IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Saeed y Alqhtani
29 jul 2024

Outline

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

Executive Summary

Summary of methodologies

- Data collection with API
- Data collection with Web Scraping
- Data Wrangling.
- Exploratory Data Analysis with SQL.
- Expoloratory Data Analysis With Visualization.
- Interactive map with Folium.
- Interactive Dashboards with Dash.
- Model prediction with Machine Learning.

Summary of all results

- Exploratory Data Analysis result.
- Interactive Analysis visuals.
- Predictive modeling results.

Introduction

- Project background and context

- SpaceX 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 SpaceX 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 SpaceX for a rocket launch

- Problems you want to find answers

- What factors determine if launch was successful?
- 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 using SpaceX API and web scraping from Wikipedia.

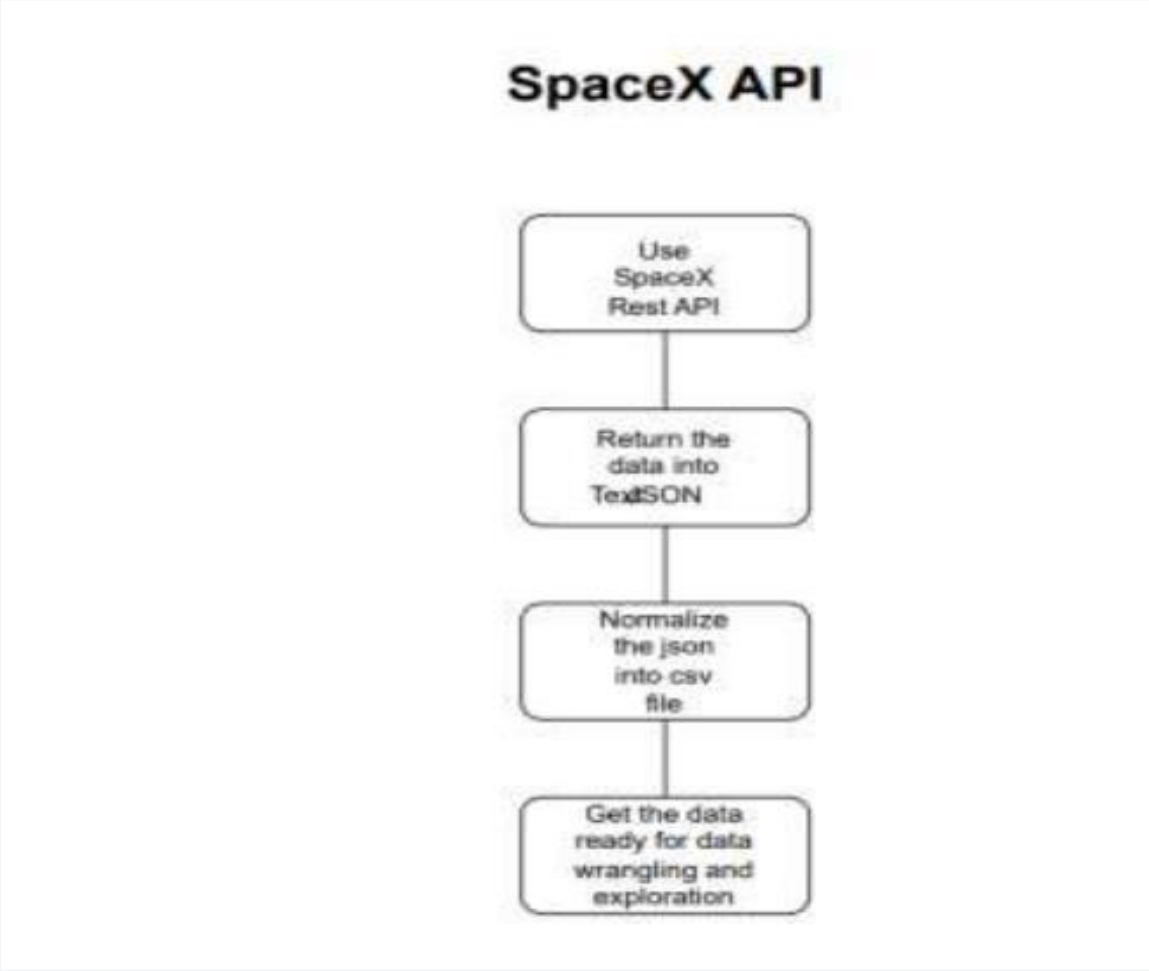
- Perform data wrangling

  One-hot encoding was applied to categorical 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
  - How to build, tune, evaluate classification models

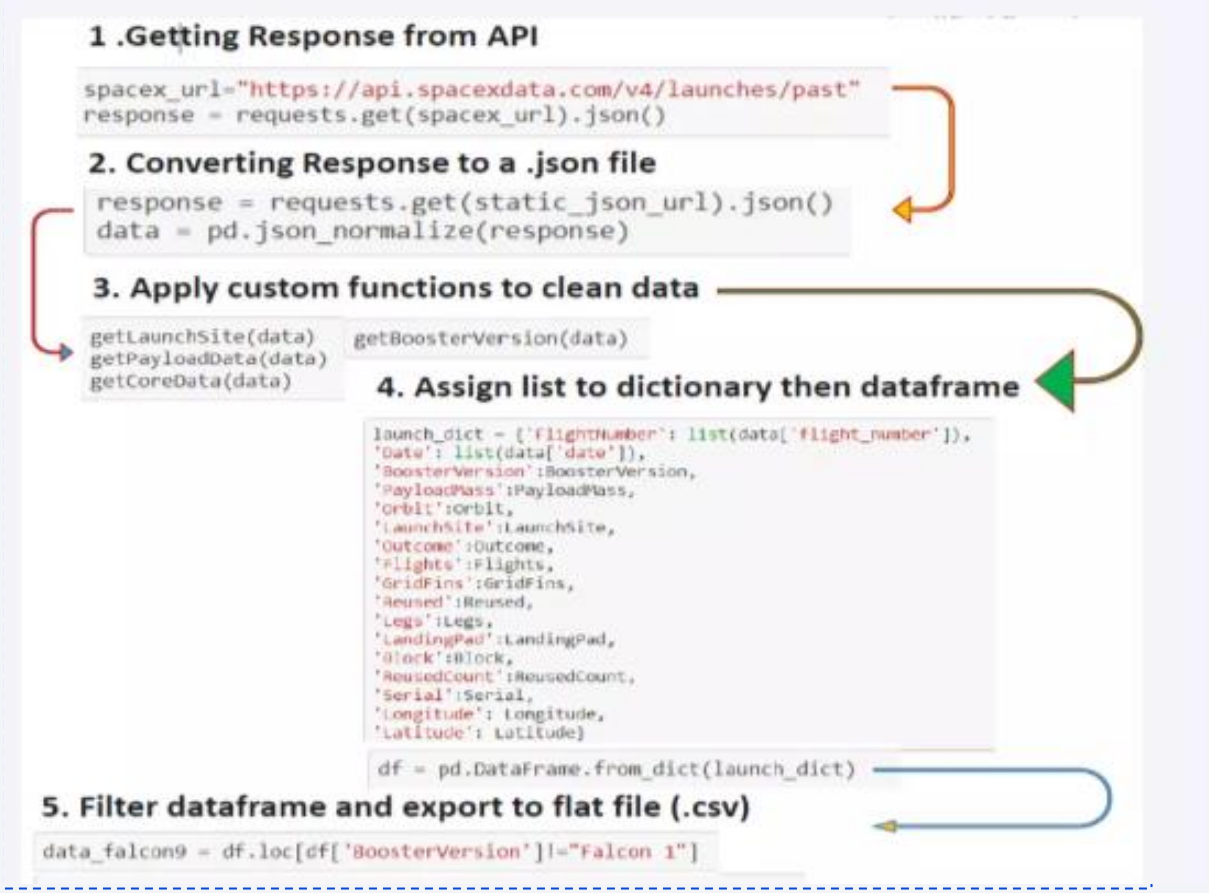
Data Collection

- The Data was collected by various methods
  - Data Collection by SpaceX API
  - Next, I decoded the response content as a Json using .json() function call and turn it into a pandas data frame using .json normalize().
  - Then I cleaned the data by checking for missing values and fill in missing values where it's necessary
  - Also, we performed web scraping from Wikipedia for Falcon 9 launch records with Beautifulsoup



Data Collection – SpaceX API

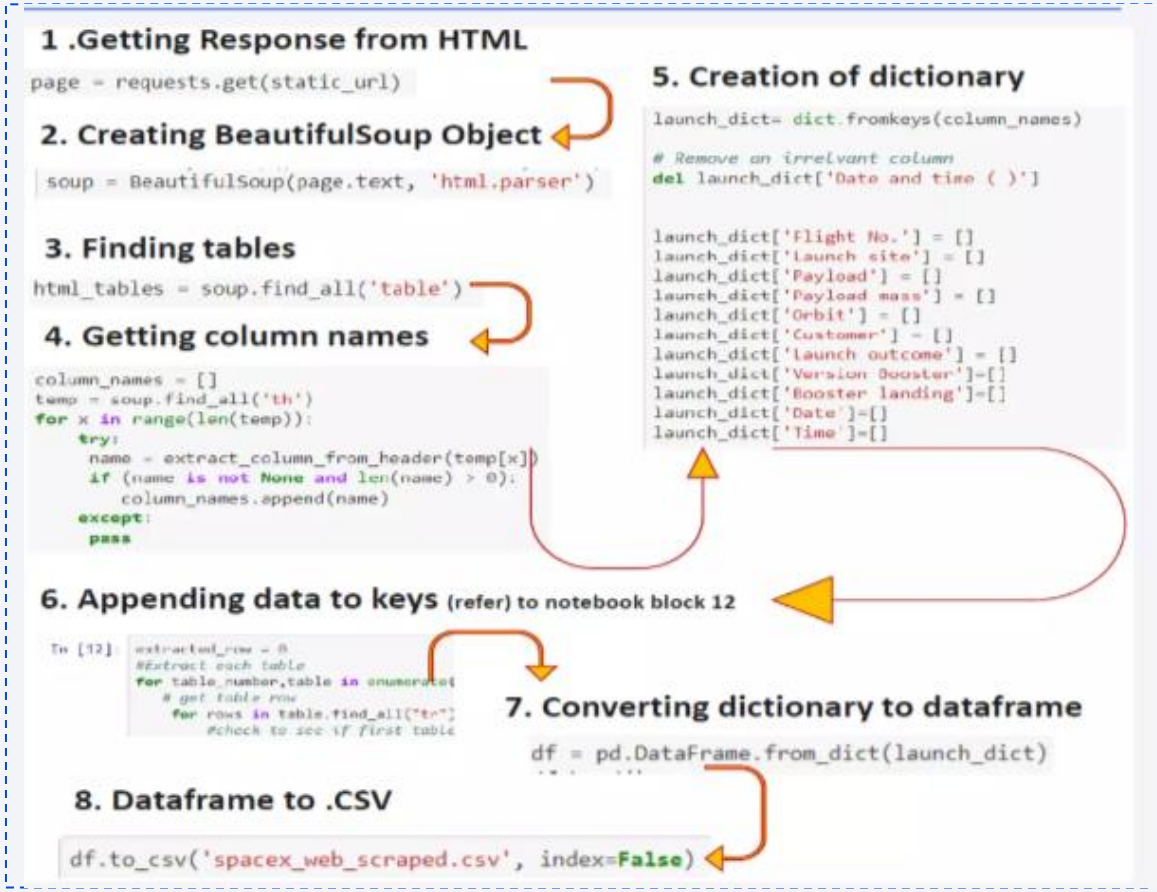
- I used the get request to the SpaceX API to collect and clean the request data
- I extract a Falcon launch records HTML table from Wikipedia and extract all columns variable name from html table.
- I convert data to data frame by using pandas library.
- The notebook github https://github.com/saeedyskasi/IBM-data-science-capstone-project -



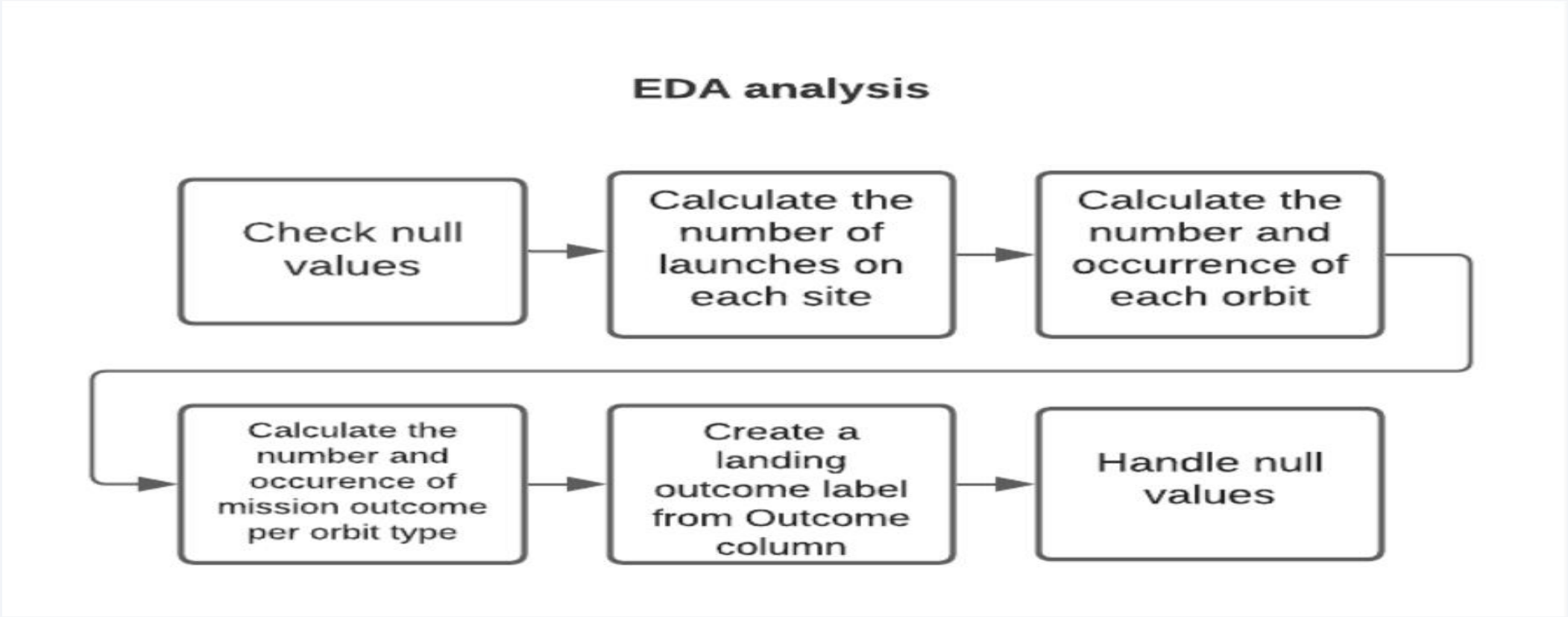
Data Collection - Scraping

- Web scraping from wikipidia

- The link
- https://github.com/saeedyskasi/IBM-data-science-capstone-project



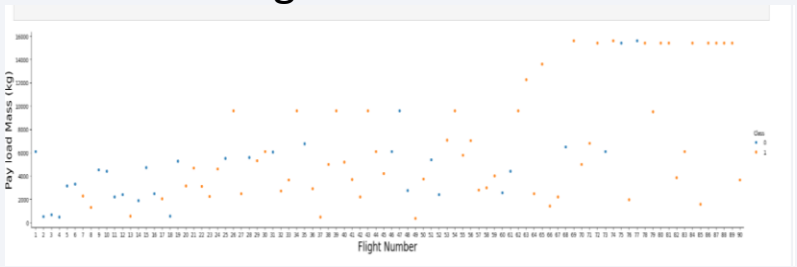
Data Wrangling



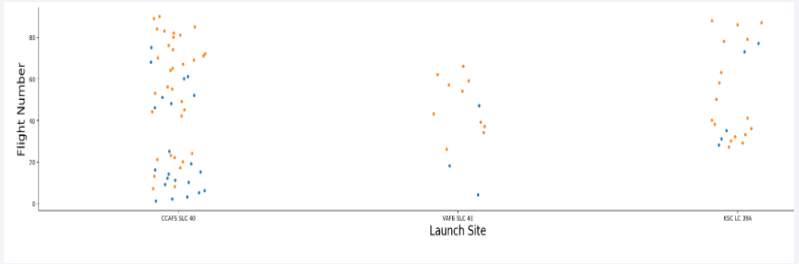
https://github.com/saeedyskasi/IBM-data-science-capstone-project/blob/main/5-lab-spacex-datavis.ipynb

EDA with Data Visualization

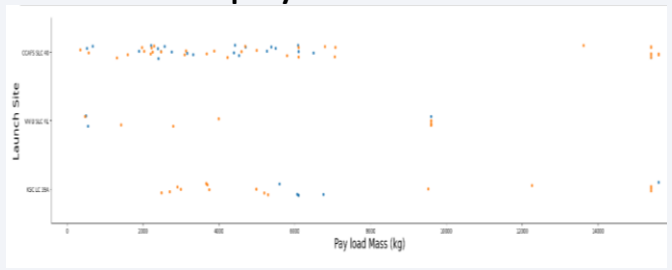
flight number



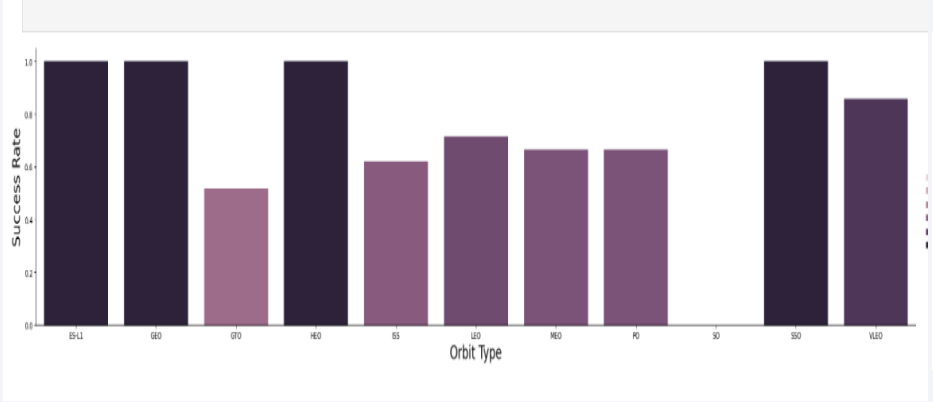
load site



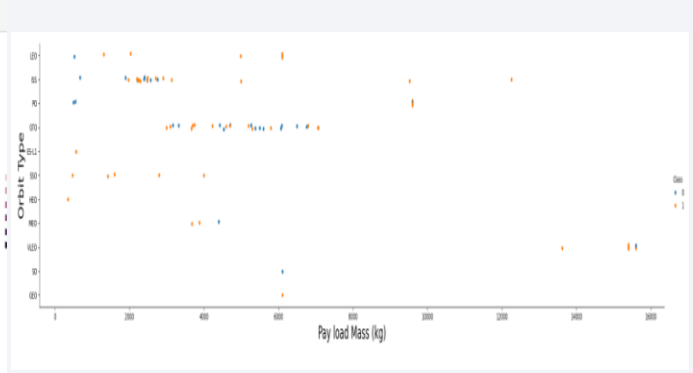
payload mas



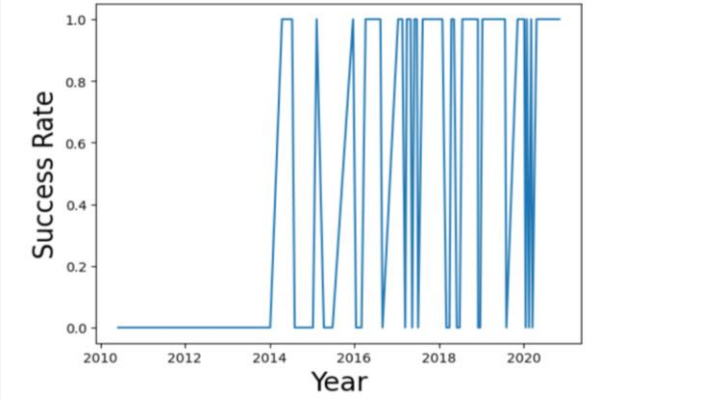
orbit



payload



year



https://github.com/saeedyskasi/IBM-data-science-capstone-project/blob/main/5-lab-spacex-datavis.ipynb

EDA with SQL

•Using bullet point format, summarize the SQL queries you performed

•Displaying the names of the unique launch sites in the space mission

• Displaying 5 records where launch sites begin with the string 'KSC'

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

• Displaying average payload mass carried by booster version F9 v1.1

• Listing the date where the successful landing outcome in drone ship was achieved.

• Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000

• Listing the total number of successful and failure mission outcomes

• Listing the names of the booster_versions which have carried the maximum payload mass.

• Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

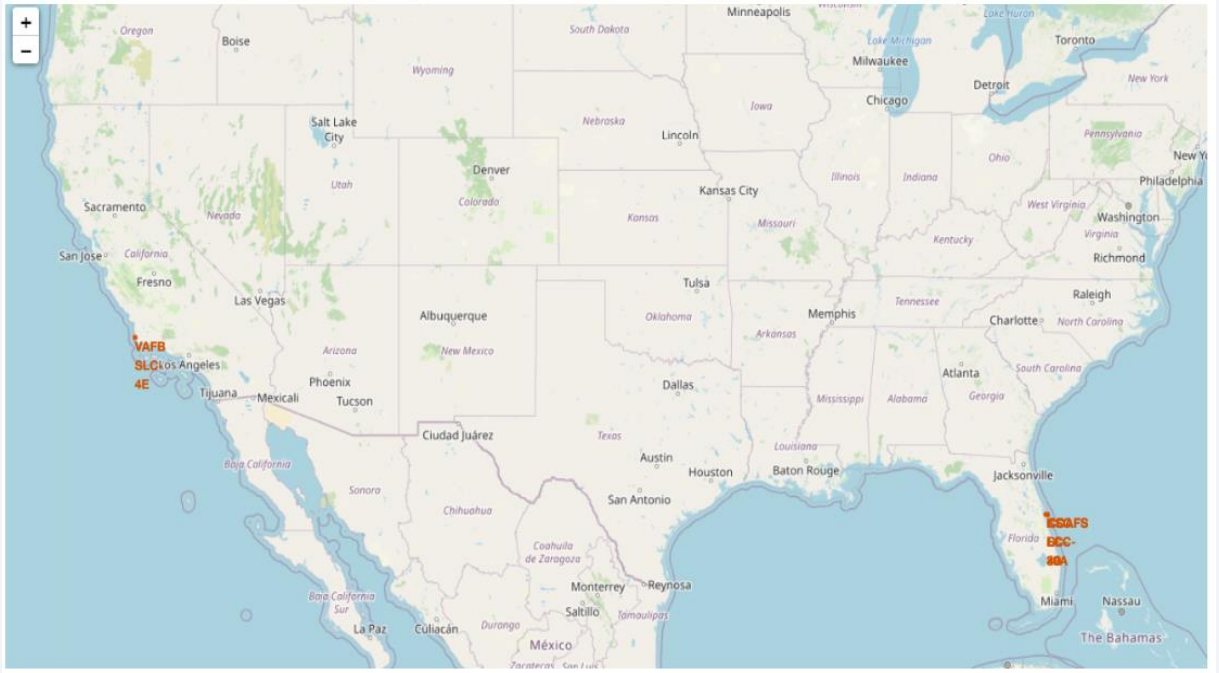
• Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

- https://github.com/saeedyskasi/IBM-data-science-capstone-project/blob/main/4-labs-spacex-EDA-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- I added markers for the aim of finding

an optimal location for building a launch site



- The link

Build a Dashboard with Plotly Dash

- I built an Interactive dashboard with plotly and dash
- I plotted pie charts showing the total launches by certain sites
- I plotted scatter plot showing the correlation between outcome and Payload Mass with different booster versions.

The link github

Predictive Analysis (Classification)

•I loaded the data using numpy and pandas.

•Transform the data.

• Split our data into training and testing.

• I built different machine learning models and tune different hyperparameters using GridSearchCV.

• I used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.

• I found the best performing classification model.

The link github

Results

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

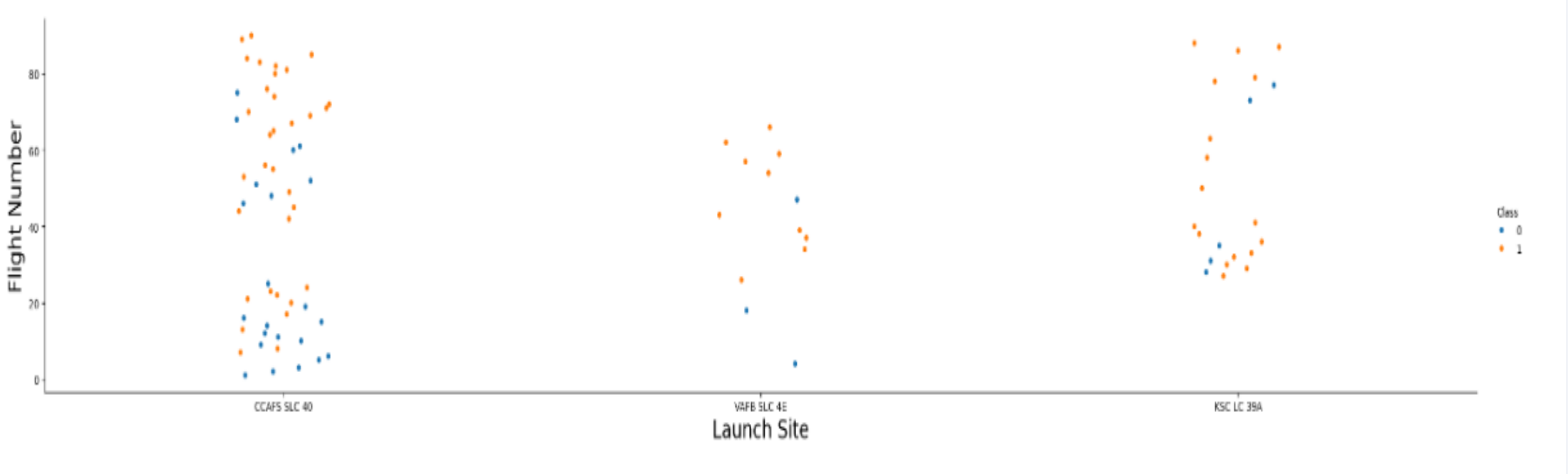
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Flight Number vs. Launch Site

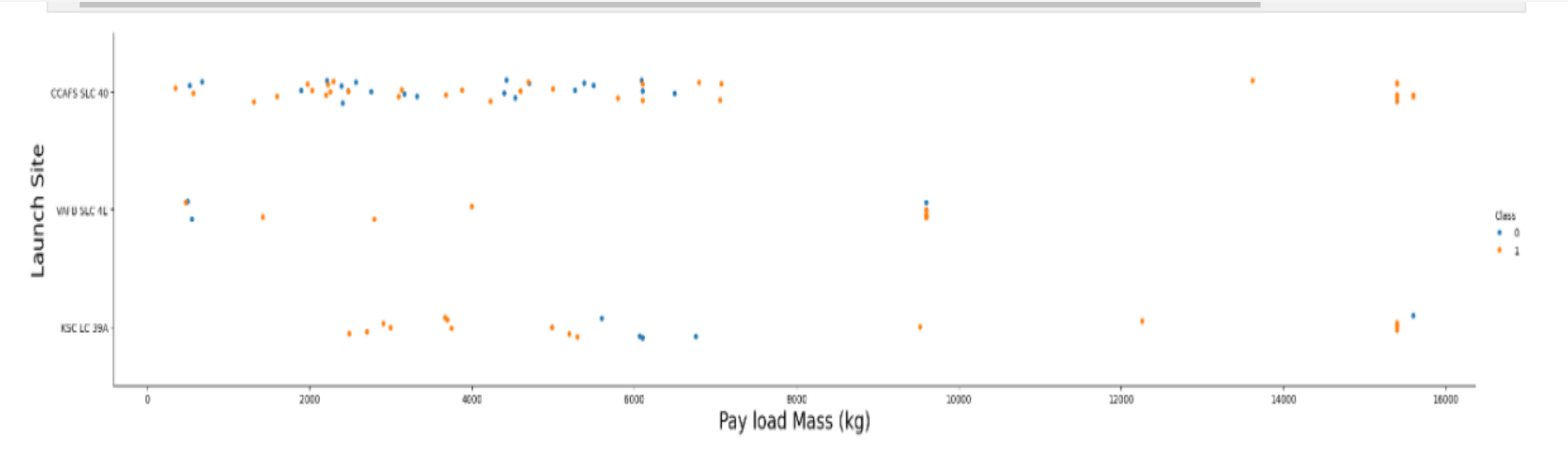
"The plot indicates that a higher number of flights at a launch site corresponds to a greater success rate at that site".



Payload vs. Launch Site

- scatter plot of Payload vs. Launch Site

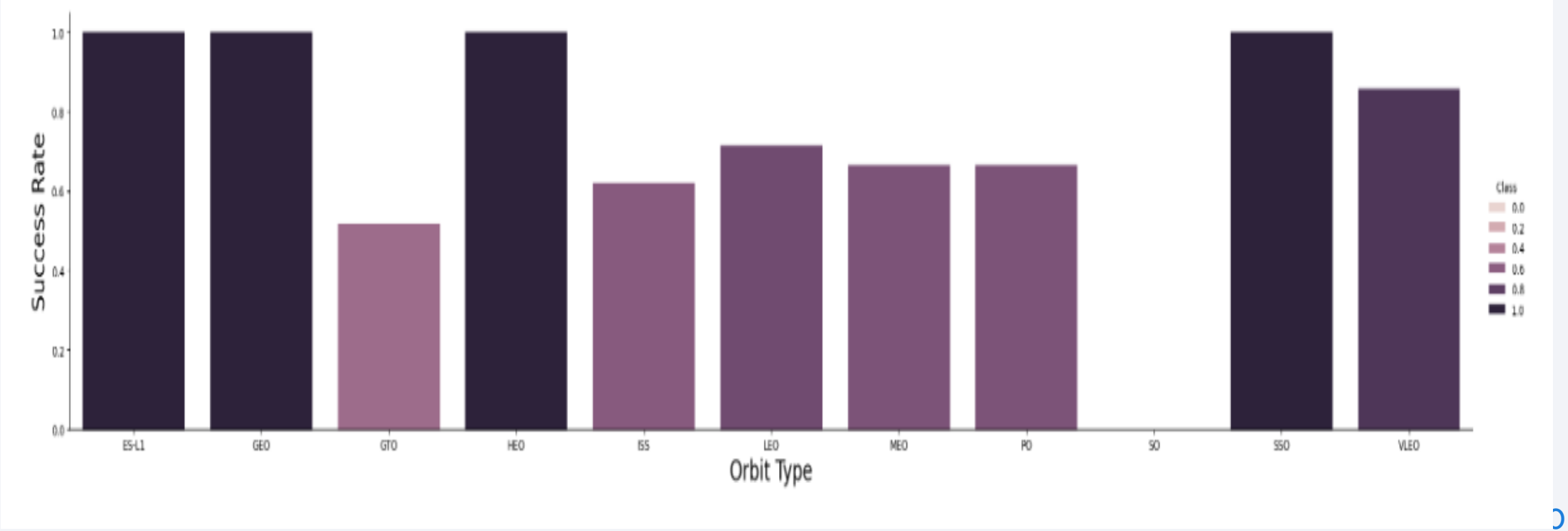
"We found that more flights at a launch site lead to a higher success rate at that site."



Success Rate vs. Orbit Type

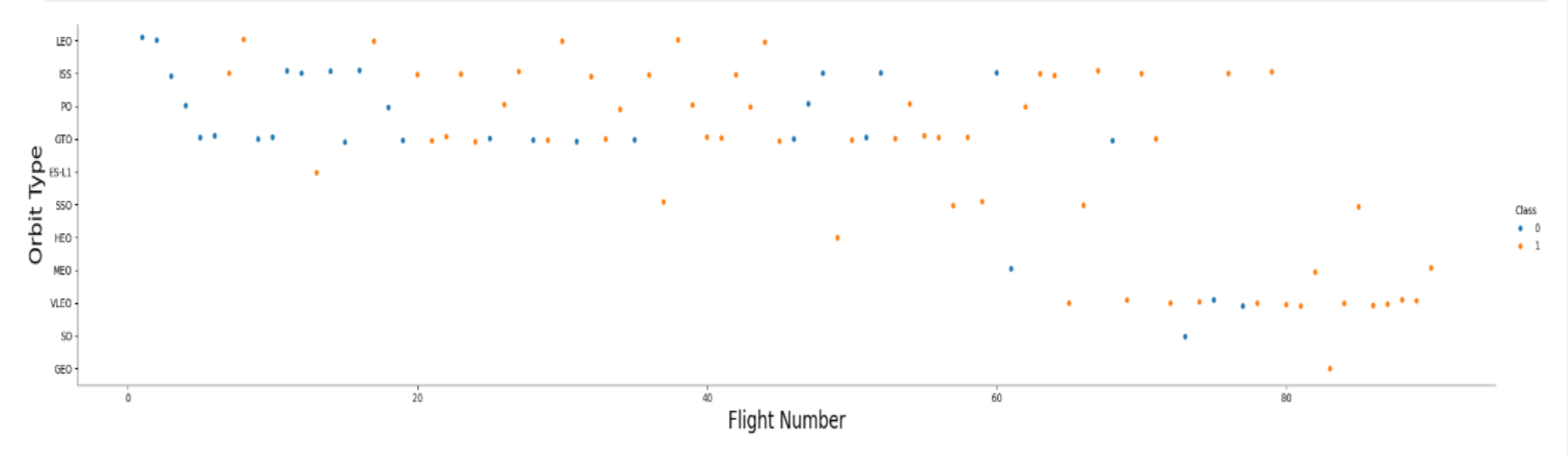
- bar chart for the success rate of each orbit type

"we see these Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate"



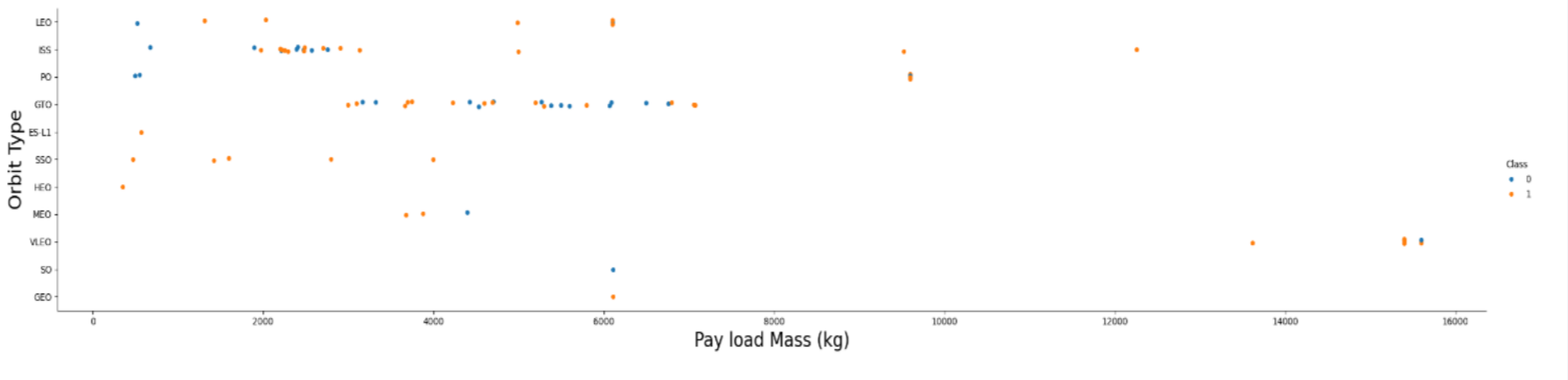
Flight Number vs. Orbit Type

“ in this below figure we observe the LEO orbit, success is related to the number of flights , so, in the GTO orbit there is no related between the flight number and the orbit ”



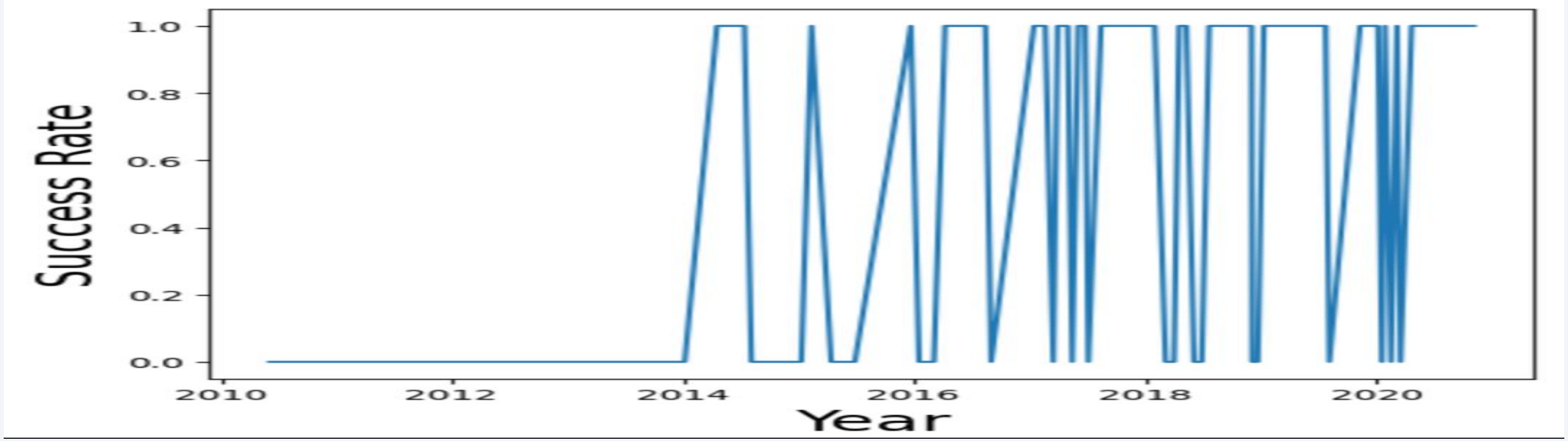
Payload vs. Orbit Type

- In this blew figure we can see that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



Launch Success Yearly Trend

- In this below figure we can observe that success rate since 2013 kept on increasing till 2020



SQL query

select distinct "LAUNCH_SITE" from SPACEXTBL ;

- Explain the query

We use distinct to show unique value in lanchsite from space xtbl.

Out[20]:

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

Launch Site Names Begin with 'CCA'

The query:

```
sql select * from SPACEXTBL where "LAUNCH_SITE" like 'CCA%' limit 5 ;
```

Explanation the query:
This query display 5 records where the launch site begin with ‘CCA’

21]:

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
2013-03-01	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 [22]: %sql select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where Customer like 'NASA%';
 * sqlite:///my_data1.db
Done.
Out[22]: sum(PAYLOAD_MASS__KG_)
                          99980
```

Query explanation :

This query calculate the total payload carried by boosters from NASA as 99980

Average Payload Mass by F9 v1.1

Task 4

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

In [141…

```
%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" like 'F9 v1.1%';
```

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

Out[141…

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

Explanation the query:

This query show average payload mass carried by booster version F9 v1.1 as 2928.4

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

In [142...

```
%sql select MIN(Date) from SPACEXTBL where "Landing_Outcome" = 'Success (ground pad)';
```

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

Out[142...

MIN(Date)
2015-12-22

Explanation the query:

This query shows that the date of the first successful landing outcome on ground pad was 22nd December 2015

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 [25]:
```
%sql select * from SPACEXTBL where "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_MASS__KG_" between 4000 and 6000;
```

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

Out[25]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Explanation query:
This query explain when landing outcome column== success (drone ship) and palyload_mass_kg between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [838...

```
%sql select "Mission_Outcome", COUNT(*) as total_number from SPACEXTBL group by "Mission_Outcome";
```

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

Out[838...

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

The query is:

%sql select "Mission_Outcome", COUNT(*) as total_number from SPACEXTBL group by "Mission_Outcome ؛"

Explanation :

This query calculate total no of mission outcome failed and success group by mission outcome.

Boosters Carried Maximum Payload

Task 8

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

```
In [839…
%sql select "Booster_Version" from SPACEXTBL where "PAYLOAD_MASS__KG_" = (select MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
Out[839…
```

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

The query:

%sql select "Booster_Version" from SPACEXTBL where "PAYLOAD_MASS__KG_" = (select MAX("PAYLOAD_MASS__KG_") from SPACEXTBL)

Explain the query :

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function

2015 Launch Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [840…

```
%sql SELECT substr(Date, 6, 2) AS Month,Date,Booster_Version,Launch_Site,"Landing_Outcome" FROM SPACEXTBL WHERE "Landing_Ou
```

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

Out[840…

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

The query:

```
%sql SELECT substr(Date, 6, 2) AS
Month,Date,Booster_Version,Launch_Site,"Landing_Outcome" FROM SPACEXTBL WHERE
"Landing_Outcome" = 'Failure (drone ship)' AND substr(Date, 1, 4) = '2015';
```

Explain the query:

The used combinations of the WHERE AND , between the date conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

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.

```
%sql SELECT (LANDING_OUTCOME) as 'count' FROM SPACEXTBL WHERE "DATE" BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY "DATE" D
```

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

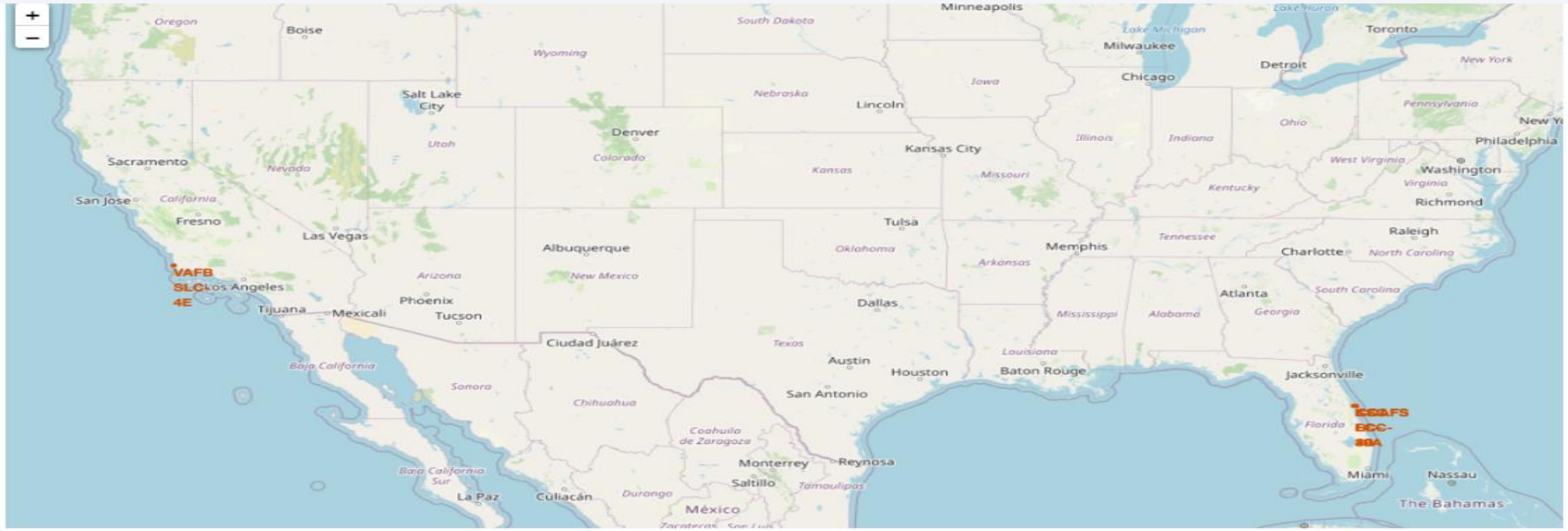
count
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt

- The query:
-
```
%sql SELECT (LANDING_OUTCOME) as 'count' FROM SPACEXTBL WHERE "DATE" BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY "DATE" DESC;
```

- Explanation the query:
- selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20 We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending orde

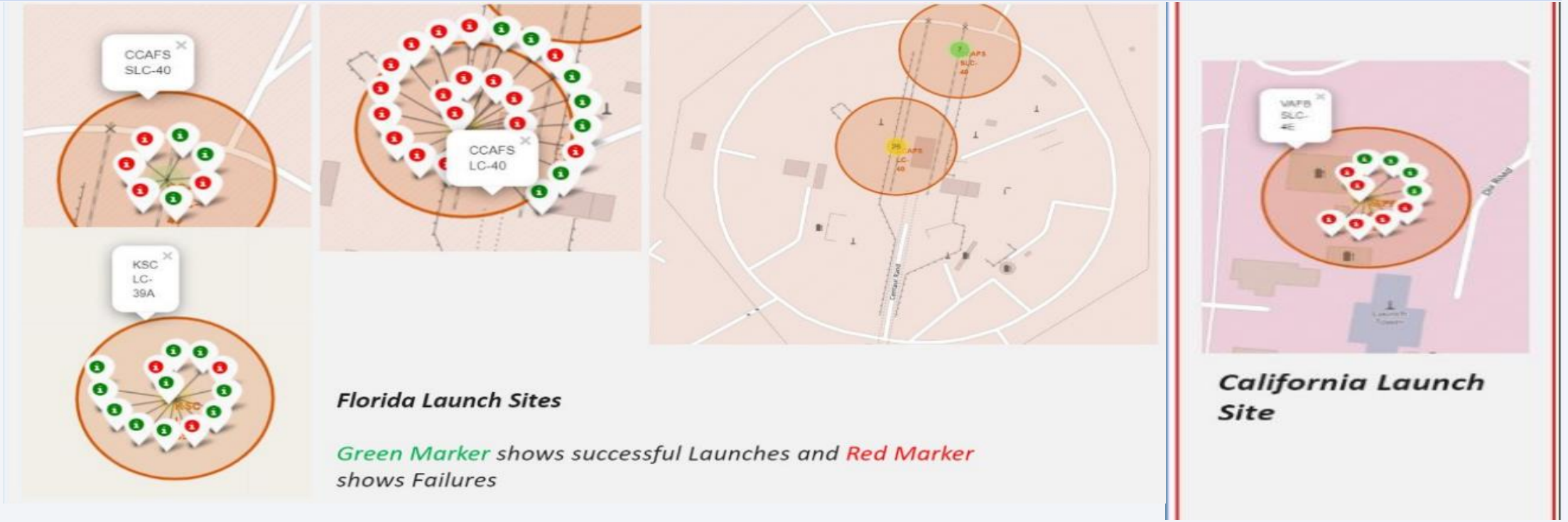

Section 3
Launch Sites
Proximities Analysis

All launch sites global map markers

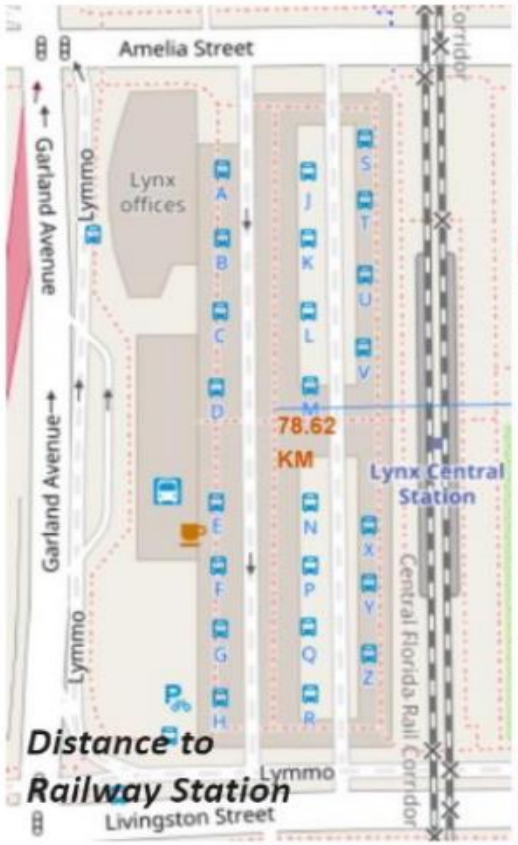


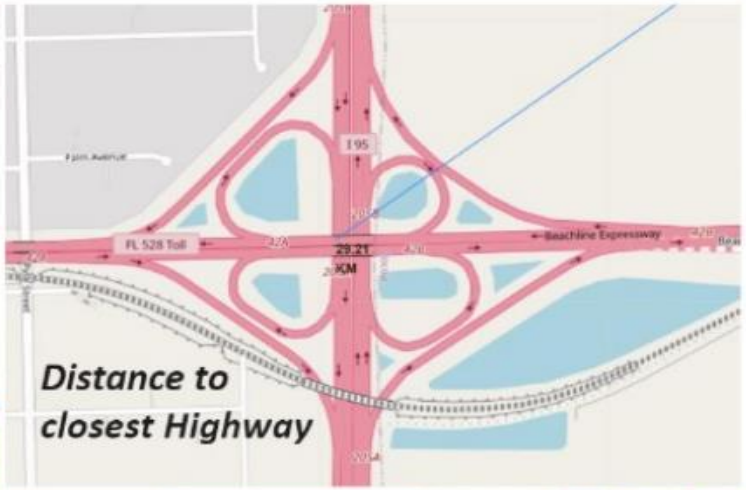
We see the SpaceX launch site location in two states in The USA Florida and California.

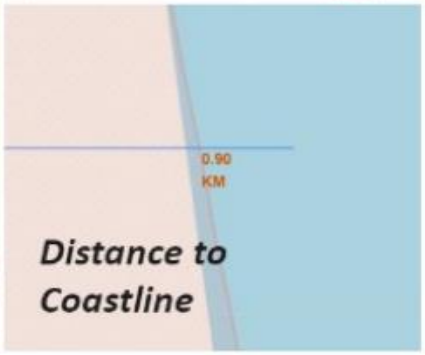
Markers showing launch sites with color labels



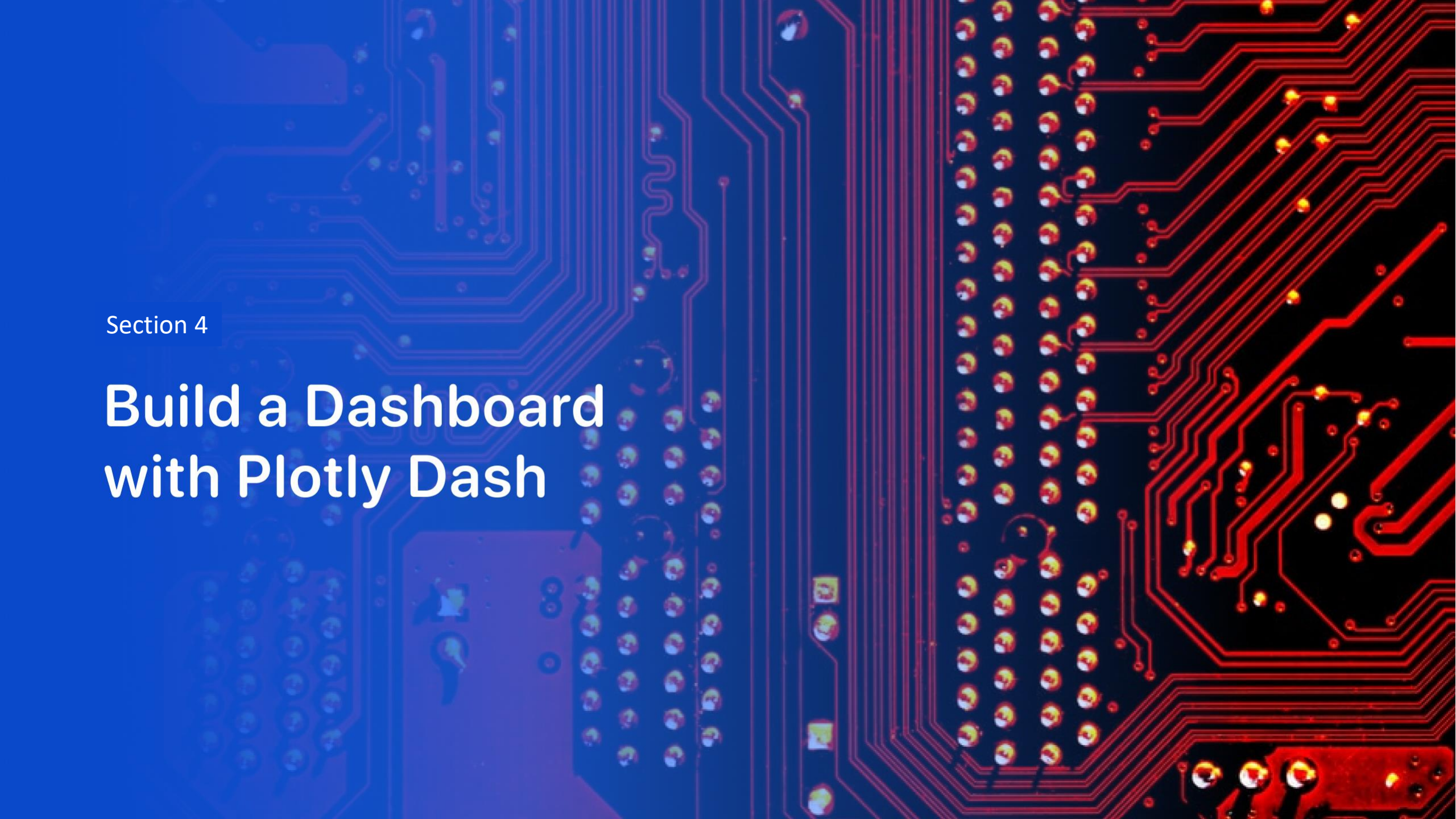
Launch Site distance to landmarks



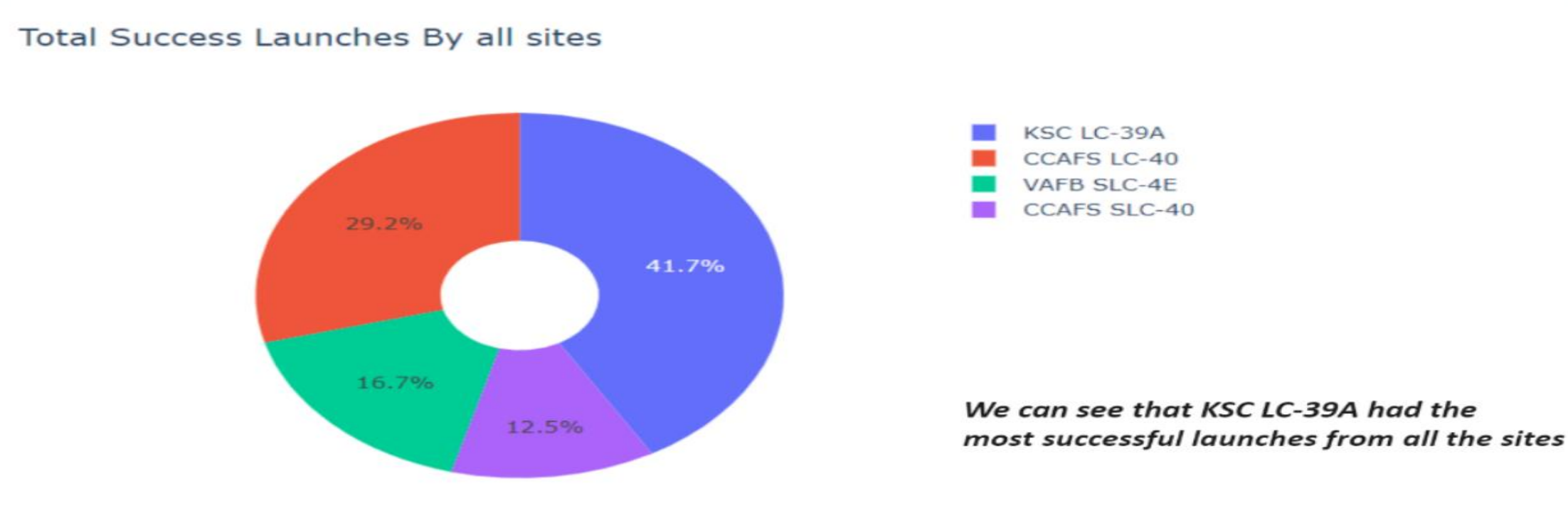




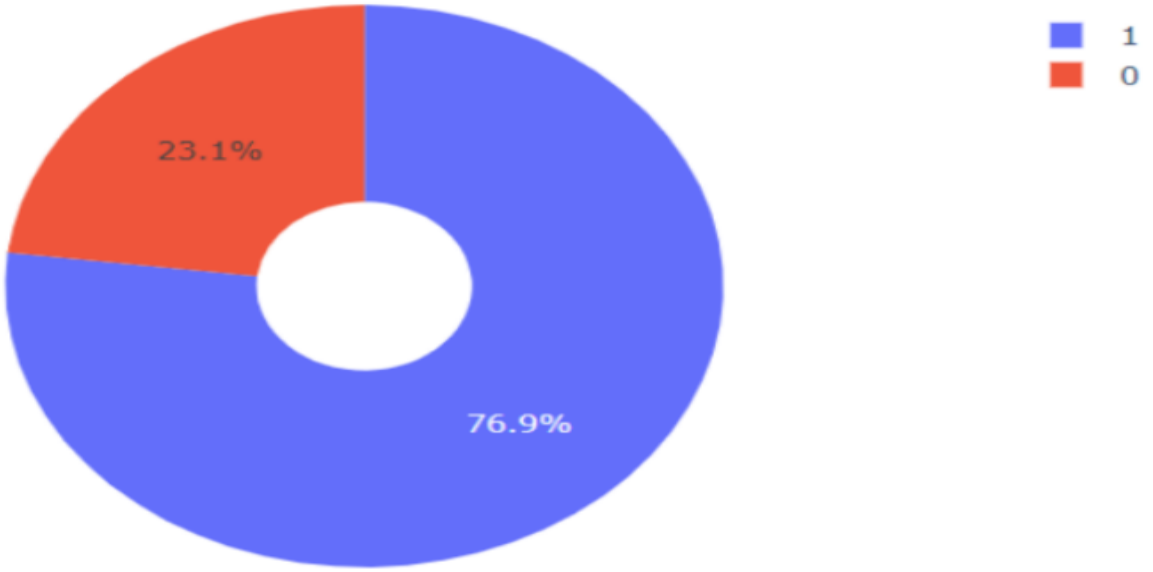
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



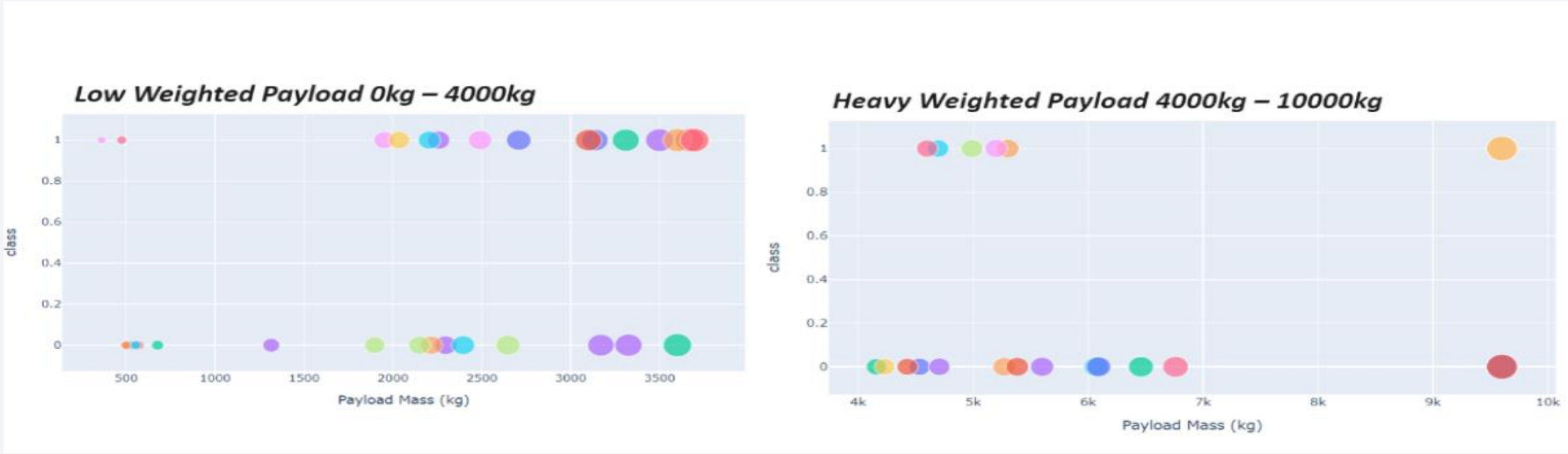
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

We can see the success rate for low weighted payloads is higher than the heavy weighted payload



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classifier is the model with the highest classification accuracy.

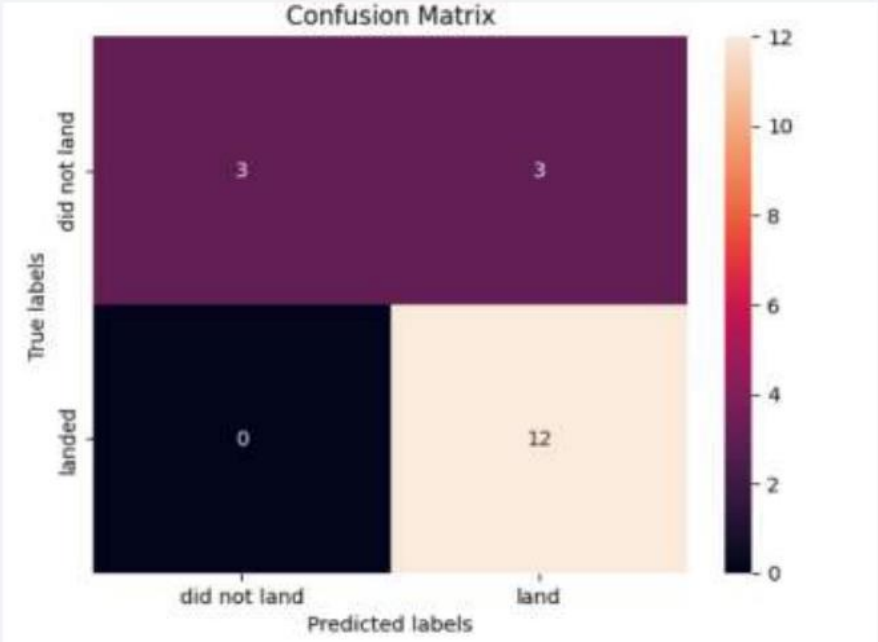
```
models = {'KNeighbors':knn_cv.best_score_,
          'DecisionTree':tree_cv.best_score_,
          'LogisticRegression':logreg_cv.best_score_,
          'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- The 4 models has the same confusion matrix as they has the same accuracy test percentage the main problem of this models is false positivity



Conclusions

In this project we can conclude that:

- The KNN model is the best in terms of prediction accuracy for this dataset.
- Low weighted payloads perform better than the heavier payloads.
- Launch success rate started to increase in 2013 till 2020.
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit (GEO,HEO,SSO,ES L1) had the best Success Rate.

Thank you!