IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Elly Kang
December 13, 2021

Outline

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

Executive Summary

- Data is collected from SpaceX Wikipedia page using the SpaceX public API. Created a new feature called ‘Class’ which classifies the landings as successful or failure. SQL, different visualization tools, folium maps and dashboards are used to explore data.
- 4 machine learning models are used to perform predictive analysis on landings. The best model is selected after using hyperparameter tuning and GridSearchCV.
- Decision Tree performed the worst with accuracy of 77.8% while all the other models performed the same with an accuracy of 83.3%.

Introduction

Background

- Space programs are now being commercialized making it possible for anyone and everyone to be able to fly to space.
- This has been made possible especially by the low cost of the SpaceX Falcon9 rocket ($62 million USD) as compared to the other providers ($165 million USD)
- The lower cost accredited to the SpaceX ability to reuse first stage.
- An analysis for a hypothetical space company is done to present as a competitor of SpaceX

Problem

- Training machine learning models to predict successful landing and recovery of stage 1.

Section 1

Methodology

Methodology

Executive Summary

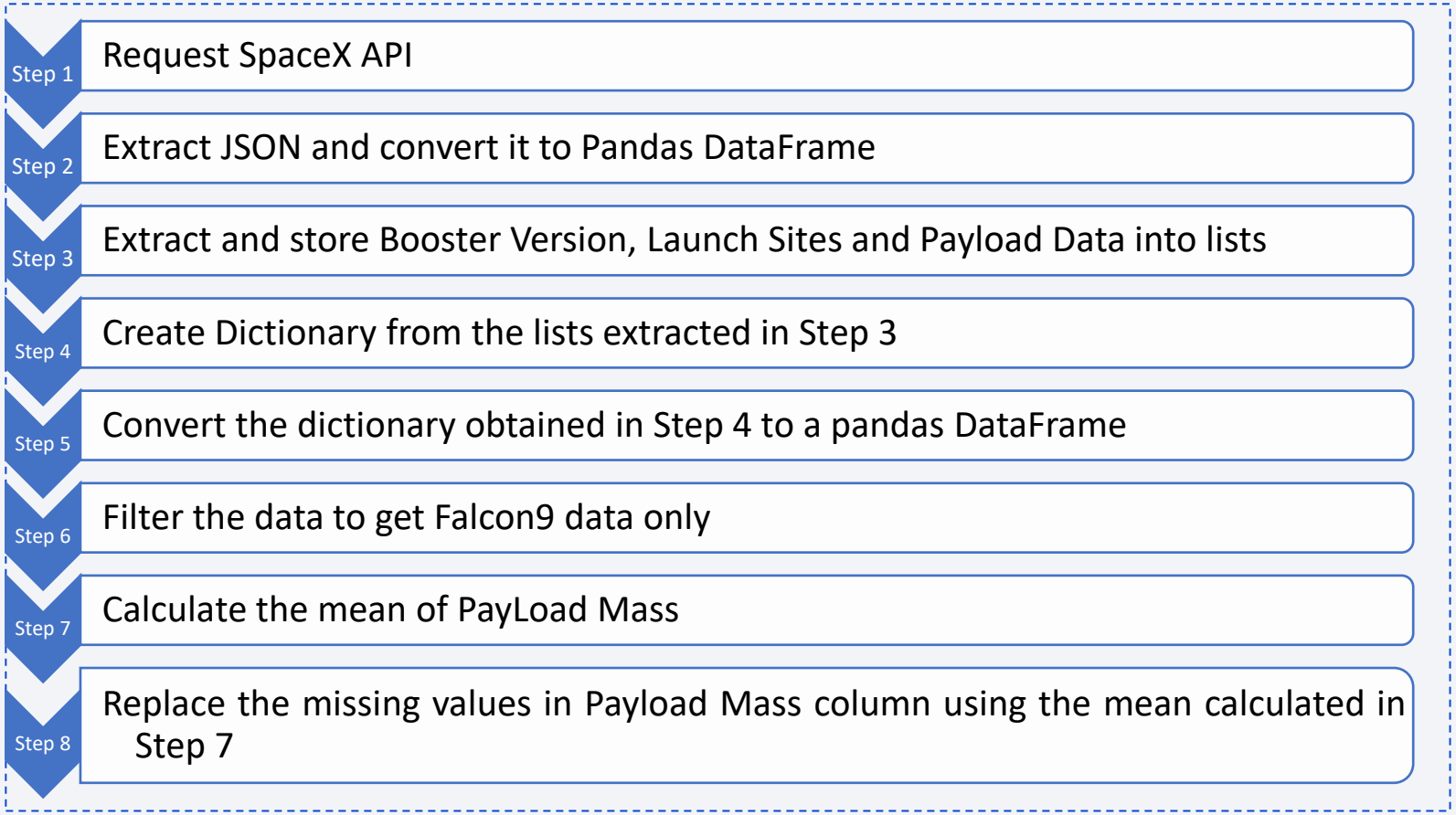
- Data collection methodology:
  - Integration of BeautifulSoup4 and SpaceX API for web scrapping data from SpaceX Wikipedia page
- Perform data wrangling
  - Normalization of data to remove variance. Imputation to replace missing value. Classifying the successful landings as successful while all other landings are classified as unsuccessful
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data is normalized and imputed. Parameter grid is defined and tuned using GridSearchCV with 10 folds

Data Collection

- The data is scrapped and collected from SpaceX Wikipedia page using the SpaceX public API.
- BeautifulSoup library in python allows to parse the data and get the actual contents from the json file obtained by the SpaceX API.
- Defined functions to get the Booster Version, Launch Site, Payload Data and Core Data

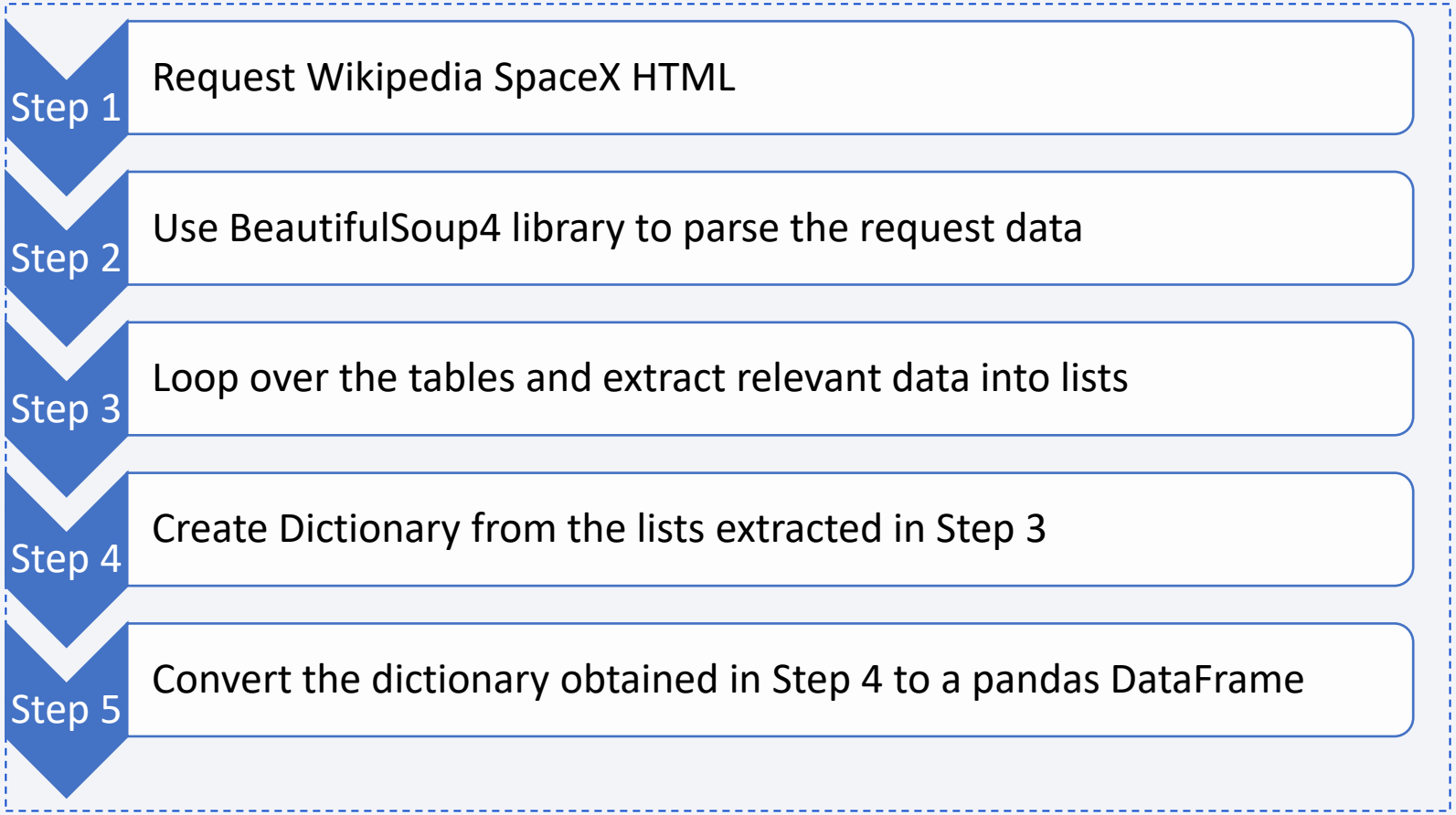
Data Collection – SpaceX API

- GitHub Link : https://github.com/elkang71/IBM-Applied-Data-Science-Professional-Certificate/blob/main/Course%2010%20%E2%94%80%20Applied%20Capstone%20with%20Python/Data%20Collection%20API.ipynb



Data Collection - Scraping

- Github Link : https://github.com/elkang71/IBM-Applied-Data-Science-Professional-Certificate/blob/main/Course%2010%20%E2%94%80%20Applied%20Capstone%20with%20Python/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Calculate number of launches on each site, number of occurrences of each orbit and number + occurrence of mission per orbit type
- Create training column ‘Class’ with successful landings mapped to 1 and unsuccessful landings mapped to 0
- Mapped features with ‘True’ as string to 1 and all the others to 0
- GitHub https://github.com/elkang71/IBM-Applied-Data-Science-Professional-Certificate/blob/main/Course%2010%20%E2%94%80%20Applied%20Capstone%20with%20Python/Data%20Wrangling.ipynb

EDA with Data Visualization

- Use scatter plot to observe the relationship of Payload vs. Outcome, Payload vs. Launch Site, Outcome vs. Orbit and Payload vs. Orbit.
  - The goal is to observe the correlation between two variables each time for which scatter plot works exceptionally well and therefore is used
- Use bar chart to visualize success rate of each type of orbit
  - Bar charts work well with categorical data hence bar chart is used
- Line chart is used to observe relationship between success rate and year
  - Line chart is used as it can show formation of trends
- GitHub
  https://github.com/elkang71/IBM-Applied-Data-Science-Professional-Certificate/blob/main/Course%2010%20%E2%94%80%20Applied%20Capstone%20with%20Python/EDA%20with%20Visualization.ipynb

EDA with SQL

- Data is loaded into the IBM DB2 Database
- Queried the data by SQL integration in Python
- Queries involved exploring the data to get more information about
  - Launch Sites Names
  - Mission Outcomes
  - Payload sizes of customers
  - Payload sizes of booster versions
  - Landing Outcomes
- GitHub
  https://github.com/elkang71/IBM-Applied-Data-Science-Professional-Certificate/blob/main/Course%2010%20%E2%94%80%20Applied%20Capstone%20with%20Python/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Built interactive maps using Folium library
- Added information about Launch sites, successful and unsuccessful landings and locations such as highways and railways located near by
- Circle markers are added for launch sites and the distance lines are added to show the distance between highway and coastline
- The objects added help better understanding the reason for the launch site locations along with being able to better visualize the successful landings corresponding to the locations
- GitHub
  https://github.com/elkang71/IBM-Applied-Data-Science-Professional-Certificate/blob/main/Course%2010%20%E2%94%80%20Applied%20Capstone%20with%20Python/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Dropdown menu, pie chart, interactive slider and line charts are added
- The dropdown menu is added to filter and show the total number of successful launches for launch sites.
- Pie chart helps visualize the results based on dropdown menu options for better understanding and interpretation
- The slider allows to adjust the payload between 0 to 10,000 kg
- The scatter plot helps visualize the correlation between the payload and the launch sites
- GitHub
  https://github.com/elkang71/IBM-Applied-Data-Science-Professional-Certificate/blob/main/Course%2010%20%E2%94%80%20Applied%20Capstone%20with%20Python/spacex_dash_app.py

Predictive Analysis (Classification)

- GitHub https://github.com/elkang71/IBM-Applied-Data-Science-Professional-Certificate/blob/main/Course%2010%20%E2%94%80%20Applied%20Capstone%20with%20Python/Machine%20Learning%20Prediction.ipynb

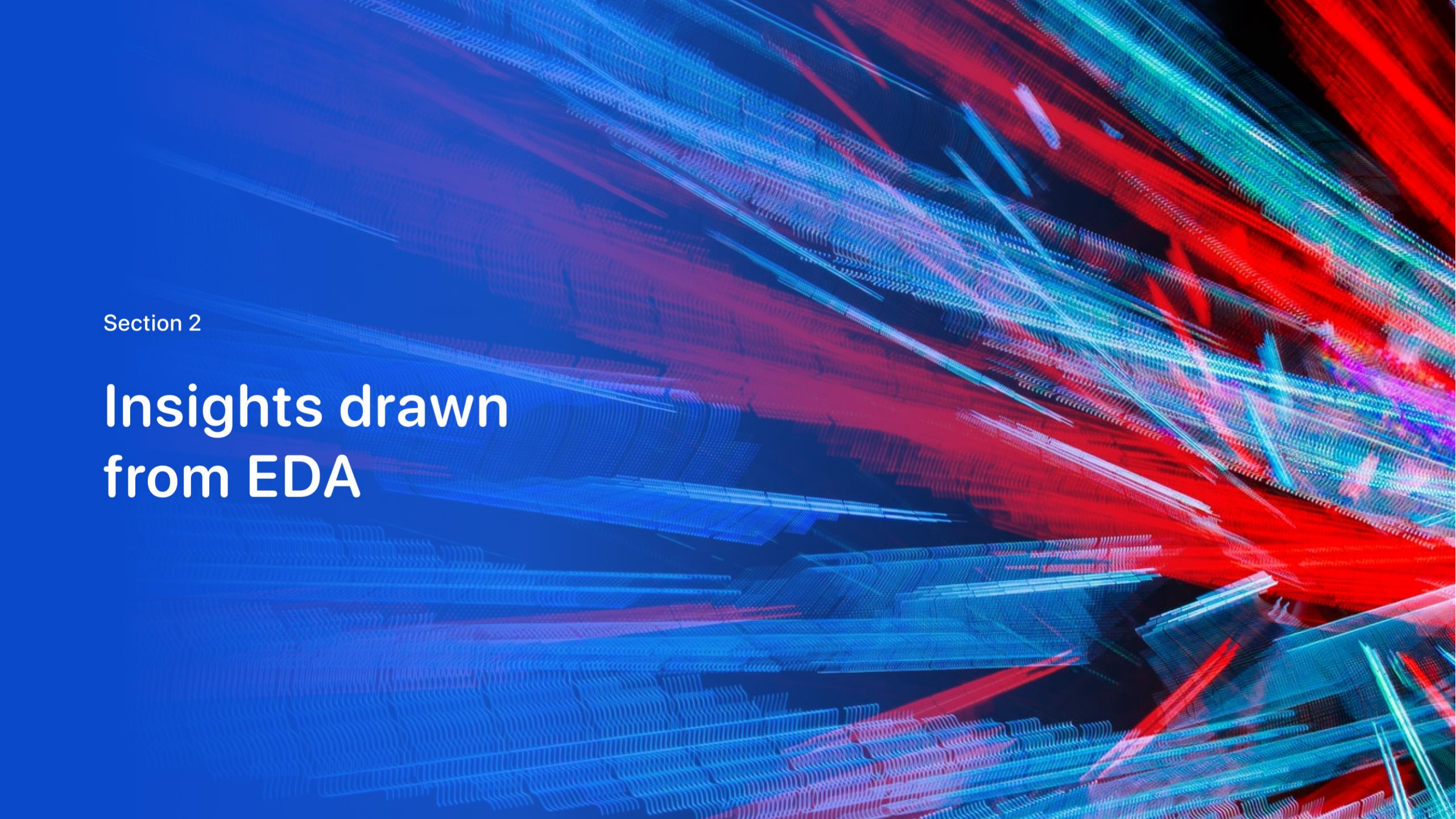
Step 1	Separate the features (X) and the target label (Y)
Step 2	Data is split into train and test sets with 75% as train and 25% as test set
Step 3	The train and test sets were scaled to normal distribution
Step 4	The models are evaluated on accuracy score
Step 5	Parameters were defined for each model to help find the best set of parameters
Step 6	GridSearchCV with 10 folds is used to tune the models with the parameters defined
Step 7	• Accuracy score and confusion matrix are obtained on the test set using best parameters found
Step 8	• The best model is found based on the accuracy score on the test set

Results

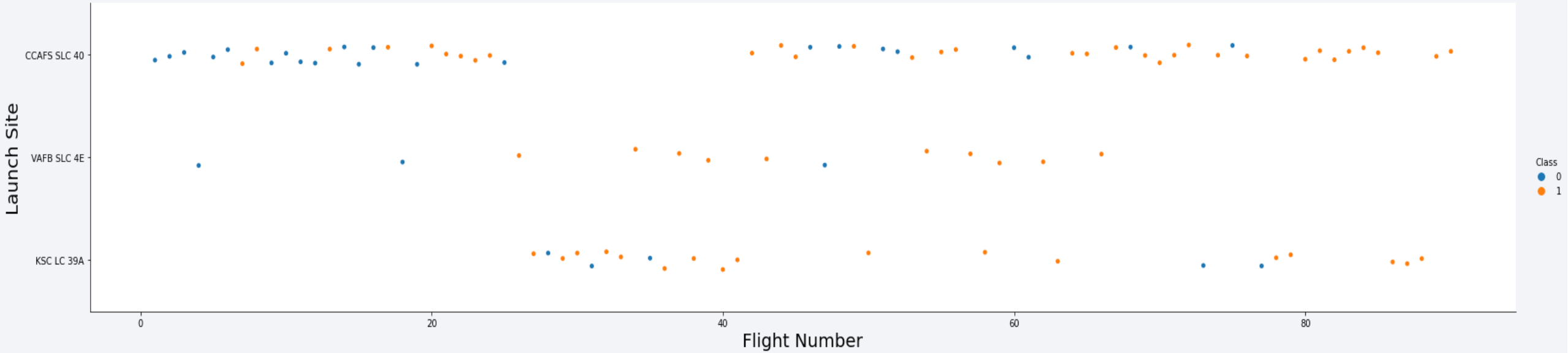
- Exploratory data analysis results
  - As the number of flights increase, the success rate increased
  - The payload is an important factor : higher payloads decrease the chances of success
  - As years passed, the success rate increased
  - The first successful landing came 5 years later in 2015
- Interactive analytics demo in screenshots
  - **CCAFS SLC-40** launch site has the shortest distance to the coastline of 0.86 KM
  - **KSC LC-39A** (center) in Florida has the most successful launches
- Predictive analysis results
  - Data collected in terms of features can be successfully used to predict the landing outcome
  - Logistic Regression, SVM and KNN were able to achieve an accuracy of 83% on the test set
  - More data is needed to improve the performance of the machine learning models and to improve results

Section 2

Insights drawn from EDA

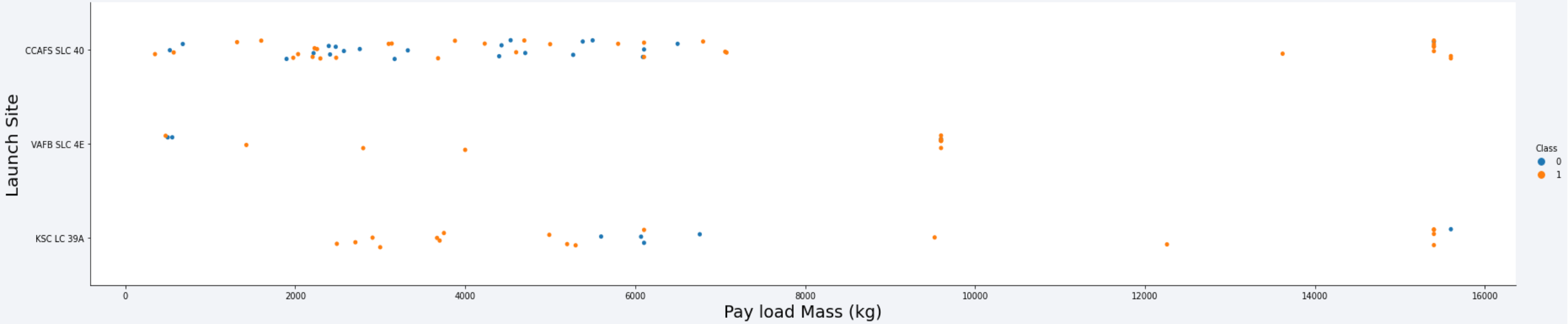


Flight Number vs. Launch Site



The blue dots represent the unsuccessful landings while the orange dots represent the successful landings. The number of unsuccessful landings are quite high in the beginning, but as there are more flights (represented by Flight Number on the X-axis) there are more successful landings. The figure also shows that the **CCAFS SCL 40** launch site has the most number of launches. **VAFB SLC 4E** launch site had the least launches and all the launches seem to be a bit further away from each other for it

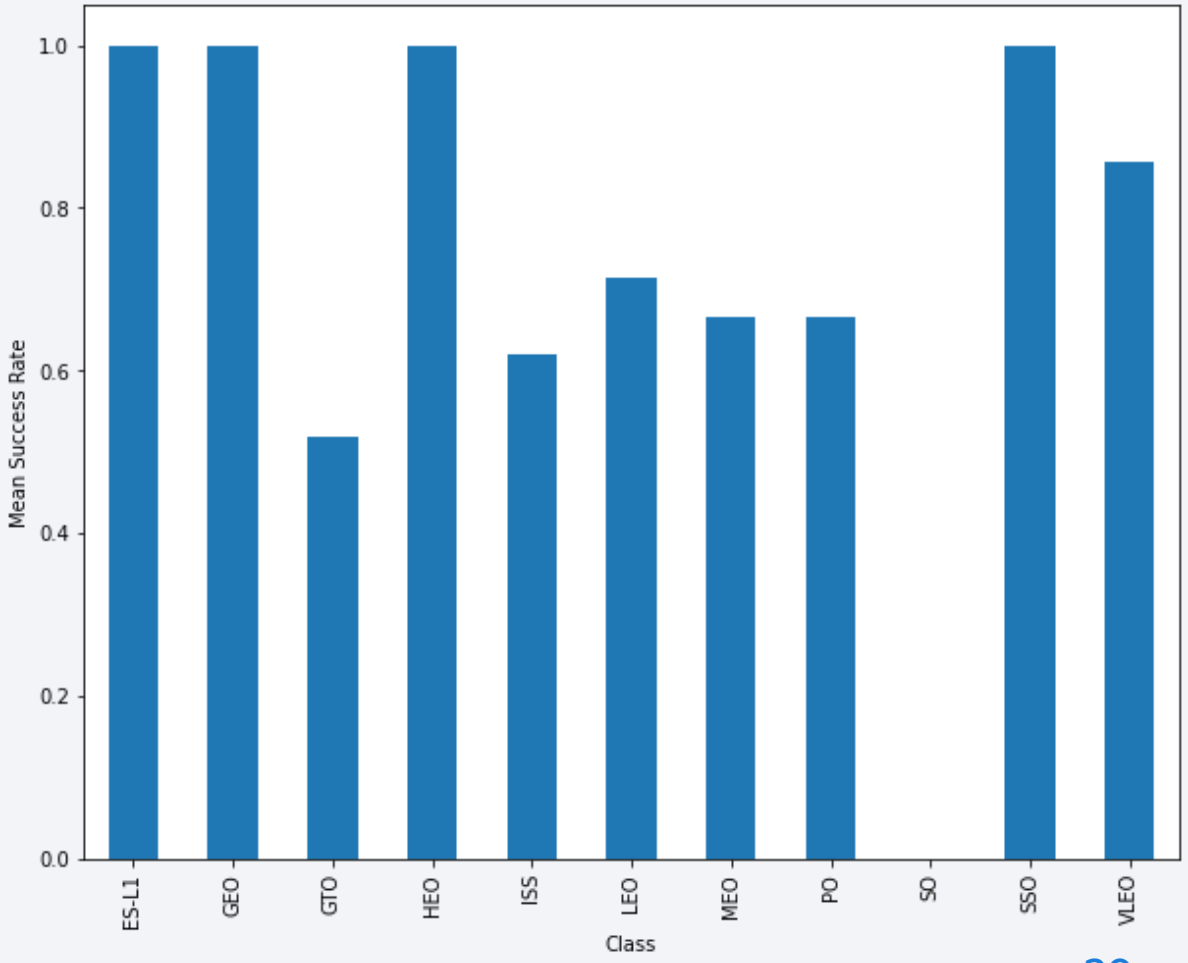
Payload vs. Launch Site



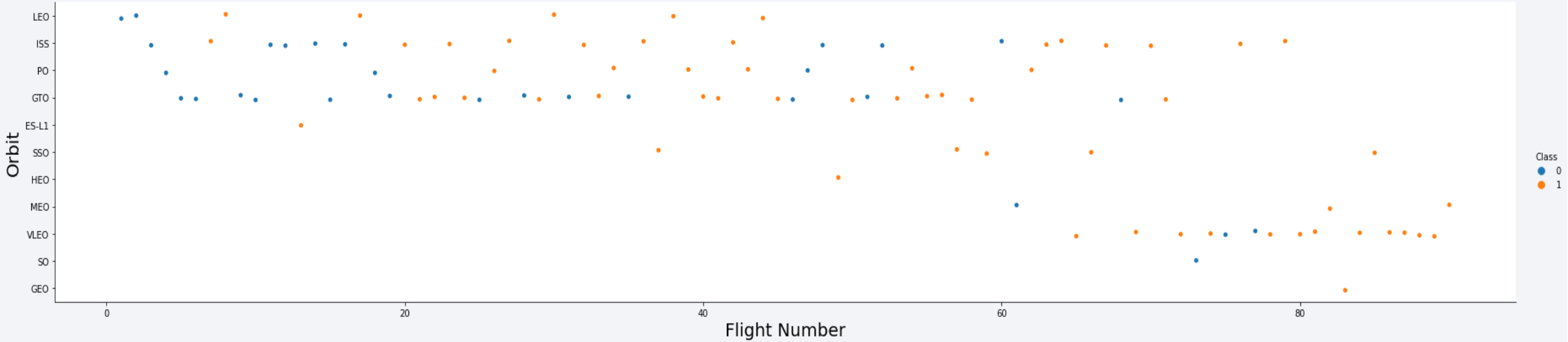
The blue dots represent the unsuccessful landings while the orange dots represent the successful landings. Vast majority of the payloads for **CCAFS SCL 40** launch site are under 8,0000 kg. The payloads greater than 8,000 kg are all successful though. Similar is he case for **KSC LC 39A** except the is one unsuccessful landing for payload greater than 8,000. **VAFB SLC 4E** launch site has no payload that exceeds 10,000 kg.

Success Rate vs. Orbit Type

- **ES-L1**, **GEO**, **HEO** and **SSO** all have the highest success rate of 100%
- **VLEO** has the second highest success rate of around 85%
- **SO** does not seem to have any launches
- All the others have a success rate lower than 80%
- **GTO** has the lowest success rate of around 55%



Flight Number vs. Orbit Type



You should see that in the **LEO** orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in **GTO** orbit.

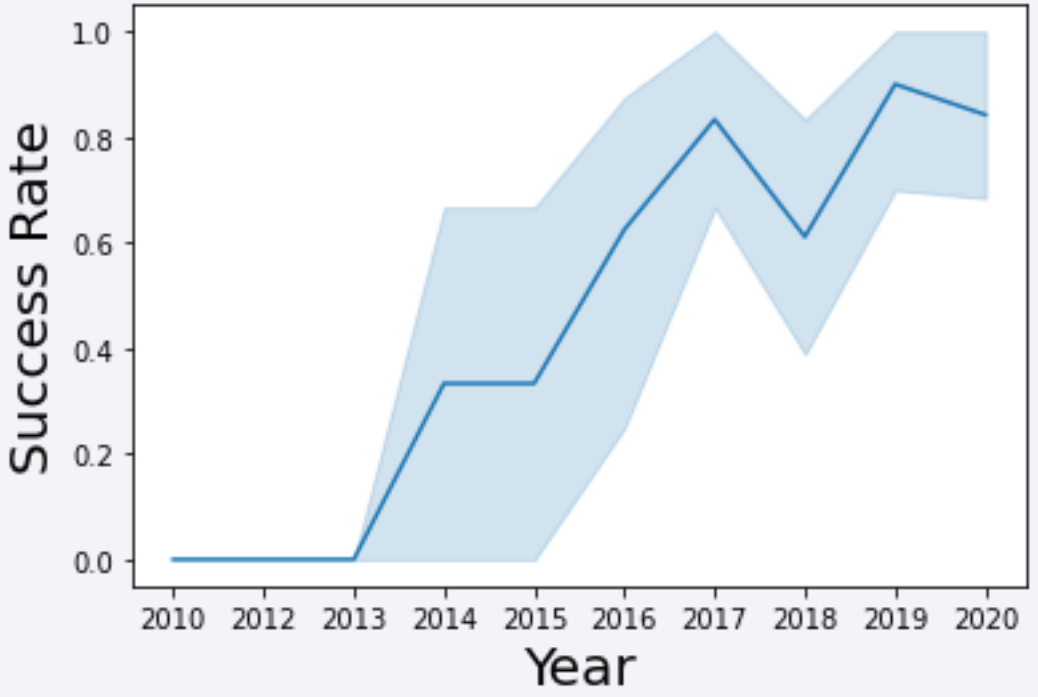
Payload vs. Orbit Type

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

However for **GTO** we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- The trend line shows and upwards trend with the years
- As the years progress, the success rate generally increases
- The exception to the upward trend is from 2013-2015 when the success rate stays constant and in 2018 when the success rate dips



All Launch Site Names

```
1 %sql SELECT UNIQUE LAUNCH_SITE FROM SPACEXTBL
```

```
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

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

- There seem to be 4 unique launch site names

- **CCAFS LC-40** and **CCAFS SLC-40** could very likely be the same launch site with typing error being made at time of entry, however, they could be 2 different launch sites

Launch Site Names Begin with 'CCA'

```
1  %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 first launch from CCAFS LC-40 is in the year 2010
- The first orbits seem to be LEO orbits
- The first five mission outcomes are successful

Total Payload Mass

```
1  %sql SELECT sum(PAYLOAD_MASS__KG_) as "Total Payload Mass by NASA (CRS)" from SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

Total Payload Mass by NASA (CRS)
45596

The total payload mass by **NASA (CRS)** is 45,596 kg

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Total Payload Mass by Booster F9 v1.1" from SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1'
```

```
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

Total Payload Mass by Booster F9 v1.1
2928

The average payload mass by **Booster F9 v1.1** is 2,928 kg

First Successful Ground Landing Date

```
1 %sql SELECT DATE AS "First successful landing outcome Date" FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE '%Success %' LIMIT 1
```

```
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

First successful landing outcome Date
2015-12-22

- The first successful landing outcome is on 2015-12-22
- This is almost 5 years after the first mission launched which launched in 2010

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT UNIQUE BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000
```

```
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

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

- There seem to be five booster versions where the landing outcome was successful by a drone ship
- All the booster versions are variations of **F9**

Total Number of Successful and Failure Mission Outcomes

```
1 %sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS "Count Mission Outcomes" FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

mission_outcome	Count Mission Outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- There seem to be only 1 failure and 1 success with payload status unclear
- The success rate seems to be 99%

Boosters Carried Maximum Payload

```
1 %sql SELECT UNIQUE BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- There are 12 booster versions
- All the booster versions seem to version of **F9**

2015 Launch Records

```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME LIKE 'Failure (drone ship)' AND DATE LIKE '2015%'
```

```
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- There are only 2 records for year 2015
- The landing outcomes seem to be failure on drone ship
- Booster version is **F9 v1.1**
- The launch site for both launches is **CCAFS LC-40**

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

- There are 31records between 2010-06-04 and 2017-03-20
- Booster version is **F9 v1.1**
- There are 8 records on successful landing
- There are 7 records of unsuccessful landing
- There are 9 records where no attempt at landing is made

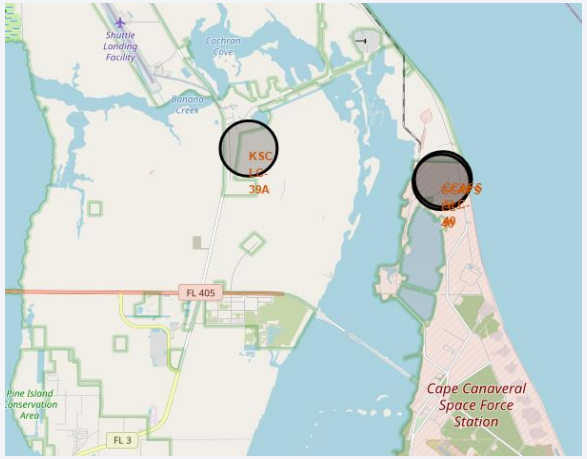
```
1  %sql SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC
 * ibm_db_sa://ktn73860:***@b0aebb68-94fa-46ec-a1fc-1c999edb6187.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:31249/bludb
Done.
```

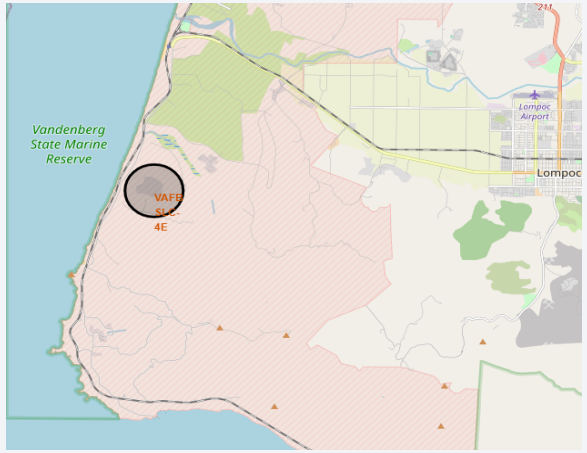
landing__outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt
Failure (drone ship)
No attempt
Controlled (ocean)
Failure (drone ship)
Uncontrolled (ocean)
No attempt
No attempt
Controlled (ocean)
Controlled (ocean)
No attempt
No attempt
Uncontrolled (ocean)
No attempt
No attempt
No attempt
Failure (parachute)
Failure (parachute)

Section 4

Launch Sites Proximities Analysis

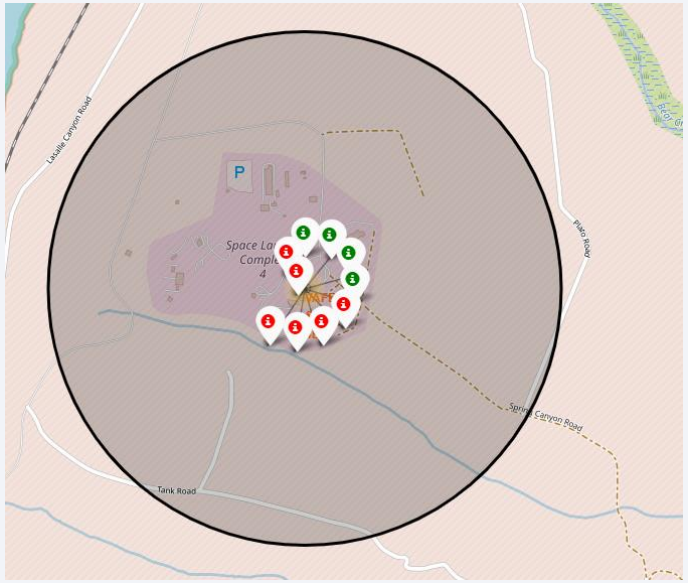
Launch Site Location

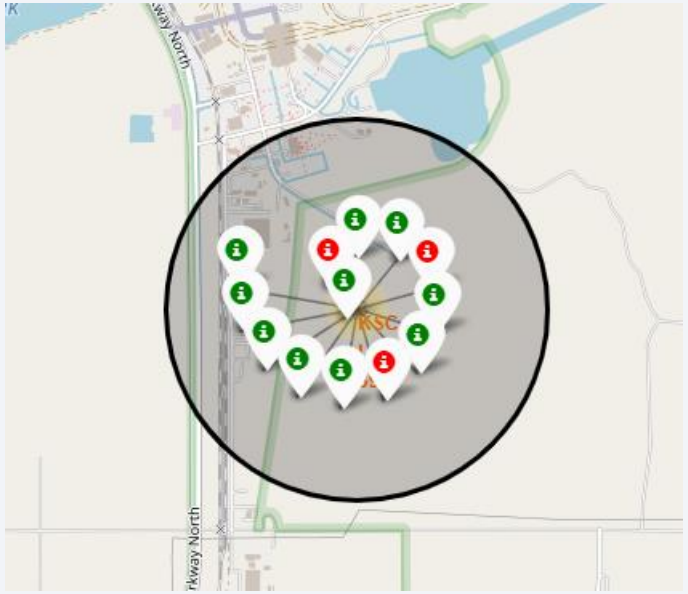


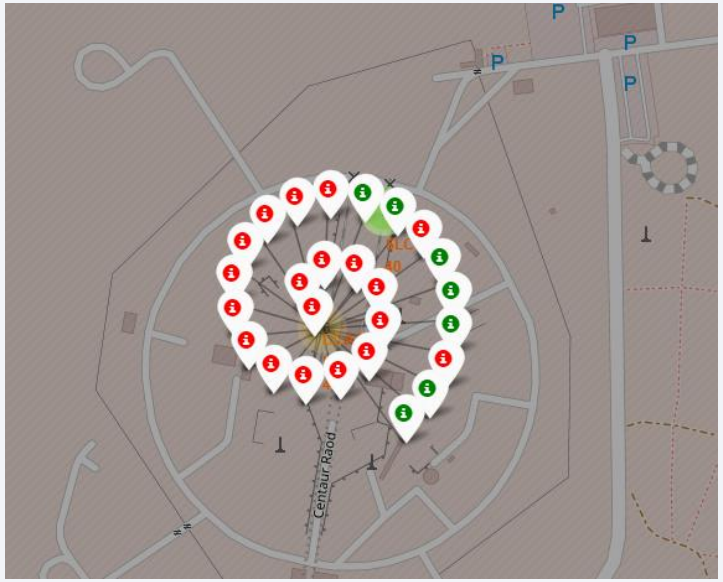


- There are 3 launch sites
- 2 are in Florida
- 1 is in Los Angeles

Landing Outcomes for Each Site

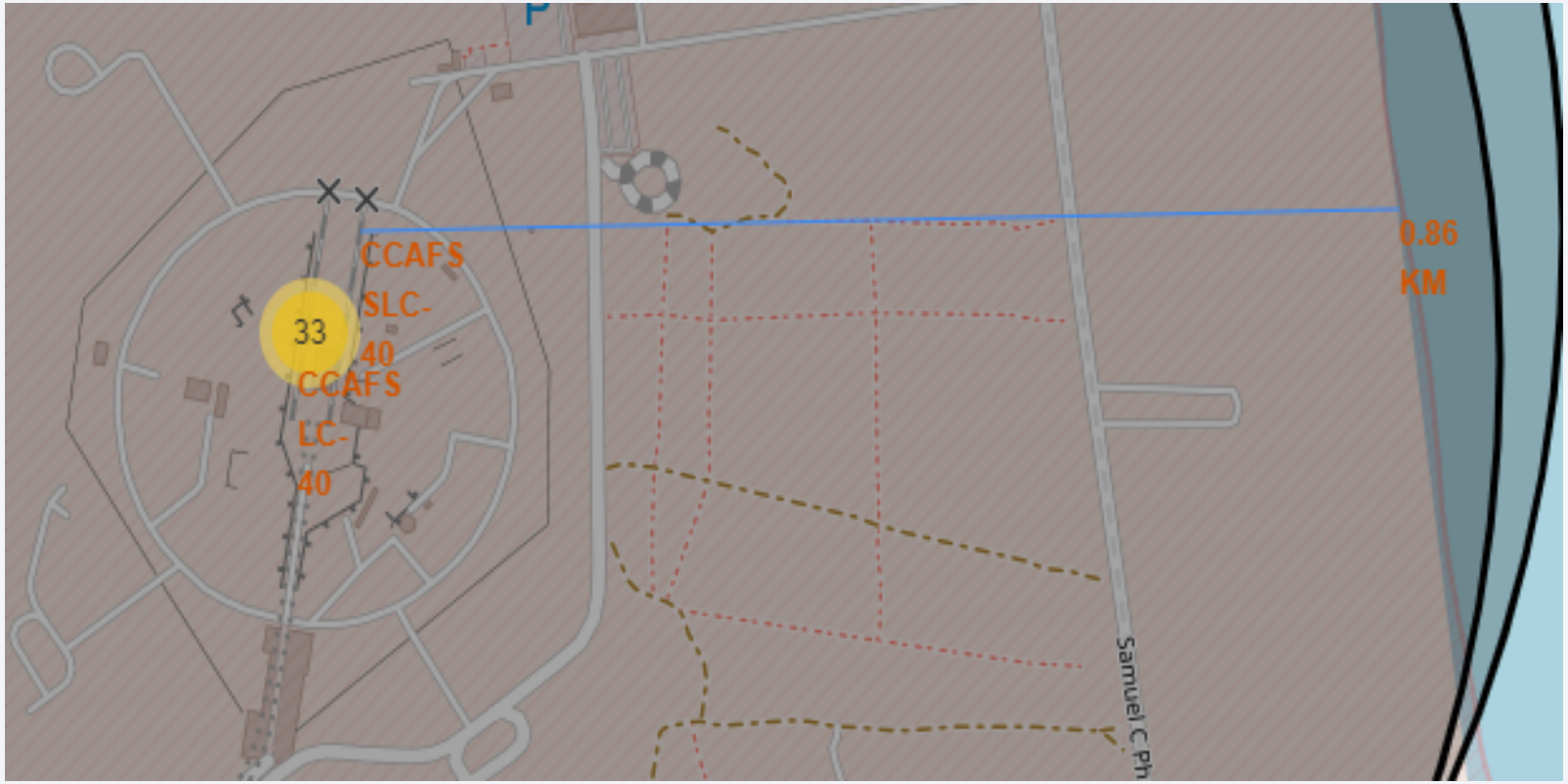






- **CCFAS SLC-40** (right) in Florida seems to have the most launches
- **VAFB SLC-4E** (left) in Los Angeles has comparatively fewer launches
- **CCFAS SLC-40** (right) has the highest launches with also the most unsuccessful launches (red)
- **KSC LC-39A** (center) in Florida has the most successful launches (green)

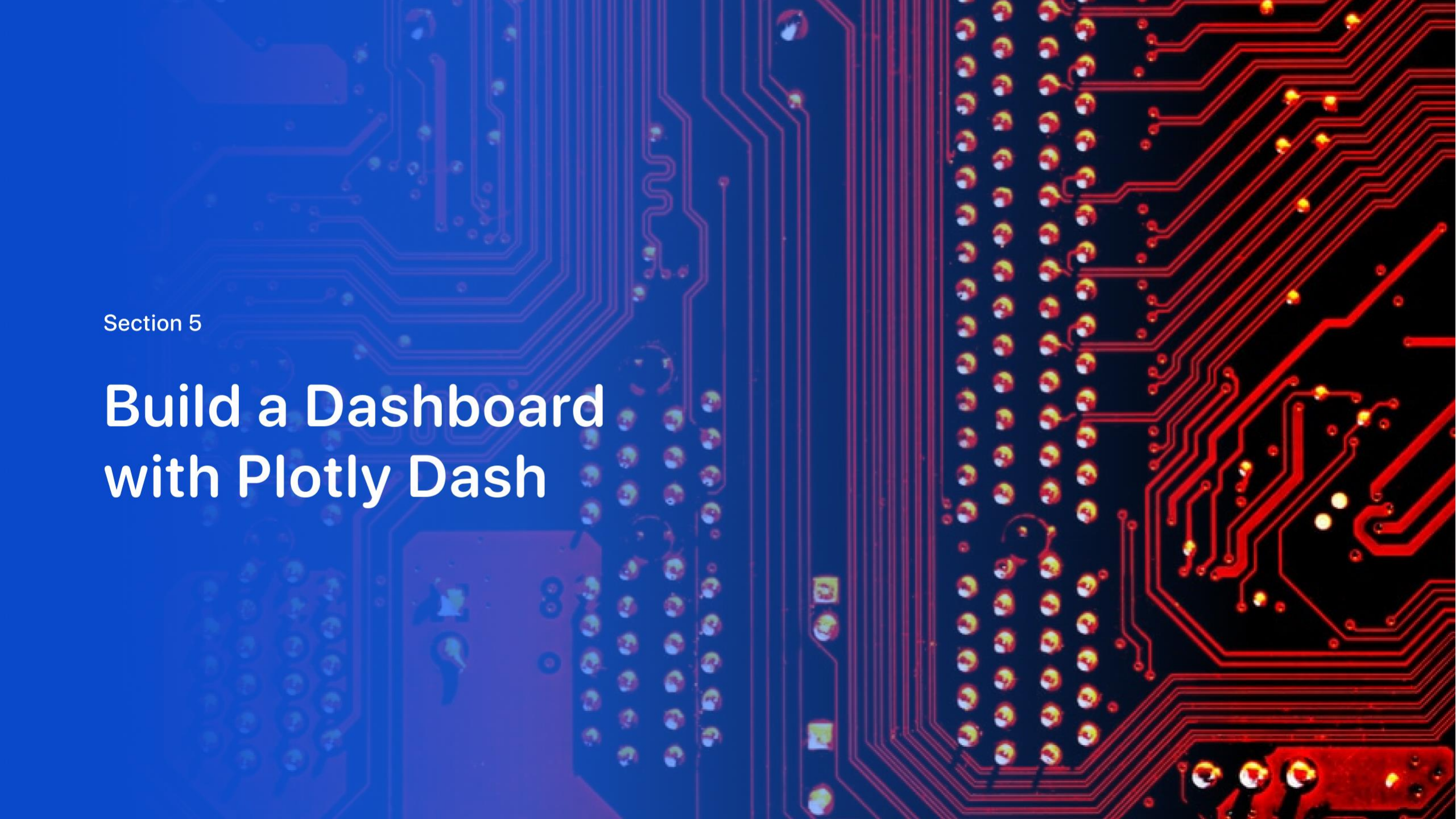
Shortest Distances from Launch Site



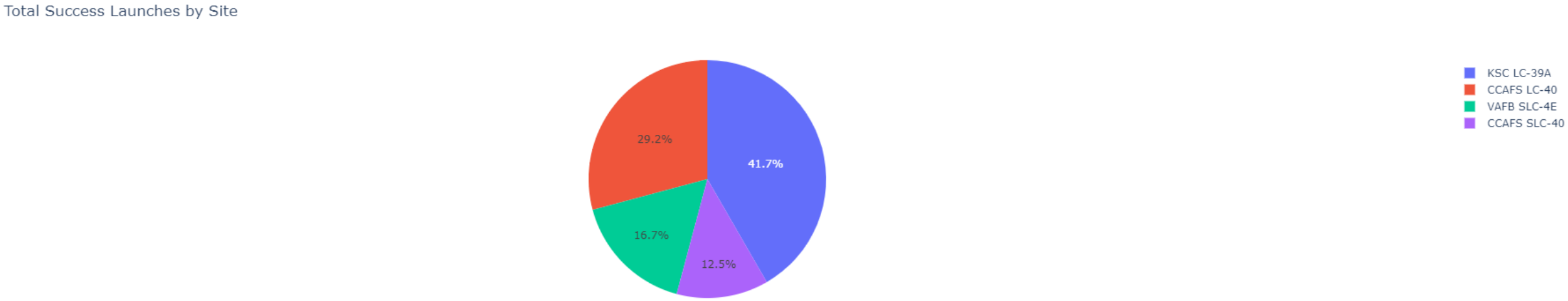
The shortest distance from the launch site **CCAFS SLC-40** to the coastline is 0.86 KM

Section 5

Build a Dashboard with Plotly Dash

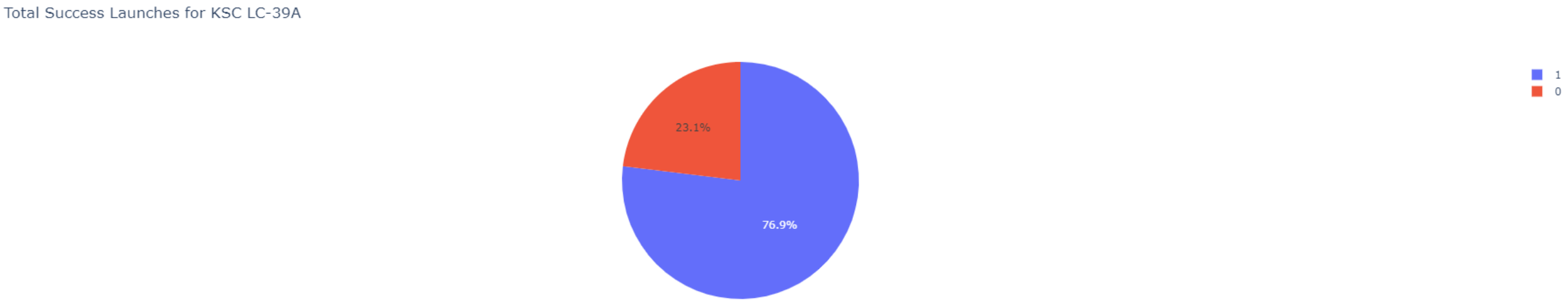


Total Success Launches for All Sites



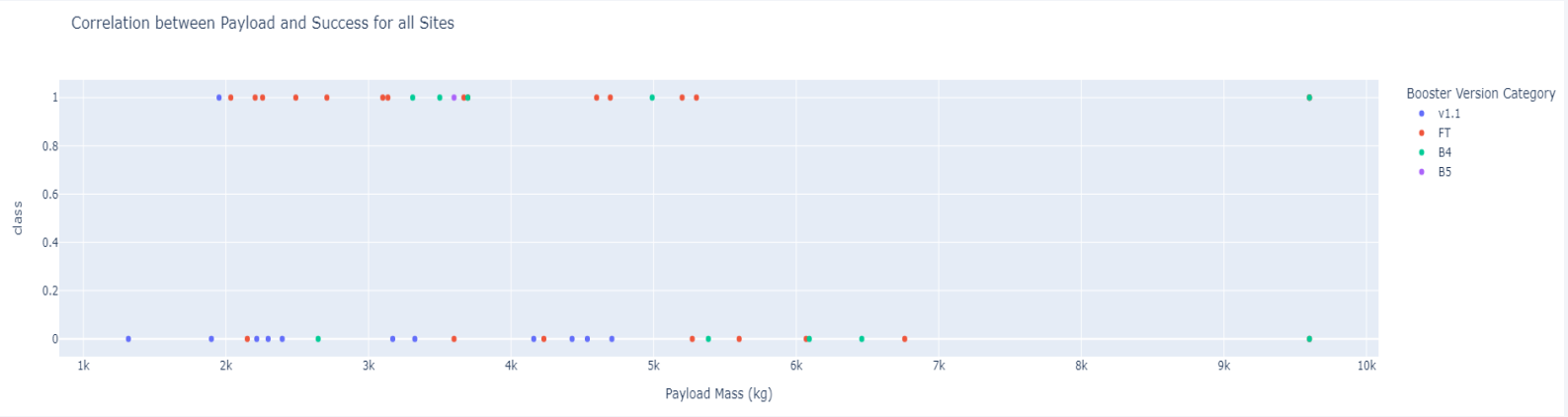
- **KSC LC-39A** has the highest success rate of 41.7%
- **CCAFS SCL-40** has the lowest success rate of 12.5%

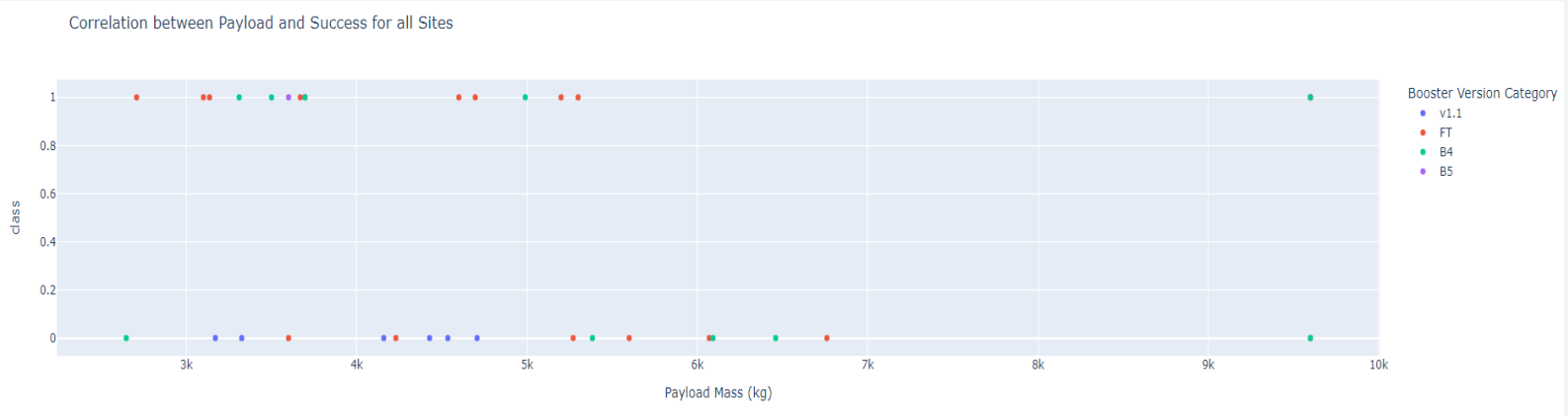
Pie Chart of KSC LC-39A



- There are 76.9% successful launches
- Only 23.1%R launches were unsuccessful

Payload vs. Launch Outcome



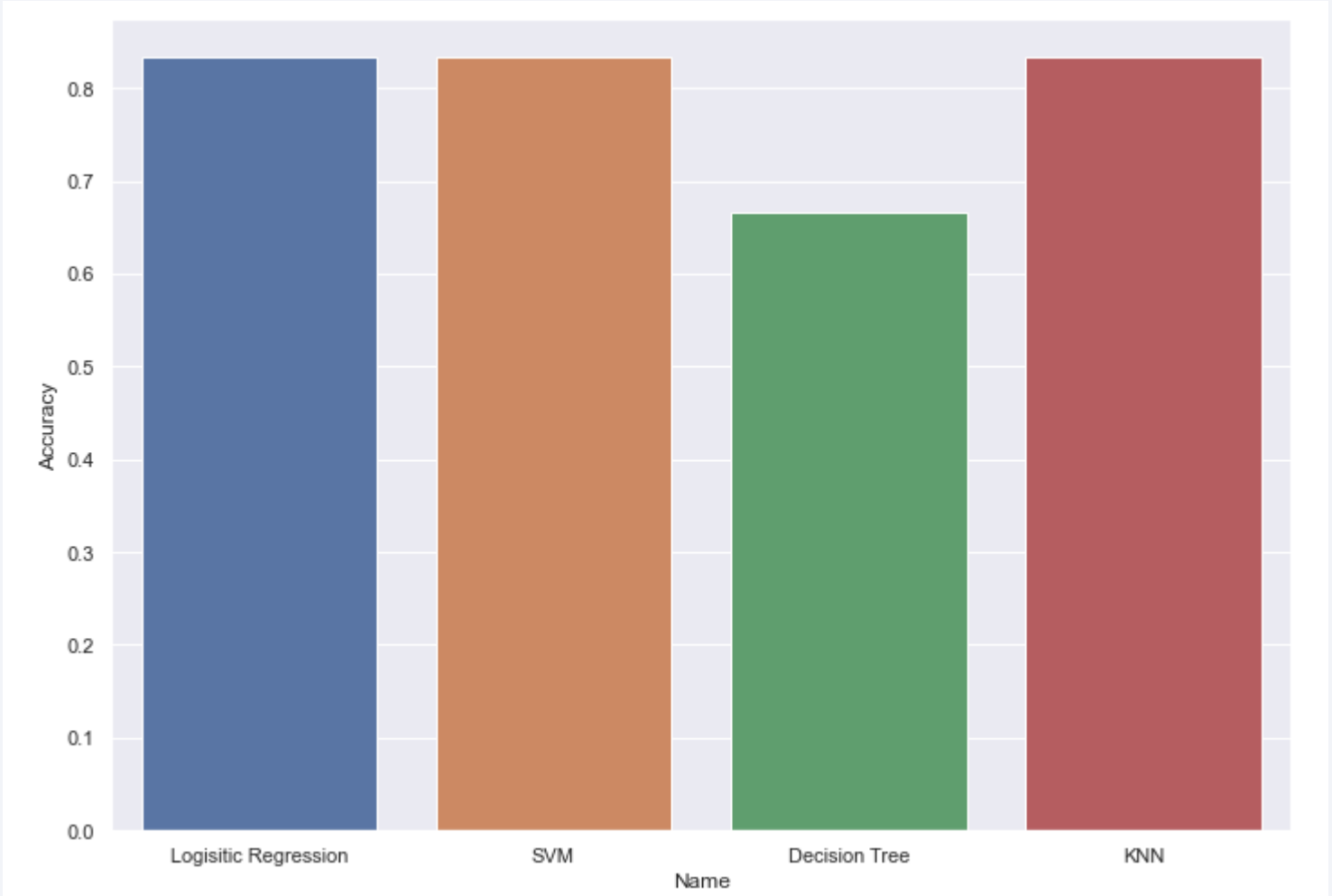


- Payload vs. Launch outcome for all sites with Payload set to 1,000 kg (top) and Payload vs. Launch outcome for all sites with Payload set to 2,500 kg (bottom)
- As the Payload increases, the success rate decreases for all sites

Section 6

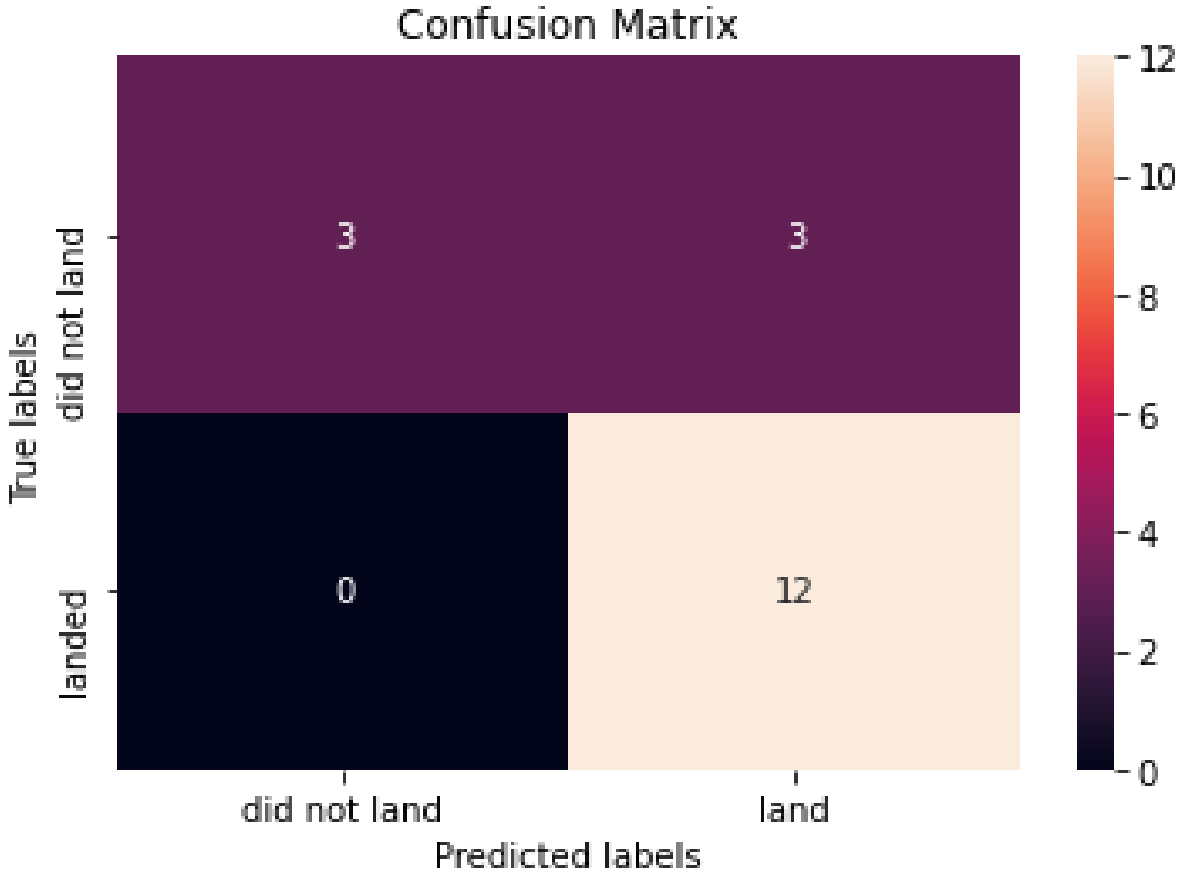
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree performs the worst on the test set with accuracy of 67%
- All the other models perform same with accuracy of 83%
- The same accuracy can be attested to the very small test size
- Training and testing on more data would be yield better insights and results

Confusion Matrix



- There are two classes : did not land (negative) and landed (positive)
- From the confusion matrix we see that there are no False Negatives (landed but predicted as did not land)
- There are only 3 False Positives (did not land but predicted as landed)
- The test data was very small with only 18 samples
- The confusion matrix is for Logistic Regression, SVM and KNN as they performed the same

Conclusions

- Payload Mass seems to be inversely proportional to the success rate : as the Payload Mass increases the success rate of landing decreases
- The site of launch is also vital. **KSC LC-39A** located in Florida has the highest success rate
- As time progressed, the number of flights increases and with that so does the success rate
- Features in the collected data can be used to successfully predict the outcome of landing using various Machine Learning algorithms
- More data needs to be collected to improve the performance of Machine Learning models

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!