IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Melanie Ng

31st January, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Methodology
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - The higher the flight number, the higher the success rate in landing first stage.
  - Lighter payload correlates with higher success rate in landing first stage.
  - Orbits with the highest success rates: ES-L1, GEO, HEO, SSO
  - Orbit with the lowest success rate: SO
  - Launch site KSC LC-39A has the highest success rate (76.9%)
  - Launch sites are close to coastlines and railways, far from highways and cities on average.
  - Decision Tree is the best algorithm for this prediction problem

Introduction

- Project background and context:
  - Reusability of rocket launch first stage determines cost of launch (e.g., SpaceX reuses first stage to save cost on rocket launches)
  - To develop a machine learning model for predicting success of SpaceX first stage landings
  - To determine price of each launch
- Problems to answer:
  - What contributes to a successful first stage landing?
  - Can we predict the success of a first stage landing?

Section 1

Methodology

Methodology

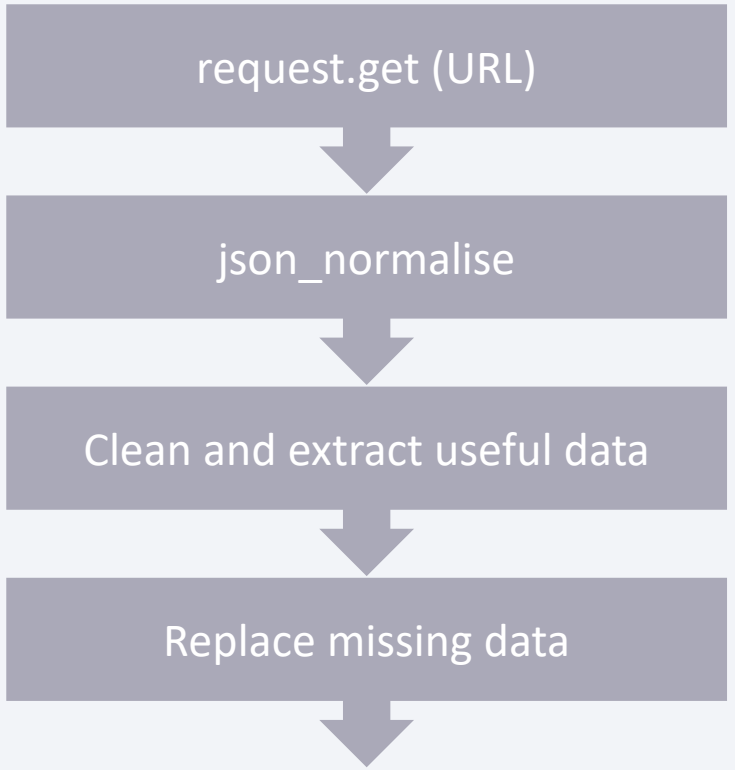
Executive Summary

1. Data collection: API, Webscraping
2. Data wrangling
3. Exploratory data analysis (EDA): SQL, visualization
4. Interactive visual analytics: Folium, Plotly Dash
5. Predictive analysis using classification models

Data Collection

Data Collection – SpaceX API

- request.get( ) from URL
- json_normalise
  - used to convert json result into dataframe
- Cleaning & extracting useful data
  - into new dict
  - converted into dataframe
  - only include Falcon9 launches
- Replacing missing values
  - Replace np.nan values with column mean

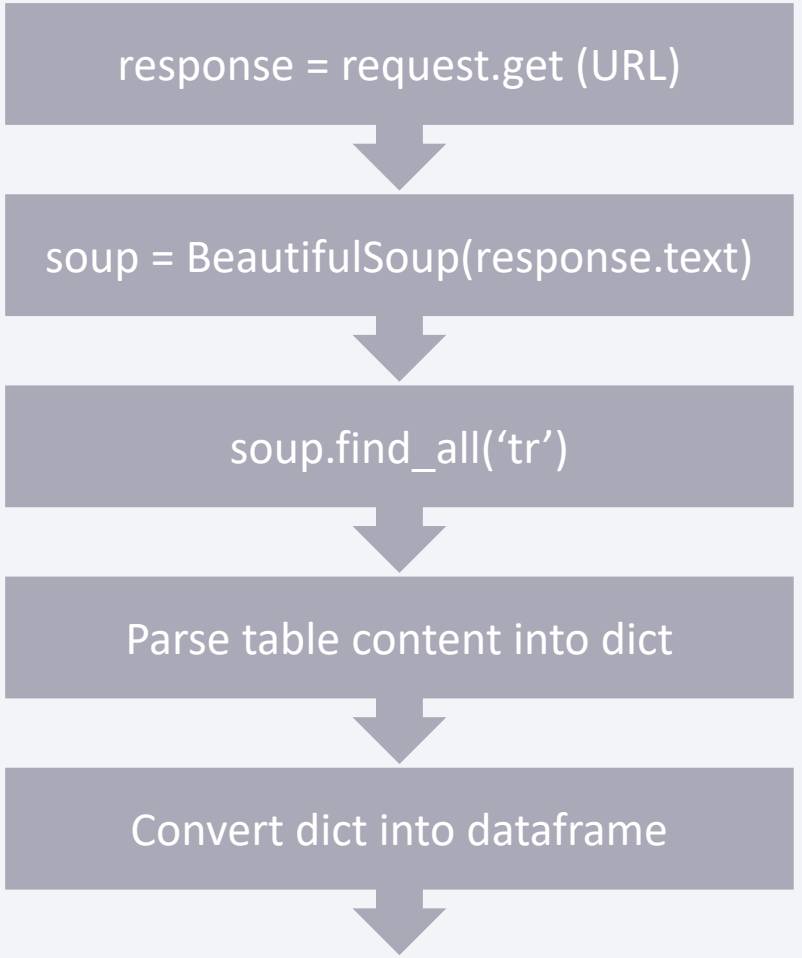


- GitHub URL: https://github.com/Melanieng401/IBM/blob/3dca6a9dcdfdf8e7d2865b1de81b8d18447a3551/IBM%20course%2010%3A%20Data%20Capstone/spacex-data-collection-api.ipynb

Data Collection – Web scraping

- Create BeautifulSoup( ) object from response text
- Find all tables
- From 3rd table, loop through rows (<th>) to extract column names
- Parse table content into dict
- Convert dict into dataframe

- GitHub URL: https://github.com/Melanieng401/IBM/blob/3dca6a9dcdfdf8e7d2865b1de81b8d18447a3551/IBM%20course%2010%3A%20Data%20Capstone/jupyter-labs-webscraping.ipynb



Data Wrangling

Data Wrangling

- Exploratory data analyses
  - Calculated number of launches on each site
  - Calculated number and occurrence of each orbit
  - Calculated number and occurrence of mission outcome of the orbits
- Created landing outcome label (class)
  - 1: success; 0: fail
- Calculated success rate based on mean of outcome label
- Github URL: https://github.com/Melanieng401/IBM/blob/cda5b7393d5ecbdc9172e956ec21de7c9ca2ab26/IBM%20course%2010%3A%20Data%20Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

Creating landing outcome label:

Separated fail landing outcomes into new list (bad_outcomes)

Append 1/0 to new list per read in landing outcomes column:
- 0 if value matches item in bad_outcomes;
- 1 if otherwise

Convert new list to dataframe and add to df as a new column

EDA with SQL

Github URL:
https://github.com/Melanieng401/IBM/blob/ef42efd0a6d2fd1cd59fff8f44642035cc2f01a5/IBM%20course%2010%3A%20Data%20Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- Queries performed:
  - Identified names of unique launch sites
  - Identified records where launch sites begin with 'CCA'
  - Identified total payload mass carried by boosters launched by NASA (CRS)
  - Identified average payload mass carried by booster version F9 v1.1
  - Identified the date when the first successful landing outcome in ground pad was achieved
  - Identified names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - Identified the names of the booster_versions which have carried the maximum payload mass, using subquery
  - Identified the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - Ranked 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

Connected to database via sqlite3

EDA with Data Visualization

- Charts plotted:
  - Category plot with Seaborn
    - Visualise the relationship between discrete and continuous data in a scatter plot format
  - Bar chart with Seaborn
    - Compare differences of continuous data (e.g., success rate) among discrete groups (e.g., orbit type)
  - Line graph with Seaborn
    - Visualise trend between data (e.g., class over years)

Github URL:
https://github.com/Melanieng401/IBM/blob/9bb1276f5110e79141502003ffe86a9a41ff0a59/IBM%20course%2010%3A%20Data%20Capstone/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Build an Interactive Map with Folium

- Created markers
  - marked success or fail launches from launch sites – green: success; red: fail
  - grouped markers into clusters for cleaner visualisation
- Added mouse position
  - to track mouse pointer on Folium map
- Added lines
  - to mark distances between two coordinates

- Github URL: https://github.com/Melanieng401/IBM/blob/ce7cc2dc9b47ef8e89b2c13e4fbe6c2c69fb8d38/IBM%20course%2010%3A%20Data%20Capstone/lab_jupyter_folium.jupyterlite.ipynb

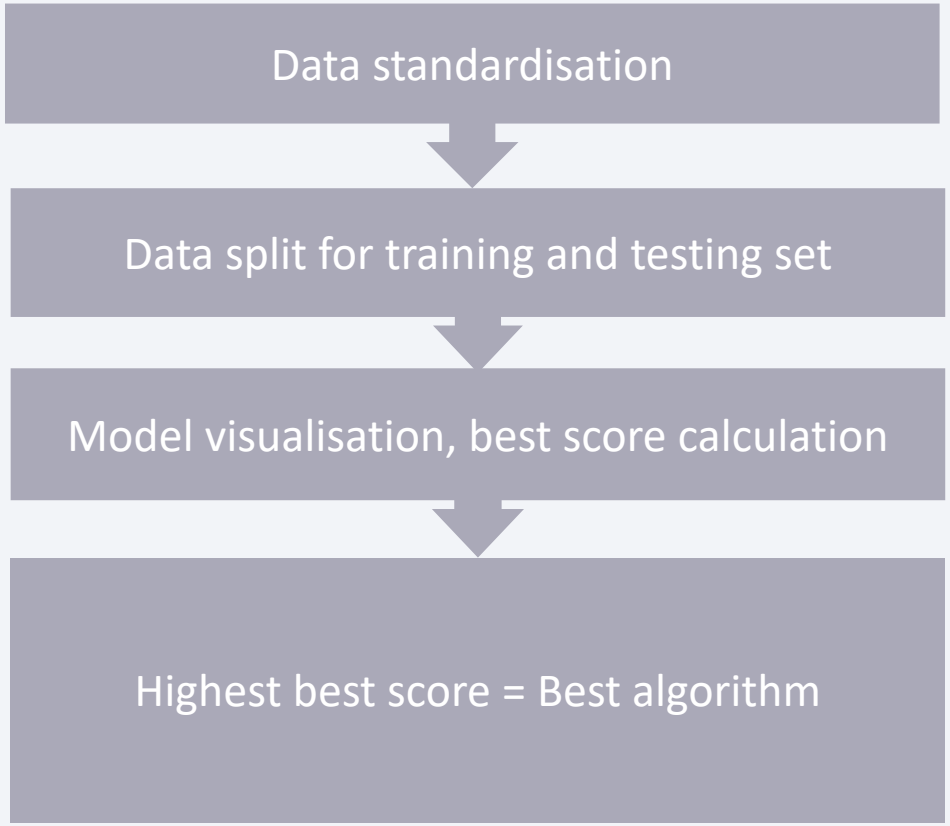
Build a Dashboard with Plotly Dash

- Interactive dashboard
  - launch site selection
  - payload mass selection via slider
- Pie charts
  - showing the total launches by a certain sites
  - allow easy-to-interpret comparisons
- Scatter graph
  - showing the relationship with Outcome and Payload Mass (Kg) for the different booster version

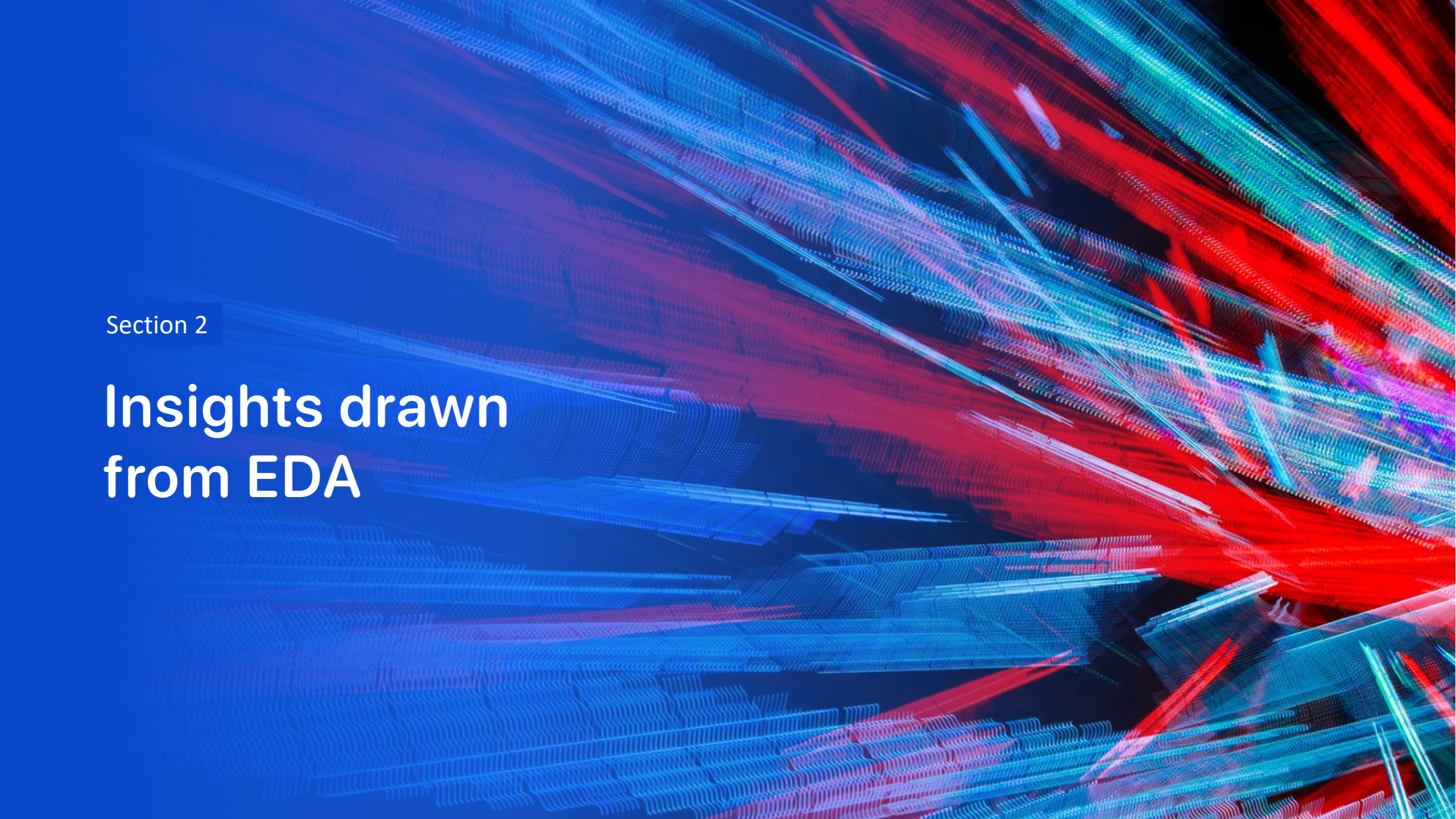
- Github URL: https://github.com/Melanieng401/IBM/blob/562ffb1927541d02300b648f6bba8cbb4dfdcdc6/IBM%20course%2010%3A%20Data%20Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

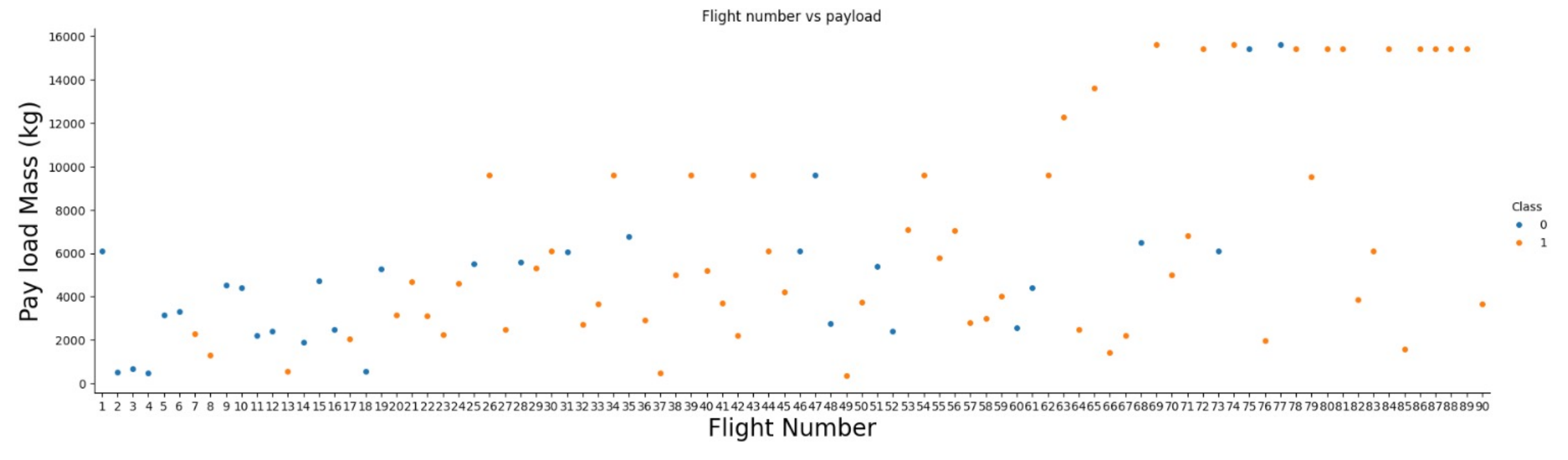
- Data processing:
  - Standardisation: StandardScaler
  - Data split: train_test_split
- Models, with GridSearchCV, cv=10:
  - Logistic regression: LogisticRegression()
  - Support Vector Machine: SVC()
  - Decision Tree: DecisionTreeClassifier()
  - K Nearest Neighbour: KNeighborsClassifier()
- Visualisation with Confusion matrix
- Compare best scores to find the best model



Github URL:
https://github.com/Melanieng401/IBM/blob/f6658d3043b8777002bdf350246b94e60a2c69b6/IBM%20course%2010%3A%20Data%20Capstone/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

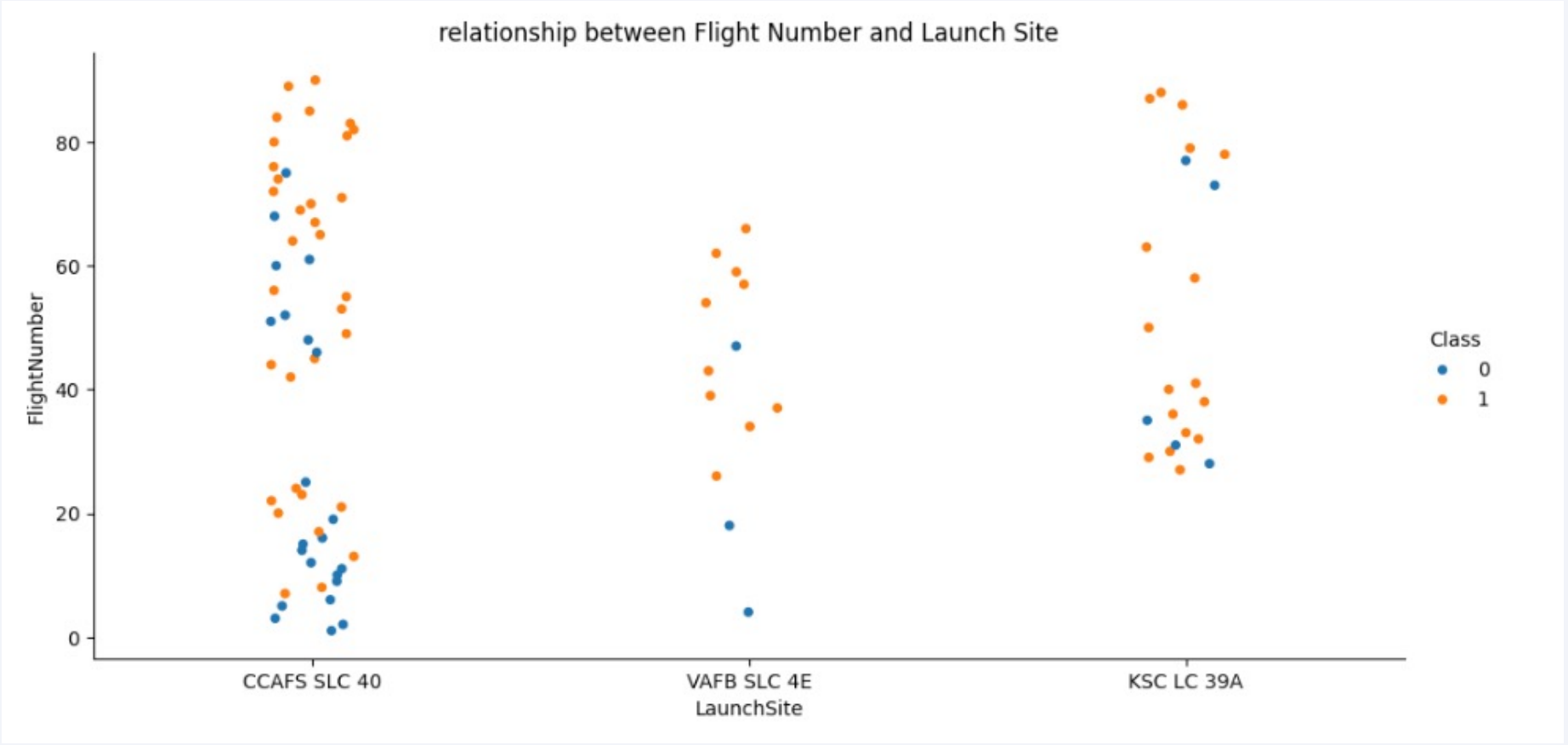

Section 2
Insights drawn from EDA

Flight Number vs. payload



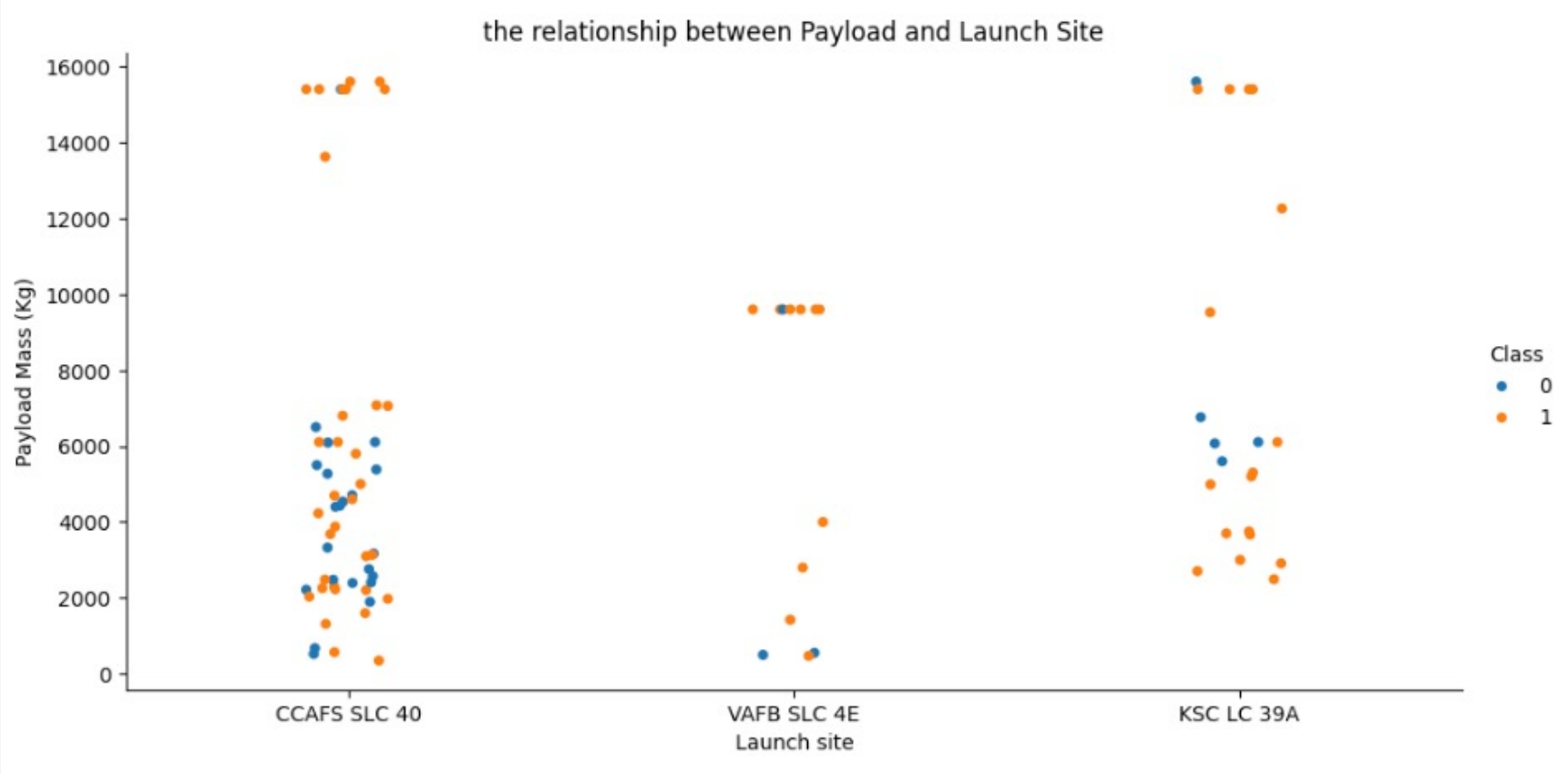
- As the flight number increases, the first stage is more likely to land successfully.
- As the payload mass increases, the less likely the first stage will be recovered.

Flight Number vs. Launch Site



The higher the flight numbers, the higher the success rate

Payload vs. Launch Site



Launch site CCAFS SLC 40 has highest success rate with heavier payloads

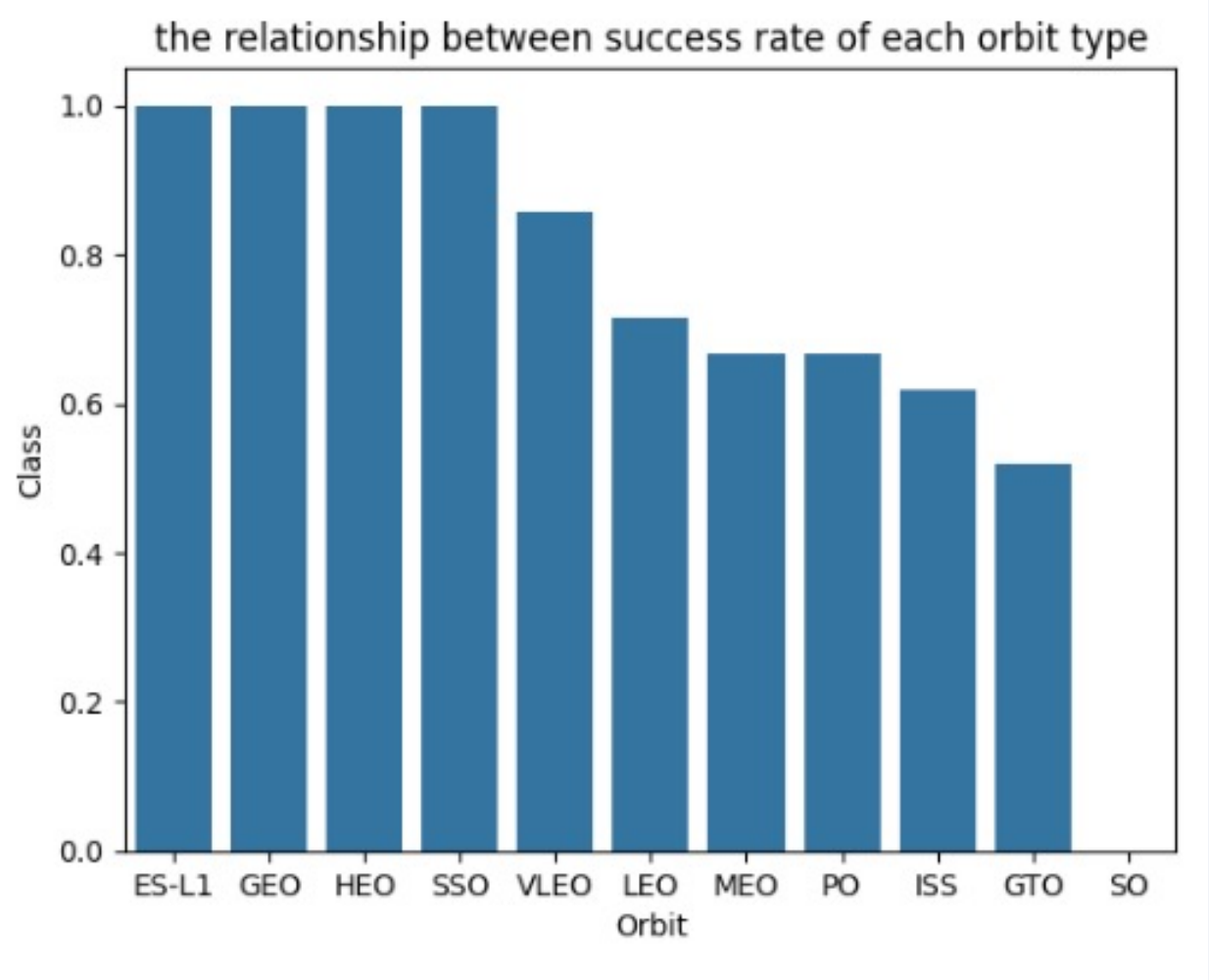
Success Rate vs. Orbit Type

Orbits with highest success rates (100%):

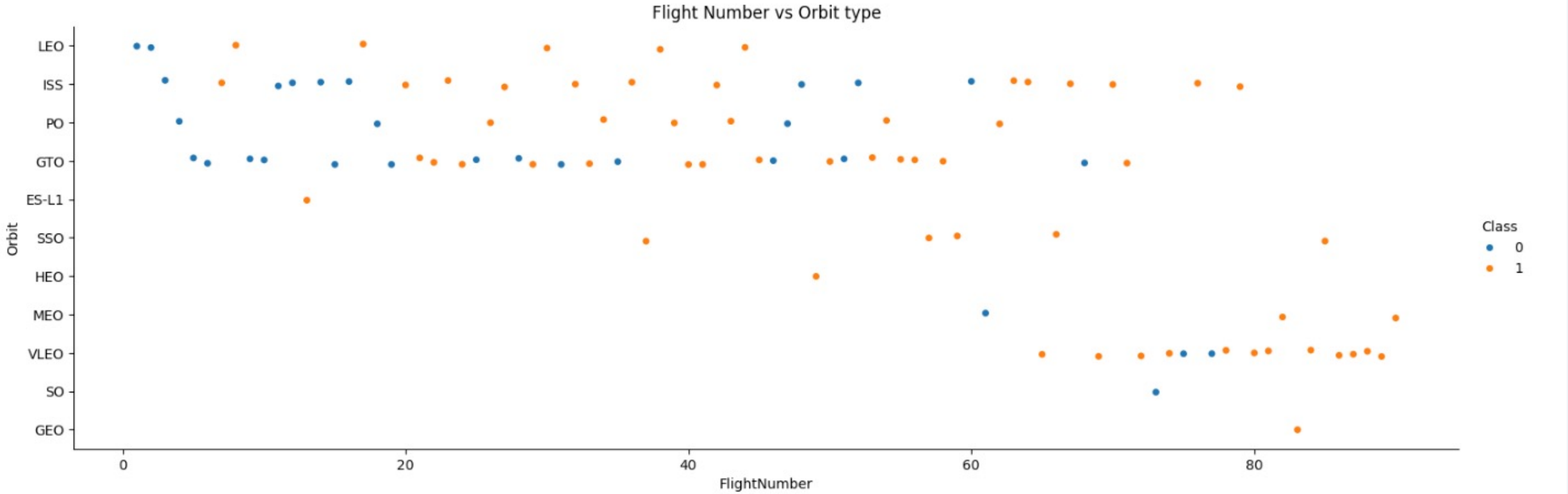
- ES-L1, GEO, HEO, SSO

Orbits with lowest success rate (0%):

- SO

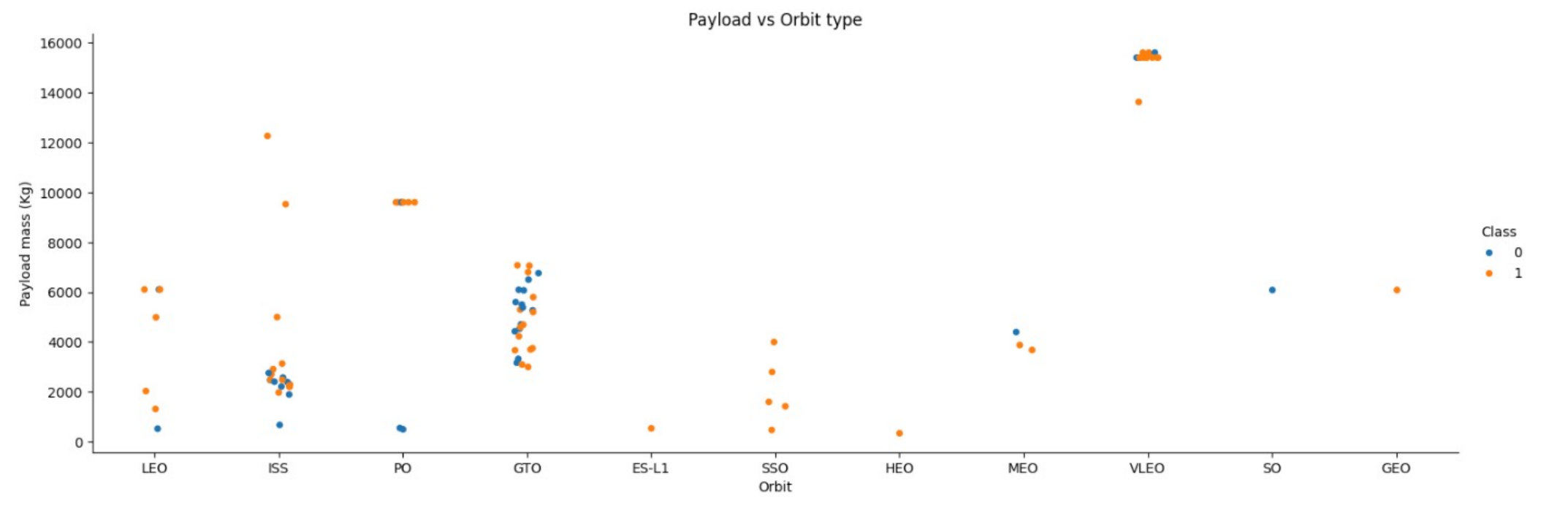


Flight Number vs. Orbit Type



- Success of the LEO orbit appears related to the number of flights
- There seems to be no relationship between flight number when in the GTO orbit

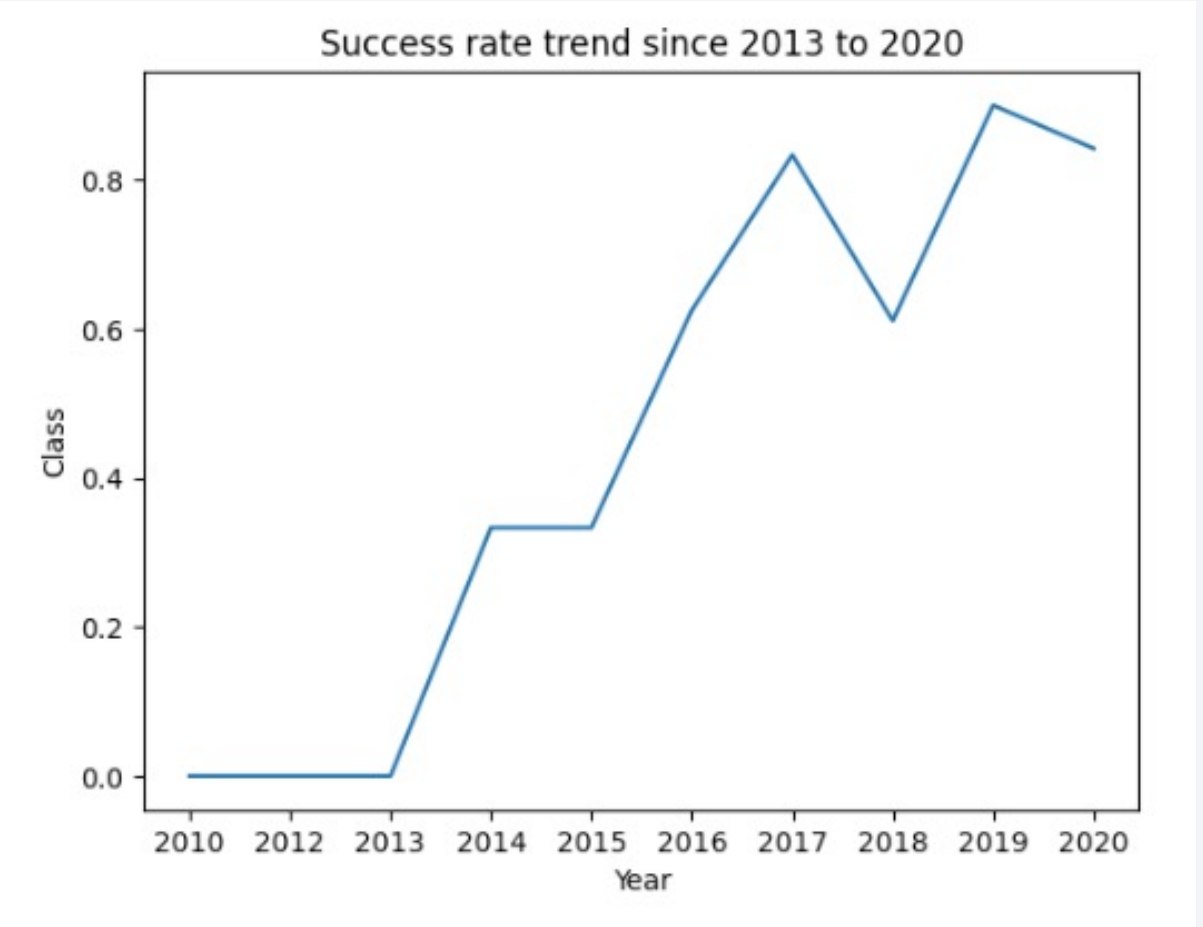
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar (PO), LEO, ISS and VLEO.
- Indistinguishable relationship between payload and success when in GTO orbit.

Launch Success Yearly Trend

- Overall increasing success rate from 2013 to 2020
- Dip in 2018



All Launch Site Names

```
%%sql
SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

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

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

SQL was used to select unique launch sites from data

Launch Site Names Begin with 'CCA'

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

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

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

Here are 5 records where launch sites begin with `CCA`

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

The total payload carried by boosters from NASA (CRS) is 45596 Kg

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

The average payload mass carried by booster version F9 v1.1 is 2928.4 Kg

First Successful Ground Landing Date

```
%%sql

SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(Date)
2015-12-22

The date of the first successful landing outcome on ground pad is 22nd December, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT Booster_Version FROM SPACEXTABLE WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000) AND Landing_Outcome ='Success (drone ship)'
```

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

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success%'
```

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

COUNT(Mission_Outcome)
100

```
%sql SELECT COUNT(Mission_Outcome) FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Failure%'
```

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

COUNT(Mission_Outcome)
1

There are 100 successful missions and 1 failed mission only

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

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

This query results show the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%%sql
SELECT substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE WHERE substr(Date,0,5)='2015' and Landing_Outcome = 'Failure (drone ship)'
```

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

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

The above query results show the names of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%%sql
-- SELECT COUNT(DISTINCT(Landing_Outcome)) as lo_count, Date FROM SPACEXTABLE
-- WHERE (Date BETWEEN 2010-06-04 AND 2017-03-20) GROUP BY Landing_Outcome ORDER BY lo_count DESC

SELECT Landing_Outcome,count(Landing_Outcome) FROM SPACEXTABLE WHERE Date <= '2017-03-20' GROUP BY Landing_Outcome ORDER BY count(Landing_Outcome) DESC
```

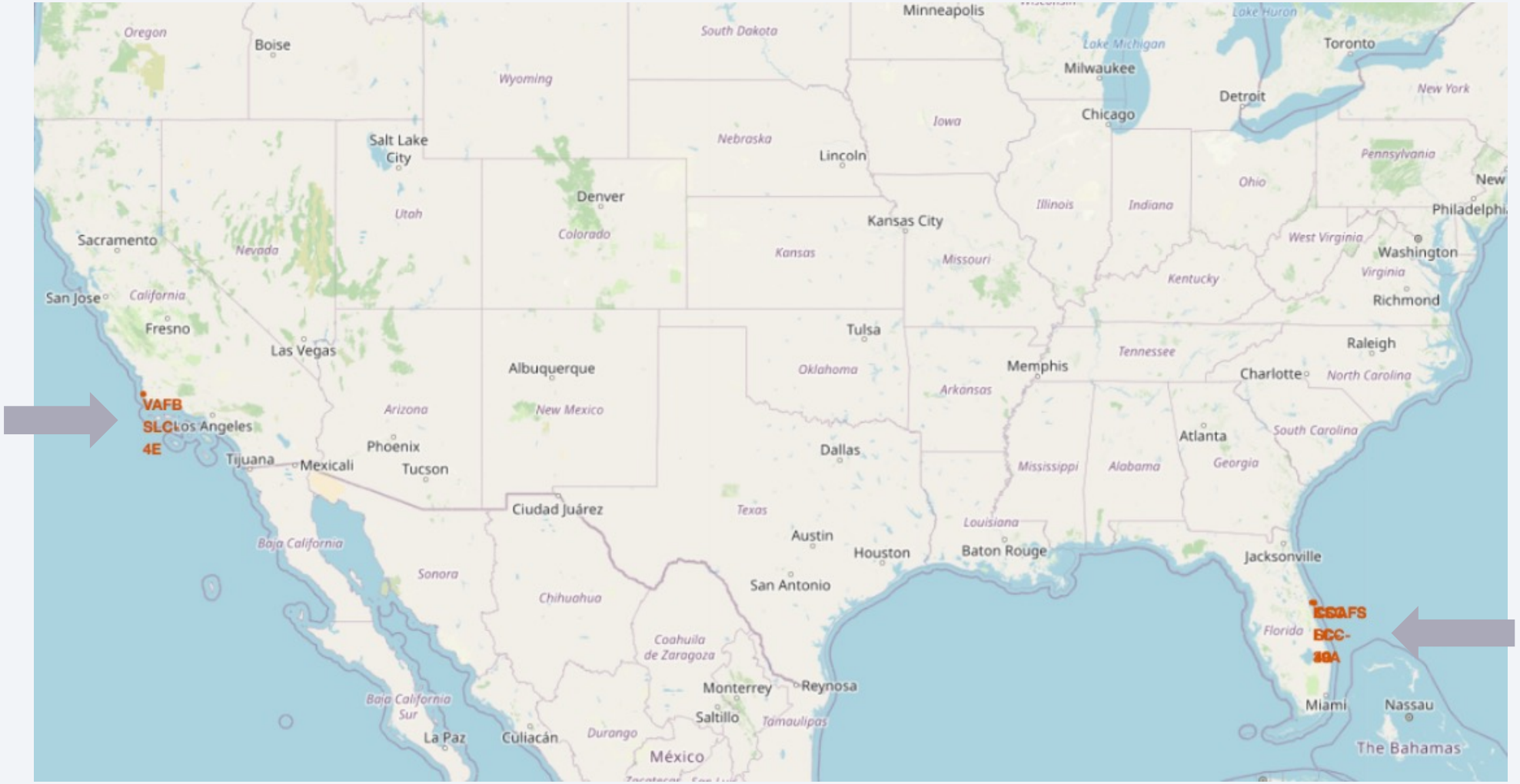
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count(Landing_Outcome)
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

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


Section 3
Launch Sites
Proximities Analysis

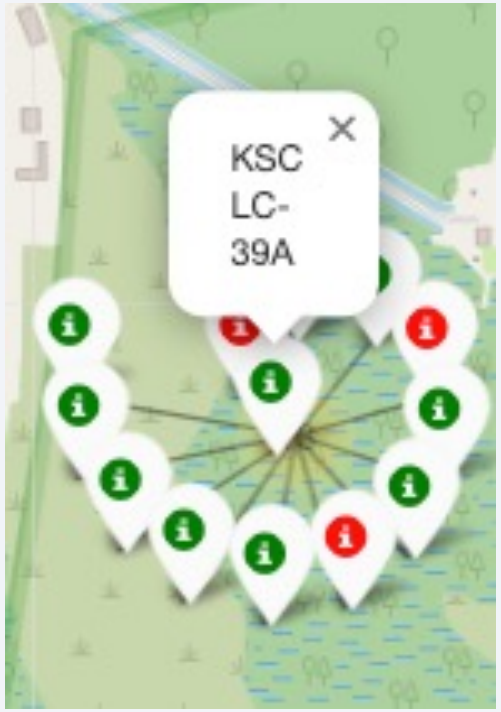
Launch sites marked

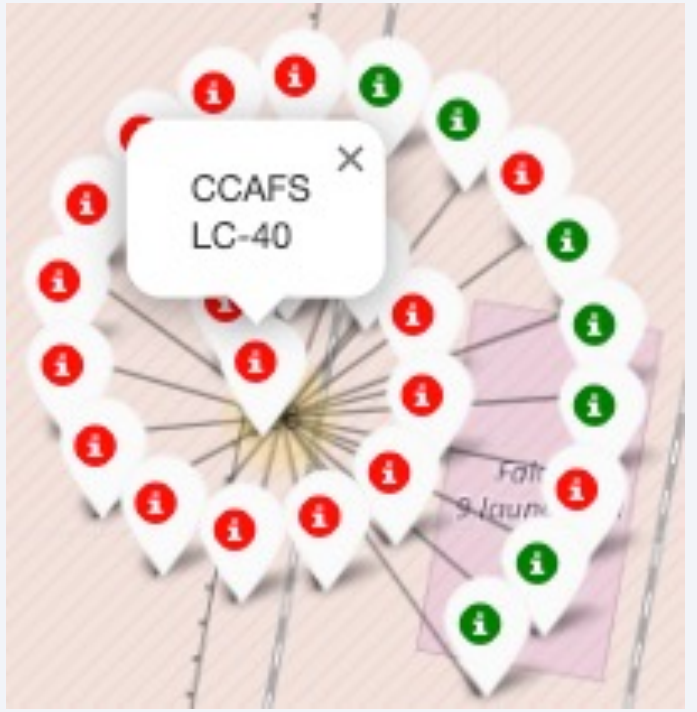


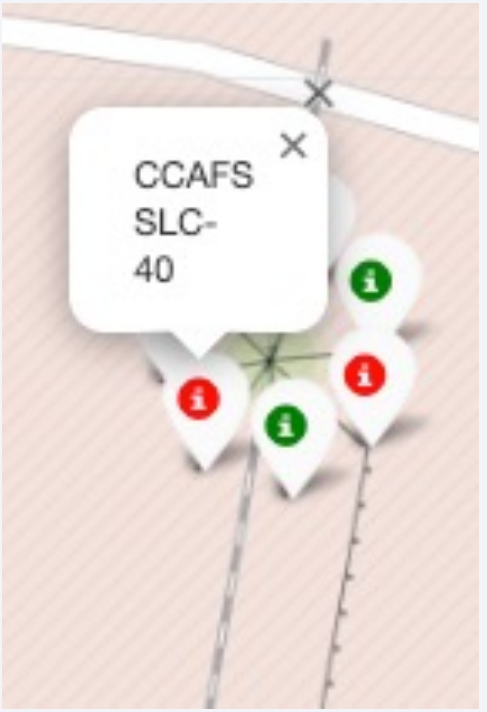
Exploratory data analysis:

Data visualisation | SQL analysis | Interactive proximity analysis | Interactive dashboard | Predictive modelling

Colour coded launch sites clusters

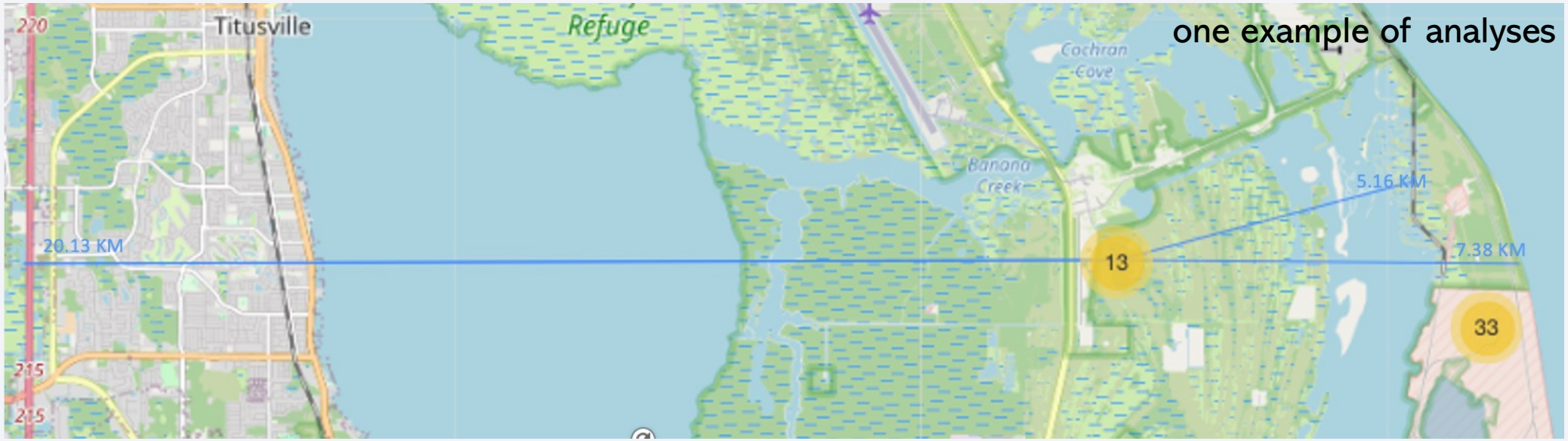






Green = successful launch
Red = non-successful launch

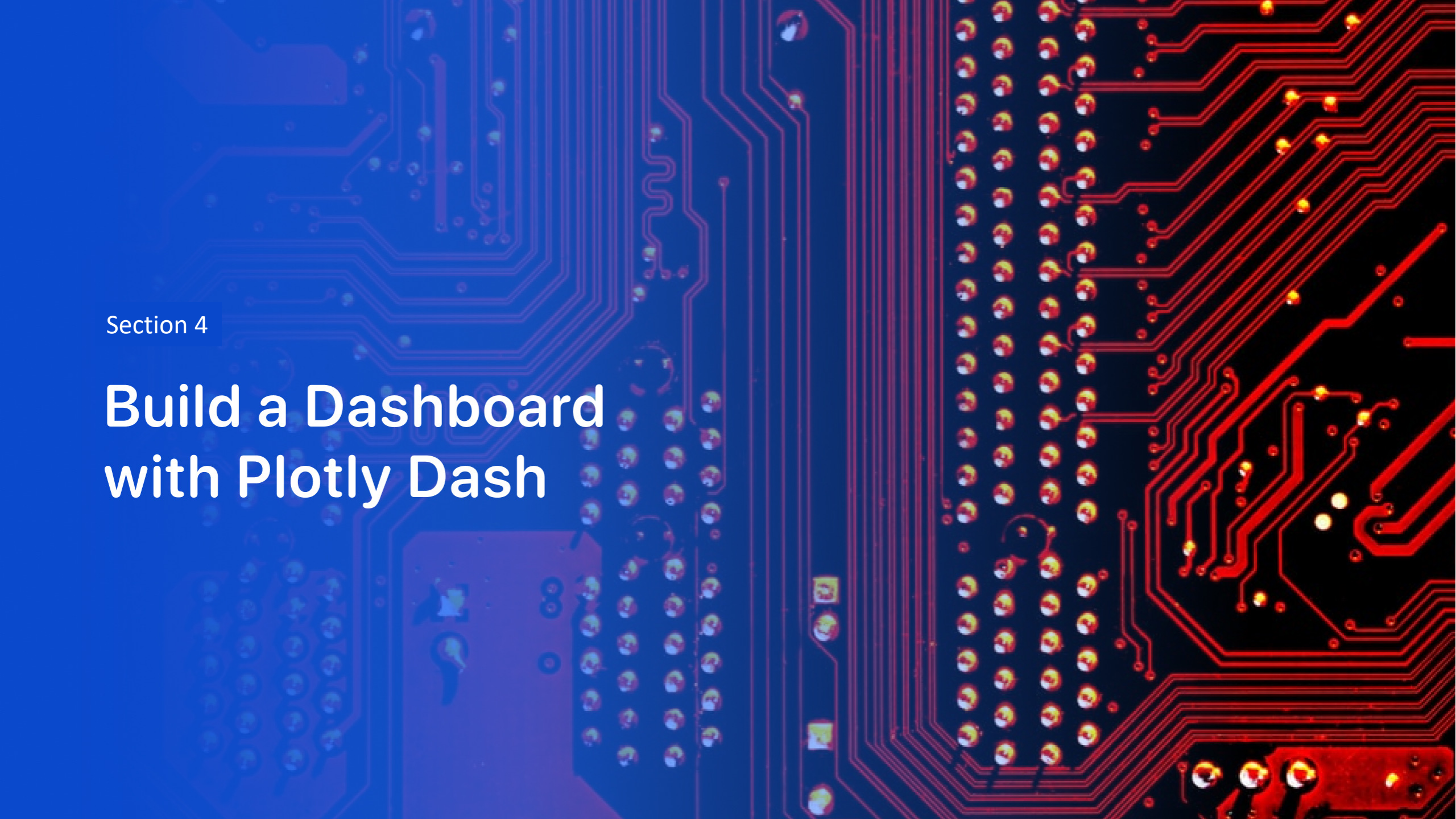
Launch site location analysis



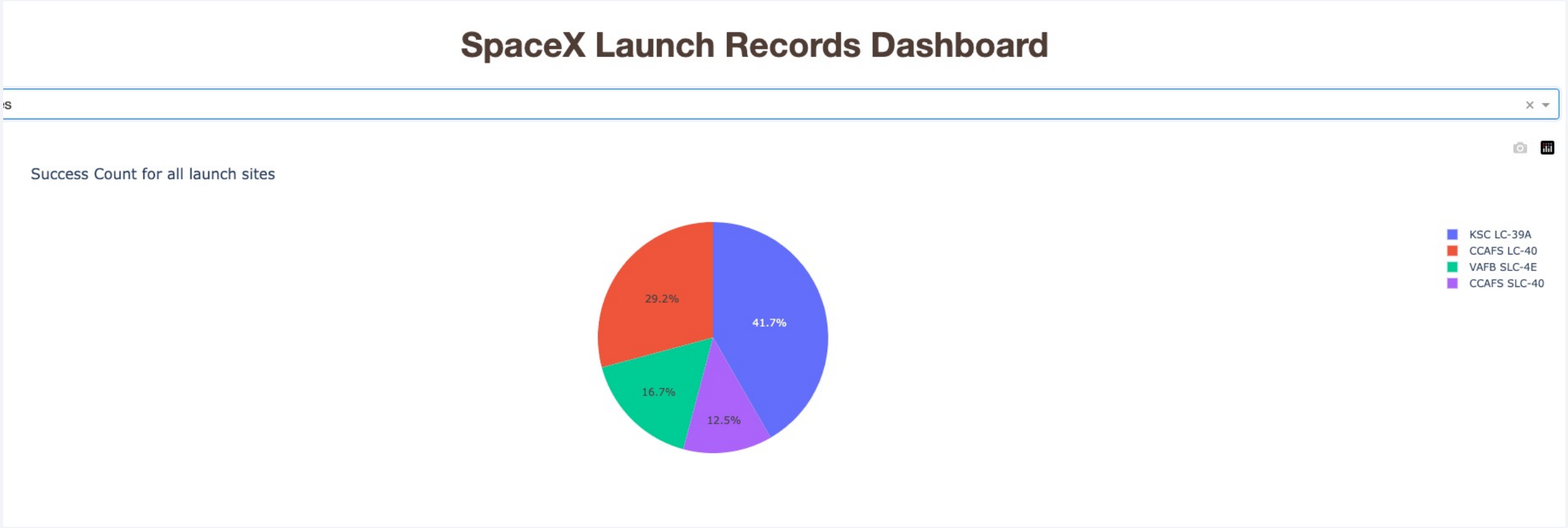
Distances between railways, highways, coastline, cities calculated and taken average

```
avg_coastlinedist 3.1714820236382O8
avg_railwaydist 2.1909241493771727
avg_highwaydist 22.004038650652 28
avg_citydist 28.11529401314546
```

Launch sites are close to coastlines and railways, and far from highways and city centres

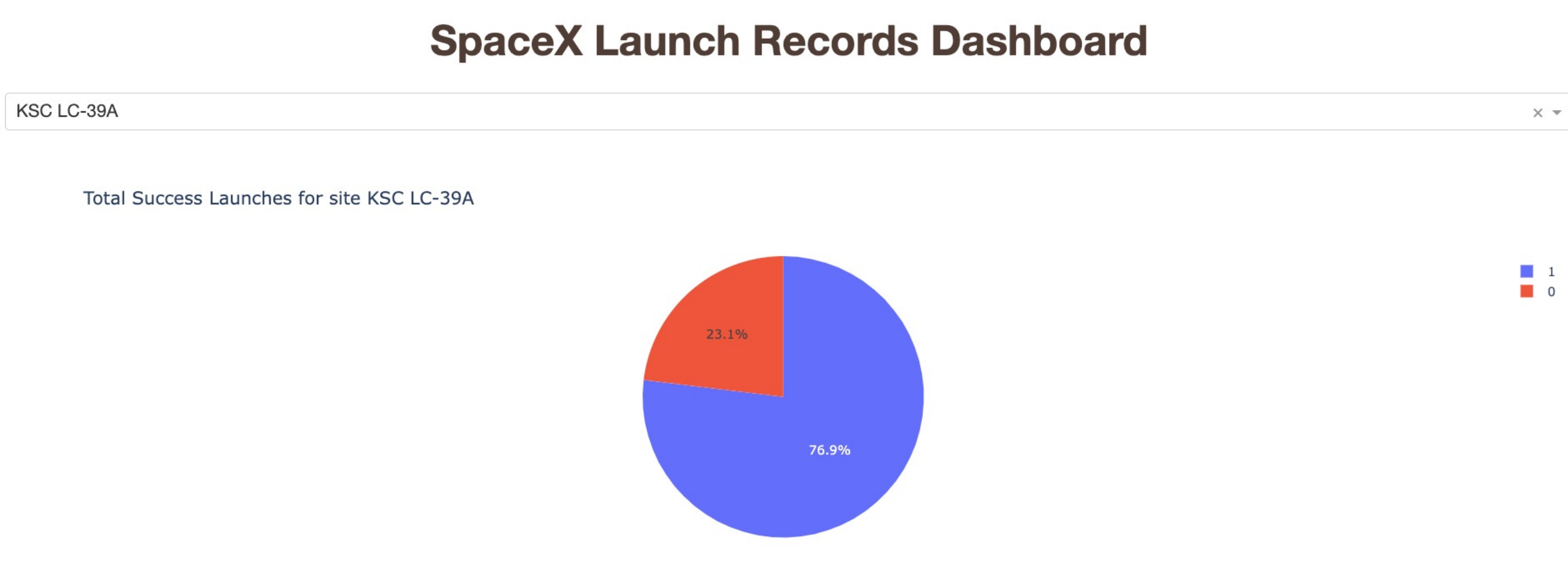

Section 4
Build a Dashboard with Plotly Dash

Comparing success rate of all launch sites



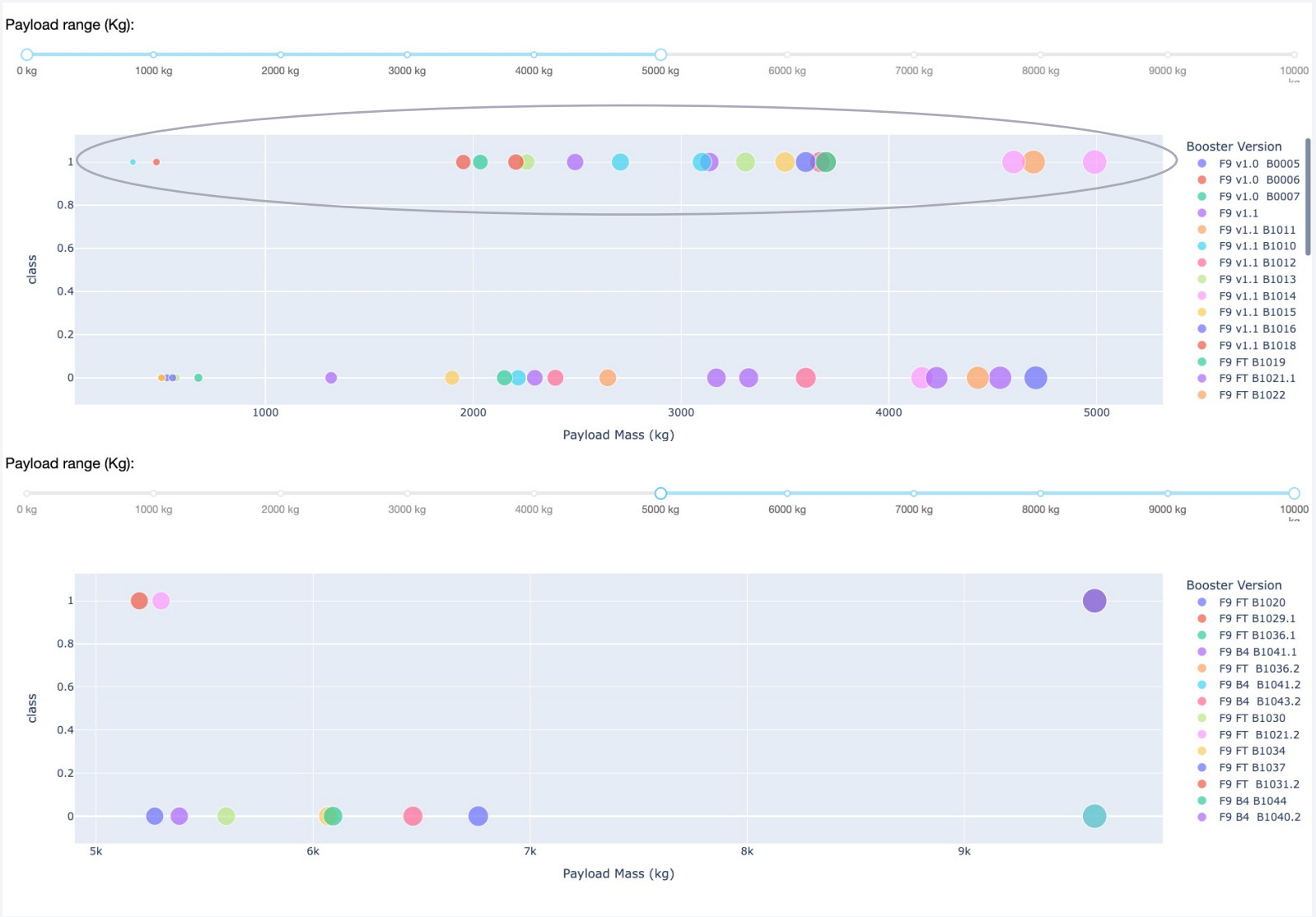
Launch site KSC LC-39A has the highest success rate. CCAFS SLC-40 has the lowest success rate.

Success:Fail ratio of the launch site with highest success rate



KSC LC-39A has a success rate of 76.9%, and a failure rate of 23.1%

Payload (Kg) vs Success rate



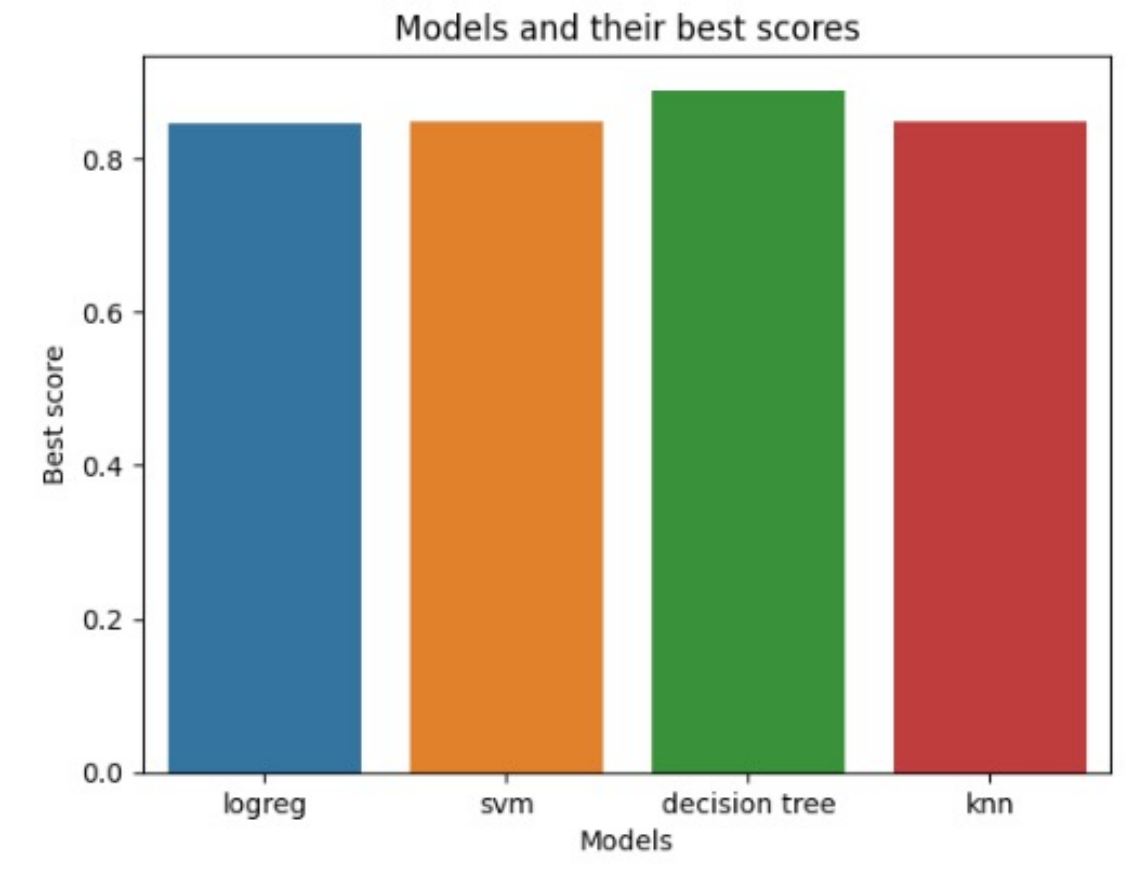
Lighter payload has a higher success rate than heavier payloads

(more class 1 data points)

Section 5

Predictive Analysis (Classification)

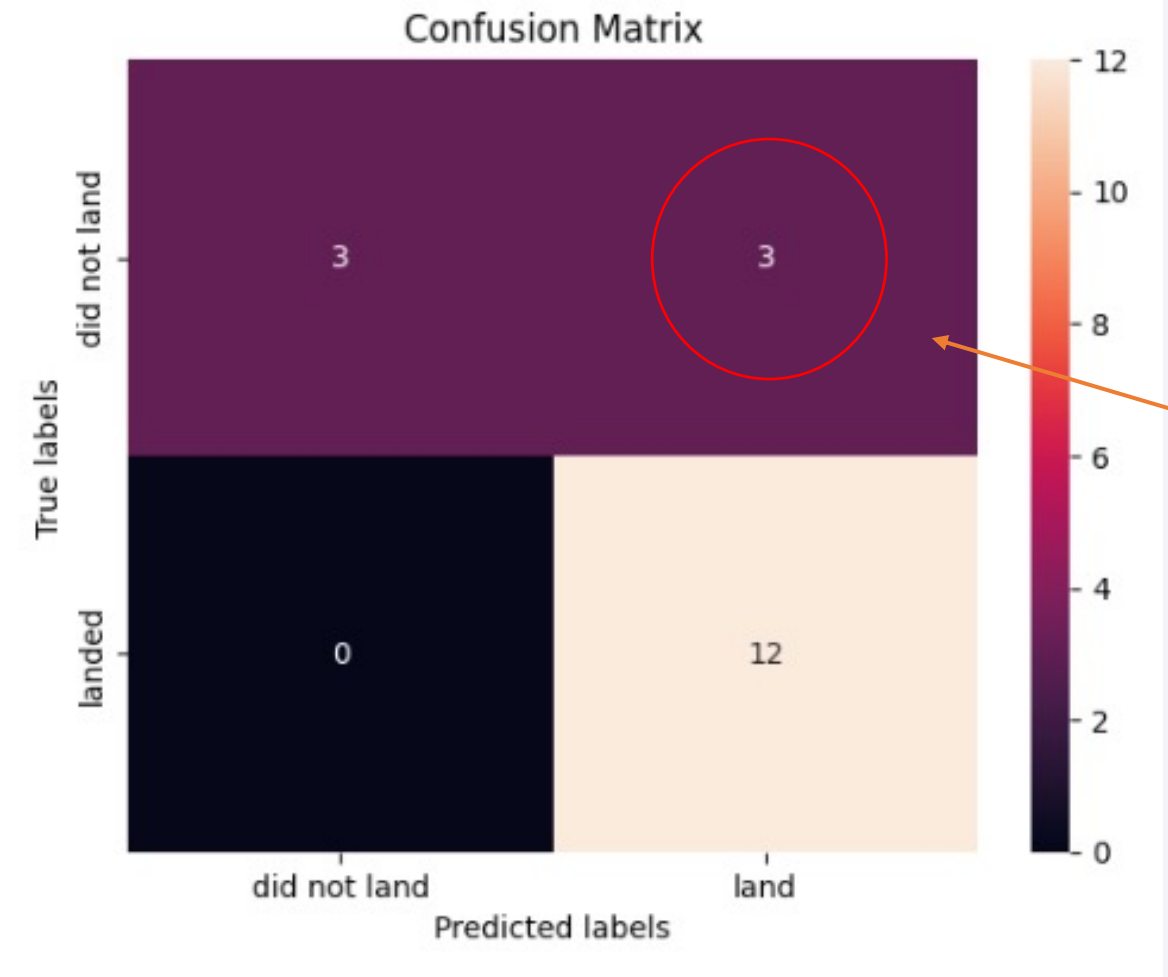
Classification Accuracy



- Decision Tree model has the highest best score
- ∴ Decision Tree model has the highest accuracy
- Best parameters:
  - criterion: gini
  - max depth: 2
  - max features: sqrt
  - min samples leaf: 1
  - min_samples_split: 10
  - splitter: best

```
best model: decision tree
best score: 0.8892857142857145 best parameters: {'criterion': 'gini', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'best'}
```

Confusion Matrix of Decision Tree Model



- Confusion matrix shows model ability to distinguish between different classes
- Main issue lies with false positives (predicted to have landed when it has not)

Conclusions

- The higher the flight number, the higher the success rate in landing first stage.
- Lighter payload correlates with higher success rate in landing first stage.
- Orbits with the highest success rates: ES-L1, GEO, HEO, SSO
- Orbit with the lowest success rate: SO
- Launch site KSC LC-39A has the highest success rate (76.9%)
- Launch sites are close to coastlines and railways, far from highways and cities on average.
- Decision Tree is the best algorithm for this prediction problem

Thank you!