Winning Space Race with Data Science

Filipe Machado
December 29th 2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data collected from SpaceX API and Wikipedia's page "SpaceX Falcon Heavy Launches"
  - Transformed using phyton, pandas and SQL
  - Analyzed with pandas, numpy, matplotlib, seaborn, folium and plotly
  - Predicted data using logistic regression, support vector machine, decision tree and K nearest neighbor models using scikit-learn
- Summary of all results
  - Launch success predictions were approximately 89% accurate in the best performing model.

Introduction

- Project background and context
  - Commercial space exploration is a new and promising field. Many companies like SpaceX and Blue Origin and Virgin Galactic are developing their own programs future development of this market.
  - One particular development from SpaceX's program draws the attention when thinking of cost and commercial viability, the capacity to recover rocket stages after a launch.
- Problems investigated by the models
  - Is it possible to predict when a rocket launch will be successful? And when a rocket stage will successfully land? Will a machine learning model be able to tell us the best conditions for a space launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Request on SpaceX API
  - Web scrapping in Wikipedia SpaceX's page
- Perform data wrangling
  - Information adjusted for better understanding and visualization (null values were treated, new values calculated using python and pandas)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

SpaceX API

• Initial data was collected by request on the SpaceX's API. Data included:

Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Finds, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, & Latitude

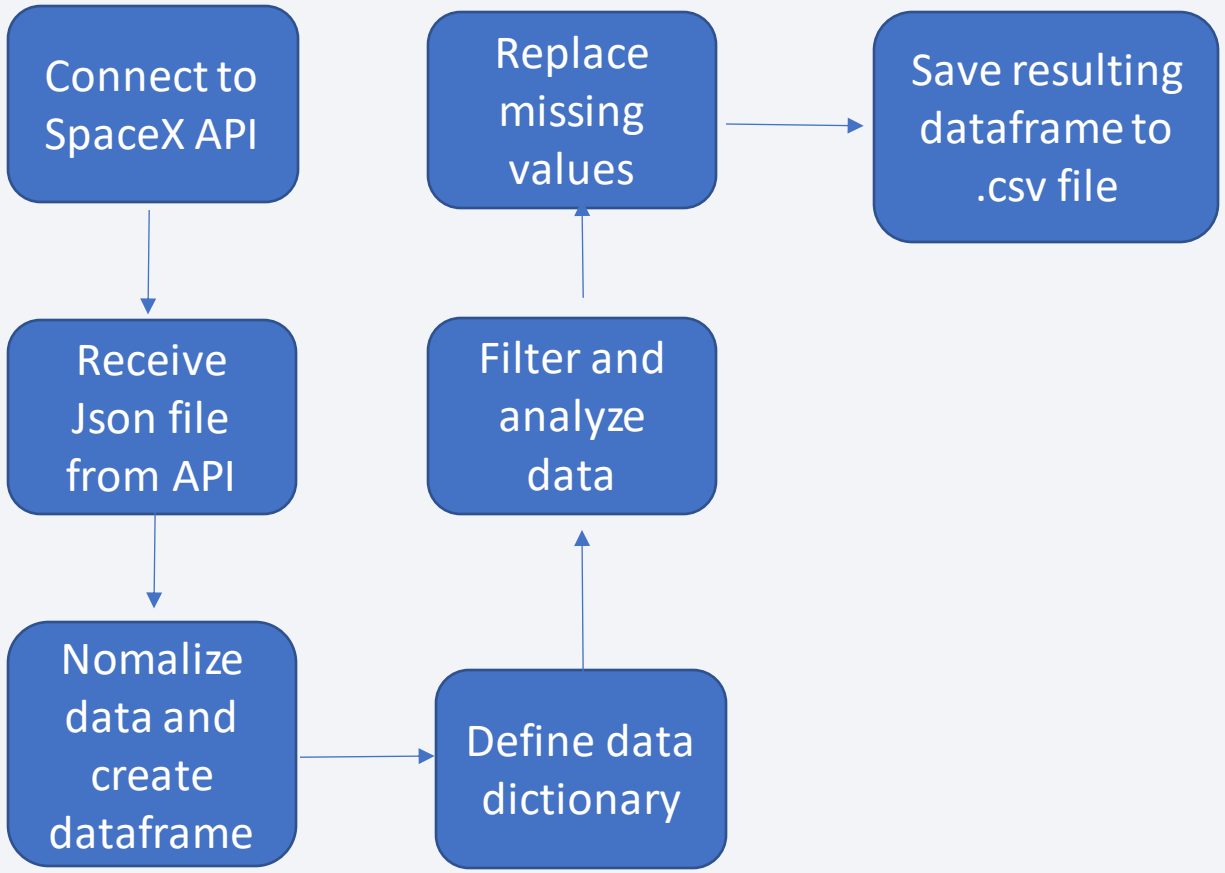
Web Scraping

• Data was captures from Wikipedia's "Falcon 9 and Falcon Heavy Launches Records" via web scraping its HTML table. Data included:

Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, & Time

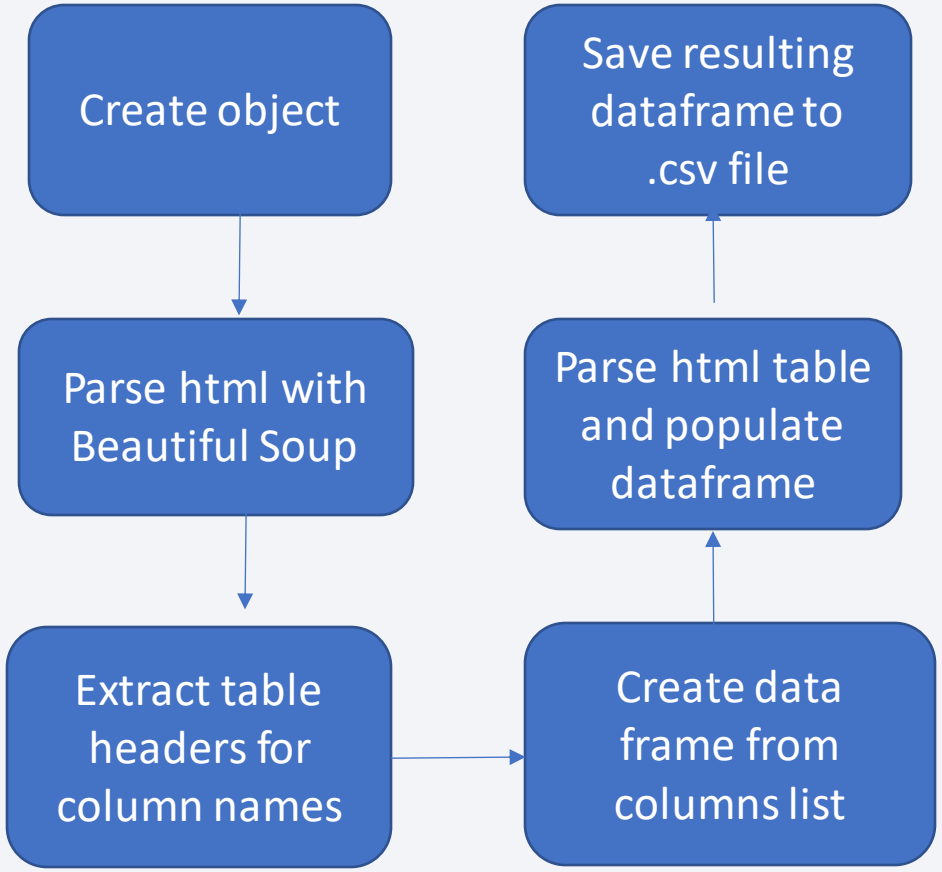
Data Collection – SpaceX API

- Information gathered from "https://api.spacexdata.com/v4/launches/past" via response method.
- View full process on the link below:
- Github repository link - DataCollection Notebook



Data Collection - Scraping

- Information gathered from Wikipedia SpaceX's page using Beautiful Soup.
- View full process on the link below:
- Github repository link - WebScraping notebook



Data Wrangling

Data was analyzed by calculating:

- The number of launches at each site
- The number and occurrences of each orbit
- The number and occurrence of mission outcomes per orbit type
- The Outcome column was created to one-hot encode the outcome of each launch (0 for failed launch, 1 for successful launch)

- Full process documented int the link below:

Github repository link

EDA with Data Visualization

Charts Created:

- Flight Number vs. Payload Mass (scatter plot)
- Flight Number vs. Launch Site (scatter plot)
- Payload vs. Launch Site (scatter plot)
- Success Rate by Orbit Type (bar chart)
- Flight Number vs. Orbit Type (scatter plot)
- Payload Mass vs. Orbit Type (scatter plot)
- Launch Success Annual Trends (line chart)

Github repository for Data Visualization

EDA with SQL

Queries performed using SQL:

- list of unique launch sites.
- Considered "CCA" launch sites.
- Calculated the total payload mass for boosters launched by NASA (CRS).
- Calculated the average payload mass (in kg) carried by launches with booster version F9 v1.1.
- Determined the first successful landing in which ground pad was achieved.
- Listed the boosters that have successful drone ship outcomes with payload mass between than 4,000 and 6,000 kg.
- Calculated the total success & failure outcomes for each type of mission outcome.
- Listed the booster versions with the greatest payload mass.
- Listed landing outcome, booster version, & launch site for failed drone ship landings.
- Ranked the outcomes of landings that occurred between 6/4/2010 & 3/20/2017 in descending order

Github repository EDA with SQL

Build an Interactive Map with Folium

- Created a map of all the launch sites.
- Marked launches at each site. When zoomed out, map displayed total number of launches at each site. when zoomed in, map displays green pop-up markers for successful launches and red pop-up markers for failed launches.
- Calculated the distance between launch sites and coastlines, railways, highways and cities. Visualized these distances with red lines connected these locations.

Github repository link for interactive map

Build a Dashboard with Plotly Dash

Using Plotly Dash:

- Created a pie chart that displays the percentage of launch successes and failures. Selecting different launch sites from the drop-down adjusts the pie chart accordingly.
- A slider bar to filter the payload in the scatter chart.
- A scatter chart visualizes the correlation between payload and launch success. Selecting different launch sites from the drop-down menu adjusts the scatter chart accordingly.

Github repository link to plotly file

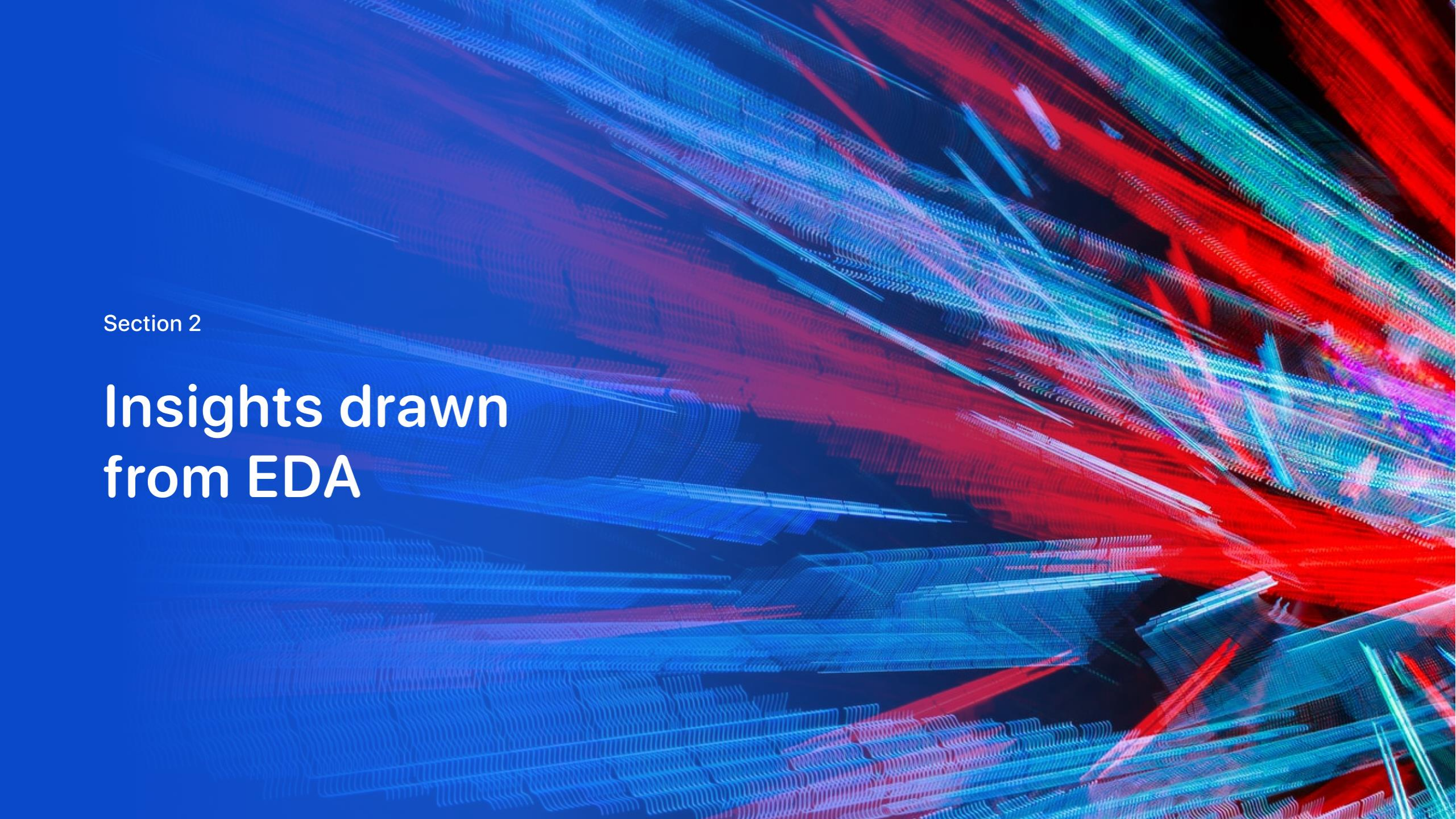
Predictive Analysis (Classification)

- Pandas was used to read a .csv file into a data frame.
- "Class" value in the data frame was assigned to a numpy array Y.
- The data frame was standardized and transformed using the StandardScaler method and assigned to data frame X.
- The X & Y data frames were split into training and testing data sets.
- Models were trained & fitted using various parameters (Logistic regression, Support Vector Machine, decision tree, & K nearest neighbors).
- Accuracy of each model was assessed for the most accurate parameters was assessed using the score method.
- A confusion matrix was plotted to assess the training test data compared to the testing data.
- The scores for each model were compared.
- The decision tree model made the most accurate predictions.

Github repository link for the MLPrediction

Results

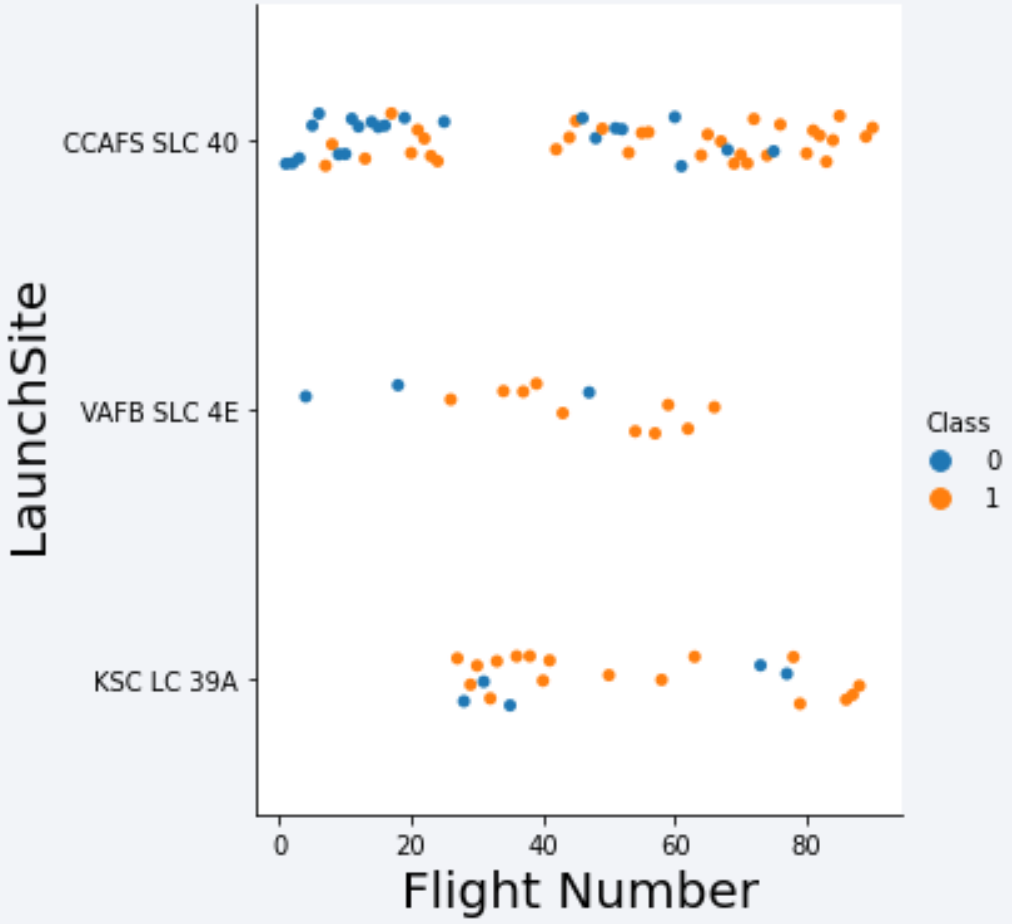
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

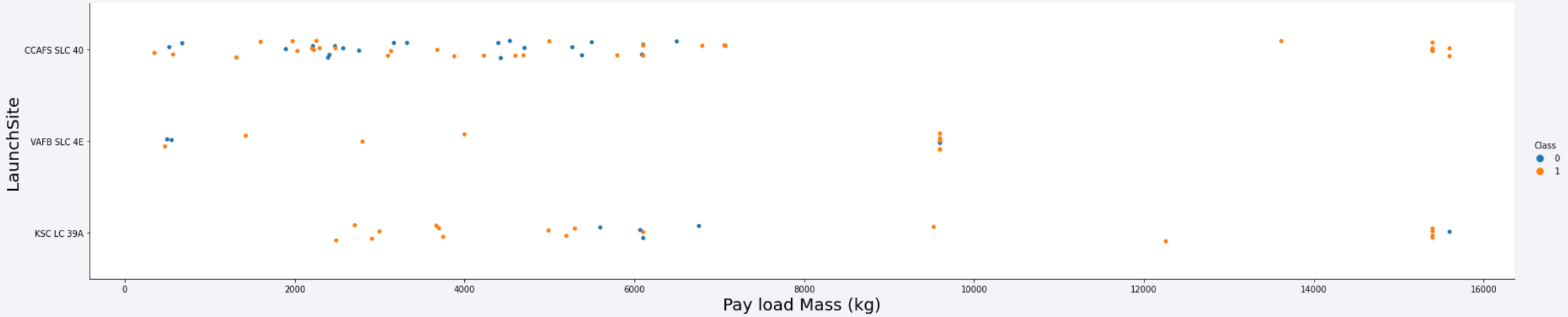
Insights drawn from EDA

Flight Number vs. Launch Site



- Orange indicates a successful launch, while blue indicates a failed launch.
- For launch site VAFB SLC 4E, it appears that higher flight numbers correspond to more successful launches.
- For the other launch sites, it appear that there is no relationship between flight number and launch site.

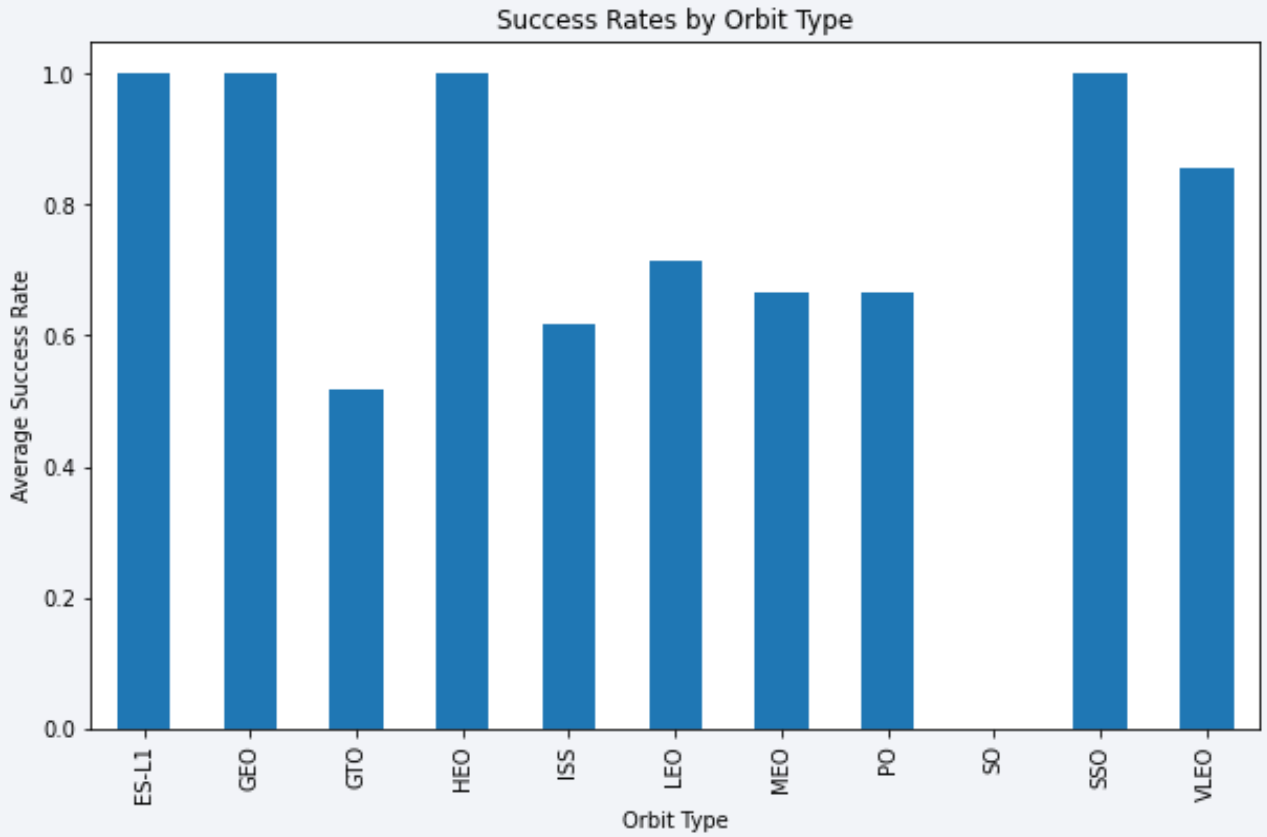
Payload vs. Launch Site



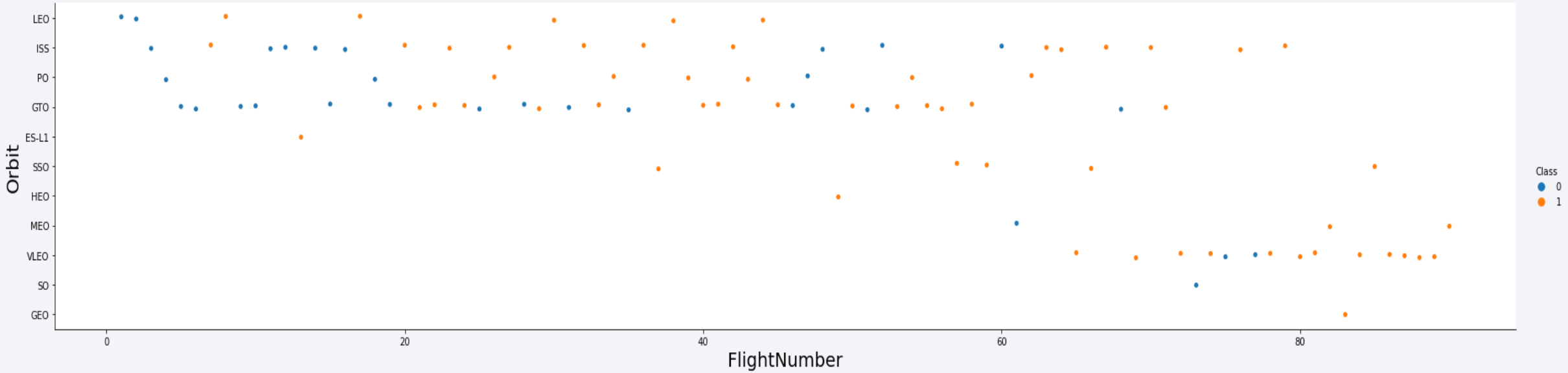
- Orange indicates a successful launch, while blue indicates a failed launch.
- Only lighter payloads were launched from site VAFB SLC 4E.
- For launch site CCAFS SCL 4G, heavier payload were more successful.

Success Rate vs. Orbit Type

- Orbit types ES-L1, GEO, HEO and SSO had the highest average success rates.
- Orbit types GTO and SO were the least successful.

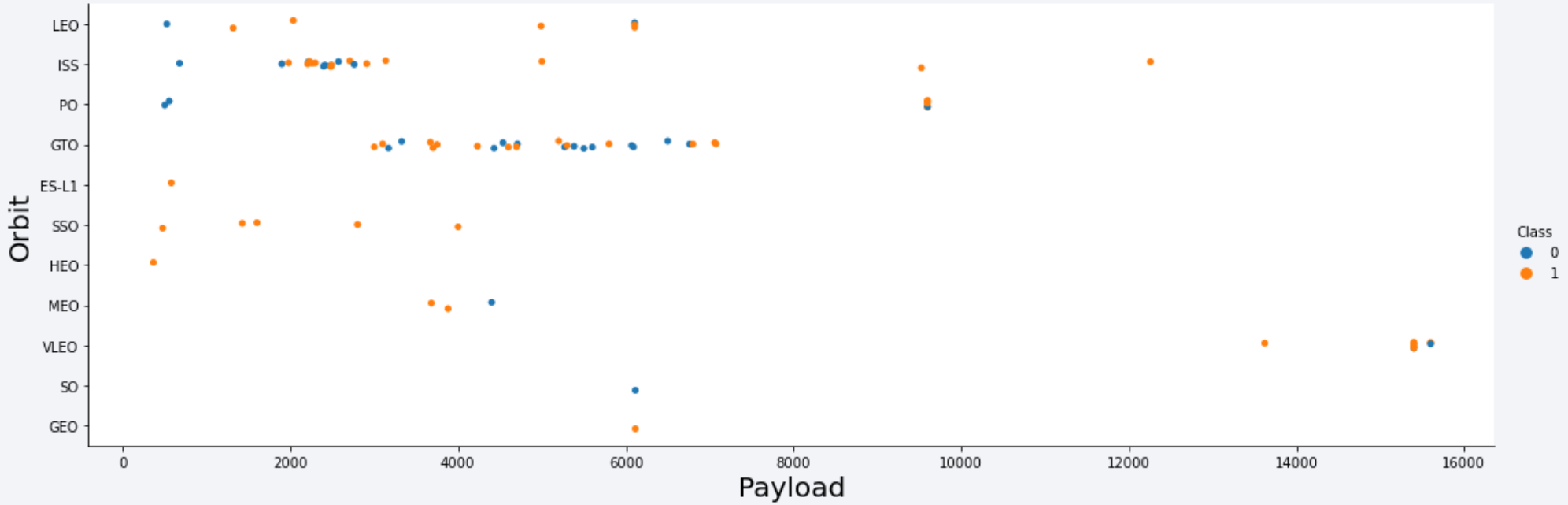


Flight Number vs. Orbit Type



- For LEO & MEO orbits, higher flight numbers were more successful.
- For ES-L1, SSO, HEO, & GEO orbit launches, all flight numbers were successful.
- All SO orbit launches were failure.

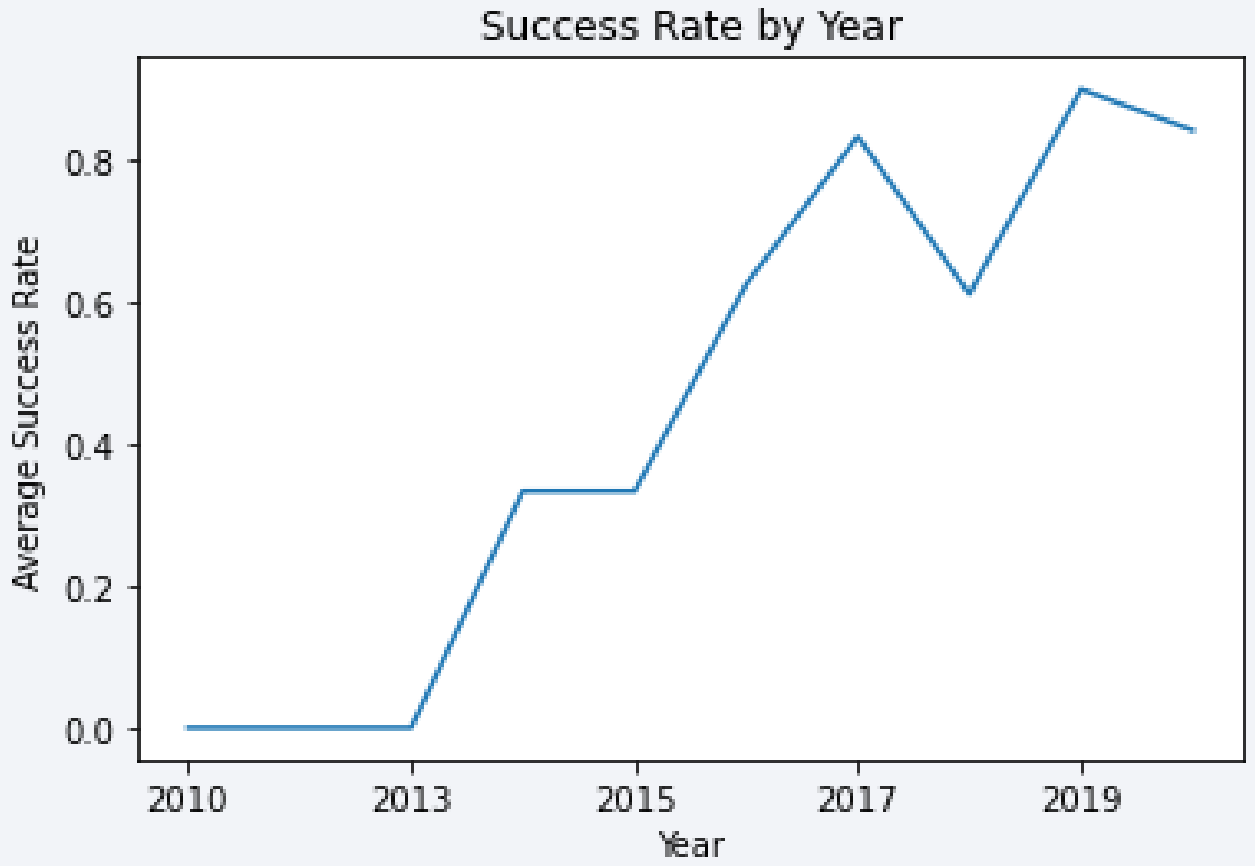
Payload vs. Orbit Type



- For LEO, ISS orbit launches, heavier payloads tend to be more successful.
- The ES-L1, SSO, HEO and GEO orbit launches, launches were successful regardless of the payload mass. However, all such payloads were lighter (i.e., less than 8,000 kg).

Launch Success Yearly Trend

- Overall, the average success rate increased from 2013 to 2019. (It decreased only in 2018 and 2020).



All Launch Site Names

- There are 4 distinct launch sites in the database.

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

Launch Site Names Begin with 'CCA'

- The first 5 launches occurring at a “CCA” launch site occurred at CCAFS LC-40.

```
In [6]: %%sql
        select * from spacexdataset where launch_site like 'CCA%'limit 5
```

```
 * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

Out[6]:

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

- The total payload mass for customer “NASA (CRS)” was 45,596 kg.

```
In [7]: %%sql
        select sum(payload_mass__kg_) as total_payload from spacexdataset where customer = 'NASA (CRS)'
         * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
        Done.
Out[7]: total_payload
        45596
```

Average Payload Mass by F9 v1.1

- The average payload mass for launches with booster version “F9 v1.1” was 2,534 kg.

```
In [8]: %%sql
        select avg(payload_mass__kg_) as avg_payload from spacexdataset where booster_version like 'F9 v1.1%'
         * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
        Done.
Out[8]: avg_payload
        2534
```

First Successful Ground Landing Date

- The first successful ground pad landing occurred on December 22nd 2015.

```
In [9]: %%sql
        select min(DATE) as first_success from spacexdataset where landing__outcome = 'Success (ground pad)'

         * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
        Done.

Out[9]: first_success
        2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are 4 distinct booster version for successful drone ship landings with payload masses between 4,000 & 6,000 kg.

```
In [10]: %%sql
         select booster_version from spacexdataset where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_ma
          * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
         Done.
```

Out[10]:

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

Total Number of Successful and Failure Mission Outcomes

- There were 100 successful missions (though one such mission had a unclear payload status). There was 1 failed mission.

```
In [11]: %%sql
         select count(date), mission_outcome  from spacexdataset group by mission_outcome
          * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
         Done.
```

Out[11]:

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

Boosters Carried Maximum Payload

- There were 12 different booster versions that had the maximum payload mass.

```
In [12]: %%sql
select distinct booster_version from spacexdataset where payload_mass__kg_ = (select max(payload_mass__kg_) from spacexdataset)
```

```
 * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

Out[12]:

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

- There were 2 different launches with failed drone ship landings in the year 2015.

In [13]:
```
%%sql
select booster_version, launch_site from spacexdataset where landing__outcome = 'Failure (drone ship)' and Year(Date)= 2015
```

```
 * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

Out[13]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- There are 31 landing outcomes between 6/4/2010 and 3/20/2017. When considered in descending date order, “No attempt” is ranked first.

```
In [14]: %%sql
         select count(date), landing__outcome  from spacexdataset where Date > '2010-06-04' and Date < '2017-03-20'
         group by landing__outcome
         ORDER BY count(date) DESC
```

```
 * ibm_db_sa://bpx11632:***@815fa4db-dc03-4c70-869a-a9cc13f33084.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30367/bludb
Done.
```

Out[14]:

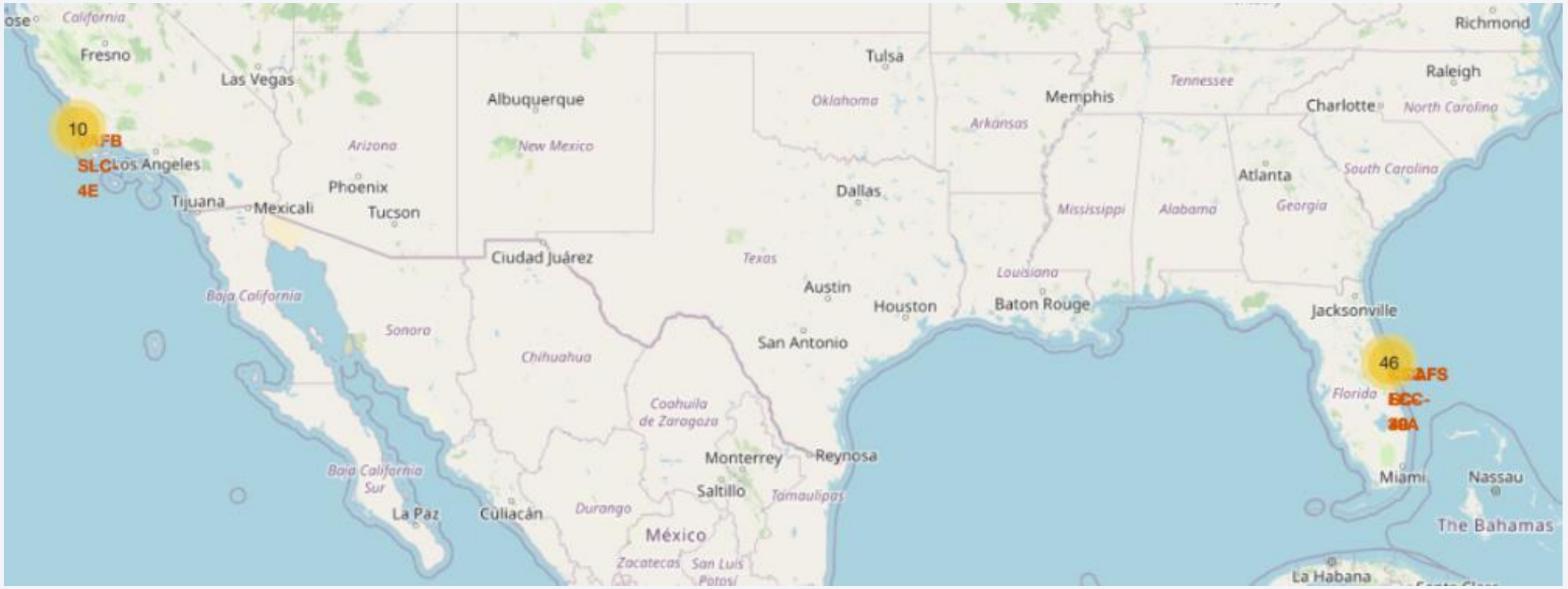
1	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Uncontrolled (ocean)
1	Failure (parachute)
1	Precluded (drone ship)

Section 4

Launch Sites Proximities Analysis

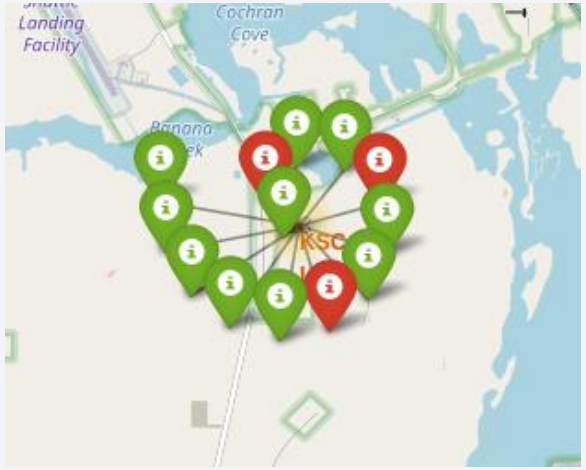
Launch Sites

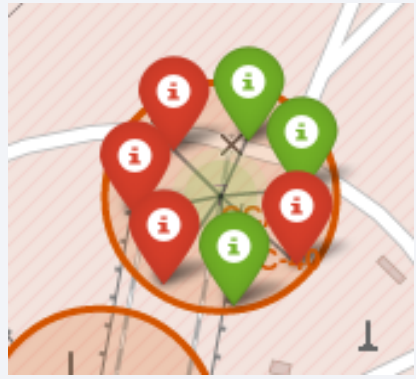
- All the launch sites appear in the map below. They are located in both costs of USA.



Success and Failure Locations

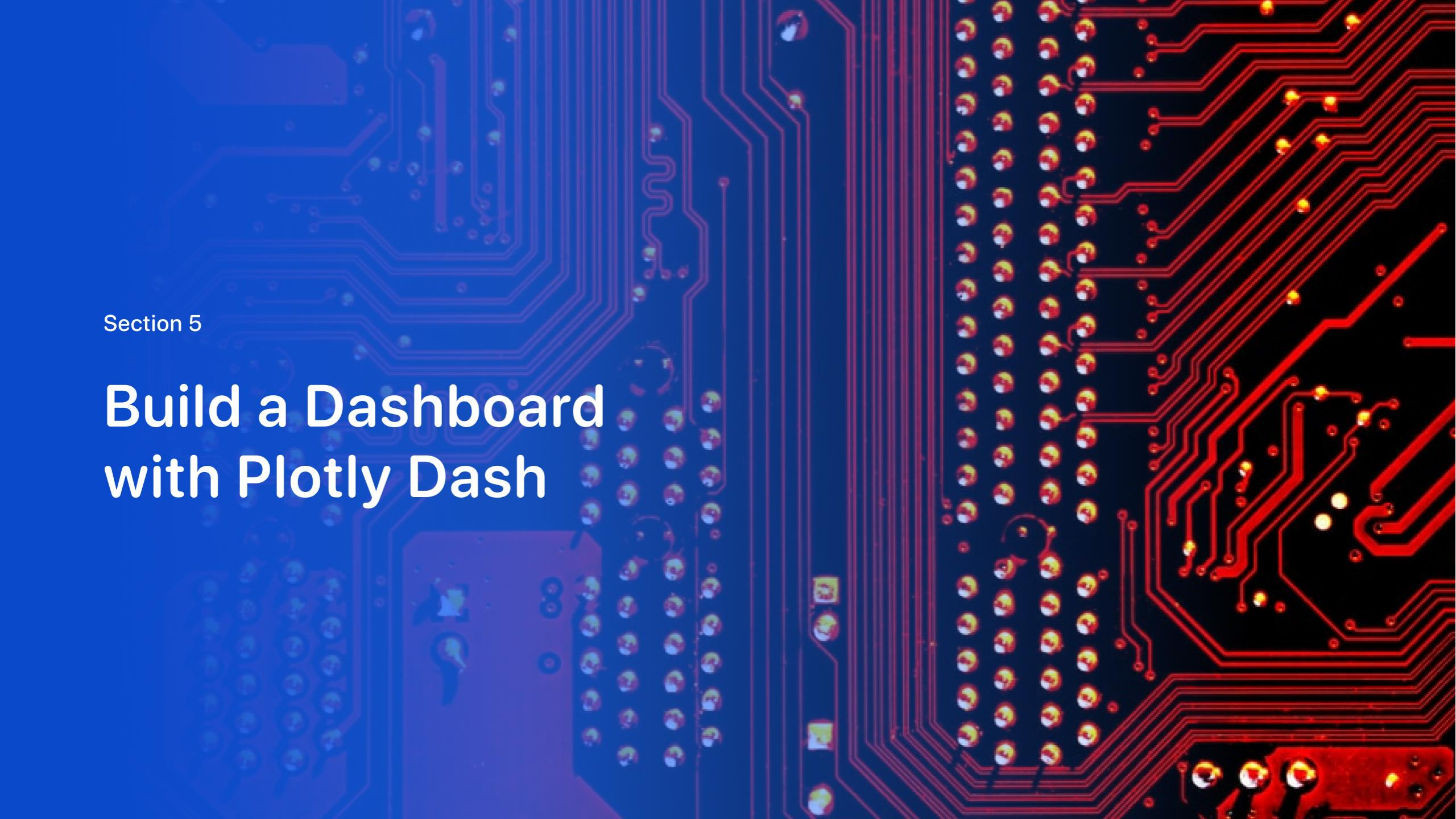
- When zoomed into a specific launch site, the specific launches that have occurred at that site become visible. Each launch is marked with a pop-up indicator that is color-coded to indicate whether the launch was a success or failure.
- We can clearly visualize a concentration of successful launches in one site (KSC LC-39A)





Launch site proximity

- The CCAFS SLC-40 launch site is 0.90 km from the coastline.
- Samuel C. Phillips Parkway highway is 0.58 km from the launch site.
- The nearest railway is 1.25 km from the launch site.
- The nearest city (Cape Canaveral) is 18.40 km from the launch site.

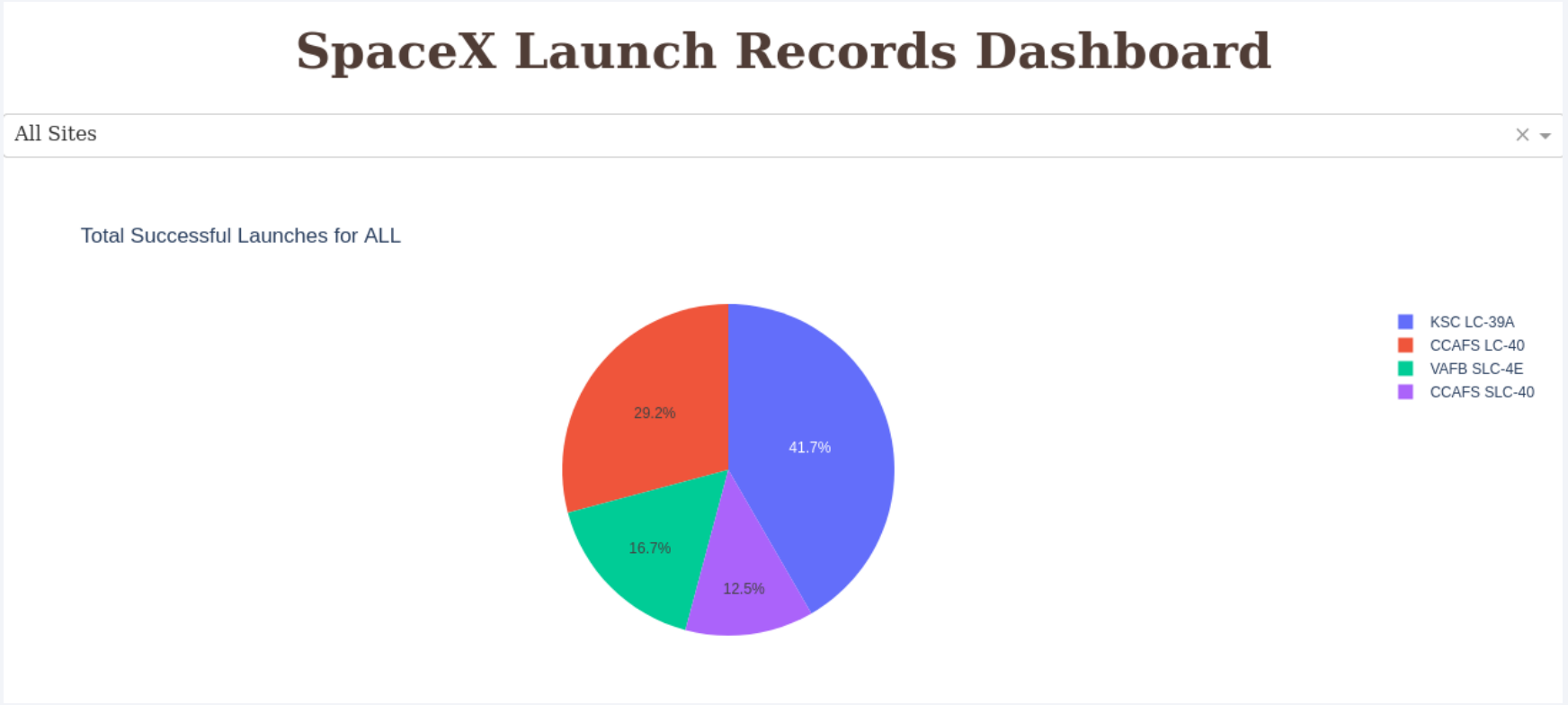


Section 5

Build a Dashboard with Plotly Dash

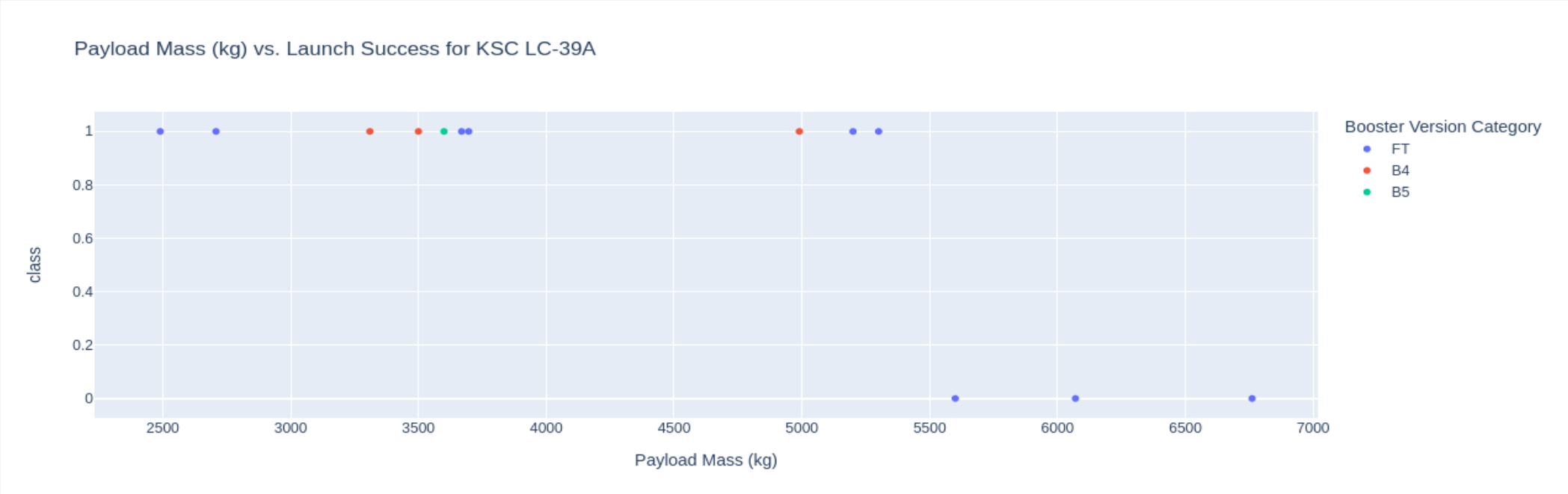
Successful Launches by Launch Site

- Site KSC LC-39A has the highest success rate, while site CCAFS SLC-40 has the lowest success rate.

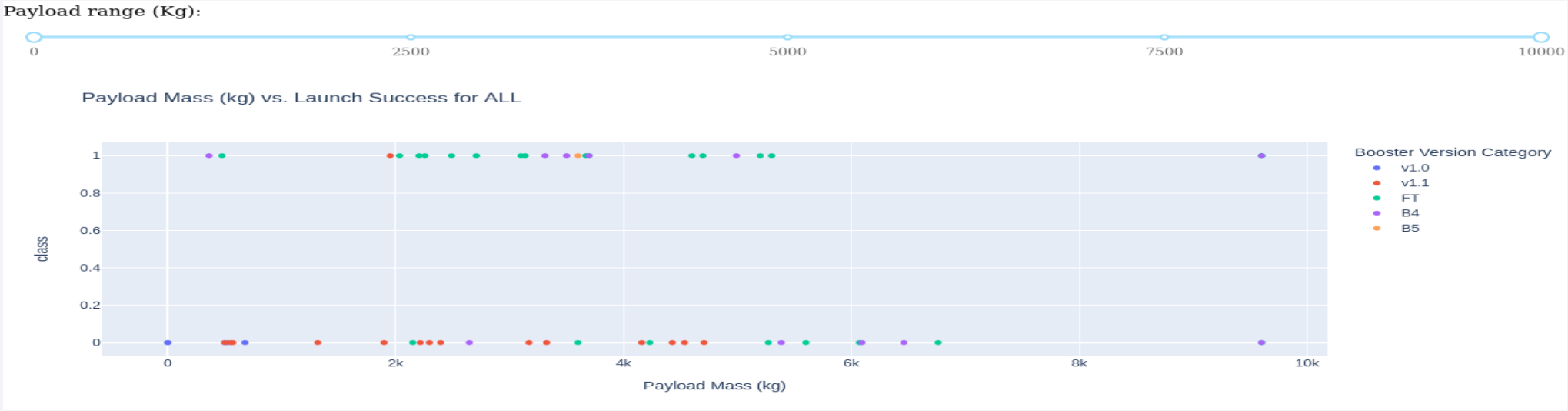


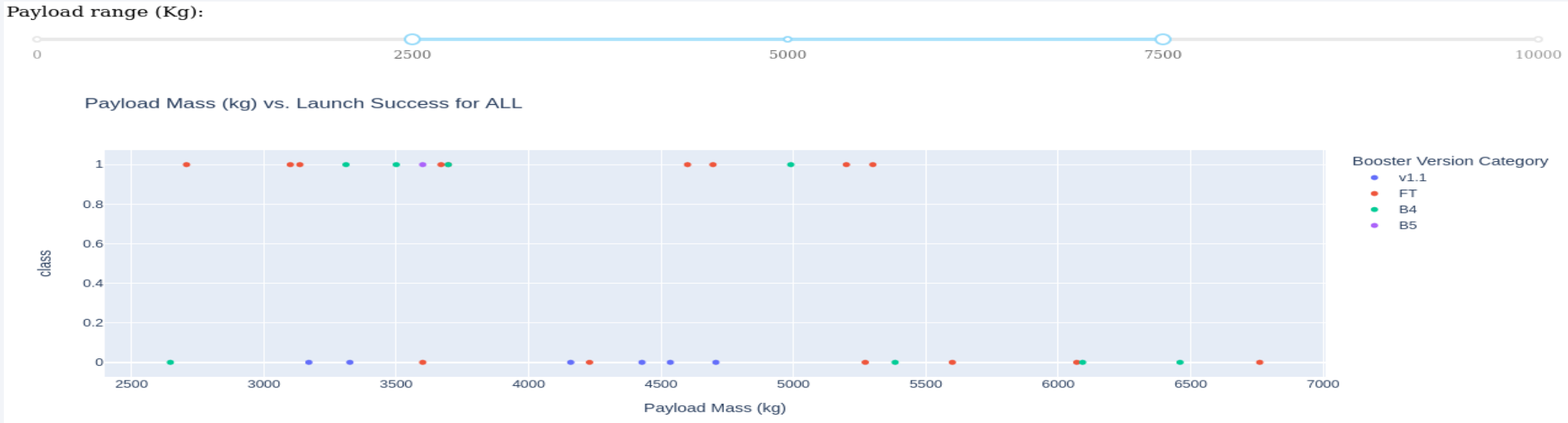
Successful Launches by KSC LC-39A

- At launch site KSC LC 39-A, 76.0% of launches were successful (class 1) , while 23% of launches failed (class 0)



Payload Mass (kg) vs. Launch Success





Section 6

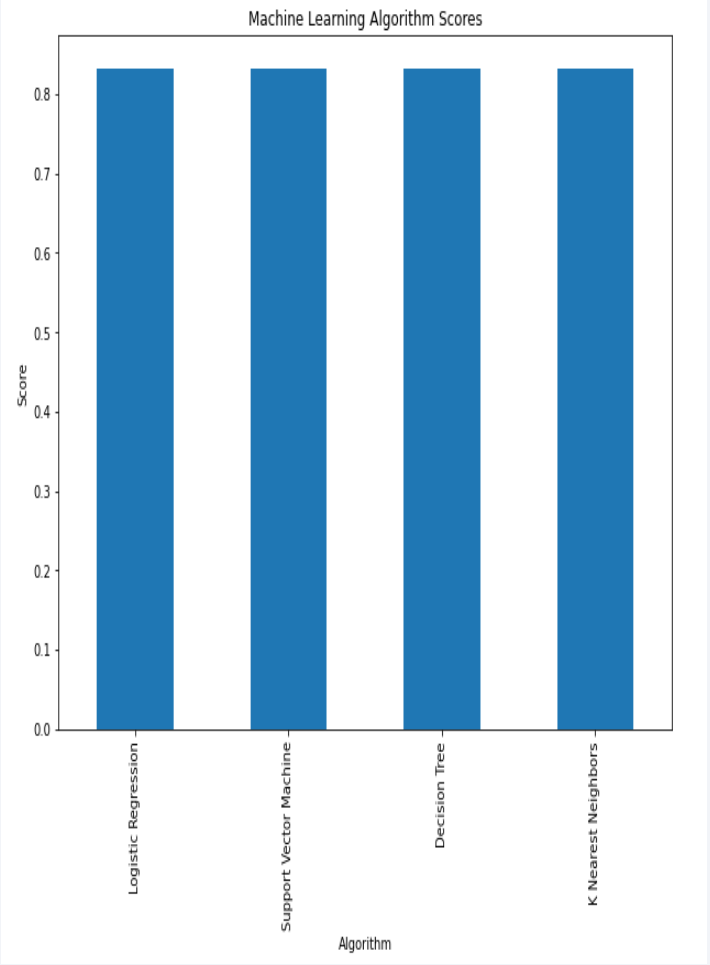
Predictive Analysis (Classification)

Classification Accuracy

Out[30]:

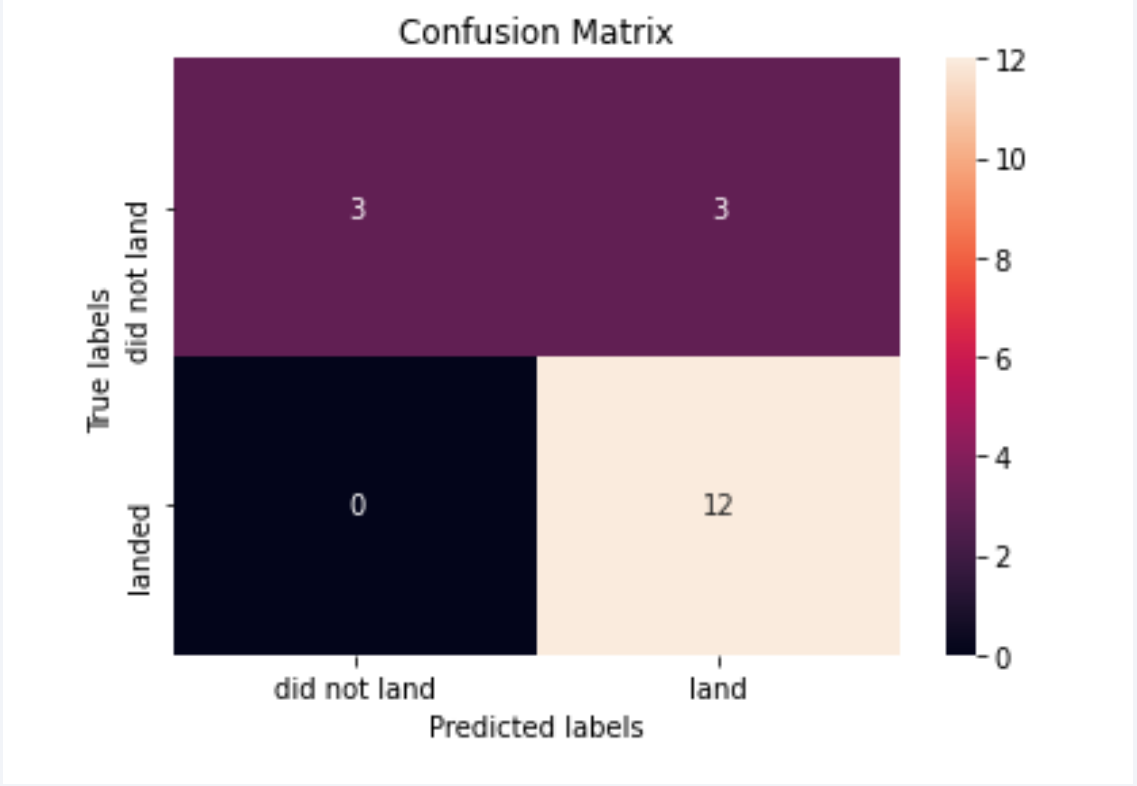
	Performance
Logistic Regression	0.846429
Support Vector Machine	0.848214
Decision Tree	0.889286
K Nearest Neighbors	0.848214

- The Decision Tree model had the best accuracy from all models



Confusion Matrix

- The decision tree confusion matrix indicates that is predicted:
  - 0 true negatives
  - 12 true positives
  - 3 false negatives
  - 3 false positives



Conclusions

- Analysis of launch data indicates that the average success rate of all launches is 66.67%.
- Querying mission outcomes indicates that only 1 out of 101 launches had an in-flight failure. The remaining missions were successful.
- The relationship between payload mass and launch site has an effect of the success of a launch. Specifically, launches with payload masses greater than 12,000 kg that occurred at sites CCAFS SLC-40 and KSC LC-39A were much more likely to succeed.
- Certain types of orbits (ES-L1, GEO, HEO, and SSO) had a higher success rate.
- The average success rate of launches increased from 2013 to 2017 and again in 2019.
- Launch sites may be in closer proximity to coastlines, highways, and railroads, but they tend to be located farther away from cities.
- Four machine learning models were used to predict the success or failure or future launches. The decision tree model had a better performance, accurately predicting 89% of the outcomes.

Thank you!