IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abrar Argya Adana
12/06/2023

Outline

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

Executive Summary

- **Summary of methodologies**
- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- **Summary of all results**
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

- **Project background and context**
- 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.
- **Problems you want to find answers**
- What's the correlation between success rate and other attributes?
- Which site and booster version has the best success rate?
- Which method perform best to predict if the Falcon 9 first stage will land successfully?

chicagotribune.com

Section 1

Methodology

convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

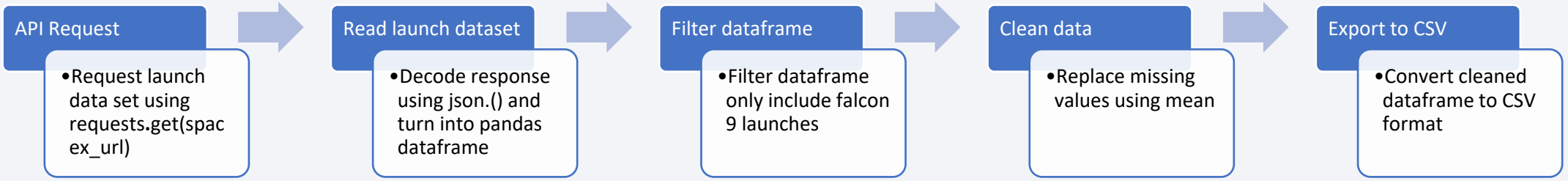
Methodology

Executive Summary

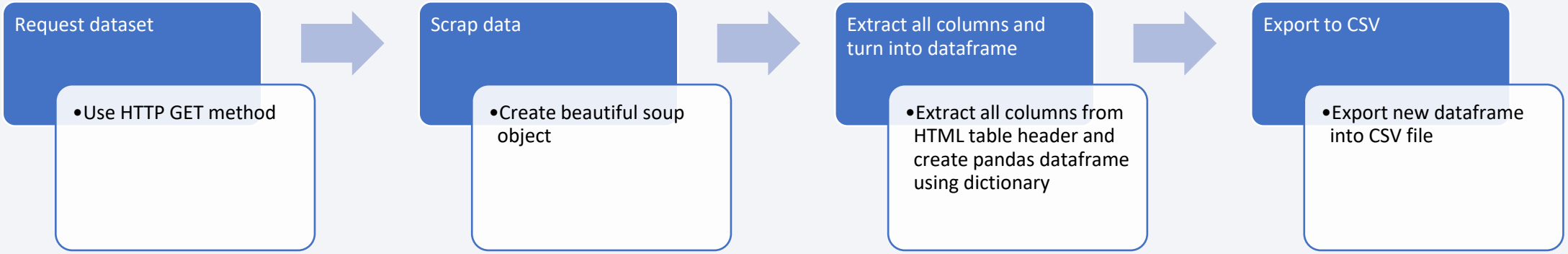
- Data collection methodology:
  - Request to SpaceX API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Convert landing outcomes to Boolean data
  - 1 means successful meanwhile 0 means unsuccesful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create column for the class, standardize the data, and split into training and test data
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the best method

Data Collection

- Datasets collected from SpaceX API and webscraping from Wikipedia
- Space X API

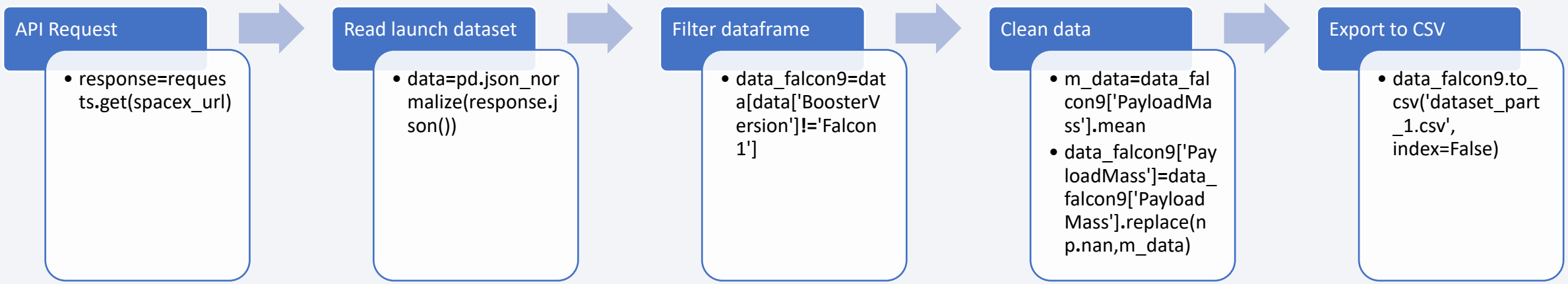


- Webscraping from Wikipedia



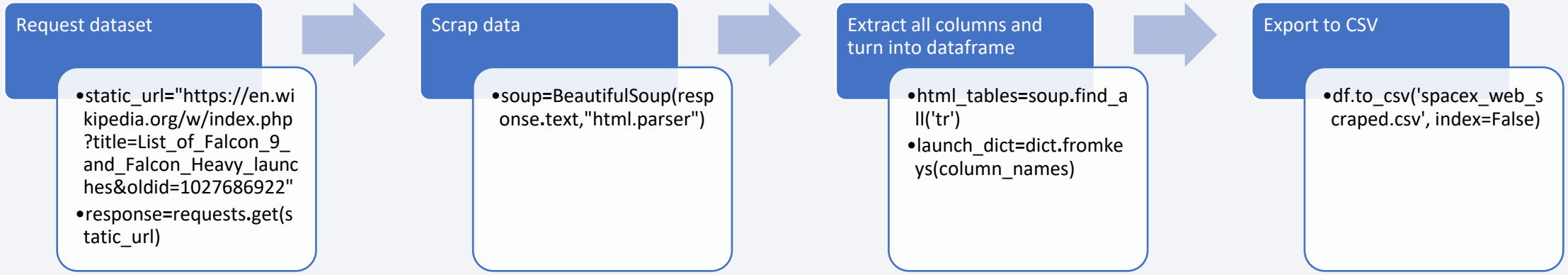
Data Collection – SpaceX API

- https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/data%20collection_1.ipynb



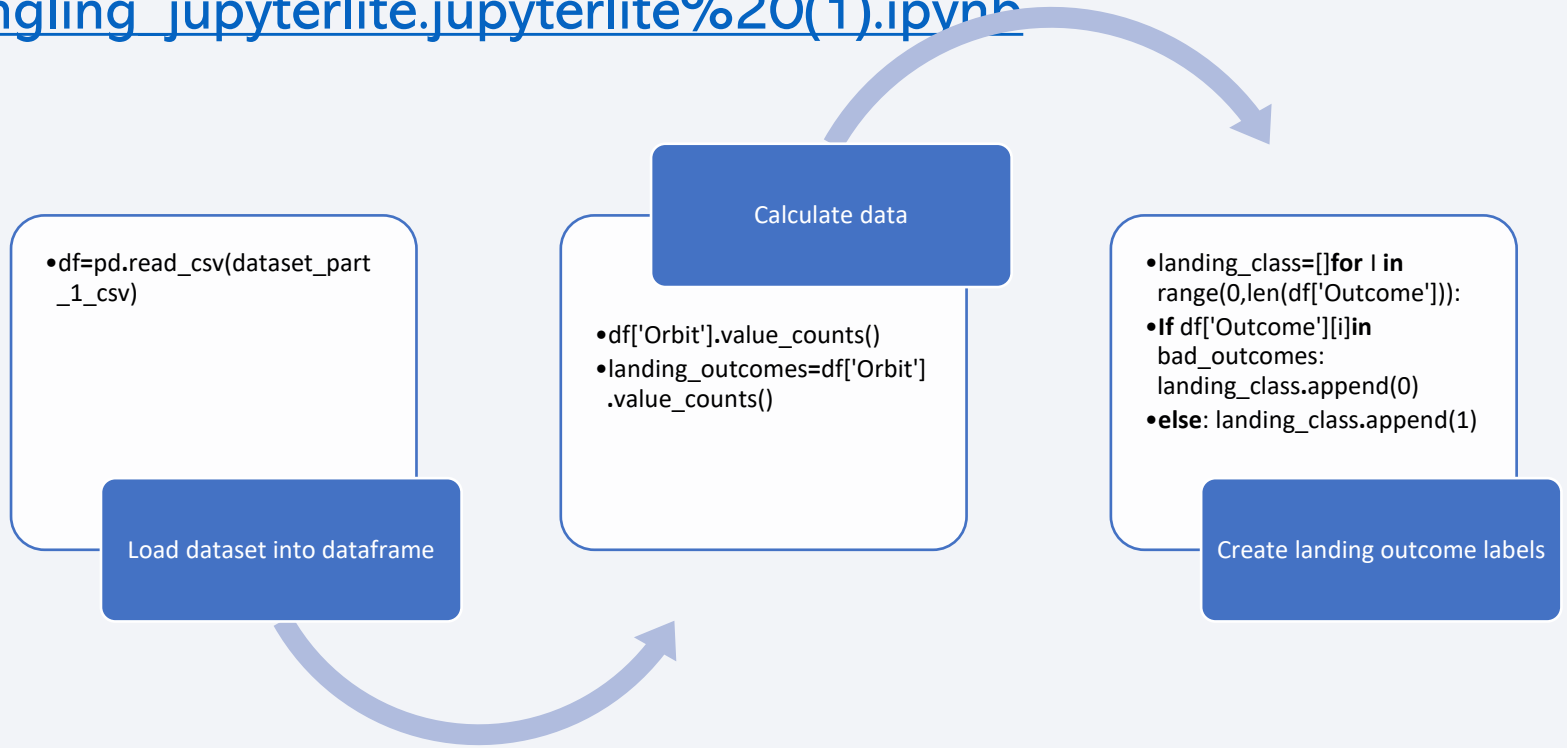
Data Collection - Scraping

- https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/Web%20scraping%20from%20Wikipedia.ipynb



Data Wrangling

- Convert landing outcomes to Boolean data (training labels), 1 means the booster landing successfully meanwhile 0 means unsuccesfully
- https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite%20(1).ipynb



EDA with Data Visualization

- **Scatter Plot**

To plot the relationship between two attributes

- Relationship between flight number and pay load mass
- Relationship between flight number and launch site
- Relationship between pay load mass and launch site
- Relationship between flight number and orbit
- Relationship between pay load mass and orbit

- **Bar Chart**

To Compare values between each data point

- Compare successful rate between each type of orbits

- **Line Chart**

To plot relationship between continuous data

- To plot average success rate in period of time (in this case, year)

- https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Connect to Db2 database
- Display all unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- 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. Use a subquery
- 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.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- Map objects that created :

1. folium.Circle dan folium.Marker

To make circle and labels to the launch sites

2. MarkerCluster()

To differentiate marker color as green if the launch success, and red if the launch unsuccessful

3. MousePosition()

Coordinate for a mouse over a point on the map.

4. folium.Polyline()

To draw polyline between launch site

- https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite%20(1).ipynb

Build a Dashboard with Plotly Dash

- 1. Added a launch site drop-down list to filter dashboard visual launch site
- 2. Added a pie chart to the dashboard to show total success by site
- 3. Added a Payload range to the dashboard to select payload ranges
- 4. Added a Scatter chart to see relationship between pay load mass and class
- https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/Dash.py

Predictive Analysis (Classification)

Load the dataframe

- data=pd.read_csv(text1)

Create Numpy array

- Create a NumPy array from the column in data
- Y=data['Class'].to_numpy()

Standardize the data

- X=preprocessing.StandardScaler().fit(X).transform(X)

Train/test/split data

- **From** sklearn.model_selection **import** train_test_split
- X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2,random_state=2)

Create Models

- parameters={"C":[0.01,0.1,1],'penalty':['l2'],'solver':['lbfgs']}*# l1 lasso l2 ridge*
- lr=LogisticRegression()
- logreg_cv=GridSearchCV(lr,parameters,cv=10)
- logreg_cv.fit(X_train,Y_train)

https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Predictive Analysis (Classification)

Test the accuracy

- print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
- print("accuracy :",logreg_cv.best_score_)

Calculate the accuracy on the test data

- yhat=logreg_cv.predict(X_test)
- plot_confusion_matrix(Y_test,yhat)

Repeat from step 4 to 7

- Using Decision tree, SVM, and KNN

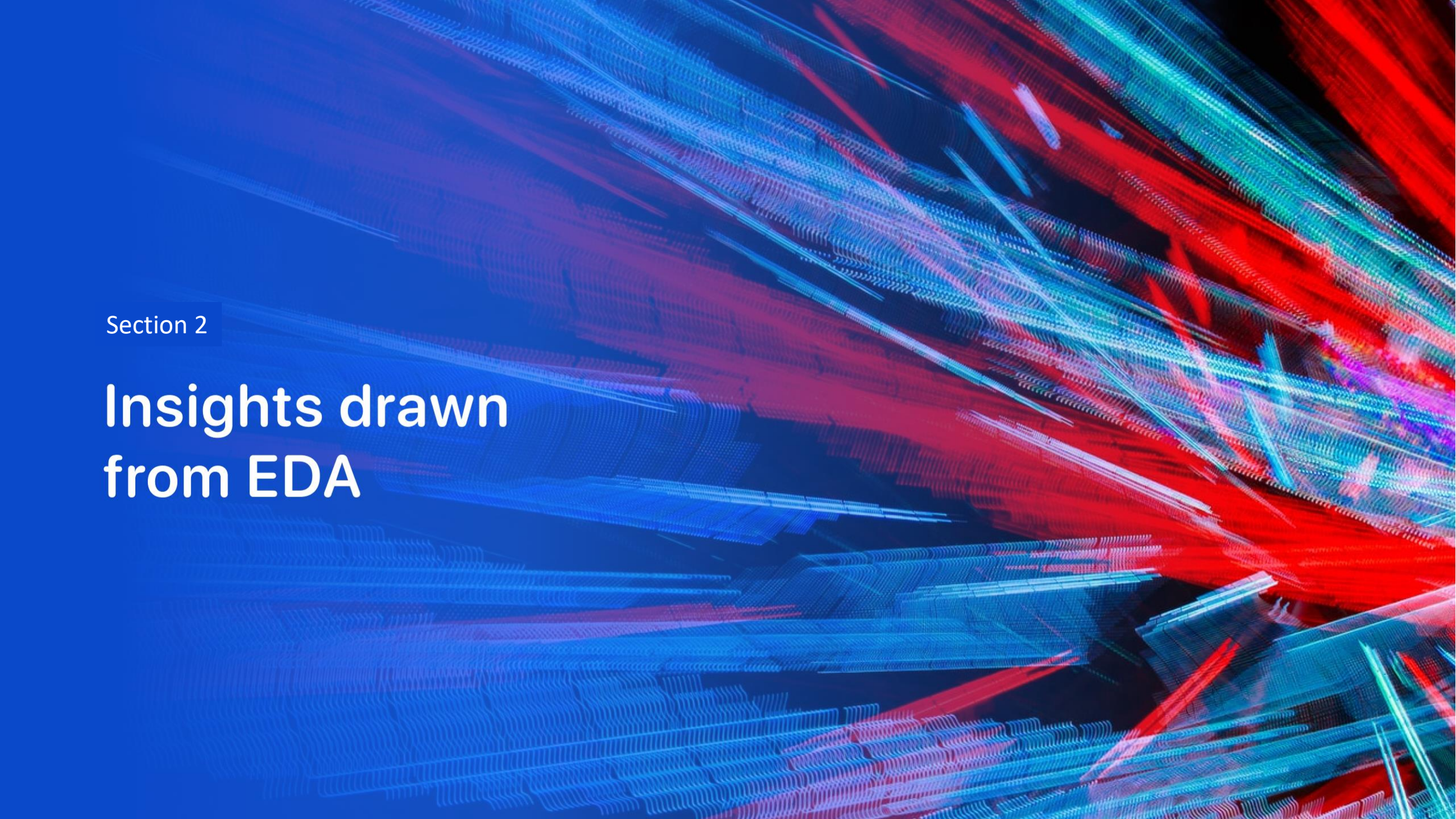
Select the best model

- Model_Performance_df=pd.DataFrame({'Algo Type':['Logistic Regression','SVM','Decision Tree','KNN'],'Accuracy Score':[logreg_cv.best_score_,svm_cv.best_score_,tree_cv.best_score_,knn_cv.best_score_],'Test Data Accuracy Score':[logreg_cv.score(X_test,Y_test),svm_cv.score(X_test,Y_test),tree_cv.score(X_test,Y_test),knn_cv.score(X_test,Y_test)]})

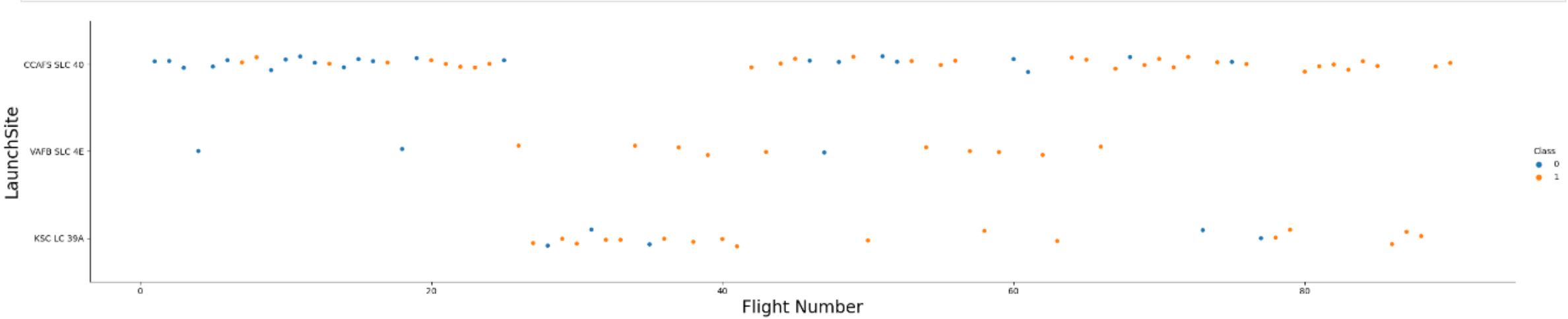
https://github.com/abrarargya/IBM-Data-Science-Capstone-Project/blob/main/IBM-DS0321EN - SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

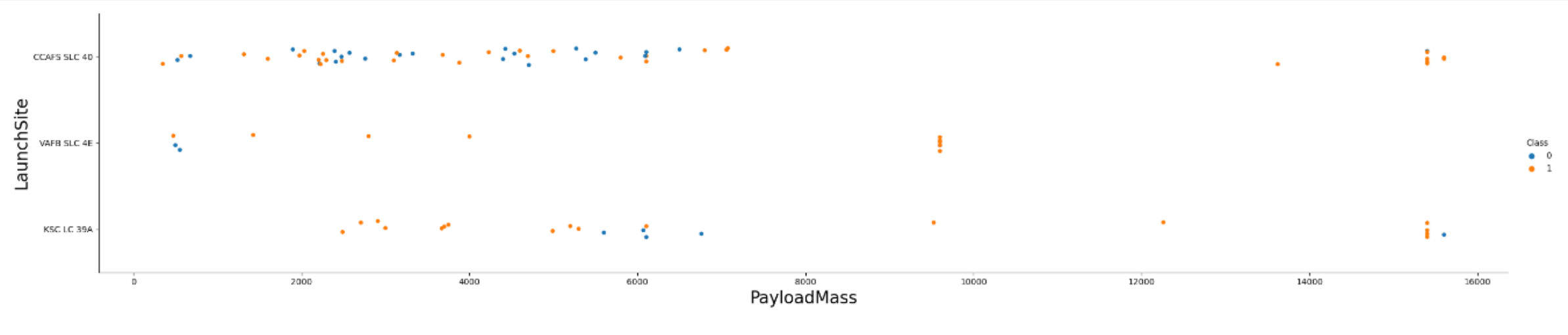

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



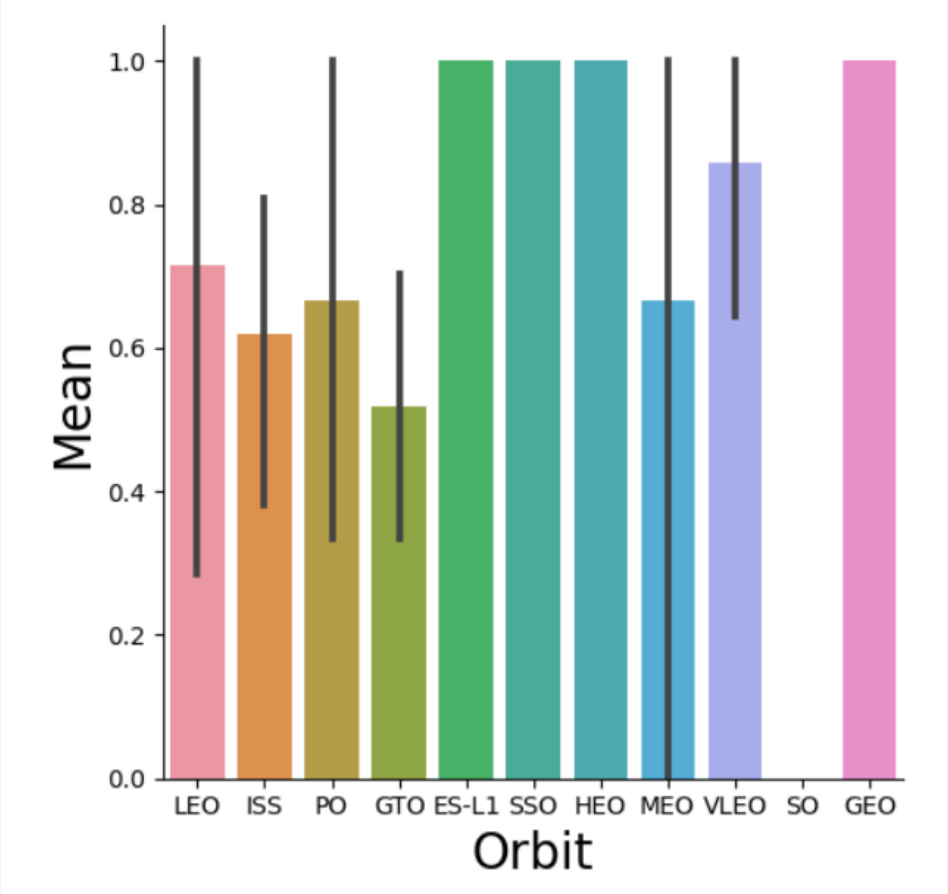
Successful launch is increasing as the flight number increasing

Payload vs. Launch Site



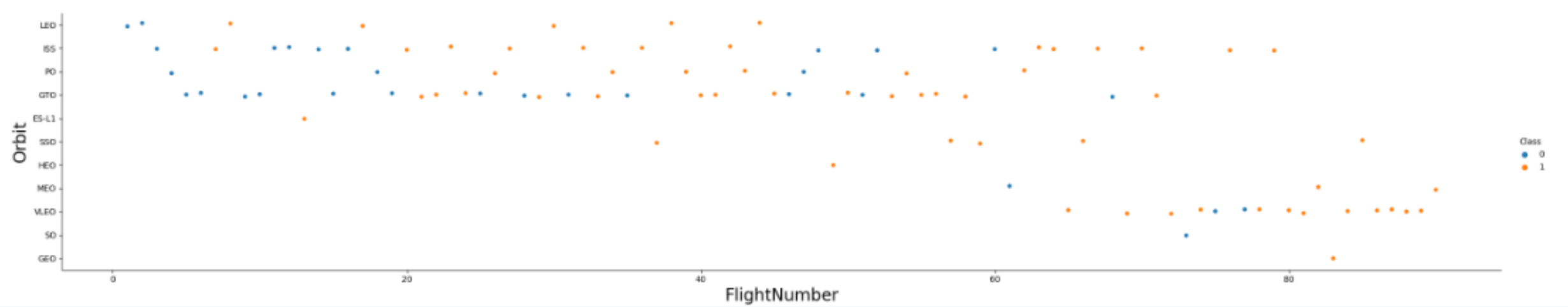
There’s no correlation between pay load mass and launch site

Success Rate vs. Orbit Type



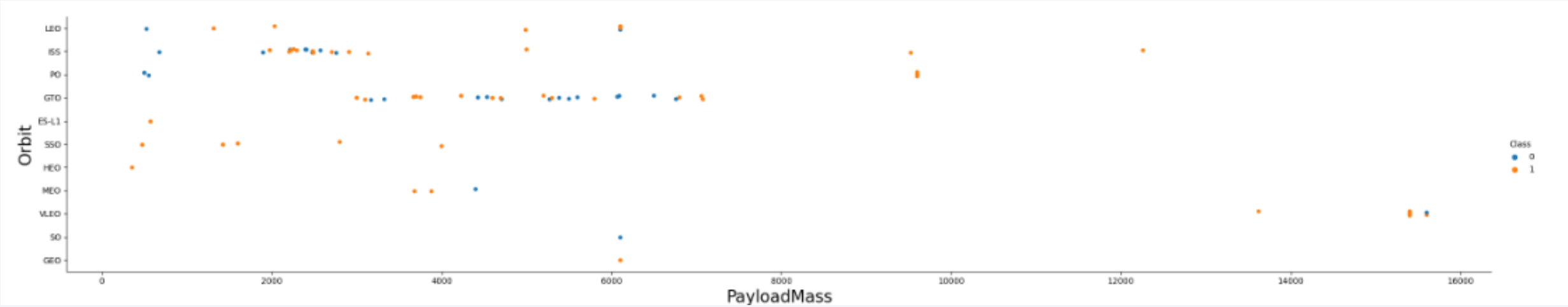
ES-L1,SSO, HEO, and GEO has the most succest rate

Flight Number vs. Orbit Type



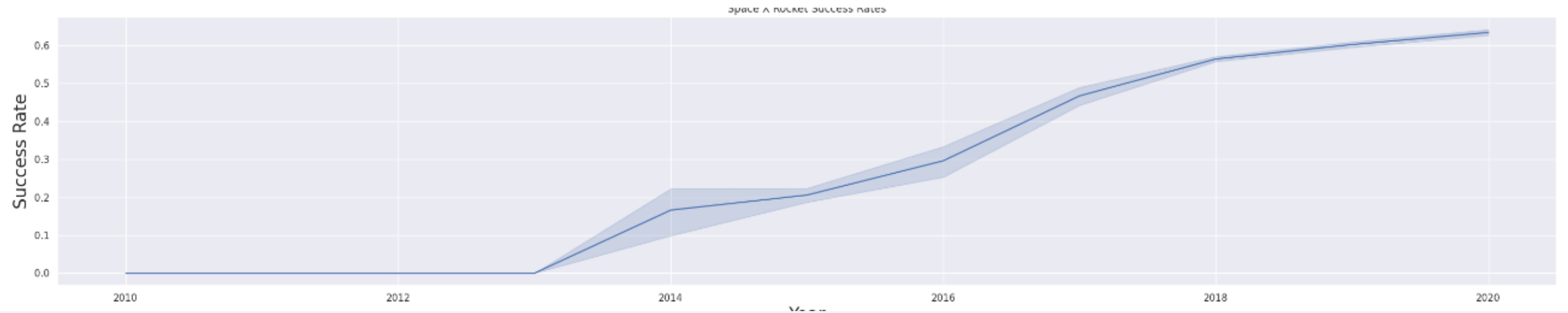
The more flight number are, the more frequent the successful landing is for orbit LEO, SSO, and VLEO

Payload vs. Orbit Type



The more pay load mass are, the more frequent the successful landing is for orbit LEO, ISS, and PO

Launch Success Yearly Trend



The success rate is generally increasing

All Launch Site Names

```
In [8]: %sql select distinct Launch_Site from SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[8]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

There are four launch site in SpaceX dataset

Launch Site Names Begin with 'CCA'

```
%%sql Select * from SPACEXTBL
where Launch_Site like 'CCA%'
limit 5
```

]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Using query like '%CCA%' to find string the same as CCA

Total Payload Mass

```
In [10]: %sql select sum(PAYLOAD_MASS__KG_) as pay_load_mass_kg from SPACEXTBL Where Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[10]:
```

pay_load_mass_kg
45596.0

The query is return total pay load mass from table SPACEXTBL and the customer is NASA(CRS); the total payload mass is 45596 kg

Average Payload Mass by F9 v1.1

```
In [11]: %sql select avg(PAYLOAD_MASS__KG_) as payloadmasskg from SPACEXTBL Where Booster_Version = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
Out[11]: payloadmasskg
         2928.4
```

The query is return average of pay load mass from table SPACEXTBL and the booster version is F9 v1.1; the result is 2928,4 kg

First Successful Ground Landing Date

```
12]: %sql select min(DATE) as mindate from SPACEXTBL where Landing_Outcome like '%ground pad%';
 * sqlite:///my_data1.db
Done.
12]:
```

mindate
01/08/2018

The query is return the date of first successful ground landing date from table SPACEXTBL and the landing outcome is ground pad like string; the date is 1st August 2018

Successful Drone Ship Landing with Payload between 4000 and 6000

```
]: %sql select Booster_Version from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 a
 * sqlite:///my_data1.db
Done.
```

]:

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

- There's five booster version that successfully landing with payload between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

```
]: %sql select Mission_Outcome , count(*) as missionoutcomes from SPACEXTBL GROUP BY Mission_Outcome
 * sqlite:///my_data1.db
Done.
```

]:

Mission_Outcome	missionoutcomes
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

There's 100 success mission, 1 failure, and 1 success with payload status unclear

Boosters Carried Maximum Payload

```
15]: %sql select BOOSTER_VERSION  from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

15]:

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

There are 12 booster that carried maximum payload

2015 Launch Records

```
42]: %%sql Select Landing_Outcome, Booster_Version, Launch_Site, (Select substring(Date,4,2) from SPACEXTBL) as Month from SPACEXTBL
     where Landing_Outcome like 'Failure%' and (Select substring(Date,7,4) = '2015')
```

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

42]:

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

In 2015, There are two failures in April from booster version F9 v.1.1 B1012 and F9 v1.1 B1015

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

```
[18]: %sql SELECT Landing_Outcome, count (*) as Landing_Counts from SPACEXTBL WHERE Date BETWEEN '04/06/2010' AND '20/03/2017' group by Landing_Outcome ORDER BY count(*) D
```

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

[18]:

Landing_Outcome	Landing_Counts
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

There are more success landing between 2010-06-04 and 2017-03-20 (total 20 success attempt) than the failures

Section 3

Launch Sites Proximities Analysis

Launch Sites Location Map

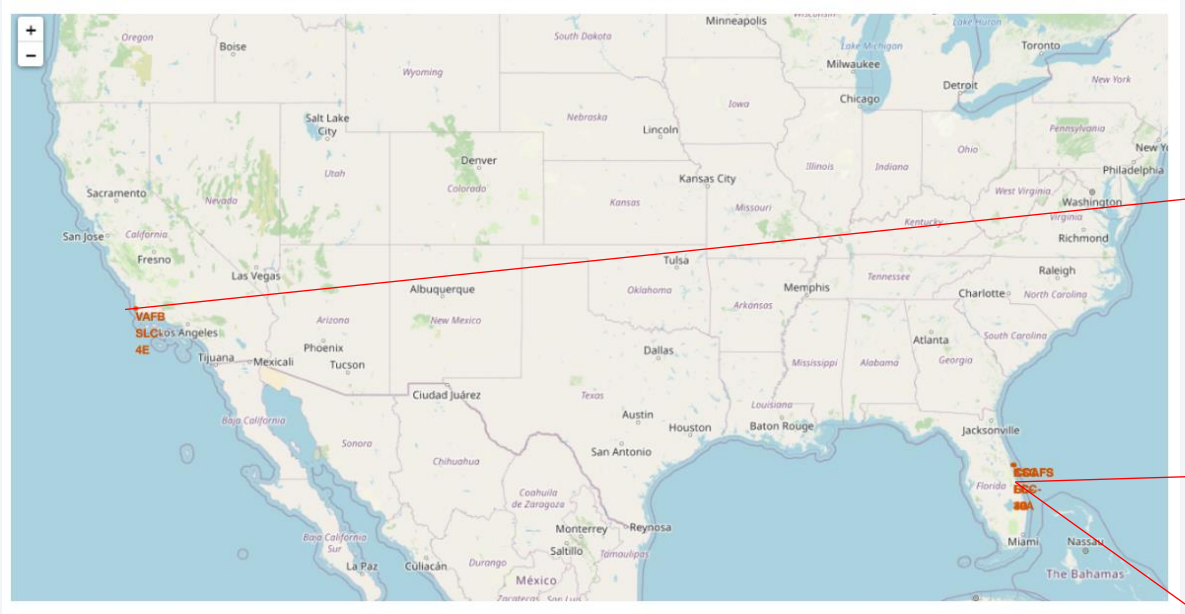


Fig 3.1. 1 Launch Sites Location Map

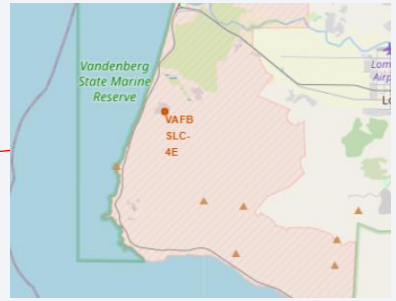


Fig 3.1.2 Launch Sites VAFB SLC-4E

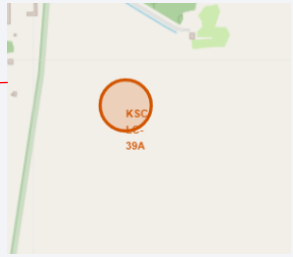


Fig 3.1.3 Launch Sites KSC LC-39A

Fig 3.1.4 Launch Sites , CCAFS LC-40, and CCAFS SLC-40

The launch sites are in United States.

There are two different location, California and Florida

Launch sites in California are VAFB SLC-4E

Launch sites in Florida are KSC LC-39A, CCAFS LC-40, and CCAFS SLC-40

Launch Sites Map Success or Unsuccess

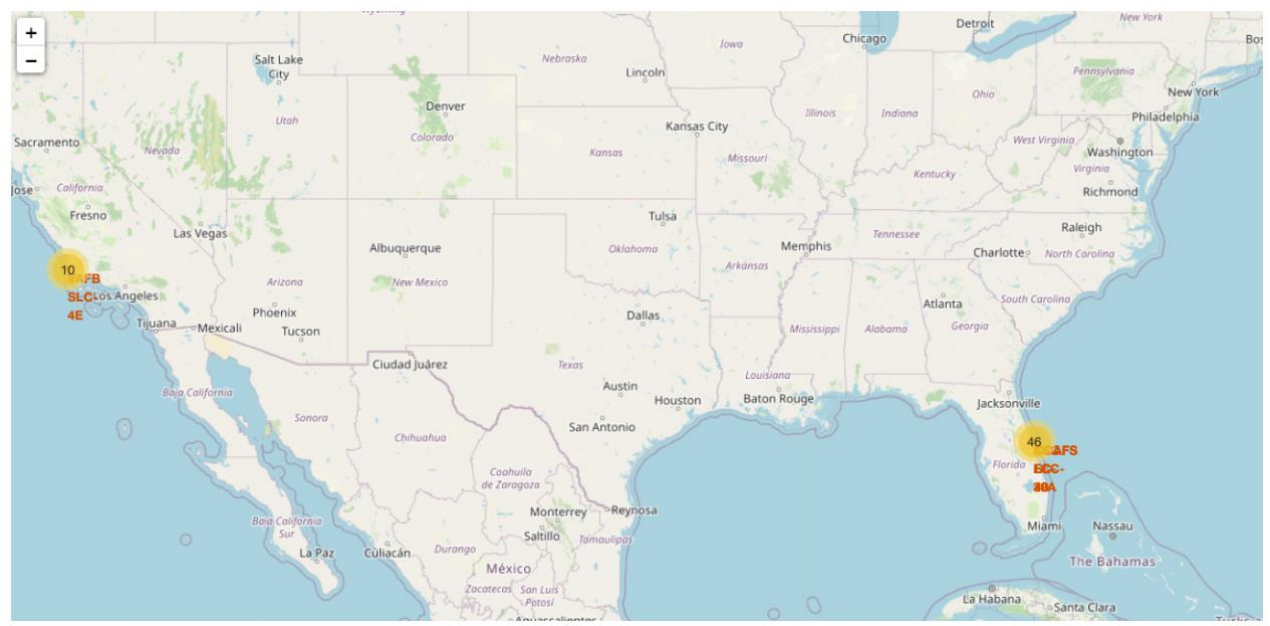


Fig 3.2.1 Launch Sites Location Map

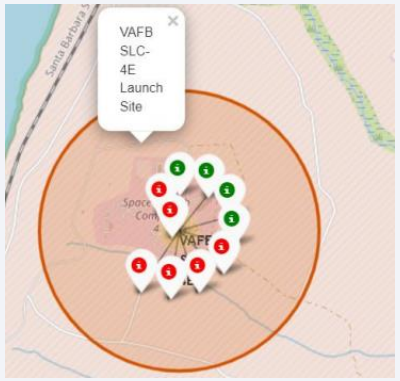


Fig 3.2.2 Success/unsuccess VAFB SLC-4E

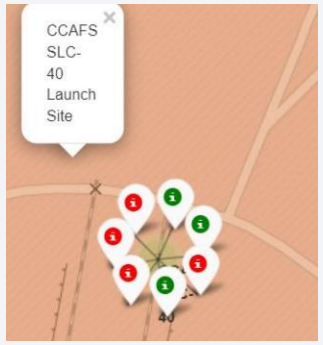


Fig 3.2.3 Success/unsuccess CCAFS LC-40

KSC LC-39A has the most success rate

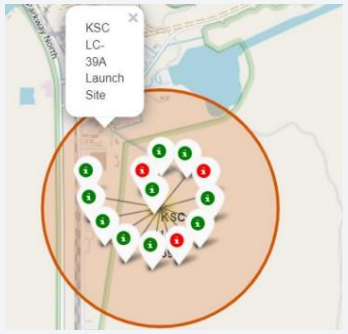


Fig 3.2.3 Success/unsuccess KSC LC-39A

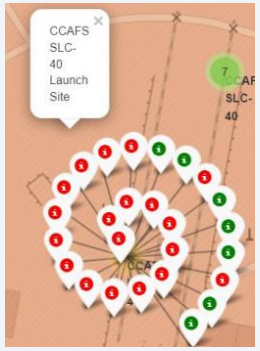


Fig 3.2.4 Success/unsuccess CCAFS SLC-40

Launch Site Distance Map

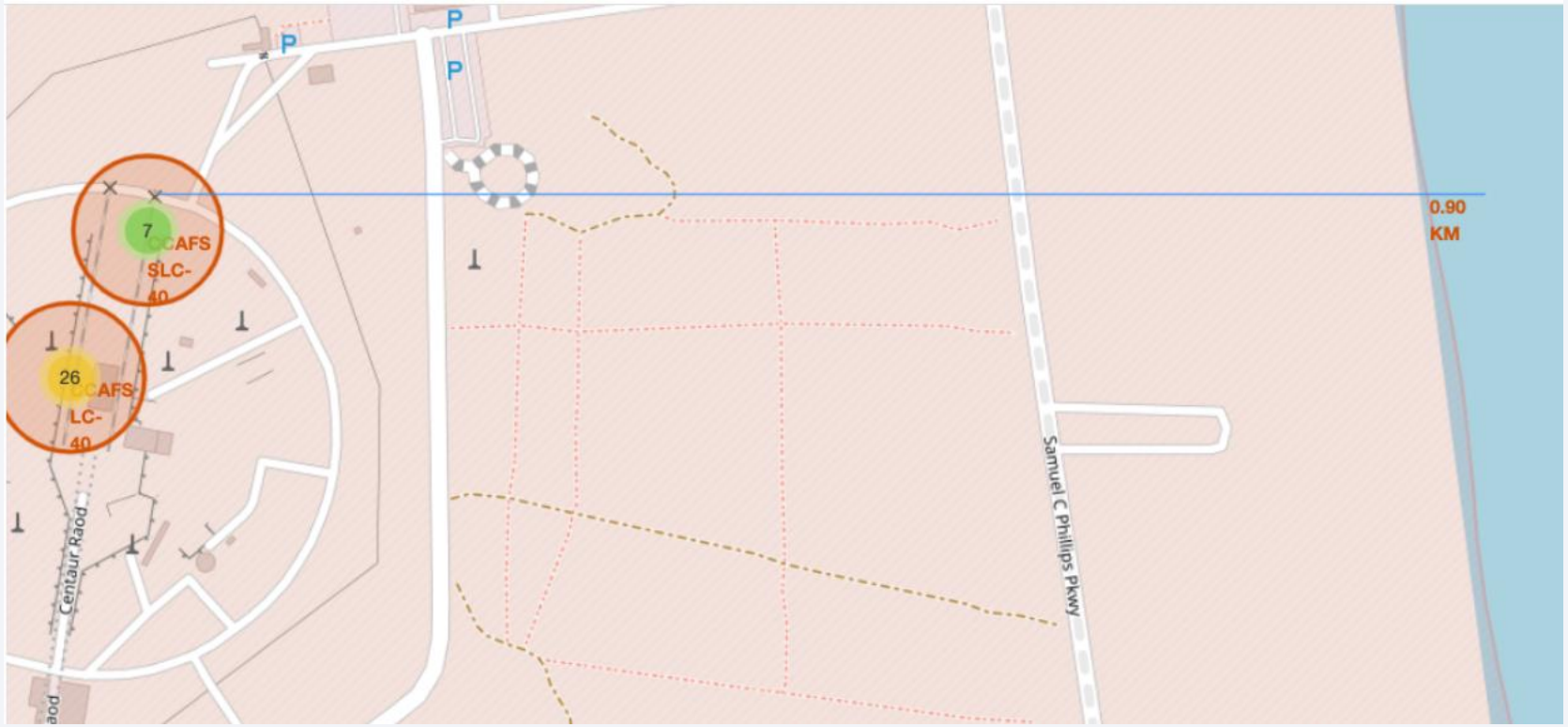
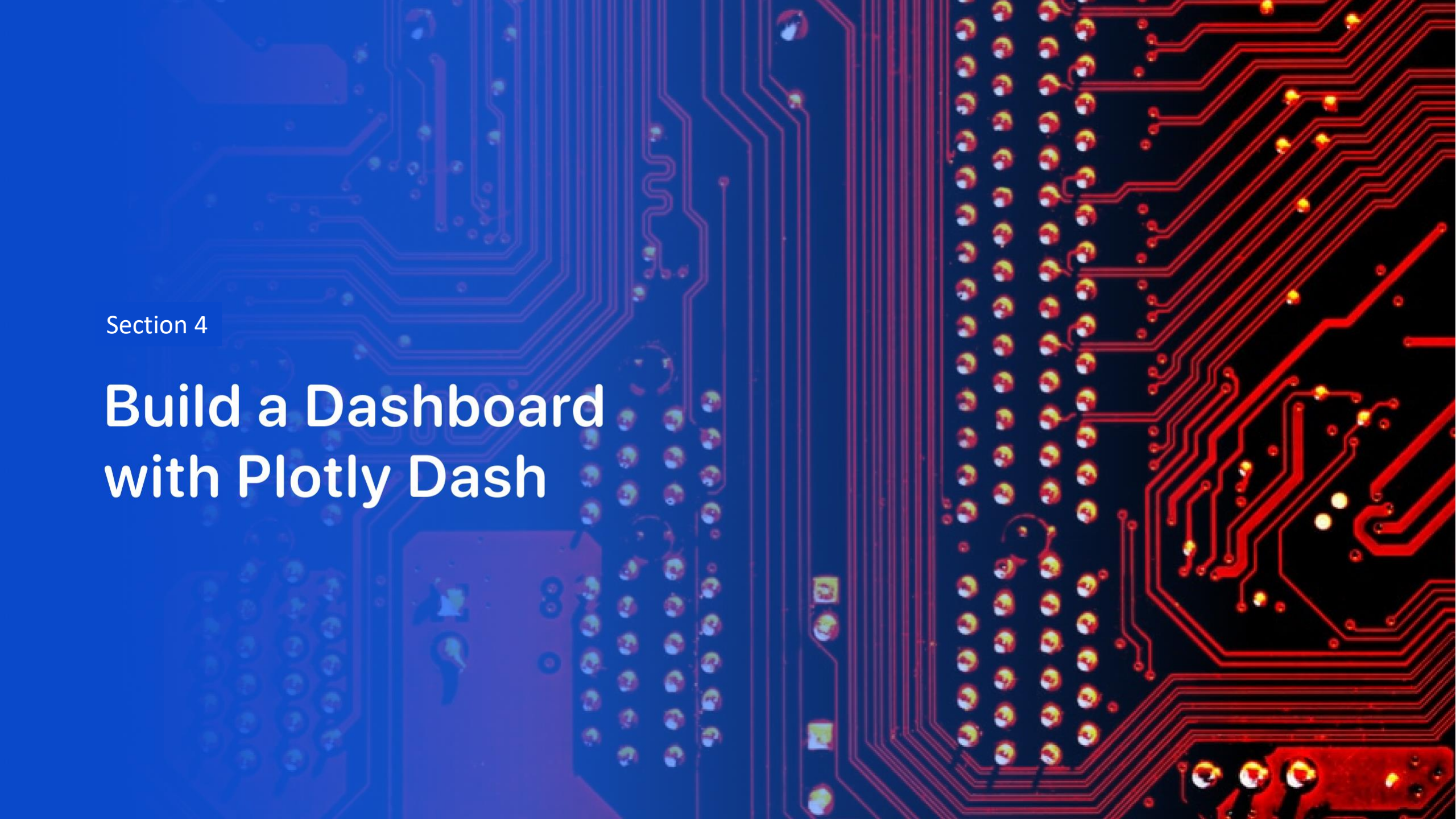
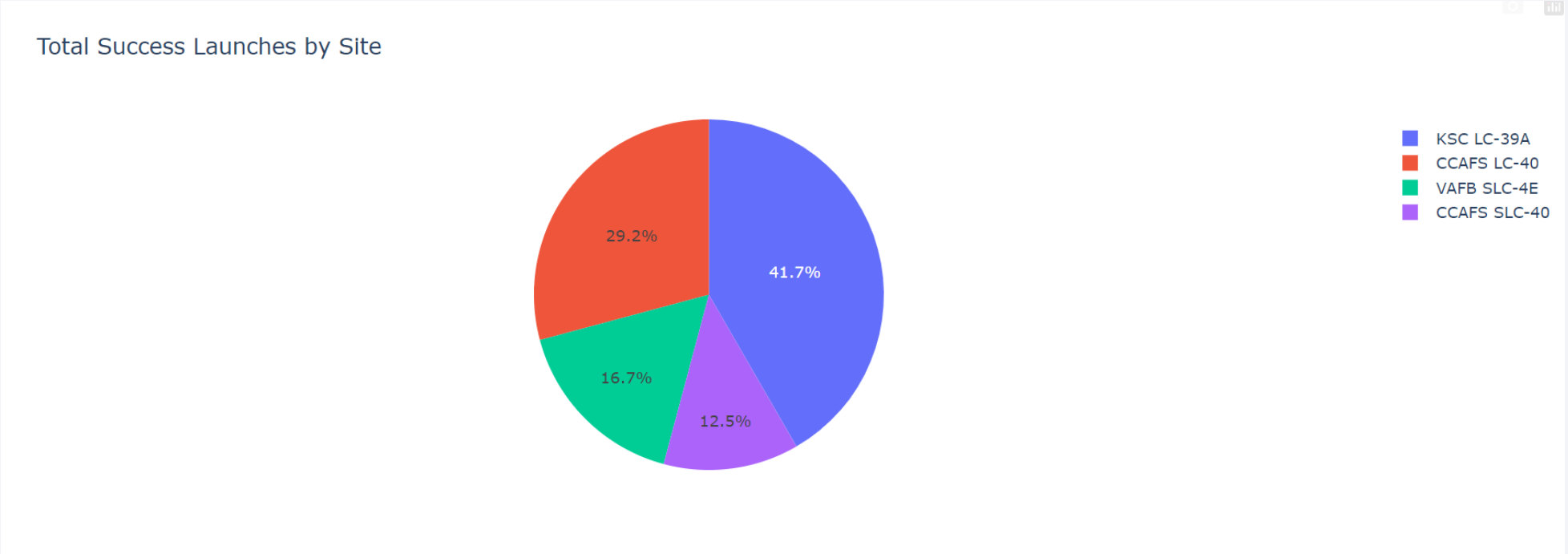


Fig 3.3.1 Distance between CCAFS SLC-40 and coastline

The launch sites are more close to coastline compared to cities

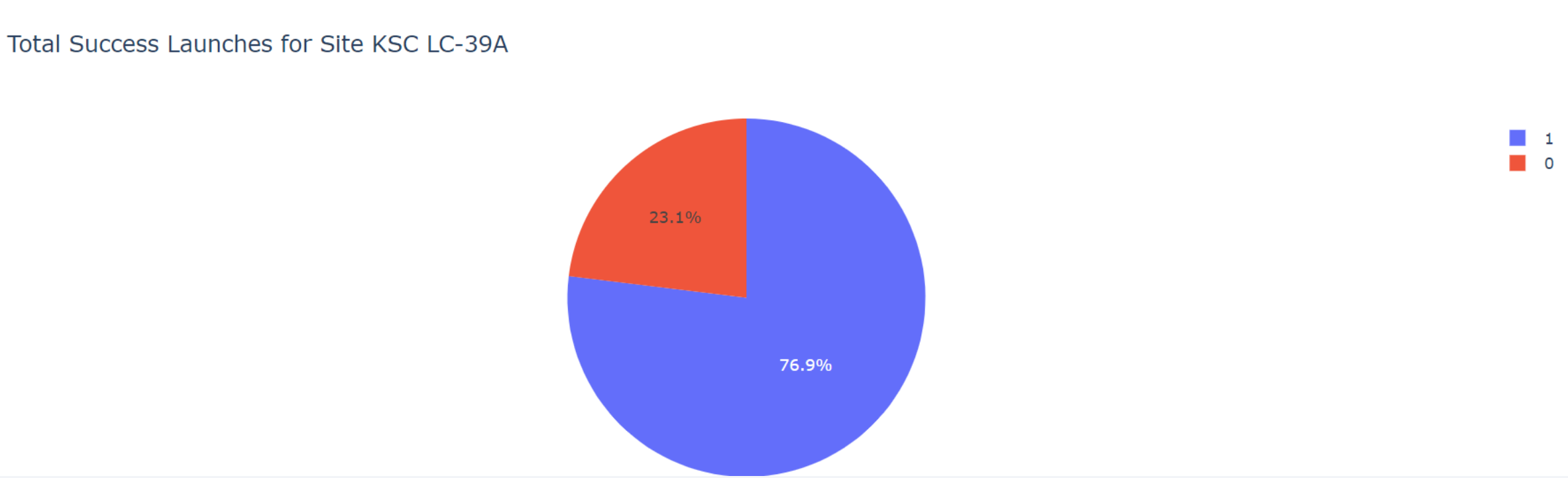

Section 4
Build a Dashboard with Plotly Dash

Success Rate by Sites



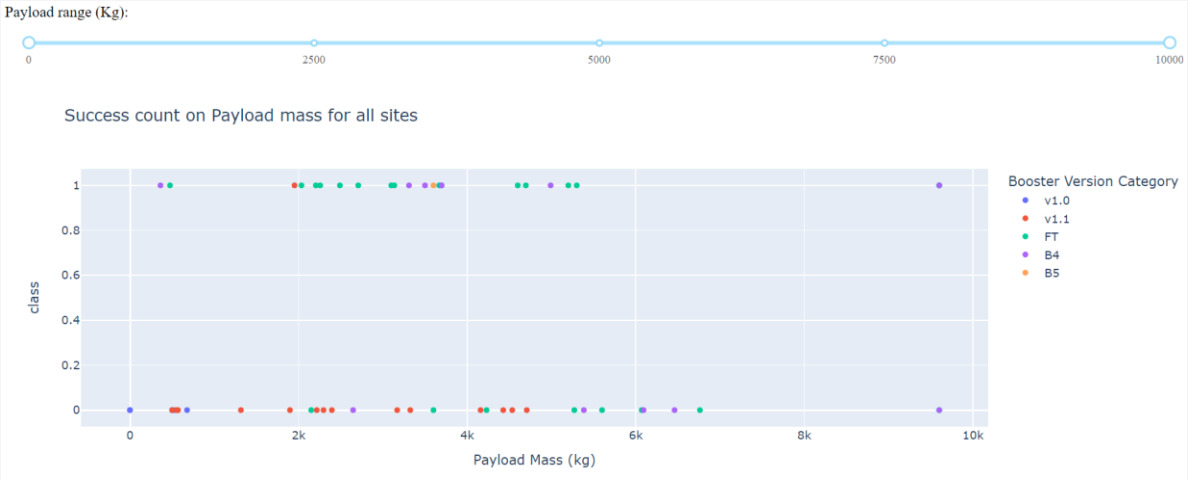
KSC LC-39A has the most launch success rate

KSC LC-39A Launch Success Ratio



The total success ratio of KSC LC-39A is 76.9%

Succes Count on Payload Mass

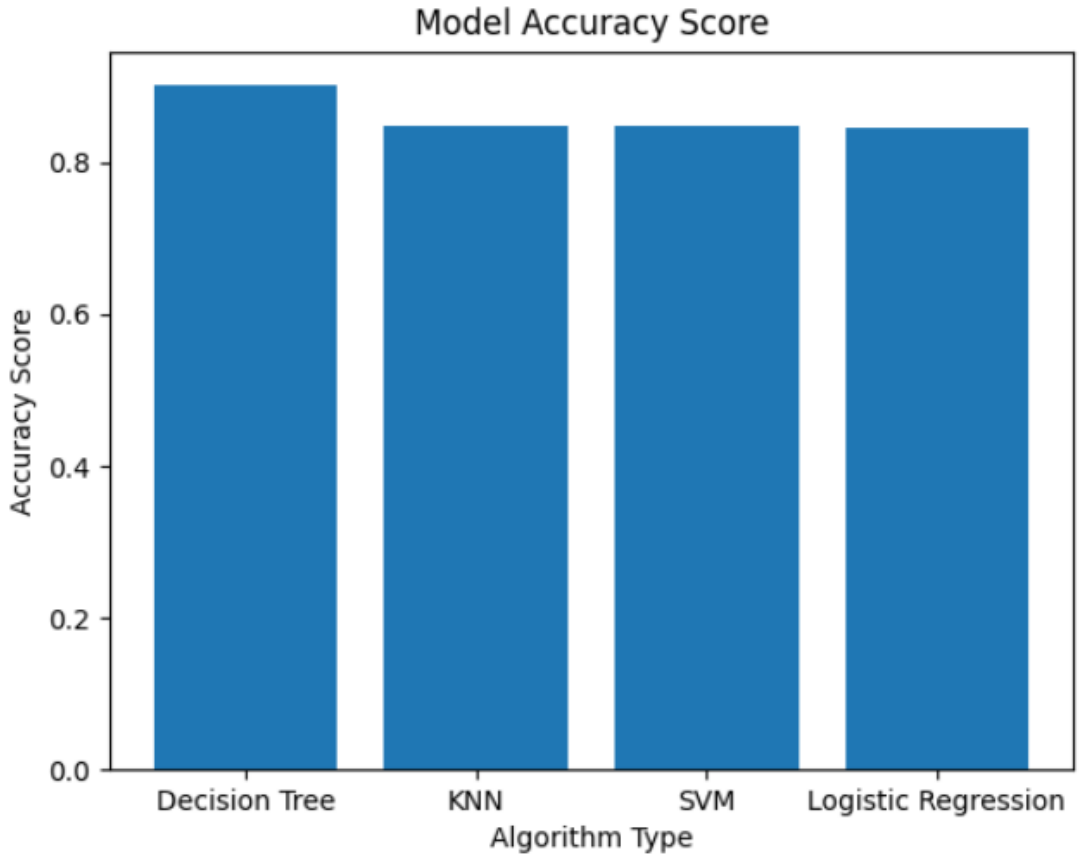


- Booster Version FT has the most success count (13)
- Payload in range of 2000 to 5500 has the most success count (13 success launch)

Section 5

Predictive Analysis (Classification)

Best Performing Model

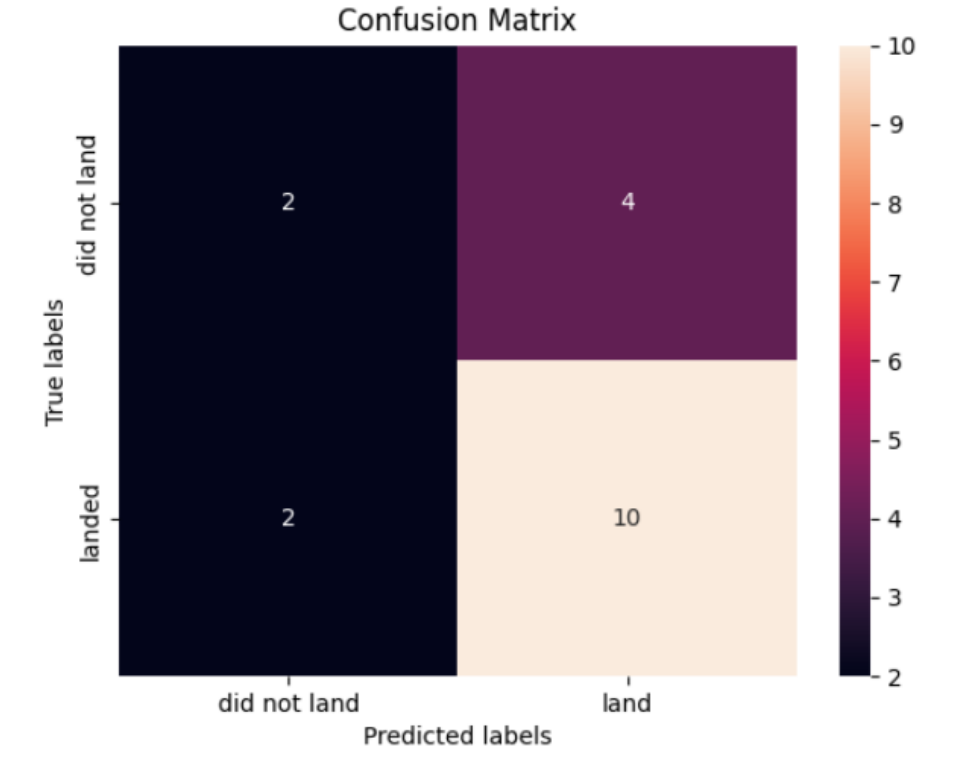


[33]:

	Algo Type	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.901786	0.666667
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

Decision Tree is the best model in this case

Confusion Matrix



Confusion Matrix of Decision Trees

- True Positive

There are 10 scenarios that the model predicted land and the outcome is true (bottom right)

- True Negative

There are 2 scenarios that the model predicted didn't land and the outcome is true (upper left)

- False Positive

There are 4 scenarios that the model predicted land and the outcome is false (upper right)

- False Negative

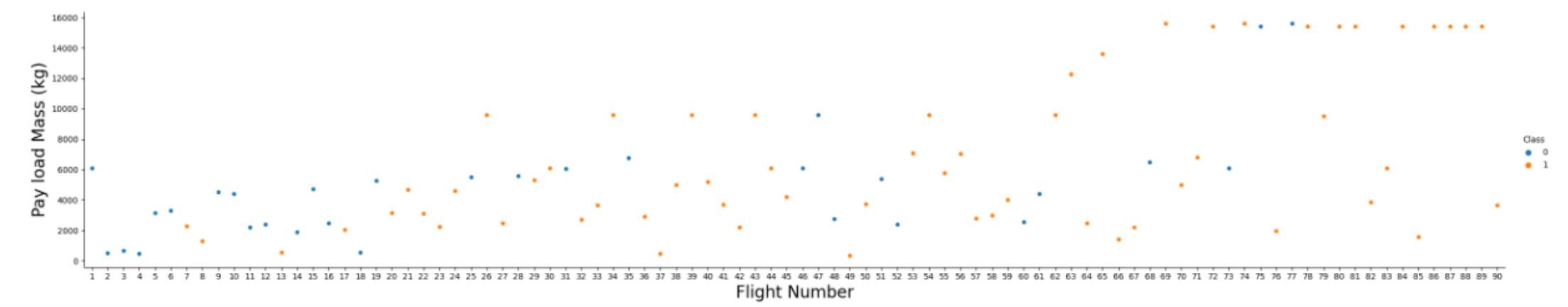
There are 4 scenarios that the model predicted didn't land and the outcome is false (bottom left)

The model's test data accuracy is 0.67%

Conclusions

- Successful launch is increasing as the flight number increasing
- The success rate is generally increasing (2013-2020)
- Site KSC LC-39A has the most launch success rate
- Booster Version FT has the most success count
- Decision Tree is the best model in this case to predict if the Falcon 9 first stage will land successfully

Appendix



Relationship between Payload mass and Flight number

Thank you!