DATA SCIENCE CAPSTONE

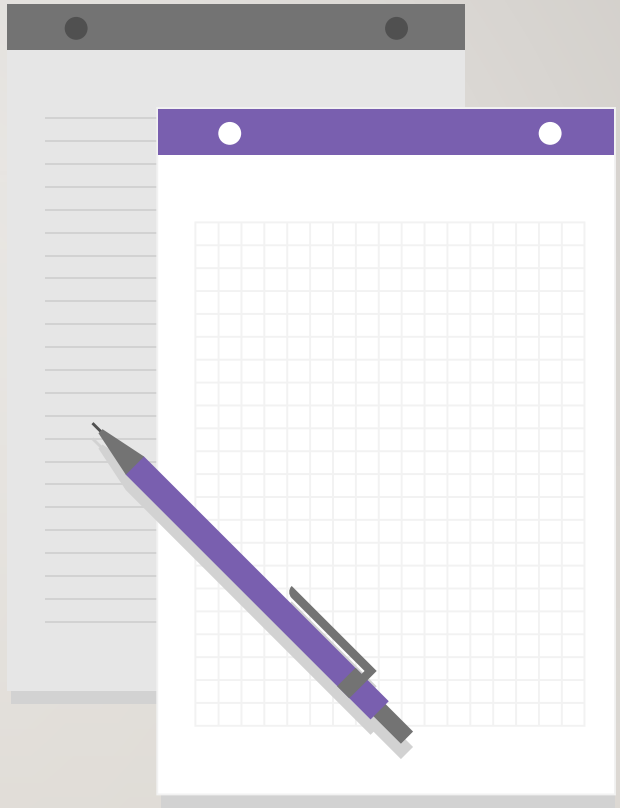
Md Mahmudur Rahman

3rd Dec. 2022

https://github.com/Tutul67/Applied-Data-Science-Capstone.git

OUTLINE

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



EXECUTIVE SUMMARY

- Data was Collected using Web- scraping and REST API queries from public SpaceX API and SpaceX Wikipedia page.
- Data Wrangling to Classify Launches based on Success and transform data into standardized numeric form.
- Exploratory Data Analysis using SQL and Visualization packages for Python.
- Interactive Plotly Web App to visualize payload and success launch data at each Launch Site.
- Exploring Launch Sites using interactive Folium Maps.
- Predictive analysis for classification of Rocket Landing Success .

INTRODUCTION

Background & Context

SpaceX triumphantly declare can launch a spaceflight with a Falcon 9 rocket with a budget of 62million dollars provided it's reusable first stage booster landed back safely. Where it's competitive counterpart can do that a whooping budget of 165million dollars. Therefore, if we can predict with a resounding 'yes' then it is possible to estimate a budget. Based on public information and with a ML we will try to reach a conclusive summary.

Challenges to be addressed

How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

Does the rate of successful landings increase over the years?

What is the best algorithm that can be used for binary classification ?

METHODOLOGY

Data Collection

- Combined data from SpaceX public API and SpaceX Wikipedia page.

Perform Data Wrangling

- Filtering The Data
- Dealing With The Missing Values
- Using One Hot Encoding and prepare the Data for Binary Classification.

Perform Exploratory Data Analysis (EDA) using Visualization and SQL.

Perform Interactive Visual Analytics using Folium, Plotly Dash.

Perform predictive analysis using Classification models

- Tuned models using GridSearchCV.

DATA COLLECTION

- Used SpaceX REST API to gather data on rocket launches:

  https://api.spacexdata.com/v4

- API provides data on rockets used: launch dates, payload masses, launch success or failure, launch sites name and location(lat. & long.),booster version(Falcon-9 booster was our interest).

- Falcon-9 launch Data was collected thru web scrapping on Wikipedia SpaceX using BeautifulSoup.

  https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- Data Columns : launch site, payload ,payloadmass,otrbit,customer,launch outcome, booster version, date time.

✓ DATA COLLECTION SPACEX API

Request(SpaceX API's

.JSON file +Lists(Launch Site, Booster version, Payload data)

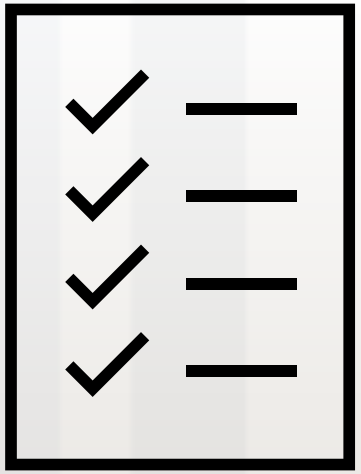
JSON_normalize to DataFrame data from JSON

Dictionary relevant data

Put dictionary to DataFrame

Filter data of our interest in Falcon 9 launches

Replace missing PayloadMass values with mean



DATA COLLECTION-WEB SCRAPPING

Request Wikipedia html

BeautifulSoup 'html5lib' Parser

Find launch info html table

Cast dict. To DataFrame

Iterate through table cells to extract data to dict.

Create Dictionary

DATA WRANGLING

- Perform exploratory Data Analysis and determine Training Labels
- Training label with landing outcomes where successful= 1 & failure=0 Outcome column has two components ‘Mission Outcome', 'Landing-Location’
- Get a training column ‘class 'with value 1 when 'Mission Outcome’ is True and 0 other way.
- Mapping
  - For True ASDS,RTLS &True-Ocean set----- 1
  - For None None, False, ASDS, None ASDS, False RTLS ,False-Ocean set----- 0
- Where ‘Ocean’ means landing some location in the sea
- Where RTLS means landing on ground pad.
- Where ASDS means landing 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

Scatter Plots

show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Bar /Line charts

show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and get a measured value.

Line charts show trends in data over time (time series)

IBM Developer

EDA WITH SQL

- Loaded data set into IBM DB2 Data Base.
- Asking Relevant queries using SQL Python integration.
- Queries were made to get better understanding about the data set. Like launch site names, mission outcomes, various payload sizes, about booster version, landing outcomes.

BUILD A DASHBOARD WITH FOLIUM

- Visualized Launch Data in an interactive Map

- Used Latitude and Longitude Coordinates of Launch Sites to Add Circle Markers with the site names labeled.
- Assigned Launch Outcome (Success/Failure) from the data frame to Classes 1 and 0 respectively and assigned the classes Green and Red markers on the map to Marker Clusters grouped by Launch Site.
- Used lines and points to measure and label the minimum distances of the launch sites to: ▪ Cities ▪ Highways ▪ Coastlines ▪ Railway
- Answered the following Questions: •
  - Are launch sites in close proximity to railways? Yes
  - • Are launch sites in close proximity to highways? Yes
  - • Are launch sites in close proximity to coastline? Yes
  - • Do launch sites keep certain distance away from cities? About 50 km

DASHBOARD WITH PLOTLY DASH

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success

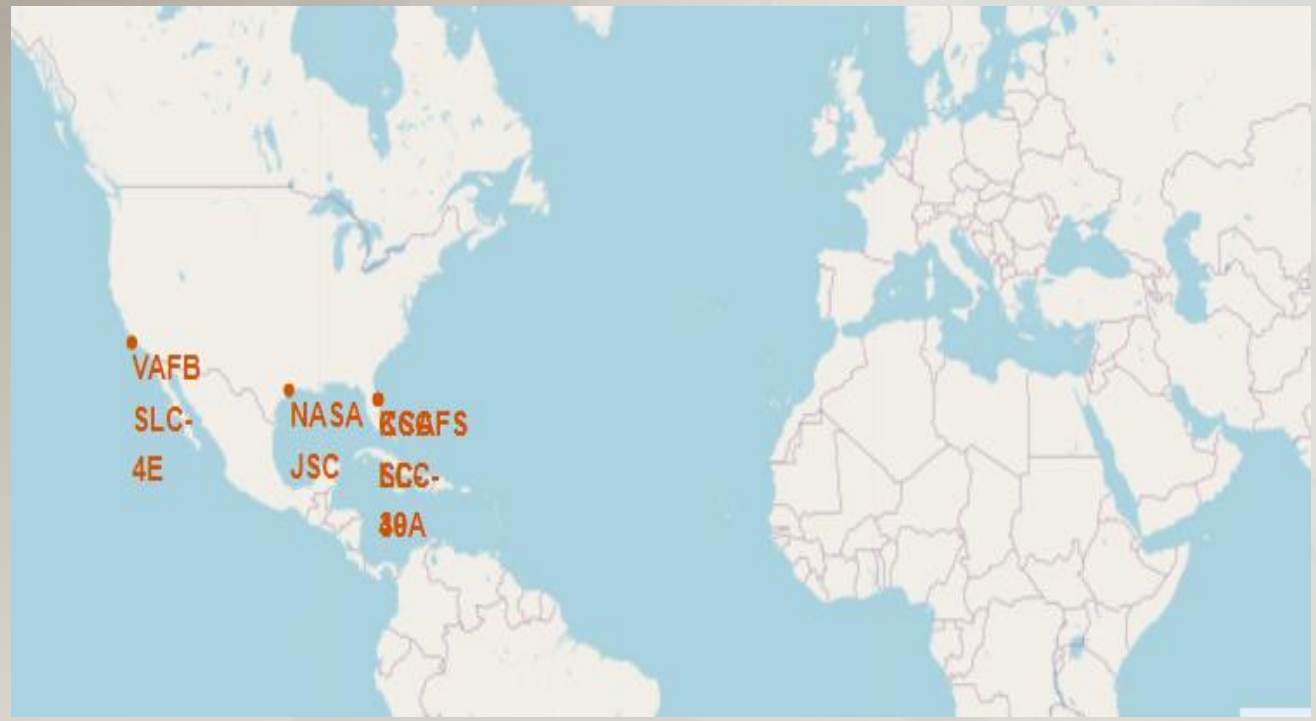
PREDICTIVE ANALYSIS (CLASSIFICATION)

- Creating a NumPy array from the column "Class" in data
- Standardizing the data with StandardScaler, then fitting and transforming it
- Splitting the data into training and testing sets with train_test_split function
- Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
- Calculating the accuracy on the test data using the method .score() for all models.
- Confusion Matrix for all models.
- Finding the method performs best by examining the Jaccard_score and F1_score metrics

RESULTS

INTERACTIVE MAP WITH FOLIUM

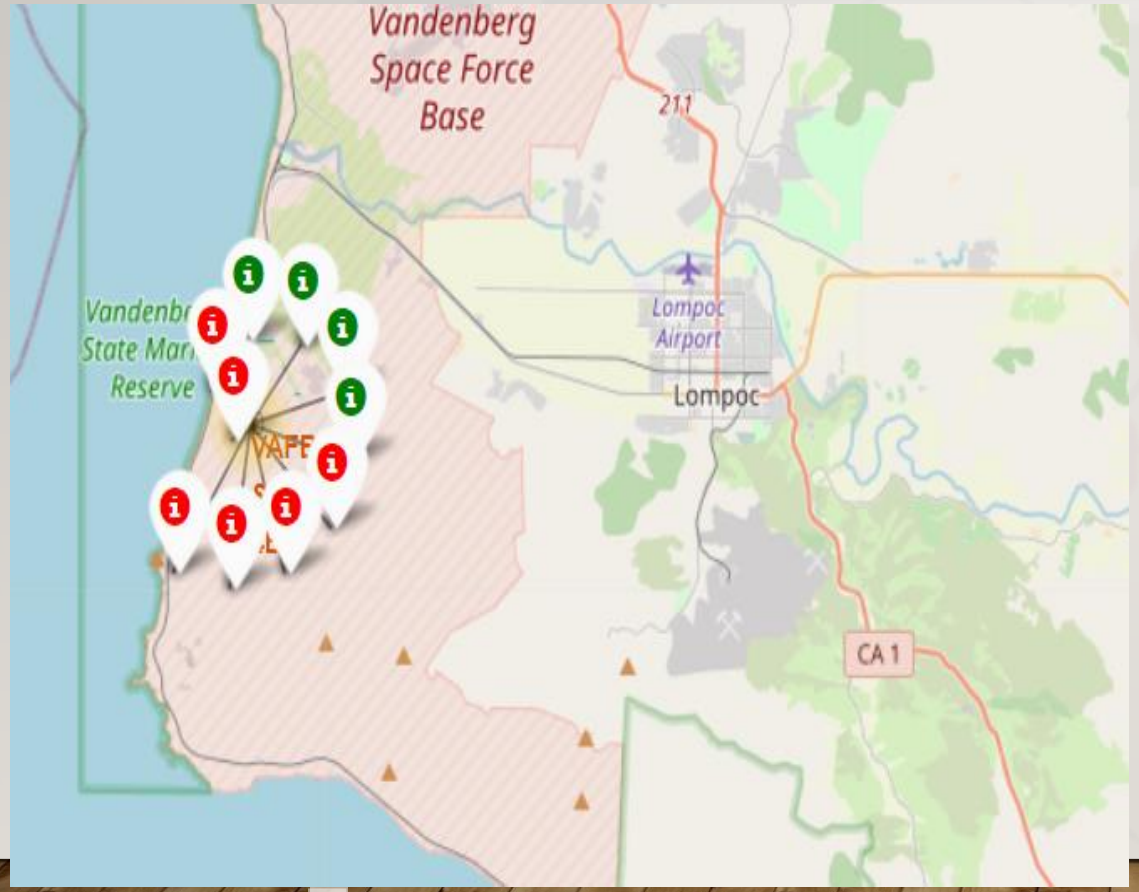
GLOBAL MAP OF SPACEX LAUNCH SITES



SPACEX VAFB SLC4E CALIFORNIA

OUR CLUSTER MARKER INDICATES 10 FALCON 9 LAUNCHES HAVE TAKEN PLACE AT THIS SITE

SPACEX VAFB SLC4E LAUNCH SITE(CALIFORNIA) LANDED OR FAILED



- **Zooming in we can see our markers color coded to indicate how many launches landed successfully**

4 Rockets Landed

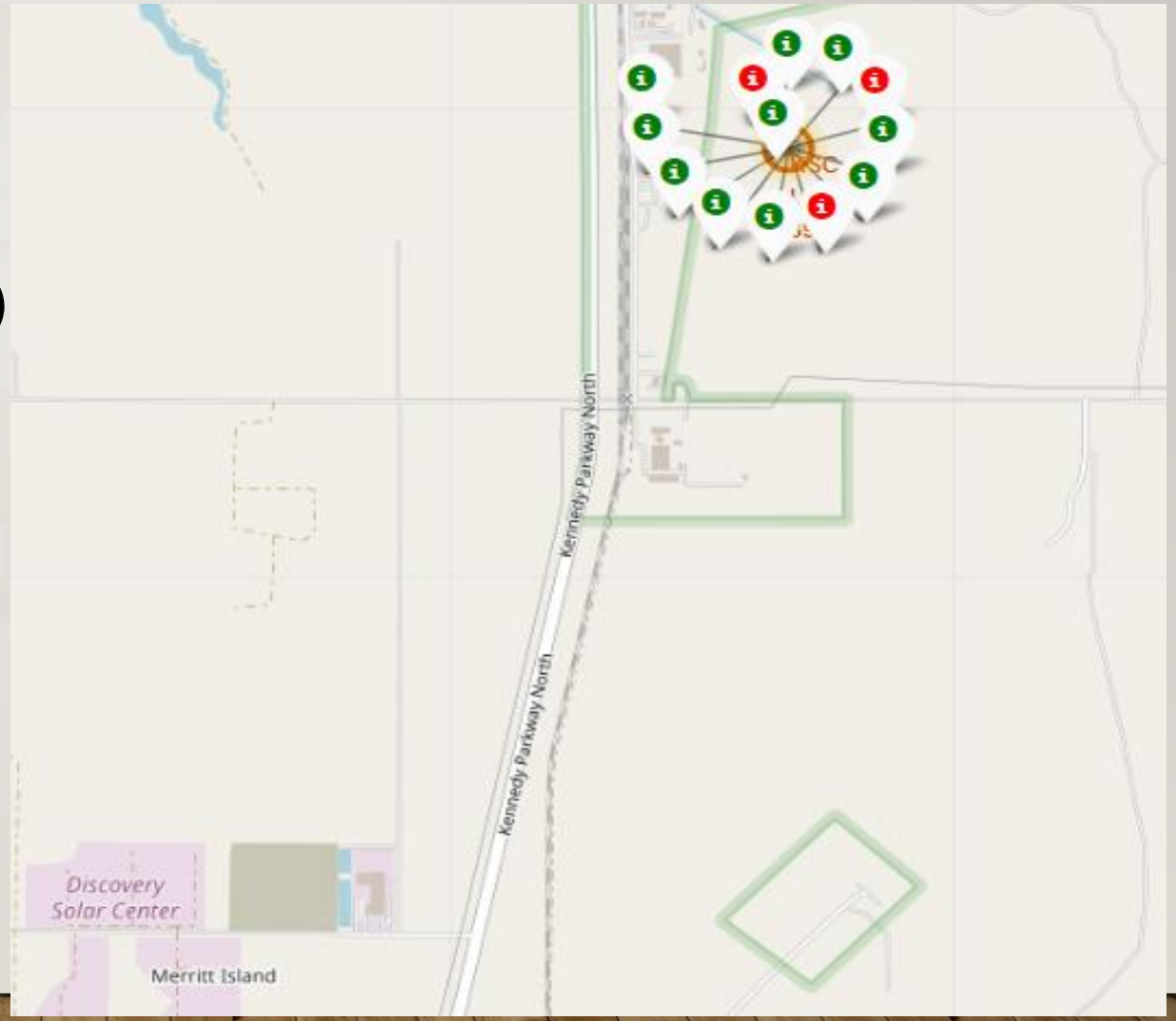
6 Rockets Failed to Land

SPACEX FLORIDA LAUNCH SITES

- **Space X has 2 separate bases with 3 total launch sites located on**
  - **Merritt Island(KSC LC -39A)**
  - **Cape Canaveral(CCAFS LC40, CCAFS SLC -40)**

SPACEX KSC LC39A LAUNCH SITE MERRITT ISLAND LANDED OR FAILED(FLORIDA)

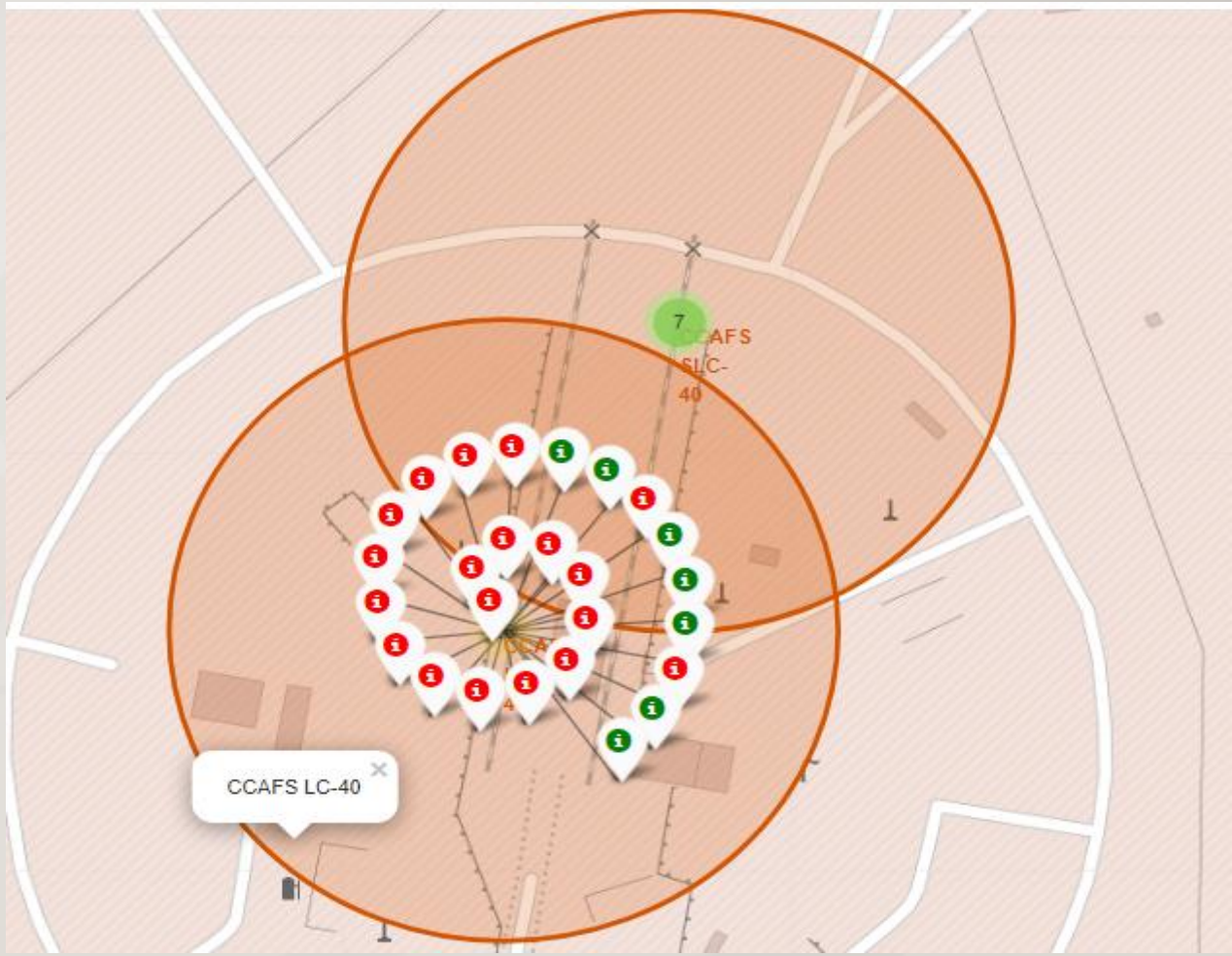
- **SpaceX KSC LC39A Launch Site Merritt Island**
  - **10 Rockets Landed**
  - **3 Rockets Failed to Land**

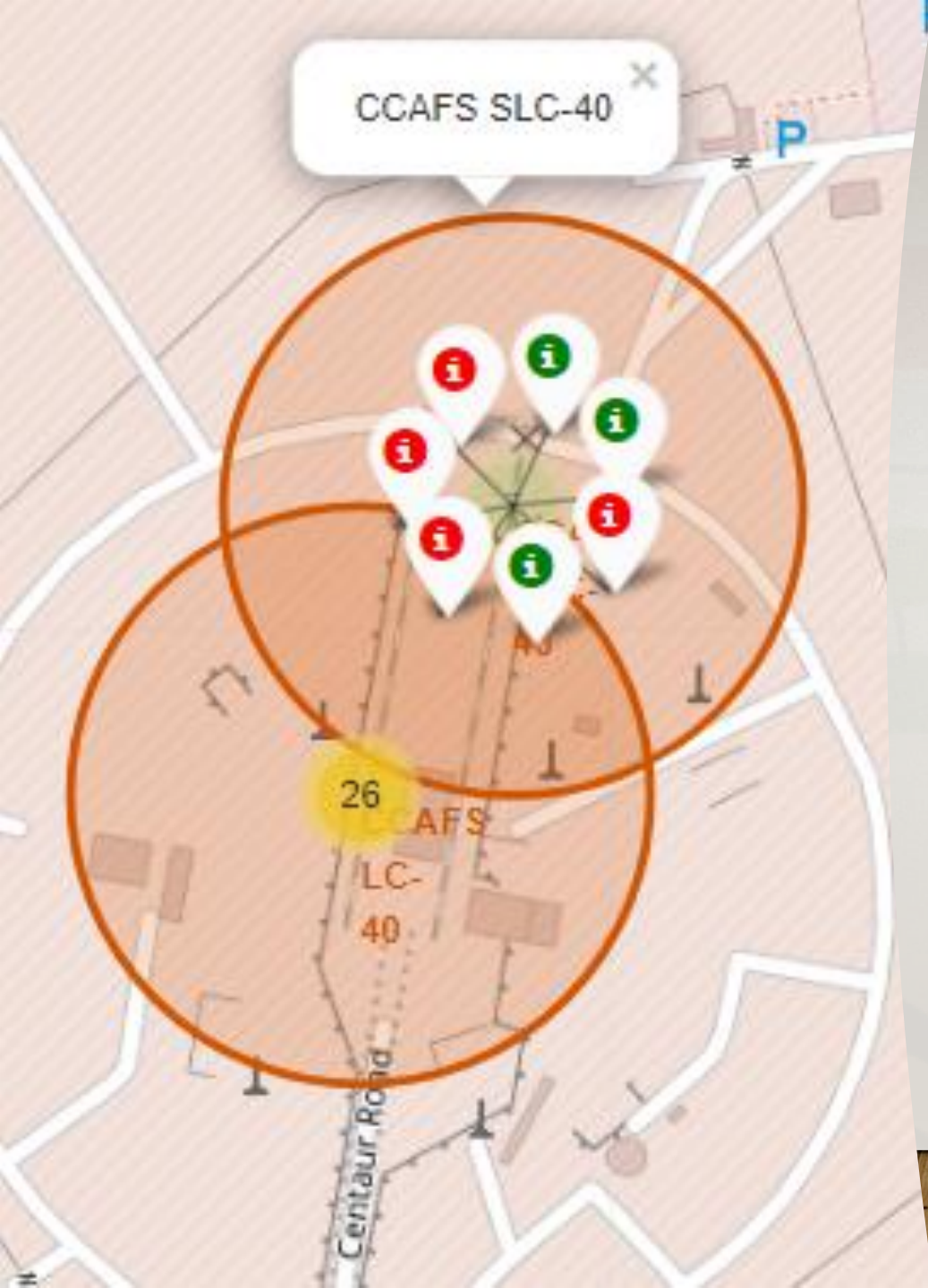


SPACEX CCAFS LC-40 LAUNCH SITE CAPE CANAVERAL, FLORIDA(LANDED OR FAILED)

- 7 Rockets Landed
- **19 Rockets Failed to Land**

- 26 Total Launches at this Site





SPACEX CCAFS SLC-40 LAUNCH SITE CAPE CANAVERAL,FLORIDA(LANDED OR FAILED)

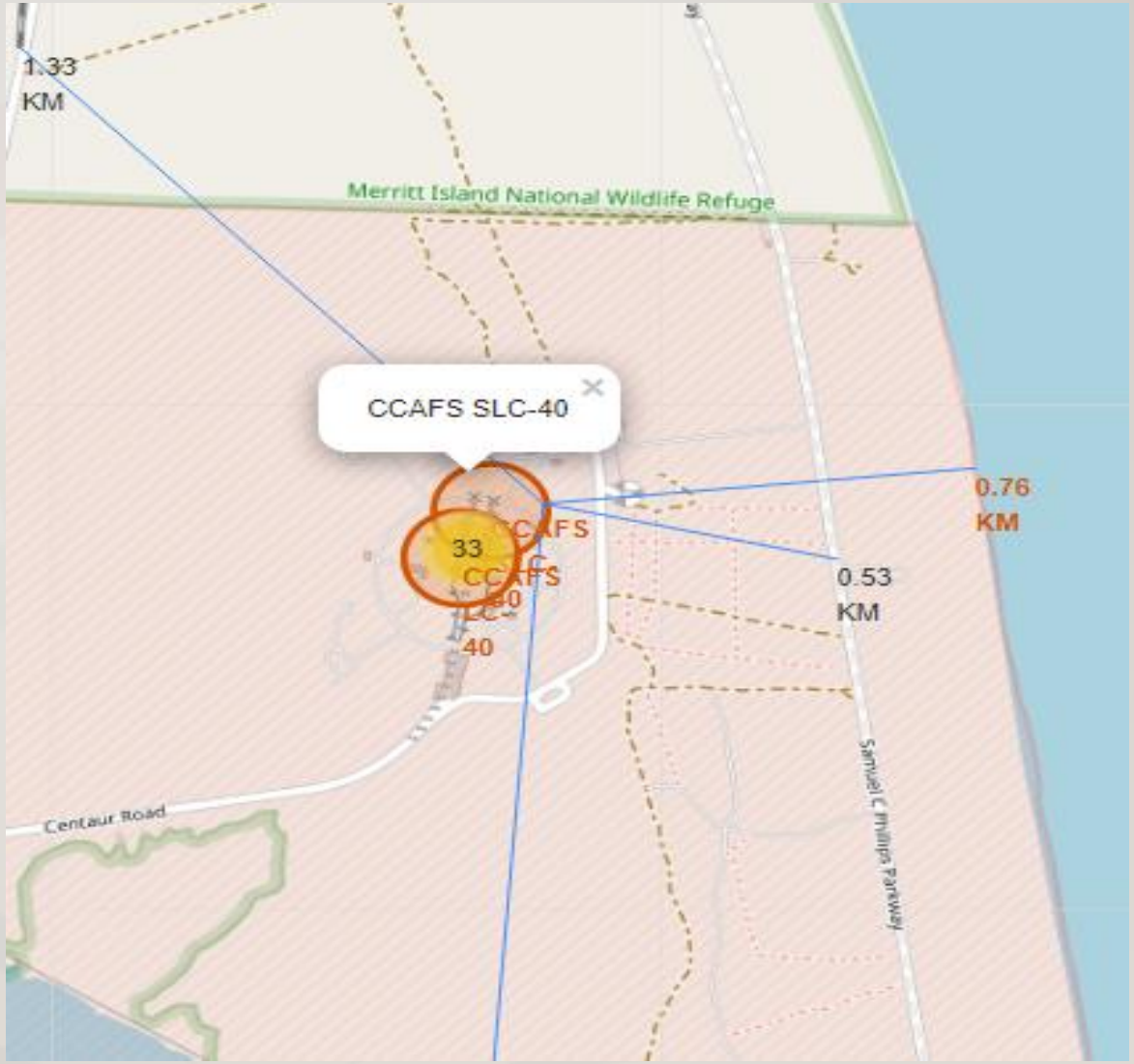
- 3 Rockets Landed
- 4 Rockets Failed to Land

- **7 Total Launches at this Site**

SKILLS NETWORK

INFRASTRUCTURE NEAR THE LAUNCH SITE

- **Nearest coastline from the launch site 0.76km**
- **Nearest Railway Station 1.33km**
- **Nearest Highway 0.53km**
- ✓ **From this we recognize that infrastructure is in close proximity to launch sites for easy access to manufactured parts**

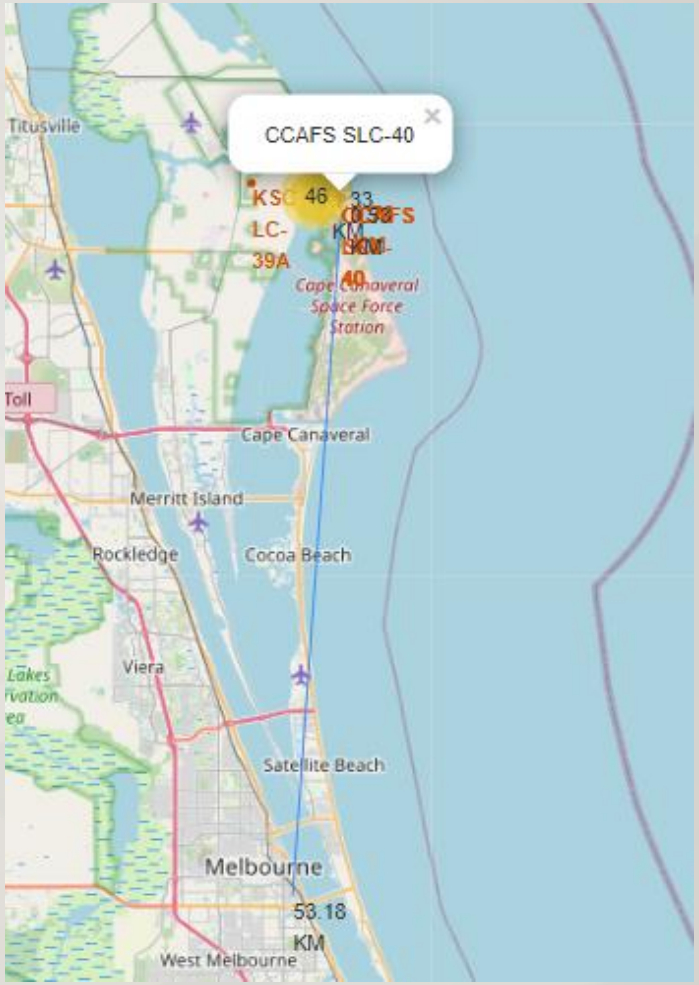


CITY DISTANCE TO LAUNCH SITES

The nearest city to the launch sites in Florida is Melbourne located

53.18Km from launch Site

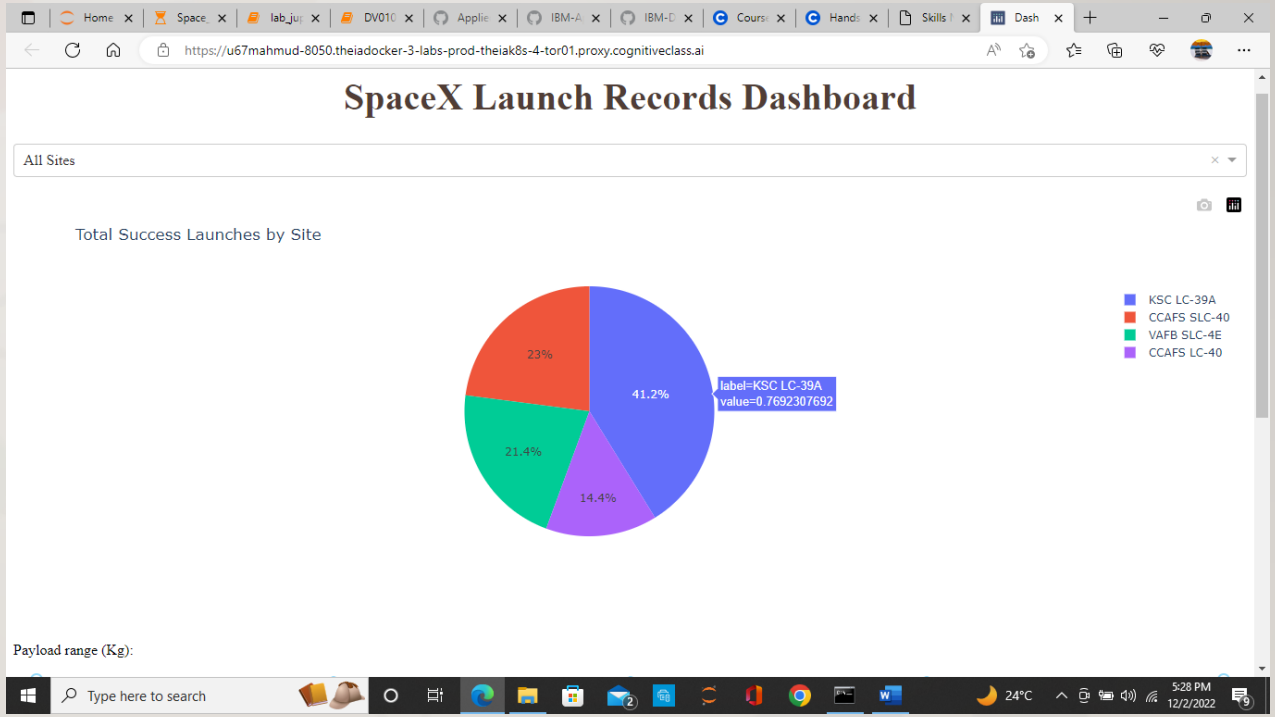
To avoid a catastrophe from a disaster arising from failed ,doomed launch ,a city usually located a far off.



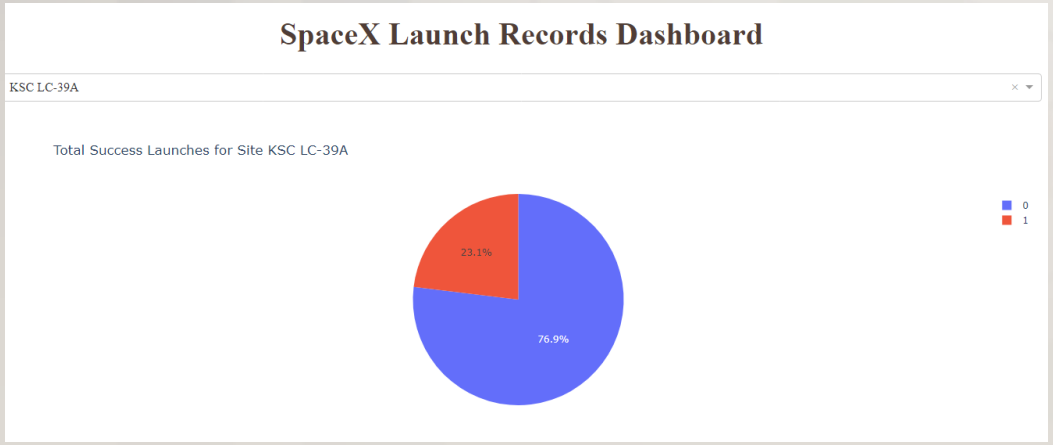

BUILD A DASHBOARD
FROM PLOTLY-DASH

PROBABILITY OF LAUNCH SITE GIVEN SUCCESSFUL LANDING

We see that the KSC LC-39A Launch site accounts for the largest percentage of the total number of successful landings at 41.2%



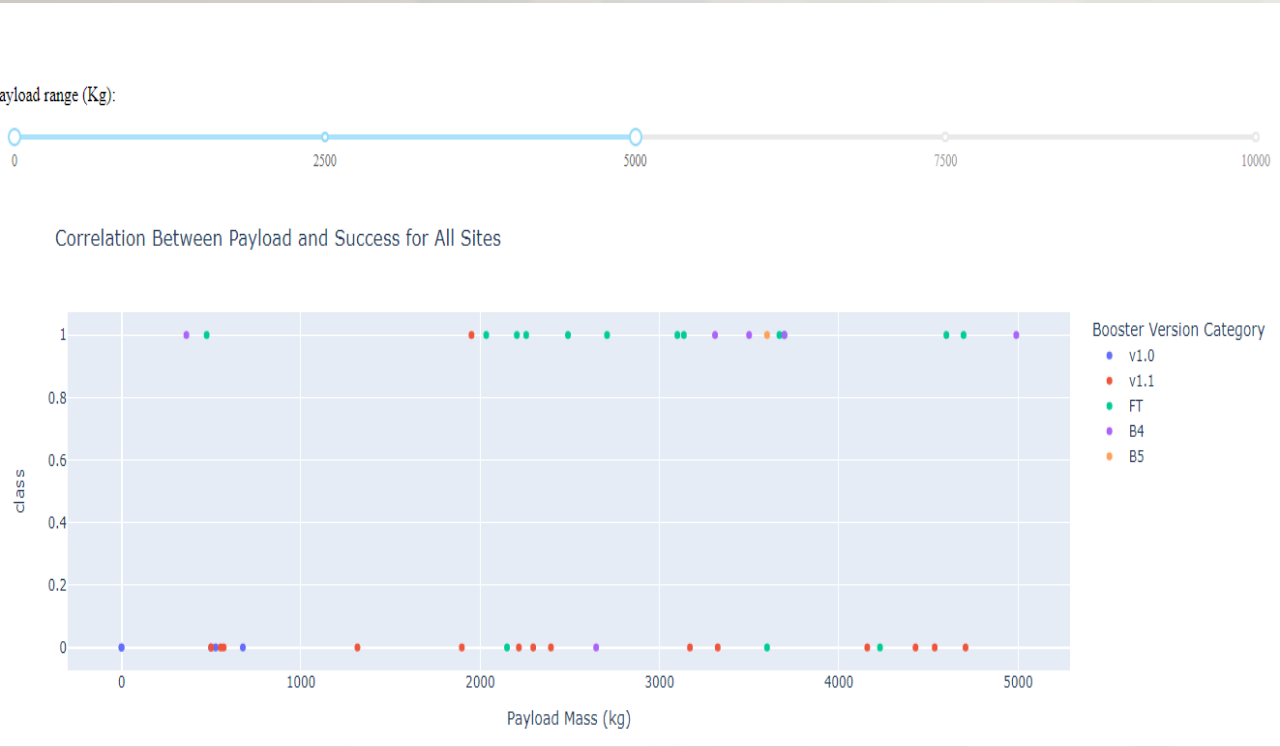
LAUNCH SITE WITH THE HIGHEST PROBABILITY OF SUCCESS



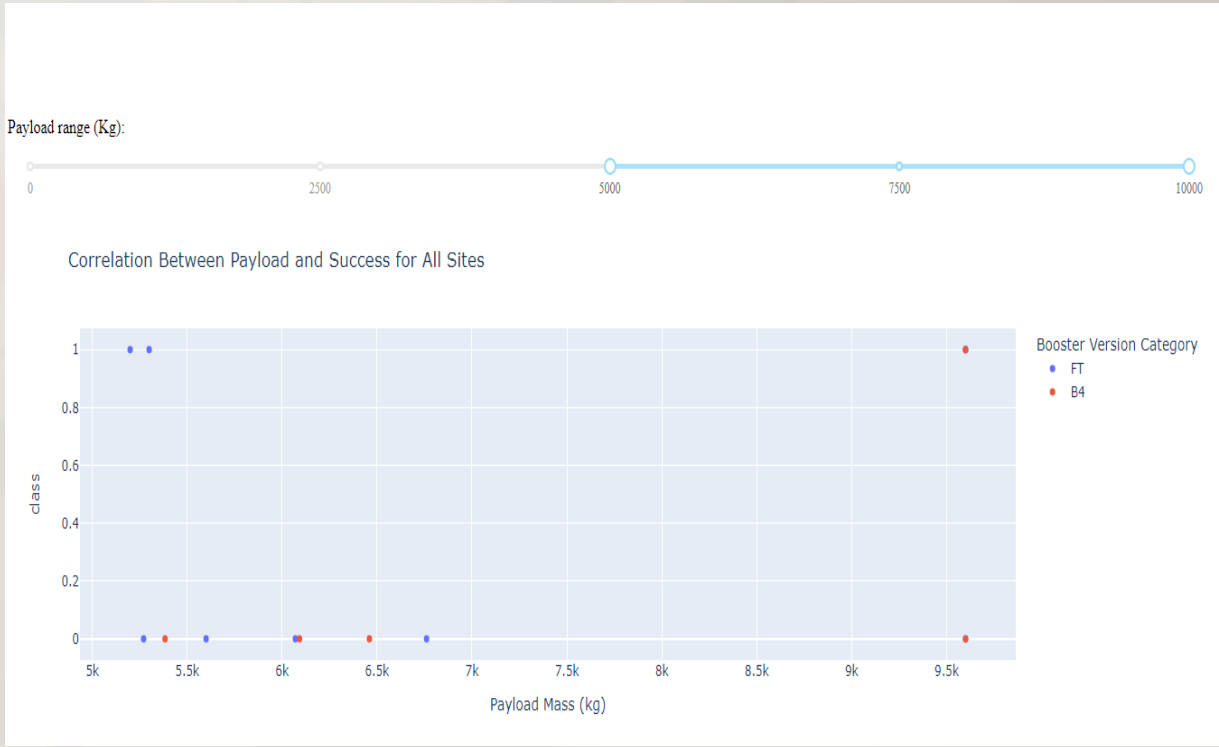
- The KSC LC-39A Launch Site also has the highest probability of success per launch
- 76.9% of all launches at the KSC LC-39A Site Land Successfully

PAYLOAD MASS VS. LAUNCH OUTCOME FOR ALL SITES

PAYLOADS UNDER 5000KG



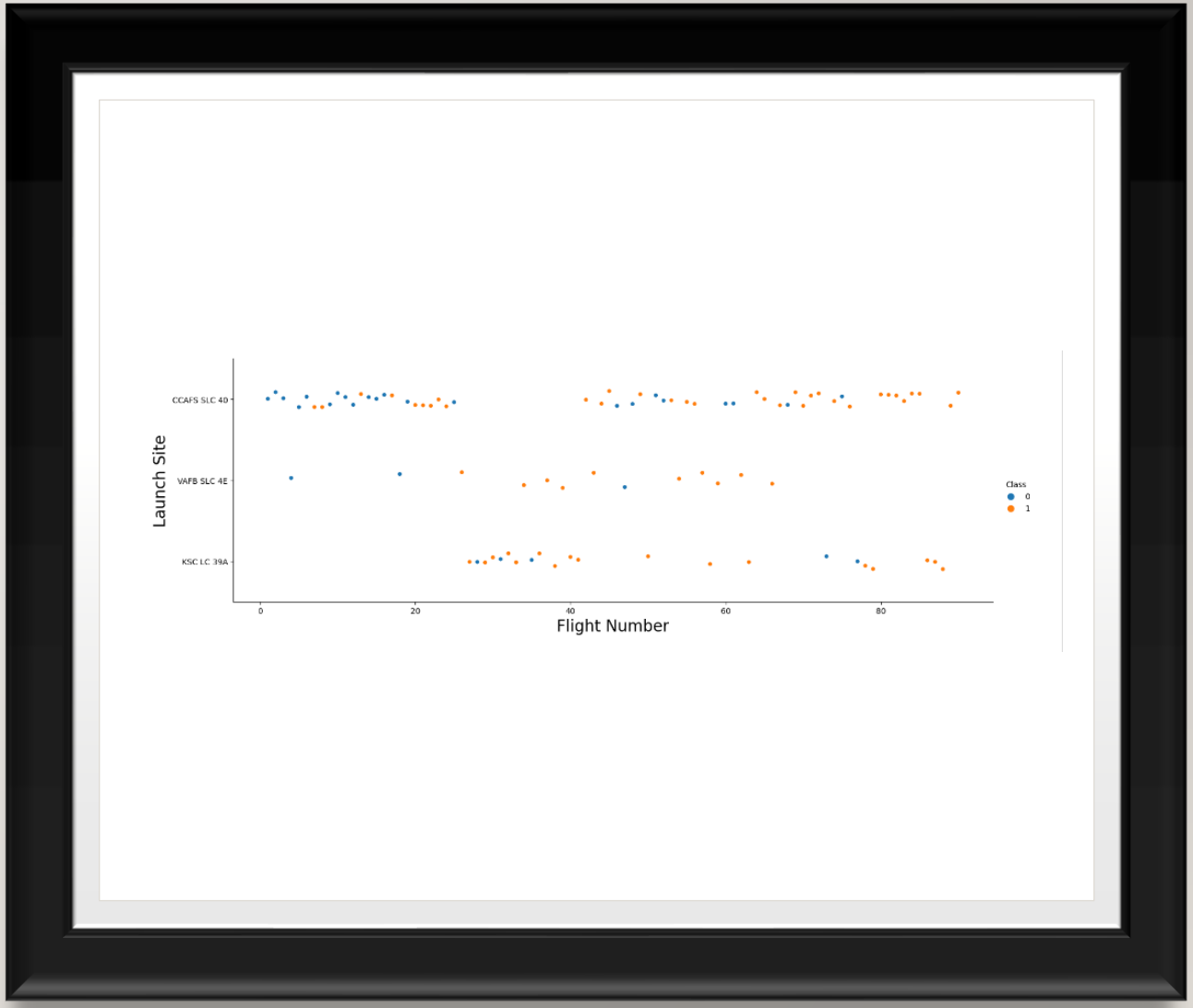
PAYLOADS OVER 5000KG

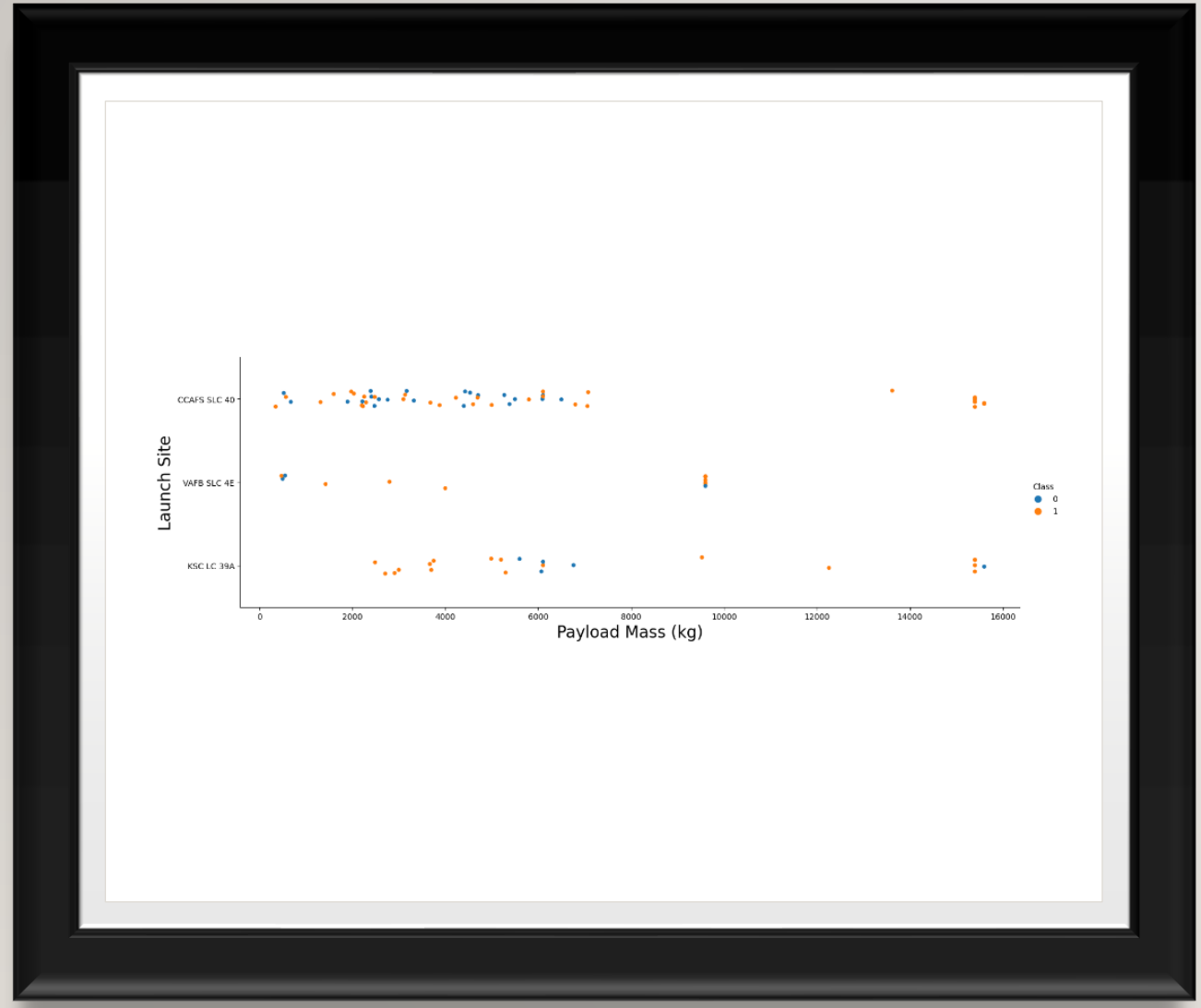


EDA WITH VISUALIZATION

FLIGHT NUMBER VS. LAUNCH SITE

- We can see that as the flight number increases the number of successful Landings increases
- CCAFS-40 appears to be the main launch site as it has the most volume. Mixed outcome(0 or 1)
- VAFB SLC 4E and KSC LC 39A have higher success rates
- . • It can be assumed that each new launch has a higher rate of success



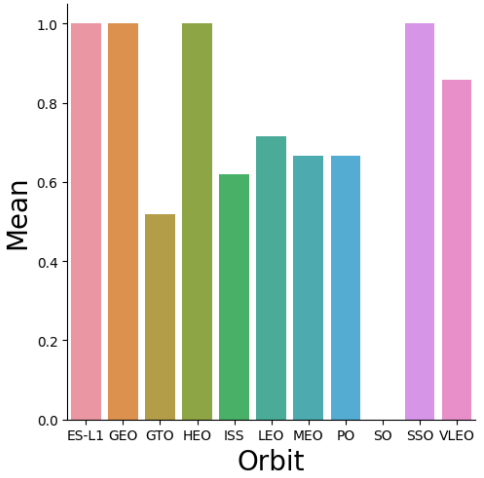


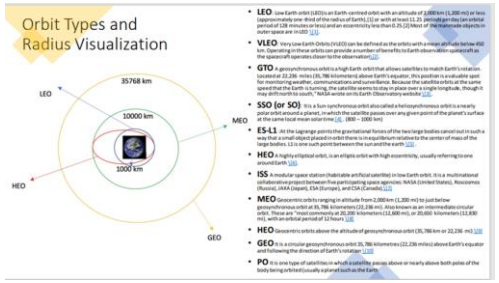
PAYLOAD VS. LAUNCH SITE

- As payload mass increases for Site CCAFS SLC 40, the probability of a successful landing increases
- Most of the launches is under 7000kg approx..
- Heavier lift off is 90000kg and above.
- CCAFS SLC 40 site in range 7000kg have mixed outcome where as KSC LC 39A site are with most successful landing though a fewer flights.

SUCCESS RATE VS. ORBIT TYPE

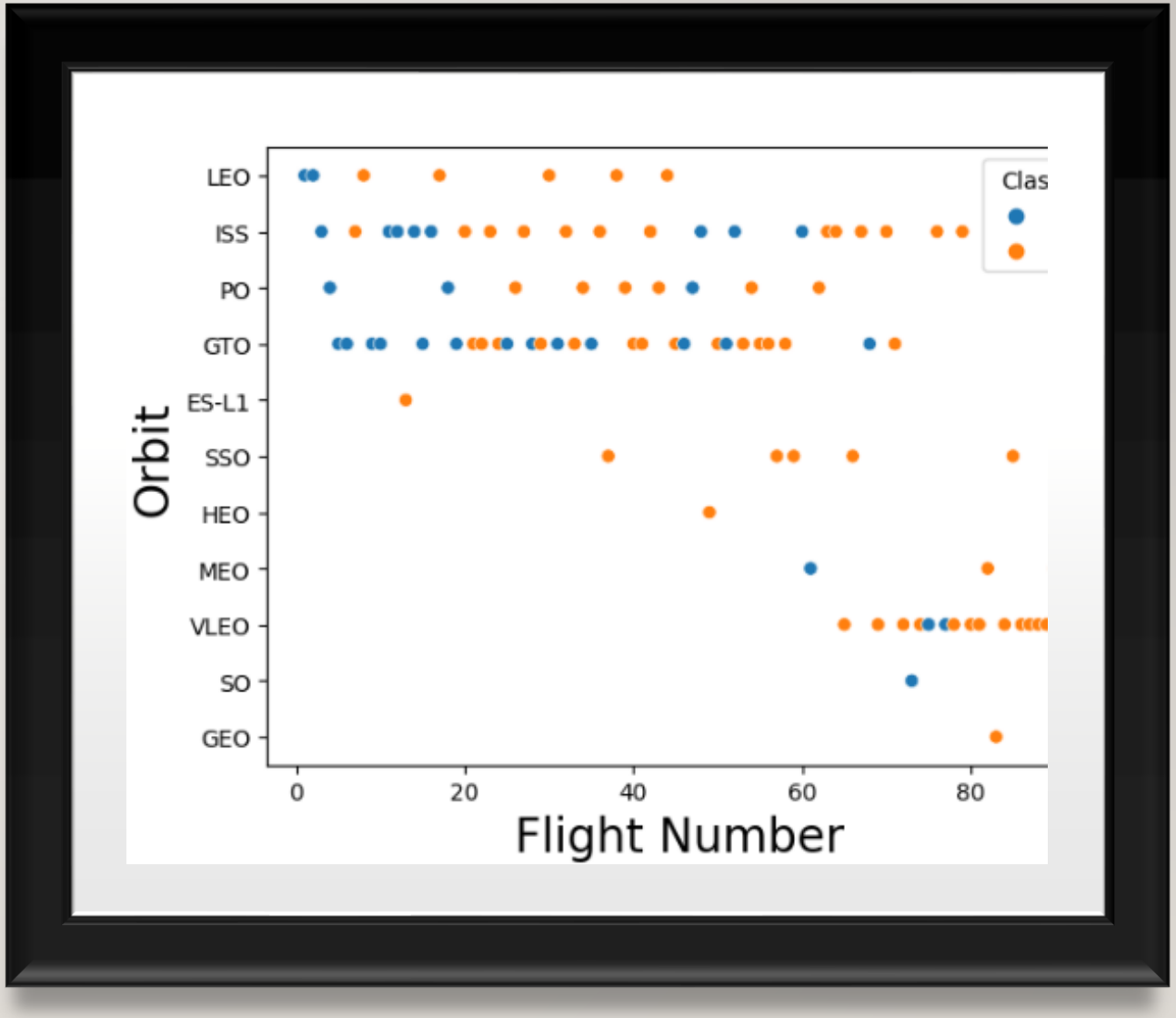
- With 100% successful orbit to get were ES-1,GEO,HEO,SSO,
- Above 80% but less then 100% was VLEO orbit.
- Between 50% to 70% success orbit were GTO,ISS,LEO,MEO,PO.
- 0% was for SO orbit

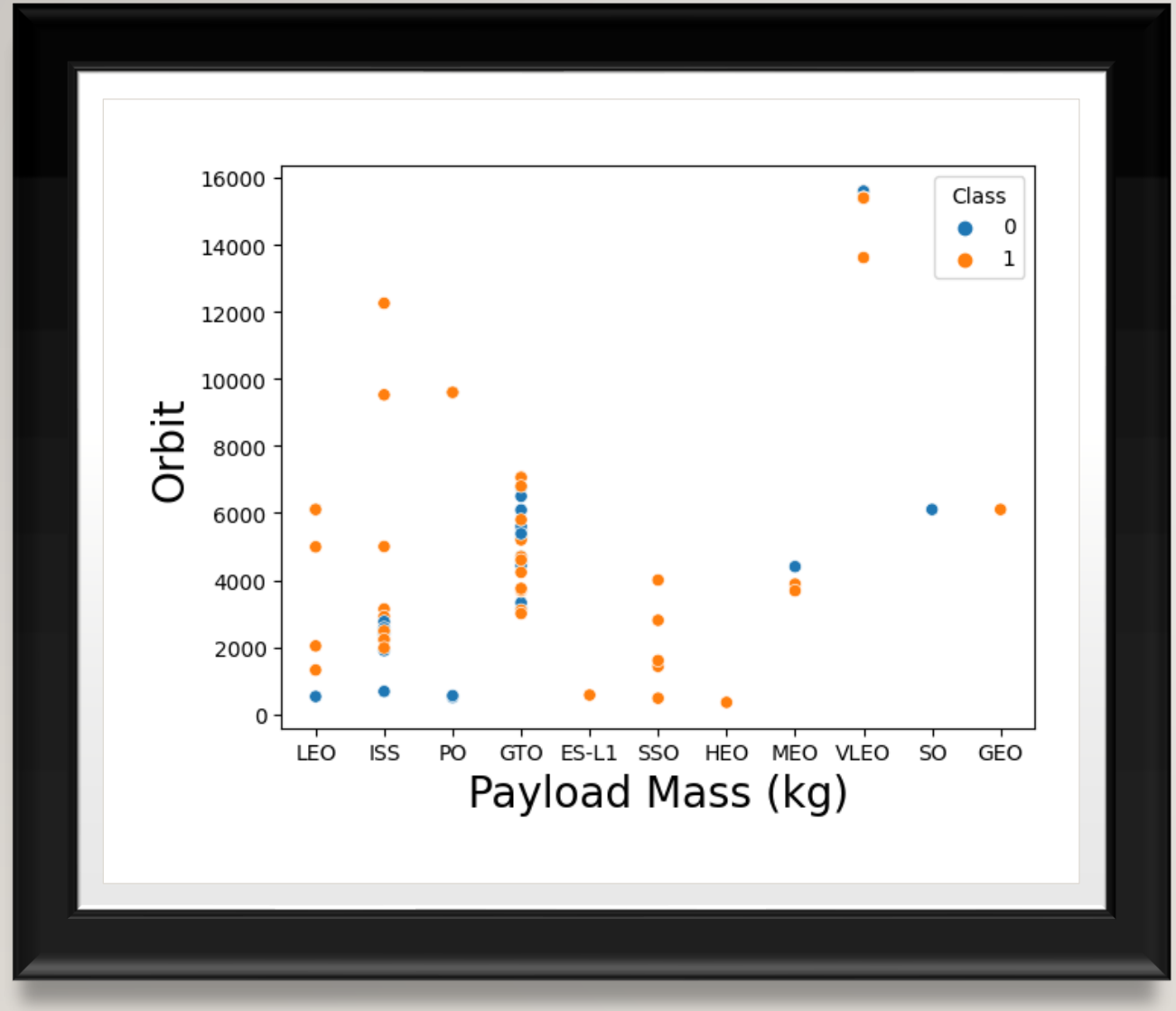




FLIGHT NUMBER VS. ORBIT TYPE

- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits.
- In 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.



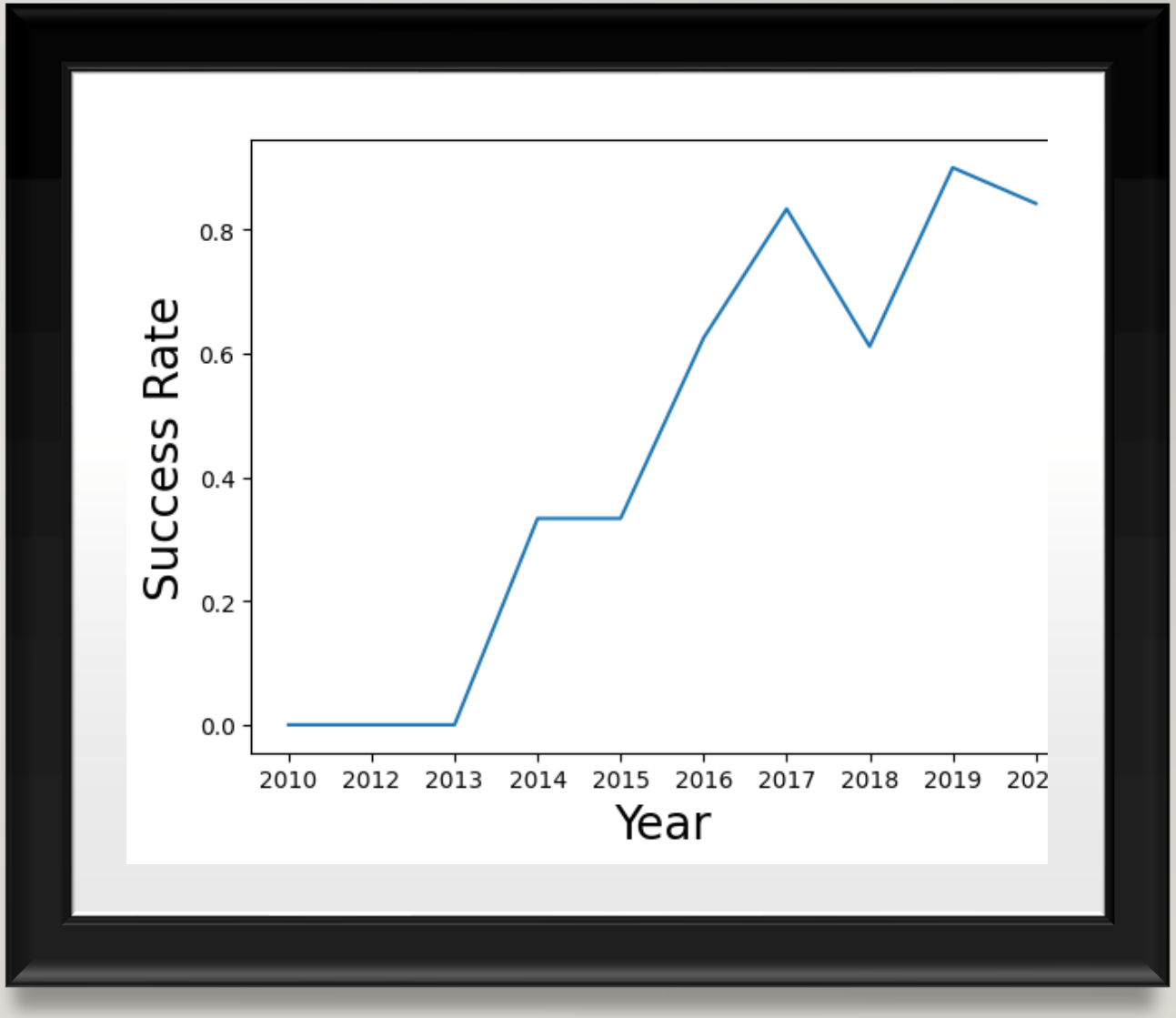


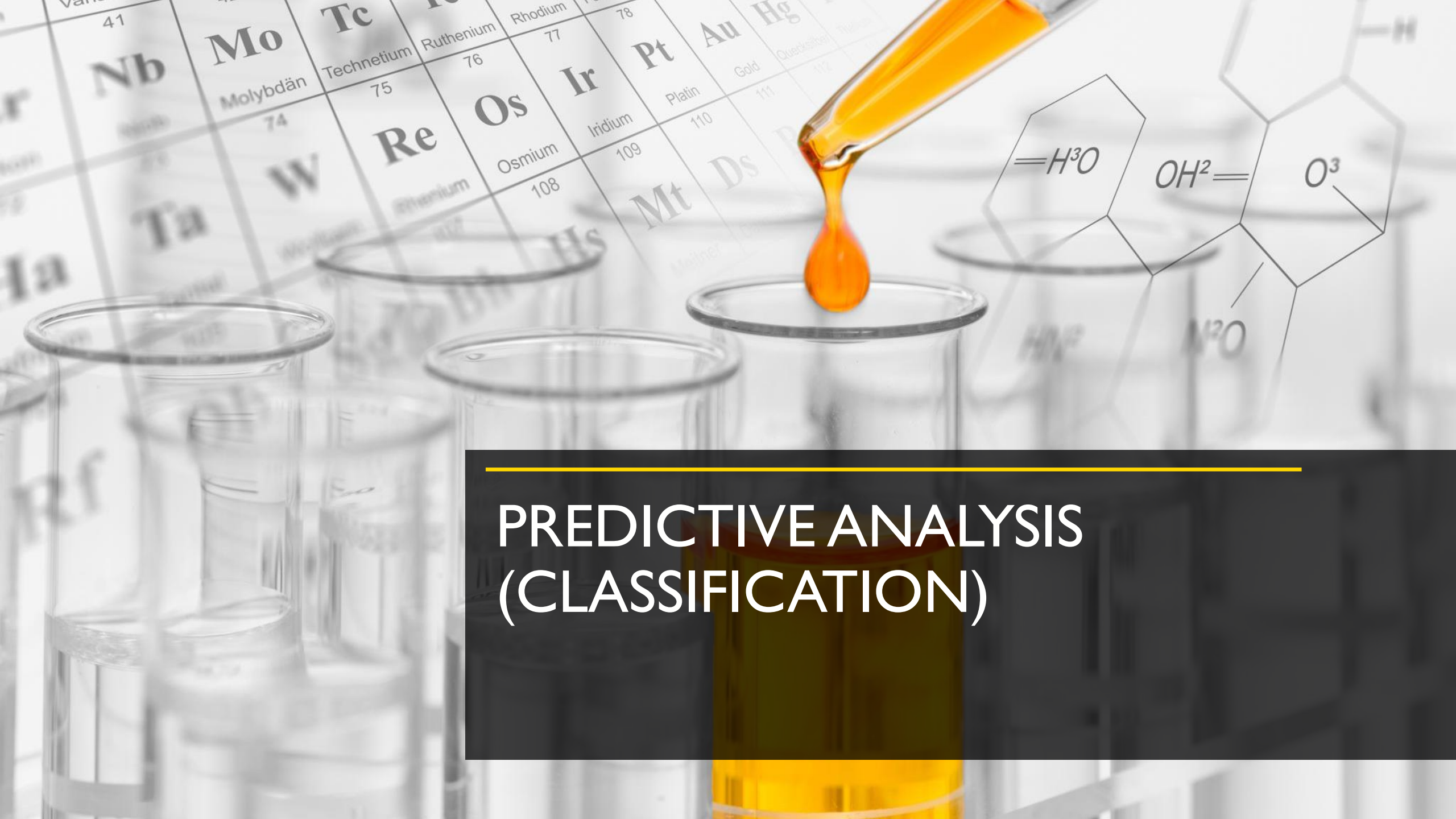
PAYLOAD VS. ORBIT TYPE

- PO, LEO and ISS Orbits have higher probability for successful landings as payload mass increases
- GTO orbits have no correlation between payload mass and landing success
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

LAUNCH SUCCESS YEARLY TREND

- there were no successful landings prior to 2013
- Yearly Trend upward from 2013 till 2017,then a dip, till mid 2018 again a peak till mid 2019 but looks COVID time it get back slide .





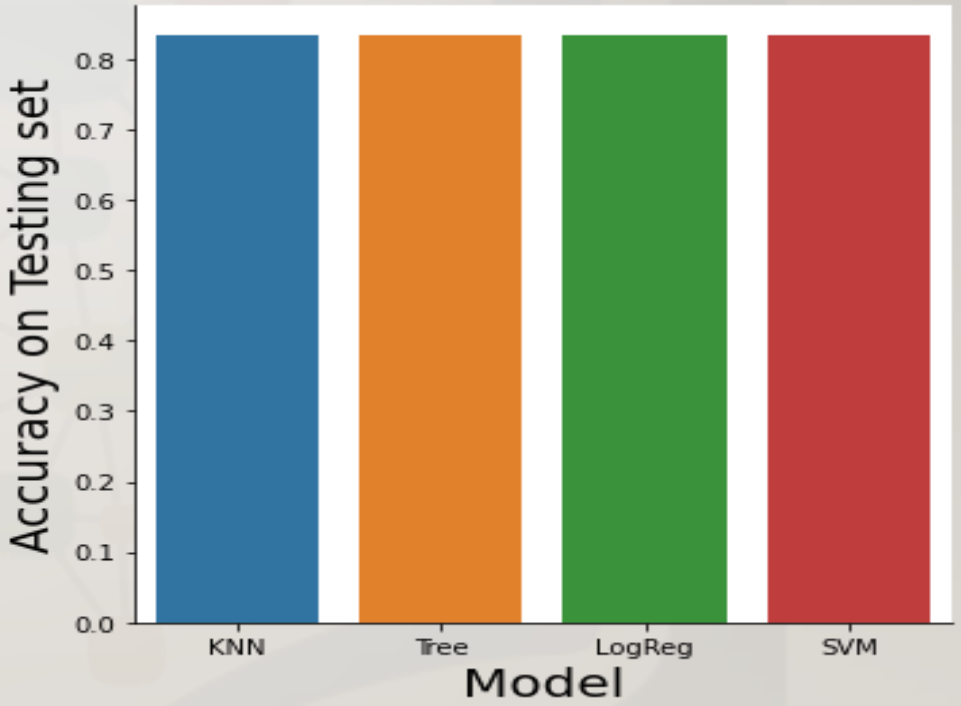
PREDICTIVE ANALYSIS (CLASSIFICATION)

CLASSIFICATION ACCURACY

- Based on the scores of the Test Set, we can not confirm which method performs best
- Same Test Set scores may be due to the small test sample size (18 samples). Bigger sample size can make a difference.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

MODELS SAME INTERPRETATIONS ON ACCURACY, PREDICTING.



EDA WITH SQL

EXPLORATORY DATAANALYSIS WITH SQL DB2

INTEGRATED IN PYTHON WITH SQLALCHEMY

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

```
In [63]: %sql select DISTINCT LAUNCH_SITE from SPACEX1
```

```
 * sqlite:///my_data1.db
Done.
```

Out[63]:

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

ALL LAUNCH SITE NAMES

DISPLAYING THE NAMES OF THE UNIQUE LAUNCH SITES IN THE SPACE MISSION

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

```
In [64]: %sql select * from SPACEX1 where launch_site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

Out[64]:

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

LAUNCH SITE NAMES BEGIN WITH `CCA`

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

```
In [16]: %sql select sum(payload_mass__kg_) as sum from SPACEX1 where customer like 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

sum
45596

TOTAL PAYLOAD MASS

- CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS)

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

In [17]:
```
%sql select avg(payload_mass__kg_) as Average from SPACEX1 where booster_version like 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[17]:

Average
2534.6666666666665

AVERAGE PAYLOAD MASS BY F9 V1.1

FIRST SUCCESSFUL GROUND LANDING DATE

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

Hint:Use min function

In [73]:
```
%sql select min(date) as First_successful_lsanding from SPACEX1 where landing_outcome like 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

Out[73]:

First_successful_lsanding
2015-12-22

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [74]: %sql select booster_version from SPACEX1 where "landing_outcome" = 'Success (drone ship)' and (PAYLOAD_MASS__KG_ between 4000 and
```

```
 * sqlite:///my_data1.db
Done.
```

Out[74]:

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

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

List the total number of successful and failure mission outcomes

In [57]:
```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEX1 GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Out[57]:

MISSION_OUTCOME	Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

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

In [58]:
```
# maxm = %sql select max(payload_mass__kg_) from SPACEX1
# maxv = maxm[0][0]

%sql select booster_version from SPACEX1 where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEX1)
```

```
 * sqlite:///my_data1.db
Done.
```

Out[58]:

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

BOOSTERS CARRIED MAXIMUM PAYLOAD

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

In [37]:
```
%%sql select landing_outcome, count(landing_outcome) as "Total" from SPACEX1
where DATE between '2010-06-04' and '2017-03-20'
group by landing_outcome
order by "Total" desc
```

```
 * sqlite:///my_data1.db
Done.
```

Out[37]:

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

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

CONCLUSION

DECISION ARRIVED

- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- More data should be collected to better determine the best machine learning model and improve the accuracy

WAITING TO LIFT OFF

APPENDIX

GitHub Repository

- https://github.com/Tutul67/Applied-Data-Science-Capstone.git

Special Thanks to All Instructors:

- https://www.coursera.org/professional-certificates/ibm-data-science
- SpaceX data ,Wikipedia