IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ALI ASHAR SHAFIQUE
10 OCTOBER 2024

Outline

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

Executive Summary

The Capstone Project focusing on SpaceX aims to analyze and understand the success factors behind SpaceX's mission outcomes, specifically its launch success rates. This project involves collecting and processing historical data related to SpaceX launches, using various data science techniques such as data visualization, machine learning, and statistical analysis.

The core objective is to predict the likelihood of successful launches, which is crucial for improving the cost-effectiveness and reliability of future missions. The project covers several phases:

1. **Data Collection**: Historical data on SpaceX launches, including details like mission date, launch site, payload, rocket type, orbit, and mission outcome, are gathered from various sources, including SpaceX's own databases and external datasets such as NASA archives.
2. **Data Wrangling and Cleaning**: The raw data is preprocessed to handle missing values, remove inconsistencies, and ensure accuracy. The data is then structured for further analysis.
3. **Exploratory Data Analysis (EDA)**: Visualizations and statistical summaries are generated to gain insights into trends and patterns related to launch outcomes. This helps in identifying key factors that may influence launch success, such as payload mass, weather conditions, and rocket type.
4. **Machine Learning Modeling**: Predictive models such as Logistic Regression, Decision Trees, and Support Vector Machines (SVM) are applied to the cleaned data to predict future launch successes. The models are trained and validated using past mission data, with performance metrics like accuracy and precision being used to fine-tune the models.
5. **Insights and Recommendations**: Based on the analysis, insights are derived to help SpaceX optimize its launch processes. For example, findings may suggest that certain rocket types or payload sizes are more likely to result in successful launches, providing actionable recommendations for SpaceX's future operations.

Introduction

Project Background and Context

Space exploration has become increasingly important with advancements in technology and the growing interest in commercial space travel. Among the leading companies in this field is SpaceX, which has revolutionized space missions by developing reusable rocket technology, significantly lowering the costs of space exploration. However, with the complexity of each mission, there are various factors that can influence the success or failure of a launch.

This project seeks to explore SpaceX's historical launch data using data science techniques to understand the variables that impact launch outcomes. By applying machine learning models, the project aims to predict the success rates of future SpaceX missions. Understanding these factors is crucial for improving operational decision-making, ensuring safety, and enhancing mission success.

Problems You Want to Find Answers To

The primary goal of this project is to address the following questions:

- What key factors contribute to the success or failure of SpaceX launches?
- How can machine learning algorithms predict the outcomes of future launches?

Section 1

Methodology

Methodology

- Data Collection using:
  - Space X API , Web scraping
- Exploratory Data Analysis (EDA) using:
  - Pandas, NumPy and SQL
- Data visualization using:
  - Matplotlib, Seaborn, folium and Dash
- Machine Learning Modeling using:
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

Data Collection

The data for this project is sourced primarily through APIs or web scraping from public repositories and databases that provide information on SpaceX launches.

Data Collection – SpaceX API

- The data for this project is sourced from the SpaceX API:
- ((https://api.spacexdata.com/v4/rockets/))
- This API provides detailed information on various SpaceX rocket launches, and for this analysis, the data is filtered to include only Falcon 9 rocket launches.
- Missing values in the dataset are handled by replacing them with the mean value of their respective columns.
- The final dataset consists of 90 instances (rows) and 17 features (columns), ready for further analysis.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection - Scraping

- This data is scraped from:

(https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

This websites only contain data about Falcons 9

It ended with 121 rows and 11 columns

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- The data is later processed so that there are no more missing entities, and categorical features are encoded using one-hot encoding
- An extra column “Class” also has been added to data frame
- It ended with 90 row and 83 columns

EDA with Data Visualization

- **Matplotlib and Seaborn**
  - Functions from the Matplotlib and Seaborn libraries are used to visualize the data through
  - scatterplots, bar charts, and line charts.
  - The plots and charts are used to understand more about the relationships between several features, such as:
  - The relationship between flight number and launch site
  - The relationship between payload mass and launch site
  - The relationship between success rate and orbit type

EDA with SQL

- SQL
  - The data is queried using SQL to answer several questions about the data such as:
  - The names of the unique launch sites in the space mission
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1

Build an Interactive Map with Folium

- **Folium**
  - Functions from the Folium libraries are used to visualize the data through interactive maps.
  - The Folium library is used to:
  - Mark all launch sites on a map
  - Mark the succeeded launches and failed launches for each site on the map
  - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

Build a Dashboard with Plotly Dash

- Dash
  - Functions from Dash are used to generate an interactive site where we can toggle the input
  - using a dropdown menu and a range slider.
  - Using a pie chart and a scatterplot, the interactive site shows:
  - The total success launches from each launch site
  - The correlation between payload mass and mission outcome (success or failure) for each launch site

Predictive Analysis (Classification)

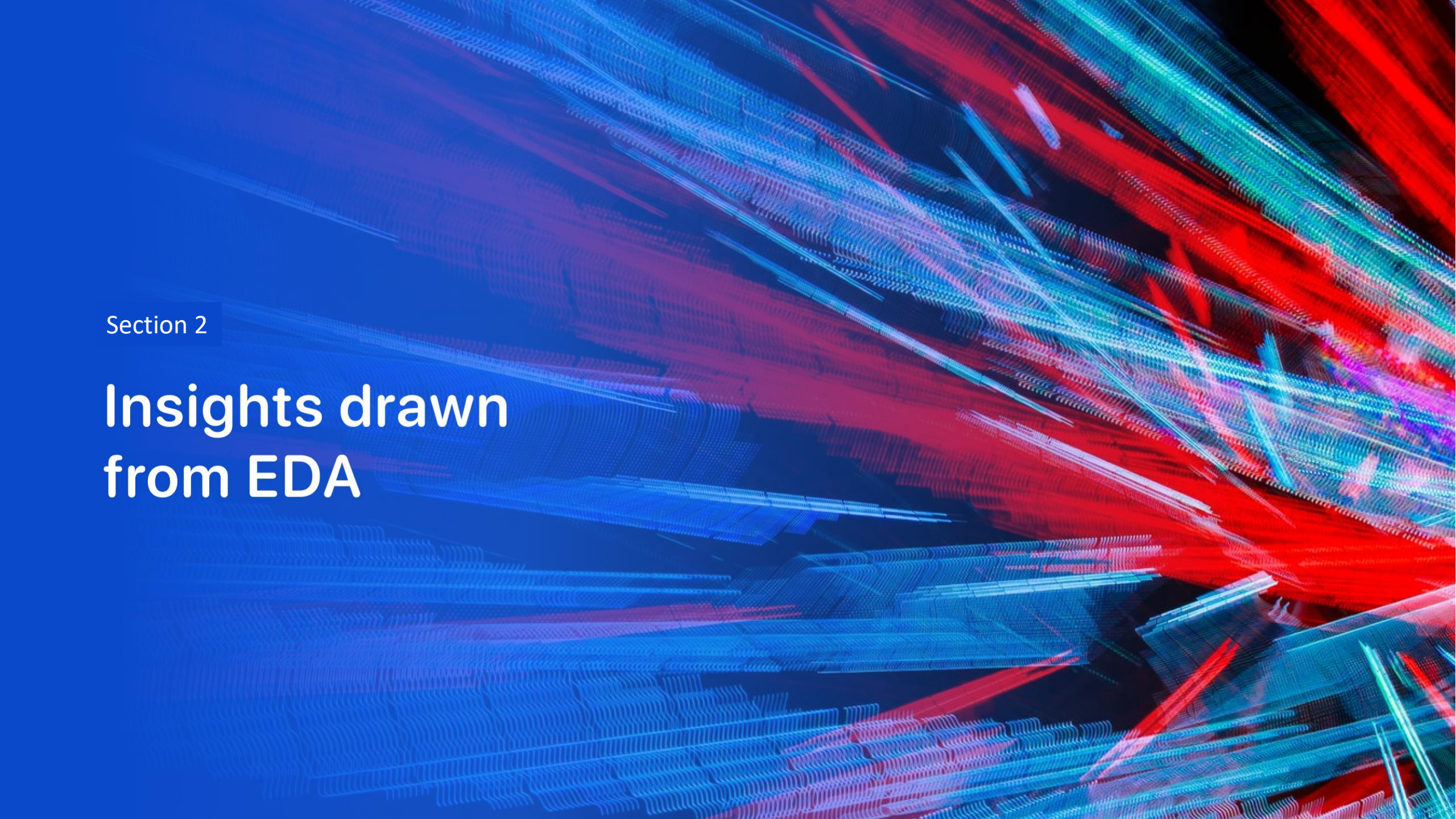
Functions from the Scikit-learn library are used to create our machine learning models.

- The machine learning prediction phase include the following steps:
- Standardizing the data
- Splitting the data into training and test data
- Creating machine learning models, which include:
- Logistic regression
- Support vector machine (SVM)
- Decision tree
- K nearest neighbors (KNN)
- Fit the models on the training set
- Find the best combination of hyperparameters for each model
- Evaluate the models based on their accuracy scores and confusion matrix

Results

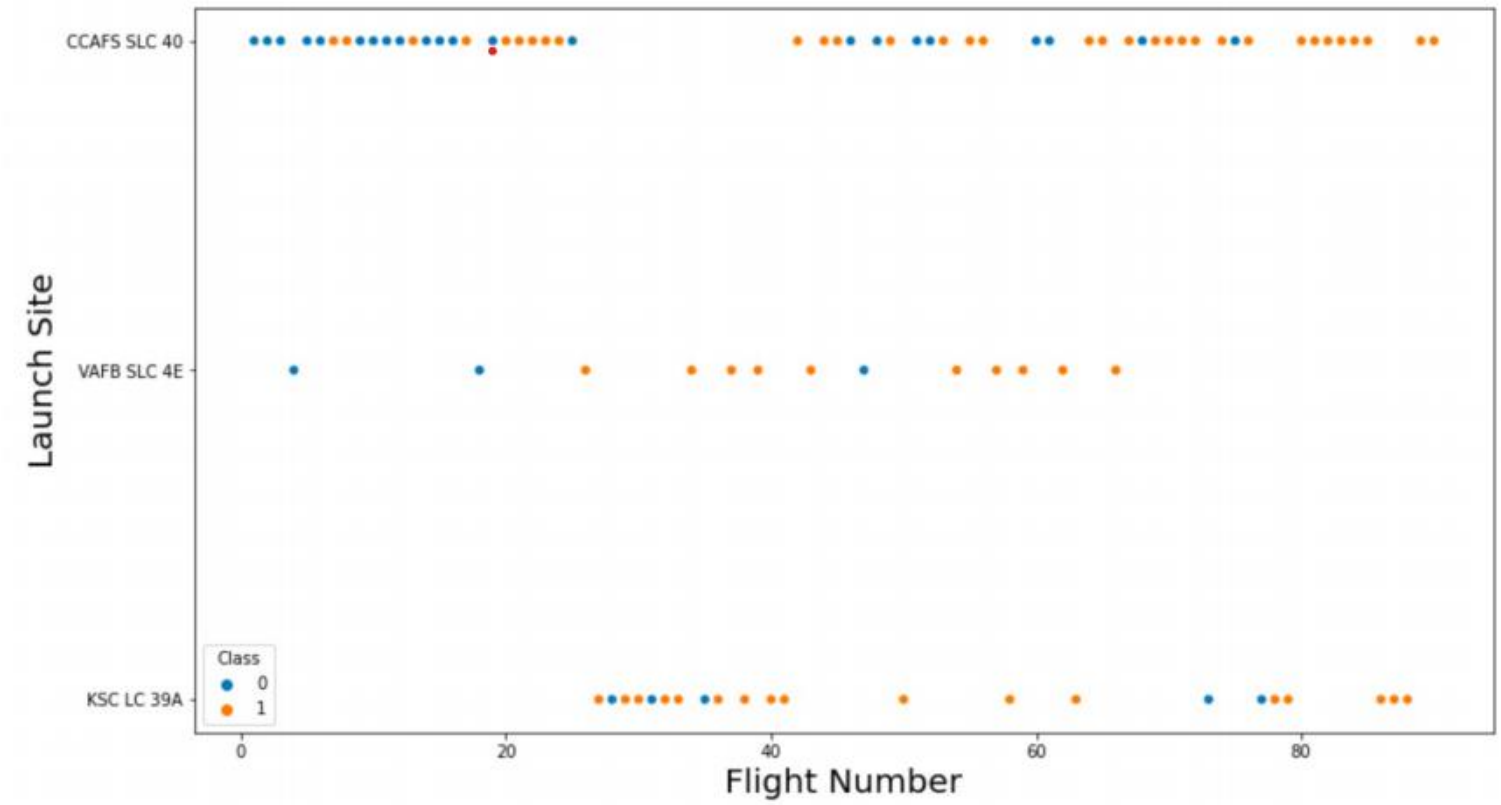
The results are split into 5 sections:

- SQL (EDA with SQL)
- Matplotlib and Seaborn (EDA with Visualization)
- Folium
- Dash
- Predictive Analysis
- In all of the graphs that follow, class 0 represents a failed launch outcome while
- class 1 represents a successful launch outcome


Section 2
Insights drawn from EDA

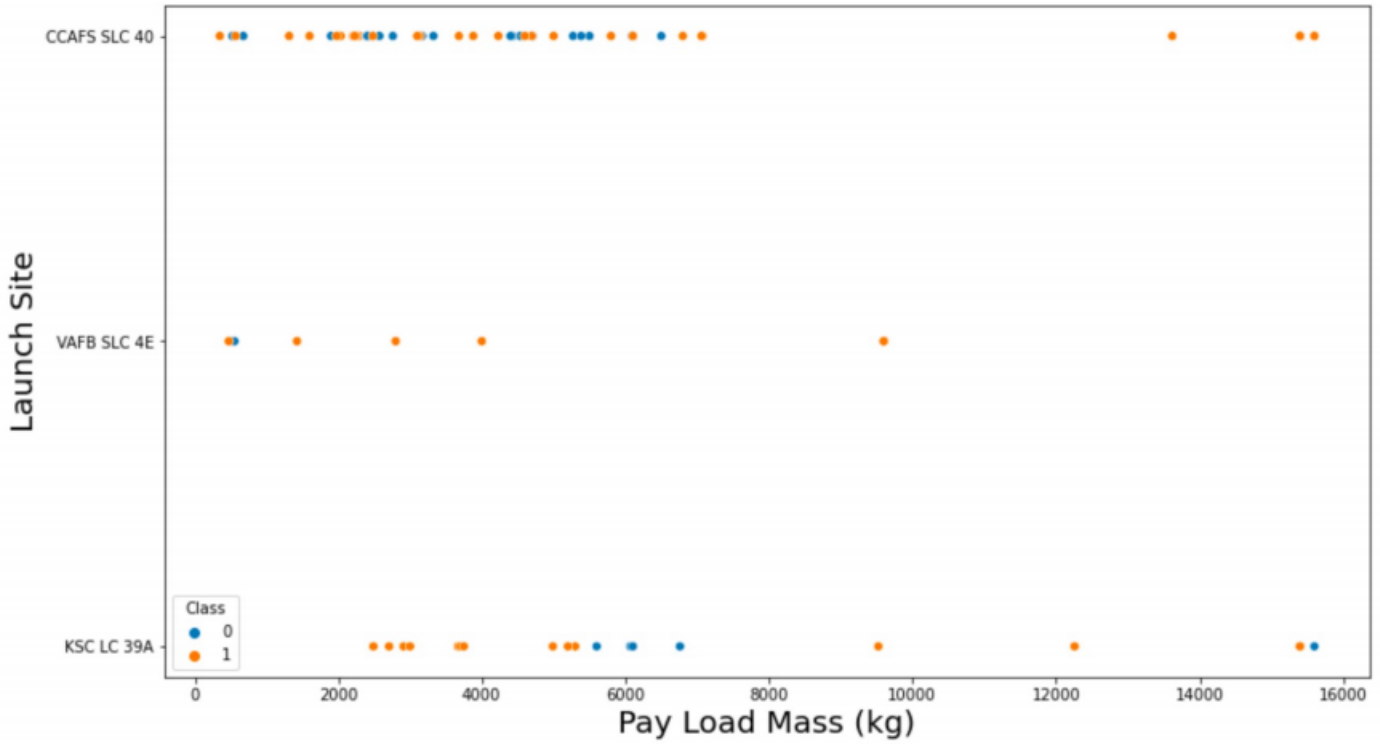
Flight Number vs. Launch Site

- The relationship between flight number and launch site



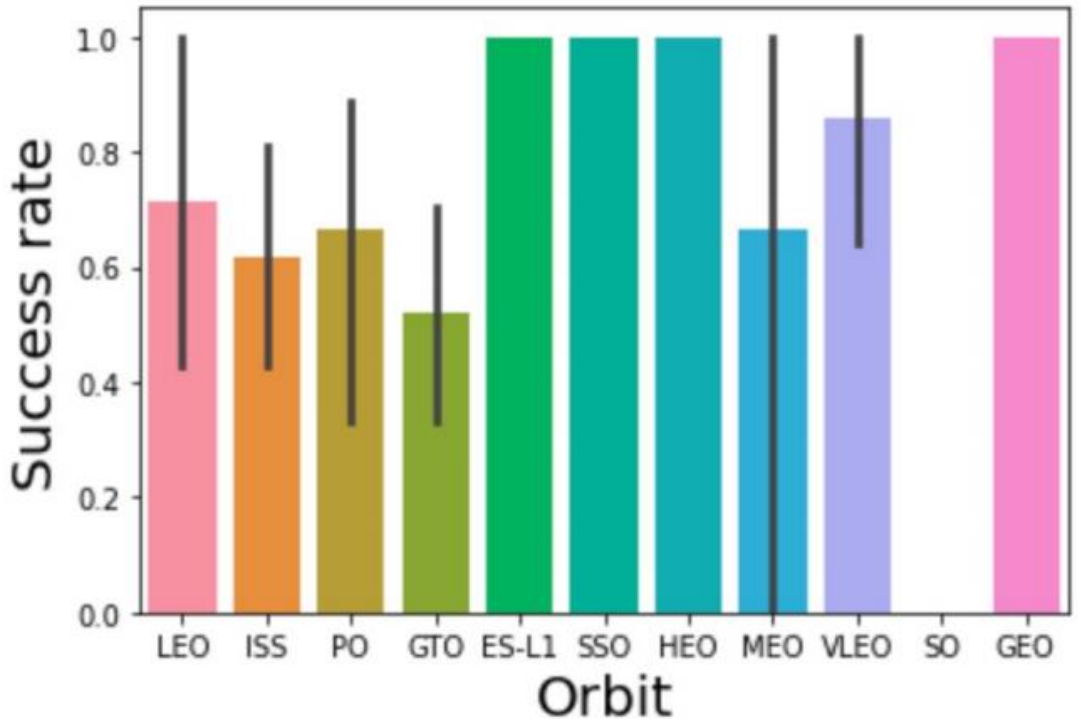
Payload vs. Launch Site

- The relationship between payload mass and launch site



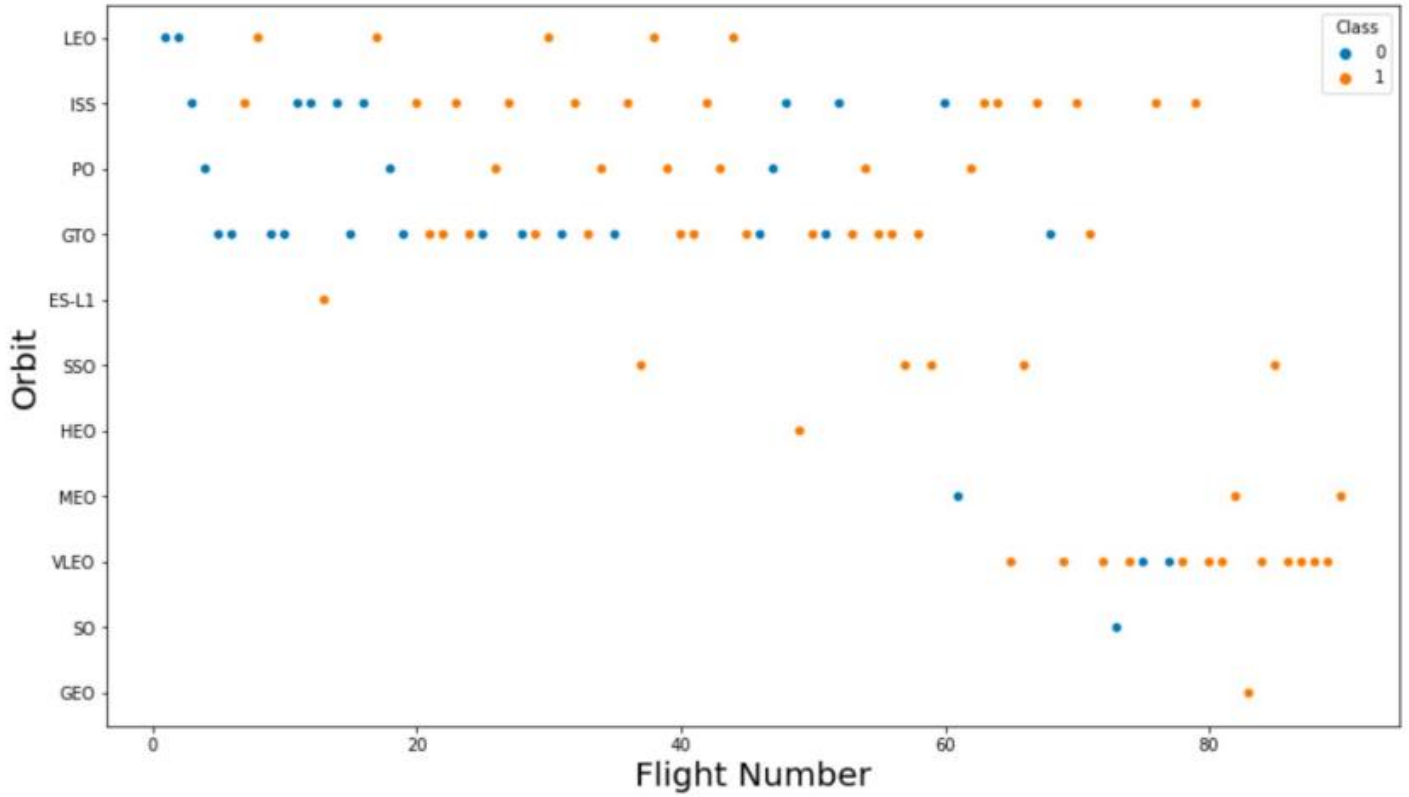
Success Rate vs. Orbit Type

- The relationship between success rate and orbit type



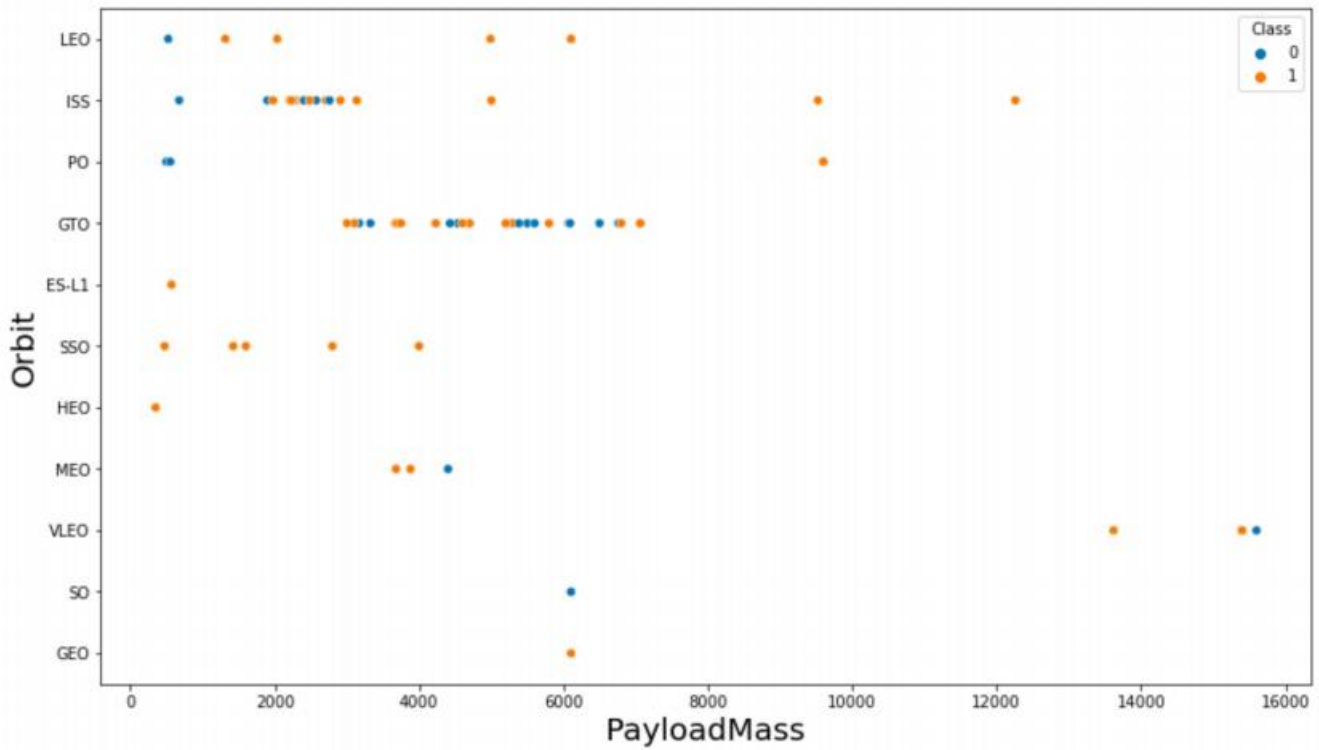
Flight Number vs. Orbit Type

- The relationship between flight number and orbit type



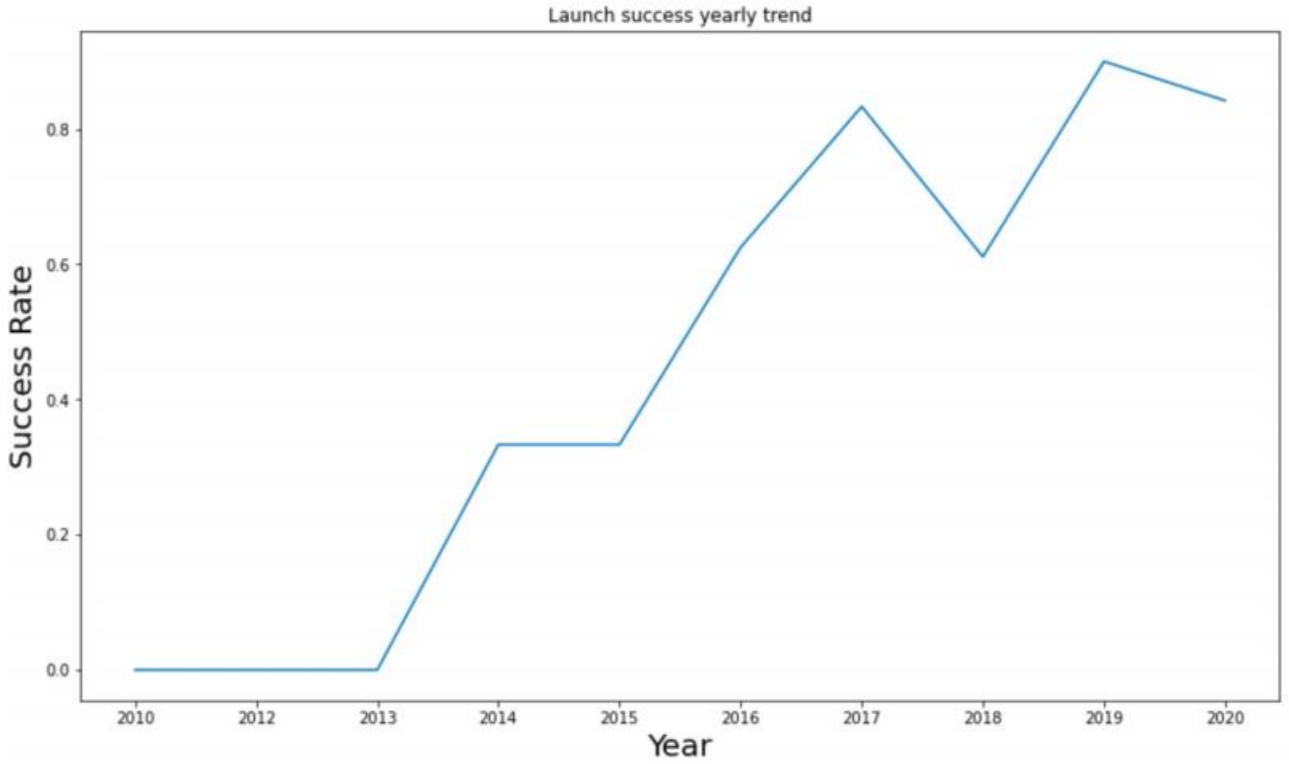
Payload vs. Orbit Type

- The relationship between payload mass and orbit type



Launch Success Yearly Trend

- The launch success yearly trend



All Launch Site Names

```
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'

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.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

```
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

```
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

```
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.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
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.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

Total Number of Successful and Failure Mission Outcomes

```
In [10]: %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.
```

Out[10]:

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

Boosters Carried Maximum Payload

```
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

2015 Launch Records

```
In [12]: %%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.
```

Out[12]:

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)

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

In [13]:
```
%%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.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

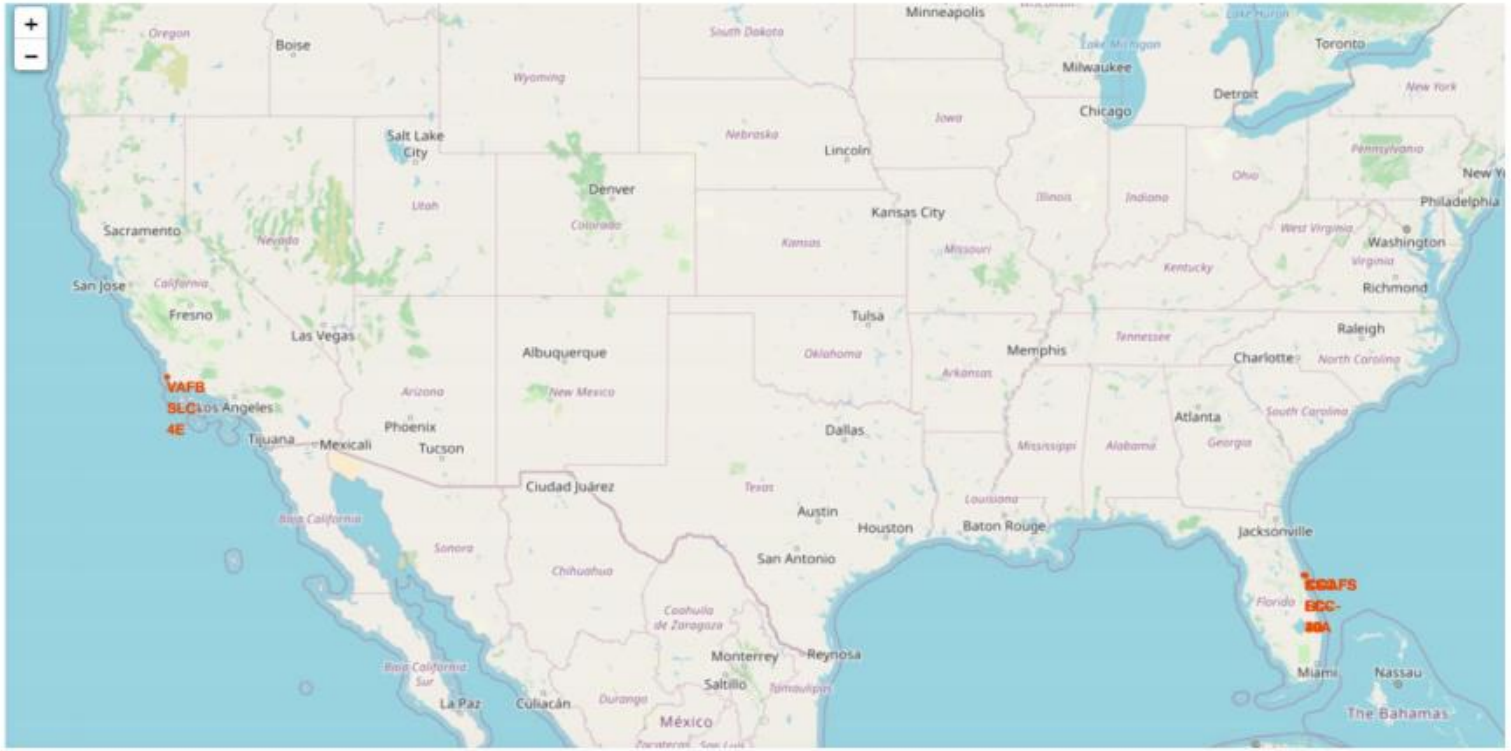
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

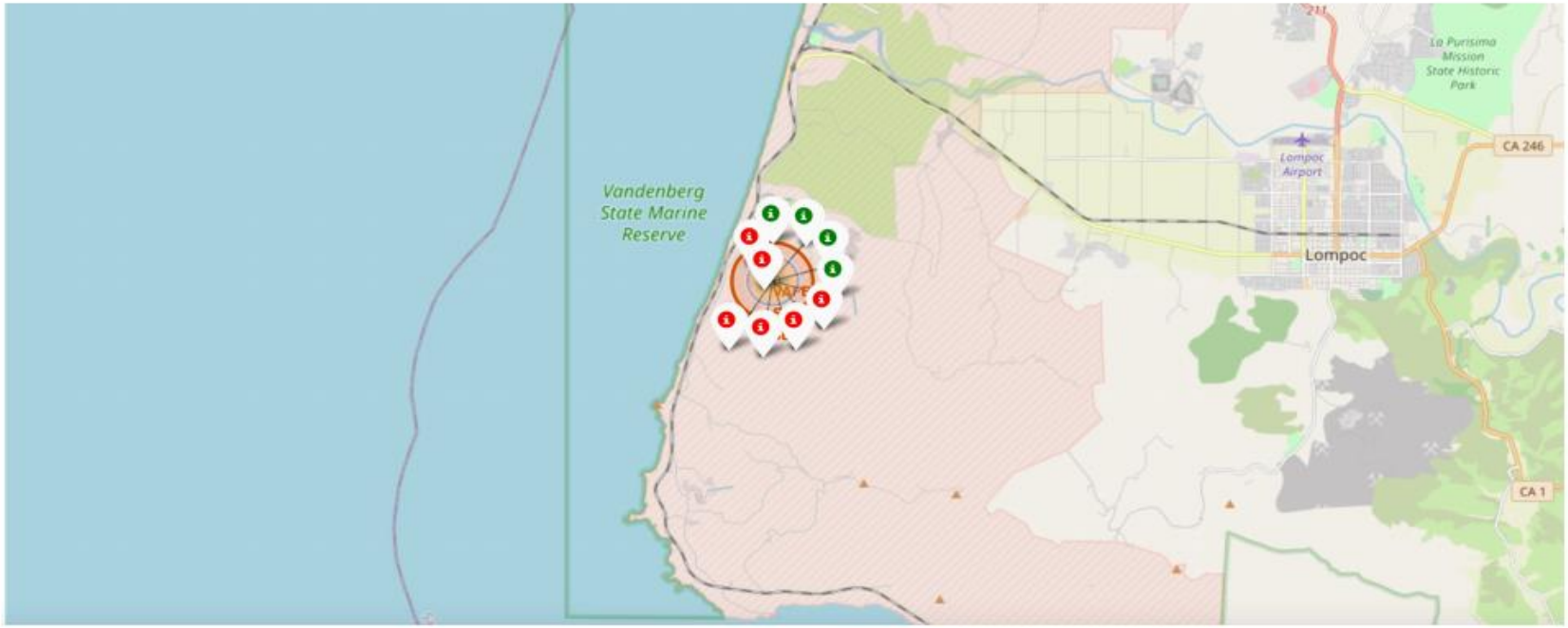
<Folium Map Screenshot 1>

- All launch sites on map



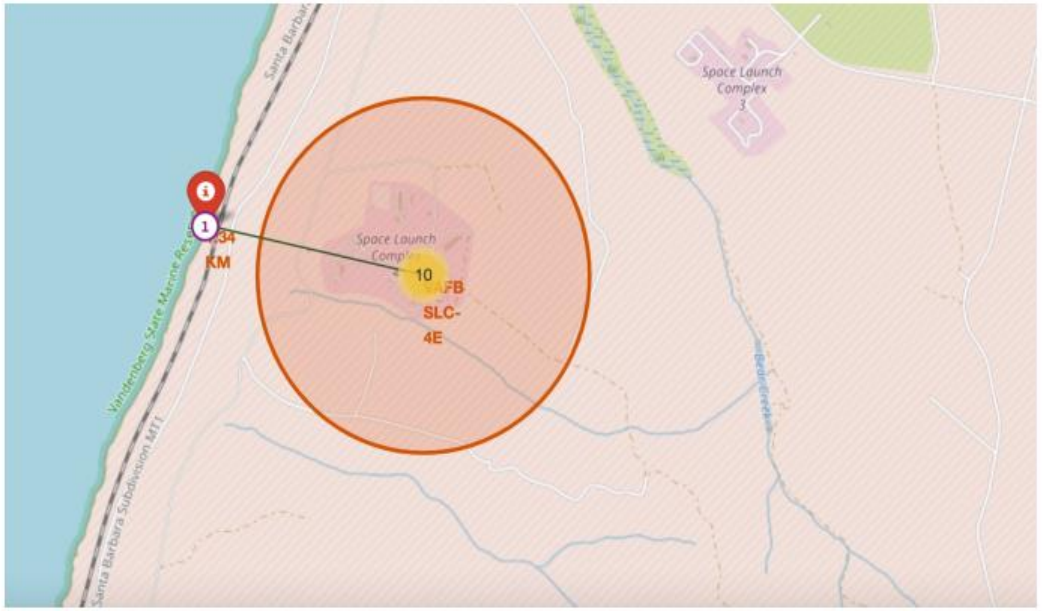
<Folium Map Screenshot 2>

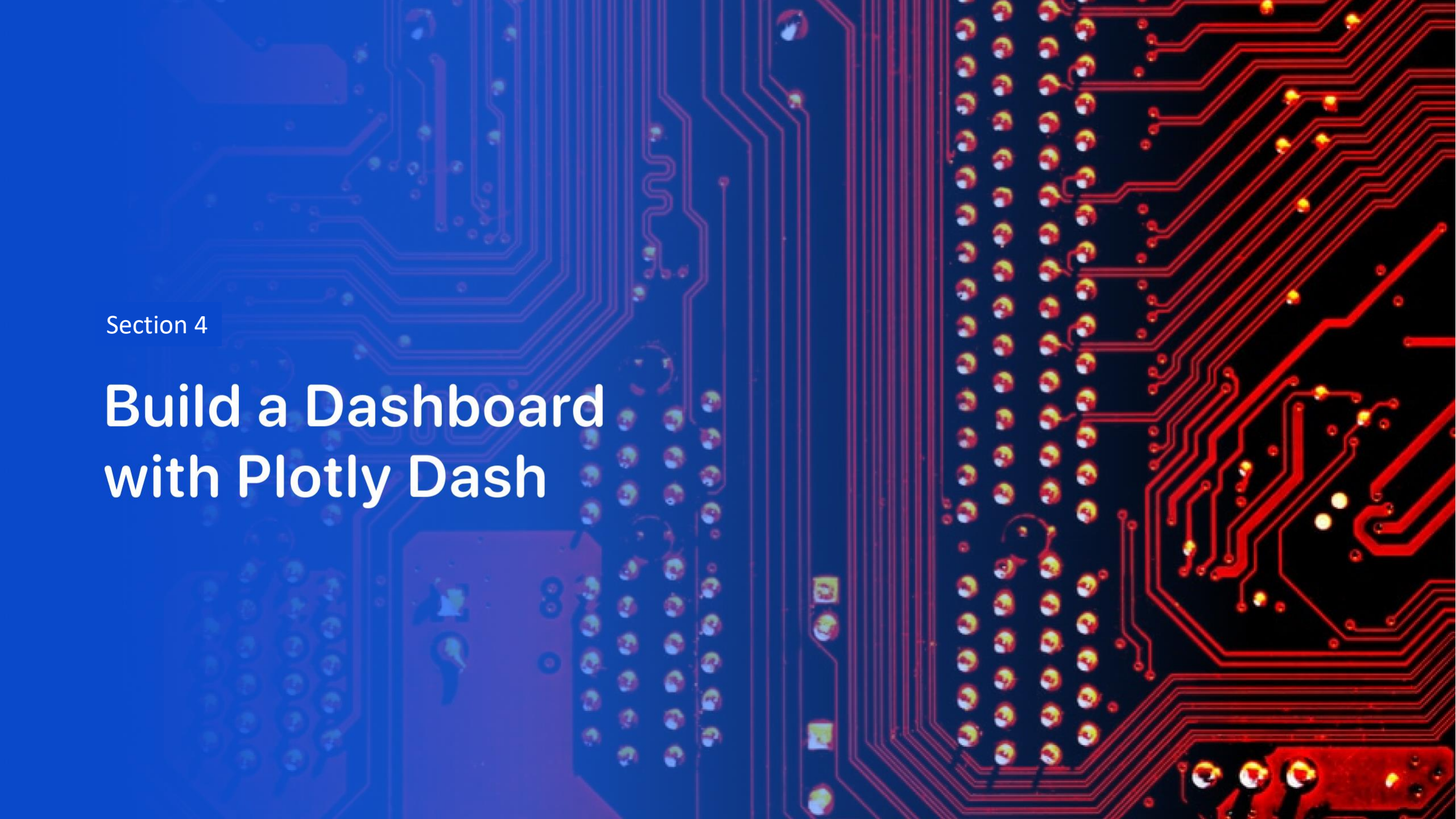
- The succeeded launches and failed launches for each site on map
  - If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch



<Folium Map Screenshot 3>

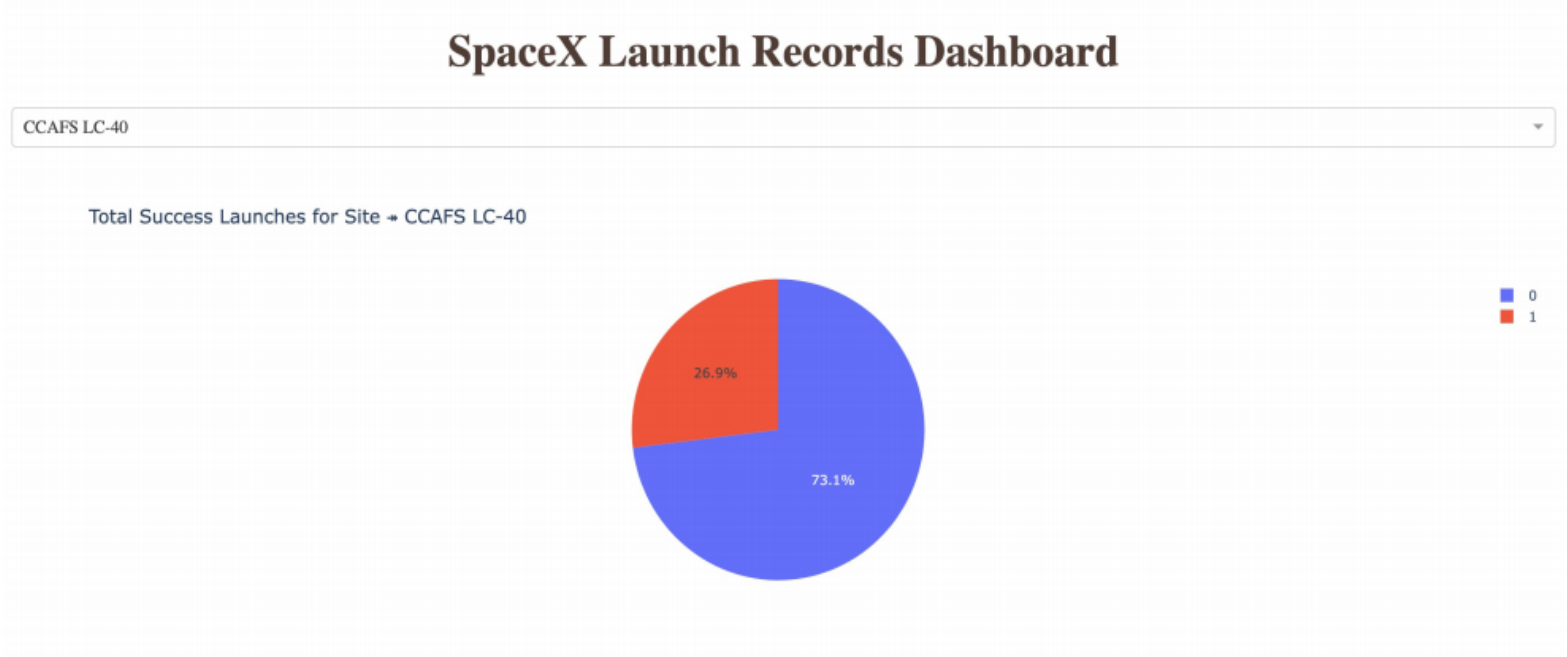
- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline




Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.



<Dashboard Screenshot 2>

- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

• Based on the scores of the Test Set, we can not confirm which method performs best.

• Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.

• The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy

Accuracy of test set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

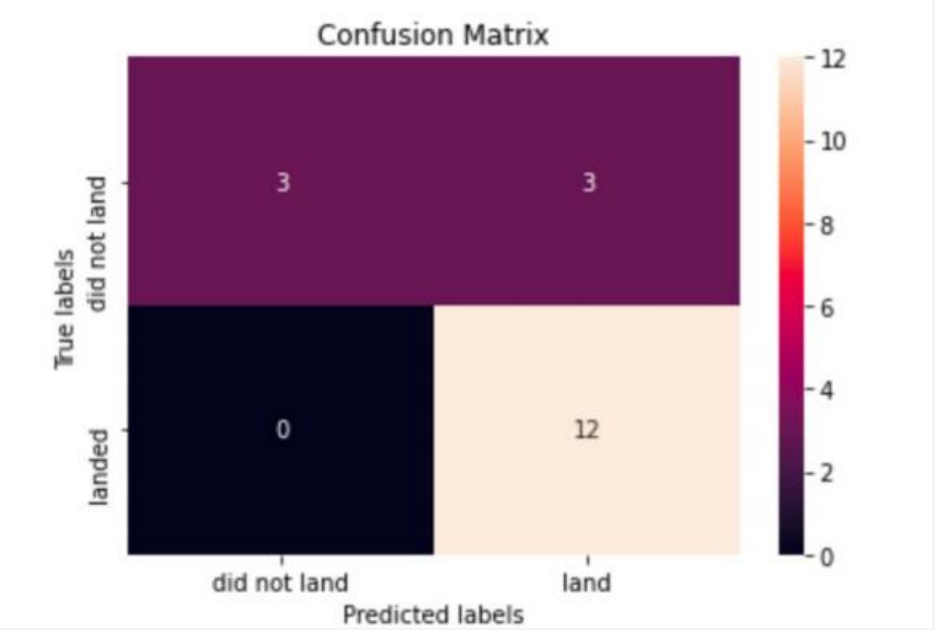
Accuracy of entire data set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation:

- 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

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

SPECIAL THANKS TO ALL COURSERA TEAM

Thank you!