IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anne-Sophie
30/06/2025

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection
  - Data Wrangloing
  - EDA with SQL and data visualization
  - Interactive map with Folium
  - Dashboard with PlotlyDash
  - Predictive analysis
- Summary of all results
  - Insight drawn from EDA
  - Launch Site proximities analysis
  - Dashboard with Plotly
  - Predictive analysis (classification)

Introduction

- Project background and context:
  - The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX.
  - One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Problems you want to find answers
  - Predict successful landing of stage 1 to get increase likelihood of success for a new company SpaceY which want to compete with Space X

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX Rest API and Webscraping from Wikipedia
    https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Data was cleaned and new features were added for categorical data, unnecessary information were dropped and a class column was creating with value 1 when the landing of stage 1 was successful and the value 0 otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Optimization and comparison of 4 different classification algorithm: logistic regression, support vector machine, decision tree and K nearest neighbour to determine the best method

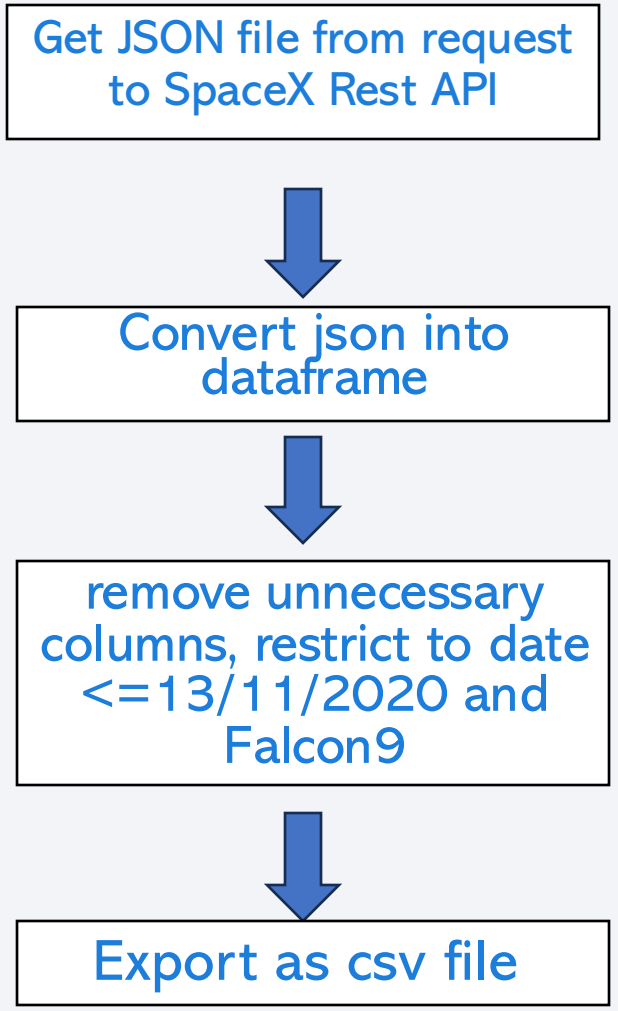
Data Collection

- Data set were collected by 2 methods:
  - using Space X API and then converting json file into a panda dataframe
  - by web scraping from wikipedia https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches using beautifulsoup
- Data were cleaned by checking for missing value and either replacing data by mean value or removing rows

Data Collection – SpaceX API

- SpaceX API was used to collect data. The data were then cleaned and the resulting dataframe was exported as .csv

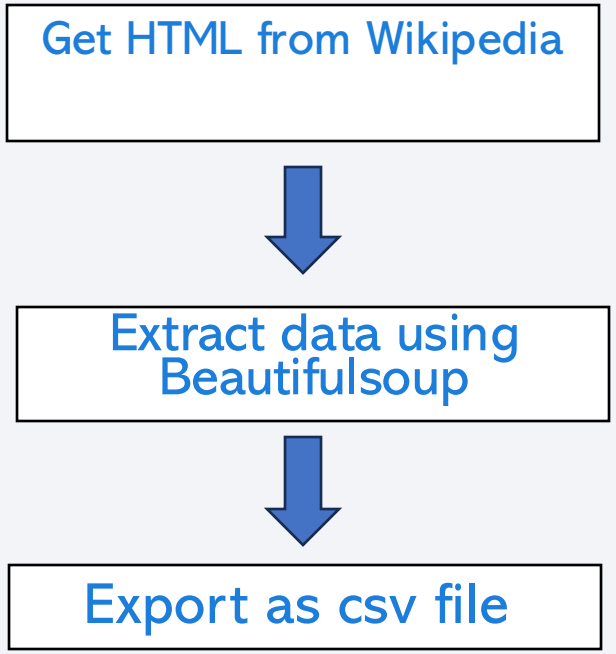
- GitHub URL https://github.com/ASR2211/DataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Webscraping technique aws used to collect data from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- GitHub URL https://github.com/ASR2211/DataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data were loaded from csv file created in part 1
- Number of lauches per site was calculated
- Number and occurrence of each orbit were calculated
- Mission outcome were examined and the entries corresponding to bad outcomes were combined
- A new column class was created and assigned to 0 when the mission outcome was in the list of bad outcome, 1 otherwise
- The success rate was determined to be 66.7%
- Resulting data was exported to .csv
- GitHub URL https://github.com/ASR2211/DataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Several graphs were generated:
  - Scatterplots FlightNumber vs. PayloadMass , Flight Number vs Launch Site and Payload Mass vs Launch Site to determine which parameter have an impact on successful landing of stage 1, FlightNumber vs Orbit type, Payload Mass vs Orbit type
  - The relationship between success rate of each orbit type was visualized with a barchart
  - The launch success yearly trend was visualized using a linechart
- Features Engineering was used to Create dummy variables to categorical columns. We obtained overall 80 features which were then Cast to float64
- GitHub URL
  https://github.com/ASR2211/DataScienceCapstone/blob/main/edadataviz.ipynb

EDA with SQL

- SQL statement were used using %sql to find distinct lauch sites, records with launch sites beginning with CCA, total payload mass launched by NASA, average payload mass carried by booster F9 v1.1, date of first successful outcome, name of boosters with successful drone ship, number of successful mission outcomes, booster version with the max payload mass, month in 2015 were successful landing occurred, landing outcome within a given perios
- GitHub URL: https://github.com/ASR2211/DataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers were added to show the launch sites, then the number of launch was indicated for each sites and the markers were colored in green for successful and red for failure. The distance of the launch site to different elements were then calculated
- GitHub URL : https://github.com/ASR2211/DataScienceCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

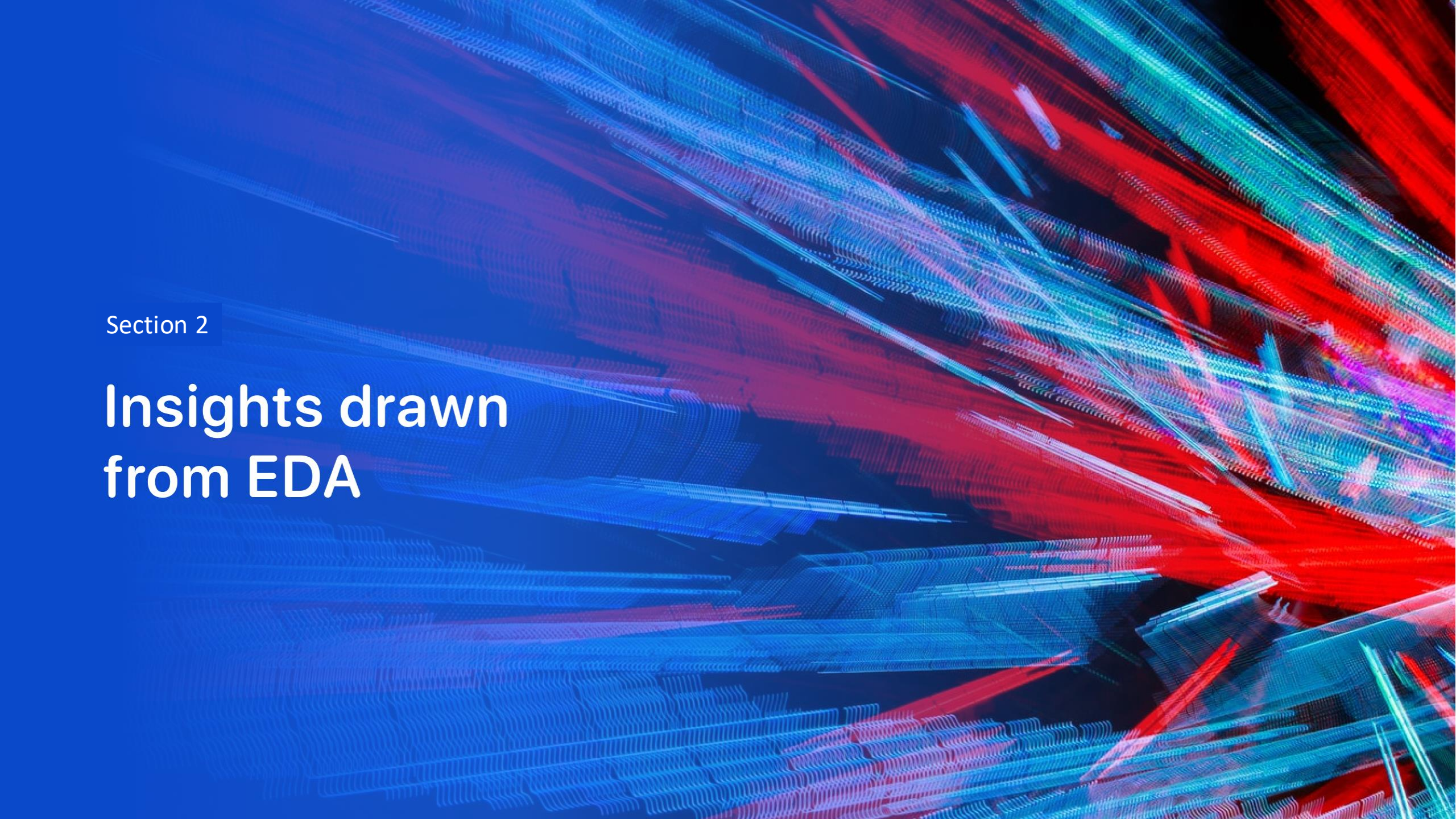
- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

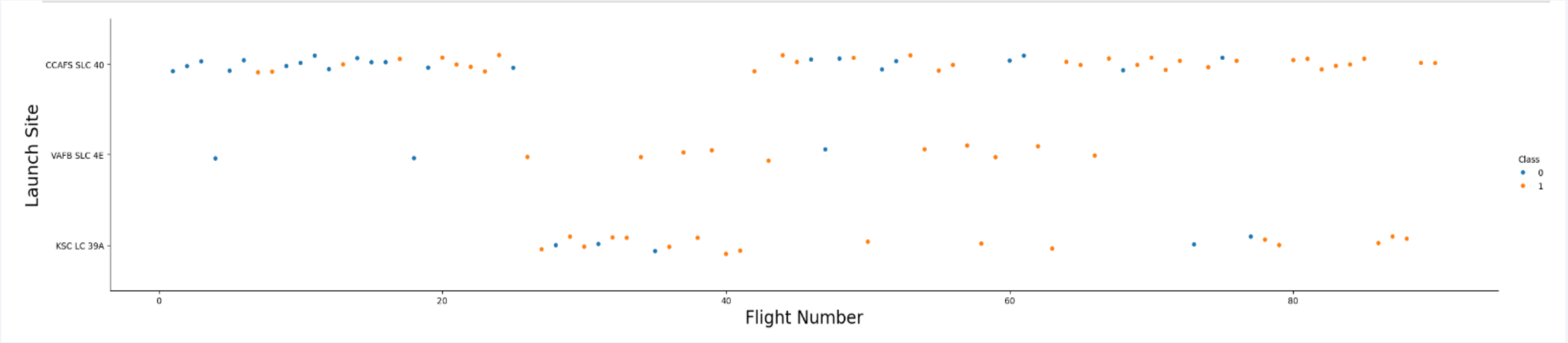
- The target Y was defined as a numpy array
- The list of feature X was standardized using StandardScaler()
- The data set was divided into train and test set using train_test_split
- For each model: logistic regression, Support vector Machine, decision tree and K nearest neighbour:
  - GridsearchCV was used to find best parameters from the given dictionary of parameters
  - The accuracy on the test data was calculated using the score method
  - The confusion matrix for test data was ploted
- The models were compared
- GitHub URL https://github.com/ASR2211/DataScienceCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

Results

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

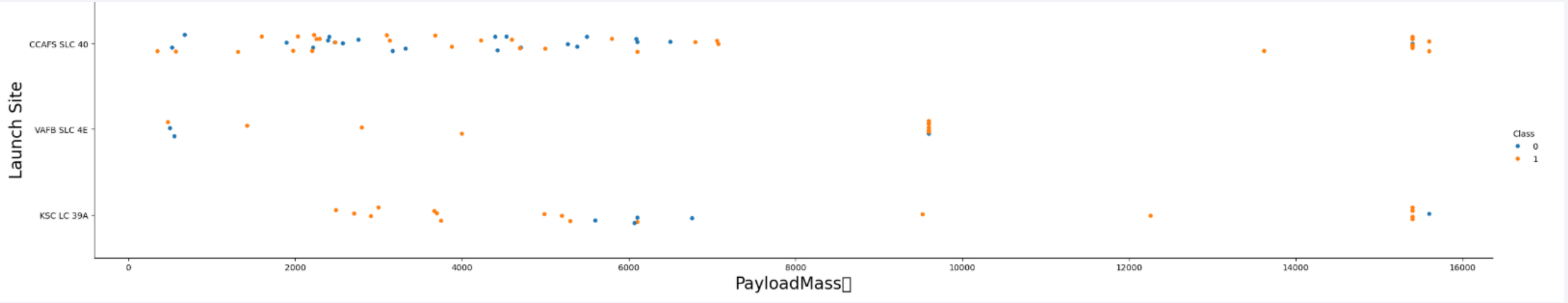

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



We see that as the flight number increases, the first stage is more likely to land successfully.

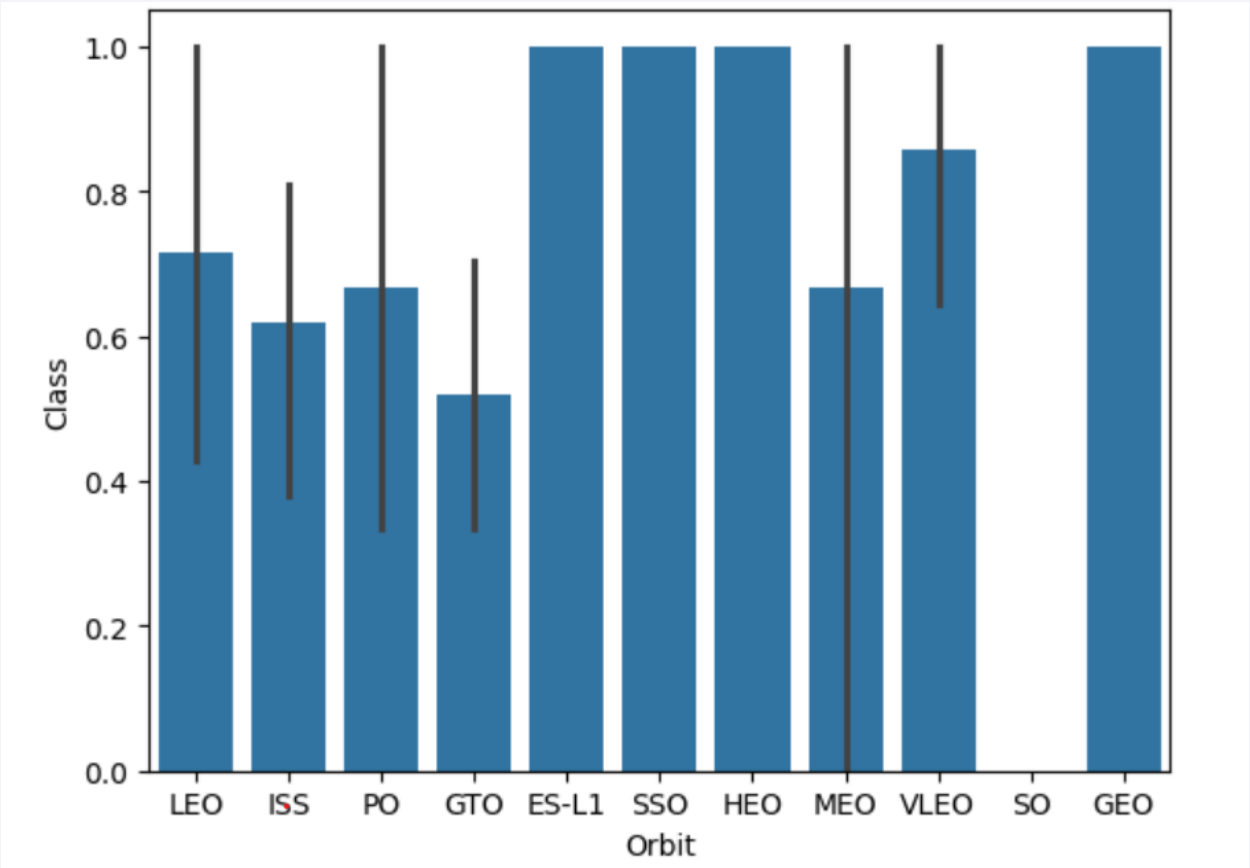
Payload vs. Launch Site



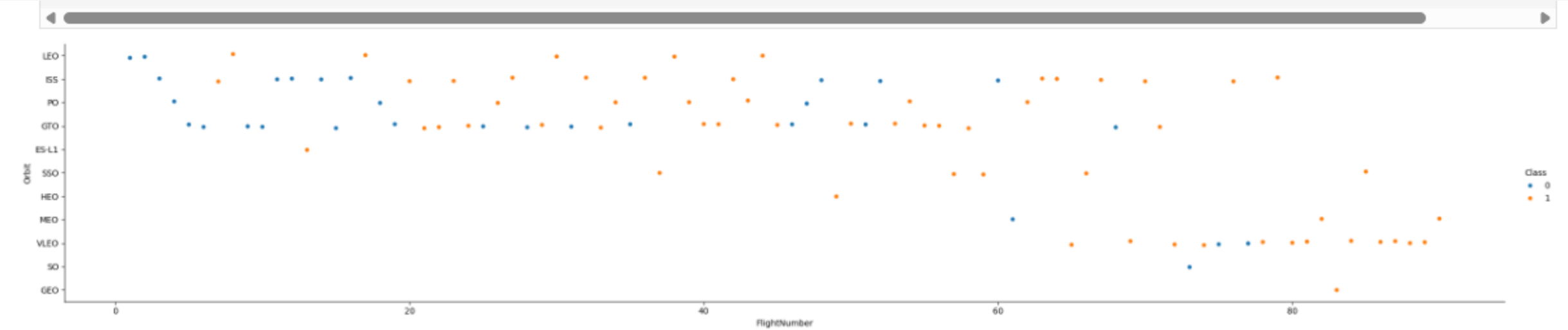
The payload mass also appears to be a factor; with more massive payloads, the first stage often returns successfully.

Success Rate vs. Orbit Type

- The highest success rate were obtained for ES-L1, SSO, HEO and GEO orbits

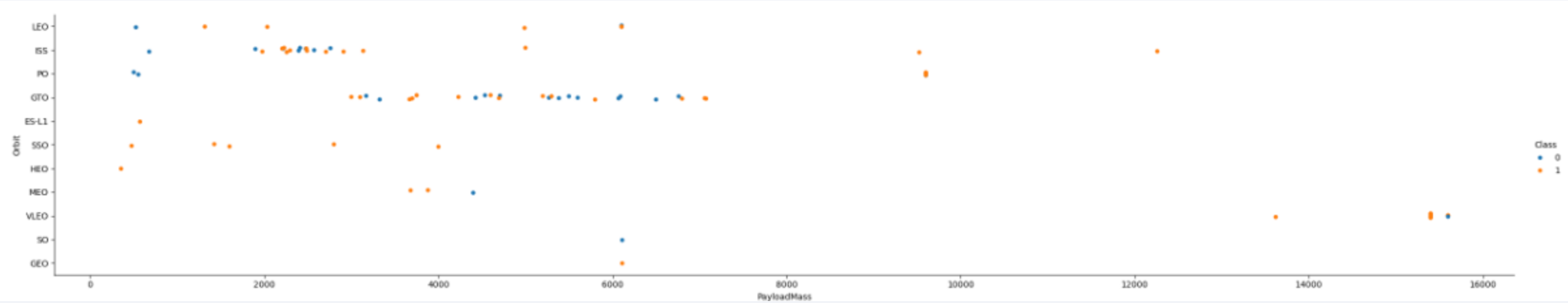


Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights.
Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

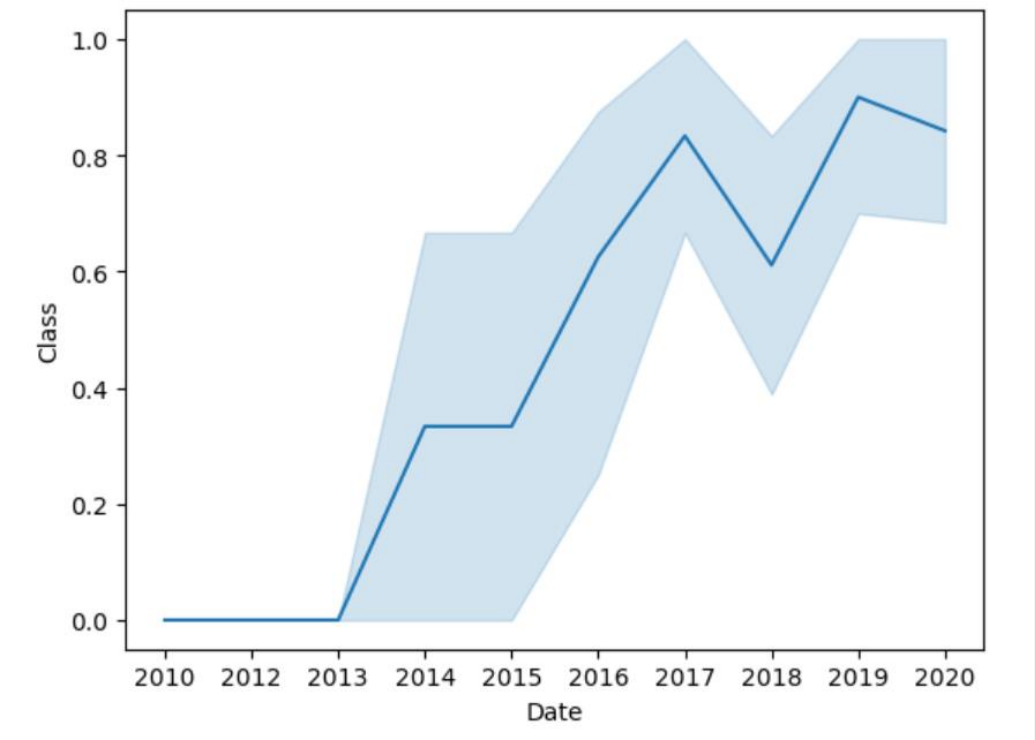
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- we observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- There are 4 distinct launch sites

- we used the following query to generate the table:

```
%sql select  Distinct Launch_Site from SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

- The query used to generate the table is:

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA: 45596 kg
- Query used:

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer is 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1: 2928.4kg
- query used: `%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version is 'F9 v1.1'`

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad: 22/12/2015
- query used: `%sql select MIN(Date) from SPACEXTBL where Landing_Outcome is '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

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

- Query used :

```
%sql select Distinct Booster_Version from SPACEXTBL where Landing_Outcome is 'Success (drone ship)' AND
PAYLOAD_MASS__KG_ between 4000 and 6000
```

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- 100 success, 1 failure

Mission_Outcome	COUNT1
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- `%sql select Mission_Outcome, COUNT(Mission_Outcome) AS COUNT1 from SPACEXTBL GROUP BY Mission_Outcome`

Boosters Carried Maximum Payload

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

```
%sql select  Distinct Booster_Version from SPACEXTBL WHERE PAYLOAD_MASS__KG_ is (SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

]:

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 the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

- Query used

```
%sql select substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL WHERE substr(Date,0,5)='2015' AND Landing_Outcome IS 'Failure (drone ship)'
```

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

Landing_Outcome	COUNT2
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

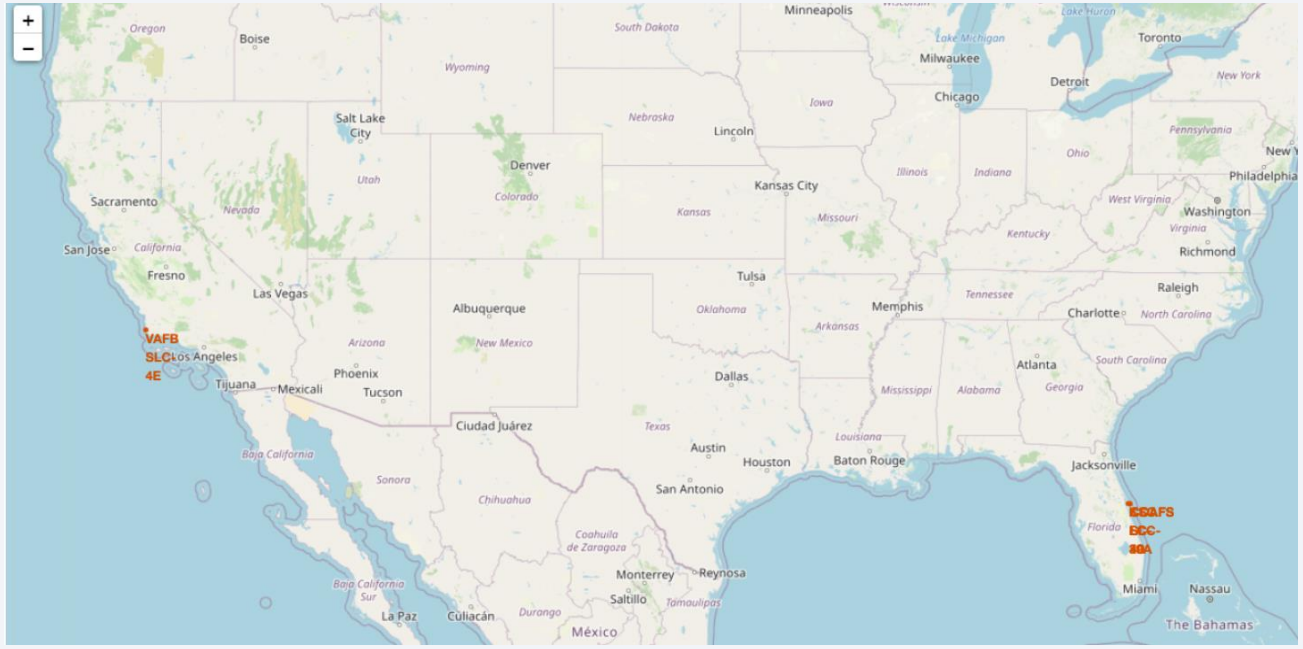
- Query used

```
%sql select Landing_Outcome, COUNT(Landing_Outcome) AS COUNT2 from SPACEXTBL WHERE Date between '2010-06-04' and '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT2 DESC
```

Section 3

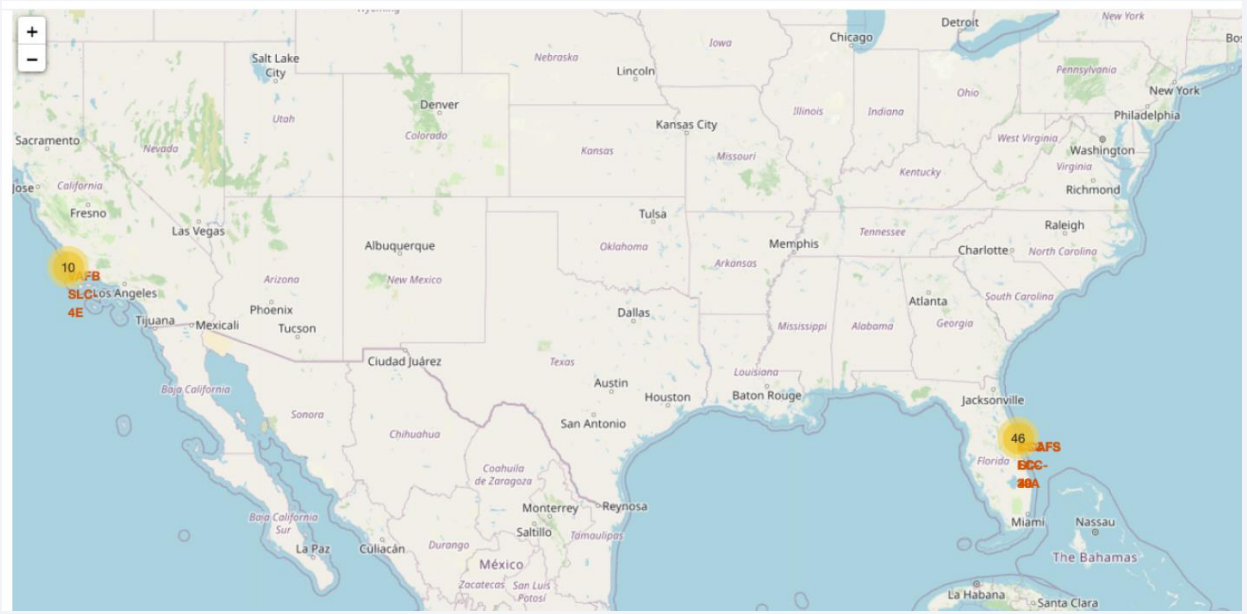
Launch Sites Proximities Analysis

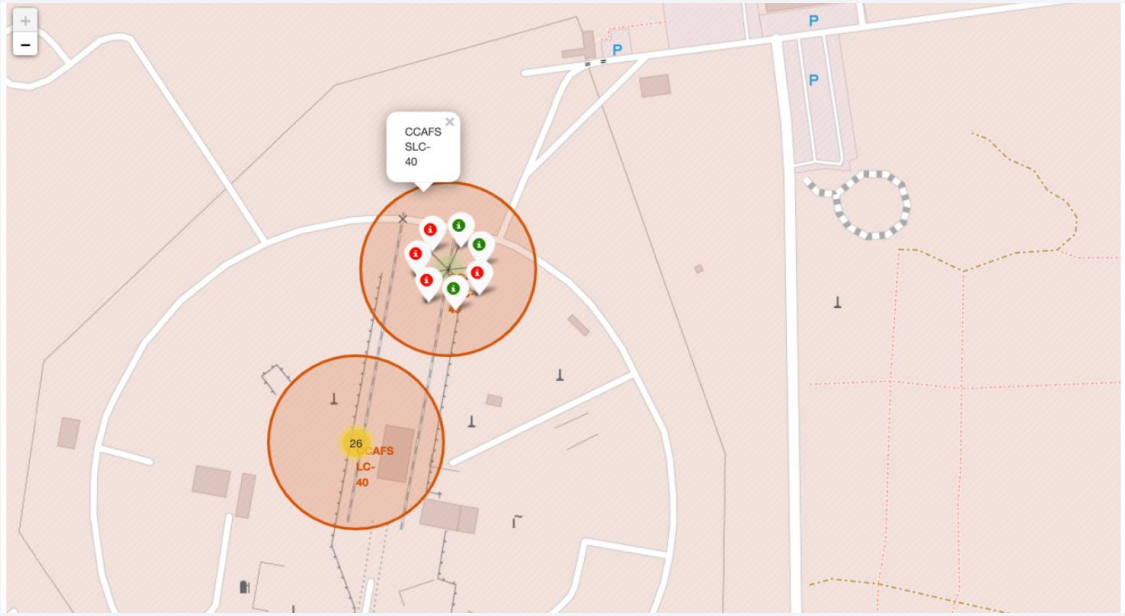
Launch Sites Locations Analysis with Folium



- The lauch sites are in proximity to the Equator line and in very close proximity to the coast

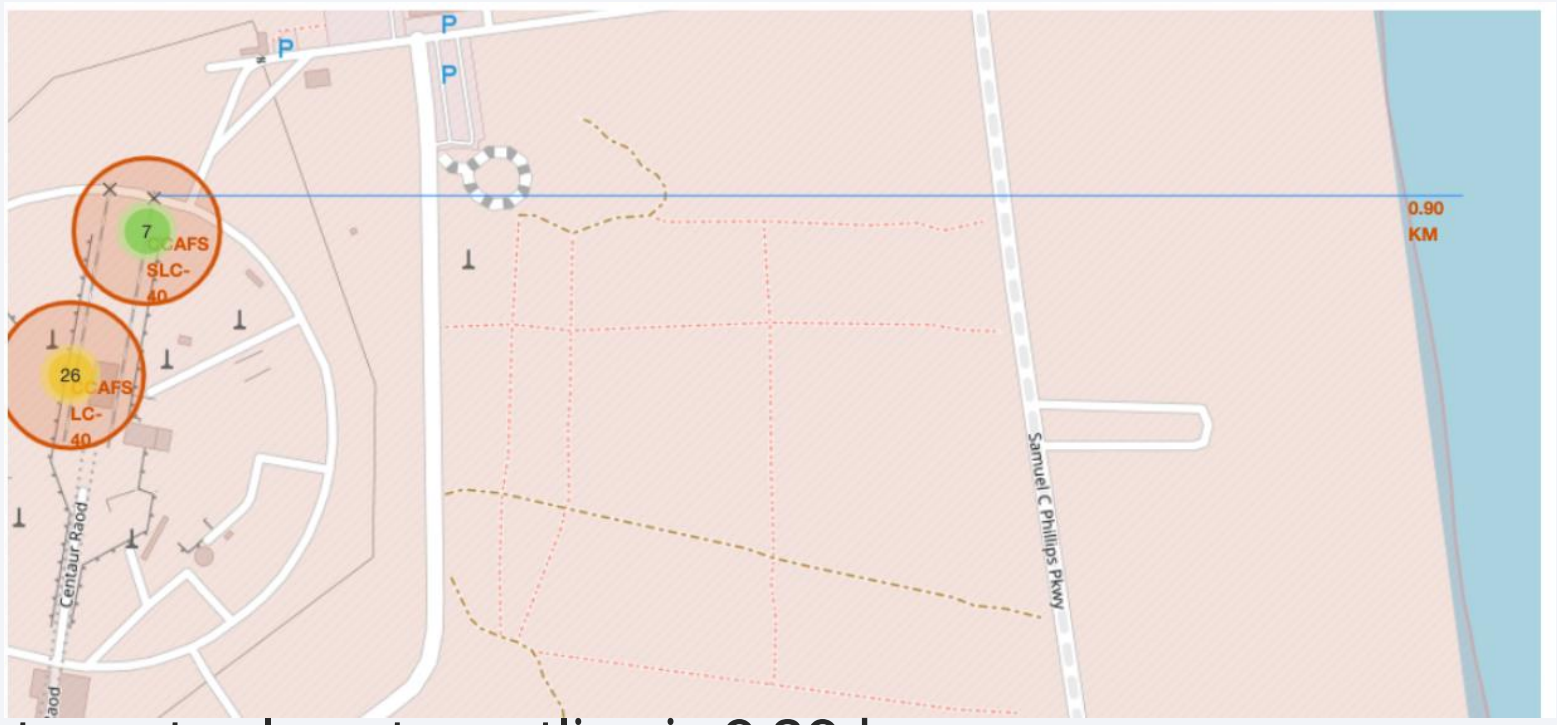
Launch sites with lauch outcome



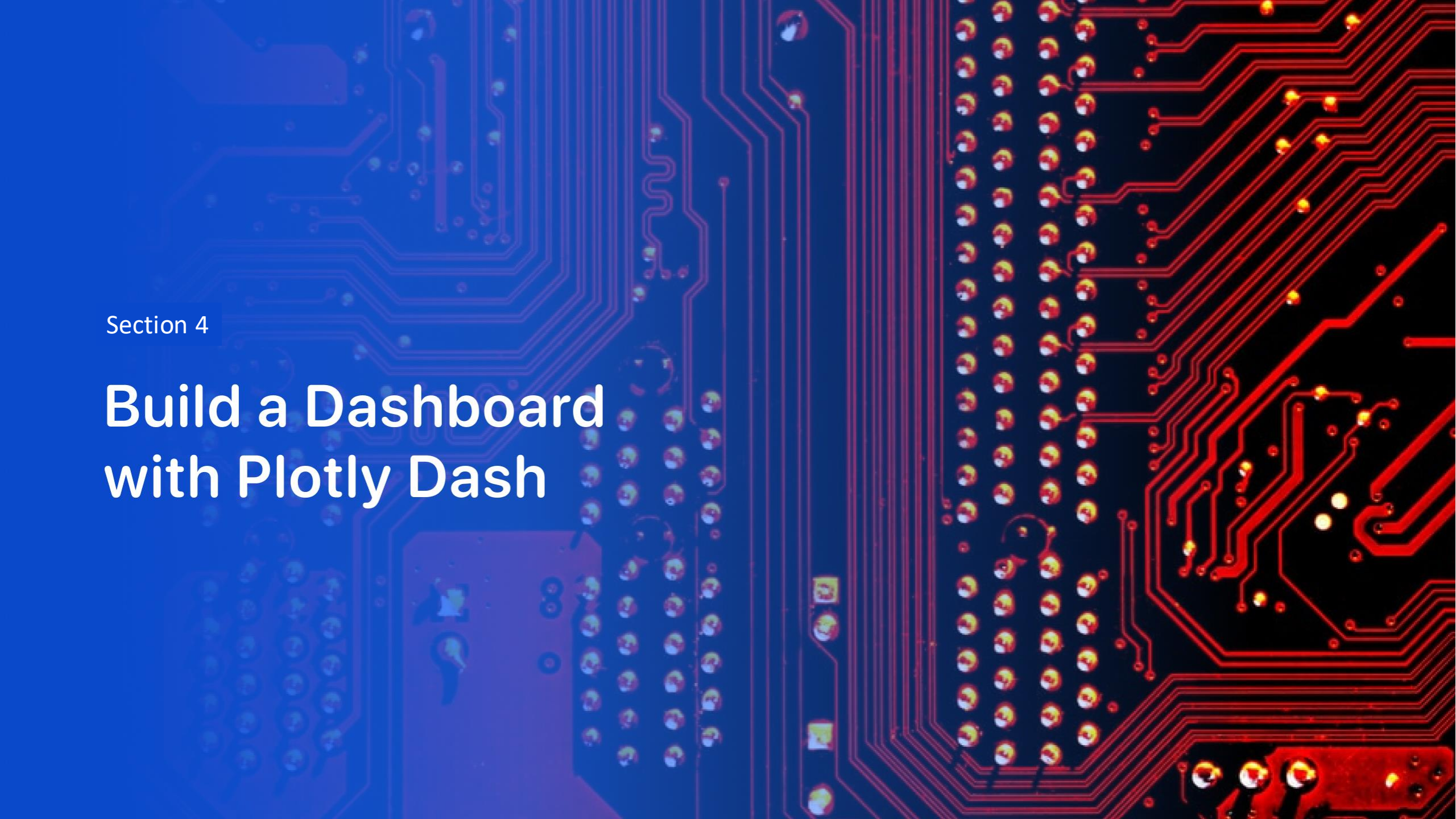


- the map now contains information about the number of launches and the launch outcome (green for successful and red for failed)

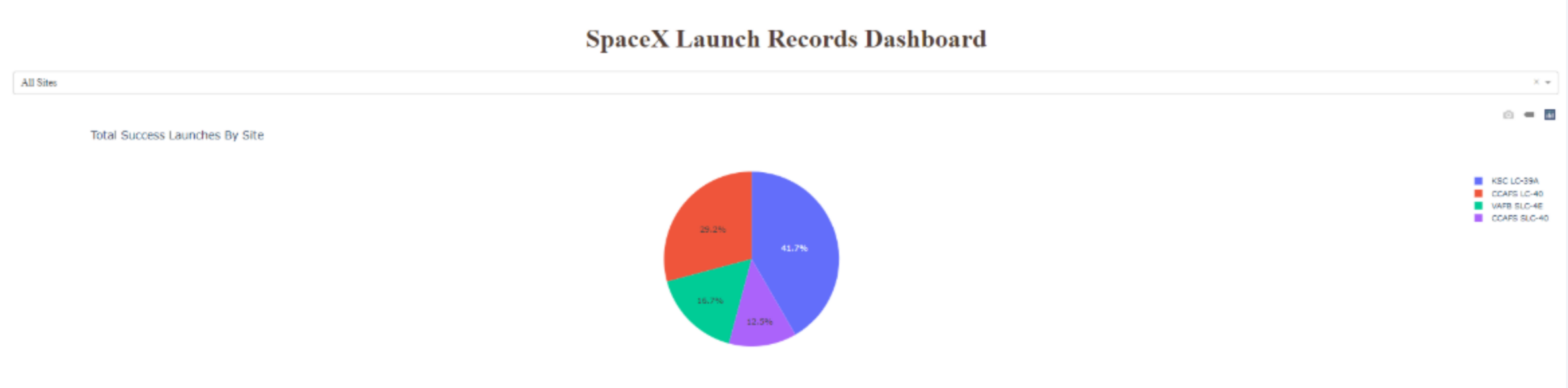
Distances between a launch site to its proximities



- Distance to closest coastline is 0.89 km

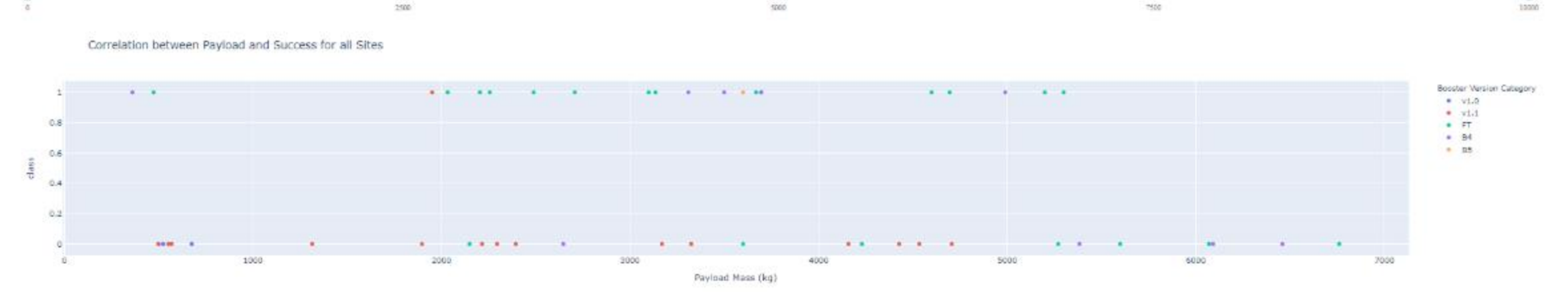

Section 4
Build a Dashboard with Plotly Dash

Total success launches for all sites



- KSC LC 39A has the highest amount of success lauch and CCAFS SLS40 the lowest

Correlation between payload and success

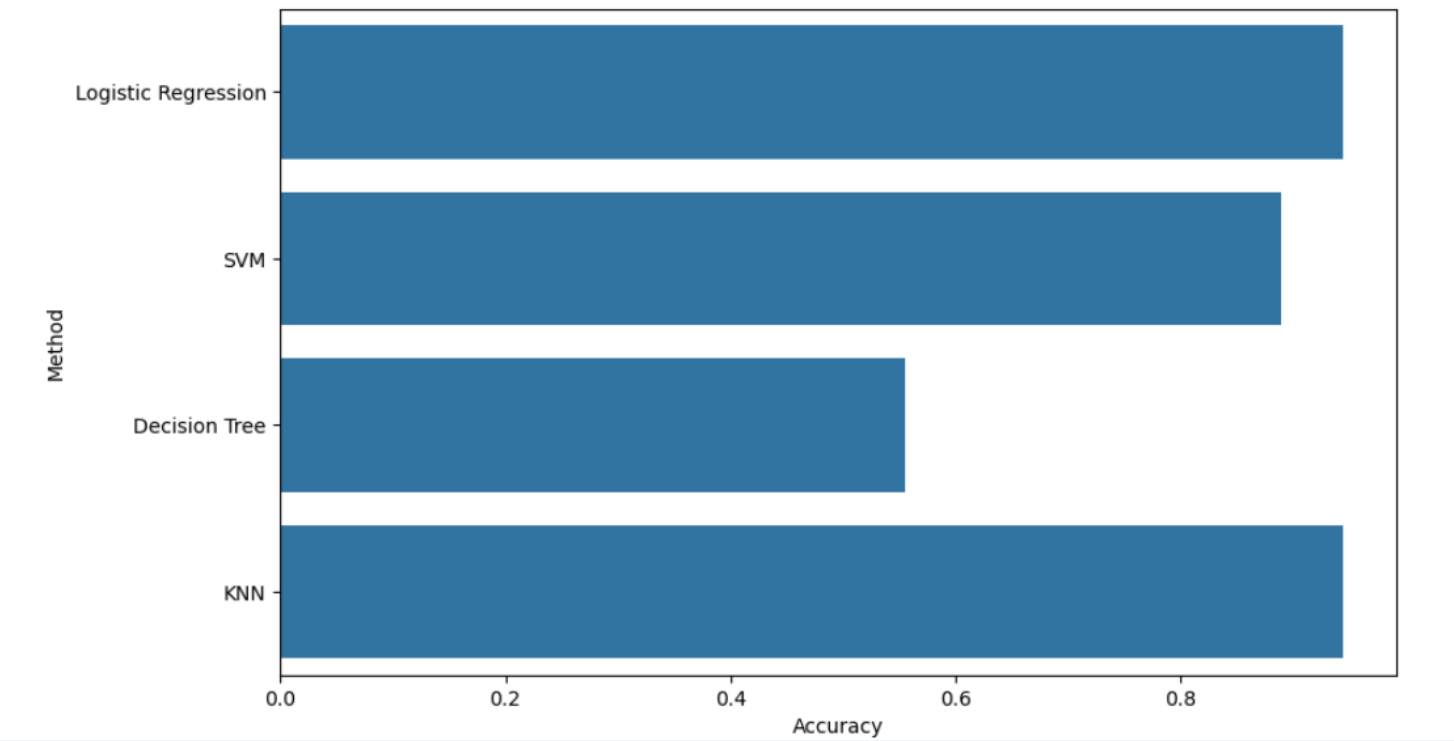


- Most success were obtained for paylaod with mass in range 2000-4000
- The highest amount of success launch was obtained for booster version FT

Section 5

Predictive Analysis (Classification)

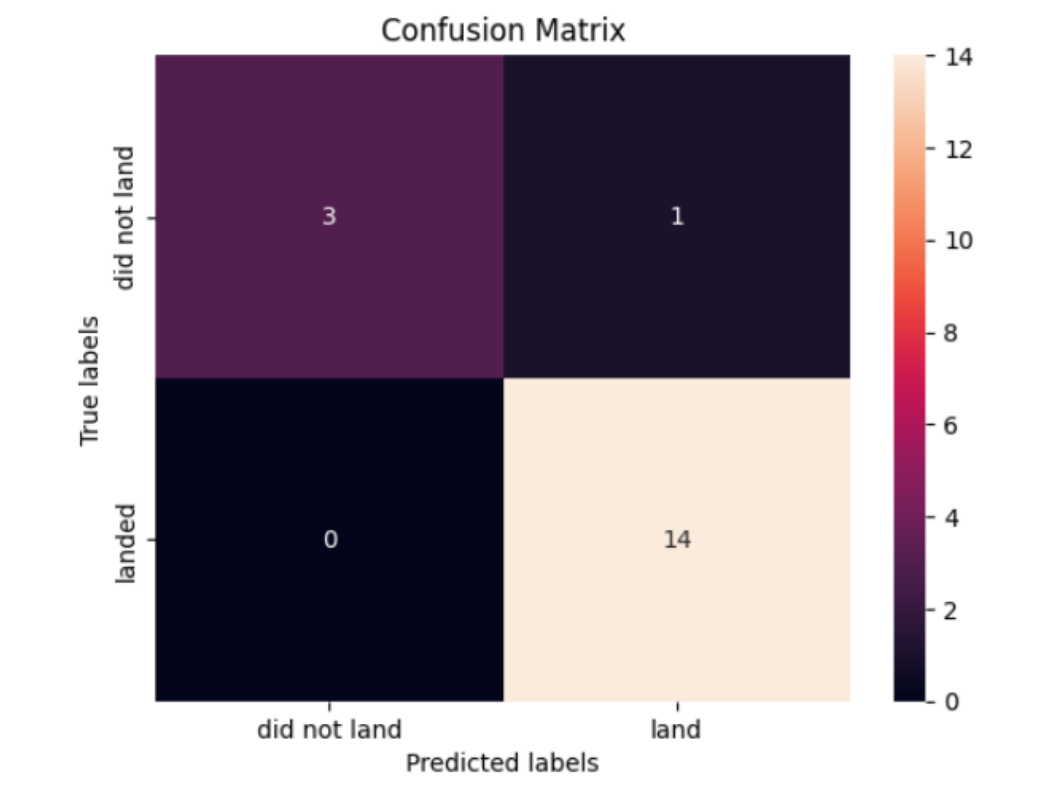
Classification Accuracy



	0
Method	Test Data Accuracy
Logistic_Reg	0.944444
SVM	0.888889
Decision Tree	0.555556
KNN	0.944444

- The methods which performed best are KNN and LR

Confusion Matrix



- The method which performed best is KNN

As only one prediction was not correct, the other 17

Were correct

Conclusions

- As the flight number increases, the first stage is more likely to land successfully.
- with more massive payloads, the first stage often returns successfully.
- The highest success rate were obtained for ES-L1, SSO, HEO and GEO orbits
- we observe that the sucess rate since 2013 kept increasing till 2020

Thank you!