Winning Space Race with Data Science

Xue Zhao
10/15/2024

Outline

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

Executive Summary

Summary of methods

- Collect data from SpaceX API and web scrapping;
- Perform EDA with SQL and Visualization
- Create interactive data visualization with Folium and Ploty Dash
- Predict launch site success rate with machine learning methods

Summary of all results

- Relationship between flight number, payload, launch site and success rate are discovered
- The time trend in success rates and launch site locations are analyzed
- Decision tree is decided as the best model for predicting launch outcome

Introduction

- **Project background and context**

Falcon 9 rocket is advertised that it launches with a cost of 62 million dollars on SpaceX websites, while other providers need more cost. The reason is that SpaceX can reuse the first stage. This gives us the motivation that if we can determing the landing outcome of the first stage, we can determine the cost. And the information would be used in helping other company to bid against SpaceX.

- **Problems you want to find answers**

1. What factors can affect the landing outcomes
2. How does landing outcomes change across years and different launch sites
3. What can we do to improve landing outcomes

Section 1

Methodology

Methodology

Executive Summary

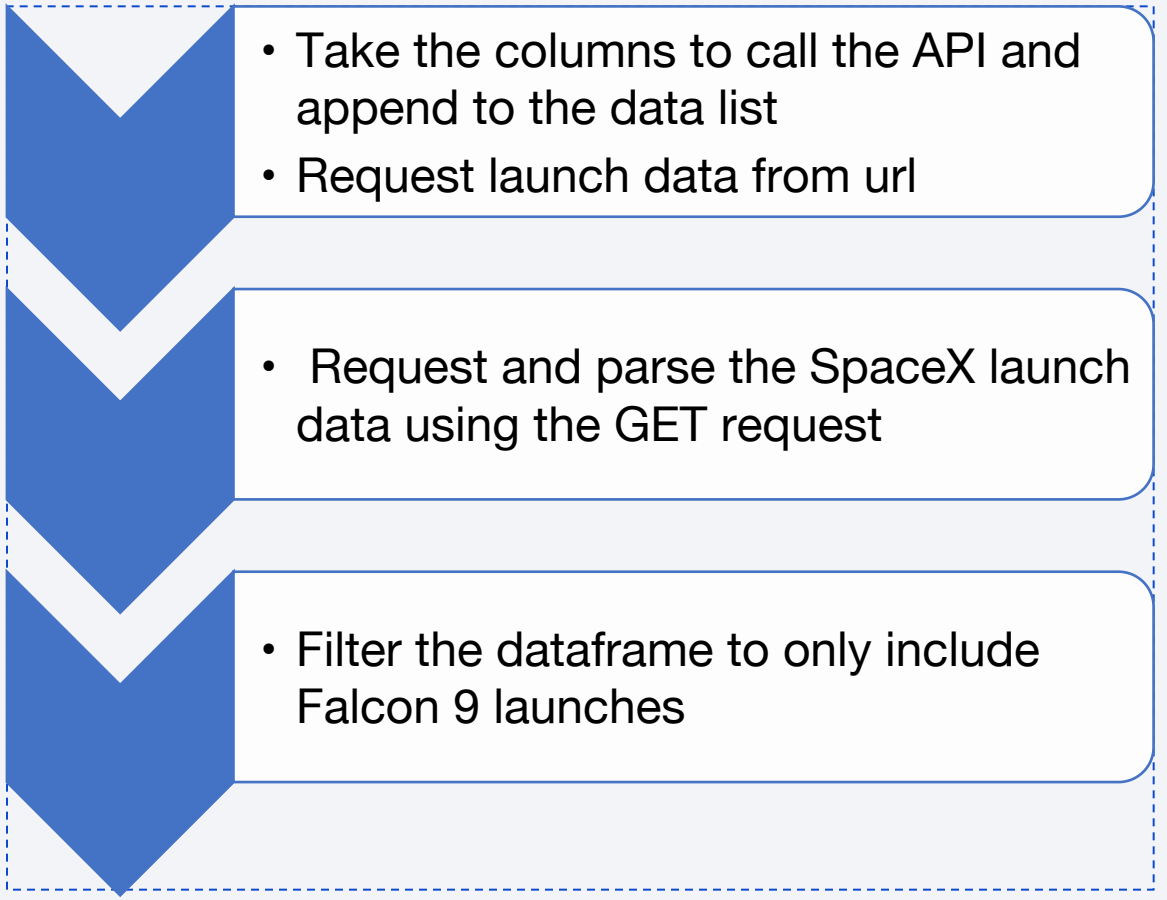
- Data collection methodology:
  - Data was collected through SpaceX API and webscrapping
- Perform data wrangling
  - Replace missing values with mean and created aggregated counts
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tune parmaeters, evalueate and choose the best peroformance models

Data Collection

- Make the request to the Spacex Launch data using Get request
- Filter the dataframe to Falcon 9 launches
- Replace missing values with mean
- Parse Falcon 9 launch records HTML table from Wikipedia using BeautifulSoup
- Request the Falcon9 Launch page from its URL; extract column names from HTML table header and create data frame

Data Collection – SpaceX API

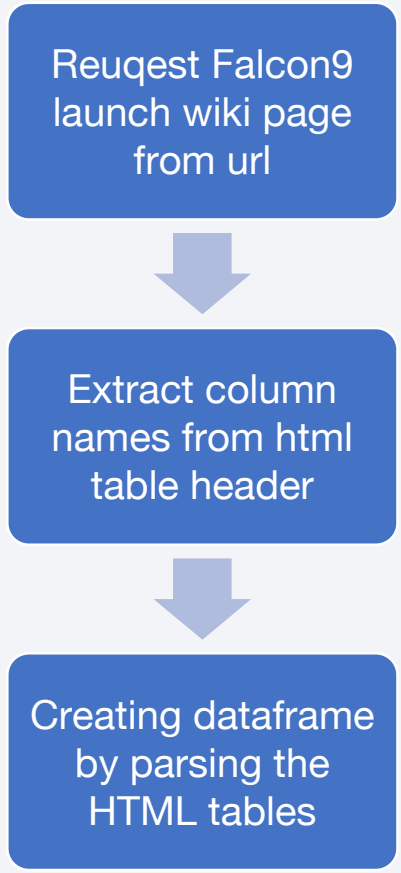
- Request launch data using url and parse data using GET request; filter to needed dataframe
- GitHub URL of the completed SpaceX API calls notebook: *https://github.com/miao678/IBM-data-science-capstone-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb*



Data Collection - Scraping

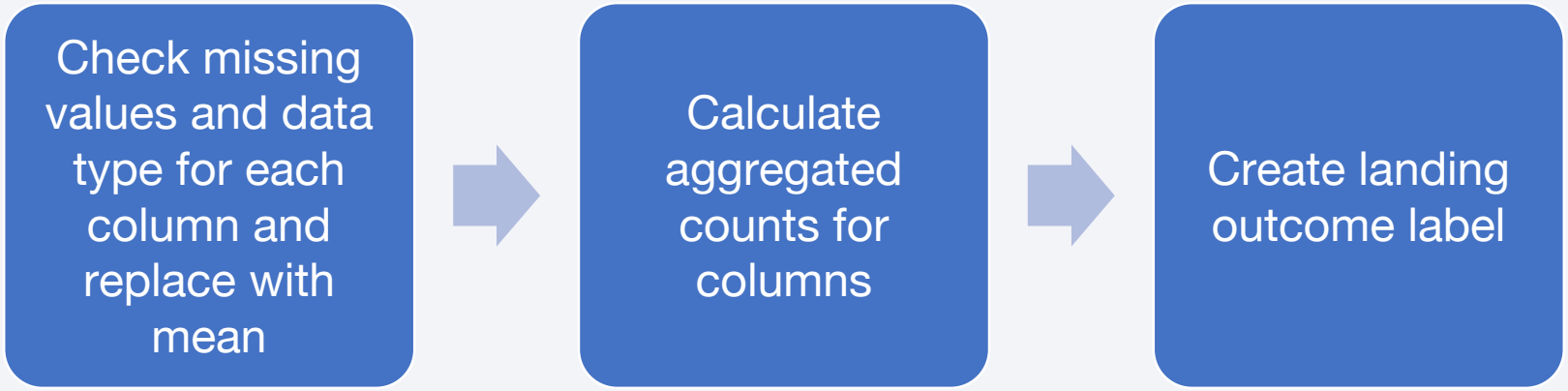
- Request from html page and get request response; extract variables from html tables and parse the tables

- GitHub URL of the completed web scraping: *https://github.com/miao678/IBM-data-science-capstone-project/blob/main/jupyter-labs-webscraping.ipynb*



Data Wrangling

- Check missing values and data type for each column and replace the missing values with mean; calculate aggregated counts for each site, each orbit, mission outcome of the orbits; create label for outcome column



- GitHub URL of data wrangling: *https://github.com/miao678/IBM-data-science-capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb*

EDA with Data Visualization

- Scatter plots and bar charts are used for investigating the relationship between different variables to find potential features for modeling; Line chart is used for analyzing the average success rate trend over the years.
- GitHub URL of EDA with data visualization notebook: *https://github.com/miao678/IBM-data-science-capstone-project/blob/main/edadataviz.ipynb*

EDA with SQL

- %sql select distinct launch_site from SPACEXTABLE
- %sql select launch_site from SPACEXTABLE where launch_site like 'CCA%' limit 5
- %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where customer='NASA (CRS)'
- %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where booster_version='F9 v1.1'
- %sql select payload from SPACEXTABLE where PAYLOAD_MASS__KG_=(SELECT max(PAYLOAD_MASS__KG_) from SPACEXTABLE)s
- GitHub URL of EDA with SQL: *https://github.com/miao678/IBM-data-science-capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb*

Build an Interactive Map with Folium

- Markers, circles, lines and mouse positions are created and added to a folium map
- Circles and markers used to show launch sites area on the map; mouse positions can show coordinates of any points of interests; lines are used to show distance between launch sites and coastaline, railway and city
- GitHub URL of Folium map: *https://github.com/miao678/IBM-data-science-capstone-project/blob/main/lab_jupyter_launch_site_location.ipynb*

Build a Dashboard with Plotly Dash

- **Added dropdown list for launch sites**

  Users are able to select plots for specific launch sites as well as all launch sites

- **Added pie charts for success rates**

  Users are able to get a general impression on success rates for launch sites

- **Added payload range slider**

  Users are able to select payload range to see changes in mission outcome

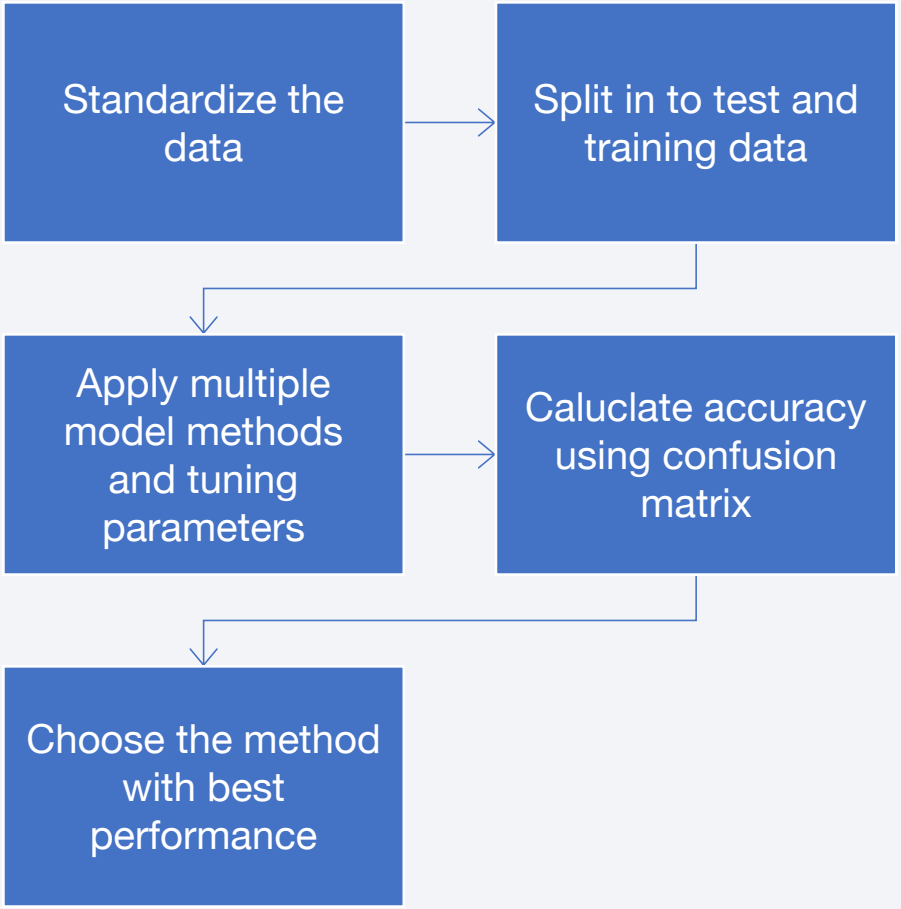
- **Added scatter plot between payload range and mission outcome**

  Users can check if there are correlation between the two factors

- Github URL of SpaceX Dash App: *https://github.com/miao678/IBM-data-science-capstone-project/blob/main/spacex_dash_app.py*

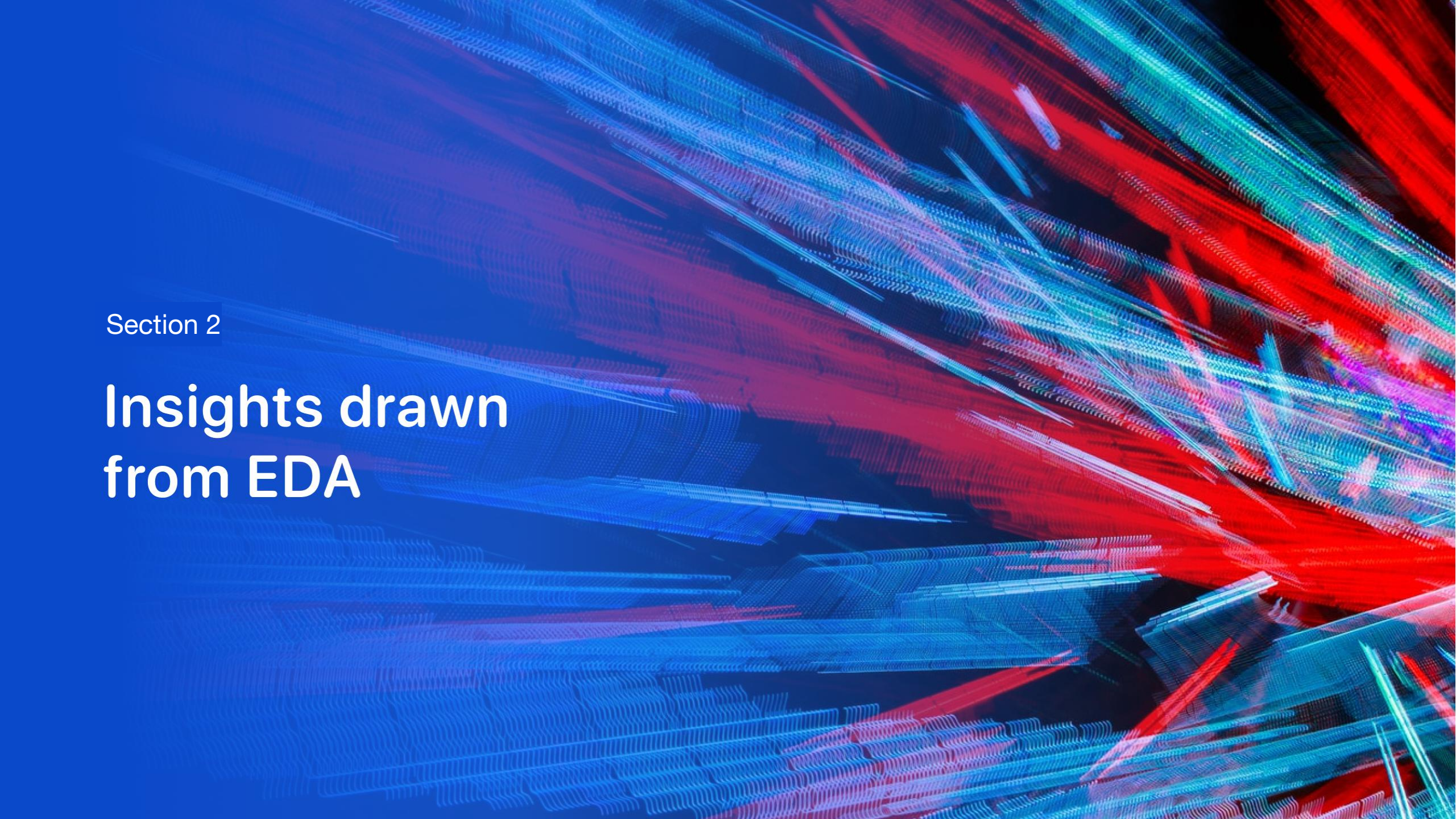
Predictive Analysis (Classification)

- Create numpy array and standardize the data; Split the data into test and training set; Try different models with multiple parameters and find the parameter set with best model performance based on confusion matrix; Choose the method that performs the best
- GitHub URL of predictive analysis: https://github.com/miao678/IBM-data-science-capstone-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

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

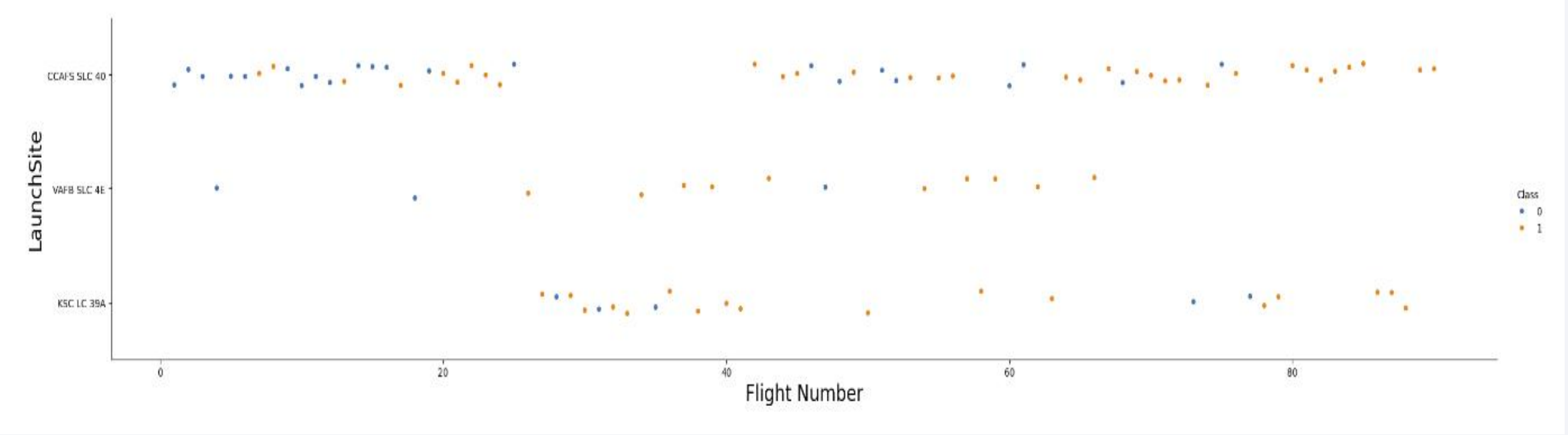


Section 2

Insights drawn from EDA

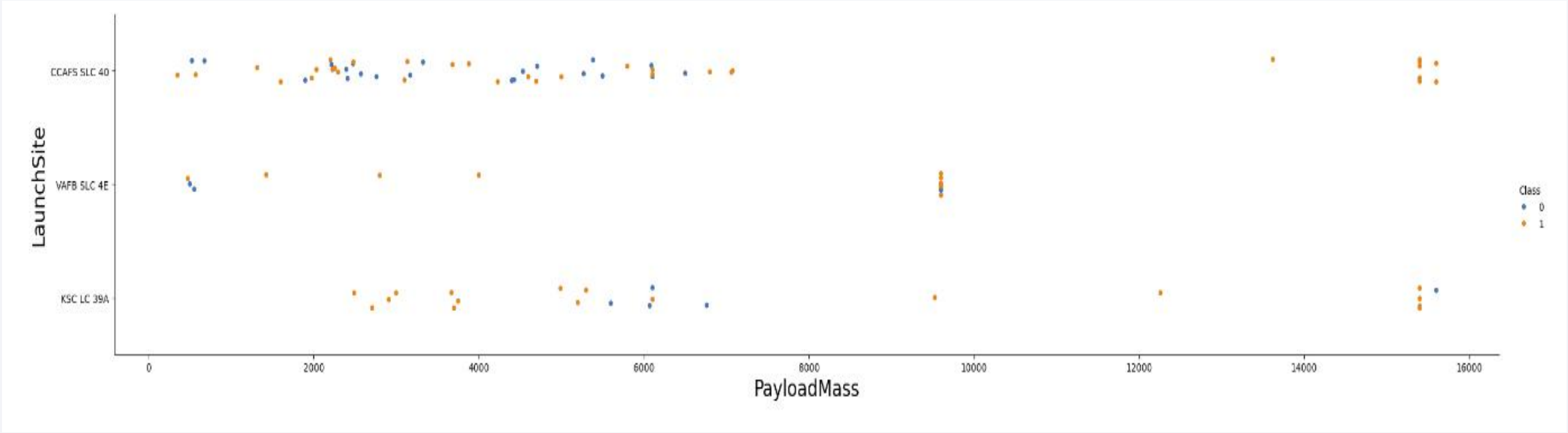
Flight Number vs. Launch Site

- For launch site of CCAFS SLC 40 and VAFB SLC 4E, larger Flight Number has a higher success rate; but for launch site of KSC LC 39A, the association between Flight Number and success rate is not obvious.



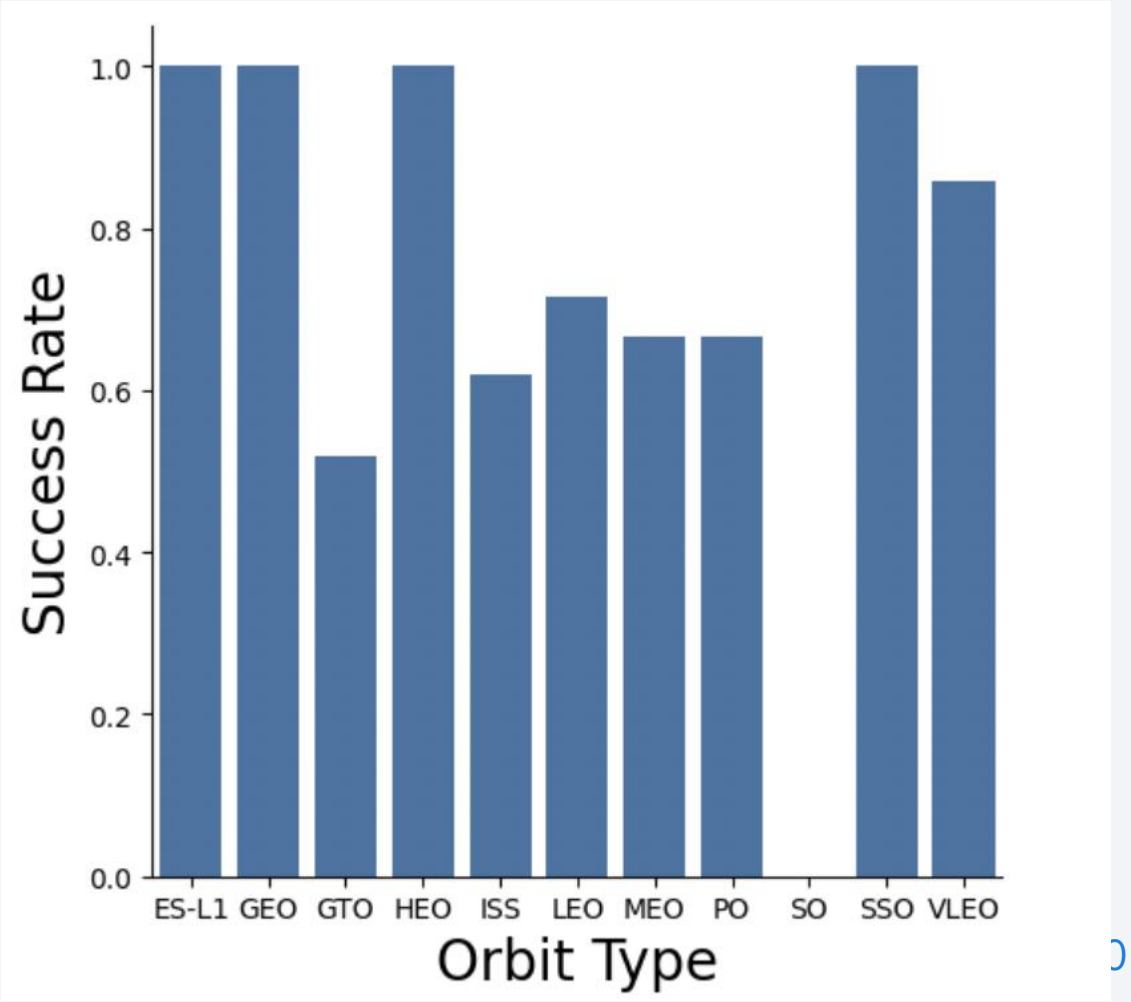
Payload vs. Launch Site

- For the launch site VAFB SKC 4E, there are no rockets launched for payload mass greater than 10000.
- There is not an obvious relationship between payload mass and success rate.



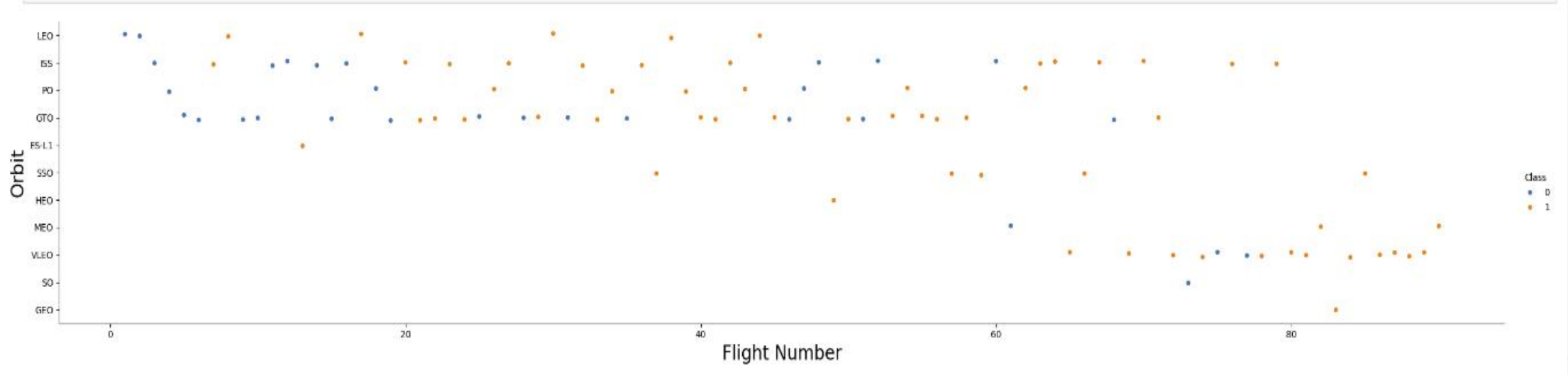
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have the highest success rate
- SO has success rate at 0%



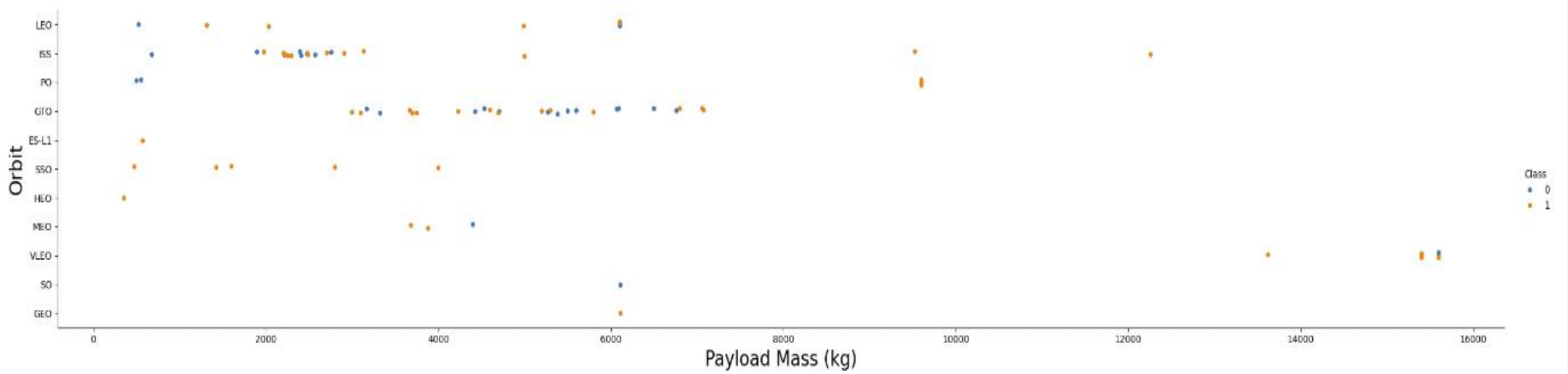
Flight Number vs. Orbit Type

- In the LEO orbit, the success rate is higher as the Flight Number increases
- In the GTO orbit, the relationship between success rate and Flight Number is not obvious



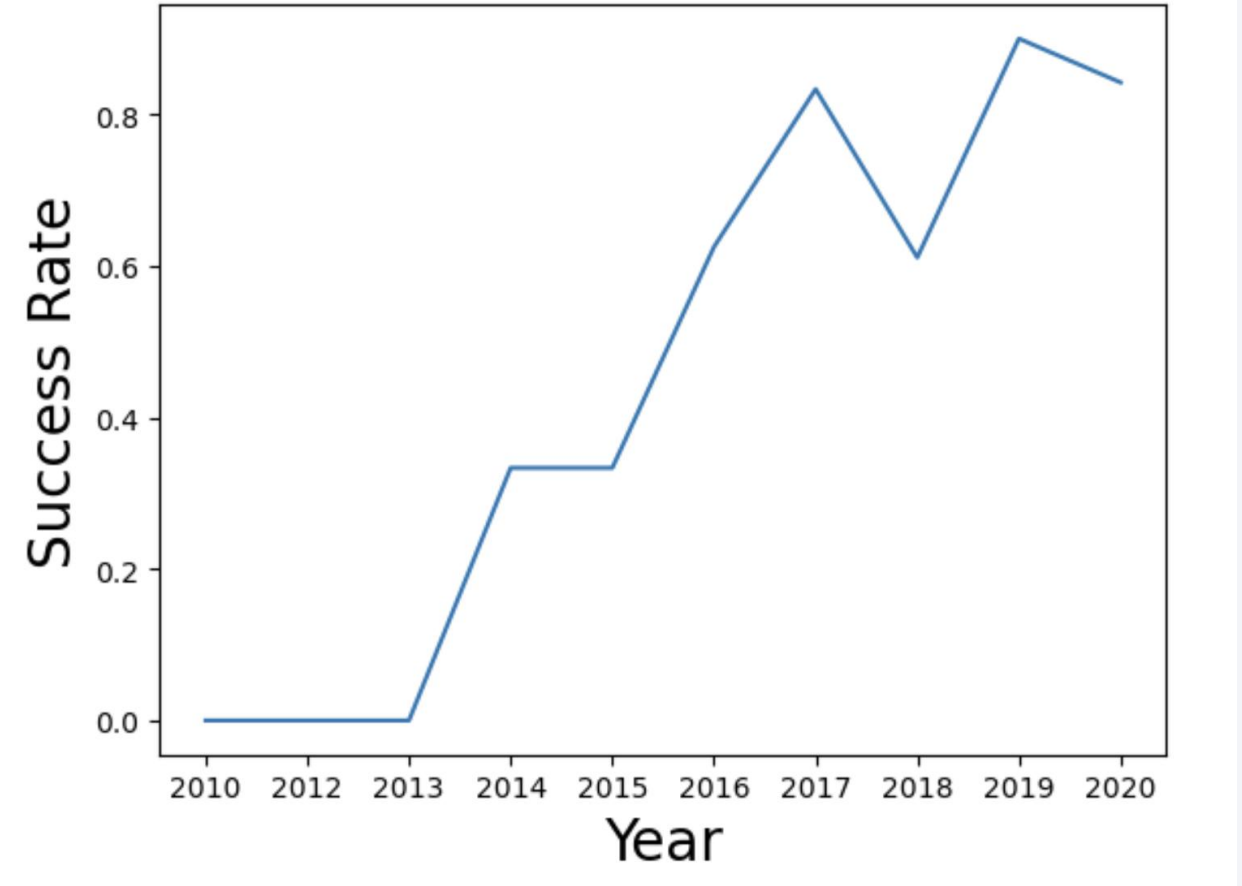
Payload vs. Orbit Type

- For orbits LEO, ISS and PO, higher success rate is related to higher payload mass
- This replationship is not obvious in other orbit type



Launch Success Yearly Trend

- The success rate increases as year goes from 2010 to 2020



All Launch Site Names

- Unique names of the launch sites are shown in the result

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

- 5 records with launch site names starting with 'CCA'

In [11]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Out[11]:

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

- The total payload mass carried by boosters launched by NASA (CRS) is 111268.

```
In [12]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
 * sqlite:///my_data1.db
Done.
Out[12]: TOTAL_PAYLOAD
         111268
```

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2928.4.

```
In [13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
Out[13]: AVG_PAYLOAD
         2928.4
```

First Successful Ground Landing Date

- The first successful ground landing date is 2015-12-22.

```
In [16]: %sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'S
 * sqlite:///my_data1.db
Done.
Out[16]: FIRST_SUCCESS_GP
         2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of booster version that have success in drone ship with payload between 4000 and 6000 are listed below

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes are shown

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOM
```

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

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

Boosters Carried Maximum Payload

- The names of the booster_versions that carried the maximum payload mass are listed.

```
In [19]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM
 * sqlite:///my_data1.db
Done.
```

Out[19]:

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- The month names and failure landing outcomes in drone ship, booster version and launch site for months in 2015 are listed.

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] FROM SPACEXTBL where
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- The counts of landing outcomes between 2010-06-04 and 2017-03-20 are listed

```
%sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROU
```

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

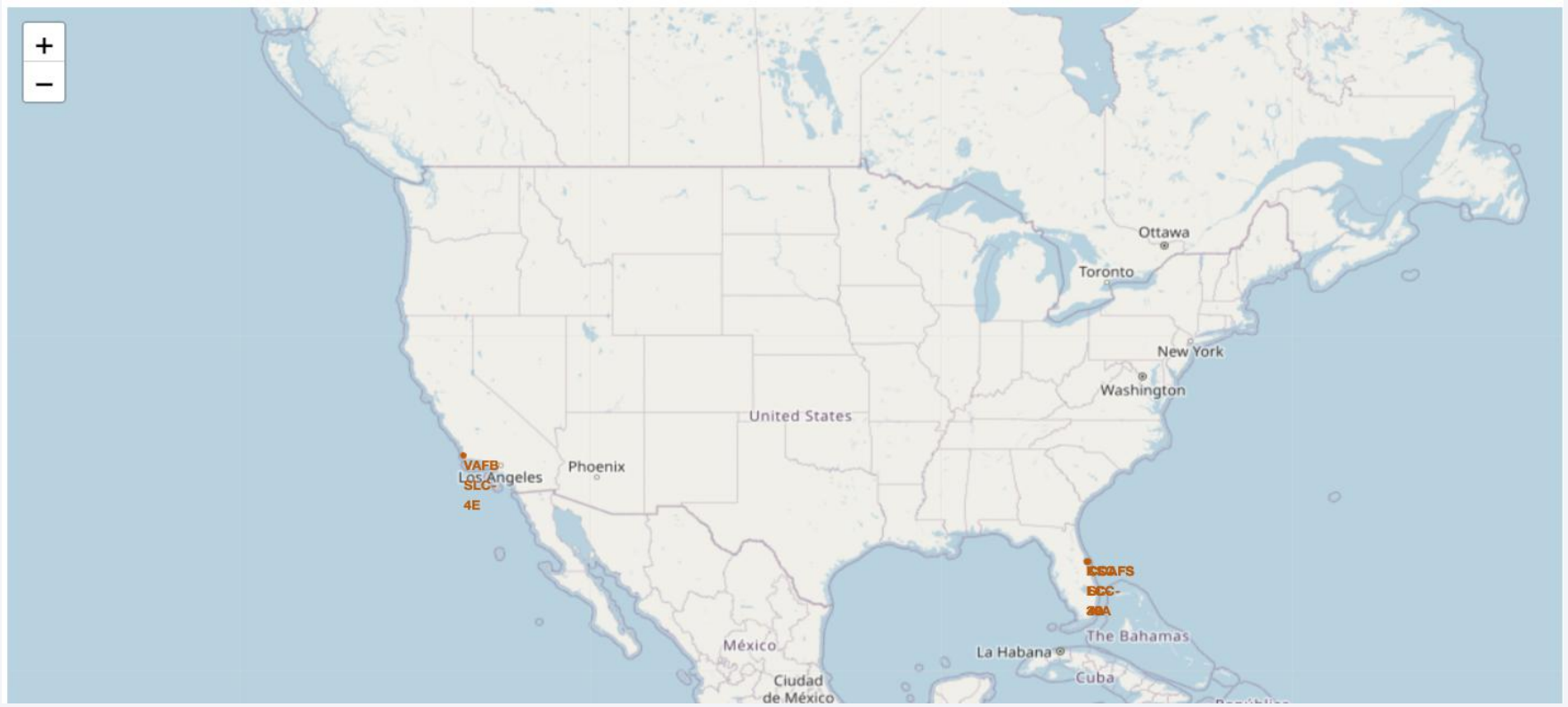
Landing_Outcome	QTY
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

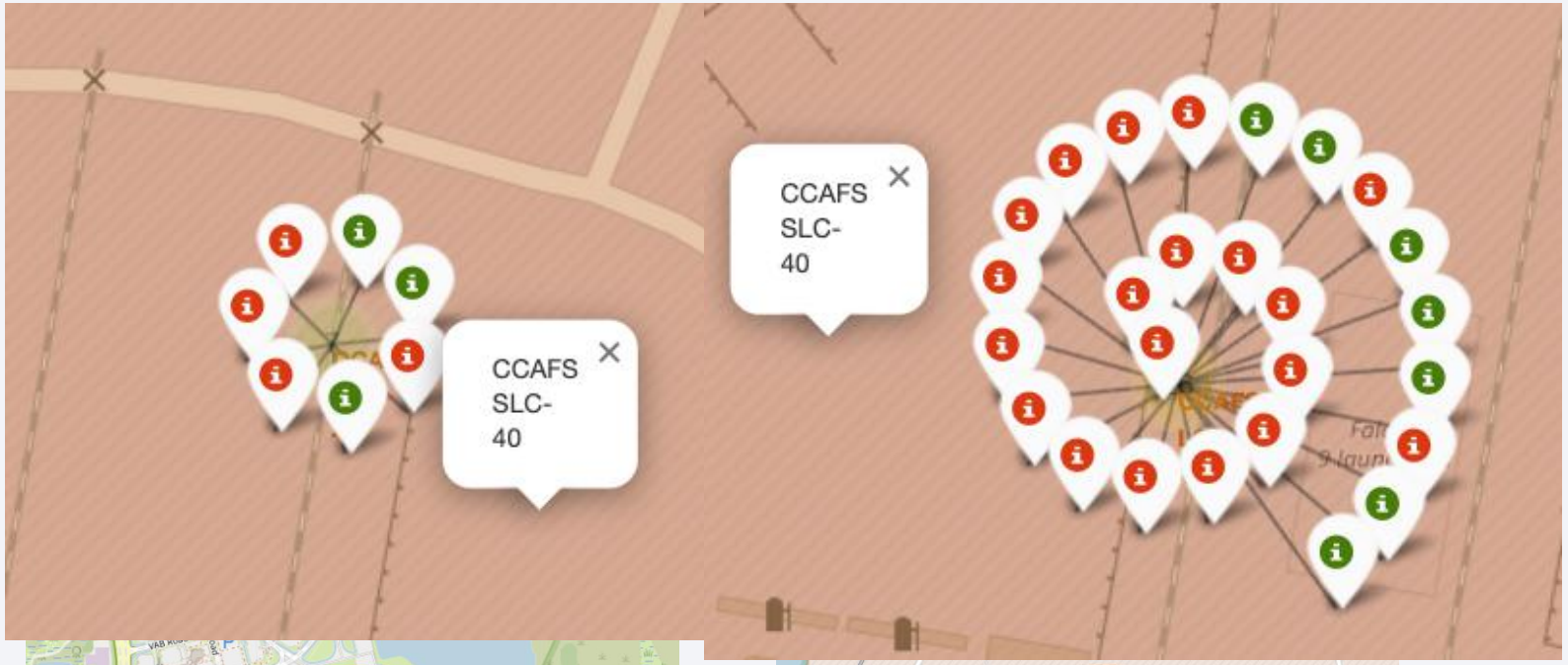
Launch sites locations with markers

- Each launch site is added a circle and labed with launch site names
- All launch sites are near the equator line and they are all very close to the coast

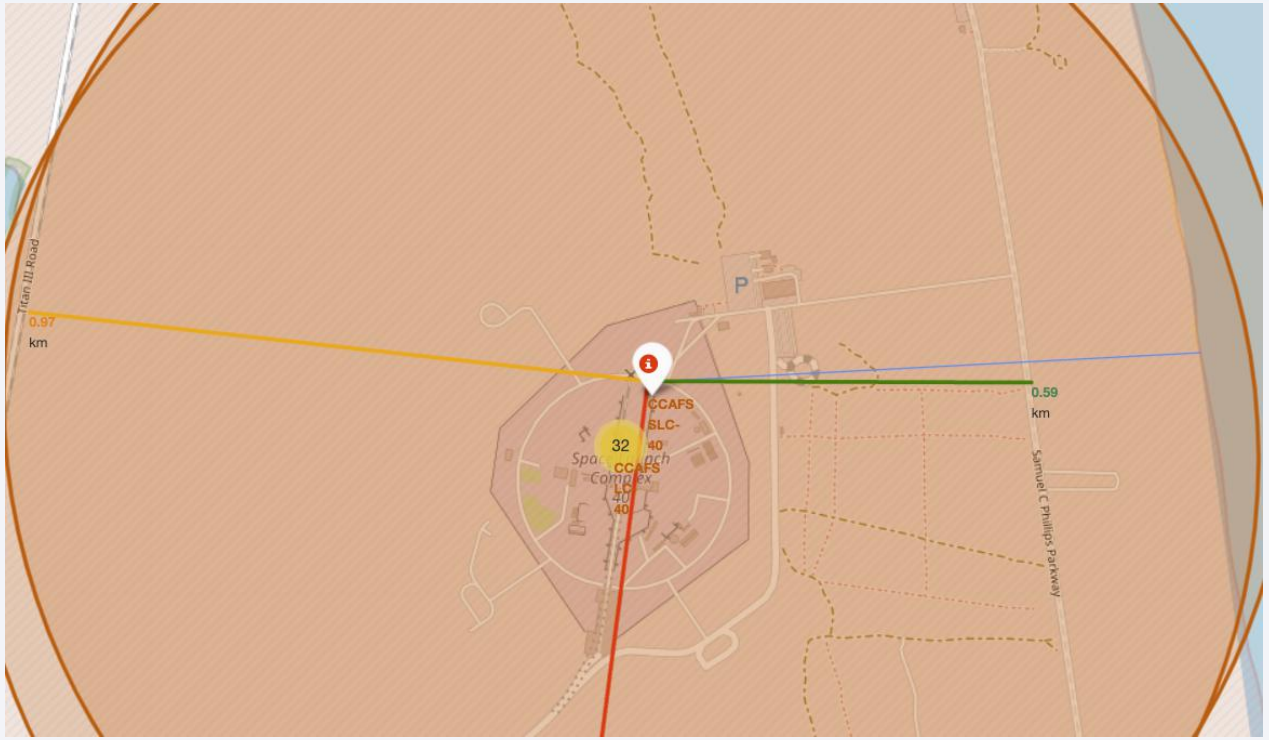


Launch Outcomes

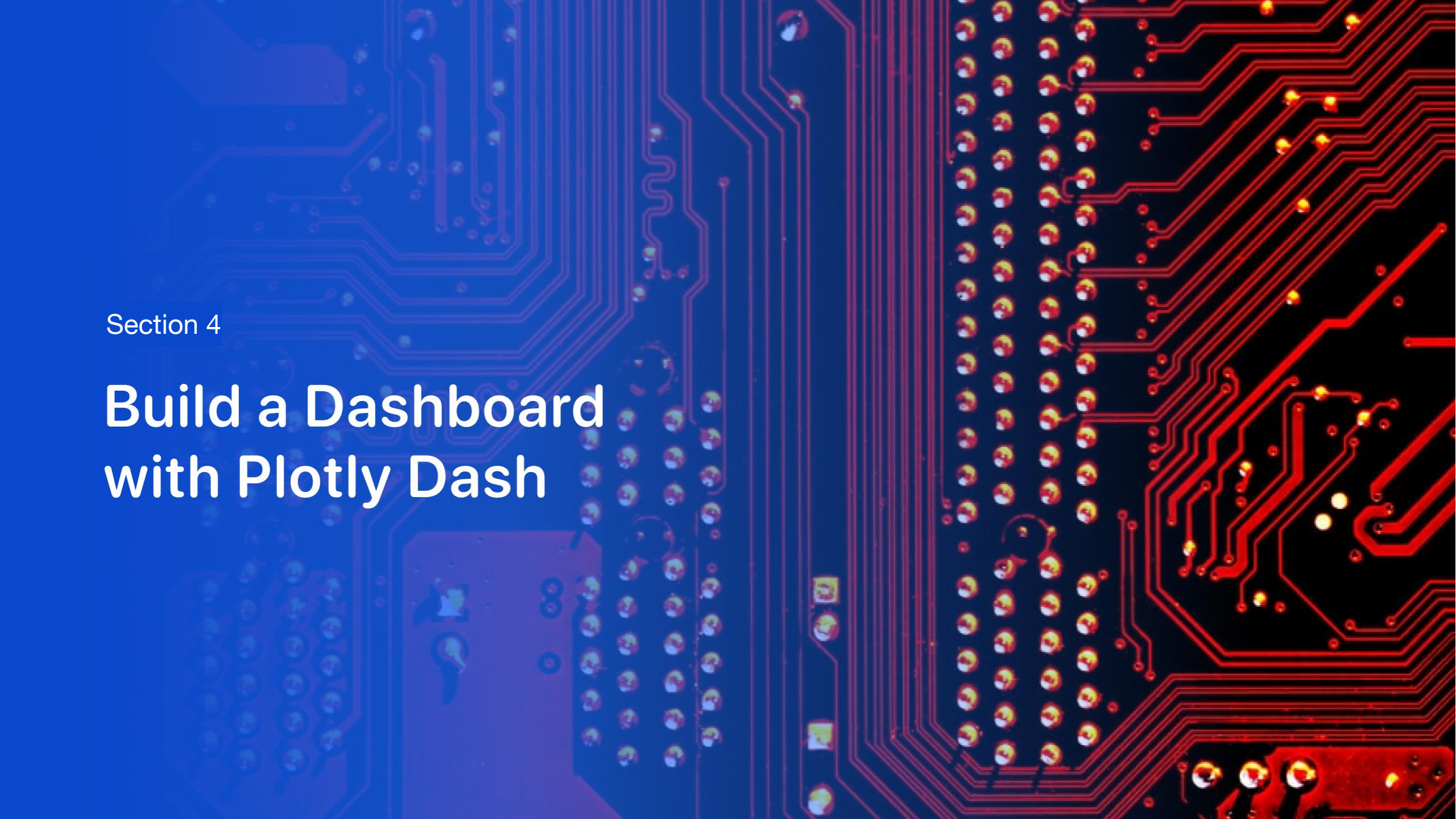
- Red markers mean successful launches
- Green markers mean unsuccessful launches
- KSC LC-33A has relatively high success rates



<Folium Map Screenshot 3>

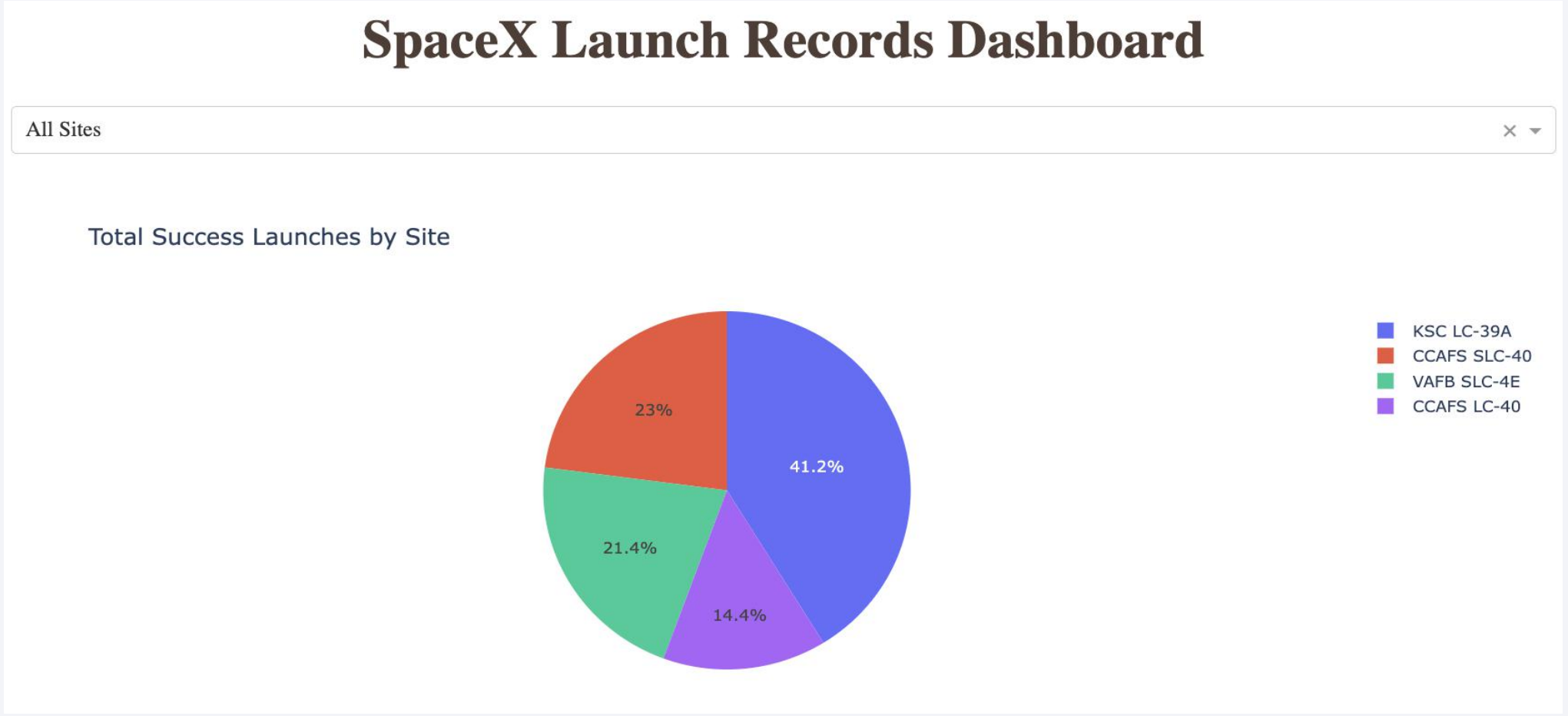


- The CCAFS SLC-40 is close to railways, highways and coastline; but not close to cities.


Section 4
Build a Dashboard with Plotly Dash

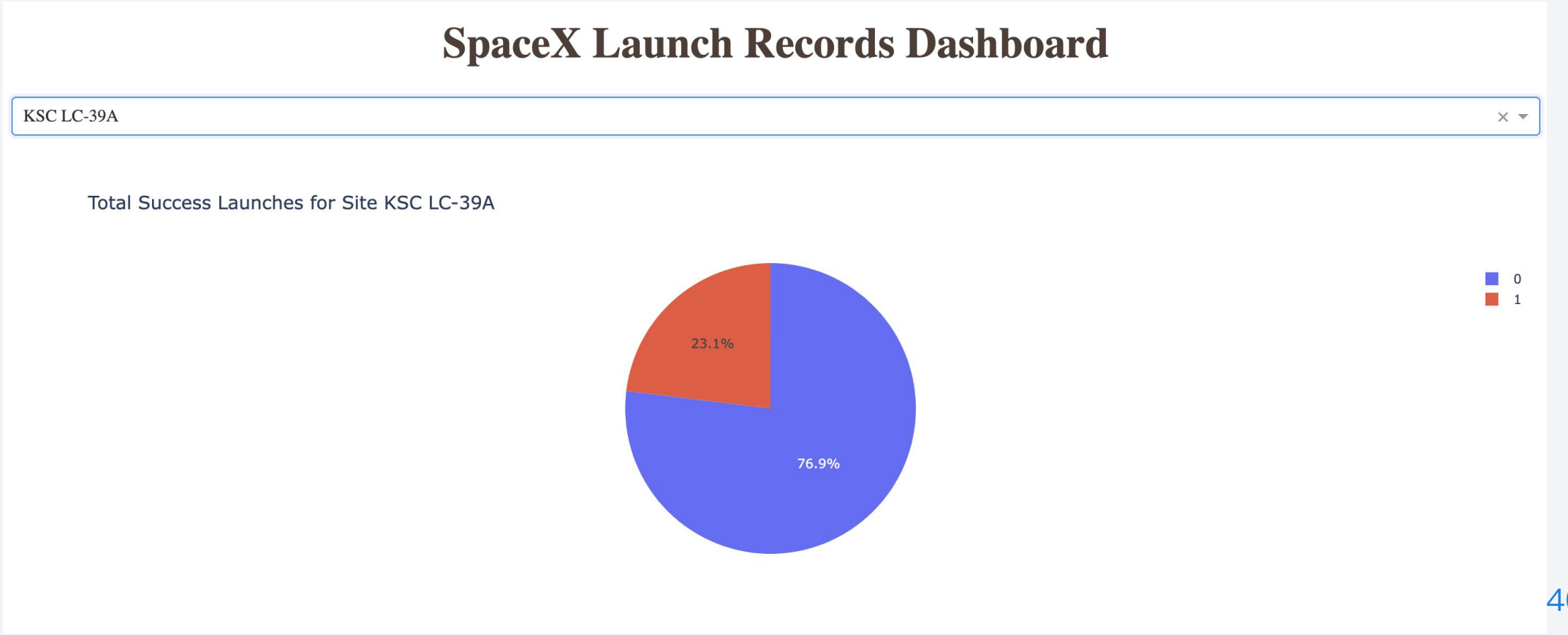
Total success launches by site

- KSC LC-39A has the most success launches

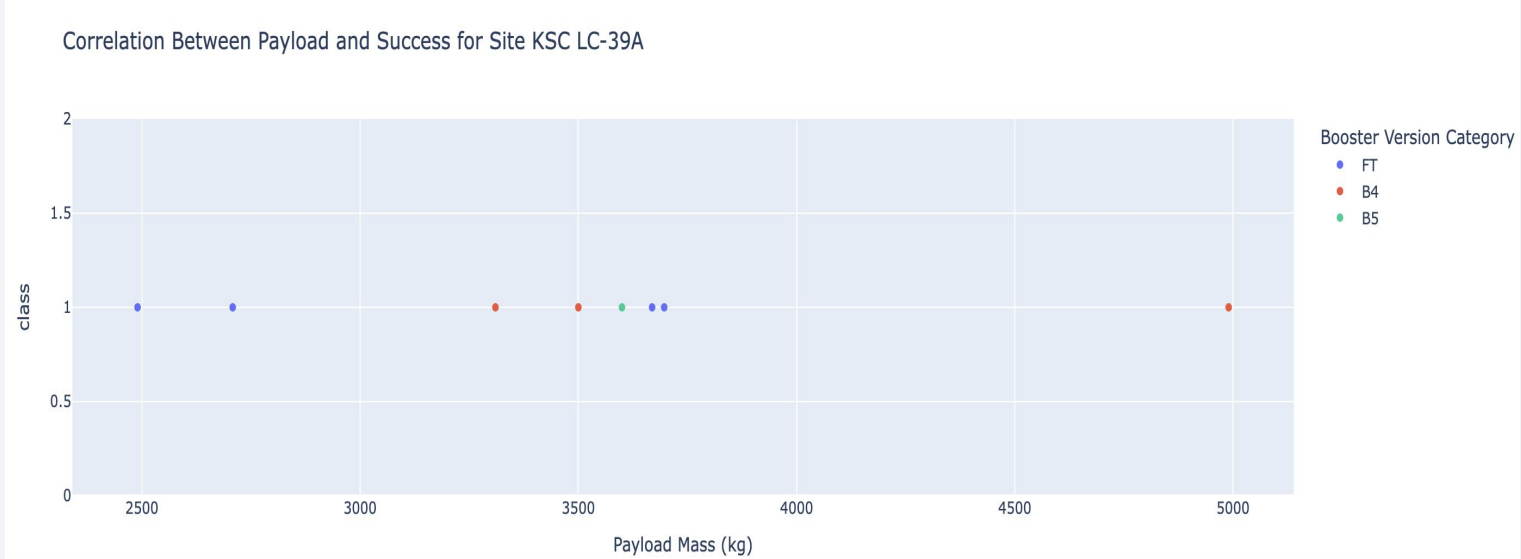


The piechart for the launch site with highest launch success ratio

- KSC LC-39A launch site has success rate of 76.9%

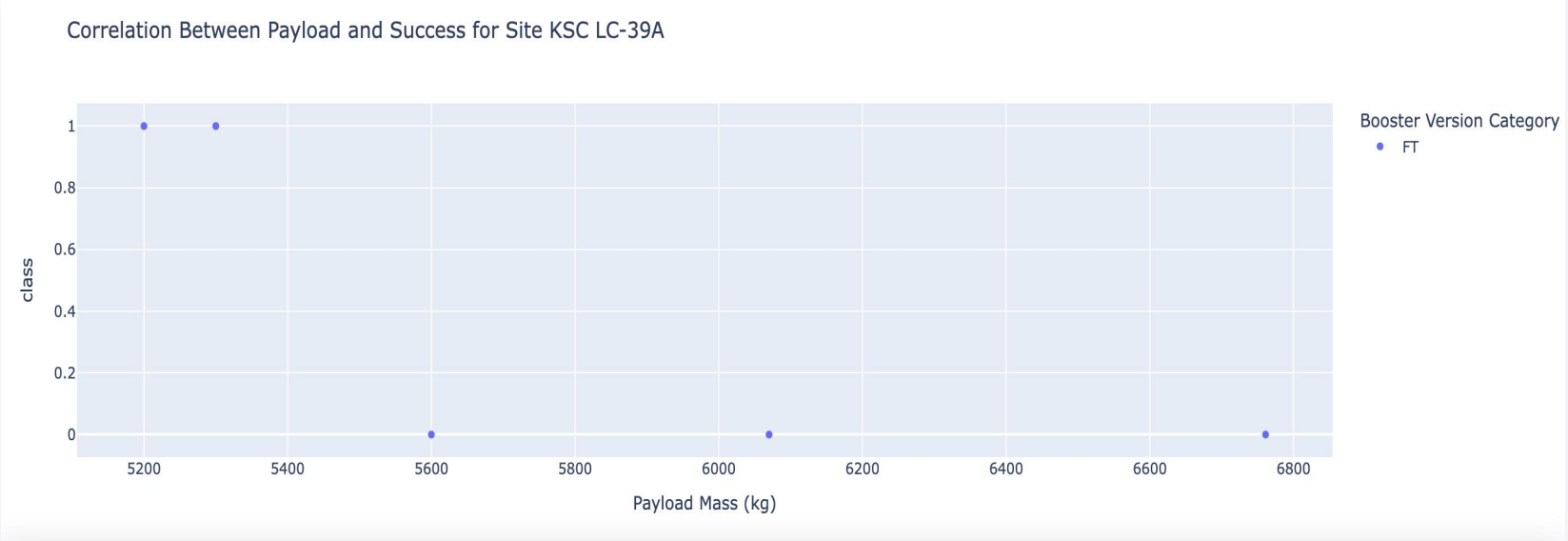


Payload vs. Launch Outcome scatter plot for all sites, with different payload



- FT and B4 booster version has the most success in payload between 2500 and 5000 kg

- Higher payload is related to less successful launches



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The DecisionTree classifer has the highest model accuracy

Find the method performs best:

In [32]:

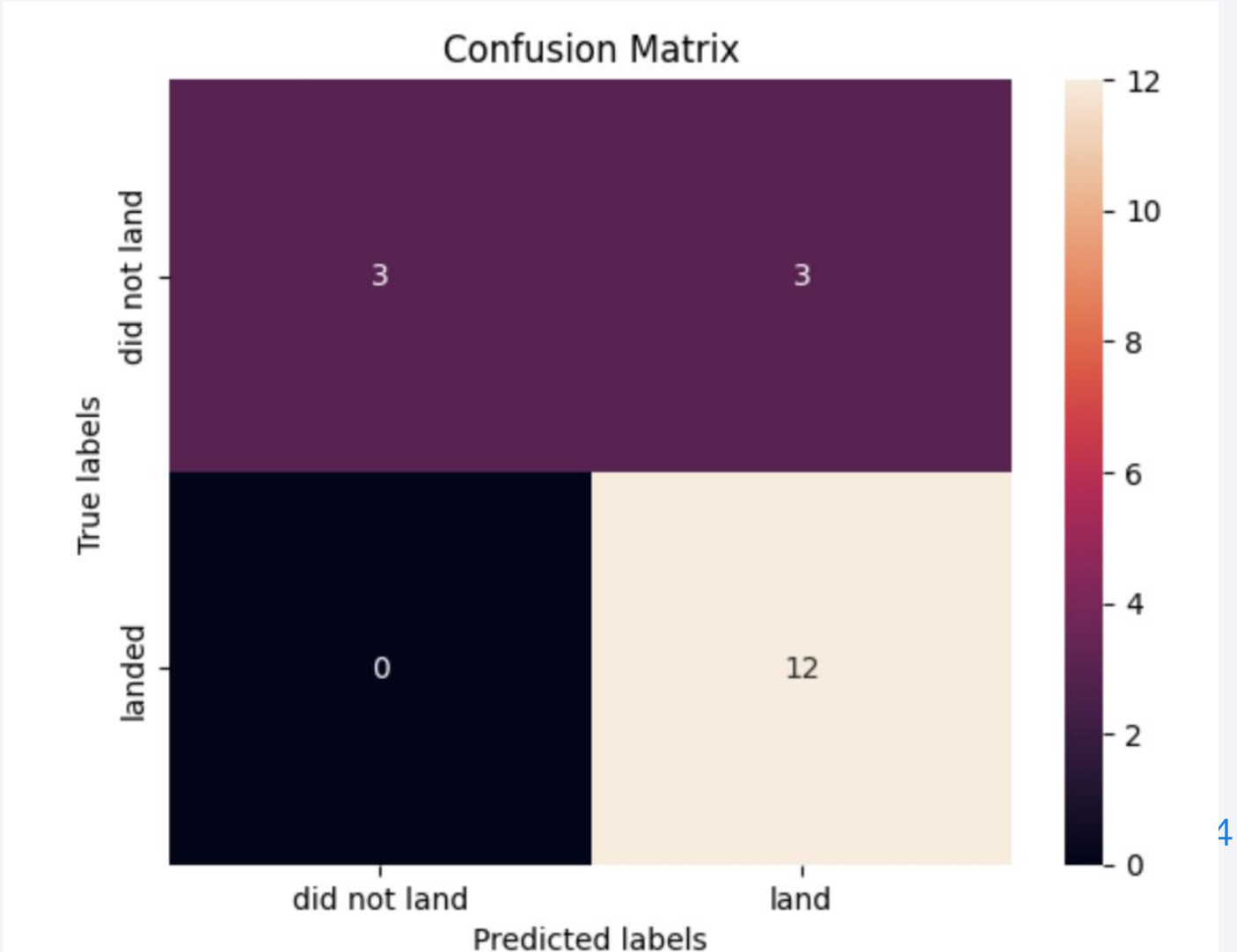
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8767857142857143
Best params is : {'criterion': 'entropy', 'max_depth': 8, 'max_features': 'sqrt', 'mi
n_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- The confusion matrix with the most success is shown
- Precision is 0.8 in the confusion matrix



Conclusions

- Flight number and launch success rate has a positive relationship
- ES-L1, GEO, HEO, and SSO have the highest success rate
- The success rate is increasing from 2010 to 2020
- All launch sites are close to equator line and coast; they are not close to cities
- KSC LC-39A has the most success launches
- FT and B4 booster version has the most success in payload between 2500 and 5000 kg
- DecisionTree has the best model performance

Appendix

- Space X dataset

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.561857
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1006	-80.577366	28.561857
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1007	-80.577366	28.561857
8	9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1008	-80.577366	28.561857
9	10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1011	-80.577366	28.561857

Thank you!