Winning Space Race with Data Science

Devank Banga
14-12-2023

Devank Banga

Outline

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

Executive Summary

- SpaceY has hired us to predict the prospective cost of the SpaceX bidding as it aims to counter bid them. SpaceX bid is heavily dependent on the successful landing of the First Stage which allow them to re-use it and heavily reduce the overall cost of the next launch and so on.
- To satisfy the requirement of SpaceY we have created multiple models to predict the landing of the First Stage and obtained a Dense Neural Network which is able to predict accurately the test dataset with 83.33% accuracy. We also have provided much simpler Machine Learning Models with accuracy of 83.33% such as Logistic Regression, SVM, KNN, and Decision Trees.
- We obtained these results by obtaining the publicly available SpaceX dataset from SpaceX and Wikipedia and tidied it up, and then performed EDA and modeling on it.

Introduction

- SpaceY is in direct competition against SpaceX and their success is based on the outbidding SpaceX by bidding a lower price for the same launch. To do this they need to predict the price SpaceX will quote, which in turn is based on if SpaceX previous launch has a successful First Stage landing or not.
- In this project we aim to understand the pattern of SpaceX First Stage landings and create models to predict them and then suggest the most suitable one to the SpaceY.

Section 1

Methodology

Methodology

In this part we will cover various Methodologies we used for following parts of the project:

- Data collection
- Data wrangling
- Exploratory Data Analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

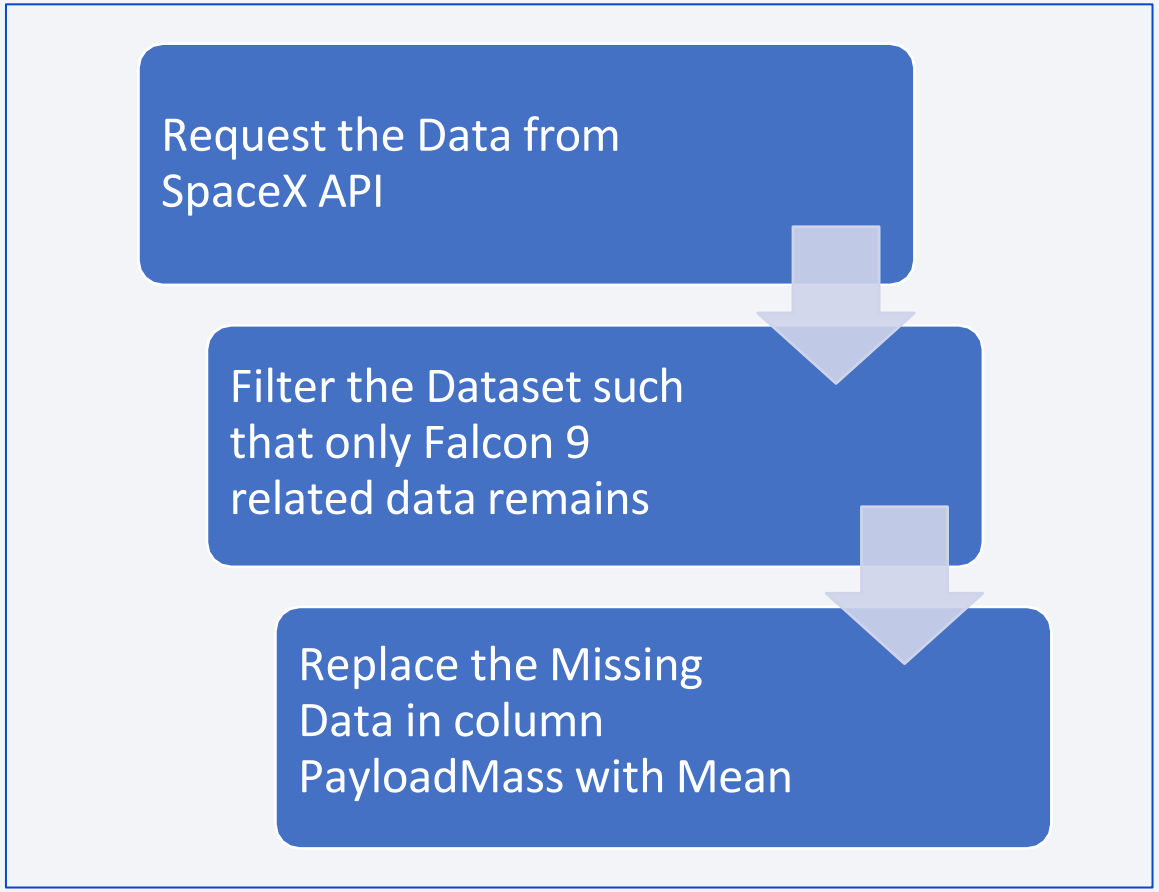
Data Collection

- SpaceX has provided the open-source API for their data, we used that to obtain the Falcon 9 related data so that we can predict if the First Stage Landing will succeed or fail and use that to make our bid. We used multiple python basic packages to complete this task.
- We will also use data related to SpaceX First Stage Success and Failure rate available publicly at Wikipedia.com.
- After obtaining the dataset required for our project we will further clean it to prepare it for EDA and Predictive Analytics.

Data Collection - SpaceX API

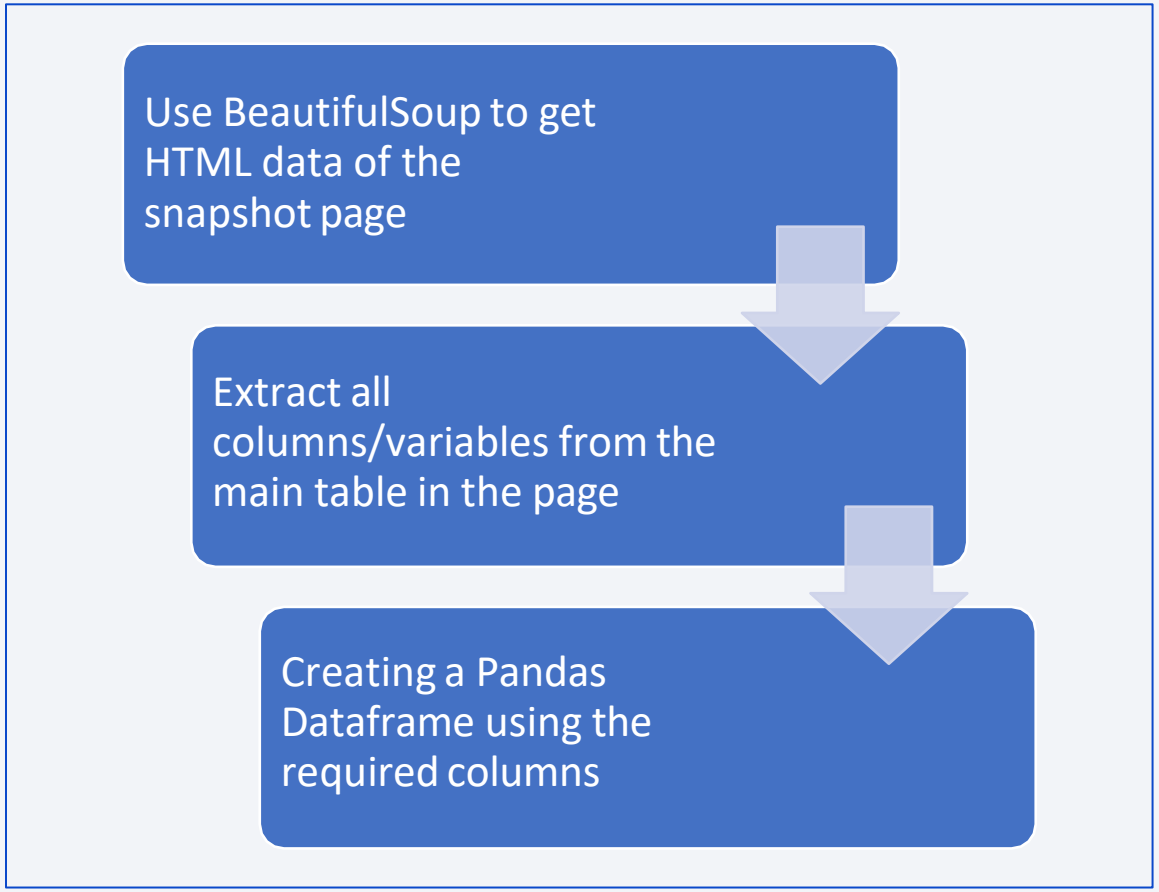
- Following the process explained in previous slide we obtained a dataset with 90 rows and 18 columns.

- GitHub URL for SpaceX API Fetch Notebook:https://github.com/ukiyoeh/IBM-Data-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- SpaceX API is not the only source of data we can use, we also used Wikipedia to enrich our dataset.
- For this project we used the static Wikipedia page with the title List of Falcon 9 and Falcon Heavy launches, it was last updated on 9th June 2021.
- GitHub URL for Web Scraping Notebook:https://github.com/ukiyoeh/IBM-Data-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- After obtaining the Data Set in previous stage we performed multiple steps to tidy the data in the Data Wrangling stage. These steps range from removing null values to one hot encoding, to feature engineering where required.
- Since dataset itself has less than 100 observations we dealt with the null values by replacing them, and for our goal we need to focus on the landing outcomes of the various Falcon 9 First Stages, we performed feature engineering to make it. This column will be the base for the target variable we will use in the Predictive analytics and EDA in later stages.
- GitHub URL for the Data Wrangling Notebook: https://github.com/ukiyoeh/IBM-Data-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- After obtaining a clean dataset we performed basic EDA. We calculated multiple summaries and used various plots to understand the data better.
- In this stage we primarily focused on the Orbit, Launch Site, Flight Number, PayloadMass, Class, and Date columns. We used multiple plots like Line Charts to Visualize Temporal Date column, Box Plot to understand the distribution of PayloadMass across different launch sites with respect to Success and Failure class for the First Stage Landing, we also performed many more analysis which can be accessed in depth using the following GitHub Notebook link.
- GitHub URL of the EDA with data visualization notebook: https://github.com/ukiyoeh/IBM-Data-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- While we used Python for most of the part of the project we also employed SQL for the EDA to not miss any form of understanding we can get from the dataset. We obtained following interesting results from it:
  - Total Payload Carried by boosters from NASA(CRS) as a customer is 45,596 Kgs.
  - Average Payload Carried by Falcon F9 v1.1 is 2534.67 Kgs
  - First Successful Ground landing was recorded on 2015-12-22
  - Between 2010-06-04 and 2017-03-20 most successes were recorded via Drone Ship
  - F9 FT B102, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2 were only Boosters which had successful launches via Drone Ship and having payload between 4,000 and 6,000 Kgs.
- More analysis can be found via this GitHub Link: https://github.com/ukiyoeh/IBM-Data-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Interactive Maps with Folium

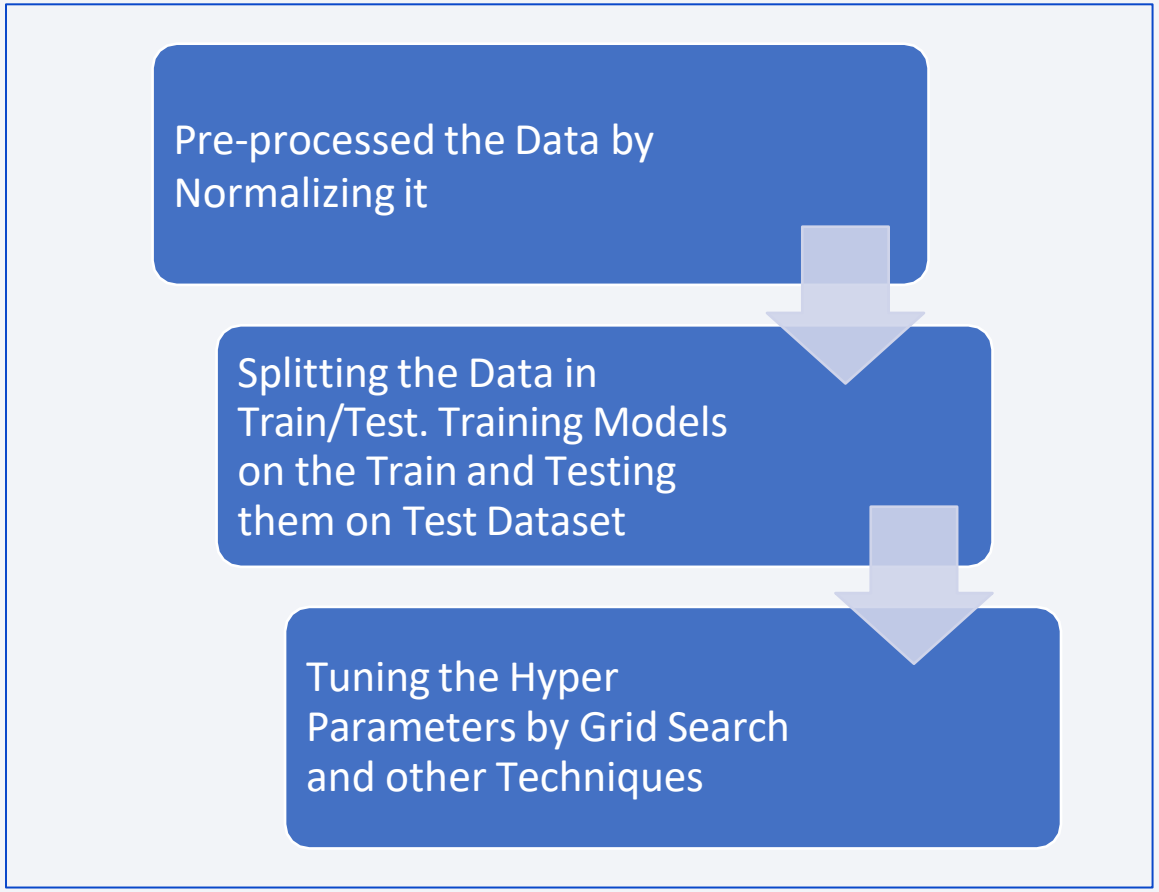
- We had geo-spatial data in the data obtained from the web which we didn't use in EDA. We used that to create Interactive World Maps using Folium Package in python.
- First, we marked all the Launch Sites, and then marked the Success/Fail depending on the launch attempts for each site by using Green/Red color respectively.
- After that we calculated distance between Site closest to the coast and the coastline, we also checked the distance between Sites and major Highways, Railway Stations, and Cities to find out if these launch were specifically established far away from public.
- More details can be found in the Folium Notebook with its GitHub URL given here: https://github.com/ukiyoeh/IBM-Data-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Dashboard with Plotly Dash

- In this Stage we used Plotly Dash Package to create interactive web based Dashboard for the some of the important findings obtained during EDA and further analyzing the Geo-Spatial data.
- In the Dashboard we created a Pie Chart and a Scatterplot. To control the Pie Chart and Scatterplot we have a Drop Down list containing the options for Launch Sites and to further interact with the Scatter chart containing Class Vs Payload Mass(Kg) we created a slider.
- We have uploaded two files for this part, the main Jupyter Notebook whose GitHub link is given below. We also created a separate Python script which can be operated on different IDEs:https://github.com/ukiyoeh/IBM-Data-Capstone/blob/main/dash_app.py

Predictive Analysis (Classification)

- After the EDA and creating a Dashboard for different Stakeholders we performed Predictive Analytics on the SpaceX dataset to predict the success/failure of the First Stage Landings so that we can use that for our own counter bidding.
- We used techniques like Logistic Regression, SVM, KNN, Decision Tree, and Dense Neural Network following the general steps given in the flowchart here. We will compare the performance of multiple models to decide on our desired model for this project.
- GitHub URL for the Predictive analysis lab: https://github.com/ukiyoeh/IBM-Data-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

- Based on the results of the Methodologies we mentioned in this section, we will cover the following parts in the report:
  - EDA Results for both SQL and Python: Insight Drawn From EDA
  - Interactive analytics demo in screenshots for Folium Maps and Dashboard
  - Predictive analysis results

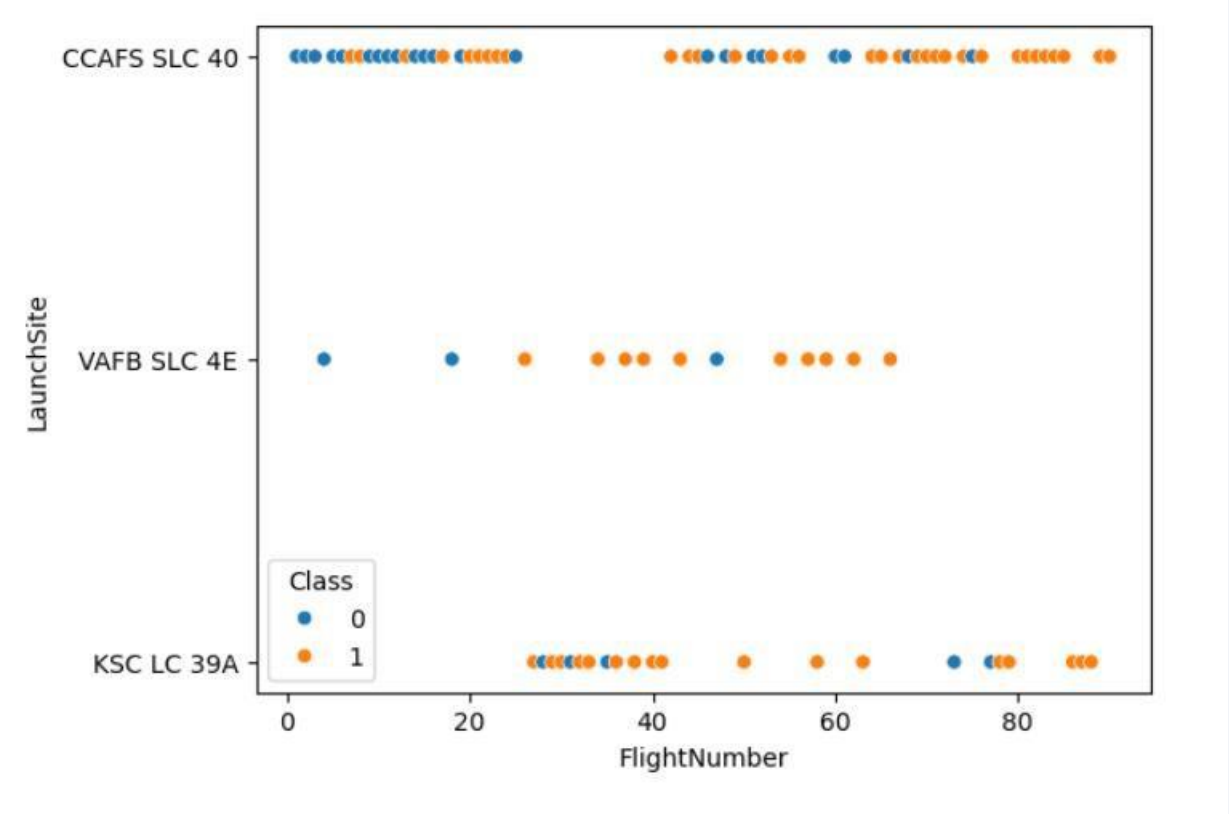
Section 2

Insights drawn from EDA

Devank Banga

Flight Number vs. Launch Site

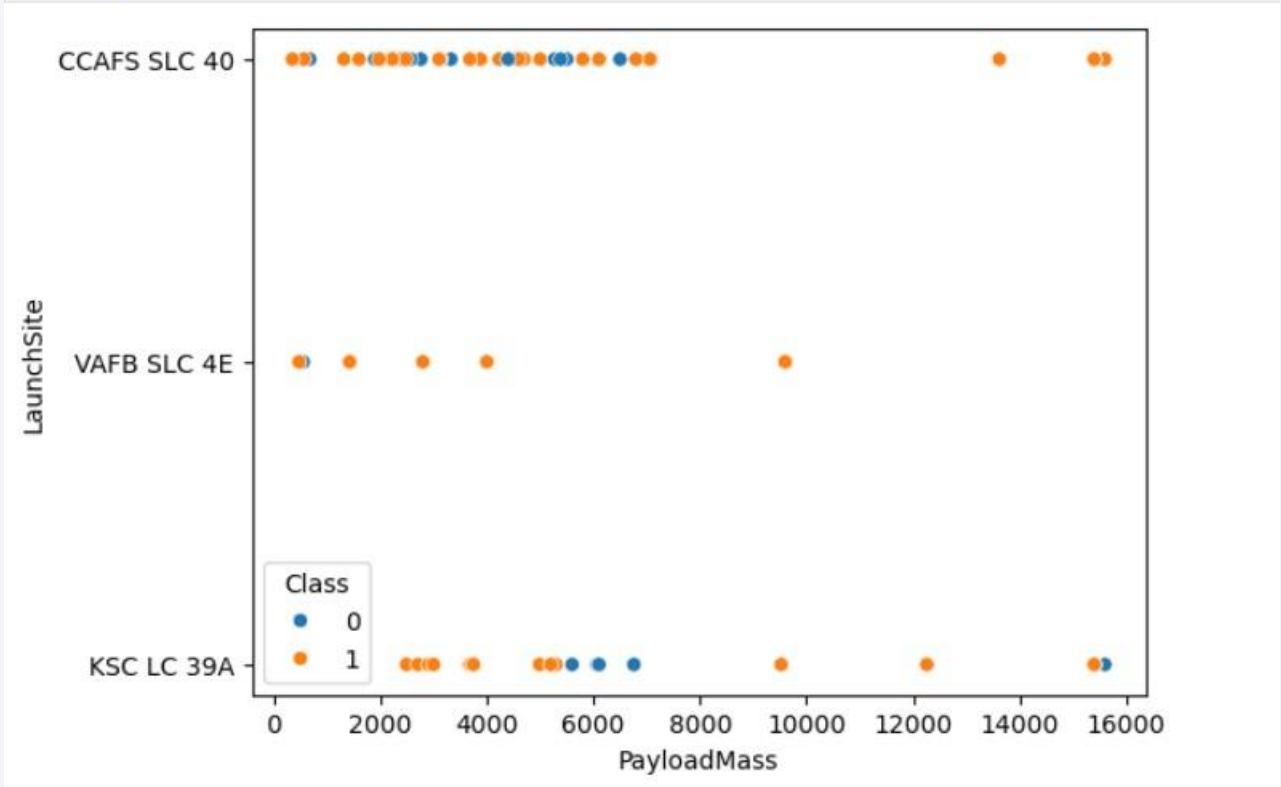
- In this plot we can observe the success/failure of multiple flights across different Launch Sites.
- Early Flights were launched from CCAFS SLC 40 most of the time and their First Stage landing failed a lot.
- However, as Flight Number increased we can see greater proportion of successes across all the launch sites given here.



Payload vs. Launch Site

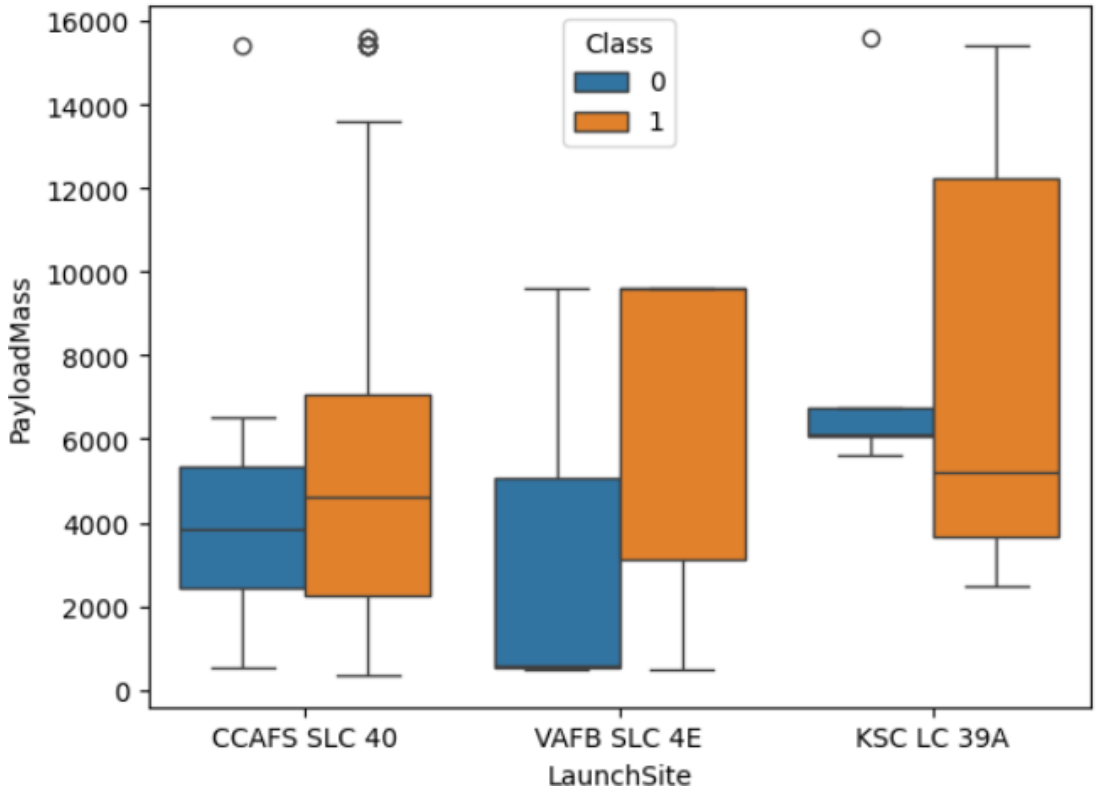
- CCAFS SLC 40 has most number of flights for PayloadMass less than 8,000 Kgs. It may be appearing so because it was the most used launch site in the beginning of SpaceX.
- Besides that, VAFB SLC 4E has least number of flights and has payload less than 11,000 Kgs but it has higher proportion of successful First Stage Landings.

```
sns.scatterplot(data = df, x = 'PayloadMass', y = 'LaunchSite', hue = 'Class')
plt.show()
```



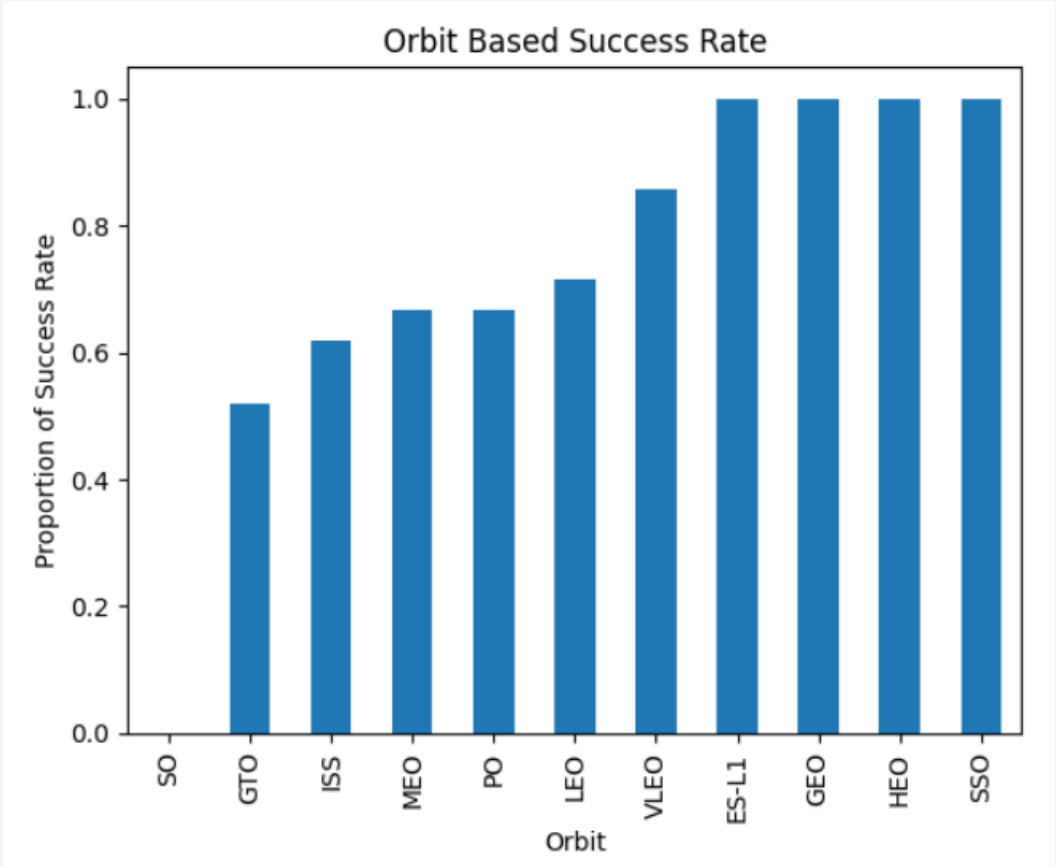
Payload vs. Launch Site - Boxplot

- We have limited data; these distributions can change with more data points. Regardless of that, from the boxplot we can see that distributions of both Classes are different across different Launch Sites.
- Most normal observation is from CCAFS SLC 40 and this is precisely because we have more data from this site. We need more data points to analyze the difference in distributions across class among the Launch Sites.



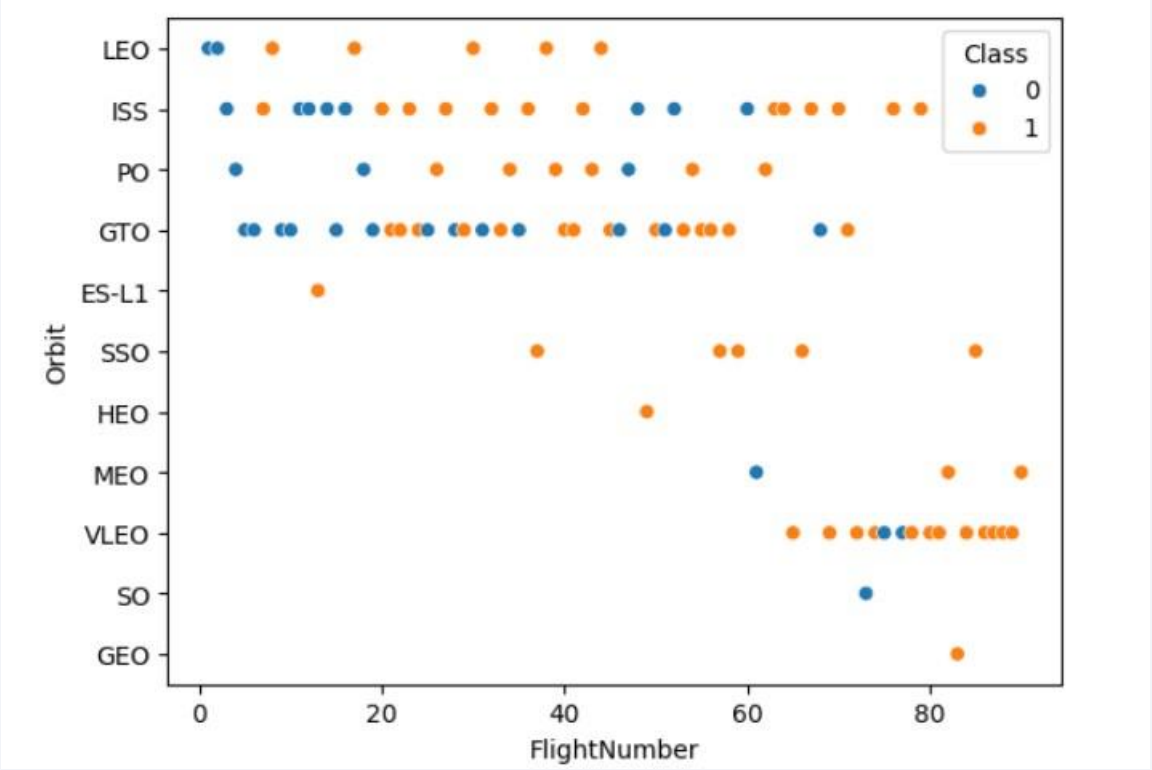
Success Rate vs. Orbit Type

- We have 0% success rate in SO and near 100% success rates in First Stage Landings where we are reaching SSO Orbit Type. And this is a point of concern because SSO and SO both relates to Sun-Synchronous Orbit so we need to further analyze data to find such a large difference.
- Besides that, SpaceX has commendable Success rates in First Stage Landings for most of the Orbit types with only GTO having around 50% success rate besides SO.



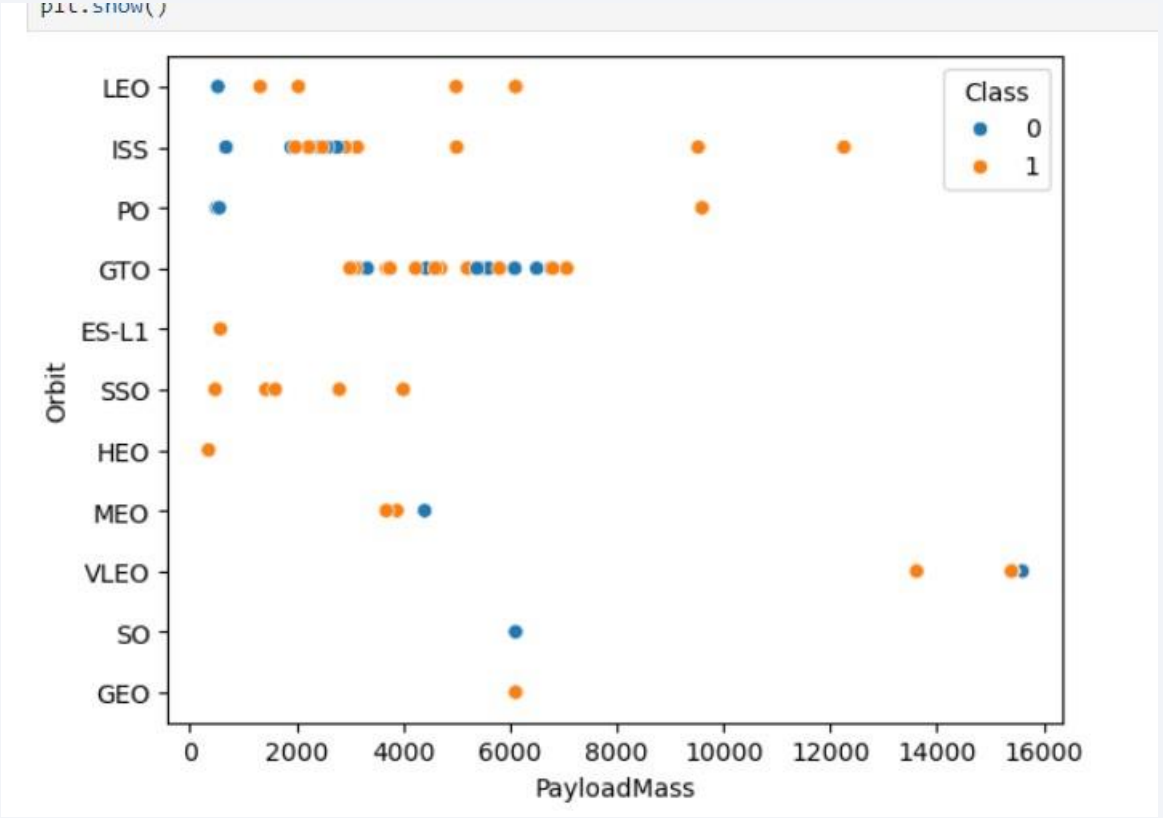
Flight Number vs. Orbit Type

- Following from previous slide we can see more details about Orbit Types.
- Earlier Flights were only focusing on the low orbit types and in Flight Number more than 40 or so we are seeing more higher altitude Launches.



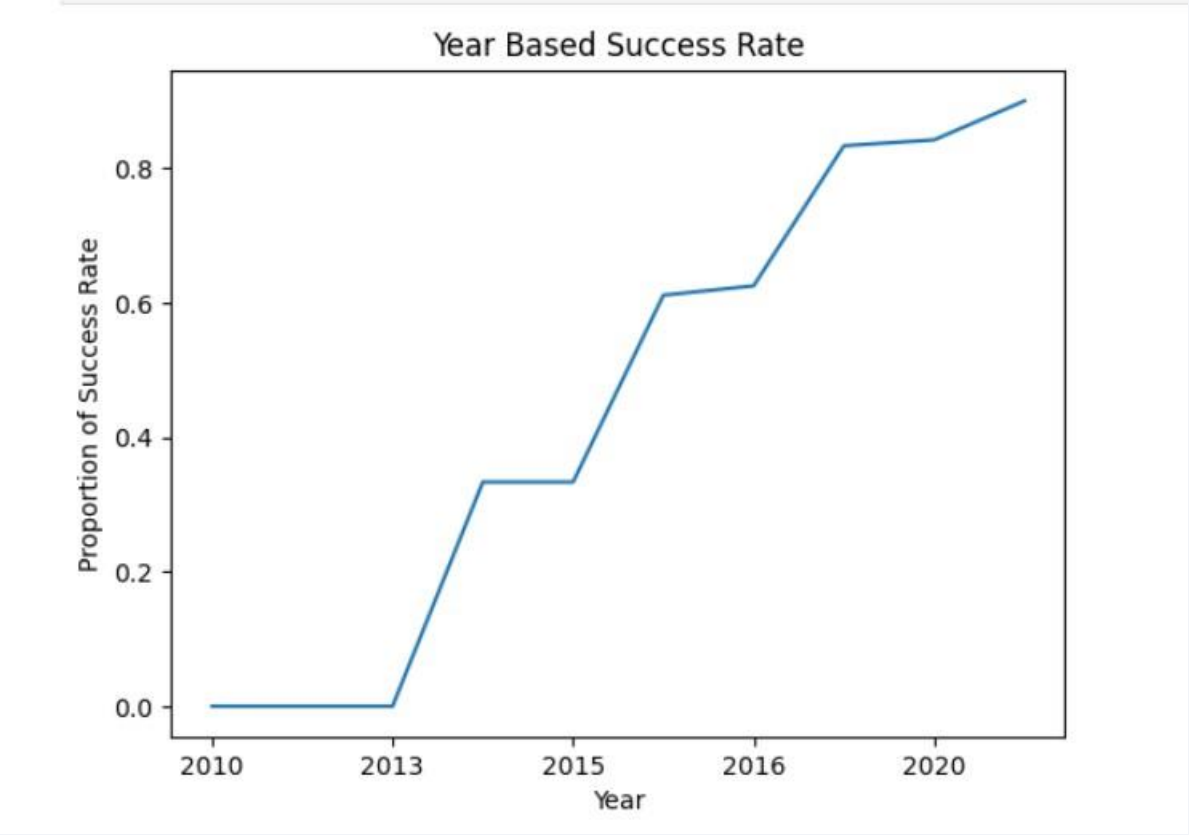
Payload vs. Orbit Type

- GTO Orbit only took payload mass of less than 8000 Kgs, this may point to some restriction in this orbit type launch.
- LEO, ISS, PO, and VLEO orbit types can support both low and high Payload Mass.



Launch Success Yearly Trend

- SpaceX has focused a lot on improving the First Stage Landings and it can be seen from the plot.
- From 2013 we can see a rising success rate in First Stage Landing, with it reaching near 90% or more after the year 2020.



All Launch Site Names

- We can identify Unique Launch Sites via Python Pandas and via SQL. Below we have shown the SQL query to obtain the result.
- SpaceX used CCAFS LC 40, VAFB SLC 4E, KSC LC 39A, and CCAFS SLC-40 as launch sites for the Falcon rockets. All of them are near the coastlines of United States.

```
]: %%sql
   SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
]:
```

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

Launch Site Names Begin with 'CCA'

- In this query we found the first 5 records containing Launch Sites starting with CCA.

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

```
 * sqlite:///my_data1.db
Done.
```

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

- In the following query we can see that SpaceX has carried a total of 45,596 Kgs worth of Payload from NASA.

```
%%sql
SELECT Customer, SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Customer	Total_Payload_Mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Average Payload Mass Carried by Falcon F9 v1.1 is 2534.67 Kgs.

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass_F9V1 FROM SPACEXTABLE WHERE Booster_Version LIKE '%F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

Average_Payload_Mass_F9V1
2534.6666666666665

First Successful Ground Landing Date

- First Successful Ground Pad Landing date was 2015-12-22. This also reflect previously observed trend which showed the low success ratio for First Stage Landings of SpaceX before 2013 as only in 2015 it got first Success Ground Landing.

```
%%sql
SELECT MIN(Date) AS first_ground_landing FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

first_ground_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- F9 FT B102, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2 were only Boosters which had successful landings via Drone Ship and having payload between 4,000 and 6,000 Kgs.

```
%%sql
SELECT  Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (drone ship)' AND
(PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Overall, we have 99 successes, 1 Success but payload status is unclear, and only 1 failure in Mission Outcome. Even if SpaceX had variable Success rate for First stage landings, it maintained a near perfect success rate in terms of the mission outcomes.

```
%%sql
SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLE GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

- Here we have a list of boosters able to carry the Maximum Payload. All of them are variants of F9 B5.

```
%%sql
SELECT DISTINCT Booster_Version FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM  SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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 this query we are can observe that in year 2015 we had two failures in landing via Drone Ships and both came from same Launch Site and similar Booster Version, that too in only 3 months gap.

```
%%sql
SELECT substr(Date, 6, 2) AS Month, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) == '2015' AND Landing_Outcome LIKE 'Failure (drone ship)'
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Most successes and failures happened via Drone Ship, reflecting that most landing attempts were made via Dron Ships during this period.
- Ground Pad has highest proportion of successes with no recorded failure in landing during this period.

```
%%sql
SELECT Landing_Outcome, COUNT(*) as Counts
FROM SPACEXTABLE
GROUP BY Landing_Outcome
HAVING Date BETWEEN '2010-06-04' AND '2017-03-20'
ORDER BY Counts DESC
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Counts
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

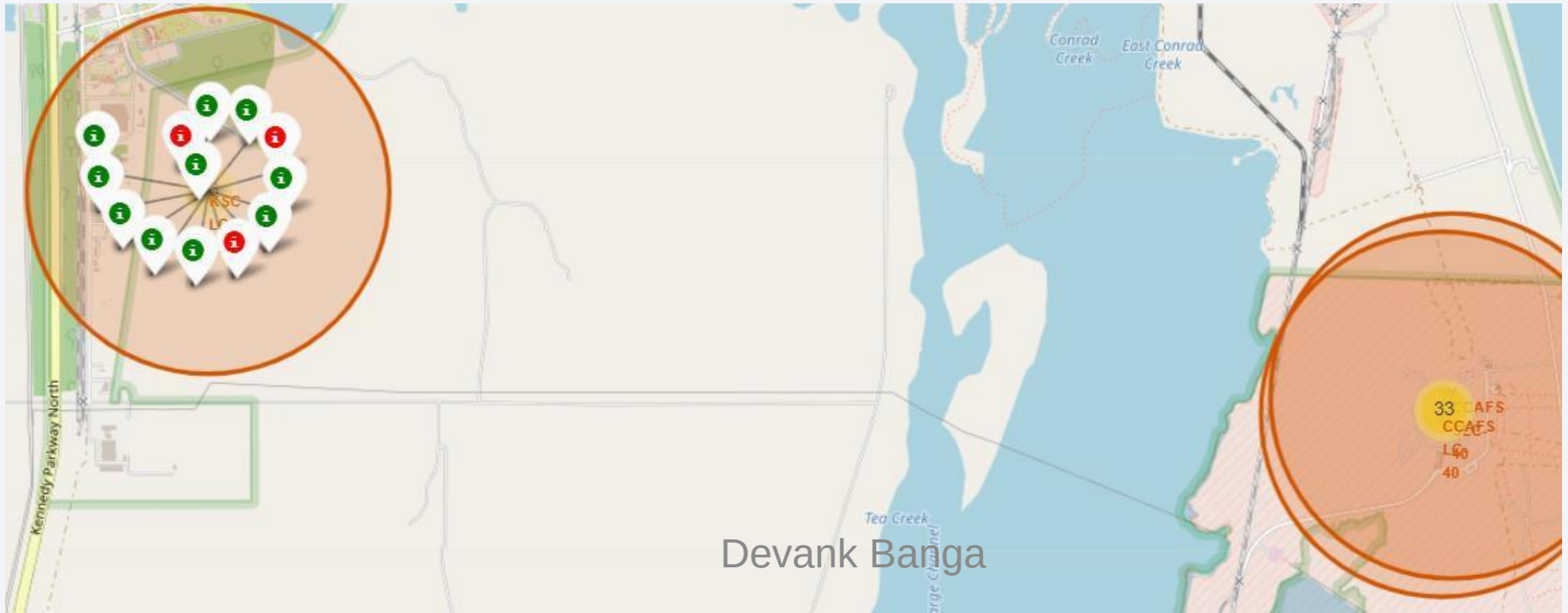
Devank Banga

Launch Sites

- All 4 sites are located in Costal Regions of United States with 3 of them located is at the East, very close to each other.

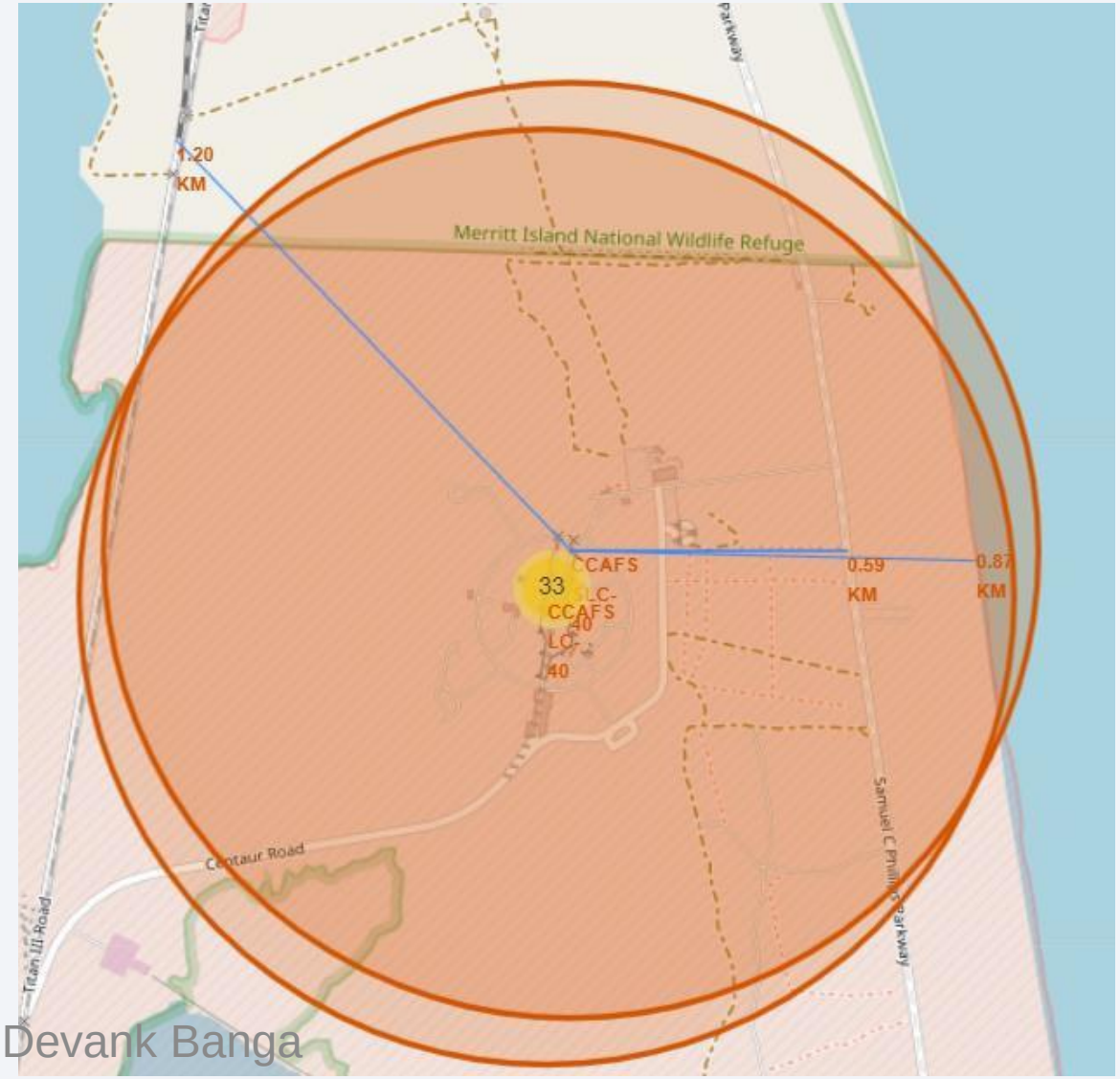
Successful/Failed Landings

- Now a closeup shot of the 3 launch sites at East. CCA launch sites are at same place while KSC is slightly far away.
- We can also observe Successful launches in Green and Failures in Red for KSC Launch Sites in form of Marker objects, it has highest success rate too.



Site Proximity Analysis

- All Launch sites are far away from cities and near costal regions with all of them within 1 km distance from costal regions.
- All of them are also near Highways.
- In this snapshot we have lines and distance from CCA SLC 40 specifically to highlight these points. It is 0.87 Km away from Costal Line, 0.59 Km from Highway, and 1.2 Km from Nearest rail line, and even beyond that for city.



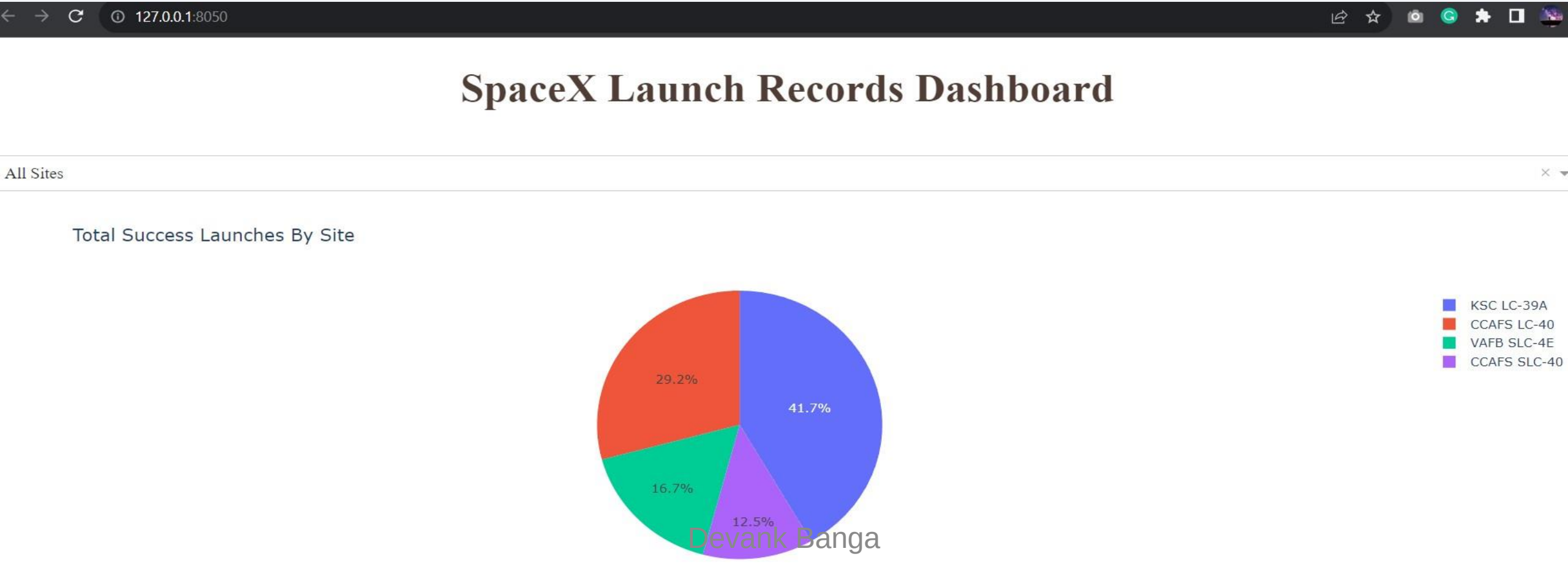
Section 4

Build a Dashboard with Plotly Dash

Devank Banga

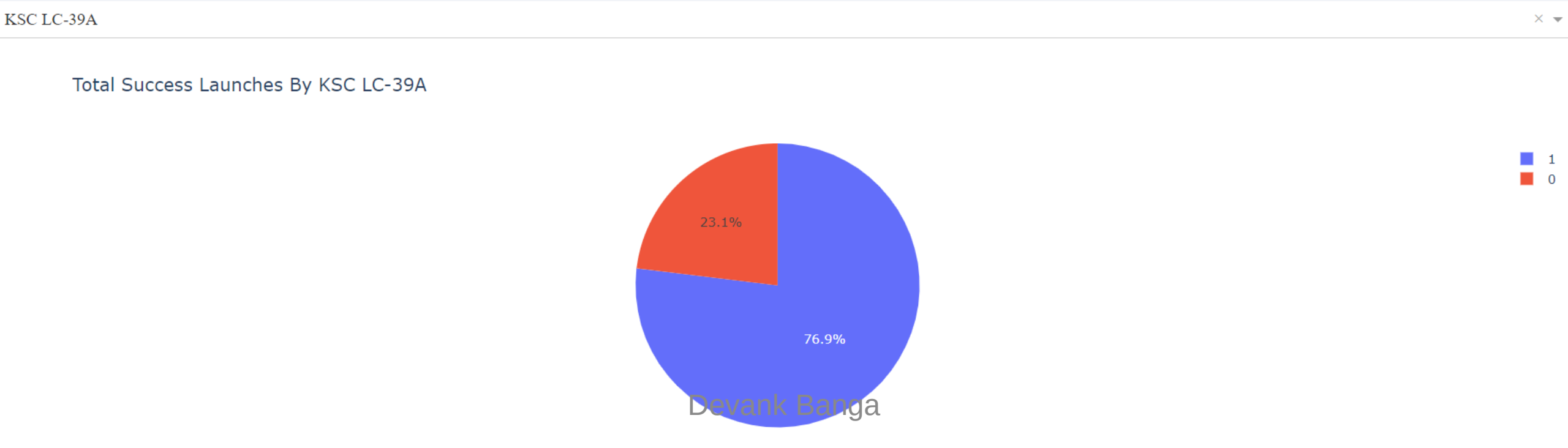
All Sites Interactive Pie Chart

- KSC LC 37A has highest Successful First Stage Landing ratio at 41.7% of total successes. In contrast CCAFS SLC 40 has lowest success ratio with 12.5%.



KSC LC-37A Success Rate Pie Chart

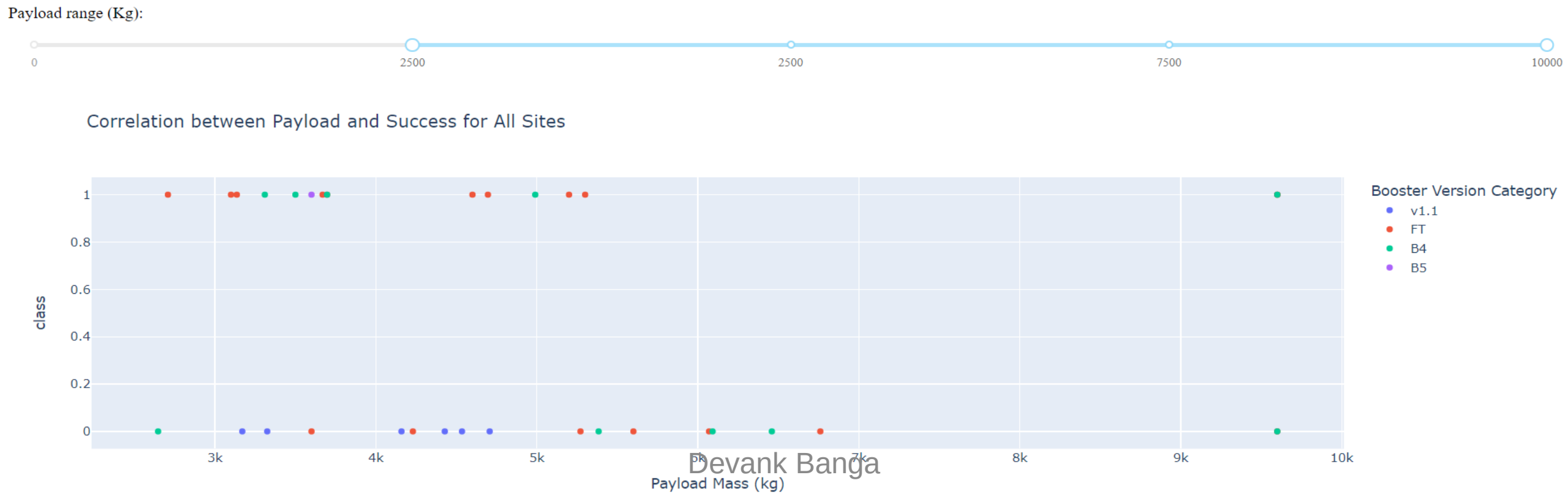
- KSC LC-39A has a success rate of 76.9% with Class 1 taking the same proportion. In contrast CCAFS SLC-40 has lowest success rate at 57.1%.



Devank Banga

Payload-Class Scatter Plot Based on Booster Version

- Within Payload range of 2,500 and 10,000 for all Sites we don't have a single successful Landing from Booster v1.1, in contrast B5 has a 100% success rate within same range.



Devank Banga

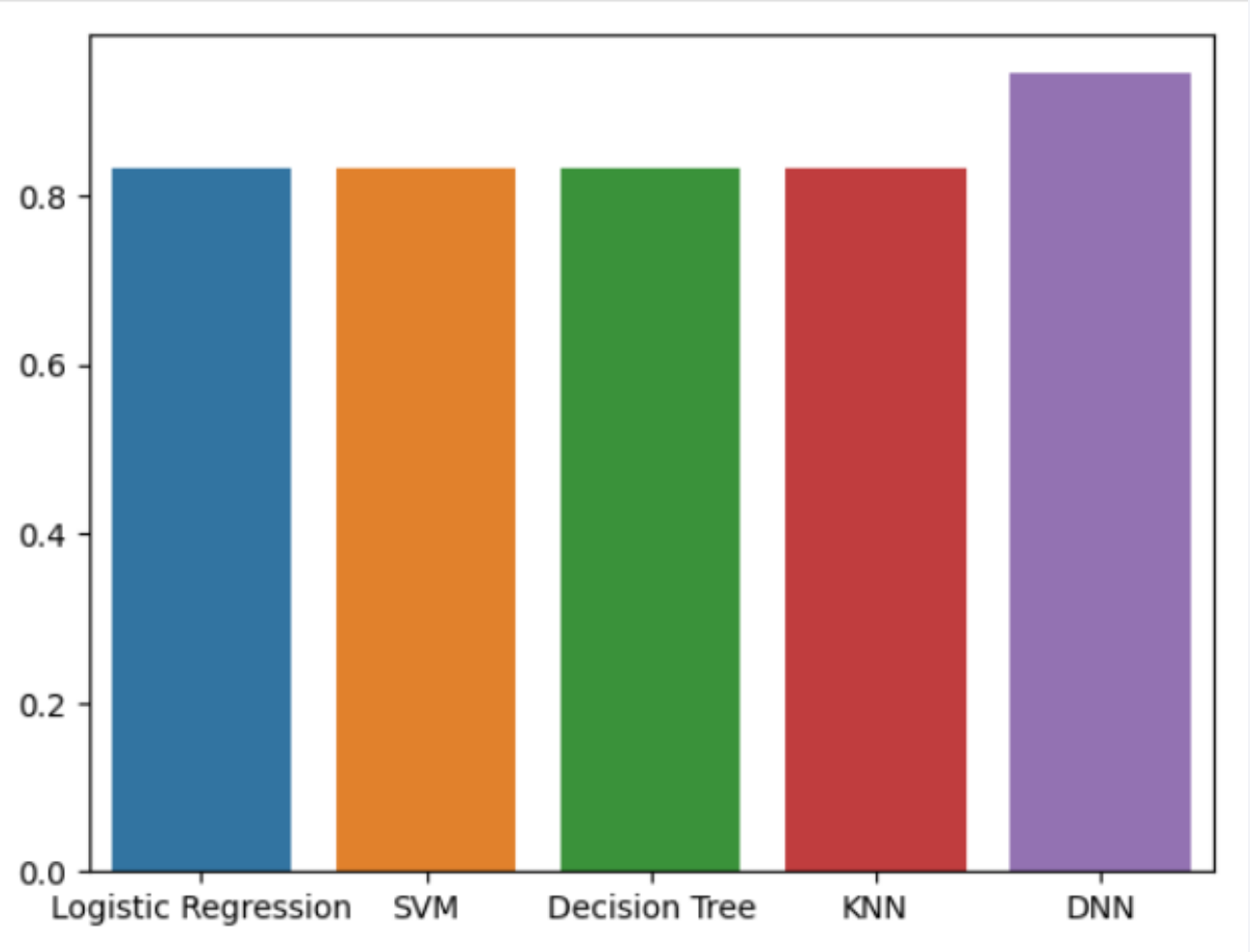
Section 5

Predictive Analysis (Classification)

Devank Banga

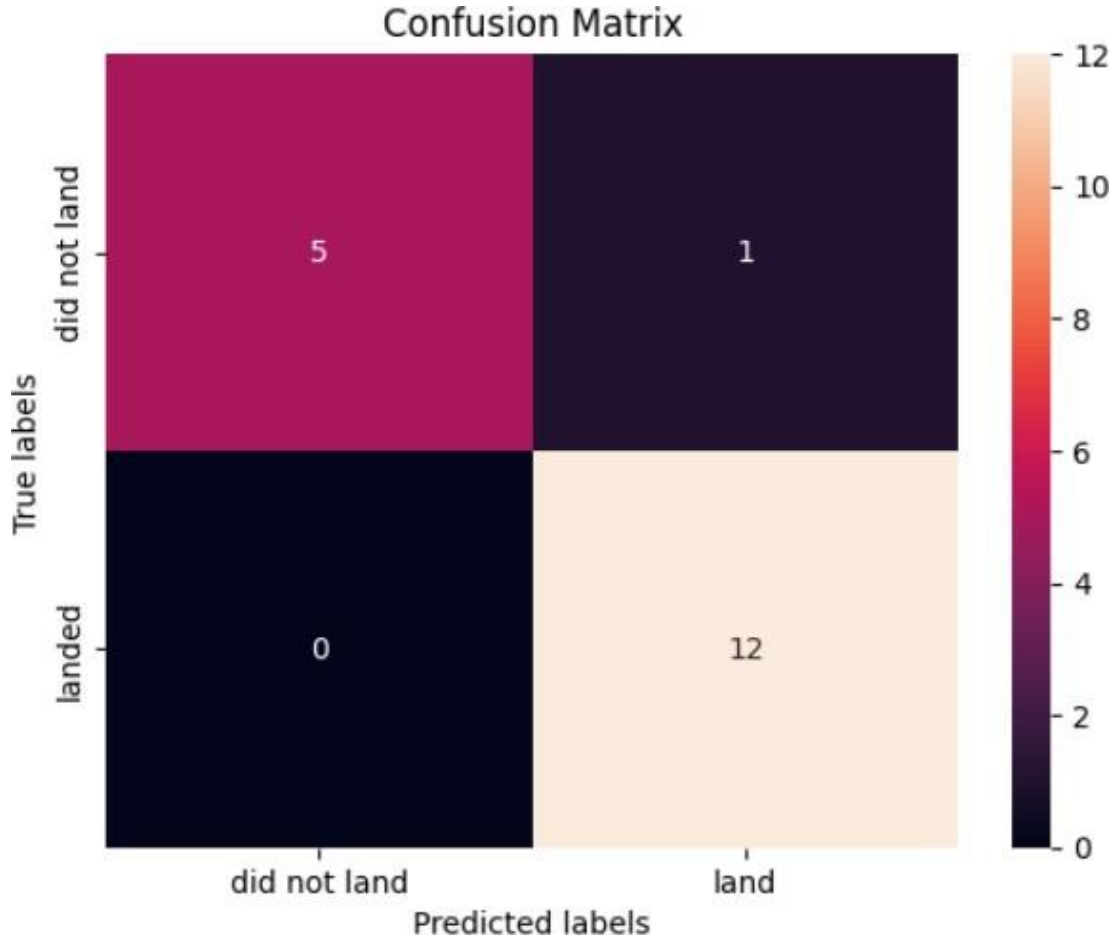
Classification Accuracy

- We compared the models on the basis of the Classification Accuracy calculated on the Test Dataset.
- During the project we used Logistic Regression, SVM, Decision Tree, and KNN for prediction and all of them reached a Test Accuracy of 0.8333.



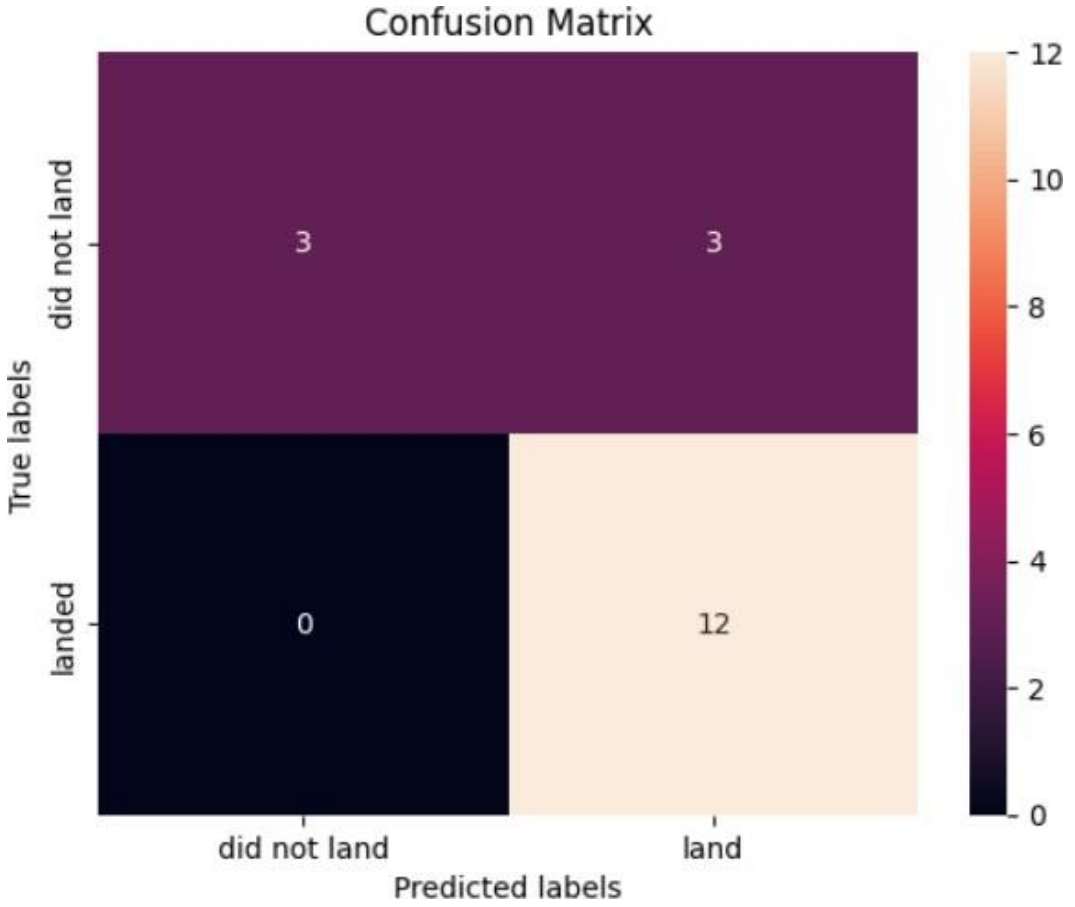
Confusion Matrix - DNN

- Our main objective was to predict the failures of SpaceX First Stage Landings so that our company can counter bid them, and DNN is able to recall at 83.3% with able to correctly classify 5 out of 6 total failed launches. And a 100% for successful launches.
- In contrast Machine Learning Methods has a recall of 50% with only 3 correct classifications out of total 6. We will analyze them too.



Confusion Matrix – LR/SVM/KNN/Decision Trees

- All the machine learning methods result in same Confusion matrix. Each one of them classifies all12 successful landings correctly but they classified 50% of the failed launches in correctly.
- More data can help us here. But till then DNN is more precise model even if we lose interpretability with it.



Conclusions

- SpaceX has an increasing trend in successful First Stage Landings from year 2013 with its earlier landings failing the most, those too mainly from CCA based Launch Sites. Newer launches are across different launch sites and with varying Payload sizes for different Orbit Types and boast a much higher success rate.
- We were able to produce a DNN model with 94.44% Test Accuracy, with it we are able to predict if the First Stage Landing will be Successful or Fail and make the counter bid as required.
- While DNN is more precise, stakeholders may prefer simpler model, especially Logistic Regression which has a Test Accuracy of 83.33%, same as other machine Learning Methods but with clear relationships between Predictors and Response.
- We suggest DNN for More Precise prediction and Logistic Regression for better interpretability.

Appendix

- Wikipedia static: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Github SpaceX project: https://github.com/ukiyoeh/IBM-Data-Capstone

Thank you!