Winning Space Race with Data Science

Taer Chan
10 December 2022

Outline

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

Executive Summary

- Summary of methodologies

Data collection using web scraping and SpaceX API

Data wrangling

Exploratory Data Analysis with Data Visualization & SQL

Building an interactive map with Folium

Building a Dashboard with Plotly Dash

Classification Prediction

- Summary of all results

Exploratory Data Analysis results

Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Delivery and presentation of the data-driven insights to determine if the first stage of Falcon 9

will land successfully

Introduction

- Project background and context

By assuming the role of a Data Scientist working for a startup intending to compete with SpaceX, we are asked to predict if the Falcon 9 first stage will land successfully.
We will have to evaluate the viability of the new company SpaceY to compete with SpaceX.

- Problems you want to find answers

We want to determine if the first stage will land in order to determine the cost of a launch.
The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
Where is the best place to make launches.

Section 1

Methodology

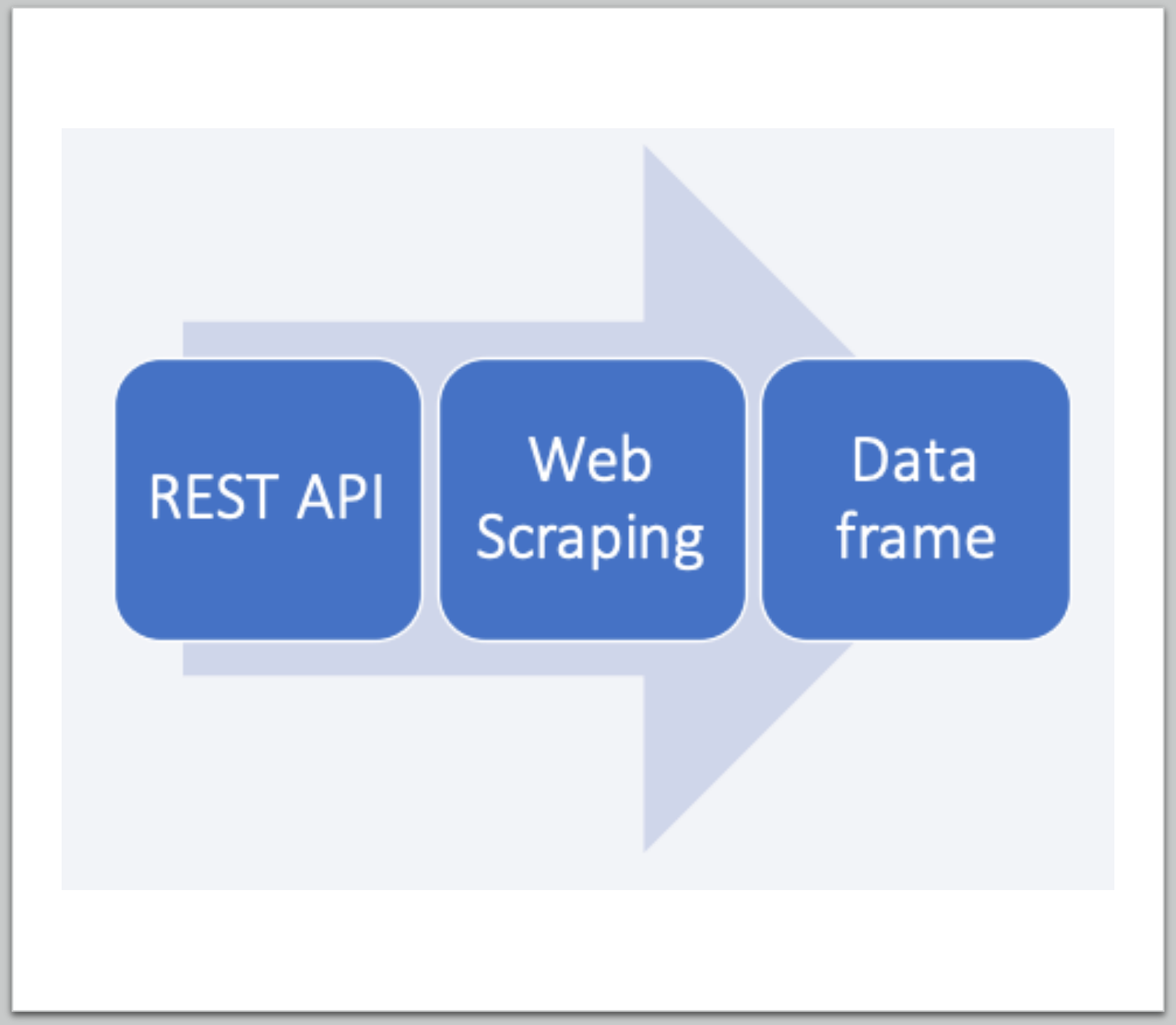
Methodology

Executive Summary

- Data collection methodology:
  - The data was collected through web scraping and the use of a REST API
  - Space X API
    (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
  - Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
    Exploratory data analysis was performed, and training labels were created
  - Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The classification models were trained on training data and tested on test data - with the
  - use of confusion matrices and the mean accuracy metric

Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/)
- and from Wikipedia
- (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches),
- using web scraping technics like beautifulsoup



Data Collection – SpaceX API

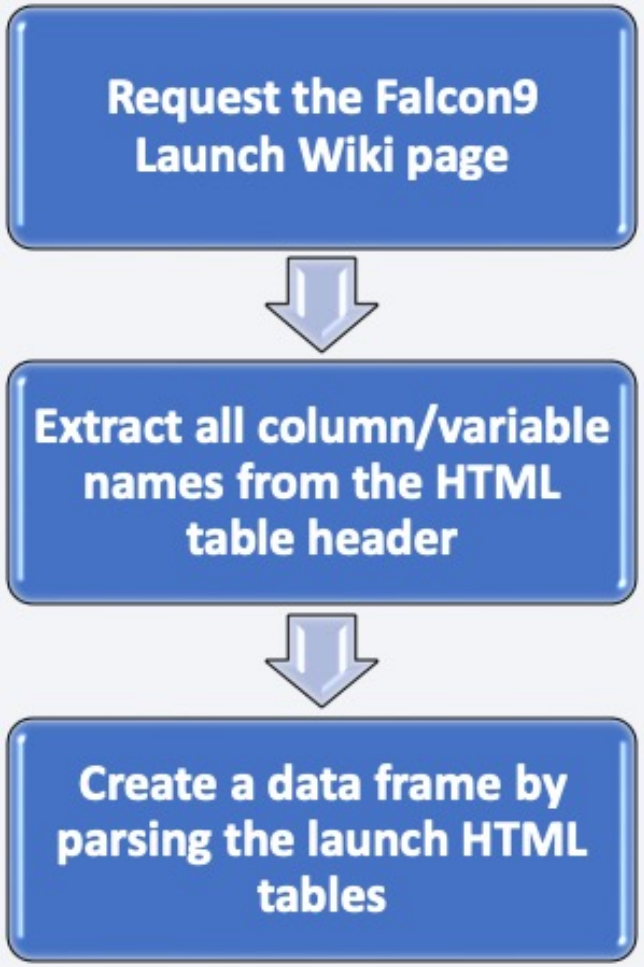
- A SpaceX public URL was used to target a specific endpoint of the API to get past launch data.
- This API was used according to the flowchart beside and then data is persisted.
- We need to clean the requested data.

- https://github.com/taerchan/Applied-Data-Science-Capstone/blob/a99ab61a1e6d0a3bd15e69b07a6f1e9d3d60ce6b/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Data from SpaceX launches can also be obtained from Wikipedia

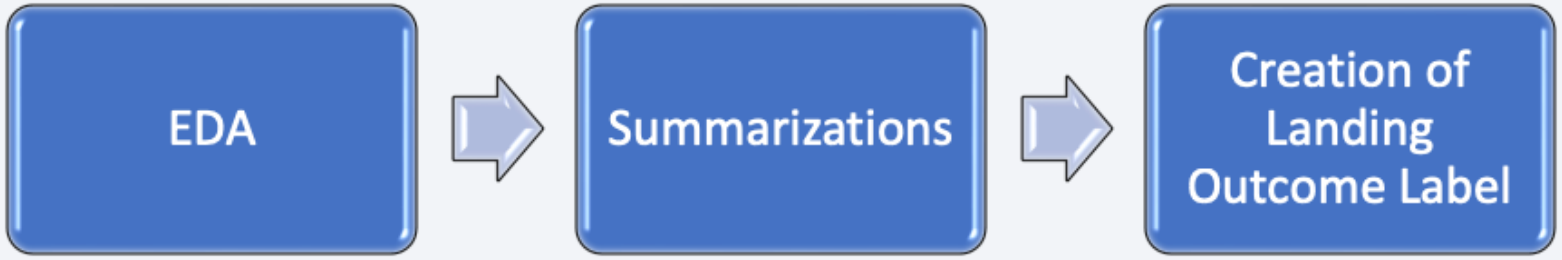
Web scrap Falcon 9 launch records with BeautifulSoup
Extract a Falcon 9 launch records HTML table from Wikipedia
Parse the table and convert it into a Pandas data frame



https://github.com/taerchan/Applied-Data-Science-Capstone/blob/9116c611c1fb7e54eef1b27f6ad68c1004e71d1f/jupyter-labs-webscraping.ipynb

Data Wrangling

- Initially Exploratory Data Analysis (EDA) was performed on the dataset
- Load SpaceX dataset
- Identify and calculate the percentage of the missing values in each attribute
- Identify which columns are numerical and categorical (Outcome)



https://github.com/taerchan/Applied-Data-Science-Capstone/blob/cs50/problems/2022/python/project/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Using basic visualization tools such as bar charts, scatter point charts, and line charts
- Scatter plots were beneficial for seeing the relationship between two variables.
- Bar charts were plotted to visualize any association between orbit type and average success rate of the booster (first stage) landing successfully.
- Line plots are useful for visualizing changes in data over time, so a line plot was used to visualize any association between year and average success rate.

- https://github.com/taerchan/Applied-Data-Science-Capstone/blob/a7e68de3a4fe9829fe3d26d7e612668b4b5912c8/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Simple SELECT queries

Names of the unique launch sites in the space mission;

Top 5 launch sites whose name begin with the string 'CCA';

Total payload mass carried by boosters launched by NASA (CRS);

Average payload mass carried by booster version F9 v1.1;

Date when the first successful landing outcome in ground pad was achieved;

Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;

Total number of successful and failure mission outcomes;

Names of the booster versions which have carried the maximum payload mass;

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;

Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

https://github.com/taerchan/Applied-Data-Science-Capstone/blob/46ac924b05146f2dcfbc2a8bba376e28a849ce00/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- In order to provide powerful visualization features, we created and added to the folium maps the following objects:
- Map: to initialize a world map representation
- Circle: to add a highlighted circle area with a text label on a specific coordinate
- Marker: to easily identify which launch sites have relatively high success rates

- https://github.com/taerchan/Applied-Data-Science-Capstone/blob/6bd718c16a5c50eb8c974a1bb5dc10e87fc2b3f9/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

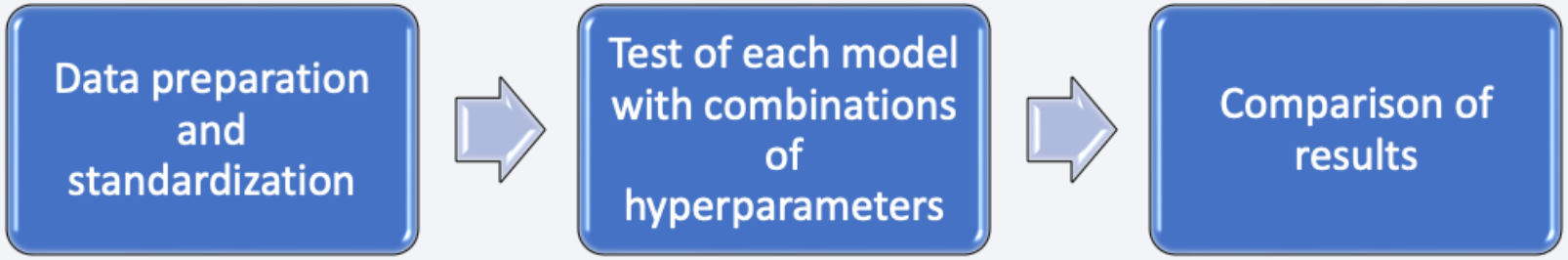
Build a Dashboard with Plotly Dash

- The dashboard application that was developed, contains a pie chart and a scatter point chart which are manipulated by dropdown and range slider functionalities.
- Percentage of launches by site
- Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- https://github.com/taerchan/Applied-Data-Science-Capstone/blob/4624b7cc14c112f66fd814fd77b57f5d56f0785c/spacex_dash_app.py

Predictive Analysis (Classification)

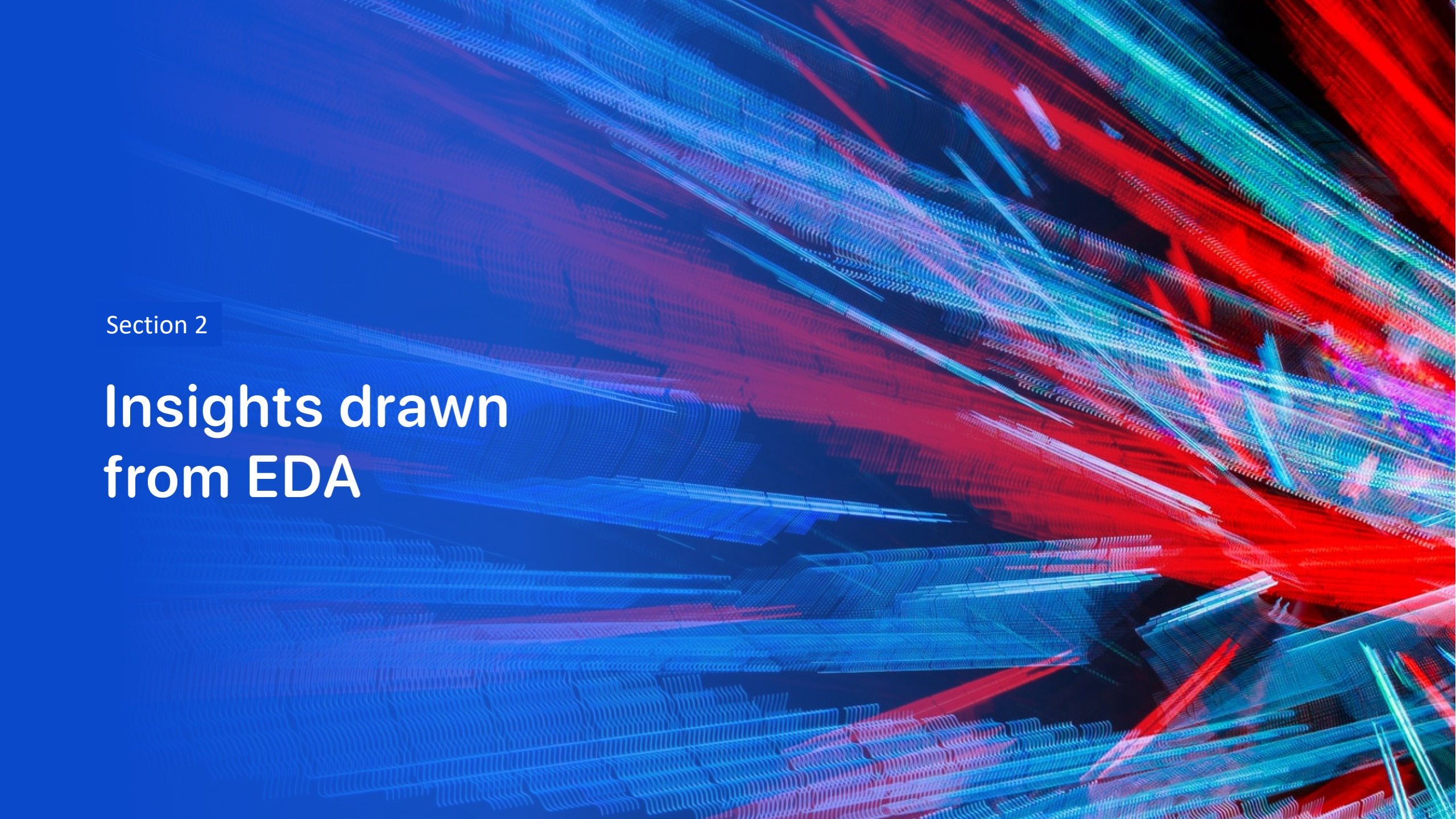
- The data was standardized, and a train-test data split was performed.
- Grid search was used to choose the best hyperparameters for each machine learning (ML) algorithm.
- Evaluate the optimum models of each method by checking the test accuracy and monitoring the confusion matrices
- Choose the best among the optimum models considering the test accuracy criterion as critical



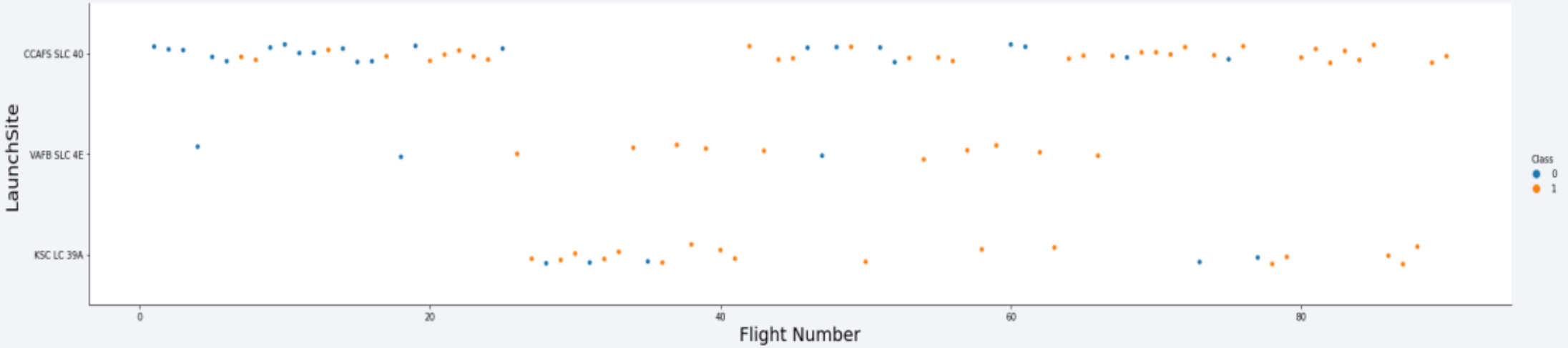
https://github.com/taerchan/Applied-Data-Science-Capstone/blob/de71a3185a376ec2bfa763023e014fe8785f184e/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- The success rate since 2013 kept increasing till 2020.

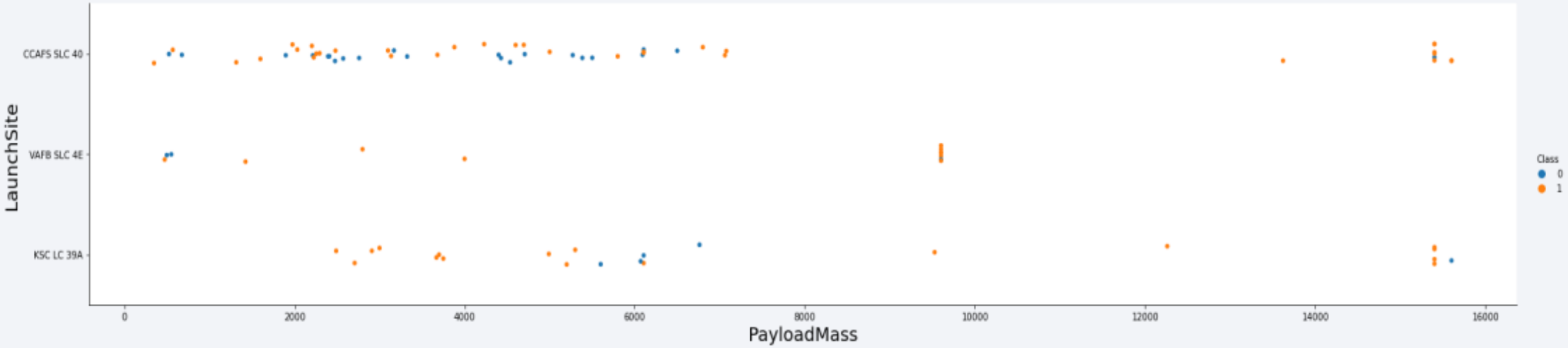

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

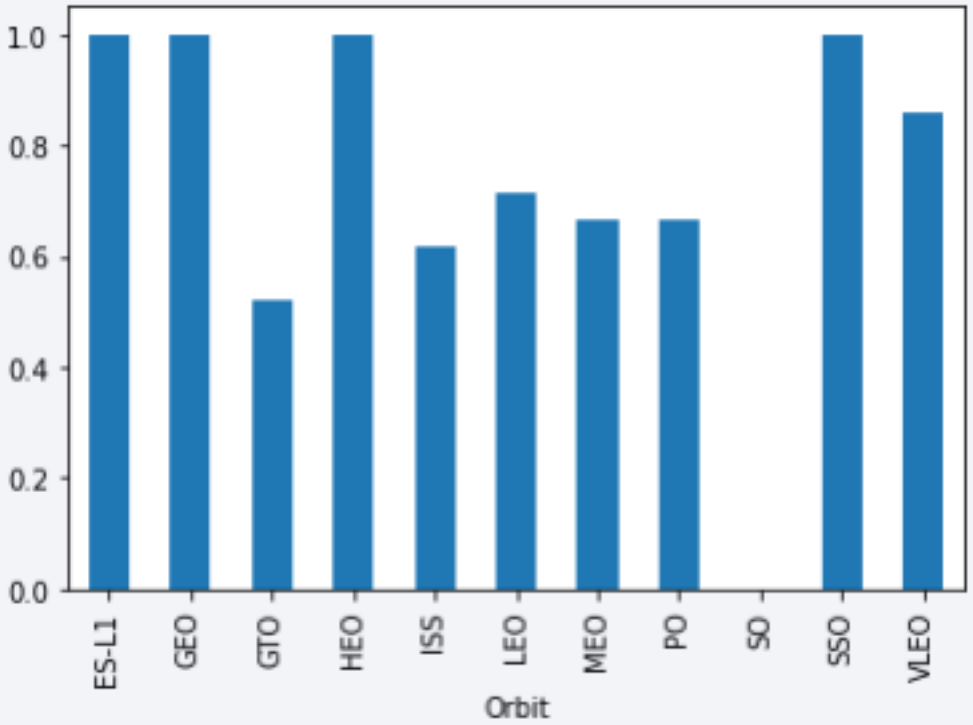
Payload vs. Launch Site



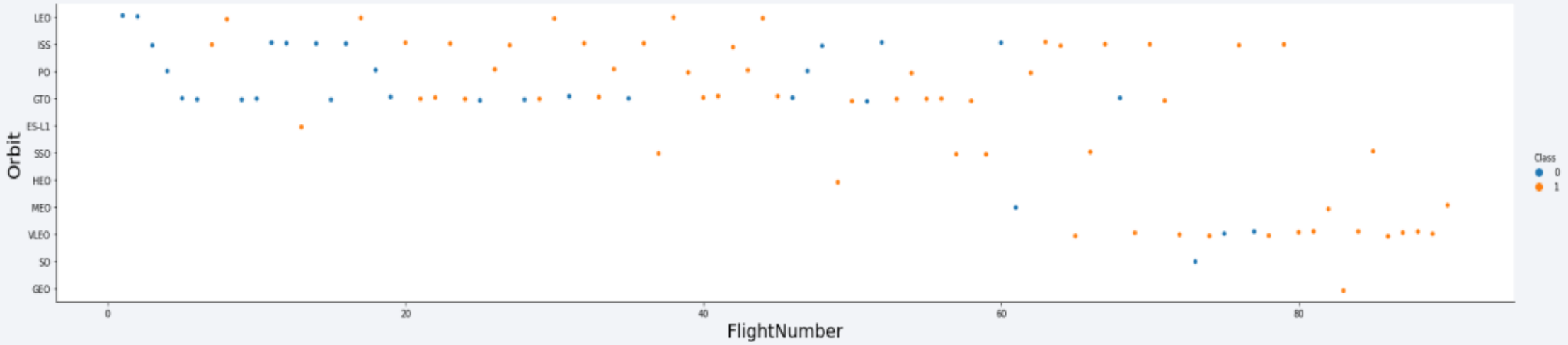
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- Orbit type GEO,ES-L1, SSO, HEO have the highest average success rate.
- Orbit type GTO has the lowest average success rate.

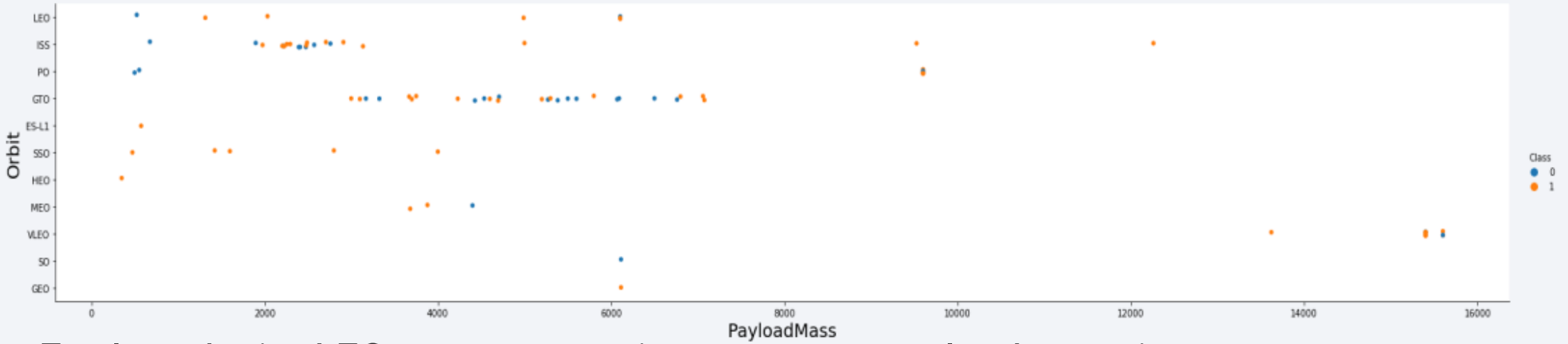


Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of flights; on the other
- hand, there seems to be no relationship between flight number when in GTO orbit.

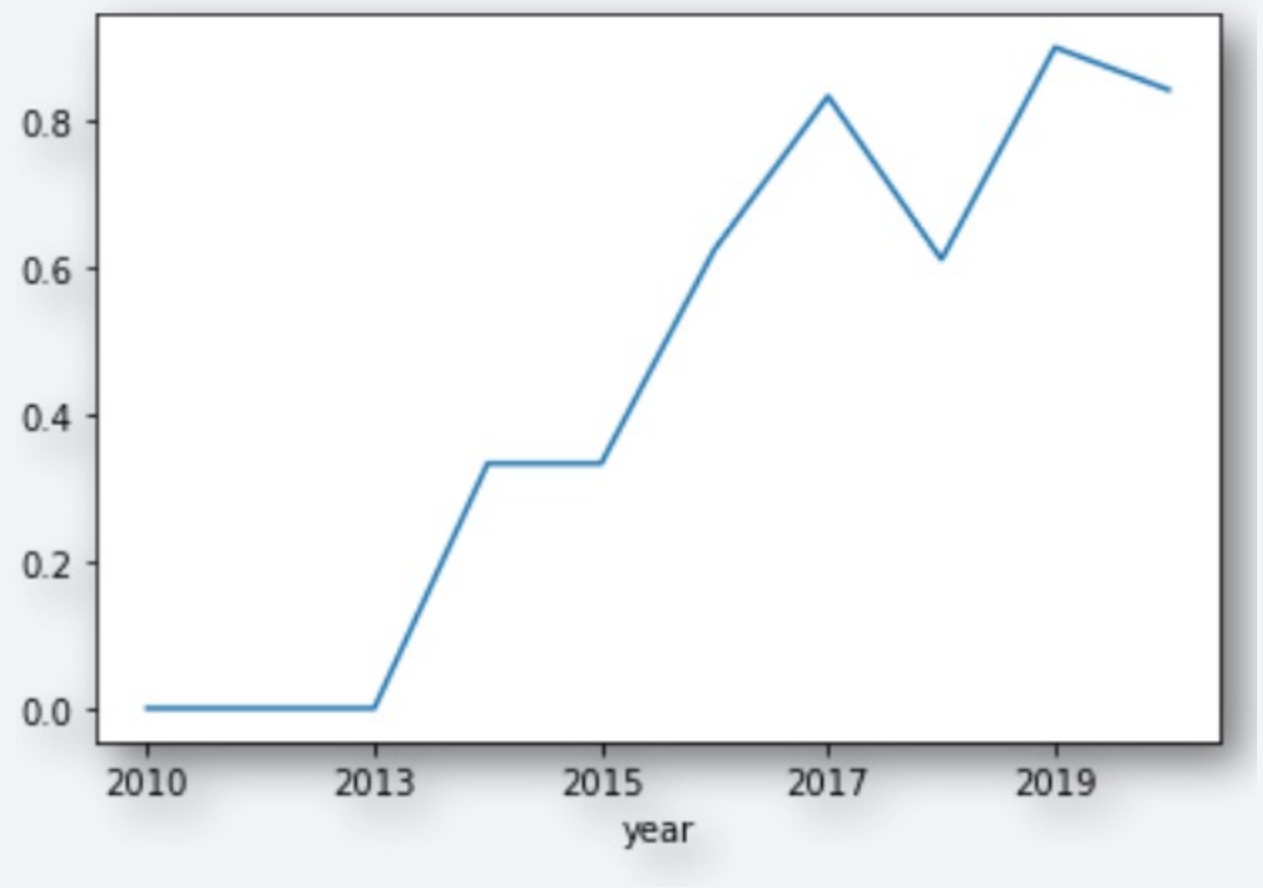
Payload vs. Orbit Type



- For launch site LEO, success rate is greater as pay load mass increases.
- For orbit type GTO, as pay load mass increases there is no effect on success rate (for better or worse).

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to data, there are four launch sites:
- Present your query result with a short explanation here

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

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

```
%%sql

SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
```

Launch Site Names Begin with 'CCA'

- 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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS 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	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

- Total payload carried by boosters from NASA:
- A SELECT statement with a SUM function and WHERE clause was used to find the total payload mass.

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_carried_by_boosters_launched_by_NASA_in_kg FROM SPACEX
WHERE CUSTOMER = 'NASA (CRS)';
```

total_payload_mass_carried_by_boosters_launched_by_nasa_in_kg
45596

Average Payload Mass by F9 v1.1

- A SELECT statement with the AVG function and a WHERE clause was used to calculate the average pay load mass by booster version F9v1.

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass_carried_by_booster_version_F9_v1_1_in_kg
FROM SPACEX
WHERE BOOSTER_VERSION LIKE 'F9 v1.1';
```

average_payload_mass_carried_by_booster_version_f9_v1_1_in_kg
2928

First Successful Ground Landing Date

- By using the MIN function and WHERE clause in the SELECT statement, the date of the first successful landing outcome in ground pad was obtained.

```
%%sql

SELECT MIN(DATE)
AS DATE_WHEN_THE_FIRST_SUCCESSFUL_LANDING_OUTCOME_IN_GROUND_PAD_WAS_ACHIEVED
FROM SPACEX
WHERE LANDING__OUTCOME LIKE 'Success (ground pad)';
```

date_when_the_first_successful_landing_outcome_in_ground_pad_was_achieved
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- A SELECT statement with a WHERE clause was used to obtain successful drone ship landings with payload between 4000 and 6000.

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

```
%%sql

SELECT BOOSTER_VERSION AS LIST_OF_BOOSTERS_NAMES
FROM SPACEX
WHERE LANDING__OUTCOME = 'Success (drone ship)' AND payload_mass__kg_<6000 AND payload_mass__kg_>4000;
```

Total Number of Successful and Failure Mission Outcomes

- • Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass
- A subquery was used to obtain the names of the booster versions that have carried the maximum payload mass.

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

Booster Version
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

- A SELECT statement was used with the required column names and WHERE clause to list the failed landing outcomes in 2015.

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
%%sql

SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX where LANDING__OUTCOME = 'Failure (drone ship)' AND year(date)=2015;
```

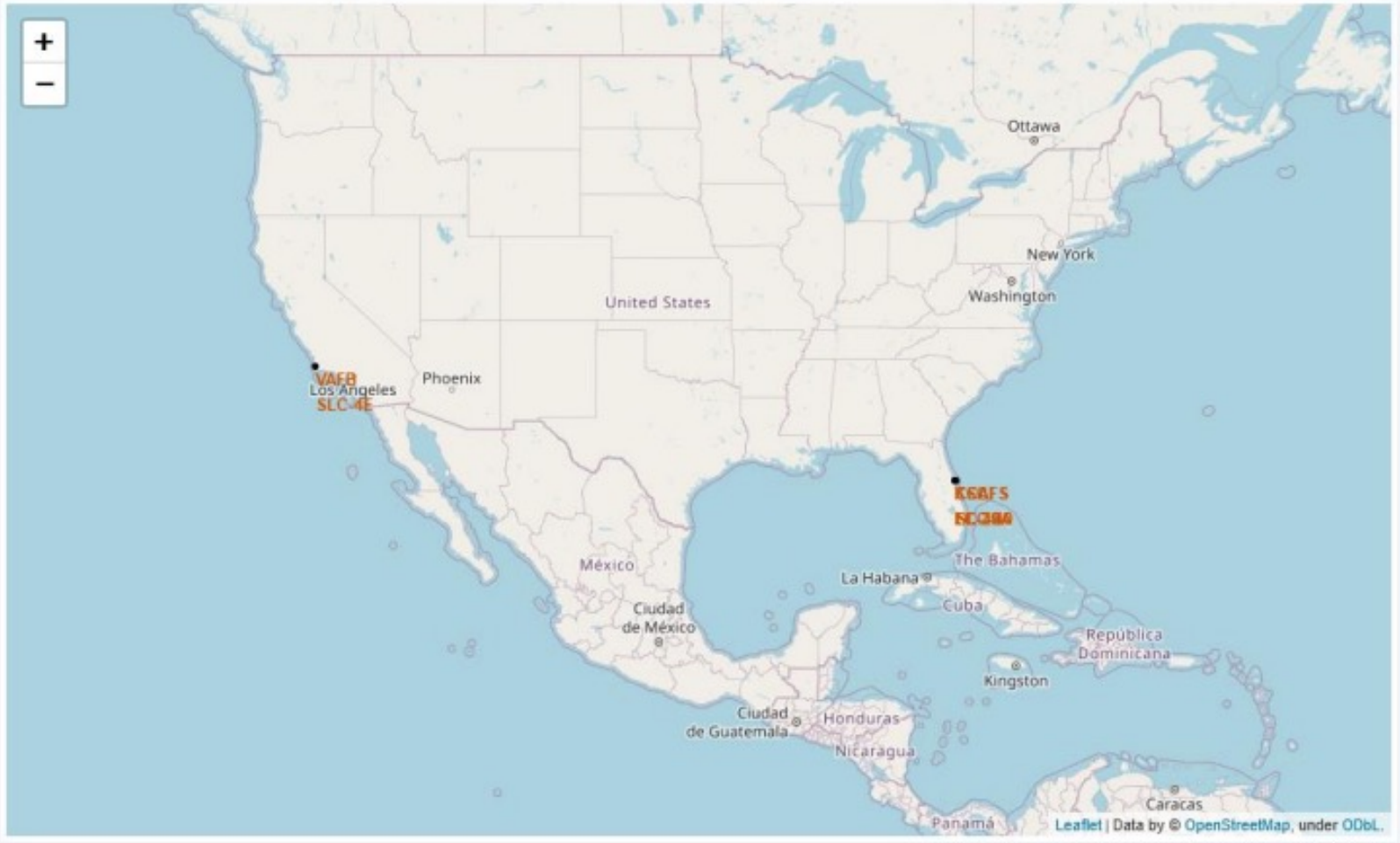
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- A SELECT statement with the WHERE, GROUP BY, and DESC commands was used to rank the landing outcomes between 2010-06-04 and 2017-03-20.

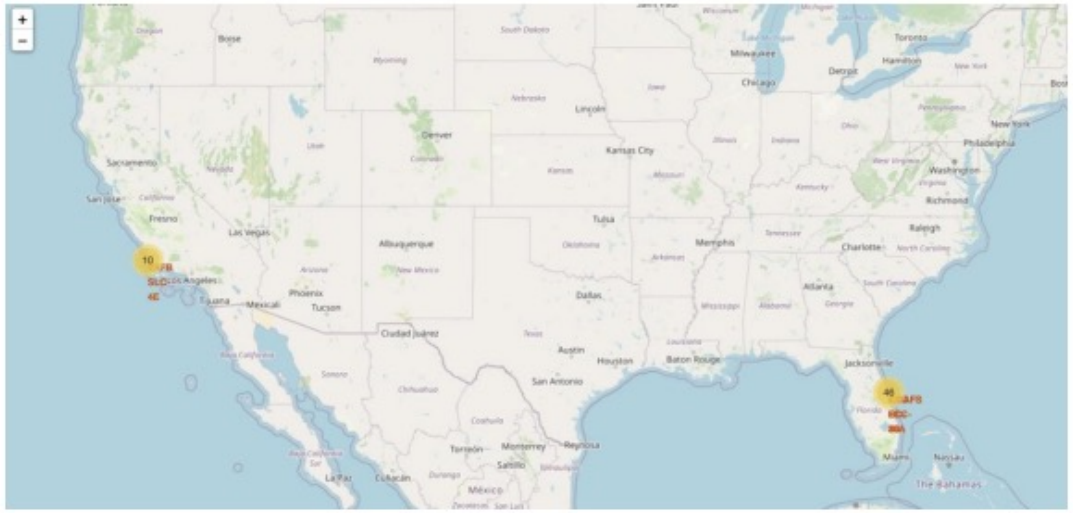
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

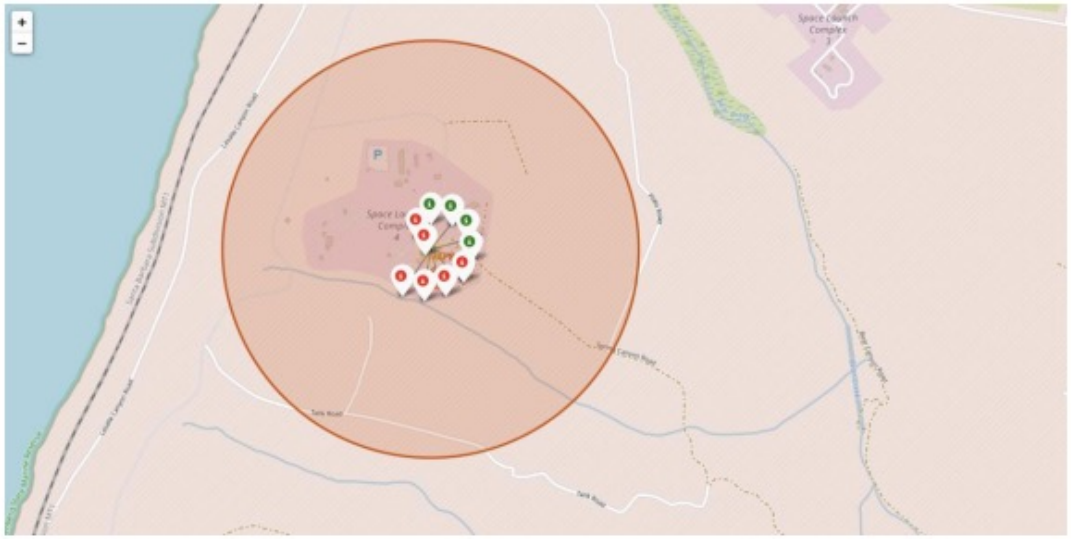

Section 3
Launch Sites
Proximities Analysis

Launch Site Locations

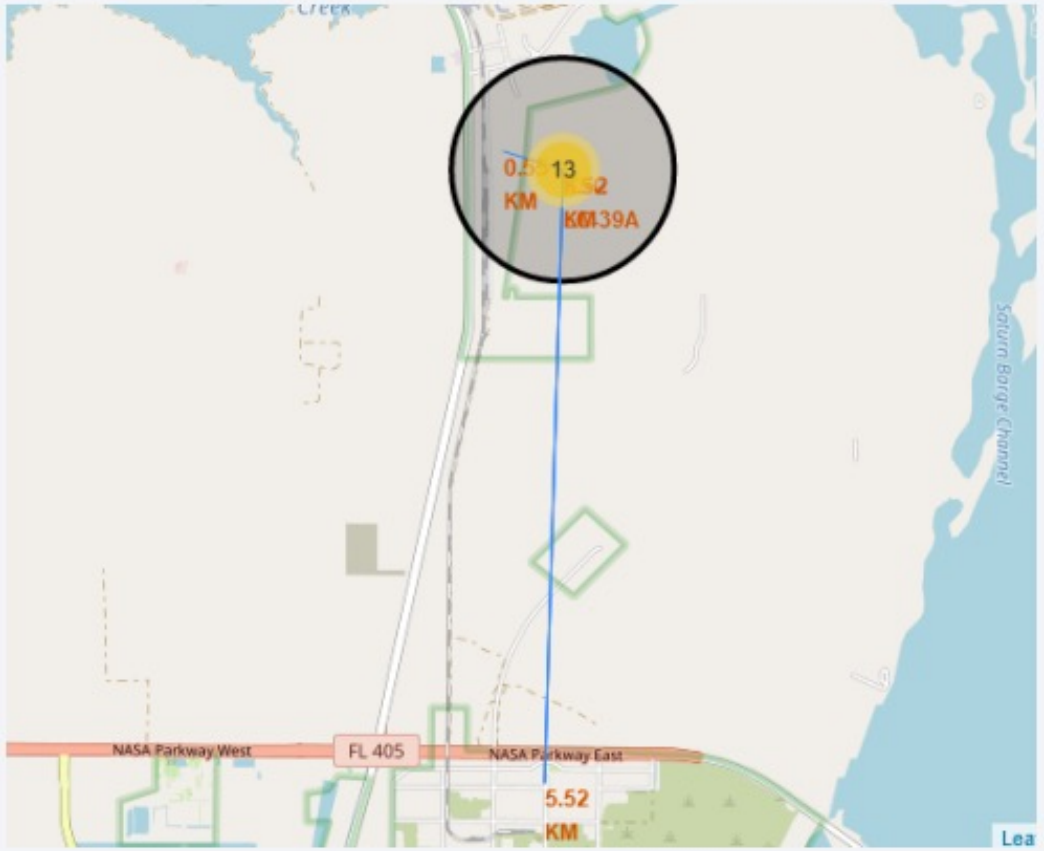


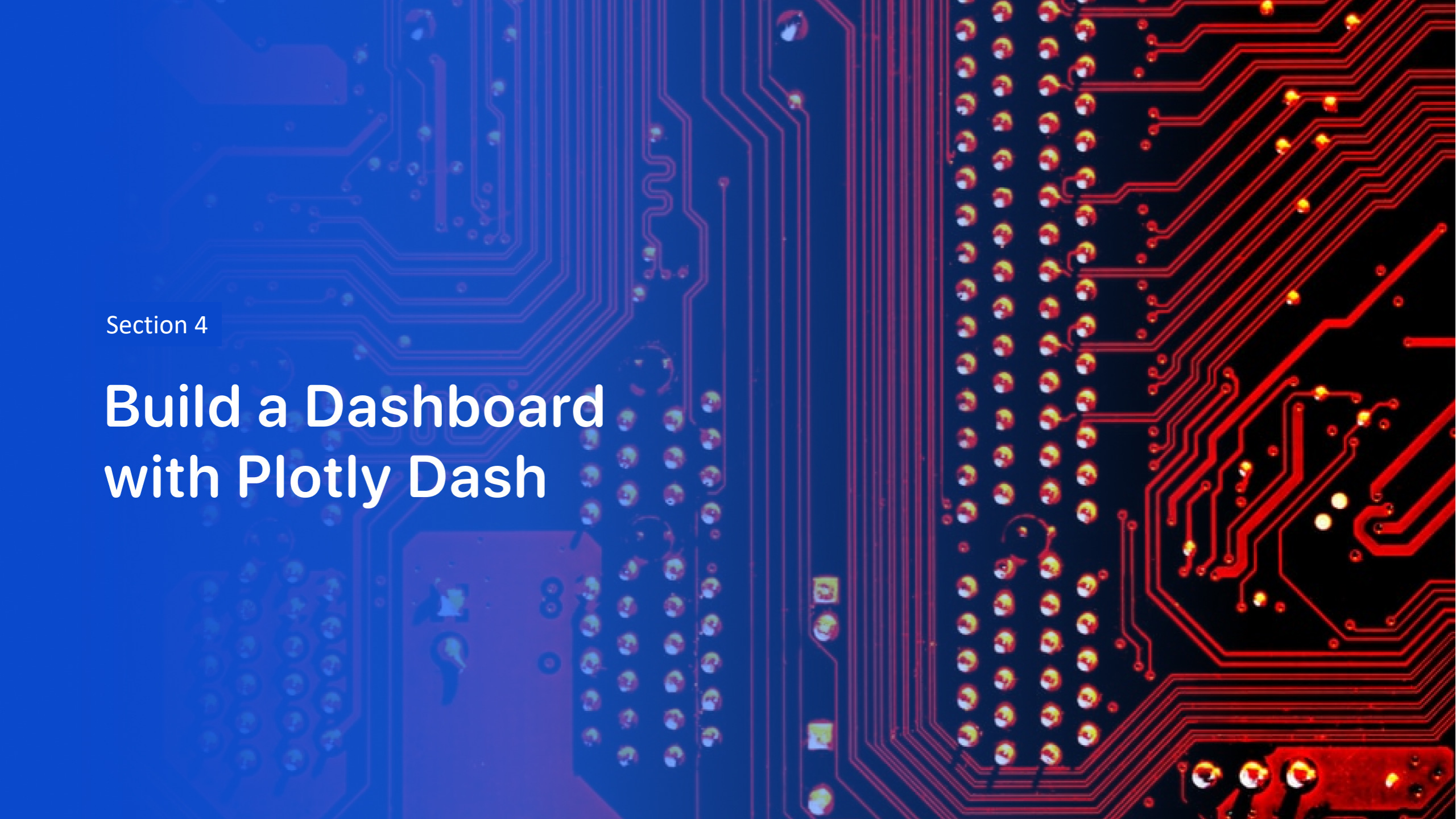
Successful and Failed Launches for each Launch Site



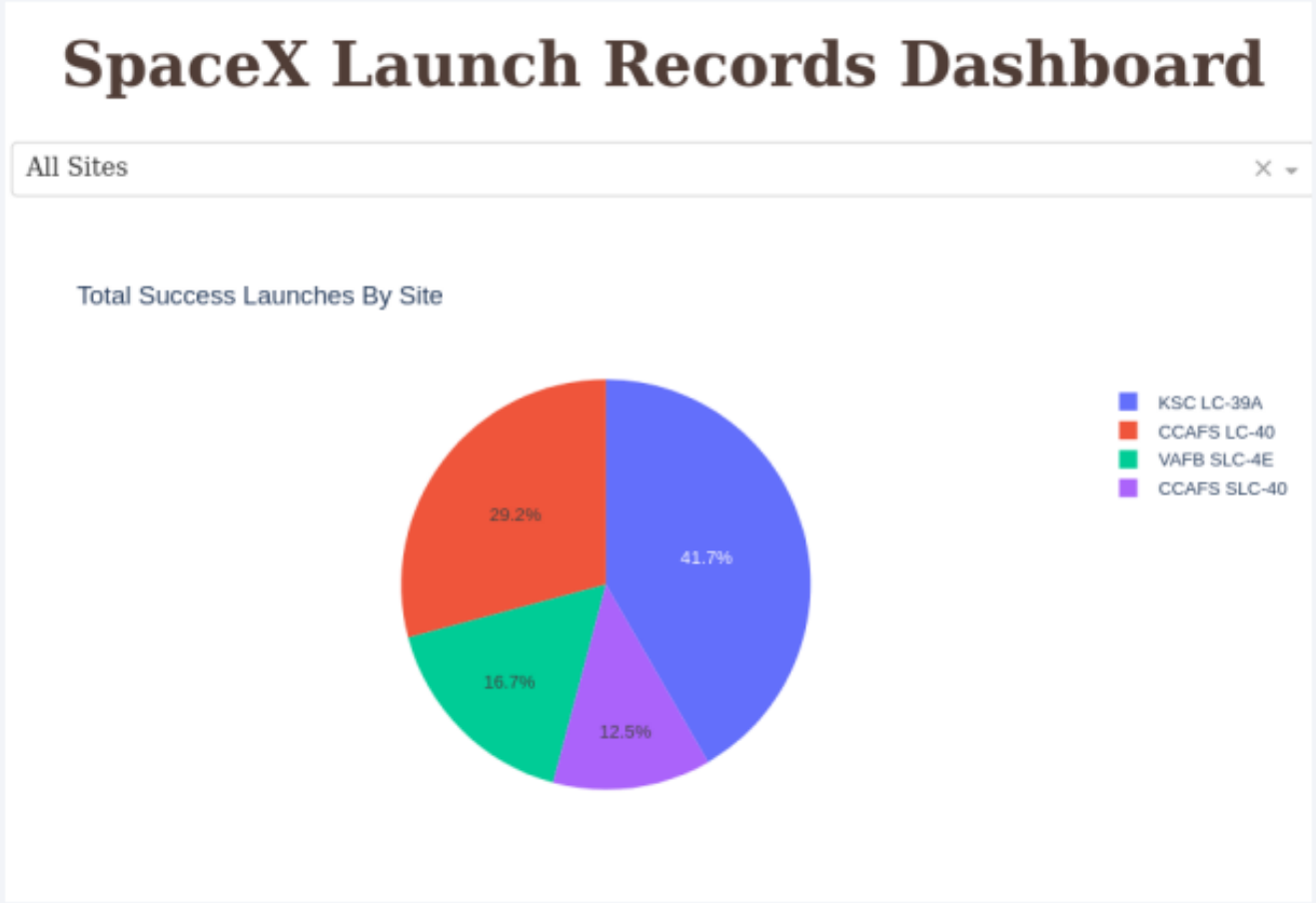


Logistics and Safety

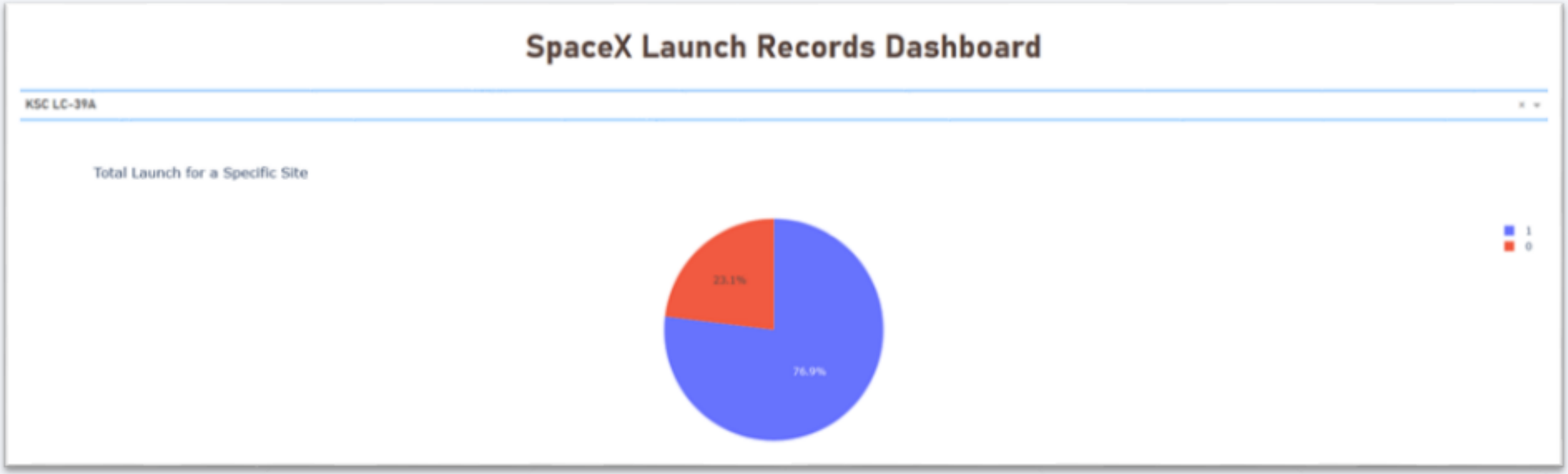



Section 4
Build a Dashboard with Plotly Dash

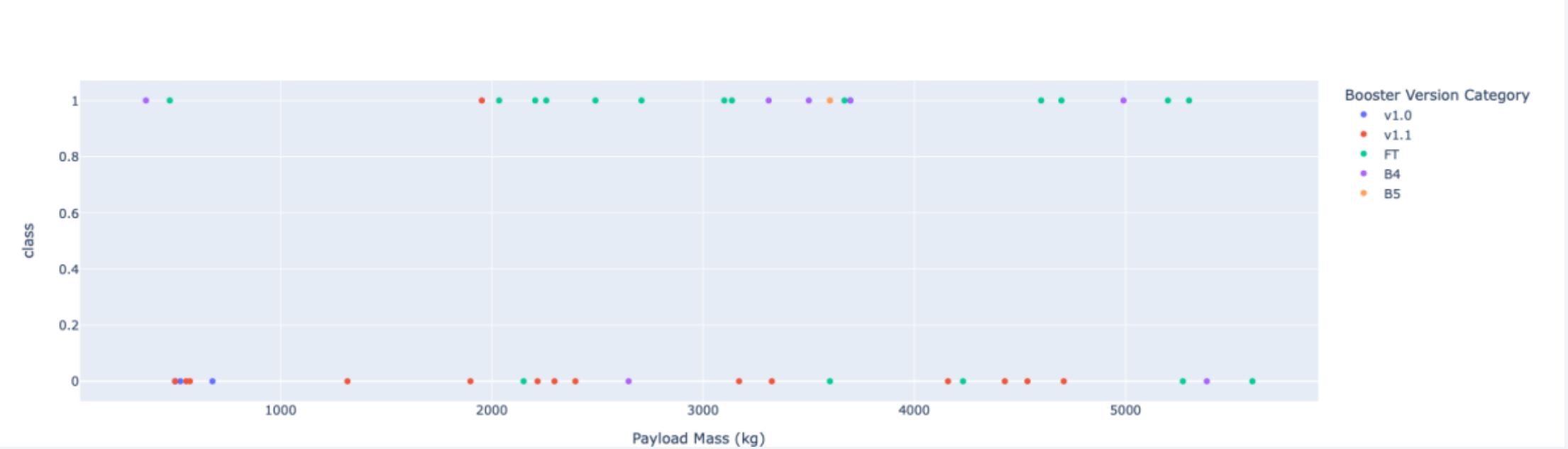
Successful Launches by Site



Success ratio for the most successful launch site



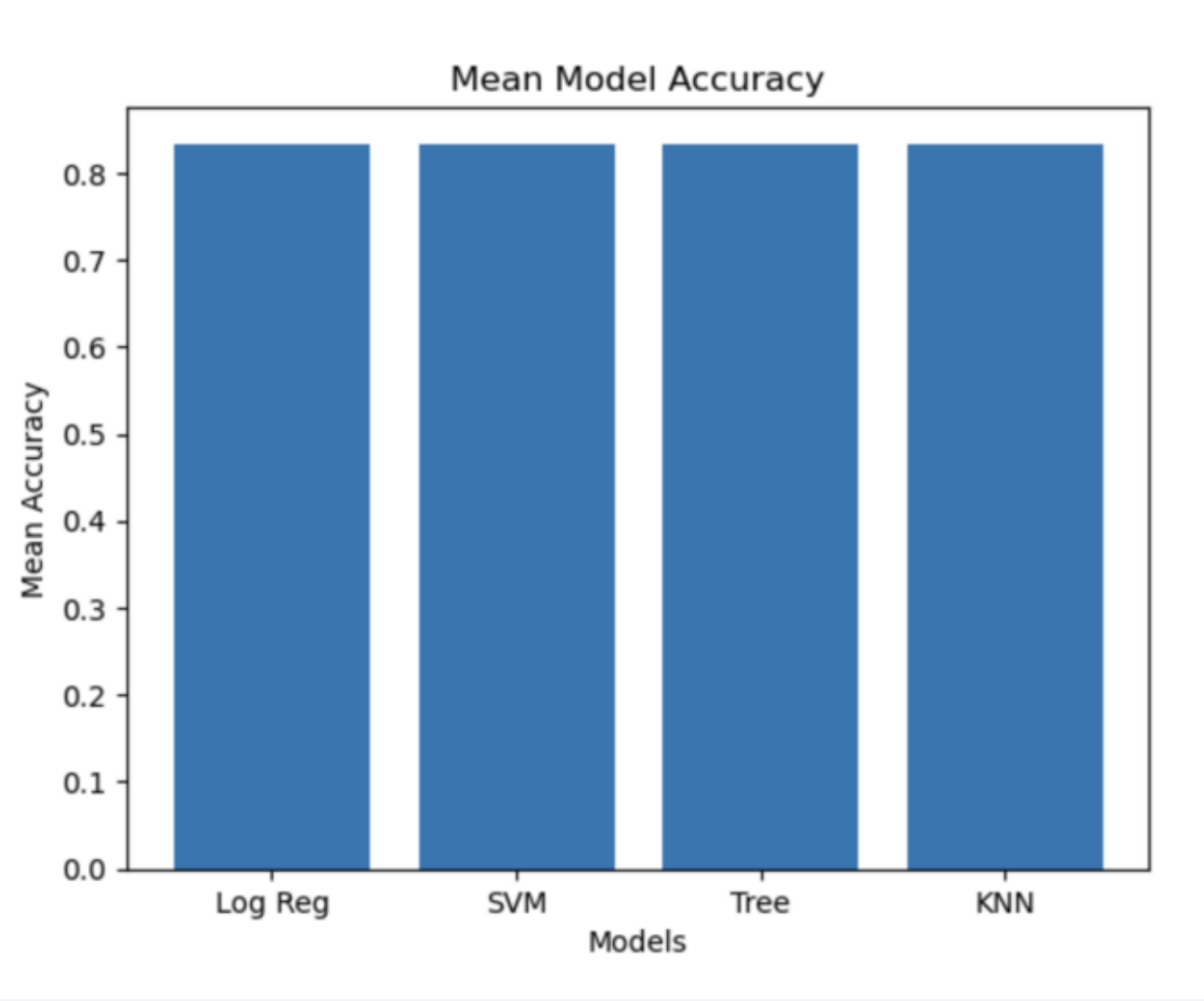
Scatter Plot of Payload vs Launch Outcome for all Sites.



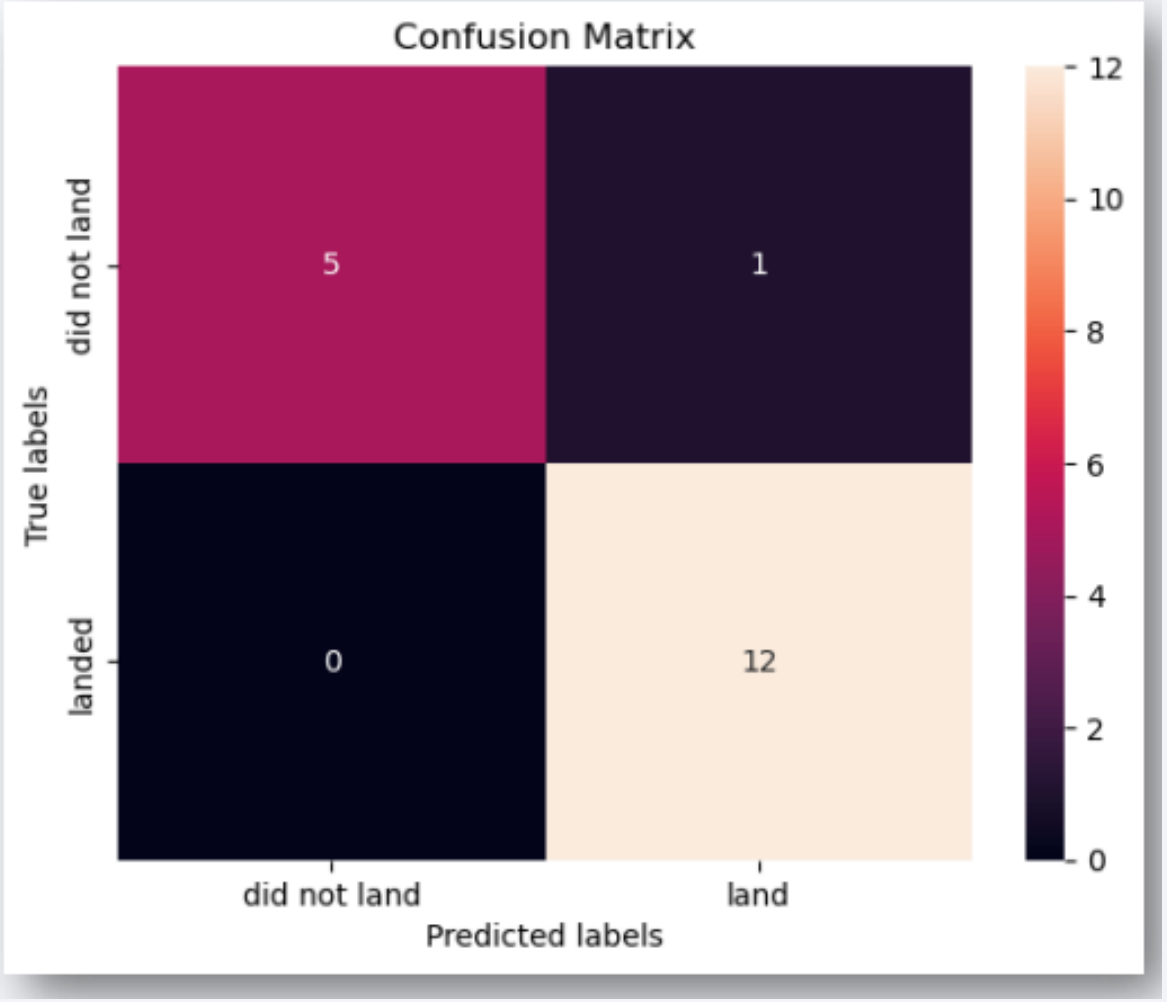
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- Overall, the models that were built were fairly accurate.
- Proximity to logistic infrastructures is a top priority.
- Launch sites might be near to the sea for security reasons.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Orbit selection is an important task as it seems that there is a relation with the outcome of the missions.

Appendix

- Link to project repository:

https://github.com/taerchan/Applied-Data-Science-Capstone

Thank you!