IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Edward Zhao>
<Mar 16 2022>

Outline

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

Executive Summary

- The methodology of this project can be split into 5 parts. First, collecting data from SpaceX API and Falcon9 Wiki page. Second, performing data wrangling to deal with missing data and determine labels for prediction. Third, practicing exploratory data analysis by SQL and dashboard to have some intuitive impression on the dataset. Fourth, conducting visual analysis with dashboard and folium map. Finally, building predictive models to carry out quantitively analysis by classification.
- The results indicates that a combination of greater flight number and payload mass with launch site KSC LC-39A, the rocket running on orbits ES-L1, GEO, HEO, SSO, and a payload range from 2000 to 7000 will lead to a higher success rate of landing.

Introduction

- SpaceX is offering to launch rockets with a cost of $62 million, while other providers charge a significantly higher price. This is because SpaceX can reuse the first stage. Whether the first stage will land or not directly affects the cost of the rocket launch.

- Therefore, in this capstone, we will analyze whether the first stage of Falcon 9 will land successfully by evaluating a series of factors.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and Falcon9 Wiki page
- Perform data wrangling
  - Fill the missing values and determine training labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, fit, and evaluate classification models

Data Collection

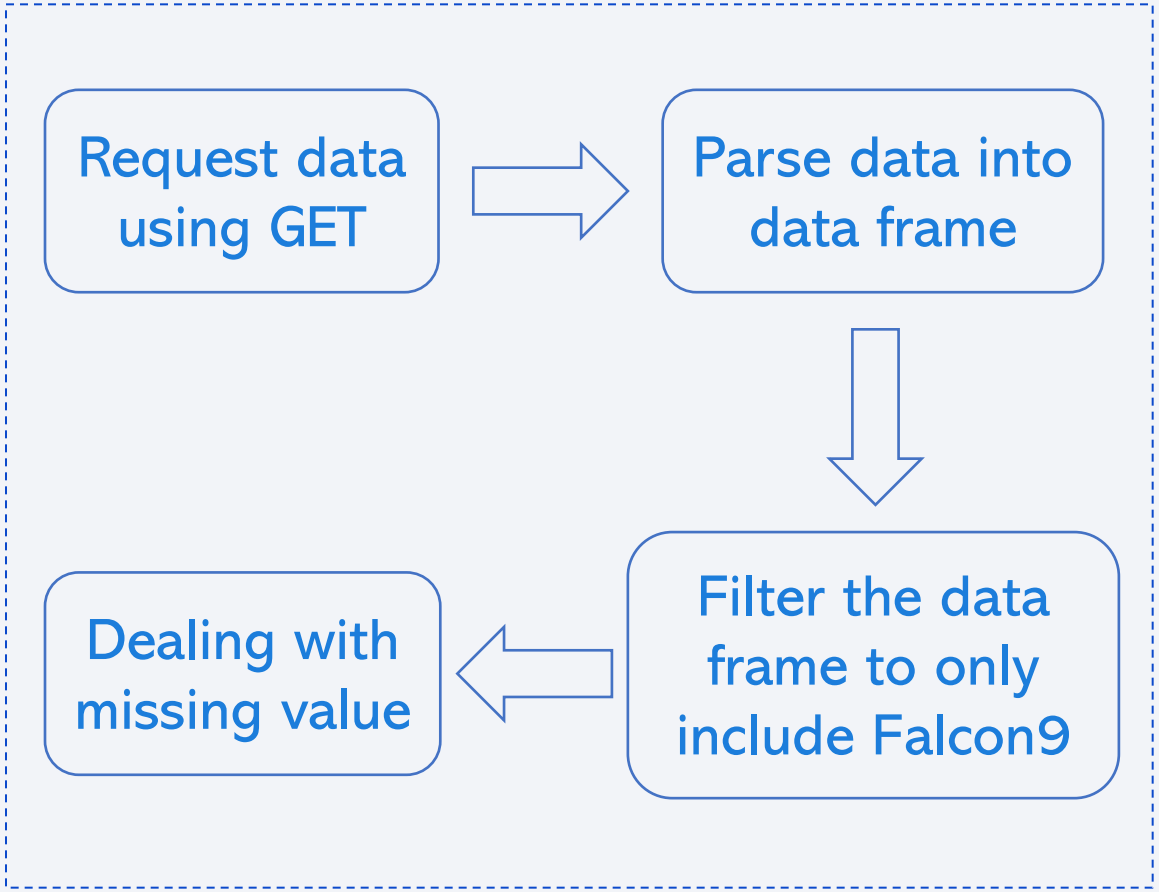
- Data sets are collected from SpaceX API and Falcon9 Wiki page
- There are two major steps for data collection. More details are given in the following slides

Request data from SpaceX API

Scrape data from Falcon9 Wiki page

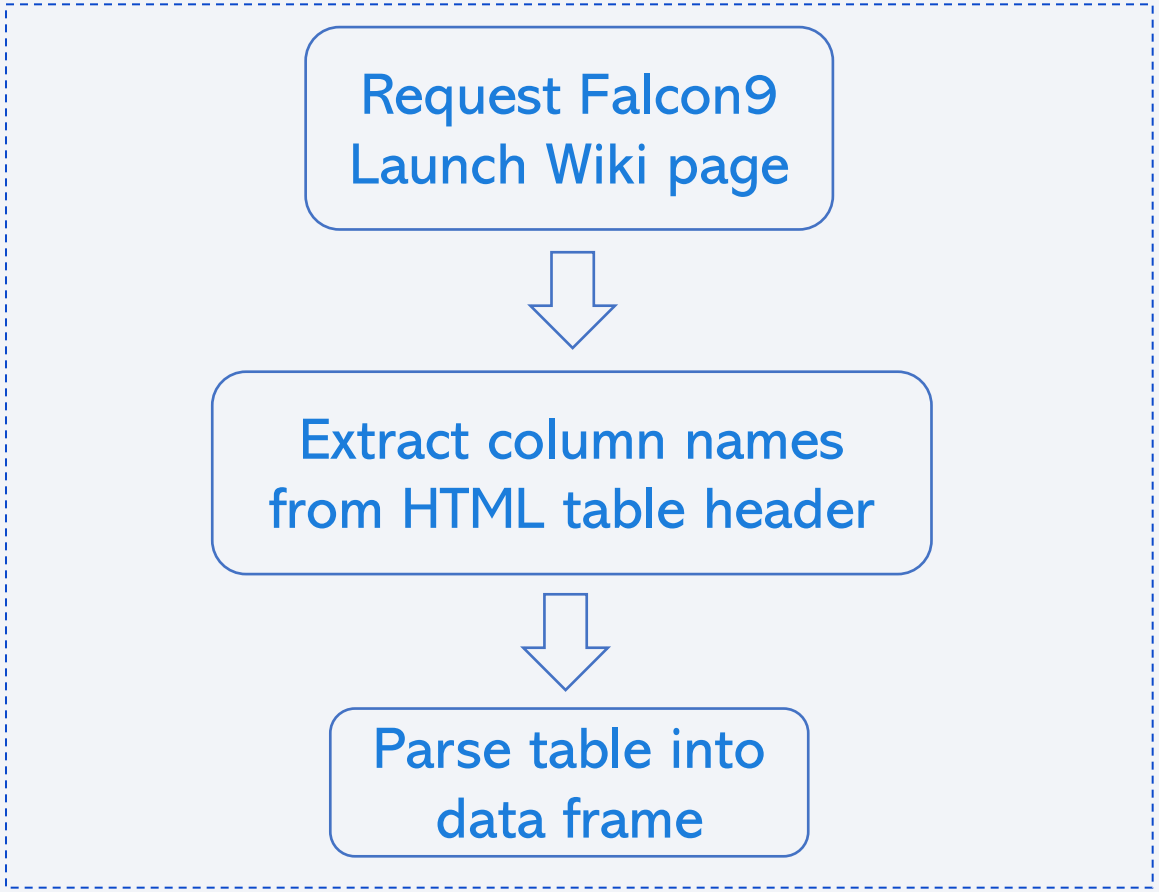
Data Collection – SpaceX API

- Request data
- Parse data
- Filter data
- Missing value
- https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/Lab1.1%20Data%20collection%20API.ipynb



Data Collection - Scraping

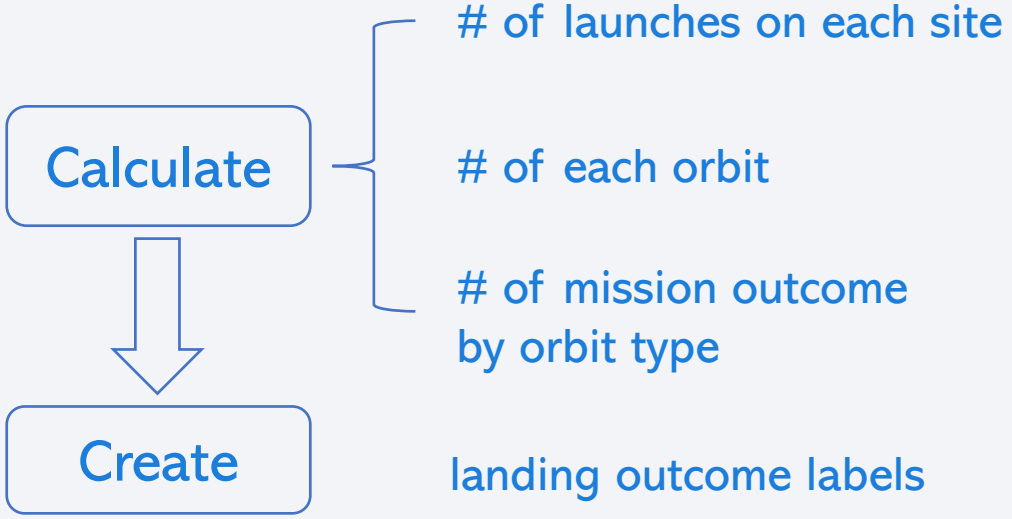
- Request Falcon9 Wiki page
- Extract data from table
- Parse table into data frame
- https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/Lab1.2%20Web%20scraping.ipynb



Data Wrangling

- In this process, I performed exploratory data analysis and determined training labels

- Calculate
  - # of launches on each site
  - # of each orbit
  - # of mission outcome by orbit type
- Create landing outcome labels



- https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/Lab1.3%20Data%20wrangling.ipynb

EDA with Data Visualization

- Three types of graphs are plotted: scatter plot, bar plot, line plot
  - Scatter plot finds out whether two variables have a relationship or correlation
  - Bar plot compares landing outcomes by different orbits
  - Line plot reveals the trend of landing outcome across years
- https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/Lab2.2%20EDA%20with%20data%20visualization.ipynb

EDA with SQL

- SQL queries performed
  - Names of unique launch sites
  - 5 launch sites begin with “CCA”
  - Total payload mass launched by NASA (CRS)
  - Average payload mass by version F9 v1.1
  - Date when 1st successful landing in ground pad
  - Boosters with success in drone ship & payload mass between 4000 and 6000
  - Total # by mission outcome
  - Booster versions with max payload mass
  - Launch site & booster version failed in drone ship in 2015
  - Rank the landing outcome between 2010-06-04 and 2017-03-20
- https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/Lab2.1%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- The following map objects are created and added to a folium map: markers, marker clusters, circles, lines, mouse location
  - Markers: add markers to the map with popup message
  - Marker clusters: add a cluster of markers to the map while avoiding overlapping
  - Circles: add circles to the specified coordinate on the map
  - Lines: add lines to connect two specified points on the map
  - Mouse location: reflects the real-time location of the mouse on the map
- https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/Lab3.1%20Interactive%20visual%20analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- There are two plots added to the dashboard: pie plot and scatter plots
  - Pie plots: reflects the successful outcomes by launch sites. It also exhibits the success rate of each launch site
  - Scatter plot: indicates the relationship between payload mass and landing outcome
- https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/Lab3.2%20Interactive%20dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

- I built the model with preprocessed data by fitting data into different machine learning models. The models are evaluated by the score() method with accuracy as its indicator and confusion matrix. The best performing model is found by comparing different models and their accuracy scores.
- There are 3 major steps: preprocessing, fitting, evaluation

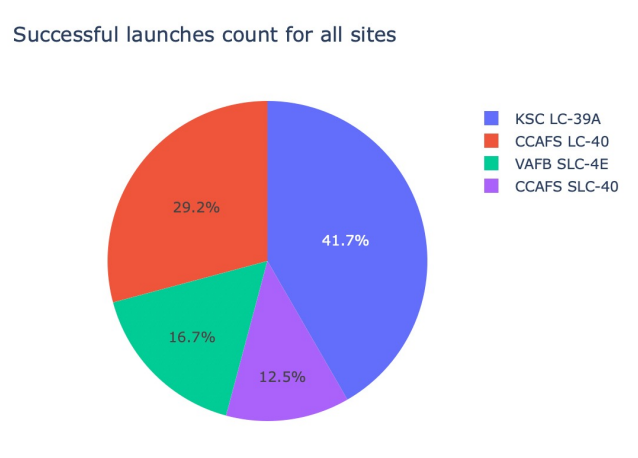
- https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/Lab4%20Machine%20learning%20prediction.ipynb

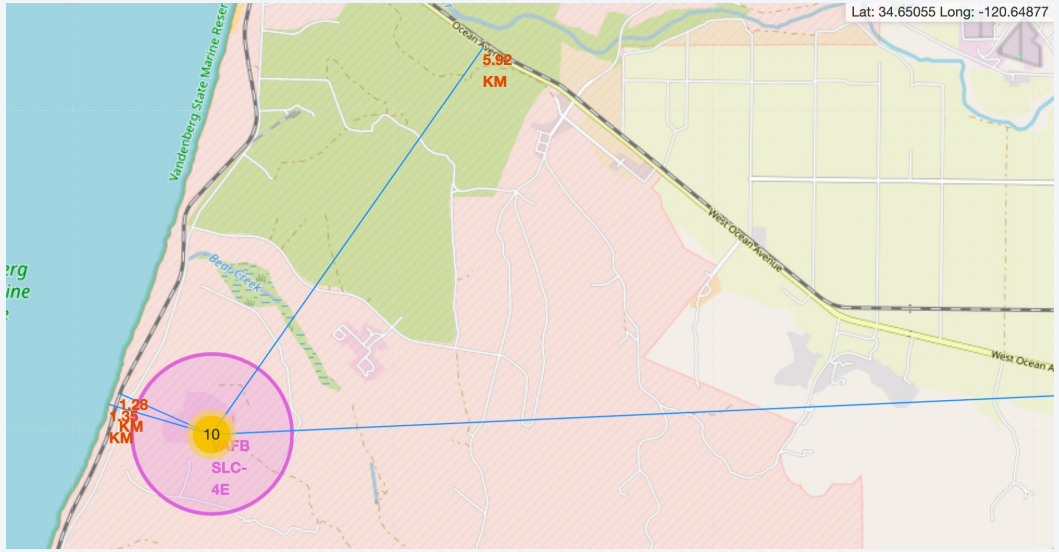
Results

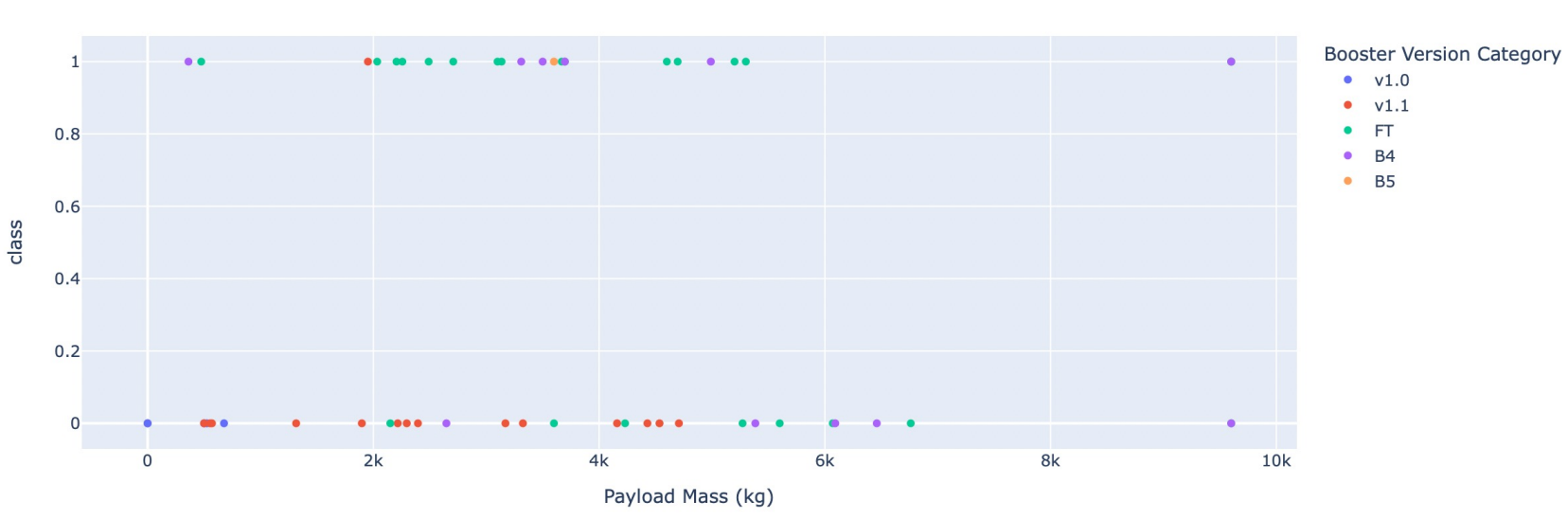
- The exploratory data analysis shows that
    1. KSC LC 39A and VAFB SLC 4E have a higher proportion of successful landings than CCAFS SLC 40
    2. The number of failed landings is decreasing as the flight number increases
    3. There are more successful landings as the payload mass increases
    4. The orbit ES-L1, GEO, HEO, SSO has a success landing rate of 100%, while the orbit SO has 0 success case
    5. The success rate gradually increases as the time goes by
- Predictive analysis results
    - The decision tree model has the highest accuracy score of 83.4%, although the output is not stable as it can increase to 91% in different runs.

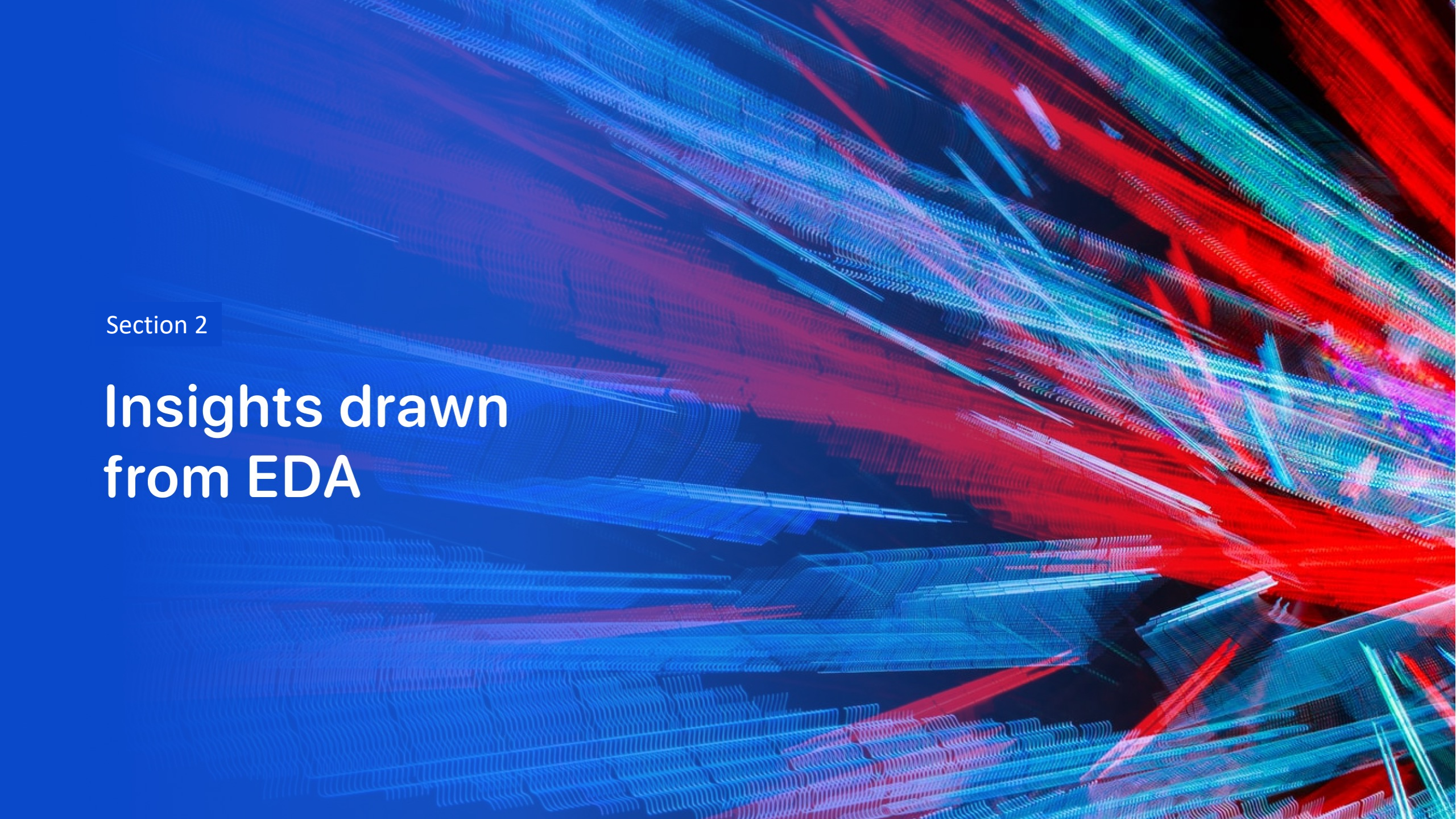
Results

- Interactive analytics demo



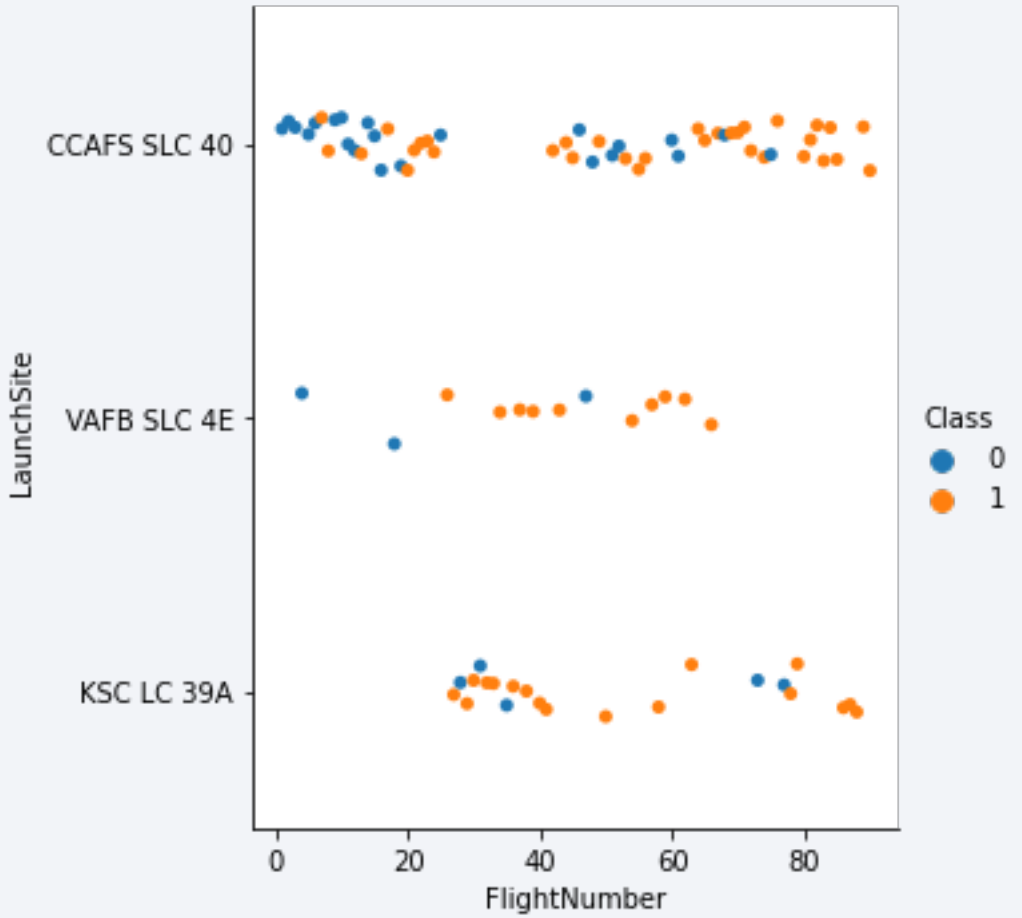





Section 2
Insights drawn from EDA

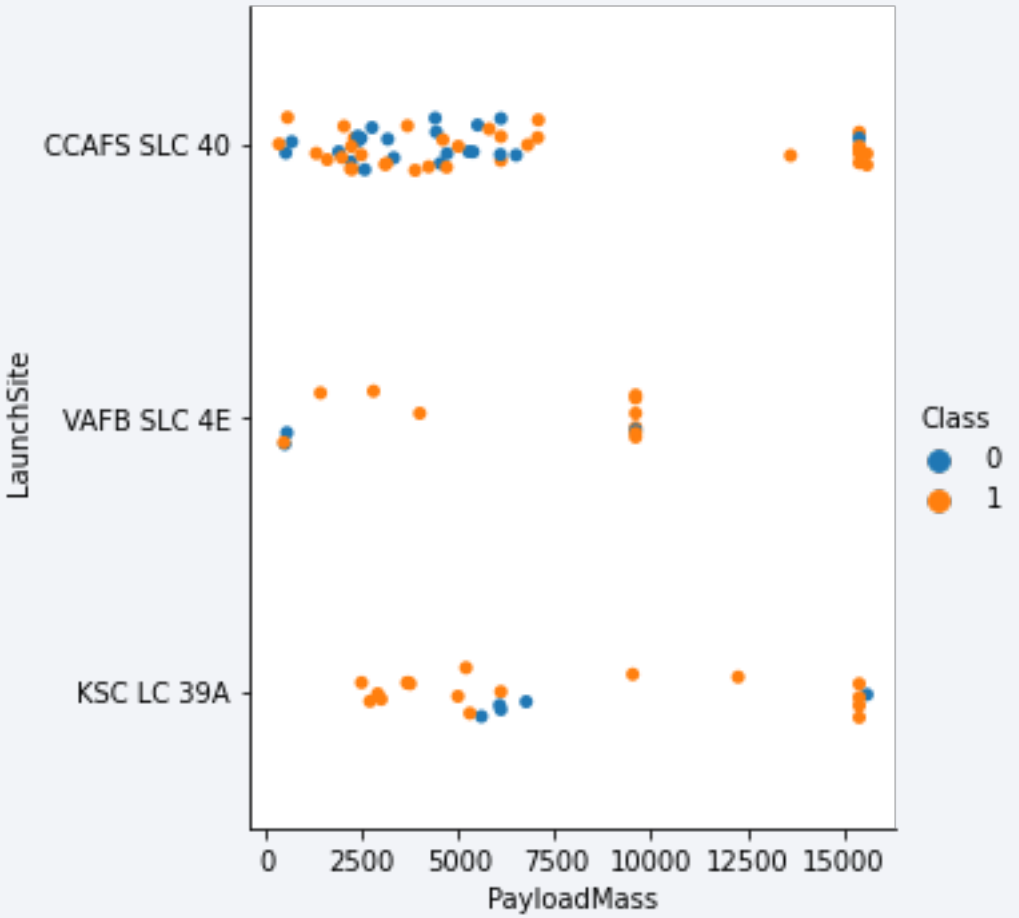
Flight Number vs. Launch Site

The scatter plot indicates that the number of failed cases decreases as the flight number increases, i.e., the two variables are negatively correlated.



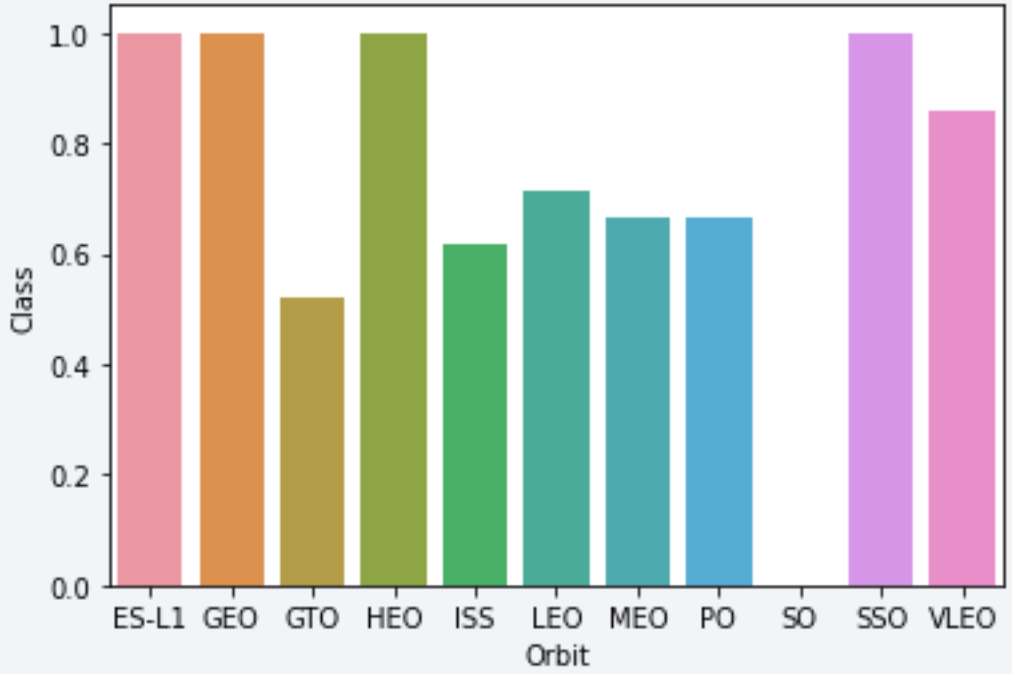
Payload vs. Launch Site

The launch site CCAFS SLC 40 has the relatively low payload mass, while KSC LC 39A almost spreads evenly across the payload mass scale. The site VAFB SLC 4E has most launches with payload mass near 9000 kg.



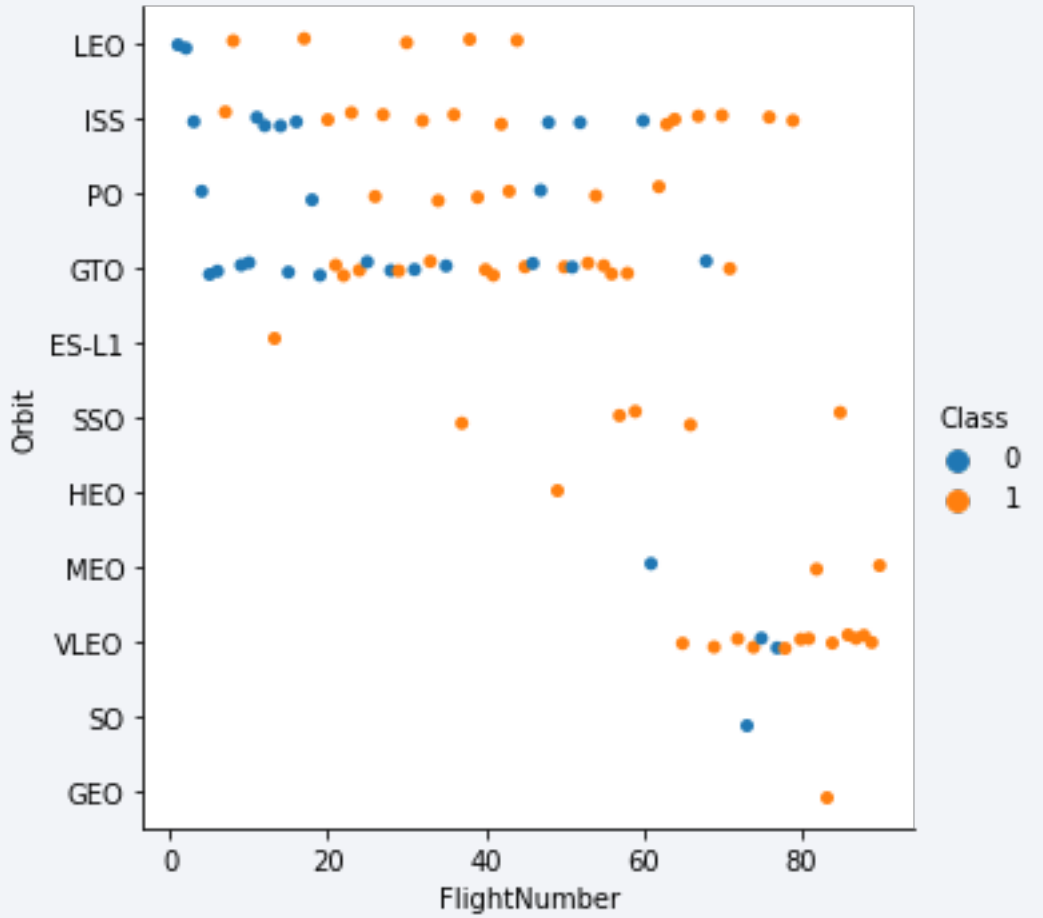
Success Rate vs. Orbit Type

The orbits ES-L1, GEO, HEO, SSO have a success landing rate of 100%, while the orbit SO has 0 success case. The orbit VLEO almost has a success rate of 100% and the others have significantly lower success rates.



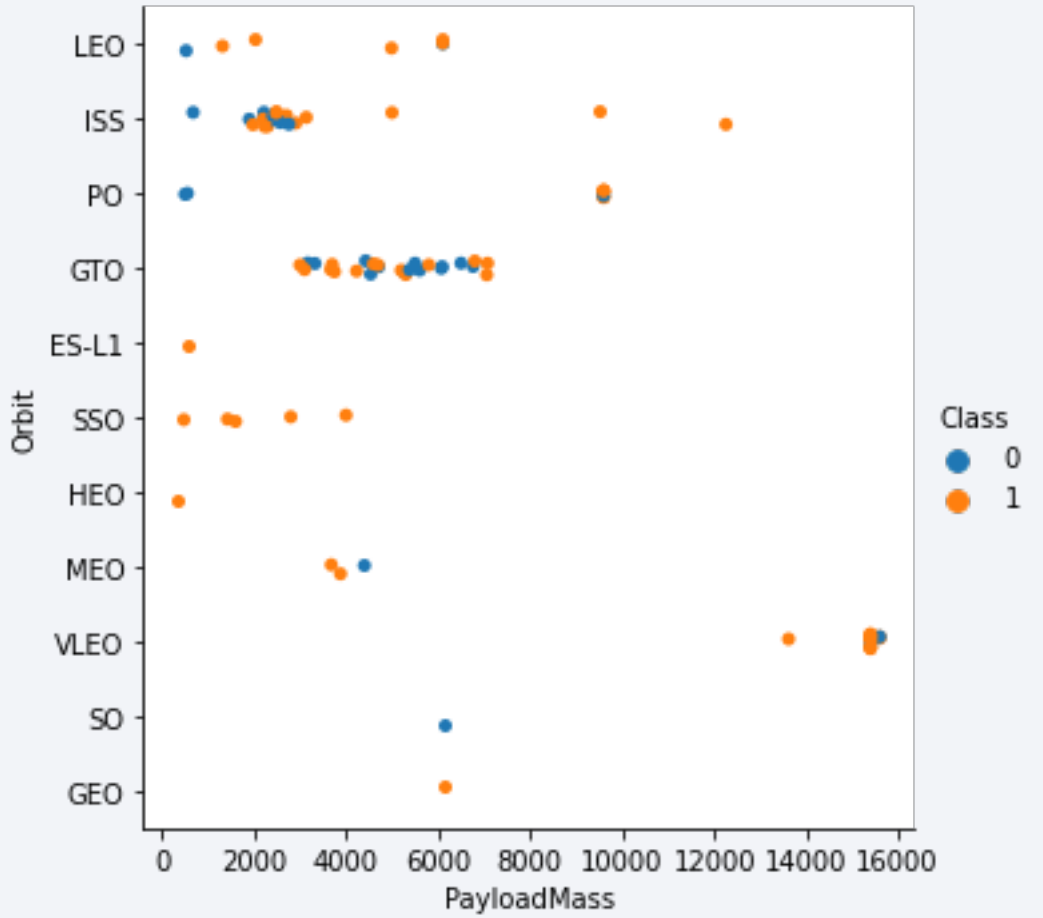
Flight Number vs. Orbit Type

The orbit VLEO has the relatively highest flight number. ISS, PO, and GTO have flight numbers spreading evenly across the scale. LEO has flight number from lowest to medium and SSO has flight number from medium to highest. Other orbits have very limited occurrence.



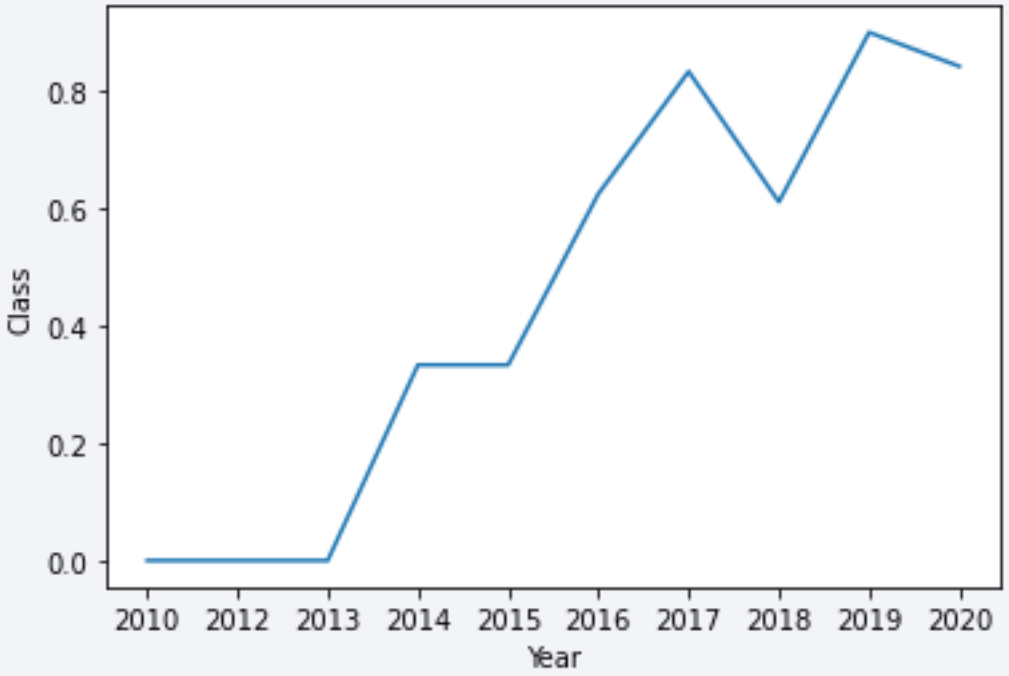
Payload vs. Orbit Type

The orbit VLEO has the highest payload mass. PO and ISS have few occurrence with a payload mass over 9000 kg, while all others orbits only have payload mass below 8000 kg. ISS has a concentration in light payload mass.



Launch Success Yearly Trend

The success rate continuously increases by year except for 2018 and 2020. There is a huge decrease of success rate in 2018 and a slight one in 2020. The success rate is 0 before 2013 implies that the first successful launch happened in 2013.



All Launch Site Names

%sql SELECT UNIQUE(Launch_Site) FROM SPACEXDATASET

The unique launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E

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

Launch Site Names Begin with 'CCA'

%sql SELECT * FROM SPACEXDATASET WHERE Launch_Site LIKE 'CCA%' LIMIT 5

5 records where launch sites begin with `CCA`

column_0	DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	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)
2	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
3	2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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

%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE Customer='NASA (CRS)’

The total payload carried by boosters from NASA is 45596.

Average Payload Mass by F9 v1.1

%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE Booster_Version = 'F9 v1.1’

The average payload mass carried by booster version F9 v1.1 is 2928.

First Successful Ground Landing Date

%sql SELECT MIN(Date) FROM SPACEXDATASET WHERE Landing__Outcome = 'Success (ground pad)'

The dates of the first successful landing outcome on ground pad: 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

%sql SELECT Customer FROM SPACEXDATASET WHERE Landing__Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000

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

- SKY Perfect JSAT Group
- SES
- SES EchoStar

Total Number of Successful and Failure Mission Outcomes

%sql SELECT COUNT(*), Mission_Outcome FROM SPACEXDATASET GROUP BY Mission_Outcome

The total number of successful and failure mission outcomes is:

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

Boosters Carried Maximum Payload

%sql SELECT Booster_Version FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET)

Names of the booster which have carried the maximum payload mass:

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 Booster_version, Launch_Site FROM SPACEXDATASET WHERE Landing__Outcome = 'Failure (drone ship)' AND YEAR(Date) = 2015

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

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

%sql SELECT Booster_Version FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET)

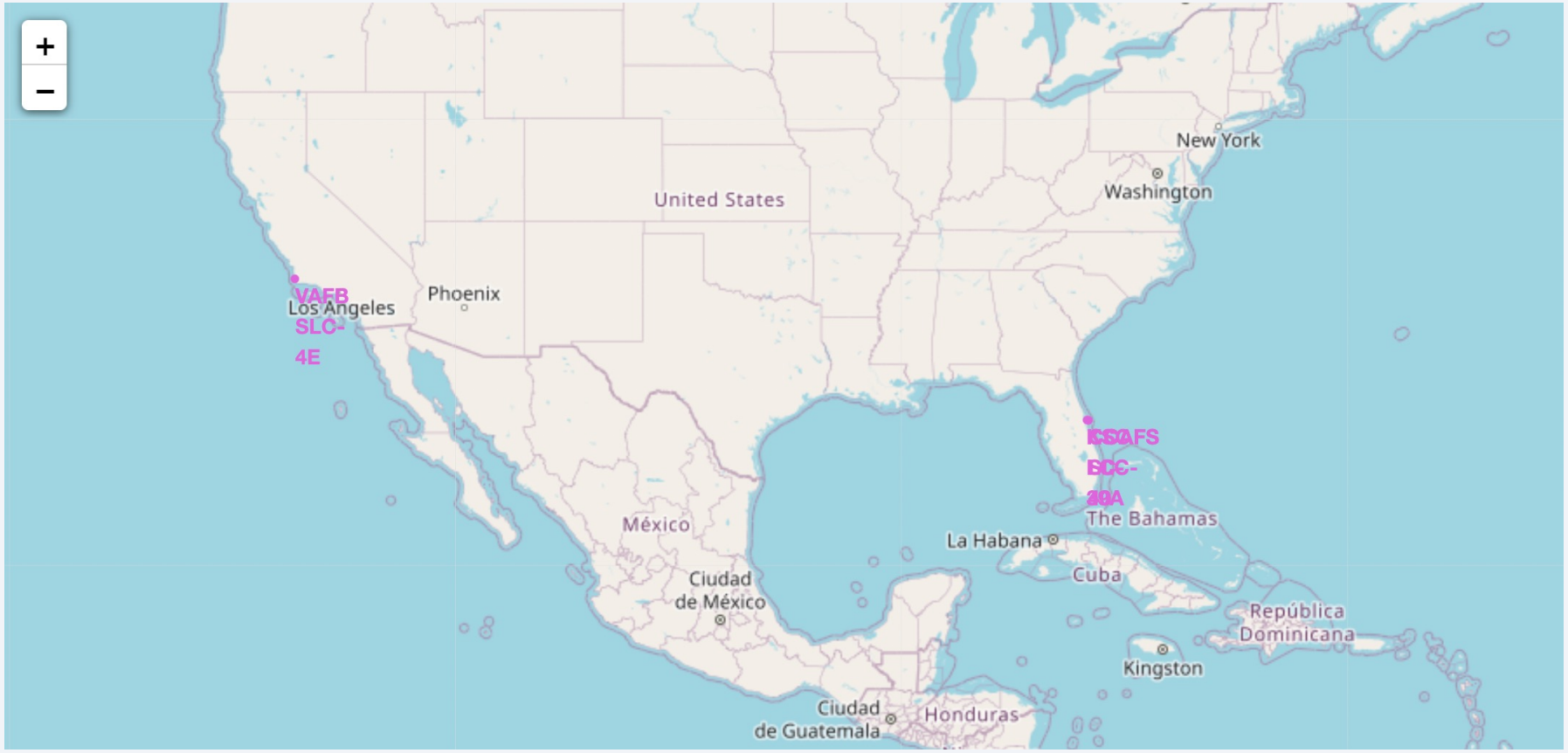
The count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order is:

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


Section 3
Launch Sites
Proximities Analysis

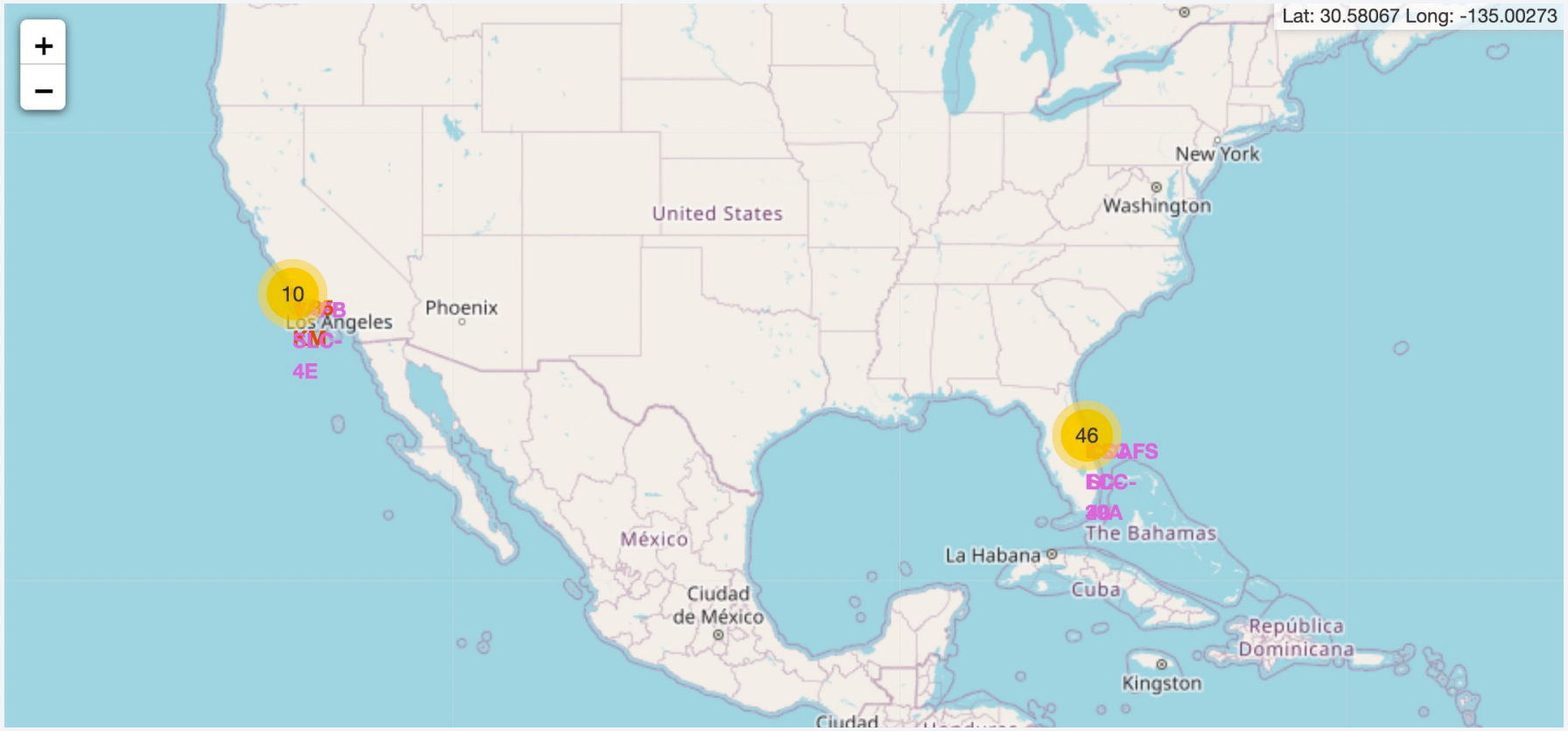
<All launch sites>

All launch sites with markers. One is on the west coast and three are on the east coast.



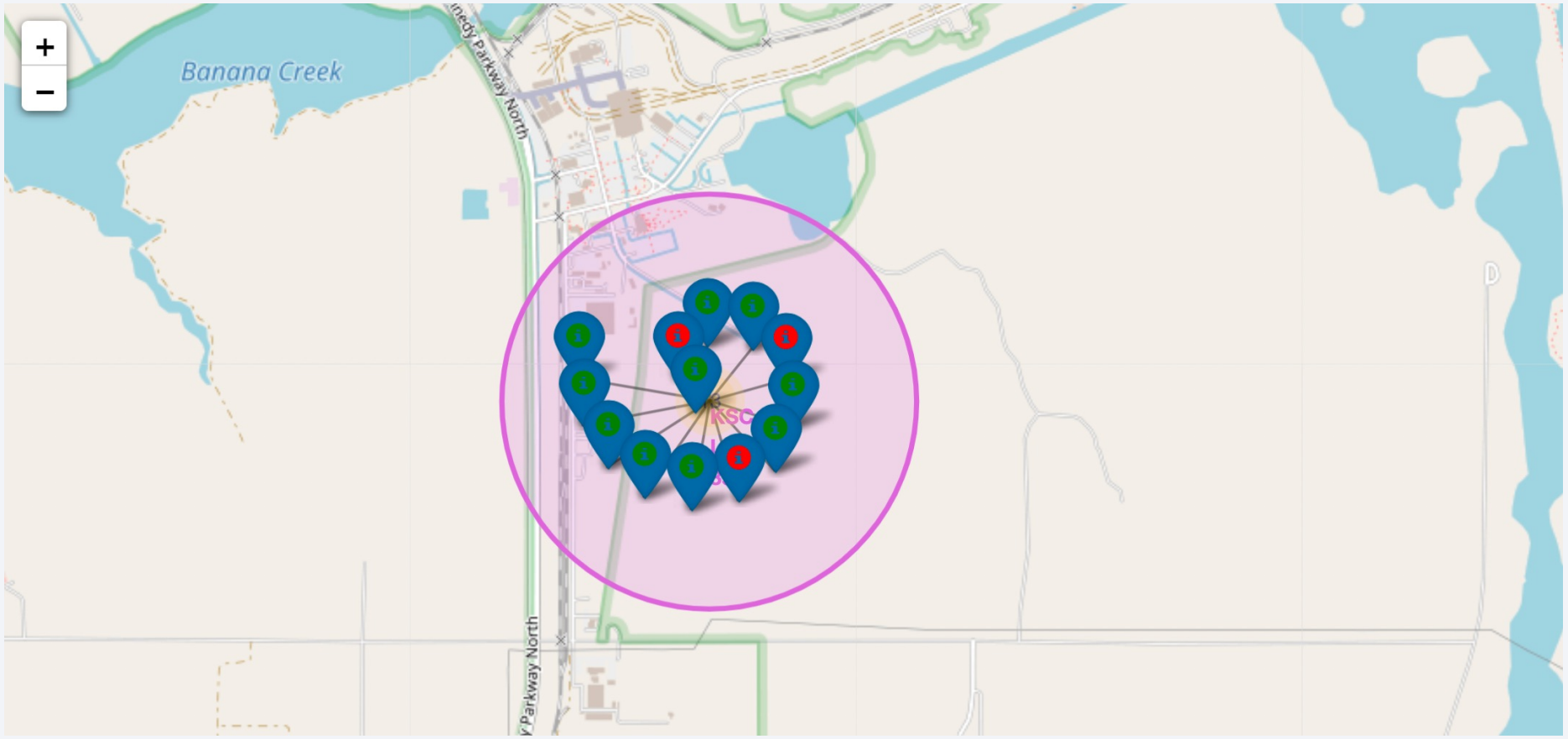
<All launch sites>

Besides four launch sites, this map also includes the number of launch outcomes at each cluster. The next page will give KSC LC-39A as an example.



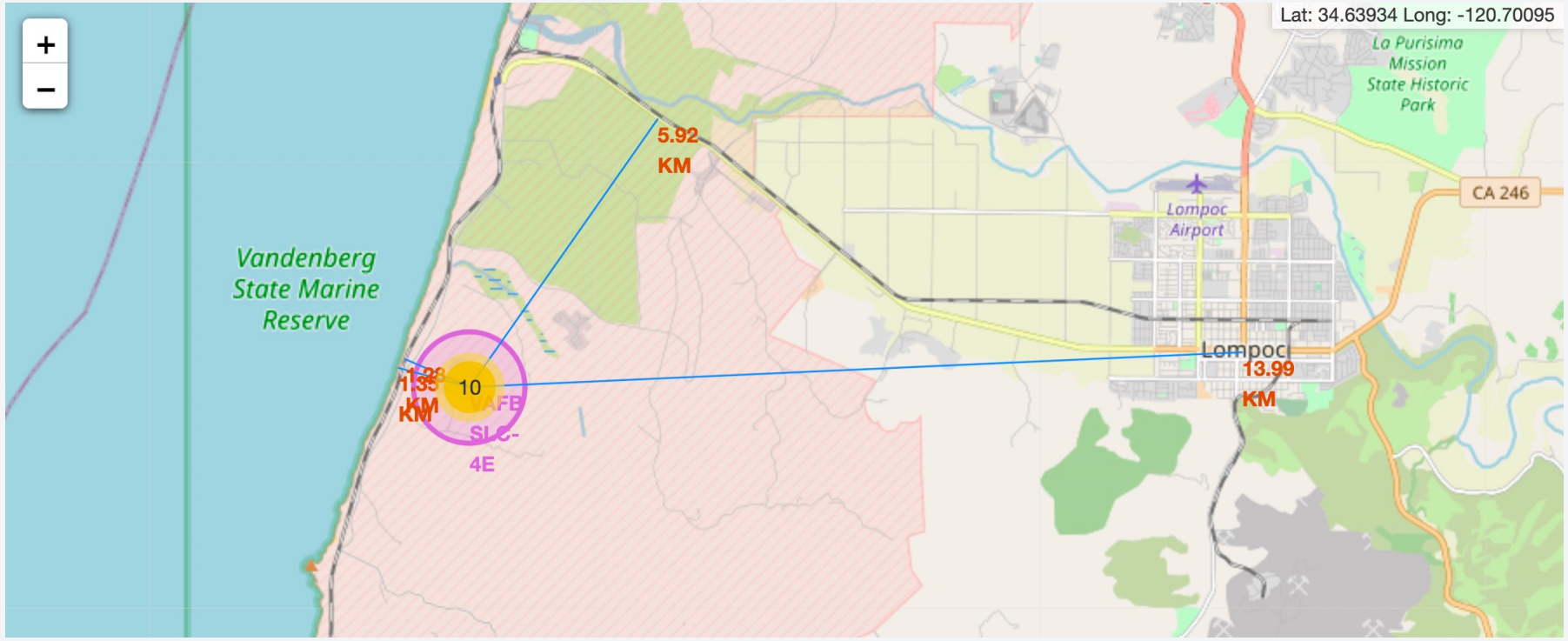
<Color-labeled Launch Outcome>

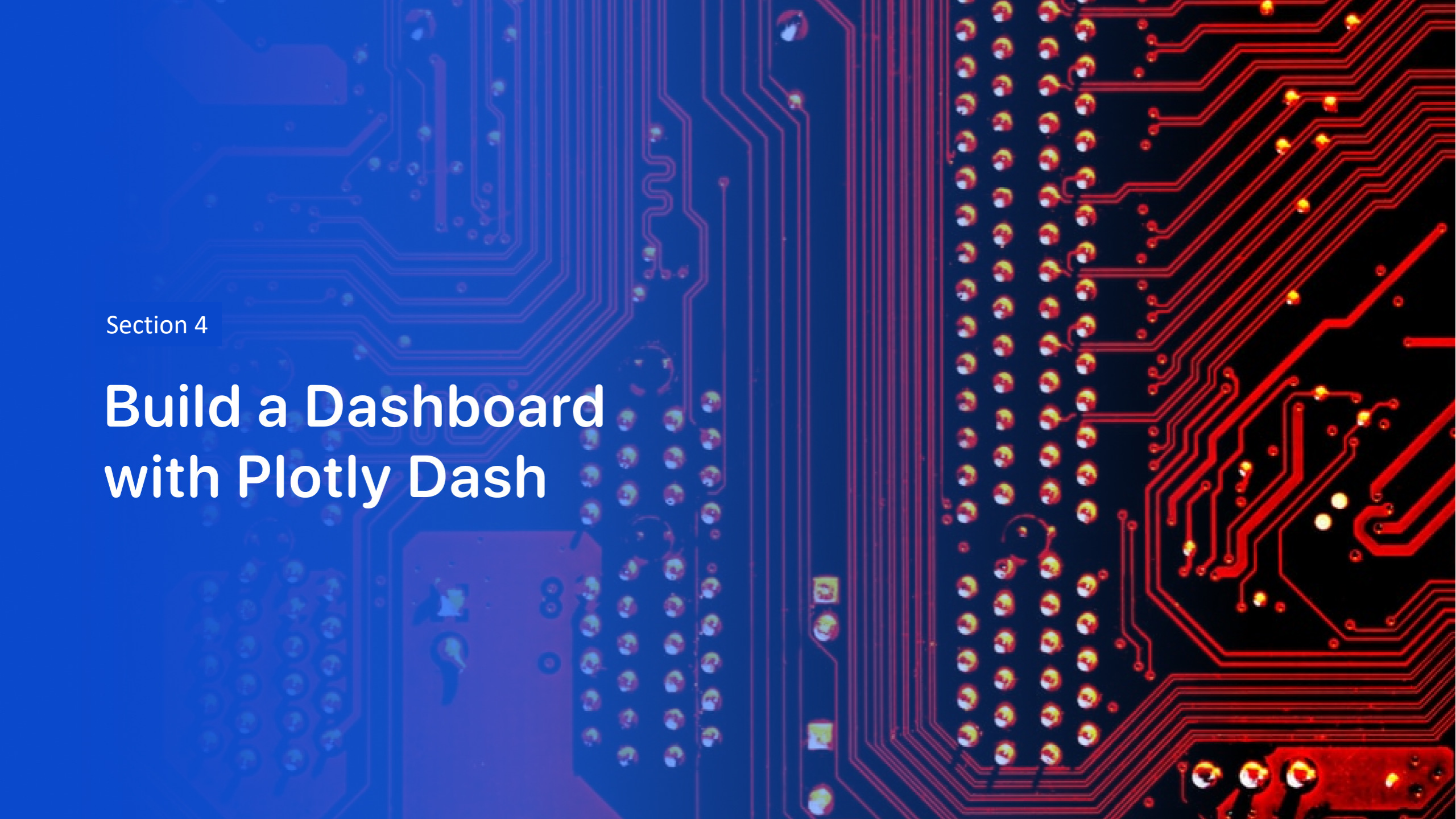
Color-labeled launch outcome for KSC LC-39A. There are 3 failed launches and 10 successful launches.



<Launch Site to Proximities>

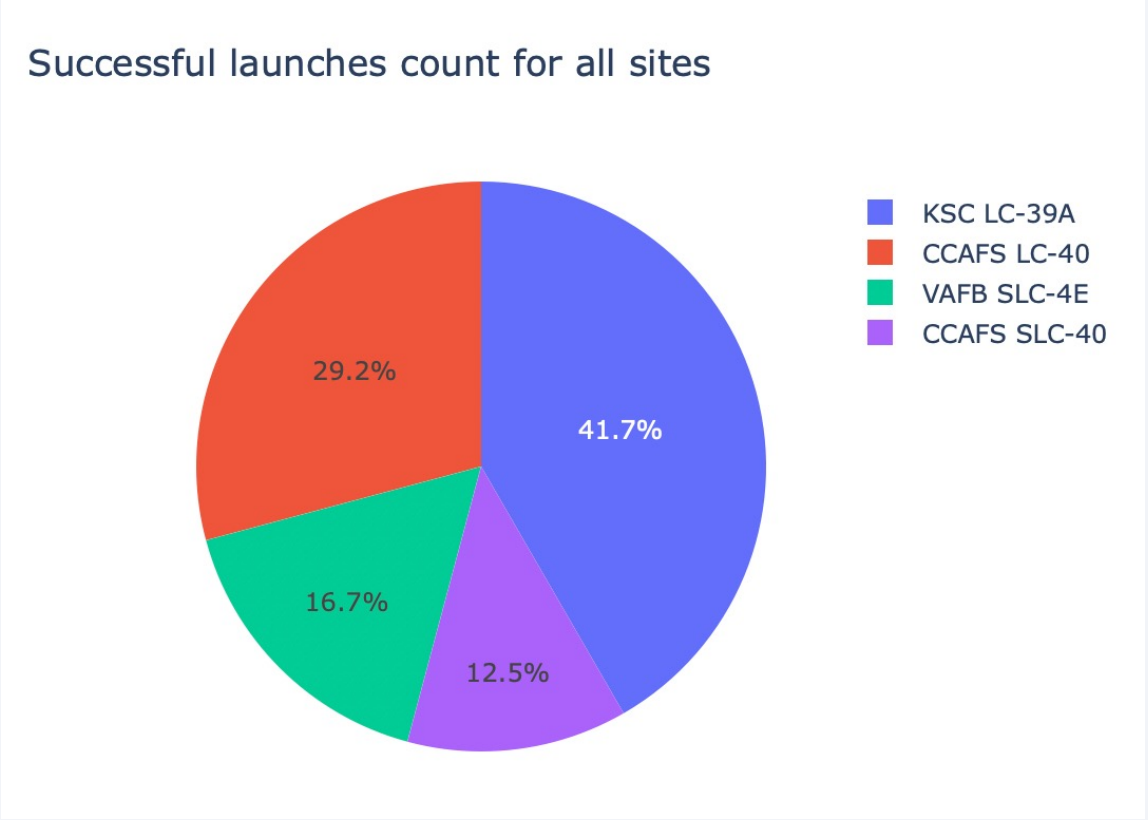
VAFB SLC-4E to its proximities with distance calculated and displayed. This screenshot shows that The distance from VAFB SLC-4E to the closest city is 13.99 KM, to the closest railway is 1.28 KM, to the closest coastline is 1.28 KM, and to the closest highway is 5.92 KM.




Section 4
Build a Dashboard with Plotly Dash

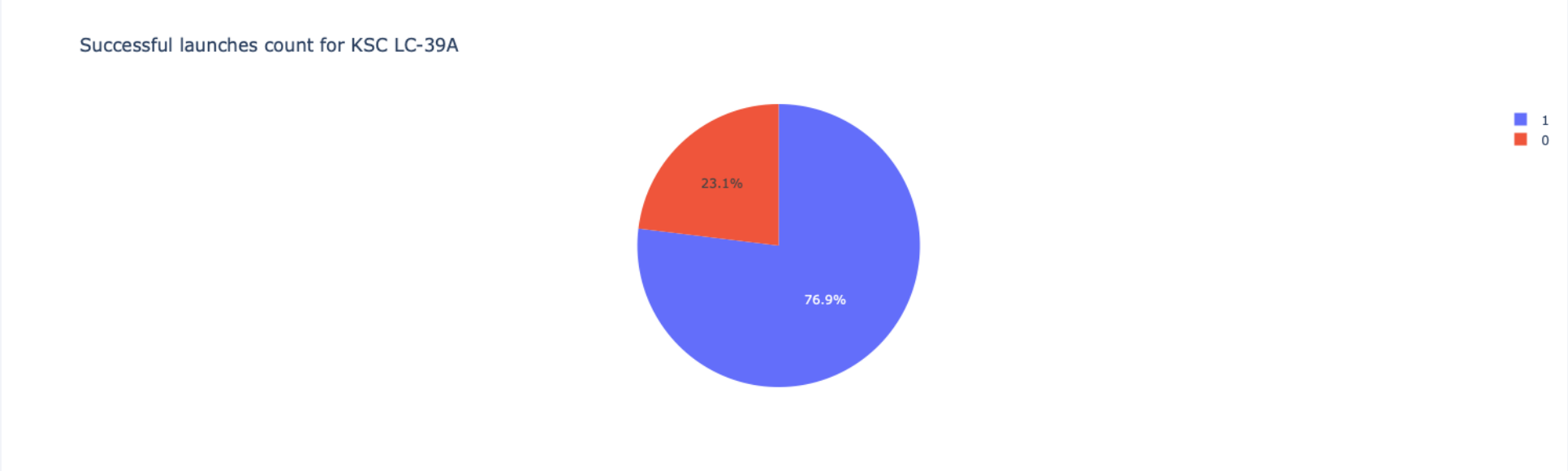
<Launch Success for All Sites>

The screenshot shows the launch success count for all sites. The KSC LC-39A has the most success cases which takes up 41.7% while CCAFS SLC-40 has the lease success cases.



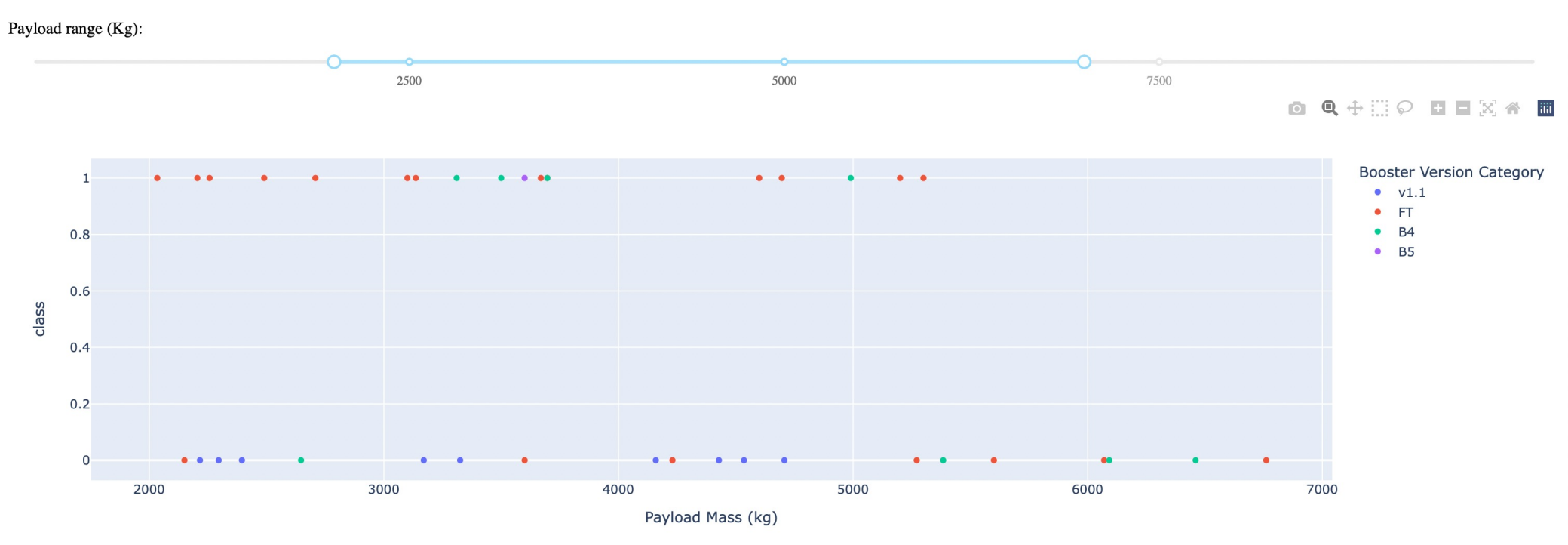
<Launch Site with Highest Success Ratio>

The launch site KSC LC-39A has the highest launch success ratio of 76.9%.



<Payload Mass vs Class by Booster Version>

The scatter plot indicates that the booster version FT has the highest success rate, and the most success cases locate in the payload range of 2000 to 7000.

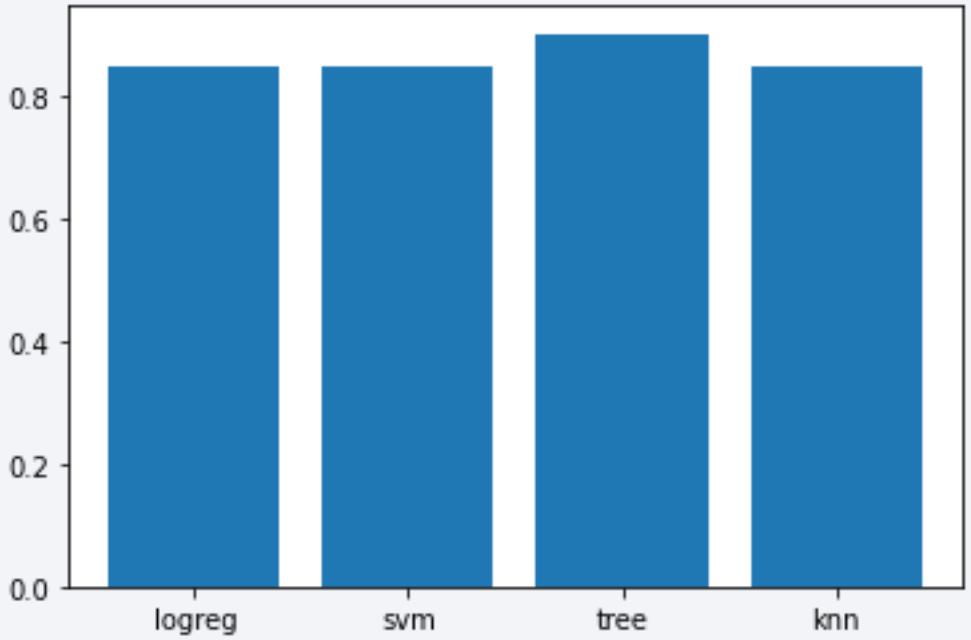


Section 5

Predictive Analysis (Classification)

Classification Accuracy

By comparing the accuracy of different models, the regression tree has a slightly higher accuracy than other models. Hence, the regression tree is more accurate than others in this question.



Confusion Matrix

The confusion matrix of regression tree model points out that there are 9 accurate prediction of landing and 4 for not landing. The model misses 5 cases with 2 did not land and 3 landed.

Conclusions

- Flight number is positively correlated to success rate
- Payload mass is positively correlated to success rate
- Orbits ES-L1, GEO, HEO, SSO have higher success rate
- Launch site KSC LC-39A has the highest success ratio
- Payload range from 2000 to 7000 has the most successful cases
- Regression tree has the highest accuracy in this project

Innovation

An innovation point I have beyond this template is that I added another widely used evaluation method for classification models – f1 score.

```
# calculate f1 score for all models
print(f1_score(Y_test, logreg_cv.predict(X_test)))
print(f1_score(Y_test, svm_cv.predict(X_test)))
print(f1_score(Y_test, tree_cv.predict(X_test)))
print(f1_score(Y_test, knn_cv.predict(X_test)))
```

```
0.888888888888889
0.888888888888889
0.888888888888889
0.888888888888889
```

Appendix

- SpaceX launch dataset: https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/spacex_launch_geo.csv
- Dashboard dataset (from IBM): https://github.com/zhao-edward/SpaceX-landing-analysis/blob/main/spacex_launch_dash.csv
- GitHub: https://github.com/zhao-edward/SpaceX-landing-analysis

Thank you!