Winning Space Race with Data Science

Bill Chen
2023.09.29

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Data Visualization
  - Data Analysis with SQL and python
  - Folium
  - Interactive Dashboard with Ploty
  - Predictive analysis
- Summary of all results
  - Data Analysis
  - Interactive Analysis
  - Predictive Analysis

Introduction

- Project background and context
  - SpaceX stands out as the leading company in the era of commercial space exploration, having played a pivotal role in rendering space travel more cost-effective. On its website, the company promotes Falcon 9 rocket launches at a price point of $62 million, which is significantly more economical than other providers whose prices exceed $165 million for each launch. A substantial portion of these savings can be attributed to SpaceX's ability to recycle the initial stage of the rocket. Consequently, by ascertaining the likelihood of a successful first stage landing, we can gauge the expense associated with a launch. Leveraging publicly available data and advanced machine learning models, we intend to make predictions regarding the reusability of SpaceX's first stage.
- Problems you want to find answers
  - What impact do variables like payload mass, launch site, number of flights, and orbits have on the probability of a successful first stage landing?
  - Is there an observable trend of increasing success rates for first stage landings over time?
  - Which binary classification algorithm is the most suitable for addressing this specific case?

Section 1

Methodology

Methodology

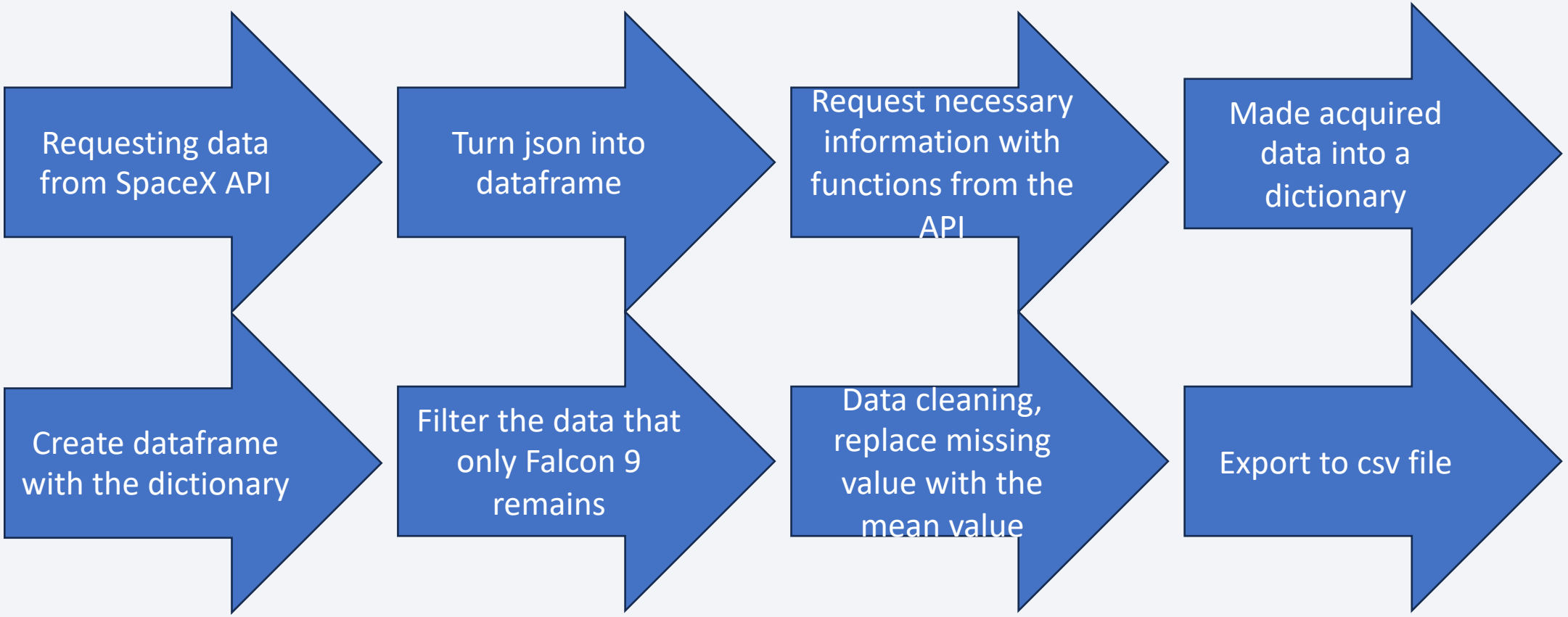
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

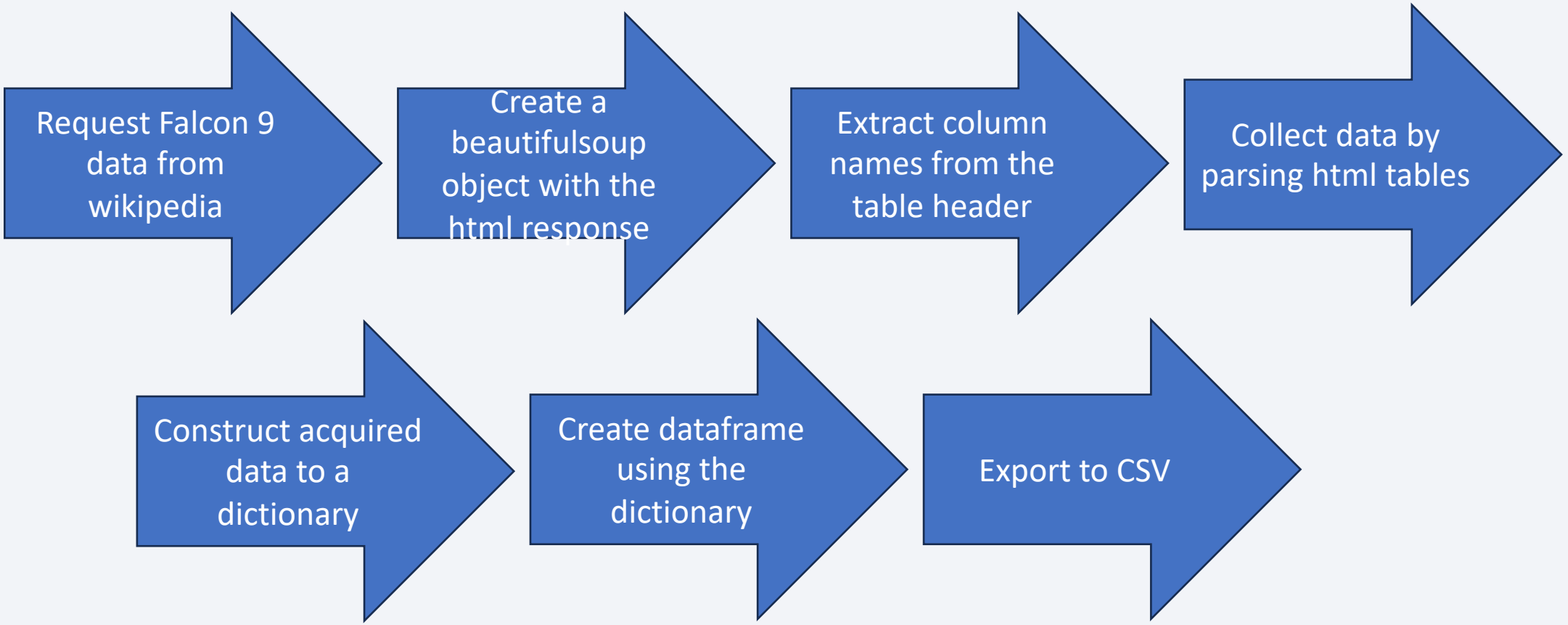
- We employed a two-pronged data collection approach, utilizing API requests from SpaceX's REST API and web scraping from a table within SpaceX's Wikipedia page. This dual method was necessary to ensure comprehensive data acquisition for a more in-depth analysis.
- Data obtained through SpaceX's REST API includes columns like FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- Data gathered through Wikipedia web scraping includes Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API



Data Collection – GitHub

Data Collection - Scraping



Data Scraping – GitHub

Data Wrangling

Within the dataset, various scenarios exist where the booster's landing did not achieve success. These instances encompass attempted landings that failed due to accidents. For instance, "True Ocean" signifies a mission outcome where the booster successfully landed in a designated ocean region, while "False Ocean" indicates an unsuccessful ocean landing. Similarly, "True RTLS" denotes a successful ground pad landing, whereas "False RTLS" represents an unsuccessful ground pad landing. "True ASDS" signifies a successful drone ship landing, and "False ASDS" denotes an unsuccessful drone ship landing. These outcomes are primarily converted into training labels, with "1" indicating a successful booster landing and "0" indicating an unsuccessful landing.

Data Wrangling – GitHub

EDA with Data Visualization

Charts:

1. - Created various charts including 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.

2. - **Scatter plots** were used to reveal relationships between variables, which can be useful for building machine learning models.

3. - **Bar charts** were employed to compare discrete categories and highlight relationships between specific categories and measured values.

4. - **Line charts** were utilized to visualize data trends over time, particularly for time series data.

Data Visualization – GitHub

EDA with SQL

1. List unique launch site names.
2. Show 5 records with launch sites starting with 'CCA.'
3. Display total payload mass for NASA (CRS) launches.
4. Show average payload mass for booster version F9 v1.1.
5. Find the date of the first successful ground pad landing.
6. List booster names with successful drone ship landings and payload mass between 4000 and 6000.
7. List the total counts of successful and failed mission outcomes.
8. Identify booster versions with the maximum payload mass.
9. List failed drone ship landing outcomes, their booster versions, and launch site names for January 2015.
10. Rank landing outcome counts (e.g., Failure - drone ship or Success - ground pad) between June 4, 2010, and March 20, 2017, in descending order.

SQL – GitHub

Build an Interactive Map with Folium

For All Launch Sites:

- Created markers with circles, popup labels, and text labels for the NASA Johnson Space Center using its latitude and longitude coordinates as the starting point.

- Added markers with circles, popup labels, and text labels for all launch sites, displaying their geographical locations and their proximity to the Equator and coasts.

Launch Outcome Markers:

- Utilized colored markers (Green for success, Red for failures) using a Marker Cluster to identify launch sites with higher success rates.

Measuring Distances from Launch Sites:

- Incorporated colored lines to depict distances between the KSC LC-39A launch site (as an example) and nearby features such as railways, highways, coastlines, and the closest city.

Folium – GitHub

Build a Dashboard with Plotly Dash

Interface Enhancements:

- Introduced a **Launch Sites dropdown list** for convenient site selection.

Visualization Features:

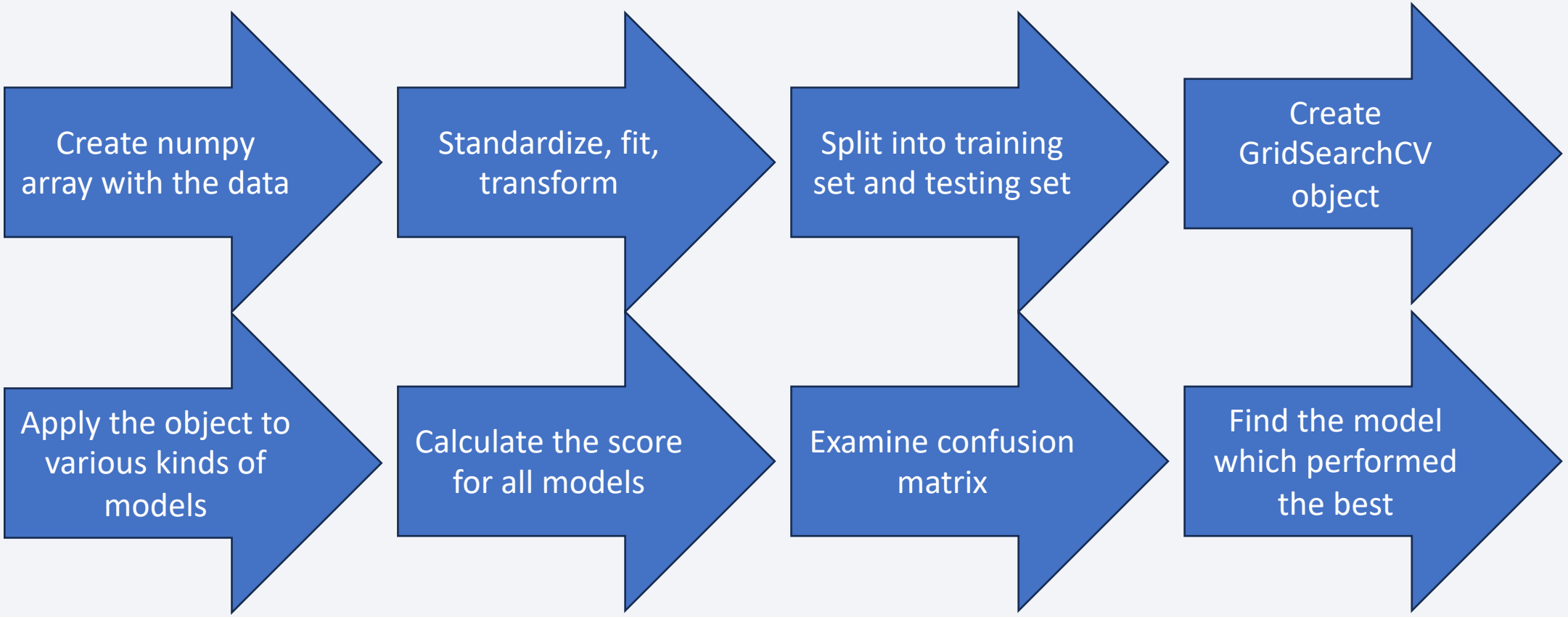
- Implemented a **Pie Chart** to display the total count of successful launches for all sites and, if a specific launch site is chosen, to illustrate the success vs. failed launch counts for that site.

- Included a **Payload Mass Range slider** for users to select specific payload ranges.

- Designed a **Scatter Chart** to depict the relationship between payload mass and launch success rate for various booster versions.

Dash App– GitHub

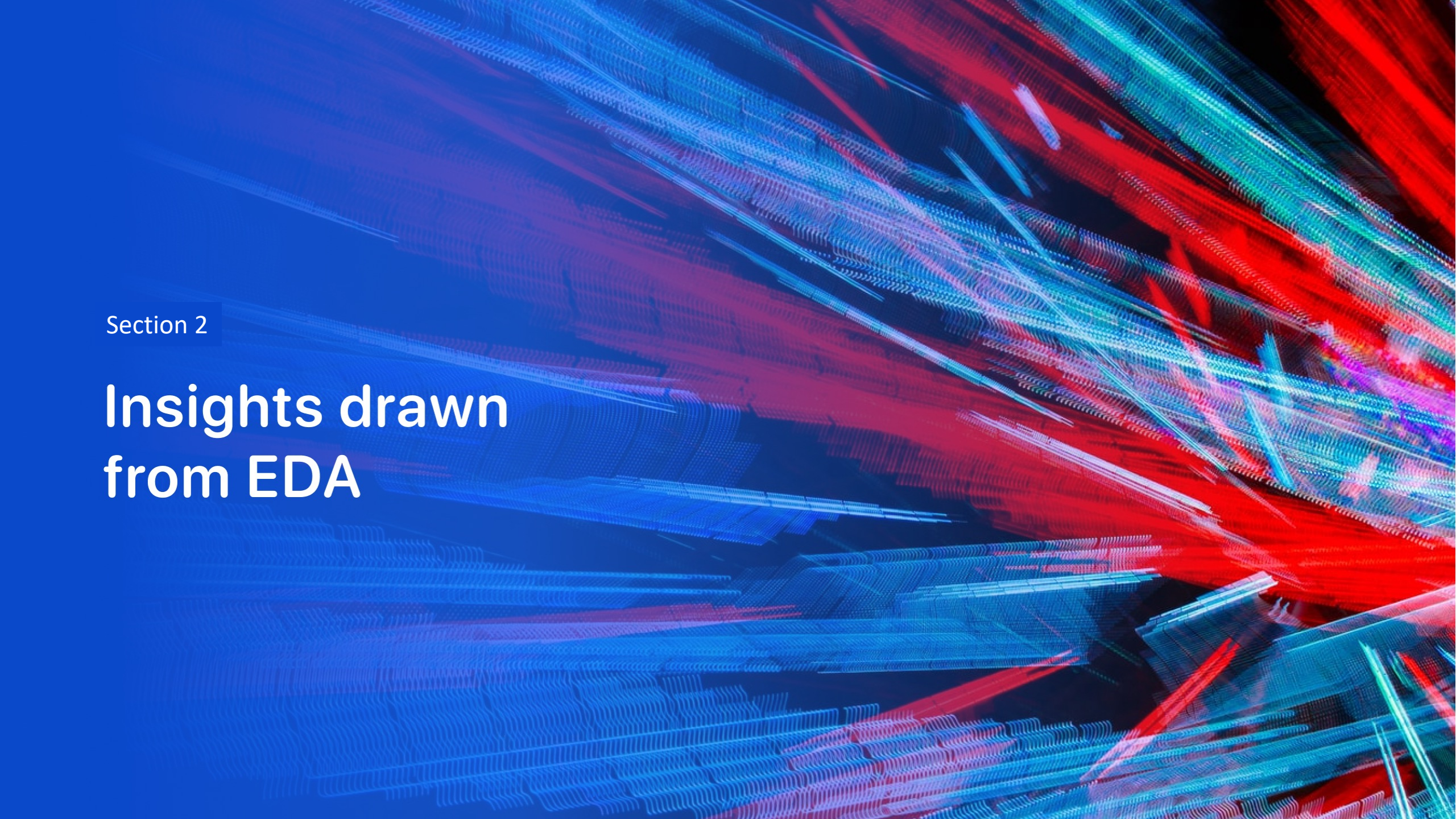
Predictive Analysis (Classification)



ML Prediction– GitHub

Results

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


Section 2
Insights drawn from EDA

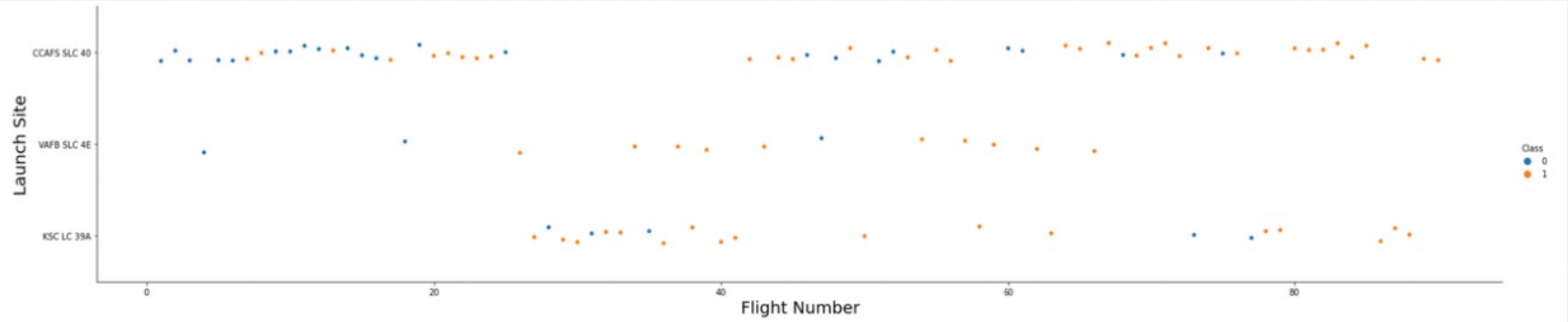
Flight Number vs. Launch Site

- Initial flights experienced a series of failures, whereas the most recent flights achieved consistent success.

- Approximately half of all launches took place at the CCAFS SLC 40 launch site.

- The VAFB SLC 4E and KSC LC 39A launch sites demonstrated notably higher success rates.

- There's a reasonable assumption that each successive launch exhibits an improved rate of success.

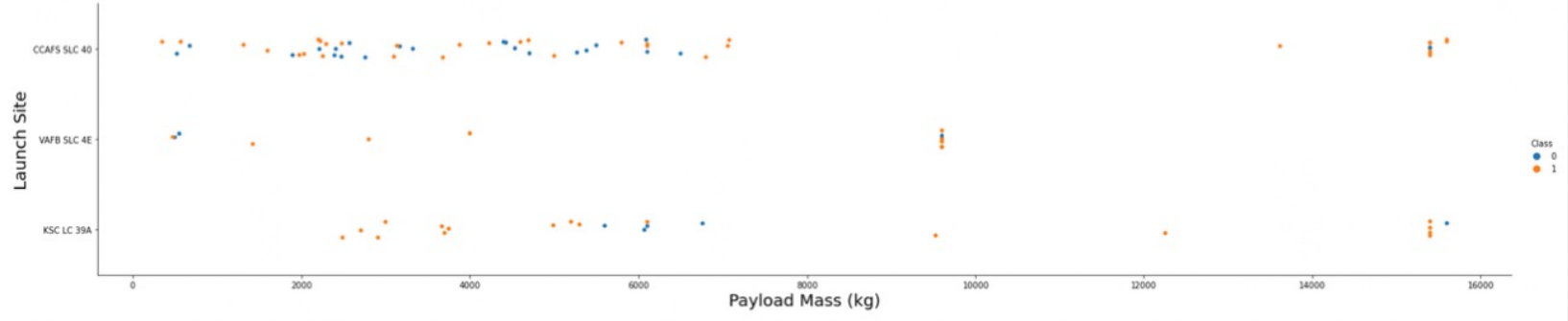


Payload vs. Launch Site

- Across all launch sites, there's a positive correlation: as payload mass increases, the success rate tends to be higher.

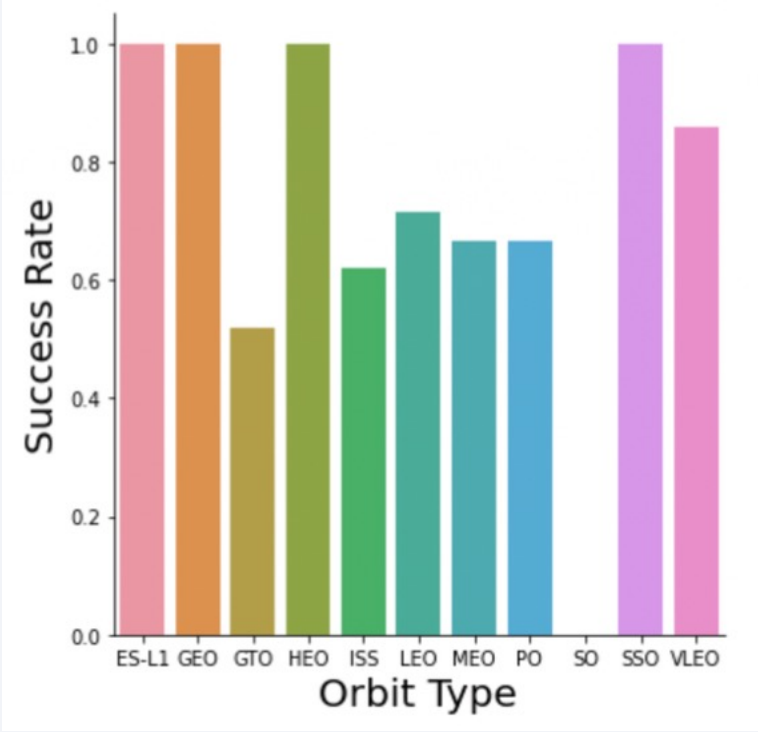
- The majority of launches with a payload mass exceeding 7000 kg achieved success.

- Notably, KSC LC 39A maintained a 100% success rate even for payload masses under 5500 kg.



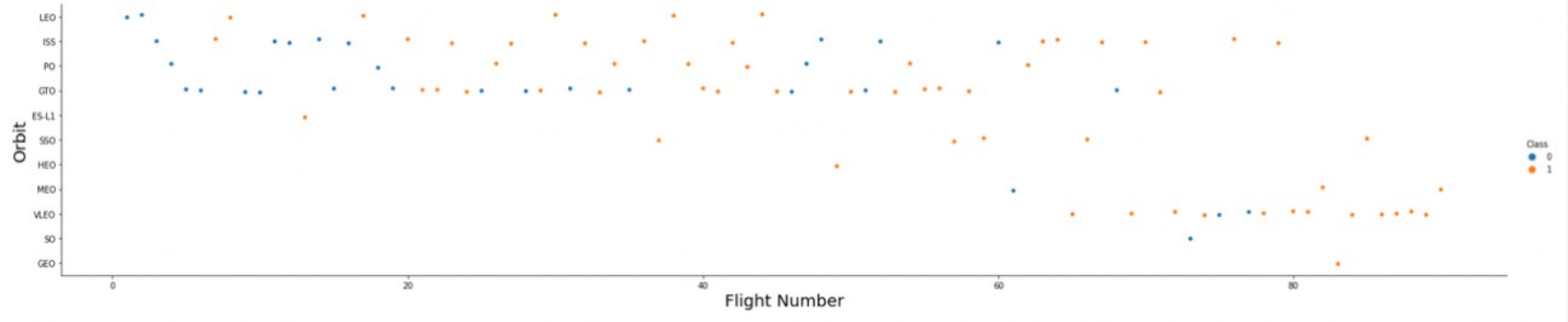
Success Rate vs. Orbit Type

- 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



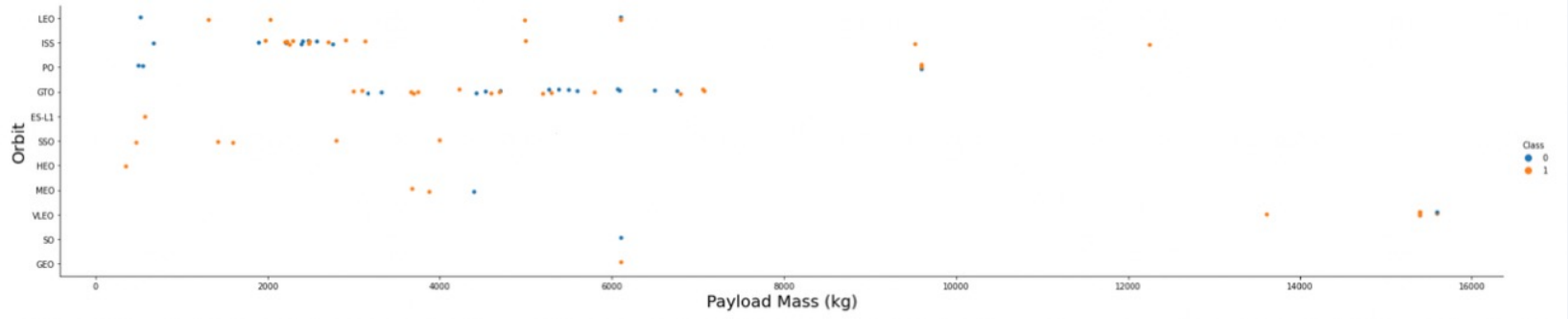
Flight Number vs. Orbit Type

- In the Low Earth Orbit (LEO), there appears to be a relationship between the number of flights and launch success. However, in the Geostationary Transfer Orbit (GTO), there doesn't seem to be any discernible relationship between flight number and success.



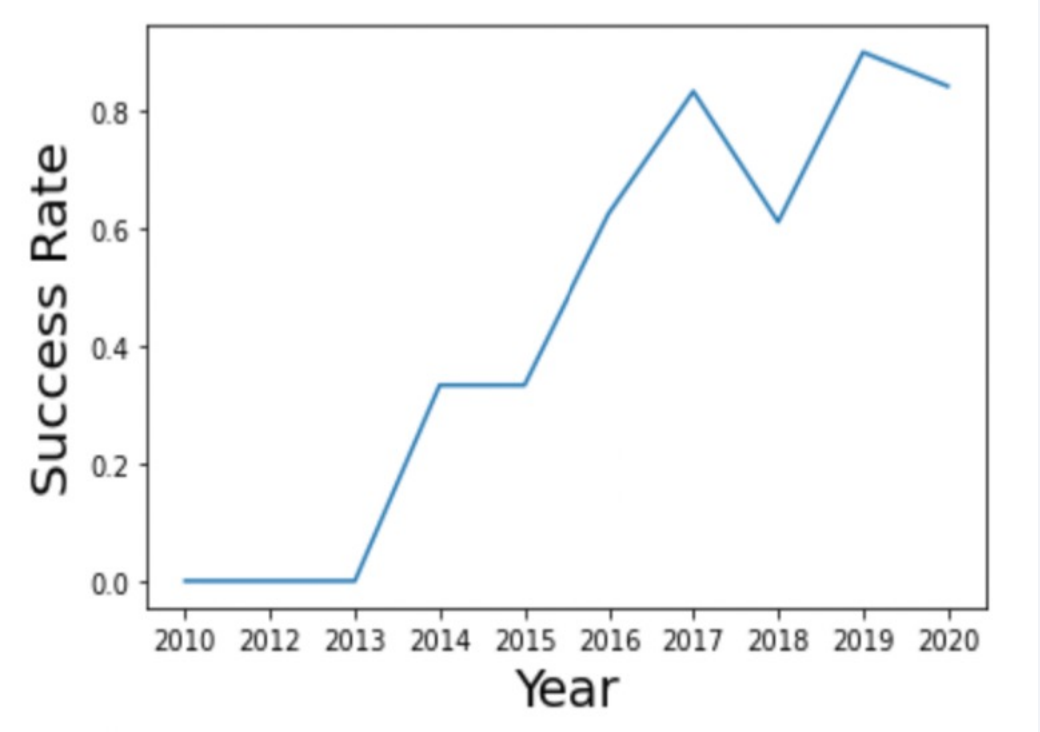
Payload vs. Orbit Type

- Heavy payloads tend to have a negative influence on success rates in Geostationary Transfer Orbits (GTO) but a positive influence on success rates in Geostationary Transfer Orbits (GTO) and Polar Low Earth Orbits (e.g., ISS orbit).



Launch Success Yearly Trend

The success rate kept increasing from 2013 to 2020 but with a small decline in 2018.



All Launch Site Names

- Display all the unique 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'

- Show 5 records of launch site names being with CCA

Task 2

Display 5 records where launch sites begin with the string '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:311
98/bludb
Done.
```

Out[5]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
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 (para
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No a
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No a
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No a

Total Payload Mass

- Show the total payload mass carrid by boosters launched by NASA

Task 3

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

```
%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.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

Task 4

Display 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.
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
98/bludb
Done.
```

Out[7]:

average_payload_mass
2534

First Successful Ground Landing Date

- List the date when the first succesful landing outcome in ground pad was acheived.

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [8]:
```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pa
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
98/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__
```

```
 * 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

- List the total number of successful and failure mission outcomes

Task 7

List 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 the names of the booster_versions which have carried the maximum payload mass.

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPAC
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:311
98/bludb
Done.
```

t[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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015

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:311
98/bludb
Done.
```

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

- 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

Task 10

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

[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;
```

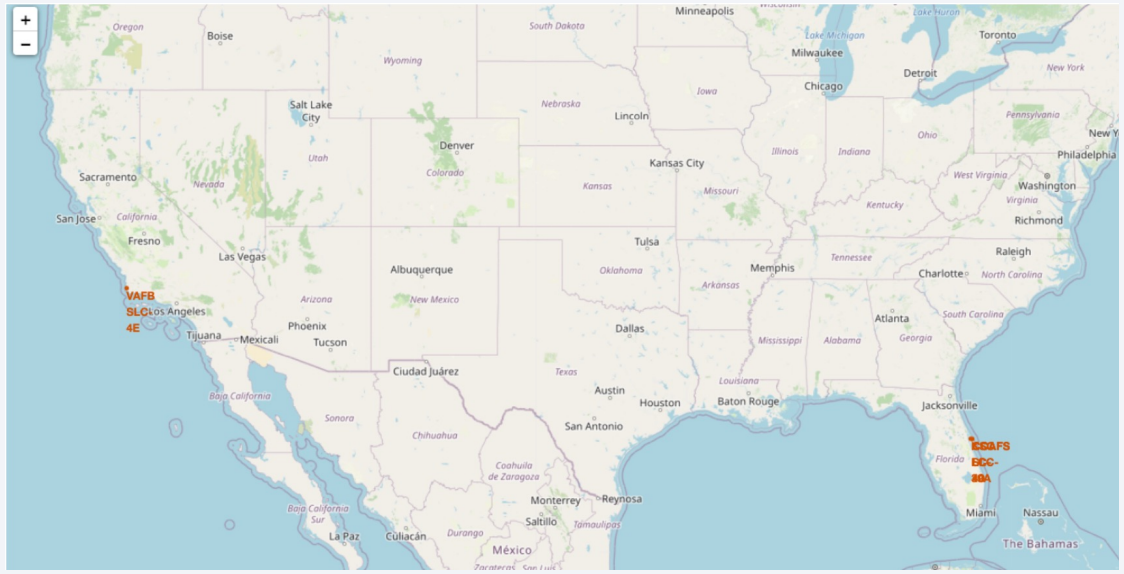
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

[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

Launch sites' locations on global map



- Most launch sites are strategically located near the Equator, where the Earth's rotation imparts a high initial velocity to anything on the surface. Objects at the Equator are already moving at a remarkable speed of 1670 km/hour. When a spacecraft is launched from the Equator, it inherits this velocity, helping it maintain the necessary speed to stay in orbit due to inertia.
- Additionally, all launch sites are situated in close proximity to coastlines. Launching rockets over the ocean minimizes the risk of debris falling or exploding near populated areas, enhancing safety during space missions.

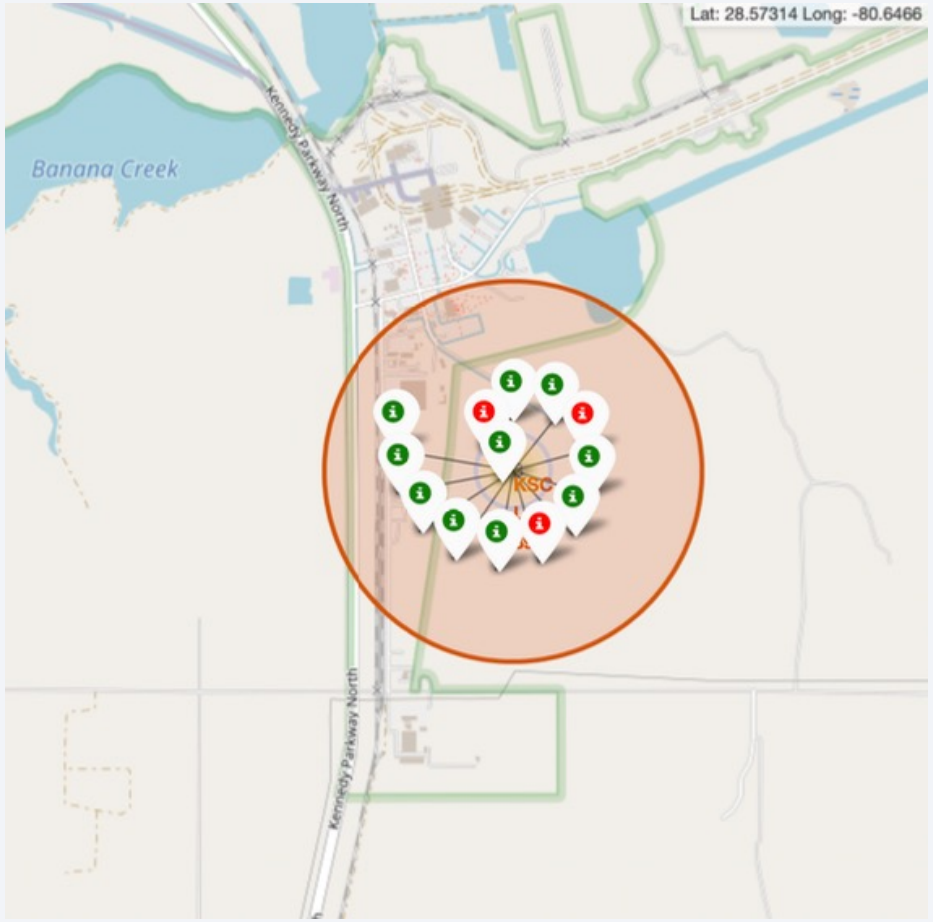
Color-labeled launch records map

By observing the color-labeled markers, it becomes evident which launch sites boast relatively high success rates.

- A Green one signifies a successful launch.

- A Red one indicates a failed launch.

- Notably, Launch Site KSC LC-39A exhibits an exceptionally high success rate, as suggested by the markers.



distance from the launch site ksc lc-39a to its proximities

i. Upon visual analysis of Launch Site KSC LC-39A, it's evident that it is in proximity to various features:

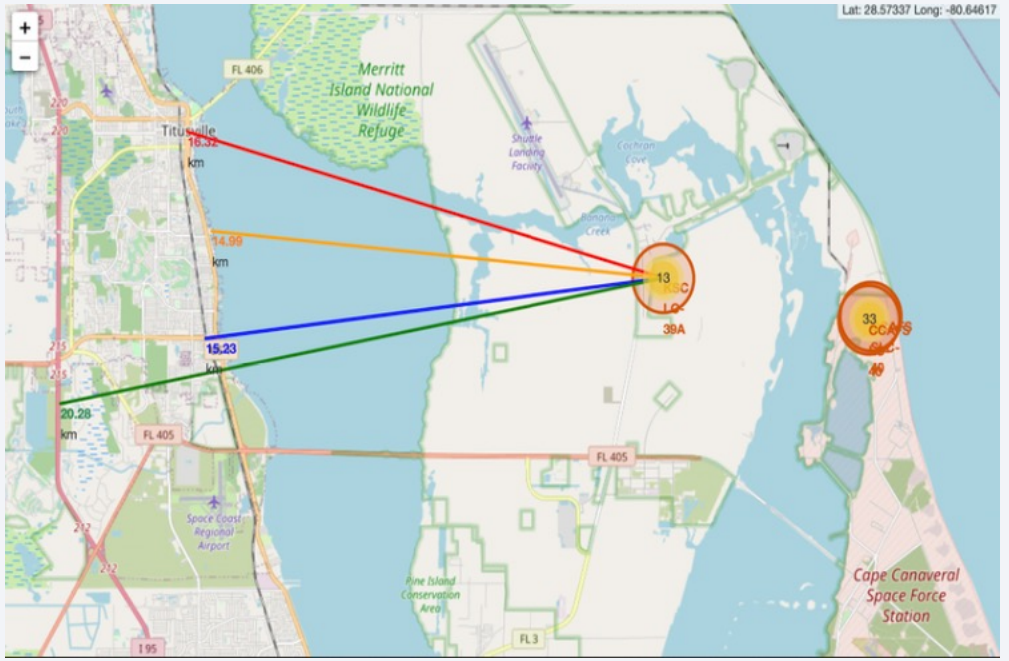
- It is relatively close to a railway, approximately 15.23 km away.

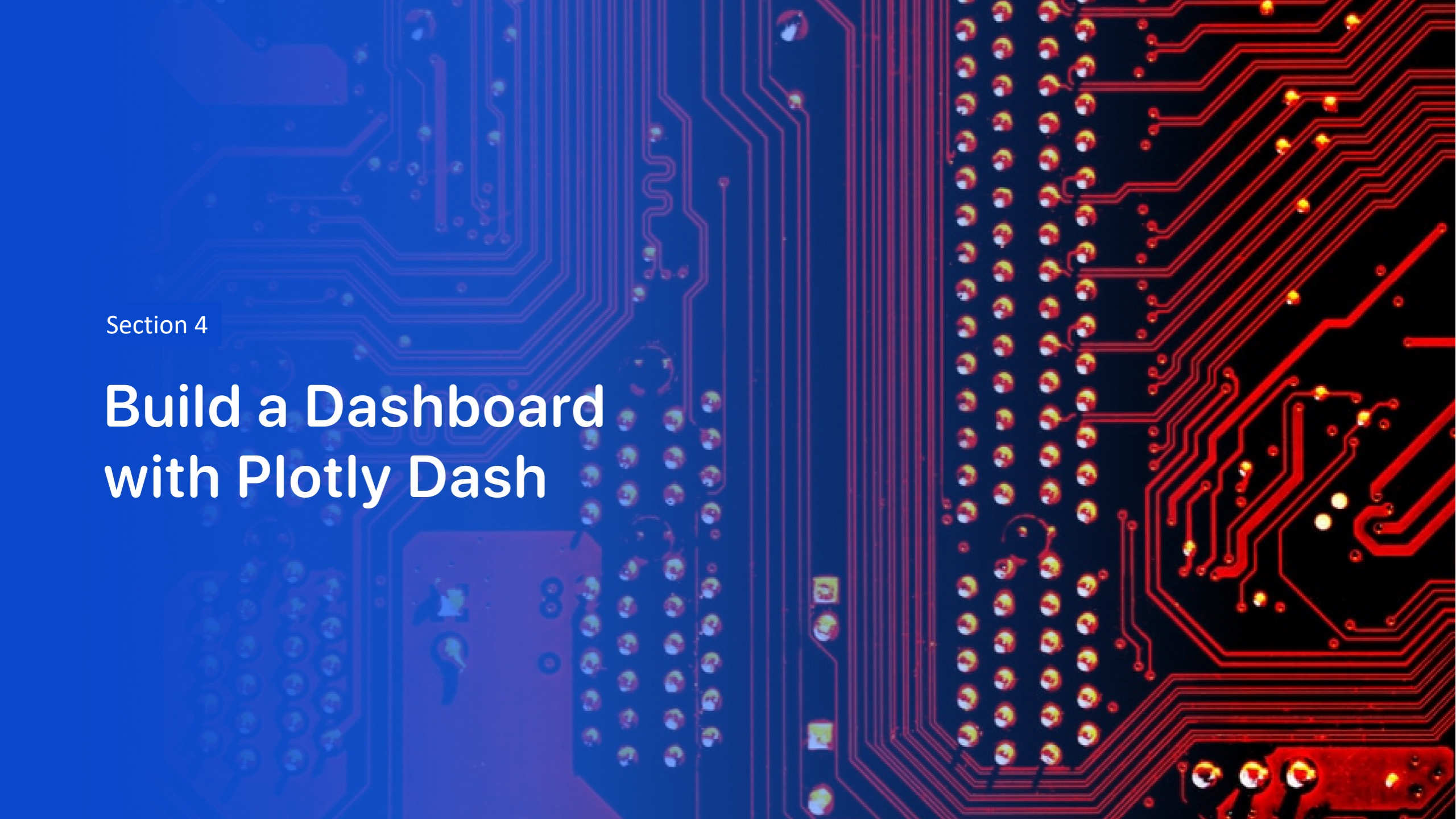
- It is also relatively close to a highway, with a distance of about 20.28 km.

- Its proximity to the coastline is approximately 14.99 km.

ii. Additionally, Launch Site KSC LC-39A is relatively close to its nearest city, Titusville, at a distance of about 16.32 km.

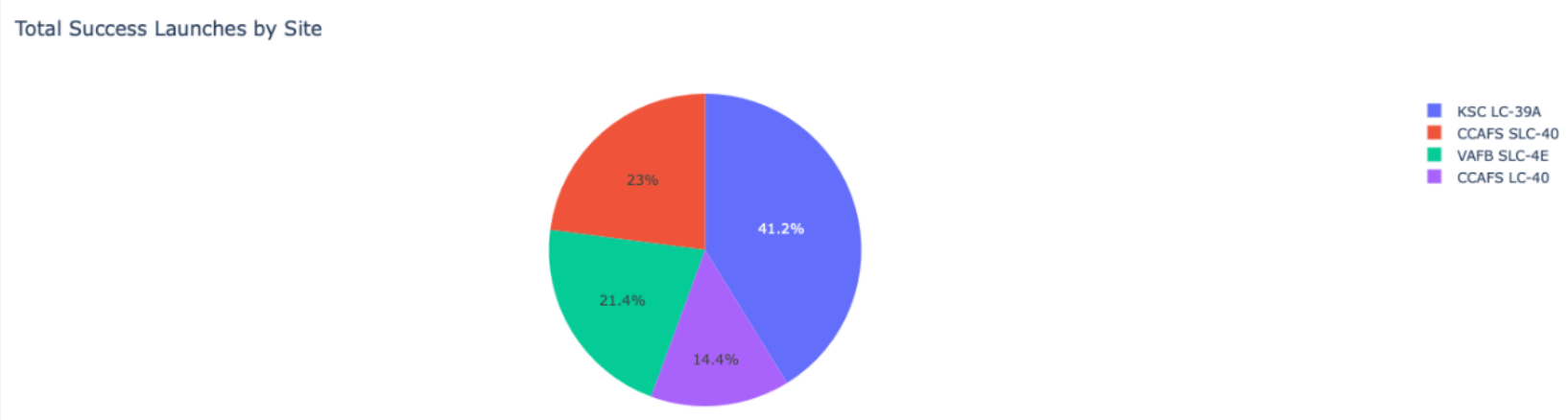
iii. It's worth noting that in the event of a failed rocket launch, a high-speed rocket could cover distances of 15-20 km in just a few seconds, which could pose potential risks to populated areas. Therefore, the proximity of launch sites to these features is a crucial safety consideration.




Section 4
Build a Dashboard with Plotly Dash

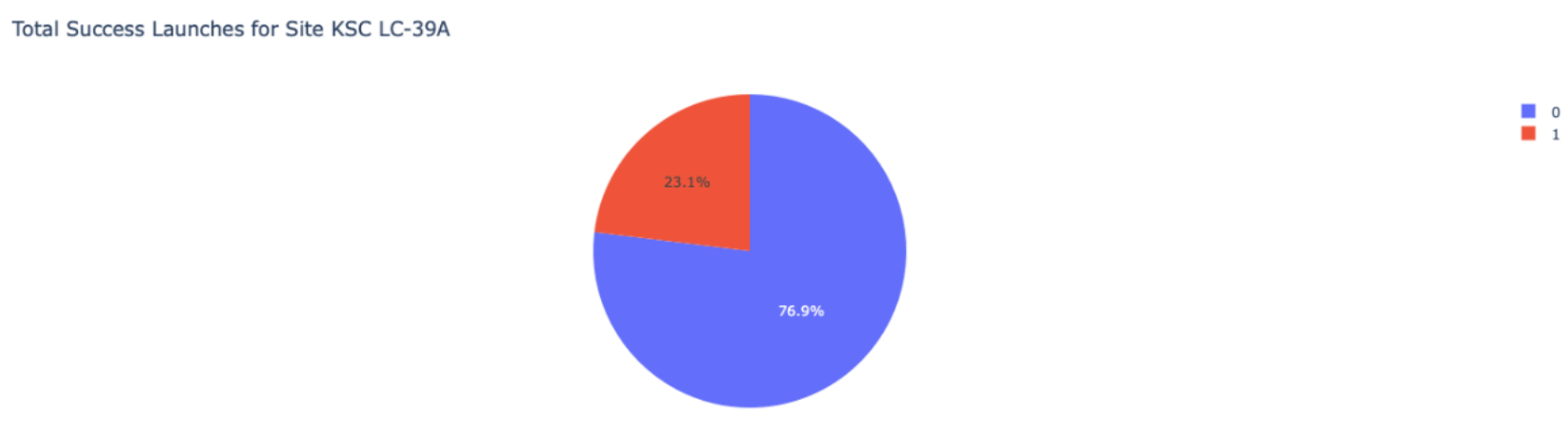
Successes Count

- KSC LC-39A has the most successes and CCAFS LC-40 has the least



Highest Successful Launch Rate

- KSC LC-39A has the highest success rate which is 10 out of 13.



Payload mass x success rate

- Payloads between 2000 – 4000 kg has higher success rate comparing to other payloads.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- It's hard to tell which method is the best with the test sets only
- From the accuracy of the complete data set we can tell that Decision Tree model has the best performance

Test Set

Out[30]:

	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

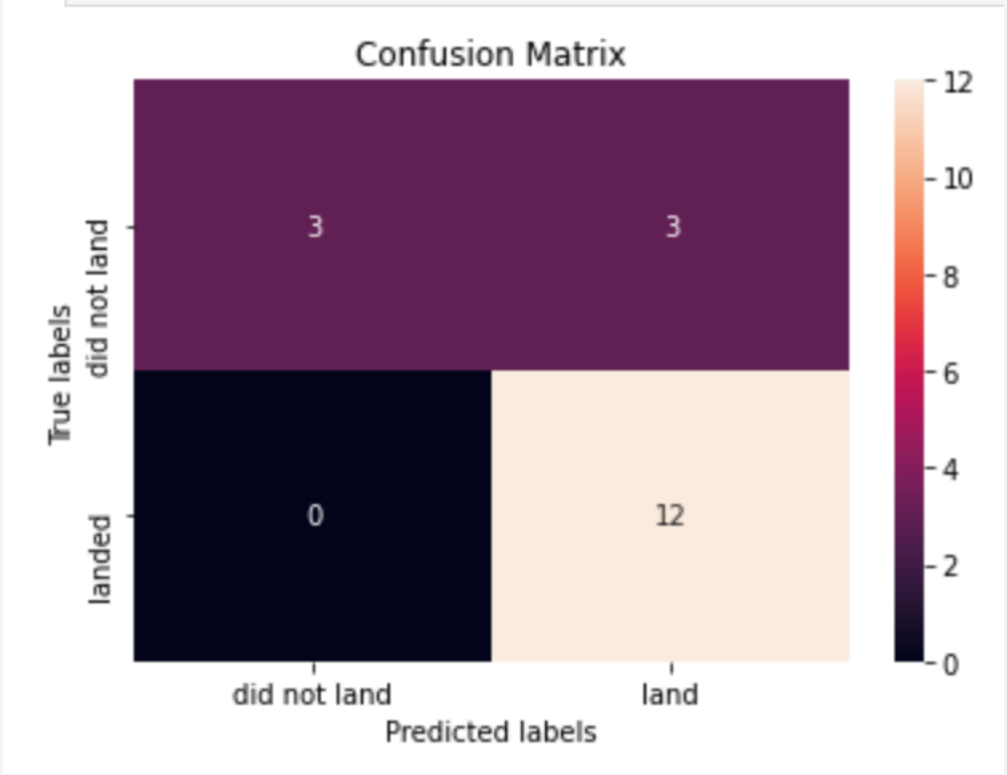
Complete Data Set

Out[31]:

	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

- Upon a comprehensive examination of the confusion matrix, it becomes apparent that the logistic regression model exhibits a notable ability to effectively differentiate between different classes within the dataset. However, a noteworthy pattern emerges from the analysis, indicating that the predominant challenge lies in the occurrence of false positives. This suggests that while the model excels at identifying positive outcomes, it may sometimes overestimate them, resulting in a higher rate of false positive predictions. Understanding and addressing this pattern is crucial for refining the model's predictive accuracy and minimizing false positive errors.



Conclusions

- The Decision Tree Model stands out as the optimal algorithm for effectively handling the dataset's complexities.

- A discernible trend within the data reveals that launches with lower payload masses consistently yield more favorable outcomes compared to launches with larger payload masses.

- Geospatially, a notable observation is that the majority of launch sites are strategically located in close proximity to the Equator, harnessing the Earth's rotational speed for launch efficiency. Furthermore, it is noteworthy that all launch sites are situated in extremely close proximity to coastal regions, minimizing potential risks associated with rocket launches near populated areas.

- An encouraging trend emerges from the historical data, indicating a gradual increase in the success rate of launches over the years, underscoring the industry's progress and maturation.

- Among all launch sites, KSC LC-39A consistently demonstrates the highest success rate, positioning it as a standout location for space missions.

- Specifically, in the case of orbits ES-L1, GEO, HEO, and SSO, a remarkable achievement is noted, with a 100% success rate, emphasizing the effectiveness of space missions targeting these specific orbital paths.

Thank you!