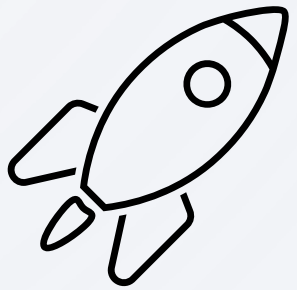
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hikaru M
November 13, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion



Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory data analysis with SQL
  - Exploratory data analysis with data visualization
  - Interactive visual analysis with Folium
  - Interactive dashboard with Plotly Dash
  - Predictive analysis with machine learning (classification)
- Summary of all results
  - Exploratory data analysis with SQL and data visualization
  - Interactive analysis with Folium and Plotly Dash
  - Predictive analysis with classification

Introduction

- Project background and context
  - SpaceX is among the most successful company in commercial space industry, making commercial space travel more affordable. One of the reason of Space X's success was able to lower the cost of the rocket launches to 62 million dollars (compared to other competitors at 165 million dollars) is because SpaceX can recover and reuse the first stage. If we could determine if the first stage will successfully land, we can accurately determine the cost of a launch.
- Problems you want to find answers
  - Predict if the first stage of the SpaceX Falcon 9 rocket will land successfully.
  - What variables contribute to successful launches?
  - What is the best machine learning model for this prediction?

Section 1

Methodology

Methodology

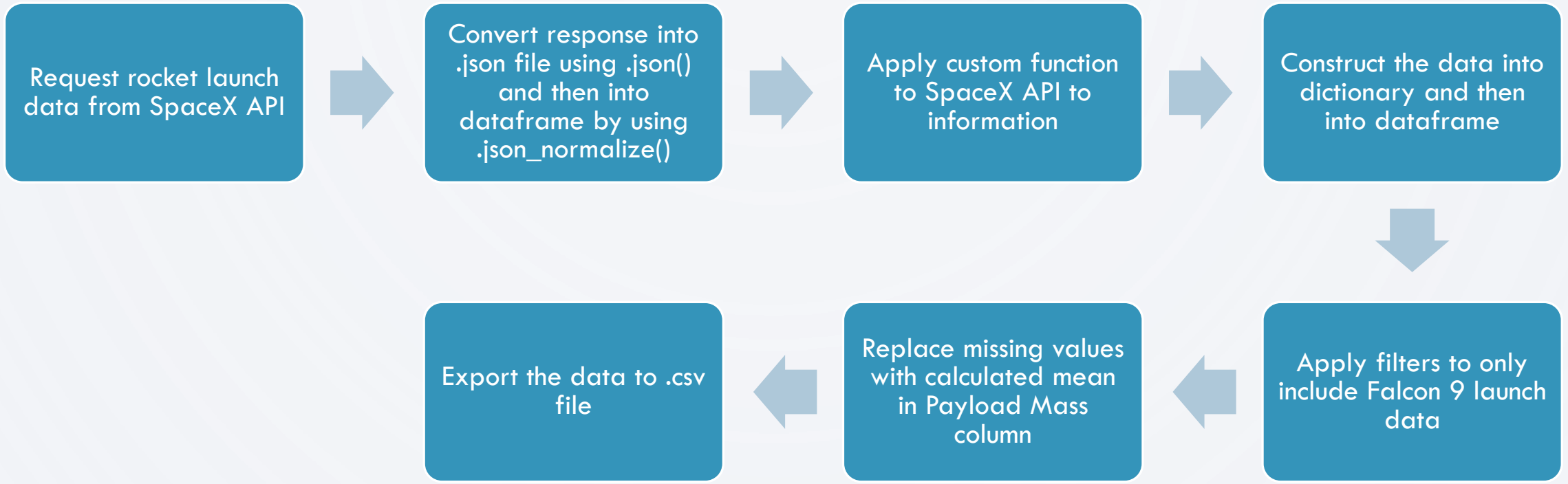
Executive Summary

- Data collection methodology:
  - SpaceX rest API
  - Web scraping from Wikipedia
- Perform data wrangling
  - One hot encoding to prepare the data (dealing with null values and irrelevant columns) for further analysis with classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, and evaluate four classification models (Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbor

Data Collection

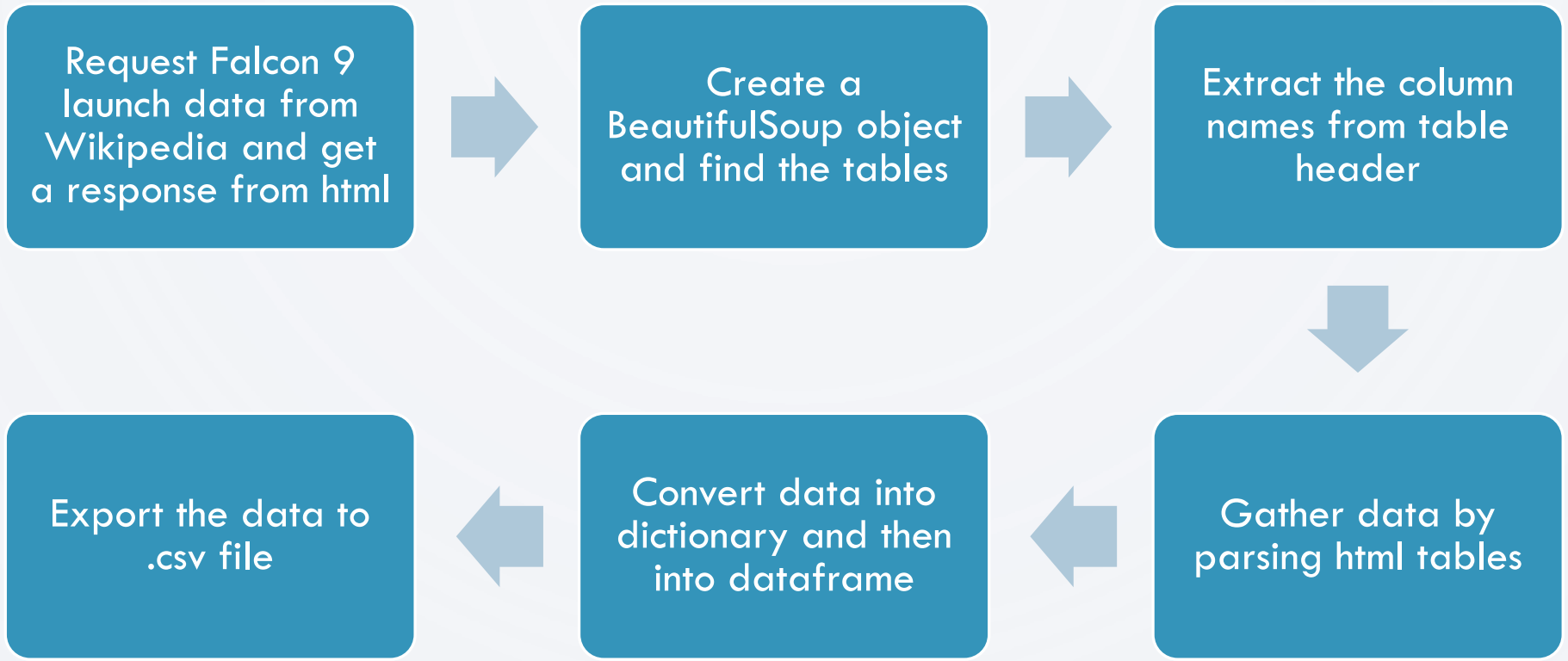
- Rocket launch data was collected using two methods to gather complete dataset for analysis:
  - API requests from SpaceX REST SPI
  - Web scraping data from table in SpaceX's Wikipedia page "List of Falcon 9 and Falcon Heavy launches"
- Following slides will explain the data collection process using flowcharts

Data Collection – SpaceX API



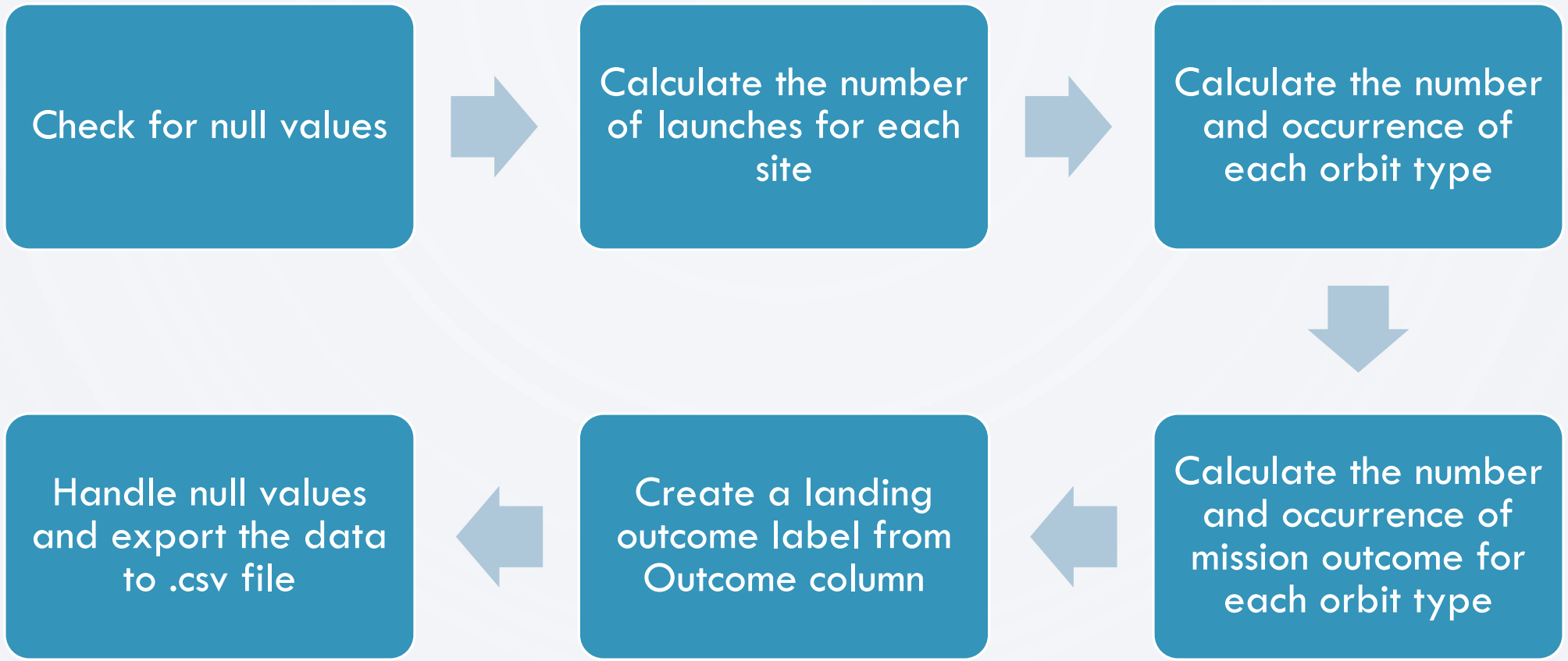
https://github.com/hikarum/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping



https://github.com/hikarum/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling



https://github.com/hikarum/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- Scatter plots show the relationship between two variables
  - Charts plotted → Flight number vs. Launch Site, Payload mass vs. Launch site, Flight number vs. Orbit type, Payload mass vs. Orbit type
- Bar charts show relationship between two discrete categorical variables
  - Chart plotted → Success rate vs. Orbit type
- Line charts show a trend in data over time period
  - Chart plotted → Launch success yearly trend

https://github.com/hikarum/Applied-Data-Science-Capstone/blob/main/EDA%20for%20Data%20Visualization.ipynb

EDA with SQL

- Following SQL queries were performed
  - Display the names of the unique launch sited in space mission
  - Display 5 records where launch sites begin with the string “CCA”
  - Display the total payload mass carried by the boosters launched by NASA (CRS)
  - Display the average payload mass carried by the booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drop ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Rank the count of landing outcomes between the date 04-06-2010 to 20-03-2017 in descending order

https://github.com/hikarum/Applied-Data-Science-Capstone/blob/main/Exploratory%20Data%20Analysis%20using%20SQL.ipynb

Build an Interactive Map with Folium

- Markers for launch sites
  - Marker with circle, pop-up label and text label were added using latitude and longitude coordinates to show the geographical location
- Coloured markers of the launch outcomes for each launch site
  - Coloured markers (success = green, failure = red) were added using marker cluster to visualize which site had high success rate
- Distance between each launch site to its proximities
  - Coloured lines were added to show the distances between the launch site to its proximities such as railway, highway, coastline and closest city

https://github.com/hikarum/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

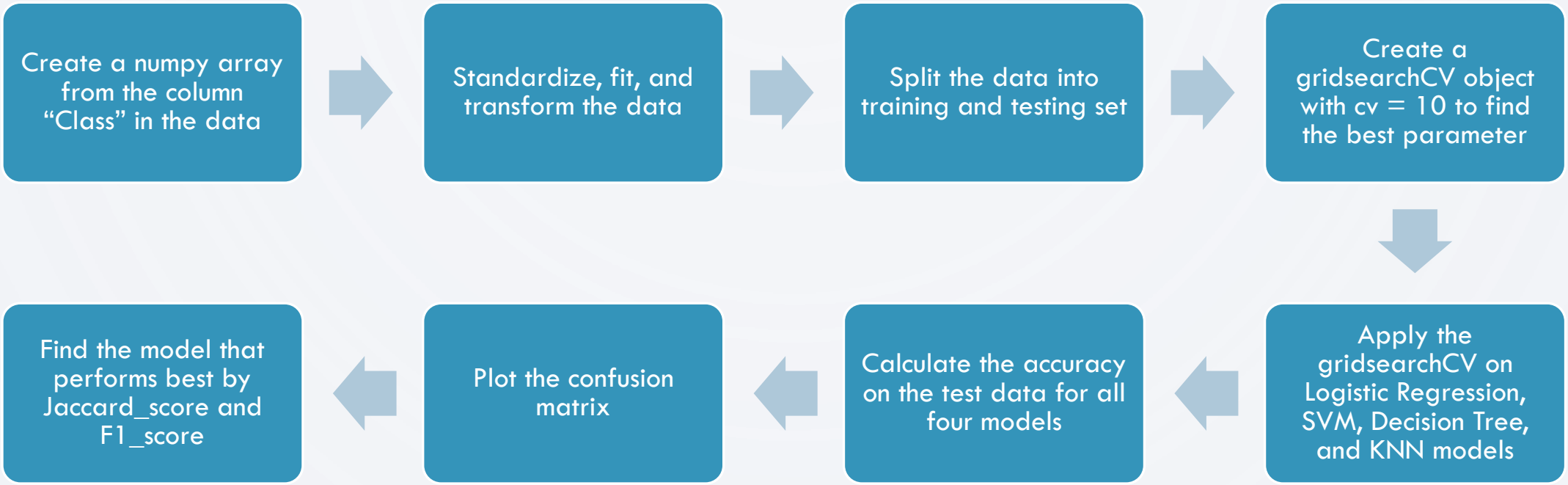
Folium maps will not load in GitHub, so the screenshots are uploaded instead. Please see folder named “Folium Map Screenshots”!

Build a Dashboard with Plotly Dash

- Launch sites dropdown list
  - Dropdown list was added to enable launch site selection, including each site and “All sites” selection
- Pie chart showing successful launches for all sites and selected site
  - Pie chart was added to showcase the total count of successful launches for all sites and success ratio for each site
- Slider of payload mass range
  - Slider was added to enable payload range selection
- Scatter chart of payload mass vs. success rate for different booster versions
  - Scatter char was added to visually showcase the correlation between payload mass and launch success on booster versions

https://github.com/hikarum/Applied-Data-Science-Capstone/blob/main/Interactive%20Dashboard%20with%20Plotly%20Dash.ipynb

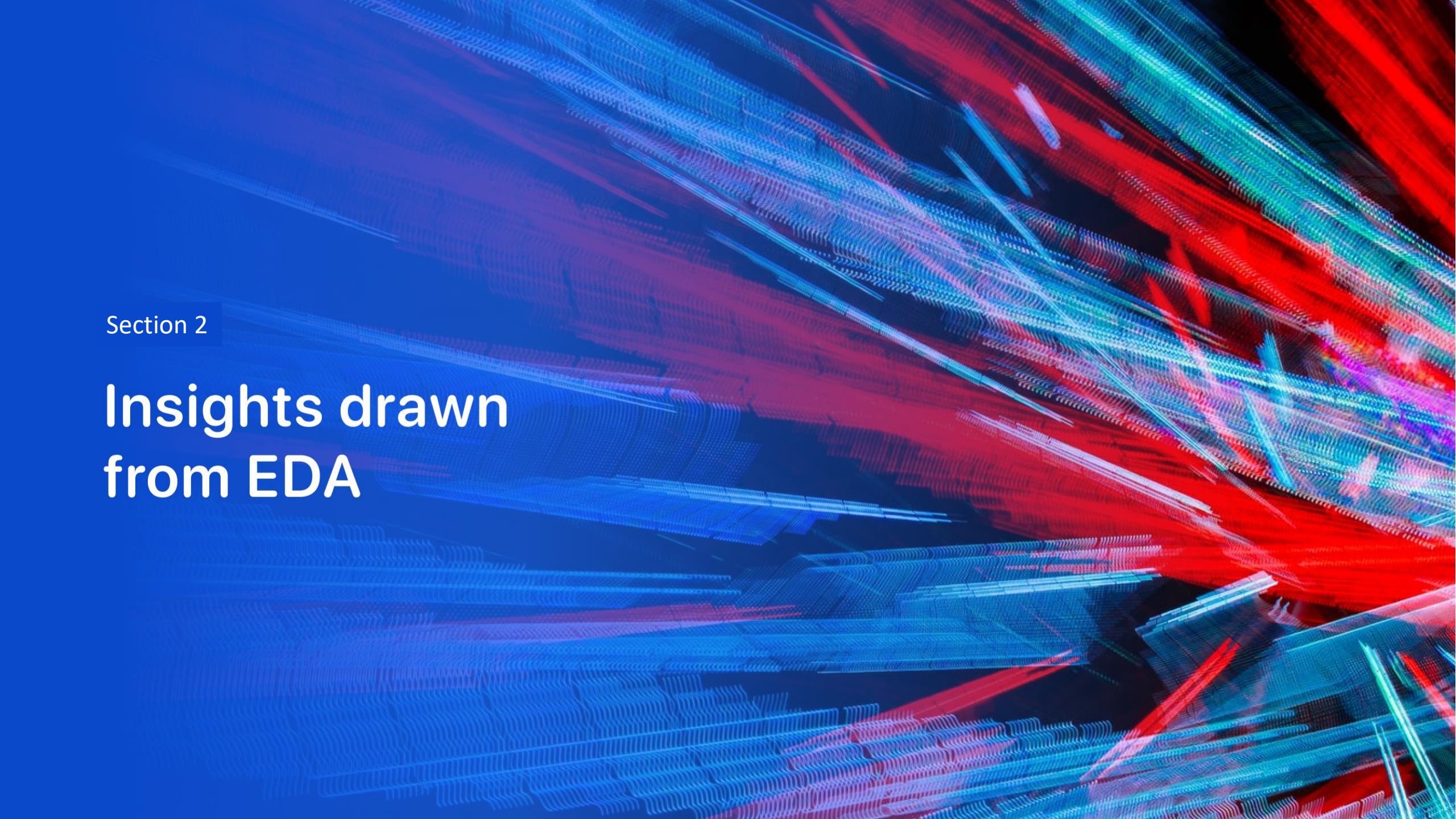
Predictive Analysis (Classification)



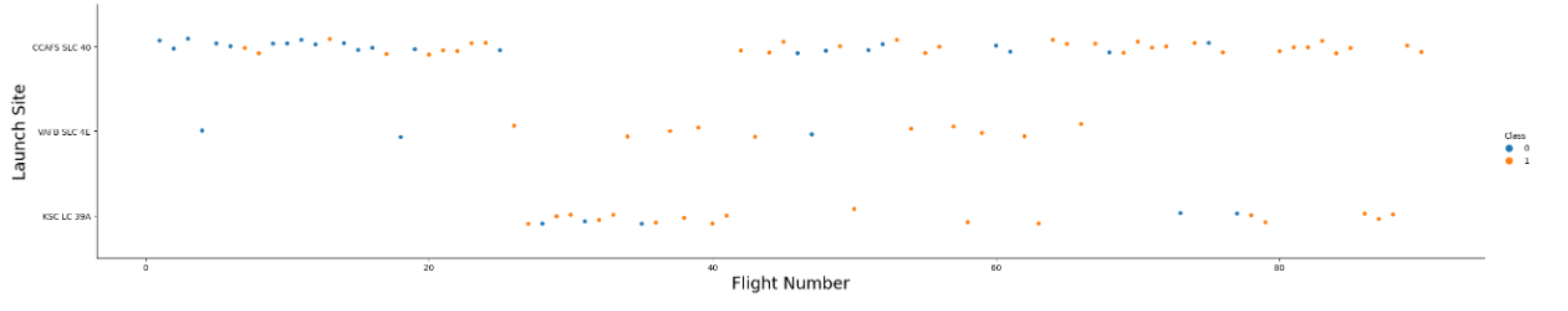
https://github.com/hikarum/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results

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

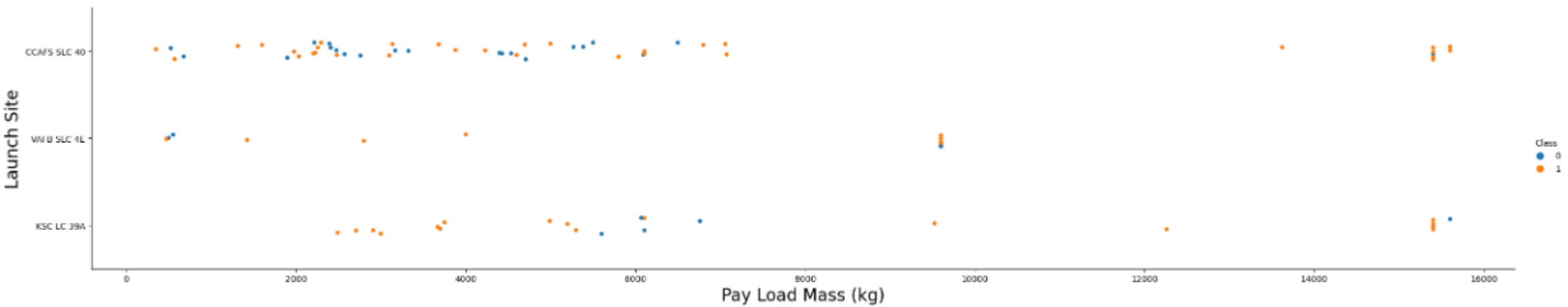

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Earlier launches have high failure count (blue dots) and latest launches have high success count (orange dots)
- CCAFS SLC 40 launch site have approx. 50/50 success rates
- VAFB SLC 4E and KSC LC 39A sites have higher success rates

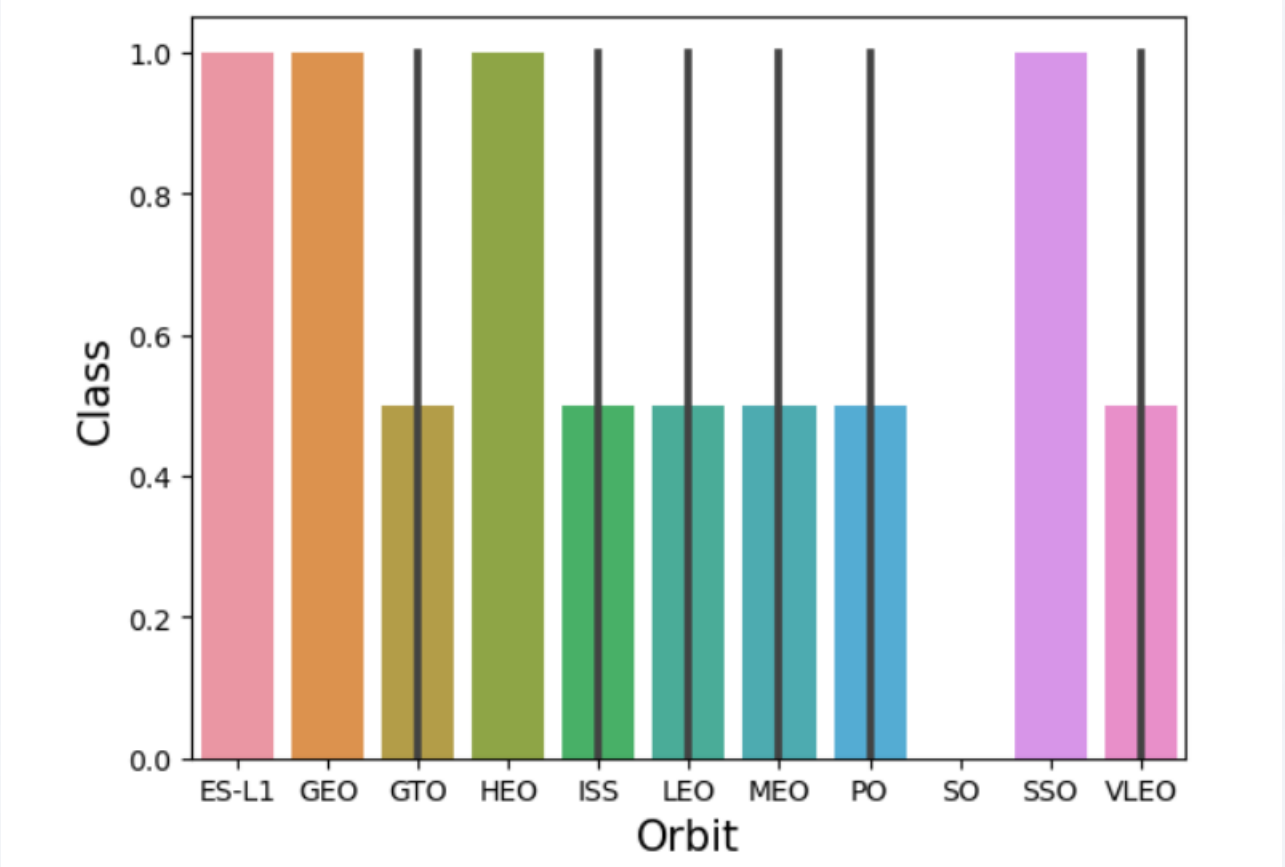
Payload vs. Launch Site



- VAFB SLC 4E site has not had a launch of payload mass more than 10000kg
- Most of rocket launches have the payload mass less than 10000kg

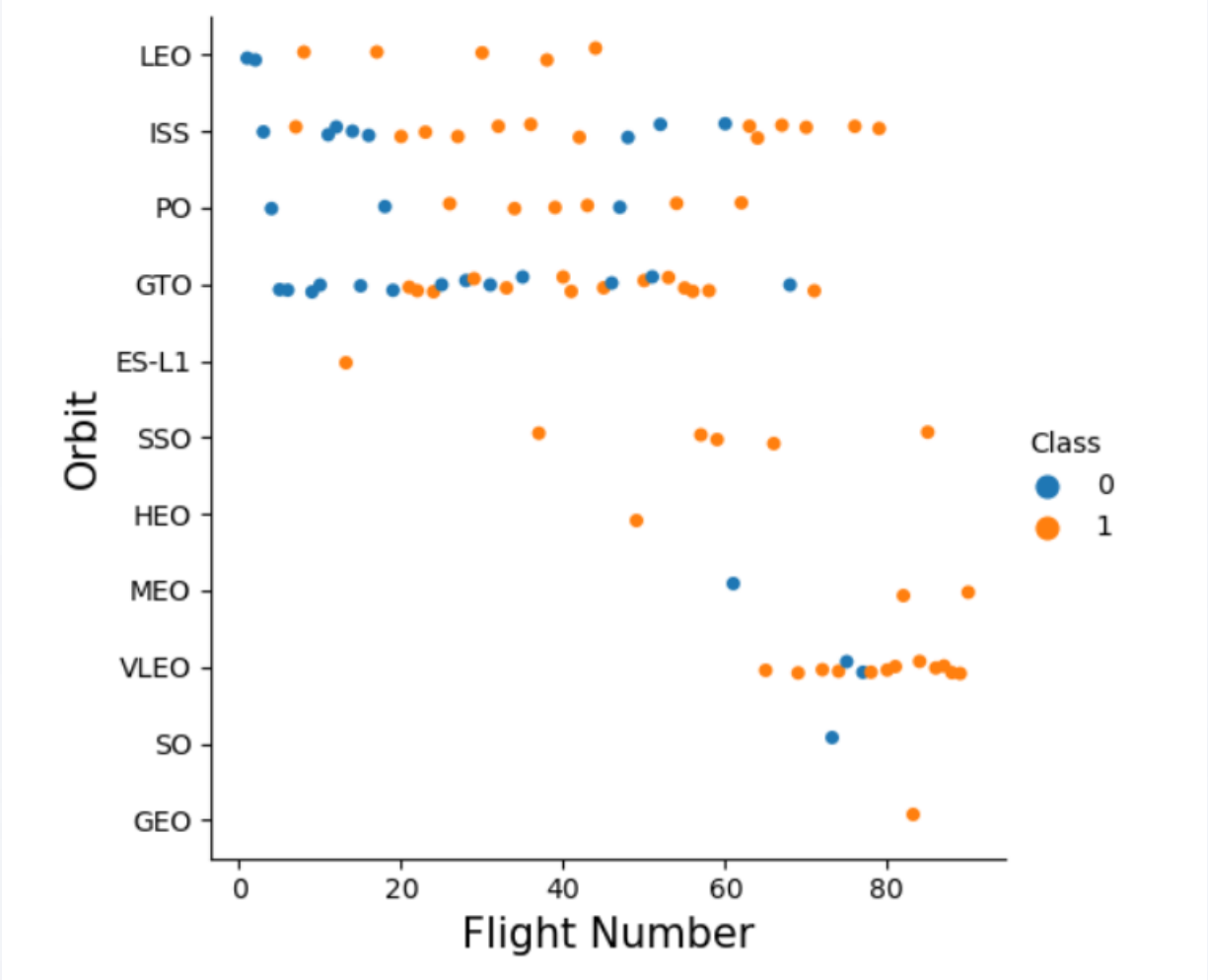
Success Rate vs. Orbit Type

- Orbit type with 100% success rate
  - ES-L1, GEO, HEO, SSO
- Orbit type with 50% success rate
  - GTO, ISS, LEO, MEO, PO, VLEO
- Orbit type with 0% success rate
  - SO



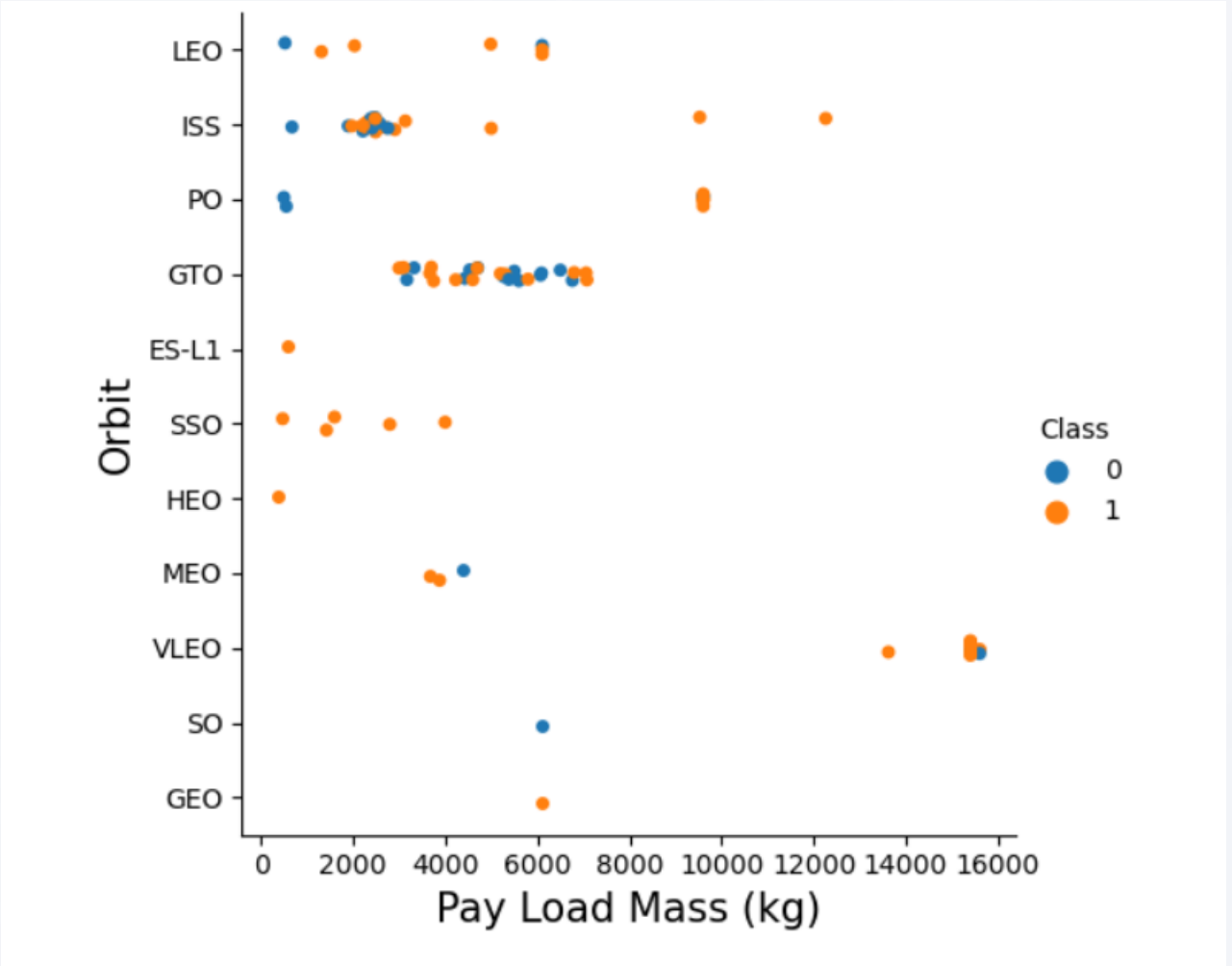
Flight Number vs. Orbit Type

- LEO orbit seems to be related to the number of flights
- There seems to be no relationship between GTO orbit and number of flights



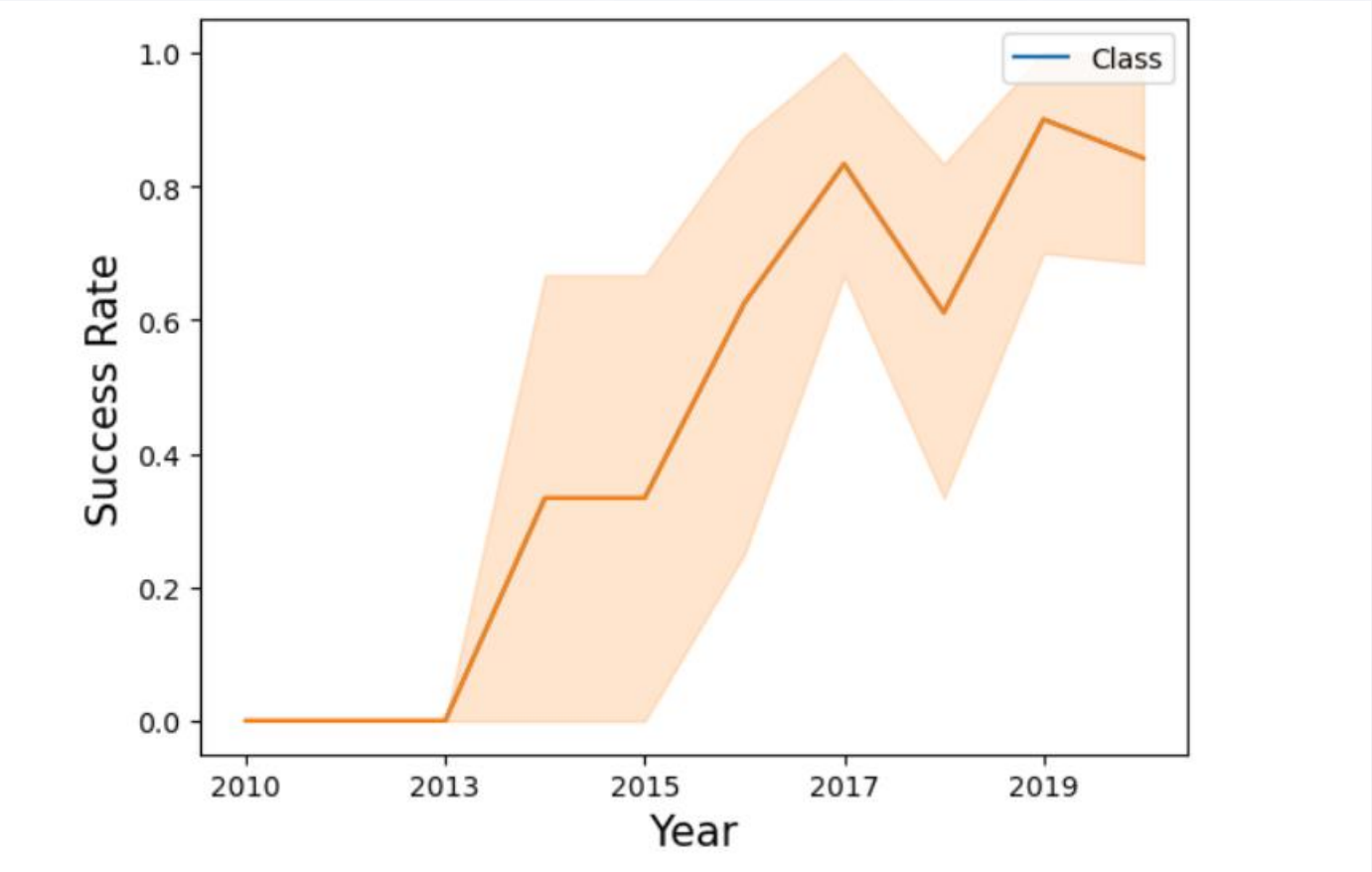
Payload vs. Orbit Type

- Heavy payload have a negative effect on GTO orbit type
- Heavy payload have a positive effect on PO, LEO, and ISS orbit types



Launch Success Yearly Trend

- Launch success rate has increased since 2013 until 2020, likely due to technology development and more data being gathered from previous launches



All Launch Site Names

```
In [6]: %sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[6]:

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

- Above query was executed to display the names of the unique launch sites in the space mission. 4 launch site names were returned.

Launch Site Names Begin with 'CCA'

```
In [56]: %sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
 * sqlite:///my_data1.db
Done.
```

Out[56]:

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

- Above query was executed to display 5 records where launch sites begin with `CCA`

Total Payload Mass

```
In [57]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS PAYLOADMASS FROM SPACEXTBL WHERE (Customer) LIKE 'NASA (CRS)';
          * sqlite:///my_data1.db
         Done.
```

Out[57]:

PAYLOADMASS
45596

- Above query was executed to display the total payload carried by boosters launched by NASA (CRS).
- Total payload mass was 45596kg.

Average Payload Mass by F9 v1.1

```
In [58]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS PAYLOADMASS FROM SPACEXTBL WHERE (Booster_Version) LIKE 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

Out[58]:

PAYLOADMASS
2928.4

- Above query was executed to display the average payload mass carried by booster version F9 v1.1.
- Average payload mass was 2928.4kg

First Successful Ground Landing Date

```
In [23]: %sql SELECT MIN("DATE") FROM SPACEXTBL WHERE [LANDING _OUTCOME] LIKE 'Success (ground pad)'
          * sqlite:///my_data1.db
         Done.
```

Out[23]:

MIN(DATE)
01-05-2017

- Above query was executed to display the dates of the first successful landing outcome on ground pad.
- First successful landing was 01-05-2017.

Successful Drone Ship Landing with Payload between 4000kg and 6000kg

In [29]:
```
%sql select BOOSTER_VERSION from SPACEXTBL\
where [LANDING _OUTCOME] ='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

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

Out[29]:

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

- Above query was executed to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000kg but less than 6000kg.
- 4 booster versions were returned as result.

Total Number of Successful and Failure Mission Outcomes

```
In [8]: %sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTBL GROUP BY Mission_Outcome;
 * sqlite:///my_data1.db
Done.
```

Out[8]:

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

- Above query was executed to display the total number of successful and failure mission outcomes.
- There was 1 in-flight failure, 99 success, ad 1 success where payload status was unclear.

Boosters Carried Maximum Payload

```
In [9]: %sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
        WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

Out[9]:

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

- Query was executed to list the names of the booster which have carried the maximum payload mass.
- List of 12 booster versions were returned as a result.

2015 Launch Records

In [26]:
```
%sql SELECT substr("DATE", 4, 2) AS MONTH, [Landing _Outcome], "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE [LANDING _OUTCOME] = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

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

Out[26]:

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

- Above query was executed to list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- There was one launch in January, and one launch in April that matches to the condition specified.

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

In [28]:
```
%sql SELECT [LANDING _OUTCOME], COUNT([LANDING _OUTCOME]) FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and [LANDING _OUTCOME] LIKE '%Success%'\
GROUP BY [LANDING _OUTCOME] \
ORDER BY COUNT([LANDING _OUTCOME]) DESC ;
```

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

Out[28]:

Landing _Outcome	COUNT([LANDING _OUTCOME])
Success	20
Success (drone ship)	8
Success (ground pad)	6

- Above query was executed to rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.
- There was 20 successful landings, 8 successful drone ship landings, and 6 successful ground pad landings.

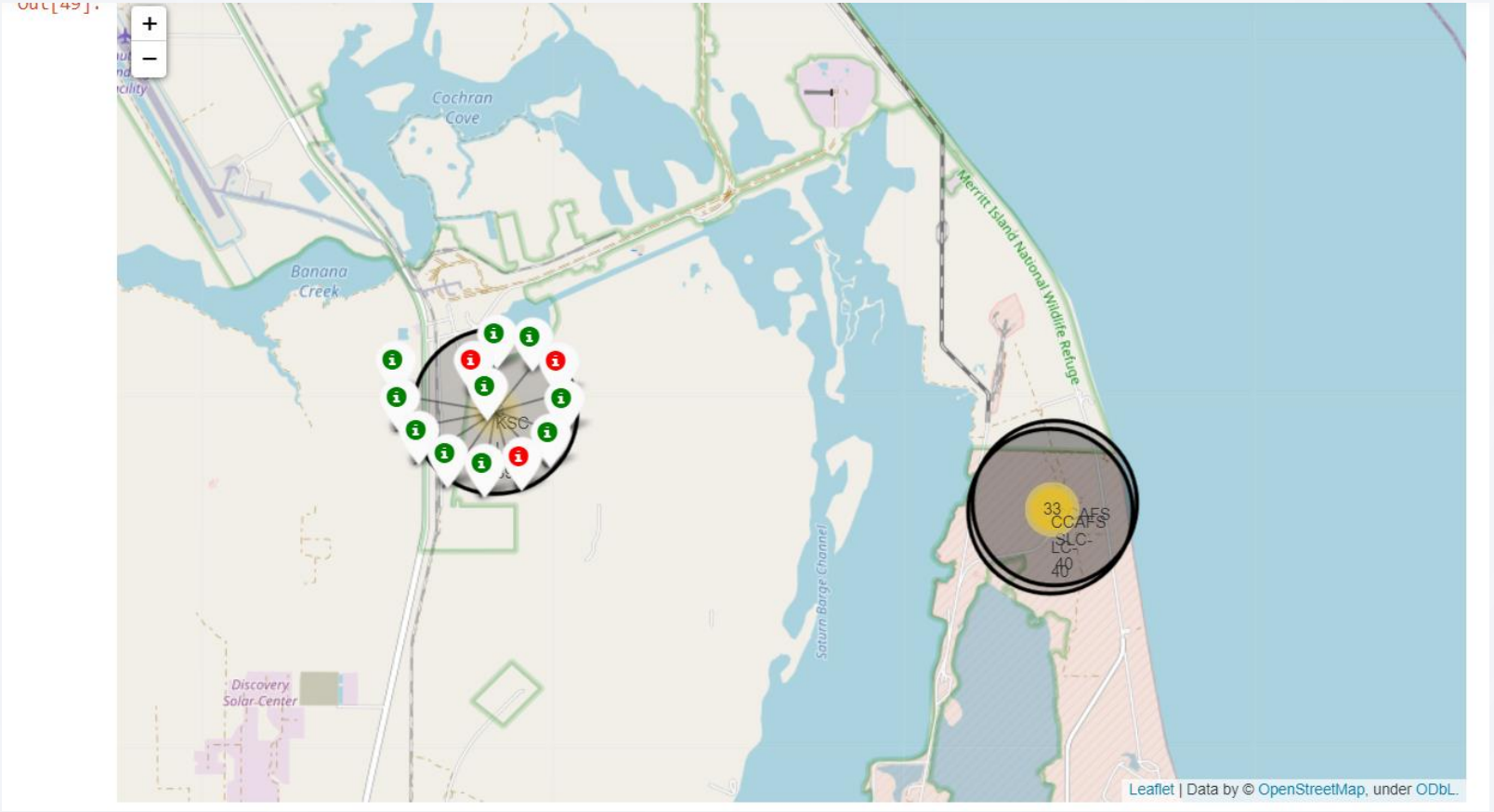
Section 3

Launch Sites Proximities Analysis

Map with All Launch Sites Marked

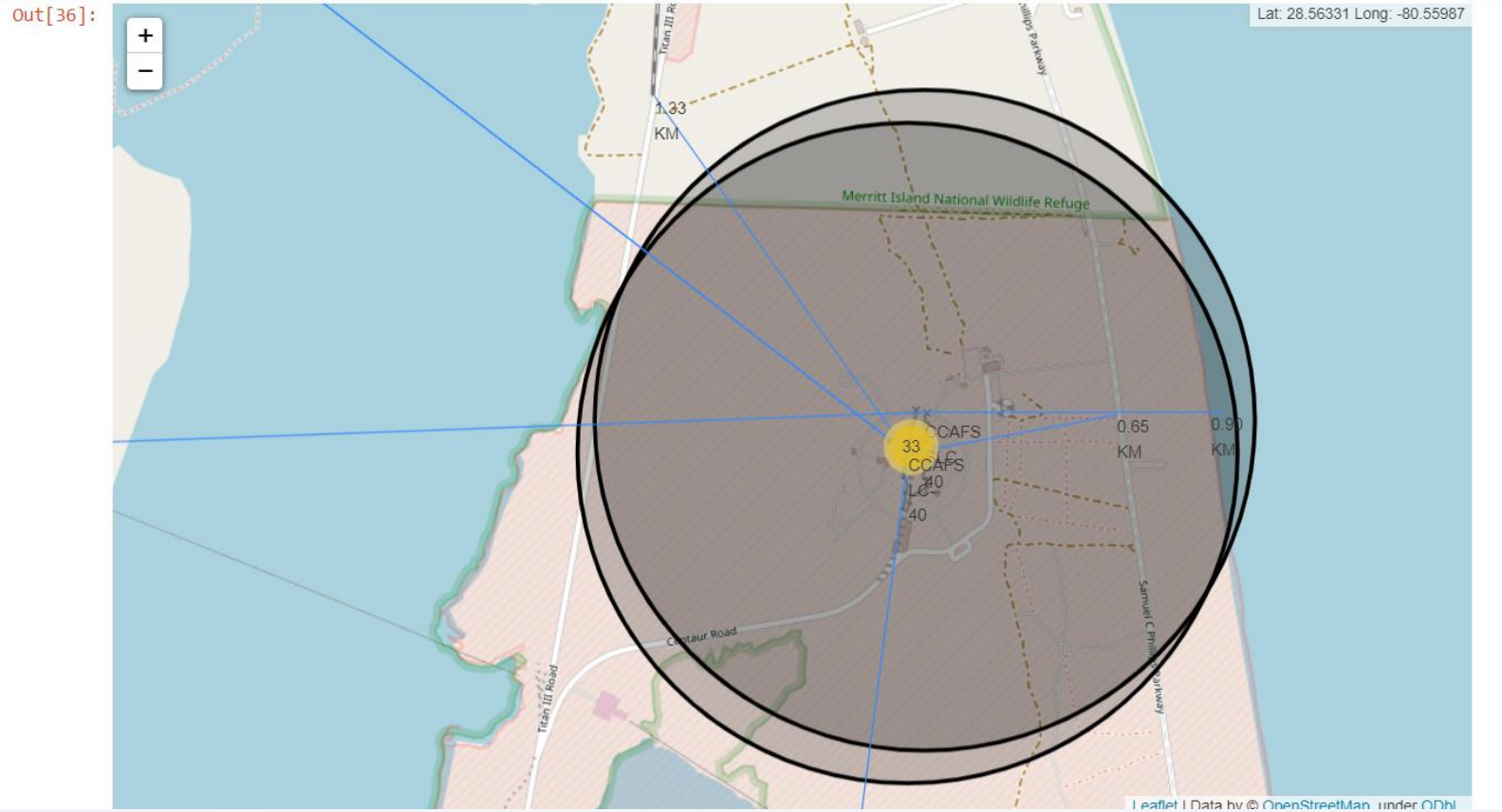
- Rocket launch sites are marked on the map with black marker.

Map with Colour-labeled Launch Outcomes



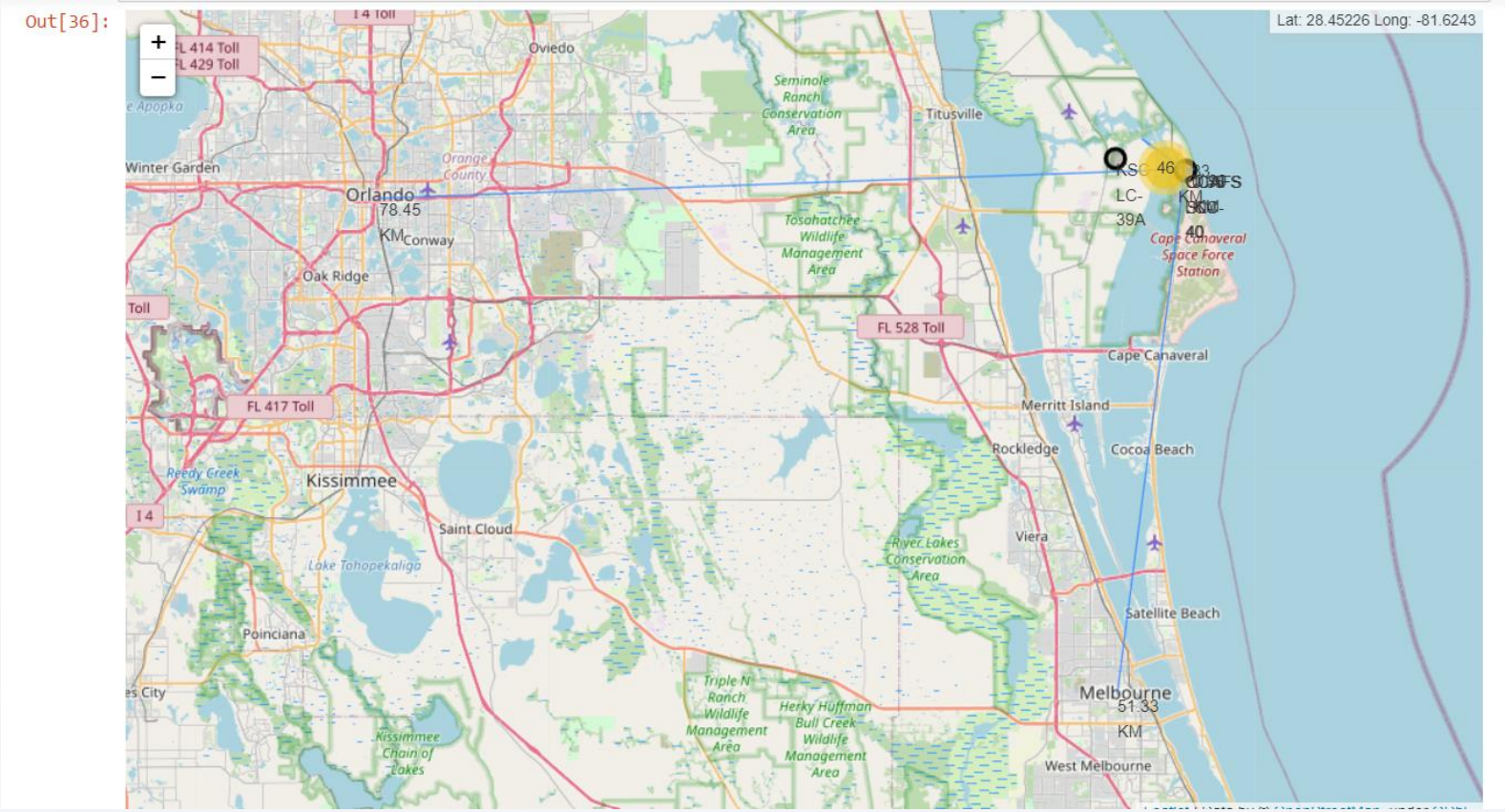
- Launch site KSC LC 39A is showed above as an example of colour-labeled launch outcome. Green indicates success, and red indicates failure.
- KSC LC 39A had 13 launches total, with 10 successful outcomes and 3 unsuccessful outcomes.

Map of Launch Site to its Proximities: Highway, Railway and Coastline

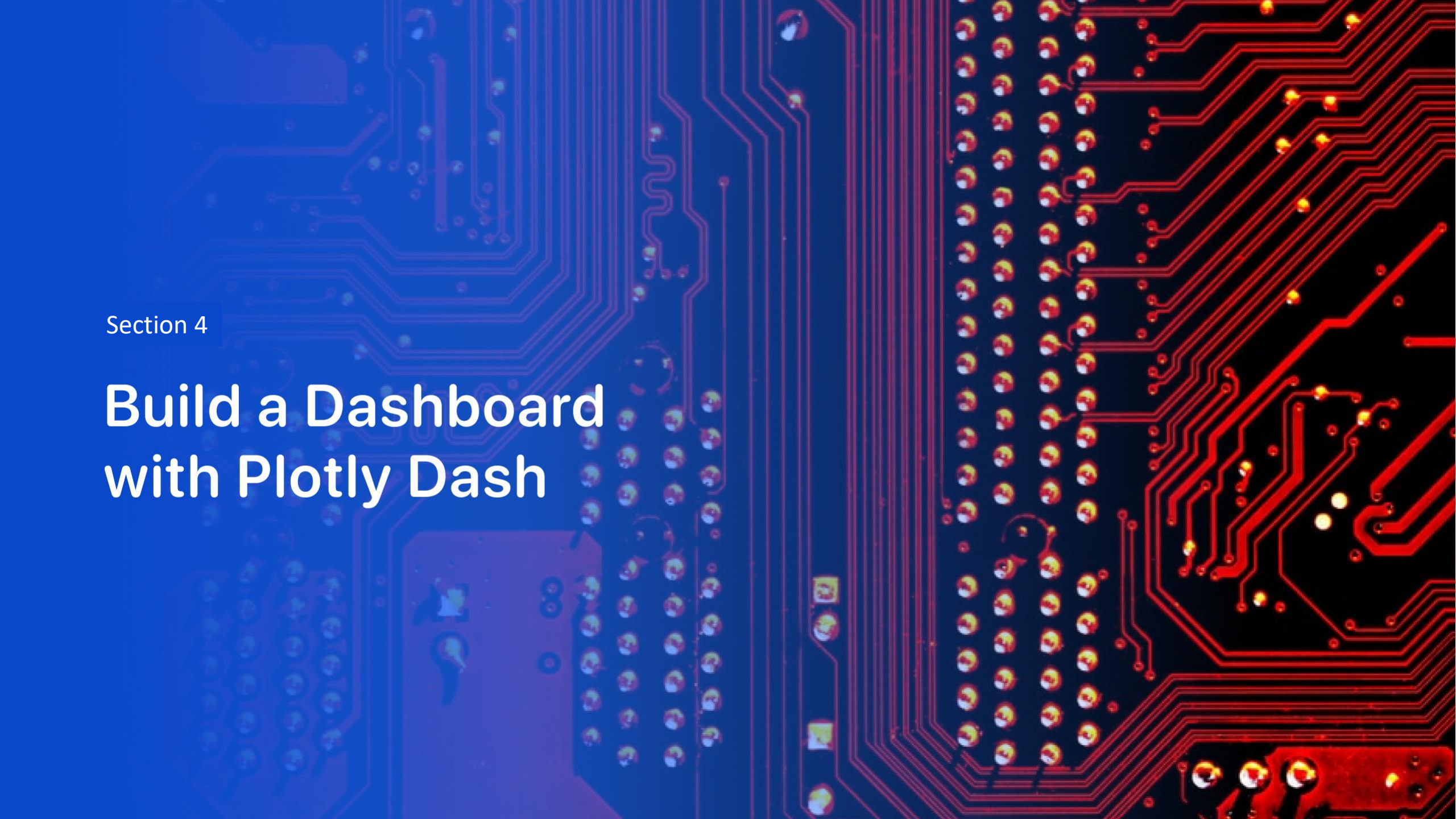


- Distance between Launch site KSC LC 39A and nearest highway, railway and coastline are mapped.
- KSC LC 39A is in close proximities to highway (0.65km), railway (1.33km) and coastline (0.90km).

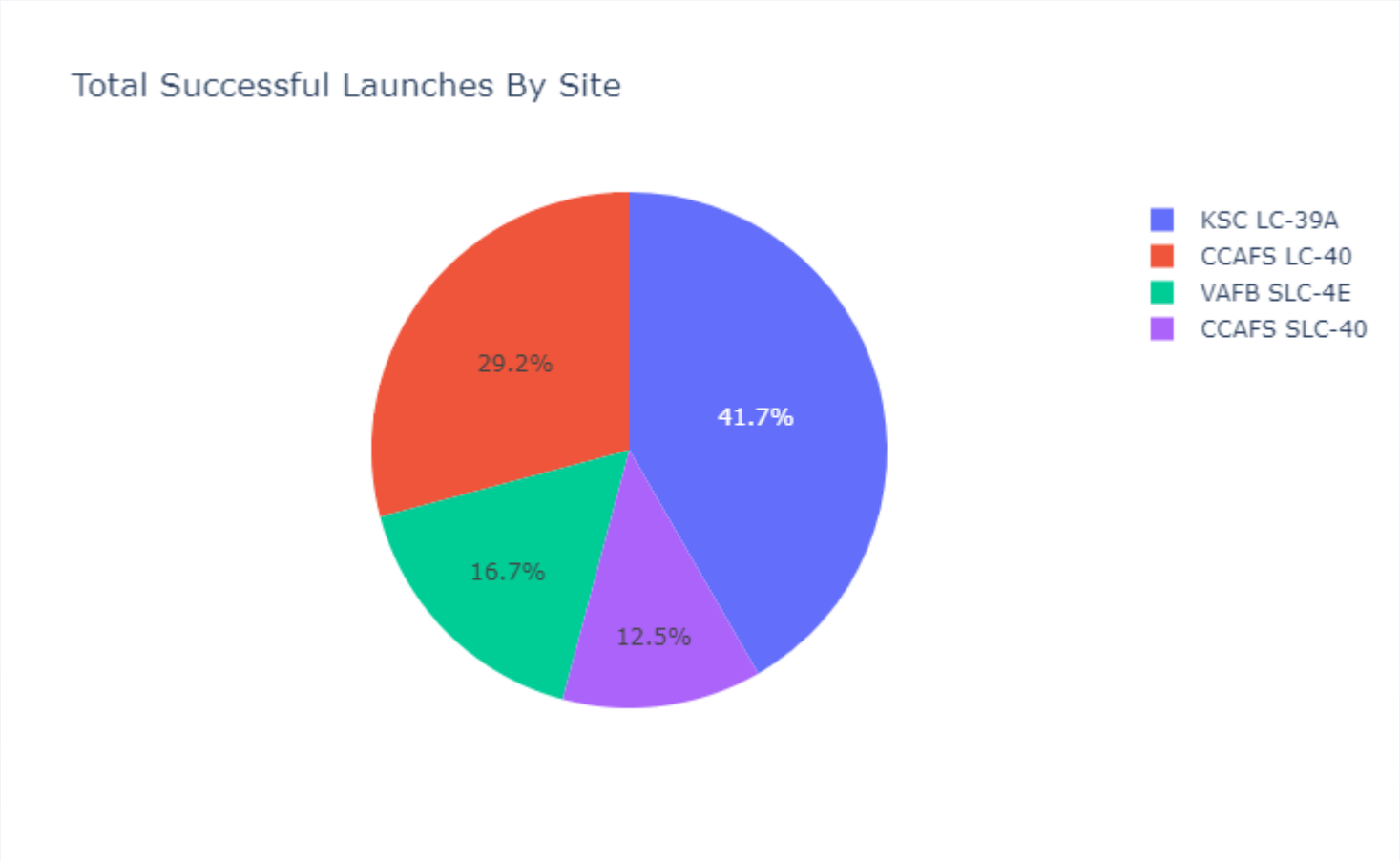
Map of Launch Site to its Proximities: City



- Distance between Launch site KSC LC 39A and nearest city is mapped.
- KSC LC 39A is located 78.45km away from Orlando, and 51.33km away from Melbourne.

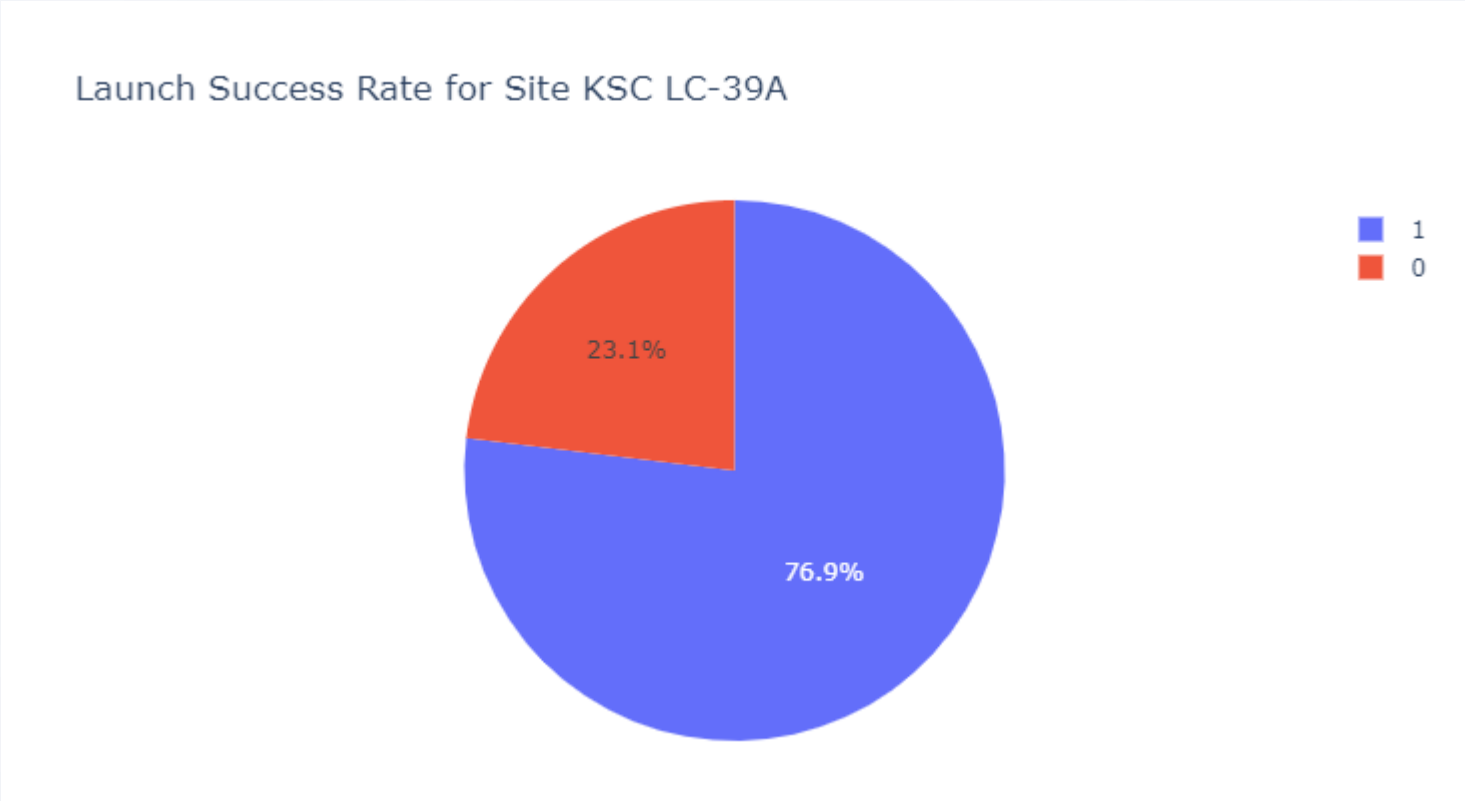

Section 4
Build a Dashboard with Plotly Dash

Successful Launches from All Sites



- KSC LC-39A launch site had the highest successful launch rate of **41.7%**. We will take a look at the launch success ratio in the next slide.

Launch Success Ratio of KSC LC-39A Launch Site



- KSC LC-39A achieved 76.9% launch success rate, while 23.1% of launches failed.

Payload vs. Launch Outcome

- **Lightweight payload (0kg to 5000kg) range** had a higher success rate, with **FT booster version** being the most successful booster version

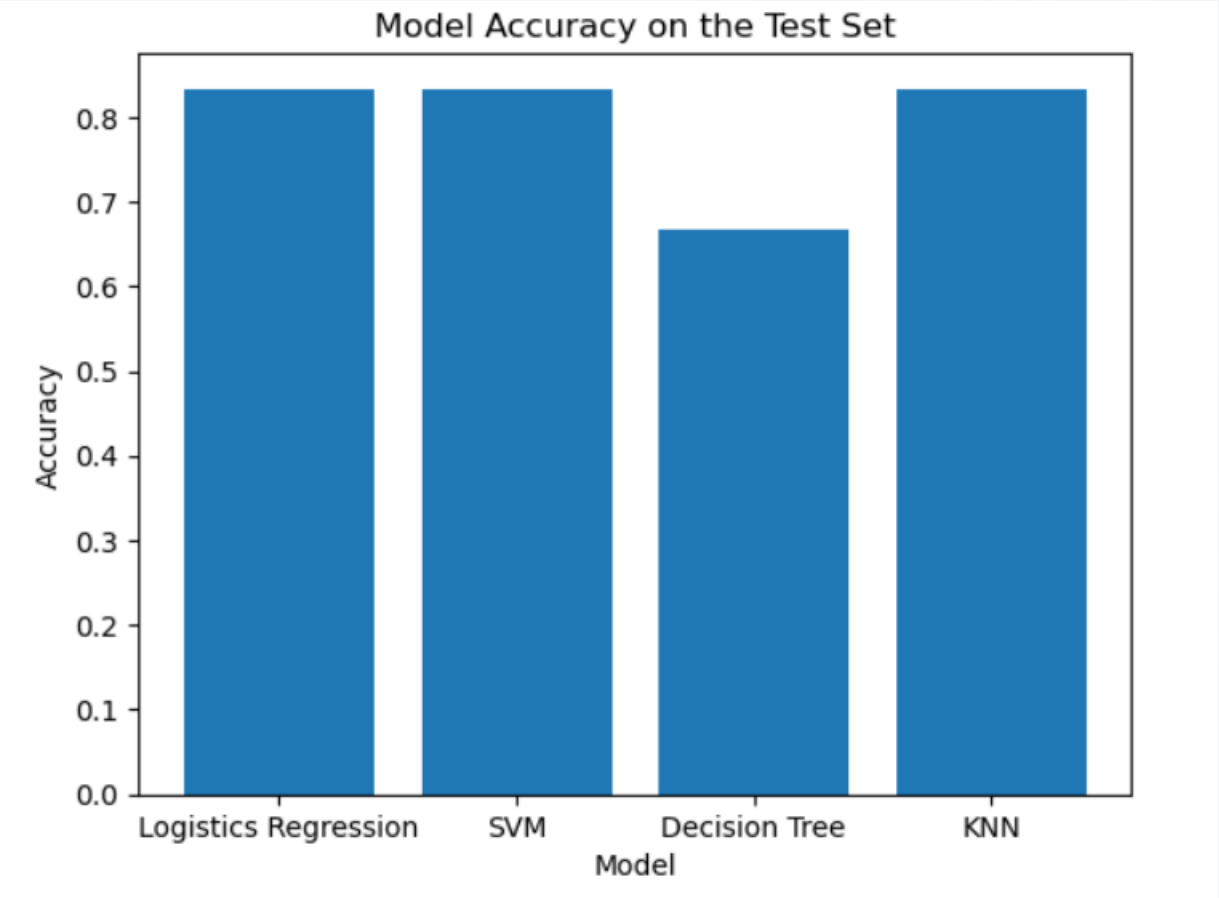
Section 5

Predictive Analysis (Classification)

Classification Accuracy

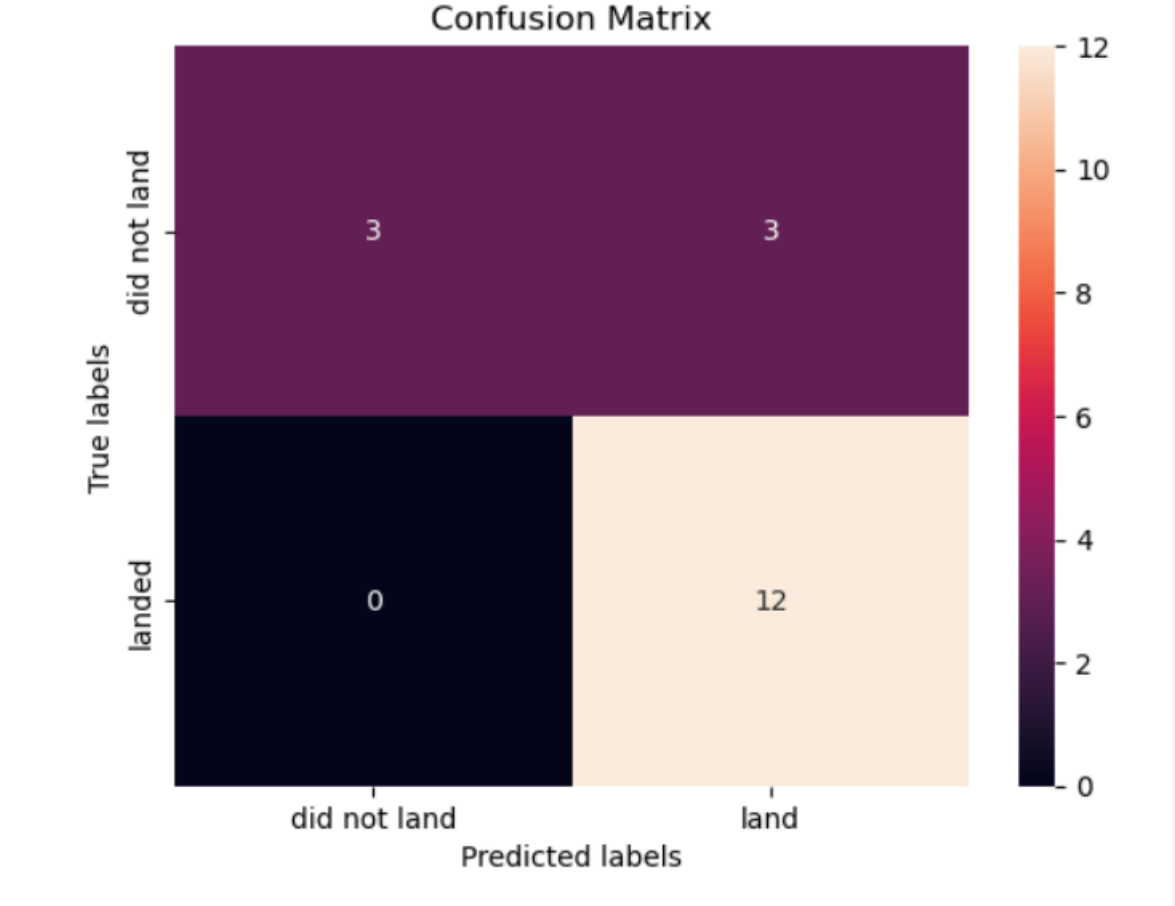
- Logistic Regression, Support Vector Machine, and K Nearest Neighbor have the highest accuracy of 0.833

Model	Accuracy
Logistic Regression	0.833
Support Vector Machine	0.833
Decision Tree	0.667
K Nearest Neighbor	0.833



Confusion Matrix

- Confusion matrix of the three best performing model shows:
  - True positive = 12
  - True negative = 3
  - False positive = 3
  - False negative = 0
- Models showed true positive rate of 80%
- Models showed true negative rate of 100%



Conclusions: Do We Have Answers To Our Problems?

- What variables contribute to successful launches?
  - Lightweight payload of less than 5000kg.
  - KSC LC 39A launch site had highest success rate.
  - Orbit types ES-L1, GEO, HEO, SSO had 100% success rate.
  - Success rate is correlated to time and number of flights.
- What is the best machine learning model for this prediction?
  - Logistic Regression, Support Vector Machine, and K Nearest Neighbor have the highest accuracy of 0.833.
- Predict if the first stage of the SpaceX Falcon 9 rocket will land successfully.
  - **Rocket will likely land successfully if launched from KSC LC 39A site with payload of less than 5000kg, with orbit types ES-L1, GEO, HEO, or SSO.**

Thank you!