IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Steve Bauer
01/29/22

Outline

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

Executive Summary

- The goal of the project was to develop a classification model to predict the success of a SpaceX Falcon 9 first stage rocket landing
- The two main data sources were the SpaceX REST API and the Wiki page for the launch results
- The data was cleaned and prepared mainly using the Pandas library in Jupyter Notebooks, which was where most of the exploratory data analysis took place
- The cleaned data was used to fit 4 different classification models – Logistic Regression, Decision Tree, Support Vector Machine, and K-nearest Neighbours
- After fitting each model with the optimal parameters, it was found that there were not enough samples in the data to distinguish between each model; each model had 83.3% accuracy on the test data

Introduction

- SpaceX has cheaper rocket launches than their competition because they can reuse the first stage of their rockets
- If we can predict the success of the first stage landing, we can determine the cost of a launch and therefore bid more accurately against SpaceX for a future launch
- The goal of this capstone project is to be able to develop a classification model to predict the success of a Falcon 9 first stage rocket landing

Source: https://www.space.com/spacex-transporter-2-rocket-landing-tracking-camera-video

Section 1

Methodology

Methodology

Executive Summary

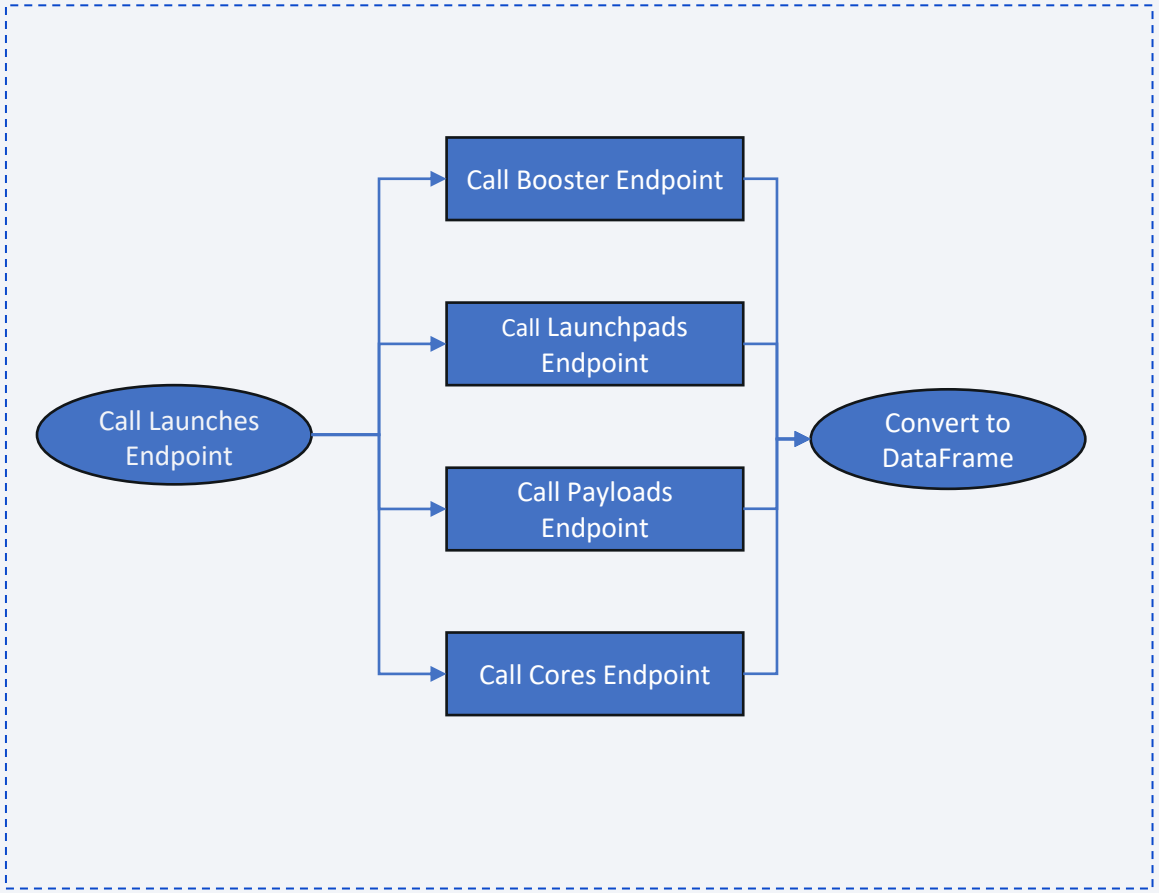
- Data collection methodology:
  - Data was collected using the SpaceX REST API and Wiki page for the SpaceX launches
- Perform data wrangling:
  - Data was processed mainly using the Pandas library
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - 4 different classification models were fit: Logistic Regression, Decision Trees, SVM, KNN

Data Collection

- Data was collected from two sources: SpaceX API and Wikipedia page for SpaceX Falcon 9 launches
- SpaceX API – http://api.spacexdata.com/v4/
- Wikipedia List of Falcon 9 Launches (archived June 2021): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

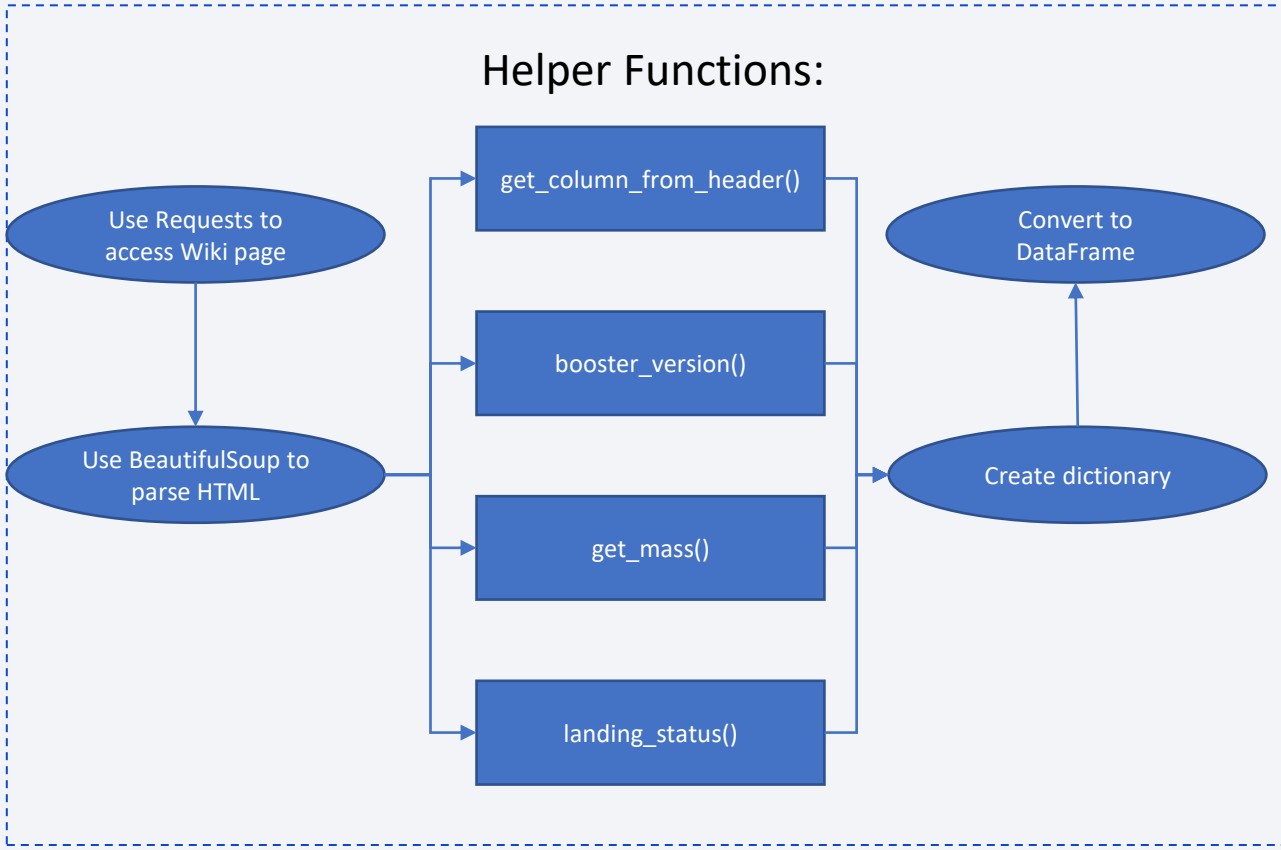
Data Collection – SpaceX API

- Make a call to SpaceX REST API (launches endpoint)
- Make API calls to other endpoints to replace IDs of boosters, launchpads, payloads, cores, etc with actual names
- Jupyter Notebook:
  https://github.com/smbauer/ibm_ds_capstone/blob/master/Data%20Collection%20API.ipynb



Data Collection - Scraping

- Use the Requests library to access the Wiki page
- Use the BeautifulSoup library to parse HTML tables
- Use helper functions to further parse the HTML tables into a Python dictionary
- Convert dictionary into Pandas DataFrame
- Jupyter Notebook: https://github.com/smbauer/ibm_ds_capstone/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Use the API again with different endpoints to replace ID numbers with actual data/names
- Remove all data that doesn't belong to Falcon 9 launches (ex remove Falcon 1 data)
- Remove nulls – replace empty values for payload mass with the mean value
- Jupyter Notebook: https://github.com/smbauer/ibm_ds_capstone/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- Created 6 charts during the exploratory data analysis stage:
  - Flight Number vs Launch Site – to see how success varies over time by launch site
  - Payload vs Launch Site – to see how success varies based on payload size
  - Success Rate vs Orbit Type
  - Flight Number vs Orbit Type – shows how success varies over time by orbit type
  - Payload vs Orbit Type – shows how successful each payload size is for each orbit type
  - Launch Success Yearly Trend – shows success rate by year
- Jupyter Notebook:
  https://github.com/smbauer/ibm_ds_capstone/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

- Loaded data into DB2 database and performed the following queries:
  - Find the names of the unique launch sites
  - Calculate the total payload carried by boosters from NASA
  - Calculate the average payload mass carried by booster version F9 v1.1
  - Find the dates of the first successful landing outcome on ground pad
  - Find the boosters which have successfully landed on drone ship and had payload mass between 4000-6000kg
  - Find 5 records where launch sites begin with `CCA`
  - Calculate the total number of successful and failure mission outcomes
  - List the names of the booster which have carried the maximum payload mass
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank Landing Outcomes Between 2010-06-04 and 2017-03-20
- Jupyter Notebook:
  https://github.com/smbauer/ibm_ds_capstone/blob/master/Exploratory%20Data%20Analysis.ipynb

Build an Interactive Map with Folium

- Created 3 interactive maps of the launch sites using Folium library:
  - First map shows zoomed out version of all 4 launch sites to get the big picture of their location
  - Second map shows zoomed in version of the launch sites with each launch marked and colour-coded by result (success/failure)
  - Third map shows distance markers from launch sites to areas of interest (such as coastline, railroads, etc)
- Jupyter Notebook: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/1c080b7d-ed84-4da6-a80b-84bb99f61367/view?access_token=d9d0ea8082fcf3fe2b3a874267f03b5f9439bbf831a001622be4851e9996987e

Build a Dashboard with Plotly Dash

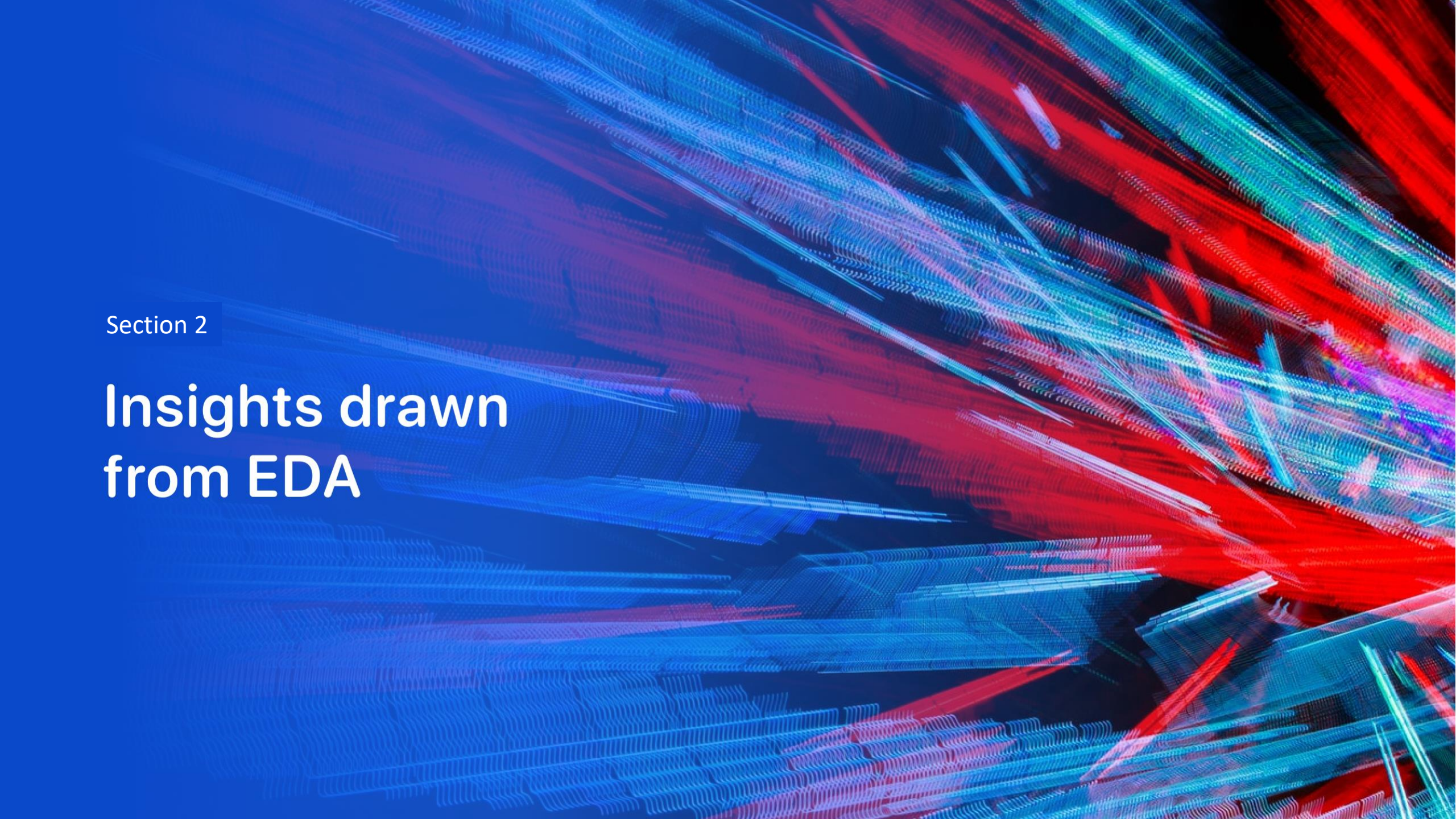
- Components include a pie chart which shows the percentage of successful launches by launch site, and scatter chart that shows the relationship between payload size and launch success
- Dashboard allows user to filter both plots by launch site as well as selecting a range of payload sizes in the scatter chart
- These plots highlight two of the most important factors in launch success; payload size and launch site
- GitHub URL: https://github.com/smbauer/ibm_ds_capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Fit 4 different classification models: Logistic Regression, Decision Trees, Support Vector Machine, K-Nearest Neighbours
- The first step was to standardize the data and split it into training and testing datasets (90 total samples – 72 training, 18 test)
- For each model, used GridSearchCV to find the best performing hyperparameters and used those to fit the models using the training dataset
- Calculate the accuracy on the test data using the method Score
- Produce a Confusion Matrix for each model
- Jupyter Notebook:
  https://github.com/smbauer/ibm_ds_capstone/blob/master/Machine%20Learning%20Prediction.ipynb

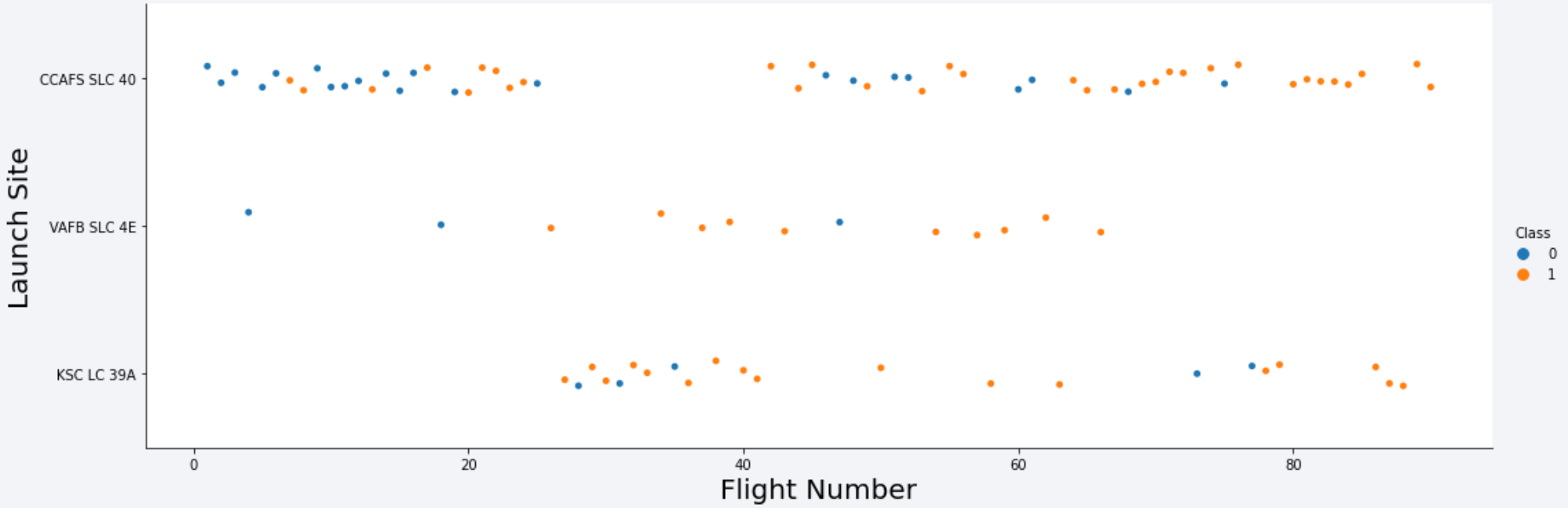
Results

- Exploratory data analysis results
  - Success Rate improves over time
- Interactive analytics demo (see screenshots in next section)
  - Can see that all launch sites are in close proximity to coastline, railroads, and highways
  - Launch site KSC LC-39A has the highest success rate of all launch sites
- Predictive analysis results:
  - Each of the 4 classification models had the same accuracy score (83.33%)
  - Looking at the Confusion Matrix (shown in next section), each model produced 3 false positives


Section 2
Insights drawn from EDA

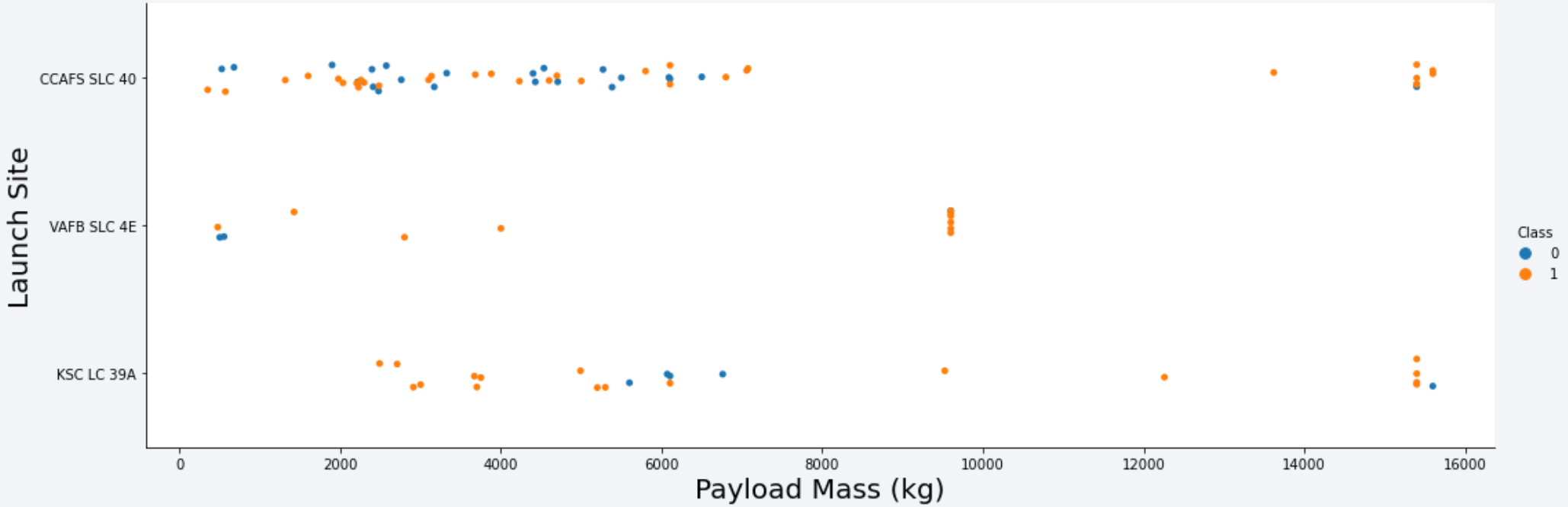
Flight Number vs. Launch Site

- You can see that success rate improves over time for each launch site



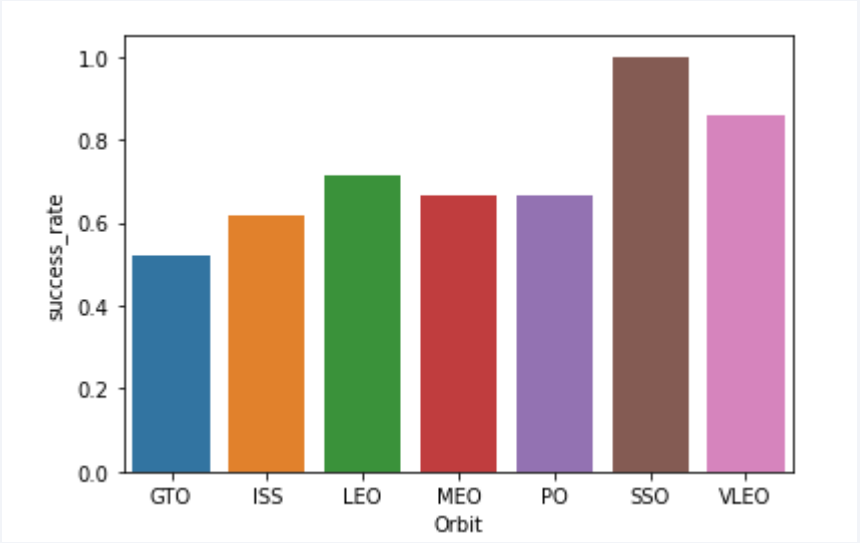
Payload vs. Launch Site

- Looking at the scatter plot below, it appears success rate increases with heavier payloads
- However, once you observe other charts, it's more likely SpaceX only used heavier payloads in later flights, once they have corrected mistakes from earlier launches



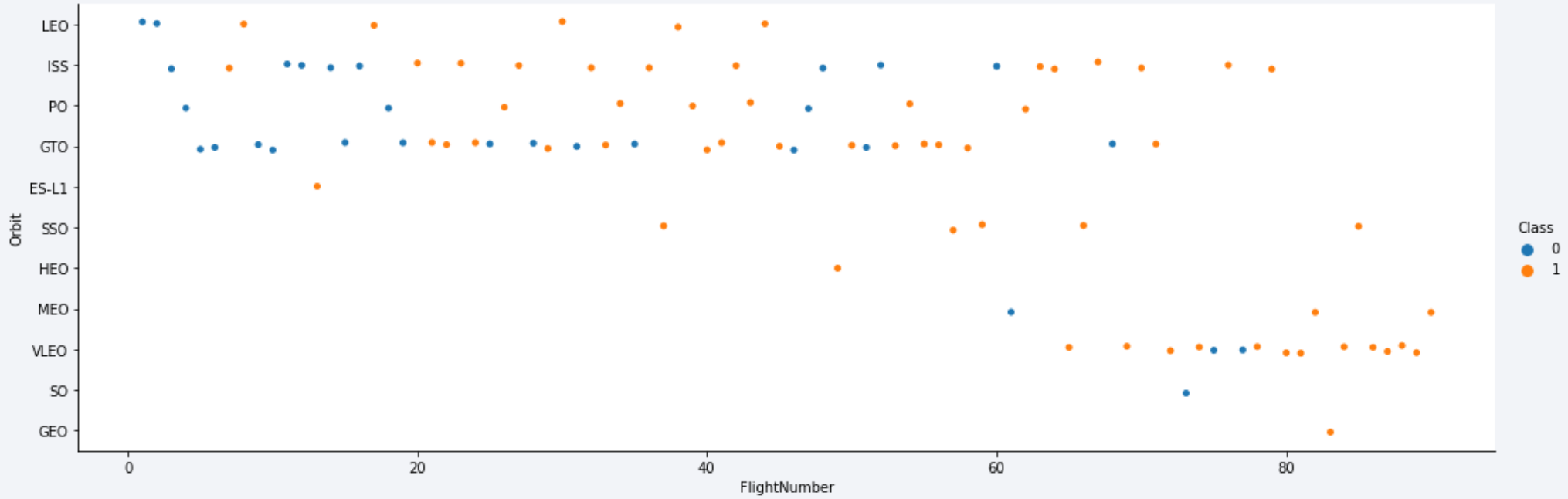
Success Rate vs. Orbit Type

- Once you remove the orbits with only one launch, the SSO orbit has the highest success rate (100%)



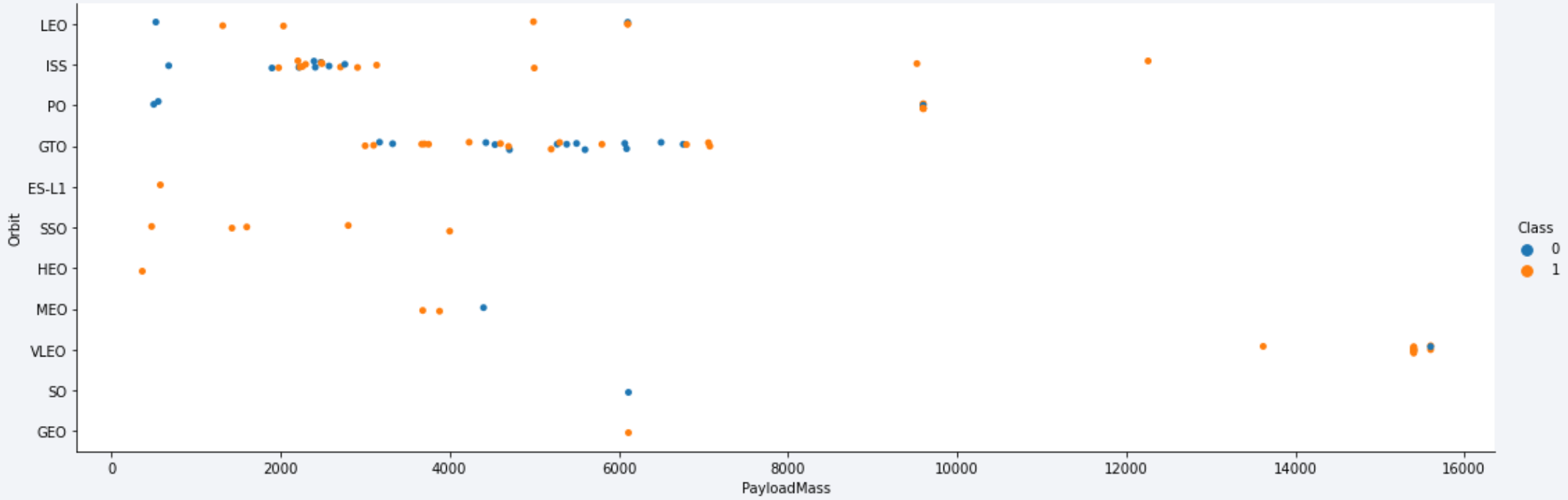
Flight Number vs. Orbit Type

- SSO and VLEO orbits have the highest success rate, but you can also see that most of those launches took place more recently
- Time still appears to have the highest impact on success rate



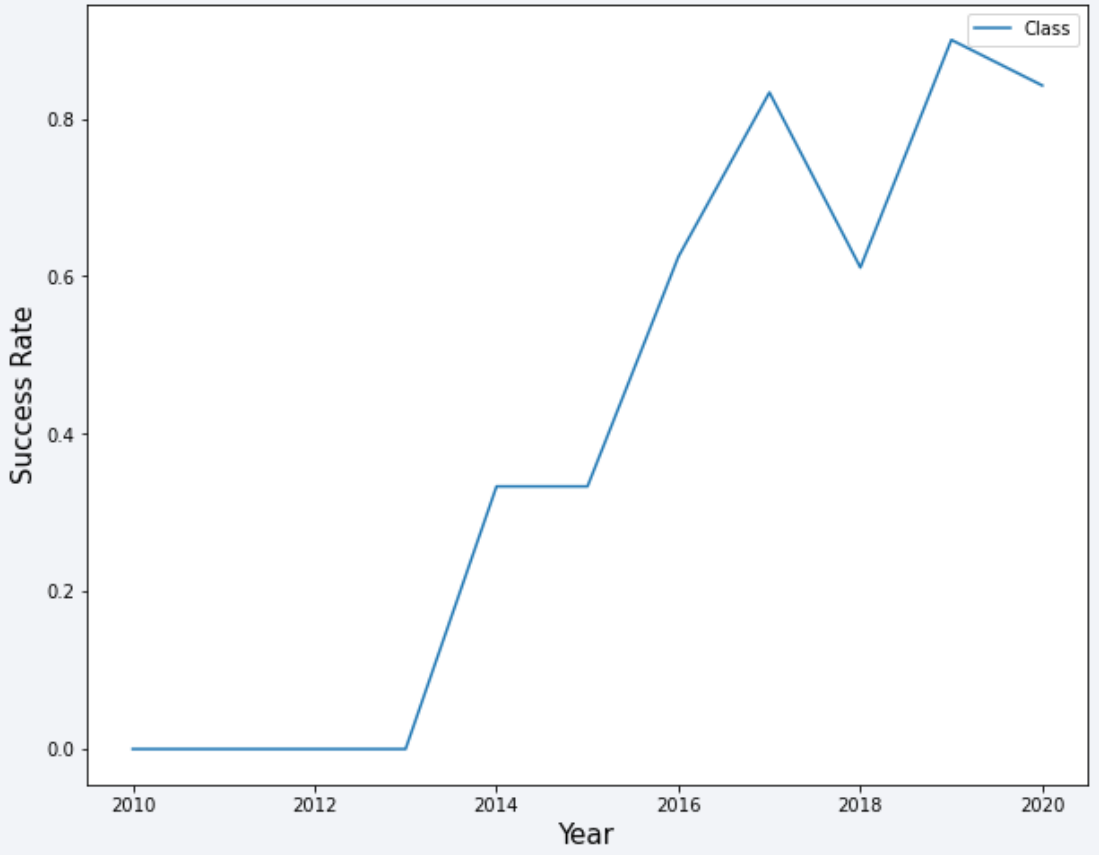
Payload vs. Orbit Type

- It appears certain orbits are better suited for different payload sizes



Launch Success Yearly Trend

- This plot strongly supports the earlier assumption that success rate increases over time



All Launch Site Names

- Find the names of the unique launch sites
- SpaceX used 4 different launch sites

```
%%sql
SELECT DISTINCT(launch_site)
FROM spacex
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

```
SELECT *
FROM spacex
WHERE launch_site LIKE 'CCA%'
LIMIT 5;
```

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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- The total payload carried by NASA boosters was 45,596kg

```
SELECT SUM(payload_mass__kg_)
FROM spacex
WHERE customer='NASA (CRS)'
```

1
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload carried by booster version F9 v1.1 was 2928kg

```
SELECT AVG(payload_mass__kg_)
FROM spacex
WHERE booster_version='F9 v1.1'
```

1
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The first successful ground pad landing was on Dec 22, 2015

```
SELECT MIN(DATE)
FROM spacex
WHERE landing__outcome='Success (ground pad)'
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000 kg

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The F9 B4, B5, and FT boosters successfully landed on drone ships with a payload mass in that range

```
SELECT DISTINCT(booster_version)
FROM spacex
WHERE (landing__outcome='Success (drone ship)') & (4000 < payload_mass__kg_ < 6000)
```

booster_version
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1
F9 FT B1029.2
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1038.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Almost all the missions were successful; only one mission failed in flight

```
SELECT mission_outcome, COUNT(mission_outcome) AS num_outcomes
FROM spacex
GROUP BY mission_outcome
```

mission_outcome	num_outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- The F9 B5 booster is the only one which has carried the max payload mass

```
SELECT DISTINCT(booster_version)
FROM spacex
WHERE payload_mass__kg_ = (
    SELECT MAX(payload_mass__kg_) FROM spacex
    )
;
```

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

2015 Launch Records

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- There were two failed landings on the drone ship in 2015; both were from launch site CCAFS LC-40 with booster version F9 v1.1

```
SELECT YEAR(DATE) AS year_, booster_version, landing__outcome, launch_site
FROM spacex
WHERE (landing__outcome='Failure (drone ship)') & (YEAR(DATE)=2015)
```

year_	booster_version	landing__outcome	launch_site
2015	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
2015	F9 v1.1 B1015	Failure (drone ship)	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
- There was a wide variation of landing outcomes during that time frame

```
SELECT landing__outcome, COUNT(landing__outcome) AS num_outcomes
FROM spacex
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY num_outcomes DESC;
```

landing__outcome	num_outcomes
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

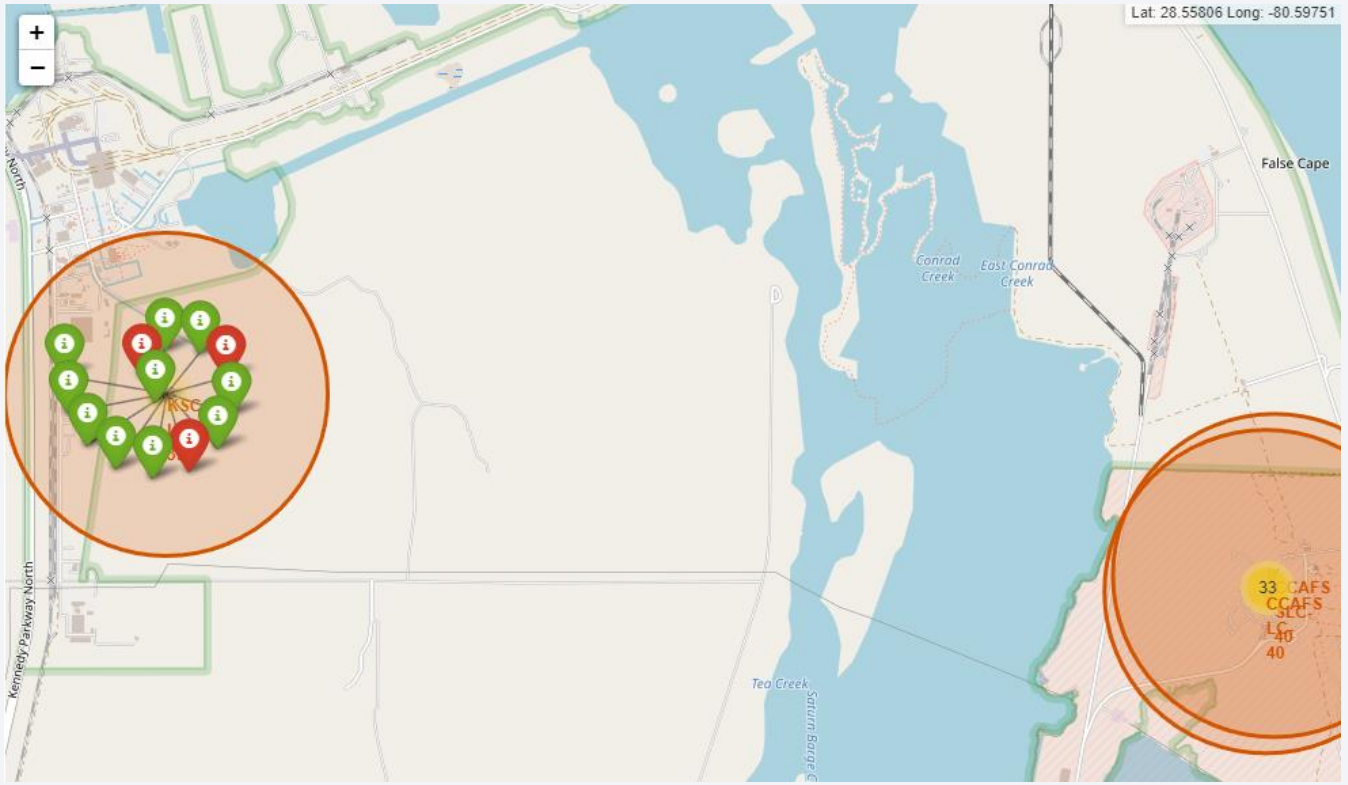
Launch Sites Proximities Analysis

Folium Map 1 – All Launch Sites

- The following screenshot shows the locations of all 4 launch sites (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)

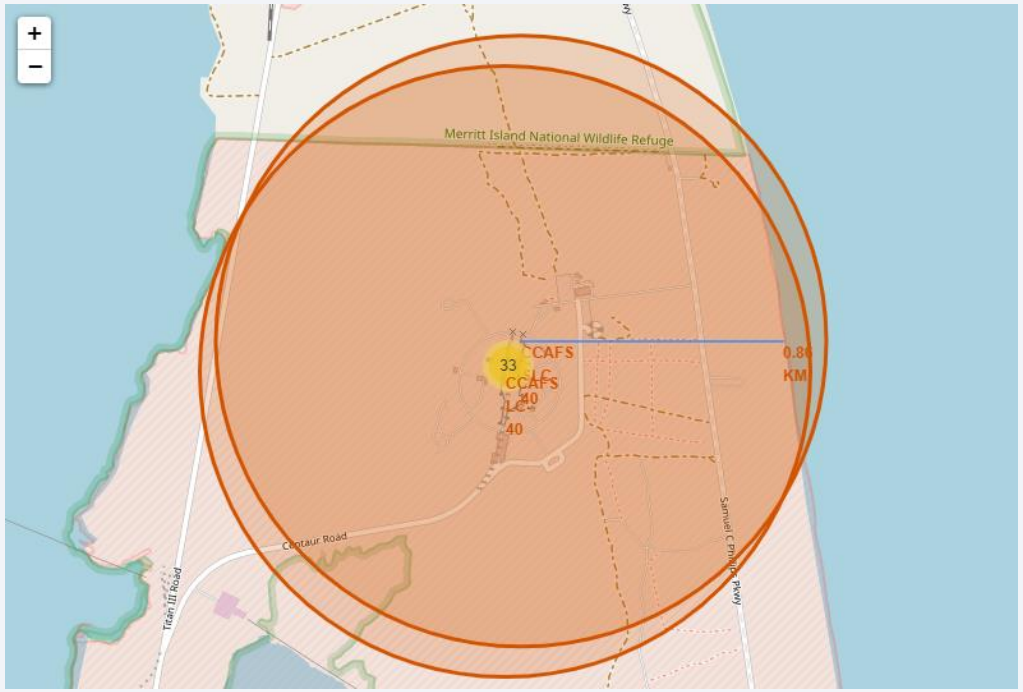
Folium Map 2 – Successful/Failed Launches for Each Site

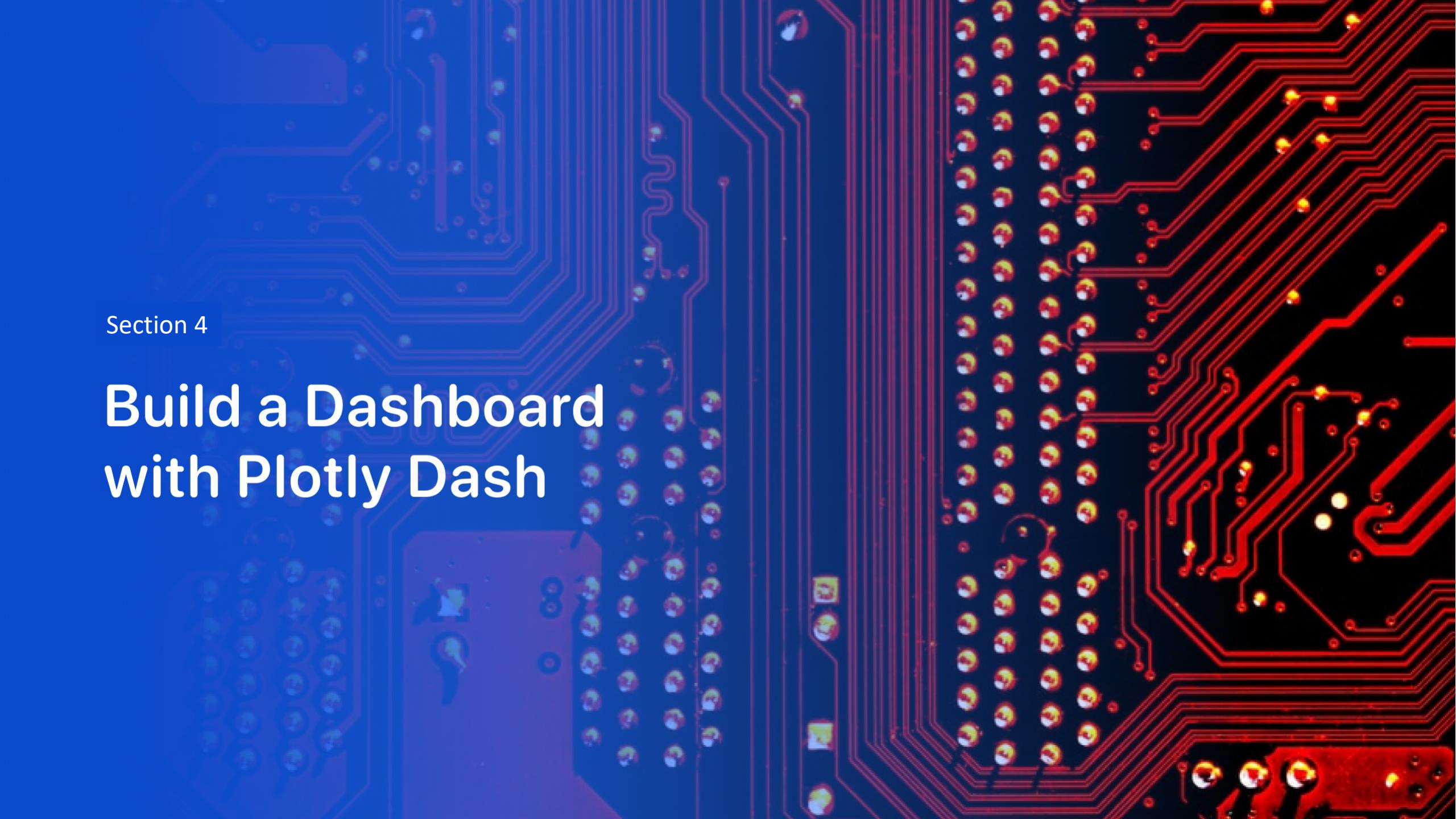
- The following screenshot shows the clustered launches for site KSC LC-39A colour-coded by launch outcome (green = success, red = failure)



Folium Map 3 – Launch Site Proximity to Coastline

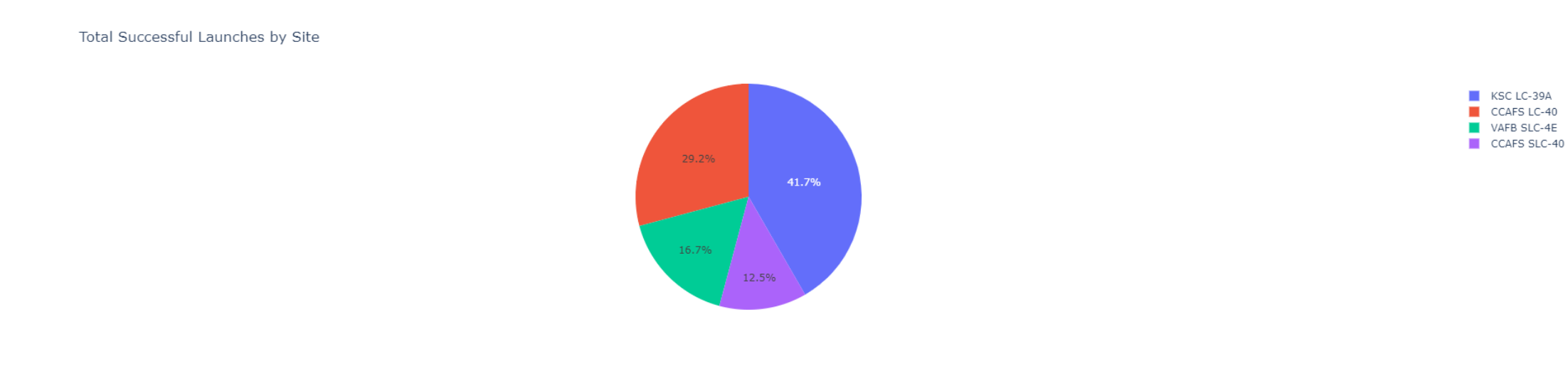
- The following screenshot shows the proximity from the CCAFS launch sites to the nearest coastline (0.86 km)
- All launch sites are in close proximity to coastline, railroads, and highways




Section 4
Build a Dashboard with Plotly Dash

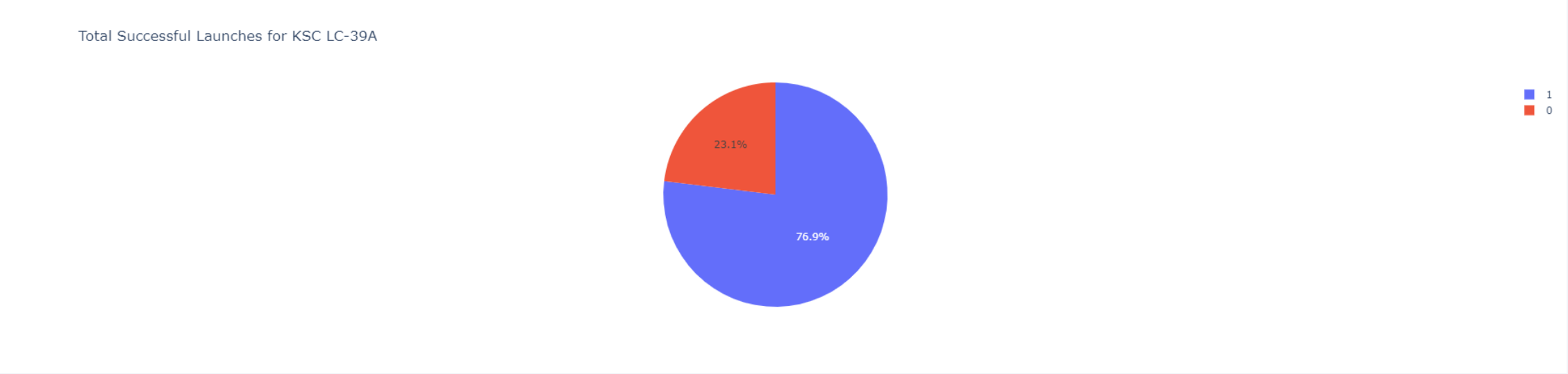
Total Successful Launches by Site

- The following pie chart shows the breakdown of successful launches as a percentage by launch site
- KSC LC-39A has the highest percentage of successful launches



Pie Chart for Site with Highest Launch Success Ratio

- Site KSC LC-39A has the highest launch success ratio of all sites
- You can see from the following pie chart that 76.9% of the launches from that site are successful (while 23.1% have failed)



Scatter Plot of Payload Mass vs Success Rate (All Sites)

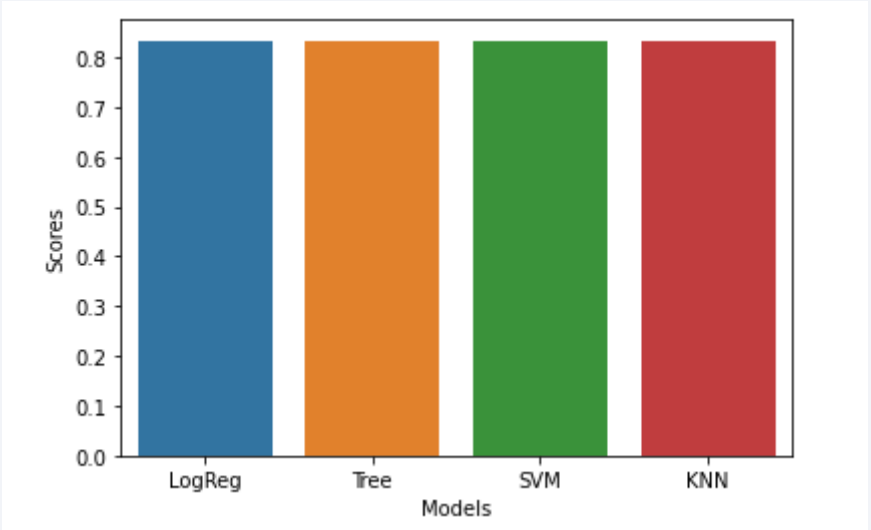
- You can see from the following scatter plot that the FT booster version appears to have the highest success rate and that the payload mass range with the highest success rate is approximately 2000-4000kg

Section 5

Predictive Analysis (Classification)

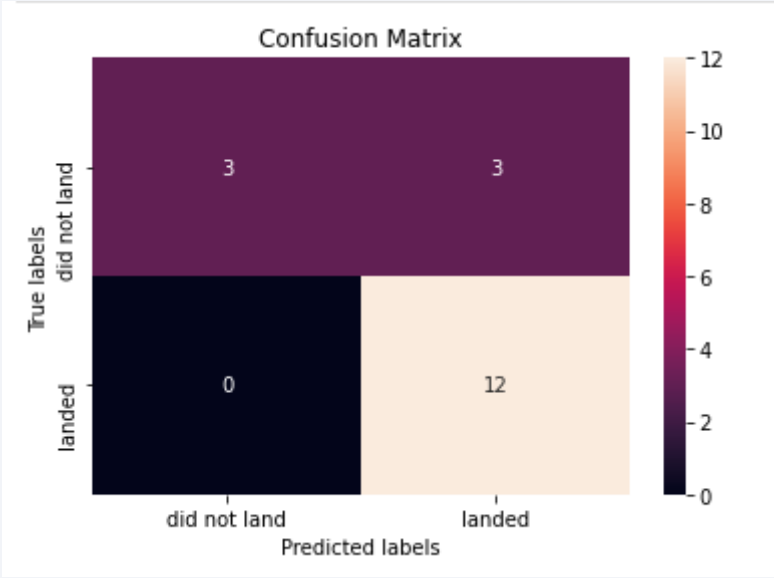
Classification Accuracy

- All 4 classification models have the exact same accuracy (83.33%)
- Not enough data to differentiate (only 18 test set samples)



Confusion Matrix

- The confusion matrix is the exact same for each of the 4 models tested
- The model can distinguish between different classes, but has an issue with false positives (3 predicted to have landed, which did not actually land)



Conclusions

- Time (flight number) has by far the largest impact on improving success rate; it can be assumed that the more flights run by SpaceX, the better they become at fixing earlier mistakes
- Either of the 4 classification models (logistic regression, decision trees, SVM, KNN) are sufficient for predicting future success as each model shows the same accuracy on the available data
- To improve model performance, we need to collect more launch data or perhaps experiment with more advanced machine learning models and methods

Appendix

- See links to individual Jupyter Notebooks posted within previous sections for detailed code snippets and visualizations

Thank you!