IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kai Machida
24 August 2025

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection: Web Scraping and API
  - EDA: SQL and Visualisation
  - Visual Analytics: Folium map and Dashboard
  - Predictive Analysis: Machine Learning Classification Models
- Summary of all results:
  - Relationship of different features related to rocket launch
  - Location attributes of launch sites
  - Classification models for predicting a successful landing

Introduction

- Project background and context:
  - The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful is SpaceX. One reason for SpaceX's success is its rocket launches are relatively inexpensive. This is due to their Falcon 9 rockets, which can reuse their first stage.
- Problems you want to find answers
  - Determine the cost of a launch, by predicting if the first stage will land. We will do this by using public information to train a machine learning model that will predict if the first stage will land successfully.

Section 1

Methodology

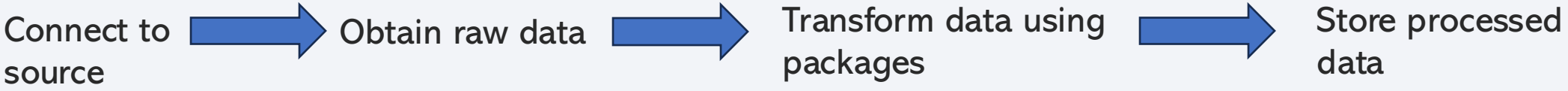
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by using the SpaceX API and web scraping from Wikipedia articles.
- Perform data wrangling:
  - Using the collected data, created a feature that recorded whether a landing was successful or not. This will be the dependent variable in our classification model.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - We first standardized our data and split the data into training and testing data. We considered four classification models: Logistic Regression, SVM, Decision Tree, and KNN. For each model we used cross-validation and a grid search to choose the best parameters for our model. We finally evaluated our optimized models using the test data, computing accuracy scores and confusion matrices.

Data Collection

- Collection method:
  - Data sets were collected by two methods, using the SpaceX API and web scraping a Wikipedia article.



Data Collection – SpaceX API

- https://github.com/machida97/IBM-Coursera-Applied-Data-Science-Capstone-Project/blob/45f2a0a753f8fb1d7e60a374d9ba563afb5dacbb/jupyter-labs-spacex-data-collection-api.ipynb

Data collection with SpaceX REST calls:

Request and parse SpaceX launch data using GET request

Get data frame with required features

Filter for only Falcon 9 rockets

Deal with Missing values

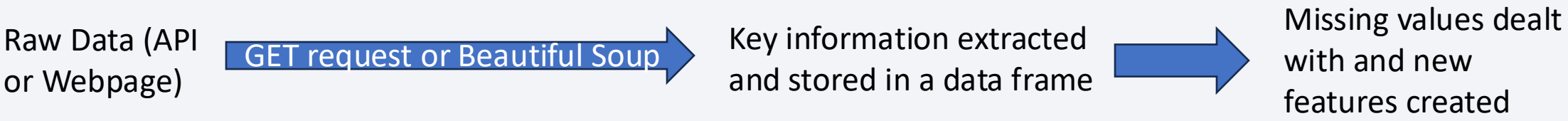
Data Collection - Scraping

- https://github.com/machida97/IBM-Coursera-Applied-Data-Science-Capstone-Project/blob/45f2a0a753f8fb1d7e60a374d9ba563afb5dacbb/jupyter-labs-webscraping.ipynb

Web scraping process:

Data Wrangling

- Data Wrangling processes:



- https://github.com/machida97/IBM-Coursera-Applied-Data-Science-Capstone-Project/blob/45f2a0a753f8fb1d7e60a374d9ba563afb5dacbb/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts plotted:
  - Scatterplots: Visualize the relationships between Flight number vs Launch Site, Payload Mass vs Launch site, Flight number vs Orbit type, Payload Mass vs Orbit type, across successful and unsuccessful relationships
  - Bar chart: Visualize the successful landing rate for each orbit type
  - Line chart: Visualize the relationship between average successful landing rate and year.

- https://github.com/machida97/IBM-Coursera-Applied-Data-Science-Capstone-Project/blob/45f2a0a753f8fb1d7e60a374d9ba563afb5dacbb/edadataviz.ipynb

EDA with SQL

- SQL queries performed:
  - Display names of unique launch sites
  - Display relation between payload mass and booster
  - Date analysis for landing outcomes

- https://github.com/machida97/IBM-Coursera-Applied-Data-Science-Capstone-Project/blob/45f2a0a753f8fb1d7e60a374d9ba563afb5dacbb/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map:
  - Added markers and circles to a folium map to mark the locations of each launch site in our data on a map of the United States.
  - Added icons and marker clusters to each launch site to mark successful and unsuccessful landings, giving us the ability to easily tell which launch sites have high success rates.
  - Added lines from launch sites to locations of interest to visualize proximity from launch sites, to understand the chosen locations of launch site relative to surrounding areas. These items included:
    - Nearest coastline
    - Nearest City
    - Nearest Highway
    - Nearest Railway

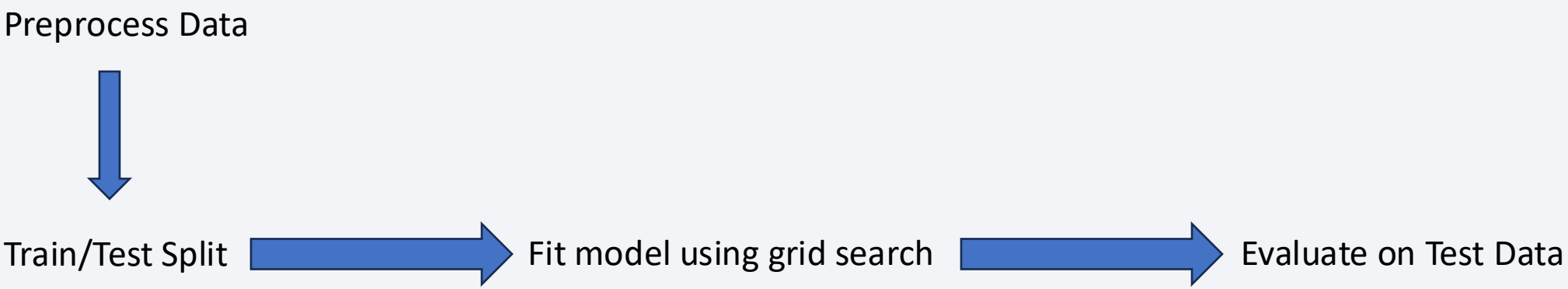
- https://github.com/machida97/IBM-Coursera-Applied-Data-Science-Capstone-Project/blob/0c3c758777cf309dc2b0f35dc5363f45b4cb76aa/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard:
  - Launch site drop down so that we could filter by specific launch site, with default option being all sites
  - Pie chart that showed the success count (for all sites), and the success rate (for selected launch site) so we could see which site had the largest success count, and which site had the highest success rate
  - Range slider to select a payload range
  - Scatter plot with x axis payload and y-axis the success outcome, which can be filtered by the payload range slider and launch site drop down. This let us visually see how payload can be related to mission outcome based on selected launch sites. The points are also colour labelled by booster version, so we can observe mission outcome with different boosters
- Add the GitHub URL of your completed Plotly Dash lab: https://github.com/machida97/IBM-Coursera-Applied-Data-Science-Capstone-Project/blob/45f2a0a753f8fb1d7e60a374d9ba563afb5dacbb/spacex-dash-app.py

Predictive Analysis (Classification)

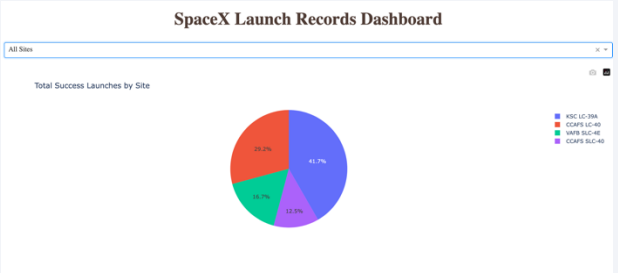
- We considered 4 models: Logistic Regression, SVM, Decision Tree and KNN
- Our process:

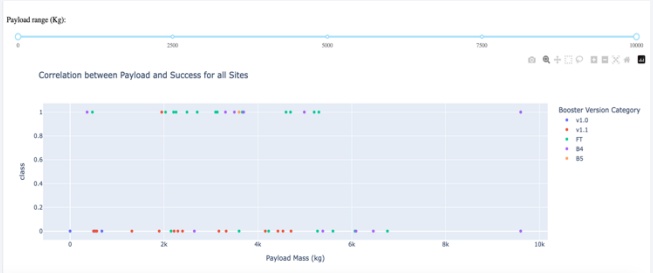


- https://github.com/machida97/IBM-Coursera-Applied-Data-Science-Capstone-Project/blob/45f2a0a753f8fb1d7e60a374d9ba563afb5dacbb/SpaceX_Machine%20Learning%20Prediction_Part_5(1).ipynb

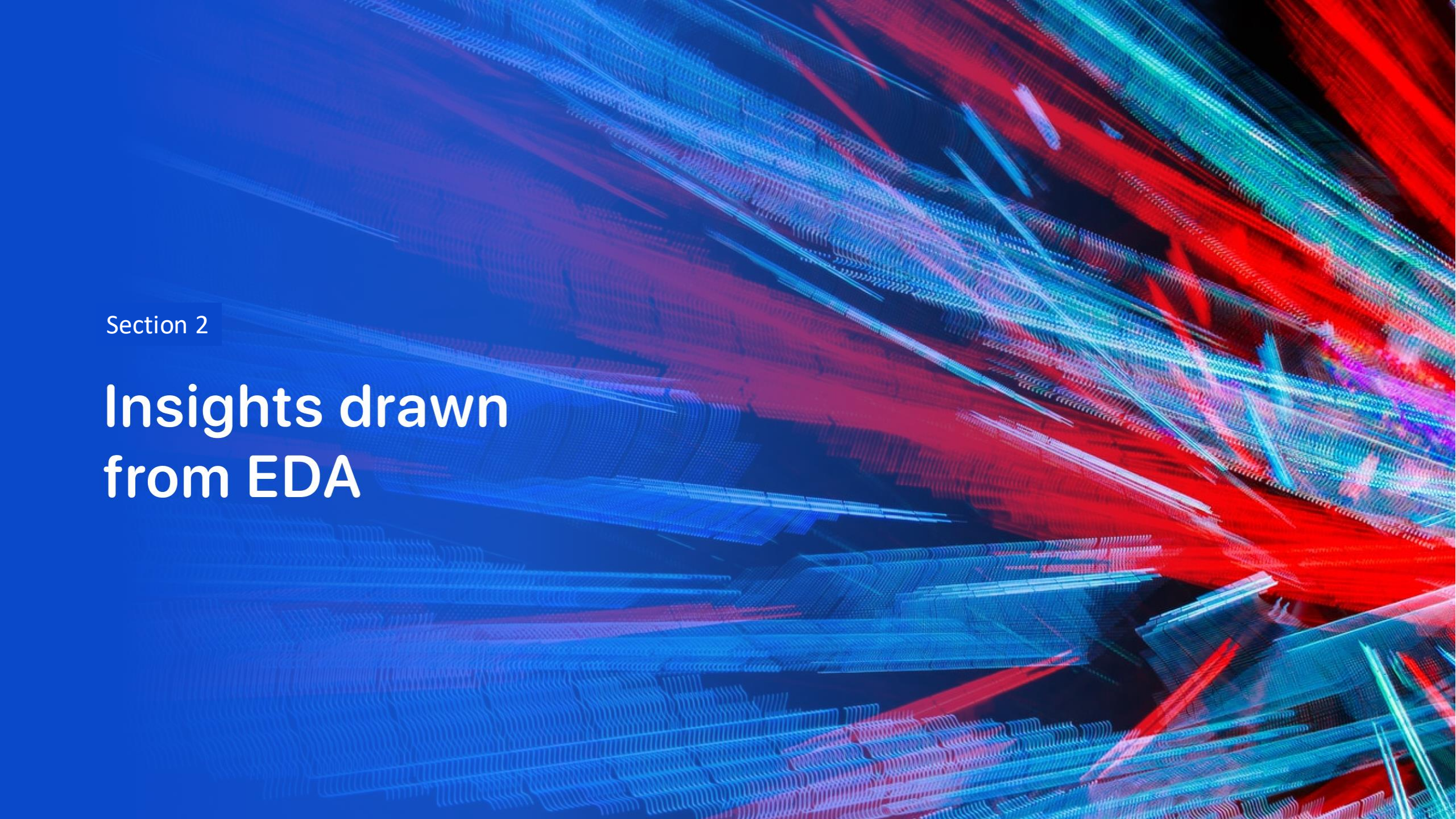
Results

- Exploratory data analysis results:
  - The findings under each task from each Data analysis (Vis and Sql) notebooks
  - From our visual data analysis we determined that the features Flight Number, Payload Mass, Orbit, Launch Site, Flights had relations with class outcomes and so would be used in our predictive analysis
- Interactive analytics demo in screenshots:



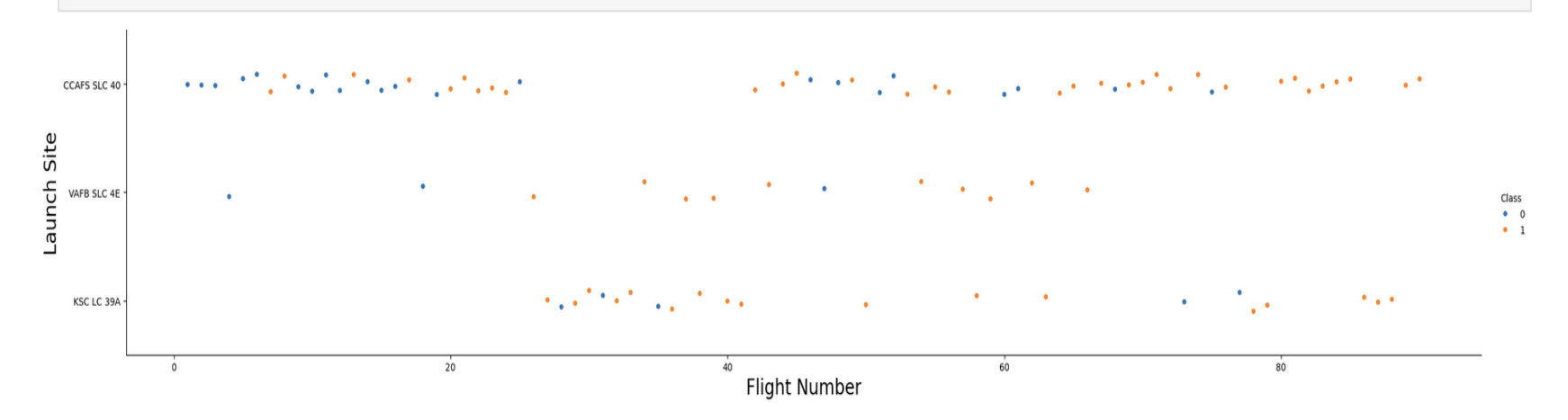


- Predictive analysis results:
  - All models performed equally well, with an accuracy score of 0.83 for predicting the success of a landing


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

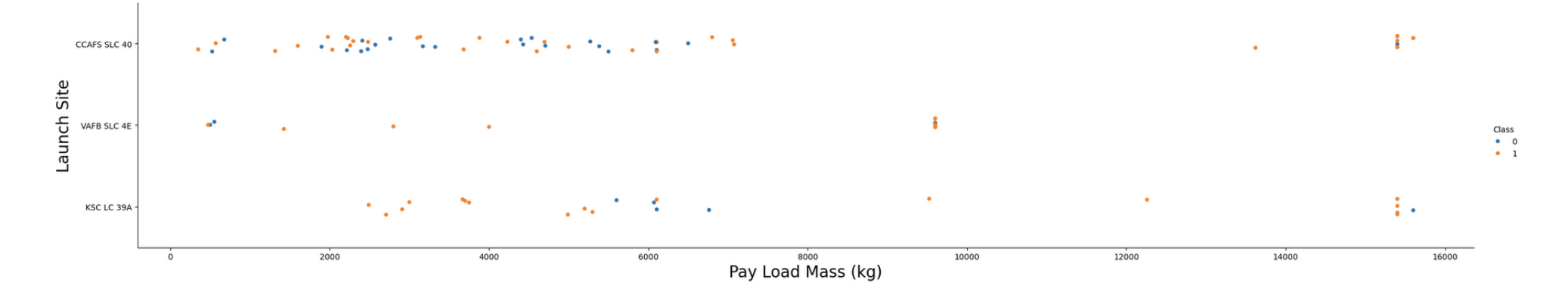
- Scatter plot of Flight Number vs. Launch Site



- For Launch site CCAFS SLC 40 it seems that for flight numbers 60 and above, there are much more successful landings.
- For VAFB SLC 4E there are no successful landings between 0 and 20 flight numbers, but from 20 and above, there are much more successes, but no results above 60 flight numbers.
- For KSC LC 39A there are far more successes when the flight number is greater than 40. In general, higher flight numbers (50 or above) seem to give more successes across all launch sites.

Payload vs. Launch Site

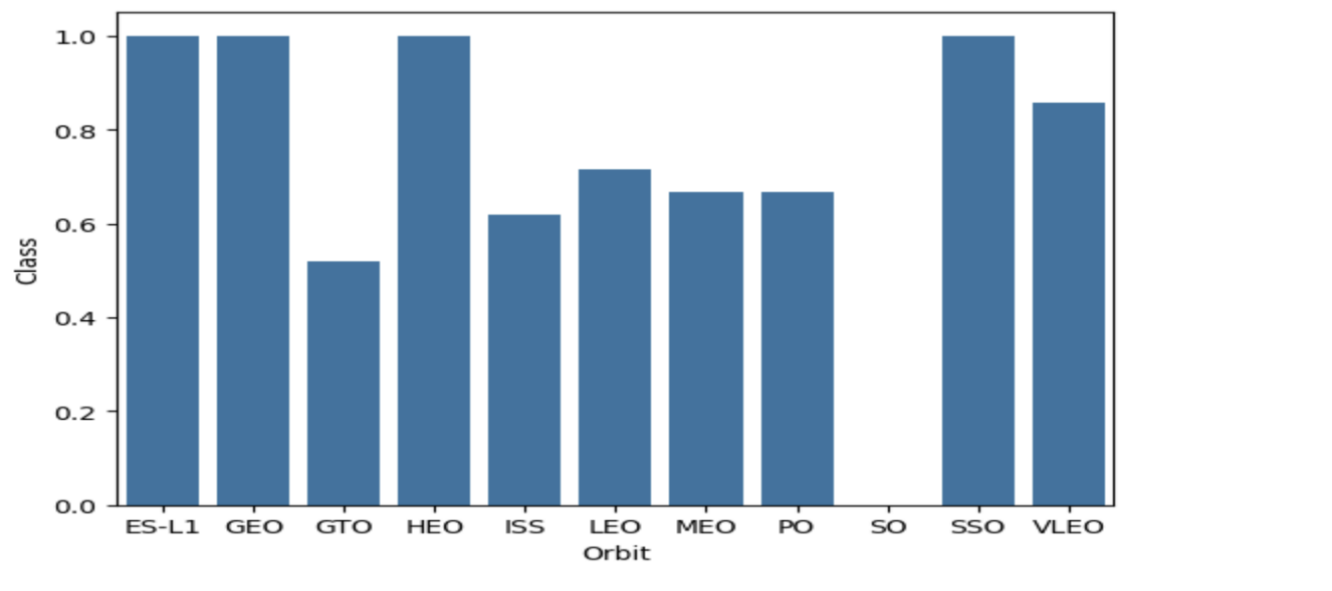
- Scatter plot of Payload vs. Launch Site



- Higher Payload tends to have higher success, but there are no results for VAFB SLC aboe 10000.

Success Rate vs. Orbit Type

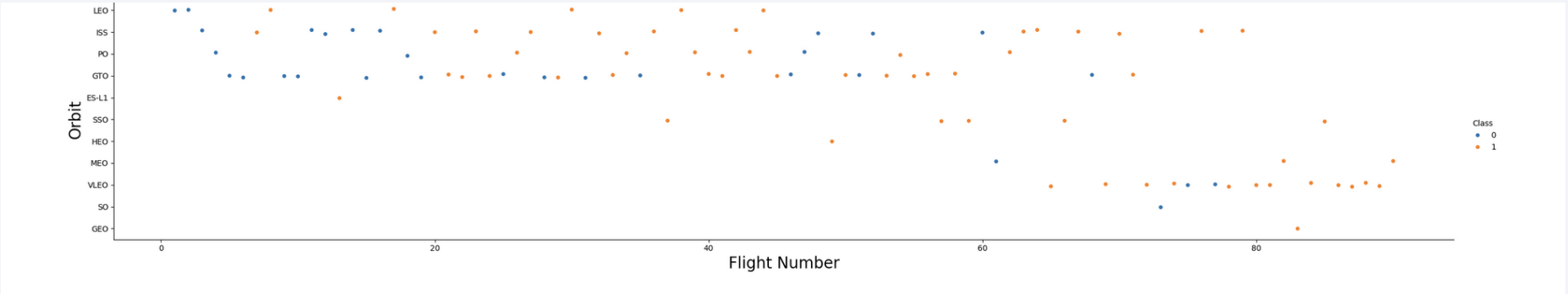
- Bar chart for the success rate of each orbit type



- ES-L1, GEO, HEO, SSO had the highest success rate at 1.

Flight Number vs. Orbit Type

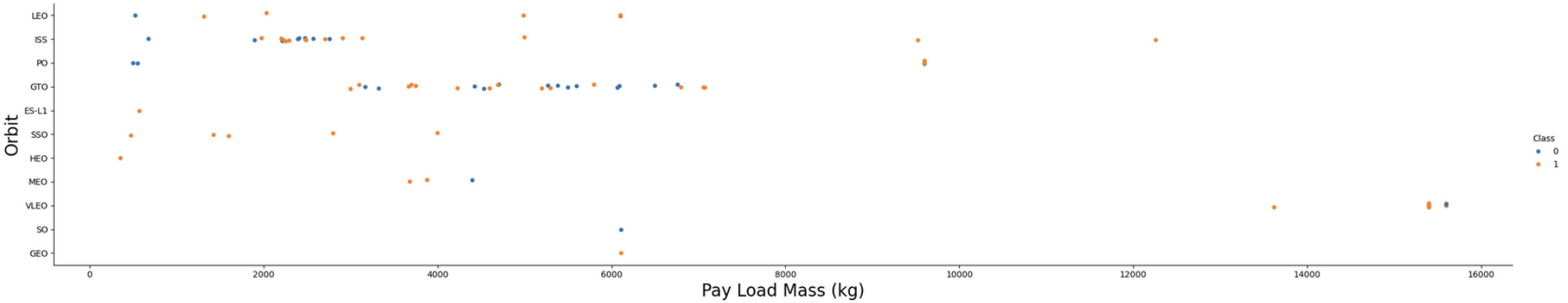
- Scatter point of Flight number vs. Orbit type



- In the LEO orbit, Flight Number seems to be related to success. Conversely, in GEO there does not seem to be any relationship.

Payload vs. Orbit Type

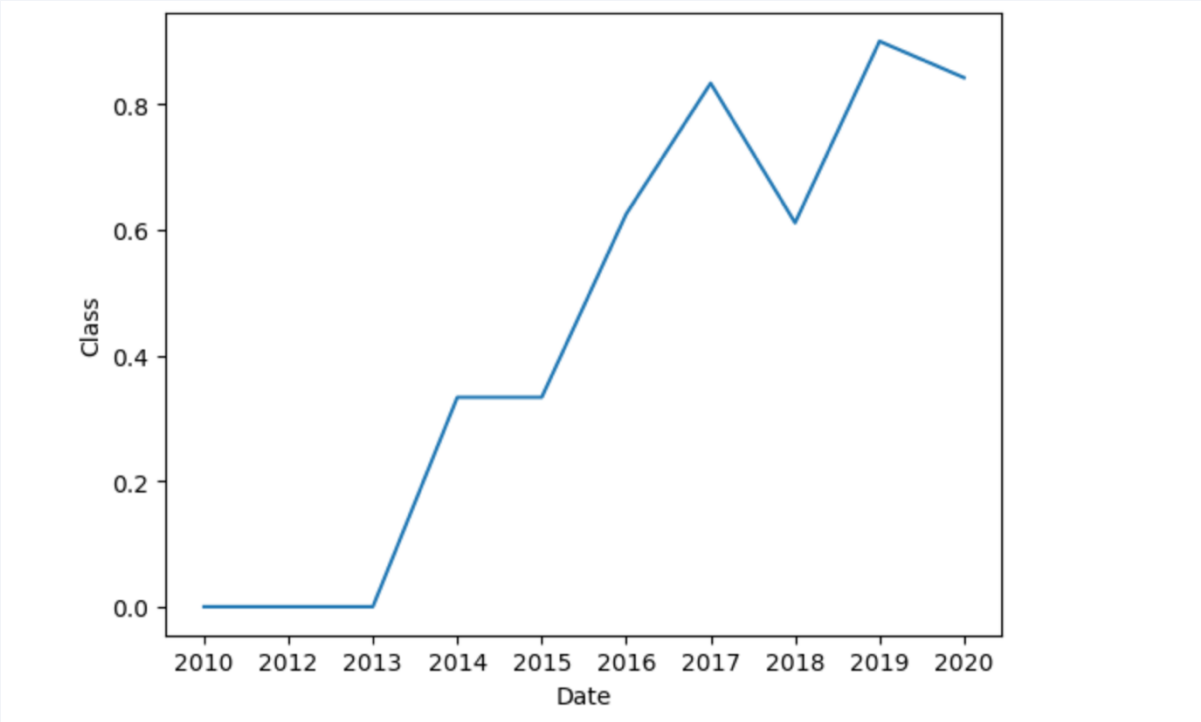
- Scatter plot of payload vs. orbit type



- For heavier payloads, the successful landing rate is higher for LEO, ISS, and Polar. However, for GTO its difficult to distinguish between successful and unsuccessful landings as both are present.

Launch Success Yearly Trend

- Line chart of yearly average success rate



- The landing success rate continued increasing from 2013 to 2020

All Launch Site Names

- Find the names of the unique launch sites

```
4]: %sql select distinct Launch_Site from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

4]:

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

- The launch sites are 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`

```
In [19]: %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

Out[19]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lar
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Fa
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	Fa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
[21]: %sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
[21]: total_payload_mass
45596
```

- The total payload is 45596 (kg)

Average Payload Mass by F9 v1.1

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

```
[22]: %sql select avg(PAYLOAD_MASS__KG_) as avg_payload_mass from SPACEXTABLE where Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
[22]: avg_payload_mass
2928.4
```

- The average payload mass carried by booster version F9 v1.1 is 2928.4 (kg)

First Successful Ground Landing Date

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

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

min(Date)
2015-12-22

- The first successful landing outcome on ground pad was 22/12/2015

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 SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

File display

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

]:

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

- The boosters are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcome

```
%sql select count(*) from SPACEXTABLE
```

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

count(*)
101

- There were a total of 101 missions

Boosters Carried Maximum Payload

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

```
%%sql select distinct Booster_Version
from SPACEXTABLE
where PAYLOAD_MASS__KG_
= (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * 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

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

```
%%sql select date, substr(Date, 6, 2) as month, 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.
```

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

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

```
%%sql select Landing_Outcome, count(*) as total_number from SPACEXTABLE
group by Landing_Outcome
having Date between '2010-06-04' and '2017-03-20'
order by total_number DESC
```

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

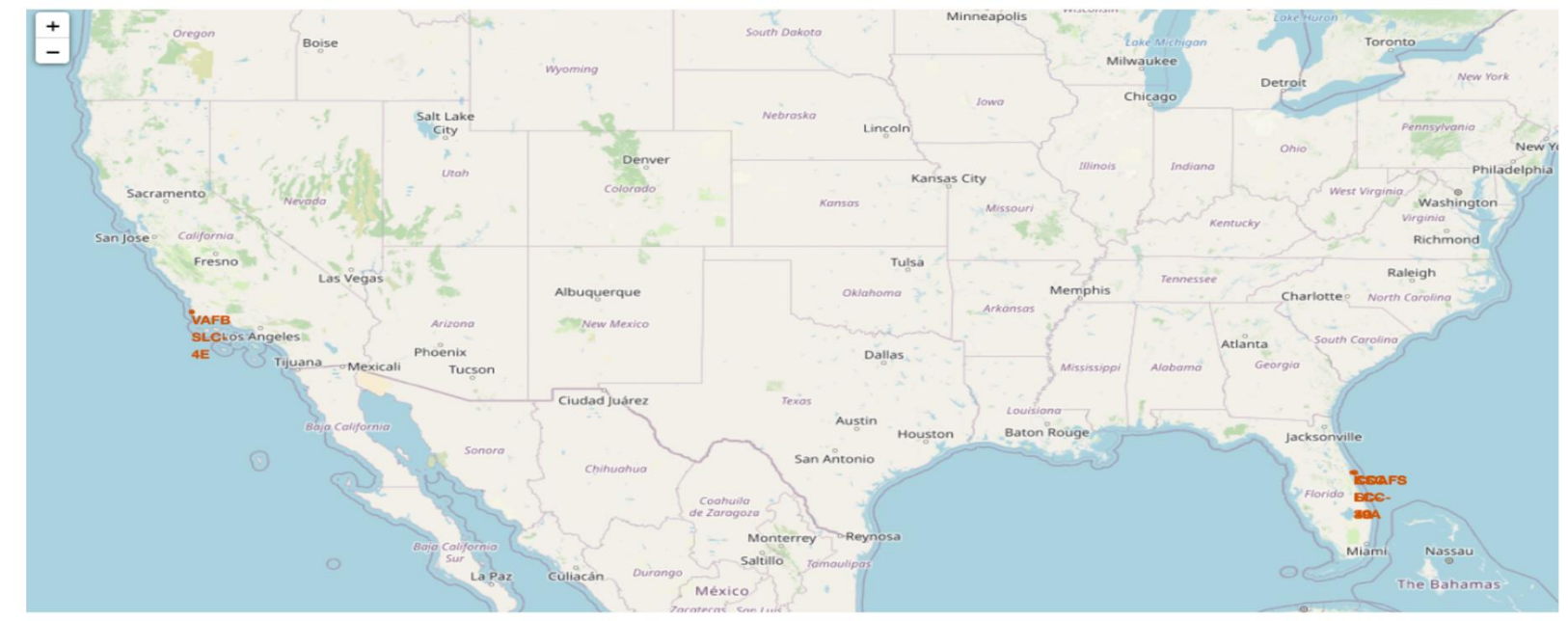
Landing_Outcome	total_number
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch site locations

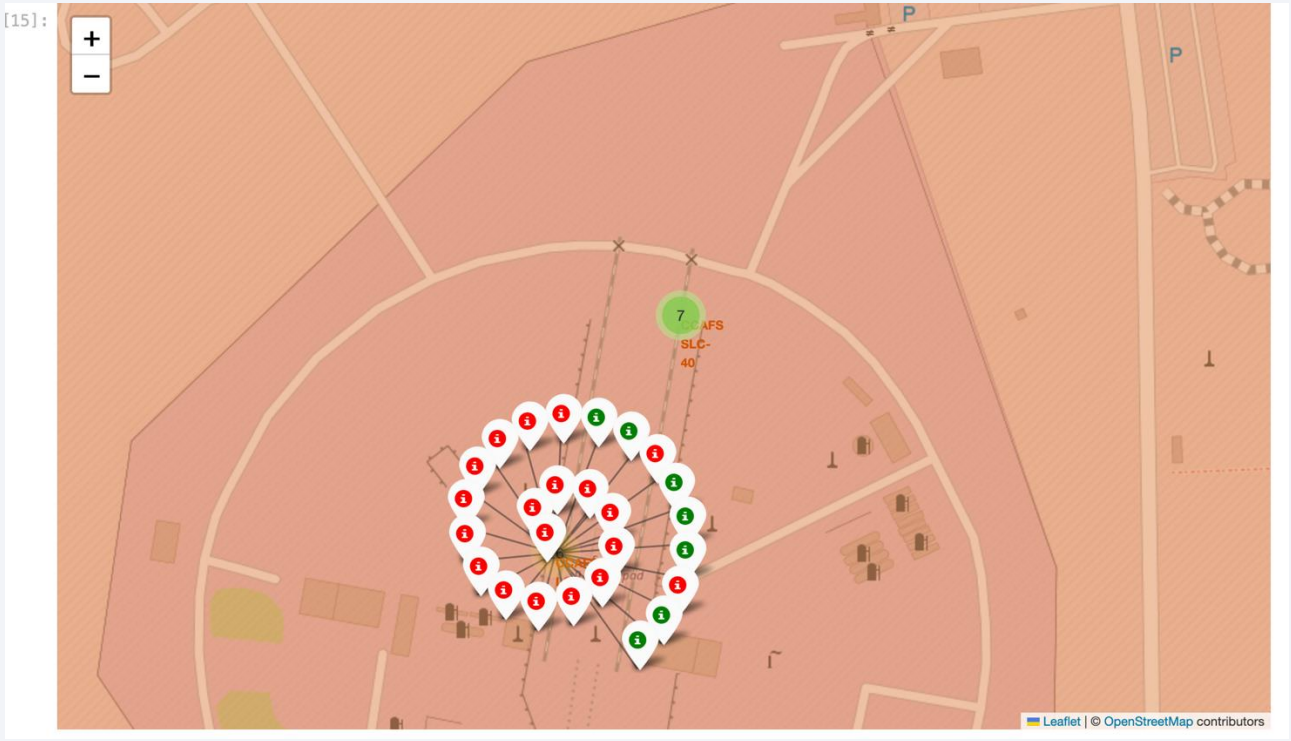
- Launch sites' location markers on a global map



- Launch sites are near the equator, due to better average weather, which is important for rocket launches
- Launch sites are near coastlines for safety reasons. Should a rocket launch fail, any debris will only hit the ocean rather than built up areas

Launch Outcomes

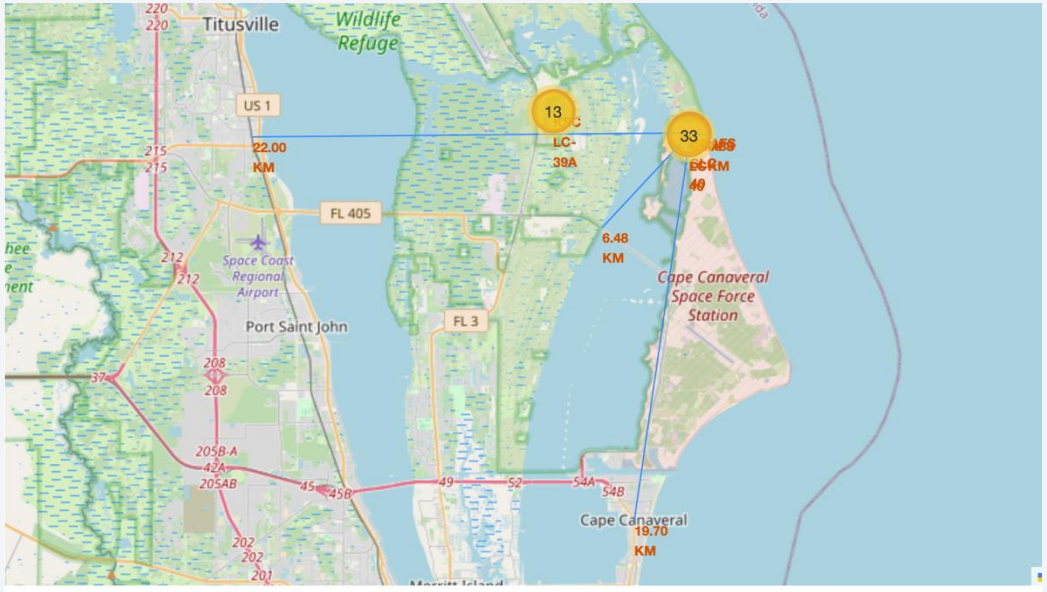
- Color-labeled launch outcomes on the map. Green indicates a successful landing, red an unsuccessful landing



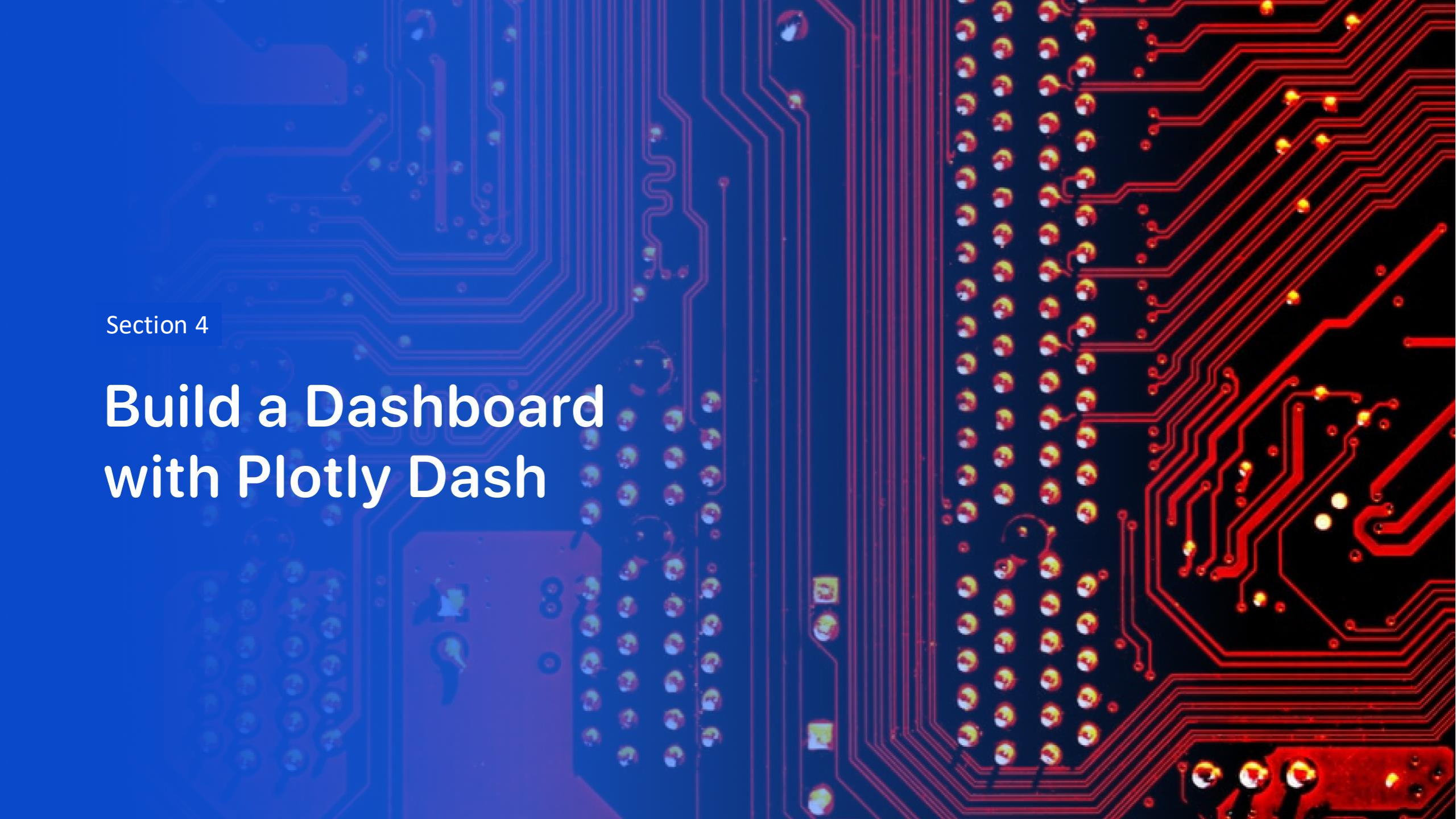
- Using this visualization, KSC LC-39A is the launch site with the highest success rate

Launch site proximities

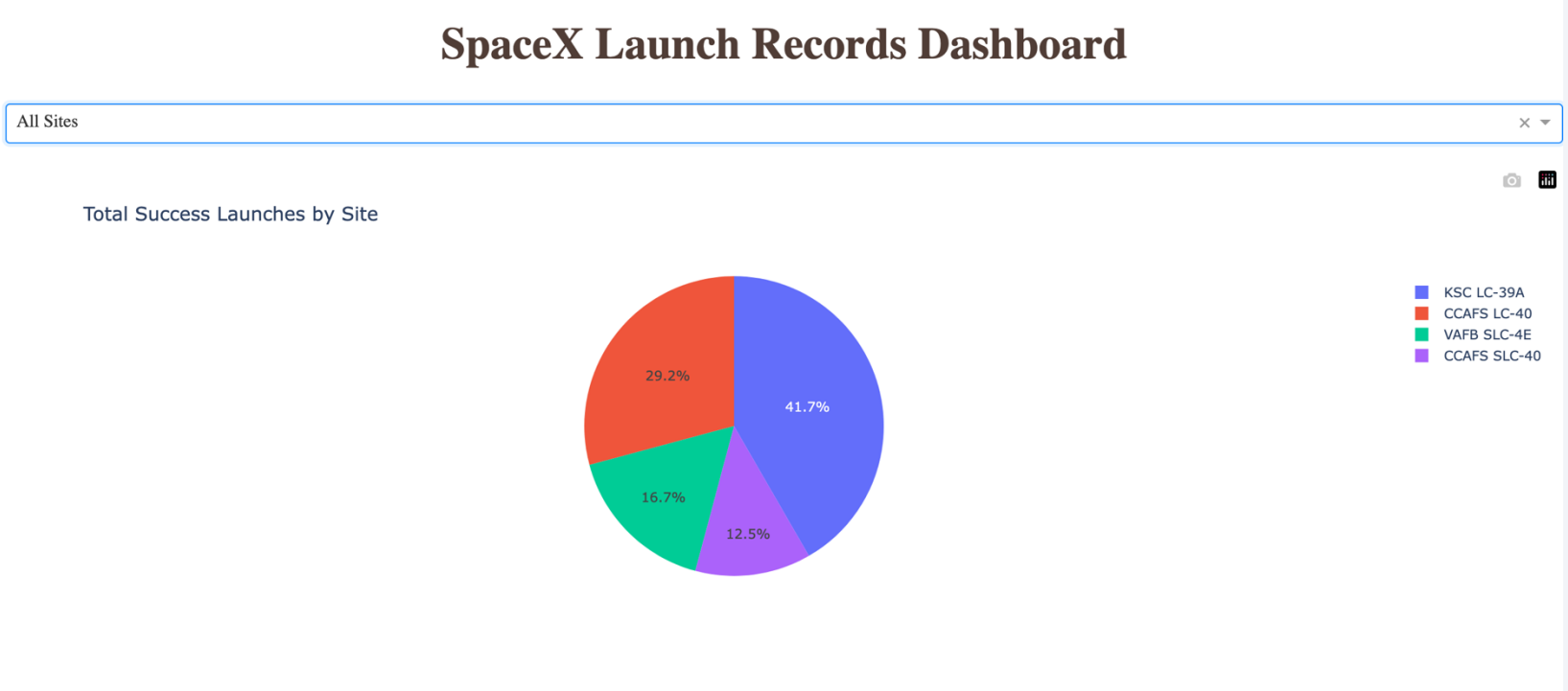
- Launch site and its proximity (shown in blue line) to railway, highway, coastline and nearest city.



1. In general launch sites are not near railways as supplies etc can be brought in by highway and a failed launch can risk damaging a railway, an important supply route
2. Launch sites are relatively close to highways for logistics purposes, as they need to bring in fuel, people to near by launch stations. The highway outlined seems mostly to service the launch station
3. Launch sites are in general close to coastlines for safety reasons. Should a launch malfunction and a rocket need to abort, it can safely crash into the nearby sea without harming people
4. Launch sites are kept a certain distance away from cities. The nearest city is Cape Canaveral, which is 20km away. This is for obvious safety reasons. Should a rocket malfunction, having a city nearby could lead to catastrophic damage

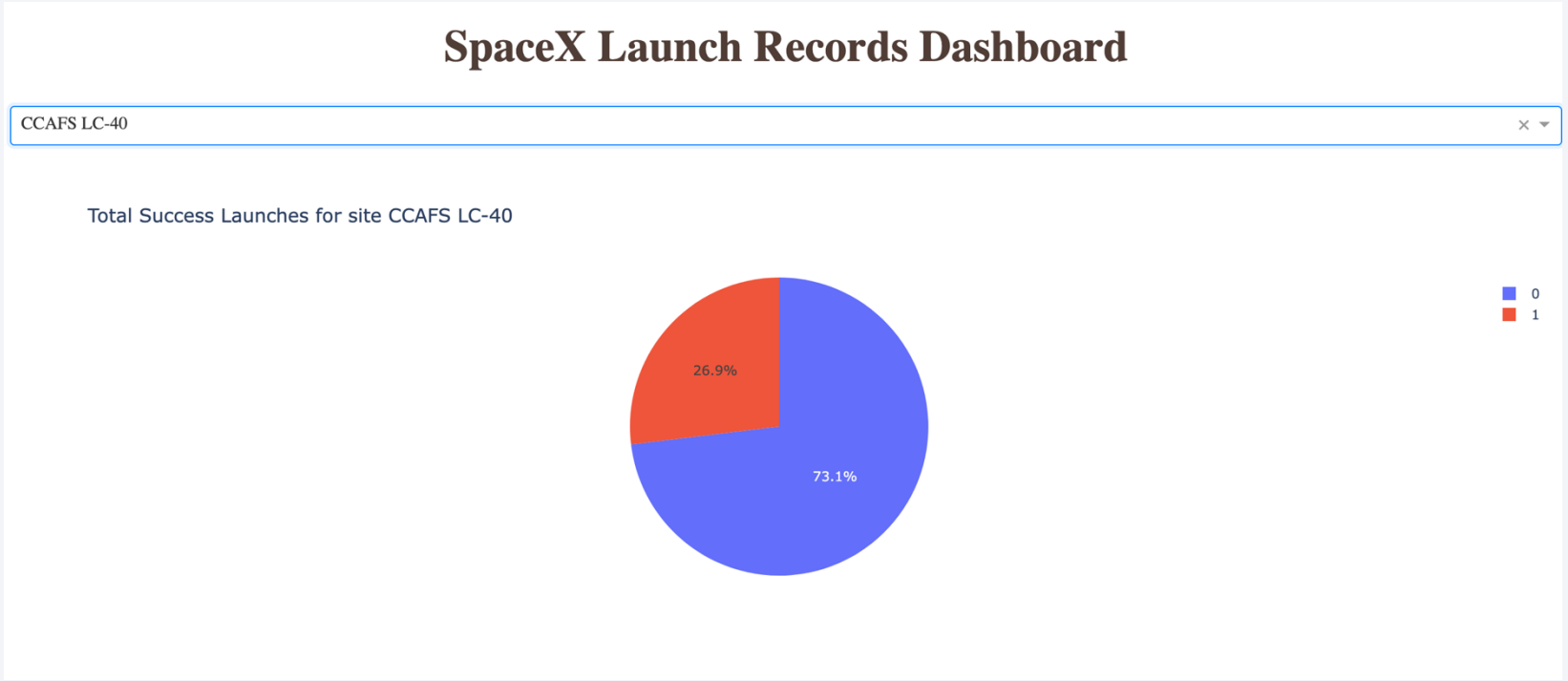

Section 4
Build a Dashboard with Plotly Dash

Launch success counts across all sites



- Launch site KSC LC-39A has the largest number of successful launches, with 41.7% of all successful launces

Highest launch success ratio



- Launch site CCAFS LC-40 has the highest launch success ratio, with 73.1% of its launches being successful

Payload vs Launch Outcome

- The payload range with the highest success rate is 3000-4000: 70% (7 successes, 3 failures)

- The payload range with the lowest success rate is 6000-10000: 17% (5 failures, 1 success)

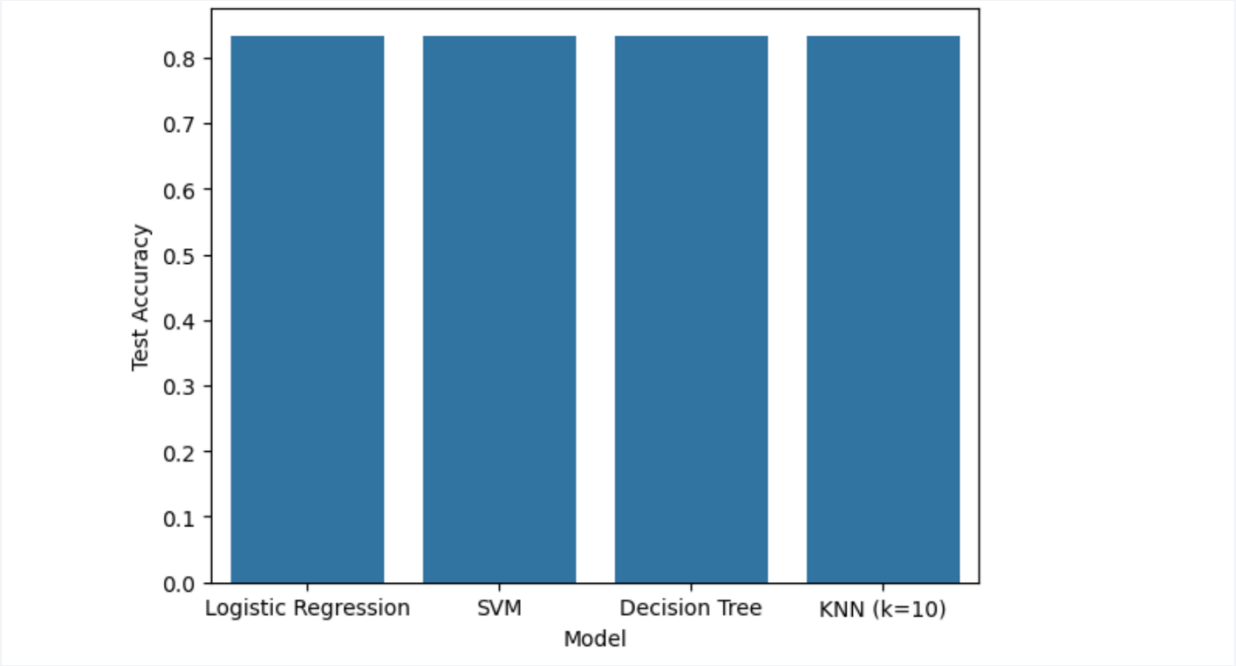
- The F9 booster version with the highest launch success rate is B5 with 100% success, though there is only 1 launch so there is not a lot of data. The next best success rate, with significantly more data is FT with a 65% success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

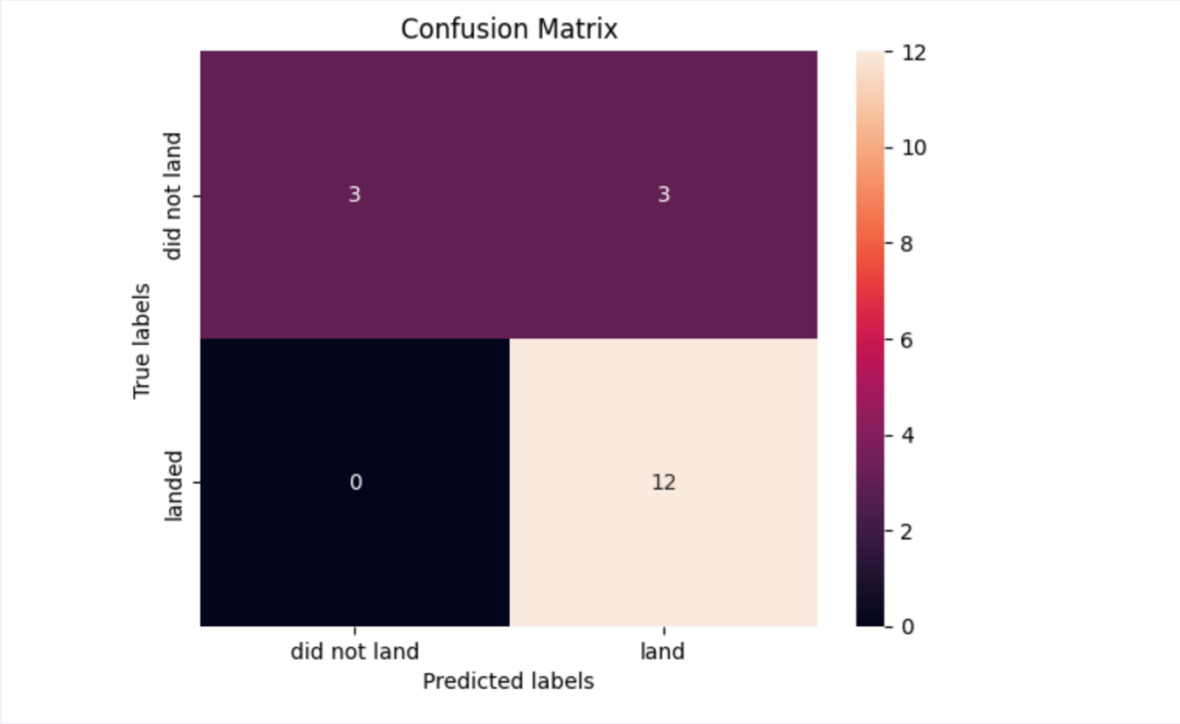
- Accuracy vs Model



- All models had the same accuracy score of 0.83.

Confusion Matrix

- Confusion matrix of KNN (k=10)



- This model correctly predicts 15 landings, only producing 3 False positive results

Conclusions

- Using a classification model, we could predict the success of a reasonable accuracy of 0.83. Our model successfully predicted all outcomes except 3 False Positives, where it incorrectly predicted 3 launches as successful, when they were not
- We can deploy our classification to predict the success of a rocket launch and so determine the cost of a launch

Appendix

- Data sets used:
  - Data from SpaceX API
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Thank you!