IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

MUHAMMAD ABDULLAH ABRAR
30TH JUNE, 2024

Outline

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

Executive Summary

SUMMARY OF METHODOLOGIES

- **Collect:** First process is of collect data form different sources like websites or Wikipedia.
- **Wrangling:** Cleaning and extracting valuable data from the collection.
- **Exploring:** Using vast Exploratory Data Analysis techniques and achieve your questions using SQL and Python.
- **Insights:** Getting useful insights form exploring and understanding the dataset.
- **Prediction:** Using various Machine Learning algorithms to predict outcomes.

RESULTS

- **Insights:** Using exploratory data analysis to get useful insights for data regarding rockets, landing and space batches.
- **Predictive Learning:** From the machine learning algorithm you can change variables to get more useful insights.

Introduction

PROJECT BACKGROUND

- Predicting Falcon 9 Landing Launches
- Cost Effectiveness of Launches
- Cost of First Stages Launches
- Prediction of First Stages Launches
- Using Lesser Beget than Other Launches

QUESTIONS TO ASK

- Catching the most Recurrent Advancement.
- Will we able to predict Falcon 9 launches?
- What should be used for cost effectiveness?
- Will this be successful based on probability?

Section 1

Methodology

Methodology

EXECUTIVE SUMMARY

- DATA COLLECTION METHODOLOGY:
    - Web Scraping
- DATA WRANGLING
    - Standardizing Data
    - Normalizing Data
- EXPLORATORY DATA ANALYSIS USING VISUALIZATION & SQL
- INTERACTIVE VISUALS USING FOLIUM & PLOTLY DASH
- PREDICTIVE ANALYSIS USING CLASSIFICATIONS MODELS
    - Tuning Model
    - Selecting Best Model & Parameters

Data Collection

DATA COLLECTION PROCESS

- **Data Collection API**
  - Using API based Scraping
  - Python Libraries to integrate into rows and columns
  - Getting ready for prediction
  - **Source:** `https://api.spacexdata.com/v4/rockets/`
- **Data Collection Web-Scraping**
  - Wikipedia Data Collection
  - Method of Web Scraping
  - Extract Falcon 9 launches
  - **Source:** `https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922`

Data Collection – SpaceX API

DATA COLLECTION API

- Requesting Data from Source
- Encoding to JSON
- Saving source Response
- JSON to CSV
- Storing Variable to Columns
- Using Dictionary to Place Data Frame
- Data Wrangling
- Dealing with Missing Values

REQUEST → JSON →
RESPONSE → CSV →
VARIABLES NAMES →
DATA FRAME → DATA WRANGLING →
MISSING VALUES →
DATA STORING

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection - Scraping

DATA COLLECTION WEB SCRAPING

- Extracting Falcon 9 Launches
- Sourcing form Wikipedia
- Requesting Access
- Extracting all variables
- Phrasing HTML Tables
- Creating Data Frame
- Exporting to CSV

Data Wrangling

DATA WRANGLING PROCESS

- Loading Dataset
- Basic Data Analysis
- Understanding Dataset
- Checking Missing Values
- Checking Null Values
- Number of Launches on Each Site
- Checking Landing Outcomes
- Removing Unnecessary Data
- Removing or Dropping Missing Or Null Values
- Clean Data
- Exporting Dataset

EDA with Data Visualization

- Exploring And Energizing Data
- Finding Relation Ship Between Different Variables
  - Payload Mass to Flight Number
  - Launch Site to Flight Number
  - Orbit to Class
- Making Dummies
- Feature Engineering
- Outcomes
  - Predictable Data
  - Understandable Data

EDA with SQL

EXPLORATORY DATA ANALYSIS WITH SQL

- Watching Unique Values
- Average Payload Mass for Rockets
- Success Rate
- Booster With Success Rate
- Relationship Between Variables
- Mission Outcomes
- Payload Masses for Rockets
- Failure on Factors

Build an Interactive Map with Folium

FOLIUM MAPS

- Watching for Launch Spaces
- Encircling Major Rockets Launches
- Falcon 9 Success Launches
- Launches Sites on Railways
- Coastline Factors on Rockets
- Cities to Site Distances
- Success Launches Space

Build a Dashboard with Polly Dash

PLOTLY DASHBOARD

- Launch Site Requirement
  - Dropdown for specific launch sites
- Pie Chart
  - Showing Successful vs Unsuccessful Launches
- Payload Mass Slider
  - Adjust According to Self
- Mass and Success Booster Version
  - How many Booster Version are Successful with Different Masses

Predictive Analysis (Classification)

CLASSIFICATION ANALYSIS

- Grouping Data into Training Testing & Splitting
- Standardizing Data
- Using Different Classification Models
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbor
- Selecting Models
  - Parameters
  - Grid Search CV
- Model Accuracy
- Confusion Matrix

Results

EXPLORATORY DATA ANALYSIS

- Launches Improved Over Period of Time
- KSC LA-39A has Highest Launch Success Rate
- All Orbits have very high Success Rate

Data Visualization

- Most Launches Near Equator or Coastline
- Launch Failure can Damage Vast Majority of Civilization

Predictive Analysis

- Decision Tree is most Accurate Model
- Model can not be 100% Accurate

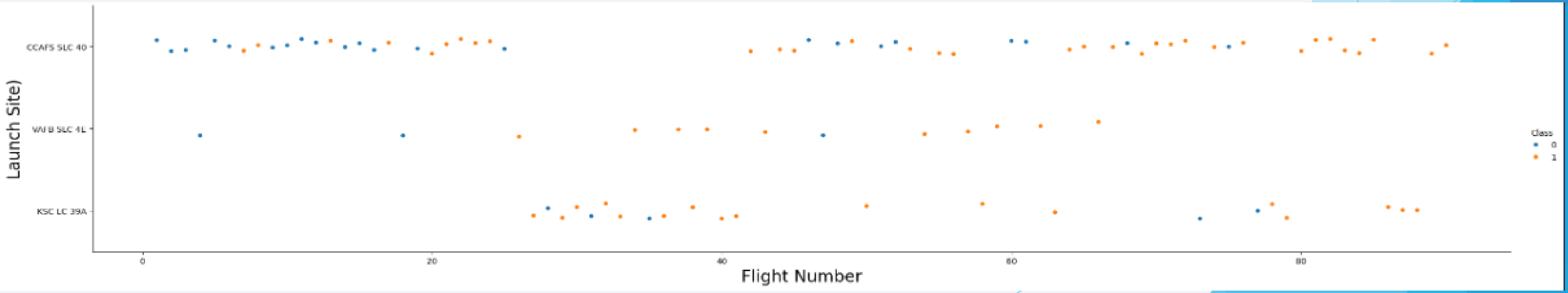
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

EXPLORATORY DATA ANALYSIS

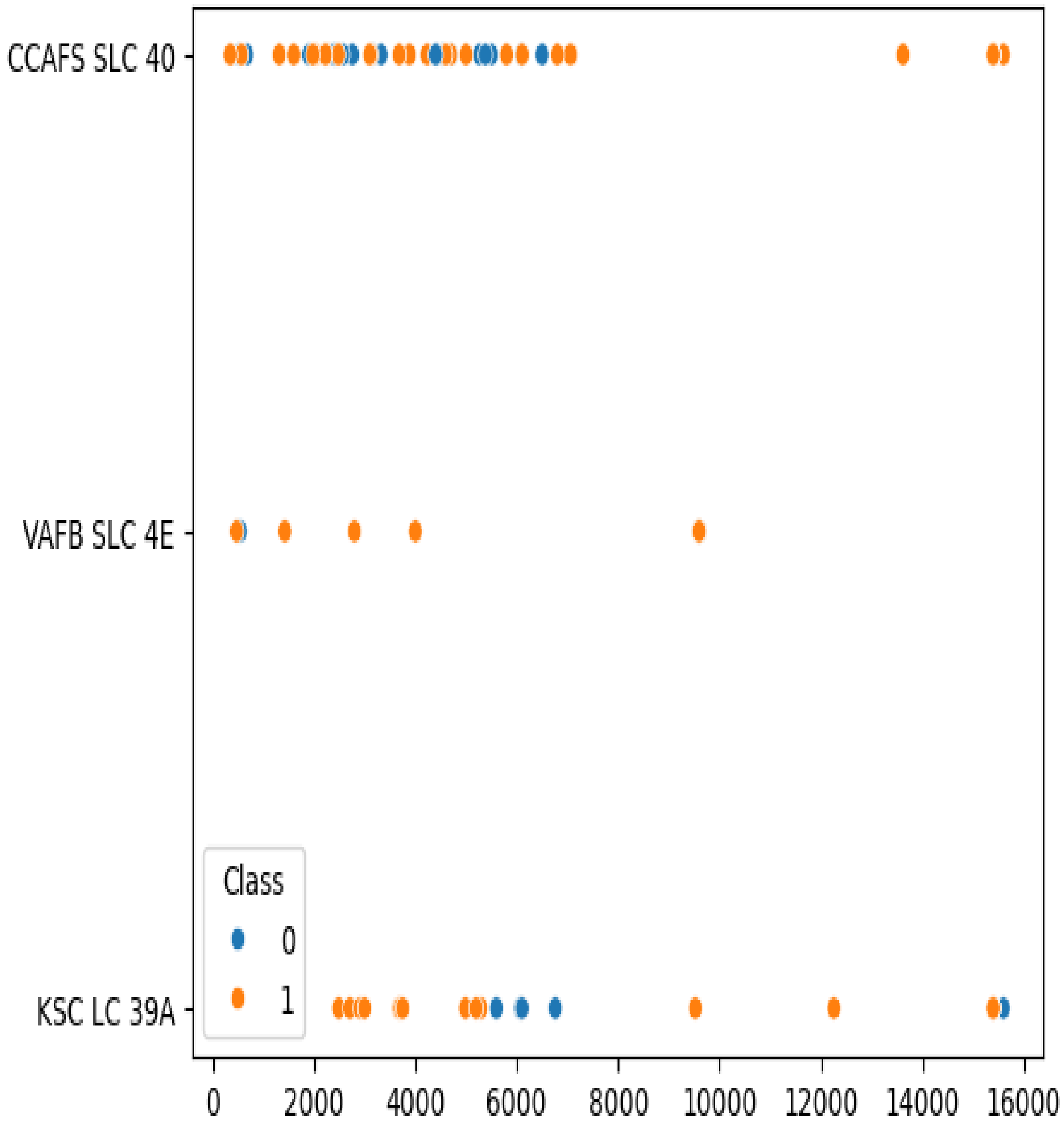
- Flight Number vs Launch Number (blue = fail) (orange = success)
- CCAFS SLC 40 has highest failure and success flights
- It is the most used flight
- Lowest used and High failure launch site is WAN SLC 4N



Payload vs. Launch Site

EXPLORATORY DATA ANALYSIS

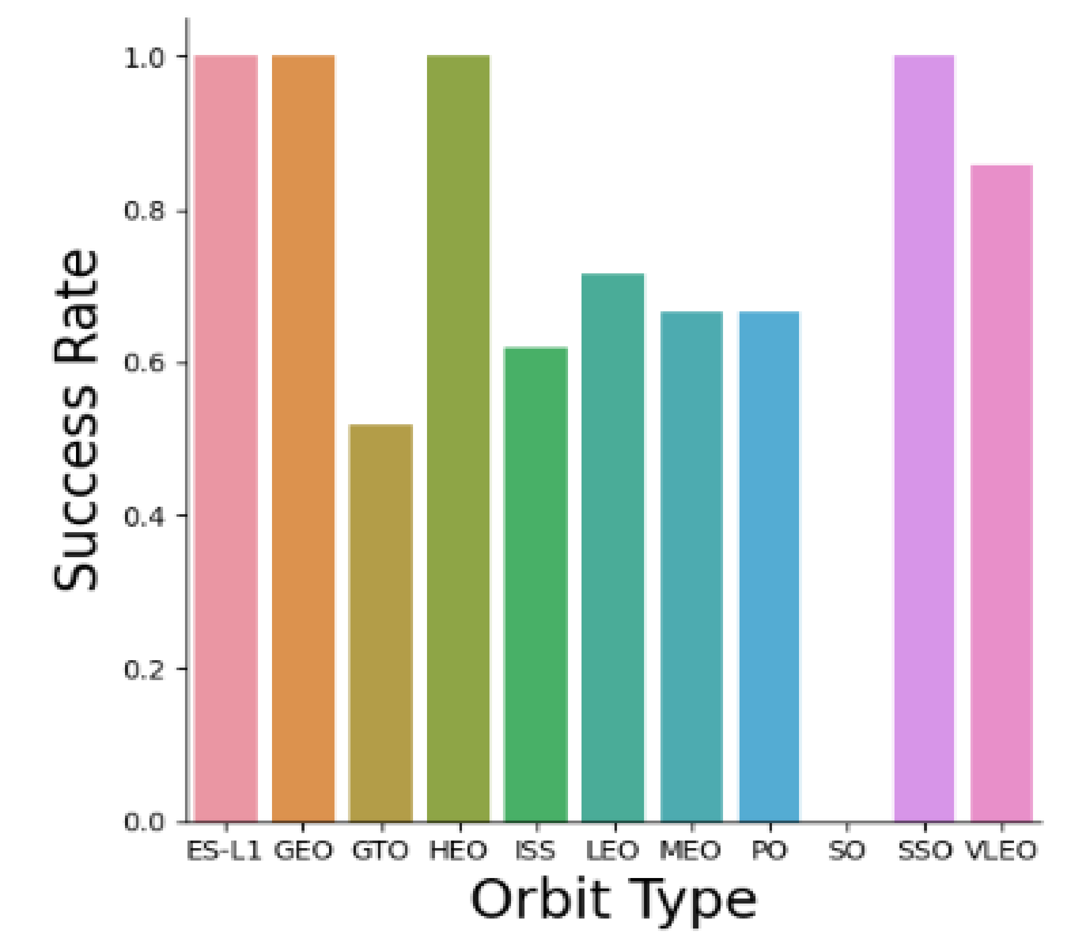
- Typically, Higher the Mass higher the success rate.
- CCAFS SLC 40 has 100% success rate for mass greater than 10000
- Most failures are between 5000 to 0 Mass
- VAFB SLC 4E has most failure rate from all launch sites



Success Rate vs. Orbit Type

EXPLORATORY DATA ANALYSIS

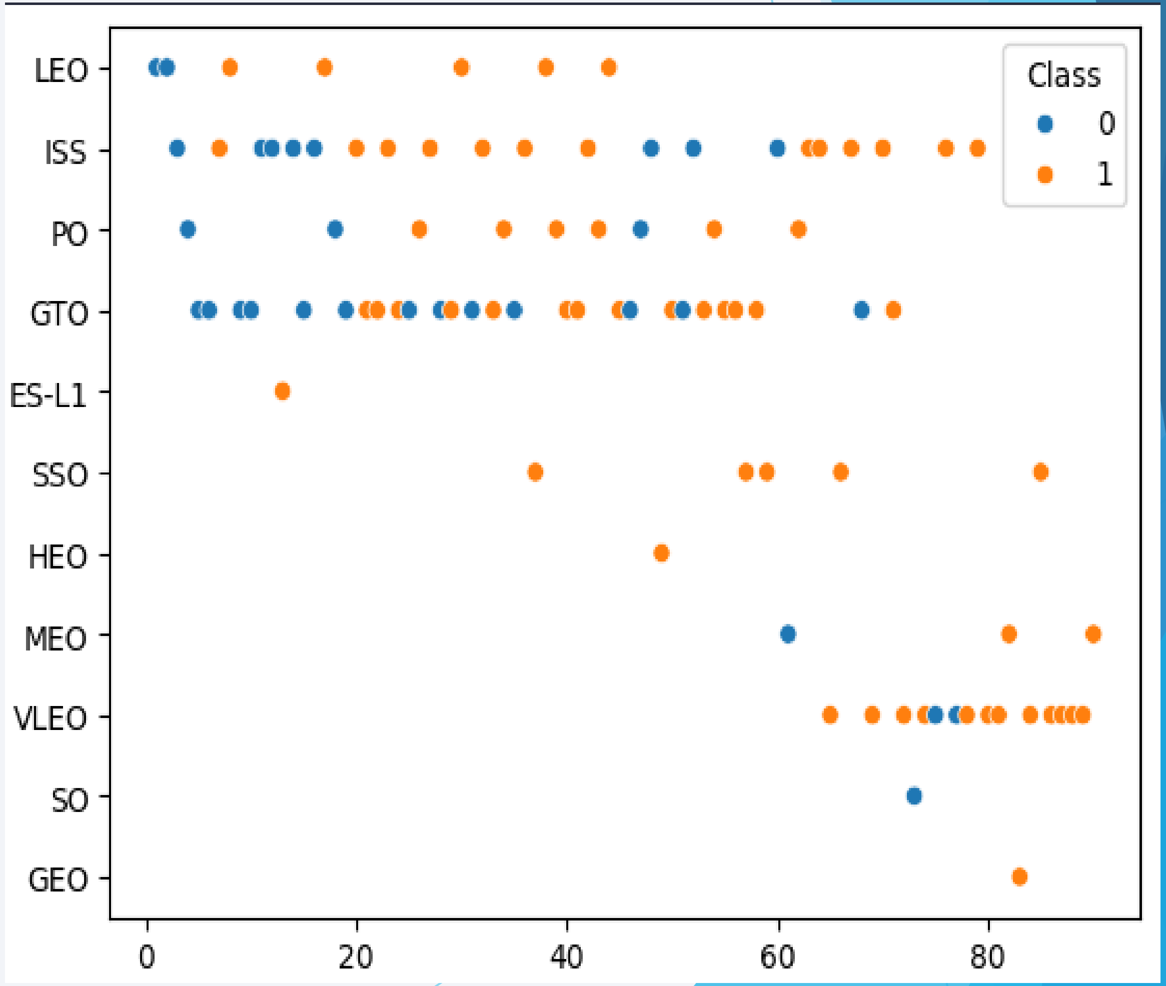
- 100% Success Rate: Orbits ES-L1, GEO, and HEO
- 60% - 90 % Success Rate: Orbits SS9, VELO, LEO, MEO, and PO
- 10% Higher Success Rate: GTO, and ISS



Flight Number vs. Orbit Type

EXPLORATORY DATA ANALYSIS

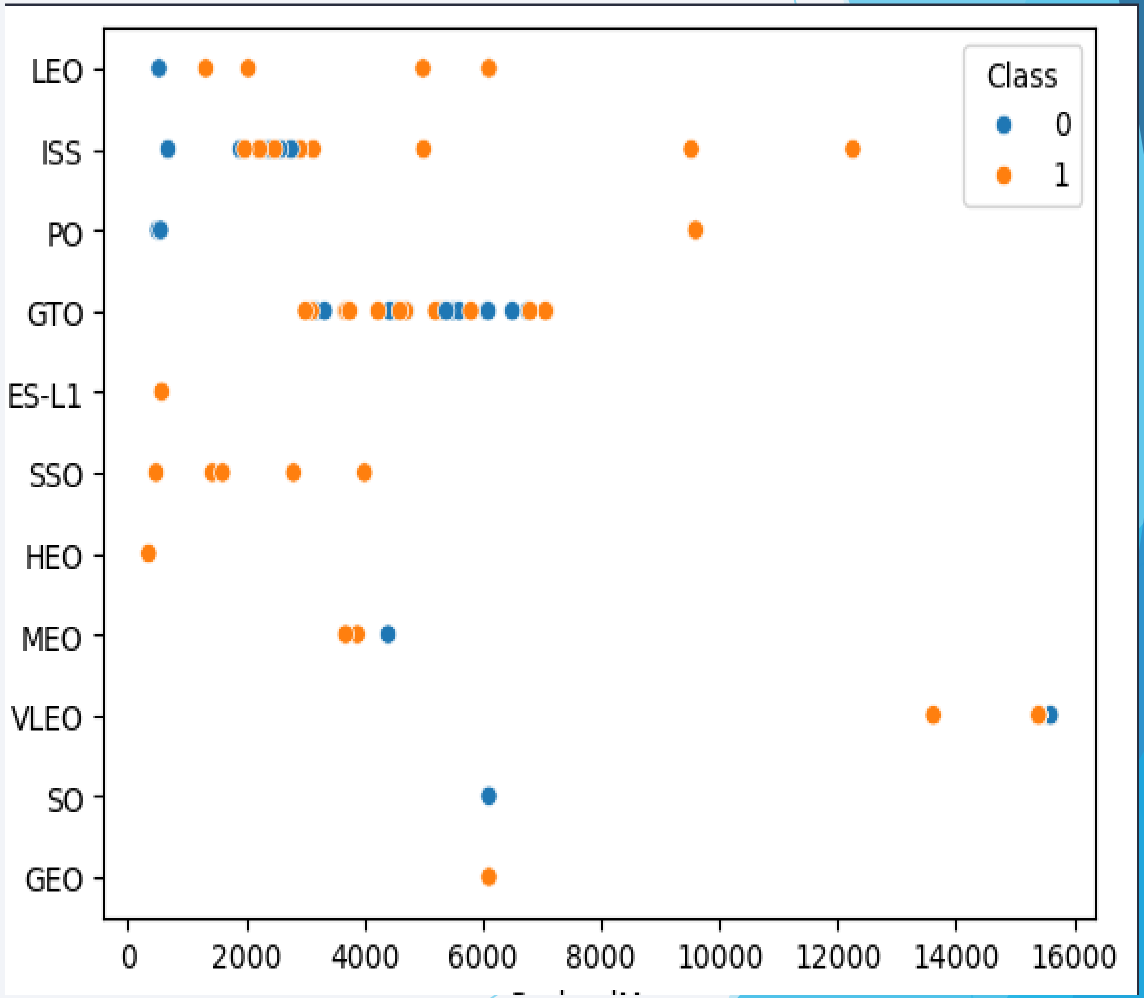
- Success Typical Increases Over Time
- Earlier Launches were a complete failure.
- VELO is producing most success launches recently
- GTO orbit is most disturbed orbit.



Payload vs. Orbit Type

EXPLORATORY DATA ANALYSIS

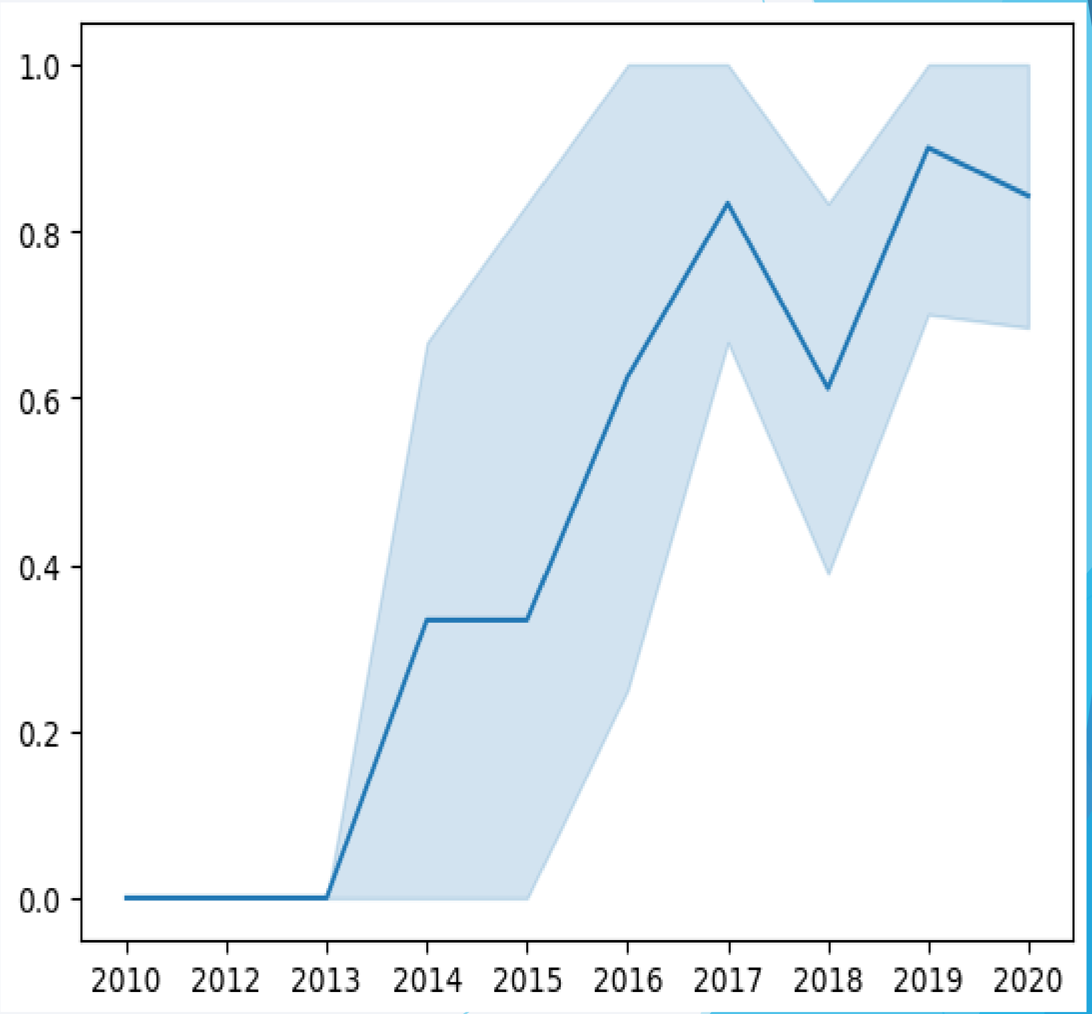
- Middle Range Masses have Higher success Rate than lower.
- GTO is most unpredictable with Payload Mass
- SSO has 100% success rate



Launch Success Yearly Trend

EXPLORATORY DATA ANALYSIS

- Over the time success is radically growing
- After 2013 success rate spiked upwards
- It decreases between years of COVID-19



All Launch Site Names

EXPLORATORY DATA ANALYSIS

Launch Sites

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

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

```
 * 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'

EXPLORATORY DATA ANALYSIS

Launch Sites With "CCA"

Task 2

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

```
%sql SELECT * FROM SPACEXTBL 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

EXPLORATORY DATA ANALYSIS

Total Payload Mass

- 45596

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

EXPLORATORY DATA ANALYSIS

Average Payload Mass by F9 v1.1

- 2928.4

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

EXPLORATORY DATA ANALYSIS

First Successful Ground Landing Date

- 2010-06-04

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date) FROM SPACEXTBL
WHERE MISSION_OUTCOME = 'Success';
```

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

MIN(Date)
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

EXPLORATORY DATA ANALYSIS

Successful Drone Ship Landing With Payload Between 4000 & 6000

Booster Versions

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

List the names of the boosters which have success

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
WHERE LANDING_OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

0.0s

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

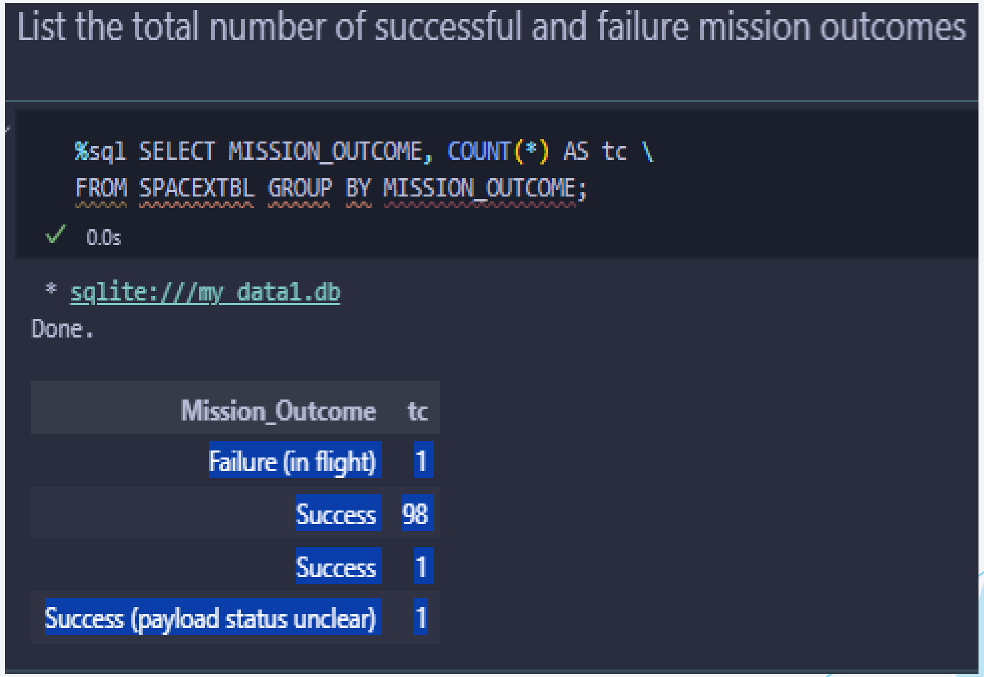
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

EXPLORATORY DATA ANALYSIS

Total Number of Successful and Failure Mission Outcomes

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


List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS tc \
FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

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

Boosters Carried Maximum Payload

EXPLORATORY DATA ANALYSIS

Boosters Carried Maximum Payload

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

List the names of the booster_versions which have ca

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ \
FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = \
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)LIMIT 5;
```

✓ 0.0s

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

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

2015 Launch Records

EXPLORATORY DATA ANALYSIS

2015 Launch Records

01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the m substr(Date,7,4)='2015' for year.

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Task 10

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

EXPLORATORY DATA ANALYSIS

Rand Landing Outcomes Between 2010-06-04 & 2017-03-20

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

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

Landing _Outcome	count_outcomes
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

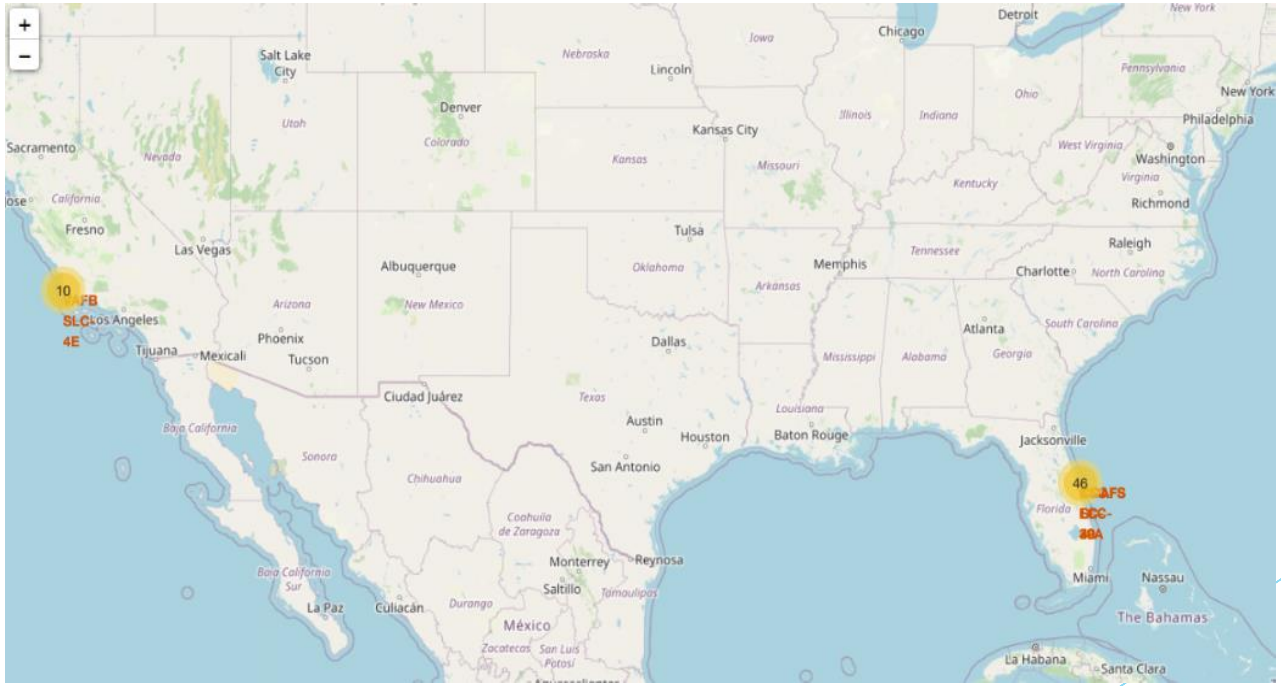
Section 3

Launch Sites Proximities Analysis

Launch Sites

LAUNCH SITES DETAILS

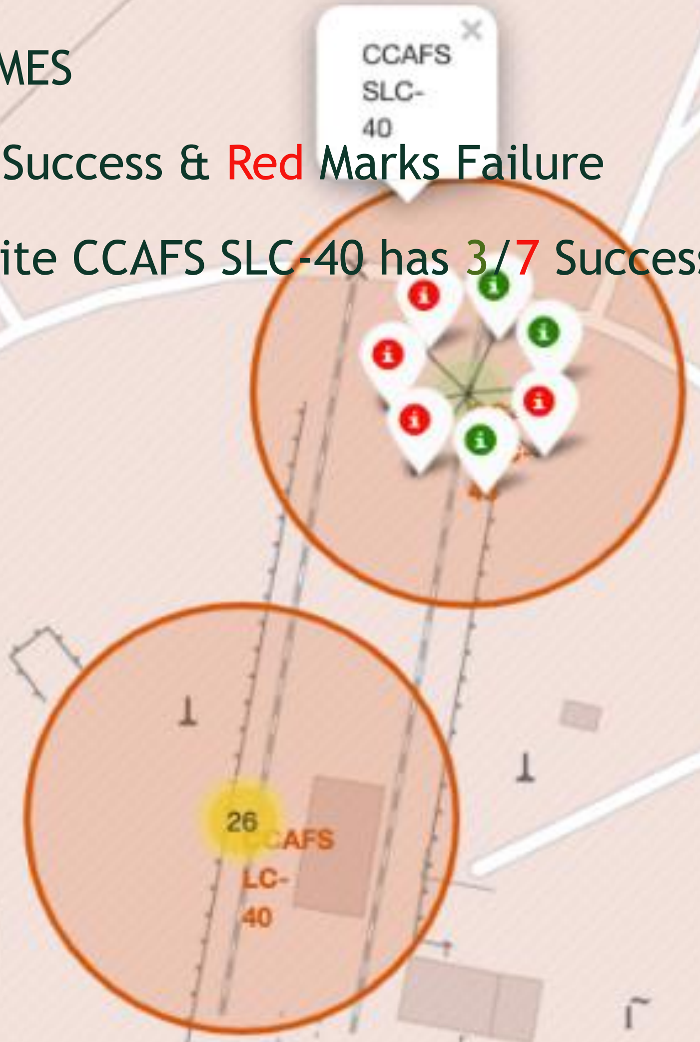
- It lies near equator to give them extra push and take lesser route to their specified orbit.
- They also lie beside coastal area to give them extra boost for their route.



Launch Outcomes

LAUNCH OUTCOMES

- Green Marks Success & Red Marks Failure
- This launch site CCAFS SLC-40 has 3/7 Success Launches. (42%)



Launch Site Distances

LAUNCH SITES DISTANCES

CCAFS SLC-40

- 1km from Coastline
- 21km from Nearest Railway
- 23km from Nearest City
- 26km from Nearest Highway.

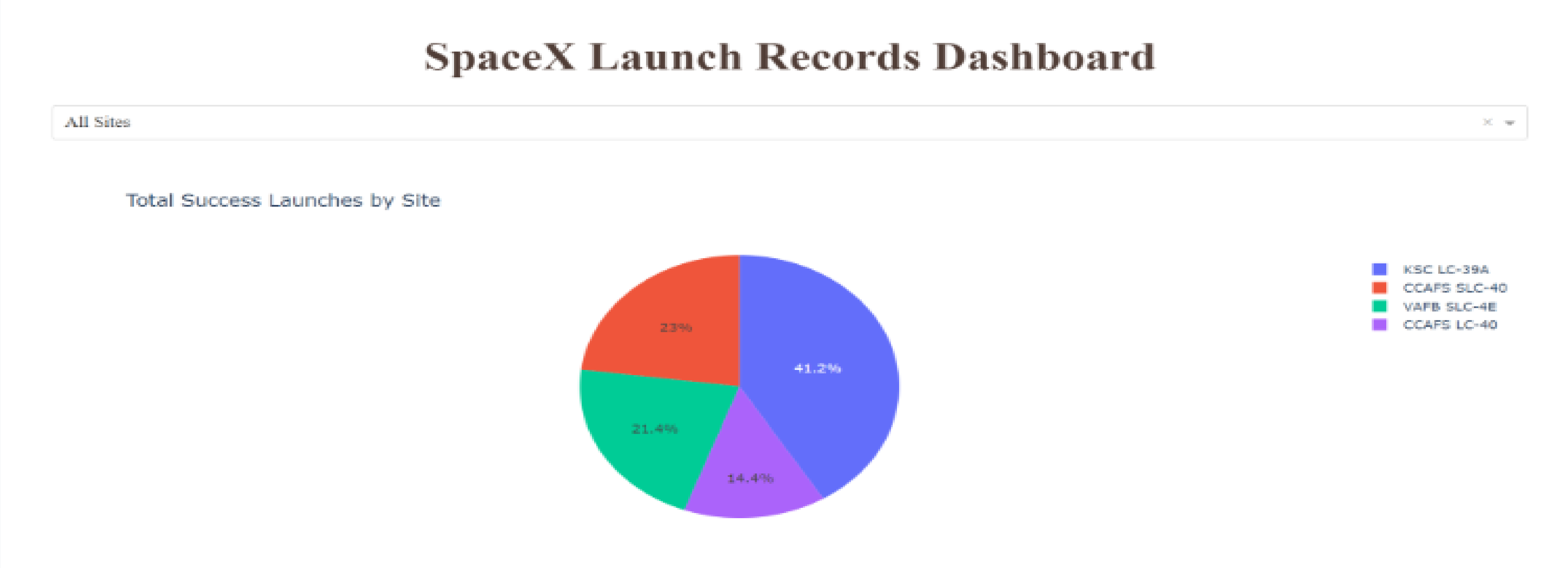
Section 4

Build a Dashboard with Plotly Dash

Pie For Launch Site Success

PIE CHART FOR LAUNCH SITES SUCCESS RATE

- Highest is for KSC LC-39A which is 42%

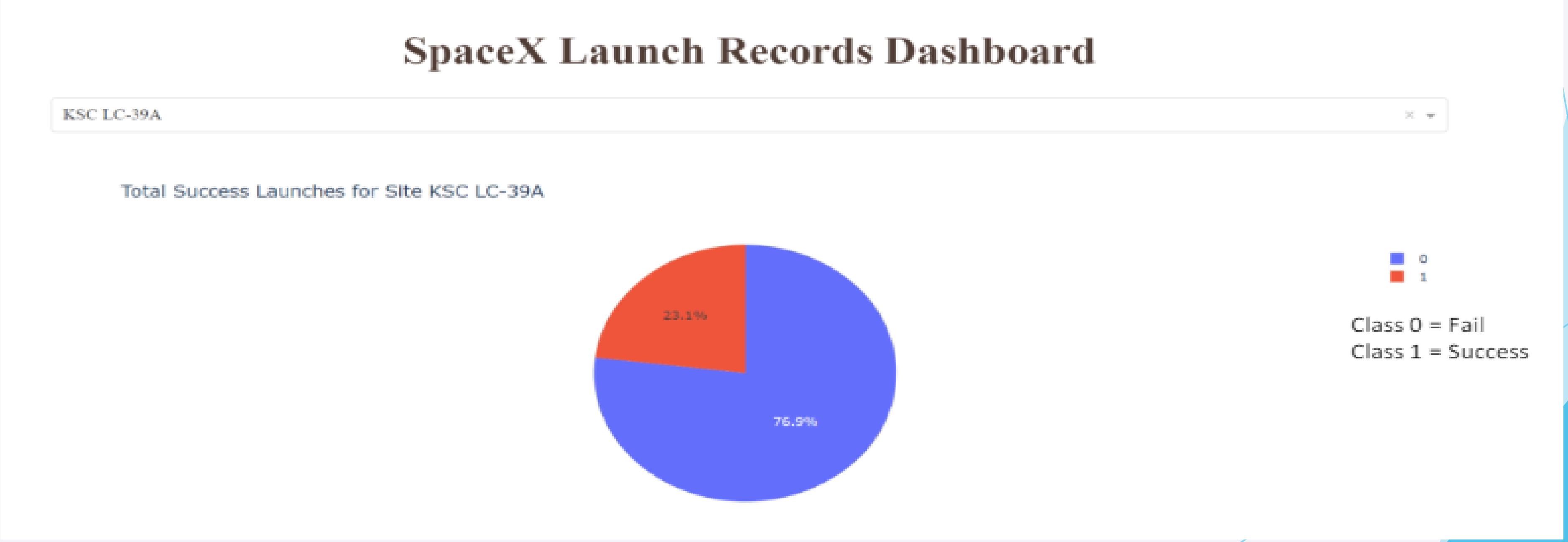


Launch Success

LAUNCH SITES SUCCESS RATE

- KSC LC-39A has highest launch success rate. (70%)

1. 3 out of 10 Launches Fail.



Payload Mass Success

PAYLOAD MASS SUCCESS RATE

- Mass between 2000 – 5000 has highest success rate among all.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

CLASSIFICATION ACCURACY

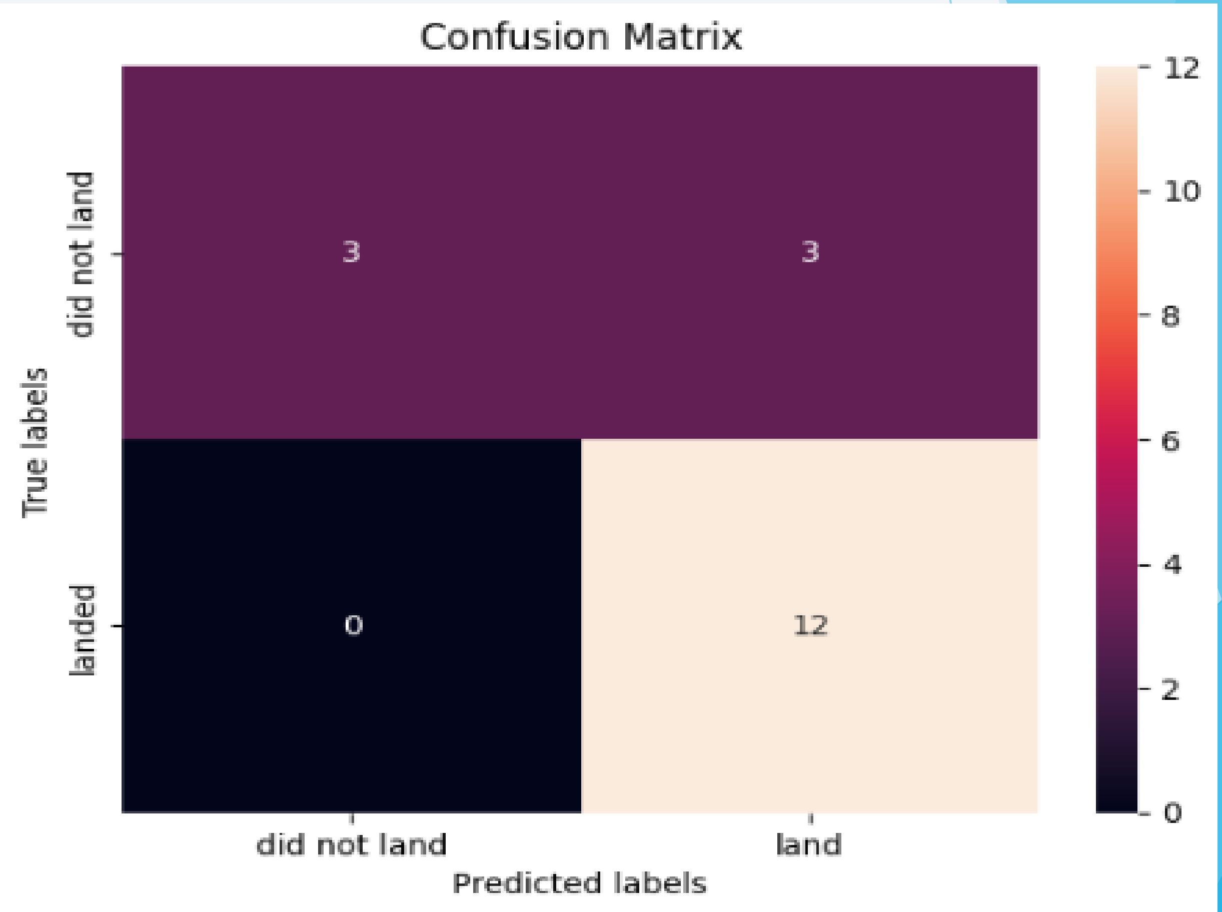
All model K-Nearest Neighbor, Support Vector Machine, Logistic Regression, & Decision Tree produced good results but Decision Tree stands out from the other due to too good model training with accuracy of 94%

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.923077	0.800000
F1_Score	0.888889	0.888889	0.960000	0.888889
Accuracy	0.833333	0.833333	0.944444	0.833333

Confusion Matrix

Best Performing Model is Decision Tree Classifier

With the accuracy of 94% it did not predicted false landed. It only produced 3 wrong outputs which predicted land but actual were not land



Conclusions

- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator:** Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast:** All the launch sites are close to the coast
- **Launch Success:** Increases over time
- **KSC LC-39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits:** ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass:** Across all launch sites, the higher the payload mass (kg), the higher the success rate

Appendix

Things to Consider

- Dataset: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set
- Feature Analysis / PCA: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy
- Boost: Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models

Thank you!