Winning Space Race with Data Science

Ephrem Getachew

September 30/2022

Outline

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

Executive Summary

- The data I used for this data science project came from SpaceX API and SpaceX Wikipedia page. I explored the data using SQL, and various visualization tools including folium maps and dashboards. Furthermore, I identified the relevant columns to be used as features for my data analysis endeavor. Additionally, I changed all categorical variables to binary using one-hot encoding and standardized the data. Finally, I used GridSearchCV to find the best parameters for various machine learning models and visualize the accuracy score of all models.
- I implemented four machine learning models on the data: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All the models produced similar results with an accuracy rate of about 83.33%. Although all our models over-predicted successful landings, more data is needed for better model determination and accuracy.

Introduction

- Project background and context
  - SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - Does the rate of successful landings increase over the years?
  - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

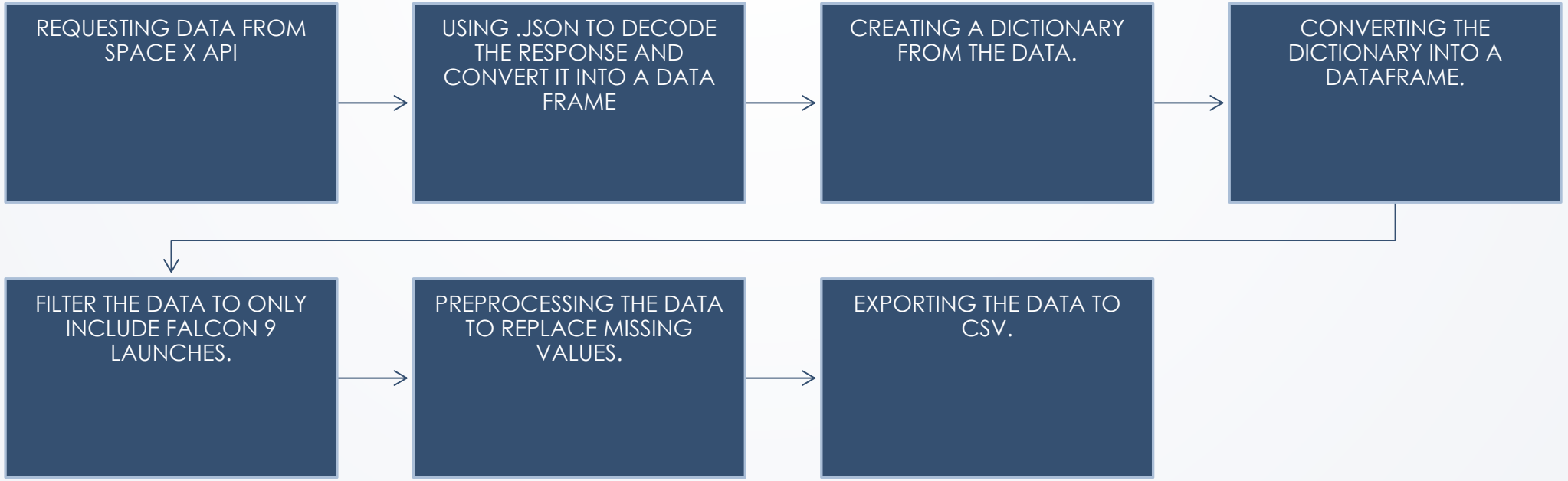
Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Performed data wrangling
- Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

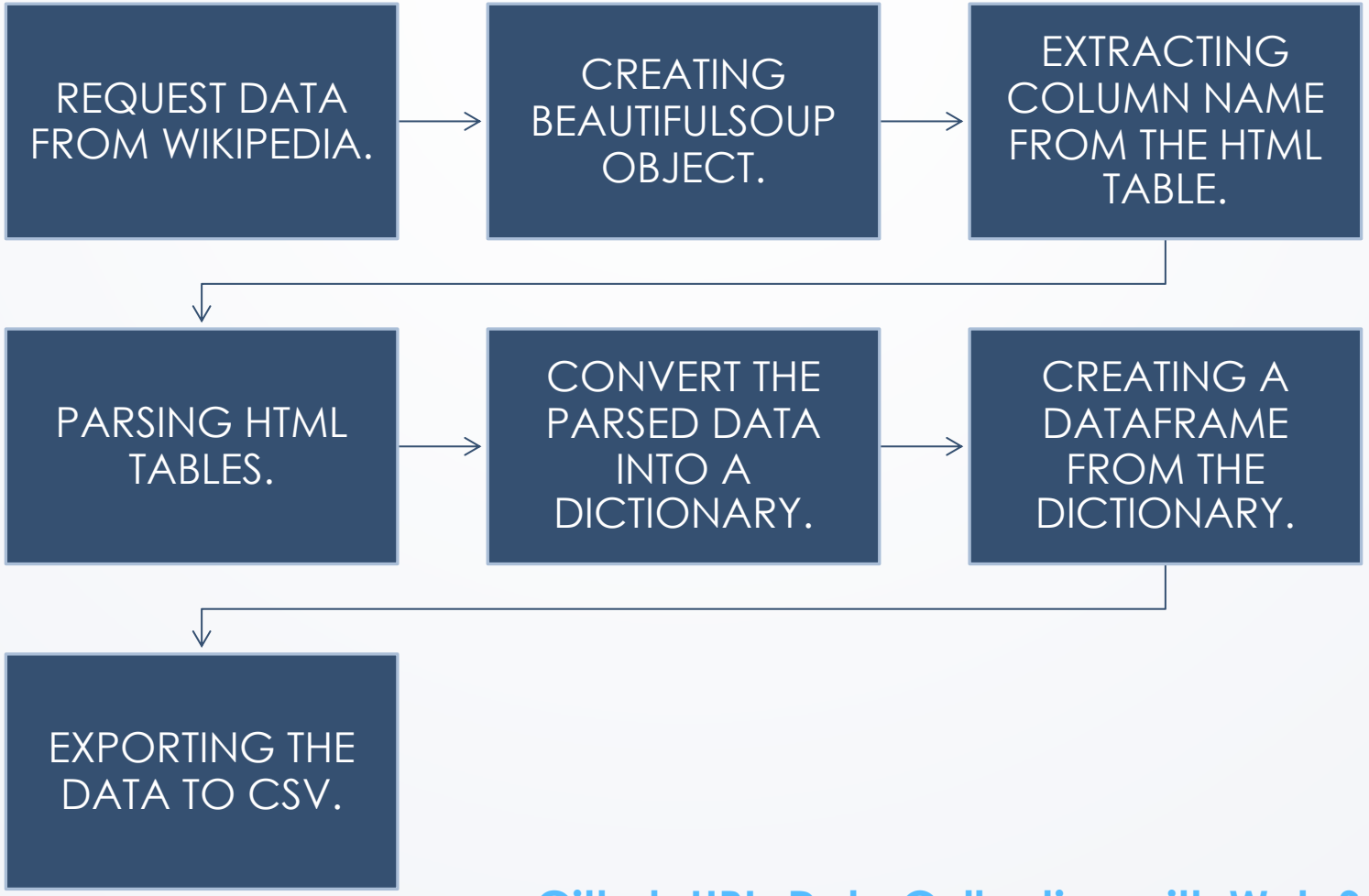
- **The data collection process involved a combination of API requests from SpaceX REST API and web scraping data from a table in SpaceX's Wikipedia entry. I used both of these data collection methods in order to get complete information about the launches for a more detailed analysis.**
- **Data columns obtained by using SpaceX RESTapi:**
  - ***Flightnumber, date, boosterversion, payloadmass, orbit, launchsite, outcome, flights, gridfins, reused, legs, landingpad, block, reusedcount, serial, longitude, latitude***
- **Data columns obtained by using Wikipedia web scraping:**
  - ***Flight no., Launch site, payload, payloadmass, orbit, customer, launch outcome, version booster, booster landing, date, time***

Data Collection – SpaceX API



Github URL: Data Collection API

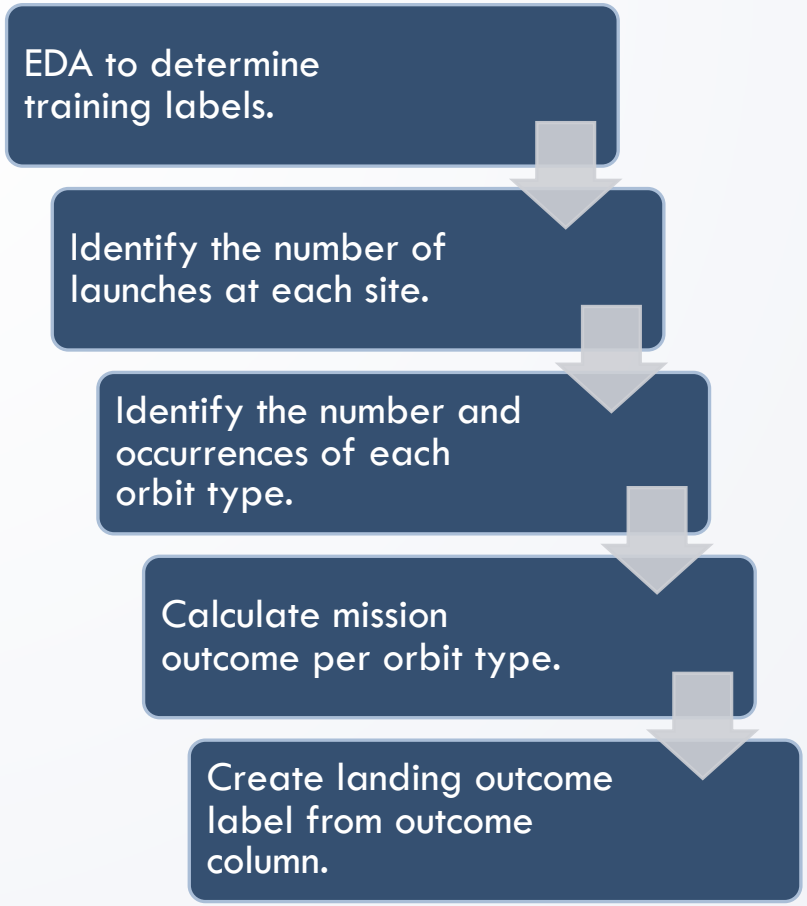
Data Collection - Scraping



[Github URL: Data Collection with Web Scraping]

Data Wrangling

- The outcome column has two components:
  - Mission outcome
  - Landing location
- Mission Outcomes:
  - True ocean means the mission outcome was successfully landed to a specific region of the ocean. Whereas false ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.
  - True RTLS means the mission outcome was successfully landed to a ground pad. Whereas false RTLS means the mission outcome was unsuccessfully landed to a ground pad.
  - True ASDS means the mission outcome was successfully landed on a drone ship. While false ASDS means the mission outcome was unsuccessfully landed on a drone ship.
- I created a training label with landing outcomes where with "1" means the booster successfully landed, "0" means it was unsuccessful.



[Github URL: Data Wrangling]

EDA with Data Visualization

- **Scatter plots show the relationship between variables. If a relationship exists, it could be used in the machine learning model.**
- **Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.**
- **Line charts show trends in data over time (time series).**
- **Charts were plotted:**
  - ***Flight number vs. Payload mass, flight number vs. Launch site, payload mass vs. Launch site, orbit type vs. Success rate, flight number vs. Orbit type, payload mass vs orbit type, and success rate yearly trend***

Github URL: EDA with Data Visualization

EDA with SQL

- Performed SQL queries:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking 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

Github URL: EDA with SQL

Build an Interactive Map with Folium

- Markers of all launch sites:
- Added marker with a circle, popup label, and text label of NASA Johnson space center using its latitude and longitude coordinates as a start location.
- Added markers with a circle, popup label, and text label of all launch sites using their latitude and longitude coordinates to show their geographical locations and proximity to the equator and coasts.
- Colored markers of the launch outcomes for each launch site:
- Added colored markers of success (green) and failed (red) launches using marker clusters to identify which launch sites have relatively high success rates.
- Distances between a launch site to its proximities:
- Added colored lines to show distances between the launch site KSC lc-39a (as an example) and its proximities like railway, highway, coastline, and closest city.

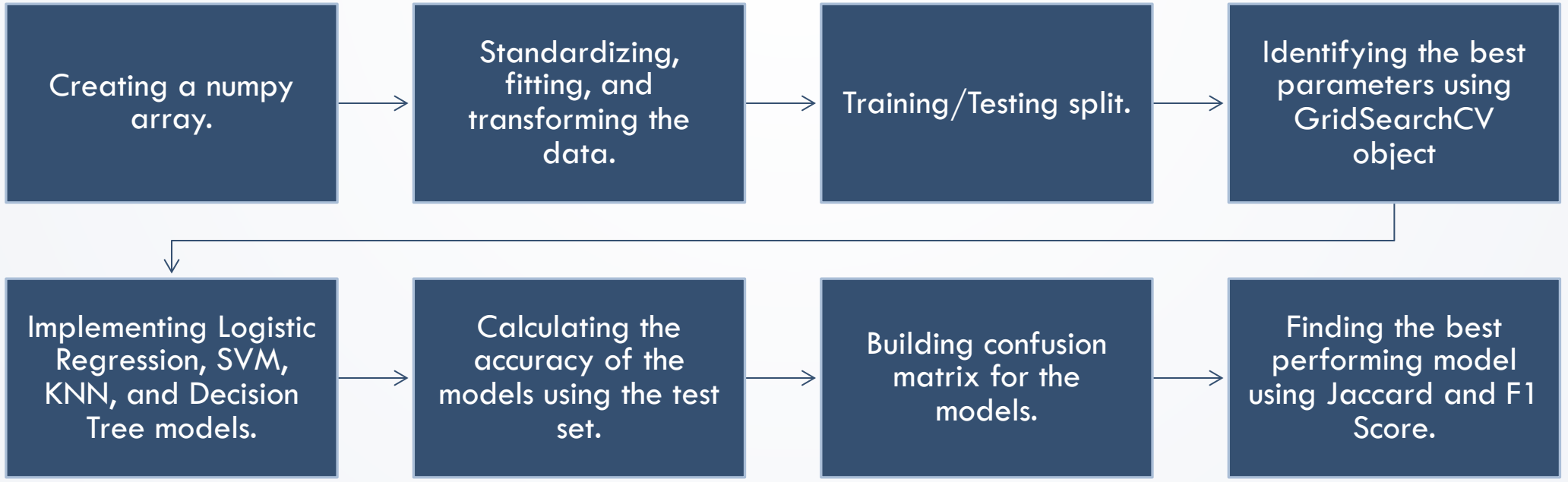
Github URL: Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- Launch sites dropdown list:
  - Added a dropdown list to enable launch site selection.
- Pie chart showing success launches (all sites/certain site):
  - Added a pie chart to show the total successful launches count for all sites and the success vs. Failed counts for the site, if a specific launch site was selected.
- Slider of payload mass range:
  - Added a slider to select payload range.
- Scatter chart of payload mass vs. Success rate for the different booster versions:
  - Added a scatter chart to show the correlation between payload and launch success.

Github URL: Space X Dash App

Predictive Analysis (Classification)

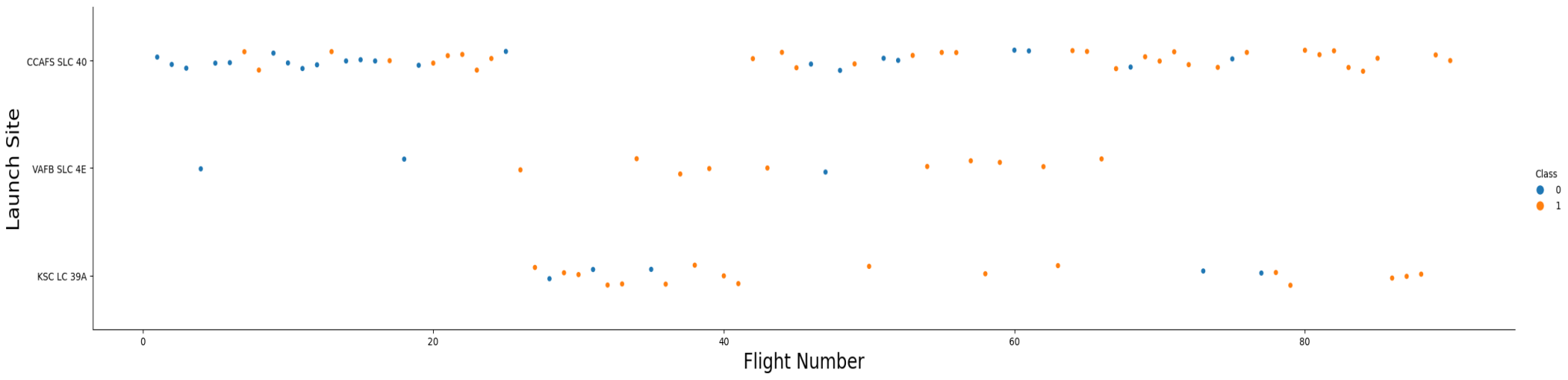


[Github URL: Machine Learning Prediction]

Section 2

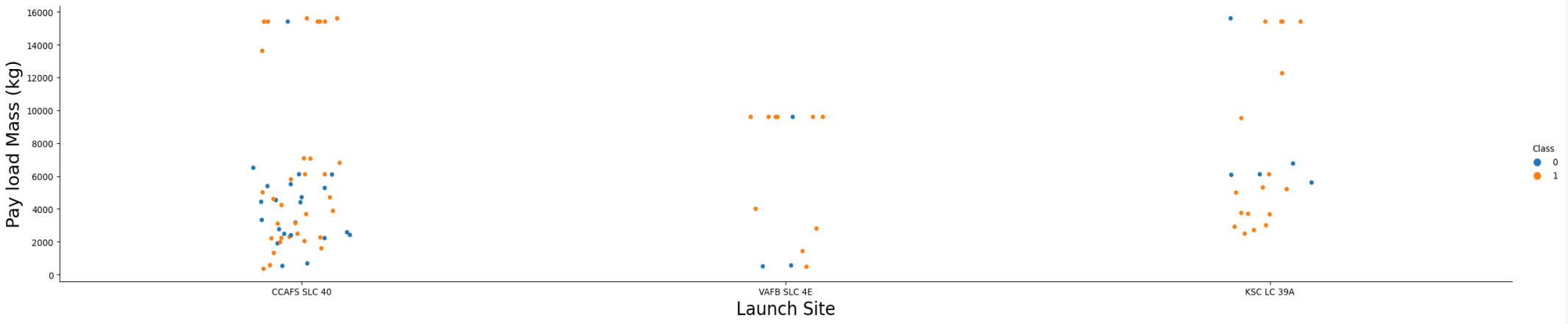
Insights drawn from EDA

Flight Number vs. Launch Site



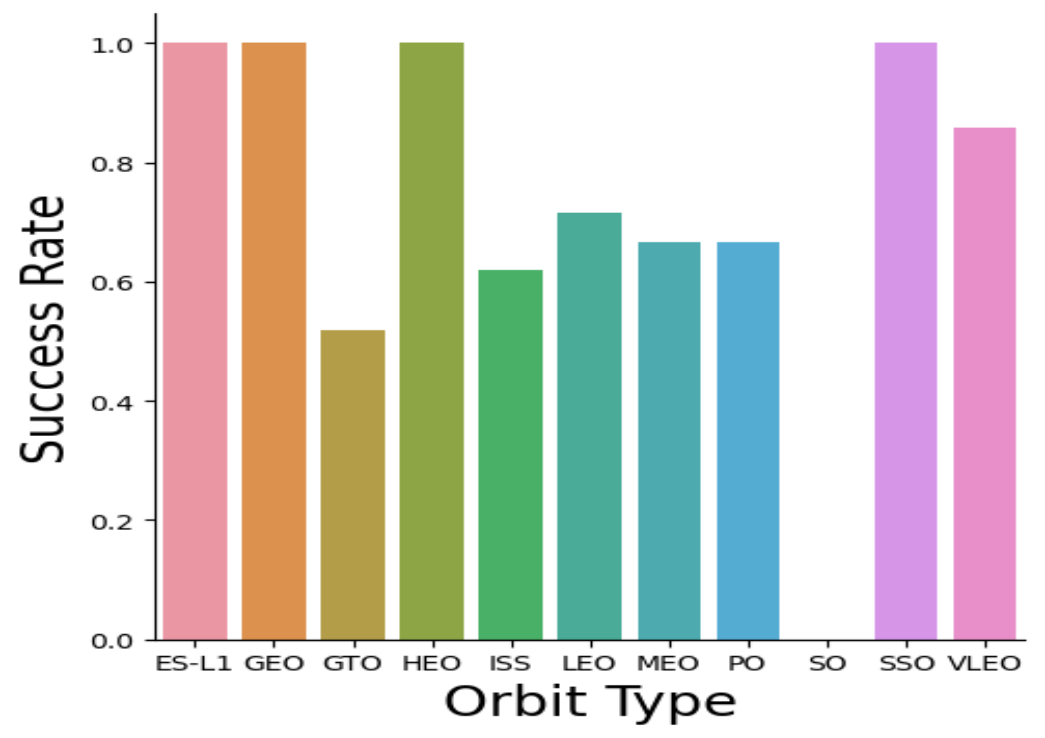
- **Insights:**
  - ***The earliest flights all failed while the latest flights all succeeded.***
  - ***The CCAFS SLC 40 launch site has about a half of all launches.***
  - ***VAFB SLC 4E and KSC LC 39A have higher success rates.***
  - ***It can be assumed that each new launch has a higher rate of success.***

Payload vs. Launch Site



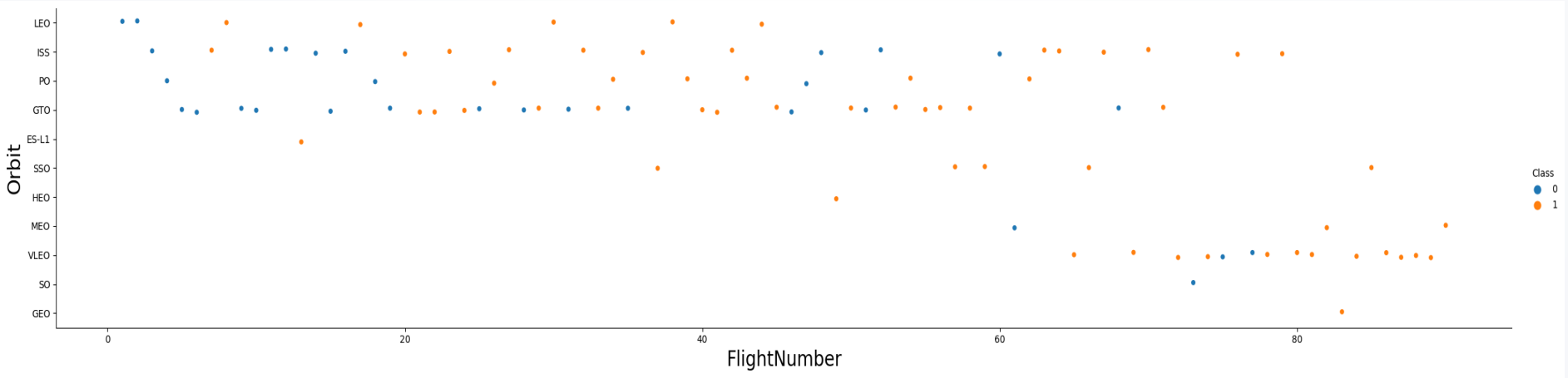
- **Insights:**
  - ***For every launch site the higher the payload mass, the higher the success rate.***
  - ***Most of the launches with payload mass over 7000 kg were successful.***
  - ***KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.***

Success Rate vs. Orbit Type



- **Insights:**
  - ***Orbits with 100% success rate:***
    - ***ES-L1, GEO, HEO, SSO***
  - ***Orbits with 0% success rate:***
    - ***SO***
  - ***Orbits with success rate between 50% and 85%:***
    - ***GTO, ISS, LEO, MEO, PO, VLEO***

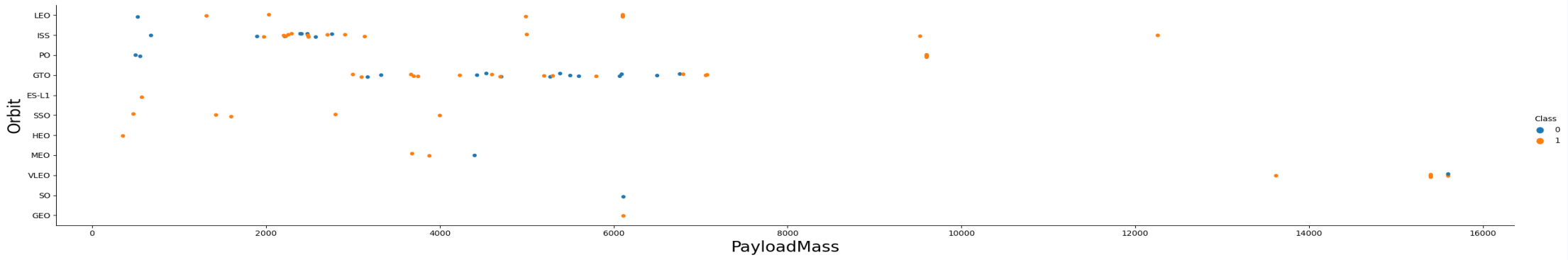
Flight Number vs. Orbit Type



Insights:

- ***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.***

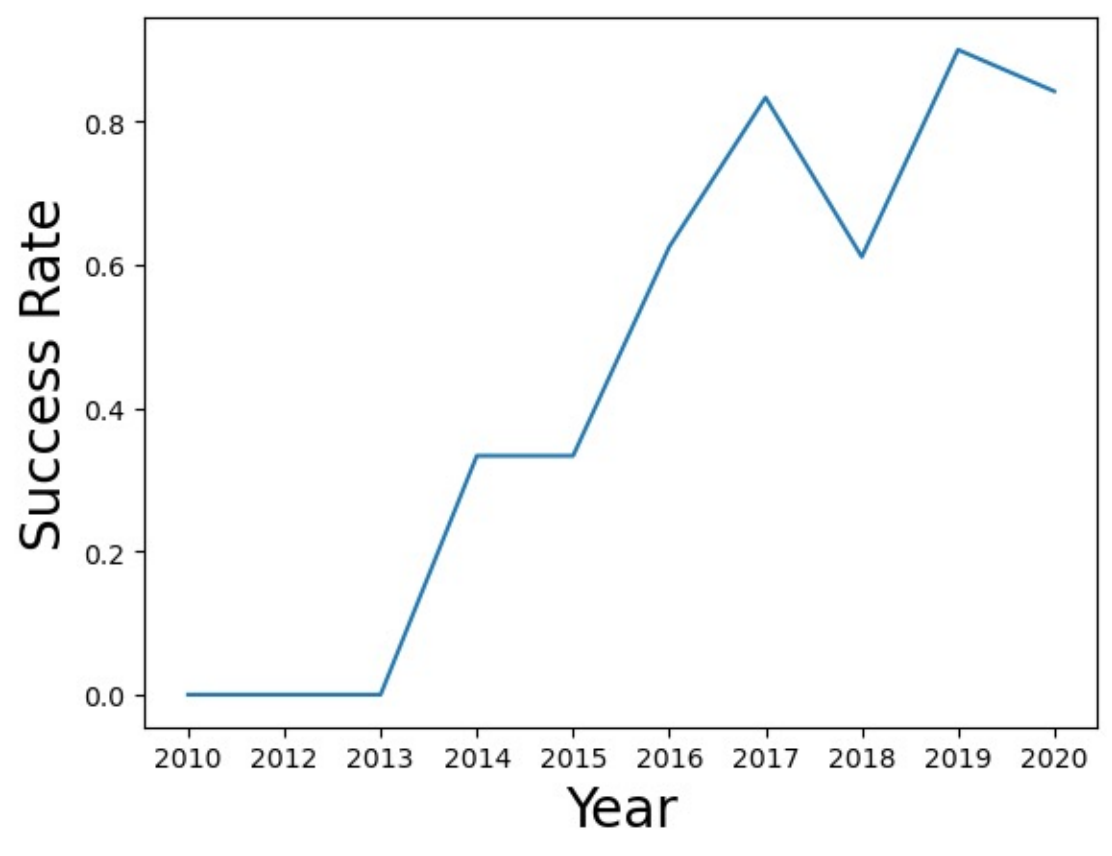
Payload vs. Orbit Type



Insights:

- ***Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.***

Launch Success Yearly Trend



Insights:

- ***The success rate since 2013 kept increasing till 2020.***

All Launch Site Names

DESCRIPTION:

- ***Displaying the names of the unique launch sites in the space mission.***

In [10]:
```
%sql select distinct LAUNCH_SITE from SPACEX
```

```
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb2
Done.
```

Out[10]:

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

Launch Site Names Begin with 'CCA'

In [12]:
```
%sql select * from SPACEX where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[12]:

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

DESCRIPTION:

- **DISPLAYING 5 RECORDS WHERE LAUNCH SITES BEGIN WITH THE STRING 'CCA'.**

Total Payload Mass

```
In [14]: %sql select sum(payload_mass__kg_) as sum from SPACEX where customer like 'NASA (CRS)'
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdon
Done.
Out[14]: SUM
         45596
```

Description:

- ***Displaying the total payload mass carried by boosters launched by NASA (CRS).***

Average Payload Mass by F9 v1.1

```
In [15]: %sql select avg(payload_mass__kg_) as Average from SPACEX where booster_version like 'F9 v1.1%'

 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.clo
Done.
Out[15]: average
         2534
```

Description

- ***Displaying average payload mass carried by booster version F9 v1.1.***

First Successful Ground Landing Date

```
In [16]: %sql select min(date) as Data from SPACEX where mission_outcome like 'Success'

 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.data
Done.
Out[16]:
DATA
2010-06-04
```

Description:

- ***Listing The Date When The First Successful Landing Outcome In Ground Pad Was Achieved.***

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [20]: %sql select booster_version from SPACEX where (mission_outcome like 'Success') and (payload_mass__kg_ BETWEEN 4000 and 6000) and (landing__outcome like 'Success (drone ship)
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
Out[20]:
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2
```

- **Description:**
  - ***Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.***

Total Number of Successful and Failure Mission Outcomes

In [21]:
```
%sql select mission_outcome, count(*) as Count from SPACEX GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[21]:

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

Description:

- **Listing the total number of successful and failure mission outcomes.**

Boosters Carried Maximum Payload

In [22]:
```
max_payload_mass = %sql select max(payload_mass__kg_) from SPACEX
max_version = max_payload_mass[0][0]
%sql select booster_version from SPACEX where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEX)
```

```
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[22]:

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

- **Description:**
  - ***Listing the names of the booster versions which have carried the maximum payload mass.***

2015 Launch Records

In [24]:
```
%sql select landing__outcome, booster_version, launch_site from SPACEX where DATE like '2015%' and landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[24]:

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

Description

- ***Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.***

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

In [25]:
```
%sql select landing__outcome, count(*) from SPACEX where Date >= '2010-06-04' and Date <= '2017-03-20' group by landing__outcome order by count desc
```

```
 * ibm_db_sa://cll34109:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[25]:

landing__outcome	2
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

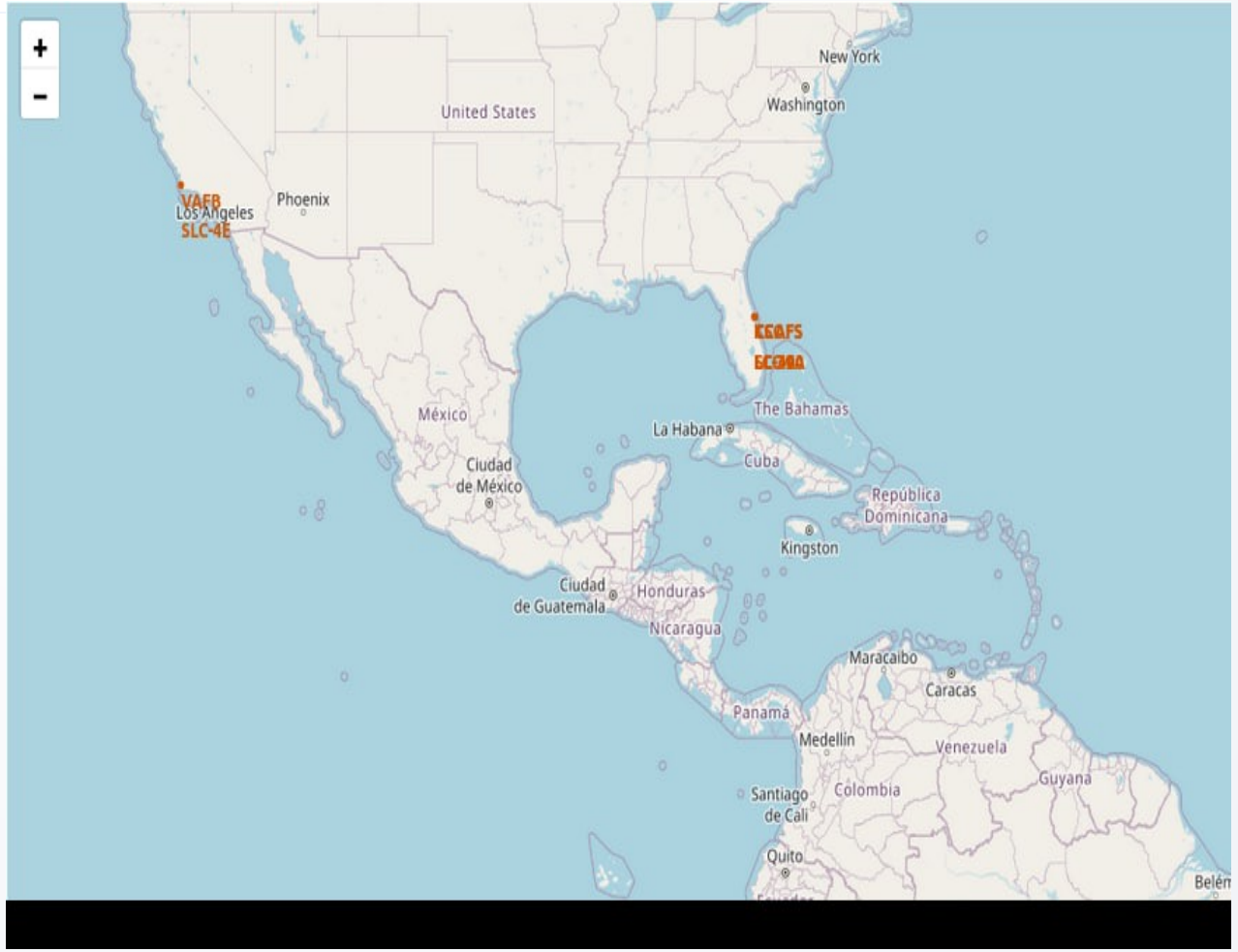
Descritpion

- ***Ranking 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.***


Section 3
Launch Sites
Proximities Analysis

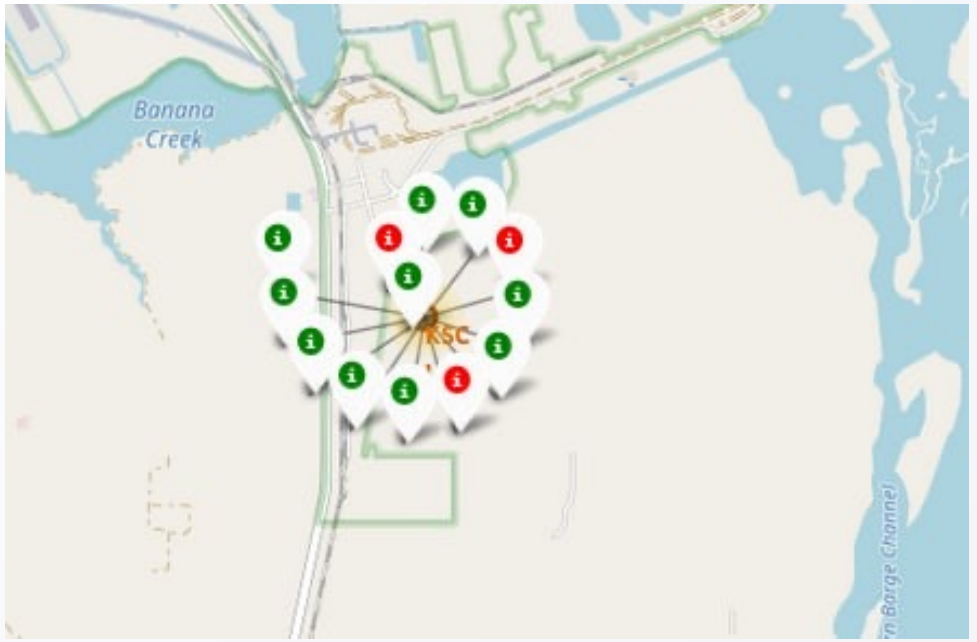
Launch Site Location Markers

- Most of launch sites are in proximity to the equator line. The land is moving faster at the equator than any other place on the surface of the earth. Anything on the surface of the earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



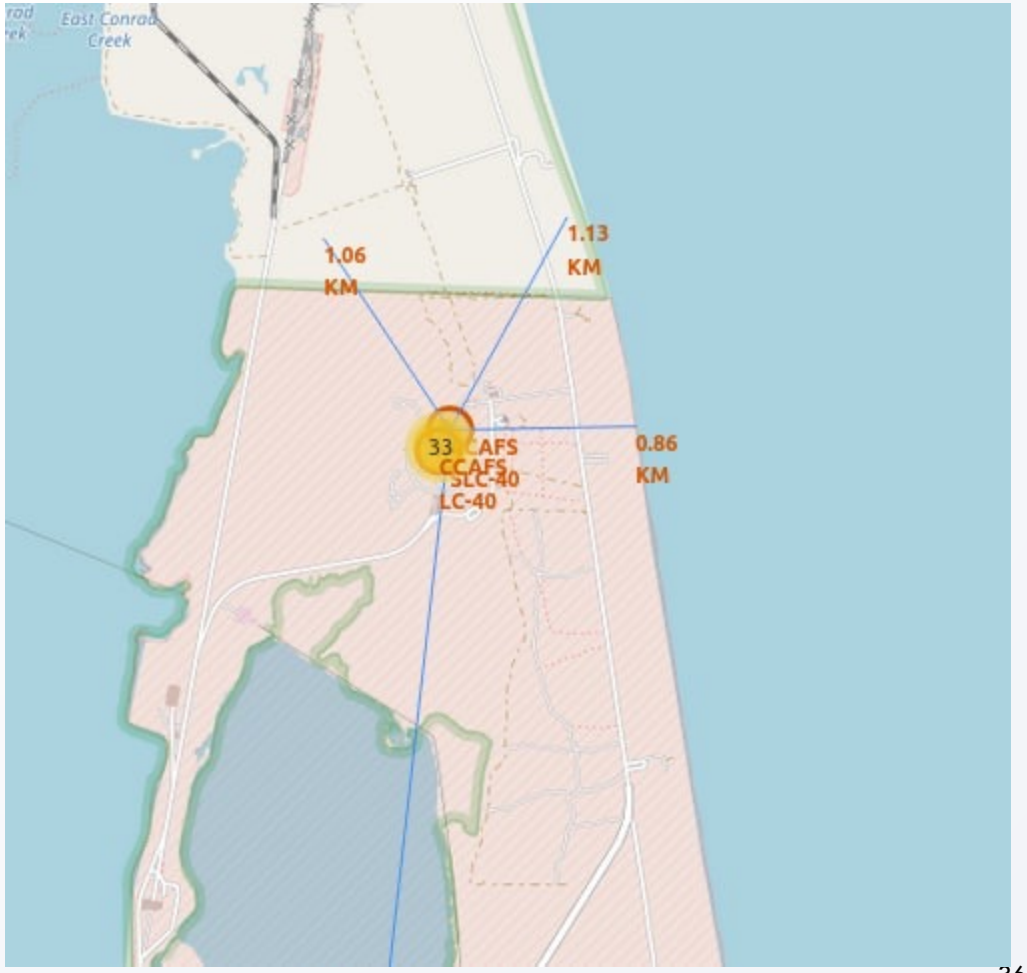
Color-labeled Launch Markers

- **Explanation:**
  - **From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.**
    - **Green marker = successful launch**
    - **Red marker = failed launch**
  - **Launch site KSC LC-39A has a very high success rate.**



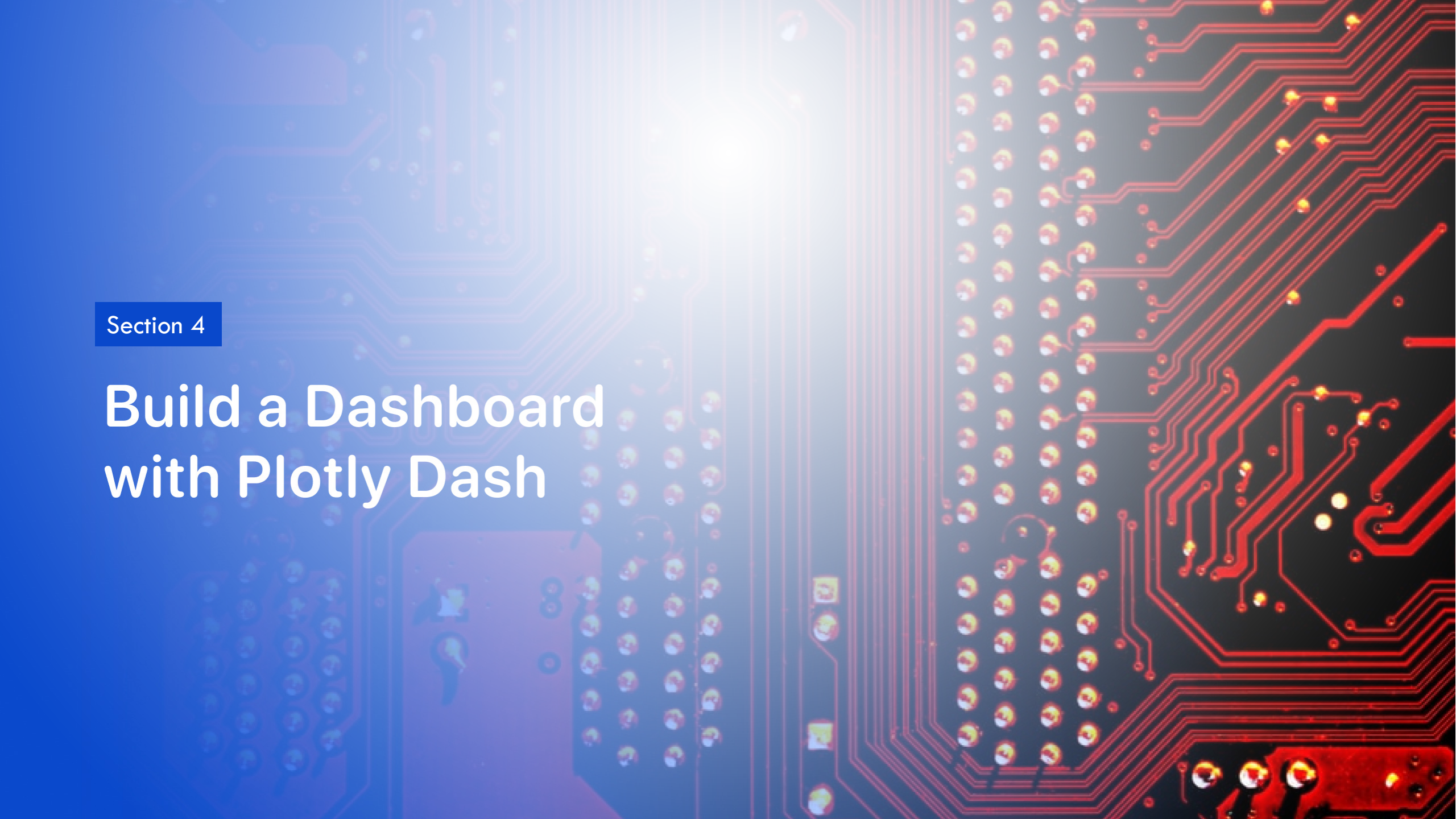
Distance From Launch Site

- Explanation:
  - From the visual analysis of the launch site CCAFS we can clearly see that it is:
    - Relative close to railway (1.06 km)
    - Relative close to highway (1.13 km)
    - Relative close to coastline (0.86 km)
  - Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

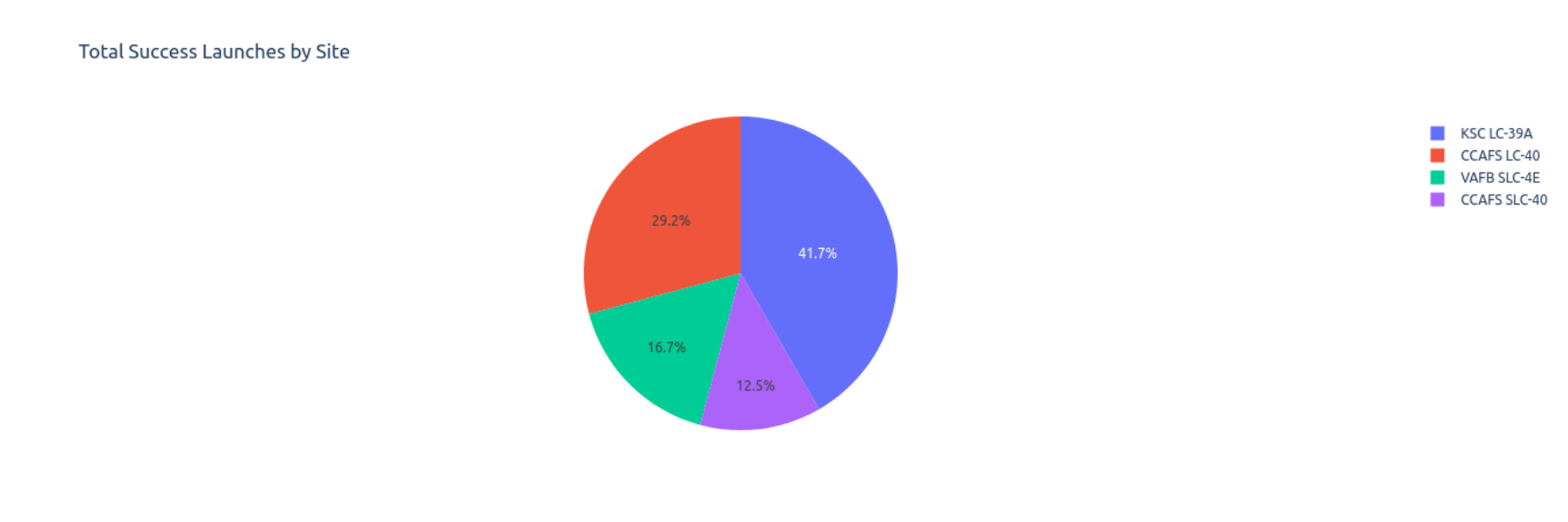


Section 4

Build a Dashboard with Plotly Dash



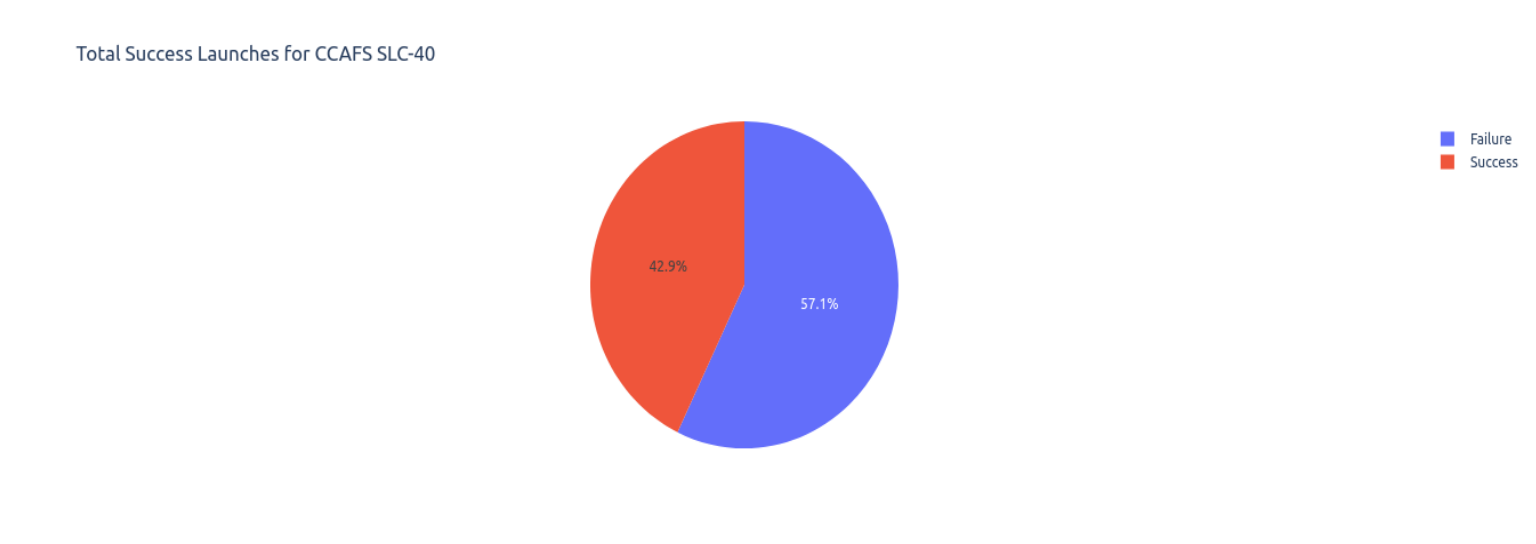
Launch Success Count for All Sites



Insight:

- ***The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.***

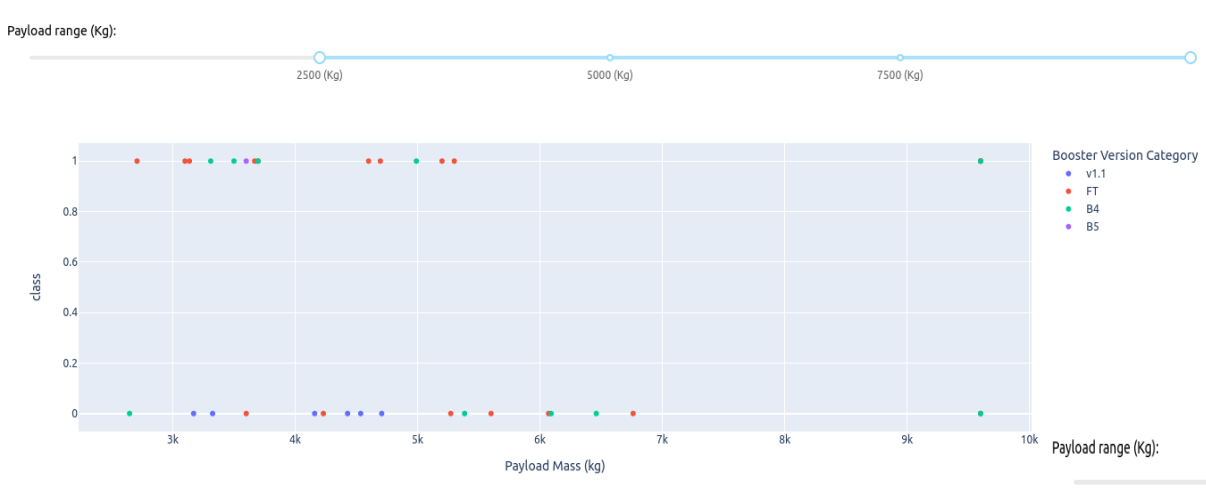
Launch Site With Highest Launch Success Ratio

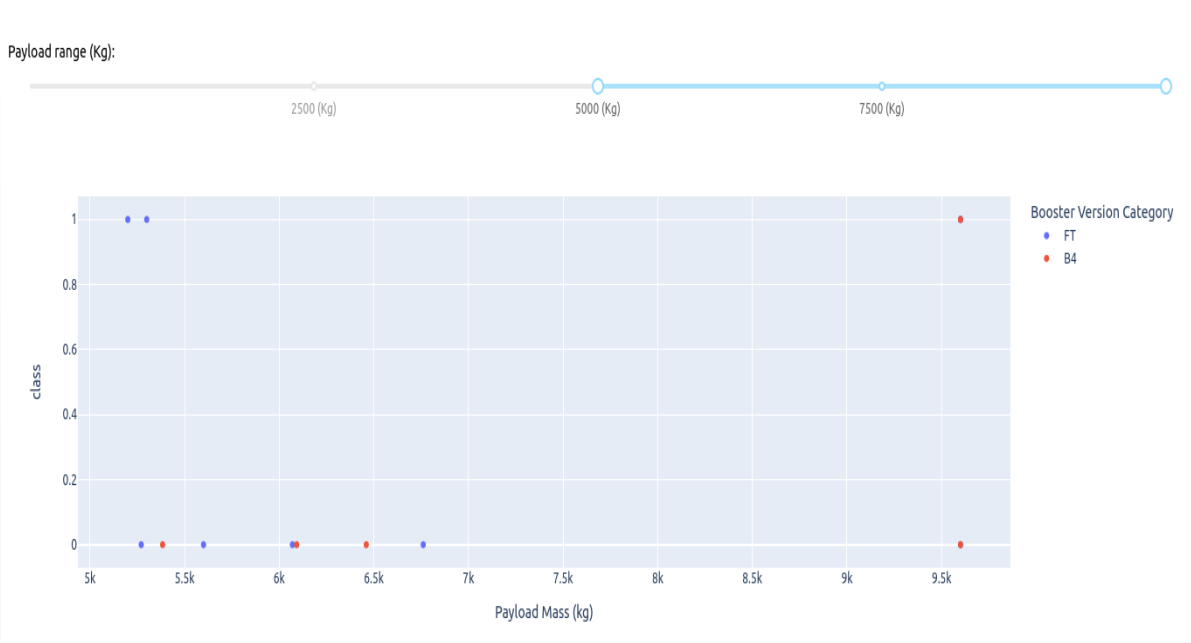


Explanation:

- *CCAFC SLC-40 is the launch site with the highest launch success ratio (42.9%).*

Launch Outcome vs. Payload Mass





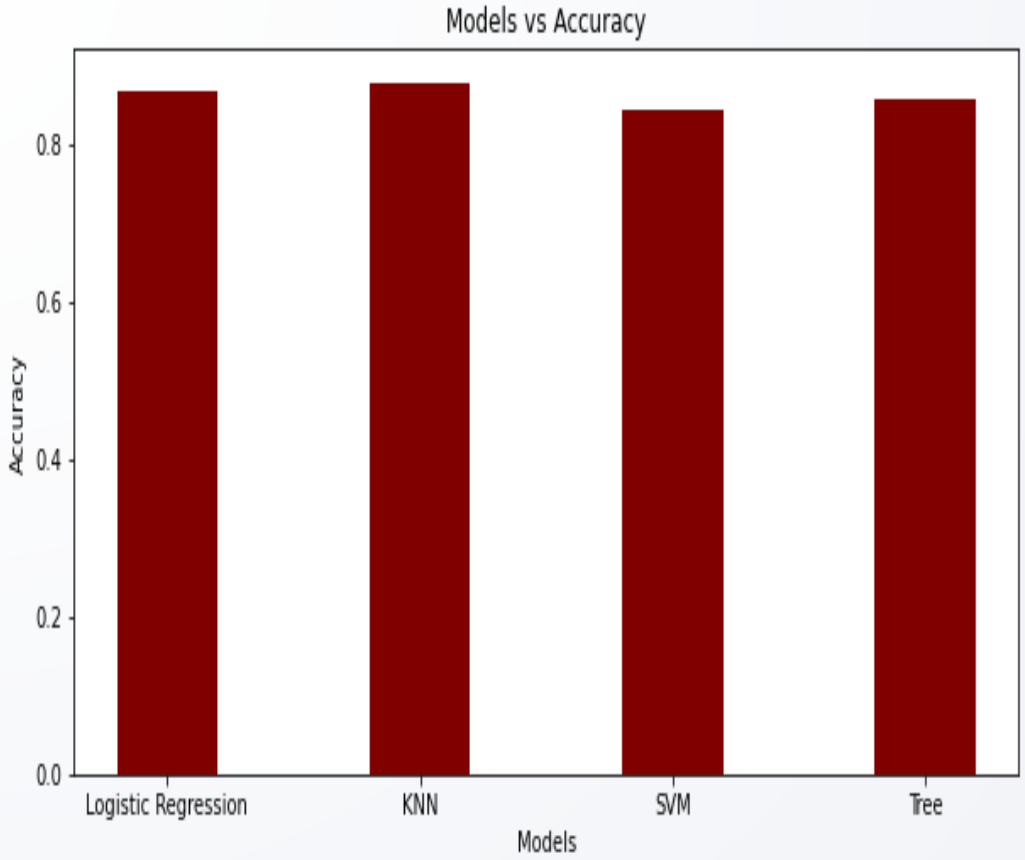
- **Explanation:**
  - **The charts show that payloads between 2000 and 5500 kg have the highest success rate.**

Section 5

Predictive Analysis (Classification)

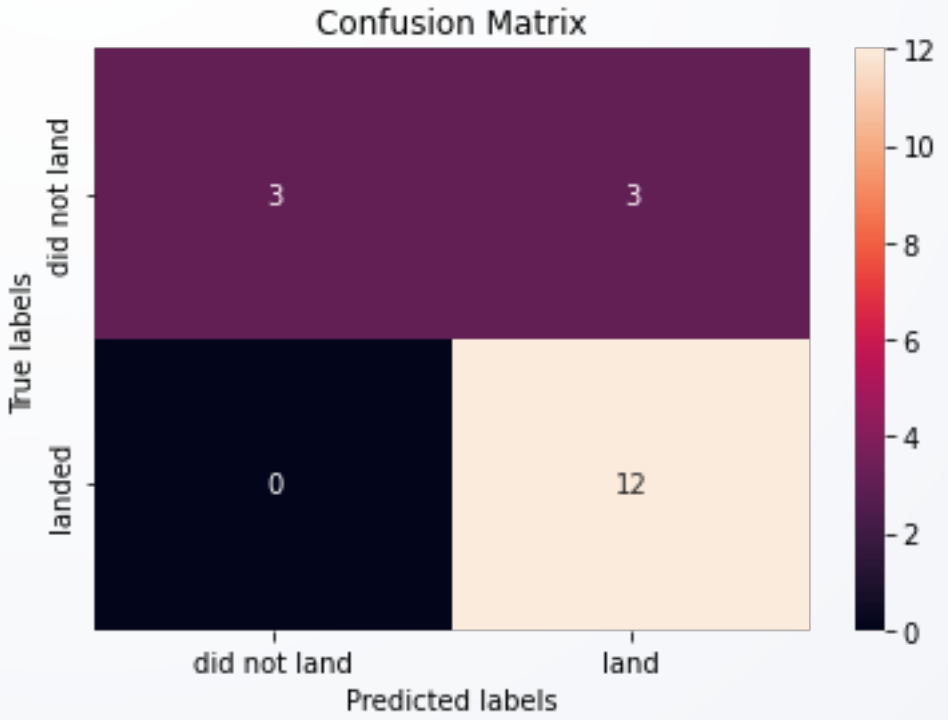
Classification Accuracy

- All models have the same accuracy on the test set at about 83%.
- However, the test size is too small to predict which model is more accurate.
- There is a need for more data to determine the best model.



Confusion Matrix

- *Since all models performed the same for the test set, the confusion matrix is the same across all models. The models predicted 12 successful landings when the true label was a successful landing.*
- *The models predicted 3 unsuccessful landings when the true label was an unsuccessful landing.*
- *The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over-predict successful landings.*



Conclusions

The main task for this capstone was to develop a machine learning model for Space Y company which wants to compete with Space X in the interstellar business. The main goal I aimed in building this model was to predict when stage 1 will successfully land to save ~$100 million USD. To that end, I used data from a public Space x API and web scraping Space X Wikipedia page and created data labels, and stored data in a DB2 SQL database. Additionally, I created a dashboard for visualization. Finally, I created a machine learning model with an accuracy of about 83%.

In conclusion, I hope that Allon mask of Space y can use this model to predict with relatively high accuracy whether a launch will have a successful stage 1 landing before launch to determine whether the launch should be made or not. Finally, I want to recommend here that in the future more data be collected and analyzed to better determine the best machine learning model as well as improve its accuracy.

Appendix

- **Github repository:**
- https://github.com/Ephrem2166/ibm_applied_data_science_capstone

- **SPECIAL THANKS TO:**
  - INSTRUCTORS
  - COURSERA
  - IBM

Thank you!