A Detailed Report on **SpaceX**'s Launches and Success

Author: Pulak Pallab Kashyap

Date: 10th October, 2023

Executive Summary

- SpaceX, a private company founded by Elon Musk, is one of the prominent commercially viable space agencies in the world which has made space travel affordable. They help send manned spacecrafts as well as satellites for internet services link Starlink.

- SpaceX are the most economically efficient of the handful of private space agencies in the world, it 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 which plays a telling role in decreasing the cost of the entire programme.

Table of Contents:

1. Introduction
2. Methodology
3. Results
4. Discussion
5. Conclusion
6. Acknowledgement
7. Appendices

Introduction

- As a starting point of almost all data science projects, data will be collected, as much and relevant as possible.
- After the raw data has been collected, data wrangling will be performed to improve the quality of the data.
- Then the processed data will be explored using SQL as the data is queried and insights gathered from it.
- Further insights into the data will be gathered by applying some basic statistical analysis and data visualization to see how variables are affected by one another.
- Predictive models, then, will be built, evaluated, and refined for discovering more exciting insights.
- The final task of this capstone project is to create a presentation that will be developed into stories of all your analysis.

Methodology (Data Collection)

- We start by using an API to retrieve the relevant data, we request for the response from the API and then get the JSON content from the same.
- We then normalize the JSON file to collect the necessary information from it.
- We drop the unimportant columns from the table and then organize and clean the data in to table for future use.
- We also use the Wikipedia page "List of Falcon 9 and Falcon Heavy Launches" to collect data important data that could be useful later. We use Webscrapping with the help of BeautifulSoup for the same.
- We request the necessary content from the page and then pass it to BeautifulSoup element for parsing through the info.
- After which we use the functions of Beautiful Soup to retrieve the necessary Tables, Columns and other info that we might require in the future
- Finally we organize and clean the data to make sure they can be stored in the desired format in Tables which could later be used for predicting and analyzing.

Methodology (Data Wrangling)

- We use Python for Data Wrangling our dataset once the relevant data is collected for the Falcon 9 Booster Version.
- We start the Data Wrangling procedure by first checking the "Null Values" in each of the Columns and data types of each field.
- Then, we check the value counts (number of times each value comes up) for "Launch Site", "Orbit" and "Outcome".
- After which we group the categories together based on whether the first stage landed back successfully or failed to (including the times the First Stage was not attempted to be landed).
- We make a new column for the Successful landing of the First Stage and the failed attempts to land the first stage.
- Finally we find the average times the First Stage landed back successfully as well as the average Payload Mass (Kg) of the First Stages.

Methodology (EDA Visual Analytics)

- First we import the necessary libraries to use SQL in Python like sqlite3 and connect it to the database, with a cursor object and load the SQL extensions
- Then using SQL commands we perform the necessary Exploratory Data Analysis on it like finding out the first date on which the first successful landing of the first stage took place, different spots where landing are attempted, extract the date, month, year from the Date Field, etc.
- After that we use Pandas and Matplotlib libraries to check how different Categories are correlated with one another and the Class field.
- We check the relations between Flight Number, Payload Mass Launch Site, Orbit and Date with the Class Attribute for the data.
- Finally we perform Feature Engineering on the dataset; drop out the irrelevant fields and apply One Hot Encoding on the necessary Categorical columns to make their data discrete as well as change all their data types to float.

Methodology (Interactive Visual Analytics)

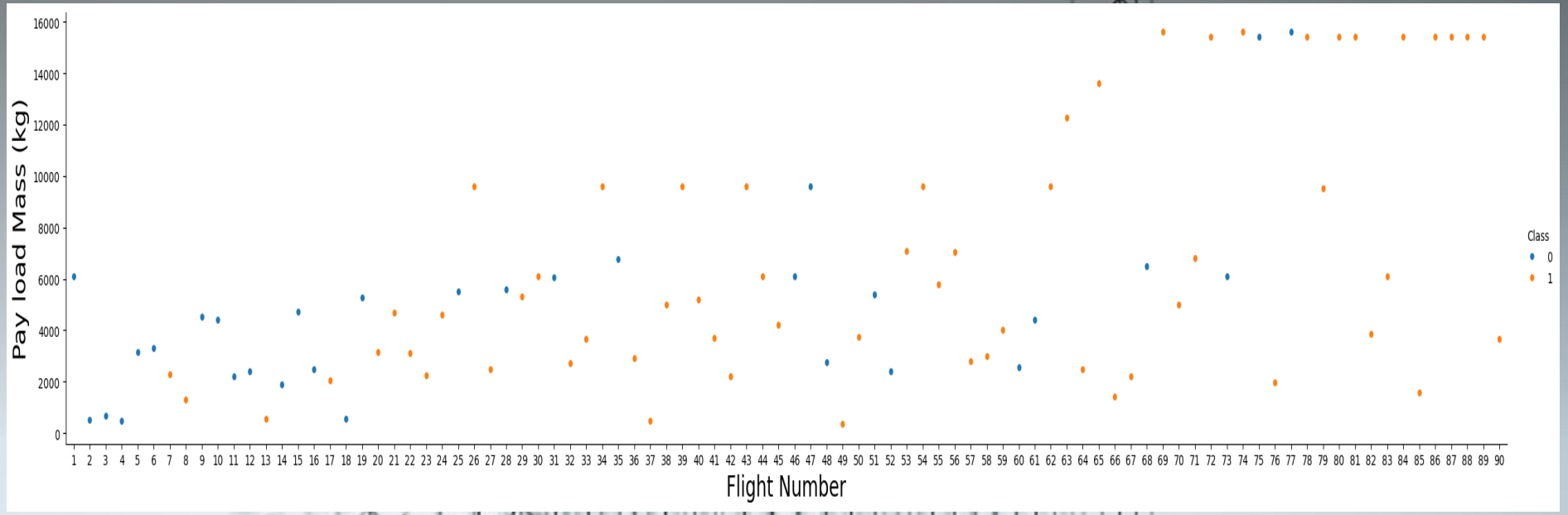
- First we import the relevant libraries and Plug-ins for the maps like Folium and zoom the same to so the map centers around our key locations.
- After that we use Marker and Circle to mark the Launch Sites and MarkerCluster to visualize the number of successful/failed launches for each site with specific colors for each category of the Class Attribute.
- We then Mark the nearest City, Highway, Coastline, Railway from a Site and draw a straight line to each from the site with their calculated distances from the site.
- After that we made a Interactive Dashboard using Plotly with Callbacks for different sites which displays the Success rate of each Site (including for All Sites combined).
- The Dashboard also includes the correlation between Payload Mass and Success Rate for each Site (including for All Sites combined) where the Payload Mass could be changed using a RangeSlider.

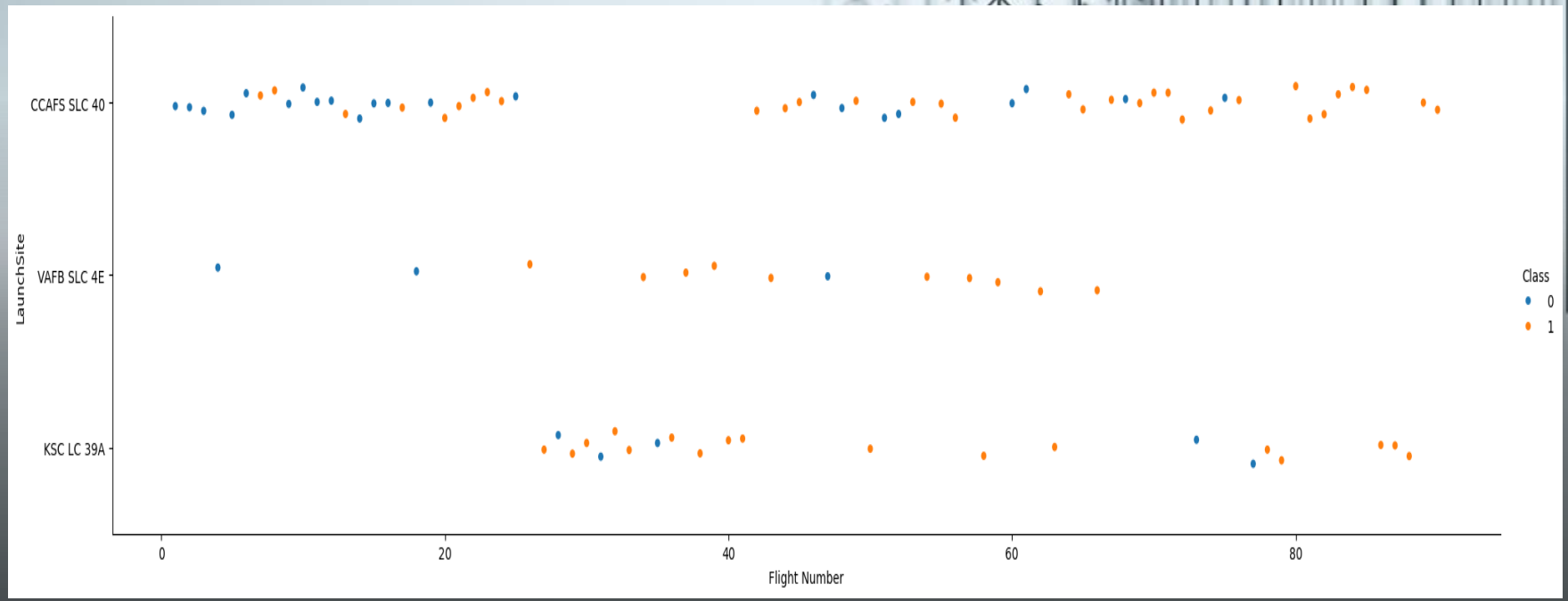
Methodology (Predictive Analysis)

- For the Predictive Analysis part we first import the necessary libraries for the Classification models along with their evaluation matrix/parameters.
- The Classification Models we are to perform for this instance to predict the class are Logistic Regression, Support Vector Machines, Decision Tree Classifier and K-Nearest Neighbour with GridSearch CV to find the best Hyperparameters for the prediction.
- We standardize the independent variables, after which we split the data to Train (80%) and Test sets (20%).
- Then, we apply the above mentioned algorithms on the Train data and test their accuracy on the Test data after the Train data is fitted on the algorithm.
- Finally we retrieve the Confusion Matrix, Evaluation Parameters, Classification Report for each Model to determine its accuracy with the Untrained (Test) data.

Results (Visualization)

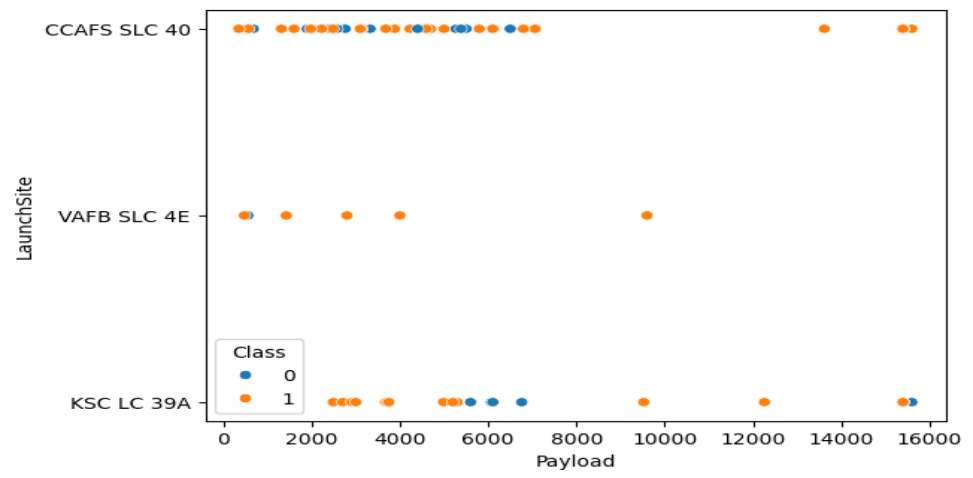
It appears that the payload has increased with higher Flight Numbers with also the Landing Outcomes being more Successful (represented as Orange dots) with respect to rising Flight Numbers.





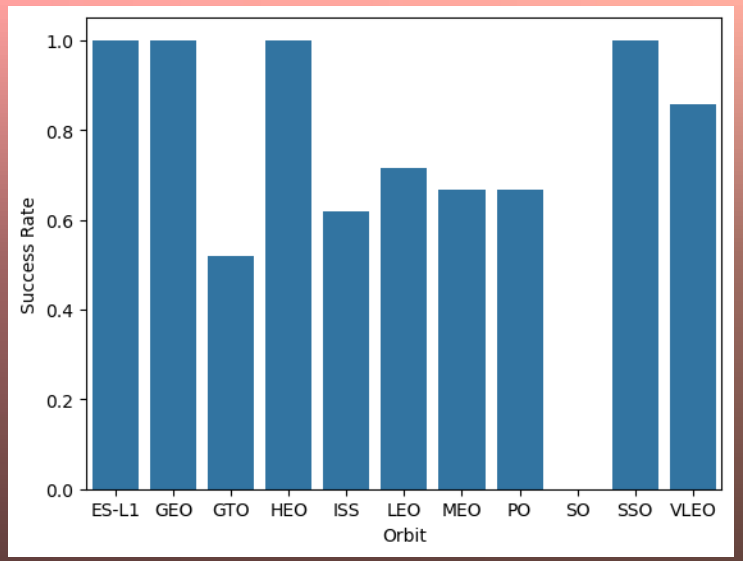
It appears that most of the Launches were from the CCAFS SLC-40 Site, though during the mid-range of the Flight Numbers KSC LC-39A and VAFB SLC-4E also had a decent amount of Launches. The Launches from KSC 39A have continued till the end but VAFB SLC-4E have piped down. Success is seen more often towards the end.

Results (Visualization II)

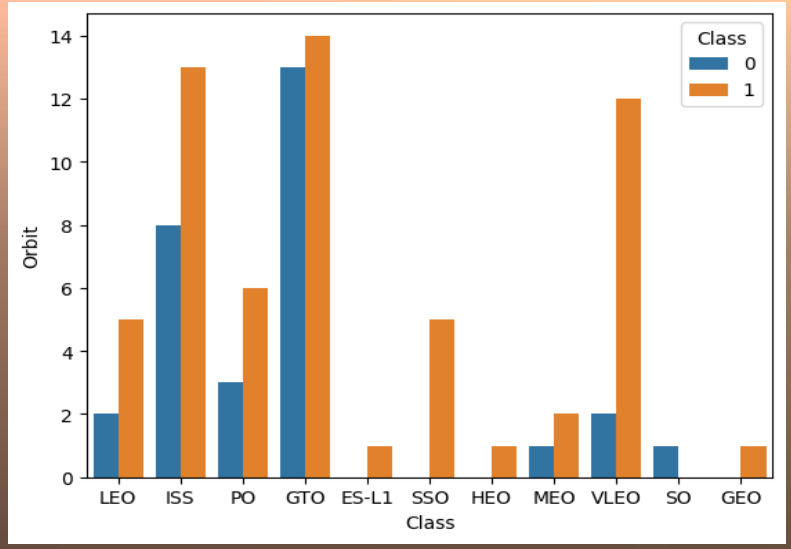


The Scatter Plot shows that most of the Launches are with Payload Mass less than 6000, and in that range the Landing Outcomes are usually Successful with most of the Launches being from CCAFS SLC-40 followed by KSC LC39A and VAFB SLC-4E. The number of Launches from the latter dies down for over 6000 Payloads with KSC LC 39A having the highest Launches for those above 6000 Payload followed by CCAFS SLC-40. The success rate of the latter is the highest for the Launches with Payload higher than 6000 Kg.

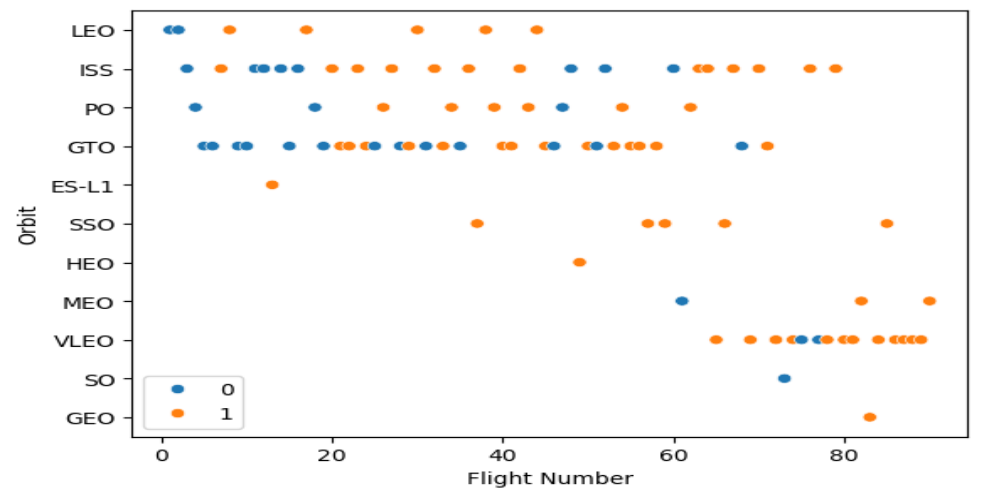
From the below two graphs it appears that VLEO, ISS and GTO Orbit has the highest number of Success followed by LEO, PO and SSO but the Success Rate of some of those like GTO, ISS, LEO, PO are not very good compared to the likes of SSO and VLEO.



Though GEO, ES-L1, HEO have fewer number of Launches along with SSO they have all Successful Landing Outcomes and No Failures which makes them highly efficient. SO is the only Orbit which has ano Success with any Landing Outcomes but it has only had 1 Launch. The highest Launches are to Orbits GTO and ISS with 27 and 21 Launches resp.

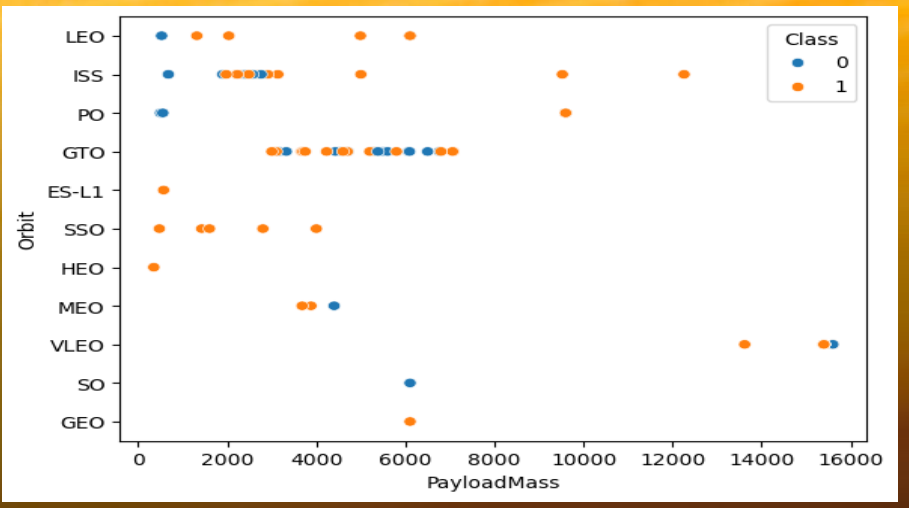


Results (Visualization III)

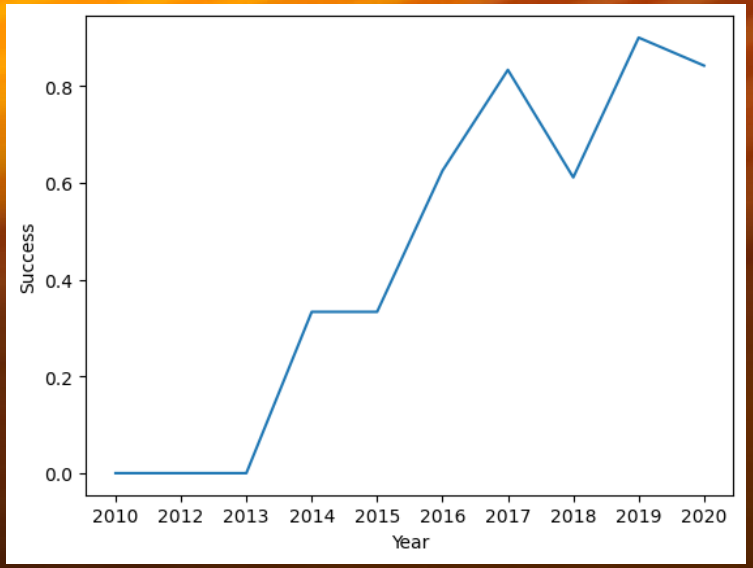


The first Scatter Plot shows that most of the early Launches were for the Orbits LEO, ISS, PO and GTO with many having Failed in their Landing Outcomes and beyond the 60th Flight, The number of Launches for the ISS and VLEO are the highest with a stark increase in Success Rate for Landing Outcomes. Also the Orbits have increased after the median value of the Flight Number and the Launches to different Orbits are more distinguished.

From the second Scatter Plot it appears that most of the launches are with Payload of less than 8000 and those above 8000 are usually launched to ISS, PO or VLEO Orbits. Those launches below 8000 Kg Payload are to more diverse Orbits with GTO and ISS being the Orbits that have been Launched to the highest times.



In the last Graph which is a Line Plot we observe that the number of Successful Landing Outcomes for the First Stage of a mission have increased with timewith the time until 2013 recording no success and those around 2019-20 recording a very high success rate.



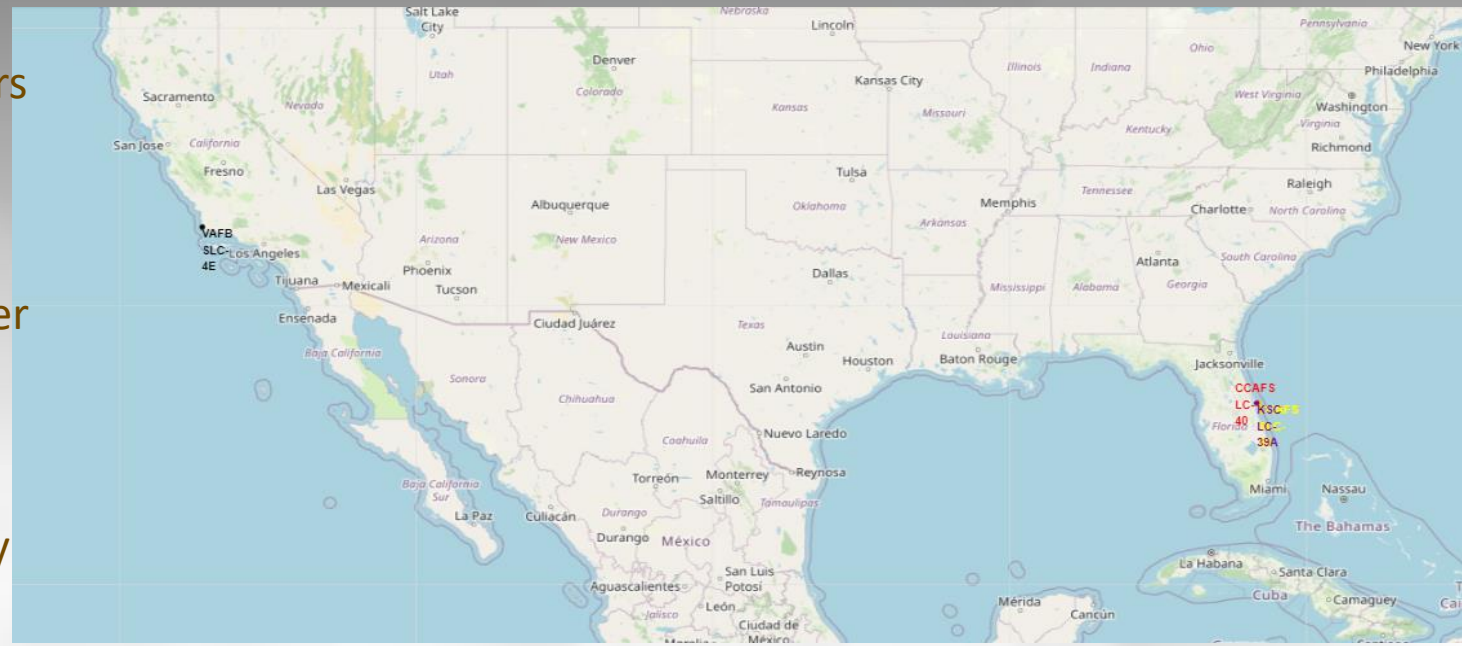
Results (EDA using SQL)

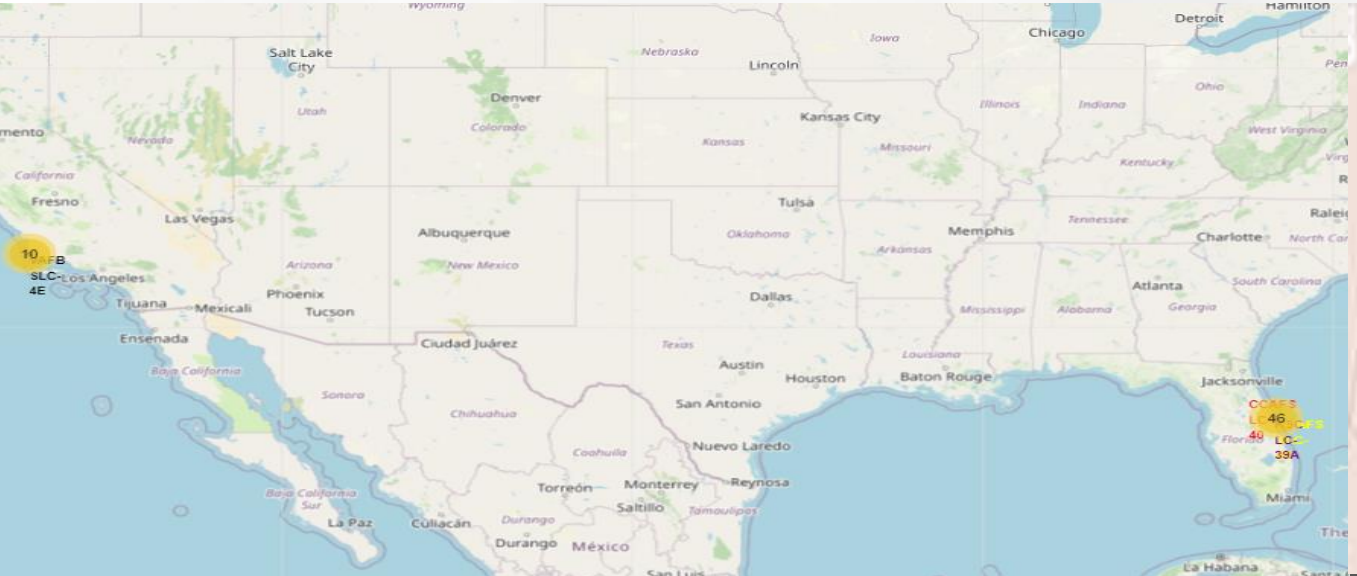
- First we select all the rows that have no Null Values in them and drop the rest.
- We check the unique Launch Sites then, which in this case are 4 and from those Launch Sites we determine the ones that start with 'CCA' which comes out to be 2 and siplay their first 5 occurances.
- Then we check the Sum of the Payload Mass of every Launch by NASA which equals 99980 and the Average Payload Mass for Falcon 9 v1.1% which gives an output of 2534.667 Kg.
- Then we check for the first (earliest) date when there was a successful landing (for ground pad) of the first stage: The output is dated 2015-12-22.
- We check all the different Landing Outcomes after that and retrieve the Booster Version, Payload Mass and Landing Outcome for the successful landings of the first stage for Payloads with Mass between 4000-6000 Kg.
- After that we count all the Mission Outcomes based on their Failure or Success Rate. The result is 1, 98 resp.
- Then we count the number of times the following outcomes have occurred with respect to the landing of the First Stage: Failure, Failure (drone ship), Failure (parachute), Success, Success (drone ship) and Success (ground pad).
- Then we determine the five highest Payload Mass and the Booster Versions that had carried them using a sub-query. All of the 5 Booster Versions turn out to be Falcon 9 B5
- Then we retrieve the Months, Booster Version, Launch Site for Failure(drone ship) as Outcome for the landing of the first stage for the year 2015. The Output is 2, with both Launch Sites being CCAFS LC-40 and Booster Versions Falcon 9 v1.1.
- Finally we count all the different Landing Outcomes between2010-06-04 and 2017-03-20. We get the output of 10 No Attempts, 10 Success, 6 Failures, 3 Controlled and 2 Uncontrolled with 1 Precluded.

Results (Interactive Map with Folium)

In the first map we can see the Markers of the 4 different Launch Sites being Marked with different colors for each. The Launch Sites are also Circled using the Circle function of Folium Library with a certain radius around the Launch Sites.

In the second and third of the two maps we used Marker Clusters to mark the number of times each site had a Successful/Failed Landing Outcome for the First Stage. The Marker Clusters are seen on the unzoomed map as Circles and can be more prominently seen when they are zoomed.



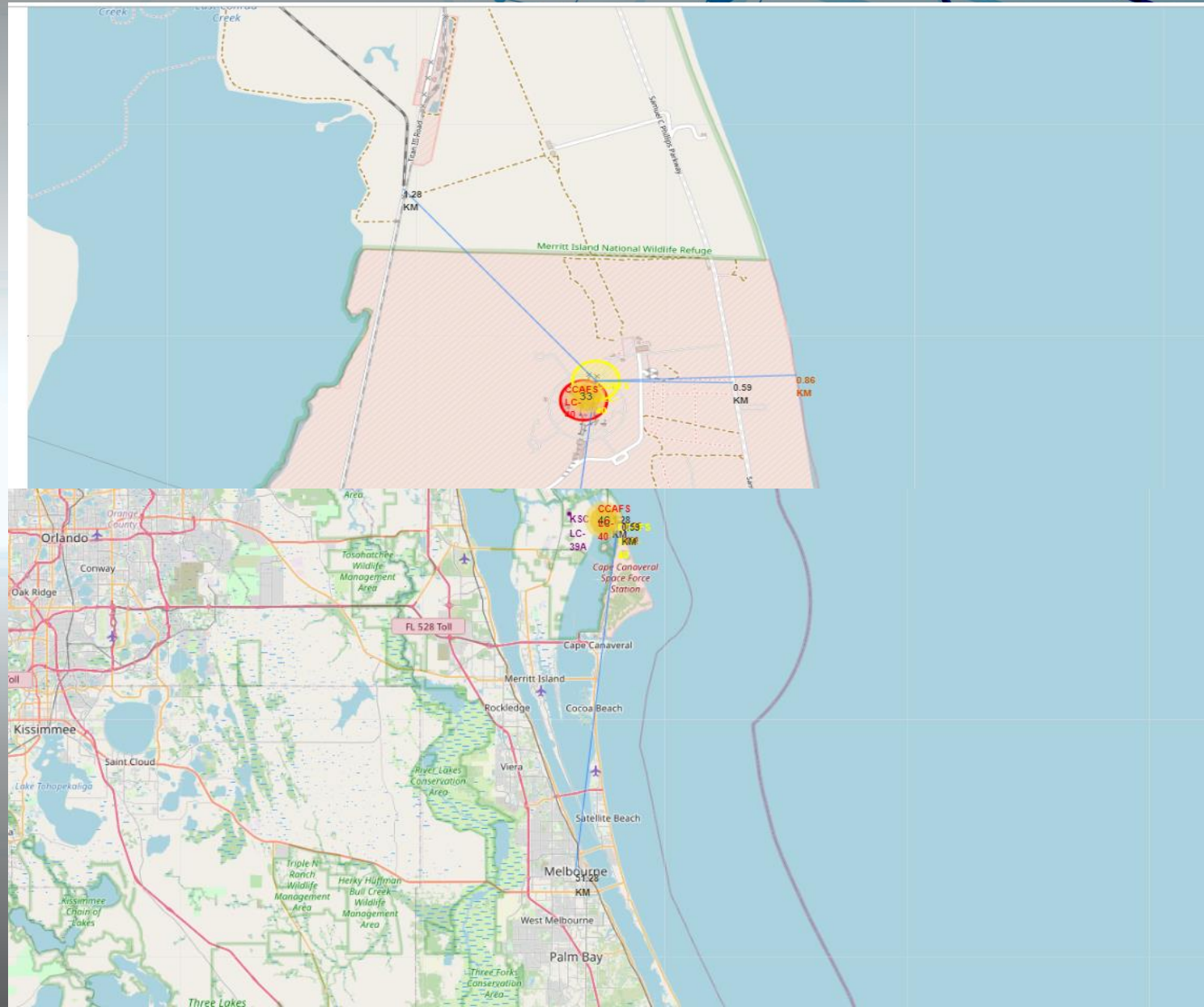


Results (Interactive Map with Folium II)

In the last map we have calculated the distance from a Launch Site CCAFS SLC-40 to the nearest points of the Coastline, Highway, Railway and City. And conneceted those points use PolyLines to the Launch Site CCAFS SLC-40.

It appears that the nearest point of the coastline is at a distance of 0.86KM and the nearest highway is 0.59KM away from the site. The nearest Railway Station/Line also appears to be close and at a distance of 1.28KM from the Launch Site.

Although those points are close to the Launch Site it appears that the Launch Site is pretty far away from the nearest City which is Melbourne at a distance of 51.28KM from it.

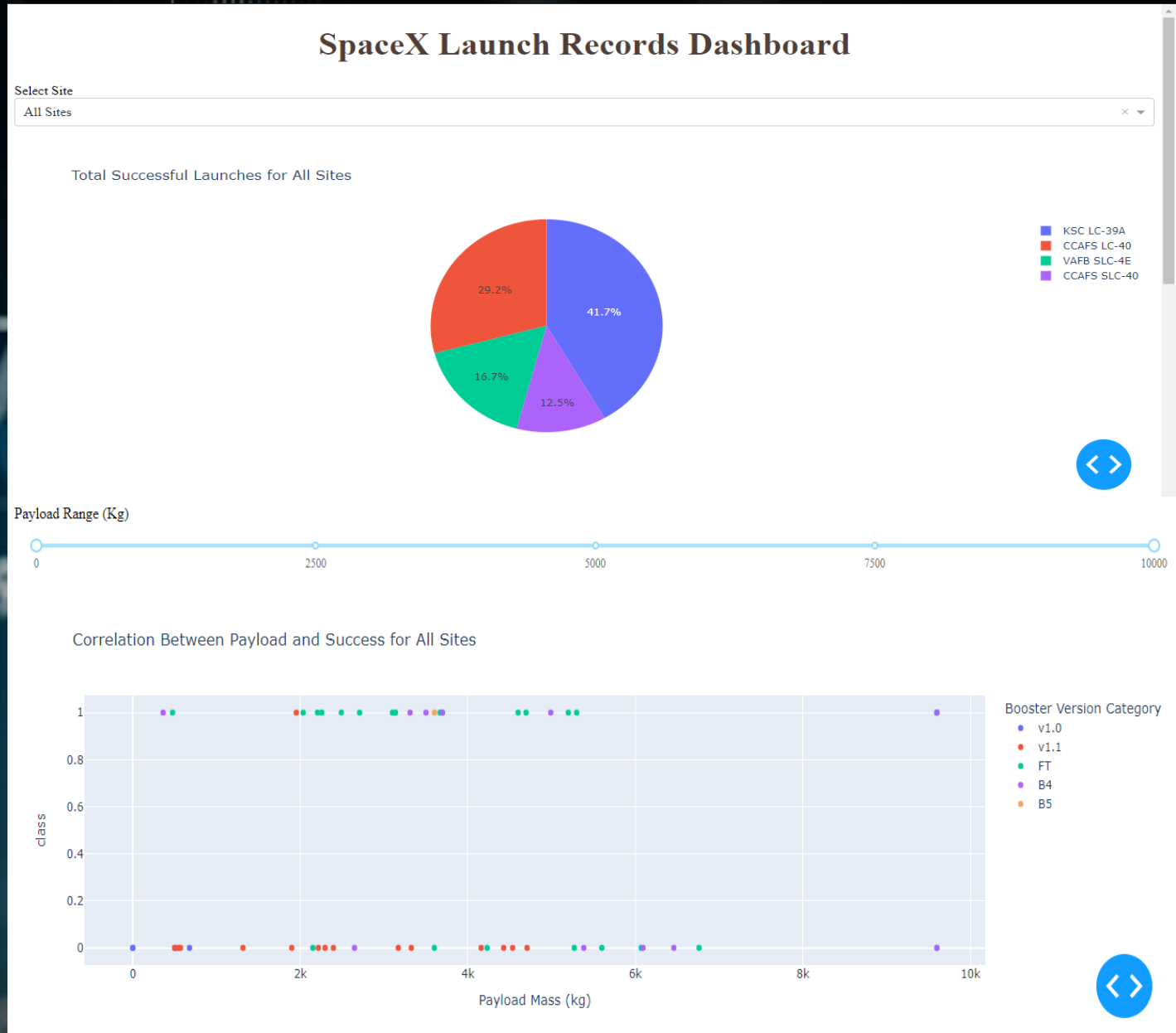


Results (Dashboard with Plotly)

We make a Dashboard using Plotly with the Callbacks for The different Launch Sites as well as All the Sites put together to retrieve the Success Rate of Landing Outcomes for each Launch Site using a Pie Chart.

We also add a RangeSlider to the Dashboard to be able to alter the Payload Mass (Kg) and it turn helps us to generate a Scatter Plot with the Success and Failure of each Landing Outcome for All Sites together as well as each Site separately.

Here, we can see the Percentage Success of All Sites in the Pie-Chart and Success and Failed Landing Outcomes for All Sites combined in the Scatter Plot. The different colors of the dots indicate the Booster Version used in the specific Launch.
The link to the Dashboard is: https://pulakpkashya-8050.theiadocker-1-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai/

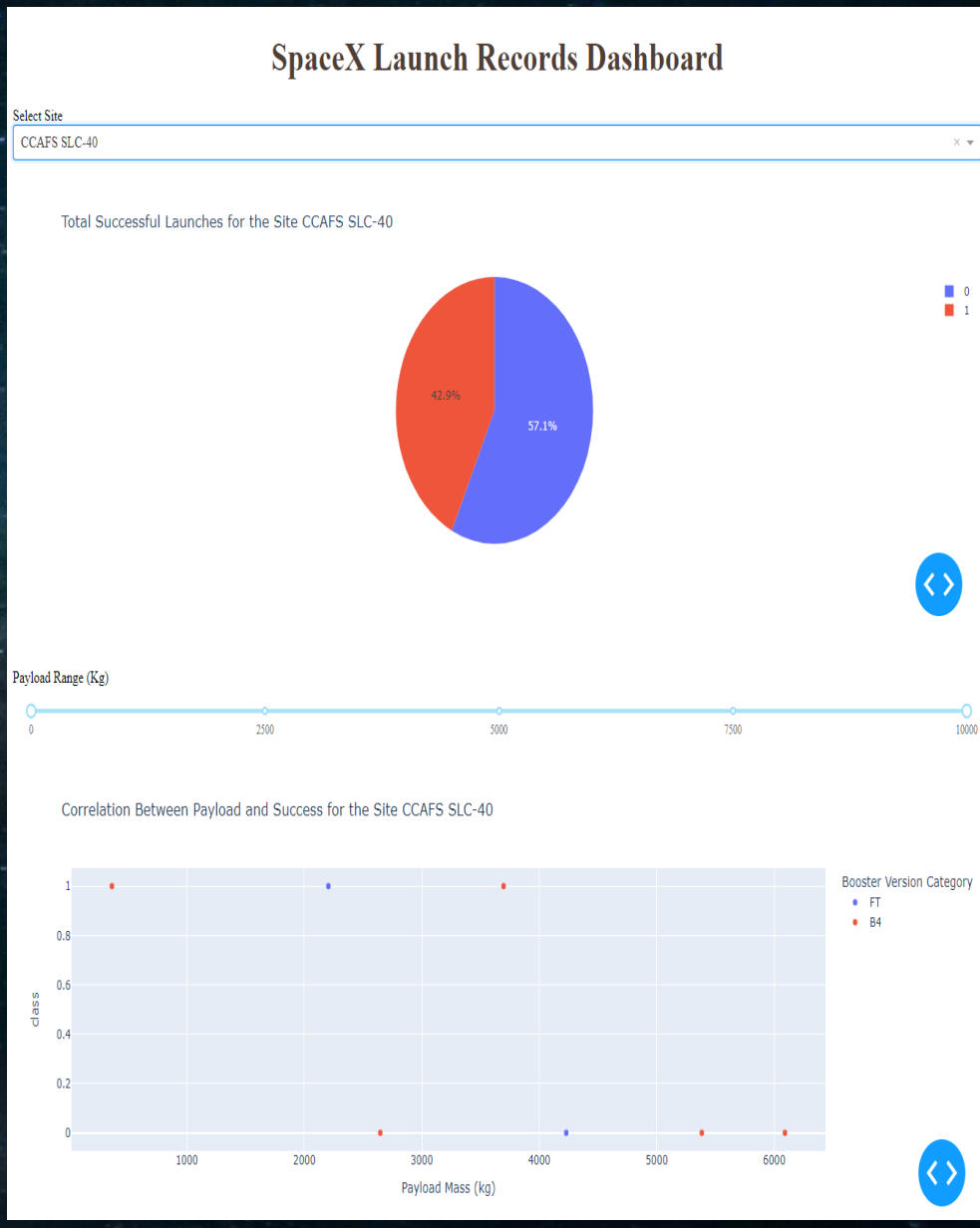


Results (Dashboard with Plotly II)

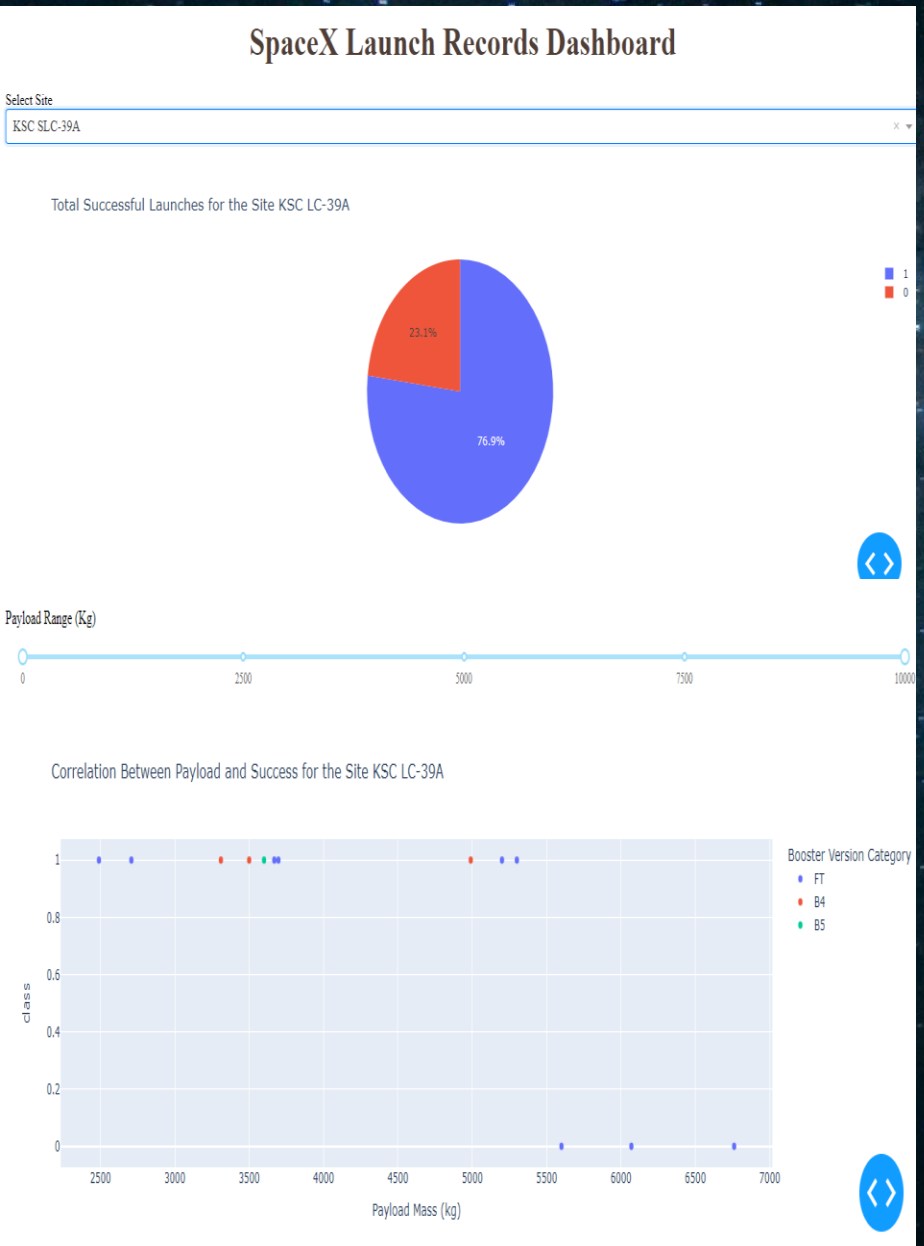
In the two Dashboards in the slide we have the Success and Failure Rate of the Landing Outcomes of the First Stage for the Sites CCAFS LC-40 and CCAFS SLC-40 where the Success Rates are 26.9% and 42.9% respectively in the Pie-Chart.

In the Scatter Plots for both the Sites we see that for CCAFS LC-40 the highest Launched Boosters are v1.1, followed by FT and v1.0 with FT having the highest Successful Landing Outcomes.

For CCAFS SLC-40 we see that there have been only 7 Launches of which 2 are FT and 5 are B4 with 1 and 2 successful landings resp.



Results (Dashboard with Plotly III)

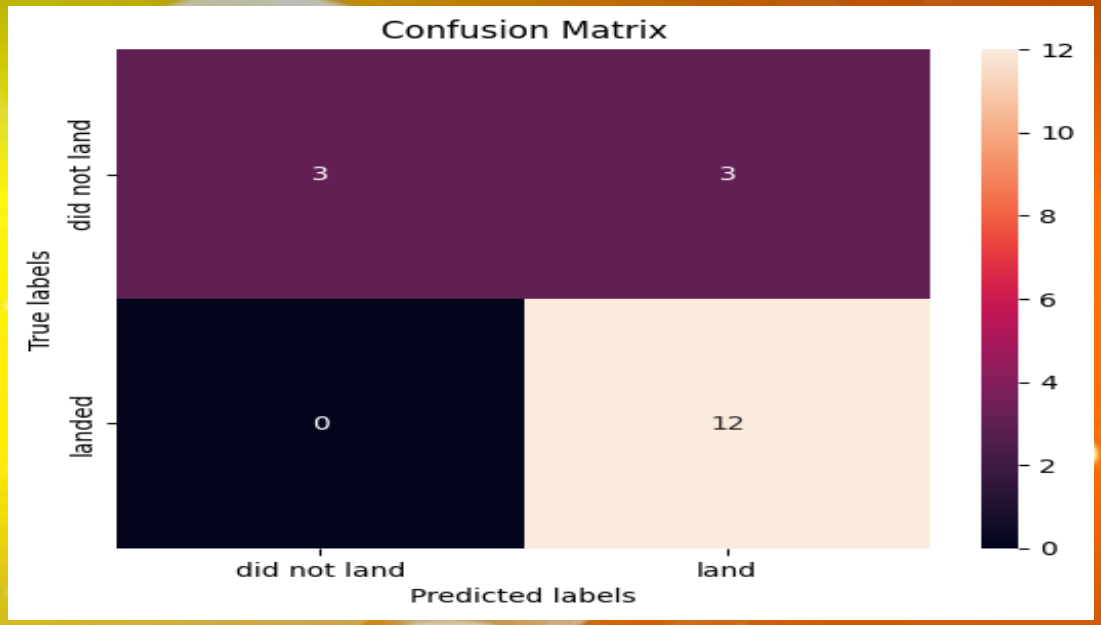


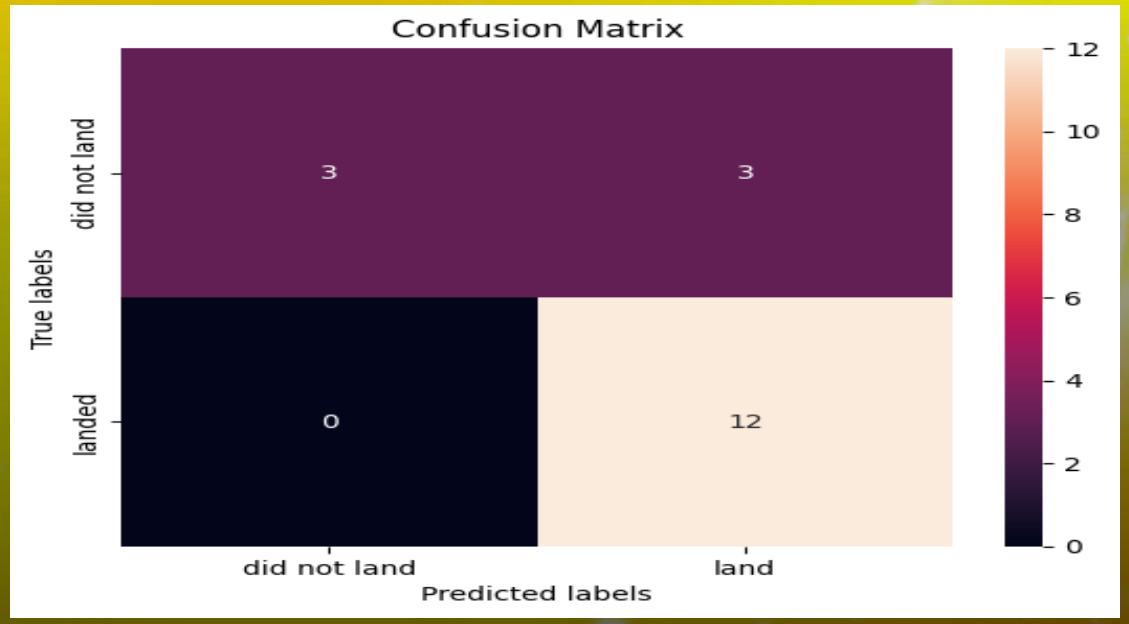
In the two Dashboards in the slide we have the Success and Failure Rate of the Landing Outcomes of the First Stage for the Sites KSC LC-39A and VAFB SLC-4E where the Success Rates are 76.9% and 40% respectively in the Pie-Chart.

In the Scatter Plots for both the Sites we see that for KSC SLC-39A have only had all the Launches below 5500 Kg Successful and the others as Failure. For VAFB SLC-4E we see that there has only been 7 launches with 2 being successful.

Results (Predictive Analysis)

The **Logistic Regression** using GridSearch CV gives a score of 0.84643 with the best hyperparameters of "Penalty": 'Ridge', 'Inverse of Regularization': 0.01 and Solver: 'lbfgs'. The accuracy in the Test Data comes out to be 83.333% and the Confusion Matrix, attached for reference, gives 3 wrong outputs (False Negative) with rest 15 Outputs being accurate. 3 True Positive; 12 True Negative.

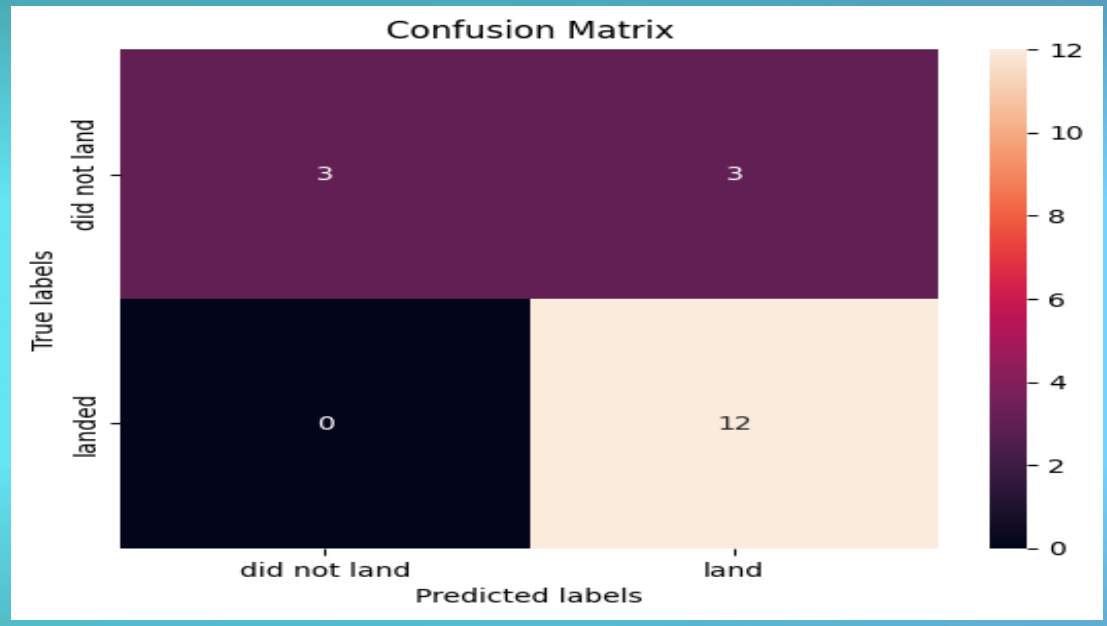


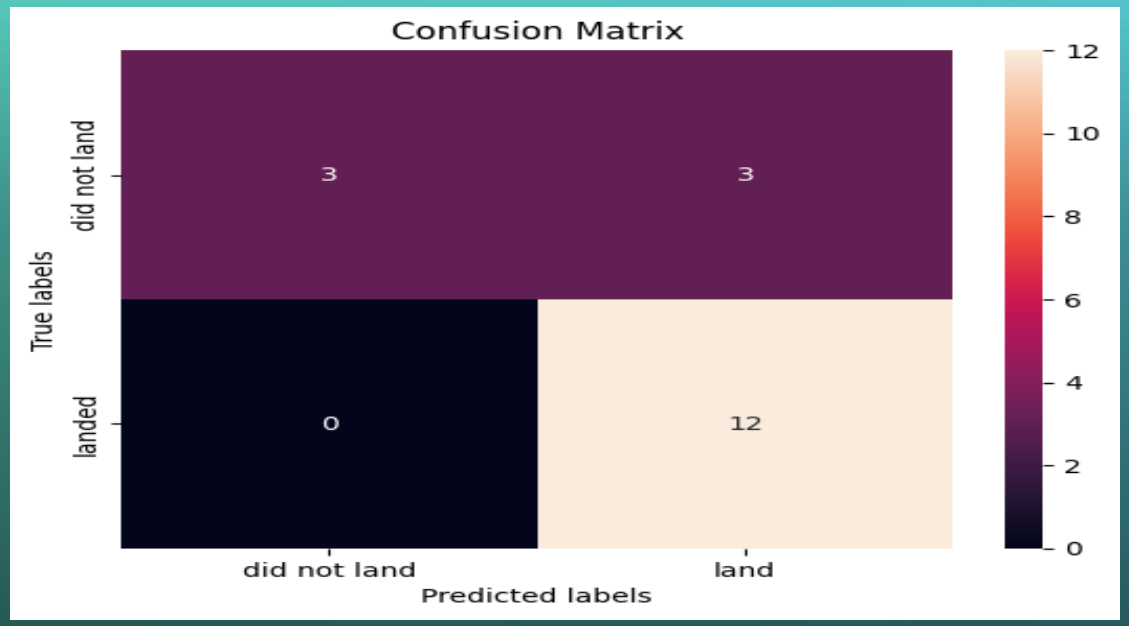


The **Support Vector Machine** algorithm using GridSearch CV gives a score of 0.82143 with the best hyperparameters of "Gamma": 0.031623, 'Inverse of Regularization': 1.0 and Kernel: 'Sigmoid'. The accuracy in the Test Data comes out to be 83.333% and the Confusion Matrix, attached for reference, gives 3 wrong outputs (False Negative) with rest 15 Outputs being accurate. 3 True Positive; 12 True Negative.

Results (Predictive Analysis II)

The **Decision Tree Classifier** using GridSearch CV gives a score of 0.889286 with the best hyperparameters of 'criterion': 'gini', 'max depth': 18, 'max features': 'auto', 'min samples leaf': 4, 'min samples split': 10, 'splitter': 'random'. The accuracy in the Test Data comes out to be 83.333% and the Confusion Matrix, attached for reference, gives 3 wrong outputs (False Negative) with rest 15 Outputs being accurate. 3 True Positive; 12 True Negative.



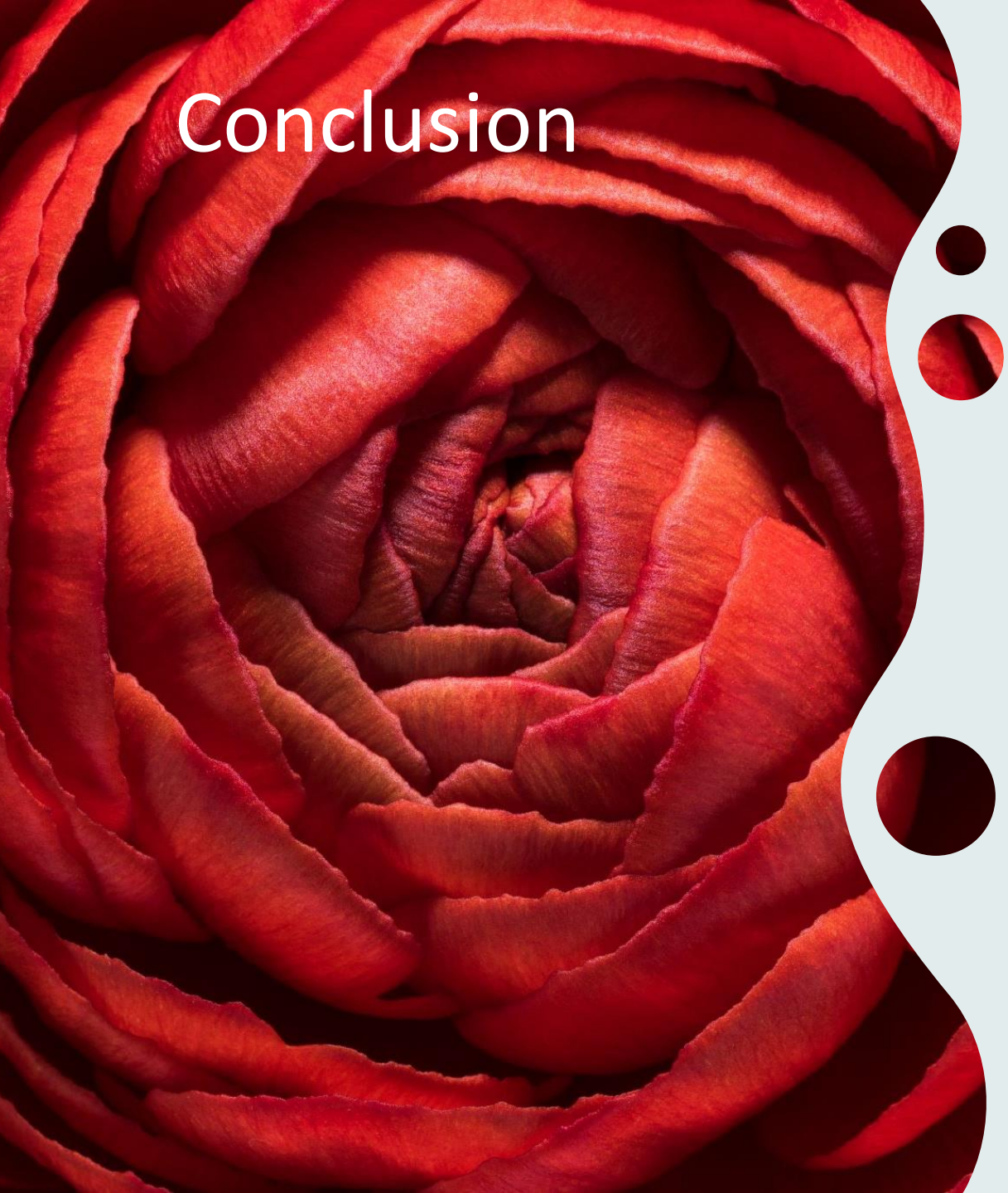


The **K-Nearest Neighbour** algorithm using GridSearch CV gives a score of 0.84821 with the best hyperparameters of 'algorithm': 'auto', 'n neighbors': 10, 'p': 1. The accuracy in the Test Data comes out to be 83.333% and the Confusion Matrix, attached for reference, gives 3 wrong outputs (False Negative) with rest 15 Outputs being accurate. 3 True Positive; 12 True Negative.

Discussion

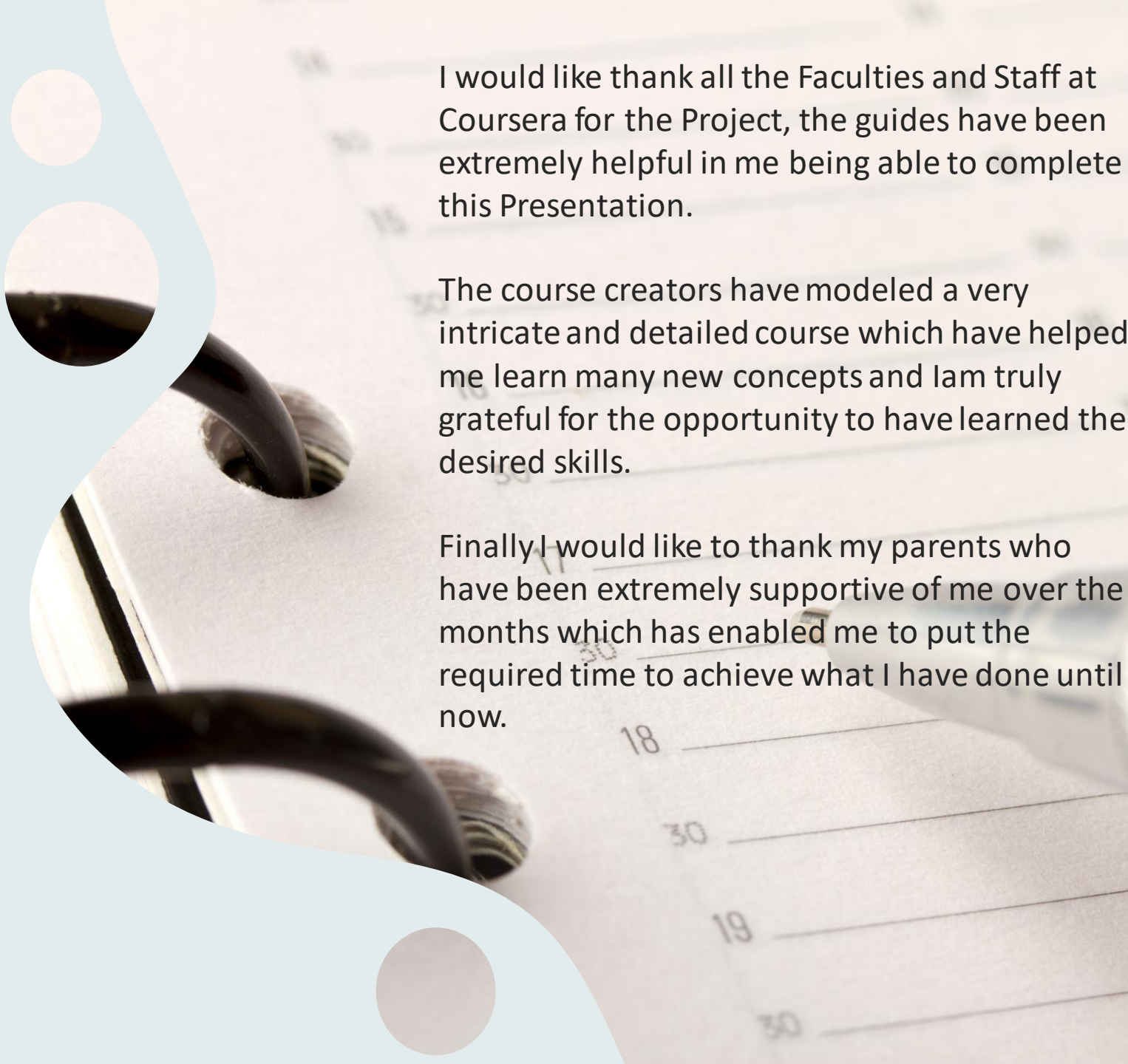
- We have seen that there are 4 Launch Sites for SpaceX out of which CCAFS LC-40 and KSC SLC-39A have seen the highest number of Launches and the other two having 7 each.
- The Payload Mass plays a vital role in determining the Landing Outcome of the First Stage and those with less Payload tend to be more successful in Landing.
- We have also observed that the ISS, GTO, VLEO, LEO are the Orbits to which the highest number of Launches has taken place though some of them has had less Success Rate of the Landing Outcome for the First Stage. The Orbits play a pivotal role in determining the Landing Outcome of the First Stage as well.
- The first successful Landing Outcome of the First Stage was in 2015 for ground pad after which the rate of success have been on the rise steadily making the missions more desirable and efficient for the company.
- We have observed that the Landing Outcome has also been affected by the Flight Number as with more experience the company has been able to achieve higher rate of success in Landing their First Stage.
- We have also seen that the Launch Sites tend to be further away from Cities, to not affect the general population in case something goes wrong. While they tend to be close to the Coastline, Highway and Railway so as to have a good connectivity

Conclusion



- We have seen that with time and experience SpaceX has been able to fundamentally increase their efficiency in Landing the First Stage of the Launch successfully which in turn has directly impacted their cost and have helped lower the same.
- There are various important factors that determine whether the First Stage would land back successfully; including Orbit to which it is Launched, Payload, Experience and the type of Landing Equipment used to land back the First Stage.
- The company has been able to immensely increase their Success Rate once they figured out what works in 2015 in which they recorded their first ever successful Landing Outcome for the First Stage.
- They usually use a definitive kind of Booster Version for every range of Payload Mass with Booster Version BT5 being preferred for the Launch of the highest Massed Payloads.

Acknowledgement



I would like thank all the Faculties and Staff at Coursera for the Project, the guides have been extremely helpful in me being able to complete this Presentation.

The course creators have modeled a very intricate and detailed course which have helped me learn many new concepts and Iam truly grateful for the opportunity to have learned the desired skills.

Finally I would like to thank my parents who have been extremely supportive of me over the months which has enabled me to put the required time to achieve what I have done until now.