IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yovan Ramyead
June 13, 2023

Outline

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

Executive Summary

- **Summary of methodologies:**

1. Data Collection - SpaceX API and Webscraping
2. Data Wrangling
3. Exploratory Data Analysis - SQL, Pandas and Matplotlib
4. Interactive Visual Analytics - Folium and Ploty Dash
5. Predictive Analysis - Machine Learning

- **Summary of all results**

1. Exploratory Data Analysis Results
2. Visual Analytics Results
3. Predictive Analysis Machine Learning Results

Introduction

- **Project background and context**

  SpaceX Falcon 9 launches at $62 million

  Reusable first stage rocket - major savings

  Competitors launches > $165 million

  Data collected used for competition bid

- **Problems you want to find answers**

  Predicting first stage rocket successful landing

Section 1

Methodology

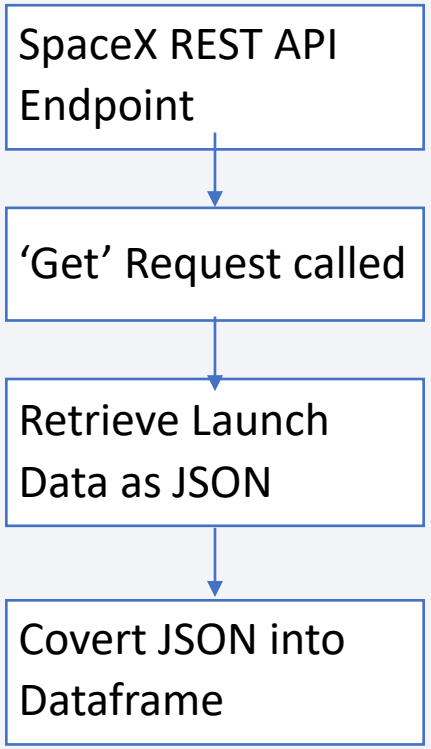
Methodology

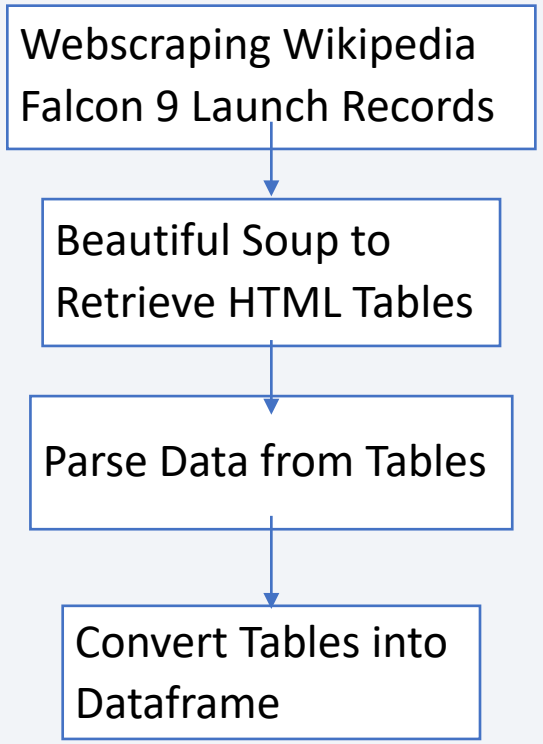
Executive Summary

- Data collection methodology:
  - Using SpaceX REST API and Webscraping Wikipedia
- Perform data wrangling
  - Data converted to Pandas Dataframe, filtered, missing values replaced, one hot encoding
  - Then Data analysis and Modelling performed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Machine Learning - Logistic Regression, Decision Tree, K Nearest Neighbour and Support Vector Machine
  - Models Built to predict landing outcomes

Data Collection

- Data was gathered through SpaceX REST API and Webscraping Wikipedia page
-



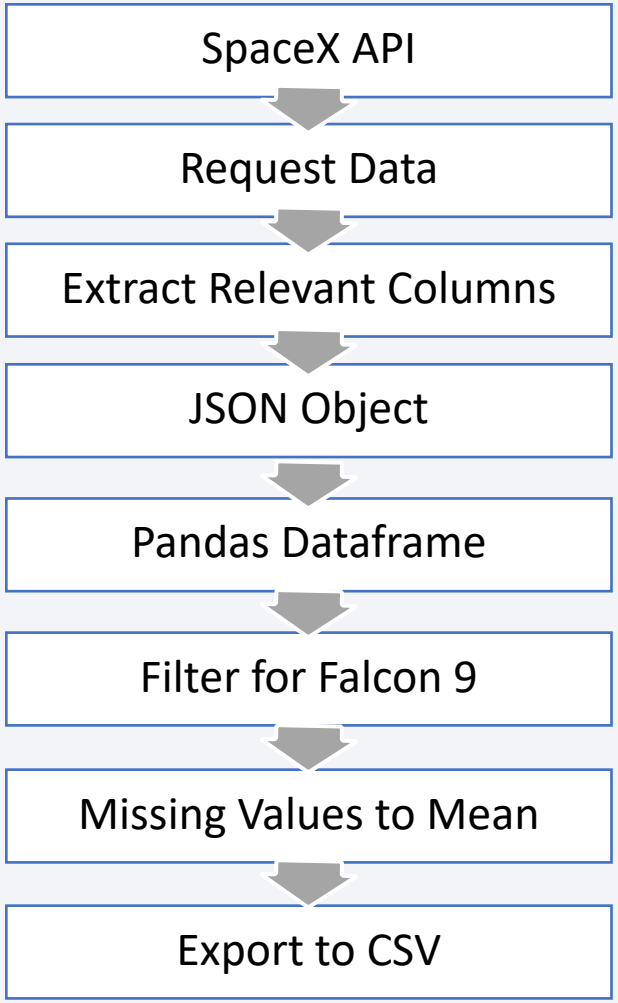


Data Collection – SpaceX API

- Request Data - SpaceX API launch data('Get' request
- Extract info about booster name, launch site, payload and landing site
- Retrieve data using JSON object
- Normalize JSON to Pandas Dataframe
- Filter Dataframe for Falcon 9 launch data
- Replace missing values with .mean()
- Export Data to CSV file

SpaceX API calls notebook:

https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/01-spacex-data-collection-api.ipynb

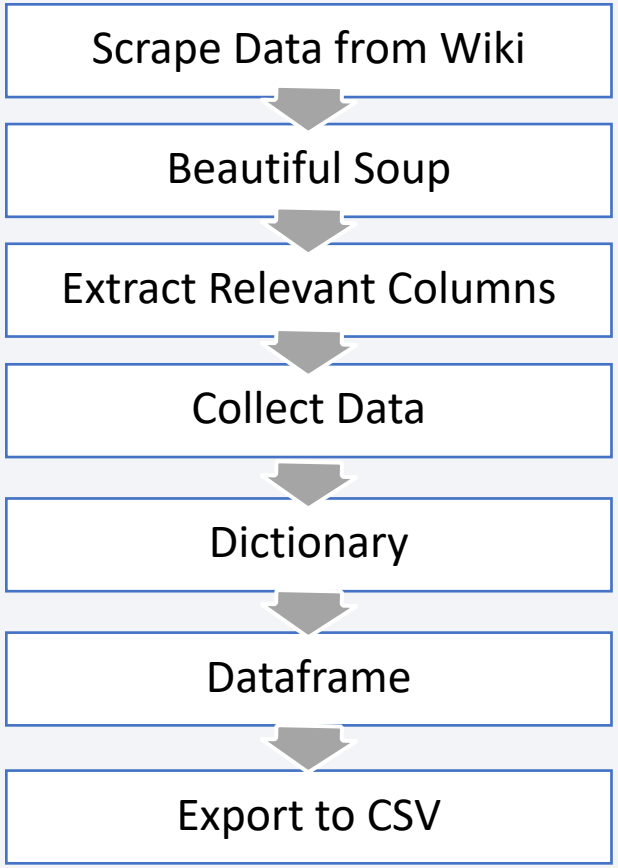


Data Collection - Scraping

- Request Falcon 9 launch data from Wikipedia
- Create Beautiful soup object from HTML
- Extract column names from HTML tables
- Collect data by parsing HTML tables
- Create dictionary
- Create Dataframe
- Export data to CSV

GitHub URL web scraping notebook:

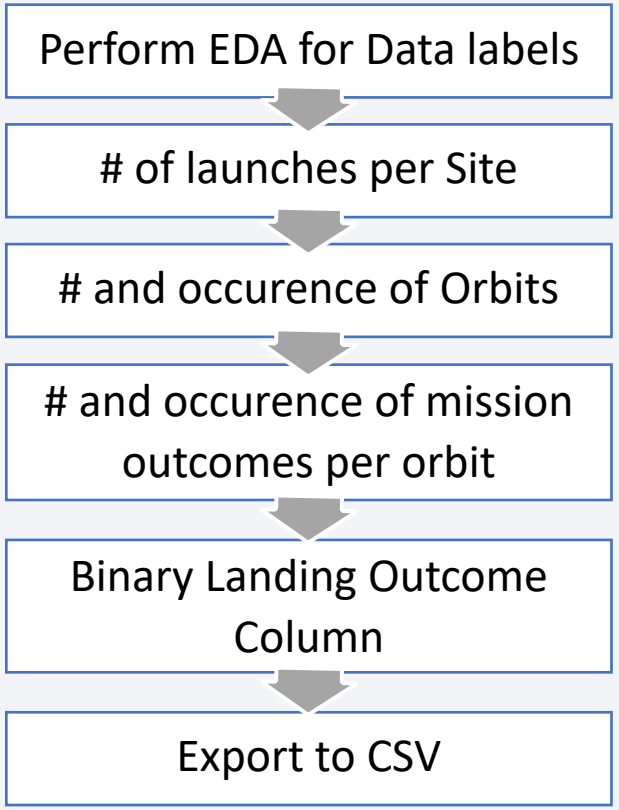
https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/02-labs-webscraping.ipynb



Data Wrangling

- Perform EDA to determina data labels
- Calculate the number of launches for each site
- Calculate the number and occurence of each orbit type
- Calculate the number and occurence of the mission outcome for each orbit type
- Create a landing outcome label from the 'Outcome' column using one-hot encoding
- Export to CSV

Data wrangling notebook: https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/03-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

- Scatter Plots to find relationship between 2 variables:
  - Flight Number vs Launch Site
  - Payload vs Launch Site
  - Flight Number vs Orbit Type
  - Payload vs Orbit Type
- Bar chart to compare values among different groups:
  - Success Rate vs Orbit Type
- Line chart to illustrate data trend over time:
  - Success Rate vs Orbit Type

EDA with data visualization notebook: https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/05-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL queries performed:

Display queries:

- Unique launch site names
```
%%sql select DISTINCTLaunch_Site from SPACEXTBL
```
- 5 launch sites beginning with ‘CCA”
```
%%sql select * from SPACEXTBL where Launch_Site LIKE'CCA%'
LIMIT 5
```
- Total weight payload launched at NASA CRS
```
%%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE
Customer = 'NASA (CRS)'
```
- Average payload mass in booster F9 v1.1
```
%%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE
Booster_Version LIKE "F9 v1.1%"
```

EDA with SQL notebook: https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/04-eda-sql-coursera_sqllite.ipynb

EDA with SQL (cont'd)

SQL queries performed:

Listing queries:

- Date of first successful landing on ground pad

```
%%sql SELECT min(Date) FROM SPACEXTBL
WHERE "Landing_Outcome" = "Success (ground pad)"
```

- Names of boosters landing successfully on drone ship with payload greater than 4000 and less than 6000

```
%%sql select Booster_Version, Payload_Mass__Kg_ from SPACEXTBL
where "Landing_Outcome" = "Success (drone ship)" and
"Payload_Mass__Kg_" between 4000 and 6000
```

- Total number of successful and failed missions

```
%%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

EDA with SQL notebook: https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/04-eda-sql-coursera_sqllite.ipynb

EDA with SQL (cont'd)

SQL queries performed:

Listing queries:

- Names of boosters with max payload

```
%%sql select Booster_Version, Payload_Mass__Kg_ from SPACEXTBL
    where Payload_Mass__Kg_ = (select MAX(Payload_Mass__Kg_) from SPACEXTBL)
```

- Failed landings on drone ship, booster version and launch site in 2015

```
%%sql select substr(Date, 4, 2) as month,
"Landing_Outcome", Booster_Version, Launch_Site from
SPACEXTBL where substr(Date, 7, 4) = '2015' and
"Landing_Outcome" = 'Failure (drone ship)'
```

- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

```
%%sql SELECT [Landing_Outcome], count(*) as Count_Outcomes
FROM SPACEXTBL
WHERE Date between '04-06-2010' and '20-03-2017' group by
[Landing_Outcome] order by count_outcomes DESC;
```

EDA with SQL notebook: https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/04-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Summarize of map objects and function:

- Circle: to highlight area on map
- Marker: to add text label to area on map
- MarkerCluster: to view multiple markers around the same are in an organized way
- MousePosition: to get coordinates on map using the mouse
- Lines: to illustrate relative position of different locations with respect to each other

Interactive map with Folium map notebook: https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/06_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

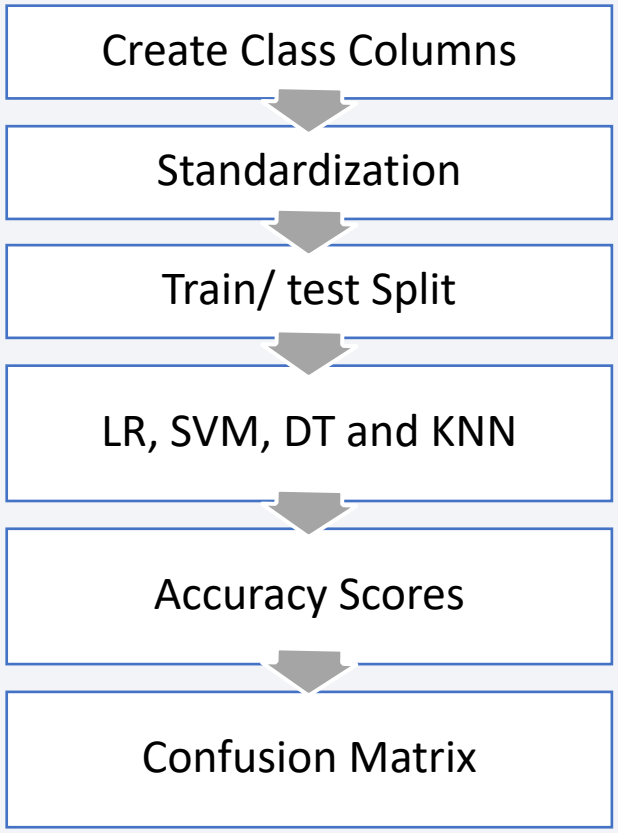
- Dropdown list with Launch Sites: To select launch sites all at once or individually
- Pie Charts: To illustrate proportion of successful and unsuccessful launches
- Slider of Payload Mass: to select varying payload mass ranges and see the success rate
- Scatter Plot: to show Payload Mass vs Success Rate by booster version

Plotly Dash lab notebook: https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/07_spacex_dash_app.py

Predictive Analysis (Classification)

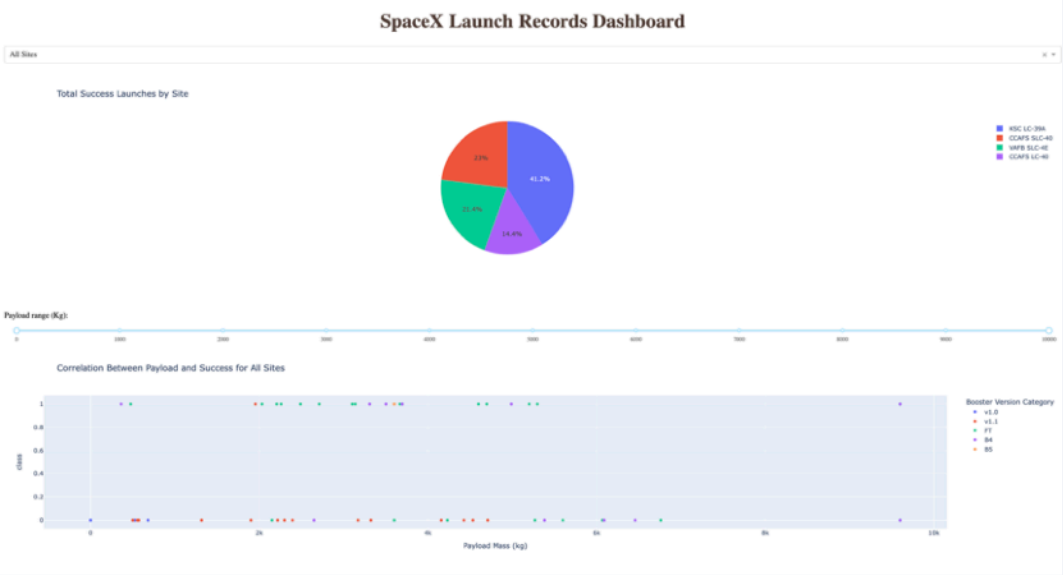
- Created Column for Class
- Standardized the data
- Split data into training and test sets
- Found best parameters for Logistoc Regression, SVM, Decision Tree and KNN
- Used test data to evaluate models based on accuracy scores and confusion matrix

Predictive Analysis lab notebook: https://github.com/yovan86/Data-Science-Capstone-SpaceX/blob/main/08_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

- Exploratory data analysis results: Success rate of Falcon rocket was 66.67%
- Predictive analysis results: Decision Tree was the best classification method with accuracy score of 87.32%
- Interactive Analysis demo:

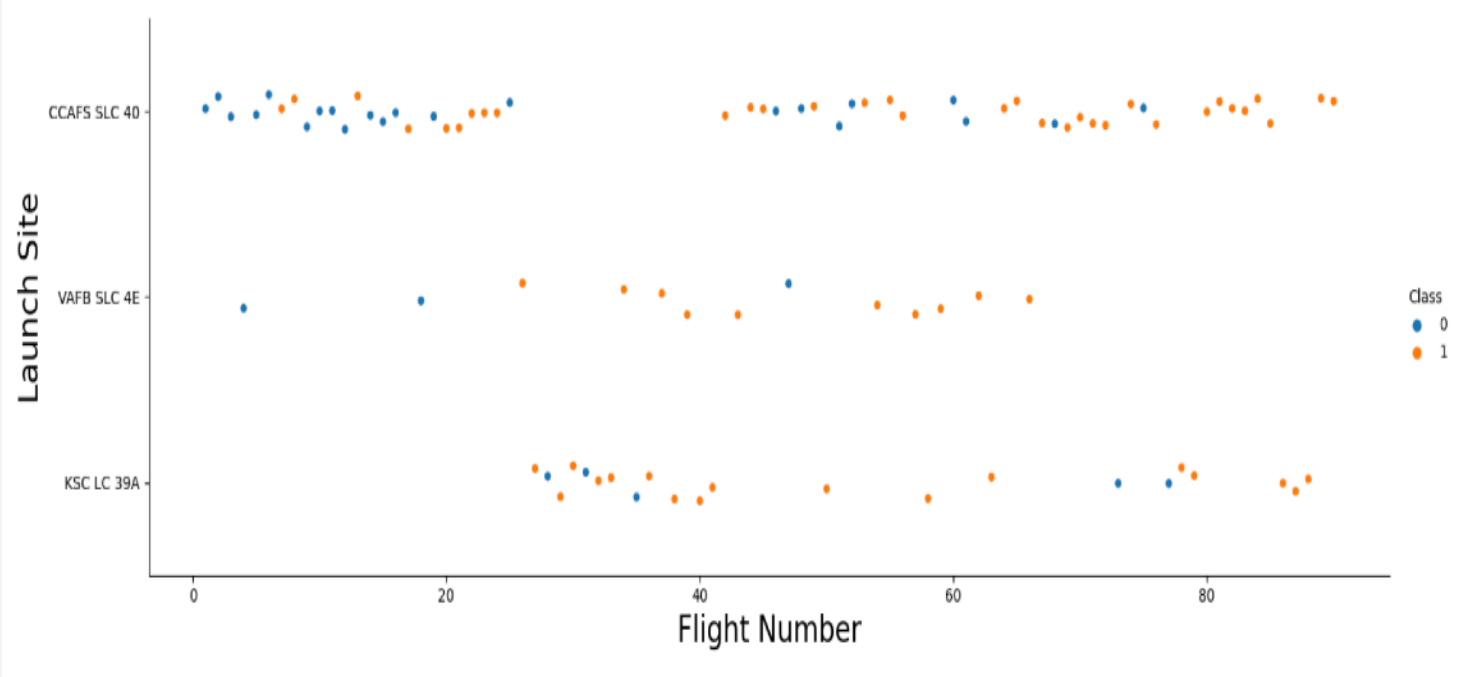


Section 2

Insights drawn from EDA

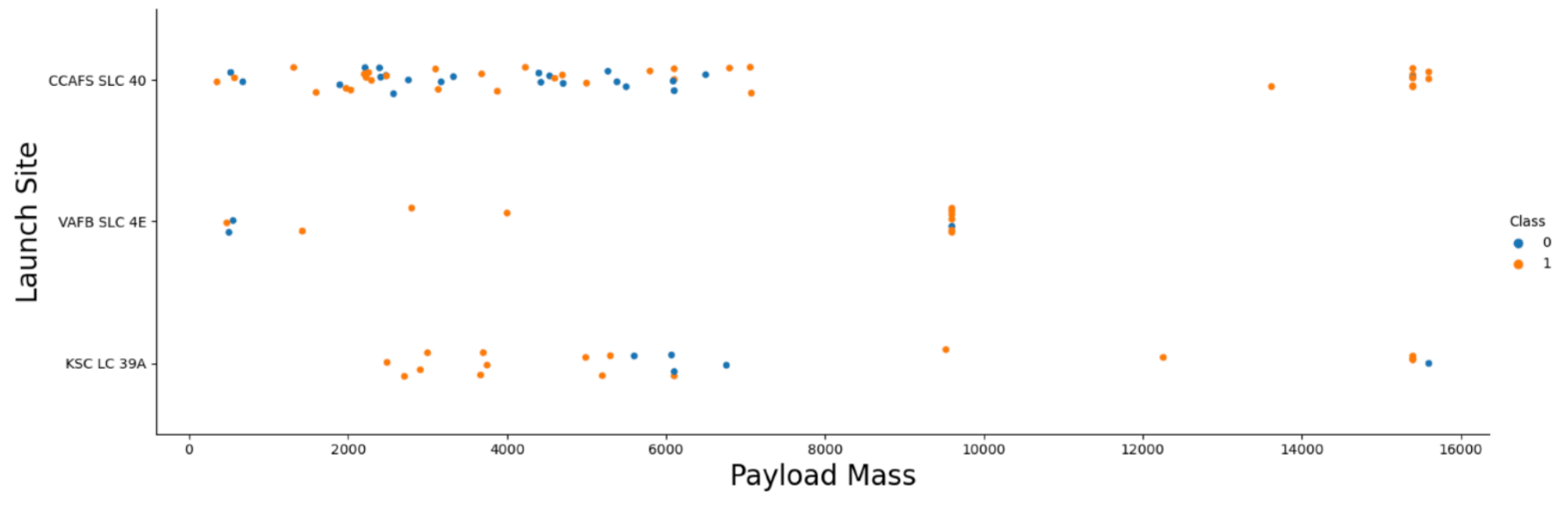
Flight Number vs. Launch Site

- The plot shows that as the flights number increased, the success rate also increased
- The orange red dots represent successful flights and the blue dots failed flights



Payload vs. Launch Site

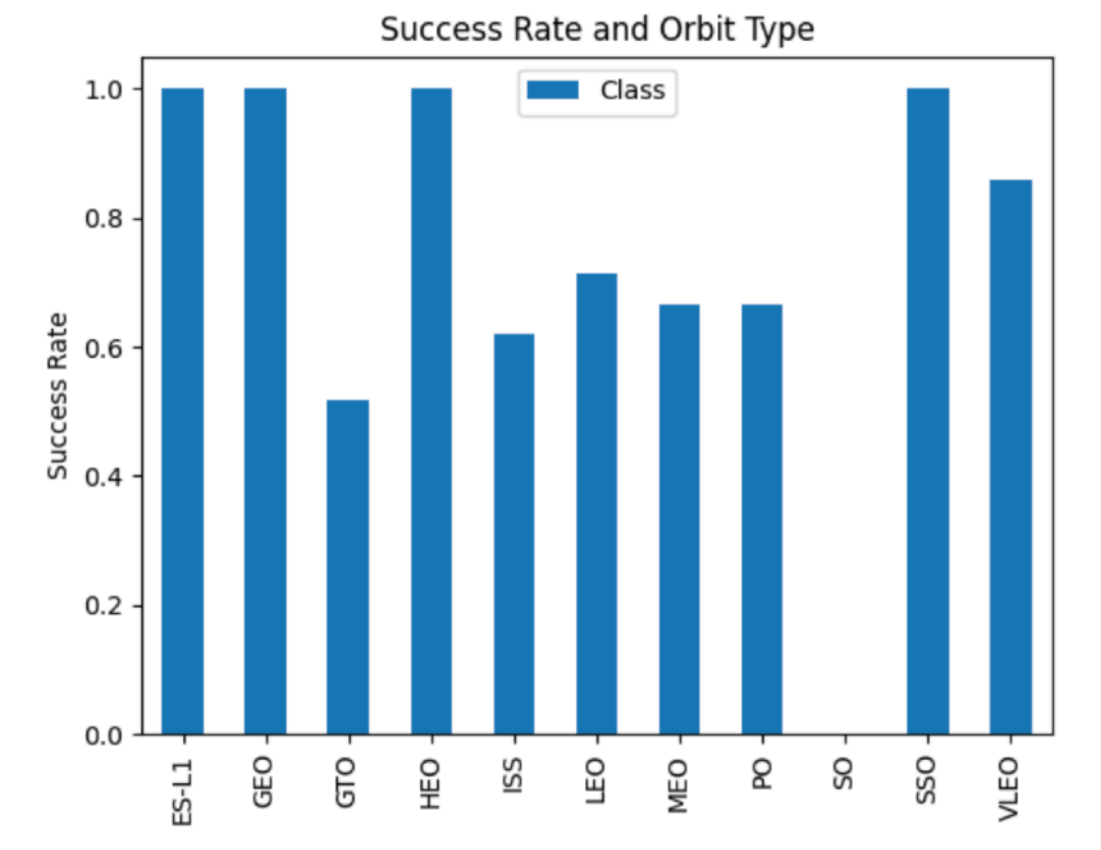
- No payloads over 10000kg at VAFB-SLC launch site
- No payloads less than 2500kg at KSC LC site
- Payloads less than 7500kg and over 13000kg at CCAFS SLC



Success Rate vs. Orbit Type

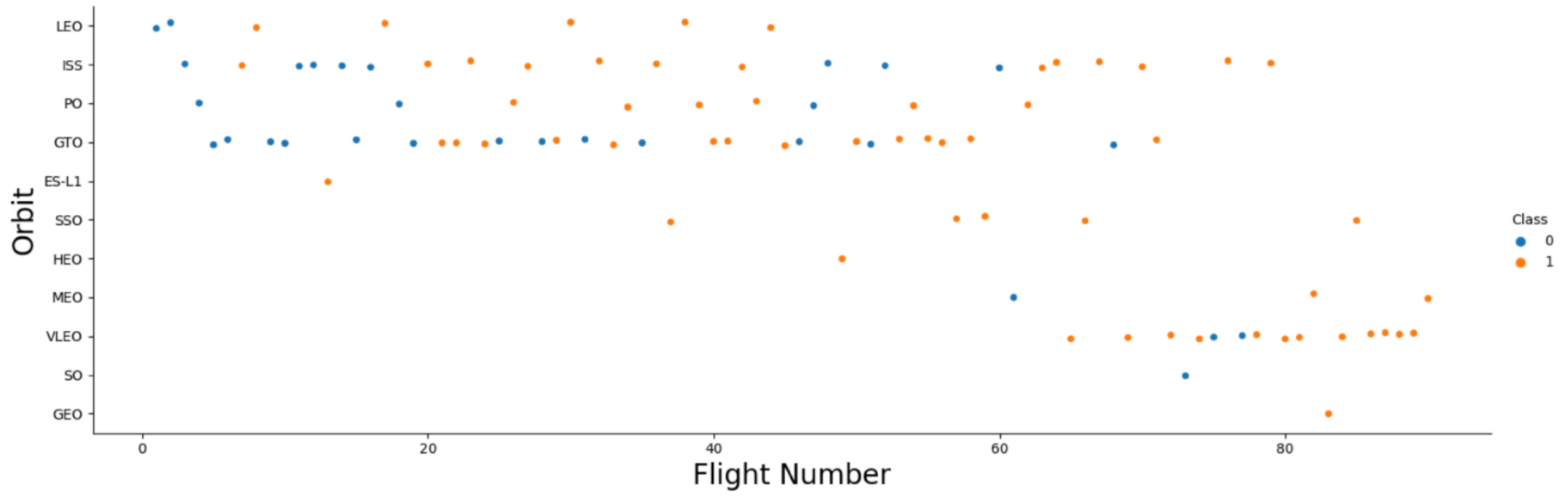
- 100% Success Rate: ES-L1, GEO, HEO, SSO

- S0 had zero success rate



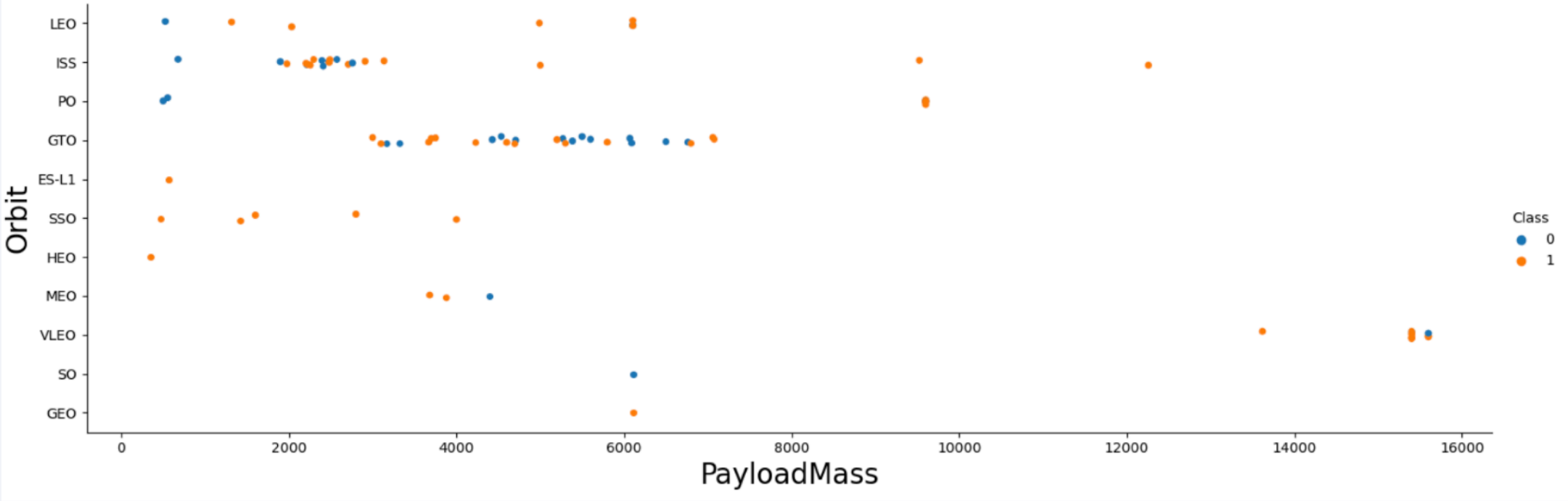
Flight Number vs. Orbit Type

- For LEO orbit the number of flights has a positive correlation with the success rate
- For GTO orbit there is no clear correlation
- For SSO orbit, there is a 100% success rate, but much less launches
- Flight numbers over 40 have a higher success rate



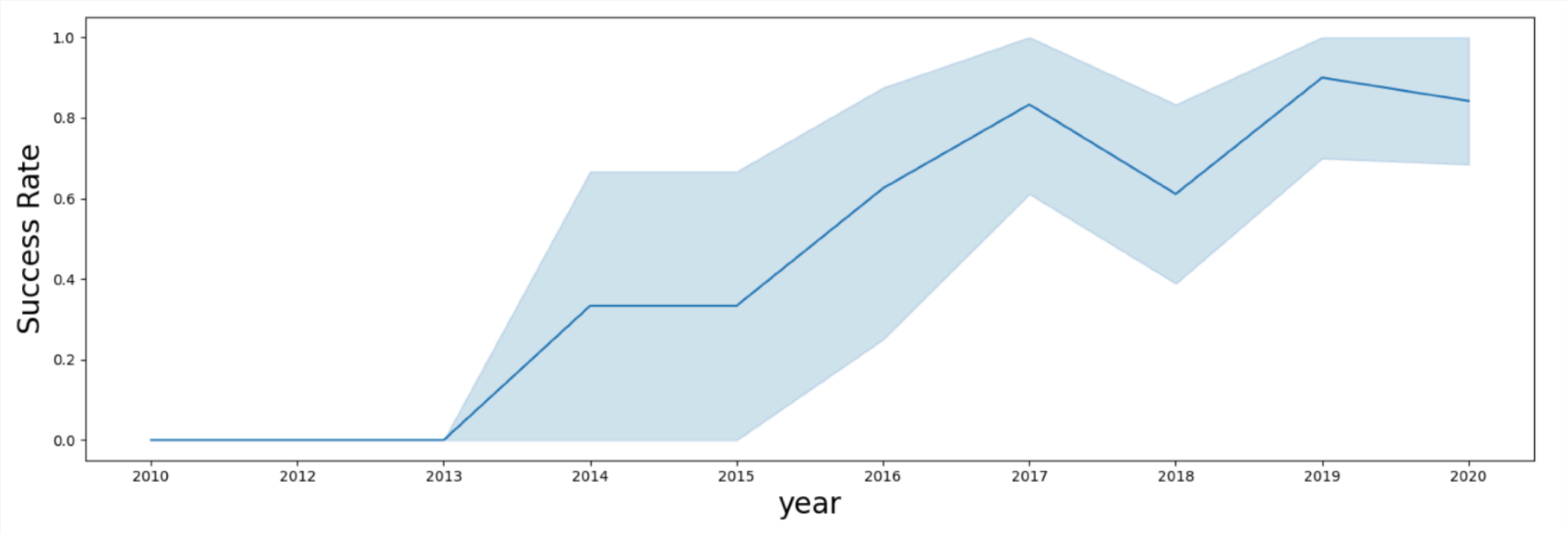
Payload vs. Orbit Type

- Heavier payloads have a higher success rate for LEO, ISS and PO
- No relationship between payload mass and success rate for GTO



Launch Success Yearly Trend

- Success rate increased in general as years increased since 2013
- Slight dip in success between 2017 and 2019



All Launch Site Names

- Launch Site Names: CCAFS LC-40,
  CCAFS SLC-40,
  KSC LC-39A,
  VAFB SLC-4E

- “DISTINCT” statement used to find unique values from the launch site column

Out[8]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

- The “LIKE”and “LIMIT” clauses were used to show the top 5 results with launch site name starting with ‘CCA’

Done.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- SUM() function used to calculate total payload mass for NASA CRS booster from the ‘PAYLOAD_MASS_KG_’ column

0]:

SUM(PAYLOAD_MASS__KG_)
45596.0

Average Payload Mass by F9 v1.1

- AVG() function and WHERE clause used to calculate the average payload mass and filter results respectively for booster F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- min(Date) function and WHERE clause used to find the first successful landing date and filter for landing outcome success respectively

min(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- BETWEEN clause used to select results with payload masses between 4000 and 6000
- WHERE clause used to filter for boosters successfully landing on drone ship

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696.0
F9 FT B1026	4600.0
F9 FT B1021.2	5300.0
F9 FT B1031.2	5200.0

Total Number of Successful and Failure Mission Outcomes

- COUNT() function and GROUP BY clause used to count the successful and unsuccessful missions
- Total of 99 successful and 2 failures

Mission_Outcome	total_number
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Max() function used as subquery to display list of boosters with max payload mass

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- SELECT clause with substr() used to retrieve month, landing outcome, booster version and lauch site for year 2015

month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- COUNT() function used to count landing outcomes
- Where and between clauses used to filter results for 2010-06-04 and 2017-03-20
- GroupBy clause to group outcome
- OrderBy and Desc clauses used to sort result in descending order

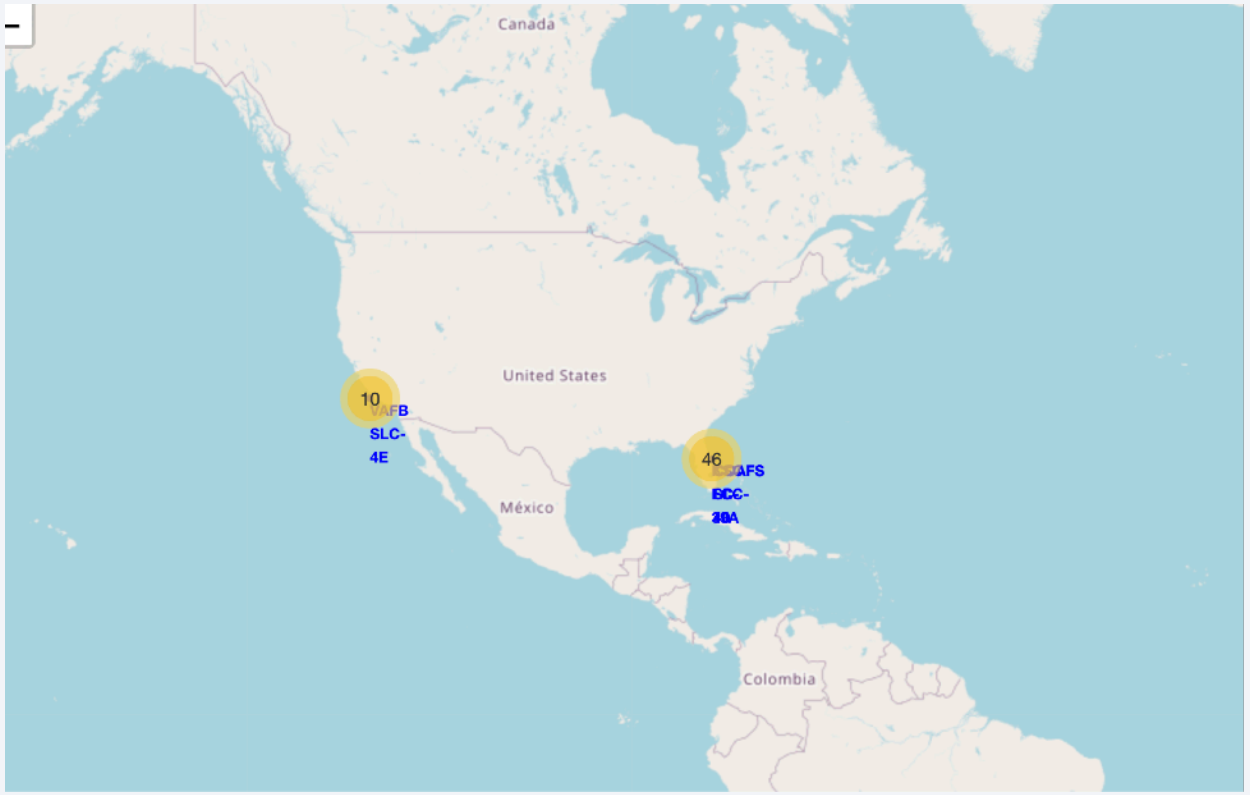
Landing_Outcome	Count_Outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Location Markers

- Yellow spots indicate the SpaceX launch sites in the US
- Icons in blue showing launch site names
- Launch sites near equator and coast line for easy launch and safety respectively

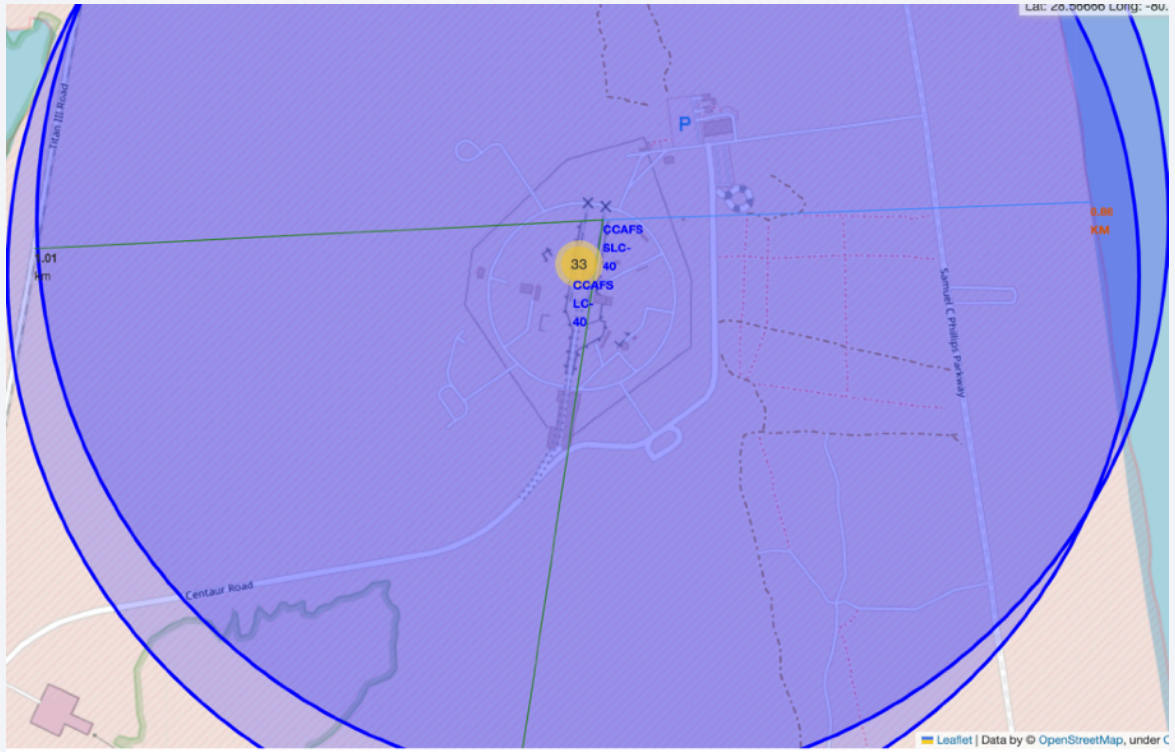


Launch Outcomes

- Green markers for successful launches and Red markers for failed launches
- CCAFS LC40 launch site has 7 successful launches out of 26

Launch Site Proximities

- Launch Site is relatively close to coast line and far away from cities for safety reason to prevent debris from causing damage and injuries
- Launch site close to railway, highway and coastline for easy access to material and personnel

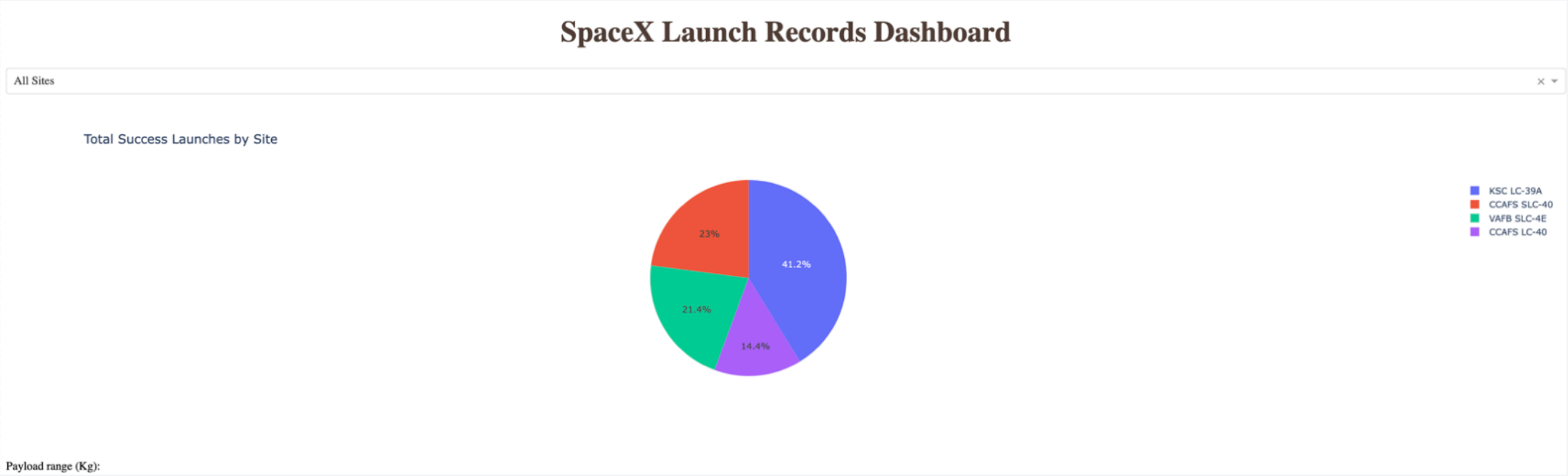


Section 4

Build a Dashboard with Plotly Dash

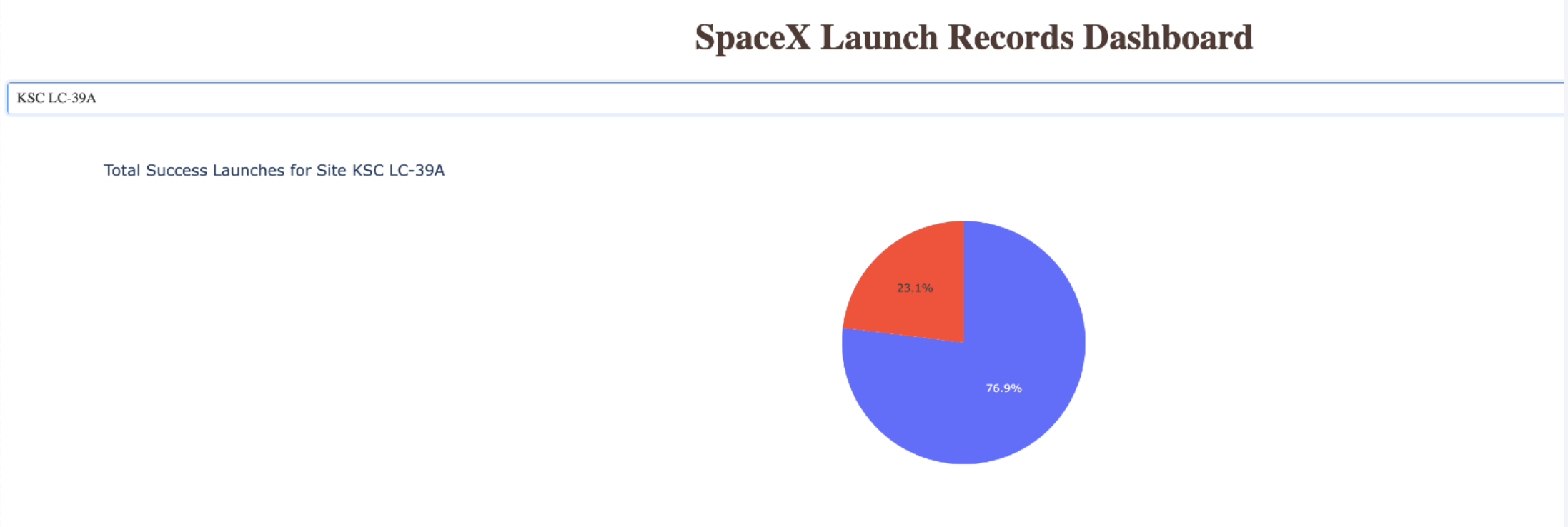
Launch Success Rate - All Sites

- KSC LC-39A has highest launch success at 41.2%
- CCAPS LC-40 has lowest launch success at 14.4%



Highest Success Rate Site

- Site KSC LC-39A has 76.9% success rate, highest among all sites



Payload vs Launch Outcome

- Booster version FT seem to have the highest success rate
- Highest success rate between 2500-5000kg
- Under 2500kg most failed launches

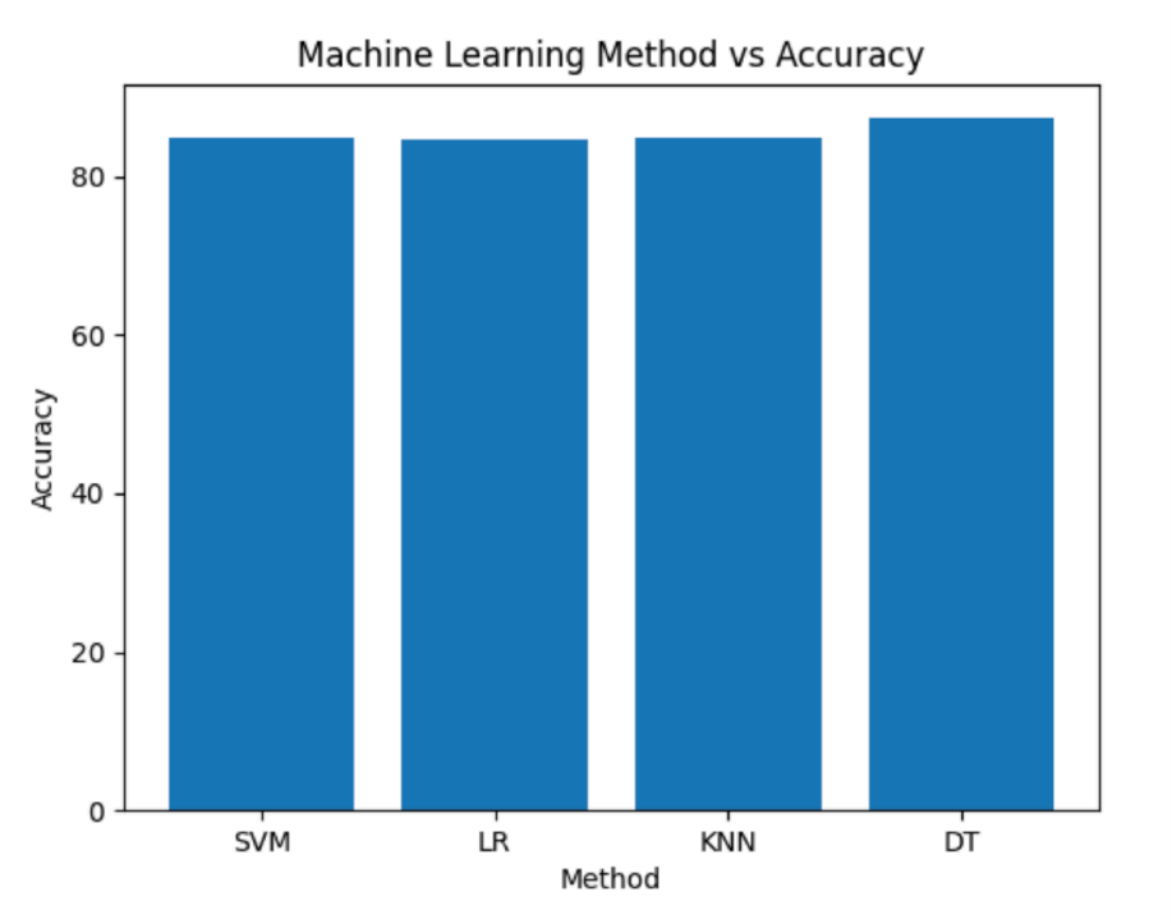
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree has a slightly higher accuracy at 87.32

	ML Method	Accuracy Score (%)
0	SVM	84.821429
1	LR	84.642857
2	KNN	84.821429
3	DT	87.321429

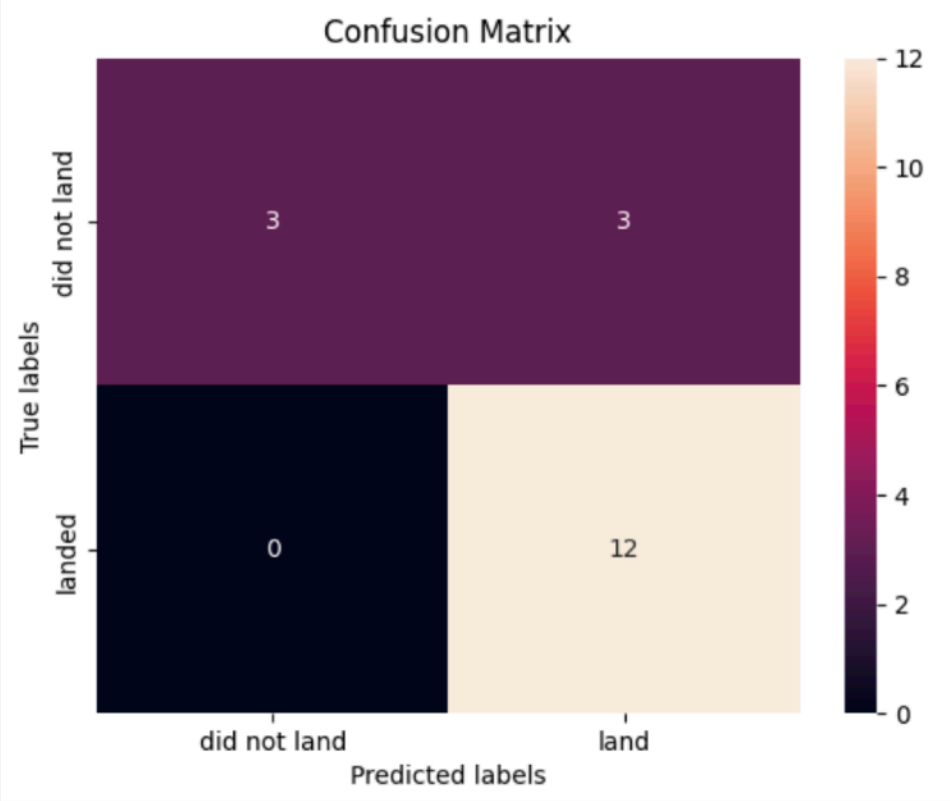


Confusion Matrix

The confusion matrix shows:

- 12 True Positive
- 3 True Negative
- 3 False Positive
- 0 False Negative

The number of false positives is somewhat concerning



Conclusions

- All machine learning algorithms about the same accuracy, with Decision Tree having slightly higher accuracy
- Launch sites are located near highways, railways and coastline and far from cities
- The overall success rate of launches increased with years
- In general, lighter payloads had higher success rate than heavier ones
- Some orbits were more successful like SSO, HEO, GEO, and ES-L1

Appendix

- Wikipedia Falcon 9 launches page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- GitHub Repository: https://github.com/yovan86/Data-Science-Capstone-SpaceX

Thank you!