IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Filip

Outline

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

Executive Summary

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

Introduction

- Project background and context

  Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and webscraping
- Perform data wrangling
  - The gathered data was enhanced by generating a landing result classification using outcome data, subsequent to summarizing and analyzing various features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was standarized, divided into two groups: training data and test data. Then four different classification models were performed to find the best model.

Data Collection

- Datasets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/)
- from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using webscraping

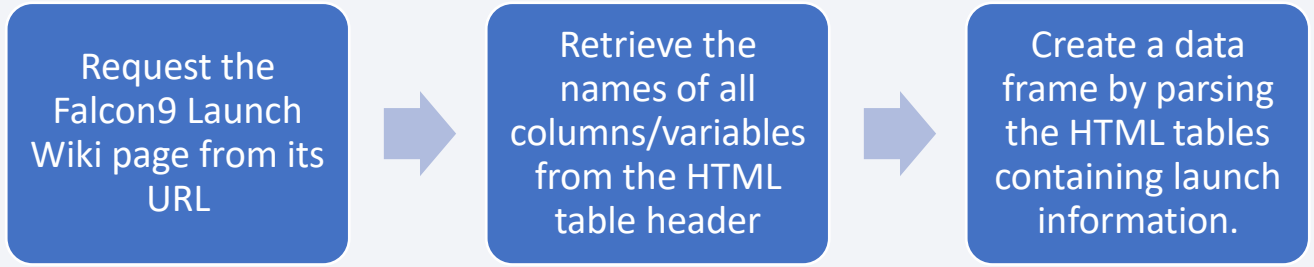
Data Collection – SpaceX API

- SpaceX provides a public API for accessing and utilizing data. The mentioned API was employed in accordance with the adjacent flowchart, followed by the persistence of data.
- https://github.com/FilipK206/Applied-Data_Science_Capstone/blob/bca5f6143bb0ee7f38499e1aab8713a206e928b5/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- https://github.com/FilipK206/Applied-Data_Science_Capstone/blob/bca5f6143bb0ee7f38499e1aab8713a206e928b5/jupyter-labs-webscraping.ipynb



Data Wrangling

- We conducted exploratory data analysis to establish the training labels. We computed the count of launches at each site, as well as the frequency of each orbit. Additionally, we generated a landing outcome label based on the outcome column and exported the results to a CSV file.
- https://github.com/FilipK206/Applied-Data_Science_Capstone/blob/bca5f6143bb0ee7f38499e1aab8713a206e928b5/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We examined the data by visualizing the connections between flight number and launch site, payload and launch site, success rate for each orbit type, flight number and orbit type, and the yearly trend in launch success.
- https://github.com/FilipK206/Applied-Data_Science_Capstone/blob/c753b01e2209e5ff56c56600780276f0fb41cecf/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

We utilized exploratory data analysis (EDA) with SQL to extract insights from the data. We formulated queries to discover, for example:

- The total payload mass carried by boosters launched by NASA (CRS).
- The total number of successful and failed mission outcomes.
- The names of unique launch sites in the space mission.
- The average payload mass carried by booster version F9 V1.1.
- The failed landing outcomes on the drone ship, along with their booster version and launch site names
- https://github.com/FilipK206/Applied-Data_Science_Capstone/blob/bca5f6143bb0ee7f38499e1aab8713a206e928b5/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We labeled all launch sites and incorporated map elements such as markers, circles, and lines to signify the success or failure of launches for each site on the Folium map. We assigned the launch outcomes to classes 0 and 1, where 0 represents failure and 1 represents success. Utilizing color-labeled marker clusters, we identified launch sites with relatively high success rates. cities.
- https://github.com/FilipK206/Applied-Data_Science_Capstone/blob/c753b01e2209e5ff56c56600780276f0fb41cecf/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- https://github.com/FilipK206/Applied-Data_Science_Capstone/blob/3af668170d0b95b057639f68862cfed75f1ddde5/spacex_dash_app.py

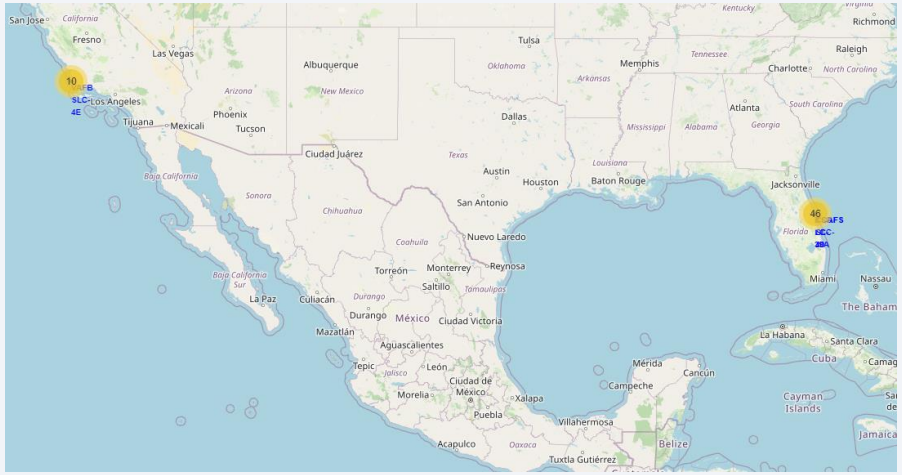
Predictive Analysis (Classification)

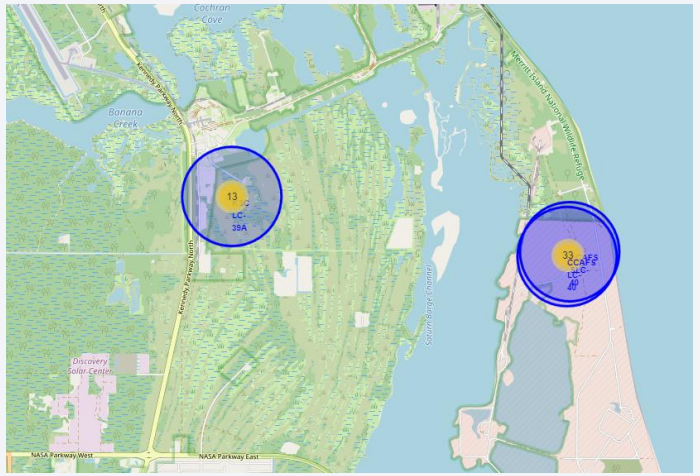
- We loaded the data using NumPy and Pandas, performed data transformation, and divided it into training and testing sets. We constructed various machine learning models and fine-tuned different hyperparameters using GridSearchCV. Accuracy served as the metric for our model evaluation, and we enhanced the model through feature engineering and algorithm tuning. Ultimately, we identified the best-performing classification models.
- https://github.com/FilipK206/Applied-Data_Science_Capstone/blob/c753b01e2209e5ff56c56600780276f0fb41cecf/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
  - There are four different launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
  - Total payload mass carried by boosters launched by NASA (CRS): 45596 KG
  - Average payload mass carried by booster version F9 v1.1: 2535 KG
  - First succesful landing outcome in ground pad was achieved: 2015-12-22
  - total number of successful and failure mission outcomes: 101

Results





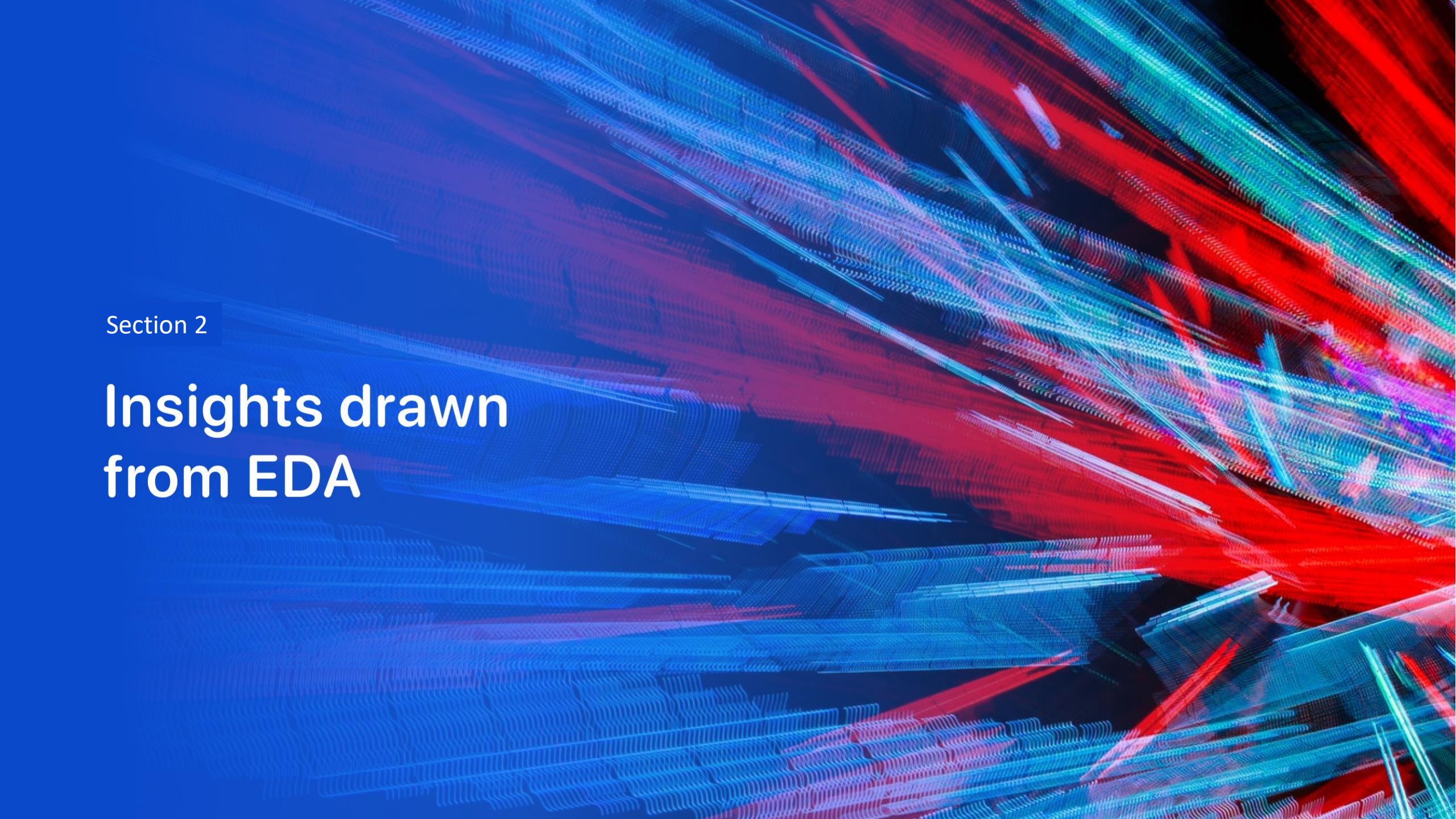
Every launch site was located close to facilities like the coastline, highway and railway with a safe distance to the cities. Which increases the likelihood of a positive landing outcome.

Results

Predictive analysis results

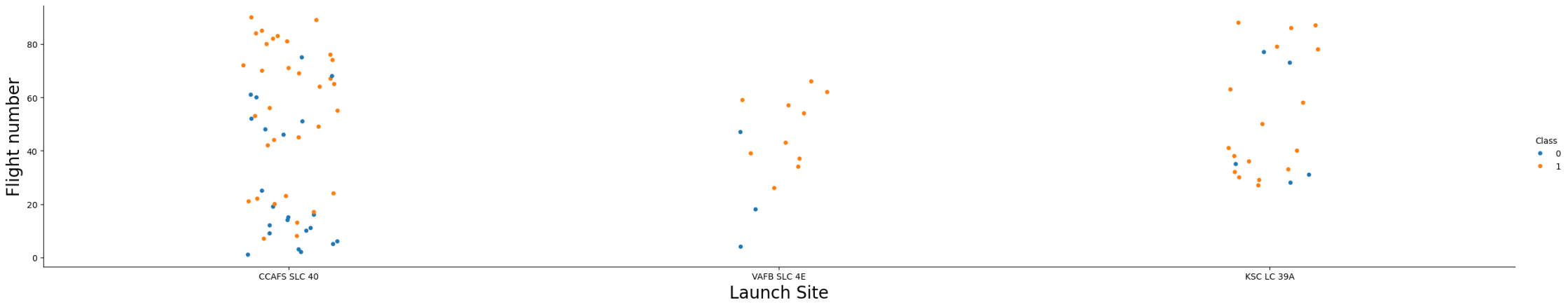
- Accuracy for Logistics Regression method: 0.834
- Accuracy for Support Vector Machine method: 0.834
- Accuracy for Decision tree method: 0.778
- Accuracy for K nearest neighbours method: 0.834

The best models for this task are logistics regression, support vector machine method and K nearest neighbours.


Section 2
Insights drawn from EDA

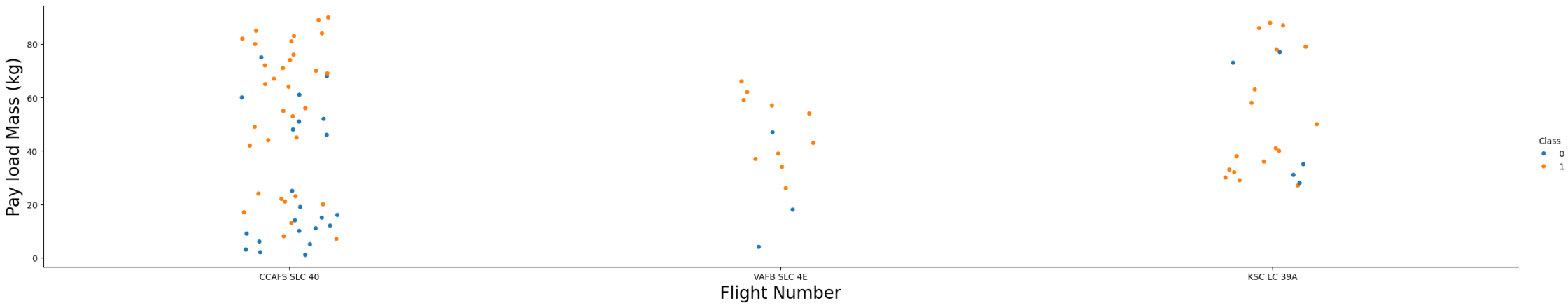
Flight Number vs. Launch Site

It appears that the greater the number of flights at a launch site the greater the succes rate at a launch site

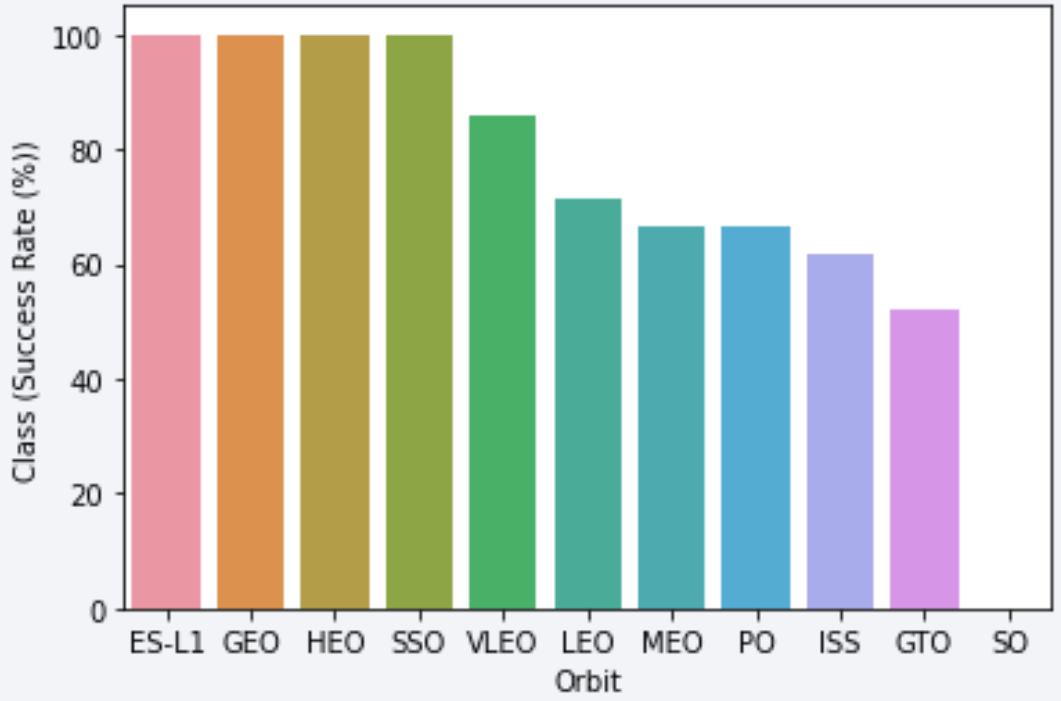


Payload vs. Launch Site

The most heavy payloads were conducted on a launch site CCAFS SLC 40.

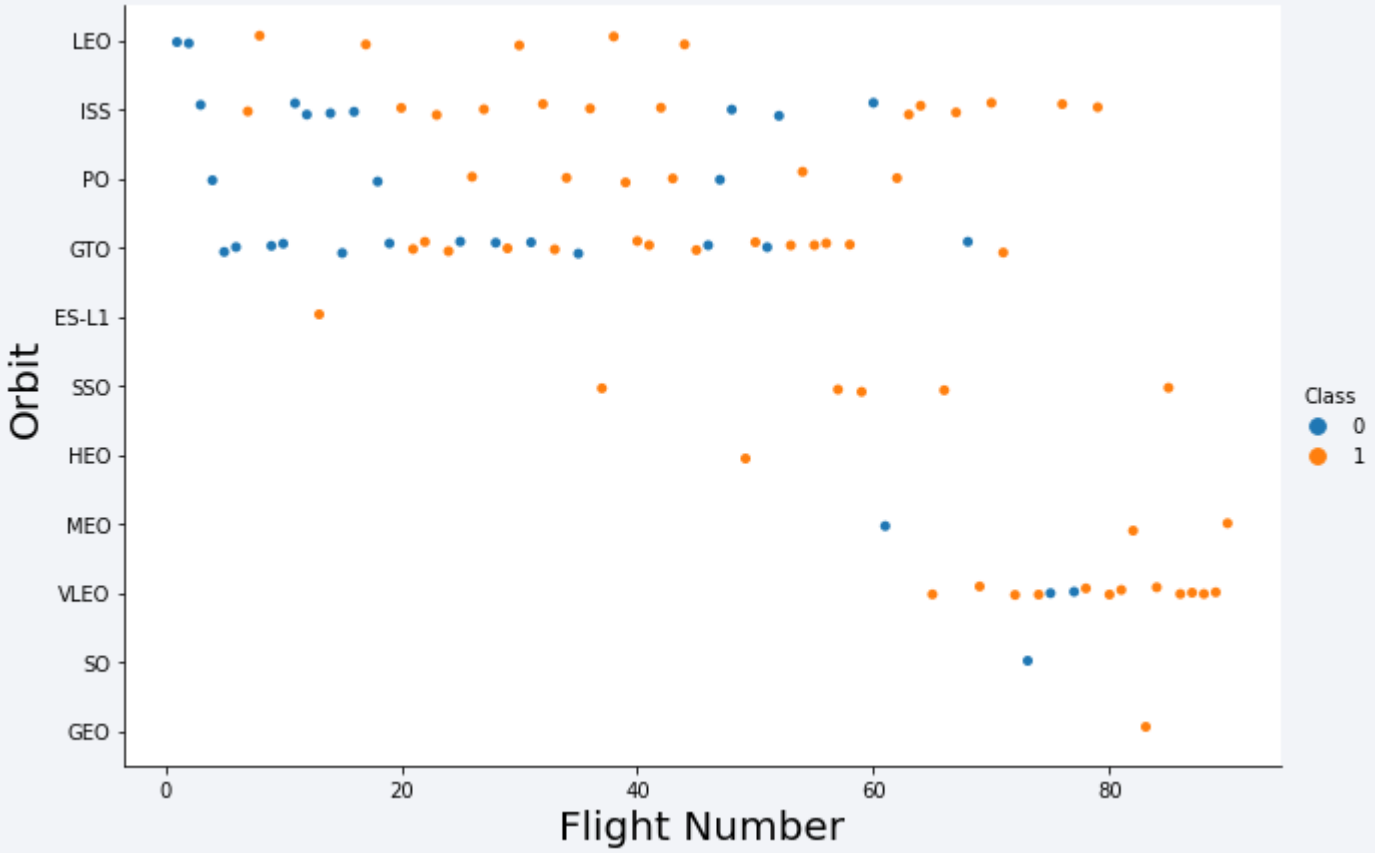


Success Rate vs. Orbit Type



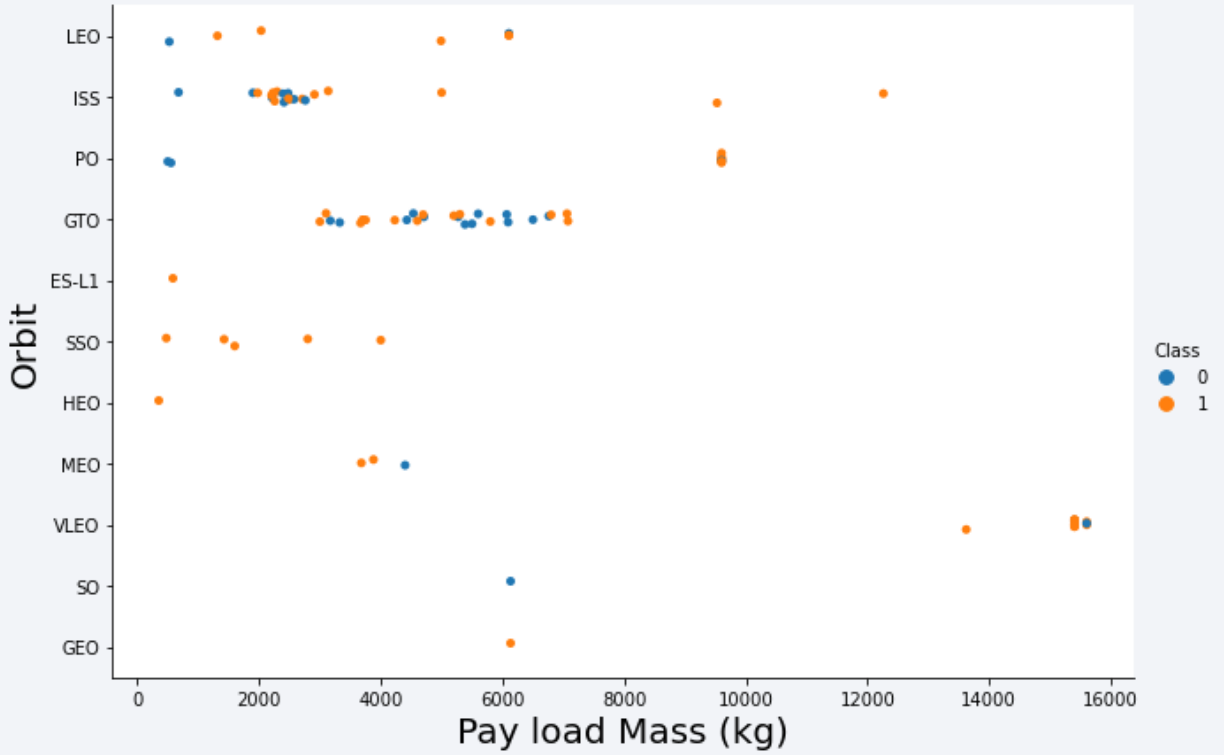
The most success rate had ES-L1, GEO, HEO, SSO

Flight Number vs. Orbit Type



The plot shows the Flight number vs Orbit type. We can see that the success is related to the number of flights for Orbit LEO and MEO.

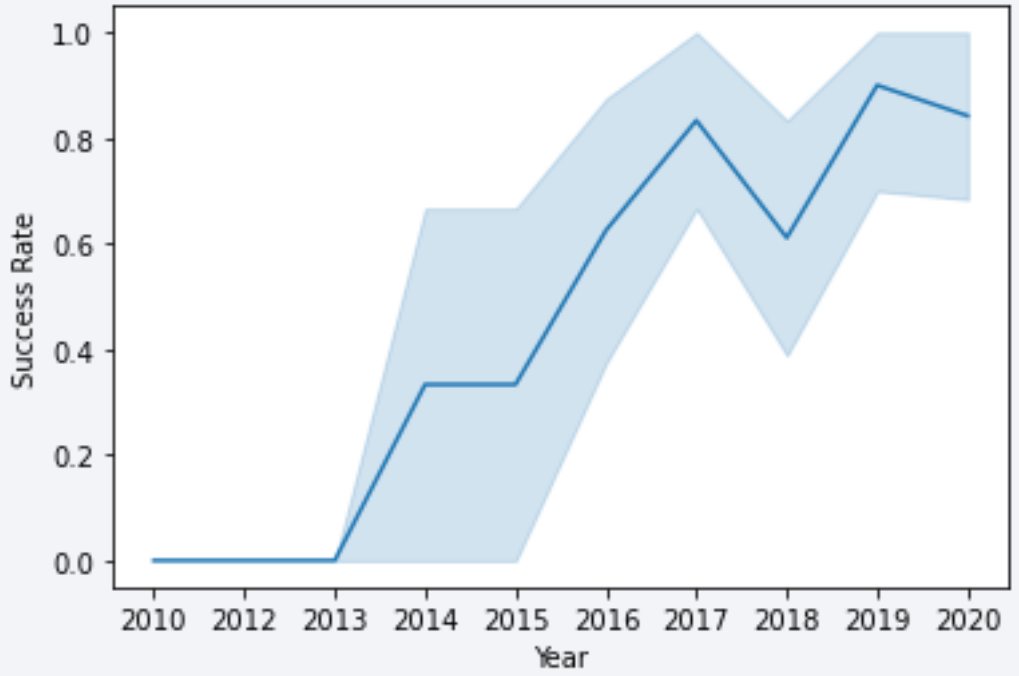
Payload vs. Orbit Type



The plot shows the Payload vs Orbit type. We can see that there is no relationship for the GTO orbit. There could be a correlation with heavy payload and success landing for PO, LEO and ISS orbits.

Launch Success Yearly Trend

The plot below shows that since 2013 success rate increased untill 2017



All Launch Site Names

We used the query with the key word DISTINCT to show only unique launch sites from the SpaceX data.

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

We used the query below to display 5 records where launch sites begin with „CCA”

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

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	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

We used the query below with a word SUM to calculate total payload carried by boosters from NASA.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

We used the query below with a word AVG to get average payload carried by booster version F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

We used the query below with word MIN to get first date for a succesful landing and word WHERE to filter the data.

```
%sql SELECT MIN(Date), Mission_Outcome, Landing_Outcome FROM SPACEXTBL WHERE Mission_Outcome = 'Success' and Landing_Outcome like '%ground%' ;
```

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

MIN(Date)	Mission_Outcome	Landing_Outcome
2015-12-22	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

We used the query below with word WHERE to filter the data and obtain list of landings with Payload between 4000 and 6000

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535
F9 v1.1 B1011	4428
F9 v1.1 B1014	4159
F9 v1.1 B1016	4707
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1030	5600
F9 FT B1021.2	5300
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 FT B1031.2	5200
F9 B4 B1043.1	5000
F9 FT B1032.2	4230
F9 B4 B1040.2	5384
F9 B5 B1046.2	5800
F9 B5 B1047.2	5300
F9 B5B1054	4400
F9 B5 B1048.3	4850
F9 B5 B1051.2	4200
F9 B5B1060.1	4311
F9 B5 B1058.2	5500
F9 B5B1062.1	4311

Total Number of Successful and Failure Mission Outcomes

We used the query with a word COUNT to obtain total number of succesful and failure mission outcomes.

```
%sql select count(Mission_Outcome) from SPACEXTBL;
```

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

count(Mission_Outcome)
101

Boosters Carried Maximum Payload

We used the query with a word MAX to get maximum payload carried by boosters and word WHERE to filter the data.

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

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

We used the query with a phrase substr(Date, 6,2) to filter months and substr(Date, 0,5) to filter the year.

```
%sql SELECT substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,0,5)='2015' and Landing_Outcome LIKE 'Failure%';
```

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

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

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

We used the query with a word WHERE to filter the data and GROUP BY to obtain grouped data

```
%sql SELECT Landing_Outcome, count(Landing_Outcome) as Total_Number from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Total_Number desc
```

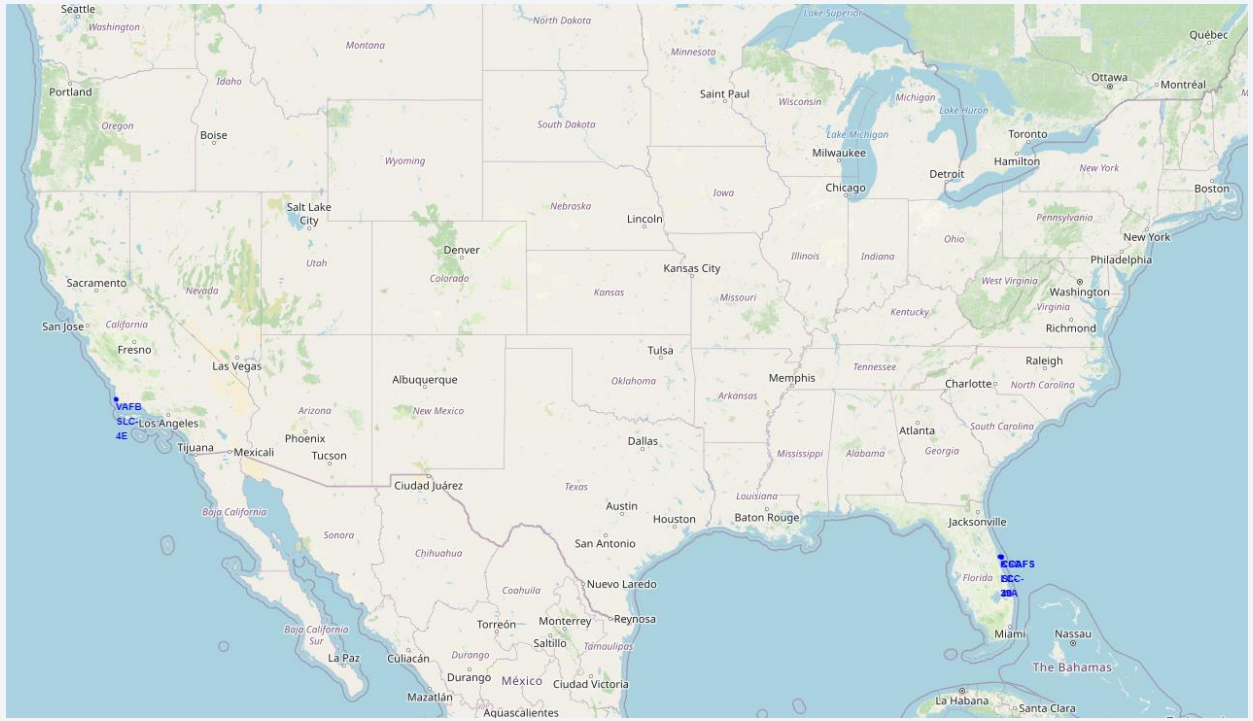
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Total_Number
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

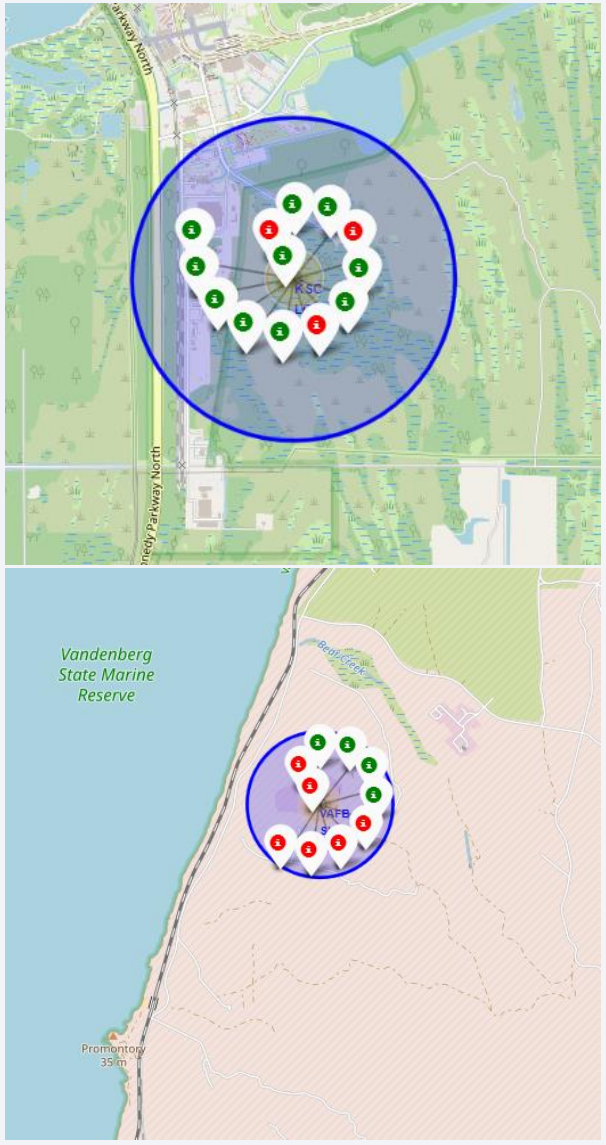

Section 3
Launch Sites
Proximities Analysis

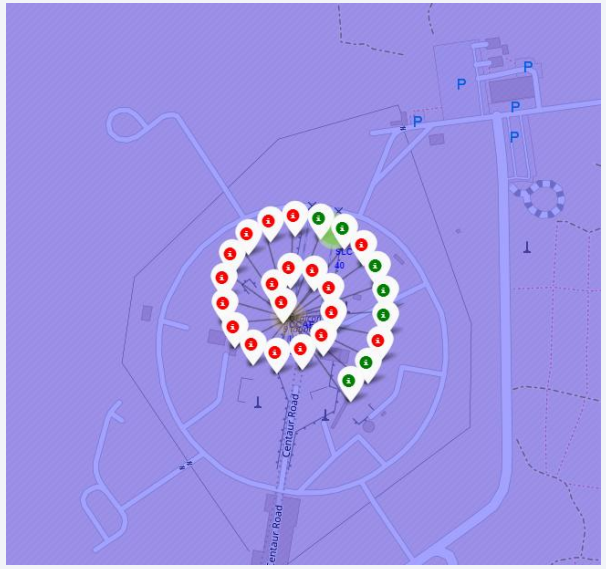
All launch sites global map markers

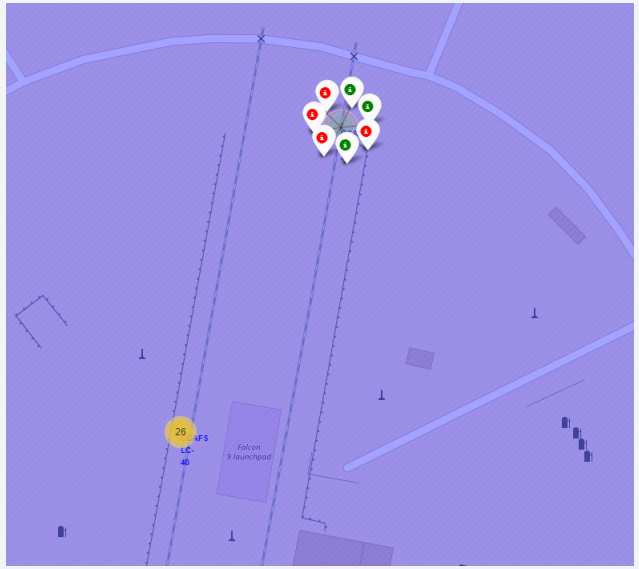
On the map below we can see that launch sites are located in United States of America in States California and Florida.



Markers showing launch sites

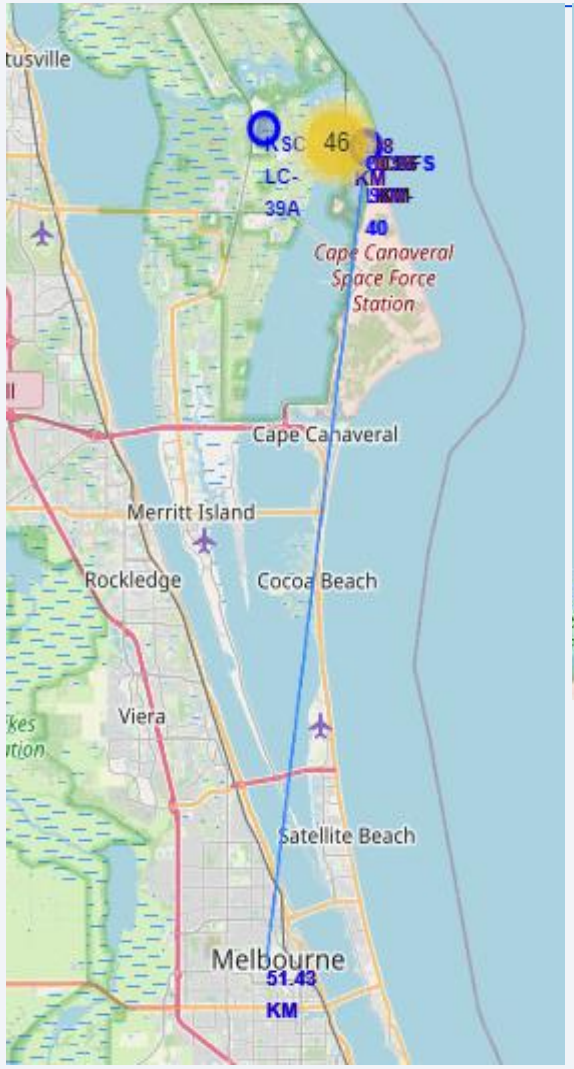


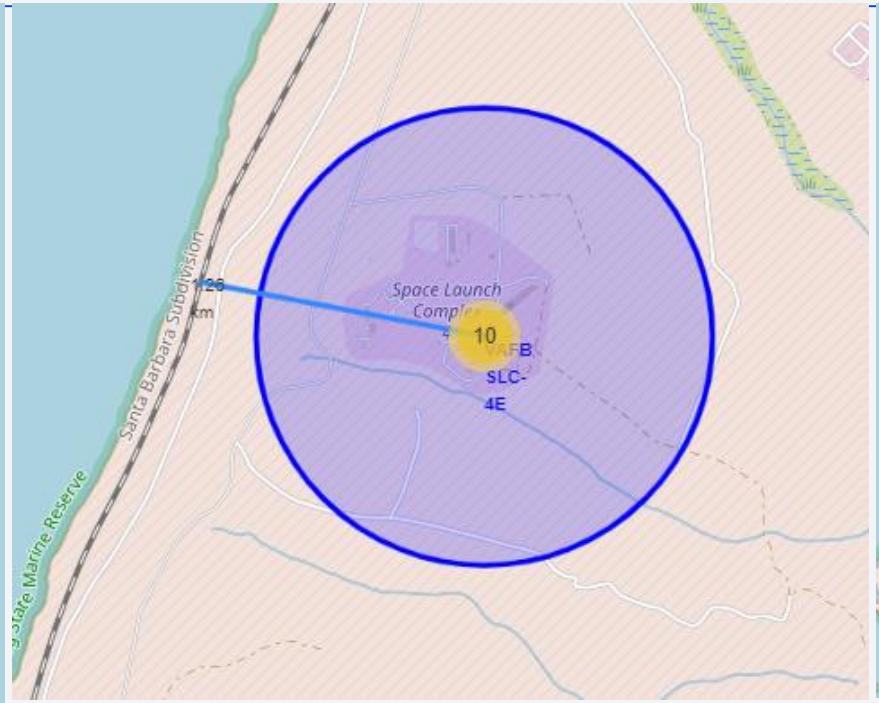


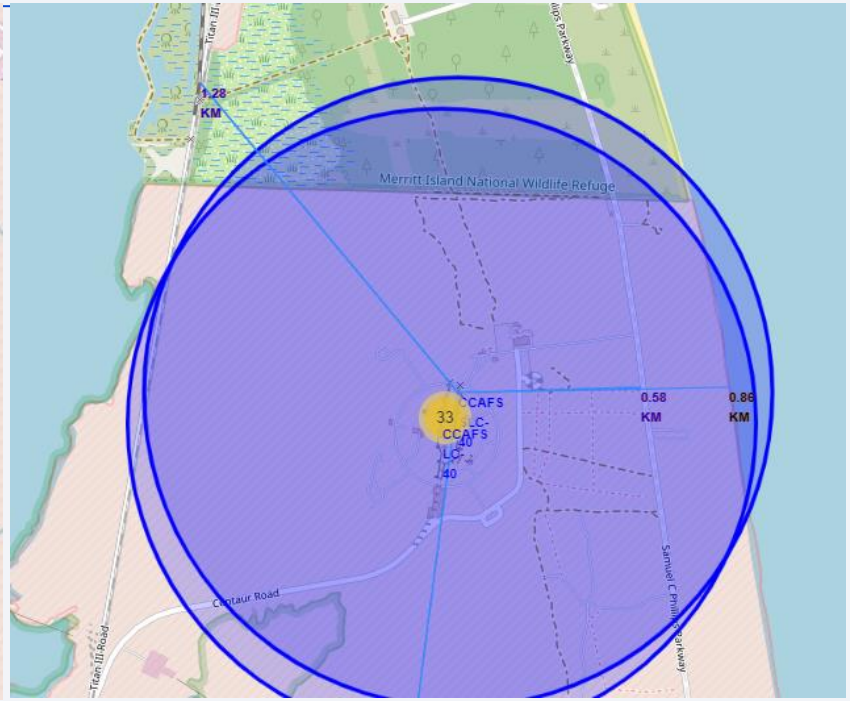


Green markers shows successful launches and Red markers shows failures.

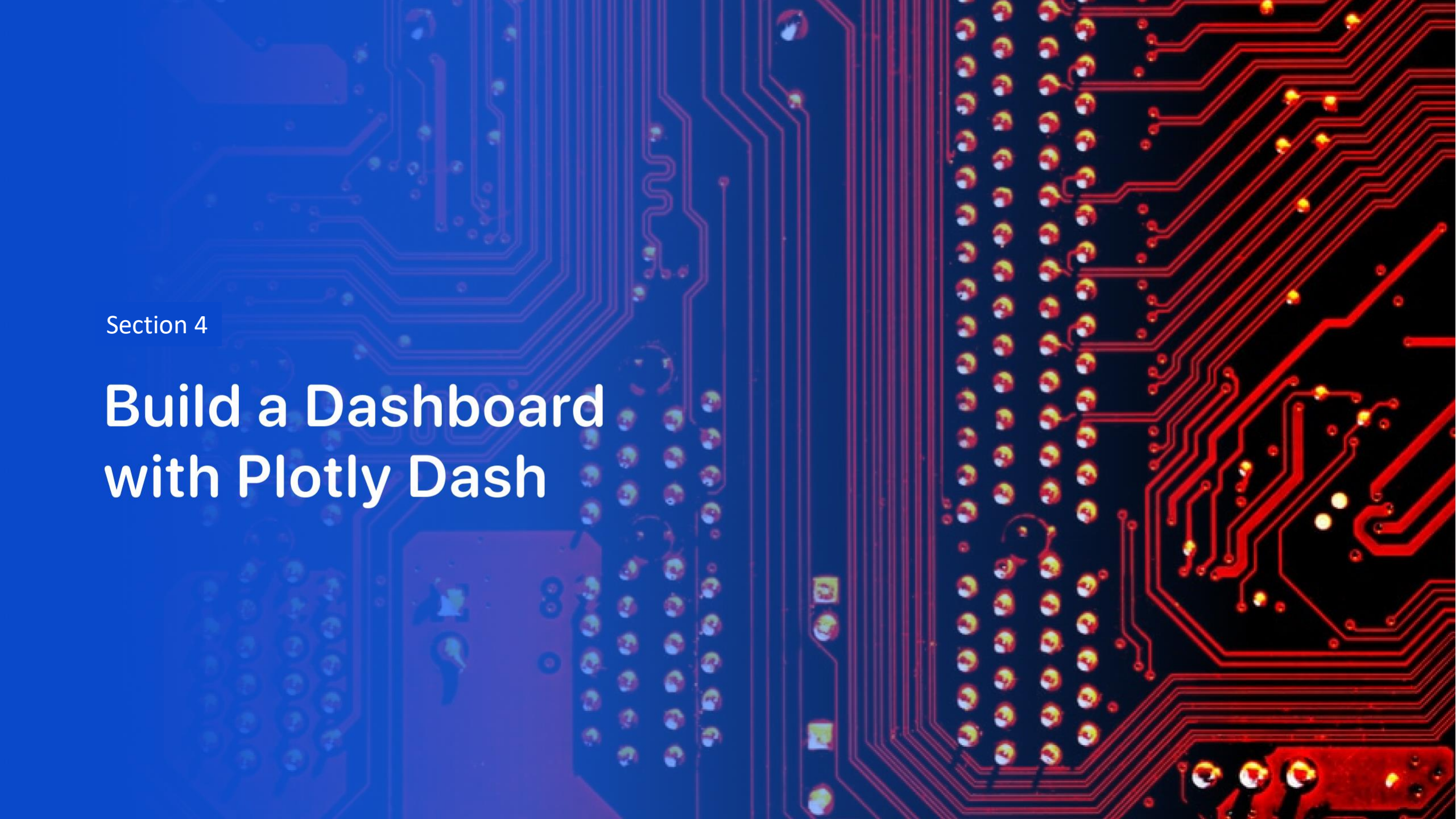
Launch Site distance to landmarks







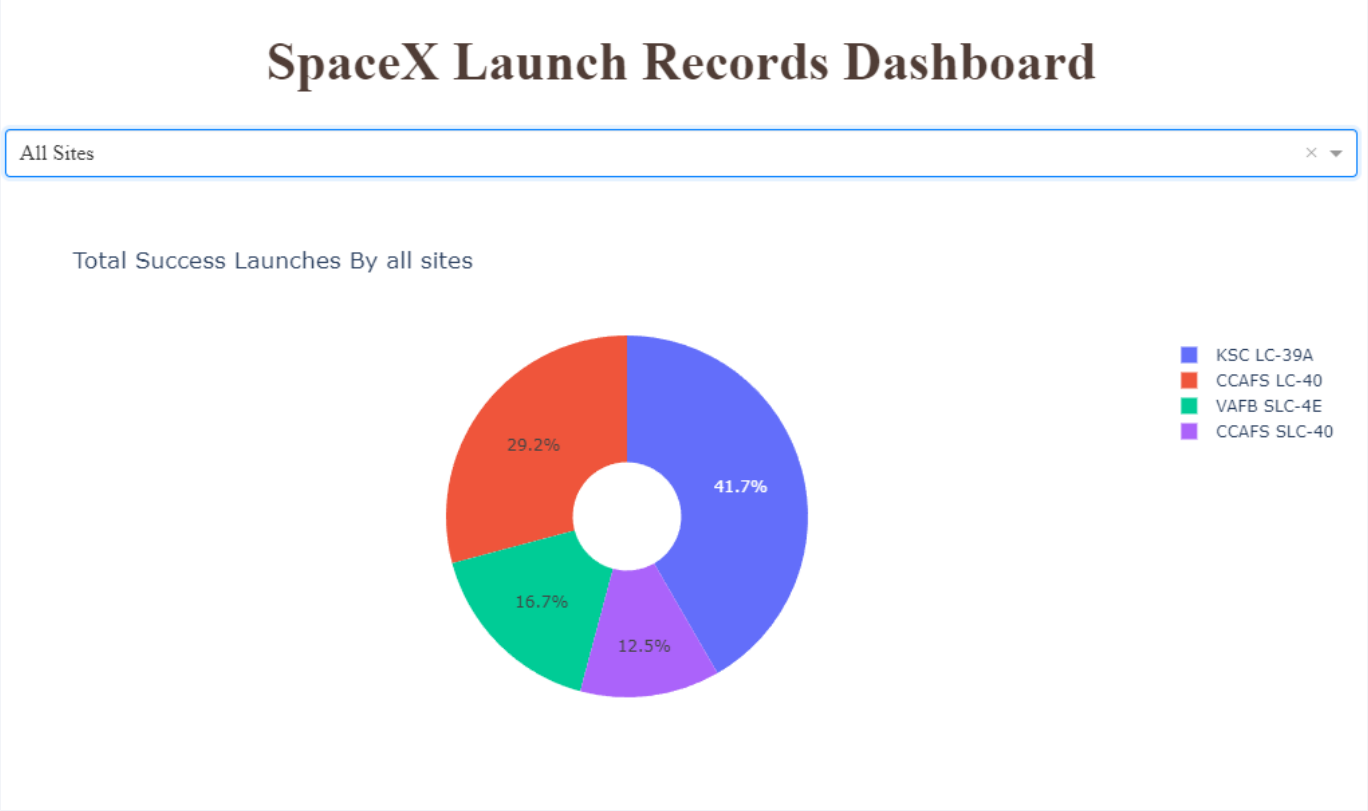
On the maps shown above we can see the distance between the launch sites and the railway, coastline, highway and the city.


Section 4
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

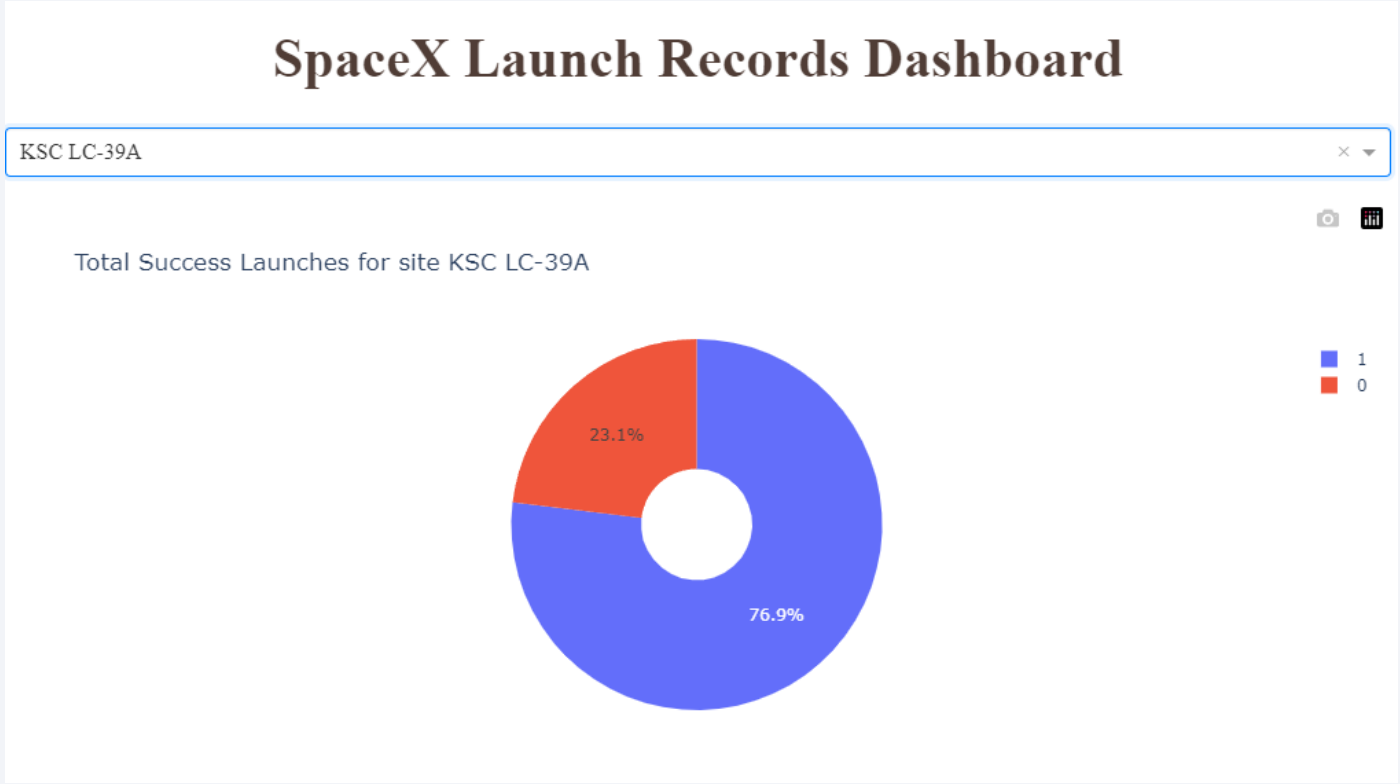
We can see below the pie chart with a total success launches by all sites.

KSC LC-39A had the best performance from all launch sites.



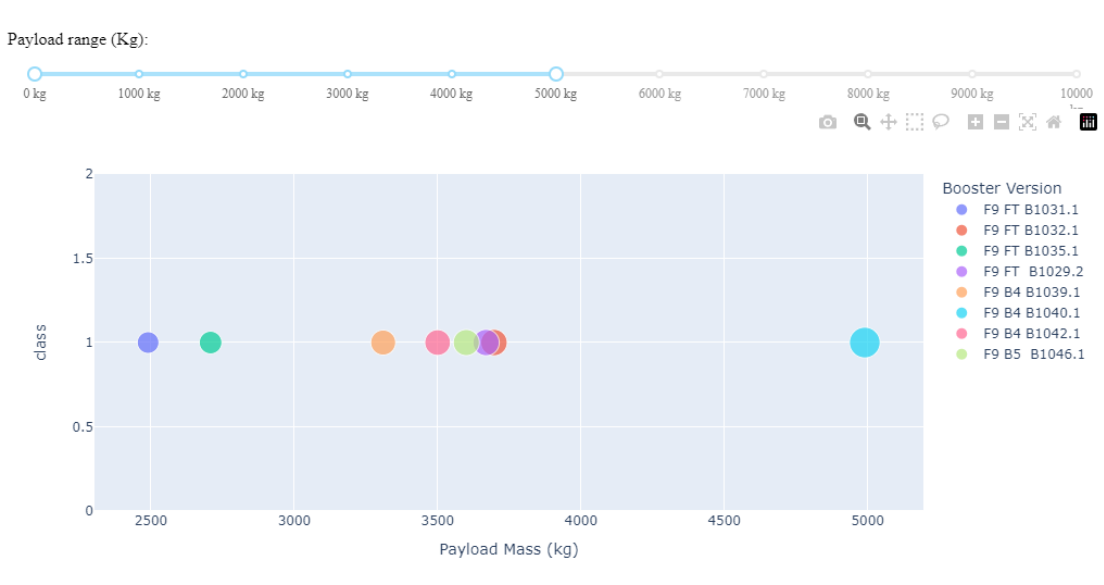
Total success launches for site KSC LC-39A

On the pie chart below we can see that a success rate for this launch sites was 76.9% and failure rate was 23.1%



Payload vs Launch Outcome

We can see below performance of all launch sites divided into groups, low weighted payloads (0kg – 5 000kg). and heavy weighted payloads (5 000kg – 10 000kg).



Section 5

Predictive Analysis (Classification)

Classification Accuracy

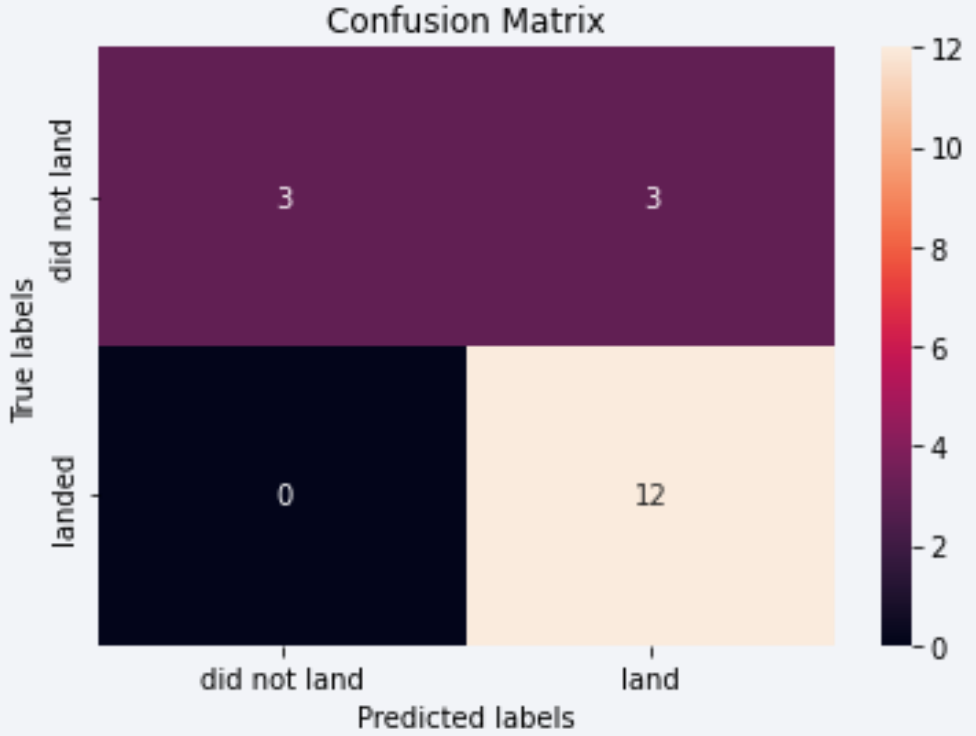
Predictive analysis results

Accuracy for Logistics Regression method: 0.834
Accuracy for Support Vector Machine method: 0.834
Accuracy for Decision tree method: 0.778
Accuracy for K nearest neighbours method: 0.834

The best models for this task are logistics regression, support vector machine method and K nearest neighbours.

Confusion Matrix

We can see below confusion matrix. All three best-performing models got equally the same scores, therefore there is only one confusion matrix.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 untill 2017.
- Orbits ES-L1, GEO, HEO, SSO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The best models for this task are logistics regression, support vector machine method and K nearest neighbours.

Appendix

- None

Thank you!