IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matthew Engelbert Bastiaan
13/11/2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
  SpaceX's website advertises Falcon 9 rocket launches with a cost of 62 million dollars; other providers cost upwards of 165 million dollars for each launch, much of the saving is because SpaceX can reuse the first stage. The objective of the project is to determine the cost of each launch by determining if the first stage will land. This information can be used by an alternate company that wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - What are the factors that contributes the most to the success of a first stage landing?
  - How can machine learning and predictive modelling be used to predict the success of a first stage landing based on historical data?
  - What machine learning algorithms works best in predicting the success of a first stage landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data about the launch details is collected from SpaceX REST API (https://api.spacexdata.com/v4/)
  - The data about Falcon 9 launch records is web scraped from (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - The features of the data is first summarized and analyzed, and then labelled based on findings regarding the success of the landing.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - The data will be standardized and the split into training and testing sets. Four classification models will be trained and tuned based on the training data using different combinations of parameters.

Data Collection

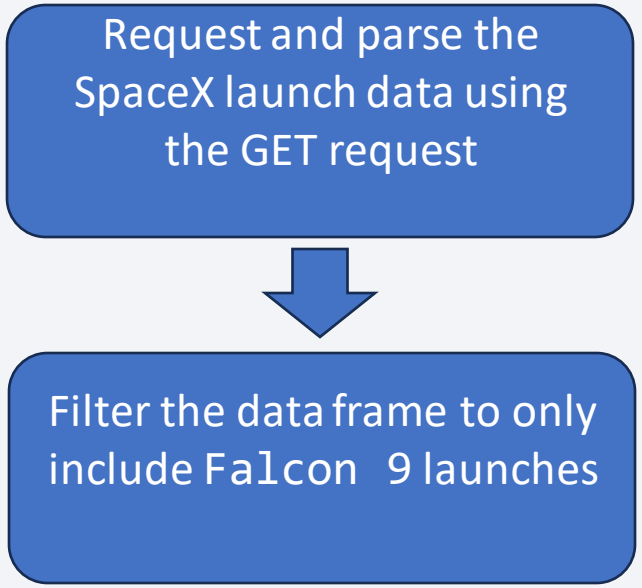
- Describe how data sets were collected.
  - The datasets was collected from SpaceX API (https://api.spacexdata.com/v4/)
    - From the rocket, the data the booster name was collected. (https://api.spacexdata.com/v4/rockets/)
    - From the launchpad, the data about the launch site being used, the longitude, and the latitude was collected. (https://api.spacexdata.com/v4/launchpads/)
    - From the payload, data about the mass of the payload and the orbit that it is going to was collected. (https://api.spacexdata.com/v4/payloads/)
    - From cores, data about the outcome of the landing, the type of the landing, number of flights with that core, whether grid fins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to separate version of cores, the number of times this specific core has been reused, and the serial of the core was collected. (https://api.spacexdata.com/v4/cores/)
    - All of the data will complement the rocket launch data collected from past launches (https://api.spacexdata.com/v4/launches/past)
    - A new data frame with columns consist of details of each Falcon 9 launches was created using the collected data.

Data Collection

- Describe how data sets were collected.
  - Additional datasets was also collected from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
    - A data frame consists of records of each Falcon 9 launches were also created during this process by using the data obtained by implementing web scraping techniques.

Data Collection – SpaceX API

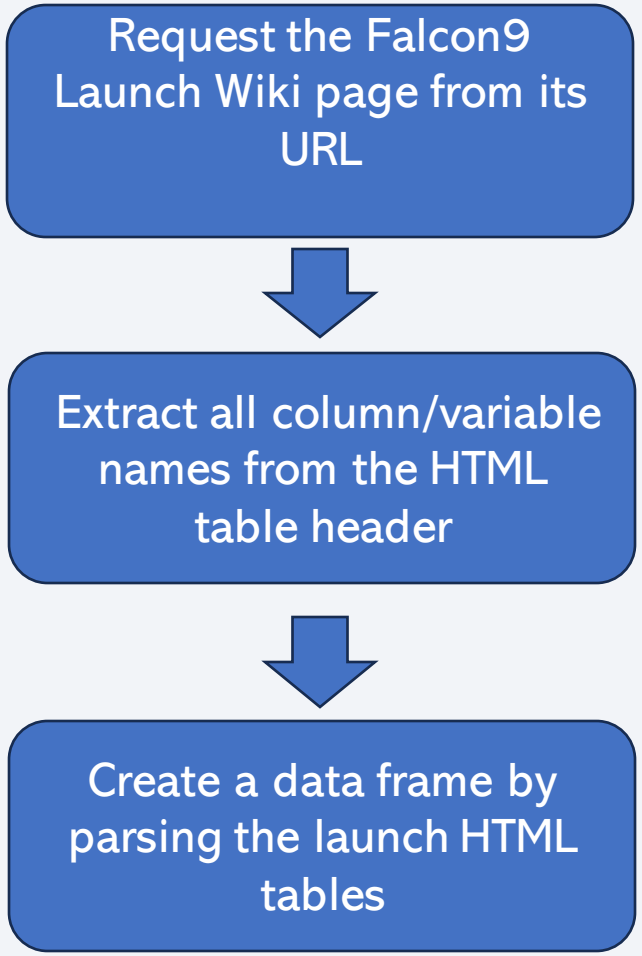
- Data collection using SpaceX Public API calls were done according to the flowchart beside.
- Notebook Github link: https://github.com/BastiaanMatt/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data collection using web scraping techniques on Falcon 9 launches Wikipedia page were done according to the flowchart beside.

- Notebook Github link: https://github.com/BastiaanMatt/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Initial data analysis was done on the filtered data obtained from the SpaceX API. Missing payload mass values are replaced with the mean of the column. Then the number of launches from each site was calculated, also the number and occurrences of each orbit and of each mission outcome is also summarized. Lastly, a landing outcome label was assigned for each launch based on whether the landing was a success or not.
- Notebook Github link: https://github.com/BastiaanMatt/IBM-Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with SQL

- The following list contains the SQL queries that have been performed in this project:

1. Names of the unique launch sites in the space mission
2. 5 records where launch sites begin with the string 'CCA'
3. The total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1
5. The date when the first successful landing outcome in ground pad was achieved.
6. The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
7. Total number of successful and failure mission outcomes
8. Names of the booster_versions which have carried the maximum payload mass.
9. The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
10. 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.

- Notebook Github link: https://github.com/BastiaanMatt/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with Data Visualization

- Exploration with data visualization was done using scatter plots, bar plots and line charts.
- Scatter plots was used to visualize the correlation between the variables (flight number, payload mass, launch site, and orbit type)
- A bar plot was used to visualize the success landing rate of each orbit type.
- A line chart was used to visualize change in the success landing rate over time.
- Notebook Github link: https://github.com/BastiaanMatt/IBM-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Build an Interactive Map with Folium

- The following list contains the map objects that have been created in this project:
    1. A set of a circle and marker of each launch sites on the map.
    2. A marker cluster of colored markers for each launch result.
    3. Lines and markers for calculating the distances between a launch site to its proximities.

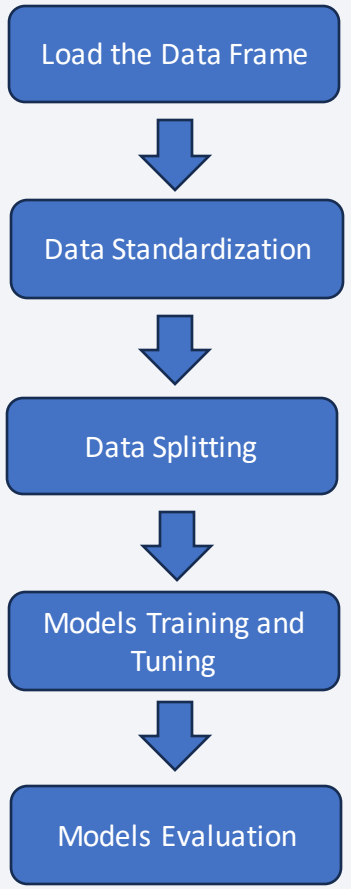
- Notebook Github link: https://github.com/BastiaanMatt/IBM-Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- In this project, the dashboard used pie charts and scatter plots to visualize either the summary of all launch sites or for a specific site that can be selected from a dropdown.
- Pie charts were used to visualize the summary of total success launches by site and the success rate for each individual site.
- Scatter plots were used to visualize the correlation between payload and success for all sites and correlation between payload and success for each individual site.
- Notebook Github link: https://github.com/BastiaanMatt/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- After the data frame from the data collection step have been loaded. Features were extracted from the web scraped data frame and the response variable was extracted from 'Class' in the SpaceX API data frame.
- The features then would be standardize before split into training and testing dataset. Models will then be trained and tuned using the standardized features from the training dataset.
- Models that were built in this project includes Logistic Regression, SVM, Decision Trees and KNN.
- After the models have been trained, then the accuracy of each model predictions were tested using the testing dataset. Confusion matrix of each models were created too.
- Notebook Github link: https://github.com/BastiaanMatt/IBM-Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

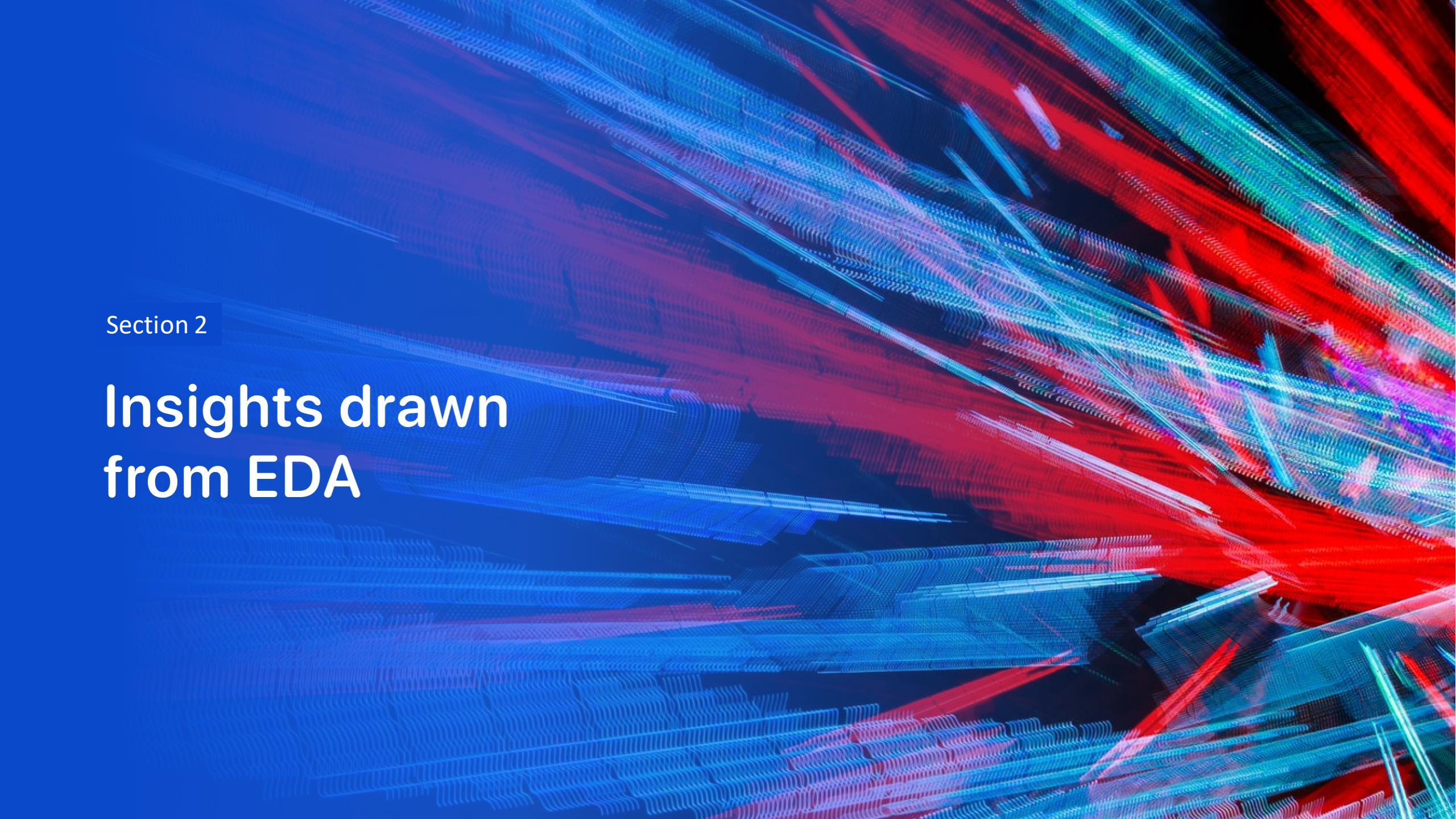


Results

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

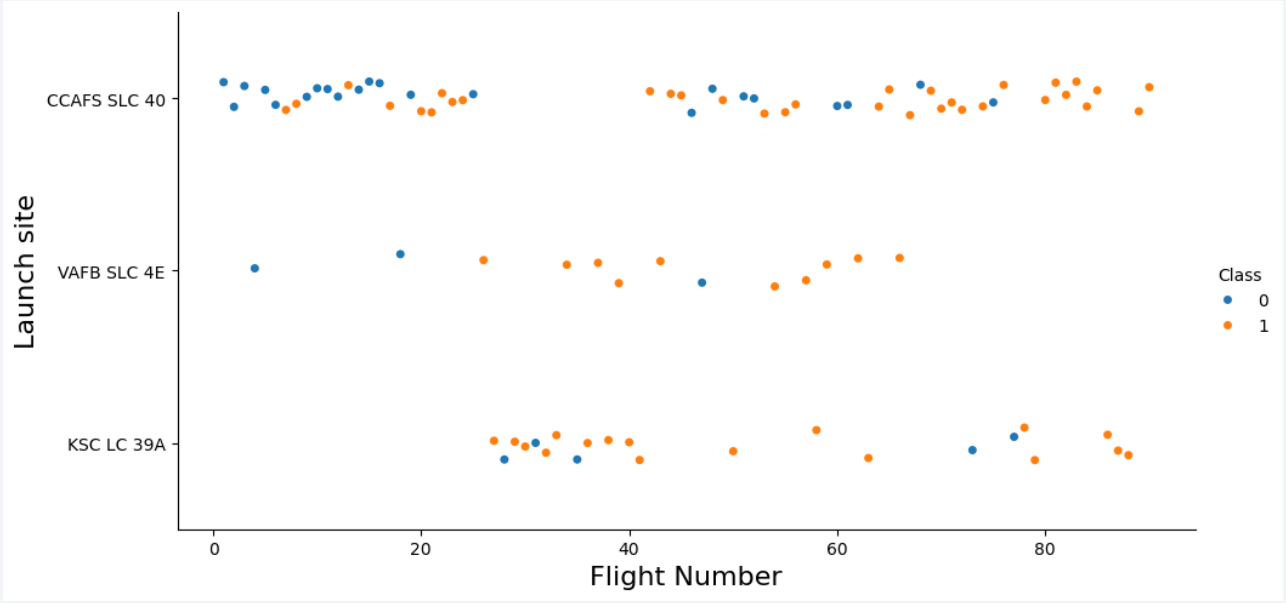
Section 2

Insights drawn from EDA



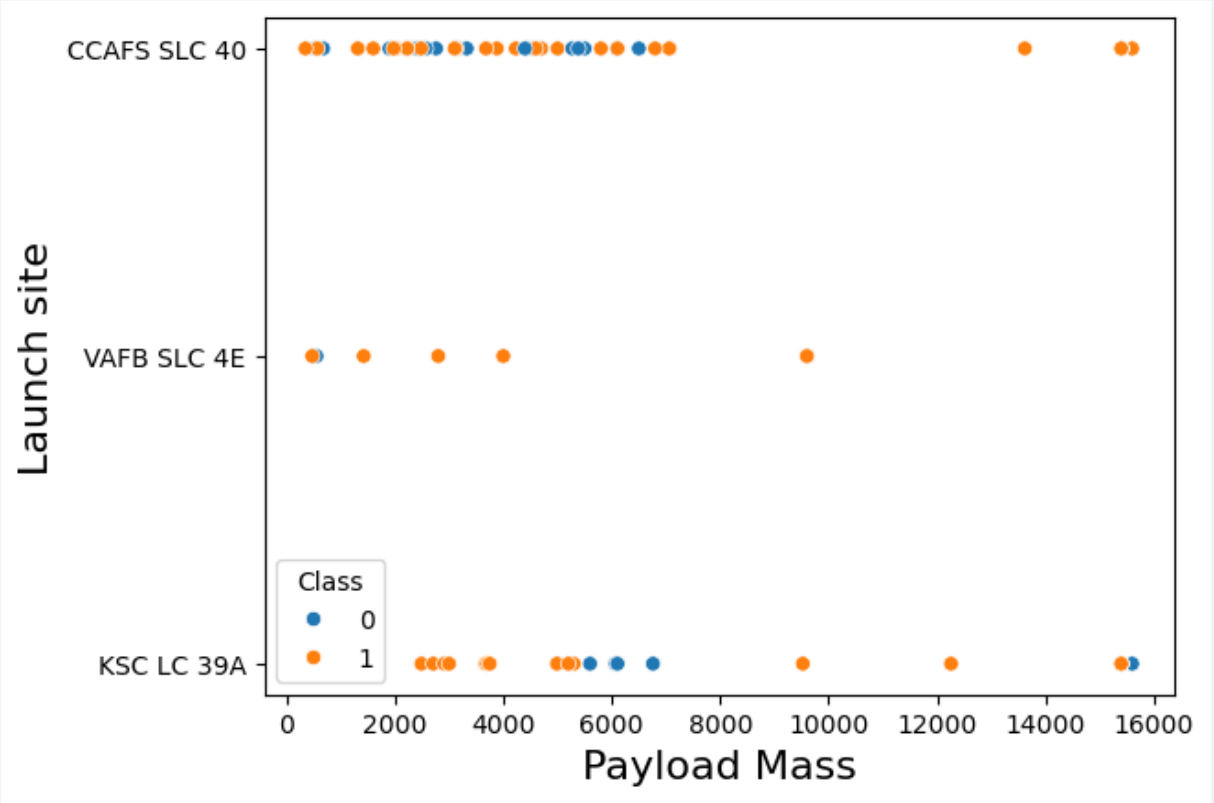
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site.
- Most of the launch were from CCAFS SLC-40.
- Launches with **higher flight number** have a **higher chance of success**.



Payload vs. Launch Site

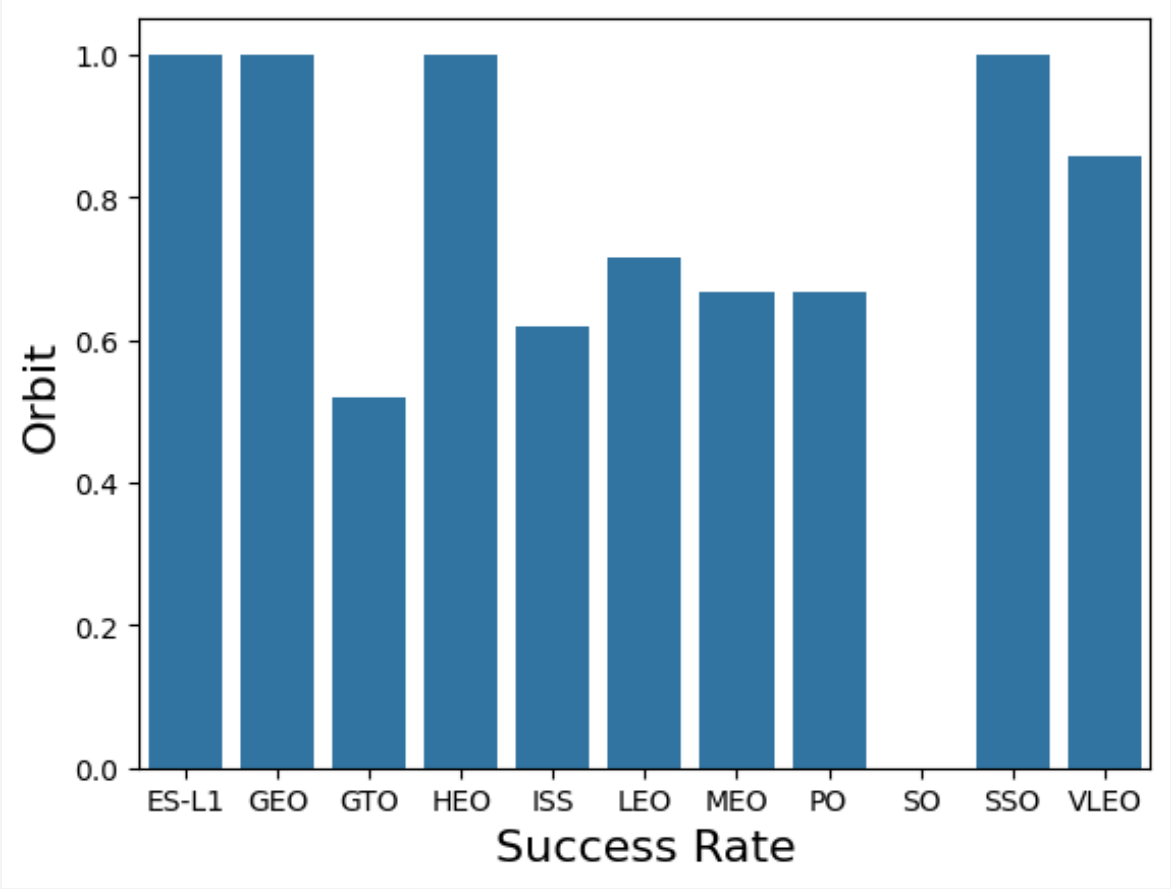
- Scatter plot of Payload vs. Launch Site
- There are no rockets launched with a payload mass greater than 10000 for the VAFB-SLC launch site.



Success Rate vs. Orbit Type

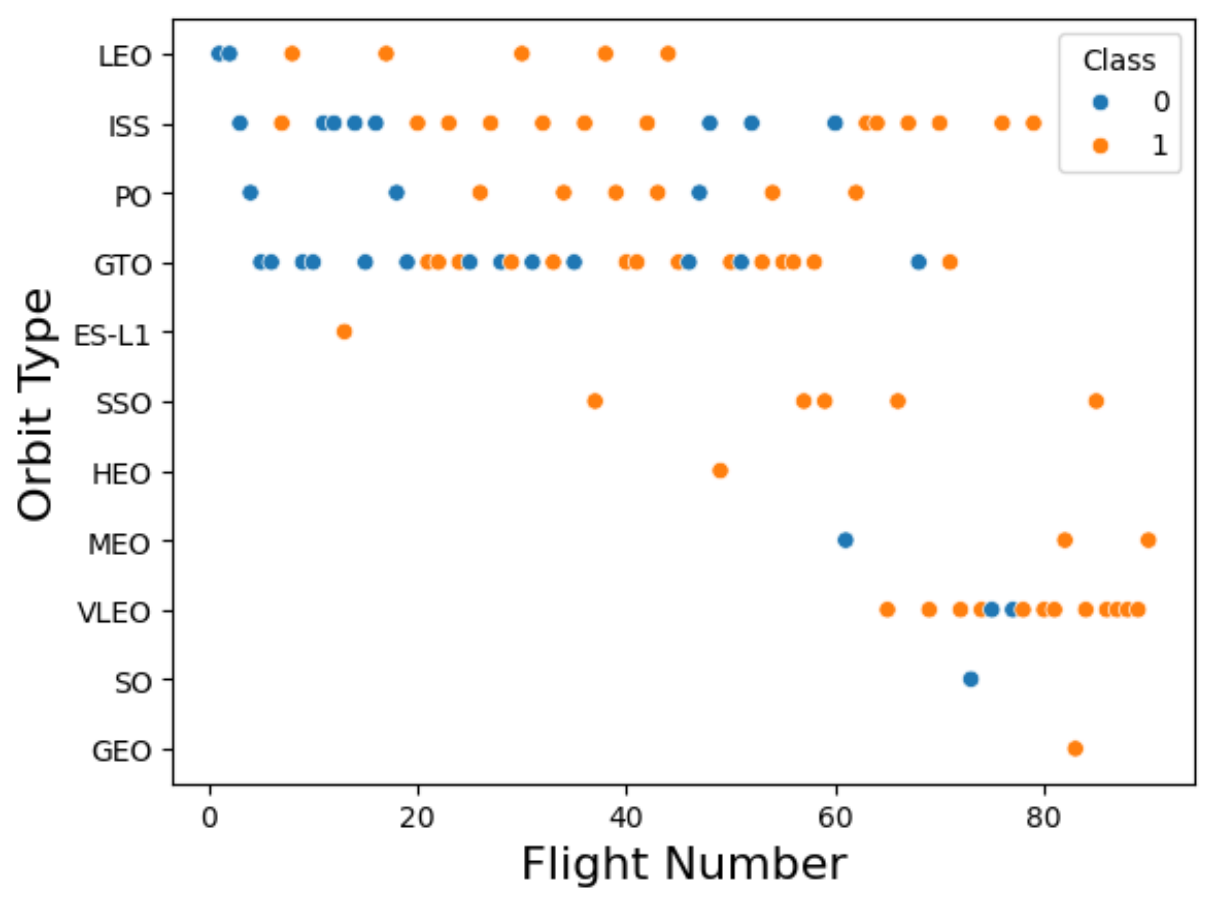
- Bar chart for the success rate of each orbit type

- Launches with the highest success rate were launches with orbit type ES-L1, GEO, ISS, and SSO.



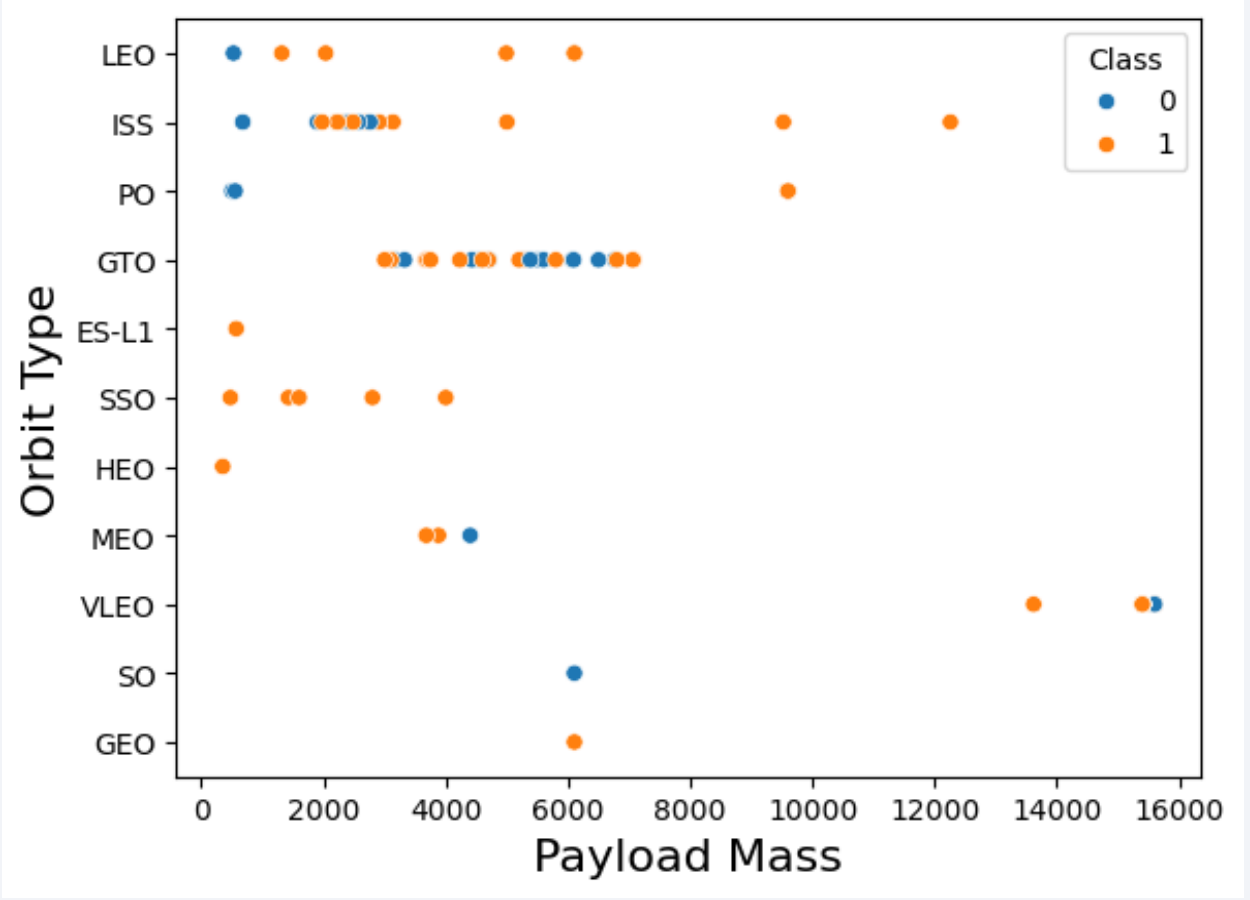
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type
- LEO orbit launches successes seems to be related with the flight numbers.
- However GTO orbit launches successes seems to be unrelated to the flight numbers.



Payload vs. Orbit Type

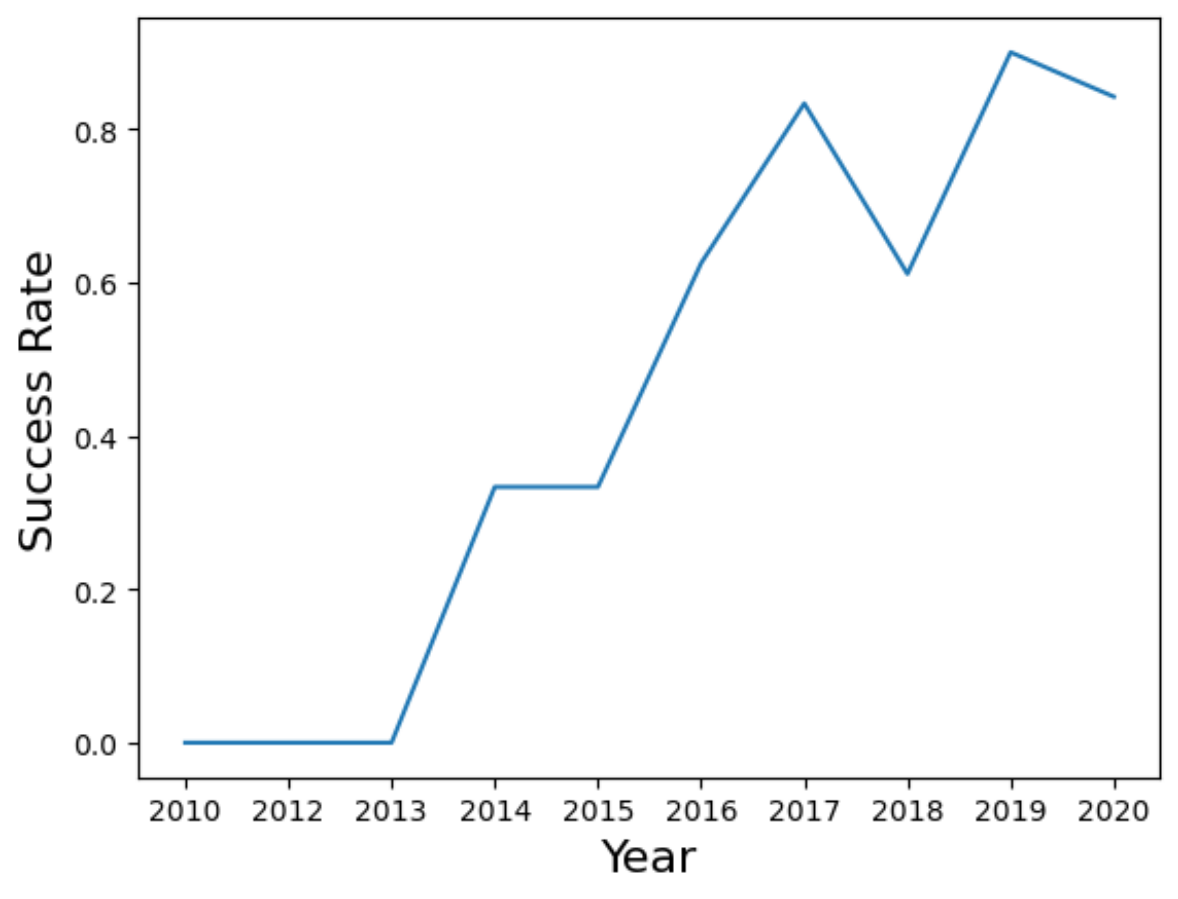
- Show a scatter point of payload vs. orbit type
- Orbit type PO, LEO and ISS seems to have more successful landings for launches with heavy payloads.
- However GTO orbit launches successes seems to have no relationship with payload mass.



Launch Success Yearly Trend

- Line chart of yearly average success rate

- The **success landing rate** for rocket launches **increases** from 2010 until 2020.



All Launch Site Names

- The names of the unique launch sites are as follows:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40
- The query used the **distinct** command line to avoid double entries and provide a list of unique launch site names.

```
In [9]: %sql select distinct(launch_site) from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[9]:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` are displayed beside.
- The first customers of the launches were from SpaceX itself and NASA.
- The query retrieve 5 records that have launch sites begins with 'CCA' .

In [10]:
```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5
```

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

Out[10]:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	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 carried by boosters from NASA (CRS) was 45596 kgs.
- The query used an aggregate function **sum** to sum all of the payload carried by the boosters with NASA (CRS) as the customers.

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [11]: %sql select sum(PAYLOAD_MASS__KG_) as sum_payload_mass from SPACEXTABLE where customer ='NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

Out[11]:

sum_payload_mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2534.67 kgs.
- The query used the aggregate function **avg** to retrieve the mean of the payload mass on booster versions that starts with 'F9 v1.1'.

```
In [12]: %sql select avg(PAYLOAD_MASS__KG_) as avg_payload_mass from SPACEXTABLE where booster_version like 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
```

Out[12]:

avg_payload_mass
2534.6666666666665

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were all starts with 'F9 FT'.
- The query retrieve the booster versions of all landing outcome that contains the words 'success' and 'drone' that have a payload mass between 4000 and 6000.

```
In [14]: %sql select booster_version from SPACEXTABLE where (lower(landing_outcome) like 'success%drone%' and (PAYLOAD_MASS__KG_ betv
 * sqlite:///my_data1.db
Done.
```

Out[14]:

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 outcomes and 1 failure mission outcome.
- The query groups the records by mission outcomes and retrieves the frequency of each mission outcomes.

List the total number of successful and failure mission outcomes

In [15]:
```
%sql select mission_outcome,count(mission_outcome) as total_number from SPACEXTABLE group by mission_outcome
```

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

Out[15]:

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

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was on December 22 2015
- The query use the aggregate function **min** to retrieve the earliest record where the landing outcome contains the word 'ground'.

```
In [13]: %sql select min(date) as first_success_date from SPACEXTABLE where lower(landing_outcome) like '%ground%'
 * sqlite:///my_data1.db
Done.
Out[13]: first_success_date
         2015-12-22
```

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass all starts with 'F9 B5', the maximum payload mass was 15600 kgs.
- The query retrieves the unique names of booster versions that have carried payloads of 15600 kgs.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [16]:
```
%sql select distinct(booster_version),PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS_
```

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

Out[16]:

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

2015 Launch Records

- There are 2 records in 2015 with a landing outcome of 'Failure (drone ship)'. Both have were launched from the same site with 2 different booster versions at 2 different months.
- The query used **substr** to retrieves months and year of the records.

In [17]:
```
%sql select substr(Date,6,2) as Month,Date,landing_outcome,booster_version,launch_site \
from SPACEXTABLE where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

Out[17]:

Month	Date	Landing_Outcome	Booster_Version	Launch_Site
01	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Between the date 2010-06-04 and 2017-03-20, the landing outcome with the highest frequency were records that have not attempt to land.
- The query groups the records by landing outcomes and orders them by the frequency of each outcome in a descending manner.

In [18]:
```
%sql select landing_outcome,count(*) as outcome_count from SPACEXTABLE where date between '2010-06-04' and '2017-03-20'\
group by landing_outcome order by outcome_count desc
```

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

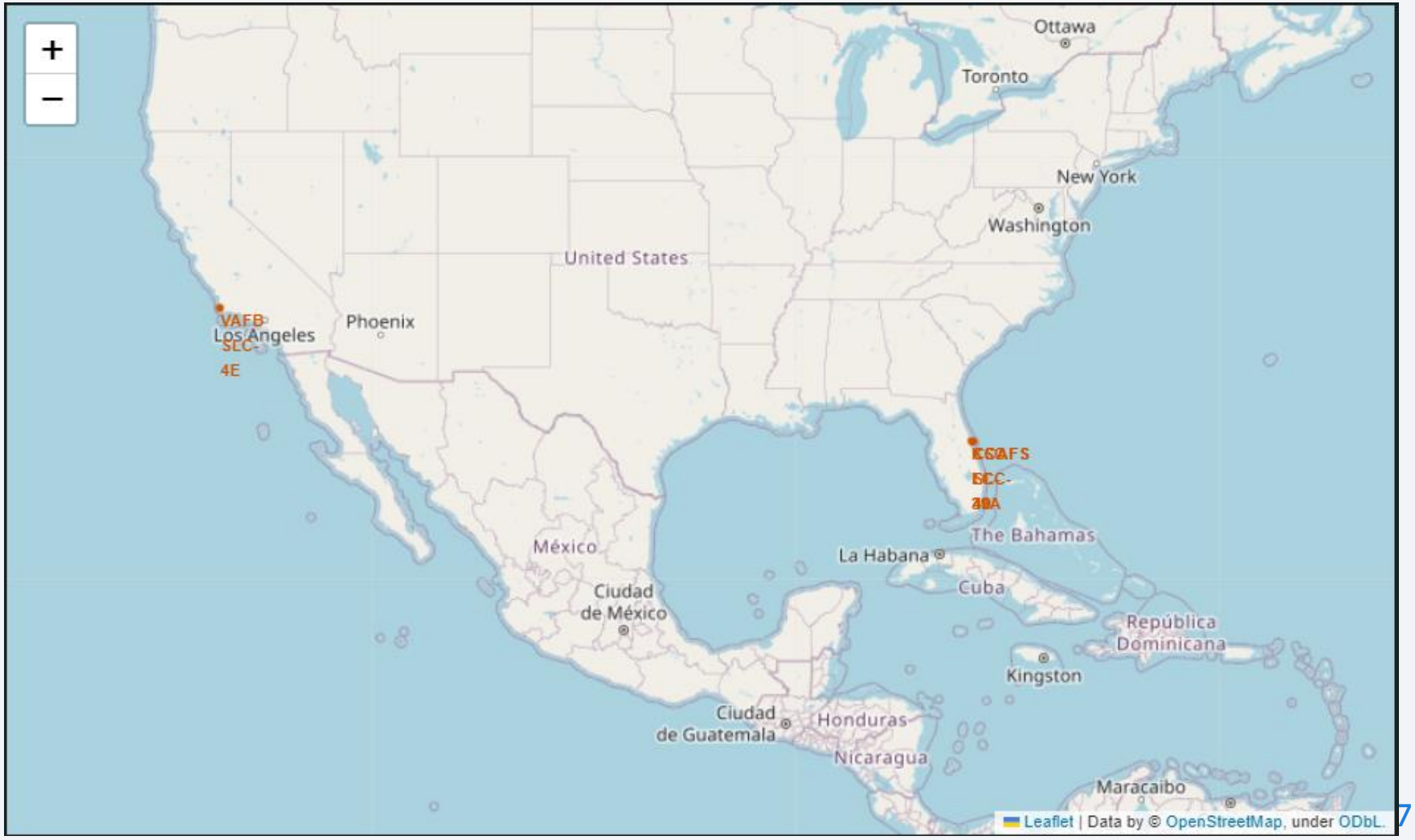
Out[18]:

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


Section 3
Launch Sites
Proximities Analysis

All Launch Sites Location Markers

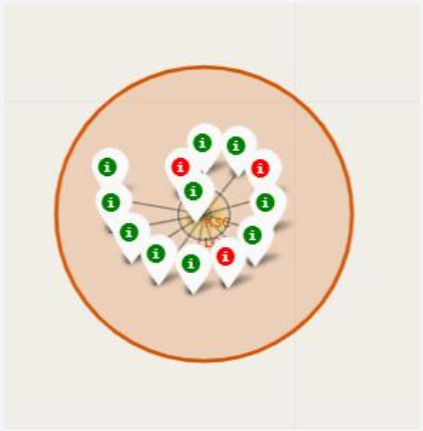
All of SpaceX launch sites are located on USA coasts near Los Angeles and Florida.



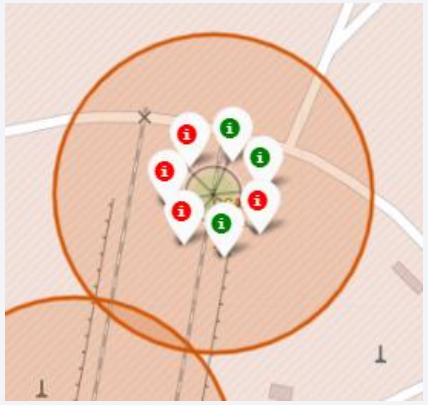
Success/Failed Launches for Each Site Markers

- VAFB SLC-4E

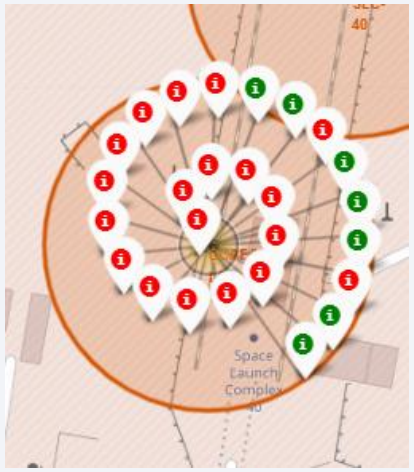
- KSC LC-39A



- CCAFS SLC-40



- CCAFS LC-40



Where **green** markers indicate successful landings and **red** markers indicate unsuccessful landings.

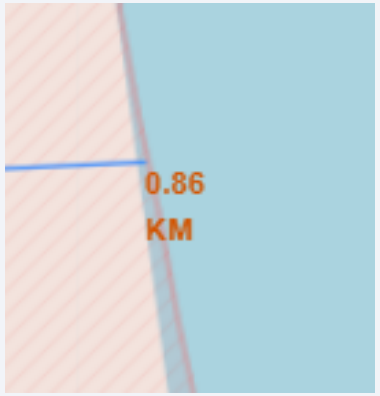
Distance between a launch site to its Proximities

- Launch site CCAFS SLC-40 to its proximities:
  - Are launch sites in close proximity to railways? Yes, the closest railway is at 1.30 km away from the launch site.
  - Are launch sites in close proximity to highways? Yes, the closest highway is at 0.59 km away from the launch site.
  - Are launch sites in close proximity to coastline? Yes, the closest coastline is at 0.86 km away from the launch site.
  - Do launch sites keep certain distance away from cities? Yes, the closest city is at 18.17 km away from the launch site.

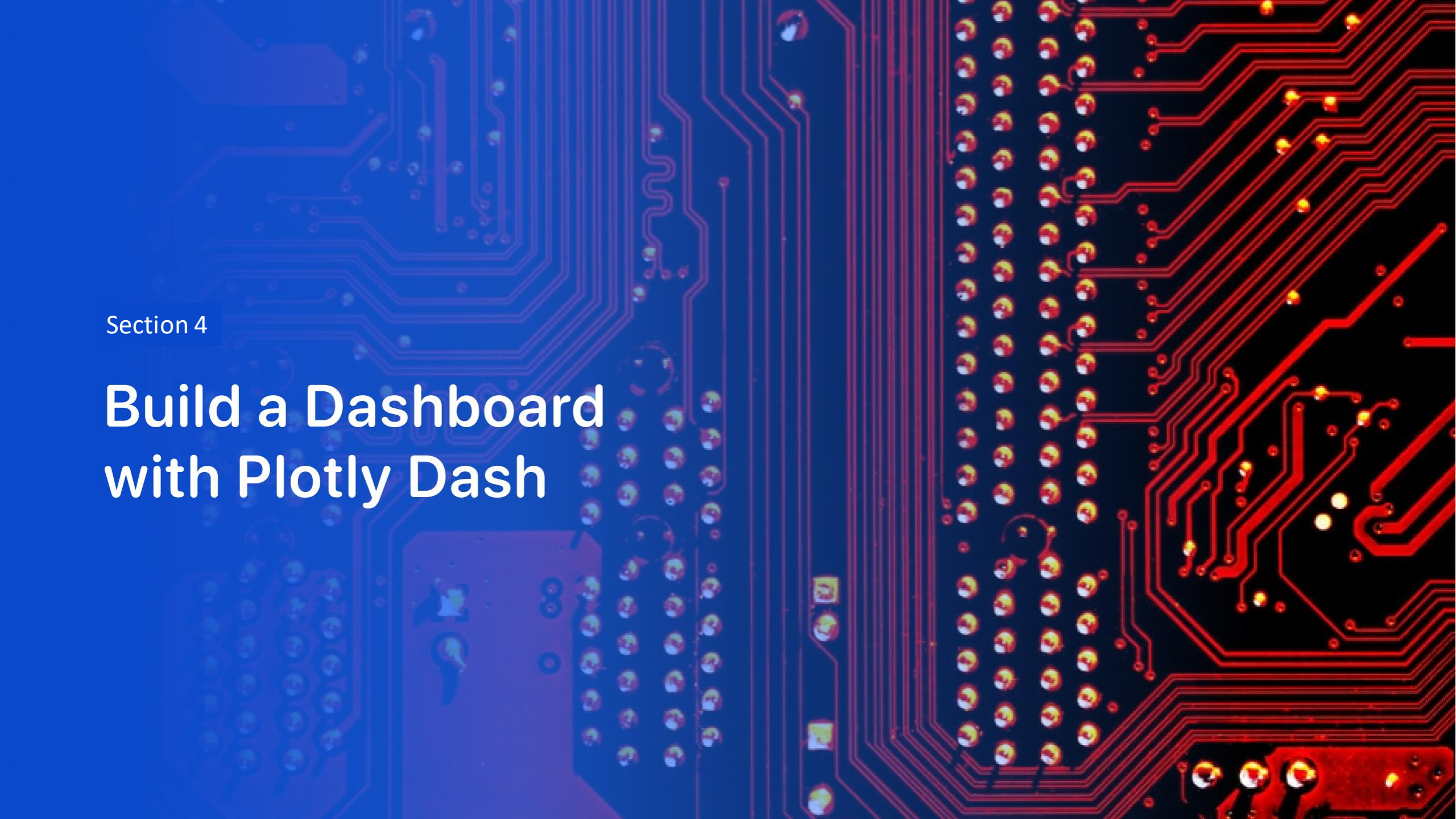
Nasa Railroad

Samuel C Phillips Parkway

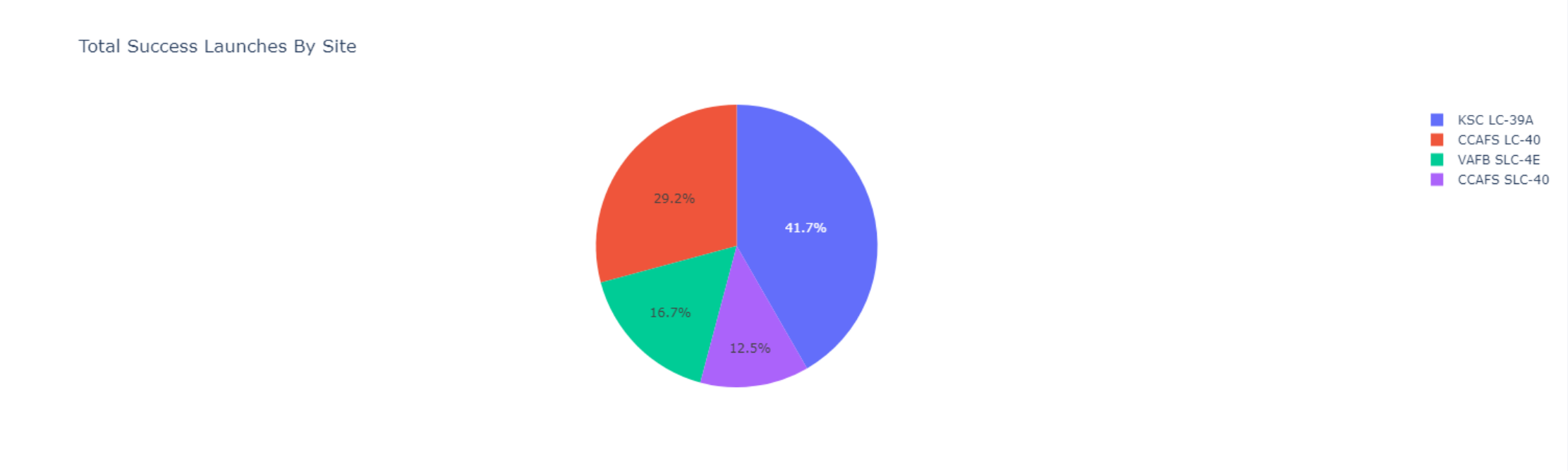
Florida Coastline



Cape Canaveral

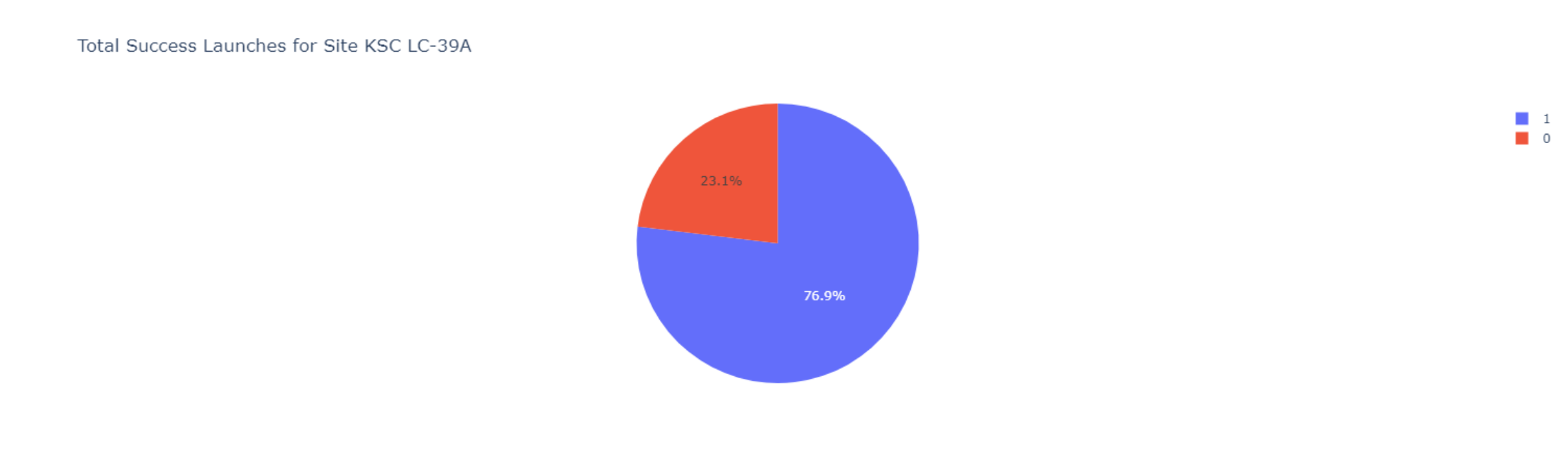

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches By Site



- The launch site **KSC LC-39A** have the highest number of successful launches at **41.7%** out of total successful launches from all sites.

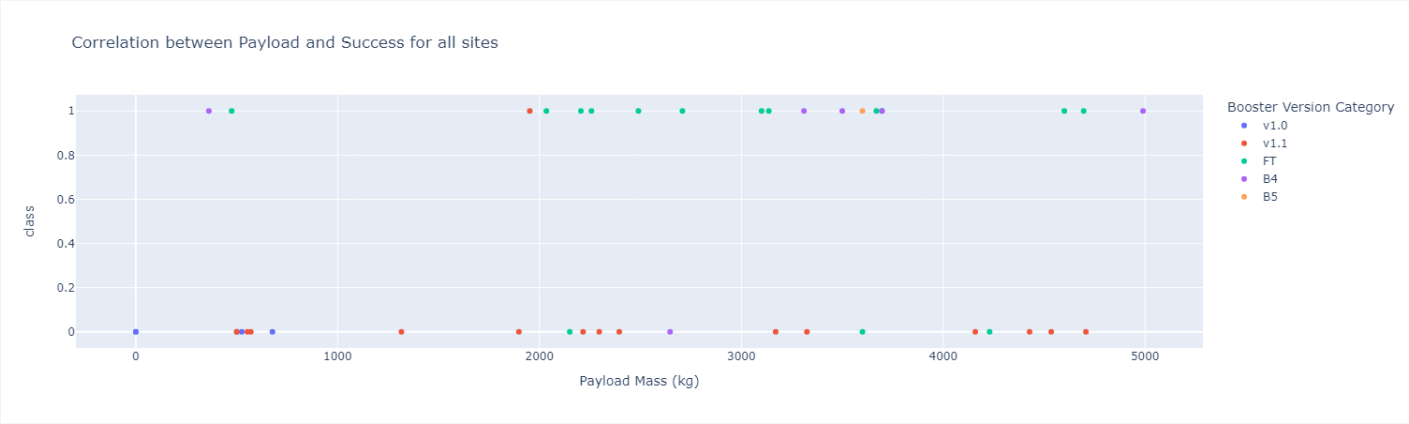
Launch Site with Highest Launch Success Ratio



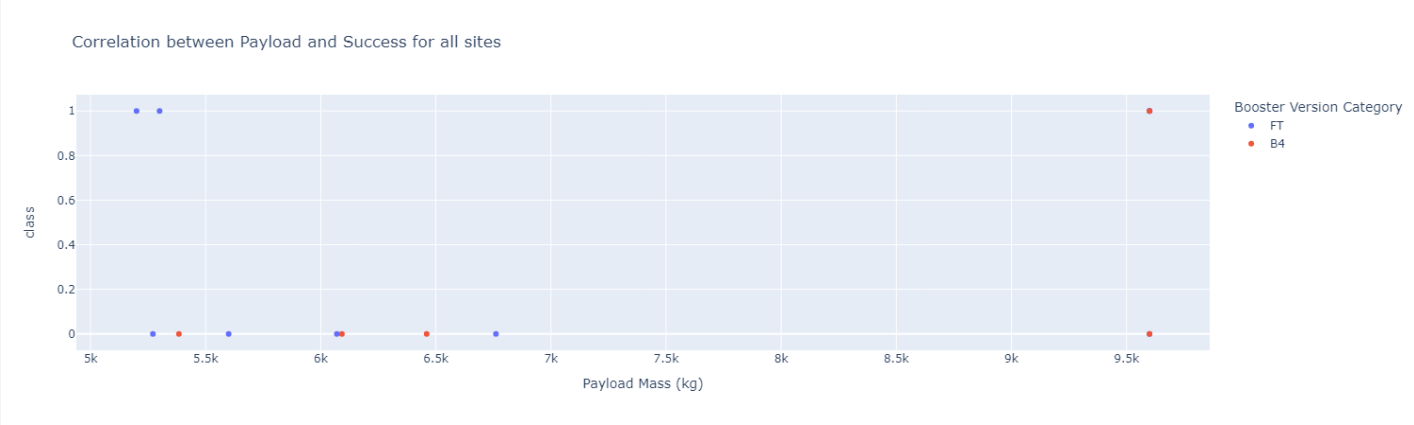
- Not only the launch site **KSC LC-39A** have the highest number of successful launches, it also have the highest launch success ratio of **76.9%**.

Payload vs. Launch Outcome

- Launches with payload between 0 kg and 5000 kgs.



- Launches with payload between 5000 kg and 10000 kgs.



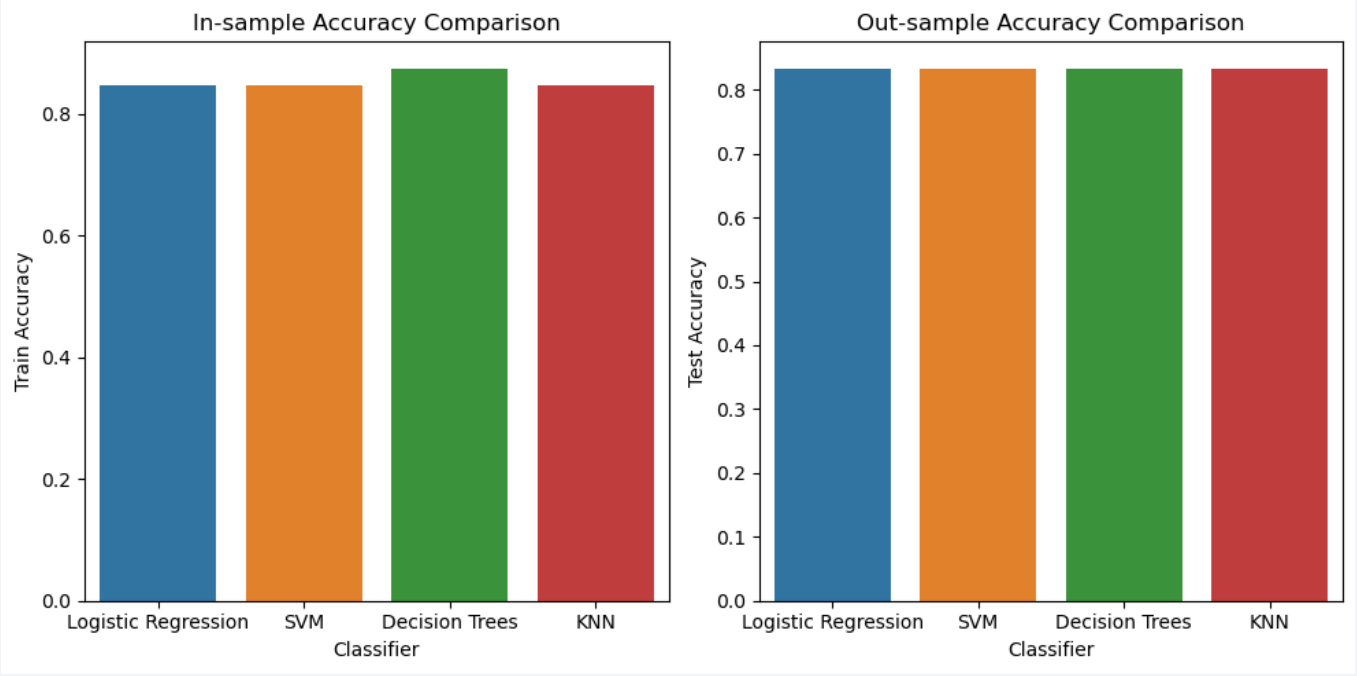
- Launches with **heavier payloads** have **lower launch success ratio** compared to the ones with lighter payload mass.

Section 5

Predictive Analysis (Classification)

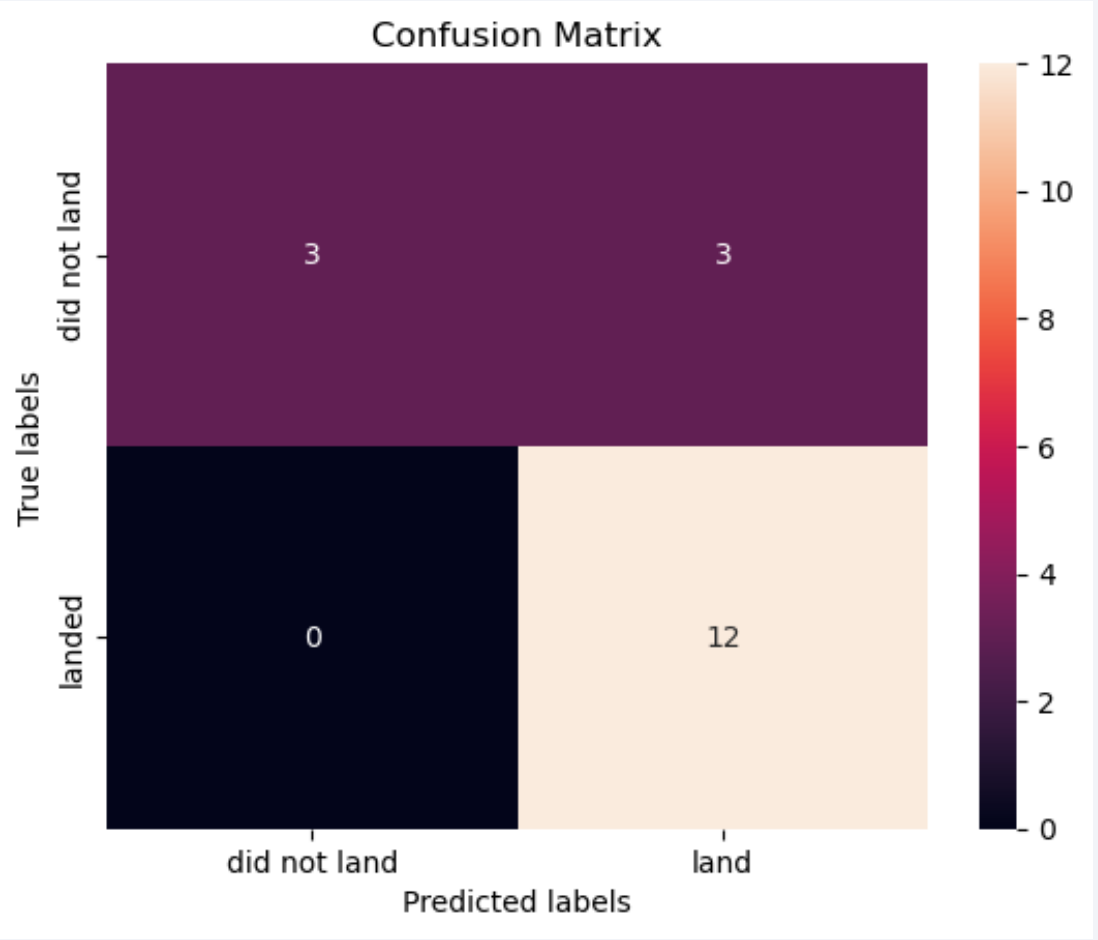
Classification Accuracy

- The **Decision Trees** model have the best performance for **in-sample (training) accuracy** at **88.75%**.
- **All models** have the same performance for **out-sample (testing) accuracy** at **83.33%.**



Confusion Matrix

- As all of the models have similar performance in out-sample accuracy score. All of the models also produced a similar Confusion Matrix.
- One thing to consider is the false positive error where there are 3 launch records where the landing was predicted as a successful land when in reality it was an unsuccessful land.



Conclusions

- Launches with **higher flight number** have a **higher chance of success.**
- The **success landing rate** for rocket launches **increases** from 2010 until 2020.
- Launches with the **highest success rate** were launches with orbit type **ES-L1**, **GEO**, **ISS**, and **SSO**.
- Launches with **heavier payloads** have **lower launch success ratio** compared to the ones with lighter payload mass.
- Not only the launch site **KSC LC-39A** have the highest number of successful launches at **41.7%** out of total successful launches from all sites, it also have the highest launch success ratio of **76.9%**.
- Out of the 4 classification models tested in this project, the were **no best** model that results in the best out-sample accuracy compared to others at **83.33%**. However, the decision trees classification models performed the best in the in-sample accuracy test at **88.75%.**

Thank you!