IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sarit Hiranyaphinant
14 Aug 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Link to Project Repo

- Summary of methodologies
  - Data Collection via SpaceX API and Webscraping from HTML tables using BeautifulSoup;
  - Exploratory Data Analysis (EDA) with Visualization and SQL Queries
  - Interactive Map with Folium
  - Dashboards with Plotly Dash
  - Predictive Analysis with Classification Models
- Summary of all results
  - Landing success rate improves over time
  - Launch site with the highest success rate is KSC LC-39A
  - Orbits with the highest success rate are ES-LS1, GEO, HEO, SSO
  - All 4 classifiers yield the same accuracy score on test data (83.33%) with decision tree classifier scoring the highest on train data (88.92%)

Introduction

- Project background and context
  - SpaceY, a new fictional player in the aerospace industry, is speculating whether it is viable for them to compete with SpaceX reusable rockets. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch as well as the conditions of a successful launch. This information will be the thing that gives SpaceY the edge it needs to compete with SpaceX.
- Problems you want to find answers
  - What are the key characteristics of a successful landing?
  - What are the relationships among independent variables and their effects on the landing success?

Section 1

Methodology

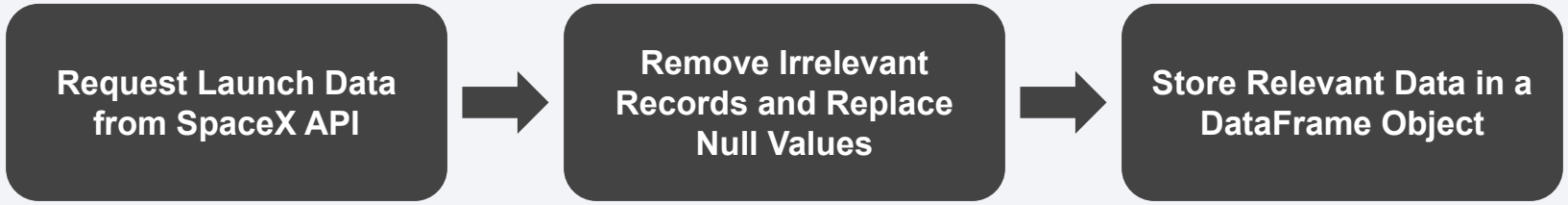
Methodology

Executive Summary

- Data collection methodology:
  - Collected initial data from SpaceX API via REST calls then use column-specific IDs to make further requests from specific endpoints and putting response data into a pandas DataFrame
  - Collected further data from an HTML table from a Wikipedia page via webscraping using BeautifulSoup
- Perform data wrangling
  - Transformed multiple categorical values of landing outcomes into a numerical binary variable of either fail (0) or success (1) which would become the target label for training the 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
  - Built and trained 4 different types (logistic regression, decision tree, svm, and knn) of classification models with multiple hyperparameters using GridSearch from scikit-learn and cross-validation technique for best result
  - Compared the 4 classifiers and selected the one with highest average accuracy, jaccard, and f1 scores on the test data.

Data Collection

- Collected initial data from SpaceX API via REST calls then use column-specific IDs to make further requests from specific endpoints and storing the data into a pandas DataFrame after removing irrelevant records and replacing null values



- Collected further data from an HTML table from a Wikipedia page via webscraping using BeautifulSoup

Link to Source Code

Data Collection – SpaceX API

Request initial data from SpaceX API and store in a DataFrame

```
spacex_url = "https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
data = pd.json_normalize(response.json())
```

Remove launch records with multiple cores and payload values

```
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]
```

Use column-specific IDs to make further requests from specific API endpoints. The response data are then turned into a dictionary `launch_dict`

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

Parse `launch_dict` into a DataFrame object for further analysis

```
launch_data = pd.DataFrame.from_dict(launch_dict)
```

Link to Source Code

Data Collection - Scraping

Load and store HTML response in a BeautifulSoup object

```
response = requests.get(static_url)
soup = BeautifulSoup(response.content)
```

Extract column names from table headers and store them in a list

Then use the list to generate a dictionary with those names as keys and initialize the values as empty lists.

```
html_tables = soup.find_all('table')
first_launch_table = html_tables[2]

for row in first_launch_table.find_all('th'):
    header = extract_column_from_head(row)

    if header != None and len(header) > 0:
        column_names.append(header)

launch_dict = dict.fromkeys(column_names)
```

Loop through each row in the HTML table and append each corresponding cell value to the list as value pair to its header key in the dictionary.

Finally, parse the dictionary into a DataFrame Object.

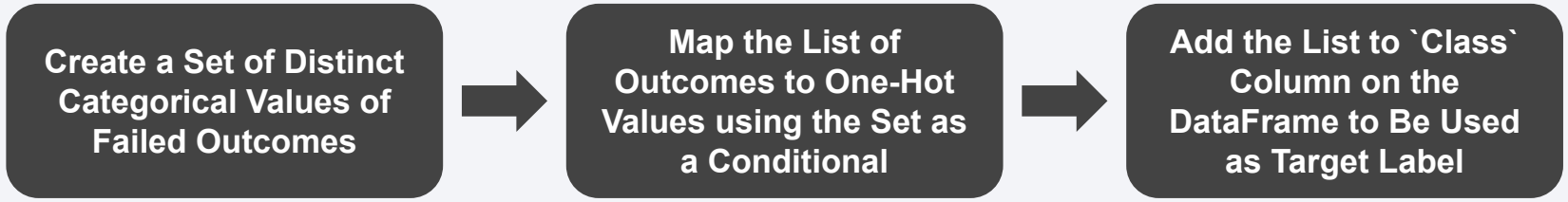
```
...
    launch_site = row[2].a.string
    launch_dict['Launch site'].append(launch_site)
...

df = pd.DataFrame.from_dict(launch_dict)
```

Link to Source Code

Data Wrangling

- Transformed multiple categorical values of landing outcomes into a numerical binary variable of either fail (0) or success (1) which would become the target label for training the model



Link to Source Code

EDA with Data Visualization

Scatter Plots

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Launch Site vs Payload Mass
- Flight Number vs Orbit Type
- Payload Mass vs Orbit Type

Scatter plots allow us to get the initial impression on the correlations between variables as well as clustering of data.

Bar Chart

- Success Rate vs Orbit Type

Bar charts show us the aggregated numerical value of each distinct categorical value.

Line Graph

- Yearly Success Trend

Line graphs show us the trend of a numerical data over a time series.

Link to Source Code

EDA with SQL

SQL Queries Performed:

1. Fetch the names of distinct launch sites
2. List 5 records whose launch site names begin with ‘CCA’
3. Fetch the total payload mass by boosters launched by NASA (CRS)
4. Fetch average payload mass carried carried by booster version F9 v1.1
5. Fetch the date of the first successful ground-pad landing
6. List the names of the boosters with payload mass between 4000–6000 kg that have success in drone-ship landing
7. Fetch the total number of executed mission attempts (both successful & failed)
8. List the names of booster versions which have carried the maximum payload mass
9. List the records displaying the month names, outcomes in drone ship, booster versions, and launch sites during the year 2015
10. Rank the count of successful landings between the date 04-06-2010 and 20-03-2017 in descending order.

Link to Source Code

Build an Interactive Map with Folium

Launch Sites

- Circles for drawing circumferences around 1000-radius-areas of launch sites centered at their coordinates
- Markers for labelling launch site names
- Marker Clusters for interactive display of the count of launches, their individual outcomes color-coded with green for successful landing, and red for failed ones

Key Locations & Distances

- Markers for displaying the distances between the launch site CCAFS SLC-40 and key locations (railway, highway, coastline, city) at their coordinates
- Poly Lines for marking the crow distances between the launch site CCAFS SLC-40 and key locations
- Pop-ups at key locations' distance markers for displaying the type of the location.

Link to Source Code

Build a Dashboard with Plotly Dash

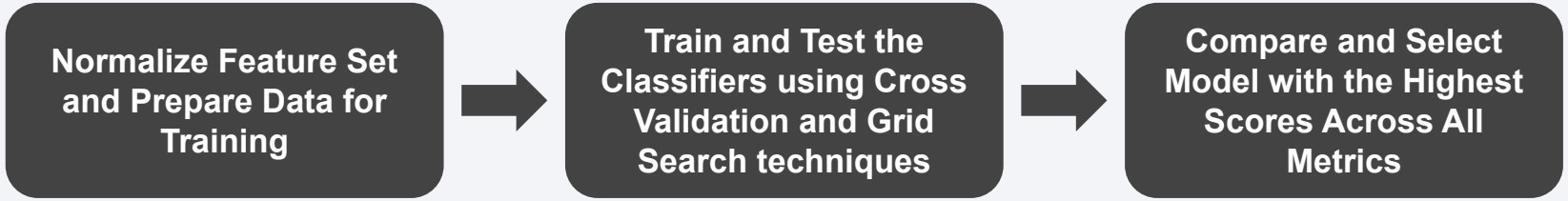
Interactive Dashboard
with Dropdown, Pie Chart, Range Slider, and Scatter Plot

- Dropdown menu allows users to view data from either a specific launch site or all of them
- Pie chart displays the success rate of the individual launch sites intuitively, and the success ratio of each site in comparison among themselves if ALL SITES is selected in the dropdown
- Range slider allows the user to control the range of payload mass of launches plotted on the scatter plot
- Scatter plot displays the relationship between success and payload mass, color-coded by booster versions.

Link to Source Code

Predictive Analysis (Classification)

- Built and trained 4 different types (logistic regression, decision tree, svm, and knn) of classification models with multiple hyperparameters using GridSearch from scikit-learn and cross-validation technique for best result
- Compared the 4 classifiers and selected the one with highest average accuracy, jaccard, and f1 scores on the test data.

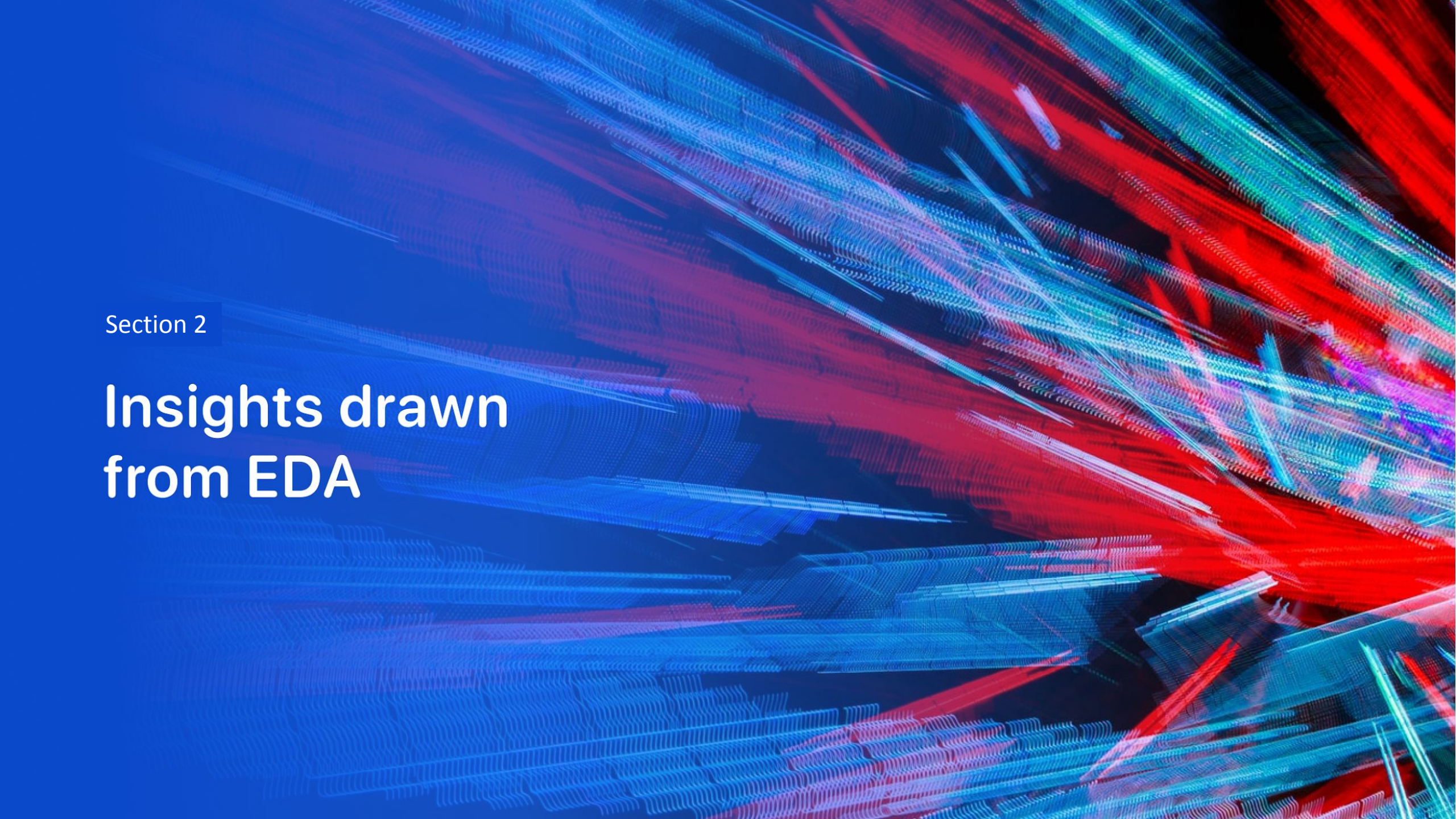


Results

- Landing success rate improves over time
- Launch site with the highest success rate is KSC LC-39A
- Orbits with the highest success rate are ES-LS1, GEO, HEO, and SSO, with SSO being the only one with multiple successful landings
- Launches with payload mass over 9000 kg are more likely to land successfully
  - However, data is scarce and inferences made from payload over 10000 kg aren’t as strong
- The first successful landing happened on 22nd December, 2015
- There are 71 successful and failed landing outcomes in total
- The maximum payload carried by a booster is 15600 kg
- All 4 classifiers perform and yield the exact same accuracy score on test data (83.33%)

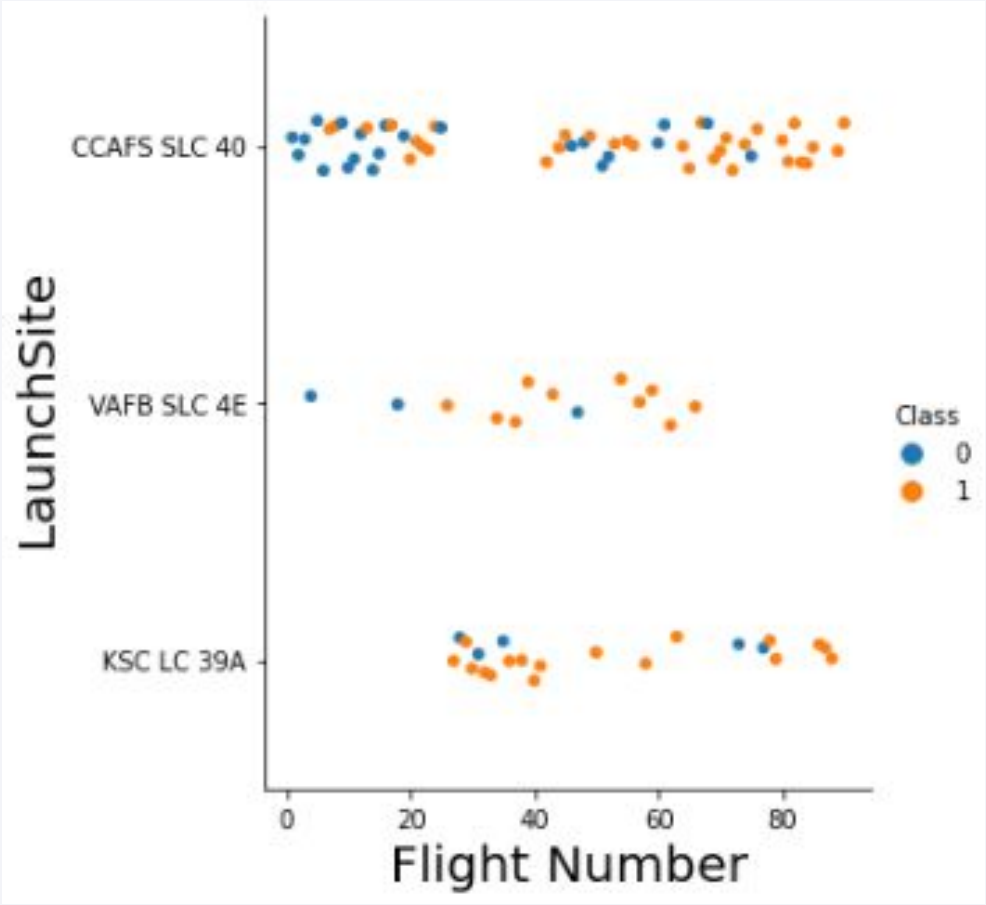
Section 2

Insights drawn from EDA



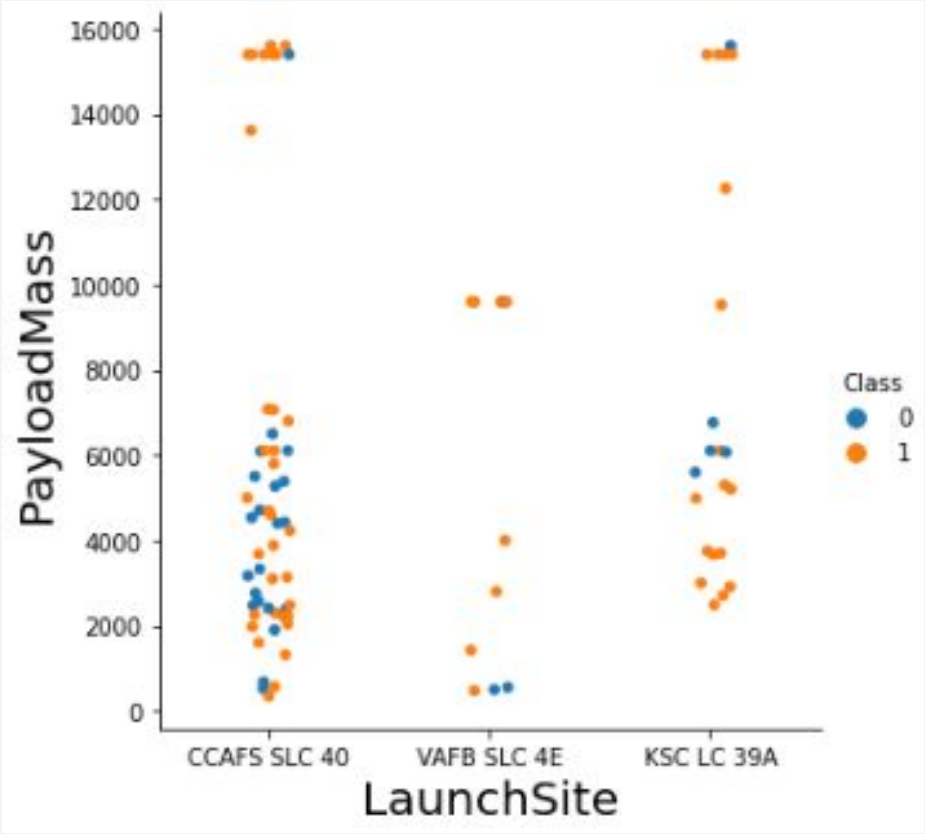
Flight Number vs. Launch Site

- The site CCAFS SLC 40 is the most popular with the highest number of total attempted landings
- The inference can be made from all 3 sites that landing success rate improved over time as class-1 data points skew toward the right of the graph with higher flight number



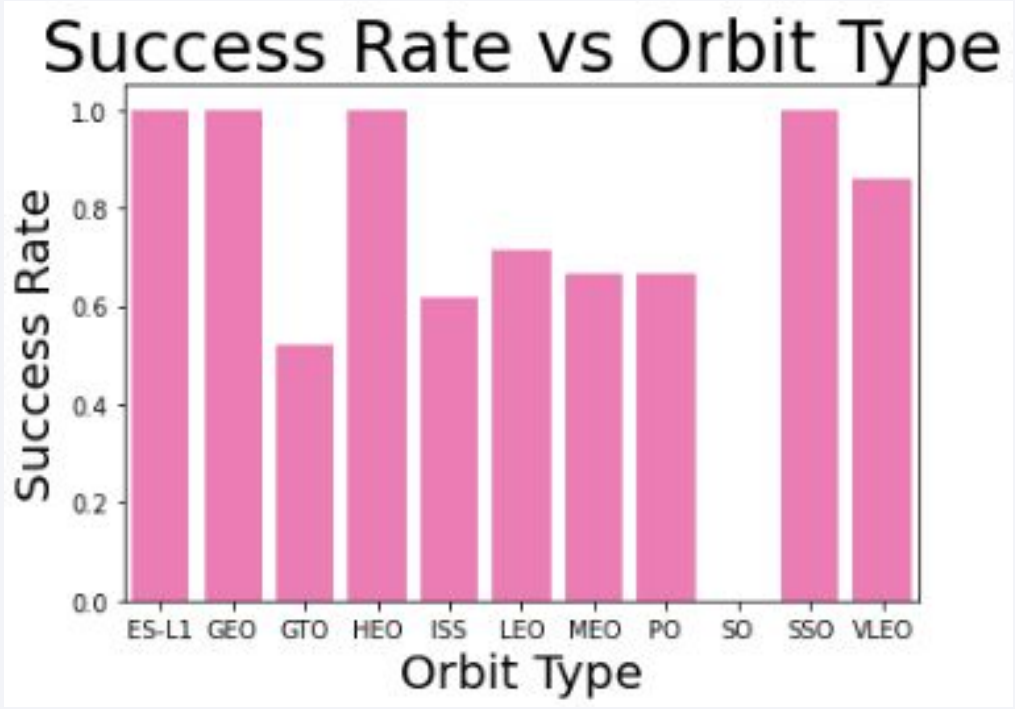
Payload vs. Launch Site

- Launches with payload mass higher than 9000 kg are more likely to land successfully
- The maximum payload carried by boosters launched from site VAFB SLC 4E is around 9500 kg
- Success rate of site KSC LC 39A plummets at the payload range between 5000-7000 kg



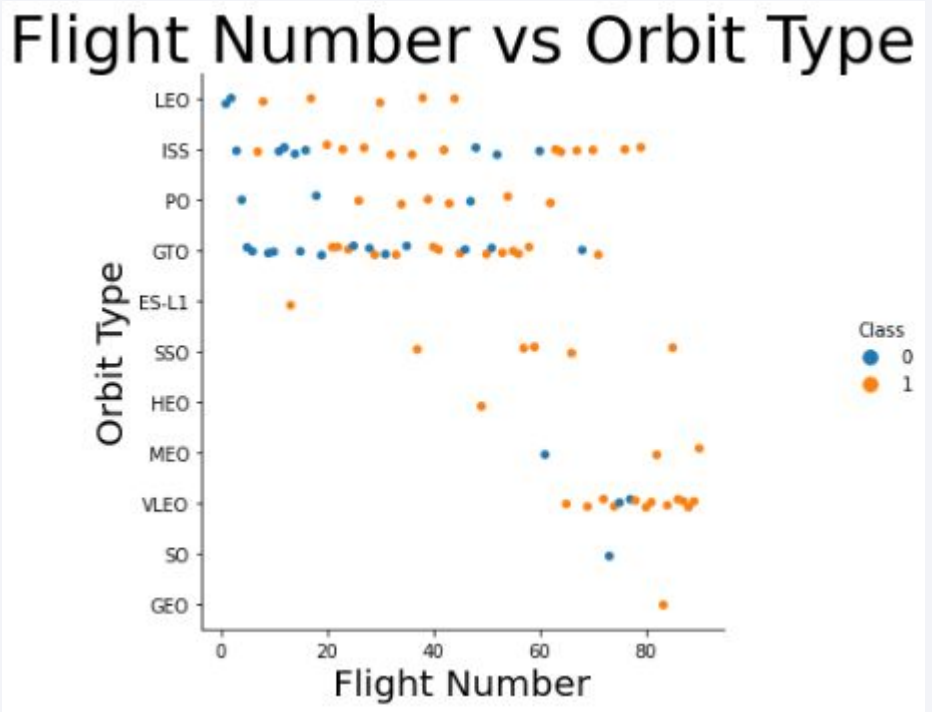
Success Rate vs. Orbit Type

- The orbit types with highest success rate of 100% from observed data are ES-L1, GEO, HEO, and SSO



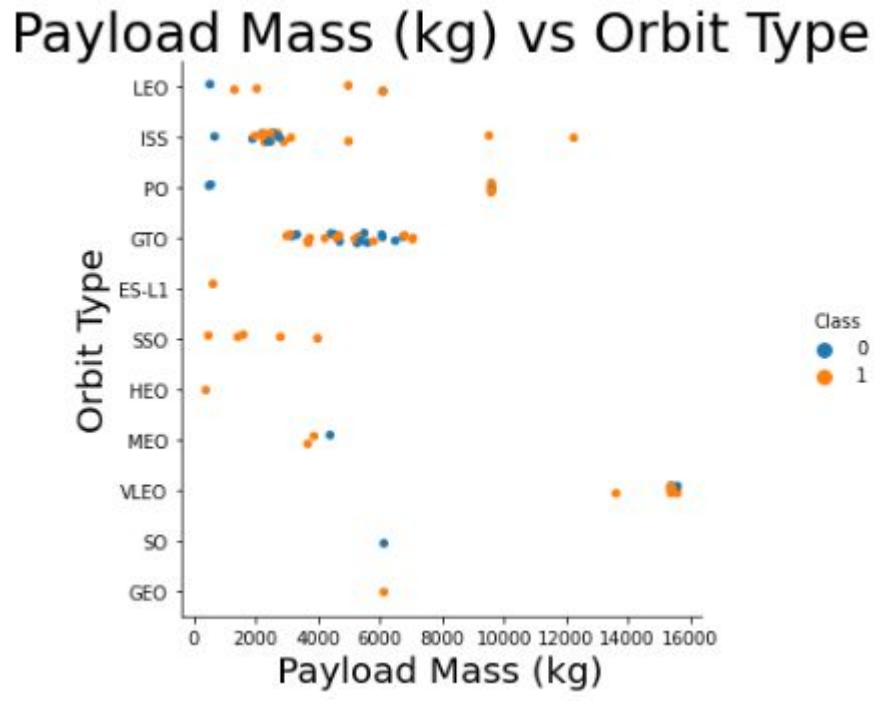
Flight Number vs. Orbit Type

- This plot confirms our previous observation that success rate improves over time as class-1 data points skew toward higher flight number across most orbit types
- Beside SSO, all other orbit types with the highest success rate (ES-L1, GEO, HEO) have only a single successful landing each.
- The most popular orbit type after flight number 60 is VLEO



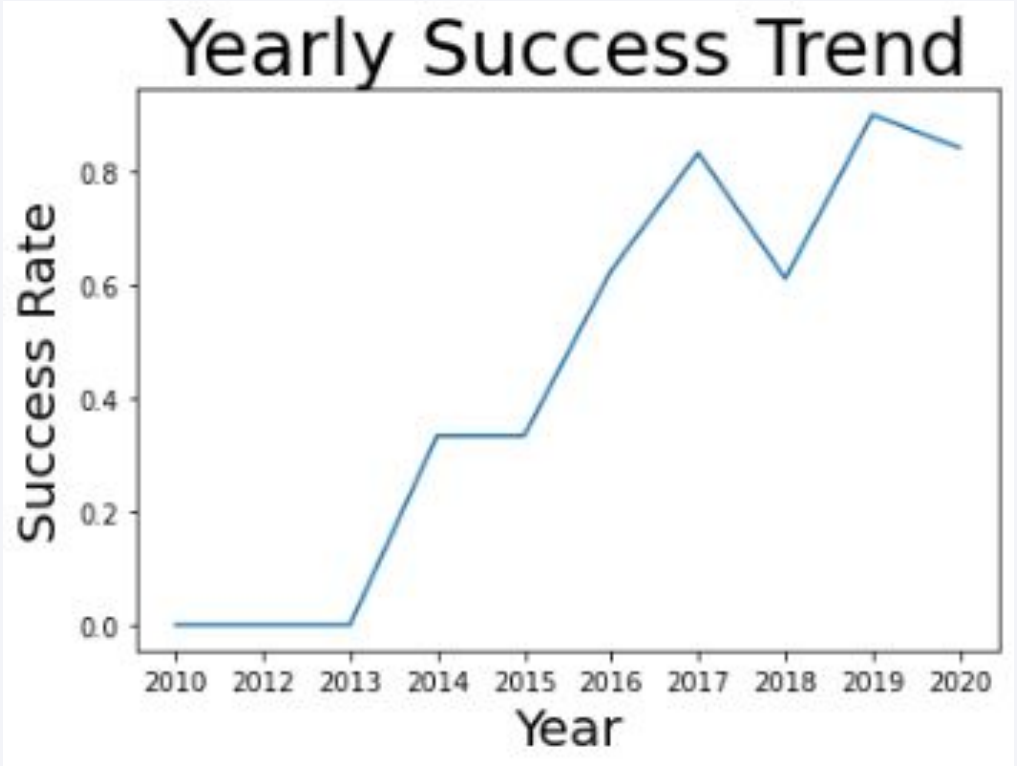
Payload vs. Orbit Type

- For LEO, ISS, and PO orbits, launches with higher payload mass are more likely to land successfully
- All successful landings with SSO orbit have payloads of less than 4000 kg
- Launches with VLEO orbit have a payload higher than 13000 kg



Launch Success Yearly Trend

- Success rate was at 0 up until 2013
- After 2013, the success rate has been increasing at a steady pace
- The success rate took a dive at the year 2018



All Launch Site Names

- Use `DISTINCT` to select only unique values from the `Launch_Site` column
- Names of the unique launch sites
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

```
%%sql
SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
 Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

- Select 5 records whose launch sites begin with `CCA` using string pattern
- 5 records where launch sites begin with CCA

```
[ ] %%sql
    SELECT * FROM SPACEXTBL 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 _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

Total Payload Mass

- Use `SUM` function to calculate the total payload of all launches made by NASA (CRS) as customer
- Total payload carried by boosters from NASA is 45596 kg

```
[ ] %%sql
    SELECT SUM(payload_mass__kg_) AS nasa_crs_payload_total
    FROM spacextbl WHERE "Customer" = "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
nasa_crs_payload_total
45596
```

Average Payload Mass by F9 v1.1

- Use `AVG` function to calculate the mean payload carried by Falcon 9 v1.1
- Average payload mass by F9 v1.1 is 2928.4 kg

```
[ ]  %%sql
     SELECT AVG(payload_mass__kg_) AS "F9 v1.1 Payload Avg."
     FROM spacextbl WHERE "Booster_Version" LIKE "F9 v1.1"

      * sqlite:///my_data1.db
     Done.
     F9 v1.1 Payload Avg.
     2928.4
```

First Successful Ground Landing Date

- Use `LIMIT` clause to select the first successful ground-pad landing
- First successful ground landing date is 22-12-2015

```
[ ] %%sql
    SELECT date AS first_successful_ground_landing_date, "Landing _Outcome"
    FROM spacextbl WHERE "Landing _Outcome" = "Success (ground pad)" LIMIT 1

     * sqlite:///my_data1.db
    Done.
    first_successful_ground_landing_date  Landing _Outcome
    22-12-2015                            Success (ground pad)
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Use `WHERE` clause to specify landing outcome to successful drone-ship landing and payload range to between 4000 and 6000 kg
- List of booster versions which have successfully landed on drone ship with payload within 4000–6000 kg range
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

```
[ ] %%sql
    SELECT DISTINCT "Booster_Version", payload_mass__kg_
    FROM spacextbl
    WHERE
        ("Landing _Outcome" = "Success (drone ship)") AND
        (payload_mass__kg_ BETWEEN 4000 AND 6000)
    ;
```

```
 * sqlite:///my_data1.db
Done.
Booster_Version PAYLOAD_MASS__KG_
F9 FT B1022     4696
F9 FT B1026     4600
F9 FT B1021.2   5300
F9 FT B1031.2   5200
```

Total Number of Successful and Failure Mission Outcomes

- Use `COUNT` function and string pattern to calculate the total successful and failure mission outcomes
- Total executed landing attempts is 71

```
[ ] %%sql
    SELECT COUNT(*) AS total_executed_attempts
    FROM spacextbl
    WHERE
        "Landing _Outcome" LIKE "Success%" OR
        "Landing _Outcome" LIKE "Failure%"
    ;
```

```
 * sqlite:///my_data1.db
Done.
total_executed_attempts
71
```

Boosters Carried Maximum Payload

- Use `DISTINCT` to select unique booster versions that have carried the maximum payload mass
- List of boosters with max payload

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

```
[ ] %%sql
    SELECT DISTINCT "Booster_Version" AS boosters_with_max_payload, payload_mass__kg_
    FROM spacextbl
    WHERE payload_mass__kg_ = (
        SELECT MAX(payload_mass__kg_)
        FROM spacextbl
    );
```

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

boosters_with_max_payload	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

2015 Launch Records

- Use `CASE` and `SUBSTR` function to replace date format with month names
- Months with failed drone-ship landing
  - January
  - April

```
[ ] %%sql
    SELECT
        (CASE
            WHEN SUBSTR(date, 4, 2) = "01" THEN "January"
            WHEN SUBSTR(date, 4, 2) = "02" THEN "February"
            WHEN SUBSTR(date, 4, 2) = "03" THEN "March"
            WHEN SUBSTR(date, 4, 2) = "04" THEN "April"
            WHEN SUBSTR(date, 4, 2) = "05" THEN "May"
            WHEN SUBSTR(date, 4, 2) = "06" THEN "June"
            WHEN SUBSTR(date, 4, 2) = "07" THEN "July"
            WHEN SUBSTR(date, 4, 2) = "08" THEN "August"
            WHEN SUBSTR(date, 4, 2) = "09" THEN "September"
            WHEN SUBSTR(date, 4, 2) = "10" THEN "October"
            WHEN SUBSTR(date, 4, 2) = "11" THEN "November"
            WHEN SUBSTR(date, 4, 2) = "12" THEN "December"
        END) AS month,
        "Landing _Outcome",
        "Booster_Version",
        "Launch_Site"
    FROM spacextbl
    WHERE
        "Landing _Outcome" = "Failure (drone ship)" AND
        SUBSTR(Date, 7, 4) = "2015"
    LIMIT 10;
```

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

month	Landing _Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Use `COUNT` function, comparisons, and string pattern to list the counts of successful landings between 04-06-2010 and 20-03-2017
- There are a total of 34 success landings, 8 in drone ship, 6 in ground pad, and 20 of unspecified type

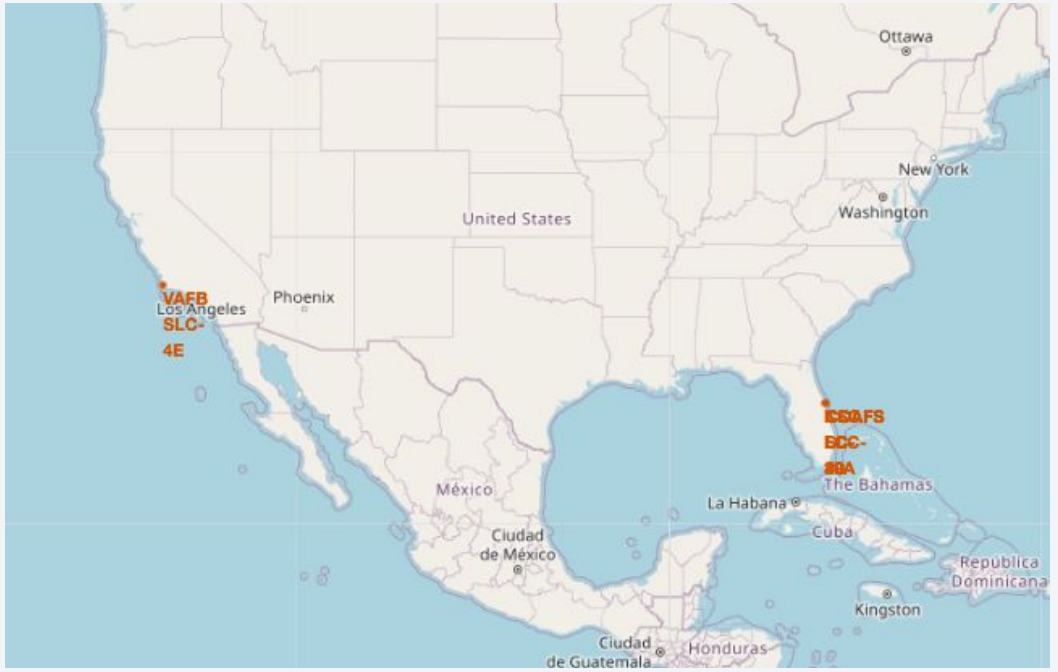
```
[18] %%sql
     SELECT "Landing _Outcome", COUNT("Landing _Outcome")
     FROM spacextbl
     WHERE
         date >= "04-06-2010"
         AND date <= "20-03-2017"
         AND "Landing _Outcome" LIKE "%Success%"
     GROUP BY "Landing _Outcome"
     ORDER BY COUNT("Landing _Outcome") DESC;

      * sqlite:///my_data1.db
     Done.
      Landing _Outcome  COUNT("Landing _Outcome")
     Success                20
     Success (drone ship)   8
     Success (ground pad)   6
```


Section 3
Launch Sites
Proximities Analysis

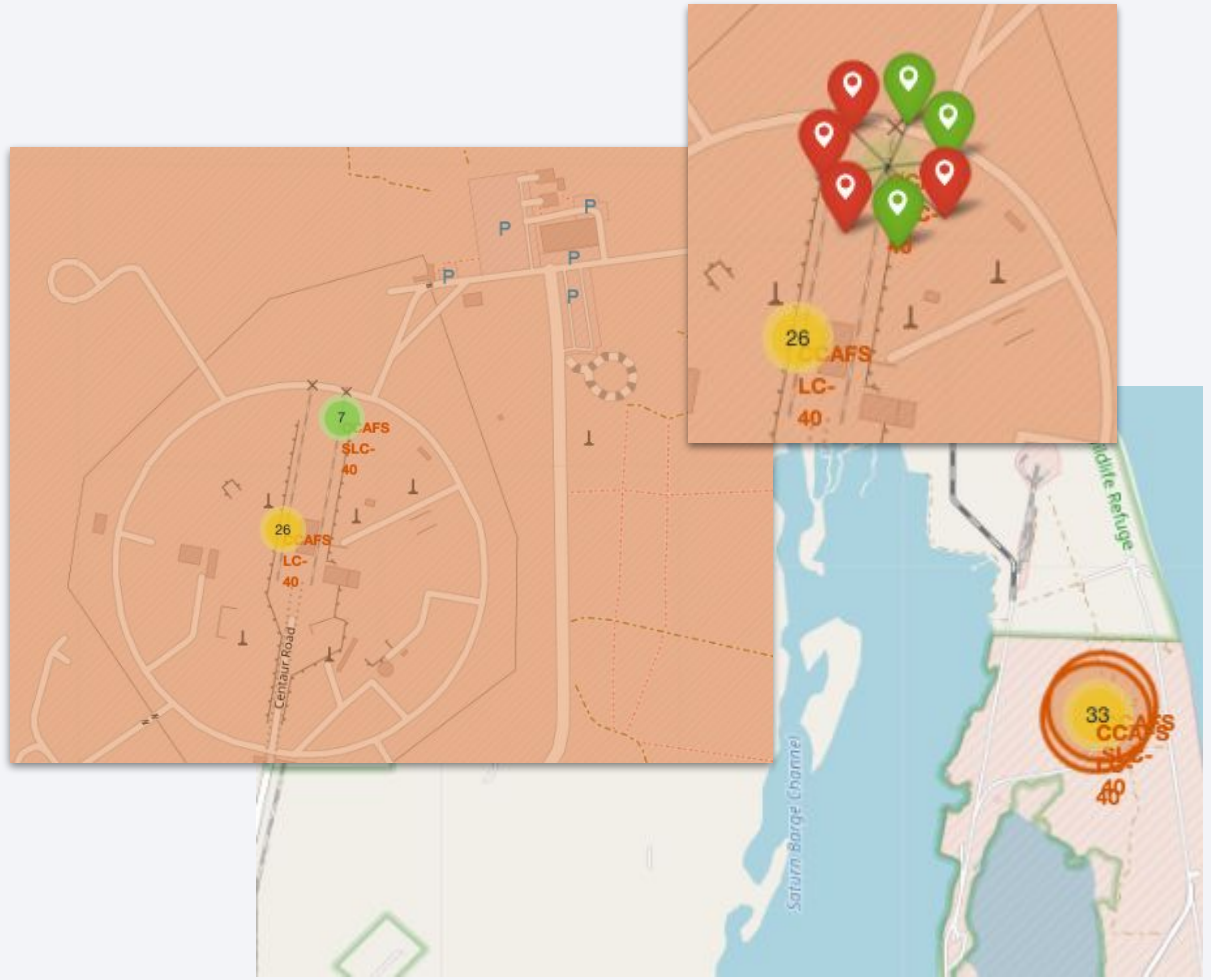
All Launch Sites in USA

- All 4 launch sites are located near the coastal ends of the US; 1 on the west coast and the other 3 on the east coasts
- All sites are considerable far from big cities but still in accessible range to railways and highways



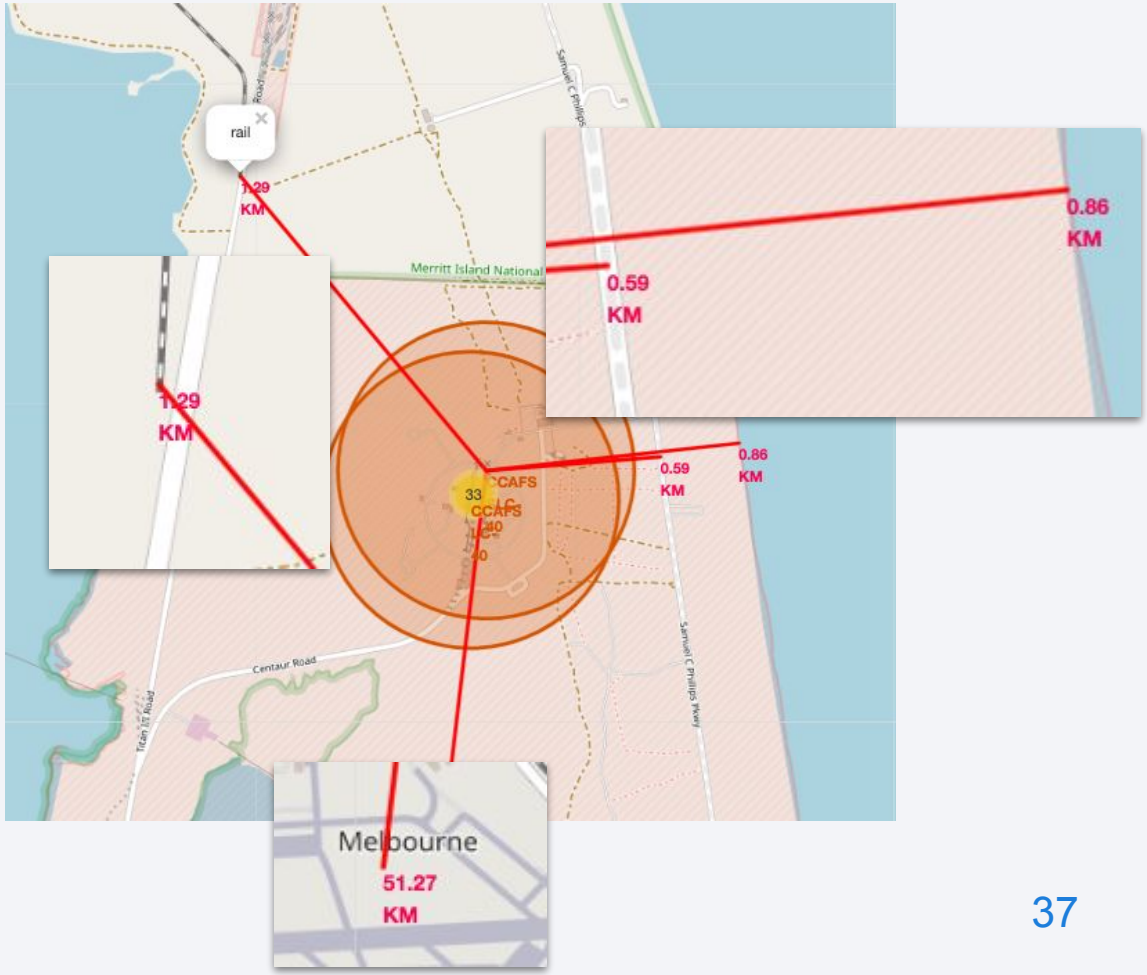
Landing Outcomes by Site & Color Coded Markers

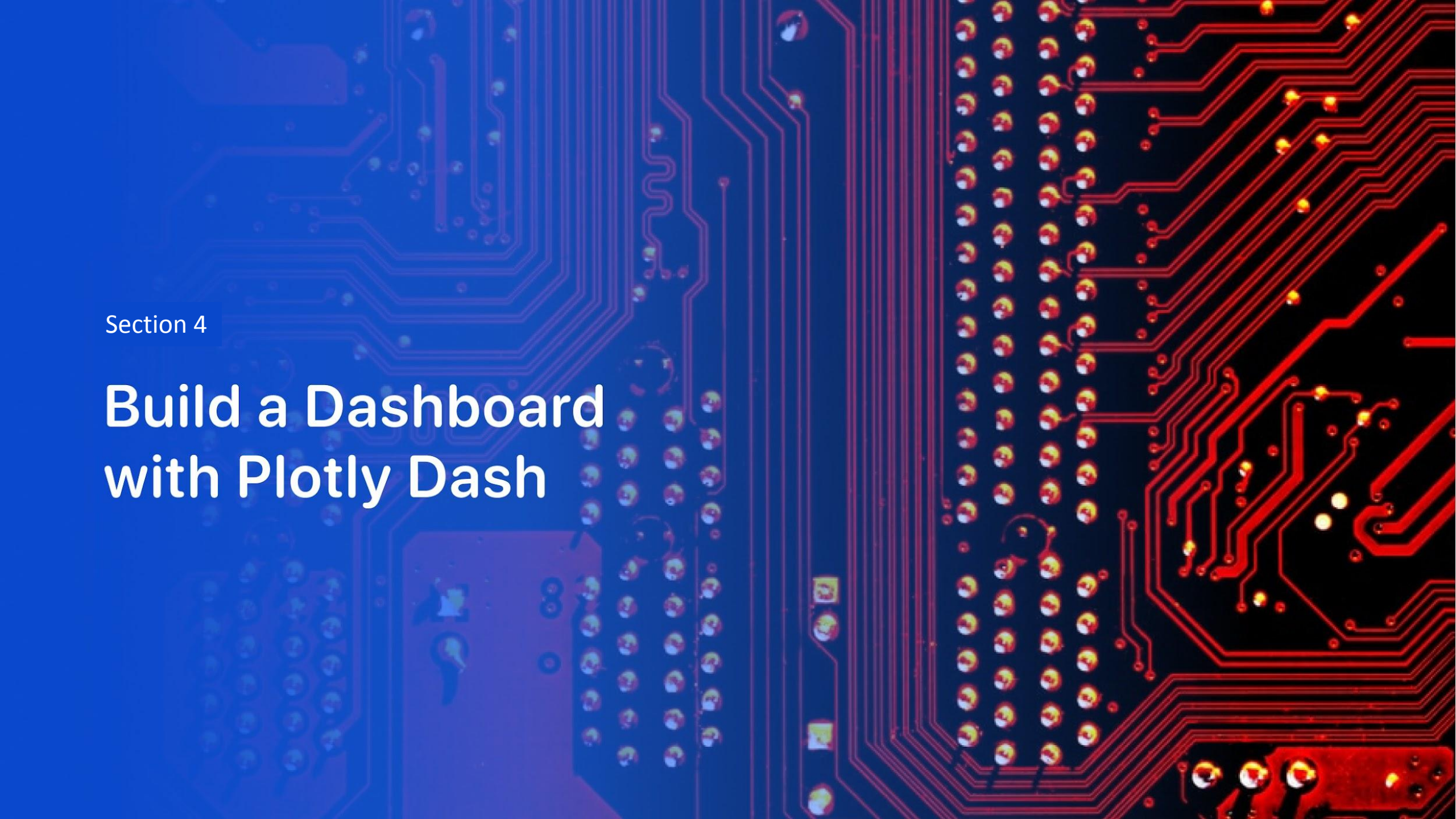
- The number at the center of each circle indicates total launches at each site
- Green markers represent successful landing outcomes and red markers represent failed landing outcomes



CCAFS SLC-40 Distances to Key Locations

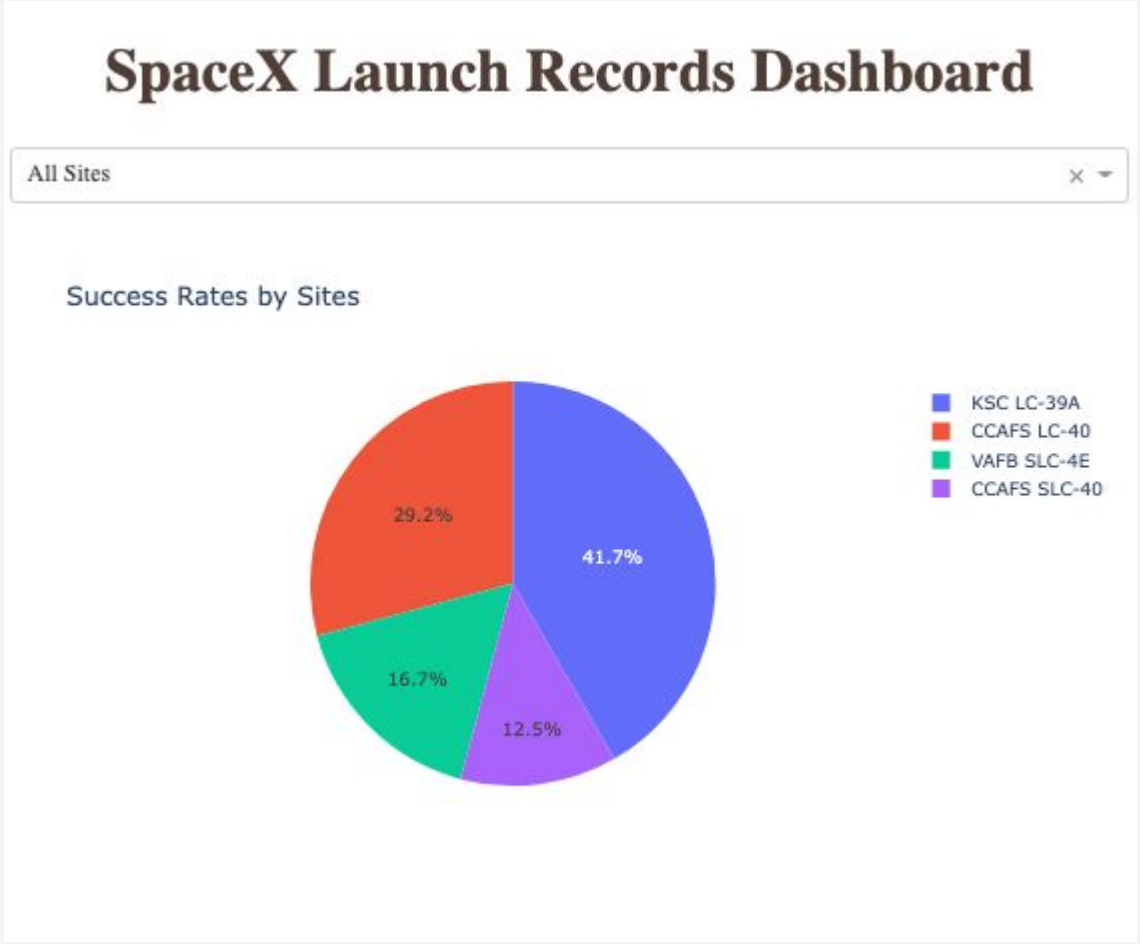
- Each line represent the crow distance from the site CCAFS SLC-40 to each key location. This is crucial to know because it contributes to the logistics of the operations as well as safety measures
  - Close to railways (1.29 km)
  - Very close to highways (0.59 km) and coastline (0.86 km)
  - Safely distant from big cities (51.27 km)




Section 4
Build a Dashboard with Plotly Dash

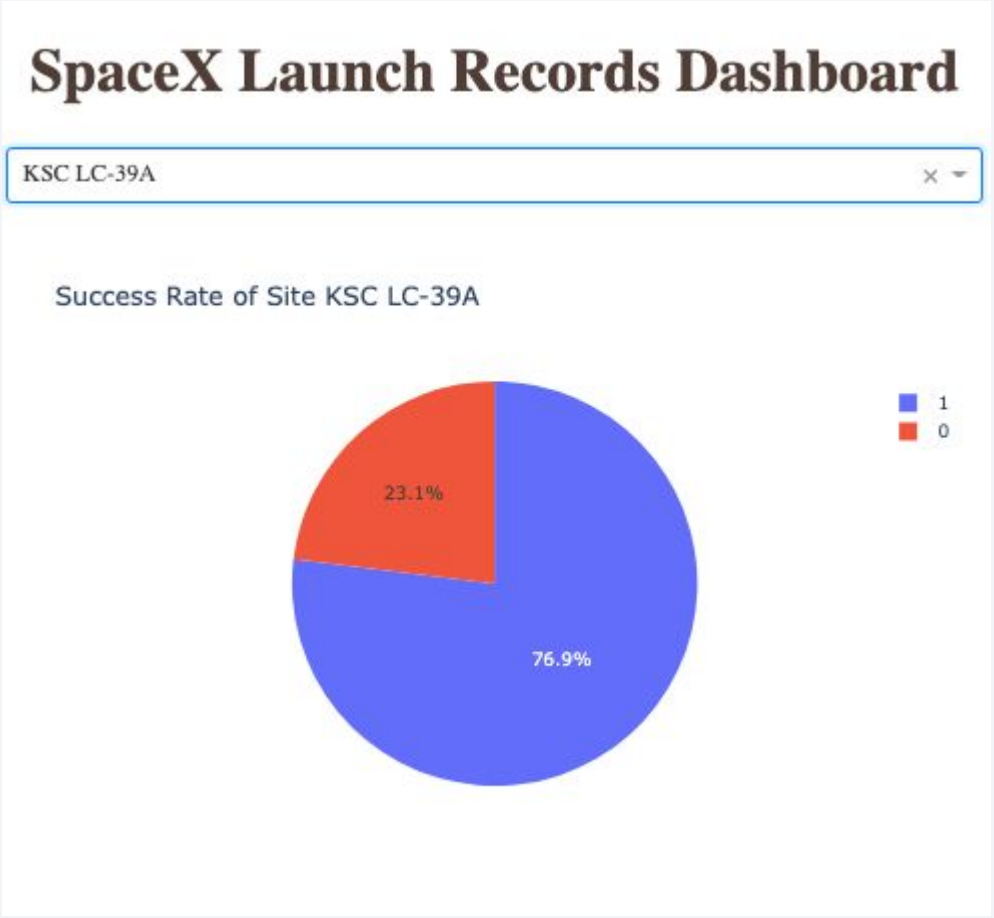
Success Rates by Sites

- From the pie chart, the inference can be made that launch site does contribute to whether a launch will be successful
- Majority of successful launches happened at KSC LC-39A and CCAFS LC-40



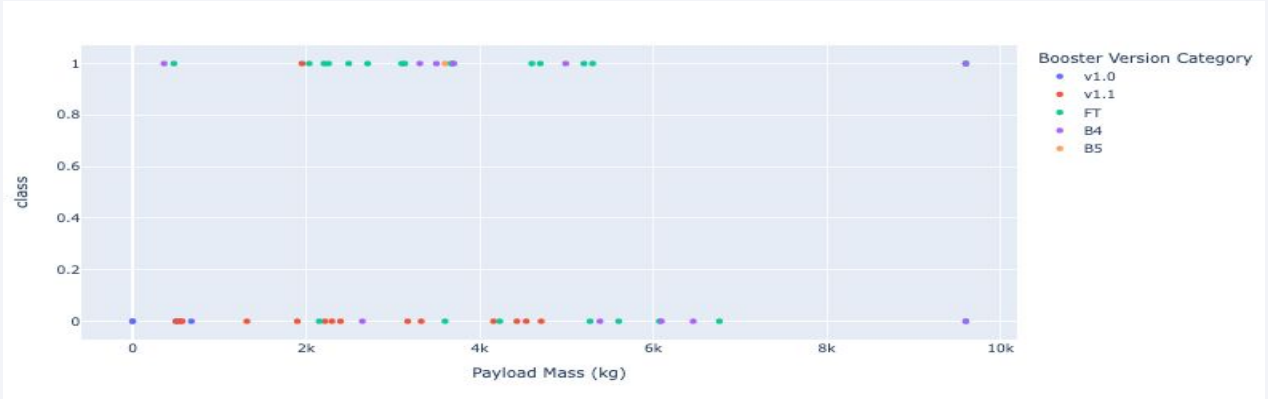
Success Rate of KSC LC-39A

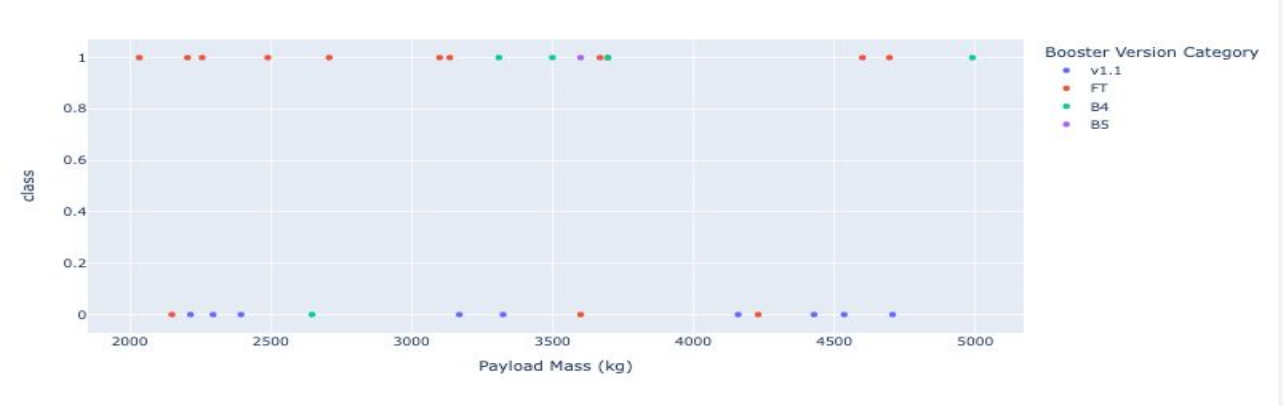
- Site KSC LC-39A has the highest launch success ratio out of the 4 sites with the success rate of 76.9% and 23.1% chance of failure.



Payload Mass vs Landing Outcome

- Data is scarce in payload range above 7000 kg
- Majority of successful launches falls within the payload range of 2000–5000 kg with `FT` being the most common booster version category





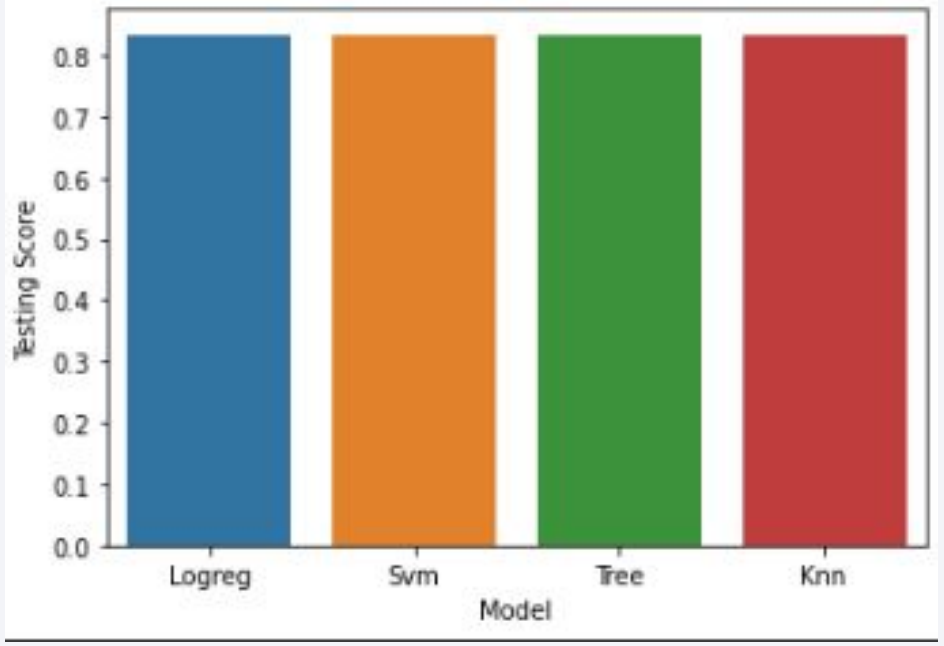
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All 4 classifiers have the exact same accuracy of 83.33%; more data is needed to make a definite distinction in accuracy
- However, the best choice for now is the decision tree classifier as it scores the highest with training data (88.92%)

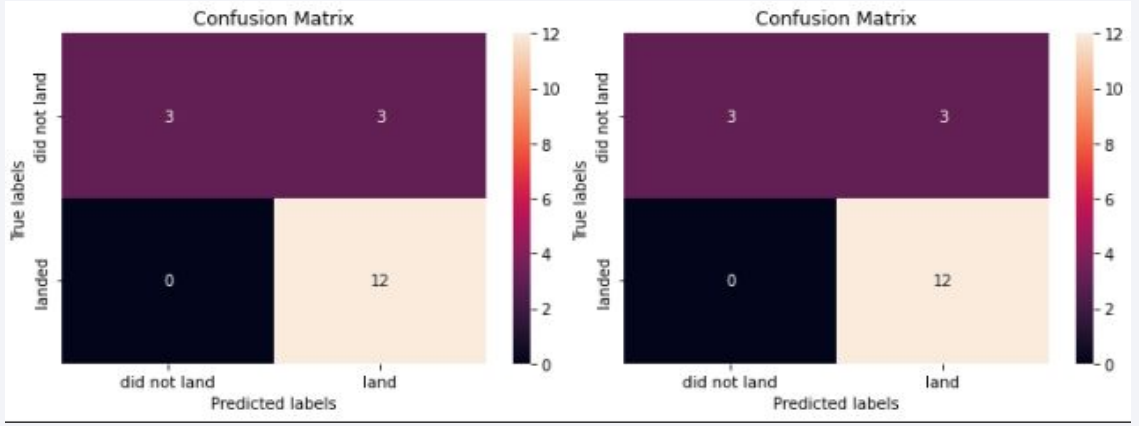
	Model	Training Score	Testing Score	Jaccard Score	F1 Score
0	Logreg	0.846429	0.833333	0.8	0.888889
1	Svm	0.848214	0.833333	0.8	0.888889
2	Tree	0.889286	0.833333	0.8	0.888889
3	Knn	0.848214	0.833333	0.8	0.888889



Confusion Matrix

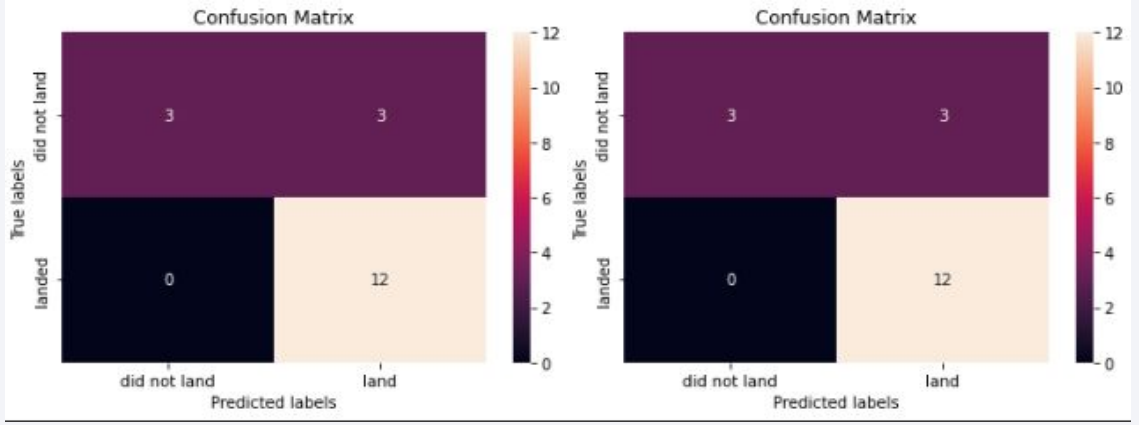
- All 4 classifiers also produces the exact same confusion matrix of [[3, 0], [3, 12]]
- The matrices show that all classifier have trouble dealing with false positives
    - Further improvement should be focused on reducing this as it directly translates to unaccounted increase in cost if implemented (model falsely predicts a successful landing)

Logistic Regression



Support Vector Machine

Decision Tree



K-Nearest Neighbor

Conclusions

- As time passes, the chance of boosters successfully landing back with their first stages increases as technology in rocket science advances
- The best launch site with the highest success rate is KSC LC-39A
- Launches with payloads within 2000-5000 kg range or above 9000 kg have a better chance of landing successfully
- Orbits with the highest success rate are ES-L1, GEO, HEO, and SSO
- All classifiers score the same on test data (83.33%) and, therefore, decision tree classifier is chosen by default for its highest score on train data (88.92%)

Thank you!