Winning Space Race with Data Science

Faizan Farooqui
October 5th 2023

Outline

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

Executive Summary

- The Purpose of this study was to identify the factors associated with saving money and being efficient in a successful space program.
- Summary of Methodologies:
  - Data was **Collected** from SpaceX REST Api and web scrapping, then was **Wrangled** to have a clean dataset.
  - Data was **Explored, Visualized** and **Analyzed** using various techniques to understand the problem at hand.
  - Later the data was trained through **Models like logistic regression, SVM, decision tree and K nearest neighbor** to predict the best factors to be used for successful landing.

- Summary of All Results:
  - Launch success improved overtime.
  - KSC LC 39A landing site have the highest success rate, similarly Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.
  - Decision Tree model was just only slightly better than other models in predicting the outcome for landing.

Introduction

Background:

- SpaceX is leading the space expedition and compared to their competitors they are very inexpensive, costing $62 million per launch.
- That's because of their reuse of the first stage rocket.
- SpaceY wants to compete with SpaceX in space exploration, by determining whether the first stage rocket will be reused.

Problem:

- SpaceY wants to understand the factors that are responsible for a successful launch.
- Payload mass, launch sites, number of flights and orbits.
- We want to predict what factors will allow us to have a successful first-stage landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using the SpaceX REST API and web scraping techniques
- Perform data wrangling
  - By filtering the data, handling missing values and applying one hot coding to prepare 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
  - We use sklearn to get the models, then transform, split data to be run through the predictive models.

Data Collection

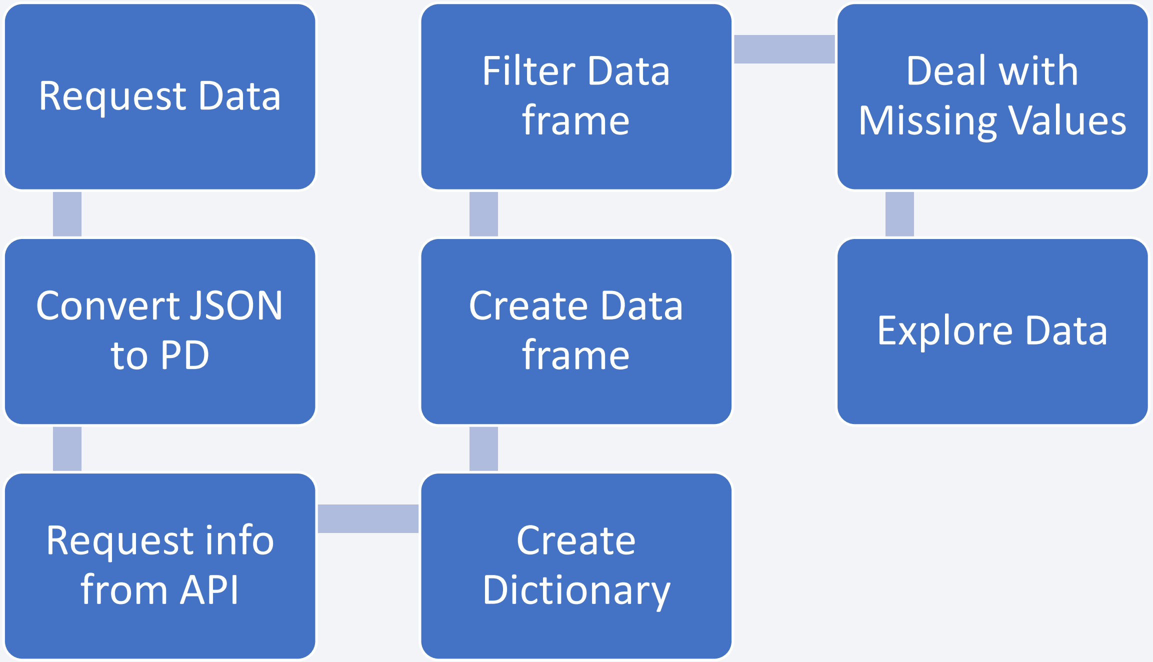
- Data was collected through a couple of processes.
- SpaceX API was used to collect some data by through jupyter notebook
- More data was collected through web scrapping Wikipedia page for SpaceX.
- In both instances a request was sent through the notebook to get the data.
- It was then converted into a pandas dataframe.
- Details for each process are in the following slides.

Data Collection – SpaceX API

- Data was collected as shown here.

- GitHub URL

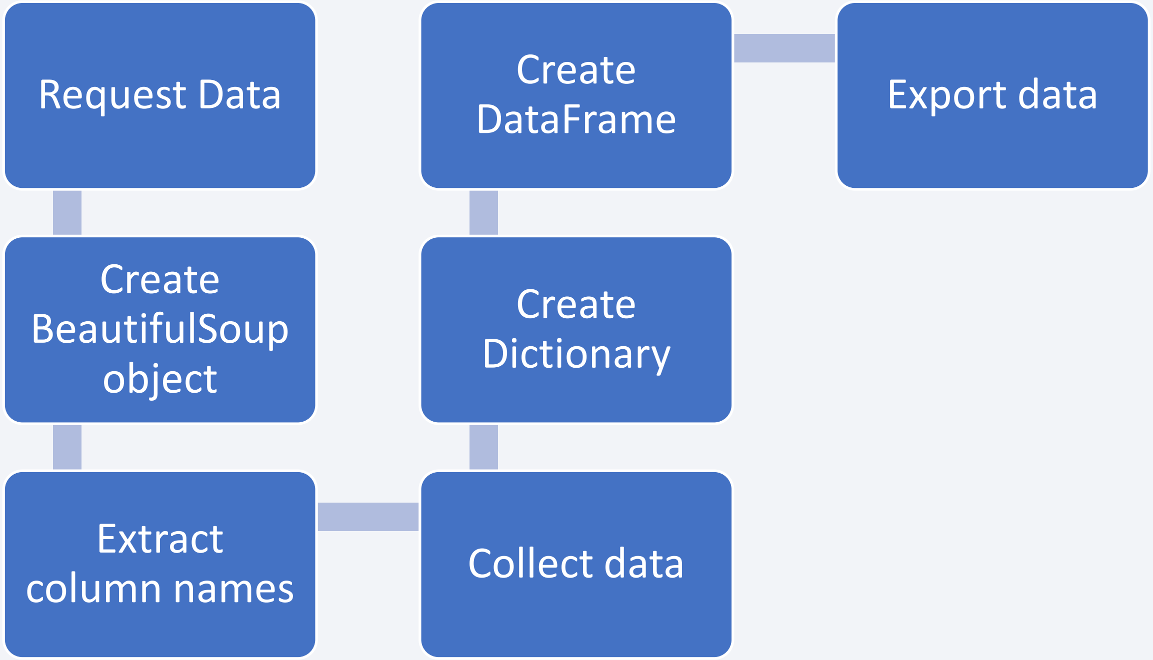
  https://github.com/faizan152/Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web Scrapping was done as shown here
- GitHub URL:

https://github.com/faizan152/Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Performed visualization techniques to understand data
- Create binary for landing outcome as 1 and 0
- 1 means success, and 0 means it doesn't land.
- Which launch sites, orbits and payloads were successful
- GitHub URL
  - ://github.com/faizan152/Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Following Charts were made for visualizing data
  - Flight number vs payload
  - Flight number vs Launch site
  - Payload Mass (kg) vs Launch site
  - Payload Mass (kg) vs Orbit type
  - Success rate vs Yearly trend

- GitHub URL:
  - ://github.com/faizan152/Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite%20(1).ipynb

EDA with SQL

- Queries were made to obtain the following
  - Names of Launch sites, payload mass, and average payload mass
  - Date of first successful mission
  - Names of boosters with success rate and with maximum payload.
  - Failed and unsuccessful missions and count of landing

- GitHub URL:
  - ://github.com/faizan152/Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Blue circle was added at NASA, Red circles were added to all launch sites.
- Lines were drawn from the launch site with distance to the coast, nearest city, highway and railroad.
- These markers and objects were added to make map more interactive and also to show that launch sites were close to coast and away from cities.
- GitHub URL:
  - .https://github.com/faizan152/Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Plotly Dash was created with a dropdown menu for all launch sites.
- Pie chart to show the Successful launch from each site.
- Scatter chart showing Payload mass vs Success rate by booster version with a slider for payload mass range.
- These were added so that data can be presented in more presentable way to the Client.
- GitHub URL:
    - https://github.com/faizan152/Capstone-Project/blob/main/dash_interactivity.py

Predictive Analysis (Classification)

- Create NumPy array from the Class column.
- Standardize and the Spit data into Test and train.
- A GridsearchCV was created with cv=10 and then it was applied to the following algorithms:
    - Logistic regression, support Vector machine(SVM), decision tree classifier and k nearest neighbor (KNN).
- Calculated Accuracy by score and then used confusion matrix for more accuracy.
- Identified the best model.
- GitHub URL:
    - https://github.com/faizan152/Capstone-Project/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

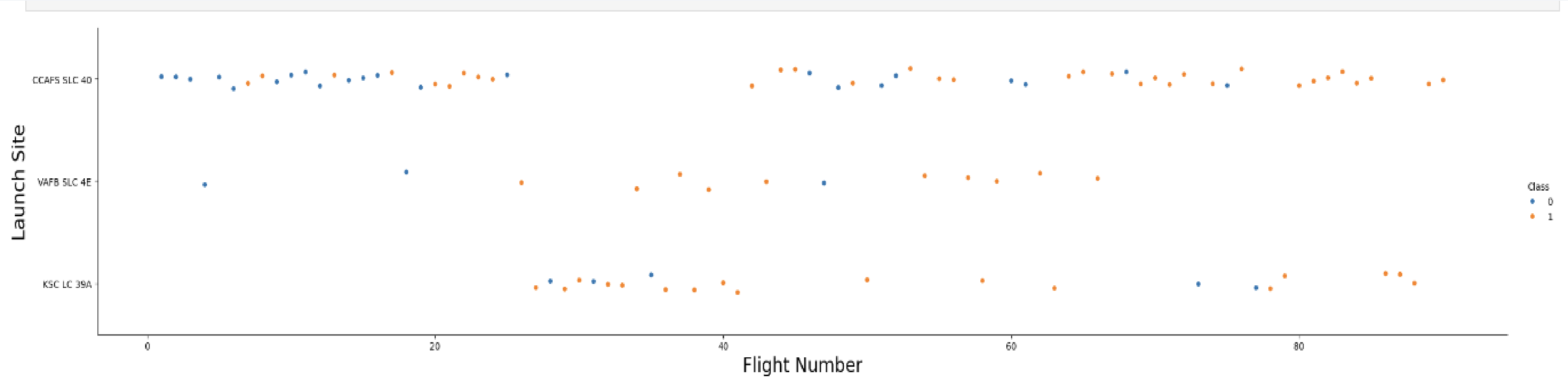
Results

- Exploratory data analysis results:
  - Launch Success increase overtime.
  - KSC LC39A site has the highest success rate.
  - GEO, HEO, SSO and ES-L1 orbits have a 100% Success rate.
- Interactive analytics demo in screenshots
  - In the following slides.
- Predictive analysis results
  - Decision tree classifier performed better amongst the models.

Section 2

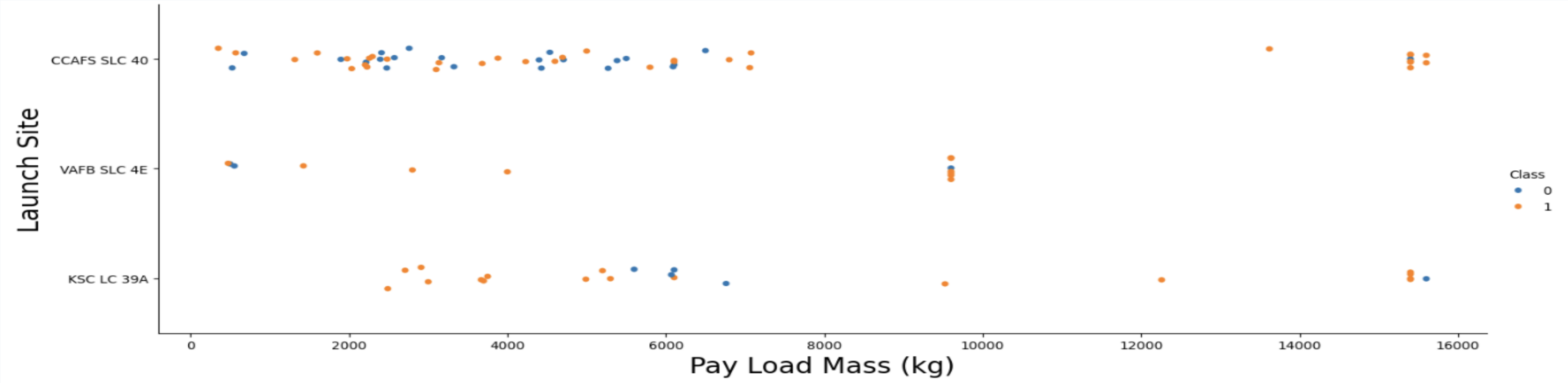
Insights drawn from EDA

Flight Number vs. Launch Site



- Blue dots represents unsuccessful missions and orange represents success.
- As the flight number increases, the success rate increases
- KSC LC 39A have a higher success rate then other sites.

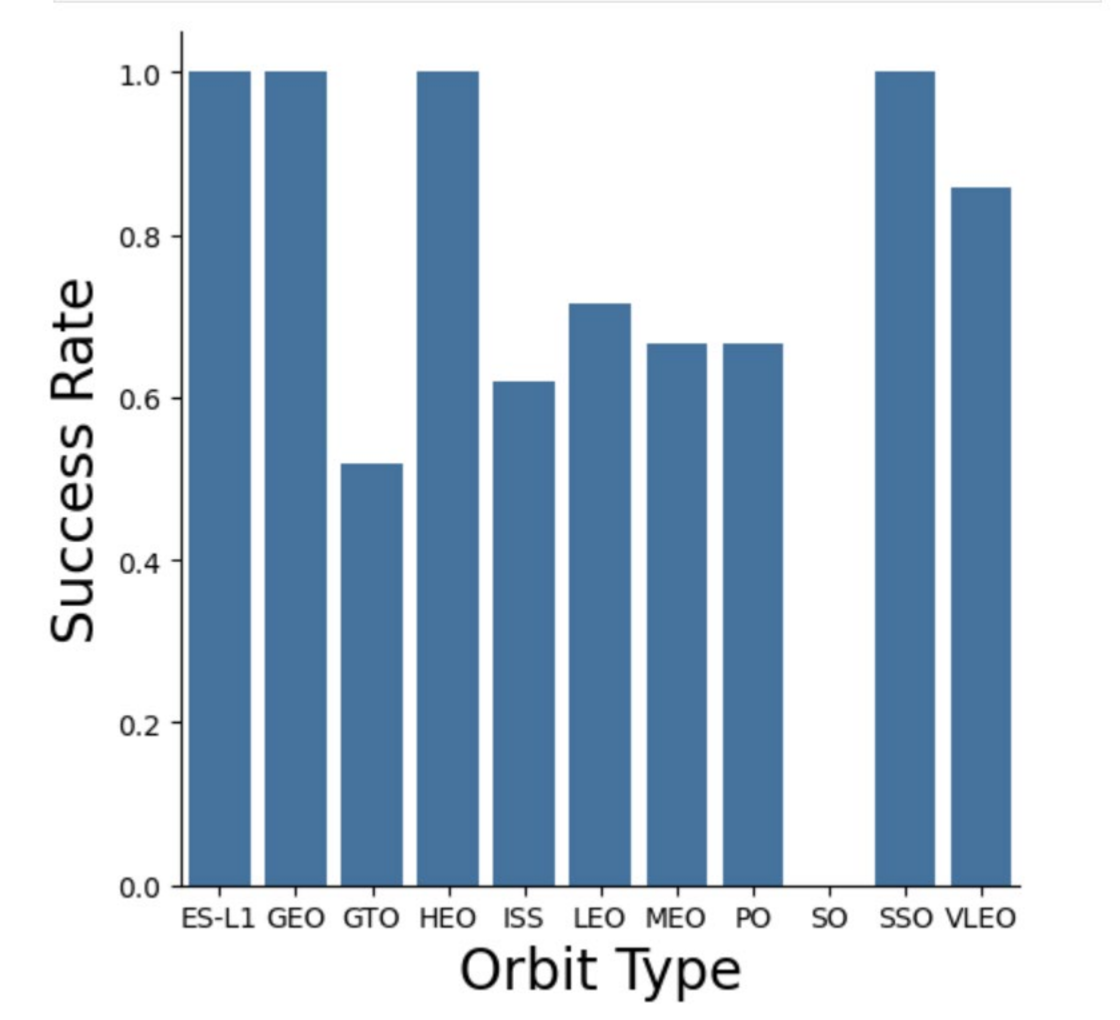
Payload vs. Launch Site



- Heavy payload have a better success rate.
- KSC LC 39A have a better success rate than others.

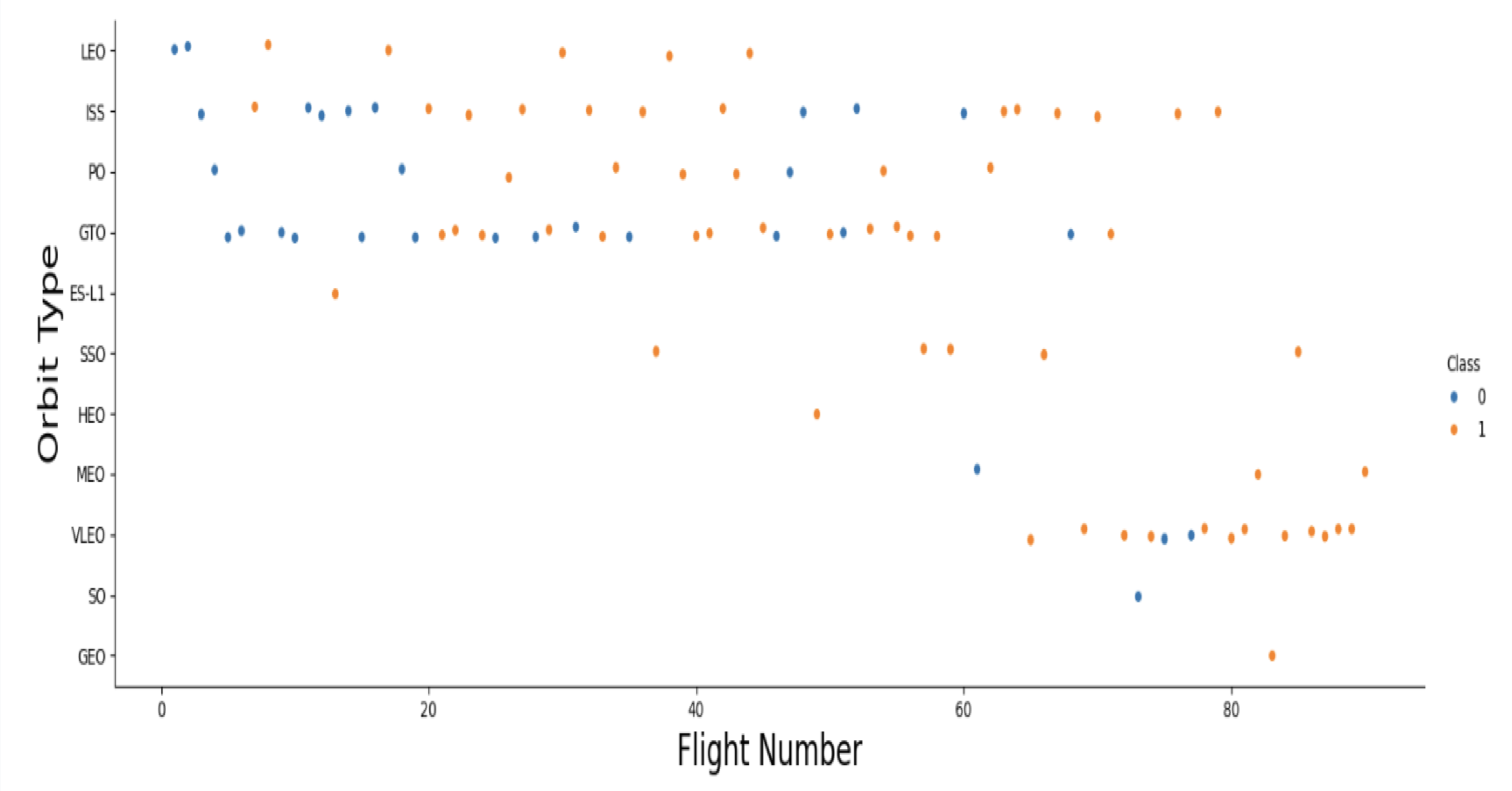
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have higher success rate.
- SO is zero.
- All other orbits are more or less 50 %
- Orbits with 100% success rate are recommended.



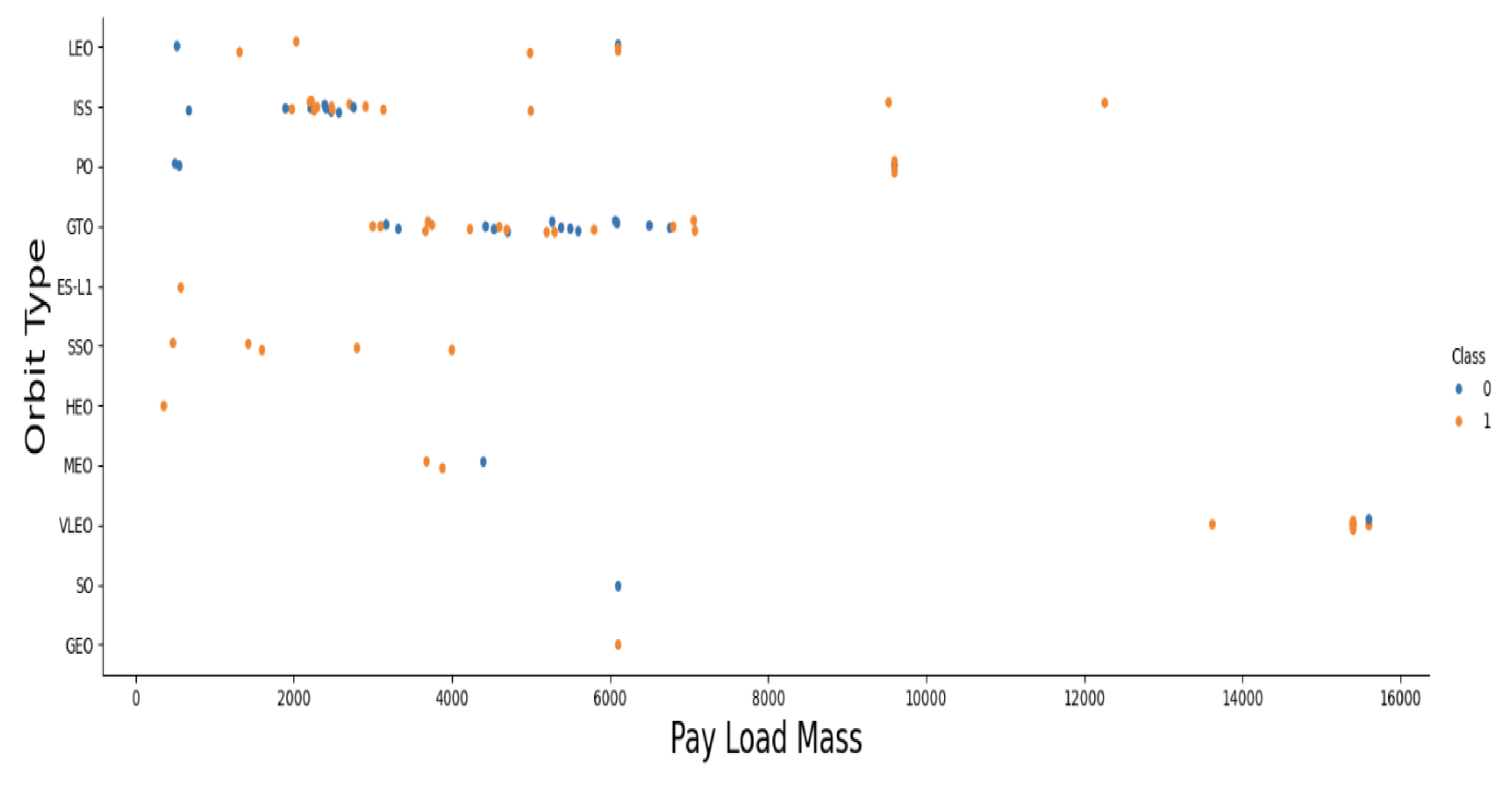
Flight Number vs. Orbit Type

- Success increased over time.
- Leo Orbit have all successful attempts after the first two



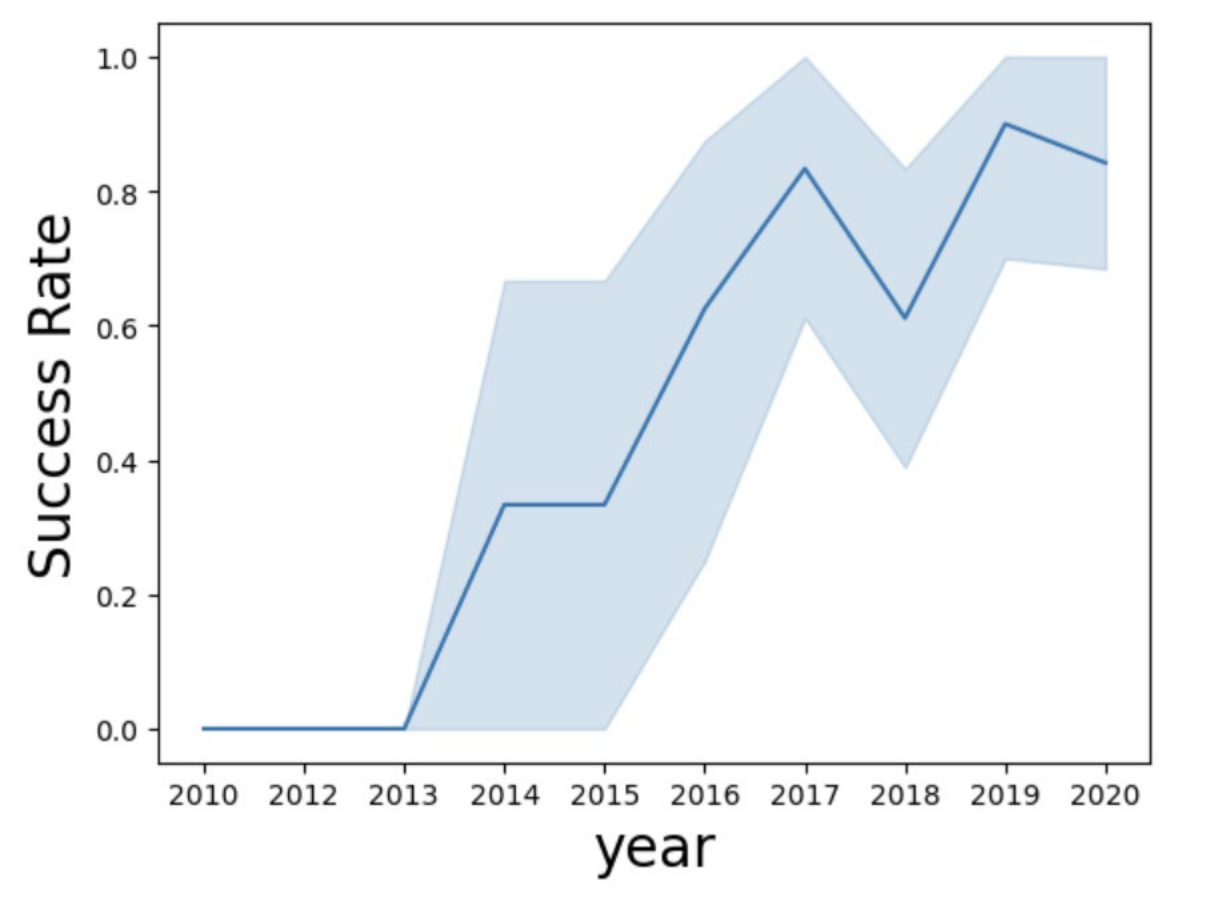
Payload vs. Orbit Type

- Heavy Payloads were good for VLEO, PO, ISS and LEO
- SSO did better with lighter payload.
- GTO orbit have mixed results.



Launch Success Yearly Trend

- Success rate increased after 2013.
- There was a dip from 2017-2018 but then there was a rise again.



All Launch Site Names

- List of all Launch Site were queried.

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE
```

```
 * 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'

- Find 5 records where launch sites begin with `CCA`
- 5 records are shown.

```
%sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' limit 5
```

```
 * 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

- Calculate the total payload carried by boosters from NASA
- Total mass was 45,596kg for the NASA booster.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS SUM FROM SPACEXTABLE WHERE CUSTOMER IS "NASA (CRS)"
```

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

SUM
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average weight was 2534.66kg for the booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE FROM SPACEXTABLE WHERE BOOSTER_VERSION LIKE 'F9 V1.1%'
```

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

AVERAGE
2534.6666666666665

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was 2010-04-06

```
%sql SELECT min(date) AS DATE FROM SPACEXTABLE WHERE MISSION_OUTCOME IS "Success"
```

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

DATE
2010-04-06

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE (Mission_Outcome IS "Success") AND  (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6
```

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

- The total number of successful and failure mission outcomes.

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTABLE GROUP by mission_outcome ORDER BY mission_outcome
```

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

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

Boosters Carried Maximum Payload

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

```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
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 of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select DATE as Month, Landing_Outcome, booster_version, launch_site from SPACEXTABLE where DATE like '2015%' AND Landir
```

```
* sqlite:///my_data1.db
)one.
```

Month	Landing_Outcome	Booster_Version	Launch_Site
2015-10-01	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

- Rank 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(*) as count from SPACEXTABLE where Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP by
```

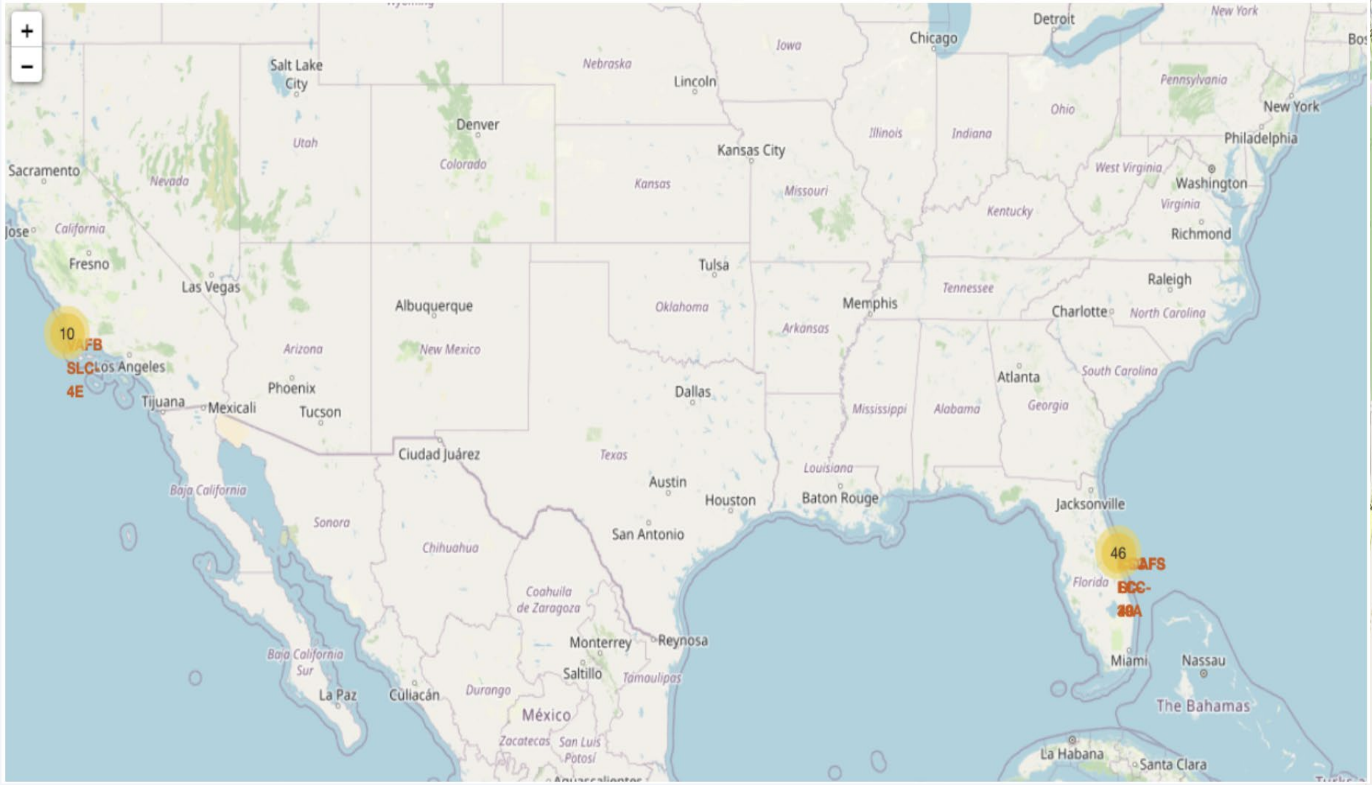
```
 * sqlite:///my_data1.db
Done.
```

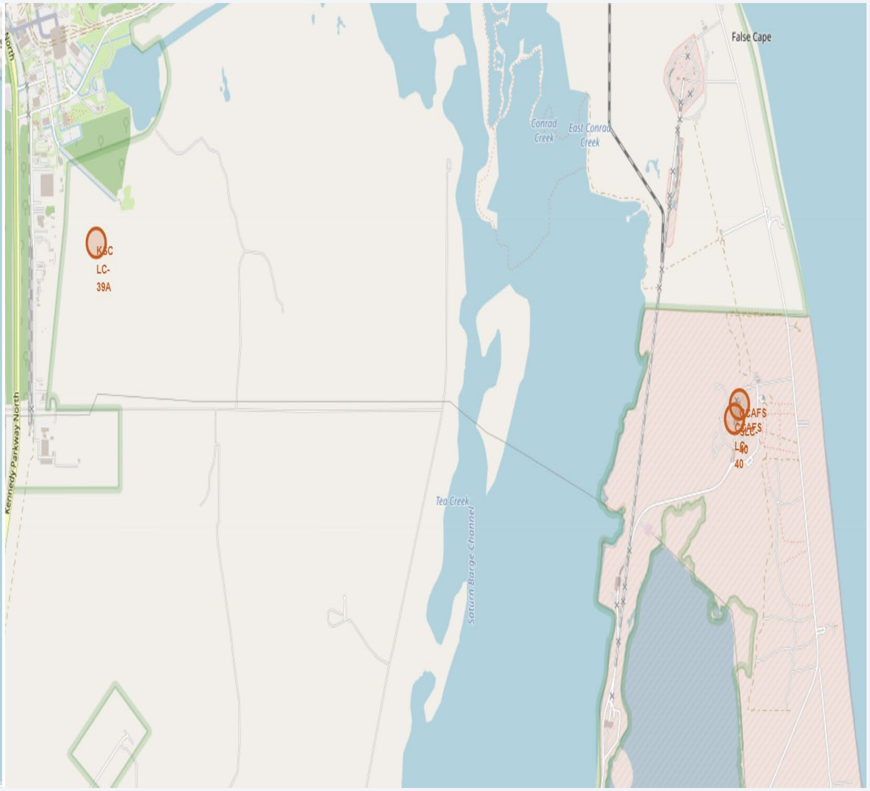
Landing_Outcome	count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

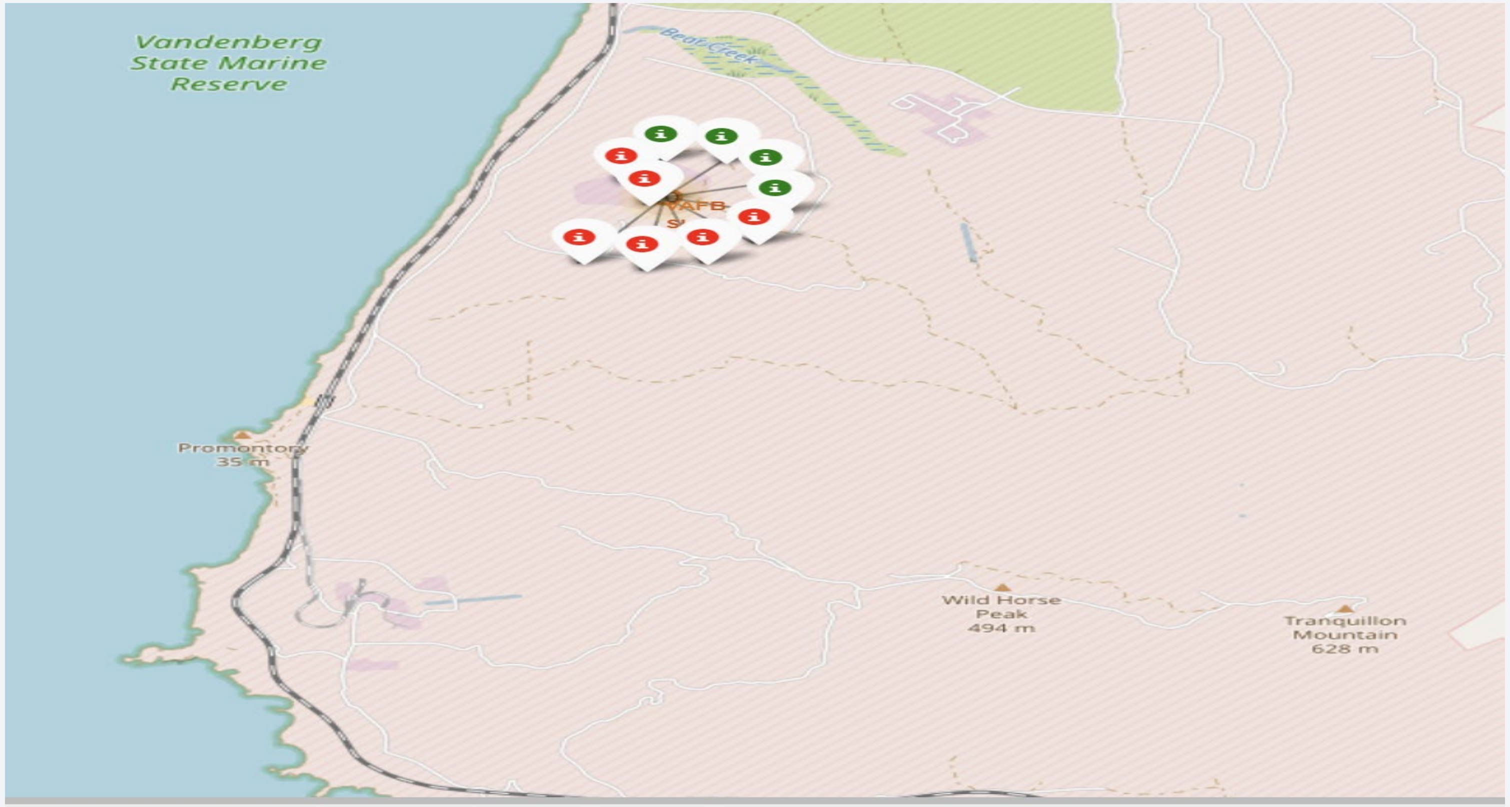
Launch Site Locations





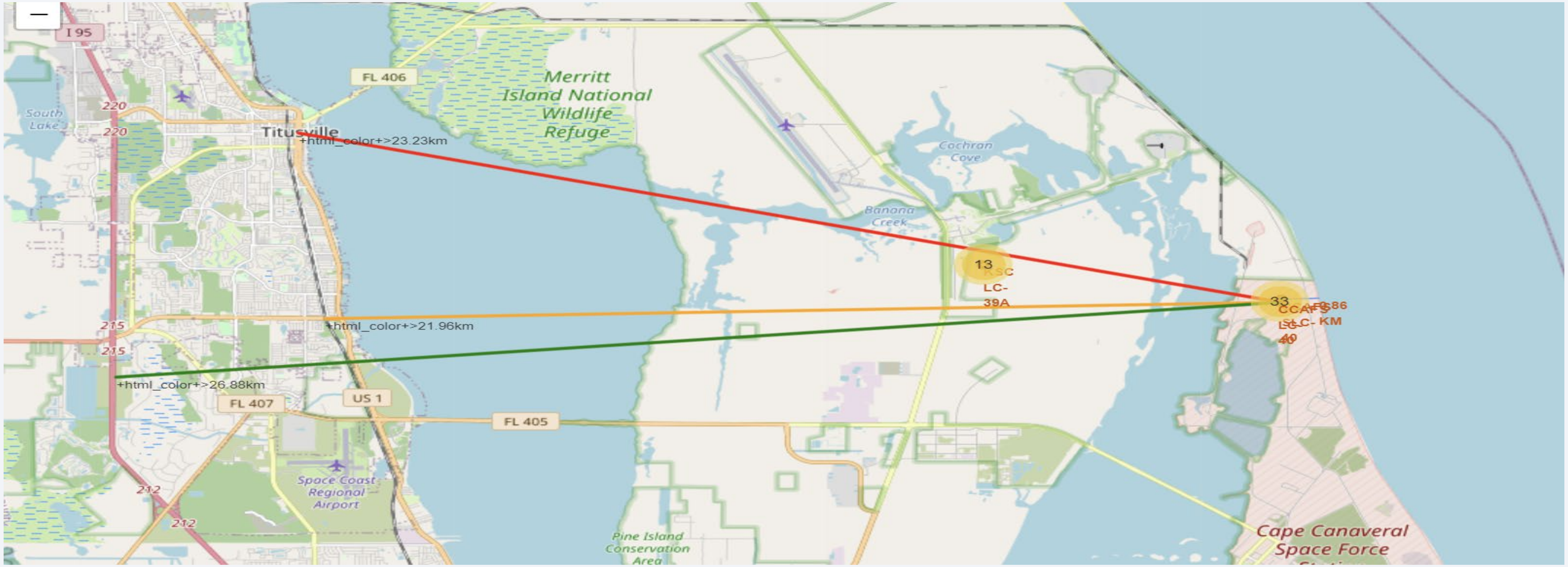
- The left shows all launch sites and the right shows close up of two sites in florida.

Color-Coded Launch Markers



- Markers showing successful and unsuccessful landing. (green= successful landing, red = failed landing)

Key location Proximities

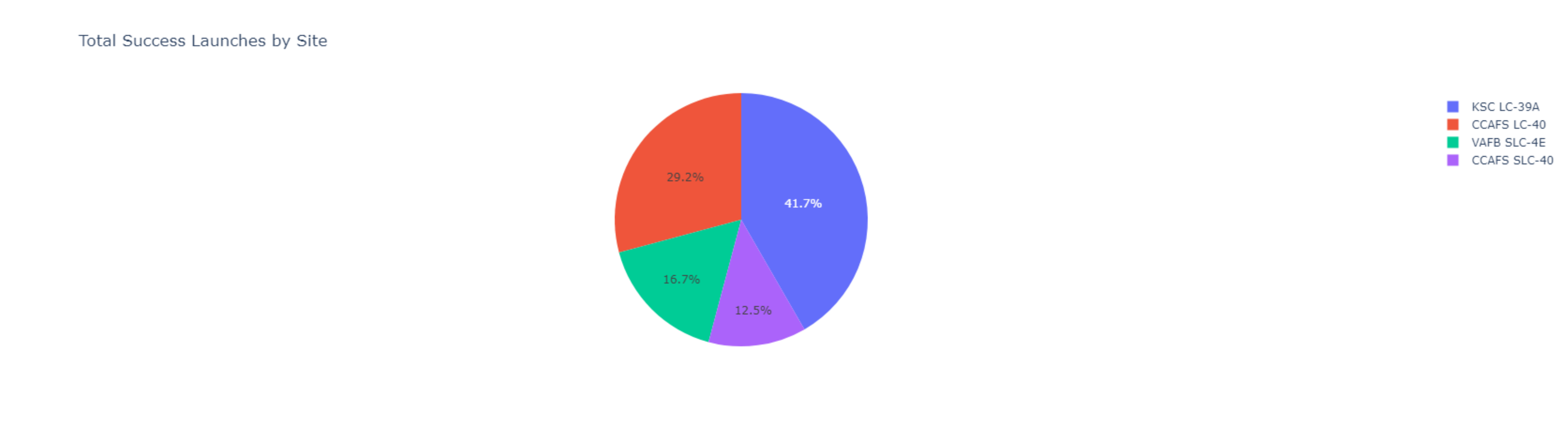


Markers showing proximity to City, Hwy and railroad.

Section 4

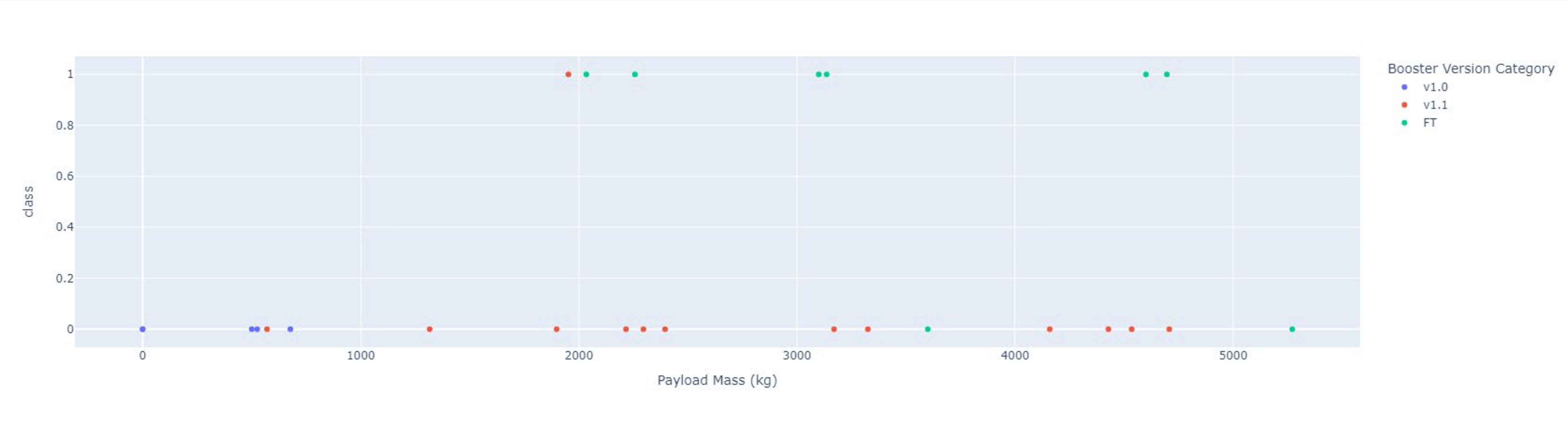
Build a Dashboard with Plotly Dash

Pie Chart showing Successful Launches



- Pie Chart showing successful launches from the sites by their percentage.

Payload Mass vs Launch Site.



- Plotly Dashboard showing a scatter plot of launch site vs payload mass by booster version. Launch site can be selected from the drop-down menu and the range can be modified.

Section 5

Predictive Analysis (Classification)

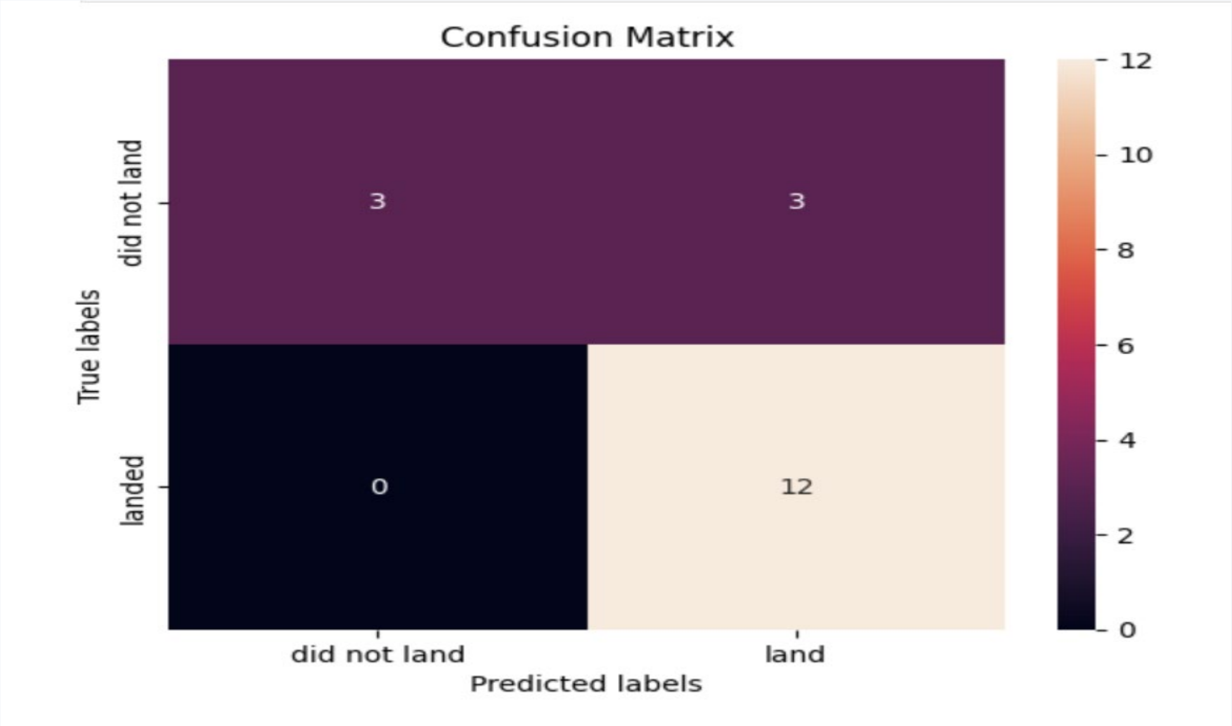
Classification Accuracy

- Model Accuracy was calculated and is shown here.

- All Models had the same accuracy score.
- Decision Tree had slightly better performance than others.

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

Confusion Matrix



- The confusion matrix was same for all the models.

BEST Model

```
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 0.875
Best params is : {'criterion': 'entropy', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split':
10, 'splitter': 'best'}
```

Conclusions

- The Best was the Decision Tree Classifier for prediction.
- SpaceY can use this predictive model along with:
  - Higher payloads
  - ISS and PO orbit.
  - F9 B5 boosters
- More data is required for better results.
- More analytics can be done on the Data to get better results.

Appendix

- Use the GitHub links to see all the code for this project

Thank you!