IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hugh Nyasha Sado
04 June 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection API with Webscraping
  - Data Wrangling
  - Exploratory Analysis with SQL
  - Exploratory Data Analysis for Data Visualization
  - Interactive Visual Analytics and Dashboard
  - Machine Learning Prediction
- Summary of all results
  - Launch success has improved over time. KSC LC-39A has the highest success rate among landing sites. Orbits ES-L1, GEO, HEO, and SSO have a high success rate. Most launch sites are near the equator, and all are close to the coast. All models performed similarly on the test set. The decision tree model

Introduction

- Project background
  - Advertisements found on SpaceX website for Falcon 9 rocket launches show a cost of 62 million dollars, while other providers' cost upward of 165 million dollars each. Much of the difference in savings for SpaceX is because of the reuse of the first stage.
- Problems
  - This project aims to predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data was collected by sending a request and parsing the SpaceX API data using the get request. The response content was a Json format which was decoded using .json() and turned into a Pandas dataframe using .json_normalize ()
- Some basic data wrangling and formatting was done. Dealt with missing values by using the mean and the .replace() function to replace np.nan values in the data.
- Analyzed the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- Explored data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- The data was analyzed using appropriate statistical classification models to identify patterns, trends, and relationships.

Data Collection

- SpaceX API: JSON data converted to pandas DataFrame
- Web Scraping: BeatifulSoup to parse Falcon 9 table from Wikipedia
- Cleaning: Handling missing values, formatting fields

Data Collection – SpaceX API

- Dataset was collected using the SpaceX API, using get requests in Python. Then we normalize the json contents into a dataframe and then using functions and pandas we extract relevant information, clean the data, and export the cleaned data.
- Refer to Notebook with GitHub URL:
  - Applied_Data_Science_Capstone_HNS/jupyter-labs-spacex-data-collection-api.ipynb at main · Hughsazw/Applied_Data_Science_Capstone_HNS

Task 1: Request and parse the SpaceX launch data using the GET request

Task 2: Filter the dataframe to only include Falcon 9 launches

Task 3: Dealing with Missing Values

Data Collection - Scraping

- Web Scraping to collect Falcon 9 historical launch records from a Wikipedia page titled 'List of Falcon 9 and Falcon Heavy launches
- Refer to Notebook with GitHub URL:
  - Applied_Data_Science_Capstone_HNS/jupyter-labs-webscraping (1).ipynb at main · Hughsazw/Applied_Data_Science_Capstone_HNS

TASK 1: Request the Falcon9 Launch Wiki page from its URL

TASK 2: Extract all column/variable names from the HTML table header

TASK 3: Create a data frame by parsing the launch HTML tables

Data Wrangling

- Converted outcomes into 1 for a successful landing and 0 for an unsuccessful landing
- The launch sites, orbit types and mission outcomes were cleaned up and new classification was added to the DataFrame for further analysis
- Used One Hot Encoding to prepare the data to a binary classification
- Refer to Notebook with GitHub URL:
  - [Applied_Data_Science_Capstone_HNS/labs-jupyter-spacex-Data wrangling.ipynb at main · Hughsazw/Applied_Data_Science_Capstone_HNS]

TASK 1: Calculate the number of launches on each site

↓

TASK 2: Calculate the number and occurrence of each orbit

↓

TASK 3: Calculate the number and occurrence of mission outcome of the orbits

↓

TASK 4: Create a landing outcome label from Outcome column

EDA with Data Visualization

- The following charts were plotted to see relationships between parameters
  - Flight number vs Launch Site
  - Payload vs Launch Site
  - Success Rate vs Orbit
  - Launch Success Rate by Year
- Refer to Notebook with GitHub URL:
  - Applied_Data_Science_Capstone_HNS/Exploratory Data Analysis and Feature Engineering.ipynb at main · Hughsazw/Applied_Data_Science_Capstone_HNS

EDA with SQL

- The following SQL queries were performed
    - Listing unique launch Sites
    - Calculating total payload for NASA Launches
    - Finding average payload for booster F(v1.1)
    - Successful and failed missions count
    - First successful landing query
- Refer to Notebook with GitHub URL:
    - Applied_Data_Science_Capstone_HNS/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · Hughsazw/Applied_Data_Science_Capstone_HNS

Build an Interactive Map with Folium

- Created and added the following to folium map
  - Markers for each Launch Site
  - Color-coded outcome markers
- Which would aid in
  - Distance calculations to railways, highways, coastlines etc
    - Regarding Kennedy Space Center: Distance to highways, cities and railways is good enough, also regarding the fact that rockets will head towards the east.
    - Regarding Vandenberg AFB: The only larger place to the East of the AFB is Lompoc, CA.
- Refer to Notebook with GitHub URL:
  - Applied_Data_Science_Capstone_HNS/lab_jupyter_launch_site_location.ipynb at main · Hughsazw/Applied_Data_Science_Capstone_HNS

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to dashboard
  - Pie charts for launch count per sight
  - Scatters plots: Payload vs Landing Outcome
- Plots and interactions can assist by
  - Showing site with the largest successful launches or site with the highest launch success rate.
  - Explaining which payload range(s) has the highest or lowest launch success rate.
  - Visualizing the F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) that has the highest launch success rate.
- Refer to Notebook with GitHub URL:
  - Hughsazw/Applied_Data_Science_Capstone_HNS

Predictive Analysis (Classification)

- Models built include:
  - Logistic Regression
  - Decision Tree (Best Accuracy)
  - KNN, SVM
  - Tuned with GridSearchCV
- Refer to Notebook with GitHub URL:
  - [Applied_Data_Science_Capstone_HNS/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · Hughsazw/Applied_Data_Science_Capstone_HNS]()

- TASK 1: Created a NumPy array from the column Class in data, by applying the method to_numpy() then assign it to the variable Y
- TASK 2: Standardized the data in X then reassign it to the variable X using the transform -- transform=preprocessing. StandardScaler()X=transform.fit_transform(X)--
- TASK 3: Used the function train_test_split to split the data X and Y into training and test data.
- TASK 4: Created a logistic regression object then create a GridSearchCV object logreg_cv with cv = 10
- TASK 5: Calculated the accuracy on the test data using the method score:
- TASK 6: Created a support vector machine object then create a GridSearchCV object svm_cv with cv = 10.
- TASK 7: Calculated the accuracy on the test data using the method score
- TASK 8: Create a decision tree classifier object then create a GridSearchCV object tree_cv with cv = 10
- TASK 9: Calculated the accuracy of tree_cv on the test data using the method score
- TASK 10: Created a k nearest neighbors object then create a GridSearchCV object knn_cv with cv = 10.
- TASK 11: Calculated ` the accuracy of knn_cv on the test data using the method score
- TASK 12: Compared the methods to see which performs best:

Results

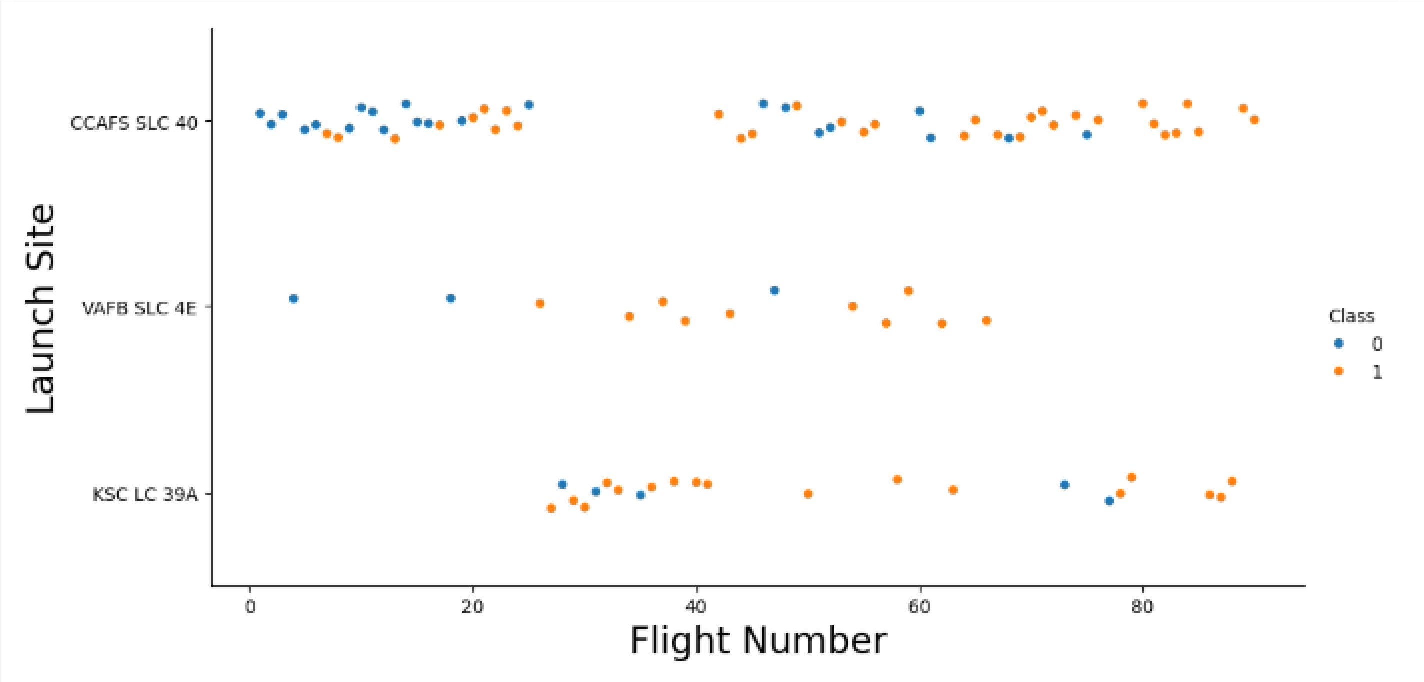
- SpaceX's Falcon 9 first stage landing outcomes have been trending towards greater success as more launches are made.
- The machine learning models can be used to predict future SpaceX Falcon 9 first stage landing outcomes
- Payloads over 8000kg have high success rate
- Space X uses 4 different launch sites
- VLEO orbit has 14 launches and 85% success rate
- With booster F9, almost every mission outcome was successful.
- Around 70 landing outcomes were successful, while there were 22 no attempts, and around 10 failed.

Section 2

Insights drawn from EDA

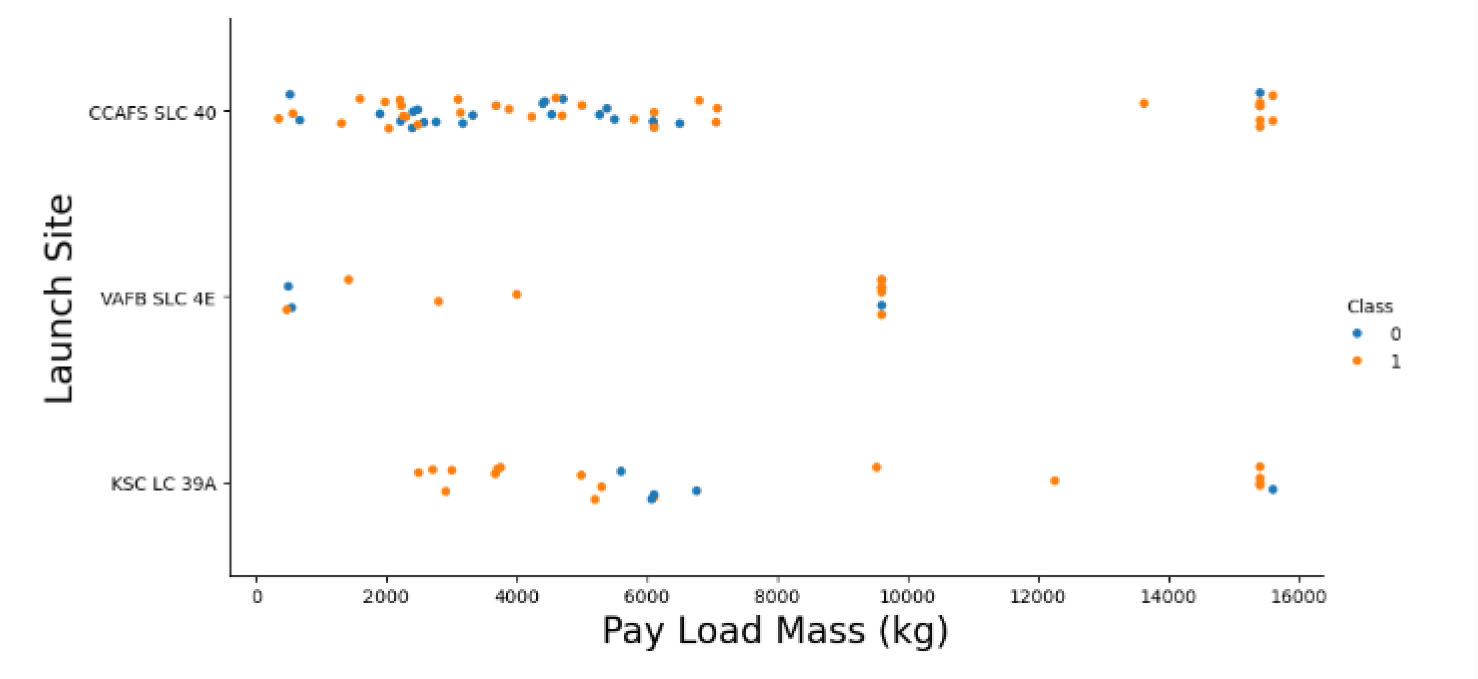
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site



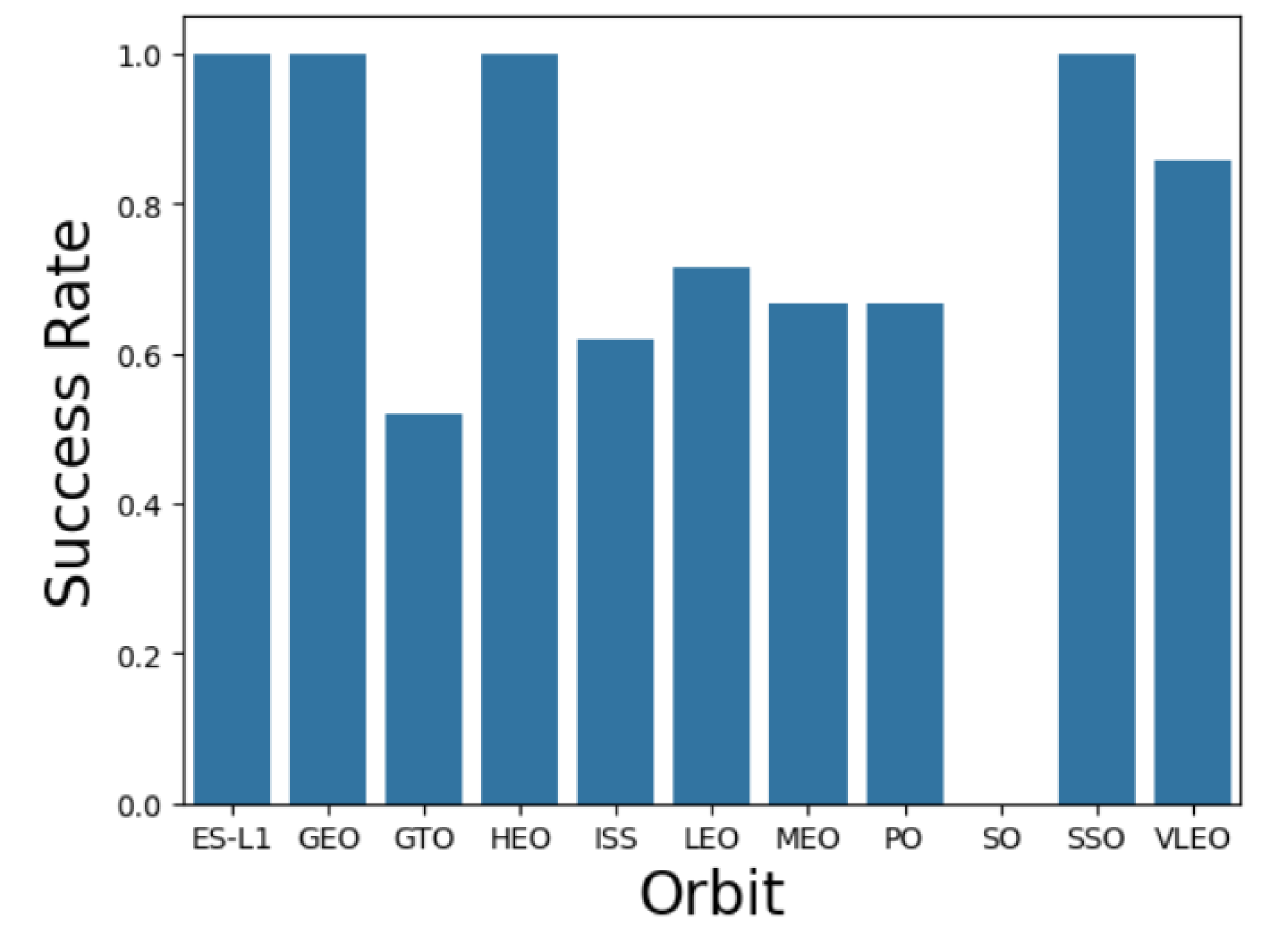
Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site



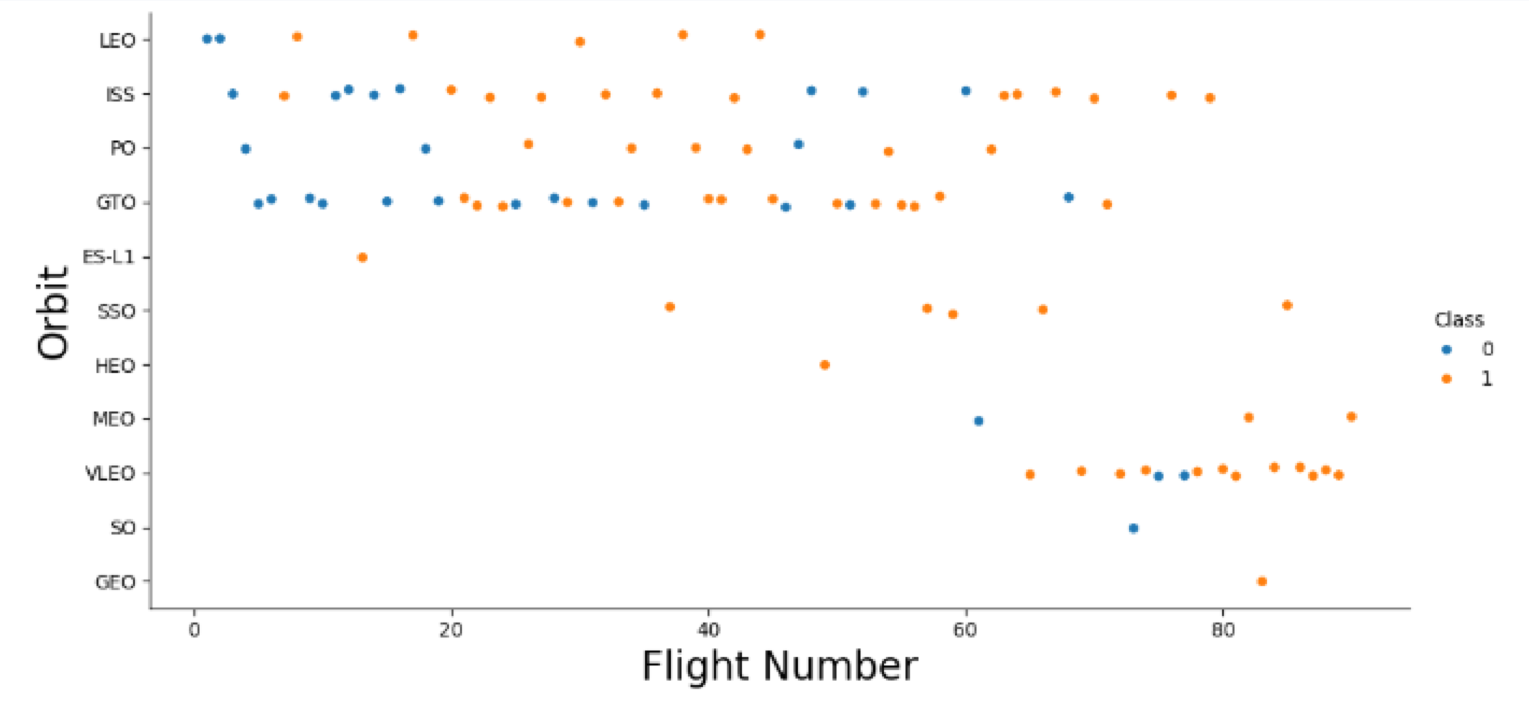
Success Rate vs. Orbit Type

- Bar chart for the success rate of each orbit type



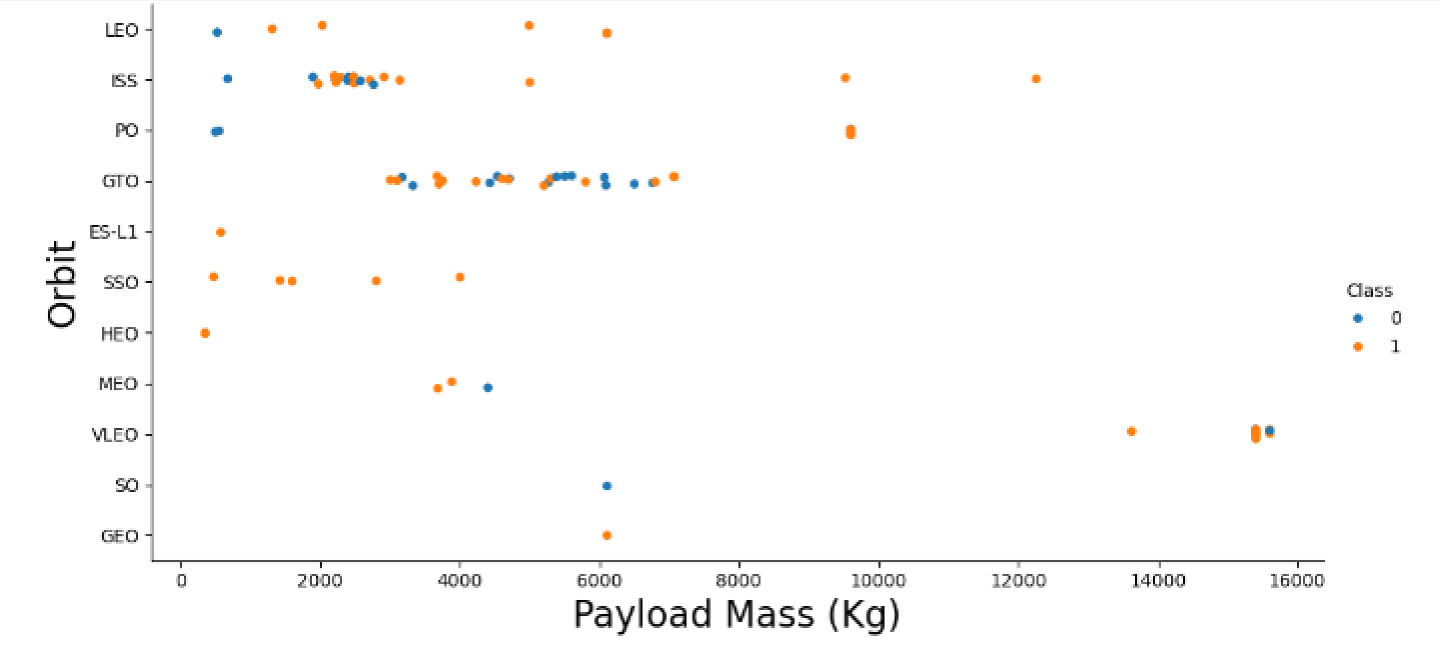
Flight Number vs. Orbit Type

- Scatter plot of Flight number vs. Orbit type



Payload vs. Orbit Type

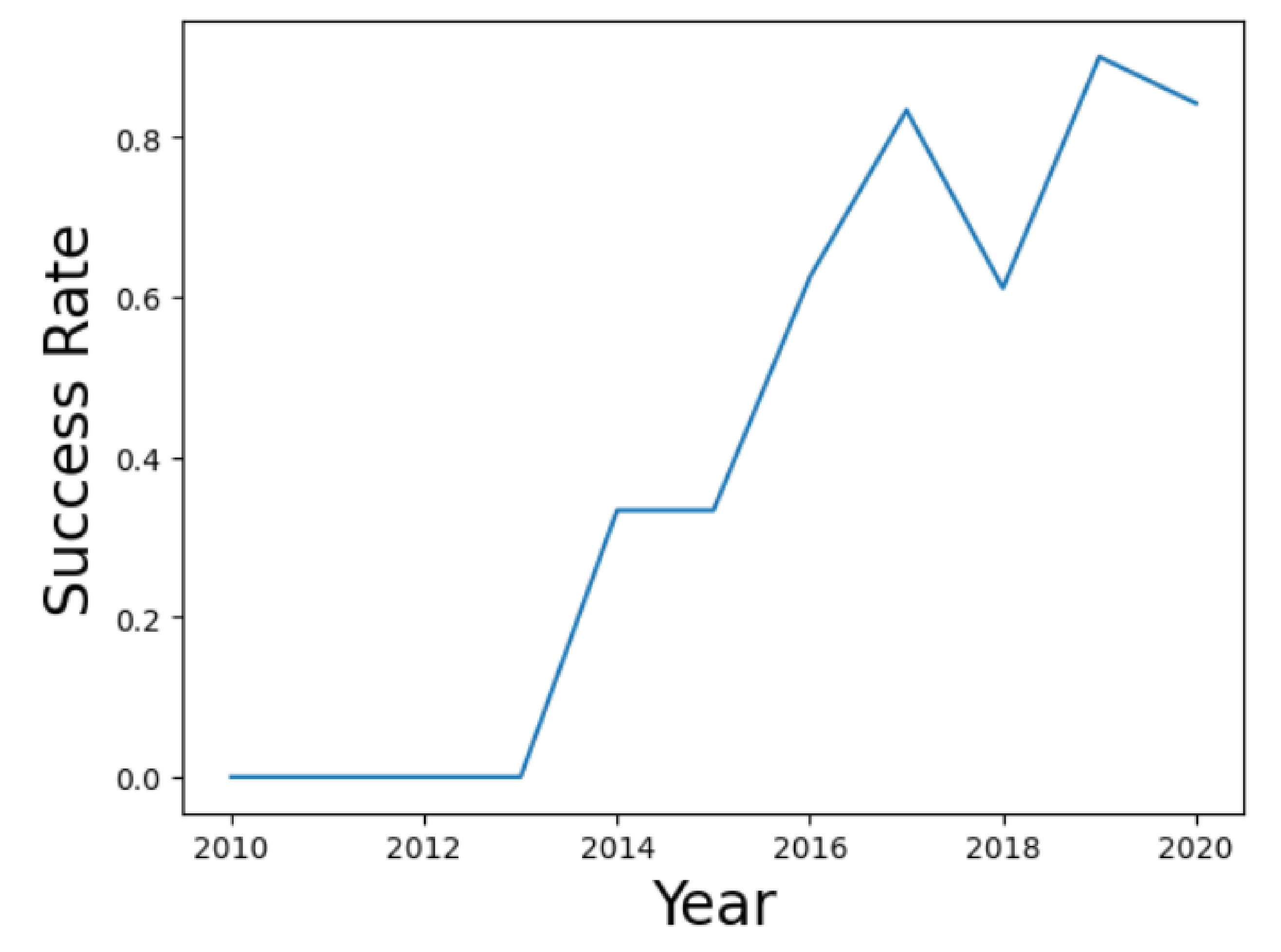
- Scatter point of payload vs. orbit type



Launch Success Yearly Trend

- Line chart of yearly average success rate

- Line plot to see mission outcome trend by year.



All Launch Site Names

- All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
Out[10]:
```

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

Launch Site Names Begin with 'CCA'

- Launch Site Names Begin with 'CCA'

In [11]:
```
%sql SELECT* FROM SPACEXTABLE WHERE Launch_site LIKE 'CCA%' LIMIT 5;
```

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

Out[11]:

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

- Total Payload Mass

```
In [12]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';

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

Out[12]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average Payload Mass by F9 v1.1

```
In [13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE (Booster_Version = 'F9 v1.1') ;

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

Out[13]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- First Successful Ground Landing Date

```
In [14]: %sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE (Landing_Outcome = 'Success (ground pad)') ;

 * sqlite:///my_data1.db
Done.

Out[14]:
```

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Successful Drone Ship Landing with Payload between 4000 and 6000

In [15]:

```
%sql SELECT* FROM SPACEXTABLE WHERE (Landing_Outcome = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

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

Out[15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Total Number of Successful and Failure Mission Outcomes

In [16]:
```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE WHERE Mission_Outcome IN ('Success', 'Failure (in fligh
```

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

Out[16]:

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98

Boosters Carried Maximum Payload

- F9 Boosters that carried maximum payload

```
In [17]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

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

Out[17]:

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

2015 Launch Records

- Records for 2015 Launches

```
In [18]: %sql SELECT substr(Date,6,2) as Month, Date,Booster_Version, Launch_Site, [Landing_Outcome] FROM SPACEXTABLE WHERE [Landing_
```

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

Out[18]:

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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

In [19]:
```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes FROM SPACEXTBL WHERE DATE between '2010-06-04 ' and '2017-03-20' g
```

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

Out[19]:

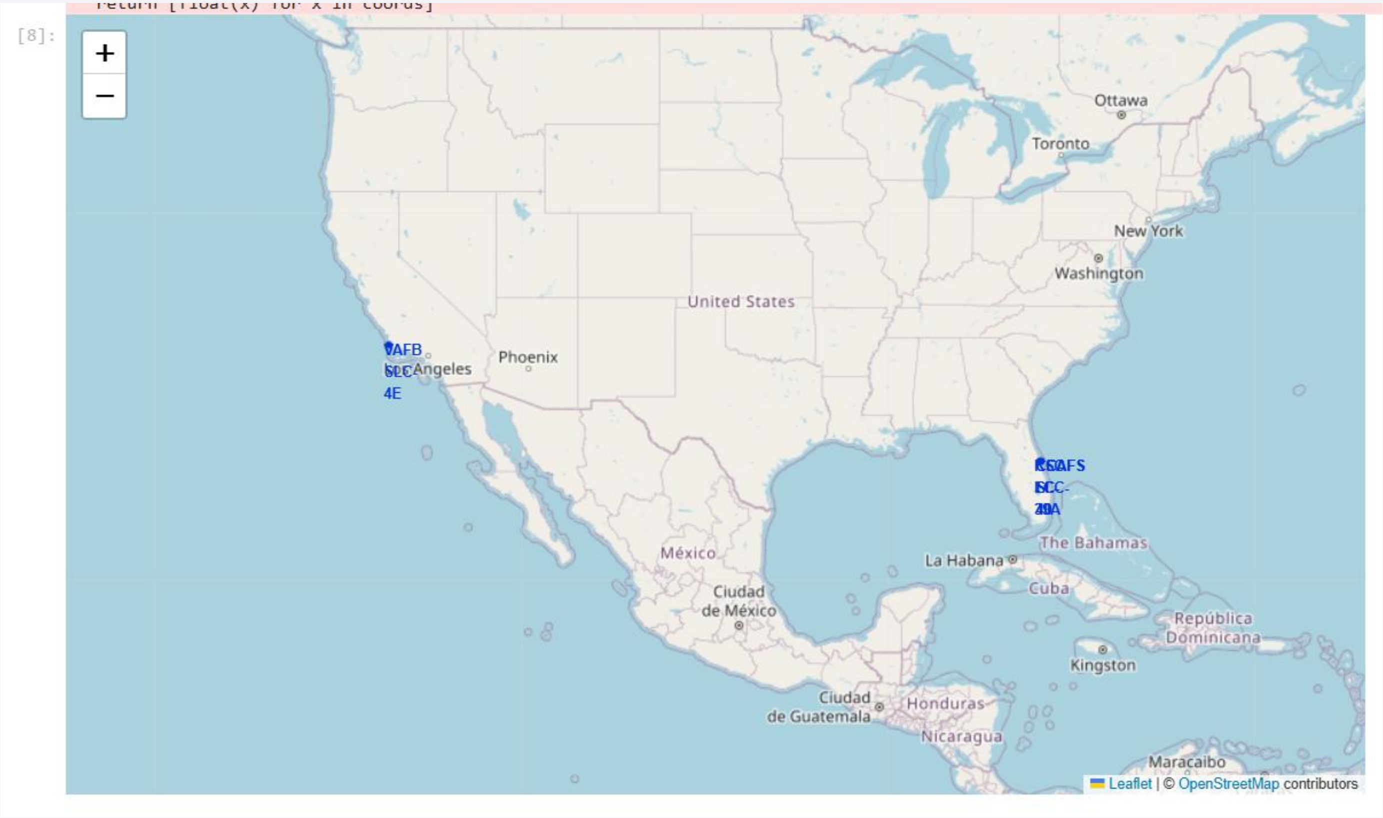
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

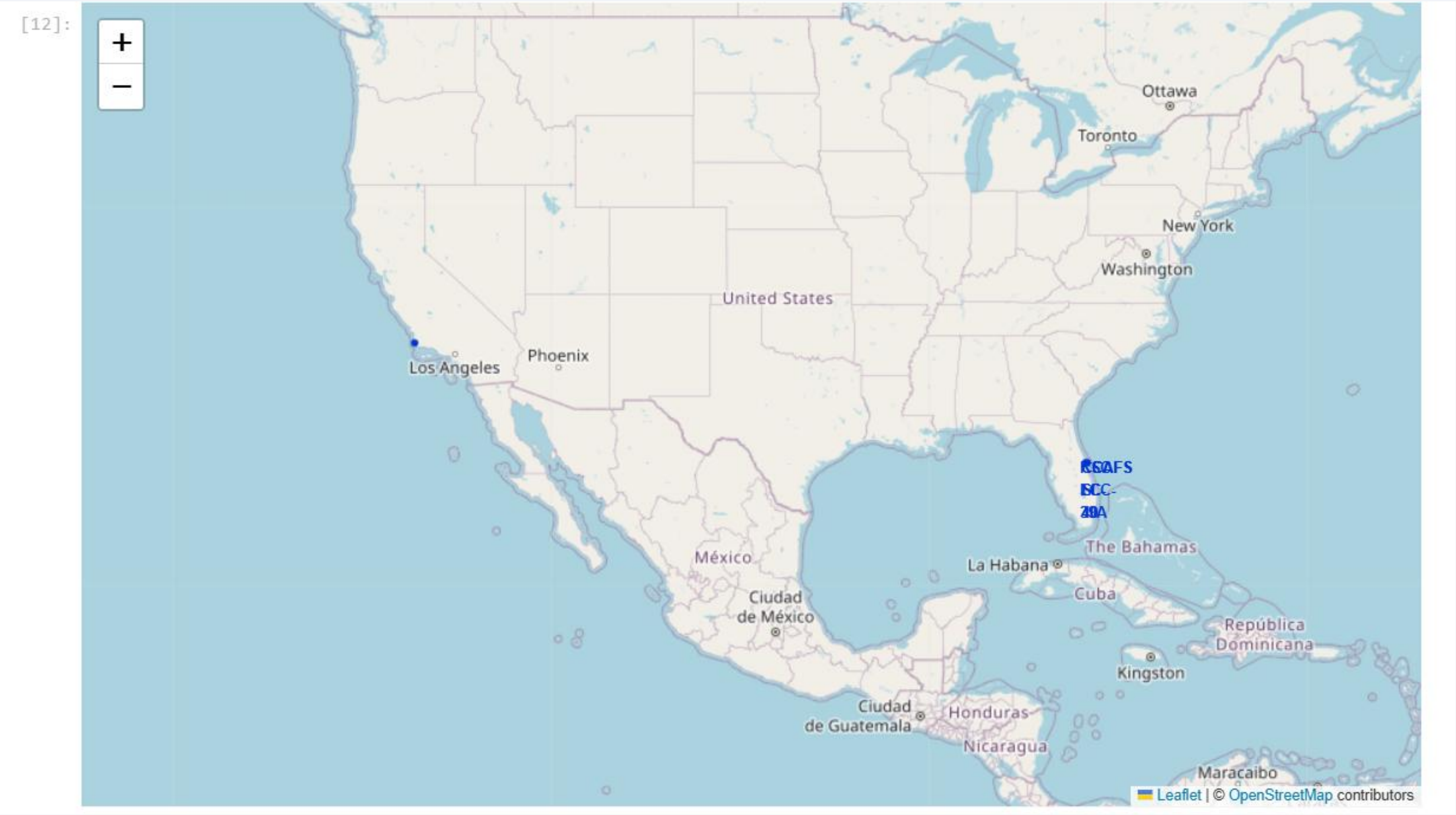
Launch sites' location markers on a global map

- Launch sites' location markers on a global map



Color-labeled launch outcomes

- Color-labeled launch outcomes



Launch site location to its proximities

- Kennedy Space Center: Distance to highways, cities and railways is good enough, also regarding the fact that rockets will head towards the east.

- Vandenberg AFB: The only larger place to the East of the AFB is Lompoc, CA.

Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title

- Show the screenshot of the piechart for the launch site with highest launch success ratio

- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

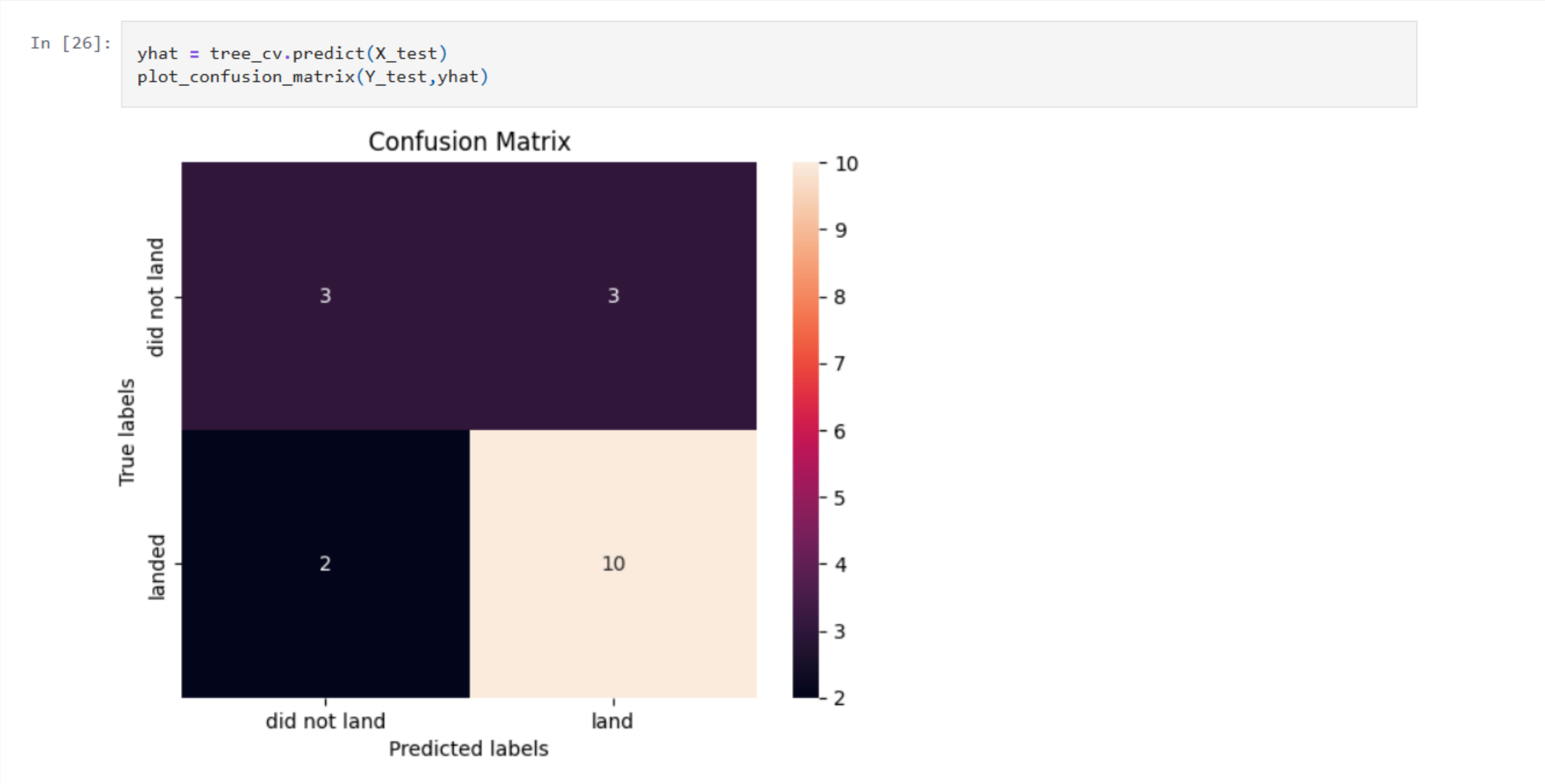
- Bars for Lr , SVM and KNN accuracy would be higher than the decision tree on when ploted with x=probability from 0-1 and y=accuracy index

```
In [32]:
#After comparing accuracy of above methods, they all preformed practically the same.
#Exception was the tree which fit train data slightly better but test data worse.
scores = [lr_score,svm_score,tree_score,knn_score]
print(scores)
#print(scores.index(max(scores)))

[0.8333333333333334, 0.8333333333333334, 0.7222222222222222, 0.8333333333333334]
```

Confusion Matrix

- Decision Tree confusion matrix was the best performing model



Conclusions

- Larger number of flights correlate with higher success rate
- Launch success rate increases steadlily since 2013
- Orbits ES-L1, GEO, HEO, SSO, VLEO had highest success
- Launch Site KSC LC-39A had most successful launches
- Decision tree classifier was best performing Machine learning model

Thank you!