Winning Space Race with Data Science

Mirco Höhne
30/12/2021

Outline

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

Executive Summary

- The Data was collected using the public SpaceX API and Web Scraping the Wikipedia Page of SpaceX
- The Data was transformed in order to use it efficiently (Created new labels and encoded categorical labels)
- statistics, visualization, SQL, folium maps and dashboards were used to explore the data and find relevant columns to be used as features for the machine learning models
- For machine learning models Logistic Regression, Support Vector Machines, Decision Trees and K-Nearest-Neighbors were used. They all performed similarly with an accuracy of 83.33%. For a better accuracy of the model more data would be needed

Introduction

- Space Flight is a very costly endeavor
- SpaceX successfully cut the cost of its rockets (62 Million USD in comparison to 165 Million USD that other providers charge)
- Much of the savings come from reusing the first stage of the rocket
- can we use data to predict if the first stage of a rocket lands successfully to use this data to create a company that can compete with SpaceX prices?

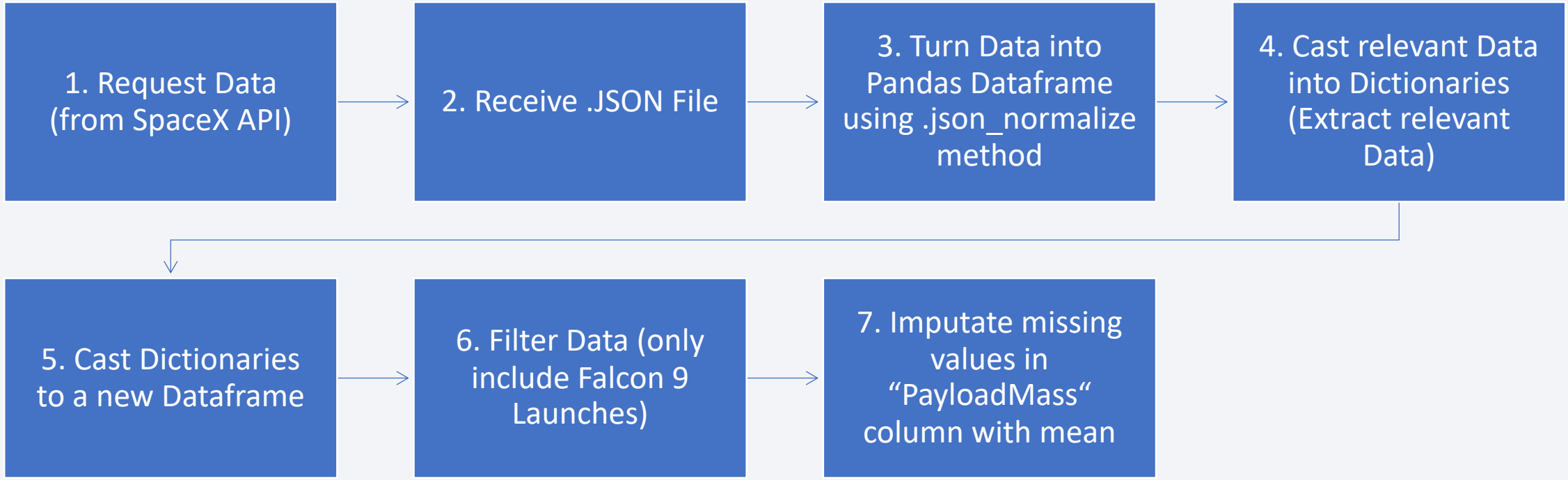
Section 1

Methodology

Data Collection

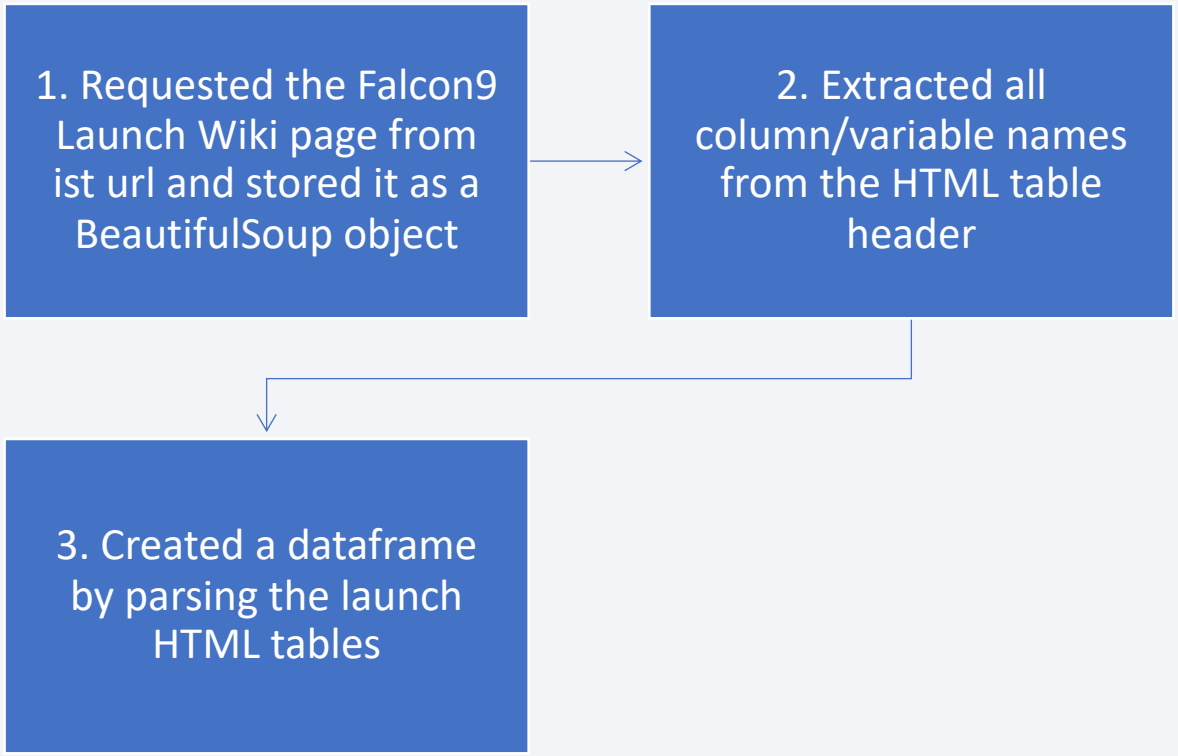
- Two approaches to Data Collection were used
  - SpaceX API
  - Web Scraping the Wikipedia Page of SpaceX

Data Collection – SpaceX API



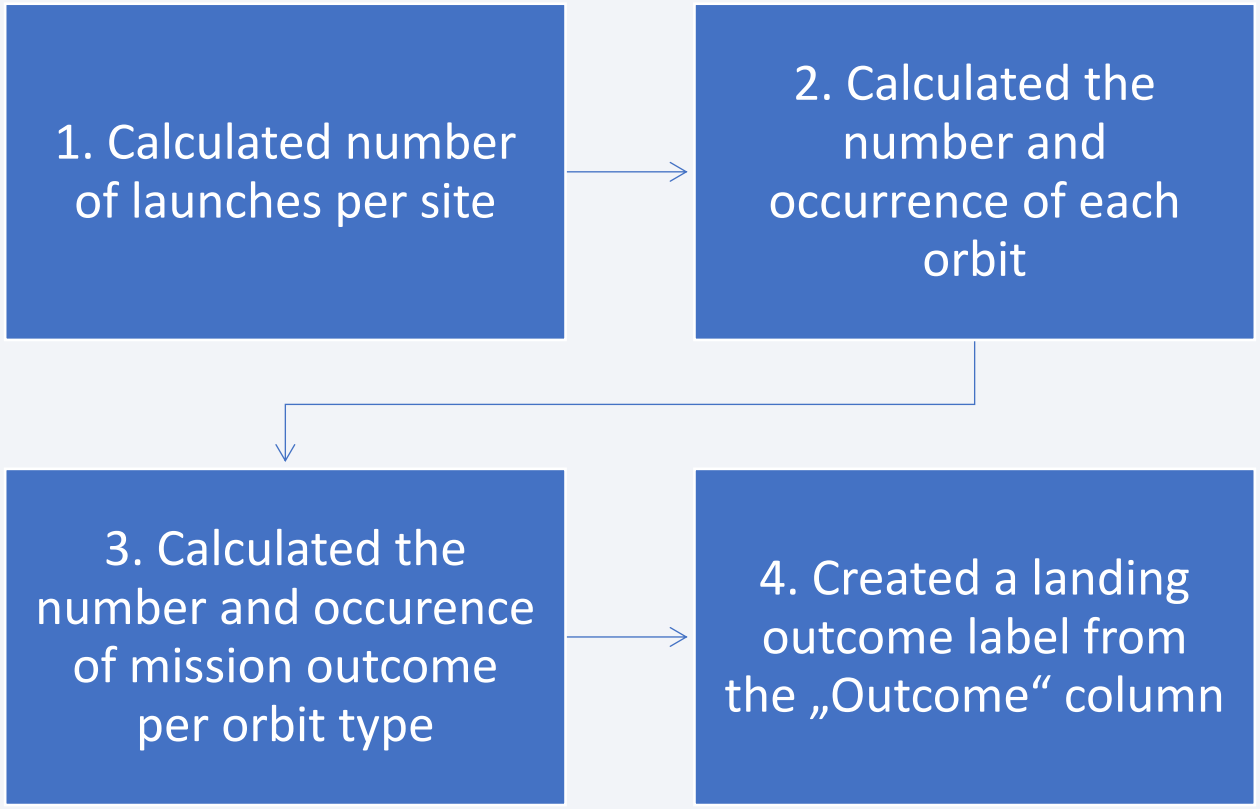
https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone/blob/master/1.%20Data_Collection_with_API.ipynb

Data Collection - Scraping



https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone/blob/master/2.%20Data_Collection_with_Webscraping.ipynb

Data Wrangling



https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone/blob/master/3.%20Data_wrangling.ipynb

EDA with Data Visualization

- For the visualization Bar Charts were used, when showing two-dimensional Data, Scatterplots for three-dimensional data and line charts for time series data in order to visualize the relationship between the variables
- The Plots where:
  - Relationship between Launch Site, Flight Number and Class
  - Relationship between Payload Mass, Launch Site and Class
  - Relationship between Success Rate and Orbit Type

https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone/blob/master/5.%20eda_dataviz_python.ipynb

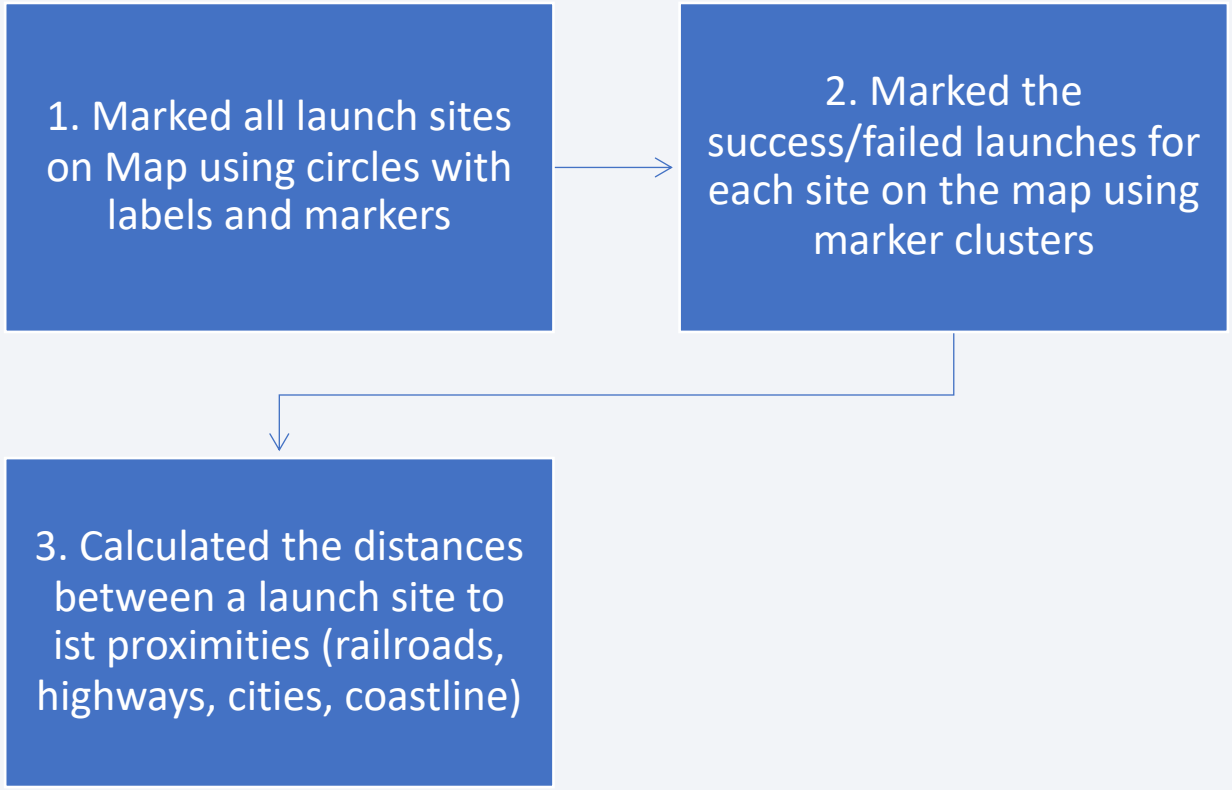
EDA with SQL

- Performed SQL queries:
  - names of unique launch sites in the space mission
  - five records where launch sites begin with the string 'CCA'
  - total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - date when first successful landing outcome in ground pad was achieved
  - list names of bossters which have success in drone ship and have payload mass between 4000 and 6000
  - total number of successful and failure mission outcomes
  - names of bosster_versions which have carried the maximum payload mass
  - List failed landing_outcomes in drone ship, their booster versions, and laucnch site names in year 2015
  - Rank count of landing outcomes between date 2010-06-04 and 2017-03-20, in descending order

https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone/blob/master/4.%20eda-sql.ipynb

Build an Interactive Map with Folium

- Different Points of interest were marked in order to gain new knowledge from the visualizations



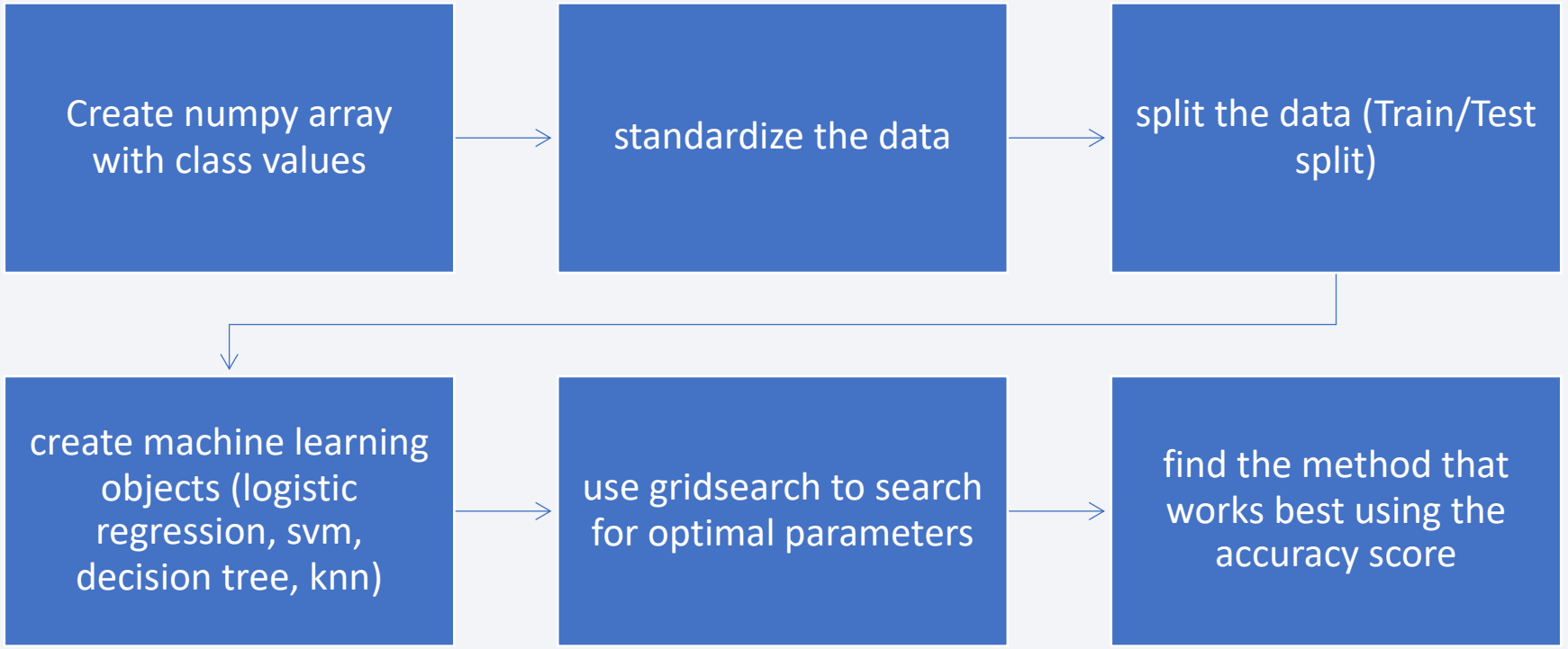
https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone/blob/master/6.%20launch_site_geo_visualization.ipynb

Build a Dashboard with Plotly Dash

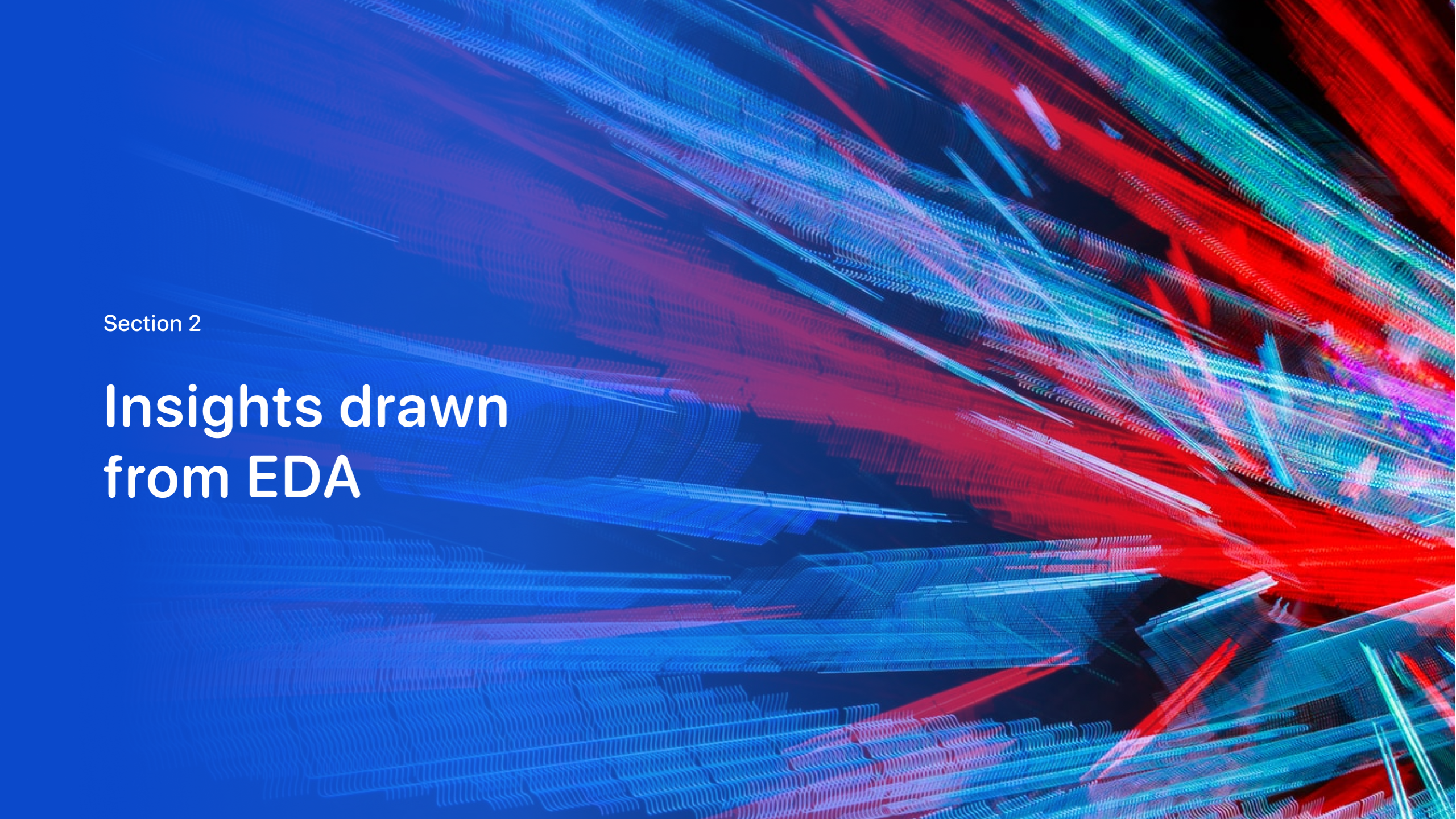
- Dashboard consists of:
  - Pie Chart for successful launches (all sites in one pie chart and a pie chart for every site)
  - Scatterplot showing the relationship between Payload Mass, class and Booster Version of all sites

https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone/blob/master/7.%20spacex_dash_app.py

Predictive Analysis (Classification)



https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone/blob/master/8.%20ml_prediction.ipynb

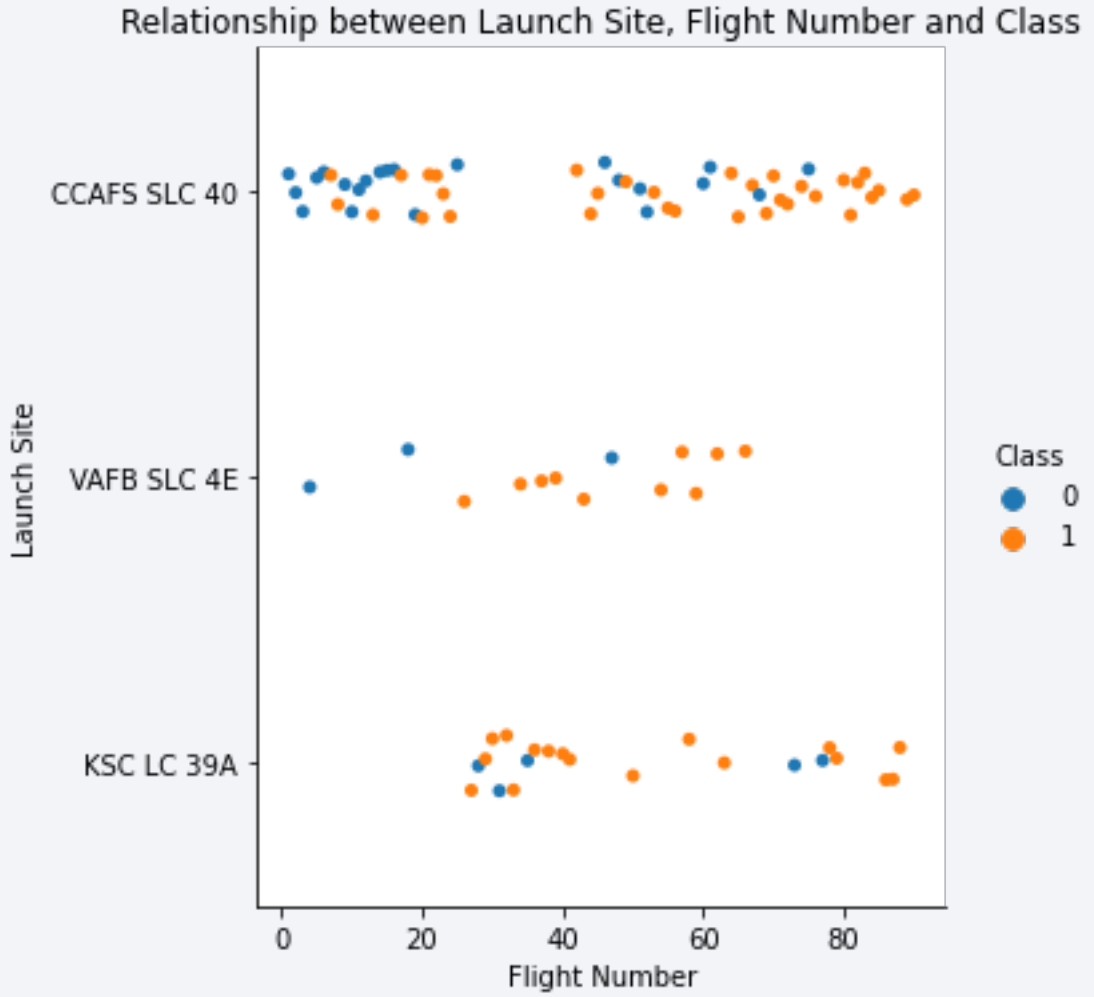


Section 2

Insights drawn from EDA

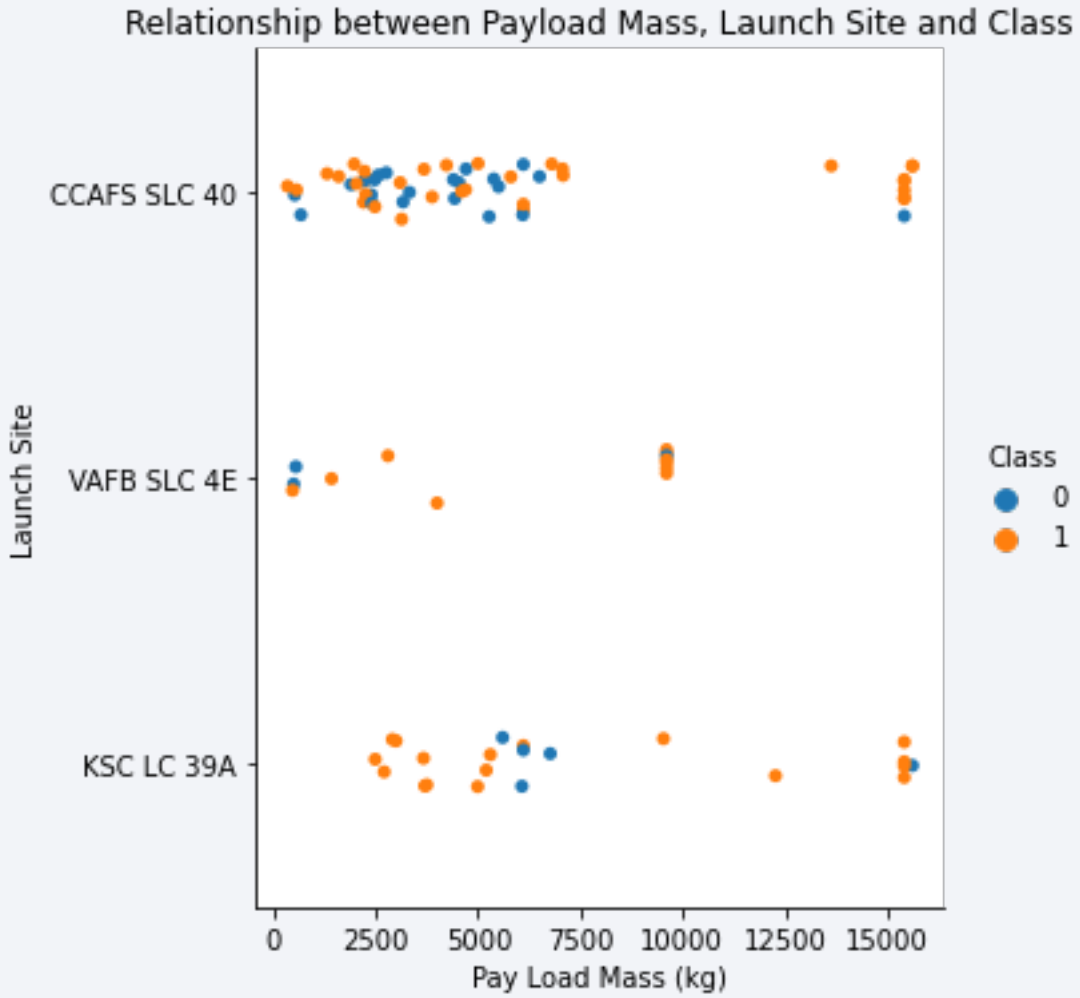
Flight Number vs. Launch Site

- *we can see, that the number of failed landings (of the first stage) decreases with higher flight numbers.*
- *This makes sense, because with more experience better systems may be implemented in the first stage*



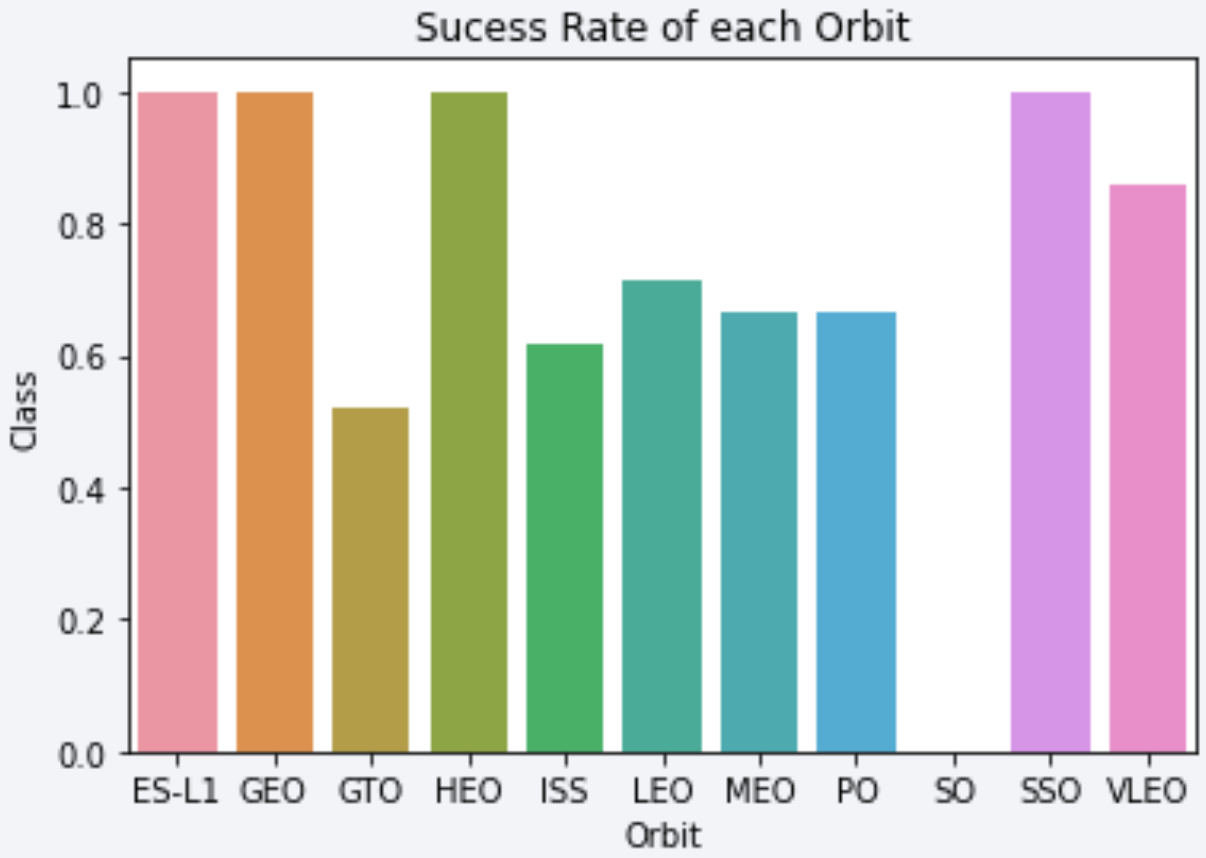
Payload vs. Launch Site

- with very high Payload Mass the success rate is also higher than with lower payload mass



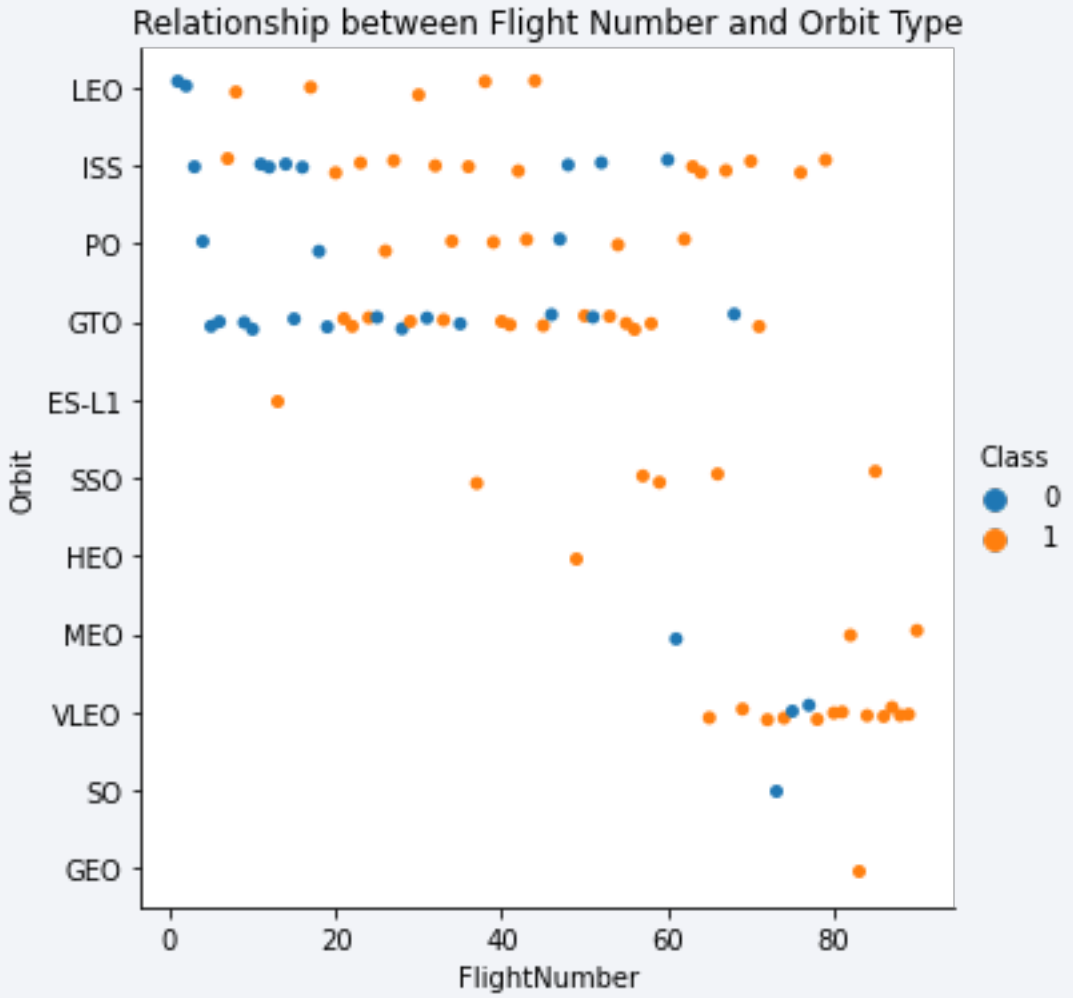
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have the highest success rates



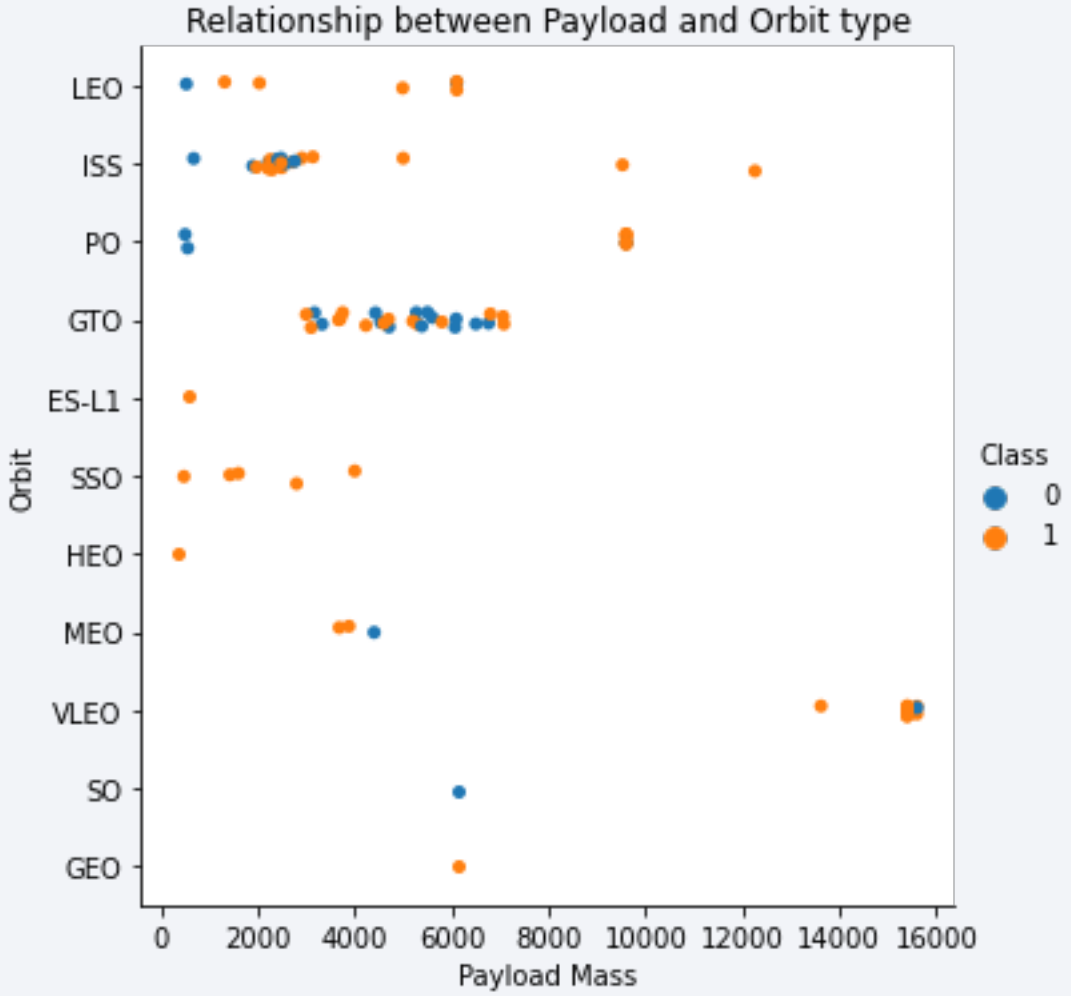
Flight Number vs. Orbit Type

- 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



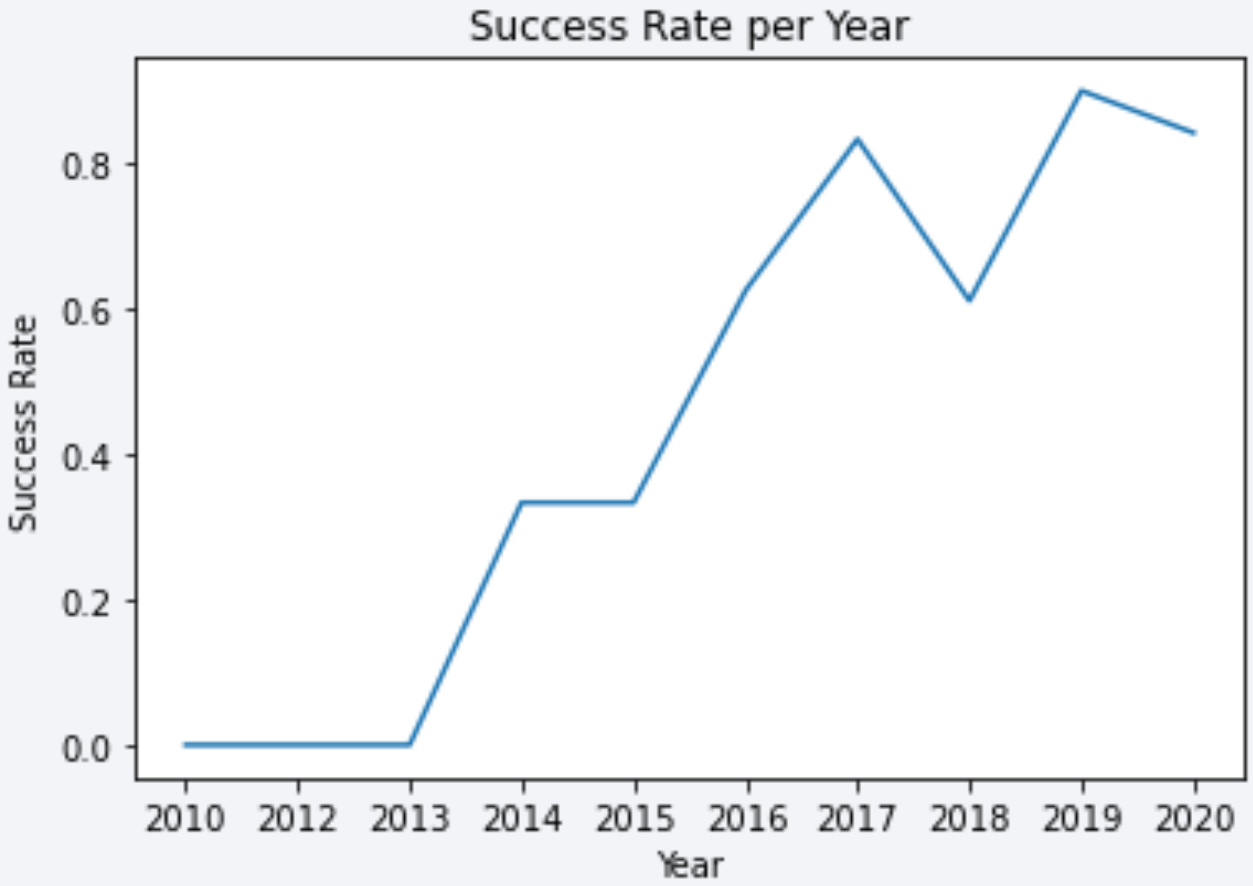
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- you can observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT(launch_site) AS names FROM SPACEXTBL;
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

names
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`

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

```
%sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

DATE	Time (UTC)	booster_version	launch_site	payload_mass_kg	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

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

```
%sql SELECT SUM(payload_mass__kg_) AS total_payload_mass__kg_ FROM SPACEXTBL WHERE customer='NASA (CRS)';
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

total_payload_mass__kg_
45596

Average Payload Mass by F9 v1.1

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

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

```
%sql SELECT AVG(payload_mass__kg_) AS avg_payload_mass__kg__F9 FROM SPACEXTBL WHERE booster_version= 'F9 v1.1';
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

avg_payload_mass__kg__f9
2928

First Successful Ground Landing Date

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

```
%sql SELECT MIN(DATE) AS first_succ_landing_ground FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

first_succ_landing_ground
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

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

```
%sql SELECT booster_version FROM SPACEXTBL WHERE ("Landing _Outcome"= 'Success (drone ship)') AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT (mission_outcome), COUNT(*) AS "count" from SPACEXTBL GROUP BY mission_outcome;
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

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

Boosters Carried Maximum Payload

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

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

```
%sql SELECT booster_version FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

booster_version
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

2015 Launch Records

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

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

```
%sql SELECT booster_version,launch_site, "Landing _Outcome", DATE FROM SPACEXTBL WHERE ("Landing _Outcome" = 'Failure (drone ship)') AND (I
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

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

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

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

```
%sql SELECT "Landing _Outcome", COUNT("Landing _Outcome") AS cnt FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY '
```

```
 * ibm_db_sa://fgv10761:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

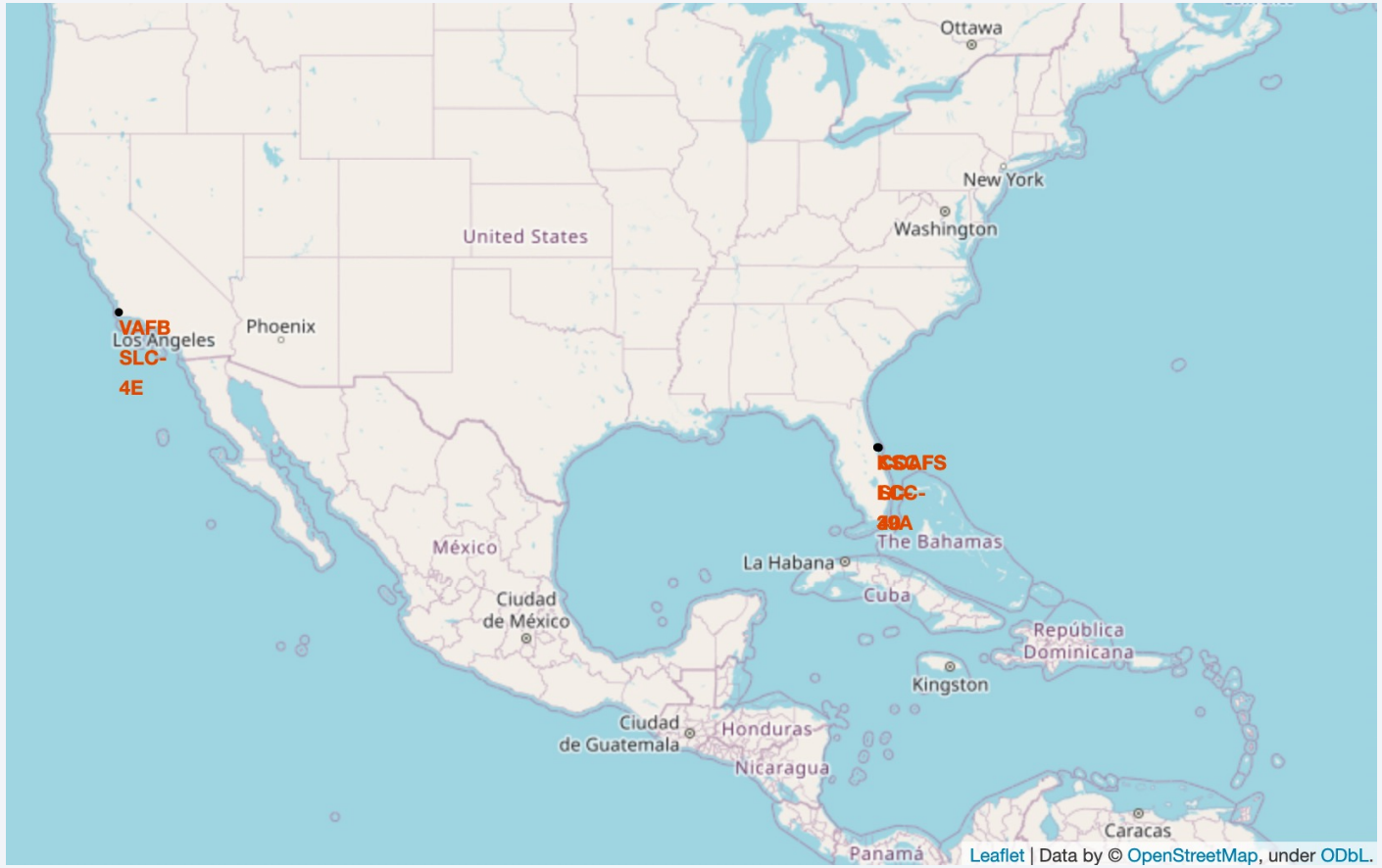
Landing _Outcome	cnt
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

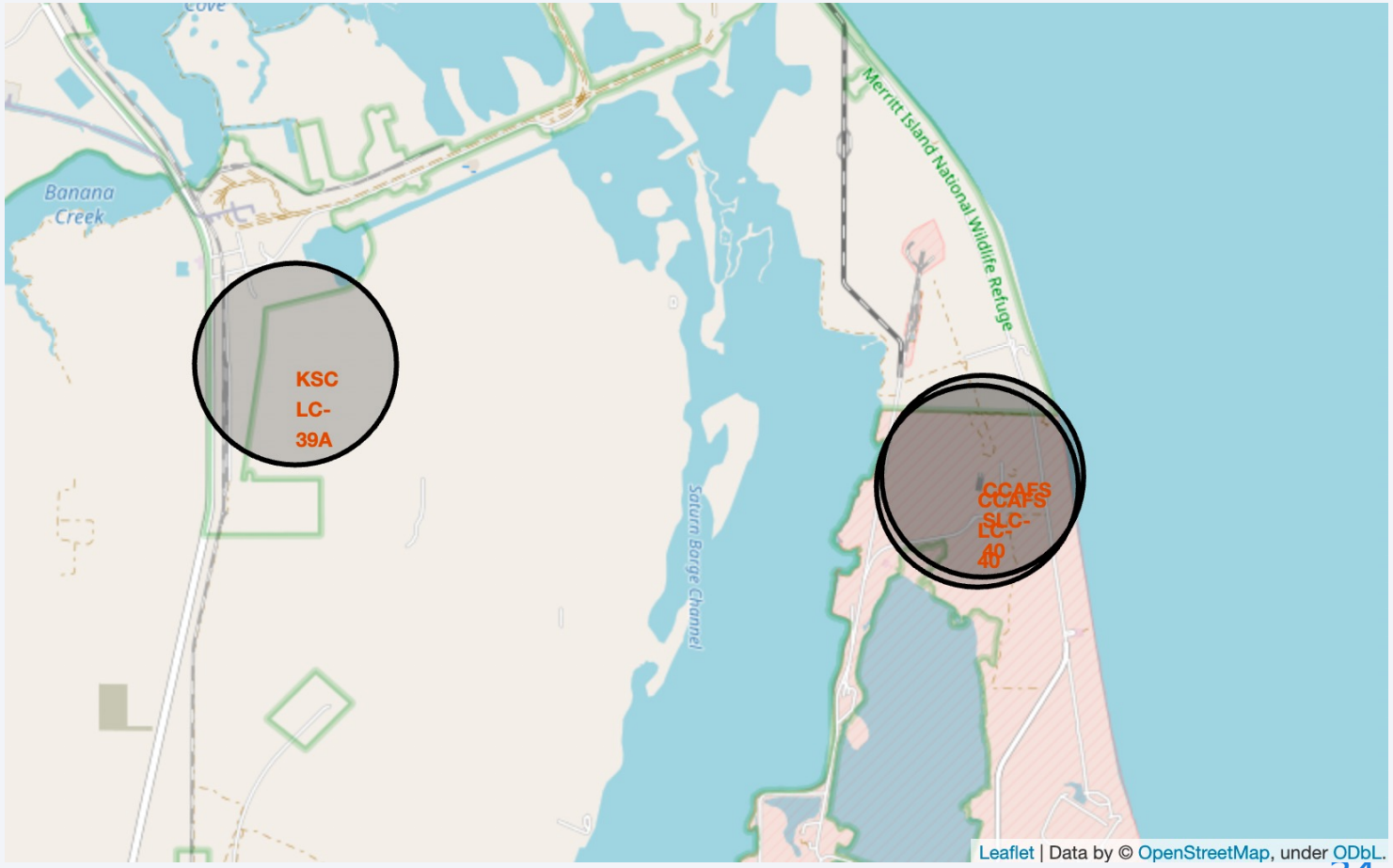
Launch Site Locations I

- Map shows the Launch Site Locations
- All Launch Sites are near the ocean



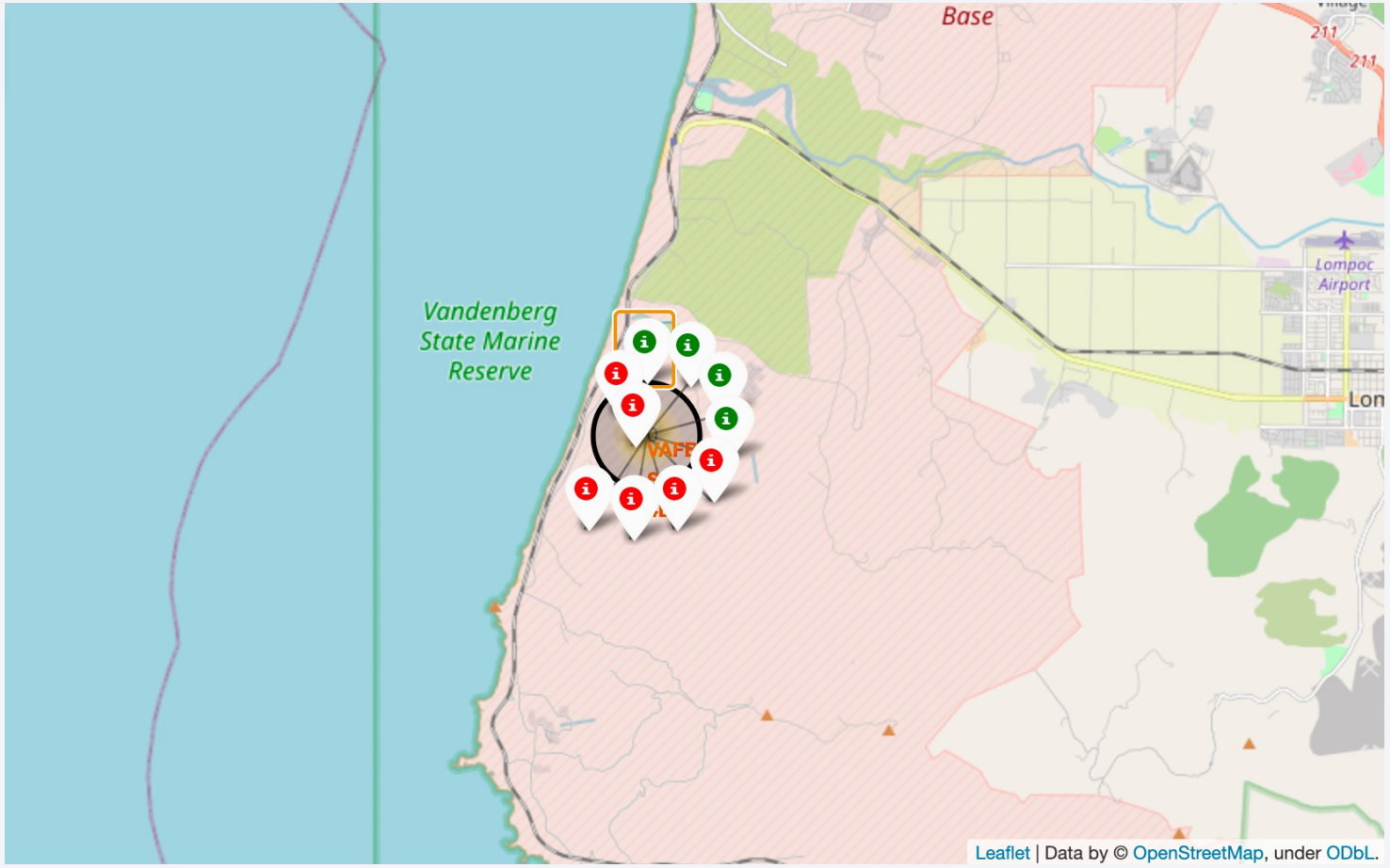
Launch Site Locations II

- Map shows the launch sites in Florida, since they are close together



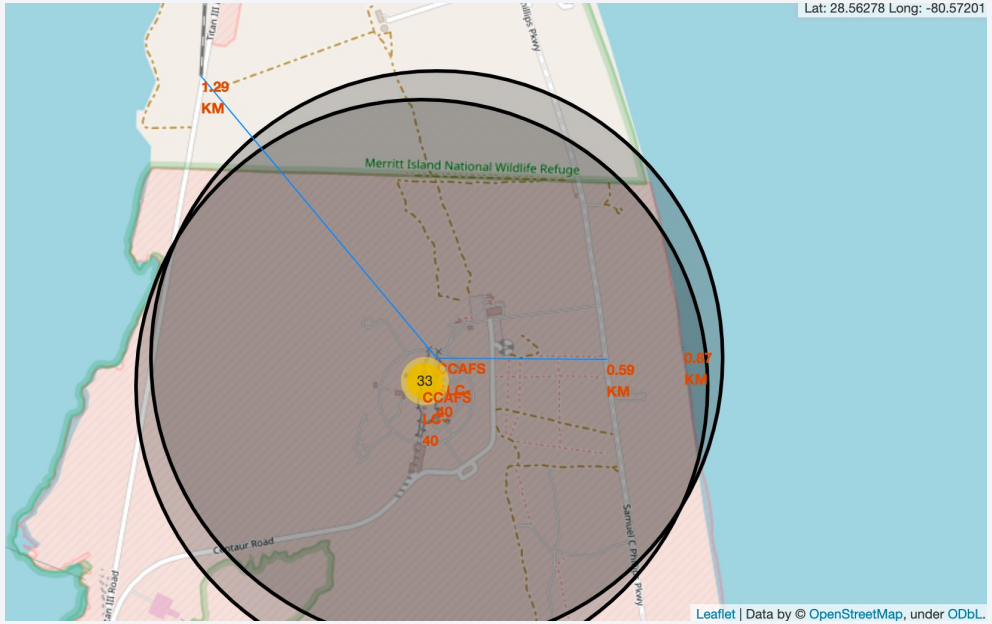
Color-labeled launch outcomes

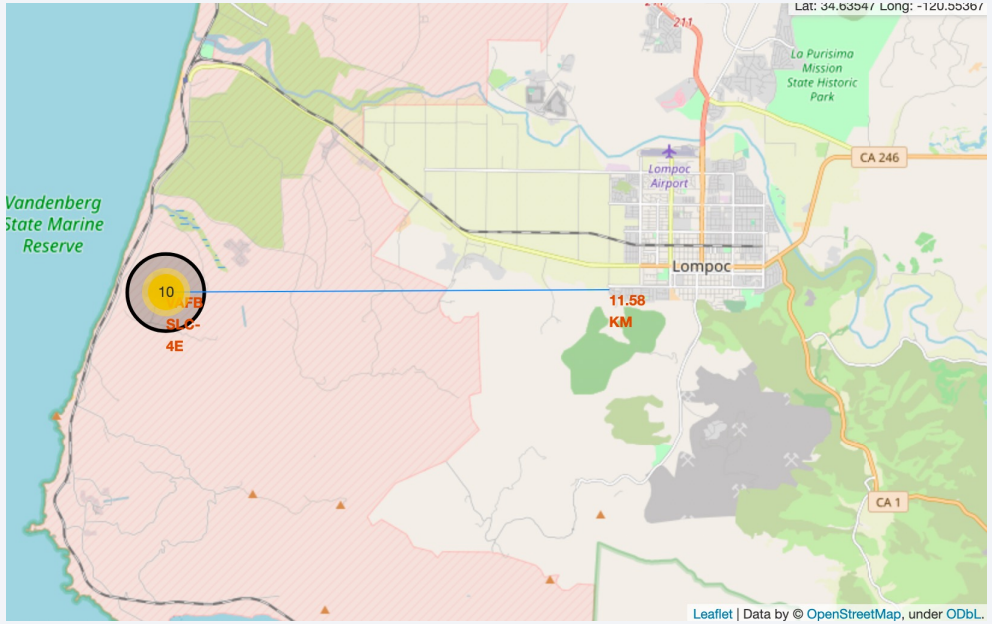
- color-labeled marker clusters where placed on the map to show the launch was successful (green) or unsuccessfull (red)

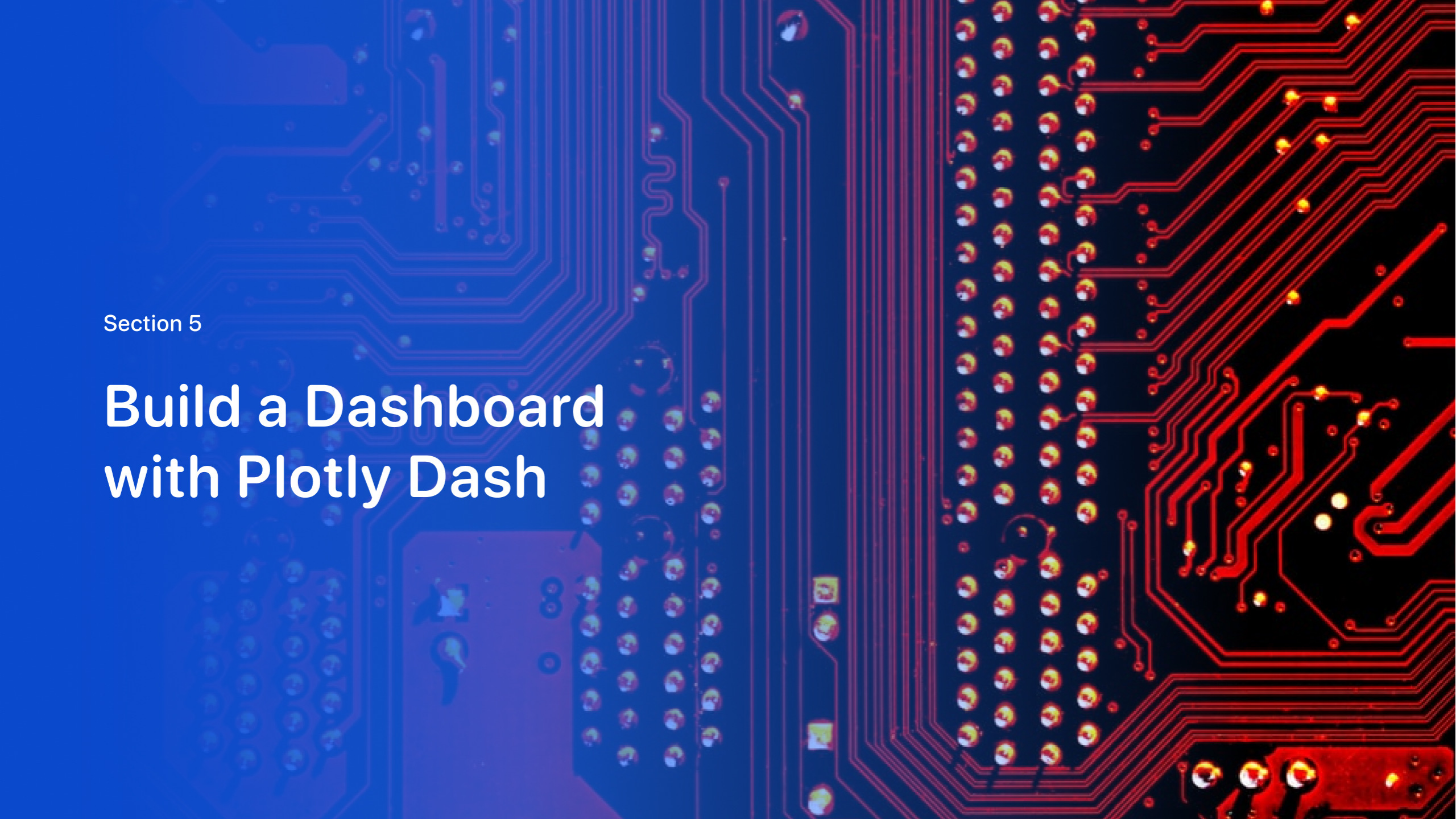


Proximity Analysis

- The Screenshots show the proximity to the nearby cities, railways and highways



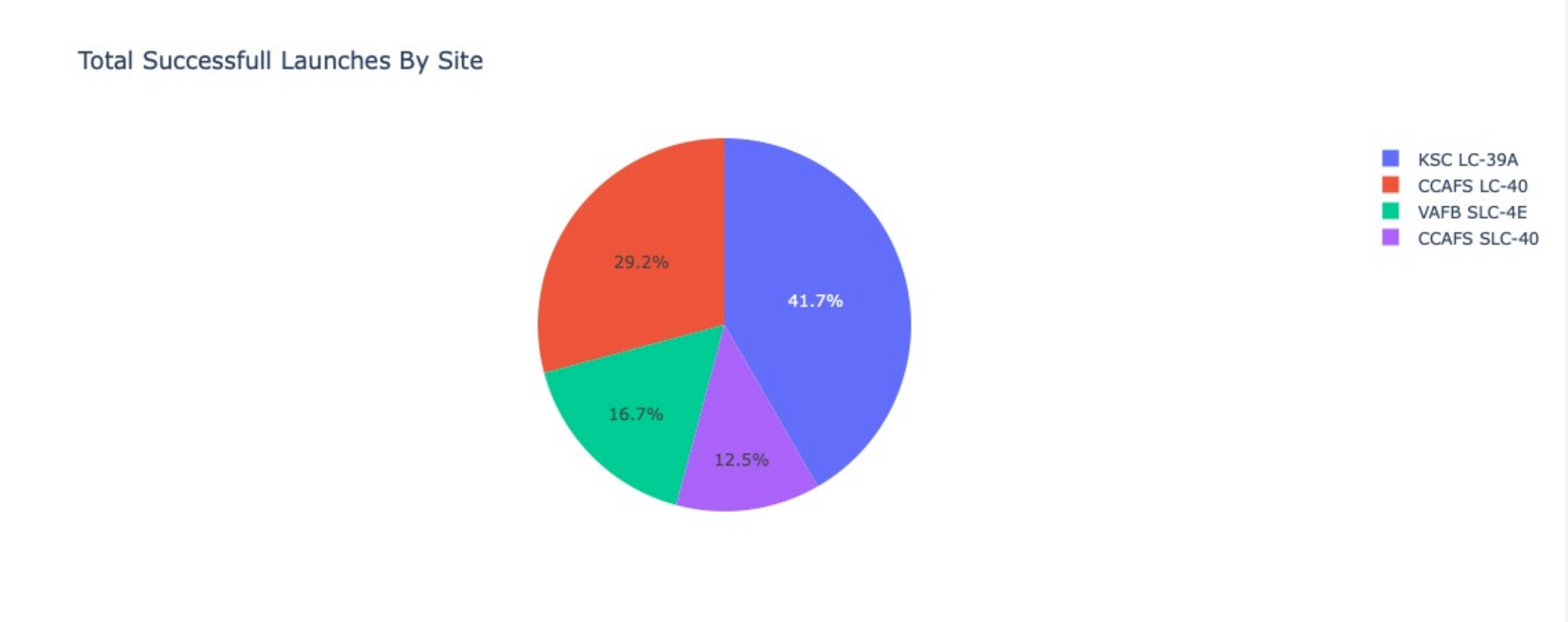




Section 5

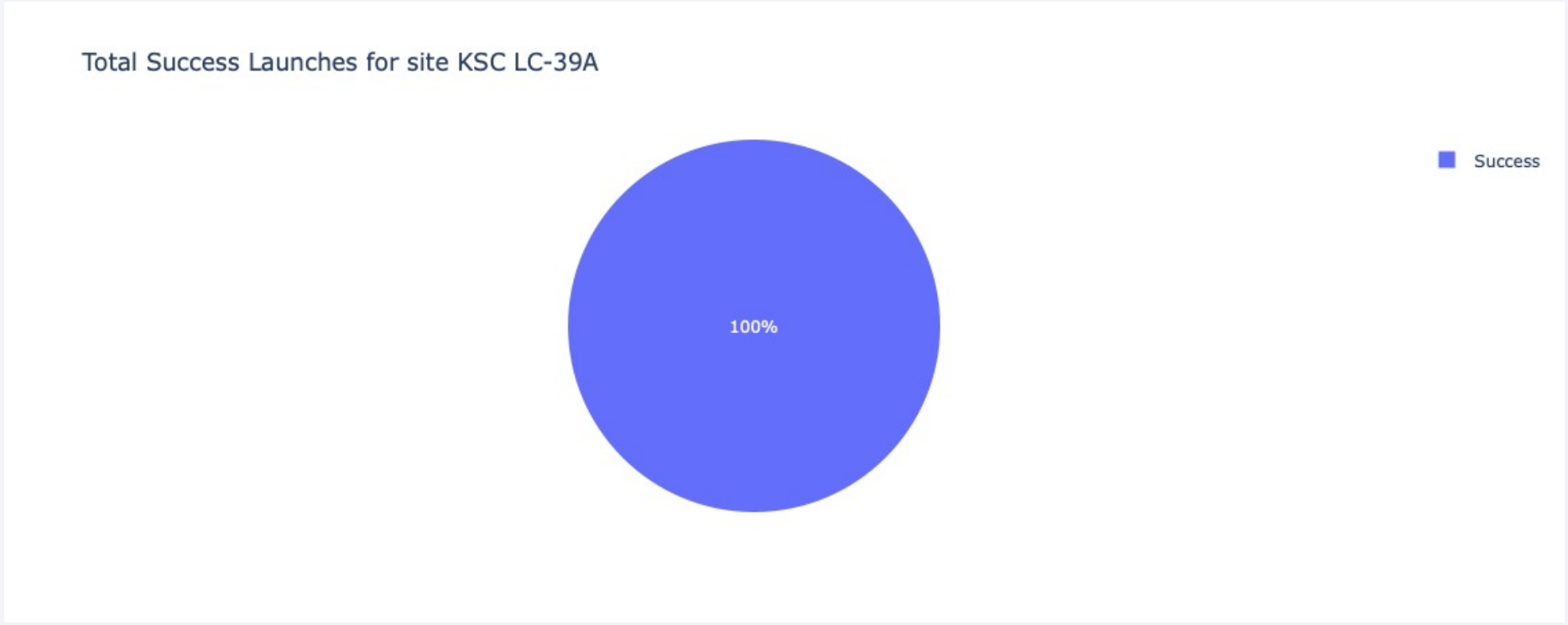
Build a Dashboard with Plotly Dash

Total successful Launches ratio per Site



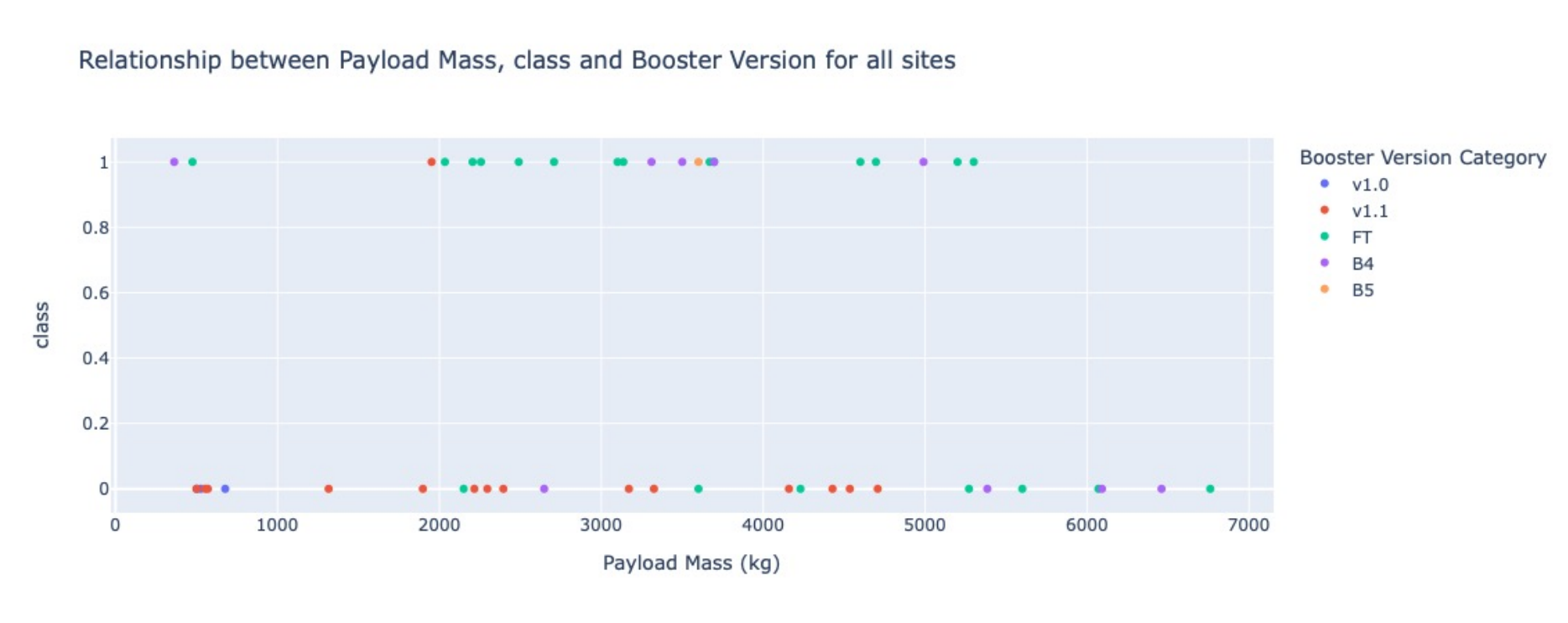
- KSC LC-39A has the best ratio for successful launches

Successful Launch Ratio for KSC LC-39A



The succesful launch Ratio for KSC LC-39A is 100%

Successful/Failed Landings for different Payload Sizes and Booster Versions



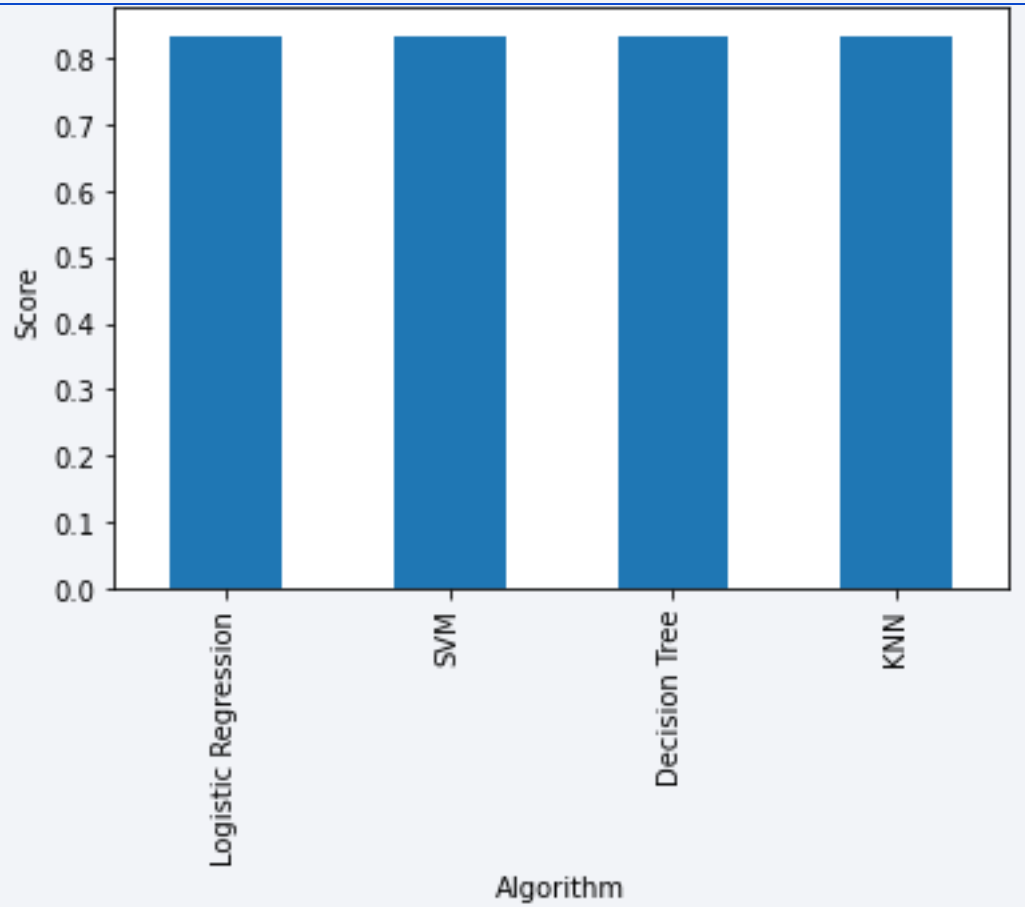
Plotly dashboard has a Payload range selector. However, this is set from 0-7500 instead of the maximum Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color.

Section 6

Predictive Analysis (Classification)

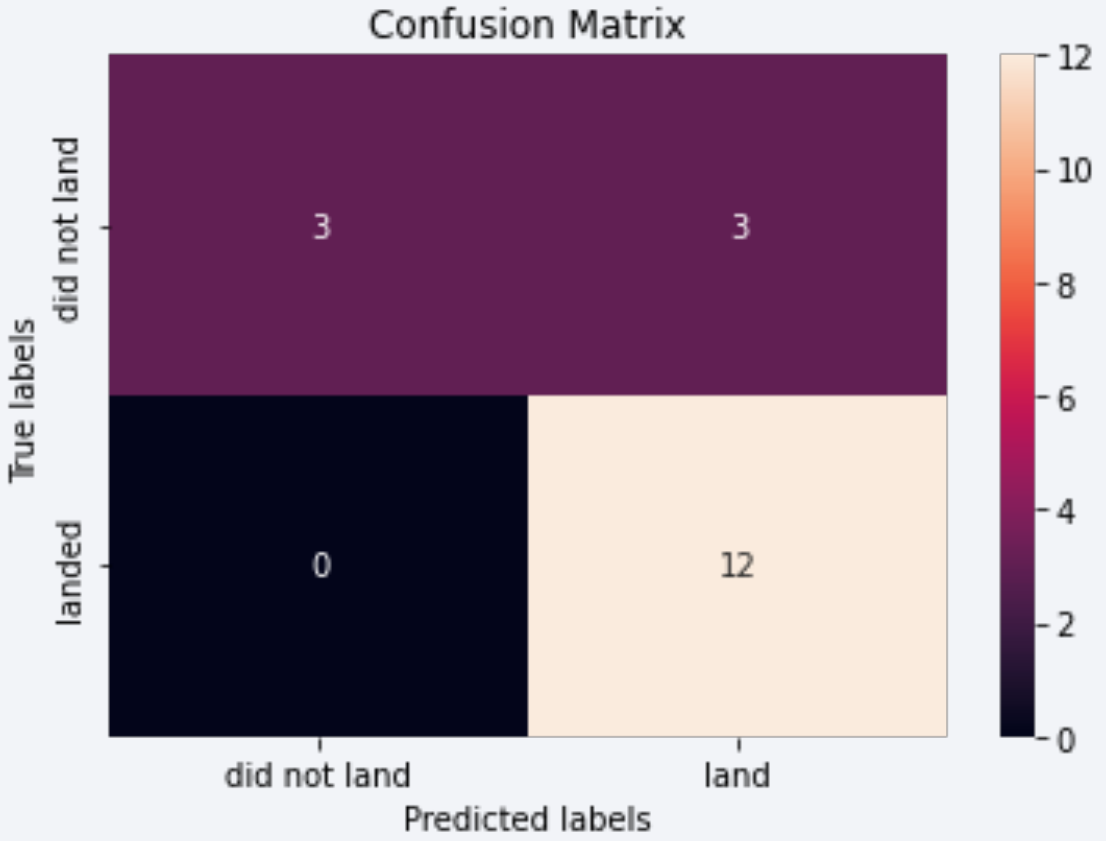
Classification Accuracy

- This Visualization shows the accuracy score of the different Algorithms
- As you can see all Algorithms perform equally good, so every one of this models could be chosen



Confusion Matrix

- All models perform equally good, and the confusion matrices are the same
- the screenshot shows the confusion matrix of the k-Nearest Neighbors Algorithm
- You can see that the Algorithm predicted 15 Instances correctly (3 did not land and 12 landed)
- three Instances where predicted false as landed while they did not land (false positive)



Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save cost
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

Appendix

- Github Url:
- https://github.com/mircohoehne/IBM_Applied_Data_Science_Capstone

Thank you!