IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Outline

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

Executive Summary

- Summary of methodologies
    - Data collection using API and Web Scraping
    - Data Wrangling
    - EDA with Data Visualization and SQL
    - Interactive visual analytics using Folium and Plotly Dash
    - Predictive analysis using Classification

- Summary of all results
    - EDA Results
    - Interactive analytics results
    - Predictive analysis results

Introduction

- Background and context
  - Space X has the best launch prices for Falcon 9 rockets ($62 million vs. $165 million USD for others) due to the reuse of the first stage.
  - Space Y wants to compete with Space X

- Problem
  - It is necessary to study the success stories Space X and the best possible launch sites so that Space Y will have lower failure rate.

Section 1

Methodology

Methodology

Executive Summary

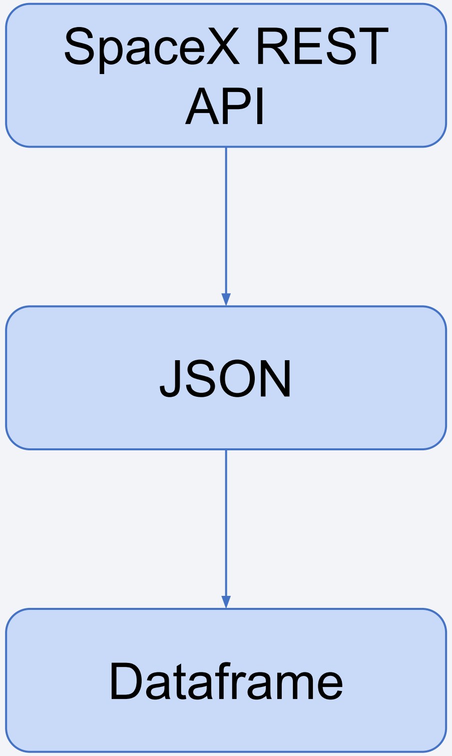
- Data collection methodology:
    - Data was collected from REST API and web scraping.
- Perform data wrangling
    - The collected data were cleaned up and converted into a format that can be summarized in the final data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - The collected data were splited, different classification models were built, and their accuracy was evaluated.

Data Collection

- The process of data collection
  - Data collection with SPACE X API using the Rest API, then data normalization to transform it to the dataframe.
  - Data collection from Wikipedia using Web Scraping Beautiful Soap technique, then data normalization to transform it to the dataframe.

Data Collection – SpaceX API

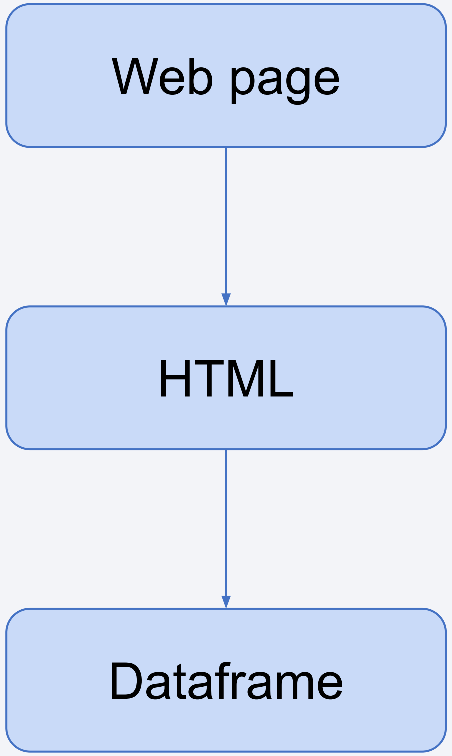
- SPACE X API: https://api.spacexdata.com/v4



https://github.com/Nazalekser/IBM-Data-Science/blob/main/10-1-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

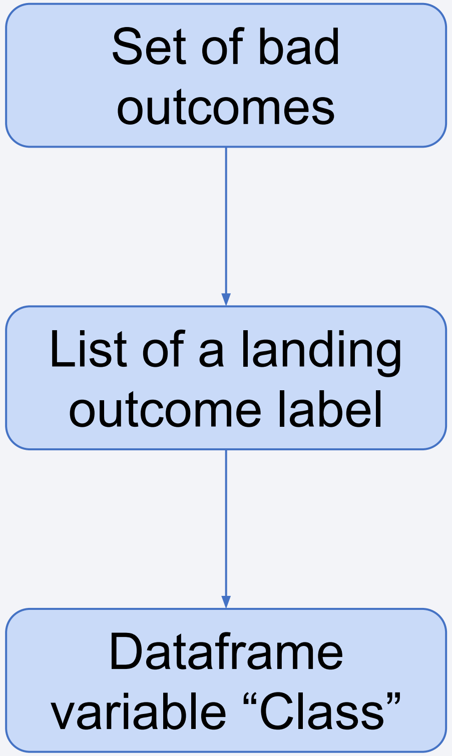
- Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



https://github.com/Nazalekser/IBM-Data-Science/blob/main/10-2-jupyter-labs-webscraping.ipynb

Data Wrangling

- Creating target variable “Class”, that represents the outcome of each launch (“0” - bad outcome, “1” - otherwise)



https://github.com/Nazalekser/IBM-Data-Science/blob/main/10-3-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- SCATTER GRAPHS: represents the data from two or more variables.
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload mass vs. Launch Site
  - Flight Number vs. Orbit
  - Payload mass vs. Orbit

- BAR GRAPHS: compare things between different groups.
  - Orbit vs. Success rate

- LINE GRAPHS: show information that changes over time.
  - Year vs. Success rate

https://github.com/Nazalekser/IBM-Data-Science/blob/main/10-5-jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Displaying the names of the unique launch sites in the space mission;
- Displaying 5 records where launch sites begin with string ‘CCA’;
- Displaying the total payload mass carried by boosters launched by NASA (CRS);
- Displaying average payload mass carried by booster version F9 v1.1;
- Listing the date where the successful landing outcome in drone ship was achieved;
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 400 and less than 6000;
- Listing the total number of successful and failure mission outcomes;
- Listing the names of the boosters_versions which have carried the maximum payload mass;

EDA with SQL

- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
- Ranking the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

https://github.com/Nazalekser/IBM-Data-Science/blob/main/10-4-jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Markers were used to indicate launch locations.
- Circles were used to highlight coordinates.
- Marker clusters were used in groups of events.
- Lines were used to calculate the distance between the launch sites.

https://github.com/Nazalekser/IBM-Data-Science/blob/main/10-6-lab_jupyter_launch_site_location.ipynb

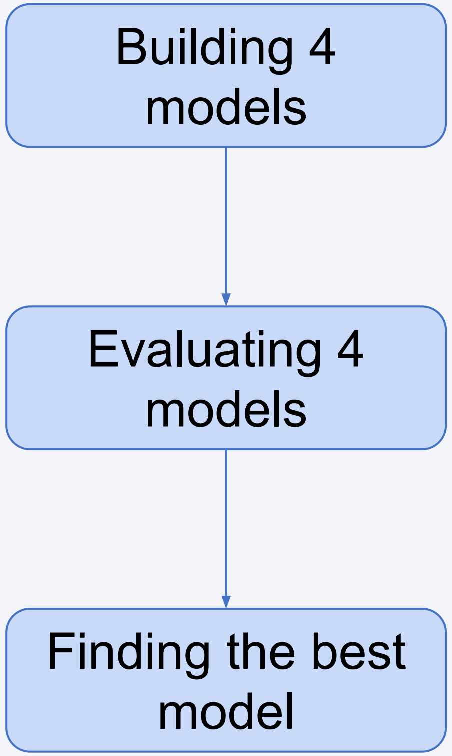
Build a Dashboard with Plotly Dash

- An interactive Pie Chart was used to show the Total Successful Launches by Site.
- An interactive Payload Range from 0 to 10,000 kg was used to see how the Success count on Payload Mass for All Sites depends on the Payload Mass, as well as on the Booster Version Category.

https://github.com/Nazalekser/IBM-Data-Science/blob/main/10-7-lab_theia_plotly_dash.ipynb

Predictive Analysis (Classification)

- BUILDING MODEL
  - Load and transform the data
  - Split the dataset
  - Implementation of the Grid Search algorithm and model training
- EVALUATING MODEL
  - Calculate accuracies and plot confusion matrices
- FINDING THE BEST MODEL
  - Choose the model with the best accuracy score



https://github.com/Nazalekser/IBM-Data-Science/blob/main/10-8-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

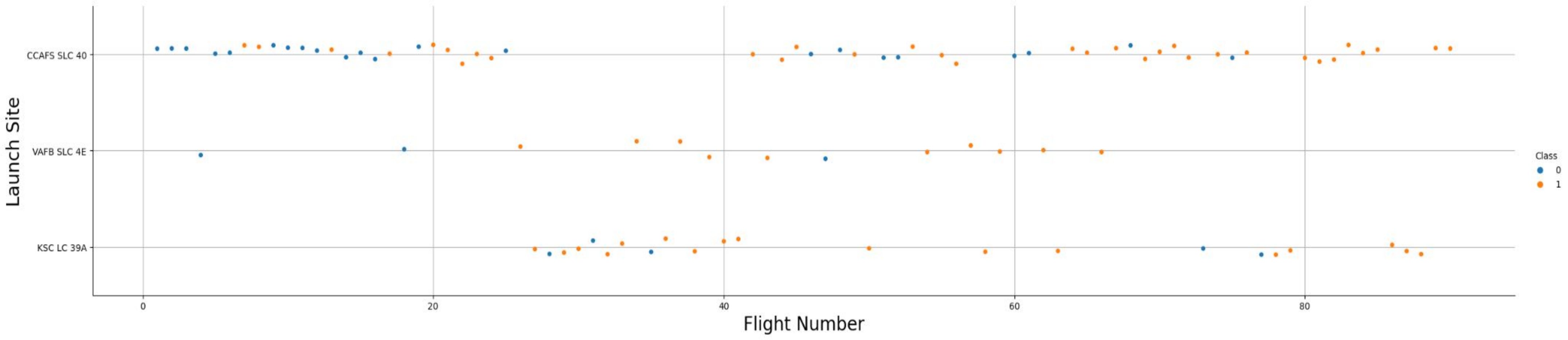
Results

- Exploratory data analysis results
- Interactive analytics results
- Predictive analysis results

Section 2

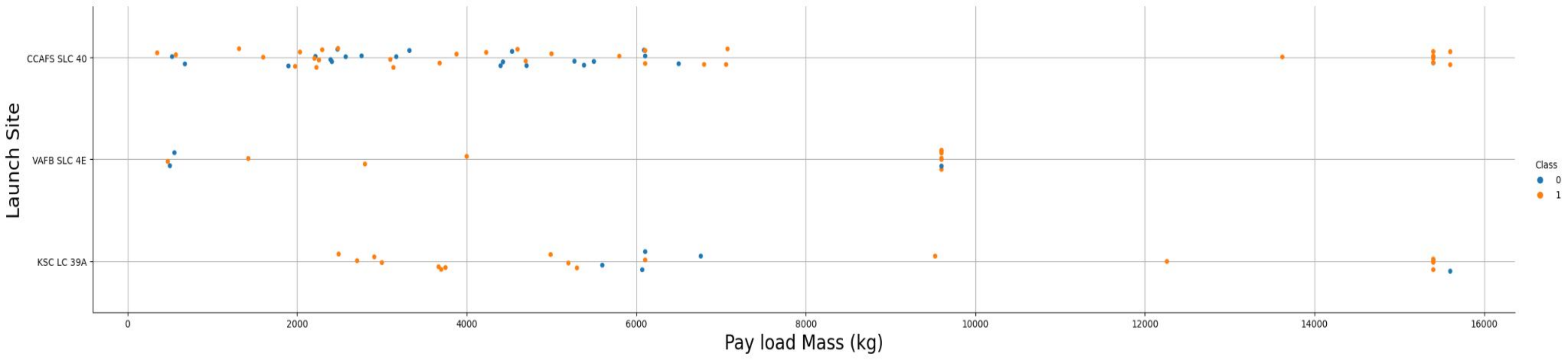
Insights drawn from EDA

Flight Number vs. Launch Site



- As the number of flights increases, so does the number of successful flights, indicating that the flaws in the rocket have been corrected.

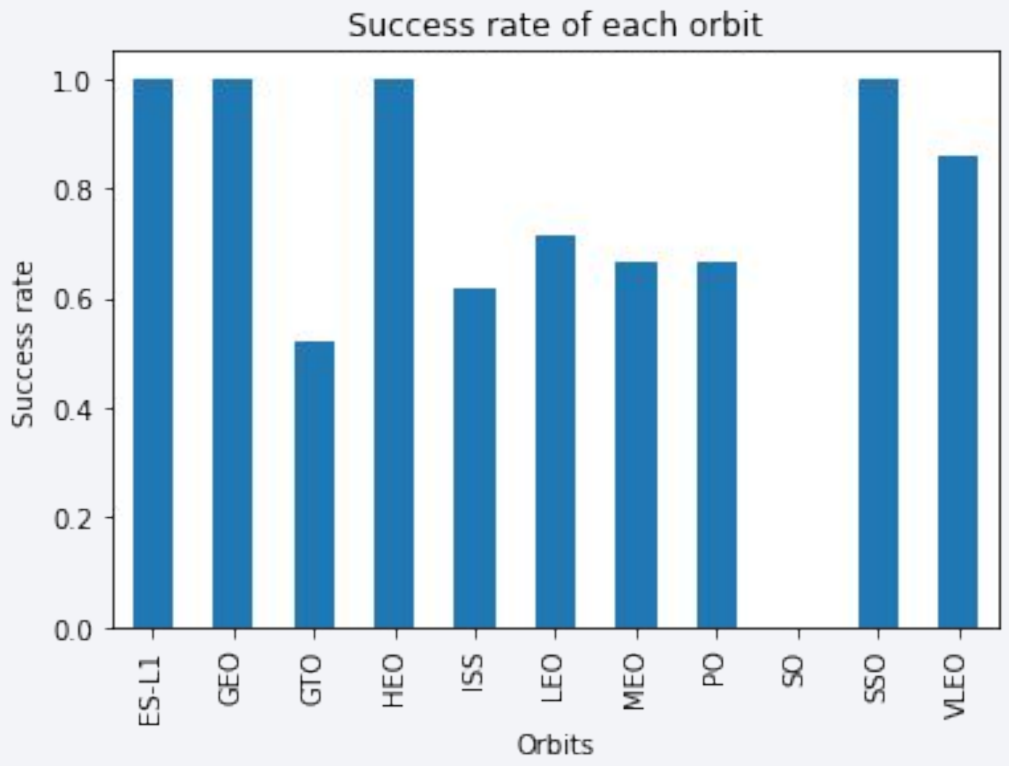
Payload vs. Launch Site



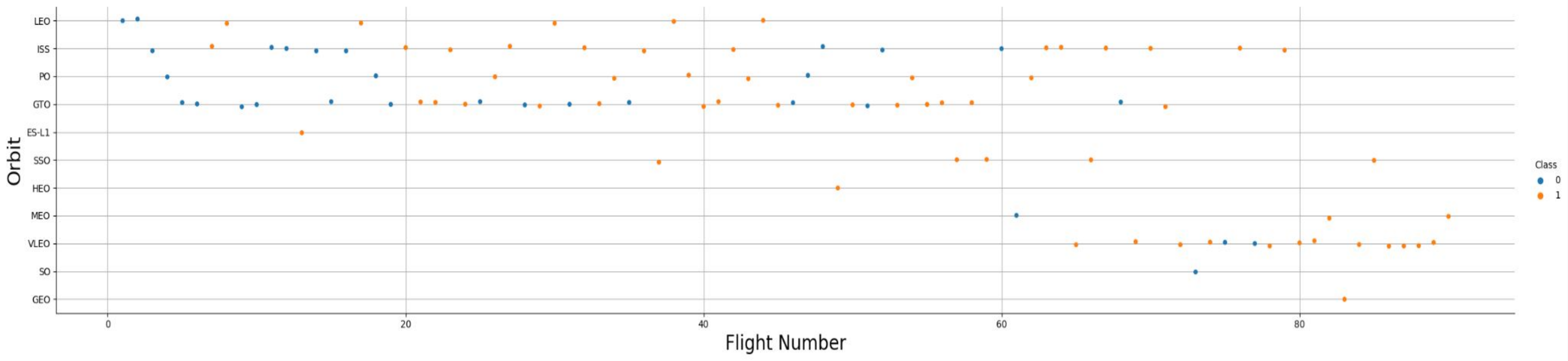
- A large proportion of successful launches have a payload of more than 9000 kg. There are no rockets launched for mass greater than 10000 kg for VAFB SLC 4E launchsite.

Success Rate vs. Orbit Type

- The highest proportion of successful launches was to ES-L1, GEO, HEO, and SSO orbits, and the lowest to GTO.

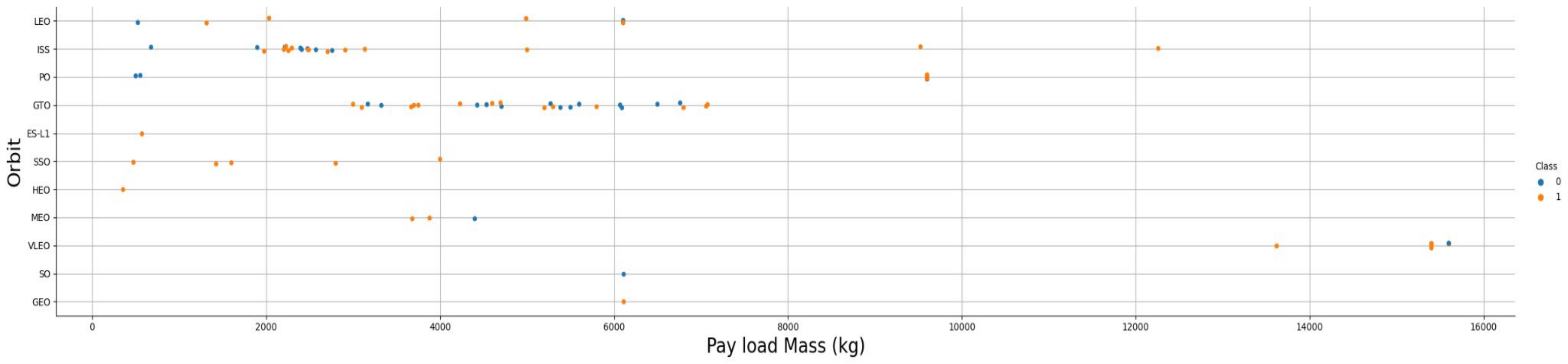


Flight Number vs. Orbit Type



- In the LEO orbit the success appears related to the number of flights; there seems to be no relationship between flight number when in GTO orbit. On 1-flight orbits (GEO, SO, HEO, ES-L1) the success rate cannot be accurately estimated.

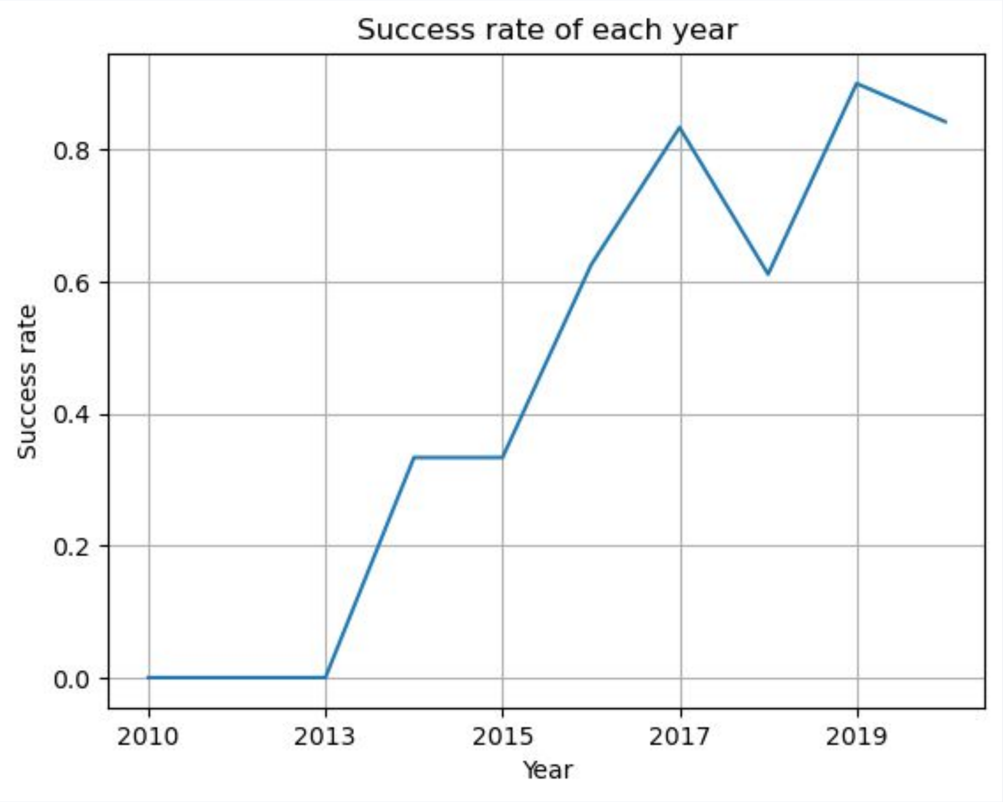
Payload vs. Orbit Type



- With heavy payloads the successful landing rate are more for PO,LEO and ISS. However for GTO we cannot distinguish this well.

Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names

- Find unique Launch Sites

With ‘DISTINCT’ we can return only unique values

```
%sql select distinct Launch_Site from SPACEXDATASET
```

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

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`

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

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

Date	Time(UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

With ‘LIMIT 5’ and ‘LIKE ’ we can return only 5 records, the ‘CCA%’ means that the Launch Site must begin with ‘CCA’.

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXDATASET where Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

The 'SUM' returns the total payload and the 'WHERE' clause filters the dataset by the appropriate condition.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXDATASET where Booster_Version like 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

The 'AVG' returns the average payload and the 'WHERE' clause filters the dataset by the appropriate condition.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql select MIN(Date) from SPACEXDATASET where Landing_Outcome = 'Success (ground pad)'
```

✓ 0.0s

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

MIN(Date)
2015-12-22

The ‘MIN’(date) returns the earliest date and the ‘WHERE’ clause filters the dataset by the appropriate condition.

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

```
%sql select Booster_Version from SPACEXDATASET where Landing_Outcome = 'Success (drone ship)' \
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

✓ 0.0s

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

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

'WHERE' clause filters the dataset by the appropriate conditions.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) from SPACEXDATASET group by Mission_Outcome
```

✓ 0.0s

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

Mission_Outcome	count(Mission_Outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Group the mission outcomes and count the records.

Boosters Carried Maximum Payload

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

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXDATASET \
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXDATASET);
```

✓ 0.0s

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0

‘MAX’ returns the maximum, ‘WHERE’ filters the dataset

2015 Launch Records

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

```
%sql select Booster_Version, Launch_Site, Landing_Outcome, Date from SPACEXDATASET \
where Date like '2015%' and Landing_Outcome = 'Failure (drone ship)';
```

✓ 0.0s

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

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

Two filters in 'WHERE' clause - year - 2015 and Landing Outcome - failure

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

‘Group BY’ - groups, ‘ORDER BY… DESC’ - orders in descending order, ‘WHERE’ clause filters the dataset by the appropriate condition.

```
%sql select Landing_Outcome, count(Landing_Outcome) \
as count from SPACEXDATASET \
where Date >= '2010-06-04' and Date <= '2017-03-20' \
group by Landing_Outcome order by count desc;
```

✓ 0.0s

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

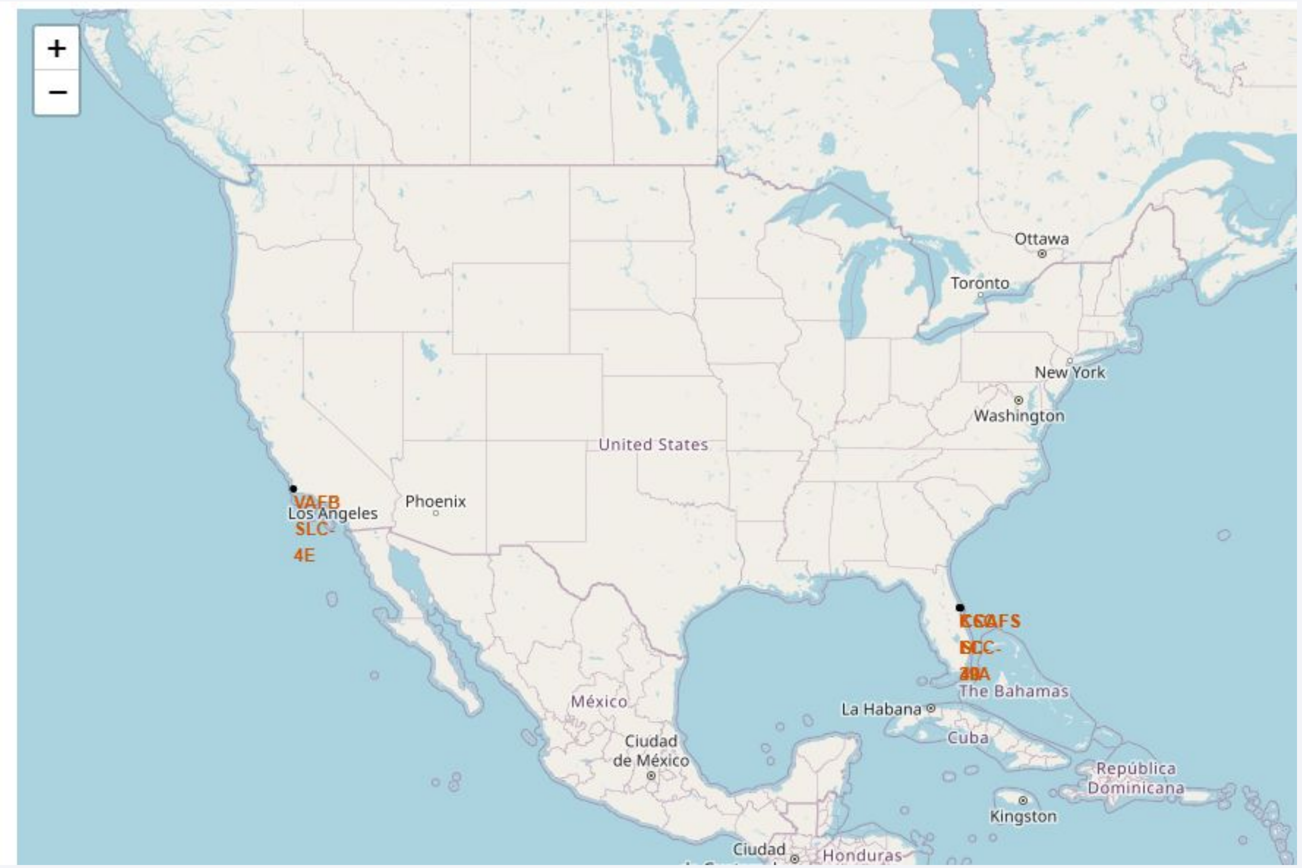
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

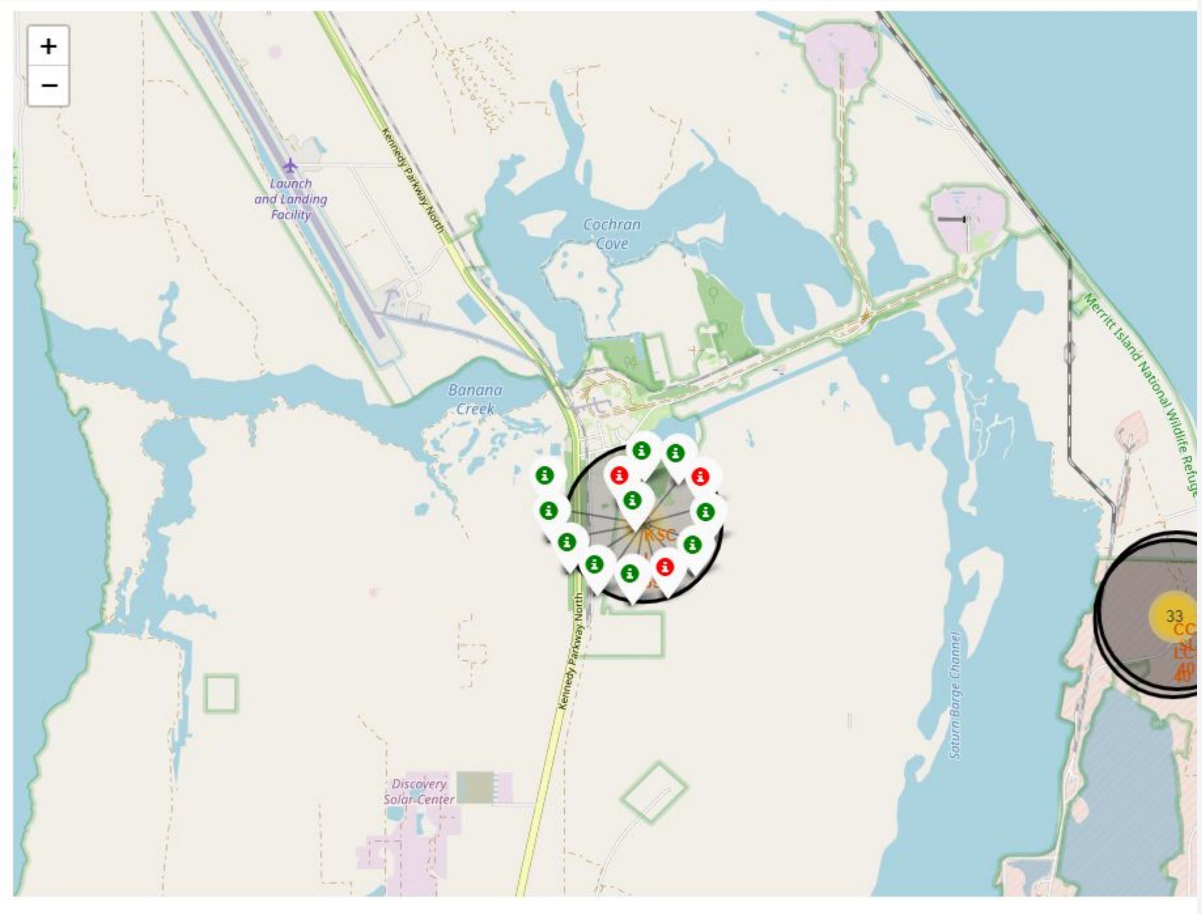
All Launch Sites

- All Space X launch pads are located on the U.S. coast



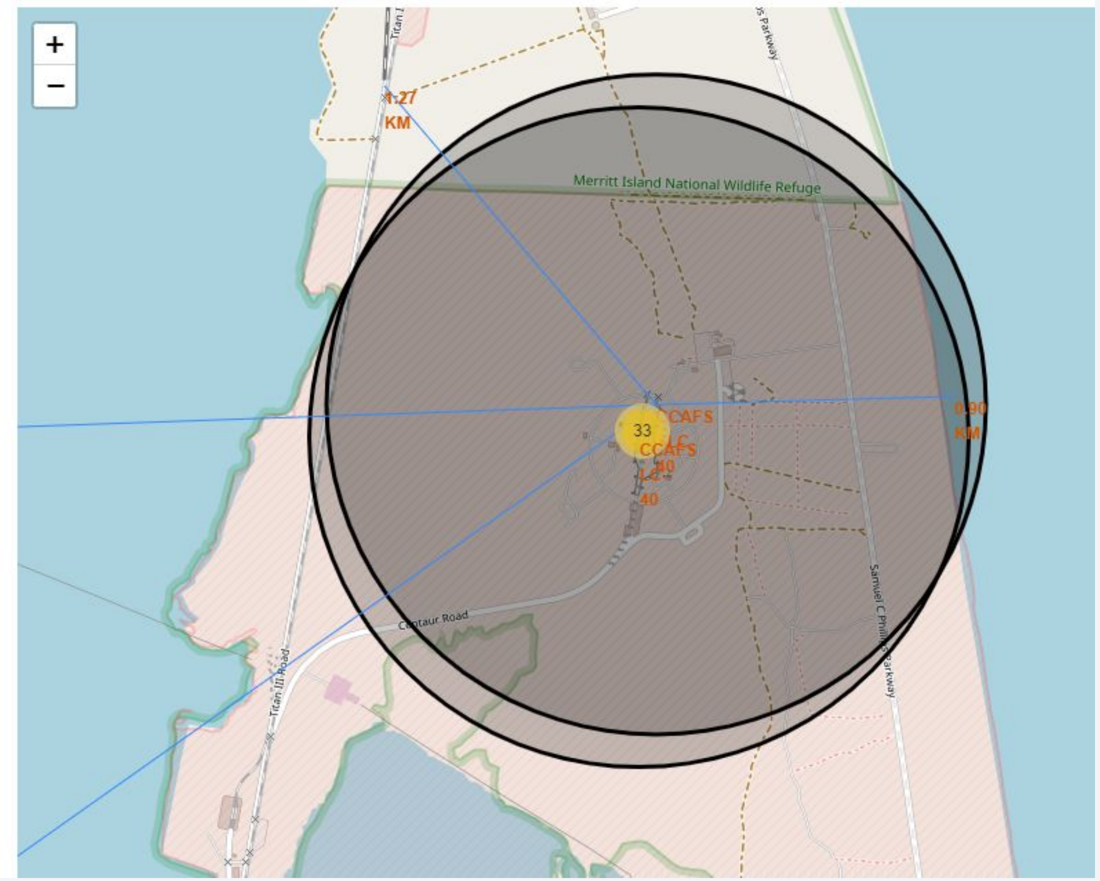
The color-labeled launch outcomes

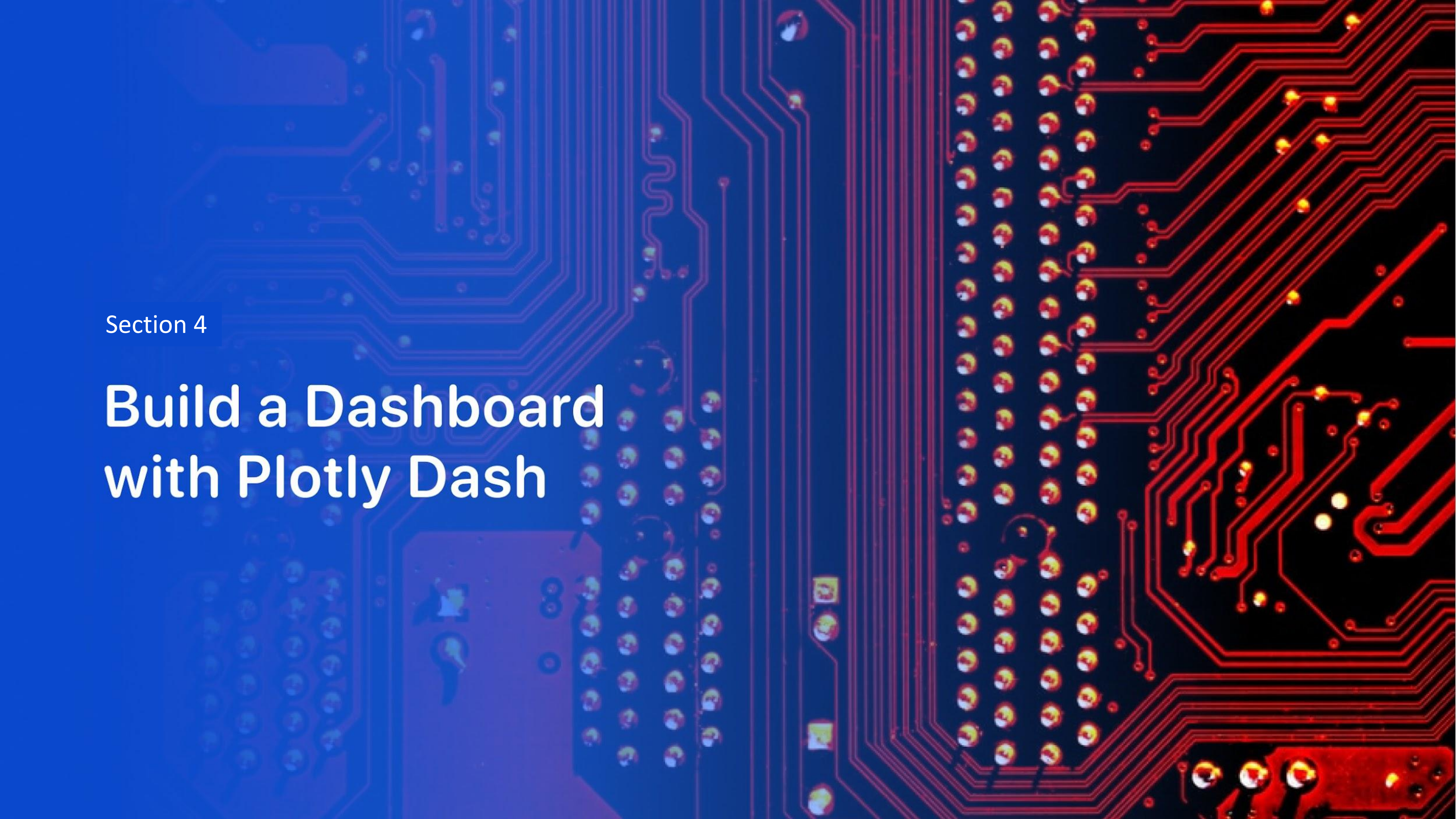
- RED labels means ‘failure’ and GREEN labels means ‘success’



Launch site proximities

- This launch sites are located in close proximity to the railroad (1.27 km), so they have good logistics



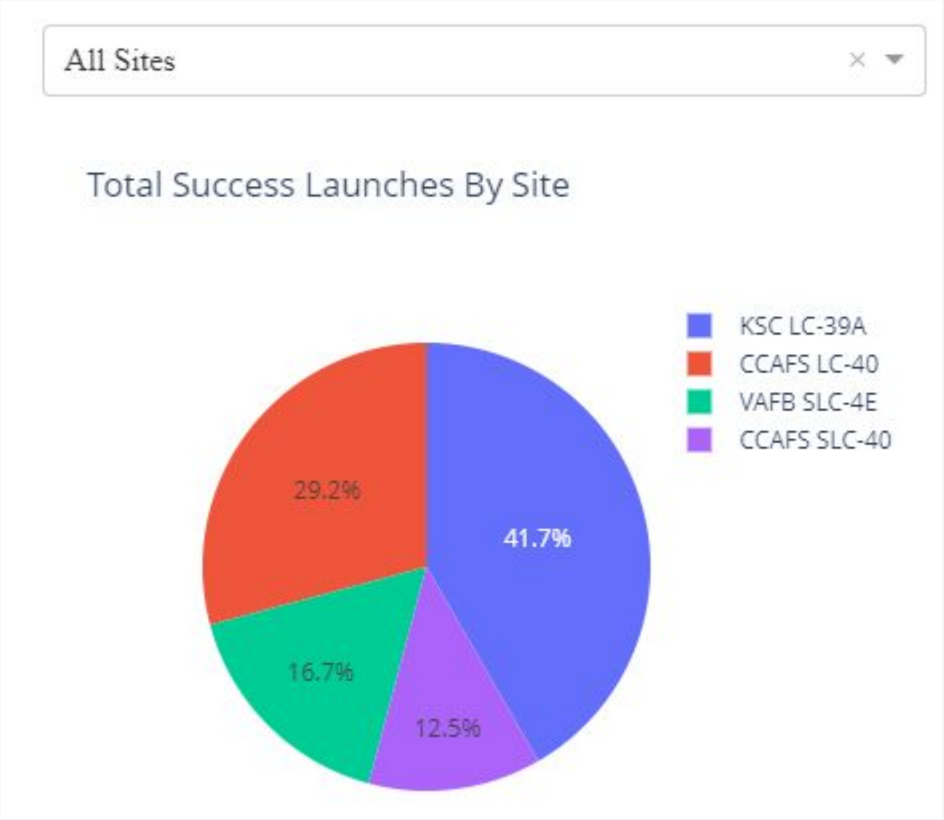


Section 4

Build a Dashboard with Plotly Dash

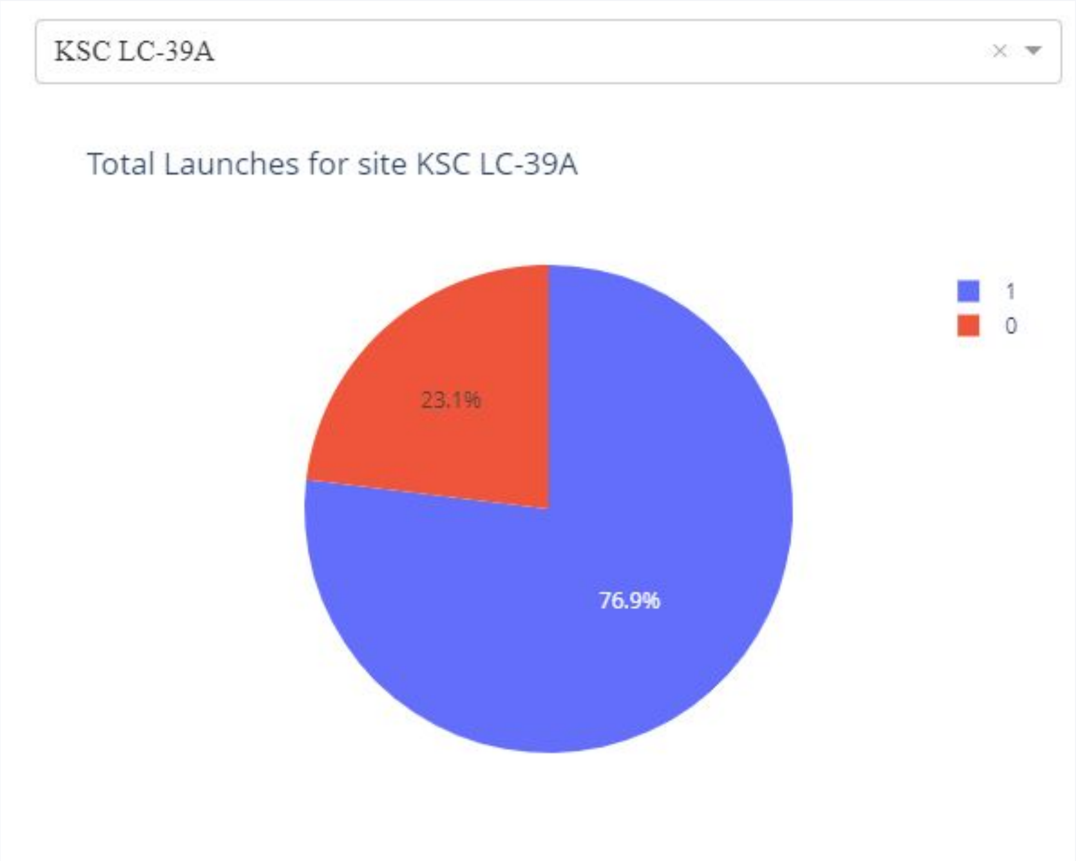
Launch success rate for all sites

- The KSC LC-39 A has the highest success rate and the CCAFS SLC-40 has the lowest.



The highest ratio success launch site

- KSC has a 76.9% success rate and a 23.1% failure rate.



Payload vs. Launch Outcome

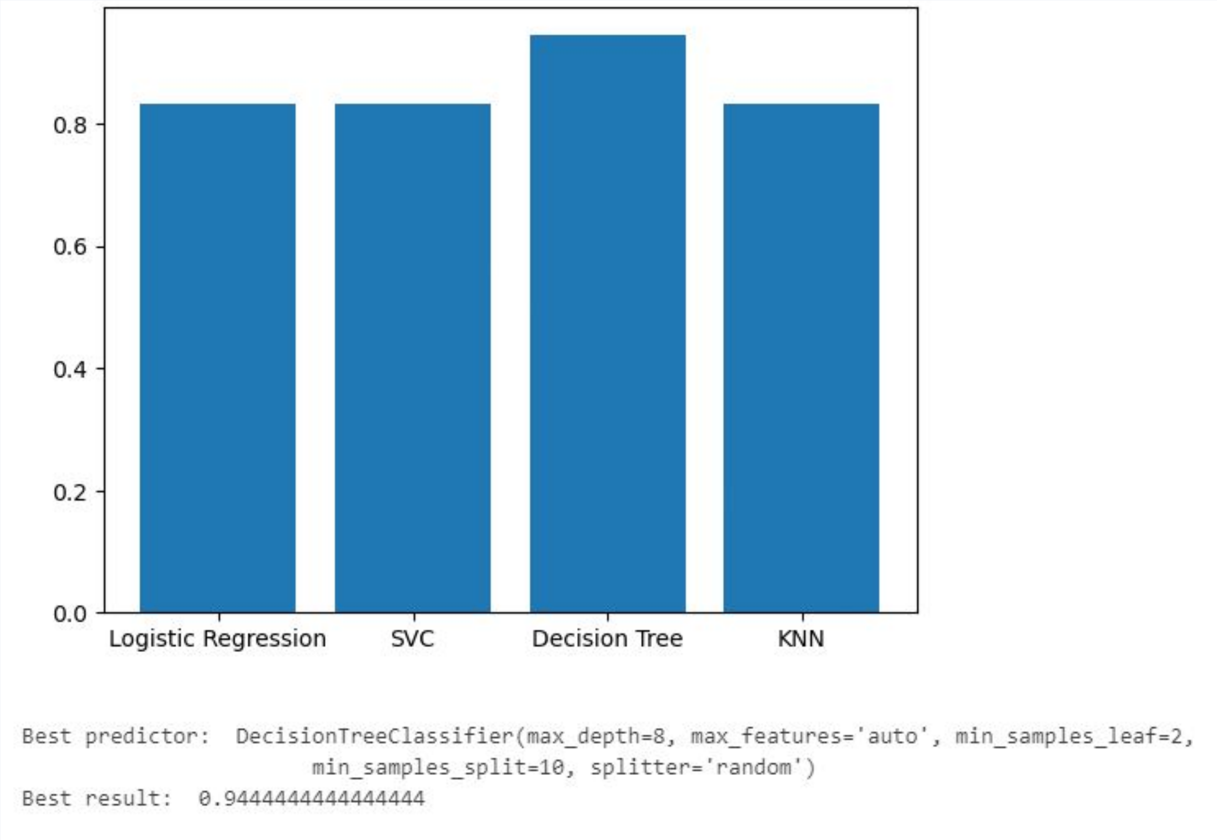
Success rates for lower weight payloads (< 5000 kg) are higher than for higher weight payloads (> 5000 kg).

Section 5

Predictive Analysis (Classification)

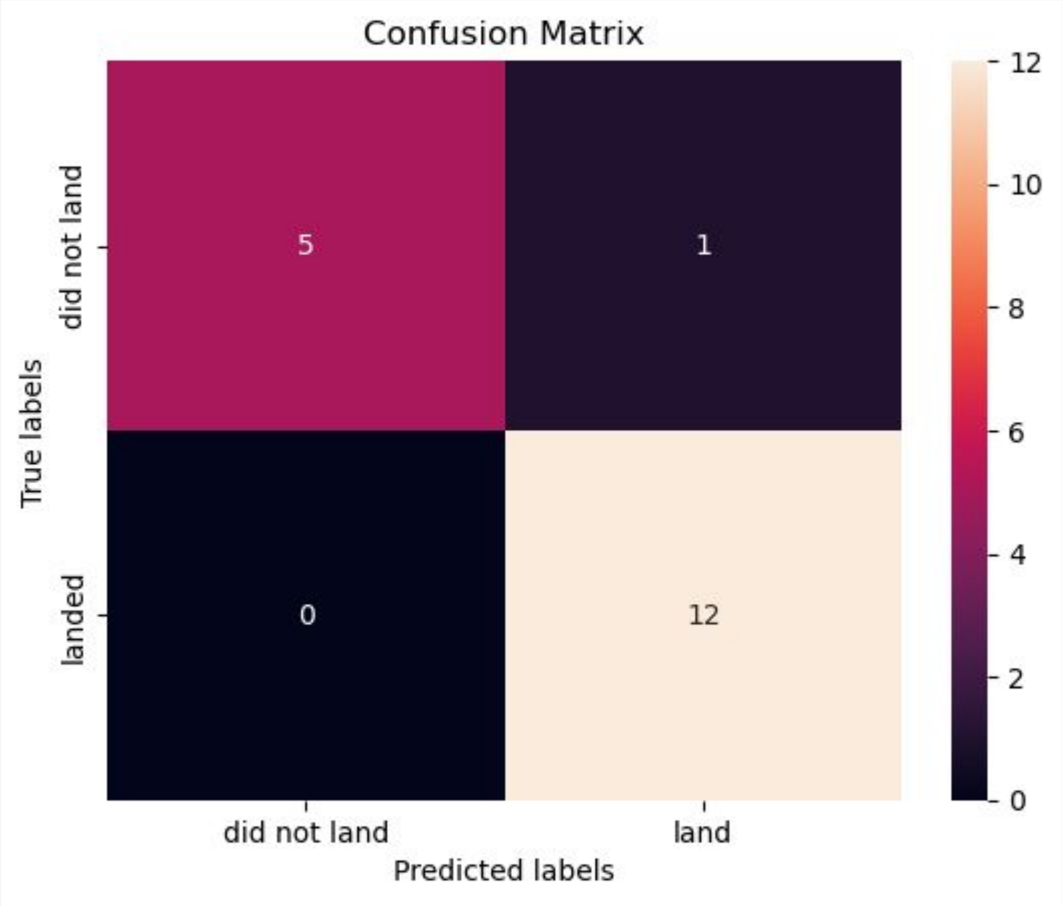
Classification Accuracy

- Decision tree has the best accuracy with above 94.4%



Confusion Matrix

- The decision tree classifier gave a wrong prediction for only one case - when the rocket launch was not successful, and it predicted its success. In the other 12+5=17 cases, his predictions were correct.



Conclusions

- SPACE Y companies might be advised to start by launching not very large loads, as they are the ones with the highest success rate based on the SPACE X story.
- It should launch from the KSC-LC 39A launch site, which has had more successful launches than the others.
- It should choose orbits with higher launch success rates: ORBIT GEO, HEO, SSO, ES-L1
- The longer SPACE Y takes to make launches, the higher its success rate will be.
- Based on the history of SPACE X, our best model, the decision tree classifier, can predict the outcome of launches with about 94.4% accuracy.

Appendix

Thank you!