Winning Space Race with Data Science

Basil

13.11.2021

Outline

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

Executive Summary

- Data collection, wrangling, then do an EDA, visualize the data.
- Data analysis showed certain factors that depend greatly on the success of stage 1.
- Use this features to create a predictive model.

Introduction

- SpaceX has reduced cost of space launches by reusing second stage of the rocket.
- If we can predict what factors affect the success, then we can take care of that and save money.
- For doing so, we need data.
- Also, we need to do data mining and create a predictive model.

Section 1

Methodology

Methodology

Executive Summary

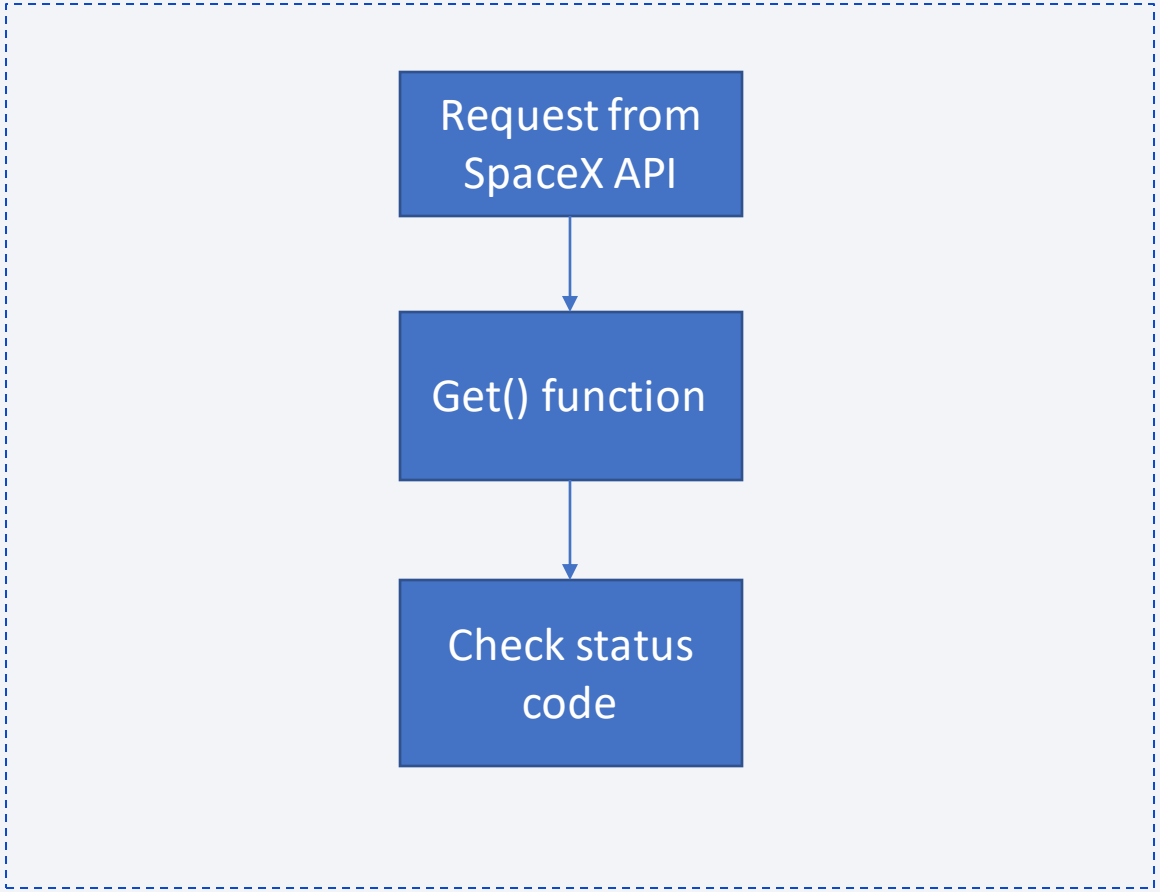
- Data collection methodology:
  - Collected from SpaceX API
- Perform data wrangling
  - Used pandas to clean, filter and evaluate quality of data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used sklearn to create models, tune hyperparameters and then train the models and evaluate their accuracy.

Data Collection

This part gives insights into how data is collected and handled.

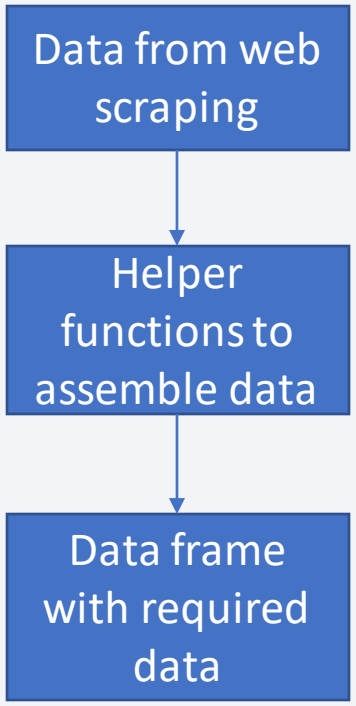
Data Collection – SpaceX API

- Used requests library.
- Use get() function from requests to request and parse all the data.
- Check for status code for successful request.
- Git url: https://github.com/neo0311/DS_capstone_spacex/blob/main/Data%20Collection%20API.ipynb



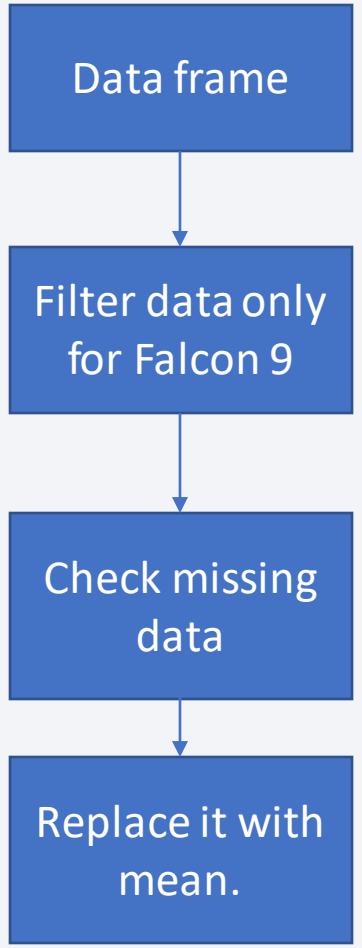
Data Collection - Scraping

- Obtained required data from multiple URLs.
- Used helper functions to assemble data frame.
- Store it to a data frame data.
- Git url: https://github.com/neo0311/DS_capstone_spacex/blob/main/Data%20Collection%20API.ipynb



Data Wrangling

- Collected data stored into a data frame.
- Kept only Falcon 9 data
- Check for missing values using pd.isnull() method.
- Replaced missing values with mean.
- Git url: https://github.com/neo0311/DS_capstone_spacex/blob/main/Data%20Collection%20API.ipynb



EDA with Data Visualization

- Used various charts to explore the data and their correlations.
- Plotted charts:
  - Scatter-Flight No. vs Payload - as payload is important feature.
  - Scatter-Flight No. vs Launch Site – check for site relevancy.
  - Scatter-Payload vs Launch Site – helps to get insights about payload dependence on site.
  - Bar-Orbit vs Success Rate – Orbit is a very relevant feature for launches.
  - Scatter – Flight No. vs Orbit – helps to find correlation between flight number and orbit types.
  - Line – Year vs Success rate – checks the progress along years.
- Git url: https://github.com/neo0311/DS_capstone_spacex/blob/main/EDA%20with%20Data%20Visualisation.ipynb

EDA with SQL

- Used SQL queries to understand dataset and look for relevant insights.
- Used queries:
  - Unique launch sites
  - Launch sites beginning with 'CCA'
  - Total payload carried by NASA
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved.
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass – using sub query.
  - Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Count of landing outcomes ranked descending order.
- Git link: https://github.com/neo0311/DS_capstone_spacex/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

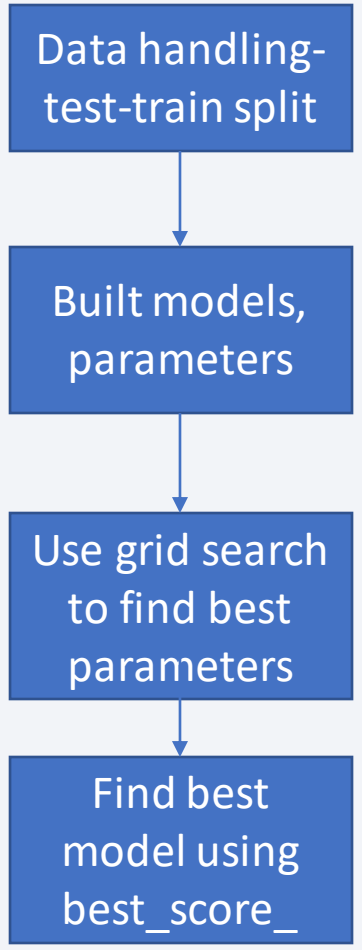
- Using markers(circle) to mark launch sites on the map – check proximity to equator or coast.
- Used marker clusters to mark success or fail attempts at each site.
- Used lines to check distance of launch sites from various land features.
- All this enabled us to get insights on various correlations.
- Git link: https://github.com/neo0311/DS_capstone_spacex/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Used filtered data to render various dashboard elements to get more insights.
- Elements added:
  - Drop down menu – various launch sites.
  - Pie chart to depict success and failure
  - Range slider – to check a specific range of payload.
  - Scatter chart – check correlations between payload and success rate.
- All this helped to get a better overview of the data easily.
- Git link: https://github.com/neo0311/DS_capstone_spacex/blob/main/spacex_launch_dash.py

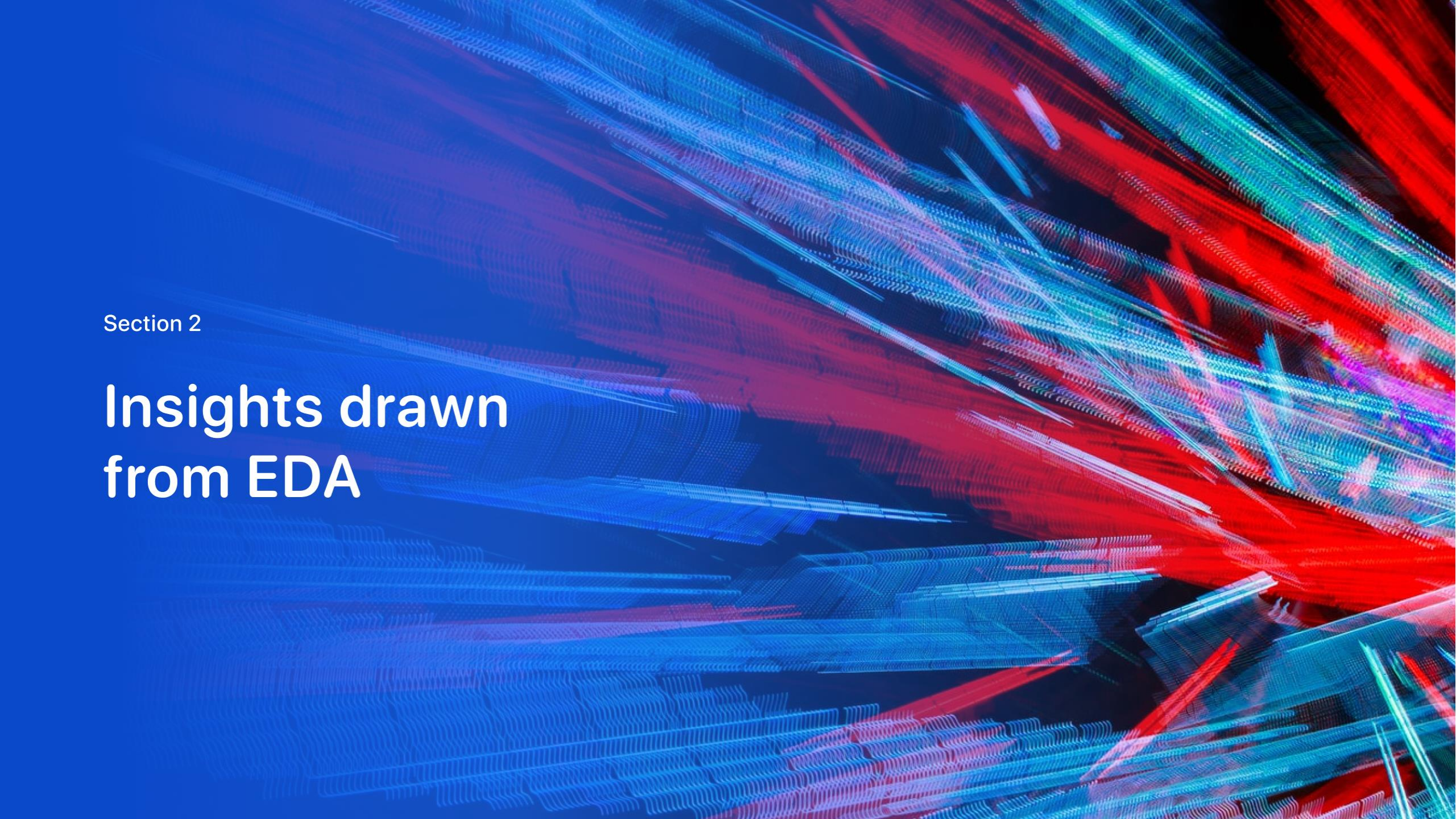
Predictive Analysis (Classification)

- Built various models.
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- Found best model using grid search on a set of parameters
- Used GridSearchCV method.
- Best score is acquired using .best_score_ method
- Git link: https://github.com/neo0311/DS_capstone_spacex/blob/main/Machine%20Learning%20Prediction.ipynb



Results

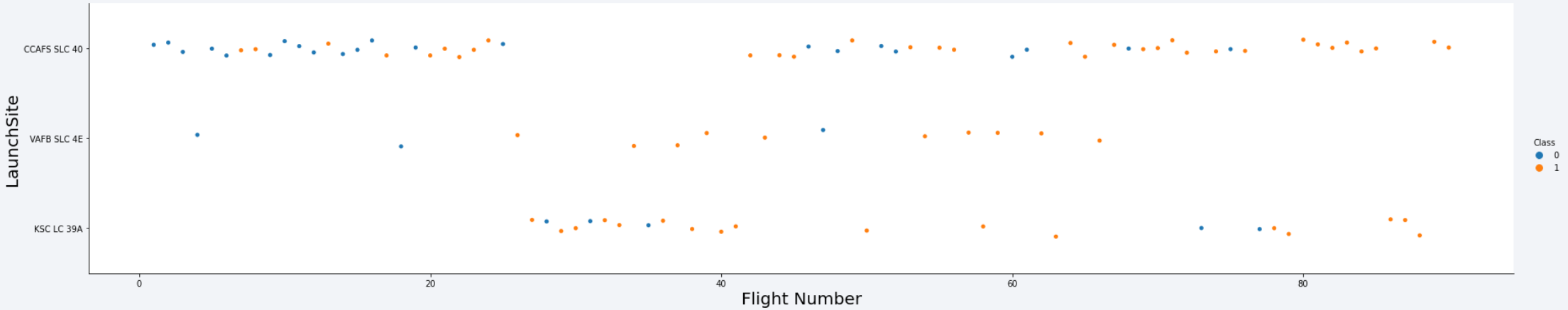
- This section depicts the results using screenshots and other data from the analysis.



Section 2

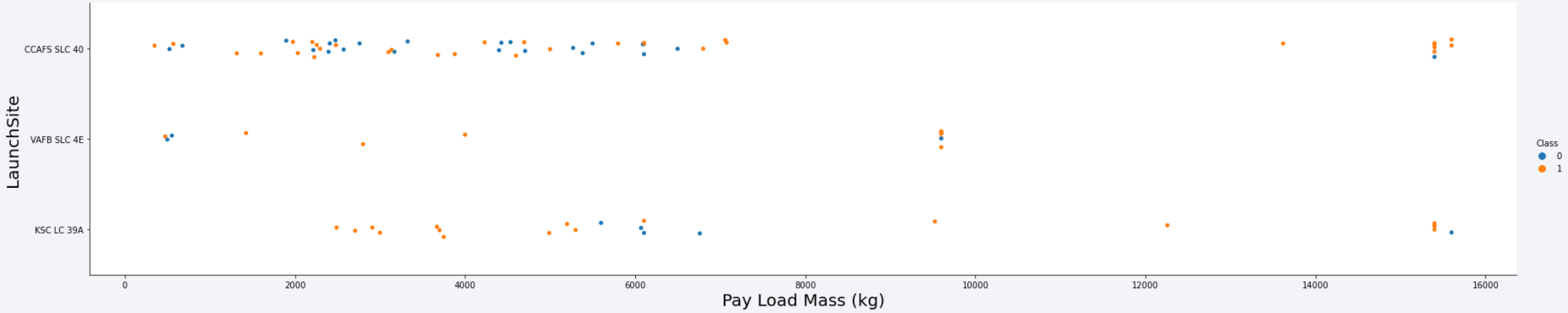
Insights drawn from EDA

Flight Number vs. Launch Site



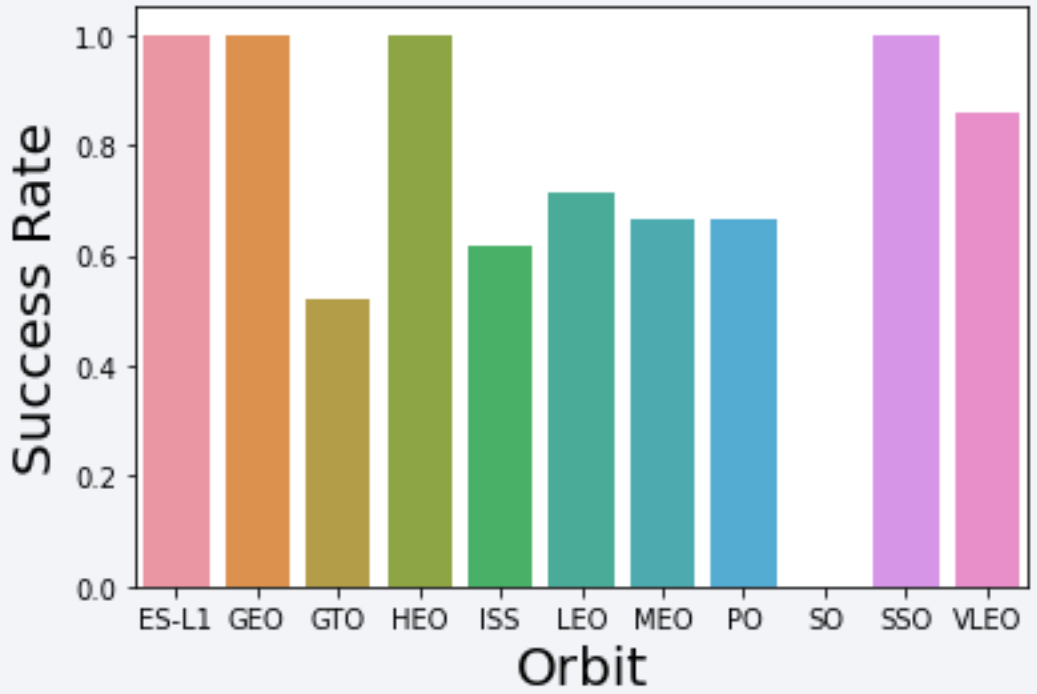
- Scatter plot of Flight Number vs. Launch Site.
- As flight number increases the success rate at CCAF5 SLC 40 site increases very much.
- Less flights done at VAFB SLC 4E towards the end.

Payload vs. Launch Site



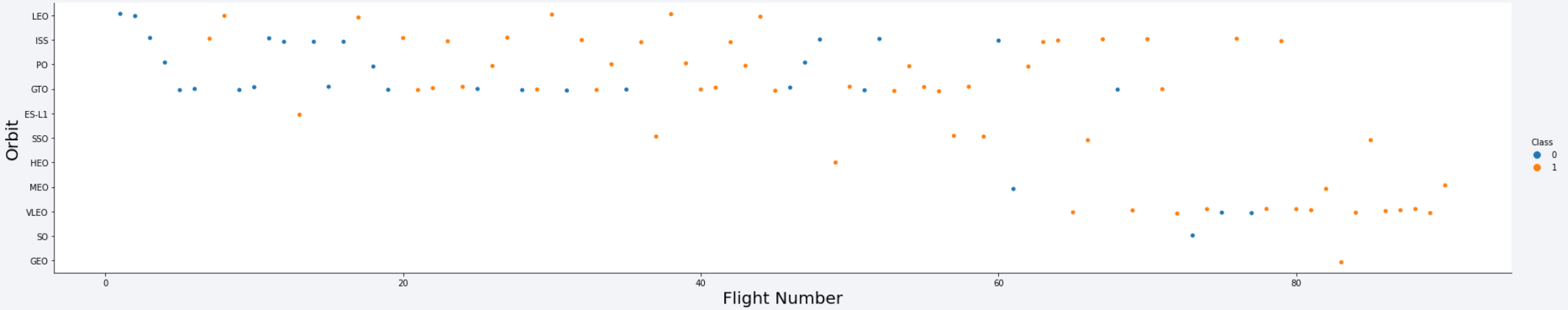
- Scatter plot of Payload vs. Launch Site.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass

Success Rate vs. Orbit Type



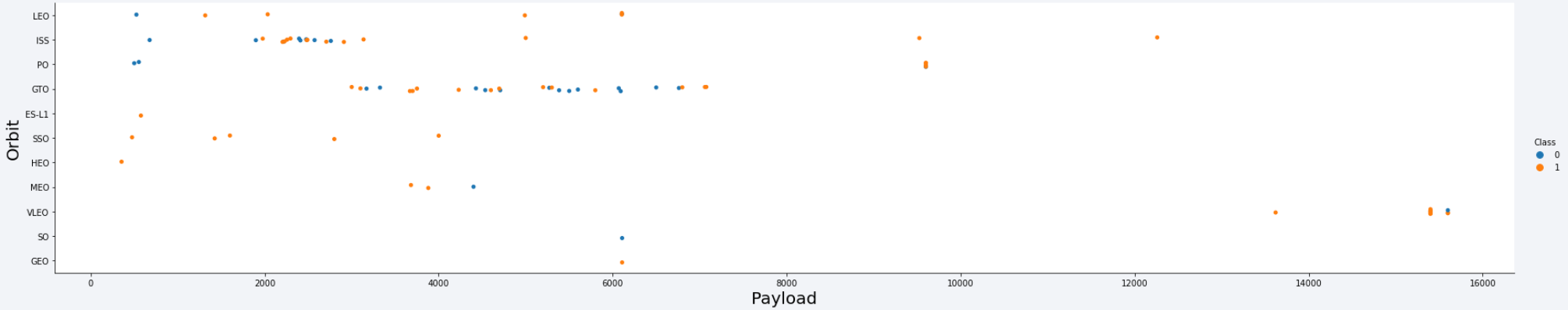
- Bar chart for the success rate of each orbit type
- Launches to SSO, HEO, GEO and ES-L1 shows high success rates.

Flight Number vs. Orbit Type



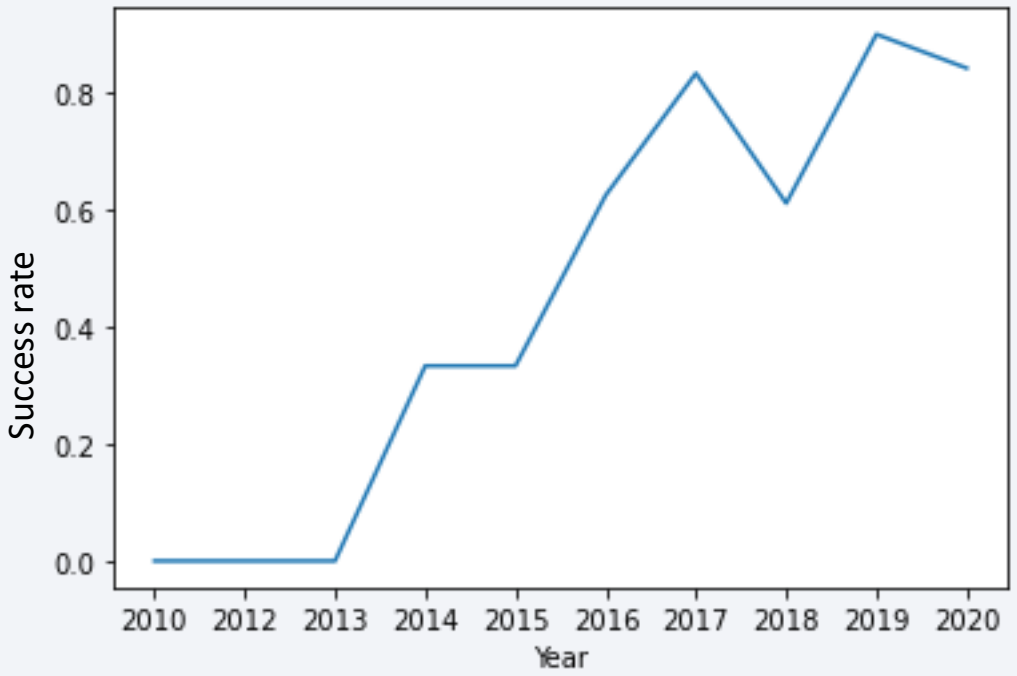
- Scatter point of Flight number vs. Orbit type
- In the LEO orbit the Success appears related to the number of flights.
- No relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- Scatter point of payload vs. orbit type
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing are both there here.

Launch Success Yearly Trend



- Line chart of yearly average success rate
- Success rate keeps on improving over the years.
- Huge improvement after 2013.

All Launch Site Names

- Find the names of the unique launch sites
- %sql select distinct(launch site) from SPACEXTBL

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

Launch Site Names Begin with 'CCA'

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

```
%%sql
select * from SPACEXTBL
where launch_site like 'CCA%'
limit 5
```

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Query:

```
%%sql
select SUM(payload_mass__kg_) from SPACEXTBL
where customer like 'NASA%'
```

1
99980

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Query:

```
%%sql
select AVG(payload_mass__kg_) from
(select * from SPACEXTBL where booster_version like '%v1.1%')
```

1
2534

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Query:

```
%%sql
select MIN(DATE) from (select * from SPACEXTBL
where "Landing _Outcome" like '%Success (ground pad)')
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query:

```
%%sql
select booster_version from SPACEXTBL
where "Landing _Outcome" like '%Success (ground pad)' and payload_mass__kg_ between 4000 and 6000
```

booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Query:

```
%%sql

select * from SPACEXTBL
group by "Landing _Outcome" like '%Success%')
```

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass.
- Query:

```
%%sql
select booster_version, payload_mass__kg_ from SPACEXTBL
where payload_mass__kg_ = (select MAX(payload_mass__kg_) from SPACEXTBL)
```

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

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

```
%%sql
select DATE, booster_version, launch_site, "Landing _Outcome" from SPACEXTBL
where "Landing _Outcome" like 'Failure (drone ship)' and DATE like '2015%'
```

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

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Query:

```
%%sql
select count("Landing _Outcome"), "Landing _Outcome" from SPACEXTBL
group by "Landing _Outcome"
order by count("Landing _Outcome") desc
```

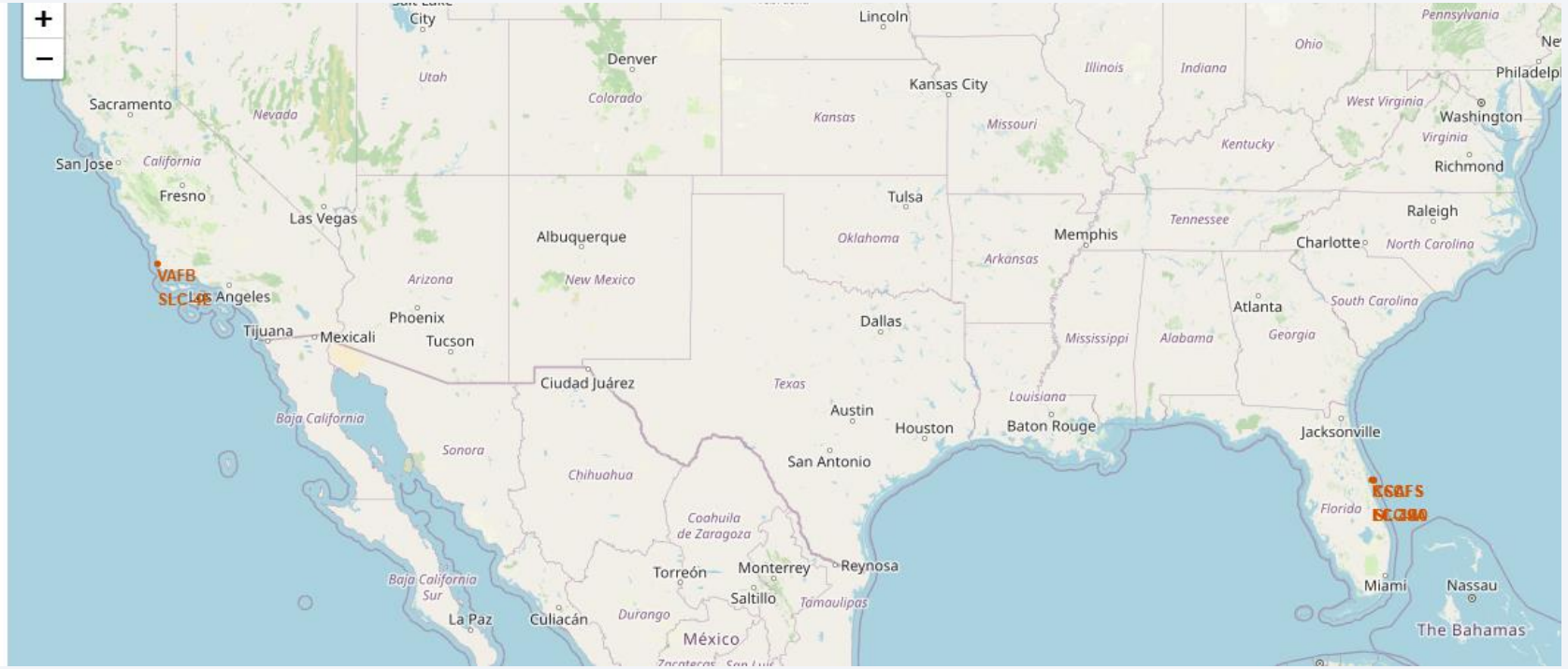
1	Landing _Outcome
38	Success
22	No attempt
14	Success (drone ship)
9	Success (ground pad)
5	Controlled (ocean)
5	Failure (drone ship)
3	Failure
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 4

Launch Sites Proximities Analysis

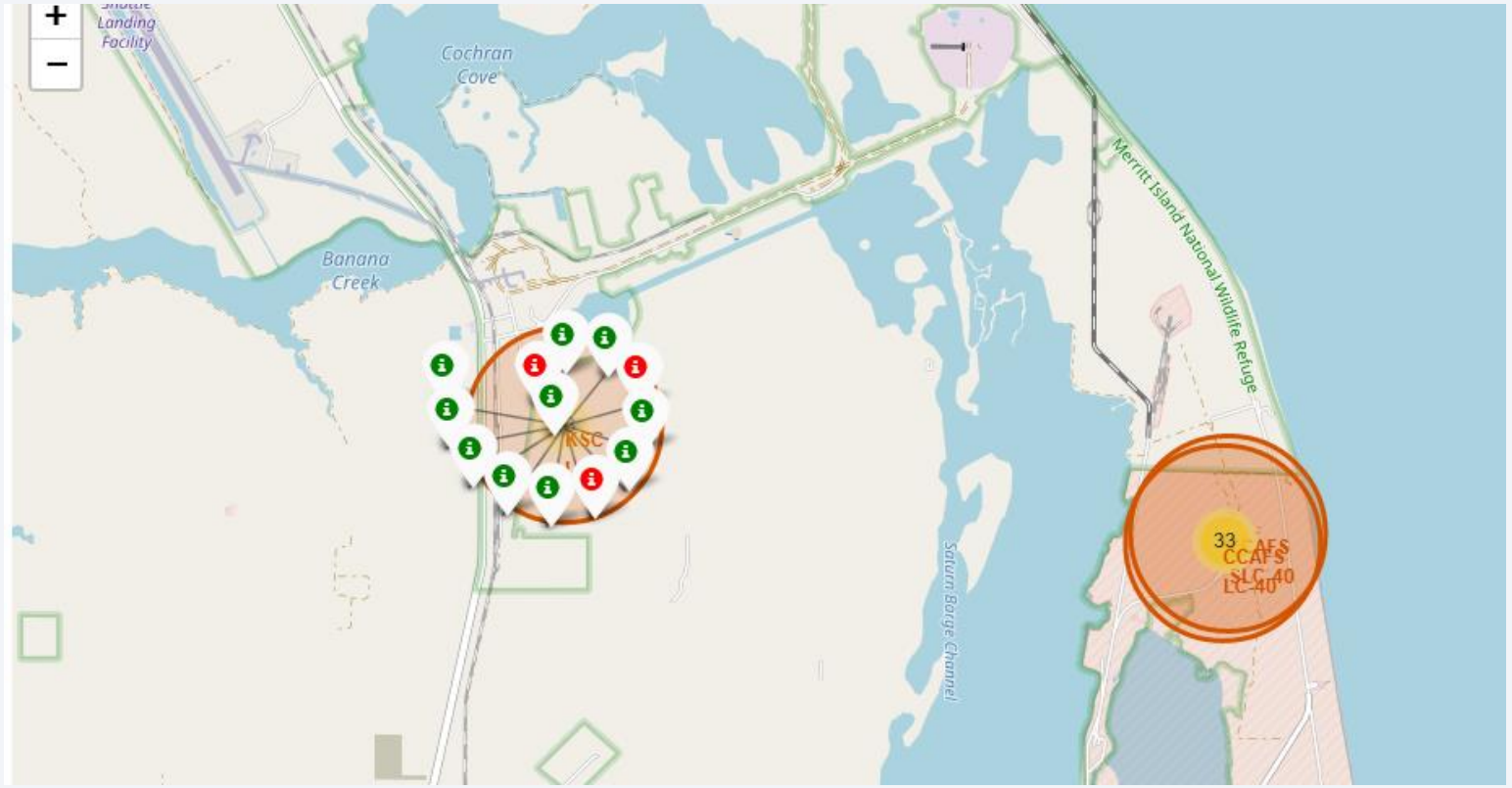
Launch Sites

- Map depicting all the launch sites are depicted.
- All sites are very close to coast.



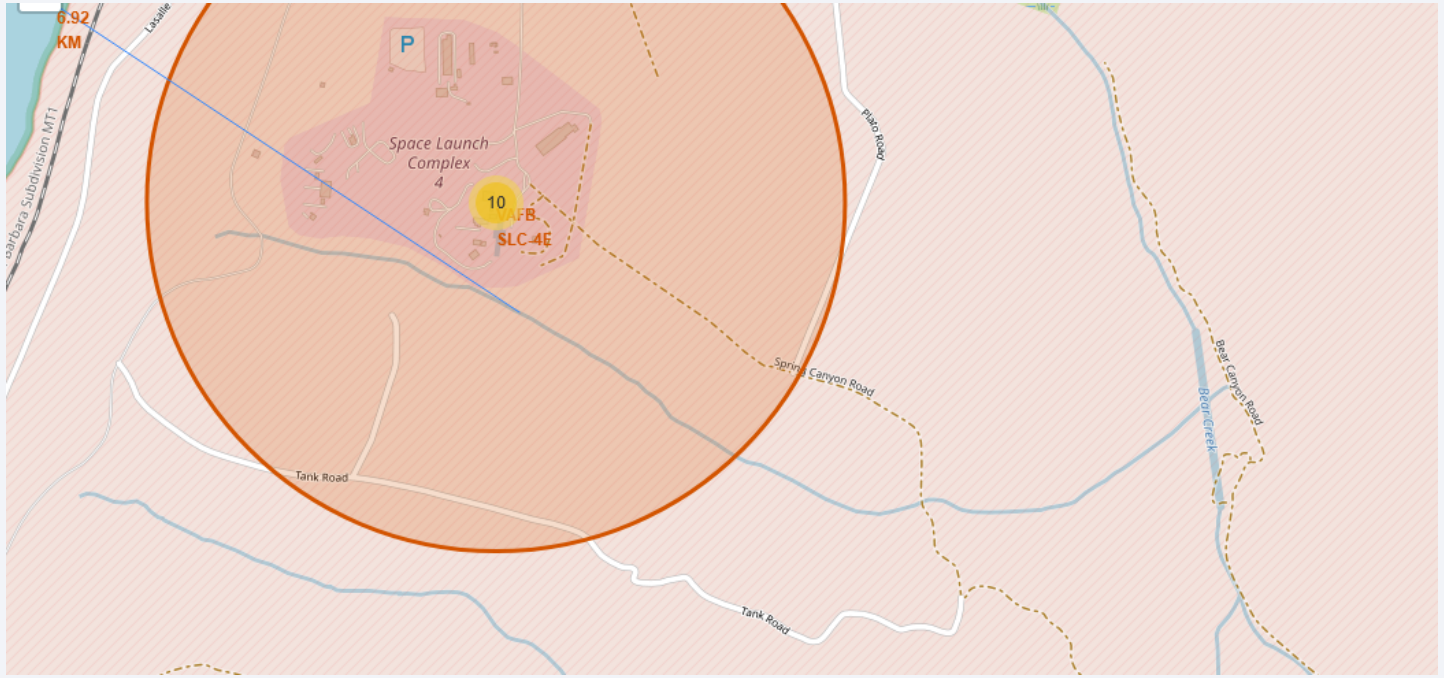
Launch outcomes at each site.

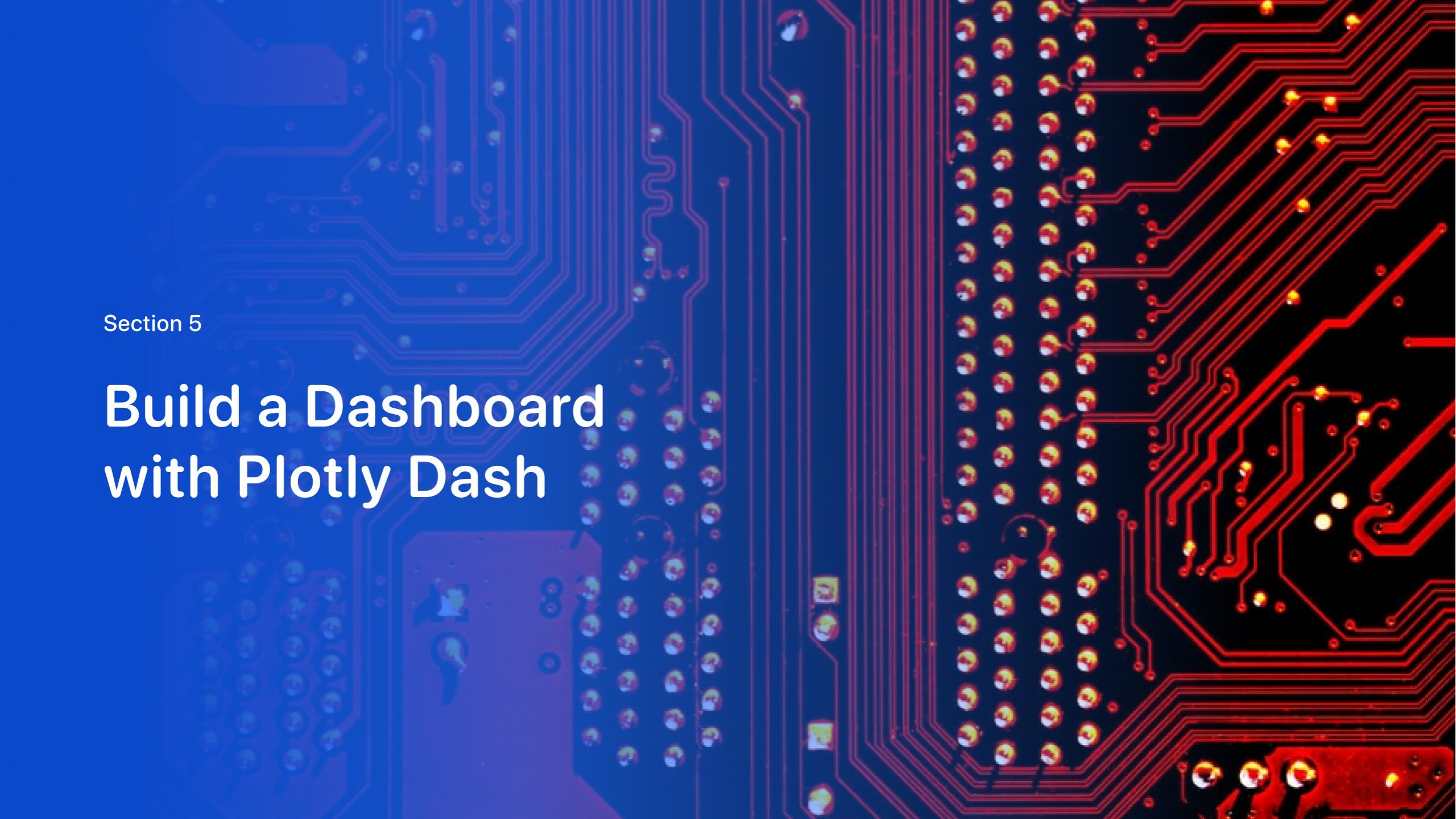
- Launch outcomes at each site are color labeled and depicted as shown.



Distance to land features

- Distance of launch sites to different land features like coastline, railways, etc. are checked.
- This helps to find the most favorable launch site for reducing coast.



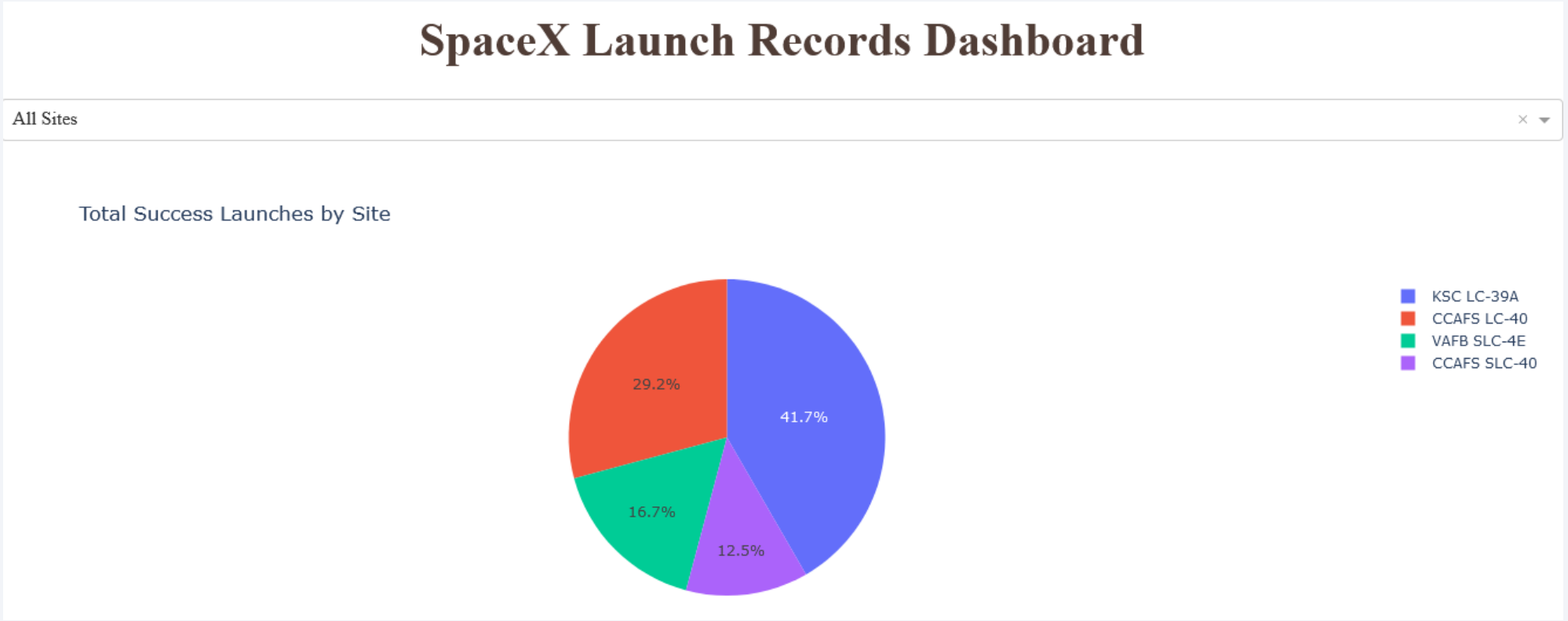


Section 5

Build a Dashboard with Plotly Dash

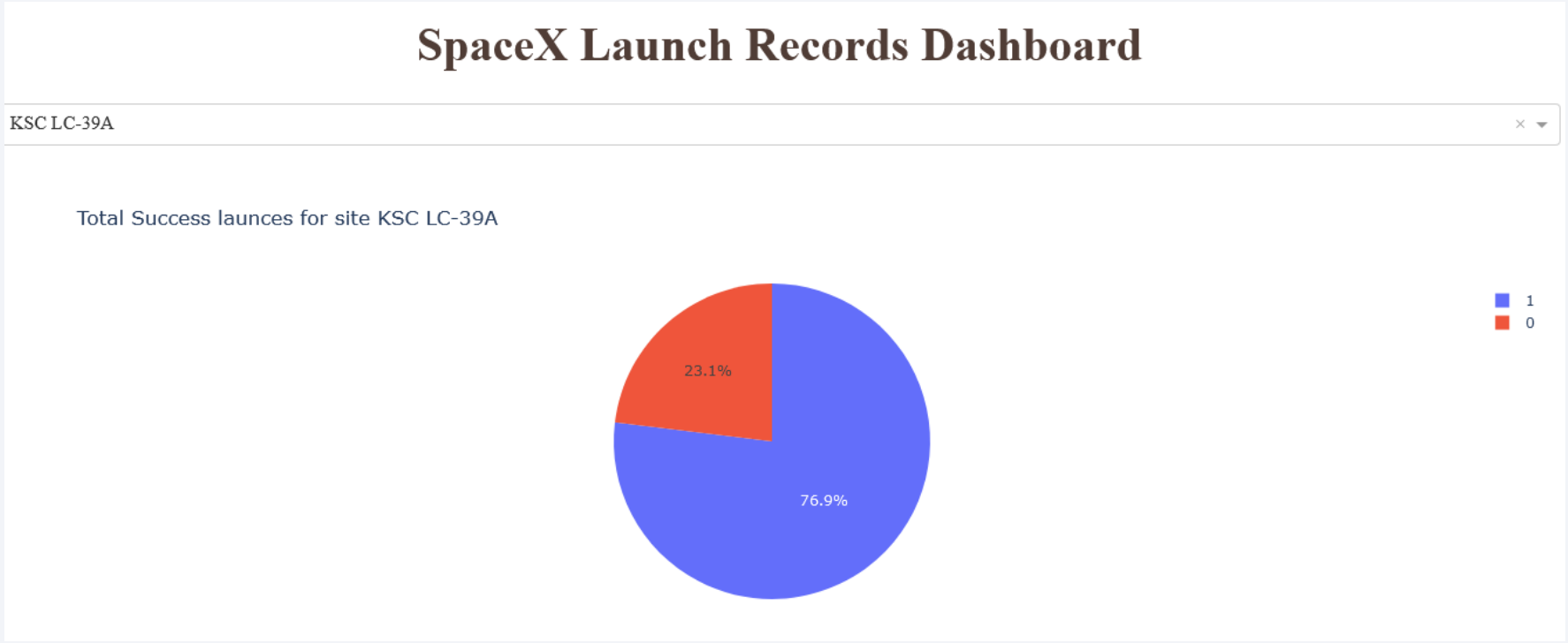
Launch success count for all sites

- The KSC LC-39A site has the most success rate followed by CCAFS LC-40



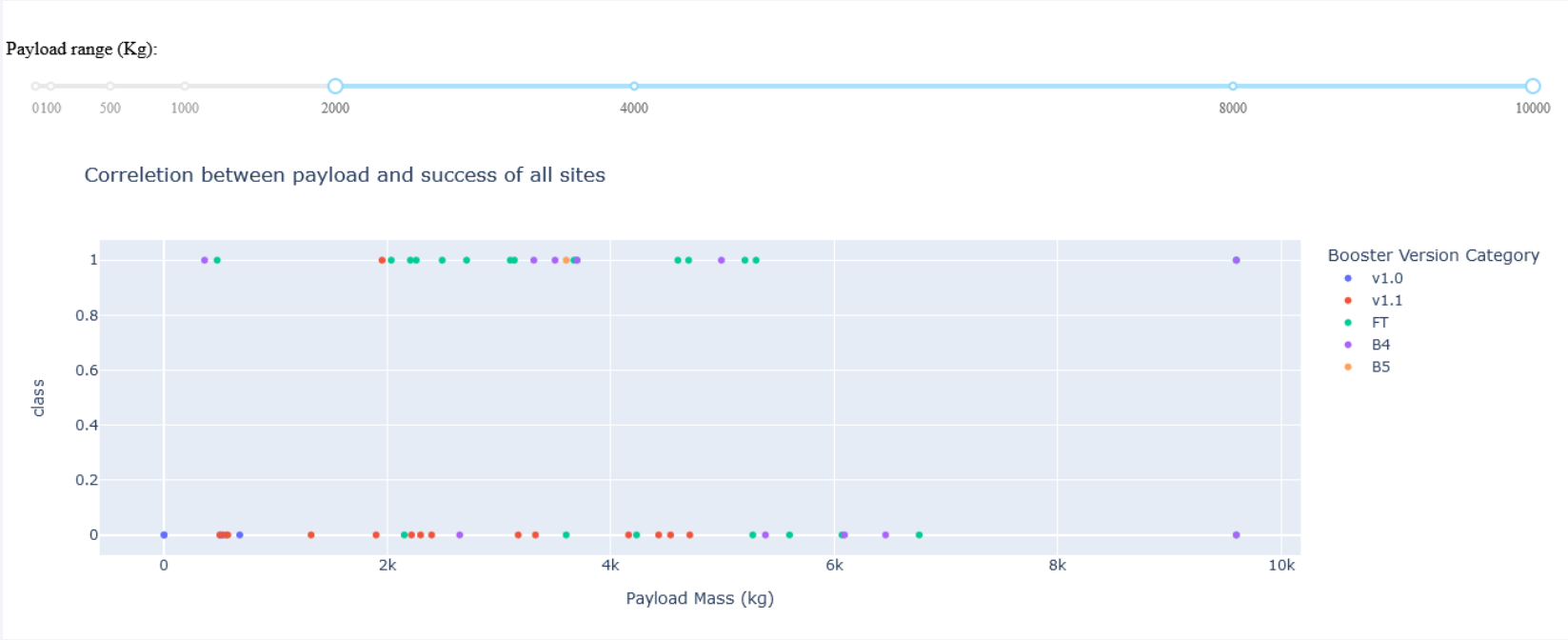
Site with highest success rate

- Figure shows the KSC LC-39A site



Payload vs. Launch Outcome

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider.



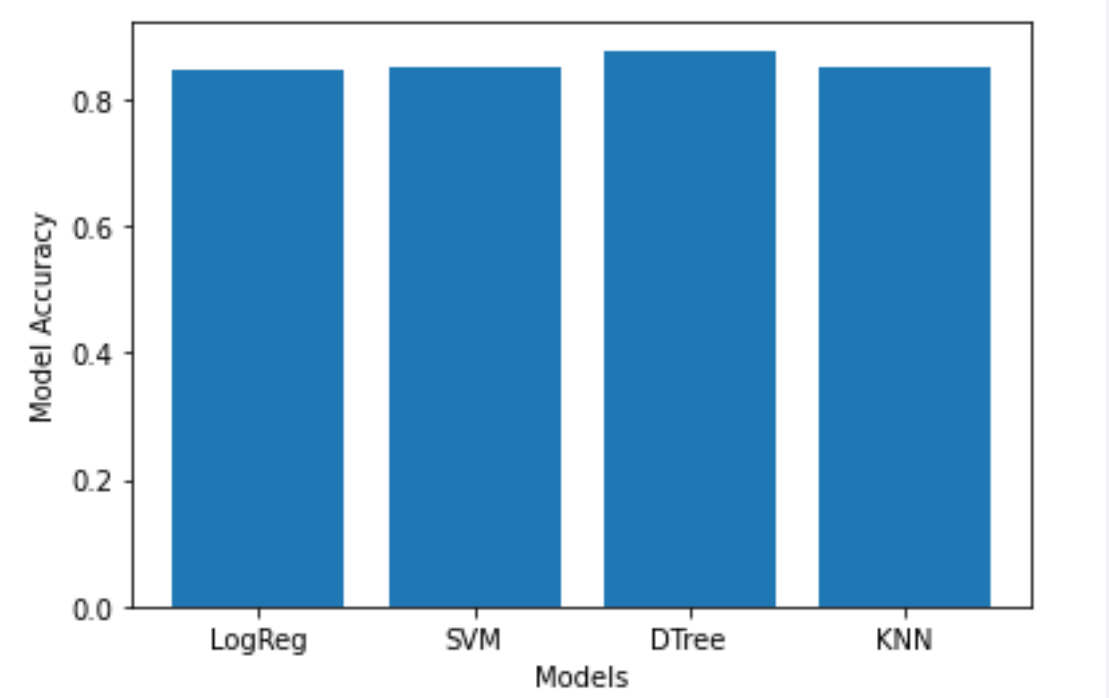
- For high payloads success rate falls.

Section 6

Predictive Analysis (Classification)

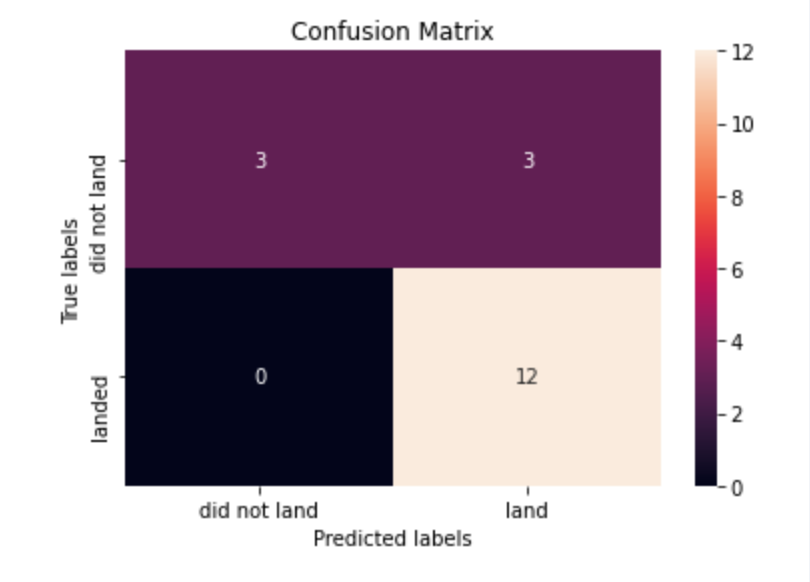
Classification Accuracy

- Models used:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- Decision Tree model has the highest classification accuracy



Confusion Matrix

- Confusion matrix of Decision Tree algorithm.
- Although there are false positives, they are very less compared to the correct predictions.



Conclusions

- Several factors affect the success rate.
- This includes, orbit types, flight number, payload mass, etc.
- By visualizing data and using EDA we were able to find the right features.
- This is evident from the high model accuracy scores.
- Next, we need to test it on upcoming launches.

Appendix

- The figure shows part of the final data set we used for training the model.

	FlightNumber	PayloadMass	Flights	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	...	Serial_B1058	Serial_B1059	Serial_B1060	Serial_B1062	GridFins_Fa
0	1.0	6104.959412	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
1	2.0	525.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
2	3.0	677.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	
3	4.0	500.000000	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
4	5.0	3170.000000	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
85	86.0	15400.000000	2.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	
86	87.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	
87	88.0	15400.000000	6.0	5.0	5.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
88	89.0	15400.000000	3.0	5.0	2.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	
89	90.0	3681.000000	1.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	

90 rows × 83 columns

Thank you!