IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eric Martín García
08/03/2023

Outline

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

Executive Summary

- The datasets were collected using API calls and web scrapping
- Then the data were cleaned and prepared to analyze it, creating the target variable from the info gathered.
- On the EDA were founded some insights of the data and the features that provide useful information to predict the target variable:
  - Payload Mass, Orbit and Launch Site are good features to predict the target variable
  - The success rate increased considerably since 2015, the year in which the first successful landing took place.
  - The launch site is normally near the coast and far from cities, with some roads and railways near.
- Then, a large predictive analysis was executed using the models LogisticRegression, SVM, DecisionTreeClassifier and K-NN with multiple hyperparameters resulting the DecisionTree the best model with an accuracy of 0.88 on the test set.

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone.
- SpaceX Falcon 9 rocket launches 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.
- The objective of this project is to determine if SpaceX will reuse the first stage.
- A machine learning model will be trained using public information to predict this outcome.

Section 1

Methodology

Methodology

Executive Summary

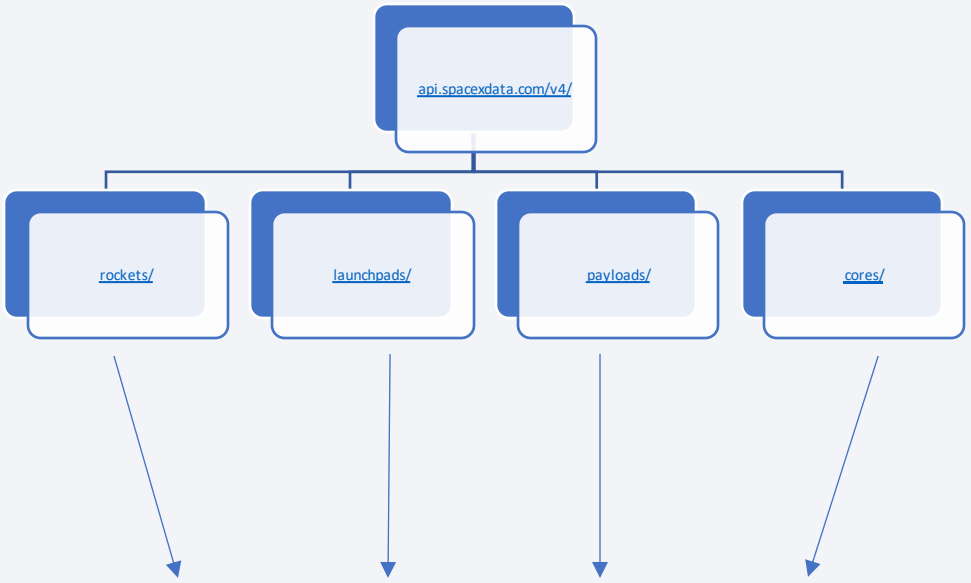
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The datasets were collected using two methods:
  - API calls to the SpaceX REST API.
    - URL: https://api.spacexdata.com/v4/
  - Web Scrapping from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches".
    - URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

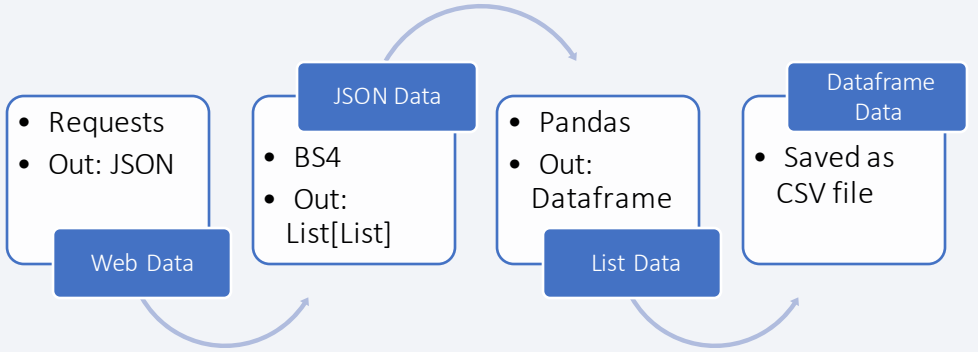
- The data were gathered using API calls through the 'requests' module on Python.
- Different endpoints were used from the SpaceX API.
- The combination of their responses was added to a Pandas Dataframe and saved to a csv file.
- https://github.com/ericmg97/ibm-data-science/blob/master/notebooks/Data%20Collection%20API.ipynb



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data Collection - Scraping

- The data were gathered from the Wikipedia page: List of Falcon 9 and Falcon Heavy launches using the 'requests' module on Python.
- The output response was cleaned and processed with the 'BeautifulSoup' (BS4) module.
- The cleaned data were added to a Pandas Dataframe and saved to a csv file.
- https://github.com/ericmg97/ibm-data-science/blob/master/notebooks/Data%20Collection%20Web%20Scraping.ipynb



Data Wrangling

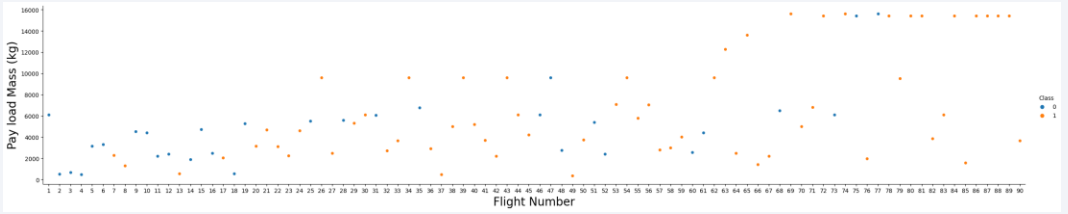
- The data were processed using the column 'Outcome' to generate the target variable named 'Class'.
- The 'Outcome' column indicates if the first stage successfully landed.
- The target variable 'Class' was added to a Dataframe corresponding to the 'Outcome' column as shown in the image.
- The new Dataframe was saved to a csv file.
- https://github.com/ericmg97/ibm-data-science/blob/master/notebooks/Data%20Wrangling.ipynb

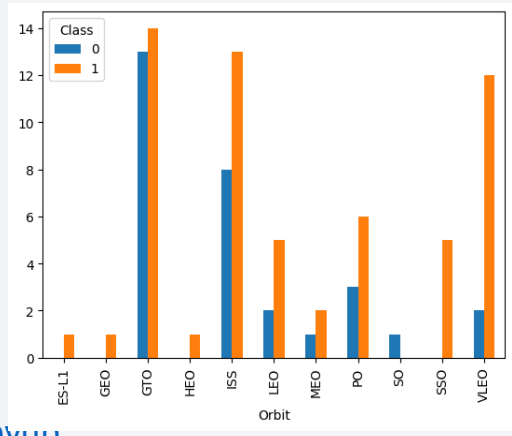
Outcome	Class
True ASDS	1
None None	0
True RTLS	1
False ASDS	0
True Ocean	1
False Ocean	0
None ASDS	0
False RTLS	0

EDA with Data Visualization

- The charts that were used to visualize the data:
  - Line Plot showing:
    - Yearly trend of success rate
  - Bar Chart showing relationship:
    - Success Rate – Orbit
  - Scatter Plot showing relationship:
    - Payload Mass (Kg) - Flight Number - Class
    - Orbit - Flight Number - Class
    - Launch Site - Flight Number - Class
- https://github.com/ericmg97/ibm-data-science/blob/master/notebooks/EDA%20Data%20Visualization.ipynb

Examples:





EDA with SQL

- Multiple SQL queries were performed to analyze the data:
  - The names of the unique launch sites in the space mission.
  - Five records where launch sites begin with the string 'CCA'.
  - The total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by booster version F9 v1.1.
  - The date when the first successful landing outcome in ground pad was achieved.
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - The total number of successful and failure mission outcomes.
  - The names of the 'booster_versions' which have carried the maximum payload mass.
  - The failed 'landing_outcomes' in drone ship, their booster versions, and launch site names for in year 2015.
  - Rank of 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.
- https://github.com/ericmg97/ibm-data-science/blob/master/notebooks/EDA%20SQL.ipynb

Build an Interactive Map with Folium

- To analyze launch site geo and proximities the module Folium was used on Python.
- The launch site locations and their close proximities were marked using Markers, Circles, MarkerClusters and PolyLines on an interactive map.
- Then, the map with those markers was explored and were discovered some patterns from them.
- The launch site is normally near the coast and far from cities, with some roads and railways near.
- https://github.com/ericmg97/ibm-data-science/blob/master/notebooks/Launch%20Site%20Visualization%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- In the pie chart the total successful launches count are shown for all sites or an specific site selected.
- In the scatter chart the correlation between payload and launch success are shown for all sites or an specific site selected. Also giving the possibility of filtering by payload mass.
- With these plots is possible to obtain some insights to answer, by example, the following questions in real-time:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
- https://github.com/ericmg97/ibm-data-science/blob/master/apps/spacex_dash_app.py

Predictive Analysis (Classification)

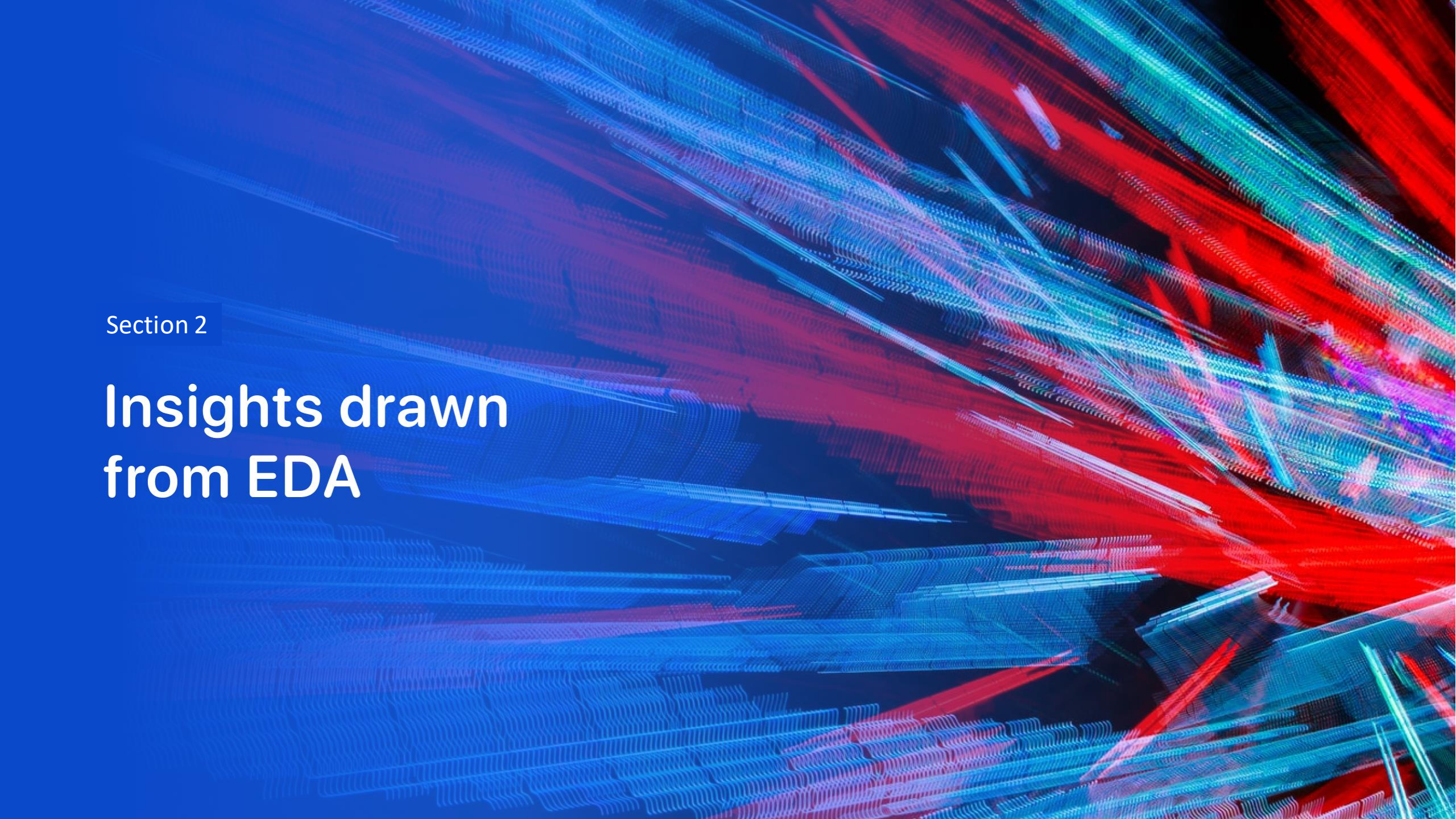
- A machine learning pipeline was built to predict if the first stage of the Falcon 9 lands successfully.
- Pipeline:
  - Standardize the data and split it into train and test sets.
  - Train the model and perform Grid Search to find the hyperparameters that allow a given algorithm to perform best.
  - Using the best hyperparameter values, determine the model with the best accuracy using the training data.
  - Output the confusion matrix and the score for each model.
  - Finally, select the model with the best score.
- Models:
  - Logistic Regression
  - Support Vector machines
  - Decision Tree Classifier
  - K-nearest neighbors.
- https://github.com/ericmg97/ibm-data-science/blob/master/notebooks/Machine%20Learning%20Prediction.ipynb

Results

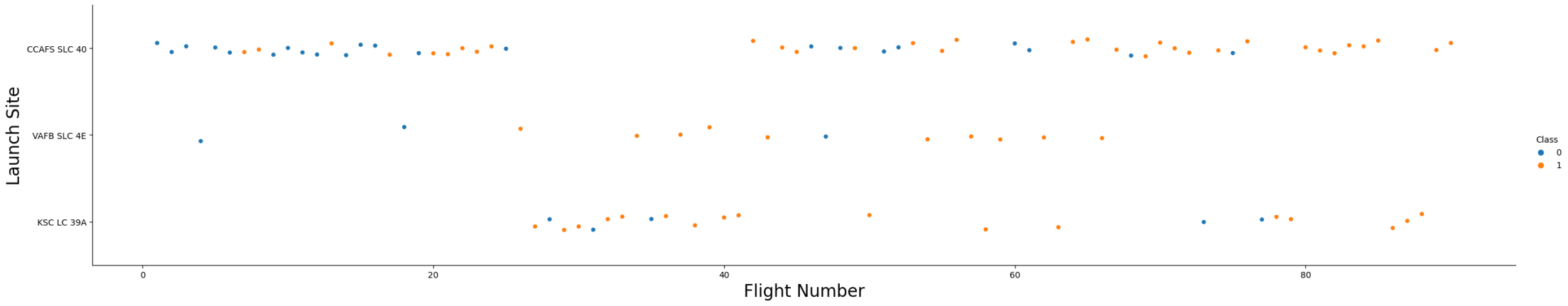
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

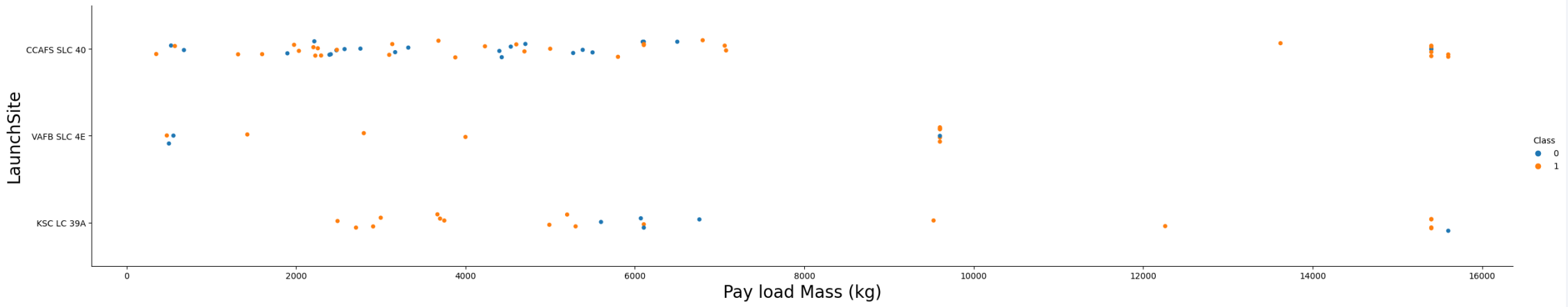


Flight Number vs. Launch Site



- At launch site 'VAFB SLC 4E' there are fewer launches than at sites 'CCAFS SLC 40' and 'KSC LC 39A' but has a great success rate.
- Globally, when the flight number increases, the success rate increases too.

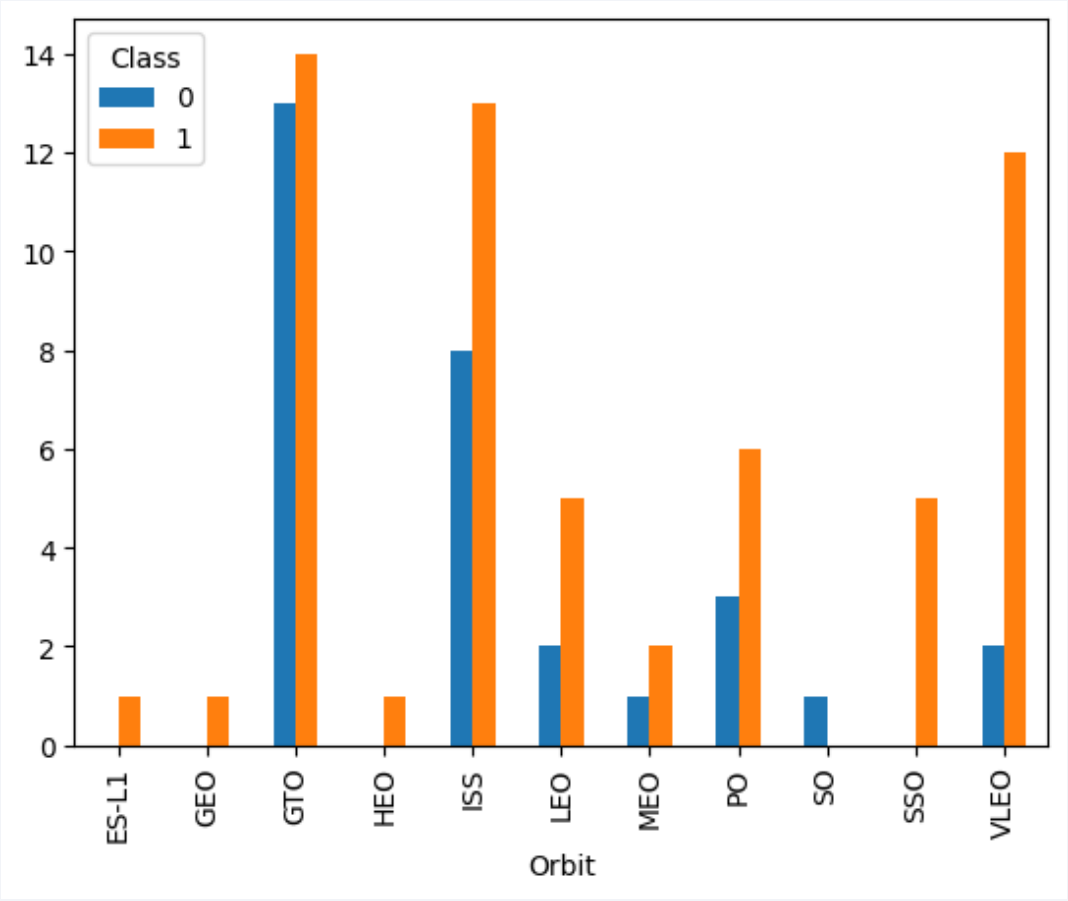
Payload vs. Launch Site



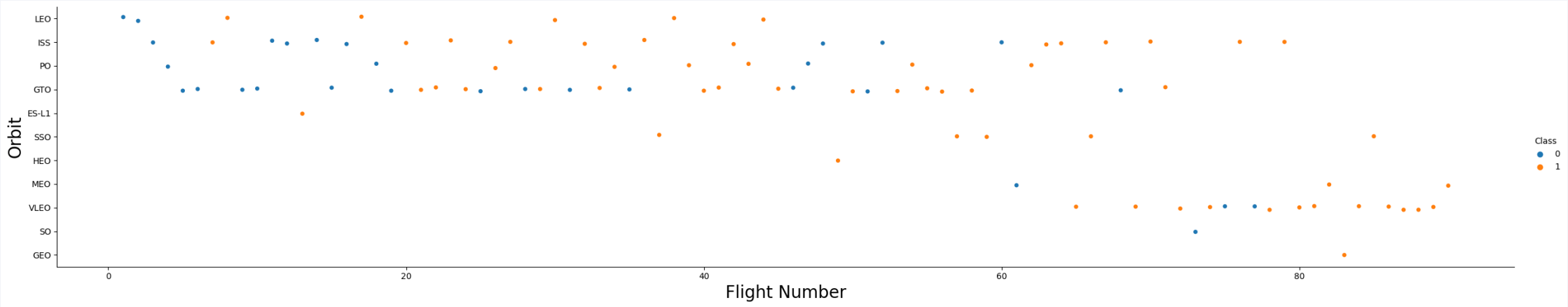
- For the 'VAFB SLC 4E' launch site there are no rockets launched for heavy payload mass (greater than 10000).
- When the payload mass increases, the success rate increases too.

Success Rate vs. Orbit Type

- The orbits GTO and ISS are the most used but have a lower success rate (<65%) compared to the others.
- The orbit VLEO has a good success rate even being one of the most used.
- The orbit SO was used only once and failed.

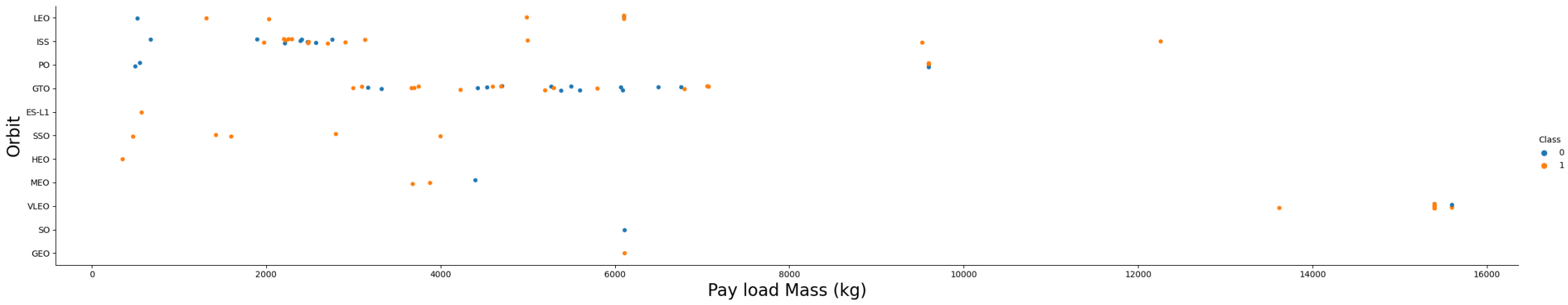


Flight Number vs. Orbit Type



- In the LEO orbit, the success rate appears related to the few number of flights made.
- There seems to be no relationship between flight number when in GTO orbit.
- Since the flight No. 78 approx, there are no failure launches.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO is not possible to distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- The success rate increased considerably since 2014.
- In the period between 2018 and 2019 the success rate temporarily decreased.

All Launch Site Names

```
launch_sites = %sql select launch_site from spacex group by launch_site
launch_sites
```

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

Launch Site Names Begin with 'CCA'

```
launch_sites_cca = %sql select * from spacex where launch_site like 'CCA%' limit 5
launch_sites_cca
```

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

Total Payload Mass

```
payload_mass = %sql select SUM(payload_mass__kg_) from spacex where customer = 'NASA (CRS)'
payload_mass
```

1
45596

Average Payload Mass by F9 v1.1

```
payload_avg = %sql select AVG(payload_mass__kg_) from spacex where booster_version like 'F9 v1.1%'
payload_avg
```

1
2534

First Successful Ground Landing Date

```
first_landing = %sql select min(DATE) from spacex where landing__outcome = 'Success (ground pad)'
first_landing
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
boosters_success = %sql select booster_version from spacex where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
boosters_success
```

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

Total Number of Successful and Failure Mission Outcomes

```
mission_outcomes = %sql select mission_outcome, count(*) as count from spacex group by mission_outcome
mission_outcomes
```

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
boosters_max_payload = %sql select booster_version from spacex where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
boosters_max_payload
```

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

```
failed_landings = %sql select booster_version, launch_site from spacex where landing__outcome = 'Failure (drone ship)' and year(DATE) = 2015
failed_landings
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

```
landing_rank = %sql select landing__outcome as outcome, count(*) as count from spacex where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome order by count desc
landing_rank
```

outcome	COUNT
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

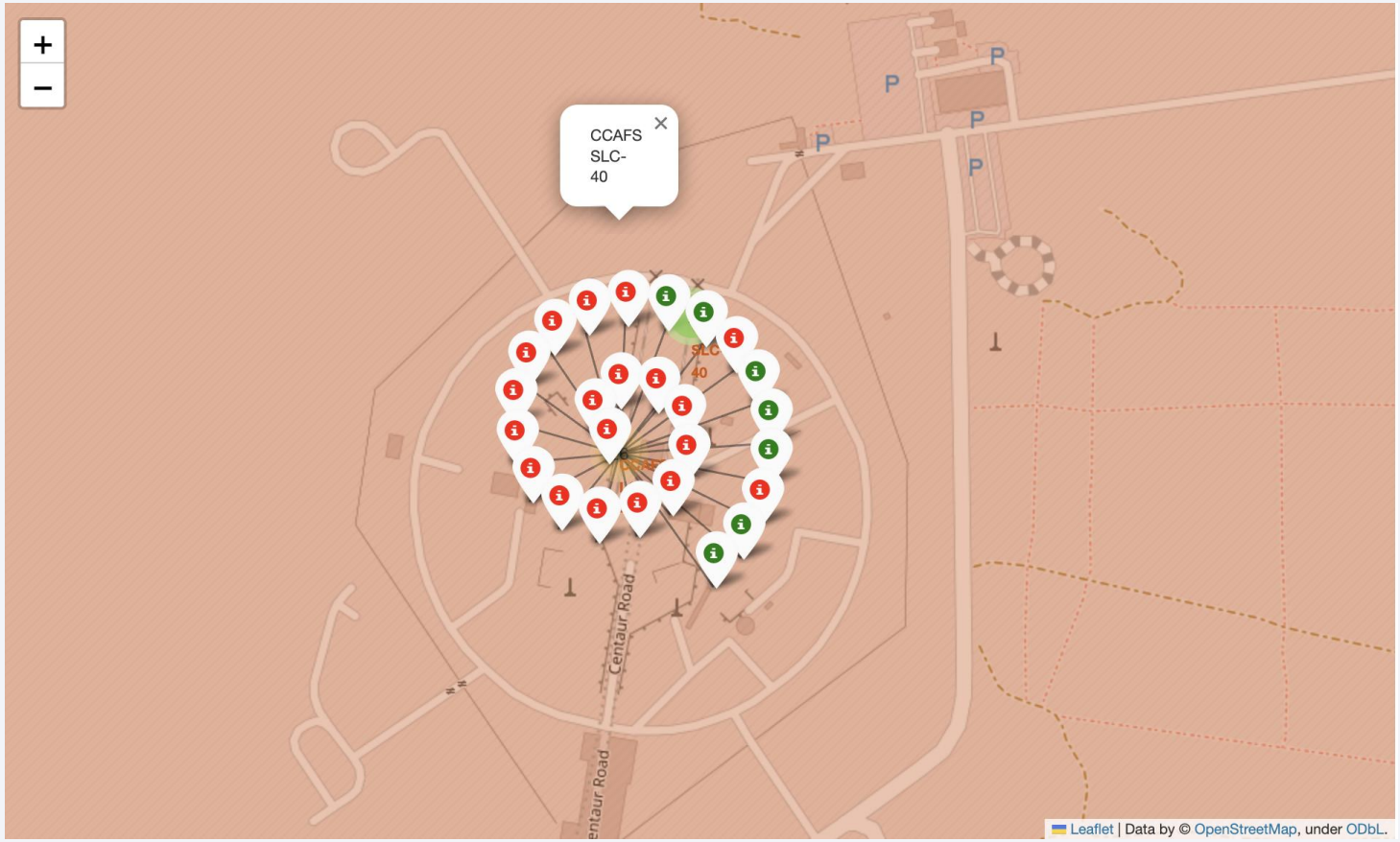

Section 3
Launch Sites
Proximities Analysis

Launch Sites Locations

- All launch sites are close to the coast
- All launch sites are close to the equator line

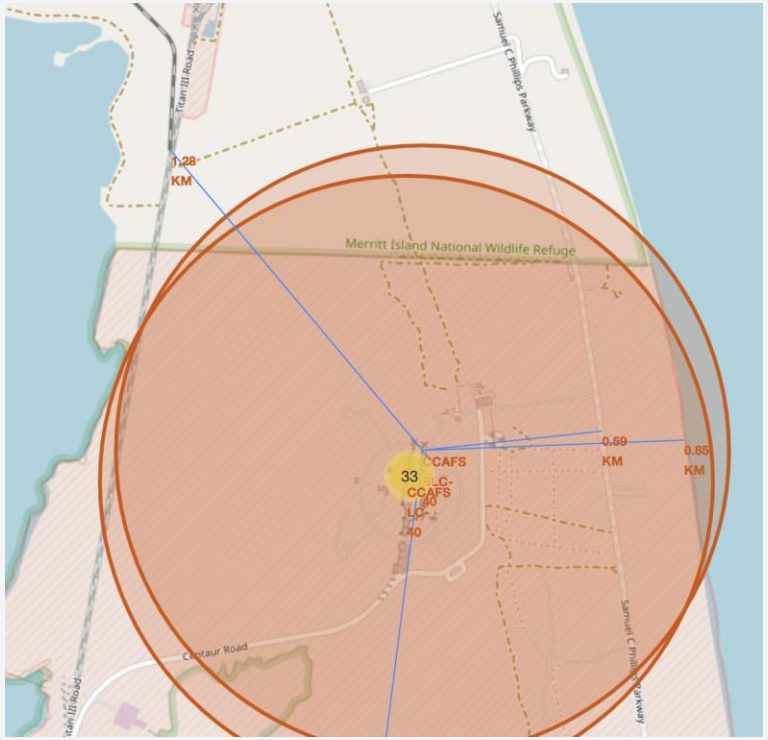
Launch Outcomes at CCAFS SLC-40

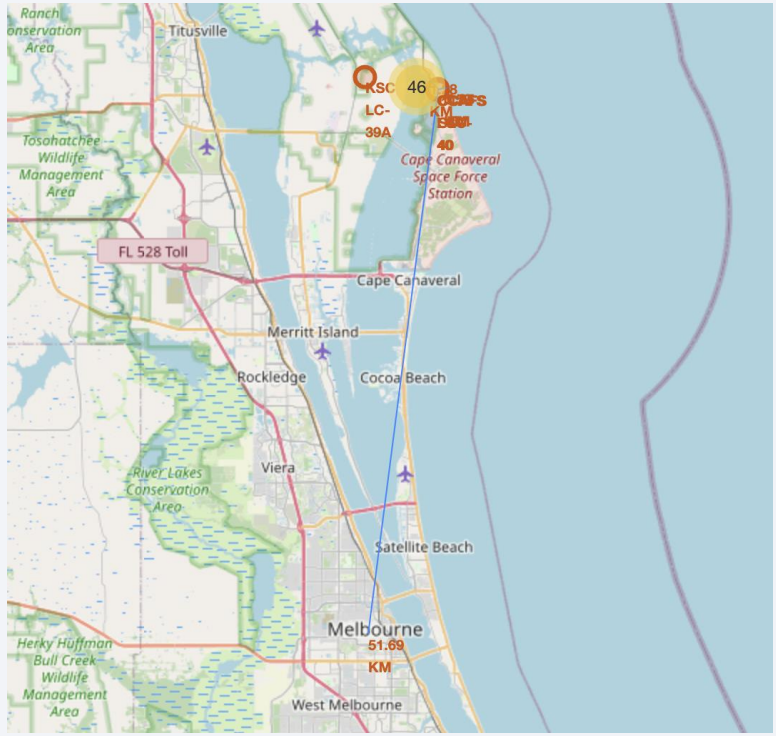
- The success rate for CCAFS SLC-40 is relatively low.

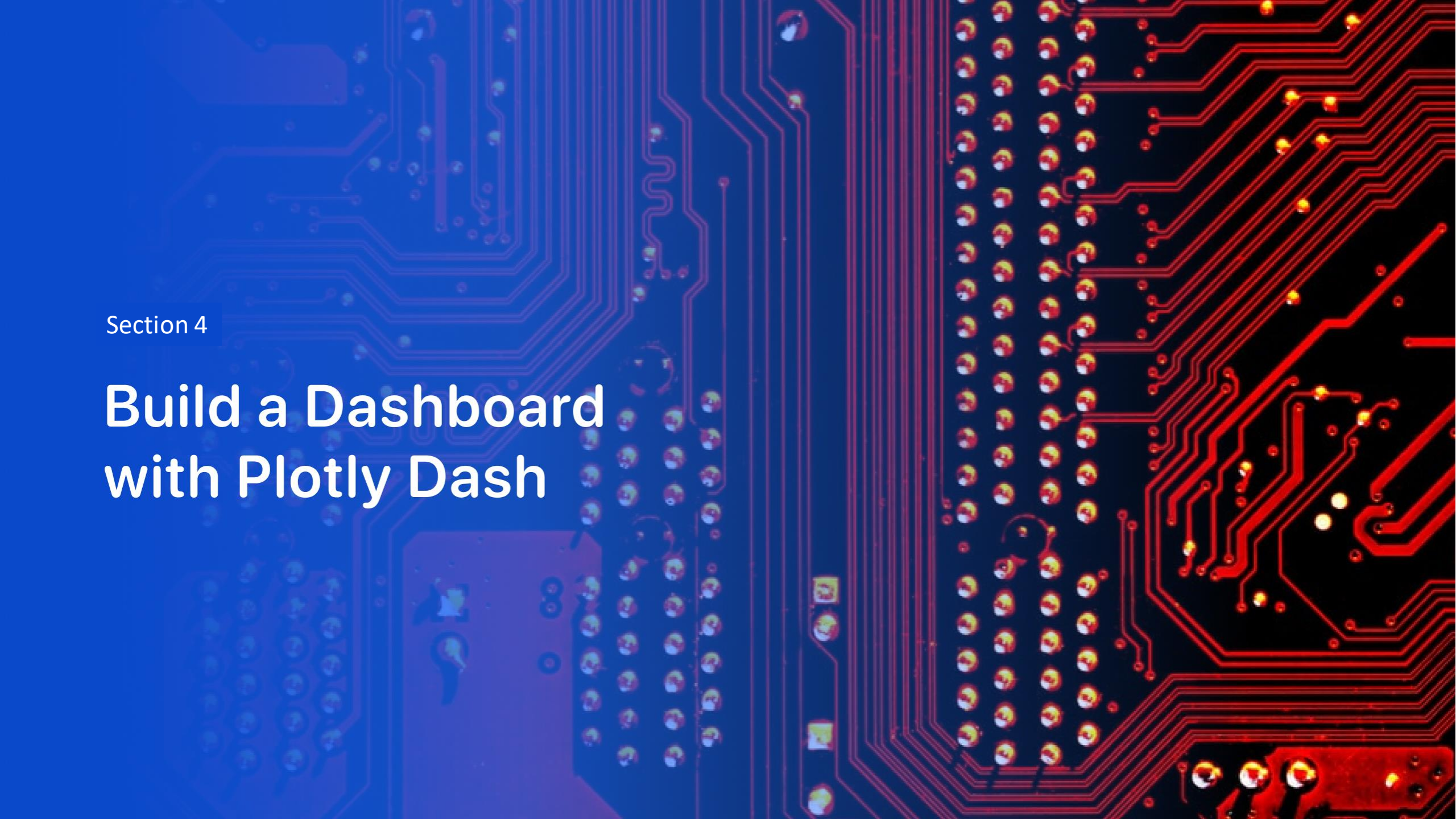


Launch Site Proximities at CCAF5 SLC-40

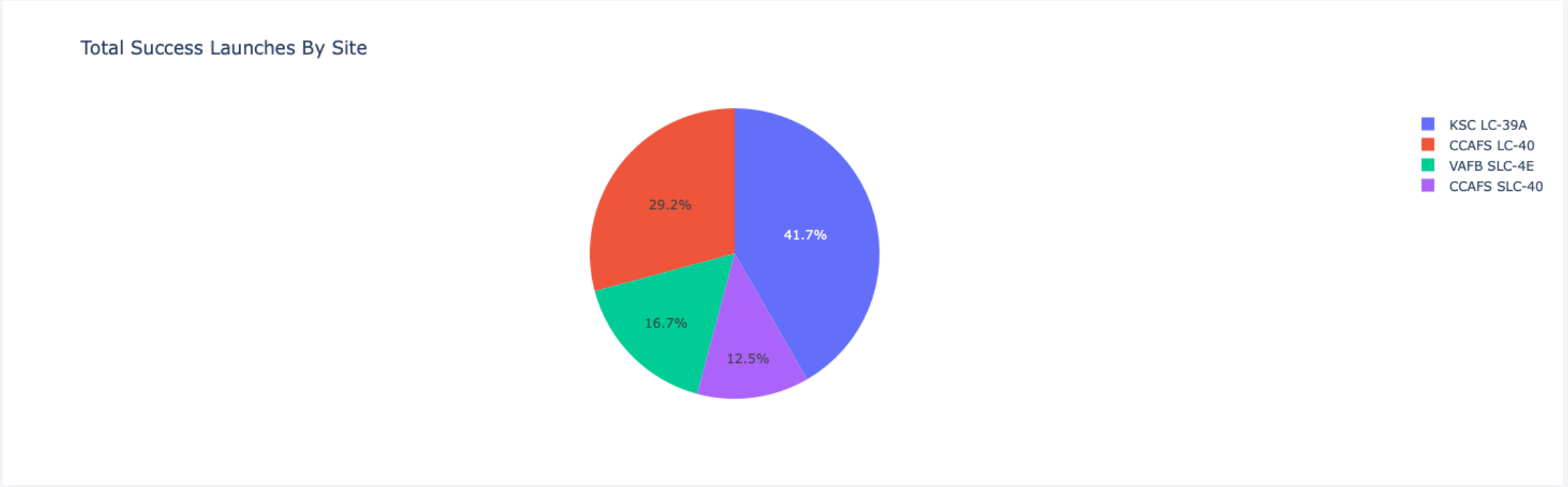
- The launch site is near to the coast (0.85KM)
- The launch site is near to a highway called Samuel C Phillips (0.59KM)
- The launch site is near to a railway called Nasa Railroad (1.28KM)
- The closest city to the launch site is Melbourne at 50KM of distance.





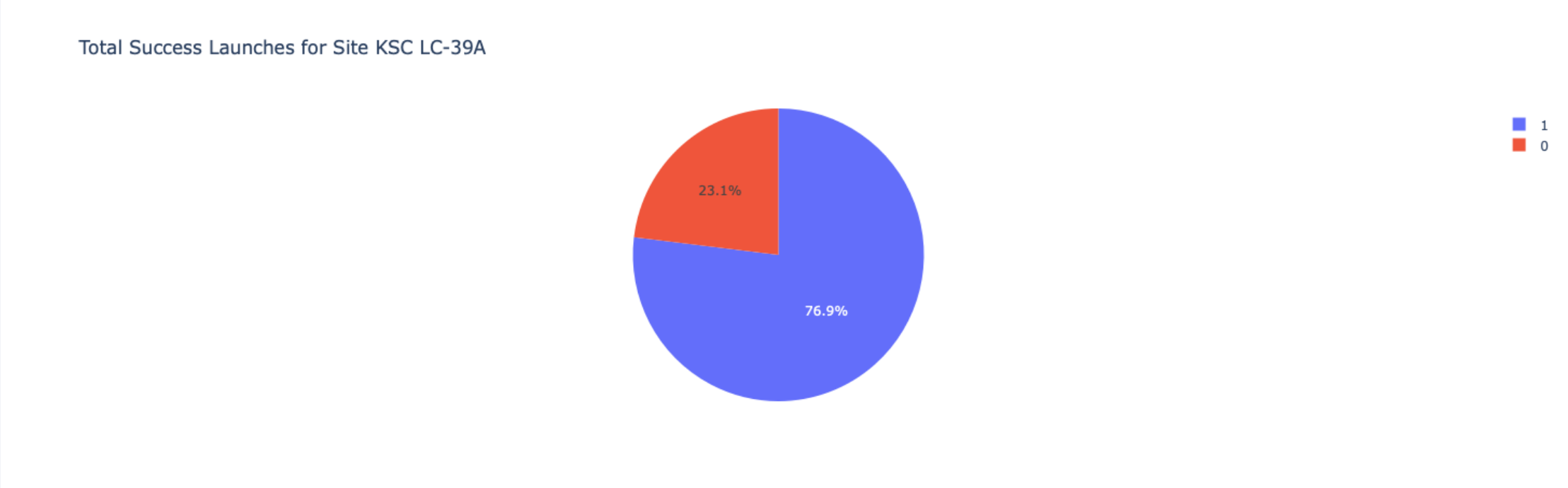

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for all Sites



- The highest number of successful launches (10) are achieved at KSC LC-39A
- The fewest number of successful launches (3) are achieved at CCAFS SLC-40

Launch Outcome for site KSC LC-39A



- KSC LC-39A has the highest launch success ratio.
- It has been successful in 10 occasions and has only failed in 3

Correlation between Payload and Success for all Sites

- The number of launches in the payload range from 0 to 5000 is significatively higher than in the payload range from 5000 to 10000.
- The payload range from 0 to 5000 has the largest success rate.
- The success rate on the payload range from 5000 to 10000 is very low.

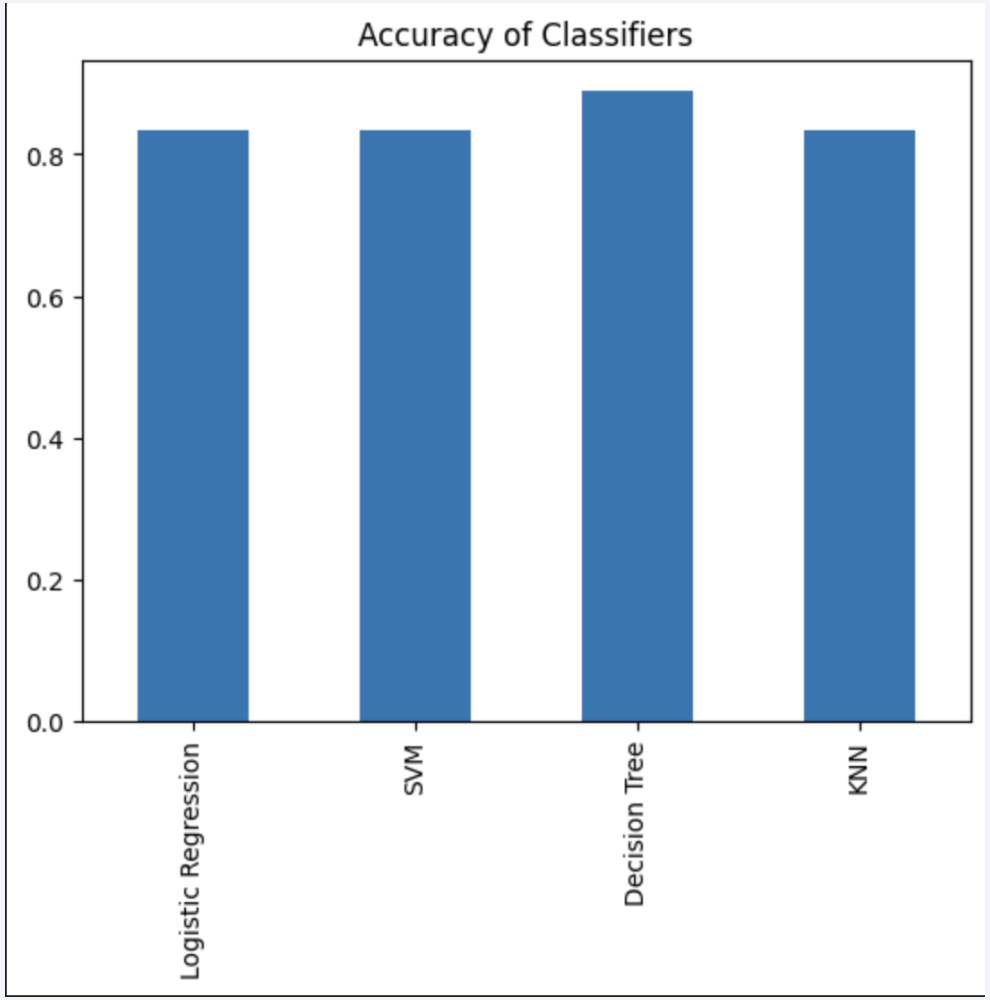
Section 5

Predictive Analysis (Classification)

Classification Accuracy

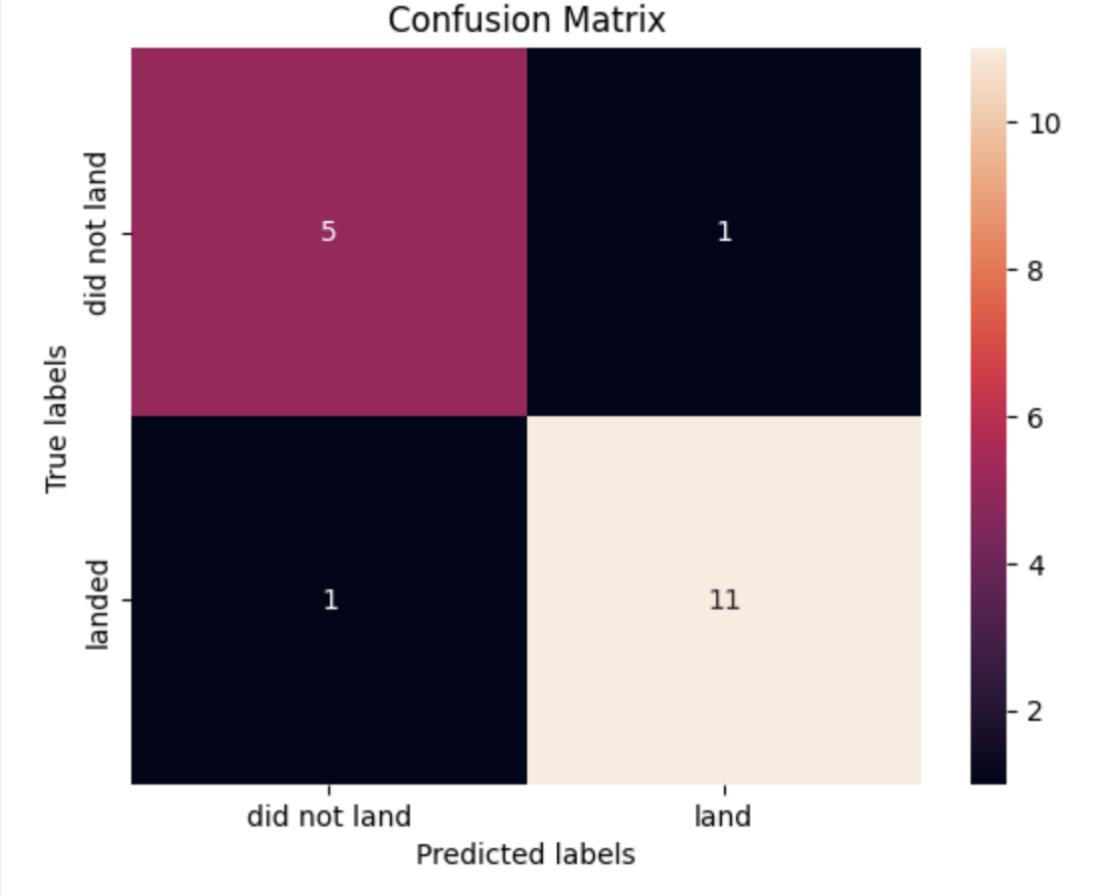
- The Decision Tree model has a slightly better performance and also a better accuracy compared to the Logistic Regression, SVM and KNN models.

	Accuracy
Logistic Regression	0.833333
SVM	0.833333
Decision Tree	0.888889
KNN	0.833333



Confusion Matrix

- The best performing model is the Decision Tree and the Confusion Matrix shows that it only has 1 false positive from 6 samples and 1 true negative from 12 samples.



Conclusions

- Due to the small size of the dataset:
  - All models perform well with an accuracy better than 0.83.
  - The results on predicting new data can be bad because the models may be underfitted
- More data is needed in order to improve the models fitting and therefore, the prediction outcome.

Thank you!