IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name : Yassine El Marhraoui>
<Date : Fri 28/09/2021>

Outline

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

Executive Summary

- First, we collect data using two methods:
  - ➢ Calling SpaceX REST API call
  - ➢ Webscraping on Wikipedia.
- Then, we prepare our data for analysis using techniques like data preprocessing feature engineering.
- After that, we rely on SQL and visualization techniques to manage the Exploratory Data Analysis
- Furthermore, we explore launch sites data on maps using the famous package Folium and we build powerful interactive charts by Dash.
- Finally, we build a classifier that predicts the outcome of the landing with an accuracy higher than 83%

Introduction

Space X 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 Space X can reuse the first stage.

Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX API and Webscraping.
- Perform data wrangling
  - Dealing with missing values and creating a binary column of successful landings and unsuccessful landings
- 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

Data Collection

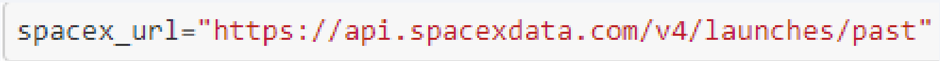
- We chose two methods:
  - Using an API call, namely the Open Source SpaceX REST API. We will use the link below which contains data about SpaceX past launches :

    https://api.spacexdata.com/v4/launches/past
  - The second method is Webscraping, where we will gather our data from the famous encyclopedia Wikipedia using Python's package BeautifulSoup :

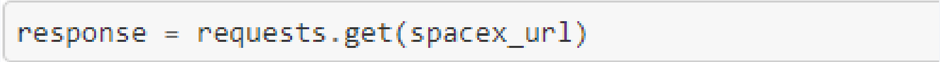
    https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- In both cases, we focus only on data related to Falcon 9 launches

Data Collection – SpaceX API

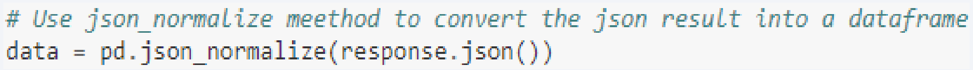
- After running the following lines of code, our data will look like this -->



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

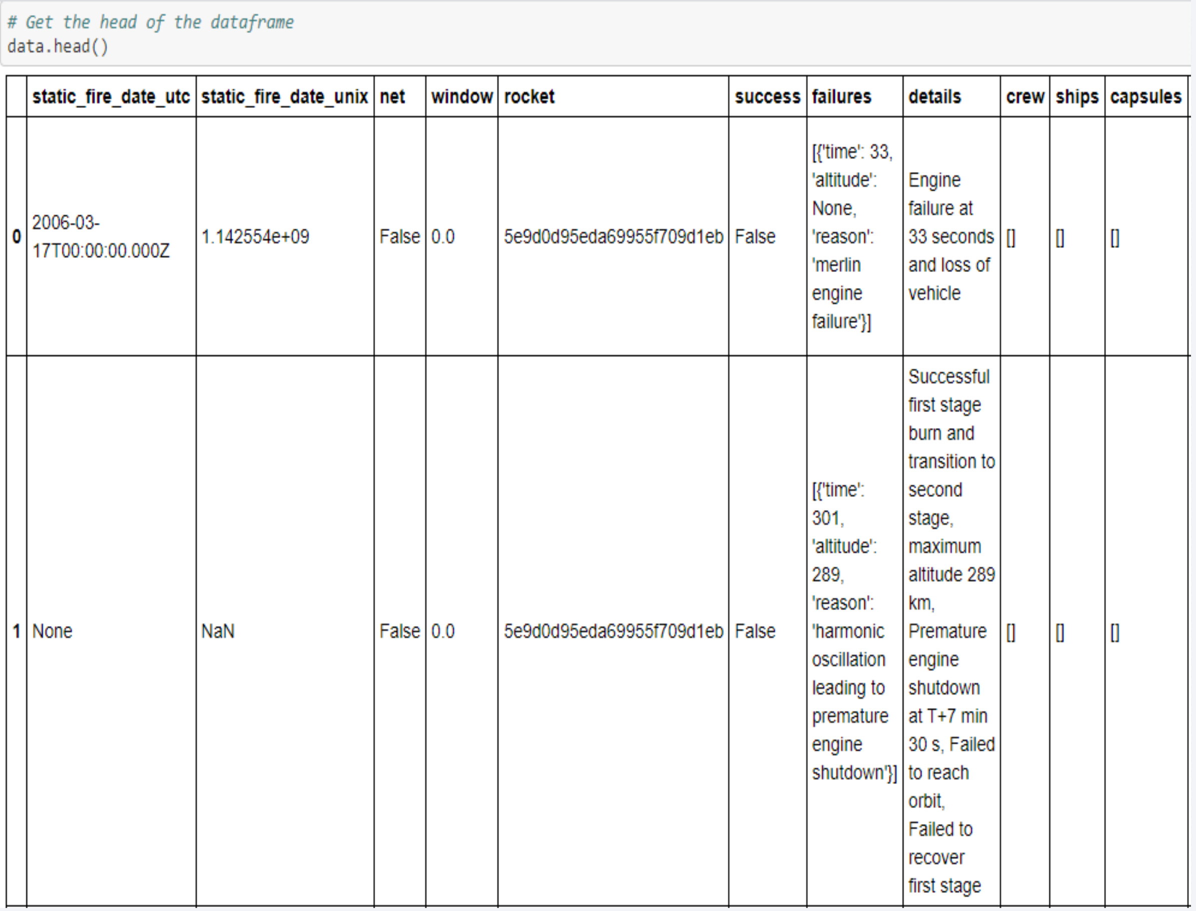


```
response = requests.get(spacex_url)
```



```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

[GitHub - Data Collection API Notebook]



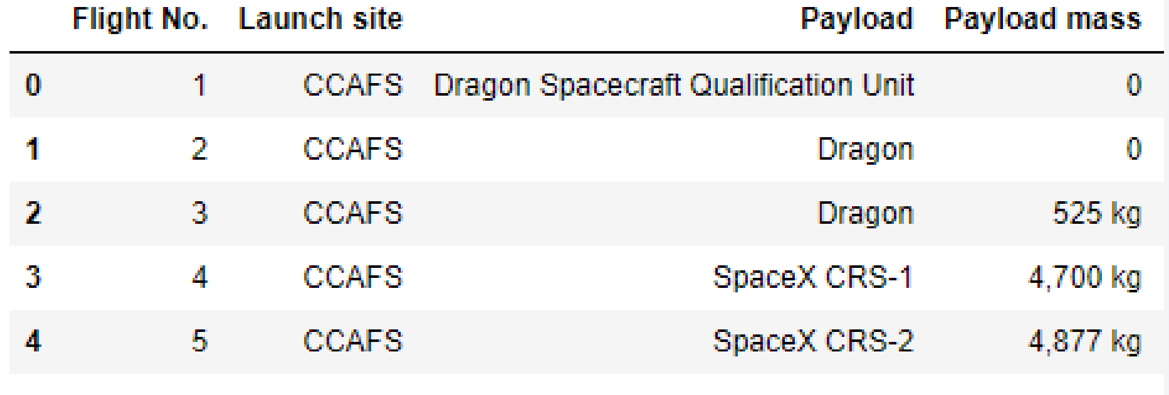
```
# Get the head of the dataframe
data.head()
```

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details	crew	ships	capsules
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	Engine failure at 33 seconds and loss of vehicle	[]	[]	[]
1	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 301, 'altitude': 289, 'reason': 'harmonic oscillation leading to premature engine shutdown'}]	Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage	[]	[]	[]

Data Collection - Scraping

- After scraping SpaceX launches page on wikipedia, our data looks like this :

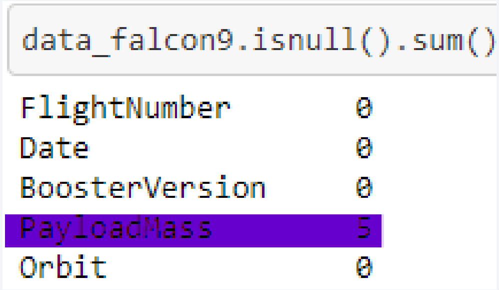
- [GitHub - Data Collection Scraping Notebook]



	Flight No.	Launch site	Payload	Payload mass
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0
1	2	CCAFS	Dragon	0
2	3	CCAFS	Dragon	525 kg
3	4	CCAFS	SpaceX CRS-1	4,700 kg
4	5	CCAFS	SpaceX CRS-2	4,877 kg

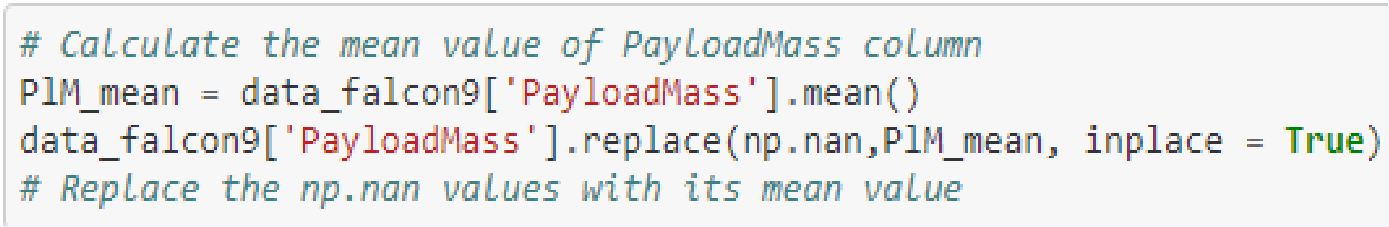
Data Wrangling

- First we deal with missing values by replacing them with the mean as follow:

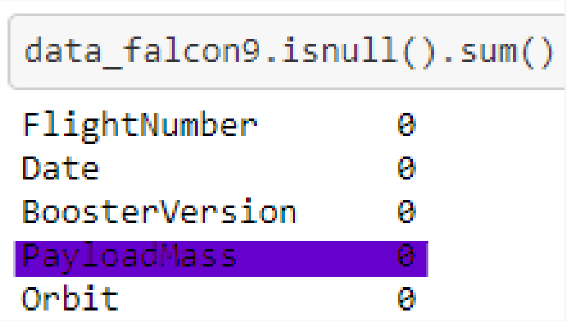


```
data_falcon9.isnull().sum()
```

```
FlightNumber      0
Date              0
BoosterVersion    0
PayloadMass       5
Orbit             0
```



```
# Calculate the mean value of PayloadMass column
PlM_mean = data_falcon9['PayloadMass'].mean()
data_falcon9['PayloadMass'].replace(np.nan,PlM_mean, inplace = True)
# Replace the np.nan values with its mean value
```

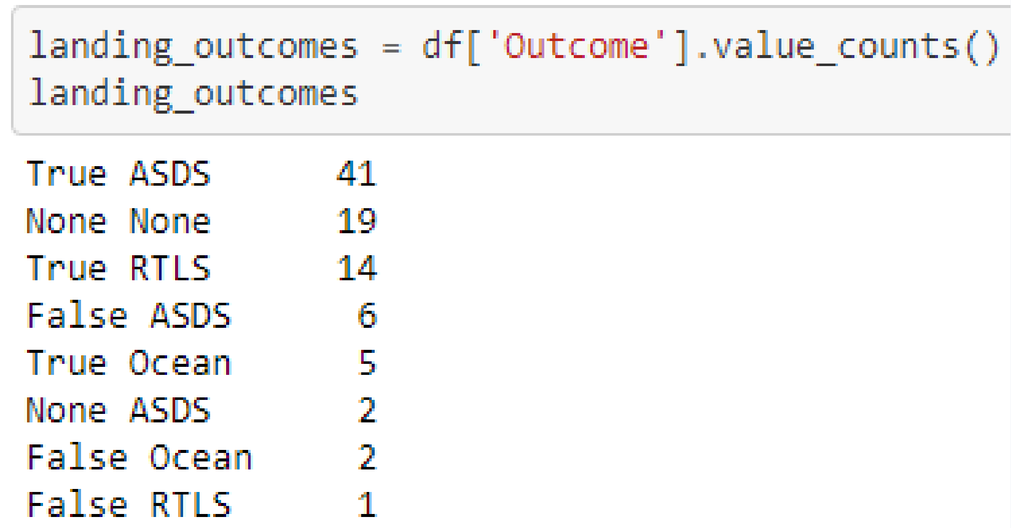


```
data_falcon9.isnull().sum()
```

```
FlightNumber      0
Date              0
BoosterVersion    0
PayloadMass       0
Orbit             0
```

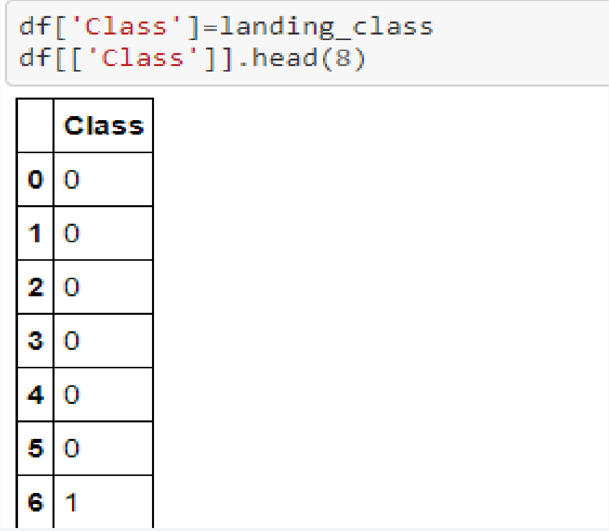
Data Wrangling

- We also chose the appropriate variable for training the models, and we convert it into training labels with 1 for successful landing and 0 for an unsuccessful one.



```
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
None ASDS       2
False Ocean     2
False RTLS      1
```



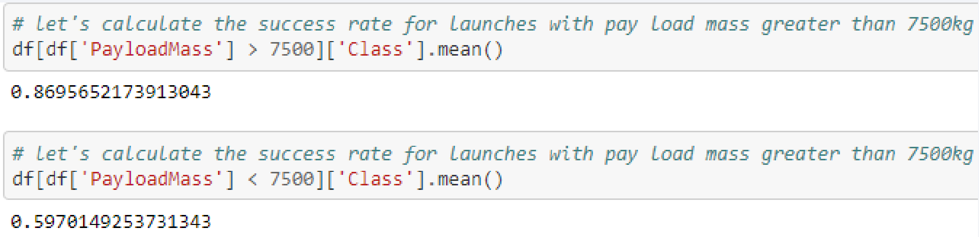
```
df['Class']=landing_class
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1

- Please check the following link: GitHub - Data Wrangling Notebook

EDA with Data Visualization

- We noticed that the greater the Payload mass the higher the landing success rate.



```
# let's calculate the success rate for launches with pay load mass greater than 7500kg
df[df['PayloadMass'] > 7500]['Class'].mean()
```

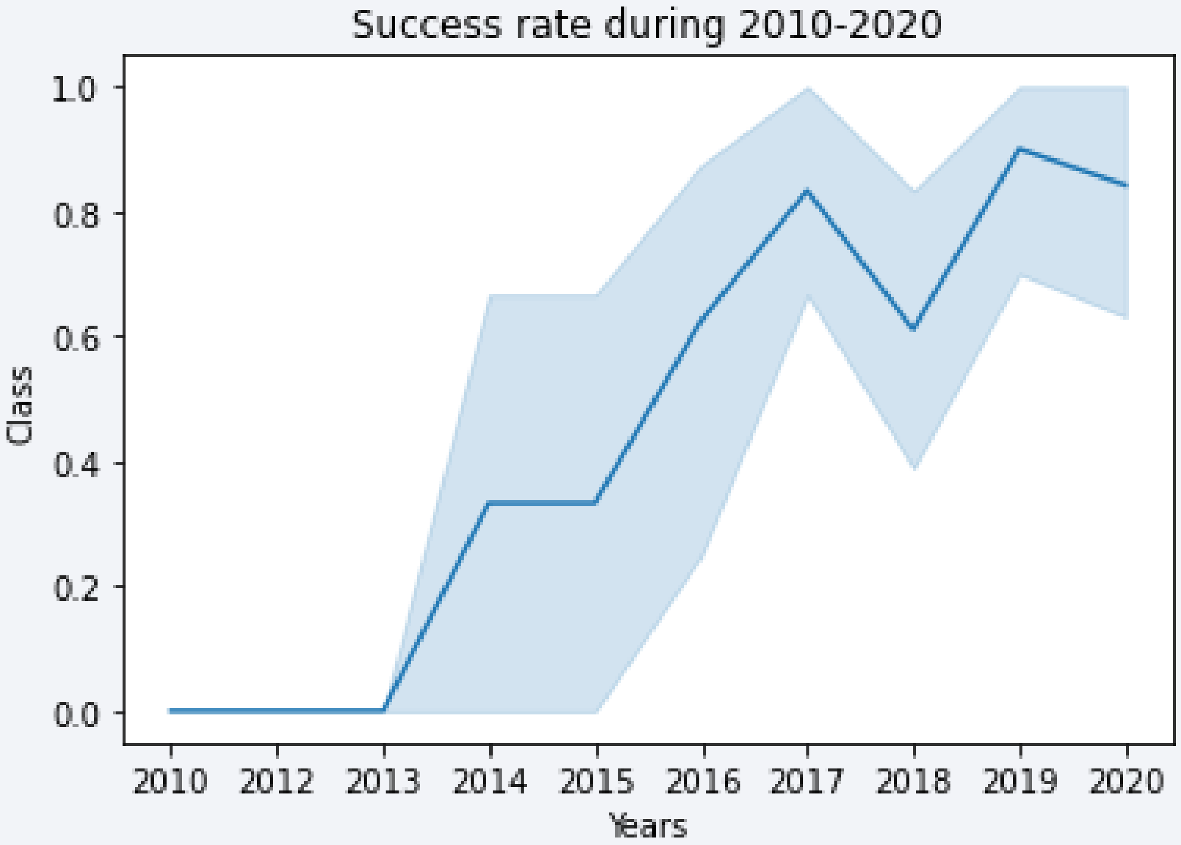
```
0.8695652173913043
```

```
# let's calculate the success rate for launches with pay load mass greater than 7500kg
df[df['PayloadMass'] < 7500]['Class'].mean()
```

```
0.5970149253731343
```

- This was mainly due to the company's mastery of the launching process which was accompanied with the increase of the total mass throughout the years (full notebook below)

[GitHub - EDA with Visualizations Notebook]()



EDA with SQL

- *Display the names of the unique launch sites in the space mission*
- *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 acheived.*
- *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 between the date 2010-06-04 and 2017-03-20, in descending order*

GitHub - EDA with SQL Notebook

Build an Interactive Map with Folium

- We made markers on the map to visualize launch sites locations.
- To get a good sense of the distance between launch sites we added stright lines with a popup showing the measured distance.
- We also created marker clusters that are label-colored to identify easily the number of successful and unsuccessful landing outcomes per launch site.

[GitHub - Analysis with Interactive Map]

Build a Dashboard with Plotly Dash

- We made a dropdown menu to let us select between all sites or a specefic one and display the graphs accordingly.
- When all sites are selected a map is shown where all sites locations are detected, a pie chart is displayed and exhibit insights about the successful launches rates per launch site. There is also a scatter plot showing the correlation between booster versions and landing outcome for all sites.
- When a specific site is selected, it's location is revealed in the map and a pie chart is rendered where the landing outcomes percentages are shown. Similarly, the scatter plot visualize landing outcomes per booster versions, but this this time for the specified launch site.
- For the sake or flexibility we added a range slider to make it easier for the user to specify the Pay Load Mass range.

GitHub - Analysis using Dash

Dashboard - Live link

Predictive Analysis (Classification)

- In this section, we try to find the best classifier for our data based on four chosen models : K-Nearest-Neighbors (KNN), Decision Trees, Support Vector Machine (SVM), and Logistic Regression
- First we prepare our data by standardizing and split it into training data and test data.
- We create a Grid Search CV for each model, and feed it with appropriate dictionaries to look for the best hyperparameters
- Finally, we display a confusion matrix along with the accuracy of the prediction

[GitHub - Predictive Analysis Notebook]

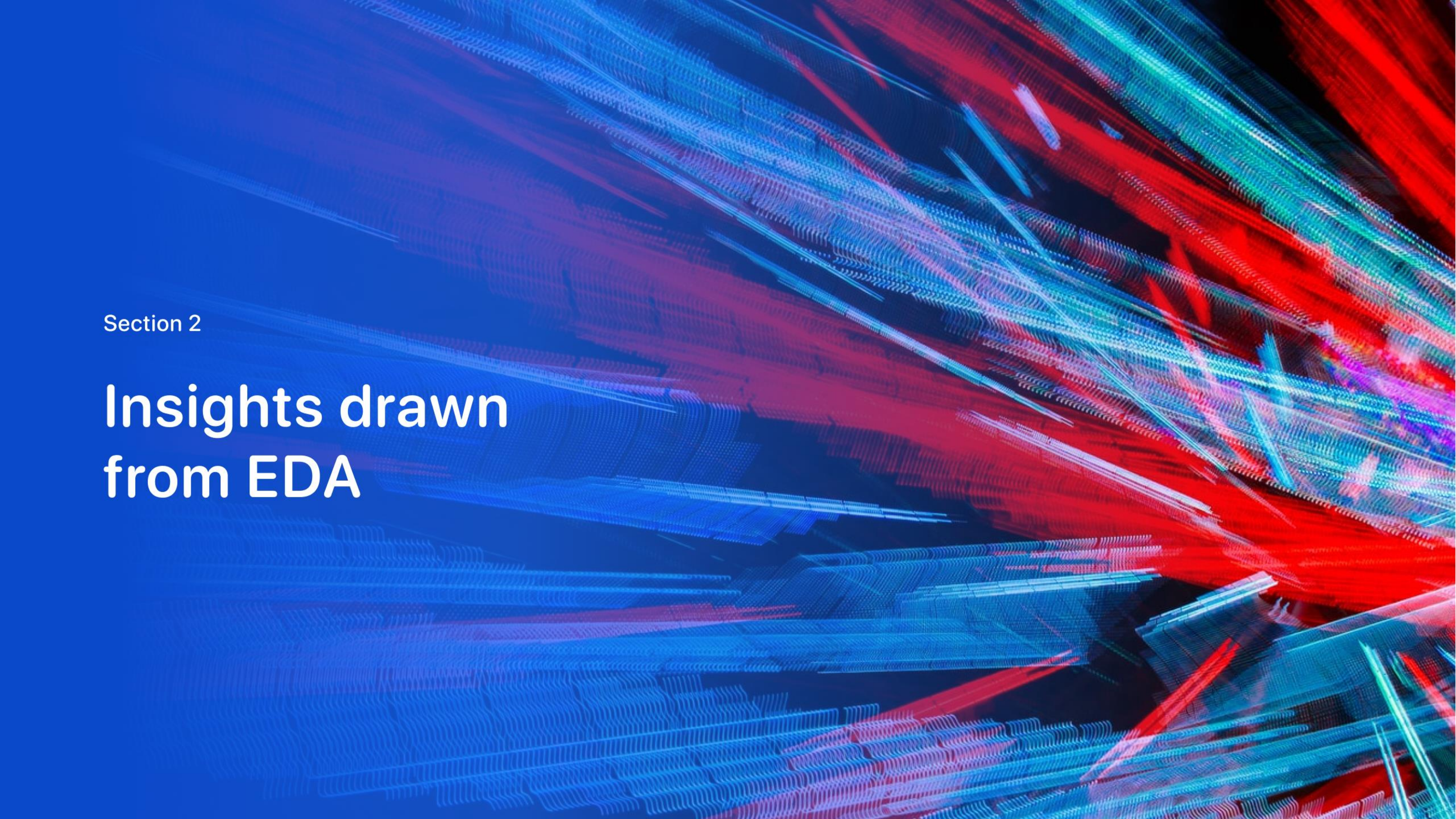
Results

We exhibit our analysis results in three parts:

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

Results

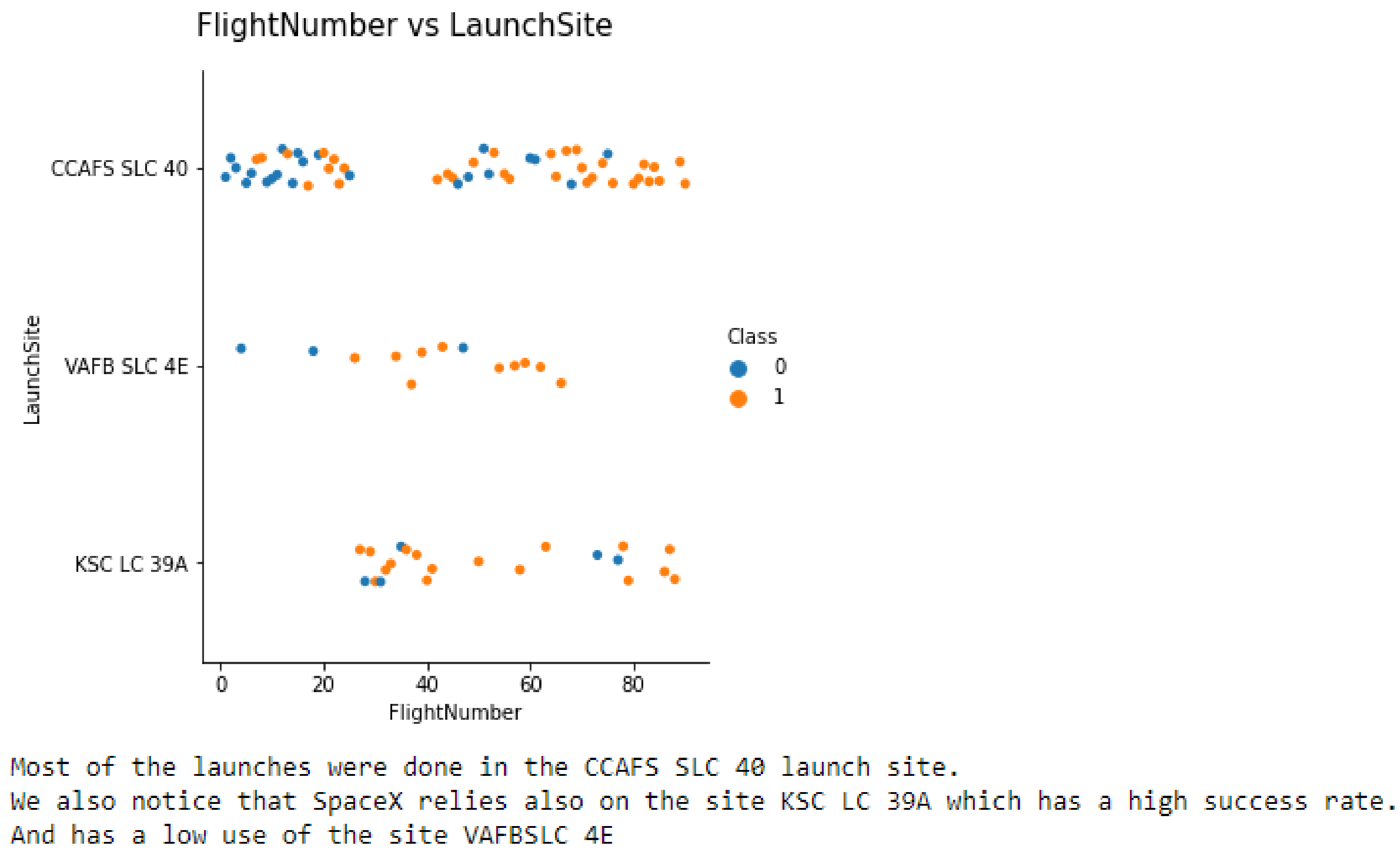
- After conducting exploratory data analysis, we find out important insights about landing outcomes and its relation with orbit types booster versions and launch sites.
- We get to explore in powerful map visualizations the launch sites proximities, distance between launch sites, and of course the landing outcomes per site.
- Lastly we build a classifier that can predict with an accuracy of more than 83%, the landing outcome.



Section 2

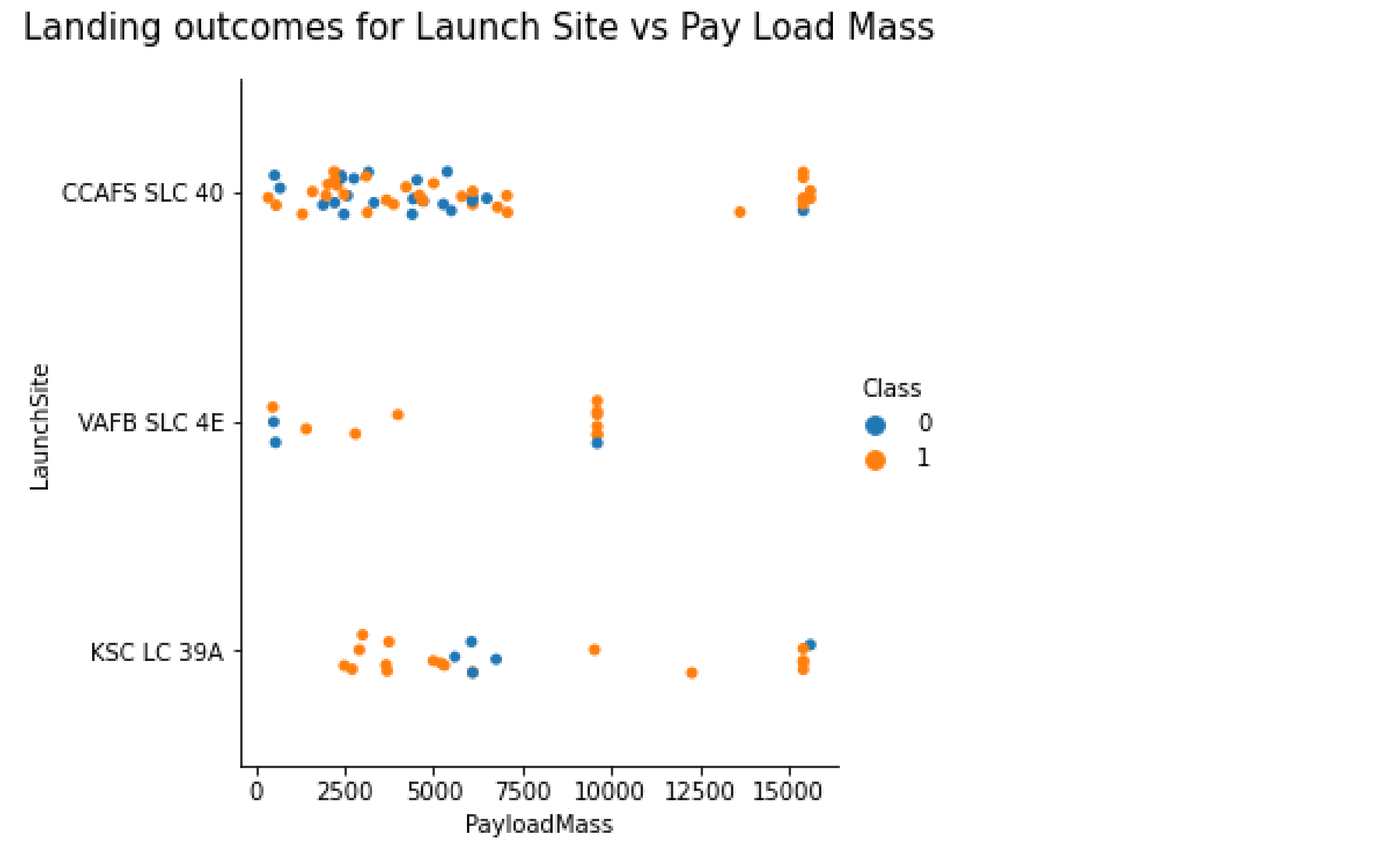
Insights drawn from EDA

Flight Number vs. Launch Site



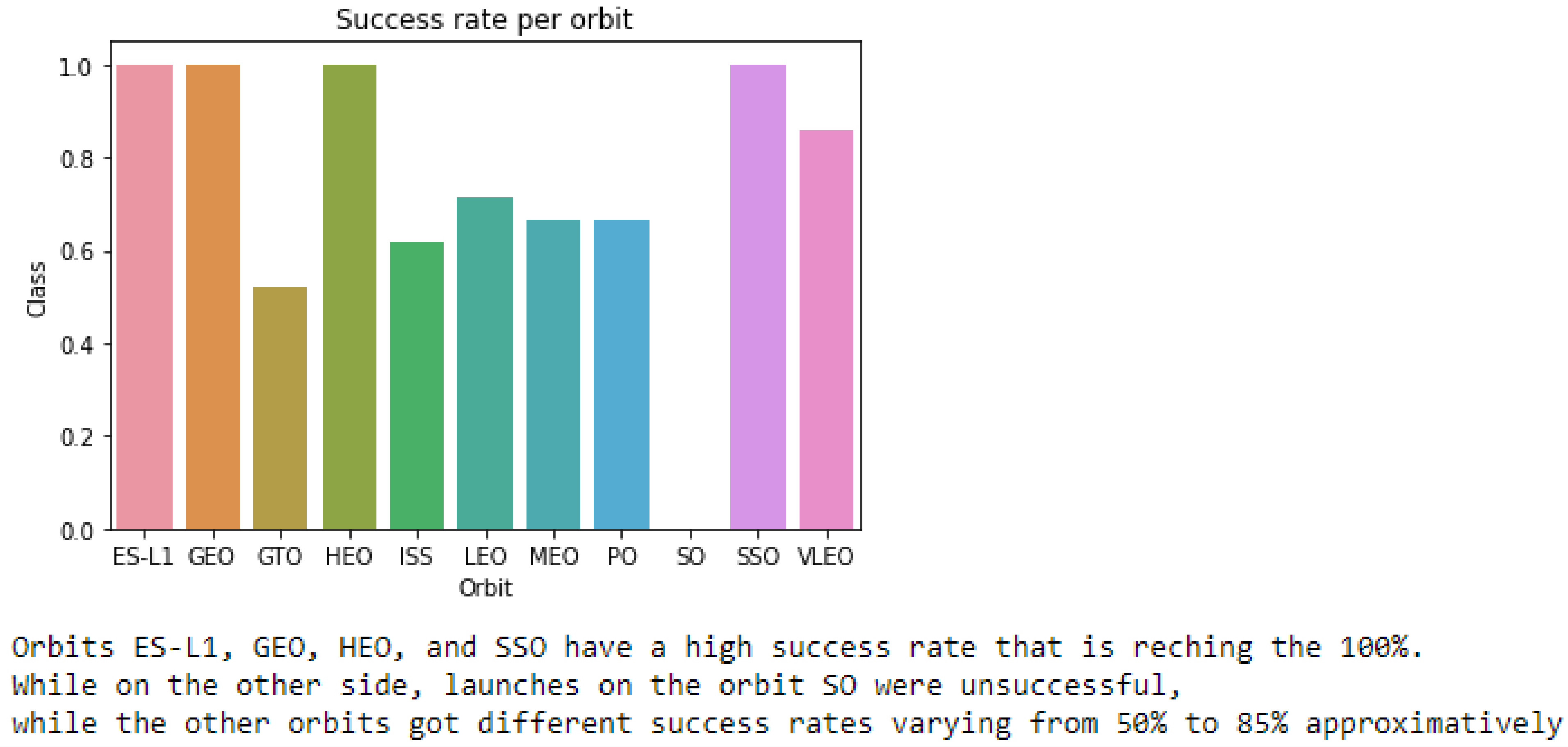
```
Most of the launches were done in the CCAFS SLC 40 launch site.
We also notice that SpaceX relies also on the site KSC LC 39A which has a high success rate.
And has a low use of the site VAFBSLC 4E
```

Payload vs. Launch Site



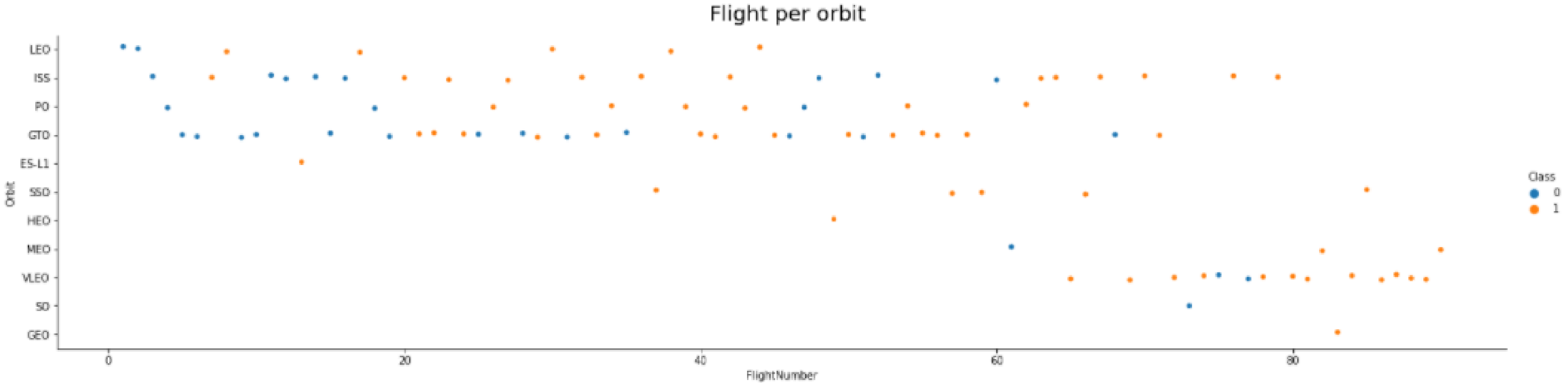
Surprisingly, most of the data points have a payload mass less than 7500kg, almost all the flights with payload mass above 7500kg have successful landing.

Success Rate vs. Orbit Type



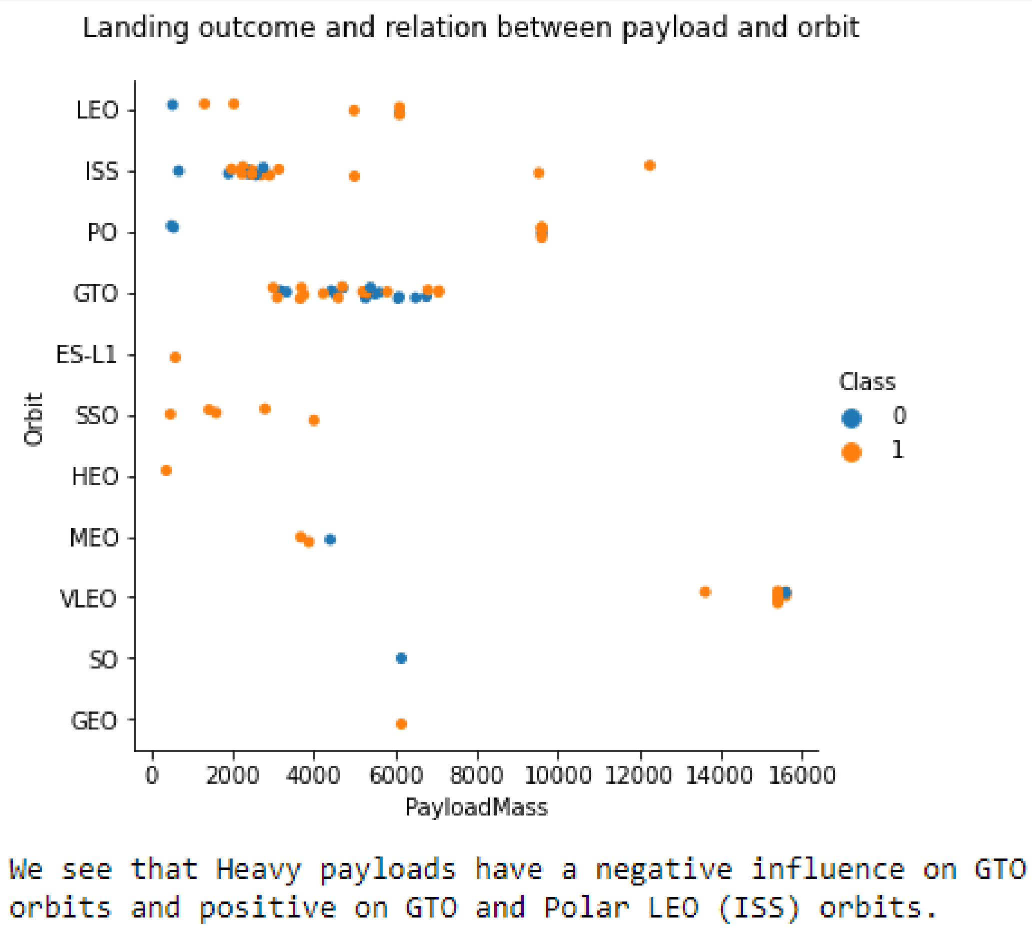
Orbits ES-L1, GEO, HEO, and SSO have a high success rate that is reching the 100%.
While on the other side, launches on the orbit SO were unsuccessful,
while the other orbits got different success rates varying from 50% to 85% approximatively

Flight Number vs. Orbit Type



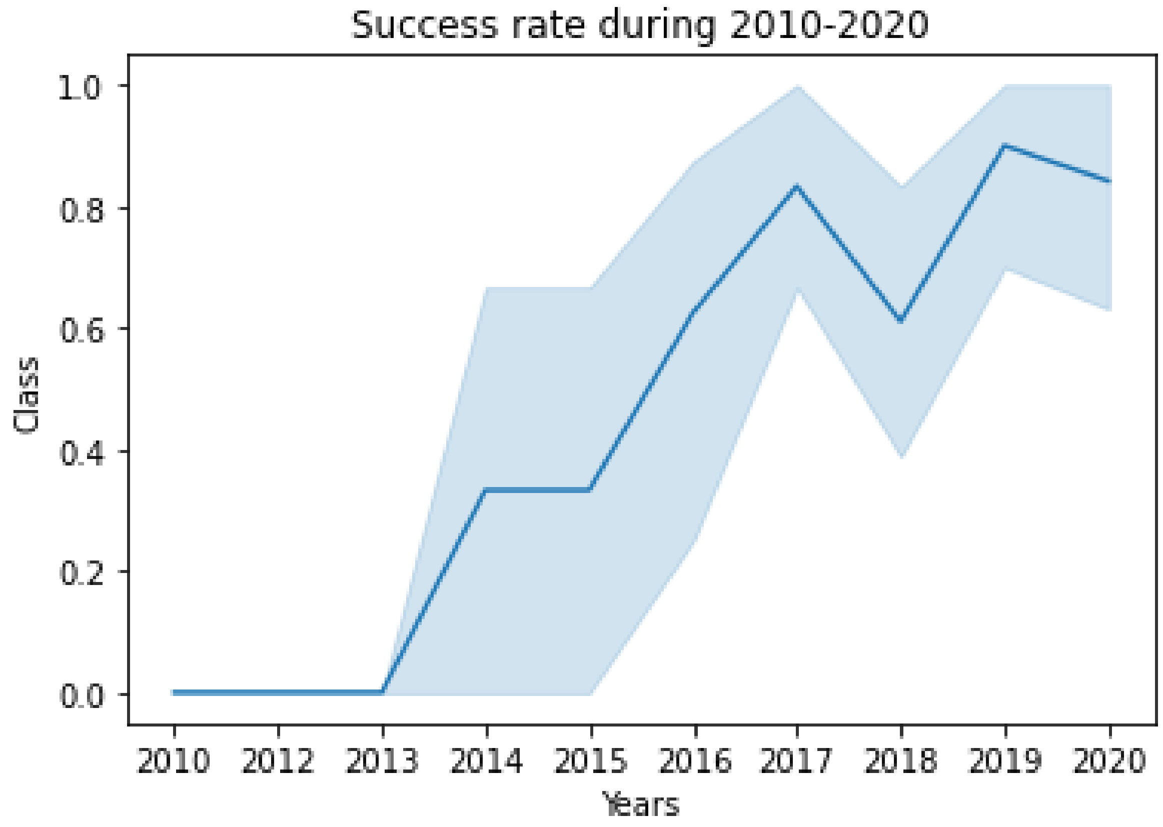
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



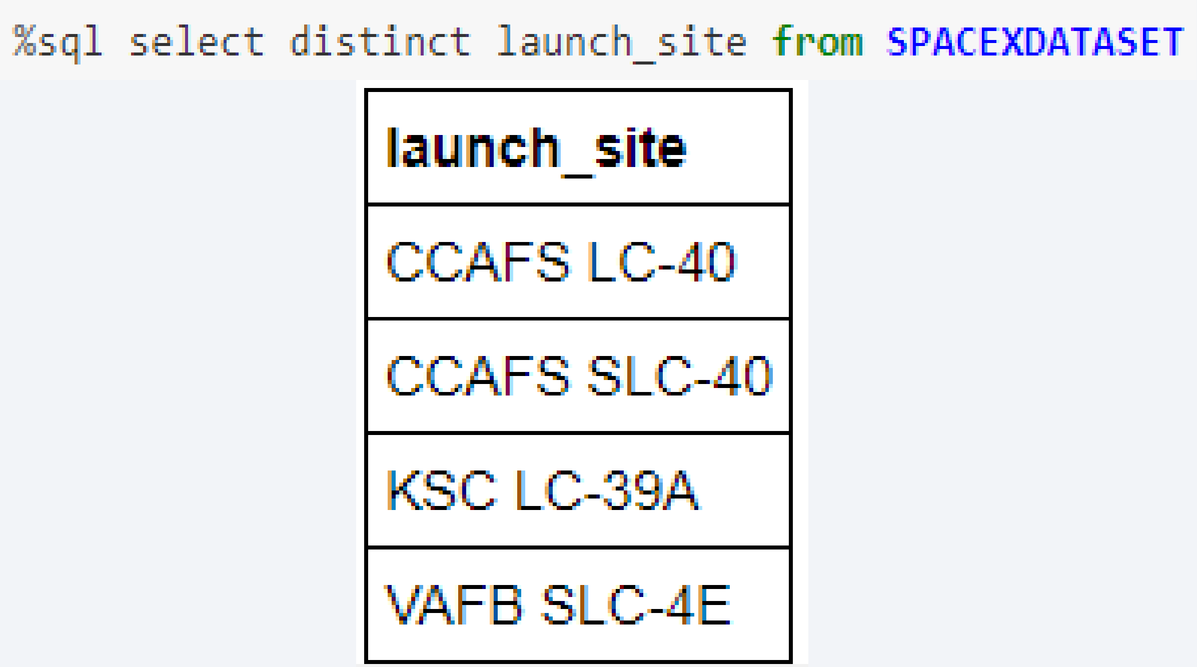
We see that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



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

All Launch Site Names

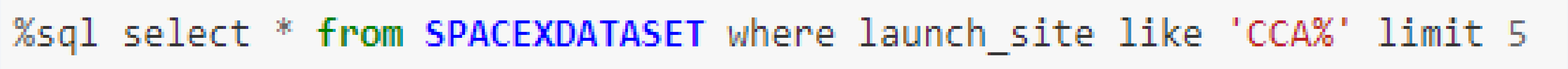


```
%sql select distinct launch_site from SPACEXDATASET
```

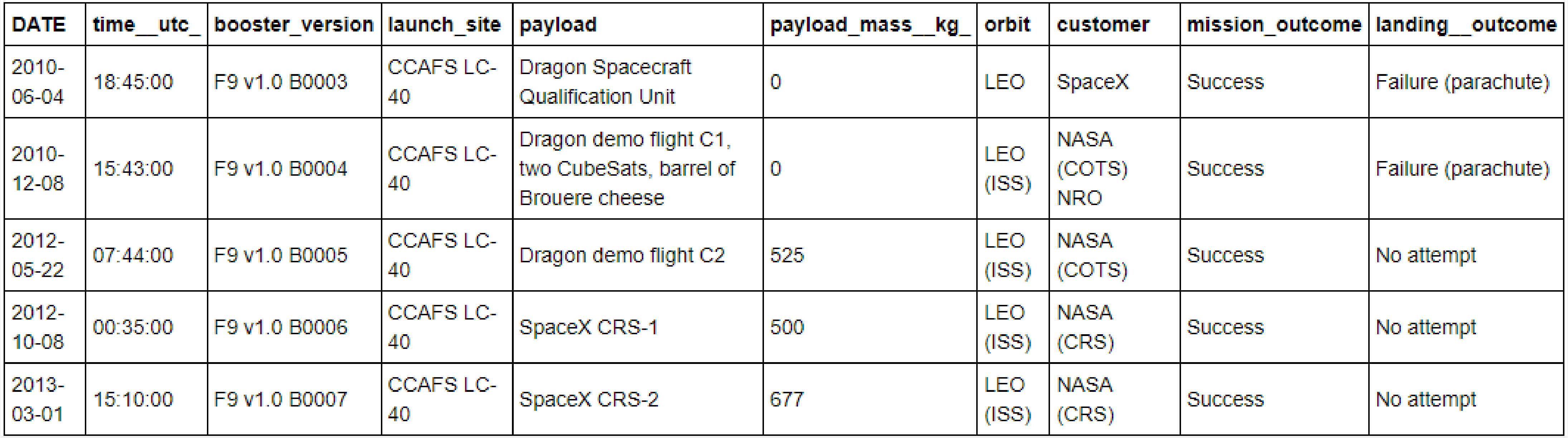
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Mainly SpaceX perform its Falcon 9 rocket launches on the sites mentioned above.

Launch Site Names Begin with 'CCA'



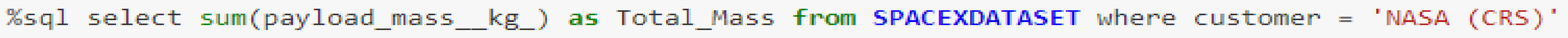
```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```



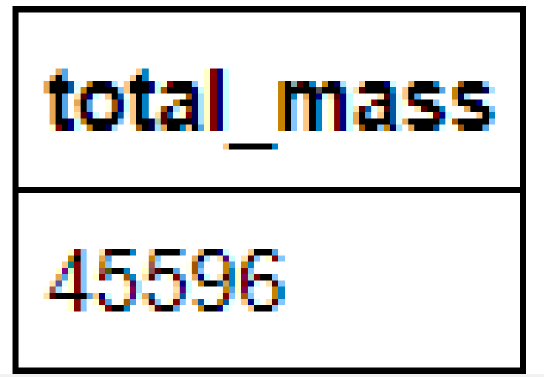
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Here are the records of 5 sites starting with the letters 'CCA'

Total Payload Mass



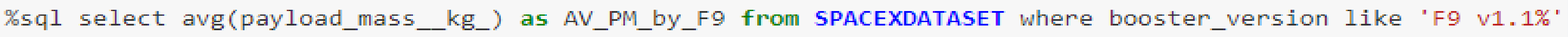
```
%sql select sum(payload_mass__kg_) as Total_Mass from SPACEXDATASET where customer = 'NASA (CRS)'
```



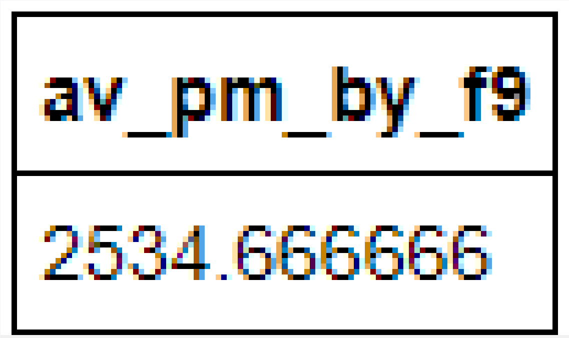
total_mass
45596

We see here the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1



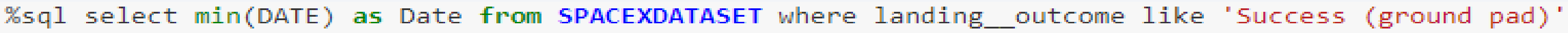
```
%sql select avg(payload_mass__kg_) as AV_PM_by_F9 from SPACEXDATASET where booster_version like 'F9 v1.1%'
```



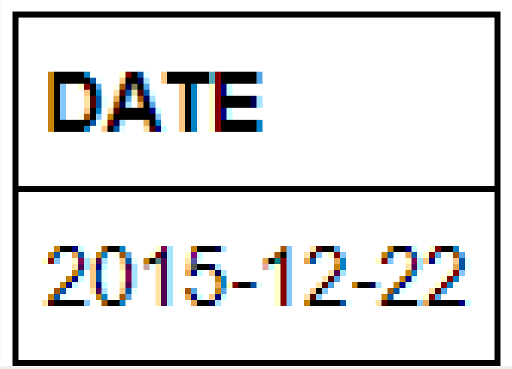
av_pm_by_f9
2534.666666

This is the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date



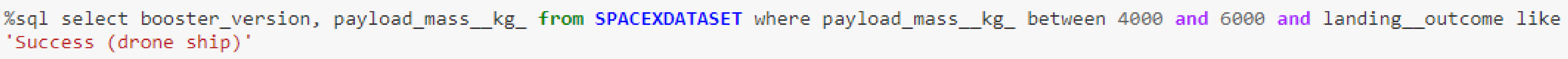
```
%sql select min(DATE) as Date from SPACEXDATASET where landing__outcome like 'Success (ground pad)'
```



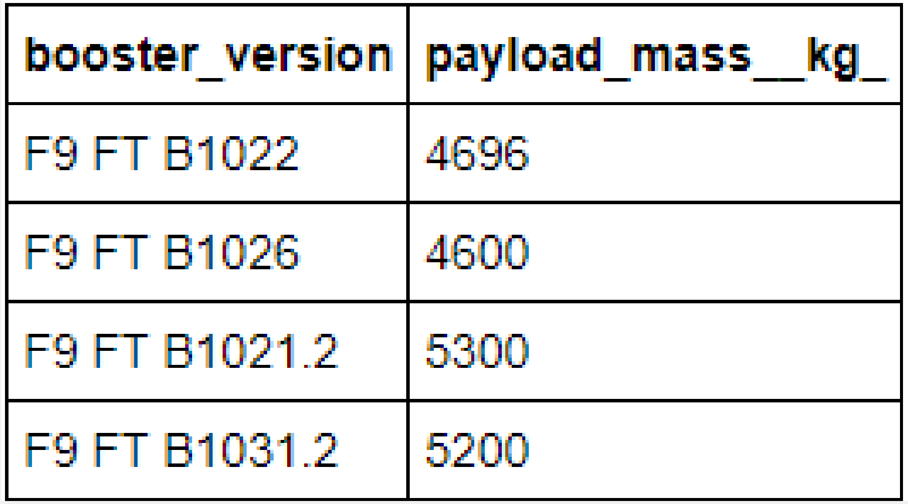
DATE
2015-12-22

This is the date of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000



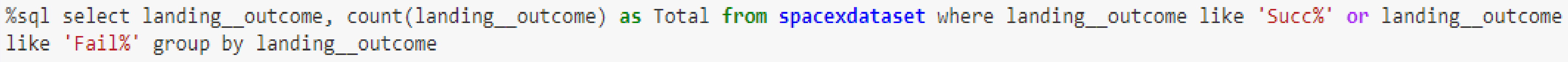
```
%sql select booster_version, payload_mass__kg_ from SPACEXDATASET where payload_mass__kg_ between 4000 and 6000 and landing__outcome like
'Success (drone ship)'
```



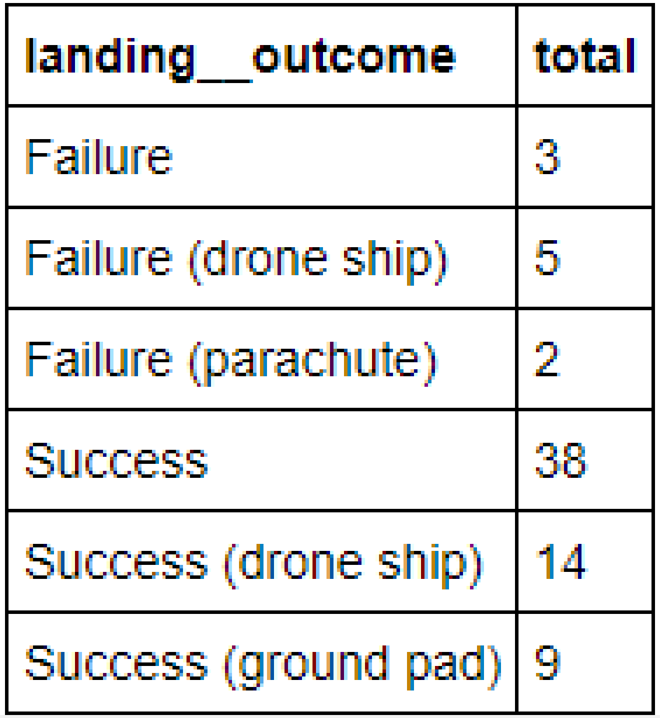
booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

This is a list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes



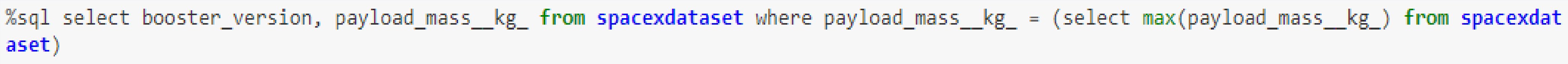
```
%sql select landing__outcome, count(landing__outcome) as Total from spacexdataset where landing__outcome like 'Succ%' or landing__outcome
like 'Fail%' group by landing__outcome
```



landing__outcome	total
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
Success	38
Success (drone ship)	14
Success (ground pad)	9

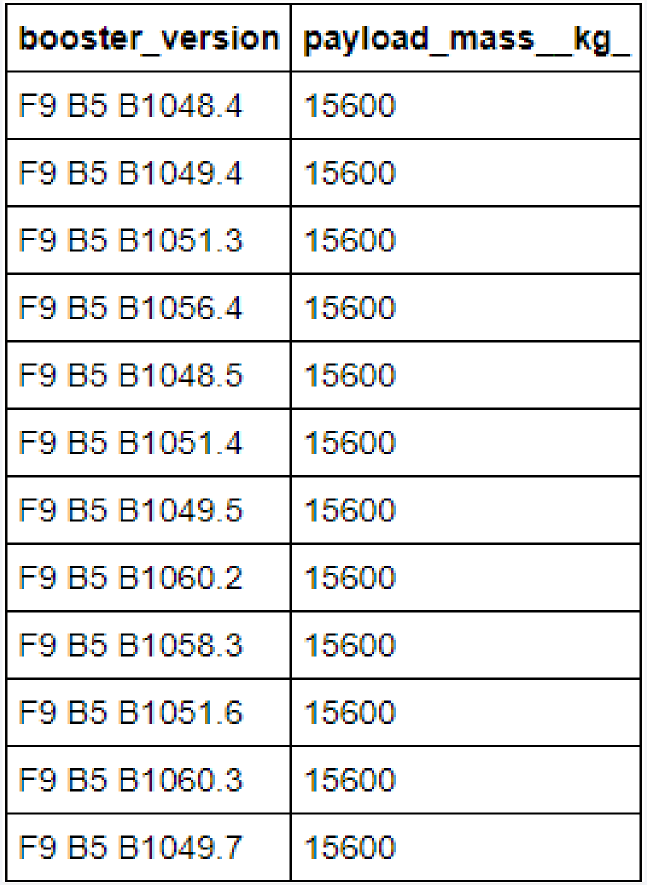
This table shows the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload



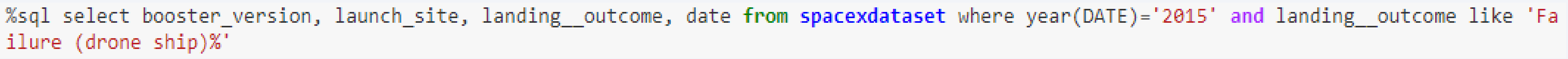
```
%sql select booster_version, payload_mass__kg_ from spacexdataset where payload_mass__kg_ = (select max(payload_mass__kg_) from spacexdataset)
```

The following table lists the names of the booster which have carried the maximum payload mass

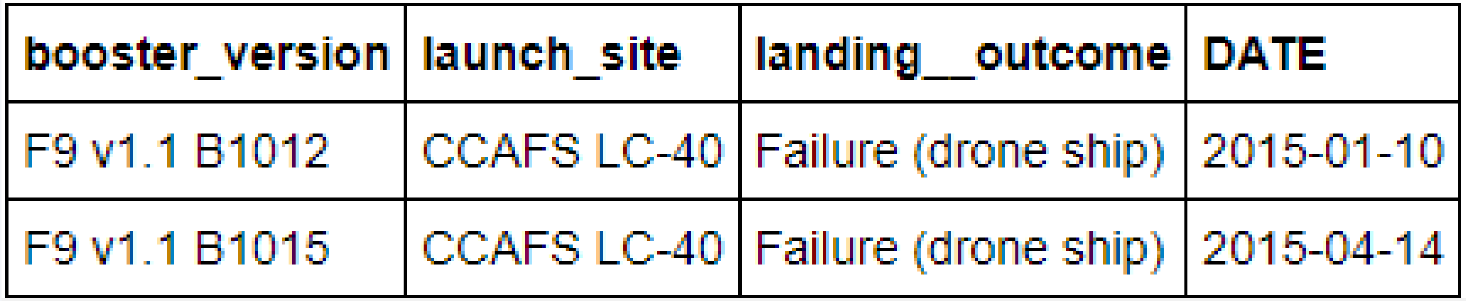


booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records



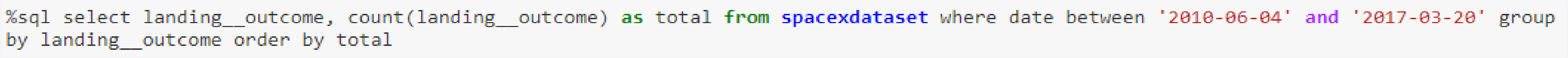
```
%sql select booster_version, launch_site, landing__outcome, date from spacexdataset where year(DATE)='2015' and landing__outcome like 'Fa
ilure (drone ship)%'
```



booster_version	launch_site	landing__outcome	DATE
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-01-10
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14

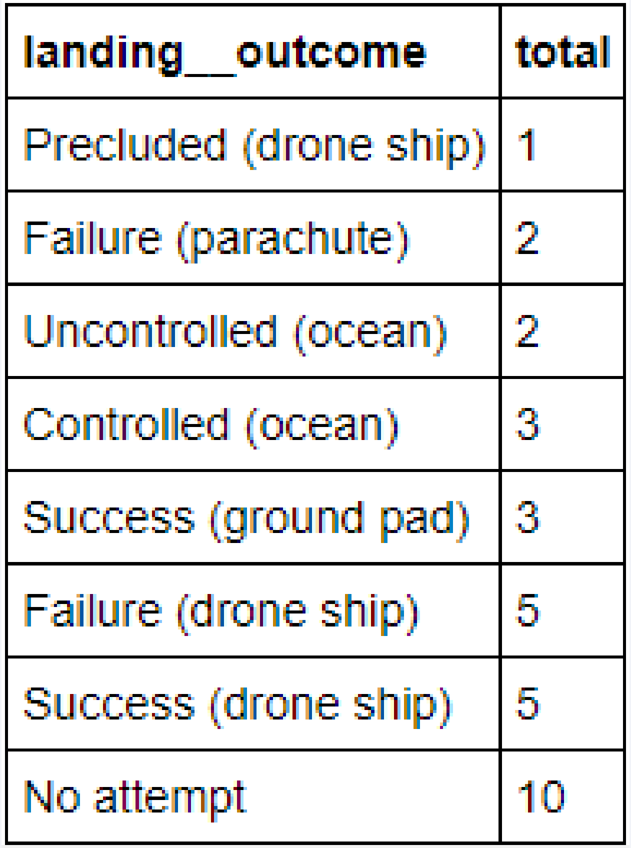
These are the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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



```
%sql select landing__outcome, count(landing__outcome) as total from spacexdataset where date between '2010-06-04' and '2017-03-20' group
by landing__outcome order by total
```

Here we can check the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

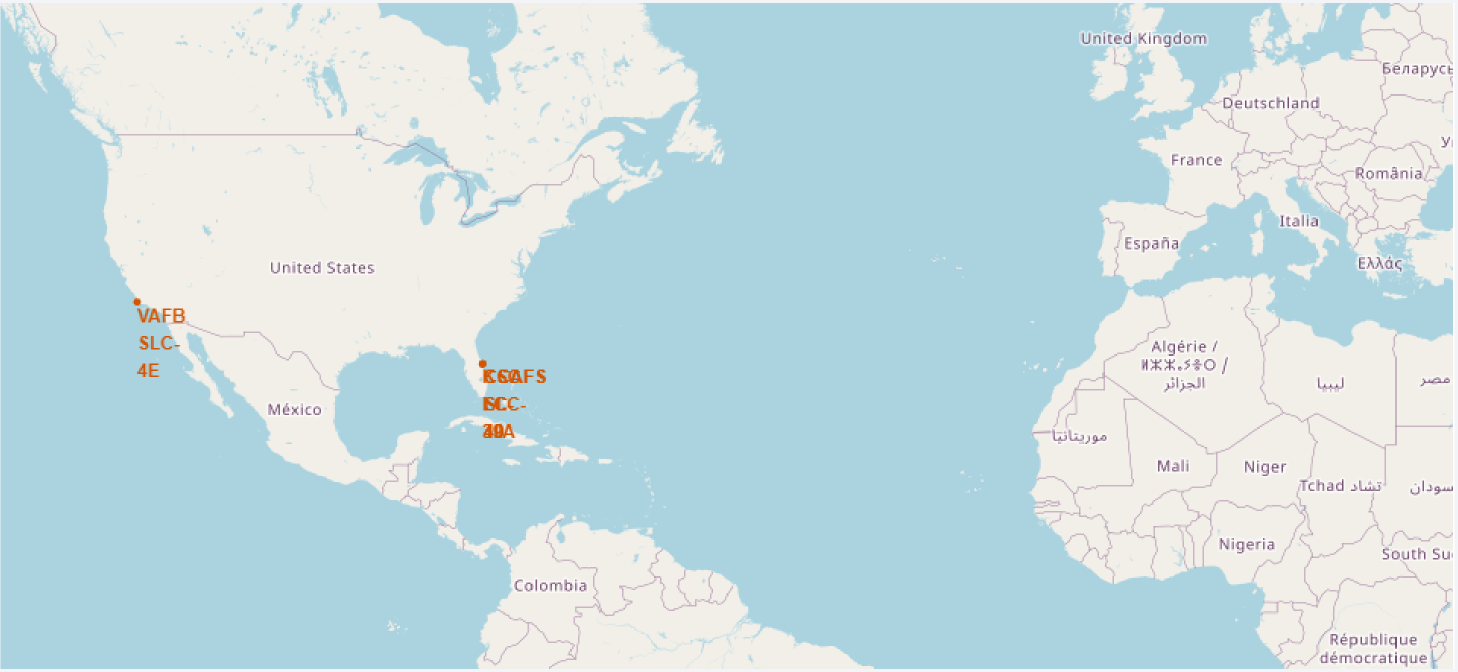


landing__outcome	total
Precluded (drone ship)	1
Failure (parachute)	2
Uncontrolled (ocean)	2
Controlled (ocean)	3
Success (ground pad)	3
Failure (drone ship)	5
Success (drone ship)	5
No attempt	10

Section 4

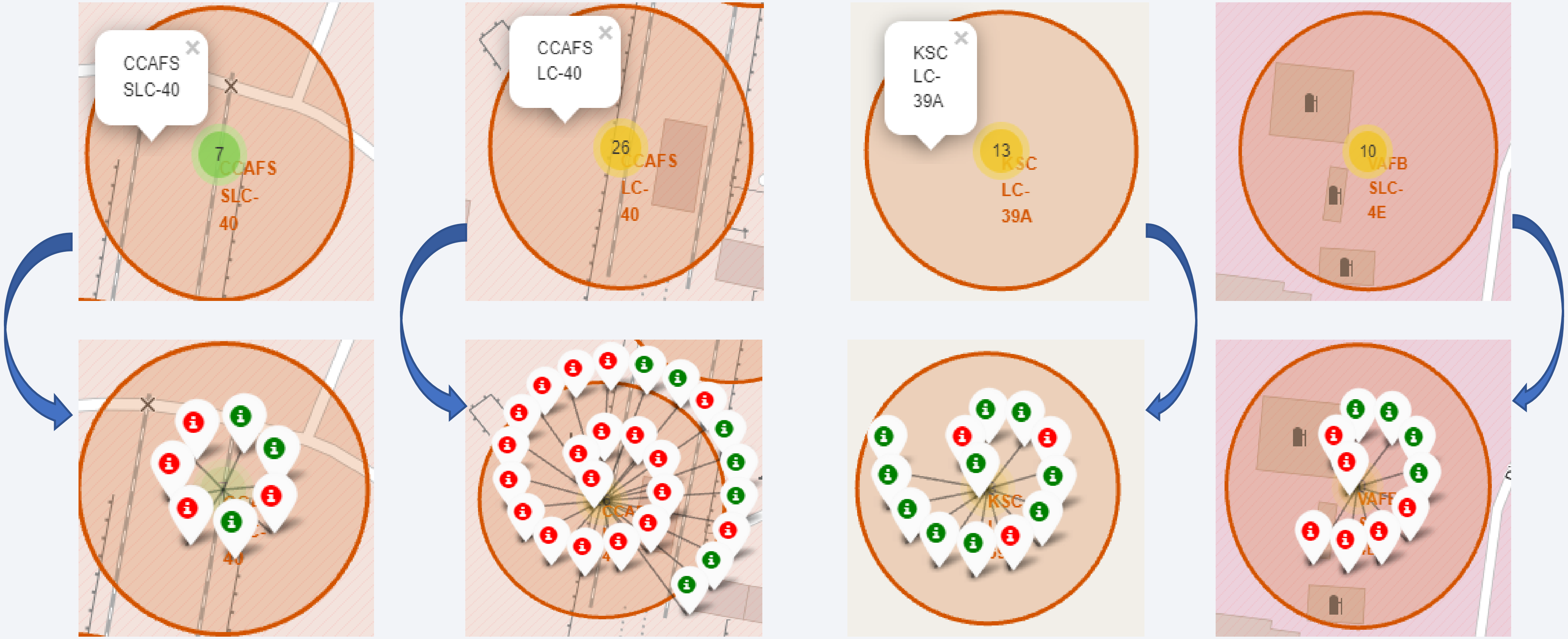
Launch Sites Proximities Analysis

Launch sites Marked on the Map



We can see that all launch sites are located in the USA and are close to coastlines.

Launches outcomes per site



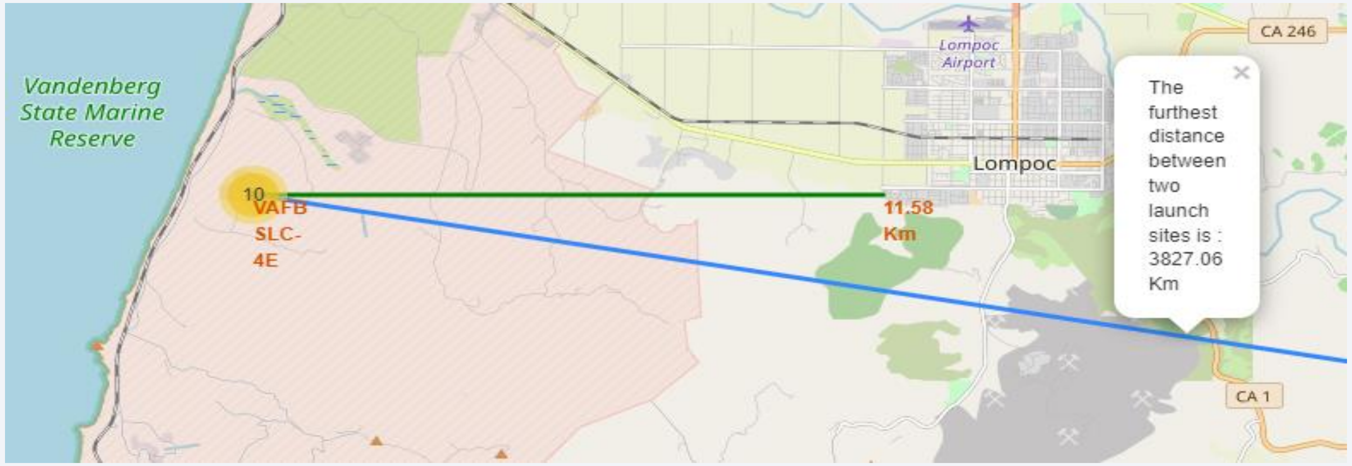
We can see properly total launches per site, and even the successful ones in green and unsuccessful ones in Red.

Launch site and proximities

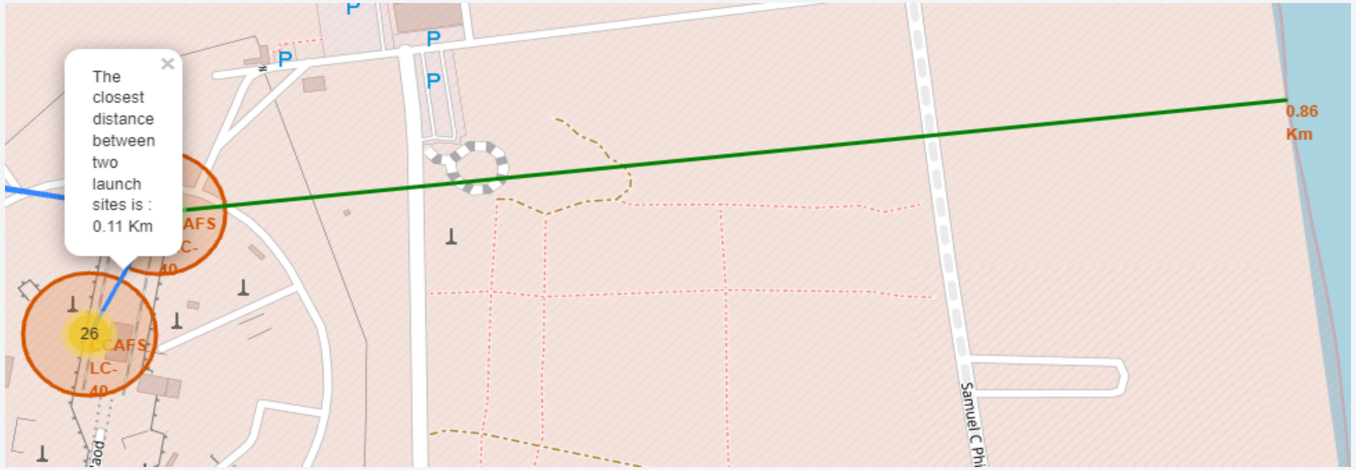
Here we get a better idea about the distance between launch sites.

Also, by exploring proximities we can tell that there are no residencial facilities in less than 11Km from launch sites.

Dashboard - Live link



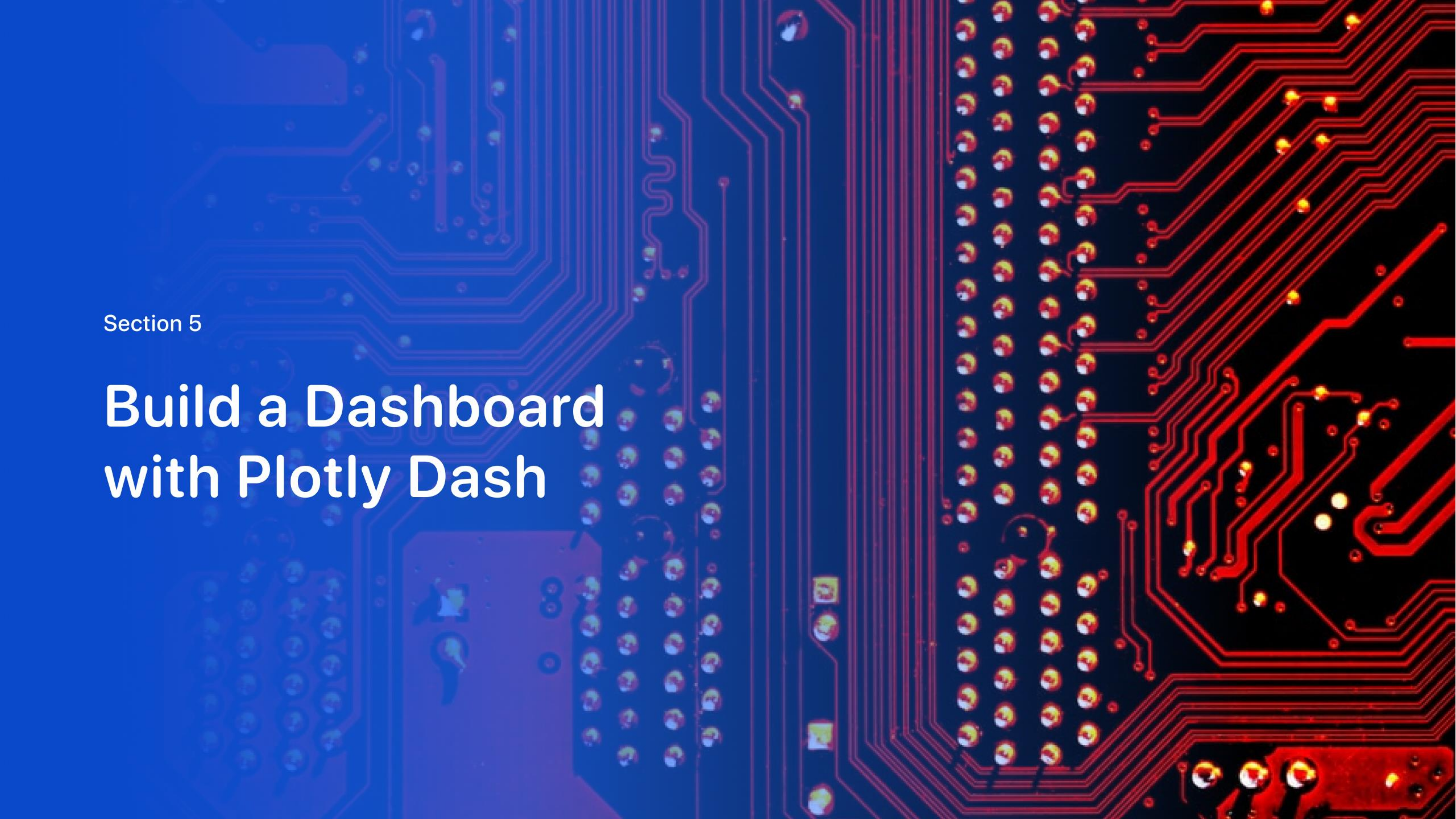
The furthest distance between two launch sites is : 3827.06 Km



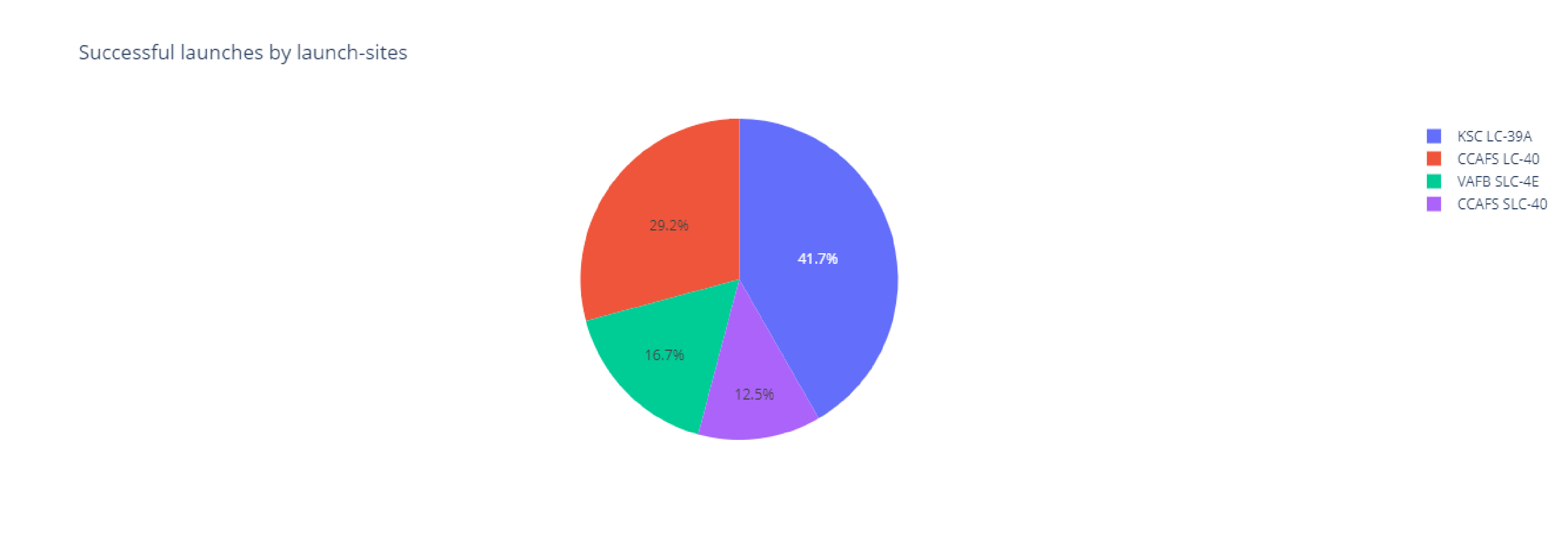
The closest distance between two launch sites is : 0.11 Km

Section 5

Build a Dashboard with Plotly Dash

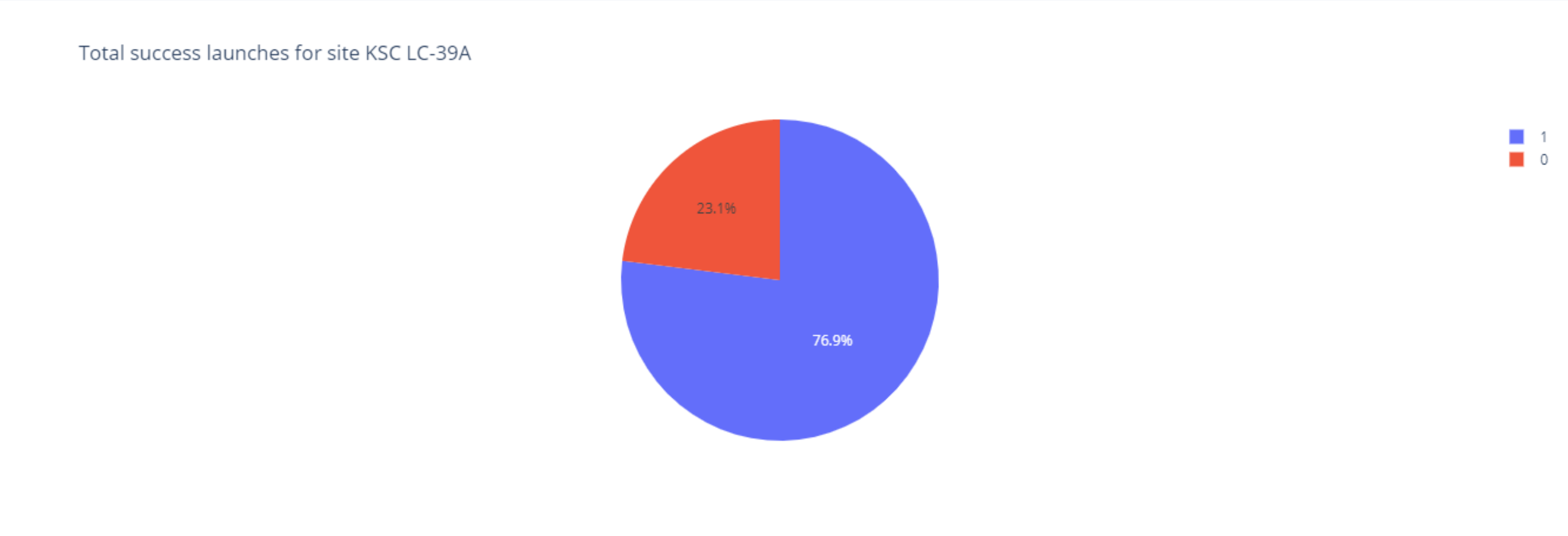


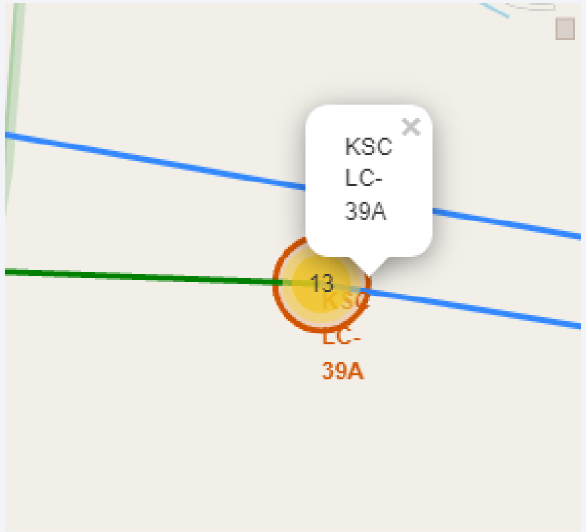
Success rate per site



We can see that success rates differ largely from site to another, while KSC LC-39A has the largest success rate that is 41,2%, only 12.5% of successful launches were on CCAFS SLC-40 launch site.

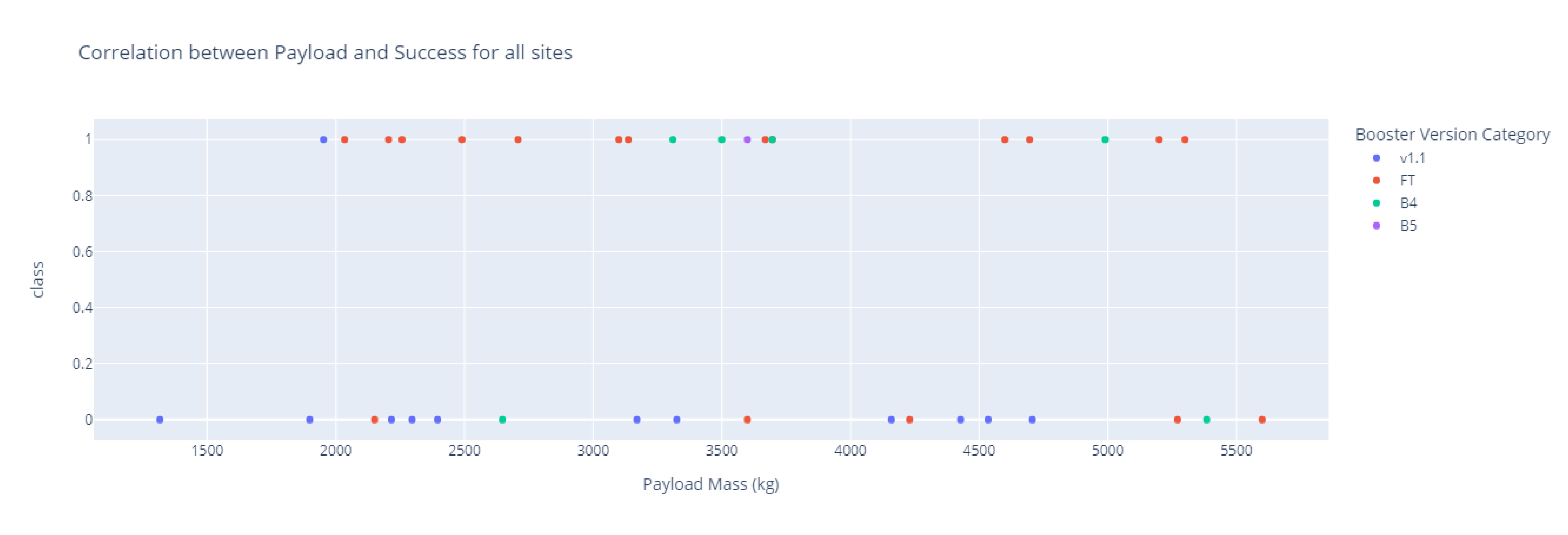
The site with the highest success rate





When we check closely we see that the KSC LC-39A launch site has the highest success rate (76.9%) from all the launches that were performed in each site and only 23.1% of the launches fail.

Landing outcome per booster version vs Payload Mass



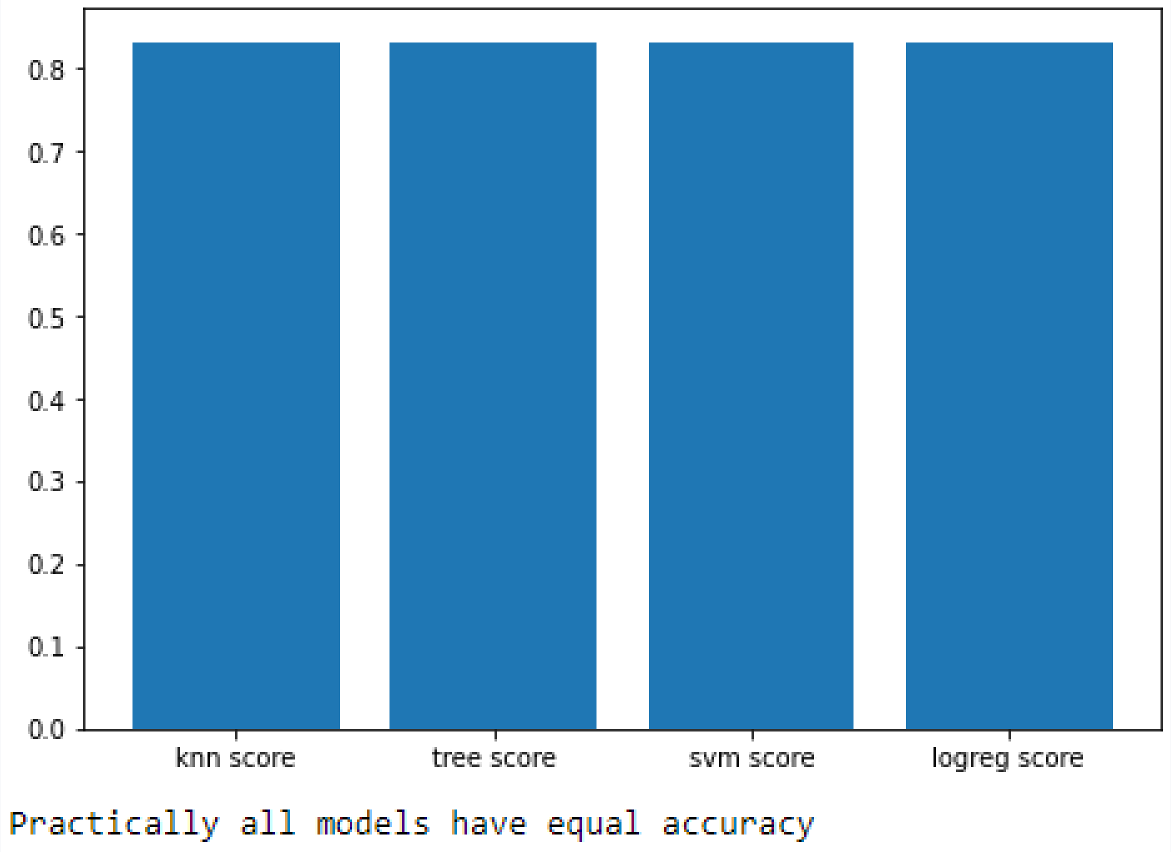
For Payload Mass ranging 1000kg to 6000kg, the booster version FT has an outstanding success rate

Dashboard - Live link

Section 6

Predictive Analysis (Classification)

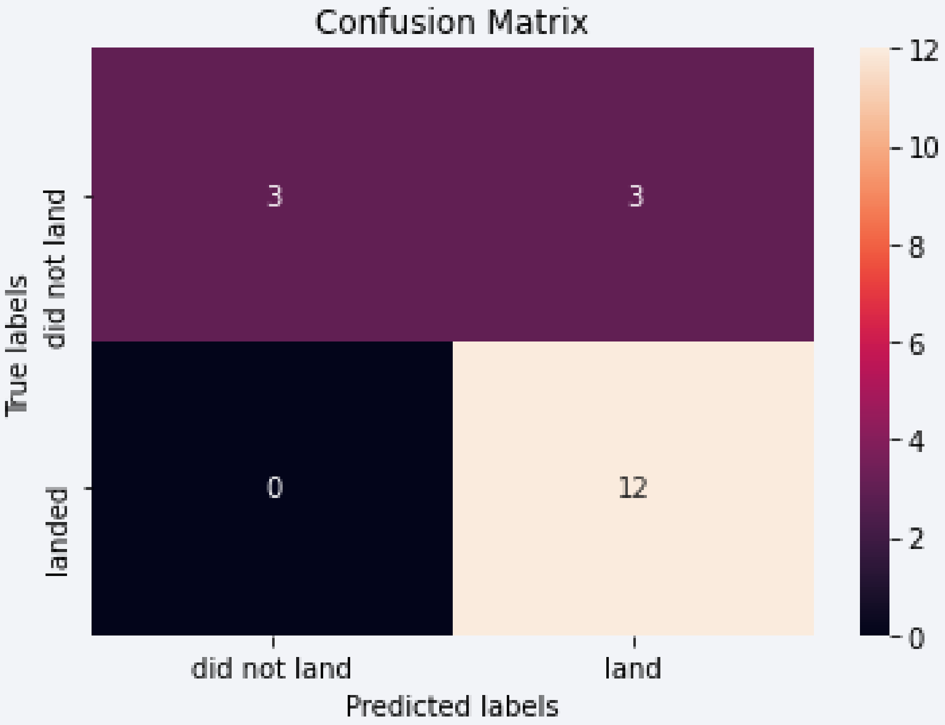
Classification Accuracy



```
Practically all models have equal accuracy
```

After finding the best hyperparameters for each model, the prediction's accuracy that was reached for all models was the highest (83.33%)

Confusion Matrix



The models were able to predict all the launches that 'landed' successfully, but wrongly predicted 3 launches that 'did not land' as 'landed'.

Conclusions

- Space X Falcon 9 launches had a low success rate (0%) in its first years.
- The success rate and the total payload mass have a rising trend throughout the years.
- While CCAFS LC-40 is the launch site with the most launches, SpaceX has got its greatest records in KSC LC-39A launch site.
- ES-L1, SSO, HEO and GEO orbits have 100% success rate.
- The established models can predict with an accuracy of more than 83% the outcome of the launches where it labels correctly 15 data points out of 18 data points used for testing.

Acknowledgements

I would like to express my special thanks to all of IBM's Data Science Professional Certificate's amazing instructors *and also to Coursera's state-of-the-art platform who made this golden opportunity available to work on this wonderful project on the topic which practically gathers everything learned from unit 1.*

Secondly i would also like to thank course moderators whose post were very helpful and allowed me to finalize this project within the limited time frame.

Thank you!