Winning Space Race with Data Science

Parker Jackson
5/24/22

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection through Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
- Summary of all results
  - Exploratory Data Analysis results
  - Machine Learning Prediction results

Introduction

- SpaceX has a cost advantage over its competitors because it has reusable first stage rockets like the Falcon 9.
- Can we use data collected on the internet to predict the successful recovery rate of the reusable rockets?

Section 1

Methodology

Methodology

Executive Summary

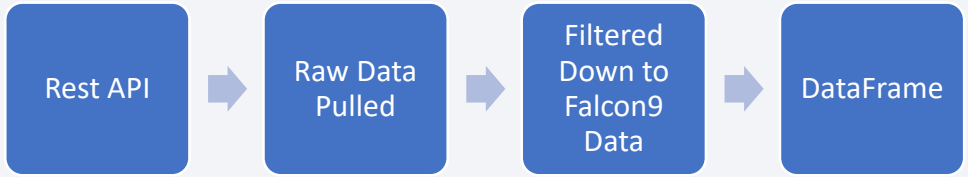
- Data collection methodology:
  - The data was scraped from the internet using the SpaceX API
- Perform data wrangling
  - The data was put into a Pandas data frame with each column given a proper label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was split into test and training datasets put through a grid search to pick the best tuning parameters to predict and fit multiple classification models

Data Collection

- The data was collected through a SpaceX Rest API and scraped from the internet
- Python was used with the Pandas library to populate Dataframes using the data collected

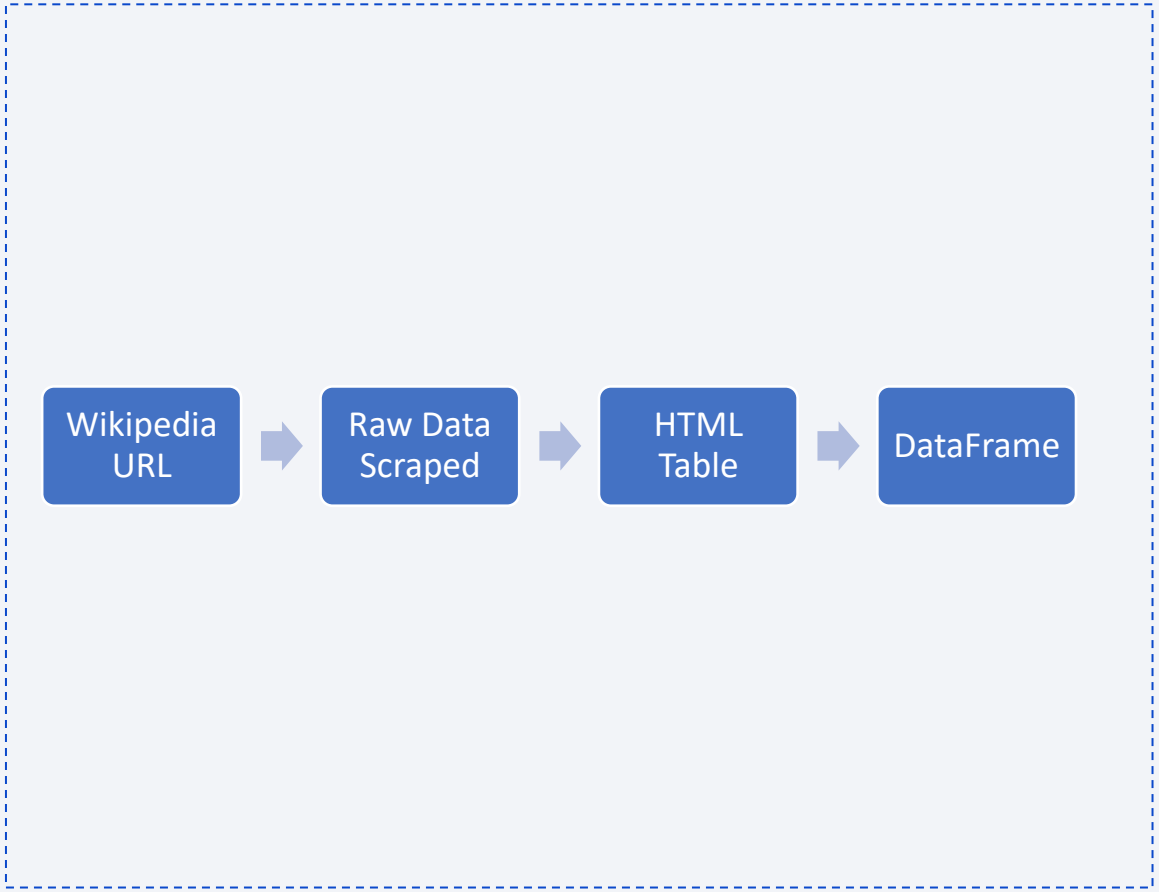
Data Collection – SpaceX API

- https://github.com/parksandrecs/solid-dollop/blob/d9e83f40099f048e0f8a67e6f6d9ce0e791ce7b8/jupyter_labs_spacex_data_collection_api.ipynb

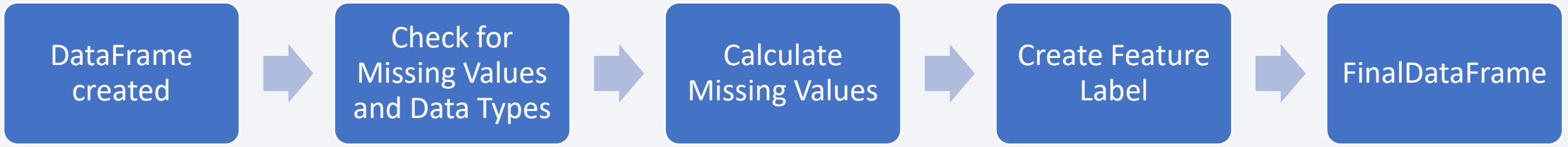


Data Collection - Scraping

- https://github.com/parksandrecs/solid-dollop/blob/aceecde4bfdc149959e012f81bab68ef9d71aeb0/jupyter_labs_webscraping.ipynb



Data Wrangling



- https://github.com/parksandrecs/solid-dollop/blob/d9e83f40099f048e0f8a67e6f6d9ce0e791ce7b8/labs_jupyter_spacex_Data_wrangling.ipynb

EDA with Data Visualization

- I graphed the relationship between flight number, payload weight, and the success rate
- https://github.com/parksandrecs/solid-dollop/blob/d9e83f40099f048e0f8a67e6f6d9ce0e791ce7b8/jupyter_labs_eda_dataviz.ipynb

EDA with SQL

- The queries found
  - the launch site names
  - Total payload weight carried by NASA(CRS) missions
  - Average payload carried by booster version F9 v1.1
  - Total number of missions by grouped by outcome type
  - Failures by drone ship and the details of the mission
- https://github.com/parksandrecs/solid-dollop/blob/d9e83f40099f048e0f8a67e6f6d9ce0e791ce7b8/sql_eda.ipynb

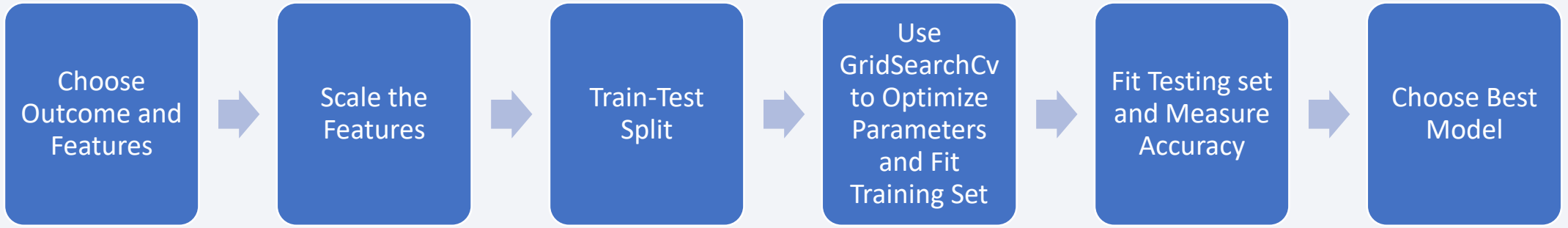
Build an Interactive Map with Folium

- The interactive map has a marker for every launch site
- Each launch site has a green marker for every successful mission and a red marker for every failure
- There is a measurement line from the eastern most launch site to the nearest coastline
- https://github.com/parksandrecs/solid-dollop/blob/d9e83f40099f048e0f8a67e6f6d9ce0e791ce7b8/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- I included pie charts representing success rate by launch site to show the relationship between the two
- I included a scatter plot to show the relationship between the payload weight and the success rate
- https://github.com/parksandrecs/solid-dollop/blob/d9e83f40099f048e0f8a67e6f6d9ce0e791ce7b8/Dashboard.py

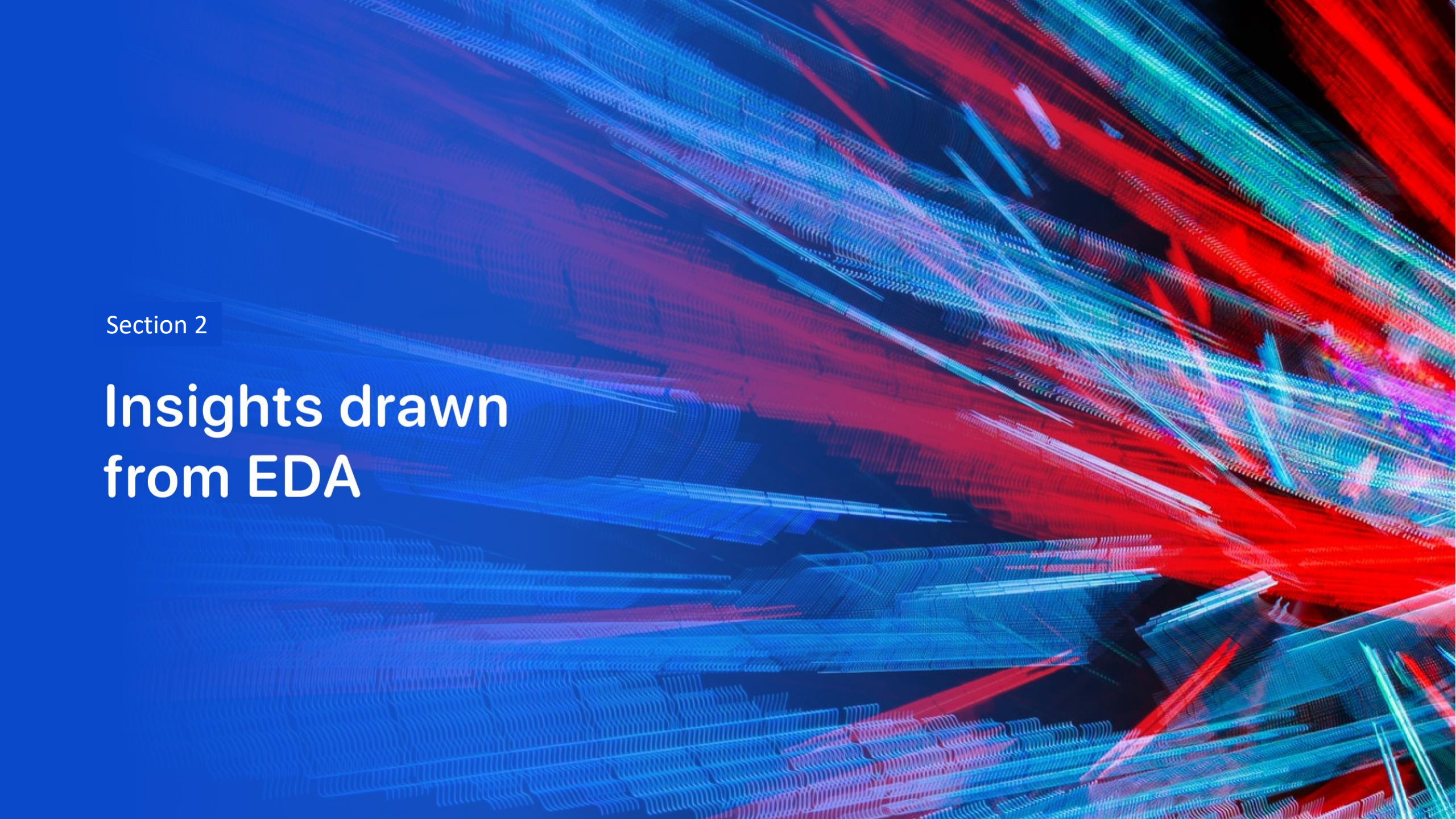
Predictive Analysis (Classification)



- https://github.com/parksandrecs/solid-dollop/blob/d9e83f40099f048e0f8a67e6f6d9ce0e791ce7b8/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Results

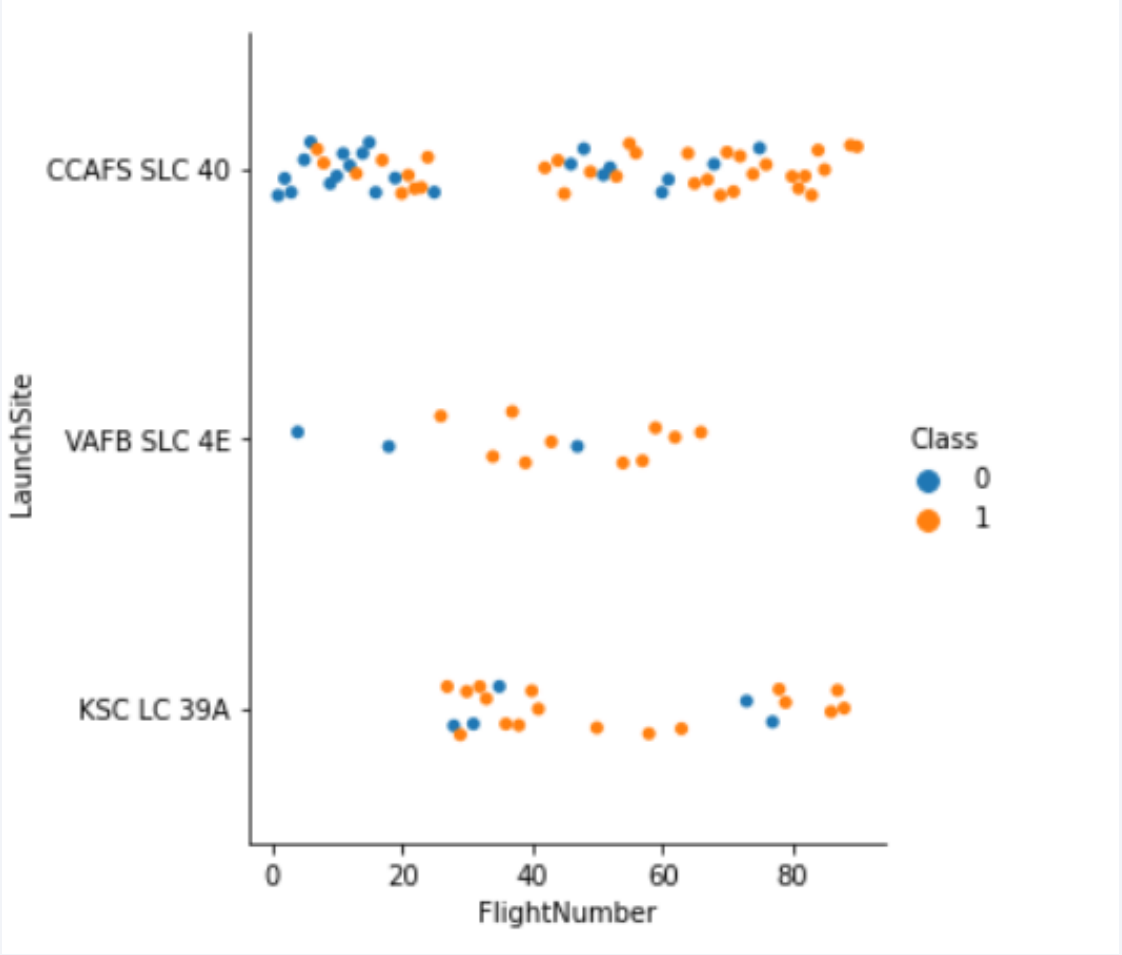
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

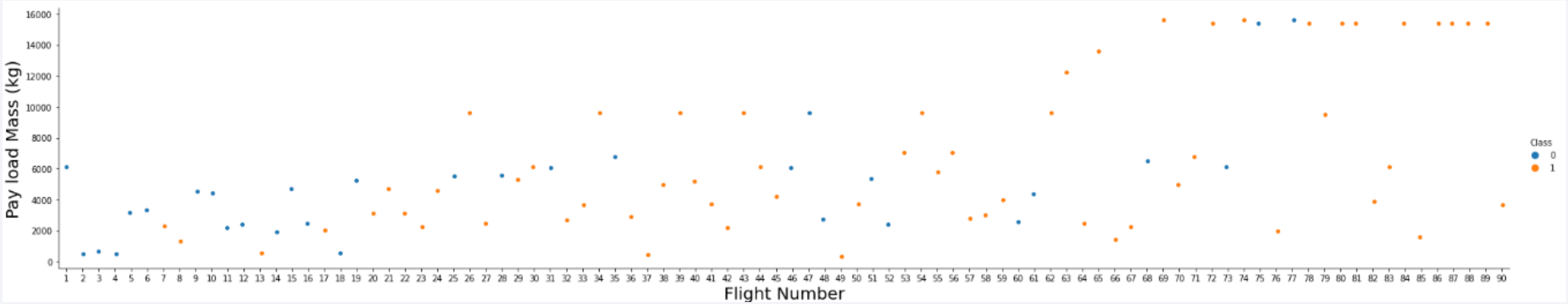
- Blue dots show rocket recovery failures while orange shows successes

- SpaceX has improved its recovery success rate over time



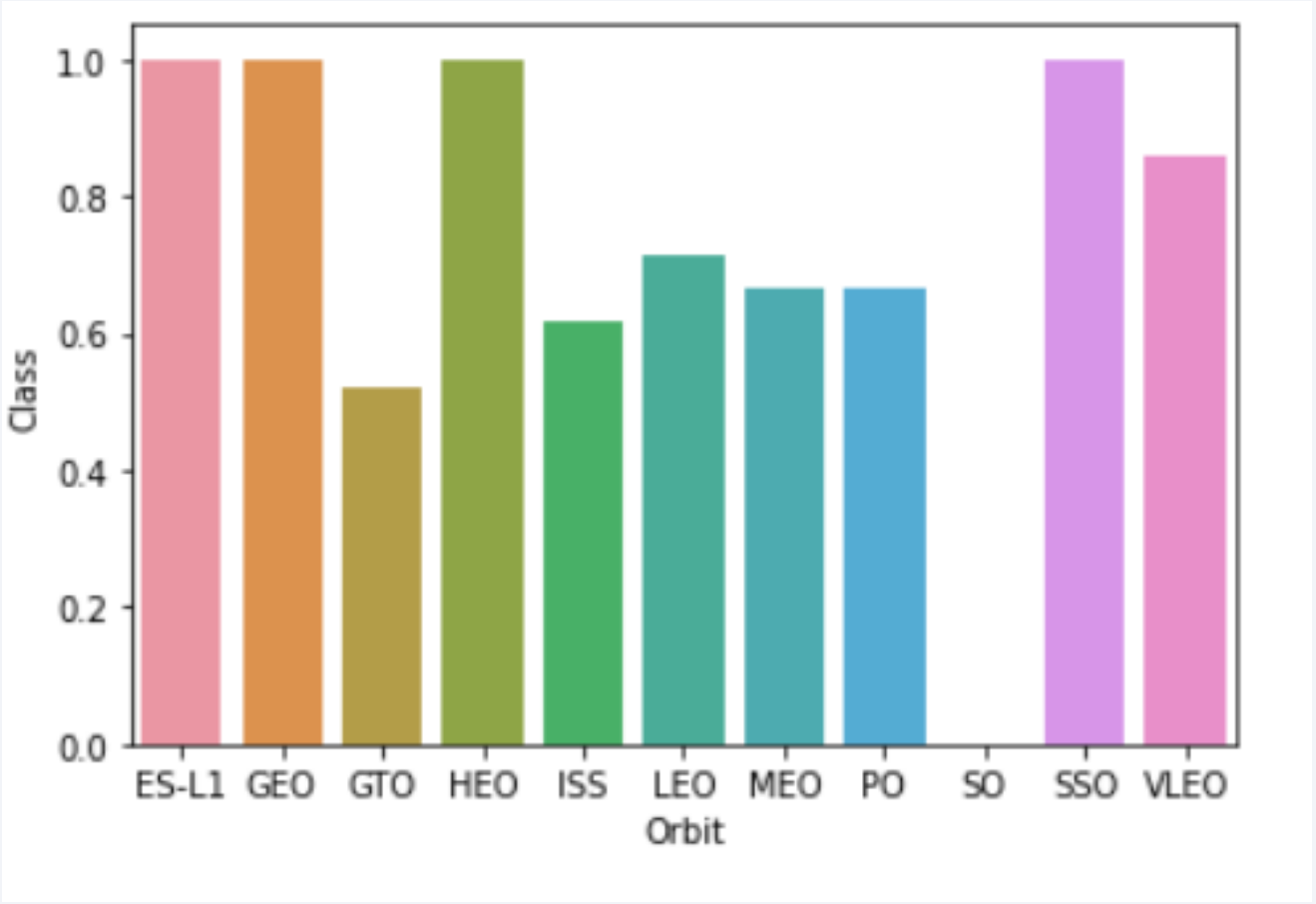
Payload vs. Launch Site

- Blue dots show rocket recovery failures while orange shows successes
- We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



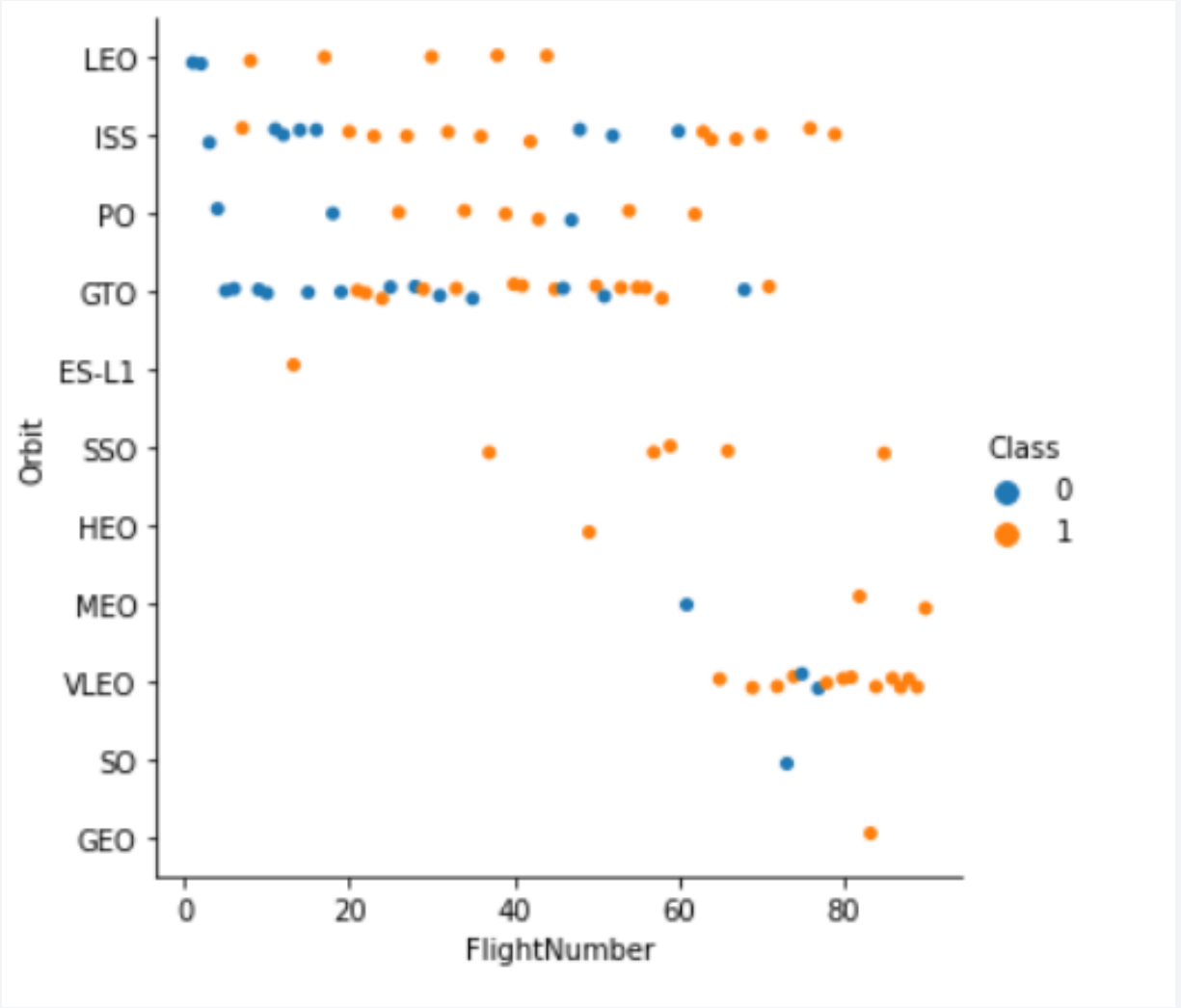
Success Rate vs. Orbit Type

- Orbit types affect success rate
- ES-L1, GEO, HEO, and SSO have the highest success rate



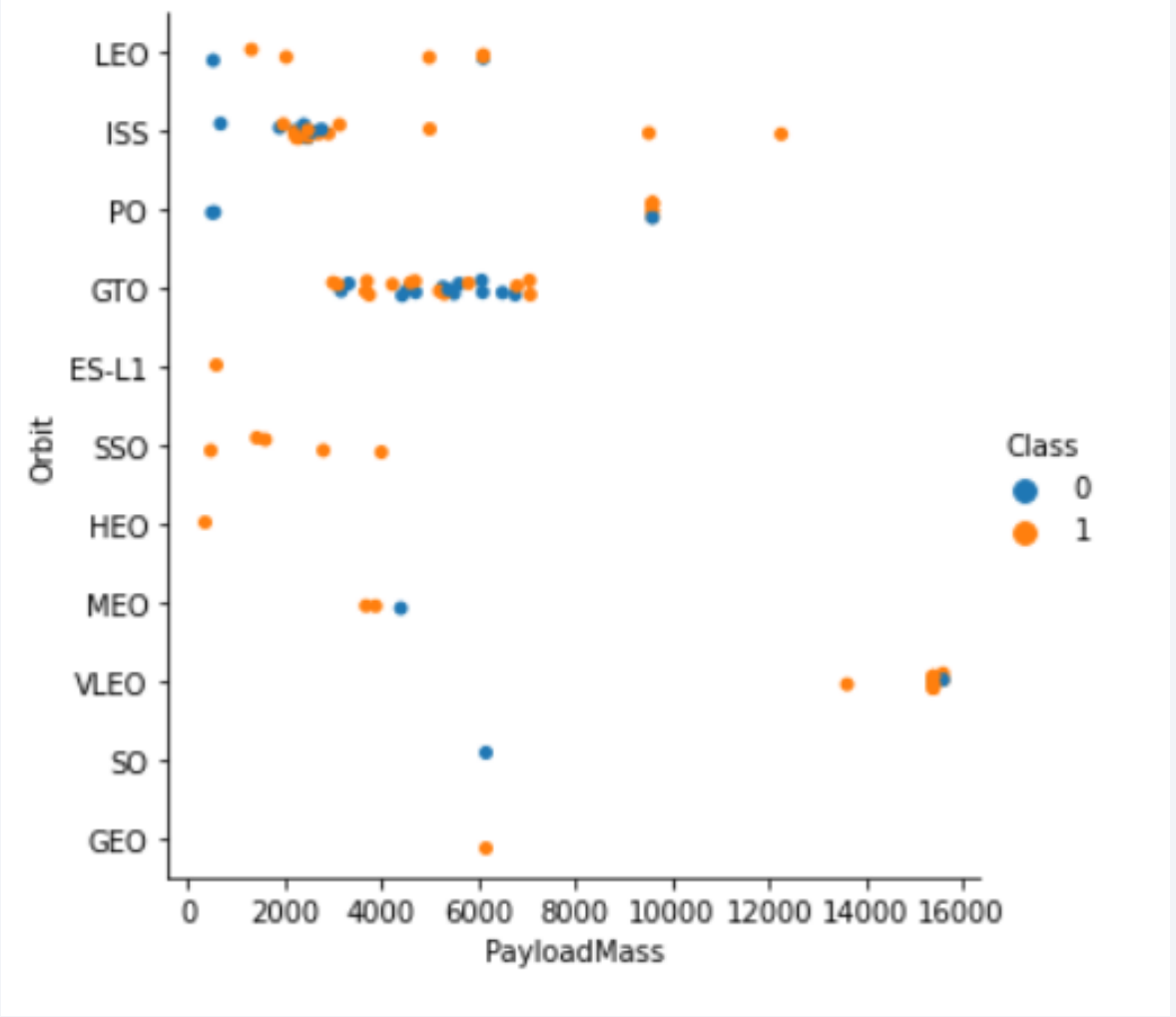
Flight Number vs. Orbit Type

- Blue dots show rocket recovery failures while orange shows successes
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit



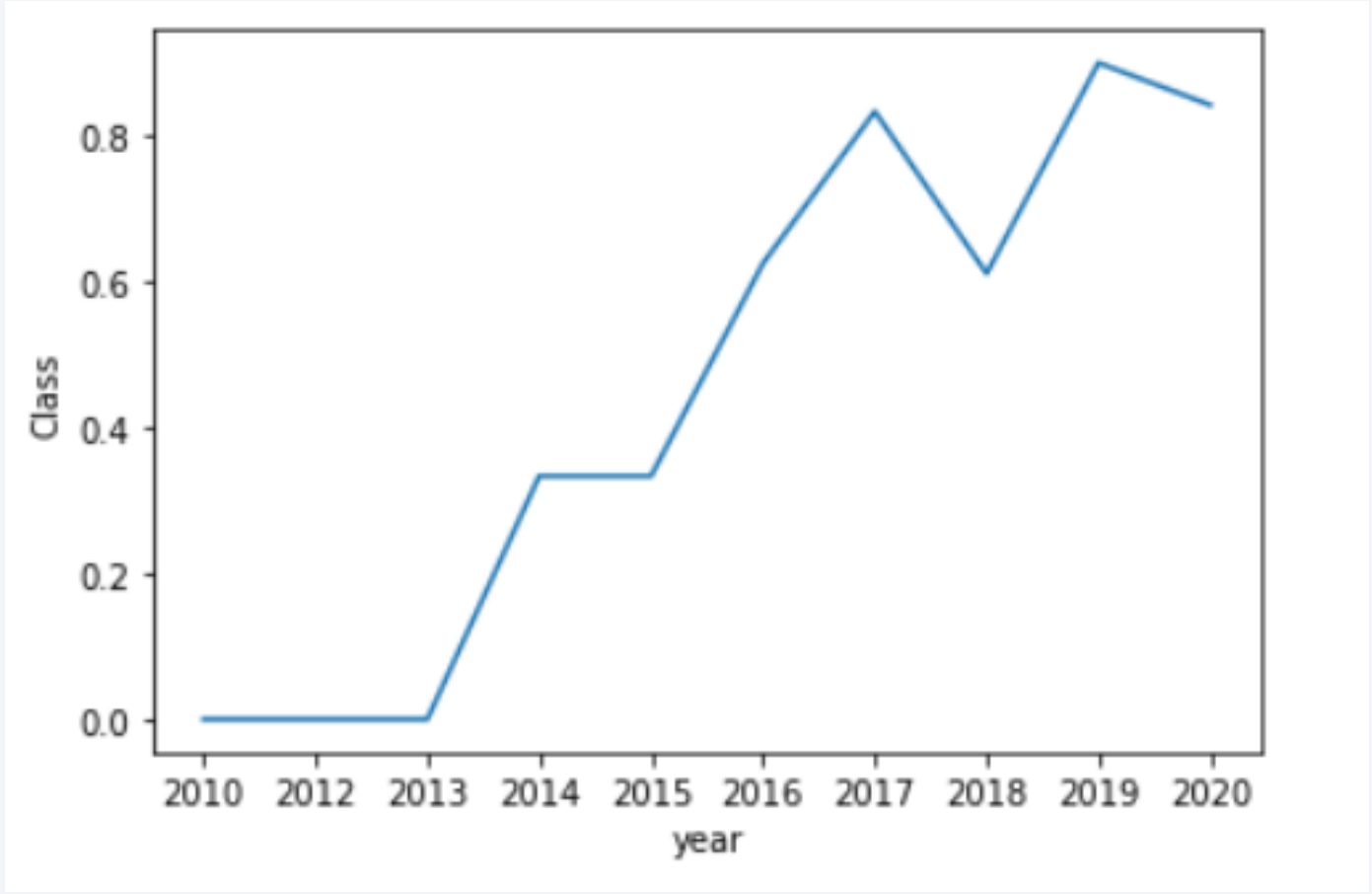
Payload vs. Orbit Type

- Blue dots show rocket recovery failures while orange shows successes
- With heavy payloads the successful landing or positive landing rate are more for Po, LEO, and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and (unsuccessful mission) are both there here



Launch Success Yearly Trend

- You can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- Here is a list of all the unique launch sites that SpaceX uses

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [23]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' Limit 5
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[23]:

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

Task 3

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

In [28]:
```
%sql SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0
Done.
```

Out[28]:

1
45596

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

In [29]:
```
%sql SELECT avg(payload_mass__kg_) FROM SPACEXTBL WHERE booster_version  like 'F9 v1.1%'
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.dat
Done.
```

Out[29]:

1
2534

First Successful Ground Landing Date

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [35]:
```
%sql SELECT min(DATE)  as date FROM SPACEXTBL WHERE landing__outcome  = 'Success (ground pad)'
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.
Done.
```

Out[35]:

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [36]:
```
%sql SELECT distinct(booster_version)  as booster_name FROM SPACEXTBL WHERE landing__outcome  = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[36]:

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [40]:
```
%sql SELECT distinct(mission_outcome) , count(booster_version) as number FROM SPACEXTBL group by mission_outcome
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud
Done.
```

Out[40]:

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

Boosters Carried Maximum Payload

Task 8

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

In [41]:
```
%sql SELECT distinct(booster_version)  as booster_name FROM SPACEXTBL WHERE payload_mass__kg_  = (select max(payload_mass__kg_) from SPACEXTBL )
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[41]:

booster_name
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

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [50]:
```
%sql SELECT date, landing__outcome , booster_version , launch_site FROM SPACEXTBL WHERE DATE like '2015%' and landing__outcome like 'Failure%'
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[50]:

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

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

Task 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

In [48]:
```
%sql SELECT distinct(landing__outcome) , count(landing__outcome) as count FROM SPACEXTBL
%WHERE DATE between '2010-06-04'and  '2017-03-20' group by landing__outcome order by count desc
```

```
 * ibm_db_sa://hpv88982:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[48]:

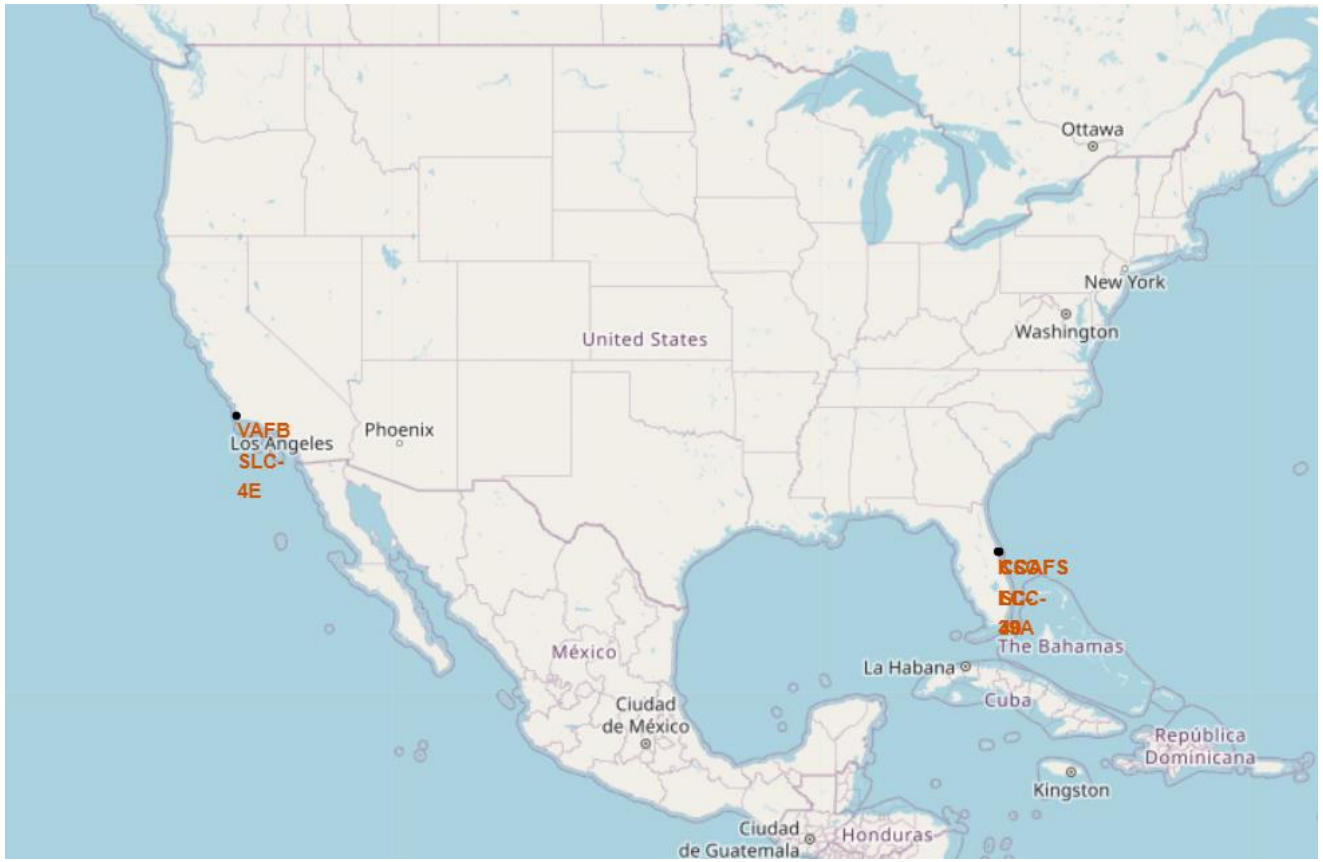
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

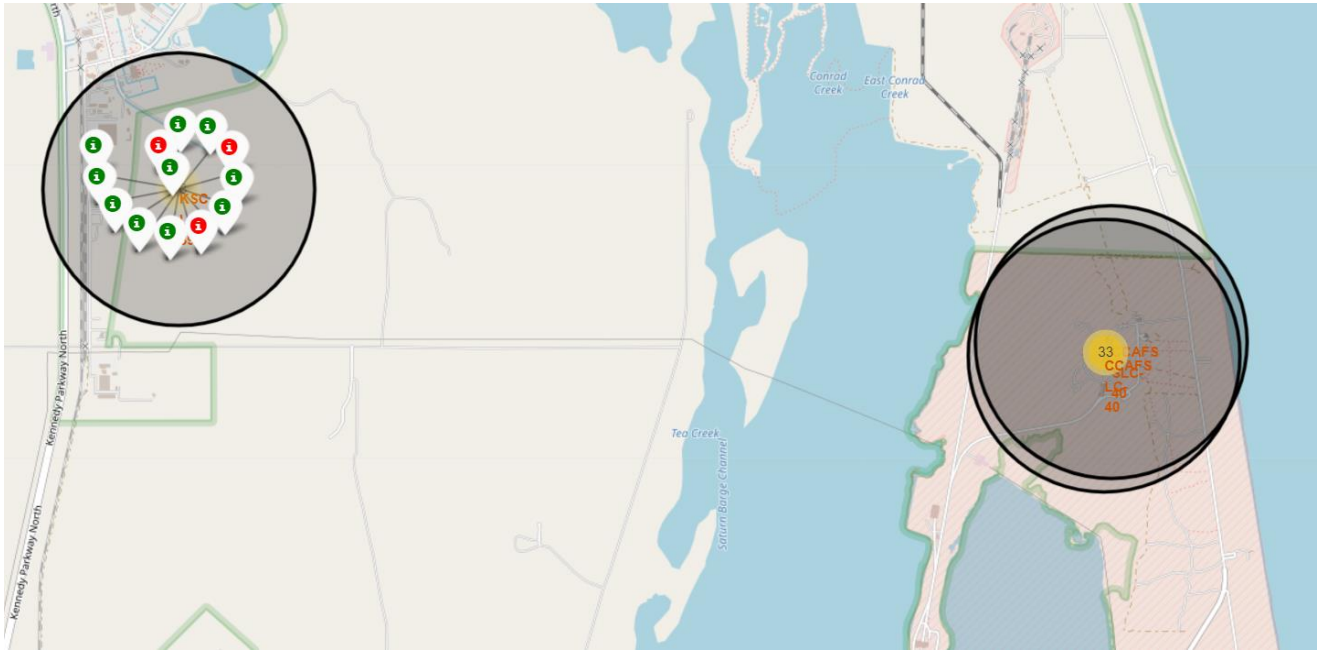
SpaceX Map of Launch Sites

- There are four launch sites in total
- The VAFB is on the west coast in southern California while the other three are on the east coast is northern Florida



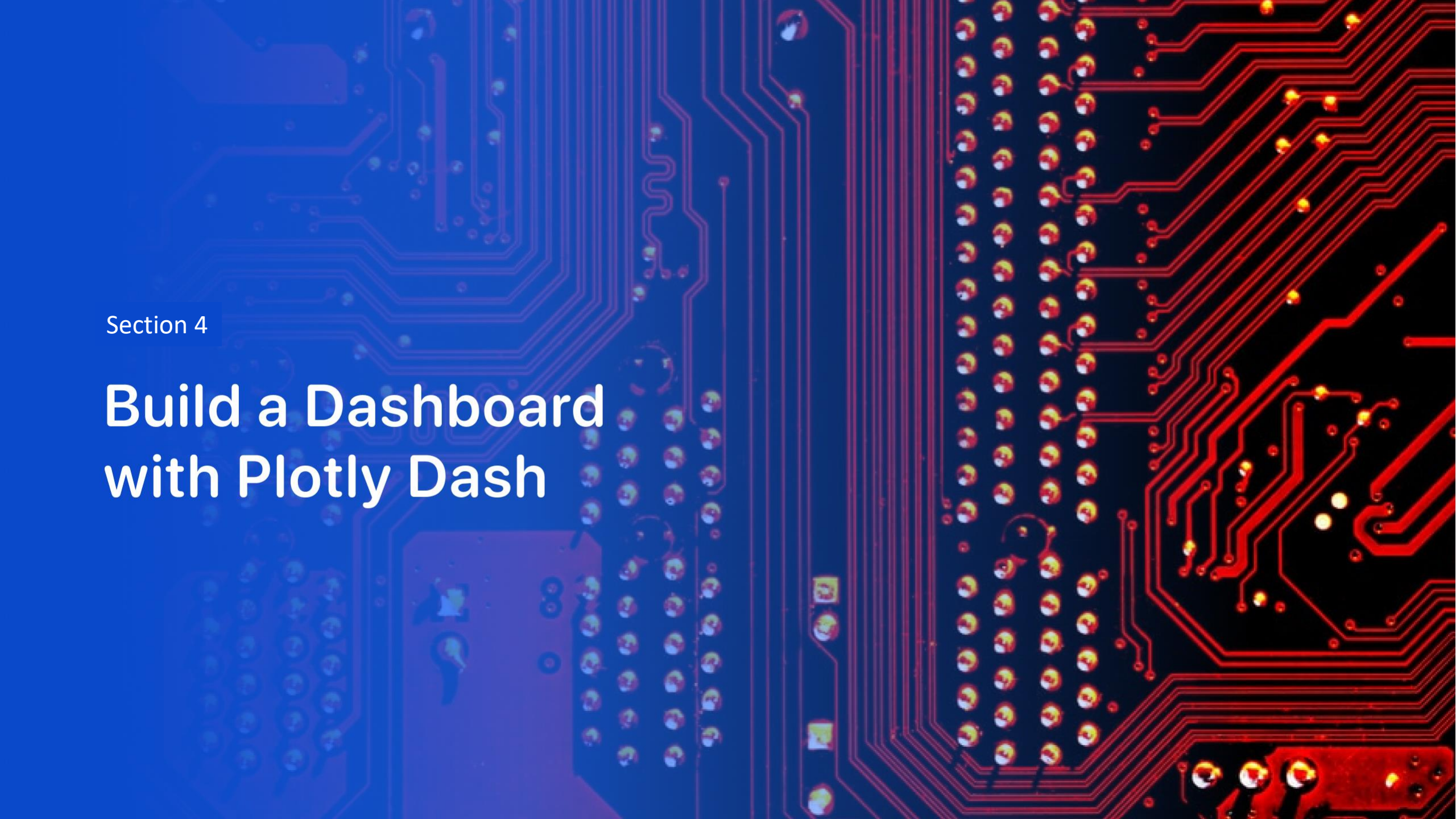
SpaceX Success and Failure Map

- This is a close-up map of the three launch sites in Florida. Only the KSC LC-39A launch site is toggled to show the success and failure incidents.
- The green markers represent successful recovery of a first stage rocket. The red represent failures.



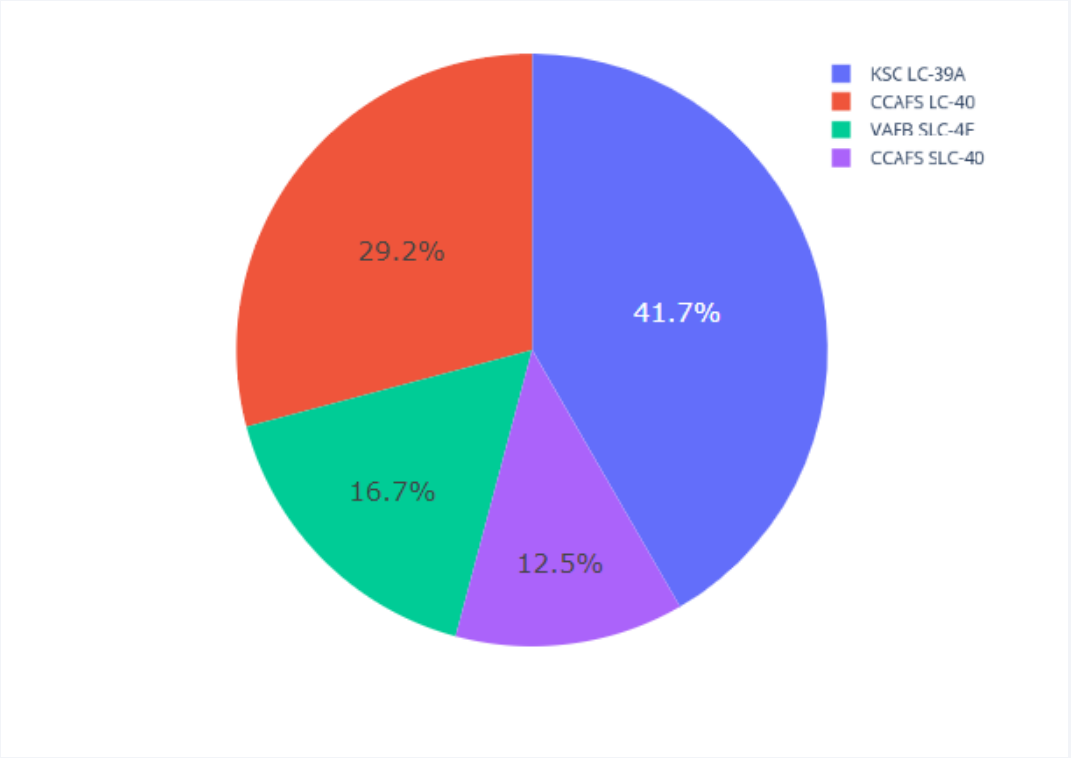
Distance Marker on Launch Site

- This map shows a distance marker from the CCAFS SLC-40 launch site to the nearest coastline


Section 4
Build a Dashboard with Plotly Dash

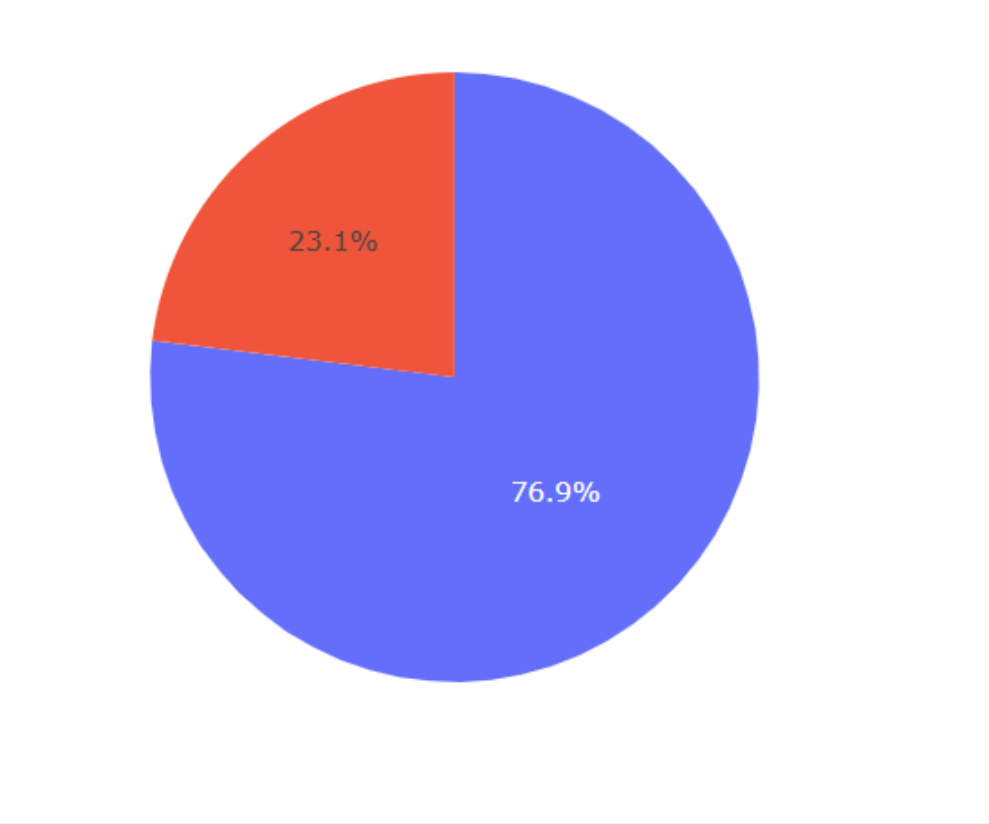
Success Rate of Launch Sites

- This pie chart shows the recovery success rate for all SpaceX launch sites



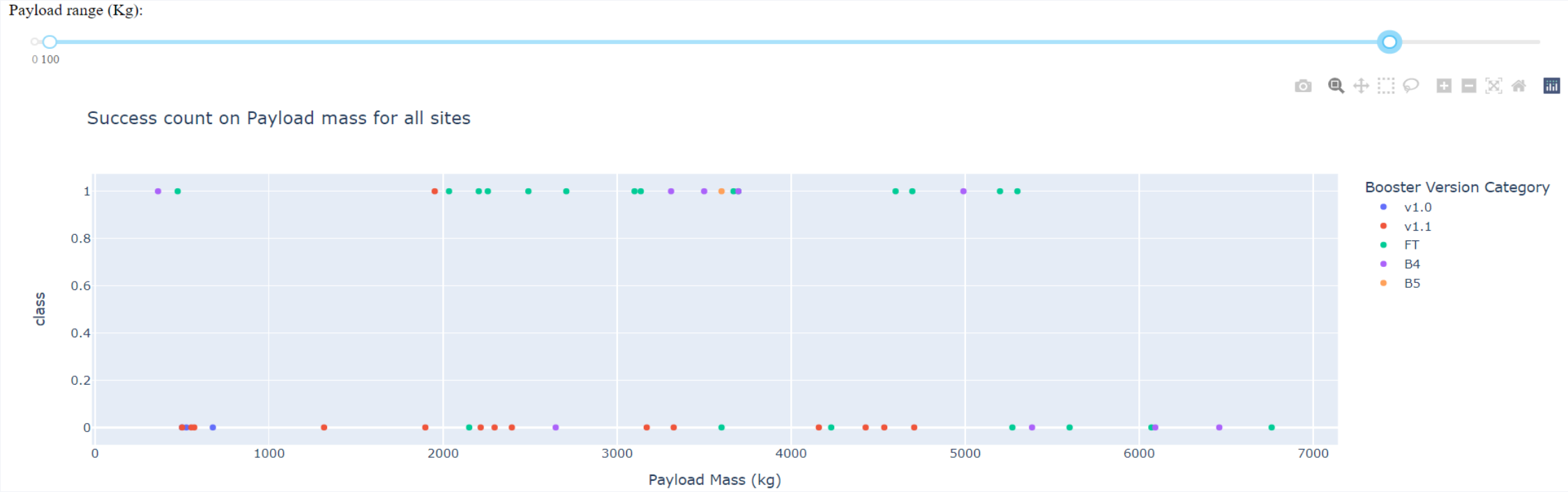
KSC LC-39A Success Rate

- This pie chart shows KSC LC-39A's recovery success rate, which is the highest of all the SpaceX launch sites.
- Blue represents success rate, red is failure rate



Payload vs. Launch Outcome

- The FT booster type has the most successes, and those successes are concentrated in the 2000 to 4000 kg payload range

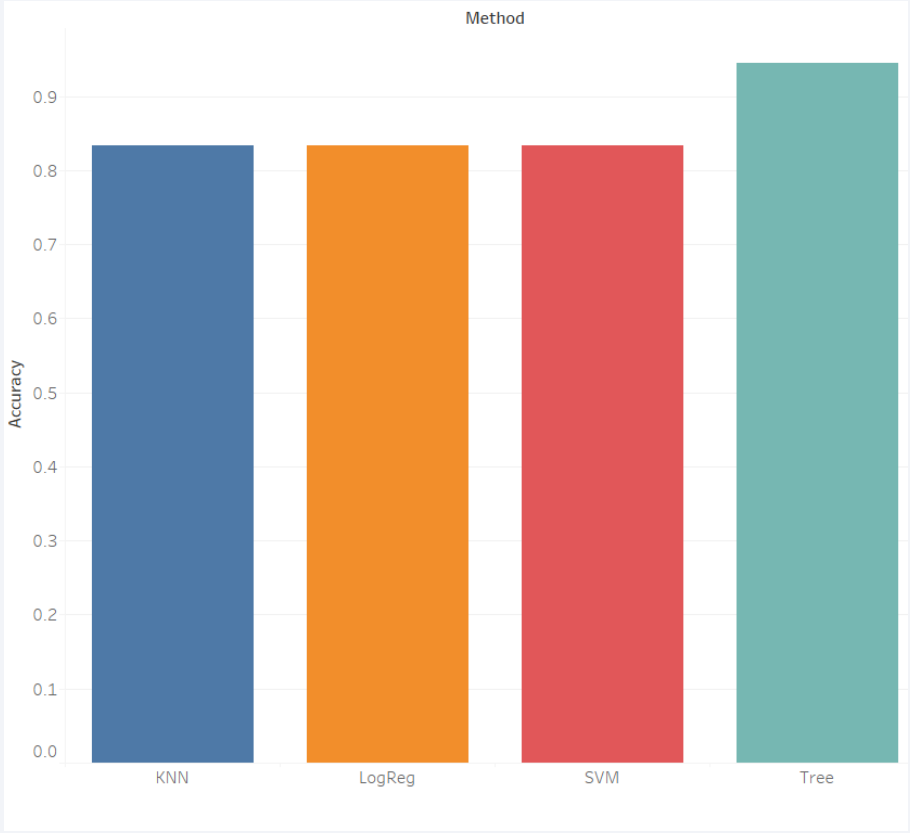


Section 5

Predictive Analysis (Classification)

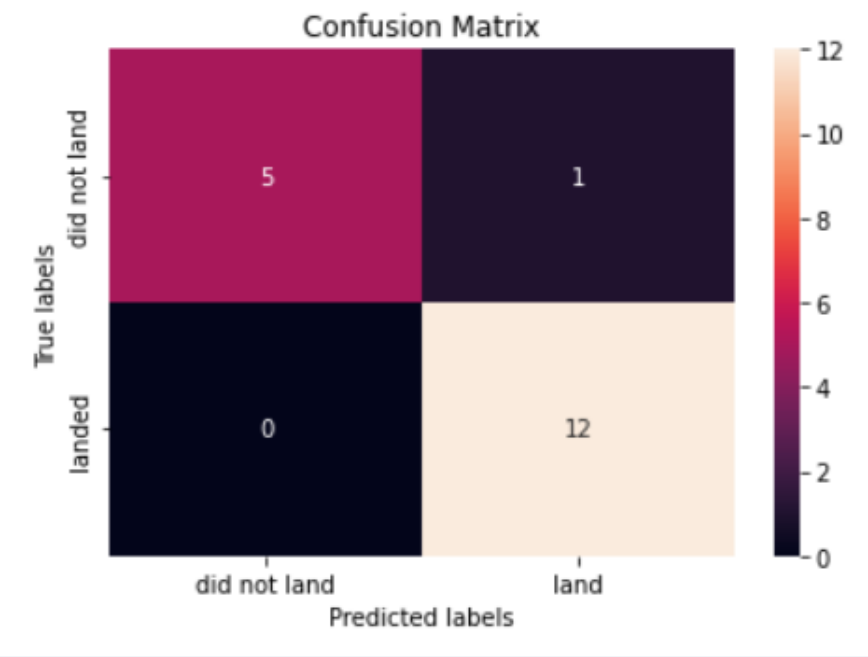
Classification Accuracy

- The decision tree gives the highest out of sample accuracy of all the methods.



Confusion Matrix

- The decision tree classification model gave an out-of-sample accuracy rate of 94%. The model only gave one false positive prediction as seen in this confusion matrix



Conclusions

- It is possible to use data available online to predict which booster jets SpaceX will recover.
- The highest out of sample prediction accuracy is produced by a decision tree prediction method.
- Important features that predict a successful recovery is the type of orbit the rocket is sent in and the payload of the rocket.
- SpaceX's recovery rate increased overtime.

Appendix

- Github repository can be found at https://github.com/parksandrecs/solid-dollop.git

Thank you!