IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kyar Hsing Ma
2022-12-26

Outline

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

Executive Summary

- In this project, we will use the SpaceX APIs historical launch data to predict whether the first stage of the rocket will land successfully or not. SpaceX has a much cheaper offer to launch rockets for customers due to its reusable first stage. This data insight from this project might be useful for competitors.
- Python programming language and its related suitable libraries, query language SQL are used in this project.
- Requests, Pandas, Numpy libraries are used for data collection and wrangling, matplotlib, folium, plotly are used for static and interactive visual analysis, and SQL is used to query relational database.
- For machine learning model, sckit-learn is the main library applied in this project

Executive Summary

- From analysis, we found that different launch sites have different success rates; CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%, while CCAFS has a highest launch number, accounting for 2/3 of launches
- The success rates of first stage landing depends on its payload mass; heavy payload mass over 10000kg has the highest success rate of all, followed by payload between 3000 and 4000 kg
- Orbit types of PO, LEO and ISS also have positive relationship with Heavy payloads on success rates. Flight number also have a profound effect on the success rates.
- The success rates have increased since 2015 and the first successful ground landing was achieved in 2017

Executive Summary

- From geographical data, launch sites for SpaceX are mainly located in Florida. All launch sites are in close proximity with coastline. They are also close to railways and highways. However, they have a certain distance from populated cities.
- Four classification models are created. However, they perform almost the same on our test data sets, while Decision Tree has the highest accuracy score for training set.

Introduction

- Project background

  In this project, we will predict whether the Falcon 9 first stage will land successfully. Compared to other provided, SpaceX announced rocket launches fee were nearly 3/5 lower_ 62 millions for SpaceX vs 165 millions upwards by others. This is because the first stage of rockets SpaceX can be reused if it has landed successfully. We will predict this based on historical data provided by SpaceX via several APIs. This information can be useful for other competitors to bid against SpaceX.

- Problems to find answers
  - Will the first stage of a particular rocket land successfully?
  - What might be the possible causal factors for a successful landing?
  - What might be the relationship of each causal factors to each launch?
  - Will the station proximities affect the success of landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launches data are collected via a few RestAPIs using Python requests libraries
- Perform data wrangling
  - The collected data is transformed in Pandas data frames which are then studied, cleaned and 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
  - We will use scikit-learn library and build SVM, Classification Trees and Logistic Regression and find the best one for prediction.

Data Collection

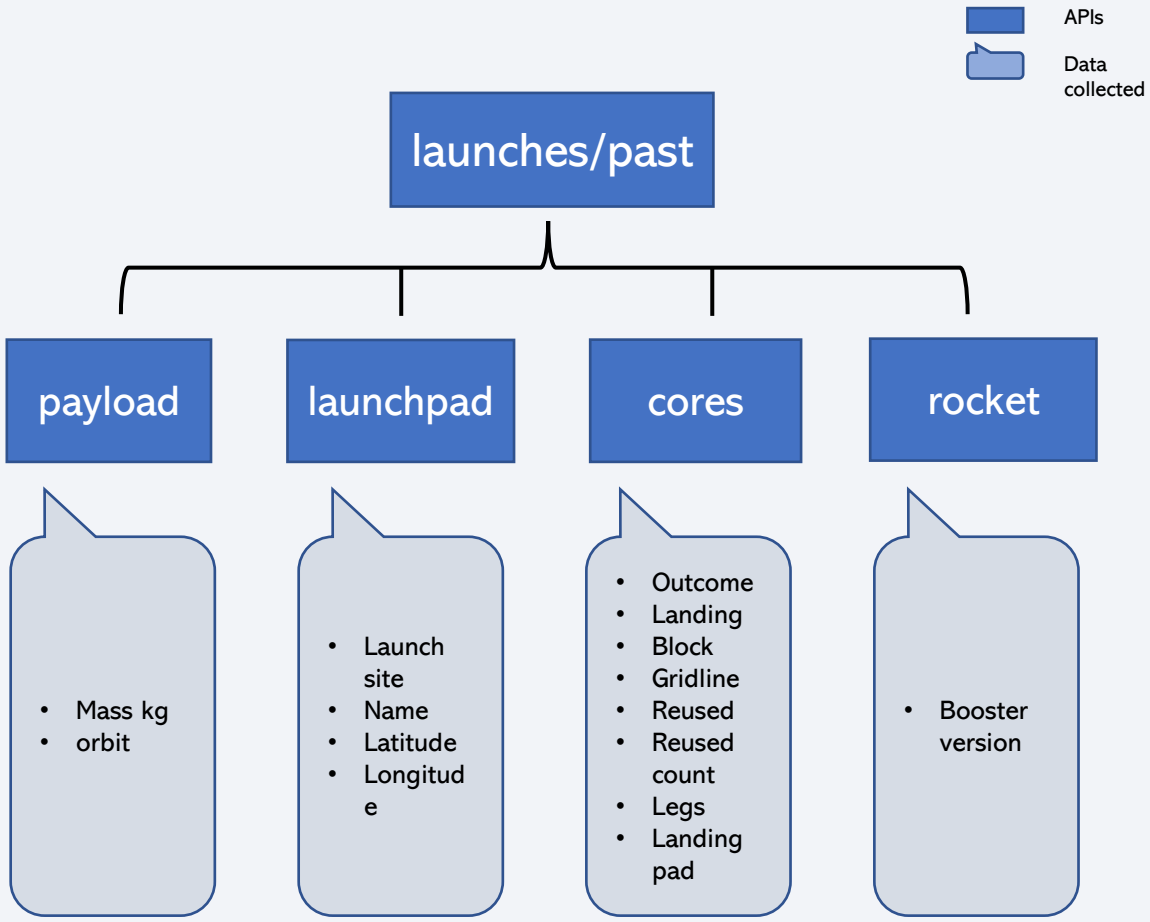
Data is collected from various SpaceX API which is then parse and scrap for more data from more APIs.

Python requests library is used for data collection and web scrapping.

Pandas is used to organized the data into dataframes.

Data Collection – SpaceX API

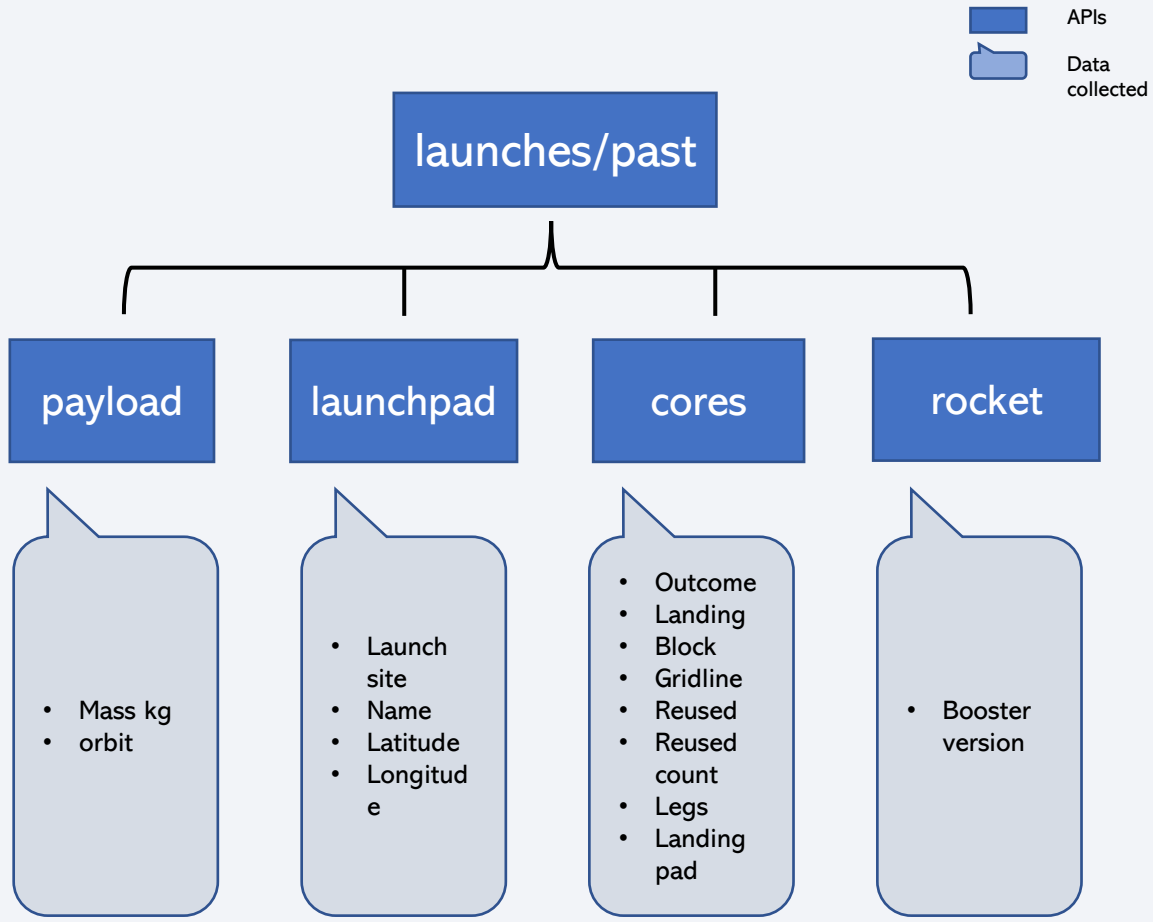
- The APIs used for the projects are shown in the flow charts. (Appx 1)
- Firstly, we used the past launches APIs to gain information about payload, launchpad, cores, and rocket data.
- We then parsed the data collected and reused them for further web-scrapping using respective APIs.



https://github.com/Aunghein09/ibm_data_science/blob/main/lab1-data-collection.ipynb

Data Collection - Scraping

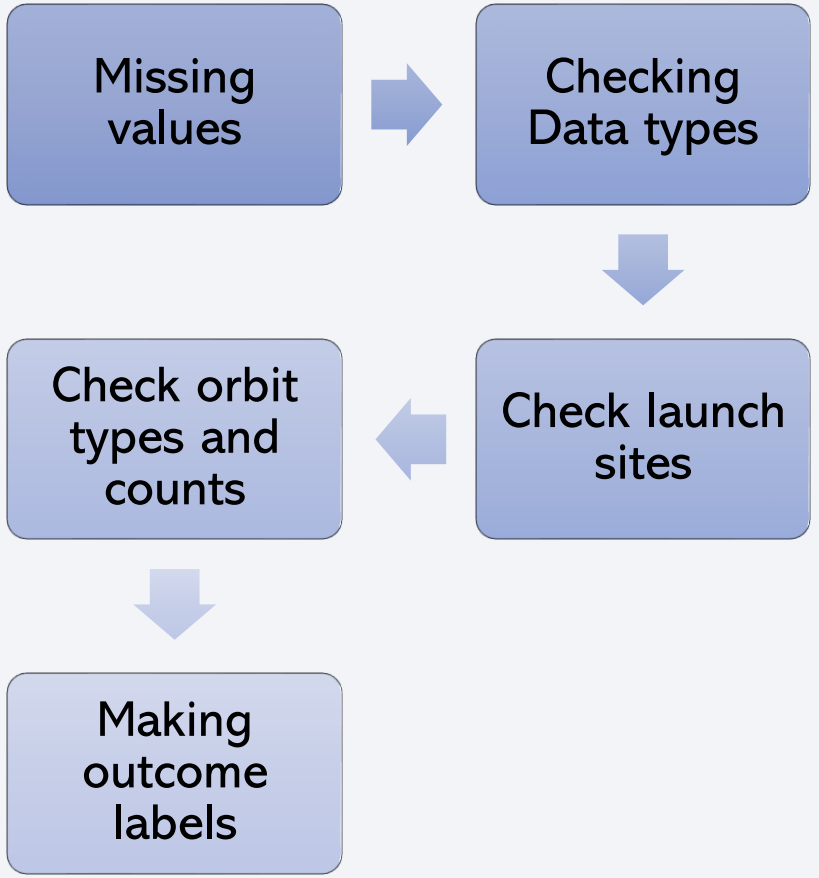
- From information of past launches APIs, we scrap:
  - Payload → mass, orbit
  - Launchpad → launch site, location
  - Cores → outcome, landing type, block, legs, landing pad, etc.
  - Rocket → booster version
- The data is then saved.



https://github.com/Aunghein09/ibm_data_science/blob/main/lab1-data-collection.ipynb

Data Wrangling

- Deal with the missing values from payload mass and launch pad
- Check and correct data types
- Overview insight of orbit types and counts, launch sites and launch counts.
- From outcome from scrapping, differentiate good and bad outcomes and produce a new label as Class with (1,0).



https://github.com/Aunghein09/ibm_data_science/blob/main/lab2-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- In summary, we visualize the relationships of data of interest as follows.
  - Flight numbers vs Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Success Rate vs. Orbit Type
  - Flight Number vs. Orbit Type
  - Payload vs. Orbit Type
  - Launch Success Yearly Trend
- Because payload mass seems to be a big factor for landing success, we take a step further to analyze its relations with other features by visualization

https://github.com/Aunghein09/ibm_data_science/blob/main/lab4-eda-dataviz.ipynb

EDA with SQL

- Summary of SQL queries
  - All Launch Site Names
  - Launch Site Names Begin with 'CCA'
  - Total Payload Mass
  - Average Payload Mass by F9 v1.1
  - First Successful Ground Landing Date
  - Successful Drone Ship Landing with Payload between 4000 and 6000
  - Total Number of Successful and Failure Mission Outcomes
  - Boosters Carried Maximum Payload
  - 2015 Launch Records
  - Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

https://github.com/Aungheinn09/ibm_data_science/blob/main/lab3-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map objects such as markers, circles, lines, etc. are created and added to a folium map

- Markers and circles are mainly used to locate different launche sites and their surrounding areas, different colored icon markers are also used to mark success or failure of each launch

- Lines are added to calculate the distance between launch sites and their surrounding proximity structures such as coastline, railways, highways and cities.

https://github.com/Aunghein09/ibm_data_science/blob/main/lab5-jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart for success rate of different sites is added to view the success rates of different sites

- Scatter plot to show the relationship of payload mass and outcome is added with a slider to view which payload mass range has the highest success rates

Predictive Analysis (Classification)

- The data are cleaned and processed first, then data frame for features and labels are divided.
- After that, the data is preprocessed by standard scaler before splitting into training and test sets
- The training data is trained by 4 models, Logistic Regression, Decision Tree, SVC, and KNN, applying different parameters and kernels, the best of which is then chosen by Grid Search CV.

https://github.com/Aunghein09/ibm_data_science/blob/main/lab6-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Different launch sites have different success rates; CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Flight number and payload mass also have positive relationship with success landing rates
- Flight number also has effect on ISS orbit type.
- Heavy payloads effect on successful landing or positive landing rate are more for PO, LEO and ISS.
- The landing success rate have been increasing since 2015
- According to interactive dashboard, KSC LC-39A has the most success rate

Results

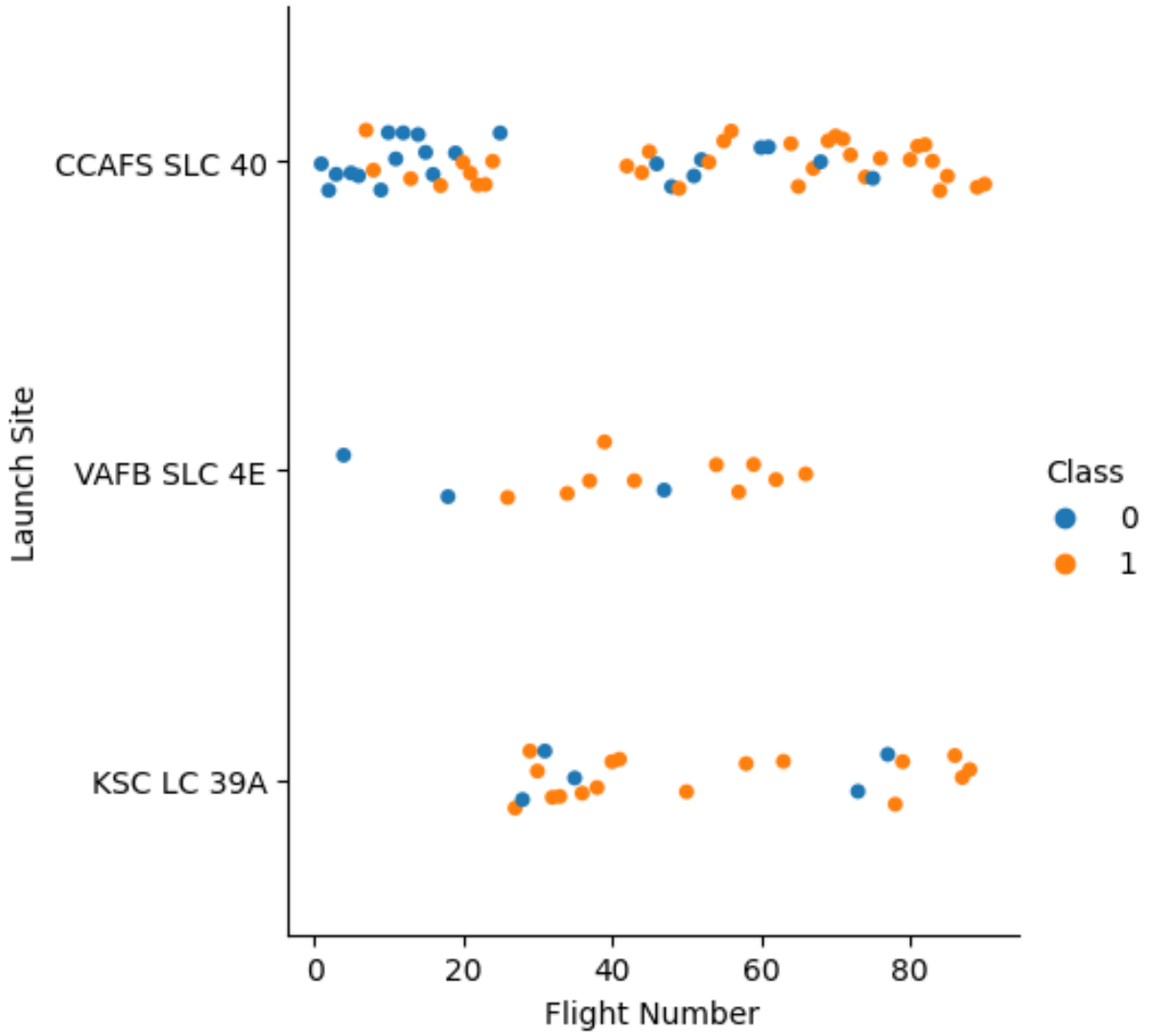
- Heavy payload mass (over 10000kg) has the highest success rate of all ,but for payload under 10000kg, payload mass between range of 3000 -4000 kg has the highest success first stage landing rate.
- Logistic Regression, Decision Tree, SVC, and KNN are built and validated to predict the test set_ they all perform the same accuracy on the test set, with Decision Tree a little bit better on the training set

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

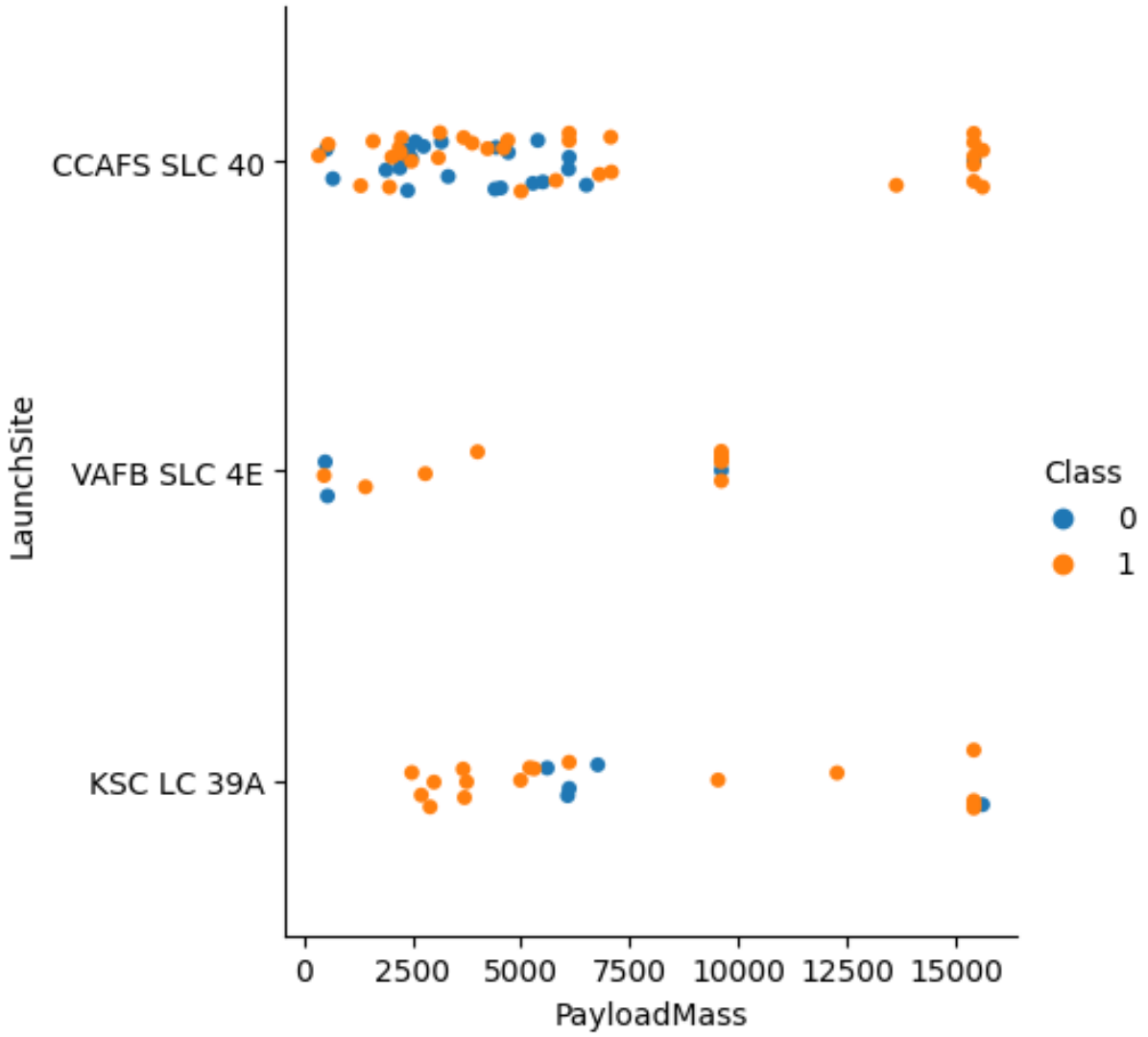
- Success landing rates of all three sites are improved with increasing flight number with VAFB obvious beyond 20, KSC approximately above 35 and CCAFS above 75



Payload vs. Launch Site

- The first stages of launches with pay load mass over 7500 kg are more likely land successfully.

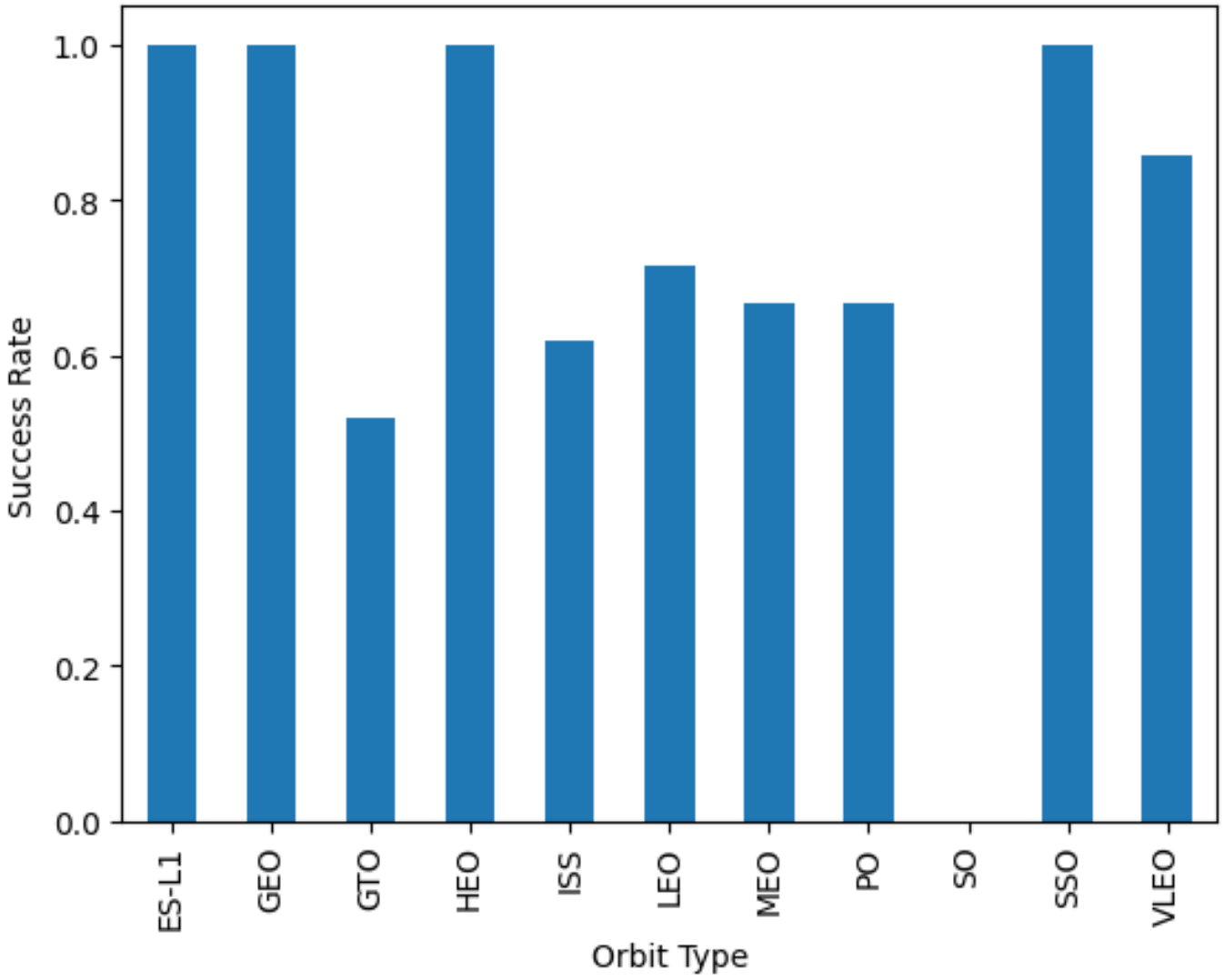
- VAFB site did not have rockets launched for heavy payload mass(greater than 10000).



Success Rate vs. Orbit Type

- SSO and VLEO orbits have a very high success rate especially SSO with 100%

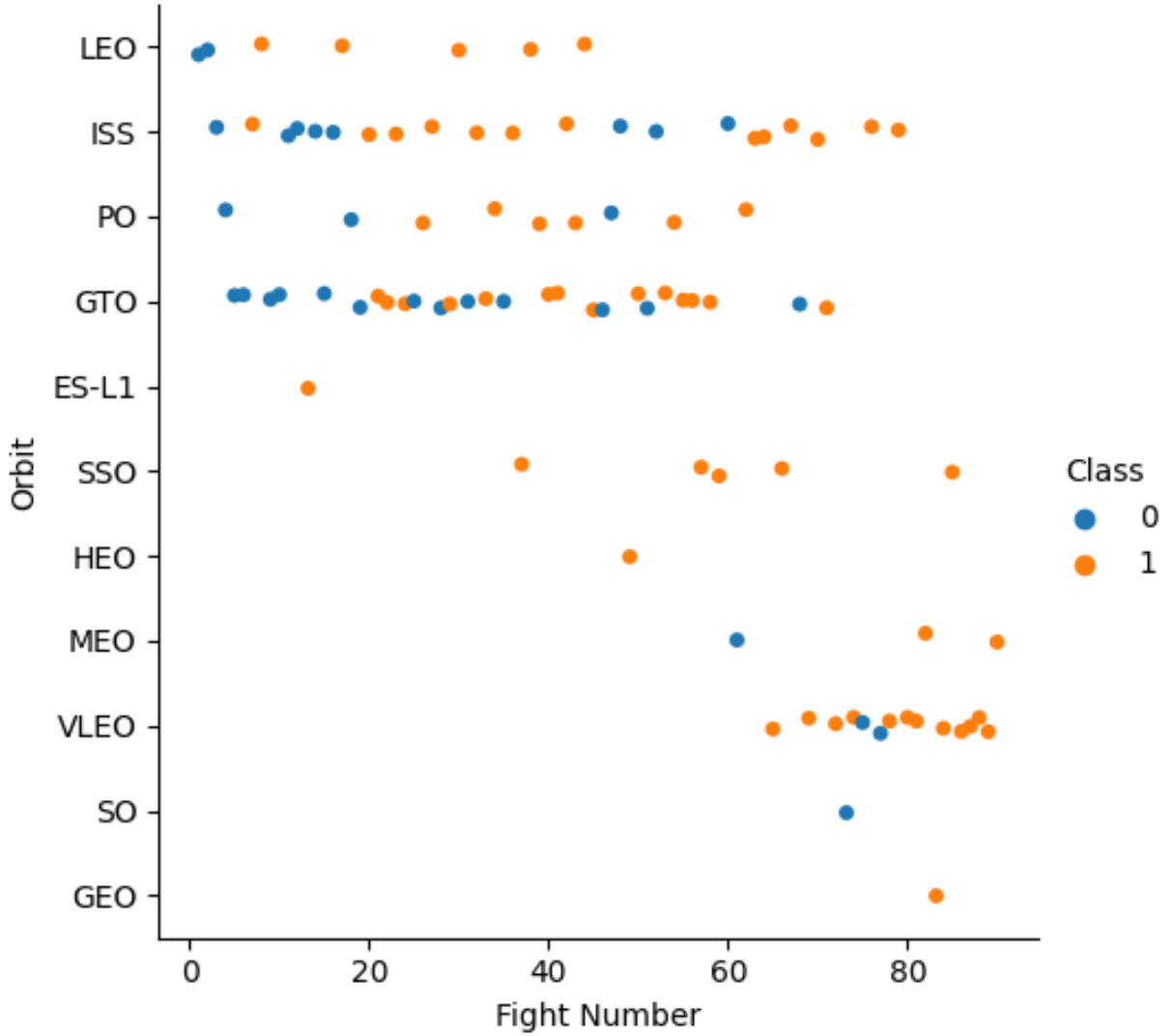
- There are also a few other with very high success rate and very low success rate (due to very few launches which we can see in next slide)



Flight Number vs. Orbit Type

- Orbits LEO, ISS and VLEO appear related to number of flights.

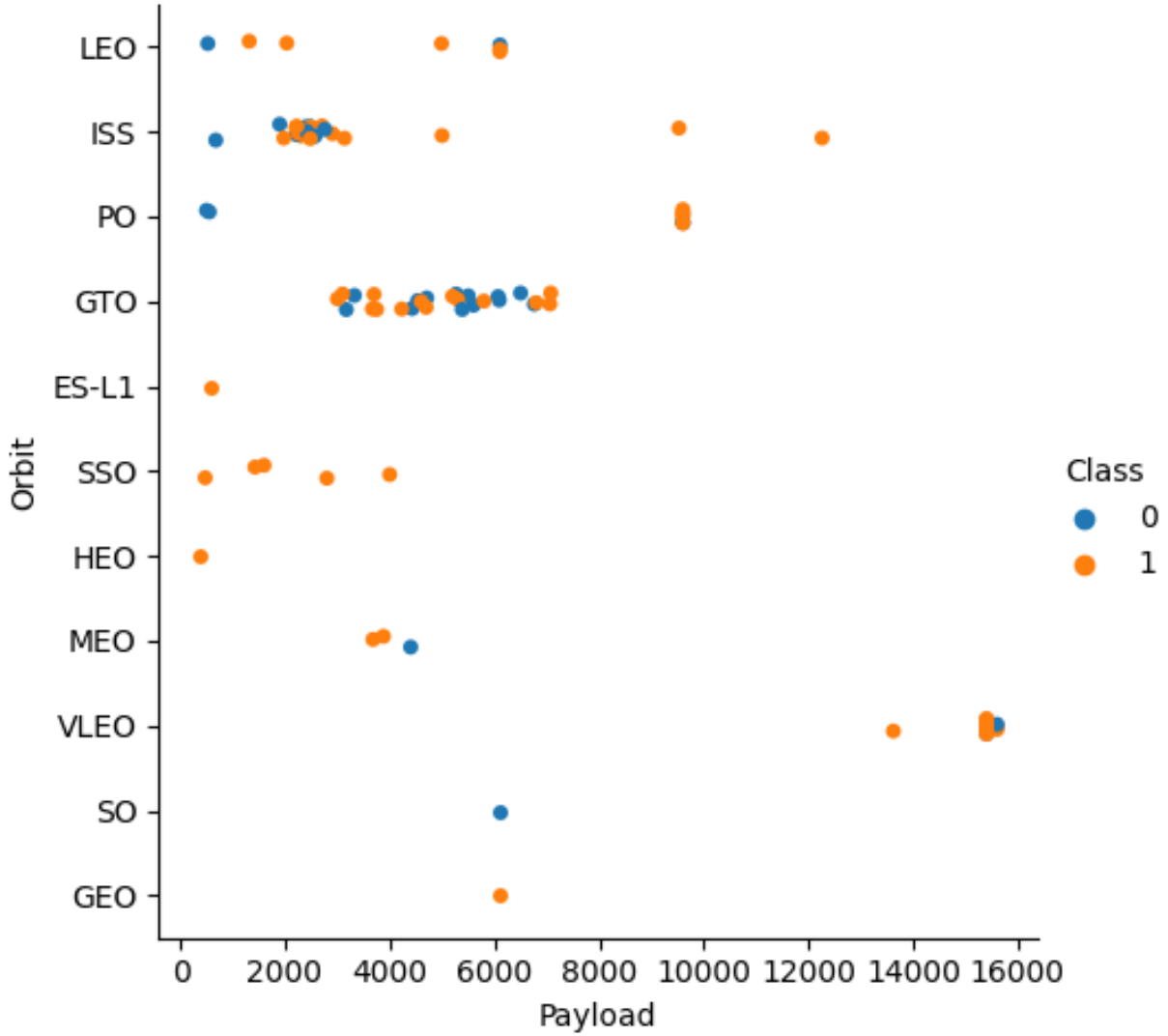
- There seems to be no relationship between flight number in other orbits.



Payload vs. Orbit Type

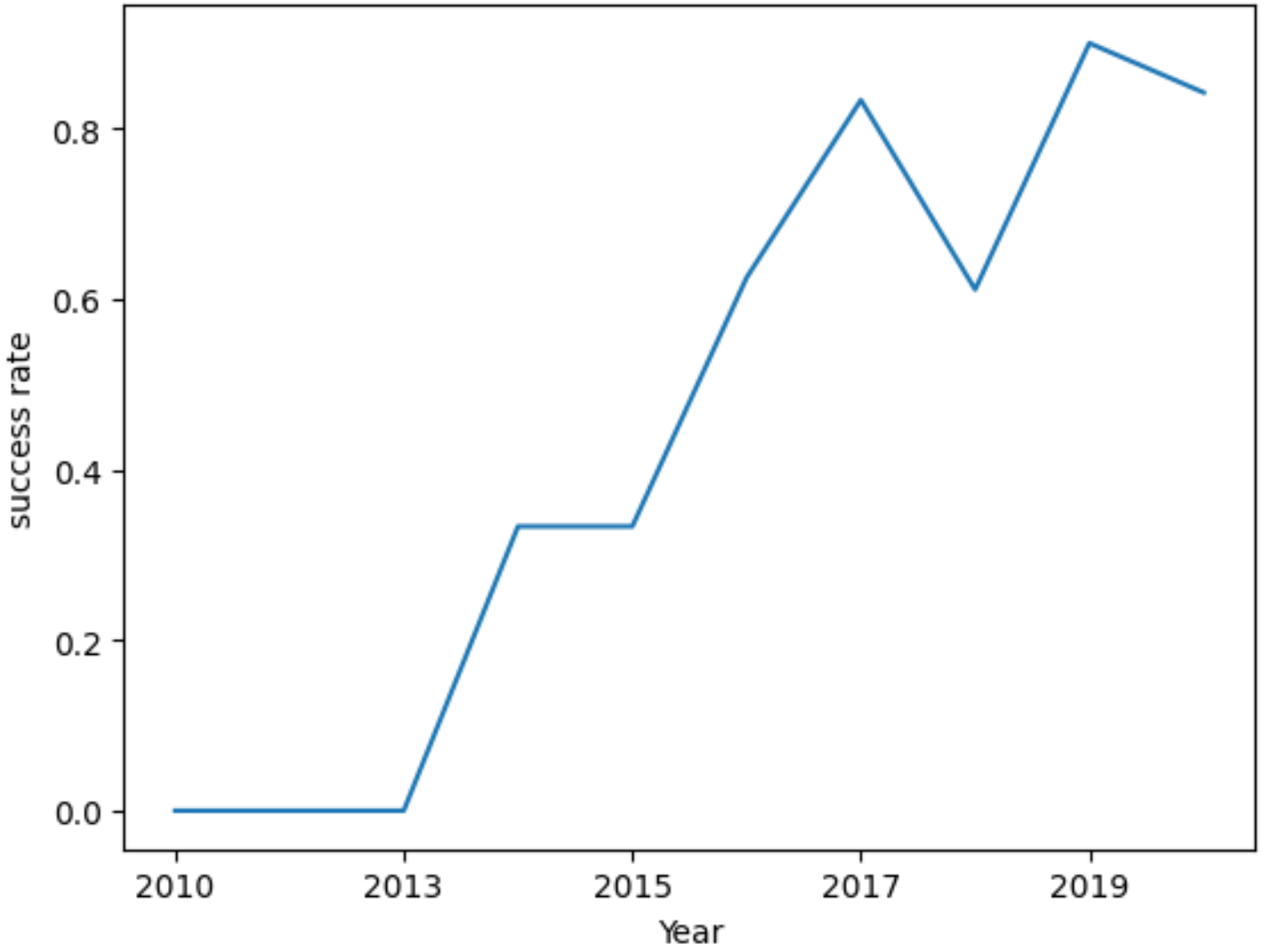
- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS

- For GTO, the boundary is not that obvious to relate to payload.



Launch Success Yearly Trend

- The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- SpaceX rockets are launched in 4 different launch sites.

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

Launch Site Names Begin with 'CCA'

- Launches are done at CCA launch sites since 2010, the earliest year in our data.
- CCA launch sites are the major launch sites in the world since 1950s(Howell, 2016)

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

Total Payload Mass

- According to data, the total payload that NASA(CRS) have using SpaceX rockets is 45596 kg.

Customer	total payload mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The average payload mass by Falcon9 v1.1 is 2534

- This might suggest that NASA(CRS) have used SpaceX service for 10-20 times.

average payload mass
2534.6666666666665

First Successful Ground Landing Date

- Successful Ground Landing of SpaceX first stage was achieved 5 years ago where overall landing success rate increase from 30% in 2015 to over 80% in 2017

First Successful Ground Landing Date
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- The booster version that had successful drone ship landing within the observed payload range are all F9 FT versions.

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

Total Number of Successful and Failure Mission Outcomes

- Although first stage landing outcomes can be not desiring, mission outcomes from SpaceX launches have over 99% success rate

Mission_Outcome	count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The booster version that carried maximum payload mass are all B5 version

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

2015 Launch Records

- In 2015, 2 of the drone ship landing attempts failed
- Both were launched in the same site and had the same booster version of 1.1

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

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

- Overall success rate of first stage landing between the observed date is 65%
- Both attempts with parachute also failed.

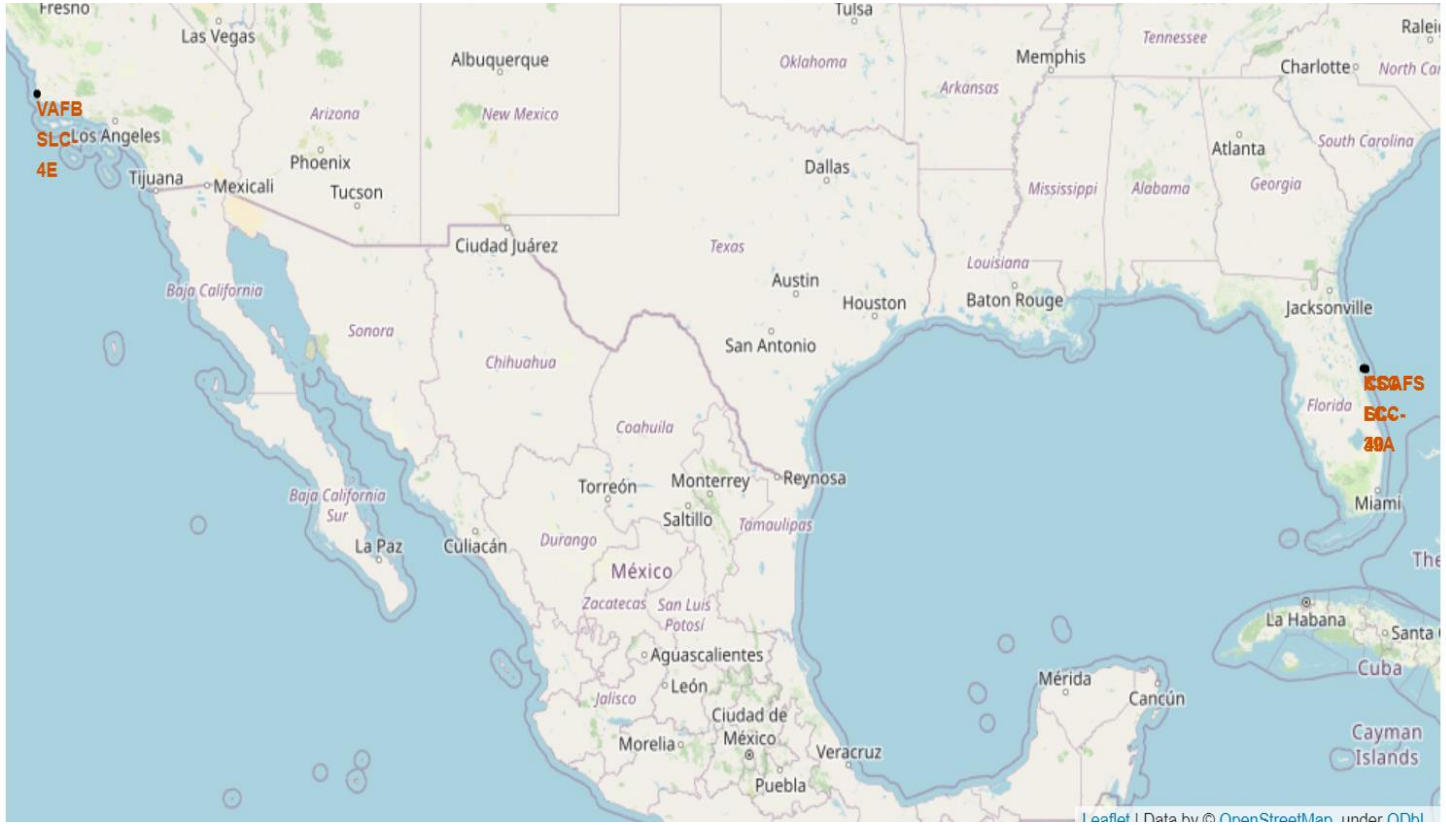
Landing_Outcome	count
Success	20
No attempt	11
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2

Section 3

Launch Sites Proximities Analysis

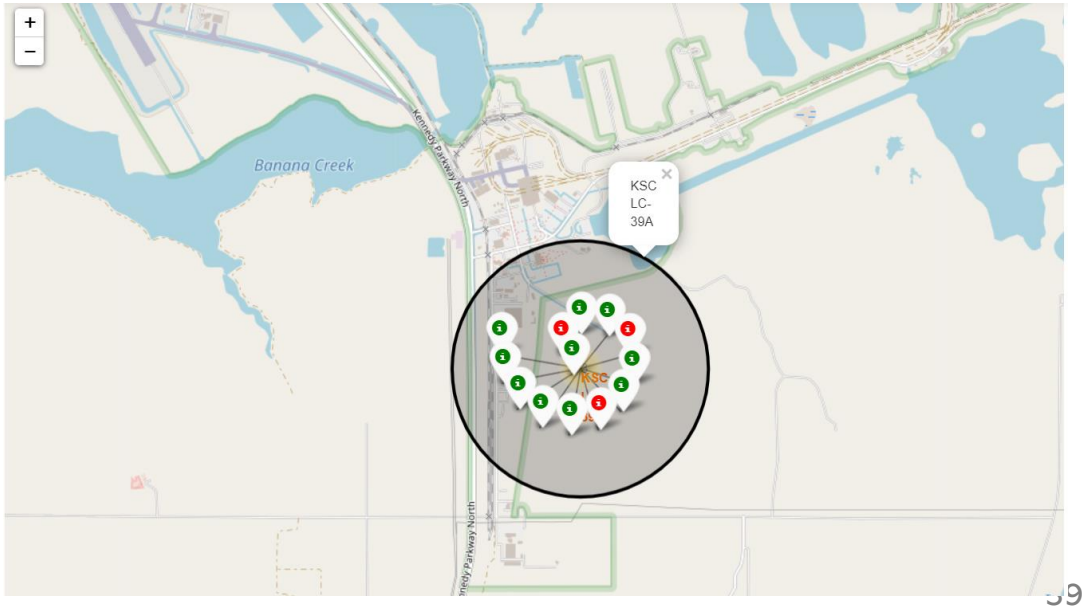
Launch Sites of SpaceX Rockets

- All the launch sites in close proximity to the coastline
- The launch sites in proximity to equator and are in Florida except VAFB SLC-4E



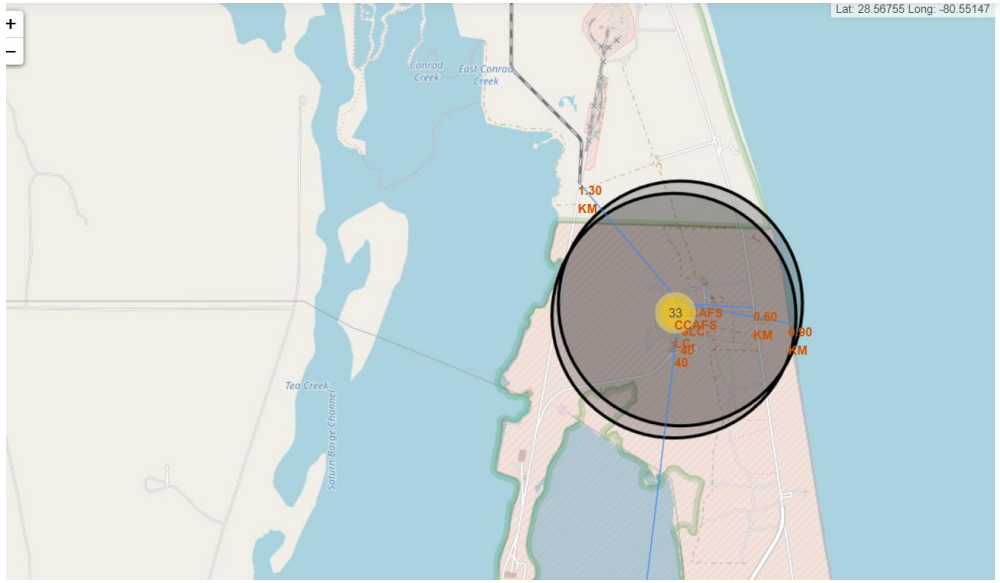
Success/Failed launches for each site

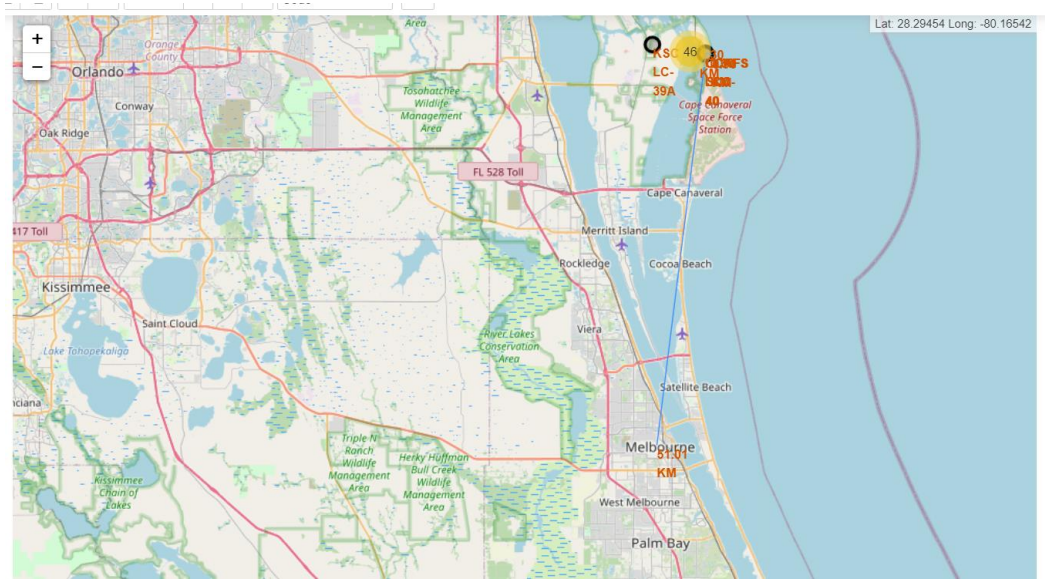
- 80% of launches were in Florida
- KSC LC 39A has the most success rate, with 10 success out of 13 launches
- Most of the SpaceX rockets were launched in CCAFSs accounting for 33 launches

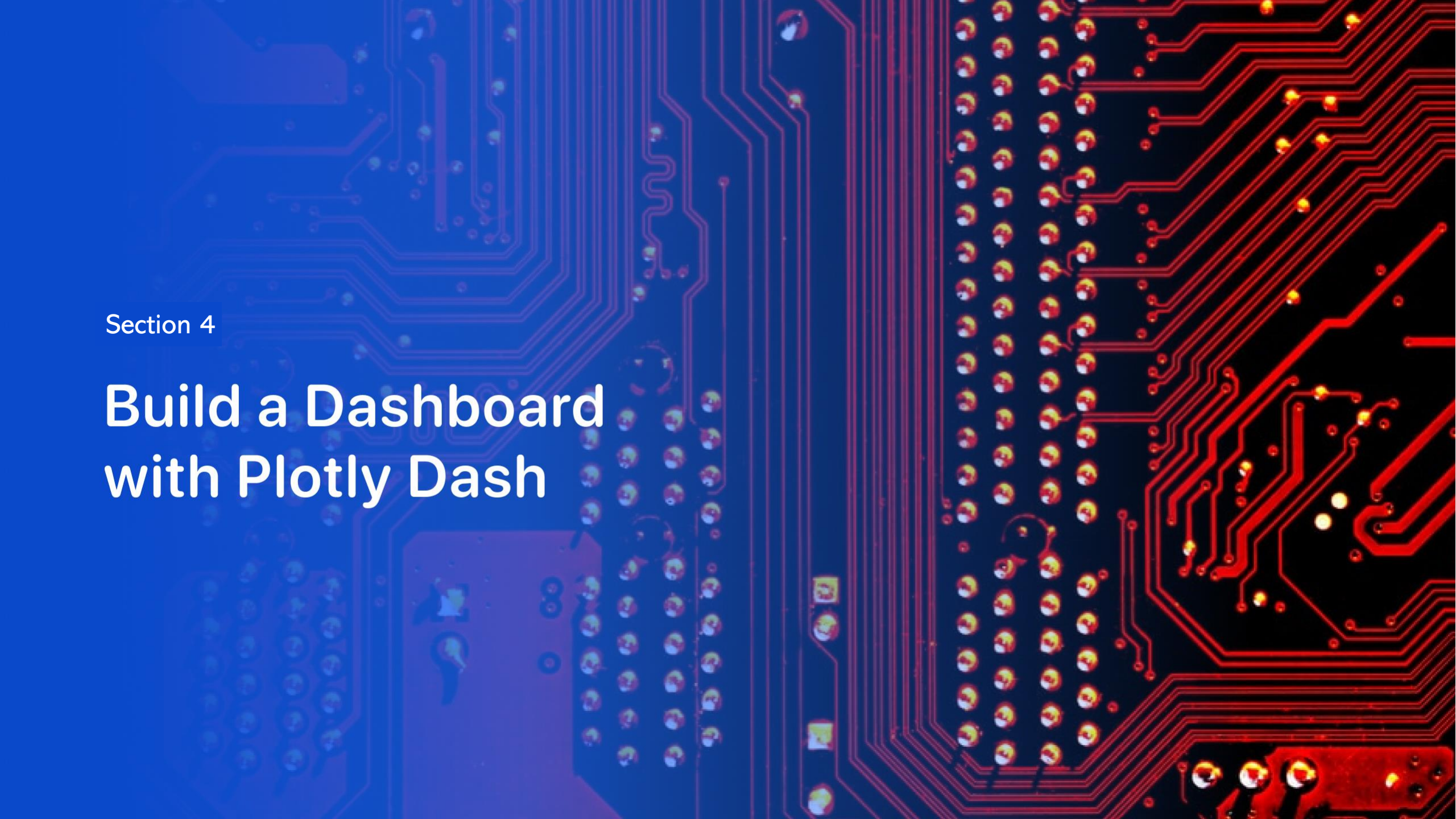


Launch sites and their proximities

- All launch sites are in very close proximity with the coastline with the shortest being less than 1 km
- They are also in proximity with railways and highways
- However, they have a moderate distance from big city with the closet over some 30 km





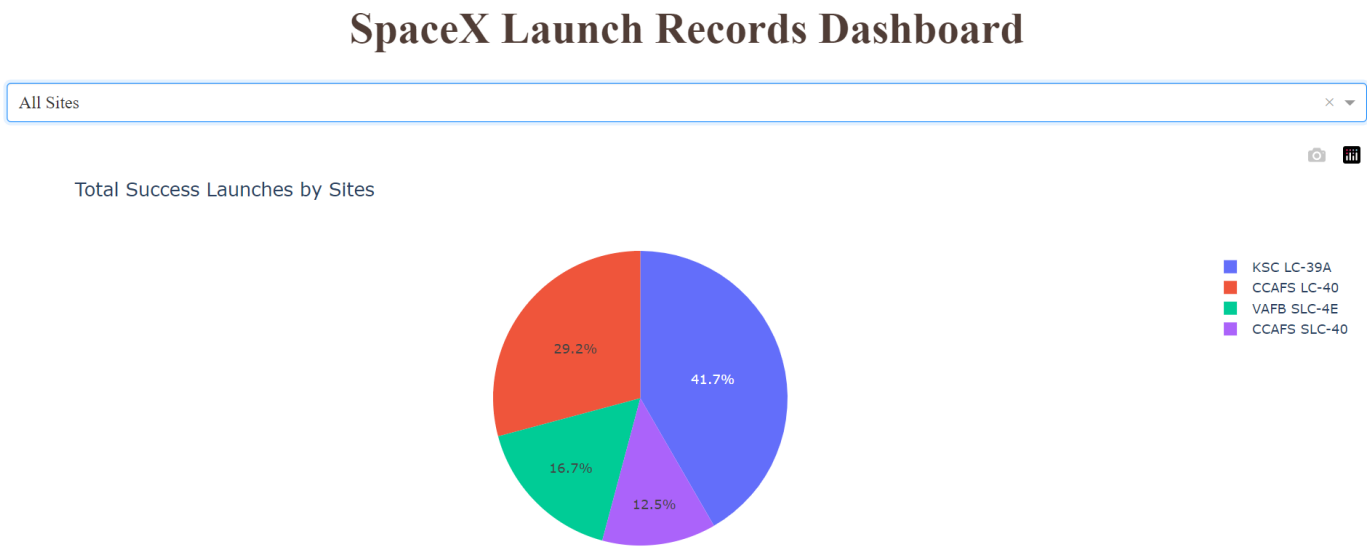


Section 4

Build a Dashboard with Plotly Dash

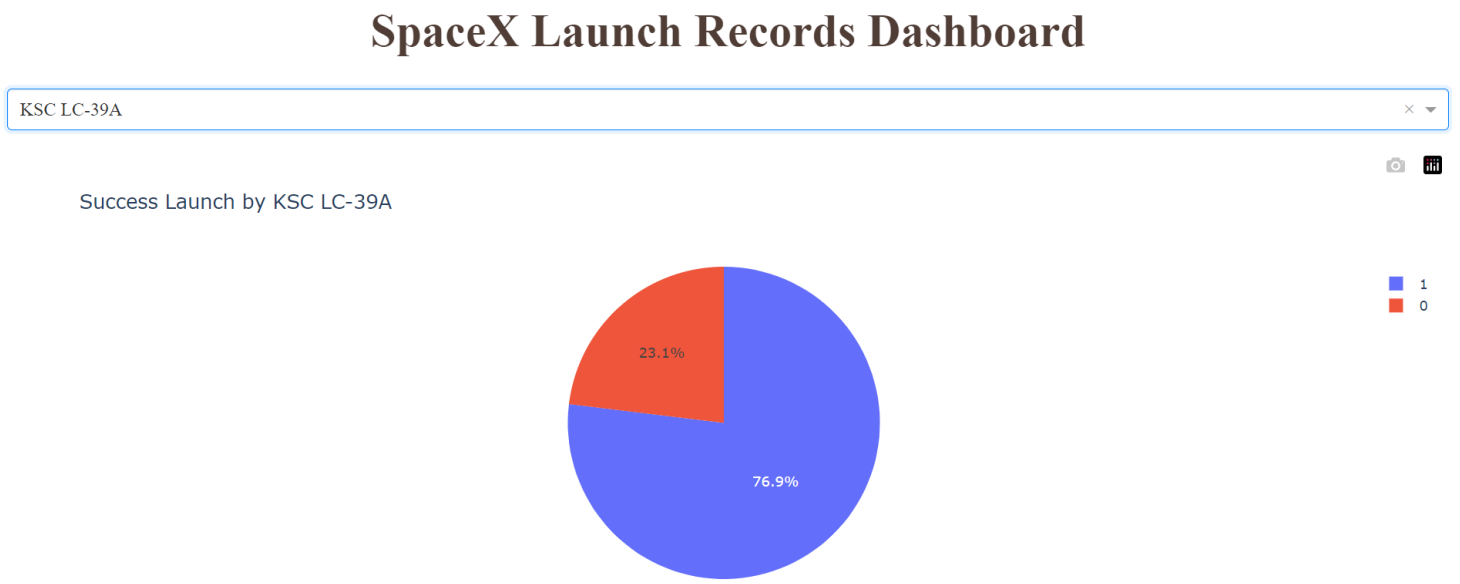
Total Success Launches by Sites

- The sites with the most success rates is KSC LC-39A
- This can also be related to previous data from folium map



Success Launch by KSC LC-39A

- The site has a success rate of over 75% with a moderate portion of total launches



Payload Mass VS Launch Success

- Payload mass of rockets high varies the success rates of landing first stage

- From the dashboard, payload between 3000 and 4000 kg have the greatest success rate

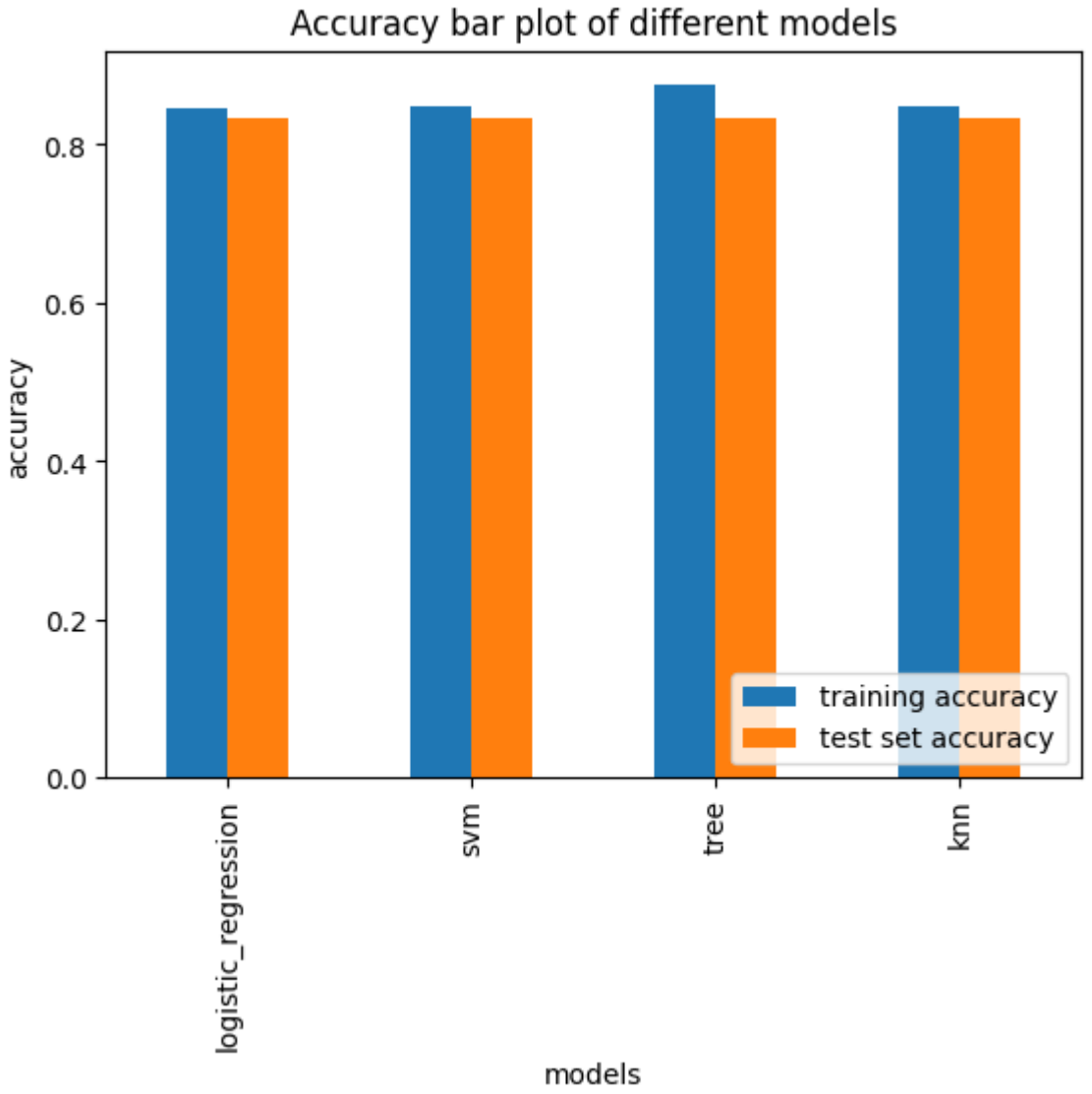
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The accuracy on the test set by all 4 models seems to be similar

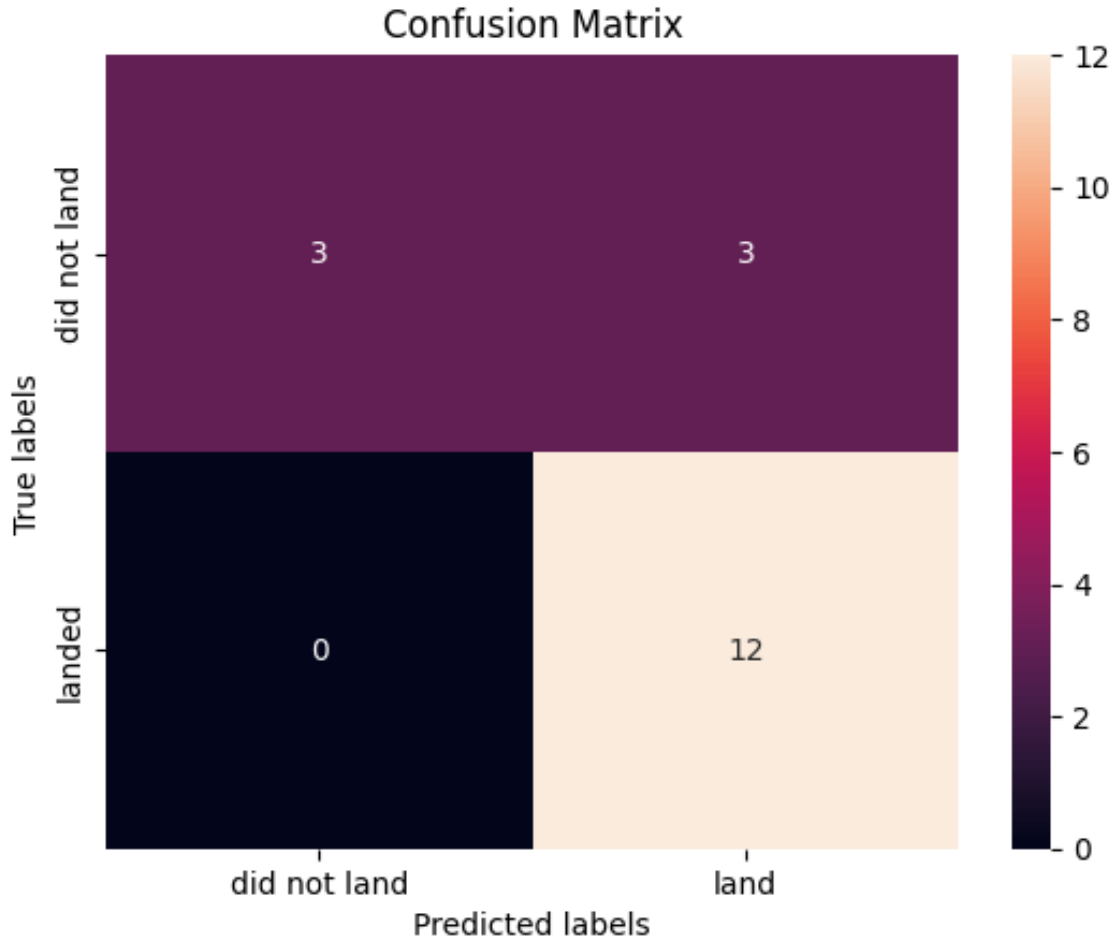
- According to the best score from the test set, decision tree model is the best



Confusion Matrix

- Decision tree model has a 100% accuracy when the first stage was successfully landed

- However, it give 50% false positive in cases where first stage was not successfully landed



Conclusions

- Different launch sites have different success rates; CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- Payload mass seems to be a major factor in achieving a successful landing of the first stage
- Flight number and certain orbit types also have positive relationship with success landing rates
- The landing success rate have been increasing since 2015
- All the launch sites in close proximity to the coastline
- All 4 models tested have the same accuracy on test set, while decision tree has a little bit better accuracy on training set.

Appendix

1. APIs

- https://api.spacexdata.com/v4/launches/past
- https://api.spacexdata.com/v4/rockets/
- https://api.spacexdata.com/v4/launchpads/
- https://api.spacexdata.com/v4/payloads/
- https://api.spacexdata.com/v4/cores/

2. Data set

- https://github.com/Aunghein09/ibm_data_science/tree/main/datasets

3. References

- E. Howell.(2016). Cape Canaveral: Launch Pad for U.S. Space Program. https://www.space.com/33926-cape-canaveral.html#:~:text=The%20area%20has%20two%20active,assassinated%20U.S.%20president%2C%20John%20F.

Thank you!