IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matthew O'Connor
19 February 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualisation
  - Exploratory Data Analysis with SQL
  - Building interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive Analytics
  - Predictive Analysis Results

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on their website at a cost of 62 million dollars. Competitors provide a cost of rocket launches in the region of 162 million dollars and above. Much of the savings are accounted for by the reusability of the first stage of the rocket launch. Therefore, if it can be determined whether or not the first stage will land then the cost of the launch can also be determined. Through using publicly available information and machine learning models, this project aims to predict the success of the first stage of SpaceX rocket launches.

- Questions to be answered
  - What variables have the biggest impact on the success of the first stage landing?
  - Does the rate of success increase over time?
  - Which algorithm can be best used to determine the success of the first stage landing?

Section 1

Methodology

Methodology

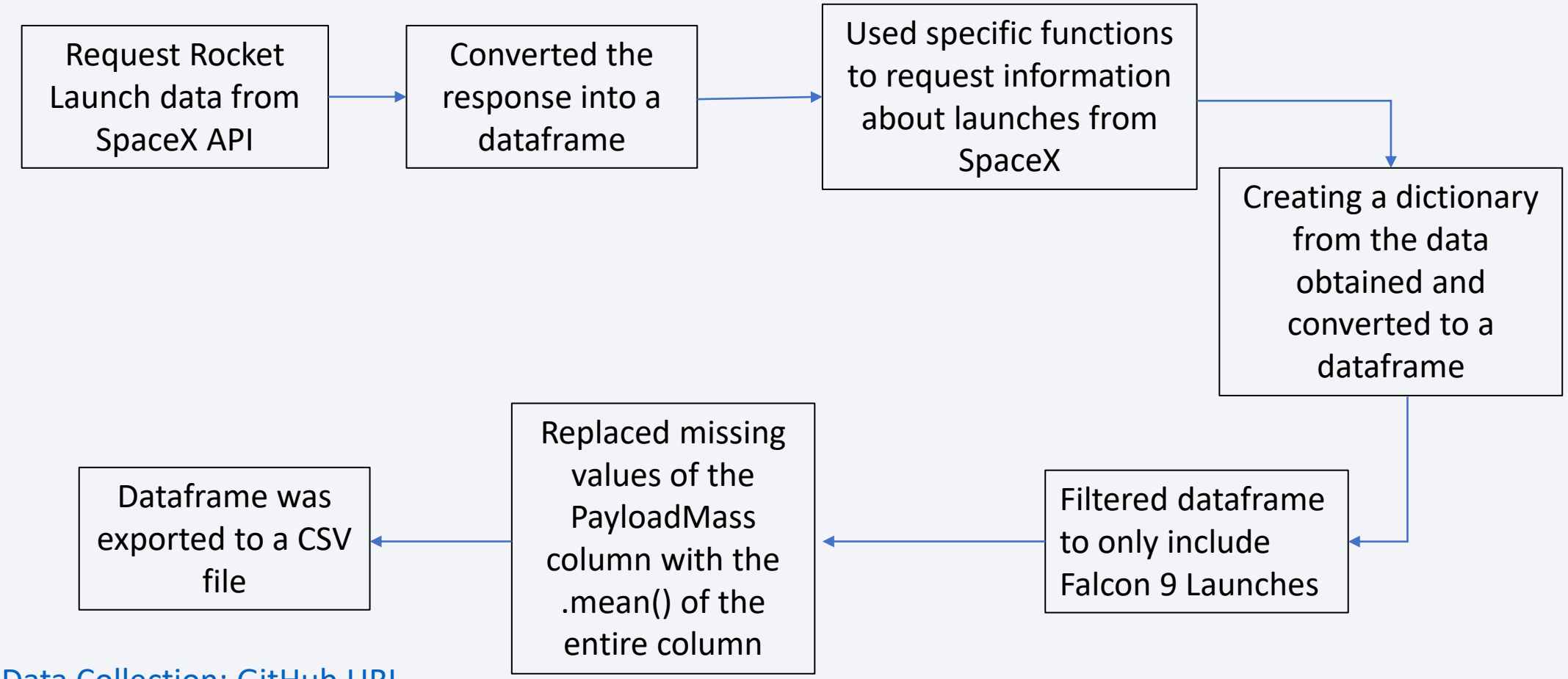
Executive Summary

- Data collection methodology:
  - Data collection was undertaken via a request of the SpaceX API and through web scraping of SpaceX data from Wikipedia.
- Perform data wrangling
  - Filtering data, dealing with missing and invalid values, One Hot Encoding.
- 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 evaluating classification models to find the greatest level of agreement.

Data Collection

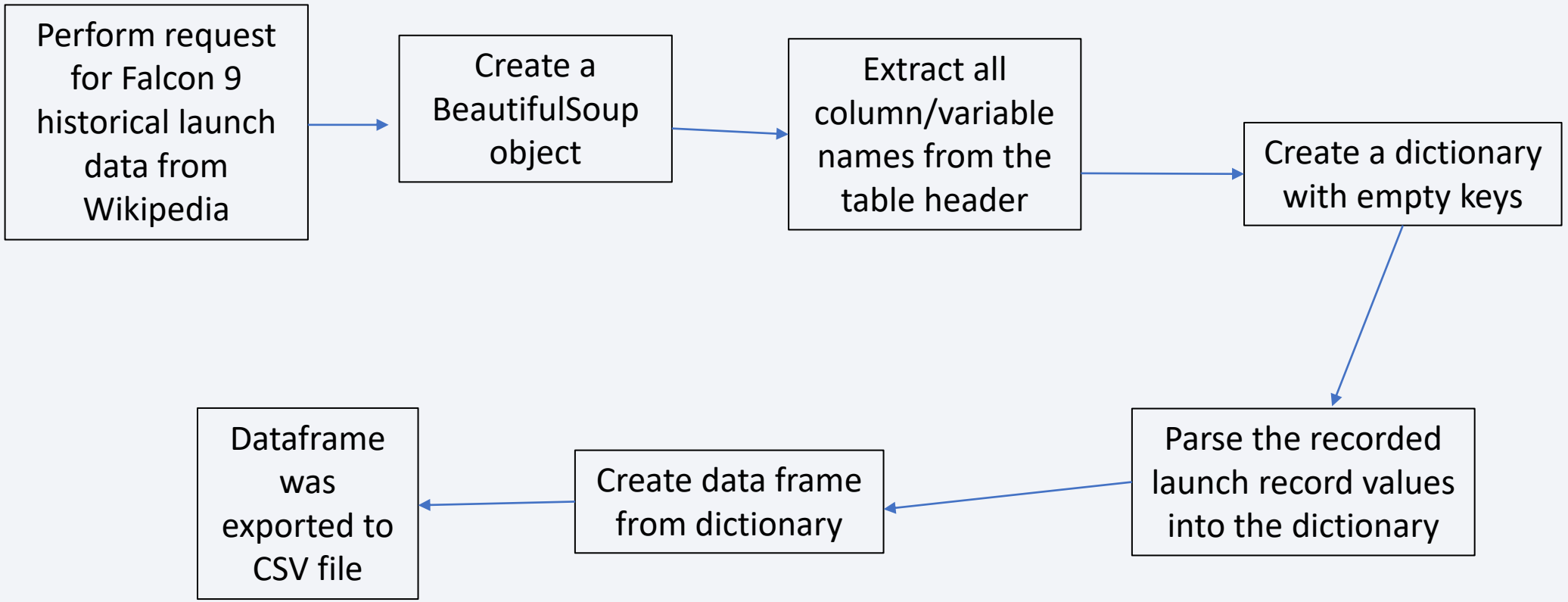
- Data collection involved API requests from SpaceX REST API and Web Scraping from the SpaceX's Wikipedia.
- The data was collected and put into a data frame using a normalization method.
- Data Columns obtained by using SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- Data Columns obtained by Wikipedia Webscraping:
  - FlightNo., LaunchSite, Payload, PayloadMass, Orbit, Customer, LaunchOutcome, Version Booster, BoosterLanding, Date, Time.

Data Collection – SpaceX API



Data Collection: GitHub URL

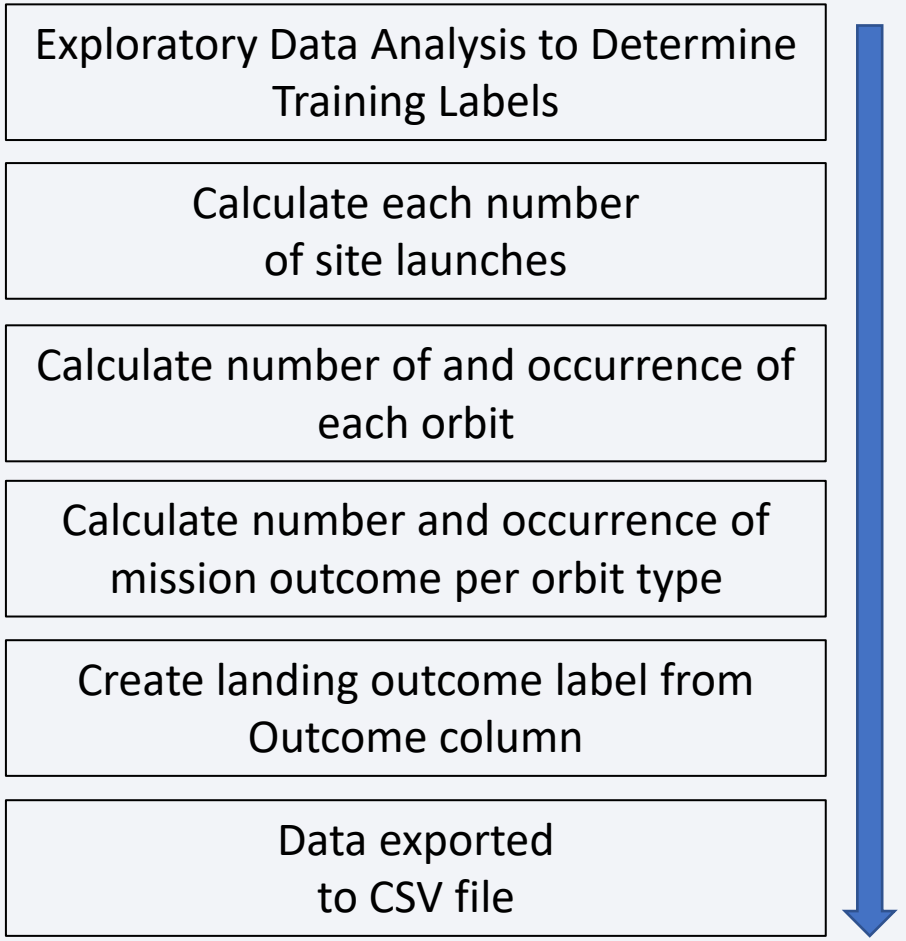
Data Collection - Scraping



Data Collection: GitHub URL

Data Wrangling

In this data set, there are several times where the booster did not land successfully. Sometimes a landing was attempted by 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.



Data Wrangling: GitHub URL

EDA with Data Visualization

- Plotted Charts:
  - Flight Number vs Launch Site, Payload vs Launch Site, Orbit Type vs Success Rate, Flight Number vs Orbit Type, Payload vs Orbit Type, Launch Success Rate and Yearly Trend
- Scatter Plots were used to show the relationship between the variables. Machine learning models can be used to determine whether a relationship exists between variables.
- Bar charts allow comparison between discrete variables. They aim to compare between specific categories and a specific value.
- Line charts (generally) show changes over time.

EDA with Data Visualisation: GitHub URL

EDA with SQL

SQL Queries Performed:

- Display names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with ‘CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List names of successful boosters in drone ships which have payload mass >4,000 and <6,000
- List total number of successful and failed mission outcomes
- List the names of booster versions which have carried the maximum payload mass
- List failed landing outcomes in drone shop, their booster versions and launch site names for 2015
- Rank the count of landing outcomes or success between 2010 and 2017

EDA with SQL: GitHub URL

Build an Interactive Map with Folium

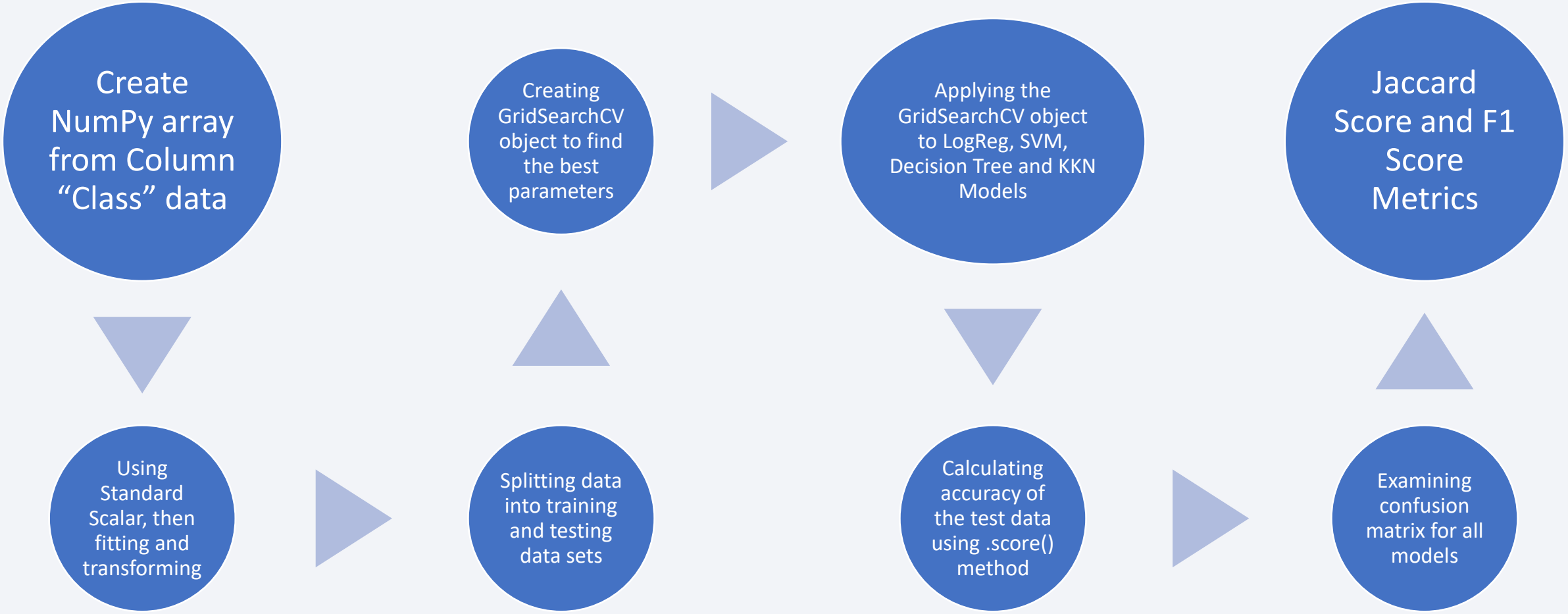
- Launch Site Markers:
  - Added Markers with Circle, Popup, Label and Text Label of NASA Johnson Space Center using latitude and longitude values as origin.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using latitude and longitude coordinates for approximate geographical locations.
- Colour Markers of launch outcomes for Launch Sites:
  - Added coloured markers of success (Green) and failed (Red) launches using Marker Cluster to identify launches with high success rates.
- Distances between a Launch Site to proximities:
  - Added Lines to show distances between Launch Sites to notable surrounding sites and identifying infrastructure.

Interactive Visual Analytics with Folium: GitHub URL

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Sites):
  - Added pie chart to show total successful launches count for all sites and the Success vs. Failed counts for the site when a specific Launch Site was selected.
- Slider of Payload Mass Range:
  - Added slider to select Payload range.
- Scatter Chart of Payload Mass vs. Success Rate for different Booster Versions:
  - Added scatter chart for the correlation between Payload and Launch Success

Predictive Analysis (Classification)



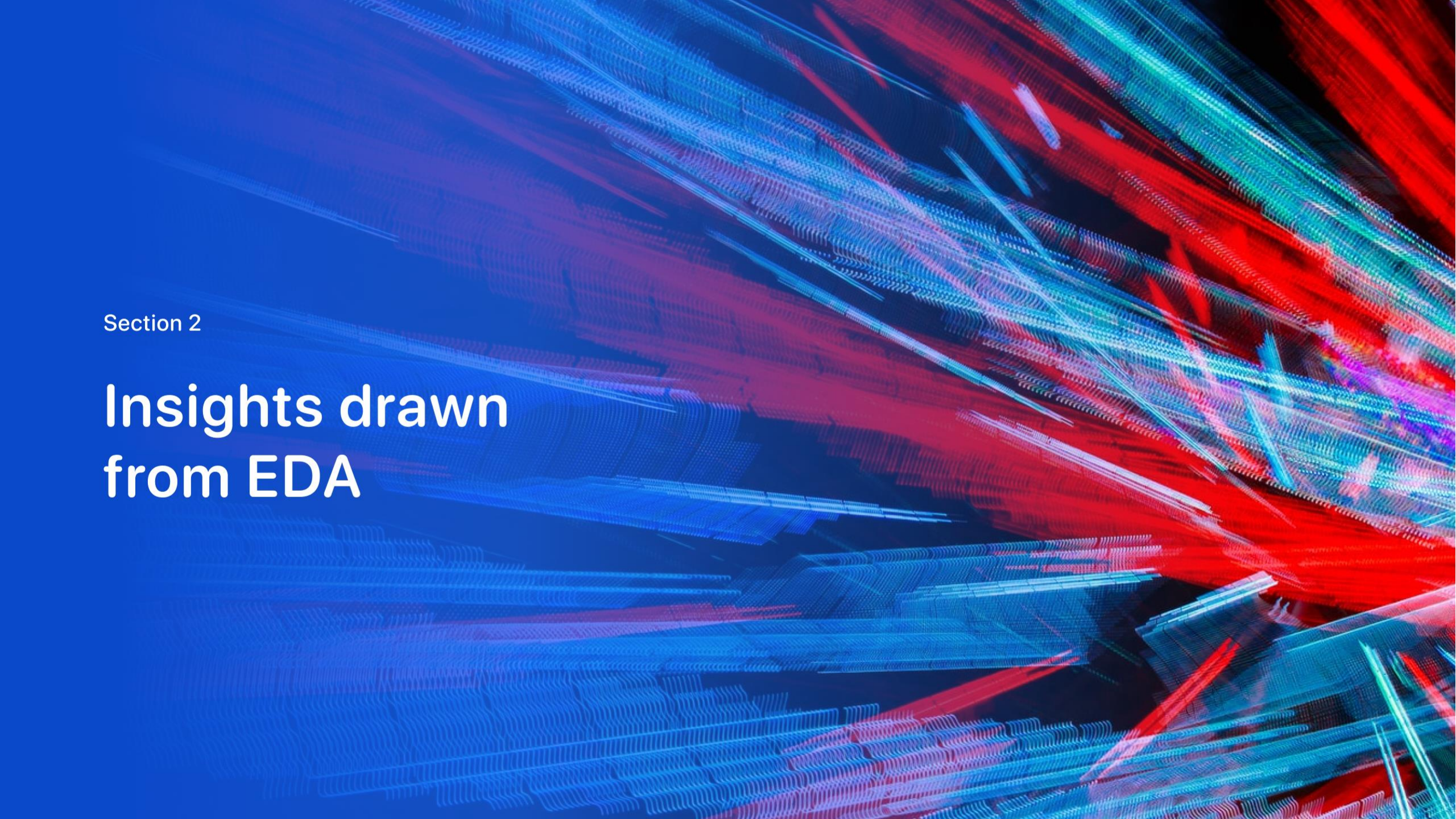
Machine Learning Prediction: GitHub URL

Results

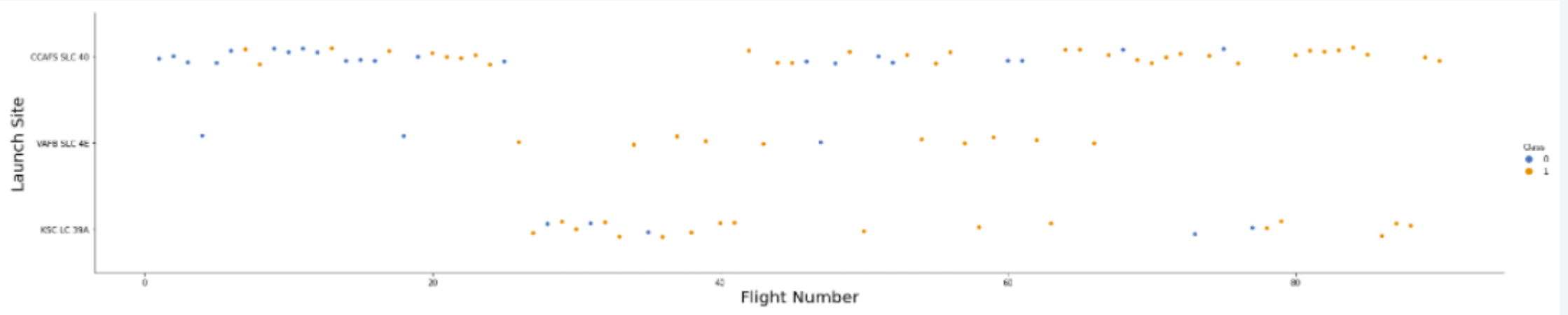
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

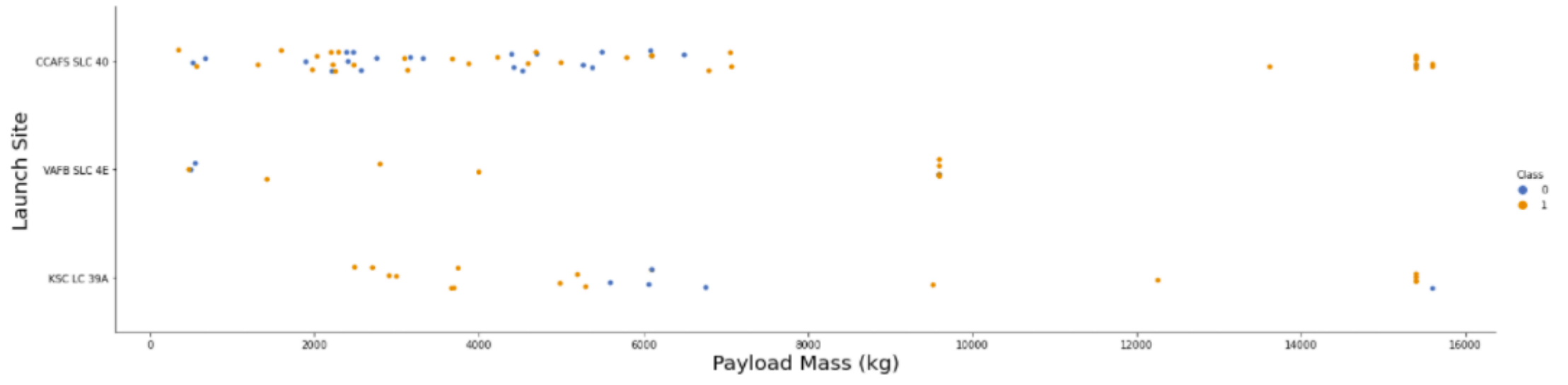


Flight Number vs. Launch Site



- Earlier flights had a higher rate of failing with more recent flights having a higher rate of succeeding.
- CCFAS SLC 40 launch site occupied around half of the launches.
- VAFB SLC 4E and KSC LC 39A had higher rates of success.
- Increase in launch number increases the success of future launches.

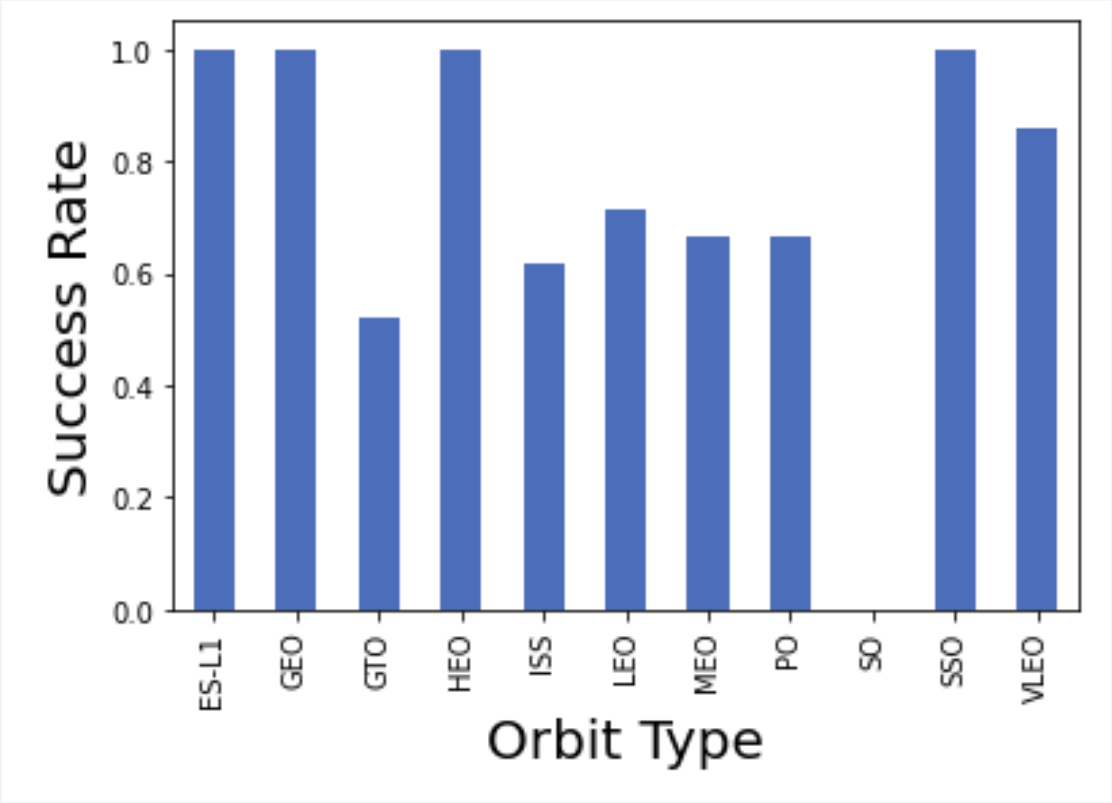
Payload vs. Launch Site



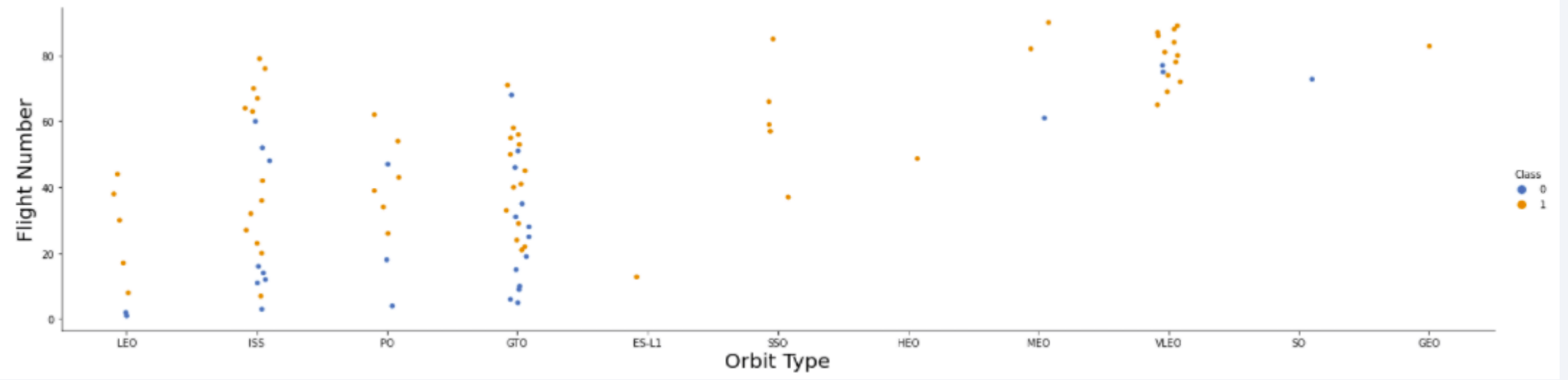
- Higher payload mass resulted in a higher success rate.
- Majority of launches over 7,000 kg payload mass were successful.
- KSC LC 39A resulted in a 100% launch success rate with a payload mass <5,500 kg.

Success Rate vs. Orbit Type

- Orbits with a 100% success rate:
  - ES-L1, GEO, HEO, SSO.
- Orbits with 0% success Rate:
  - SO.
- Orbits with varying success rates:
  - GTO, ISS, LEO, MEO, PO

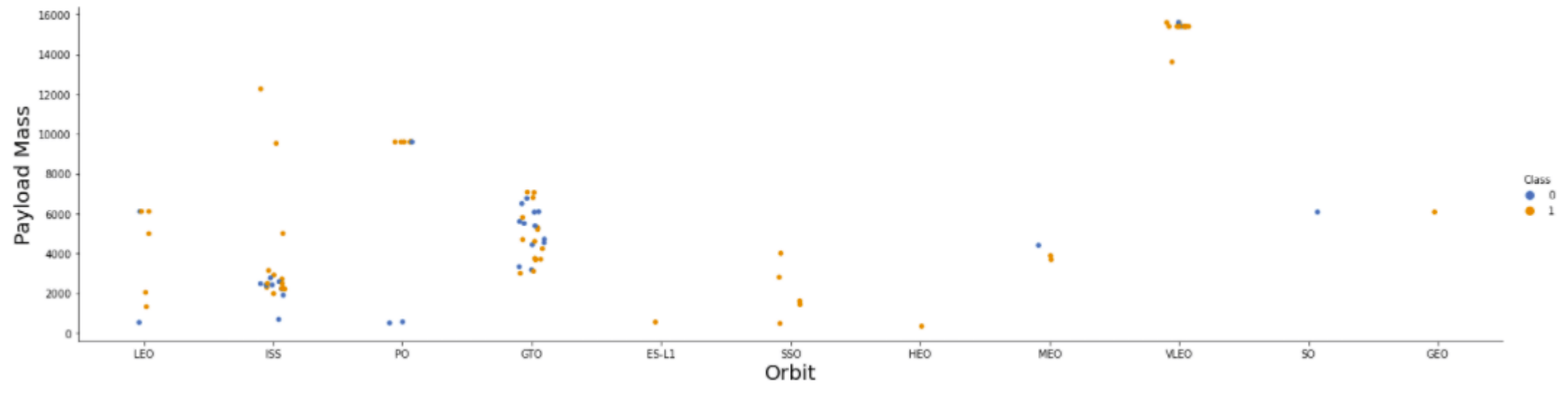


Flight Number vs. Orbit Type



- LEO orbit appears to have higher success with related with the number of flights. There is no apparent relationship between flight number when in GTO orbit.

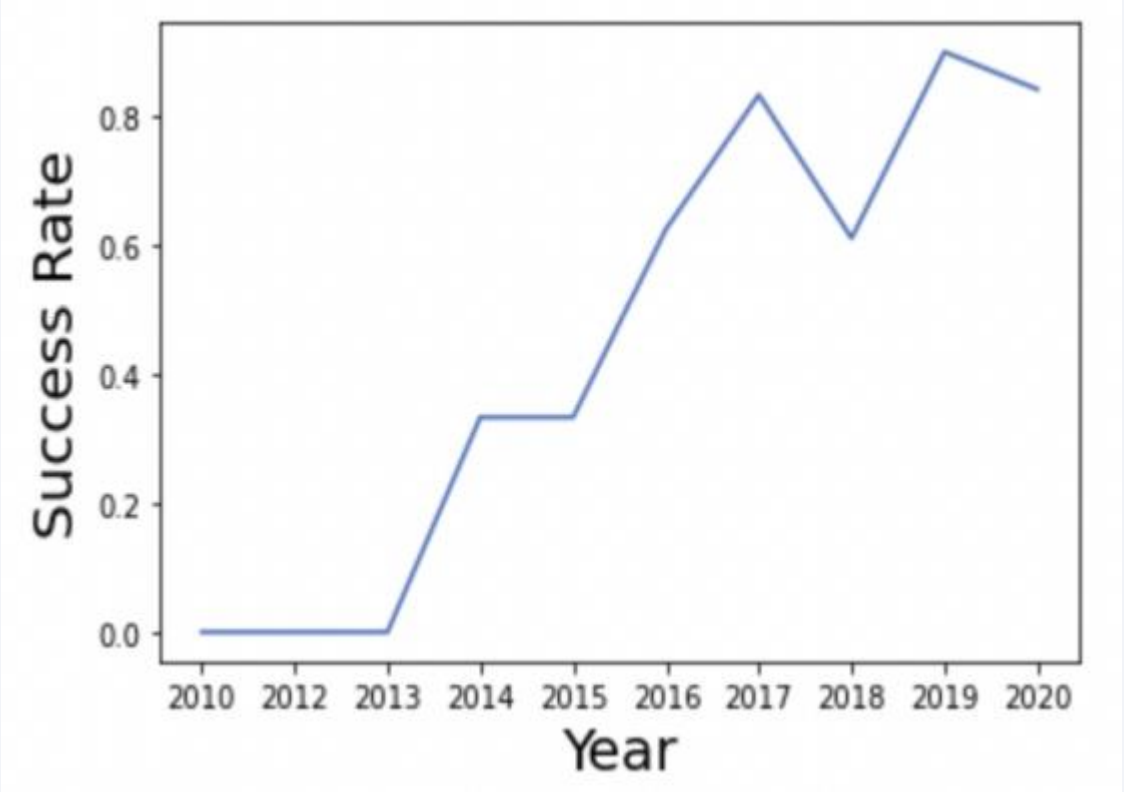
Payload vs. Orbit Type



- Heavier payloads negatively impact the GTO orbits, where as heavier payloads have a positive impact on VLEO, PO, LEO and ISS orbits. More data required for the remaining orbits.

Launch Success Yearly Trend

- General trend for launch success has been increasing since 2013 to 2020, with slight dips around 2017-2018.



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT launch_site FROM SPACEXTBL;
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

- Unique launch site names were collected with a SELECT DISTINCT query in SQL.

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' ORDER BY launch_site LIMIT 5;
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/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)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	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
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
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)

- First 5 records beginning with ‘CCA’ were found using a select query SQL.

Total Payload Mass

Task 3

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

```
%sql SELECT sum(payload_mass__kg_) as total_payload FROM SPACEXTBL WHERE customer LIKE 'NASA (CRS)%' Group By customer;
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

total_payload
45596

- Total payload mass was found using a select query in SQL while ordering by the customer (in this case, it was NASA).

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql SELECT AVG(payload_mass__kg_) as mean_payload FROM SPACEXTBL WHERE booster_version= 'F9 v1.1';
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

mean_payload
2928

- Average payload mass was found using an AVG() query in SQL for the booster version F9 v1.1

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
%sql SELECT landing__outcome, date FROM SPACEXTBL WHERE landing__outcome= 'Success (ground pad)' ORDER BY date ASC LIMIT 1;
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

landing__outcome	DATE
Success (ground pad)	2015-12-22

- First successful ground landing date was found using a select query in SQL.

Successful Drone Ship Landing with Payload between 4,000 and 6,000

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 landing__outcome, payload_mass__kg_ FROM SPACEXTBL WHERE landing__outcome= 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 A
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

landing__outcome	payload_mass__kg_
Success (drone ship)	4600
Success (drone ship)	4696
Success (drone ship)	5200
Success (drone ship)	5300

- Successful drone ship landings between the weights of 4,000 and 6,000 were found using a select query in SQL.

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(*) FROM SPACEXTBL GROUP BY mission_outcome;
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Mission outcomes (success/failures) were found with a select query in SQL.

Boosters Carried Maximum Payload

Task 8

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

```
%sql SELECT booster_version FROM SPACEXTBL WHERE payload_mass__kg_=(SELECT MAX(payload_mass__kg_) from SPACEXTBL);
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/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

- Booster versions that carried maximum payload mass were found using a select query and subquery with SQL.

2015 Launch Records

Task 9

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

```
%sql SELECT booster_version, landing__outcome, launch_site, date FROM SPACEXTBL WHERE landing__outcome LIKE 'Fail%' AND date LIKE '2015%';
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

- Failed landing outcomes and their respective launch site and launch date were found using a select query in SQL.

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

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

```
%sql SELECT LANDING__OUTCOME,count(LANDING__OUTCOME)as LANDING_OUTCOME_COUNT from SPACEXTBL where DATE between '2010-06-04' and '2017-03-20' group b
```

```
 * ibm_db_sa://xmw21109:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

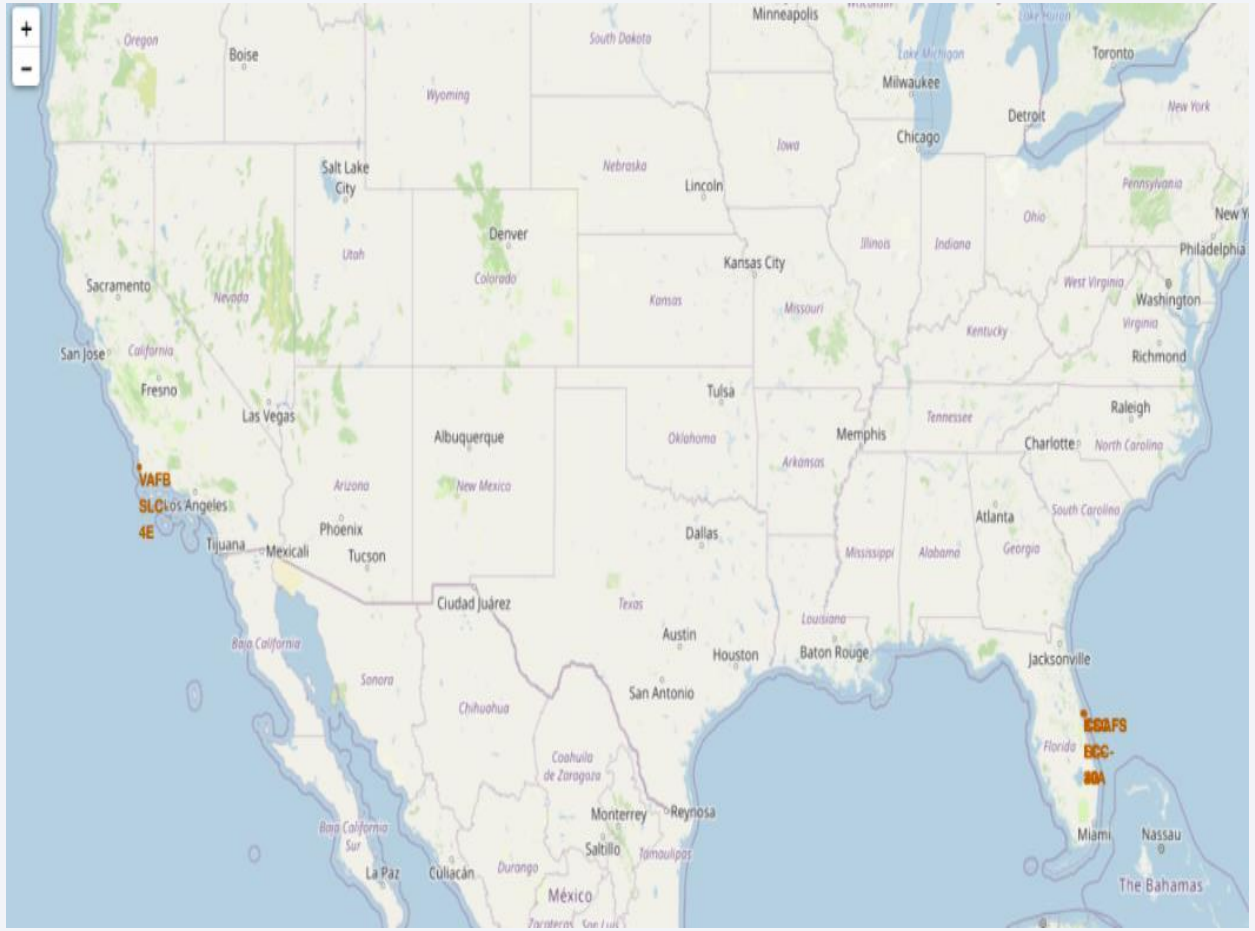
landing__outcome	landing_outcome_count
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1
Success (drone ship)	5
Success (ground pad)	3
Uncontrolled (ocean)	2

- Land outcomes from 2010 to 2017 were grouped using a select query in SQL.

Section 4

Launch Sites Proximities Analysis

Launch Site Locations within the United States



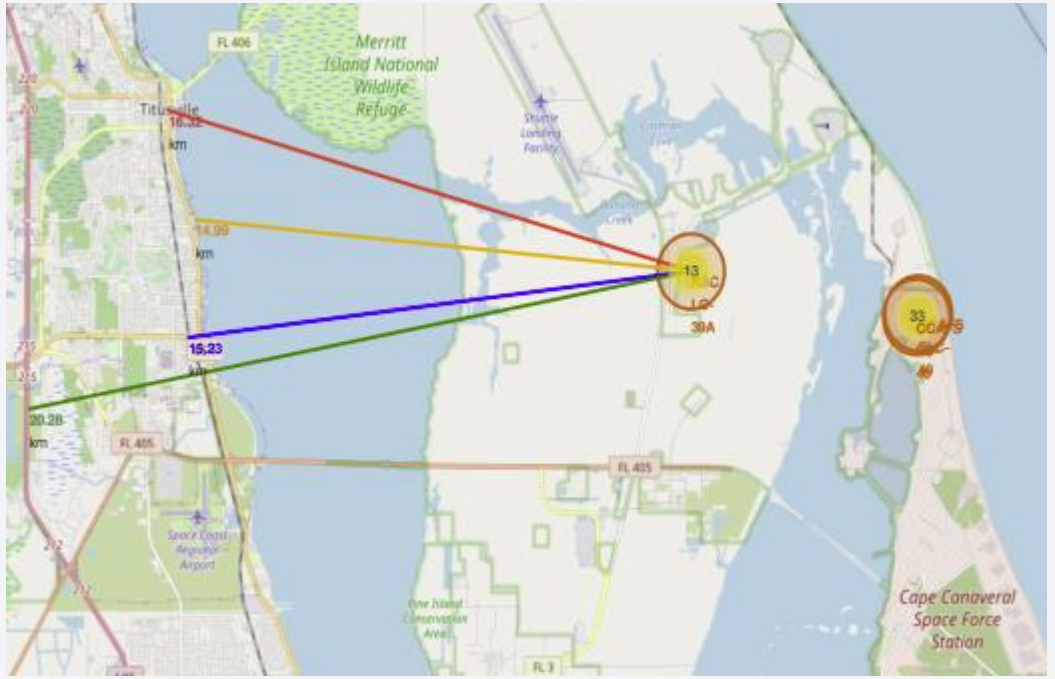
- Both launch sites are located within close proximity to the coastline. Any further inland and it would become unsafe for launches due to the risk of the public. Having the launch sites closer to the ocean means that if things do fail then the likelihood of causalities or further damage would be low.
- Launching rockets closer the equator allows the spin of the earth to further propel the rocket forward. Anything near the equator is already moving at 1,670 km/h. If a ship is launched from the equator and ends up going to space it will also be moving around the Earth at the same speed it was moving before launching due to inertia. This helps keep the aircraft in orbit.

Successful Launch Site – KSC LC-93A

- Coloured labels indicate the success of the launch.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-93A was shown to have a very high launch success rate.

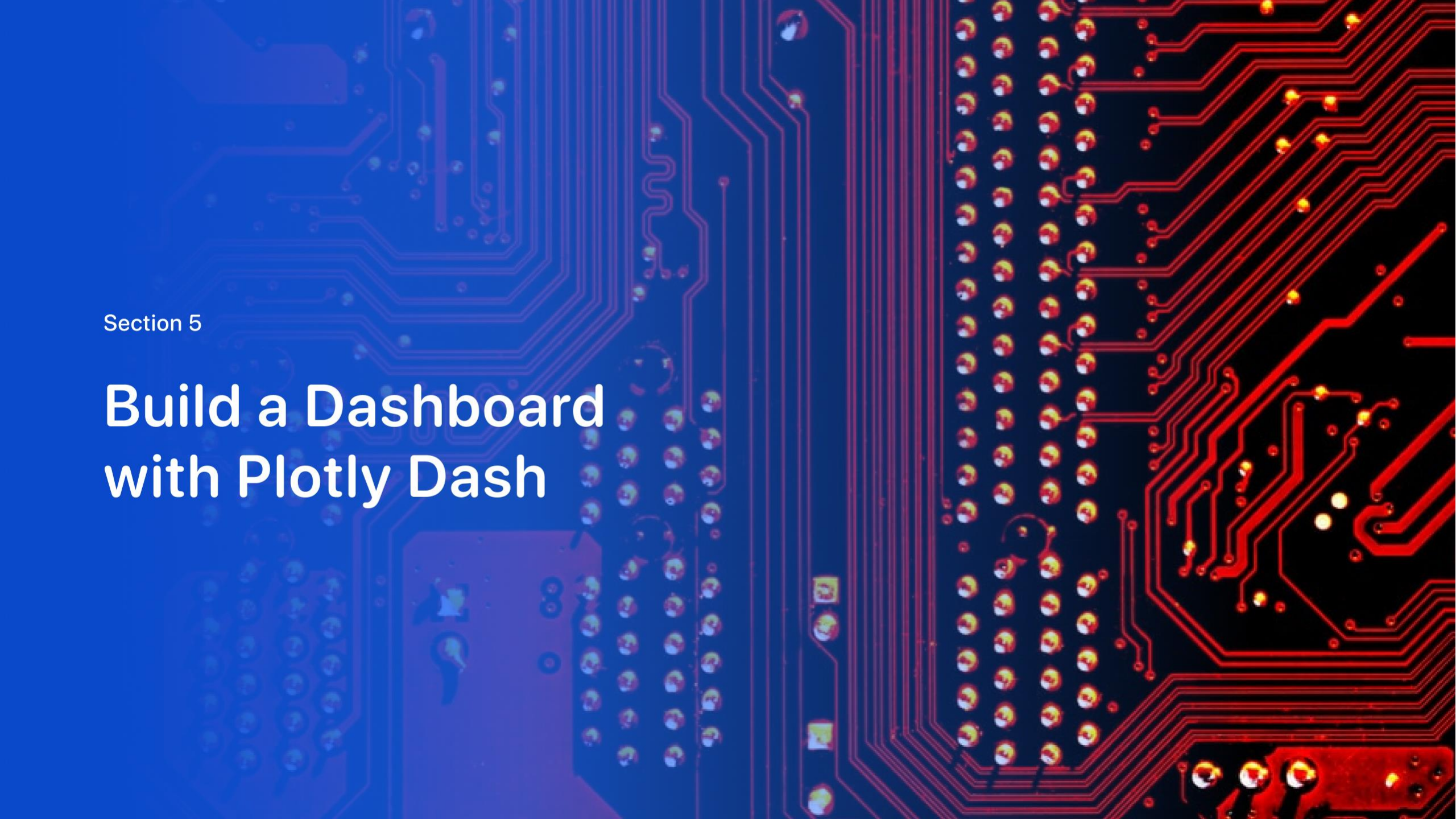
Launch Location KSC LC-39A And Surrounds

- Launch Location KSC LC-39A is located within proximity to several built areas and infrastructure:
  - Closest railway – 15.23km
  - Closest highway – 20.28km
  - Closest public coastline – 14.99km
- Due to the velocity of rocket launches, these areas are potentially at risk of failed launches.
- Additionally, the nearby city of Titusville is located 16.32km. A failed launch would be catastrophic if directed at the city.

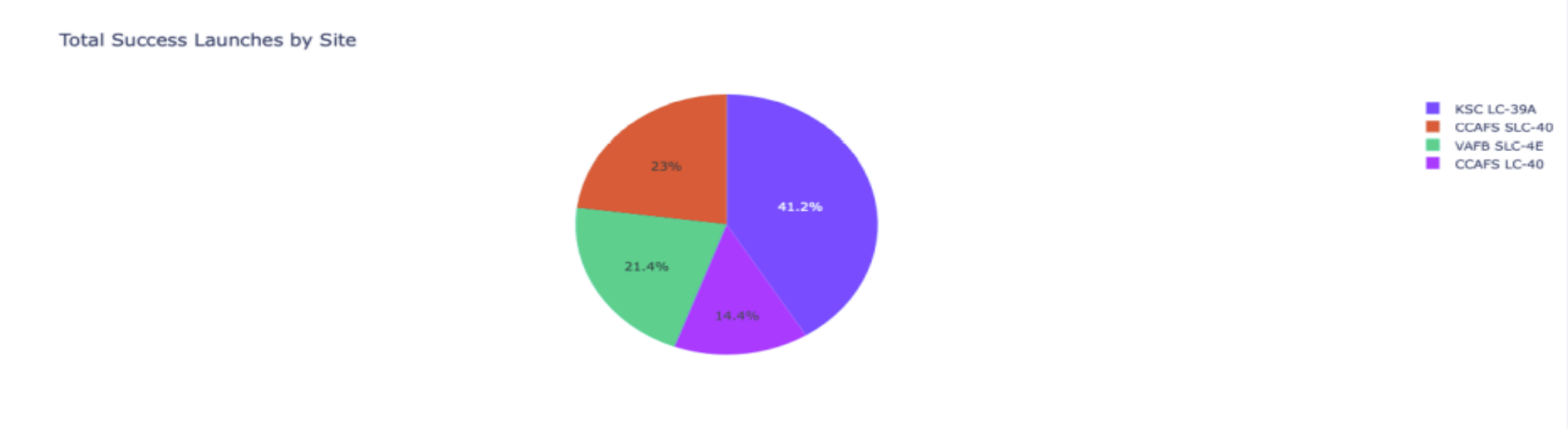


Section 5

Build a Dashboard with Plotly Dash

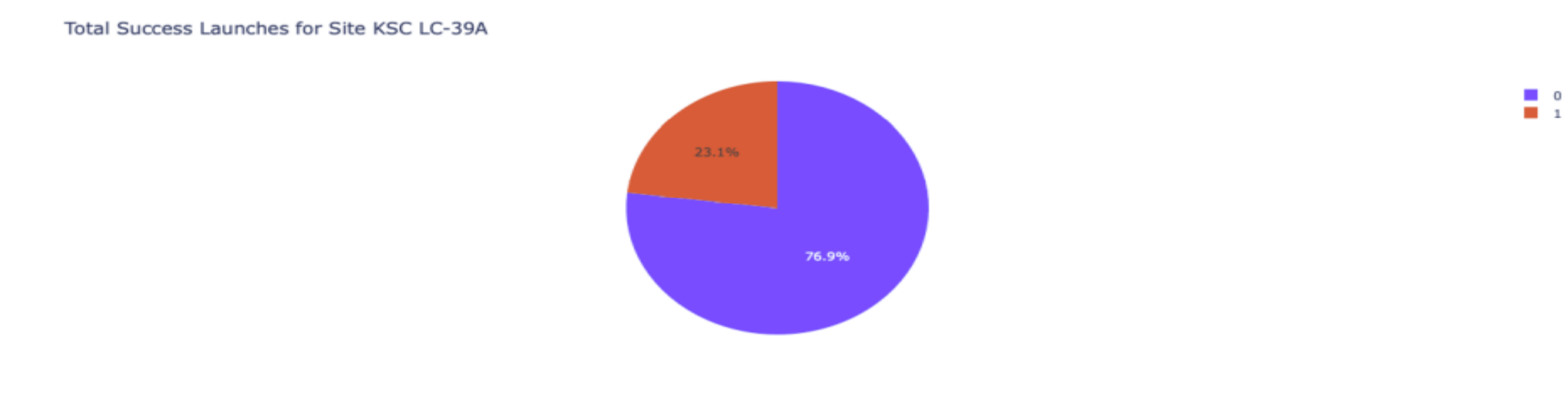


Launch Success for All Launch Sites



- This pie chart shoes the success rate for all launch sites. KSC LC-39A was shown to have the highest success rate for launches with 41.2%.

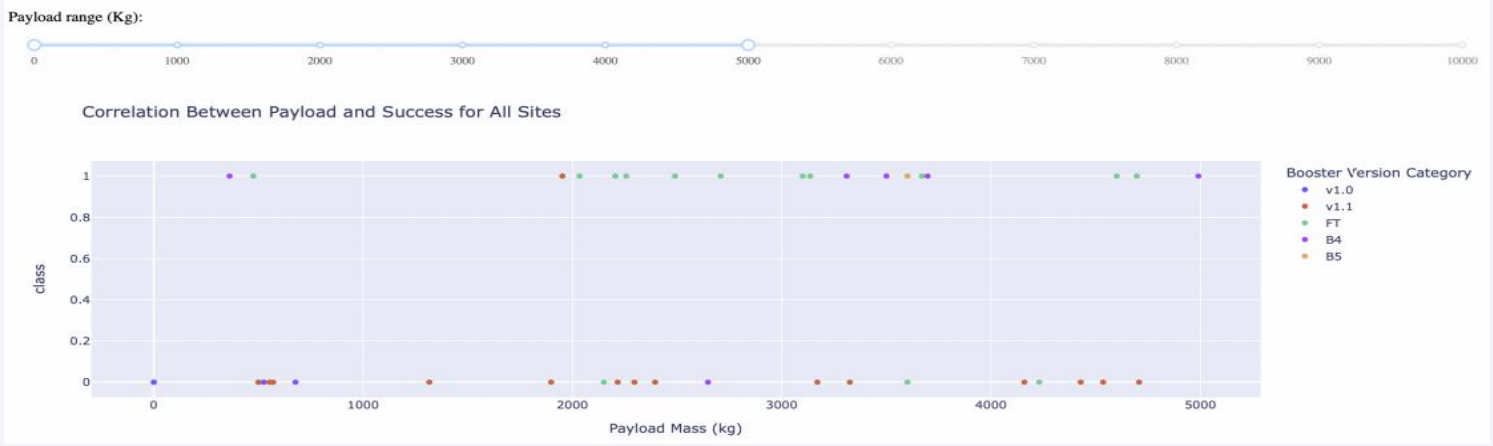
Highest Launch Success Rate



- This pie chart shows the total success launches for Launch Sites KSC LC-39A (76.9%).

Payload Mass Compared to Launch Outcome

- This dashboard screenshot shows that payloads with a mass between 2,000 and 5,500 kg have the greatest launch success rate.



Section 6

Predictive Analysis (Classification)

Classification Accuracy

Accuracy of Test Data

	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

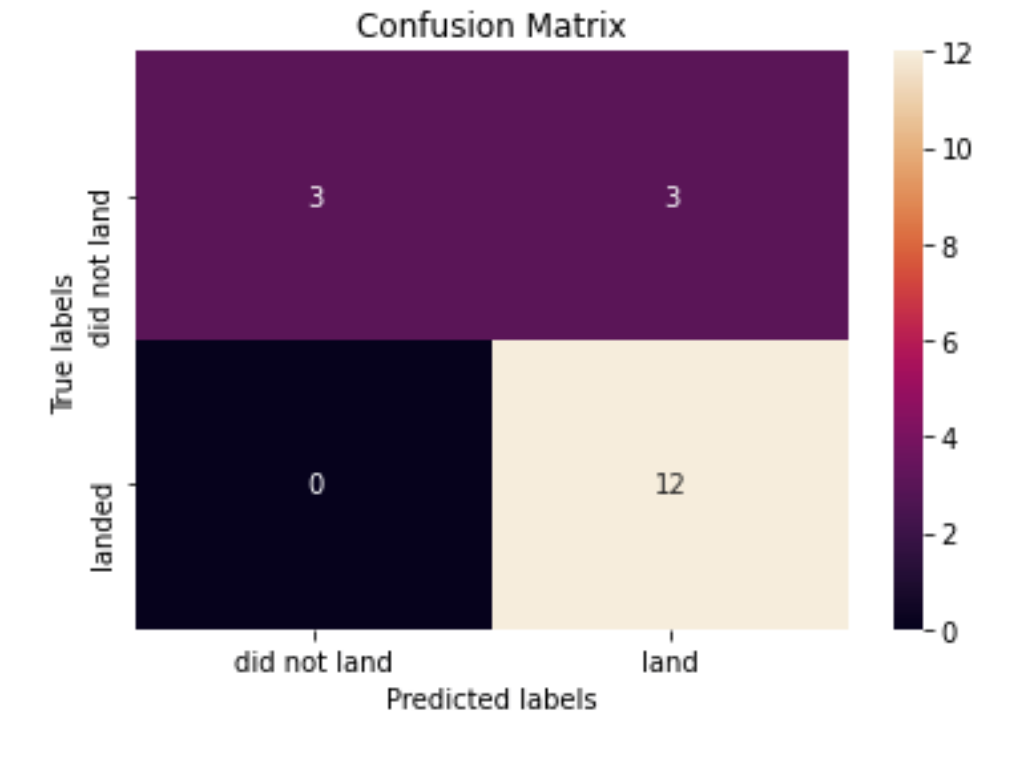
Accuracy of Data Set

	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

- The accuracy scores resultant from the test dataset were unable to confirm which model performed the best. This was likely due to the small sample size.

- The Decision Tree model resulted in the highest accuracy scores, meaning that it model is best suited for this data set.

Confusion Matrix



- Confusion matrix for the dataset shows that logistic regression can differentiated between the different classes. However, there were the same amount of false positives and true negatives.

Conclusions

- Decision Tree Model was the best algorithm for the test data set.
- Increase in payload mass resulted in higher success rate of launches.
- Launch sites are closer to the equator and are located directly adjacent the large water bodies.
- Success rate of rocket launches increased over time.
- Launch site KSC LC-39A had the highest launch success out of all launch sites.
- Orbit had a significant impact on launch success, in particular, ES-L1, GEO, HEO and SSO had 100% launch success rate.

Thank you!