IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Łukasz Czebiołko
09.01.2023

Outline

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

Executive Summary

Methodologies:

- Data collection
- Data wrangling
- Exploratory data analysis
- Interactive visual analytics
- Predictive analysis

Summary of all results:

- Exploratory data analysis results
- Proxymities analysis results
- Predictive analysis results

Introduction

Project background:

SpaceX company with their Falcon 9 rocket managed to reduced cost of launching payload into space thanks to reusing first stage. Rockets from different companes cost to 165 mln dollars, but Falcon 9 is available for 63 mln dollars.

Problems to solve:

Project task is to predict if reusable first stage of Falcon 9 will land successfully.

Section 1

Methodology

Methodology

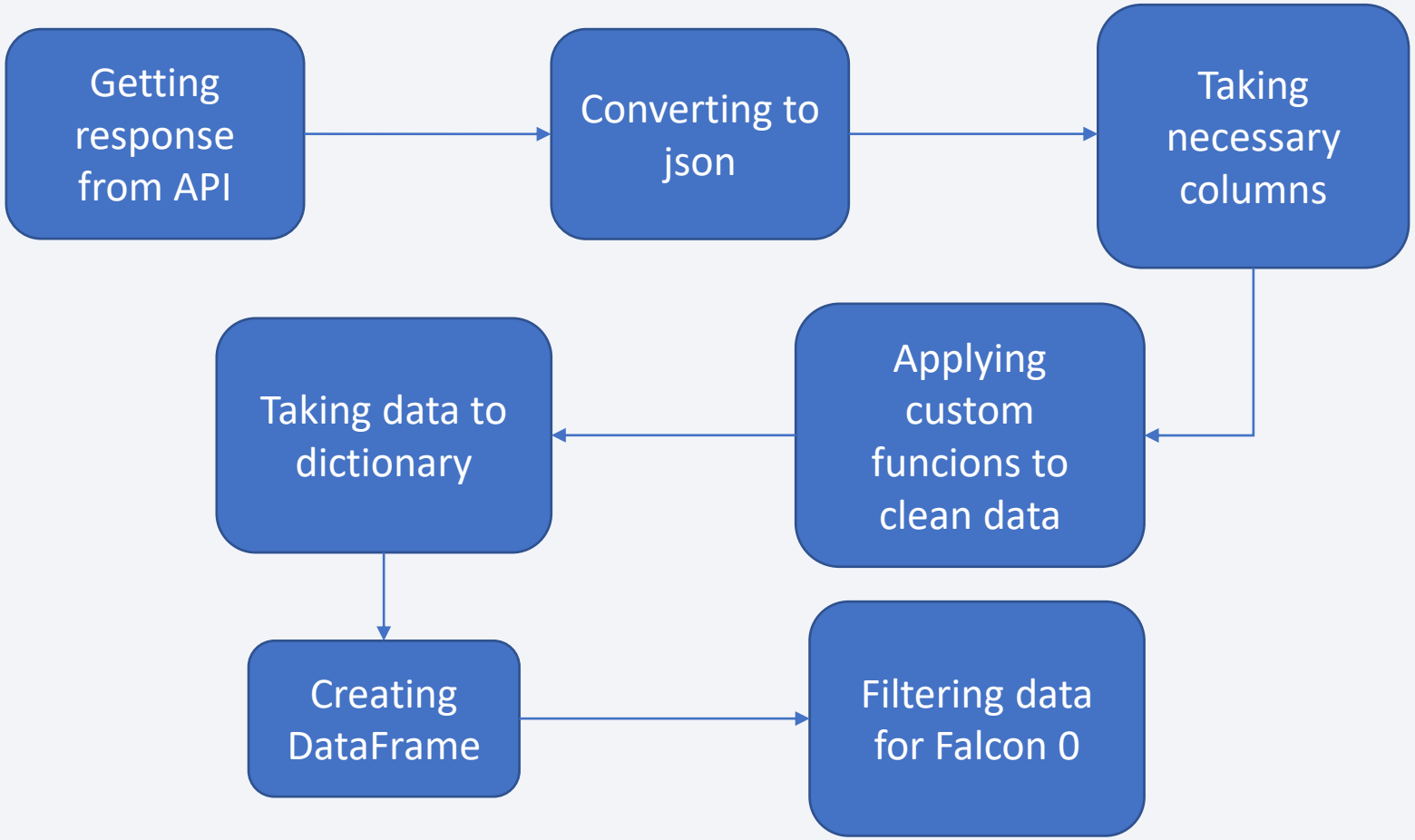
Executive Summary

- Data collection methodology:
  - Wikipedia – web scrapping
  - SpaceX Rest API
- Perform data wrangling
  - Delleting unneccessary columns, cleaning null values and One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform using classification models
  - Classification models: Linear Regression, SVM, Decision Tree, KNN

Data Collection

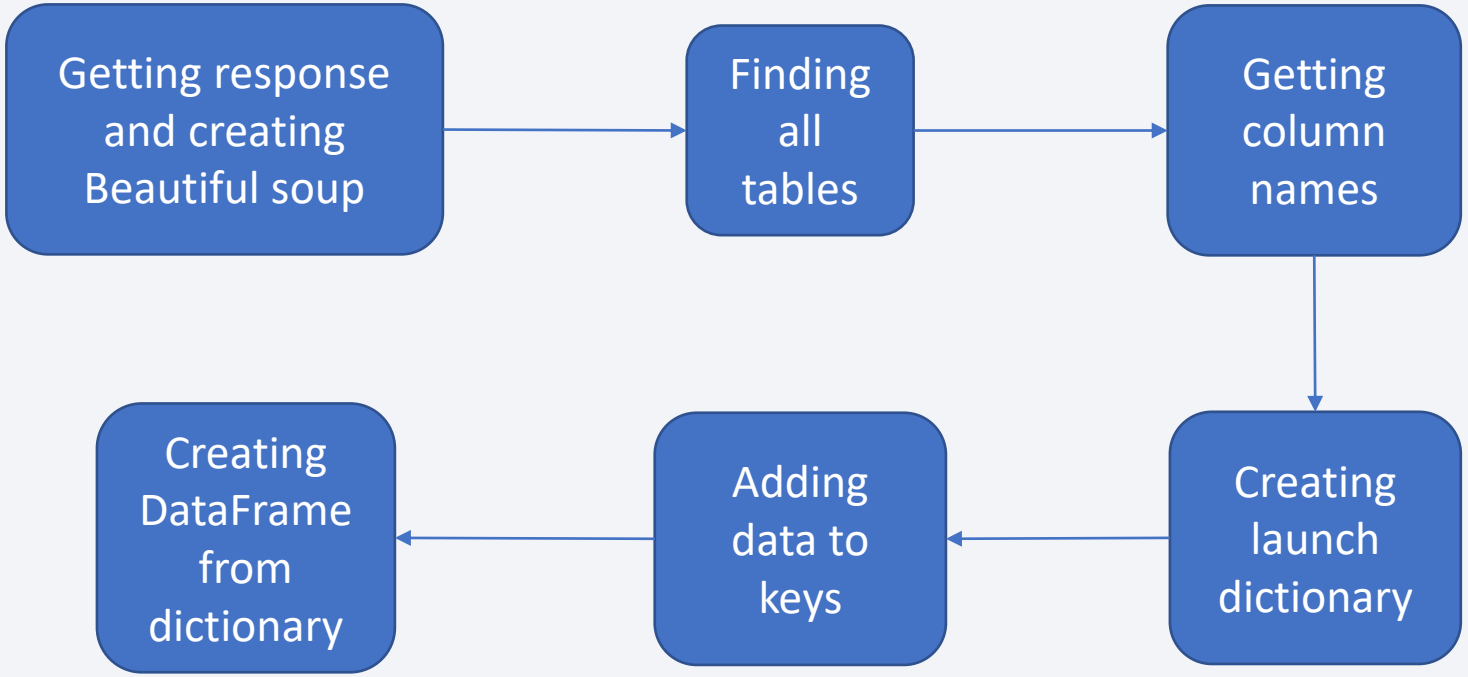
Data was collected by using SpaceX API and web scrapping wikipedia page about launches of Falcon rockets.

Data Collection – SpaceX API



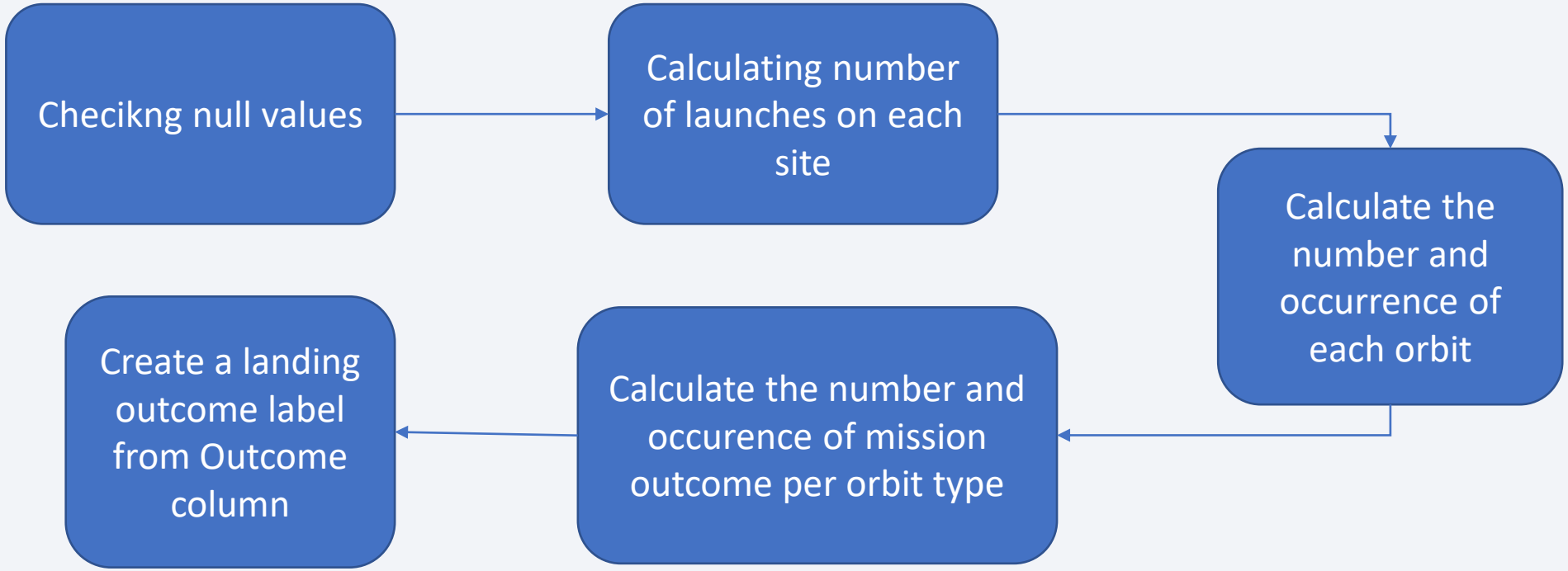
IBM_data_science_professional_certificate-/jupyter-labs-spacex-data-collection-api.ipynb at main · LukaszCzeb/IBM_data_science_professional_certificate- (github.com)

Data Collection - Scraping



[IBM_data_science_professional_certificate-/jupyter-labs-webscraping.ipynb at main · LukaszCzeb/IBM_data_science_professional_certificate- (github.com)]

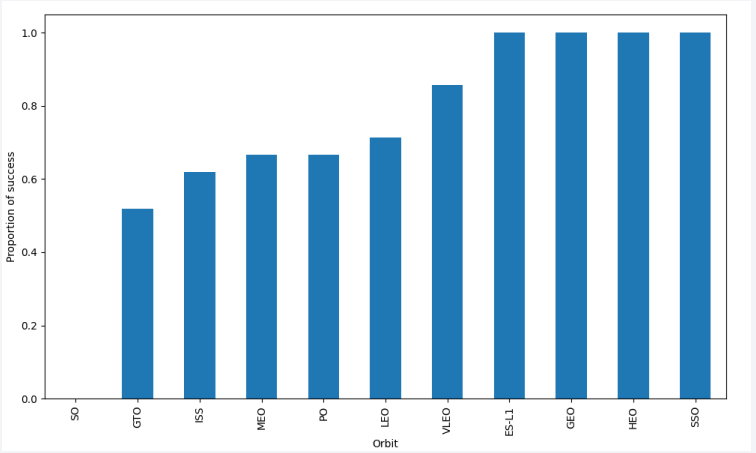
Data Wrangling

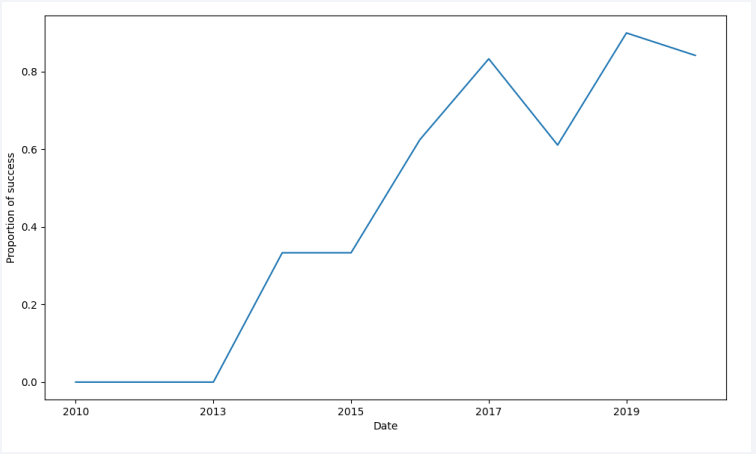


IBM_data_science_professional_certificate-/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb at main · LukaszCzeb/IBM_data_science_professional_certificate- (github.com)

EDA with Data Visualization

- Scatterplots to show correlation between different factors, like
- Bar plot to show number of successed landing among orbits
- Line plot to show number of successed landing in each year





[IBM_data_science_professional_certificate-/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb at main · LukaszCzeb/IBM_data_science_professional_certificate - (github.com)]

EDA with SQL

SQL querries:

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

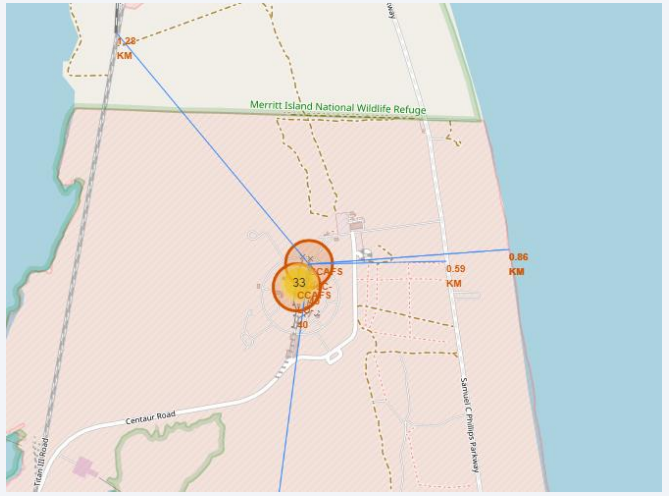
IBM_data_science_professional_certificate-/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · LukaszCzeb/IBM_data_science_professional_certificate- (github.com)

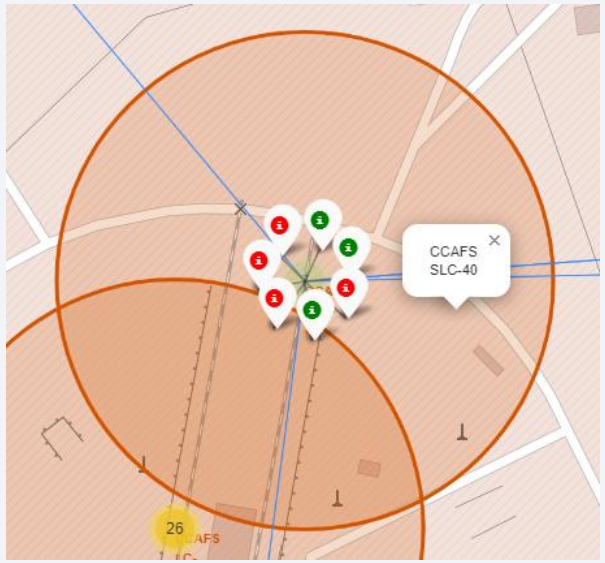
Build an Interactive Map with Folium

I added to map:

- markers to check places of Falcon 9 launches with information if was successful or not
- circles to show launch sites
- lines to show distance to closest railway, highway, coastline and city

With this markers You can see paterns about launch sites locations.





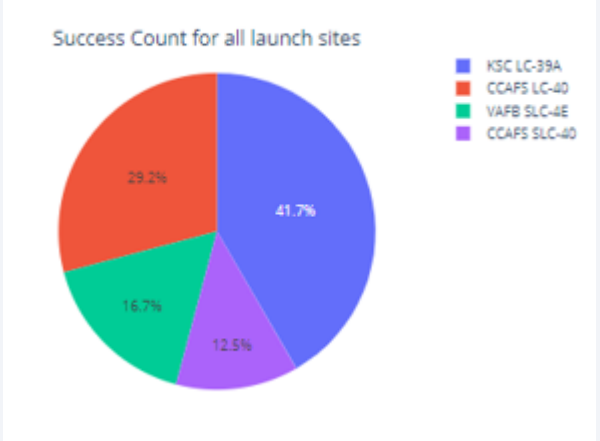
[IBM_data_science_professional_certificate-/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb at main · LukaszCzeb/IBM_data_science_professional_certificate-(github.com)]

Build a Dashboard with Plotly Dash

In Dashboard I added posibility to choose launching site (or to chose all sites) and show pie chart with success launching rate and scatter plot showing correlation between payload and success. There is also possibility to change payload range with slider.

Thanks to that we are able to see patterns between payload and success rate.

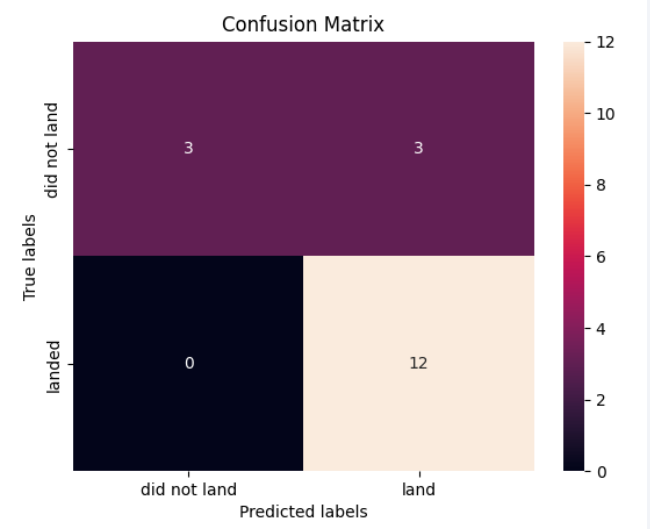
[IBM_data_science_professional_certificate-/spacex_dash_app.py at main · LukaszCzeb/IBM_data_science_professional_certificate-(github.com)]



Predictive Analysis (Classification)

After data standardization I split it on training and test set. With GridSearchCV I created four models: Logistic Regression, SVM, Decision Tree and KNN.

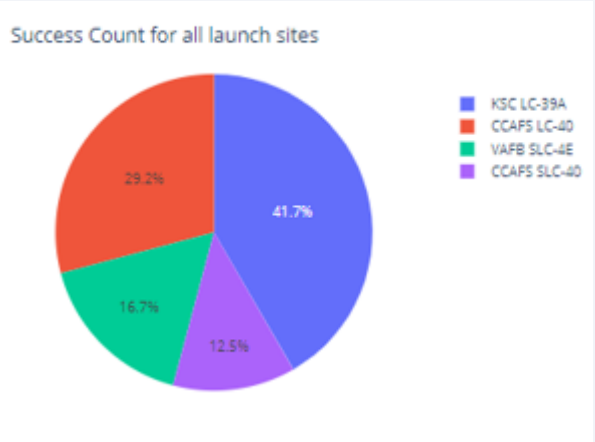
LR, SVM and KNN had best accuracy with 83,3%.

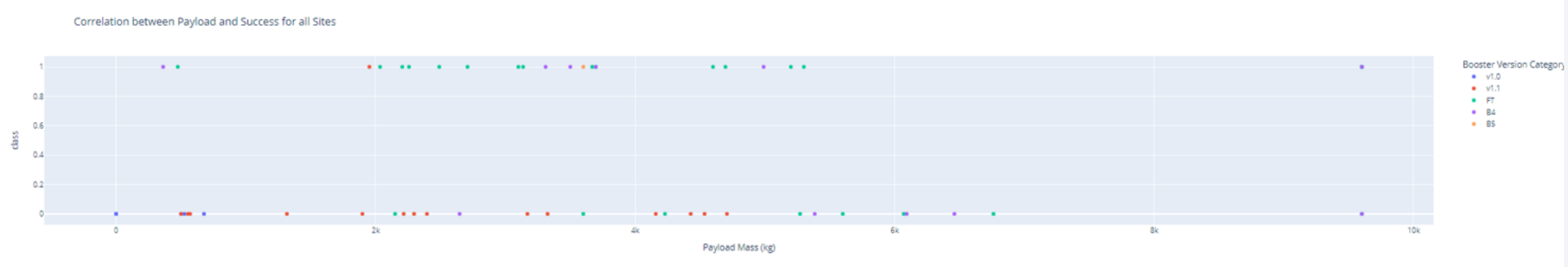


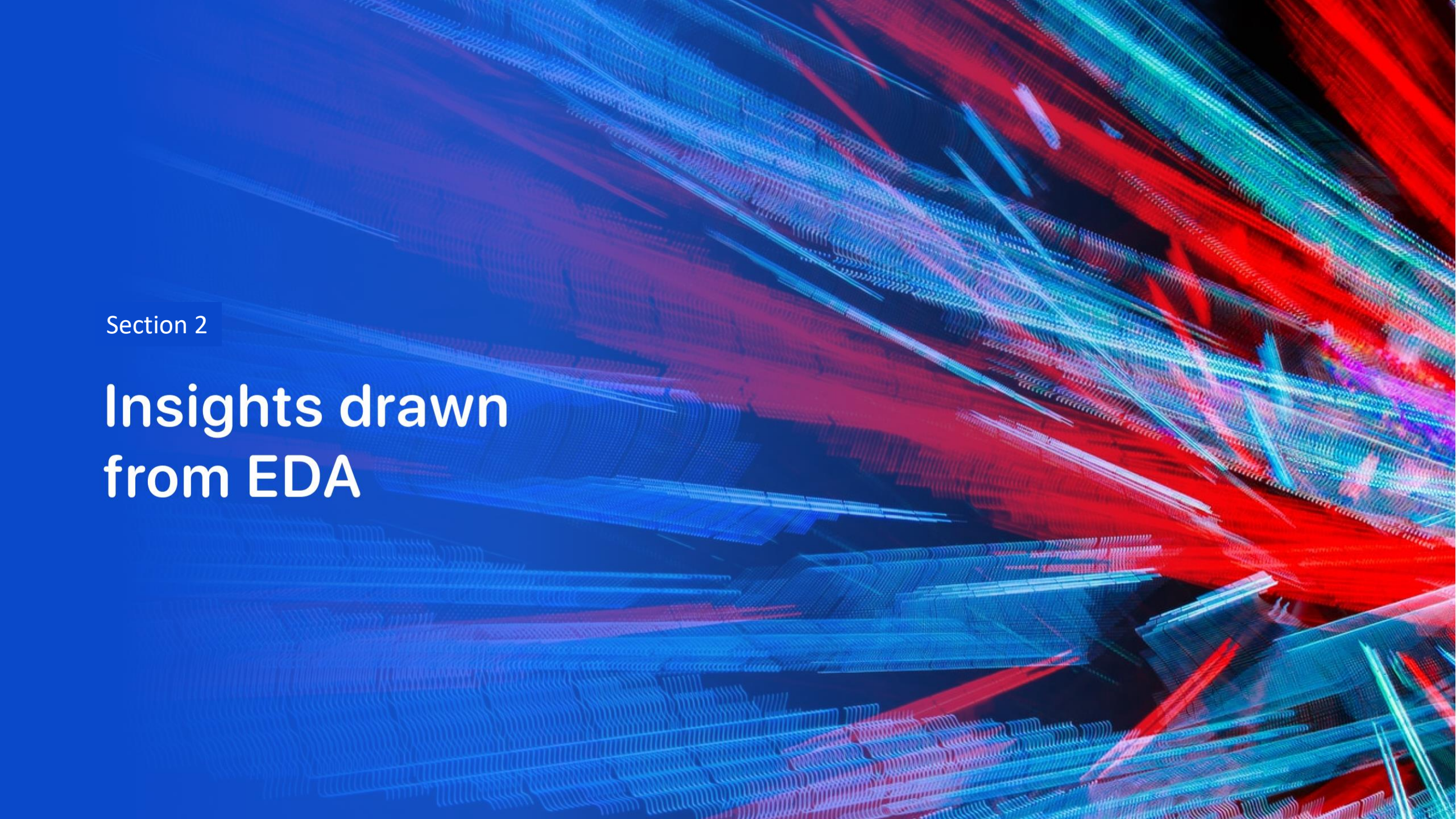
[IBM_data_science_professional_certificate-/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb at main · LukaszCzeb/IBM_data_science_professional_certificate- (github.com)]

Results

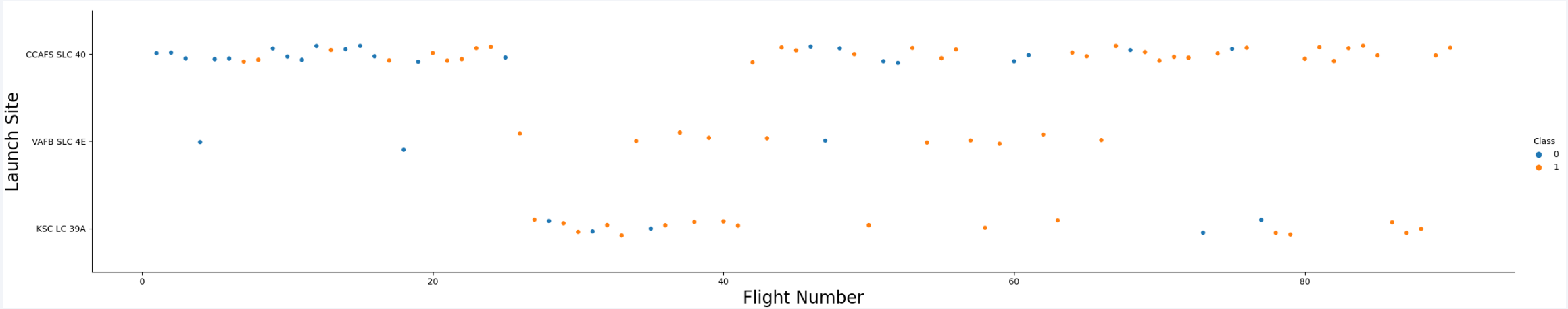
- Success rate of launching was increasing with time and now is on constant level.
- SO orbit has the worst rate of success among orbits.
- Having big payload can increase chance of failure
- Logistic Regression, KNN and SVM had best score and can predict success of launch with 83,3 % accurracy.





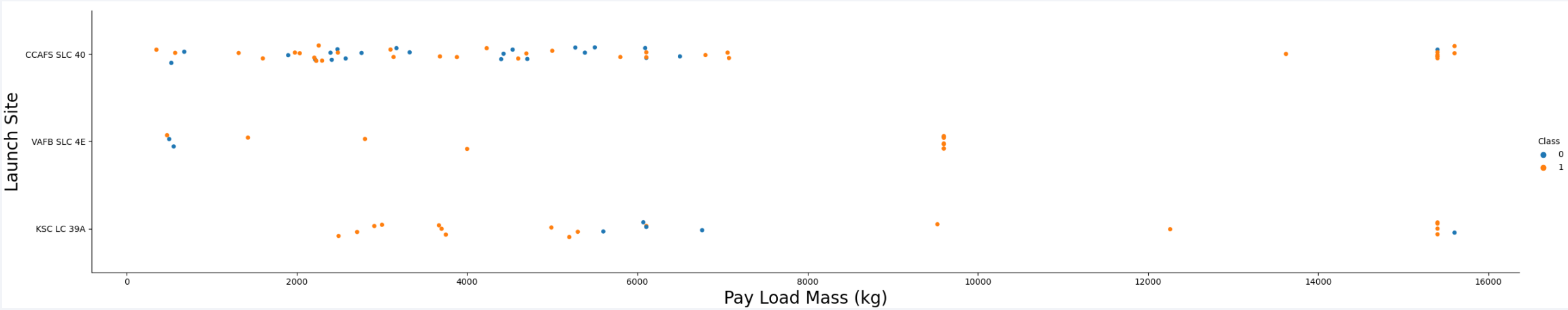

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



First launches had the biggest chances of failure, when they developed technology chance of success increased and in simmilar among diffetent launch sites.

Payload vs. Launch Site

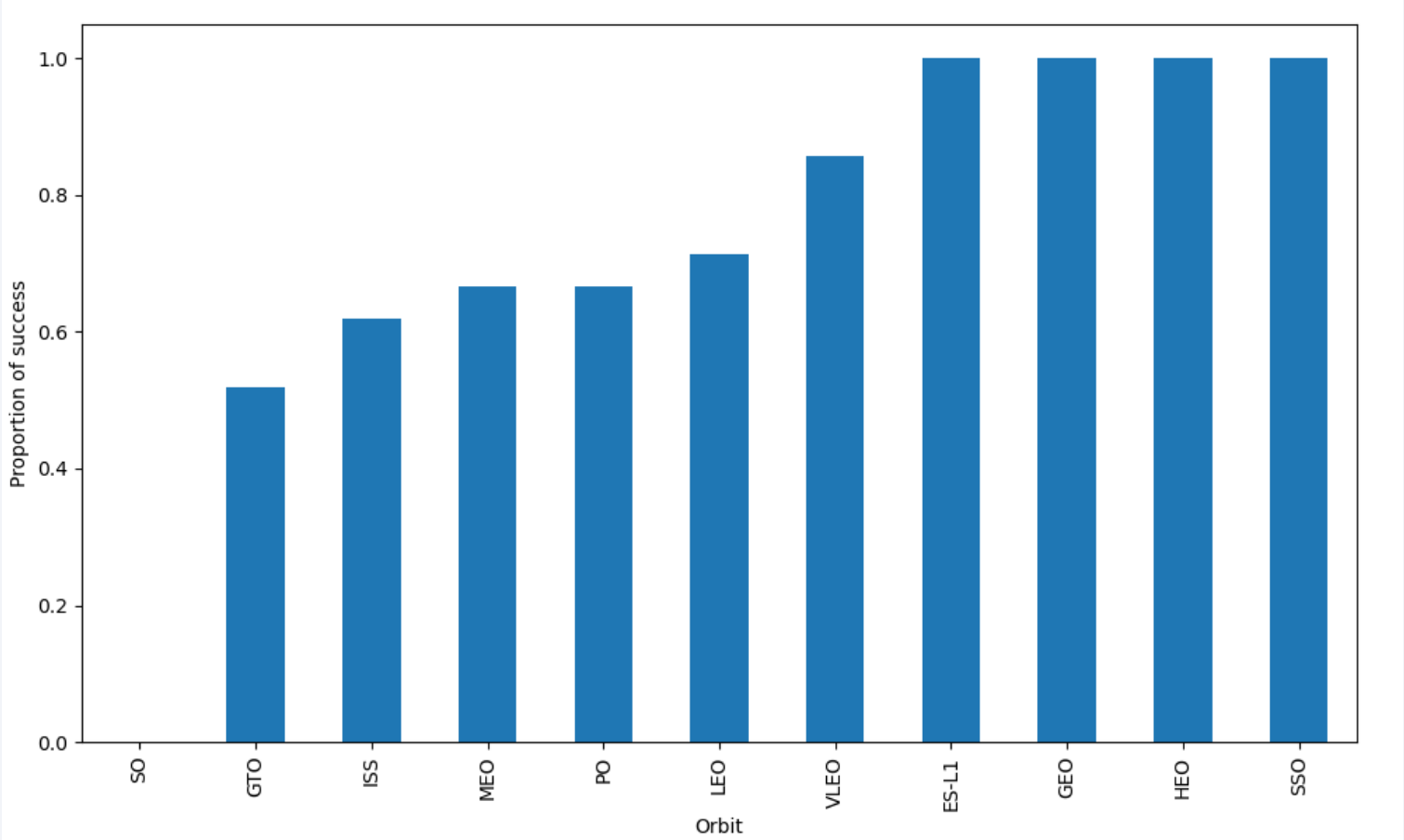


Increasing payload can increase failure chance. On VAFB SLC 4E site there was no bigger payload than 10000. Part of launches with failer and with small payload are becouse of that the first, most uncertain launches was with small payload.

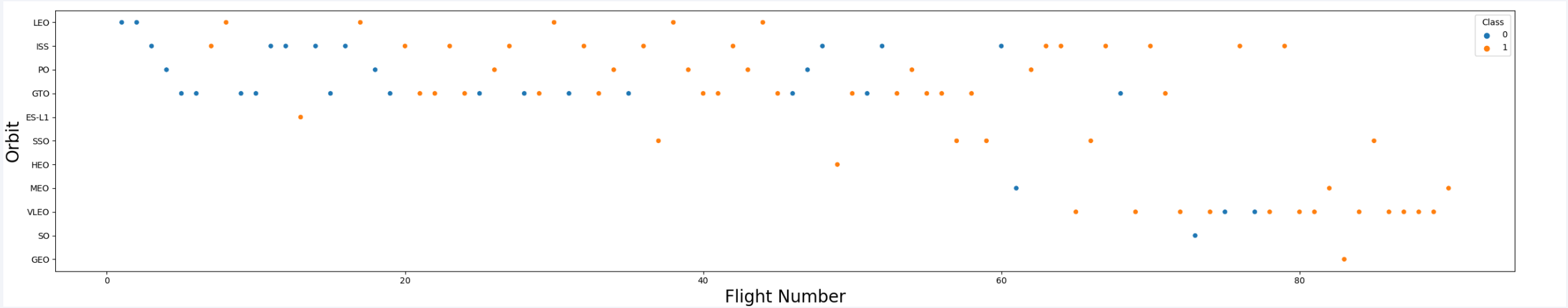
Success Rate vs. Orbit Type

ES-L1, GEO, HEO and SSO have the best success rate.

The worst success rate has SO orbit and GTO.

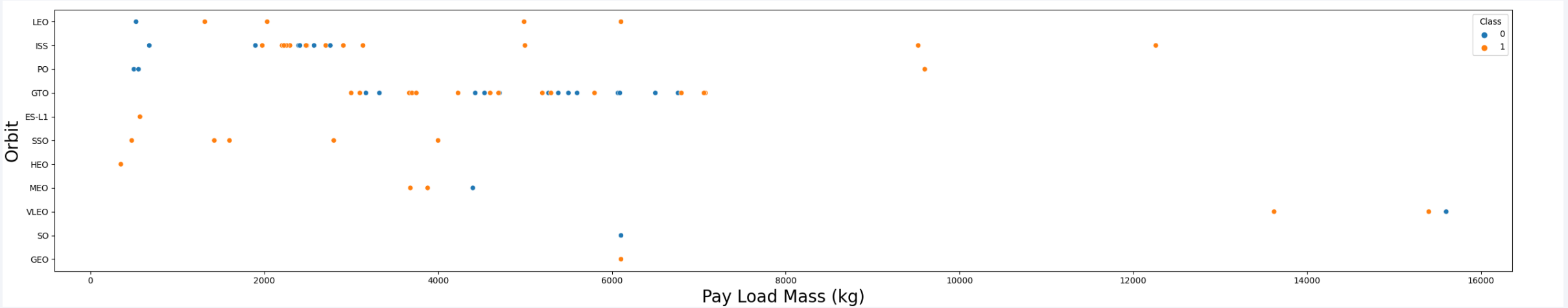


Flight Number vs. Orbit Type



Most of the orbits have small amont of Falcon 9 flights and becouse of that they can have very good or very bad success rate. The most flights was on LEO, ISS, PO, GTO orbits.

Payload vs. Orbit Type

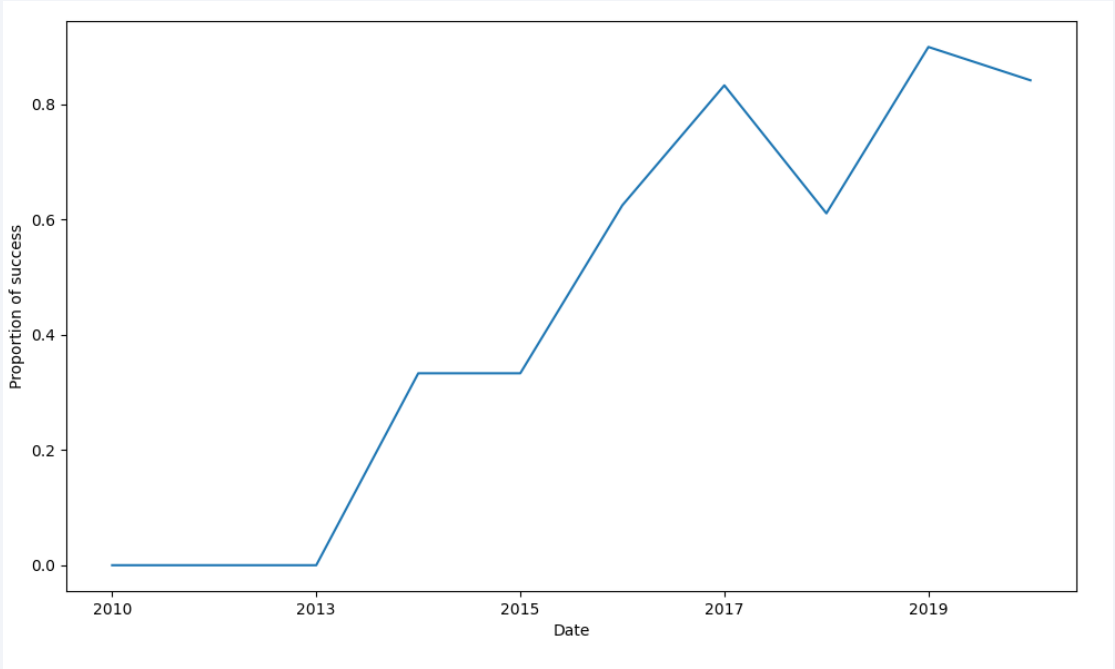


Most orbits had flights usually with medium payload. Only ISS, PO, VLEO had flights with the biggest payload.

Launch Success Yearly Trend

Proportion of success increased with time and now is on constant level.

Probably becouse they developed Falcon 9 technology and reach level when it is very hard to increase chance of success.



All Launch Site Names

```
SELECT DISTINCT(Launch_Site) FROM spacextbl
```

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

All Launch Site Names

Launch Site Names Begin with 'CCA'

```
SELECT *
FROM spacextbl
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

5 Launch Sites Names Begin with 'CCA'

Total Payload Mass

```
SELECT SUM(payload_mass__kg_)
FROM spacextbl
WHERE customer = 'NASA (CRS)'
```

SUM(payload_mass__kg_)
45596

Total Payload Mass

Average Payload Mass by F9 v1.1

```
SELECT AVG(payload_mass__kg_)
FROM spacextbl
WHERE booster_version LIKE 'F9 v1.1%'
```

AVG(payload_mass__kg_)
2534.6666666666665

Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

```
SELECT date
FROM spacextbl
WHERE "Landing _outcome" LIKE '%Success%ground pad%'
LIMIT 1
```

Date
22-12-2015

First Successful Ground Landing Date

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT booster_version, "landing _outcome", payload_mass__kg_
FROM spacextbl
WHERE ("landing _outcome" LIKE '%Success%drone ship%') AND (payload_mass__kg_ > 4000) AND (payload_mass__kg_ < 6000)
```

Booster_Version	Landing _Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Successful Drone Ship Landing with Payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

```
SELECT mission_outcome, COUNT(Mission_outcome)
FROM spacextbl
GROUP BY mission_outcome
```

Mission_Outcome	COUNT(Mission_outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Total Number of Successful and Failure Mission Outcomes

Boosters Carried Maximum Payload

```sql
SELECT booster_version, payload_mass__kg_
FROM spacextbl
WHERE payload_mass__kg_ IN (SELECT MAX(payload_mass__kg_) FROM spacextbl)
```

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

Boosters Carried Maximum Payload

2015 Launch Records

```
SELECT substr(Date, 4, 2) as month, "landing _outcome" , booster_version, launch_site
FROM spacextbl
WHERE (substr(Date,7,4) = '2015')  AND ("landing _outcome" LIKE '%Failure%drone ship%')
```

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

2015 Launch Records

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

```
SELECT "landing _outcome" ,COUNT("landing _outcome")
FROM spacextbl
WHERE date between '04-06-2010' AND '20-03-2017'
GROUP BY "landing _outcome"
ORDER BY COUNT("landing _outcome") DESC
```

Landing _Outcome	COUNT("landing _outcome")
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

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

Section 3

Launch Sites Proximities Analysis

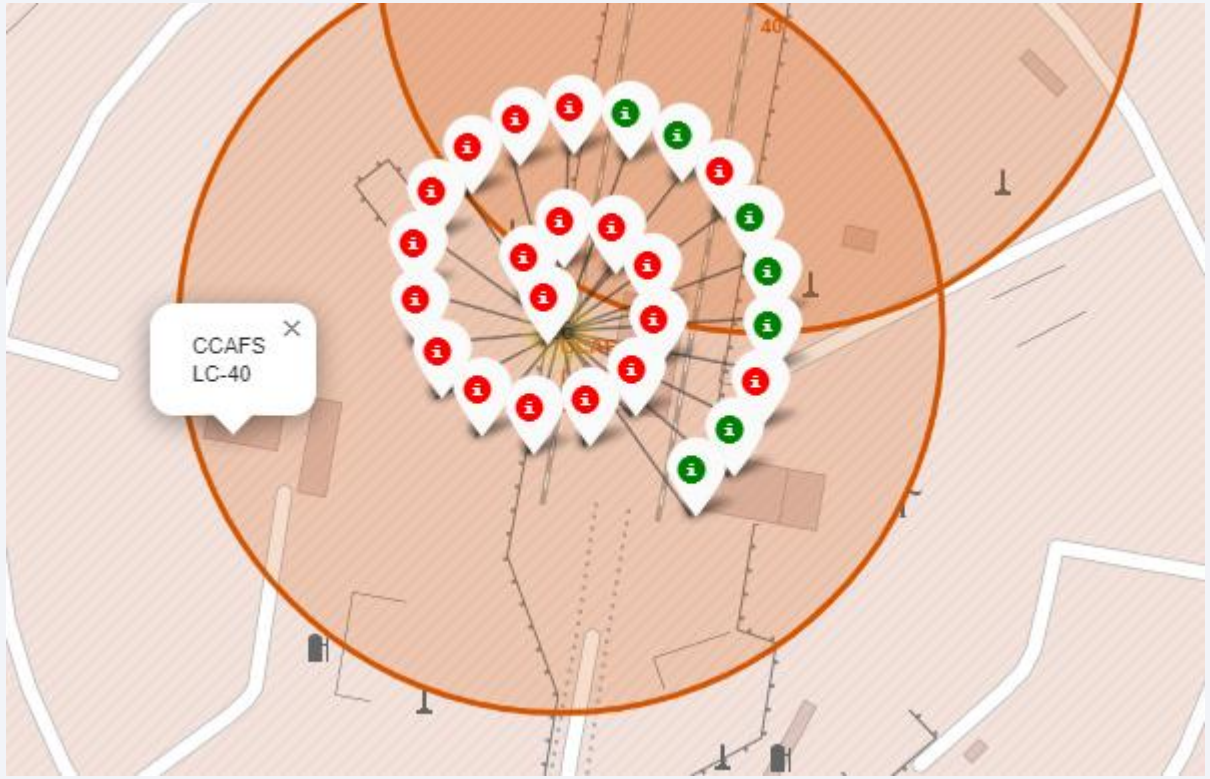
Launch Sites Location

All launch sites location are in USA, close to coastline and close to equator.

This location proviedes best weather, safety and logisctic conditions.

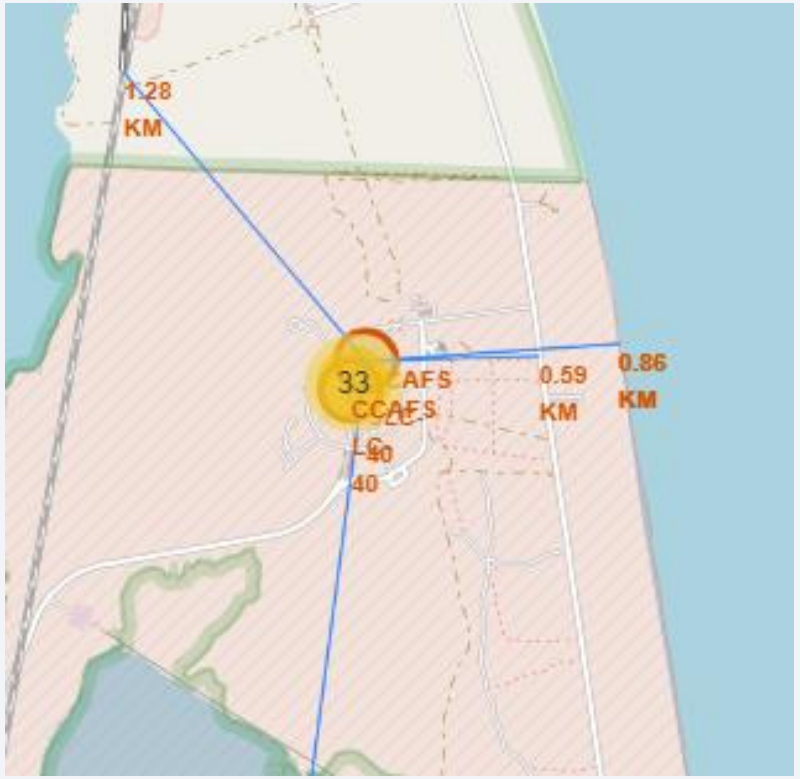
CCAFS LC-40 Launch outcomes map

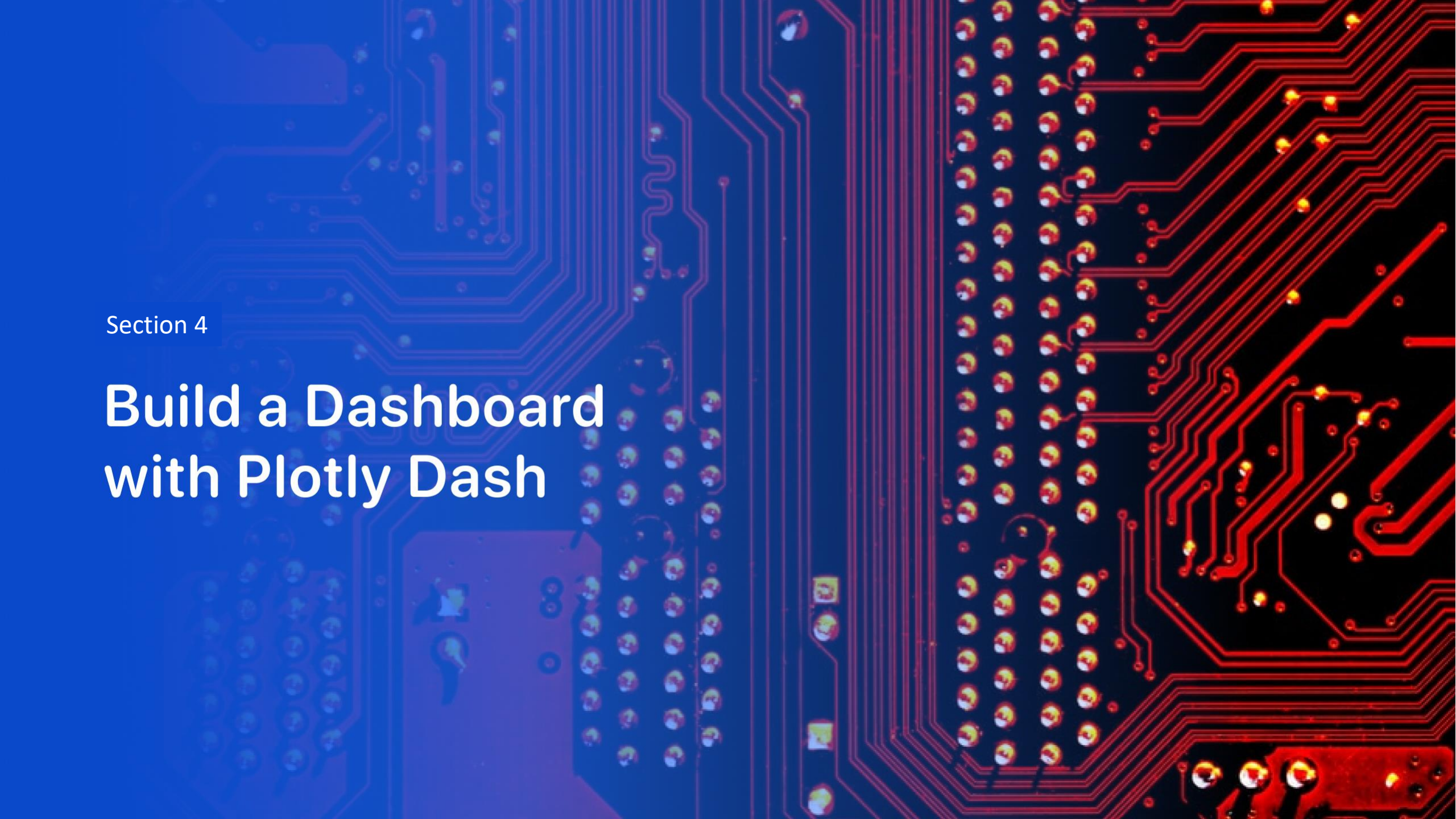
With launch outcomes markes we can easly check locaction of each launch site and see how many launch they have with success or failure outcomes.



CCAFS SLC-40 distances map

With distance to coastline, highway, railway and city we can see paterns what distances provide best condistion for launch site and provide safety for citizens.

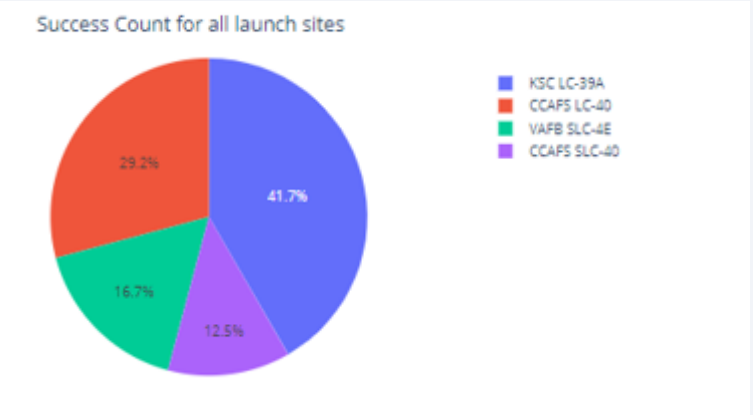



Section 4
Build a Dashboard with Plotly Dash

Success Count for all launch sites

Pie chart shows how many precentege of success from all success launches each launch site have.

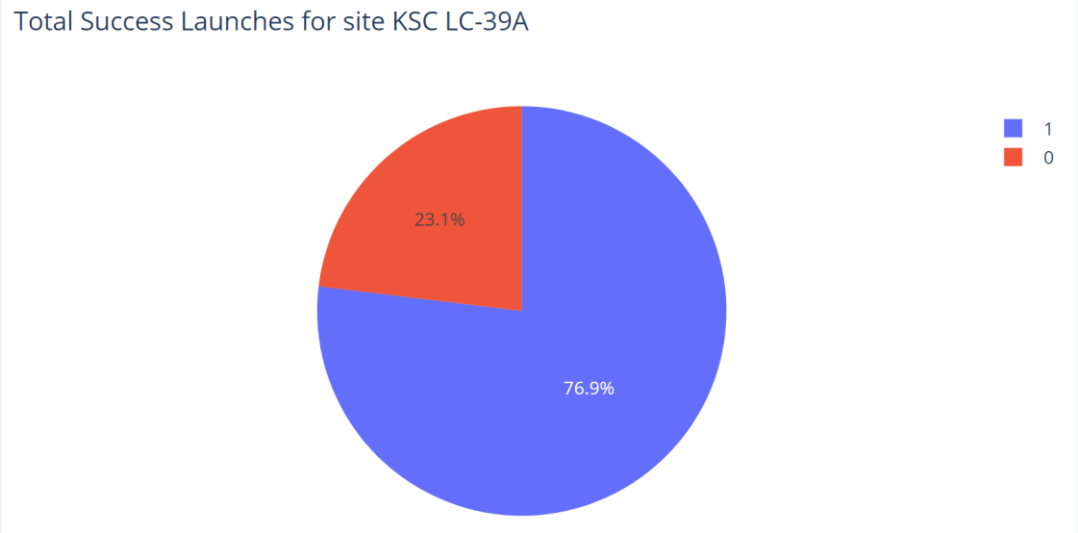
KSC LC-39A is responsible for biggest count of success launches (41,7%)



Total Success Launches for site KSC LC-39A

Pie chart shows prectentege of success launches for site KCS LC-39A.

We can see that 76,9% launches on this site was successful.



Correlation of Payload and Launch Outcome plots

Smaller Payloads has bigger chance of success than bigger payload (more than 5000 kg).

FT Booster Version has the best ratio success
From all booster versions.

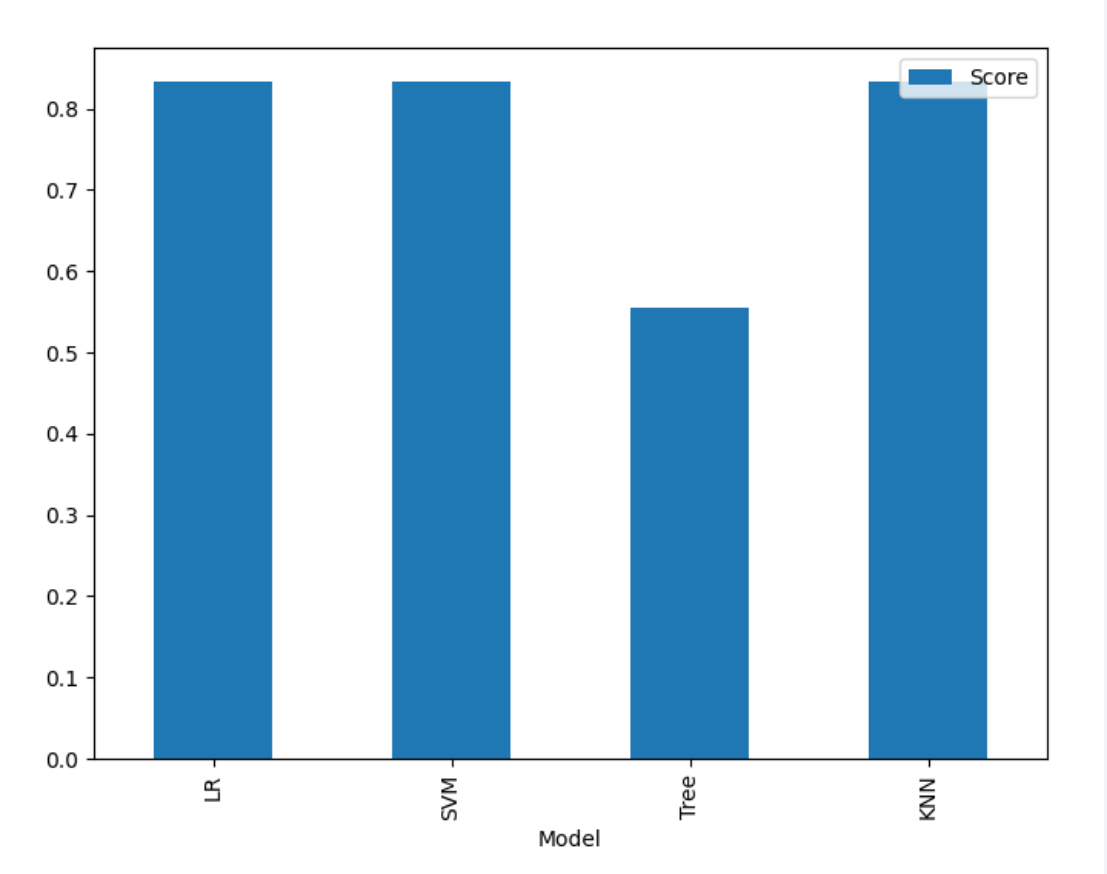
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Logistic Regression, SVM and KNN model have the best accuracy.

Decision Tree model has the lowest accuracy.

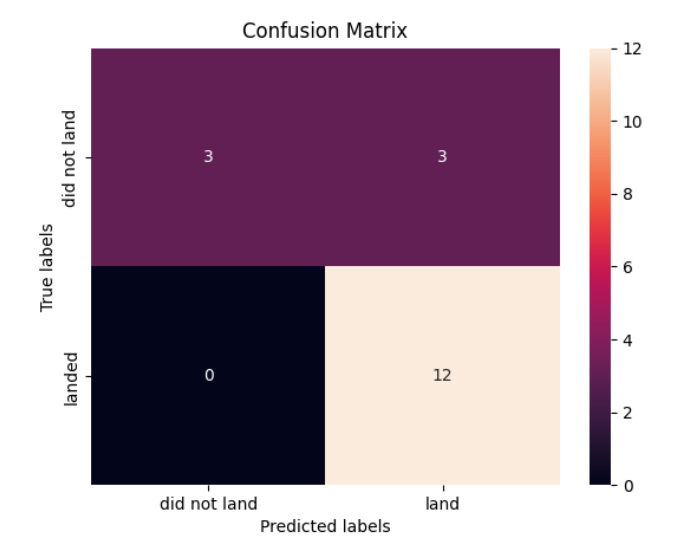


Confusion Matrix

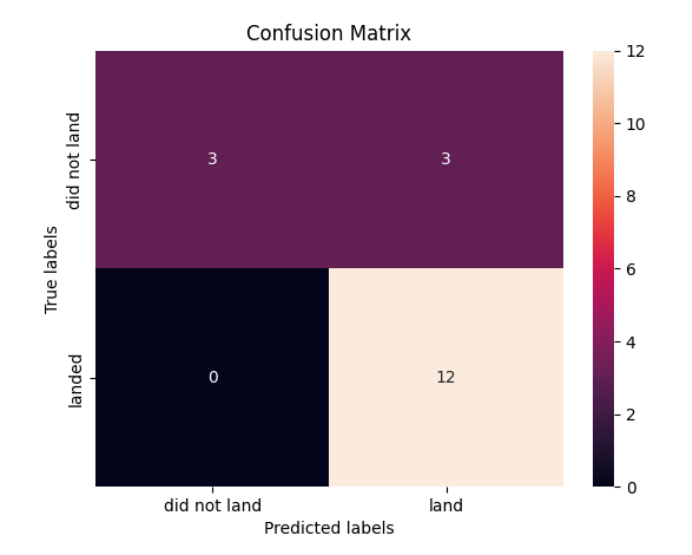
Three best models has confusion matrix with the same results.

Each model predict correctly 12 succesful landings and 3 unsuccesful landings.

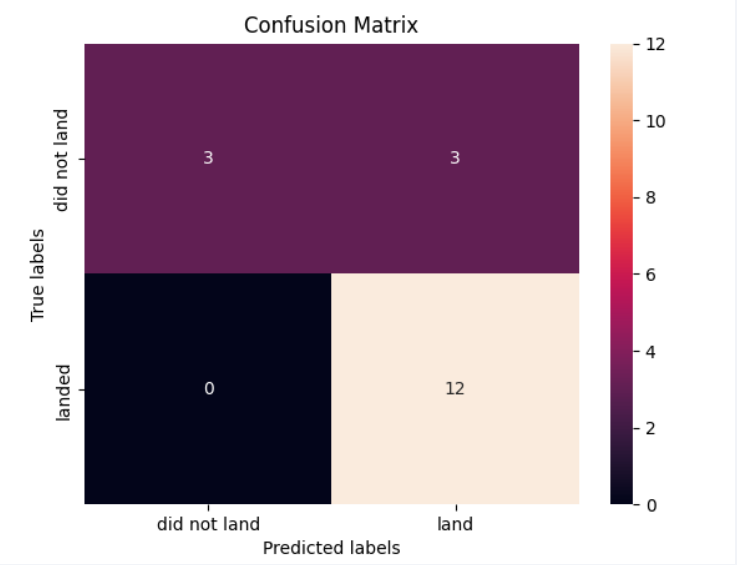
LR



SVM



KNN



Conclusions

- Success rate increased by the time and reach constant level on about 80%.
- Increasing payload weight increase chance of failure.
- Logistic Regression, SVM and KNN models have best accuracy to predict success launch of Falcon 9.
- Most successful site is KSC LC-39A.
- Most successful orbits are ES-L1, GEO, HEO and SSO, but because of small amount of launches for this orbits, their successful rates are very uncertain.

Appendix

I created additonal model. Random Forest Classifier get the same score like best models, so it not increased accuracy and because of longer time of computing there is better to use simpler models than Random Forest.

```
from sklearn.ensemble import RandomForestClassifier
```

```
parameters = {'criterion': ['gini', 'nationality'],
     'max_depth': [2*n for n in range(1,10)],
     'max_features': ['auto', 'sqrt'],
     'min_samples_split': [2, 5, 10]}

random = RandomForestClassifier()
```

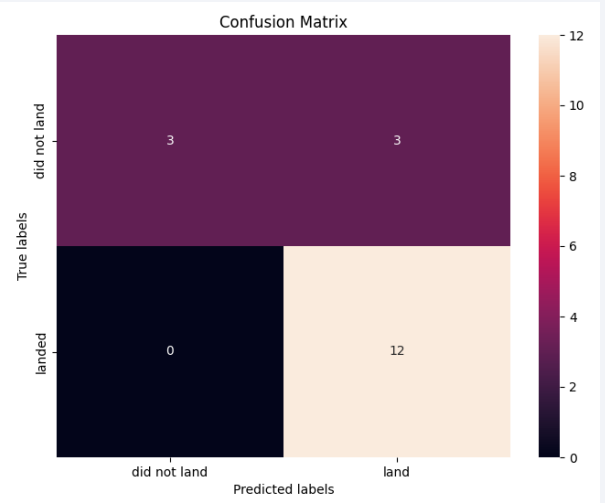
```
random_cv = GridSearchCV(random, parameters)
random_cv.fit(X_train, Y_train)
```

```
print("tuned hpyerparameters :(best parameters) ",random_cv.best_params_)
print("accuracy :",random_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth': 2, 'max_features': 'auto', 'min_samples_split': 2}
accuracy : 0.8476190476190476
```

```
Randomscore = random_cv.score(X_test, Y_test)
Randomscore
```

```
0.8333333333333334
```



Thank you!