IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

KUEH SEOW TECK

8th December 2021

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data wrangling
- Exploratory data analysis (EDA):
  - Data visualization
  - SQL
- Interactive visual analysis:
  - Folium
  - Plotly Dash
- Predictive analysis using classification models

Summary of all results

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

Introduction

Commercial space age is here and SpaceX is one of the most successful company in providing affordable space travel. This is due to its reusable first stage Falcon 9 rockets which if retrieve successfully can reduce cost of launch to mere 62 million dollars; other competitors cost upwards of 165 million dollars.

The purpose of this project is to predict if Falcon 9 first stage will land successfully. This can help us determine the cost of a launch which will be beneficial to have if we want to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Describe how data was collected
- 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

- Data is collected through to source:
  - SpaceX REST API
  - Falcon 9 and Falcon Heavy Launches Records Wikipedia page (web scrapping)
- The following two slides shows the workflow of how the data collection process is performed.

Data Collection – SpaceX API

- Data collection through SpaceX REST API is performed in three stages:
  - Get request to the SpaceX API
  - Clean the data
  - Data wrangling
- Github URL for reference.

Get Response from API

- Decode the response as Json
- Convert the Json data into Pandas dataframe

Clean up data

- Filter the required columns
- Extract data from the dataframe using custom functions
- Filter data to include only Falcon 9 launches

Data wrangling

- Replace null value for PayloadMass by its average

Data Collection - Scraping

- Web scraping SpaceX Wikipedia page is done in two stages:
  - Getting request from URL and save it as BeautifulSoup object
  - Extract data from BeautifulSoup object
- Github URL for reference

Getting request from URL

- Create a BeautifulSoup object from request

Extract data from BeautifulSoup Object

- Find all the tables
- Create dictionary from column names
- Extract data from tables and append to dictinary
- Convert dictionary to Pandas dataframe

Data Wrangling

- Data wrangling is performed to determine the missing values are filled and data types for each columns are correct.
- EDA is performed to identify patterns in the data and identify training labels
- A new column named ‘Class’ is created to label successful outcomes as 1 and failed outcomes as 0.
- Github URL for reference

Data Wrangling

- Load the SpaceX dataset
- Identify the missing values
- Identify the data types for each columns

Exploratory Data Analysis (EDA)

- Calculate the number of launches on each site
- Calculate the counts for each type of orbit
- Determine the types and its value counts for all the mission outcome

Create Training labels from the outcomes

- Create a set called bad_outcome which contains the labels for failed mission outcome (‘False ASDS’,’False Ocean’,’False RTLS’,’None ASDS’ and None None'
- Create a column named ‘Class’ by assigning successful launches as 1 and failed launches as 0 based on the bad_outcome set

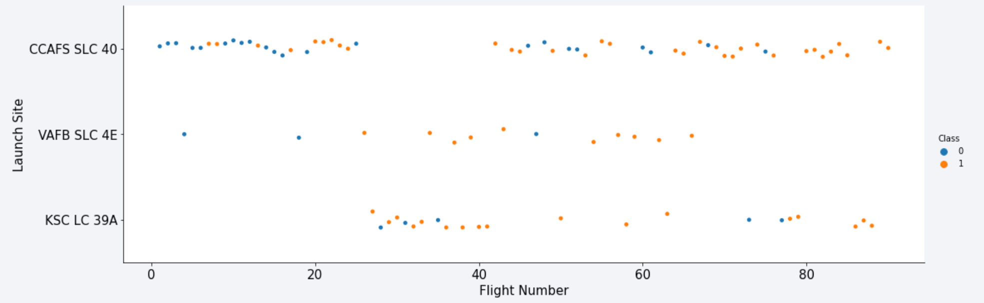
EDA with Data Visualization

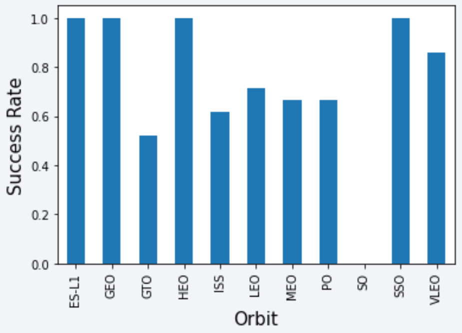
- Three types of graphs are used to perform EDA on the data
    - Scatter plots

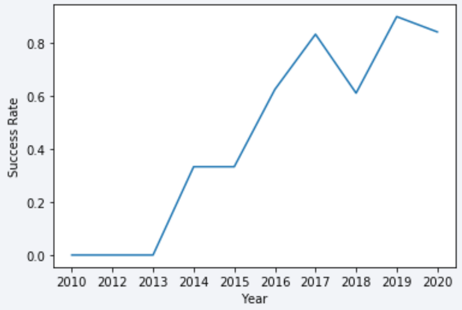
        Scatter plots are used to observe relationships between variables. In our case, we observe the relationship between flight number, payload mass, launch site, orbit and launch outcome.
    - Bar chart

        Bar charts are used to compare things between different groups. In our case, we compare the success rate of the launch for different orbit types.
    - Line chart

        Line charts are used to track changes over a period of time. In our case, we use line chart to visualize the success rate of the launches over multiple years
- Github URL for reference







EDA with SQL

- In order to understand the SpaceX Dataset, the dataset is loaded into a table in Db2 database.
- SQL queries were executed to answer queries below:
  - Names of the unique lauch sites in the space mission
  - Display 5 records where the launch site begin with the string 'CCA'
  - Find the total payload mass carried by booster launched by NASA (CRS)
  - Find the average payload mass carried by booster version F9 v1.1
  - Find the date for the first successful landing on a ground pad
  - List the names of the booster which landed successfully on drone ship and have payload mass of between 4000-6000kg
  - List the total number of successful and failed mission outcomes
  - List the names of booster_versions which have carried the maximum payload mass
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- Github URL for reference

Build an Interactive Map with Folium

- Folium is used to visualize the launch data through an interactive map.
- All the launch sites are plotted on a world map to explore the common characteristics of selecting these launch sites
- Launch outcomes are color coded with 0 (Failure) as red and 1 (success) as green. These outcomes are plotted as individual points at the launch site and clustered together using MarkerCluster ( )
- The distance between the launch sites its proximities is calculated to answer the questions below:
    - Are launch sites in close proximity to railways?
    - Are launch sites in close proximity to highways?
    - Are launch sites in close proximity to coastline?
    - Do launch sites keep certain distance away from cities?
- Github URL for reference

Build a Dashboard with Plotly Dash

- Plotly Dash is used to build interactive dashboard to visualize the launch data
- Interactive pie chart is used to how success counts for all sites, as well as success rates for each sites.
- Interactive scatter plot with a slider filter to filter the payload mass is used to visualize the relationship between success rate and payload mass.
- Github URL for reference

Predictive Analysis (Classification)

- Predictive analysis is broken into 4 phases:
  - Preparing dataset
  - Building Model
  - Improving Model
  - Identify best performing Model
- Github URL for reference

Preparing dataset

- Dataset is loaded into Pandas and NumPy
- Y dataset contains the 'Class' data
- X dataset contains the standardized feature data
- X ,Y data are split into training and test data sets

Building Model

- Below are the different types of machine learning algorithm model tested:
  - Logistic regression
  - Support vector machine
  - Decision tree classifier
  - K nearest neighbors

Improving Model

- Test out different parameters to improve the model

Find the best classification Model

- Calculate the best score for each models and compare which has be best score

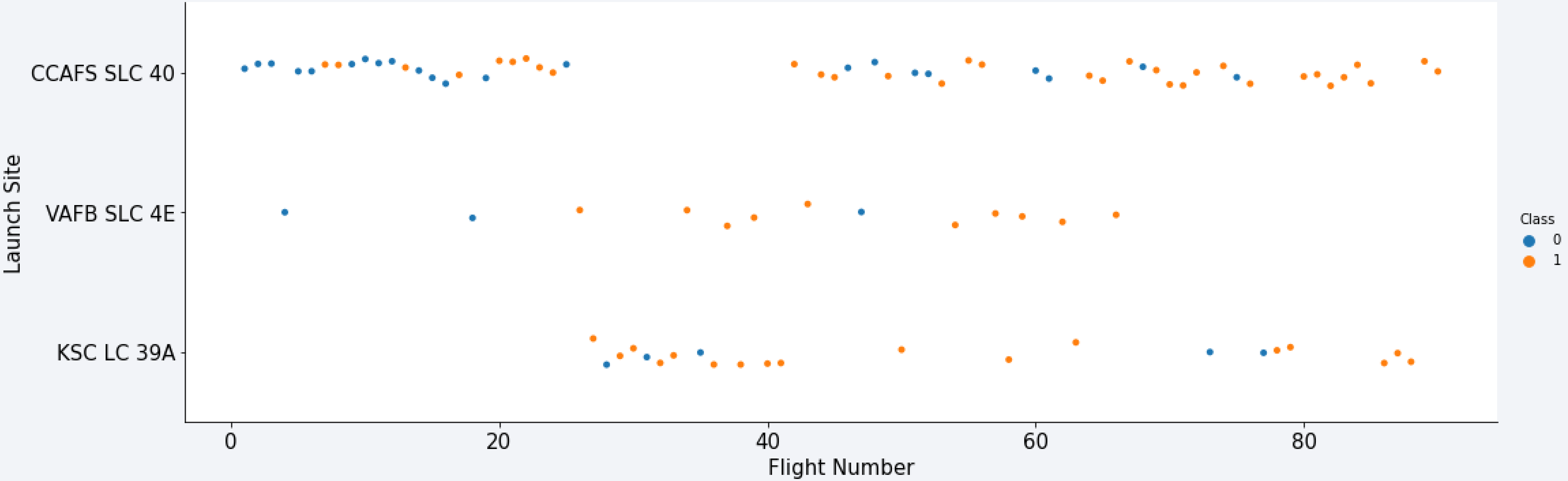
Results

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

Section 2

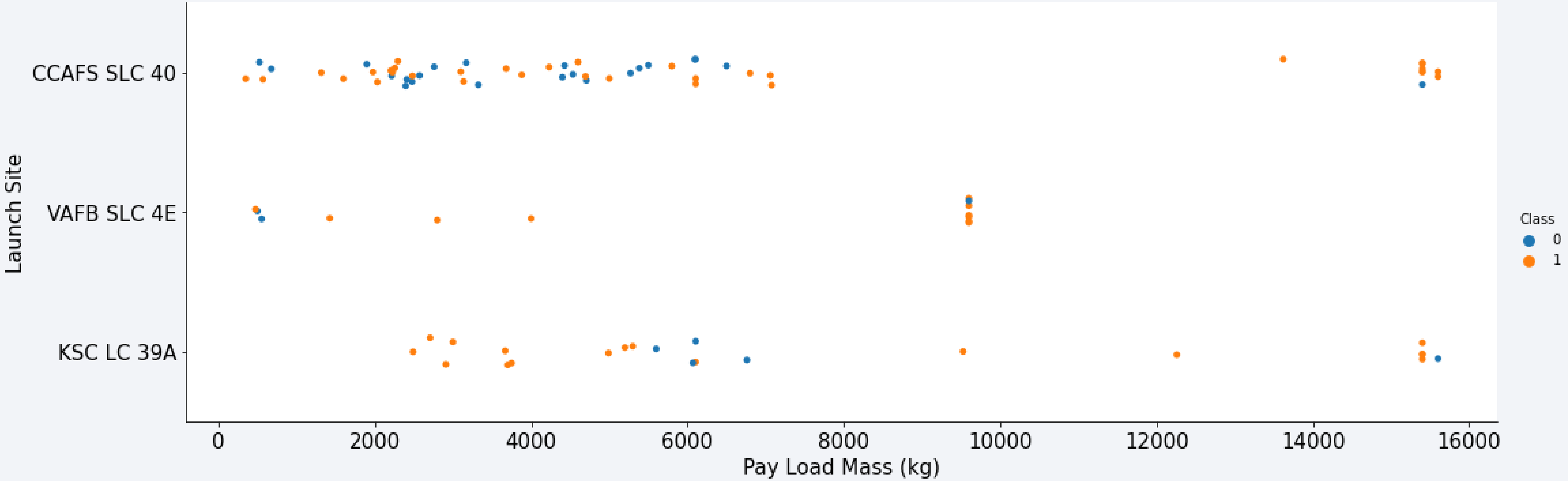
Insights drawn from EDA

Flight Number vs. Launch Site



- There are more flights launched from CCAFS SLC 40 if compared to other launch sites
- Almost all the initial 20 flights failed.
- Flights 80 onwards have 100% successful launch.

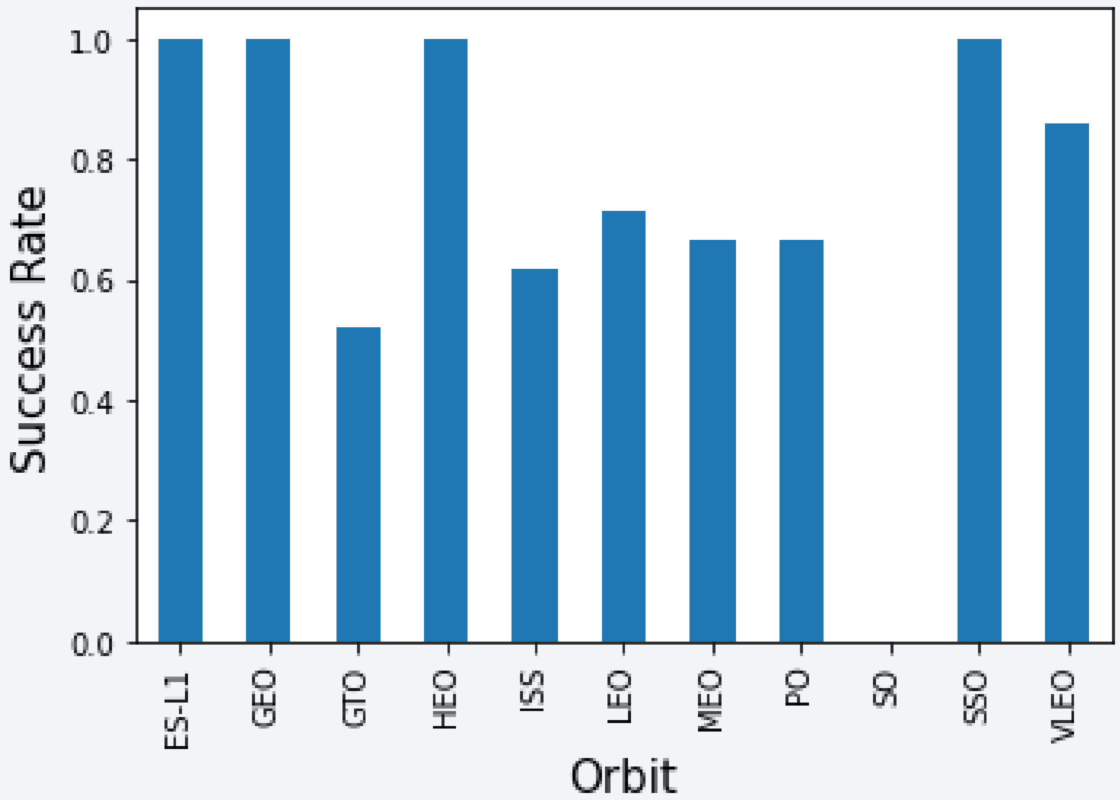
Payload vs. Launch Site



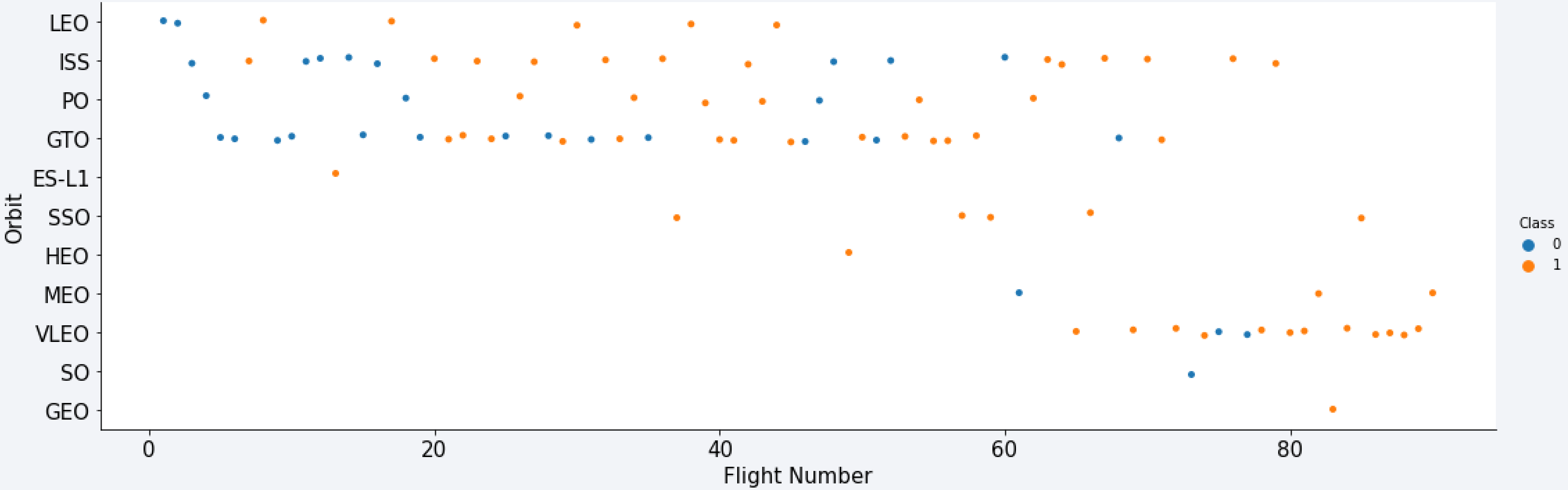
- There are more launches with pay load mass of less than 8000kg

Success Rate vs. Orbit Type

- Orbit ES-L1, GEO, HEO, SSO has 100% success rate while orbit SO has 0% success rate.

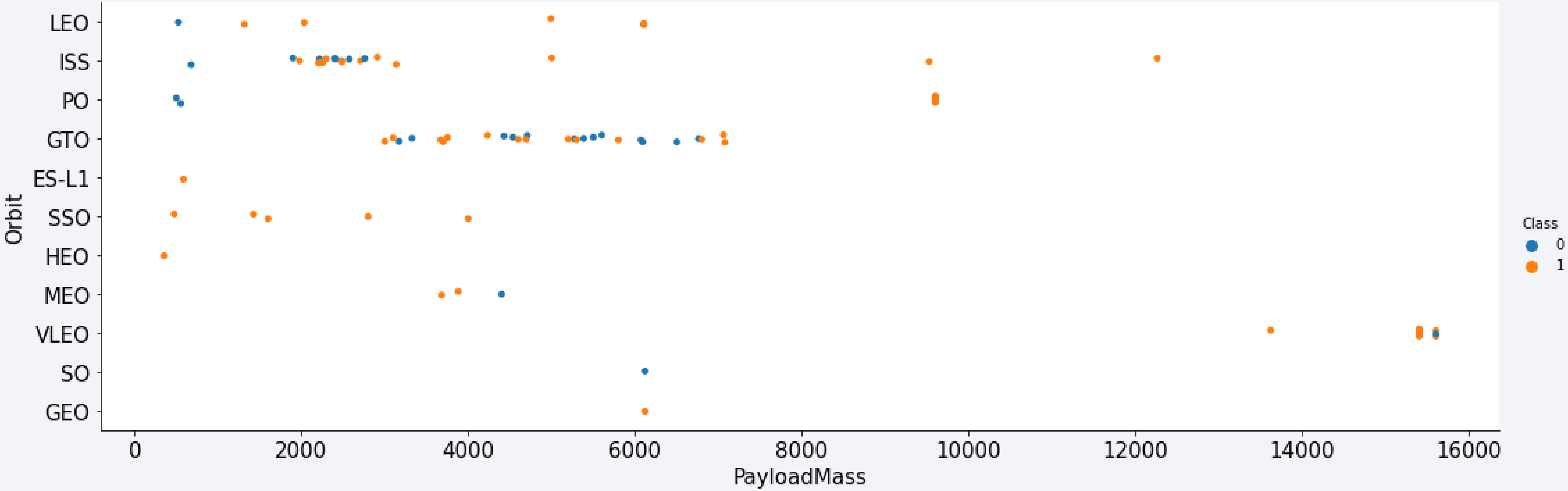


Flight Number vs. Orbit Type



- Later flights (flights 60 and above) are more focused on VLEO orbit and has a relatively high success rate.

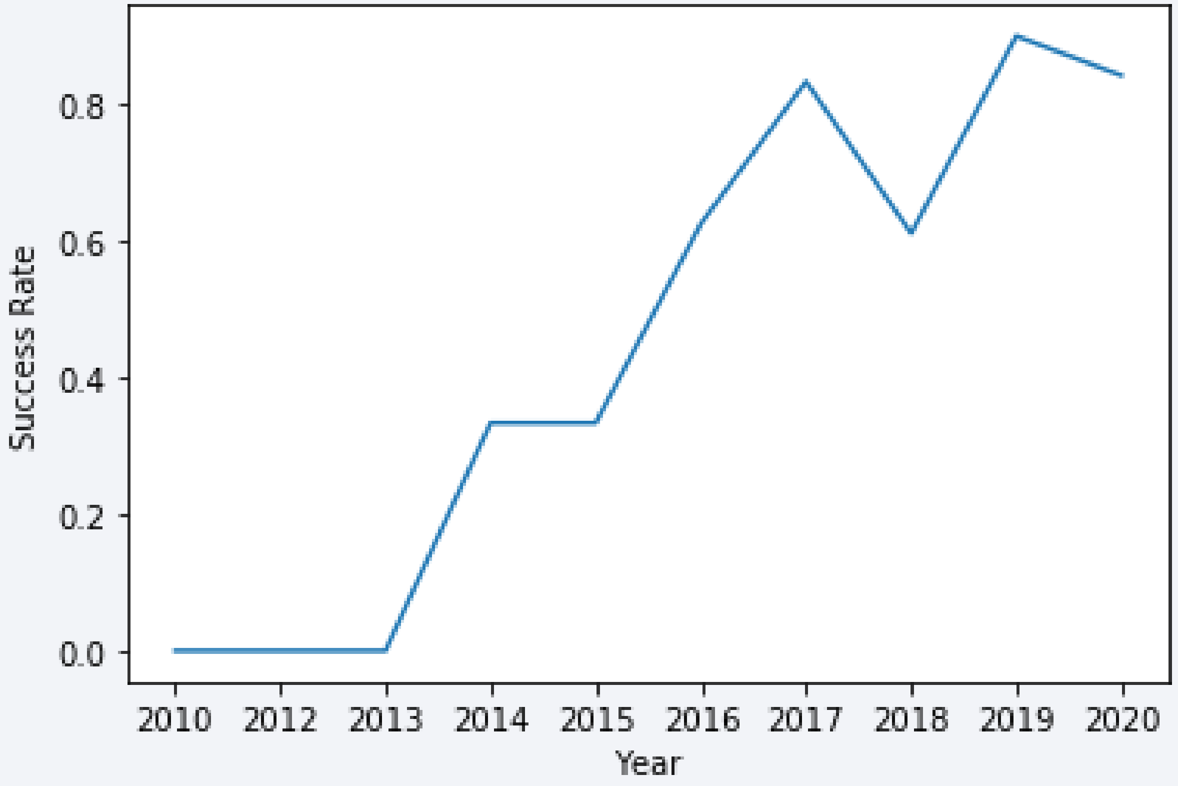
Payload vs. Orbit Type



- Launches with payload mass <8000kg are mostly launched at ISS and GTO orbit
- While the heavier payload mass launches (>14000kg) are launched towards the VLEO orbit.

Launch Success Yearly Trend

- The success rate of the launches increase over the years since 2013 till year 2020, with a minor setback in year 2018



All Launch Site Names

```
%sql SELECT DISTINCT(launch_site) FROM SPACEXTBL;
```

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

- Using **DISTINCT** to query unique values from **launch_site** column from table **SPACEXTBL.**
- The query returns 4 unique launch sites CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
%%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
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

- Using **WHERE … LIKE** to query columns where **launch_site** column has the keyword **'CCA%'** (note: % is a wild card) and limit the query result to only 5 rows with **LIMIT**

Total Payload Mass

```
%%sql
SELECT SUM(payload_mass__kg_) AS "Total Payload Mass (kg)"
FROM SPACEXTBL
WHERE customer='NASA (CRS)';
```

Total Payload Mass (kg)
45596

- Using **SUM ( )** to find the total of column **payload_mass_kg_** from the table **SPACEXTBL**
- **WHERE** is used to include the customer **NASA (CRS)** in the summation.
- The Total Payload Mass for NASA (CRS) is 45596 kg

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(payload_mass__kg_) AS "Average Payload Mass (kg)"
FROM SPACEXTBL
WHERE booster_version LIKE 'F9 v1.1%';
```

Average Payload Mass (kg)
2534

- **AVG ( )** is used to get the average of the **payload_mass_kg_** column from the table **SPACEXTBL**
- **WHERE** is used to filter the data to include only **booster_version** F9 v1.1
- **%** is used to include the variation of the F9 v1.1 booster.
- The average payload mass for F9 v1.1 is 2534 kg

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) AS "First Successful Ground Landing Date"
FROM SPACEXTBL
WHERE landing__outcome='Success (ground pad)'
```

First Successful Ground Landing Date
2015-12-22

- **MIN ( )** statement is used to find the minimum date in the **DATE** column from **SPACEXTBL** table
- **WHERE** statement is used to filter the dataset to include only **landing_outcome** of '**Success (ground pad)**'.
- The first successful ground landing date is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXTBL
WHERE landing__outcome='Success (drone ship)'
    AND (4000< payload_mass__kg_ <6000);
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

- Select only the **booster_version** column
- **WHERE** statement is used to filter out **landing_outcome = 'Success (drone ship)'**
- **AND** statement adds additional filter conditions to include only **payload_mass_kg_** between 4000 and 6000kg

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT mission_outcome, COUNT(mission_outcome) AS total_number
FROM SPACEXTBL
GROUP BY mission_outcome
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Query the **COUNT** of **mission_outcome** from **SPACEXTBL** table, **GROUP BY mission_outcome**
- There are 99 successful outcomes with 1 successful outcome with unclear payload status and 1 failure in flight.

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version
FROM SPACEXTBL
WHERE payload_mass__kg_ = (
    SELECT MAX(payload_mass__kg_)
    FROM SPACEXTBL);
```

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
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- First, perform a subquery to query the value for the maximum **payload_mass__kg_** using **MAX()** from **SPACEXTBL** table
- Then, query the booster_version from **SPACEXTBL** table, **WHERE** the **payload_mass__kg_** equals to the maximum value from the first query.

2015 Launch Records

```
%%sql
SELECT landing__outcome, booster_version, launch_site
FROM SPACEXTBL
WHERE landing__outcome='Failure (drone ship)' AND
    DATE LIKE '2015%'
```

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

- SELECT three columns; landing_outcome, booster_version and launch_site from SPACEXTBL table
- WHERE … AND statement is used to include two conditions; Failed landing on drone ship in the year 2015.

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

```
%%sql
SELECT landing__outcome, COUNT(landing__outcome) as landing_counts
FROM SPACEXTBL
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY landing_counts DESC;
```

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

- Query the **COUNT** of **landing_outcome** as **landing_counts GROUP BY landing_outcome**
- Filter the data queried **WHERE** the data is **BETWEEN** '2010-06-04' and '2017-03-20'.
- Used **ORDER BY… DESC** to order the query outcome by **landing_counts** in descending order.
- Most of the launch between those dates did not attempt landing.

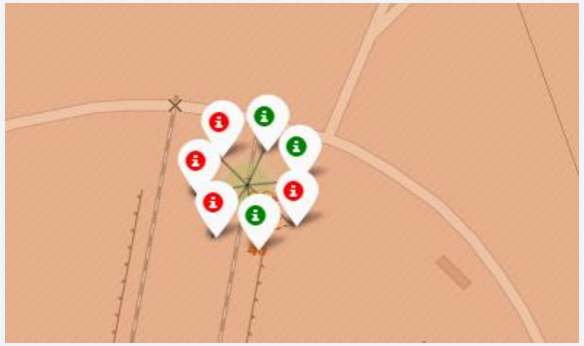
Section 4

Launch Sites Proximities Analysis

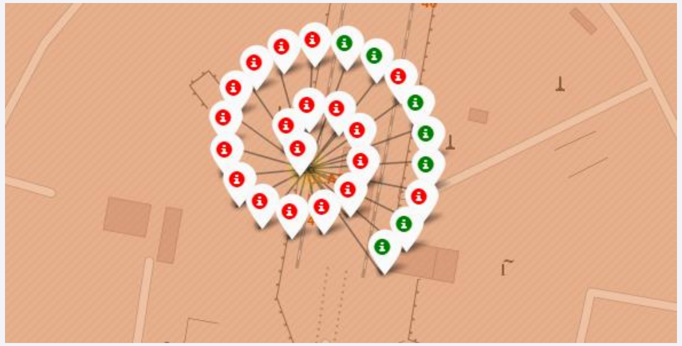
All launch sites marked on global map

- Figure on the right shows all the launch sites for Falcon-9 plotted on a global map
- All the launch sites are in the Southern part of United States of America and are all very close to the coasts.
- All the launch sites are south towards the Equator

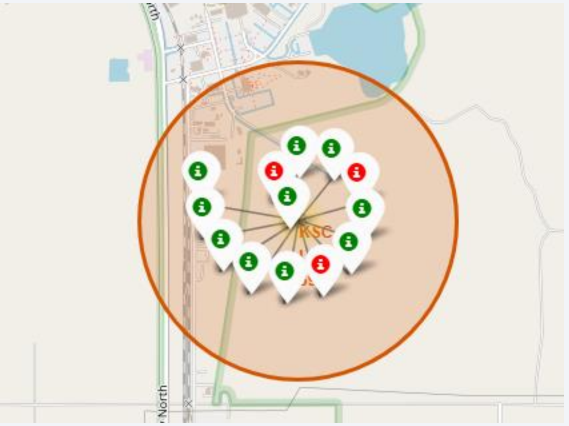
Color-labeled launch outcomes for all launch sites



Cape Canaveral Space Launch Complex 40



Cape Canaveral Launch Complex 40



Kennedy Space Center Launch Complex 39

Vandenberg Space Launch Complex 4

- The snapshots on the right shows color-labeled launch outcomes from all four Falcon-9 launch sites (Green marker: successful launch; Red marker: failed launch)
- Kennedy Space Center Launch Complex 39 shows a higher success rate of rocket launch if compared with other launch sites.

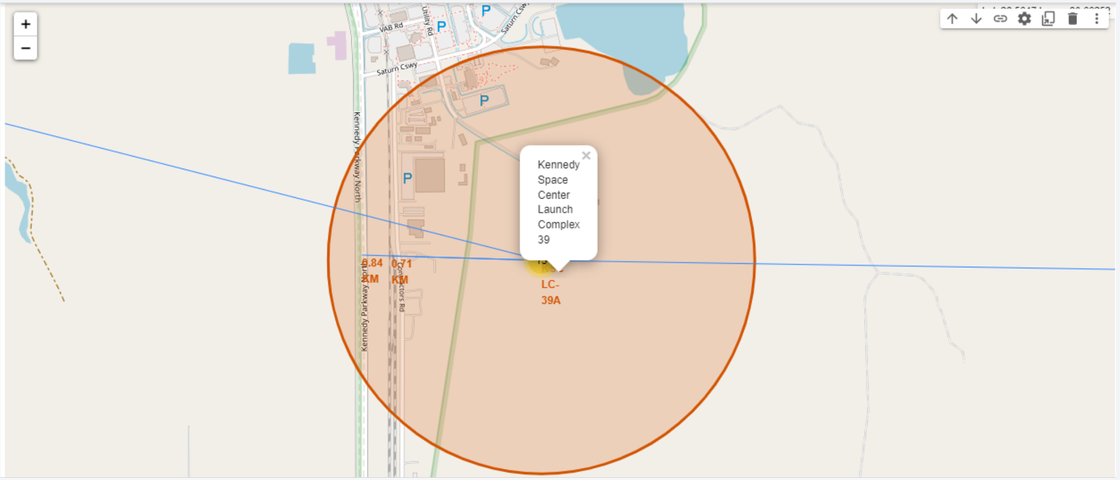
Distance between Kennedy Space Center Launch Complex 39 and its Proximities

- The two figure on the right shows the distance between Kennedy Space Center Launch Complex 39 and its proximities

- Table below shows a summary of the distance

Proximities	Distance (km)
Railway	0.71
Highway	0.84
Coastline	3.94
Nearest town	16.50

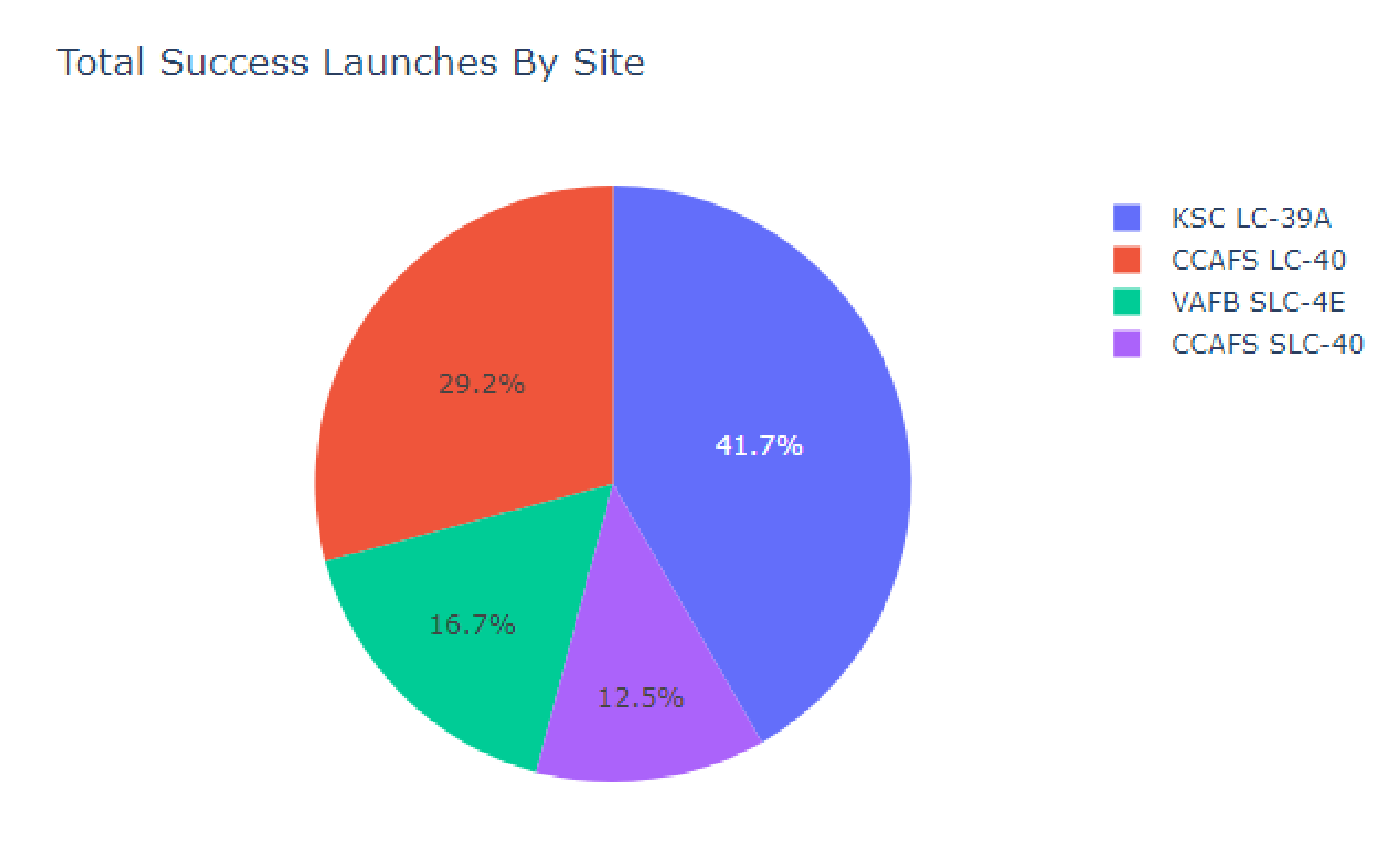
- Launch sites are in close proximity with railways, highways and coastlines but very far away from cities



Section 5

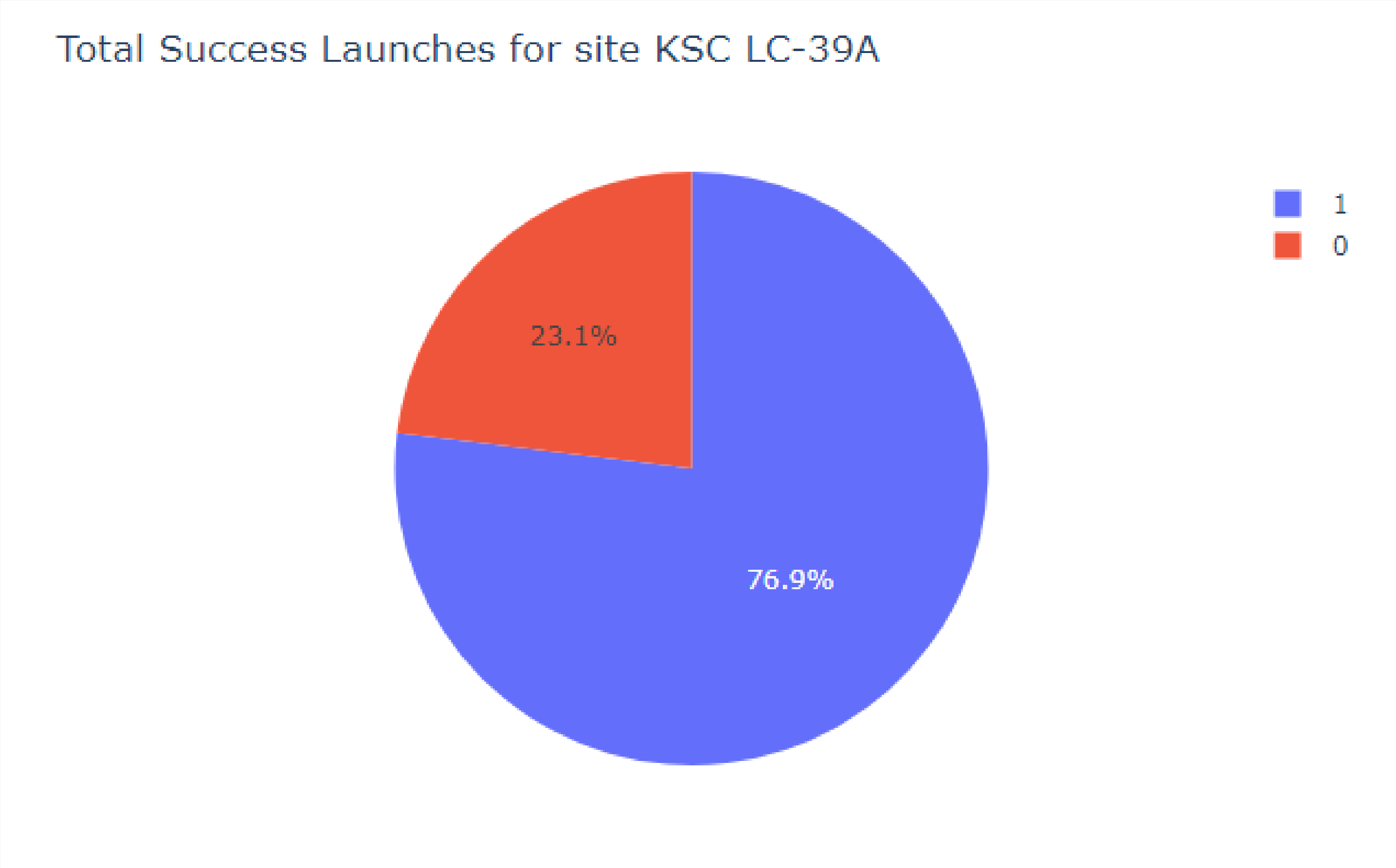
Build a Dashboard with Plotly Dash

Launch success count for all sites



- The piechart shows the success count for all launch sites.
- KSC LC-39A launch site has the most number of successful launches if compared with other sites

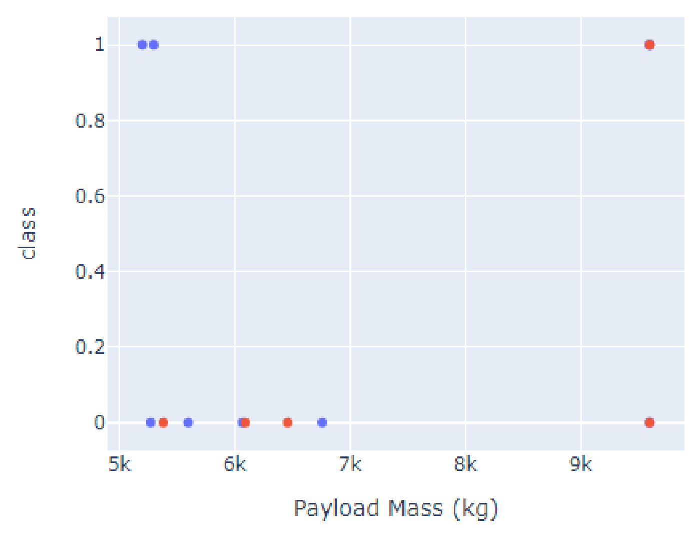
Piechart for the launch site with highest launch success ratio



- KSC LC-39A has a 76.9% success rate and 23.1% failure rate.

Payload vs. Launch Outcome scatter plot for all sites

Low weighted payload (0-5000kg)



High weighted payload (>5000kg)

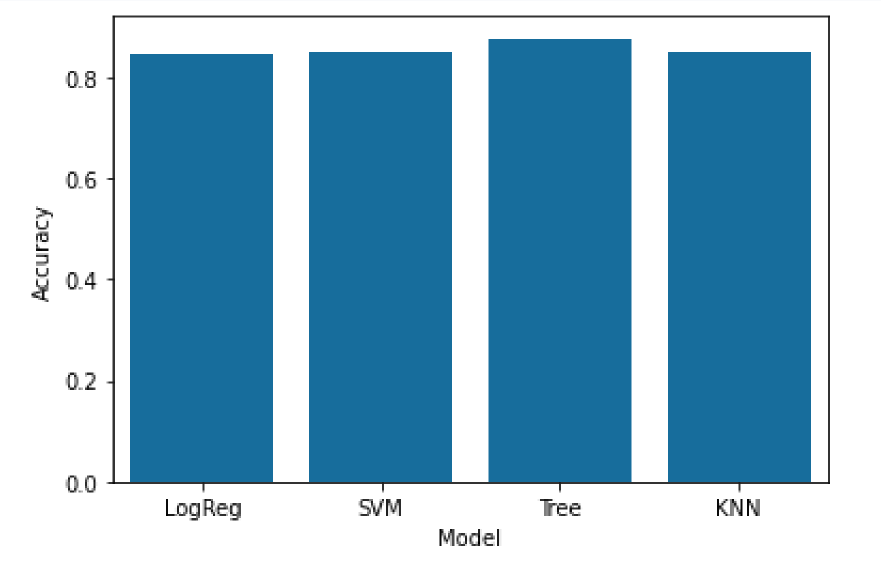
- The figures on the left shows the payload vs. launch outcome scatter plot for all sites for low weighted payloads (<5000kg) and high weighted payloads (>5000kg)
- Low weighted payloads has a higher success rate compared to heavy weighted payloads

Section 6

Predictive Analysis (Classification)

Classification Accuracy

- Bar chart on the right shows the accuracy for all four classification models; logarithmic regression, support vector machine, Decision tree, and K Nearest Neighbor model.
- Decision tree classifier model has the highest accuracy among the four models tested for this dataset.

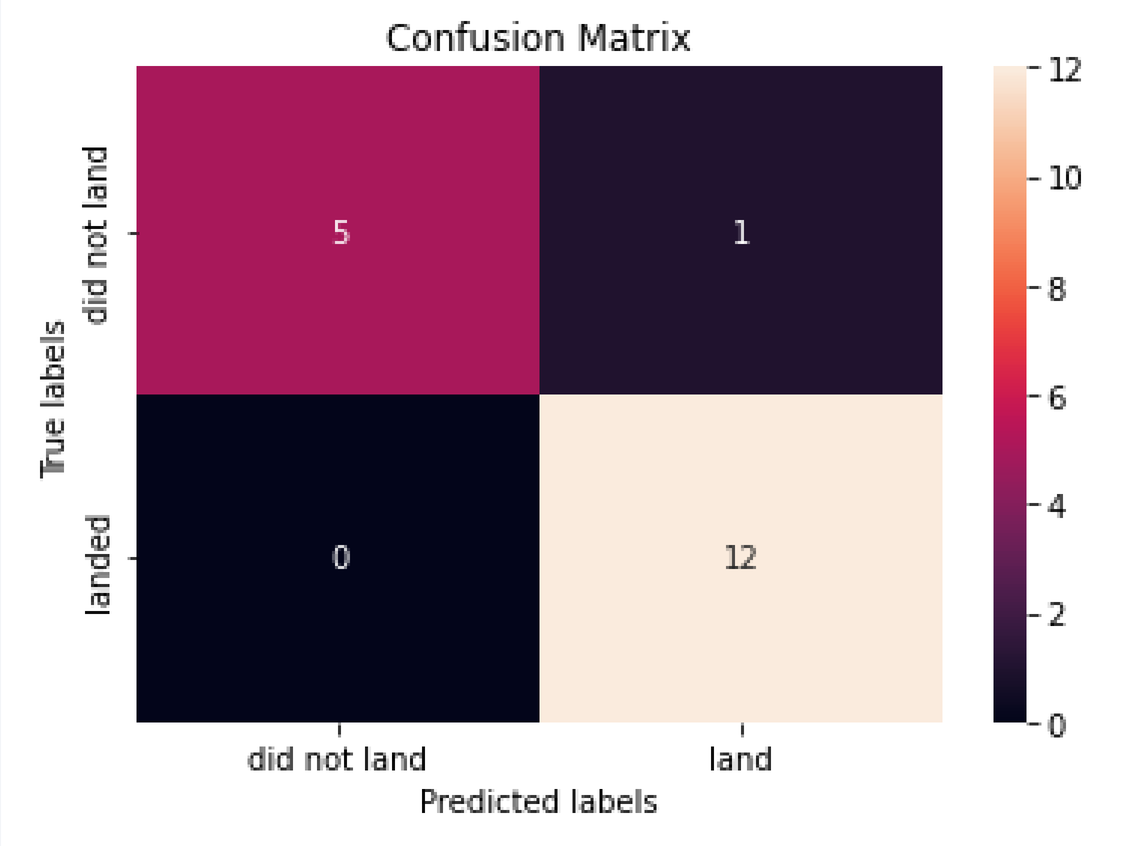


	Model	Accuracy
0	LogReg	0.846429
1	SVM	0.848214
2	Tree	0.876786
3	KNN	0.848214

```
Best performing model:  Tree
Score:  0.8767857142857143
Best paramaters:  {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'}
```

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the model almost perfectly predicted the outcomes with just one false positive.



Conclusions

- From EDA through visualization:
  - Success rates increases over the years as more launches are completed
  - More launches has pay load mass of less than 8000kg
  - Orbit ES-L1, GEO, HEO, SSO has 100% success rate
- From EDA through SQL:
  - Total payload mass for NASA (CRS) is 45596kg
  - Average payload mass for F9 v1.1 rockets is 2534kg
  - Out of all the launches in the SQL table, there are 100 successful outcome with 1 failure.

Conclusions

- From Folium map:
  - Kennedy Space Center Launch Complex 39 has the highest launch success rate
  - The launch center are near to railway, highway and coastline but far away from towns
- From Plotly Dash interactive dashboard:
  - Kennedy Space Center Launch Complex 39 has the highest launch success rate
  - Launches with low weighted payloads (<5000kg) has a higher success rate.
- From predictive analysis (Classification):
  - Decision tree model has the highest accuracy score (0.8767) among the models tested for this dataset.

Appendix

- None

Thank you!