Data Science Capstone Project

30, DECEMBER 2023

Outline

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

Summary of methodologies

- - Data collection
- - Data wrangling
- - Exploratory Data Analysis with Data Visualization
- - Exploratory Data Analysis with SQL
- - Building an interactive map with Folium
- - Building a Dashboard with Plotly Dash
- - Predictive analysis (Classification)

Summary of all results

- - Exploratory Data Analysis results
- - Interactive analytics demo in screenshots
- - Predictive analysis results

Introduction

Project background and context

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data is collected using get request from SpaceX API.
- The requested data contain information of rocket type, lunch site , mission success and lunch date etc.
- The data is cleaned and structured into a Data frame using pandas for further analysis, contain data only about falcon 9 rocket.
- Web scraping a Wikipedia page for more historical data of Falcon 9.

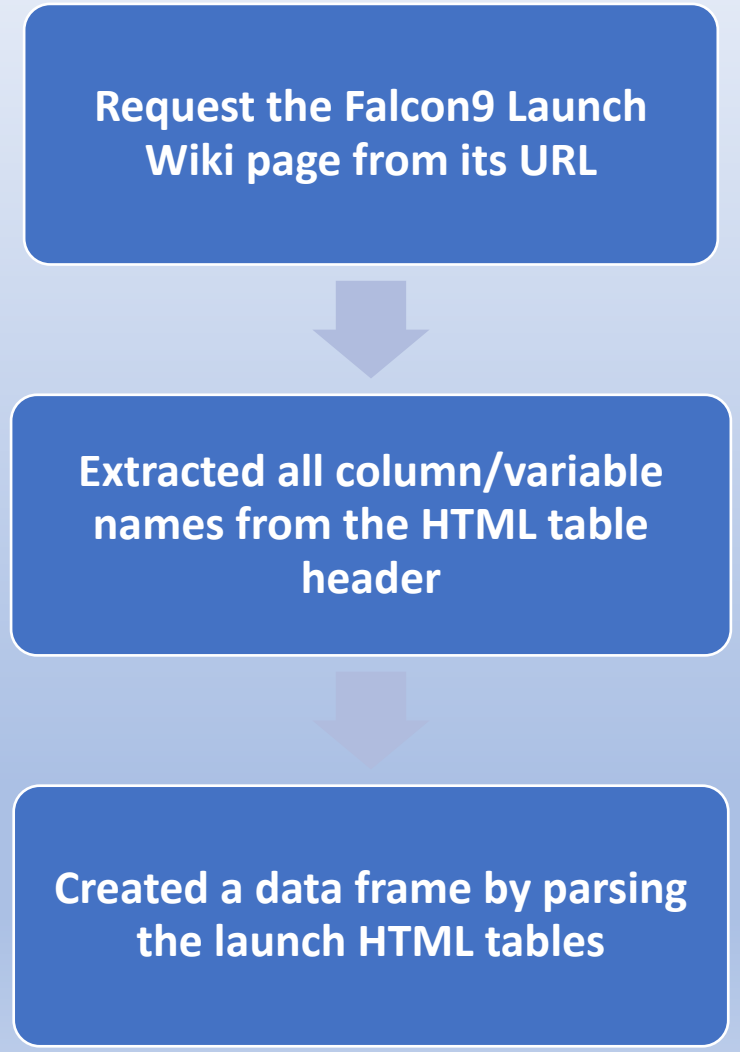
Data Collection – SpaceX API

- Flowcharts of the process

- GitHub URL of SpaceX API

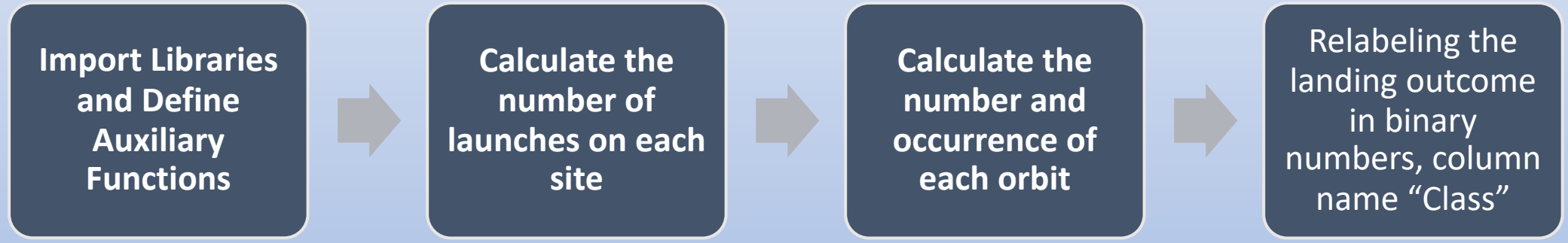
Data Collection - Scraping

- Wikipedia page URL.
- GitHub URL of web scraping notebook



Data Wrangling

- Variables which can predict the mission outcome is selected and saved as CSV file.



- The mean of column ”class” is grater than 0.5 meaning the success rate is greater compared to unsuccessful attempts.
- GitHub URL of data wrangling.

EDA with Data Visualization

- Charts plotted include : Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type,Payload Mass vs Orbit Type and Success Rate Yearly Trend.
- The scatter plot of flight number vs lunch site confirms the most used lunch site for falcon 9 rocket is CCAFS SLC 40.
- The bar chart of Orbit Type vs. Success Rate shows VLEO has the highest success rate.
- The line chart of orbit type and success showcase a continues increase in success rate over the period from 2013 to 2020.
- GitHub URL of EDA with data visualization notebook

EDA with SQL

- Important SQL queries include: Average payload mass carried by booster version F9 v1.1, First successful landing outcome in ground pad, Total number of successful and failure mission outcomes, booster versions which have carried the maximum payload mass, Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

- GitHub URL of EDA with SQL notebook.

Build an Interactive Map with Folium

Markers of all Launch Sites:

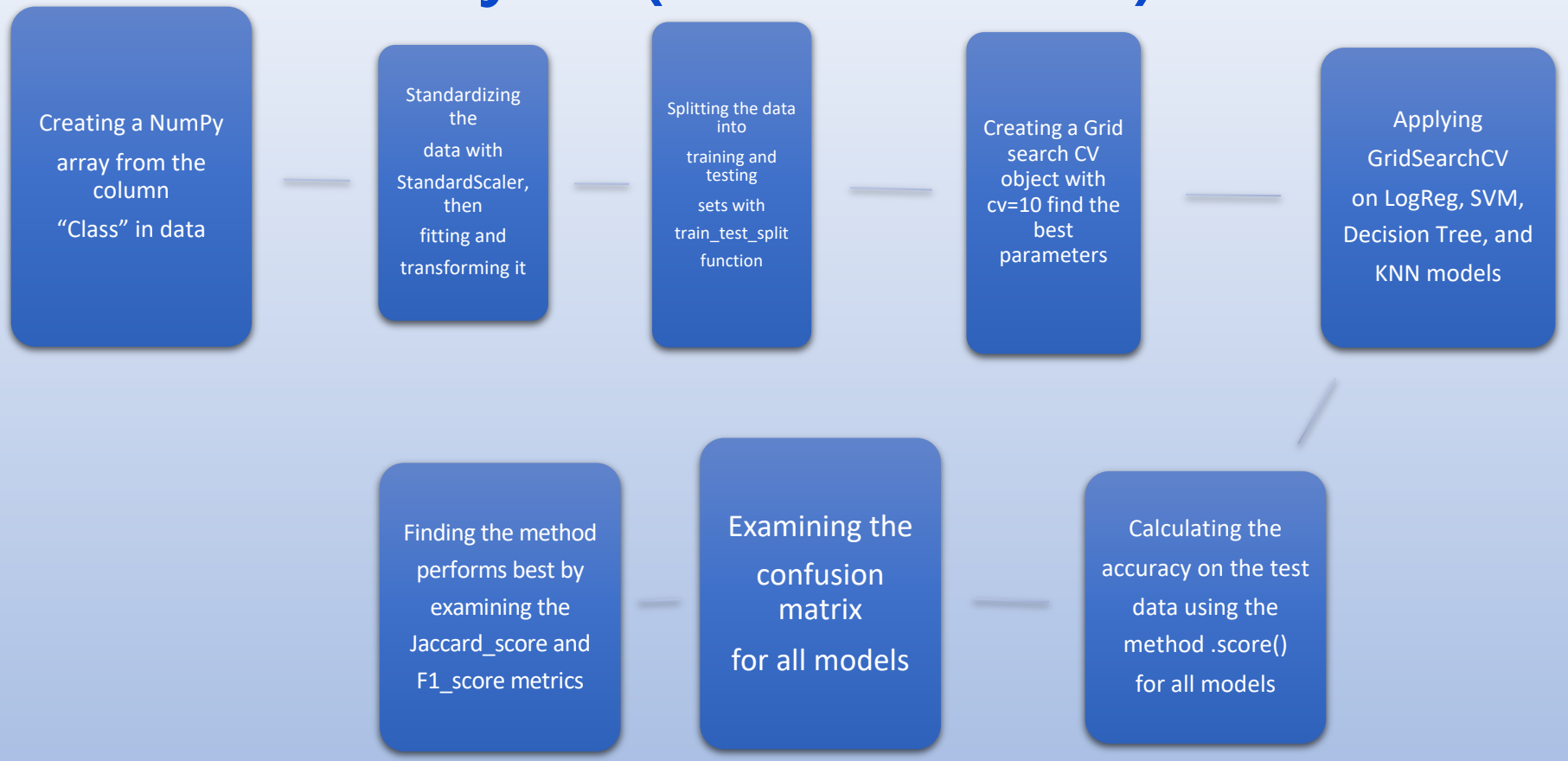
- Added Markers with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added 4 Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and Coasts.
- Within each marker, objects are created to visualize the succuss(green popup) and fail (red popup) mission to answer does lunch site location effect mission success.
- 3 distance markers are added for lunch site CCAFS SLC-40 to visualize the site distance from ocean, railway track and highway.

- GitHub URL of map with Folium map

Build a Dashboard with Plotly Dash

- Pie chart is generated with 5 inputs which include a pie chart representing total number of successful mission for each site and 4 other inputs generating pie chart, representing success to failer ratio.
- A scatter plot showcasing successful and failed mission with respect to payload mass on x-axis, where dotes are colored for respective lunch site.
- These plot help understand does payload and lunch site together has any correlation with mission output.
- GitHub URL of Plotly Dash lab.

Predictive Analysis (Classification)



- GitHub URL of predictive analysis lab

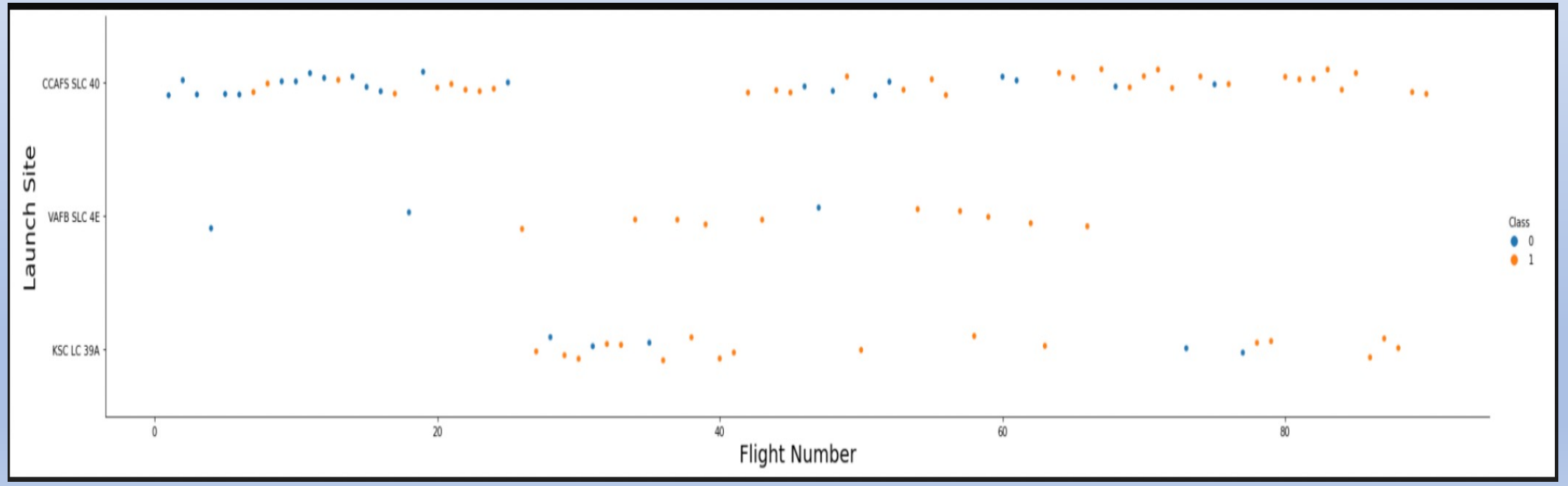
Results

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

EDA with Visualization

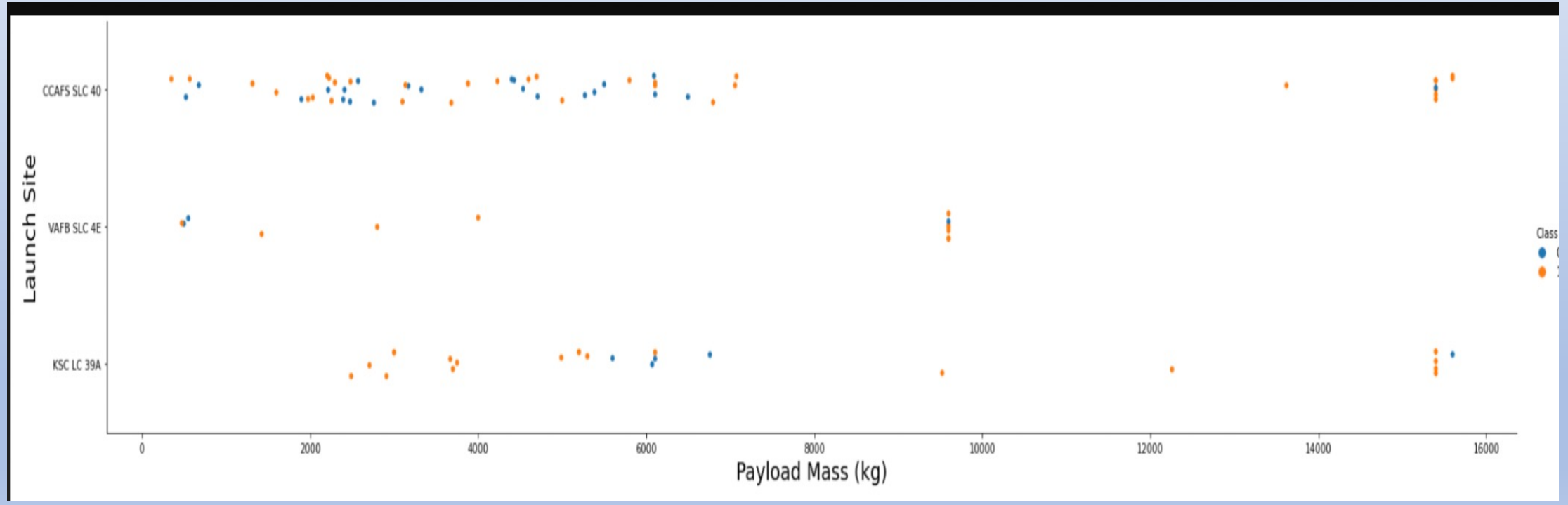
Flight Number vs. Launch Site

- The earliest flights all failed while the latest flights all succeeded.
- VAFB SLC 4E and KSC LC 39A have higher success rates.



Payload vs. Launch Site

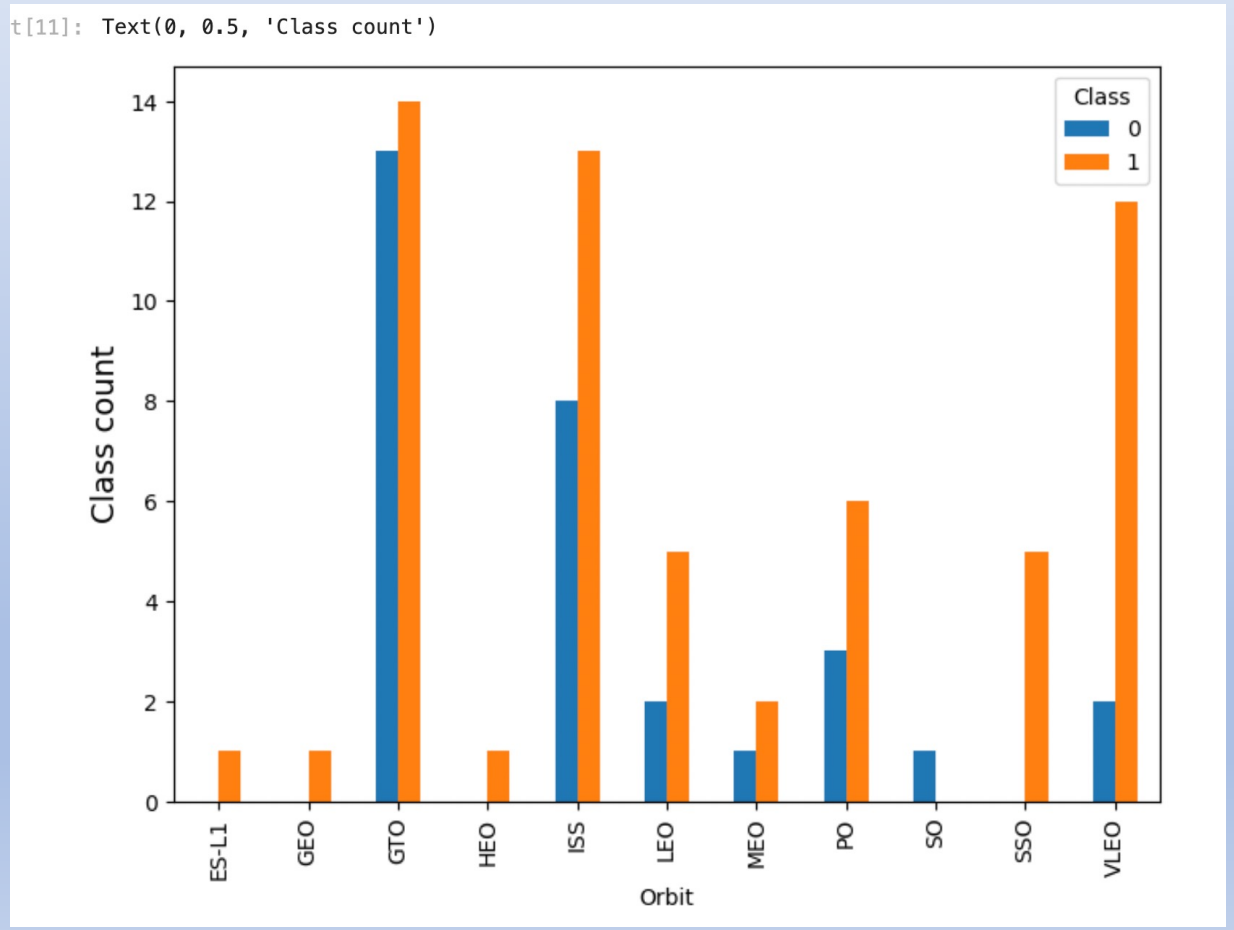
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 6000 kg were successful.



Success Rate vs. Orbit Type

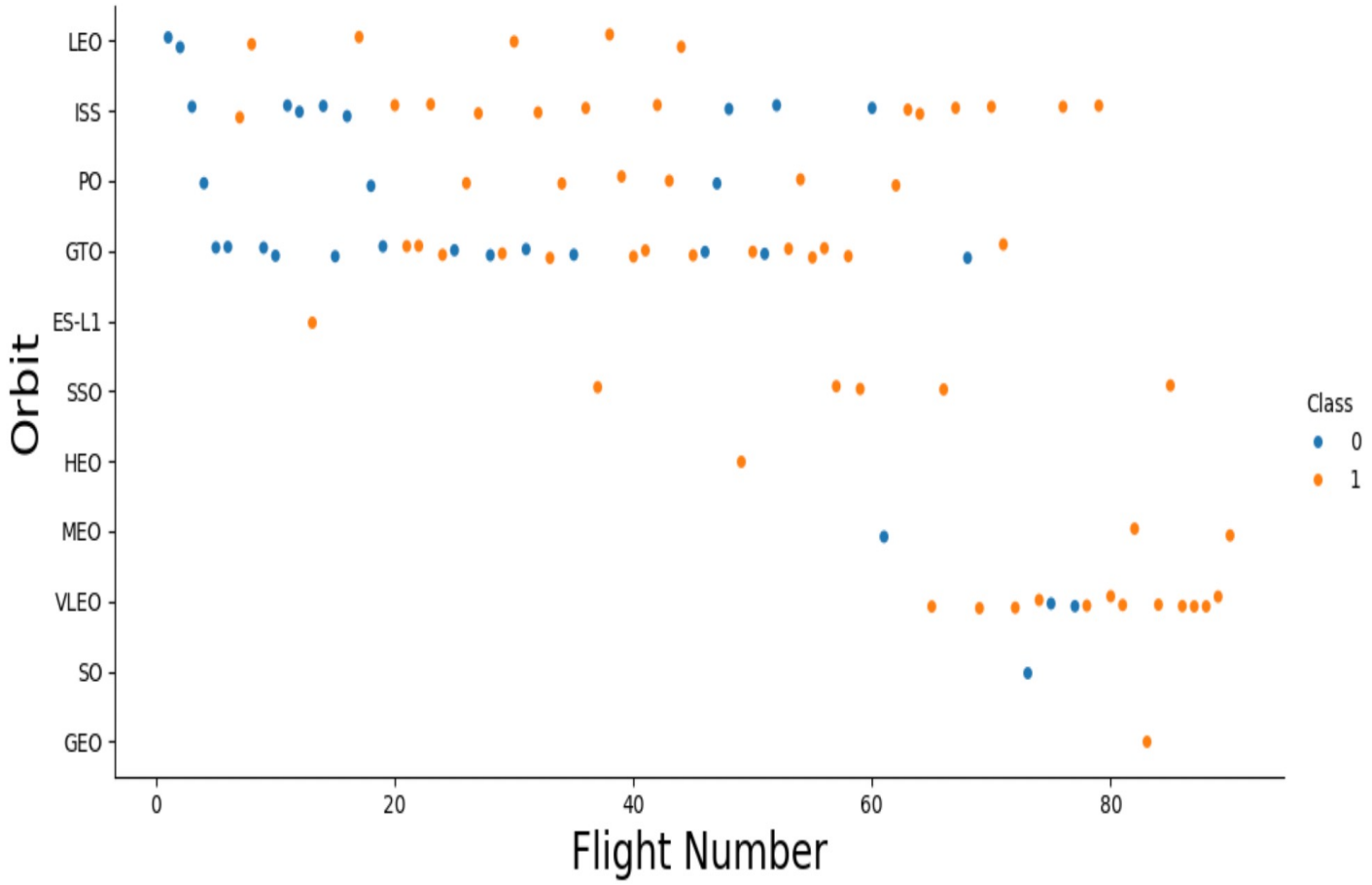
VLEO has the highest success rate.

The highest attempts done is for orbit GTO, having lowest success rate as well.



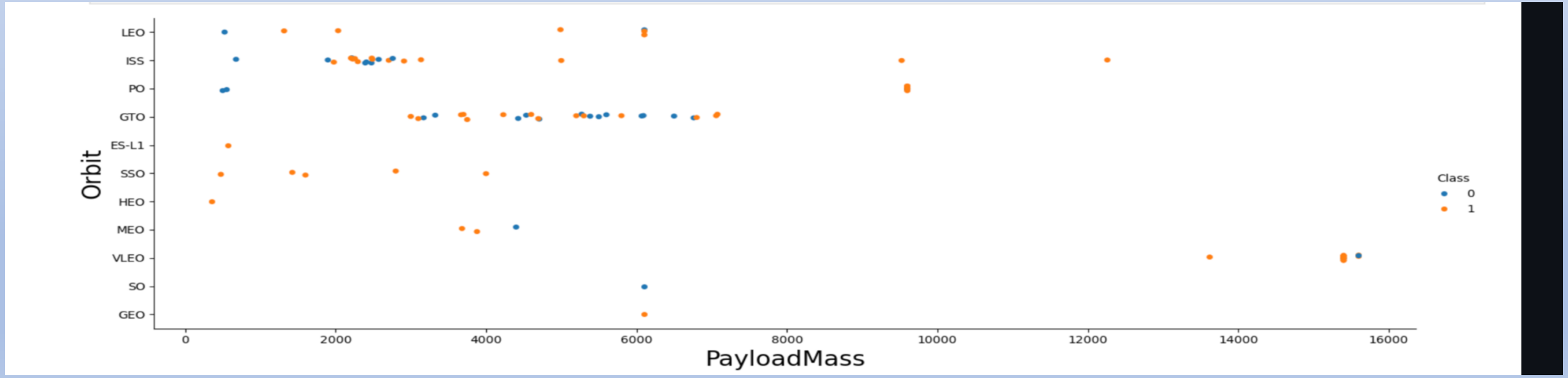
Flight Number vs. Orbit Type

- Mostly initial attempts for an orbit are unsuccessful.



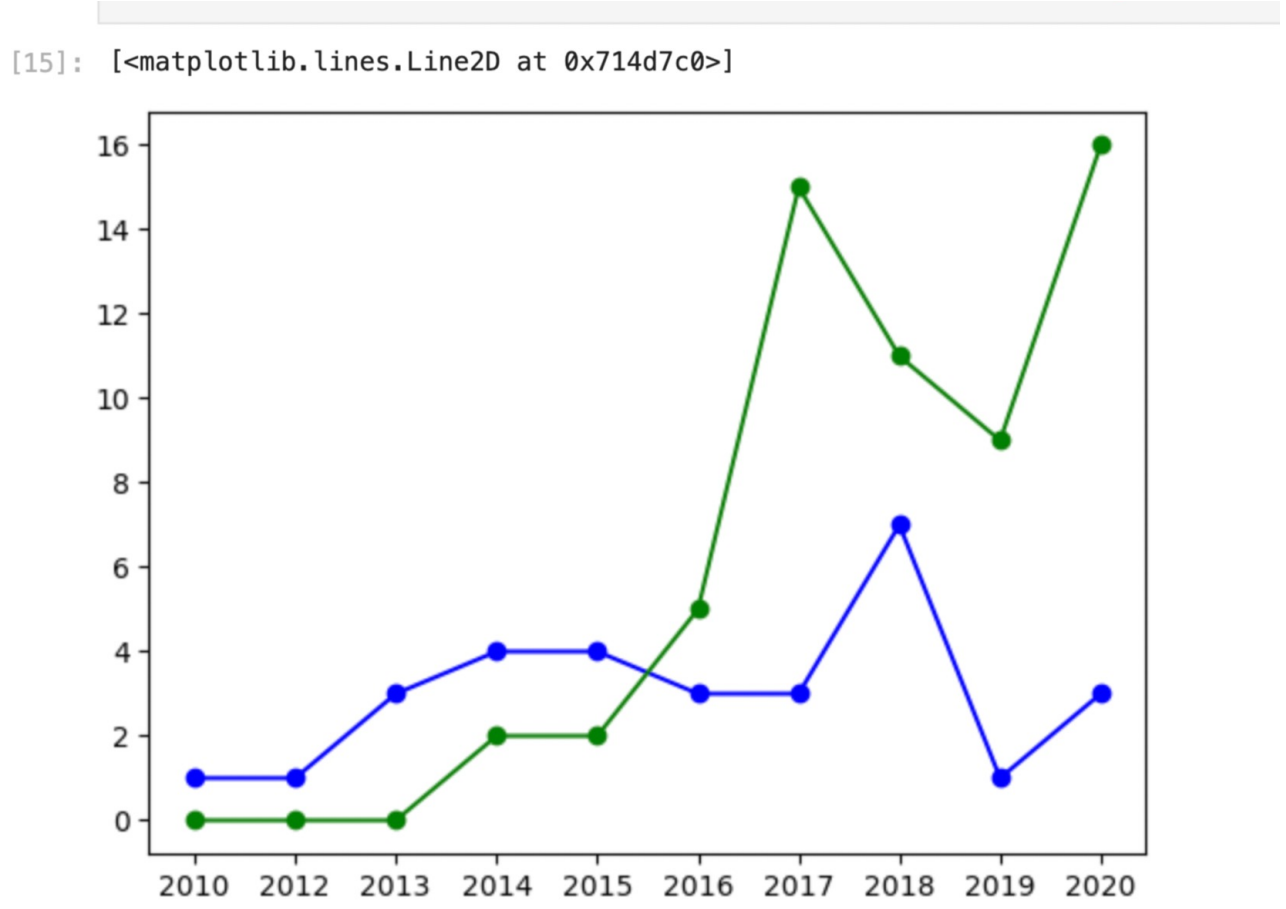
Payload vs. Orbit Type

- Payload Mass not has no strong influence in mission success.
- Most attempts for GTO has payload mass between 2000kg to 3000kg.
- Most attempts for ISS has payload mass between 3000kg to 7500kg.



Launch Success Yearly Trend

- Green line show case total count of successful attempts and blue line representing failed mission count.
- Over time the succuss rate of mission improved surpassing unsuccessful attempts count in year 2015.



EDA with SQL

All Launch Site Names

- Displaying the names of the unique launch sites in the space mission.

Done.

Out[8]:

Landing_Outcome
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt

Launch Site Names Begin with 'CCA'

- Displaying 5 records where launch sites begin with the string 'CCA'.

Done.

Out[15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

```
In [10]: %sql select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" like 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[10]: sum("PAYLOAD_MASS__KG_")
                            45596
```

Average Payload Mass by F9 v1.1

- Displaying the Average payload mass carried by booster version F9.

```
In [11]: %sql select avg("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" like 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
Out[11]: avg("PAYLOAD_MASS__KG_")
         2534.6666666666665
```

First Successful Ground Landing Date

- Listing the date when the first successful landing outcome in ground pad was achieved.

```
In [12]: %sql select min("Date"),"Landing_Outcome" from SPACEXTBL WHERE "Landing_Outcome"='Success (ground pad)' limit 1;
 * sqlite:///my_data1.db
Done.
```

Out[12]:

min("Date")	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

Total Number of Successful and Failure Mission Outcomes

- Listing the total number of successful and failure mission outcomes.

```
In [14]: %sql select "Mission_Outcome",COUNT("Mission_Outcome")from SPACEXTBL group by "Mission_Outcome" having "Mission_O
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Mission_Outcome	COUNT("Mission_Outcome")
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass.

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

- List of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
[43]: %sql select substr(Date, 6,2),"Landing_Outcome","Booster_Version","Launch_Site" from SPACEXTBL where "Landing_Out
 * sqlite:///my_data1.db
Done.
```

t[43]:

substr(Date, 6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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

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

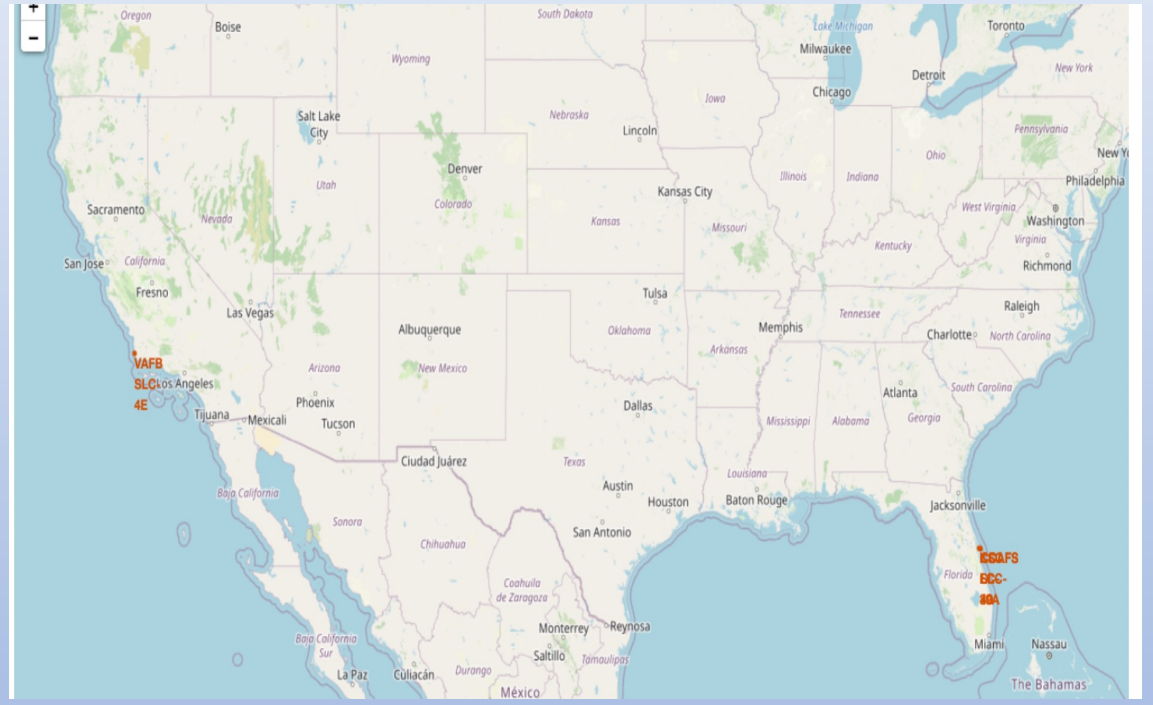
74]:

count_landing_type	Landing_Outcome	Date
10	No attempt	2012-05-22
5	Success (drone ship)	2016-04-08
5	Failure (drone ship)	2015-01-10
3	Success (ground pad)	2015-12-22
3	Controlled (ocean)	2014-04-18
2	Uncontrolled (ocean)	2013-09-29
1	Precluded (drone ship)	2015-06-28
1	Failure (parachute)	2010-12-08

Interactive map with Folium

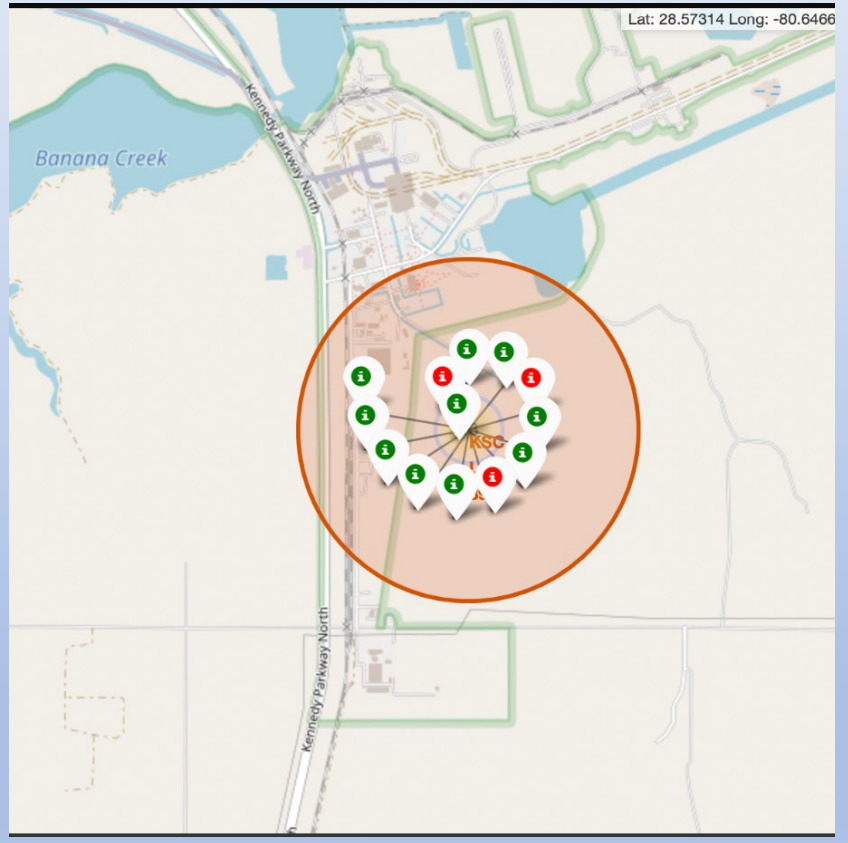
All launch sites' location markers on a global map

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.
- All lunch site are on the equator line due the fact that land is moving faster at equator dur to its spin, giving a fare Advantage to to push the rocket at outer space also reducing the fuel consumption.



Colour-labeled launch records on the map

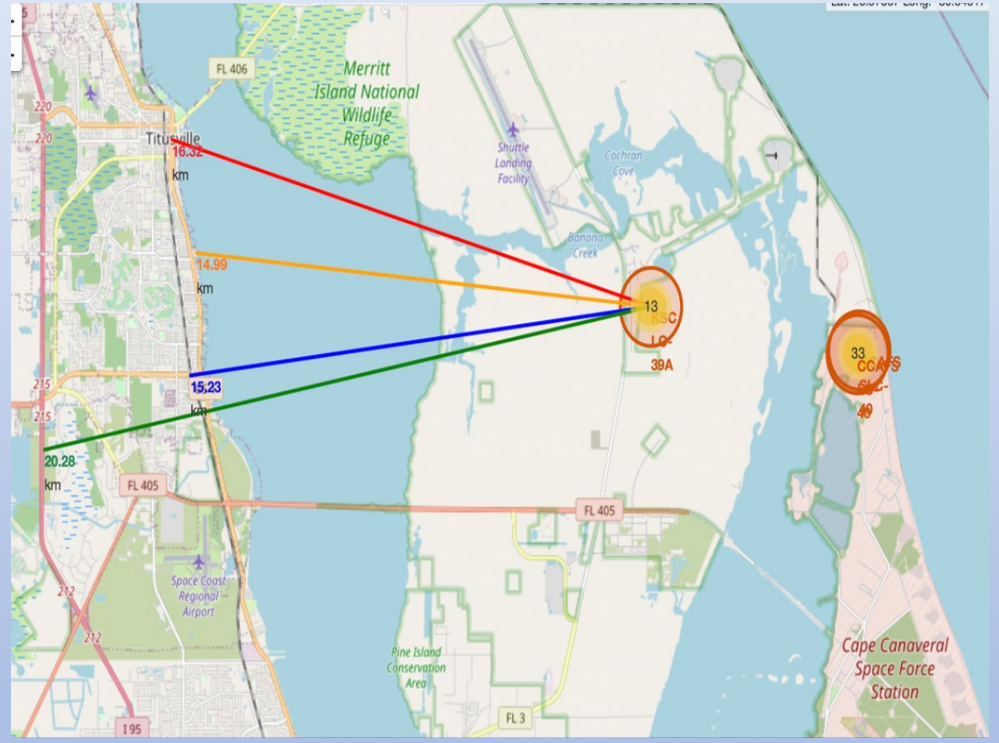
- The Green marker represents the successful attempts and Red are unsuccessful.
- Launch Site KSC LC-39A has a very high Success Rate as shown in the map screenshot.



Distance from the launch site KSC LC-39A to its proximities

- Replace <Folium map screenshot 3> title with an appropriate title

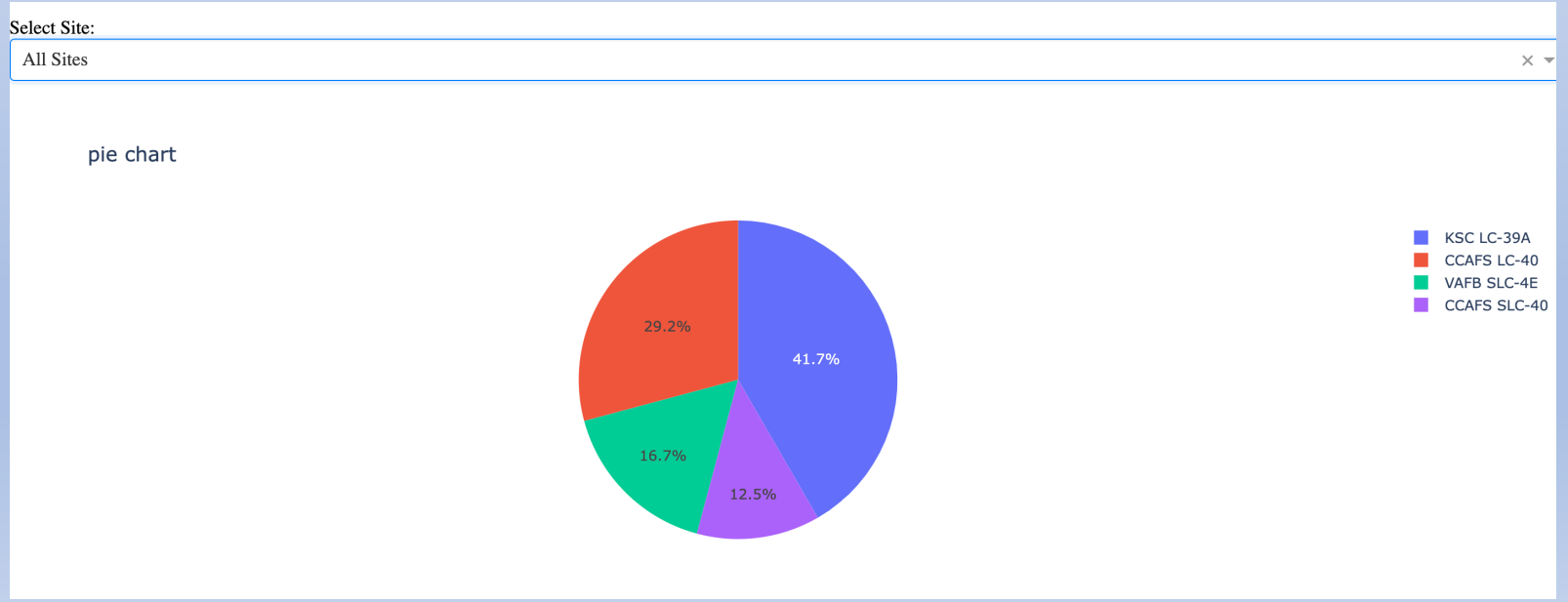
- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)



Build a Dashboard with Ploty Dash

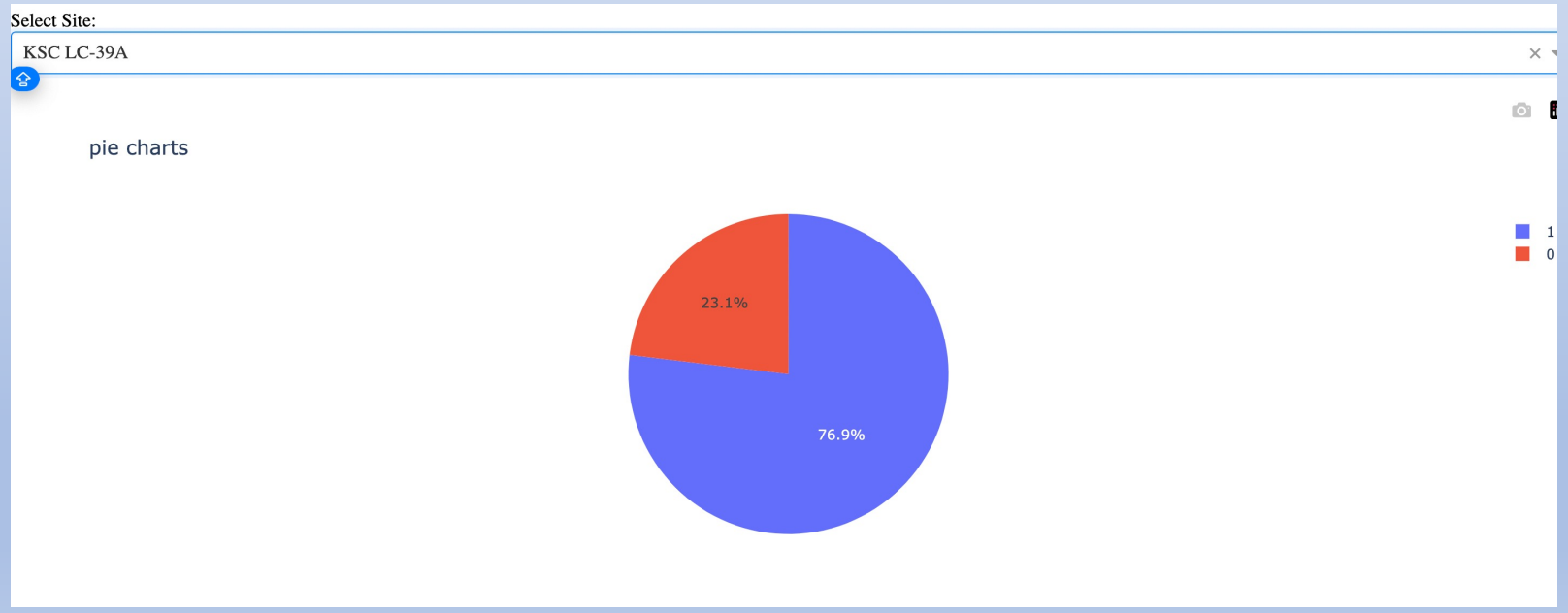
Chart of percentage of successful mission of all sites.

- This is a dashboard screenshot of pie chart explaining what percentage of each site has maximum chances of a successful mission.
- 2 sites that stand out are KSC-LC-39A and CCAFS-LC-40.



Launch site with highest launch success ratio

- When selecting KSCLC-39A on dash it is clear there are 23% chances that a mission fail.



Payload Mass vs. Launch Outcome for all sites

- When selected all sites with payload range from 0 to 10000 this scatter plot shows the highest chances of mission success is between 2000 to 5500kg payload mass, WHERE “FT Booster Version Category” has highest success rate among other category.

Predictive analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we cannot confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test Set

t[34]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.533333
F1_Score	0.888889	0.888889	0.888889	0.695652
Accuracy	0.833333	0.833333	0.833333	0.611111

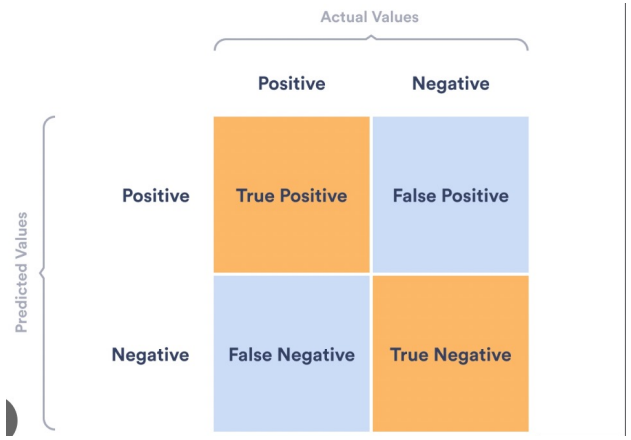
Scores and Accuracy of the Entire Data Set

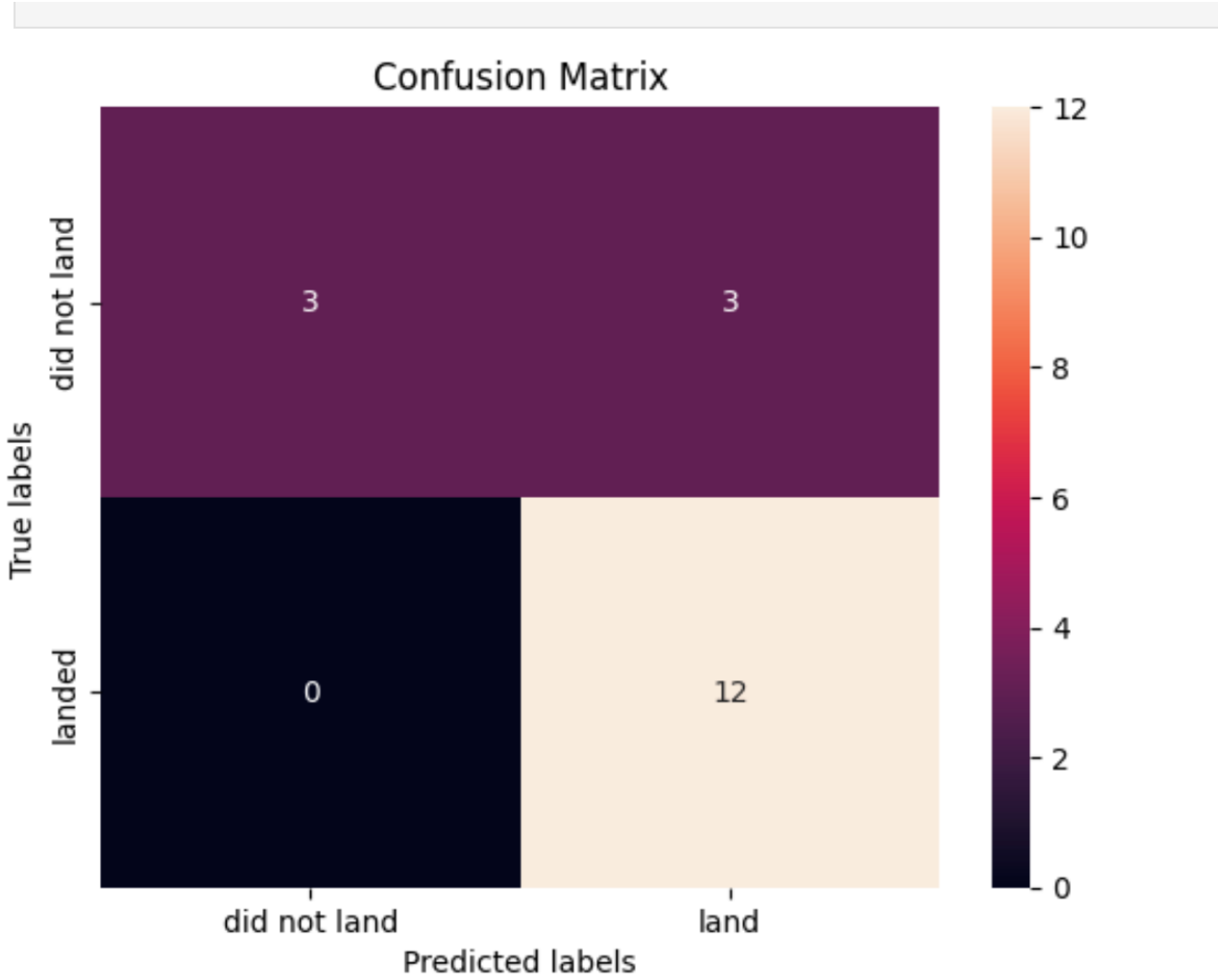
35]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.794521	0.830986	0.819444	0.718310
F1_Score	0.885496	0.907692	0.900763	0.836066
Accuracy	0.833333	0.866667	0.855556	0.777778

Confusion Matrix

- Examining the confusion matrix, we see that Decision Tree can distinguish between the different classes. We see that the major problem is false positives.





Conclusions

- In predictive analysis Decision tree outperform all other models and can be deployed to a SpaceX mission outcome.
- Site of launch pad and payload mass has considerable impacts on a space mission success.
- KSC-LC-39A has the highest success rate.
- Payload mass between 2000 to 5500kg with FT Booster Version Category positively influence the mission lunch success.
- With time the SpaceX success mission have increased especially after 2015.
- Most mission lunch site are near the line of Equator.

Appendix

- GitHub repository for all the notebooks and source code.
- Special thanks to IBM Data science coursera team.

Thank You