



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

<Name>

<Date>



# Outline

---

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

# Executive Summary

---

- Summary of methodology
  - Launch Data is gathered from the SpaceX REST API including web scrapping from a Wiki page and data wrangling
  - Exploratory data analysis (EDA) using visualization and SQL including interactive visual analytics (Folium and Plotly Dash)
  - Predictive analysis using ML classification models including model building, tuning and performance evaluation of the classification models
- Summary of all results
  - A Decision Tree classifier has the highest accuracy of the tested models (88.9 accuracy on the test data)

# Introduction

---

In this capstone project for *the Applied Data Science Capstone* module, we will predict if the Falcon 9 first stage will land successfully. This has a direct application on determining the cost of a launch by knowing if the first stage will land

All the process is detailed from collecting the data, wrangling, analysing and machine learning building to final conclusions



Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - SpaceX Launch Data is gathered from the SpaceX REST API
  - Web scrapping from a Wiki page is performed to collect Falcon 9 historical launch records
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models



# Data Collection

---

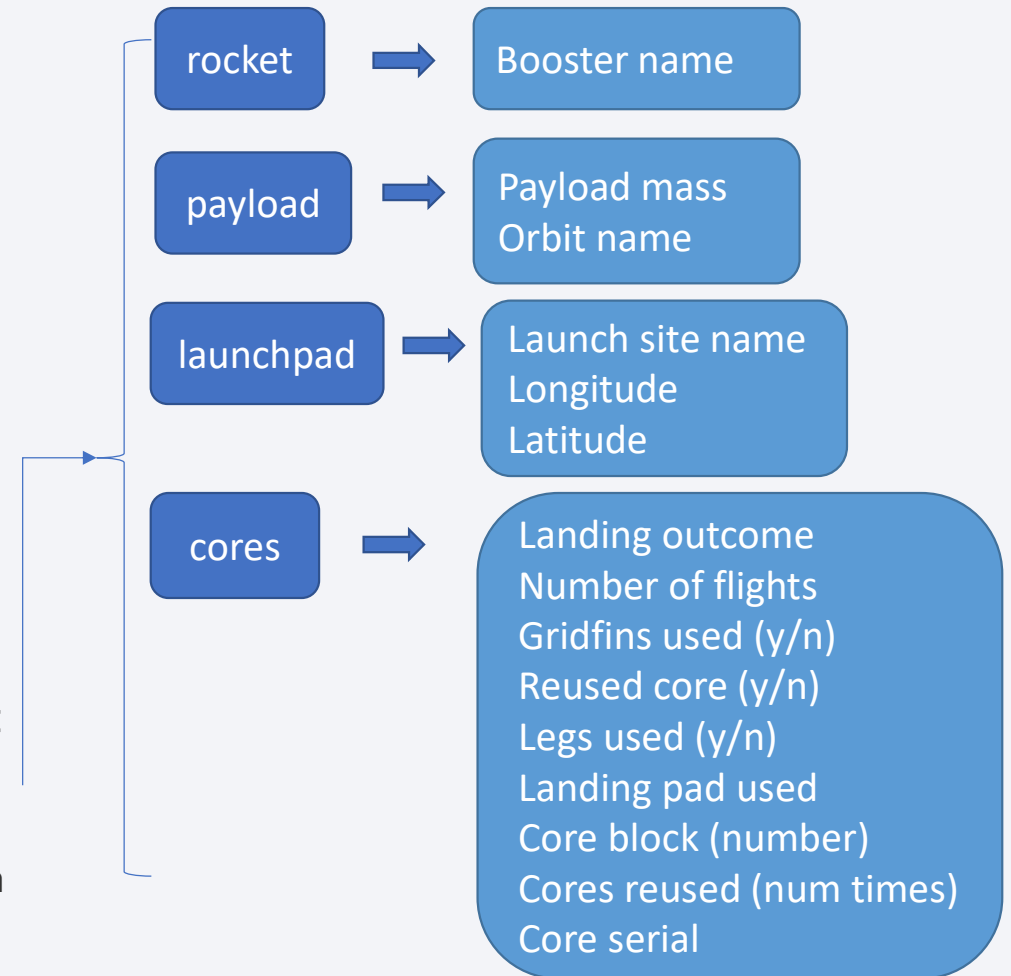
GitHub Link [here](#)

- SpaceX Launch Data is gathered from the SpaceX REST API
  - Data about launches: rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Wrangling Data using an API to convert data (json) into a pandas Dataframe
  - `Data = pd.json_normalize(response.json())`
- Web scraping is performed to obtain Falcon 9 Launch data
  - Python BeautifulSoup package is used for web scraping HTML tables (wiki pages) with Falcon 9 launch records.

# Data Collection – SpaceX API

GitHub Link [here](#)

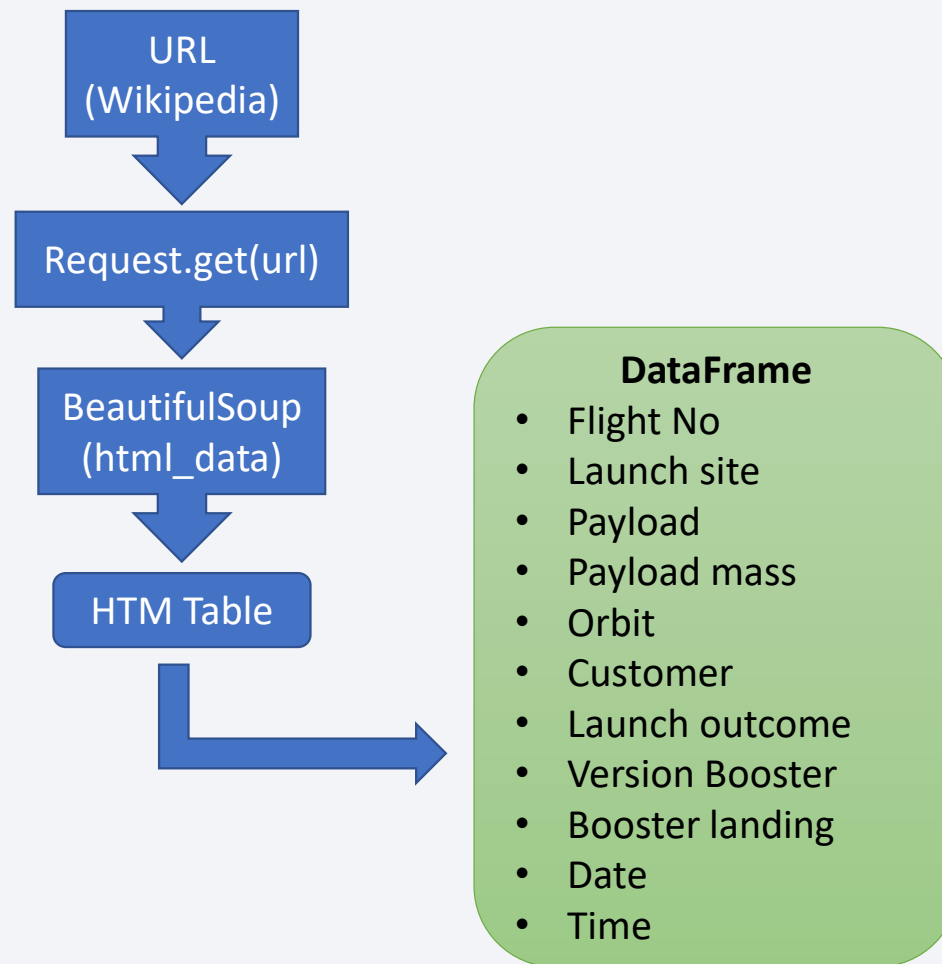
- Helper functions are used to extract API information using identification numbers
- Data is requested from the URL (endpoint):
  - <https://api.spacexdata.com/v4/launches/past>
- Data is requested and parsed using the GET request (*response = requests.get(url)*)
  - Response content is decoded as Json and turned into pandas DF (*.json\_normalize()*)
- Only Falcon 9 launches data is kept
- Data wrangling using an API is also used to target targeting another endpoint to gather specific data for each ID number
  - the mean of PayloadMass is used to replace np.nan values in the data (*replace()* function)





# Data Collection - Scraping

GitHub Link [here](#)



- Web scrapping from a Wiki page is performed using Python's BeautifulSoup
  - to collect **Falcon 9** historical launch records
  - Table is parsed and converted into a pandas DF
- [https://en.wikipedia.org/w/index.php?title=List\\_of\\_Falcon\\_9\\_and\\_Falcon\\_Heavy\\_launches&oldid=1027686922](https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

# Data Wrangling (EDA)

GitHub Link [here](#)

- There are 17 different attributes in the Data Set

- Numerical and categorical columns

|                |         |             |         |
|----------------|---------|-------------|---------|
| FlightNumber   | int64   | Reused      | bool    |
| Date           | object  | Legs        | bool    |
| BoosterVersion | object  | LandingPad  | object  |
| PayloadMass    | float64 | Block       | float64 |
| Orbit          | object  | ReusedCount | int64   |
| LaunchSite     | object  | Serial      | object  |
| Flights        | int64   | Longitude   | float64 |
| GridFins       | bool    | Latitude    | float64 |

- Missing values are identified for each attribute
- There are 3 types of launches for each site with different counts and 11 types of orbits

|              |    |             |    |
|--------------|----|-------------|----|
| CCAFS SLC 40 | 55 | True ASDS   | 41 |
| KSC LC 39A   | 22 | None None   | 19 |
| VAFB SLC 4E  | 13 | True RTLS   | 14 |
|              |    | False ASDS  | 6  |
|              |    | True Ocean  | 5  |
|              |    | False Ocean | 2  |
|              |    | None ASDS   | 2  |
|              |    | False RTLS  | 1  |

- The Outcome indicates if the first stage successfully landed. There are 8 of them

|             |    |
|-------------|----|
| True ASDS   | 41 |
| None None   | 19 |
| True RTLS   | 14 |
| False ASDS  | 6  |
| True Ocean  | 5  |
| False Ocean | 2  |
| None ASDS   | 2  |
| False RTLS  | 1  |

- The Landing Outcomes are converted to Classes
  - 0: bad outcome, the booster did not land.
  - 1: good outcome, the booster did land.
- This is the classification variable that represents the outcome of each launch (66.6% of success rate)

# EDA with Data Visualization

---

GitHub Link [here](#)

- Following charts were plotted
  - FlightNumber vs. PayloadMass
  - FlightNumber vs. LaunchSite
  - Payload mass vs. LaunchSite
  - Success rate vs. Orbit type
  - FlightNumber vs. Orbit type
  - Payload mass vs. Orbit type
  - Launch success yearly trend
- These are done to evaluate relationships between variables and with between variables and the outcome

# EDA with SQL

---

GitHub Link [here](#)

- Following information was retrieved using SQL queries
  - Unique launch sites names in the space mission
  - 5 records where launch sites name begin with the string 'CCA'
  - Total payload mass carried by boosters NASA (CRS)
  - Average payload mass carried by booster F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - Booster versions which have carried the maximum payload mass

# Build an Interactive Map with Folium

GitHub Link [here](#)

- For each Launch site, a circle marker is added into the map (*folium.Circle* and *folium.Marker*)
  - All launch sites are very close in proximity to the coast and relatively close to the Equator line (CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A being the closest in Florida)
- Markers are classified for all launch records:
  - Successful launch (class=1), a green marker is used; failed launch (class=0), a red marker is used
- For each launch result, a *folium.Marker* to *marker\_cluster* is added
  - Marker clusters are used to simplify a map containing many markers with the same coordinates
- Proximities of launch sites are explored and analyzed:
  - *MousePosition* is added on the map to get coordinates for a mouse over a point on the map
  - *PolyLine* are added between a launch site to the selected points of interests (city, railway, highway)
  - Markers with distances are also added
  - Launch sites in Florida are close to highways, railways and coastlines but further apart from a major city (70 km from Orlando)

# Build a Dashboard with Plotly Dash

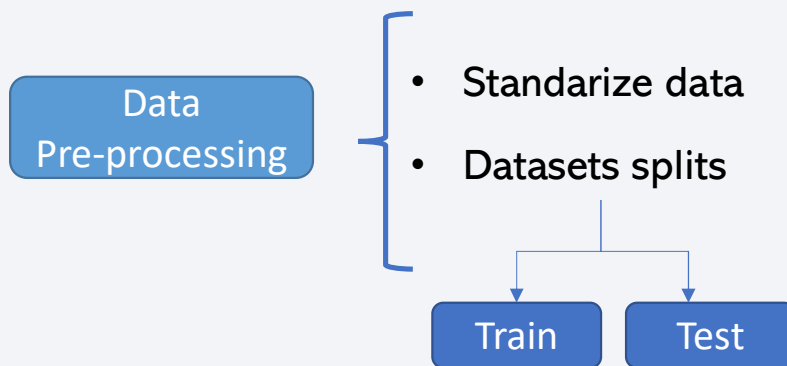
GitHub Link [here](#)

- A dashboard completed so it can be used to analyze SpaceX launch data
  - It contains input components which allow to interact with a pie chart and a scatter point chart
- Launch Site Drop-down Input Component
  - To select one specific site and check its detailed success rate (class=0 vs. class=1)
- A callback function to render success-pie-chart based on selected site dropdown
  - To get the selected launch site from site-dropdown and render a pie chart visualizing launch success count
- Add a Range Slider to Select Payload
  - To easily select different payload range and see if some visual patterns can be identified
- Add a callback function to render the success-payload-scatter-chart scatter plot
  - To observe how payload may be correlated with mission outcomes for selected site(s)

# Predictive Analysis (Classification)

GitHub Link [here](#)

- For the Predictive Analysis, a ML (machine learning) pipeline is built
  - The goal is to predict if the first stage of the Falcon 9 lands successfully
  - Scikit-learn Python library is used to pre-process the data



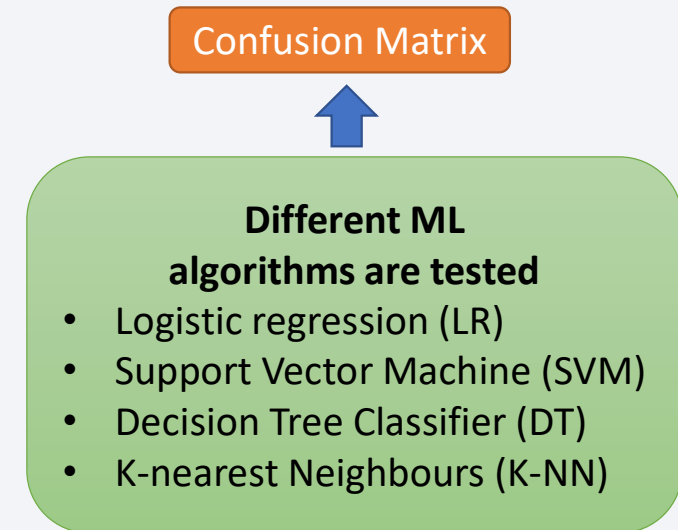
- The model is trained and a Grid Search is performed



We find the best hyperparameters



With these parameters we determine the model with the best accuracy using the test set





# Results

---

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

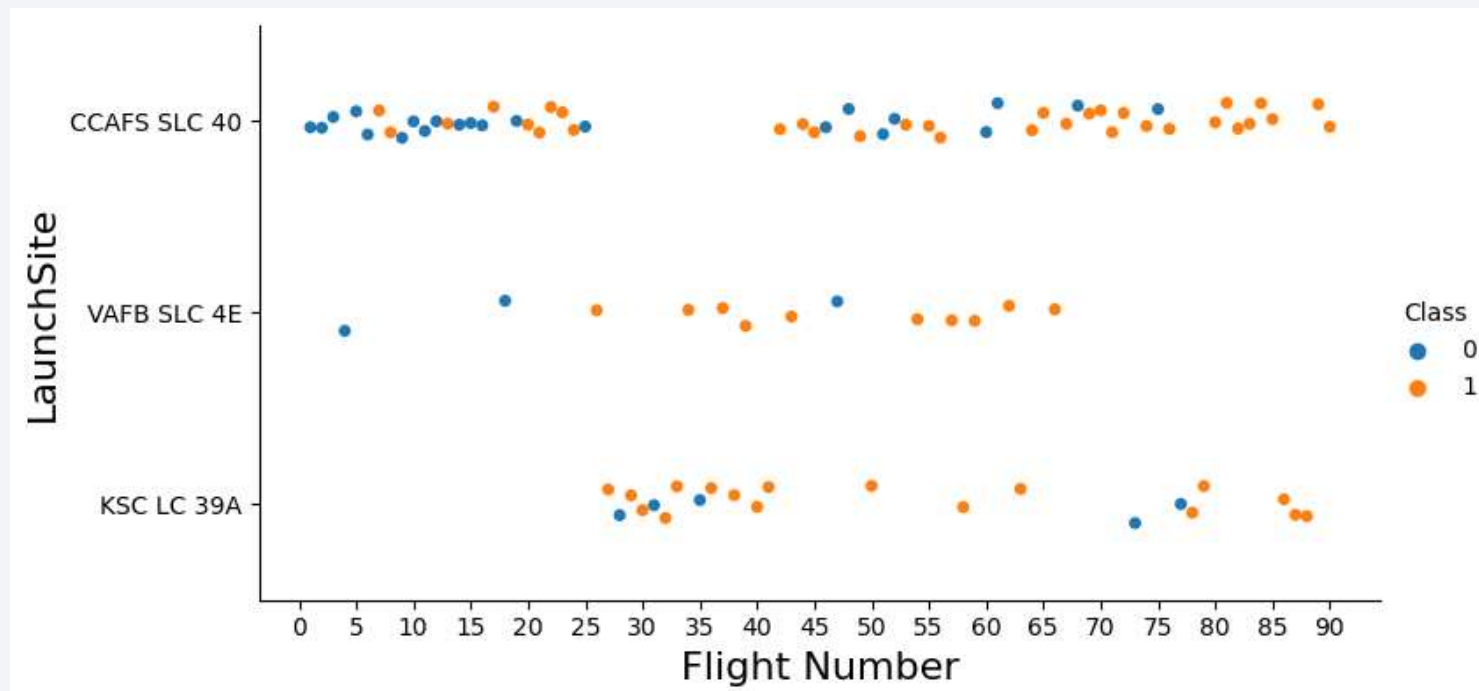


Section 2

# Insights drawn from EDA

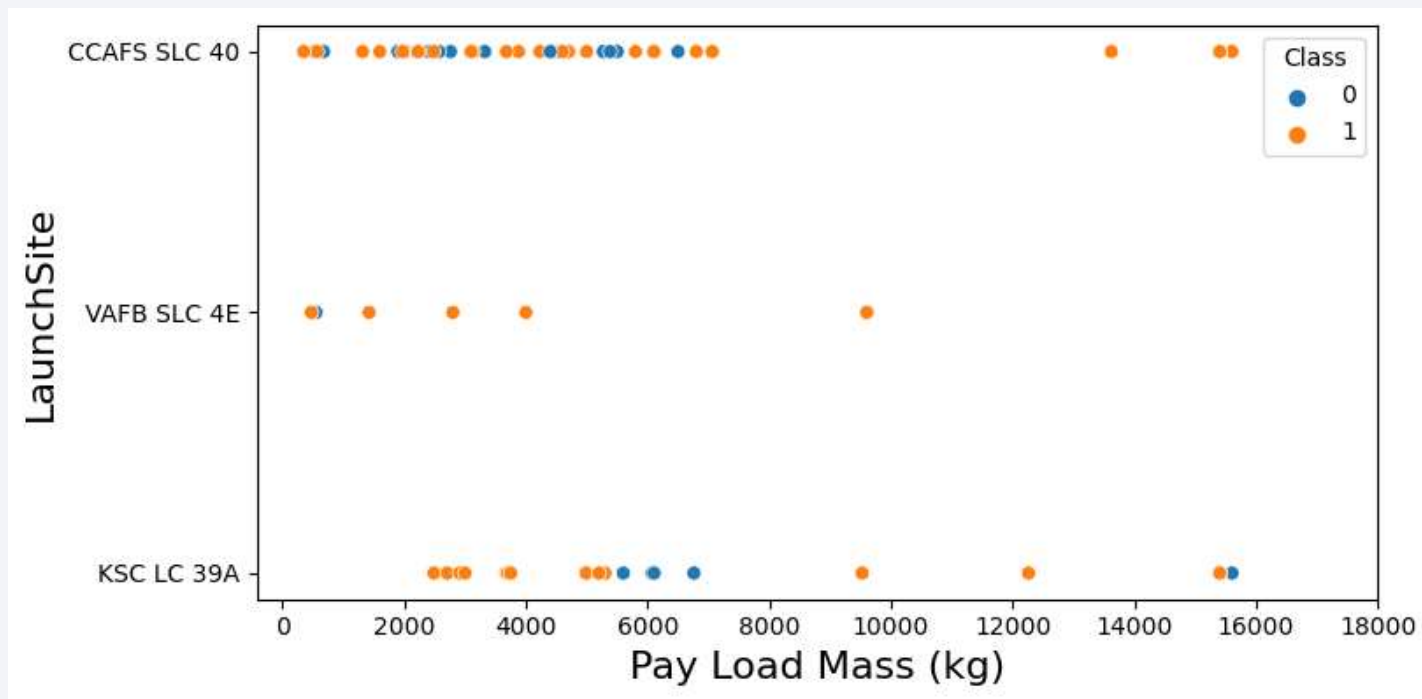
# Flight Number vs. Launch Site

- For the three launch sites, as Flight Number increases, the success rate also increases
- VAFB-SLC-4 has the highest success rate with the smallest number of flights



# Payload vs. Launch Site

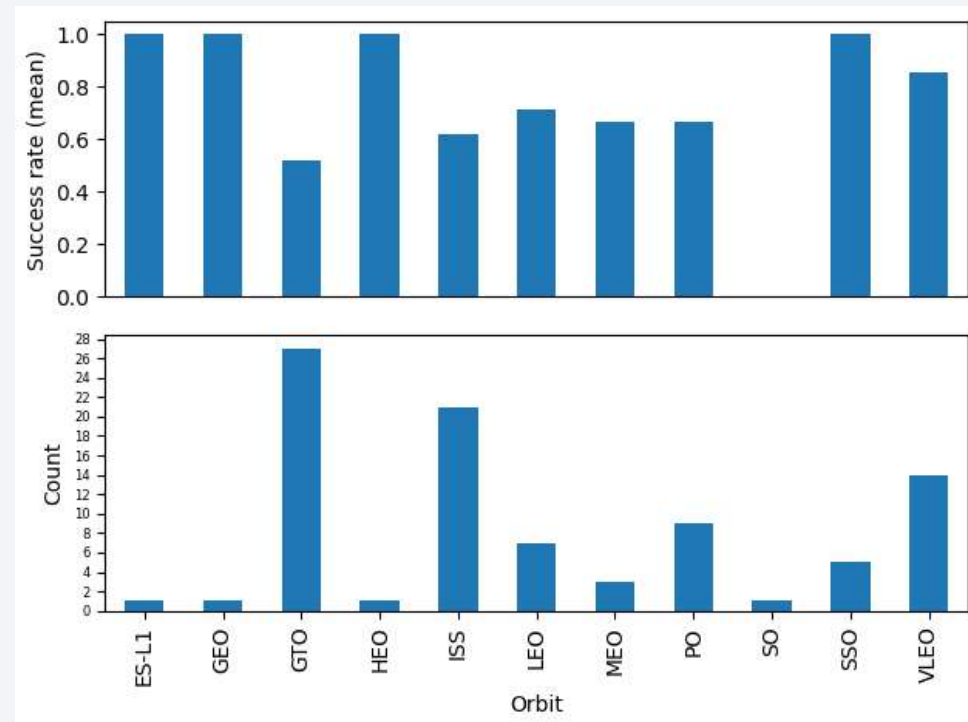
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass ( greater than 10000 Kg)
  - 96% and 86% of launches are below the 10,000 Kg mark for CCAFS-SLC and VAFB-SLC respectively





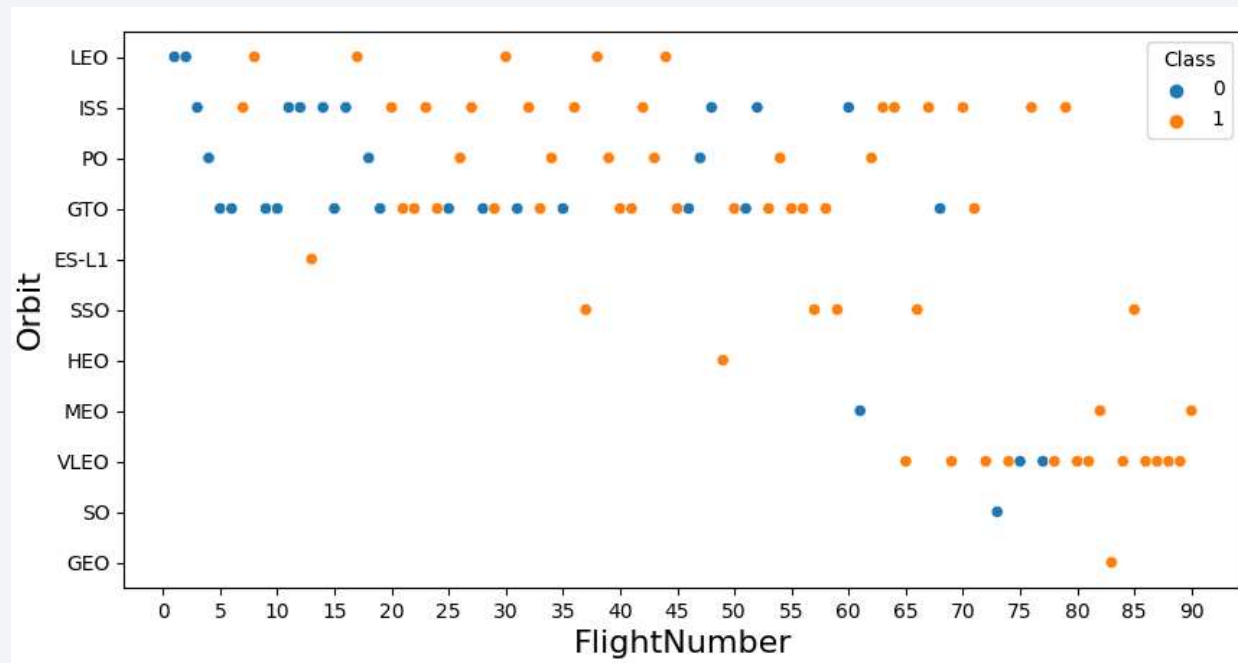
# Success Rate vs. Orbit Type

- SSO has 100% success rate, with 5 launches (count)
  - Three orbits (ES-L1, GEO, HEO) have a 100% success rate, but with only one launch
- Orbits with highest number of launches are GTO and ISS (27 and 21), with 52% and 62% success rates, respectively



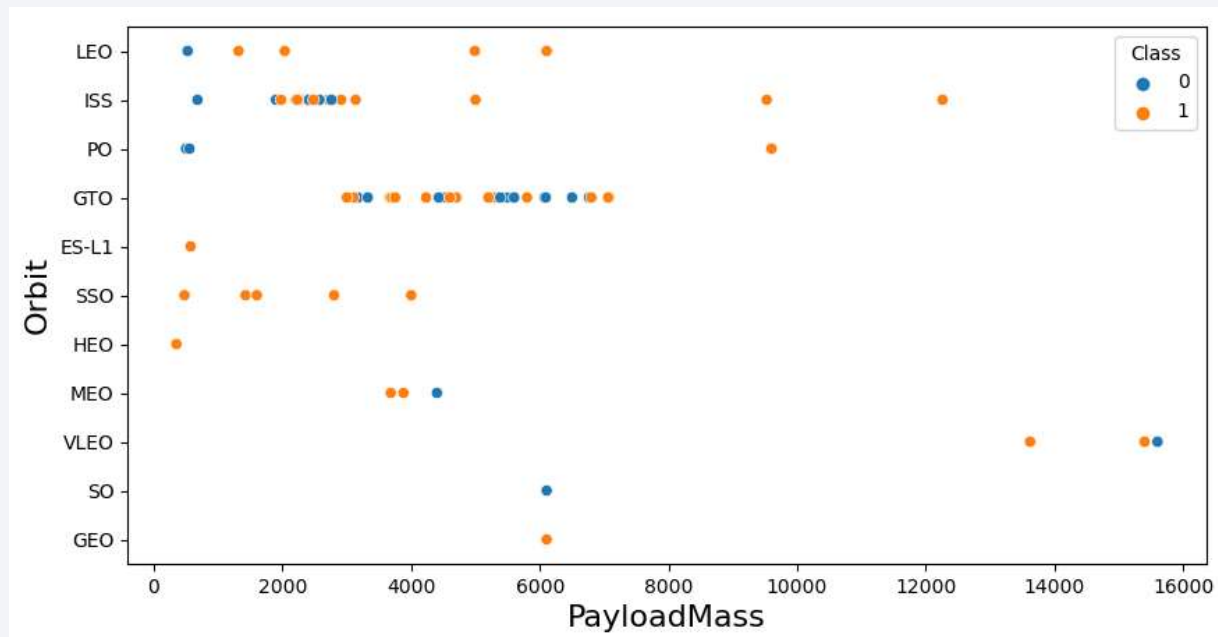
# Flight Number vs. Orbit Type

- For LEO orbit the success appears related to the number of flights;
- No clear relationship between flight number and other orbits



# Payload vs. Orbit Type

- LEO and ISS orbits show increasing number of success landing with heavier payloads

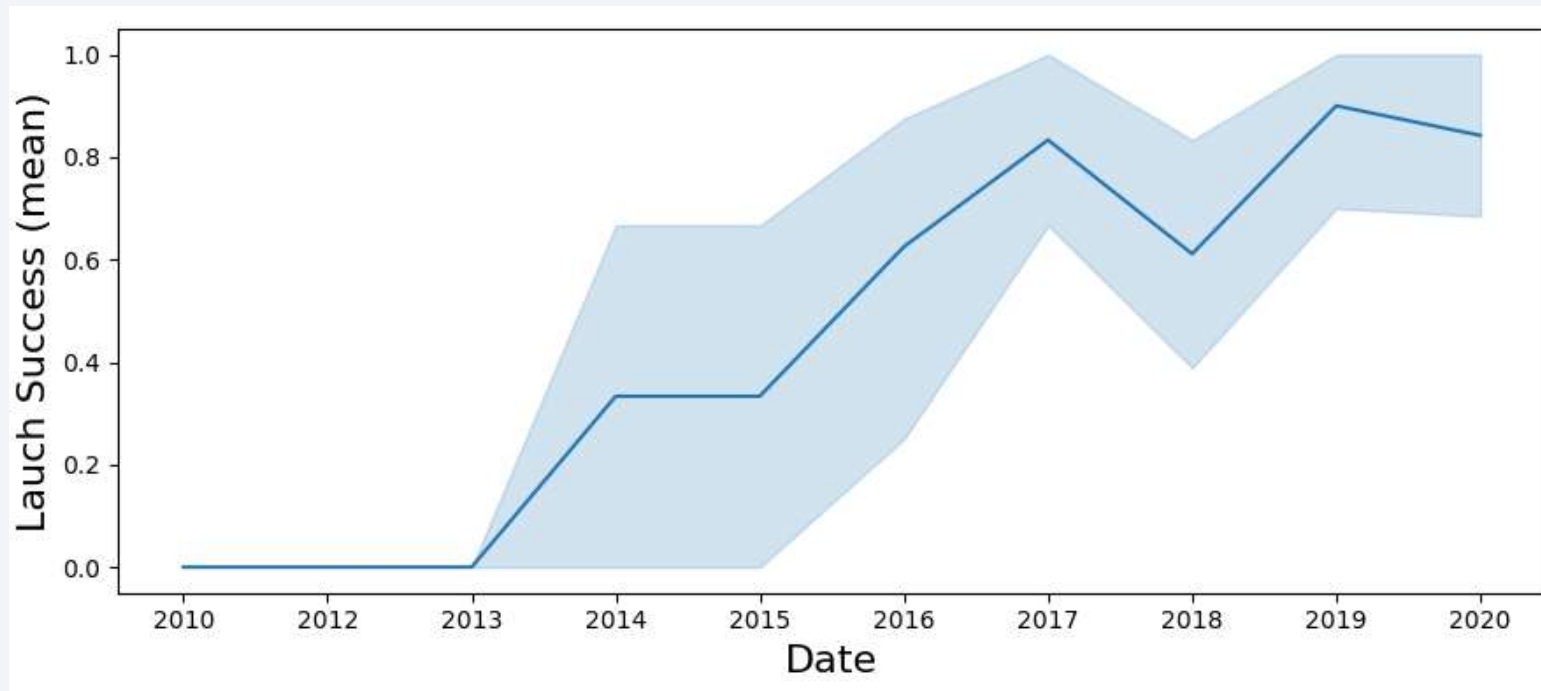




# Launch Success Yearly Trend

---

- The success rate has been increasing since 2013
  - Except between 2017 and 2018 and between 2019 and 2020



# All Launch Site Names

---

- There are 4 different Launch Sites
  - CCAFS SLC-40 has the higher count, VAFB the lowest

```
%%sql
SELECT "Launch_Site", COUNT("Launch_Site") AS "Count"
FROM SPACEXTBL
GROUP BY "Launch_Site"

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

| Launch_Site  | Count |
|--------------|-------|
| CCAFS LC-40  | 26    |
| CCAFS SLC-40 | 34    |
| KSC LC-39A   | 25    |
| VAFB SLC-4E  | 16    |

# Launch Site Names Begin with 'CCA'

- 5 records where Launch sites begin with the string “CCA”

```
%%sql  
  
SELECT *  
FROM SPACEXTBL  
WHERE "Launch_Site" LIKE 'CCA%'  
LIMIT 5
```

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

| Date                | Time (UTC) | Booster_Version | Launch_Site | Payload   | PAYLOAD_MASS_KG_ | Orbit     | Customer        | Mission_Outcome | Landing_Outcome     |
|---------------------|------------|-----------------|-------------|---|------------------|-----------|-----------------|-----------------|---------------------|
| 2010-06-04 00:00:00 | 18:45:00   | F9 v1.0 B0003   | CCAFS LC-40 | Dragon Spacecraft Qualification Unit                          | 0                | LEO       | SpaceX          | Success         | Failure (parachute) |
| 2010-12-08 00:00:00 | 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 00:00:00 | 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 00:00:00 | 00:35:00   | F9 v1.0 B0006   | CCAFS LC-40 | SpaceX CRS-1  | 500              | LEO (ISS) | NASA (CRS)      | Success         | No attempt          |
| 2013-03-01 00:00:00 | 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 total Payload Mass carried by boosters launched by NASA (CRS) is 45,596 Kg

```
%%sql
SELECT SUM("PAYLOAD_MASS_KG_") AS "Total Mass"
FROM SPACEXTBL
WHERE "Customer" IS 'NASA (CRS)'

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

| Total Mass |
|------------|
| 45596      |

# Average Payload Mass by F9 v1.1

---

- Average Payload Mass carried by booster version F9 v1.1 is 2,534.67 Kg

```
%%sql
SELECT AVG("PAYLOAD_MASS_KG_") AS "Avg Mass"
FROM SPACEXTBL
WHERE "Booster_Version" LIKE '%F9 v1.1%'

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

| Avg Mass           |
|--------------------|
| 2534.6666666666665 |

# First Successful Ground Landing Date

---

- The first successful landing outcome in ground pad was achieved in 22/12/2015

```
%%sql
SELECT MIN(Date) FROM SPACEXTBL
WHERE "Landing_Outcome" = "Success (ground pad)"
* sqlite:///my_data1.db
Done.
```

| MIN(Date)           |
|---------------------|
| 2015-12-22 00:00:00 |

## Successful Drone Ship Landing with Payload between 4000 and 6000

---

- These are the names of the boosters which have success in drone ship and have payload mass between 4,000 and 5,000 Kg

```
%%sql
SELECT Booster_Version, PAYLOAD_MASS_KG_ FROM SPACEXTBL
WHERE "Landing_Outcome" = "Success (drone ship)"
AND PAYLOAD_MASS_KG_ BETWEEN 4000 AND 6000
```

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

| Booster_Version | PAYLOAD_MASS_KG_ |
|-----------------|------------------|
| F9 FT B1022     | 4696             |
| F9 FT B1026     | 4600             |
| F9 FT B1021.2   | 5300             |
| F9 FT B1031.2   | 5200             |



# Total Number of Successful and Failure Mission Outcomes

---

- This is the total number of successful and failure mission outcomes
  - 100 success (3 types) and 1 failure

```
%%sql  
  
SELECT "Mission_Outcome", COUNT("Mission_Outcome") AS "Count"  
FROM SPACEXTBL  
GROUP BY "Mission_Outcome"
```

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

| Mission_Outcome                  | Count |
|----------------------------------|-------|
| Failure (in flight)              | 1     |
| Success                          | 98    |
| Success                          | 1     |
| Success (payload status unclear) | 1     |

# Boosters Carried Maximum Payload

---

- These are the booster versions which have carried the maximum payload mass (15,600 Kg)

```
%%sql
SELECT DISTINCT(Booster_Version), PAYLOAD_MASS_KG_ FROM SPACEXTBL
WHERE PAYLOAD_MASS_KG_ = (SELECT MAX(PAYLOAD_MASS_KG_) FROM SPACEXTBL)

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

| 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

---

- These are the records for failure landing outcomes in drone ship versions for 2015
  - Displaying month and year

```
%%sql  
  
SELECT "Booster_Version", "Launch_Site", "Landing_Outcome", substr(Date, 6, 2) AS "Month", substr(Date,1,4) AS "Year"  
FROM SPACEXTBL  
WHERE "Date" LIKE "%2015%" AND "Landing_Outcome" is "Failure (drone ship)"
```

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

```
Done.
```

| Booster_Version | Launch_Site | Landing_Outcome      | Month | Year |
|-----------------|-------------|----------------------|-------|------|
| F9 v1.1 B1012   | CCAFS LC-40 | Failure (drone ship) | 01    | 2015 |
| F9 v1.1 B1015   | CCAFS LC-40 | Failure (drone ship) | 04    | 2015 |

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

---

- This is the rank of successful landing\_outcomes between the date 04-06-2010 and 20-03-2017
  - in descending order

```
%%sql  
  
SELECT "Landing _Outcome", COUNT("Landing _Outcome") AS "Count"  
FROM SPACEXTBL  
WHERE "Landing _Outcome" LIKE "%Success%" AND "Date" BETWEEN "2010-06-04" AND "2022-03-20"  
GROUP BY "Landing _Outcome"
```

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

```
Done.
```

| Landing _Outcome     | Count |
|----------------------|-------|
| Success              | 38    |
| Success (drone ship) | 14    |
| Success (ground pad) | 9     |

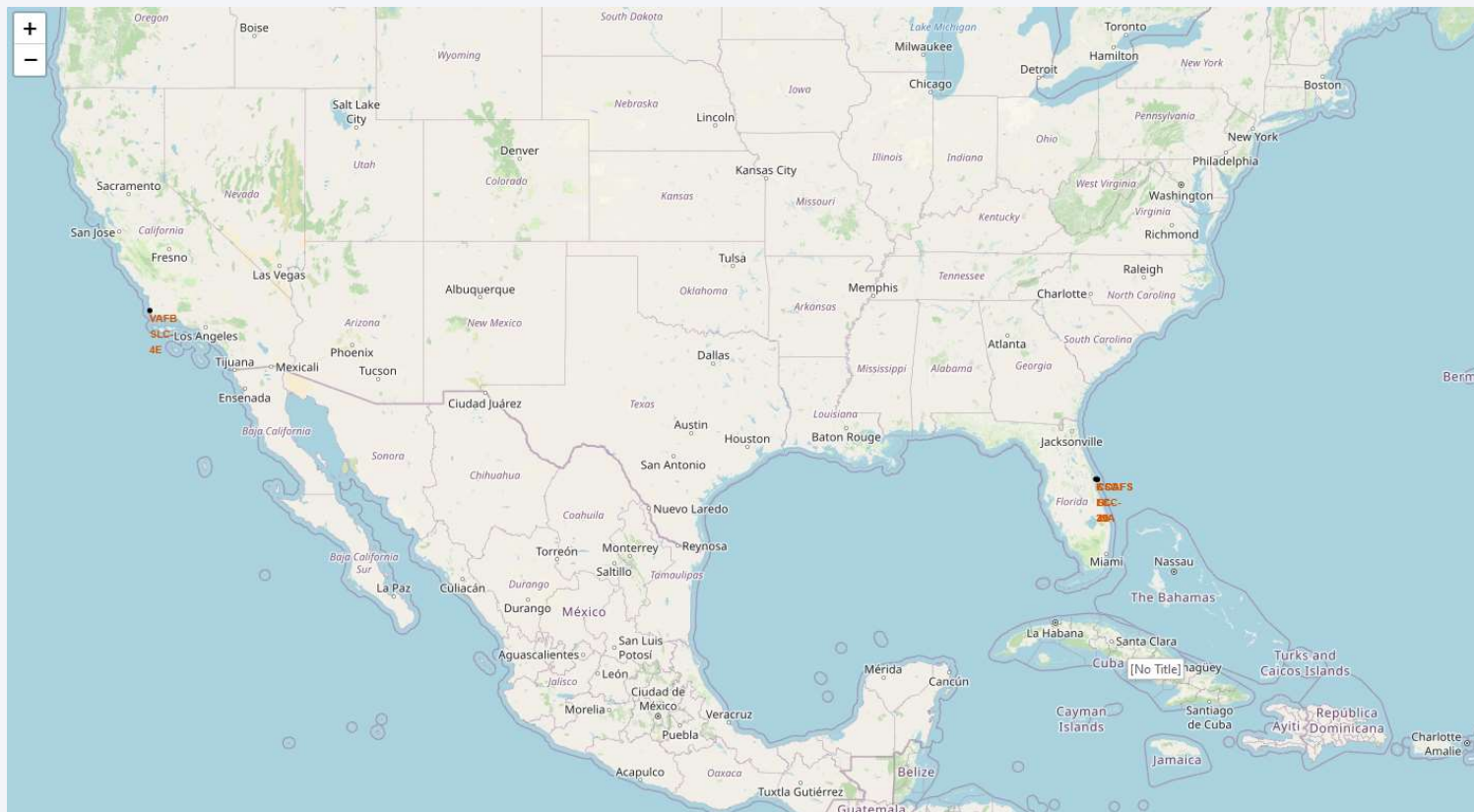
A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is used as a background for the title slide.

Section 3

# Launch Sites Proximities Analysis

# Folium Map Screenshot 1: Launch Sites on the map

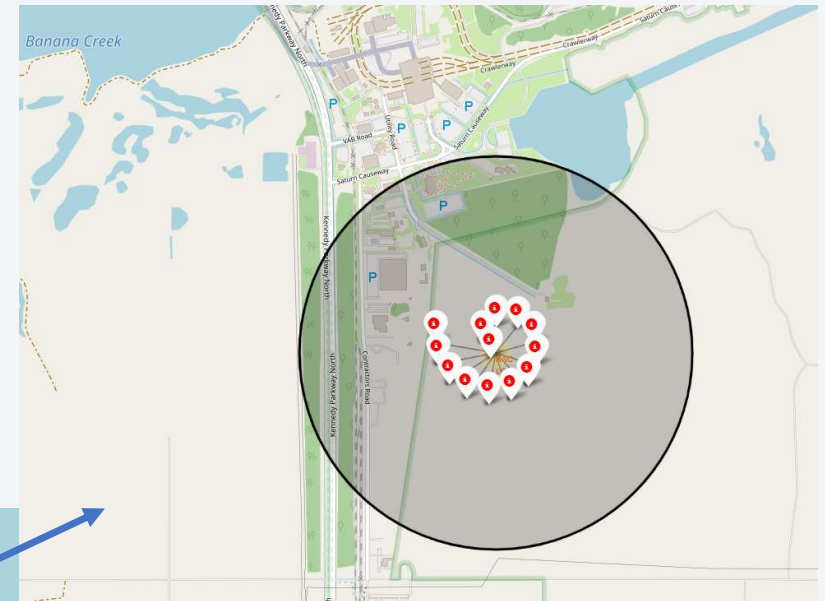
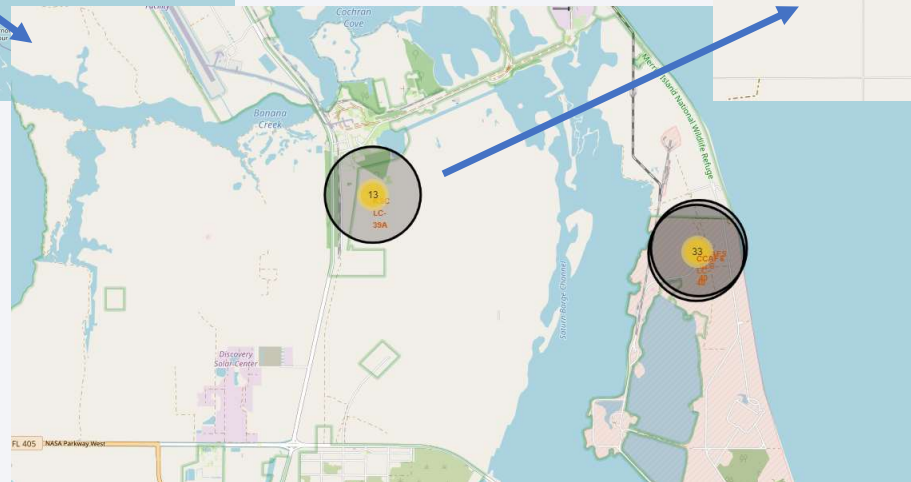
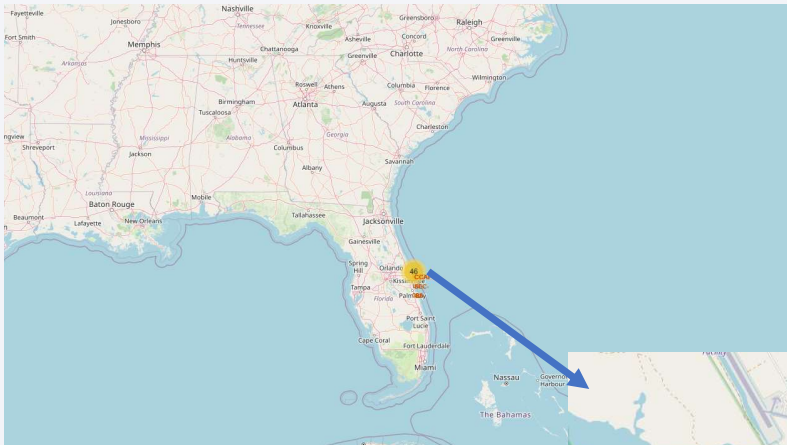
All Launch sites are situated in two main locations:





# Folium Map Screenshot 2: success/failed launches for each site

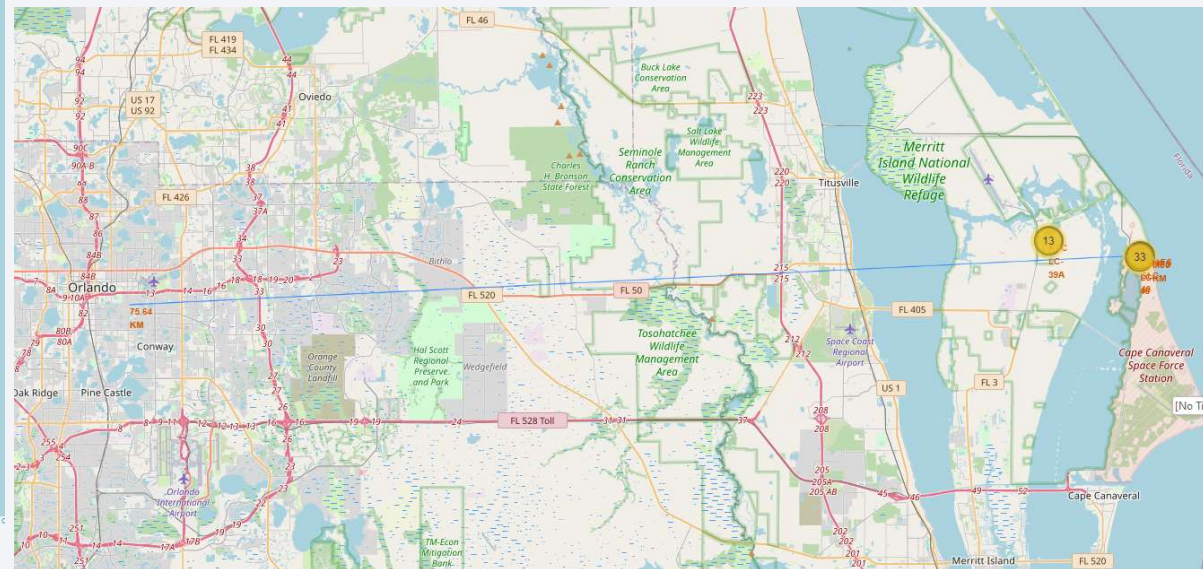
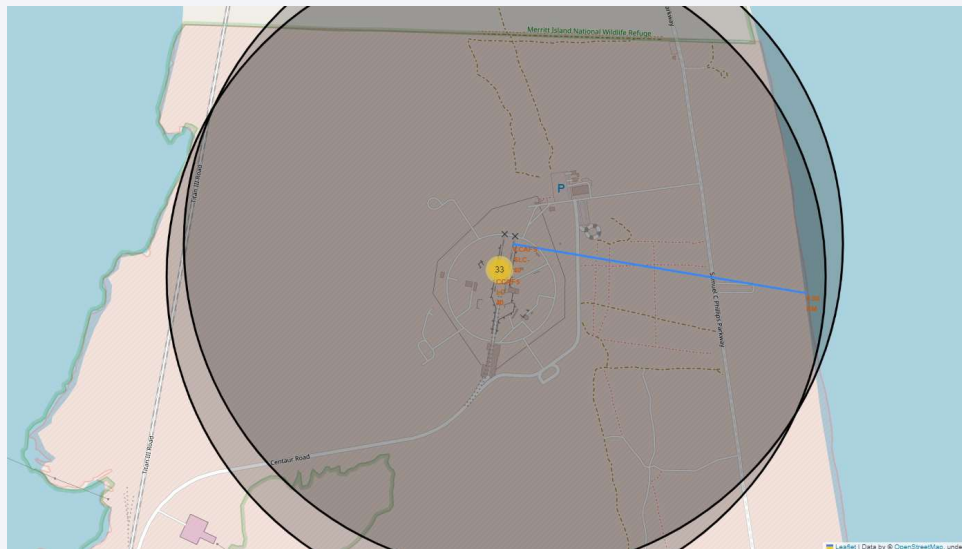
Launch outcomes for each site are marked (red/green) to identify which sites have high success rates





## Folium Map Screenshot 3: distances between a launch site to its proximities

CCAFS LC-40 and CCAFS SLC-40 Launch sites are located 0.9 Km to the coastline and 75Km to a major city (Orlando)

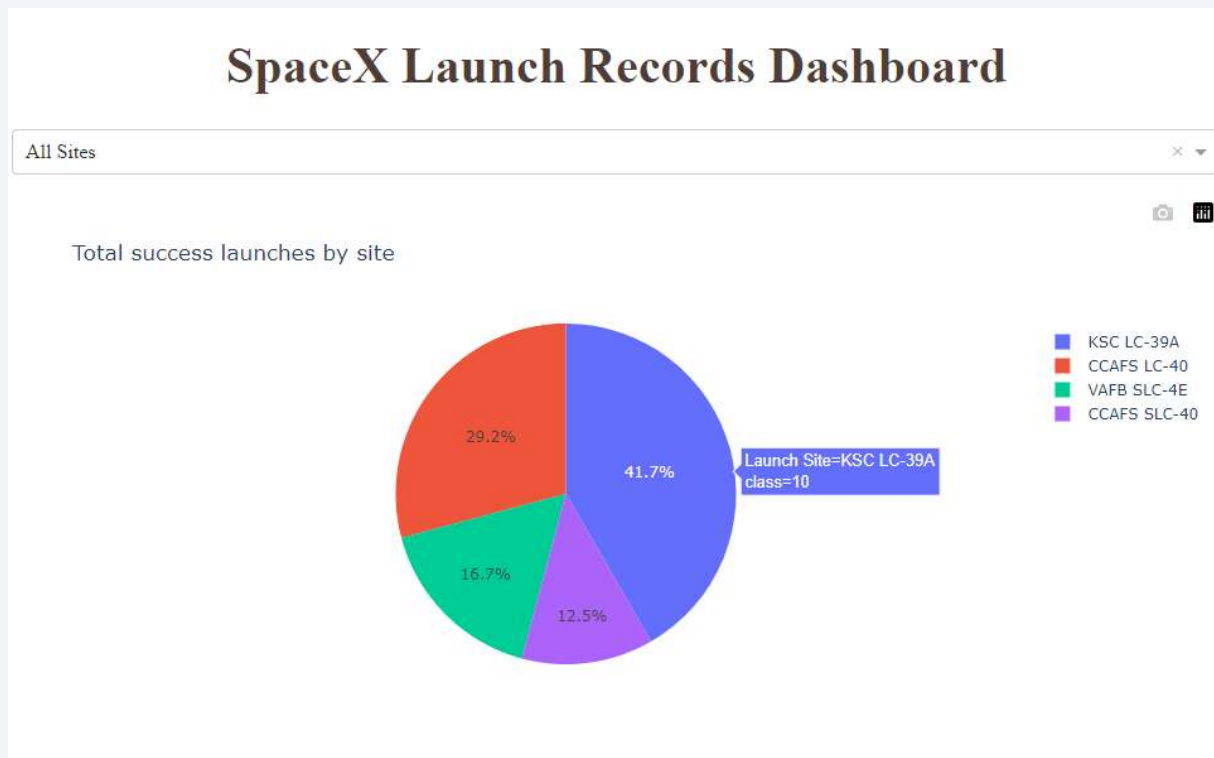




Section 4

# Build a Dashboard with Plotly Dash

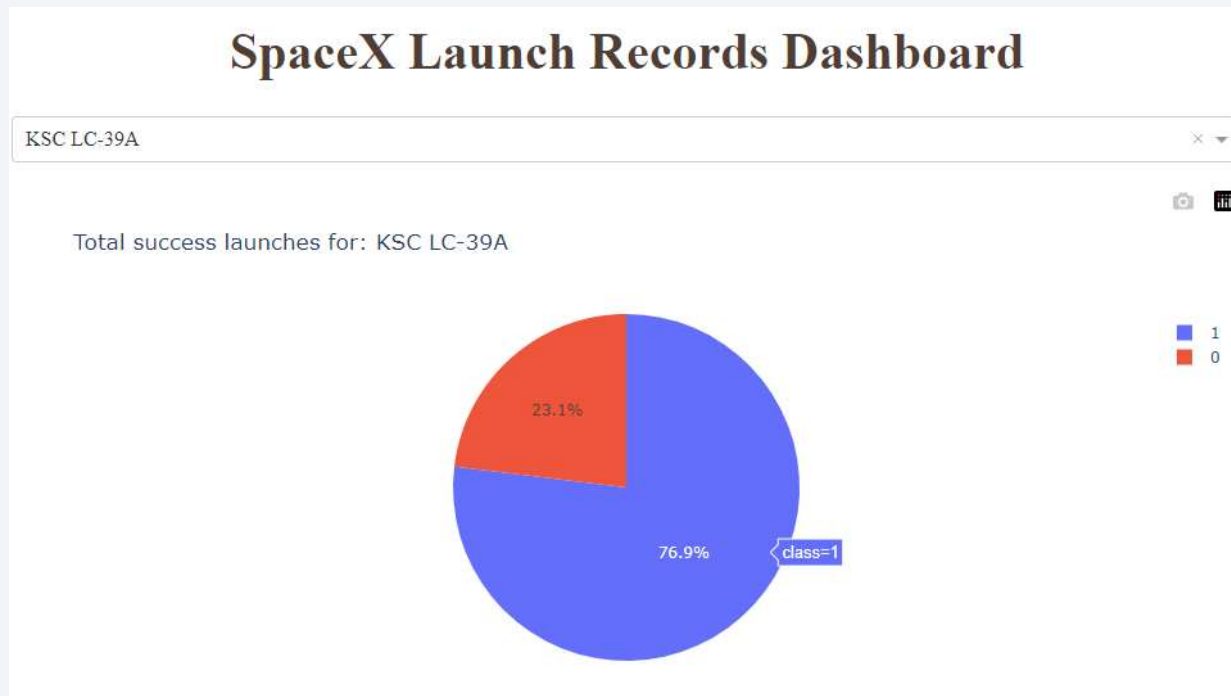
## Dashboard Screenshot 1: launch success count for all sites



KSC LC-39A Launch Site  
has the highest count of  
success launches  
(41.7%)

## Dashboard Screenshot 2: launch site with highest launch success ratio

---

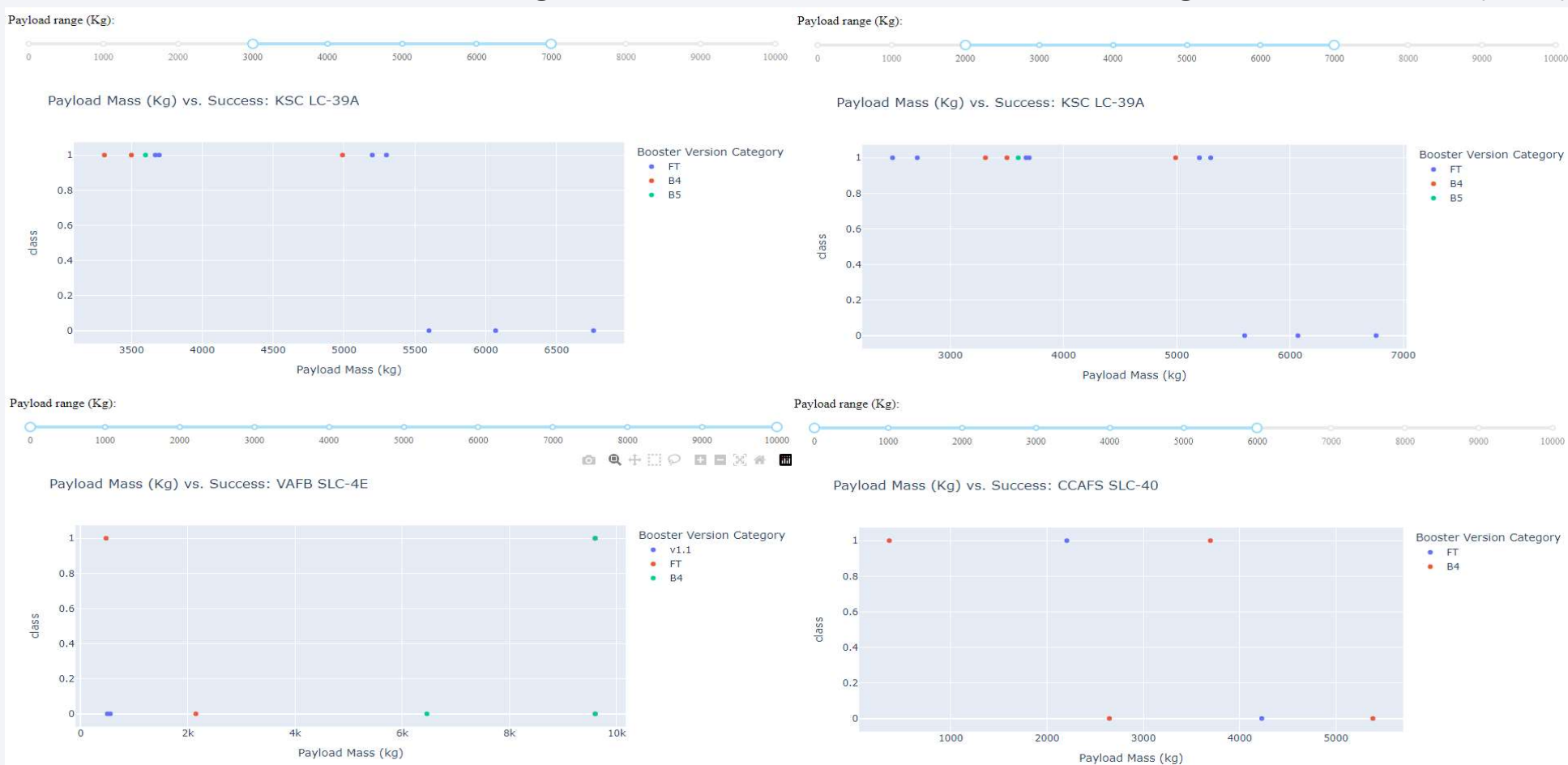


KSC LC-39A Launch Site has the highest launch success ratio (76.9%)



# Dashboard Screenshot 3: Payload vs. Launch Outcome

KSC LC-39A Launch Site has the highest number of successful launches with 10 against 3 unsuccessful (all FT)



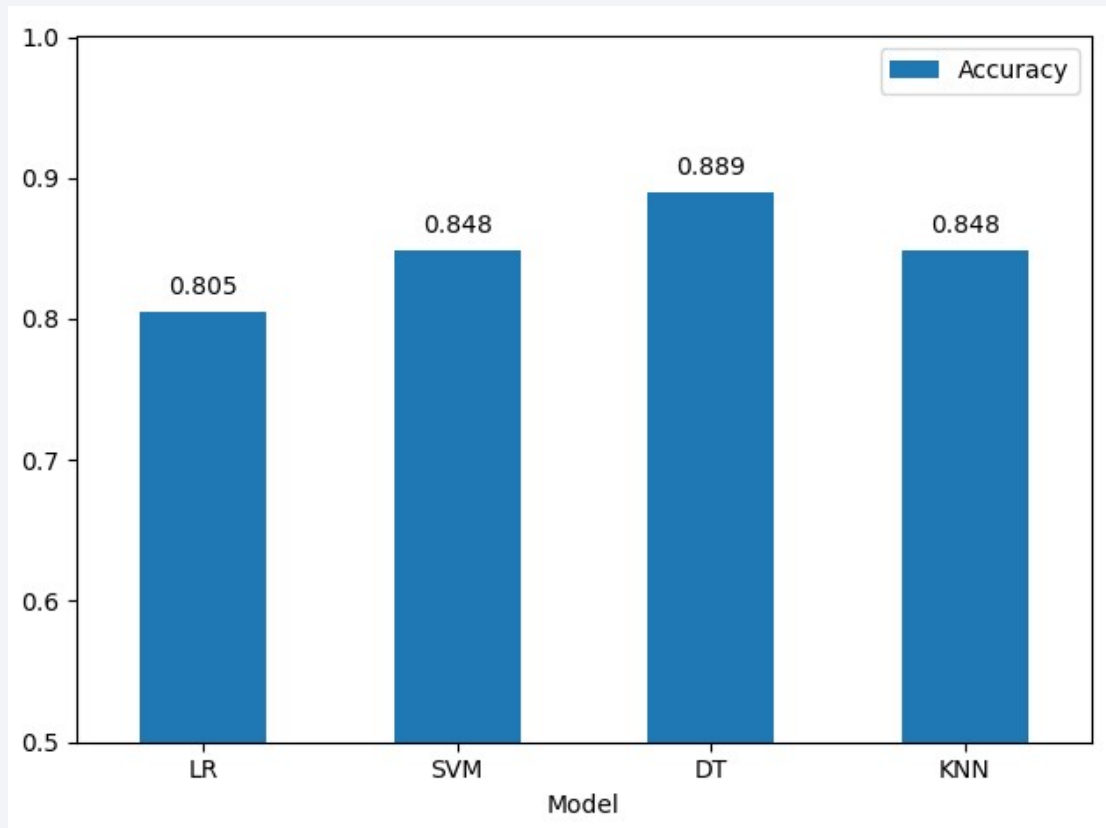


Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

---

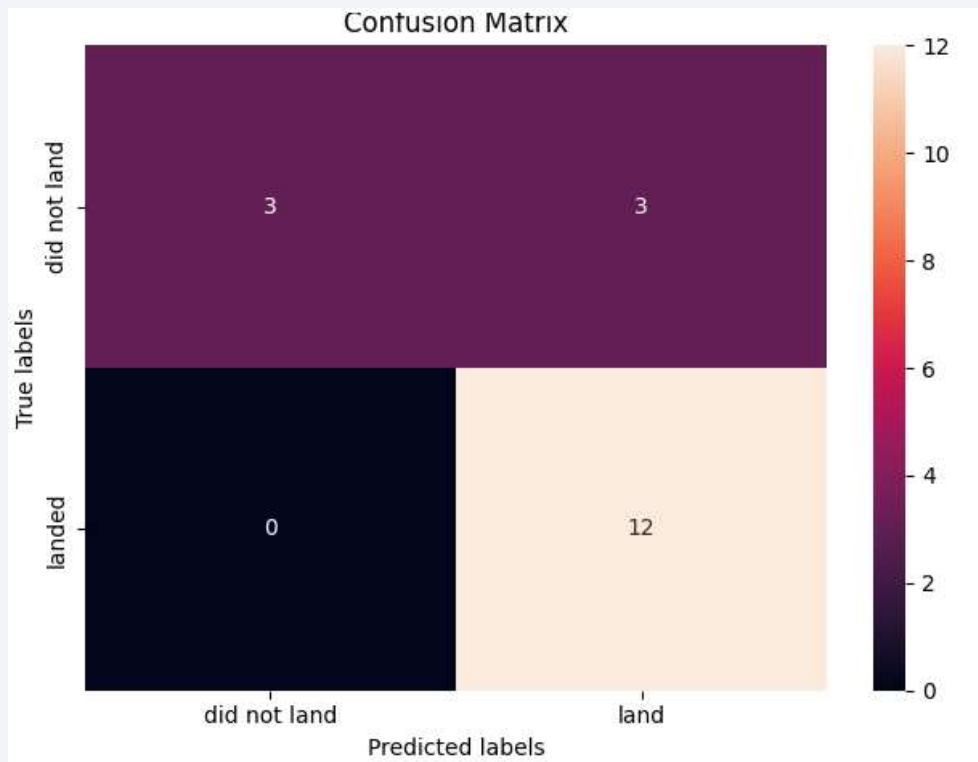


The Decision Tree classifier has the highest accuracy of the models (test data set)



# Confusion Matrix

This is the Confusion Matrix for the Decision Tree classifier:



- From the 18 test samples (test dataset):
  - Out of 6 labelled as “did not land”
    - 3 were classified as “did not land” (true negative)
    - 3 were classified as “landed” (false negative)
  - Out of 12 labelled as “landed”
    - 0 were classified as “did not land” (false positive)
    - 12 were classified as “landed” (true positive)

# Conclusions

---

- Following are the key conclusions from this project
  - The Decision Tree classifier has the highest accuracy of the models with 88.9% on the test data
  - For the three launch sites, as Flight Number increases, the success rate also increases
  - Orbits with highest number of launches are GTO and ISS (27 and 21), with 52% and 62% success rates, respectively
  - KSC LC-39A Launch Site has the highest launch success ratio (76.9%)
  - The success rate on the launches has been increasing since 2013
  - All Launch sites are situated in two main locations
  - Launch sites are located close to the coastline

# Appendix

---

- All Jupiter Notebooks containing code, SQL queries, charts and data sets created during this project are available at Github following the link below
  - [https://github.com/mjbaldomir/Capstone IBMDS](https://github.com/mjbaldomir/Capstone_IBMDS)

Thank you!

