Winning Space Race with Data Science

Charan Vengatesh
09/12/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX is the most successful commercial space firm, making space travel affordable. The company advertises Falcon 9 rocket launches on its website for 62 million dollars; other providers charge up to 165 million dollars apiece; much of the savings is due to SpaceX's ability to reuse the first stage. As a result, if we can predict whether the first stage will land, we can estimate the cost of a launch. We will forecast if SpaceX will reuse the first stage using public data and machine learning techniques.
- Problems you want to find answers
  - How do variables like payload mass, launch site, number of flights, and orbits affect first stage landing success?
  - Is the rate of successful landings increasing over time?
  - In this scenario, what is the optimal algorithm for binary classification?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- 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 data was gathered through a combination of API queries to SpaceX's REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.
- In order to obtain complete information about the launches for a more extensive study, we had to use both of these data collection methods.

Data Collection - SpaceX API

- Requesting rocket launch data from SpaceX API
- Decoding the response content using .json() and turning it into a dataframe using .json_normalize()
- Requesting needed information about the launches from SpaceX API by applying custom functions
- Constructing data we have obtained into a dictionary
- Creating a dataframe from the dictionary
- Filtering the dataframe to only include Falcon 9 launches
- Replacing missing values of Payload Mass column with calculated .mean() for this column
- Exporting the data to CSV

Data Collection – Scraping

Requesting	Requesting Falcon 9 launch data from Wikipedia
Creating	Creating a BeautifulSoup object from the HTML response
Extracting	Extracting all column names from the HTML table header
Collecting	Collecting the data by parsing HTML tables
Constructing	Constructing data we have obtained into a dictionary
Creating	Creating a dataframe from the dictionary
Exporting	Exporting the data to CSV

Data Wrangling

There are multiple examples in the data set where the booster did not successfully land. A landing may have been attempted but failed due to an accident; for example, True Ocean indicates that the mission outcome was successfully landed to a specific location of the ocean, whereas False Ocean indicates that the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS indicates that the mission has successfully landed on a ground pad. False RTLS indicates that the mission failed to land on a ground pad.True ASDS indicates that the mission result was successfully landed on a drone ship. False ASDS indicates that the mission conclusion was an unsuccessful landing on a drone ship.

- Perform exploratory Data Analysis and determine Training Labels
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Exporting the data to CSV

EDA with Data Visualization

- 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
- Scatter plots show the relationship between variables. If a relationship exists, they could be used in 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).

EDA with SQL

Performed SQL queries:

EDA with SQL

- 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

Build an Interactive Map with Folium

Markers of all Launch Sites:	• Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location. • Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to 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 Cluster 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.

Build a Dashboard with Plotly Dash

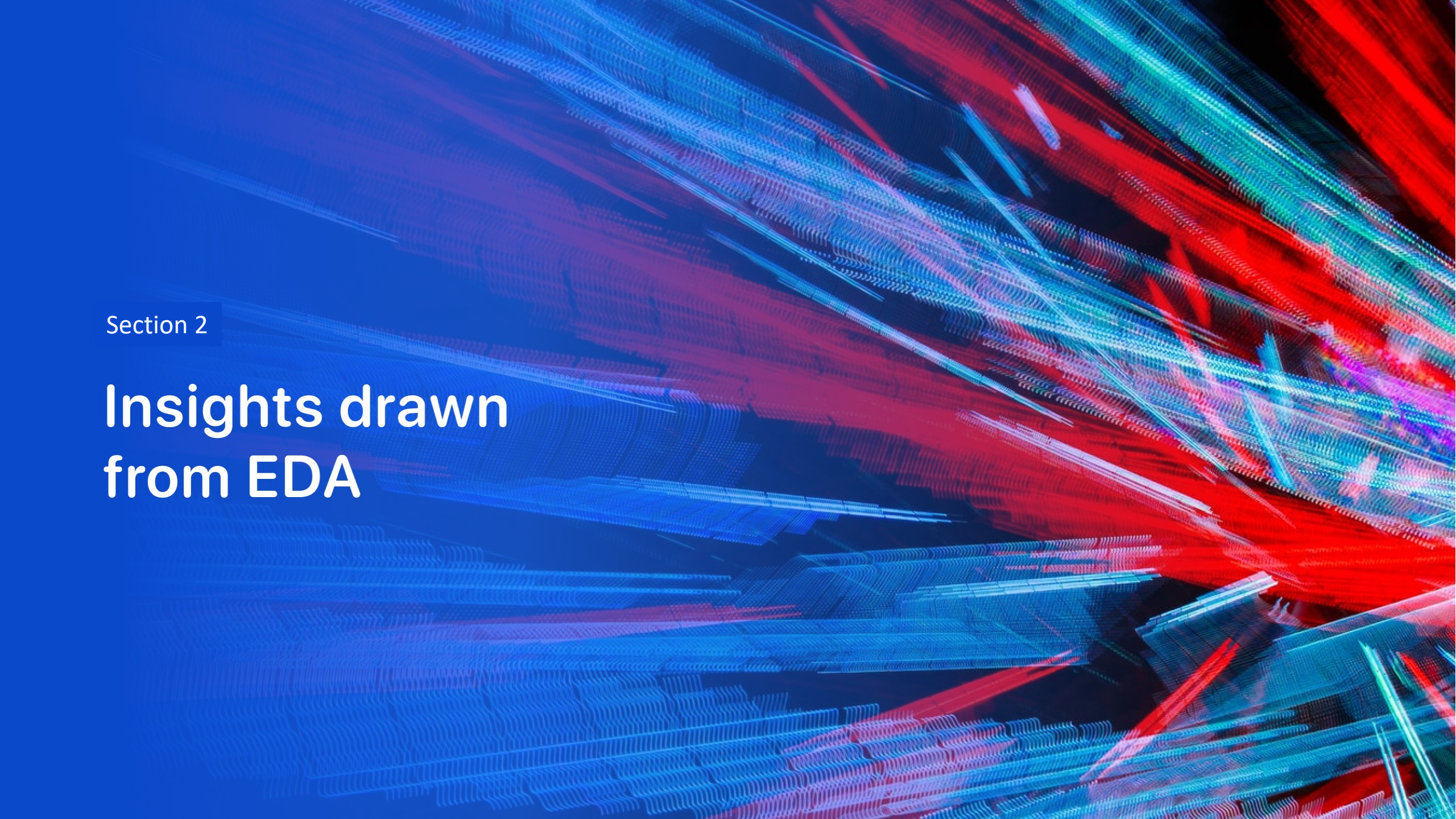
 Launch Sites Dropdown List:	Added a dropdown list to enable Launch Site selection.
 Pie Chart showing Success Launches (All Sites/Certain Site):	Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
 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.

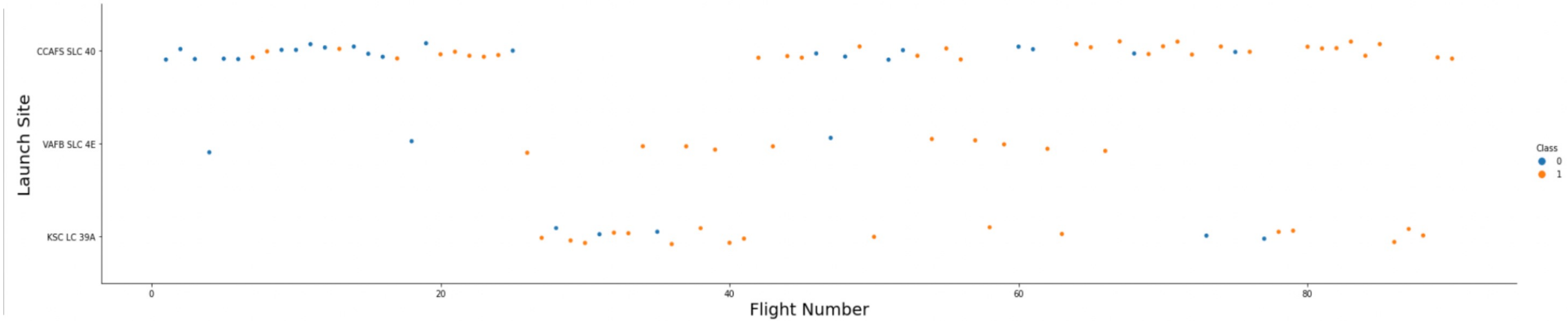
Predictive Analysis (Classification)

Creating	Creating a NumPy array from the column “Class” in data
Standardizing	Standardizing the data with StandardScaler, then fitting and transforming it
Splitting	Splitting the data into training and testing sets with train_test_split function
Creating	Creating a GridSearchCV object with cv = 10 to find the best parameters
Applying	Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
Calculating	Calculating the accuracy on the test data using the method .score() for all models
Examining	Examining the confusion matrix for all models
Finding	Finding the method performs best by examining the Jaccard_score and F1_score metrics

Results

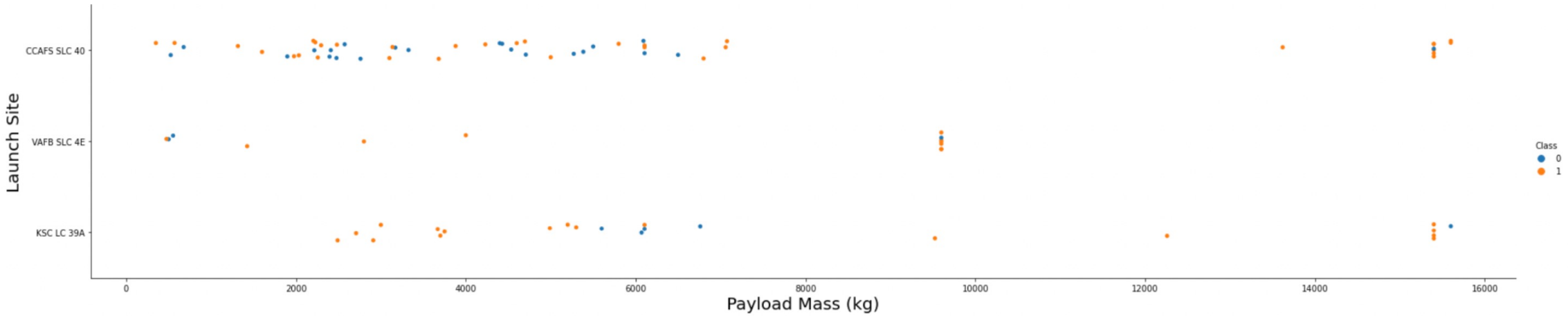
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

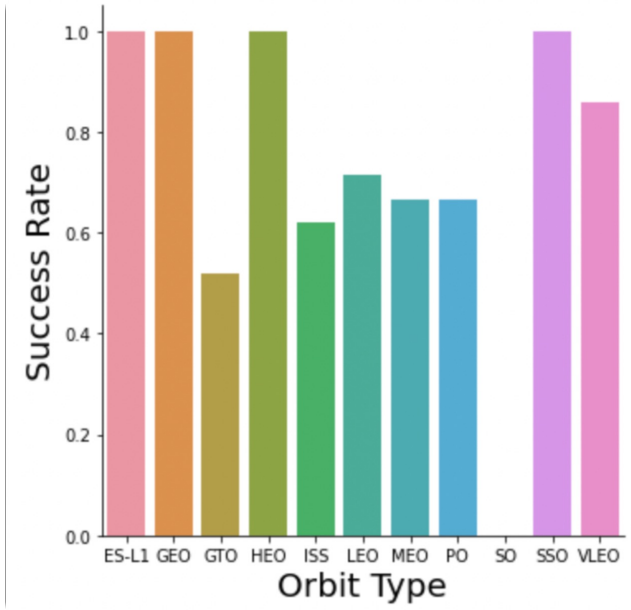
- The first flights all failed, but the last ones all succeeded.
- The CCAFS SLC 40 launch location accounts for over half of all launches.
- Success rates are greater at VAFB SLC 4E and KSC LC 39A.
- It can be assumed that each new launch will have a higher success percentage.



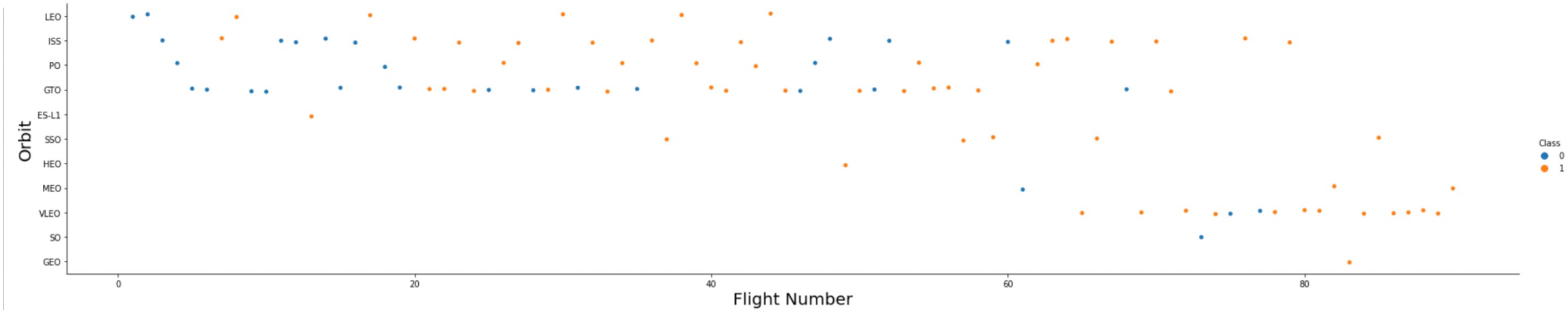
Payload vs. Launch Site

- For each launch site, the bigger the payload mass, the higher the success percentage;
- The majority of launches with payload masses greater than 7000 kg were successful.
- KSC LC 39A also has a 100% success rate for payloads weighing less than 5500 kg.

Success Rate vs. Orbit Type

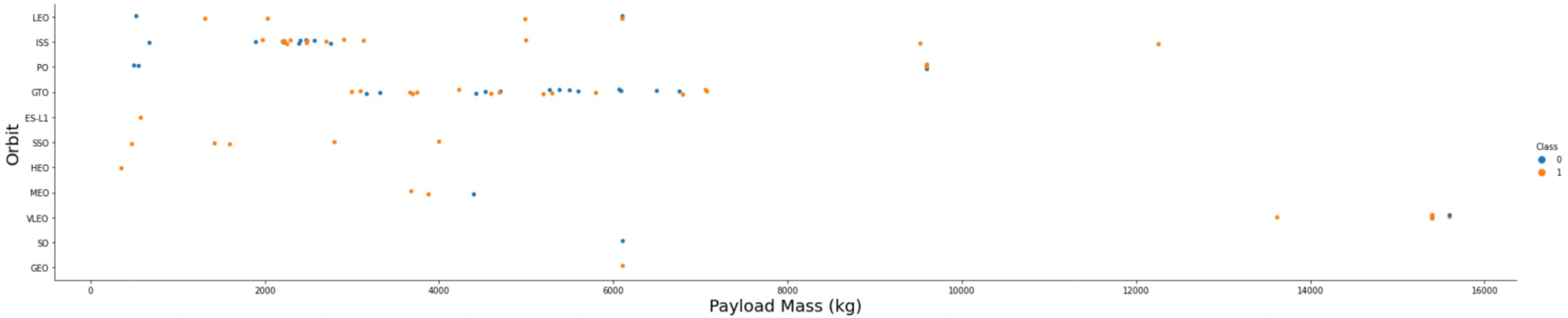


- Orbits with a perfect success rate: ES-L1, GEO, HEO, and SSO
- Orbits having a success rate of 0%: - SO
- Orbits with a success percentage of 50% to 85%:
- GTO, ISS, LEO, MEO, and PO



Flight Number vs. Orbit Type

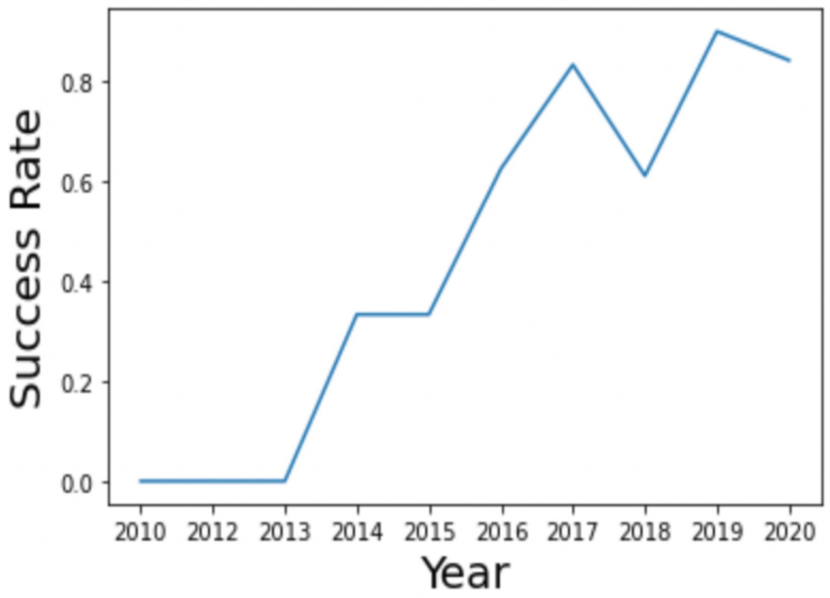
- In LEO orbit, success appears to be connected to the number of flights; however, in GTO orbit, there appears to be no relationship between flight number.



Payload vs. Orbit Type

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

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020.

All Launch Site Names

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

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

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

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Boosters Carried Maximum Payload

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

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

2015 Launch Records

- List of 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

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

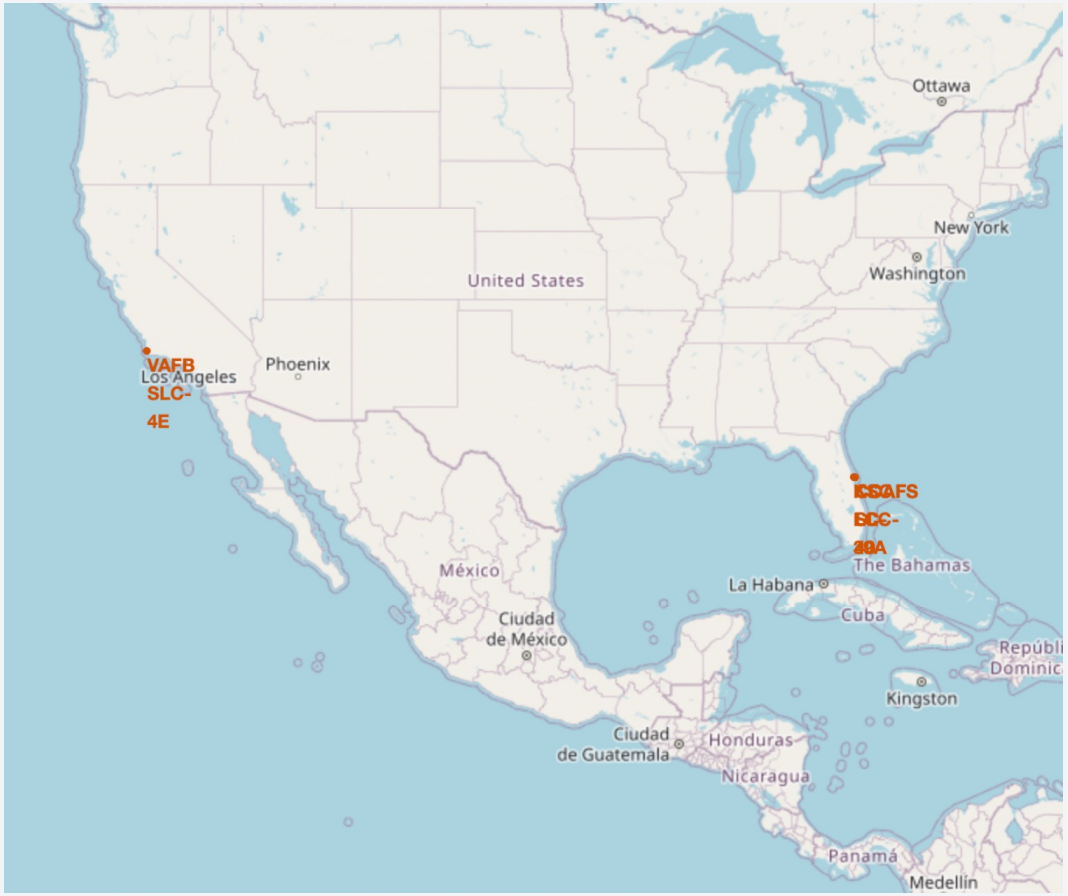
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

landing__outcome	count_outcomes
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

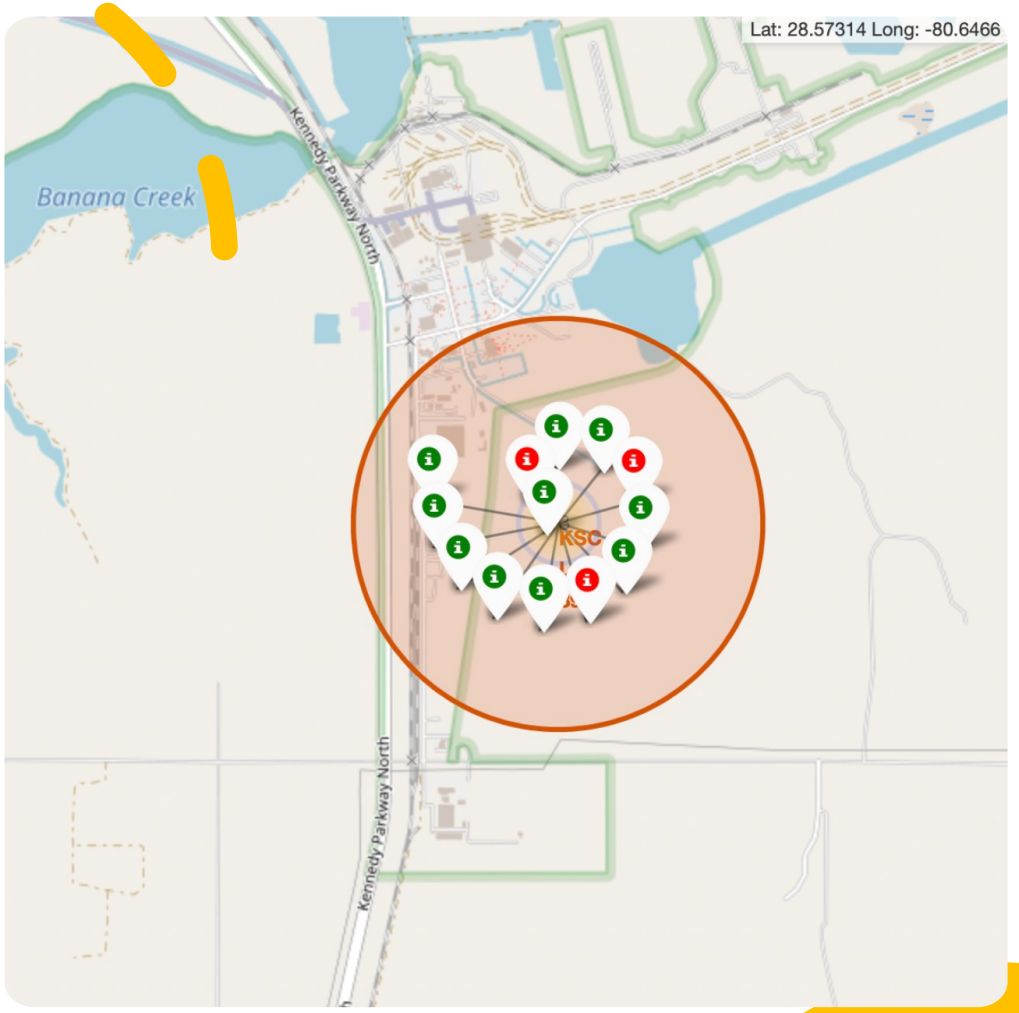
Location markers for all launch sites on a worldwide map

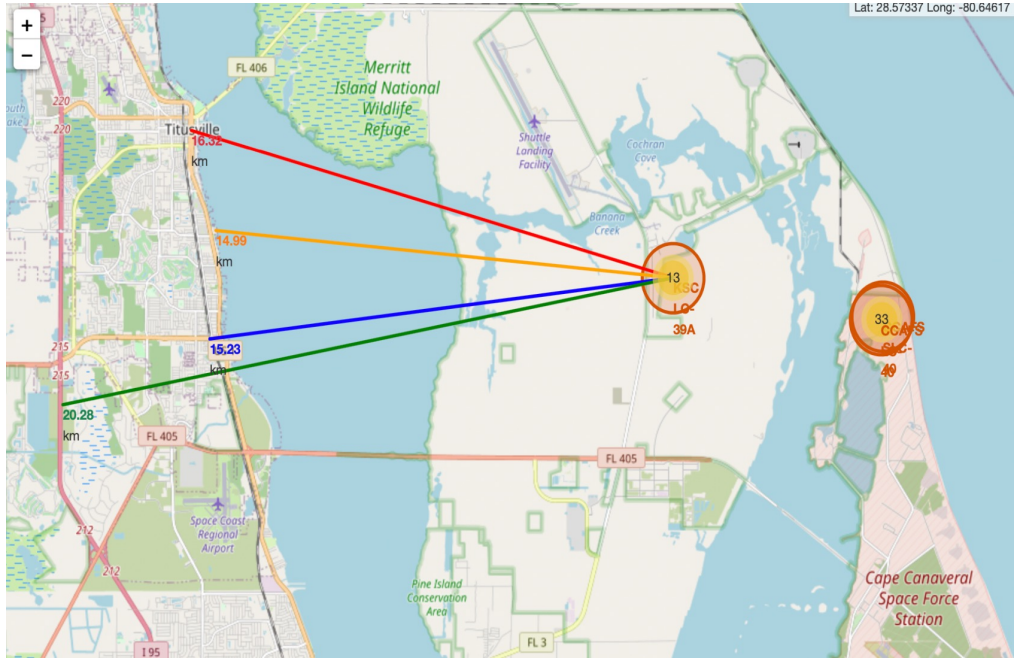
- The majority of launch sites are near the Equator. The ground moves quicker at the equator than everywhere else on the planet's surface. Anything on the Earth's surface at the equator is already traveling at 1670 km/h. When a ship is launched from the equator, it travels into space while concurrently travelling around the Earth at the same speed it was before the launch. This is due to inertia. This speed will assist the spaceship in maintaining a sufficient speed to remain in orbit.
- All launch locations are fairly close to the coast, and launching rockets towards the ocean reduces the possibility of debris falling or exploding near humans.



Colour-labeled launch records on the map

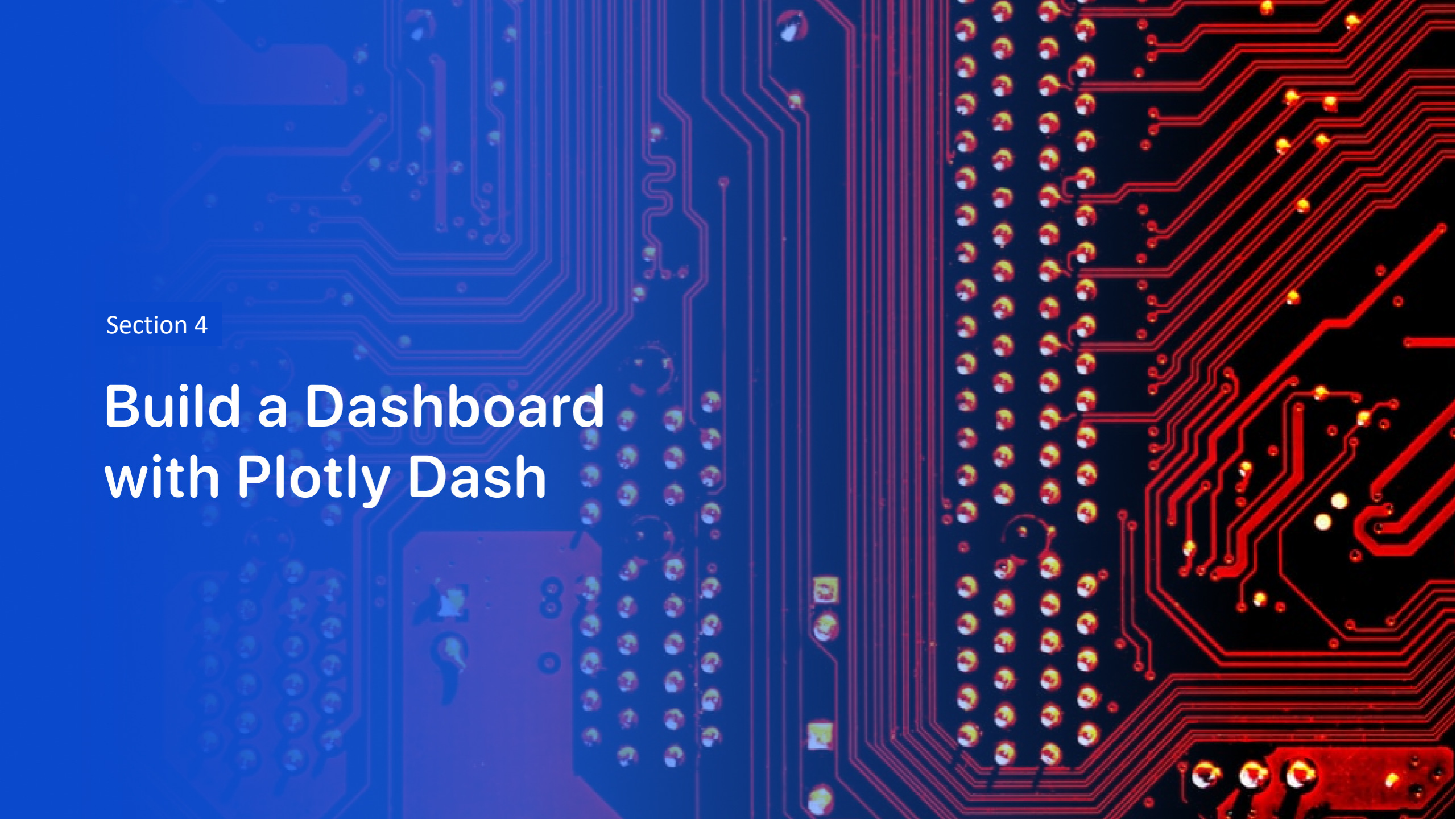
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- • Launch Site KSC LC-39A has a very high Success Rate.





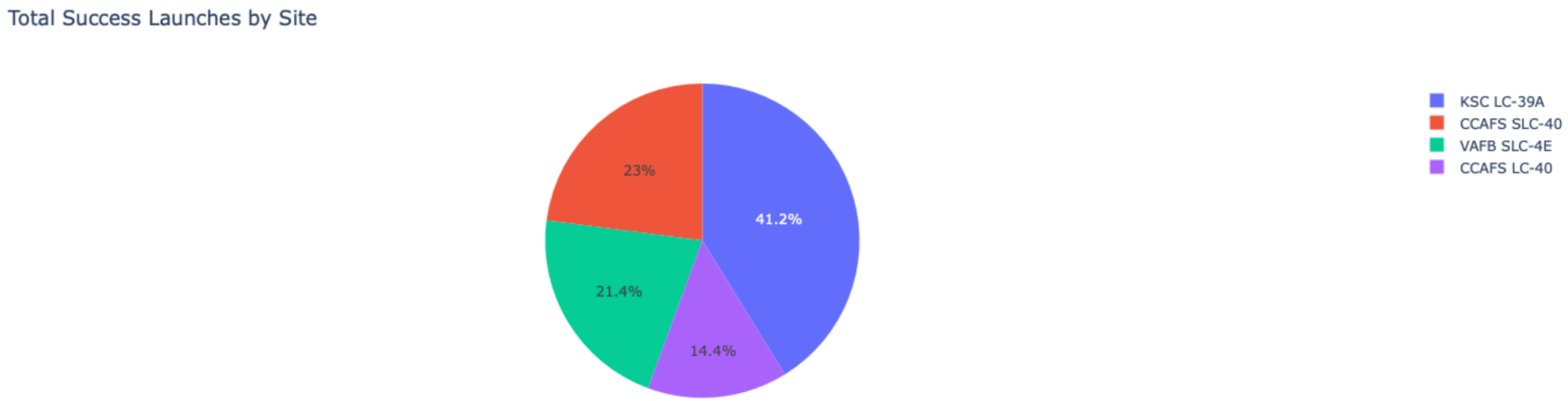
Distance from the launch site KSC LC-39A to its proximities

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km) - relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 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

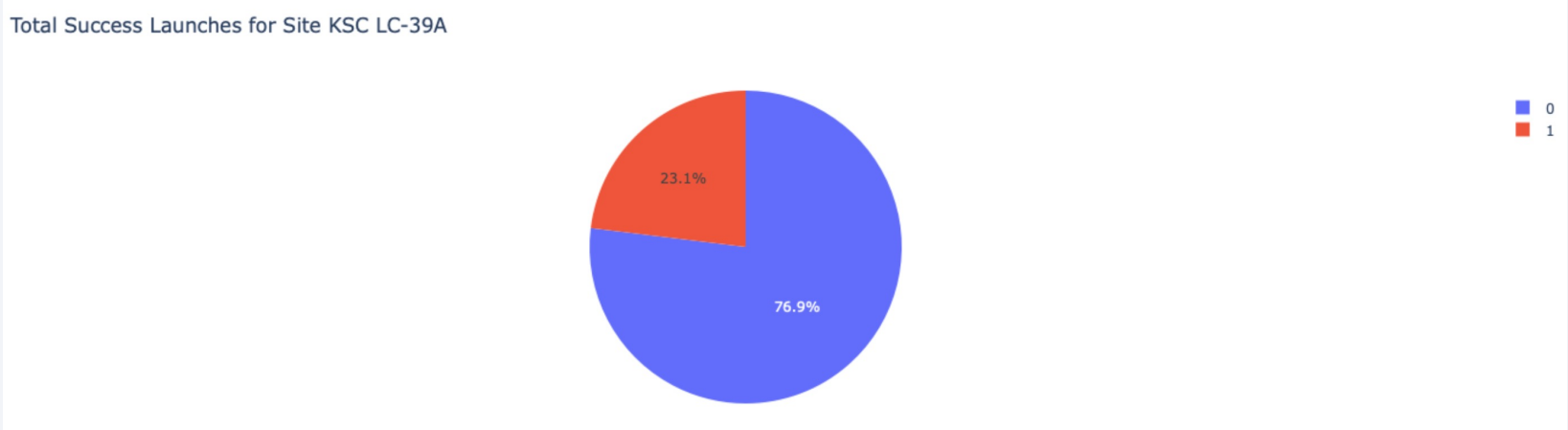
Launch success count for all sites

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



Launch site with highest launch success ratio

- Replace <KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



Payload Mass vs. Launch Outcome for all sites

- According to the figures, payloads weighing between 2000 and 5500 kg have the highest success rate.

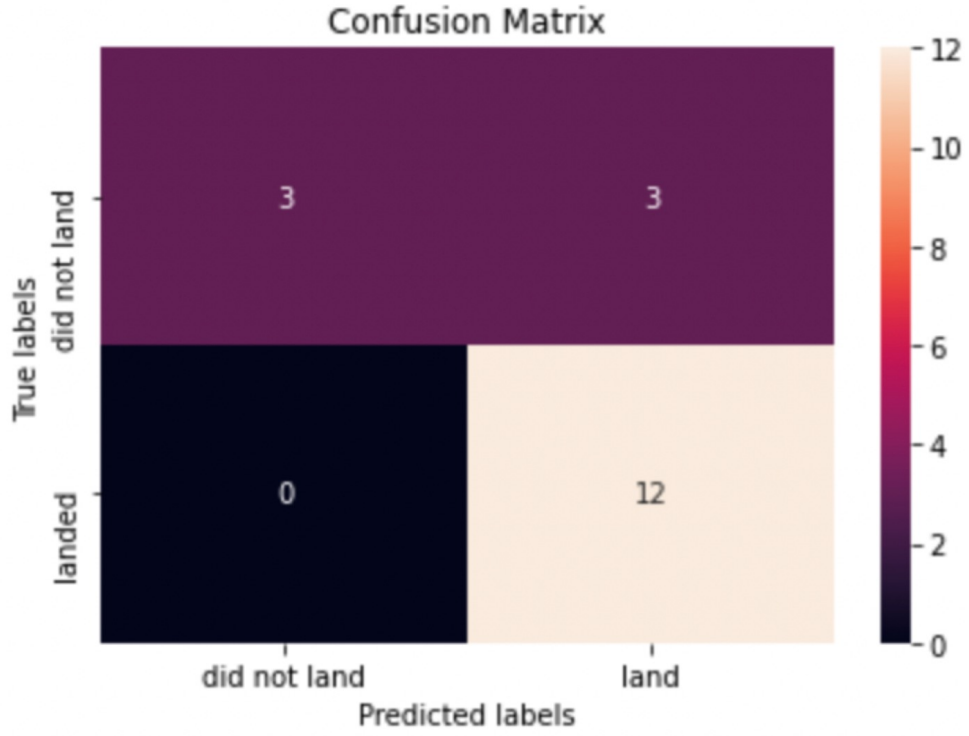
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- We cannot confirm which strategy performs best based on the Test Set results.
- Due to the short test sample size (18 samples), the same Test Set scores may occur. As a result, we tested all techniques using the entire Dataset.
- The overall Dataset scores confirm that the Decision Tree Model is the best model. This model not only has the best scores, but it also has the highest accuracy.

Confusion Matrix



- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- The best algorithm for this dataset is the Decision Tree Model.
- Launches with a low payload mass outperform launches with a higher payload mass.
- The majority of launch sites are near to the Equator line, and all of them are close to the coast.
- The success rate of launches has increased with time.
- KSC LC-39A has the best success rate of any site's launches.
- ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.

Thank you!