IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andy Huang
23 Feb 22

Outline

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

Executive Summary

- Summary of methodologies
  - Collection of Falcon 9 launch data via SpaceX API, and web scrapping on wiki, then store data into IBM Db2 for analytics.
  - Data cleaning, replace 0 or missing values with either mean.
  - Visualization of the data
  - Performance of a predictive analysis to train and evaluate a best model and give prediction over the success of future landings.
- Summary of all results
  - When Payload was greater then 7500 kg falcon rocket had a higher chance of successful landing.
  - Among 11 orbit types ES L1, GEO, HEO, SSO were 100 successful with less than 6000 kg payload.
  - SpaceX has 4 launch sites, one is near California, the other three is near Florida and South Texas. All the sites are in near proximity to ocean and all the sites are bit far away from the city.
  - All models performed similarly when trained with the data at hand.

Introduction

- Project background and context
  - SpaceX is a company that aims to make commercial space travel more affordable for everyone.
  - This company can launch rockets for a cost of around 60 million dollars. In contrast, other providers require165 million dollars for one launch. This is due to the fact that SpaceX can reuse the first stage of the rocket Falcon9.
  - The primary cost saving agent is the high success rate of stage 1 landing and thus its reusability in future launches.
  - The challenge here is to set a right costing fore cast of the rocket launches through predicting its potential to land stage 1 successfully.
- Problems you want to find answers
  - What are features that contributes the most to predict whether the stage one of the rocket will and land successfully?
  - Can we predict if we new launches will be successful based on our trained model. What will be the accuracy of our predictions?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX REST API and Wikipedia Web scrapping using python BeautifulSoup.
- Perform data wrangling
  - Data was processed using python pandas and numpy library.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models (Experiment usability and compatibility of SVM, Tree maps, KNN, Logistic
  - Regression optimizing parameters) were built, evaluated and tuned using sklearn.

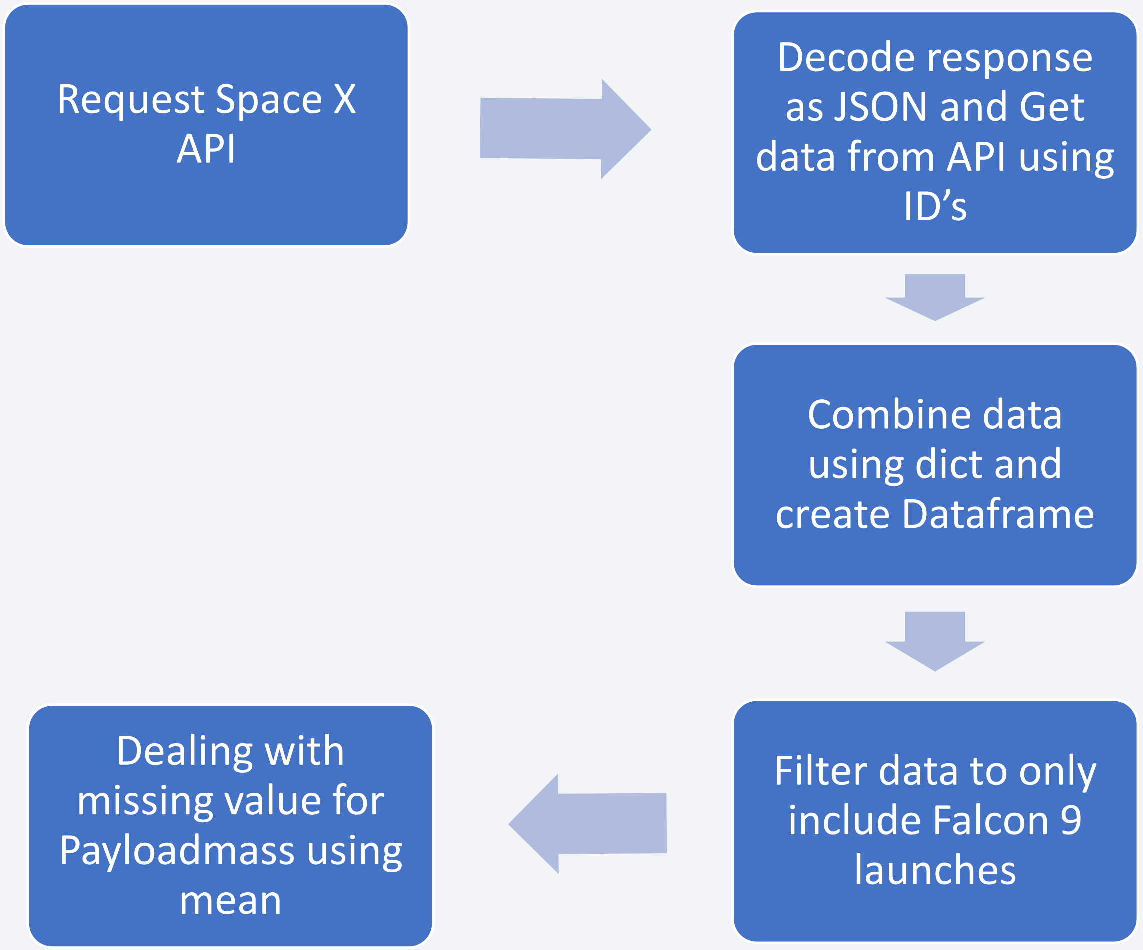
Data Collection

- Describe how data sets were collected.
    - SpaceX API.
    - BeautifulSoup library to scrap data from the Wikipedia page.

Data Collection – SpaceX API

- Request launch data from SpaceX URL using given API.
- Extract the data from the response.
- Pre-process and construct the data.
- Store the data in CSV file.

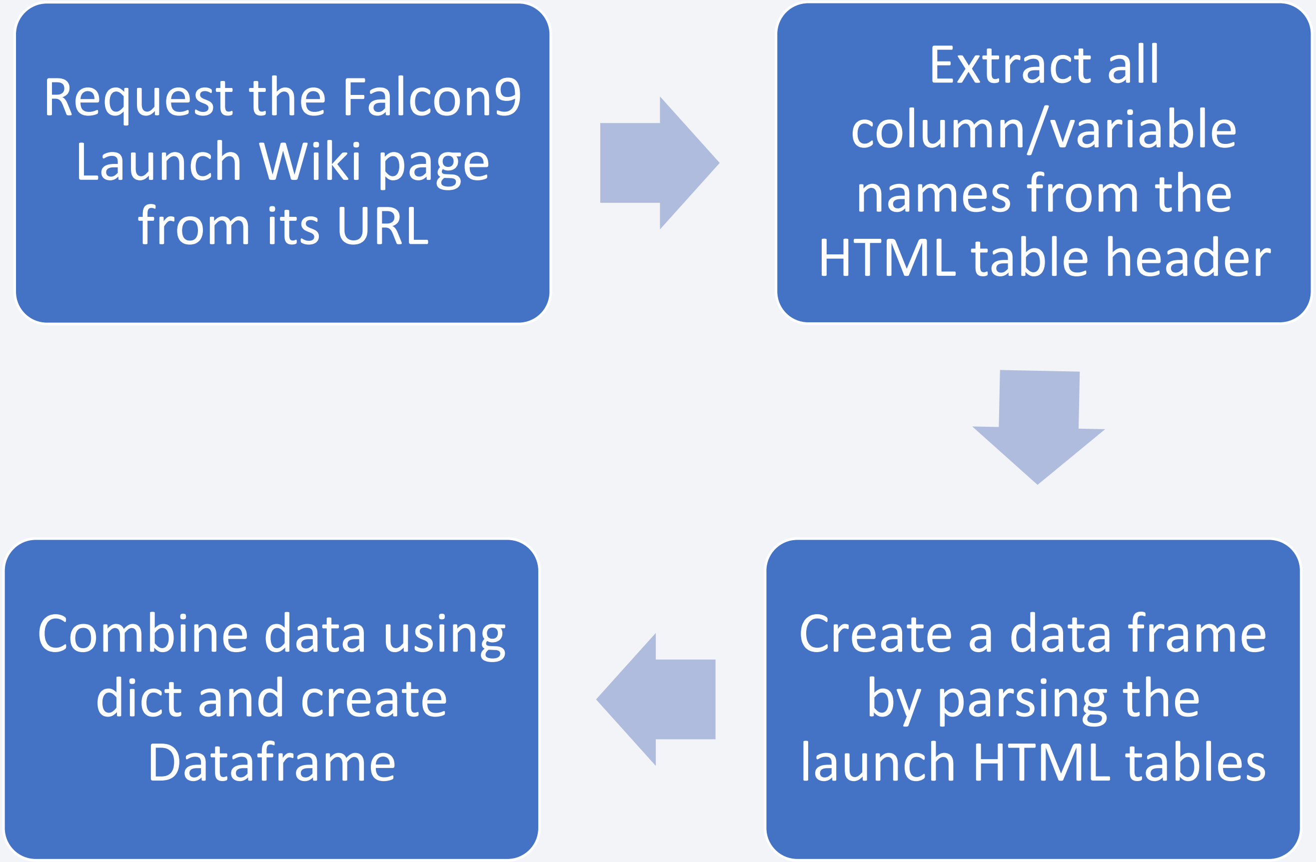
[GitHub]



Data Collection - Scraping

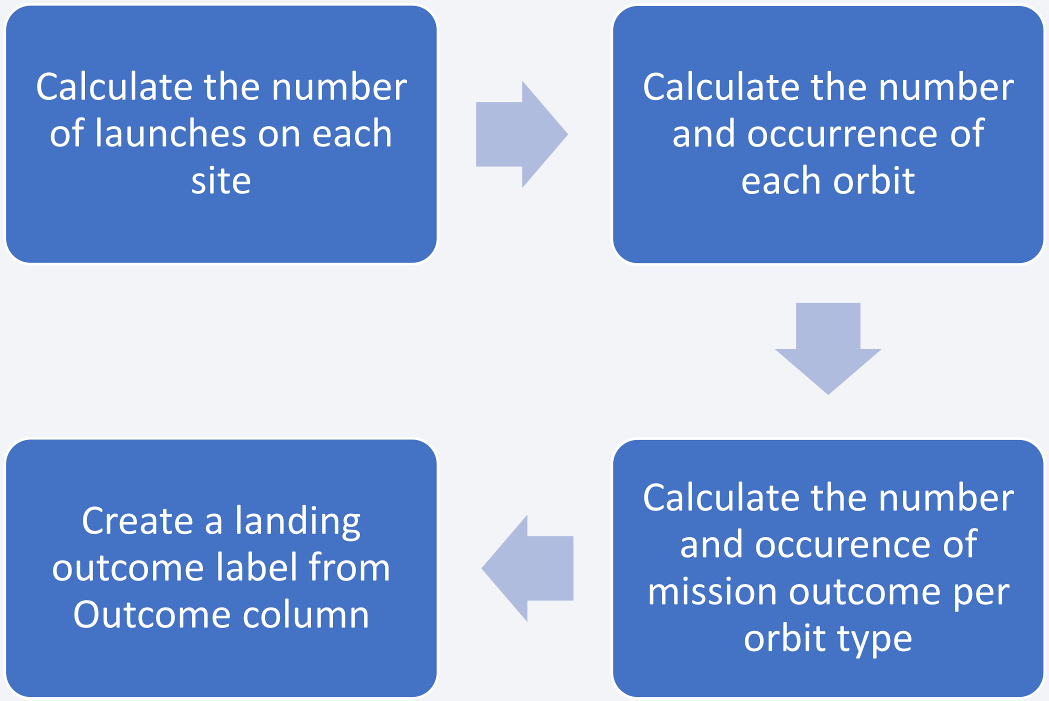
- Request for the Wikipedia page.
- Parse the table data from html text using beautifulSoup4 library.
- Create pandas data frame from table data.
- Construct the data and store it in CSV.

[GitHub]



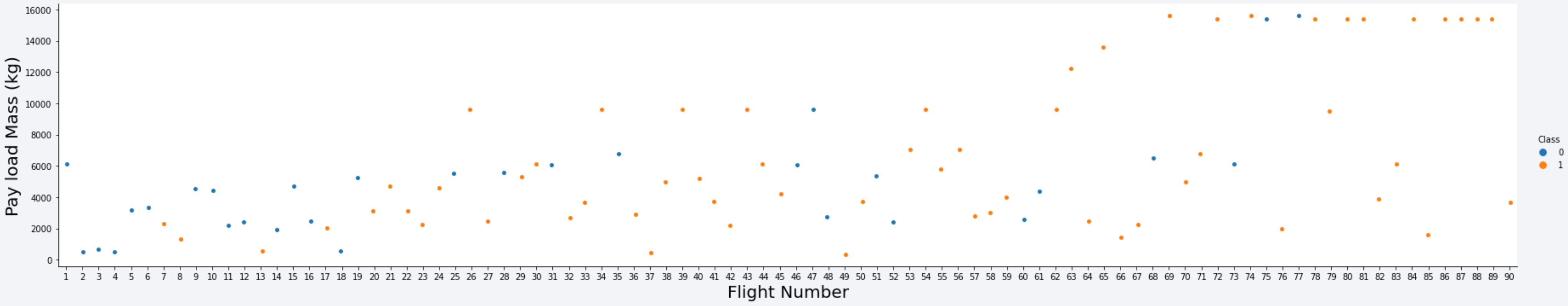
Data Wrangling

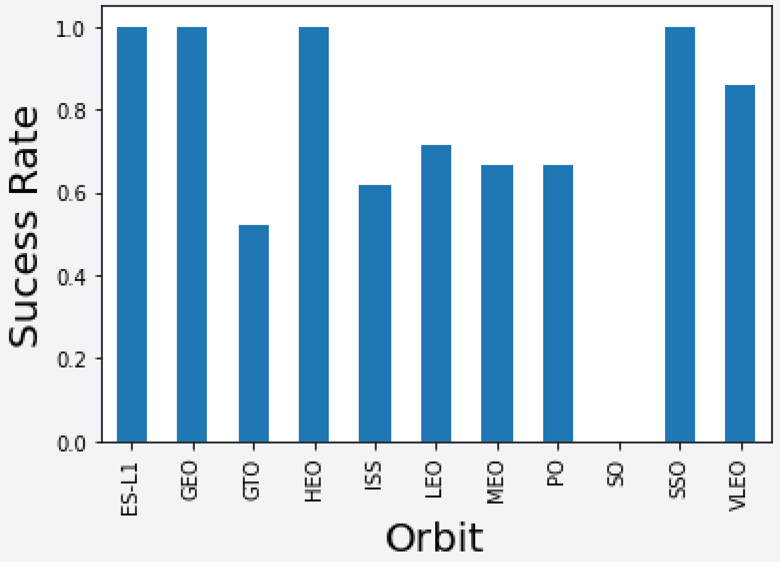
- Calculate the number of launches per site.
- Number of occurrence of each orbit.
- Number of occurrences of outcome per orbit.
- Create landing outcome label.

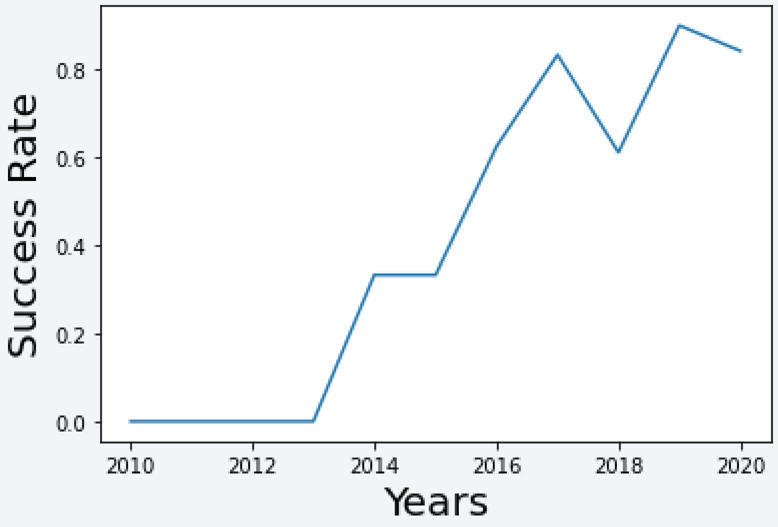


[GitHub]

EDA with Data Visualization







GitHub

EDA with SQL

- Display the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string ‘CCA’ .
- Display the total payload mass carried by boosters launched by NASA.
- Display average payload mass carried by booster version ‘F9 v1.1’.
- List the date when the first successful landing outcome in ground pad was achieved.
- 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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010 06 04 and 2017 03 20 , in descending order.

GitHub

Build an Interactive Map with Folium

- Map objects which are created and added to the folium map are given below:
  - Markers: Added to mark a specific area with a text label on a specific coordinate.
  - Circles: Added to highlight circle areas with a text label on a specific coordinate.
  - Marker Cluster: Marker clusters were used to simplify the containing many markers having the same coordinates.
  - Mouse Position: Used to get coordinate for a mouse over a point on the map (proximities). It helps to find the coordinates easily of any points of interests while exploring the map.
  - Polyline: It draws polyline overlays on a map. It was used to denote the distance between a launch site and its proximities(such as Railway station, city, etc.).

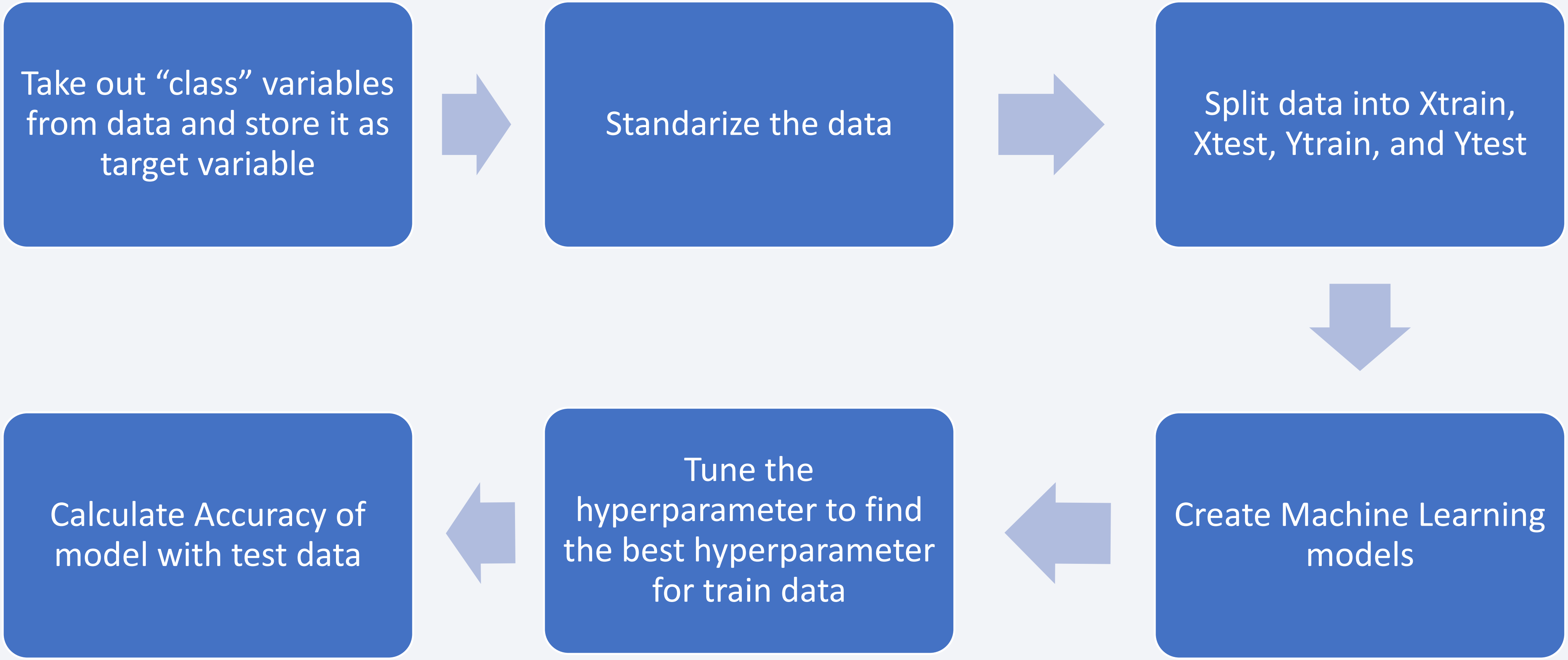
[GitHub]

Build a Dashboard with Plotly Dash

- Pie chart (total launches for a selected site or the total sites collection)
  - Shows relative proportions of different sites successful landing distribution.
  - Shows % of success vs. failure for a given site.
- Scatter Plot
  - Showing the correlation between Outcome and Payload Mass(Kg) for different Booster Versions with freedom of selection of the range of payload mass of Interest.

GitHub

Predictive Analysis (Classification)

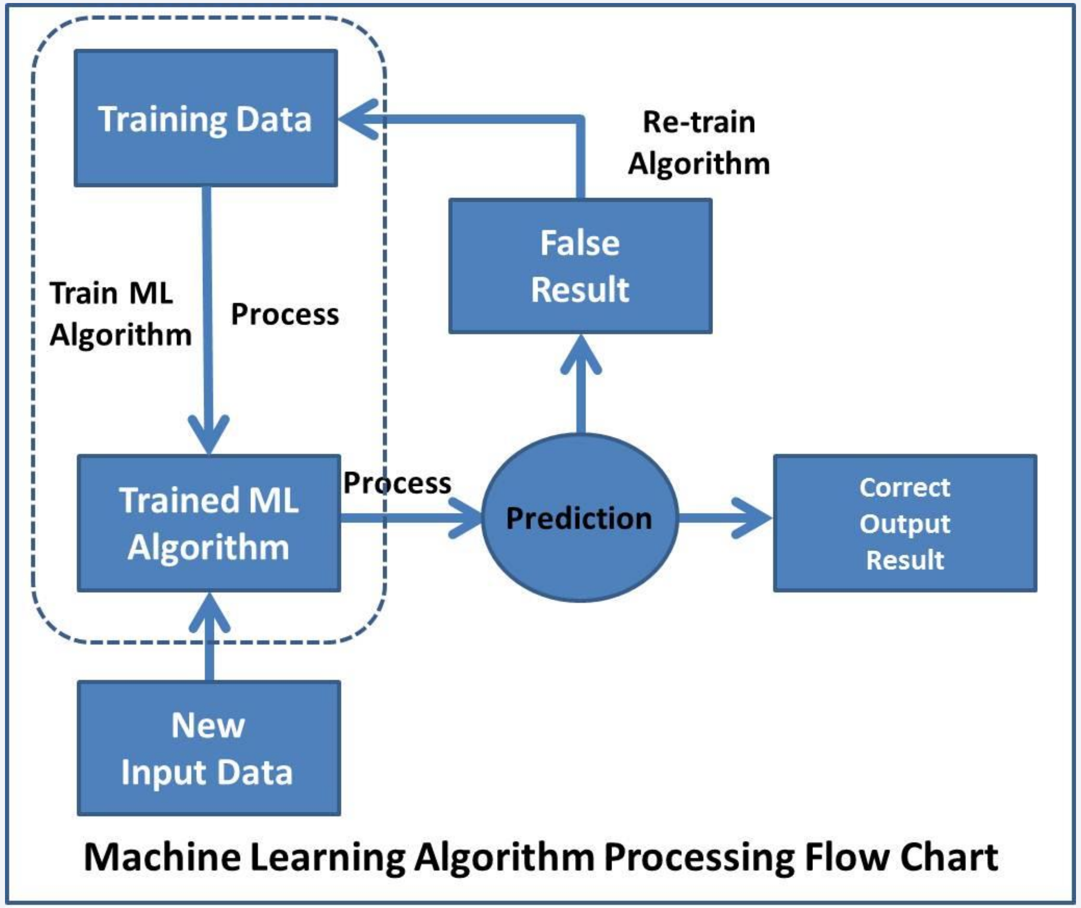


GitHub

Results

- Models were built using Scikit-Learn, data were previously normalized and models hyper parameters were found using a Grid Search with a 10 fold cross validation, in the end the best performing model has been selected based on accuracy.

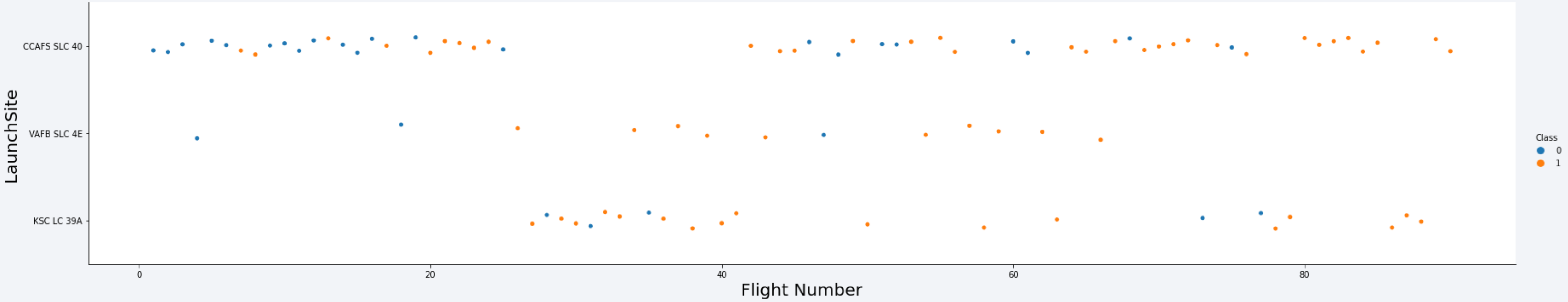
[GitHub](#)



Section 2

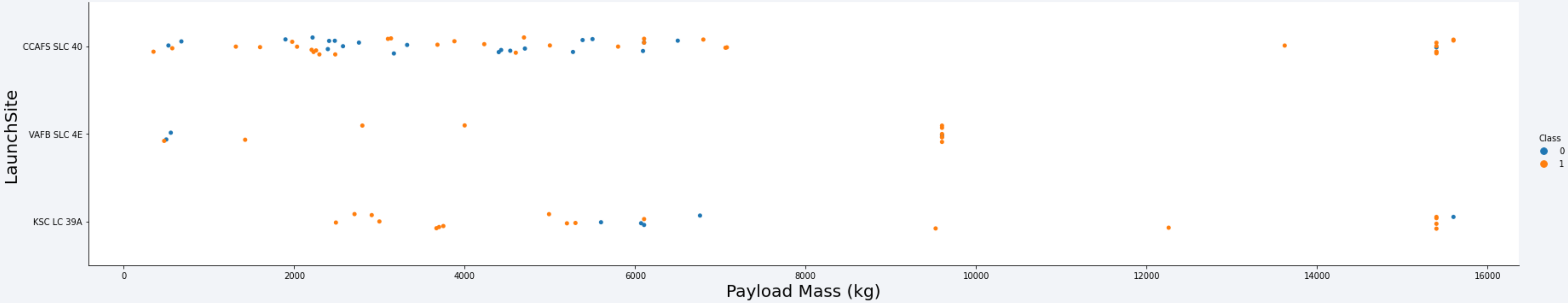
Insights drawn from EDA

Flight Number vs. Launch Site



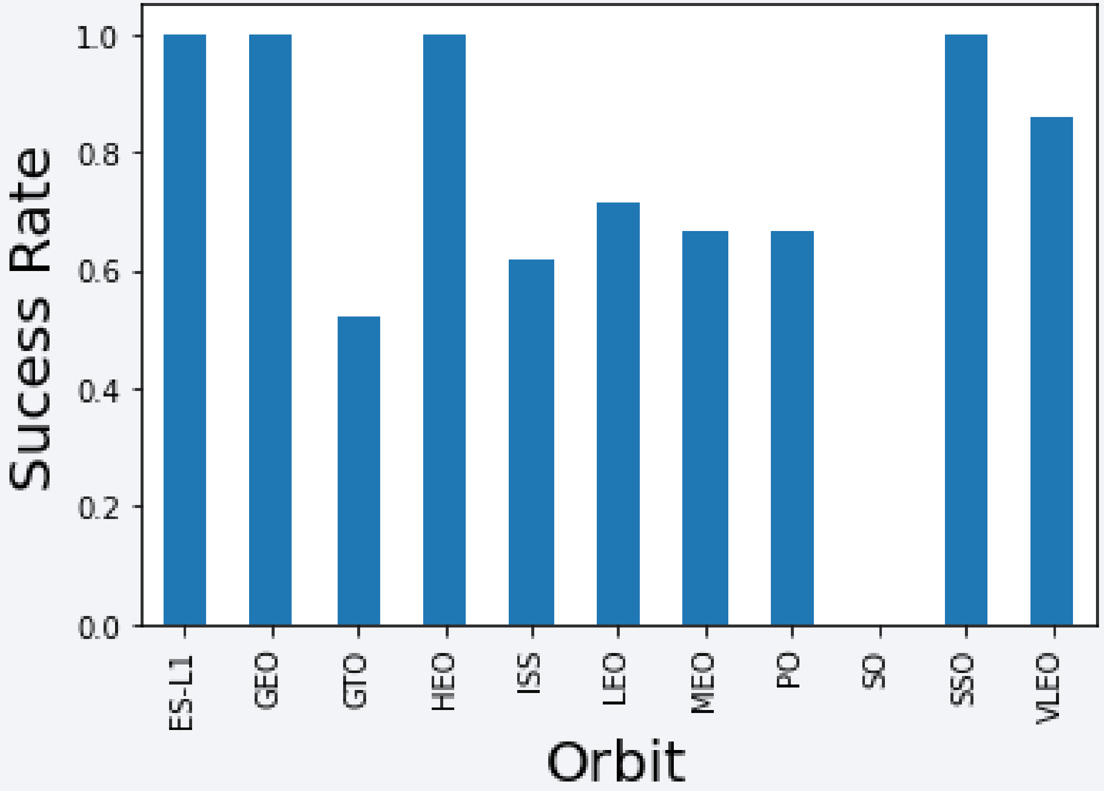
This chart seems to indicate that a “young” launching site will probably a lower success rate than one which had a lot of rocket launched from.

Payload vs. Launch Site



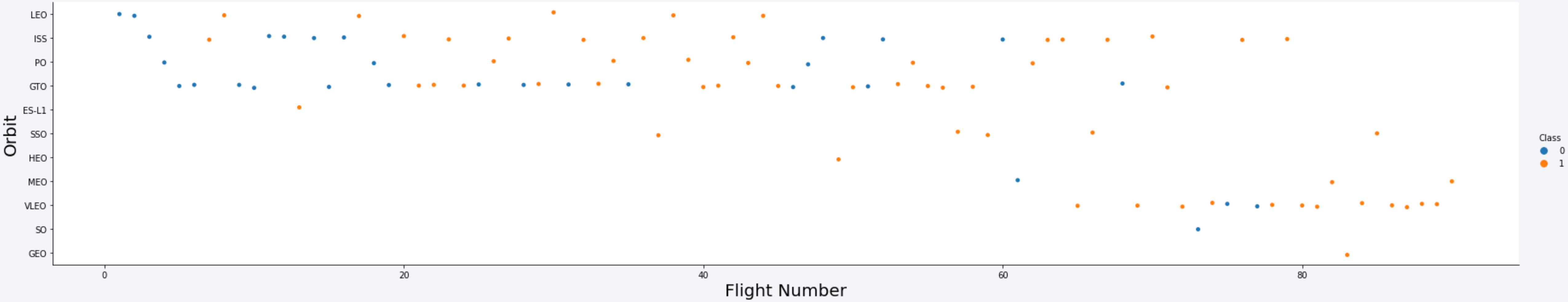
It seems like a lot of rocket launched had a payload between 500kg and 6000kg. Also, the launching site VAFB SLC 4E seems to be a site where there are not that much rocket launched. An impact of the payload could be possible but it will need further analysis.

Success Rate vs. Orbit Type



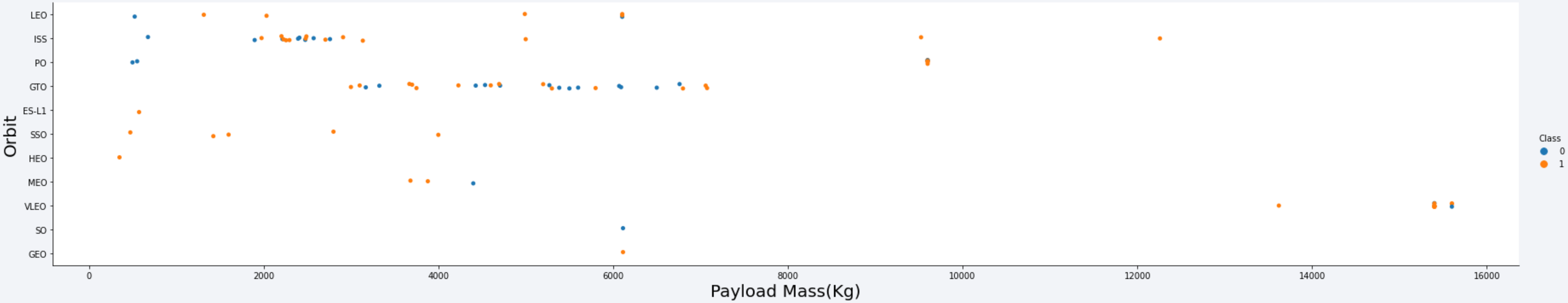
There is a strong correlation between these two indeed as we can observe the SO or GTO Orbit type are quite risky as the success rate is below 0.6. However, some orbit type provide a 1.0 success rate which is perfect but can hide suspicious data. Indeed if for this orbit type only one rocket has been launched the reliability of this hypothesis is null.

Flight Number vs. Orbit Type



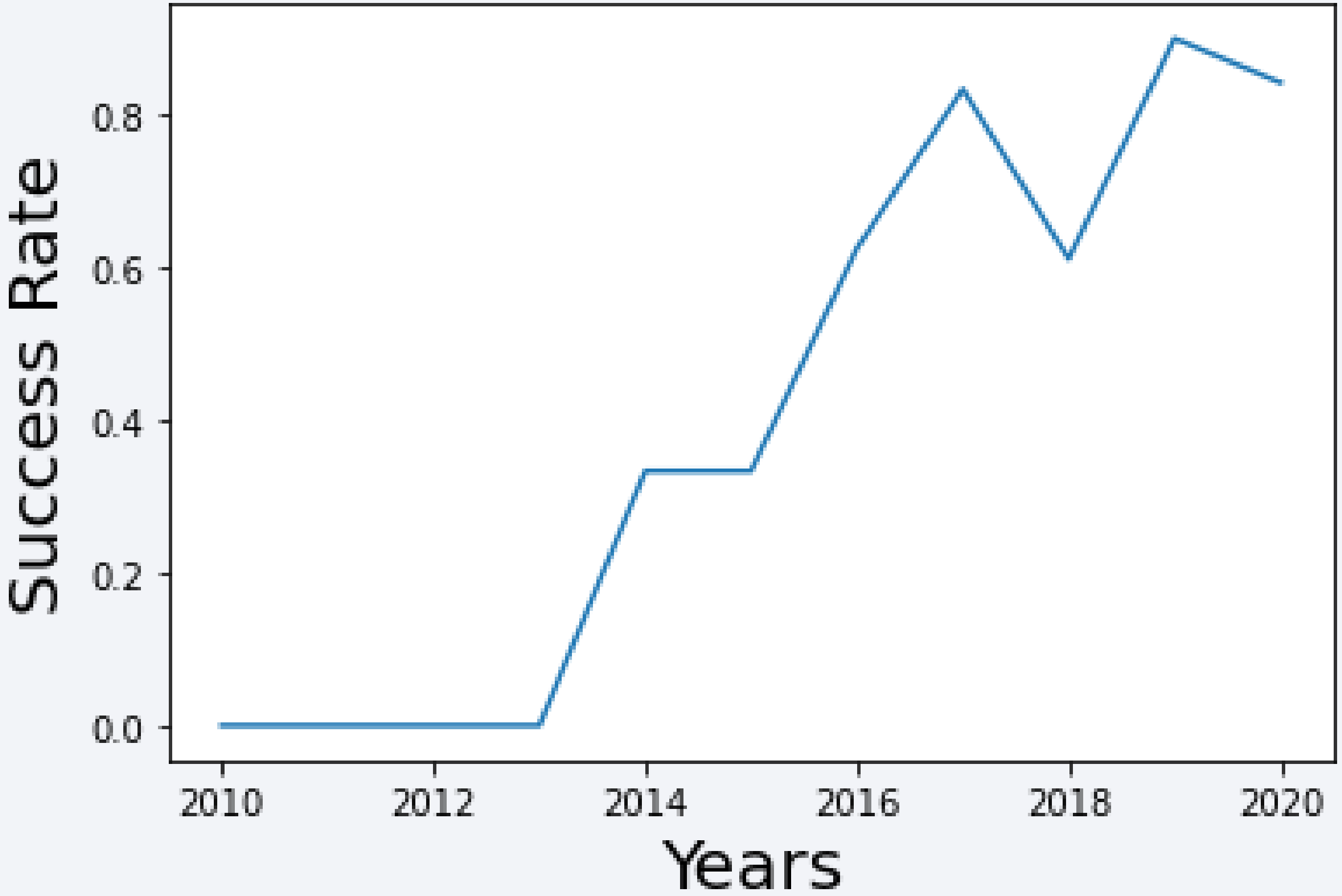
This chart confirmed what has been said before, some Orbit type have only couples of Flights in their history and thus make data quite confusing. However for the GTO,VLEO and ISS it seems like there are enough data to be confident on those data.

Payload vs. Orbit Type



Here we can observe that certain sites have a strong relation with the payload mass, for example the GTO and ISS.

Launch Success Yearly Trend



Here the chart demonstrates that as Humans learn more and more through the years thanks of Sciences, it results in a significant rocket launches success rate increasing.

All Launch Site Names

- launch sites
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Query : select distinct(launch_site) from SPACEXTBL;

Launch Site Names Begin with 'CCA'

In [4]:

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

```
* ibm_db_sa://hhb69393:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

Out[4]:

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

select * from SPACEXTBL where launch_site like 'CCA%' limit 5;

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - The total payload carried by boosters from NASA is 45596 kg.
- Query: select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';

```
In [5]: %sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';

 * ibm_db_sa://hhb69393:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.

Out[5]:
```

1
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - The average payload mass carried by booster version F9 v1.1 is 2534 kg.
- Query:
  select avg(payload_mass__kg_) from SPACEXTBL where booster_version like 'F9 v1.1%';

In [6]:
```
%sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version like 'F9 v1.1%';
```

```
 * ibm_db_sa://hhb69393:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

Out[6]:

1
2534

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - The dates of the first successful landing outcome on ground pad was 2015-12-22.
- Query:
  select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)';

```
In [7]: %sql select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://hhb69393:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

Out[7]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
  - F9 FT B1032.1
  - F9 B4 B1040.1
  - F9 B4 B1043.1
- Query: select distinct(booster_version) from SPACEXTBL where landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000;

In [8]:
```
%sql select distinct(booster_version) from SPACEXTBL where landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000;
```

```
 * ibm_db_sa://hhb69393:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

Out[8]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Query: select mission_outcome,count(*) from SPACEXTBL group by mission_outcome;

In [9]:
```
%sql select  mission_outcome, count(*) as Counter from SPACEXTBL group by mission_outcome;
```

```
 * ibm_db_sa://hhb69393:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

Out[9]:

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Query: select distinct(booster_version) from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Query: select landing__outcome,booster_version,launch_site,date from SPACEXTBL where Date like '2015%' and landing__outcome = 'Failure (drone ship)';

```
In [11]: %sql select  landing__outcome, booster_version, launch_site, date from SPACEXTBL where Date like '2015%' and landing__outcome = 'Failure (drone ship)';

 * ibm_db_sa://hhb69393:***@b1bc1829-6f45-4cd4-bef4-10cf081900bf.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32304/bludb
Done.
```

Out[11]:

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

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

- 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

- Query: select landing__outcome,count(landing__outcome) as count from SPACEXTBL where DATE BETWEEN '2010-06-04' and '2017-03-20' GROUP BY landing__outcome ORDER BY count

landing__outcome	COUNT
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

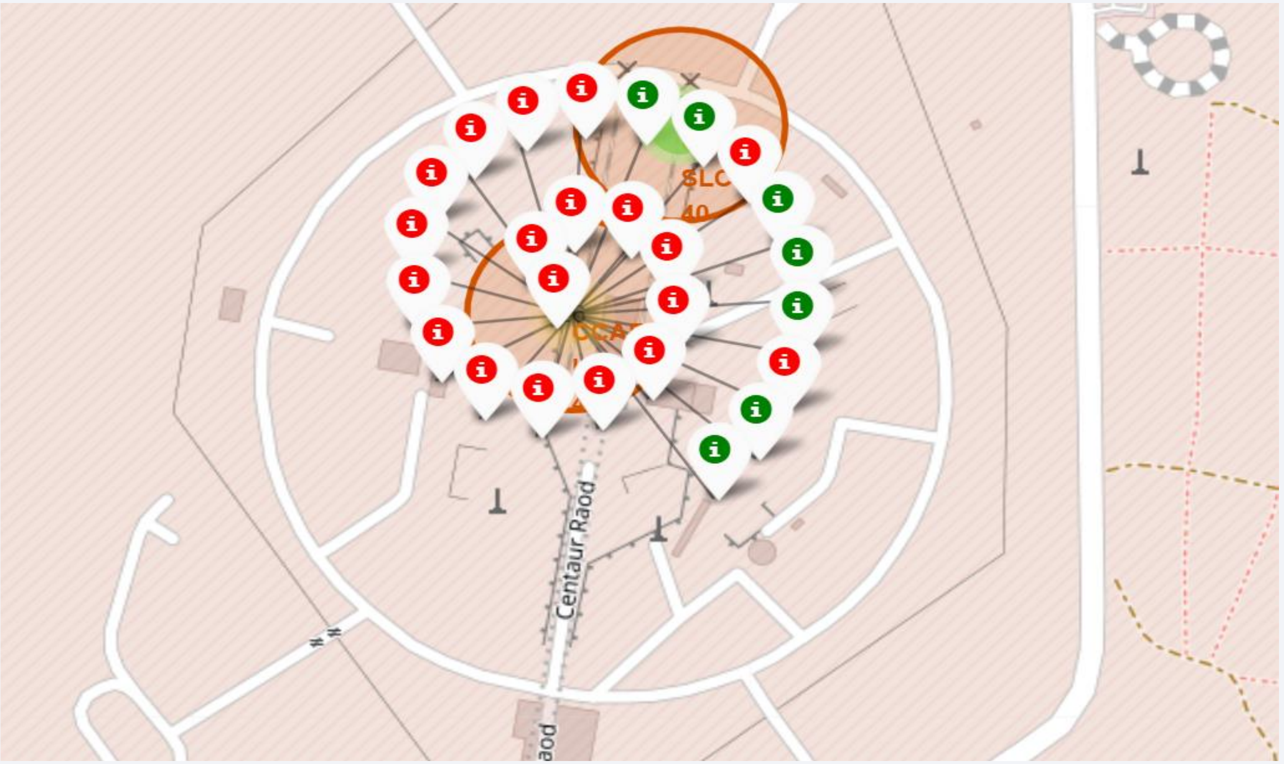
Section 3

Launch Sites Proximities Analysis

Launch Site

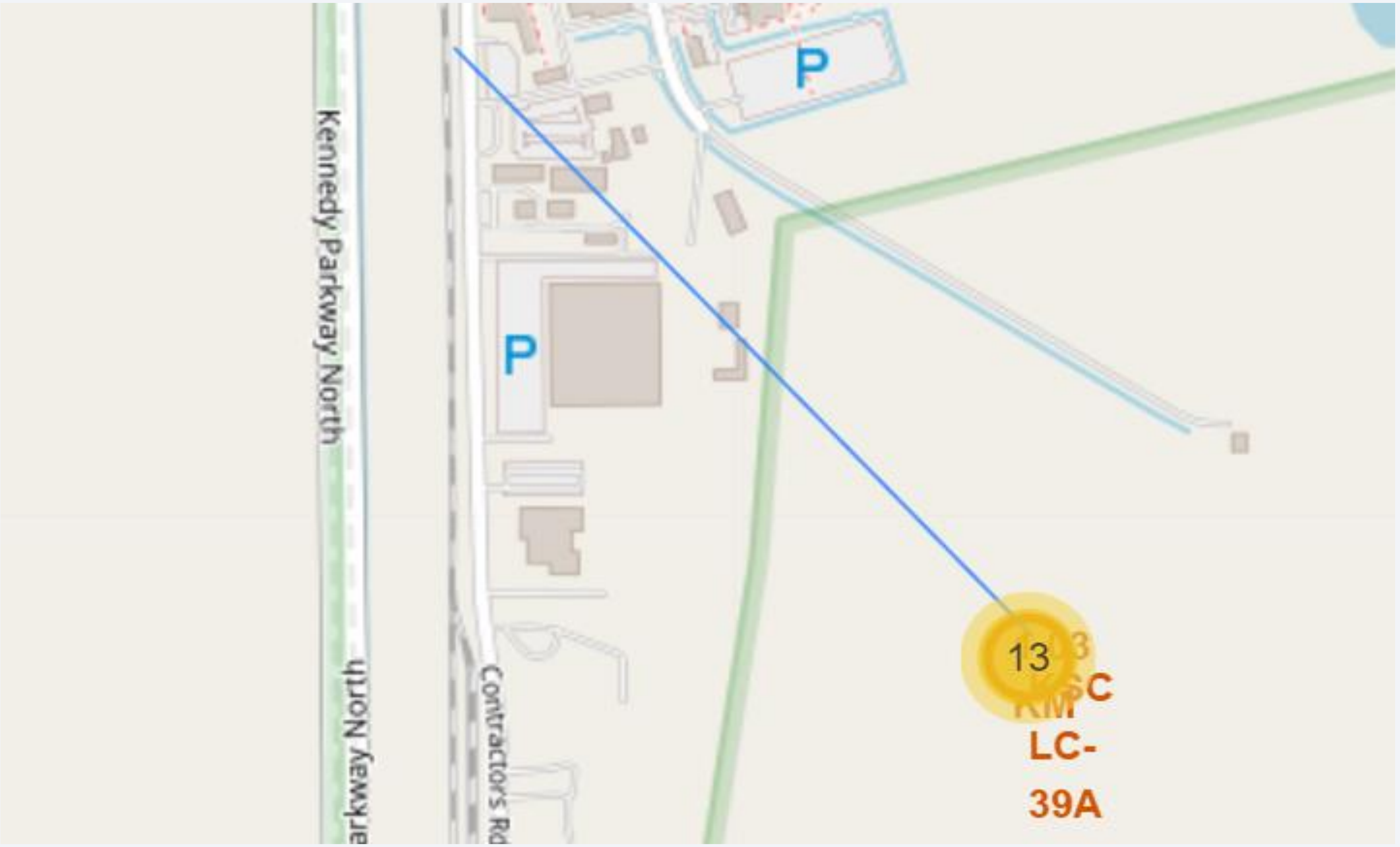
Here we can observe launching sites in the US marked in orange, it is a little bit hard to see in Florida as the three sites are very close.

Launches, Successful and Failed



- Marker clusters is used to simplify the map containing many markers having the same coordinate.
- Successful launches are marked using a green marker and failed launches are marked using a red marker.

Nearest Railway Station

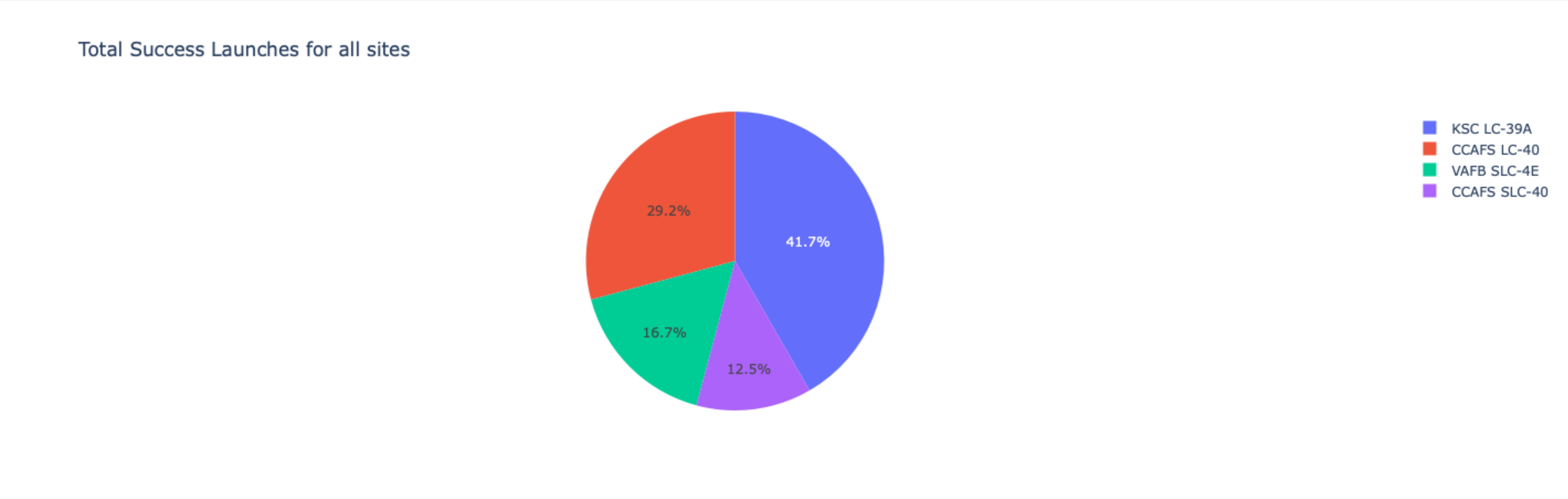


Screenshot show the nearest railway station.

Section 4

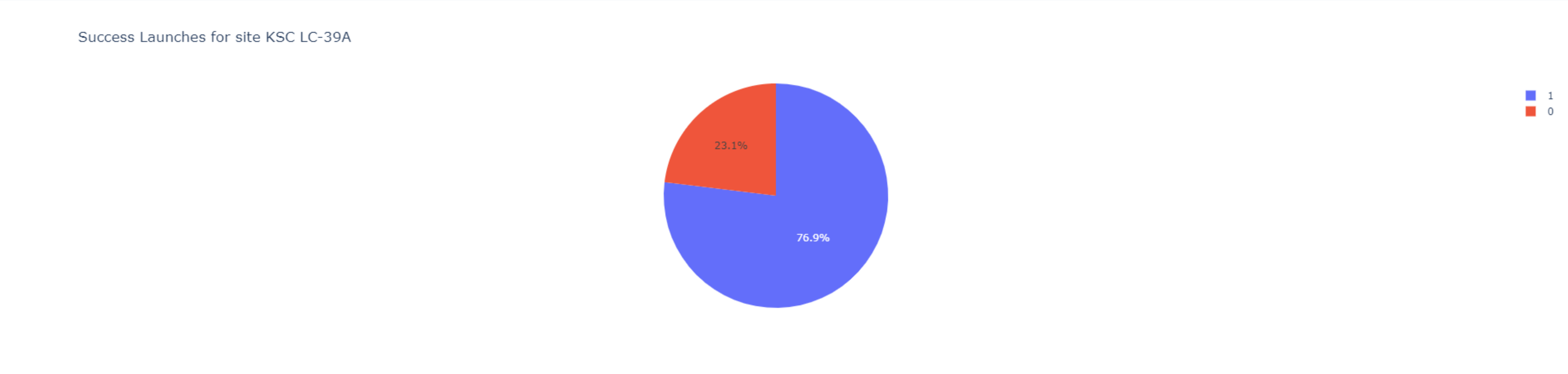
Build a Dashboard with Plotly Dash

Success count for all launch sites



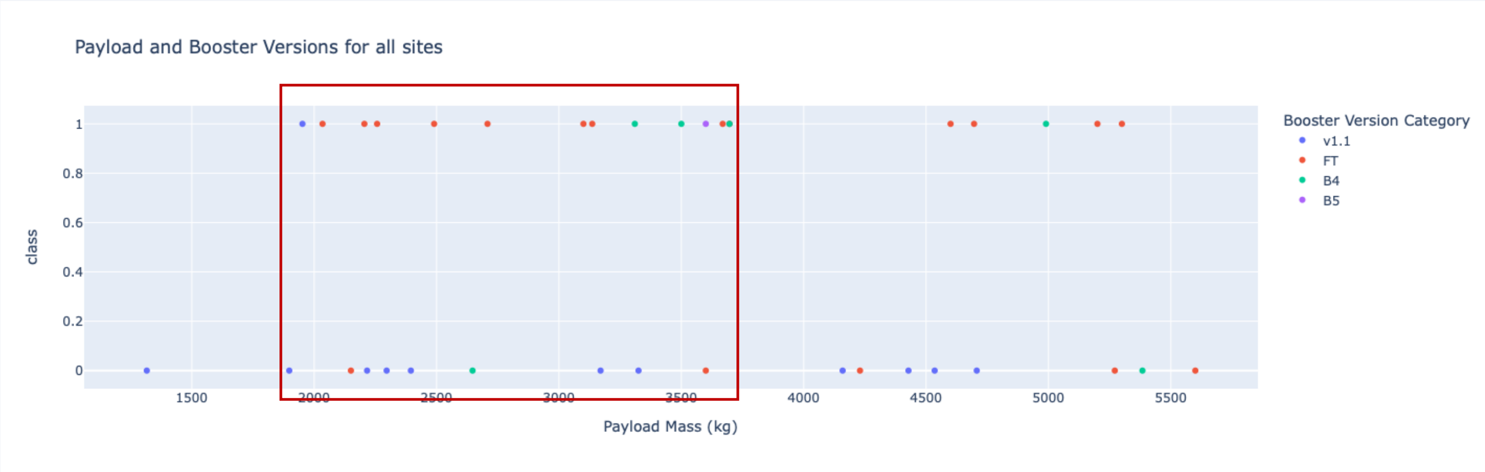
KSC LC – 39A has the largest number of successful launches at 41.7%

Highest launch success ratio

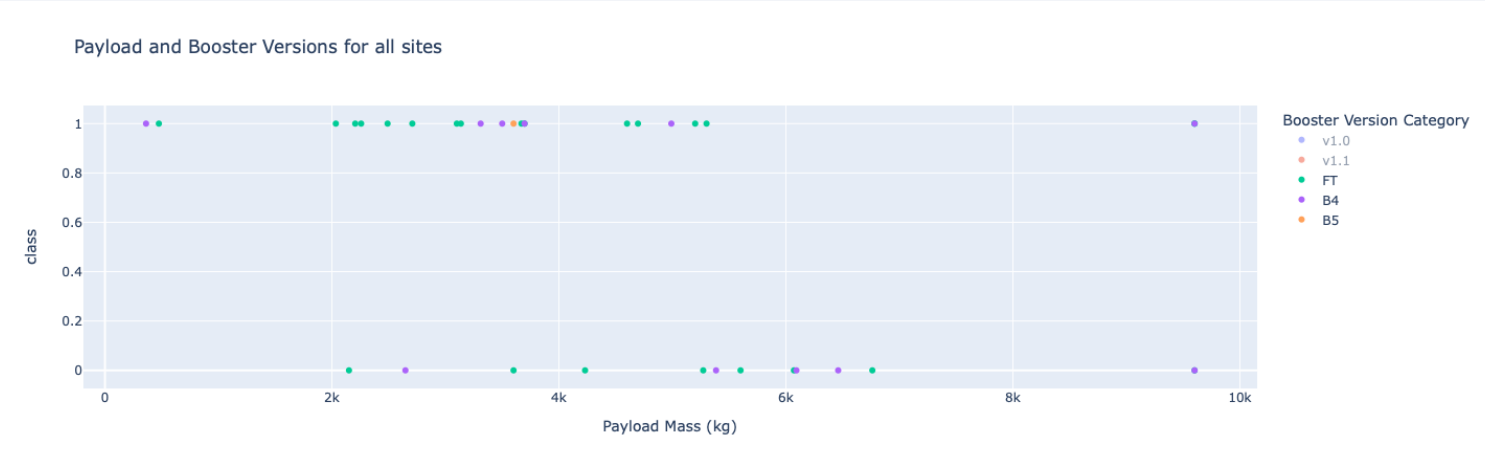


The KSC LC-39A has almost a 77% of success ratio and a 23% failure ratio.

Payload and booster versions on successful launch



We can see payload range 2000 to 3750kg has the highest rate of success as we see the congregation at the top in comparison to the bottom.

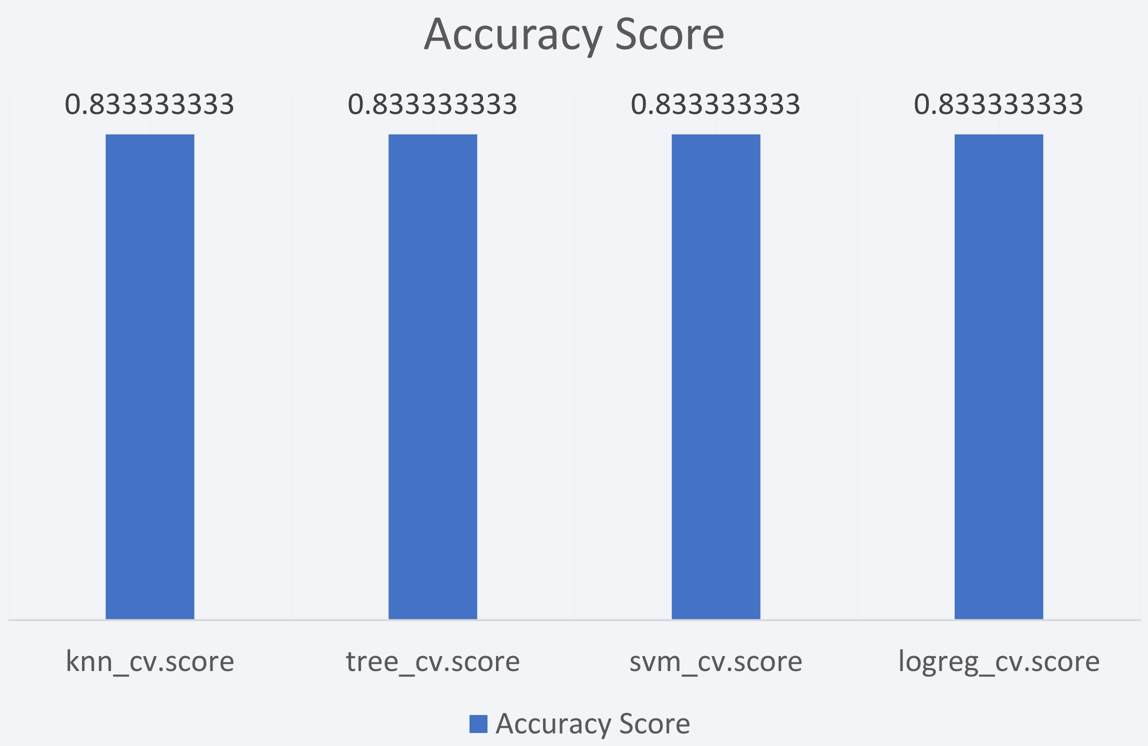


As booster B5 has only 1 launch and it's successful, it has the highest ratio. If we exclude B5 for its small sample size, FT and B4 has higher success ratio with FT having more edge.

Section 5

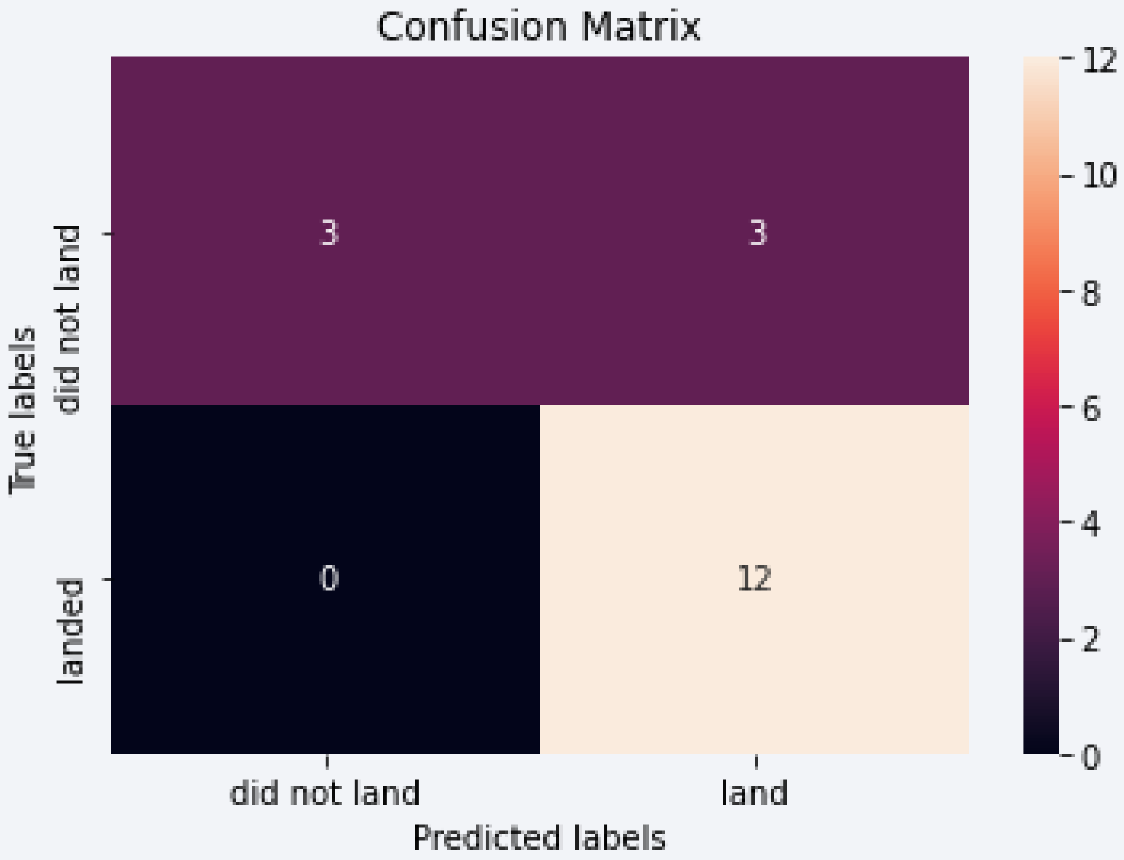
Predictive Analysis (Classification)

Classification Accuracy



Accuracy for the built classification model is 0.833334 for all models - SVM, Classification Trees, Logistic Regression and KNN. Hence they are equivalent of one another and all models can be used.

Confusion Matrix



The confusion matrix for all models is the same. The weakness is the number of false positives as seen in the top right square.

Conclusions

- Data can be extracted from API or webscraped from online sources such as Wikipedia.
- Exploratory analysis can be using various methods such as SQL queries, visualization using charts, interactive dashboards as well as maps when needed.
- Features including ‘FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount’ are important variables.
  - Later flight numbers indicate the flights launched at later timeline, hence we can observe flights at later are more successful since it builds on experience of previous flights.
- Accuracy for the built classification model is equivalent for all models - SVM, Classification Trees, Logistic Regression and KNN. Issue may rise for false positives.

Appendix

- All required can be found in this Github link: https://github.com/Andy-gtm/IBM-Data-Science-Capstone
- Thanks Dorothy Teng, Tarun .

Thank you!