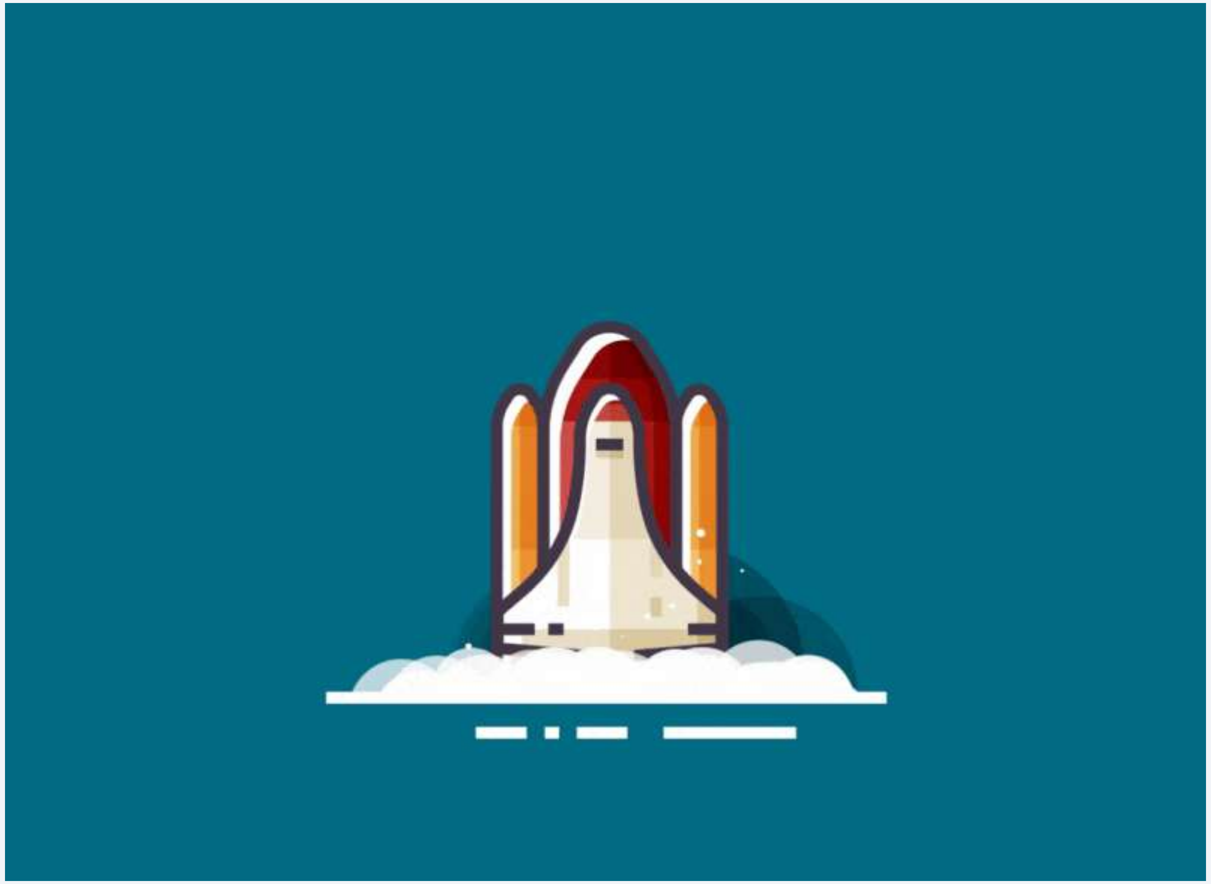
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shalini Kothuru
08/06/2023

Outline

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



Executive Summary

Summary of methodologies

- Data Collection through API and SPACEX wiki page with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with and Data Visualization
- Interactive Visual Analytics with Folium
- Dashboard using Plotly
- Machine Learning Prediction

Summary of results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

Background

- The Rocket launch costs depends on recovery of stage1 parts of rocket
- Space X will launch rockets at $62 Billon while other companies charge $165 Billion USD because of the reuse of Stage1 parts.
- Space Y want to compete will Space X to reduce cost of rocket launch by analyzing the conditions that result in successful landings

Problem we want to find answers

- Determine the factors the result in a successful landing of rocket
- Relation between variables that affect the landing of a rocket
- Use the data available on the internet and predict whether a launch

  will be successful or not using Machine Learning.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was integrated from resources like SpaceX public API and SpaceX wikipedia
- Perform data wrangling
  - We have marked ‘1’ for all categories of successful landings and ‘0’ for unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Leveraged GridSearch to find best parameters for different classification algorithms and evaluated using score method.

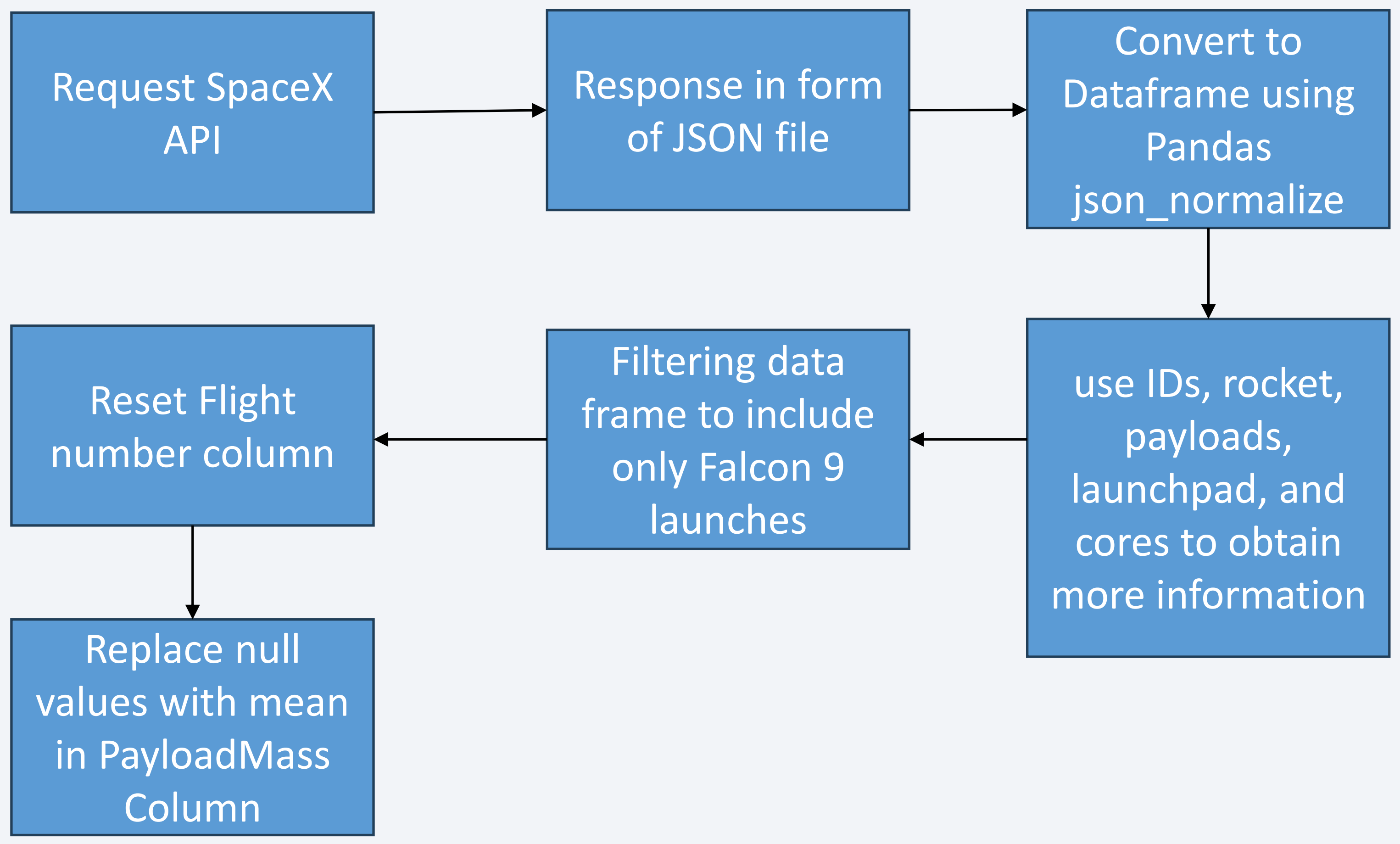
Data Collection

- Data is collected from 2 resources
  - Data Collection from SpaceX Public API
    - a combination of API requests from Space X public API
  - Data Collection from Web Scraping SpaceX Wikipedia
    - web scraping data from a table in Space X's Wikipedia entry

Data Collection – SpaceX API

Columns

- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- GitHub

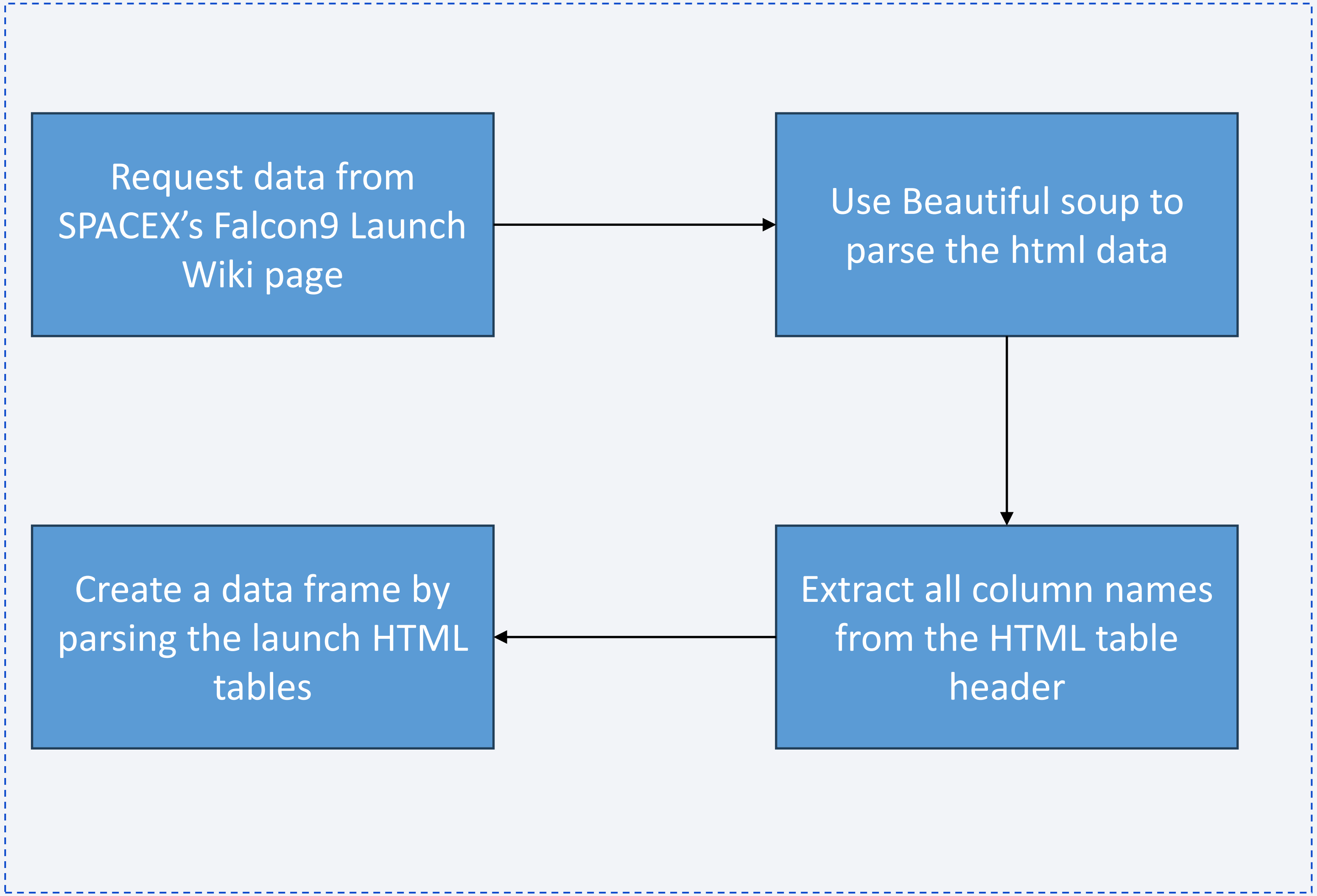


Data Collection - Scraping

Columns:

- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

GitHub



Data Wrangling

- Applied on one hot encoding to convert categorical variables to continuous variable
- Created a class or target variable from Outcome column
  - Class = 0 if Outcome equals to 'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'
  - Class = 1 if Outcome equals to 'True ASDS', 'True RTLS', 'True Ocean'
- '1' indicate successful landing while 0 indicates unsuccessful landings
- 66% of the records belongs to successful landings
- [GitHub]

EDA with Data Visualization

- Used various plots like Scatter plot, bar graph, line chart to identify relationship between different variables.
- Scatter plot(for categorical vs continuos) is used for plotting payload mass for each launch site
- Bar graph (for categorical variables) to identify relation success rate in each orbit
- Line chart is used to see the trend of success rate year wise
- GitHub

EDA with SQL

- Loaded the data into SQLite database
- Used SQL language using sql magic to explore the dataset from jupyter notebook itself
- Identified relationships between variables
- Got a better understanding on the data
- GitHub

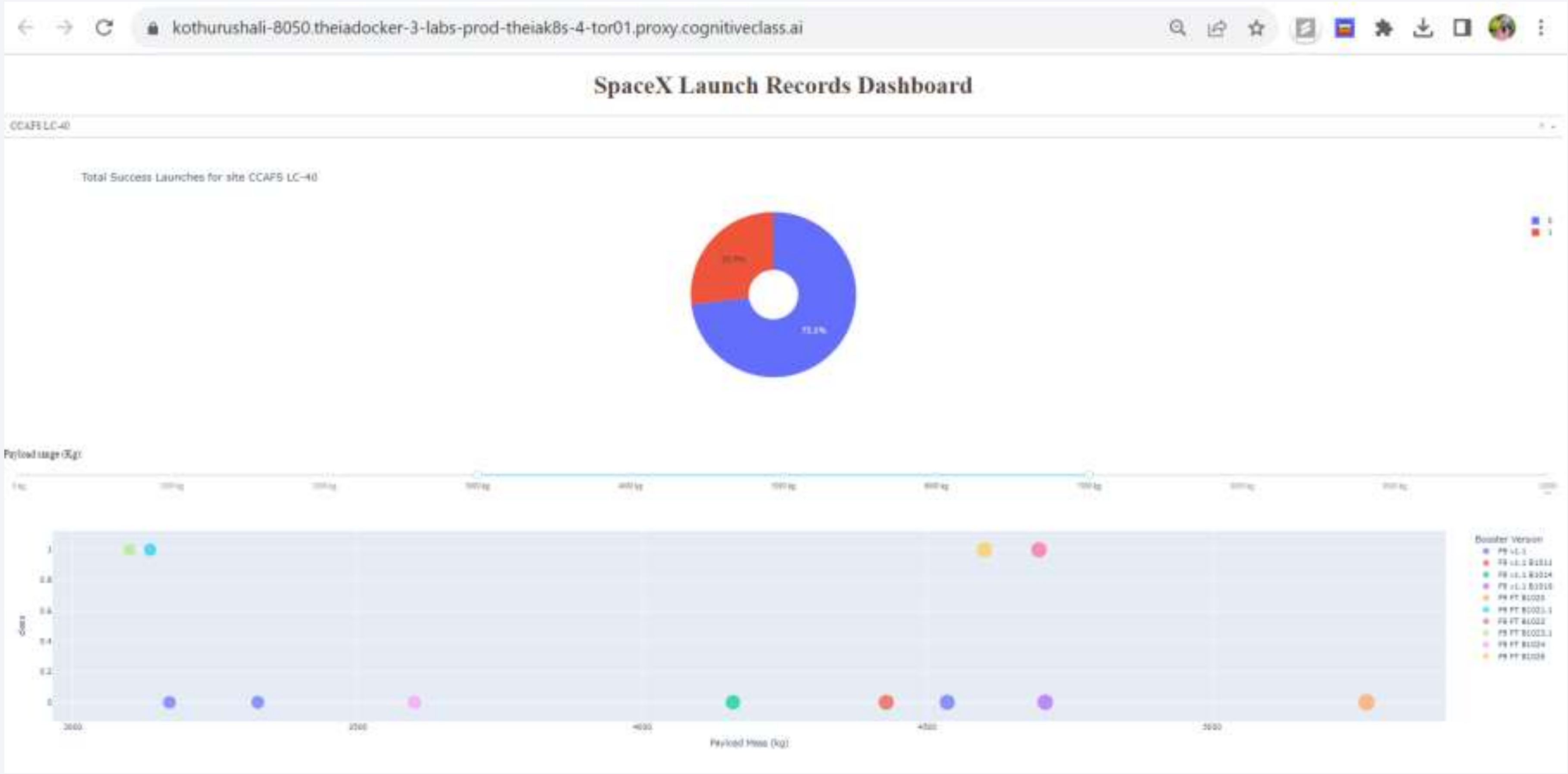
Build an Interactive Map with Folium

- Folium maps were utilized to visually represent various key elements related to SpaceX's launch sites
  - Markers for SpaceX launch sites globally.
  - Circles representing launch sites
  - Differentiating markers for successful and unsuccessful landings.
  - Proximity examples to key locations: Railway, Highway, Coast, City using lines.
- Using these maps, I was able to
  - Analyze factors influencing launch site locations.
  - Visualizing successful landings in relation to key locations.
  - Insights for future launch site selection and mission planning.
- GitHub

Build a Dashboard with Plotly Dash

- Built an interactive dashboard using Plotly dash
- Dashboard contains a dropdown for input launch site and slider for payload mass
- Selecting the launch site give the success rate of that site using pie chart
- Selecting the slider plots scatter plot to show the relation between outcome and payload mass for different booster version
- GitHub

Build a Dashboard with Plotly Dash

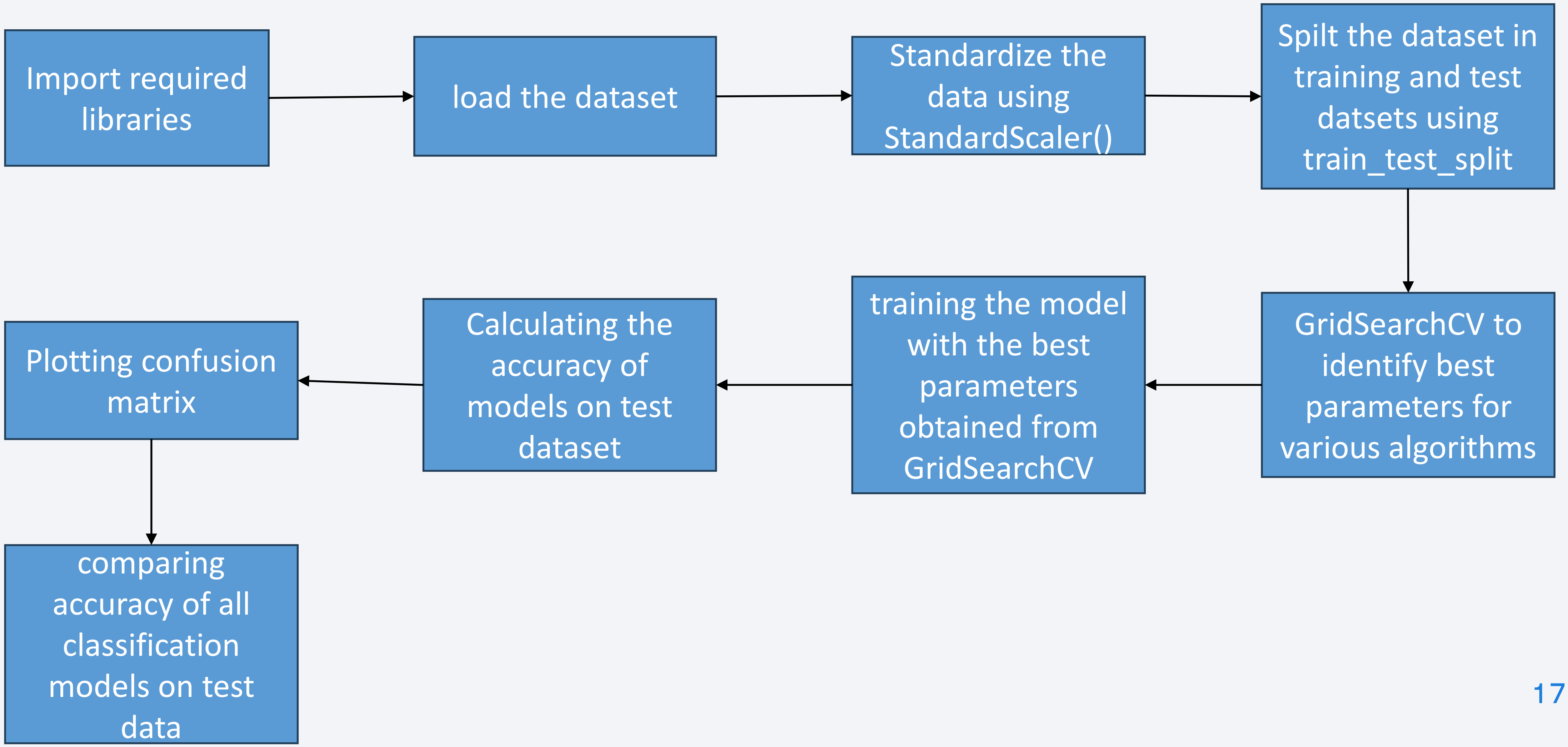


Dashboard using Plotly dash that shows success rate of CCAFS LC-40 launch site and scatter plot to show relation between class and payload mass in range 3000 to 7000 kgs

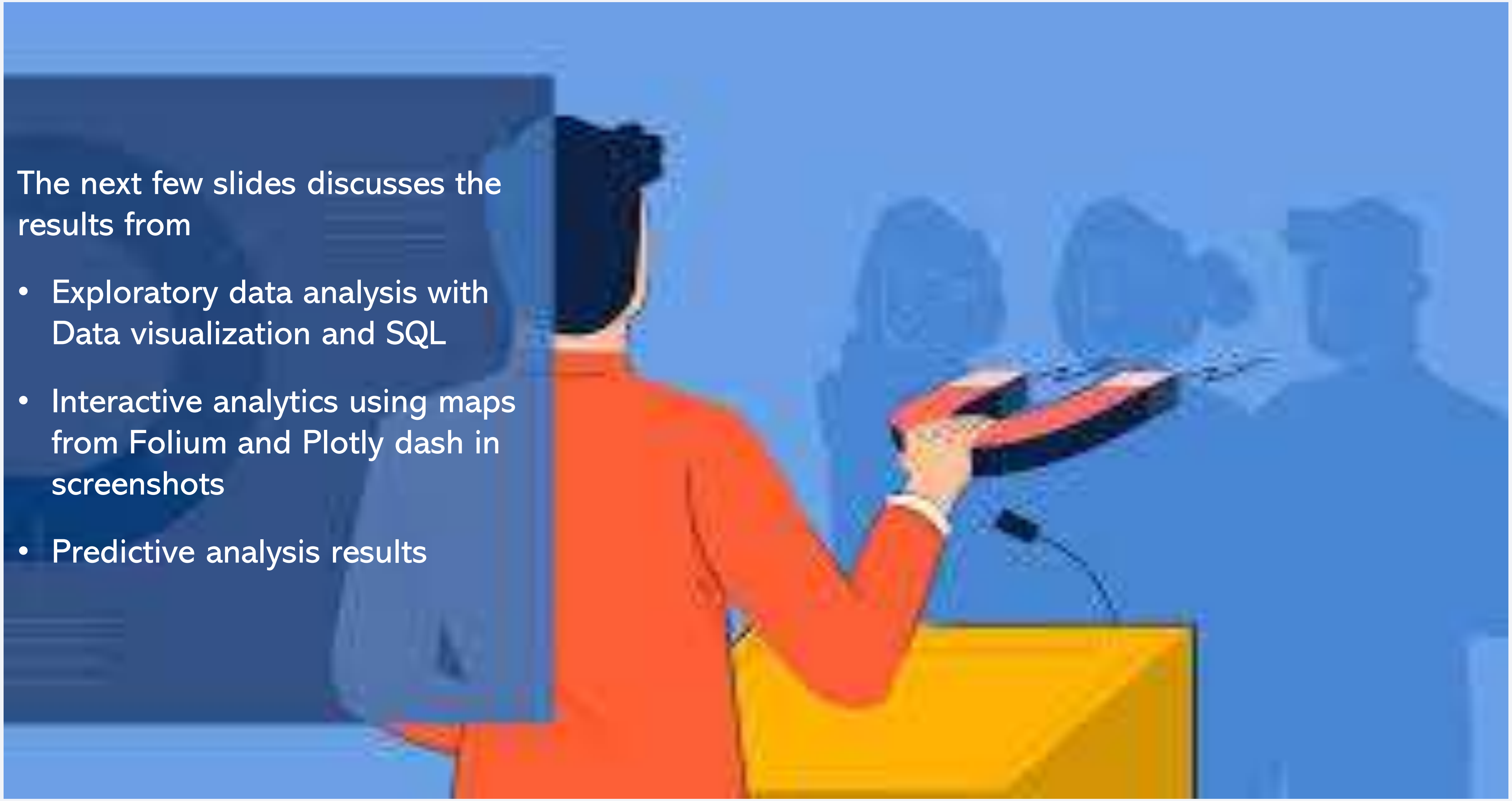
Predictive Analysis (Classification)

- Used GridSearchCV to find the best parameters for each classification algorithm like Logistic Regression, Support Vector Machine(SVM), Decision Tree Classifier, K Nearest Neighbors
- Calculated accuracy using score method
- All models resulted in accuracy of 83%
- The next slide illustrates the flow chart of the predictive analysis
- GitHub

Predictive Analysis (Classification)



Results



The next few slides discusses the results from

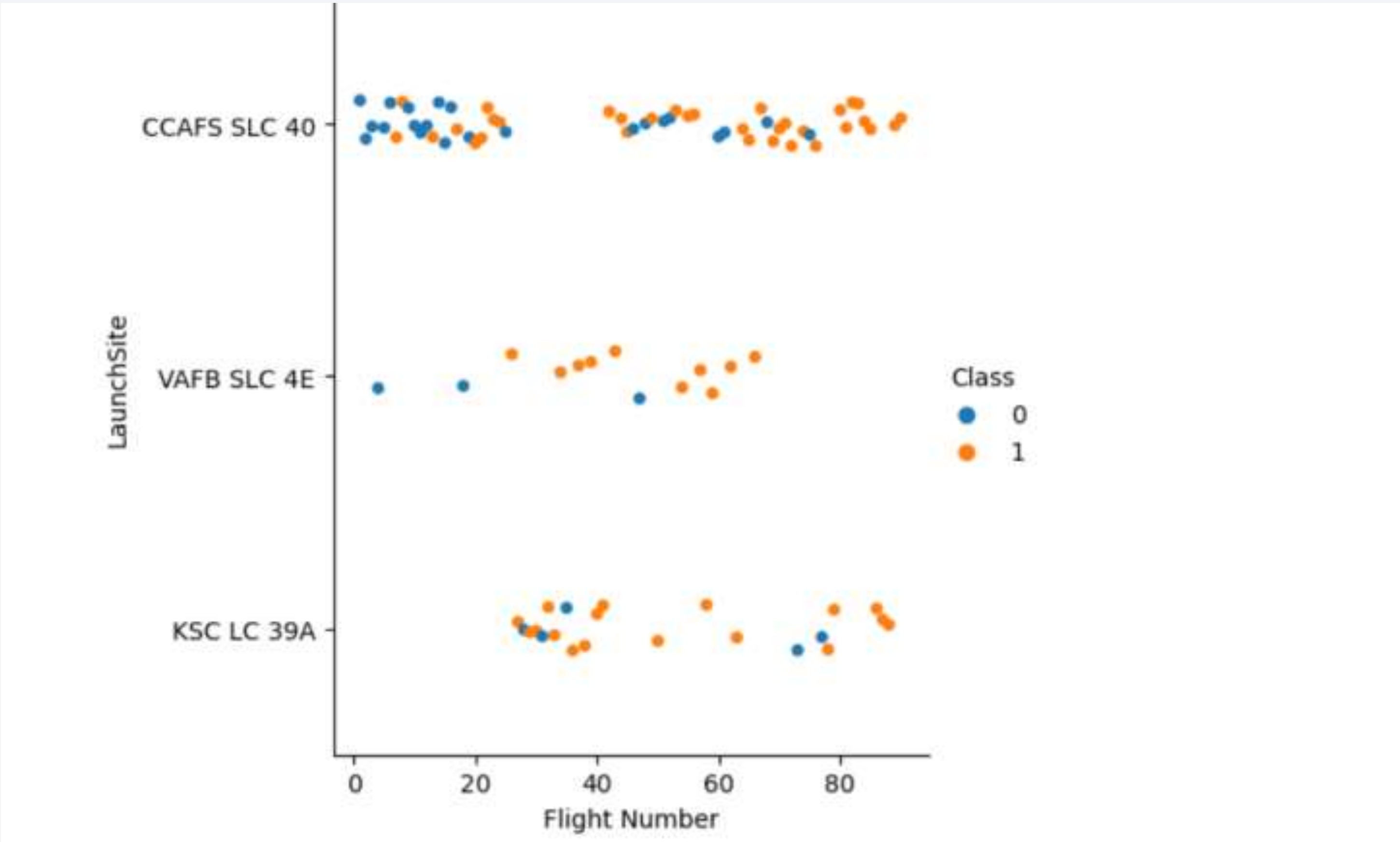
- Exploratory data analysis with Data visualization and SQL
- Interactive analytics using maps from Folium and Plotly dash in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

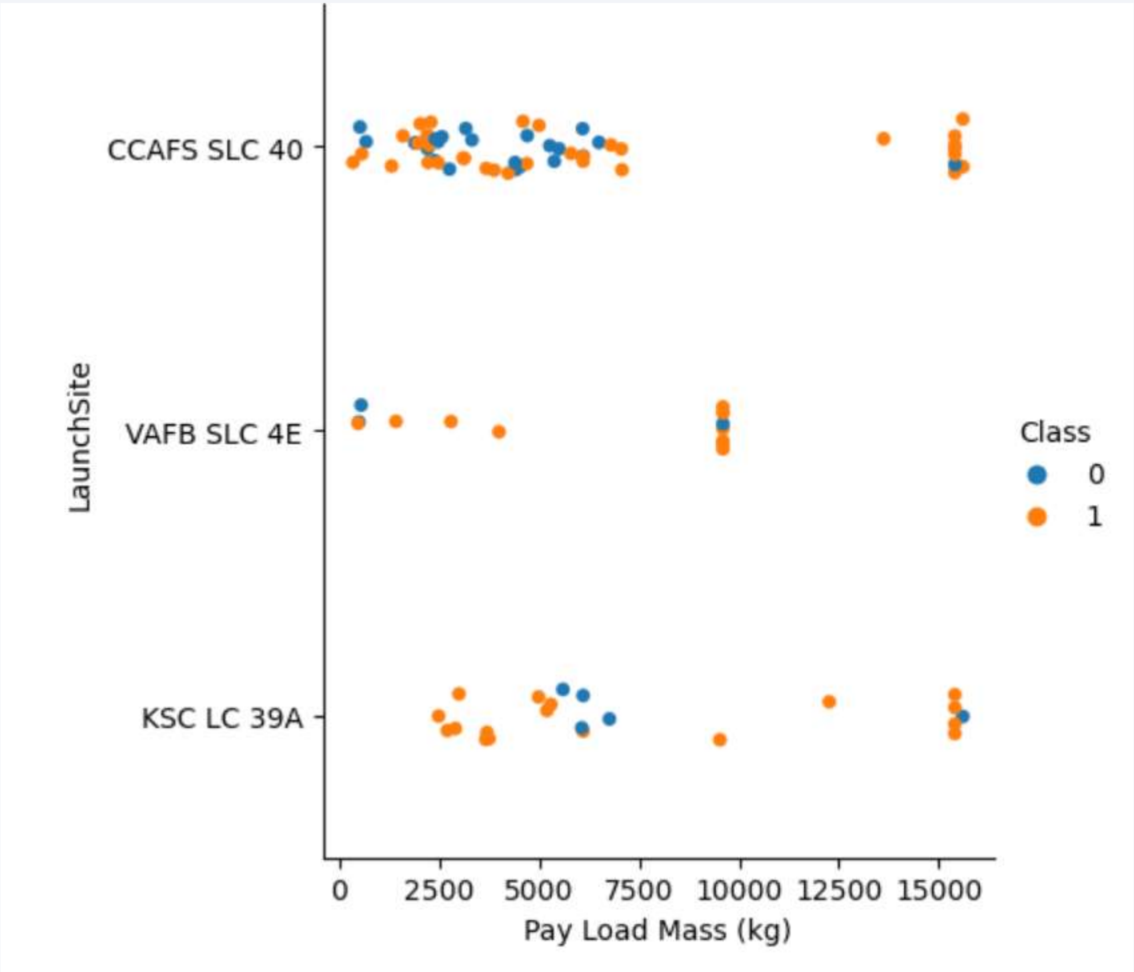
Flight Number vs. Launch Site

- CCAFS SLC 40 launch site has more number of launches while VAFB SLC 4E has less
- Also, ratio of unsuccessful launches is more in CCAFS SLC 40
- Flight number value with more 80 always resulted in successful landing



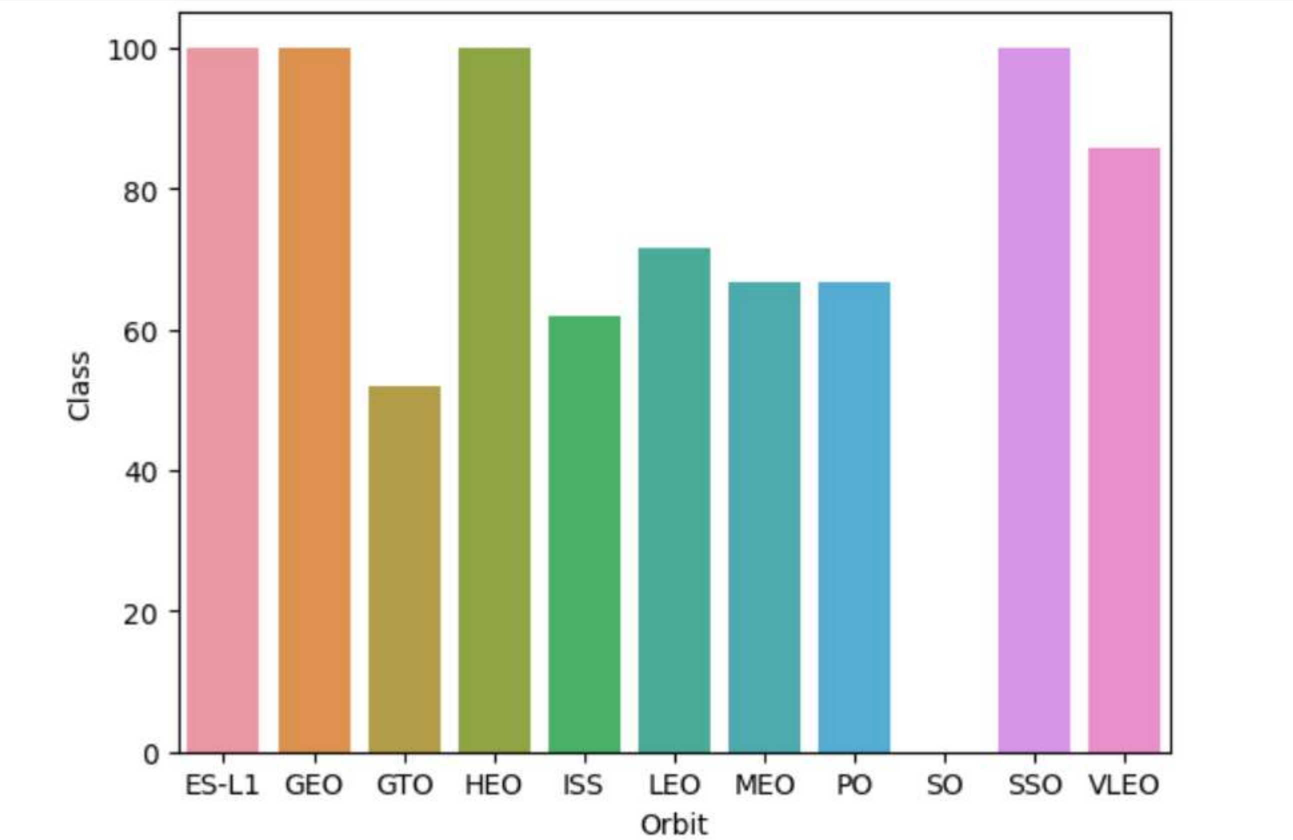
Payload vs. Launch Site

- Greater pay load mass resulted in successful landings at launch site CCAFS SLC 40
- Payload mass around 2500 lead to success at launch site KSC LC 39A, VAFB SLC 4E



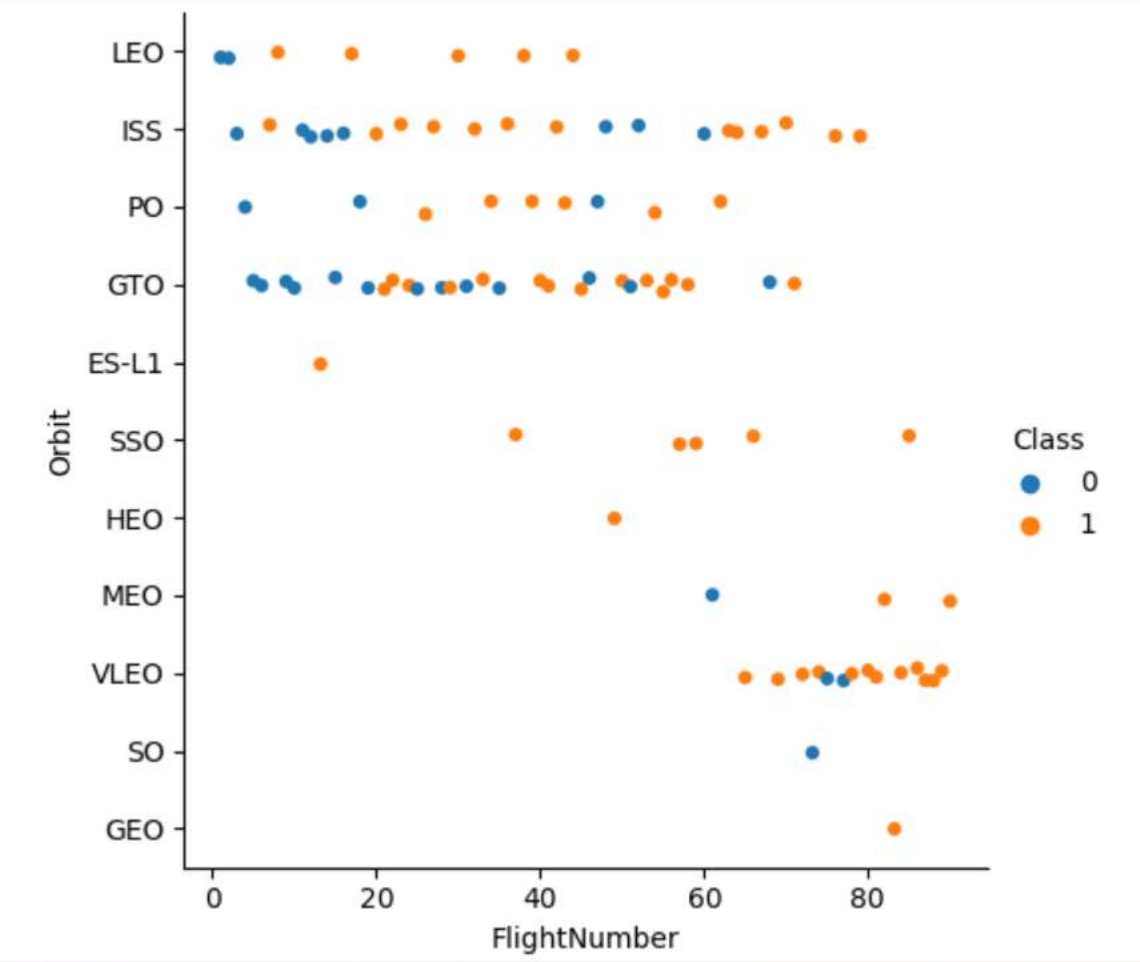
Success Rate vs. Orbit Type

- At least 40 records indicate successful landing in each orbit except SO
- ES-L1, GEO, HEO, SSO have highest number of successful landings
- SO has no successful landings at all



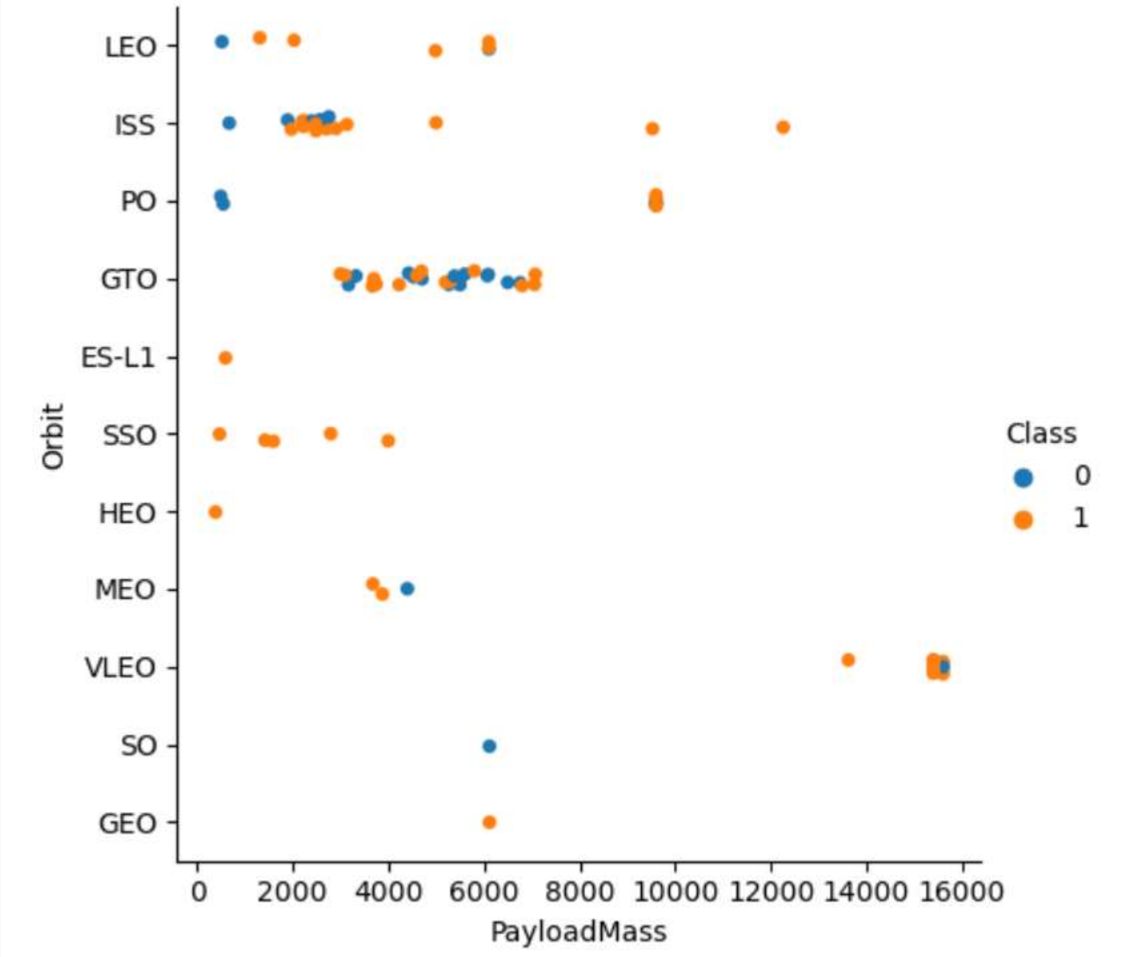
Flight Number vs. Orbit Type

- The flights operated in SSO always have successful landings
- FlightNumber with low values less than 40 are the one which resulted in unsuccessful landings or not operated at all in some orbits like SSO, HEO, MEO, VLEO, SO, GEO
- GEO and SO has only one successful and unsuccessful landing respectively



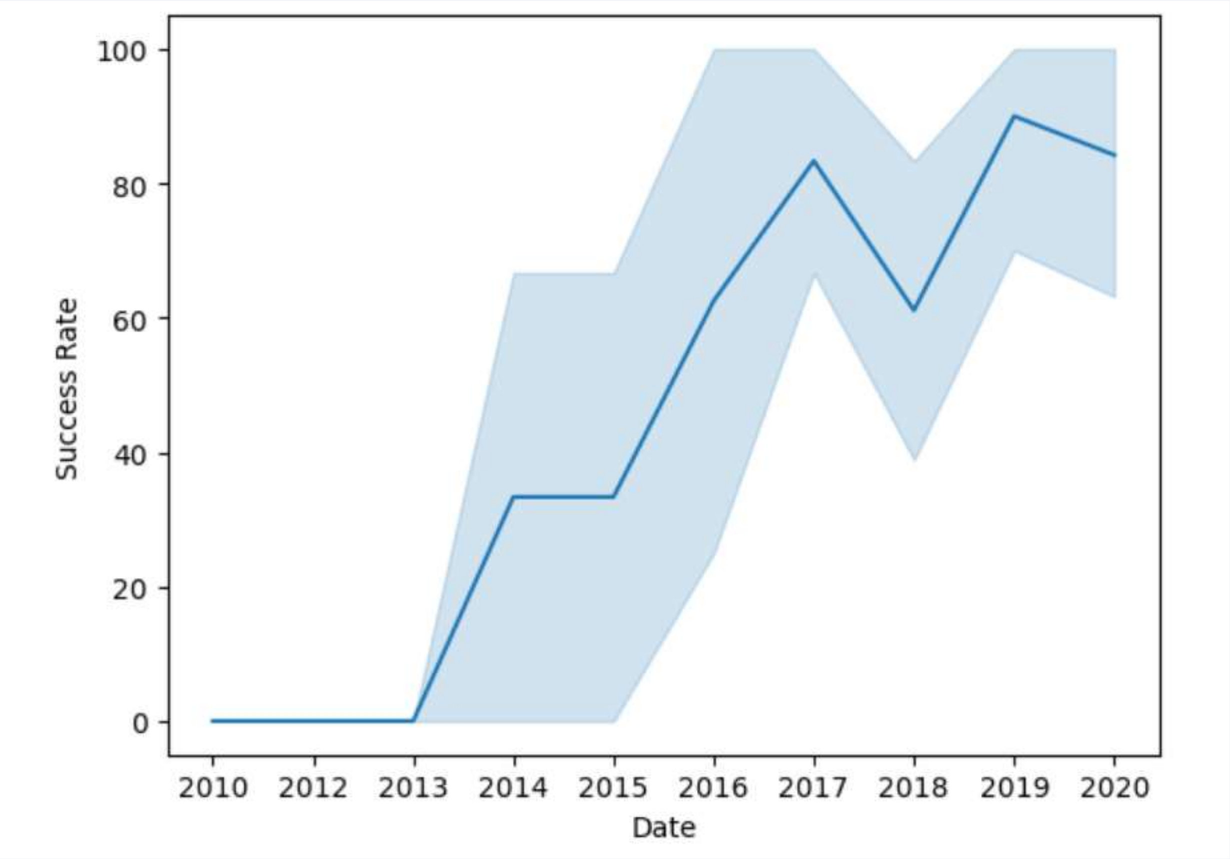
Payload vs. Orbit Type

- Heavy pay loads resulted in successful landings in orbits ISS, PO, VLEO
- With low payloads in orbits ES-L1, SSO, HEO, rockets were landed successfully



Launch Success Yearly Trend

- There is significant increase in success rate from 2013 to 2020
- 2019 had most of the success rate compared to all years



All Launch Site Names

There are four launch sites in out dataset namely

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

In [16]:
```
%sql Select distinct Launch_Site from SPACEXTABLE
```

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

Out[16]:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- From the result set, we can be these are in orbit LEO

```
%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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	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 [21]:  %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"

 * sqlite:///my_data1.db
Done.

Out[21]:  sum(PAYLOAD_MASS__KG_)
                         45596
```

- The total payload carried by boosters from NASA is 45596 Kgs

Average Payload Mass by F9 v1.1

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

In [23]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

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

Out[23]:

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- The average payload mass carried by booster version F9 v1.1 is 2534.6 Kgs

First Successful Ground Landing Date

```
%sql select min(date) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)' limit 1
```

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

min(date)
2015-12-22

- The date of the first successful landing outcome on ground pad is on 22nd December, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 40
```

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

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

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 have version F9 FT B10%

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

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

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

- There are 100 success mission outcomes, 1 failure mission outcome in our dataset.

Boosters Carried Maximum Payload

- The list of the boosters which have carried the maximum payload mass

```
%sql Select 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

```
%sql Select substr(Date, 4, 2), Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome = 'Fail
```

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

substr(Date, 4, 2)	Landing_Outcome	Booster_Version	Launch_Site
5-	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
5-	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- There are only 2 records which have failed landing_outcomes in drone ship in year 2015

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

```
%sql select Landing_Outcome, count(Landing_Outcome) as cnt from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20'
```

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

Landing_Outcome	cnt
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- The landing outcomes (such as Failure (drone ship) or Success (ground pad)) and its count between the date 2010-06-04 and 2017-03-20, in descending order
- The count of success landing outcomes are more

Section 3

Launch Sites Proximities Analysis

Folium Map highlighting launch sites

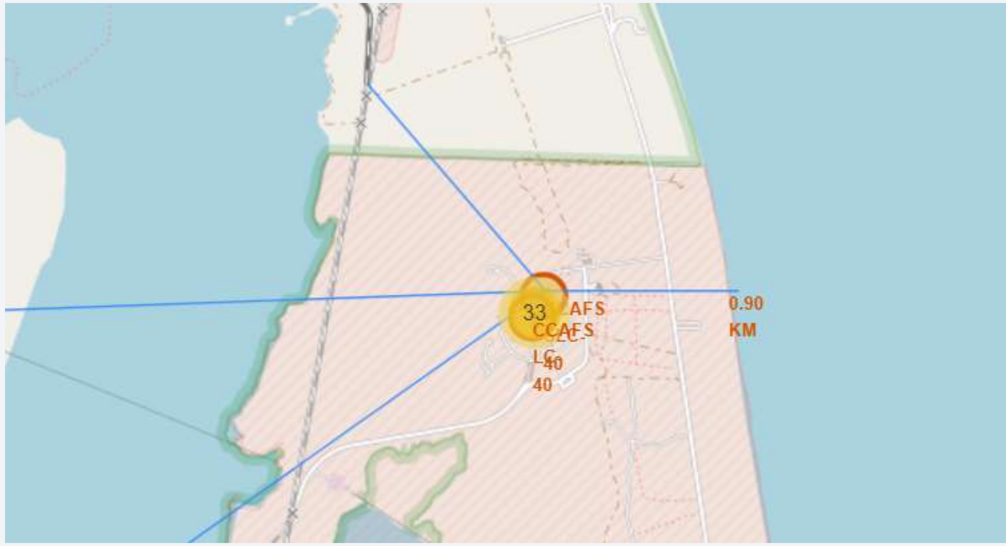
- The map on left side highlights the four different launch sites present in our dataset
- The map on right side highlights launch sites CCAFS LC-40, CCAFS SLC-40
- From map we can see that most of the launch sites are located near ocean

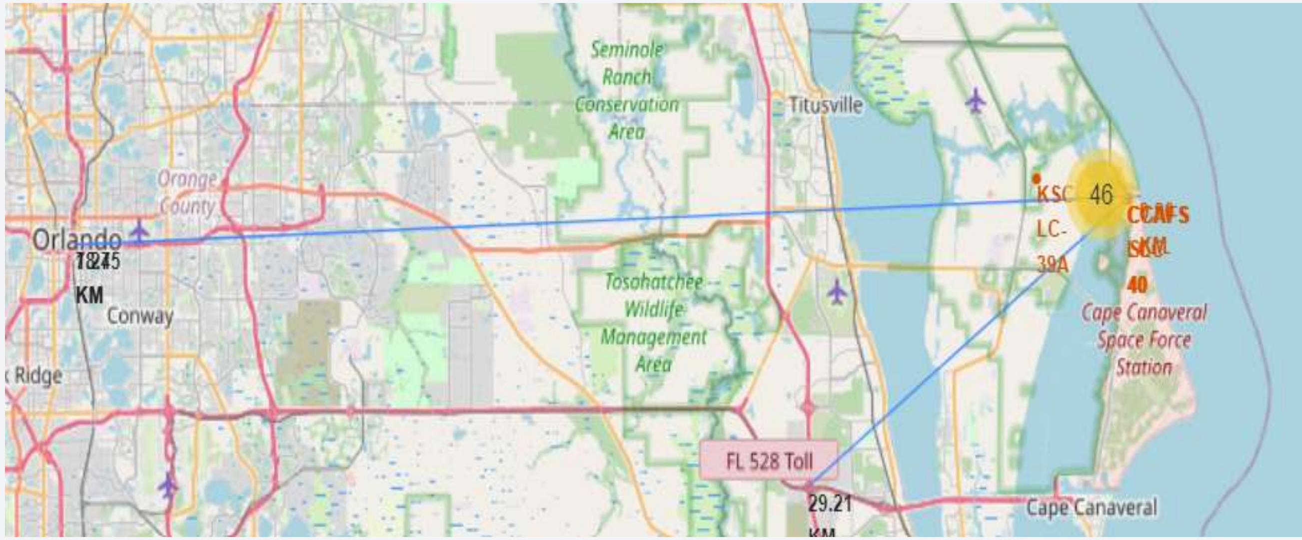
Folium Map highlighting the launch outcome for each site

- From above map, we can say that 46 launches are from launch site that are located on east coast and 10 launches are from western coast launch site

- The folium map on right side highlights the color based (green indicates success, red failure) the launch outcome at CCAFS SLC-40 launch site

proximity of highway, coastal line and city from CCAFS SLC-40 launch site





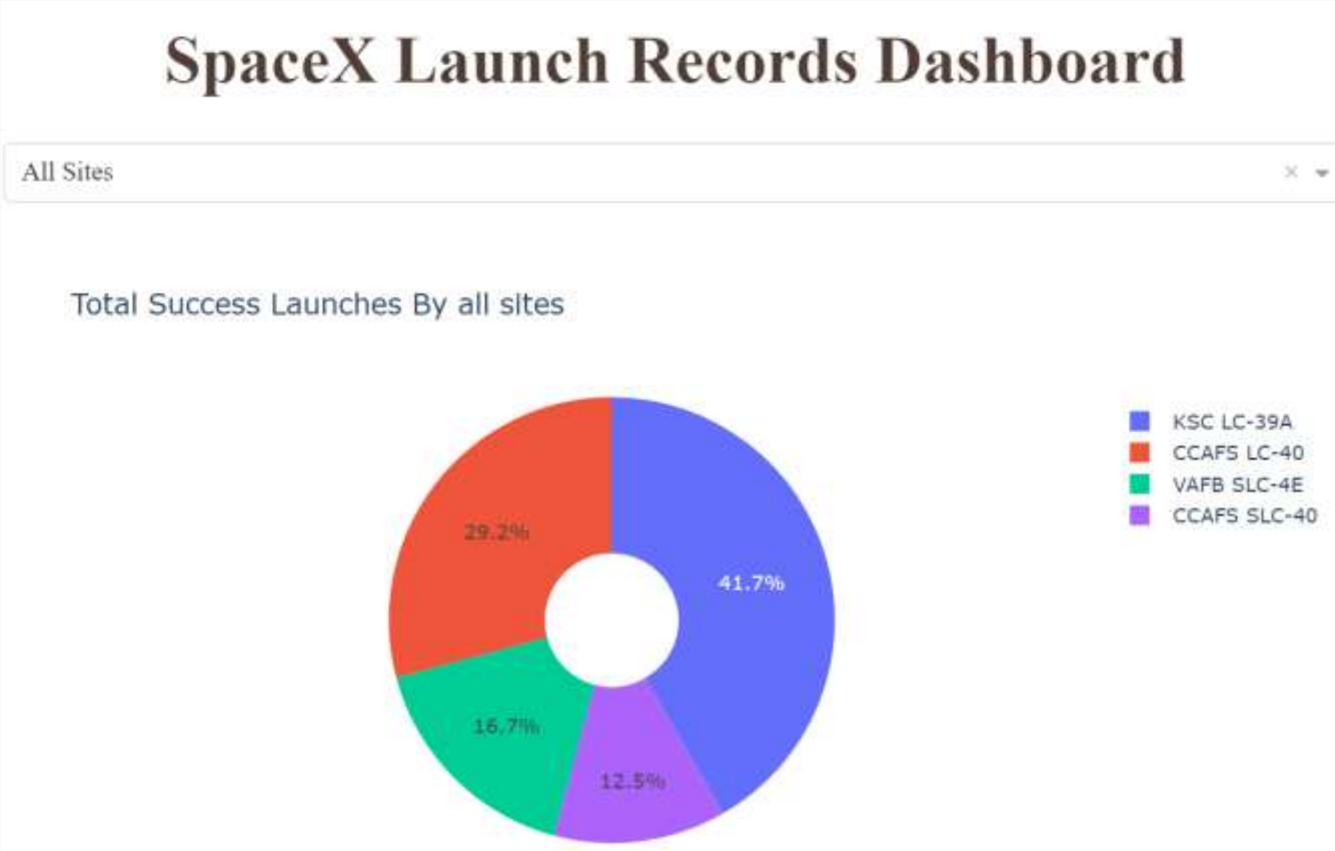
- The map on left shows the proximity of highway, railway, coastal line and Florida city from CCAFS SLC-40 launch site from close
- We can observe that the launch site is close to coastal line first and then railway
- Second image show shows the same but at zoom out level

Section 4

Build a Dashboard with Plotly Dash

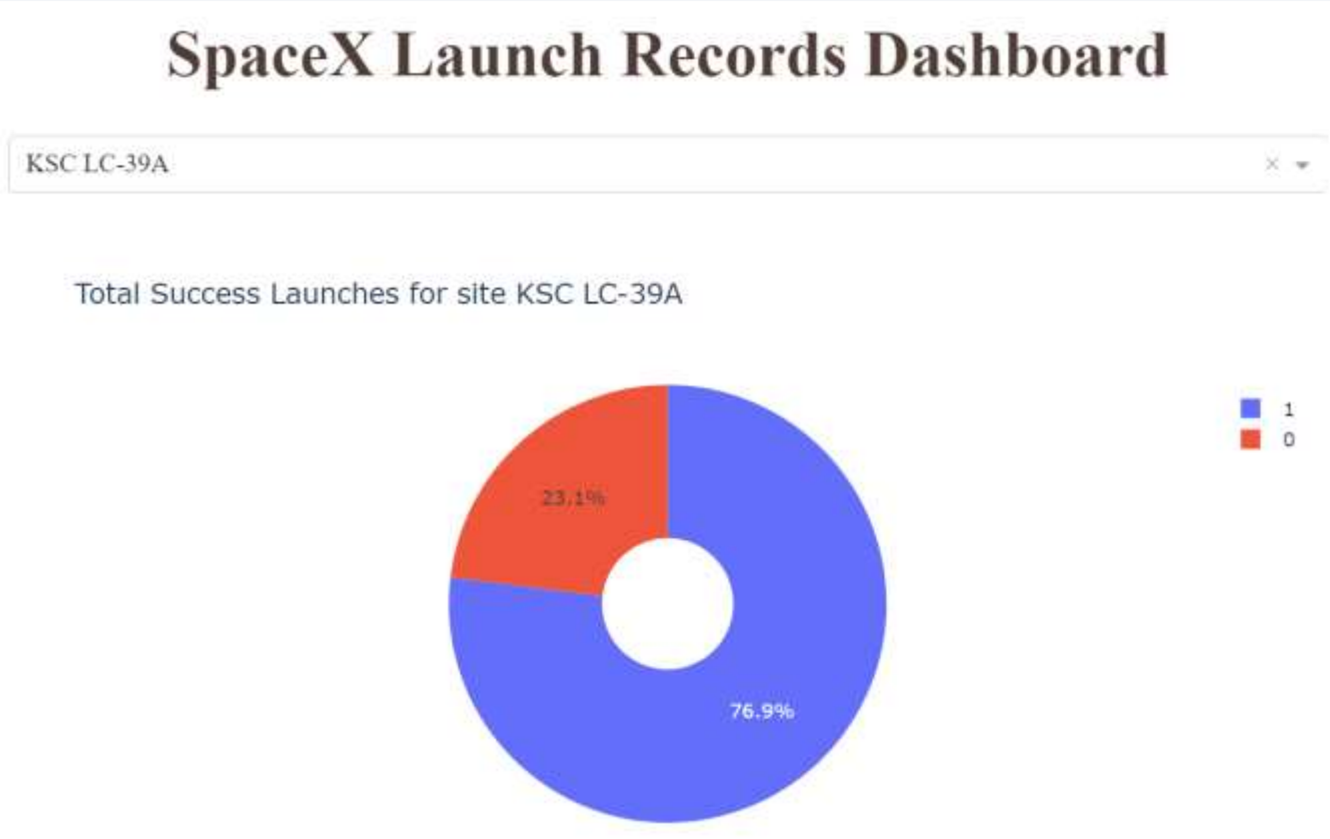
Dashboard - Success rate of all launch sites

- Pie chart shows the success rate at each site.
- We can see KSC LC-39A has highest success rate i.e 41.7%

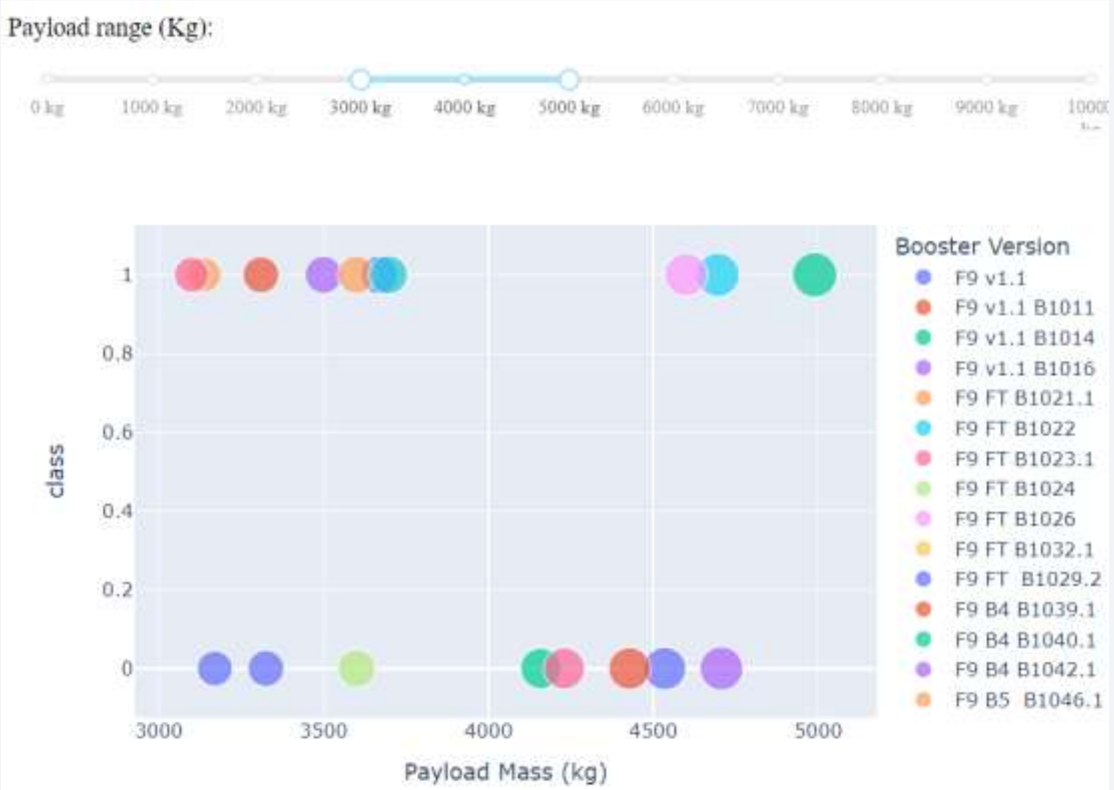


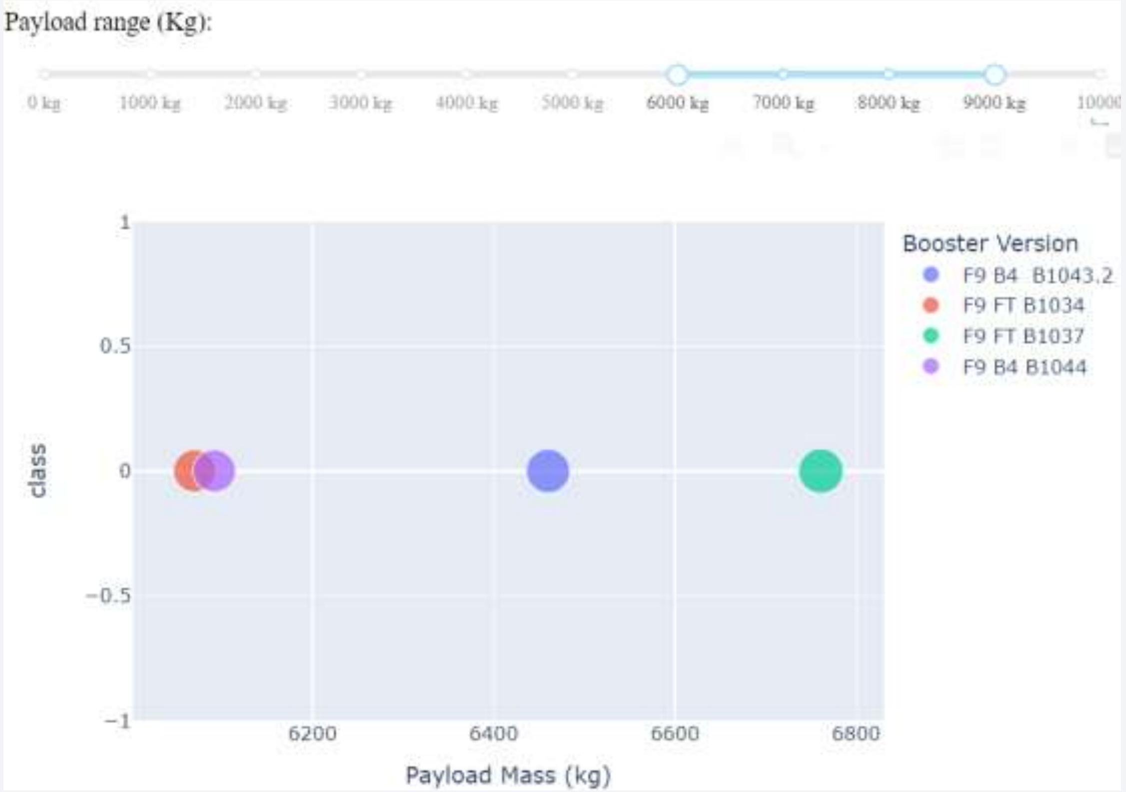
Dashboard - Success launches at KSC LC-39A

- As we know from pie chart in previous slide, KSC LC-39A have highest successful launches so we plotted one more pie chart that shows the success to failure ratio at this site
- KSC LC-39A has success ratio of 76.9%



Dashboard – Payload vs. Launch Outcome for different booster version





If payload is between 3000 to 5000 kgs, the success ratio is high but when the payload is high i.e ranging between 6000 to 9000 kgs, there are no successful cases.

Dashboard - Payload Payload vs. Launch Outcome

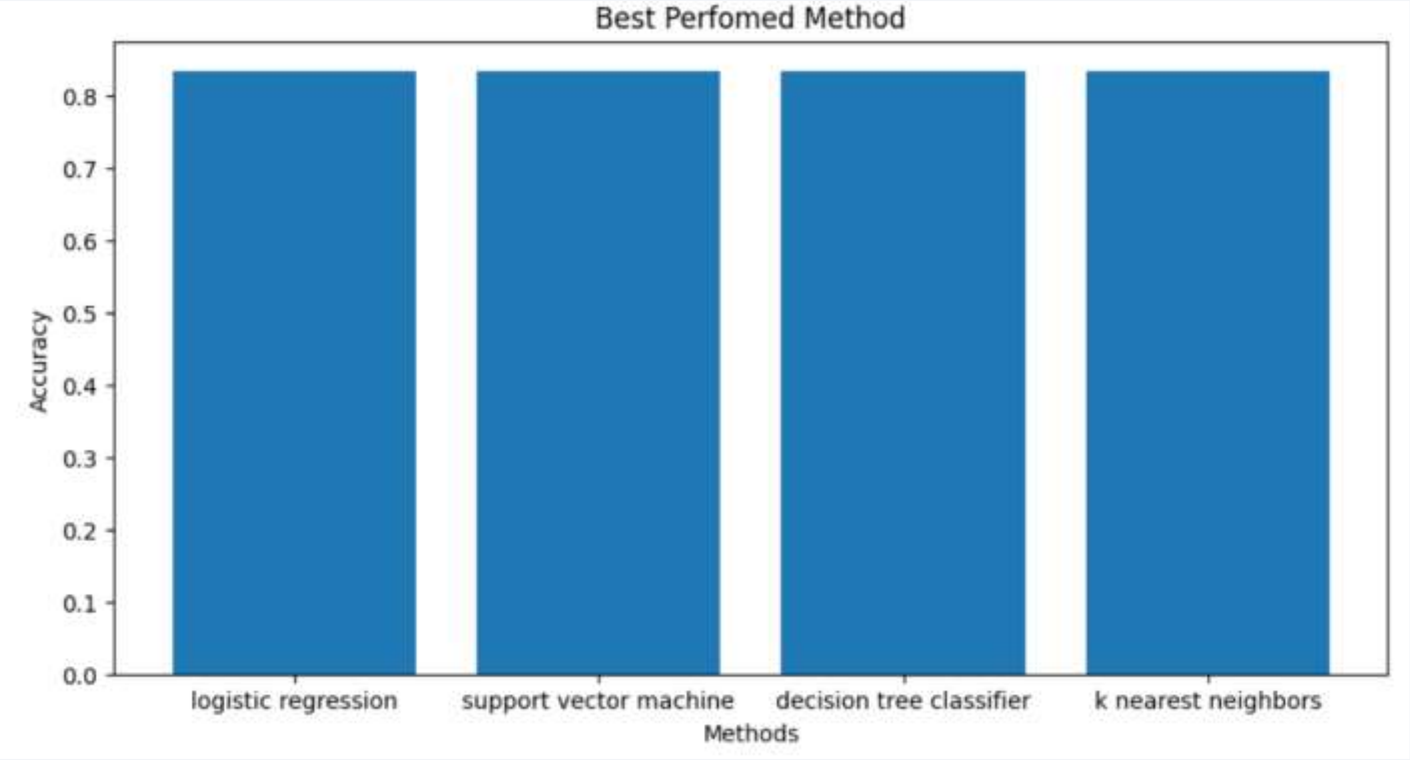
If we select the complete range for slider, we could see payload mass between 2000 to 5500 kgs resulted in more success ratio compared to higher payload

Section 5

Predictive Analysis (Classification)

Classification Accuracy

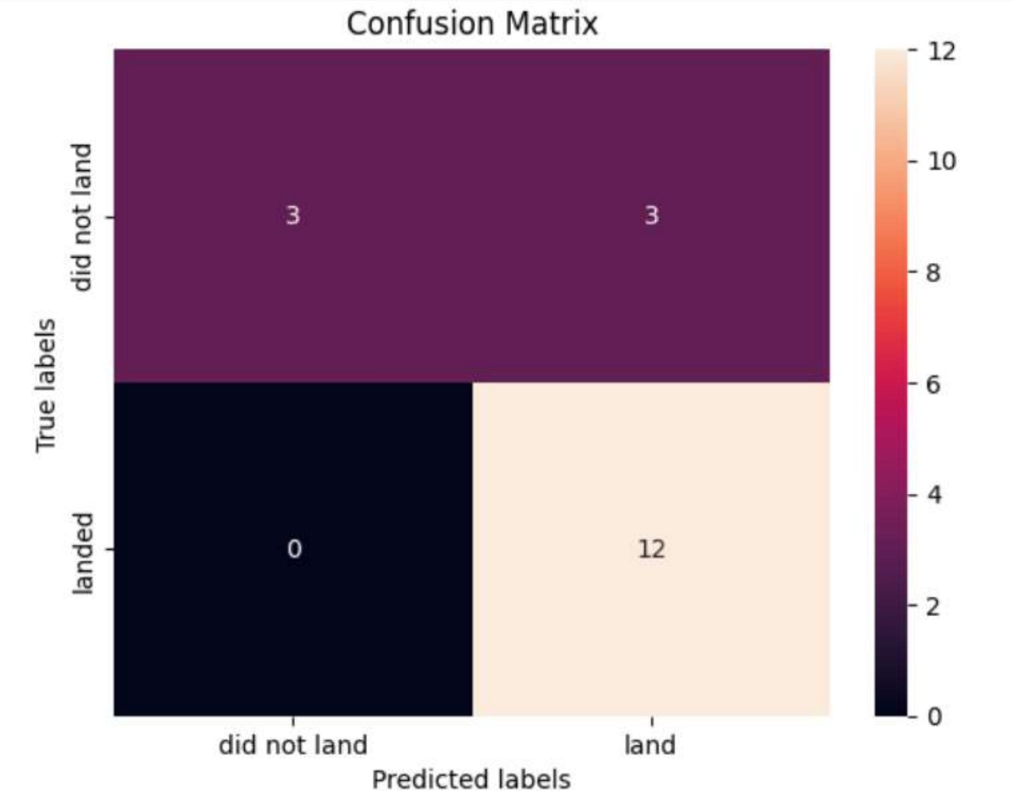
- Bar plot show casing the accuracy for different classification algorithms
- All model have the same classification accuracy i.e 83%



Confusion Matrix

The confusion matrix is same for all models as accuracy is same.

- The models predicted 12 successful landings when the true label was successful landing
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives)



Conclusions

- We collected data from various sources and performed EDA
- 66% of the landing outcomes are successful across from 4 launching sites
- The success rate increased from the year 2013
- We then created a machine learning classification model to predict the success of landing that resulted in an accuracy of 83%
- In order to find best model and increase in accuracy of the model, we need more data
- Allon Mask of SpaceY can use this model to predict outcome of a launch whether it is successful or not before launch to determine to move forward with the launch or not

Appendix

- GitHub repository for the SpaceX project

  [Applied Data Science Capstone Project]

- Special thanks to all the instructors for providing these course

Thank you!