IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AAYUSHI GUPTA
5TH JANUARY 2023

Outline

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

Executive Summary

- The objective of this project is to predict if the first stage of Falcon 9 rocket launch will be reused and conversely determining the cost of launch.
- To achieve this, data was collected and wrangled from SpaceX's API and wiki pages. Exploratory data analysis was performed on the data using SQL and visualization tools, Plotly Dash and Folium.
- Also, predictive analysis was performed using multiple classification algorithms to determine the best estimators.
- The results include
  - Interactive map of optimum launch sites
  - Dashboard depicting the success rates of each launch site
  - Confusion matrices signifying the accuracy of each classification algorithm.

Introduction

- A leading commercial space company, SpaceX has a staggering low figure for rocket launches as compared to its contemporaries.
- They offer inexpensive rocket launches as SpaceX reuses the first stage. Hence, if we can determine if the first stage will be reused, we can determine the cost of a launch.
- A new company, Space Y, wants to compete with SpaceX by understanding its strategy.
- This will be achieved by gathering information about Space X and creating dashboards.
- The project task is to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully.
- Data science methodologies have been applied to analyze launch data and optimize prediction of launches.

Section 1

Methodology

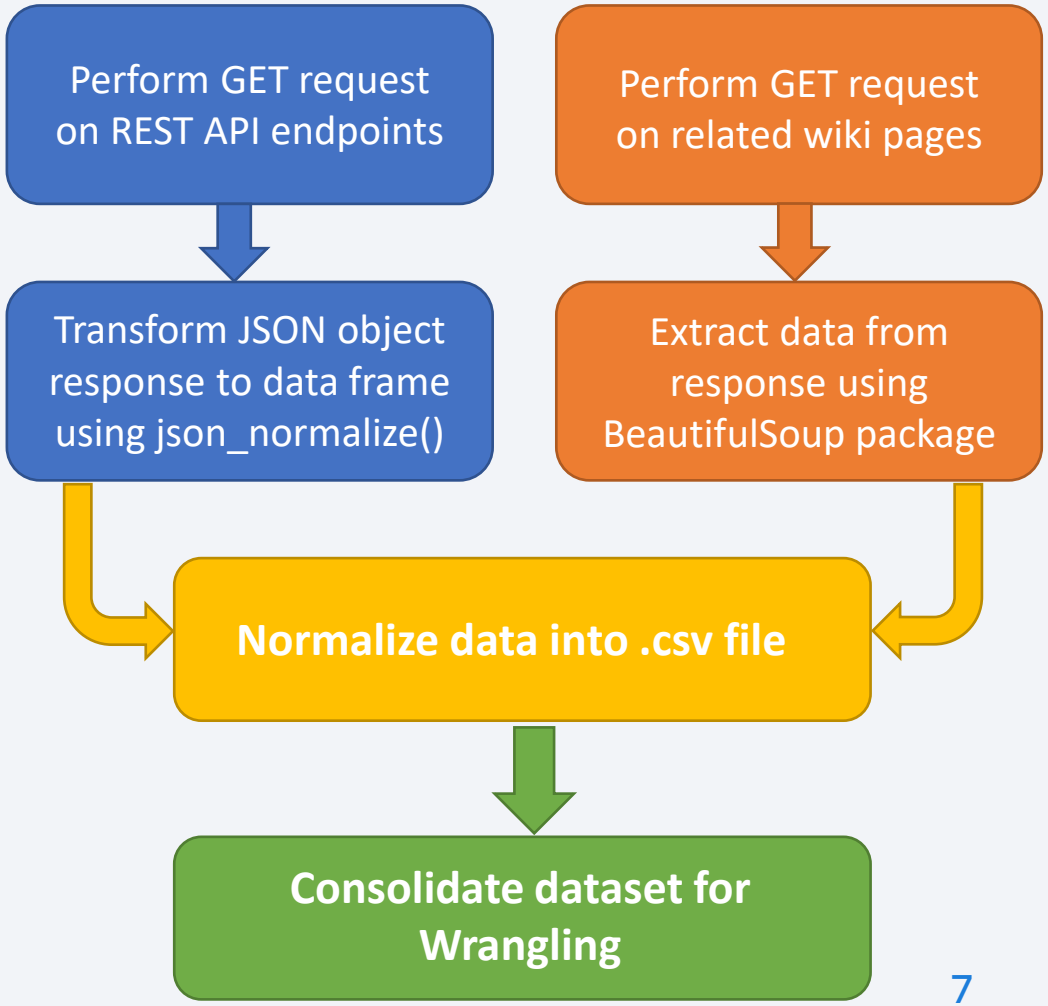
Methodology

Executive Summary

- Data Collection Methodology:
  - Consolidating data from different endpoints of SpaceX REST API
  - Web Scraping SpaceX Falcon 9 related wiki pages
- Perform Data Wrangling
  - Performed feature engineering on the data collected
- Perform Exploratory Data Analysis (EDA) using Visualization and SQL
- Perform Interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, KNN, Support Vector Machine and Decision Tree models were used to identify the best estimators

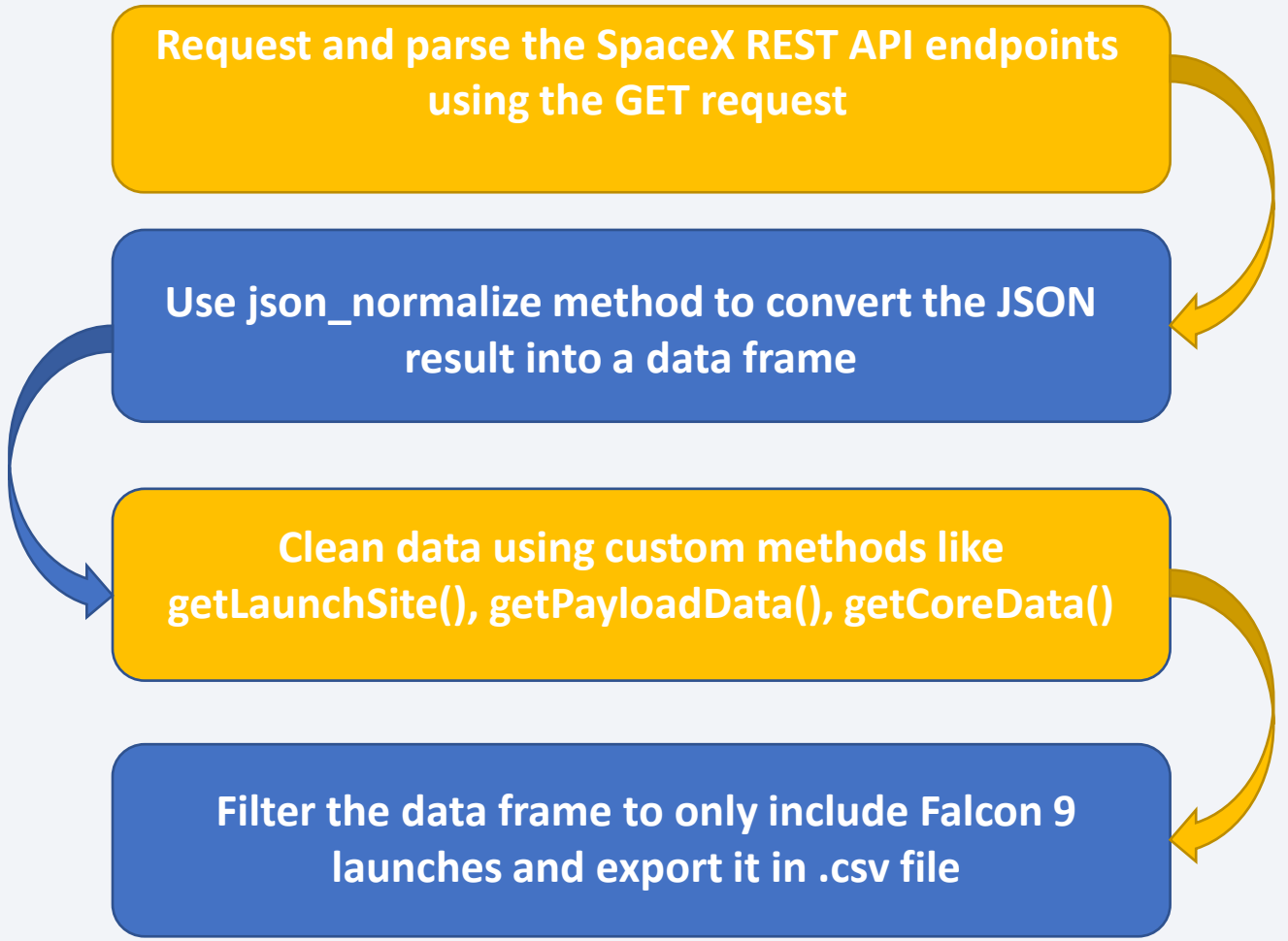
Data Collection

- SpaceX launch data is gathered from SpaceX REST API and its different endpoints.
- This API provides data about launches, the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- This data would be used to predict whether SpaceX will attempt to land a rocket or not.
- Another source of data scraping HTML tables on Wikipedia pages related to SpaceX Falcon 9 using Python BeautifulSoup package.



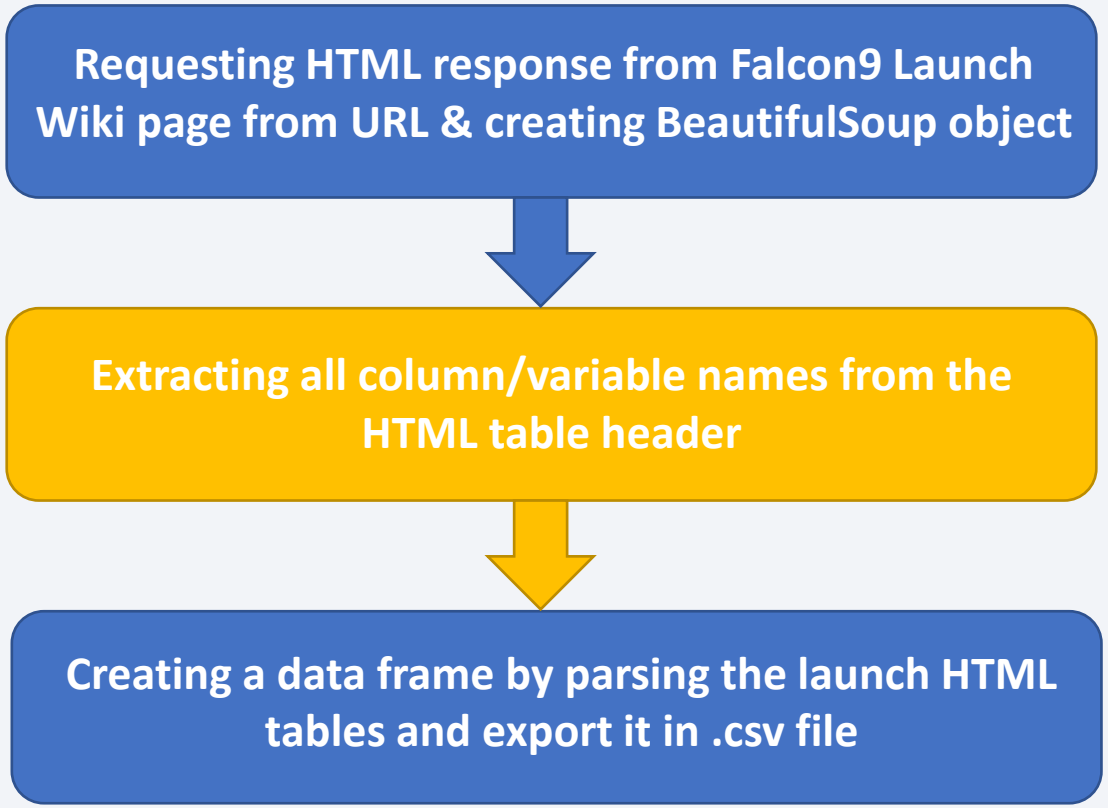
Data Collection – SpaceX API

- SpaceX REST API and its different endpoints are used to gather launch data.
- SpaceX REST API starts with api.spacexdata.com/v4/.
- Flowchart beside shows the steps followed to gather this data.
- GitHub URL: https://github.com/aayushigupta99/SpaceY/blob/main/Data%20Collection%20API.ipynb



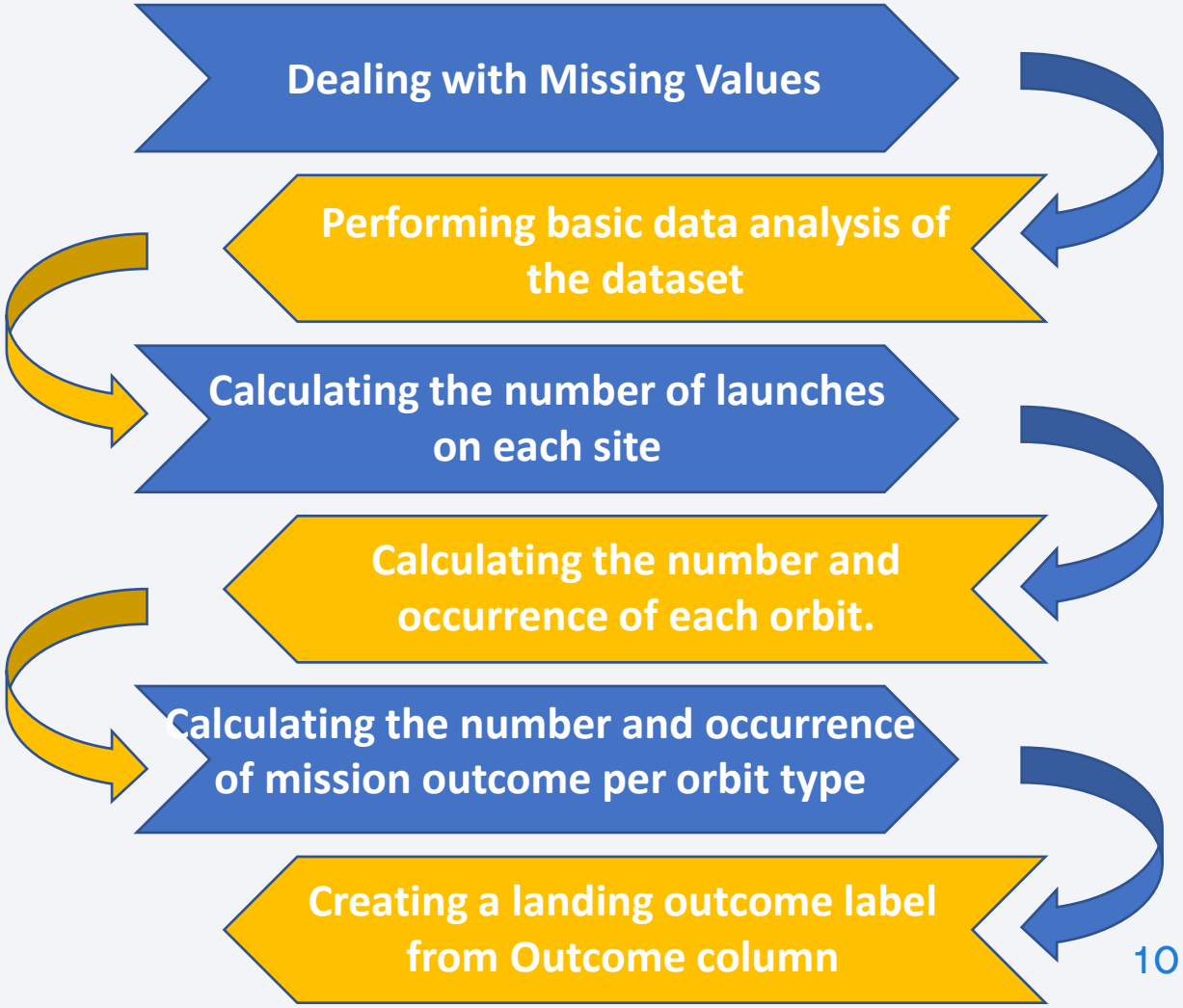
Data Collection - Scraping

- Web Scraping of SpaceX Falcon 9 related Wikipedia pages provides us with launch records.
- This process is carried out using the BeautifulSoup package of python.
- Flowchart beside shows the steps followed to gather this data.
- GitHub URL: https://github.com/aayushigupta99/SpaceY/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- Data wrangling is pre-processing the dataset before performing the analysis.
- As part of data wrangling, we performed the following (also shown in the flowchart):
  - Handle missing values
  - Data formatting
- GitHub URL: https://github.com/aayushigupta99/SpaceY/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

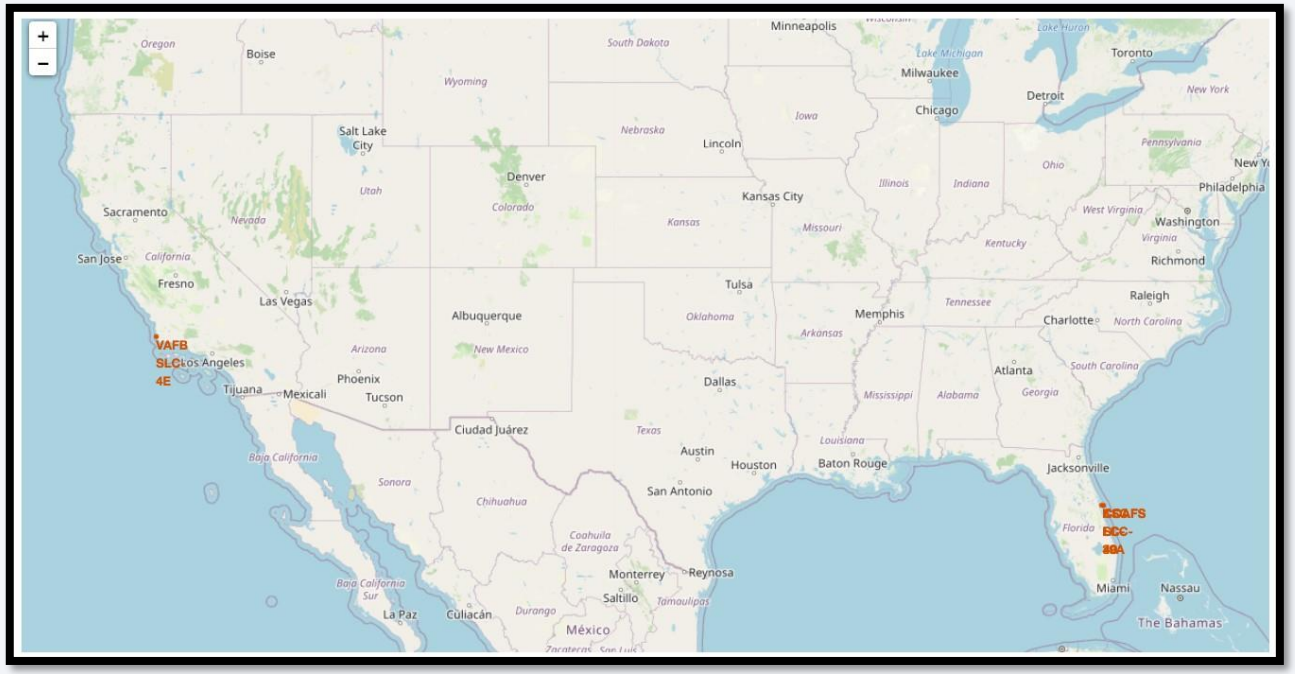
- Following charts were plotted as part of exploratory data analysis using Seaborn and Matplotlib:
  - Scatter plot: Visualizing relationship between Flight Number & Launch Site
  - Scatter plot: Visualizing the relationship between Payload & Launch Site
  - Barplot: Visualizing relationship between success rate of each orbit type
  - Scatter plot: Visualizing relationship between FlightNumber & Orbit type
  - Scatter plot: Visualizing relationship between Payload & Orbit type
  - Lineplot: Visualizing the launch success yearly trend
- GitHub URL: https://github.com/aayushigupta99/SpaceY/blob/main/EDA%20with%20Python%20%26%20Matplotlib.ipynb

EDA with SQL

- SQL queries performed as part of exploratory data analysis of the dataset:
  - 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 dates 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 using subquery
  - Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- GitHub URL: https://github.com/aayushigupta99/SpaceY/blob/main/EDA%20with%20SQL.ipynb

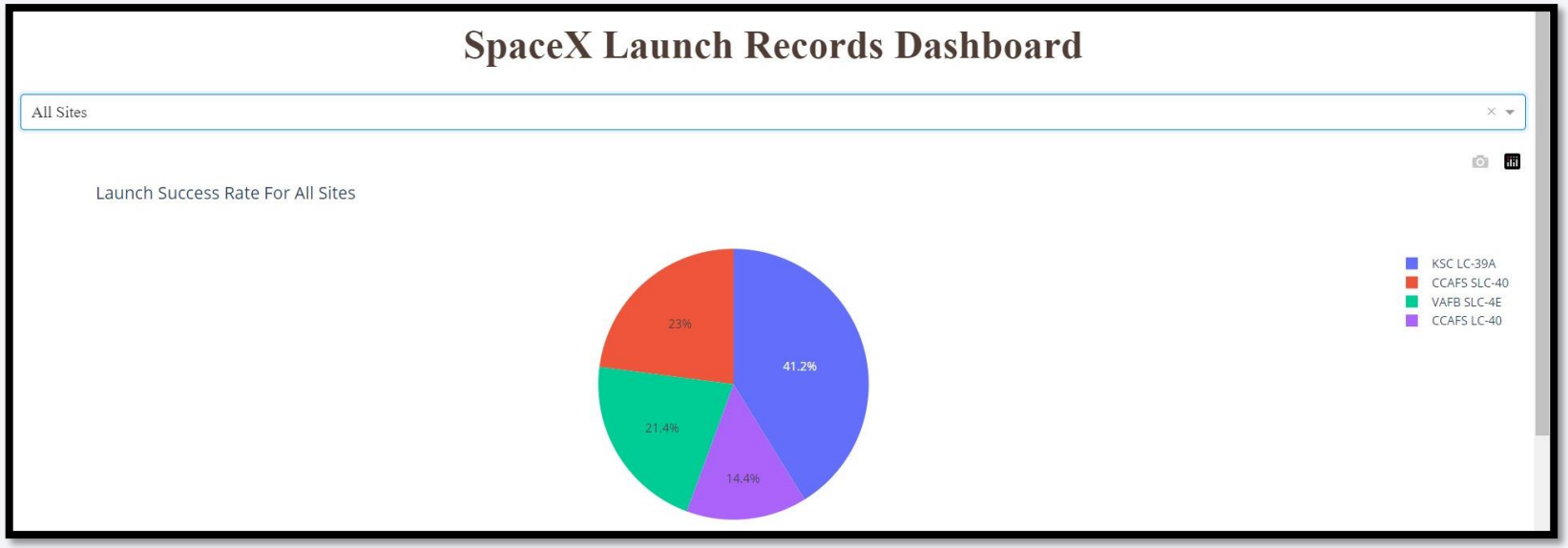
Build an Interactive Map with Folium

- Built an interactive map using Folium to represent launch sites of Falcon9 geographically.
- This map was created using marker, circle, line, markercluster objects to represent failed/successful sites and proximity to nearby areas.
- GitHub URL: https://github.com/aayushigupta99/SpaceY/blob/main/Data%20Visualization%20-%20Interactive%20Map%20Using%20Folium.ipynb



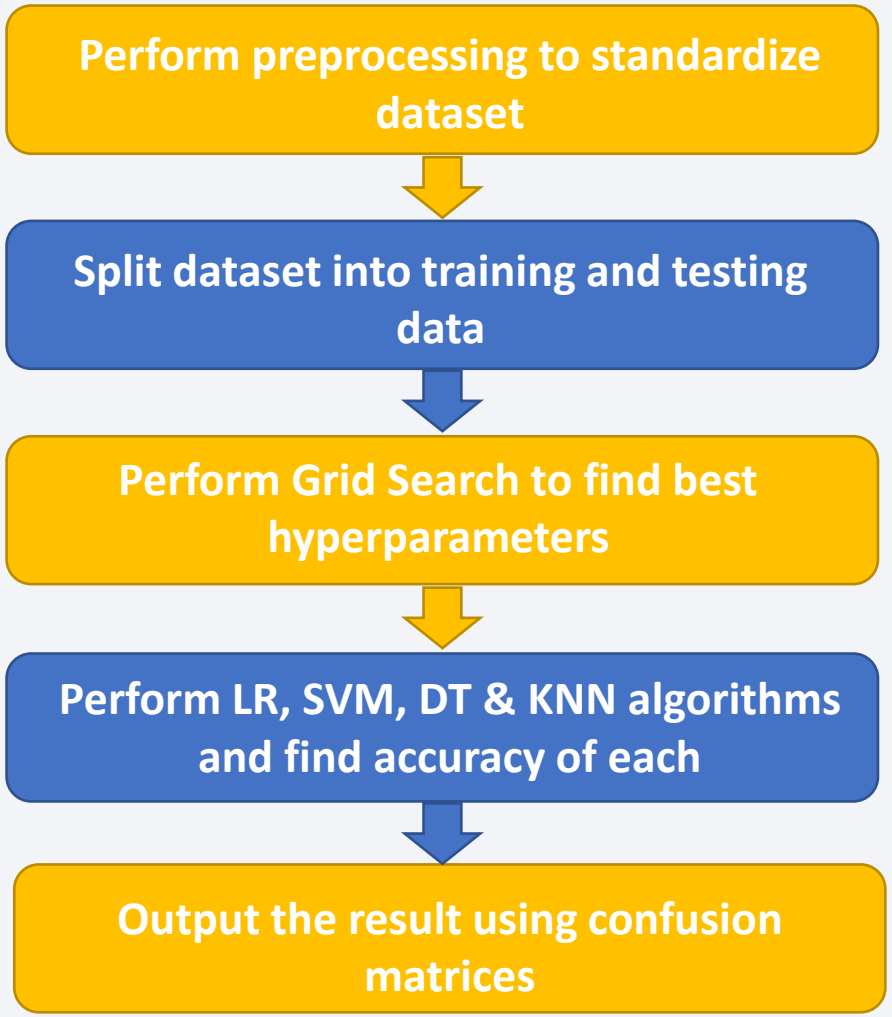
Build a Dashboard with Plotly Dash

- Built an interactive dashboard using Plotly Dash.
- These plots and interactions help to perform visual analytics to better understand data insights.
- GitHub URL: https://github.com/aayushigupta99/SpaceY/blob/main/Data%20Visualization%20-%20Dashboard%20Using%20Plotly%20Dash.py



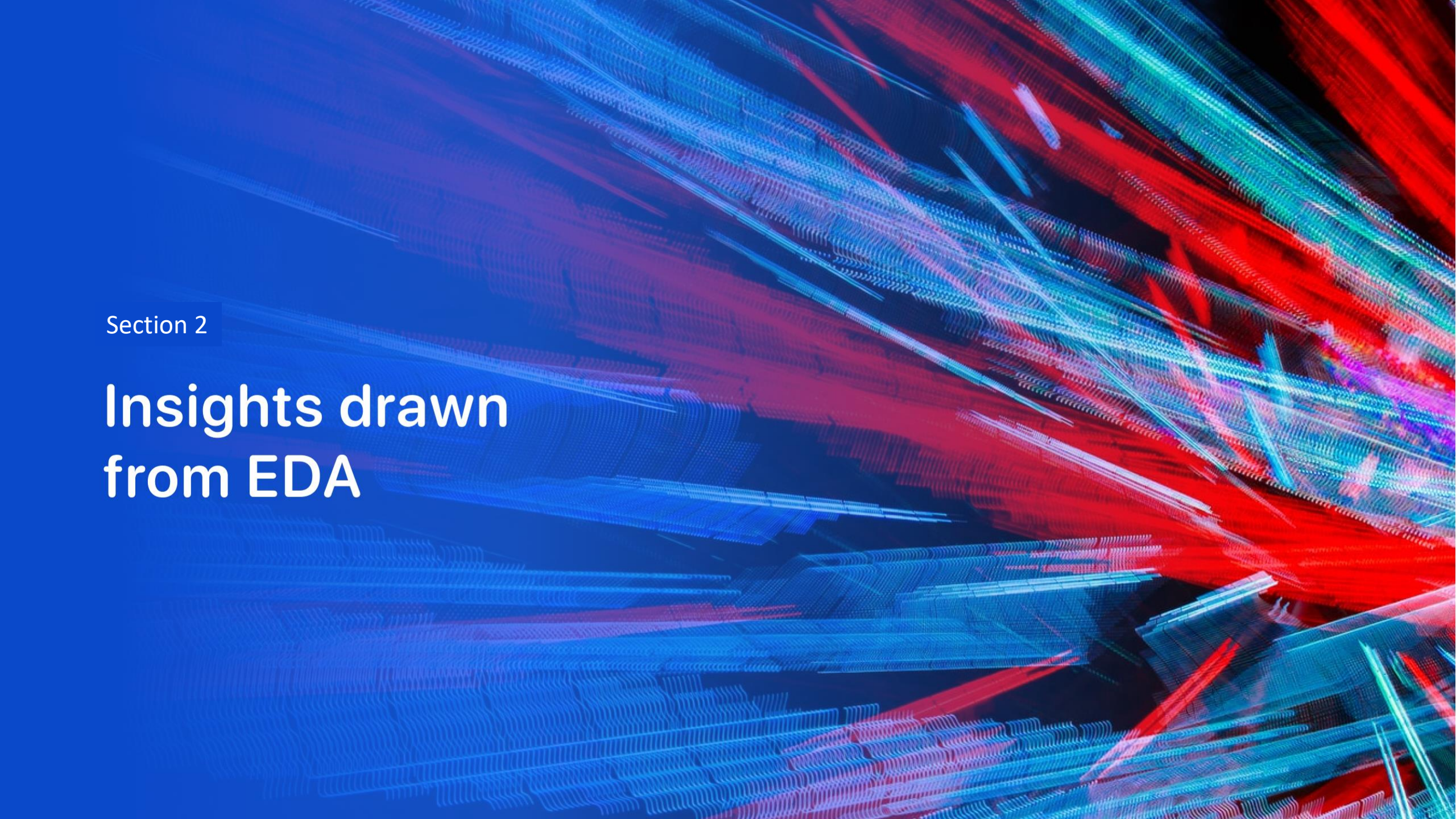
Predictive Analysis (Classification)

- Built a machine learning pipeline to predict if the first stage of the Falcon9 lands successfully.
- It included the steps as depicted in the flowchart.
- Using the best hyperparameter values found, we determined the model with the best accuracy using the training data.
- Logistic Regression, Support Vector machines, Decision Tree Classifier, and KNN algorithms were used to output confusion matrices.
- GitHub URL: https://github.com/aayushigupta99/SpaceY/blob/main/Predictive%20Analysis%20-%20Classification.ipynb



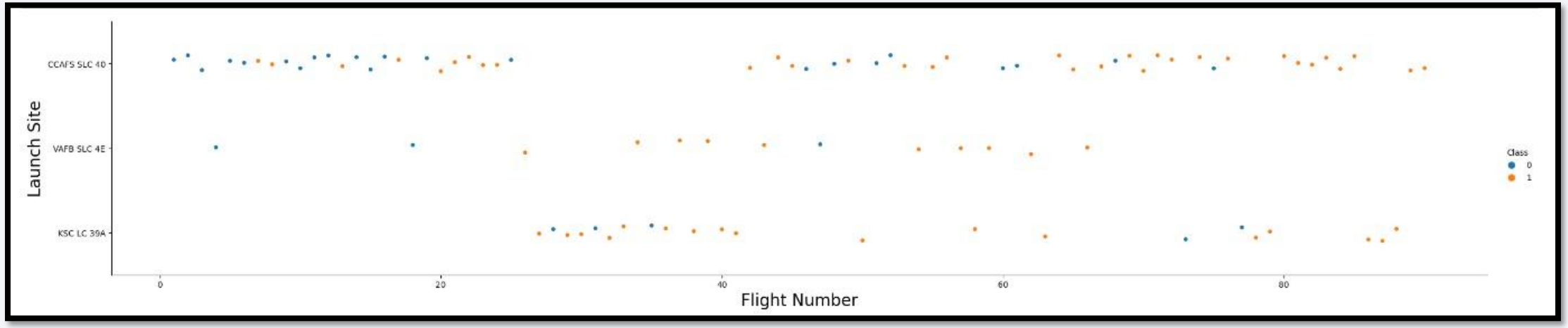
Results

- The SVM, KNN and Logistic Regression models are the best in terms of prediction accuracy nearing 83%.
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time. In years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches among all sites.
- Orbit GEO, HEO, ES L1 have the best success rate.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

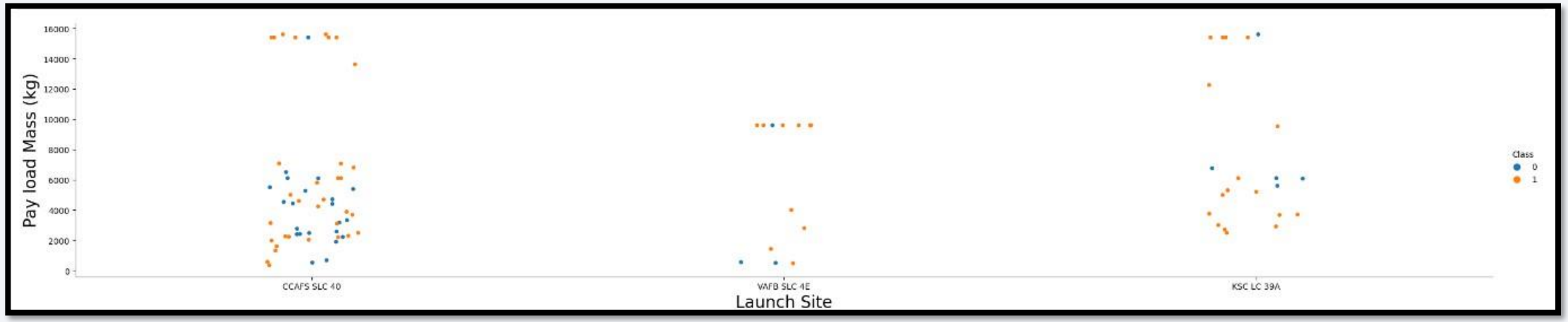
Scatter plot of Flight Number vs. Launch Site:



- Launch site CCAFS SLC 40 had the greatest number of launches among the three.
- Launch site VAFB SLC 4E has the best success rate.

Payload vs. Launch Site

Scatter plot of Payload vs. Launch Site:

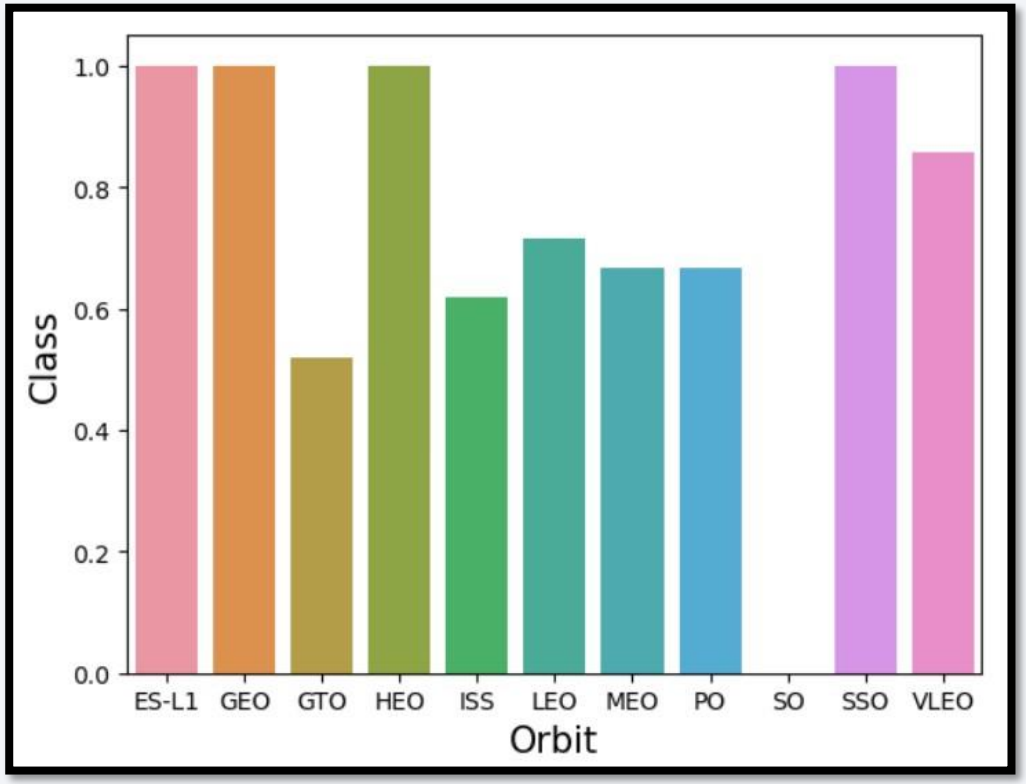


- Launch site CCAFS SLC 40 had more successful launches with high payload.
- Launch site VAFB SLC 4E did not have any launches with payload > 10,000 kgs.

Success Rate vs. Orbit Type

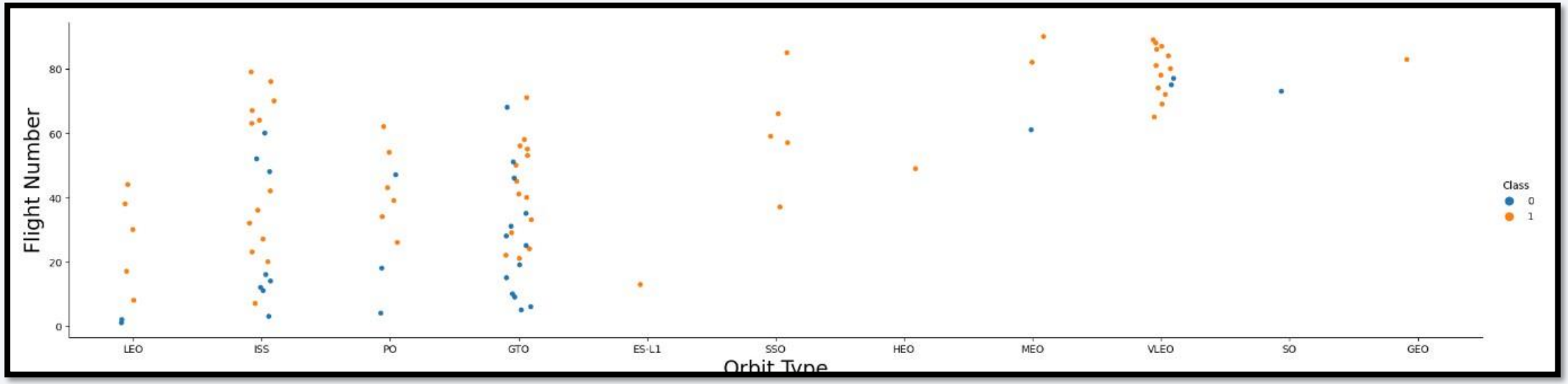
- Bar plot of Orbit Type vs. Success Rate:

- Orbits ES-L1, GEO, HEO and SSO have absolute success rate.
- Orbit SSO has not had any successful launches.



Flight Number vs. Orbit Type

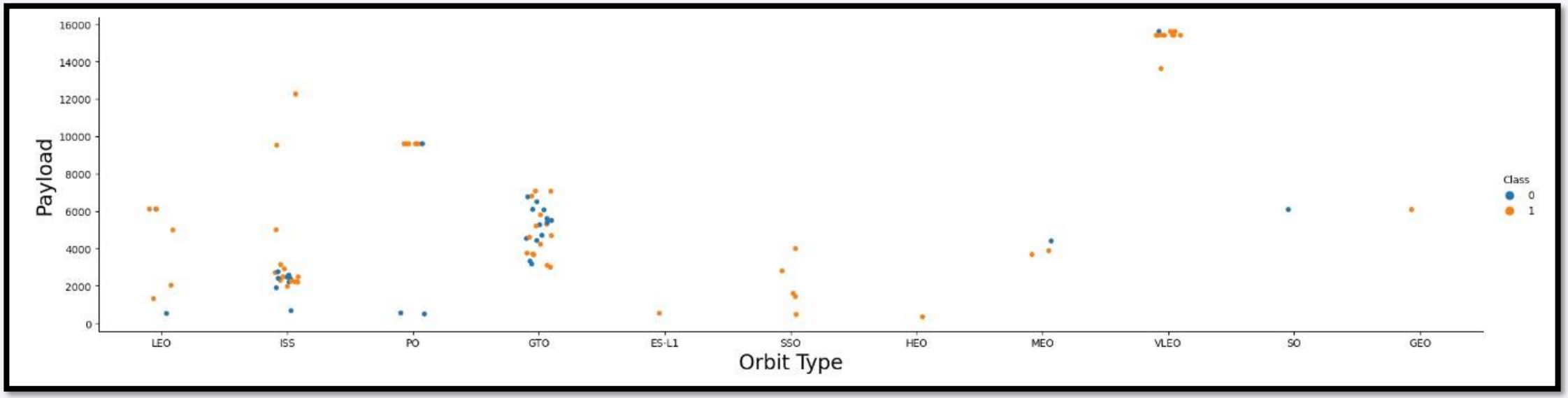
Scatter plot of Flight Number vs. Orbit Type:



- In the LEO orbit, success appears to be related to the number of flights.
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

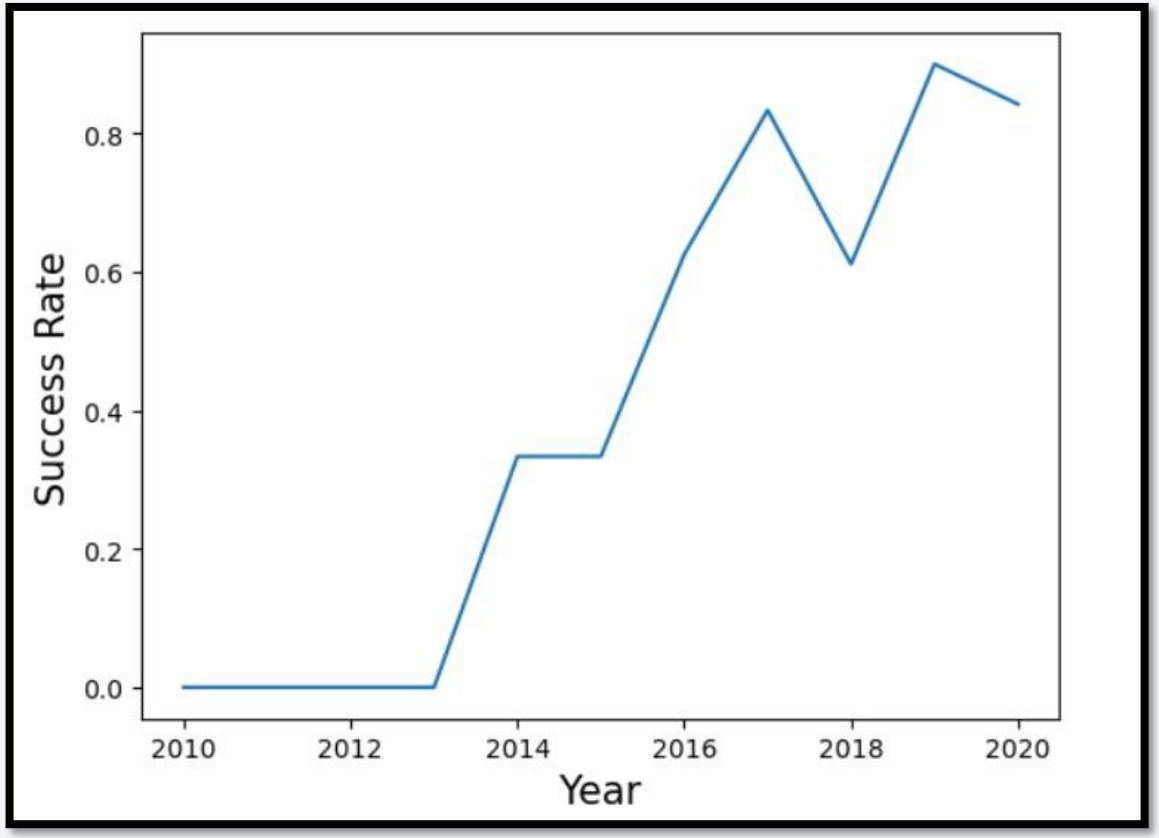
Scatter plot of Orbit Type vs. Payload:



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO, it is hard to distinguish the same as both positive landing rate and negative landing (unsuccessful mission) are present.

Launch Success Yearly Trend

- Line chart of yearly average success rate:
- We can clearly infer that the Success Rate has increased continuously after 2013.
- This implies that, in time, SpaceX will perfect its launches.



All Launch Site Names

- Query to find the names of the unique launch sites:
- There are 4 distinct launch sites.

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

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX_DATA
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

- Query to find 5 records where launch sites begin with `CCA`

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

```
%sql SELECT * FROM SPACEX_DATA WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/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

- Query to calculate the total payload carried by boosters from NASA
- The total payload that the boosters carried from NASA was 45596 kgs.

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX_DATA WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- Query to calculate the average payload mass carried by booster version F9 v1.1
- The average payload that booster version F9 v1.1 carried was 2928 kgs.

Task 4

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX_DATA WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
2928

First Successful Ground Landing Date

- Query to find the dates of the first successful landing outcome on ground pad
- The first successful landing on ground pad happened on 22nd December 2015.

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

Hint:Use min function

```
%sql SELECT MIN(DATE) FROM SPACEX_DATA WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- There were 4 booster versions which successfully landed on drone ship carrying a payload between 4000 kgs and 6000 kgs.

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 SPACEX_DATA WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/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

- Query to calculate the total number of successful and failure mission outcomes
- Total number of successful and failed missions is 71.

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(*) FROM SPACEX_DATA WHERE LANDING__OUTCOME LIKE 'Success%' OR LANDING__OUTCOME LIKE 'Failure%';
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
71

Boosters Carried Maximum Payload

- Query to list the names of the booster which have carried the maximum payload mass

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

```
%sql SELECT BOOSTER_VERSION FROM SPACEX_DATA WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX_DATA);
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/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

- Query to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- There were 2 failure landings for drone ships in the year 2015.

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

```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX_DATA WHERE LANDING__OUTCOME LIKE 'Failure (drone ship)' AND DATE LIKE '2015%'
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Query to 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

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 OUTCOME_COUNT FROM SPACEX_DATA GROUP BY LANDING__OUTCOME ORDER BY OUTCOME_COUNT DESC;
```

```
 * ibm_db_sa://stq76324:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

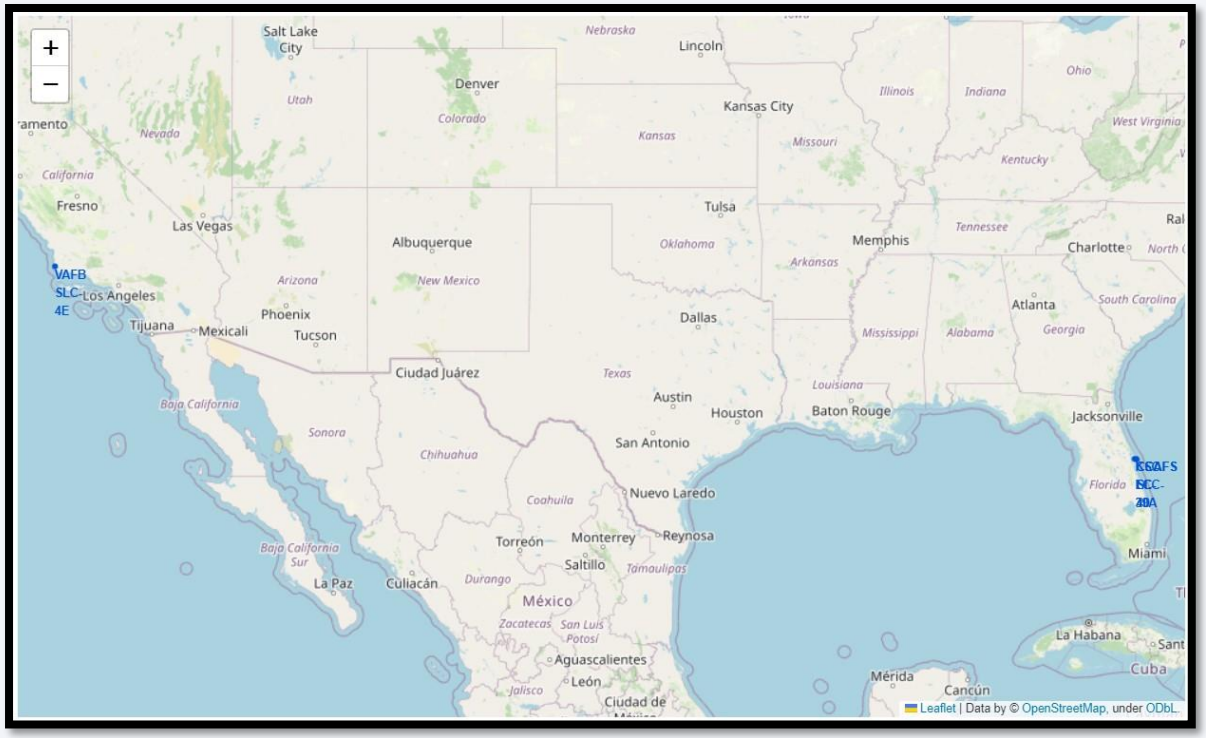
landing__outcome	outcome_count
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

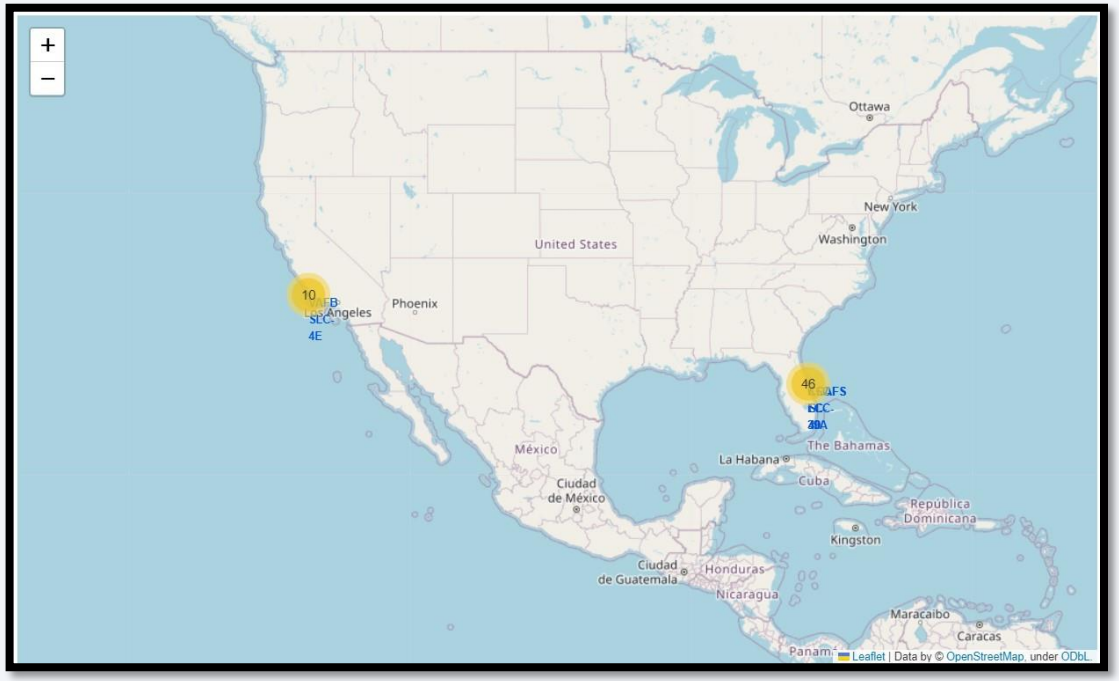
All Launch Sites On Map

- All launch sites are near the equator and the coast.
- This makes sense as it takes less fuel to get into space from the equator due to the physics of Earth's rotation.
- The launch sites' close proximity to the coast is also logical for safety reasons.



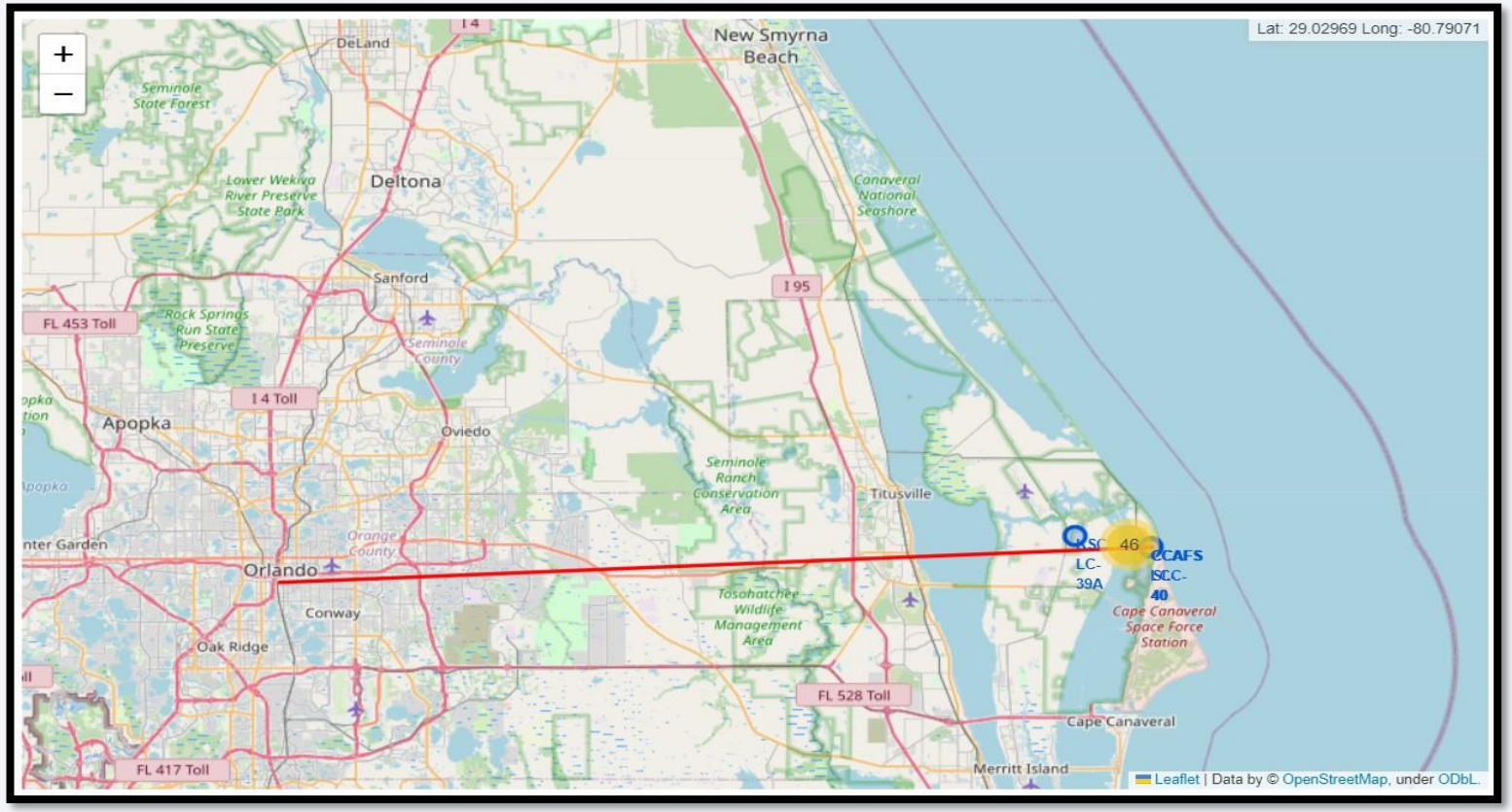
Successful/Failed Launches for Each Site

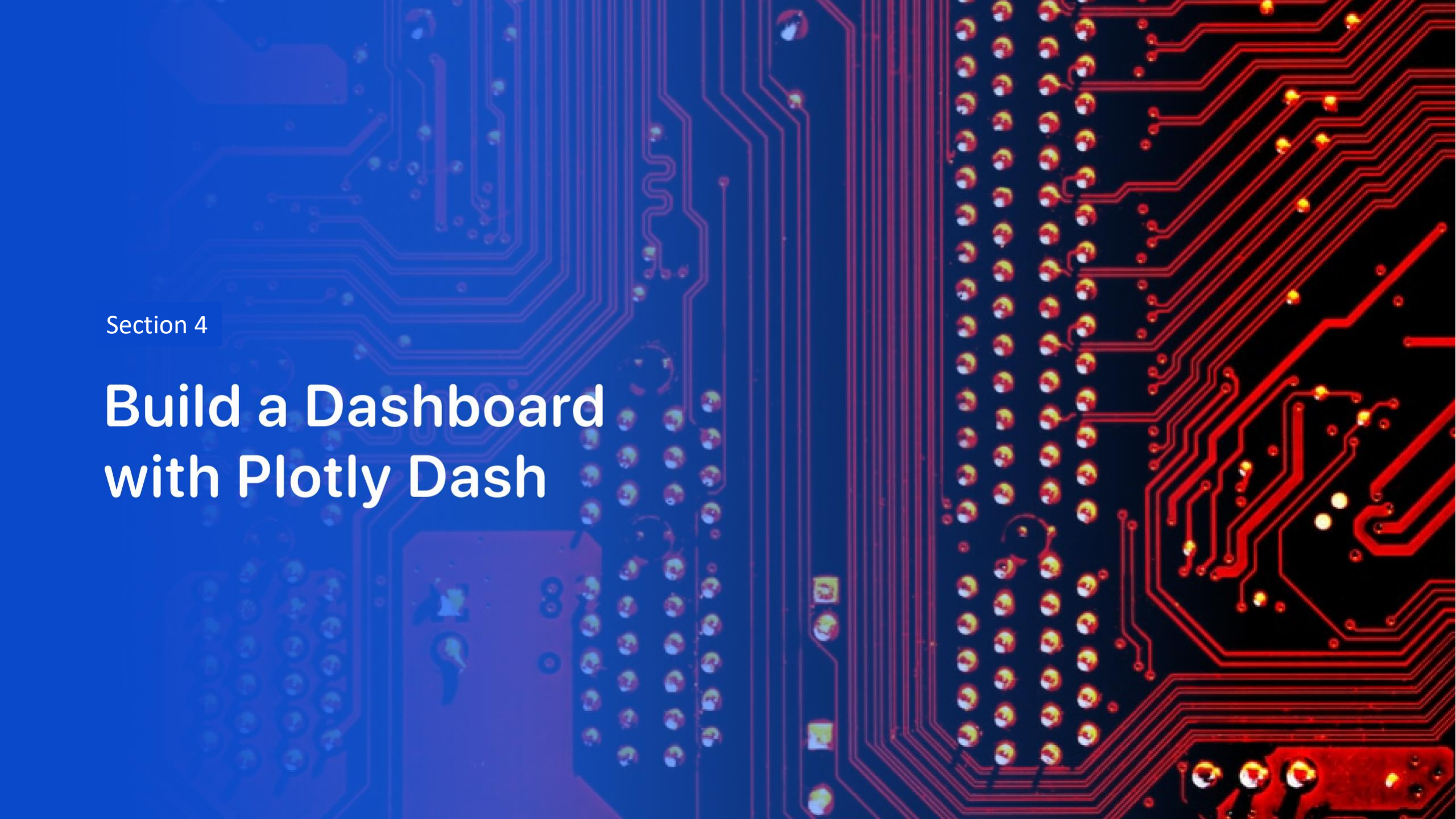
- KSC LC 39A has the best success rate for launches.
- CCAFS SLC-40 has the lowest success rate for launches.



Distances Between Launch Sites to its Proximities

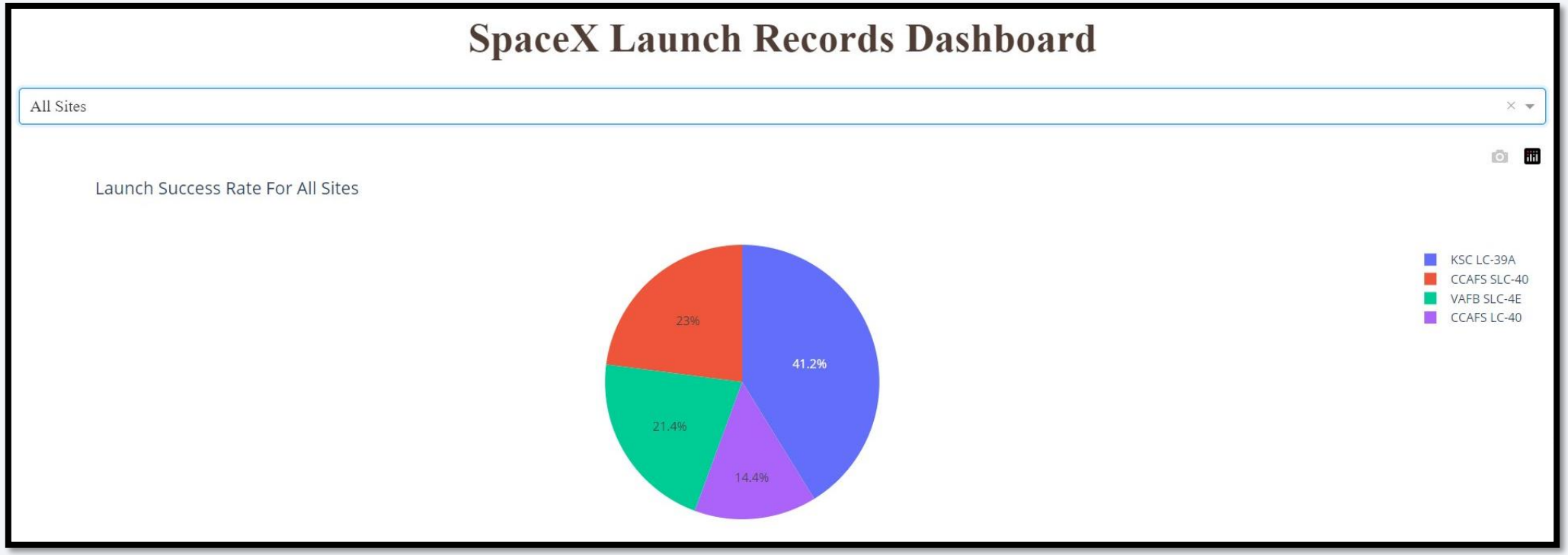
- To get the exact coordinates over any points of interests, such as railway, in the map, a `MousePosition` object is added to the map.
- Polyline object is used to plot a line from Orlando city to CCAFS SLC-40 launch site.




Section 4
Build a Dashboard with Plotly Dash

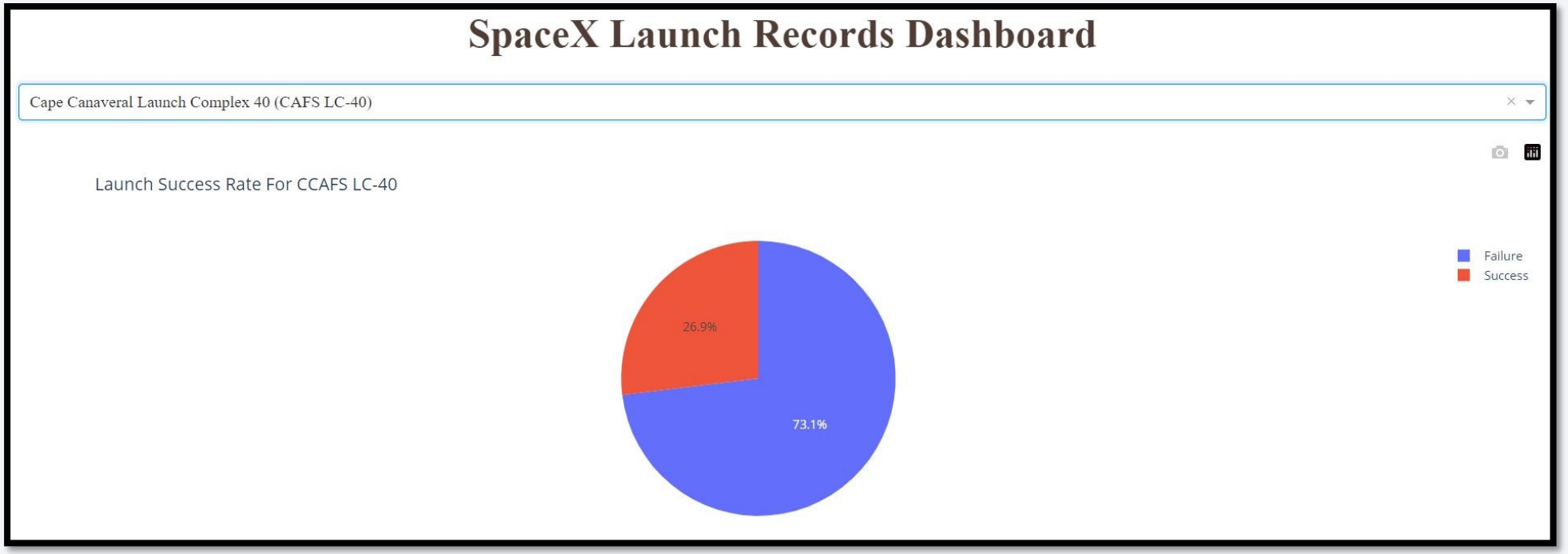
Launch Success Count for All Sites

- Pie-Chart showing launch success count for all launch sites.
- KSC LC-39A has a dominating share of success rate.
- The pie-chart confirms our previously drawn insight from scatter plots.



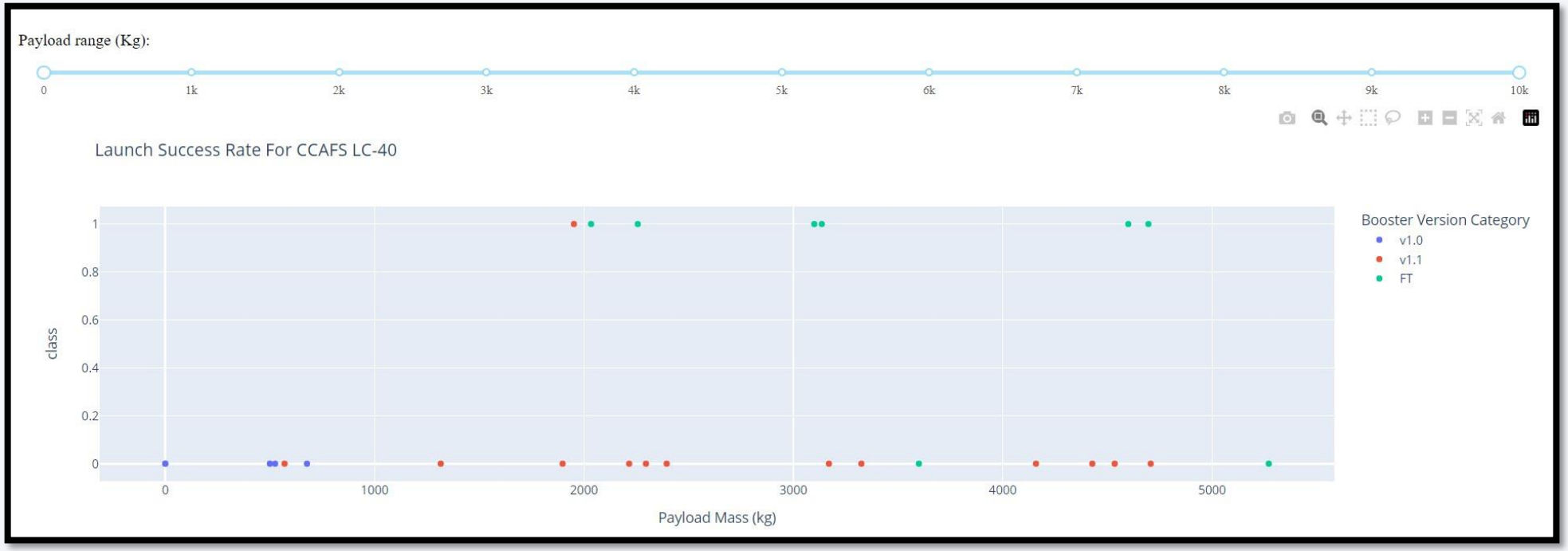
Launch Success Ratio for CCAFS LC-40

- Pie-chart showing launch success ratio for CCAFS LC-40 launch site.
- As depicted in the screenshot below, we have a dropdown with values of each launch site.
- Blue depicts the failure ratio and Red the success ratio



Payload vs. Launch Outcome Scatter Plot for All Sites

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider.
- As visible from the screenshot below, the FT booster version has the highest success rate at high payload values for CCAFS LC-40 launch site.

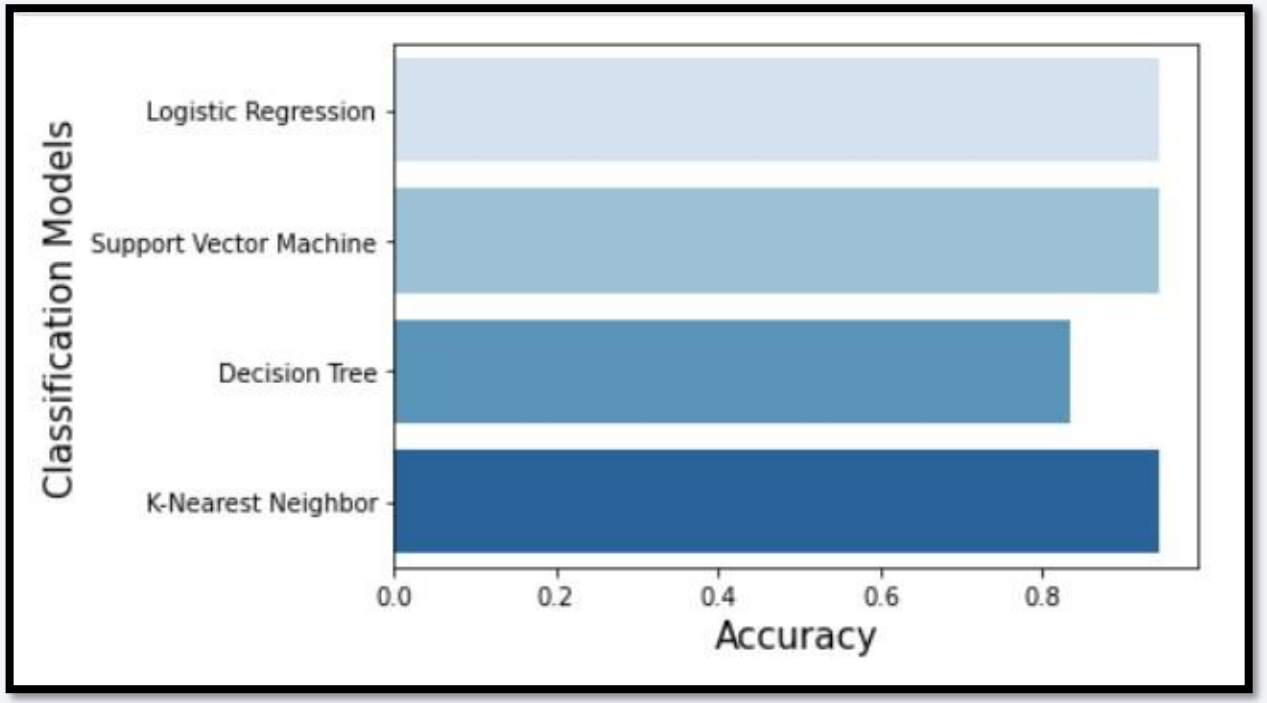


Section 5

Predictive Analysis (Classification)

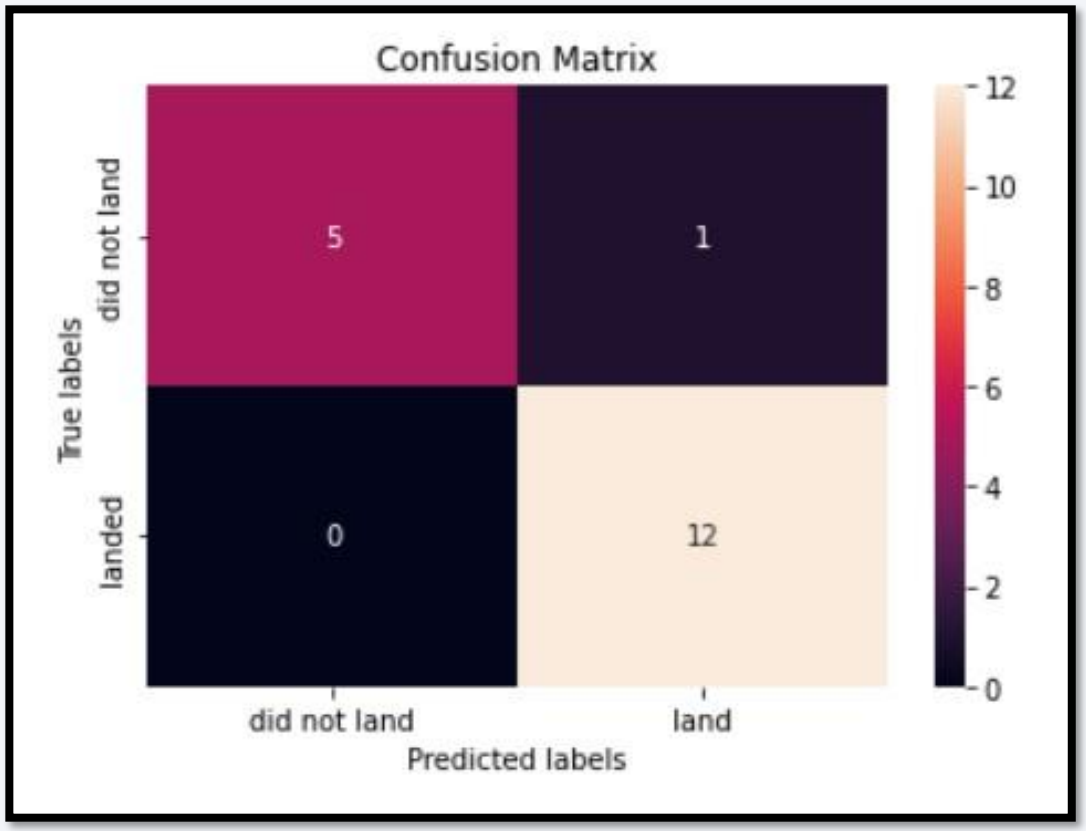
Classification Accuracy

- Visualization of the built model accuracy for all built classification models, in a bar chart.
- Logistic Regression, Support Vector Machine and K-Nearest Neighbor models have the same and highest accuracy of all the models built.
- The accuracy is close to 94.45%.



Confusion Matrix

- Confusion matrix shown is for LR, SVM and KNN models.
- Following are the metrics depicted in the confusion matrix:
  - True Positives: 12
  - False Positives: 1
  - True Negatives: 5
  - False Negatives: 0



Conclusions

- Logistic Regression, Support Vector Machine and K-Nearest Neighbor models have the highest accuracy of all the models built.
- The prediction accuracy is close to 94.45%.
- Launch site KSC LC-39A has the highest success rate of all the sites.
- Orbits ES-L1, GEO, HEO and SSO have absolute success rate in terms on launches.
- The success rates of SpaceX launches is directly proportional to time. Hence, we can say, in time, they will perfect their launches with an absolute success ratio.

Appendix

- Data set created during this project: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv

Thank you!