IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tashif Khan
August 10, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API
  - Data Collection using web-scraping
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Maps with Folium
  - Interactive Plotting with Dashboard Application
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics Results
  - Predictive Analysis Results

Introduction

Problem for Investigation

- Rockets launched by SpaceX have a reusable first stage booster. If the first stage booster lands successfully after launch, it can be reused for future missions. The reusability of the stage 1 booster allows for millions of dollars of savings in SpaceX's overall launch price as compared to its competitors. If we can predict whether a stage 1 booster landing is successful or not, we can predict the cost of a launch for a rocket. This information can be used to find the cost of a rocket launch for SpaceX's competitors on their own missions.

Questions to Answer

- How can we predict when a booster landing is successful?
- Which features of a launch contribute to a successful booster landing?
- Where should the rocket launch take place to ensure success?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collecting data from an API and web-scraping
- Perform data wrangling
  - Applying one-hot encoding to categorical features
  - Assigning classification labels for successful or failed booster landings to 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
  - Using grid search to select parameters for decision tree, SVM, KNN, logistic regression classification models, and selecting the model with the highest accuracy

Data Collection

- SpaceX launch data was collected by:
  1. Accessing the SpaceX API with get requests and parsing the JSON response
  2. Web-scraping a Wikipedia page for launch record tables using BeautifulSoup. Results were filtered to include launches for only Falcon 9 rockets
  3. Pandas data frames were used to arrange and save the data tables as csv files

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Data Collection – SpaceX API

- From the SPACEX REST API, GET requests were used to gather data through the response
- Results were filtered in a Pandas data frame of Falcon 9 launches with the necessary variable columns

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

https://github.com/TashifK/Data-Science-Capstone/blob/main/Data%20Collection%20Notebook.ipynb

Data Collection - Scraping

- A Wikipedia site URL and a GET request was used to store the HTTP as a response
- Beautifulsoup object is created from the response, parsed, and the HTML Falcon 9 launch table is extracted
- The web-scraped data is stored in a dictionary and converted into a Pandas data frame, containing variables such as:
  - Flight Number
  - Launch Site
  - Payload mass
  - Orbit

```
# use requests.get() method with the provided static_url
requests.get(static_url)
# assign the response to a object
response = requests.get(static_url).text
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response, 'html5lib')
```

https://github.com/TashifK/Data-Science-Capstone/blob/main/Webscraping%20Notebook.ipynb

Data Wrangling

- Used the 'Landing Outcome' column to make a 'Class' training label column
  - Successful booster landings are labeled as 1
  - unsuccessful booster landings are 0
- Replaced missing values with the mean of their columns
- One-hot encoding was applied to categorical features
- Numerical columns were cast to the correct data type: float64

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
None ASDS       2
False Ocean     2
False RTLS      1
Name: Outcome, dtype: int64
```

```
df['Class']=landing_class
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

```
df.dtypes
```

```
FlightNumber      int64
Date             object
BoosterVersion   object
PayloadMass     float64
Orbit            object
LaunchSite       object
Outcome          object
Flights           int64
GridFins           bool
Reused             bool
Legs               bool
LandingPad       object
Block           float64
ReusedCount       int64
Serial           object
Longitude       float64
Latitude        float64
Class             int64
dtype: object
```

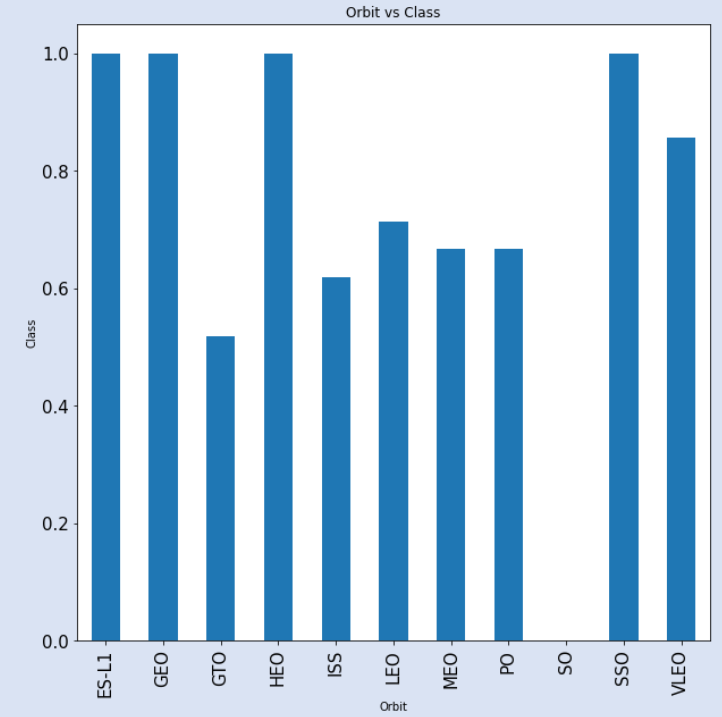
```
features_one_hot.dtypes
```

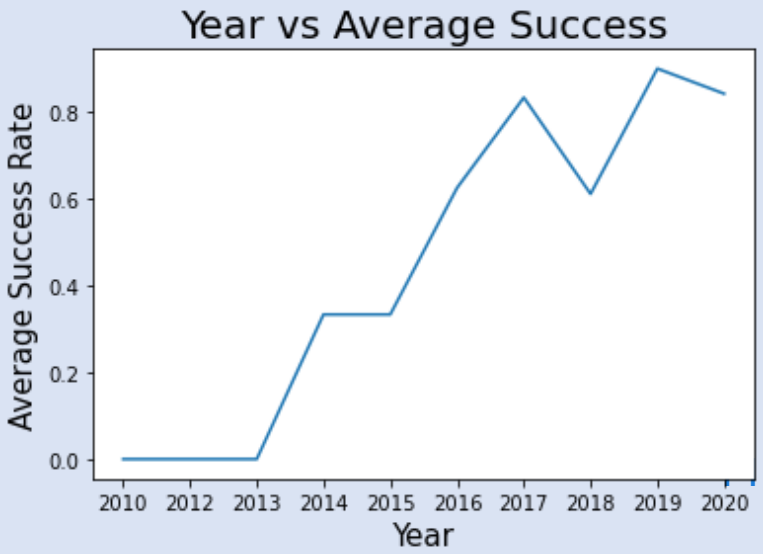
```
FlightNumber    float64
PayloadMass     float64
Flights         float64
GridFins        float64
Reused          float64
                 ...
B1056           float64
B1058           float64
B1059           float64
B1060           float64
B1062           float64
Length: 80, dtype: object
```

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

EDA with Data Visualization

- To see their effects on launch success, variables in the data were plotted against:
  - Orbit types
  - Launch site
- Average success rate of rockets were plotted to see if successful landings were becoming more common overtime
- Scatterplots, bar charts and line plots were used to visualize the data





https://github.com/TashifK/Data-Science-Capstone/blob/main/EDA%20Visualization%20Notebook.ipynb

EDA with SQL

- SpaceX dataset was loaded into IBM-Db2 as a table ‘SPACEXTBL’
- To discover which features like booster versions, launch site and payload mass contributes to successful landings, SQL queries were used to:
  - Find all launch site names
  - Find average payload mass for F9 v1.1
  - Find booster versions carrying maximum payload mass
  - Find failed landing outcomes, their booster versions and launch site name
  - and more

https://github.com/TashifK/Data-Science-Capstone/blob/main/EDA%20with%20SQL%20Notebook.ipynb

Build an Interactive Map with Folium

- Folium maps were used to analyze factors in existing launch locations to find optimal initial positions for a successful launch
- To see how launch success is affected by launch location, folium maps were used to analyze these features
  - Color coded Markers and Marker clusters were used to demarcate the position of launch sites on the world map and each landing outcome at the location
  - Measured lines were plotted between a launch site and railway, highway, coastline and city to analyze launch site locations

https://github.com/TashifK/Data-Science-Capstone/blob/main/Folium%20Interactive%20Visual%20Analytics%20Notebook.ipynb

Build a Dashboard with Plotly Dash

- On a Plotly dashboard application interactive plots were created.
  - To break down the success rate of each the launch site:
  - Pie chart of total successful stage 1 landings by site
  - Pie charts of successful/unsuccessful stage 1 landings for individual launch sites
- Scatter plots of payload mass and outcome class for each booster version
- The relationship between payload mass and landing outcome was also plotted visually to answer:
  - Which launch site has the largest successful launches?
  - Which payload range has the highest launch success?
  - Which booster version has the best success rate?

https://github.com/TashifK/Data-Science-Capstone/blob/main/Dashboard%20Application%20with%20Plotly%20Dash%20Notebook.ipynb

Predictive Analysis (Classification)

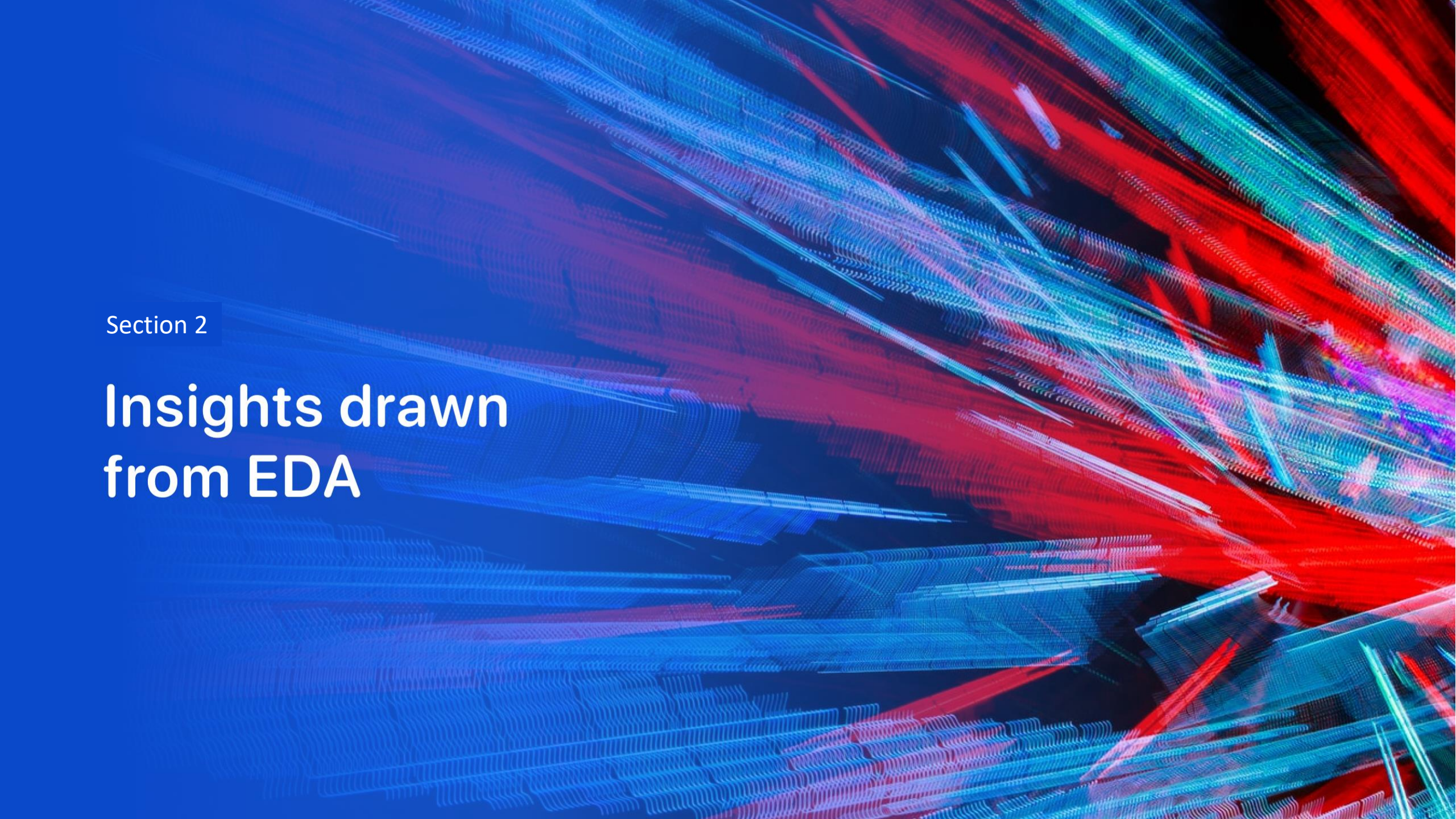
- To predict the success or failure of a launch, four classification algorithms were used:
  - *K*-nearest neighbors
  - Decision tree
  - Support Vector Machine
  - Logisitic Regression
- Grid-Search was used to select the best hyper-parameters for each model
- The best classifier was selected based on the model's confusion matrix and accuracy score; its ability to classify successful or unsuccessful launches correctly

	Model	Model Scores
0	Decision Tree	0.889286
1	KNN	0.848214
2	Logistic Regression	0.846429
3	SVM	0.848214

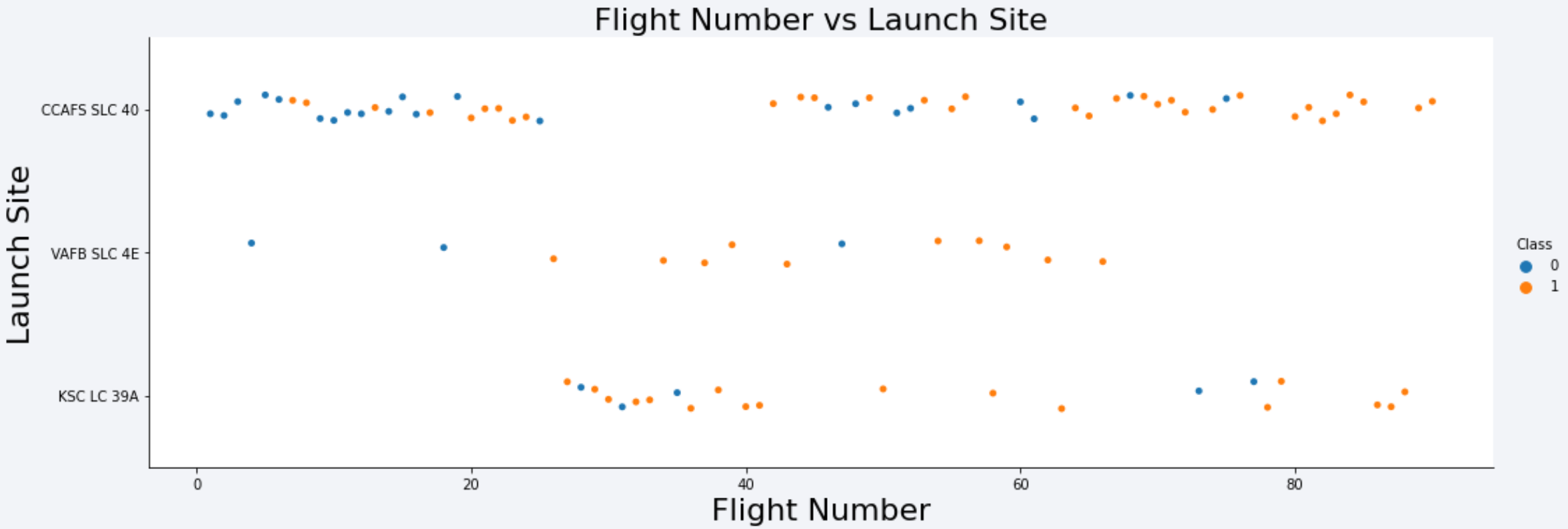
https://github.com/TashifK/Data-Science-Capstone/blob/main/Machine%20Learning%20Predictions%20Notebook.ipynb

Results

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

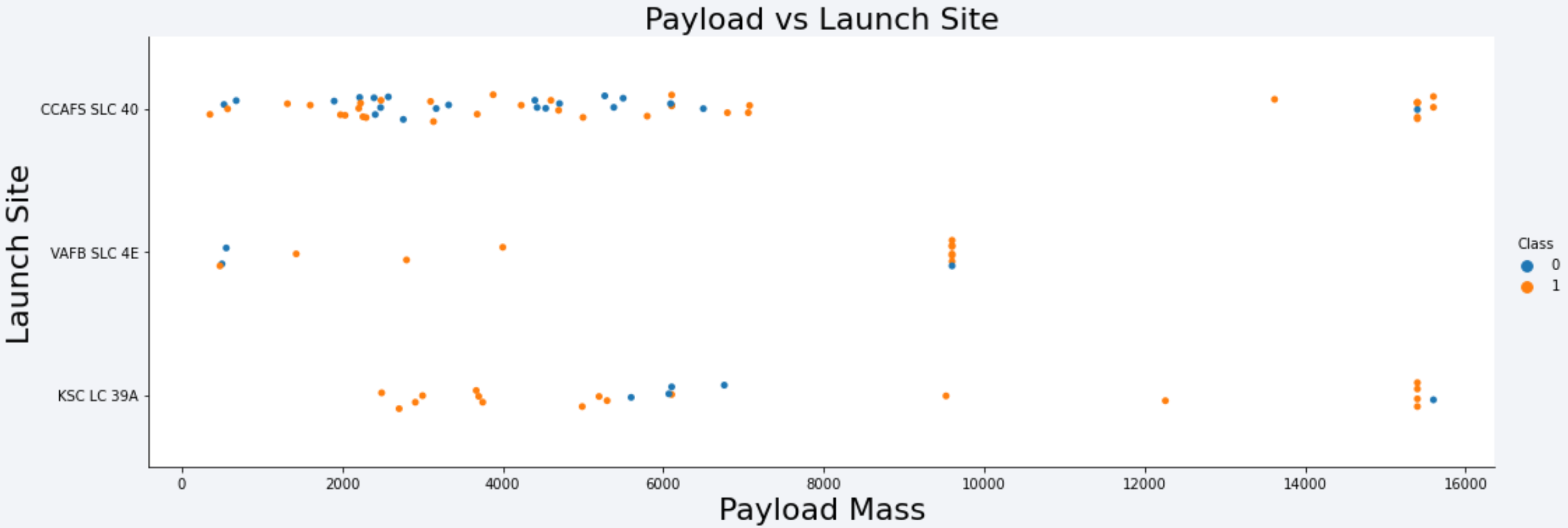

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



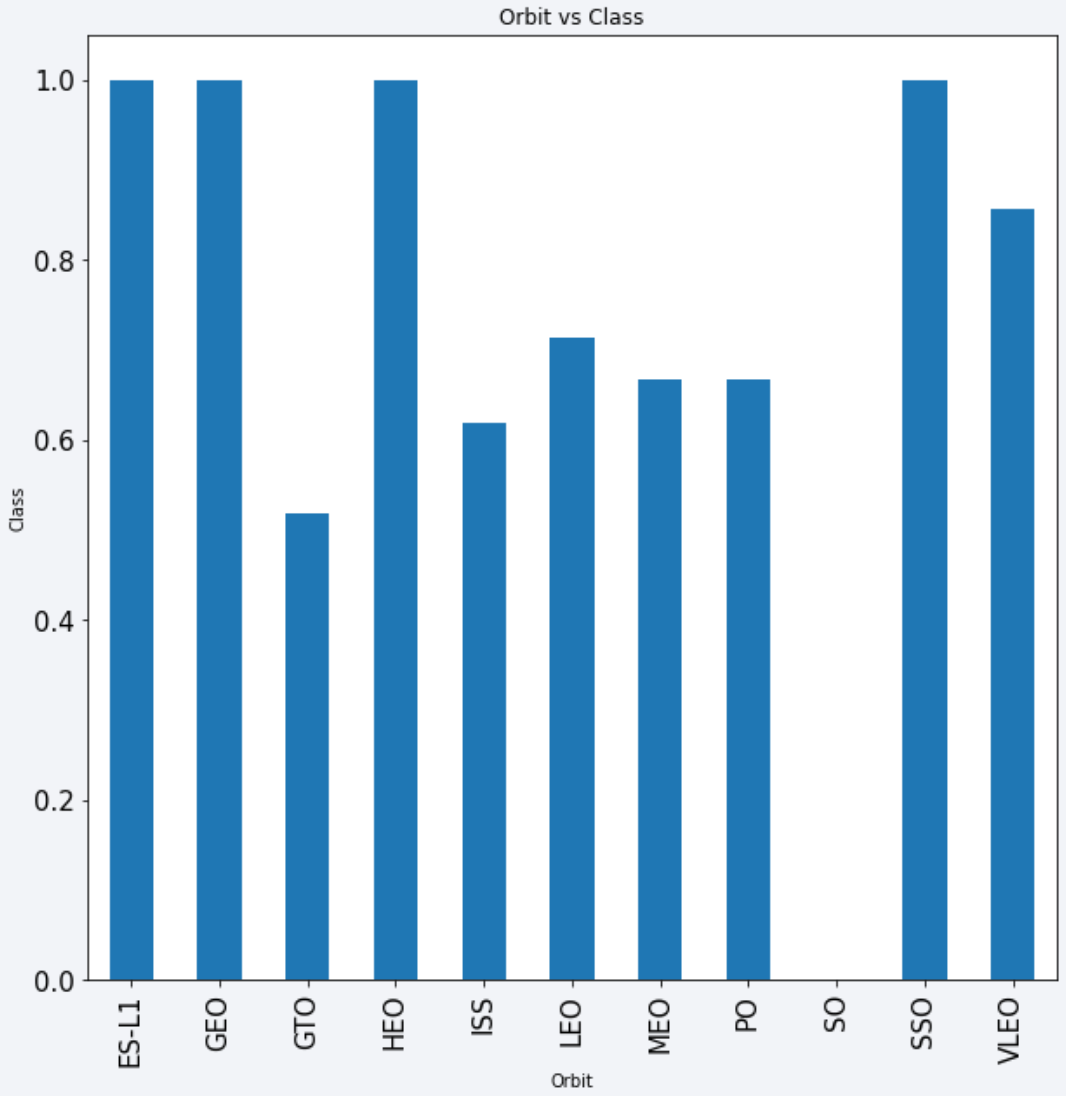
- As the number of flights for a launch site increases, the likelihood of successful stage 1 landings for the site increases

Payload vs. Launch Site



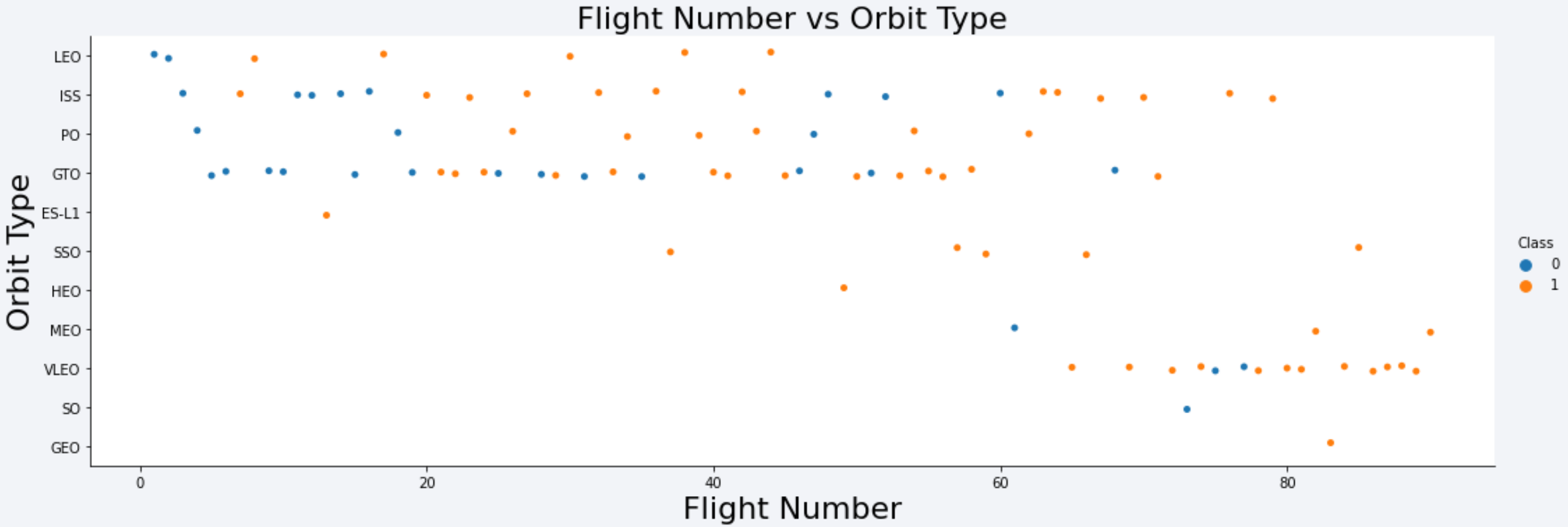
- There are no rockets launched from VAFB SLC 4E greater than 10000.
- Heavy payload launches from CCAFS SLC 40 and KSC LC 39A have a high success rate

Success Rate vs. Orbit Type



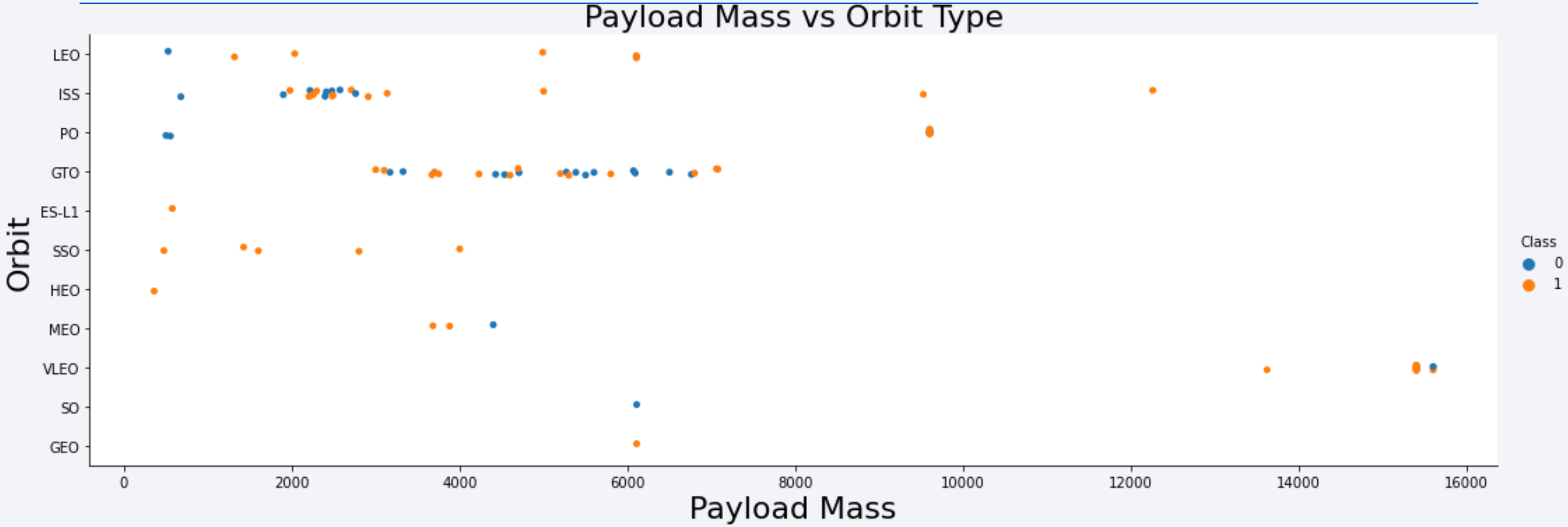
- ES-L1, GEO, HEO and SSO orbits have had a 100% success rate of stage 1 landing.
- SO orbits have had a 0% success rate of stage 1 landing

Flight Number vs. Orbit Type



- Successful landings are related to flight number for many of the orbit types such as LEO
- However, this is not the case for GTO orbit type

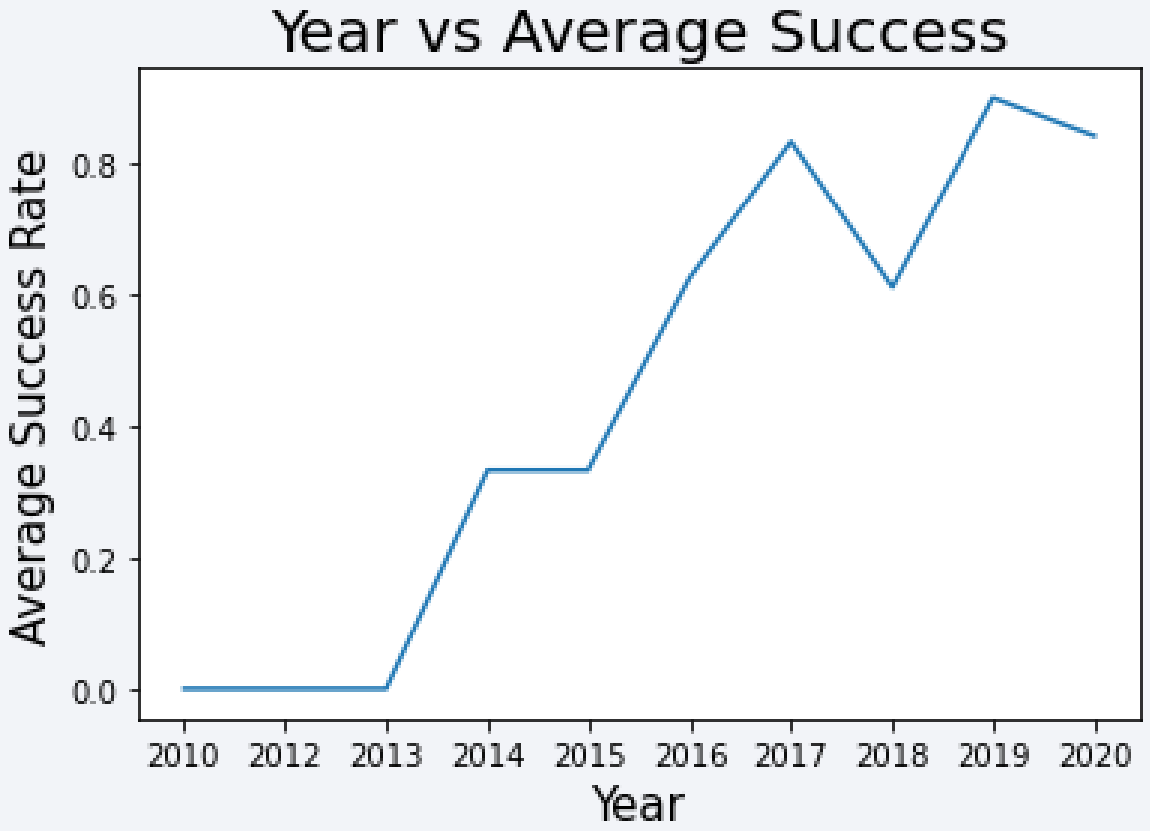
Payload vs. Orbit Type



- For heavy payloads, LEO, PO and ISS have high success rates
- For lighter payloads, SSO had a high success rate

Launch Success Yearly Trend

- The success rate of launches has increased from 2013 to 2020



All Launch Site Names

- The names of all launch sites were retrieved from the database table SPACEXTBL

```
%sql select unique Launch_site from SPACEXTBL
```

[160]:

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

Launch Site Names Begin with 'CCA'

- The names of 5 records where launch sites begin with `CCA`

```
%sql select * from SPACEXTBL where Launch_site like 'CCA%' limit 5
```

[8]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- These records were retrieved from the SPACEXTBL by searching for 5 rows where the launch sites' names begin with "CCA…"

Total Payload Mass

- Total payload carried by boosters from NASA is 45596kg

```
%sql select sum(PAYLOAD_MASS__KG_) as total_NASA from SPACEXTBL where Customer like 'NASA (CRS)'
```

[28]:

total_nasa
45596

- Total payload was retrieved from the SPACEXTBL by summing the payload for all rows in the table where customer name is ‘NASA (CRS)’

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2928kg

```
%sql select avg(PAYLOAD_MASS__KG_) as booster_average from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

booster_average
2928

- Average payload mass carried by booster version F9 v1.1 was retrieved from the SPACEX table by averaging the payload mass entries for rows where the booster version is named ‘F9 v1.1’

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was 22nd December 2015

```
%sql select min(Date) as first_landing from SPACEXTBL where landing__outcome = 'Success (ground pad)'
```

first_landing
2015-12-22

- The date of first landing on ground pad was retrieved from the SPACEX table by taking the minimum date value for rows where the landing outcome is ‘Success (ground pad)’

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

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

```
%sql select Booster_Version from SPACEXTBL \
where PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 \
AND landing__outcome = 'Success (drone ship)'
```

- The booster versions were retrieved from the SPACEX table where payload mass is between 4000 and 6000, and landing outcome is 'Success (drone ship)'

Total Number of Successful and Failed Mission Outcomes

- Total number of successful mission outcomes is 100

```
%sql select count(Mission_Outcome) as Successful_Missions from SPACEXTBL \
where Mission_outcome like 'Success%'
```

[63]:

successful_missions
100

- Total number of failed mission outcomes is 1

```
%sql select count(Mission_Outcome) as Failed_Missions from SPACEXTBL \
where Mission_outcome like 'Fail%'
```

[64]:

failed_missions
1

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass of 15600 are:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ \
FROM SPACEXTBL where PAYLOAD_MASS__KG_ = \
(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

- Booster versions that carried the maximum payload were retrieved from the SPACEXTBL rows in which the payload mass was equal to the maximum payload mass

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 were found using:

```
%sql select Date, Landing__Outcome, Booster_Version, Launch_Site \
from SPACEXTBL \
where YEAR(Date) = '2015' AND Landing__Outcome = 'Failure (drone ship)'
```

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Both booster versions used in 2015 were from CCAFS LC-40 launch site, and resulted in a failed drone ship landing

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

- Count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order are:

```
%sql select Landing__Outcome, count(Landing__Outcome) as Count \
from SPACEXTBL \
where Date between '2010-06-04' and '2017-03-20' \
group by Landing__Outcome order by count(Landing__Outcome) desc
```

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

Section 3

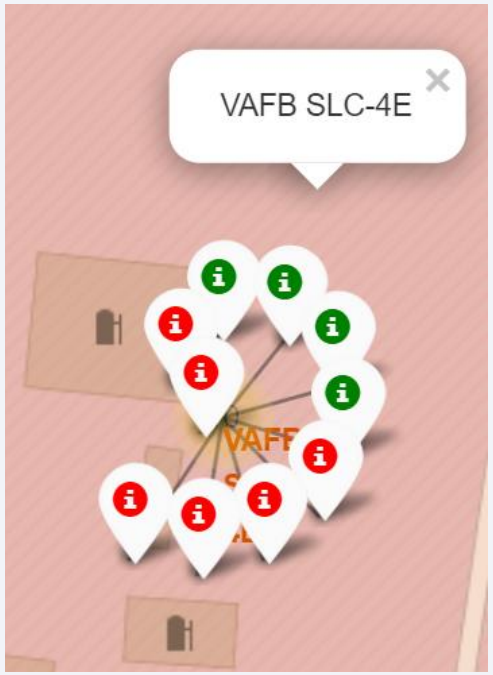
Launch Sites Proximities Analysis

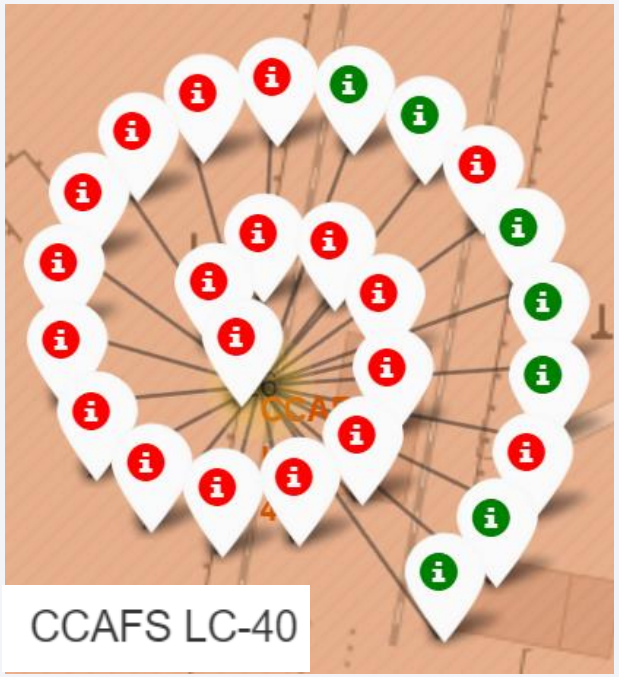
All Launch Site Locations

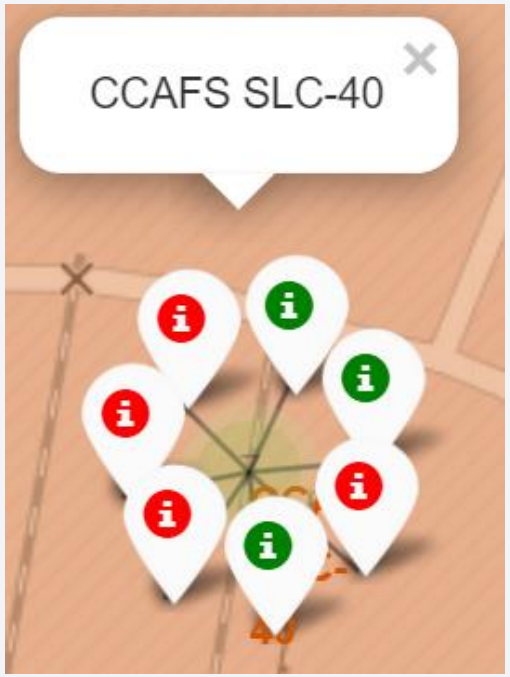
- All launch sites are coastal locations close to the equator and within the United States
- On the Florida coast, 46 total launches are from:
  - CCAFS SLC-40
  - CCAFS LC-40
  - LC-39A
- On the California coast, 10 total launches are from:
  - VAFB SLC-4E

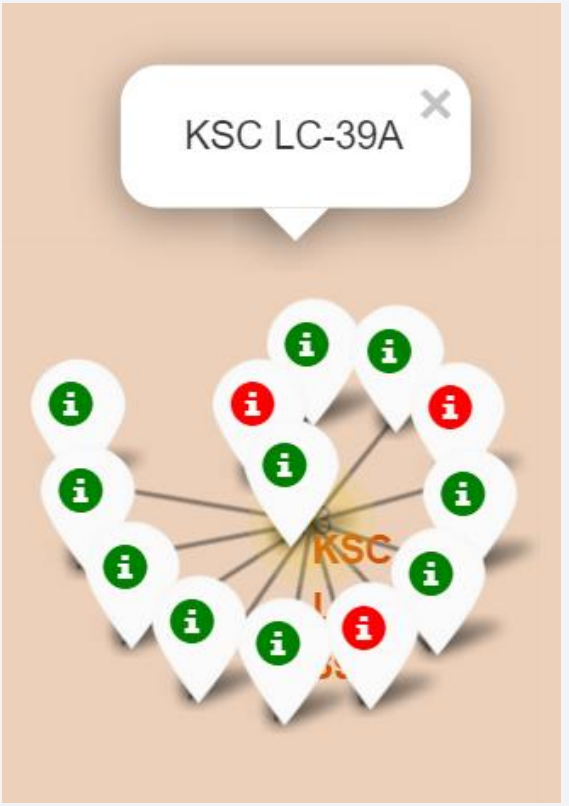
Launch Outcomes at Launch Sites

- Successful landings are green markers.
- Failed landings are red markers





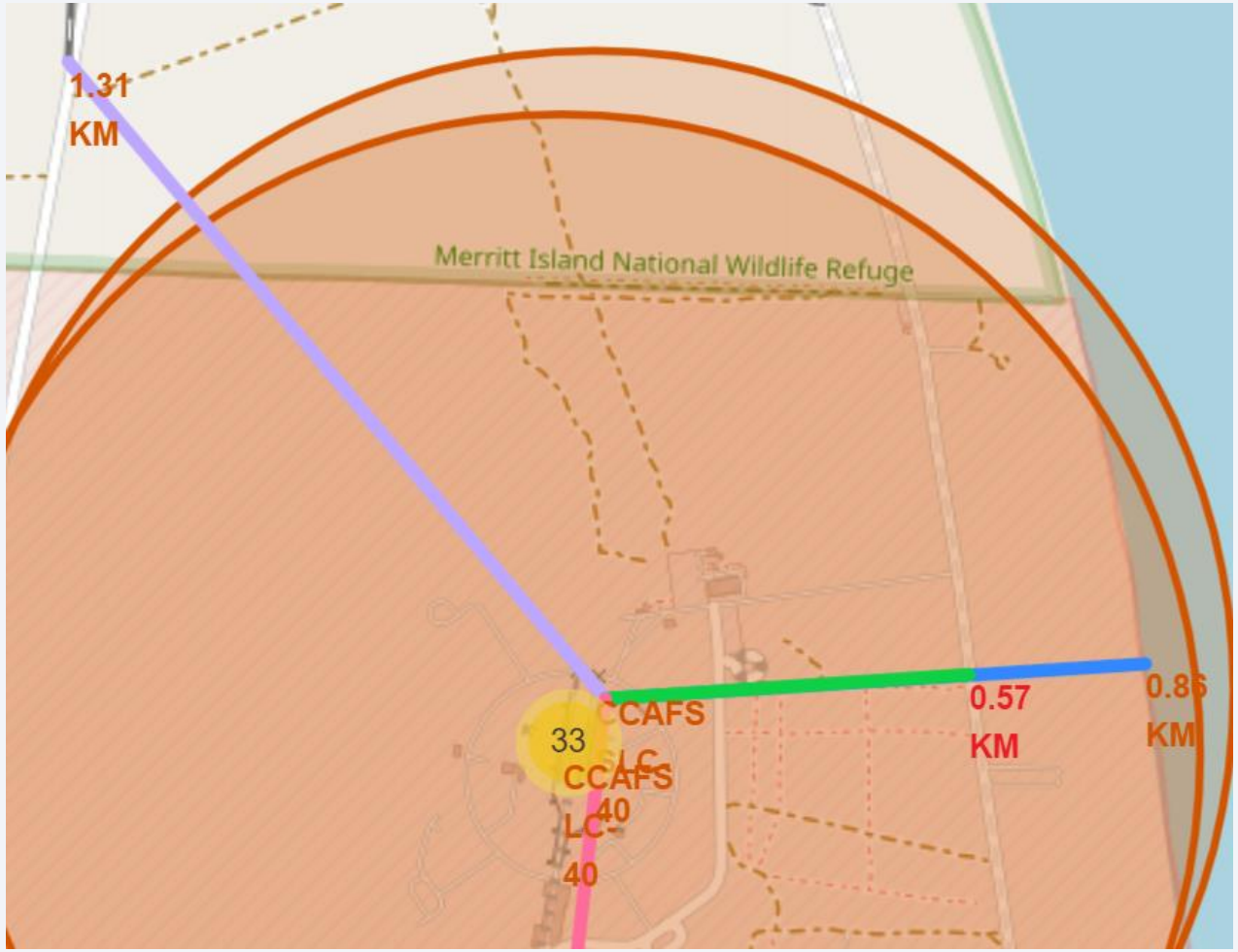


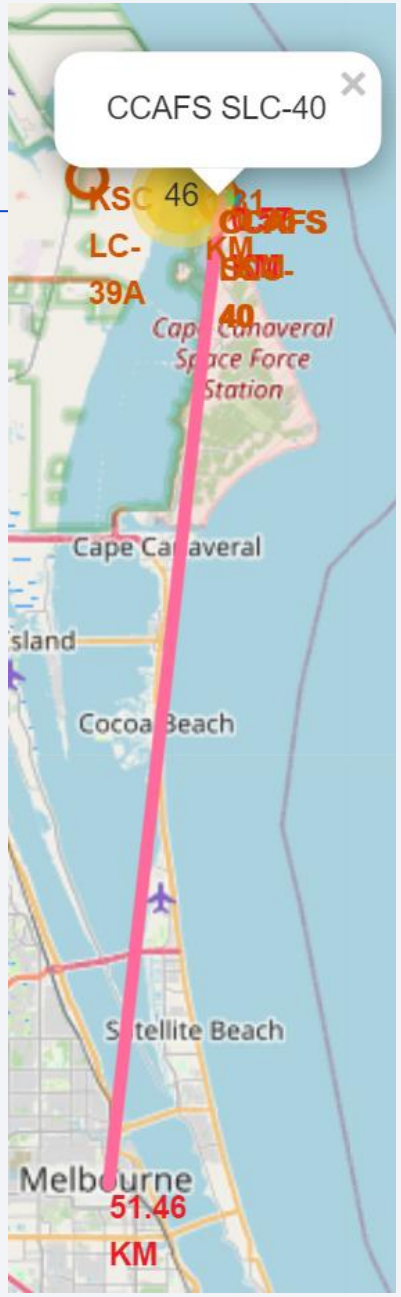


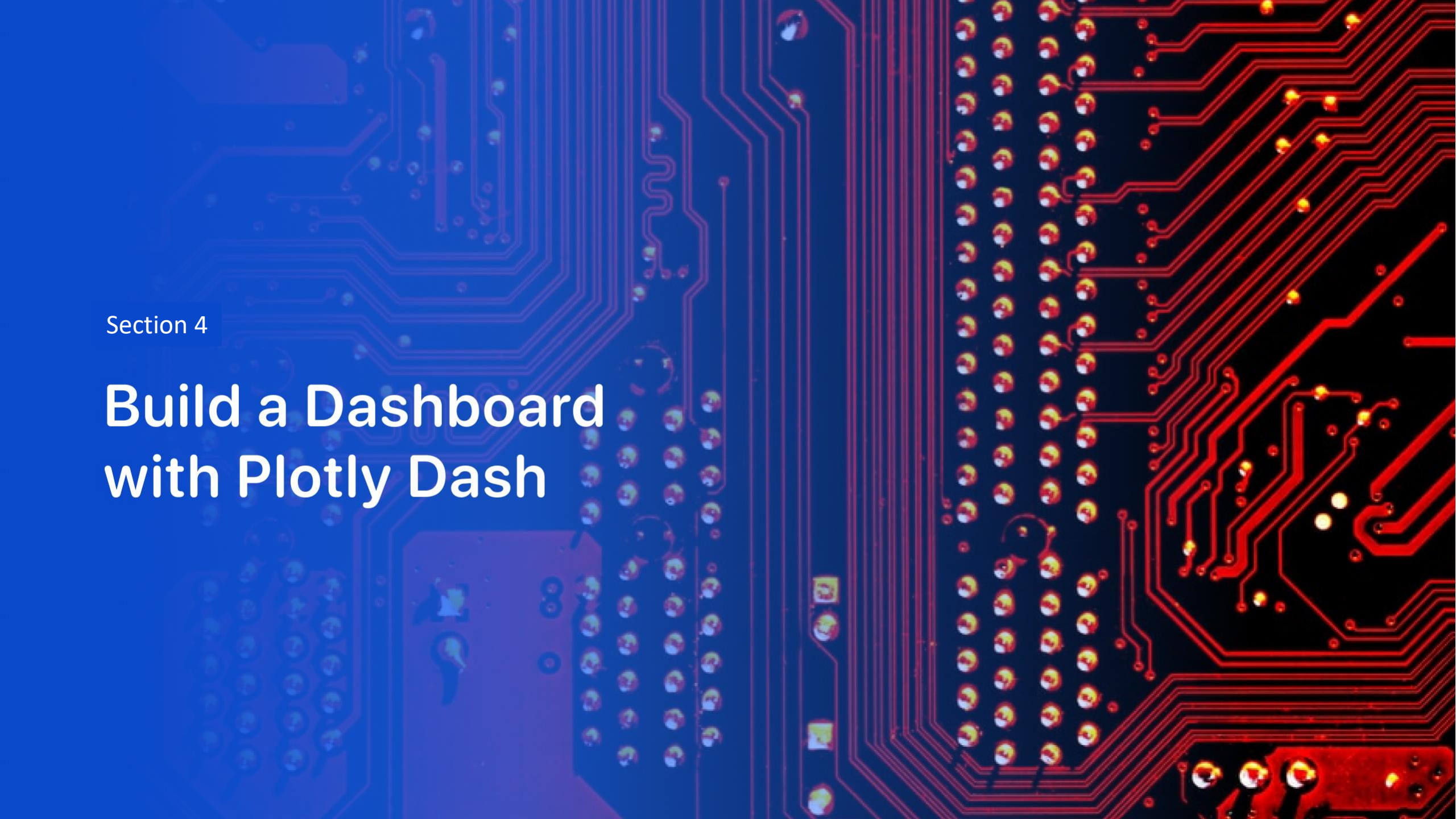
- Successful launch ratios are high at KSC LC-39A and CCAFS SLC-40
- These may be optimal launch positions for successful landings

Launch Site Proximity

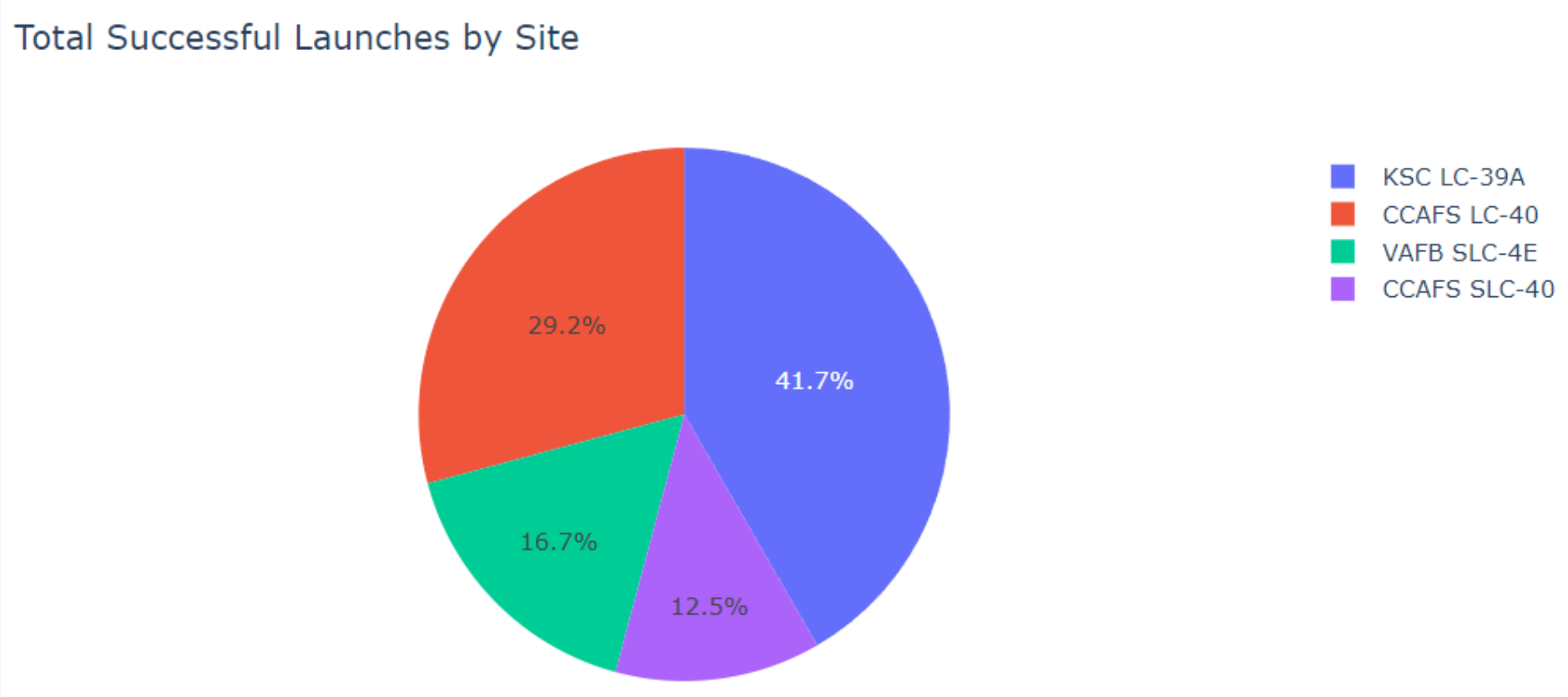
- Lines have been plotted from the CCAFS SLC-40 launch site to its closest surroundings of interest
- The launch site is not in close proximity to railways, highways, coastlines or cities.
- This may be to keep those safe from potential dangers in the launch site





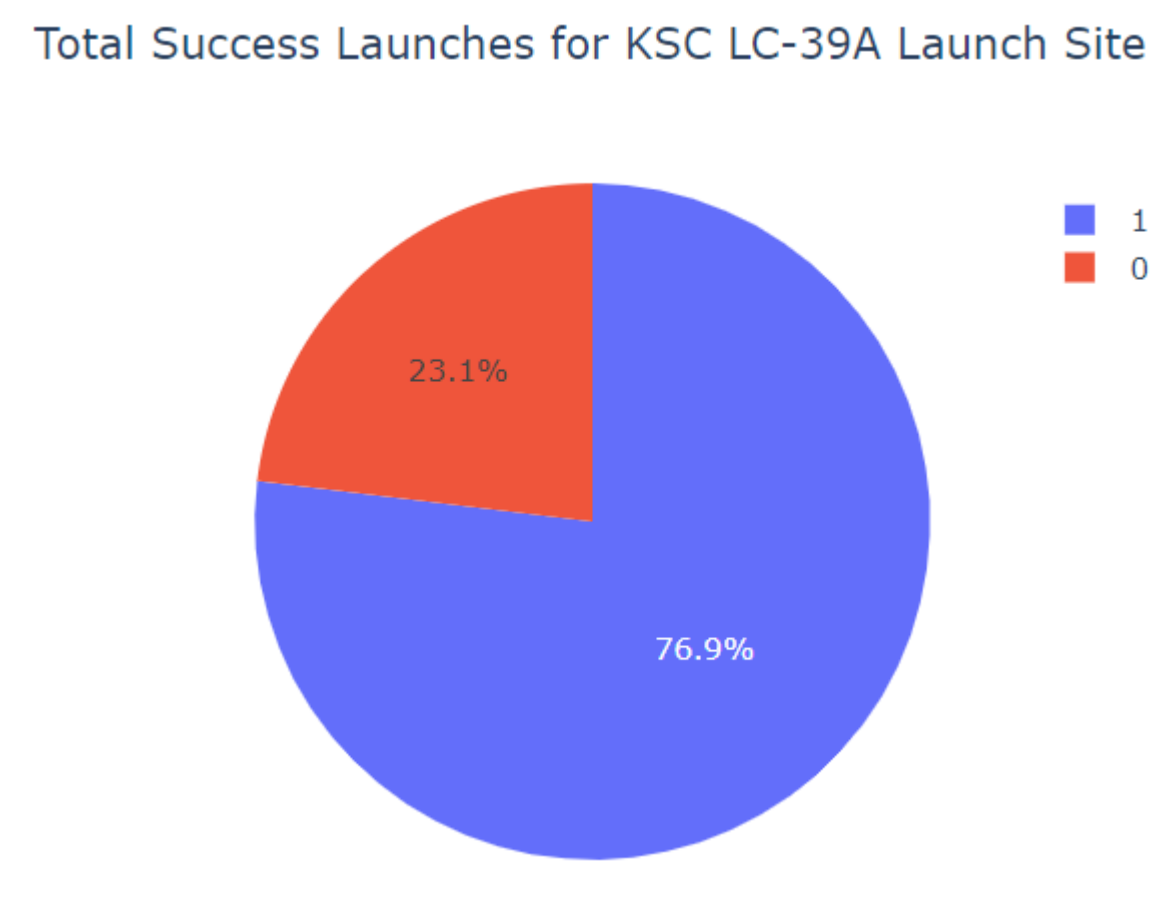

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



- KSC LC-39A site makes up most of the total successful launches.

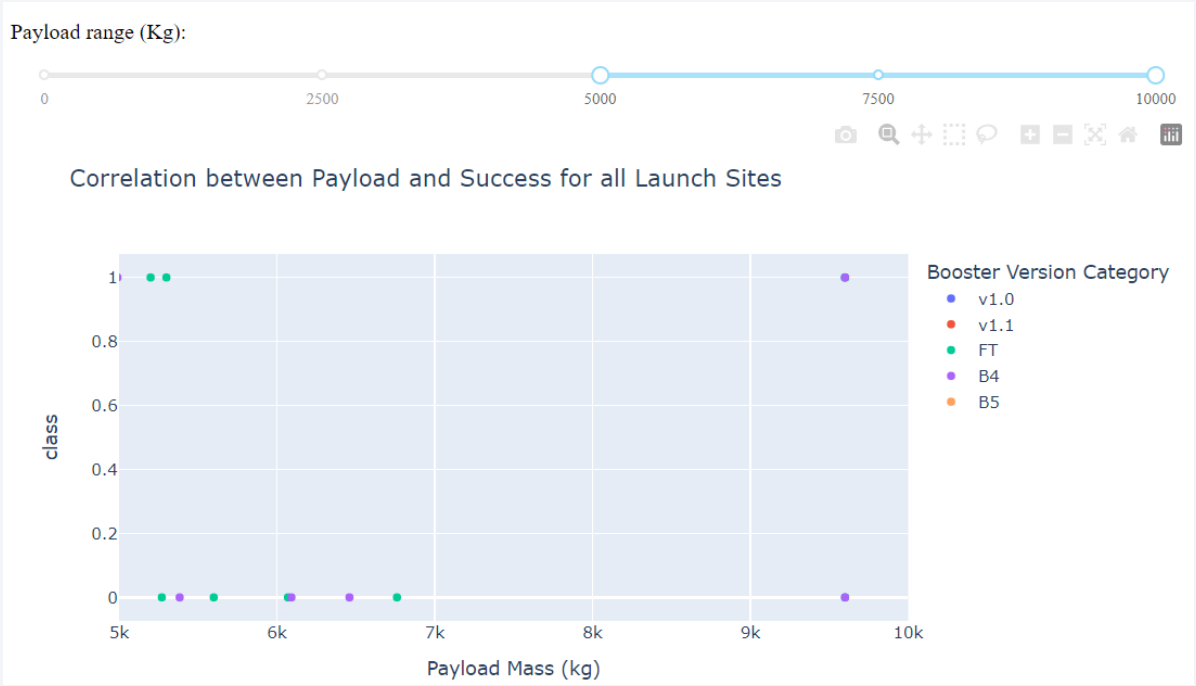
Launch Site with the Highest Launch Success Ratio



- KSC LC-39A launch site has the highest launch success ratio: 76.9%

Payload vs Launch Outcome

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider



- The success rate is much higher for payloads between 2000 and 5000kg
- The 'FT' booster version has a high success rate with low payloads

Section 5

Predictive Analysis (Classification)

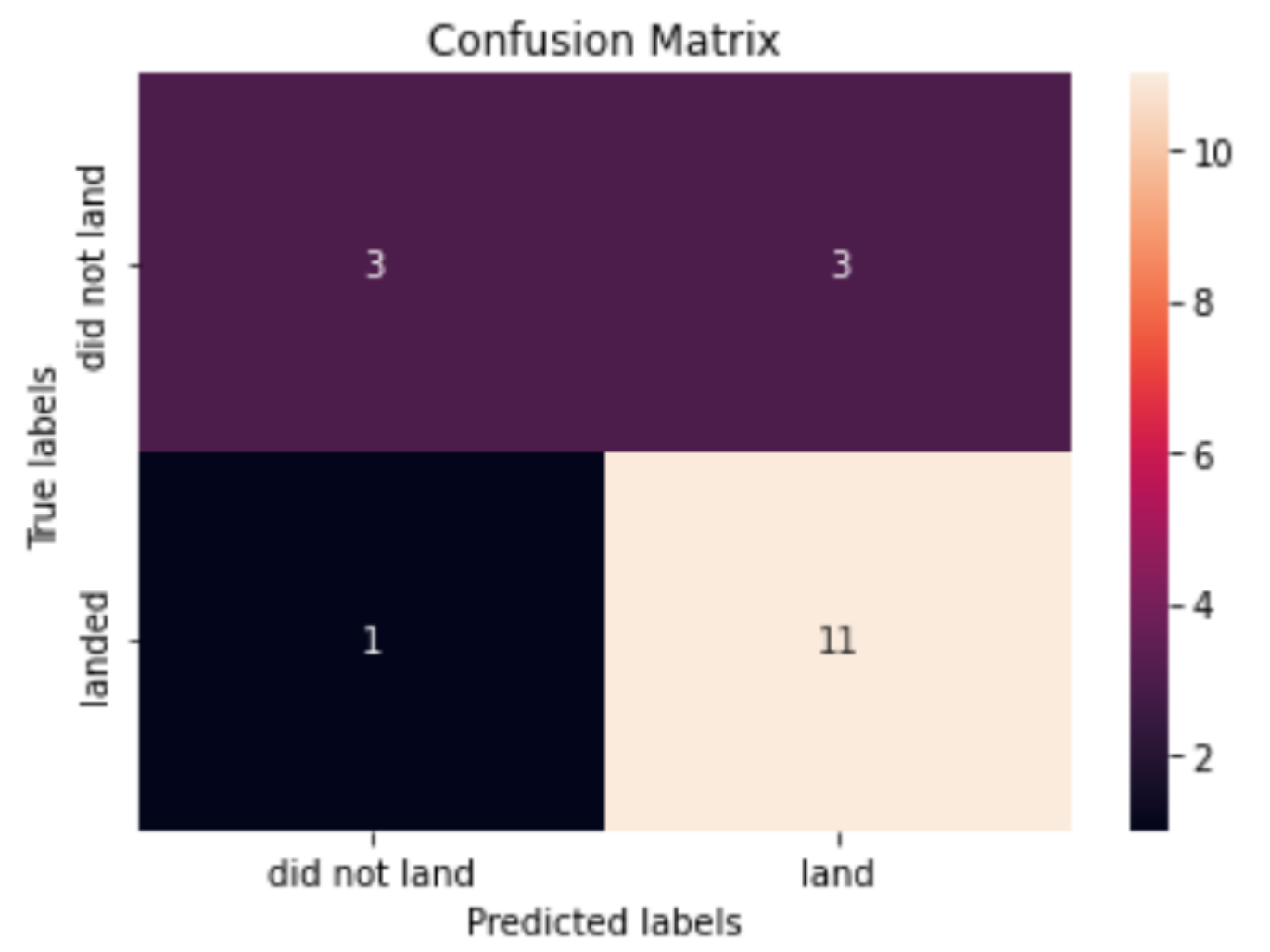
Classification Accuracy

- Model accuracy for all of the classification models

	Model	Model Scores
0	Decision Tree	0.889286
1	KNN	0.848214
2	Logistic Regression	0.846429
3	SVM	0.848214

- The decision tree model has the highest classification accuracy of 88.928%. It is the best of the four models at classifying launch outcomes correctly.

Confusion Matrix



- Confusion matrix for the decision tree classification model classifies most cases correctly
- Most of the incorrect classifications are false positives, where the model predicts that an unsuccessful landing is successful.

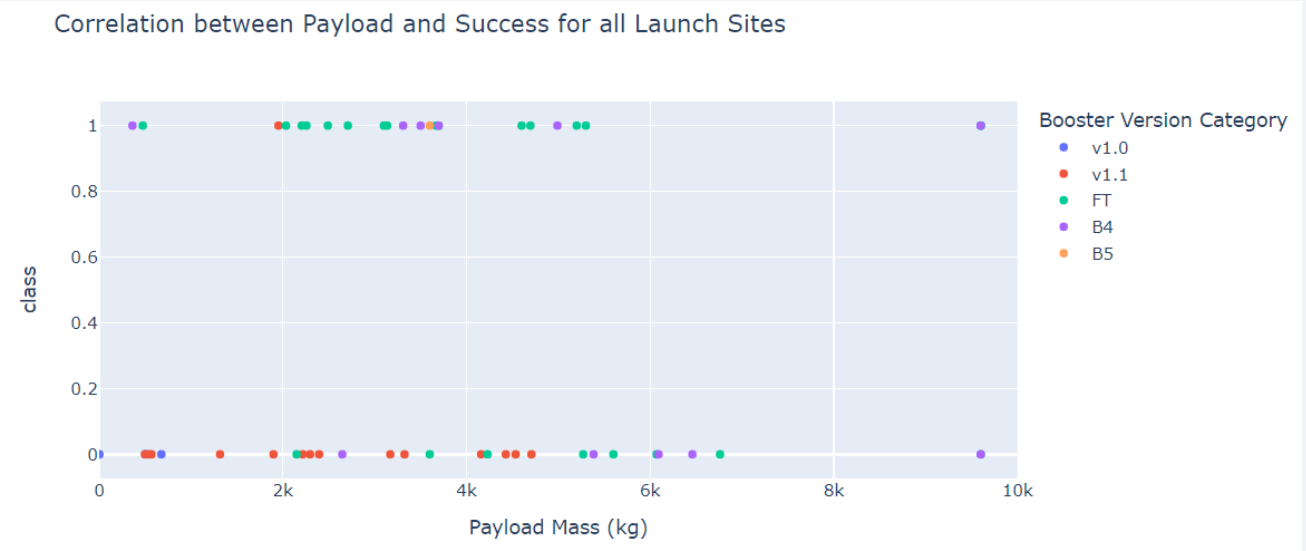
Conclusions

- **Launch Locations**
  - Launch sites are located near the equator, away from cities, highways, coasts, and railroads
  - KSC LC-39A launch site has the highest launch success ratio of 76.9% and the most total successful launches
- **Launch Specifications**
  - SSO orbits have had a very high success rate especially for light payloads
  - FT booster version has high success rate with light payloads
  - LEO, PO and ISS orbits are better for heavy payloads and have a high success rate when launched from CCAFS SLC 40 or KSC LC 39A sites
  - Successful landings are related to flight number for many of the orbit types such as LEO
- **Outcome Prediction**
  - Decision tree is the best classification algorithm for predicting launch success

Appendix

- **Project GitHub repository**
  - https://github.com/TashifK/Data-Science-Capstone
- **SpaceX API**
  - https://api.spacexdata.com/v4/launches/past
- **Wikipedia Page**
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- Payload and Class Plot for All Launch Sites



Thank you!