Winning Space Race with Data Science

Vladimir Morozov @curiousdata
5 Apr 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Methodology. This project is created on Python and SQL with the help of Python libraries for scientific computing and machine learning. Data is collected from SpaceX API and web-scraped from Wikipedia – to ensure data consistency and showcase collection skills. Then, data is wrangled and formatted. New column "Class" is a binary alternative for categorical "Landing Outcome" column.

SQL is used to explore data. Visual analytics – charts, maps, graphs and plots - are performed on Seaborn, Matplotlib and Folium. Plotly's dashboard is created to help explore data further. For predictive analysis, a Grid search iterates over four classification models from Scikit-Learn – Logistic Regression, KNN, SVM and Decision Tree.

Results. Variables like Launch site, Payload, Orbit Type, Year and others influence the success probability. Interactive analytics explored the dependencies within the data and influence of geographical factors on success rate. These variables are used in training ML models, out of which KNN model shows the best accuracy of 1.0.

Introduction

SpaceX, a renowned pioneer in space exploration, sets the tone in the industry by using cutting-edge technology to reduce launch costs. The reusable first stage technology, like that on SpaceX Falcon 9 rockets, contributes to company's success and enables two-to-three-fold launch cost reduction – while other companies' similar flights can cost up to 165 million dollars, Falcon 9's launch costs only about 62 million dollars. However, there are accounts where landing attempts of the first stage fail.

This project is dedicated to finding and analyzing the factors that affect the likelihood of successful landing. We will analyse the correlation between different variables and first stage recovery success, and create a machine learning pipeline to predict whether the Falcon 9 first stage will land successfully, based on data of previous flights and patterns and regularities found.

These findings may be helpful in more than one way: SpaceX may use it to increase the rate of successful landings, as well as predict the price and status of a future landing. Besides, both SpaceX and other space companies may draw insights on what contributes to the landing being successful.

Section 1

Methodology

Methodology

Executive Summary

- Data collection: through two different sources to ensure data consistency and showcase skills – pulled from SpaceX API via Requests and web-scraped from Wikipedia via BeautifulSoup;
- Data wrangling: loaded into dataframe, filtered, formatted, cleaned from irrelevant features, decoded (for SpaceX API). Missing values are found and handled. Data types are checked. New column "Class" is created as a binary alternative for categorical "Landing Outcome" column;
- EDA: using visualization and SQL. Several queries are formed to further explore the data. Seaborn and Matplotlib's plots help gain insights about variable importance, laying a foundation for data-driven feature selection in the machine learning model;
- Creating interactive visual analytics using Folium and Plotly Dash;
- Implementing predictive analysis using Grid search over four classification models with Scikit-Learn.

Data Collection

Data for the project is collected from two different sources:

- From SpaceX API via Requests, loaded into dataframe, with function iterating through ID-encrypted columns with API calls, and then into Pandas dataframe;
- Web-scraped from Wikipedia via BeautifulSoup, cleaned, into table-formatted dataframe:

Retrieving the same data from different sources is a validation step to ensure data accuracy and consistency. Data in both cases is prepared for EDA, cleaned and wrangled, checked for missing values, some variables are either optimised or dropped.

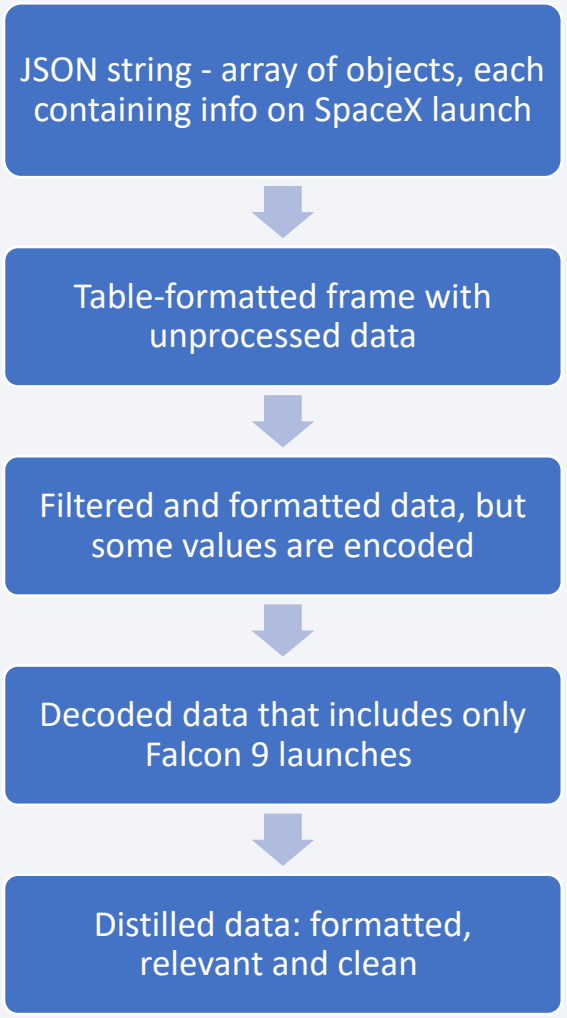
Data Collection – SpaceX API

Records of previous flights are collected from a SpaceX REST API using 'Requests' library. They need to be processed and transformed to become useful:

- 'Requests' library pulls data from SpaceX API and stores it in a variable as a JSON-formatted string. This is a raw data containing hundreds of lines of unnecessary data, and needs to be transformed and cleaned.
- The string is loaded, normalized and parsed into Pandas dataframe in a tabular format.
- Data is filtered to exclude irrelevant launches with 2+ boosters. For some features, a value is extracted and used to replace the text. "Date" column is formatted, and set to only include data within a certain time period.
- Other features contain feature ID instead of a value. A function iterates through each of these and retrieves value for every ID from API. Decoded values are appended to a corresponding feature list. With these lists, a new decoded dataframe is created.
- Decoding the 'Booster version' allowed to filter out Falcon 1 launches. Missing values are found and handled: for continuous values a mean substitutes 'Null'.

Now, after some data wrangling and formating, the final dataframe includes only clean, formatted and relevant data that can be used in next steps.

See the source code: Data Collection API



Data Collection - Scraping

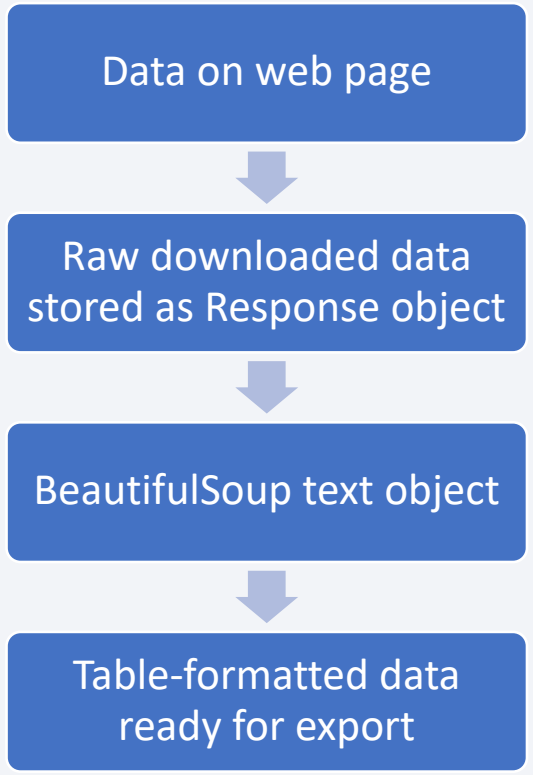
Falcon 9 historical launch records are also collected from open-source Wikipedia page, where they are stored in a HTML table.

- The page URL is requested with 'Response' command, which returns Response object – raw page data.
- BeautifulSoup object is created from a response text content. A FOR loop finds and parses the table within it, extracting all column/variable names and values and storing them.
- Table-formatted dataframe is created and filled with values from the text object.

After the table is ready, it is exported as CSV file for next steps. Collecting the data from different sources serves the purpose of validating and checking the data.

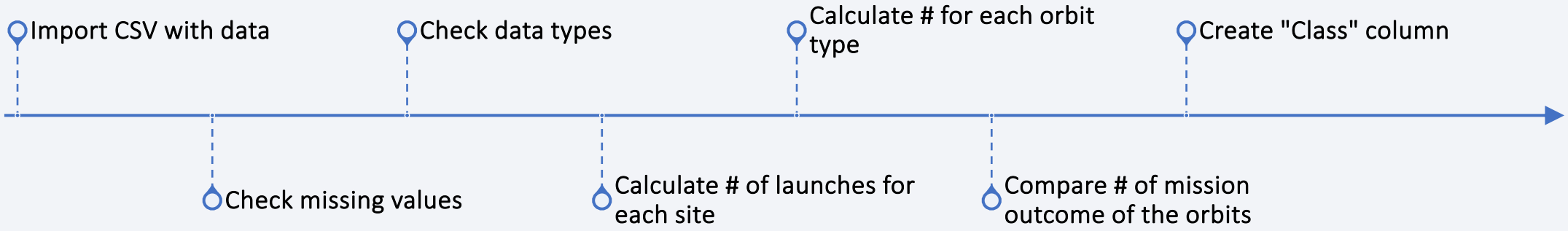
See the source code: Data Collection with Web Scraping

See the data source here.



Data Wrangling

First off, some data wrangling and exploratory analysis is done, it includes checking the data and calculating some statistics to better understand and explore the data:



Data is checked for missing values and correct data types. Number of launches from each site, occurrence of each target orbit is calculated. For each orbit, number and occurrence of outcomes is calculated.

"Landing Outcome" column contains multiple categorical variables specifying the landing type, but there is no need for this information. Instead, a new binary column “Class” is added, dividing all launch outcomes into '1' for all successful outcomes and '0' for failures. It is the target variable, and will be used to predict first step launch success in a machine learning model.

See the source code: Data Wrangling

EDA with SQL

To form a deeper understanding of the data, a number of SQL queries is formed. The main findings are visualized, and include:

- Discover all launch sites of Falcon 9s
- Calculate the total payload mass of NASA boosters
- Analyze average payload mass for a certain booster
- Uncover the date of first successful landing
- Create list of successful boosters within a set payload interval;
- Compare total number of successes and failures
- See boosters used for maximum payload mass;
- Assess failure history for select year and location;
- Visualize distribution of landing outcomes within timeframe.

The queries helped see the connections between variables and to progress to visualizing data and building a model. The highlights from the EDA are included in this presentation.

The source code is also available: EDA with SQL

EDA with Data Visualization

Further Exploratory Data Analysis is performed using Pandas and Matplotlib . Select relationships between the variables are visualized. Emphasis is made on discovering interdependencies of the shown variables:

Flight number	Launch site	Orbit type	Payload

Then, launch success yearly count is created, highlighting an upward trend in successful landing probability. Tools used for visualization include line plots, scatter plots, bar charts, scatter point charts.

Features are selected and engineered, one-hot encoding categorical variables and setting numerical data as type "float64" for compatibility and enhanced model performance.

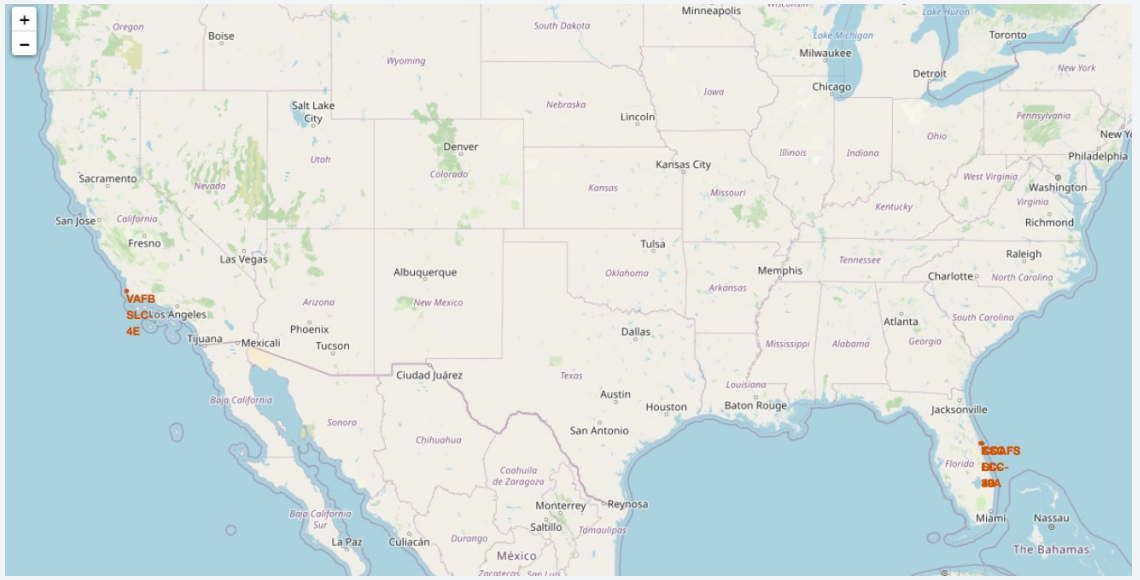
See the source code in .IPYNB file: EDA with Visualization

Build an Interactive Map with Folium

Each Falcon 9 launch site for Falcon 9 rockets is marked on the Folium world map, to visualize the use intensity and success rate for each site. For each, relevant launches are added, with colors representing successful and failed recovering of the first stage.

Hypothetically, geographical location and nearby structures (highways, railways, cities, shorelines) may influence the launch site's success rate. To explore the potential significance of these factors, a distance from each site to a closest one of these objects is calculated.

See the source code in .IPYNB file: Interactive Visual Analytics with Folium



Dashboard with Plotly Dash

The created dashboard allows stakeholders to analyze launch data and to find insights visually. It has two main elements:

- a **pie chart**, that either shows a share of each launch site in successful launches, or a success/failure percentage for a selected launch site,
- a **scatter plot**, to visualize successful launches and of different boosters for a selected site and payload.

The dashboard also has two input components to select the data to be visualized:

- **a launch site dropdown,**
- a **payload range slider.**

Both inputs and both graphs are connected to the corresponding Callback function, which updates the visual component as soon as the input changes. And here are some of the insights that can be identified visually through interactive Dashboard (source here):

Launch site:
Which has the largest history of successful launches, and which - the highest launch success rate?

Payload range:
Which has the highest, and which – the lowest launch success rate?

Falcon 9 booster version:
Which booster (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

Predictive Analysis (Classification)

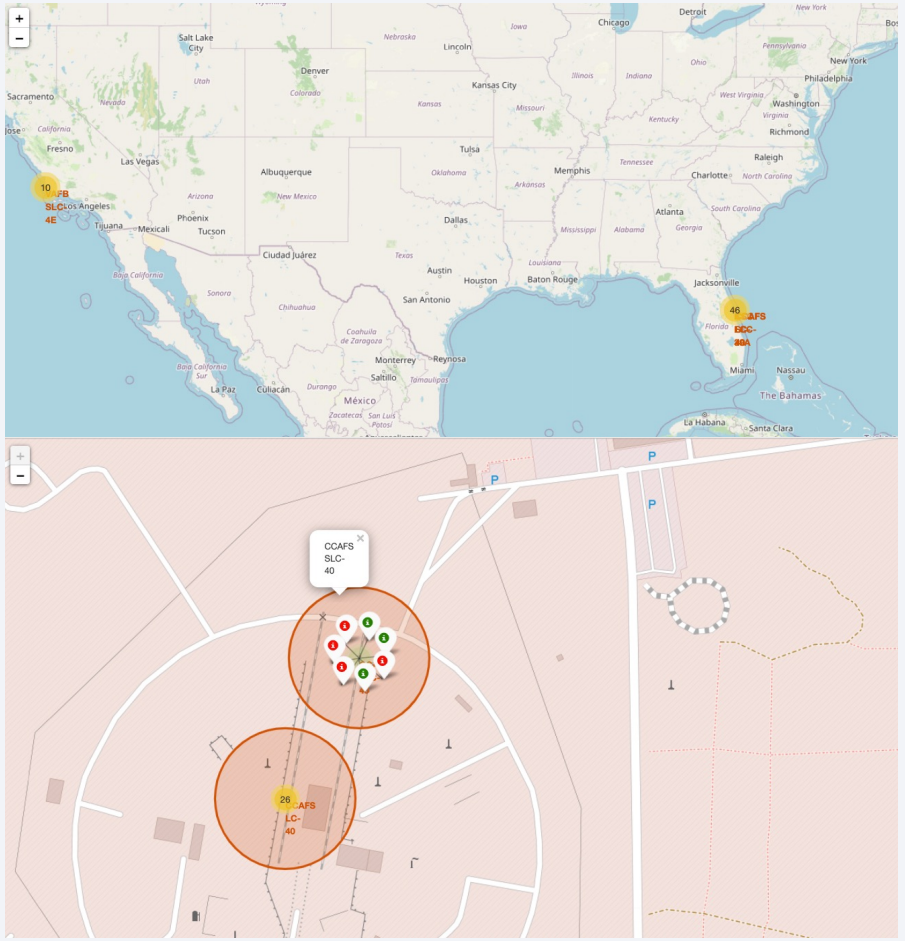
Data preparation process:

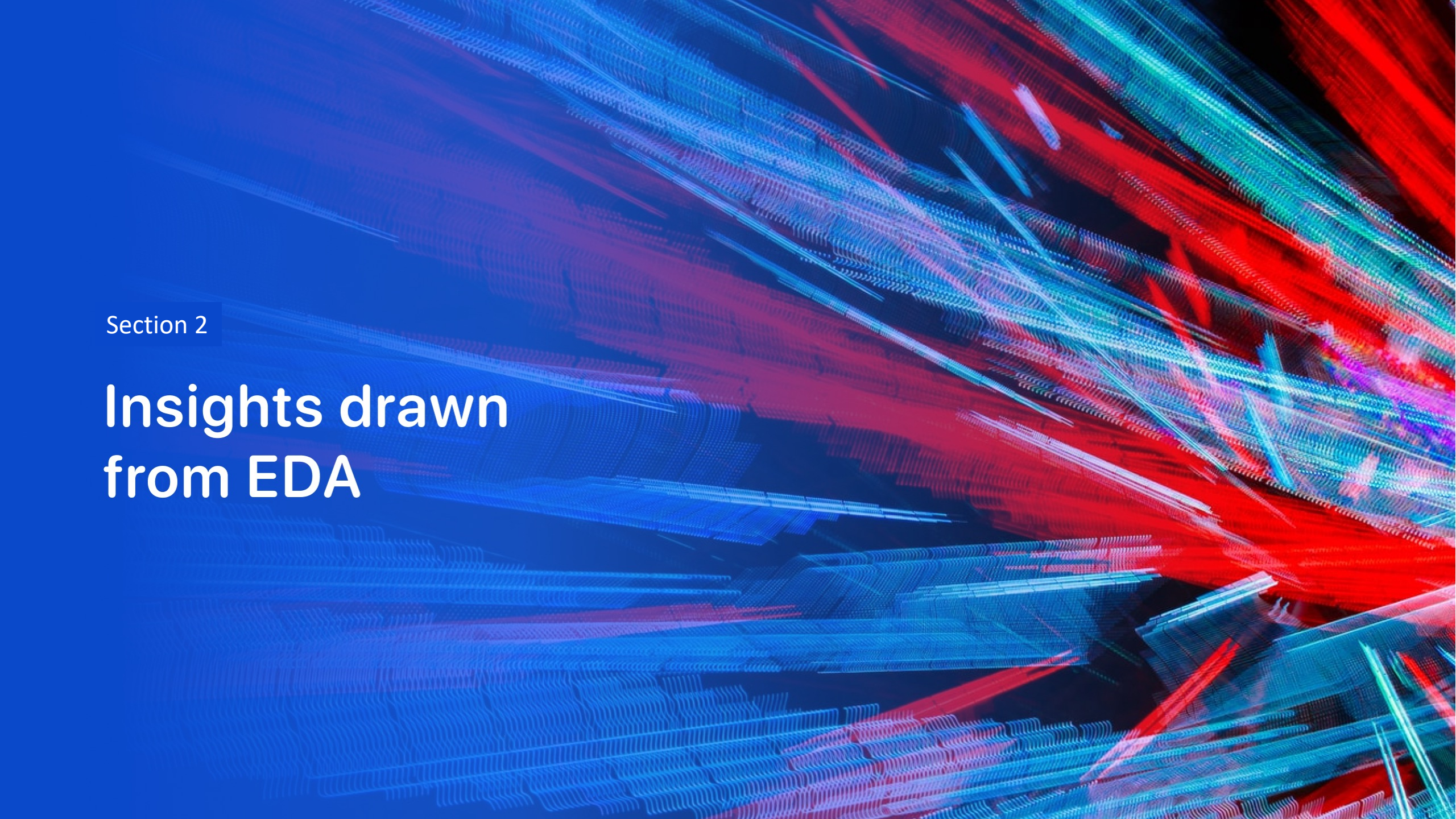
1. Dividing data into target variable ("Class") and independent predictor variables.
2. Storing predictor variables as an array and standardizing with StandardScaler to improve the model performance.
3. Splitting data into "Train", that would be used to create and train a ML model (80%), and "Test", that would be used to assess model performance (20%).
4. Creating four classes of machine learning models, in each case GridSearchCV object iterates through hyperparameters and finds the best-performing model, and for one of each class, confusion matrices and accuracy scores are calculated. One model that performs best is chosen.
    - Logistic regression,
    - Support vector machine,
    - Decision tree classifier,
    - K-nearest neighbors classifier.

See the source code in .IPYNB file: Complete the ML Prediction

Results

- EDA revealed the influence of Launch site, Payload, Orbit Type, Year and others on the success probability.
- Interactive analytics explored the dependencies within the data and influence of geographical factors on success rate – site's location does not influence the Falcon 9's success rate on it.
- Predictive analysis resulted in an accurate machine learning model. All important features found during EDA were used in training a high-accuracy regression model capable of predicting the successful recovery of Falcon 9's first stage, with absolute accuracy.



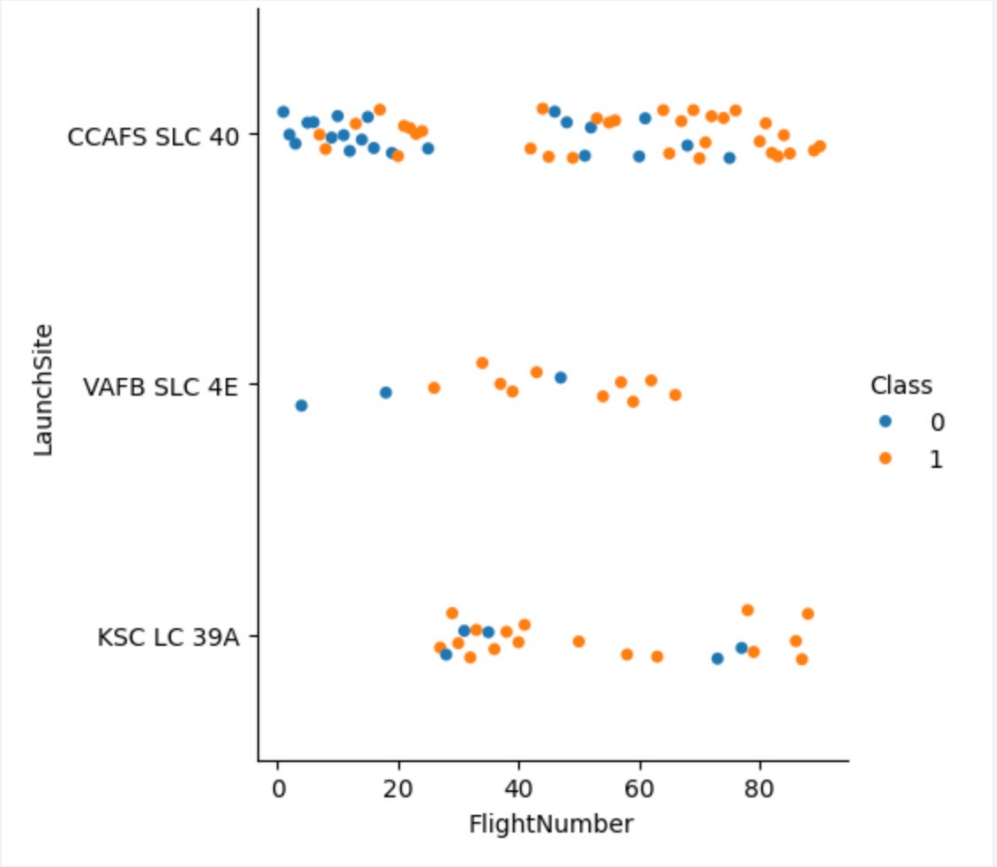

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

A scatter plot of Flight Number vs. Launch Site shows the relationship between the variables with additional differentiation on whether in was successful or otherwise.

A general tendency of launches becoming more reliable is evident on two sites out of 3.

Additionally, the plot provides info on the use of said sites – approximate time when it first was used, and when it fell out of use for Falcon 9 (VAFB SLC 4E).

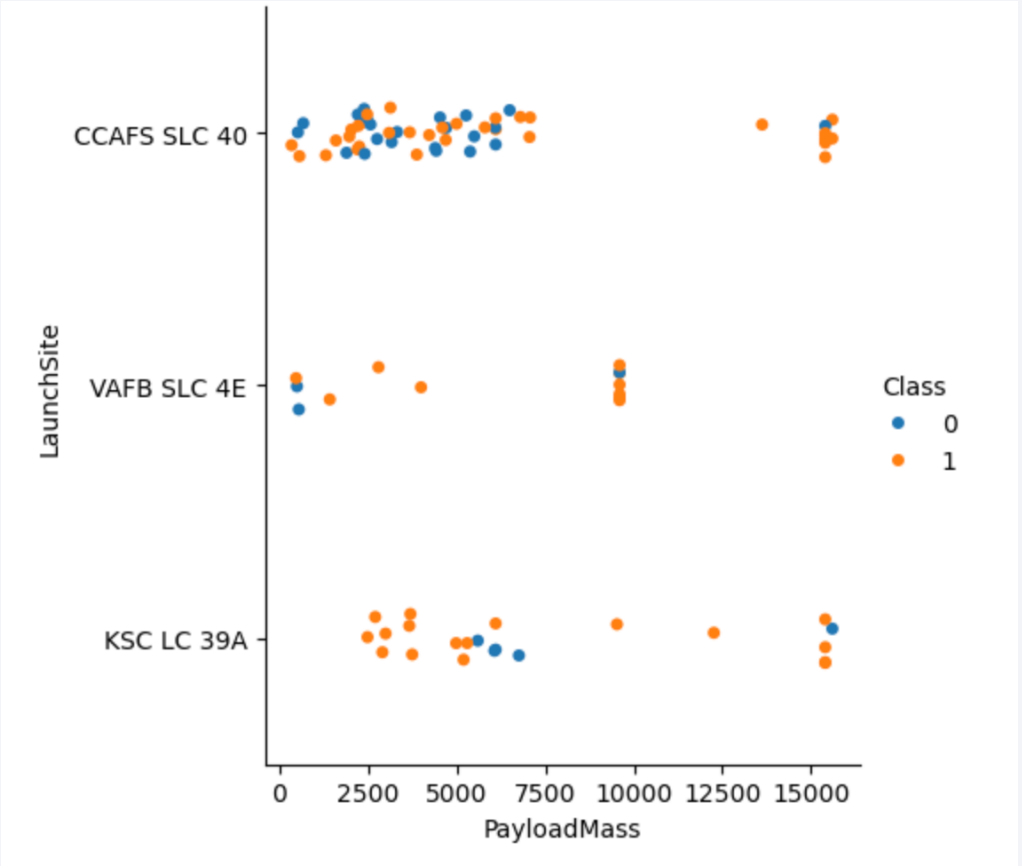


Payload vs. Launch Site

A scatter plot of Payload vs. Launch Site shows the distribution of launches with different mass, successful and not, between three sites.

The success rate is different for different payloads – on CCAFD, a very high success rate for 15000 kg payload, while the rate for 0-7500 payload is significantly lower.

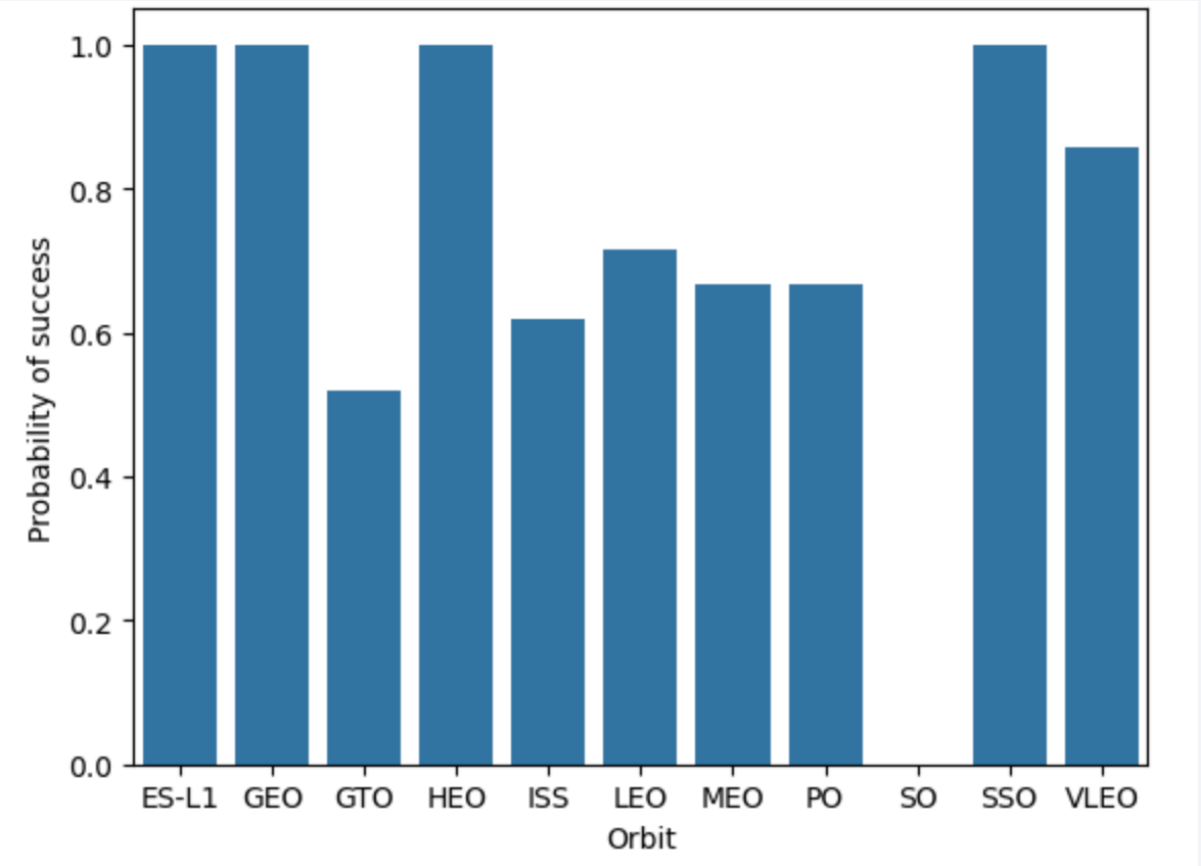
It can be seen that there possibly are limitations of rocket's payloads – the maximum cap of 10000 kg for VAFB, or the minimum cap of 2500 kg for KSC, for instance.



Success Rate vs. Orbit Type

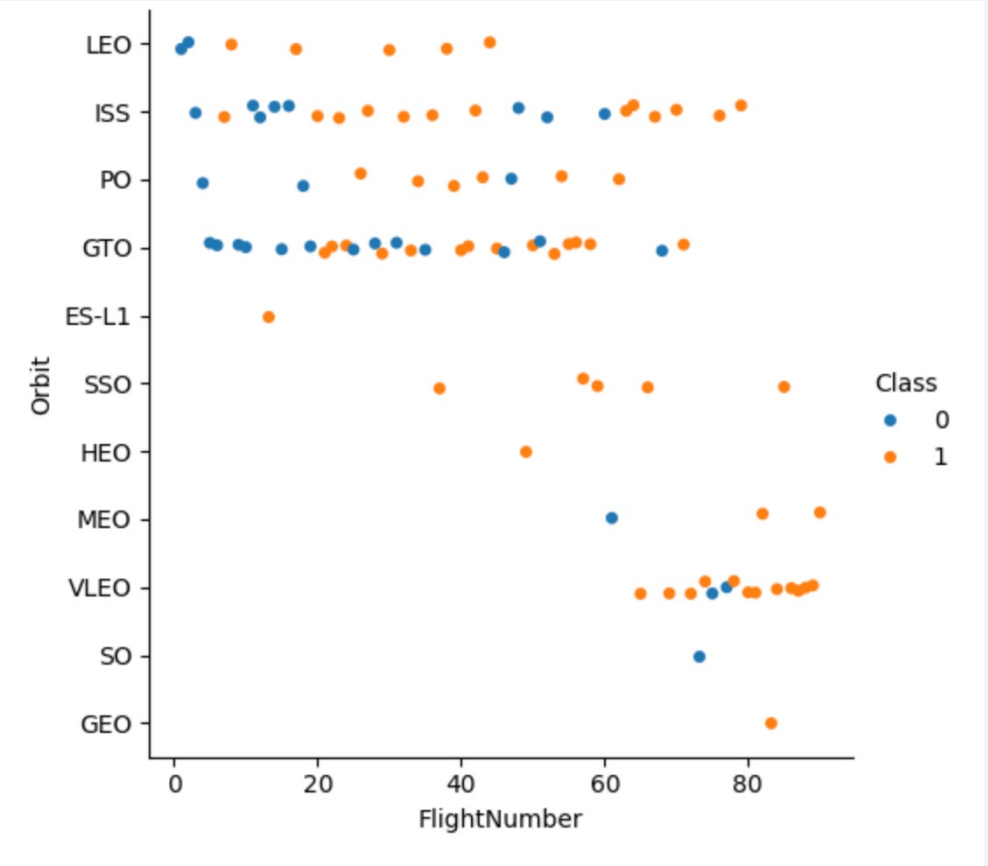
A bar chart for the success rate of each orbit type compares the probability of recovering the first stage, depending on the target orbit of the mission.

For some orbits – ES-L1, GEO, HEO, SSO – every recovery was successful. For others, the probability is lower. Clearly, target orbit affects the chances for the first stage recovery.



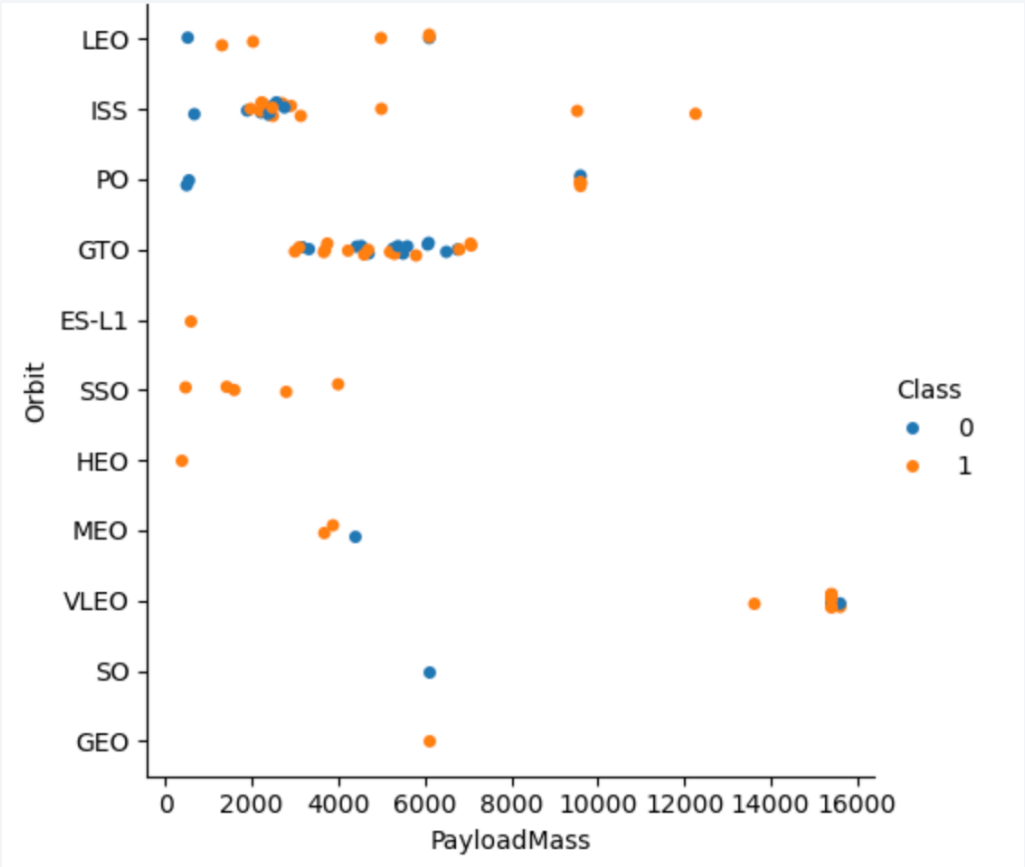
Flight Number vs. Orbit Type

- Scatter point of Flight number vs. Orbit type shows that orbits like LEO or ISS were most popular in the beginning of Falcon 9 flights history, and the attention gradually shifted to SSO, VLEO and others.
- The latter orbits have higher success rate while the first four show the tendency of launches becoming more successful with time.



Payload vs. Orbit Type

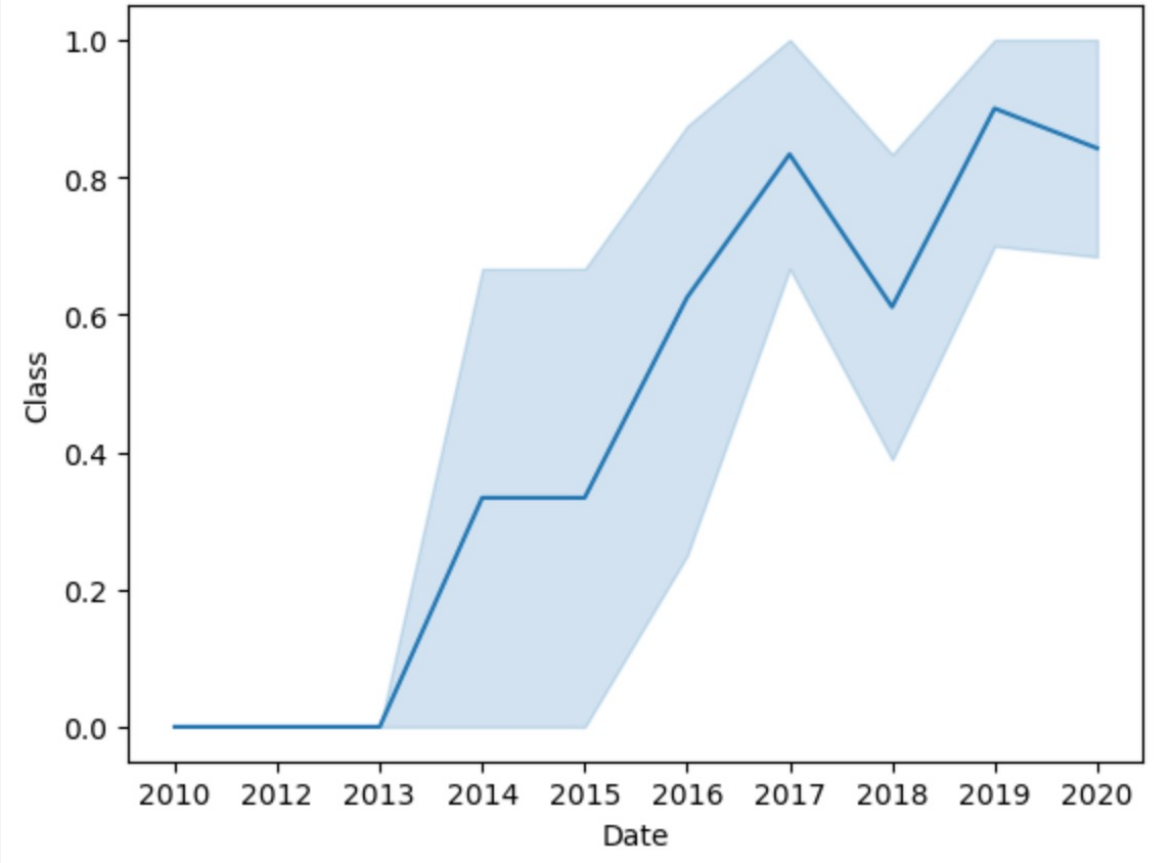
- A scatter point of payload vs. orbit type shows that for every orbit, only a range of payload is used.
- We can also explore the success rate of every orbit and payload



Launch Success Yearly Trend

A line chart of yearly average success rate shows a share of Falcon 9 launches that ended in successful recovery of the first stage.

An upward trend of the line indicates the increasing probability of recovery, whether is it due to more reliable booster version (such as “FT”) or better launch site (“KSC LC-39A”)



All Launch Site Names

Unique launch sites' names are found via an SQL query. The result is a distinct list of sites that were used for Falcon 9 launches:

- Cape Canaveral Air Force LC-40;
- Vandenberg Air Force Base SLC-4E;
- Kennedy Space Center LC-39A;
- Cape Canaveral Air Force SLC-40.

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with “CCA” were found with SQL query to the database, limited to first five results that meet the criteria:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

The total payload carried by boosters from NASA, as a foundation for more data exploration and company-specific projects based on this project. Can easily be altered to find any other space agency. Queried with SQL:

sum(PAYLOAD_MASS__KG_)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1. Can be modified to find averages of any other booster version and to compare different versions, which is out of scope for this project:

Average	Mission_Outcome
2534.6666666666665	Success

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad. Can be used to explore technical or other factors that led to that particular landing being successful.

The code can be modified to find the date of last successful landing, first failure or other date-related queries, within minutes.

Date	Mission_Outcome
2010-06-04	Success

Successful Drone Ship Landing with Payload between 4000 and 6000

List of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, as an example of a specific query.

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535
F9 v1.1 B1011	4428
F9 v1.1 B1014	4159
F9 v1.1 B1016	4707
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1030	5600
F9 FT B1021.2	5300
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 FT B1031.2	5200
F9 FT B1032.2	4230
F9 B4 B1040.2	5384
F9 B5 B1046.2	5800
F9 B5 B1047.2	5300
F9 B5 B1048.3	4850
F9 B5 B1051.2	4200
F9 B5B1060.1	4311
F9 B5 B1058.2	5500
F9 B5B1062.1	4311

Total Number of Successful and Failure Mission Outcomes

This query shows the total number of successful and failure mission outcomes.

Mission_Outcome	Customer	count(Mission_Outcome)
Failure (in flight)	NASA (CRS)	1
Success	SpaceX	98
Success	USAF	1
Success (payload status unclear)	Northrop Grumman	1

Boosters Carried Maximum Payload

List of the names of the booster which have carried the maximum payload mass.

Shows the boosters capable of handling the maximum payload mass of 15600 kg.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Failure Records

The failed in drone ship landing outcomes listed: month of the event, their booster versions, and launch site names for year 2015. As can be seen, for this specific query there are only two, in January and April of 2015.

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

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

This SQL query calculates the distinct outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 and orders them in descending order, showing the probability distribution of different outcomes.

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


Section 3
Launch Sites
Proximities Analysis

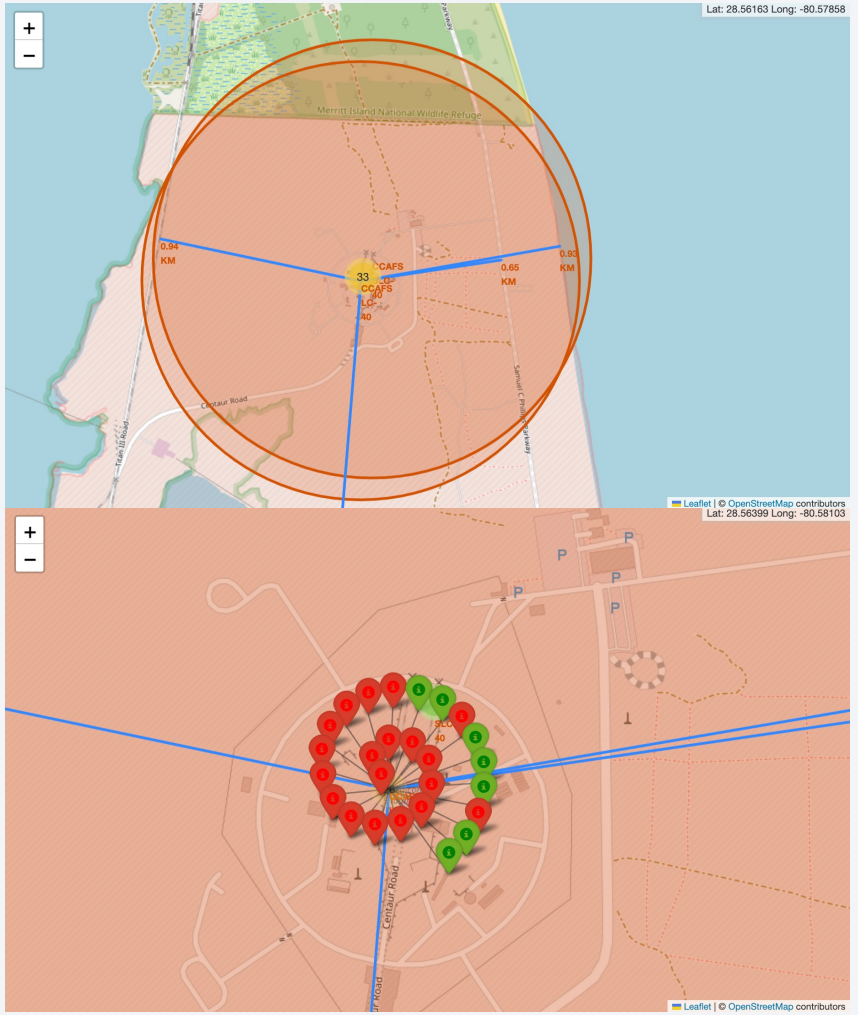
Folium interactive map: CCAFS LC-40

CCAFS LC-40 (and SLC-40) are located in Florida, on coordinates (28.5,-80.9). They are located within meters of each other, and for simplicity, distance from only one of them is calculated – LC-40 with much more flights and success rate of 26.9%.

Nearest coastline, highway and railway are located 0.9, 0.6 and 0.9 km away, respectively.

Nearest city is 17.2 km away.

Both of them are 3177 km from the Equator.



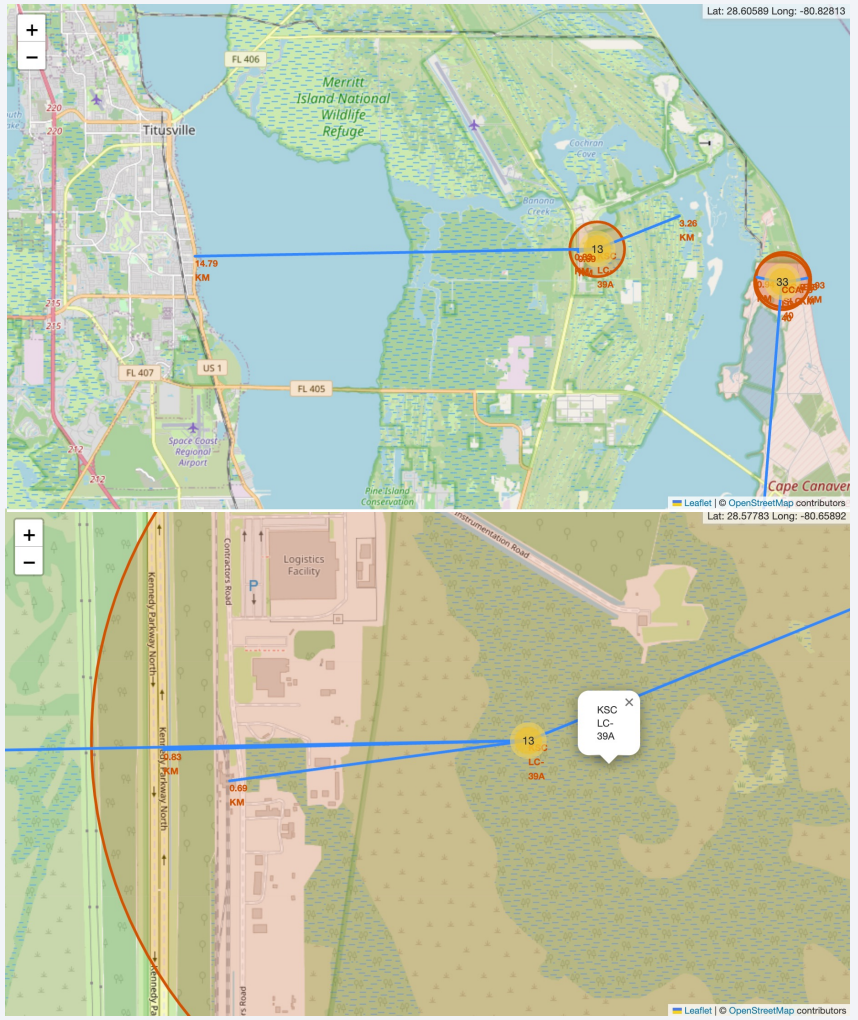
Folium interactive map: KSC LC-39A

KSC LC-39A is situated in Florida, a couple of kilometers away from CCAFS. This site’s geographical location may be the reason for its unusual success rate of 76.9%.

Nearest coastline, highway and railway are located 3.2, 0.8 and 0.7 km away, respectively.

Nearest city is 14.7 km away.

Its distance from the Equator is 3178 km.



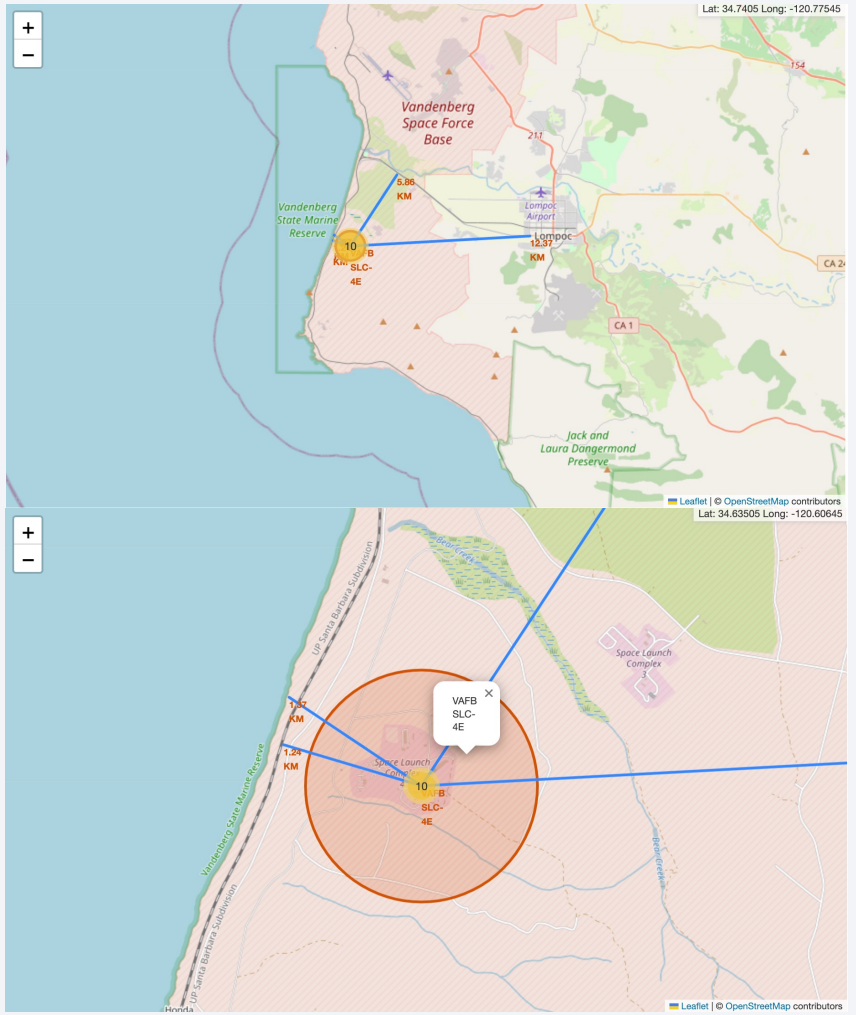
Folium interactive map: VAFB SLC-4E

VAFB SLC-4E is situated in other part of continental US, California, not far from Los Angeles and close to the shoreline. Analysis shows that all distances to potentially meaningful objects are more or less the same for it, too.

Nearest coastline, highway and railway are located 1.4, 5.9 and 1.2 km away, respectively.

Nearest city is 12.4 km away.

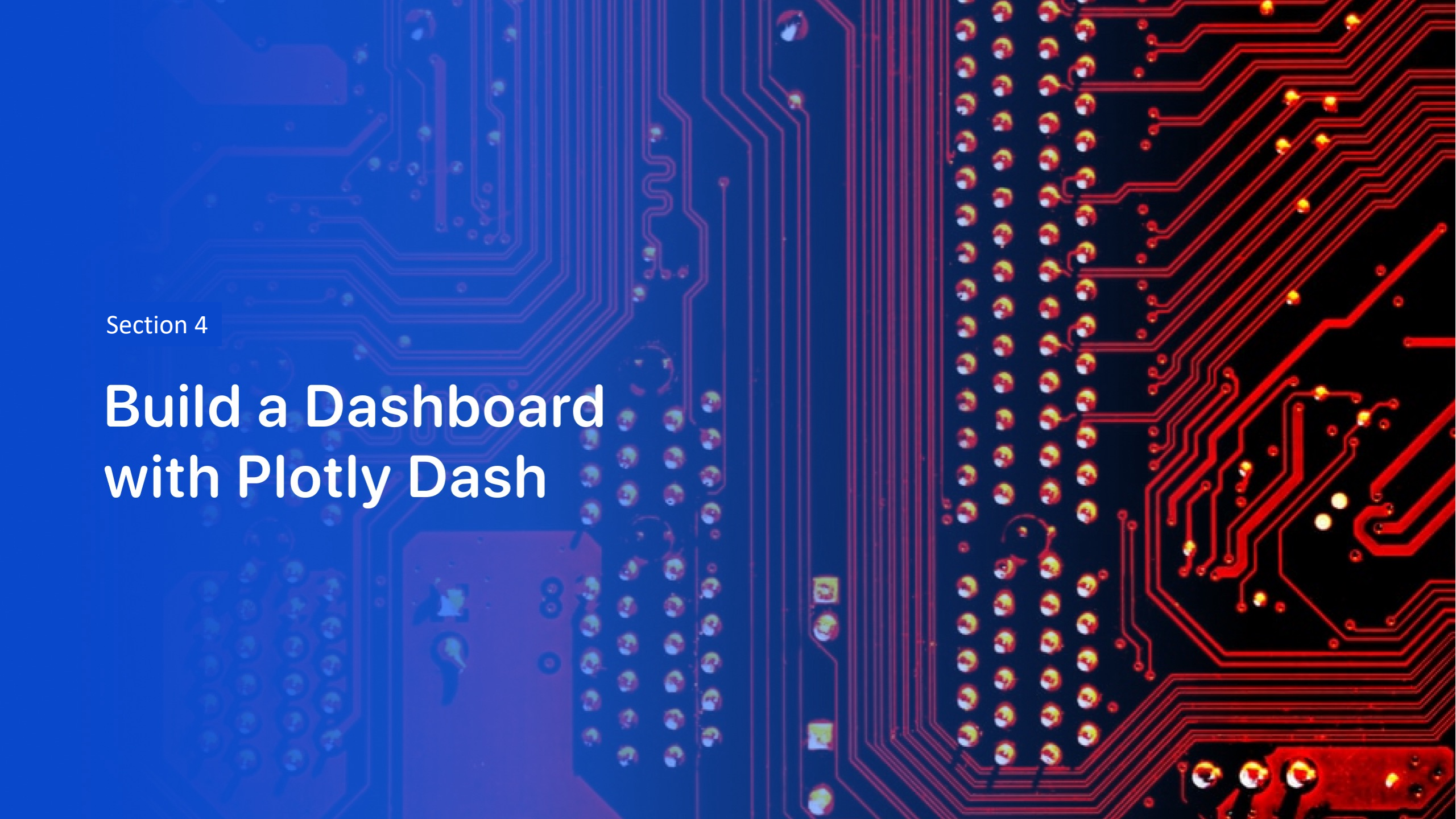
It is 3852 km from the Equator.



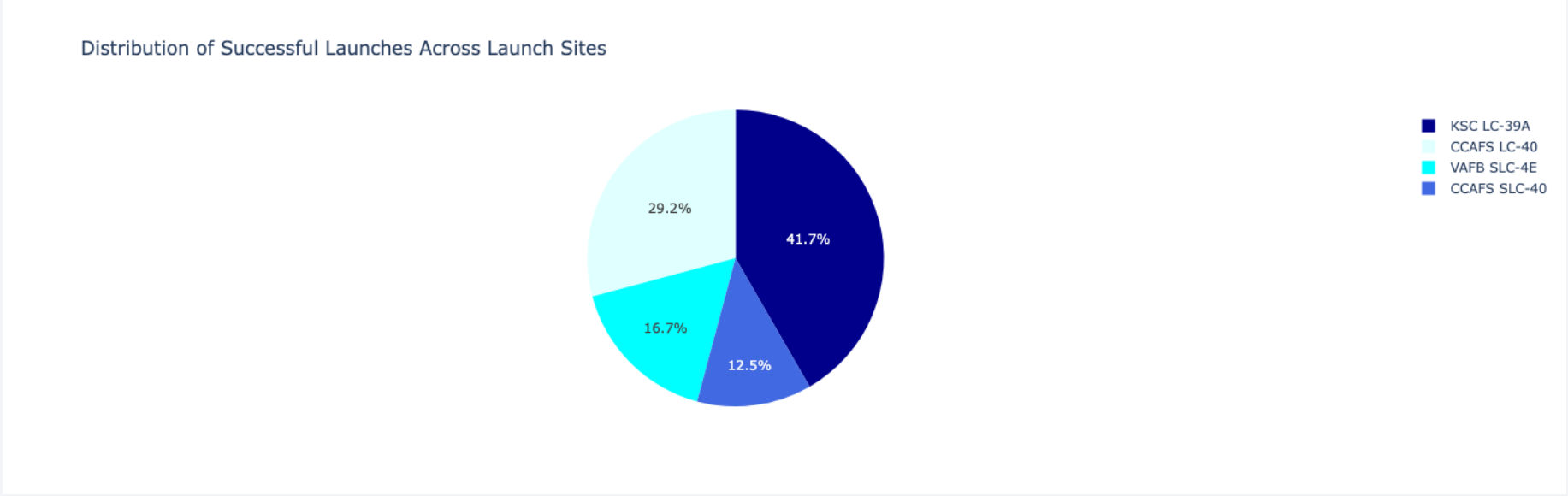
Location vs. Success rate analysis

Three launch sites were used for analysis of distances - CCAFS LC-40 and CCAFS SLC-40 are located within meters of each other, and it seemed adequate to use only one coordinates for the comparison.
Moreover, what was of the most interest is a high success rate of KSC launch site, and investigating whether or not geographical location was a contributing factor in that.
However, **no such dependency was discovered**. It might take more time – out of scope for this project - to prove it, but initial analysis showed that geographical location does not influence the Falcon 9's recovery probability:

	CCAFS LC-40	KSC LC-39A	VAFB SLC-4E
Distance to Coastline	0.9	3.2	1.4
Distance to City	17.2	14.7	12.4
Distance to Railway	0.9	0.7	1.2
Distance to Highway	0.6	0.8	5.9
Distance to Equator	3177	3178	3852
Success rate	26.9%	76.9%	40%

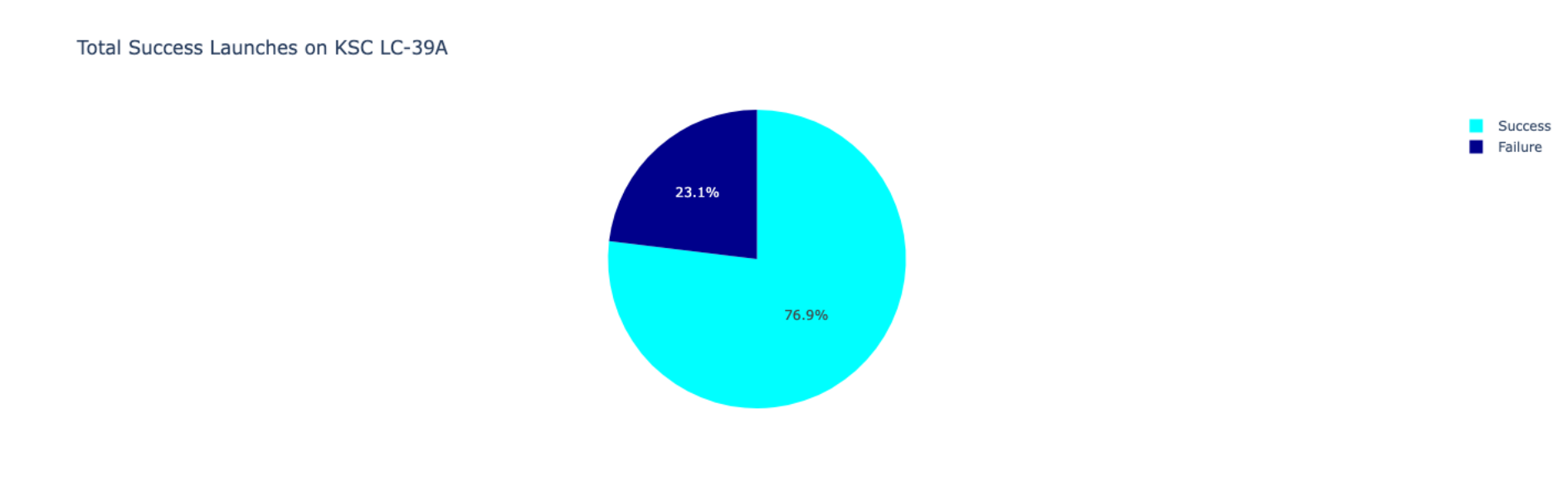

Section 4
Build a Dashboard with Plotly Dash

Dashboard: Successful launch distribution pie chart



The pie chart visualizes how successful launches are distributed across all four sites used for Falcon 9s. There is a hierarchy in the distribution, with a substantial part of all successful launches being held on Kennedy Space Centre KSC LC-39A launch site.

Dashboard: Launch success share



The pie chart is rendered for each launch site separately, and with it we discover that KSC LC-39A site has the highest relative success rate of 76.9%.

Combining it with what we have learned on success launch distribution across launches, we can see the importance of this particular site in Falcon 9 launch history.

Dashboard: Payload vs. Launch Outcome

This Payload vs. Launch Outcome scatter plot for all sites allows to visually determine:

- Most successful booster version category for a set payload range: “FT”;
- Payload range of the “FT” booster: 2000-5500 kg;
- Least successful booster version for the 2000-5500 kg range: “v1.1”.

Section 5

Predictive Analysis (Classification)

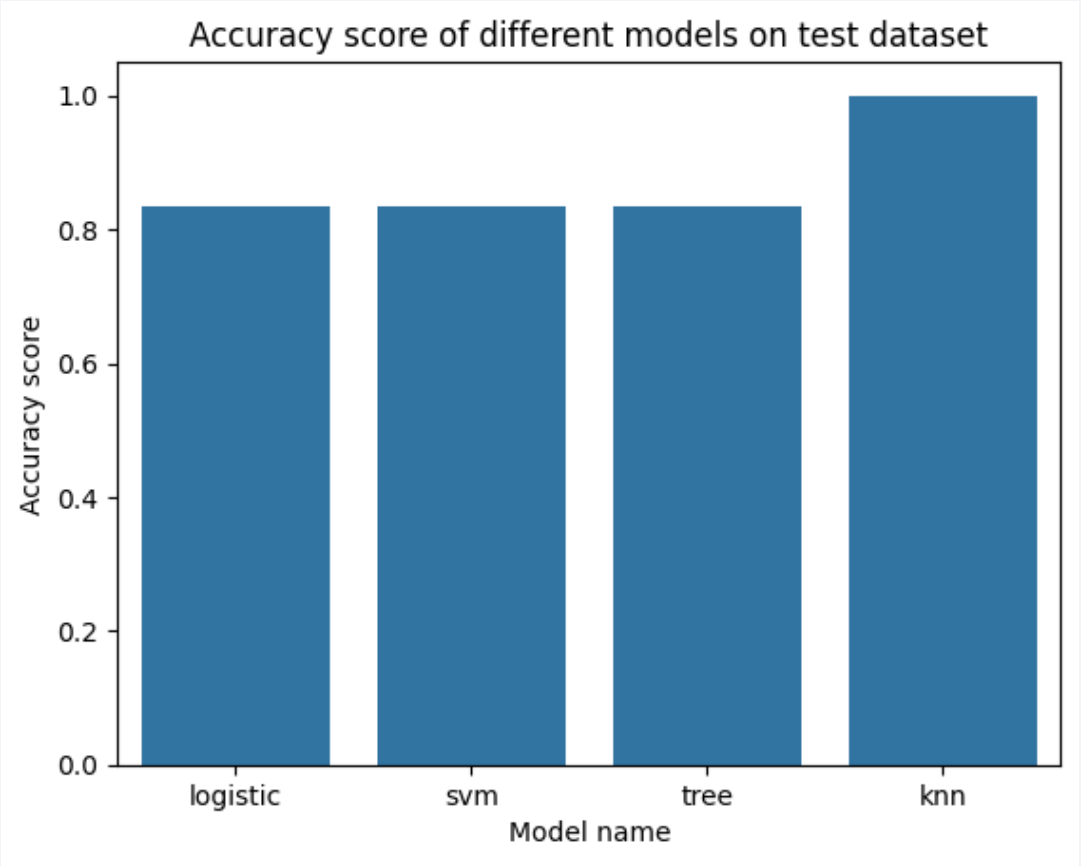
Classification Accuracy

Logistic regression, SVM and Decision Tree models showed good accuracy of 0.83 on the test dataset.

A K-Nearest neighbors model proved to have the best classification accuracy, with the score of 1.0.

The KNN version that showed the best accuracy has:

- ***n*** (number of neighbours) = 1, and
- ***p*** (power parameter for the Minkowski metric) = 1, Manhattan distance.

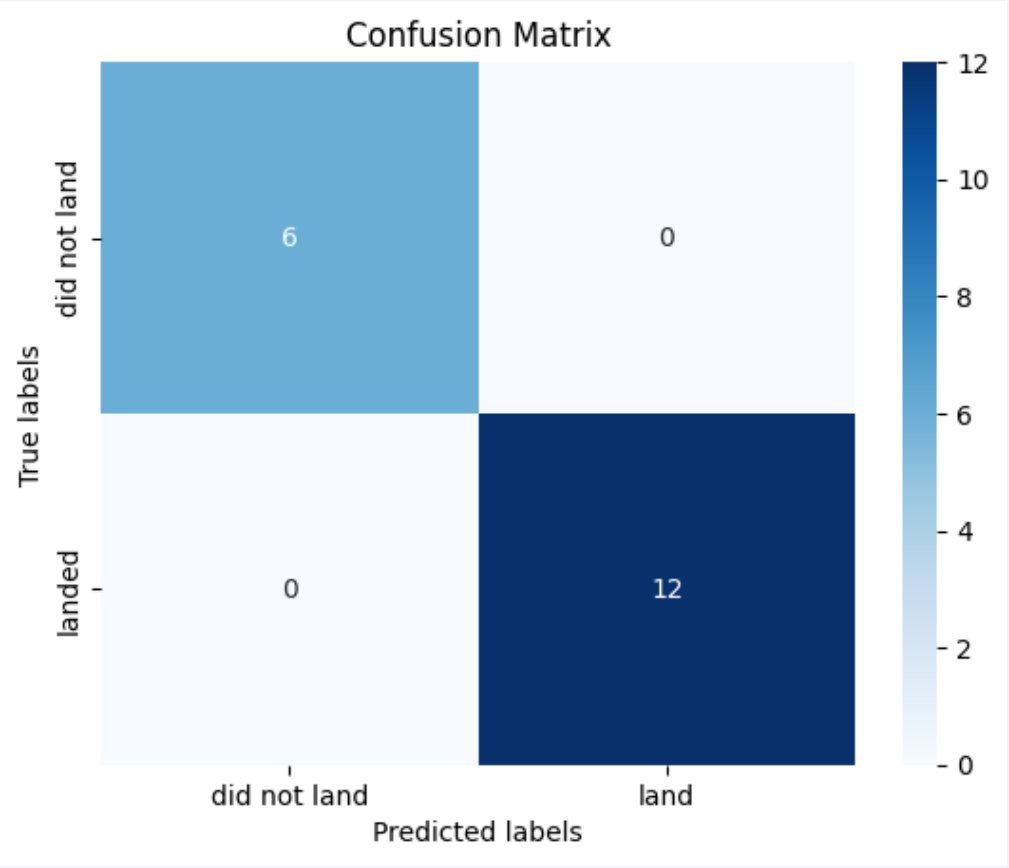


Confusion Matrix

For KNN model, the matrix represents 100% accuracy on test dataset:

- 12 landings predicted to be successful turned out to be successful:
  “True positive”
- 6 were predicted to fail and failed to land:
  “True negative”
- Both “False positive” and “False negative” cases equal to 0, meaning that the model never predicted a successful landing to fail, and vice versa.

Though it seems unlikely, the model may actually have an absolute accuracy. However, it would be best to evaluate the model on a bigger dataset to have a more confident evaluation.



Conclusions

- Launch site, Payload, Orbit Type, Year, Booster version and other variables have a great influence on a launch success rate.
- Flights become more reliable with time, with over 80% success rate in 2020.
- Of all launch sites, Kennedy Space Center has a much higher success rate than all others.
- Every recovery from orbits ES-L1, GEO, HEO, SSO was successful.
- v1.1 Booster has the highest rate of successful recovery.
- There is no connection between launch site's location and Falcon 9 success rate on it.
- Several ML models have shown good accuracy in predicting the success of a flight. KNN model has shown a 100% accuracy, correctly predicting every success and failure in test set.

Thank you!

Vladimir Morozov
Aspiring Data scientist

Reach me at:
LinkedIn | GitHub