IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AMP
12.20.24

Outline

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

Executive Summary

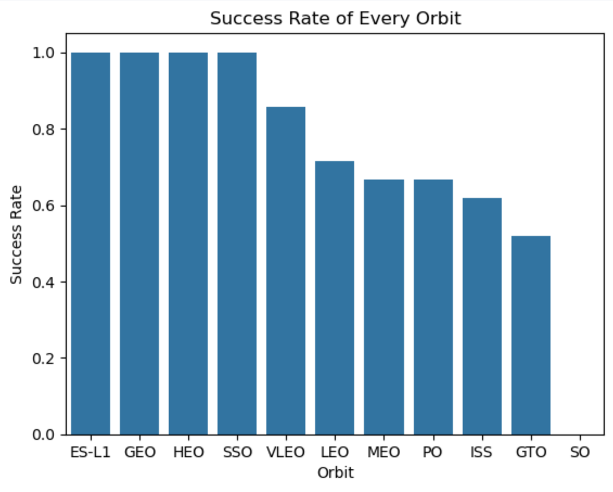
Project Overview: The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful is SpaceX. SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

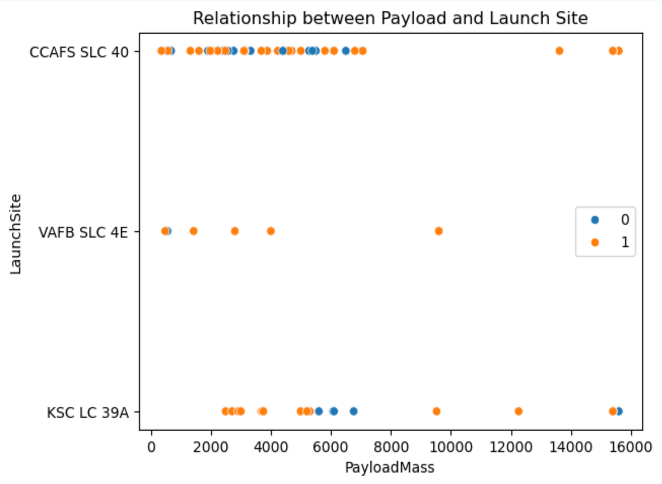
Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Key Insights:

- Launch site VAFB SLC 4E had more successful outcomes with payload under 5000Kg
- Orbits with more successful outcomes are ES-L1, GEO, HEO and SSO
- Launch site KSC LC-39A had more successful outcomes in the full range of payload
- Launch site KSC LC-39A had 41.7% of total site successful outcomes, by itself had 76.9% of successful outcomes
- Overall, there are more successful outcomes with payloads under 8000kg

Details:





Introduction

The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful is SpaceX. SpaceX's accomplishments include: Sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access. Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - -Connect to SpaceX API to get data about all the launches registered
  - -Web scrapping on Wikipedia page *'List of Falcon 9 and Falcon Heavy Launches'*
- Perform data wrangling
  - -Perform EDA on collected data to understand the features available
  - -Determine predictor and target variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - -Train and evaluate 4 different machine learning models

Data Collection

API Data Collection

Send request to SpaceX API from Python

Get a Json file from SpaceX with relevant data

Transform Json file to DataFrame

SpaceX API: https://api.spacexdata.com/v4/launches/past

Web Scrapping Data Collection

Send request to SpaceX Wikipedia page from Python

Get a HTML file from SpaceX Wikipedia page with relevant data

Transform HTML file to DataFrame

SpaceX Wikipedia page: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

1
- Import relevant library for data collection: *requests*

2
- Save API URL in a variable: *spacex_url*

3
- Send a request to SpaceX API: *response.get(spacex_url)*

4
- Get Json file with relevant data

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/1.%20Spacex%20Data%20Collection%20API%20v2.ipynb

Data Collection - Scraping

1

- Import relevant library for Web Scrapping data collection:

requests

2

- Save Wikipedia SpaceX URL in a variable:

spacex_url

3

- Send a request to SpaceX API:

response.get(spacex_url)

4

- Get HTML file with relevant data

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/2.%20Spacex%20Webscraping.ipynb

Data Wrangling – SpaceX API

1

- Convert Json file to DataFrame:

pandas.json_normalize(data)

2

- Select important IDs of DataFrame to use again the API and get relevant data:

'rocket', 'payloads', 'lanchpad', 'cores', etc.

3

- Create a new DataFrame with relevant features:

'BoosterVersion', PayloadMass', 'Orbit', etc.

4

- Handle missing values on new DataFrame:

'.isnull().sum()', '.replace(np.nan, value)'

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/3.%20Spacex%20Data%20Wrangling%20v2.ipynb

Data Wrangling – Scrapping

1
- Convert HTML file to Python object:
BeautifulSoup(response.content, ‘html.parser’)

2
- Extract feature names from HTML table header:
‘soup.find_all(‘table’), ‘.join(row.contents)’

3
- Create a new DataFrame with relevant features:
‘Flight No.’, ‘Launch Site’, ‘Payload’, etc.

4
- Fill in the new DataFrame:
‘List Comprehension’

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/2.%20Spacex%20Webscraping.ipynb

EDA with Data Visualization

The following charts were plotted in the EDA process:

- Relationship between Flight Number and Payload Mass (Scatterplot)
- Relationship between Flight Number and Launch Site (Catplot)
- Relationship between Payload and Launch Site (Scatterplot)
- Relationship between Success Rate and Orbit Type (Bar Plot)
- Relationship between Flight Number and Orbit Type (Scatterplot)
- Launch Success Yearly Trend (Lineplot)

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/5.%20EDA%20Dataviz%20v2.ipynb

EDA with SQL (1/2)

The following SQL queries were made in the EDA process:

- Display the names of the unique launch sites in the space mission: (%sql select distinct(Launch_Site) from SPACEXTABLE)
- Display 5 records where launch sites begin with the string 'CCA’: (%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5)
- Display the total payload mass carried by boosters launched by NASA (CRS): (%sql select sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTABLE where Customer = 'NASA (CRS)’)
- Display average payload mass carried by booster version F9 v1.1: (%sql select avg(PAYLOAD_MASS__KG_) as Avg_Payload_Mass from SPACEXTABLE where Booster_Version = 'F9 v1.1’)
- List the date when the first successful landing outcome in ground pad was achieved: (%sql select min(Date) as First_Landing_Success from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)’)
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000: (%sql select Booster_Version from SPACEXTABLE where (Landing_Outcome = 'Success (drone ship)') and (PAYLOAD_MASS__KG_ between 4000 and 6000)

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/4.%20EDA%20SQL.ipynb

EDA with SQL (2/2)

The following SQL queries were made in the EDA process:

- List the total number of successful and failure mission outcomes: (%sql select Mission_Outcome, count(Mission_Outcome) as Count from SPACEXTABLE group by Mission_Outcome)
- List the names of the booster_versions which have carried the maximum payload mass: (%sql select distinct(Booster_Version) from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
- 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: (%sql select substr(Date, 6, 2) as Month, substr(Date, 0,5) as Year, Booster_Version, Landing_Outcome, Launch_Site from SPACEXTABLE where (Landing_Outcome = 'Failure (drone ship)') and (Year = '2015’)
- 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 SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count Desc)

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/4.%20EDA%20SQL.ipynb

Build an Interactive Map with Folium

The following analysis with using Folium was made:

- Mark all launch sites on a map:

These location marks of the 4 different launch sites on a map, help us to know the specific region or city where the launch sites are located

- Mark the success/failed launches for each site on the map:

Adding green and red color marks for each successful and unsuccessful launch, we can easily understand the behavior of each launch site

- Calculate the distances between a launch site to its proximities:

Calculating distances of launch site proximities can give us a better understanding of the regions each launch site is located

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/6.%20Launchsite%20Location%20v2.ipynb

Build a Dashboard with Plotly Dash

The following features were added to the Plotly Dashboard:

- Add a Launch Site Drop-down Input Component:

This dropdown menu filters data according to each launch site, or aggregate all launch sites

- Add a callback function to render success-pie-chart based on selected site dropdown:

This pie chart shows the segregation between successful and unsuccessful outcomes

- Add a Range Slider to Select Payload:

This range slider filters the payload range

- Add a callback function to render the success-payload-scatter-chart scatter plot

This scatterplot shows the relationship between Payload Mass and Success

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/7.%20Spacex%20Dash%20app.py

Predictive Analysis (Classification)

1
- Assign the Y variable to a NumPy array :
Y= data['Class'].to_numpy()

2
- Standardize depended features and assigned to X:
X= transform.fit(X).transform(X.astype(float))

3
- Split the data for training and testing:
train_test_split

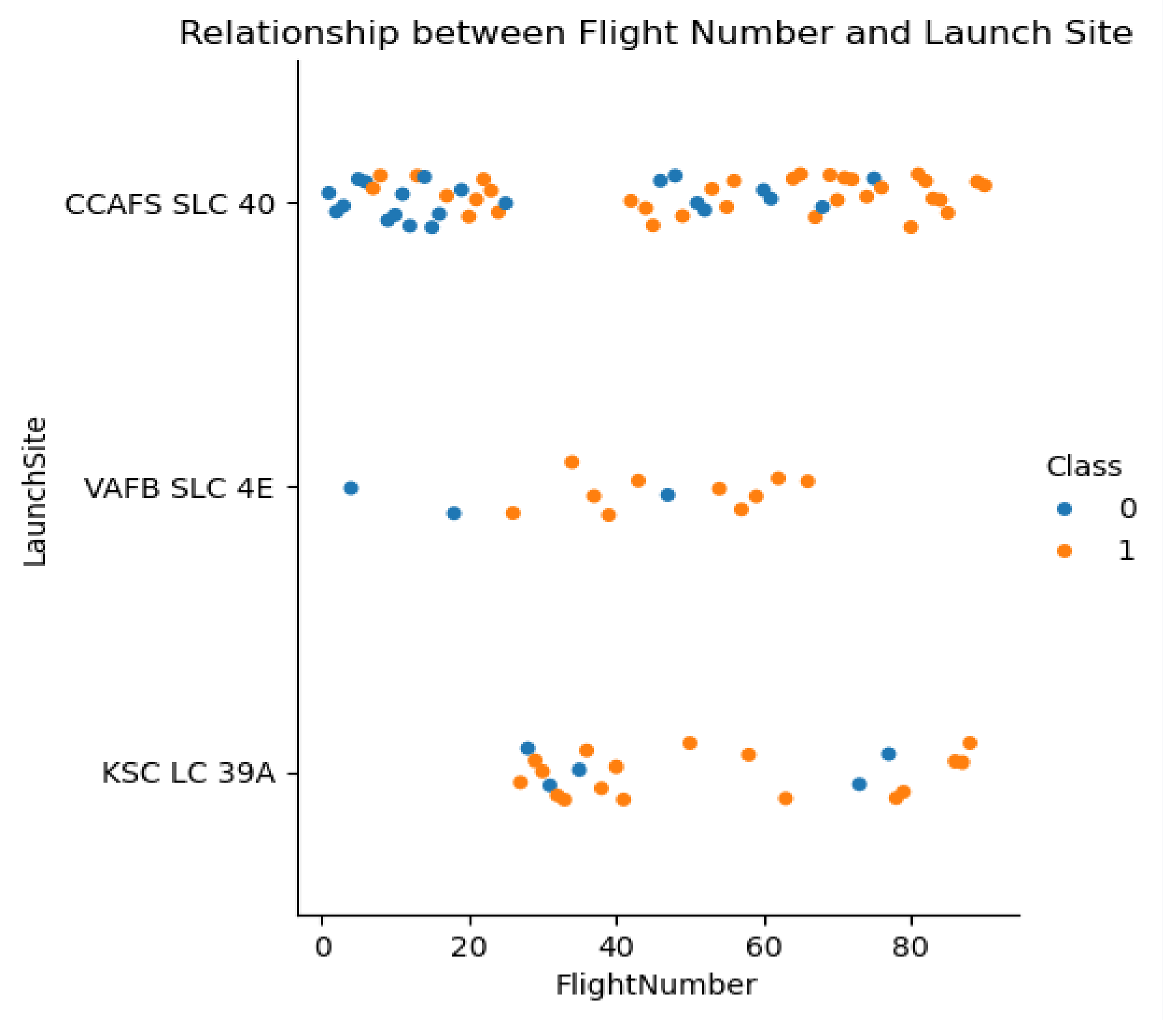
4
- Create ML models and evaluate them:
Logistic Regresion, SVM, Decision Tree, KNN

https://github.com/amp-18/IBM-Applied-Data-Science-Capstone/blob/main/8.%20SpaceX%20Machine%20Learning%20Prediction.ipynb

Section 2

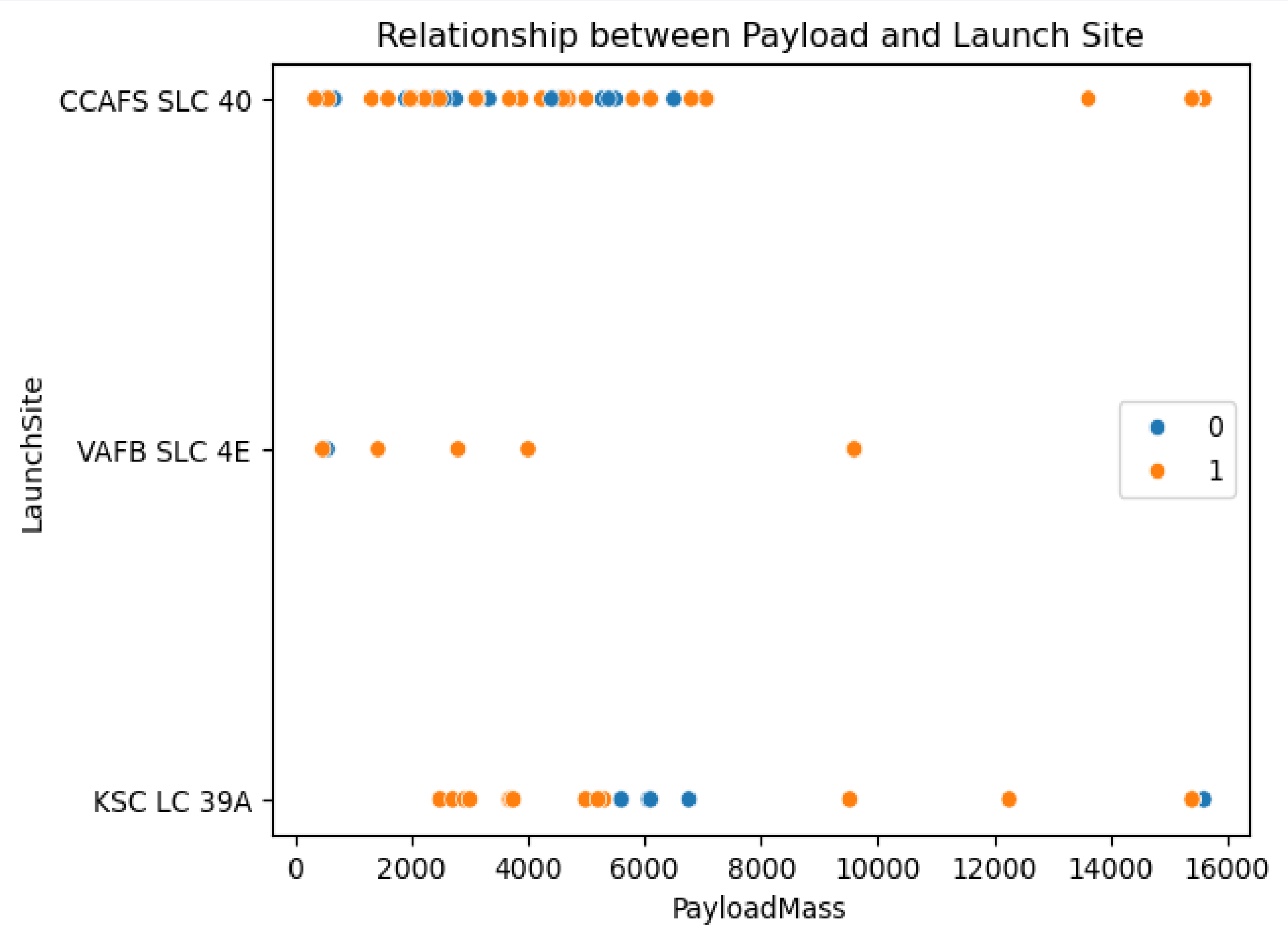
Insights drawn from EDA

Flight Number vs. Launch Site



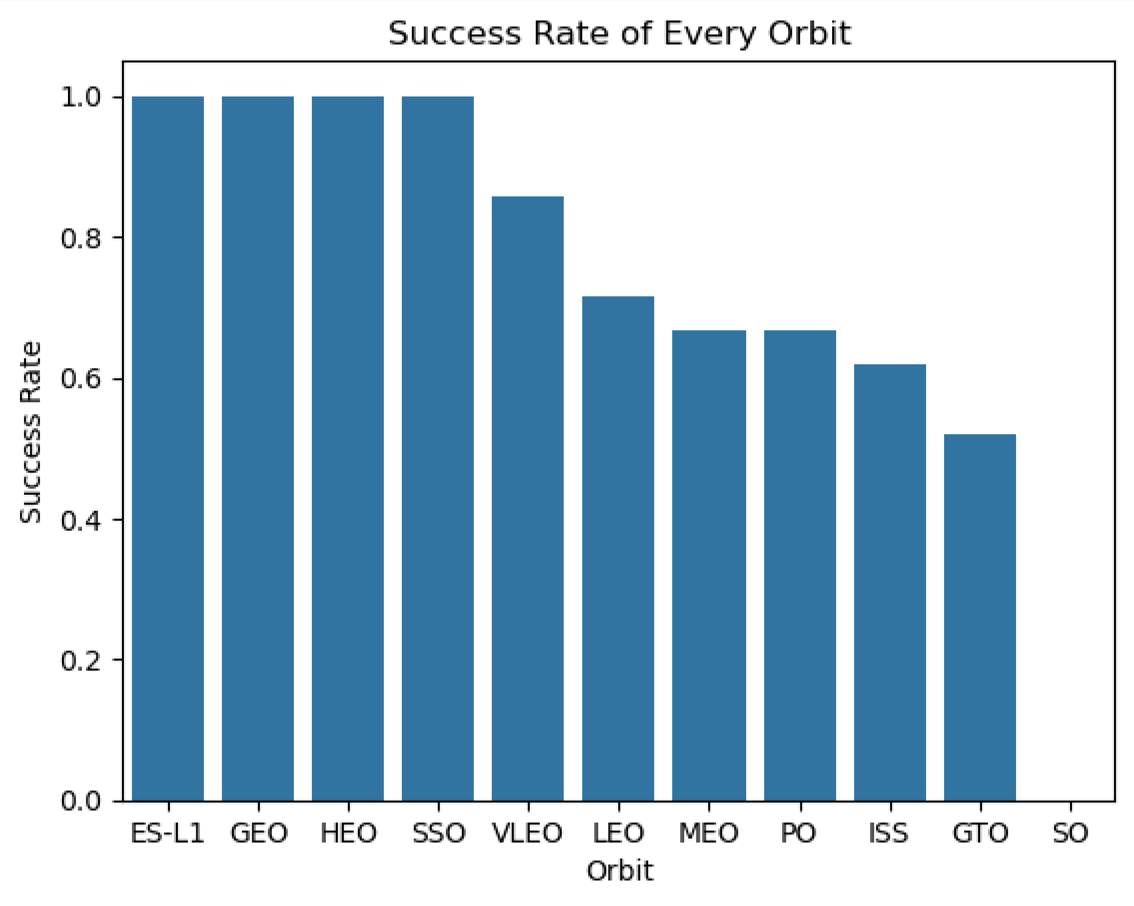
We see for the launch site VAFB SLC 4E, it seems that as the Flight Number increases the successful launch outcome increases, however for the rest launch sites, data is spread between successful and unsuccessful outcomes.

Payload vs. Launch Site



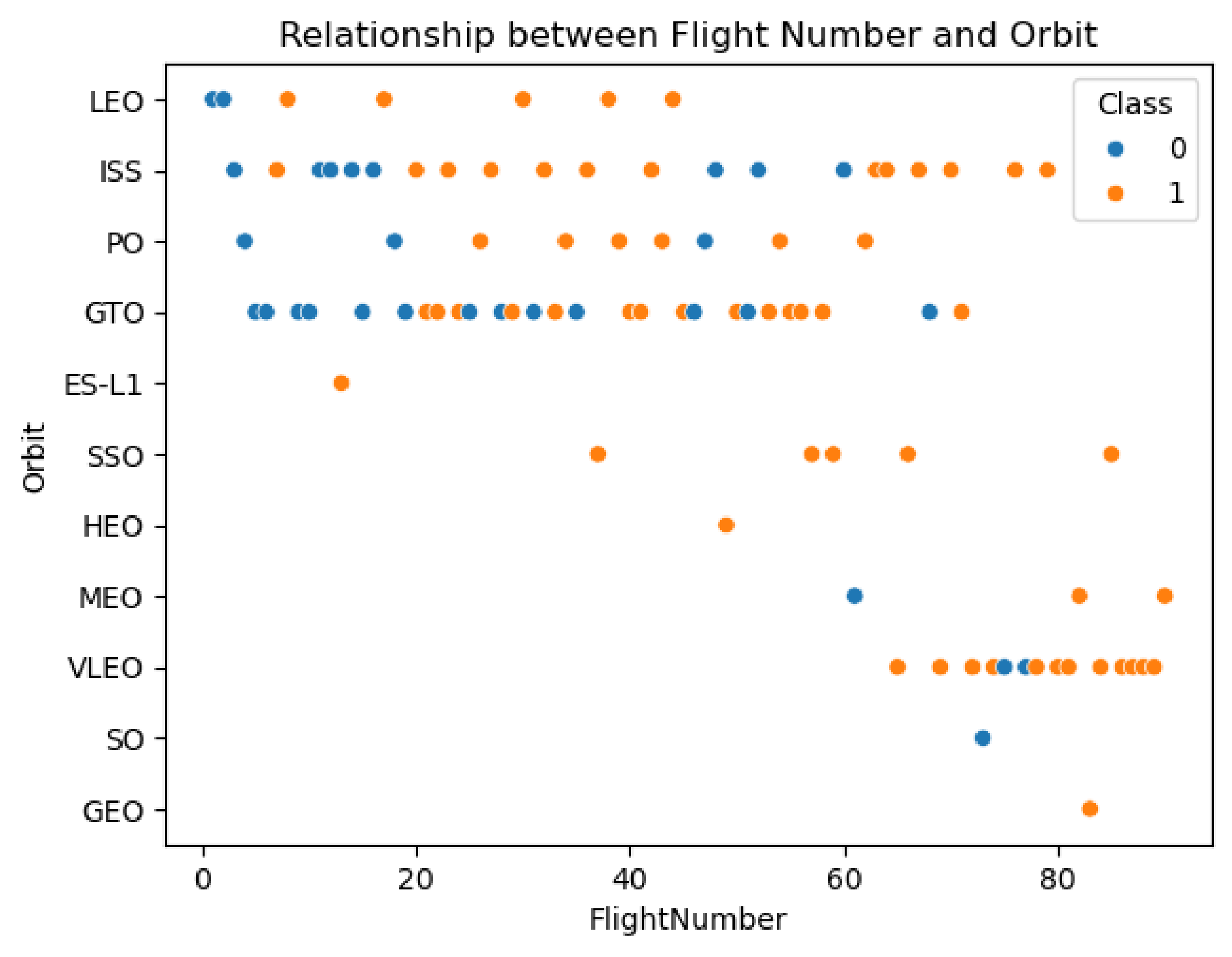
Launch site VAFB SLC 4E shows a positive correlation between Payload Mass and Success Outcomes, important to mention this site has the lowest number of launches than the other sites which Success Outcomes are spread along Payload Mass

Success Rate vs. Orbit Type



The top successful orbits are the ES-L1, GEO, HEO and SSO, the worst successful rate is in the GTO orbit

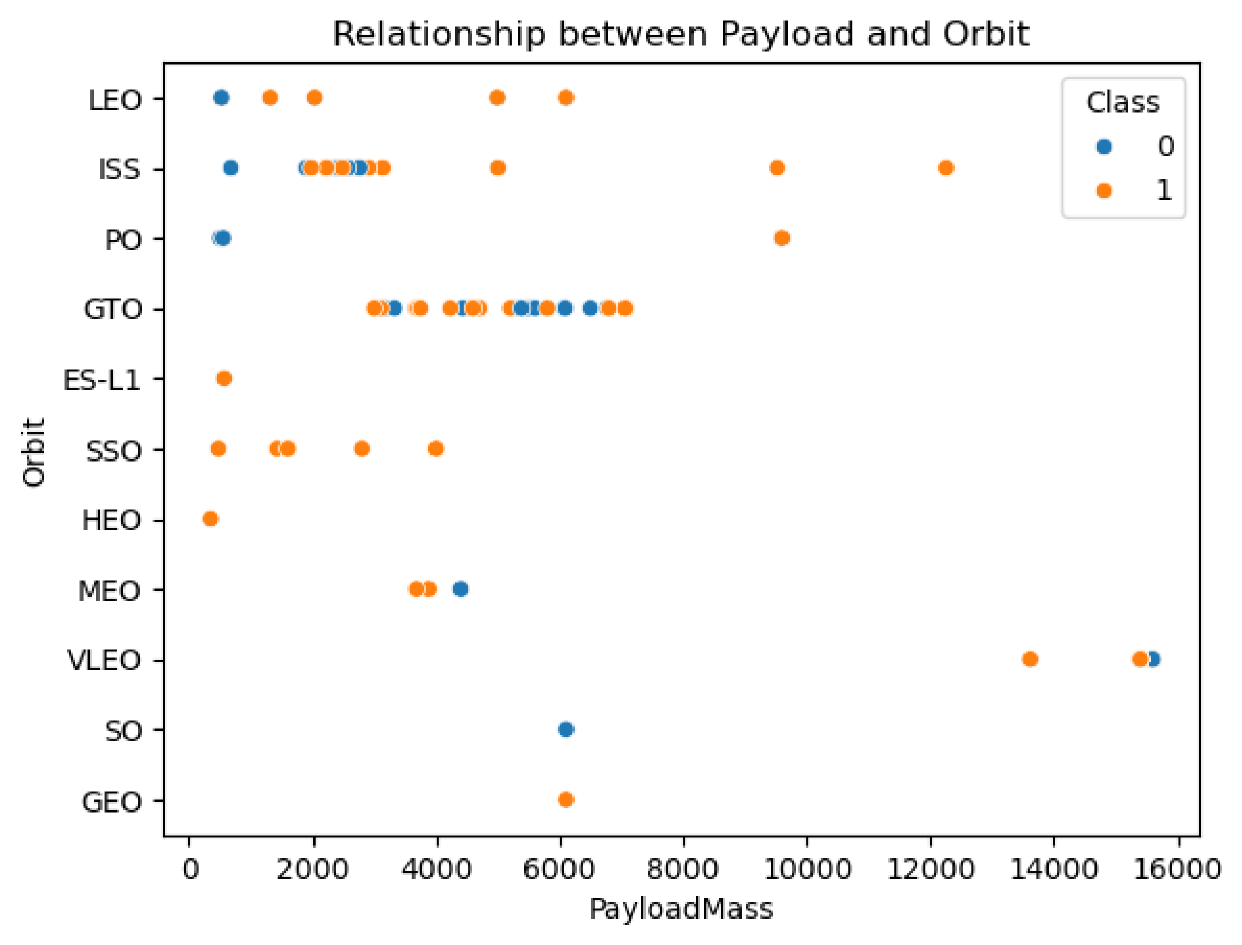
Flight Number vs. Orbit Type



LEO and SSO orbits show a positive correlation between Flights Number and Success.

The data for the relationship between Flight Number and Orbit Type is spread for the most orbits.

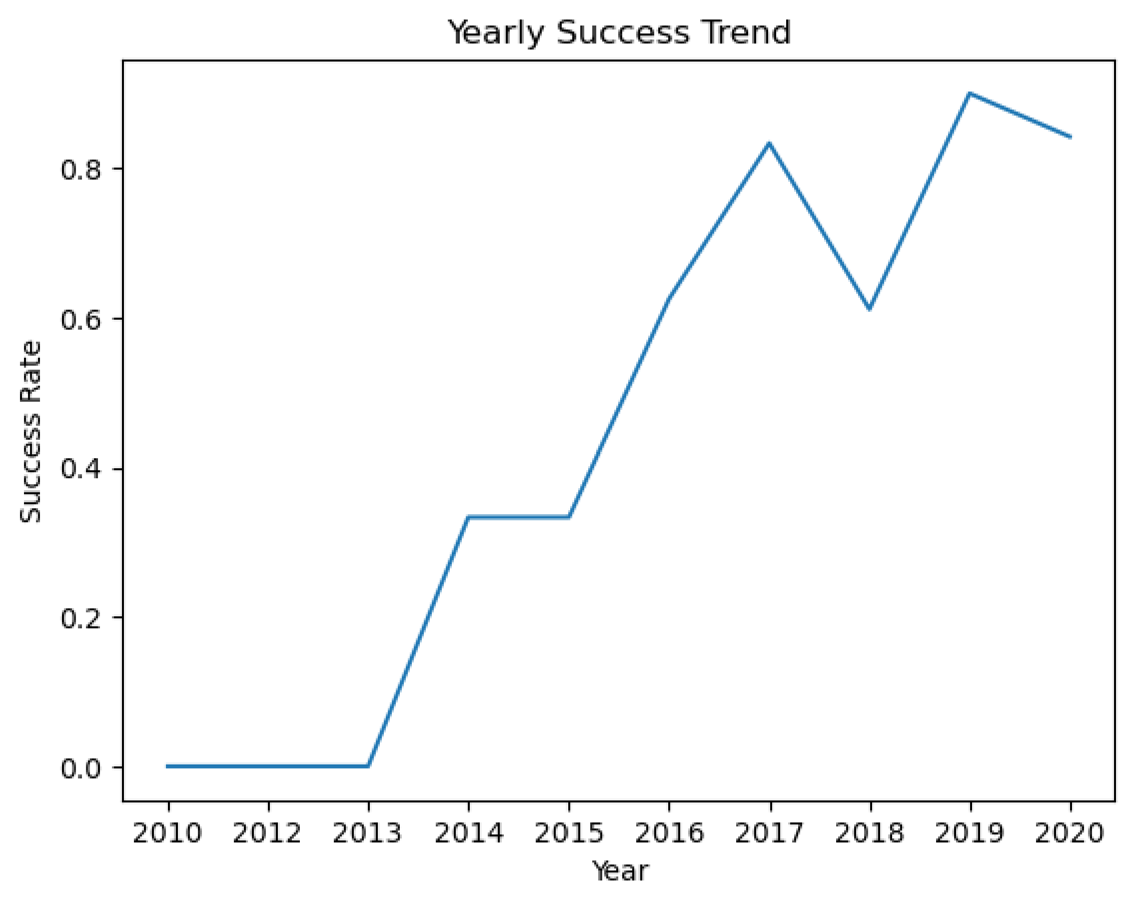
Payload vs. Orbit Type



There is a positive correlation in success between payload and orbit for the orbits LEO, ISS, PO and SSO

However, for GTO orbit launch outcomes are spread along the payload range

Launch Success Yearly Trend



The yearly success trend plot, shows an increment in 2013, then it was stable for 2014 and continue increasing until 2017.

The general trend is positive over years

All Launch Site Names

We found the names of each launch site with the following SQL query:
%sql select distinct(Launch_Site) from SPACEXTABLE

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

The query selected the distinct values in the *Launch_Site* column from the table *SPACEXTABLE*

Launch Site Names Begin with 'CCA'

We found 5 records where launch sites begin with the string ‘CCA’ using the following SQL query: *%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5*

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

The query selected all columns from the table *SPACEXTABLE* where the values in the *Launch_Site* column contain the string *CCA* and it was limited to show only 5 records

Total Payload Mass

We found the total Payload Mass carried by boosters launched by NASA (CRS) using the following SQL query:

```
%sql select sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTABLE where Customer = 'NASA (CRS)'
```

Total_Payload_Mass
45596

The query selected the *Total_Payload_Mass* and sum all the rows from the table *SPACEXTABLE* where customer was *NASA (CRS)*

Average Payload Mass by F9 v1.1

We found the average payload mass carried by booster version F9 v1.1 using the following SQL query:

```
%sql select avg(PAYLOAD_MASS__KG_) as Avg_Payload_Mass from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

Avg_Payload_Mass
2928.4

The query selected the *PAYLOAD_MASS__KG_* column and calculates the average of all rows, from table *SPACEXTABLE* where *Booster_Version* was *F9 v1.1*

First Successful Ground Landing Date

We found the date when the first successful landing outcome in ground pad was achieved using the following SQL query:

```
%sql select min(Date) as First_Landing_Success from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

First_Landing_Success
2015-12-22

The query selected the *Date* column found the minimum value from table *SPACEXTABLE* where *Landing_Outcome* was *Success (ground pad)*

Successful Drone Ship Landing with Payload between 4000 and 6000

We found the names of the boosters which has success in drone ship with payload between 4000 and 6000 using the following SQL query:

```
%sql select Booster_Version from SPACEXTABLE where (Landing_Outcome = 'Success (drone ship)') and (PAYLOAD_MASS__KG_ between 4000 and 6000)
```

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

The query selected the *Booster_Version* column from the table SPACEXTABLE where *Landing_Outcome* was *Success (drone ship)* and *Payload_Mass__KG_* was between *4000kg* and *6000kg*

Total Number of Successful and Failure Mission Outcomes

We found the total number of successful and failure mission outcomes using the following SQL query:

%sql select Mission_Outcome, count(Mission_Outcome) as Count from SPACEXTABLE group by Mission_Outcome

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

The query selected the *Mission_Outcome* and count each category of this column from table SPACEXTABLE and group the data by *Mission_Outcome*

Boosters Carried Maximum Payload

We found the names of the booster_versions which have carried the maximum payload mass using the following SQL query: `%sql select distinct(Booster_Version) from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)`

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 selected the distinct *Booster_Version* from table *SPACEXTABLE* where *PAYLOAD_MASS__KG_* is equal to maximum payload mass

2015 Launch Records

We found the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015 using the following SQL query:

%sql select substr(Date, 6, 2) as Month, substr(Date, 0,5) as Year, Booster_Version, Landing_Outcome, Launch_Site from SPACEXTABLE where (Landing_Outcome = 'Failure (drone ship)') and (Year = '2015')

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

The query subtract the month and year from *Date* column and select columns *Booster_Version, Landing_Outcomes* and *Launch_Site* from table *SPACEXTABLE* where *Landing_Outcome* is equal to *Failure (drone ship)* and *Year* is equal to *2015*

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

We ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20 using the following SQL query: %sql select Landing_Outcome, count(Landing_Outcome) as Count from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count Desc

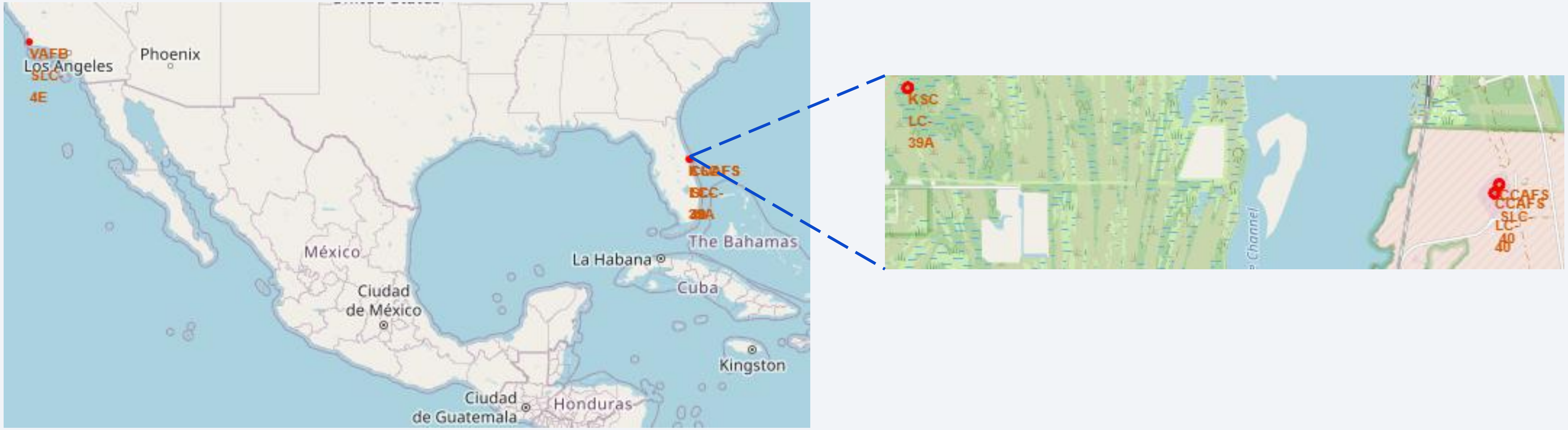
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

The query selected the *Landing_Outcome* and count each event from *SPACEXTABLE* where *Date* is between 2010-06-04 and 2017-03-20 grouping the information by *Landing_Outcome* category

Section 3

Launch Sites Proximities Analysis

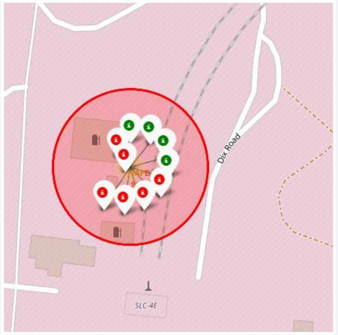
Launch Sites Locations



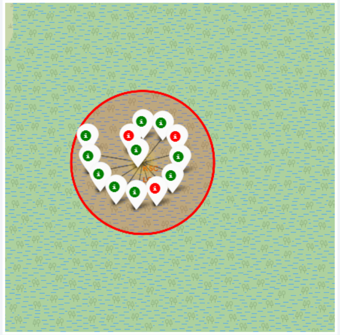
According to the launch sites coordinates the *VAFB SLC-4E* site is located on the West Coast of North America and the sites *CCAFS LC-40, KSC LC-39A* and *CCAFS SLC-40* are located on the East Coast of North America

Launch Outcomes by Site

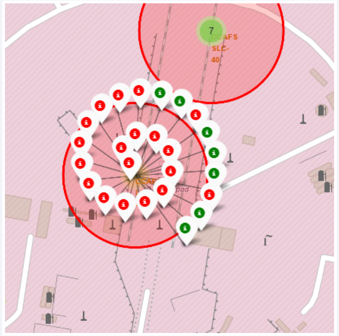
VAFB SLC-4E



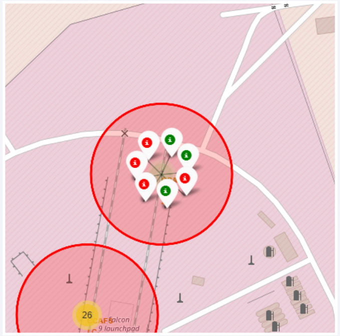
KSC LC-39A



CCAFS LC-40

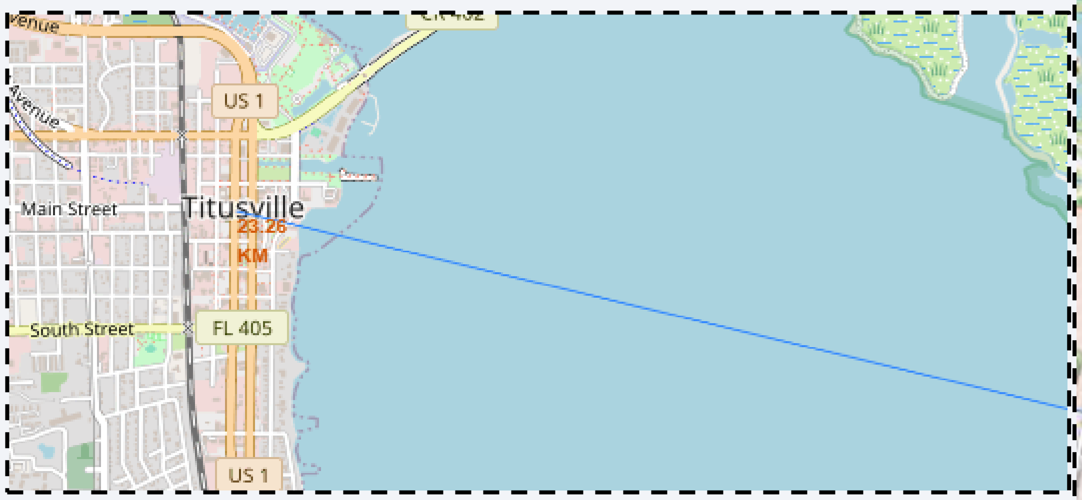


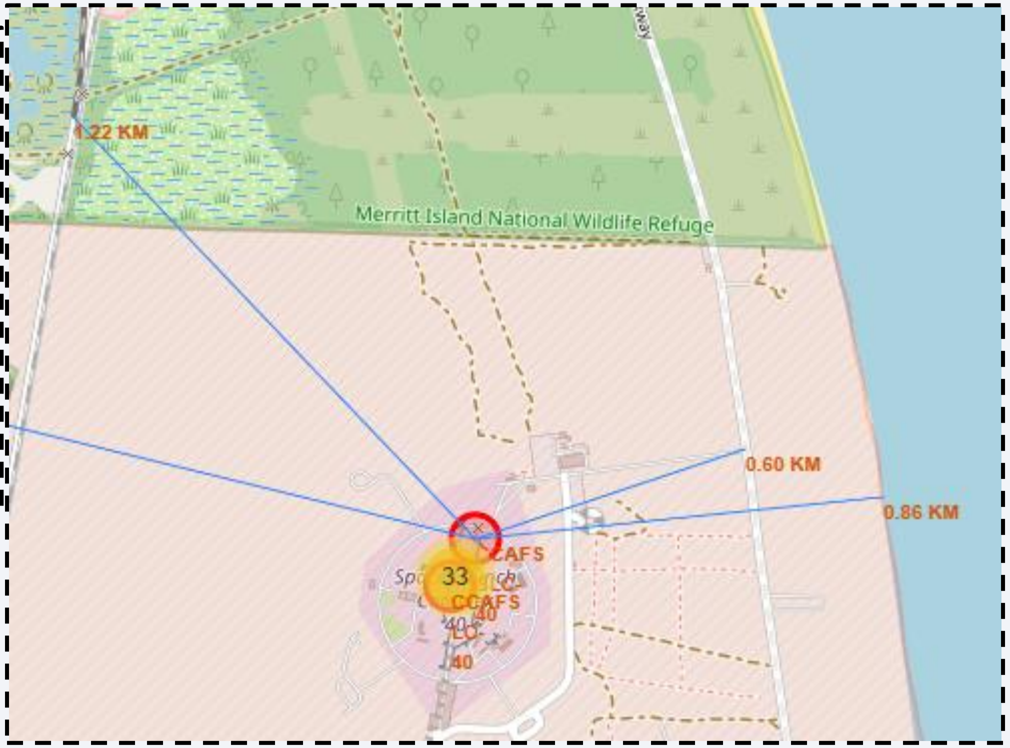
CCAFS SLC-40



These visualizations show the distribution outcomes for each launch site. Launch site KSC LC-39A had the highest number of successful outcomes

Launch Site Proximities (CCAFS SLC-40)



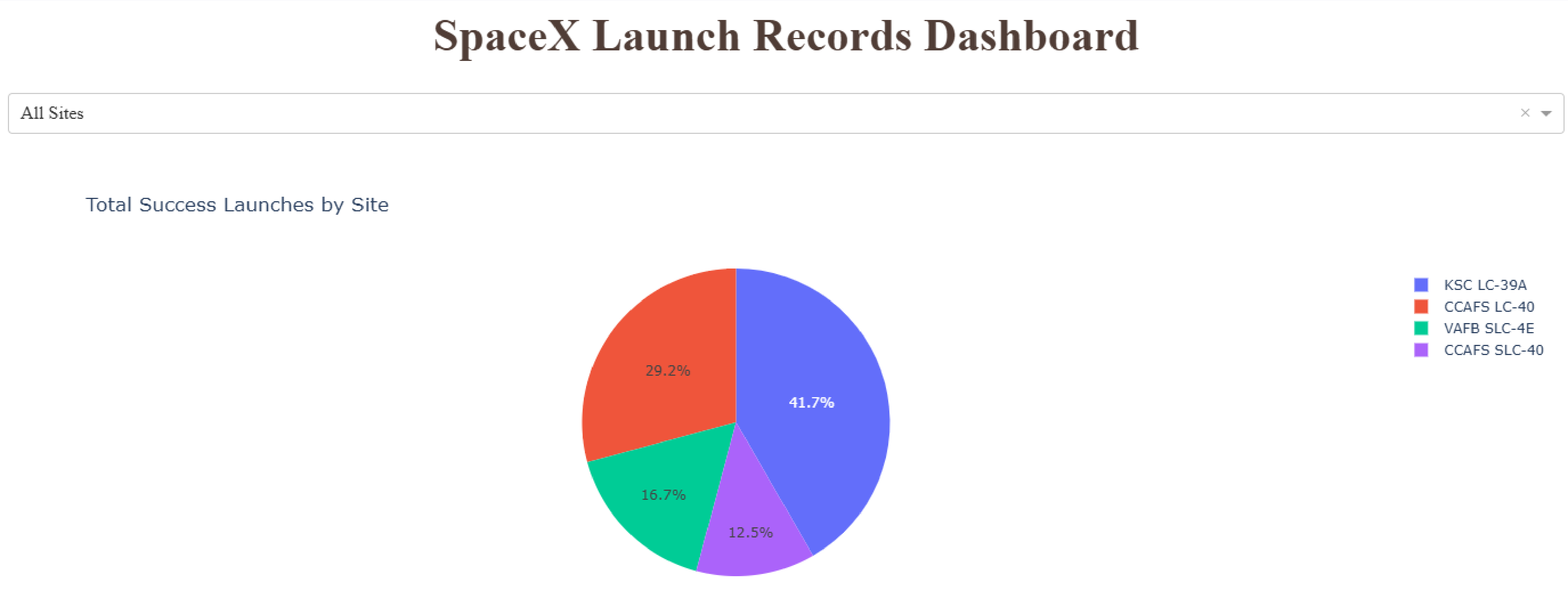


These visualizations show the proximities to the coast, highway, railroad and the nearest city from the *CCAFS SLC-40* launch station

Section 4

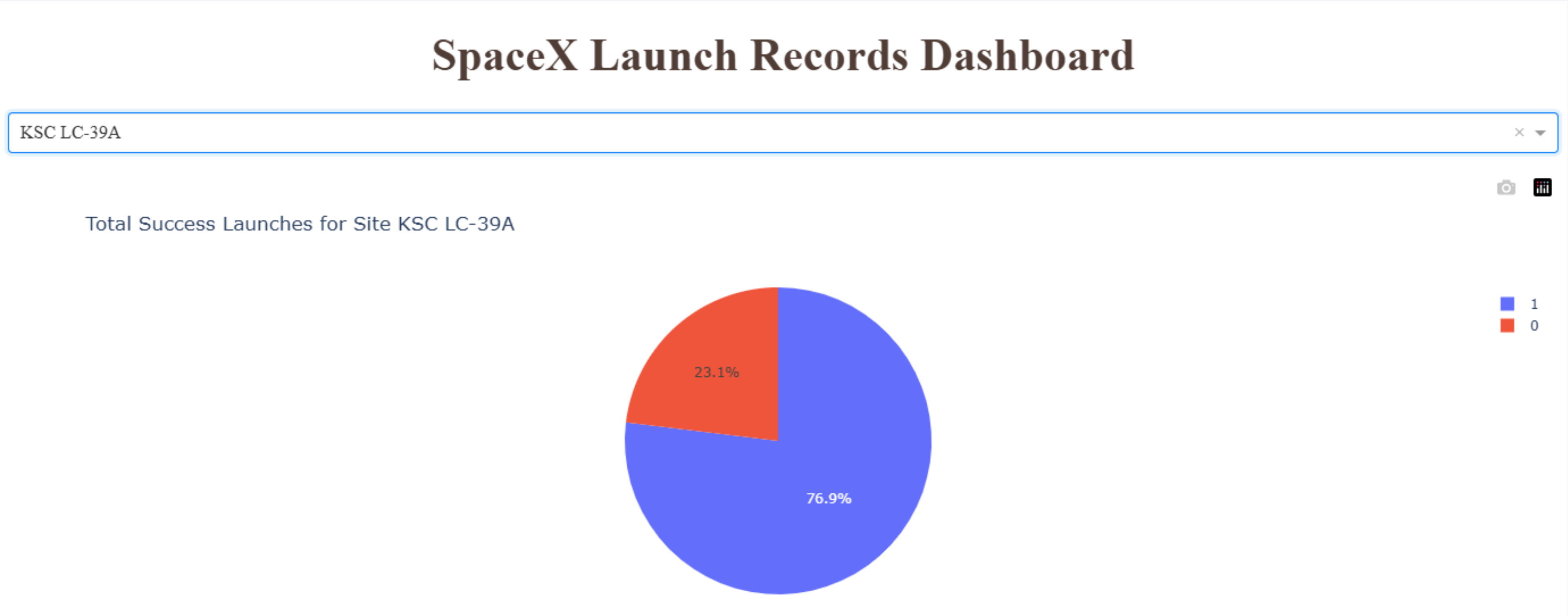
Build a Dashboard with Plotly Dash

Launch Success for all Sites



According to this Pie Chart, the launch site KSC LC-39A had the highest proportion of successful launches, on the other hand, the launch site CCAFS SLC-40 had the lowest proportion of successful launches.

KSC LC-39A Launch Site Success Ratio



As mentioned before the launch site KSC LC-39A had the highest success ratio compared with the other launch sites. KSC LC-39A had a 76.9% of successful outcomes and 23.1% of unsuccessful outcomes.

Relationship between Payload and Launch Outcome

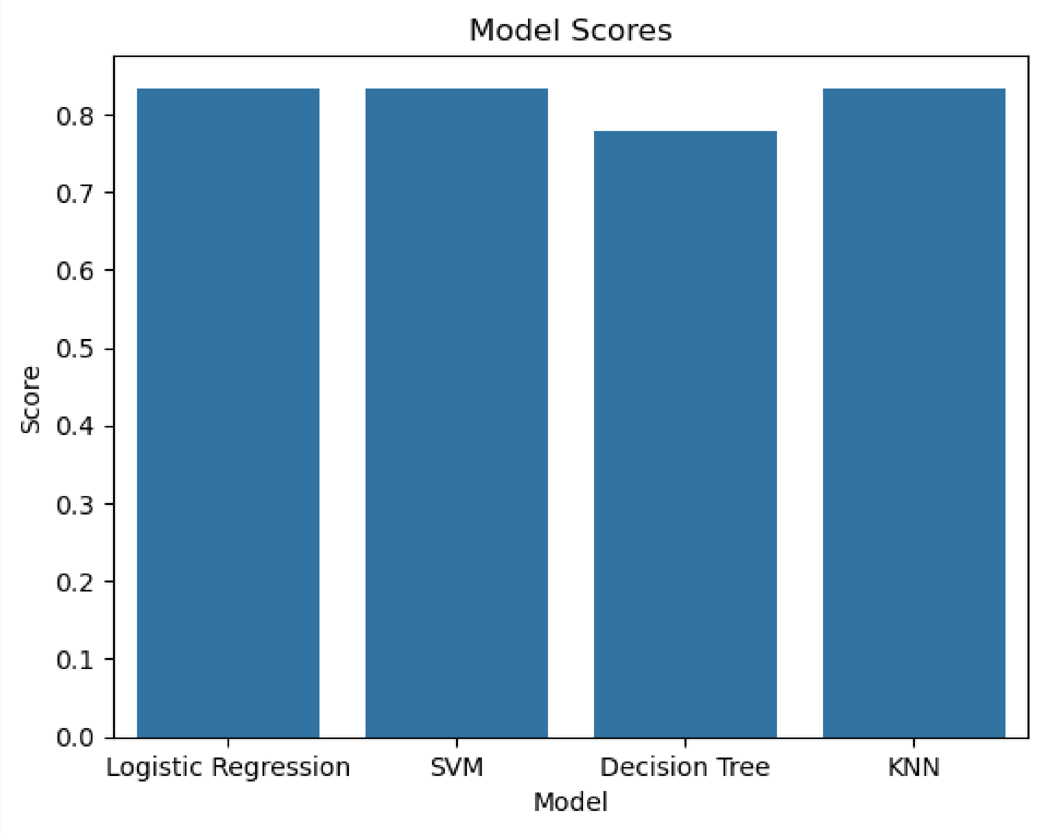
For payload under 5000Kg, the outcomes were spread between successful and unsuccessful, showing no patterns. On the other hand, for payload above 5000kg, there are more unsuccessful outcomes as payload increases, with the exception of one outlier at maximum payload

Section 5

Predictive Analysis (Classification)

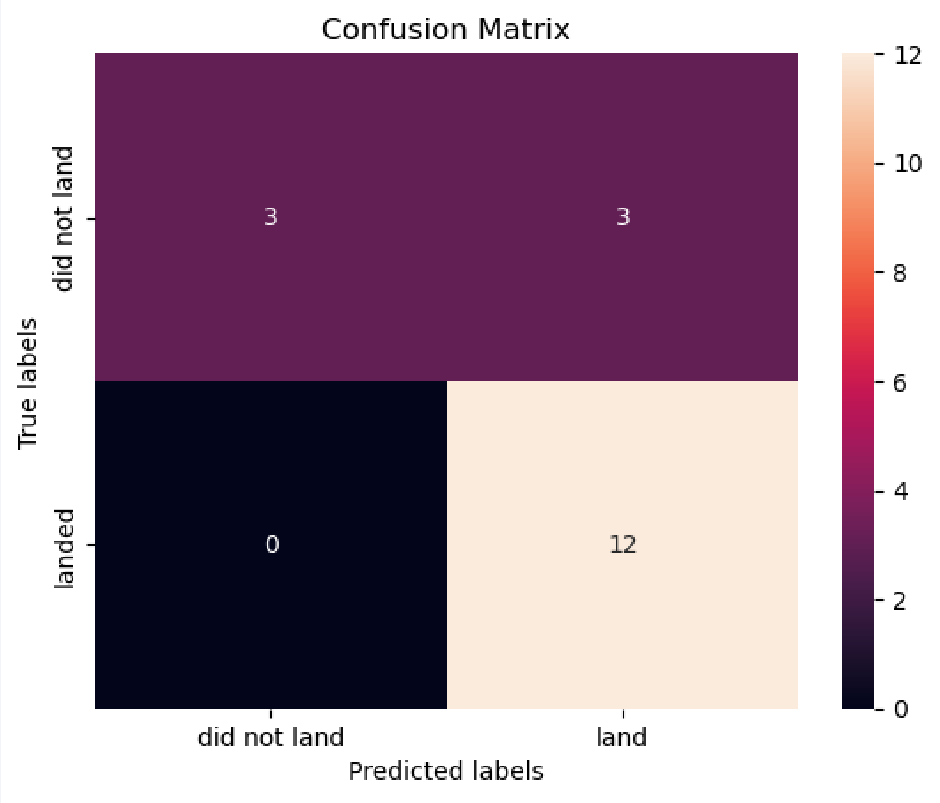
Classification Accuracy

Model	Score
Logistic Regression	0.83333
SVM	0.83333
Decision Tree	0.77778
KNN	0.83333



Logistic Regression, SVM, and KNN model obtained 83.3% accuracy levels, Decision Tree model obtained 77.7% accuracy level.

Confusion Matrix



The confusion matrix for models Logistic Regression, SVM, and KNN which obtained 83.3% accuracy levels, had the same results. 3 false positives were classified by the models.

Conclusions

- Launch site VAFB SLC 4E had more successful outcomes with payload under 5000Kg
- Orbits with more successful outcomes are ES-L1, GEO, HEO and SSO
- Launch site KSC LC-39A had more successful outcomes in the full range of payload
- Launch site KSC LC-39A had 41.7% of total site successful outcomes, by itself had 76.9% of successful outcomes
- Overall, there are more successful outcomes with payloads under 8000kg
- Best Machine Learning models had 83.3% accuracy (Logistic Regression, SVM and KNN)

Appendix

IBM-Applied-Data-Science-Capstone

Thank you!