Winning Cost Effective Race to the Space

Muhammad Ahsan Zaheer

11/05/2022

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

EXECUTIVE SUMMARY

- Methodologies used to analyze data:
  - Data Collection through web scraping and SpaceX API.
  - Exploratory Data Analysis (EDA) [Data Wrangling, Data Visualization, Interactive Visual Analysis].
  - Predictive Analysis with the help of Machine Learning (ML).
- Summary:
  - Valuable data was collected from different sources which helped in our findings.
  - EDA helped identify the best features to predict the success of the landings.
  - ML Prediction gave the best model to predict the success of the landings, showed which characteristics are important when prediction which model is the best.

INTRODUCTION

- Our goal is to evaluate how successfully can the new company Space Y compete with the industry giant SpaceX.
- A few pointers to keep in mind:
  - SpaceX has leading edge due to the lowest cost of landings.
  - How can Space Y make sure their costs are competitive to that of SpaceX's?
  - What is the best way to estimate the total cost of landings?
  - It is by predicting successful landings of the first stage of rockets.
  - The best location to make the launches from.
  - Landing in water or landing on earth.

METHODOLOGY

- Data Collection
  - SpaceX related data was obtained from two sources:
    1) SpaceX API: https://api.spacexdata.com/v4/rockets/
    2) Web Scraping:
    https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Data Wrangling
  Data was made more meaningful by introducing a landing_outcome column based on other outcome data.
- EDA using SQL and Data Visualization

METHODOLOGY

- Interactive and Visual Analysis with the help of Folium and Plotly Dash
- Predictive Analysis using Machine Learning and classification models.
- For training the models accurately, the data was normalized and split into testing and training data sets.
- Data was then evaluated by 4 different classification models using different parameters.

DATA COLLECTION

- Data was collected from the SpaceX API and Wikipedia with the help of web scraping.
- SpaceX has a public API from where data can be obtained and used to derive meaningful insights.
- SpaceX data is also available on Wikipedia which can also be obtained through web scraping.

Notebook URL for Data Collection: https://github.com/ahsanz95/Applied-Data-Science-Capstone/blob/main/Data%20Collection.ipynb

Notebook URL for Data Collection with Web Scraping: https://github.com/ahsanz95/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

DATA WRANGLING

- Initially performed some Exploratory Data Analysis (EDA) on the dataset.
- Calculated number of launches on each site, the number and occurrence of each orbit, and number and occurrence of mission outcome per orbit type.
- The landing outcome label was also created from the outcome column.

Notebook URL for Data Wrangling: https://github.com/ahsanz95/Applied-Data-Science-Capstone/blob/main/Data%20wrangling.ipynb

EDA With Data Visualization

- To explore data in more detail, scatterplots and bar plots were used to visualize the relationships between different features:

  - Flight Number and Launch Site
  - Payload and Launch Site
  - Flight Number and Orbit Type
  - Payload and Orbit Type

- Created dummy variables for the categorical columns.

Notebook URL for EDA with Data Visualization: https://github.com/ahsanz95/Applied-Data-Science-Capstone/blob/main/EDA%20With%20Data%20Visualization.ipynb

EDA With SQL

- There were some SQL queries performed to understand the data better:

  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begins with 'CCA'
  - Total Payload mass carried by boosters launched by NASA (CRS)
  - Average Payload mass carried by booster version F9v1.1
  - Date of the first successful landing outcome in ground pad
  - Total Number of successful and failure mission outcomes
  - Names of successful boosters in drone ship and payload mass between 4000 – 6000
  - Names of booster versions that have carried the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions and launch site names for year 2015
  - Rank the count of landing outcomes such as failure or success between 2010-06-04 and 2017-03-20.

Notebook URL for EDA with SQL: https://github.com/ahsanz95/Applied-Data-Science-Capstone/blob/main/EDA%20With%20SQL.ipynb

Interactive Visual Analysis with Folium

- Markers, lines, circles, marker clusters were all used with Folium Maps.
- All launch Sites were marked on the map with Markers.
- Lines were used to indicate the distance between the said coordinates. For example the distance between the launch site and its proximities.
- Circles were used to indicate highlighted areas around specific coordinates like the NASA Johnson Space Center.
- Marker Clusters help indicating different events in each coordinate, like the launches in a launch site.

Notebook URL for Interactive Analytics with Folium: https://github.com/ahsanz95/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

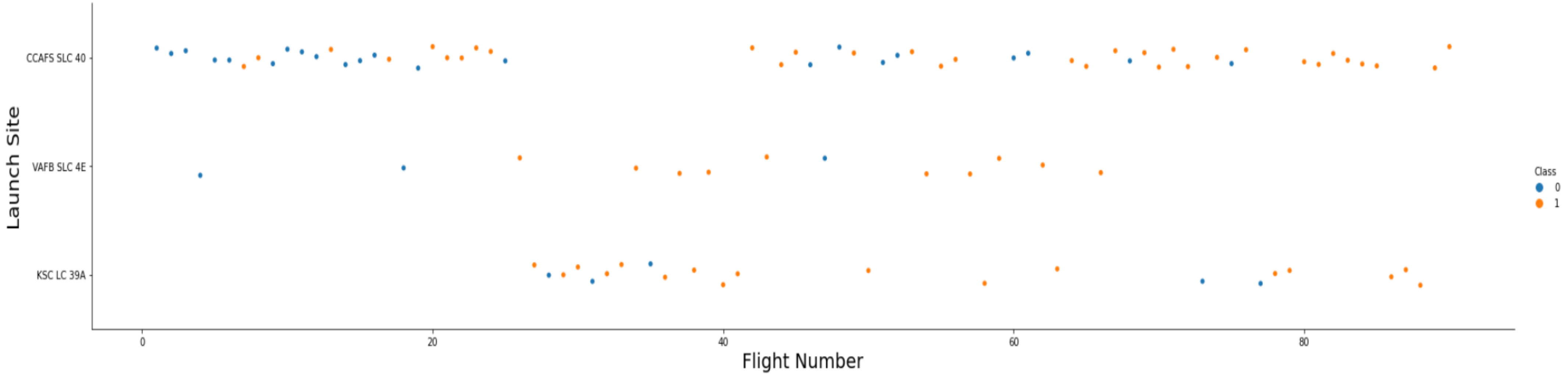
PREDICTIVE ANALYSIS

- Data was split into training and testing data.
- Four Classification models were built:
- Logistic Regression, K-Nearest Neighbor, Support Vector Machine (SVM) and Decision Tree.
- All the models were then compared to each other and the best one was picked.

Notebook URL for Predictive Analysis: https://github.com/ahsanz95/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

EDA with Data Visualization Results

Flight Number and Launch Site

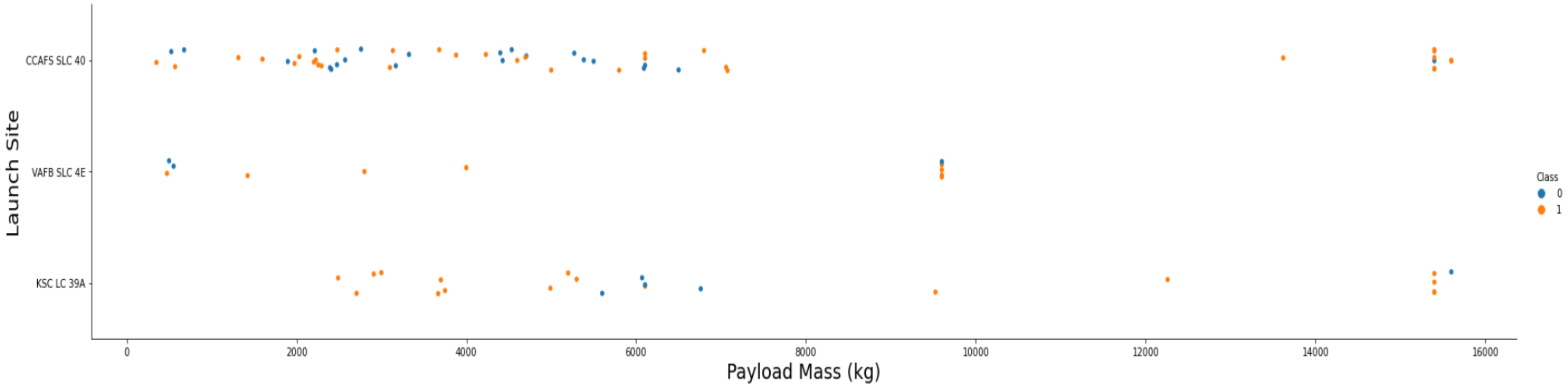


This graph shows the most successful launch site recently is the CCAFS SLC 40, and hence the best.

Followed by VAFB SLC 4E in second and KSC LC 39A in third.

The general success rate improved over time.

PAYLOAD and LAUNCH SITE



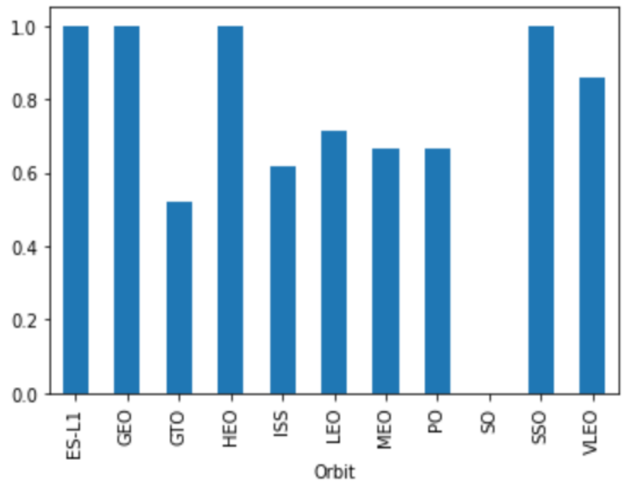
We can see that payloads over 9000(kg) have a great success rate. Almost all of them are successful launches.

Payloads that are over 12000(kg) seem to be only possible on CCAFS SLC 40 and KSC LC 39A launch sites.

The least used launch site seems to be VAFB SLC 4E

SUCCESS RATE of each ORBIT TYPE

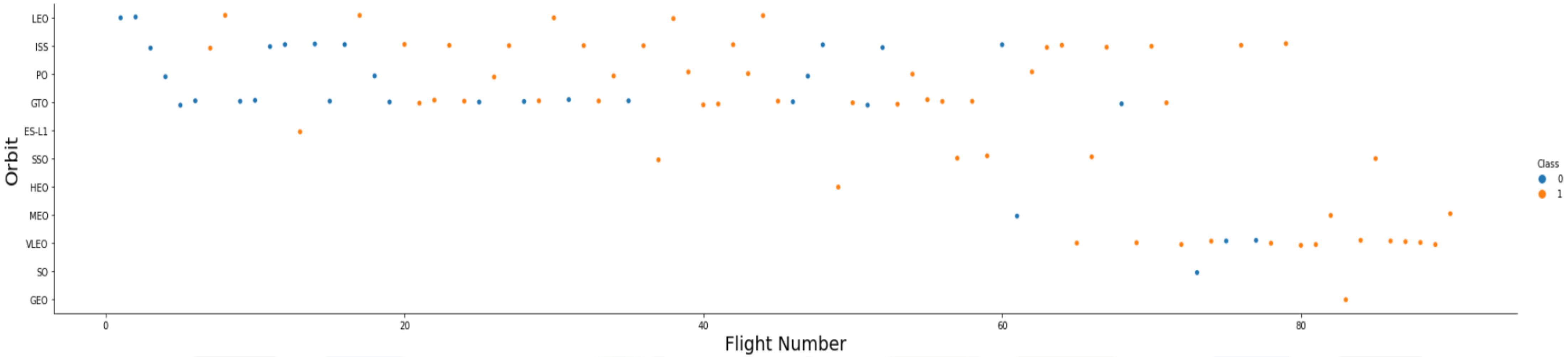
The orbits with the highest success rates are:
- ES-L1
- GEO
- HEO
- SSO



With VLEO close in second,

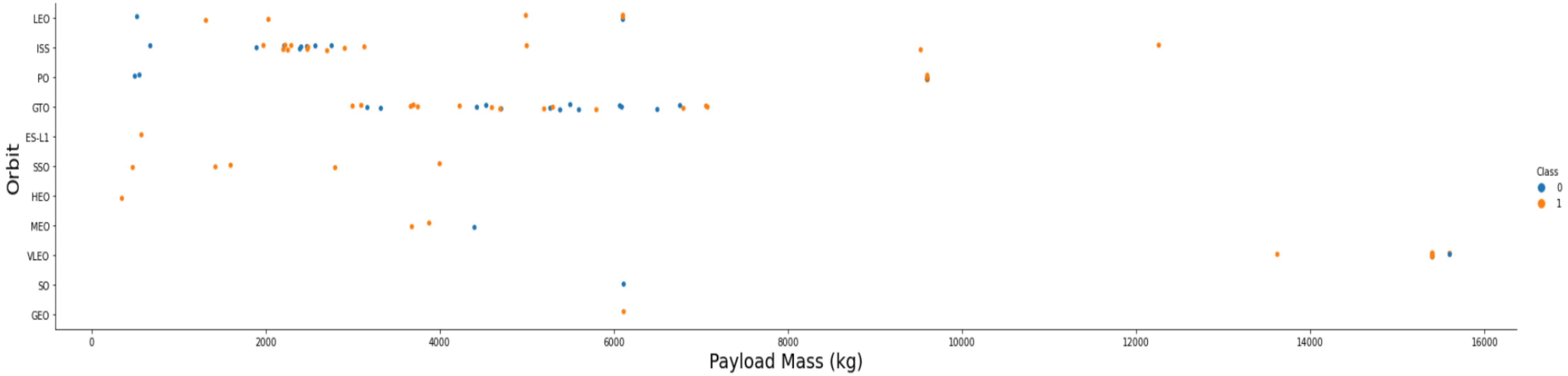
Followed by LEO in third.

FLIGHT NUMBER and ORBIT TYPE



- LEO Orbit is proving to be successful after a few failures in the beginning.
- The success rate to all the orbits have improved with time.
- The use of LEO, PO and GTO orbits have decreased with time.
- And the use of the VLEO orbit has increased with time. This can be helpful information while deciding with orbits to go with.

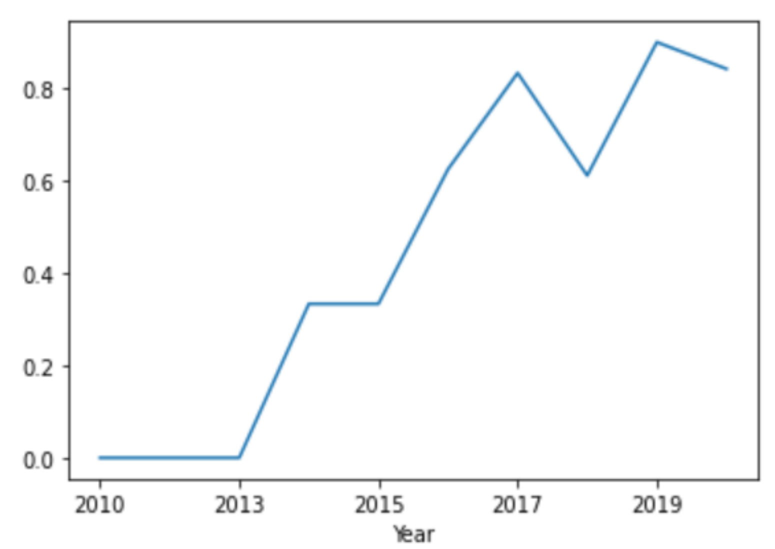
PAYLOAD and ORBIT TYPE



- No relation could be found between payload and success rate to the orbit GTO.
- There are very few launches to the orbits SO, HEO and GEO
- The orbit ISS has a very wide variety of payload and a great success rate.
- SSO orbit also has a good success rate but is not used for higher payloads.

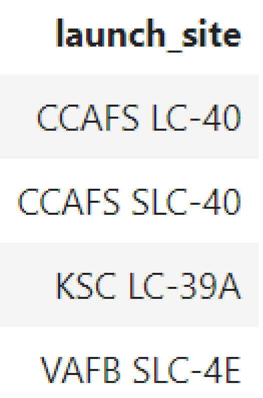
LAUNCH SUCCESS YEARLY TREND

- We can see the success rate kept increasing from 2013 all the way till 2020.
- The first few years seem to be just experiments and adjustments to newest technologies and innovations.
- The last few years were a few hiccups for the success rate.
- The latest launch success still shows a staggering 80%



EDA with SQL RESULTS

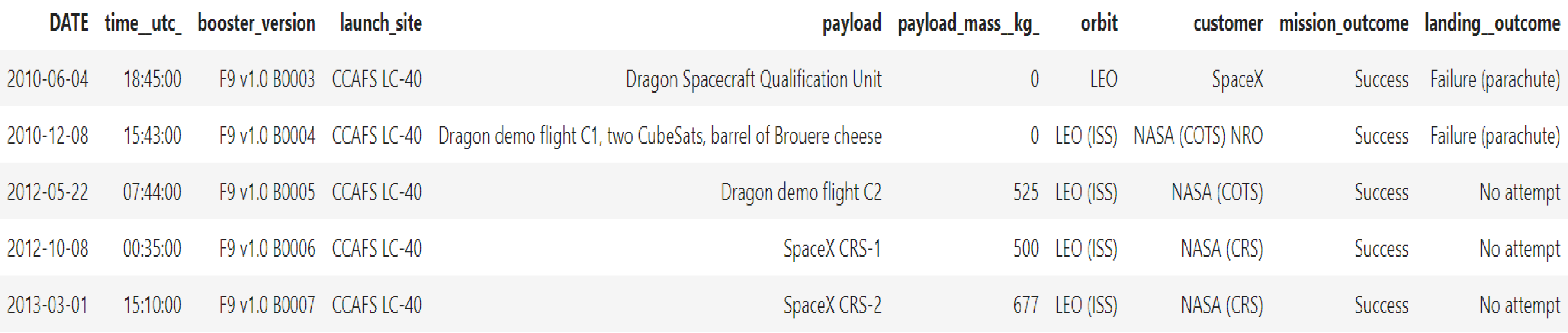
ALL LAUNCH SITE NAMES:



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

- Obtained by using the unique clause in the SQL query.
- There are only 4 unique launch sites:
- CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E.

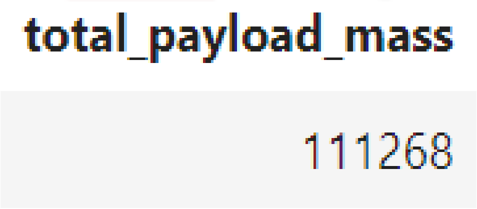
LAUNCH SITE NAMES BEGINNING WITH 'CCA'



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

- Displayed above are 5 records of launch_sites beginning with 'CCA'
- We selected other information as well in the query to retrieve more information about each launch/event.
-Rightly shows the CCAFS LC-40 launch site name in the records above.

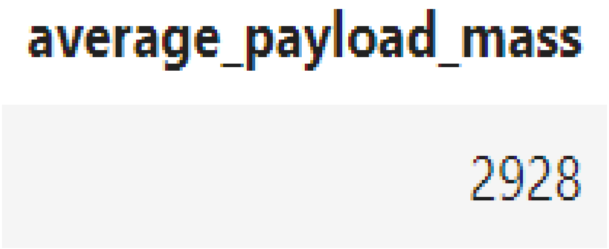
TOTAL PAYLOAD MASS



total_payload_mass
111268

- The total payload mass carried by boosters from NASA is 111268 (kg).
- Used the sum aggregate clause as total_payload_mass.
- Summed all payloads with 'CRS' which corresponds to NASA.

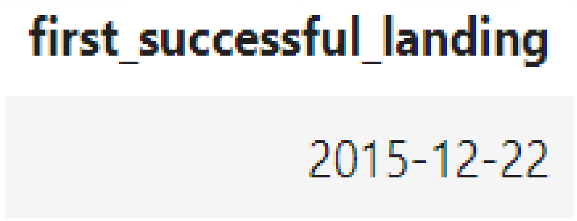
AVERAGE PAYLOAD MASS



average_payload_mass
2928

- The average payload mass carried by booster F9 v1.1 is 2928 (kg).
- Used the average aggregate clause as average_payload_mass.

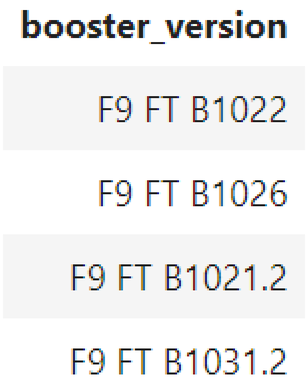
FIRST SUCCESSFUL LANDING DATE



first_successful_landing
2015-12-22

- The first successful landing outcome on ground pad was on 22nd December 2015.
- Filtered data by successful landing outcomes on ground pad and then found the minimum date value in order to get the first successful landing date.

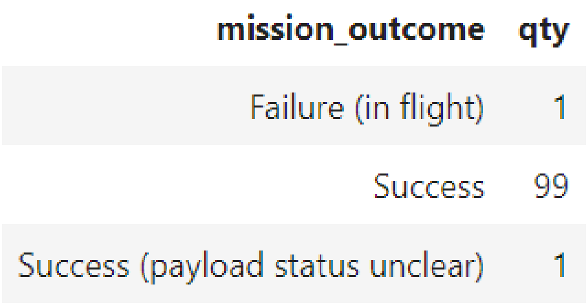
SUCCESSFUL DRONE SHIP LANDING BETWEEN 4000 AND 6000 PAYLOAD



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

- Query shows boosters that had a successful landing on drone ship and had payload between 4000 and 6000 (kg).
- Distinct clause was used to avoid any repetitive values and we got 4 boosters: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2.

TOTAL SUCCESSFUL AND FAILURE MISSION OUTCOMES

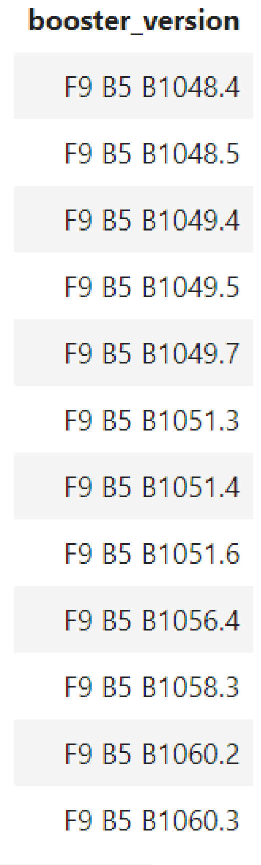


mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- There were 99 clear successful mission outcomes, and 1 where the payload status showed 'unclear'.
- There only 1 failure (in flight).
- Grouped the values for the mission_outcomes and then did a count on all the records in each grouping to achieve these values.

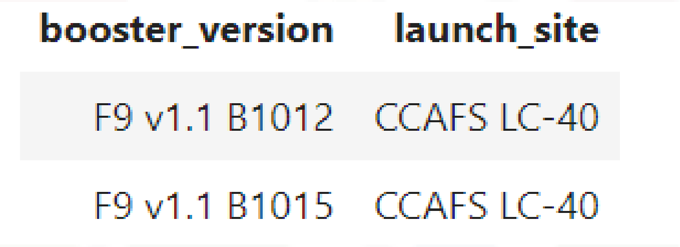
BOOSTERS CARRYING MAXIMUM PAYLOAD

- These are all the boosters that have carried the maximum payload mass.



booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

LAUNCH RECORDS FOR YEAR 2015

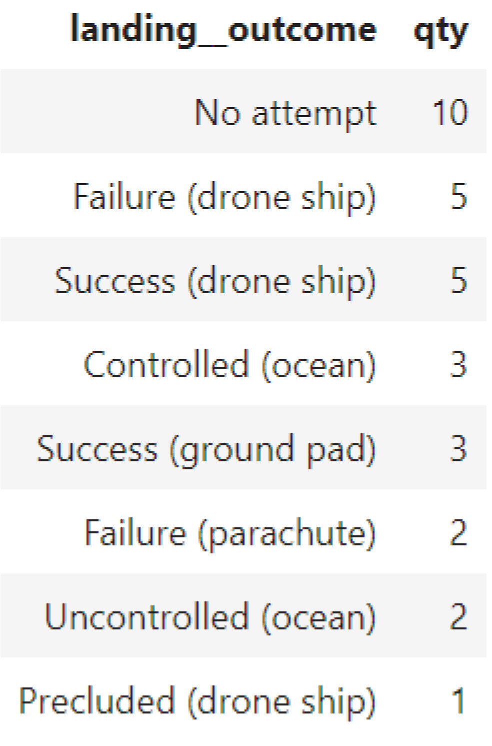


booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- These are the failed landing outcomes in drone ship, their booster versions and launch site names for year 2015.
- The query retrieved only 2 records.

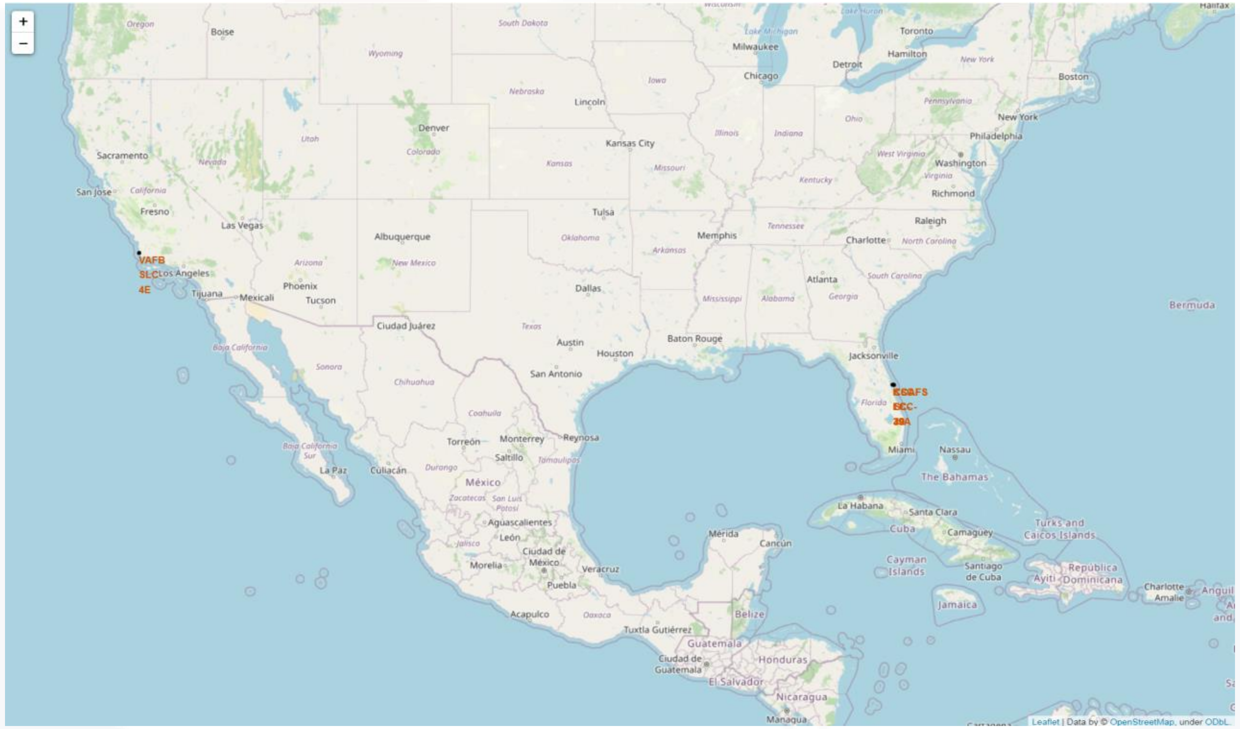
LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

- Here is the ranking of all landing outcomes between 2010-06-04 and 2017-03-20 in descending order.
- Landing outcome "No Attempt" has the most number of occurrences (10) whilst "Precluded (drone ship)" has the least number of occurrences (1) between the said time period.
- It is alerting that "No Attempt" has occurred 10 times during the said time period and should be taken into account for analysis.



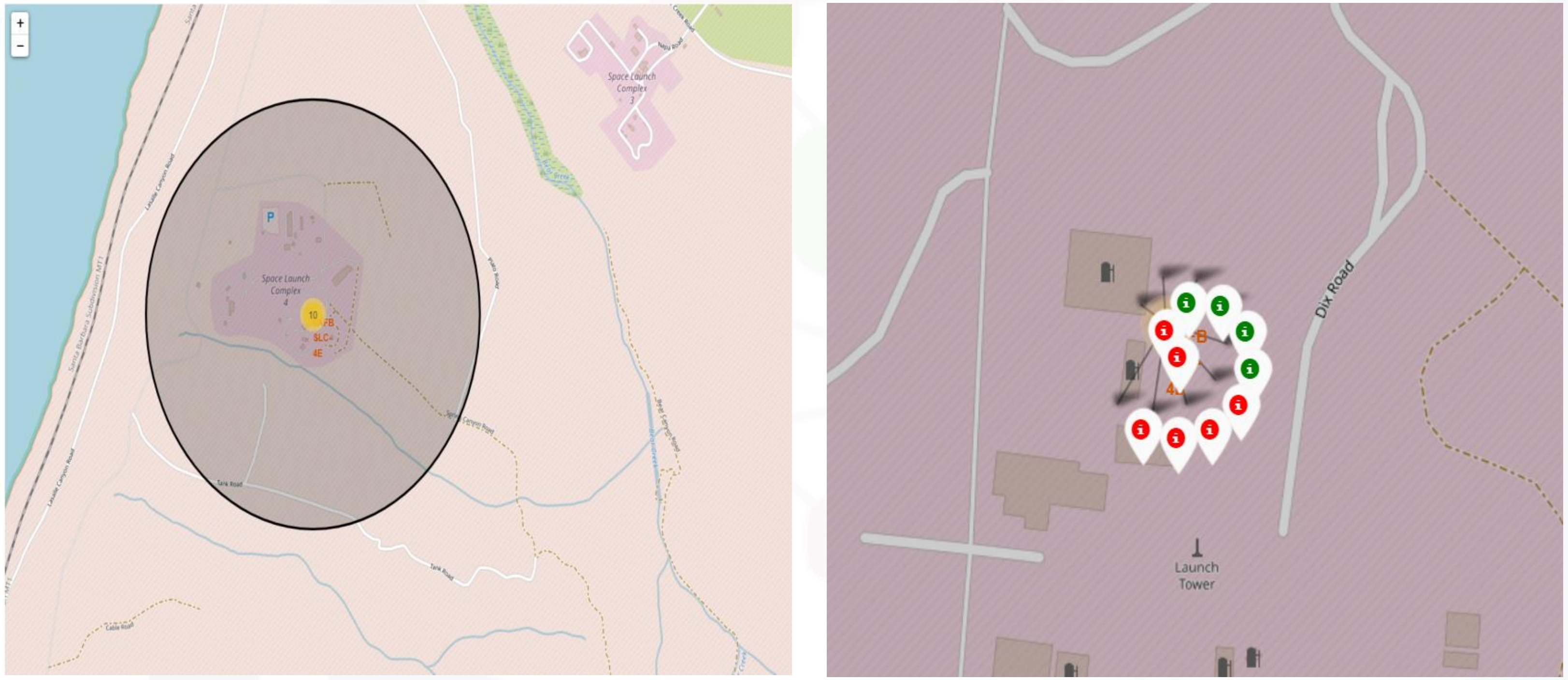
landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

INTERACTIVE VISUAL ANALYSIS WITH FOLIUM



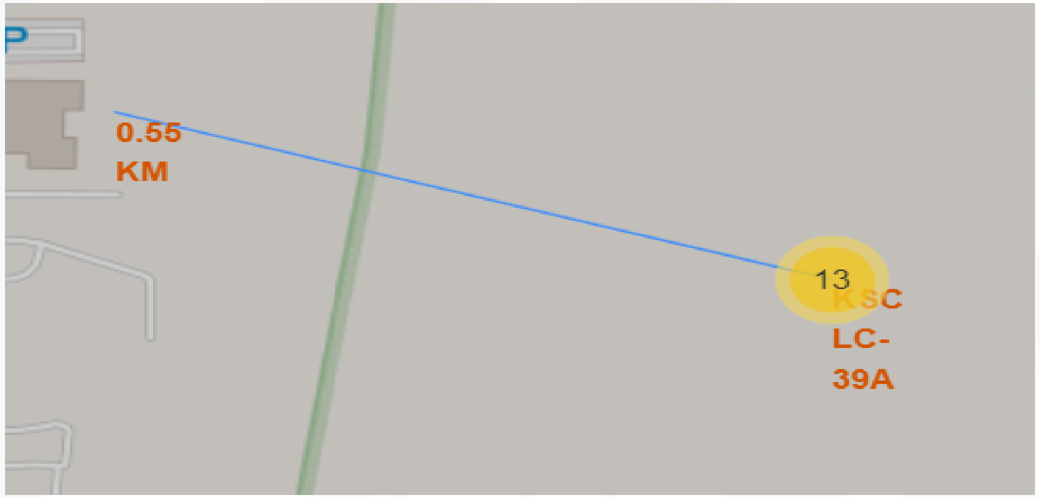
- This map shows all the launch sites.
- We can see that all of them are near the coast due to safety concerns, and we also noticed they are not too far from roads and rail tracks.

LAUNCH OUTCOMES BY SITE



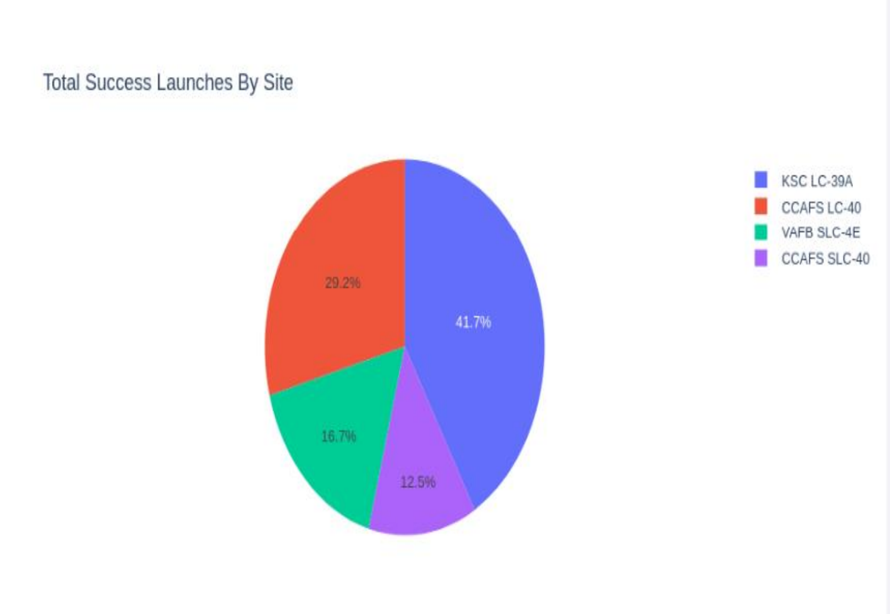
- Theses images focus on the landing outcomes of the launch site VAFB SLC-4E.
- Green exclamation marks indicate a successful landing whereas the red ones indicate a failure.

SAFETY AND LOGISTICS OF LAUNCH SITES



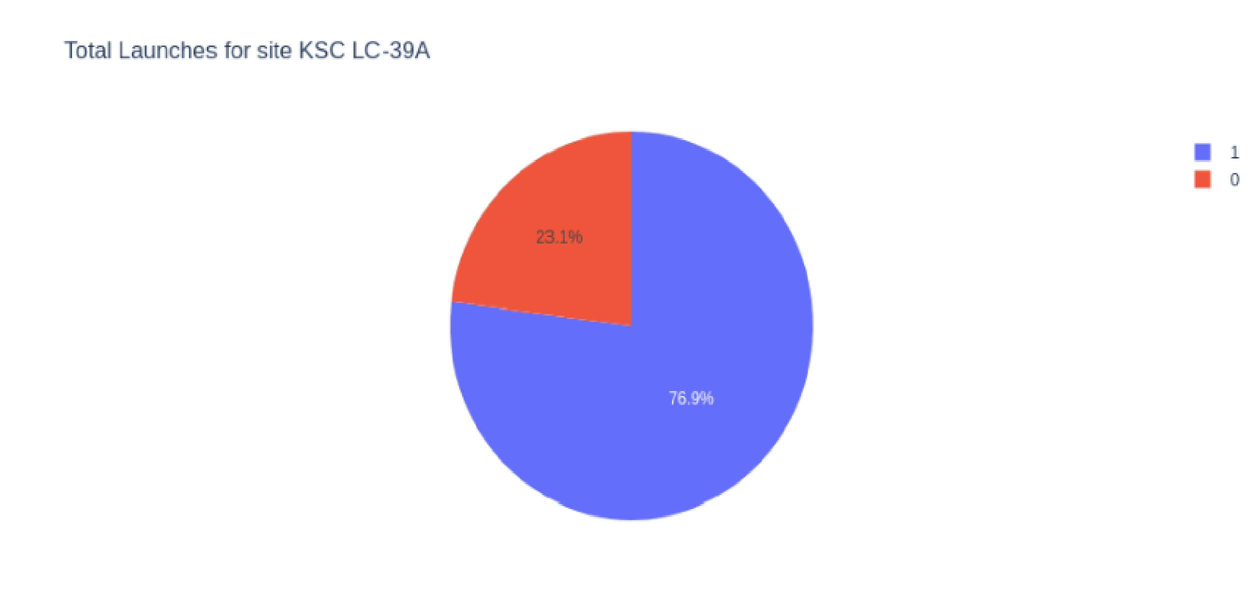
- Launch Site KSC LC-39A is pretty accessible and has good logistic aspects, as it is very near to a railroad.
- This Launch Site is somewhat far from crowded and inhabited areas, providing safety if anything were to go wrong.

DASHBOARD WITH PLOTLY DASH



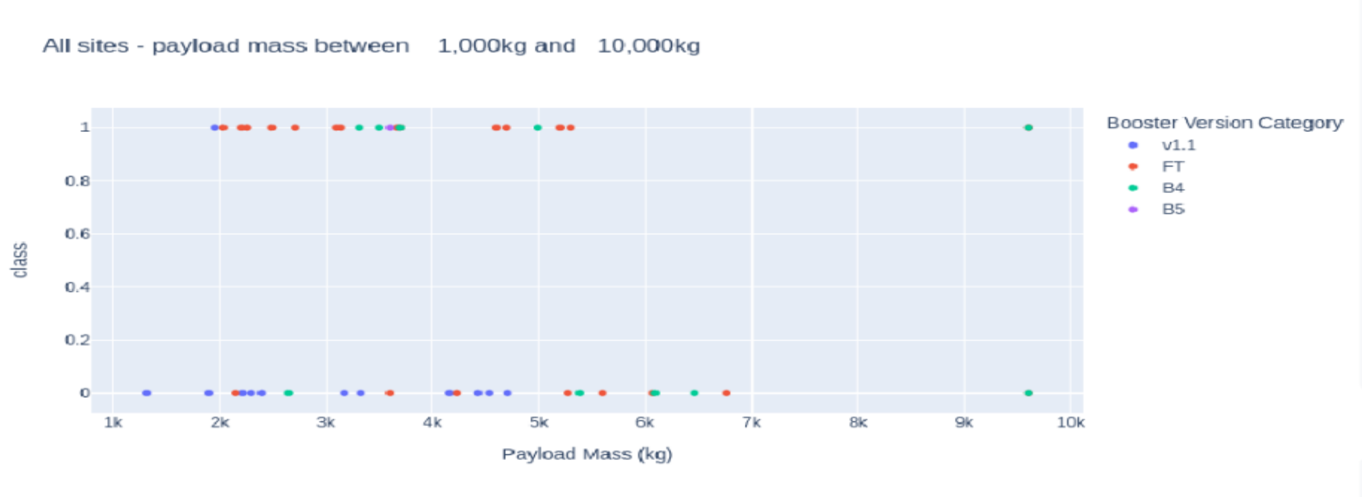
- This Pie Chart shows successful launches by each launch site.
- The launch site seems to be an important factor in the success of a launch.

DASHBOARD WITH PLOTLY DASH



- This chart shows total launches for the site KSC LC-39A
- 76.9% launches are successful at this site.

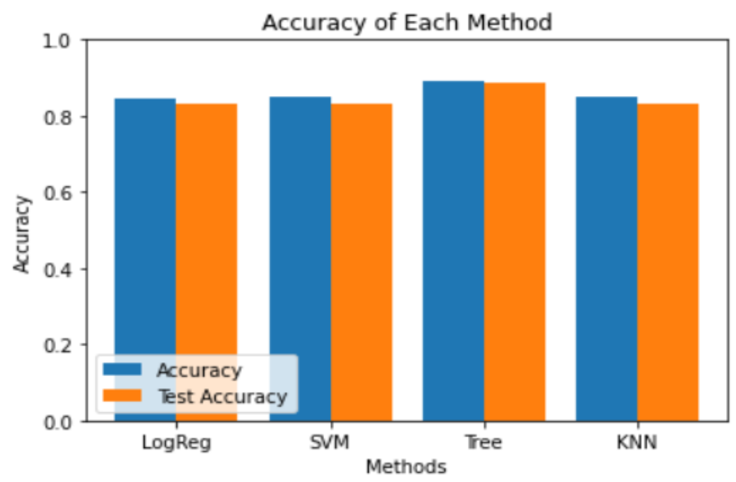
DASHBOARD WITH PLOTLY DASH



- This chart compares the launch outcomes with payload.
- Booster version V1.1 has the most number of failures as per this chart for the given range of Payload.
- FT Booster is very successful as long as it is under 6000kg of Payload.
- We can also notice there is not enough data available to estimate risks of launches over 7000kg.

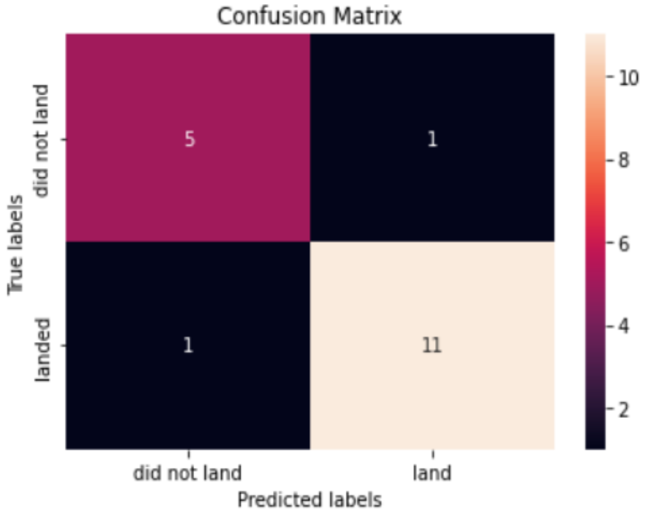
PREDICTIVE ANALYSIS RESULTS

- This bar graph shows the accuracy of each method.
- 4 different classification models were tested and their accuracies plotted next to them.
- We can see the Decision Tree has an edge with accuracies almost 89%



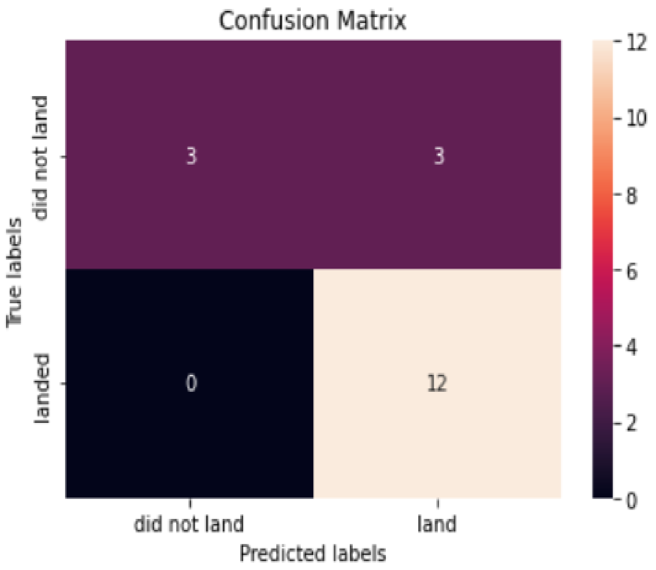
PREDICTIVE ANALYSIS RESULTS

- This graph shows **confusion matrix** for the **Decision Tree Classifier.**
- This confusion matrix proves the accuracy for the Tree Classifier, with high numbers for true positive and true negative compared to the false ones.



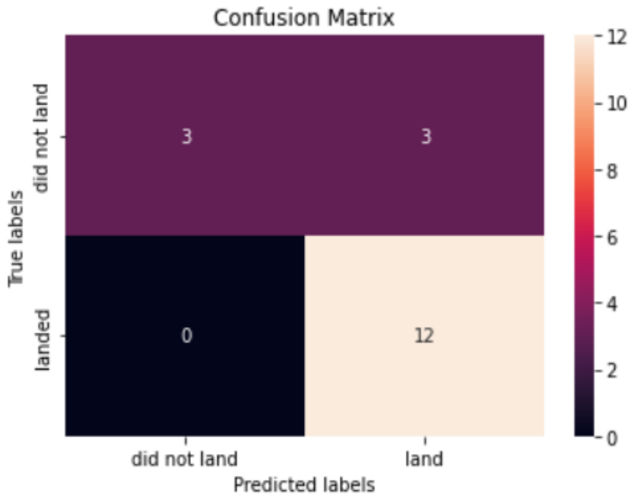
PREDICTIVE ANALYSIS RESULTS

- This graph shows **confusion matrix** for the **Logistic Regression Classifier.**
- This confusion matrix proves Logistic Regression is not as accurate as the Tree Classifier.



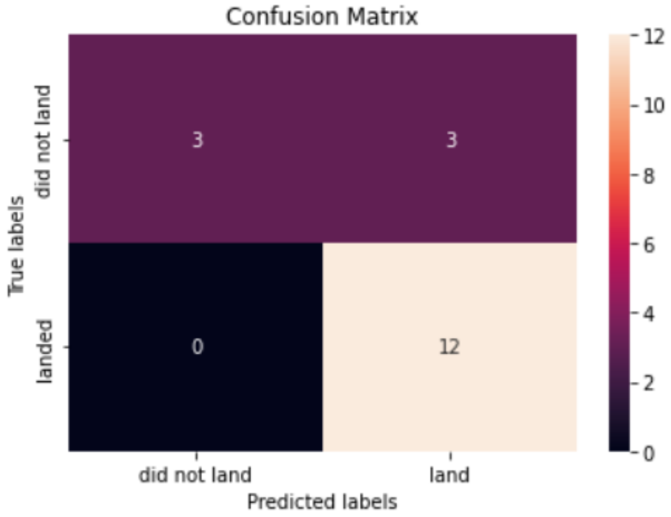
PREDICTIVE ANALYSIS RESULTS

- This graph shows **confusion matrix** for the **SUPPORT VECTOR MACHINE Classifier.**
- This confusion matrix proves SVM is not as accurate as the Tree Classifier.



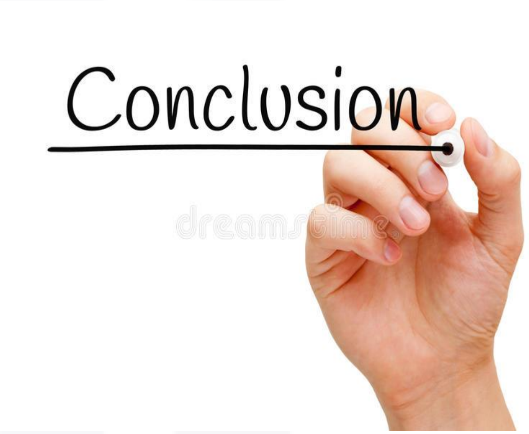
PREDICTIVE ANALYSIS RESULTS

- This graph shows **confusion matrix** for the **K Nearest Neighbor Classifier.**
- This confusion matrix proves KNN is not as accurate as the Tree Classifier.



CONCLUSIONS

- Data was collected from different sources, prepared and then analysed based on our requirements.
- The best launch site stood out to be **KSC LC-39A**
- Most of the mission outcomes are successful, but we see that the successful landing outcomes seem to improve over time.
- This can be due to more advancements and innovations in technology and better decision making.
- Launch sites are very important factor in determining the success of a landing.
- The **Decision Tree Classifier** can be used to predict the success of the landings and hence contribute towards increasing the profits.



Thank You!