IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

EMMANUEL ONAH
10 March 2023

Outline

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

Executive Summary

- **Summary of methodologies:**
  - Dataset used for the project was collected via several ways such as web scrapping from Wikipedia page and SpaceX API calls.
  - Data wrangling was performed followed by exploratory data analysis and feature selection, engineering and standardization and one-hot encoding.
  - Several interactive visualization were done to communicate the findings.
  - Classification machine learning models were built to predict the landing success of the SpaceX rocket.
- **Summary of all results:**
  - All the models - logistic regression, decision tree, k-nearest neighbor, and support vector machine exhibited high predictive accuracy of greater than 83%. However, decision tree had the highest predictive accuracy of over 88%.
  - Factors such as launch site, payload mass, and orbit contribute significantly to the successful landing of a Falcon 9 rocket.

Introduction

- **Project background and context**

  SpaceX is an American spacecraft manufacturer, launcher, and satellite communications company. Elon Musk founded it in 2002 with the stated goal of lowering space transportation costs to enable Mars colonization. The Falcon 9, Falcon Heavy, and Starship launch vehicles, as well as several rocket engines, Cargo Dragon and Crew Dragon spacecraft, and Starlink communications satellites, are all manufactured by the company. SpaceX has accomplished a number of feats in space exploration. On its website, SpaceX promotes Falcon 9 rocket launches for 62 million dollars; other suppliers charge upwards of 165 million dollars for each launch. A large portion of the savings is due to SpaceX's ability to reuse the first stage.

  Hence, in this project, we aim to figure out if the first stage will land, we can figure out how much a launch will cost. If a different business want to compete with SpaceX for a rocket launch, it may use the information provided here.

- **Problems you want to find answers**

  - Predicting whether the Falcon 9 first stage will land successfully is the main problem this project is aimed at answering.

  Other problems whose answer will ultimately lead to the above goal include:

  - What are the factors affecting the successful landing of a Falcon 9 rocket?
  - What is the accuracy of a successful landing of Falcon 9?
  - How does the site of launch, the payload mass, orbit, etc. affect the successful landing of falcon 9?

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - Request to the SpaceX API
  - Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
- **Perform data wrangling:**
  - Missing values were filled up and appropriate type conversion perfumed
  - Data were standardized and one-hot encoded to be used later as input for the machine learning model
- **Perform exploratory data analysis (EDA) using visualization and SQL**
  - Discovering new patterns in the data using various visualizations such as scatter plot
- **Perform interactive visual analytics using Folium and Plotly-Dash**
  - Dash and Folium were used to create interactive plots
- **Perform predictive analysis using classification models:**
  - Logistic regression, Decision tree, K-nearest neighbor, and support vector machine were trained on 80% of the dataset after hyper-parameter tuning and the remaining 20% was used for testing of the models

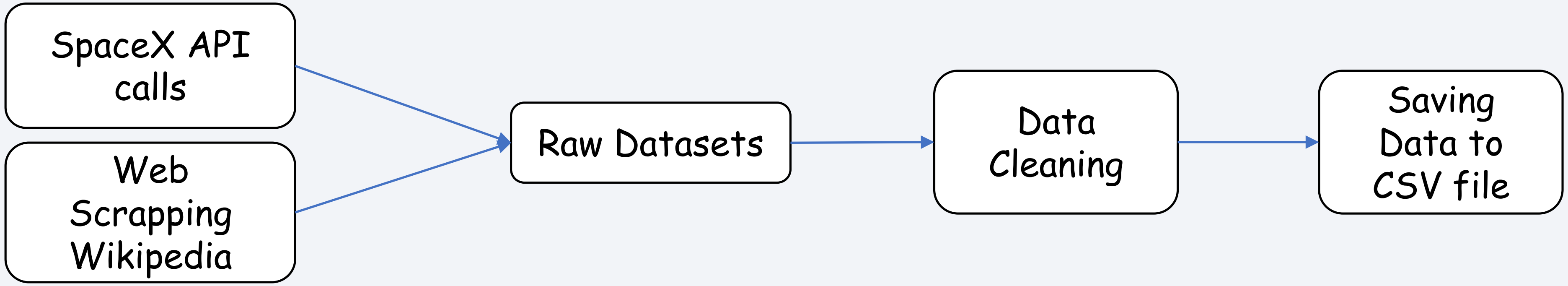
Data Collection

- **The dataset used for this project was collected from two major sources -**

  1) By making request to the SpaceX API endpoint (https://api.spacexdata.com/v4/) and

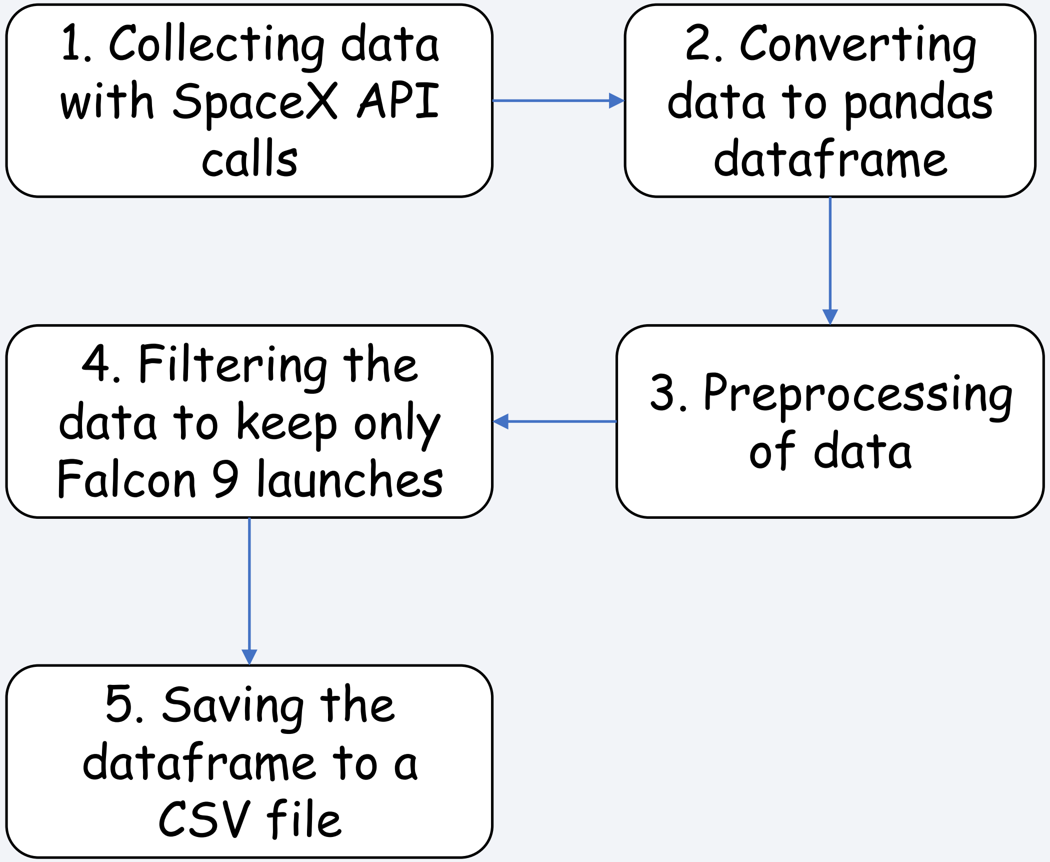
  2) By Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922).

- **The flowchart below shows the data collection process:**



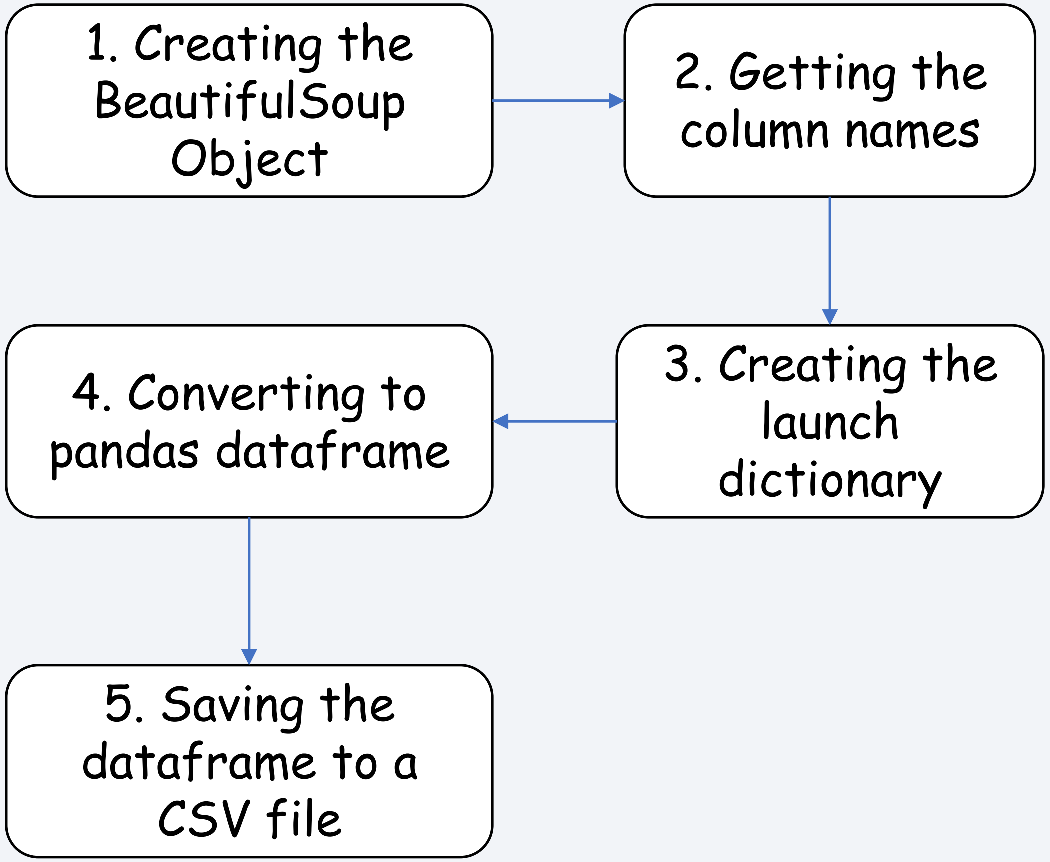
Data Collection - SpaceX API

- The “get method” of the request library was used to make the SpaceX REST calls to the API endpoint.
- The link to the GitHub repo where the completed SpaceX API calls notebook was hosted is given below: https://github.com/OnahPmi/My_Coursera_Repo/blob/77b67417658e9ae94a8396e402942ed25b494e39/Applied%20Data%20Science%20Capstone%20Project/SpaceX_Falcon_9_First_Stage_Landing_Prediction_(SpaceX_API_calls).ipynb



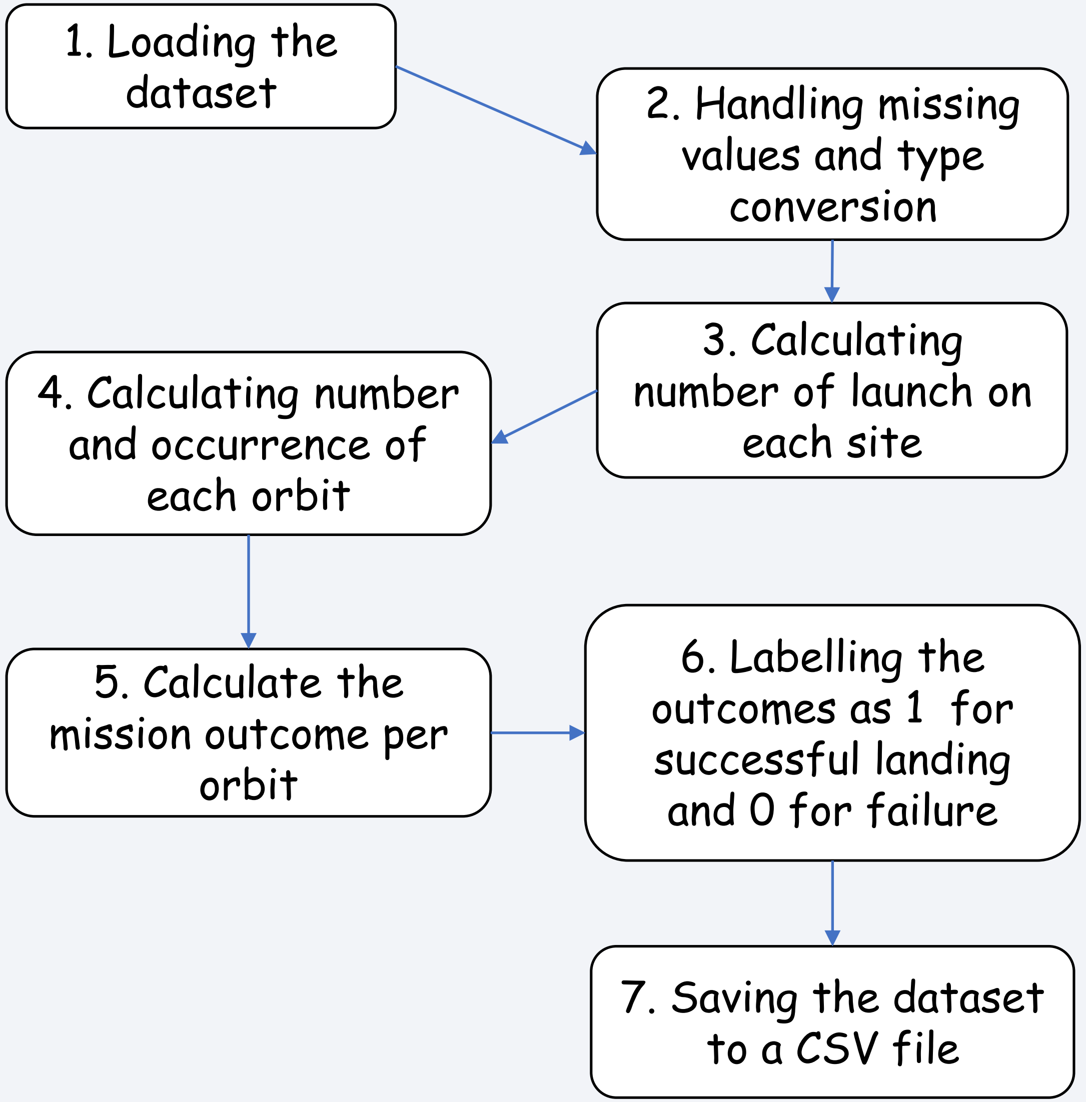
Data Collection - Scraping

- The BeautifulSoup python library was used to scrap the Wikipedia site for Falcon 9 and Falcon Heavy Launches Records.
- The link to the GitHub repo where the completed web scraping notebook was hosted is given below: https://github.com/OnahPmi/My_Coursera_Repo/blob/aabcfb02e258a6ea82529b526d8daf1ee69238e4/Applied%20Data%20Science%20Capstone%20Project/SpaceX_Falcon_9_First_Stage_Landing_Prediction_(Web_scrapping_records_from_Wikipedia).ipynb



Data Wrangling

- **The dataset was processed as follows:**
    - Missing values were replaced with the mean value of the column.
    - Appropriate type conversion were done.
    - Number of launch on each site was calculated.
    - Number and occurrence of each orbit was calculated.
    - Number and occurrence of mission outcome per orbit type was calculated
    - Outcome labels ("1" for Successful outcome and "0" for bad outcome) was created.
- The link to the GitHub repo where the completed data wrangling related notebooks was hosted is given below: https://github.com/OnahPmi/My_Coursera_Repo/blob/ef7c6a854321fbf6a7eed4ff28369606a05425ba/Applied%20Data%20Science%20Capstone%20Project/SpaceX_Falcon_9_First_Stage_Landing_Prediction_(Data_wrangling).ipynb



EDA with Data Visualization

- "Categorical", "scatter" and "line" plots were generated to find some pertains in the data. Pandas, Seaborn, and Matplotlib were used for the exploratory data analysis (EDA).
- The link to the GitHub repo where the completed exploratory data analysis, data preparation and feature engineering notebook was hosted is given below: https://github.com/OnahPmi/My_Coursera_Repo/blob/9eeb98da70a2ab44a5baec1d50423dc9c586f7a4/Applied%20Data%20Science%20Capstone%20Project/SpaceX_Falcon_9_First_Stage_Landing_Prediction_(Exploring_and_Preparing_Data).ipynb

EDA with SQL

- **The following are the SQL queries performed:**
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string '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 names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - 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.
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- The link to the GitHub repo where the completed EDA with SQL notebook was hosted is given below:
  https://github.com/OnahPmi/My_Coursera_Repo/blob/60211f6d993e7db099662ff88fac650ce5e0f269/Applied%20Data%20Science%20Capstone%20Project/SpaceX_Falcon_9_First_Stage_Landing_Prediction_(EDA_with_SQL).ipynb

Build an Interactive Map with Folium

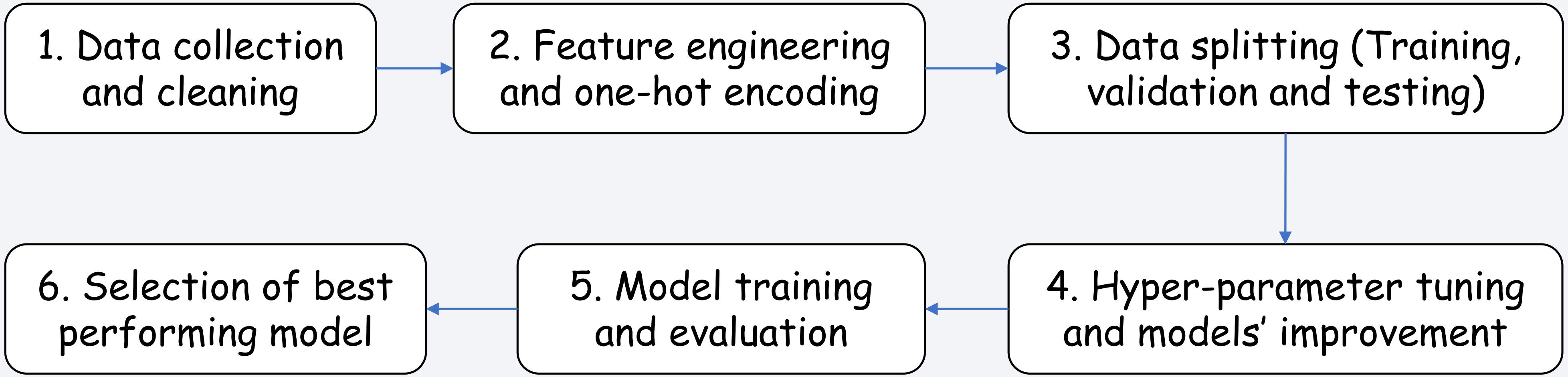
- **The map objects created and added to a folium map include:**
  - folium.Marker(): to create marks on the map.
  - MarkerCluster(): to simplify a map which contain several markers with identical coordinates.
  - MousePosition(): to display coordinate of a point as you mouse over the map.
  - folium.Circle(): to add a highlighted circle area with a text label on a specific coordinate.
  - Folium.PolyLine(): to create polynomial lines between points on the map.
- The link to the GitHub repo where the completed interactive map with Folium map notebook was hosted is given below: https://github.com/OnahPmi/My_Coursera_Repo/blob/c3a2dc4390c98a851416e9b741b0fdd23f3890f0/Applied%20Data%20Science%20Capstone%20Project/Launch_Sites_Locations_Analysis_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- **The following were used to build the dashboard with plotly dash:**
  - Dash HTML components and callbacks: Important components of a dash application. Everything such as graphs, rangeslider and dropdown, etc. depend on them.
  - Pandas was used to load dataset into a dataframe and also for processing the data before plotting.
  - Plotly was used to plot the graphs --- pie charts and scatter plots
  - Rangeslider was used for selecting payload mass range.
  - Dropdown was used for launch site selection.
- The link to the GitHub repo where the completed Plotly Dash lab notebook was hosted is given below: https://github.com/OnahPmi/My_Coursera_Repo/blob/ed94597f3d4d9f83fc8f1b063bbd4307a8bd585a/Applied%20Data%20Science%20Capstone%20Project/spacex_dash_app.py

Predictive Analysis (Classification)

- **Summary of how the models were built, evaluated, improved, and the best performing classification model selected:**



- The link to the GitHub repo where the completed predictive analysis lab notebook was hosted is given below: https://github.com/OnahPmi/My_Coursera_Repo/blob/4bd192bd158670e6a0128426319c2331f944f36d/Applied%20Data%20Science%20Capstone%20Project/SpaceX_Falcon_9_First_Stage_Landing_Prediction_(Machine_Learning_Prediction).ipynb

Results

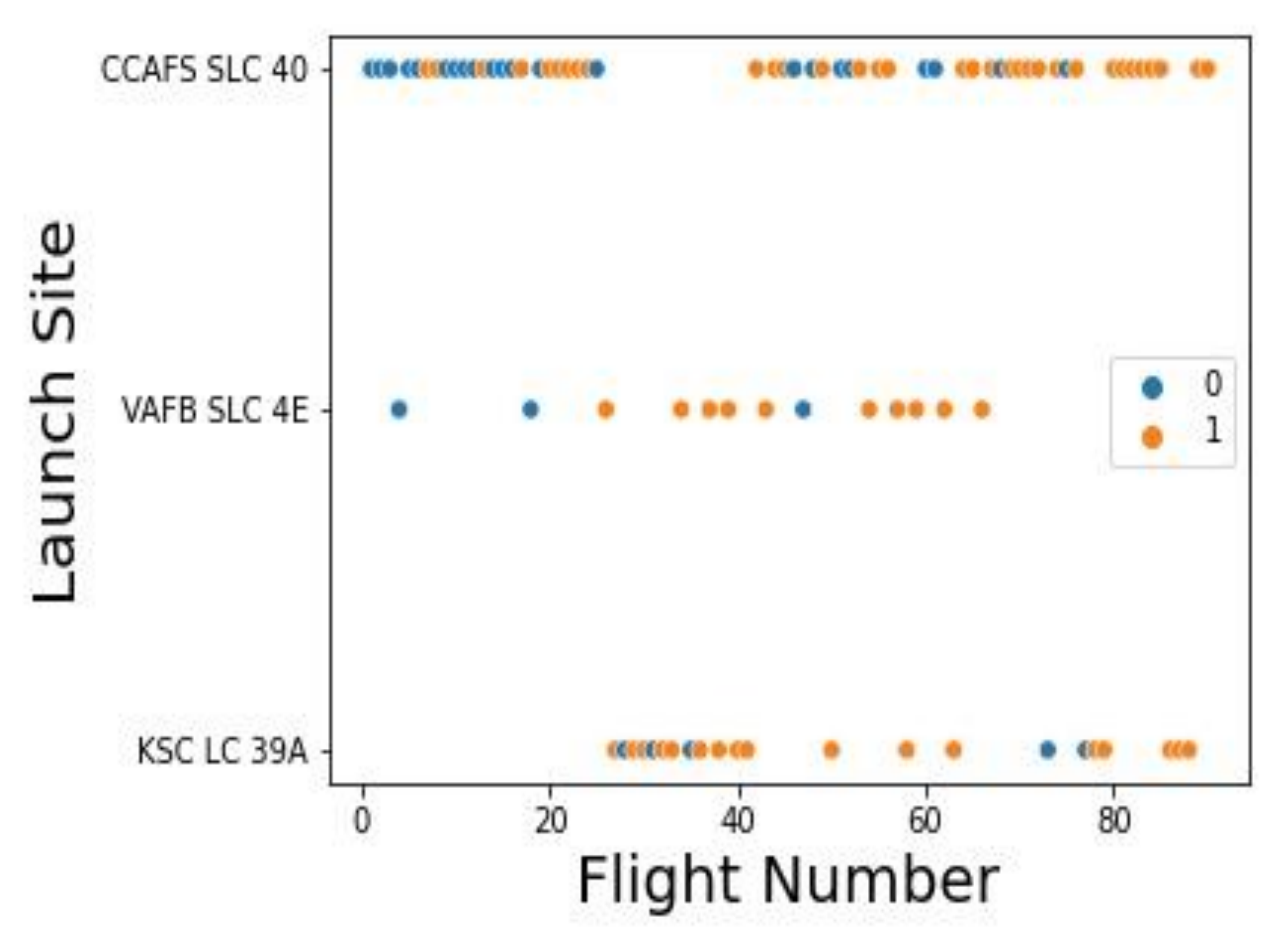
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

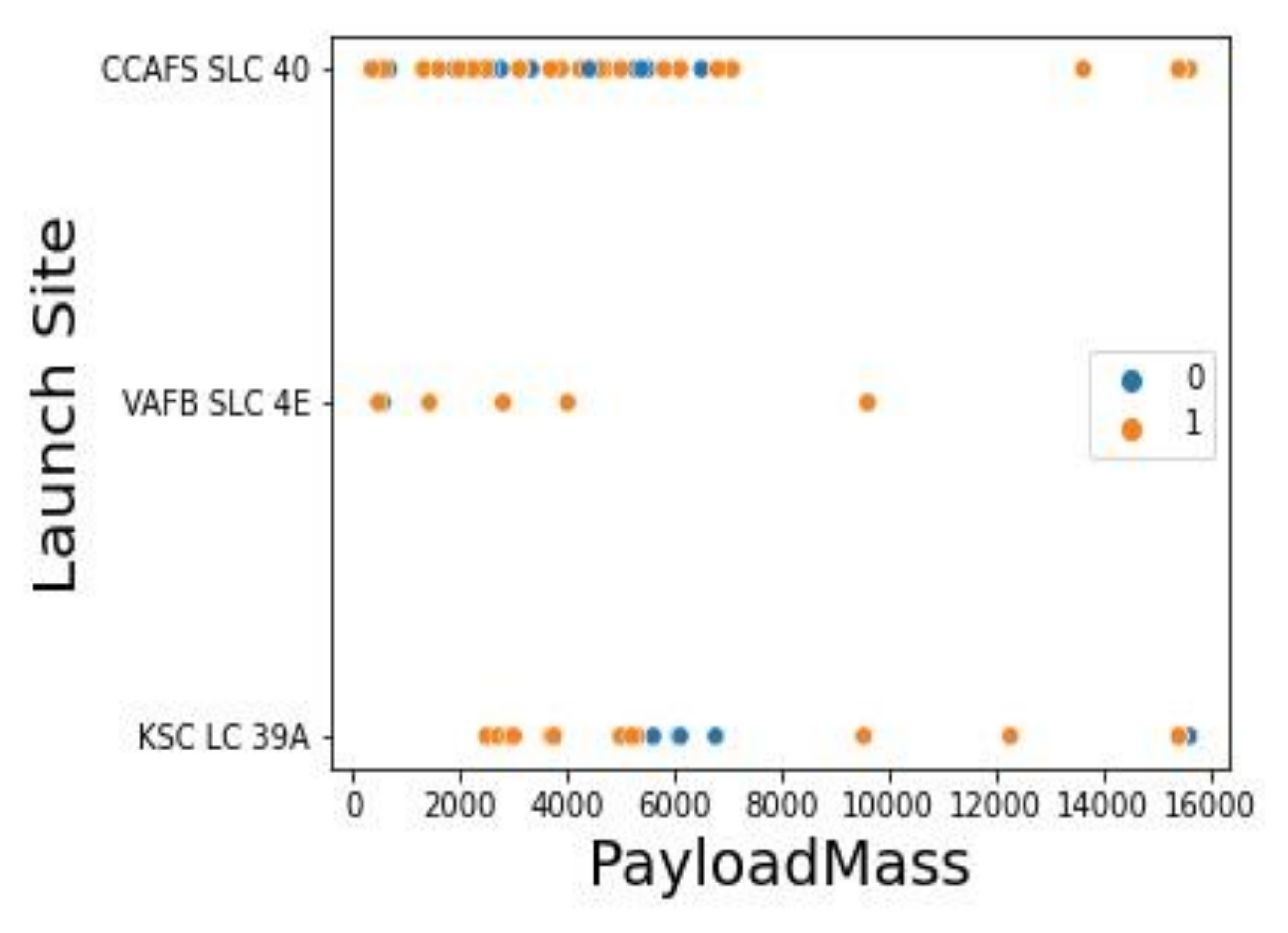
Scatter Plot of Flight Number vs. Launch Site



- 1 represent successful landing, while 0 represent landing failure.
- Among all the Launch Sites, the landing success rate are higher as the Flight Number increases.
- Launch site CCAFS SLC 40 had the largest number of launches based on Flight Number with 33 successes and 22 failures.
- Launch site VAFB SLC 4E had the least number of launches based on Flight Number with 10 successes and 3 failures.
- Launch site KSC LC 39A had the second largest number of launches based on Flight Number with 17 successes and 5 failures.

Payload vs. Launch Site

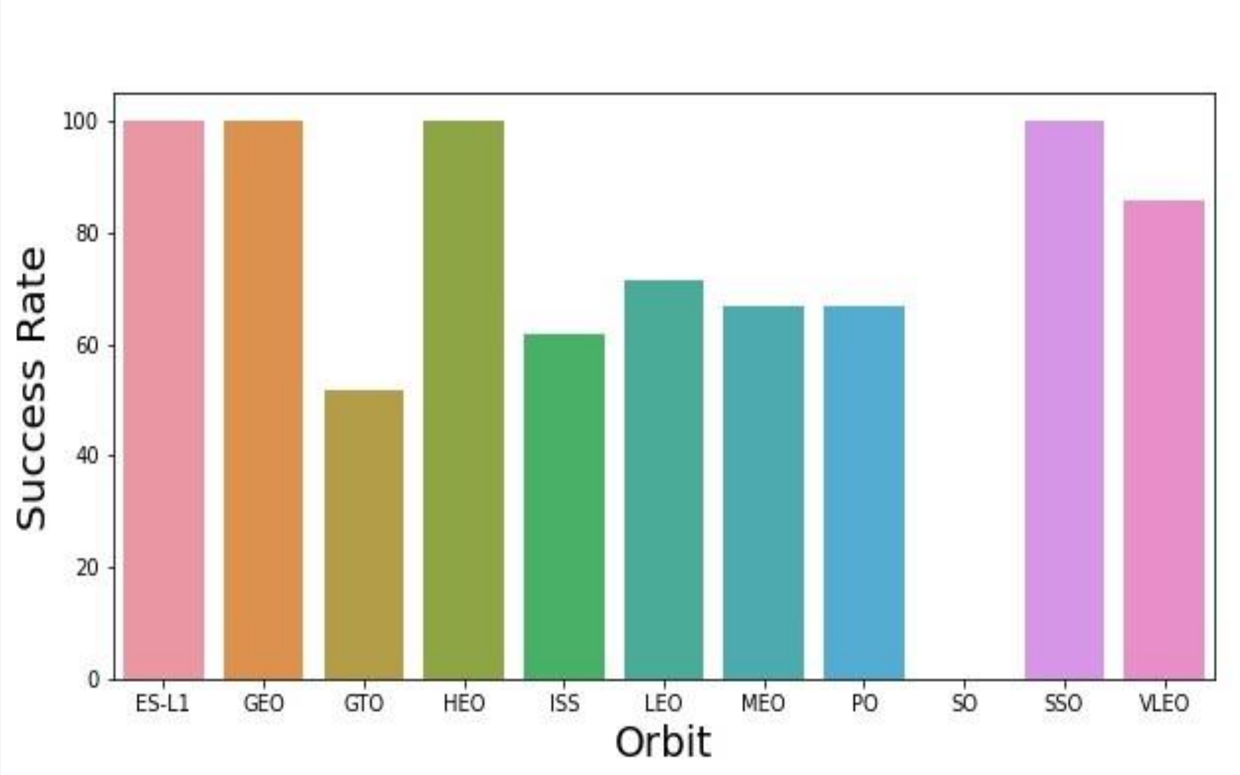
Scatter plot of Payload vs. Launch Site



- 1 represent successful landing, while 0 represent landing failure.
- Among all the Launch Sites, the landing success rate are higher for lower payload masses.
- For launch site, KSC LC 39A, the highest failure occurred for payload mass between 4000 and 8000 Kg.
- Launch Site CCAFS SLC 40 has the highest numbers of launches followed by KSC LC 39A and VAFB SLC 4E.

Success Rate vs. Orbit Type

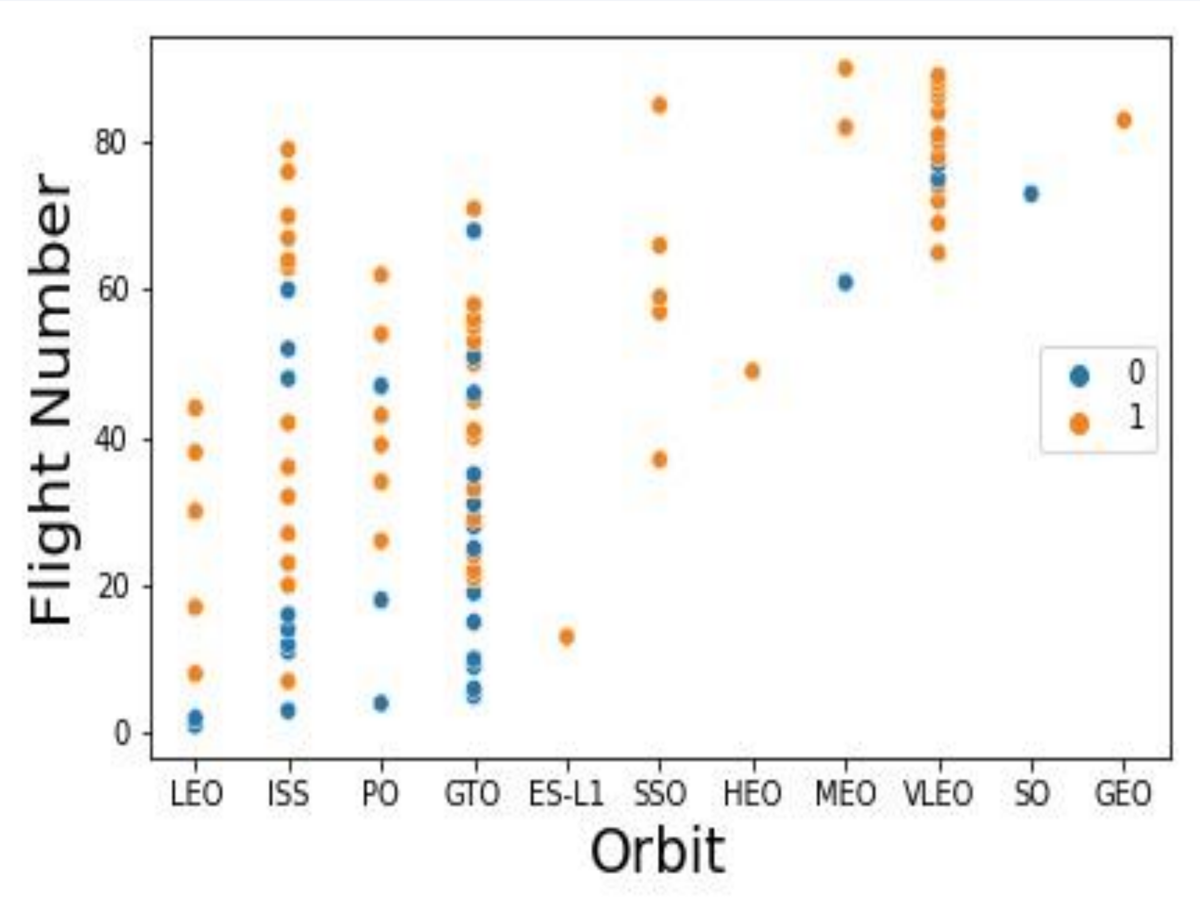
Bar chart for the success rate of each orbit type



- Rockets with Orbit types of ES-L1, GEO, HEO, and SSO have 100 % landing success rate.
- Rockets with Orbit type of VLEO have success rate greater than 80 %.
- Rochets with Orbit types of ISS, LEO, MEO, and PO have landing success rate above 60 %.
- Rockets with orbit type of GTO have landing success rate greater than 50 %.
- Rockets with Orbit type SO have 0 % landing success rate.

Flight Number vs. Orbit Type

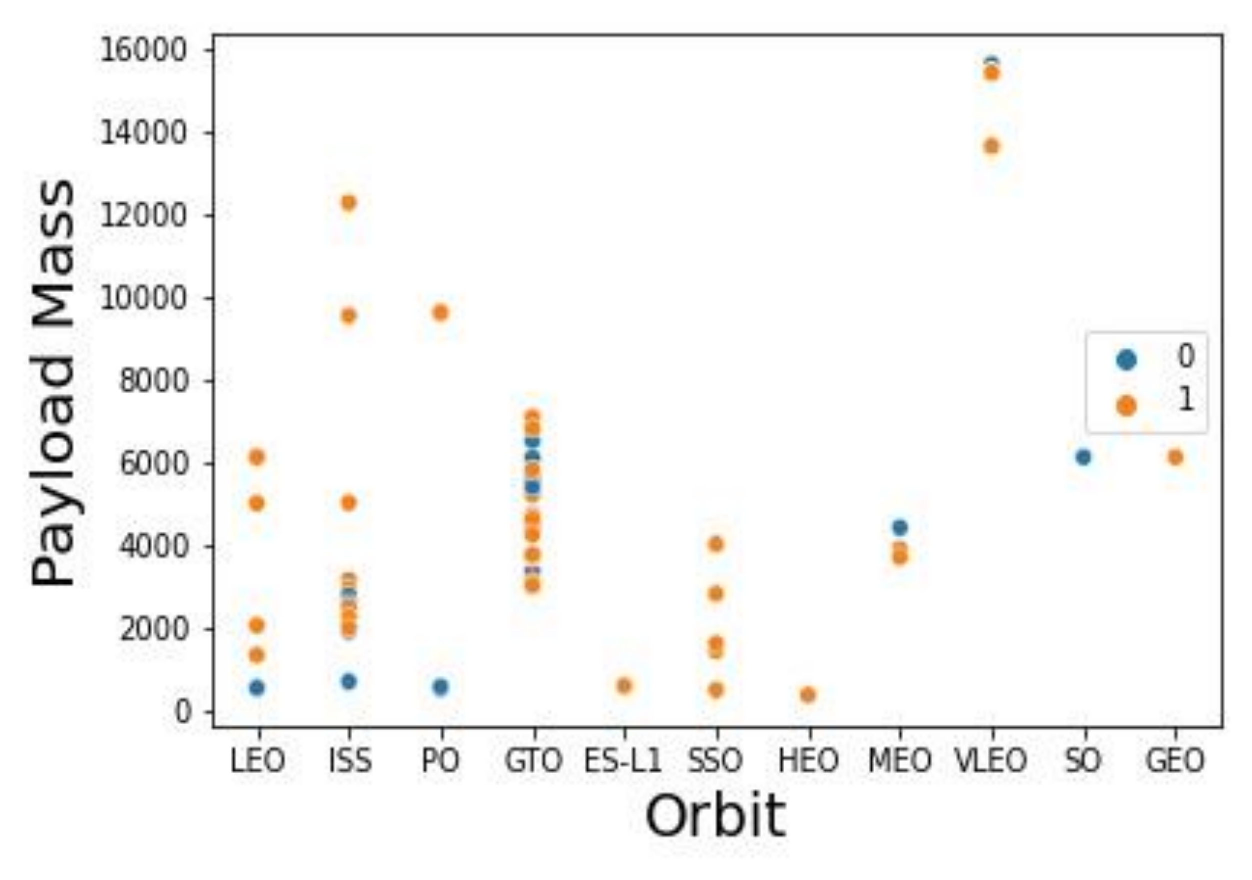
Scatter point of Flight number vs. Orbit type



- 1 represent successful landing, while 0 represent landing failure.
- Generally, it appears that the higher the Flight number of each orbit type, the higher the landing success.

Payload vs. Orbit Type

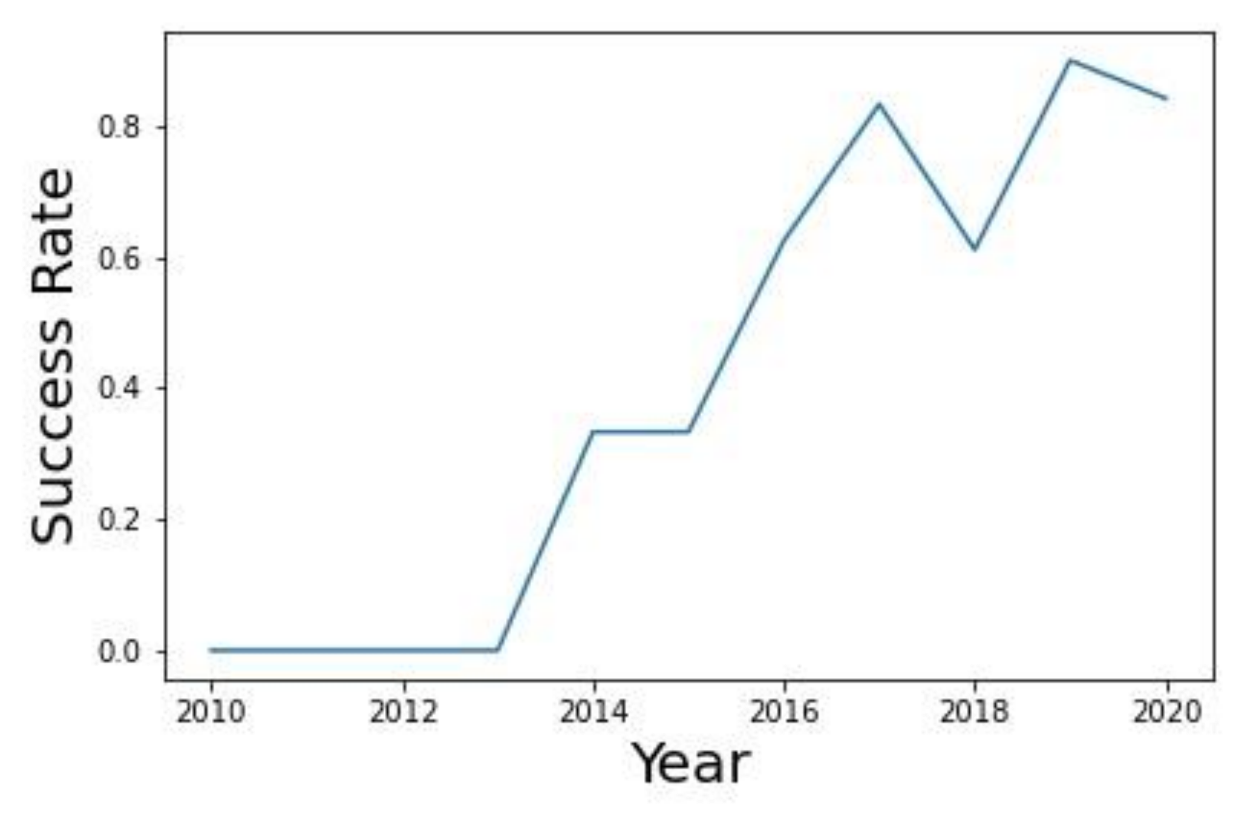
scatter point of payload vs. orbit type



- 1 represent successful landing, while 0 represent landing failure.
- Besides Rockets with Orbit types of GTO, MEO, VLEO and SO, landing success appears to increase with increase in payload mass.
- Rockets with Orbit types ES-L1, SSO, HEO, and GEO have 100 % landing success.
- Overall, there has been more successful landing than failure.

Launch Success Yearly Trend

Line chart of yearly average success rate



- The net success rate since 2013 kept increasing till 2020.
- The success rate between 2010 and 2013 remained at a constant rate of 0 %.
- There was a fall in the success rate between 2017 and 2018 followed by a subsequent rise in the succeeding years.
- Between 2014 and 2015, the success rate plateaued at a value of about 0.38 (38 %).

All Launch Site Names

Names of the unique launch sites

```
%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

- To select unique launch sites, **DISTINCT** is used in the **SELECT** statement

Launch Site Names Begin with 'CCA'

- **5 records where launch sites begin with `CCA`**

```
%%sql
select * from SPACEXTBL 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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Use **LIKE** and the "%" wildcard in the **WHERE** clause and limit the result to 5.

Total Payload Mass

- **The total payload carried by boosters from NASA**

```
%%sql
select sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTBL
where Customer = "NASA (CRS)";
```

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

Total_Payload_Mass
45596

- Use **SUM()** function and set Customer to "NASA (CRS)" in the **WHERE** clause

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.

```
%%sql
select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL
where Booster_Version = "F9 v1.1";
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Use the **AVG()** function and set Booster_Version to "v1.1" in the **WHERE** clause.

First Successful Ground Landing Date

- **Find the date of the first successful landing outcome on ground pad**

```
%%sql
select MIN(Date) as Date_of_First_Successful_Landing from SPACEXTBL
where "Landing _Outcome" = "Success (ground pad)";
```

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

Date_of_First_Successful_Landing
2015-12-22

- Since "first date" means the "minimum date" we use the **MIN()** function and set Landing outcome to "Success (ground pad)" in the **WHERE** clause

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Names of boosters which have successfully landed on drone ship and had 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;
```

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

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

- Set Landing Outcome to "Success (drone ship) and use "**BETWEEN** to filter the payload mass in the **WHERE** clause

Total Number of Successful and Failure Mission Outcomes

- **The total number of successful and failure mission outcomes**

```
%%sql
select count(Mission_Outcome) from SPACEXTBL
where Mission_Outcome like "Success%";
```

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

count(Mission_Outcome)
100

```
%%sql
select count(Mission_Outcome) from SPACEXTBL
where Mission_Outcome like "Failure%";
```

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

count(Mission_Outcome)
1

- For successful mission outcome, use like and wildcard in the **WHERE** clause to filter out mission outcome with Success.
- For successful mission outcome, use like and wildcard in the **WHERE** clause to filter out mission outcome with Failure.

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

```
%%sql
select booster_version from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * 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

- Use the **MAX()** function in the subquery to filter out the booster versions with the maximum payload mass

2015 Launch Records

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

```
%%sql
select Month, "Landing _Outcome", Booster_Version, Launch_Site from SPACEXTBL
where "Landing _Outcome" = "Failure (drone ship)" and Year = 2015;
```

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

Month	Landing _Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
4	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Use the **WHERE** clause to filter out the desired results.

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**

```
%%sql
select "Landing _Outcome", count("Landing _Outcome")
as Count from SPACEXTBL
where date between "2010-04-06" and "2017-03-20"
group by "Landing _Outcome"
order by count("Landing _Outcome") DESC;
```

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

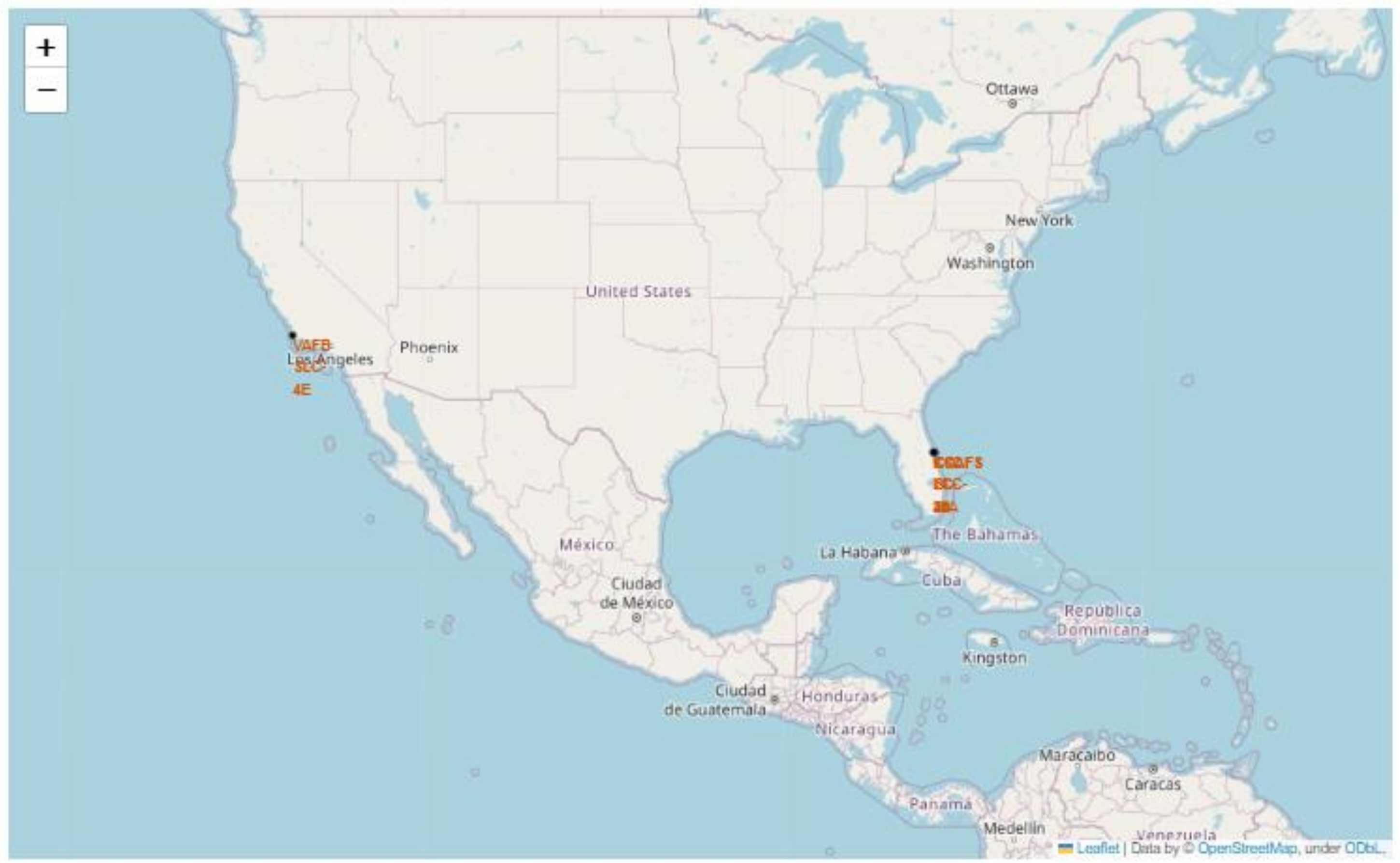
Landing _Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Use **GROUP BY** and **ORDER BY** (in descending order) to filter the “landing outcome” and “count of the landing outcome” in the **WHERE** clause.

Section 3

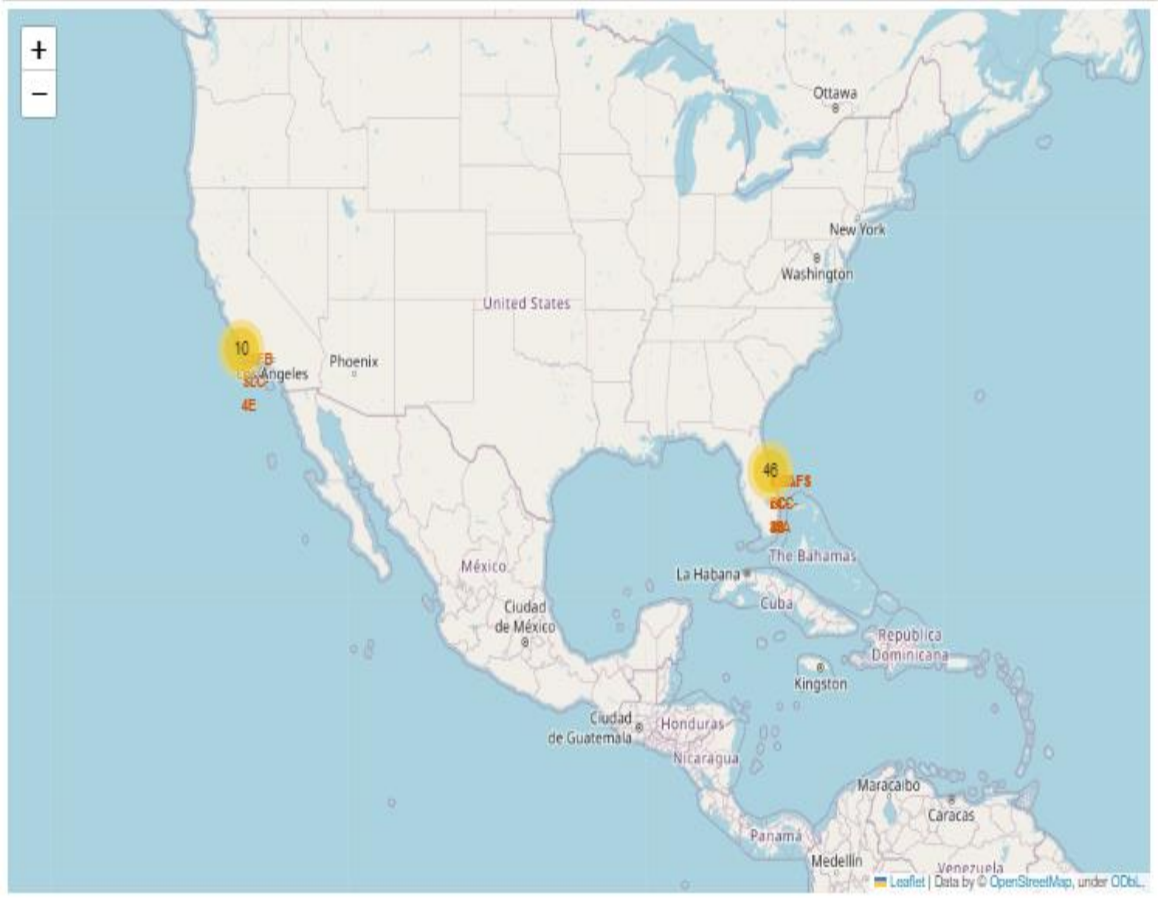
Launch Sites Proximities Analysis

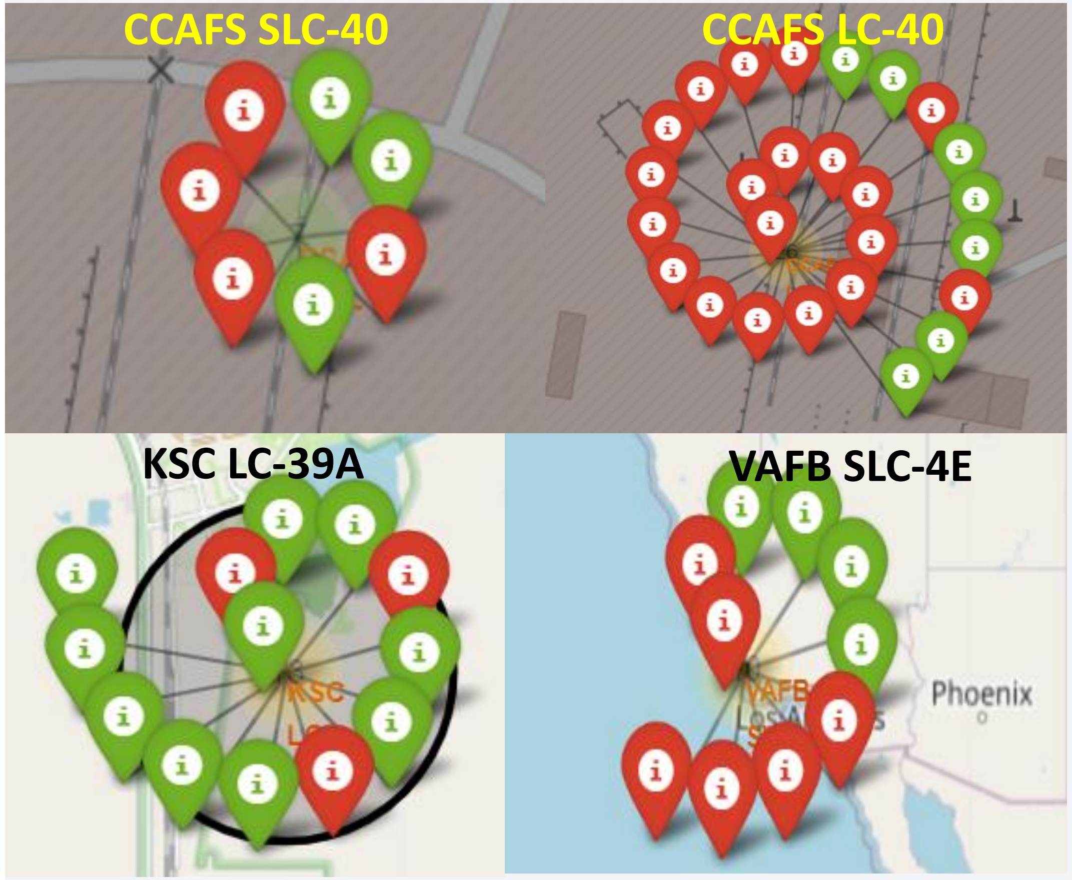
All launch sites' location markers on a global map



- All the launch sites are located close to coastlines.
- Launch site VAFB SLC-4E is located close to the Vandenberg State Marine Reserve.
- Launch sites CCAFS SLC-40, CCAFS LC-40 and KSC LC-39A are close to each other and in close proximity to Merritt Island National Wildlife Refuge.

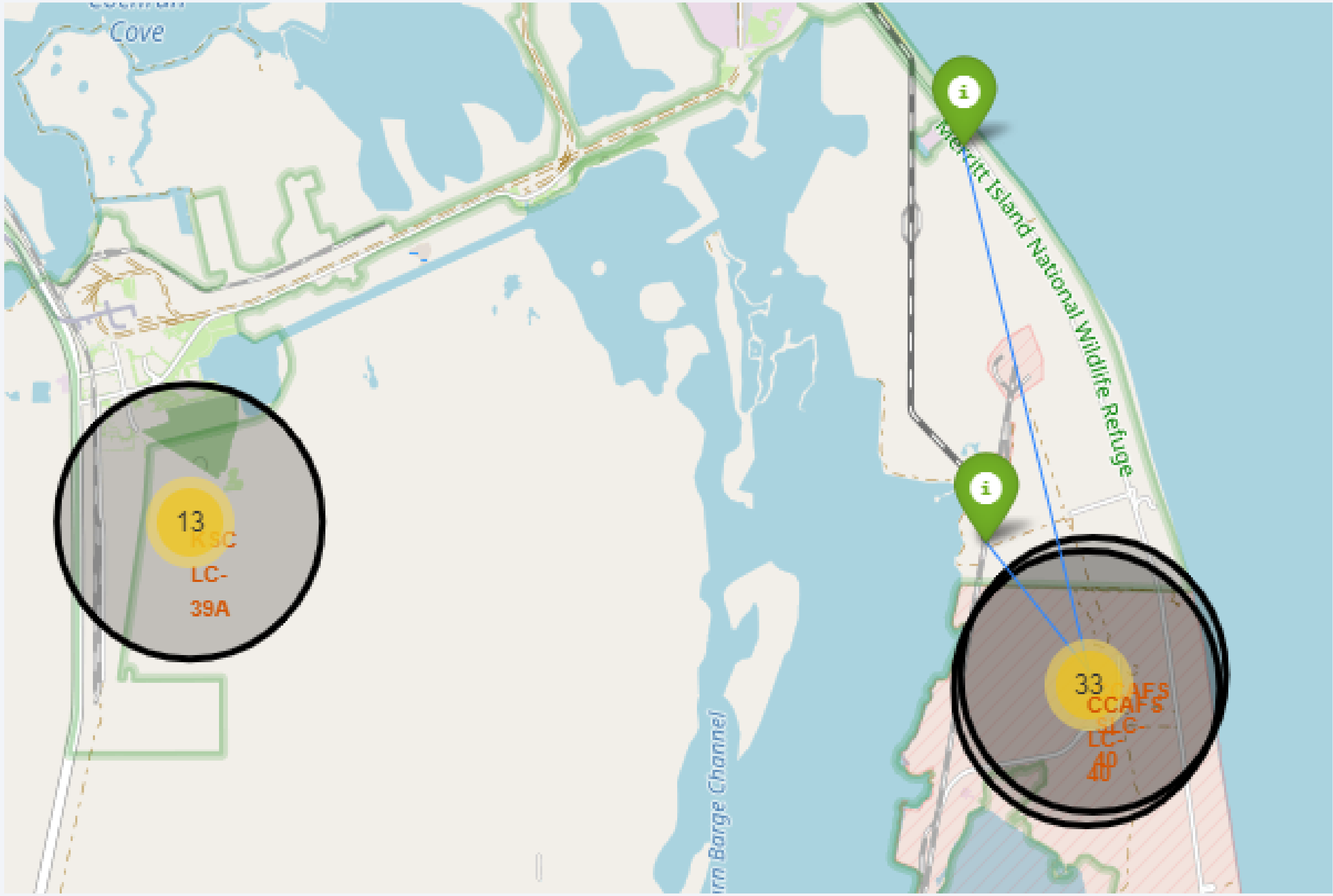
Color-labeled launch outcome from each site location on a global map





The red markers show launch sites with failure while the green markers show launch sites with success. KSC LC-39A has the highest success rate.

Selected launch sites and there proximities to railways, highways, coastlines, with distance calculated and displayed

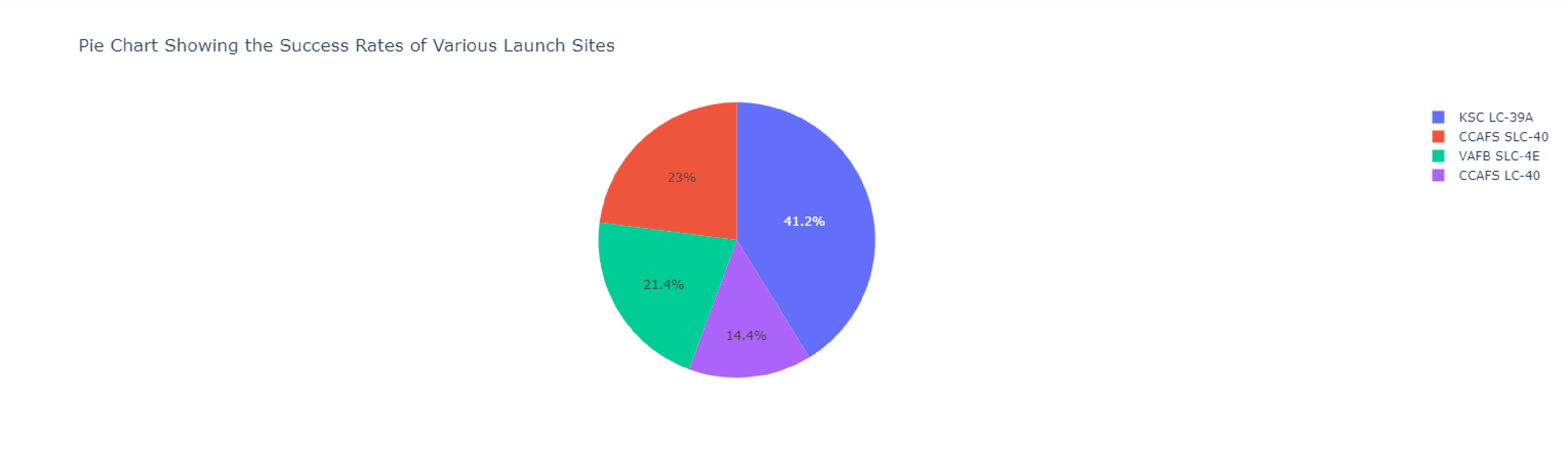


- Launch sites are in close proximity to railways, highways, and coastlines but keep certain distance away from cities.

Section 4

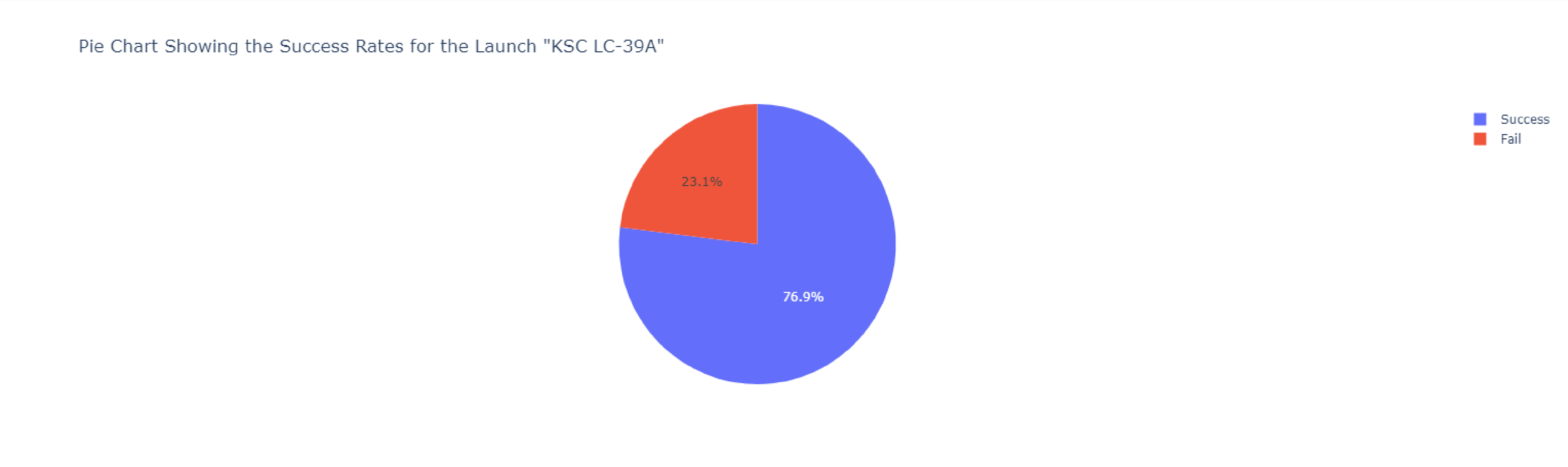
Build a Dashboard with Plotly Dash

Launch success count for all sites



- KSC LC-39A has the highest success rate followed by CCAFS SLC-40, VAFB SLC-4E and CCAFS LC-40.

Pie chart for the launch site with highest launch success ratio



- KSC LC-39A has the highest success ratio, with success rate of 76.9 % and failure rate of 23.1 %.

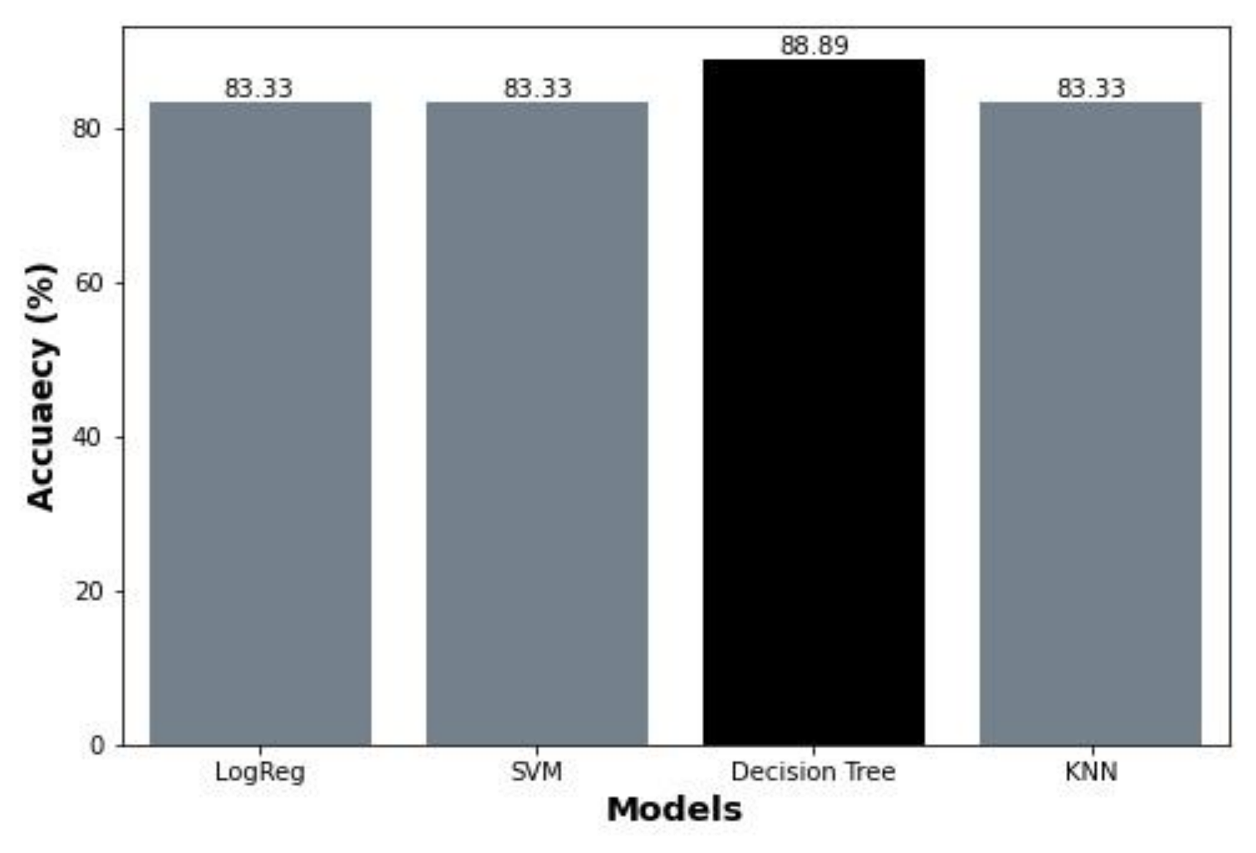
Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- From the charts above, it is easy to see that the smaller payload masses had higher success rate than the larger payload masses.

Section 5

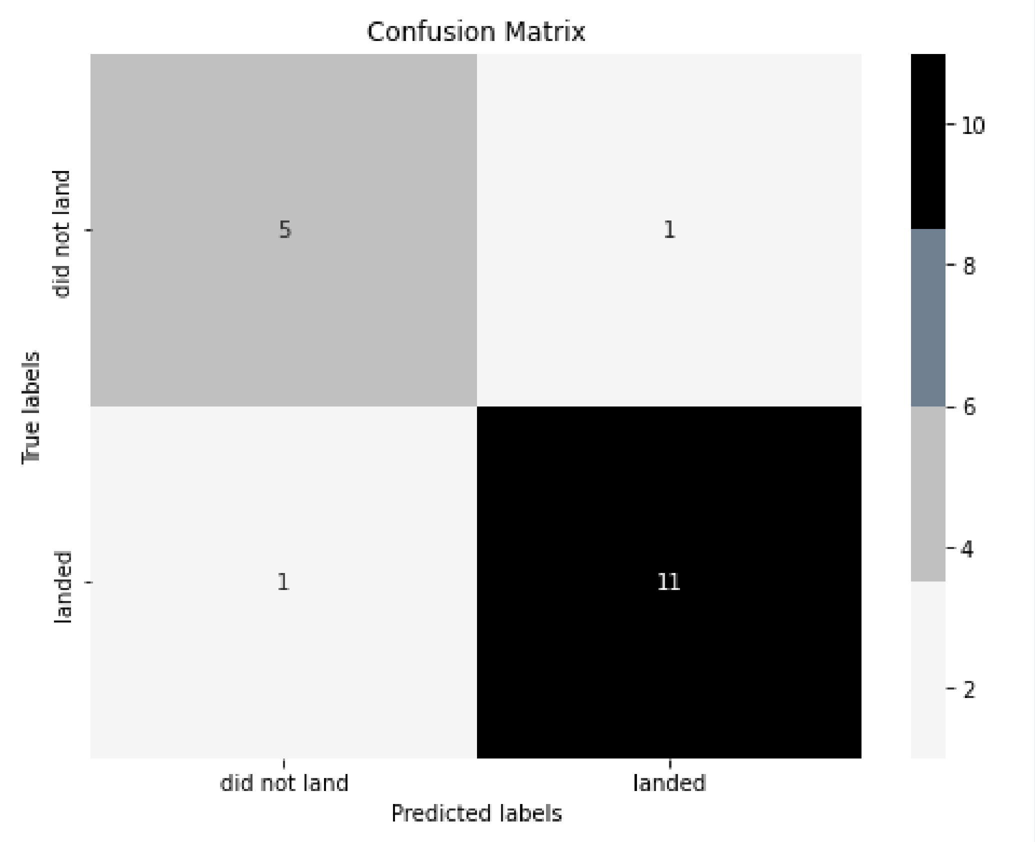
Predictive Analysis (Classification)

Classification Accuracy



- Among the 4 classification models evaluated, decision tree classifier has the highest classification accuracy of 88.89 %.
- Others have classification accuracy of 83.33 %.

Confusion Matrix of the Best Performing Model (Decision Tree)



- **TP = 11, FP = 1, TN = 5, FN = 1**
- **Sensitivity** = TP/(TP+FN) = 0.92
- **Specificity** = TN/(TN + FP) = 0.83
- **Precision** = TP/(TP +FP) = 0.92
- **Accuracy** = (TP+TN)/(TP+TN+FP+FN) = 0.89
- From the metrics above, its easy to see that there is a balance in the ability of the decision tree to classifier to identify both successful landing (high sensitivity) and failed landing (high specificity).

Conclusions

- All the models - logistic regression, decision tree, k-nearest neighbor, and support vector machine exhibited high predictive accuracy of greater than 83%. However, decision tree had the highest predictive accuracy of over 88%.
- Factors such as launch site, payload mass, and orbit contribute significantly to the successful landing of a Falcon 9 rocket.
- The success rates of SpaceX launches increased as years went by as they eventually perfect the art of rocket launching.
- Among all the Launch Sites, the landing success rate are higher for lower payload masses. For launch site, KSC LC 39A, the highest failure occurred for payload mass between 4000 and 8000 Kg. Launch Site CCAFS SLC 40 has the highest numbers of launches followed by KSC LC 39A and VAFB SLC 4E.
- KSC LC 39A had the most successful launches from all the launch sites.
- Rockets with Orbit types of ES-L1, GEO, HEO, and SSO have 100 % landing success rate. Rockets with Orbit type of VLEO have success rate greater than 80 %. Rochets with Orbit types of ISS, LEO, MEO, and PO have landing success rate above 60 %. Rockets with orbit type of GTO have landing success rate greater than 50 %. While, rockets with Orbit type SO have 0 % landing success rate.

Appendix

- Link to the cleaned dataset used for the interactive dash-plotly charts: https://github.com/OnahPmi/My_Coursera_Repo/blob/d5b77e783860df1881edfb2928a567734b6549c9/Applied%20Data%20Science%20Capstone%20Project/spacex_launch_dash.csv
- Link to the cleaned dataset used for the folium maps: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
- Link to the one-hot encoded dataset used for building the machine learning models: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv
- Link to my github repository where all the codes are hosted: https://github.com/OnahPmi/My_Coursera_Repo/tree/main/Applied%20Data%20Science%20Capstone%20Project

Thank you!