IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Smit Shah

September 7, 2023

Outline

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

Executive Summary

- Methodologies that were used to analyze data:
  - Data Collection (by web scraping and SpaceX API)
  - Exploratory Data Analysis (EDA) - including data wrangling, data visualization, and interactive visual analytics
  - Machine Learning Prediction
- Summary of all results
  - It was possible to collect important data from public sources
  - EDA identified the best features to predict the success of launchings
  - Machine Learning Prediction showed the best model to predict which characteristics were important to drive this opportunity by the best way, using all the obtained data

Introduction

- Project background and context
  - This project’s goal was to evaluate the viability of the new company Space Y to compete with Space X
- Problems you want to find answers
  - The most efficient way to estimate the total cost for the launches through prediction of successful landings of the first stage of rockets
  - The prime location to make launches

Section 1

Methodology

Methodology

Executive Summary

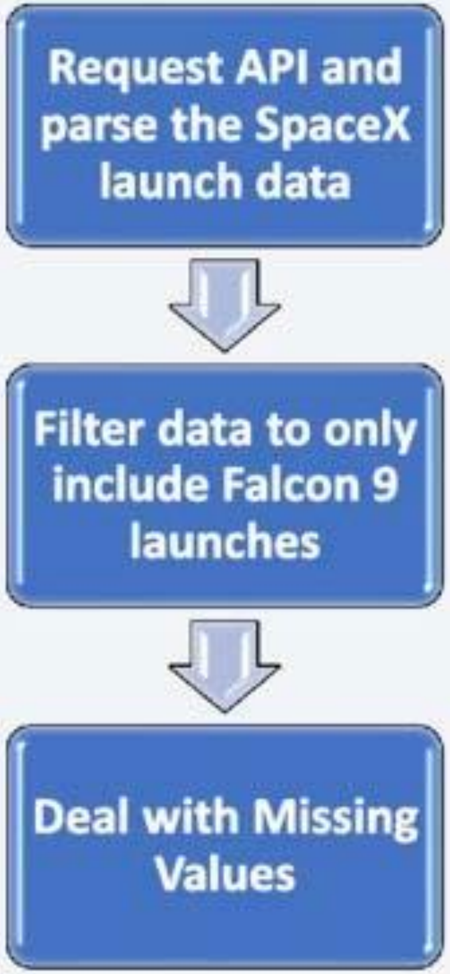
- Data collection methodology:
    - Data was obtained through 2 sources:
        - Space X API
        - WebScraping
- Perform data wrangling
        - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Data that was collected prior to this step was normalized, divided into training and test data sets, then evaluated by four different classification models

Data Collection

- Describe how data sets were collected.
  - Data sets were collected through the Space X API and from Wikipedia using web scraping technique

Data Collection – SpaceX API

- Space X offers an API where data can be obtained and then utilized by the public
  - The API was used in accordance to the flowchart (on the right)
- source code: https://github.com/smitshah02/Applied_DataScience_Capstone


Request API and parse the SpaceX launch data

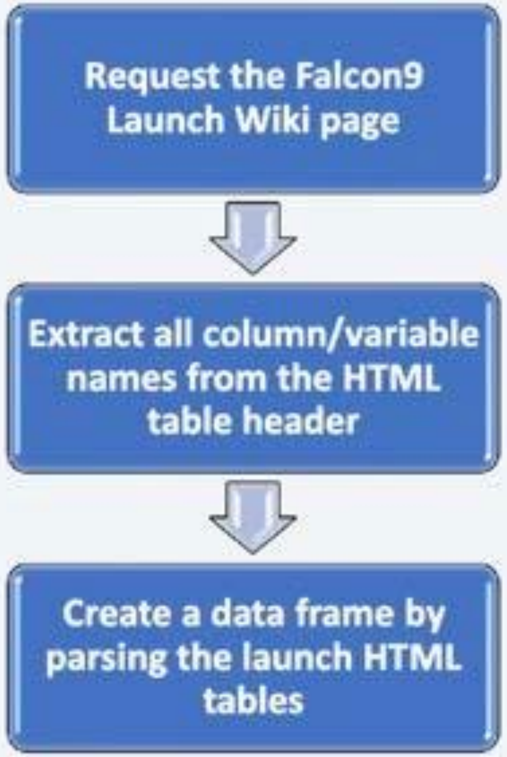
Filter data to only include Falcon 9 launches

Deal with Missing Values

Data Collection - Scraping

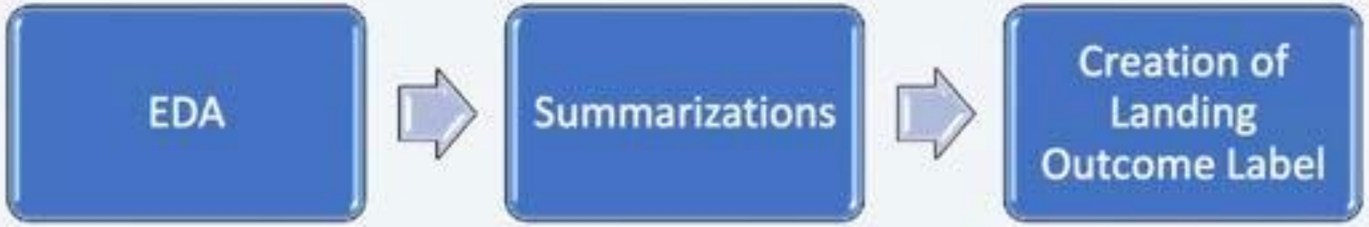
- Data from the launches by SpaceX can be found on Wikipedia
    - Data is downloaded from Wikipedia in accordance to the flowchart (on the right)

- source code: https://github.com/smitshah02/Applied_DataScience_Capstone


Request the Falcon9 Launch Wiki page
Extract all column/variable names from the HTML table header
Create a data frame by parsing the launch HTML tables

Data Wrangling

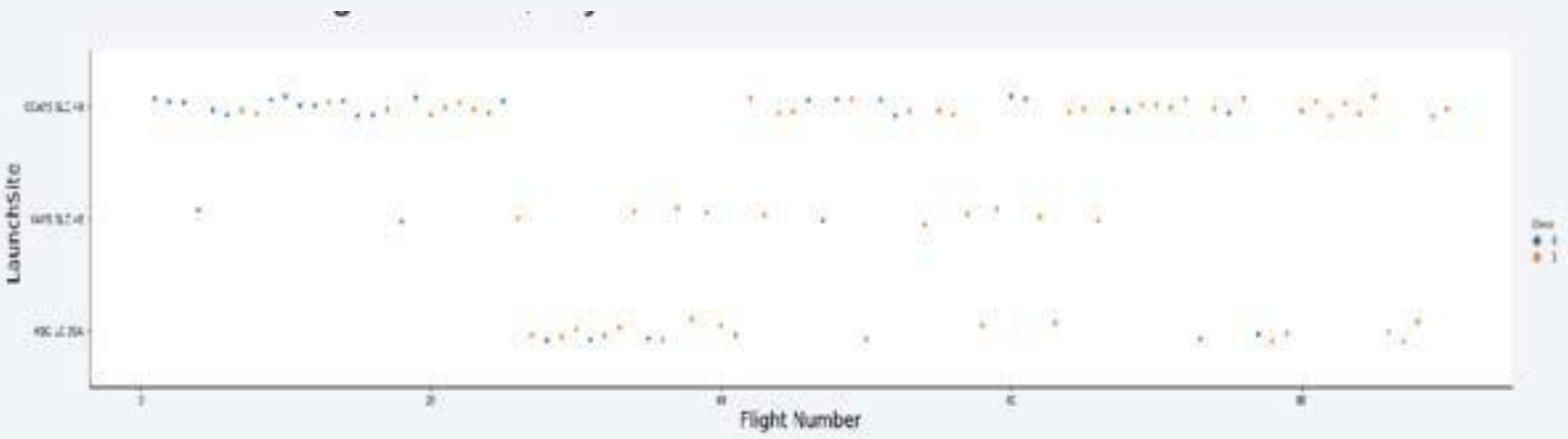
- EDA was performed on the dataset early on
- The summary launches per site, occurrence of each orbit, and occurrences of mission outcome per obit type were calculated
- The landing outcome label was created fro the Outcome column



- source code: https://github.com/smitshah02/Applied_DataScience_Capstone

EDA with Data Visualization

- Scatterplot and bar plots were used to visualize the relationship between pair of features in order to explore data



- source code: https://github.com/smitshah02/Applied_DataScience_Capstone

EDA with SQL

- Performed SQL queries:
  - names of the unique launch sites in the space mission
  - top 5 launch sites beginning with string ‘CCA’
  - total payload mass carried by boosters launched by NASA
  - average payload mass carried by booster version F9 v1.1
  - date of the first successful landing outcome in ground pad
  - names of the boosters with success in drone ship and with a payload mass between 4000 and 6000 kg
  - total number of successful and failure mission outcomes
  - names of booster versions which have carried the maximum payload mass
  - failed landing outcomes in drone ship, their booster versions, as well as launch site names
  - rank of the count of landing outcomes between two dates
- source code: https://github.com/smitshah02/Applied_DataScience_Capstone

Build an Interactive Map with Folium

- Markers, circles, lines, and marker clusters used with Folium Maps
  - markers indicate points (ex. launch sites)
  - circles indicate highlighted areas around specific coordinates (ex. NASA Johnson Space Centre)
  - marker clusters indicates groups events in each coordinate (ex. launches in a launch site
  - lines are used to indicate distances between two coordinates

- source code: https://github.com/smitshah02/Applied_DataScience_Capstone

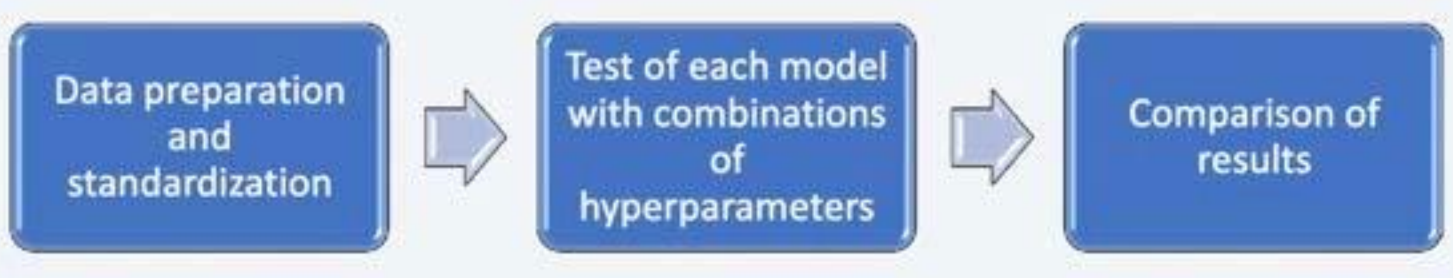
Build a Dashboard with Plotly Dash

- Graphs and Plots used to visualize data:
  - percentage of launches by site
  - payload range
- the combination of the above two allowed quick analysis between the relation of payloads and launch sites, helping to identify the best location to launch according to payloads

- source code: https://github.com/smitshah02/Applied_DataScience_Capstone

Predictive Analysis (Classification)

- the four classification models that were compared
  - logistic regression
  - support vector machine
  - decision tree
  - k-nearest neighbours



Data preparation and standardization

Test of each model with combinations of hyperparameters

Comparison of results

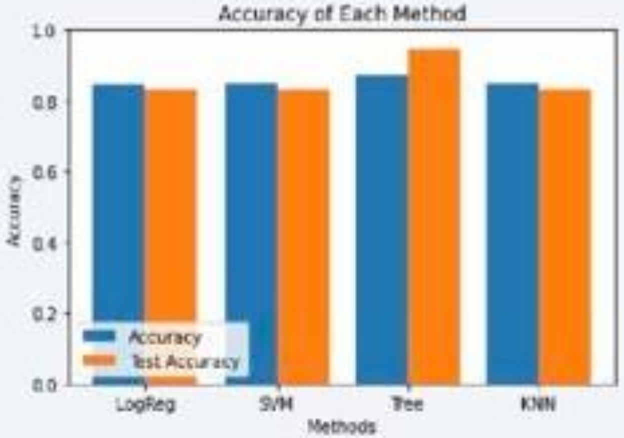
- source code: https://github.com/smitshah02/Applied_DataScience_Capstone

Results

- Exploratory data analysis results
  - SpaceX utilized four different launch sites
  - first launches were done to SpaceX and NASA
  - average payload of F9 v1.1 booster is 2928 kg
  - first success landing outcome occurred in 2015
  - many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
  - nearly all mission outcomes were successful
  - two booster versions failed at landing in drop ships
  - the number of landing outcomes has become better over time

Results

- using interactive analytics was possible to identify that launch sites used to be in safety places, near sea, for example and have a good logistic infrastructure around
- most launches occur at east coast launch sites



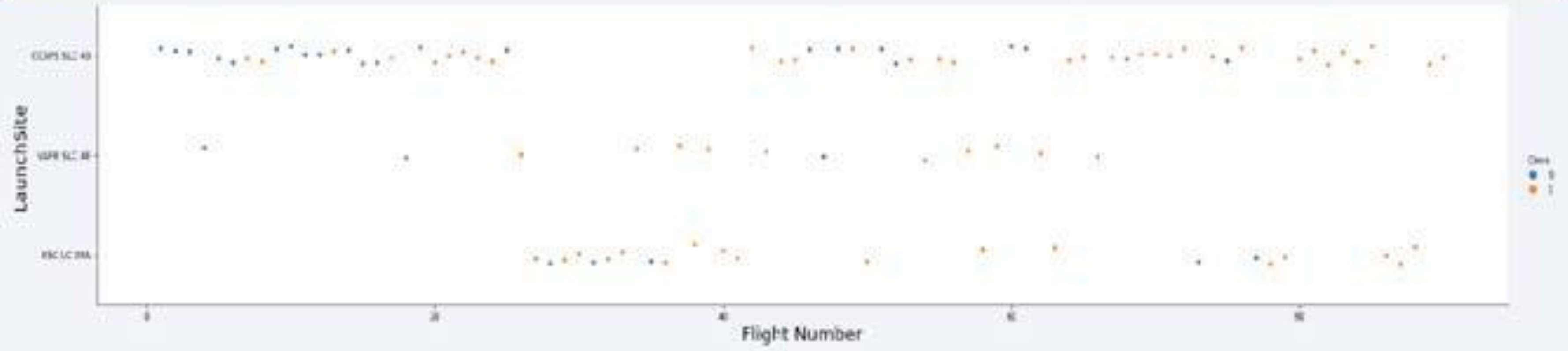
- predictive analysis showed that decision tree classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%

Section 2

Insights drawn from EDA

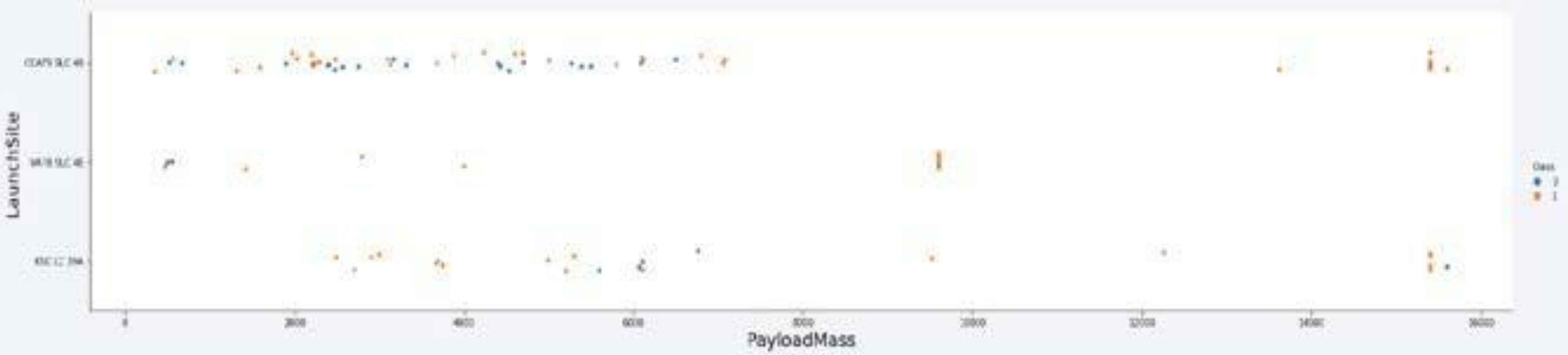
Flight Number vs. Launch Site

- In relation to the plot below, it is possible to verify that the best launch site nowadays is CCAF5 SLC 40 - where most of the recent launches were successful
- In second place: VAFB SLC 4E; In third place: KSC LC 39A
- The general success rate has improved over time



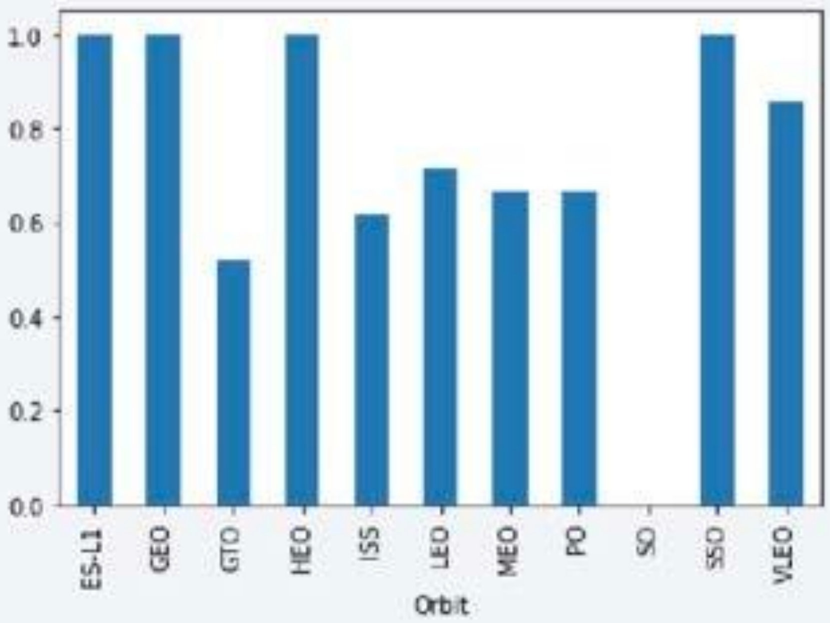
Payload vs. Launch Site

- Payloads greater than 9000 kg showed excellent success rates
- Payloads exceeding 12000 kg seemed to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites



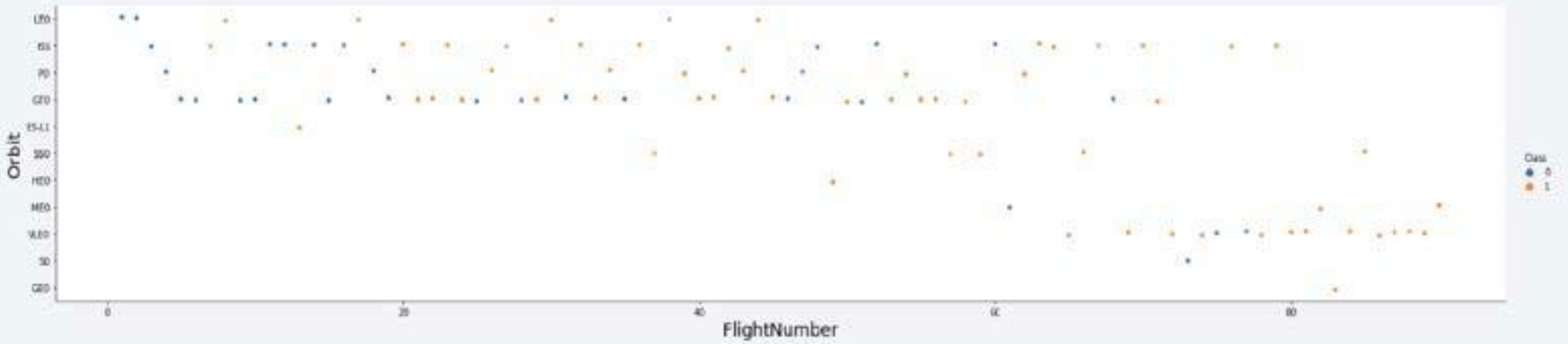
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits: ES-L1, GEO, HEO, and SSO
  - followed by: VLEO (greater than 80%) and LFO (greater than 70%)



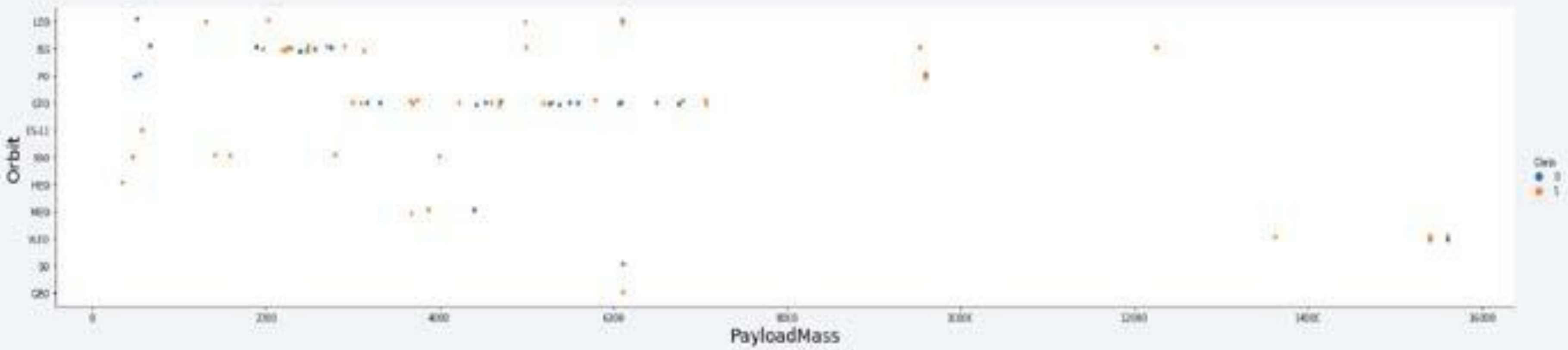
Flight Number vs. Orbit Type

- success rate improved over time to all orbits
- VLEO orbit appears to be a new business opportunity due to its recent increase of its frequency



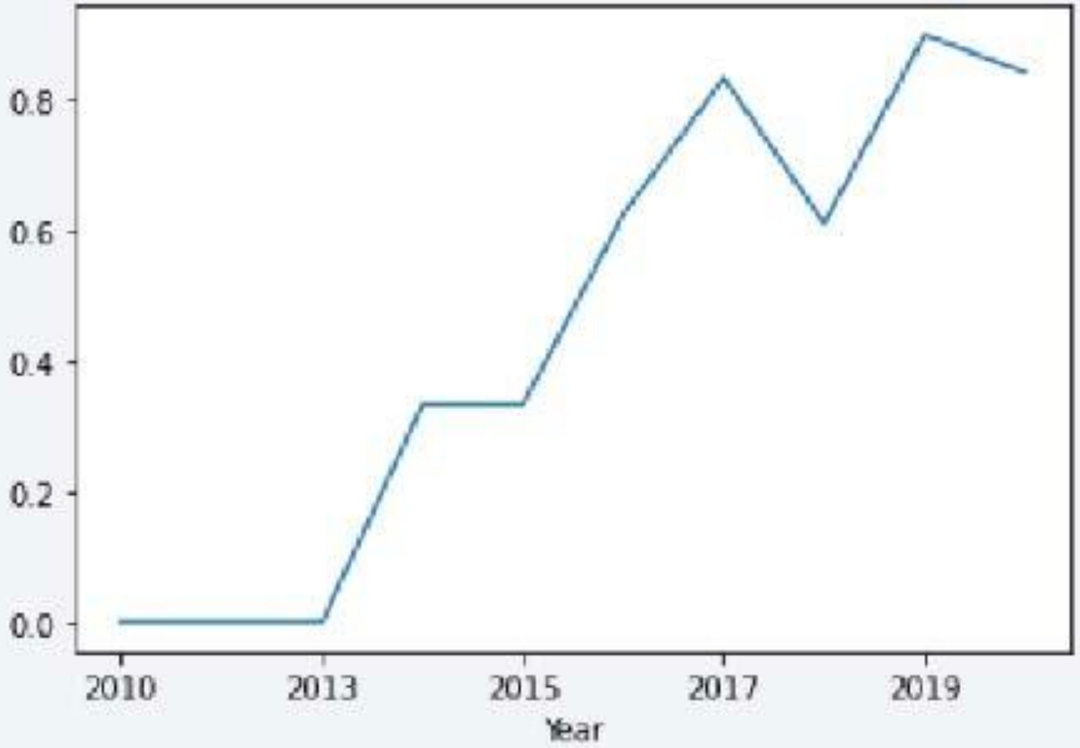
Payload vs. Orbit Type

- there is no relation between payload and success rate to orbit GTO
- ISS orbit has the vastest range of payload and a good rate of success
- there are few launches to the orbits SO and GEO



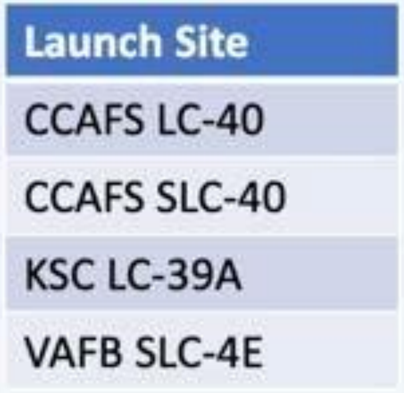
Launch Success Yearly Trend

- success rate begins to increase in 2013 and continues until 2020
- can be assumed that the first three years was a period of adjustment and improvement to their technology



All Launch Site Names

- Names of the four unique launch sites:

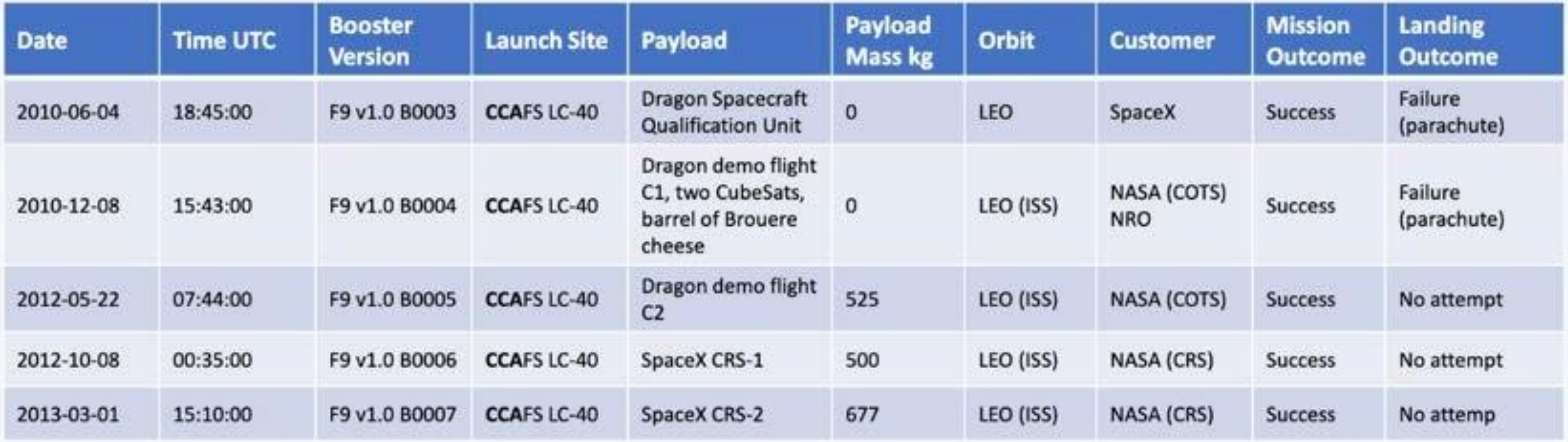


Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- obtained by selecting unique occurrences of “launch_site” values from the dataset

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with ‘CCA”

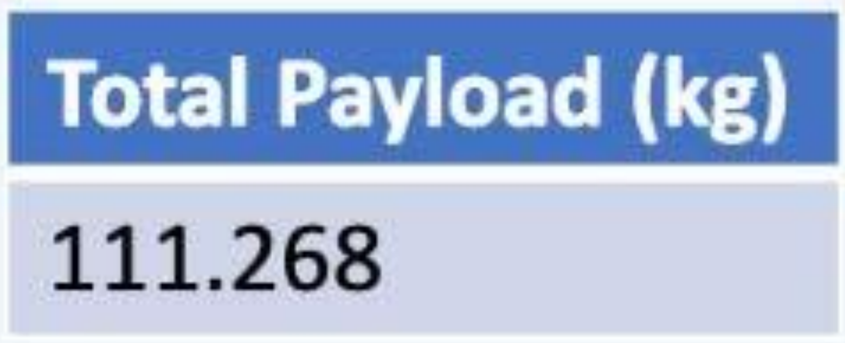


Date	Time UTC	Booster Version	Launch Site	Payload	Payload Mass kg	Orbit	Customer	Mission Outcome	Landing Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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 attemp

Total Payload Mass

- Total payload carried by boosters from NASA: 111.268 kg

- Total payload given above is calculated by summing all payloads whose codes contain ‘CRS: which corresponds to NASA

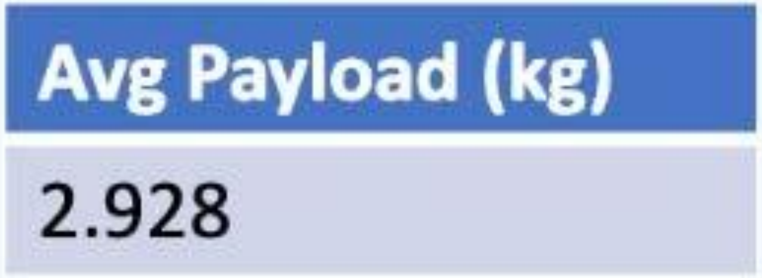


Total Payload (kg)
111.268

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: 2.928 kg

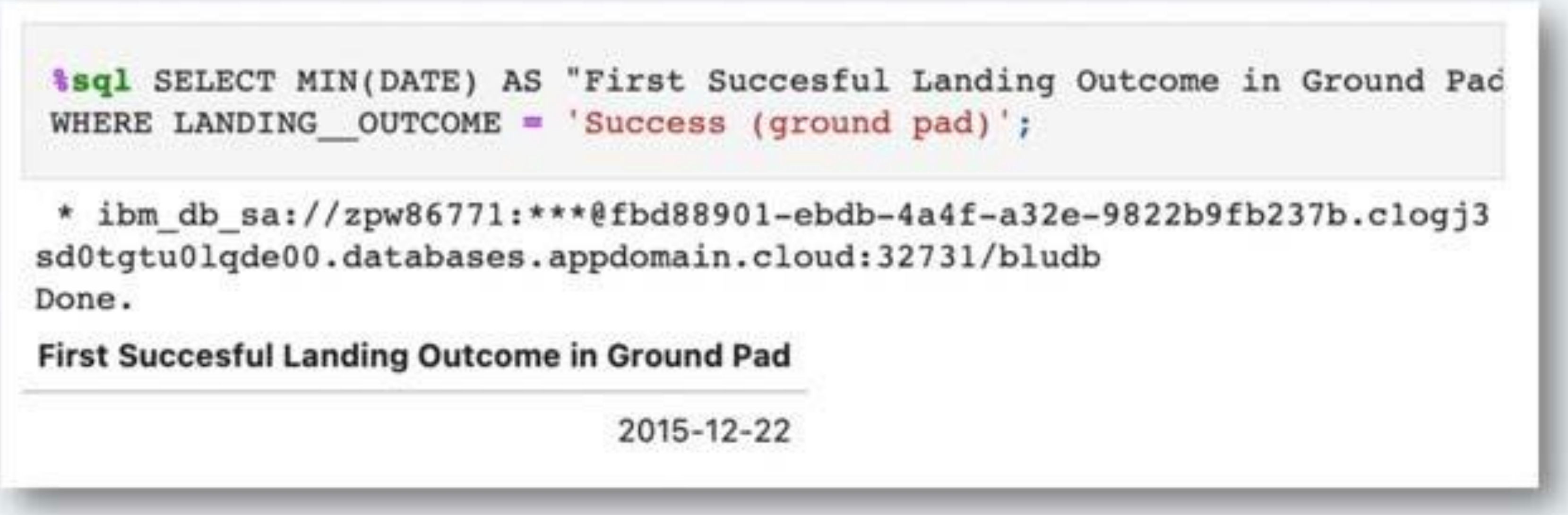
- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2928 kg



Avg Payload (kg)
2.928

First Successful Ground Landing Date

- the function min() was used to find the result
- observed that the dates of the first successful landing outcome on ground pad was December 22, 2015



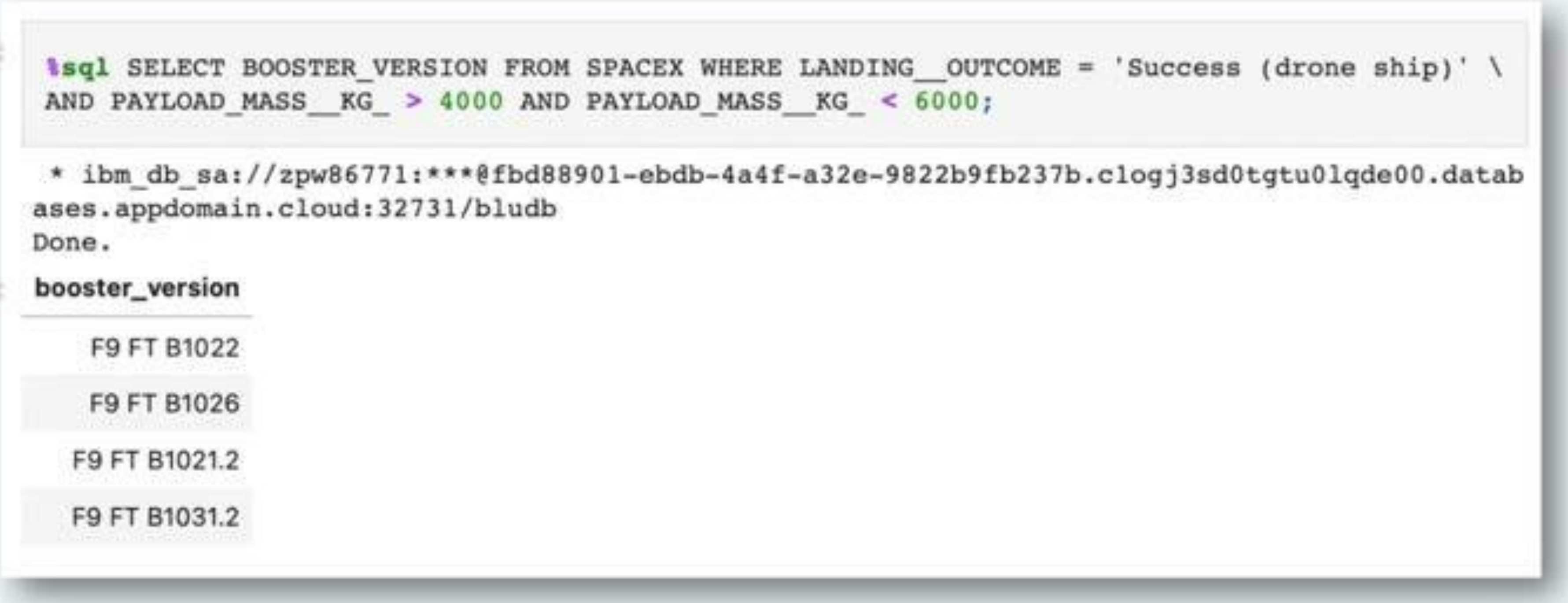
```
%sql SELECT MIN(DATE) AS "First Succesful Landing Outcome in Ground Pad
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.clogj3
sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

First Succesful Landing Outcome in Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- the WHERE clause was used to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000



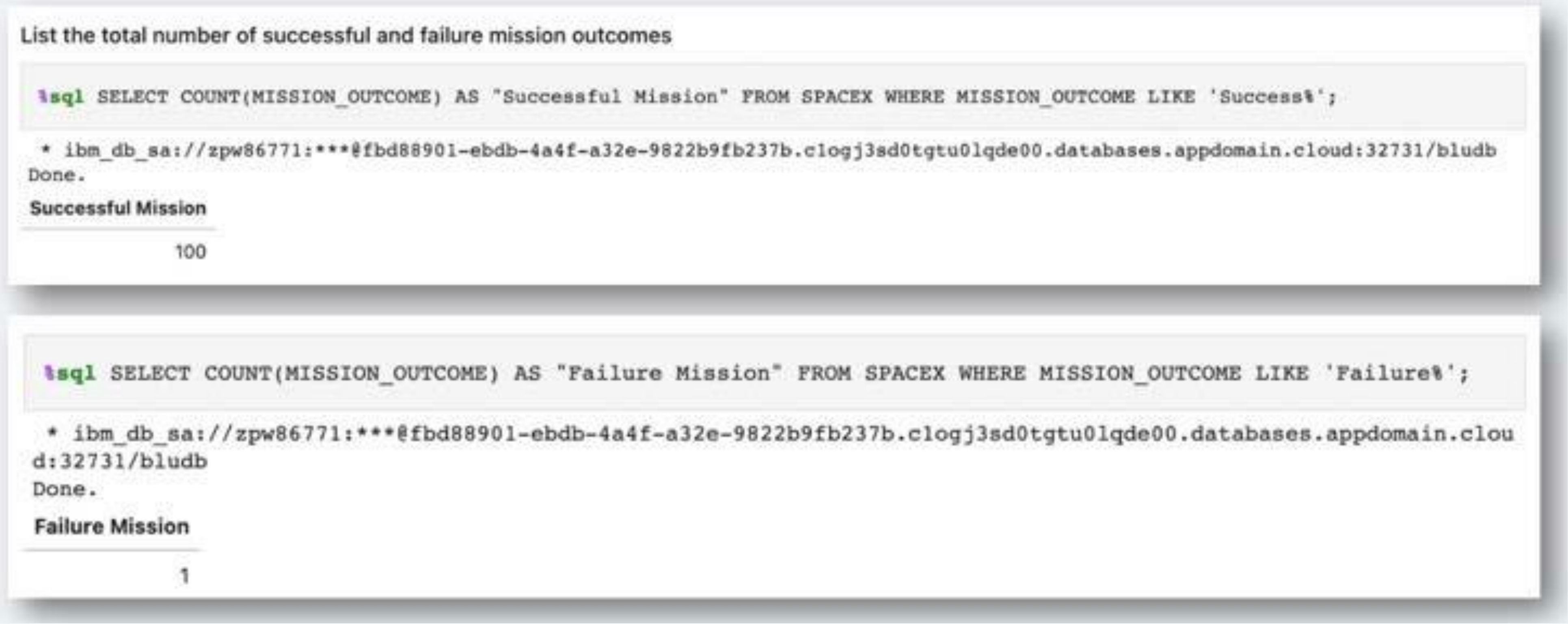
```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.clogj3sd0tgtu0lqde00.datab
ases.appdomain.cloud:32731/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

- Wildcard like ‘%’ was used to filter for WHERE MissionOutcome was a success or a failure



List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(MISSION_OUTCOME) AS "Successful Mission" FROM SPACEX WHERE MISSION_OUTCOME LIKE 'Success%';
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Successful Mission
100

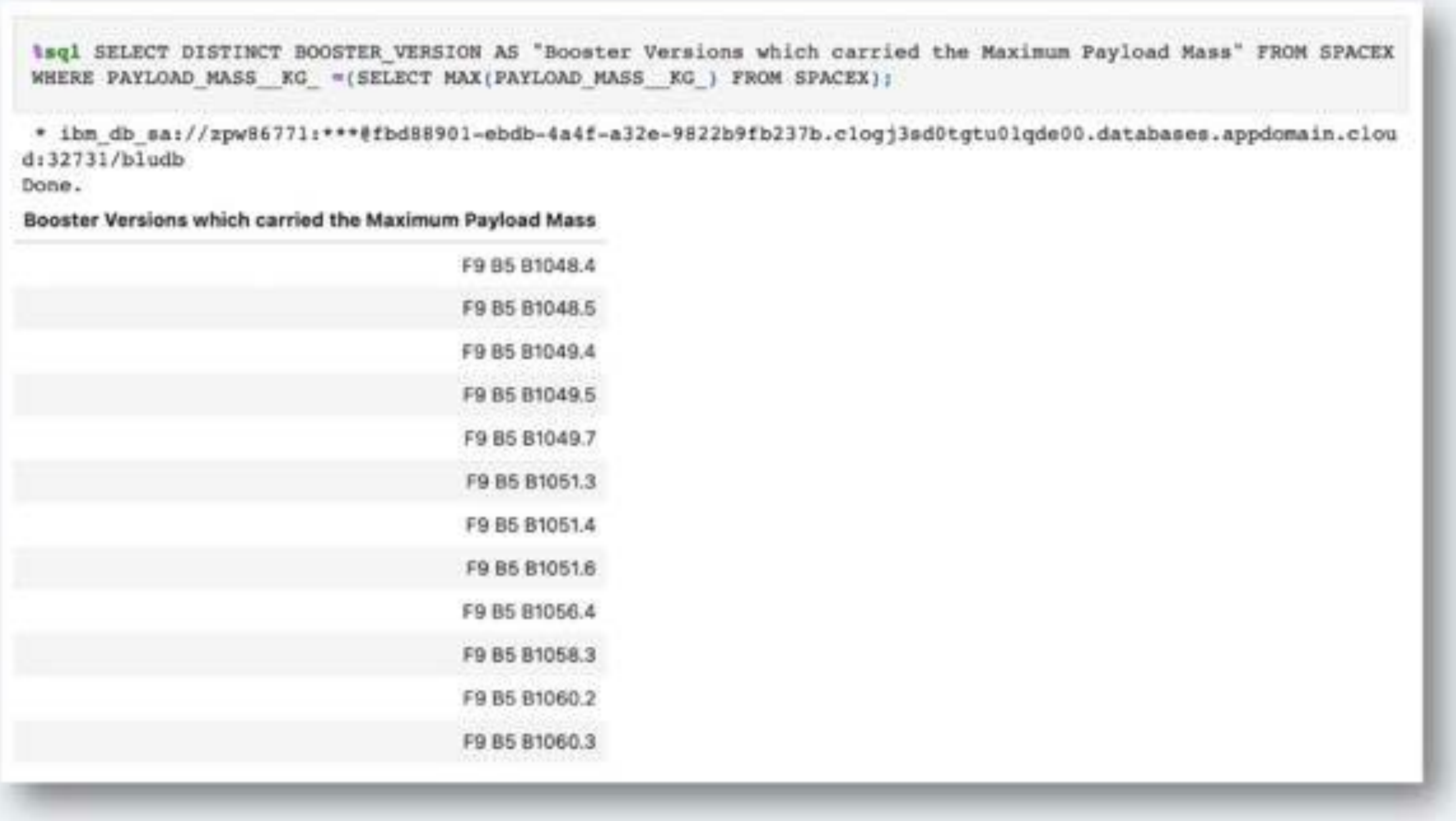
```
%sql SELECT COUNT(MISSION_OUTCOME) AS "Failure Mission" FROM SPACEX WHERE MISSION_OUTCOME LIKE 'Failure%';
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.clogj3sd0tgtu0lqde00.databases.appdomain.clou
d:32731/bludb
Done.
```

Failure Mission
1

Boosters Carried Maximum Payload

- The booster to carry the maximum payload using a subquery in the WHERE clause and the MAX() function was determined



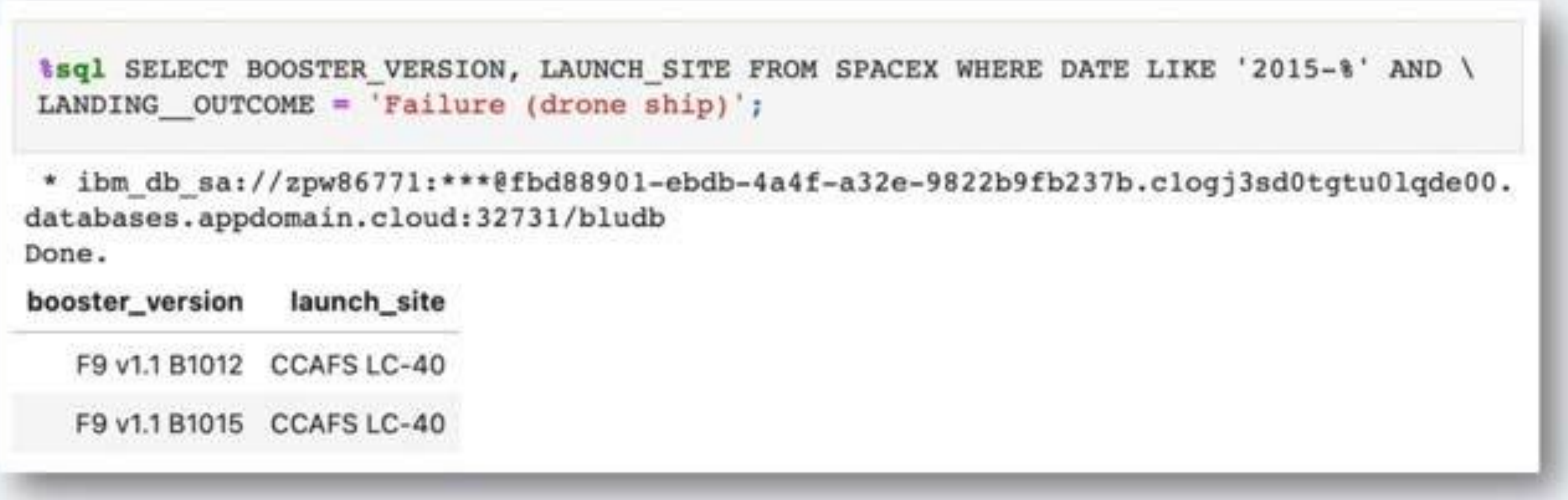
```
%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEX
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.clogj3sd0tgtu0lqde00.databases.appdomain.clou
d:32731/bludb
Done.
```

Booster Versions which carried the Maximum Payload Mass
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

2015 Launch Records

- A combination of the WHERE clause, LIKE, AND, and BETWEEN conditions was used to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015



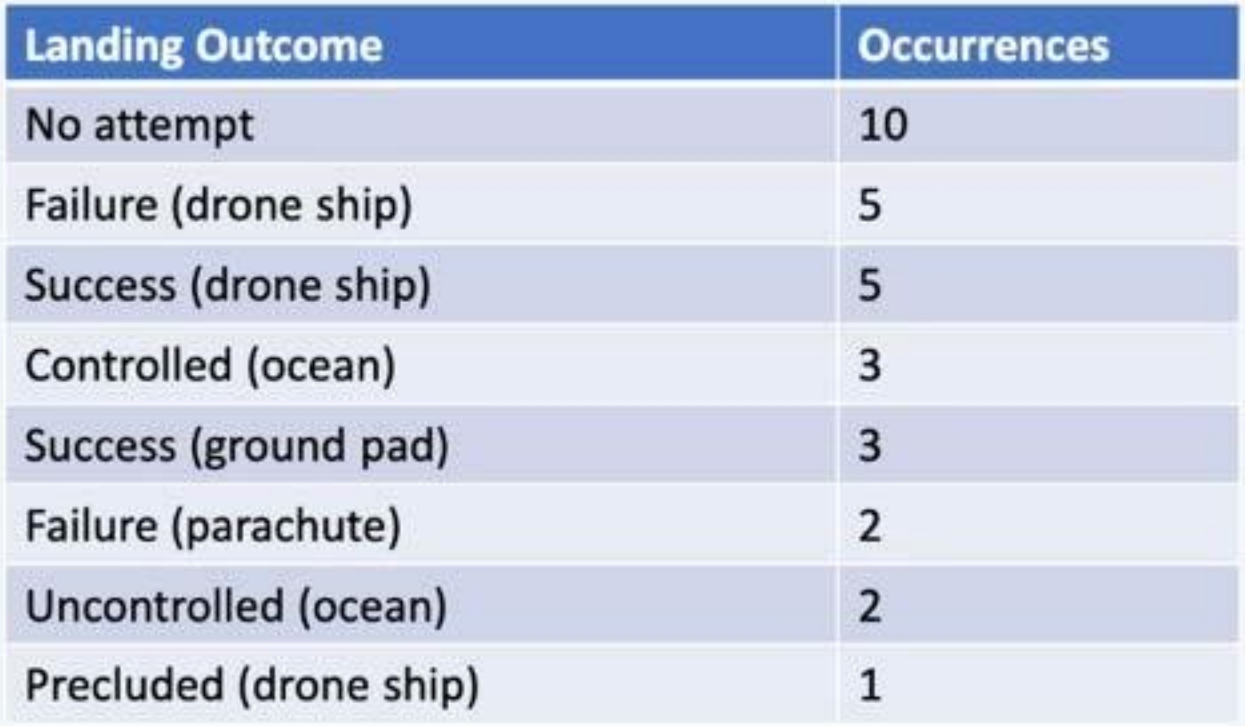
```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE DATE LIKE '2015-%' AND \
LANDING__OUTCOME = 'Failure (drone ship)';
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.clogj3sd0tgtu0lqde00.
databases.appdomain.cloud:32731/bludb
Done.
```

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

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

- Ranking of all landing outcomes between the 2010 and 2018
- This view of data gives the alert that “No attempt” must be taken into account



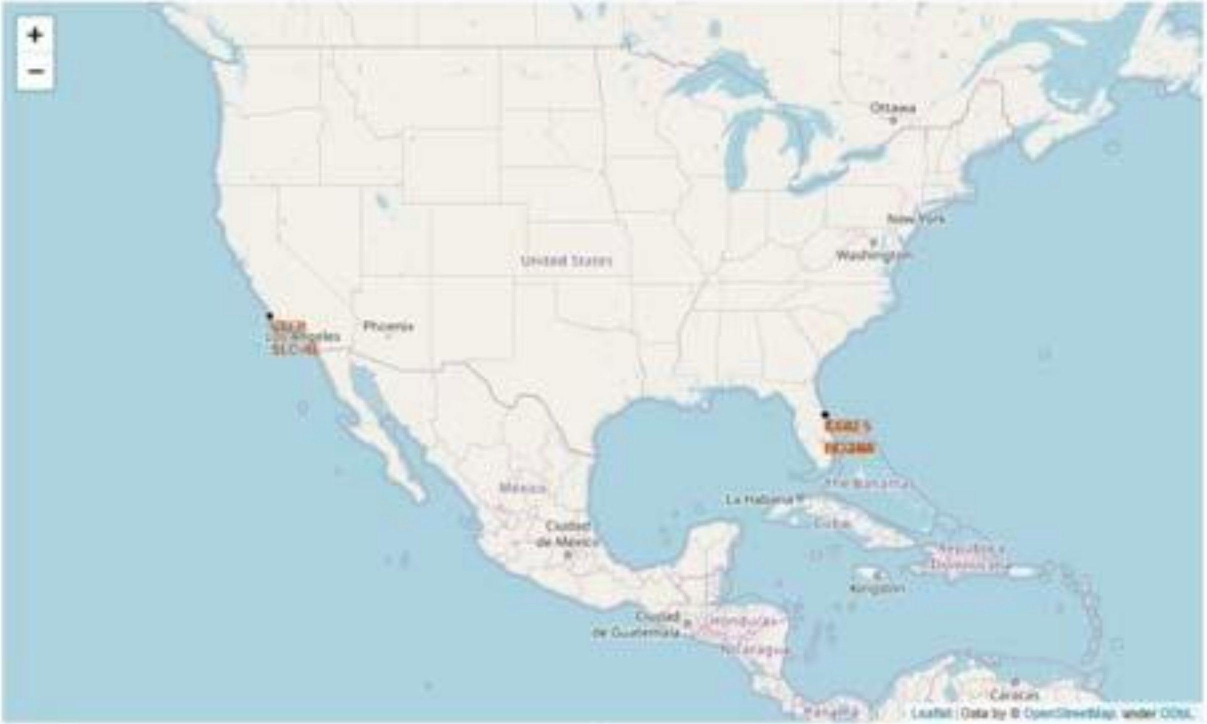
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All launch sites

- Launch sites are near the sea
  - probably by safety
  - not too far from roads and railroad



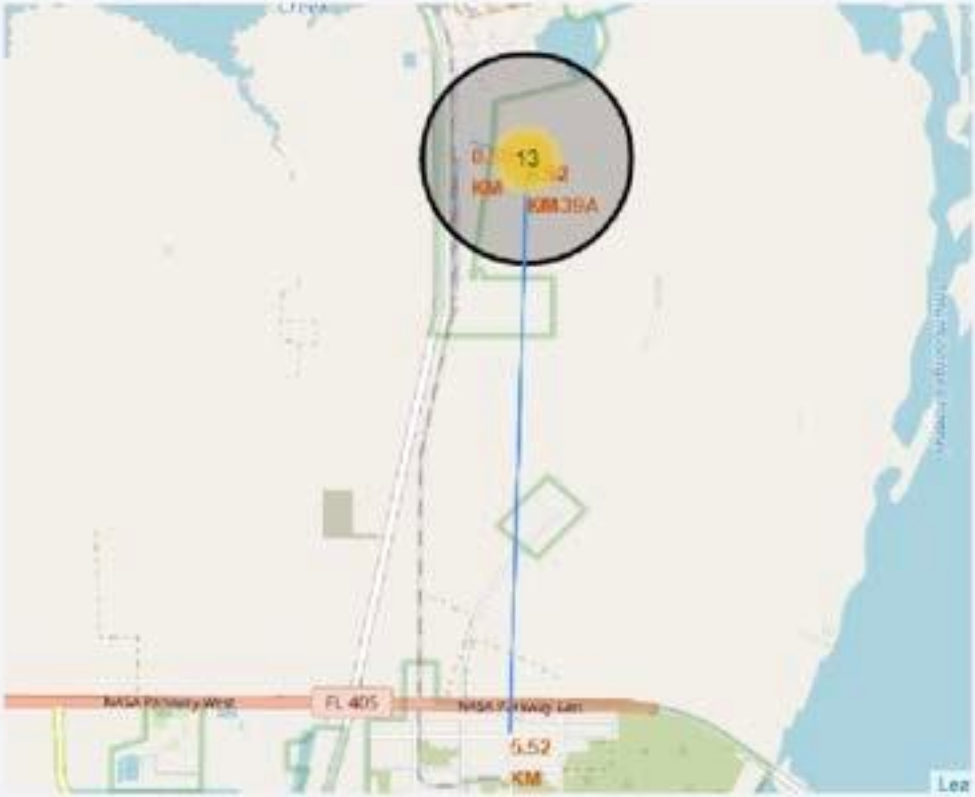
Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes
- Green markers indicate successful and red ones indicate failure

Las Vegas
10
Angeles
Tijuana
Mexicali
Phoen
13
LC-39A
Memphis
Charlotte
Atlanta
Rouge
Jacksonville
46
AFS

Logistics and Safety

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas

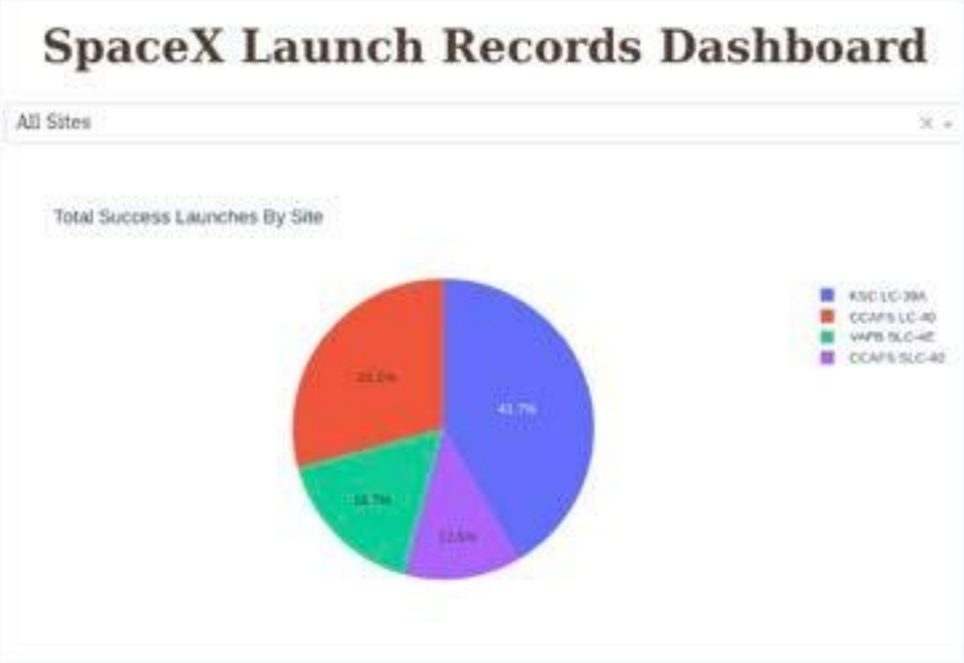


Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

- The place from where launches are done seems to be a very important factor of success of missions

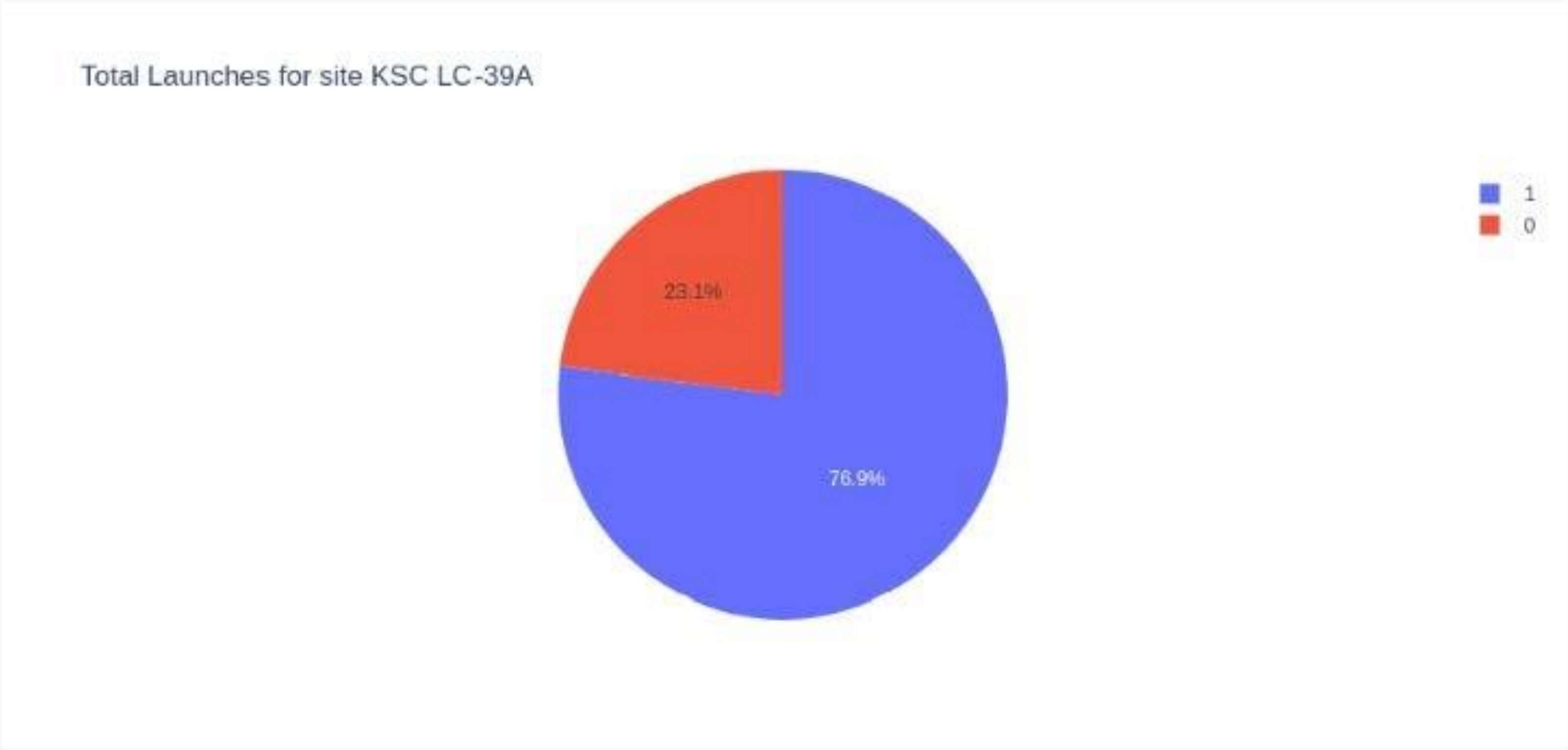


SpaceX Launch Records Dashboard

All Sites

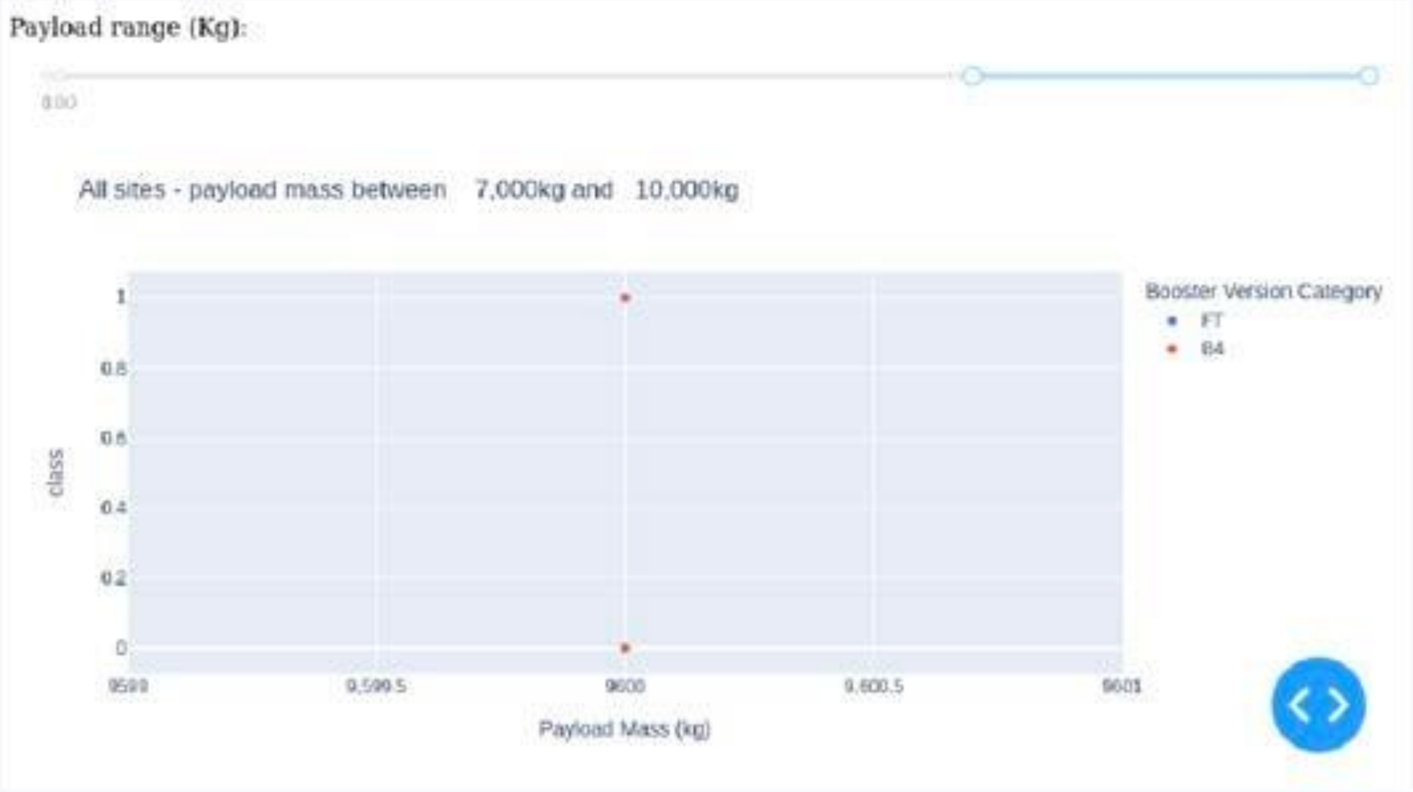
Launch Success Ration for KSC LC-39A

- 76.9% of launches are successful in this site



Payload vs. Launch Outcome

- To estimate the risk of launches exceeding 7000 kg cannot be predicted due to the lack of data



Payload range (Kg):

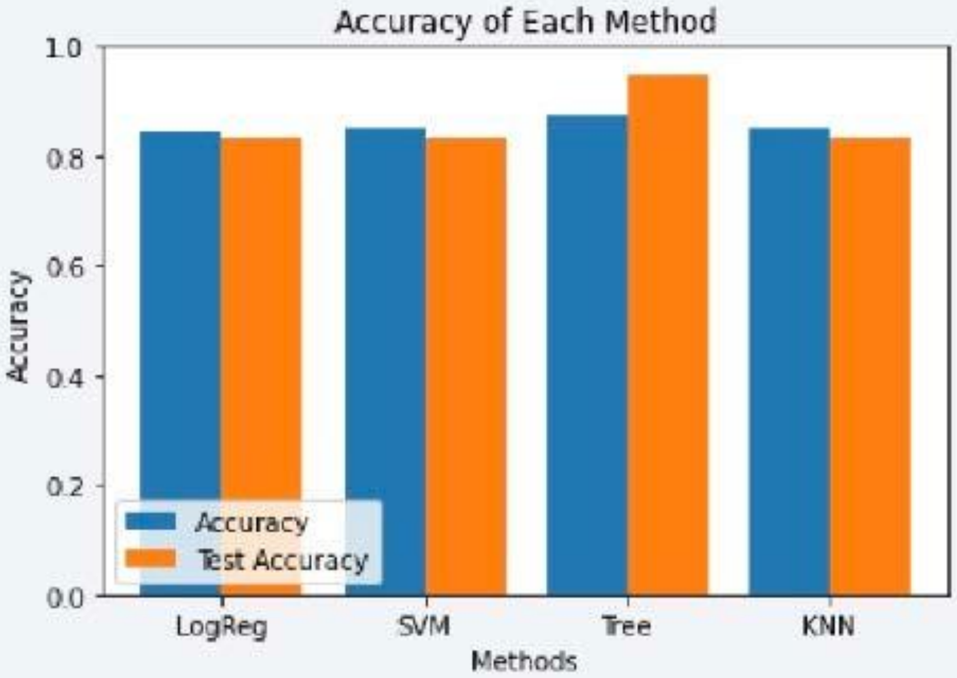
All sites - payload mass between 7,000kg and 10,000kg

Section 5

Predictive Analysis (Classification)

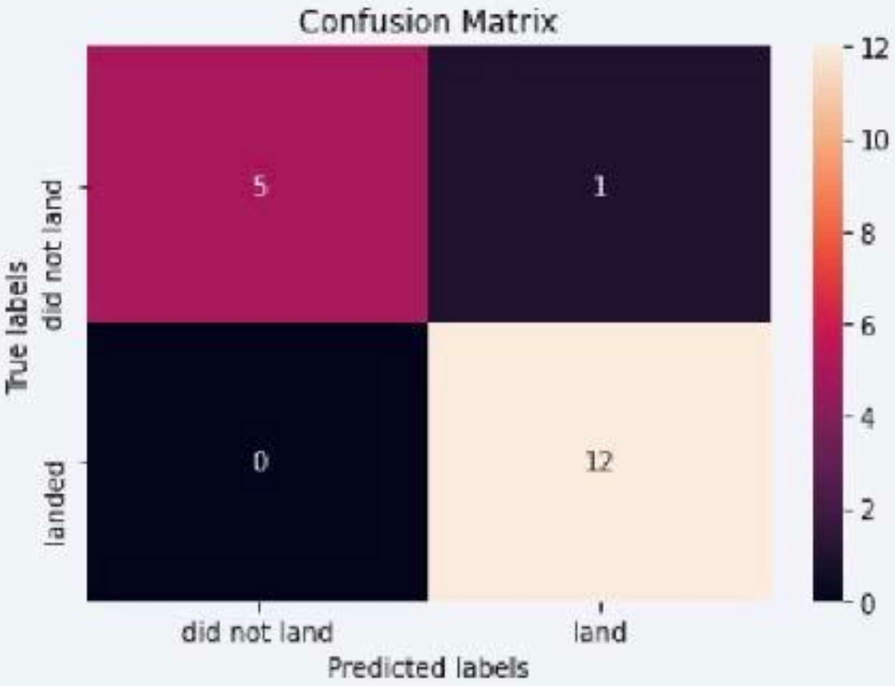
Classification Accuracy

- Four classification models were tested and their accuracies are plotted side-by-side
- The model with the highest classification accuracy is Decision Tree Classifier - an accuracy of over 87%



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones



Conclusions

- Different data sources were examined
- The prime launch site is KSC LC-39A
- Launches exceeding 7000 kg are less risky
- Although the majority of mission outcomes are successful, successful landing outcomes seem to improve over time, according to the evolution of processes and rockets
- Decision Tree Classifier can be used to predict successful landings and increase profits

Appendix

Thank you!