IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Maryam M.Alizadeh
2022/04/15

Outline

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

Executive Summary

- To do our project we went through these steps:
  - Collected Data from Data Collection API
  - Performed Data Wrangling
  - Performed Exploratory Data Analysis using SQL and Visualization
  - Built Interactive Visual Analytics and Visual Dashboards
  - Performed Predictive Analysis using Classification

- The result showed that our model was good at predicting successful landing with accuracy of 83.33%

Introduction

- Project Background:
  - Companies are making space travel affordable for everyone.
  - 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.
  - We work in SpaceY and want to compete with SpaceX

- Problems we want to find answers
  - Determine the price of each launch
  - Determine if SpaceX will reuse the first stage

Section 1

Methodology

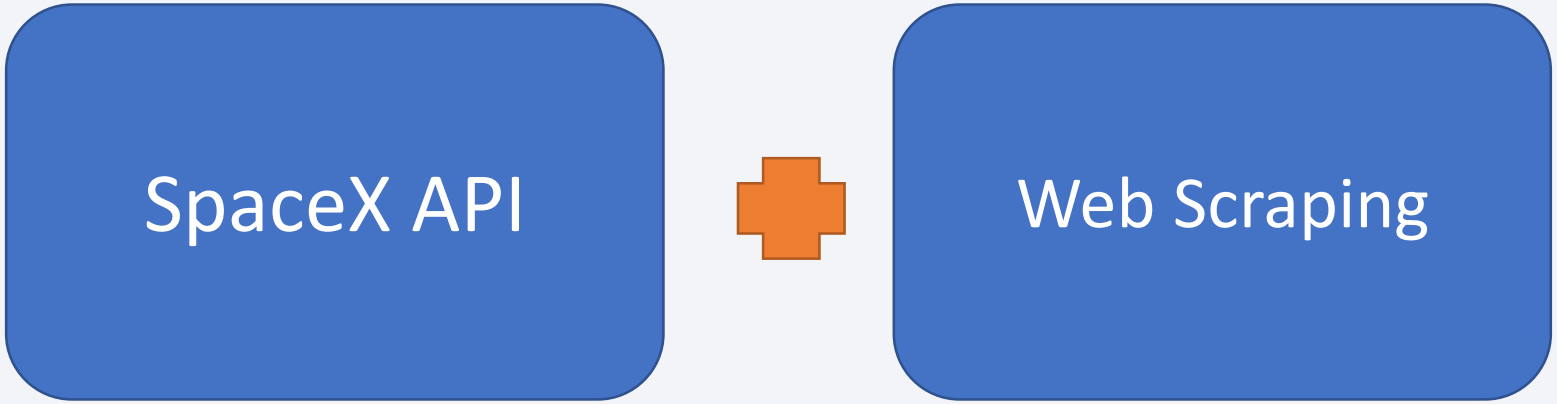
Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Load the Dataframe, Tuned models using GridSearchCV, Evaluated model's accuracy using score method

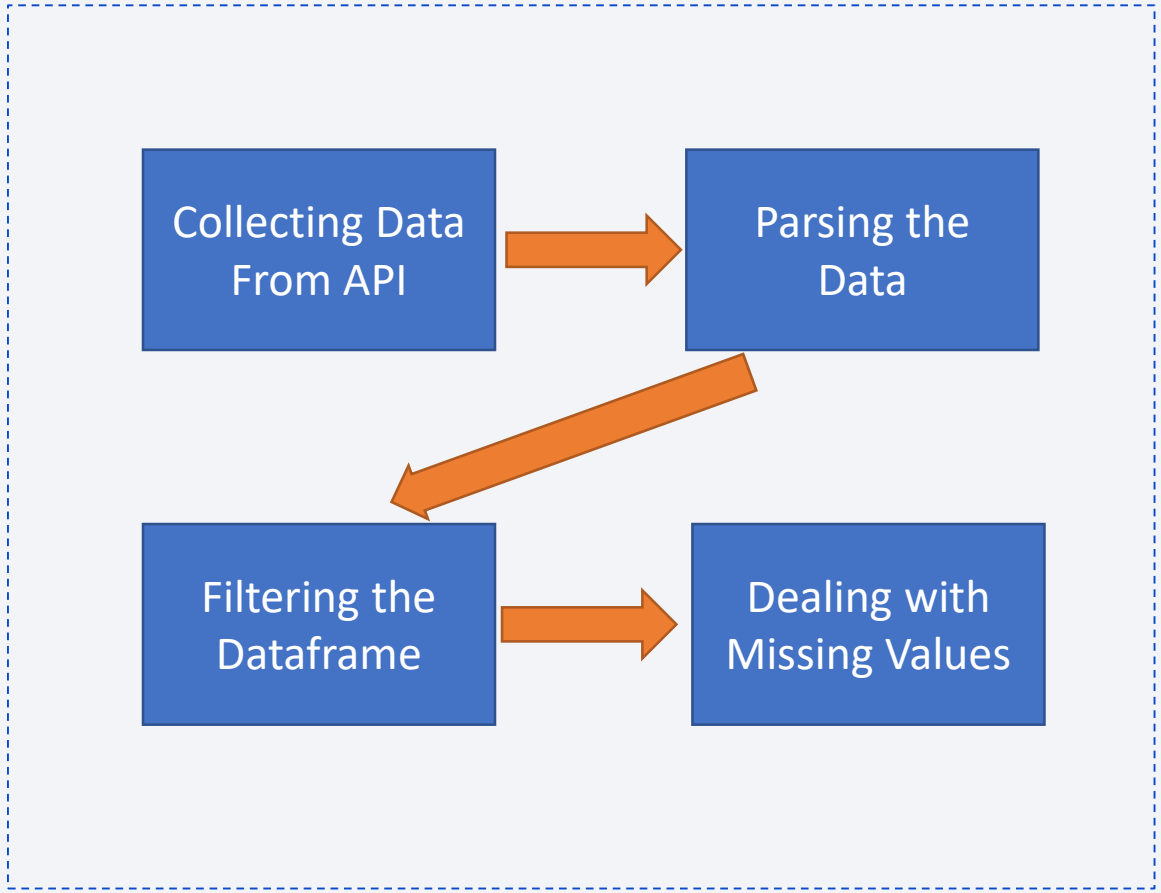
Data Collection

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.



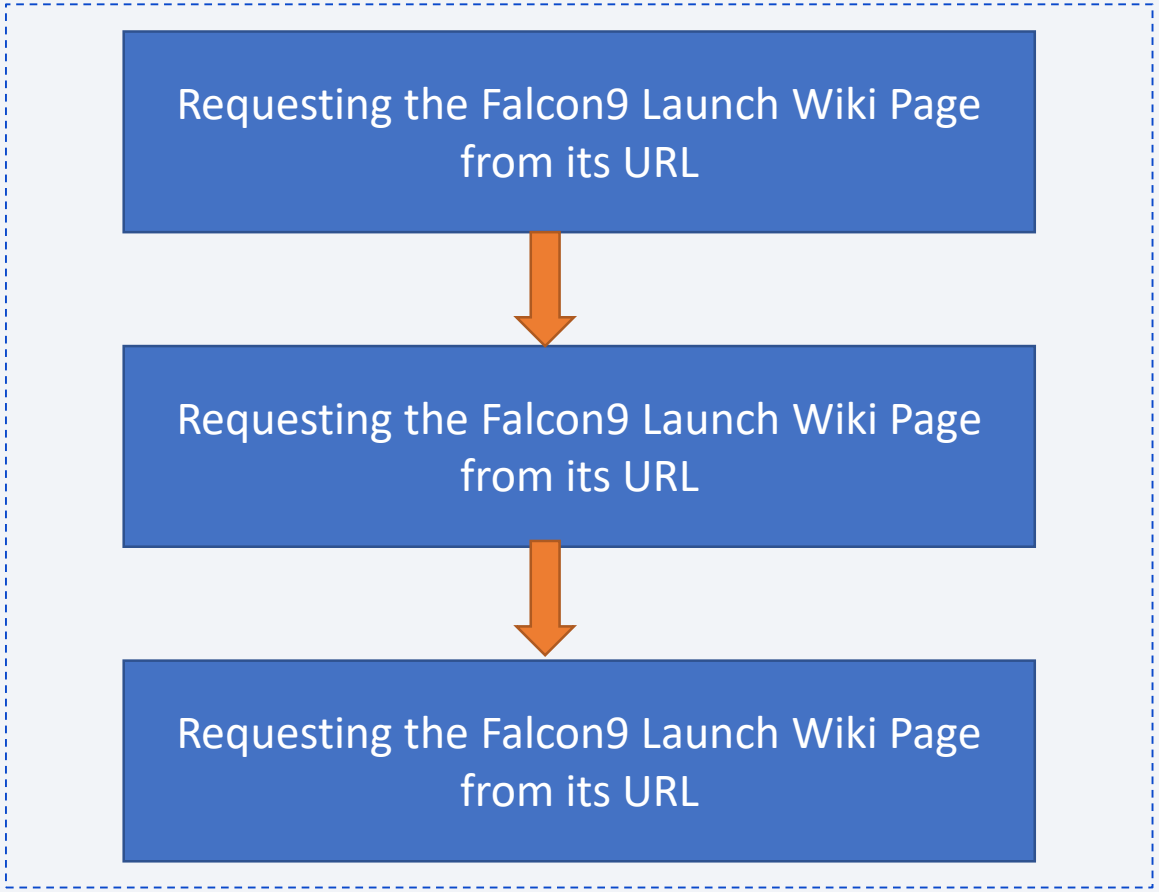
Data Collection - SpaceX API

- The process for API request included:
  - Collecting data from https://api.spacexdata.com/v4/launches/past
  - Parsing the data
  - Filtering the Dataframe
  - Dealing with Missing Values
- Github URL:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/SpaceX%20data%20collection%20api.ipynb



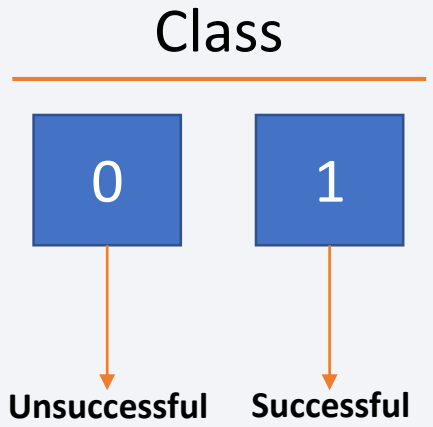
Data Collection - Scraping

- The process for Web Scraping included:
  - Request the Falcon9 Launch Wiki page from its URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - Extracting all column/variable names from the HTML table header
  - Creating a data frame by parsing the launch HTML tables
- Github URL:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/Webscraping.ipynb



Data Wrangling

- The process of Data Wrangling included:
  - Data Analysis
  - Calculating the number of launches on each site
  - Calculating the number and occurrence of each orbit
  - Calculating the number and occurrence of mission outcome per orbit type
  - Creating a landing outcome label from Outcome column
    - For each launch if column Class = 0, the landing was unsuccessful
    - For each launch if column Class = 1, the landing was successful
  - Determining the success rate
    - Success Rate = 0.6667



- GitHub URL:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- The Process of Data Visualization included:
  - Visualizing the relationship between Flight Number and Play load Mass with Scatter Plot
  - Visualize the relationship between Flight Number and Launch Site with Scatter Plot
  - Visualizing the relationship between Payload and Launch Site with Scatter Plot
  - Visualizing the relationship between success rate of each orbit type with Bar Chart
  - Visualizing the relationship between FlightNumber and Orbit type with Scatter Plot
  - Visualizing the relationship between Payload and Orbit type with Scatter Plot
  - Visualizing the launch success yearly trend with Line Chart
  - Creating dummy variables to categorical columns ('Orbit','LaunchSite', 'LandingPad', 'Serial')
  - Casting all numeric columns to float64
- GitHub URL:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- The Process of EDA with SQL included:
  - Downloading the datasets
  - Storing the dataset in database table
  - Connecting to the database
  - Displaying the names of the unique launch sites in the space mission using SQL query
  - Displaying 5 records where launch sites begin with the string 'KSC' using SQL query
  - Displaying the total payload mass carried by boosters launched by NASA (CRS) using SQL query
  - Displaying average payload mass carried by booster version F9 v1.1 using SQL query
  - Listing the date where the first successful landing outcome in drone ship was achieved using SQL query
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000 using SQL query
  - Listing the total number of successful and failure mission outcomes using SQL query
  - Listing the names of the booster_versions which have carried the maximum payload mass using SQL query and subquery
  - 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 using SQL query
  - Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order using SQL query
- GitHub URL:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- The Process of Building Interactive Visual Analytics with Folium:
  - Marking all launch sites on a map using folium.Circle and folium.Marker
  - Marking the success/failed launches for each site on the map using marker_cluster
  - Calculating the distances between a launch site to its proximities (highway, railroad, city, coastline) using folium.Marker and PolyLine
- We did all of these to be able to find some geographical patterns about launch sites.
- GitHub URL:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The Process of Building an Interactive Dashboard with Ploty Dash included:
  - Adding a Launch Site Drop-down Input Component
  - Adding a callback function to render success-pie-chart based on selected site
  - Adding a Range Slider to Select Payload
  - Adding a callback function to render the success-payload-scatter-chart scatter plot
- We did all of this to find the launch site with the largest success count and also to find the correlation between payload and mission outcome
- GitHub URL:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

- The Process of Predictive Analysis included:
  - Creating a NumPy array from the column Class in data
  - Standardizing the data in X
  - Using the function train_test_split to split the data X and Y into training and test data
  - Creating a logistic regression object then creating a GridSearchCV object logreg_cv with cv = 10
  - Calculating the accuracy on the test data using the method score
  - Creating a support vector machine object then creating a GridSearchCV object svm_cv with cv = 10
  - Calculating the accuracy on the test data using the method score
  - Creating a decision tree classifier object then creating a GridSearchCV object tree_cv with cv = 10
  - Calculating the accuracy of tree_cv on the test data using the method score
  - Creating a k nearest neighbors object then creating a GridSearchCV object knn_cv with cv = 10
  - Calculating the accuracy of tree_cv on the test data using the method score
  - Finding the method which performs best
- GitHub URL:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/Machine%20Learning%20Prediction.ipynb

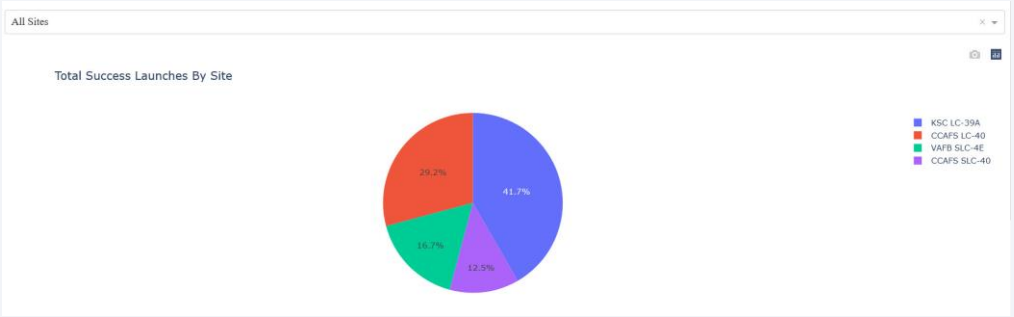
Results

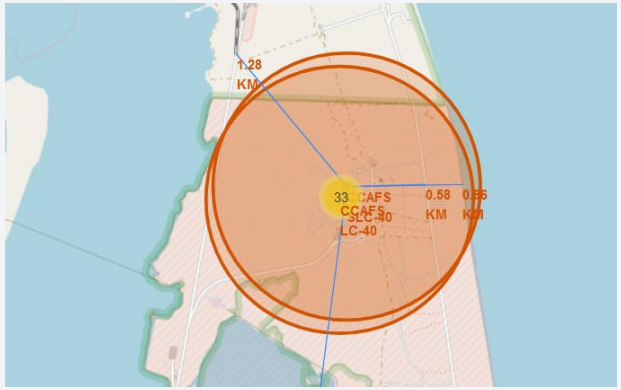
- Exploratory data analysis results
  - As Flight Numbers increases the Successful Lands increases in all Launch Sites
  - As Pay Load Mass increases the Successful Lands increases in all Launch Sites
  - Orbits ES-L1, GEO, HEO and SSO have the highest Mean Success Rate
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS
  - Launch success yearly trend shows that the success rate since 2013 kept increasing till 2020

  - Unique launch sites in the space mission are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
  - Total payload mass carried by boosters launched by NASA (CRS) is 45596
  - The first successful landing outcome in drone ship was achieved at 2015-12-22.
  - Total number of successful and failure mission outcomes: Failure (in flight) = 1, Success = 99, Success (payload status unclear) = 1

Results

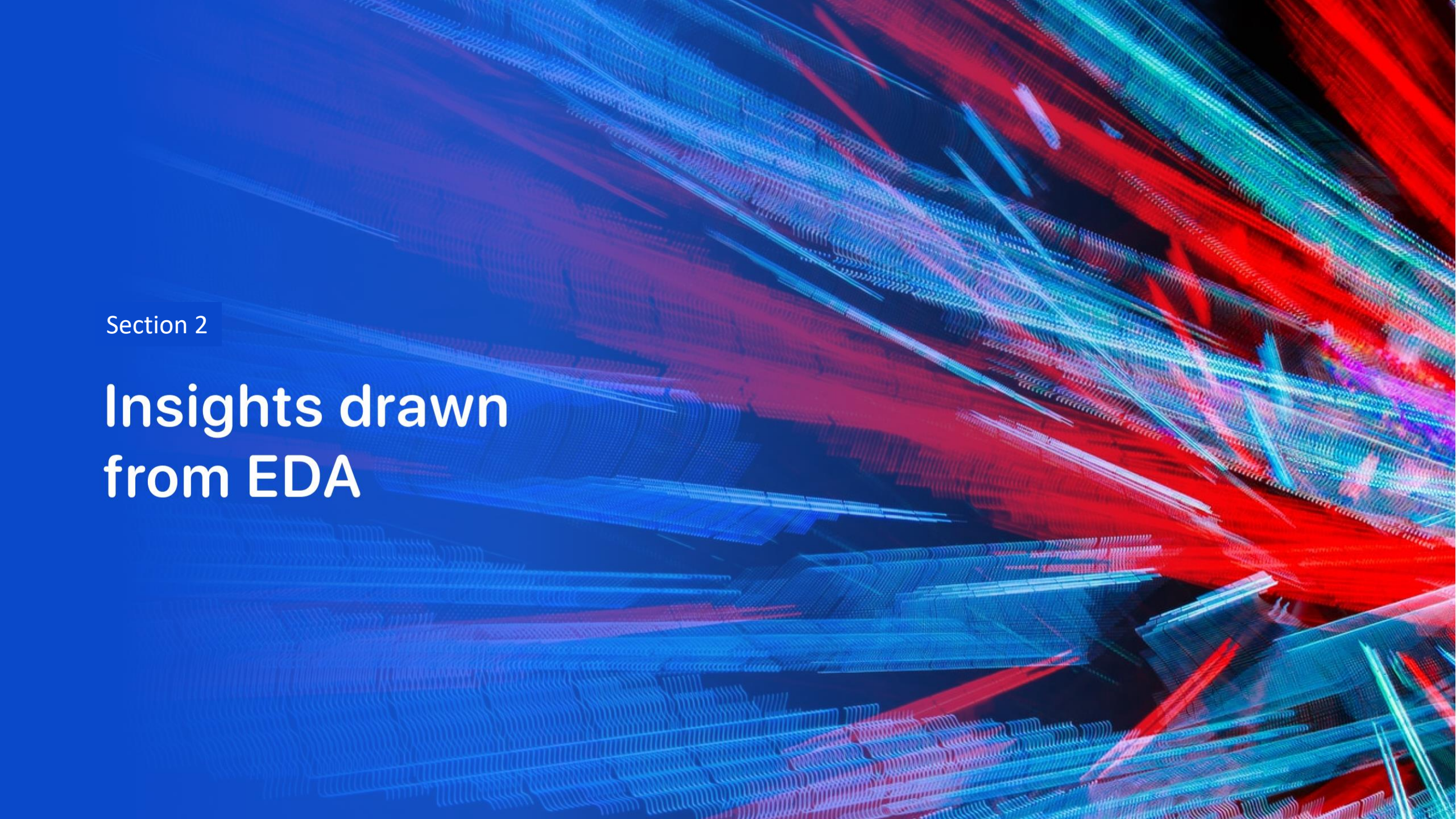
- Interactive analytics results:
  - KSCL LC-39A has the most successful launches
  - The is no correlation between Payload and Successful Lands based on Booster Version Category
  - CCAFS SLC-40 is in close proximity to highway, coastline and railway and far away from Melbourne



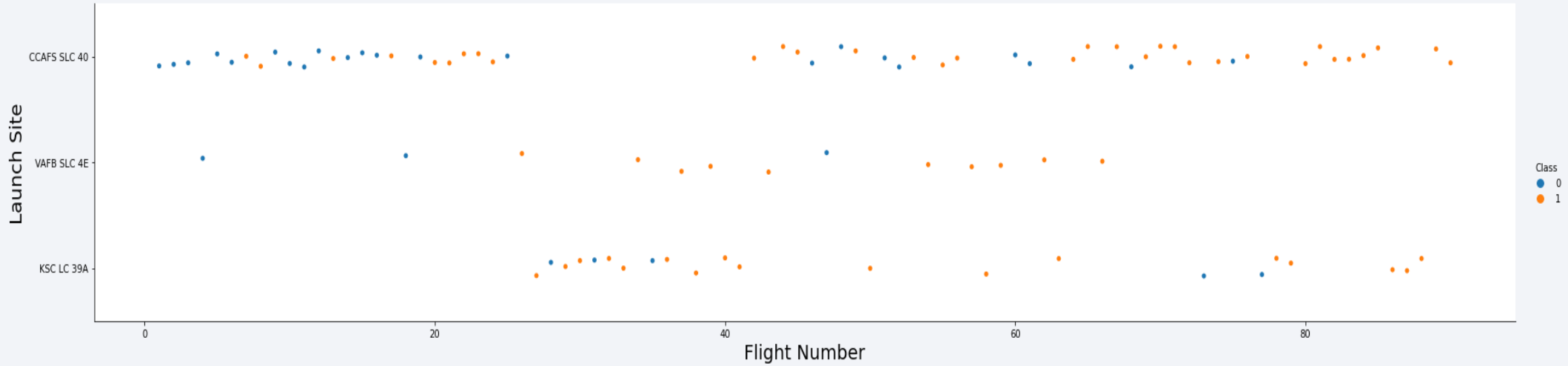


Results

- Predictive analysis results
  - For all of the below methods The highest accuracy rate is 0.8333:
    - logistic regression
    - support vector machine
    - decision tree classifier
    - k nearest neighbors

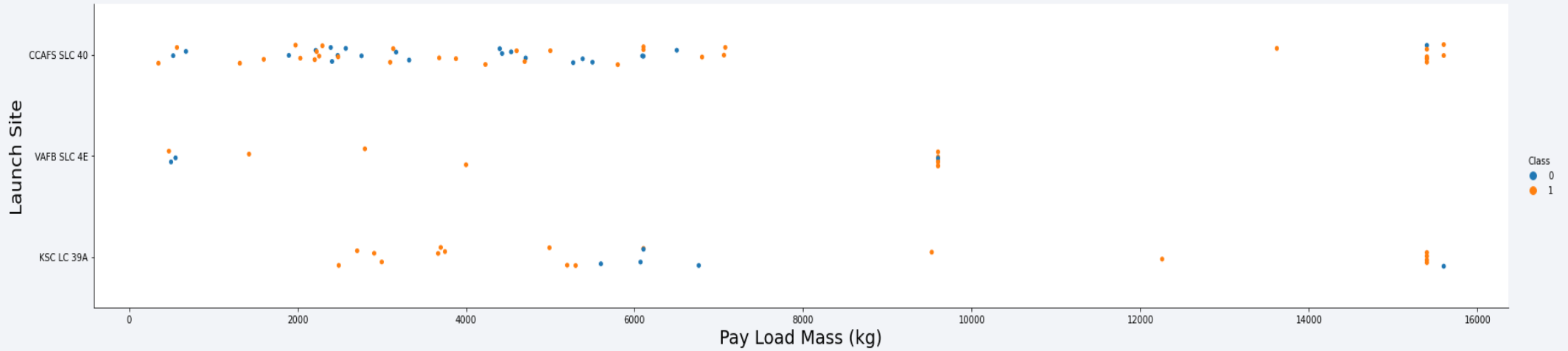

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



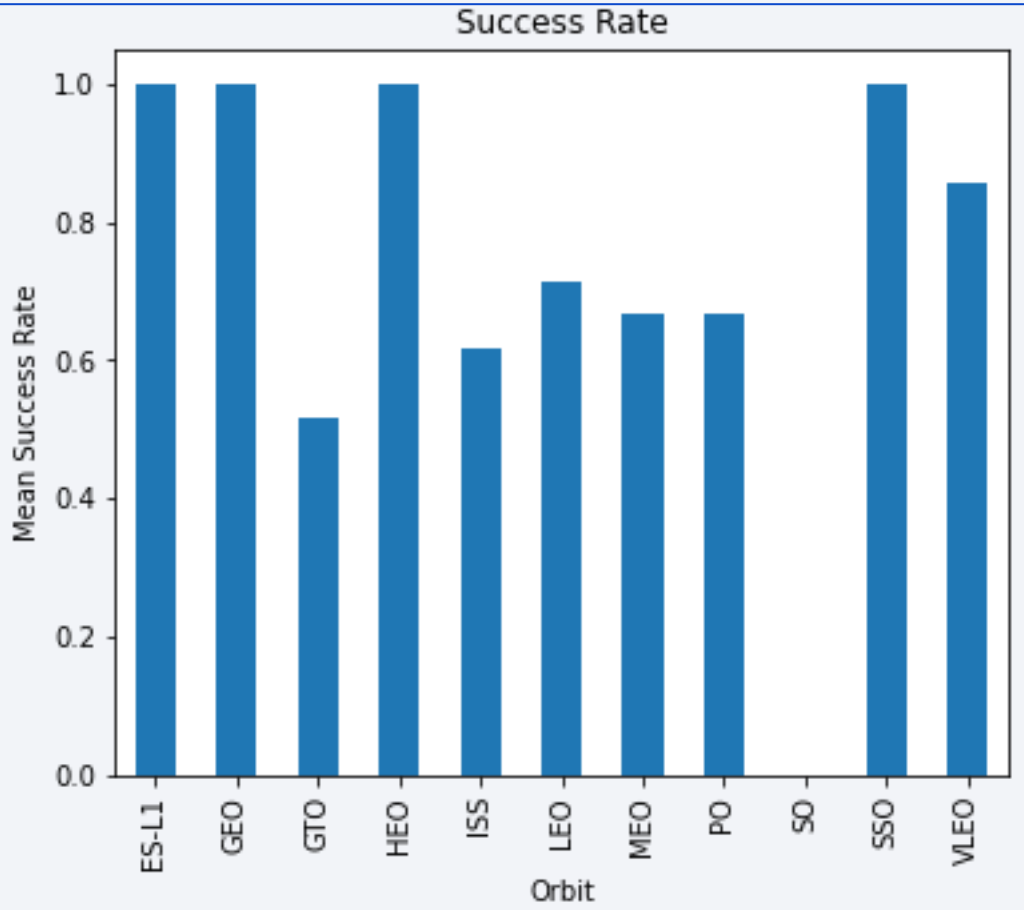
- As flight numbers increases, successful lands increases in all these three launch sites

Payload vs. Launch Site



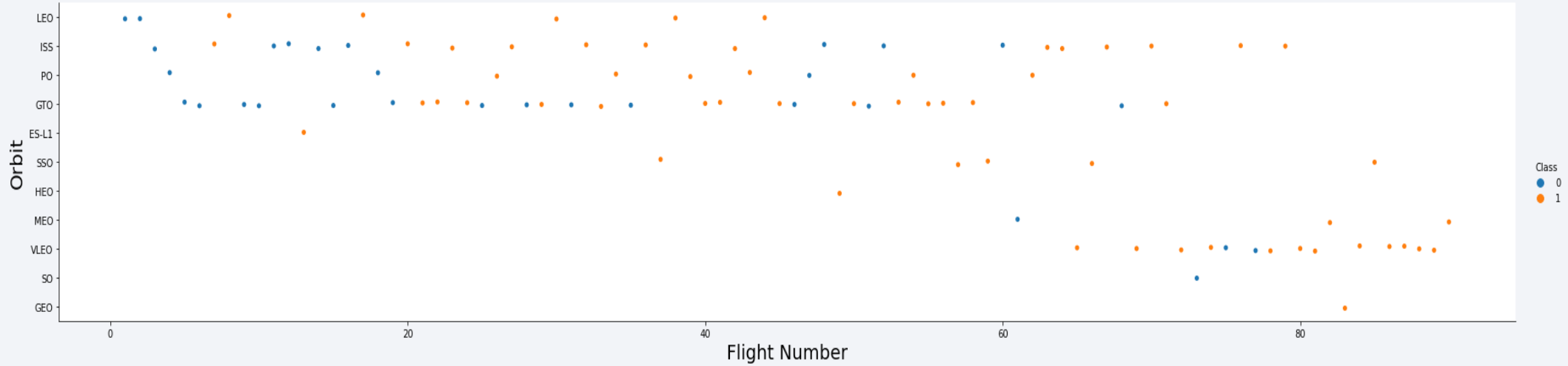
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).
- As payload mass increases, successful lands increases in all these three launch sites

Success Rate vs. Orbit Type



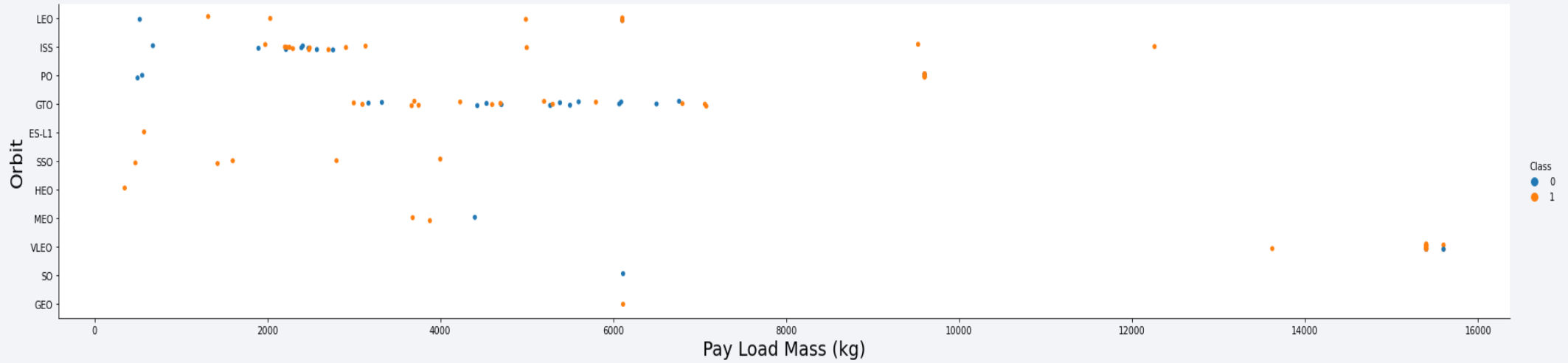
- Orbits ES-L1, GEO, HEO and SSO have the highest Mean Success Rate
- Success Rate in orbit SO is 0

Flight Number vs. Orbit Type



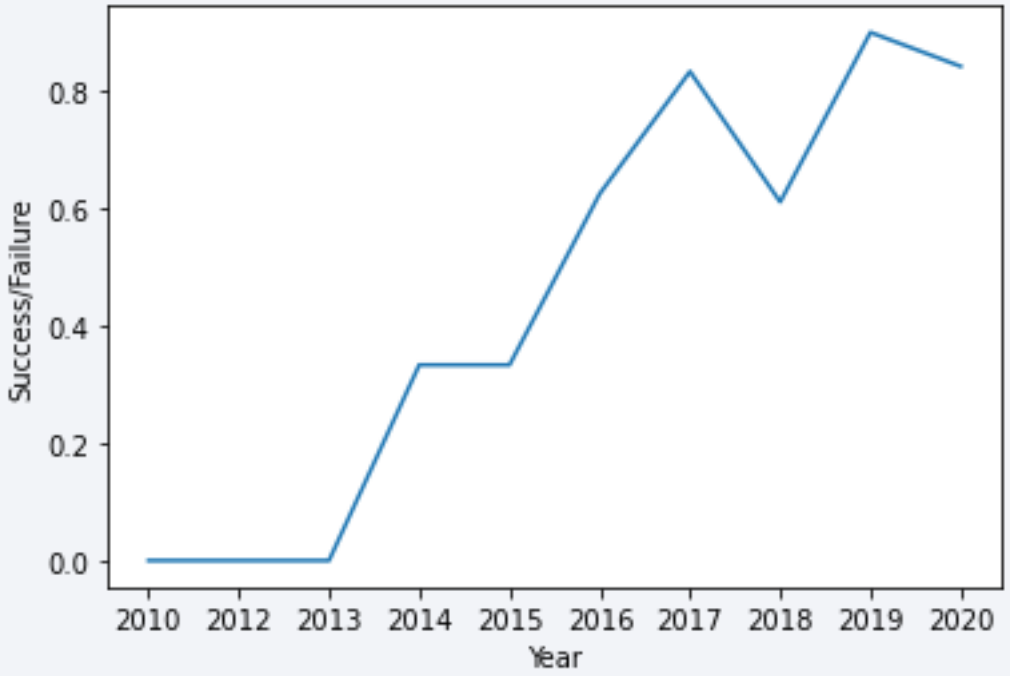
- In the LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

Launch Success Yearly Trend



- We can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites
- Names of unique launch sites are: CCAFS LC-40, CCAFS SLc-40, KSC LC-39A, VAFB SLC-4E

```
%sql select unique(launch_site) from SPACEXTBL
```

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

Launch Site Names Begin with 'KSC'

- Find 5 records where launch sites' names start with `KSC`

```
%sql SELECT * from SPACEXTBL where (LAUNCH_SITE) LIKE 'KSC%' LIMIT 5
```

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Total Payload Mass = 45596

```
%sql SELECT sum(payload_mass__kg_) as sum_payload from SPACEXTBL where (customer) = 'NASA (CRS)'
```

sum_payload
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Average payload mass carried by booster version F9 v1.1 = 2928

```
%sql SELECT avg(payload_mass__kg_) as average_payload from SPACEXTBL where (booster_version) = 'F9 v1.1'
```

average_payload
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- First successful landing outcome on ground pad was at 2015-12-22

```
%sql SELECT min(date) from SPACEXTBL where landing__outcome = 'Success (ground pad)'
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Their names are: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

```
%sql select BOOSTER_VERSION from SPACEXTBL 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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

mission_outcome	outcome
Failure (in flight)	1
Success	99
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 SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

```
%%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH')
AS MONTH_NAME, LANDING__OUTCOME AS LANDING__OUTCOME, BOOSTER_VERSION AS BOOSTER_VERSION, LAUNCH_SITE AS LAUNCH_SITE FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)' AND "DATE" LIKE '%2017%'
```

month_name	landing__outcome	booster_version	launch_site
FEBRUARY	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
MAY	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
JUNE	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
AUGUST	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
SEPTEMBER	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
DECEMBER	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

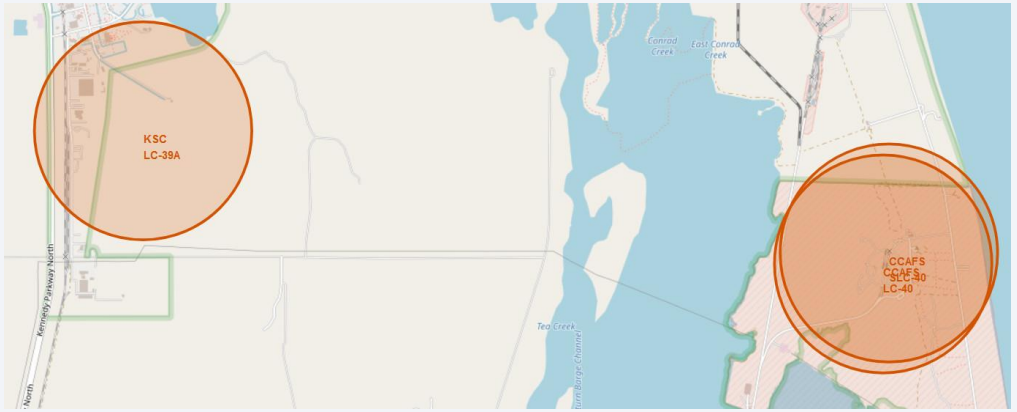
- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT_LAUNCHES DESC
```

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

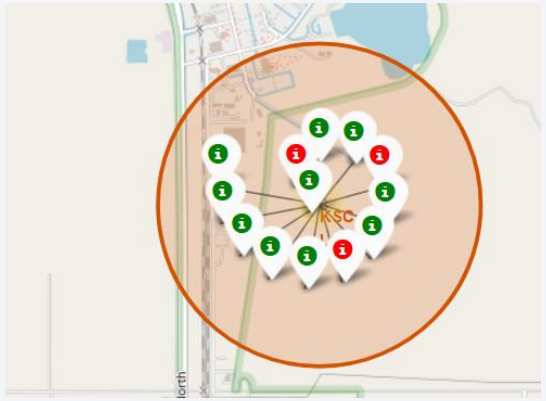

Section 3
Launch Sites
Proximities Analysis

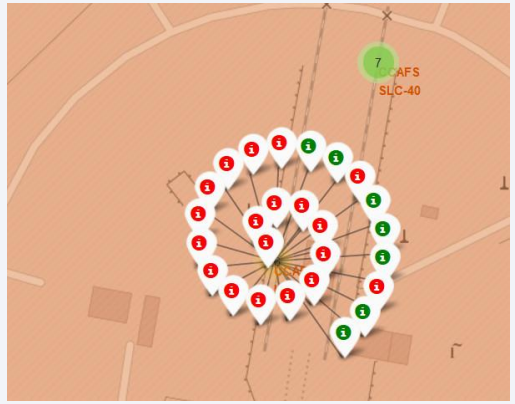
Mark all launch sites on a map

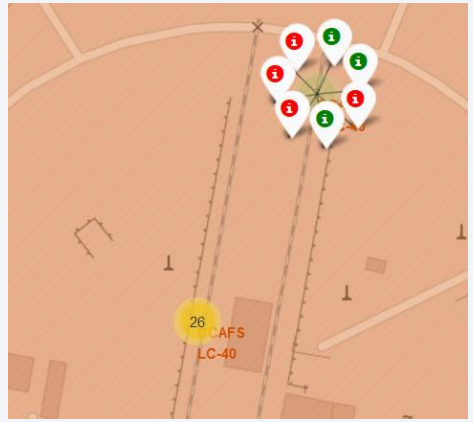


- There are 4 launch sites on the map: VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40, CCAFS LC-40
- VAFB SLC-4E is in California
- KSC LC-39A, CCAFS SLC-40, CCAFS LC-40 are in Florida
- CCAFS SLC-40 and CCAFS LC-40 are near each other

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

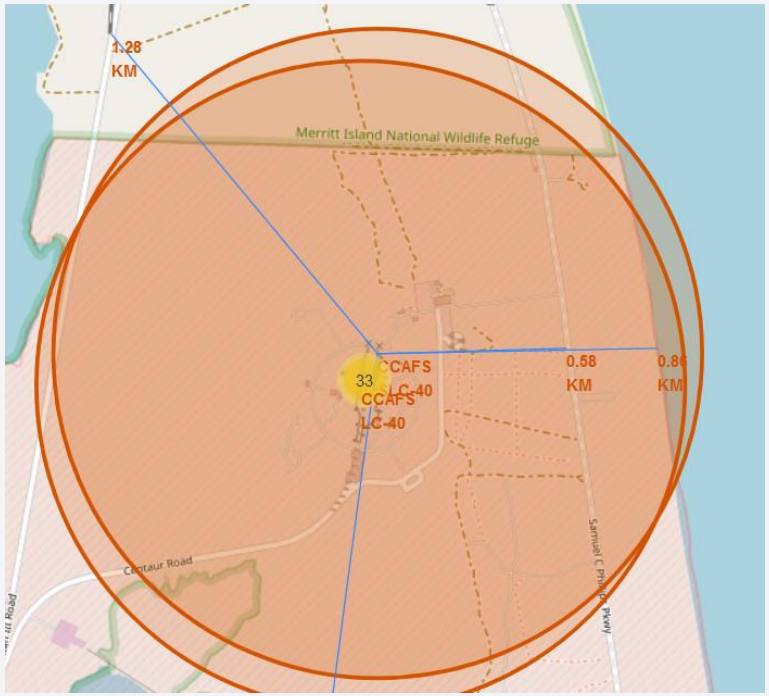


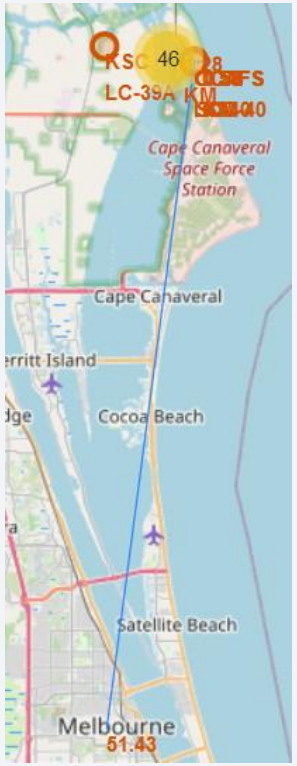




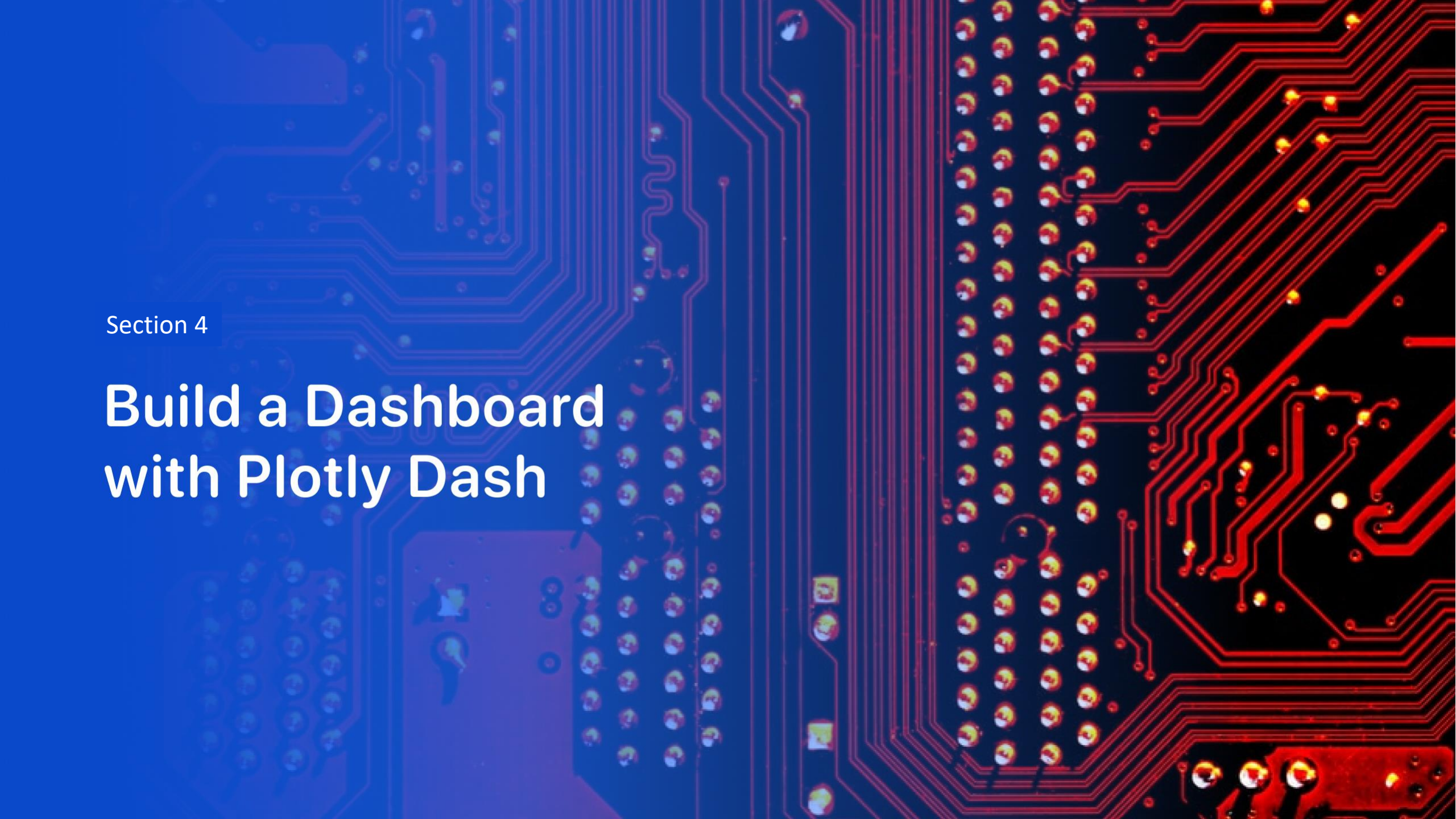
- VAFB SLC-4E has 4 successful and 6 failed landings
- KSC LC-39A has 10 successful and 3 failed landings
- CCAFS LC-40 has 7 successful and 19 failed landings
- CCAFS SLC-40 has 3 successful and 4 failed landings

Calculate the distances between CCAFS SLC-40 to its proximities

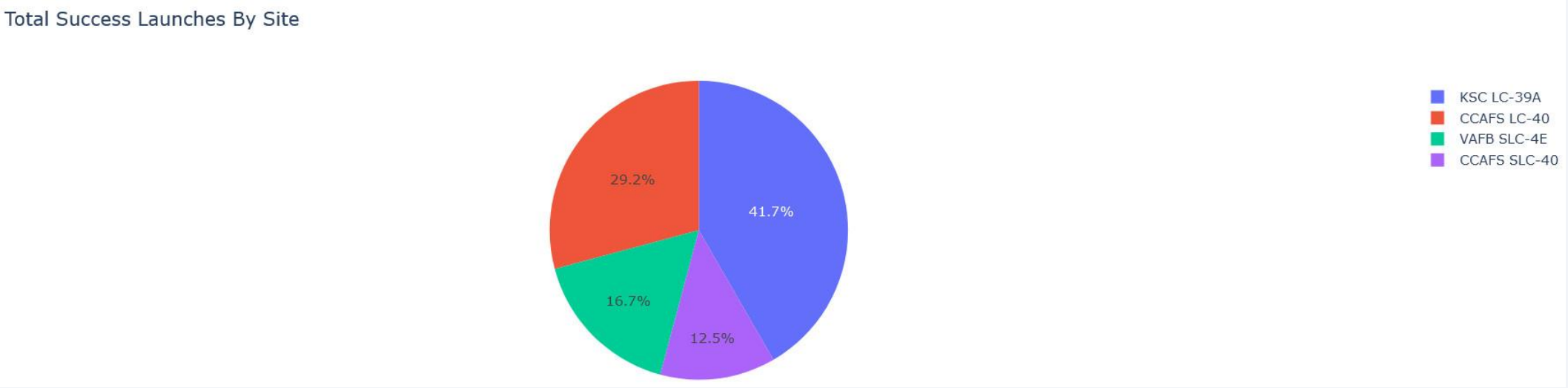




- The distance between CCAFS SLC-40 and its nearest highway (Samuel C Philips Pkwy) is 0.58 KM
- The distance between CCAFS SLC-40 and its nearest railway (NASA Railroad) is 1.28 KM
- The distance between CCAFS SLC-40 and its nearest coastline is 0.86 KM
- The distance between CCAFS SLC-40 and Melbourne is 51.43 KM

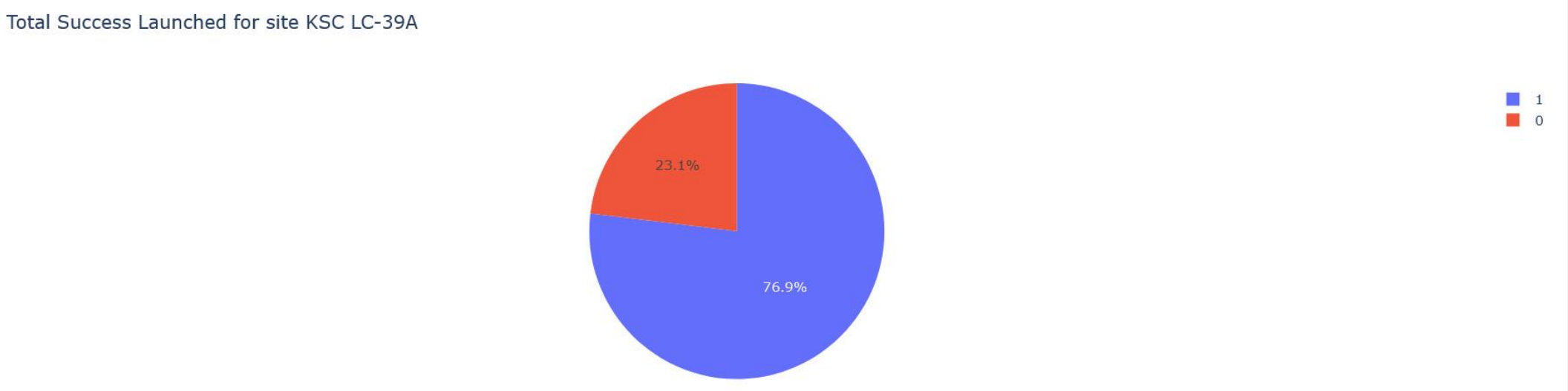

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



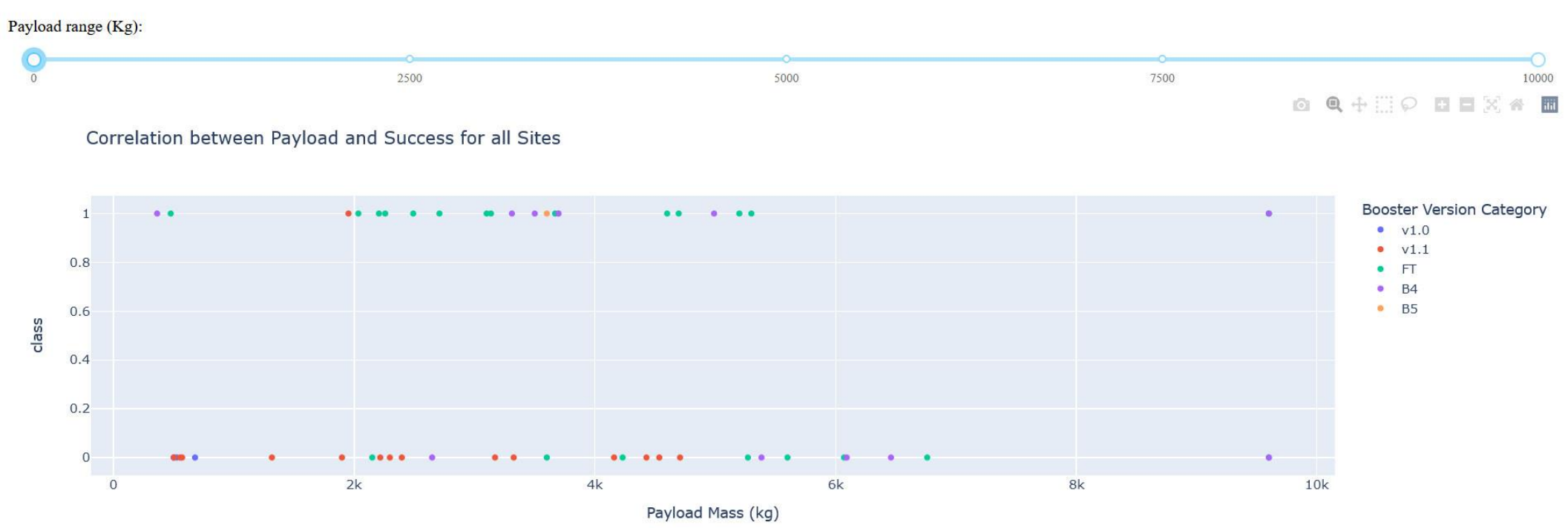
- KSC LC-39A has the highest successful land counts among the others with 10 successful lands.

Total Success Lanched for site KSC LC-39A



- Site KSC LC-39A has the highest success rate with 10 successful landings and 3 failures.

Correlation between Payload and Success for site CCAFS SLC-40

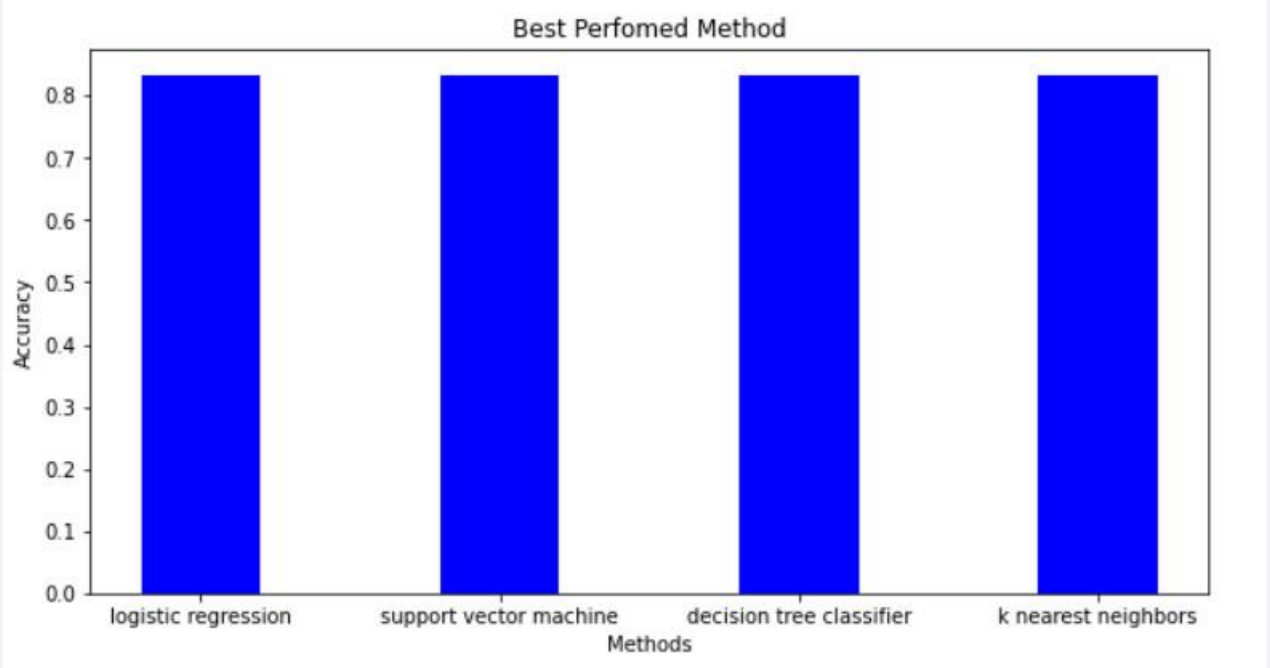


- There is no relationship between booster versions v1.1 and B5
- FT has the highest success rate when the Payload Mass is between 0 and 3500
- B4 has the highest success rate when the Payload Mass is between 3000 and 5000
- v1.0 success rate is 0

Section 5

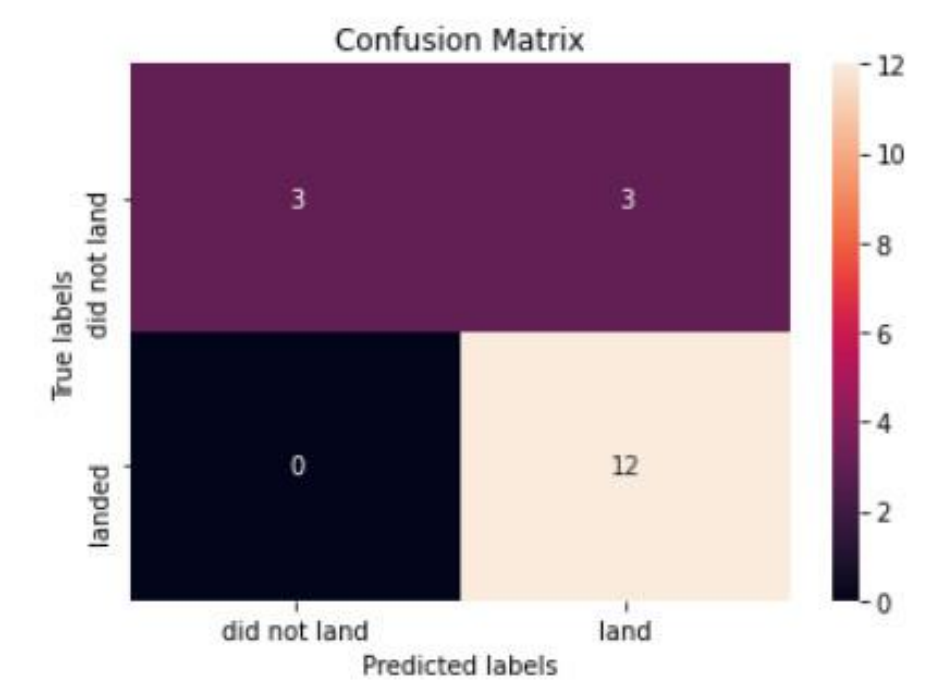
Predictive Analysis (Classification)

Classification Accuracy



- All methods have an accuracy of 0.8333

Confusion Matrix



- Examining the confusion matrix, we see that the major problem is false positives.

Conclusions

- As the number of flights increases, the success rate increases
- Orbits SSO, HEO, GEO, and ES-L1 have the highest success rate (100%).
- The launch sites are close to railways, highways, and coastline, but far from cities.
- KSLC-39A has the highest number of launch successes and the highest success rate among all sites.
- In this dataset, all classification models have the same accuracy (83.33%) but it seems that more data is needed to determine the optimal model due to the small data size.
- SpaceY can use this model to predict with high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not.

Appendix

- All Jupyter notebooks and Python Codes are in GitHub at:
  - https://github.com/maryamalizadeh91/Data-Science-and-Machine-Learning-Capstone-Project

Thank you!