Wining Space Race with Data Science

Sina Khodadadi

26 August 2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Introduction
- Summary of methodologies
- Data Collection & Data Wrangling
- EDA & Interactive Visual Analytics Methodology
- Predictive Analysis Methodology
- EDA with Visualization Results
- EDA With SQL Results
- Interactive Map with Folium Results
- Plotly Dashboard Results
- Predictive Analysis (Classification) Results
- Summary of all results

INTRODUCTION

- **Project background and context**

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- **Problems you want to find answers**

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**:
  - SpaceX Rest API
  - Web Scrapping from SpaceX Wikipedia webpage
- **Perform data wrangling**
  - Hot Encoding data fields
  - Data Cleaning: Correcting and Cleaning null values & irrelevant columns
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - Different algorithm such as: Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Decision Tree model have been deployed and evaluated to find the best method.

Data Collection

There are two methods that data are collected:

1. Used SpaceX Rest API → Returned SpaceX data → Normalized data int a csv file → Data Consolidation & Wrangling

2. Got HTML Response from Wikipedia webpage → Used BeautifulSoup to extract Data → Normalized data int a csv file → Data Consolidation & Wrangling

PROGRAMMING LANGUAGE TRENDS

- Data collection with SpaceX REST calls

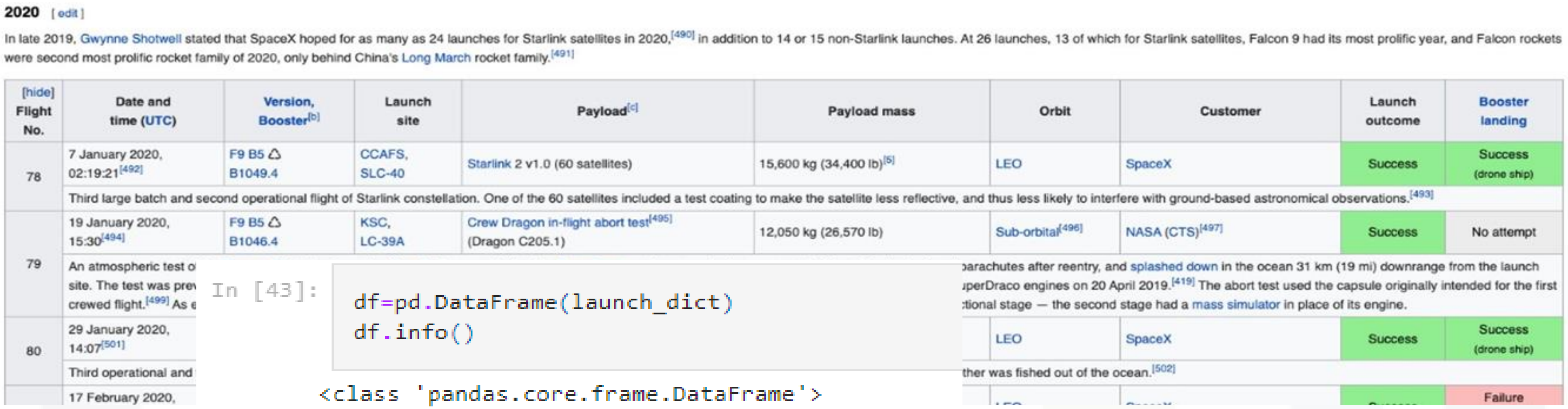
Out[35]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	La
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6b

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module1/jupyter-labs-spacex-data-collection-api%20(2).ipynb

Data collection - Scraping

- Web scraping process



2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 △ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 △ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test o … site. The test was pre… crewed flight.[499] As e… parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch … uperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first … tional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]					LEO	SpaceX	Success	Success (drone ship)
	Third operational and … ther was fished out of the ocean.[502]								
	17 February 2020,								Failure

```
In [43]: df=pd.DataFrame(launch_dict)
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121 entries, 0 to 120
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   Flight No.      121 non-null    object
 1   Launch site     121 non-null    object
 2   Payload         121 non-null    object
 3   Payload mass    121 non-null    object
 4   Orbit           121 non-null    object
 5   Customer        120 non-null    object
 6   Launch outcome  121 non-null    object
 7   Version Booster 121 non-null    object
 8   Booster landing 121 non-null    object
 9   Date            121 non-null    object
 10  Time            121 non-null    object
dtypes: object(11)
memory usage: 10.5+ KB
```

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module1/jupyter-labs-webscraping%20(2).ipynb

DATA Wrangling

Determine Null Values → Determine the number of launches on each site → Determine the number and occurrence of each orbit

→ Determine the number of Landing_outcomes → Create a landing outcome label from Outcome column → Handle Null values

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module1/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

•For machine learning models we should get insight into the relationship between parameters. Therefore, various type of plots such as: Scatter plots, line charts, and bar plots are used to show the relation between different parameters.

- 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

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module2/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- •Displayed the names of the unique launch sites in the space mission
- •Displayed 5 records where launch sites begin with the string 'CCA'
- •Displayed the total payload mass carried by boosters launched by NASA (CRS)
- •Displayed average payload mass carried by booster version F9 v1.1
- •Listed the date when the first successful landing outcome in ground pad was achieved.
- •Listed the names of the boosters which have success in drone ship and have payload mass
- •greater than 4000 but less than 6000
- •Listed the total number of successful and failure mission outcomes
- •Listed the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- •Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- •Ranked the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module2/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

With aim to finding an optimal location for building a launch site:

- Calculated the distances between a launch site to its proximities
- Marked down a point on the closest coastline using Mouse Position and calculated the distance between the coastline point and the launch site
- Drew a PolyLine between a launch site to the selected coastline point
- Created a marker with distance to a closest city, railway, highway, etc.
- Drew a line between the marker to the launch site

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module3/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

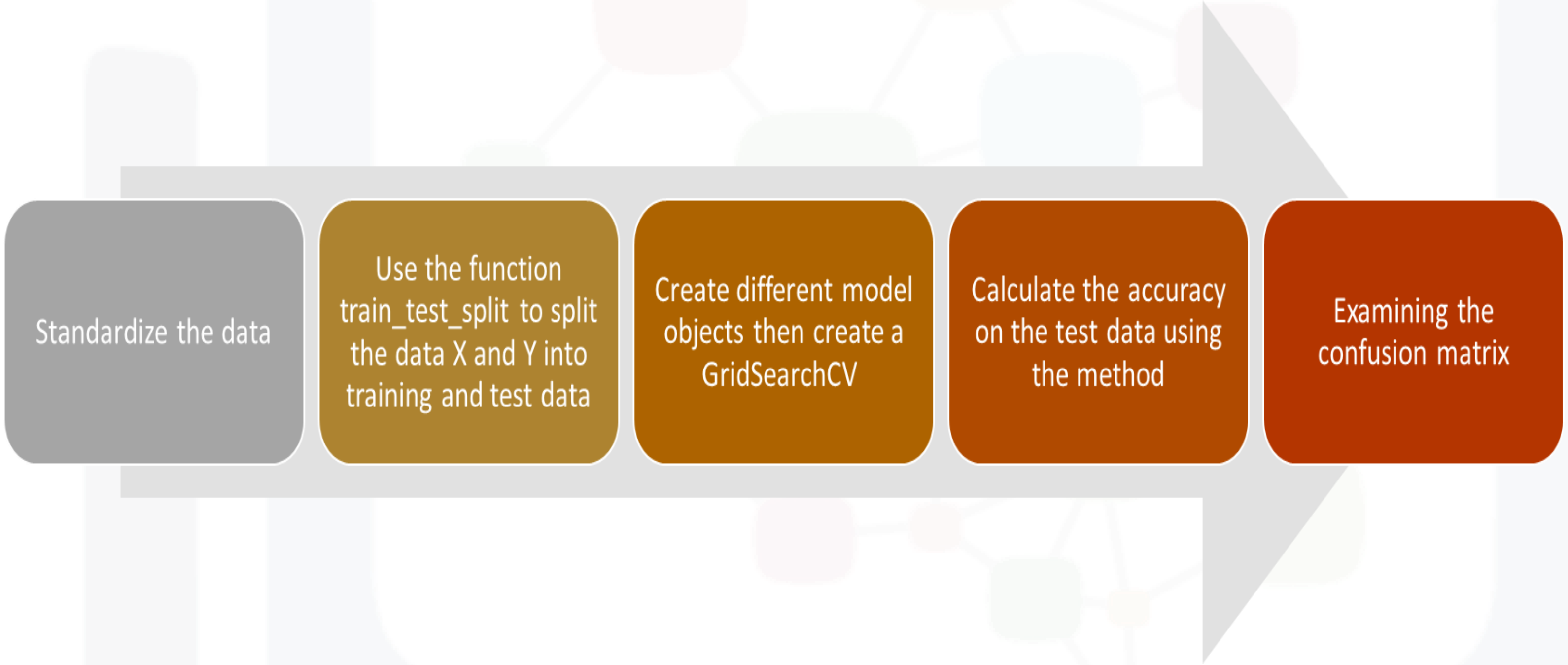
Dashboard with a selectable pie chart and a scatter plot.

- Pie chart shows distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- The pie chart is used to visualize launch site success rate.

- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module3/spacex_dash_app.py

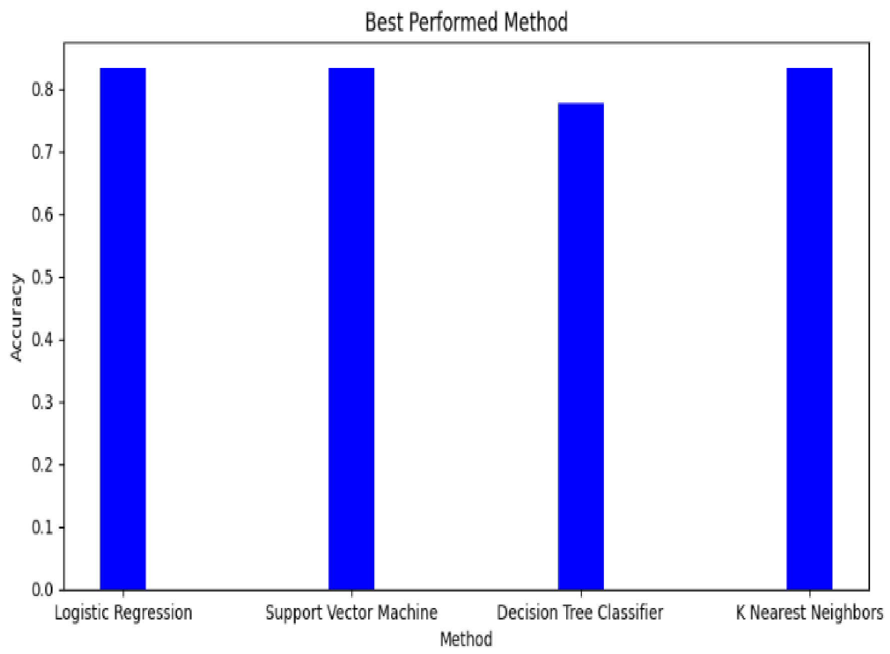
Predictive Analysis (Classification)



https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

Results

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.



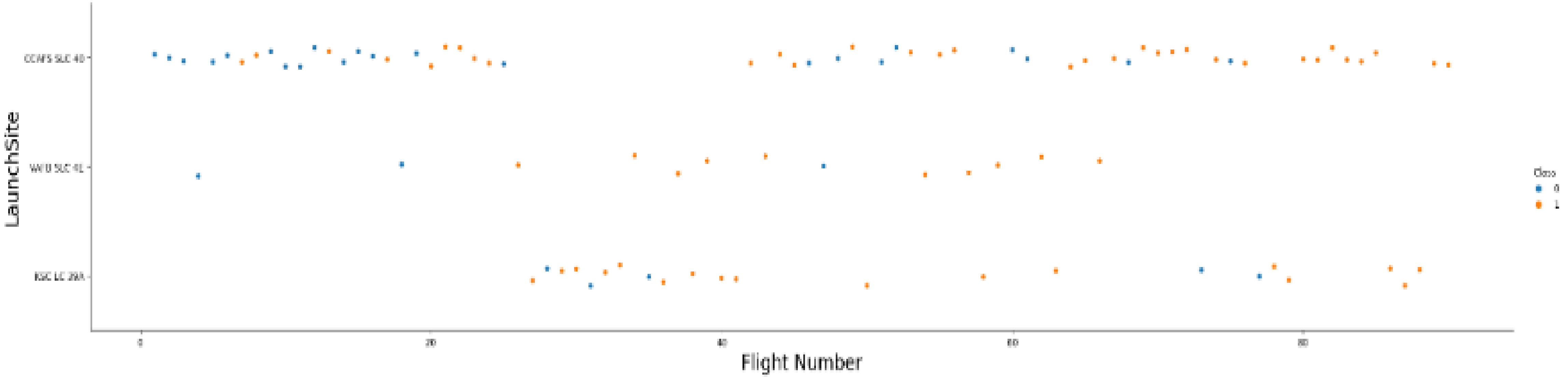
https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

Section 2

Insights Drawn from EDA

Flight Number vs. Launch Site

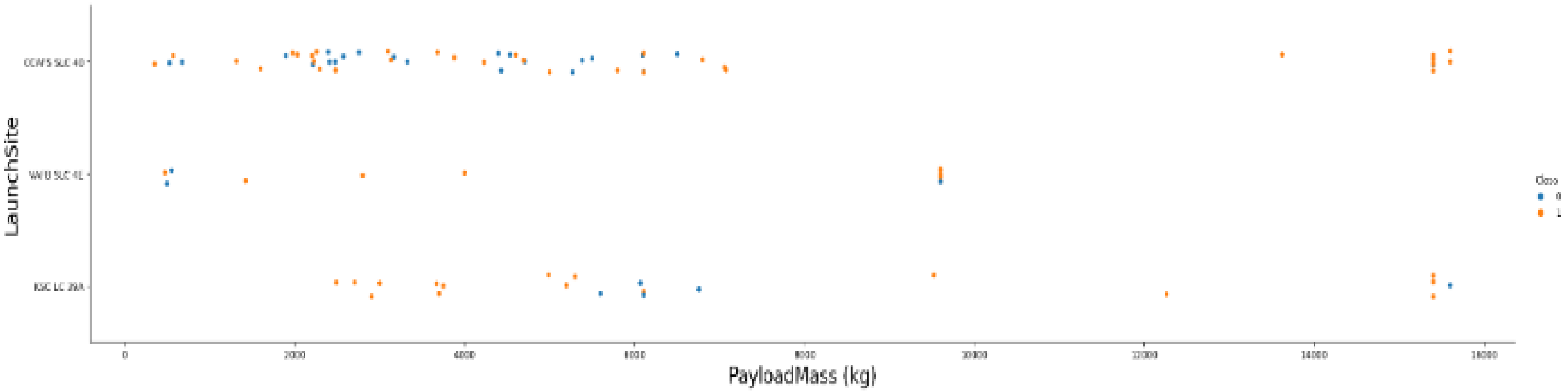
- Launches from the site of CCAFS SLC 40 are significantly higher than lunches from other sites



https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Payload vs. Launch Site

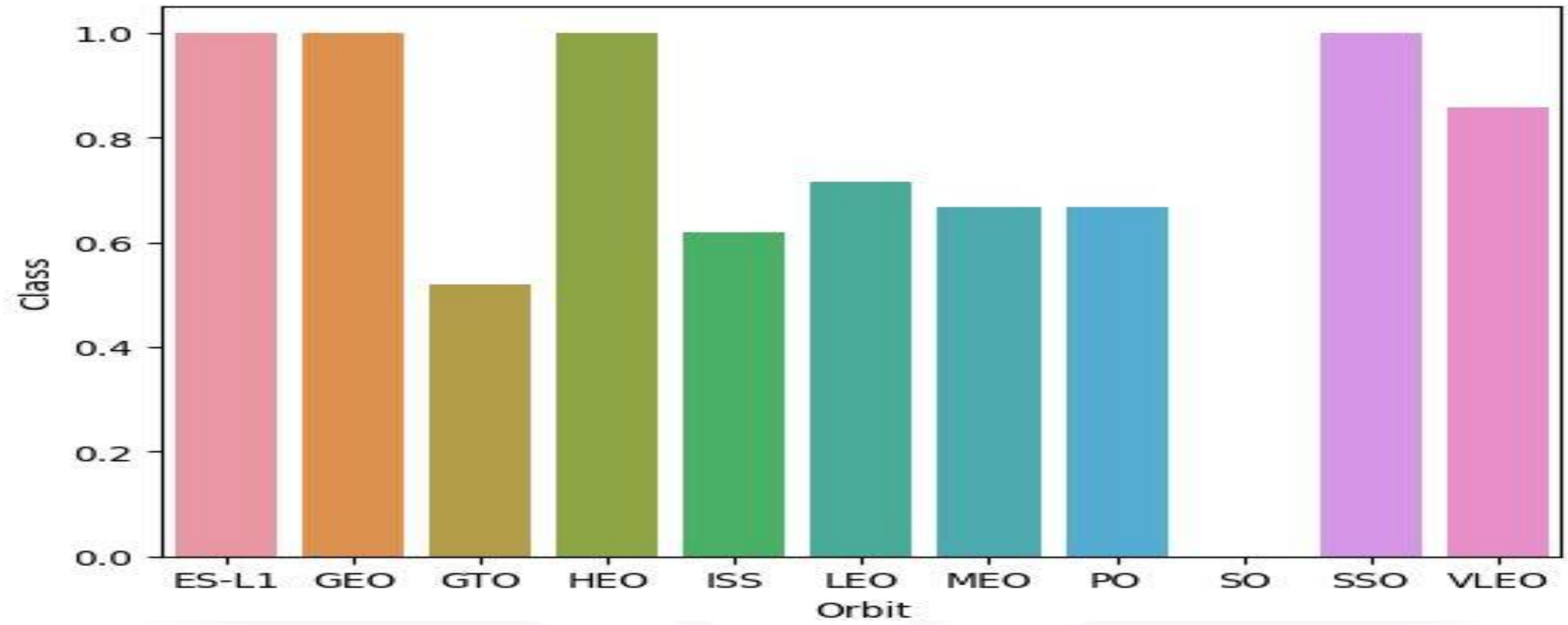
- The majority of Pay Loads with lower Mass have been launched from CCAFS SLC 40



https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Success Rate vs. Orbit Type

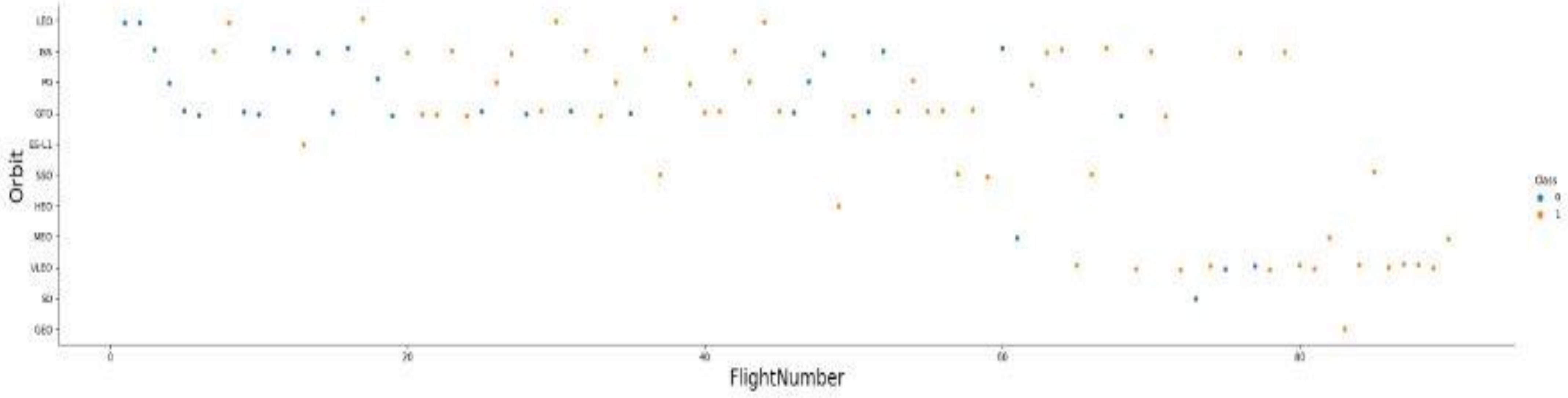
- ESL1, GEO, HEO and SSO orbits had the highest success rate.



https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Flight Number vs. Orbit Type

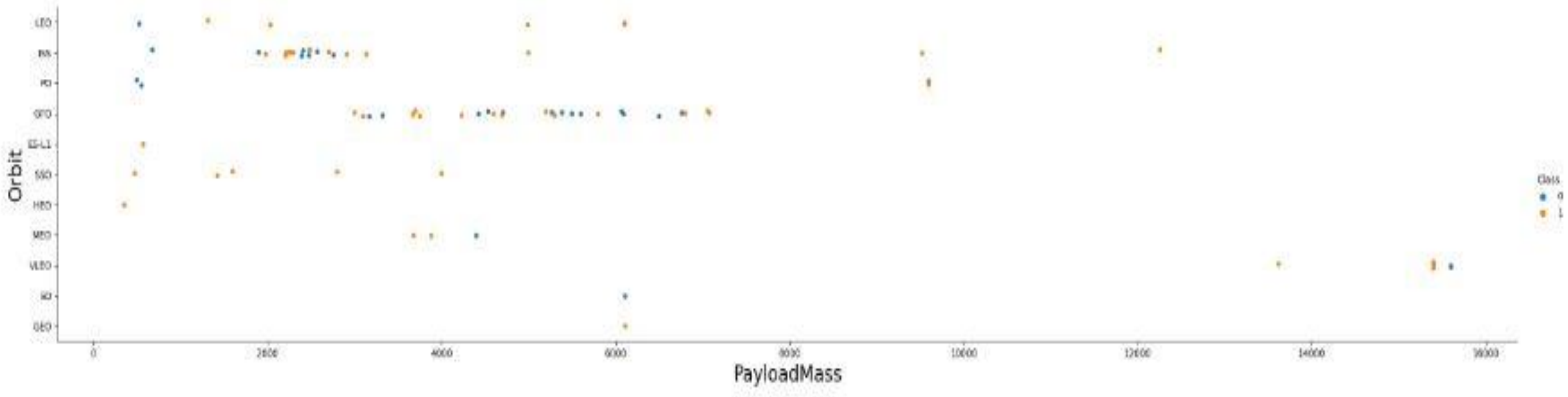
- There is an increasing rate of VLEO launches in recent years.



https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Payload vs. Orbit Type

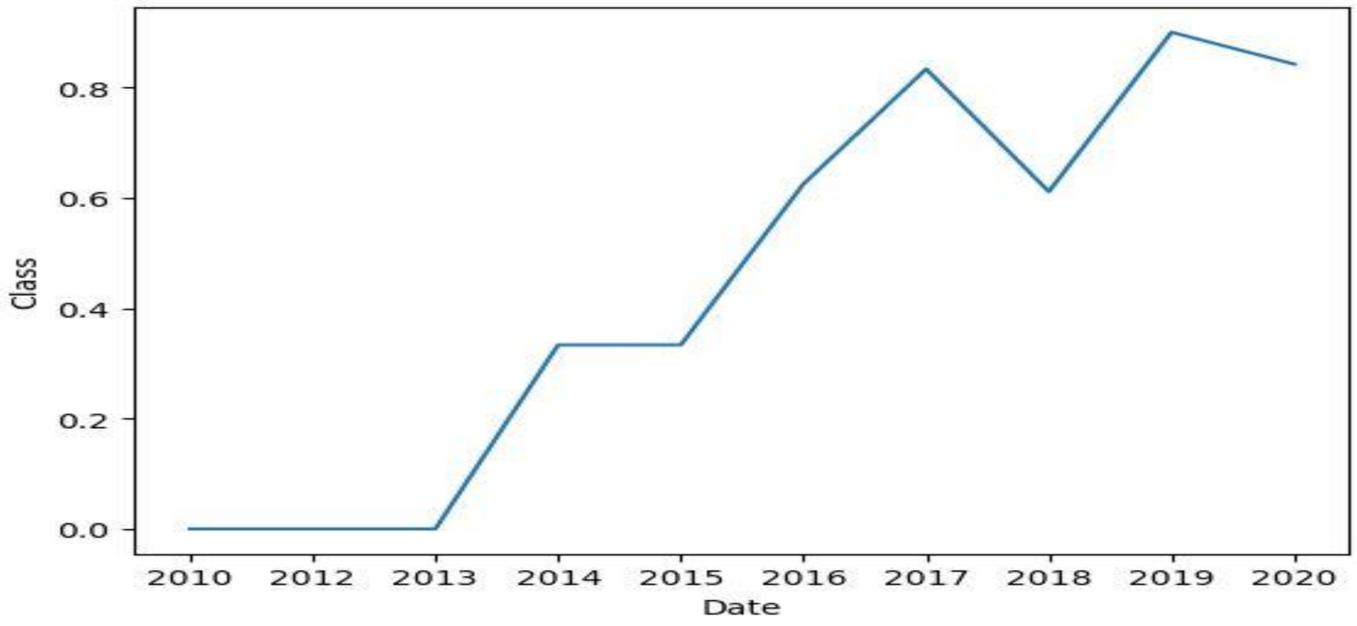
At ISS most of payloads were around 2000
At GTO most of payload were between 3000 and 7000



https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Launch Success Yearly Trend

- Success rate has increased incredibly in recent years



https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

All Launch Site Names

```
In [6]: %sql select distinct launch_site from SPACEXTBL

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

Out[6]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Launch Site Names Begin with 'CCA'

In [7]:
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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

Out[7]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (p
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (p
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	N
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	N
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	N

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Payload Mass

In [8]:

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)';
```

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

Out[8]:

sum(PAYLOAD_MASS__KG_)
45596.0

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Average Payload Mass by F9 v1.1

In [9]:

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1';
```

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

Out[9]:

avg(PAYLOAD_MASS__KG_)
2928.4

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

First Successful Ground Landing Date

```
In [10]: %sql select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';

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

Out[10]:

min(Date)
01/08/2018

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

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

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [11]: %sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000 and Landing_Outcome = 'Success

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

Out[11]:

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

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Total Number of Successful and Failure Mission Outcomes

In [25]:
```
%sql select Mission_Outcome, COUNT(*) as Count from SPACEXTBL group by Mission_Outcome
```

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

Out[25]:

Mission_Outcome	Count
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Boosters Carried Maximum Payload

```
In [21]: %sql select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACE

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

Out[21]:

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

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

2015 Launch Records

%sql select substr(Date, 4, 2) as month, "Landing_Outcome", Booster_Version , Launch_Site from SPACEXTBL where "Landing_Outcome" = 'Failure (drone ship)' and substr(Date,7,4)='2015'

```
In [24]: %sql select substr(Date, 4, 2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where Landing_Out

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

Out[24]:

month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

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

- %sql SELECT "Landing_Outcome", count(*) AS count, RANK() OVER (ORDER BY count(*) DESC) AS rank FROM SPACEXTBL WHERE "Landing_Outcome" like 'Success%' and Date between '04-06-2010' and '20-03-2017' GROUP BY "Landing_Outcome" ORDER BY count DESC

```
In [35]: %sql SELECT "Landing_Outcome", count(*) AS count, RANK() OVER (ORDER BY count(*) DESC) AS rank FROM SPACEXTBL WHERE  Da

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

Out[35]:

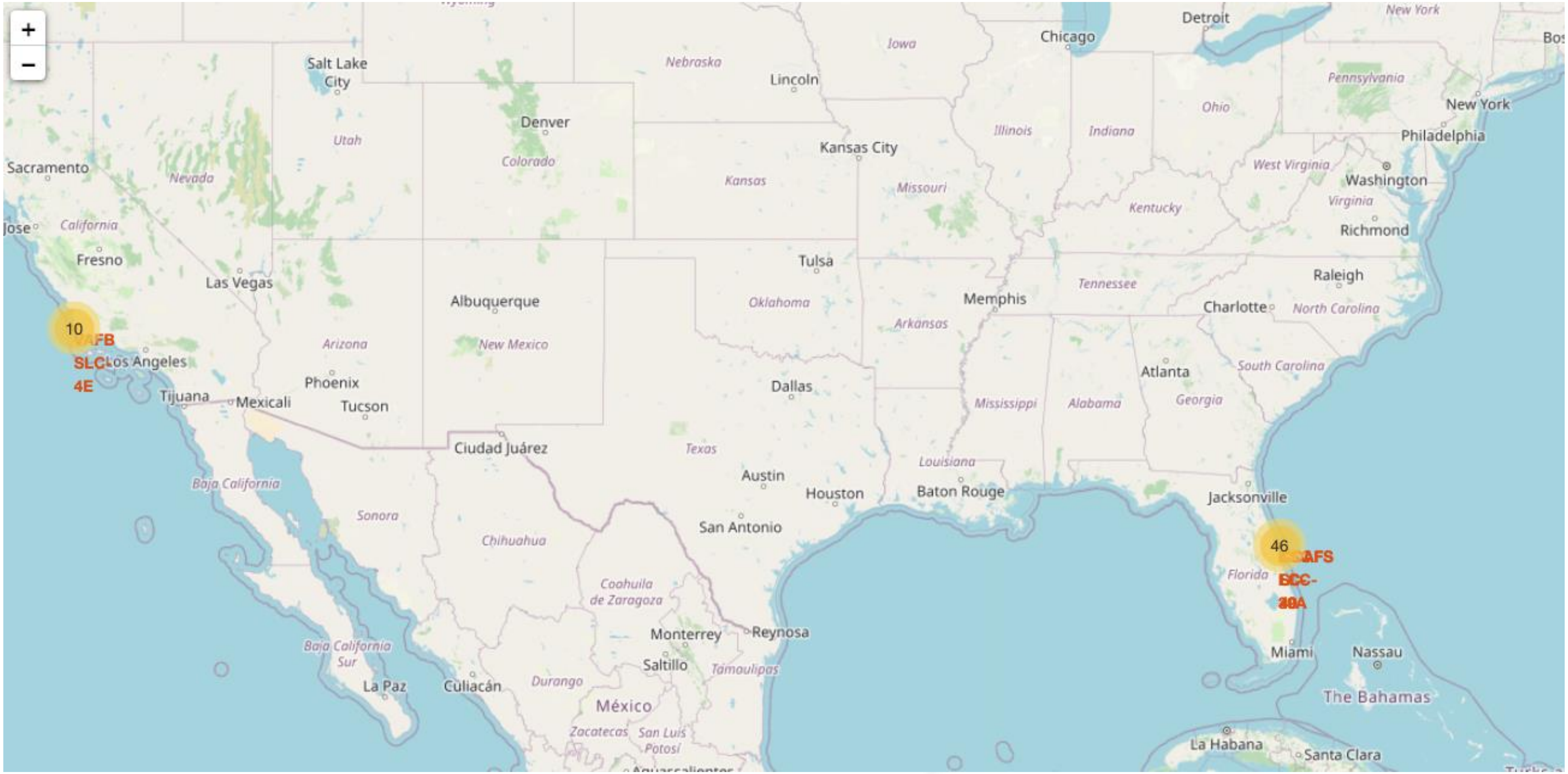
Landing_Outcome	count	rank
Success	20	1
No attempt	10	2
Success (drone ship)	8	3
Success (ground pad)	7	4
Failure (drone ship)	3	5
Failure	3	5
Failure (parachute)	2	7
Controlled (ocean)	2	7
No attempt	1	9

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module4/jupyter-labs-eda-sql-coursera_sqllite.ipynb

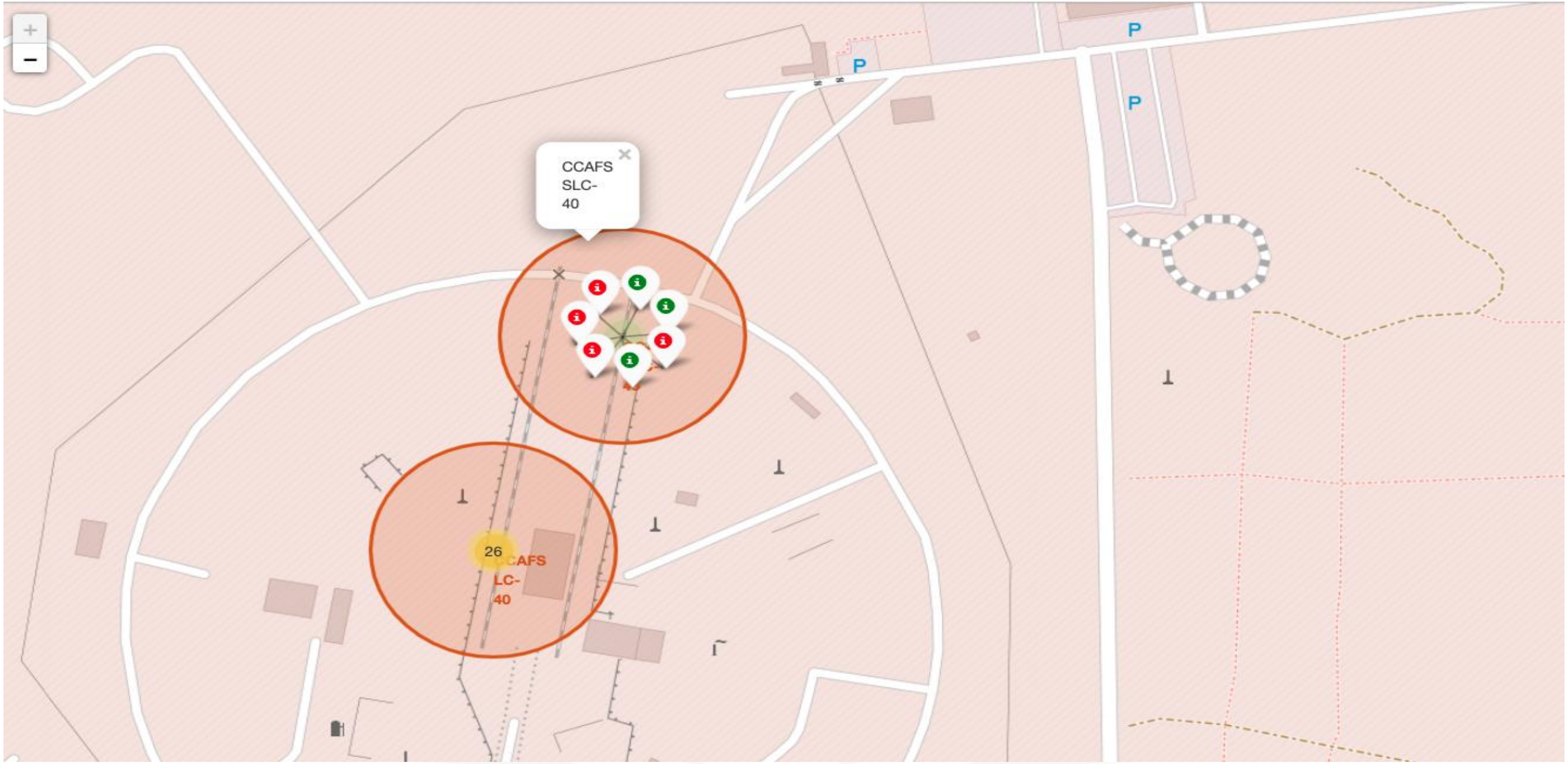
Section 3

Launch Sites Proximities Analysis

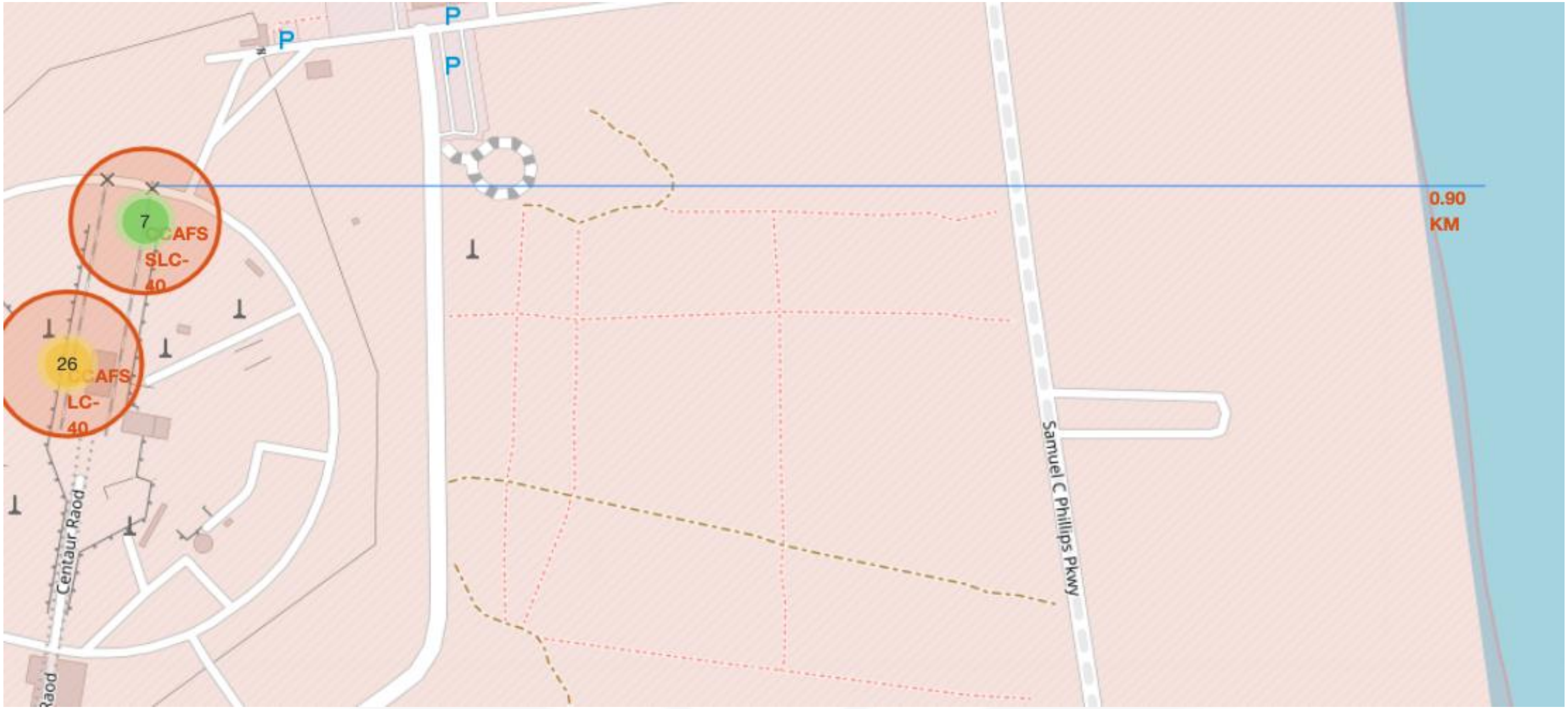
Folium Map Screenshot 1



Folium Map Screenshot 2



Folium Map Screenshot 3



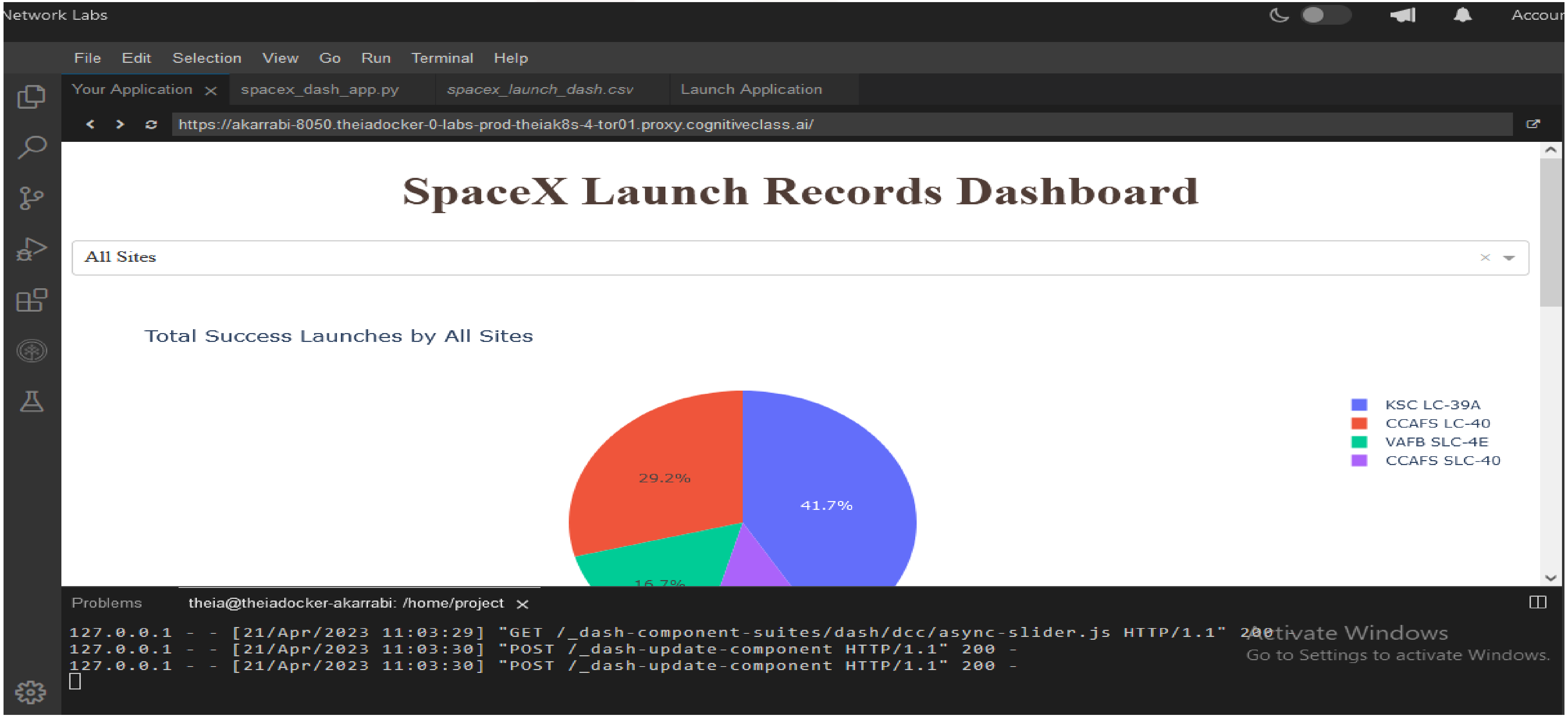
Section 4

Build a Dashboard with Plotly Dash

DASHBOARD

https://github.com/sinasorme/IBM-Data-Science-Capstone/blob/Module3/spacex_dash_app.py

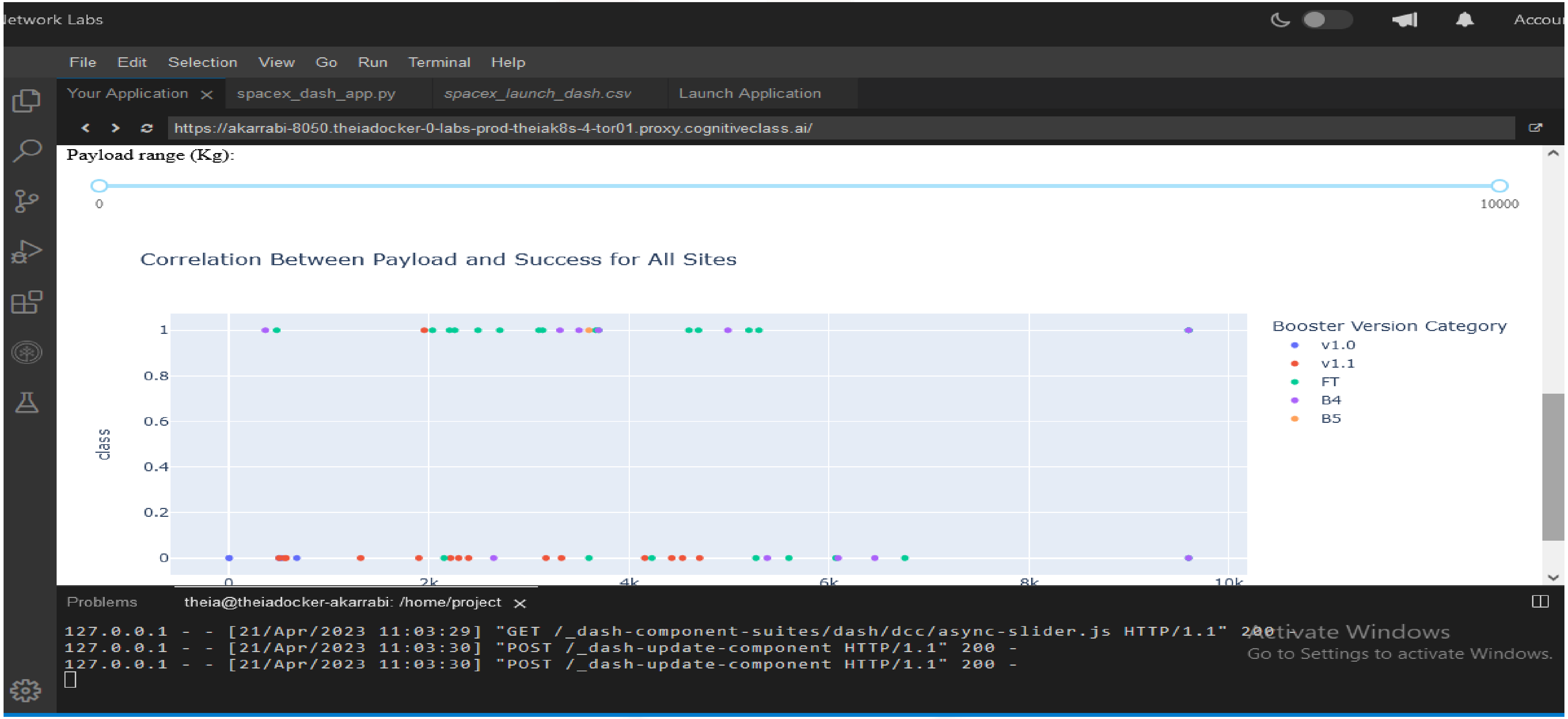
DASHBOARD TAB 1



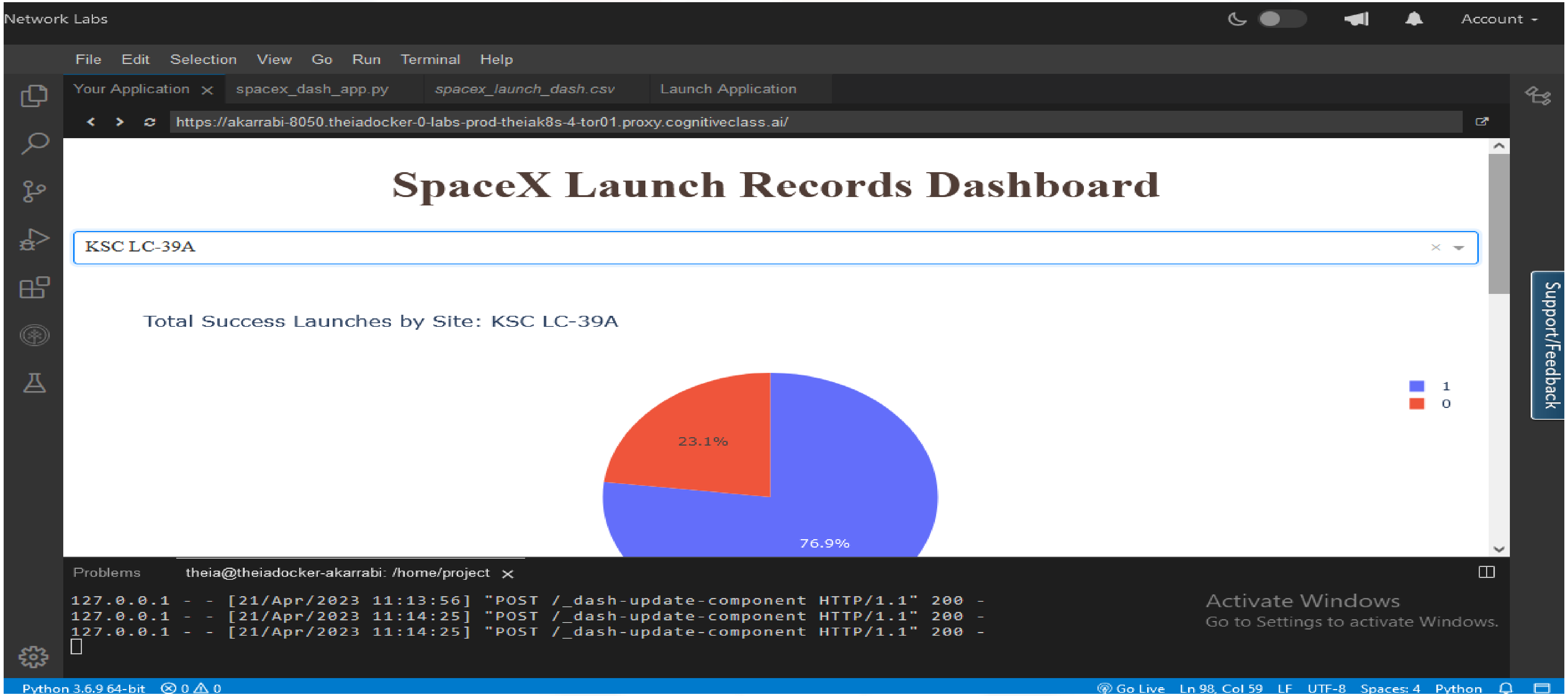
IBM Developer

SKILLS NETWORK

DASHBOARD TAB 2



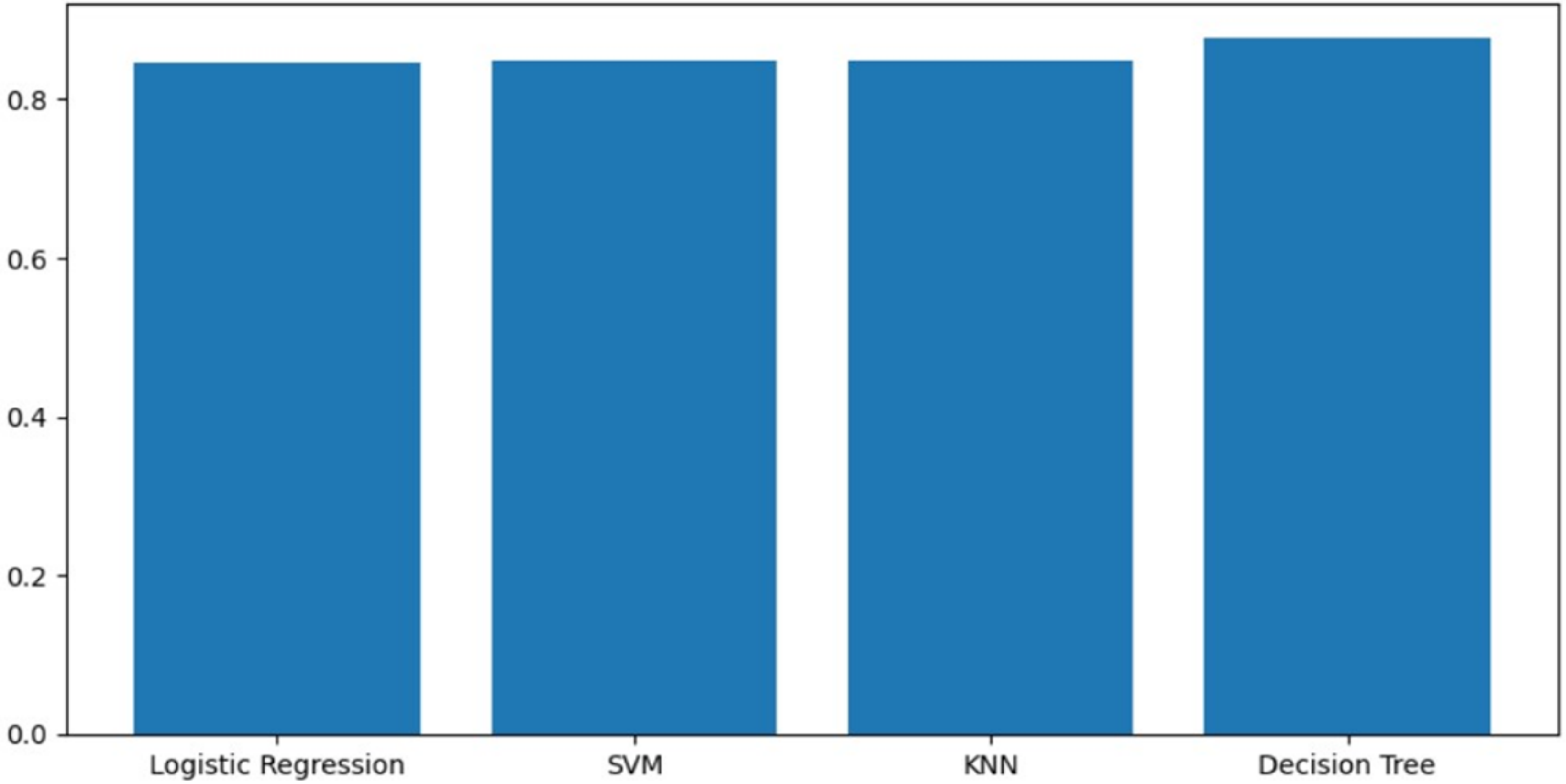
DASHBOARD TAB 3



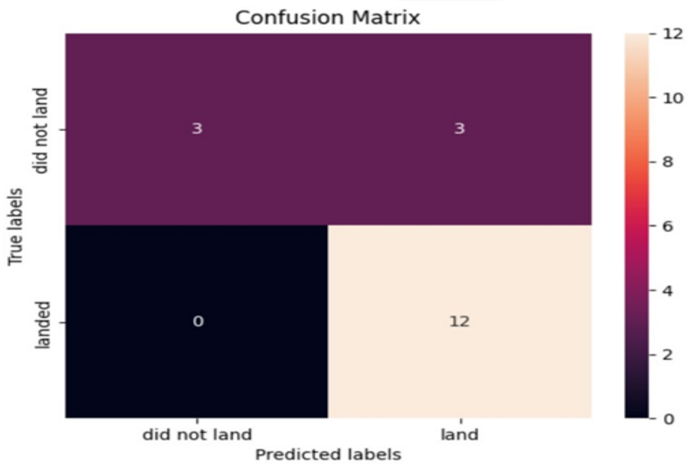
Section 5

Predictive Analysis (Classification)

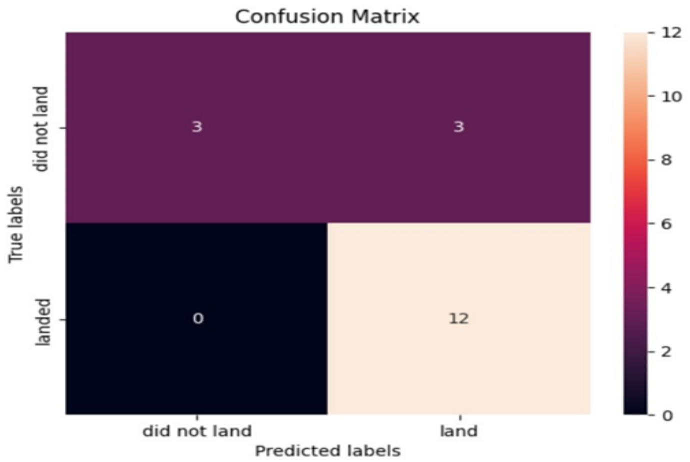
Classification Accuracy



Confusion Matrix

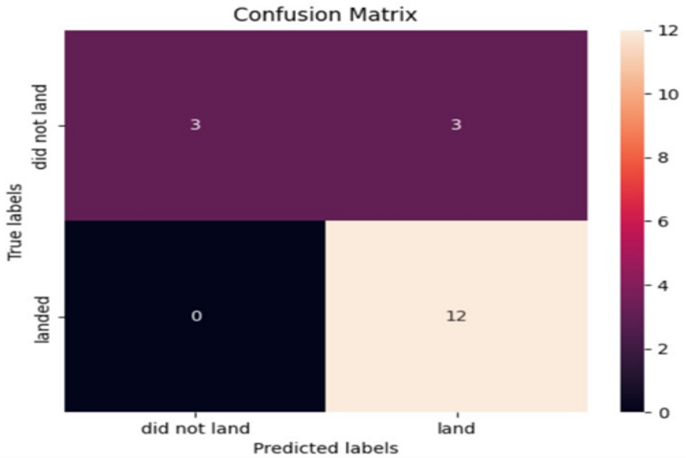


Logistic Regression

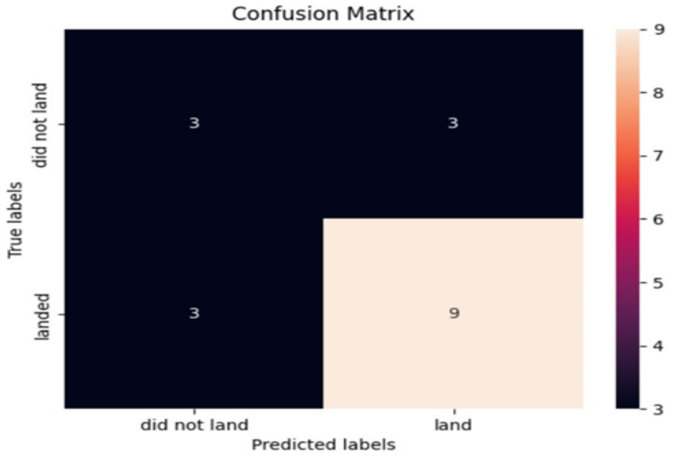


KNN

SVM



Decision Tree



Conclusions

- The Decision Tree model has the best prediction accuracy.
- Low weighted payloads has a better success rate than the heavier payloads.
- The success rates for SpaceX launches have increased significantly in recent years.
- KSC LC 39A had the most successful launches among all other sites.
- Orbit GEO,HEO,SSO,ES L1 had the best Success Rate.

APPENDIX

Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project