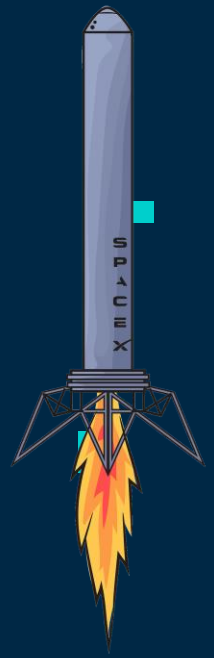
Skills Network

Maximizing landing success of the first stage flight through data

João Luz

18/03/2023



Executive Summary

This presentation has the purpose to find the best investments for a successfull launch site. To colect the data was made webscraping using two different sources with python, using principally numpy, pandas and matplotlib to EDA and scikit learn to classification algorithms.

It was found that the greatest combination for a successfull launch is: KSC LC-39A launch site with VLEO orbit, more than 8000kg of payload mass, RTLS launch and the FT Booster Version for launches.

Table of Contents

Introduction

The Company SpaceX announces that the launches of the Falcon 9 rocket have a cost of 62 million dollars, while other competitors announce something between 165 million dollars.

The price of the Falcon 9 is so much lower due, among other factors, to the reuse of the first flight stage. By determining which factors contribute to a successful landing, we can significantly lower the launch cost, thus being able to create competitive flights with SpaceX.

The purpose of this presentation is to show the factors that contribute to a successful landing through data analysis.

Data collection and data wrangling methodology

Two different web scraping jupyter notebooks was created for collecting the data, one utilizing a RESTful API and the second utilizing a wikipedia page.

First was utilized web scraping from a SpaceX API utilizing "requests", "pandas", "numpy" and "datetime" librarys in python.

Then was created four functions to extract information from the API:

- getBoosterVersion: For learning the booster name.
- getLaunchSite: For launch site, longitude and latitude.
- getPayloadData: To the mass os the payload and the orbit of the flight.
- getCoreData: For the outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, wheter the core is reused, wheter legs were used, the landing pad used, the block of the core which is a number used to seperate version of cores, the number of times this specific core has been reused, and the serial of the core.

Data collection and data wrangling methodology

Was utilized the URL: spacex_url="https://api.spacexdata.com/v4/launches/past" for GET requesting the data and after that we normalized it with json_normalize().

Then the data was passed for a pandas DataFrame format.

The DataFrame was filtered to only include Falcon 9 launchs (excluding Falcon 1).

The missing values in the data (only in the Payload Mass) was substituted for the mean of the column.

Data collection and data wrangling methodology

In the Wikipedia notebook was utilized the "sys", "requests", "BeatifulSoup", "re", "unicodedata" and "pandas" library and was created five functions to extract information from the HTML wikipedia pages:

- Date_time: For data and time
- Booster_version: For Booster Version
- Landing_Status: For Landing Status
- Get_mass: For mass
- Extract_column_from_header:

After that the data was extracted from the HTML tables using a Beatiful Soup object and parsed to a Dataframe.

The wikipedia url is the following: https://en.wikipedia.org/w/index.php title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data collection and data wrangling methodology

It is important tell that are different cases for the landing of a rocket:

- True Ocean: The mission outcome was successfully landed to a specific region on the ocean.
- False Ocean: The mission outcome was unsuccessfully landed to a specific region of the ocean.
- True RTLS: The mission outcome was successfully landed to a ground pad.
- False RTLS: The mission outcome was unsuccessfully landed to a ground pad.
- True ASDS: The mission outcome was successfully landed on a drone ship
- False ASDS: The mission outcome was unsuccessfully landed on a drone ship.

Also is important indicate the existence of different orbits: LEO, VLEO, GTO, SSO, ES-L1, HEO, ISS, MEO, GEO and PO.

EDA and interactive visual analytics methodology

From the last dataset (Wikipedia's tables), some exploratory analysis was made:

Using the value_counts() was counted:

- The number of launches on each site
- The number of launches in each orbit
- The number of different cases per orbit

Using a lambda function in the dataset and spliting the bad outcomes in a list, numerical values was atributed to success/failures (respectively 0 or 1).

The success rate was calculated using the mean().

EDA and interactive visual analytics methodology

A SpaceX CSV document passed to a dataframe was utilized with SQL in python using sqlite. So EDA was realized with queries as follows:

- Display the unique launch site labels
- Display the records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA(CRS)
- Display the average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived
- List the names of the boosters which have succes 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 (a subquery was utilized).
- List the records which will display the month names, failure landing outcomes in drone ship, booster version, launch_site for the months in year 2015
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017

EDA and interactive visual analytics methodology

Using the same dataset more EDA was made with numpy, matplotlib and seaborn, the following things was necessary (sns.catplot() for creating the scatter, plus the success class was overlayed in the scatter plot)

Plots created:

- Scatter plot of the continuous launch attempts vs the payload mass
- Scatter plot of the continuous launch attempts x launch site Scatter plot of the payload mass x launch site labels
- Bar chart of the sucess rate of each orbit (using groupby() and barplot())
- Scatter plot of the continuous launch attempts vs the orbit
- Scatter plot of the payload mass x Orbit
- A line chart to see the success rate through the years (2010-2020)

Some feature engineering was made using one hot enconding (pd.get_dummies) and the entire dataframe was passed to float64 using astype('float64')

EDA and interactive visual analytics methodology

Some exploratory data analysis was made with Folium:

- Mark all launch sites on a map (groupby() the Launch Site columns with latidude and longitude of each launch site, transform the selected data to a dictionary and pass to folium.Map(), plus using .Circle() and .Marker())
- Mark the success/failed launches for each site on the map (a MarkerCluster() object was created, a lambda function to create a column with names "red" and "green" to mark)
- The latitude and longitude of the map was showed with a MousePosition() object, and the distance between the launch sites and proximites (city, railway, hailway and coastline) was calculated with haversine equation.
- Using a PolyLine() a straight line was created conecting the launch site with proximities in the site map.

EDA and interactive visual analytics methodology

A dashboards was created using dash and plotly in python with:

- A dropdown list so select a specific or all launch sites
- A pie chart to show the totall sucess launches depending of the dropdown list
- A slider to select payload range
- A scatter chart to show the correlation between payload and launch sucess

The dashboard was made using @app.callback to take the dropdown and slider inputs and output the pie chart and scatter chart.

Was utilized html.Div(), html.H1 for the title and a server was executed for showing the dashboard.

Predictive analysis methodology

Taking in account that the desired output (Sucsess/failure launches) is a categorical variable, so for classification algoritms was tested:

- Logistic Regression
- Support Vector Machines
- Decision Tree
- K – nearest neighbors

The dataset was passed to nympt with .to_numpy(), standardized with .StandardScaler() object and fit_transform() and the data was divided in train and data (20% for test).

For every classification algoritm, a GridSearchCV() was created to find the best parameters from a dictionary, the accuracy was calculated and a confusion matrix was ploted, then the accuracy of the algoritms was compared.

EDA with visualization results

Count of launches on each site:

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Count of launches on each orbit:

In [8]:
```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()
```

Out[8]:
```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

EDA with visualization results

CCAFS SLC 40 has the most number of launches, 2x more than the 2th place and VAFB SLC 4E has the least.

GTO, ISS and VLEO together have 68% of all the launches, the other 8 orbits only has 32%

EDA with visualization results

Count of mission outcome per orbit type:

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

The success rate of the missions:

```
df["Class"].mean()
```

```
0.6666666666666666
```

EDA with visualization results

Generally the outcome of the mission is ASDS, more than 2x the second place. The most failure outcome is in ASDS too. The rate of failure is 14%

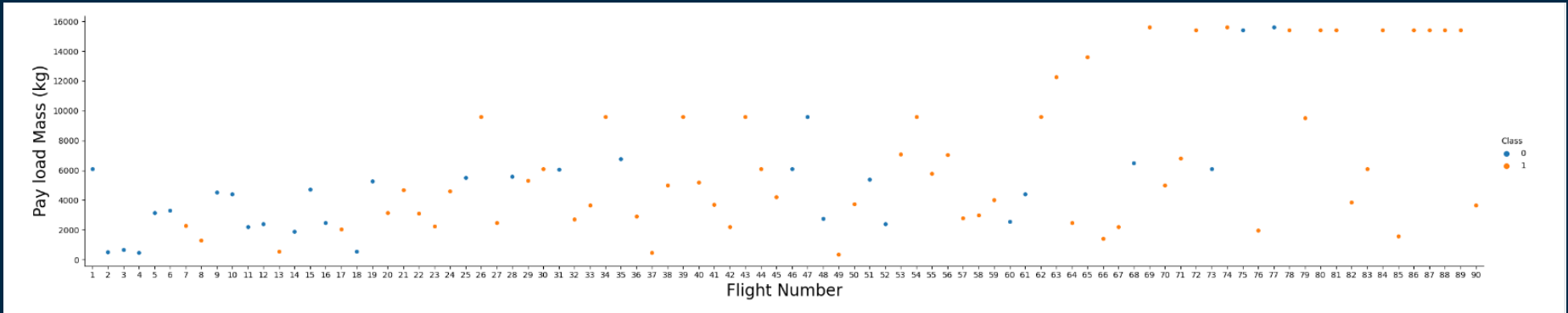
Ocean has a big failure rate: 40%, bigger than the mean failure ratio.

RTLS has the least failure rate: 7% plus is in the top tree number of outcomes. It is the half of the failure ratio of the ASDS

The success rate of all the mission is 66% (so the failure ratio is 34%)

EDA with visualization results

Scatter plot of the Flight Number x Pay load mass(kg)



EDA with visualization results

The success rate increases through number of flights.

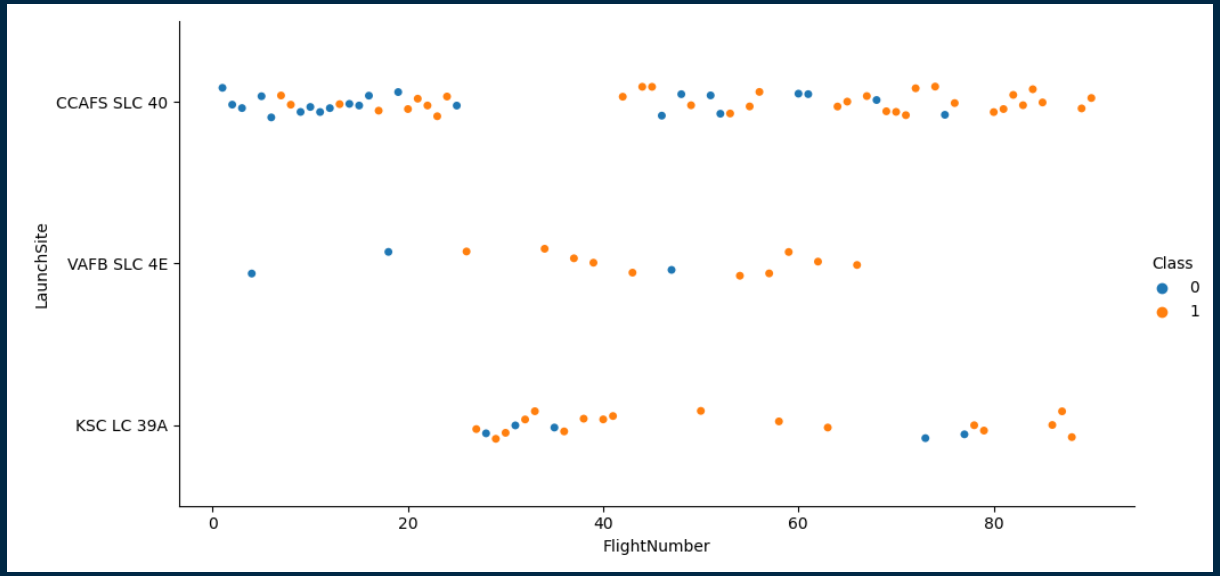
The 4000 – 6000 kg payload has the biggest number of failures.

Payloads bigger than 8000 kg have the smaller number of failures.

Payload with 8000-10000 kg is the range with the smaller number of failures.

EDA with visualization results

Scatter plot of the Flight Number x Launch Site



EDA with visualization results

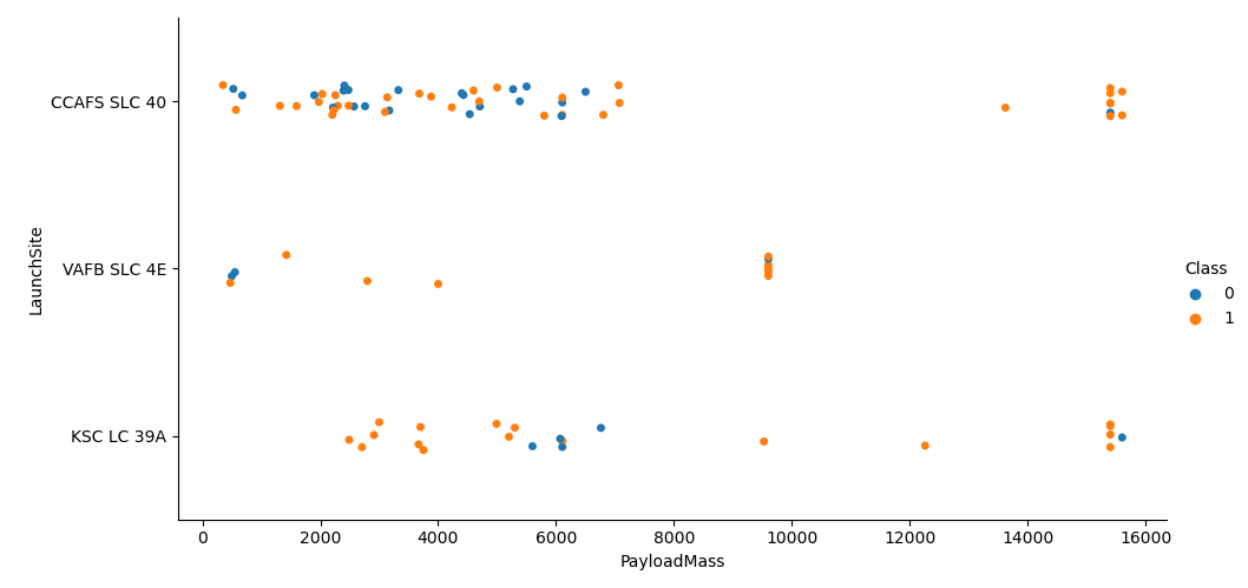
In CCAFS and VAFB the success rate increased with the number of flights, but in KSC that wasn't the case.

KSC has a failure rate between 40+ number of flights of 18%, CCAFS has 33% and VAFB has 11% (the least).

CCAFS has 3 times the number of failures than VAFB.

EDA with visualization results

Scatter plot of the Payload mass x Launch Site



EDA with visualization results

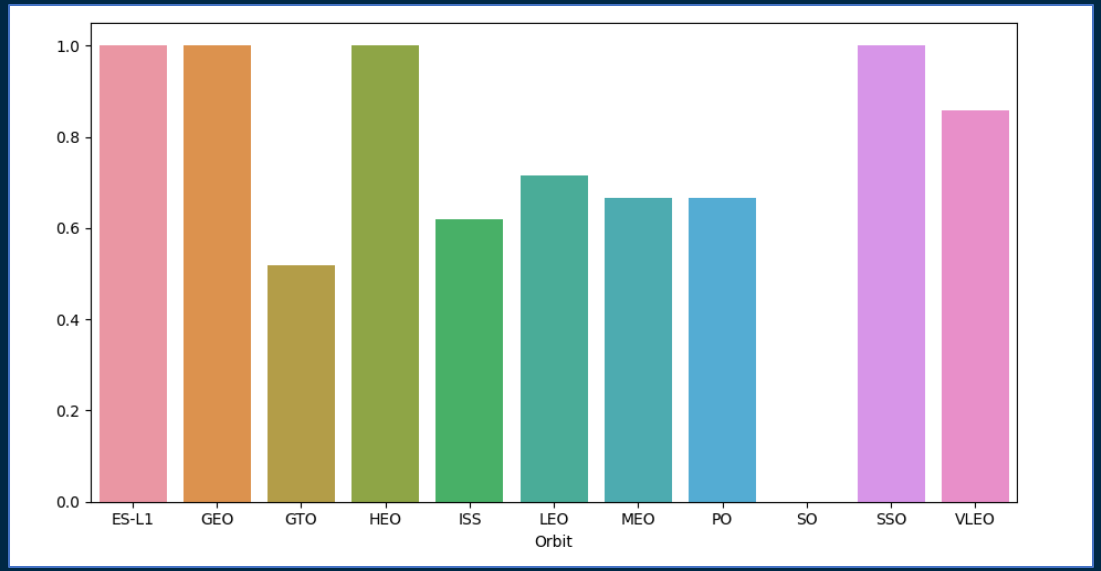
The payload between 0kg and 8000kg has a giant number of failures in all launch sites compared to the payload bigger than 8000kg has much less failures, but it has significatively less flights.

The failure rate in CCAFS of light weight (less than 8000Kg) is big compared to the other two Launch Sites.

VAFB has the least number of failures in the light weight.

EDA with visualization results

Bar Chart showing the success rate of each orbit type



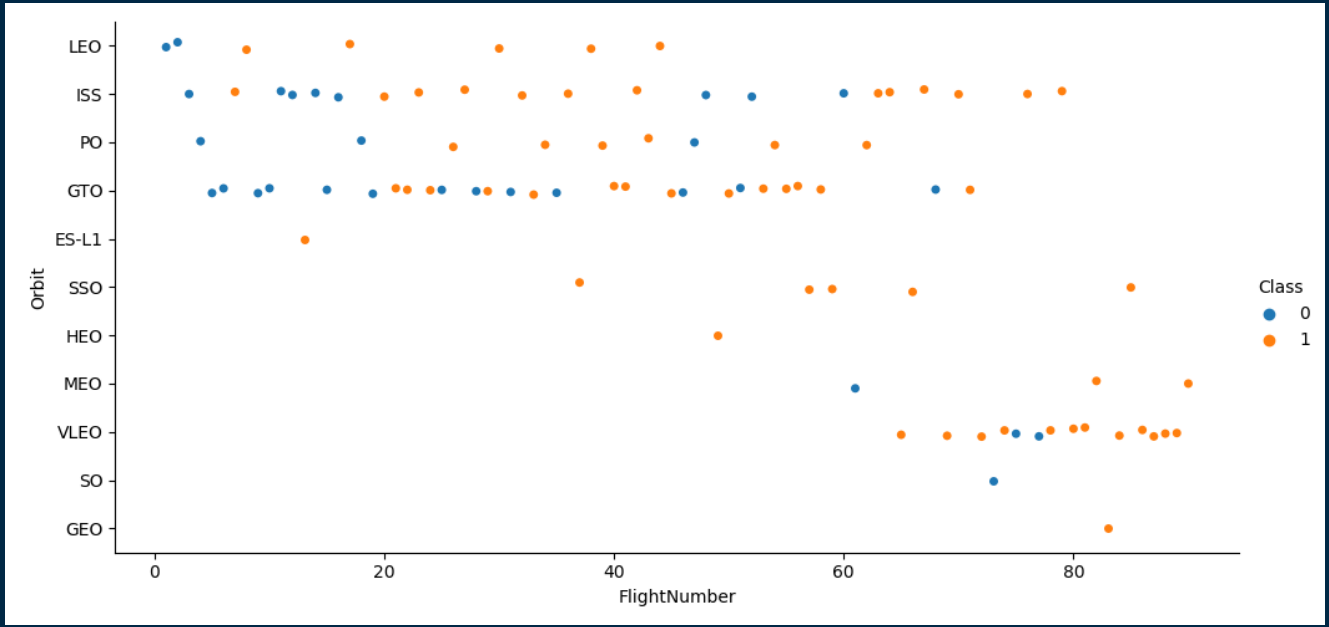
EDA with visualization results

ES-L1, GEO, SSO and HEO have a 100% rate of success.

The most failure orbit is GTO with 50% rate, and ISS with 60%. VLEO has a 90% success rate.

EDA with visualization results

Scatter plot of the Flight Number x Orbit type



EDA with visualization results

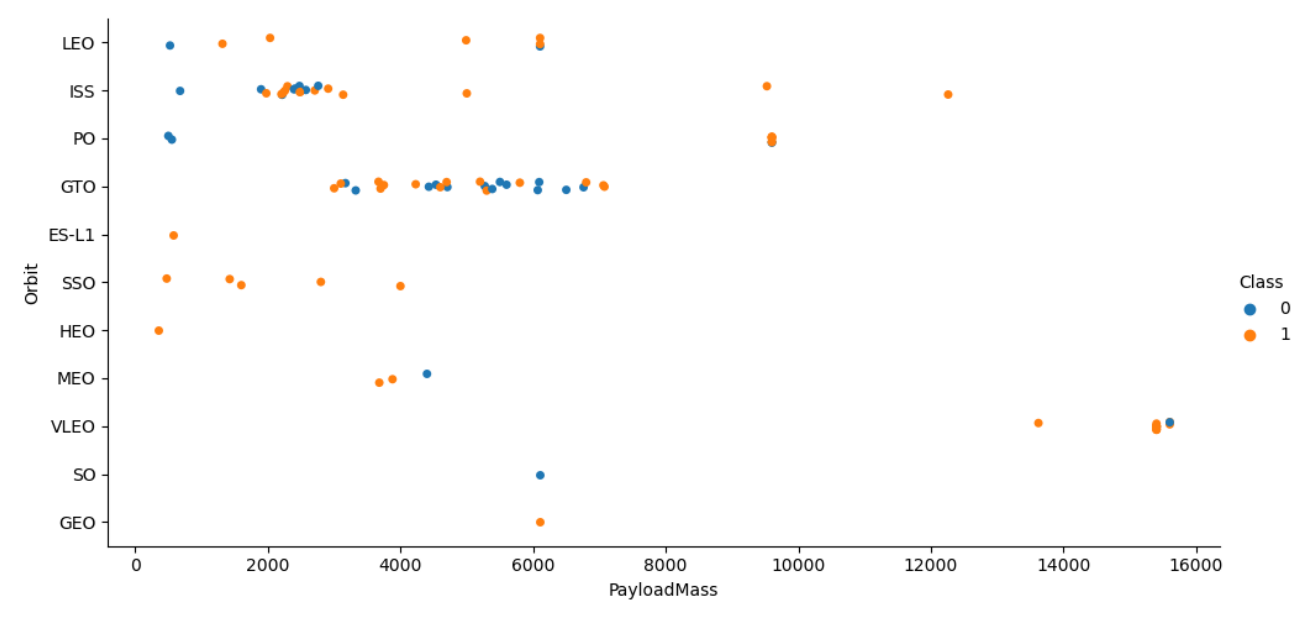
In GTO, PO and ISS the success rate doesn't seem to increase with the number of flights. In LEO this isn't the case.

SSO is the only with a razoable number of flights that have 100% of success. ES-L1, HEO and GEO only have one flight.

VLEO has a big number of flights and the least number of failures.

EDA with visualization results

Scatter plot of the payload mass x Orbit type



EDA with visualization results

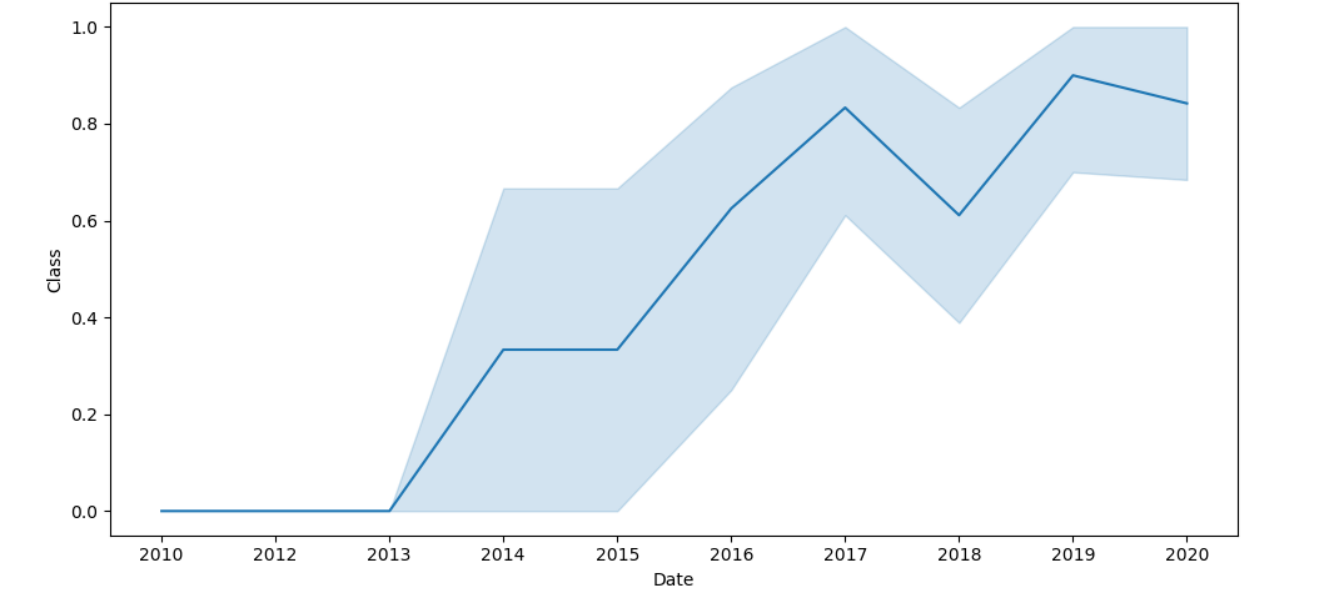
Specific orbits have a range of payload mass, for example VLEO doesn't have flights with less than 14000 kg, GTO only have flights between 3000-8000.

In the light weight SSO has a perfect success rate and goes in the range 0-4000.

In the range 4000-8000kg only GTO was tested.

EDA with visualization results

Line chart showing the sucess rate of flights through 2010-2020



EDA with visualization results

The success rate tends to grow through the years, stagnating in 2014-2015 and only decreasing in 2017-2018 and more recently 2019-2020.

But overall the rate of success is 80-90%.

EDA with SQL results

Display the names of the unique launch sites in the space mission

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

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

EDA with SQL results

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

sum(PAYLOAD_MASS__KG_)
619967

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

payloadmass
6138.287128712871

List the date when the first succesful landing outcome in ground pad was acheived.

min(DATE)
01-03-2013

EDA with SQL results

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

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

List the total number of successful and failure mission outcomes

count(MISSION_OUTCOME)
1
98
1
1

EDA with SQL results

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

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 the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

month	Booster_Version	Landing _Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

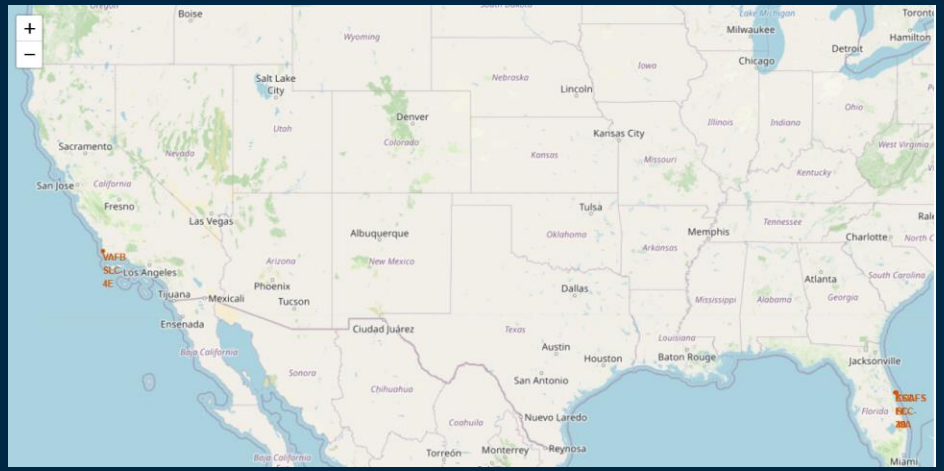
EDA with SQL results

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

Date	Landing _Outcome	LANDING_OUTCOME_COUNT
22-05-2012	No attempt	10
08-04-2016	Success (drone ship)	5
10-01-2015	Failure (drone ship)	5
22-12-2015	Success (ground pad)	3
18-04-2014	Controlled (ocean)	3
29-09-2013	Uncontrolled (ocean)	2
04-06-2010	Failure (parachute)	2
28-06-2015	Precluded (drone ship)	1

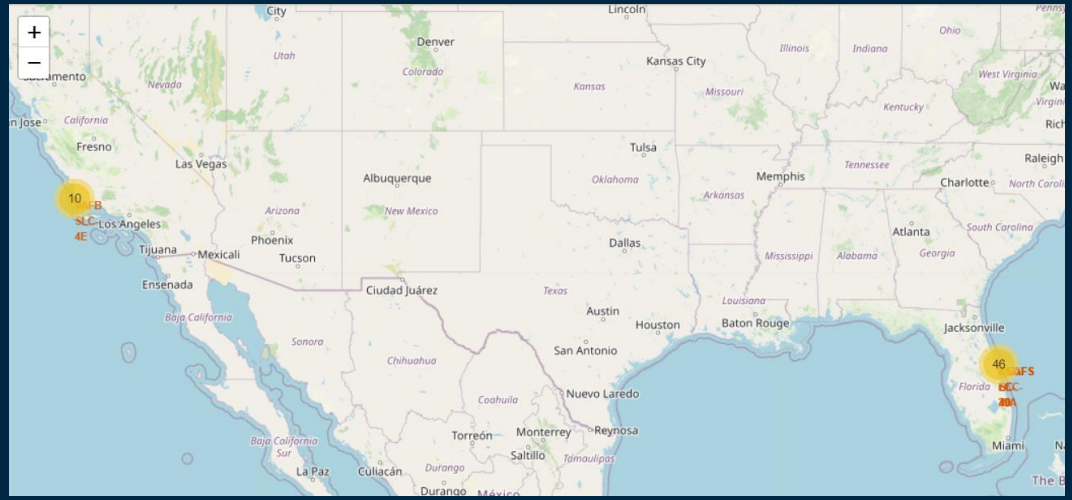
Interactive map with Folium results

Launch Sites marked with orange



Interactive map with Folium results

Successfull/failures launch counts per launch site

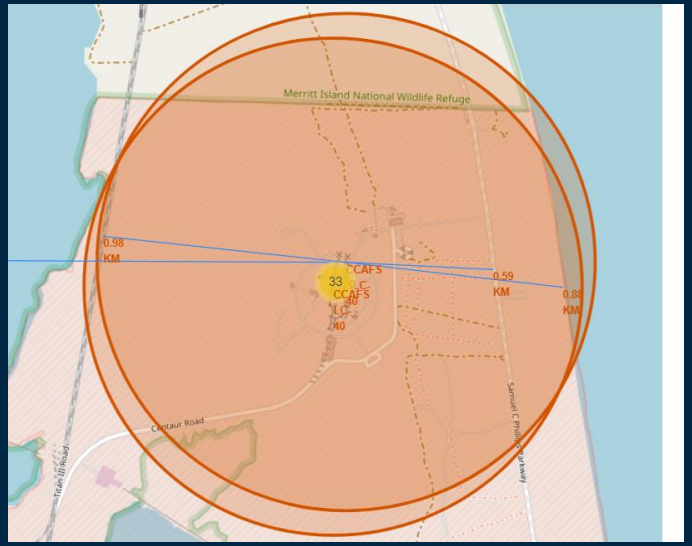


Interactive map with Folium results

Cursor in the map with latitude and longitude

Interactive map with Folium results

Straight line and haversine distance showed in the map

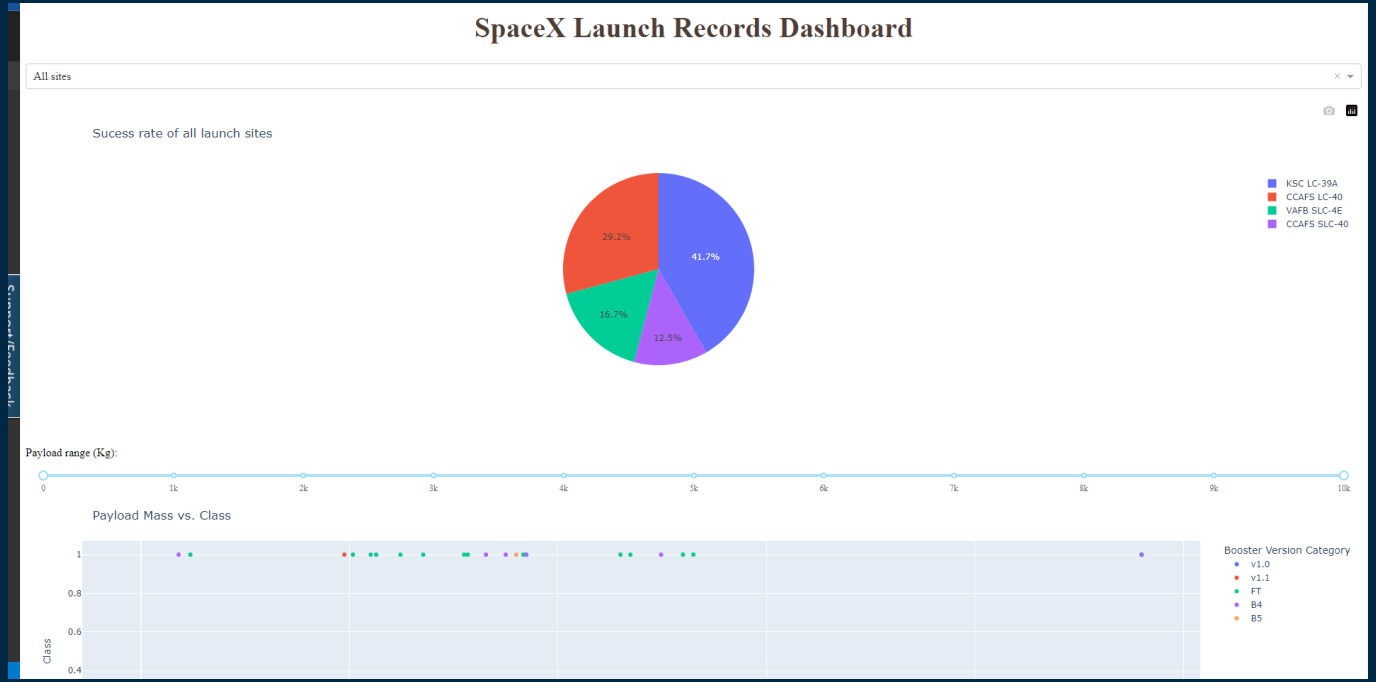


Interactive map with Folium results

All the launch sites have to be close to a railway, hailway and a coastline, plus, have to be some bigger distance to cities, but don't too far.

Plotly Dash dashboard results

General Vision of the dashboard



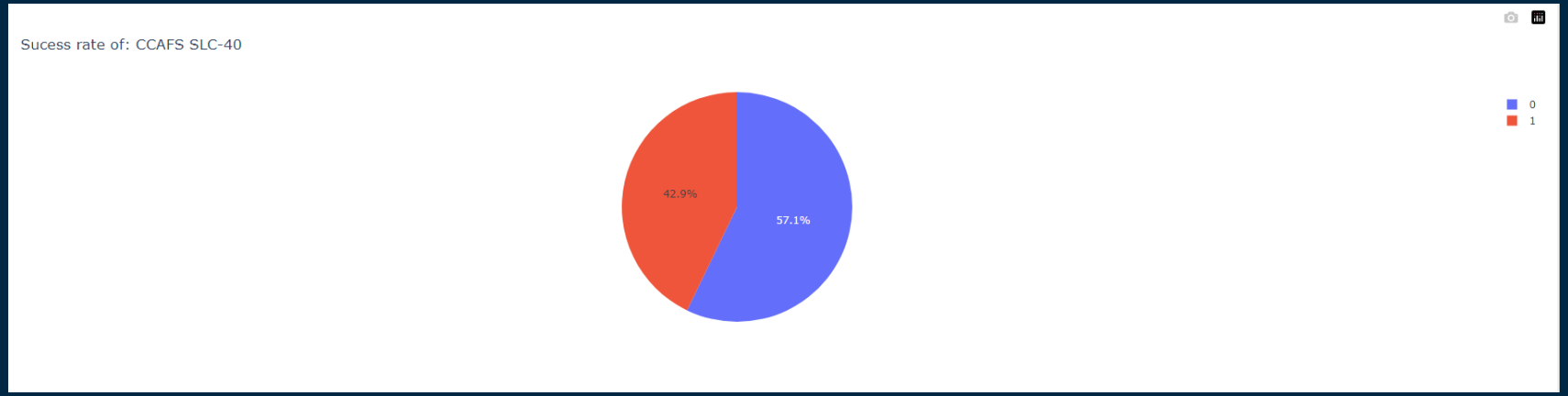
Interactive map with Folium results

CCAFS LC-40 have 2x the success rate of CCAFS SLC-40, that has the biggest number of failures. The two launch sites are very close each other.

41.7% of the success comes from KSC LC-39A, that have the biggest proportion of success compared to other launch sites.

Plotly Dash dashboard results

Sucess rate of a launch site



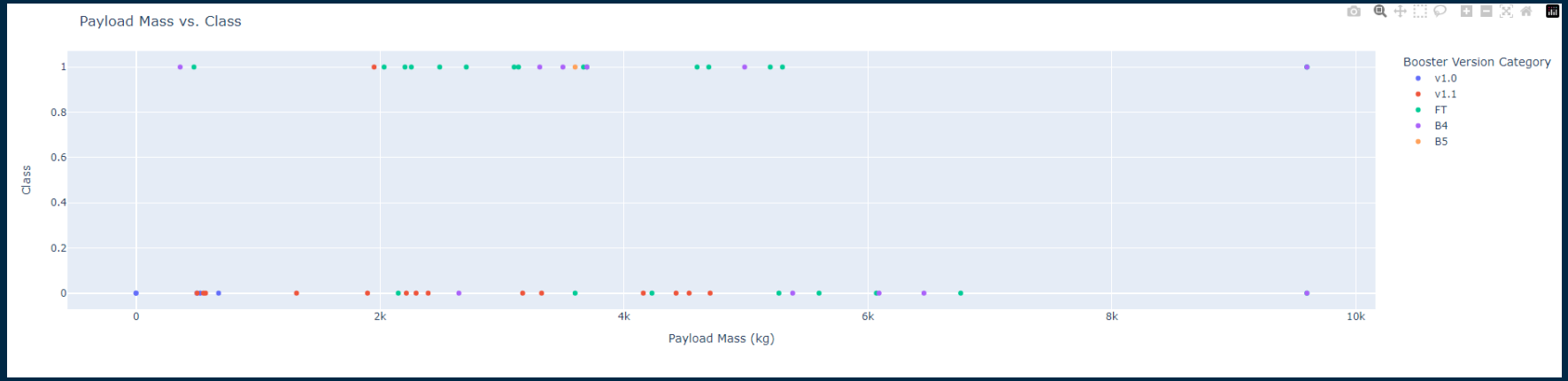
Interactive map with Folium results

KSC LC-39A has a success rate of 76.9%, and CCAFS LC-40 has a success rate of 73.1% with the biggest rate of success.

CCAFS SLC-40 has the smaller number of success rate, only 57.1%

Plotly Dash dashboard results

Scatter plot of the Payload Mass x Class



Interactive map with Folium results

The Booster Version with biggest success rate is B5, but only one flight was made. Taking into account number of flights, FT has the biggest success rate.

The B4 booster has a 50% success rate, the smallest success rate taking into account the significatively number of flights.

Predictive analysis (classification) results

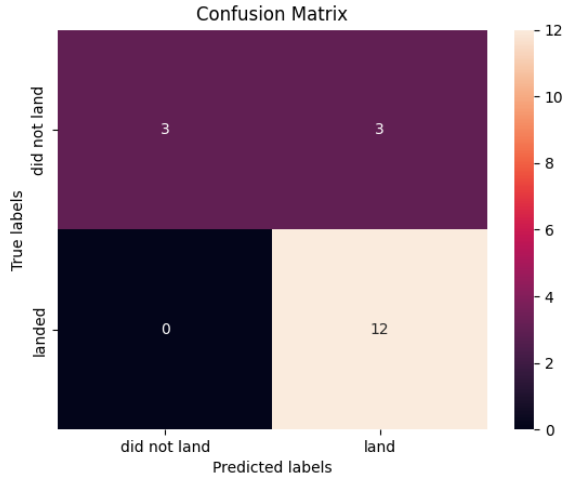
Logistic Regression Results

```
logreg_cv.score(X_test,Y_test)
```

```
0.8333333333333334
```

Lets look at the confusion matrix:

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Predictive analysis (classification) results

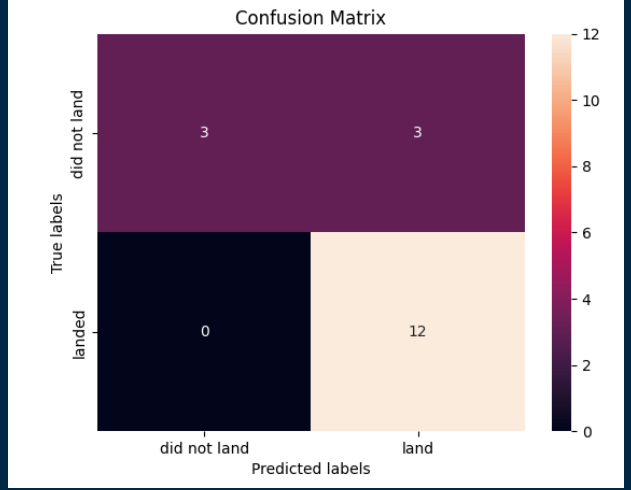
SVM results

```
svm_cv.score(X_test,Y_test)
```

```
0.8333333333333334
```

We can plot the confusion matrix

```
yhat=svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Predictive analysis (classification) results

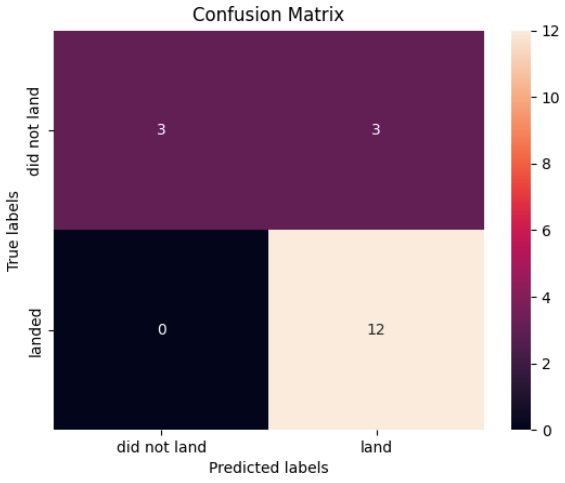
Decision tree results

```
tree_cv.score(X_test, Y_test)
```

```
0.8333333333333334
```

We can plot the confusion matrix

```
yhat = svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Predictive analysis (classification) results

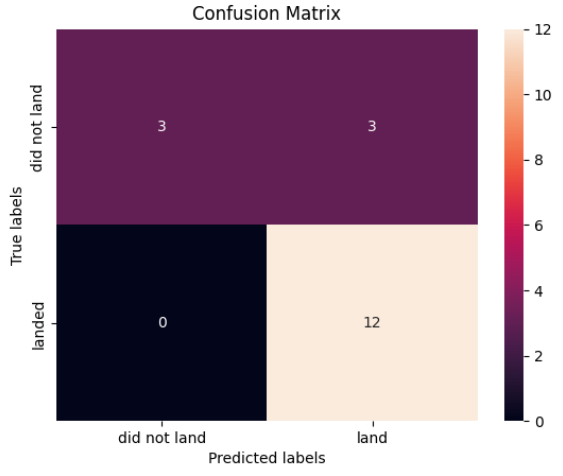
KNN results

```
knn_cv.score(X_test, Y_test)
```

```
0.8333333333333334
```

We can plot the confusion matrix

```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Predictive analysis (classification) results

Comparison between algorithms

	Accuracy
Logistic Regression	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

All the algorithms have a accuracy of 0.83.

Conclusion

Taking into account that the number of success launches has a strong correlation with the number of flights, it is preferable chosing the launch sites that have more frequency of flights.

However, the success rate of CCAFS SLC 40 is significatively smaller than KSC LC-39A, that doesn't have the same number of flights, but has a razoable number of launches. So in general, **at the present moment, it is more worth investing in KSC LC-39A to maximize the chances of success.** It is important note too that the success rate of CCAFS SLC can increase with the time, so monitoring the data continuosly is necessary.

Talking about orbits, the same logic aplies to number of launche sites, i.e, more launches in a orbit, more chances of geting sucess. But GTO has a small chance of sucess compared with VLEO that doesn't has 100% of sucess like ES-L1, GEO and HEO, but in compensation VLEO has a big number of flights, so **is more worth investing in VLEO flights.**

Taking into account the payload mass**, the orbit that have to be prioritize in the range of 2000-5000kg is the SSO,** and in general, launches that have more than 8000kg have more chance of landing successfuly, so **the heavy weights should be prioritize.**

Conclusion

Also is important noting that RTLS has the least failure rate and the same time has a razoable number of flights, so **RTLS launches should be prioritize and ocean launches should be avoided**.

For constructing a launch site it should be close to a railway, hailway and a coastline at the same time mantaining a distance from a city but not so distant (like desert places).

Taking into account the number of flights, **the Booster Version that should be prioritized is the FT.**

Any classification model (KNN, SVM, Logistic Regression, Decission Tree) **can be choosen randomly** to predict if the launch will fail or not because the four have the same accuracy.

Appendix

Description of different orbits: https://en.wikipedia.org/wiki/List_of_orbits

Github with all the notebooks: https://github.com/JV-Luz/AppliedDataScienceCapstone