IBM Data Science Capstone Project

Shril Patel

02/01/2025

OUTLINE

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

Executive Summary:

The goal of this research is to analyze SpaceX Falcon 9 data collected through various sources and employ Machine Learning models to predict the success of the first stage landing that provides other space agencies the ability to decide if they did against SpaceX.

- Summary of Methodologies:
  - Following concepts and methods were used to collect and analyze data, and evaluate machine learning models, and make predictions:
    - Collect data through API and Web Scraping
    - Data Wrangling
    - Conduct exploratory data analysis with SQL and data visuals
    - Build interactive map with Folium to analyze launch site proximity
    - Build a dashboard to analyze launch records interactively
    - Build predictive model to predict if the first stage of Falcon 9 will land successfully
- Summary of all Result:
  - Exploratory Data Analysis results
  - Interactive analytics
  - Predictive Analysis Results

Introduction:

- Project background and context:
  - SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Questions to be Answered:
  - Determine if the first stage of SpaceX Falcon 9 will land successfully
  - Impact of different parameters/variables on the landing outcomes
  - Correlations between launch and success rate

Methodology

Methodology:

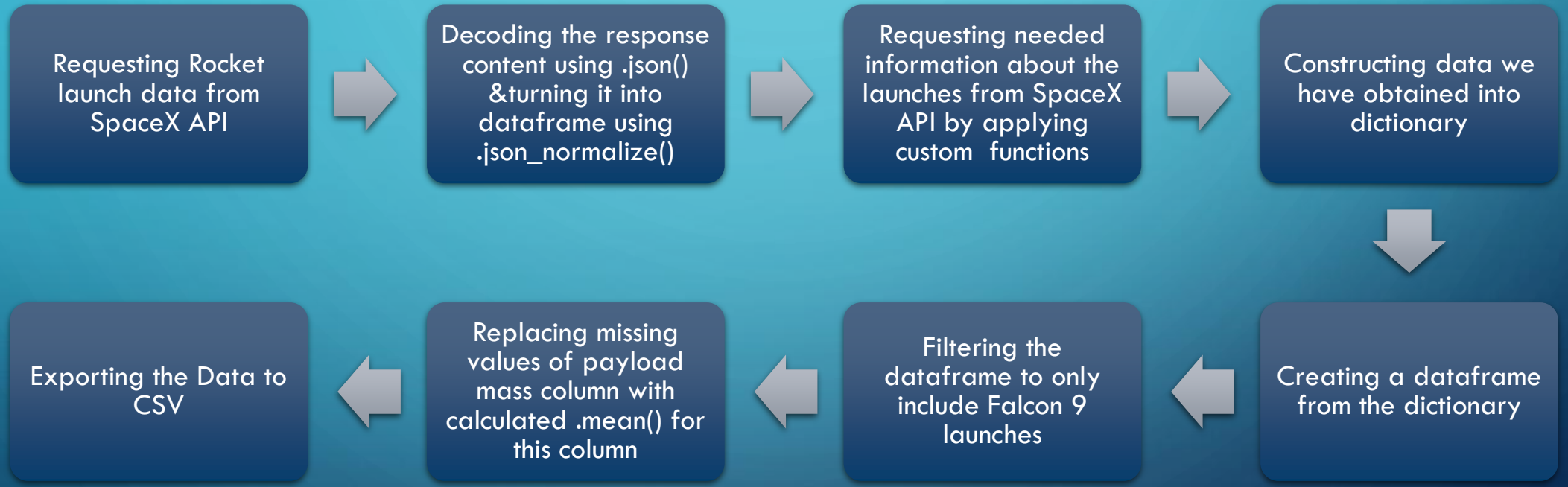
- **Data collection methodology**:
  - Using SpaceX API
  - Using Web Scrapping from Wikipedia
- **Performed data wrangling**:
  - Filtering the Data
  - Dealing with missing Values
  - Using one hot encoding to prepare the data to binary classification
- **Performed exploratory data analysis (EDA) using visualization and SQL**:
- **Preformed interactive visual analytics**:
  - Using Folium and Plotly Dash Visualization
- Performed predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best result

DATA COLLECTION

Data collection involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

- Data Columns are obtained by using SpaceX REST API:
  - Flight No, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
  - Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data collection-SpaceX API



Data collection-Web Scraping

Requesting Falcon 9 launch data from Wikipedia

Creating a beatifulsoup object from HTML response

Extracting all column names from the HTML table header

Collecting the data by parsing HTML table

Constructing data we have obtained into a dictionary

Creating a Dataframe from the dictionary

Exporting the data to CSV

DATA WRANGLING

- Conducted Exploratory Data Analysis (EDA) to find patters in data and define labels for training supervised models
- The data set contained various mission outcomes that were converted into Training Labels with 1 meaning the booster successfully landed and 0 meaning booster was unsuccessful in landing. Following landing scenarios were considered to create labels:
  - True Ocean means the mission outcome was successfully landed to a specific region of the ocean
  - False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean
  - 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

EDA with Data Visualization

Charts were plotted:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend.

1. **Scatter plots** show the relationship between variables. If a relationship exists, they could be used in machine learning model.
2. **Bar charts** show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
3. **Line charts** show trends in data over time (time series)

EDA with SQL

SQL is an indispensable tool for Data Scientist and Analysts as most of the real-world data is stored in databases. It's not only the standard language for Relational Database operations, but also a incredibly tool for analyzing data and drawing useful insights from it.

- We performed SQL queries to gather information from give dataset:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass. Use a subquery
9. List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
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

Build an interactive map with Folium

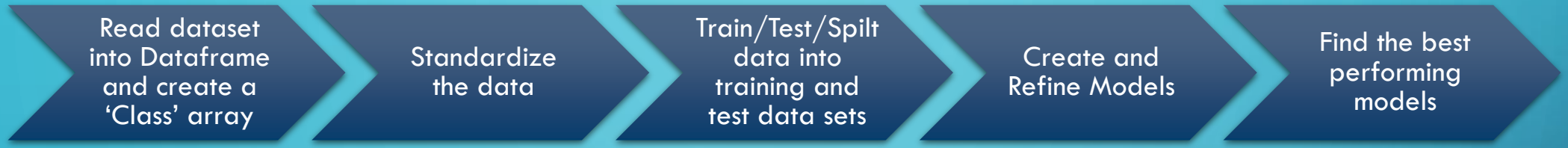
Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map.

Map Objects	Code	Result
Map Marker	folium.marker(	Map object to make a mark on map
Icon Marker	folium.icon(	Create am icon on map
Circle Marker	folium.circle(	Create a circle where marker is being placed
PolyLine	folium.polyline(	Create a line between points
Marker Cluster Object	Markercluster()	This is good way to simplify map containing markers having the same coordinate
AntPath	Folium.plugins.antpath(	Create an animated line between points

Build a Dashboard with Plotly Dash

Map Objects	Code	Result
Dash and its Components	Import dash Import dash_html_components as html Import dash_core_components as dcc From dash.dependencies import Input, Output	The dash core component library contains a set of higher–level components like sliders, graphs, dropdown, tables and more.
Pandas	Import pandas as pd	Fetching values from CSV and creating a dataframe
Plotly	Import plotly.express as px	Plot the graphs with interactive plotly library
Dropdown	dcc.Dropdown(	Create a dropdown for launch sites
Rangeslider	dcc.RangeSlider(	Create a rangleslider for payload mass range selection
Pie Chart	Px.pie(	Creating the pie graphs for success percentage display
Scatter Chart	Px.scatter(	Creating the scatter graph for correlation display

Predictive Analysis (Classification)

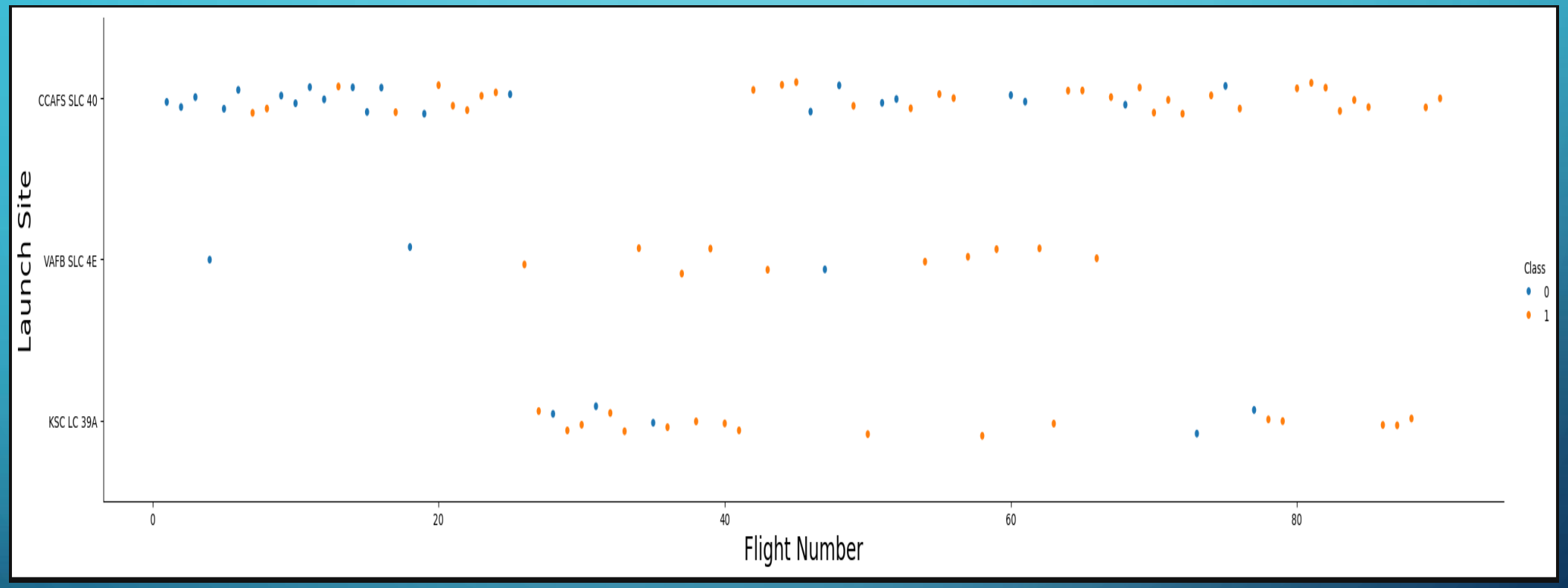


Results

- Exploratory Data Analysis Results
- Interactive Analytics demo in Screenshot
- Predictive Analysis Results

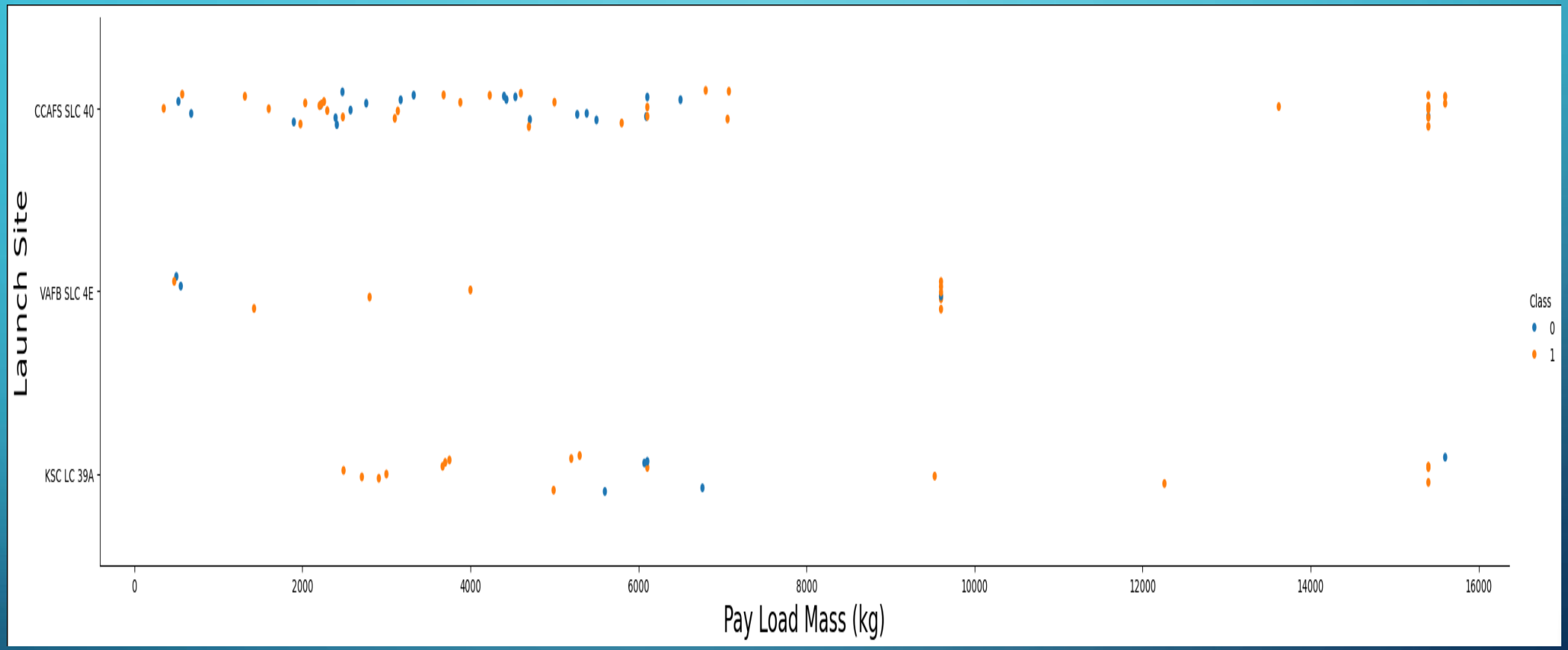
INSIGHTS DRAWN FROM EDA

FLIGHT NUMBER VS. LAUNCH SITE



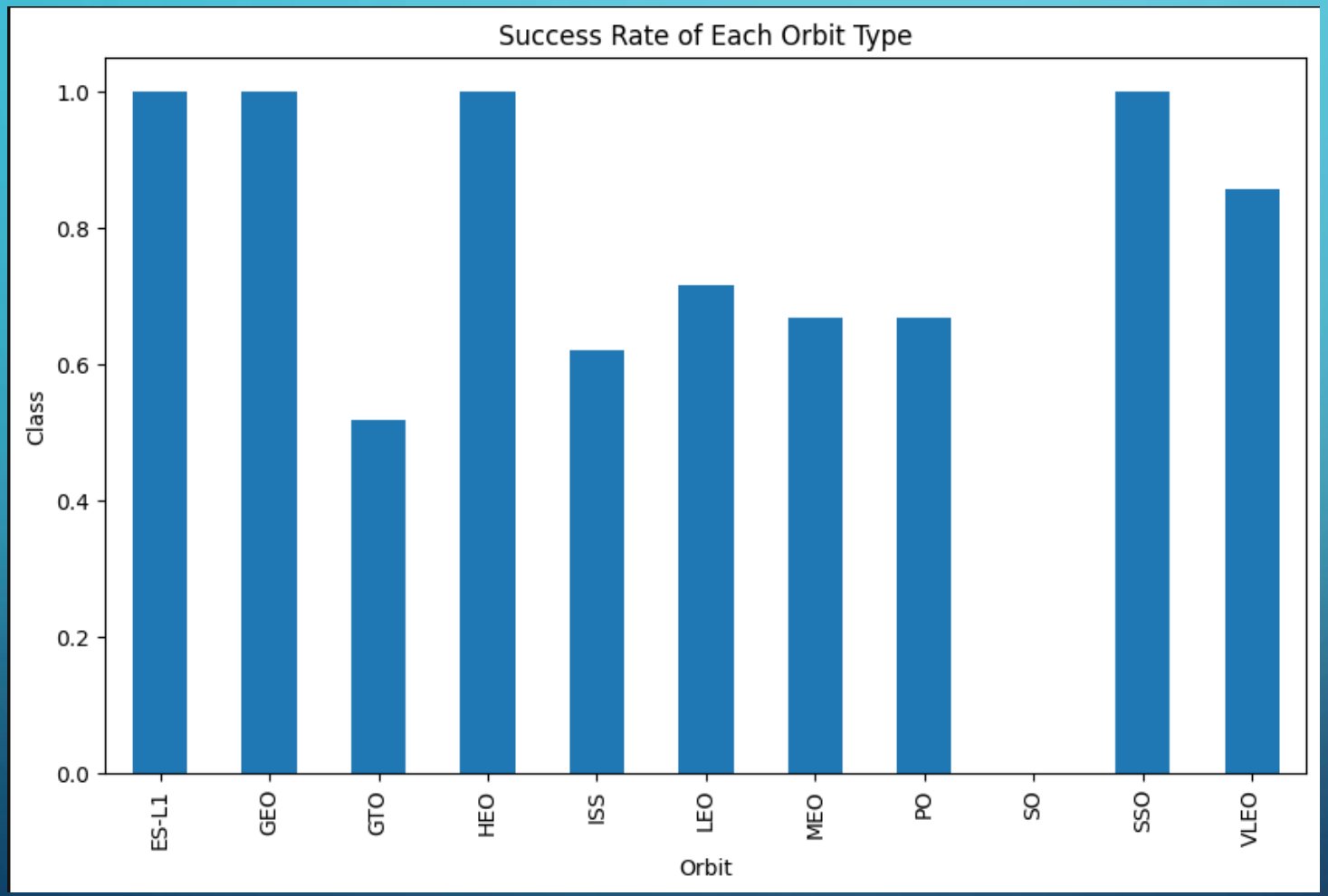
- Success rates (Class=1) increases as the number of flights increase
- For launch site 'KSC LC 39A', it takes at least around 25 launches before a first successful launch

PAYLOAD VS. LAUNCH SITE



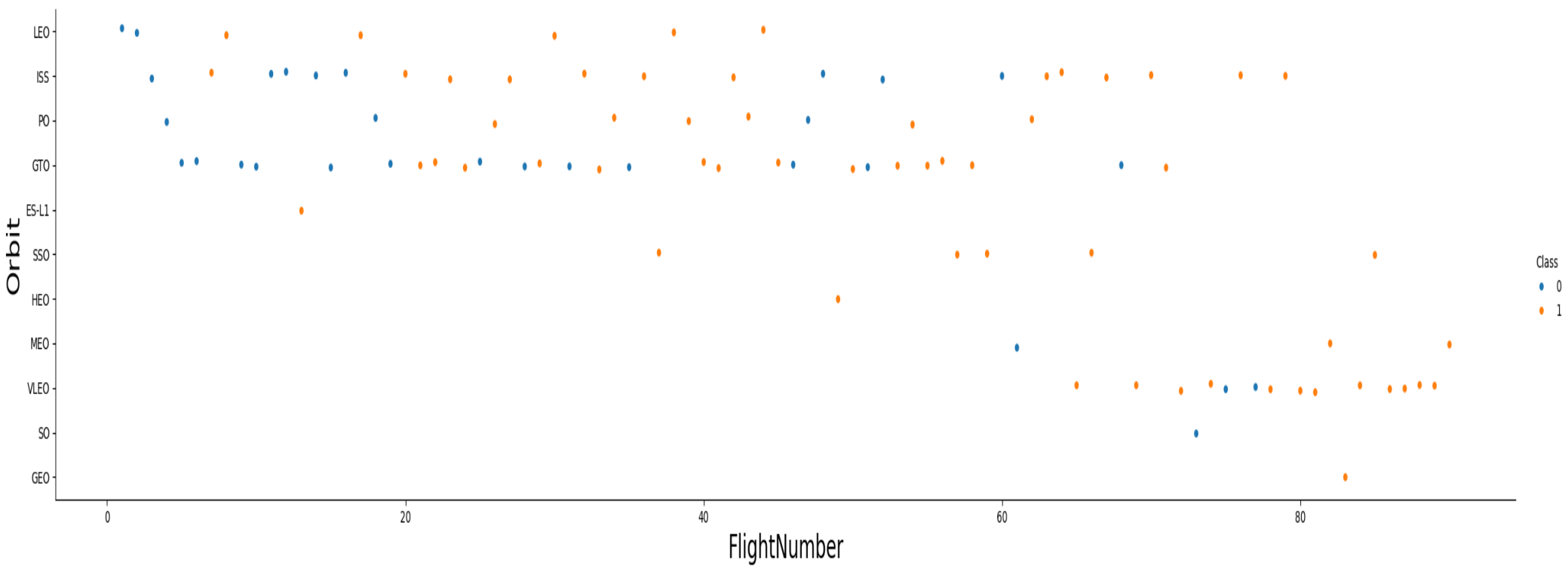
- Percentage of successful launch (Class=1) increases for launch site 'VAFB SLC 4E' as the payload mass increases
- There is no clear correlation or pattern between launch site and payload mass

SUCCESS RATE VS. ORBIT TYPE



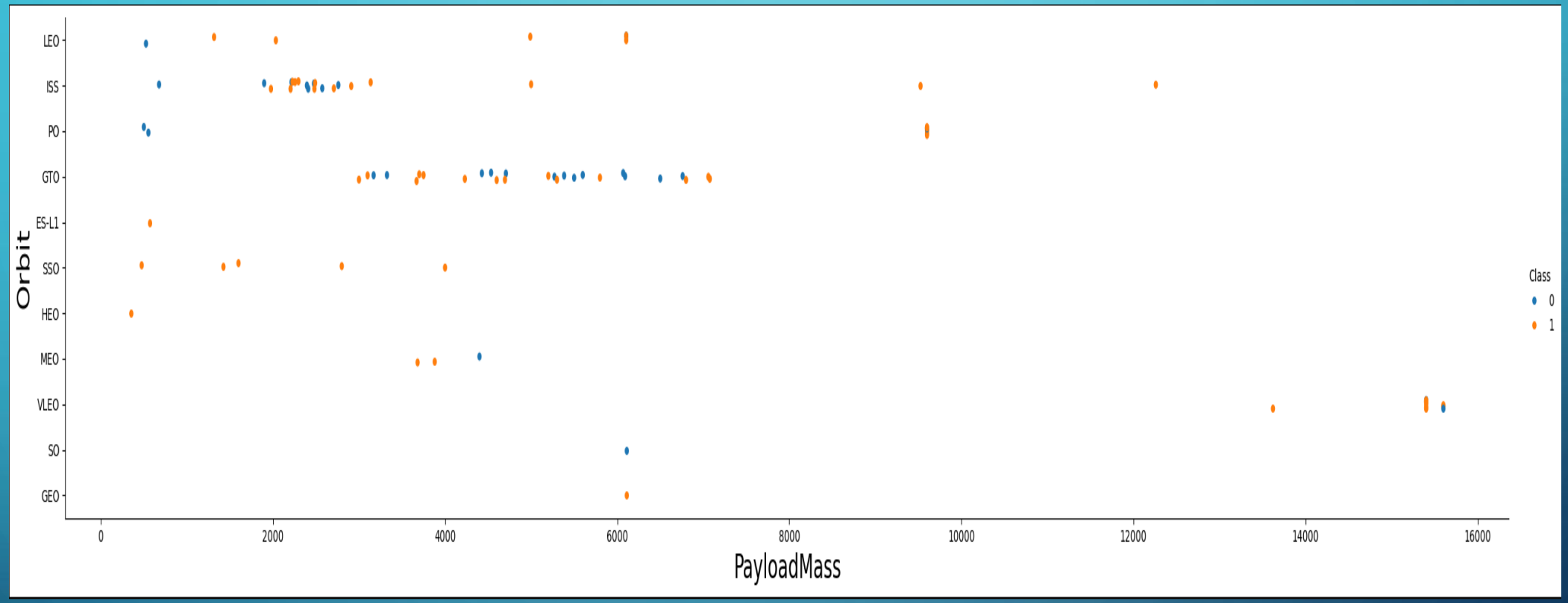
- Orbits ES-LI, GEO, HEO, and SSO have the highest success rates
- GTO orbit has the lowest success rate

FLIGHT NUMBER VS. ORBIT TYPE



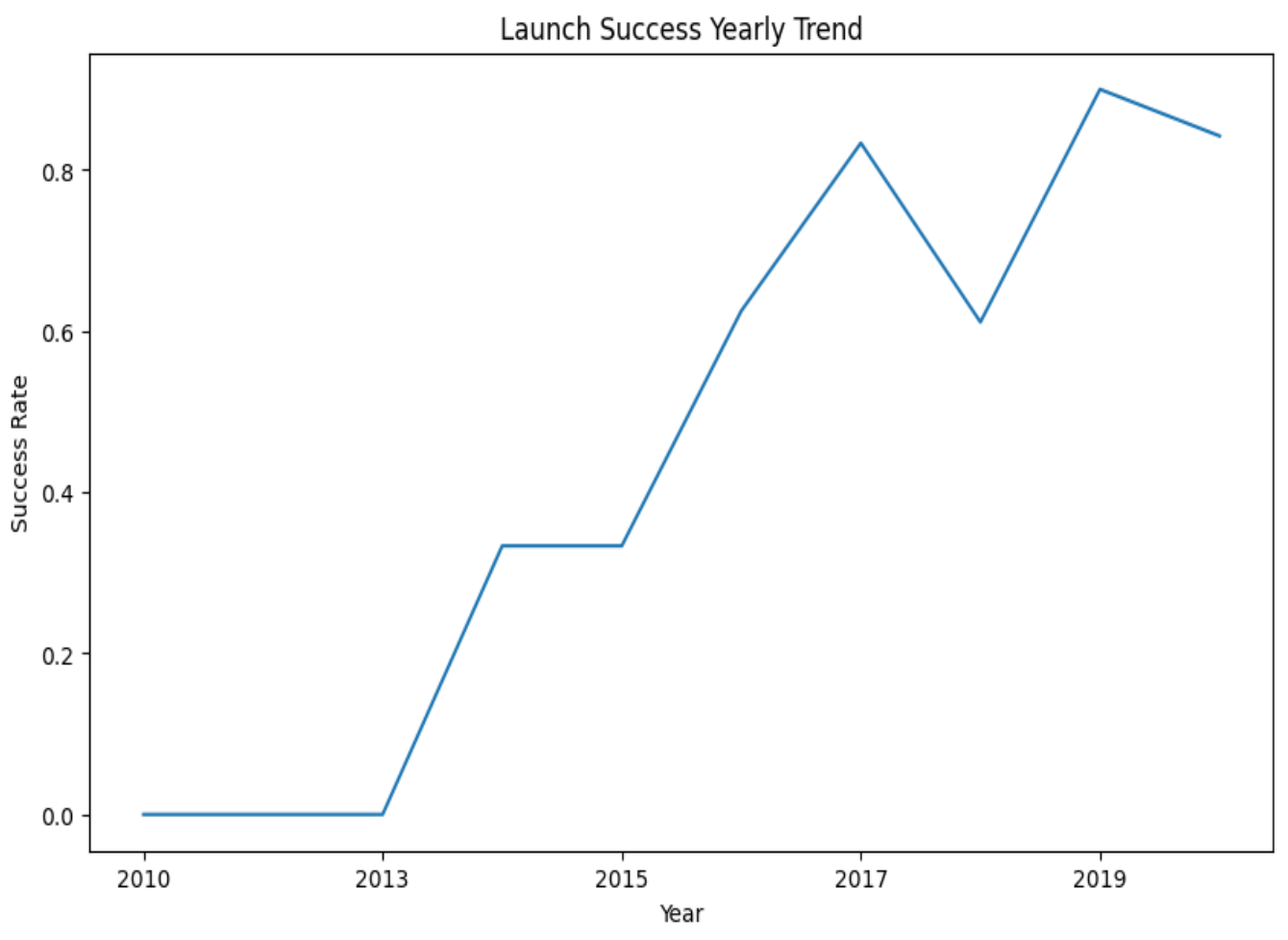
- For orbit VLEO, first successful landing (class=1) doesn't occur until 60+ number of flights
- For most orbits (LEO, ISS, PO, SSO, MEO, VLEO) successful landing rates appear to increase with flight numbers

PAYLOAD VS. ORBIT TYPE



- Successful landing rates (Class=1) appear to increase with pay load for orbits LEO, ISS, PO, and SSO
- For GEO orbit, there is not clear pattern between payload and orbit for successful or unsuccessful landing

LAUNCH SUCCESS YEARLY TREND



- Success rate (Class=1) increased by about 80% between 2013 and 2020
- Success rates remained the same between 2010 and 2013 and between 2014 and 2015

ALL LAUNCH SITE NAMES

- Query:

```
%%sql

select distinct Launch_Site from spacextbl
```

- Description:
  - ‘distinct’ returns only unique values from the queries column (Launch_Site)
  - There are 4 unique launch sites
- Result:

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

LAUNCH SITE NAMES BEGIN WITH ‘CCA’

- Query:

```
%%sql

select * from spacextbl where Launch_Site LIKE 'CCA%' limit 5;
```

- Description:
  - ‘Using keyword ‘Like’ and format ‘CCA%’, returns records where ‘Launch_Site’ column starts with “CCA”.
  - Limit 5, limits the number of returned records to 5Result:
- Result:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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:

```
%%sql

select sum(PAYLOAD_MASS__KG_) from spacextbl where Customer = 'NASA (CRS)'
```

- Description:
  - ‘sum’ adds column ‘PAYLOAD_MASS_KG’ and returns total payload mass for customers named ‘NASA (CRS)’
- Result:

sum(PAYLOAD_MASS__KG_)
45596

AVERAGE PAYLOAD MASS BY F9 V1.1

- Query:

```
%%sql

select avg(PAYLOAD_MASS__KG_) from spacextbl where Booster_Version LIKE 'F9 v1.1';
```

- Description:
  - ‘avg’ keyword returns the average of payload mass in ‘PAYLOAD_MASS_KG’ column where booster version is ‘F9 v1.1’
- Result:

avg(PAYLOAD_MASS__KG_)
2928.4

FIRST SUCCESSFUL GROUND LANDING DATE

- Query:

```
%%sql

select min(Date) as min_date from spacextbl where Landing__Outcome = 'Success (ground pad)';
```

- Description:
  - ‘avg’ keyword returns the average of payload mass in ‘PAYLOAD_MASS_KG’ column where booster version is ‘F9 v1.1’
- Result:

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

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

- Query:

```
%%sql

select Booster_Version from spacextbl where (PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_ < 6000)
and (Landing__Outcome = 'Success (drone ship)');
```

- Description:
  - The query finds the booster version where payload mass is greater than 4000 but less than 6000 and the landing outcome is success in drone ship
  - The ‘and’ operator in the where clause returns booster versions where both conditions in the where clause are true
- Result:

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:

```
%%sql

select Mission_Outcome, count(Mission_Outcome) as counts from spacextbl group by Mission_Outcome;
```

- Description:
  - The ‘group by’ keyword arranges identical data in a column in to group
  - In this case, number of mission outcomes by types of outcomes are grouped in column ‘counts’
- Result:

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

BOOSTERS CARRIED MAXIMUM PAYLOAD

- Query:

```
%%sql

select Booster_Version, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl);
```

- Description:
  - The sub query returns the maximum payload mass by using keywork ‘max’ on the pay load mass column
  - The main query returns booster versions and respective payload mass where payload mass is maximum with value of 15600

- Result:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 LAUNCH RECORDS

- Query:

```
%%sql

select Landing__Outcome, Booster_Version, Launch_Site from spacextbl where Landing__Outcome = 'Failure (drone ship)' and year(Date) = '2015'
```

- Description:
  - • The query lists landing outcome, booster version, and the launch site where landing outcome is failed in drone ship and the year is 2015
  - • The ‘and’ operator in the where clause returns booster versions where both conditions in the where clause are true
  - • The ‘year’ keywork extracts the year from column ‘Date
  - • The results identify launch site as ‘CCAFS LC-40’ and booster version as F9 v1.1 B1012 and B1015 that had failed landing outcomes in drop ship in the year 2015
- Result:

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:

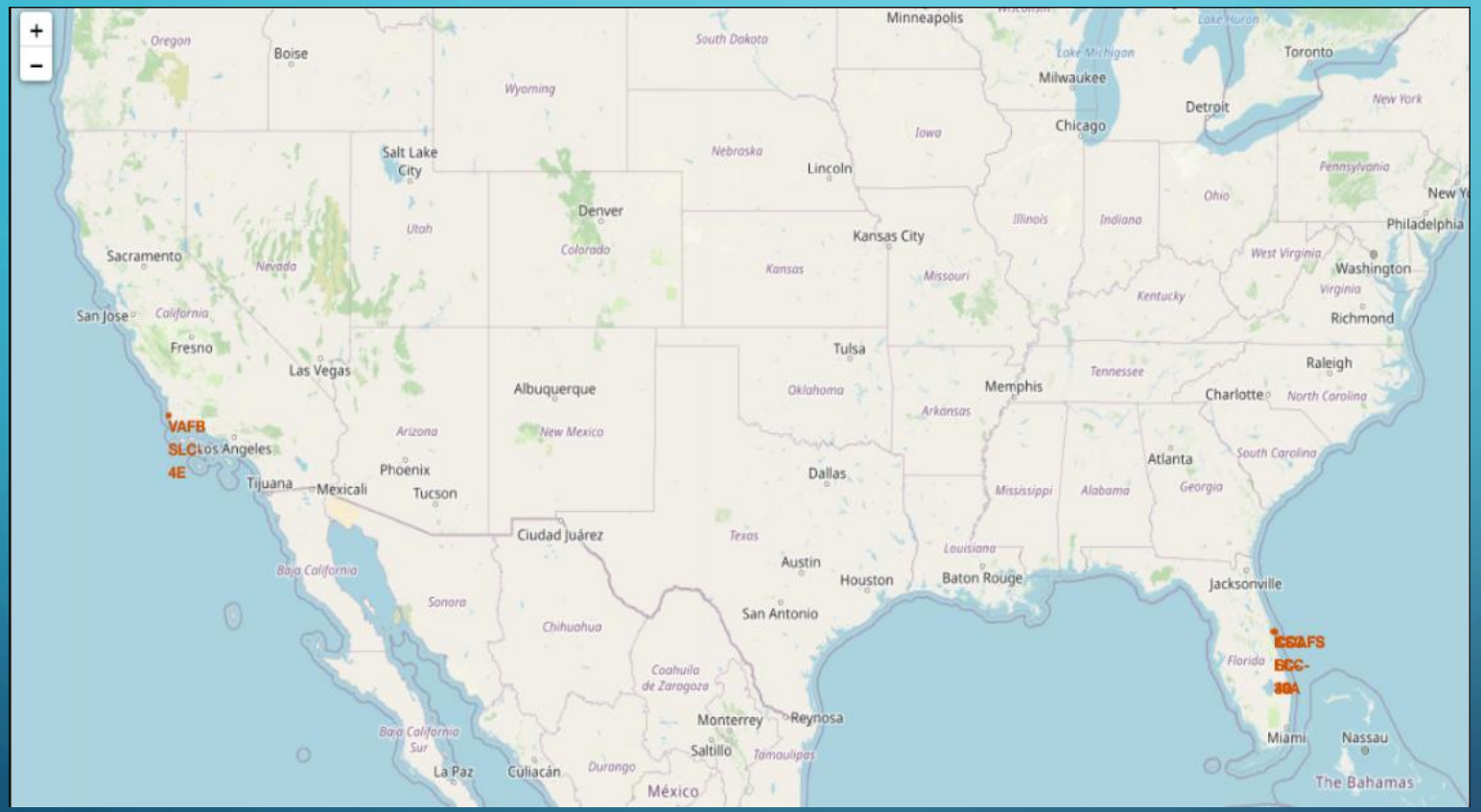
```
%%sql

select Landing__Outcome, count(*) as LandingCounts from spacextbl where Date between '2010-06-04' and '2017-03-20'
group by Landing__Outcome
order by count(*) desc;
```

- Description:
  - The ‘group by’ key word arranges data in column ‘Landing__Outcome’ into groups
  - The ‘between’ and ‘and’ keywords return data that is between 2010-06-04 and 2017-03-20
  - The ‘order by’ keyword arranges the counts column in descending order
  - The result of the query is a ranked list of landing outcome counts per the specified date range
- Result:

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

INTERACTIVE MAP WITH FOLIUM

SPACEX FALCON9 - LAUNCH SITES MAP



Each launch site contains a circle, label, and a popup to highlight the location and the name of the launch site. It is also evident that all launch sites are near the coast.

SPACEX FALON9 – SUCCESS/FAILED LAUNCH MAP FOR ALL LAUNCH SITES

Fig 2 – VAFB Launch Site with success/failed markers

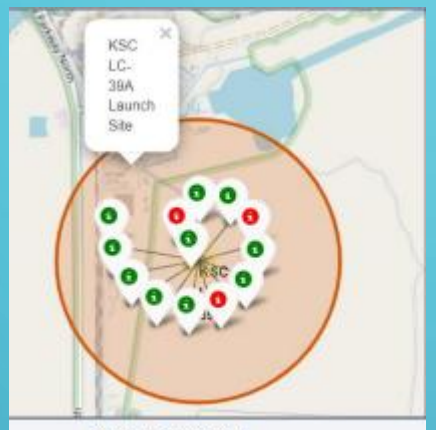


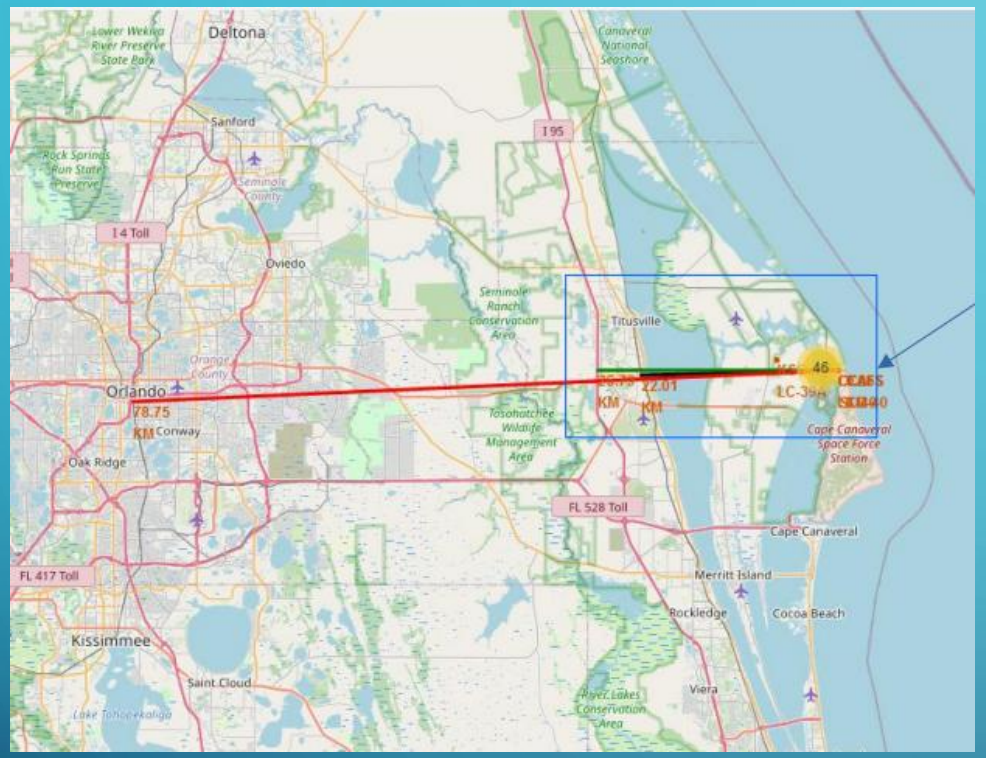
Fig 3 – KSC LC-39A success/failed markers

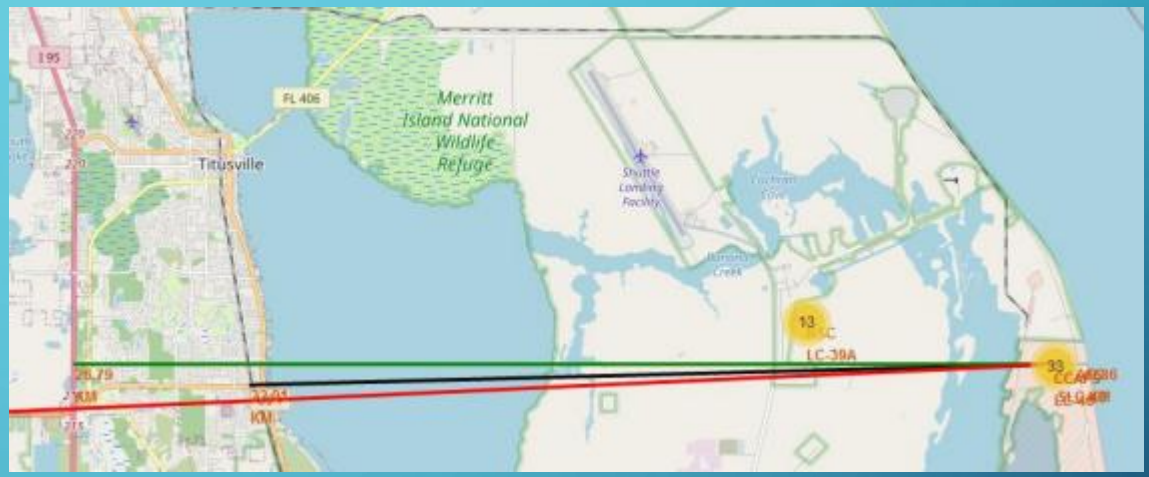
Fig 4 – CCAFS SLC-40 success/failed markers

Fig 5 – CCAFS SLC-40 success/failed markers

- Figure 2, 3, 4, and 5 zoom in to each site and displays the success/fail markers with green as success and red as failed
- By looking at each site map, KSC LC-39A Launch Site has the greatest number of successful launches

SPACEX FALCON9 – LAUNCH SITE TO PROXIMITY DISTANCE MAP





Distance from CCAFS_SLC40 to:

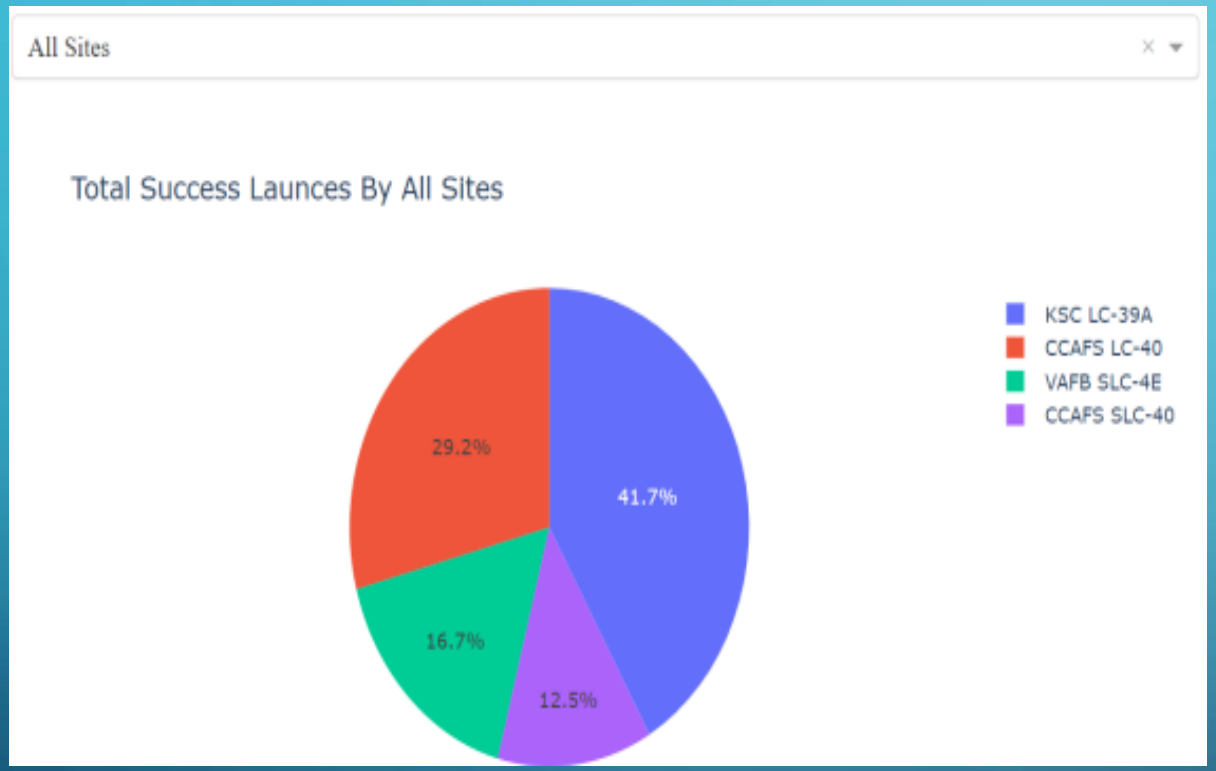
- Closest coast: ~900 m
- Florida East Coast Railway: 22.0 km
- Highway I 95: 26.8 km
- Orlando: 78.75 km

Launch sites are close to coasts. For safety issues if launcher is lost in the early stage of the flight. Rockets are launched:

- from West to East over the ocean in Florida.
- North or South bound over the ocean in California.

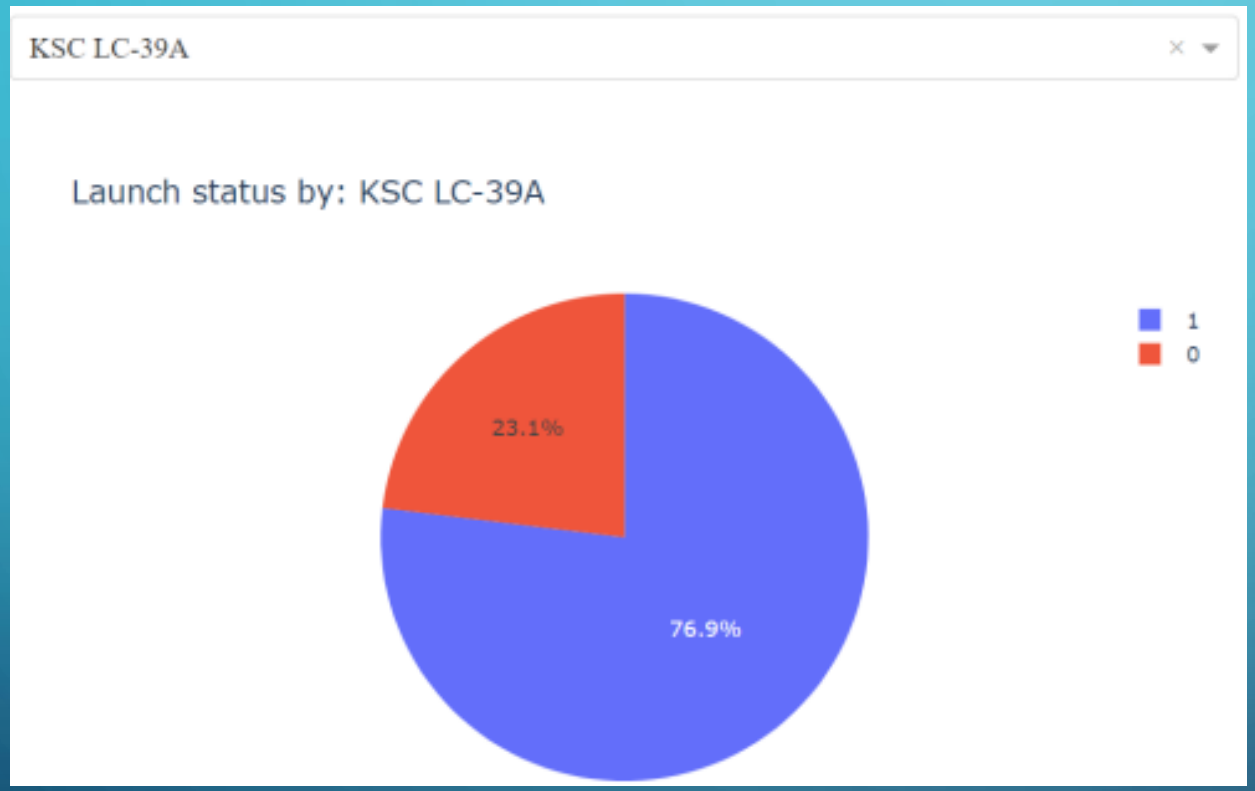
BUILD A DASHBOARD WITH PLOTLY DASH

LAUNCH SUCCESS COUNTS FOR ALL SITES



- Launch Site ‘KSC LC-39A’ has the highest launch success rate
- Launch Site ‘CCAFS SLC40’ has the lowest launch success rate

LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO



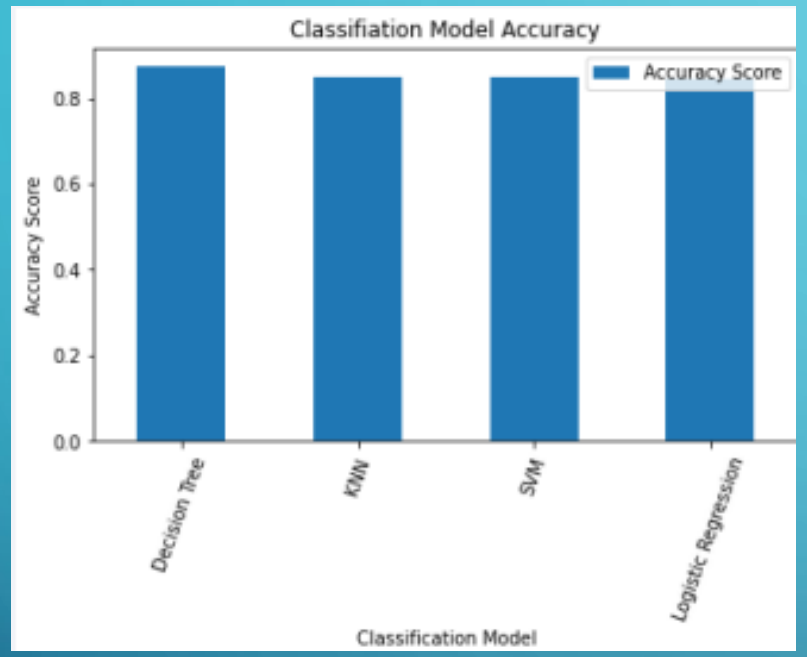
- KSC LC-39A Launch Site has the highest launch success rate and count
- Launch success rate is 76.9%
- Launch success failure rate is 23.1%

Payload vs. Launch Outcome Scatter Plot for all Sites

- Most successful launches are in the payload range from 2000 to about 5500
- Booster version category ‘FT’ has the most successful launches
- Only booster with a success launch when payload is greater than 6k is ‘B4’

PREDICTIVE ANALYSIS (CLASSIFICATION)

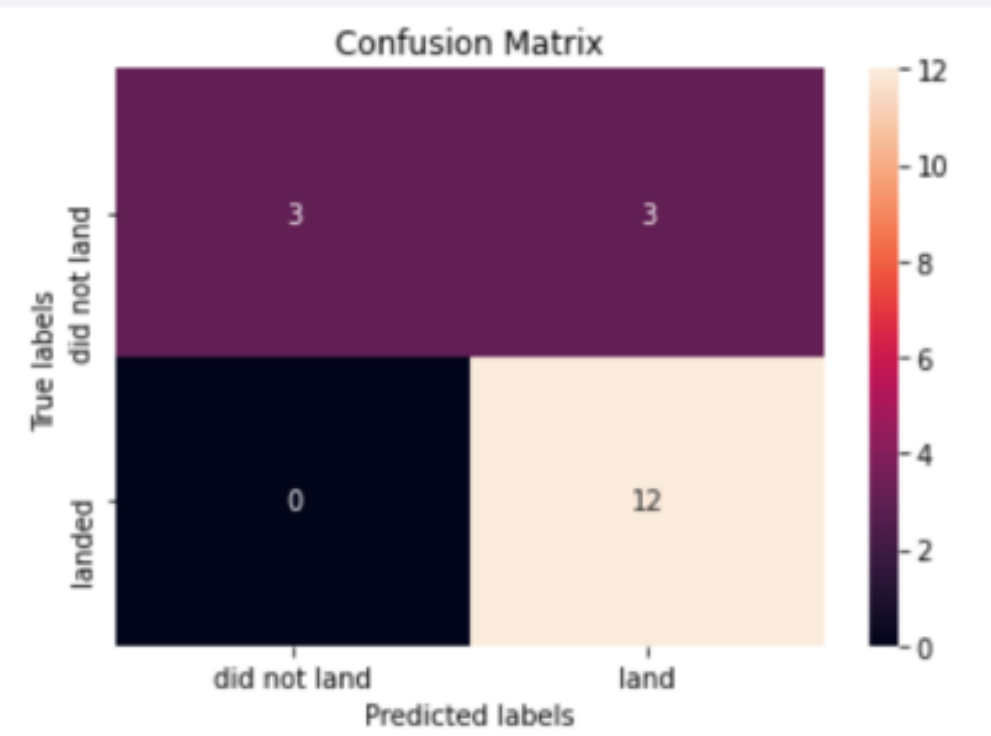
CLASSIFICATION ACCURACY



	Algo Type	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.875000	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

- Based on the Accuracy scores and as also evident from the bar chart, Decision Tree algorithm has the highest classification score with a value of .8750
- Accuracy Score on the test data is the same for all the classification algorithms based on the data set with a value of .8333
- Given that the Accuracy scores for Classification algorithms are very close and the test scores are the same, we may need a broader data set to further tune the models

CONFUSION MATRIX



- Per the confusion matrix, the classifier made 18 predictions
- 12 scenarios were predicted Yes for landing, and they did land successfully (True positive)
- 3 scenarios (top left) were predicted No for landing, and they did not land (True negative)
- 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive)

CONCLUSION

- As the numbers of flights increase, the first stage is more likely to land successfully • Success rates appear go up as Payload increases but there is no clear correlation between Payload mass and success rates
- Launch success rate increased by about 80% from 2013 to 2020
- Launch Site ‘KSC LC-39A’ has the highest launch success rate and Launch Site ‘CCAFS SLC40’ has the lowest launch success rate
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest
- Lunch sites are located strategically away from the cities and closer to coastline, railroads, and highways
- The best performing Machine Learning Classification Model is the Decision Tree with an accuracy of about 87.5%. When the models were scored on the test data, the accuracy score was about 83% for all models. More data may be needed to further tune the models and find a potential better fit.

APPENDIX

Special Thanks to:

Instructor

Coursera

IBM