IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

David Santiago Martinez
9 august 2024

Outline

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

Executive Summary

- In this project we try to predict if rocket of Space Y can landing successfully or not in a platform having in mind a lot of variables like weight, orbit, booster version and so on. We collect information using web scrapping and using API's, handling the missing values with pandas and searching information in the data set with SQL commands, after that we compare the different classification models (KNN, SVM, Tree & log regression) with Accuracy Score to find that the best Model is tree model and the orbits with more successful rate are ES-L1, GEO, HEO, SSO

Introduction

- We will predict if the Falcon 9 first stage will land successfully. 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. 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 SpaceX for a rocket launch.

Section 1

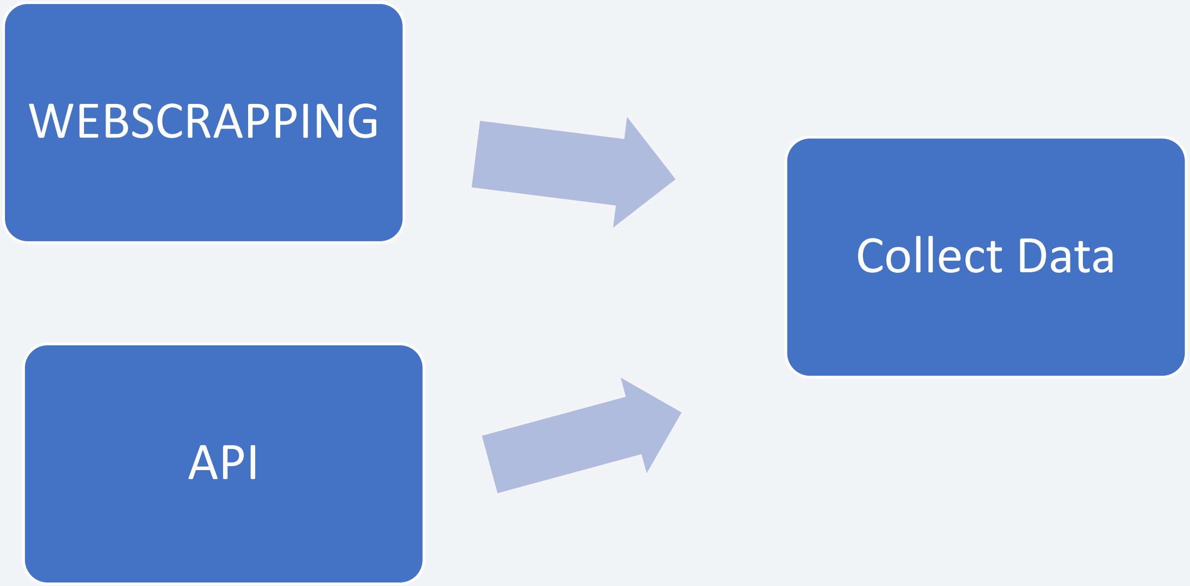
Methodology

Methodology

- In this project we try to predict if rocket of Space Y can landing successfully or not in a platform having in mind a lot of variables like weight, orbit, booster version and so on. We collect information using web scrapping and using Space X API's, after that handling the missing values with pandas and searching information in the data set with SQL commands to find some answers, after that wedivide the sample in two, one for training the models and one for testing the models, we compare the different classification models (KNN, SVM, Tree & log regression) with Accuracy Score to find that the best Model and using confusion matrix to answer our questions

Data Collection

- The data were collected in two ways, the first one was with webscrapping on Wikipedia and the second one was with the Space x API

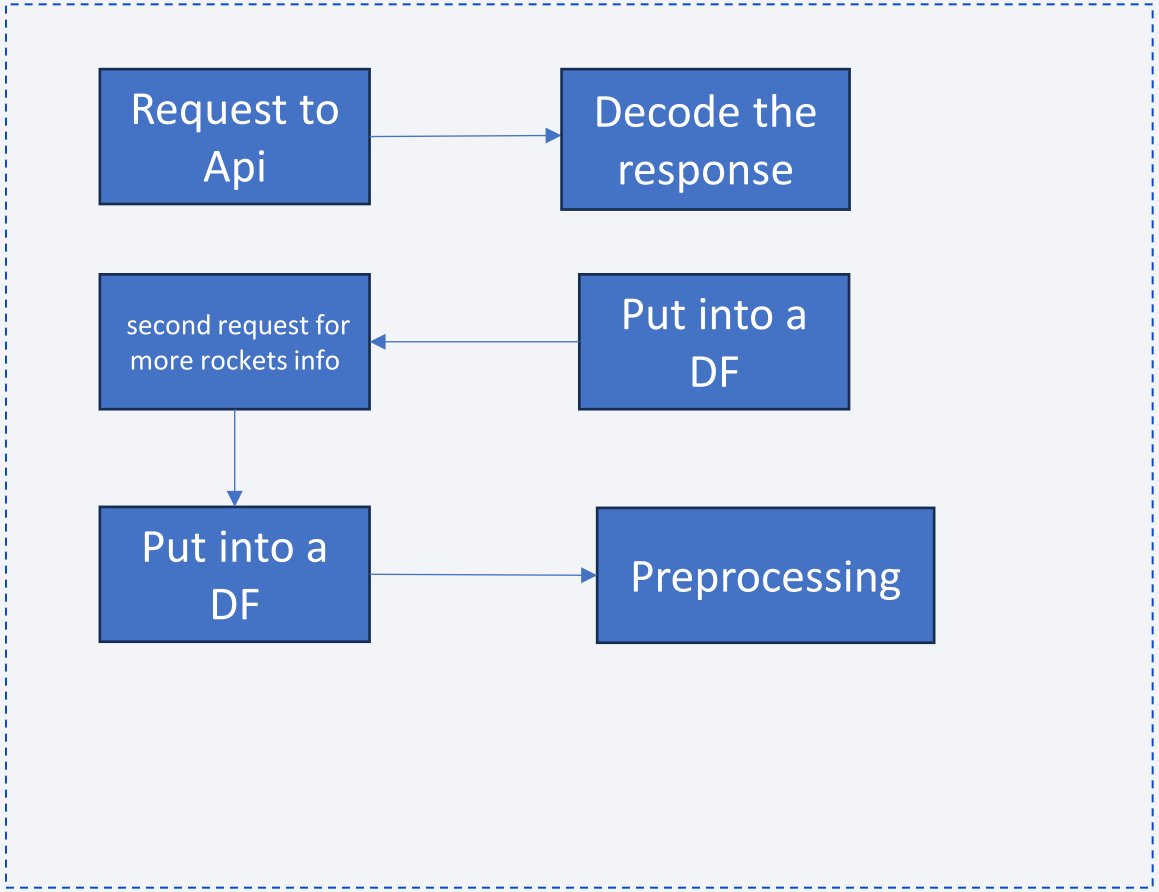


Data Collection – SpaceX API

- The process to obtain the info with SpaceX API was:

Link on GitHub :

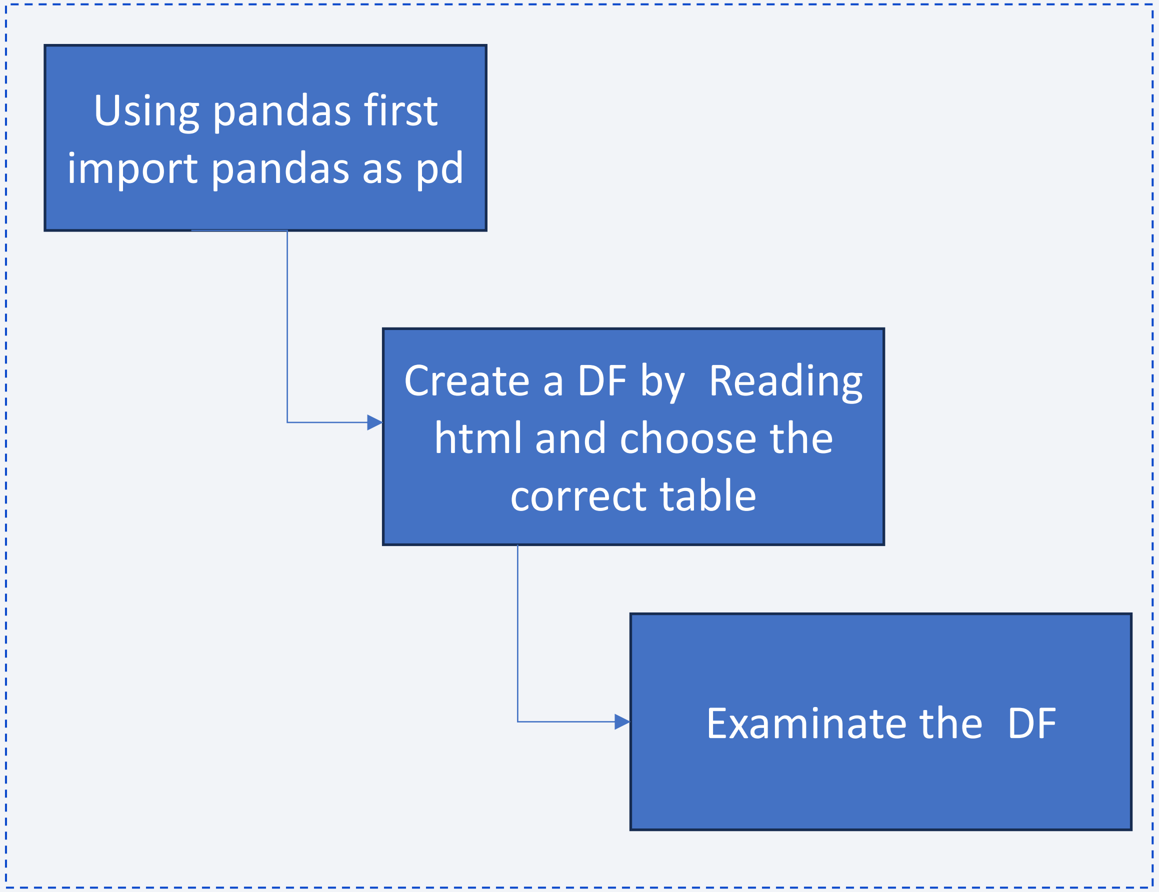
- https://github.com/santiago2025/Capstone-Project/blob/a198dda2f61ae1181ff904c851228e88fbd61277/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

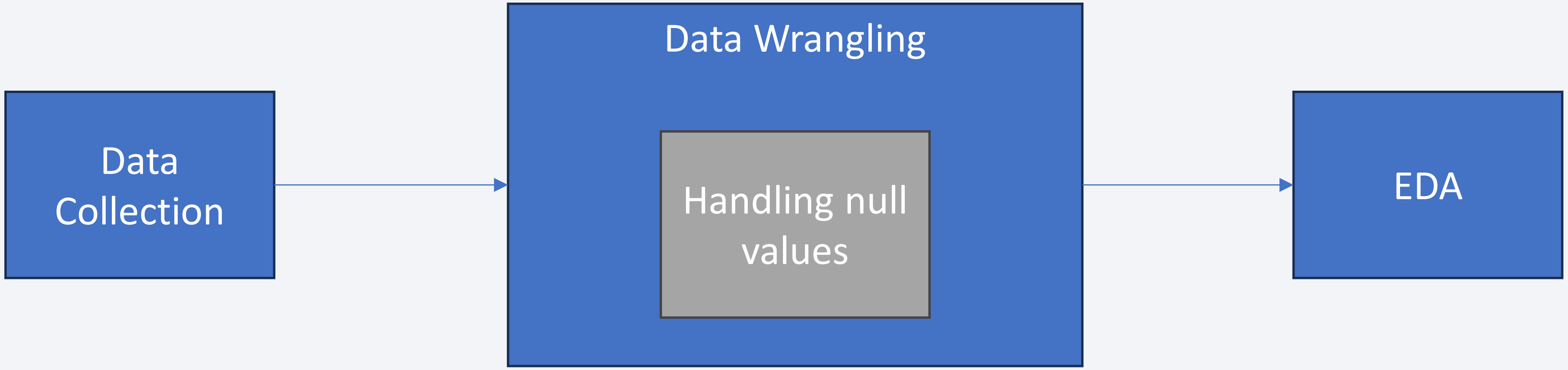
- For WebScrapping we use The Wikipedia page and we Use the option from pandas to do this job with the function pd.read_html

- Git hub link: https://github.com/santiago2025/Capstone-Project/blob/d59e4890529b0403003abdcef811136a6bf35b8e/jupyter-labs-webscraping.ipynb



Data Wrangling

- Int this project we deal with some null values, and we changed by the mean, because we can't find the real values and doesn't affect the model
- Link: https://github.com/santiago2025/Capstone-Project/blob/a198dda2f61ae1181ff904c851228e88fbd61277/jupyter-labs-spacex-data-collection-api.ipynb



EDA with Data Visualization

- We use Scatter plots to visualize the relations between some variables, the blue points show the flights that not landing, and the orange points the flights that landing successfully in the different launch sites

- Link:https://github.com/santiago2025/Capstone-Project/blob/0e104e20bd7b8dca502b927720c19274f2f9b15e/edadataviz.ipynb

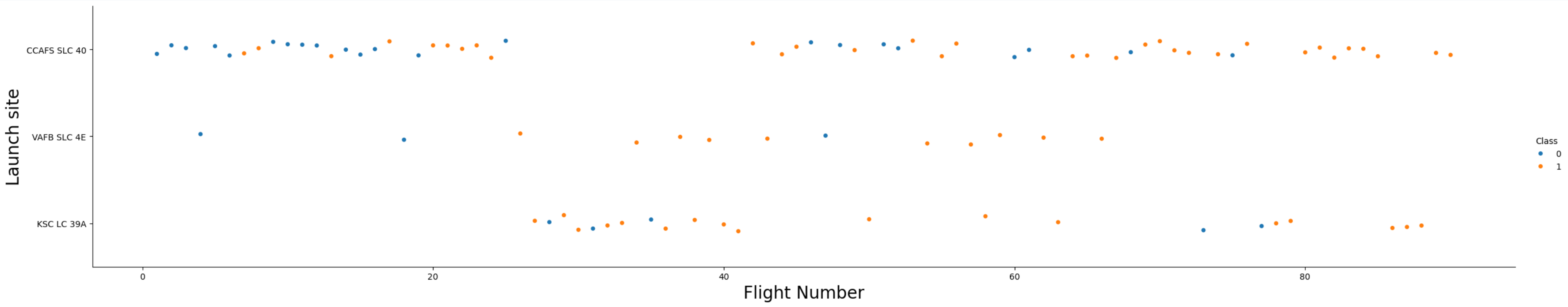


Fig1: Launch site vs Flights

EDA with Data Visualization

- We use Scatter plots to visualize the relations between some variables, in this case the Launch site vs Payload Mass in Kg
- Link:https://github.com/santiago2025/Capstone-Project/blob/0e104e20bd7b8dca502b927720c19274f2f9b15e/edadataviz.ipynb

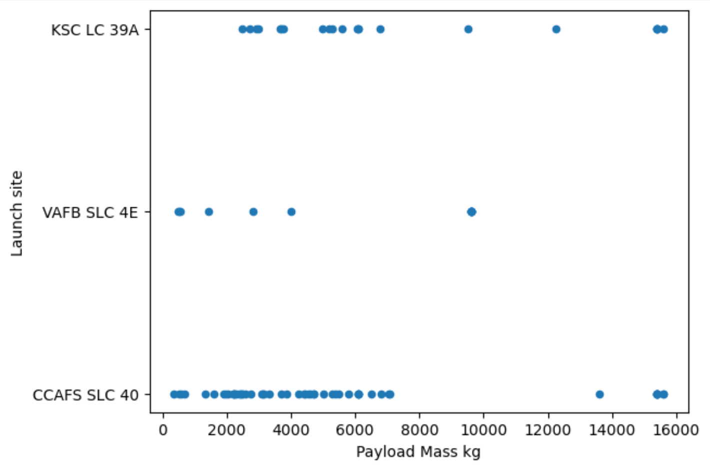


Fig1: Launch site vs Flights

EDA with Data Visualization

- We use bar plot to show the successful rate for each orbit flight
- Link:https://github.com/santiago2025/Capstone-Project/blob/0e104e20bd7b8dca502b927720c19274f2f9b15e/edadataviz.ipynb

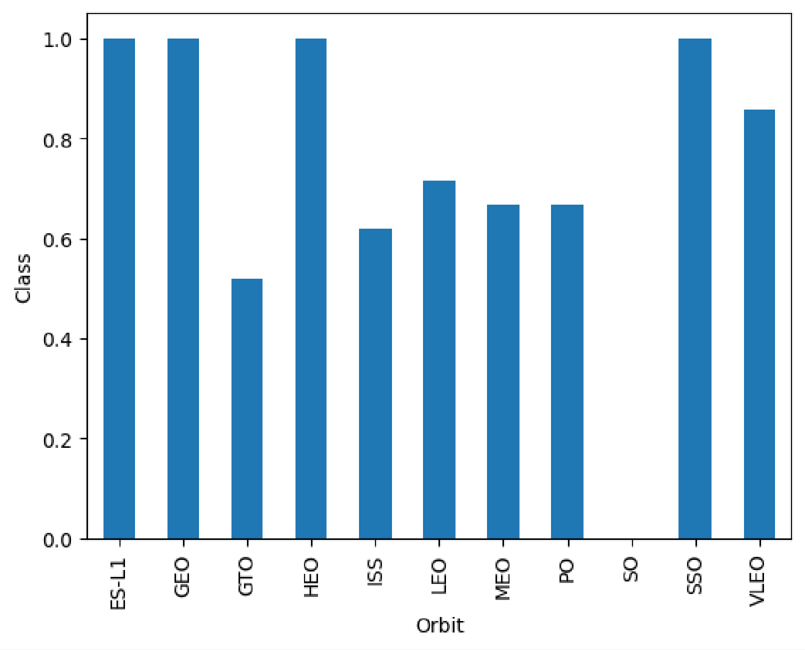


Fig3: succesfull rate for each orbit

EDA with SQL

- First using SQL made the consultant list on the next list:
- Link:https://github.com/santiago2025/Capstone-Project/blob/d804a2ba855fa2bc11ba0aedad0a7c5358f6acdd/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- ✓ Find the names of the launch sites
- ✓ Display th total Payload Mass carried by Boosters Launched by NASA
- ✓ Display average payload mass carried by booster version F9 v1,1
- ✓ 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 records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

Predictive Analysis (Classification)

- After all visualizations decide to use one Classification model, because we need to classify labeled data and need to find the best model, for that reason I decide to train different models and compare with the accuracy score

- https://github.com/santiago2025/Capstone-Project/blob/a7ffff53f79203e9096afc5143b510b250b457bf/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Normalize data → Split the data into test and train → Choose the model to evaluate (tree model, logistic regression, knn) → Train the model and made a gridsearch to find the best parameters → Find the accuracy of each model and compare

Results

- The best model to predict the data set is knn model with an accuracy score of 0,848
- The parameters of this model are the followings: neighbours number=10, ‘p’=1

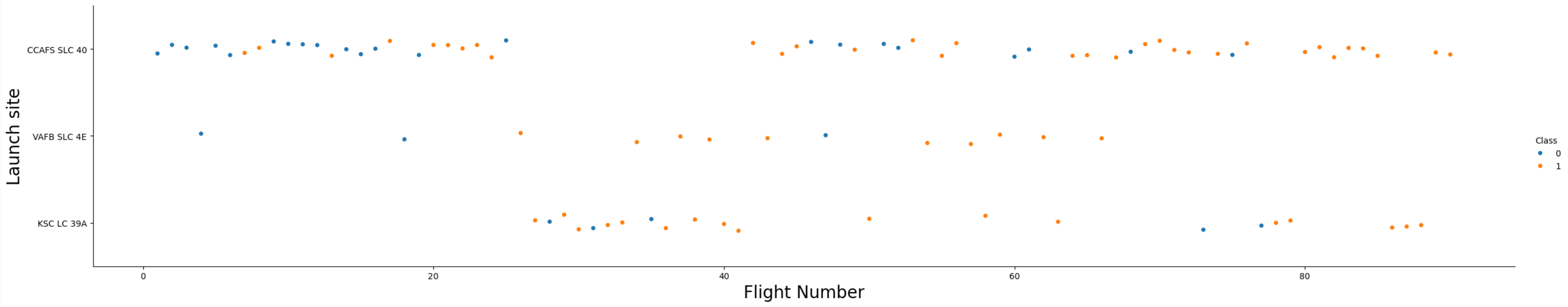
- https://github.com/santiago2025/Capstone-Project/blob/a7ffff53f79203e9096afc5143b510b250b457bf/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

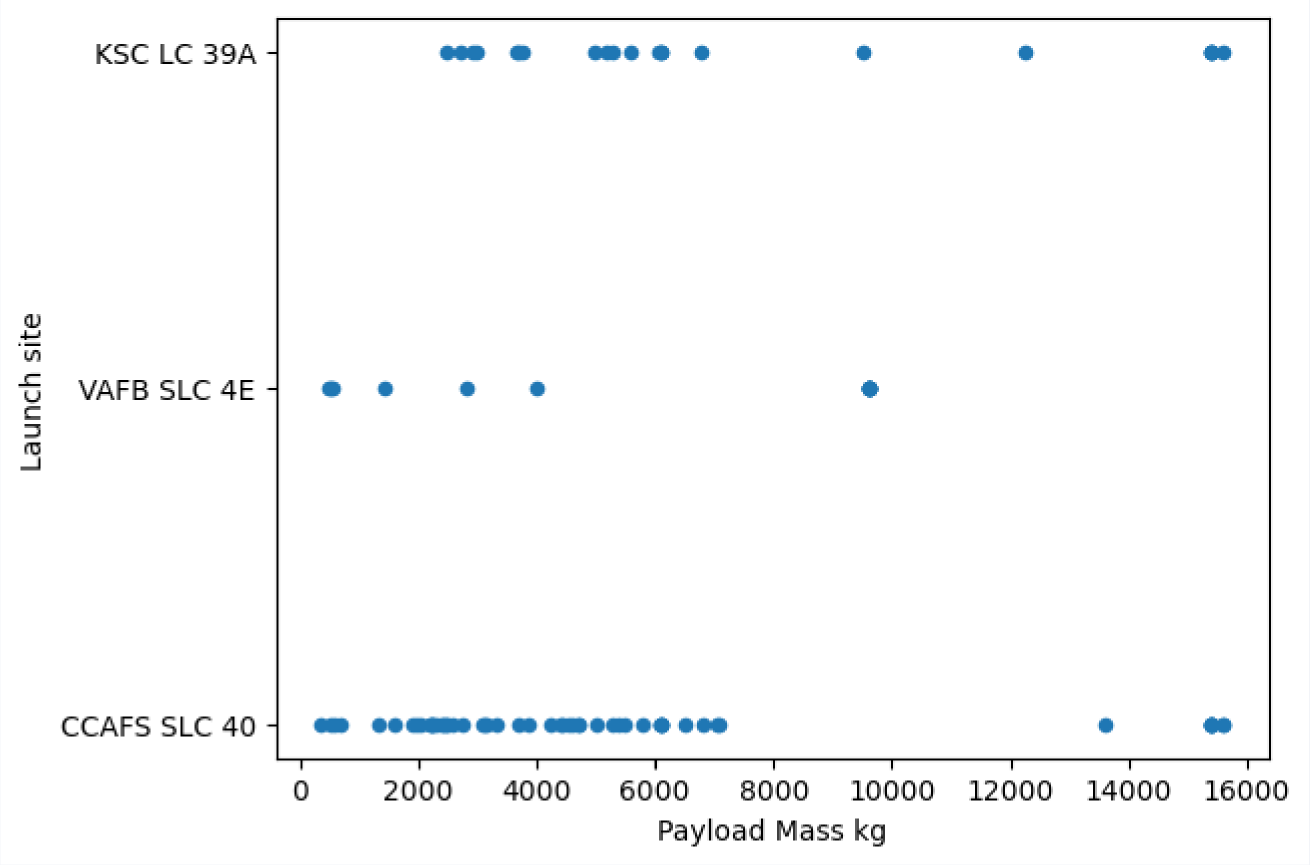
- Flight Number vs. Launch Site
- Where blue dots are successful landings, and the orange are unsuccessful landings.
- Could see in CCAFS SLC 40 were more launches



Payload vs. Launch Site

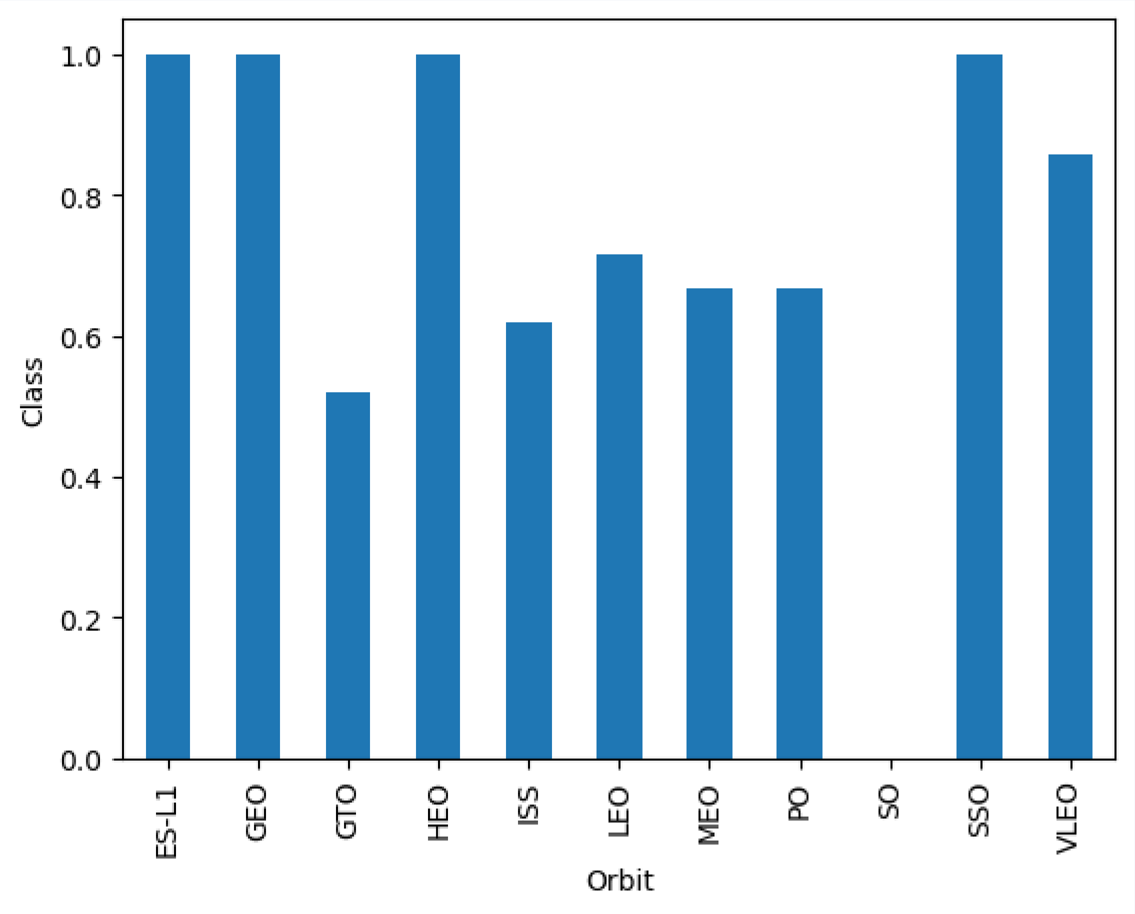
- Payload vs. Launch Site

- In this graph we can observe all launchs on different platforms, and we can observe there are not many heavy lunches above the 10,000 kg



Success Rate vs. Orbit Type

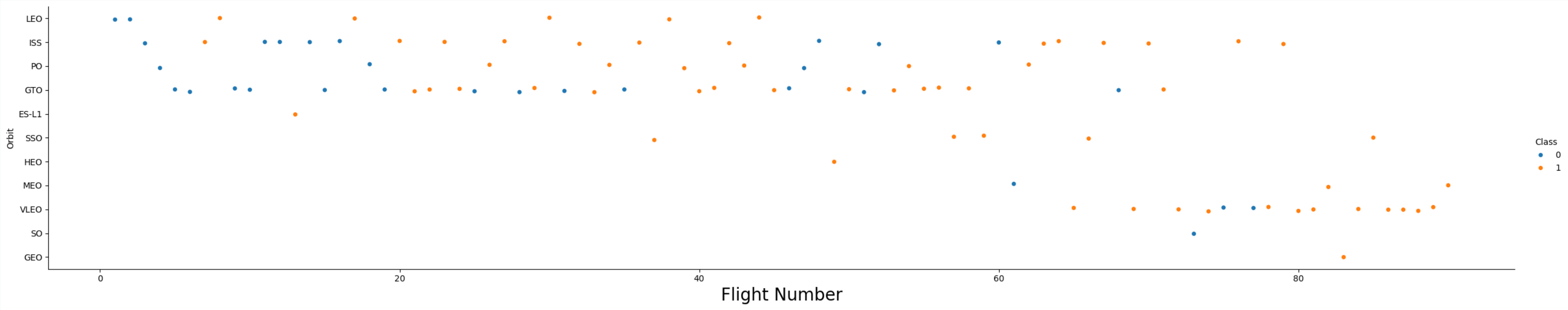
- Success rate of each orbit type
- We can observe there are 4 orbits with a 100% successful rate (ES-L1, GEO, HEO, SSO)
- There is no launches on SO and the worst successful rate i



Flight Number vs. Orbit Type

- Flight number vs. Orbit type where the orange points are successful and the blue ones unsuccessful

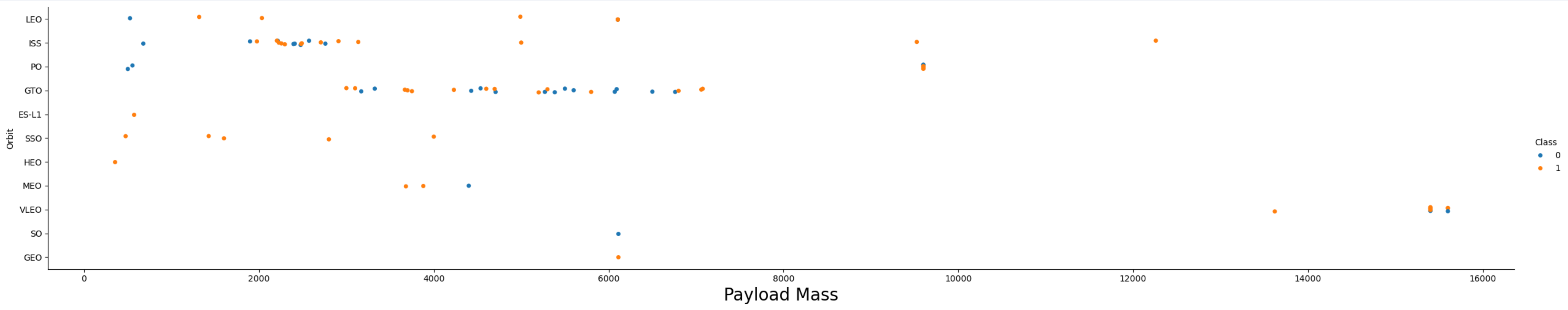
- You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



Payload vs. Orbit Type

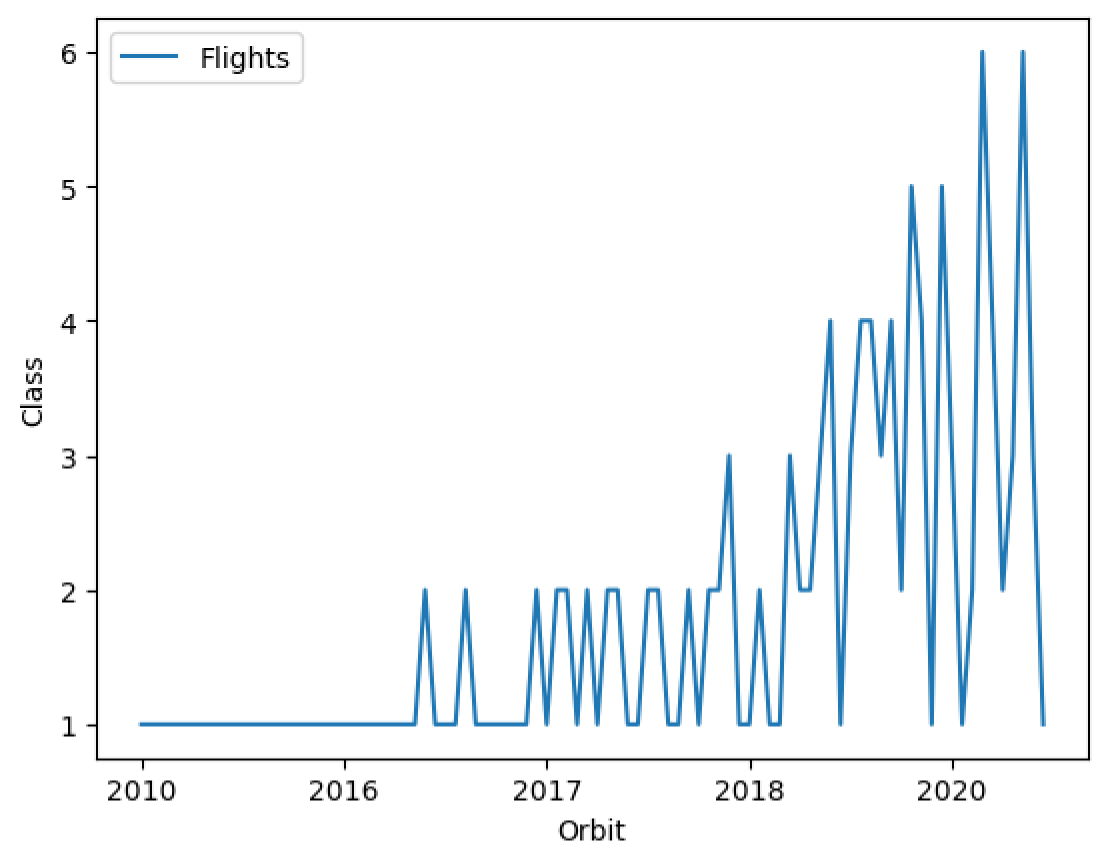
- Payload vs. Orbit type where the orange points are successful and the blue ones unsuccessful

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- Success yearly landing trends
- We can observe after 2018 increase the number of successful landings



All Launch Site Names

Find the names with a SQL consult using function distinct

```
%%sql
select  Distinct "Launch_site"  from SPACEXTABLE
```

Launch Sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- Using SQL commands find 5 site names that start witt CCA

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

```
%%sql
select * from SPACEXTABLE 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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Calculate the total payload carried by boosters from NASA
- We sum the payload carried by booter from NASA in SQL

```
%%sql
select sum(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'
```

PAYLOAD MASS KG total
45.596

Average Payload Mass by F9 v1.1

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

```
%%sql
select avg(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version" Like 'F9 v1.1%'
```

AVG PAYLOAD MASS KG
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad

Date	Time (UTC)	Booster Version	Launch Site	Payload	PAYLOAD MASS Kg	Orbit	Customer	Mission Outcome	Landing Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

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

Booster Version	PAYLOAD MASS Kg	Landing_Outcome
F9 FT B1022	4.696	Success (drone ship)
F9 FT B1026	4.600	Success (drone ship)
F9 FT B1021.2	5.300	Success (drone ship)
F9 FT B1031.2	5.200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Succesfull	Failure
61	40

Boosters Carried Maximum Payload

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

Booster Version	Date	Time (UTC)	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission Outcome	Landing Outcome
F9 B5 B1048.4	2019-11-11	14:56:00	CCAFS SLC-40	Starlink 1 v1.0, SpaceX CRS-19	15600	LEO	SpaceX	Success	Success
F9 B5 B1049.4	2020-01-07	2:33:00	CCAFS SLC-40	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600	LEO	SpaceX	Success	Success
F9 B5 B1051.3	2020-01-29	14:07:00	CCAFS SLC-40	Starlink 3 v1.0, Starlink 4 v1.0	15600	LEO	SpaceX	Success	Success
F9 B5 B1056.4	2020-02-17	15:05:00	CCAFS SLC-40	Starlink 4 v1.0, SpaceX CRS-20	15600	LEO	SpaceX	Success	Failure
F9 B5 B1048.5	2020-03-18	12:16:00	KSC LC-39A	Starlink 5 v1.0, Starlink 6 v1.0	15600	LEO	SpaceX	Success	Failure
F9 B5 B1051.4	2020-04-22	19:30:00	KSC LC-39A	Starlink 6 v1.0, Crew Dragon Demo-2	15600	LEO	SpaceX	Success	Success
F9 B5 B1049.5	2020-06-04	1:25:00	CCAFS SLC-40	Starlink 7 v1.0, Starlink 8 v1.0	15600	LEO	SpaceX, Planet Labs	Success	Success
F9 B5 B1060.2	2020-09-03	12:46:14	KSC LC-39A	Starlink 11 v1.0, Starlink 12 v1.0	15600	LEO	SpaceX	Success	Success
F9 B5 B1058.3	2020-10-06	11:29:34	KSC LC-39A	Starlink 12 v1.0, Starlink 13 v1.0	15600	LEO	SpaceX	Success	Success
F9 B5 B1051.6	2020-10-18	12:25:57	KSC LC-39A	Starlink 13 v1.0, Starlink 14 v1.0	15600	LEO	SpaceX	Success	Success
F9 B5 B1060.3	2020-10-24	15:31:34	CCAFS SLC-40	Starlink 14 v1.0, GPS III-04	15600	LEO	SpaceX	Success	Success
F9 B5 B1049.7	2020-11-25	2:13:00	CCAFS SLC-40	Starlink 15 v1.0, SpaceX CRS-21	15600	LEO	SpaceX	Success	Success

2015 Launch Records

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

- Present your query result with a short explanation here

month	Date	Time (UTC)	Booster Version	Launch Site	Payload	PAYLOAD MASS KG	Orbit	Customer	Mission Outcome	Landing Outcome
01	2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
04	2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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

```
%%sql
select count(Landing_outcome), * from SPACEXTABLE  where Landing_outcome not like 'Success%' and substr(Date,0,5)<='2017' order by date asc
```

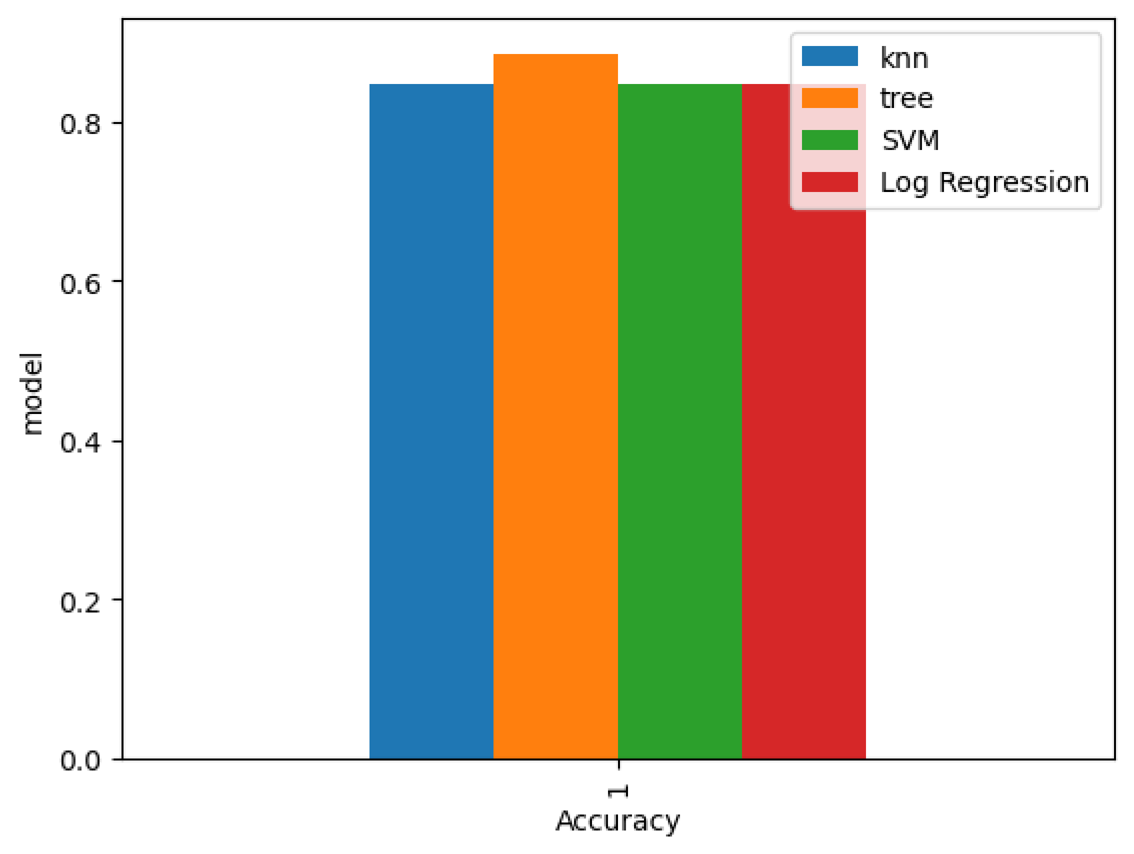
count(Landing_outcome)	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
26	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)

Section 3

Predictive Analysis (Classification)

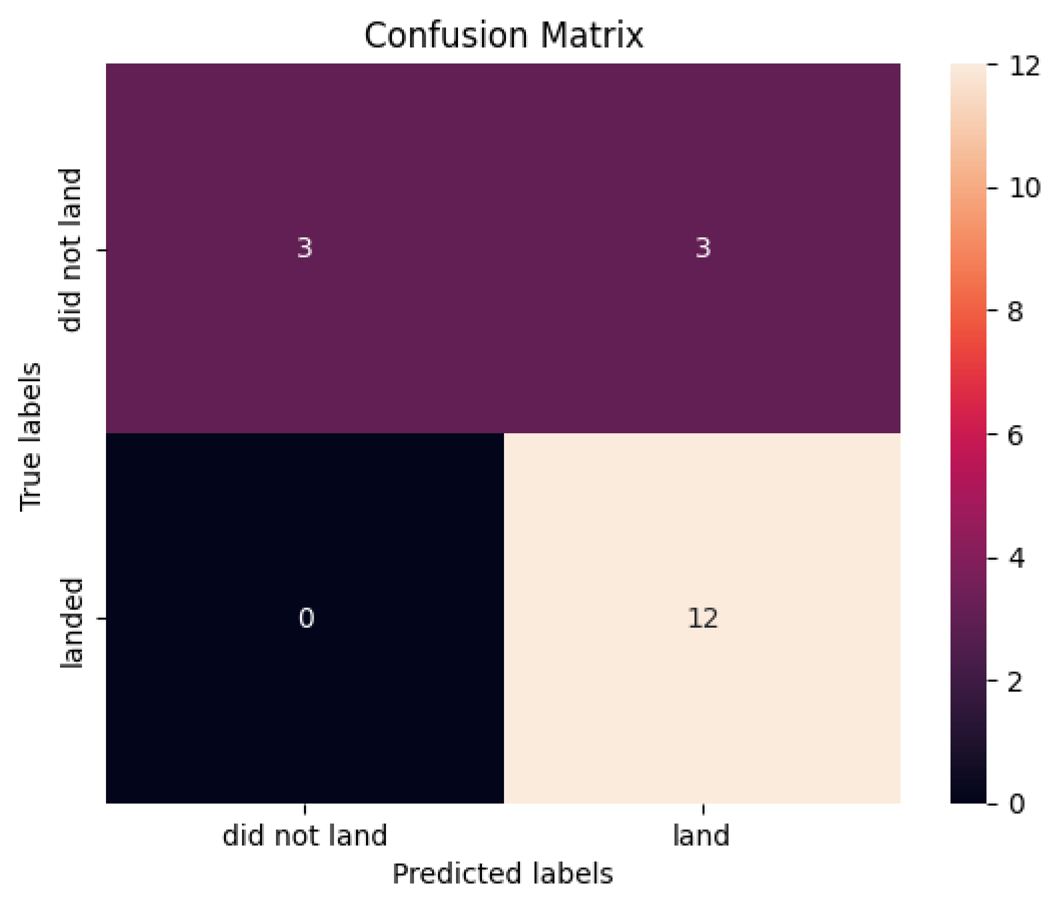
Classification Accuracy

- Visualization the built model accuracy for all built classification models, in a bar chart
- The model with the highest accuracy is Tree model



Confusion Matrix

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



Conclusions

- The best model to predict if a landing going to be successfully is tree model
- Sql help us to find some special cases in our data set
- We can observe there are 4 orbits with a 100% successful rate (ES-L1, GEO, HEO, SSO)

Thank you!