IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Edmund Nesveda
15.06.2022

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

SpaceY is looking to enter the market for rocket launches as direct competitor to SpaceX. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Our project is to try to determine the likelihood of first stage will successfully land and therefore if we can accurately predict the likelihood of the first stage rocket landing successfully, we can determine the cost of a launch

to add text

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - How data was collected
- Perform data wrangling
  - How data was processed
- 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 used two methods to collect the required data:

Webscraping the Wikipedia page titled 'List of Falcon 9 and Falcon Heavy launches', using the requests and beautiful soup Python libraries

Directly from the Spacexdata website using the REST API provided by SpaceX

Data Collection – SpaceX API

Form the whole dataset we wanted only specific information:

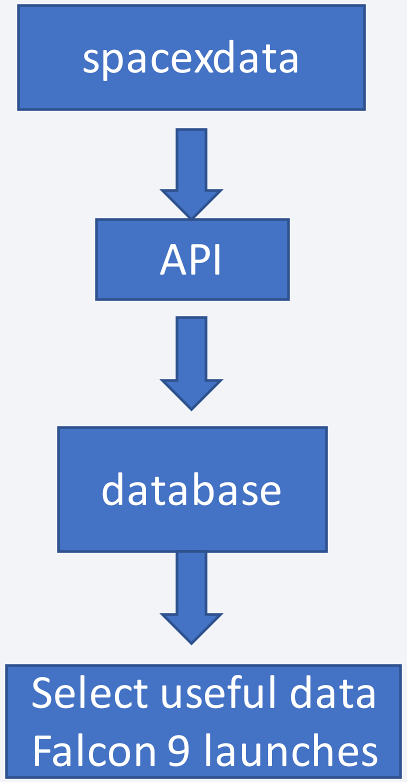
From the rocket column we would like to learn the booster name.

From the launchpad we wanted the name of the launch site being used, the longitude, and the latitude

From the payload, the mass of the payload and the orbit that it is going to.

From cores we would like to learn the outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to separate version of cores, the number of times this specific core has been reused, and the serial of the core.

Finally we filtered only the Falcon 9 launces.

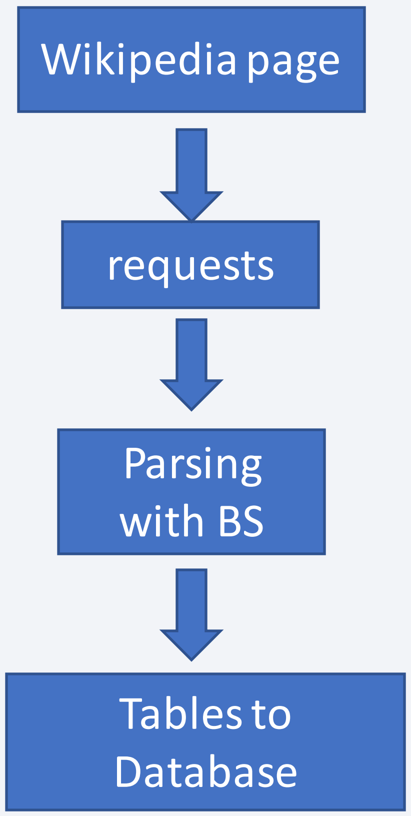


https://github.com/ednes76/Capstone-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Webscraping Falcon 9 launch records with Requests and BeautifulSoup libraries:

1. First we extracted the Falcon 9 launch records HTML table from Wikipedia
2. Next we parsed the table and convert it into a Pandas data frame
3. Finally we exported the dataframe to CVS file



https://github.com/ednes76/Capstone-project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

In the EDA stage, we will try to find some patterns in the data and determine what would be the label for training supervised models. In the data set, there are several different cases where the booster did not land successfully. These outcomes were converted into Training Labels with 1 meaning the booster successfully landed 0 meaning it was unsuccessful.

These are the steps we followed in obtaining the Training Labels:

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column

https://github.com/ednes76/Capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

At this Stage we looked at the relationship between:

1. Flight Number and Launch Site
2. Payload and Launch Site
3. Success rate of each orbit type
4. Flight Number and Orbit type
5. Payload and Orbit type
6. Launch success yearly trend

https://github.com/ednes76/Capstone-project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

For this stage we use SQL database for a better understanding of the data and carried out some inquiries:

1. Displayed the names of the unique launch sites in the space mission
2. Displayed 5 records where launch sites begin with the string 'CCA'
3. Displayed the total payload mass carried by boosters launched by NASA (CRS)
4. Displayed average payload mass carried by booster version F9 v1.1
5. Listed the date when the first successful landing outcome in ground pad was achieved.
6. Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

7. Listed the total number of successful and failure mission outcomes
8. Listed the names of the booster versions which have carried the maximum payload mass
9. Listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. Ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20

https://github.com/ednes76/Capstone-project/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and we could discover some of the factors by analysing the existing launch site locations.

We marked all the launch locations and then we marked the success/failed launces for each site, then we calculated the proximity of the sites to certain geographical features for example the coast or transport systems.

https://github.com/ednes76/Capstone-project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

To get an even cleared picture of the launch sites and the success/failure of each one we build an interactive dashboard using Plotly.

The aim is to find the answers to the following:

- Which site has the largest successful launches?
- Which site has the highest launch success rate?
- Which payload range(s) has the highest launch success rate?
- Which payload range(s) has the lowest launch success rate?
- Which F9 Booster version has the highest launch success rate?

https://github.com/ednes76/Capstone-project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

For the predictive analysis of the success/failure of launches we employed four classification models:

1. Logistic regression
2. Support Vector Machine
3. Decision Tree
4. KNN

We created the classifier object and then using GridSearchCV we looked for the best parameters. Finally we calculated the accuracy and plated a confusion matrix in order to find out the false positives and negatives.

https://github.com/ednes76/Capstone-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

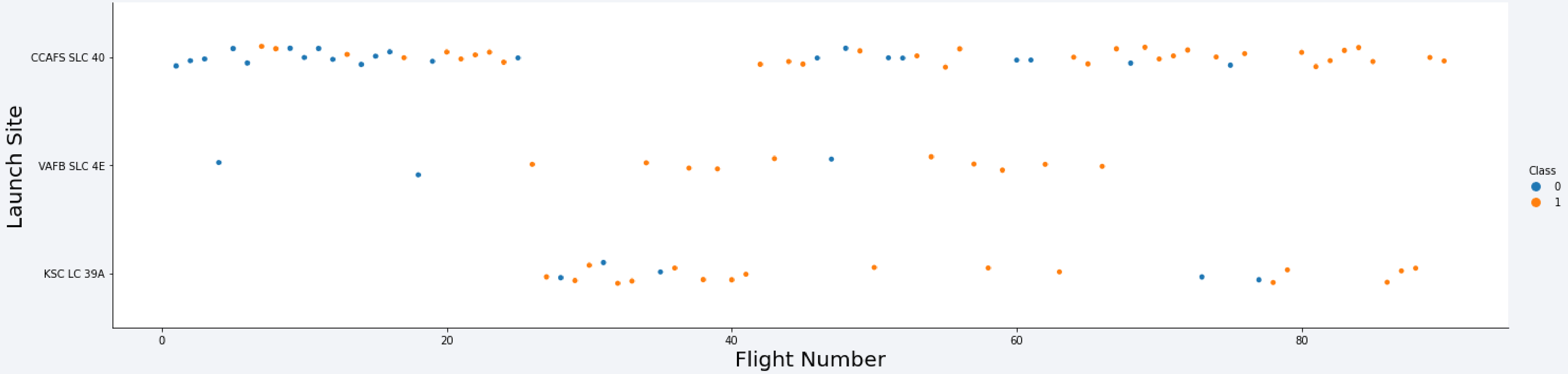
Results

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

Section 2

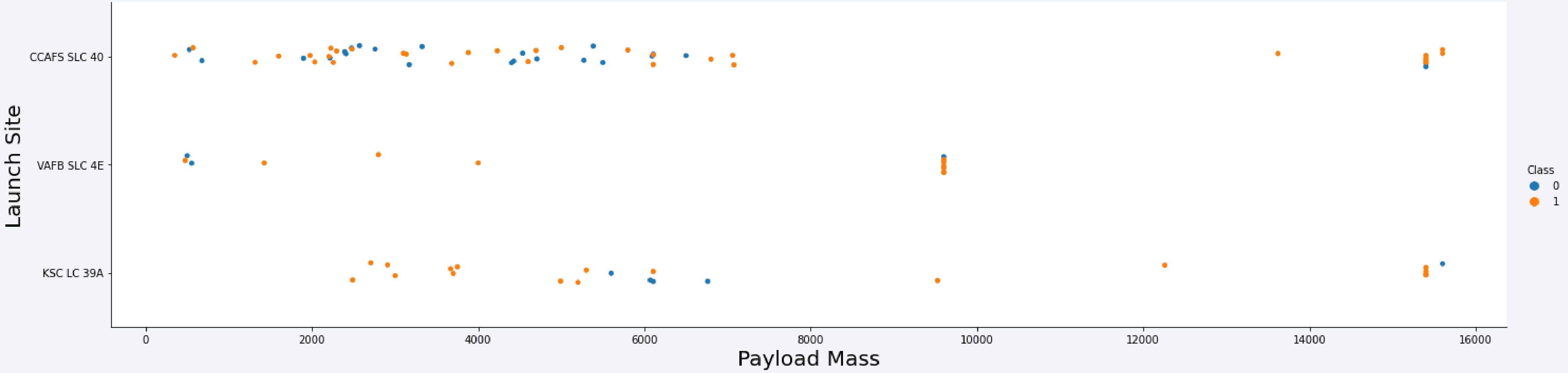
Insights drawn from EDA

Flight Number vs. Launch Site



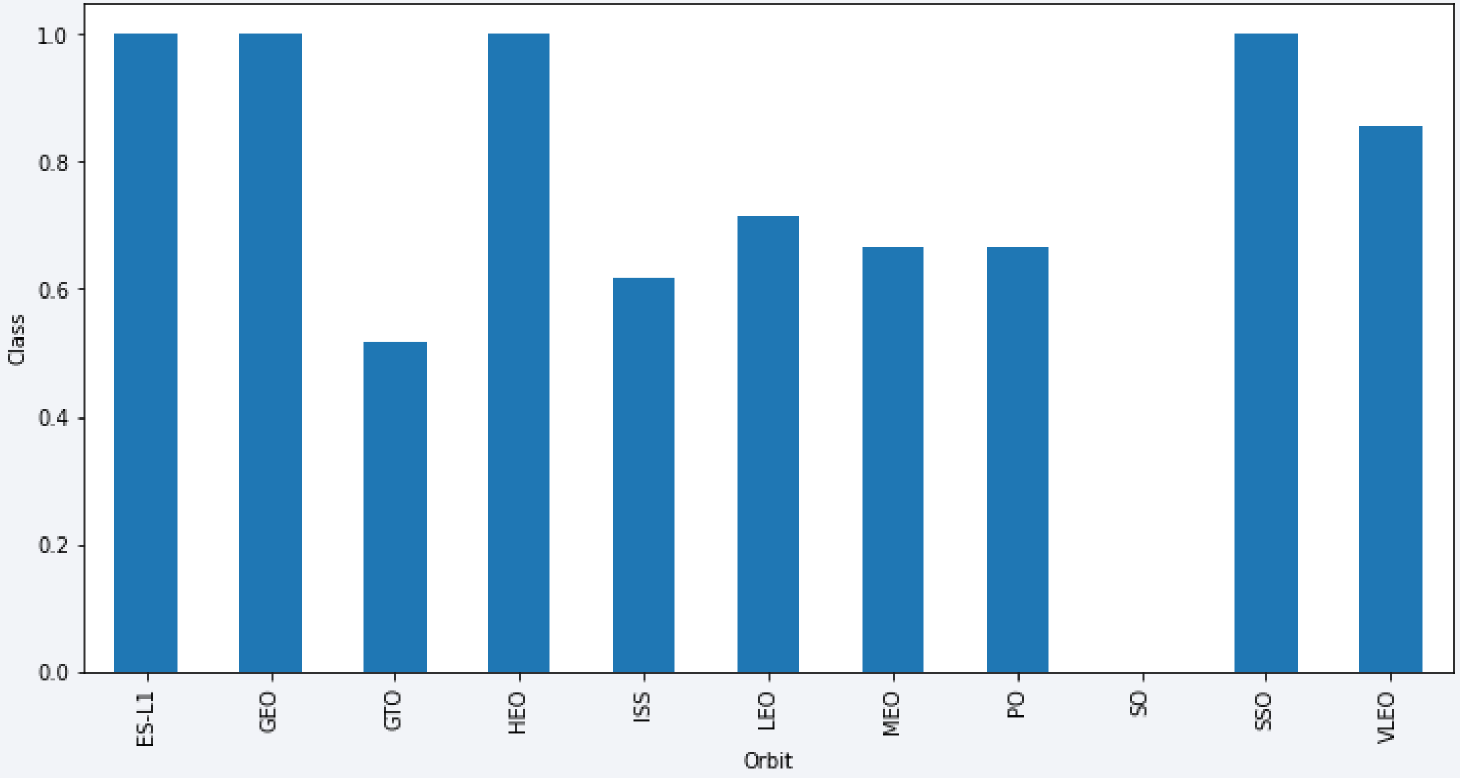
There seems to be a correlation between the number of flights and the success rate of the landings.

Payload vs. Launch Site



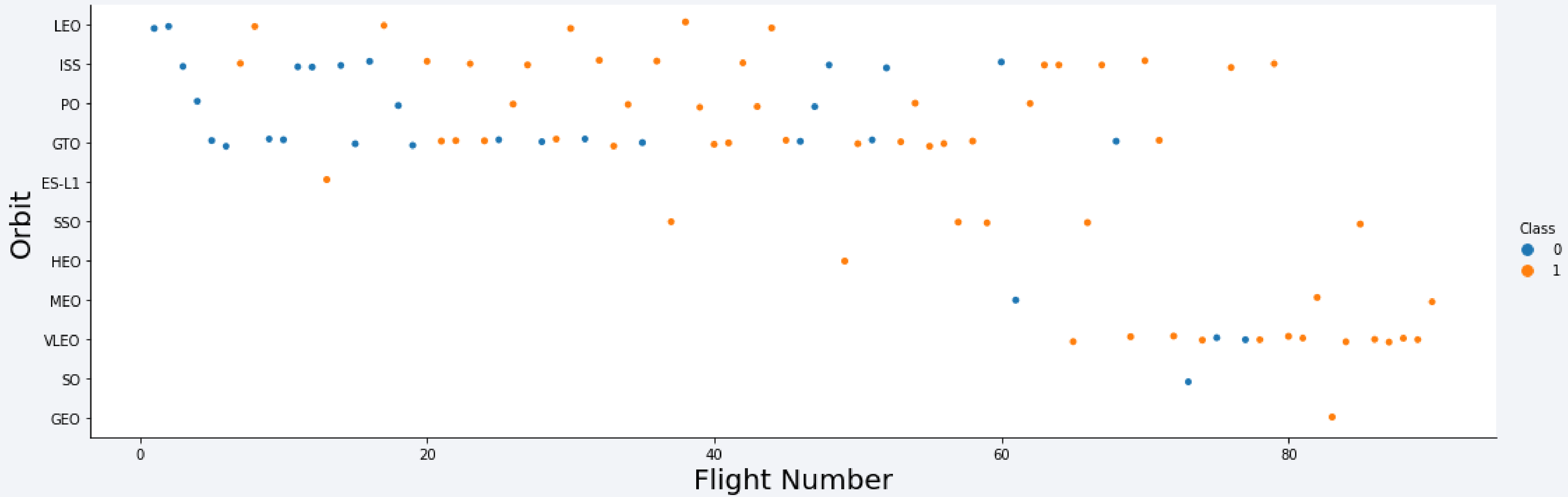
There seem to be a correlation between the payload mass and the success rate of the landings especially for VAFB SLC 4E and KSC LC-39A

Success Rate vs. Orbit Type



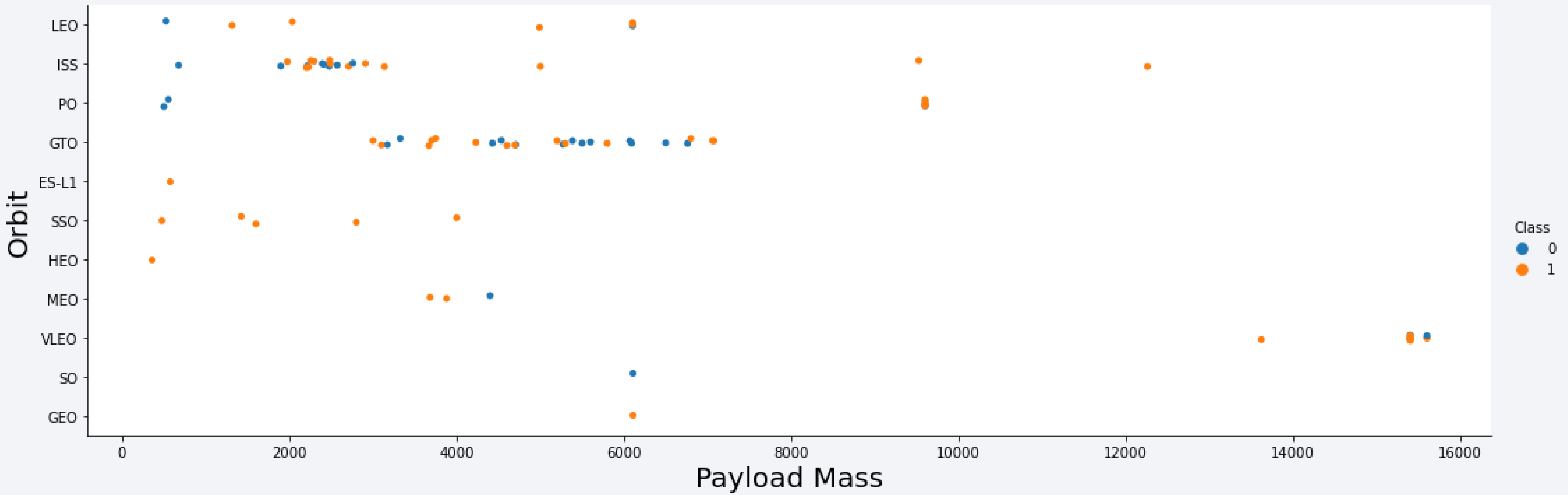
It seems obvious there are certain orbits that have a greater success rate that others
With GEO, HEO, SSO and ES-L1 having 100%

Flight Number vs. Orbit Type



Again there seems to be a connection between the number of flights and the success rate for each orbit except SSO, MEO, ES-L1 and HEO

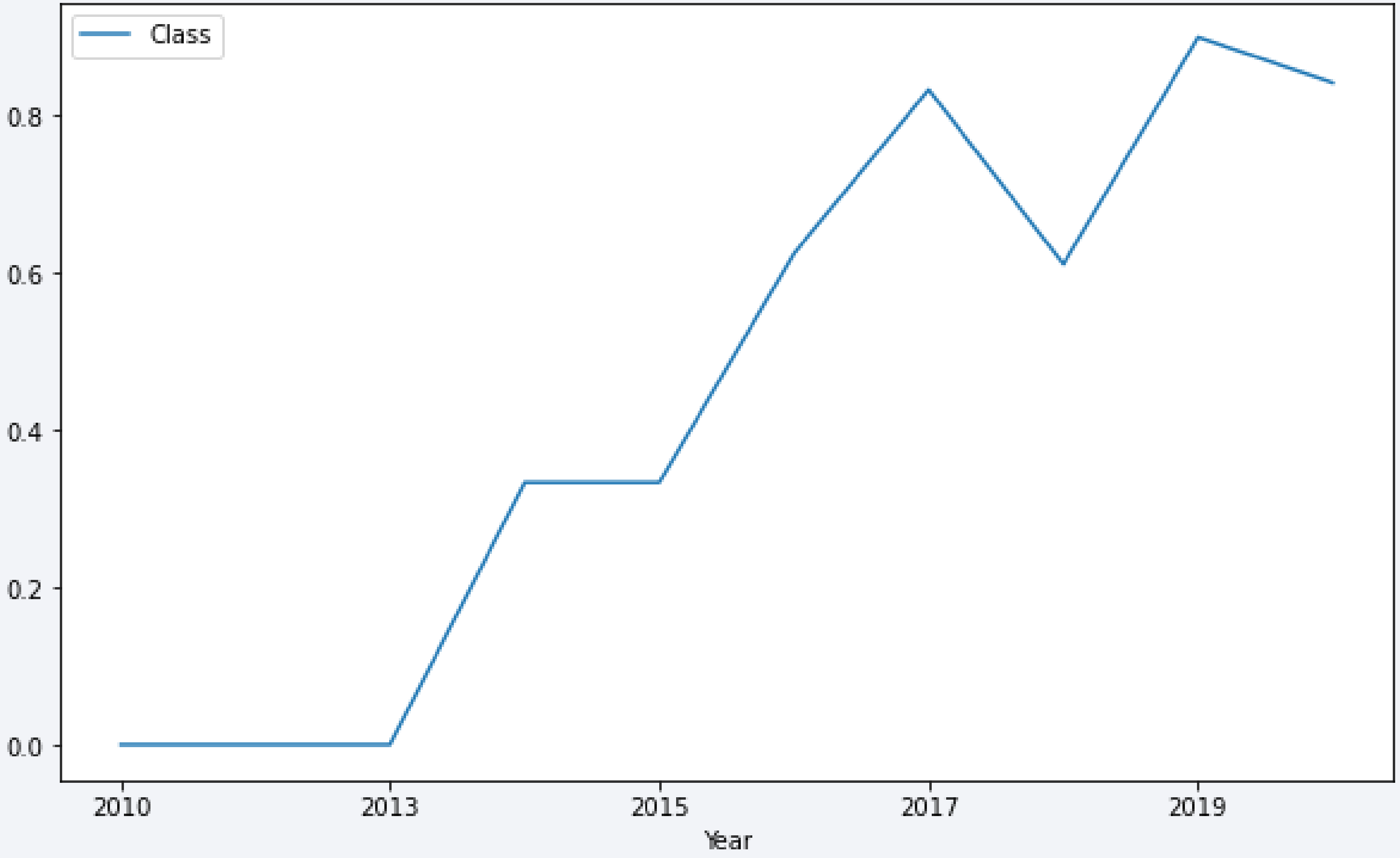
Payload vs. Orbit Type



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

We can observe an increase in success rate over the years



All Launch Site Names

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

There are four launch sites, one in California (VAFB SLC-4E) and three in Florida (CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A)

Launch Site Names Begin with 'CCA'

Out[39]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

In [41]:
```
%%sql
select sum(payload_mass__kg_) from SPACEXTBL
where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://jdj33138:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[41]:

1
45596

Total Payload Mass from NASA is 45596

Average Payload Mass by F9 v1.1

The average Payload Mass for the F9 v 1.1 booster is 2928

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

In [42]:
```
%%sql
select avg(payload_mass__kg_) from SPACEXTBL
where booster_version like 'F9 v1.1'
```

```
 * ibm_db_sa://jdj33138:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[42]:

1
2928

First Successful Ground Landing Date

```
In [43]: %%sql
         select date from SPACEXTBL
         where landing__outcome like 'Success (ground pad)'
         order by date asc
         limit 1

 * ibm_db_sa://jdj33138:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[43]:

DATE
2015-12-22

First successful landing date is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In [44]:

```
%%sql
select booster_version from SPACEXTBL
where landing__outcome like 'Success (drone ship)' and (payload_mass__kg_>4000 and payload_mass__kg_<6000)
```

```
 * ibm_db_sa://jdj33138:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[44]:

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

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

Total Number of Successful and Failure Mission Outcomes

In [45]:

```
%%sql
select mission_outcome, count(*) from SPACEXTBL
group by mission_outcome
```

```
 * ibm_db_sa://jdj33138:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[45]:

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

Number of successful and failed mission outcomes

Boosters Carried Maximum Payload

Out [46]:

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

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

2015 Launch Records

Out [47]:

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

List of the failed landing 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

Out[48]:

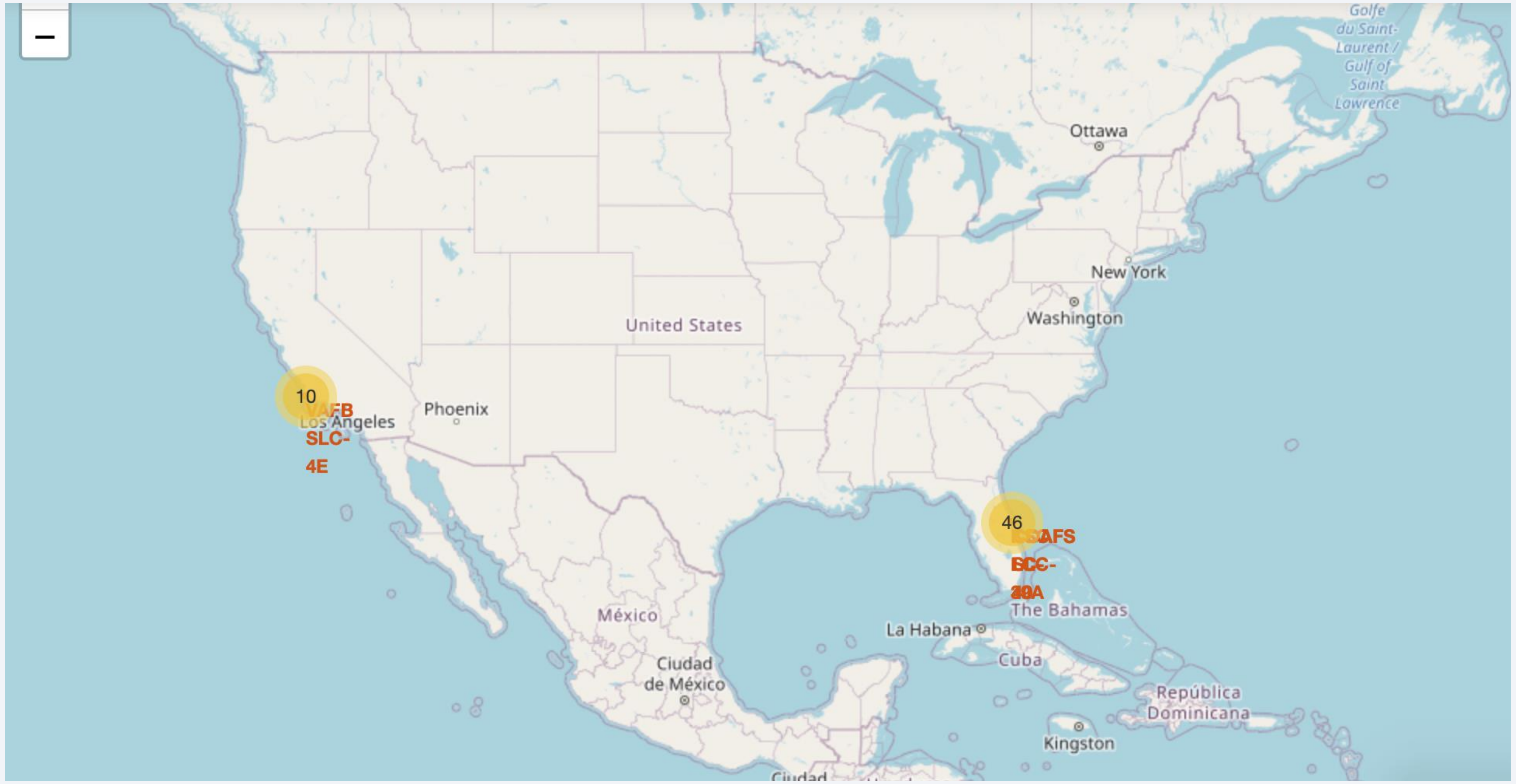
landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Ranking 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

Section 3

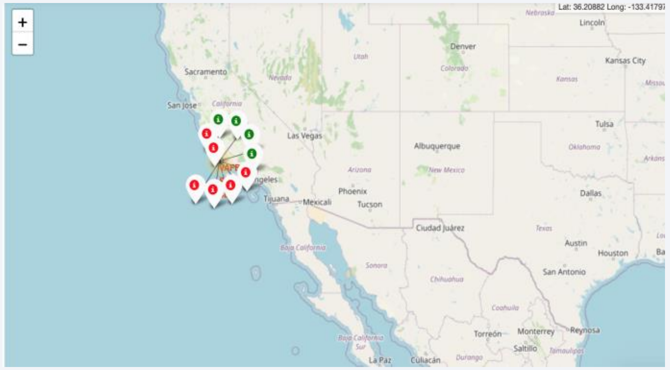
Launch Sites Proximities Analysis

General map of launch sites

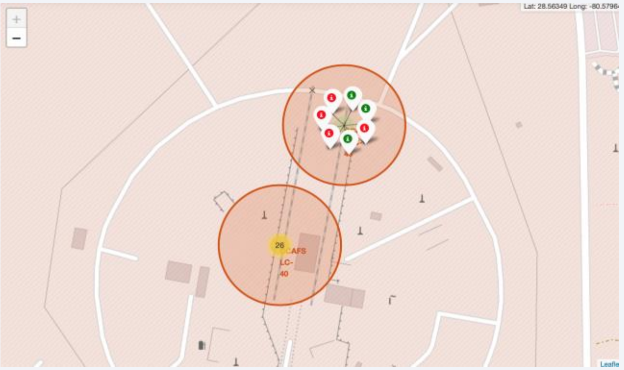


All launch sites are close to the coastline at approximately same latitude

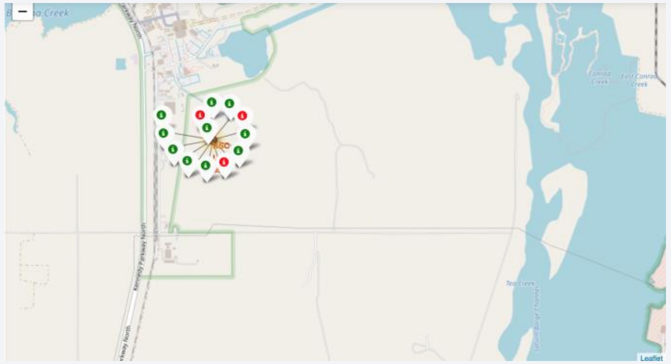
Booster successful landing for each launch site



VAFB SLC-4E



CCAFS SLC-40

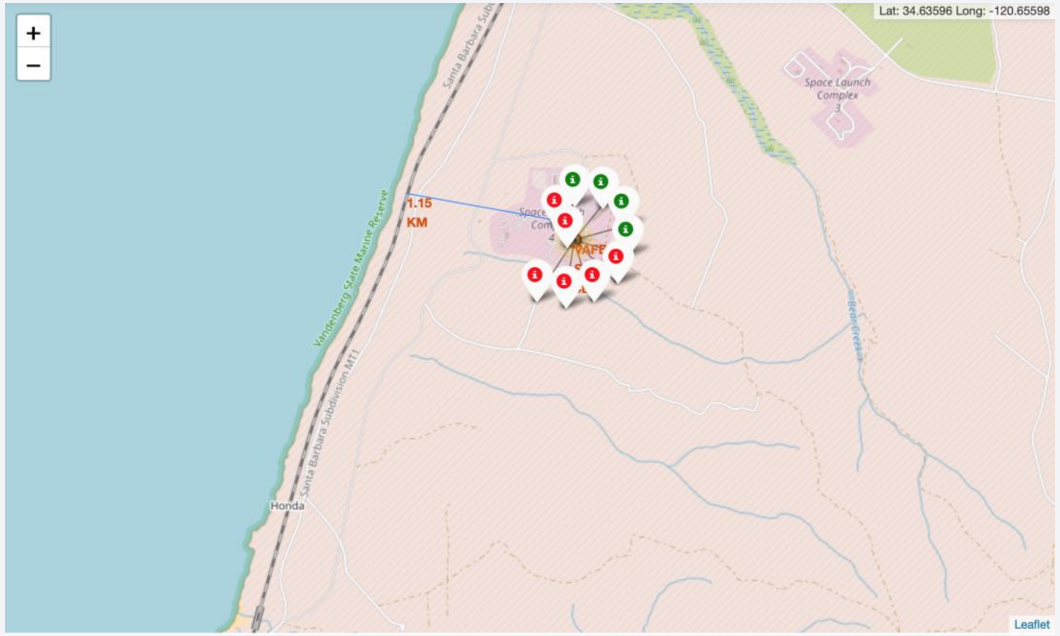


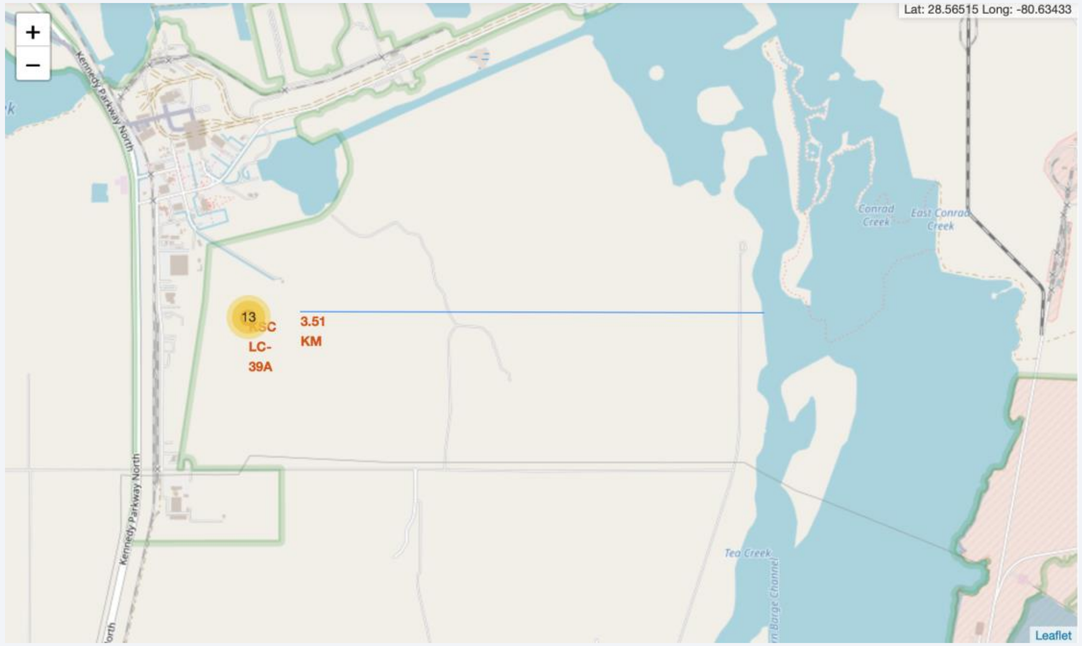
KSC LC-39A

CCAFS LC-40

We can see that the most successful site is KSC LC-39A and we explore further why

Distance from coastline



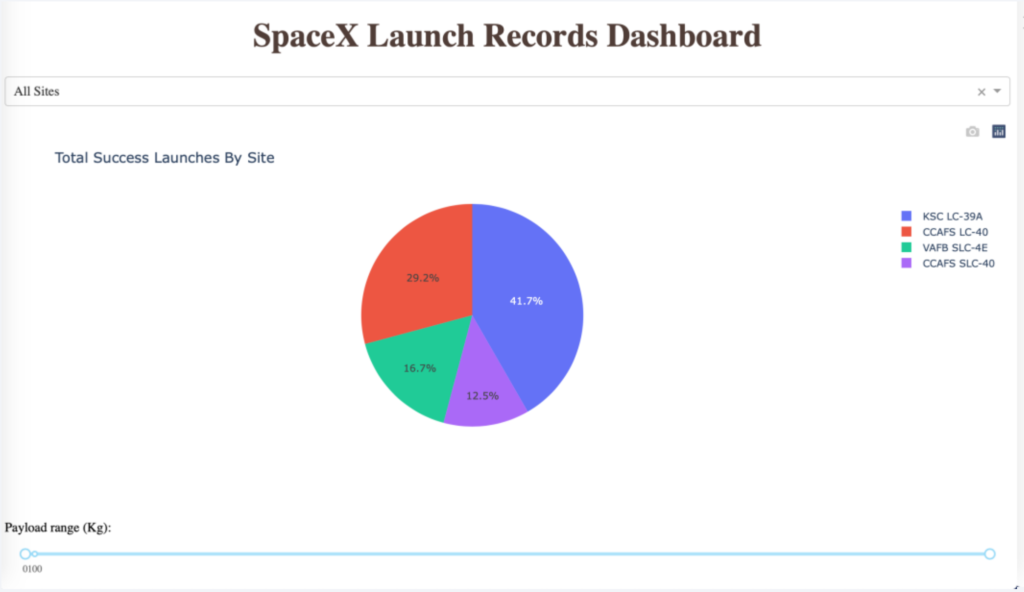


We can observe that the most successful launch site (KSC LC-39A) is further inland than the other sites, 3.51km vs 1.15km.

Section 4

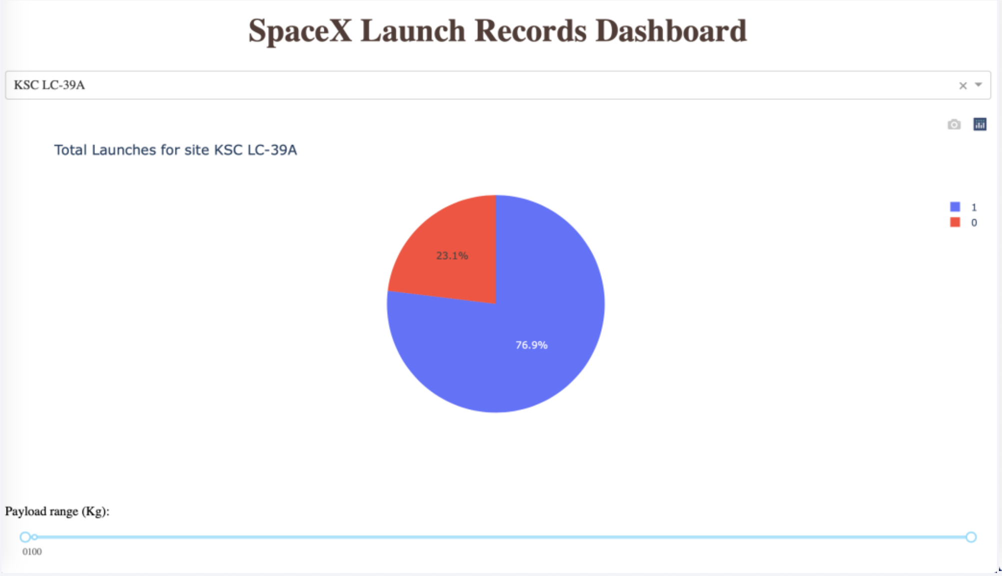
Build a Dashboard with Plotly Dash

Success count for all sites



Looking at the payload vs booster version we seen that booster version FT has more successful landings while version v1.1 more unsuccessful landings

KSC LC-39A has the highest success rate



KSC LC-39A is the most successful site with 76.9% success rate. Looking at the payload vs booster we see that Payloads under 5500 kg were all successful and over 5500 kg failures were with booster version FT. It is worth further investigation into payload vs booster version

Payload and Launch Outcome

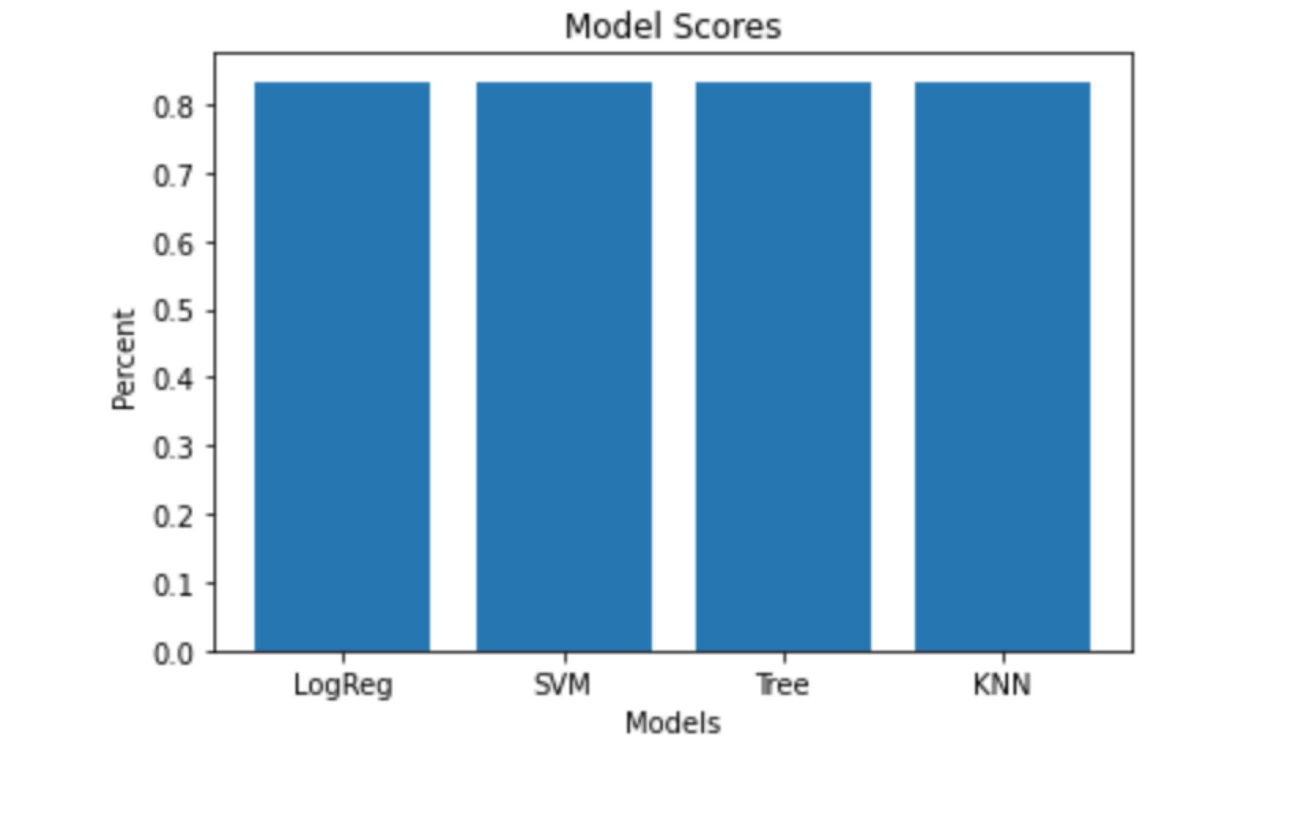
Payloads under 6000kg have greater success rate that those over 6000kg where only one was a success. For payloads under 6000kg the most successful booster version was FT while least successful was v.1.1.

The emerging picture is that the combination of booster version FT and payloads under 6000kg has the biggest success rate.

Section 5

Predictive Analysis (Classification)

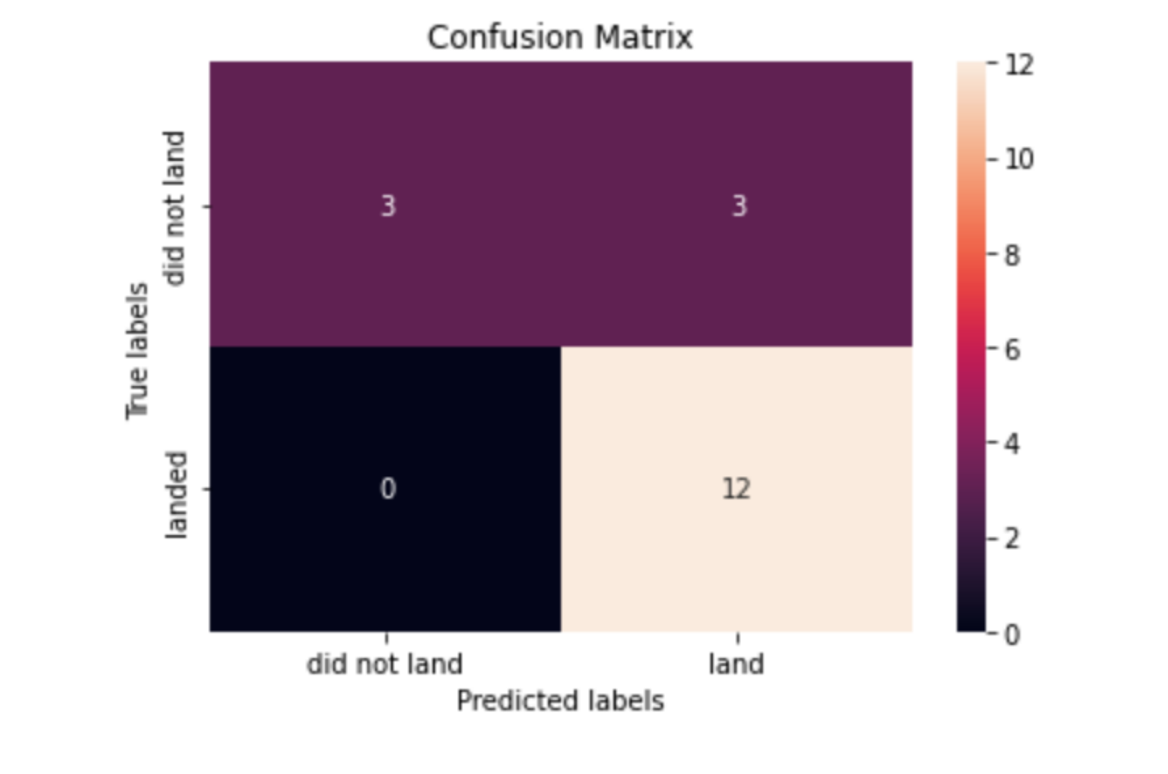
Classification Accuracy



After building and testing all models for accuracy we come to the conclusion that all have the same accuracy of 0.8333333333333334.

Confusion Matrix

- Show the confusion matrix of the best performing model with a explanation



The confusion matrix for all Classification models shows 3 false positives which is quite a lot relative to such a small dataset, but in there lies the problem with this attempt at predicting a successful landing.

Conclusions

After carrying out the Exploratory Data Analysis and Predictive Analysis some conclusions can be drawn:

1. The most successful launch site was the one further from the coastline (KSC LC-39A). This may raise the issue of high coastal winds having an impact on the booster landings.
2. Payloads under 6000kg had higher booster landing success rate when combined with booster version FT.
3. There are certain orbits that have a greater success rate that others with GEO, HEO, SSO and ES-L1 having 100%
4. The trend for the success rate show a positive growth year after year which shows the value of learning from experience.
5. Predicting a successful booster landing is proving challenging at this point due to the small dataset. All models showed the same accuracy and false positive rates.

Appendix

All notebooks, apps and dataset can be found at the link below.

https://github.com/ednes76/Capstone-project

Thank you!