IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hoang Minh Hanh
26. November 2022

Outline

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

Executive Summary

The purpose of this project is to determine the price of each rocket lauch for a new rocket company, SpaceY. SpaceY wants to use the information about SpaceX's launches to approximate the price of the rocket launch as currently SpaceX's launches seem to be the most cost efficient. Rockets are launched in 2 stages. The first stage does most of the work and the more expensive. If we can determine if the first stage will land, we can also determine the cost of the launch. We try to determine this by using 4 machine learning methods.

The following steps are taken during the analyses: Data Dollection, Data Wrangling and Preprocessing, Exploratory Data Analysis, Data Visualization with Folium and Plotly Dashboard, and Machine Learning Prediction.

The results of our analysis indicate that the success or failure of rocket launches can be predicted by some of the features under examination. As for which model to use for future rocket launch predictions, we can conclude that the Decision Tree algorithm performs best as it has the highest best accuracy among the four methods.

Introduction

- The purpose of this project is to determine the price of each rocket lauch for a new rocket company, SpaceY. SpaceY wants to use the information about SpaceX's launches to approximate the price of the rocket launch as currently SpaceX's launches seem to be the most cost efficient. In fact, 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.

- Rockets are launched in 2 stages. The first stage does most of the work and the more expensive. SpaceX's Falcon 9 rocket launches are relatively inexpensive since they can save a lot by being able to reuse the first stage. So if we can determine if the first stage will land, we can also determine the cost of the launch. SpaceY can use this information to see whether they want to bid against SpaceX for a rocket launch.

- The concrete question we want to answer: Based on the information and data about the Falcon 9 rocket launch, will the first stage of the Falcon 9 rocket launch will land?

Section 1

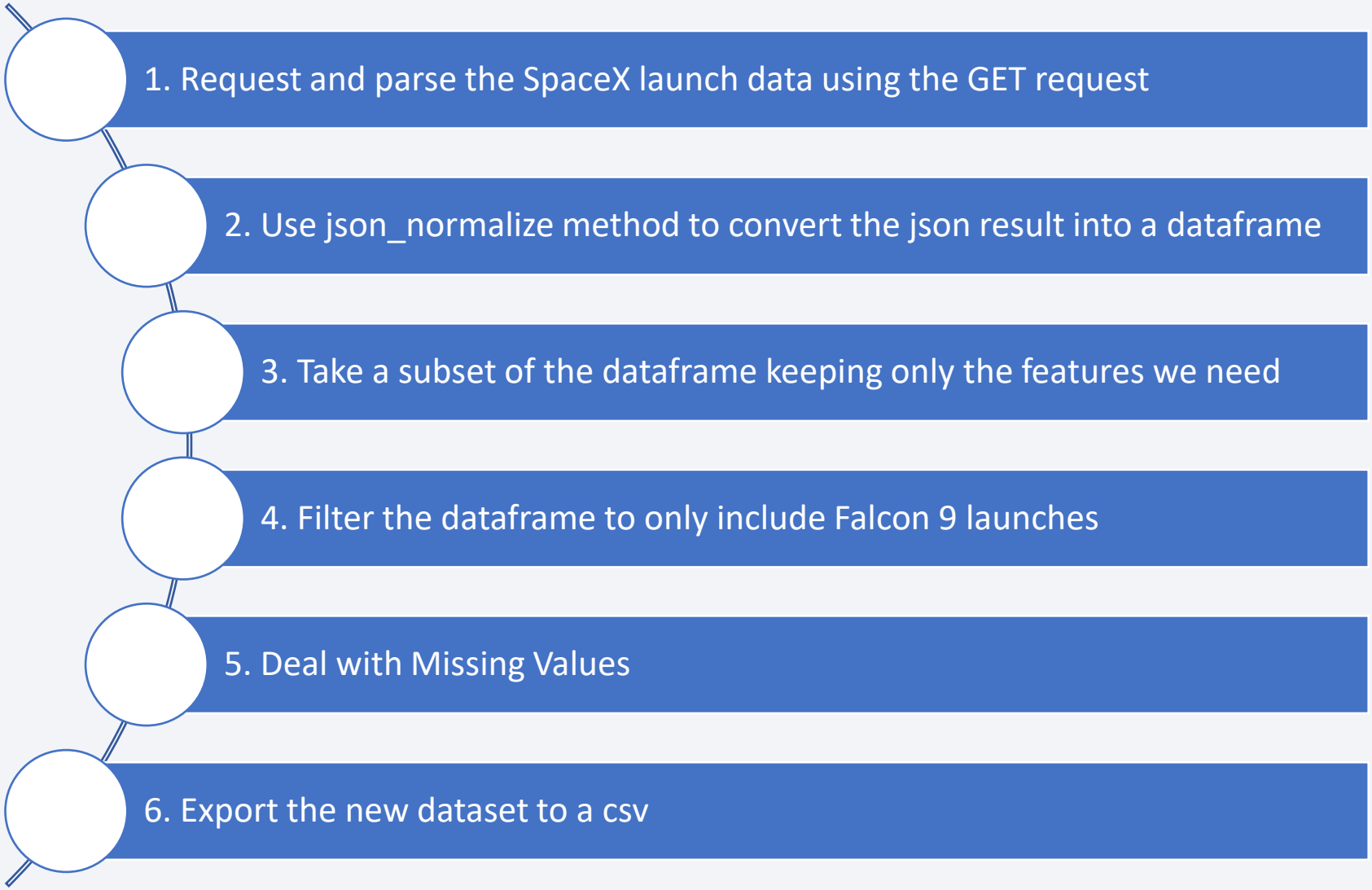
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect rocket launch data via the SpaceX API
- Perform data wrangling
  - Decode the response content by using a json parser
  - Treat the missing values on a case by case basis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune and evaluate 4 classification models

Data Collection – SpaceX API



- SpaceX API for Data Collection notebook GitHub link

Data Collection – Scraping

1. Perform an HTTP GET method to request the Falcon9 Launch HTML page
2. Use the BeautifulSoup object to find the third table that contains the actual launch records
3. Use the helper functions to process the web scraped HTML table
4. Create a data frame by parsing the launch HTML table
5. Export the new dataset to a csv

- SpaceX API for Web Scraping notebook GitHub link

Data Wrangling

1. Calculate the number of launches on each site using the launch table obtained in the previous step
2. Calculate the number and occurrence of each orbit since each launch aims to an dedicated orbit
3. Calculate the number and occurence of mission outcome (landing success/failure) per orbit type
4. Create a landing outcome label from Outcome column (landing success = 1 , landing failure = 0)
5. Determine the success rate
6. Export the new dataset to a csv

- SpaceX Data Wrangling notebook GitHub link

EDA with Data Visualization

- To visualize the data the following type of charts are used:
  - **Scatter plots:** Scatter plots are used to observe relationship between variables and uses dots to represent the relationship between them. In this case we examine the following relationships:
  - Flight Number vs. Payload Mass
  - Launch Site vs. Payload Mass
  - Flight Number vs. Orbit type
  - **Bar charts:** A bar chart represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. We use the bar chart to find which orbits have high success rate. To do that we create a bar chart for the success of each orbit.
  - **Line charts:** Line charts are used to represent the relation between two data X and Y on a different axis. Here we create a line chart with x axis to be Year and y axis to be average success rate, to get the average launch success trend.
- SpaceX EDA + Matlab GitHub link

EDA with SQL

- The following SQL queries are executed in order to answer assignment questions:
  - Display the names of the unique launch sites in the space mission
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- EDA using SQL Notebook GitHub link

Build an Interactive Map with Folium

The launch success rate may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Our goal in this exercise is to discover some of the factors for finding an optimal launch location by analyzing the existing launch site locations using the Folium library.

- The following tasks are performed:
  - Visualize launch sites by pinning them on a map
  - Add a highlighted circle area with a text label on a specific coordinate
  - Explore its proximity to see if we can easily find any railway, highway, coastline, etc.
  - Draw a line between a launch site to its closest city, railway, highway, etc.
- The questions answered:
  - Are all launch sites in proximity to the Equator line?
  - Are launch sites in close proximity to railways?
  - Are launch sites in close proximity to highways?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?
- Launch Site Analysis with Folium Notebook GitHub link

Build a Dashboard with Plotly Dash

- In order to perform interactive visual analytics on SpaceX launch data in real-time we use Plotly Dash to build a SpaceX Launch Record Dashboard
- After visual analysis using the dashboard, we obtain some insights to answer the following questions:
  - 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?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
- To visualise the launch success counts a pie chart is used is
- To find out whether the variable payload is correlated to the mission outcome a range slider is used
  - We can easily select different payload range and see if we can identify some visual patterns.
- SpaceX Dashboard code GitHub link

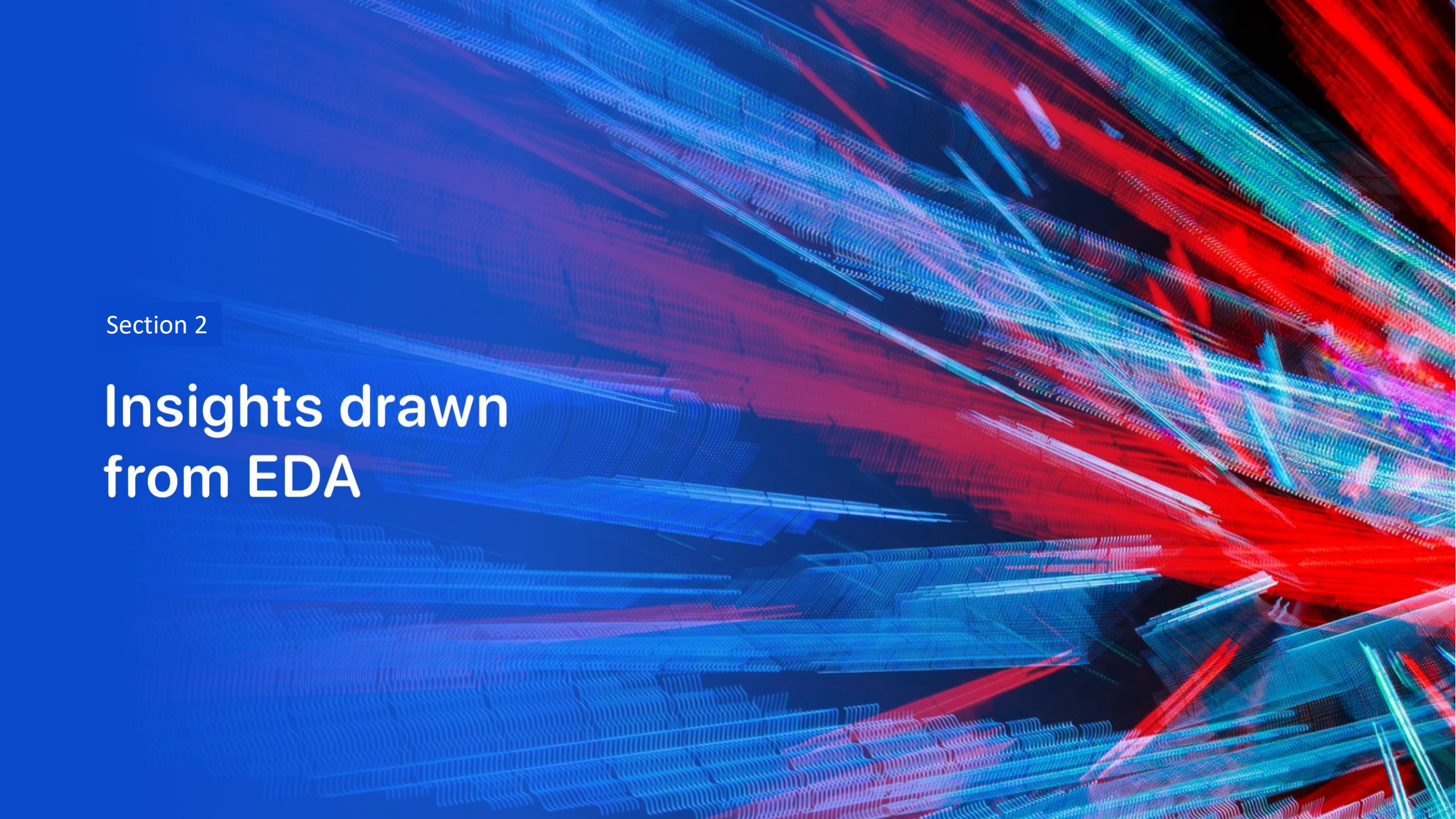
Predictive Analysis (Classification)

1. Create a “Class” feature that indicates whether the launch was a success or a failure (binary)
2. Standardize the dataset that contains the independent variables (X)
3. Split the data X and Y into training and test data
4. By using a logistic regression and GridSearchCV objects find the best parameters from the predefined dictionary parameters
5. Repeat step 4 with the SVM, Decision Tree and KNN methods
6. Calculate the accuracy on the test data using each method
7. Find the method that performs best

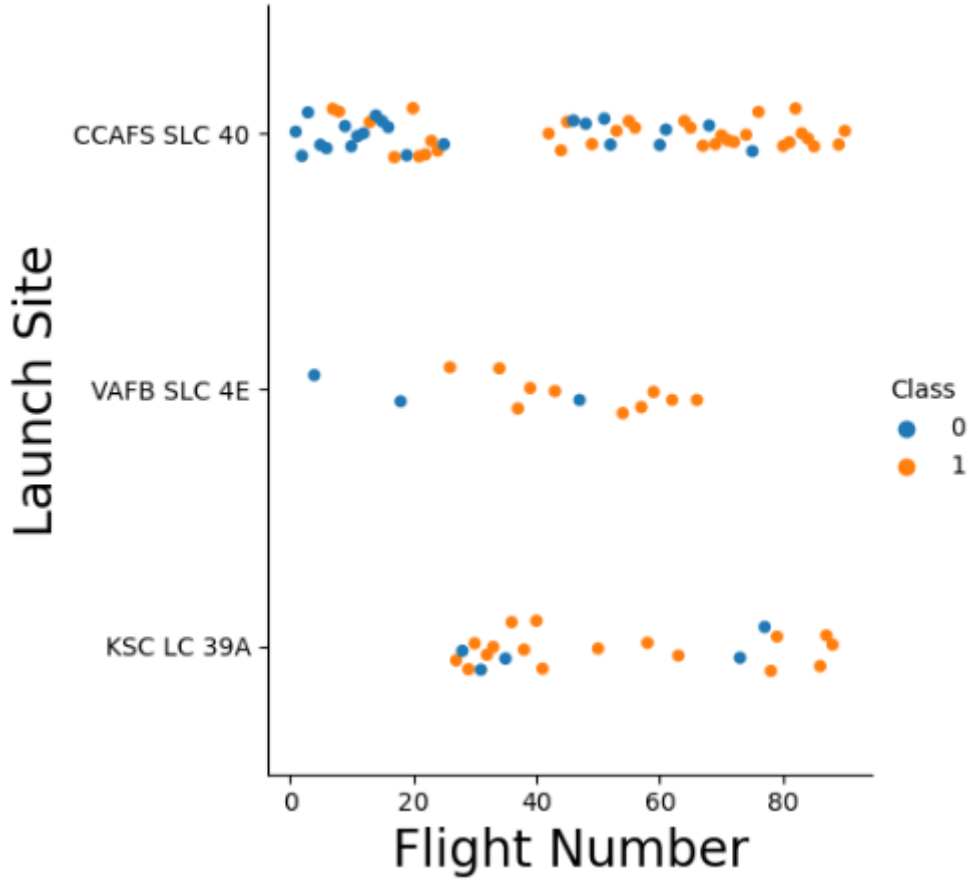
- SpaceX Machine Learning notebook GitHub link

Results

- During the Exploratory data analysis we discovered that
  - as the flight number increases, the first stage is more likely to land successfully
  - the more massive the payload, the less likely the first stage will return
  - the success rate since 2013 kept increasing till 2020
- Launch sites seem to be in proximity to the Equator line
- Launch sites are in close proximity to railways, highways, coastline and keep certain distance away from cities
- Out of the 4 Machine Learning algorithms (logistic regression, SVM, decision tree, KNN) none of them seem to be superior to the others
  - They all have the same accuracy that is 83.33%

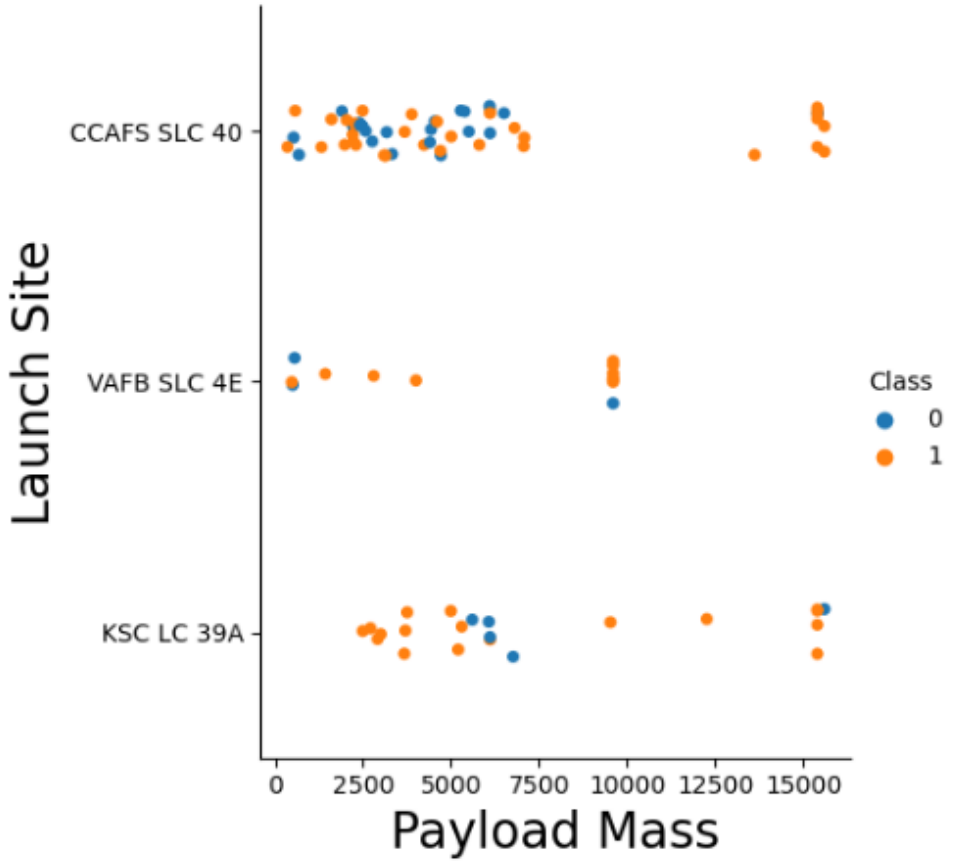

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



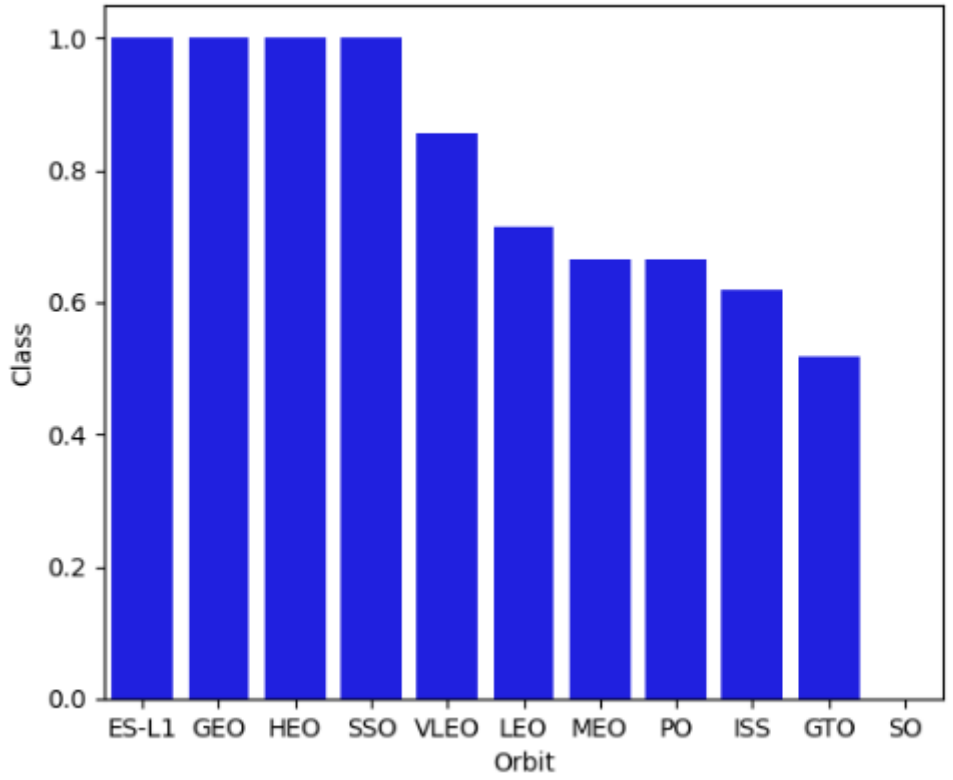
- The scatter plot of Flight Number vs. Launch Site indicates that the larger the Flight Number is, the higher the probability of a successful launch (class = 1)
- We can also notice that KSC LC 39A Launch Site has the highest success rate whereas CCAFS SLC 40 Launch Site seems to have the lowest success rate

Payload vs. Launch Site



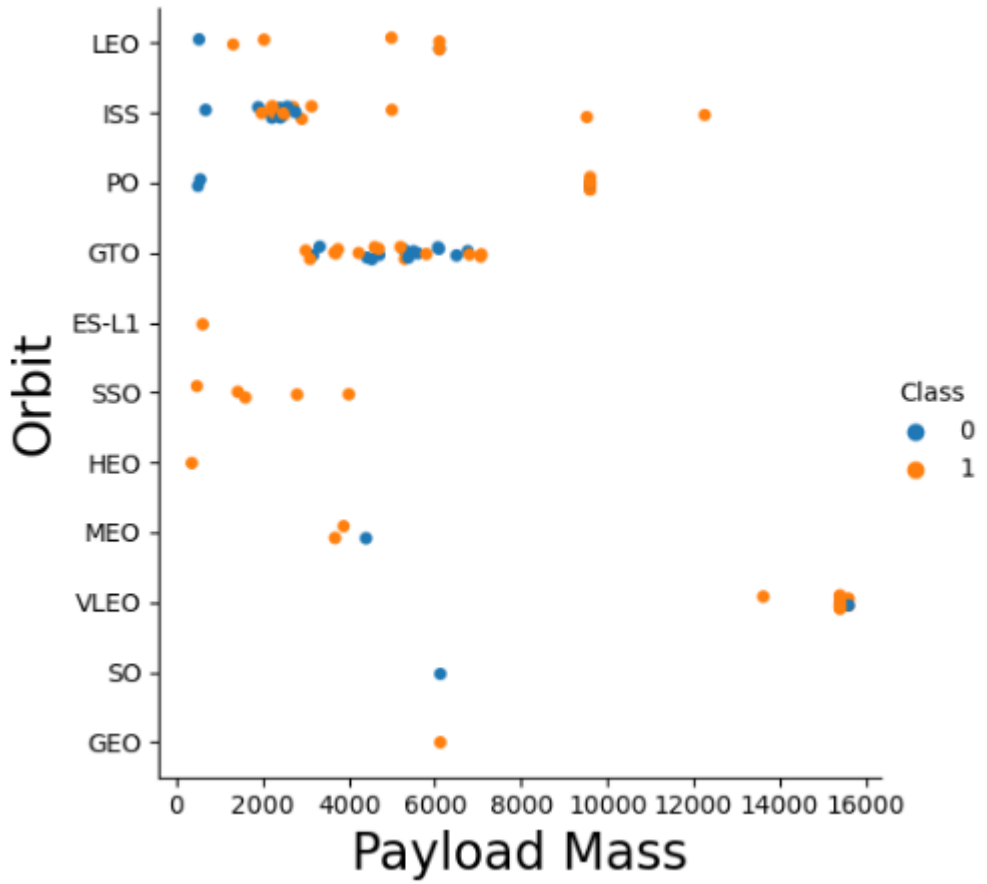
- The scatter plot of Payload Mass vs. Launch Site indicates that the larger the Payload Mass is (>=10000), the probability of a successful launch (class = 1) is very close to 100%
- We can also notice that KSC LC 39A Launch Site has the highest success rate whereas CCAFS SLC 40 Launch Site seems to have the lowest success rate

Success Rate vs. Orbit Type



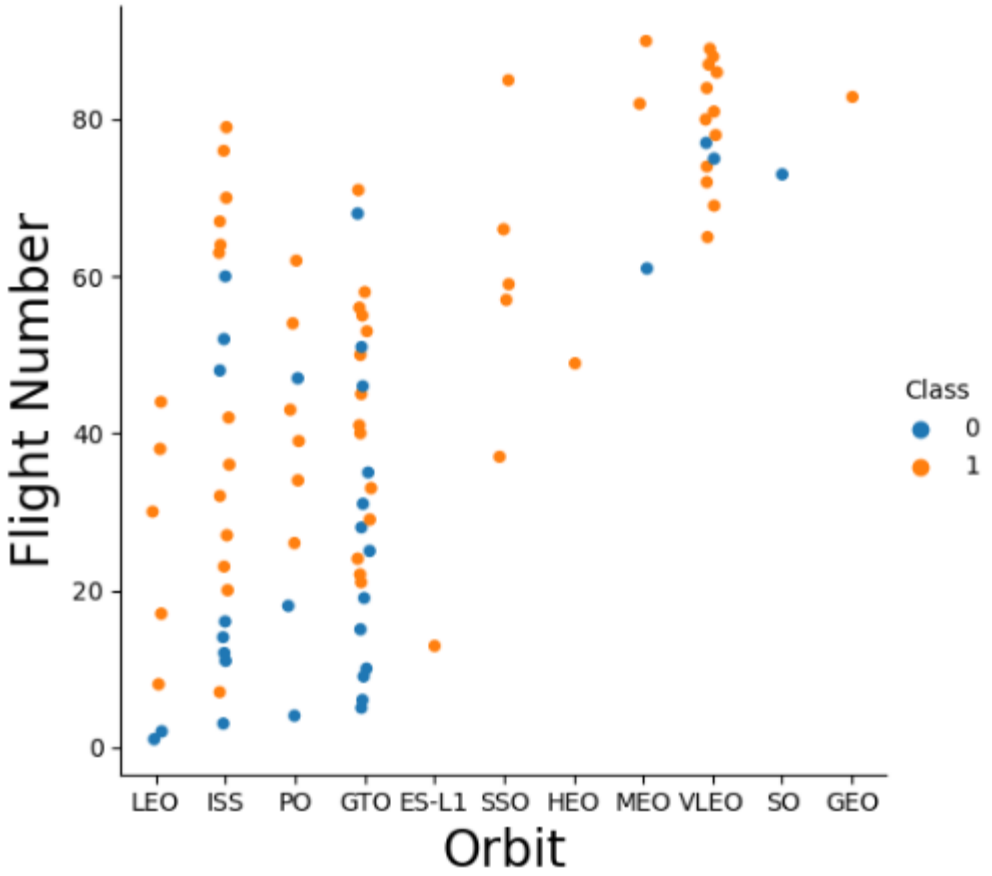
- The bar plot for the success rate of each orbit indicates that ES-L1, GEO, HEO and SSO orbit types have the highest success rate
- SO orbit type has zero success with the launches

Payload vs. Orbit Type



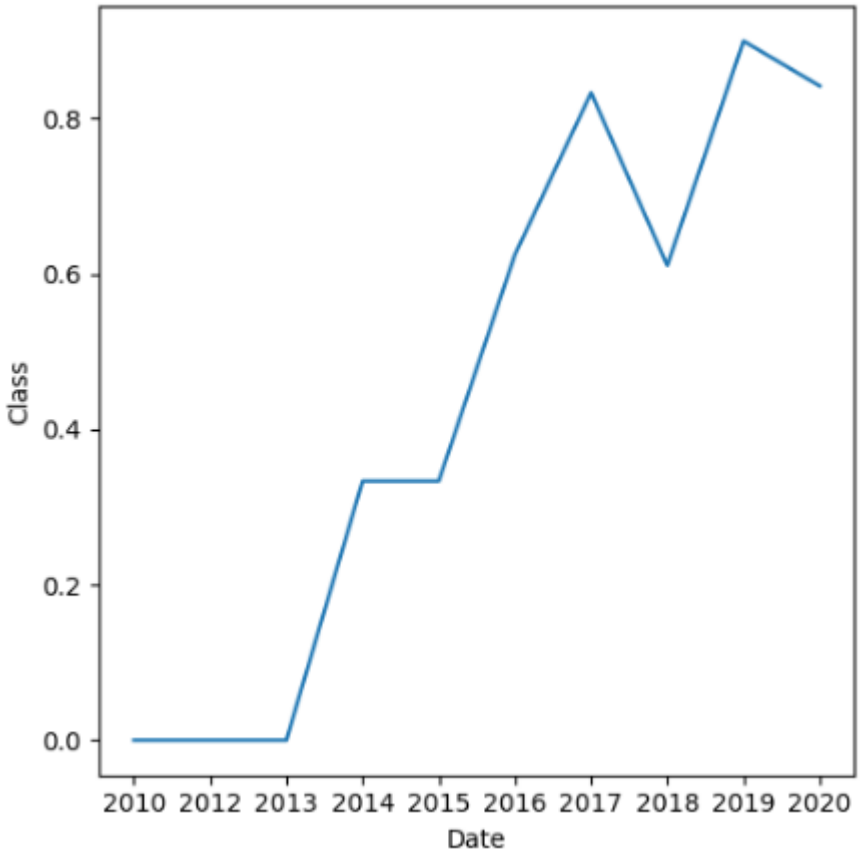
- The scatter plot of Payload Mass vs. Orbit indicates that for orbit SSO the launch was successful regardless of the Payload Mass
- GTO Orbit has mixed results (success rate is approx. the same as failure rate (50%)) so we cannot conclude anything
- Overall there seems to be no correlation between Payload Mass and Orbit type

Orbit Type vs. Flight Number



- The scatter plot of Orbit vs. Flight Number indicates that for the majority of orbit types, the larger the flight number is the higher the probability of a successful launch (class = 1)
- Exceptions are GTO and SO, here we cannot conclude a relationship between Orbit type and Flight Number

Launch Success Yearly Trend



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

All Launch Site Names

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

- These are the launch sites for SpaceX Falcon 9 Launches
- We use SQL “Select Distinct” statement on column “launch_site” to obtain this list

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- These are the launch sites for SpaceX Falcon 9 Launches where site name begins with “CCA”
- We use SQL “Like” statement in the “WHERE” clause on column “launch_site” to obtain this list

Total Payload Mass

sum_of_payload
45596

- The Total Payload Mass carried by boosters launched by NASA (CRS) is 45596 kg
- We use SQL ”SELECT SUM” statement on column “payload_mass__kg_” to obtain this figure
- In the “WHERE” clause we put customer=“NASA (CRS)”

Average Payload Mass by F9 v1.1

Average_payload
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg
- We use SQL ”SELECT AVG” statement on column “payload_mass__kg_” to obtain this figure
- In the “WHERE” clause we put booster_version = “F9 v 1.1”

First Successful Ground Landing Date

Date	Landing _Outcome
01-05-2017	Success (ground pad)

- The date when the first successful landing outcome in ground pad was achieved is 01-05-2017
- We use SQL ”SELECT MIN” statement on column “Date” to obtain this information
- In the “WHERE” clause we put "Landing _Outcome" = “Success (ground pad)”

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	PAYLOAD_MASS__KG_	Landing _Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- The table shows the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- We obtain this list when in the “WHERE” clause we filter on successful landing_outcome and set payload_mass__kg_ to be between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

MISSION_OUTCOME	Landing _Outcome
1	Precluded (drone ship)
98	Failure (parachute)
1	No attempt
1	Success (ground pad)

- The table counts the total number of successful and failure mission outcomes
- We use SQL ”SELECT COUNT” and GROUP BY statements on column “mission_outcome” to obtain this information

Boosters Carried Maximum 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

- The table shows the list of the names of the boosters which have carried the maximum payload mass
- We obtain this list when in the “WHERE” clause we use a subquery;
- In the subquery we create a sub-table: payload_mass__kg_= select avg(payload_mass__kg) from SpaceX

2015 Launch Records

Landing _Outcome	Booster_Version	Launch_Site	Date
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	10-01-2015
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	14-04-2015

- The table shows the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- We obtain this list (in this case only one occurance) when in the “WHERE” clause we filter on failed landing outcomes on drone ships and the year is set to be 2015

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

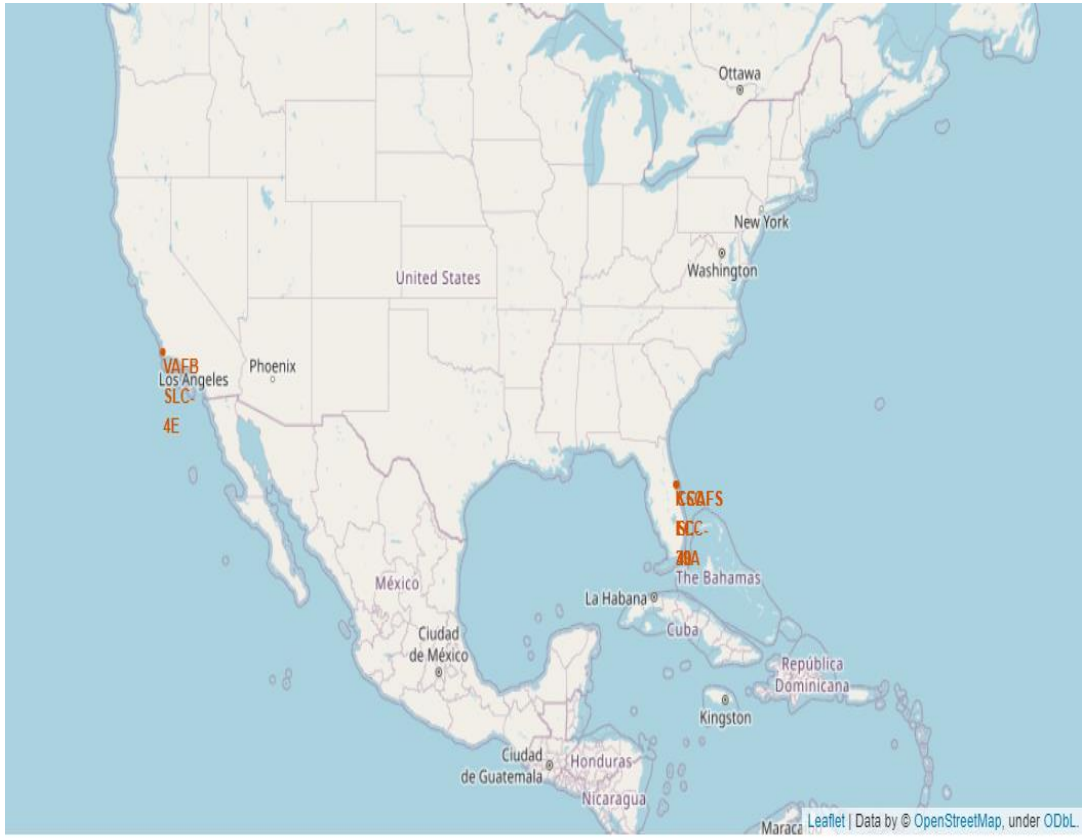
landing _outcome	Landing _Outcome
20	Success
10	No attempt
8	Success (drone ship)
6	Success (ground pad)
4	Failure (drone ship)
3	Failure
3	Controlled (ocean)
2	Failure (parachute)
1	No attempt

- The table shows 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
- We use SQL "SELECT COUNT" and "GROUP BY" statement on the column "landing_outcome" to obtain this list
- We use a subquery in the "FROM" clause
- We use ORDER BY on the first column

Section 3

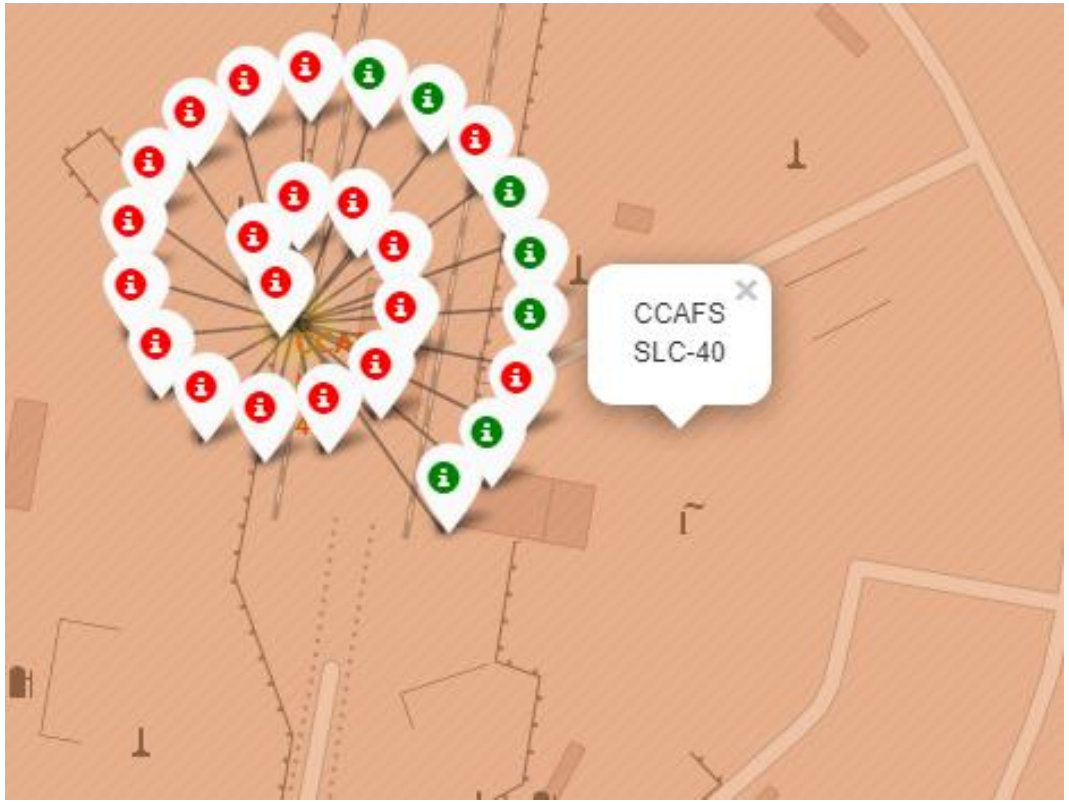
Launch Sites Proximities Analysis

Distance between a launch site to the selected coastline point



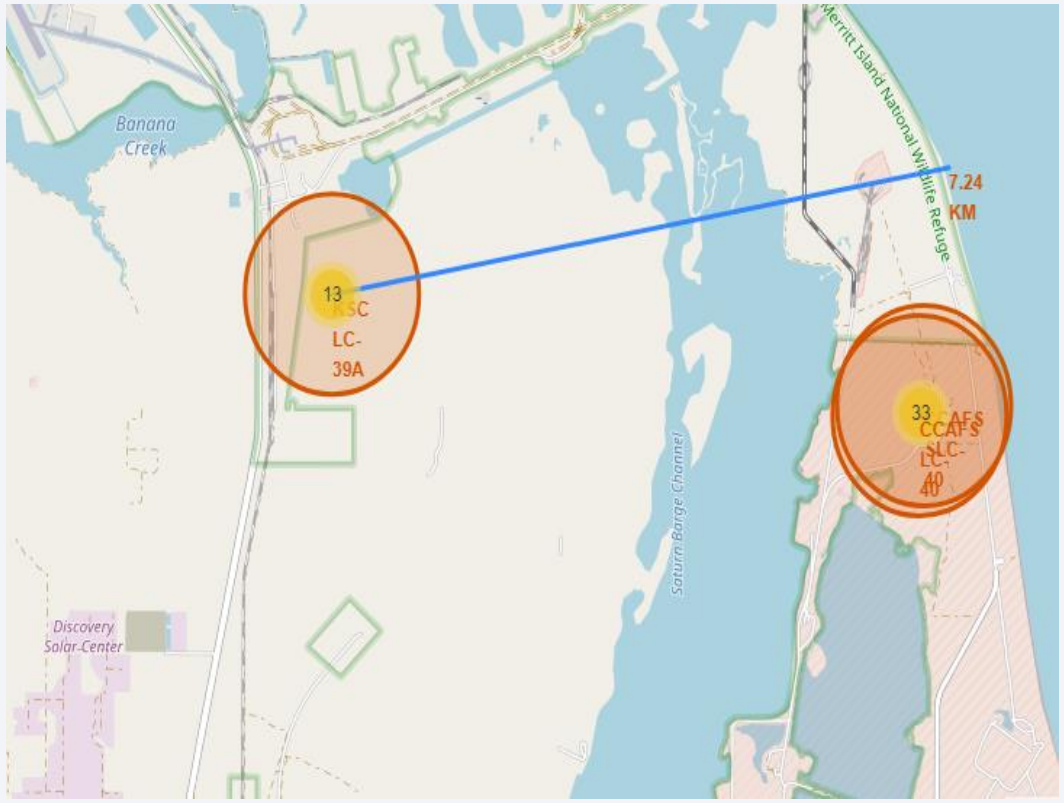
- With Folium we can generate a map with marked launch sites
- We can explore the map by zoom-in/out the marked areas, and try to answer the following questions:
  - Are all launch sites in proximity to the Equator line?
  - Are all launch sites in very close proximity to the coast?

Identify launch site success rates using color-coding

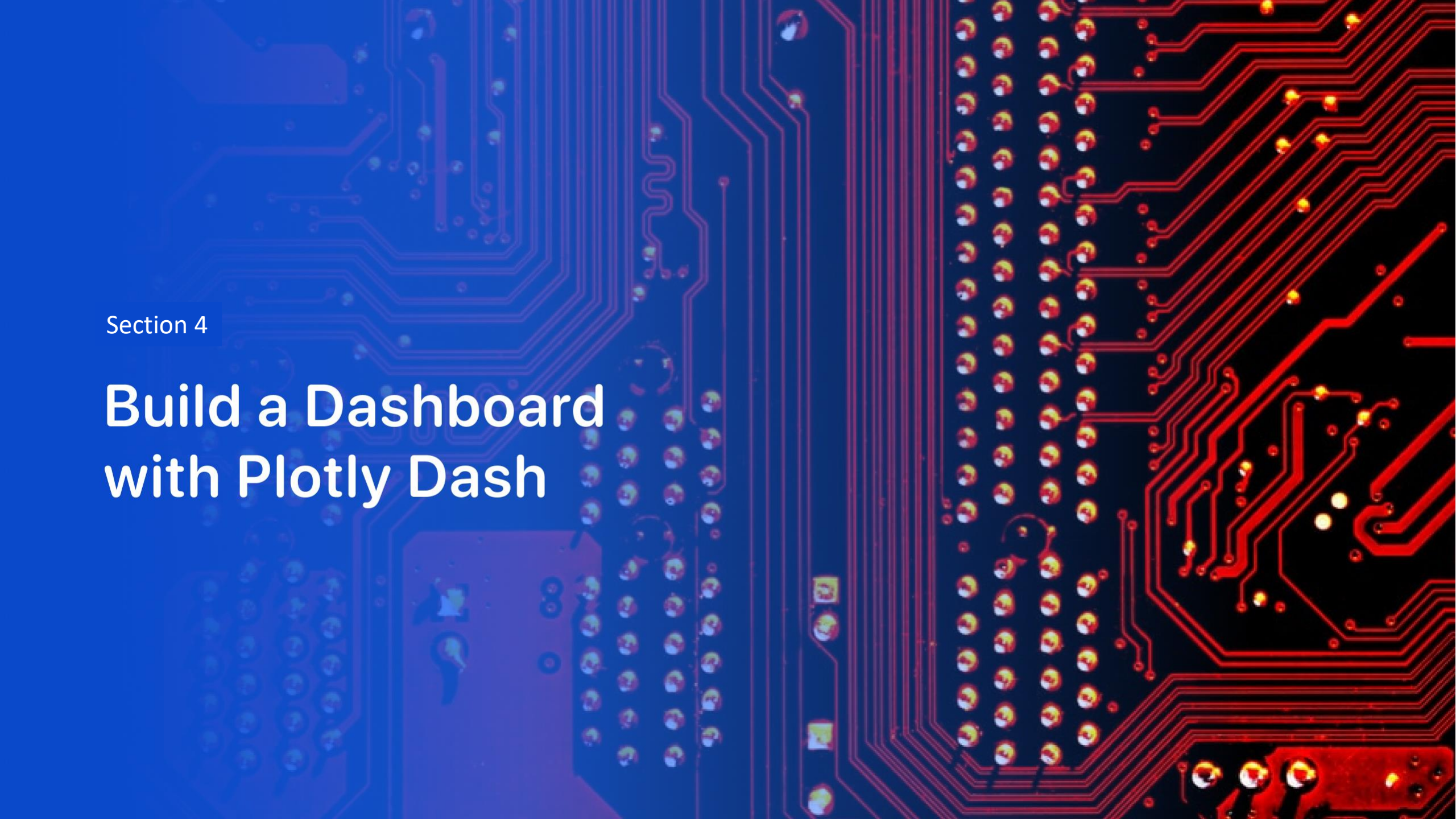


- For each launch site, we can create color-labeled launch outcomes on the map
  - Green: successful landing
  - Red: unsuccessful landing
- This is an example that shows that the majority of launches on CCAFS SLC 40 Launch Site was unsuccessfull

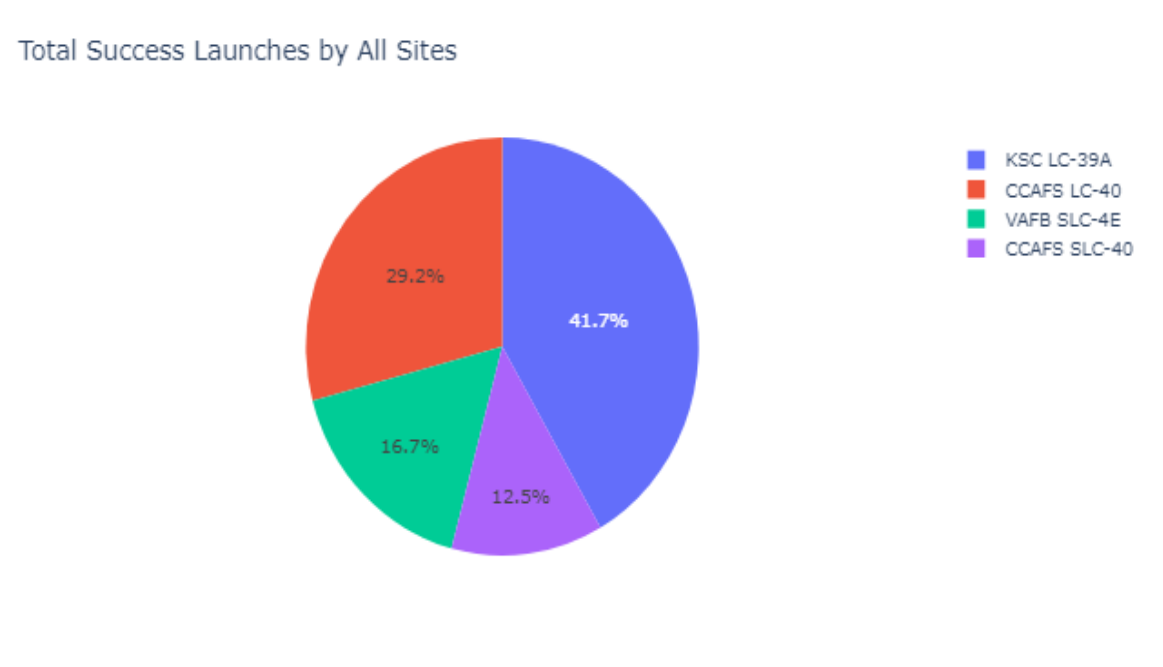
Distance between a launch site to the selected coastline point



- For each launch site, we can calculate the distance between the launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- This is an example that shows that the coastline is approx. 7.24 km away from KSC LC – 39A Launch Site

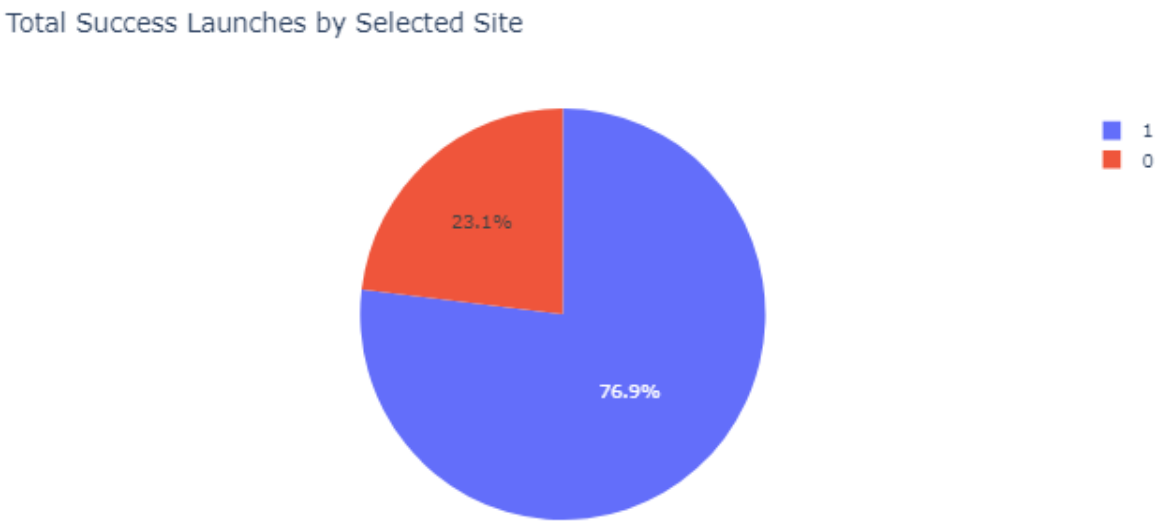

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



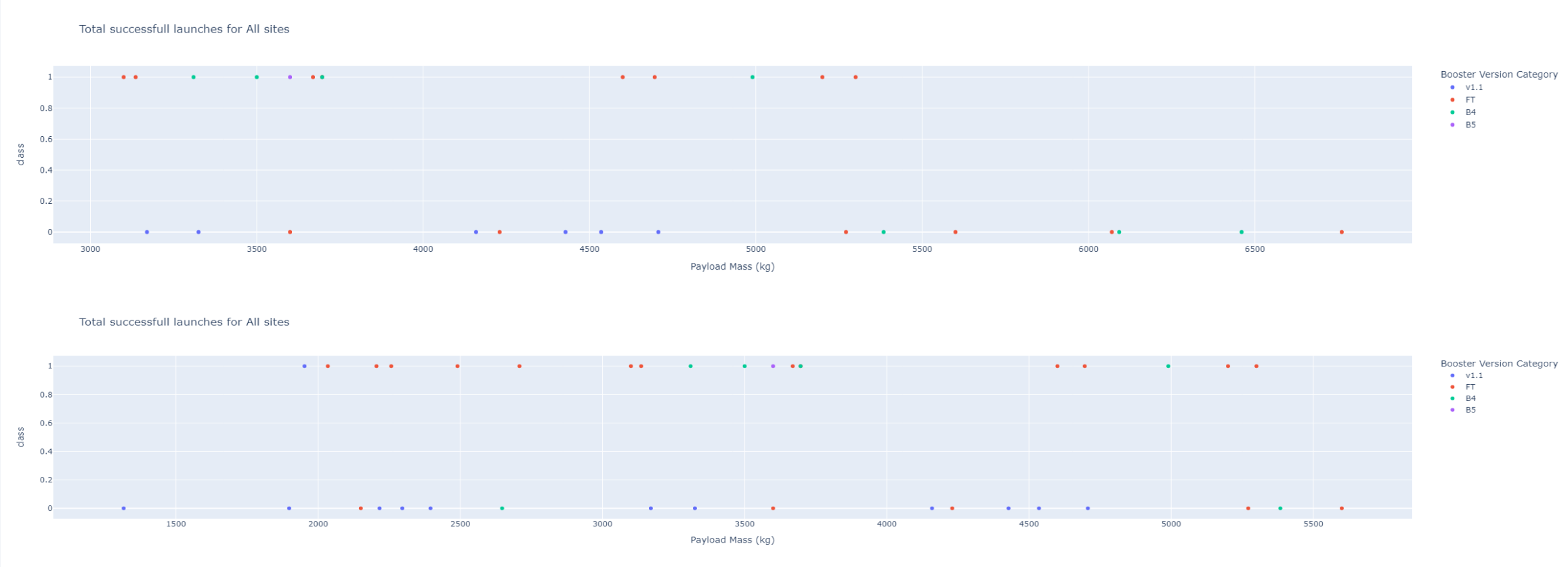
- KSC LC-39A Launch site has the most successful launches with 41.7% success ratio
- CCAFS LC-40 Launch site has the second most successful launches

Launch success count for KSC LC-39A



- KSC LC-39A Launch site has launch site with highest overall launch success ratio
- This is due to the fact that 77% of the a launches from this site landed successfully

Payload Mass influence on the outcome

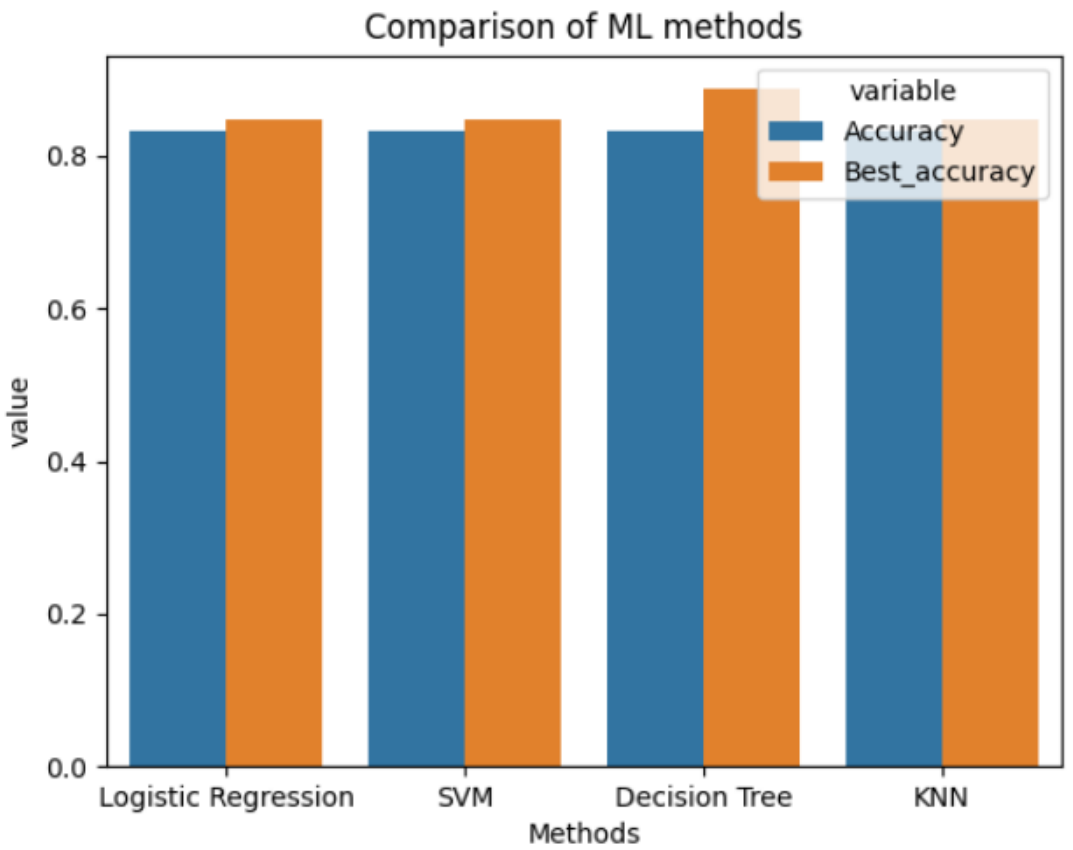


- The launch success rate is the highest when the Payload Mass is between 3000 and 3500 or between 4500 and 5500
- FT Booster version has the highest success rate

Section 5

Predictive Analysis (Classification)

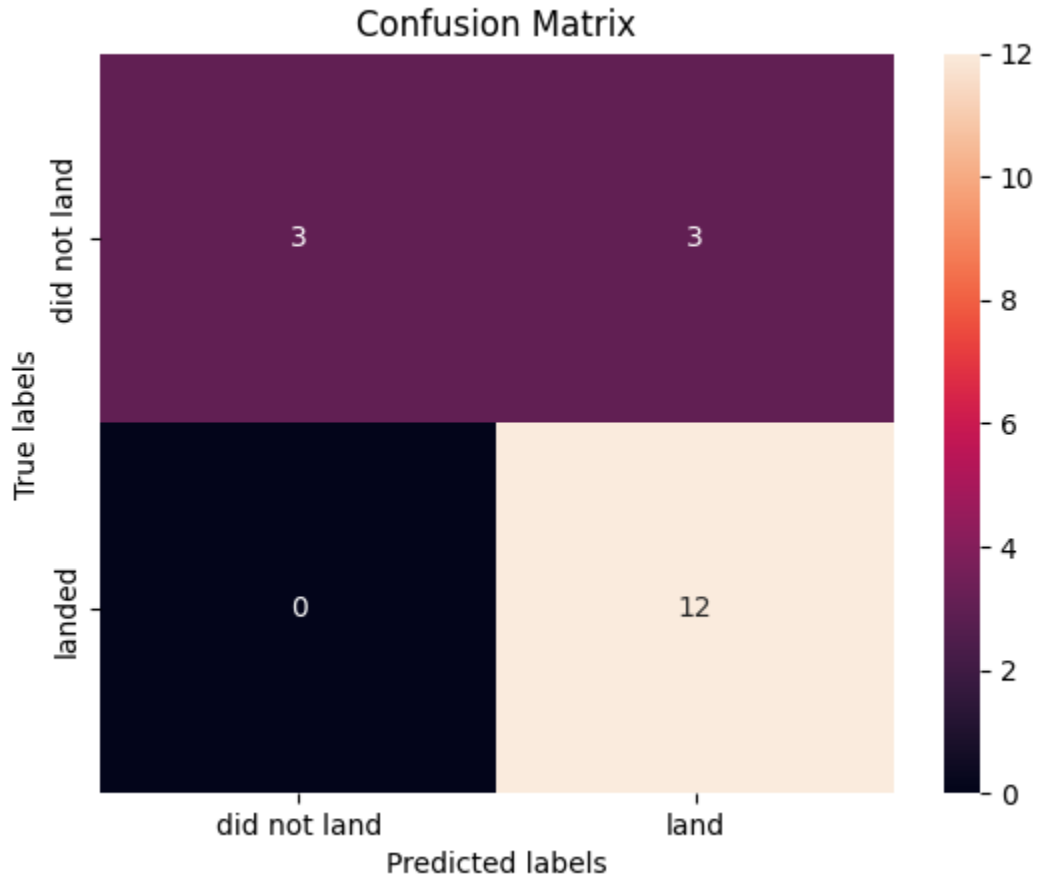
Classification Accuracy



- All 4 Machine Learning algorithms have the same model accuracy that is 83.33%
- However, Decision Tree has the highest best accuracy with 88.75% therefore we choose this method

	Methods	Accuracy	Best_accuracy
0	Logistic Regression	0.833333	0.846429
1	SVM	0.833333	0.848214
2	Decision Tree	0.833333	0.887500
3	KNN	0.833333	0.848214

Confusion Matrix



- True positive : Out of the 12 successful landings and the model also predicted 12 successful landings
- True negative: Out of the 6 unsuccessful landings the model predicted 3 unsuccessful landings
- False positive : Out of the 6 unsuccessful landings the model predicted 3 successful landings
- There were no false negative results

Conclusions

- Flight number and payload mass seem to be a good indicator of a successful landing
  - As the flight number increases the first stage is more likely to land successfully
  - As the payload mass increases the first stage is more likely to land successfully
- ES-L1, GEO, HEO and SSO orbit types have the highest success rate.
- The success rate of the launches since 2013 kept increasing till 2020.
- Launch sites are in close proximity to railways, highways, coastline and keep certain distance away from cities
- Out of the 4 Machine Learning algorithms (logistic regression, SVM, decision tree, KNN) Decision Tree has the highest best accuracy

Appendix

- GitHuB Repository link: https://github.com/hmh0490/IBM-Data-science-course

Thank you!