IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emma Hansson
March 4, 2024

Outline

Executive Summary

- Summary of methodologies
  - Data collection via Rest API and web scraping followed by data wrangling
  - EDA using visualization and SQL queries
  - Interactive analytics using folium and Plotly DASH
  - Machine learning models using training and test data. Four models were tested.
- Summary of results:
  - Launch sites should be near the coast and not next to a major population center
  - Success increases with the number of flights
  - The top landing site, KSC LC-39A, has a success rate of around 77%
  - Booster version FT and B4 should be preferred
  - Type of orbit should be carefully considered, some types have a 100% success rate while others are as low as 50%
  - The machine learning models are fairly accurate at predicting failed or successful mission with the highest accuracy being 87.5% (decision tree model)

Introduction

- Project background and context
  - SpaceY is a new space company funded by billionaire Allon Musk
  - The goal is to be able to compete with SpaceX
  - SpaceX will be studied to learn more about the key to success and to determine if SpaceX will reuse the first stage

- What will be looked at:
  - If the mission and launch are successful, what do these successful launches have in common?
  - Several factors will be looked at: payload, orbit type, number of flights, launch site and booster type.
  - A machine learning model will be trained to try and predict the outcome

Section 1

Methodology

Methodology

Executive Summary

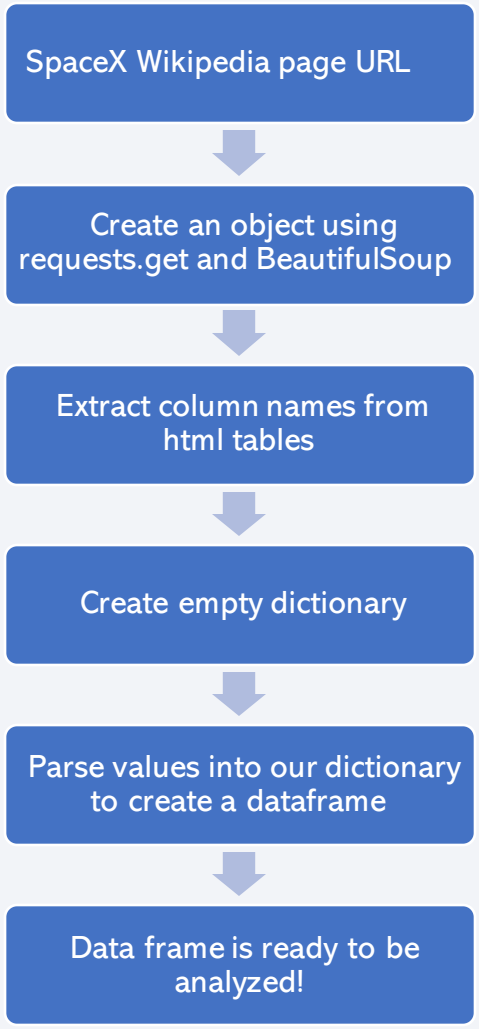
- Data collection methodology:
  - SpaceX API and web scraping were used to find the relevant data
- Perform data wrangling:
  - Null values were removed and 2 landing classes were created: 1 for success and 0 for failure
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Created scatter plots to see the relationship between different parameters and how they affect a successful outcome
  - Used SQL queries to find important information such as total and average payload and dates of successful landings
- Perform interactive visual analytics using Folium and Plotly Dash
  - Mapped the launch sites to see if there was an obvious relationship to where they are located and the success of the mission
  - Created an interactive dashboard to look at the successful launches for each site
- Perform predictive analysis using classification models
  - Looked at 4 different machine learning models and how effective they were at predicting the outcome of the mission

Data Collection

Data Collection – SpaceX API

- This API will give us data about SpaceX launches: information about the rocket used, payload, launch specifications, landing specifications, and landing outcome.
- Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.
- GitHub URL https://github.com/esvhansson/spaceX_API_lab

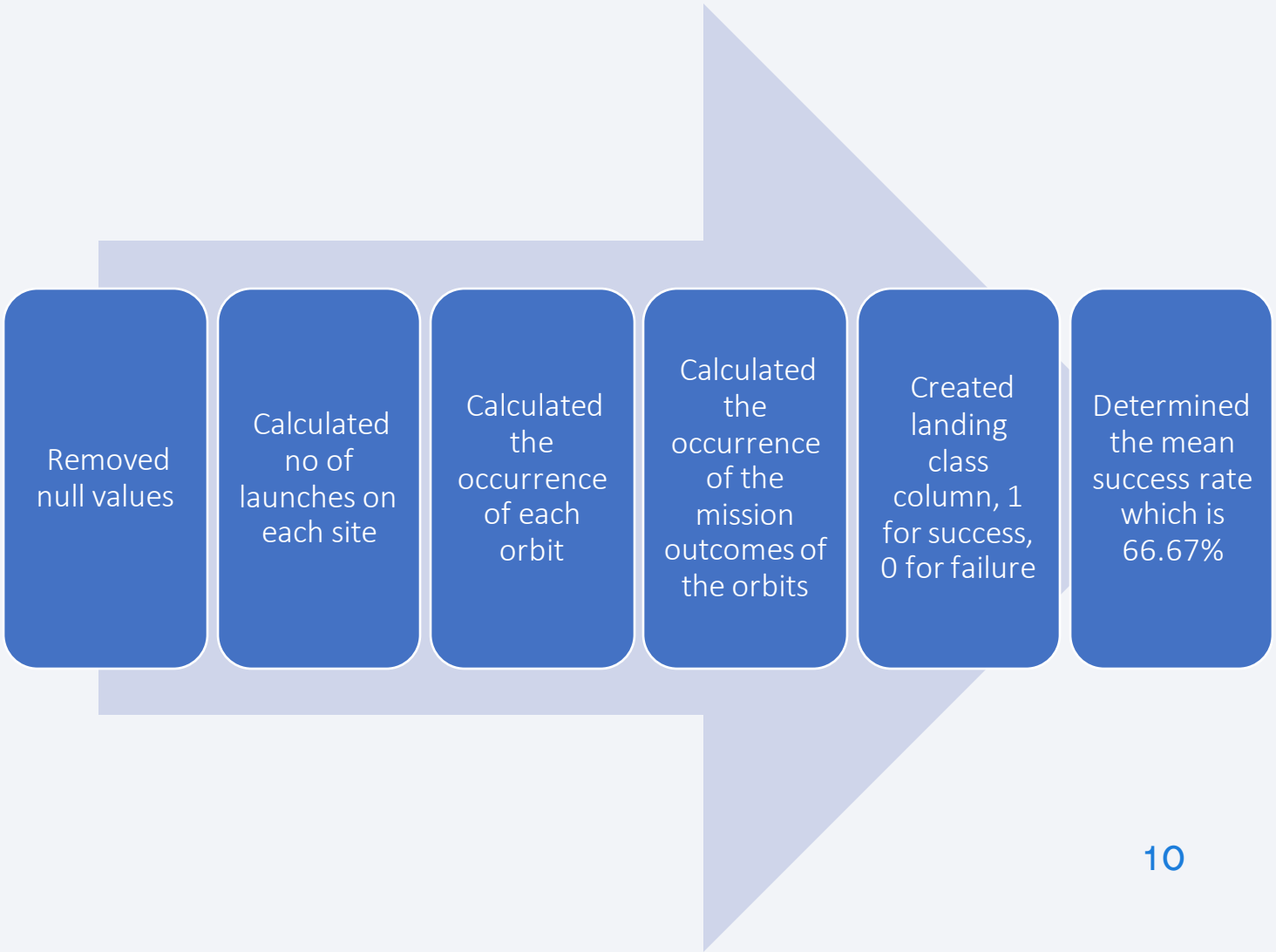
Data Collection - Scraping



GitHub URL: https://github.com/esvhansson/webscraping1/tree/main

Data Wrangling

- Flow chart for the data wrangling process can be seen on the right
- Missing values: calculated the mean for the payload using .mean(). Then used the .replace() function to replace missing values in the data with the mean calculated.



- GitHub URL: https://github.com/esvhansson/Data_wrangling_spaceX/tree/main

EDA with Data Visualization

- Exploratory data analysis was performed using pandas and matplotlib
- Scatterplots were created in order to check relationships between the successful landings and other variables such as payload and flight number.
  - See graphs on page 18-23
- A split for each site was made to check if different variables affected different sites
  - Success at each site vs flight number: a clear trend that later flight numbers has a higher success rate at each site
  - Payload mass: the higher the payload, the less likely it is that the first stage will return.
  - Success rate of each orbit : some orbits have a much higher success rate than others, namely ES-L1, GEO, HEO and SSO. See bar chart on page 20
- For the yearly trend, a line chart was created (see page 23).
  - The success rate increased every year until 2019-2020
- GitHub URL: https://github.com/esvhansson/Visulaization-SpaceX/tree/main

EDA with SQL

- Using SQL and the SpaceX table, the following queries were performed:
  - Names of unique launch sites
  - Launch site starting with "CCA"
  - Total payload mass carried by NASA booster
  - Average payload mass carried by booster F9 v1.1
  - Date of first successful ground pad landing
  - Boosters successful in drone ships with a payload between 4000 and 6000
  - Total number of successful and failed missions
  - For 2015, list of failed landings on drone ships
  - Ranking landing outcomes in a date span
- GitHub URL: https://github.com/esvhansson/EDA_SQL_spaceX

Build an Interactive Map with Folium

- All launch sites were plotted on a map.
  - Each yellow circle shows how many launches have been done at each site
  - Zooming in further we can click on each launch site and see the number of launches and each of those marked in green for success and red for fail
  - Distances to railway lines, major highways and the coast are also added
- The interactive map makes it easier understand and visualize where the successful launches have been and what those locations have in common
- GitHub URL: https://github.com/esvhansson/folium-spaceX/tree/main

Build a Dashboard with Plotly Dash

- Interactive dashboard shows the successful launches per site in percentage terms
- It plots the successful and failed missions and its correlation to payload mass and booster version
  - See graphs on page 39-41
- Why?
  - There are large variations in terms of sites and booster versions and how they impact a successful mission.
  - Helps us draw conclusions if weight has an impact on successful missions.
- GitHub URL https://github.com/esvhansson/DASH_app_code

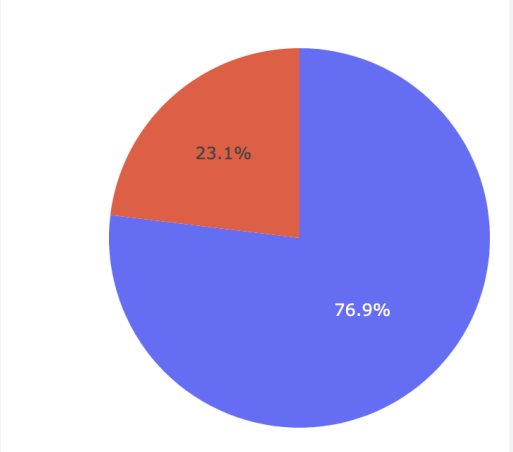
Predictive Analysis (Classification)

- Process followed:
    1. Imported necessary libraries and loaded the data frame
    2. Created numpy arrays
    3. Split the data into training and testing sets
    4. Performed logistic regression analysis
    5. Performed SVM analysis
    6. Performed a decision tree analysis
    7. Performed KNN analysis
    8. Used a confusion matrix and score to check the accuracy of the data

- GitHub URL: https://github.com/esvhansson/MachineLearning_SpaceX

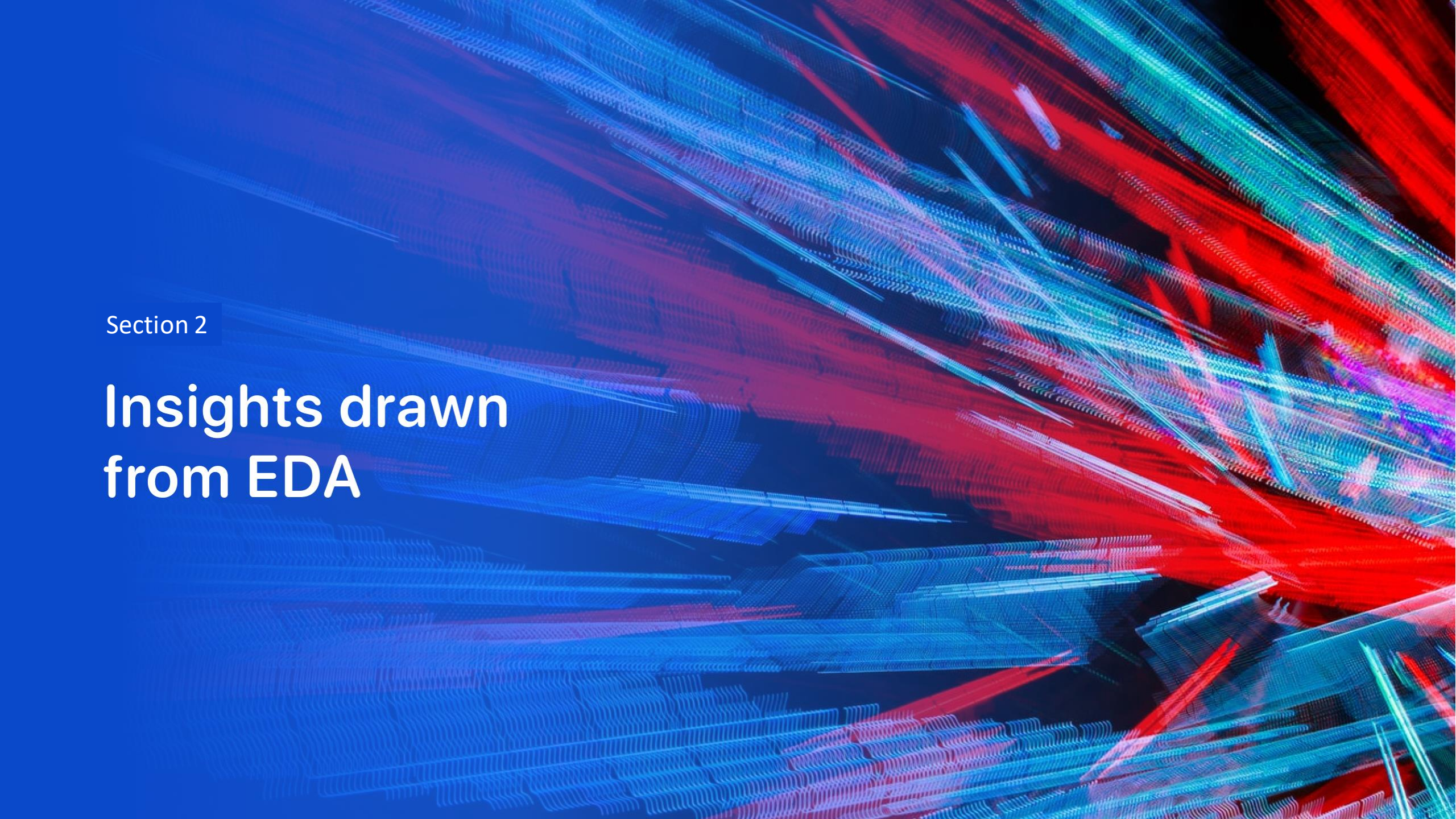
Results

- Exploratory data analysis results:
  - Correlation between multiple factors were explored and plotted
  - The strongest correlation was between launch site and number of flights, showing a clear increase in success with the number of flights. This ties in well with the yearly success trend which has gone up every year until 2019-2020
  - The SQL query of fist successful ground landing in 2015 supports this finding
- Interactive analytics:
  - The dashboard shows us the success of each site, the best being KSC LC-39A (see chart)
  - The Booster analysis shows a clear picture that FT and B4 are the best choice but failure increases across the board of the boosters with a higher payload
- Predictive analysis results:
  - From the machine learning models we tried, all 4 had an accuracy of over 80%
  - The decision tree proved to be the best model with an accuracy of 87.5%
  - The maps plotted with Plotly shows us the best launch sites are near the ocean so rockets can be launched out over open sea



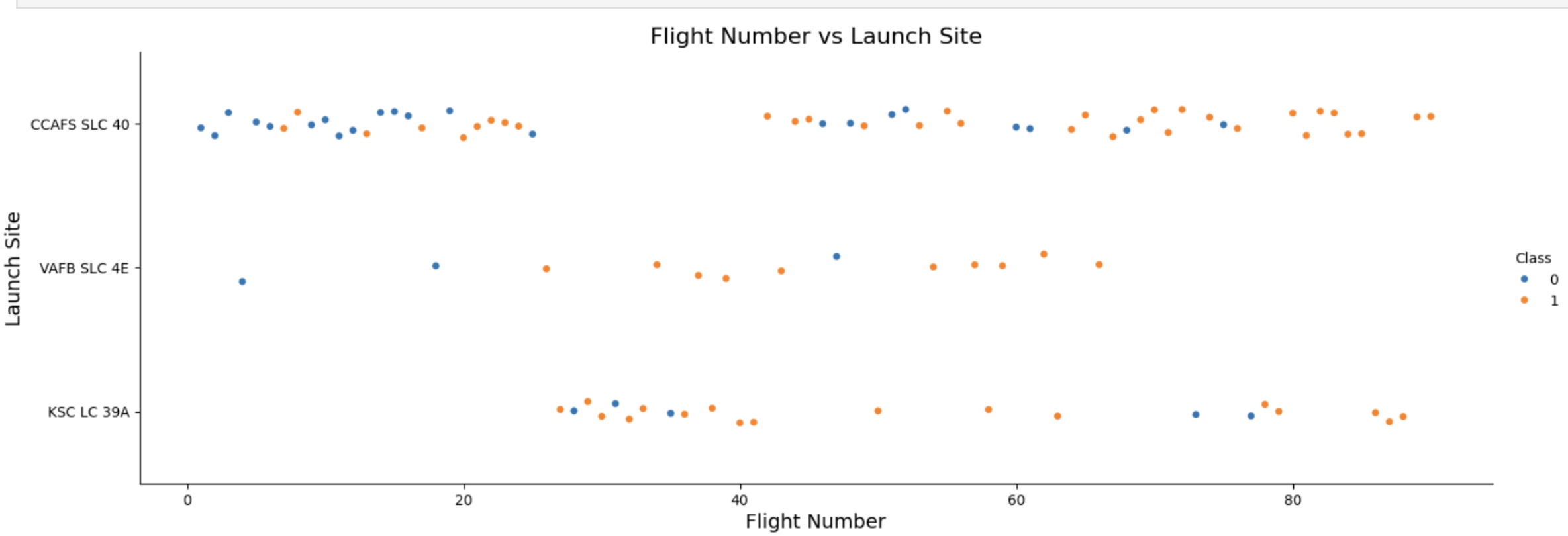
Section 2

Insights drawn from EDA



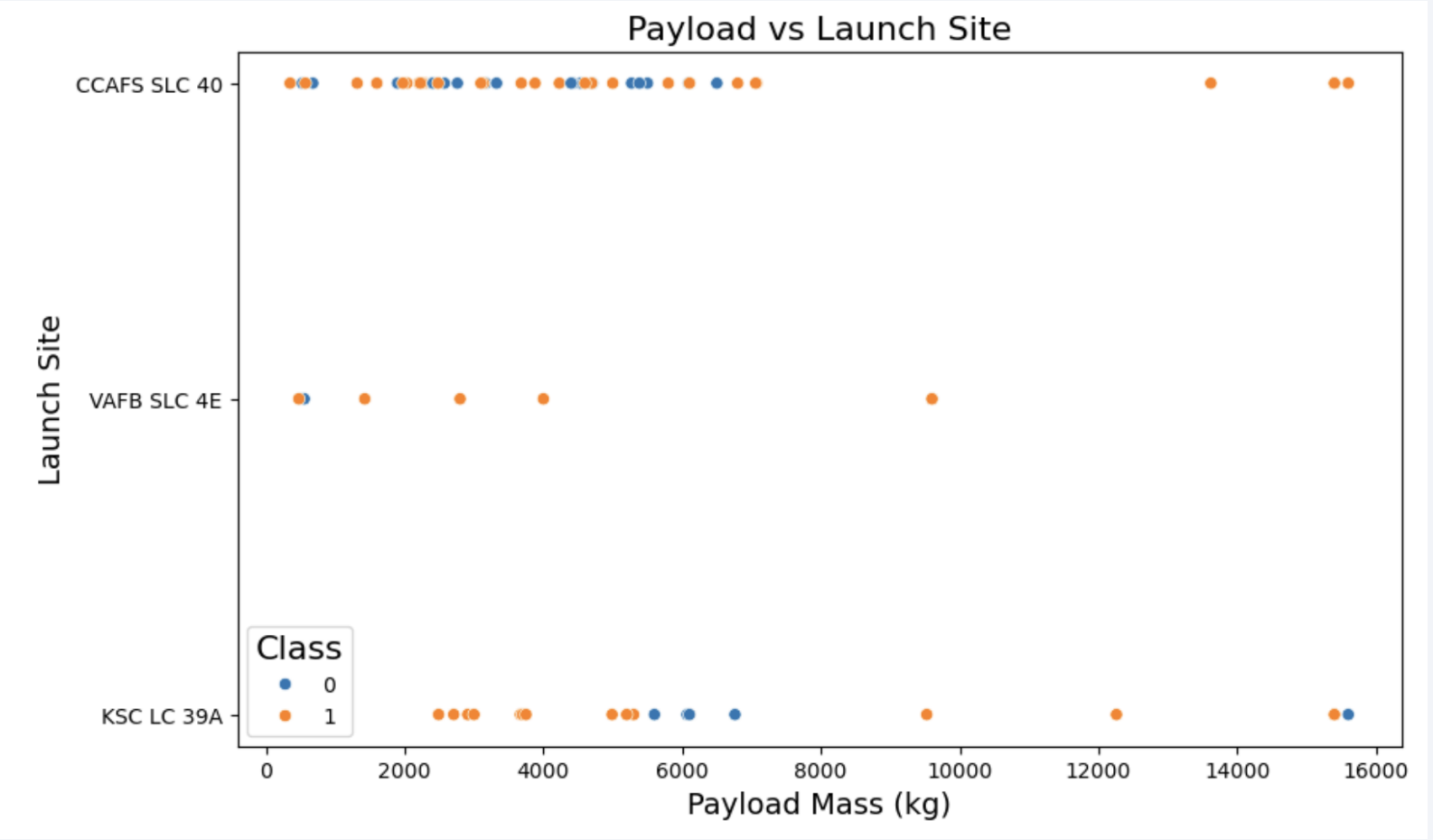
Flight Number vs. Launch Site

- At all launch sites the number of successful launches increase with the number of flights



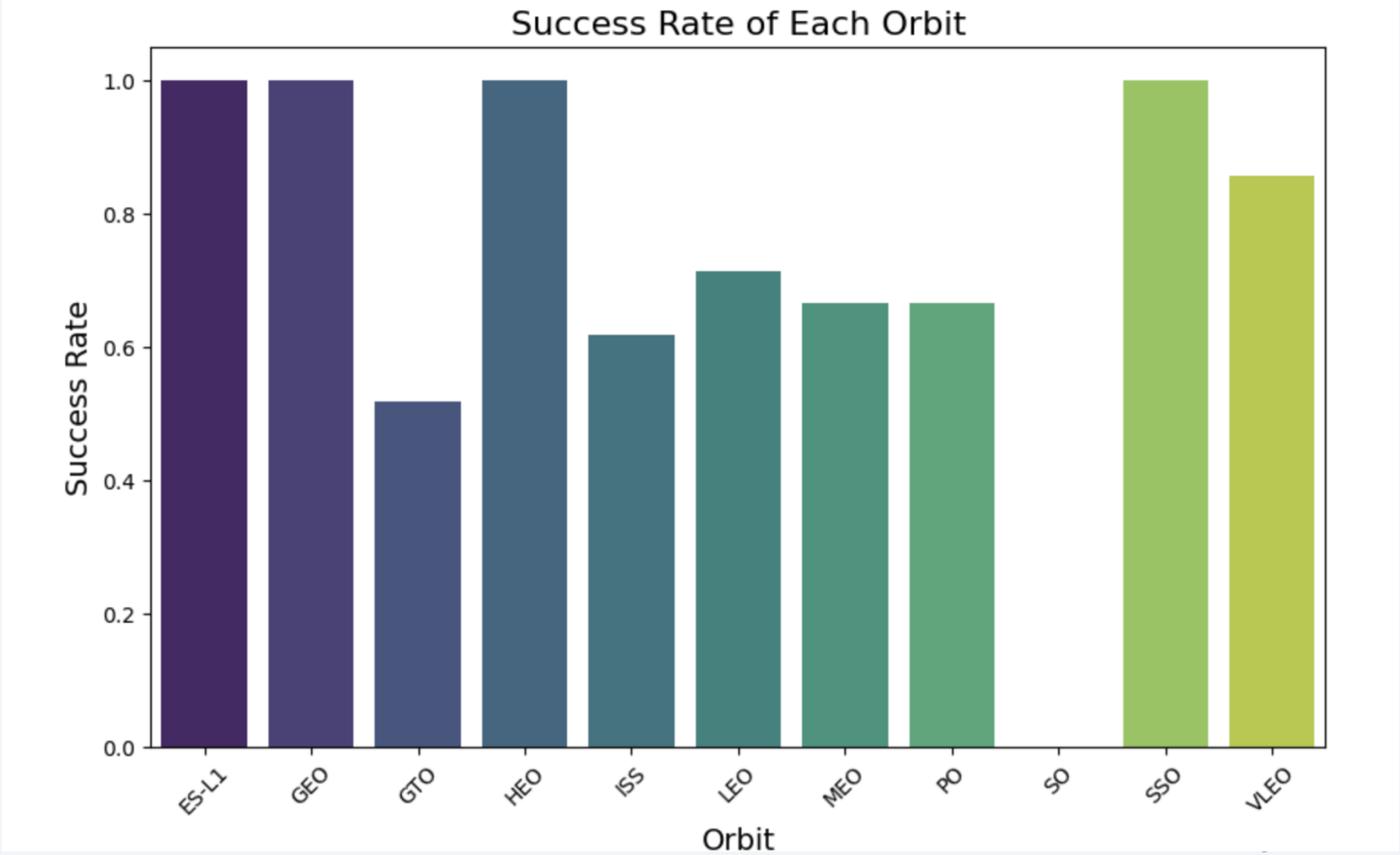
Payload vs. Launch Site

- There does not seem to be much of a correlation for payload vs launch site
- There are mixed results for all payloads on most sites



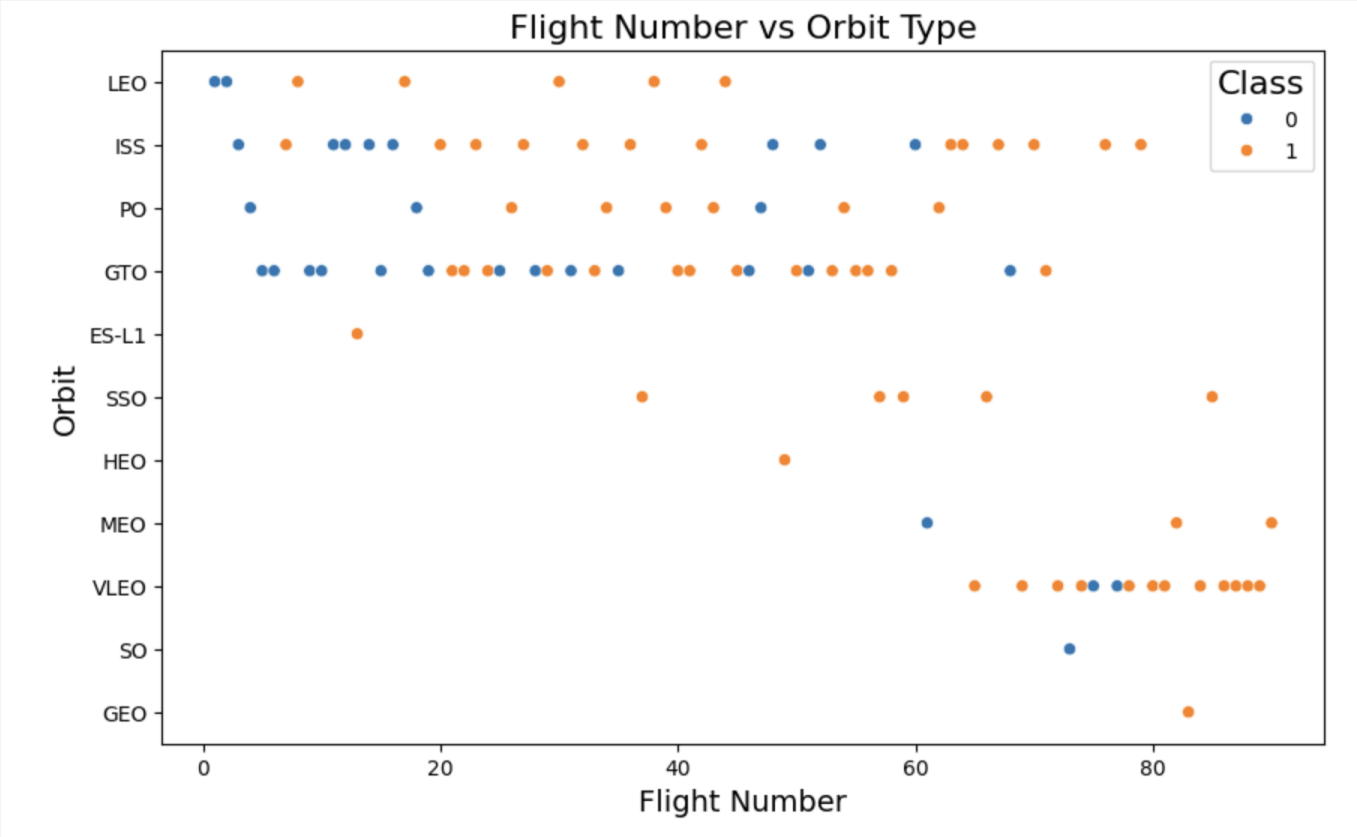
Success Rate vs. Orbit Type

- The bar chart shows clearly that some orbits have a higher success rate than others
- ES-L1, GEO, HEO and SSO have a 100% success rate while GTO is as low as 50%



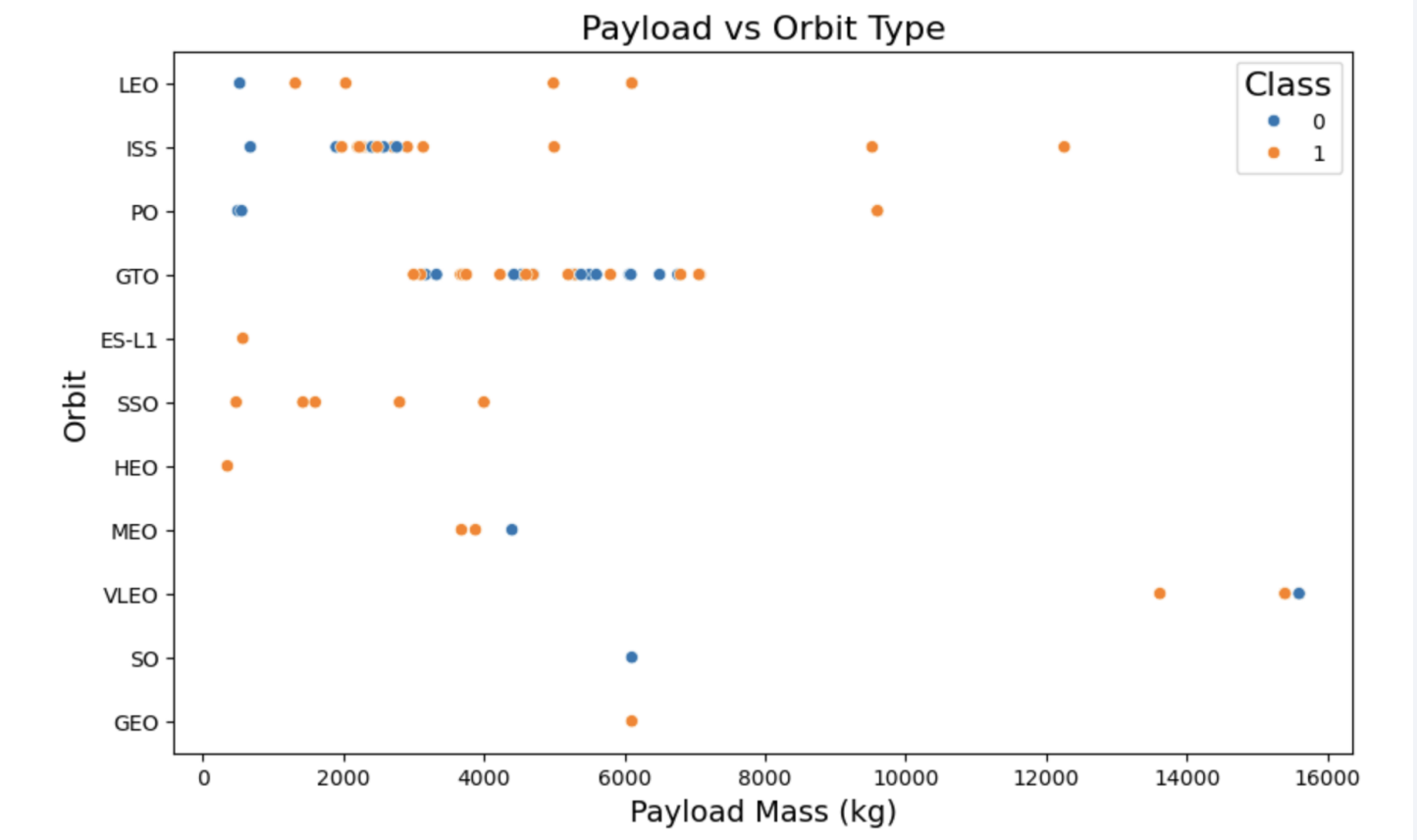
Flight Number vs. Orbit Type

- Most orbit types show a trend of increased success with the number of flights.
- GTO is an exception where there is no clear pattern
- SO, GEO, MEO and ES-L1 have too few flights to determine a trend



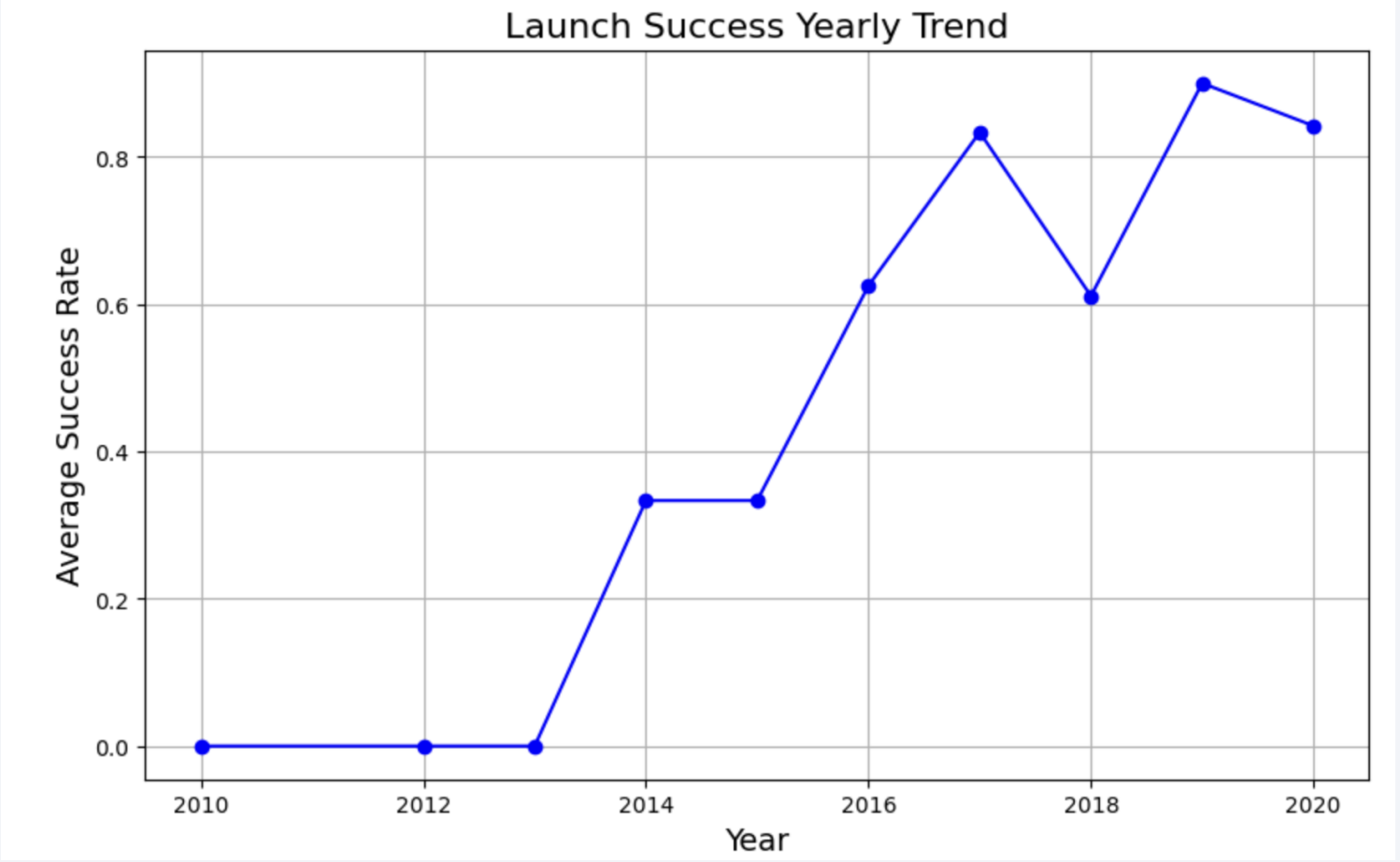
Payload vs. Orbit Type

- There is no clear correlation for most orbit types, for example GTO and ISS
- This does not seem to be a significant factor



Launch Success Yearly Trend

- The line chart shows a clear trend of increased success every year until 2019-2020
- The success rate increased the most percentage wise from 2015-2017



All Launch Site Names

- Used SELECT DISTINCT to find the unique names of the launch sites.
- There are four in total, listed below

```
%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'

- To find 5 records where launch sites begin with `CCA`, WHERE LIKE was used
- LIMIT 5 was used to limit the results to only 5

```
%sql SELECT * FROM SPACEXTBL 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

Total Payload Mass

- To calculate the total payload carried by boosters from NASA the command WHERE LIKE was used in combination with SUM

```
%sql PRAGMA table_info(SPACEXTBL);

%sql SELECT SUM("PAYLOAD_MASS__KG_") AS "Total Payload Mass (kg)" FROM SPACEXTBL WHERE "Customer" LIKE '%NASA (CRS)%';
```

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

Total Payload Mass (kg)
48213

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1 we calculted the AVG of the payload mass column and selected WHERE the booster version equals F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload Mass (kg)" FROM SPACEXTBL WHERE "Booster_Version" = 'F9 v1.1';
```

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

Average Payload Mass (kg)
2928.4

First Successful Ground Landing Date

- To find the date of the first successful landing outcome on a ground pad we used the MIN(Date) function in combination with WHERE landing outcome was a success
    - The MIN function gives us the earliest possible date where the landing outcome was a success

```
%sql SELECT MIN(Date) AS "Date of First Successful Landing on Ground Pad" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

Date of First Successful Landing on Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, the query below was used:
  - Selected only the landings that were a success on a drone ship
  - Limited the payload mass to only range between 4000 and 6000kg

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

```
 * 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

- To calculate the total number of successful and failure mission outcomes we used the COUNT function
  - Then GROUP BY the mission outcome to get a better overview in a table

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) AS "Total" FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

- To see the names of the booster which have carried the maximum payload mass we used the MAX function of the payload mass column

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

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

Booster_Version
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

- To list the failed landing outcomes on a drone ship, their booster versions, and launch site names for in year 2015 we used the query to the right
- Dates were converted as SQLite does not understand date names
- The results is that only 2 failed landings on drone ships occurred in 2015

```
%%sql
SELECT
    CASE substr("Date", 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS "Month",
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM
    SPACEXTBL
WHERE
    substr("Date", 1, 4) = '2015'
    AND "Landing_Outcome" LIKE '%Failure (drone ship)%';
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- To rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order the query to the right was used
- The COUNT function counts the number of occurrences
- The GROUP BY function groups the results into the different landing outcome categories

```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS "Count"
FROM SPACEXTBL
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY COUNT(*) DESC;
```

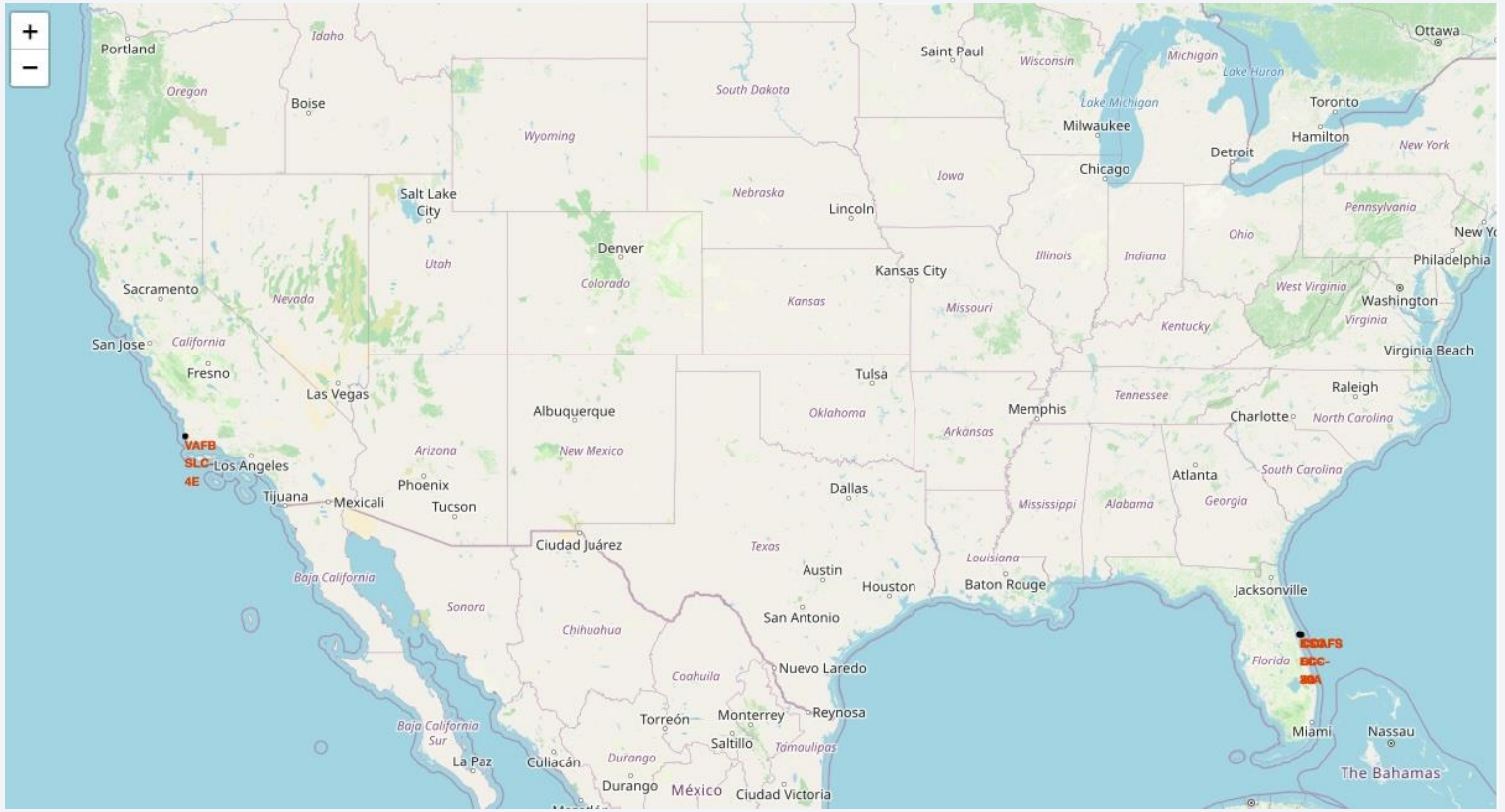
```
 * sqlite:///my_data1.db
Done.
```

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

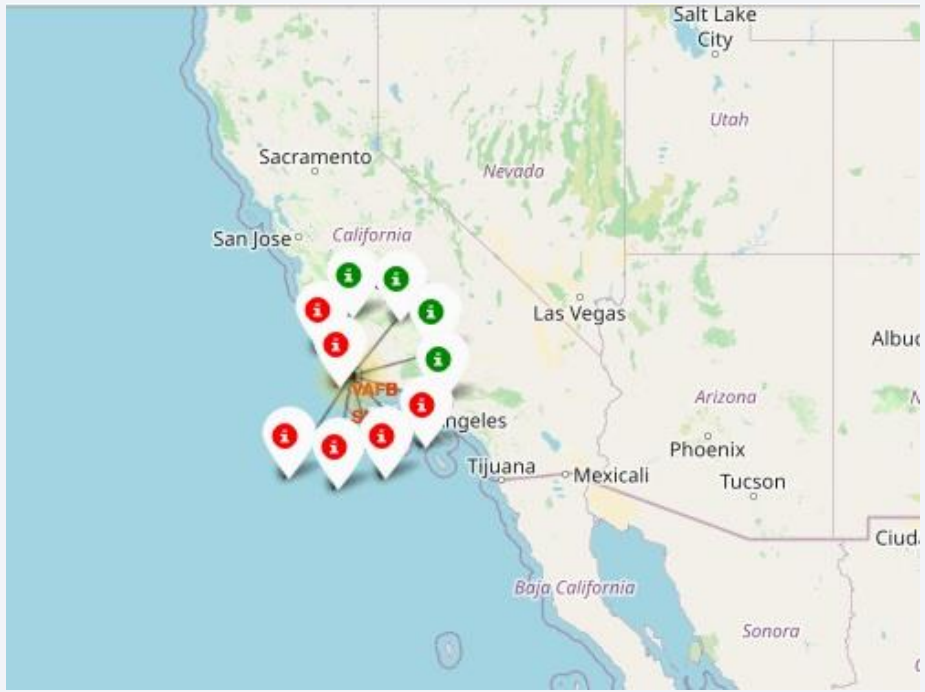
SpaceX launch sites

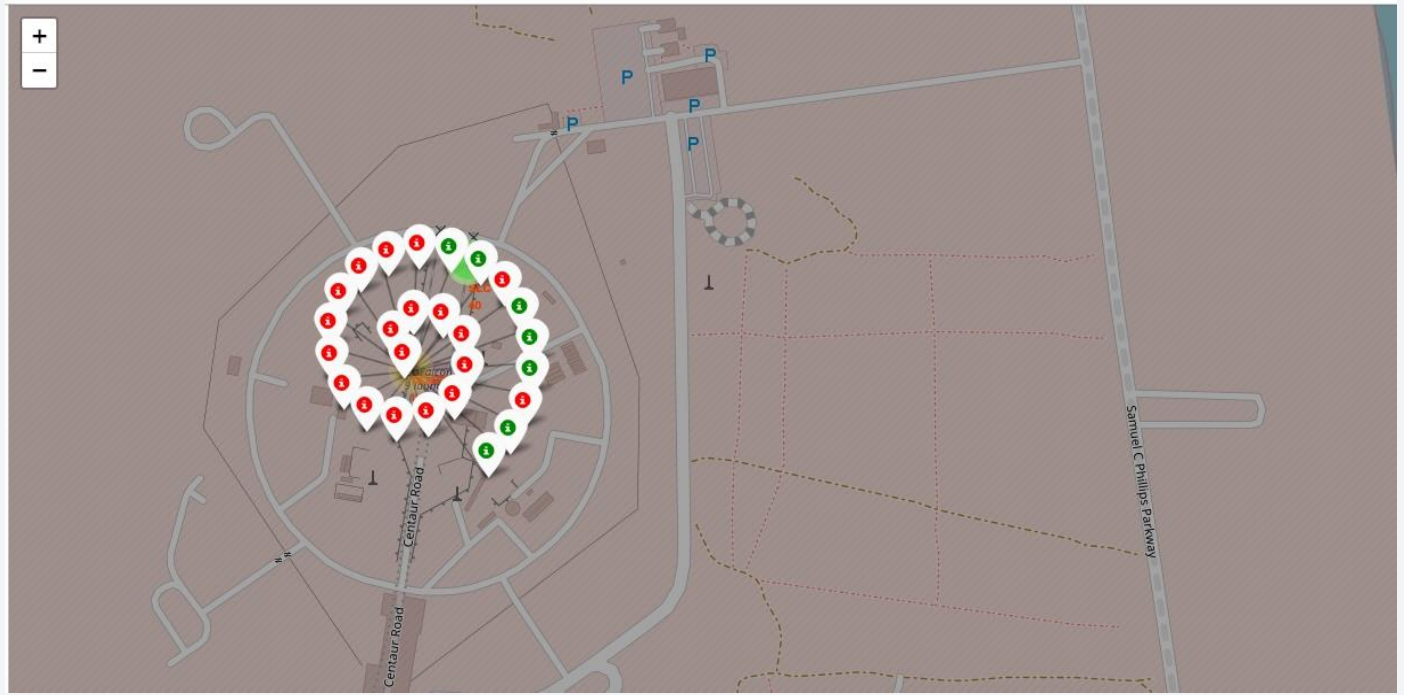
- The maps shows the global launch sites of SpaceX
- Launch sites are all located close to the coast in areas with favorable weather conditions (California and Florida)
    - Coastal areas are chosen as rockets can be launched out over open water, giving safety to those on the ground



Map of successful and failed launches

- Each map is labeled with green for successful launches and red for failed launches at all sites.
  - Zooming in we can deduce that some sites, such as Cape Canaveral in Florida, have more successful outcomes compared to others.
  - The left map shows VAFB SLC-4E and the right map shows CCAFS LC-40

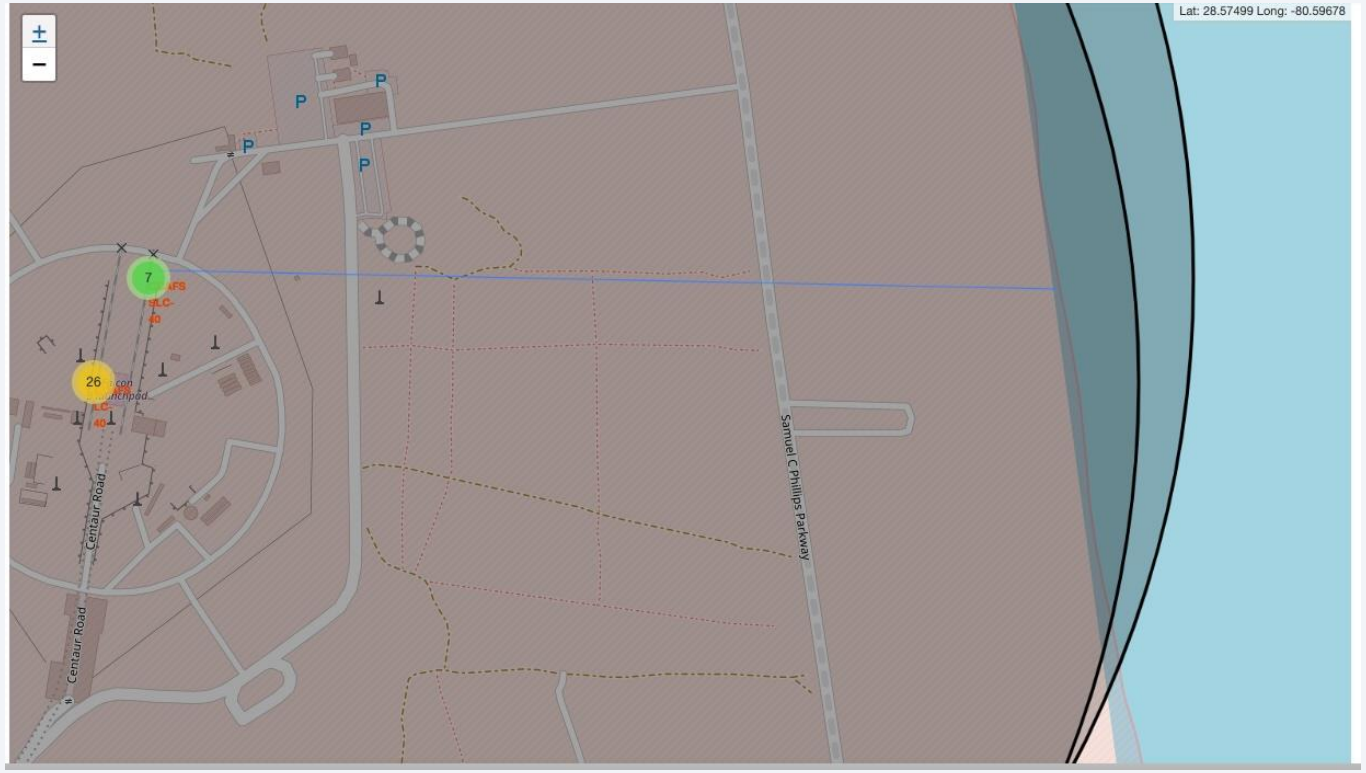


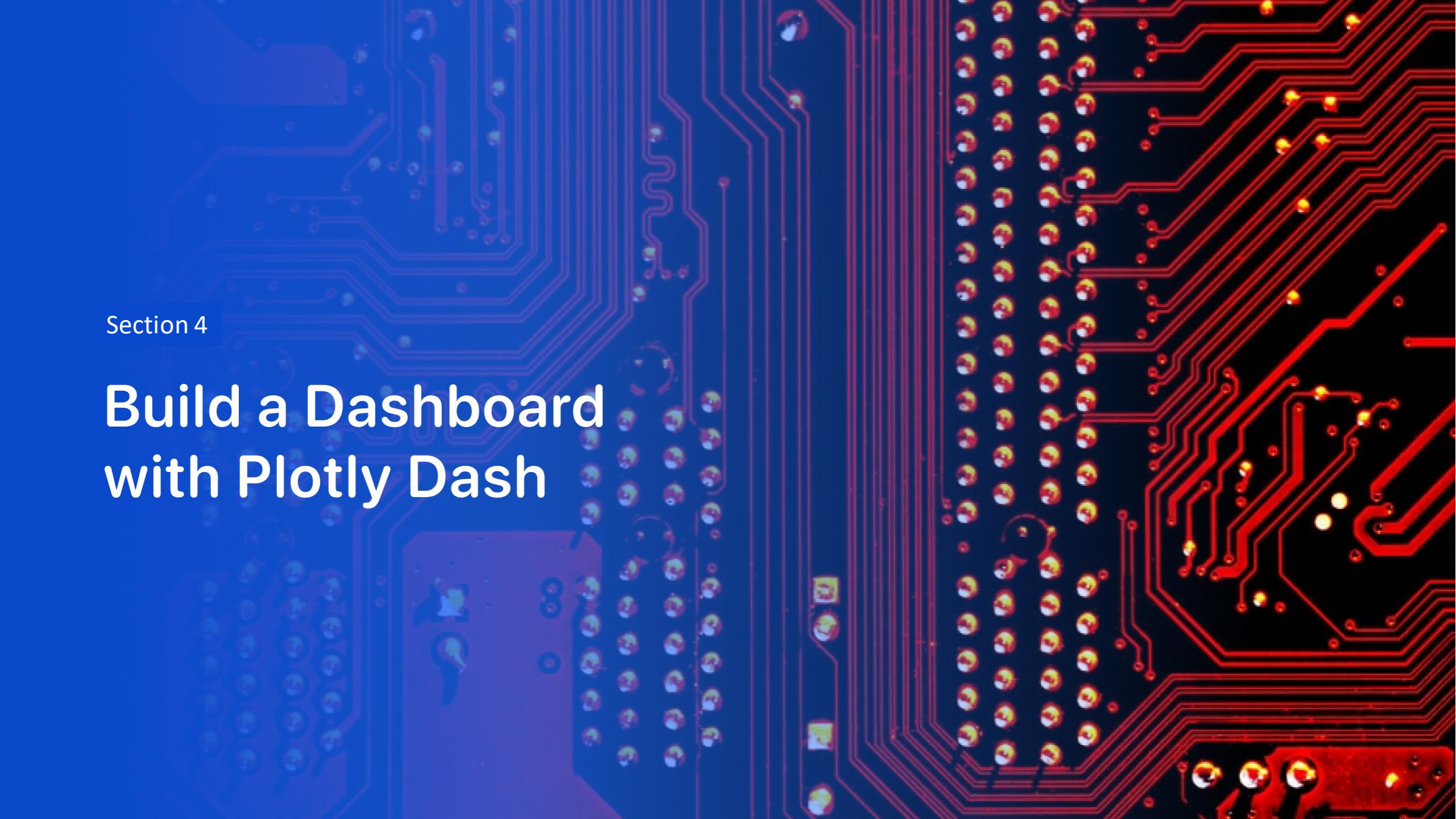


Coast proximity for launch sites

- Zooming in on Cape Canaveral in Florida we can see how close the launch site is to the coast line

- Rockets are usually not launched over a major road or town, launch sites tend to stay away from built up areas to ensure safety



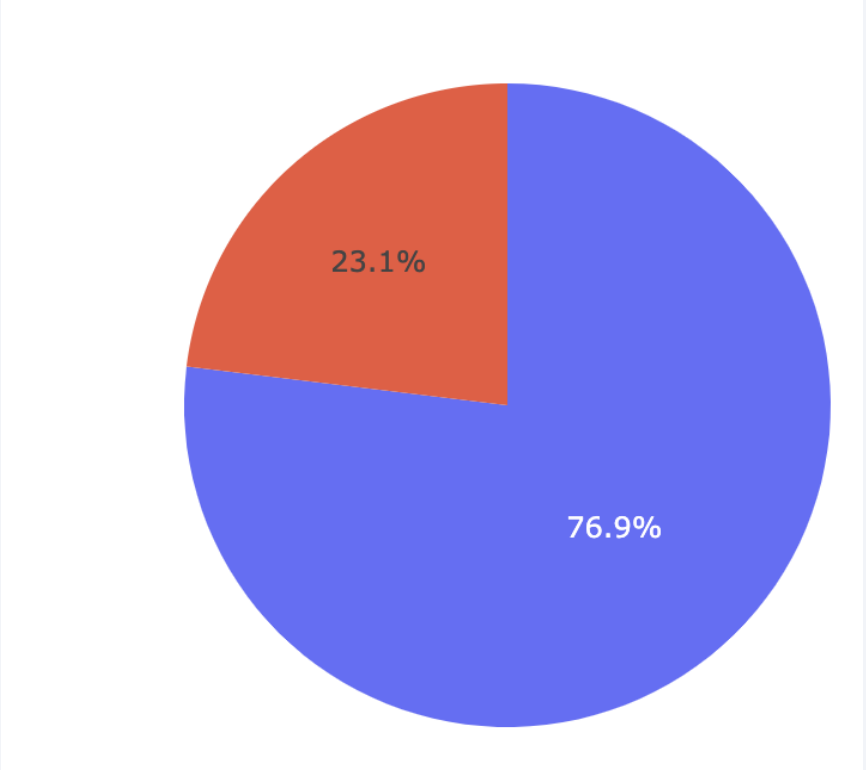

Section 4
Build a Dashboard with Plotly Dash

Total successful launches by site

- The chart shows all successful launches and how they are divided by site
- KSC LC-39A has the highest number of successful launches
- The lowest share of successful landings are at CCAFS SLC-40

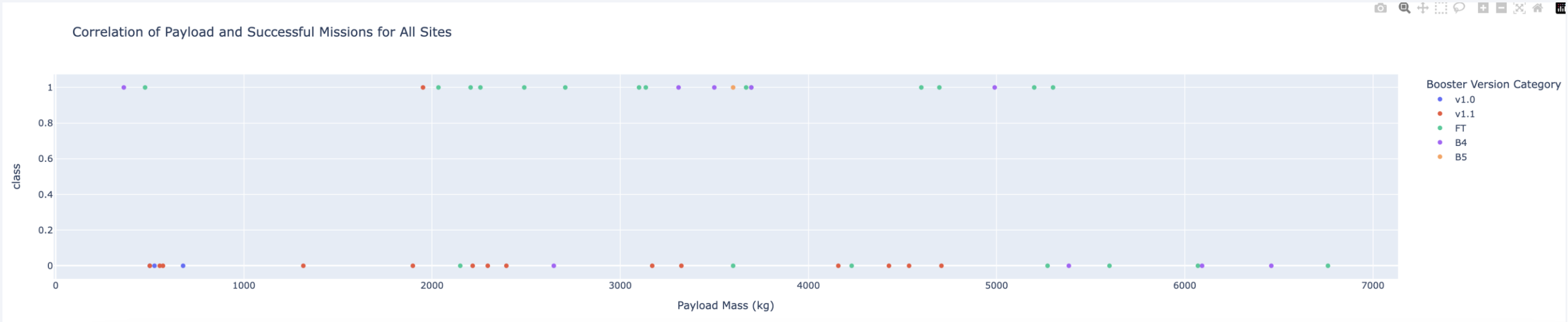
Most successful landing site

- KSC LC-39A shows a success rate of almost 77% in terms of landings
- It is the highest among all launch sites
- This could indicate this is a site the company should focus on.



Payload vs Booster version correlation

- The correlation chart below shows the successful and failed missions sorted by payload mass and booster version.
  - Class 1 = success and class 0 = failure
- We can see that booster version v1.0 and v.1.1 almost only have failed missions
- The top booster versions seem to be FT and B4, however, as payload mass increases, so does failure

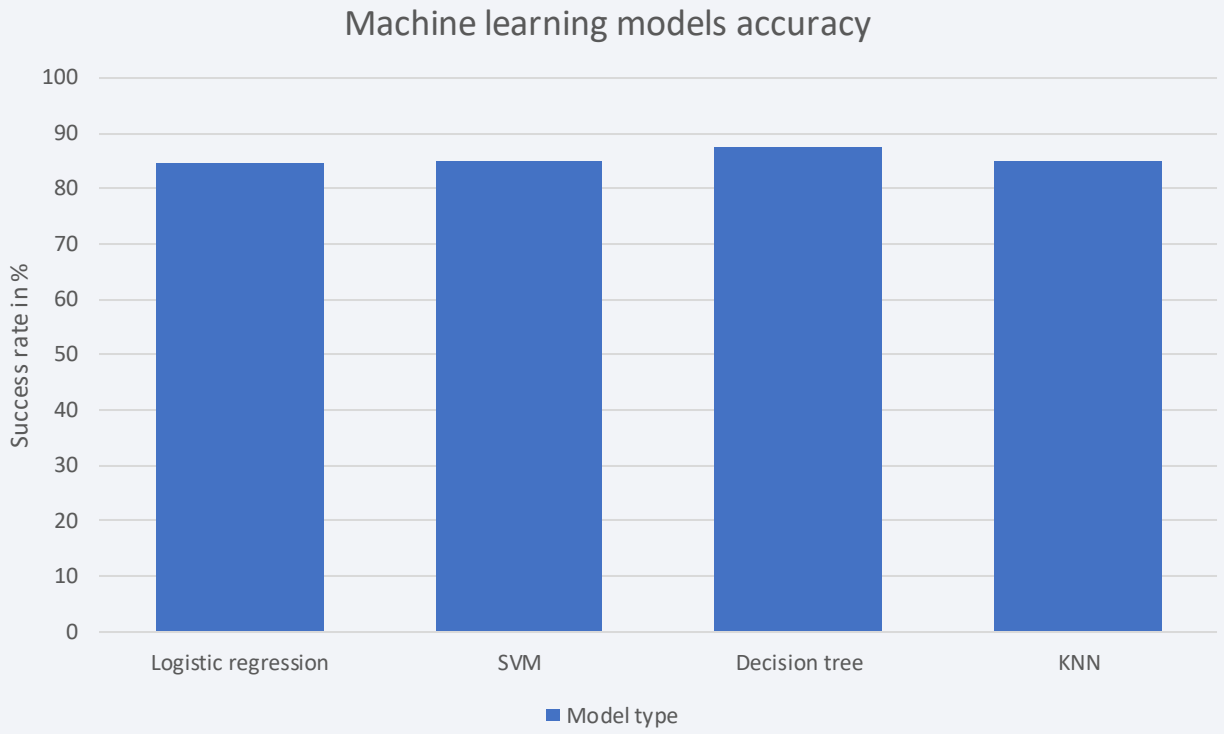


Section 5

Predictive Analysis (Classification)

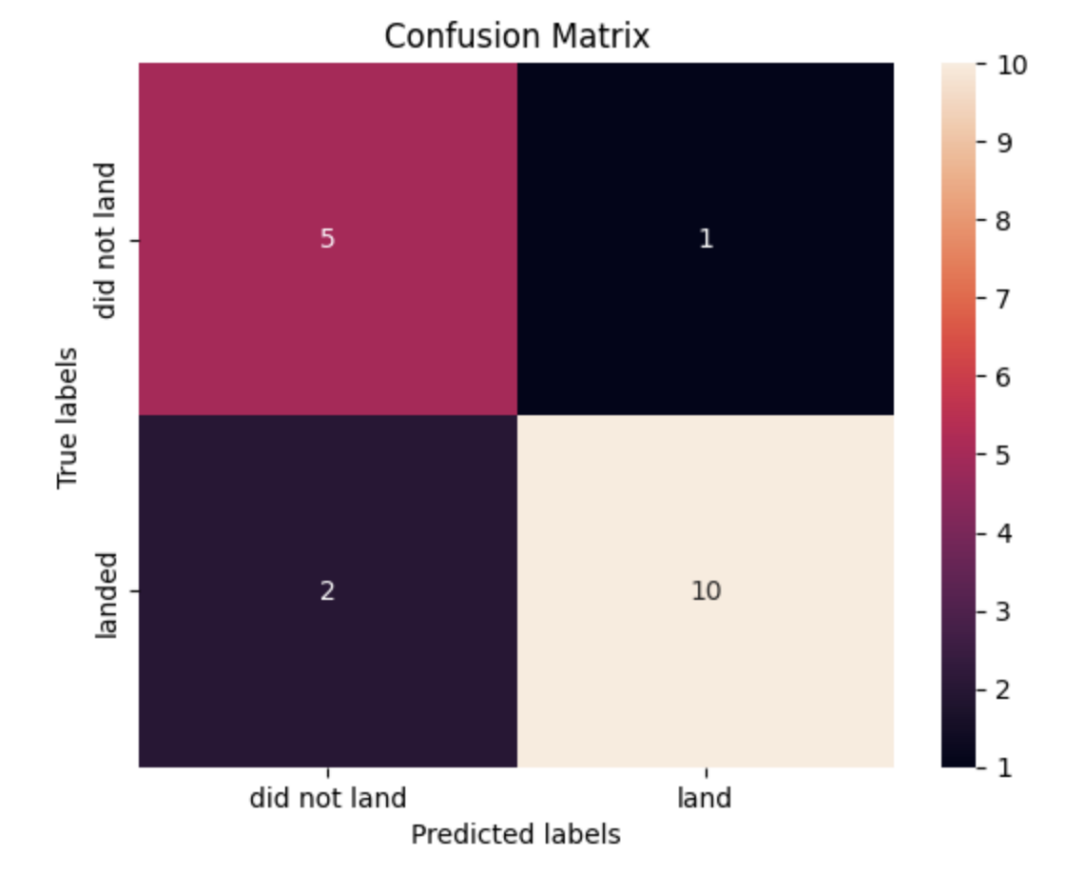
Classification Accuracy

- The bar chart shows the model accuracy for the different models we tried on the data
- The highest accuracy was the decision tree of 87.5% but there were no major differences



Confusion Matrix

- This shows the decision tree model confusion matrix
- Predicted landing and actually landed has a score of 10 vs just 1 false prediction
- Predicted "did not land" and actually did not land has a score of 5 vs 2 wrongly predicted



Conclusions

- Launch sites should be near the coast and not next to a major population center
  - However it needs to be connected by road and rail in order to transport parts and employees
- Success increases with the number of flights
  - Prepare to put significant investment in the learning process to reach the results wanted
- The top landing site, KSC LC-39A, has a success rate of around 77%
- The higher the payload mass, the less likely it is that the rocket will return
- Booster version FT and B4 should be preferred
- Type of orbit should be carefully considered, ES-L1, GEO, HEO and SSO have a 100% success rate while GTO is as low as 50%
- The machine learning models are fairly accurate at predicting failed or successful mission with the highest accuracy being 87.5% (decision tree model)
  - This is a solid model that can be used for our own data

Thank you!