IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Julian Carro Verdia
05/09/2023

Outline

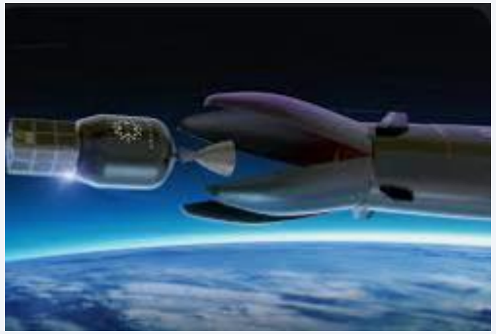
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies:
  - Data collection from CVS, SQL and web (webscrapping).
  - Data wrangling: dealing with missing values, standardize values.
  - Exploratory Data Analysis (EDA):
    - Visualization methods like ploting scatter plots, bar charts, interactive charts.
  - Predictive analysis: Calssification.
- Summary of all results
  - Success of mission based on mass, launch site, orbit, etc.
  - From 2013 to 202 the missions' outcomes improved.

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone: Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets.
    - Perhaps the most successful is **SpaceX**.
        - SpaceX's accomplishments: Sending spacecraft to the International Space Station.
        - Differentiation: rocket launches are relatively inexpensive (reuse the first stage). SpaceX advertises Falcon 9 rocket launches cost 62 million dollars; other providers cost upwards of 165 million dollars each. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Spaces X's Falcon 9 launch like regular rockets



Introduction

- Study will be based on another company called “SpaceY” and below are the problems to find answers to:
    - Determine the price of each lunch.
    - Determine if SpaceX will reuse the first stage.
    - Determine if Falcon’s 9 second stage will land.
    - Explain how to choose an optimal launch site.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from an API and loaded into coding environment (Jupyter Notebook) using Python language and several libraries such as Pandas, Numpy, Matplotlib, Seaborn, Sci-kit Learn, etc.
- Perform data wrangling
  - Data was processed by the following general steps:
    - Identify percentage of missing values and fill them with mean values from rest of data.
    - Identify type of data for each occurrence.
    - Calculate several key information (i.e.: number of launches per site, number and occurrence of each orbit, etc.).

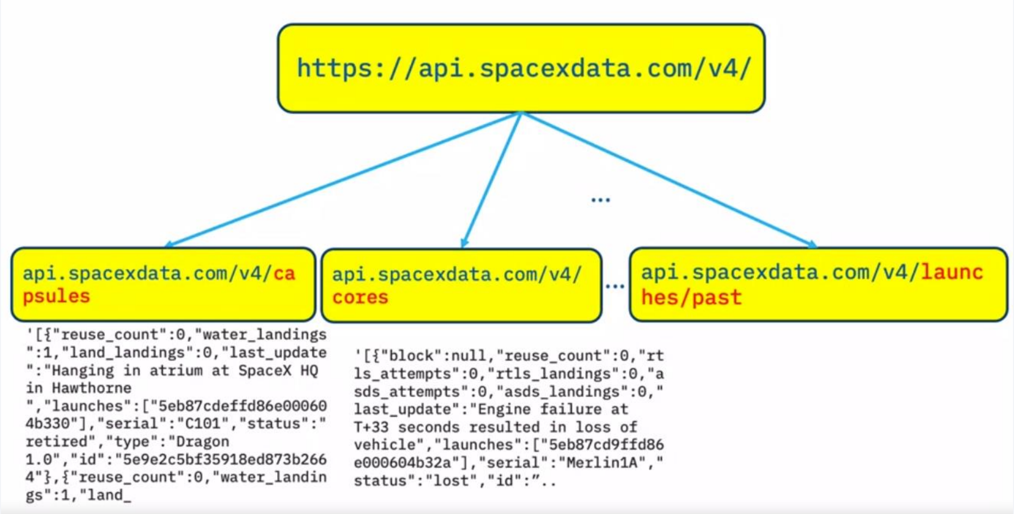
Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Standarize data in order to get valuable insights.
    - Split data into training and testing subsets.
    - Train and evaluate accuracy of different classification models.

Data Collection

- Data sets were collected via an endpoint or URL from REST API from SpaceX with launch data and from a specific web page via web-scrapping methods.
- This endpoint is targeted via Python commands and loaded into the coding environment as a csv (comma separated values) file and then converted to a table that this language can manage.



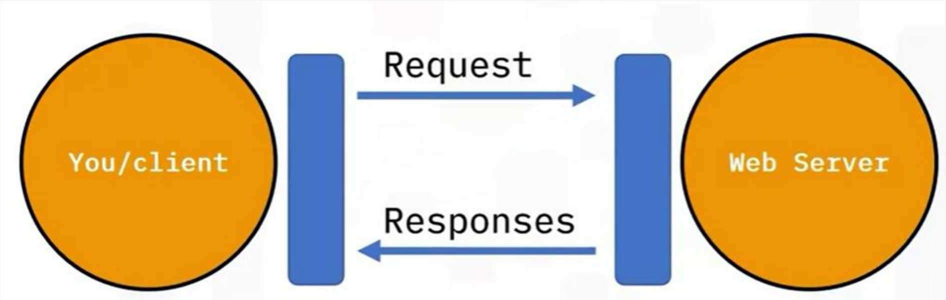
Data Collection – SpaceX API

- Basically a series of communication standards and functions from libraries are structured and performed to retrieve data from a server or database into the Jupyter notebook where the data can be filtered, structured and calculations can be made.

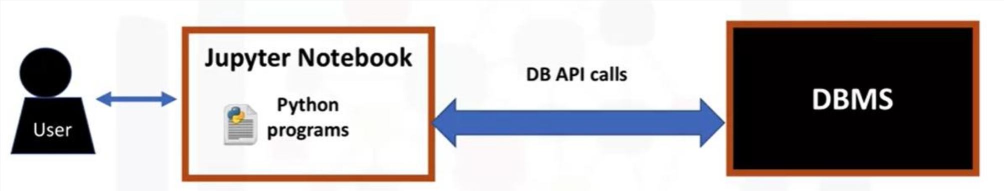
- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation

Basic diagram of REST API call.



Basic diagram of database call.

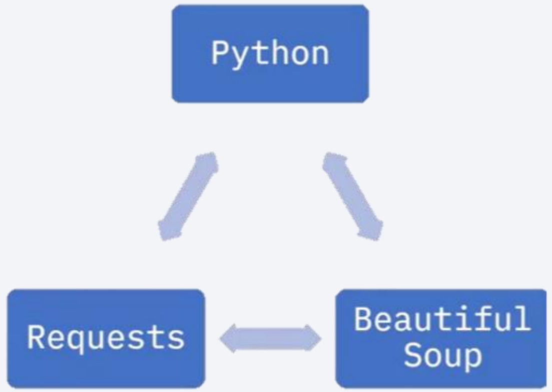


Data Collection - Scraping

- Basically a series of communication standards and functions from libraries are structured and performed to retrieve data from a webpage (mainly tables) into the Jupyter notebook where the data can be filtered, structured and calculations can be made.

- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation



Data Wrangling

- The main techniques for processing data were:
  - Replacing missing values in columns with mean of the rest of the data from that column.
  - Create a new column or list with binary values: 0 for one specific case and 1 for the rest of the cases.
- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation

EDA with Data Visualization

- The following charts were used to see if there is a correlation between those variables and with that information, estimate outcomes:
  - Flight Number vs. Payload Mass in a scatter plot and overlay the outcome of the launch.
  - Flight Number vs. Launch Site in a scatter plot and overlay the outcome of the launch.
  - Payload Mass vs. Launch Site in a scatter plot and overlay the outcome of the launch.
  - Flight Number vs. Orbit in a scatter plot and overlay the outcome of the launch.
  - Payload Mass vs. Orbit in a scatter plot and overlay the outcome of the launch.
  - Success rate over the years in a line plot.
- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation

EDA with SQL

- Below is the summary of the SQL queries performed:
  - select distinct "Launch_Site" from SPACEXTBL
  - SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 20
  - SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE Customer="NASA (CRS)“
  - SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" LIKE 'F9 v1.1%’
  - SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome"="Success (ground pad)“
  - SELECT distinct "Booster_Version" FROM SPACEXTBL WHERE "Landing _Outcome"="Success (drone ship)" AND "PAYLOAD_MASS__KG_">4000 AND "PAYLOAD_MASS__KG_"<6000
  - SELECT "Mission_Outcome", count(*) as total_outcomes FROM SPACEXTBL GROUP BY "Mission_Outcome“
  - SELECT DISTINCT"Booster_Version" FROM SPACEXTBL where (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)

EDA with SQL

- SELECT

  CASE

  WHEN substr(Date, 4, 2) = '01' THEN 'January'

  WHEN substr(Date, 4, 2) = '02' THEN 'February'

  WHEN substr(Date, 4, 2) = '03' THEN 'March'

  WHEN substr(Date, 4, 2) = '04' THEN 'April'

  WHEN substr(Date, 4, 2) = '05' THEN 'May'

  WHEN substr(Date, 4, 2) = '06' THEN 'June'

  WHEN substr(Date, 4, 2) = '07' THEN 'July'

  WHEN substr(Date, 4, 2) = '08' THEN 'August'

  WHEN substr(Date, 4, 2) = '09' THEN 'September'

  WHEN substr(Date, 4, 2) = '10' THEN 'October'

  WHEN substr(Date, 4, 2) = '11' THEN 'November'

  WHEN substr(Date, 4, 2) = '12' THEN 'December'

  END as month_name,

EDA with SQL

"Landing _Outcome",

"Booster_Version",

"Launch_Site"

FROM SPACEXTBL

WHERE substr(Date,7,4)='2015' AND "Landing _Outcome" LIKE '%Failure%' AND "Landing _Outcome" LIKE '%(drone ship)%’

- SELECT * COUNT(*) as successful_landings

  FROM SPACEXTBL

  WHERE "Landing _Outcome" LIKE '%Success%’ AND "Date" BETWEEN '2010-06-04' AND '2017-03-20’

  GROUP BY "Landing _Outcome“

  ORDER BY successful_landings DESC

- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation

Build an Interactive Map with Folium

- Below are the map objects such as markers, circles, lines, etc. created and added to the folium map:
    - Circle.
    - Marker.
    - Marker cluster.
    - Mouse position.
    - Polyline.
- These objects were added in order to get insights from the map from the launches such as on a mouse position get the coordinates quickly, having straightforward information over the map like success launch, and having distance to main railroads, coastlines, etc.
- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation

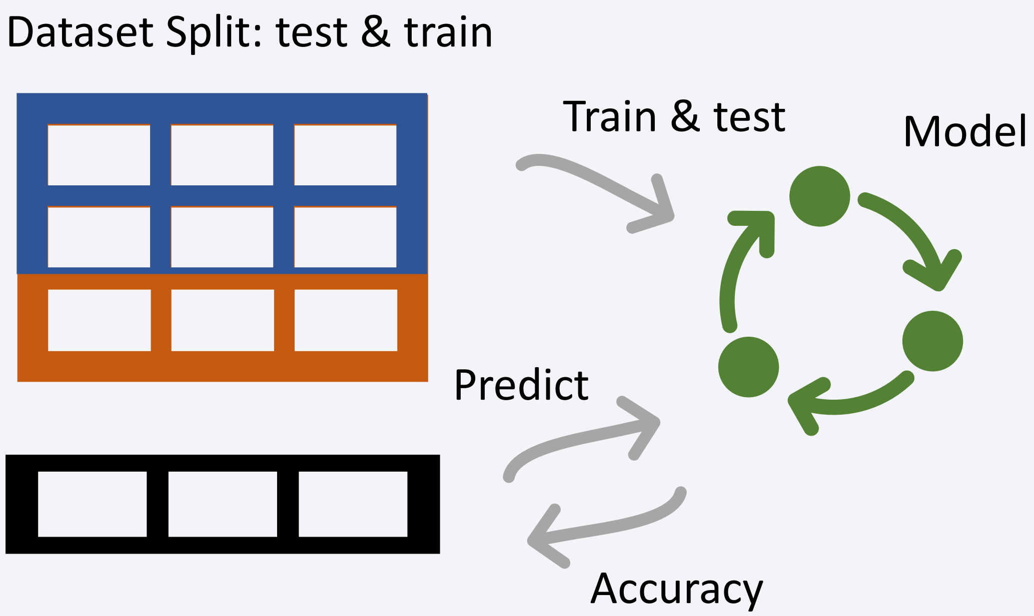
Build a Dashboard with Plotly Dash

- Below is the summary of the plots/graphs and interactions added to a dashboard:
  - Launch site drop-down input component.
  - Pie chart with a callback function based on selected site from drop-down input component.
  - Range slider to select payload mass (kg).
  - Scatter plot success vs. payload.
- These graphs and interactions were added in order to enable stakeholders to explore and manipulate data in an interactive and real-time way. In this way, instead of presenting findings in static graphs, interactive data visualization can always tell a more appealing story.
- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation

Predictive Analysis (Classification)

- Below is the summary building, evaluating, improvement and finding the best performing classification model:
  - Creation of Numpy array from selected columns for performing model testing and evaluation.
  - Standardize data for the range to be consistent.
  - Split dataset into test and train subsets.
  - Create objects for each model to be fitted and getting the accuracy for each one.



- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation

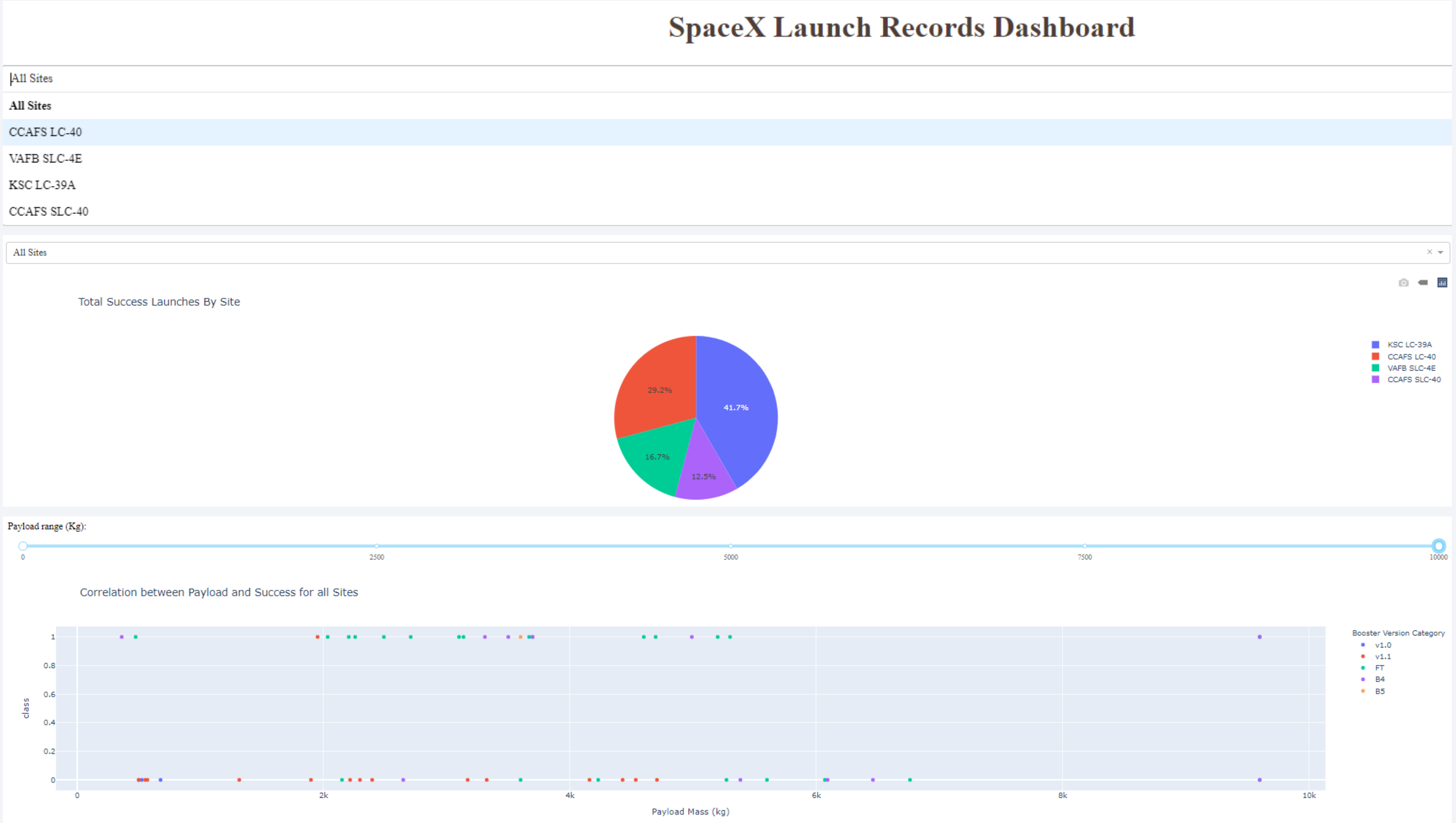
Results

- By analizing the information in the exploratory data analysis, we se the following results:
  - As the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
  - For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000) and generally speaking for this heavy payload mass there is a high success rate.
  - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
  - The success rate since 2013 kept increasing till 2020.
  - There was a very high success rate for all missions. See table →

Mission_Outcome	total_outcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Results

- Below are some screenshots from the interactive analytics:



Results

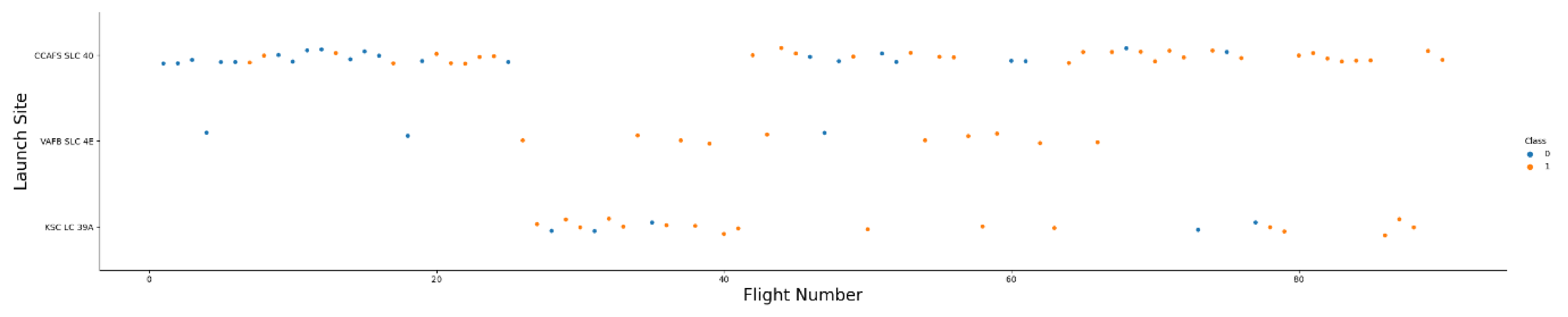
- Below are the predictive analysis results:
  - Overall all methods testes for making a predictive analysis had a relatively high accuracy (>83%), except for the decision tree methos which perfomed with a 72% of accuracy.

Section 2

Insights drawn from EDA

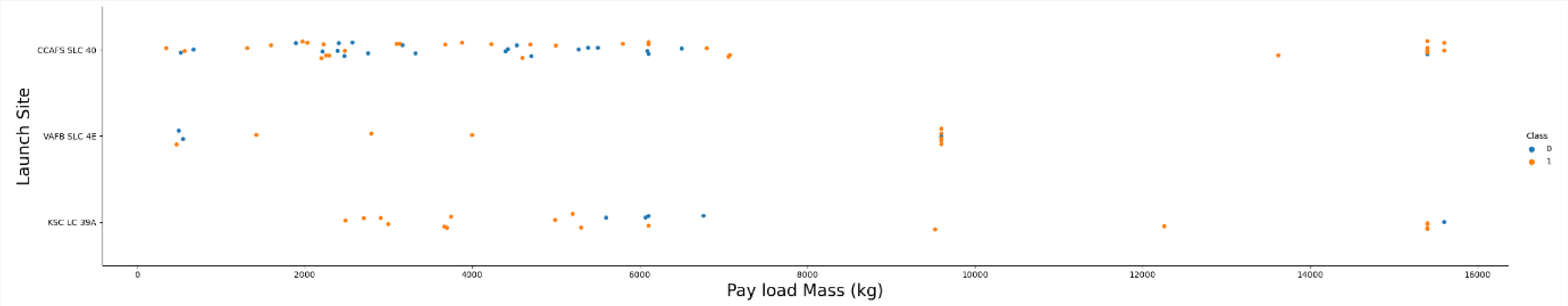
Flight Number vs. Launch Site

- As it can be seen from the plot, the highest succes rate was from the launch site VAFB SLC 4E, followed by the KSC LC 39A and last CCAFS SLC 40. (Class 0 = failure, class 1 = success).



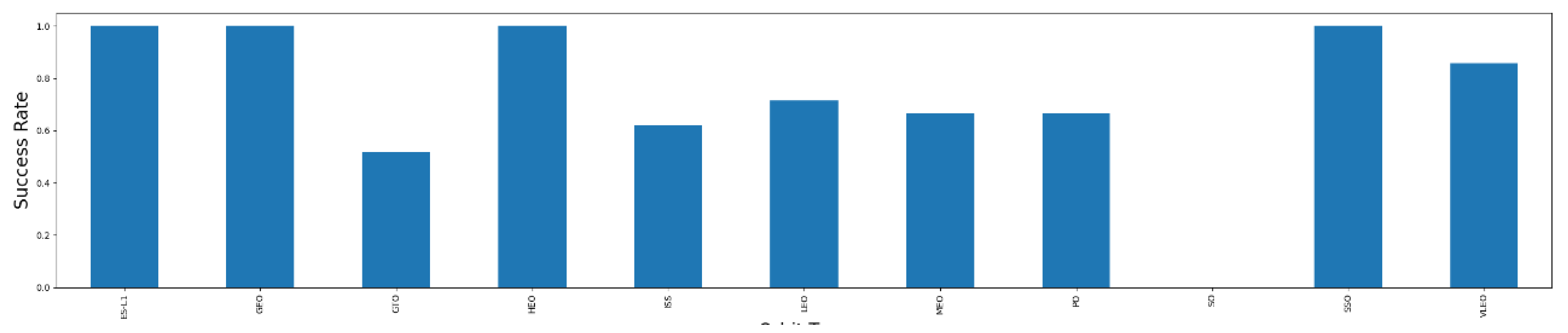
Payload vs. Launch Site

- As it can be seen from the plot, generally speaking for big payload mass the success rate was bigger. (Class 0 = failure, class 1 = success).



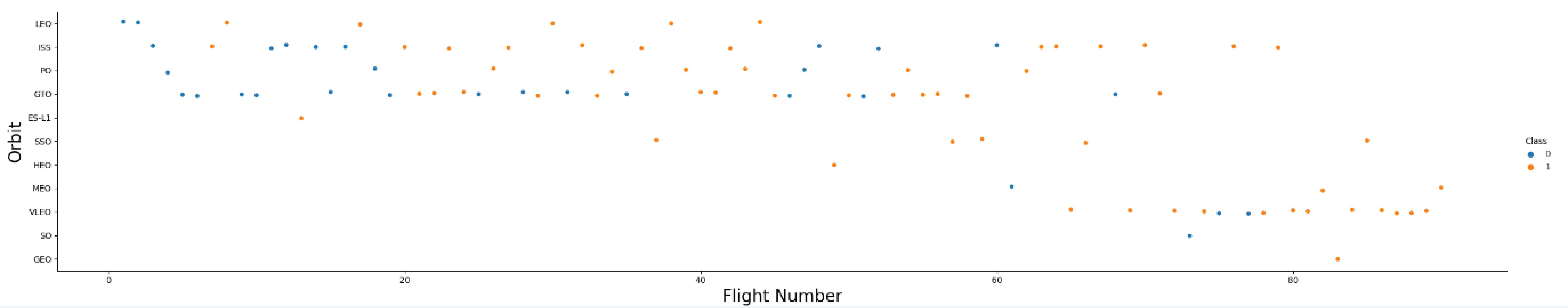
Success Rate vs. Orbit Type

- As it can be seen from the plot, the orbits ES-L1, GFO, HEO. SSO and VLEO had the best success rate.



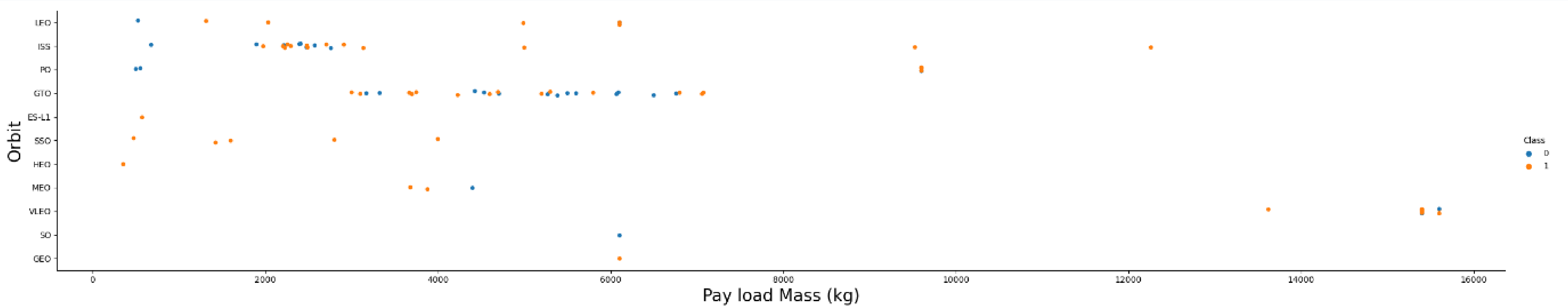
Flight Number vs. Orbit Type

- As it can be seen from the plot, the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



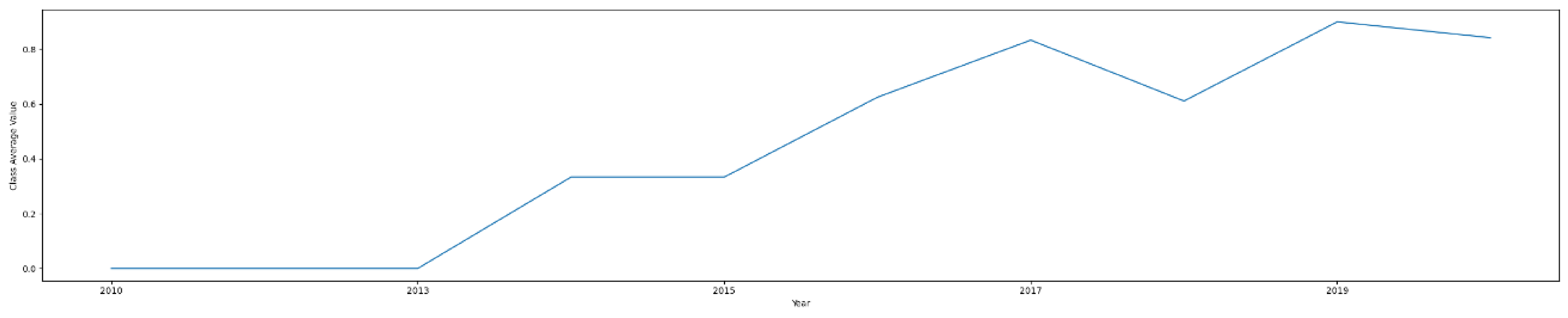
Payload vs. Orbit Type

- As it can be seen from the plot, with heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- As it can be seen from the plot, the success rate since 2013 kept increasing till 2020



All Launch Site Names

- The unique launch site in the space missions were retrieved with a SQL query to get these unique values, where the column Launch_Site is selected and a function named DISTINCT is applied to that column to return only unique values.
- Query: %sql select distinct "Launch_Site" from SPACEXTBL

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

Launch Site Names Begin with 'CCA'

- This query is used to retrieve all columns from SQL database and filter the column Launch_Site where the names begin with CCA by applying LIKE ‘CCA%’.
- Query: %sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The query uses SUM function to sum all values from the column PAYLOAD_MASS__KG_ where customer = NASA (CRS).
- Query: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE Customer="NASA (CRS)"

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- The query uses AVG function to calculate de average of all values from the column PAYLOAD_MASS__KG_ where Booster_Version has the characters F9 v.1.1.
- Query: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" LIKE 'F9 v1.1%'

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- The query uses MIN function to calculate de minimum vale from the column Date (first date) where Outcome = Success (ground pad).
- Query: %sql SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome"="Success (ground pad)"

MIN(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query uses DISTINCT function retrieve unique values from Booster_Version column, with Landing_Outcome=Success and Payload between 4000 and 6000.
- Query: %sql SELECT distinct "Booster_Version" FROM SPACEXTBL WHERE "Landing _Outcome"="Success (drone ship)" AND "PAYLOAD_MASS__KG_">4000 AND "PAYLOAD_MASS__KG_"<6000

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

Total Number of Successful and Failure Mission Outcomes

- The query selects the Mission outcome column and adds a column where it fills with the count of each case of mission outcome.
- Query: %sql SELECT "Mission_Outcome", count(*) as total_outcomes FROM SPACEXTBL GROUP BY "Mission_Outcome"

Mission_Outcome	total_outcomes
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The query selects unique values from Booster Version column and then uses a sub-query to select the max Payload Mass. (The table is very long, so a portion of it was included).
- Query: %sql SELECT DISTINCT"Booster_Version" FROM SPACEXTBL where (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)

Booster_Version
F9 v1.0 B0003
F9 v1.0 B0004
F9 v1.0 B0005
F9 v1.0 B0006
F9 v1.0 B0007
F9 v1.1 B1003
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1010
F9 v1.1 B1012
F9 v1.1 B1013
F9 v1.1 B1014

2015 Launch Records

- The query uses a CASE function to translate number-type month date into letter-type month date, and filter: 2015 year failure launches, their bosster version and launch sites.
- Query:

```sql
%%sql
SELECT
  CASE
    WHEN substr(Date, 4, 2) = '01' THEN 'January'
    WHEN substr(Date, 4, 2) = '02' THEN 'February'
    WHEN substr(Date, 4, 2) = '03' THEN 'March'
    WHEN substr(Date, 4, 2) = '04' THEN 'April'
    WHEN substr(Date, 4, 2) = '05' THEN 'May'
    WHEN substr(Date, 4, 2) = '06' THEN 'June'
    WHEN substr(Date, 4, 2) = '07' THEN 'July'
    WHEN substr(Date, 4, 2) = '08' THEN 'August'
    WHEN substr(Date, 4, 2) = '09' THEN 'September'
    WHEN substr(Date, 4, 2) = '10' THEN 'October'
    WHEN substr(Date, 4, 2) = '11' THEN 'November'
    WHEN substr(Date, 4, 2) = '12' THEN 'December'
  END as month_name,
  "Landing _Outcome",
  "Booster_Version",
  "Launch_Site"
FROM SPACEXTBL
WHERE substr(Date,7,4)='2015' AND "Landing _Outcome" LIKE '%Failure%' AND "Landing _Outcome" LIKE '%(drone ship)%'
```

month_name	Landing _Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The query selects the Landing_Outcome column and counts the cases for the specified date range, groups it by Landing_Otume case and orders it by what was counted.
- Query:%%sql

  SELECT "Landing _Outcome", COUNT(*) as 'Quantity' FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY "Landing _Outcome" ORDER BY "Quantity"

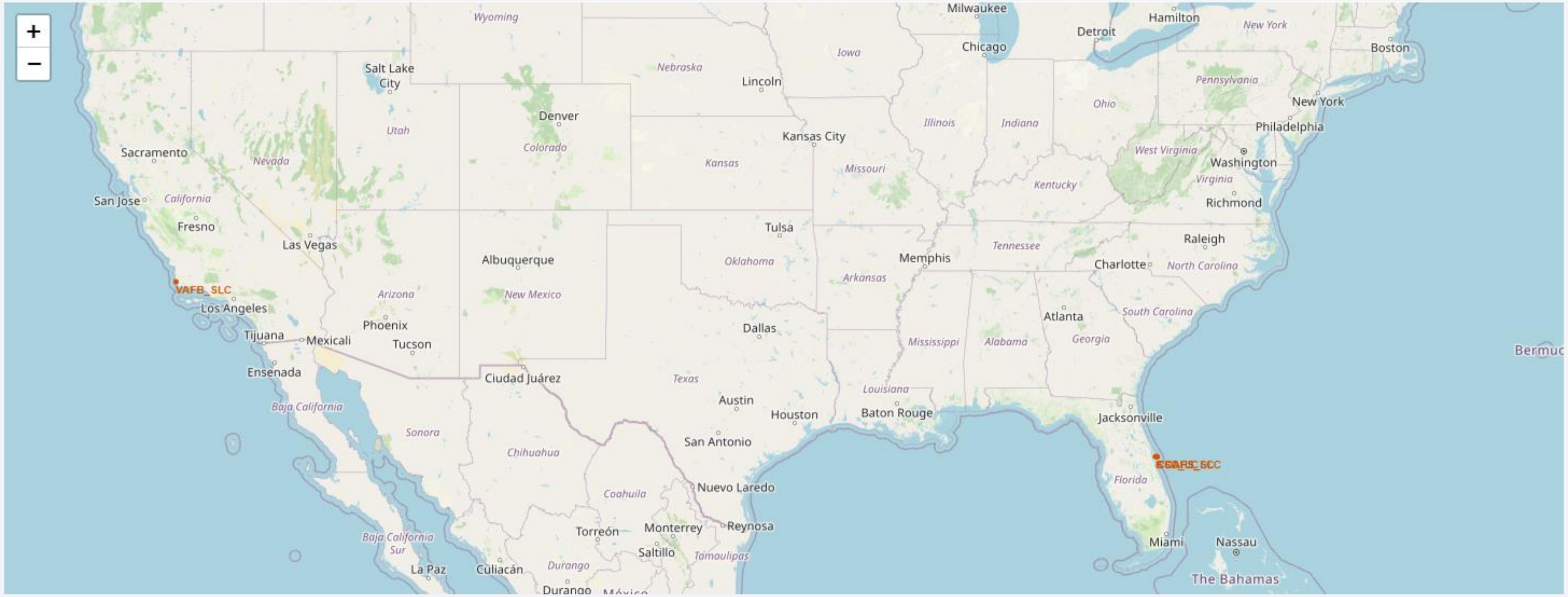
Landing _Outcome	Quantity
No attempt	1
Failure (parachute)	2
Controlled (ocean)	3
Failure	3
Failure (drone ship)	4
Success (ground pad)	6
Success (drone ship)	8
No attempt	10
Success	20

Section 3

Launch Sites Proximities Analysis

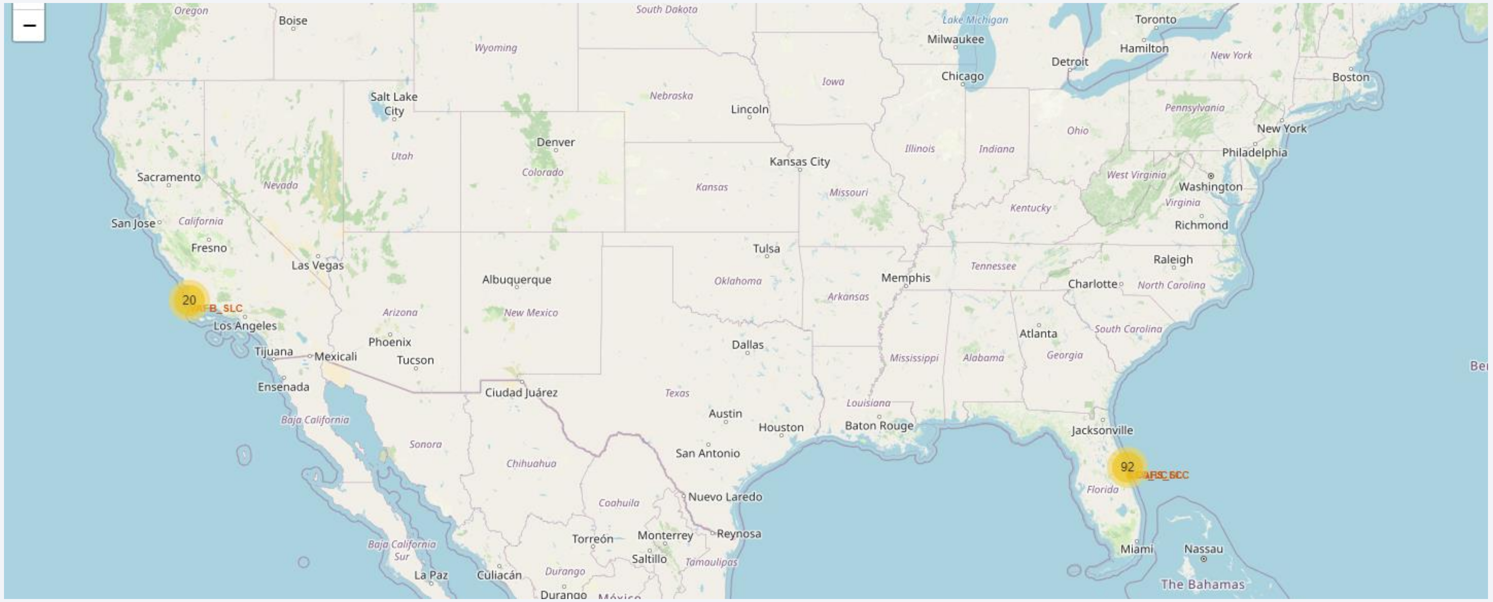
Map with markers and circles.

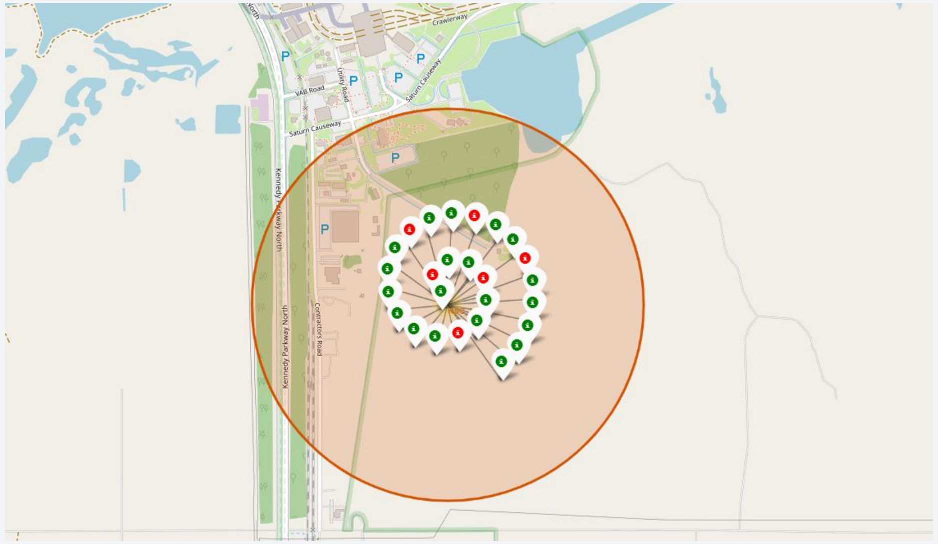
- In the following screenshot of the map, a first layer of circles and markers were added to it. This is important in order to detect how close each launch site is from each other.



Map with markers, circles with success outcome.

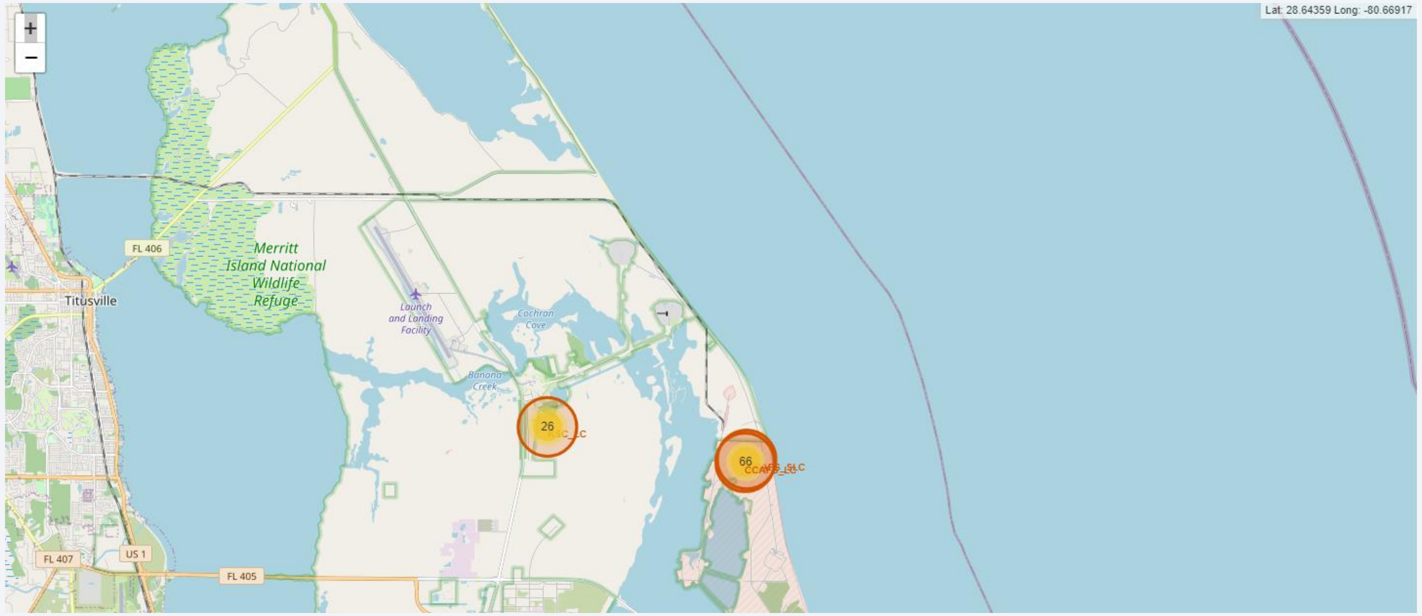
- In the following screenshot of the map, an additional layer was added to display the quantity of successful launch outcomes. This is important in order to detect areas were successful outcomes were positive.





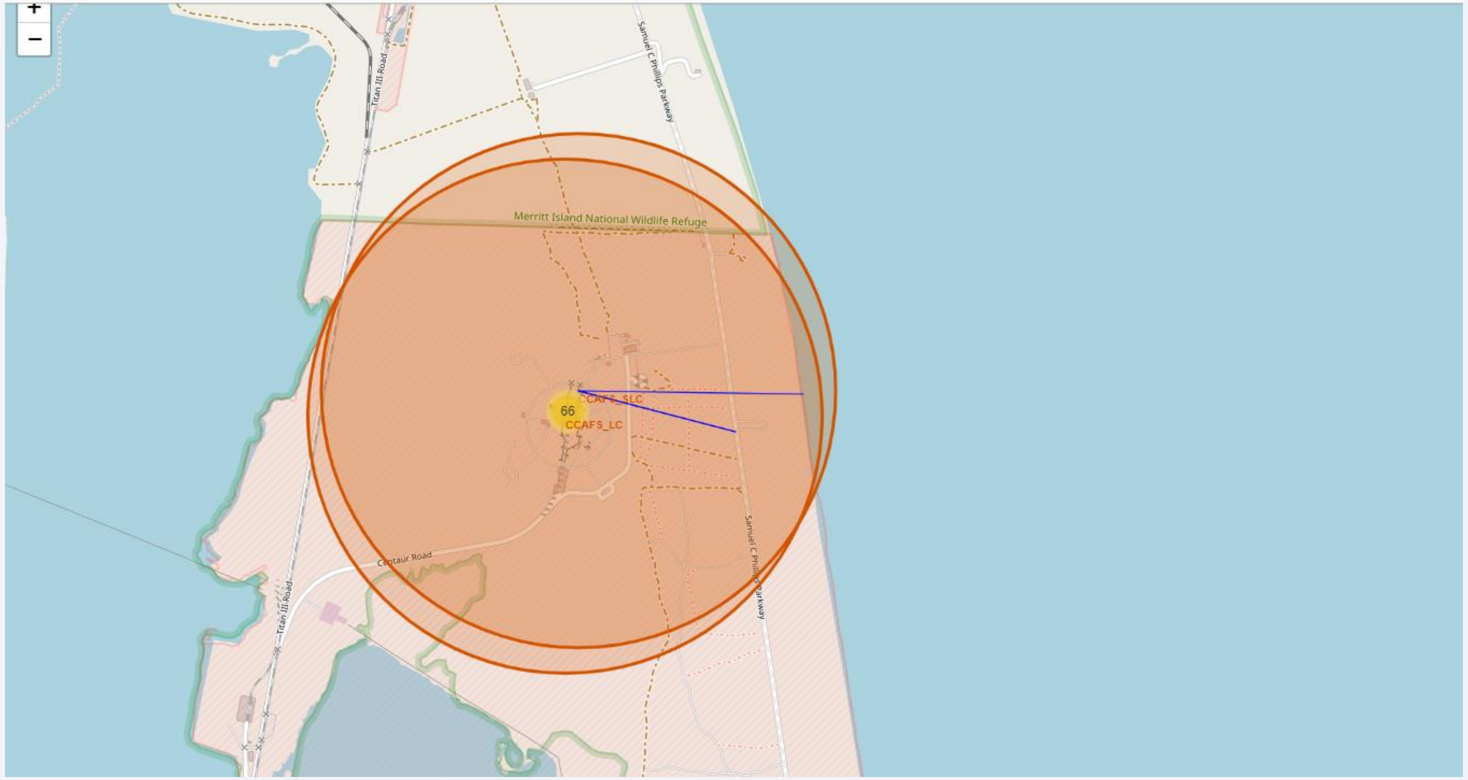
Map with markers, circles with success outcome and coordinates.

- This map is the same as the one in the previous slide but it has a functionality that it indicates the coordinates as you move the mouse over the map for giving the instantaneous coordinate ion a specific point if the map.



Map with distance lines

- The last map has all the functionalities from previous maps but lines with distances are drawn over it that show how long a line is from one point to another to gain insightful information whether if a launch site is close to a road, railroad, shore line, etc.

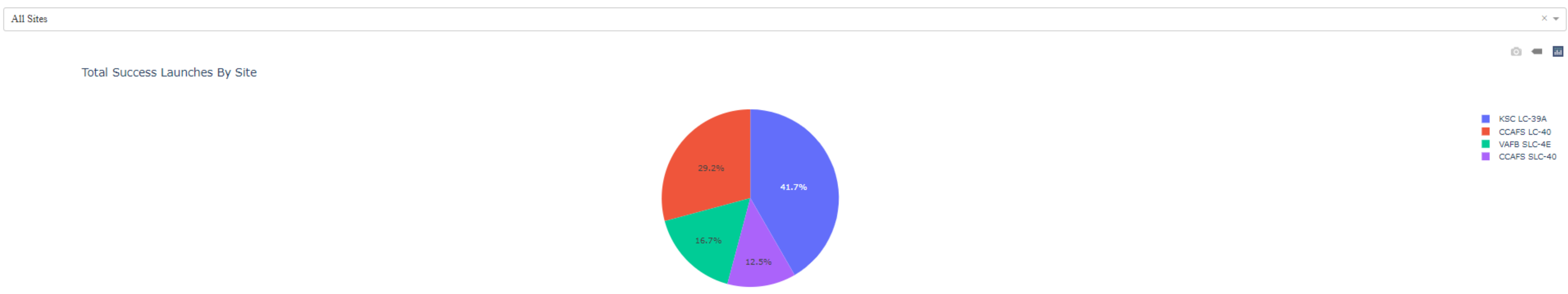


Section 4

Build a Dashboard with Plotly Dash

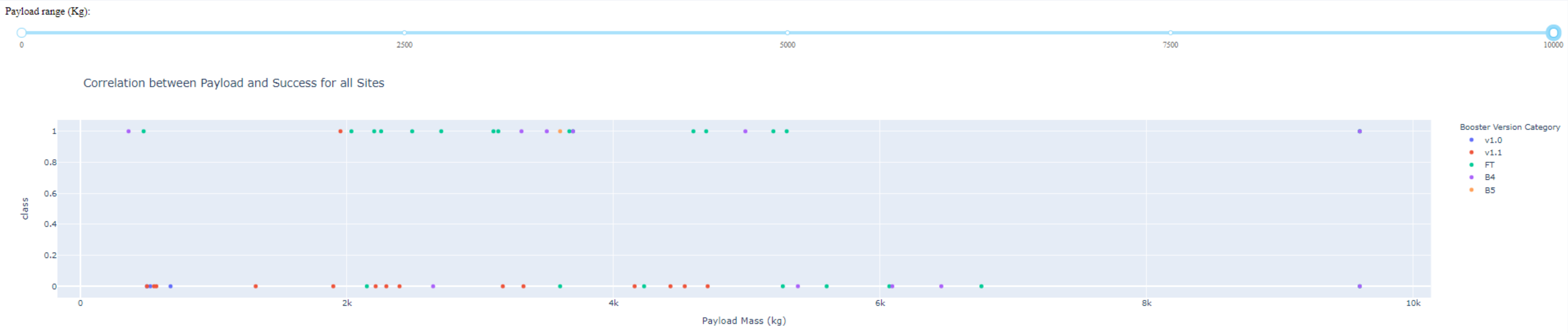
Success Launch for all sites

- This pie chart clearly shows which site has the highest and lowest success rates, indicating their percentages.



Payload vs. Launch Outcome scatter plot

- This scatter plot shows the correlation between Payload mass and launch outcome. The range slider is very useful because it can filter the payload mass range for a desired one and get instant insightful information.

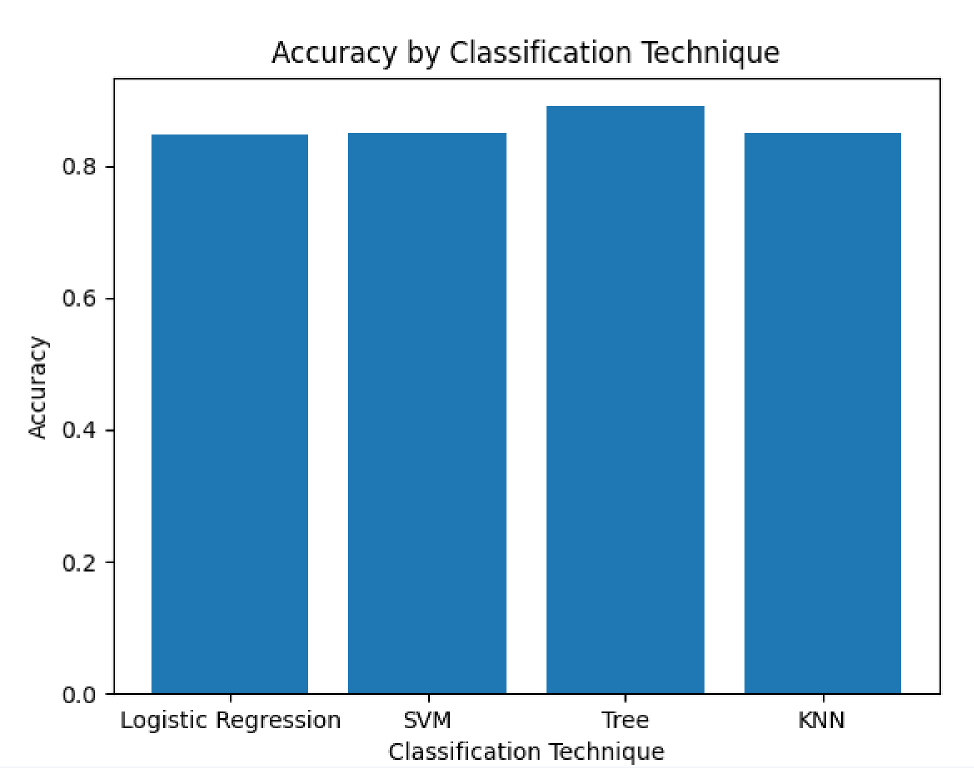


Section 5

Predictive Analysis (Classification)

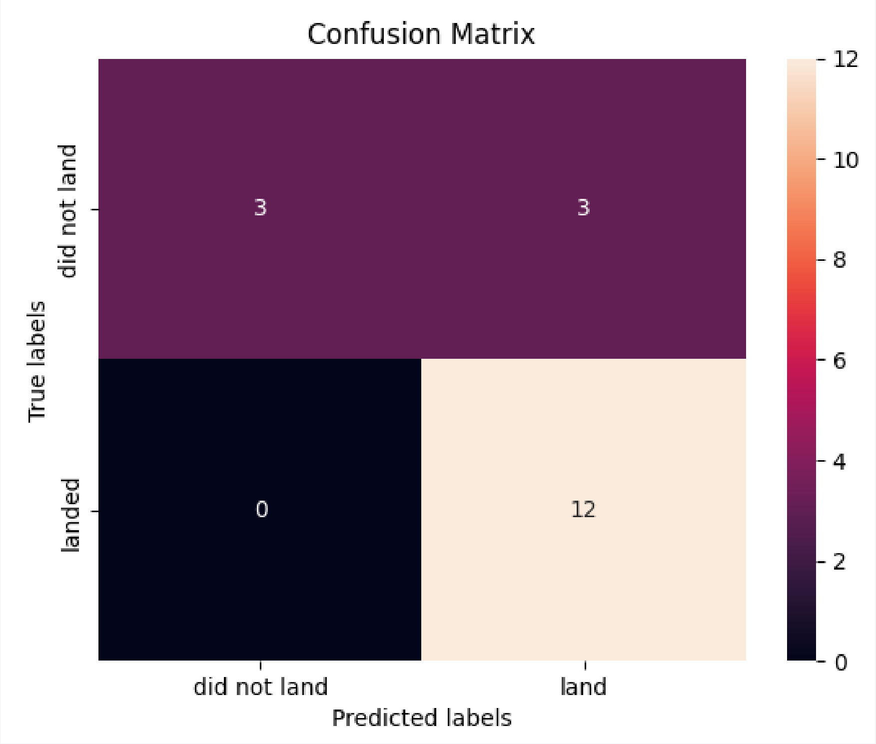
Classification Accuracy

- By comparing the accuracy of each model, we found out that the tree classification technique is the most suitable one, even though the others are not far from it.



Confusion Matrix

- The confusion matrix for the decision tree technique is shown and it reveals that for the actual landed occurrences, the model predicted all the cases and for the not landed actual ocurrences the model predicted half of them.



Conclusions

- As the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000) and generally speaking for this heavy payload mass there is a high success rate.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
- The success rate since 2013 kept increasing till 2020.
- There was a very high success rate for all missions.
- The decision tree technique was the best technique for the model.

Appendix

- Below is the link to the several steps for this study:

https://github.com/JulianCarroVerdia/IBM_Data_Science_Capstone_Presentation

Thank you!