IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adam Mason
21st February 2024

Outline

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

Executive Summary

SpaceX is currently the largest private company which specialises in space rocket launches and hopes to achieve the historical feat of sending the first person to Mars. A key feature that separates SpaceX from their competitors is being able to reuse the first stage of their launches, reducing costs drastically.

This presentation will attempt to utilise available data on SpaceX to understand the key considerations that must be understood in order to develop our new space rocket company.

We have utilised a combination of data analysis and machine learning techniques during our investigation, and the results and findings will be covered in this presentation.

Introduction

- **Project background and context**
  - Space X is the largest private space company in the world.
  - A large part of their success is being able to keep the price of their space launches low (62 mil vs 165 mil from competitors).
  - Being able to reuse the first stage of a launch is a large factor in why they save so much money.
- **Our Aims?**
  - Determine cost of each space launch, so can determine if the first stage will land.
  - Determine if Falcon9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

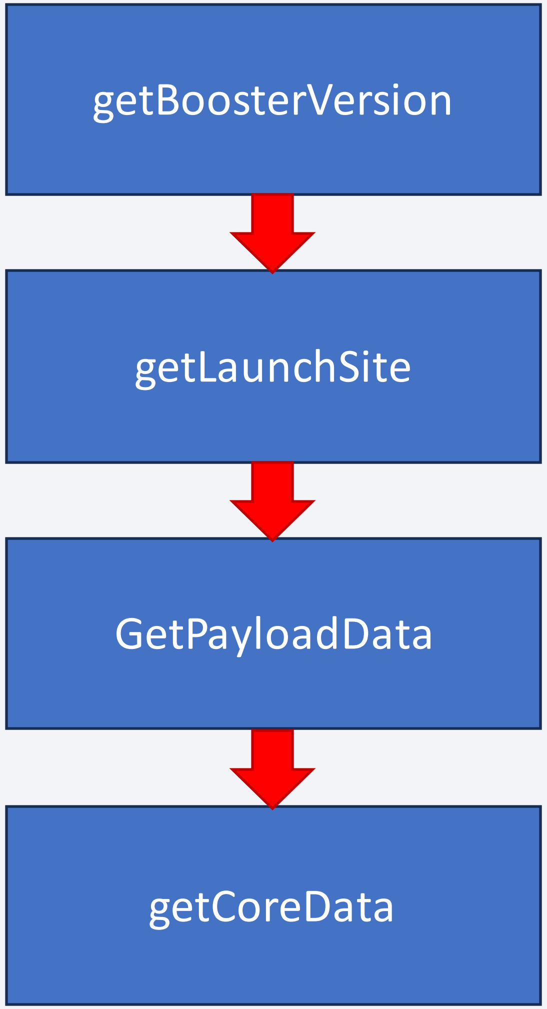
- Data collection methodology:
  - We will discuss how the data was collected.
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification model

Data Collection – SpaceX API

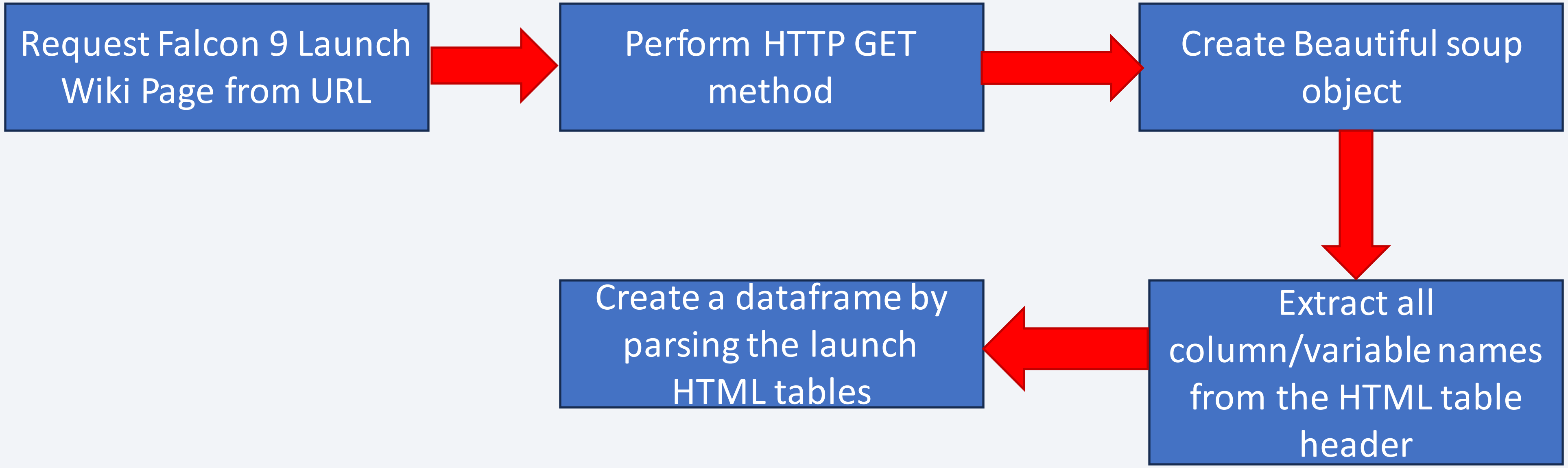
- Data was collected from the SpaceX API.
- Provides info about booster name, launch site coordinates, payload mass, designated orbit, landing outcome, landing type, number of flights and much more.
- Use Pandas to create a dataframe with all of the relevant data.

- Link to python code used to collection the data:

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



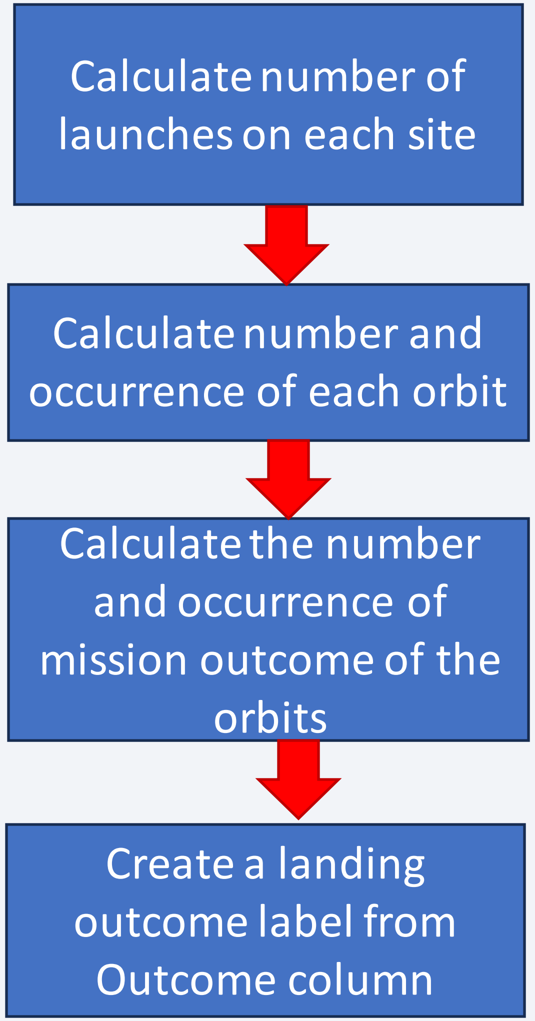
Data Collection - Scraping



- Link to python code used to webscrape the data: Https://github.com/AdamMasonGH/Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Value_counts() was used to determine the number of launches on each site.
- Value_counts() was also used to determine the number and occurrence of each orbit
- Value_counts() was also used to determine the number and occurrence of mission outcome of the orbits.
- Created a list where 1 represents a successful landing and 0 represents a failed landing. We then create a new column in the dataset containing the data in this list.
- Link to python code used to wrangle the data: https://github.com/AdamMasonGH/Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- The names of the scatterplots produced are shown on the right.
- The scatterplots help us to determine the relationship between variables.
- Provides valuable information regarding the most important factors that need considered when launching our own rockets.
- Link to python code to create the plots: https://github.com/AdamMasonGH/Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Scatterplot for Flight number vs Launch Site

Scatterplot: Payload Mass vs Orbit Type

Scatterplot: Launch Site Vs Payload Mass

Lineplot: Launch Success Rate yearly trend

Barlplot: Success rate of each orbit

Scatterplot: Flight Number vs Orbit Type

EDA with SQL

SQL queries performed

- Display names of the unique launch sites
- Display 5 records where launch sites begin with 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the boosters which have drone ship success and have payload mass between 4000 and 6000 kg.
- List the total number of successful and failure mission outcomes
- List the booster versions which have carried the maximum payload mass.
- List the records which will display the month names, failure landing_outcomes in 2015.
- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20.

- Link to code: https://github.com/AdamMasonGH/Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Summary of Task	Explanation
Used Folium.Circle & Folium.Marker to mark all unique launch sites on the map.	Gives us insight into where the best locations are to build our own launch sites.
Used marker clusters to mark successful and failed launches at each launch site on the map. Green markers = successful. Red markers = failure.	This tells us which launch sites have the most successful and most failed launches. Clusters allow us to mark multiple entries that have the same coordinates on our map without overlapping.
Used Polyline to draw lines between launch sites and key geographical features, such as the nearest coastline, city, railway and highway.	Provides insight into considerations that have to be taken when choosing the optimal location for our launch site.

- Link to code: https://github.com/AdamMasonGH/Data-Science-Capstone-Project/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- **Plots/graphs and interactions added to dashboard**
  - Pie charts which show the launch success rate of each launch site
  - Pie chart which shows the proportion of successful launch rates from each site
  - Interactive scatterplot which shows correlation between payload mass and launch success rate for each site
- **Want to determine:**
  - Which site has the most successful launches?
  - Which site has the highest launch success rate?
  - Which payload ranges have the highest and lowest launch success rate?
  - Which F9 booster version has the highest launch success rate?
- Link to code: https://github.com/AdamMasonGH/Data-Science-Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

1. Create NumPy array from Class column in dataframe
2. Standardise the data
3. Split data into training and testing data
4. Create logistic regression object and find the best parameters
5. Create an SVM and find the best parameters
6. Create a decision tree classifier object and find the best parameters
7. Create a k nearest neighbors object and find the best parameters
8. Calculate the accuracy of the models on the test data
9. Find the method which performs best

- Link to code: https://github.com/AdamMasonGH/Data-Science-Capstone-Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

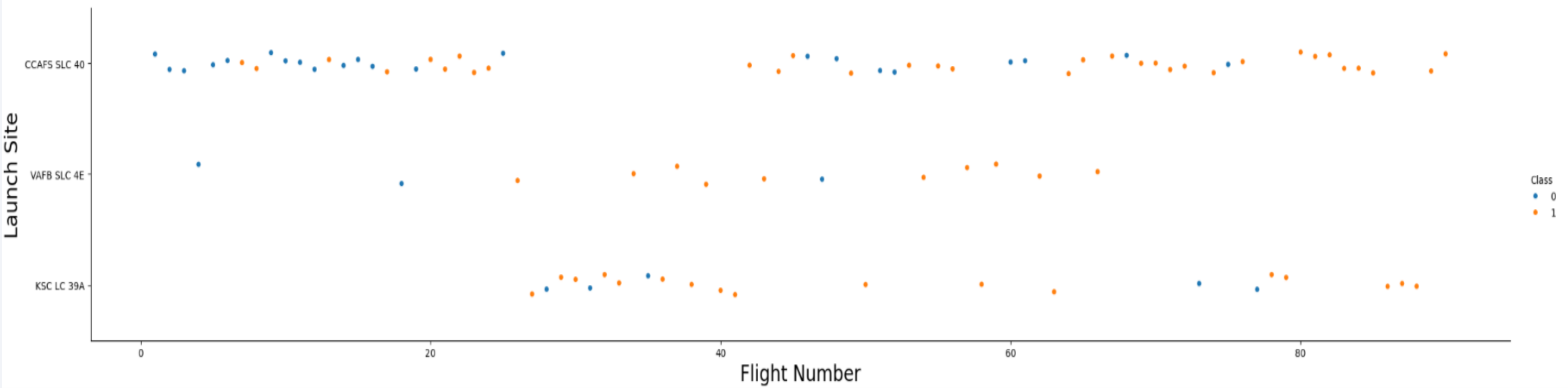
Results

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

Section 2

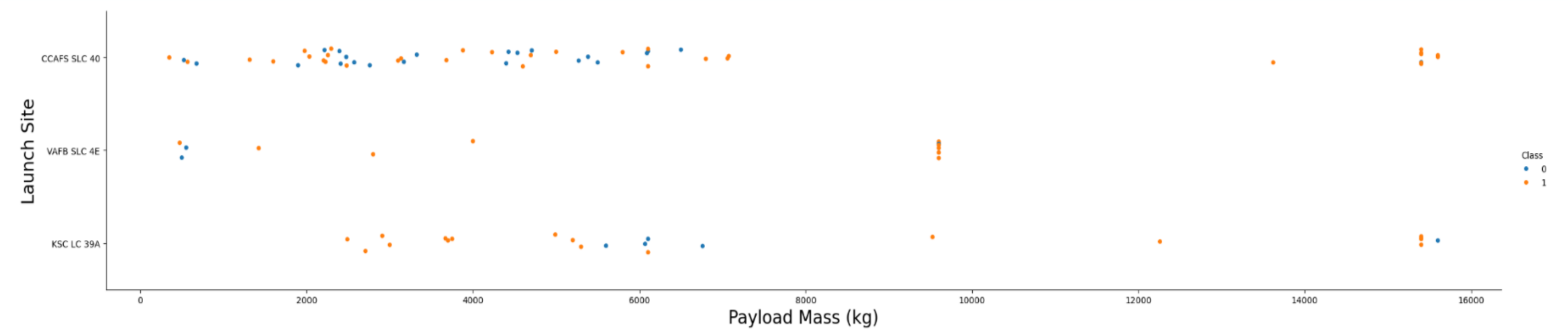
Insights drawn from EDA

Flight Number vs. Launch Site



- Number of successful launches appears to increase as the flight number increases. Especially noticable in CCAFS SLC 40.
- Majority of the failures occurred when the flight number was <30.

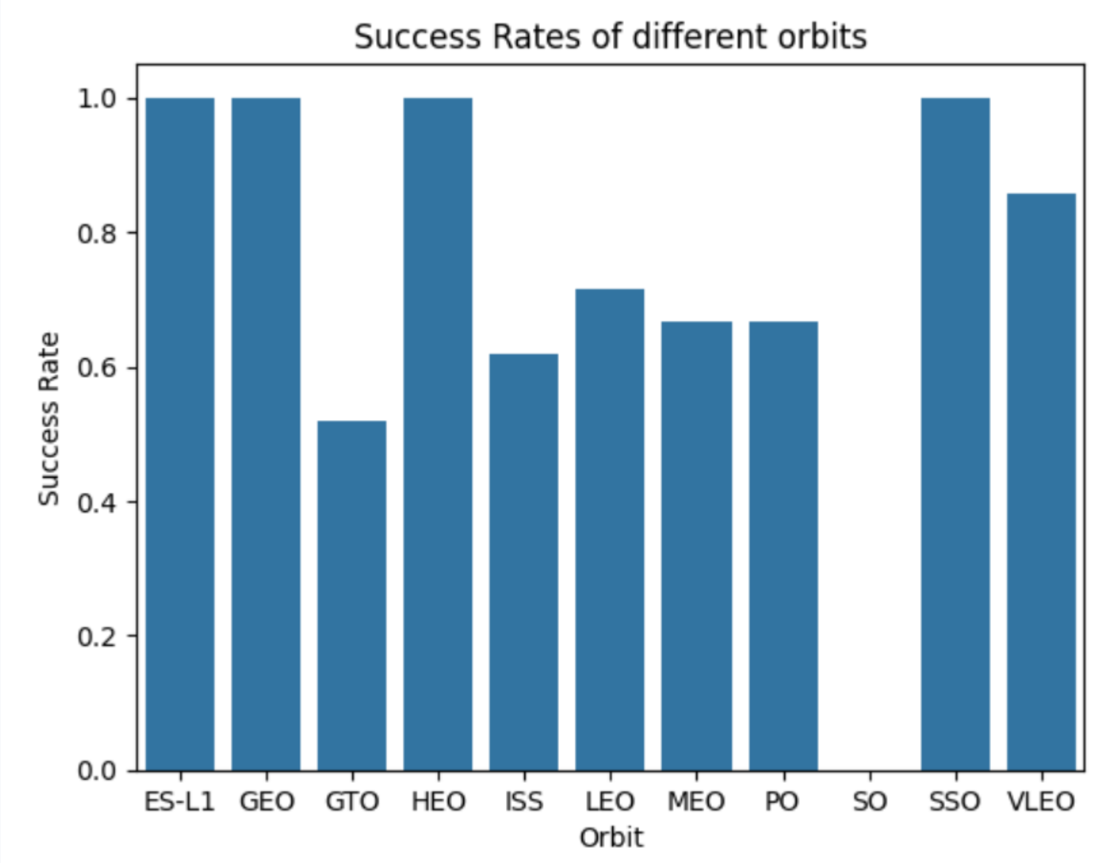
Payload vs. Launch Site



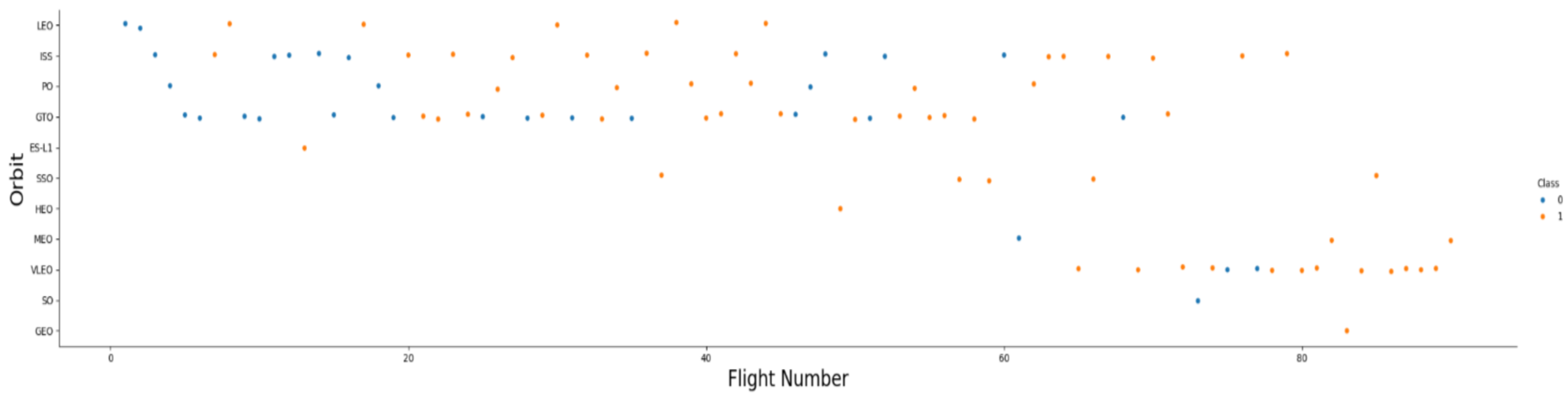
- Successful landings tended to be higher at higher payload masses.
- Majority of the failed landings occurred when the payload mass was < 5000kg.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO had the highest success rates, with 0 failed landings between them.
- All orbit types had a success rate >50 %, except for SO.

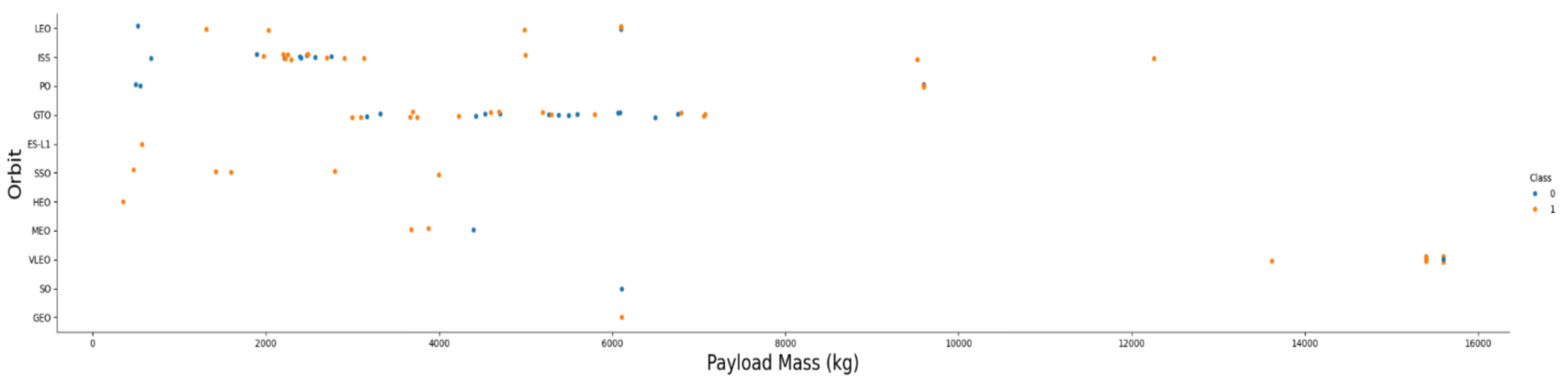


Flight Number vs. Orbit Type



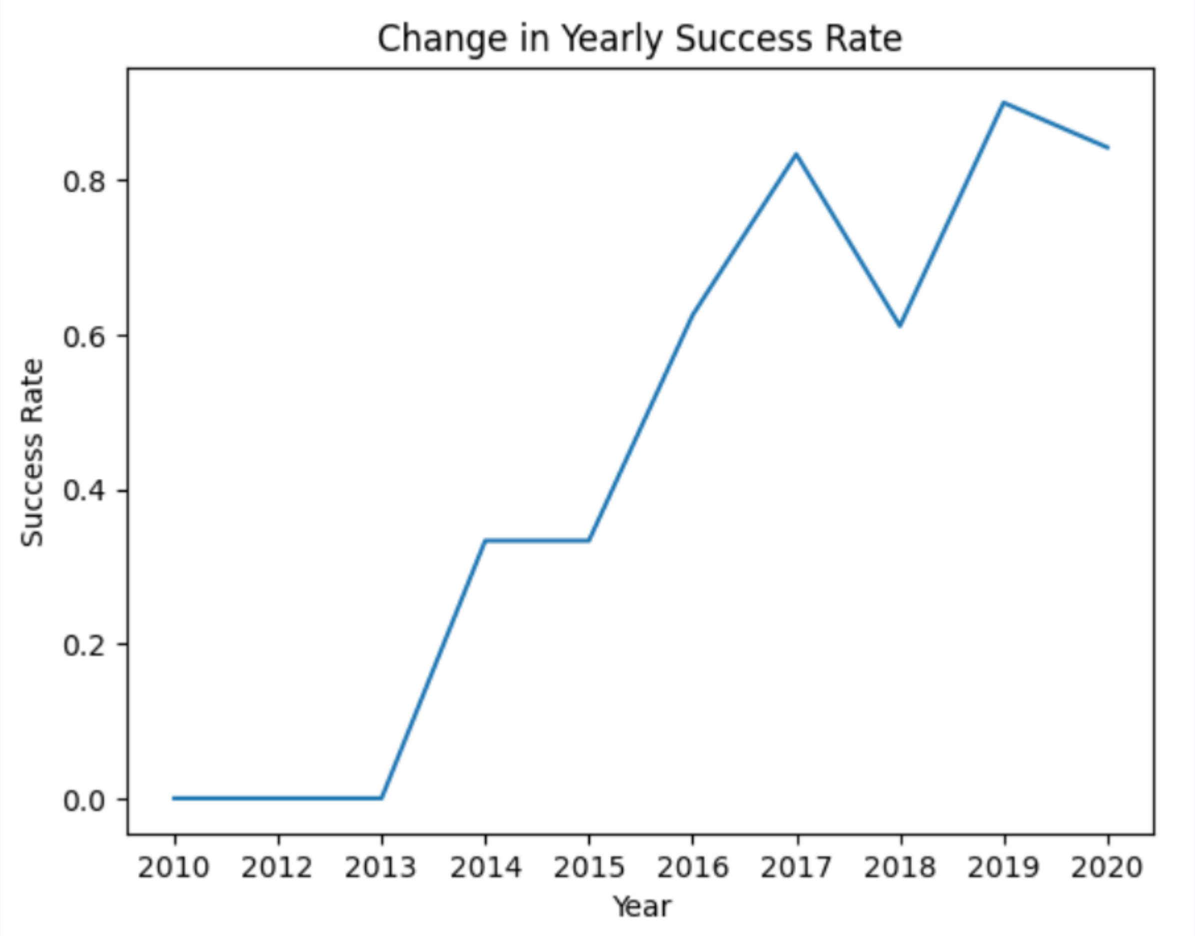
- Majority of orbit types experienced higher success rates at higher flight numbers.
- SO and GEO only had one flight. LEO, ISS, PO and GTO had the most.

Payload vs. Orbit Type



- Majority of launches occurred at payload masses < 6000kg.
- The rate of successful launches tended to increase at higher payloads, but there is a small sample size at higher payload masses.

Launch Success Yearly Trend



- Success rate from 2010 onwards has steadily increased, reaching its highest rate in 2019

All Launch Site Names

- We can obtain the names of the unique launch sites with the following line(s) of code:

```
In [8]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

- This produces the following output on the right ->

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

Launch Site Names Begin with 'CCA'

- To produce 5 records where launch sites begin with 'CCA', we can use the following code:

```
In [9]: %%sql SELECT * FROM SPACEXTABLE
        WHERE Launch_Site LIKE 'CCA%'
        LIMIT 5;
```

- This produces the following output on the right ->

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

- To calculate the total payload carried by boosters from NASA, we can use the following code:

```
In [10]: %%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
         WHERE Customer LIKE "%NASA%";
```

- Which produces the following output:

SUM(PAYLOAD_MASS__KG_)
107010

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1, the following code was used:

```
In [11]: %%sql SELECT avg(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
         WHERE Booster_Version = 'F9 v1.1';
```

- This gives the following output:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- To find the dates of the first successful landing outcome on ground pad, the following code was used:

```
In [12]: %%sql SELECT Min(Date), Landing_Outcome FROM SPACEXTABLE
         WHERE Landing_Outcome = 'Success (ground pad)';
```

- This produces the following output:

Min(Date)	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- To produce the list of names of boosters which have successfully landed on drone ships and had a payload mass between 4000 and 6000 kg, the following code was used:

```
In [13]:
%%sql SELECT DISTINCT(Booster_Version), PAYLOAD_MASS__KG_, Landing_Outcome from SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

- This produces the following output:

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failed mission outcomes, the following code was used:

```
In [14]: %%sql SELECT DISTINCT(Mission_Outcome),COUNT(Mission_Outcome) FROM SPACEXTABLE
         GROUP BY Mission_Outcome;
```

- This produces the following output:

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

Boosters Carried Maximum Payload

- To create a list of the names of the boosters which have carried the maximum payload mass, the following code was used:

```
In [15]:  %%sql SELECT DISTINCT(Booster_Version), PAYLOAD_MASS__KG_ FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

- This produces the following code on the right ->

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

- To list the failed landing outcomes in drone ship, their booster versions, and launch site names for 2015, the following code was used:

```
In [21]:  %%sql SELECT substr(Date,6,2), Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE
          WHERE substr(Date,0,5)='2015'
          AND Landing_Outcome = 'Failure (drone ship)';
```

- This produces the following output:

substr(Date,6,2)	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- To rank the count of landing outcomes between 2010-06-04 and 2017-03-20, the following code was used:

In [32]:
```
%%sql SELECT Date, Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT(Landing_Outcome) DESC;
```

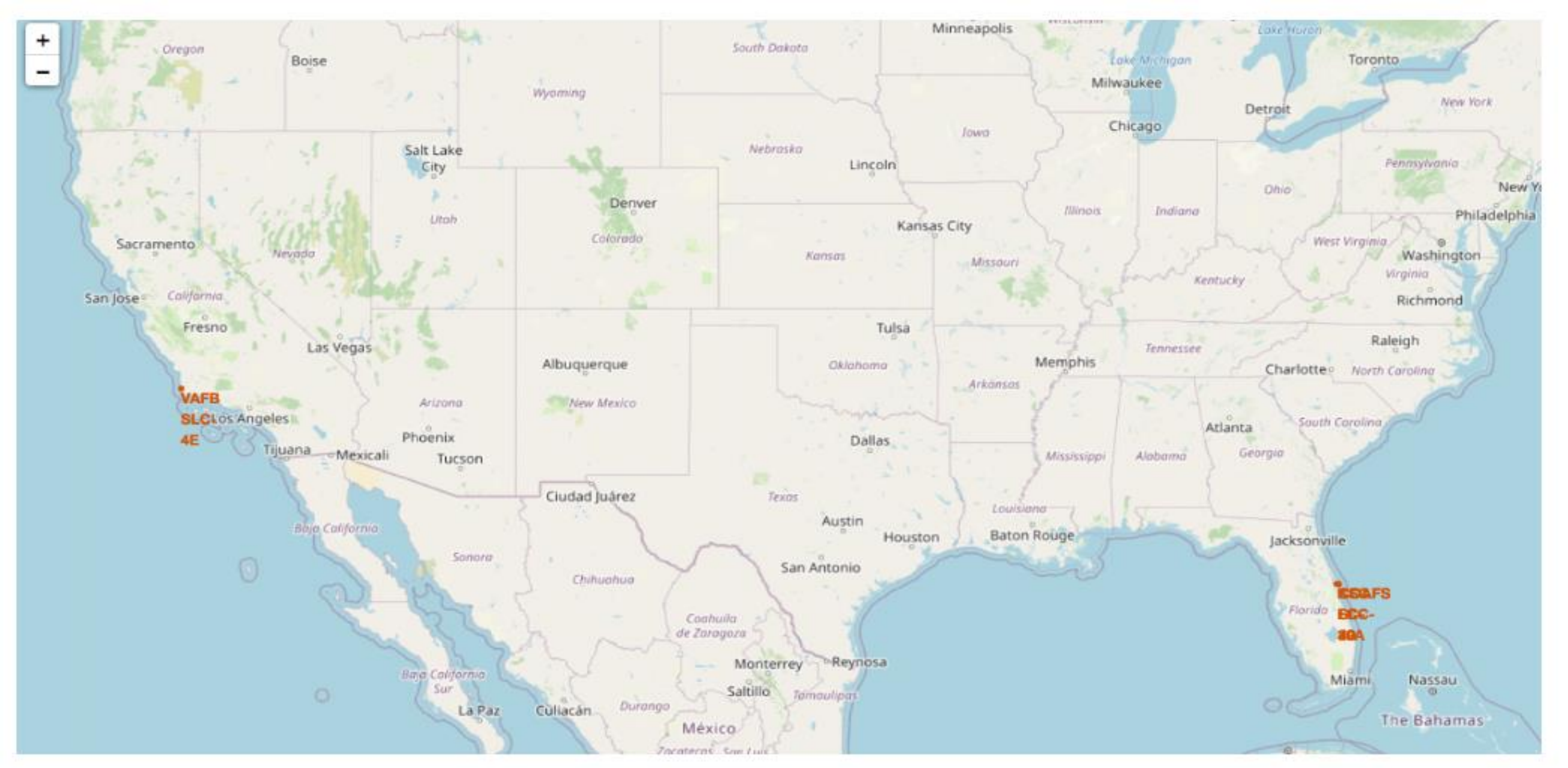
- This produces the following output on the right:

Date	Landing_Outcome	COUNT(Landing_Outcome)
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2010-06-04	Failure (parachute)	2
2015-06-28	Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Map with all launch sites marked

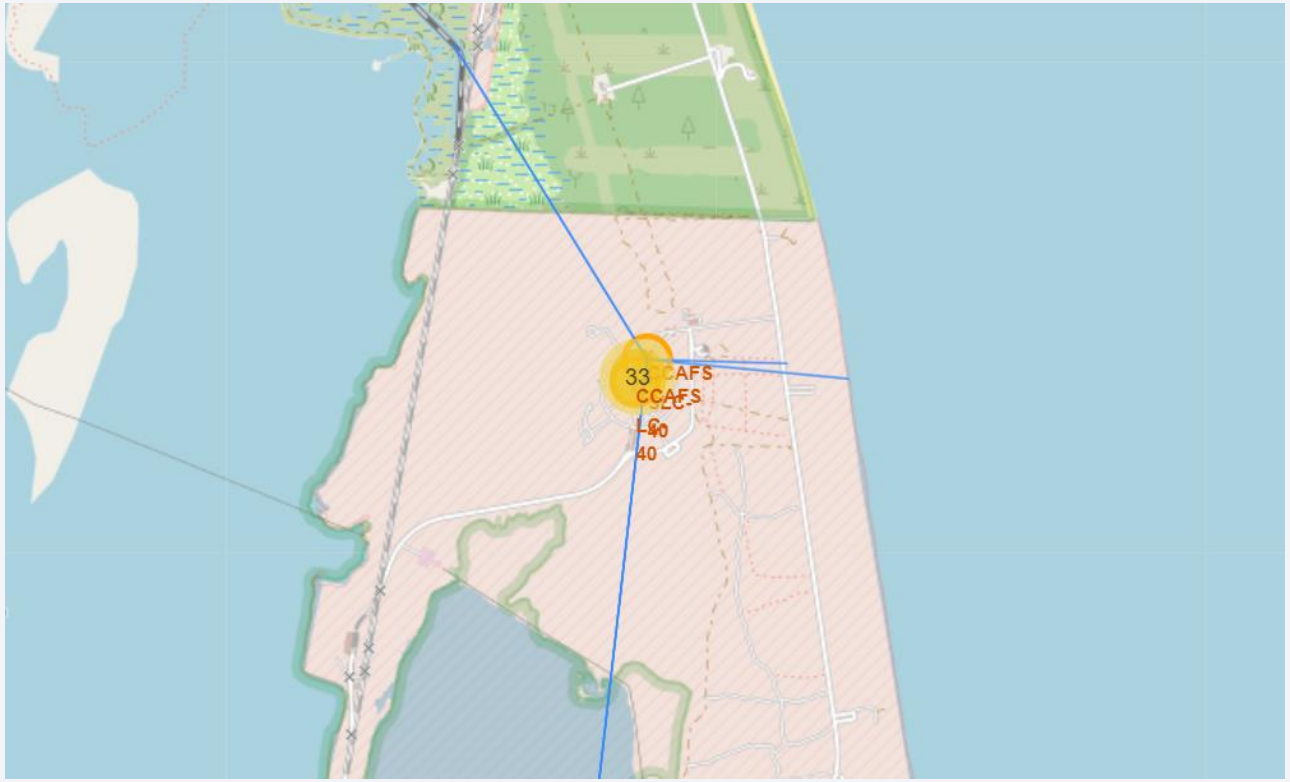


- From this we can see that the launch sites located in the United States are situated close to both the western and eastern coastlines. This is so failed launches can land in the water and away from any towns or cities in the vicinity.

Map of launch outcomes at a specified launch site

- The green markers indicate a successful outcome and the red markers indicate a failed outcome. We can see that most of the launches failed, but there is a very small sample size.

Map showing the distance to the closest landmarks



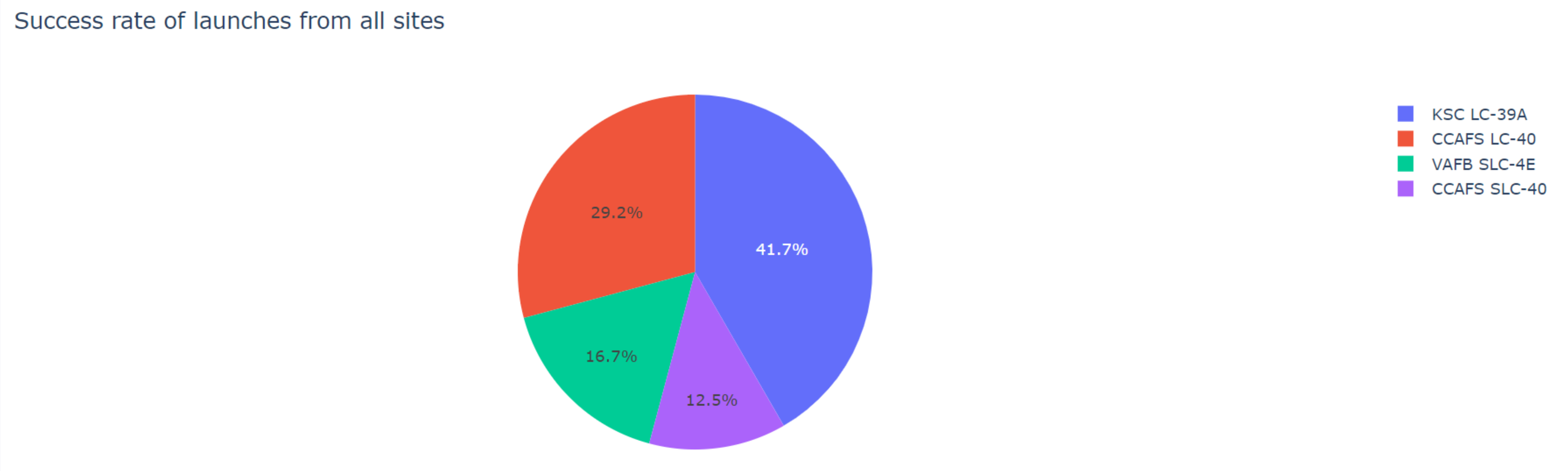

```
Distance to closest coastline (km): 0.8882231888122465
Distance to closest city (km): 51.45929218526419
Distance to closest railway (km): 1.611082369921484
Distance to closest highway (km): 0.6097626019594105
```

- The map shows that the closest coastline, railway and highway are within a very short distance of the launch site, which helps for transportation purposes and allowing launches to land safely in the water.
- The closest city is located very far away in order to minimize disturbances to the general population

Section 4

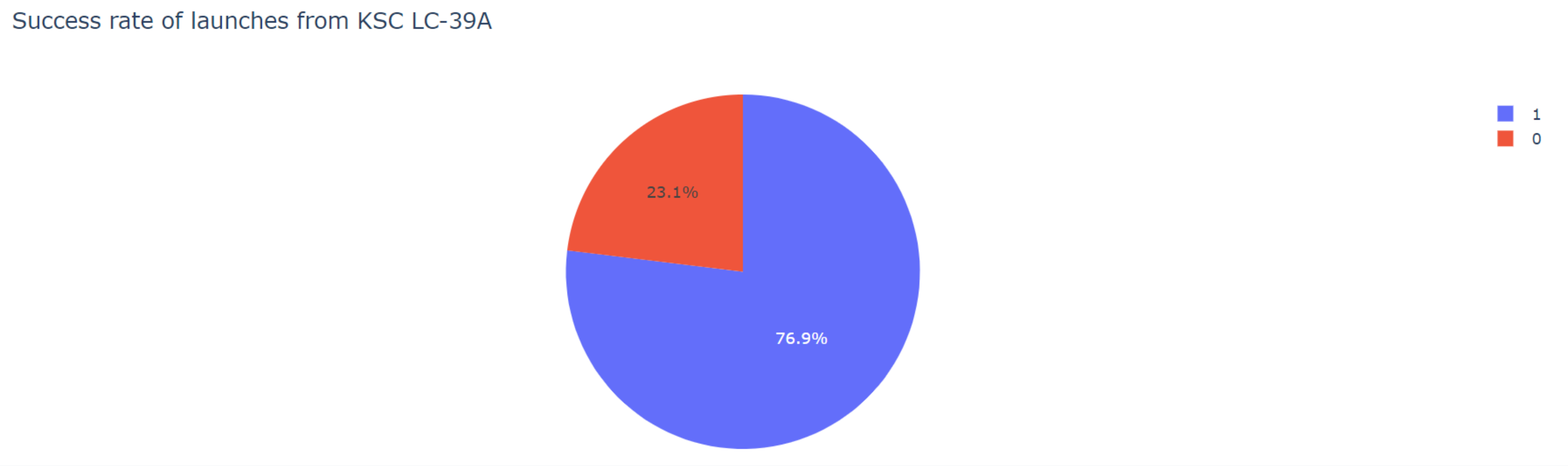
Build a Dashboard with Plotly Dash

Success rate of launches from all sites



- From the piechart, we can see that the KSC LC-39A had the largest proportion of successful launches with 41.7%, whilst CCAFS SLC-40 had the lowest at 12.5%.

Success rate of launches from KSC LC-39A



- The piechart above shows the success rate of the launch site with the highest proportion of successful launches: KSC LC-39A.
- Over ¾ of the total launches from this launch site had a successful outcome.

Effect of payload mass on success rate for all launch sites

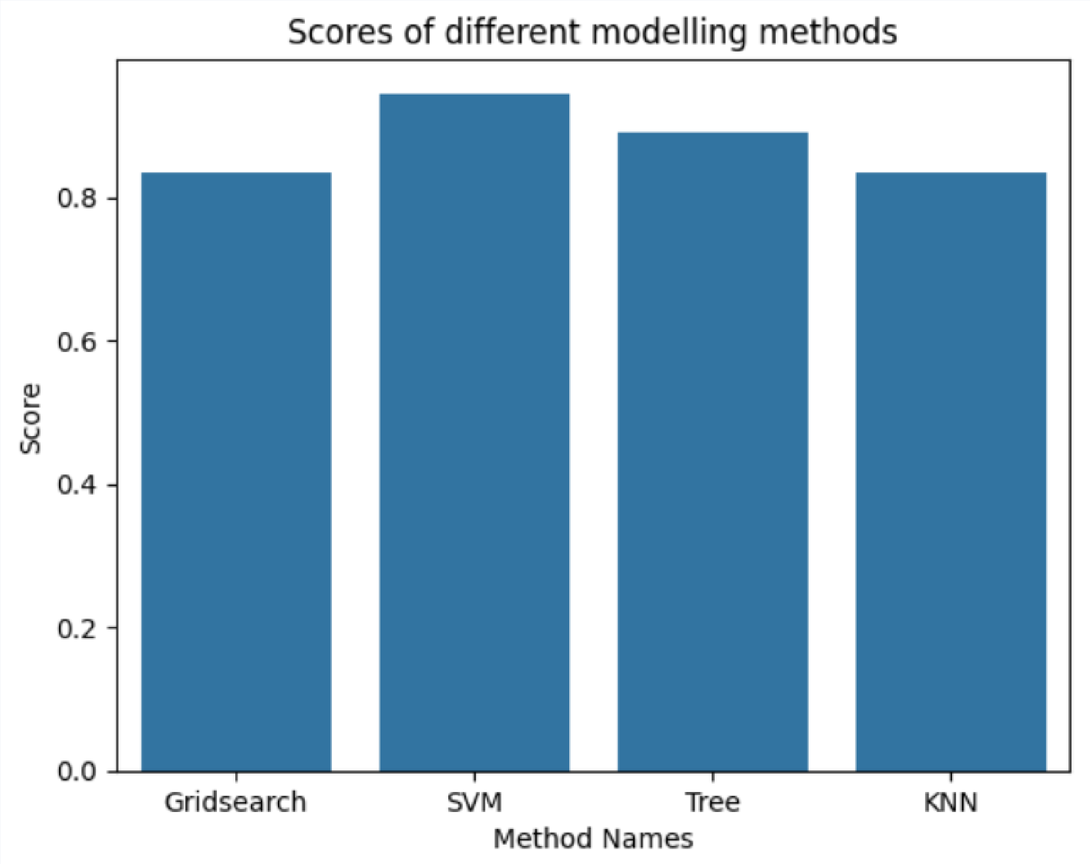
- From the scatterplot above, we can see that the 'FT' booster version category had the most successful launch outcomes, whilst 'v1.1' had the most failed outcomes.
- Only the 'v1.0' had launches where the payload mass was greater than 9000kg, and had mixed success, although there is a very small sample size.

Section 5

Predictive Analysis (Classification)

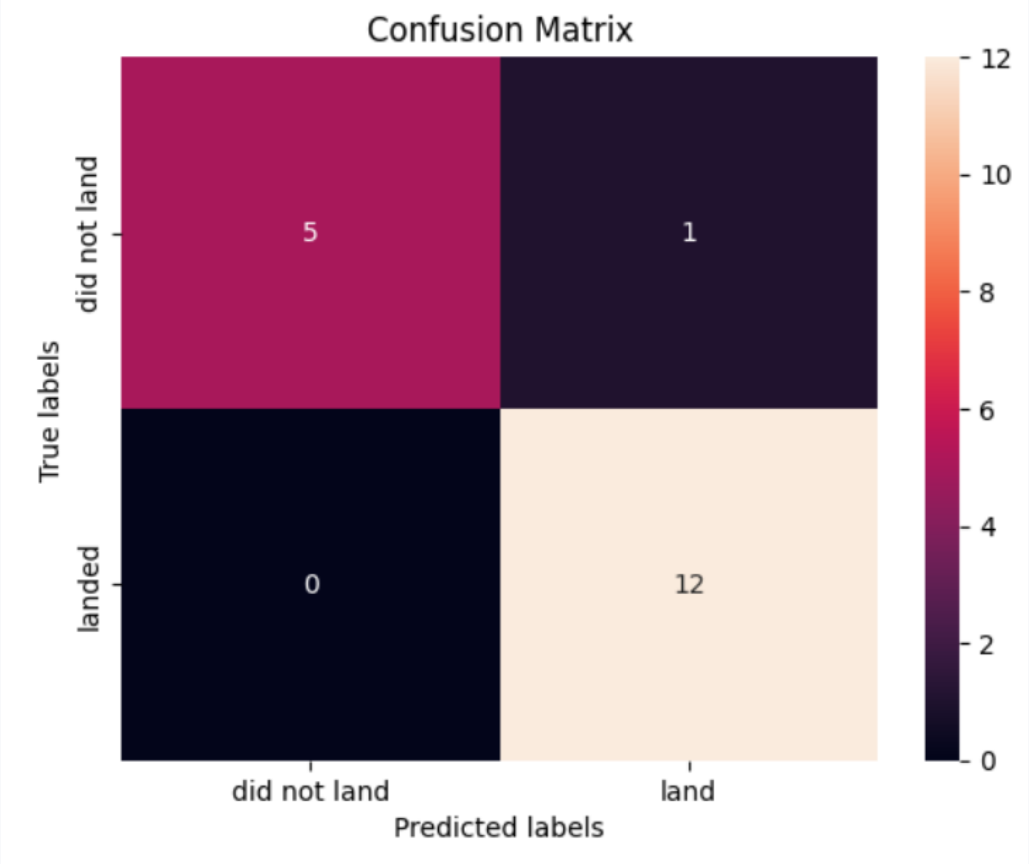
Classification Accuracy

- From the bar chart, we can see the Support Vector Machine method scored the highest accuracy, with a score of over 90 %.
- The Gridsearch and KNN methods scored the lowest accuracy with ~84 %.



Confusion Matrix

- The confusion matrix helps to visualise the performance of the model by showing the true and false positives/negatives in the sample.
- We can see that the model does a good job at predicting the outcomes of the testing samples, with only 1 false negative and 0 false positives.

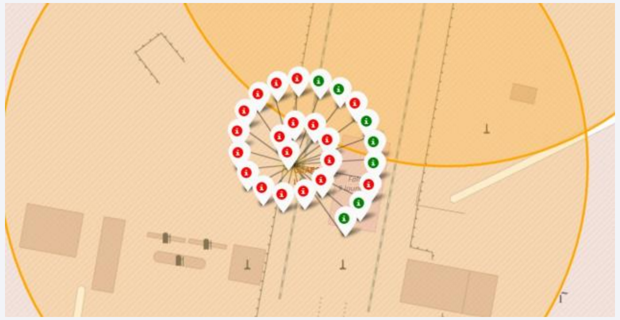


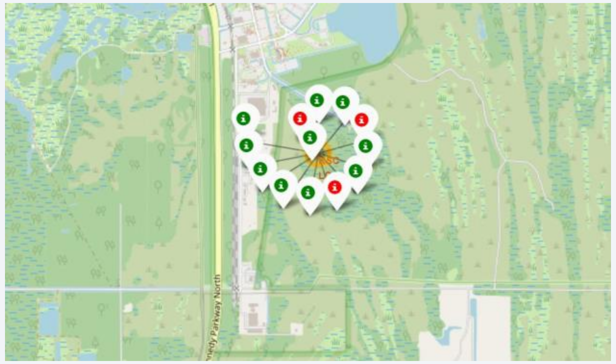
Conclusions

- Success rates of launches have steadily increased each year from 2010 onwards.
- Based on existing launch sites, the optimal location for launch sites appears to be areas close to railways, highways and coastlines, whilst also being far away from the nearest city.
- The KSC LC-39A launch site had the highest number of successful launch sites
- The Support Vector Machine method scored the highest classification accuracy of all of the models utilised, producing an overall accuracy score of 94.4 %

Appendix

Additional Folium Maps





Additional Payload Range Slider Maps

Thank you!