IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abhishek Kumar
Beernelli
06-11-2023

Outline

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

Executive Summary

- Summary of methodologies:

1. Data collection
2. Data wrangling
3. Exploratory analysis with data visualization
4. Exploratory analysis with SQL
5. Building an interactive map with folium
6. Building a dashboard with plotly
7. Predictive analysis

Summary of all results:

Exploratory data analysis, Interactive analytics and Predictive analysis results.

Introduction

- SpaceX is the most successful company in the United States with commercial space and making space travel affordable to the world. SpaceX advertises Falcon 9 rockets on its website with money lesser than other rocket providing companies. Therefore, if we can determine if the first stage lands then we can predict its cost. Based on information with their APIs and Wikipedia we make machine learning models and predict the landing and reuse of the first stage of it.
- Questions to be answered:

1. How do payload mass, launch sites, flights and orbit occurrence will affect the success of first stage??
2. Rate of successful landings increase over the following years with lesser costs??
3. What is effective algorithm to use in this case??

Section 1

Methodology

Methodology

Executive Summary

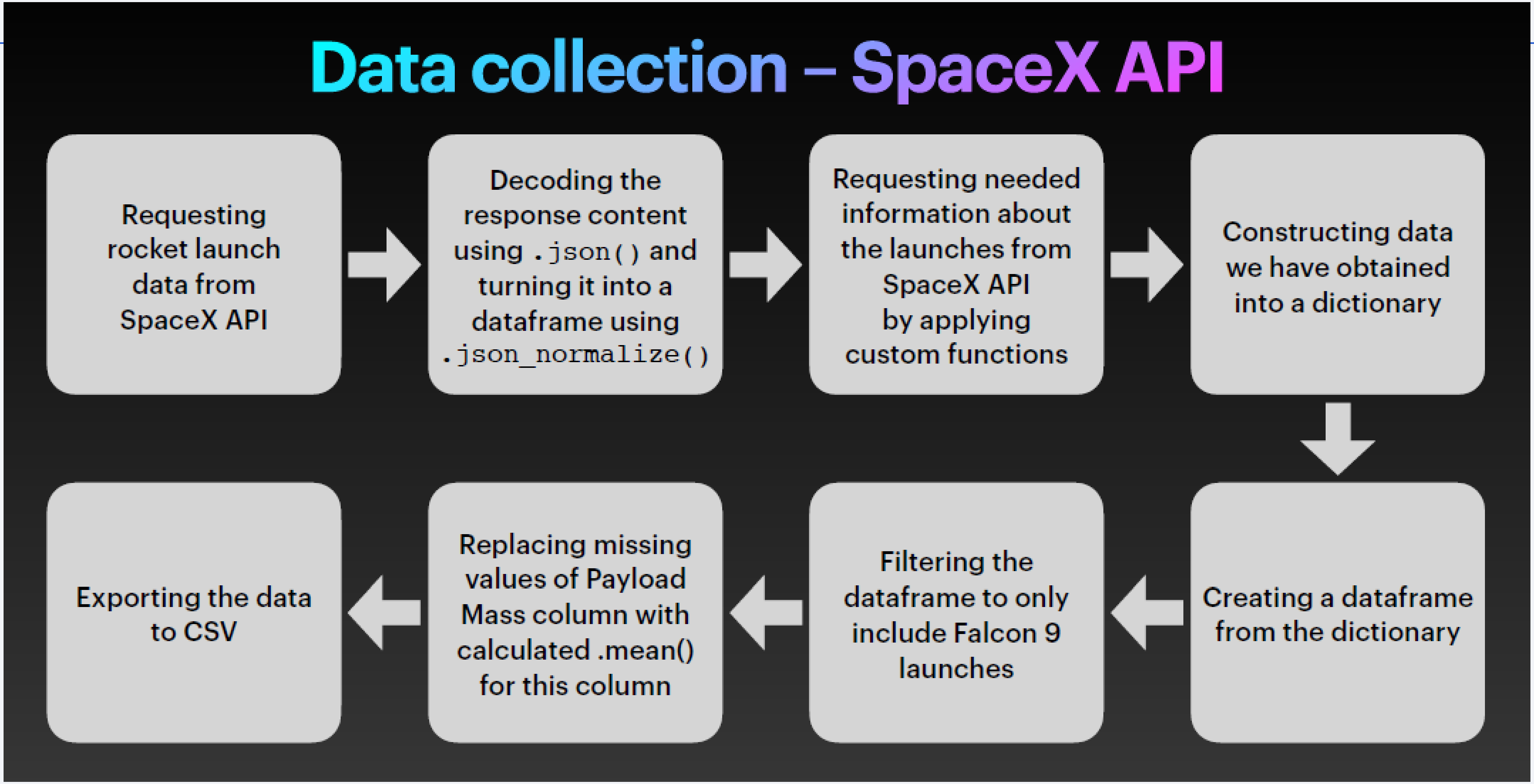
- Data collection methodology:
  - Data was collected using the SpaceX API and Wikipedia with the following are the links
  - Rockets data-`https://api.spacexdata.com/v4/rockets/`
  - Launchpads data-`https://api.spacexdata.com/v4/launchpads/`
  - Payload data-`https://api.spacexdata.com/v4/payloads/`
  - Cores data-`https://api.spacexdata.com/v4/cores/`
  - Past Launches-`https://api.spacexdata.com/v4/launches/past`
- Perform data wrangling
  - Data was first checked if there are any null values
  - Missing values are then replaced with mean values

Data Collection

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models-

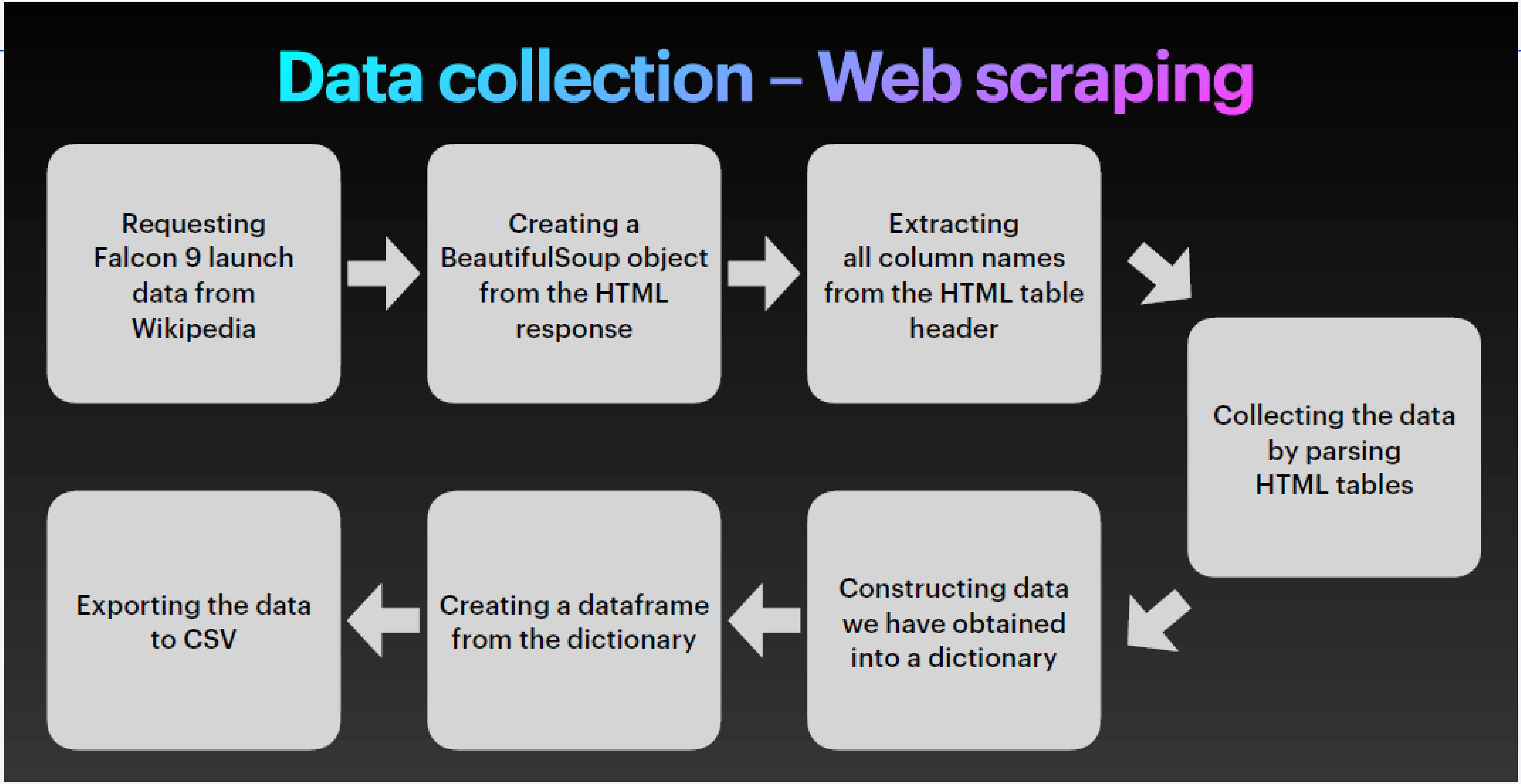
Building, Tuning and evaluation of classification models to ensure the best results. Data that was collected until this step were normalized, divided on training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

Data Collection – SpaceX API



https://github.com/BeernelliAbhishek/APPLIED-DATA-SCIENCE-PROJECT/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

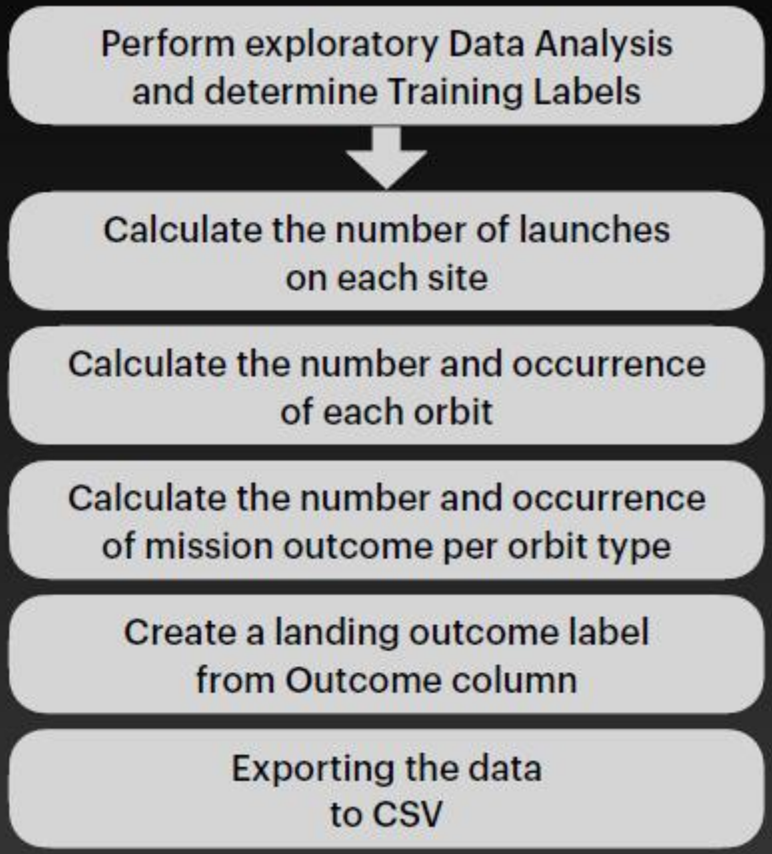


https://github.com/BeernelliAbhishek/APPLIED-DATA-SCIENCE-PROJECT/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

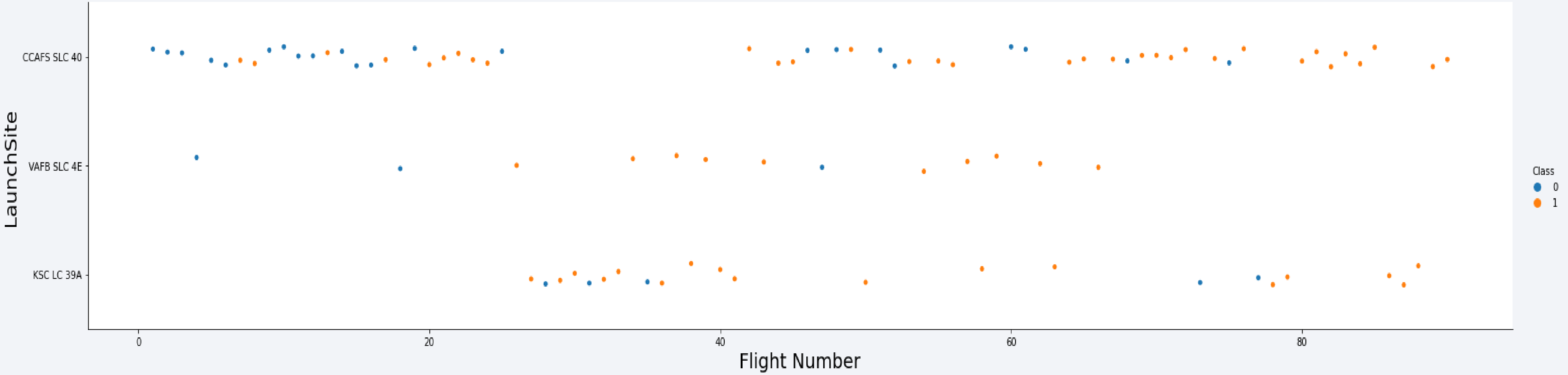
In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad 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 False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with "1" means the booster successfully landed, "0" means it was unsuccessful.



https://github.com/BeernelliAbhishek/APPLIED-DATA-SCIENCE-PROJECT/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization



CHARTS THAT 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-https://github.com/BeernelliAbhishek/APPLIED-DATA-SCIENCE-PROJECT/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Displaying the unique launch sites in the space mission;
  - Listing top 5 launch sites whose name begin with the string 'CCA';
  - Displaying total payload mass carried by boosters launched by NASA (CRS);
  - Displaying average payload mass carried by booster version F9 v1.1;
  - Listing date when the first successful landing outcome in ground pad was achieved;
  - Displaying the names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - List of all the successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

https://github.com/BeernelliAbhishek/APPLIED-DATA-SCIENCE-PROJECT/blob/main/EDA%20with%20SQL.ipynb

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.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured 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 coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

https://github.com/BeernelliAbhishek/APPLIED-DATA-SCIENCE-PROJECT/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

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.

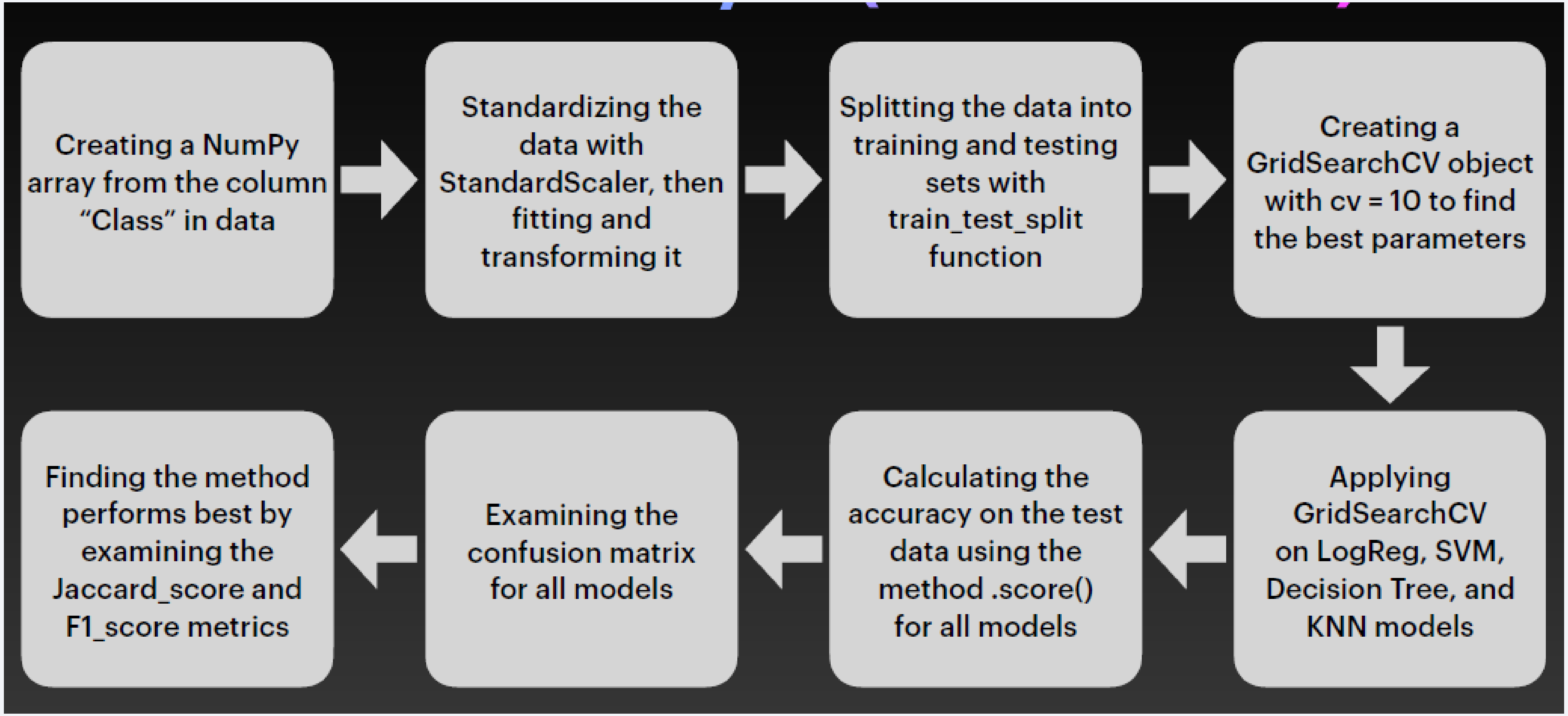
Slider of Payload Mass Range:

- 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)



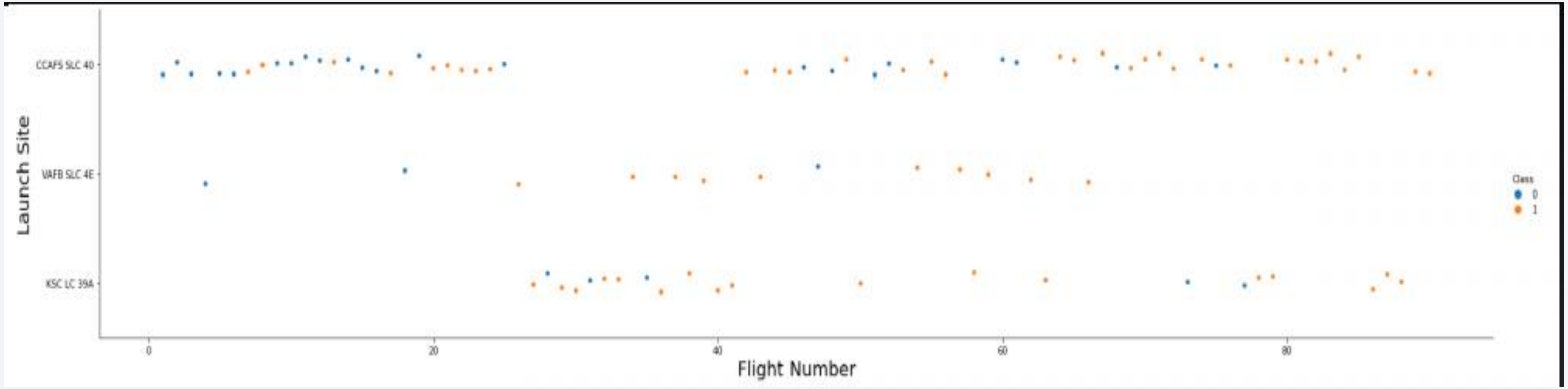
Results

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

Section 2

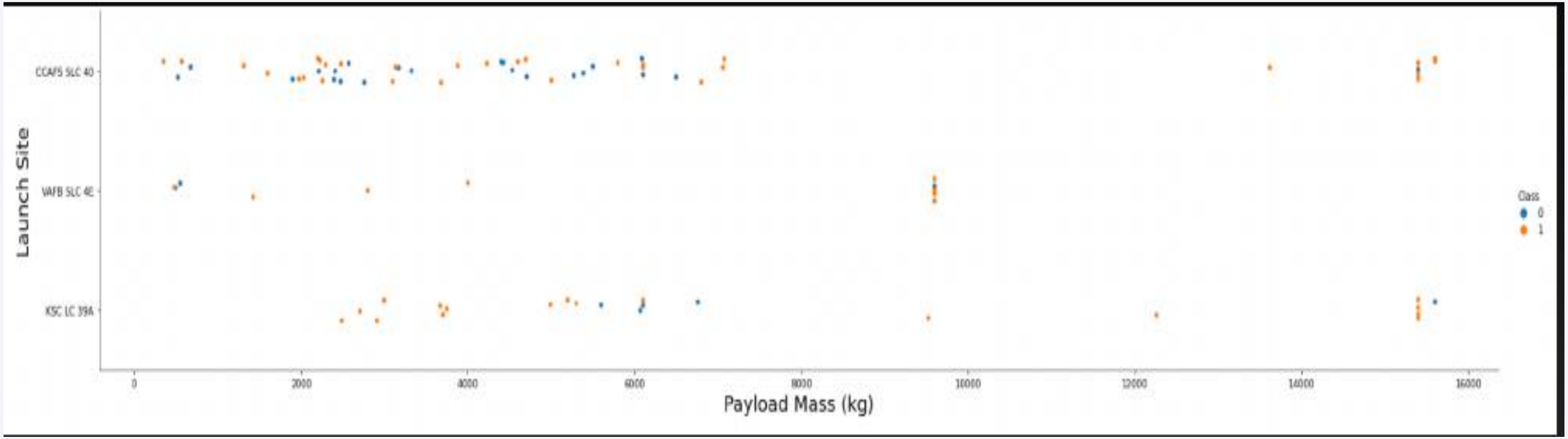
Insights drawn from EDA

Flight Number vs. Launch Site



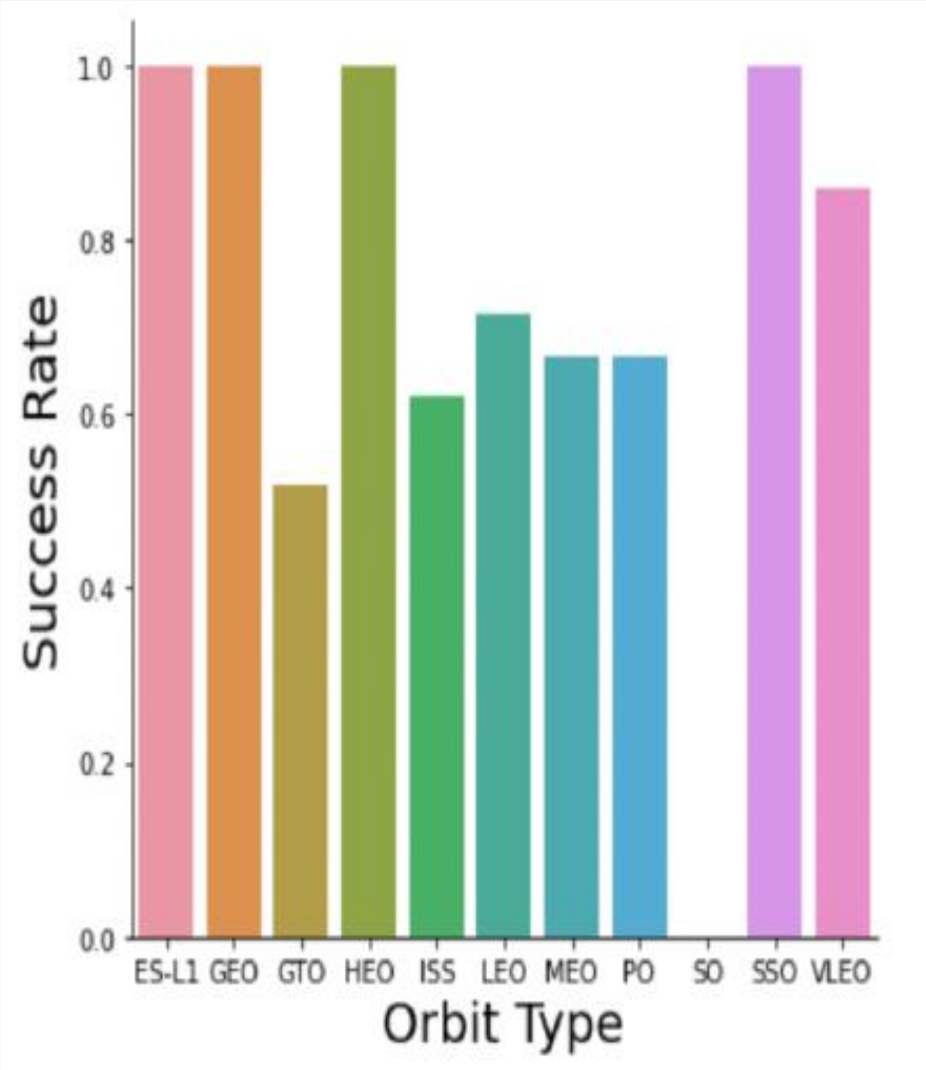
- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSCLC 39A;
- It's also possible to see that the general success rate improved over time.

Payload vs. Launch Site



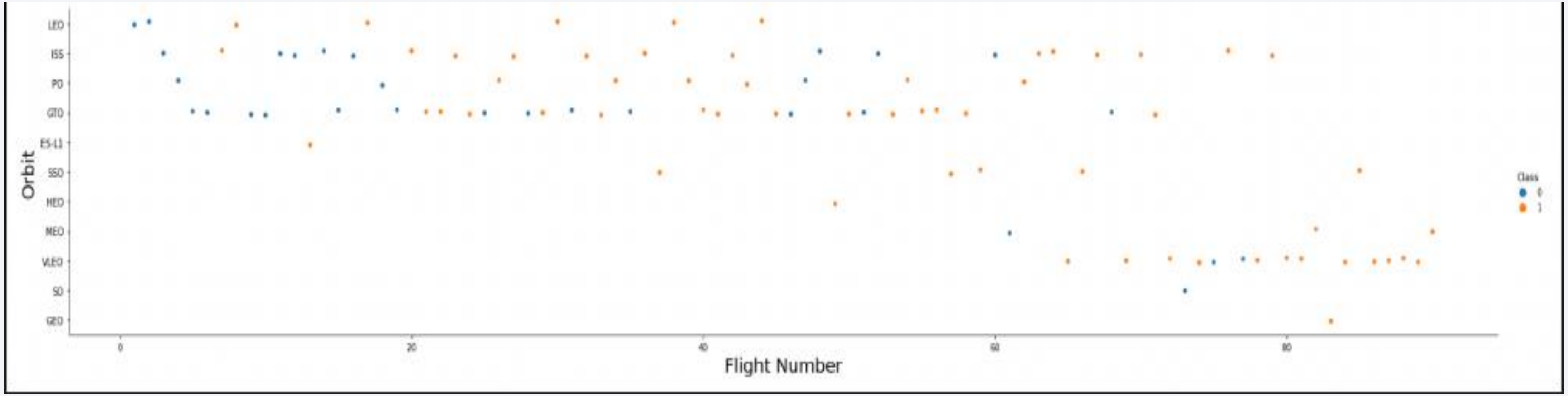
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type



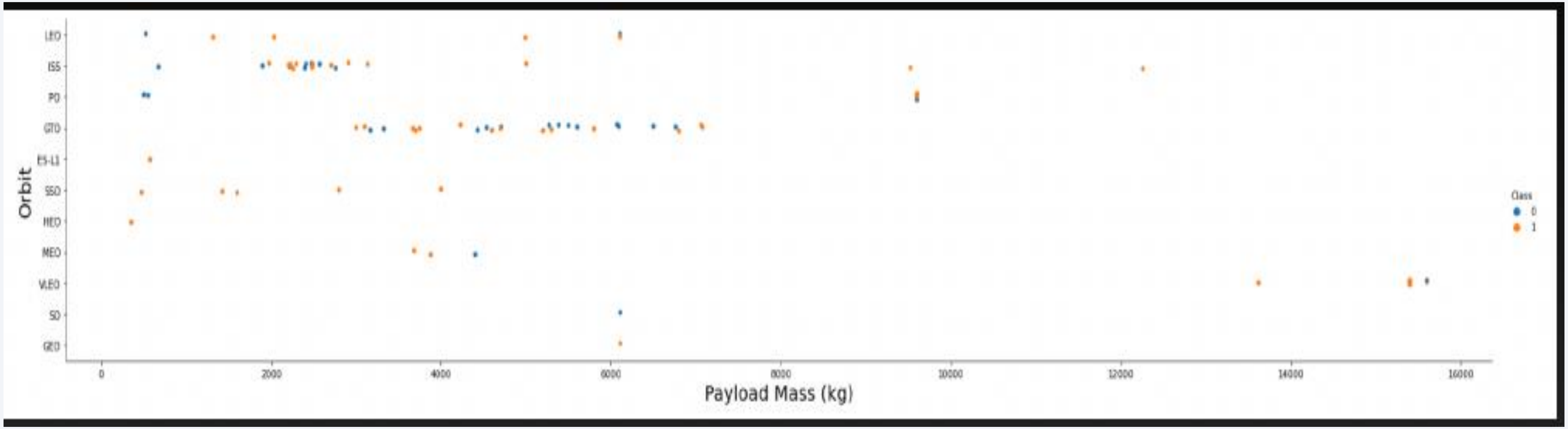
- The biggest success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).

Flight Number vs. Orbit Type



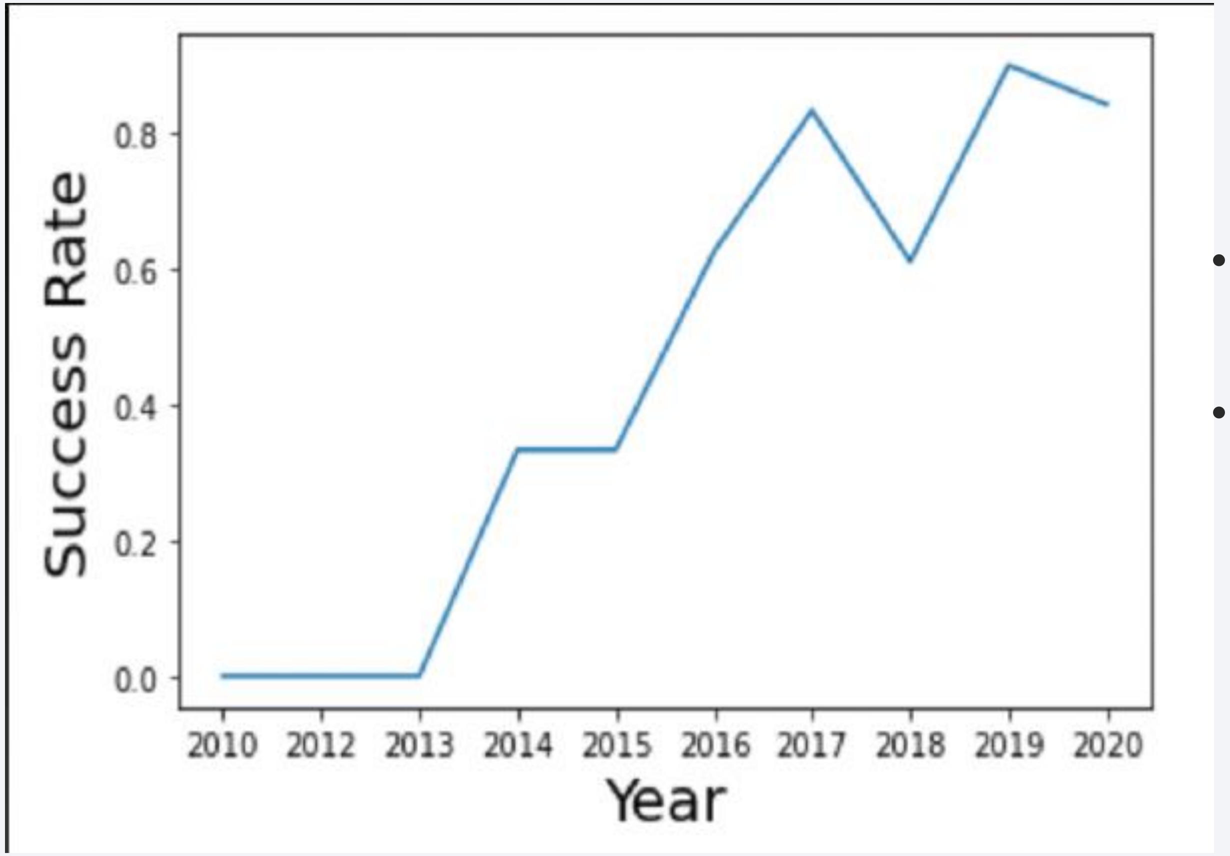
- Apparently, success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

Payload vs. Orbit Type



- Apparently, there is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend



- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.

All Launch Site Names

```
%sql select distinct launch_site from SPACEXDATASET;
```

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

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

THESE ARE THE LAUNCH SITES

Launch Site Names Begin with 'CCA'

```
%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.
```

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

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

average_payload_mass
2534

First Successful Ground Landing Date

```
%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.
```

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%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.
```

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

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

```
%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.
```

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

```
%%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)

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

```
%%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.
```

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>

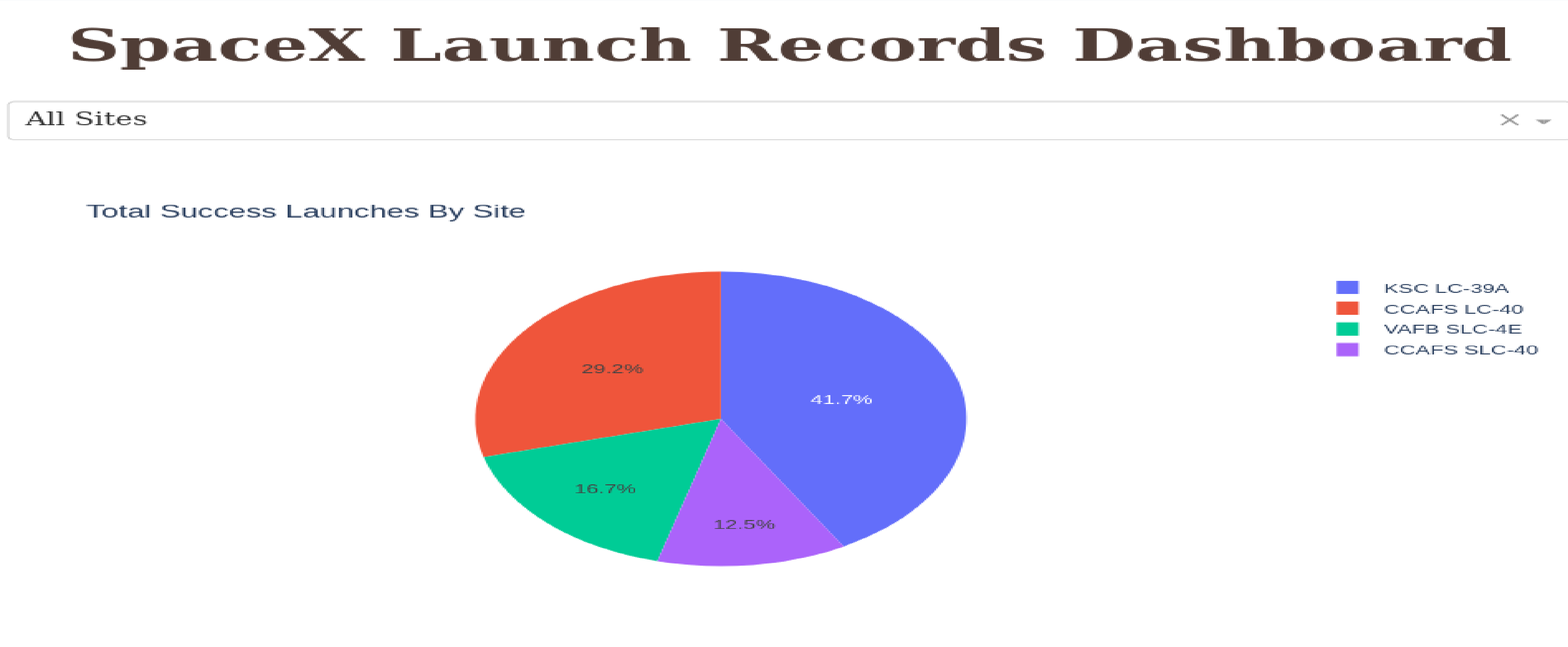
<Folium Map Screenshot 2>

<Folium Map Screenshot 3>

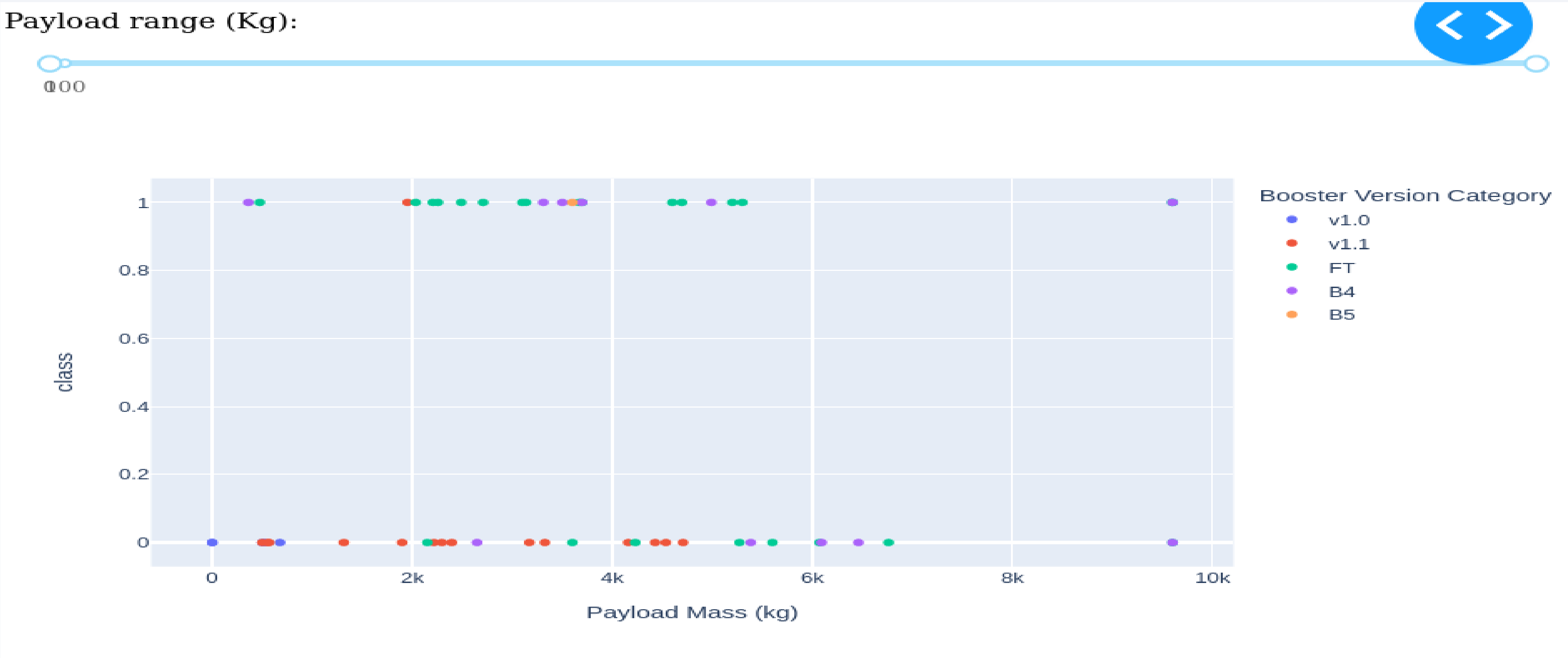
Section 4

Build a Dashboard with Plotly Dash

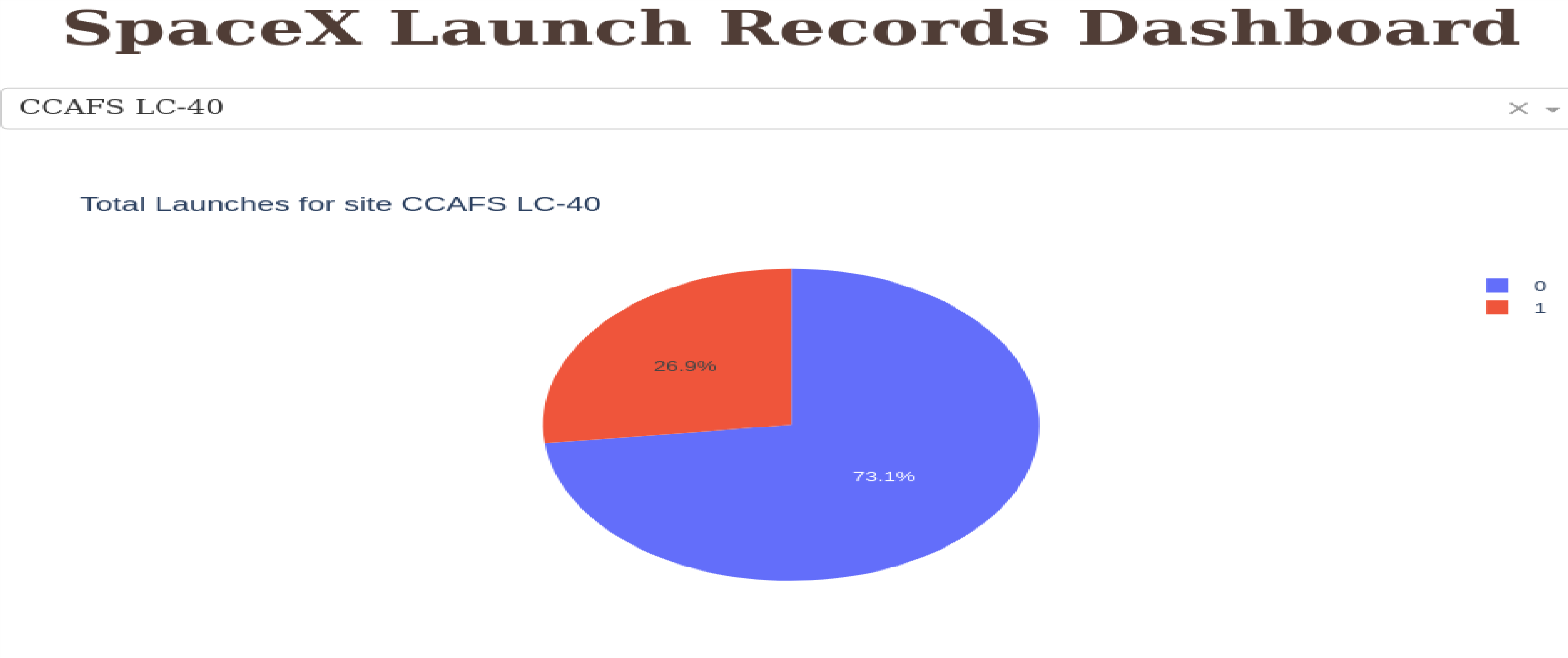
<Dashboard Screenshot 1>



<Dashboard Screenshot 2>



<Dashboard Screenshot 3>



Section 5

Predictive Analysis (Classification)

Classification Accuracy

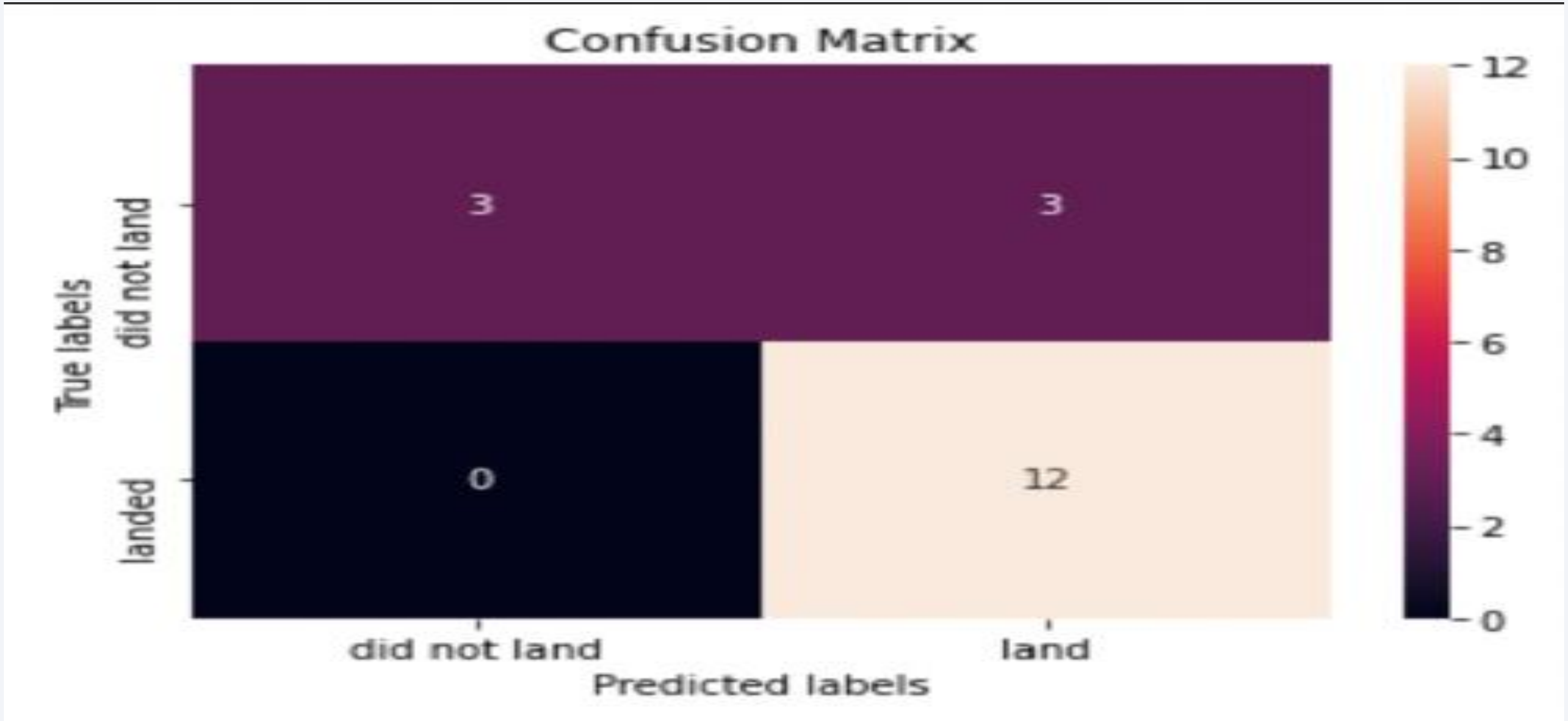
Scores and Accuracy of the Test Set

	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

Scores and Accuracy of the 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



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

1. **Data Insights and Exploration:** Detail how the project uncovered valuable insights through data analysis, helping to understand various aspects of space exploration such as celestial bodies, orbits, and cosmic events.
2. **Predictive Modeling and Forecasting:** Highlight the predictive models developed using IBM Data Science tools that helped in forecasting space phenomena like satellite movements, space debris, or planetary events with increased accuracy.
3. **Optimization and Efficiency:** Discuss how the project contributed to optimizing space exploration processes, potentially reducing costs or improving mission efficiency by leveraging data-driven insights and predictive analytics.
4. **Machine Learning Applications:** Explain the application of machine learning algorithms in analyzing complex space data, such as image recognition for celestial objects or natural language processing for understanding scientific literature.
5. **Technological Innovations:** Highlight any novel technological solutions or methodologies developed within the project that could be applicable not only to space exploration but also to other domains where data analysis is crucial.

APPENDIX

1. **Challenges Overcome:** Mention the challenges faced during the project, how they were addressed, and what lessons were learned in dealing with complex and dynamic space data.
2. **Future Prospects and Recommendations:** Discuss potential future applications or expansions of the project, suggesting how the insights gained can be utilized for further advancements in space exploration or related scientific fields.
3. **Collaboration and Interdisciplinary Work:** Emphasize the interdisciplinary nature of the project, how different expertise areas like data science, astrophysics, and engineering collaborated to achieve the project goals.
4. **Impact and Significance:** Discuss the potential impact of the project on the field of space exploration, whether in terms of scientific discoveries, technological advancements, or cost-efficient methods.
5. **Acknowledgments and Gratitude:** Show appreciation for the collaborative efforts of the team, mentors, data providers, and the support of IBM's resources and tools in successfully executing the project.

INSIGHTS

1. **Data Security and Confidentiality:** Ensure that any sensitive or proprietary data obtained from SpaceX or other sources is handled securely, following all relevant data protection and confidentiality protocols.
2. **Legal and Ethical Compliance:** Adhere to legal and ethical guidelines concerning space-related data, especially when dealing with data that might have export restrictions or intellectual property considerations.
3. **Data Integrity and Quality:** Validate the quality and integrity of the data used. Check for any errors, inconsistencies, or biases in the data to avoid misleading analysis and conclusions.
4. **Understanding the Context of Space Data:** Recognize the uniqueness of space data and its potential limitations. Space data might have inherent challenges, such as signal noise, calibration issues, or inherent uncertainties due to the nature of space environments.
5. **Collaboration with Experts:** Collaborate with domain experts in space science and engineering to ensure a deep understanding of the data and its context. Engage with professionals who have experience in handling space-related data.

INSIGHTS

1. **Use Reliable Tools and Methodologies:** Utilize trusted and validated IBM Data Science tools and methodologies for data analysis and interpretation. Be cautious about the tools used and ensure they are appropriate for space data analysis.
2. **Risk Assessment and Mitigation:** Identify and address potential risks associated with space data analysis. Be aware of the potential impact of incorrect analysis or misinterpretation of space-related information.
3. **Regular Validation and Review:** Continuously validate findings and interpretations with experts and peers. Peer review and validation are crucial, particularly in complex fields such as space science.
4. **Documentation and Transparency:** Maintain comprehensive documentation of methodologies, data sources, and analytical processes. Ensure transparency in the analysis and conclusions reached.
5. **Backup and Recovery Plan:** Given the critical nature of space data, have backup plans for data storage, ensuring robust recovery procedures in case of data loss or system failures.
6. **Regulatory Compliance:** Consider any regulatory requirements related to space data analysis, especially if the project involves collaboration with government agencies or private space companies.
7. **Communication and Dissemination:** Clearly communicate the limitations and uncertainties associated with space data analysis. Avoid overgeneralization and communicate findings responsibly, especially when addressing the public or media.

Thank you!