Data Science Capstone project

Pierre Sookiew

August 2021

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)

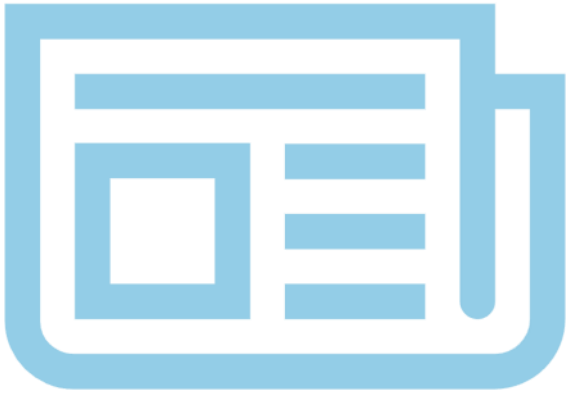
- **Summary of all results**
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- We predict if the Falcon 9 first stage will land successfully. 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.

- What influences if the rocket will land successfully?
- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

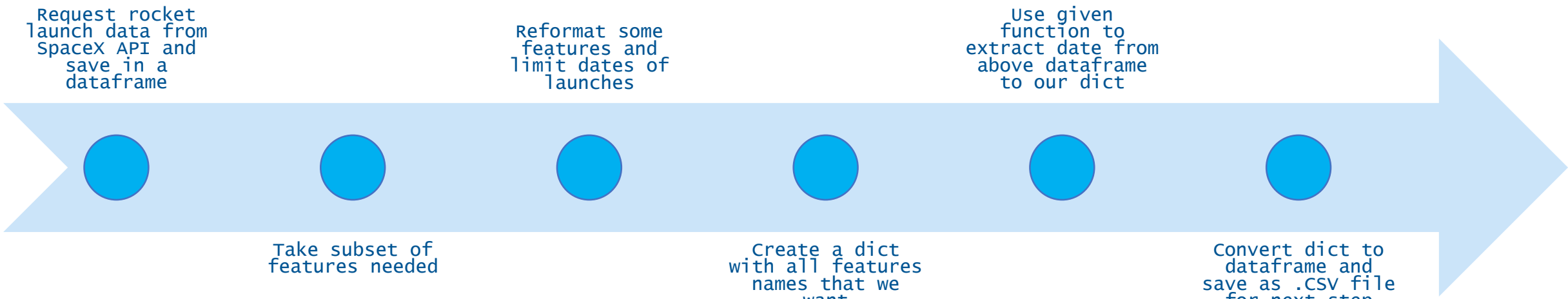
Methodology



- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping
- Perform data wrangling
  - Transforming, cleaning and mapping data
- 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

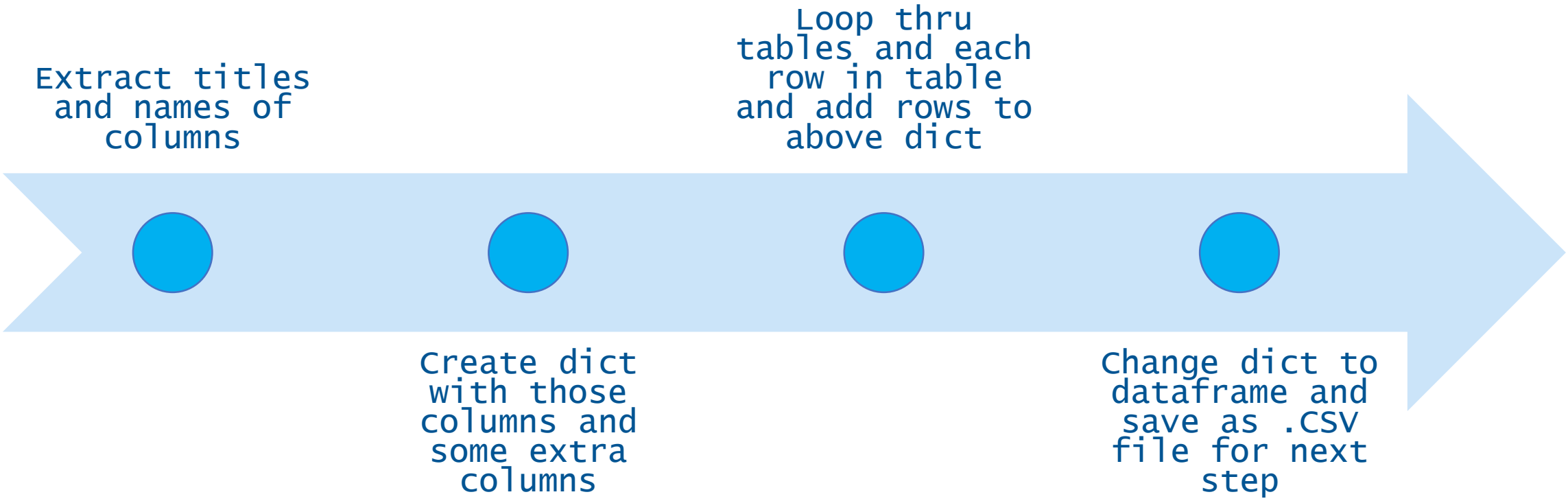
Methodology

Data collection – SpaceX API



https://github.com/peterpouf/python_data_science/blob/main/Data%20Collection%20API.ipynb

Data collection – Web scraping



https://github.com/peterpouf/python_data_science/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data wrangling

Check for number of nulls and types of each column

Count values of each launch site and types of Orbit for better understanding of our dataset

Create landing outcome label for Outcome column with 0 for failure and 1 for success mission

Save datafame as .CSV file for next step

https://github.com/peterpouf/python_data_science/blob/main/EDA.ipynb

EDA with data visualization

- Summarize what charts were plotted and why used those charts
  - Scatter plot - To see the correlation of two parameters on a dependent variable via color coding
  - Bar chart – To visualize effect of one categorical parameter on a continuous one
  - Line graph – To observe trend of dependent variable based on an independent one.

https://github.com/peterpouf/python_data_science/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Summarize performed SQL queries using bullet points
  - ‘Select’ queries to display variety of combination of columns
  - Use of ‘where’ and ‘like’ commands to use filters
  - Use of subquery to allow complex filters
  - Use of mathematical operations like AVG(), MIN(), COUNT(), etc.
  - Use of ‘Order BY’, ‘GROUP BY’, ‘RANK()’ functions to apply pivot like functions of table.

Build an interactive map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Circle Markers created for each launch site
  - Marker clusters created for multiple markers generated on each site
  - Color coding of markers
  - Lines
  - Labels
- Explain why you added those objects
  - Circle Markers – to highlight area of site
  - Marker clusters – to increase visibility and comprehension of markers at a point
  - Color coding of markers to indicate success and failure in first stage landing for each site in our dataset
  - Lines – to show distance from closest railway or coastline to launchsite
  - Labels – to increase readability of maps

https://github.com/peterpouf/python_data_science/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Dropdown box
  - Pie chart
  - Range Slider
  - Scatter Plot

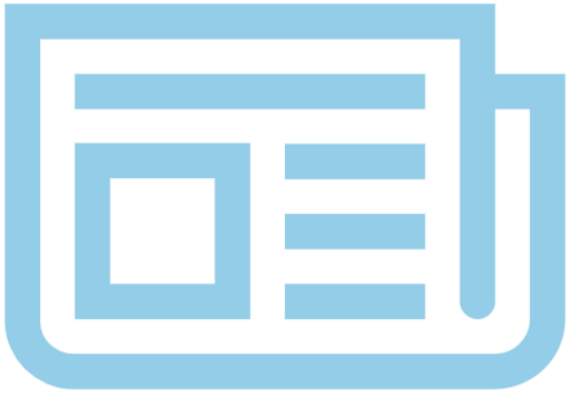
- Explain why you added those plots and interactions
  - Dropdown box – to select the site or sites for analysis
  - Pie chart – to provide view of success rate at each site
  - Range Slider – to choose the payload mass range for analysis
  - Scatter Plot – to depict the relation between payload mass and launch site on the success rate of landing

Predictive analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Preprocess the data using StandardScaler()
  - Split dataset into training and test data
  - Create variety of models, fit them with training data, predict using test data.
  - Evaluate score of each model and pick model with highest accuracy

https://github.com/peterpouf/python_data_science/blob/main/Machine%20Learning%20Prediction.ipynb

Results

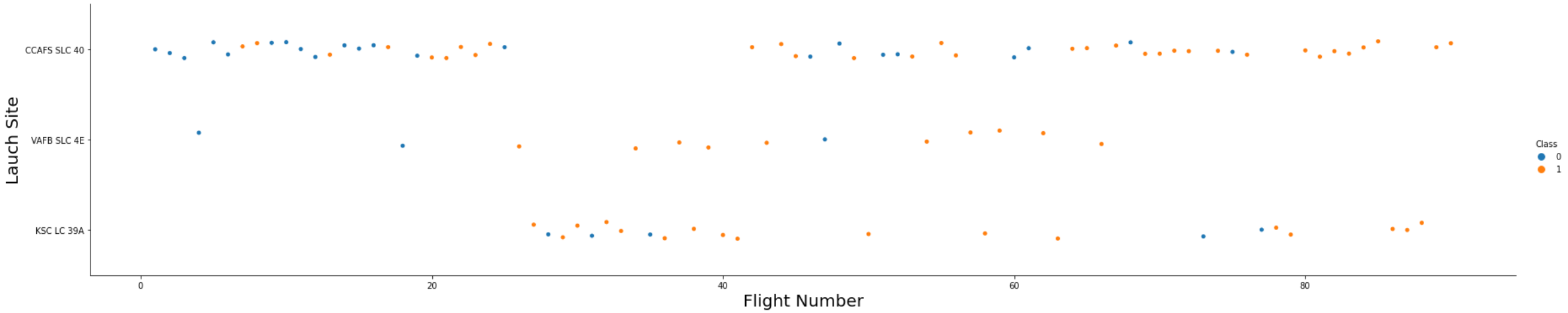


- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

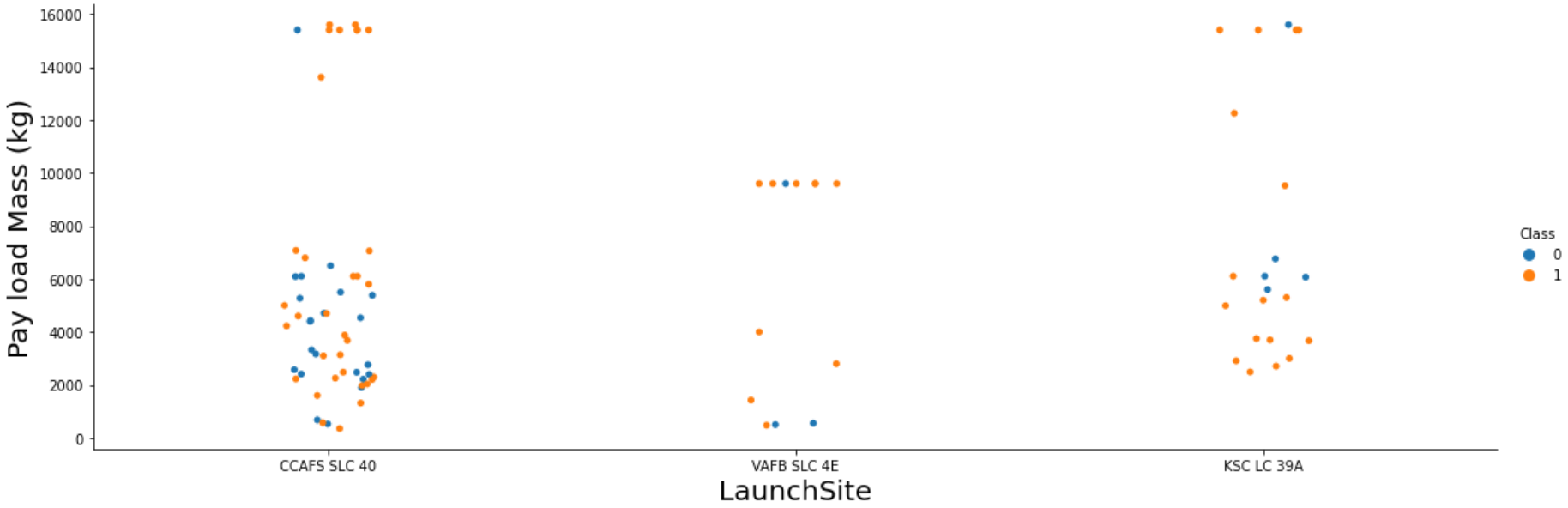
Flight Number vs. Launch Site

- As flight numbers have increased, chances of success has increased.
- Further, CCAFS SLC 40 is predominantly used as the launch site, so it has a lower success rate, but the other two sites have a higher success ratio due to relatively lower flight numbers.



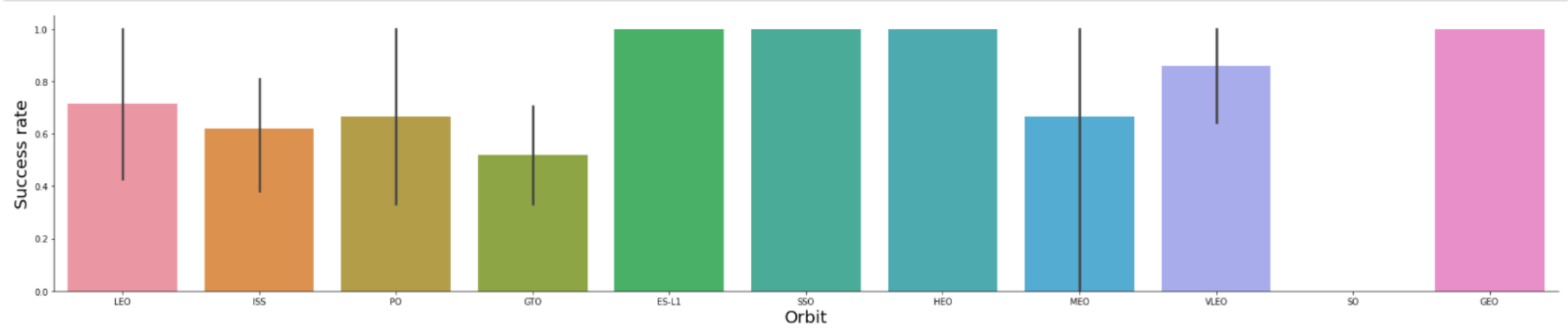
Payload vs. Launch Site

- Success rate seems to increase with higher payload mass; meanwhile CCAFS SLC 40 & KSCLC 39A outperform VAFB SLC4E at higher payload mass points



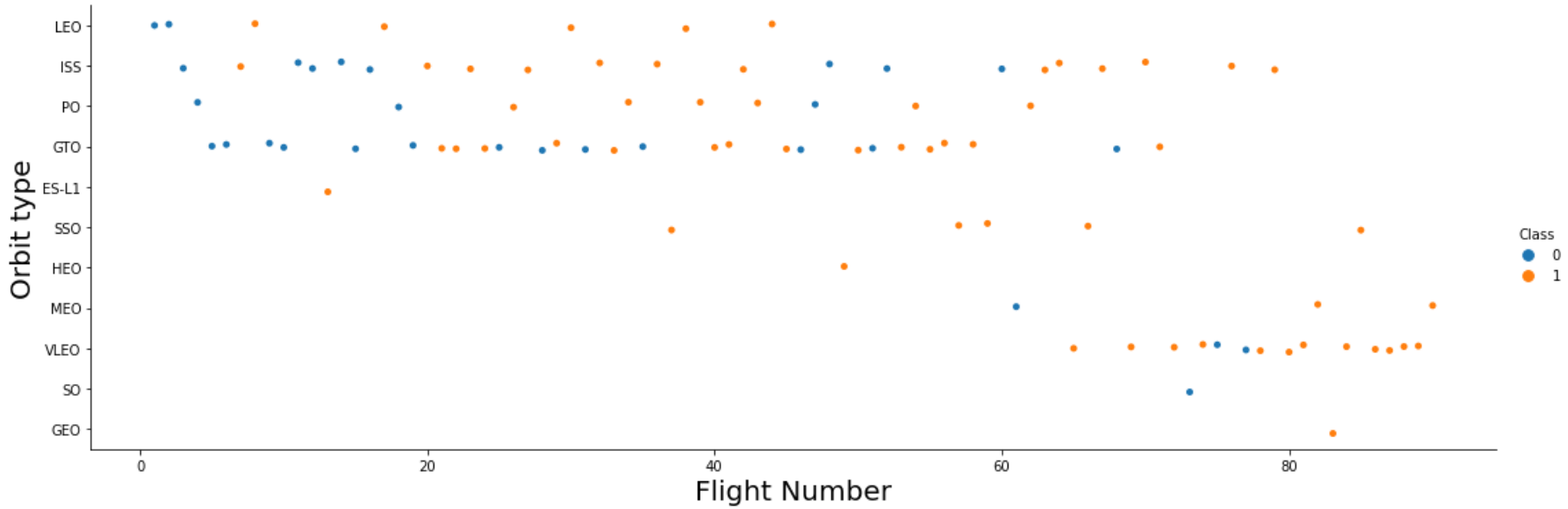
Success rate vs. Orbit type

- Orbits ES-L1, GEO, HEO and SSO have best success rates
- GTO orbit has lowest success rate



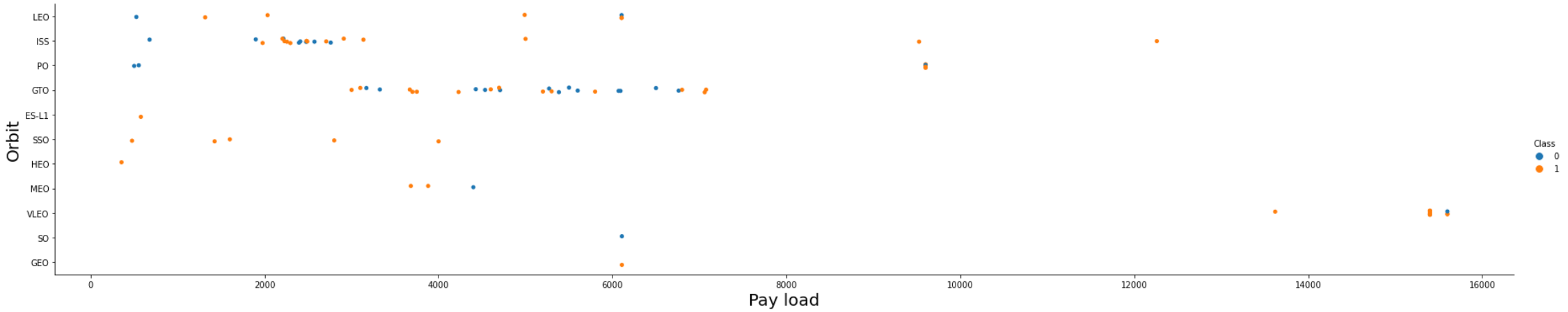
Flight Number vs. Orbit type

- The LEO orbit the Success appears related to the number of flights.
- No relationship between flight number when in GTO orbit.



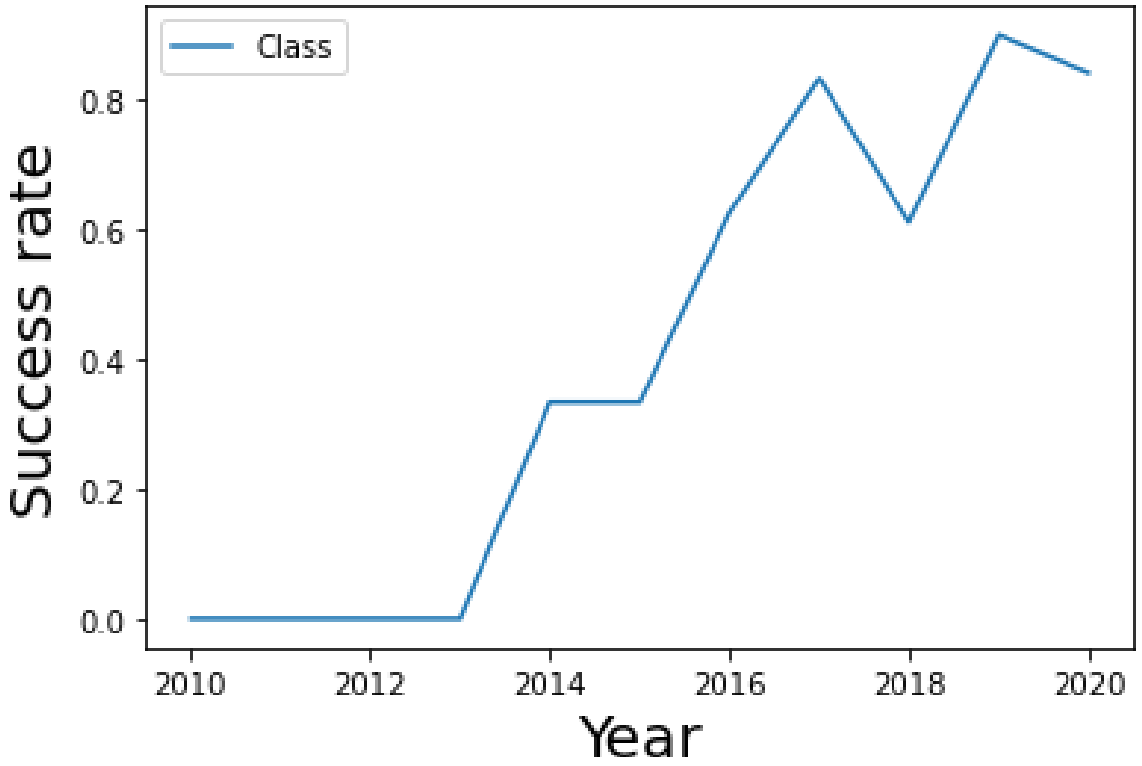
Payload vs. Orbit type

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



Launch success yearly trend

- Success rate starts rising since 2013 and kept increasing till 2020
- Success rate peaks in 2019 with above 80%



EDA with SQL

All launch site names

- Find the names of the unique launch sites

```
select Launch_Site from SPACEXTBL
group by Launch_Site;
```

- Present your query result with a short explanation here

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

Launch site names begin with `CCA`

- Find all launch sites begin with `CCA`

```
select DISTINCT Launch_Site from SPACEXTBL
where Launch_Site like 'CCA%'
```

- Present your query result with a short explanation here

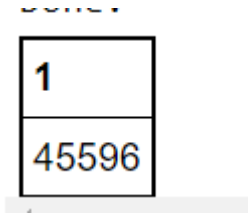
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(payload_mass__kg_) as Total_Payload_Mass FROM SPACEXTBL where CUSTOMER ='NASA (CRS)'
```

- Present your query result with a short explanation here



1
45596

Average payload mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

selectAVG(PAYLOAD_MASS_KG_)AveragePayloadMassfromtblSpaceX whereBooster_Version='F9v1.1'

- Present your query result with a short explanation here

1
2928

First successful ground landing date

- Find the date when the first successful landing outcome in ground pad

```
%sql SELECT * FROM SPACEXTBL WHERE Landing__Outcome LIKE 'Success%' ORDER BY DATE
```

- Present your query result with a short explanation here

1
2015-12-22

Successful drone ship landing with payload between 4000 and 6000

- List the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

*%sql SELECT * FROM SPACEXTBL WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ >4000 AND payload_mass__kg_ <6000*

- Present your query result with a short explanation here

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total number of successful and failure mission outcomes

- Calculate the total number of successful and failure mission outcomes

```
select distinct(mission_outcome), count(*) from SPACEXTBL
group by MISSION_OUTCOME;
```

- Present your query result with a short explanation here

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters carried maximum payload

- List the names of the booster which have carried the maximum payload mass

```
%sql SELECT DISTINCT booster_version, payload_mass__kg_ FROM SPACEXTBL
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

- Present your query result with a short explanation here

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 launch records

- 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

```
%sql SELECT DATE, MONTH(DATE) AS MONTH, landing__outcome, booster_version, launch_site FROM SPACEXTBL WHERE YEAR(DATE)=2015 AND landing__outcome = 'Failure (drone ship)
```

- Present your query result with a short explanation here

DATE	MONTH	landing__outcome	booster_version	launch_site
2015-01-10	1	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	4	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

```
%sql SELECT landing__outcome, COUNT(landing__outcome) AS landing_outcome_counts FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND landing__outcome LIKE '%Success%' GROUP BY landing__outcome
```

- Present your query result with a short explanation here

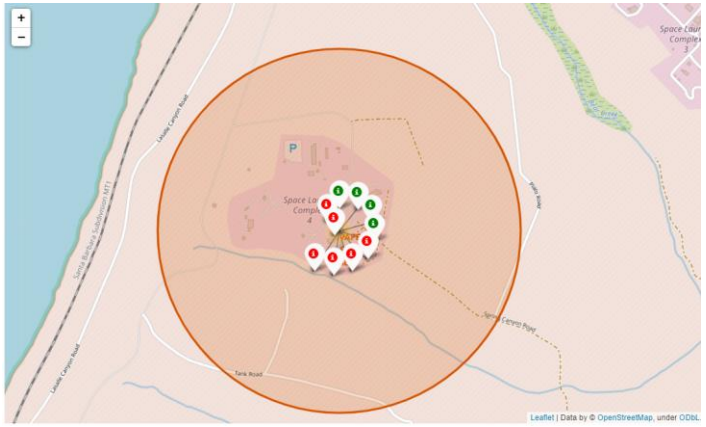
landing__outcome	landing_outcome_counts
Success (drone ship)	5
Success (ground pad)	3

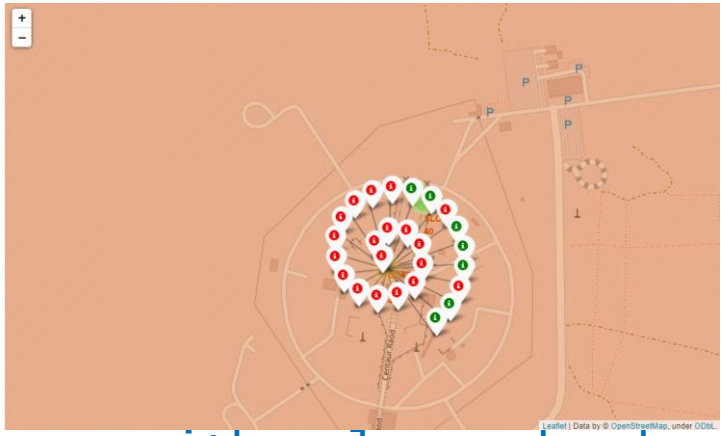
Interactive map with Folium

Locations of all Launches Sites

- The unique launch sites with circle markers
- Labels with corresponding site names

Success/ Failure launches for each site





Each launch record has a pointer on map with color code basis success or failure of corresponding data point.

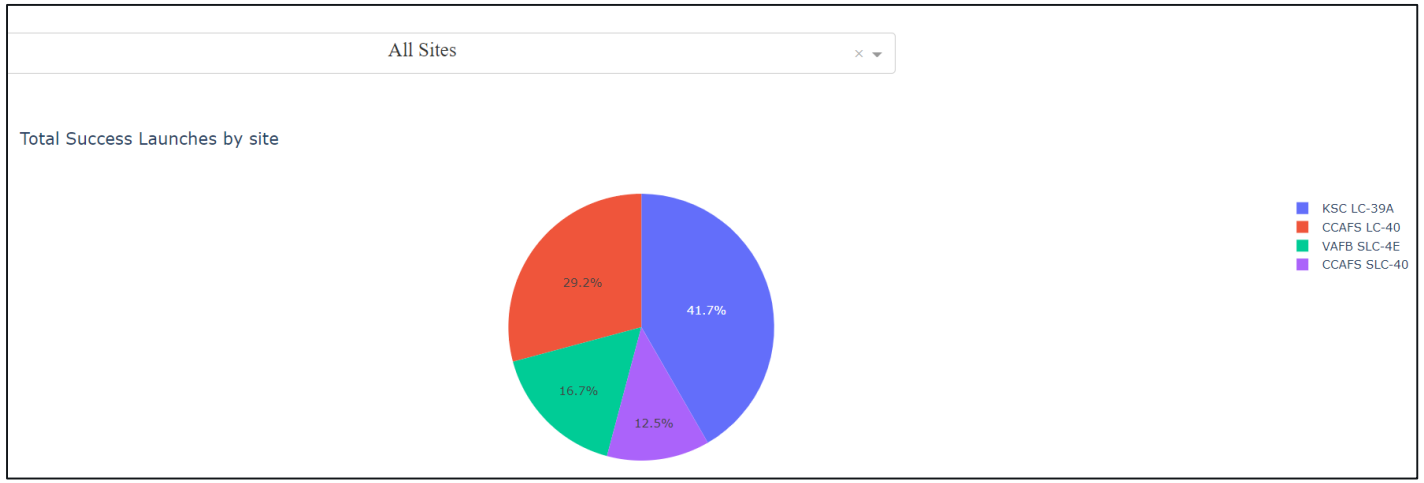
Green marker shows successful launches and red market indicates failures.

Distance between launch site and its proximities

- It shows the straight distance from the two points on map and label depicts the distance shown by line.

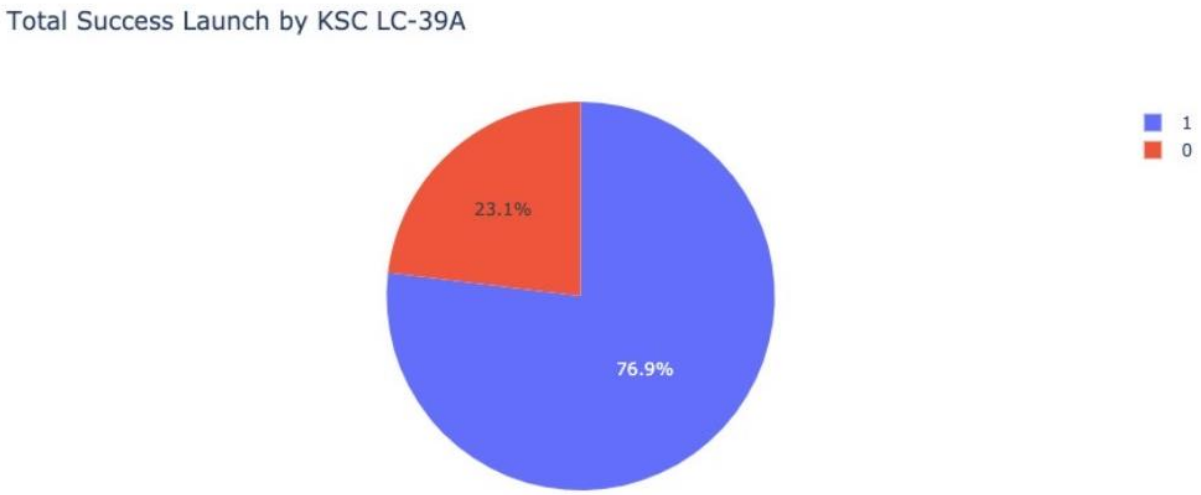
Build a Dashboard with Plotly Dash

Total Success Launches by Site



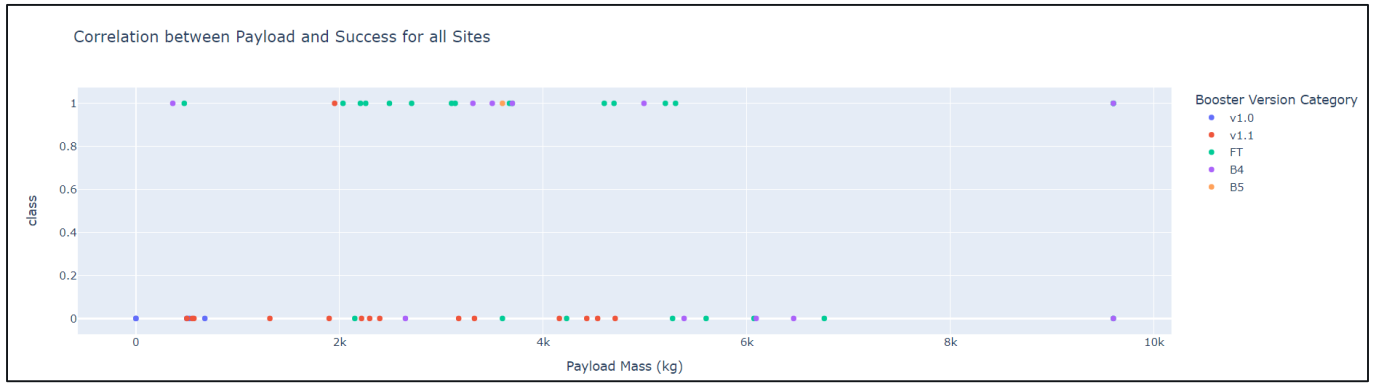
- The launch site ‘KSC LC -39A’ has the highest success rate of 41.7% among all sites

Success Launches for Individual sites



- The launch site ‘KSC LC -39A’ a success rate of 76.9% vs a failure percentage of 23.1%

Correlation between payload & success

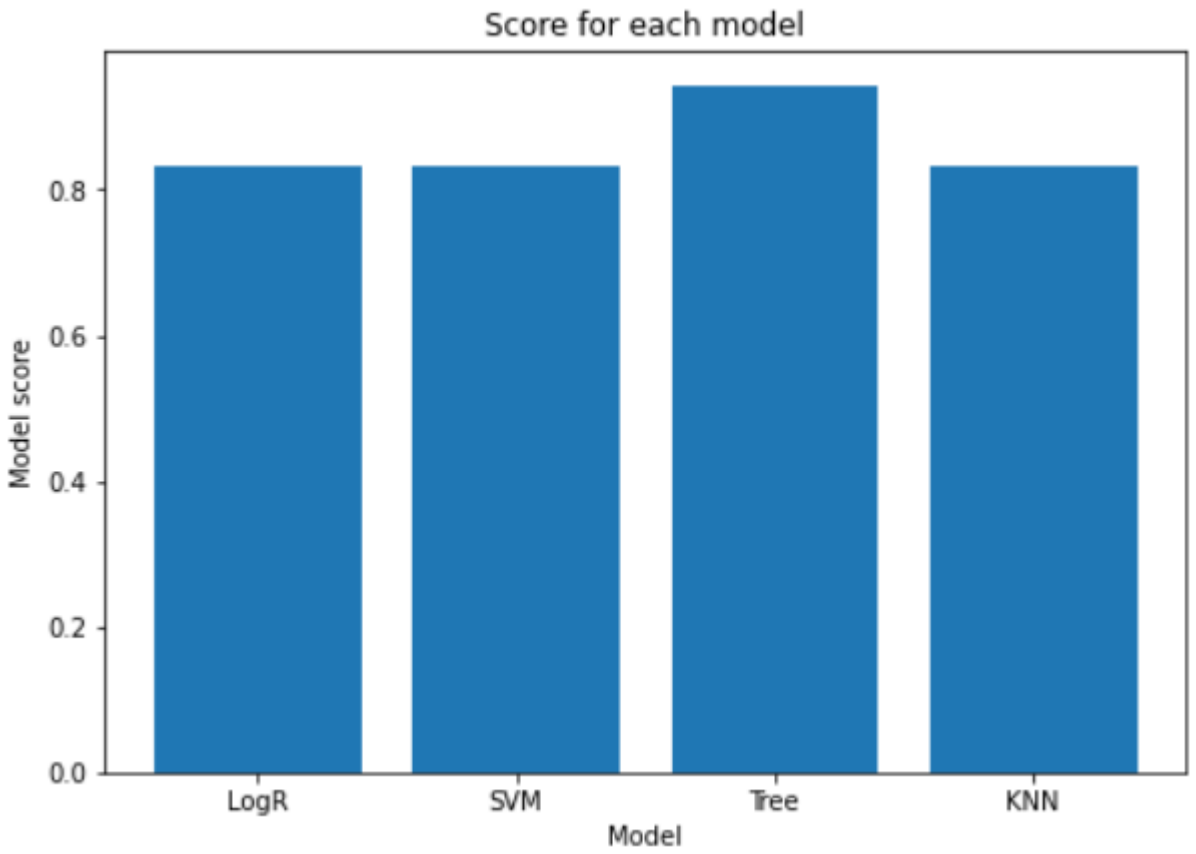


- There is a lower chance of success at lower payload mass.

Predictive analysis (Classification)

Classification Accuracy

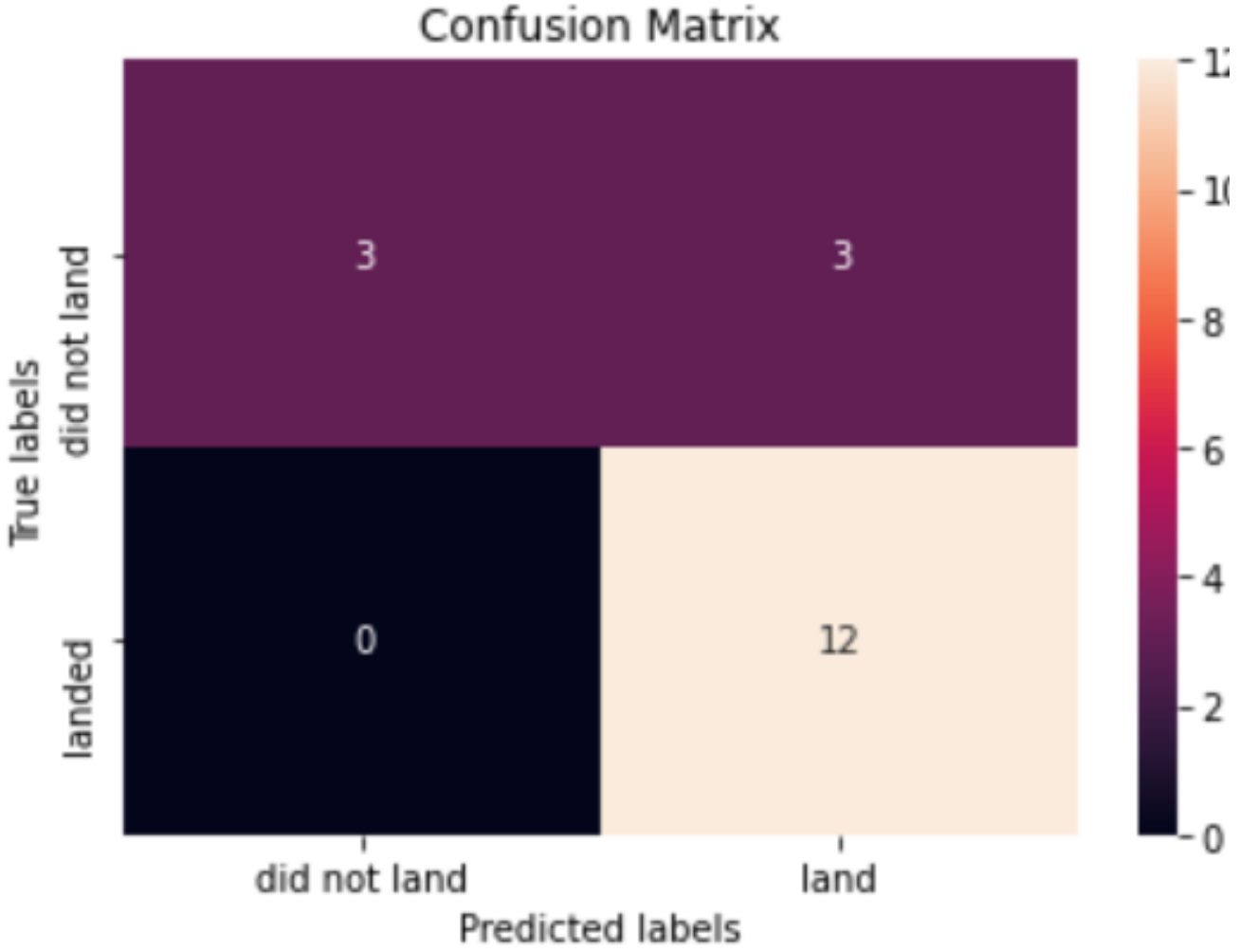
Best Algorithm is Tree with a score of 0.8928571428571429



Confusion Matrix

We have test set of 18 samples:

1. Our model predicted correctly for all “Did not land” launches
2. It predicted correctly for 12 landed launches, and faultly for 3 launches which were in fact did not land



CONCLUSION

- The Tree Classifier Algorithm is the best predictive model classification with highest accuracy
- Low weighted payloads perform better than the heavier payloads