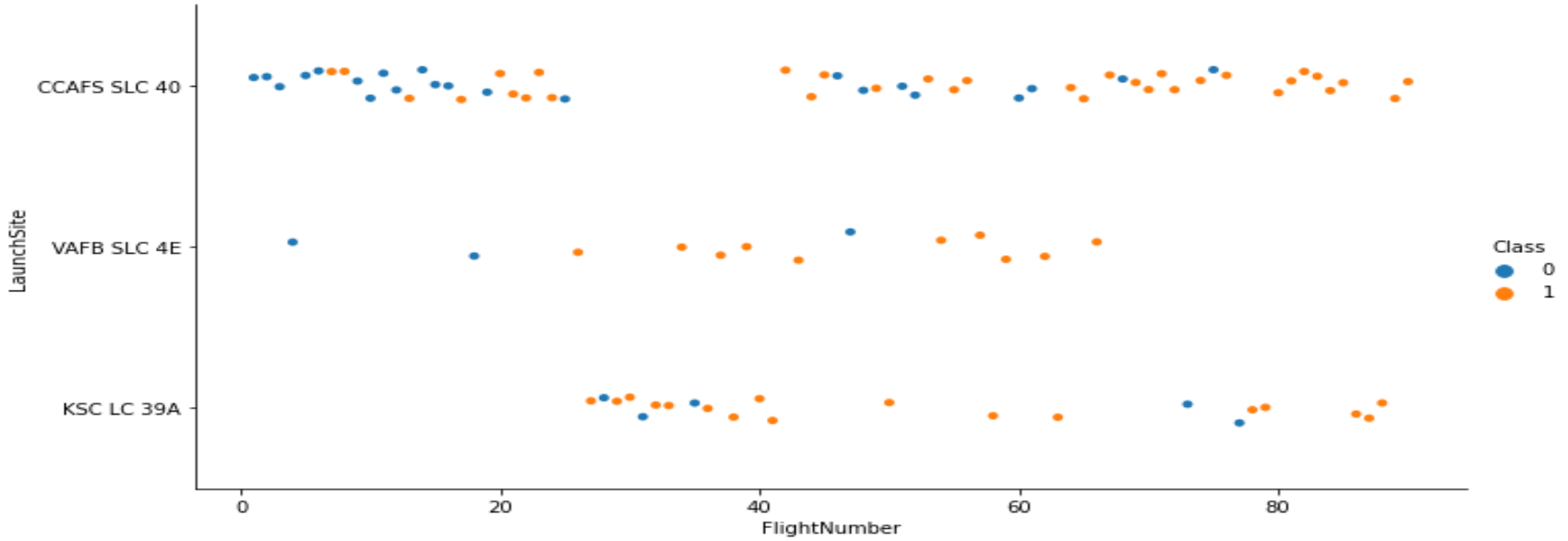
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

Flight Number vs. Launch Site

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y='LaunchSite',
            x='FlightNumber',
            hue='Class',
            data=df,
            aspect= 2)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f48e89da130>
```

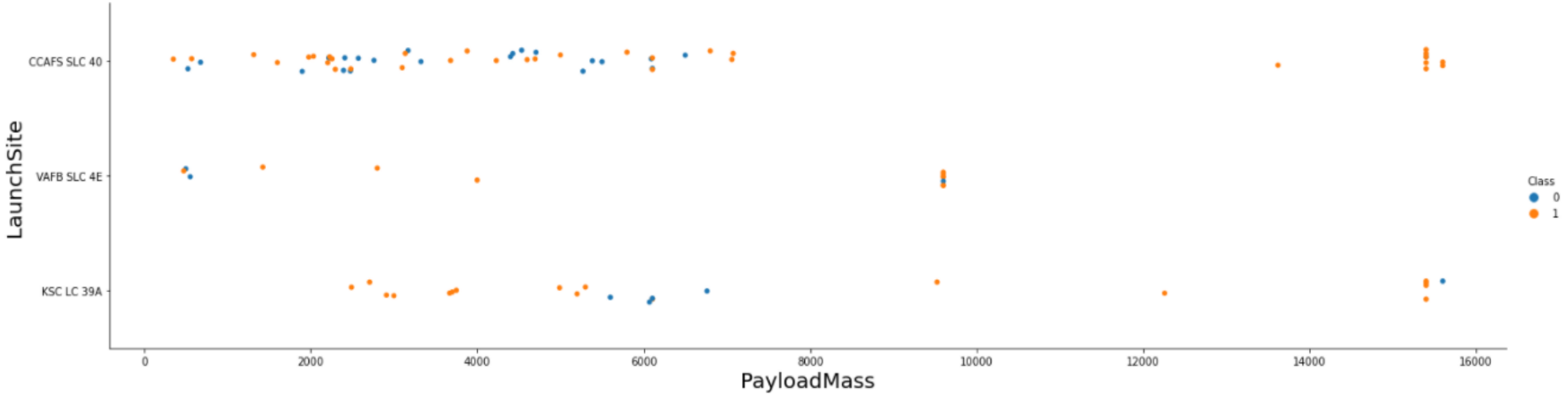


KSC LC 39A has relatively more class 1 values than its class 0 values in comparison to other launch sites. Moreover, as the flight number increases, the success rate increase.But the chart does not still give a clear idea about the relation between launch flight number and clas .

Payload vs. Launch Site

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(x='PayloadMass', y='LaunchSite', hue='Class',data=df, aspect=4)
plt.xlabel("PayloadMass",fontsize=20)
plt.ylabel("LaunchSite",fontsize=20)
```

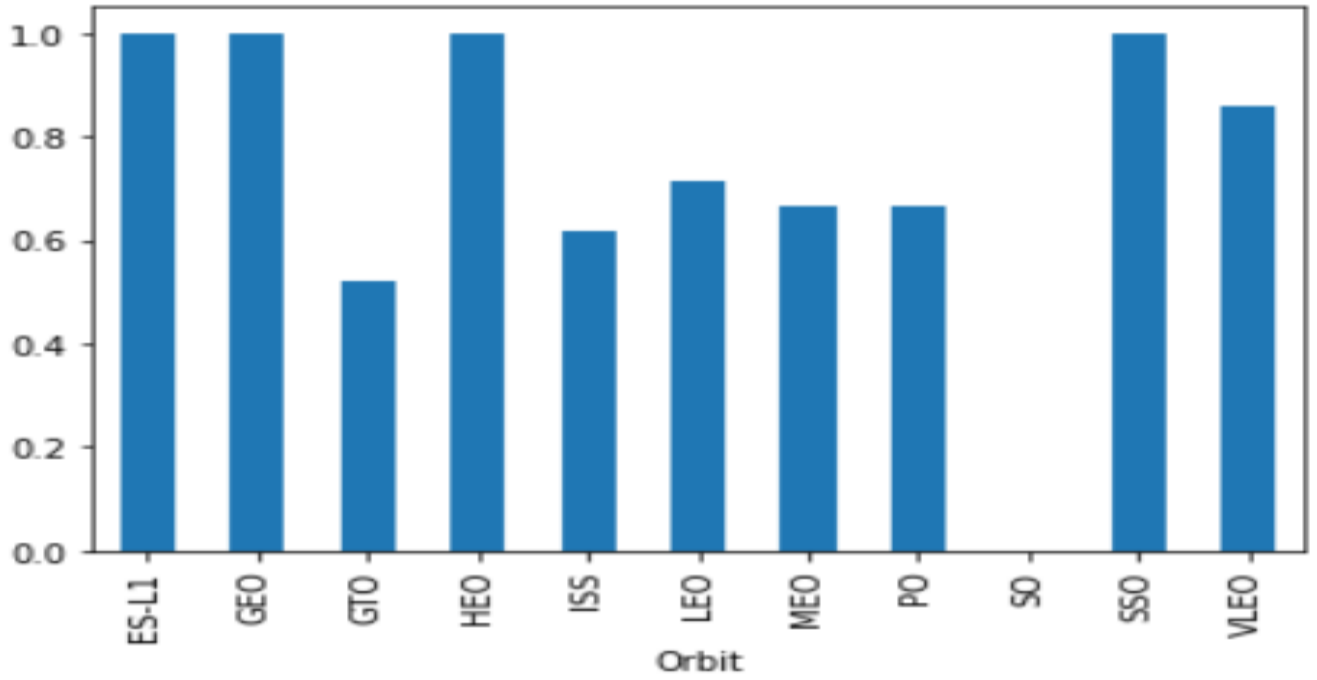
```
Text(9.38737630208334, 0.5, 'LaunchSite')
```



There is no tendency that explains well about the relationship between PayloadMass and succes rate. But between 2000 and 5000 kg PayloadMass, KSC LC39A shows a good performance.

Success rate vs. Orbit type

```
: df1.plot(kind='bar')
: <AxesSubplot:xlabel='Orbit'>
```

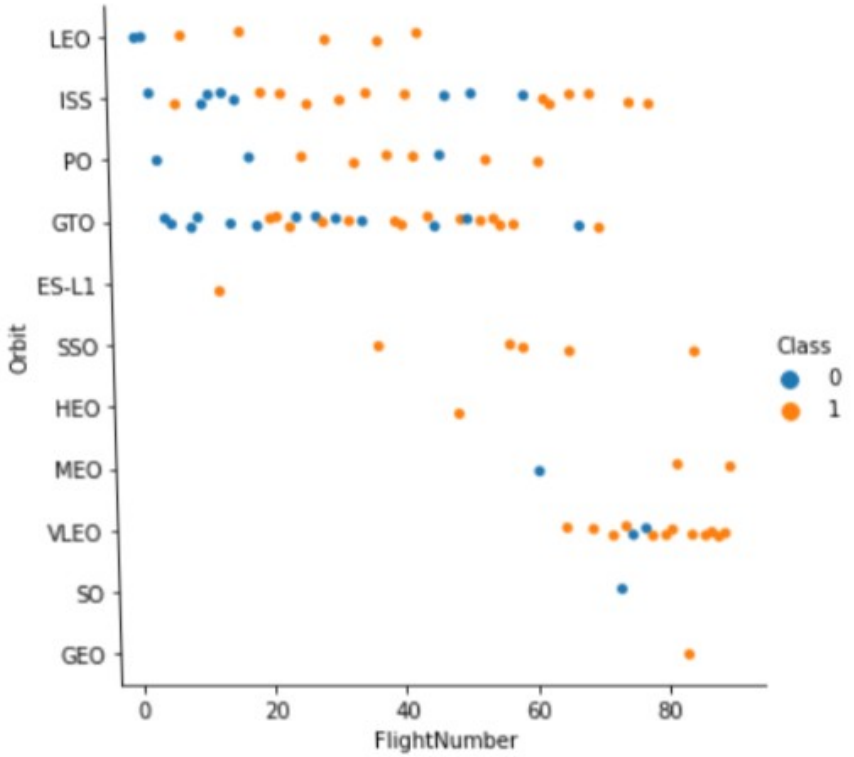


ES-L1, GEO, HEO and SS0 have the better success rates close 1.

Flight Number vs. Orbit type

```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(x='FlightNumber', y='Orbit', data=df, hue='Class')
```

```
<seaborn.axisgrid.FacetGrid at 0x7f48e8c547c0>
```

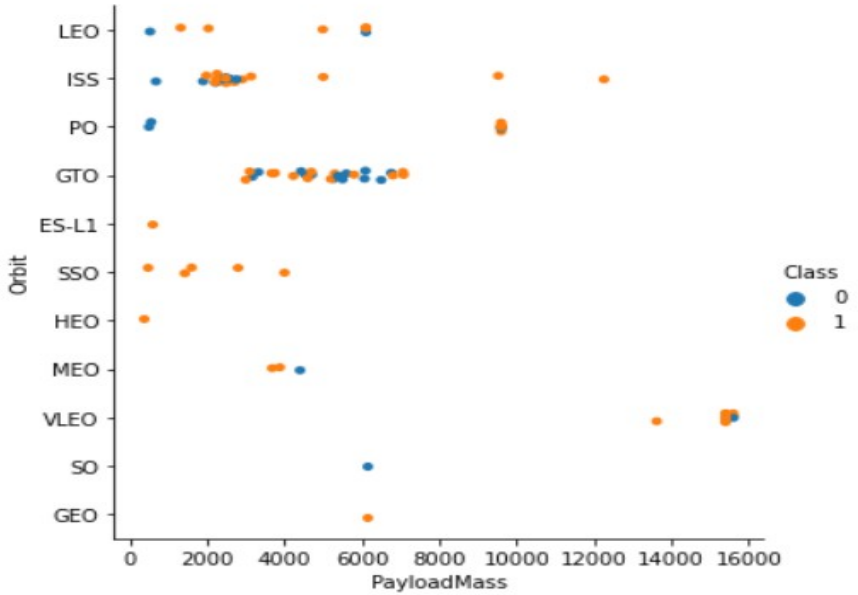


As the flight number increases, almost in all orbits, the success rates also increase which may be explained by the experience in launching.

Payload vs. Orbit type

```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(x='PayloadMass',y='Orbit', hue='Class', data=df)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f48e8664580>
```



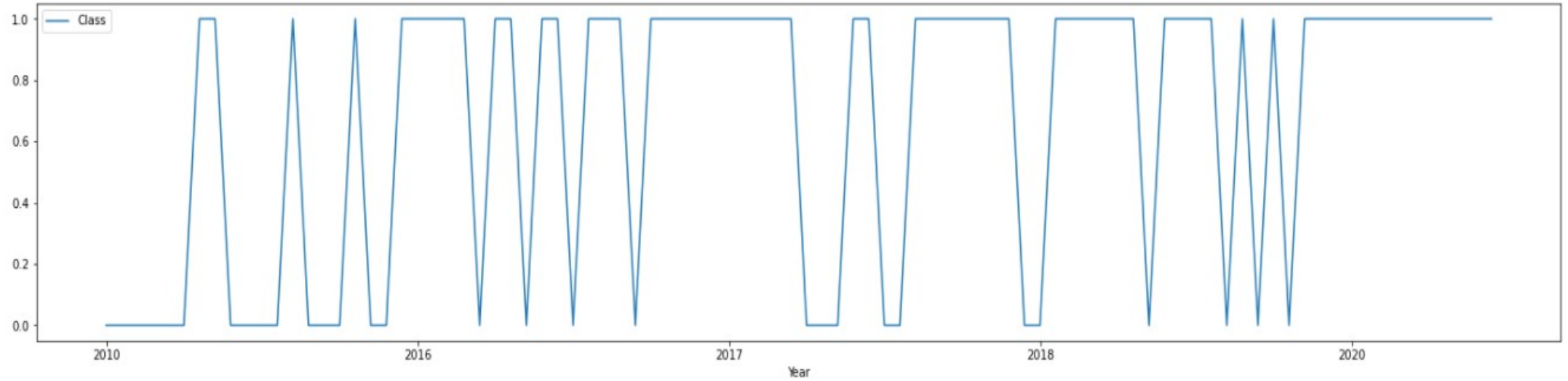
Payload Mass does not seem to explain well whether an orbit succeed. But it is quite possible to say that between 2000 and 6000 PayloadMass, all orbits performing well.

Launch success yearly trend

```
df['Year']=pd.DataFrame(Extract_year())
```

```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
df.plot(kind='line', x='Year', y='Class', figsize=(25,5))
```

```
<AxesSubplot:xlabel='Year'>
```



you can observe that the sucess rate since 2013 kept increasing till 2020

A very fluctuating trend across the years.

EDA with SQL

All launch site names

```
%sql SELECT distinct LAUNCH_SITE FROM SPACEDTX;
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

4 different launch sites are given back

Launch site names begin with `CCA`

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

```
%sql select * FROM SPACEDTX where LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

LIKE is very useful query when the exact name of the value is not remembered or the names of the values are similar.

Total payload mass carried by boosters launched by NASA (CRS)

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

```
%sql select PAYLOAD_MASS__KG_  from SPACEDTX WHERE customer='NASA (CRS)';
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

payload_mass__kg_
500
677
2395
3136
2708
2647
2500
2495
1977
2972

Average payload mass by F9 v1.1

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

```
%sql select AVG(PAYLOAD_MASS__KG_) FROM SPACEDTX WHERE BOOSTER_VERSION='F9 v1.1' ;
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
3676

Average payload mass (kg) of F9 v1.1

First successful ground landing date

```
%sql SELECT min(DATE) from SPACEDTX where LANDING__OUTCOME='Success (ground pad)';
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
2017-01-05

The date of first successful grounding landing

Successful drone ship landing with payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION from SPACEDTX WHERE LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and  6000;
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

booster_version
F9 FT B1022
F9 FT B1031.2

The a.m. booster versions are successfully landed with payload mass between 4000 and 6000.

Total number of successful and failure mission outcomes

```
%sql select COUNT(MISSION_OUTCOME) FROM SPACEDTX WHERE MISSION_OUTCOME='Success' OR MISSION_OUTCOME='Failure';
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
44

All the counts of outcome

Boosters carried maximum payload

```
%sql select BOOSTER_VERSION from SPACEDTX WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) from SPACEDTX);
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

The booster with maximum payload

2015 launch records

```
%sql select monthname(DATE) as DATE1, BOOSTER_VERSION, LAUNCH_SITE from SPACEDTX where LANDING__OUTCOME='Failure (drone ship)' and DATE L
IKE '2015%';
```

```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

date1	booster_version	launch_site
October	F9 v1.1 B1012	CCAFS LC-40

The fail landed drone ship in October 2015 is listed with its booster_version and launch site.

Rank success count between 2010-06-04 and 2017-03-20

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

```
%sql SELECT (LANDING__OUTCOME) FROM SPACEDTX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' and LANDING__OUTCOME LIKE 'Success%' ORDER BY (LANDING__OUTCOME) DESC;
```

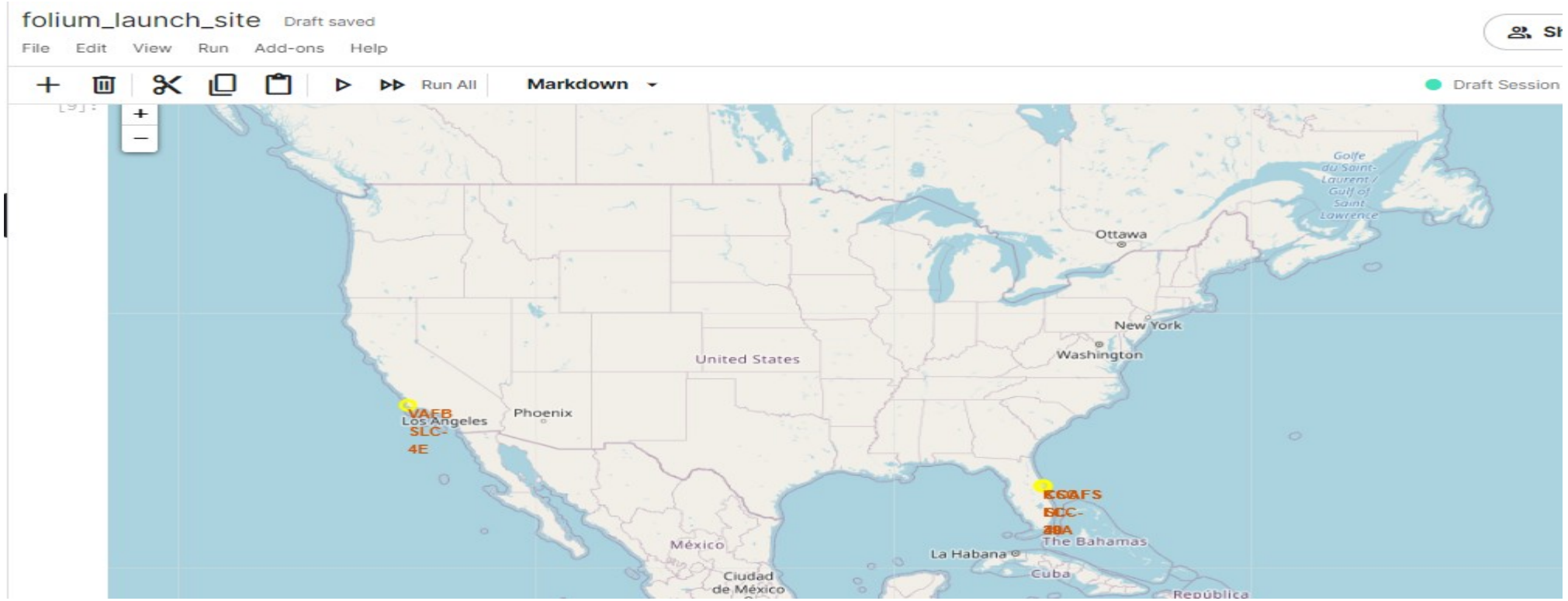
```
 * ibm_db_sa://ddt82037:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

landing__outcome
Success (ground pad)
Success (ground pad)
Success (drone ship)
Success (drone ship)

2 successful ground pad and 2 successful drone ship landings are occurred between 2010-06-04 and 2017-03-20.

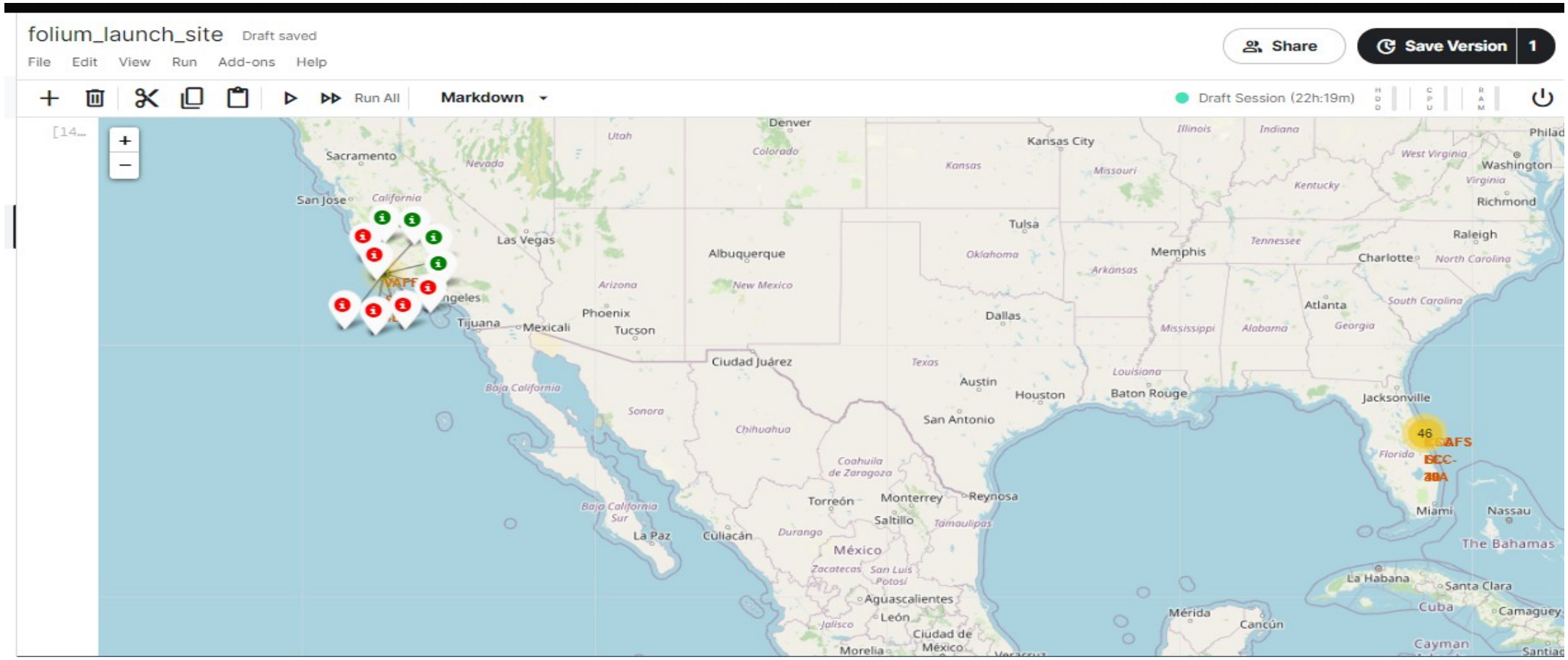
Interactive map with Folium

Launch Sites



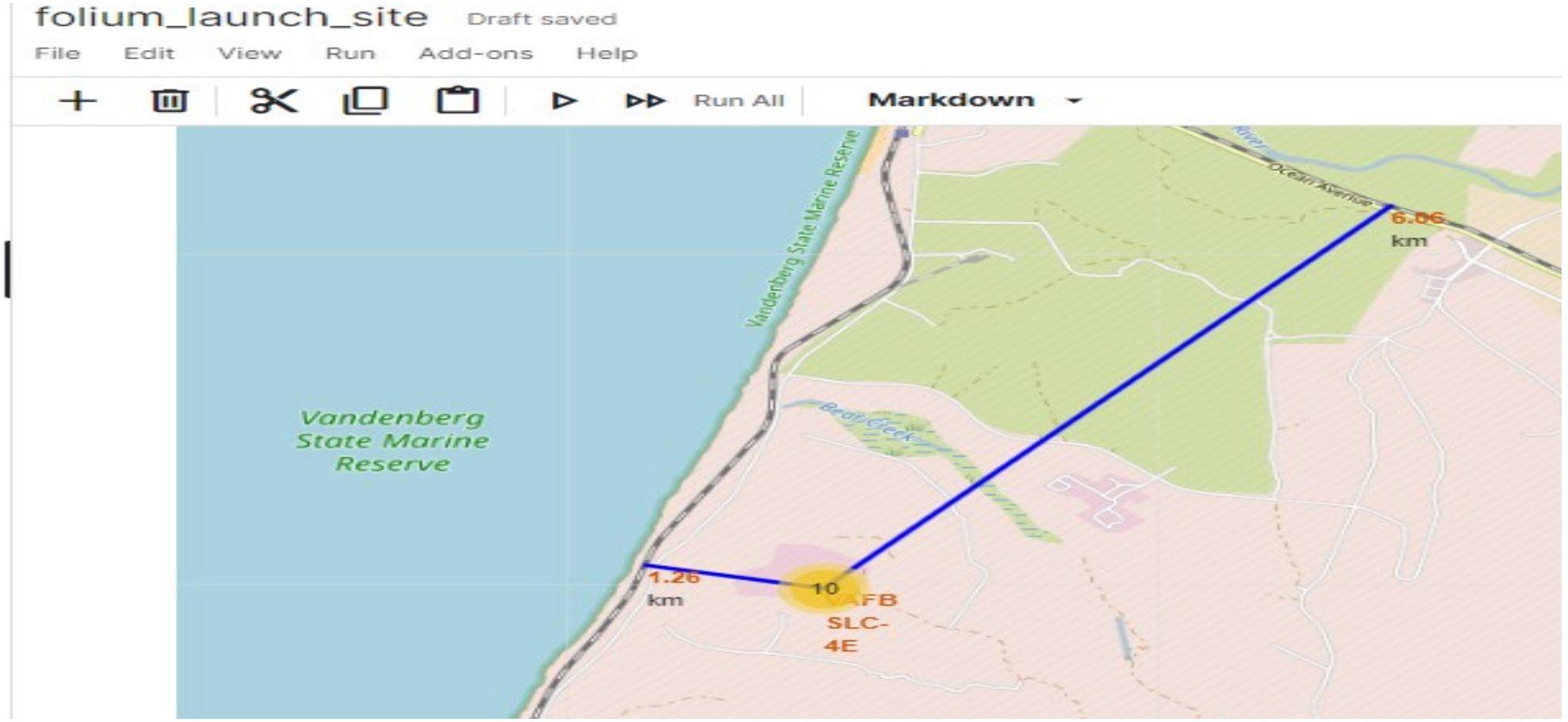
You can see here the launch sites in the global map. They are located mainly in the west and east coast of USA.

Color Labeled Launch



The launch sites here are marked again with respect to their success in landing. If the landing was successfull, it is colored with red otherwise with green.

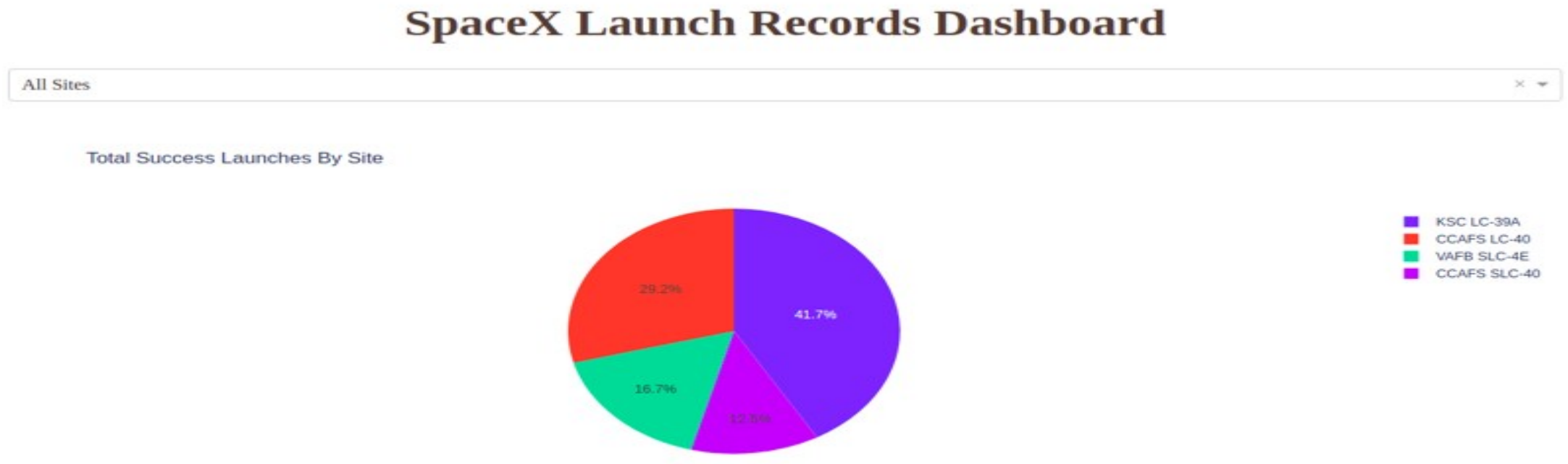
Distance to Proximities



The selected site is 1.26 km away from its closest railway and 6.06 km away from its closest highway

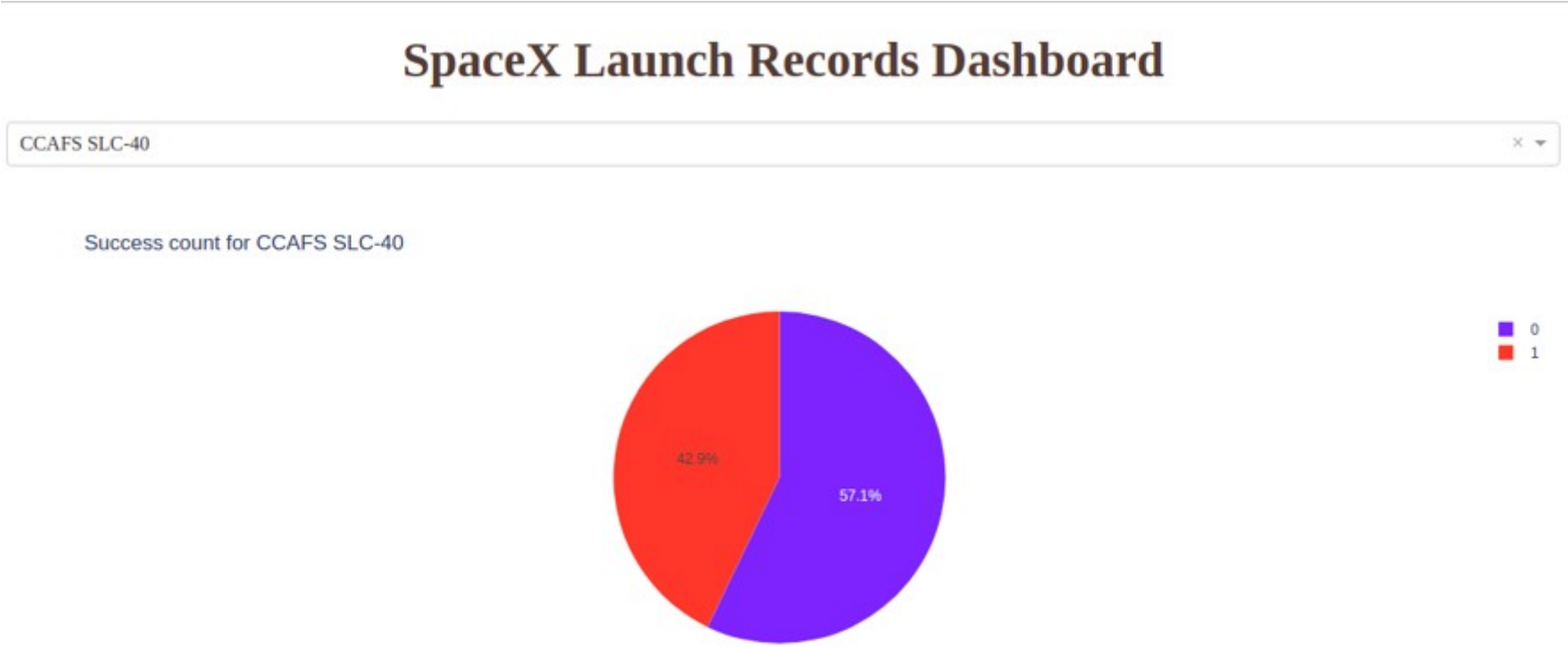
Build a Dashboard with Plotly Dash

Total Success Launches by Site



With regards to the success counts of launch sites, KSC LC-39A is most successful regardless of how many failure counts it has. It has 10 successful counts.

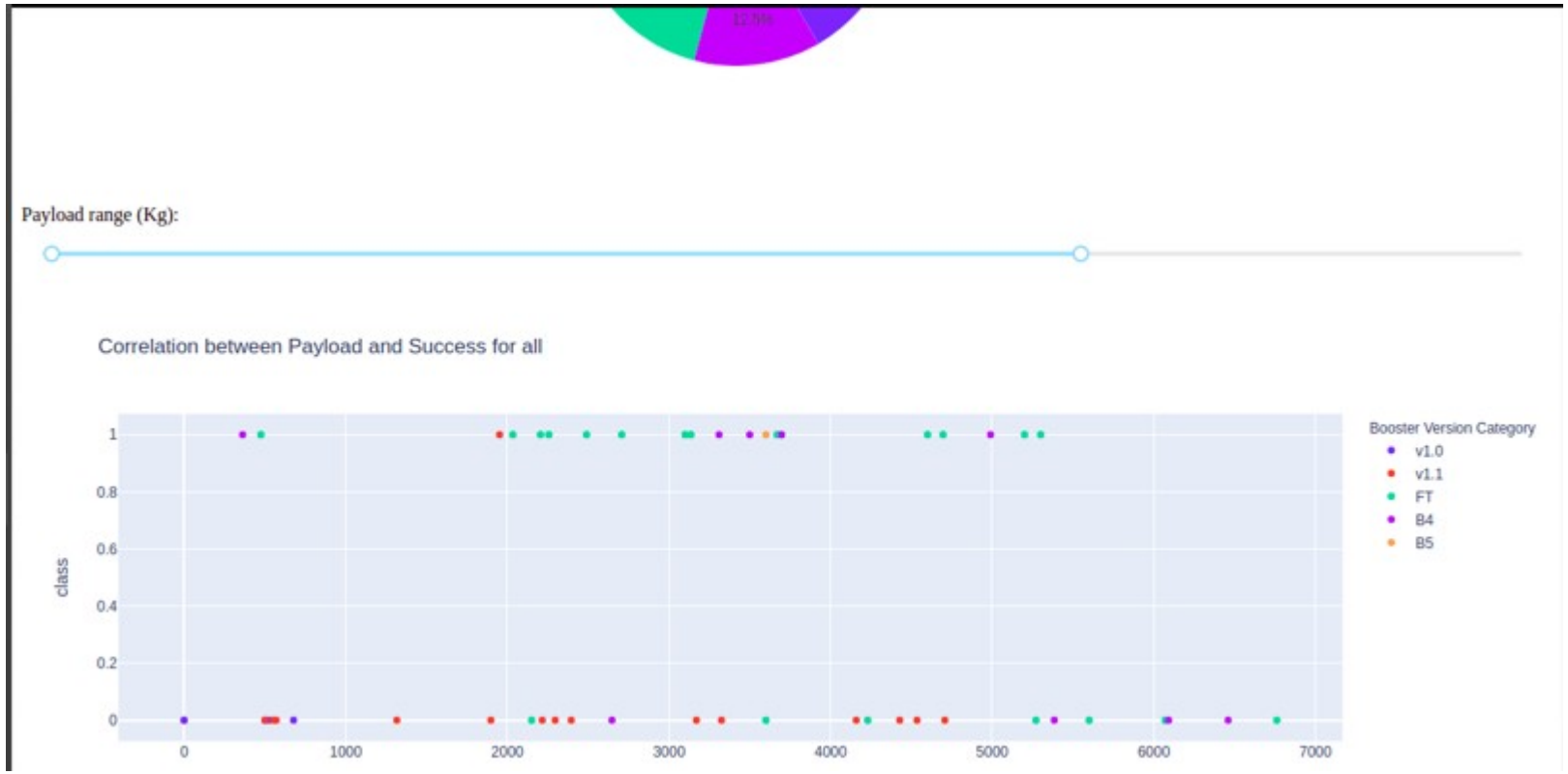
Launch Site– Highest Success Ratio



CCAFS SLC-40 is the most successful launch site with respect to its ratio (42,09 %) of class 1 counts to class 0 counts. Bu it was the worst one in the ranking of the previous pie chart as it has just 3 class 1 values.

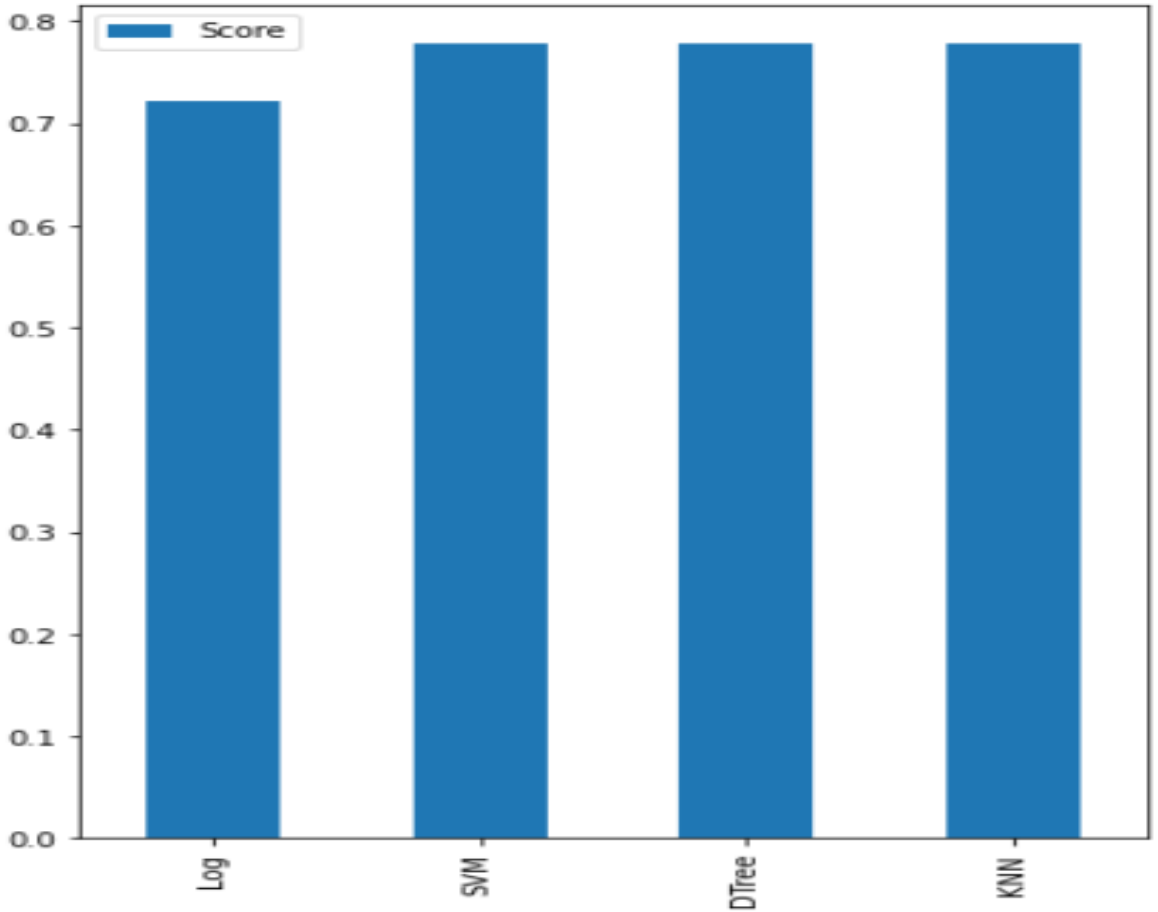
Predictive analysis (Classification)

Payload vs.Launch Outcome - All Sites



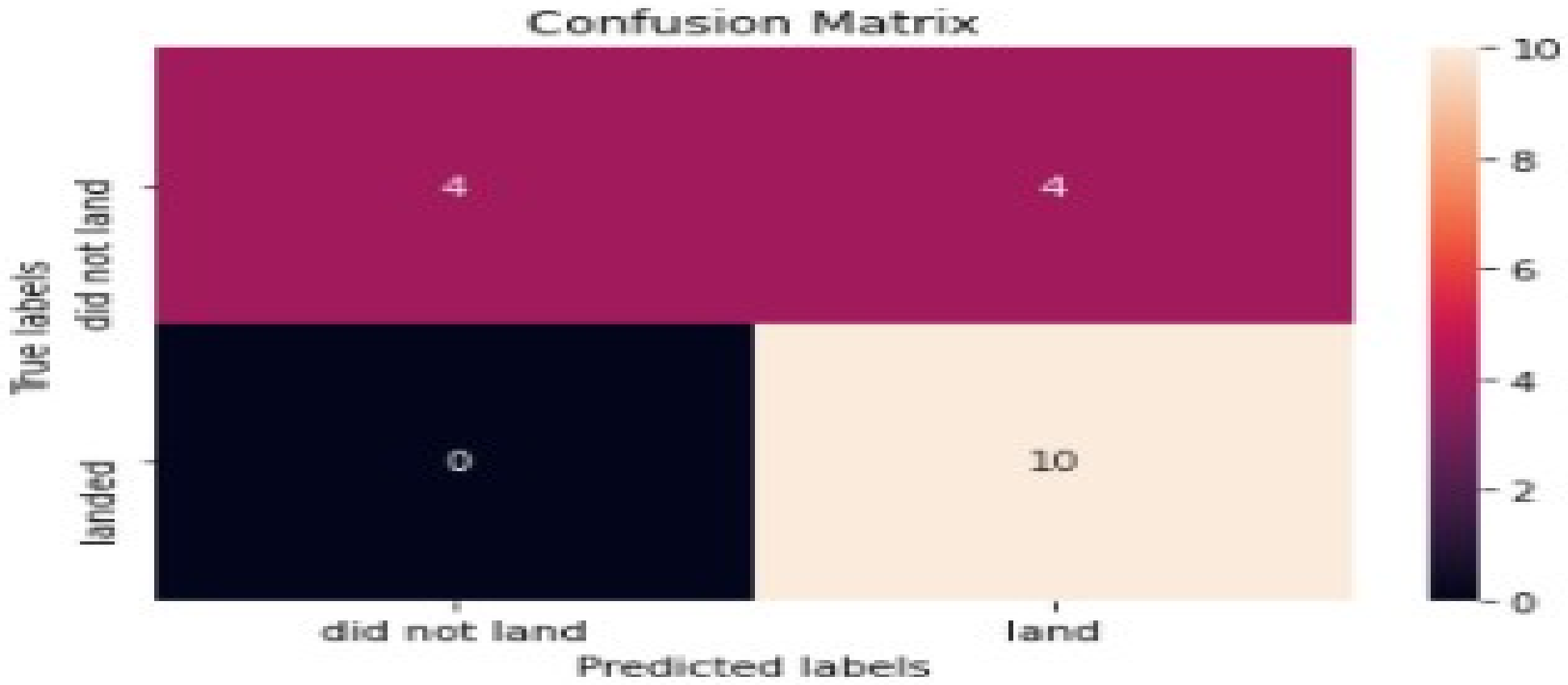
It is not easy to say that there is strong relationship between the payload and Lauch Succes. As Payload Mass (kg) gets more than 6000 kg, the success value is mostly 0 for all launch sites.

Classification Accuracy



KNN, SVM and Decision Tree are the best predictor methods.

Confusion Matrix – KNN, SVM, Logical Regression



CONCLUSION

- KNN, SVM, Decision Tree are methods found equally successful
- KSCLC-39A ist most successful launch site in counts but CCAFS SLC 40 is the one in ratio of success rate
- Payload Mass is not successful to determine the launch outcome