IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Steve Daly
June 15th 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  -Exploratory Data Analysis (visualization & SQL)
  -Interactive Map with Folium
  -Interactive dashboard with Plotly Dash
  -Predictive analysis (classification models)
- Summary of all results

  - Exploratory Analysis Results

  -Dashboard Outputs

  -Predictive Analysis Results

Introduction

- SpaceX is leading the way in commercial space exploration. They have set themselves apart from their competitors in the commercial space due the affordability they provide. This affordability results from SpaceX being able to reuse the first stage of their Falcon 9 rocket.
- In this project we were able to identify different factors relating to each rocket launch to predict if a rocket will successfully land.

Section 1

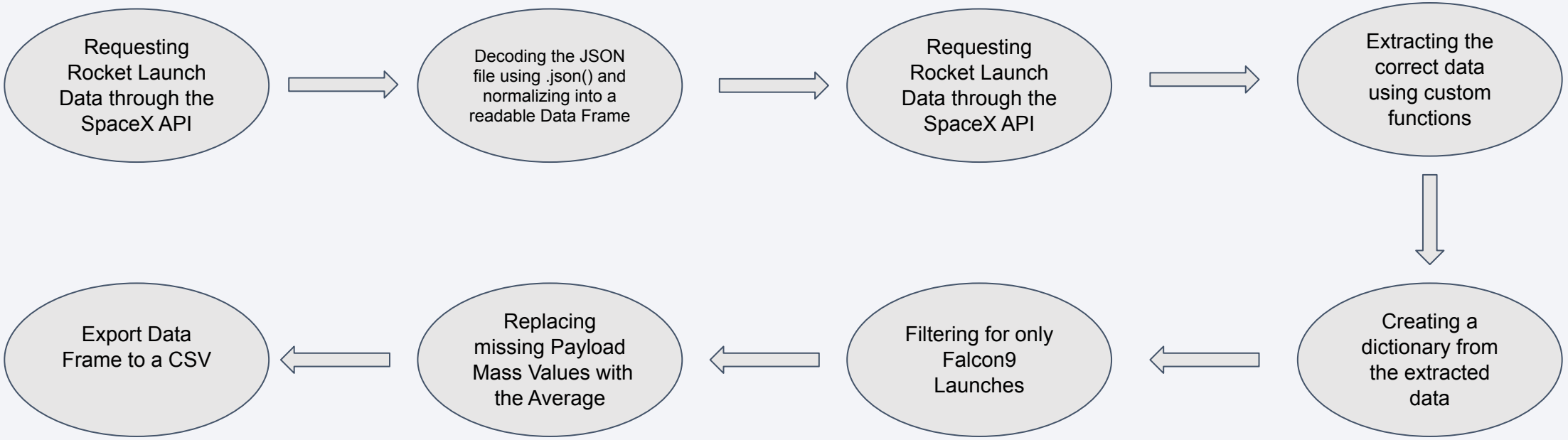
Methodology

Methodology

Executive Summary

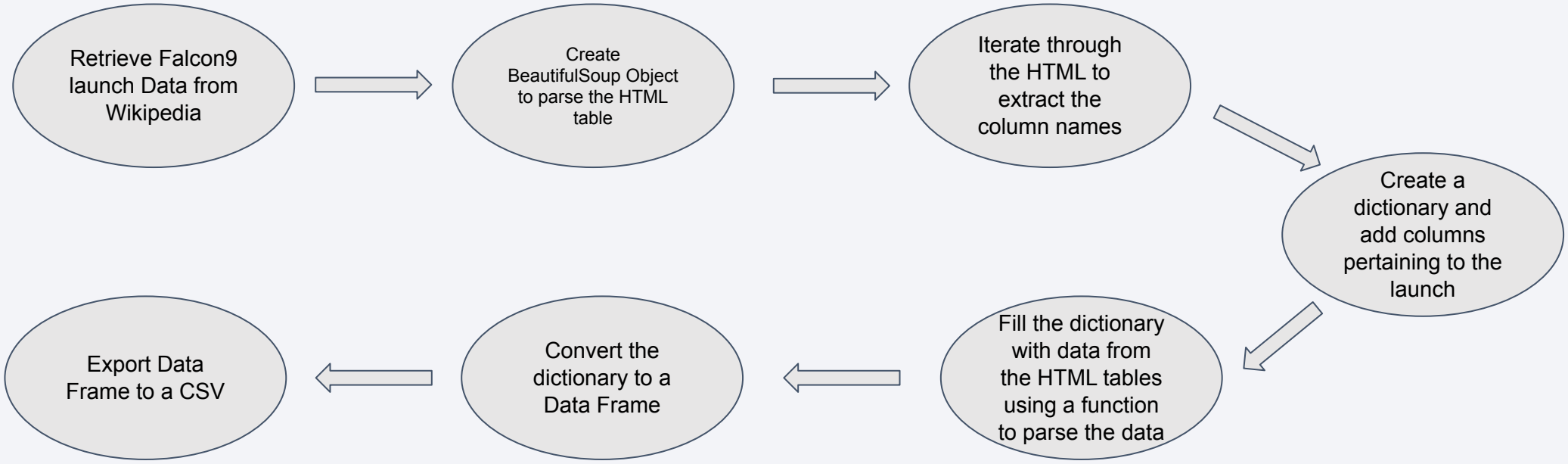
- Data collection methodology:
  - The data was collected using SpaceX's API. Starting in a JSON file, the data was then normalized to a readable Data Frame. Data was also collected using web-scraping on wikepedia.
- Perform data wrangling
  - NULL values in the data for Payload Mass with the averages amongst the dataset. Landing outcomes were converted to a binary 1 or 0 value in the data set to be best suited for our classification models.
- 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 normalized and split the data into training and test sets to run in multiple predictive models to identify which had the most accurate outcome.

Data Collection- SpaceX API



https://github.com/Steve-Dal/Data-Science-IBM-Captstone/commit/baf55fc08f55af4ec2f67f437410c59337380f70#diff-b6b8abd9ceea0d9bbbdccf909540f0723966abe4f4be4cb62455c2a3165816e6

Data Collection- Scraping



https://github.com/Steve-Dal/Data-Science-IBM-Captstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Each row in the data frame represented a particular launch. The columns in the data frame contained data about the launch like the orbit type the launch was aimed to, what the mass of the payload was and the landing outcome.
- The desired predicted variable was Landing outcome but there were multiple categorical values that equated to a successful landing and a failure to land
- Each the landing outcomes was converted to a numeric value to feed into the classification model later on. A value of 1 represented a successful landing and a 0 represented a failure to land

Data Frame Original Outcomes	Original Landing Outcome Meanings
True ASDS	Successful landing on drone ship
None None	Failure to land
True RTLS	Successful landing on a ground pad
False ASDS	Unsuccessful landing on drone ship
True Ocean	Successful landing in specific part of the ocean
False Ocean	Unsuccessful landing in a specific part of the ocean
None ASDS	Failure to land
False RTLS	Unsuccessful landing on ground pad

Class
1
0

https://github.com/Steve-Dal/Data-Science-IBM-Captstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Multiple different types of charts were created to provide insights to the data.
  - Scatter Plot Created: Flight Number vs Payload Mass, Launch SIte vs Flight Number, Payload Mass vs Launch Site, Flight Number vs Orbit & Payload Mass vs Orbit.
  - Bar Chart Created: Orbit type vs Average Success Rate
  - Line plot Created: Average Success Rate over time

- Scatter plots displayed the relationship between variables to determine if they would be a good fit for a machine learning model.
- Line charts identified trends over time.
- Bar Charts showed comparisons of discrete values to display the relationship between categories and a measured value.

https://github.com/Steve-Dal/Data-Science-IBM-Captstone/blob/main/edadataviz.ipynb

EDA with SQL

- SQL Queries performed on the SpaceX data set
  - Identifying distinct Launch Sites
  - Identifying Distinct Landing Outcomes
  - Select data from the table where the Launch site contains “CCA”
  - Calculated the Sum of Payload Masses in the dataset
  - Calculated the average payload mass in the dataset
  -Listed the date of the first successful drone ship landing
  -Listed the successful drone ship landings with a payload mass between 4000 and 6000 kg
  -Calculated the total amount of successful and failed landings
  -Listed the booster versions that have carried the maximum payload mass

https://github.com/Steve-Dal/Data-Science-IBM-Captstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Using folium, a map was created to display the different launch sites from the SpaceX data. Circles and labels were added to each launch site to see from a larger scale. Clusters were also added to each launch site depicting how many failed and successful landings took place. Lines were created to depict the distance from one of the launch sites to major infrastructure like highways, railways and cities.

https://github.com/Steve-Dal/Data-Science-IBM-Captstone/blob/main/lab_jupyter_launch_site_location.ipynb

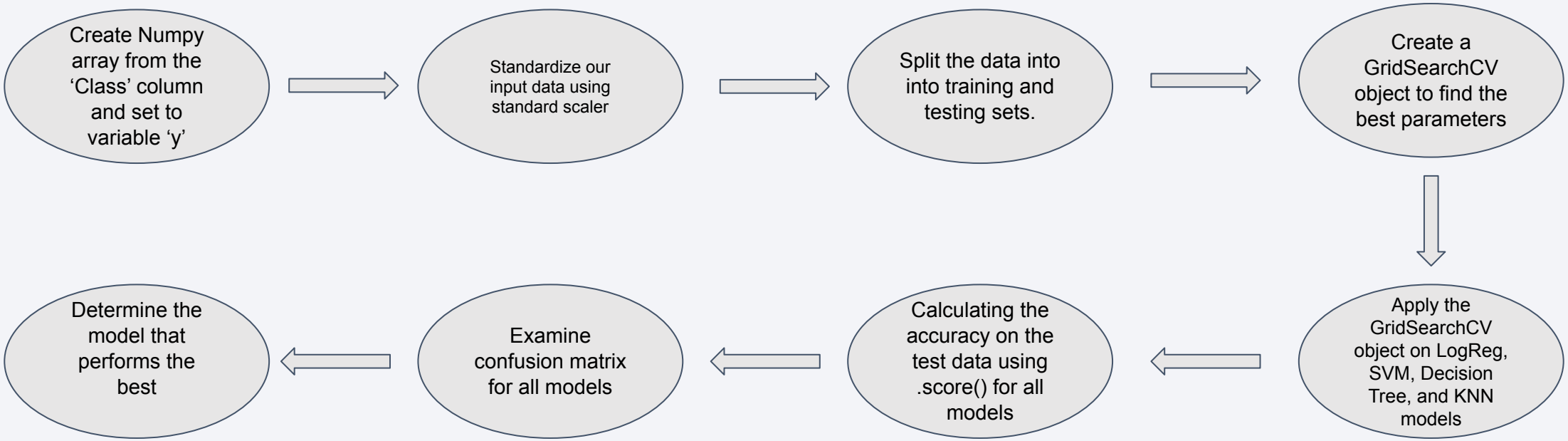
Build a Dashboard with Plotly Dash

- Our dashboard contains a drop down menu containing all of the launch sites in the SpaceX Data to include an option for all of the sites.
- When a site on the drop down is selected, it displays a pie chart displaying the percentage of failed/ successful landings.
- There is also a Payload Mass Range that allows the user to select a range for payload mass and display a scatter plot for landing outcome and booster version category.
- These particular plots were used to display the relationship between certain features and the landing outcome.

https://github.com/Steve-Dal/Data-Science-IBM-Captstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- To conduct predictive analysis on the data set, we standardized and split our data into training and test sets. We then found the best parameters for multiple classification models to include a logistic regression model, KNN, Decision Tree and SVM



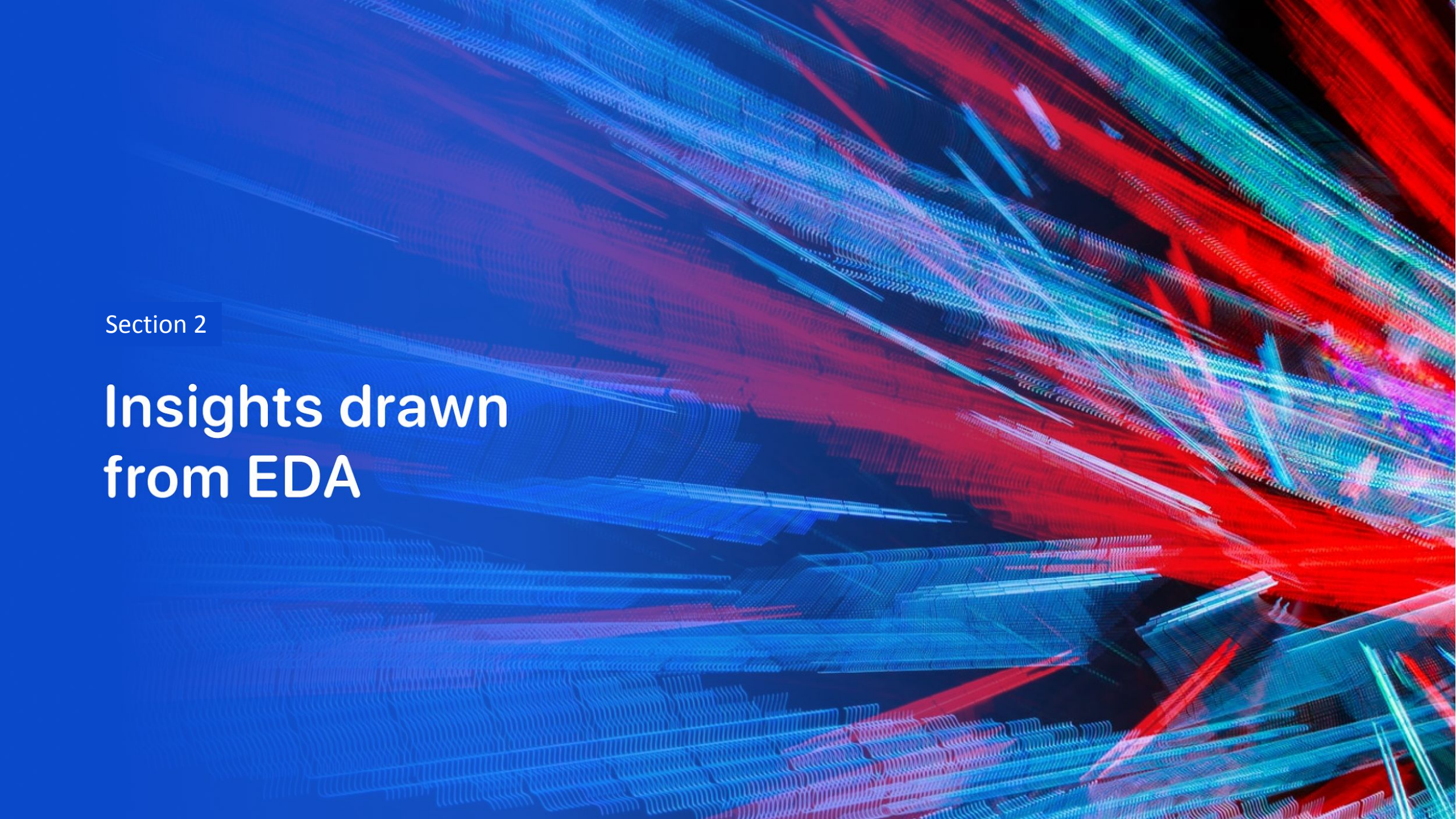
https://github.com/Steve-Dal/Data-Science-IBM-Captstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

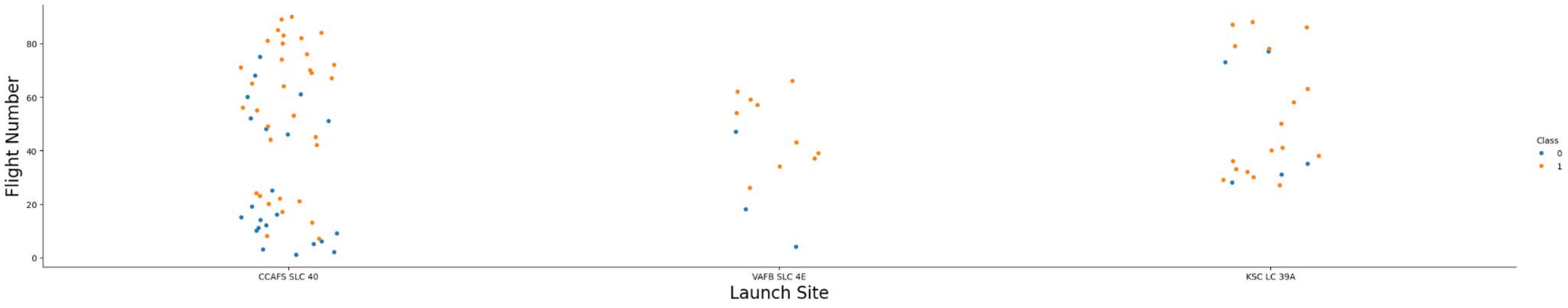
Section 2

Insights drawn from EDA



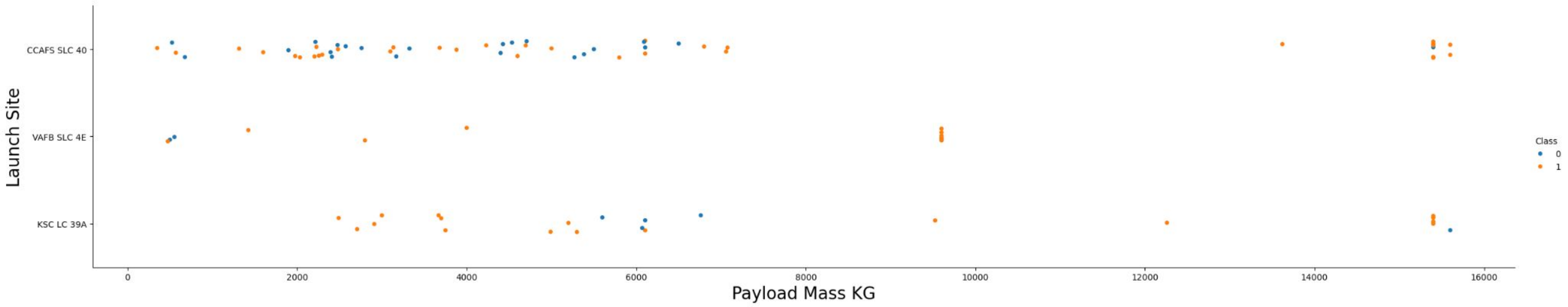
Flight Number vs. Launch Site

- This plot displays the relationship between flight number, launch site and landing outcome.
- We can see that at each launch site the success rate increases as the amount of flights increases.
- The CCAFS SLC 40 launch site has most of the launches amongst the rest
- The highest success rates are at VAFB SLC 4E and KSC LC 39A launch sites



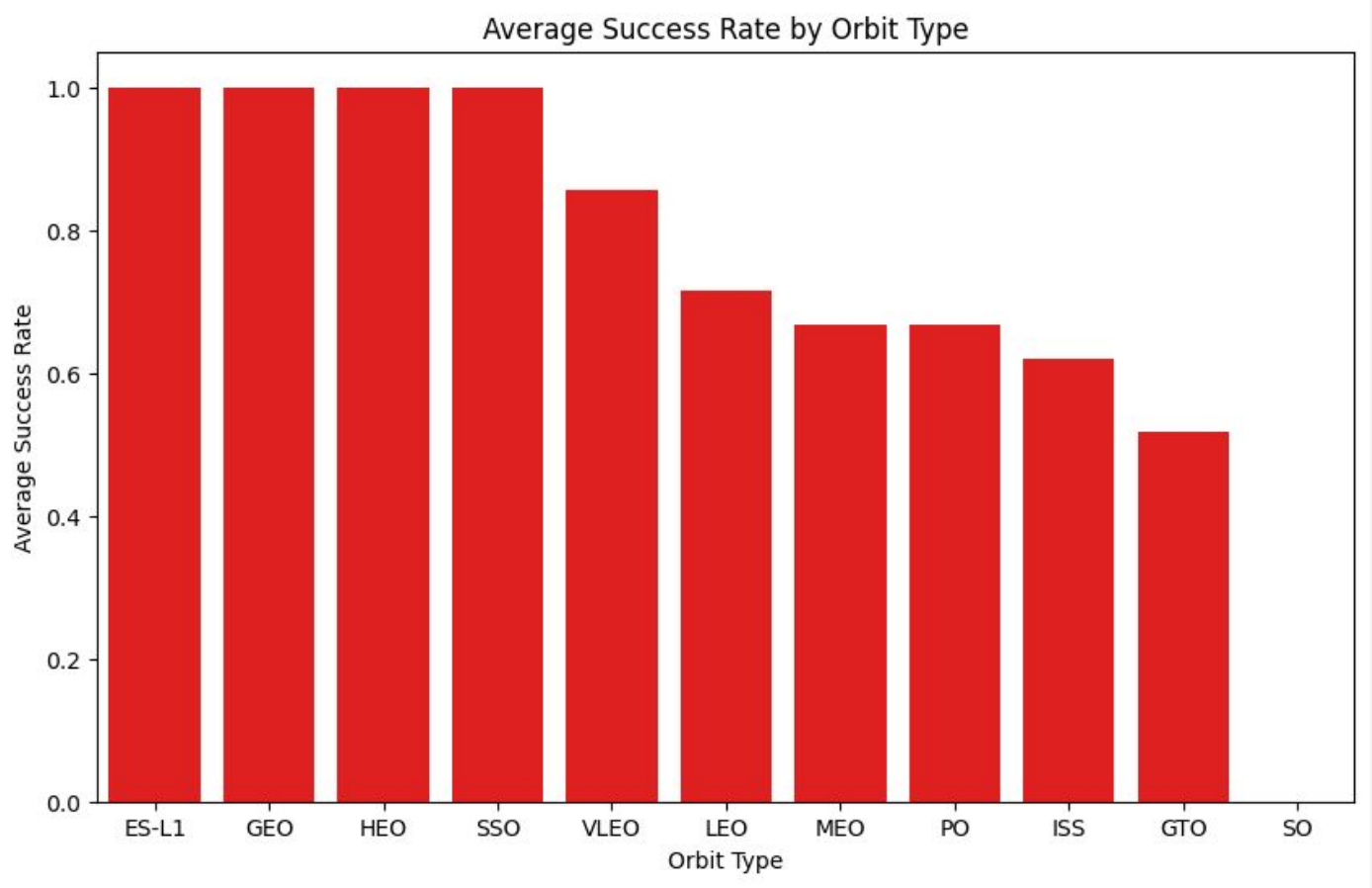
Payload vs. Launch Site

- This plot displays the relationship between Payload Mass, launch site and landing outcome.
- We can immediately determine that the success rate of landings increases with the payload mass .
- We can also see that the Launch Site VAFB SLC 4E has no launches with a payload mass more than approximately 10,000 kg.
- The Launch Site KSC LC 39A seemed to have a 100 percent success rate until they reached payload masses between 5000 and 7000 kg.



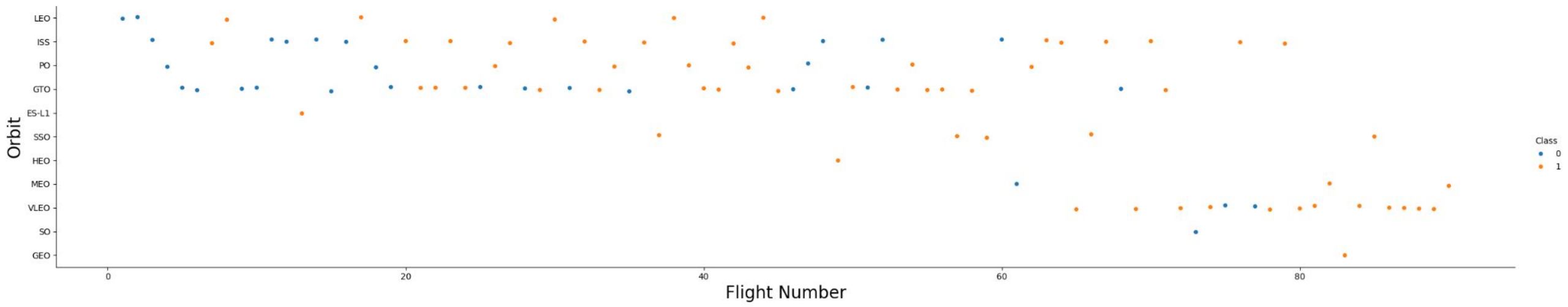
Success Rate vs. Orbit Type

- This bar chart displays the average success rate by orbit type.
- This chart shows that the highest success rates come from launches that enter the ES-L1, GEO, HEO and SSO orbits.
- The success rates for launches that enter SO orbits have a average success rate of 0.



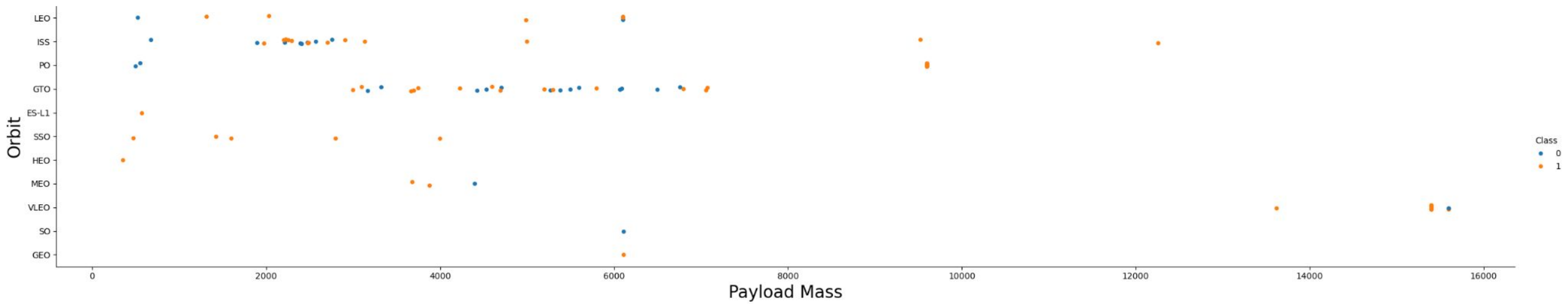
Flight Number vs. Orbit Type

- This plot displays the relationship between flight number, orbit type and landing outcome.
- We can see that the GEO, SO, VLEO and MEO orbits weren't attempted until about flight number 50.
- The GTO, PO, ISS and LEO orbits were launched consistently throughout the dataset.



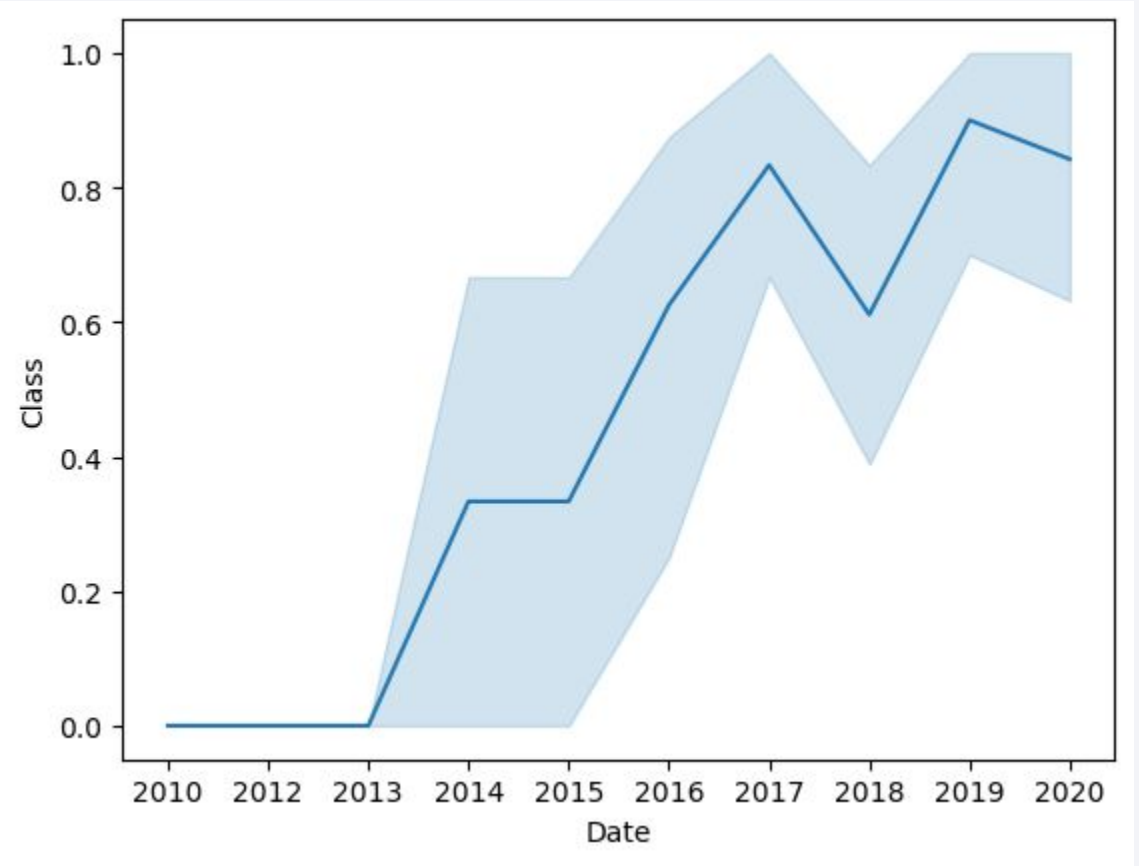
Payload vs. Orbit Type

- This plot displays the relationship between payload mass, orbit type and landing outcome.
- This plot shows that GTO orbits did not have a payload mass exceeding 8000 kg.
- Success rates increase as payload mass increases for GTO and ISS orbits.



Launch Success Yearly Trend

- The line plot displays a steady increase in success until 2020



All Launch Site Names

- Query to display unique launch site names

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

[11]:
```
%%sql
select distinct Launch_Site from SPACEXTBL;
```

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

[11]:

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

Launch Site Names Begin with 'CCA'

- Query to see all data where the launch site contains “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_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parac
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 (parac
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No att
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No att
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

- Query to calculate the total payload mass for NASA launched boosters

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

[33]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass
FROM SPACEXTABLE
WHERE customer LIKE 'NASA (CRS)%';
```

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

[33]:

total_payload_mass
48213

Average Payload Mass by F9 v1.1

- Query to calculate the average payload mass carried by booster version F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

[35]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS total_payload_mass_1
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

[35]:

total_payload_mass_1
2534.6666666666665

First Successful Ground Landing Date

- Query to determine the date of the first successful ground pad landing was achieved

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

[13]:
```
%%sql
select min(Date) from SPACEXTABLE where Landing_Outcome == "Success (ground pad)"
```

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

[13]:

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query to list the boosters that successfully landed with payload masses between 4000 and 6000 kg

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

[39]:
```
%%sql
select Booster_Version from SPACEXTABLE
where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
and Landing_Outcome == "Success (drone ship)"
```

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

[39]:

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

Total Number of Successful and Failure Mission Outcomes

- Query to calculate the total successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

[31]:
```
%%sql
select Mission_Outcome, Count(*) as total_count
from SPACEXTABLE
group by Mission_Outcome
```

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

[31]:

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

Boosters Carried Maximum Payload

- Query to show the boosters that carried the maximum payload

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

[18]:
```
%%sql
select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

[18]:

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

- Query that displays the failed drone ship landings in 2015

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[35]:
```
%%sql
select
    substr(Date, 6, 2) as month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
from SPACEXTABLE
where Landing_Outcome like 'Failure (drone ship)%'
  and substr(Date, 1, 4) = '2015';
```

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

[35]:

month	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

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

- Query that displays the outcomes and counts by type between June 4th 2010 and March 20th 2017

[50]:
```
%%sql
select Landing_Outcome, count(*) as outcome_count
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by outcome_count desc
```

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

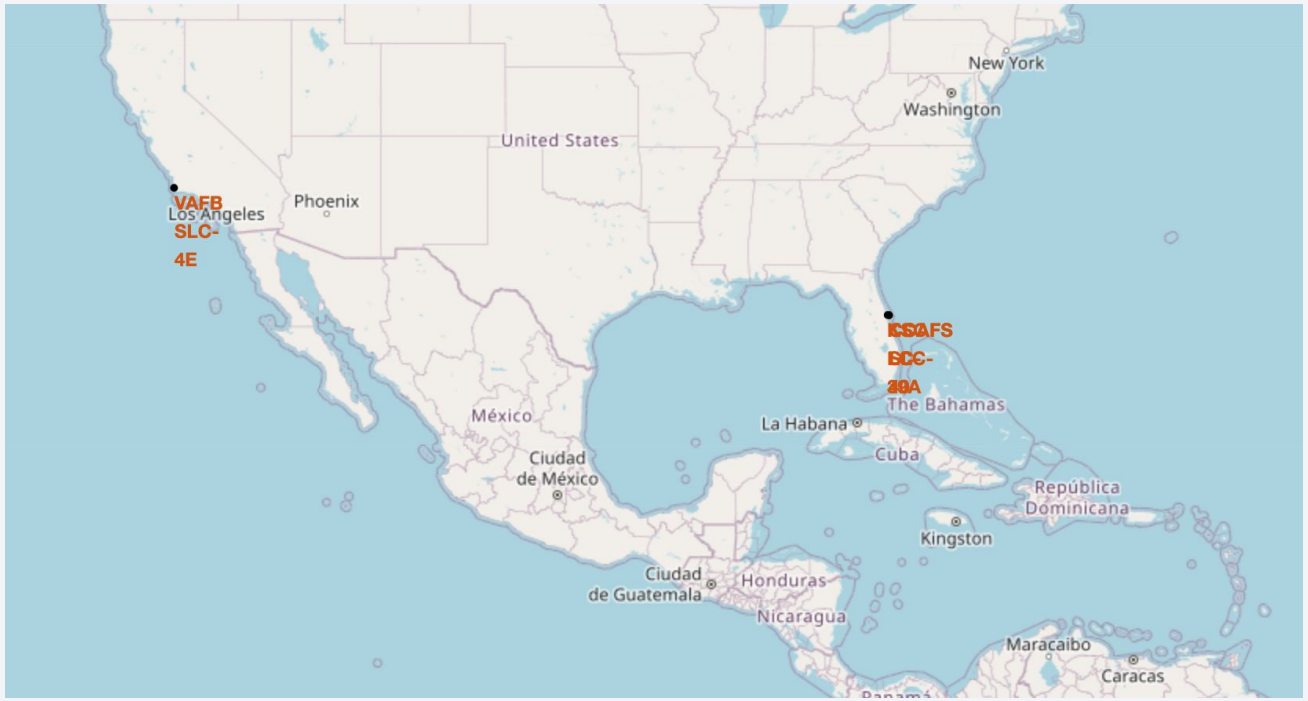
[50]:

Landing_Outcome	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

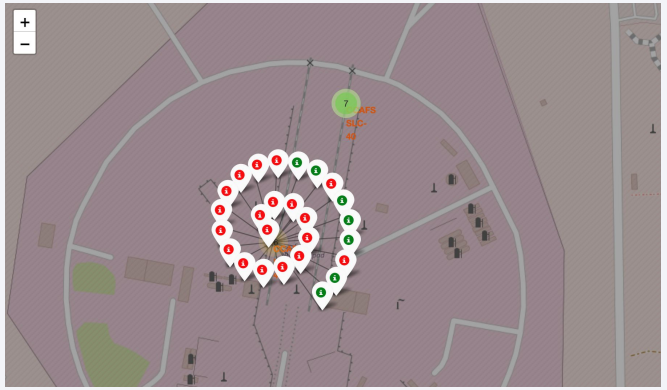
Launch Sites Map

• Launch sites are in the southern portions of the United States along the east and west coasts. The positions of these launch sites are optimal for launching ships into orbit.



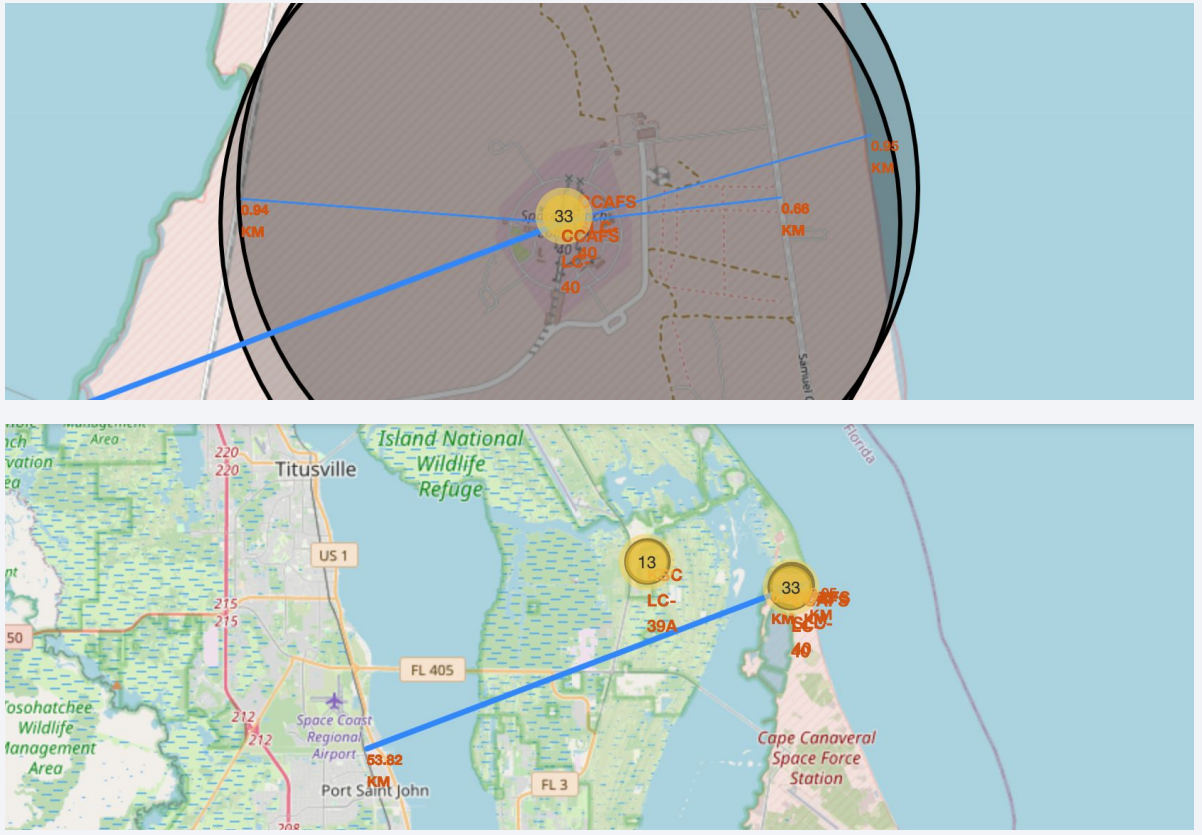
Success and Failure Clusters

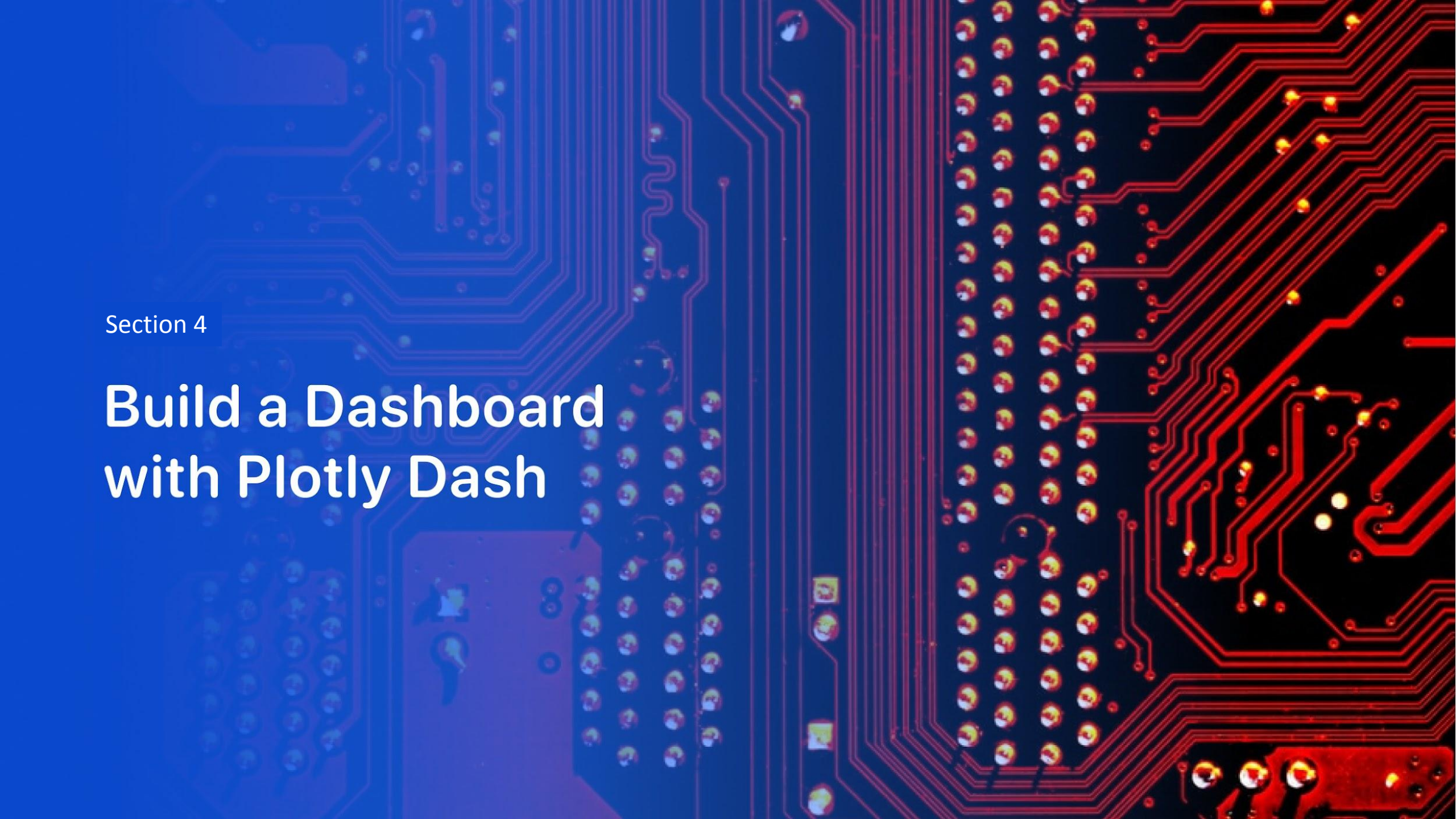
- Using folium we created clusters displayed the amount of successful and failed landings.
- Red marker = Failed Launch
- Green Marker = Successful Launch
- We can see that KSC LC-39A has a high success rate.



Launch Sites Distances

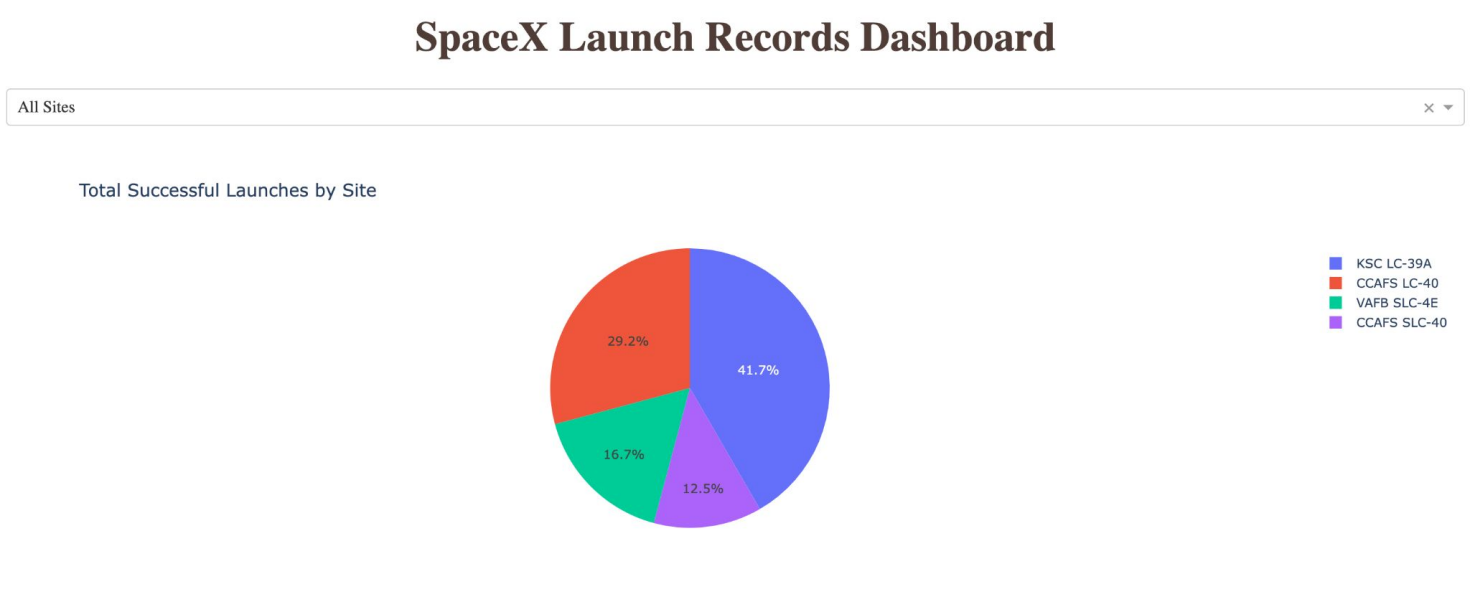
- These maps display the distances of coast line, highways railways and cities to the launch site.
- Launch sites are situated in close proximity to highways, railways and coastline but maintain distance from population centers.




Section 4
Build a Dashboard with Plotly Dash

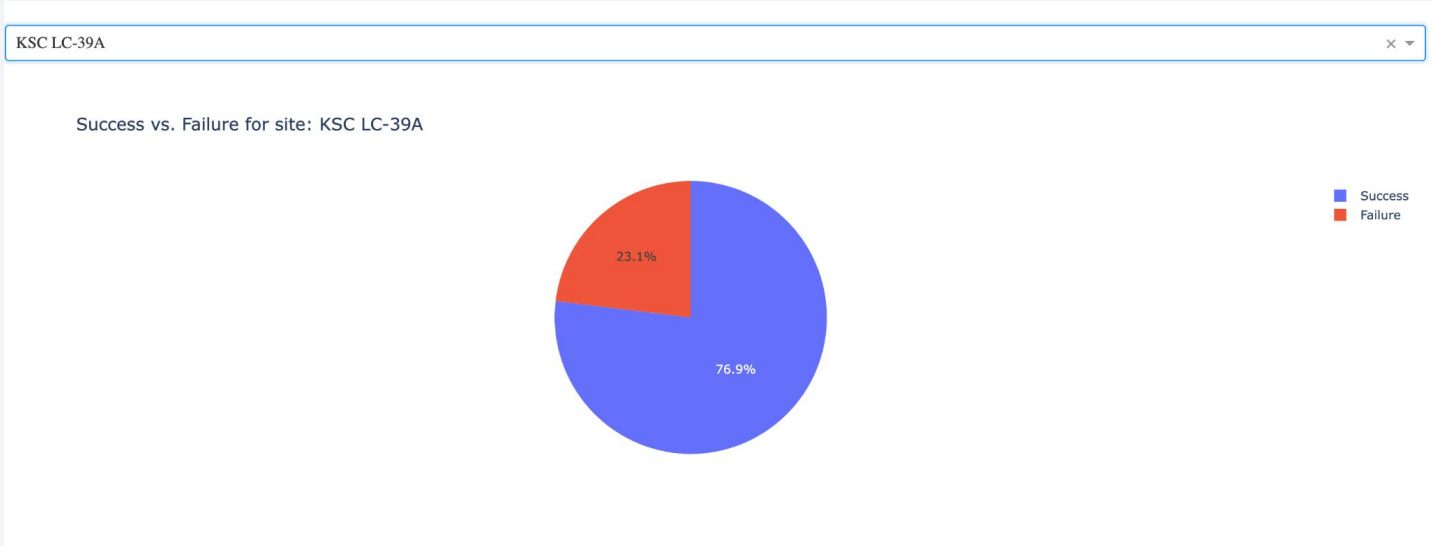
Successful Launches by Site

- The first section of this dashboard displays an interactive pie chart that shows the successful launches by site. We can see that KSC LC-39A accounts for 41% of successful outcomes.



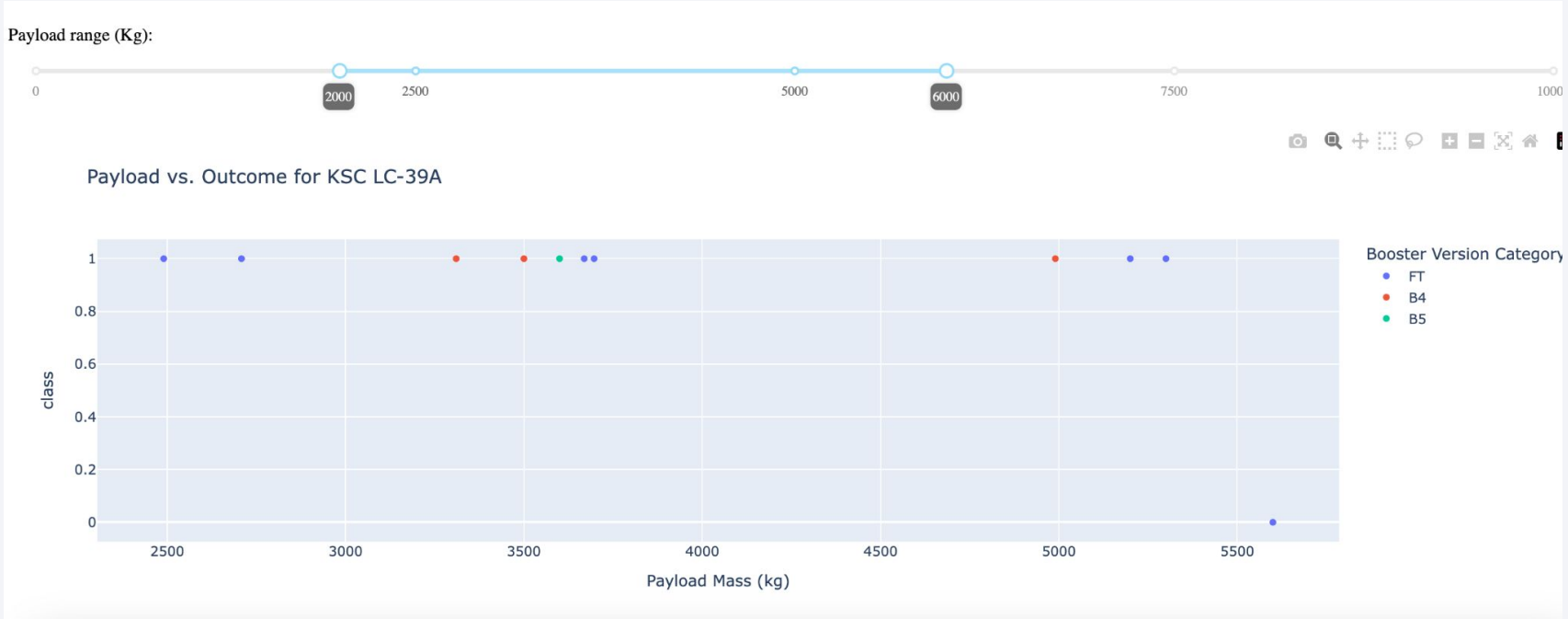
<Dashboard Screenshot 2>

- The launch site that accounts for most of the successful launches in the dataset has an overall success rate of 76.9%.



<Dashboard Screenshot 3>

- Chart shows that payloads between 2000 and 5500 kg have the highest success rates.



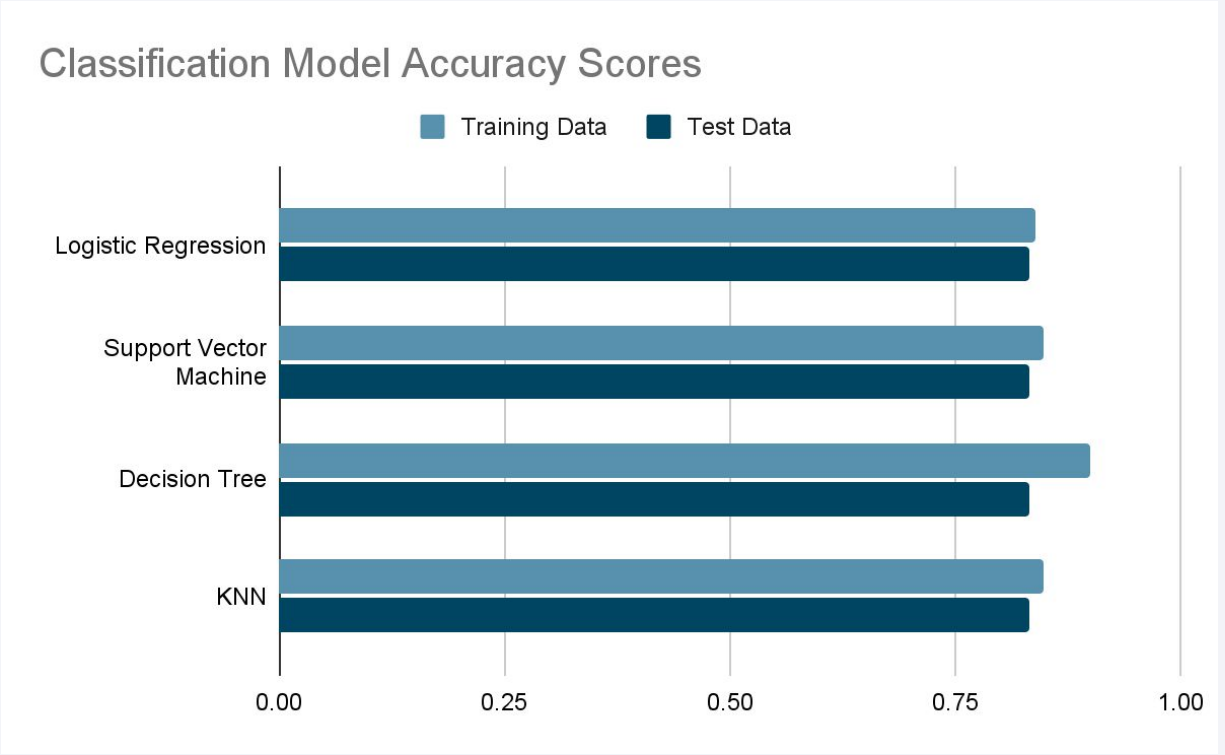
Section 5

Predictive Analysis (Classification)

Classification Accuracy

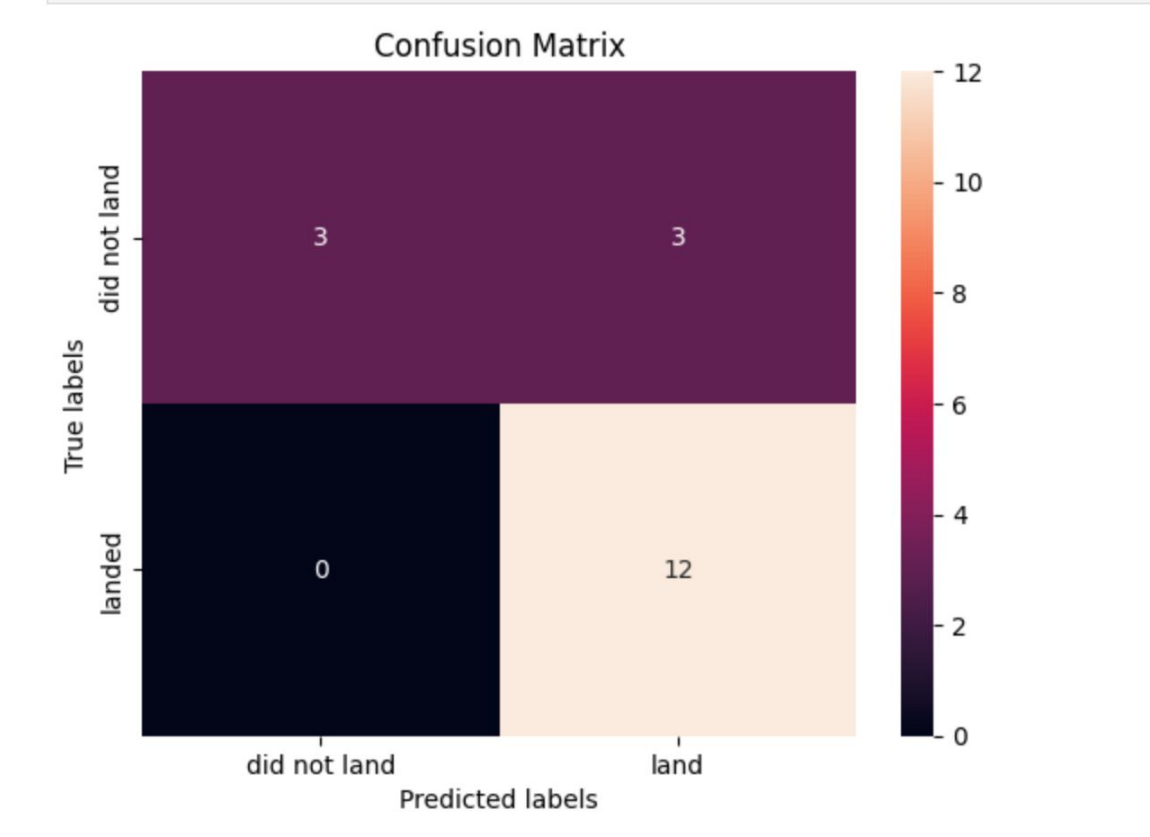
- After training and testing the models we found that all 4 classification models had the same accuracy outcome. This could be due to the small sample size.

- The decision tree model performed the best on the training data despite having the same outcome as the others when applied to the testing data.



Confusion Matrix

- The confusion matrix shows that the decision tree model had 12 true positives and 3 false positives.



Conclusions

- All of the classification models had similar accuracy scores on the test data but the decision tree model had the highest accuracy on the training data.
- KSC LC-39A has the highest success rate amongst the rest of the launch sites
- Launch sites typically maintain a close proximity to the equator and distance themselves from population centers
- Launches that intend to enter ES-L1, GEO, HEO and SSO orbits have a 100% success rate.

Appendix

- Github link that contains all code used to derive these insights
  https://github.com/Steve-Dal/Data-Science-IBM-Captstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

Thank you!