IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luis Payán
December 6, 2021

Outline

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

Executive Summary

- Summary of methodologies
    1. Data Collection API
    2. Data Collection with Web Scrapping
    3. Data wrangling EDA
    4. EDA with SQL
    5. EDA visualization
    6. Interactive Visual Analytics with Folium
    7. Machine learning
- Summary of all results
    - Exploratory Data Analysis and visualization
    - Screenshots of Interactive analytics dashboard
    - Machine Learning prediction results

Introduction

- Project background and context

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

- Problems you want to find answers

1. What variables define if the rocket will land successfully?
2. How are variables related with the success rate of landing?
3. What prescription of variables does SpaceX have to achieve to ensure the reuse of the first stage of the rocket?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - I worked with SpaceX launch data that was gathered from the SpaceX REST API.
  - Using the Python BeautifulSoup package, I web scraped HTML tables from Wikipedia that contain Falcon 9 launch records.
- Perform data wrangling
  - I performed Exploratory Data Analysis to find some patterns in the data and determine what would be the label for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Several types of plots were used to show relationships between variables and quantitative performance for some particular scenarios were evaluated by using SQL scripts.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The best hyperparameters for SVM, Classification Trees and Logistic Regression were determined through grid search methods.

Data Collection

- I worked with SpaceX launch data gathered from the SpaceX REST API.

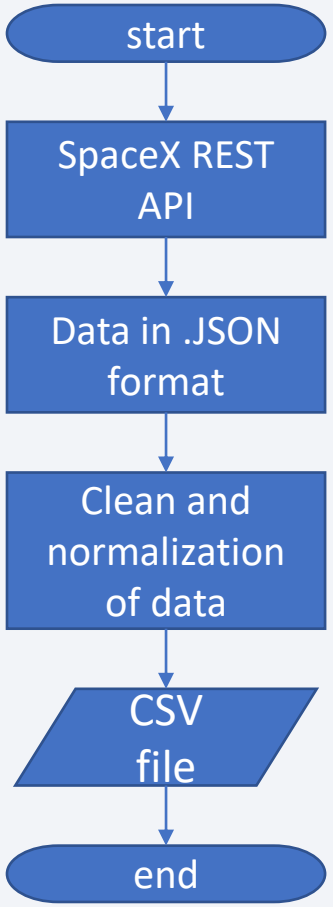
- Another data source that I used for obtaining Falcon 9 Launch data was Wikipedia.
- For web scraping from Wikipedia, I used the BeautifulSoup package.

Data Collection – SpaceX API

This API provides information related to launches, including the rocket used, payload mass delivered, launch and landing specifications, as well as the landing outcome.

- GitHub URL:

https://github.com/payanrg/testrepo_capstone/blob/62a0064a72b12626d2cac26502ce2d40c63f3ba0/Data%20Collection%20API%20Lab.ipynb



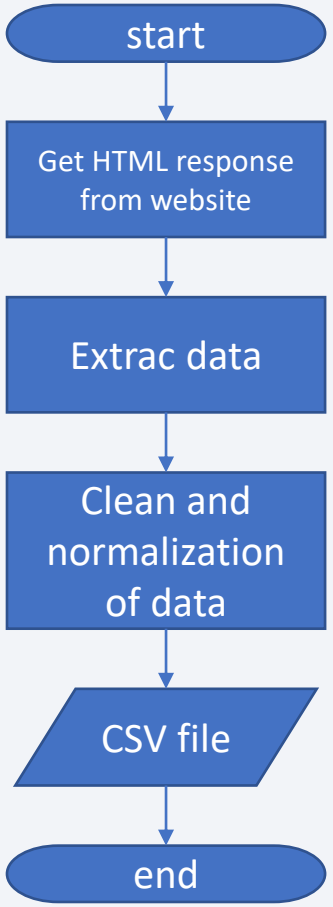
Data Collection - Scraping

Web scrap Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

GitHub URL:

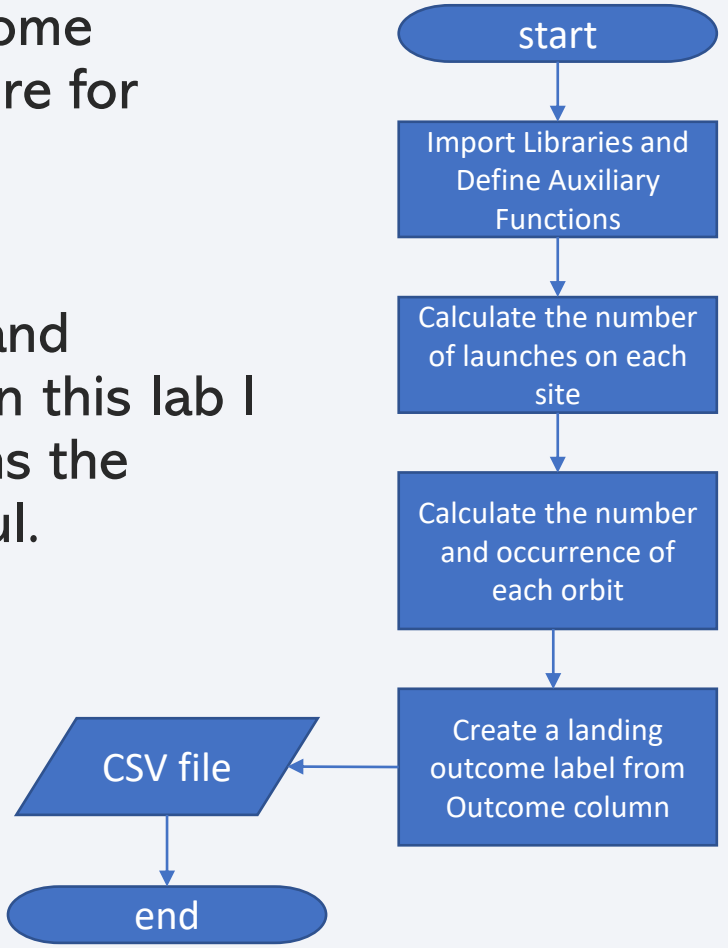
https://github.com/payanrg/testrepo_capstone/blob/62a0064a72b12626d2cac26502ce2d40c63f3ba0/Data_Collection_with_Web_Scraping_lab.ipynb



Data Wrangling

- I performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the feature for training supervised models.

- In the data set, there are cases where the booster did not land successfully, but other where the landing was successfully. In this lab I converted those outcomes into Training Labels with 1 means the booster successfully landed and 0 means it was unsuccessful.



GitHub URL:

https://github.com/payanrg/testrepo_capstone/blob/62a0064a72b12626d2cac26502ce2d40c63f3ba0/Data%20wrangling%20EDA%20lab.ipynb

EDA with Data Visualization

- Scatter plots to visualize the relation between variables
  - FlightNumber vs. PayloadMass
  - FlightNumber vs LaunchSite
  - Payload vs Launch Site
  - Flight Number vs Orbit type
- Bar plot to count success outcomes
  - Success rate of each orbit type

GitHub URL:

https://github.com/payanrg/testrepo_capstone/blob/62a0064a72b12626d2cac26502ce2d40c63f3ba0/%20EDA_Visualization_lab.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the 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
- 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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

GitHub URL:

https://github.com/payanrg/testrepo_capstone/blob/62a0064a72b12626d2cac26502ce2d40c63f3ba0/EDA_with_SQL_lab.ipynb

Build an Interactive Map with Folium

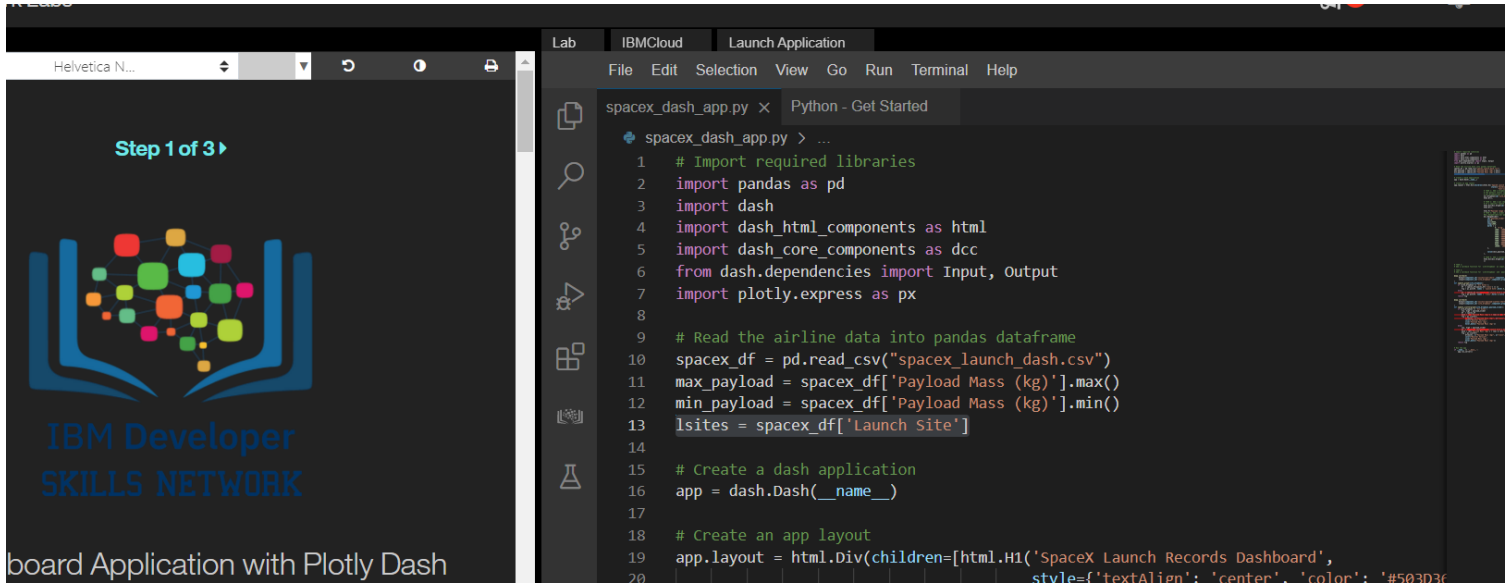
- Markers for all launch sites on a map
- Markers the success/failed launches for each site on the map
- Distances between a launch site to its proximities
- Some geographical patterns about launch sites were found.

GitHub URL:

https://github.com/payanrg/testrepo_capstone/blob/257411058a6292d9afbba6515dbf9c5c8da81477/Interactive_Visual_Analytics_Folium_lab.ipynb

Alternative link:
https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/aba2ce61-36e8-40ae-a85c-e9dcdab5435f/view?access_token=2df85ec25e7e3ffadf2eab02e097c305efdca193cbf0094264692b436f1572ca

Build a Dashboard with Plotly Dash



Pie chart: total success launches by site

Scatter plot: correlation between payload and success for all sites

https://github.com/payanrg/testrepo_capstone/blob/7c76da4ed7278b0b3ce82a8eacfe7b9fd3b0e5d4/dashboard_plotly.py

Predictive Analysis (Classification)

- Model building
- Load dataset and transform with NumPy and Pandas methods
- Split data into training and test data sets
- Set hyperparameters for GridSearchCV
- Fit datasets into the GridSearchCV objects and train.
- Check accuracy for each model using test data
- Get tuned hyperparameters for each type of algorithms
- Plot Confusion Matrix
- The model with the best accuracy score was Decision Tree

GitHub URL:

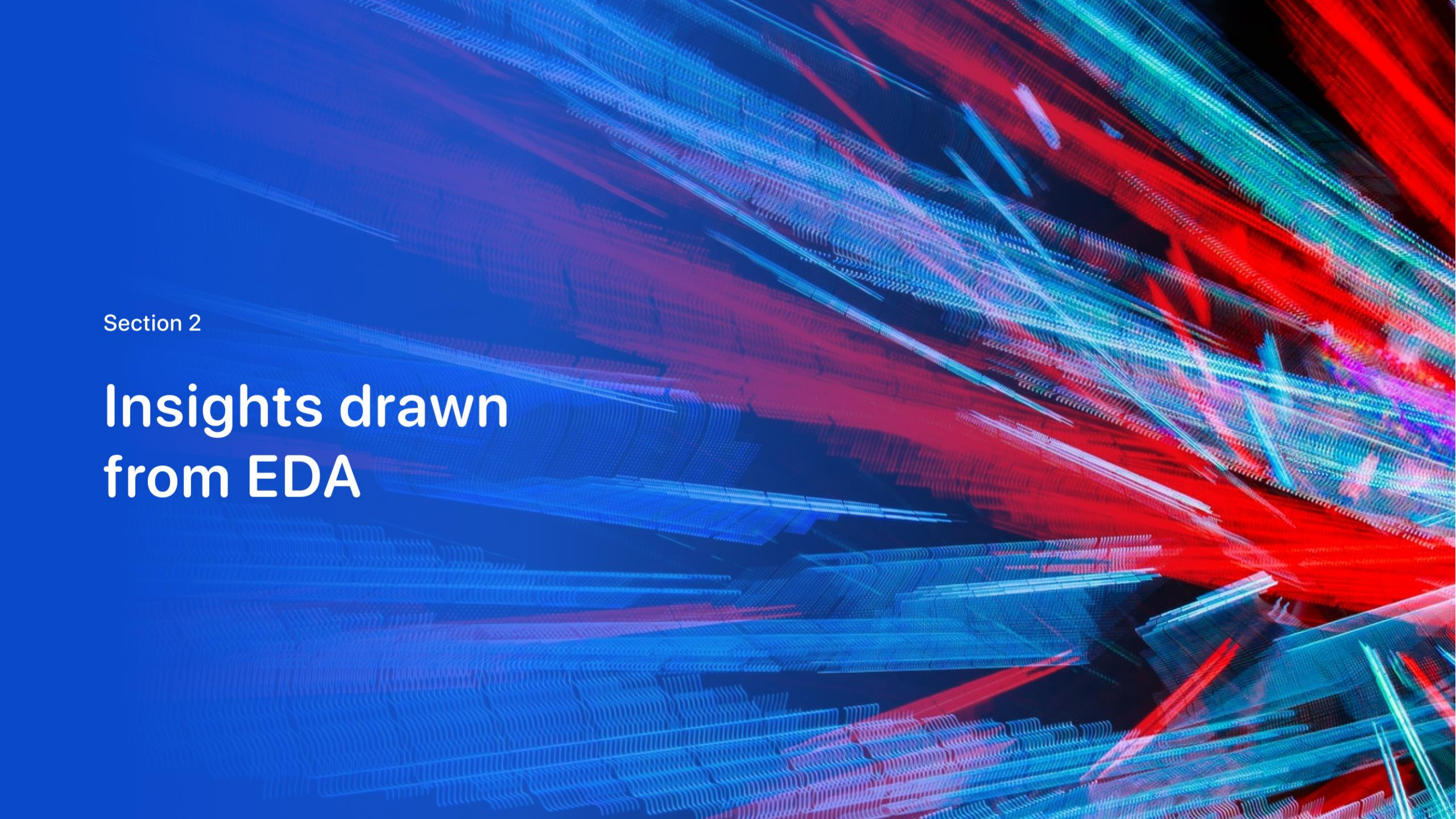
https://github.com/payanrg/testrepo_capstone/blob/62a0064a72b12626d2cac26502ce2d40c63f3ba0/Machine_Learning_Prediction_lab.ipynb

Results

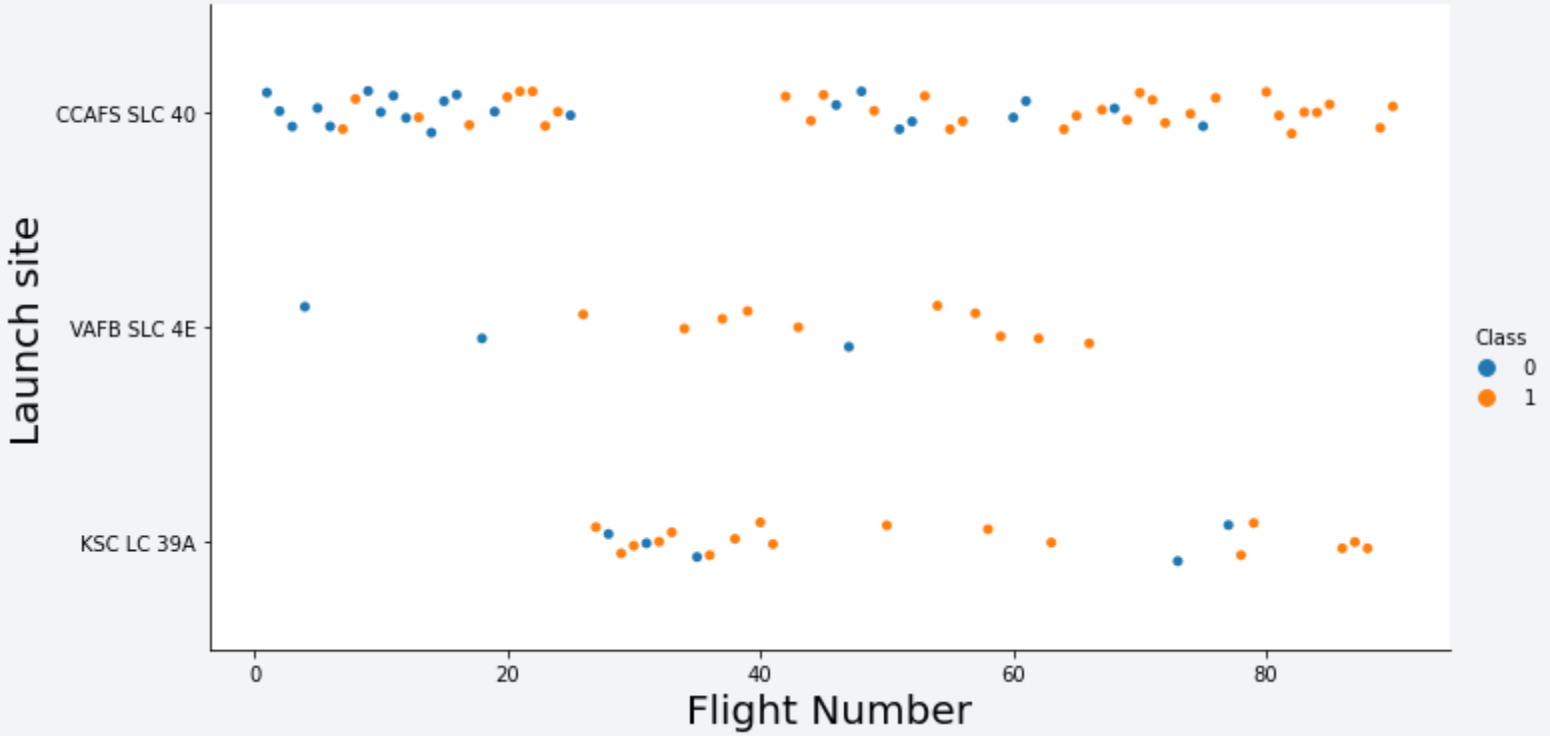
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

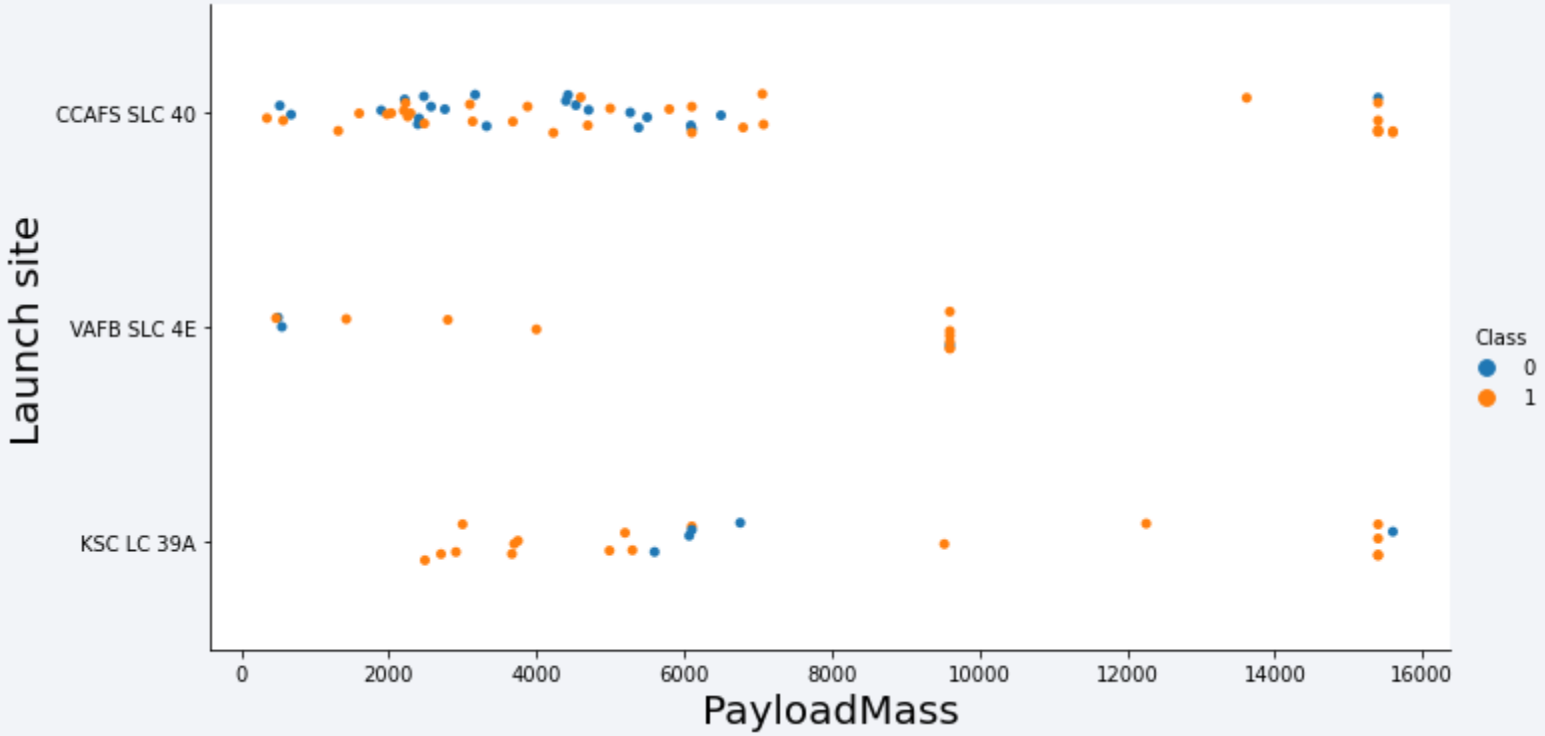


Flight Number vs. Launch Site



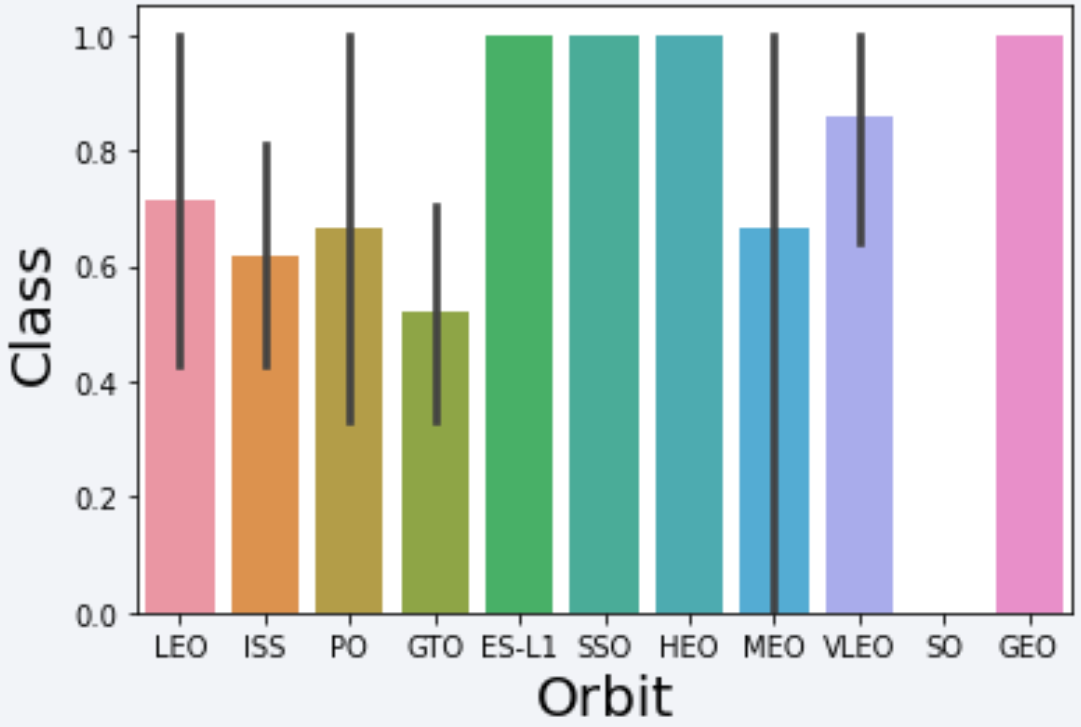
- CCAFS SLC 40 has the greater number of launches, so the success rate is greater also

Payload vs. Launch Site



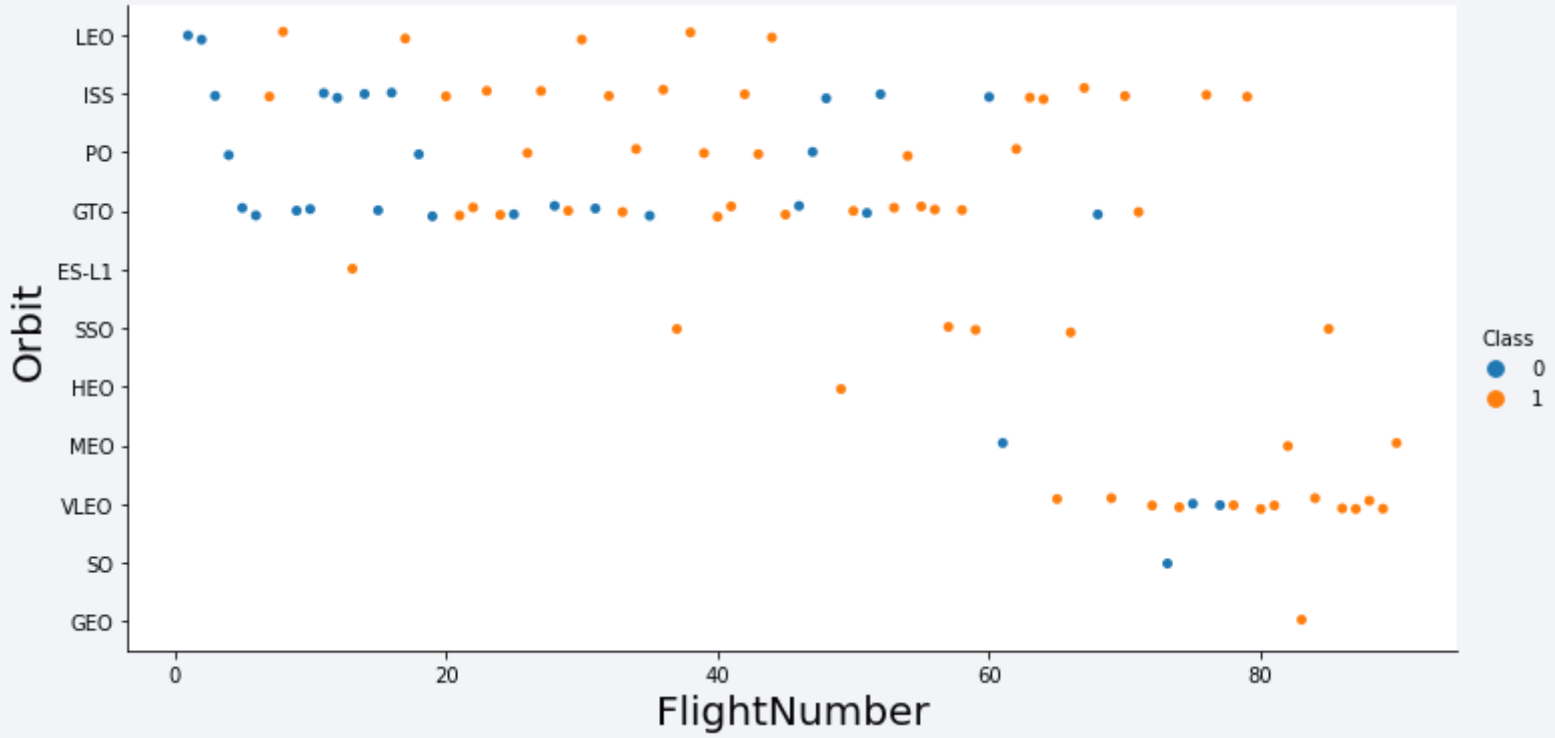
- CCAFS SLC 40 has the greater number of launches but also almost all launches has less than 7000 kg of payload mass

Success Rate vs. Orbit Type



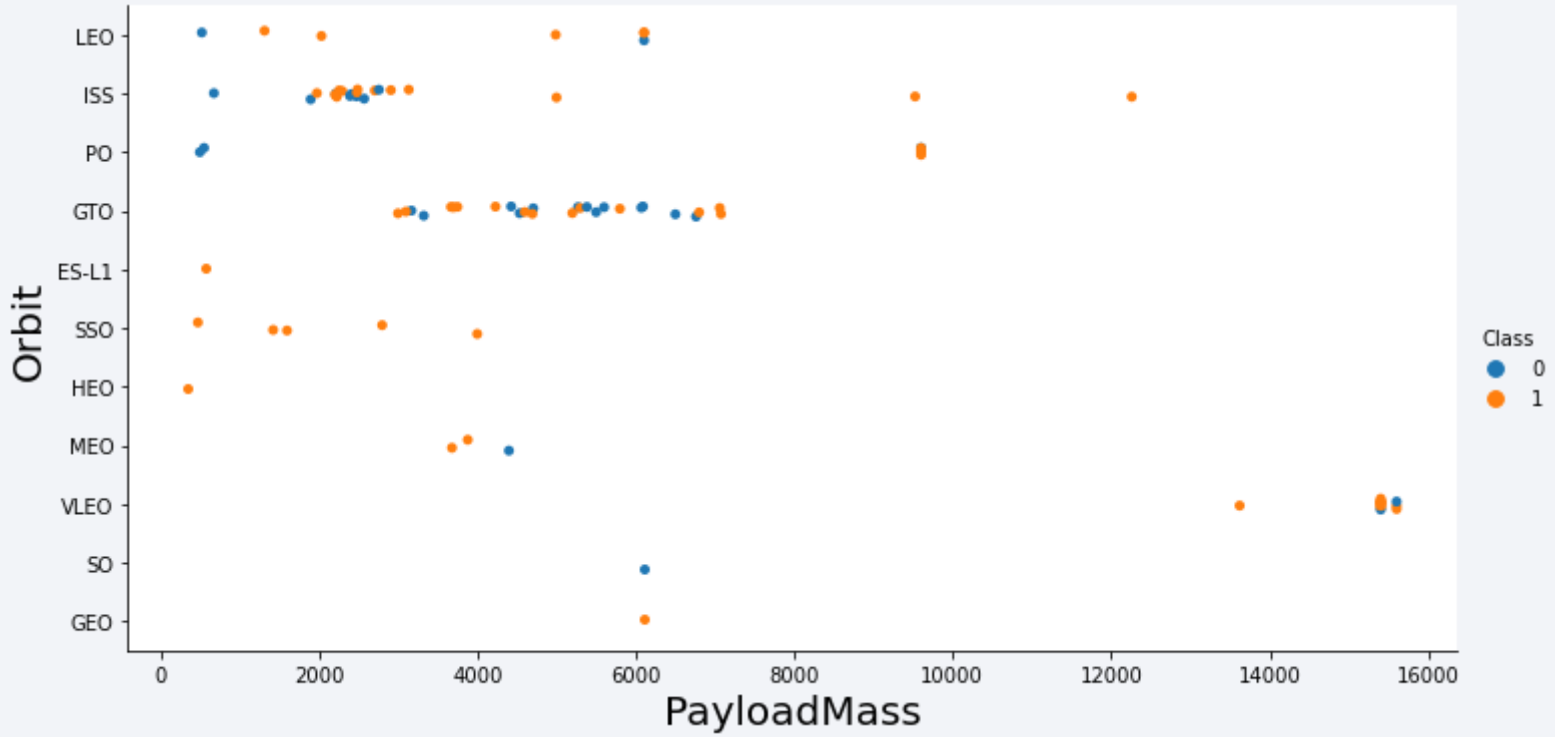
- Orbits GEO, HEO, SSO, ES-L1 have the best success rate

Flight Number vs. Orbit Type



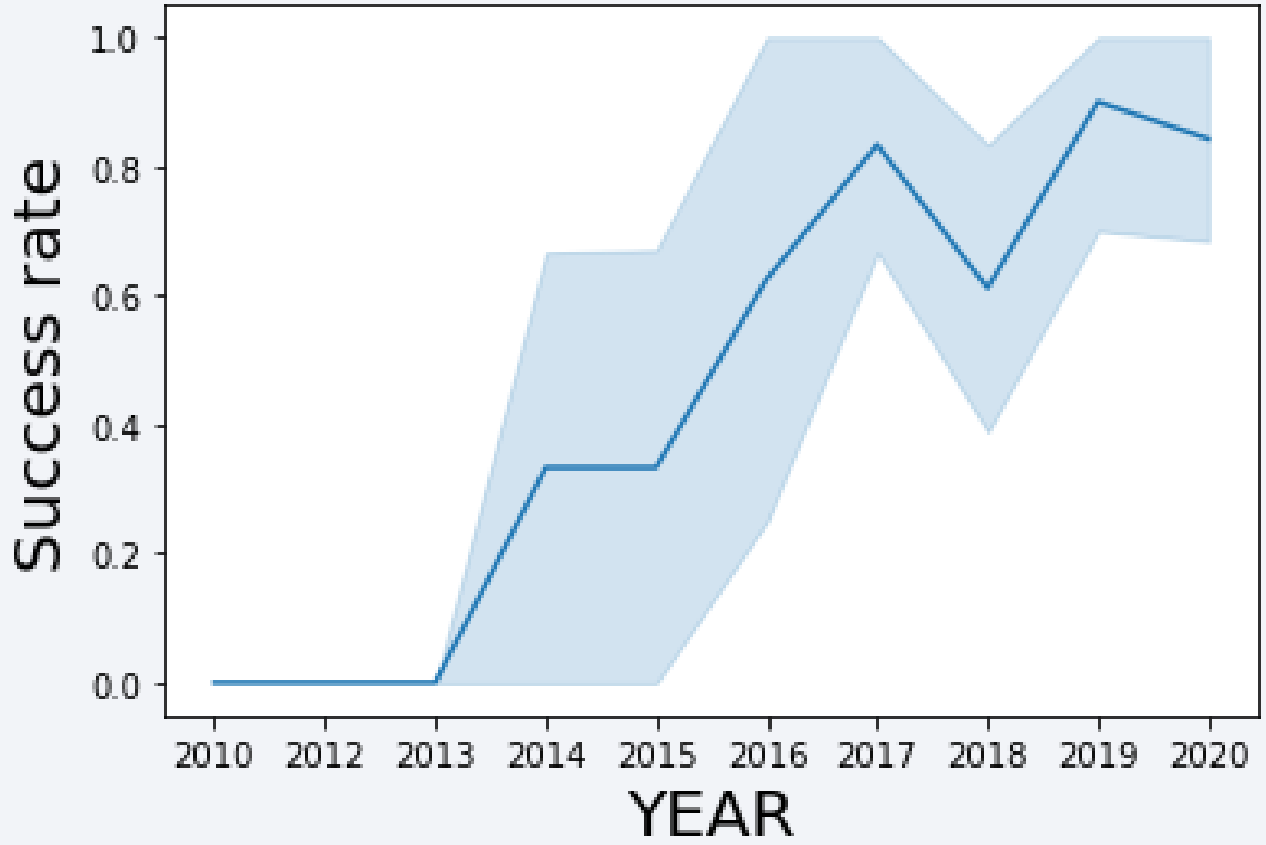
- GTO orbit shows more unsuccessful flight and VLEO shows good success rate

Payload vs. Orbit Type



- VLEO orbit shows the higher payload mass, GTO shows payload mass in the interval of 3000 to 7000 kg with a mixture of success and unsuccessful landings

Launch Success Yearly Trend



- Success rate increases significantly since 2013 with slightly drop between 2019 and 2020

All Launch Site Names

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

```
# ok
%sql select distinct LAUNCH_SITE from SPACEXTB;
```

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA', I used where and limit instructions to carry on this task

```
# ok
%sql SELECT LAUNCH_SITE from SPACEXTB where LAUNCH_SITE like 'CCA%' LIMIT 5;
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS). For this task, a used the instruction SUM with a condition structured with the instruction WHERE

```
# ok
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTB where CUSTOMER ='NASA (CRS)';
```

1
45596

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1. A combination of AVG and WHERE instructions were used to obtain the average payload mass for a specific booster version.

```
# ok
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTB where Booster_Version = 'F9 v1.1' ;
```

1
2928

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was achieved. The query was completed using min(DATE) to obtain the specific date of an event from a list.

```
# OK
%sql select min(DATE) from SPACEXTB where Landing__Outcome like 'Success (ground pad)';
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
# OK
%sql select BOOSTER_VERSION from SPACEXTB where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes. COUNT and GROUP BY, were the instructions used in this query to obtain the list.

```
# OK
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME)  from SPACEXTB GROUP BY MISSION_OUTCOME;
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
# OK
%sql select BOOSTER_VERSION from SPACEXTB where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTB);
```

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

List the names of the booster_versions which have carried the maximum payload mass. A subquery was used to obtain the list of names.

2015 Launch Records

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015. A date conditioned query was scripted for this task.

```
%sql SELECT DATE, LANDING__OUTCOME, Booster_Version, Launch_Site from SPACEXTB where LANDING__OUTCOME='Failure (drone ship)' and YEAR(DATE) = 2015
```

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

```
%sql SELECT LANDING__OUTCOME FROM SPACEXTB WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```

landing__outcome
No attempt
Success (ground pad)
Success (drone ship)
Success (drone ship)
Success (ground pad)
Failure (drone ship)
Success (drone ship)
Success (drone ship)
Success (drone ship)
Failure (drone ship)
Failure (drone ship)
Success (ground pad)
Precluded (drone ship)
No attempt

Failure (drone ship)
No attempt
Controlled (ocean)
Failure (drone ship)
Uncontrolled (ocean)
No attempt
No attempt
Controlled (ocean)
Controlled (ocean)
No attempt
No attempt
Uncontrolled (ocean)
No attempt
No attempt
No attempt
Failure (parachute)
Failure (parachute)

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.

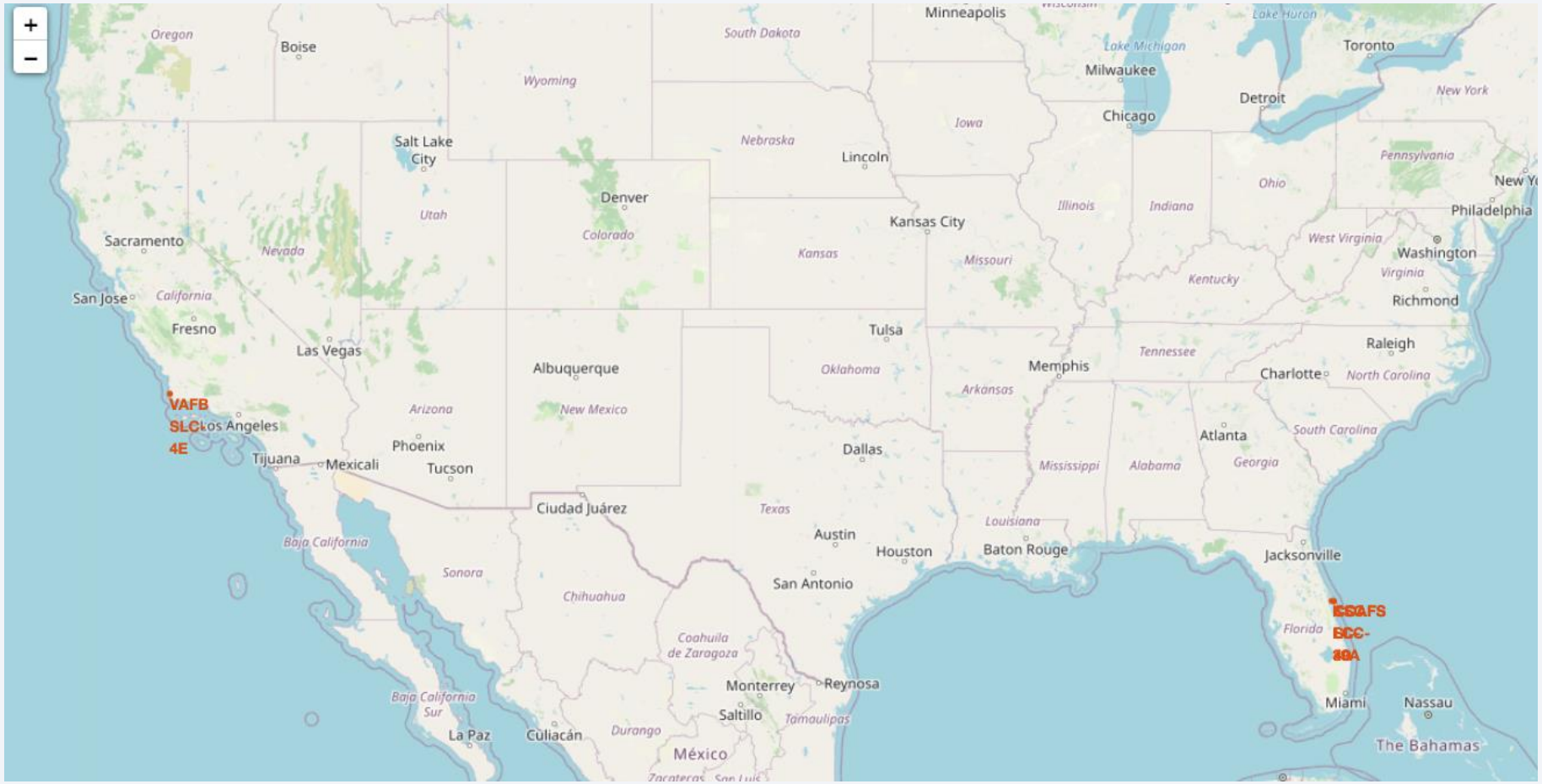
ORDER BY DATE DESC were the main instructions used to obtain this query.

Section 4

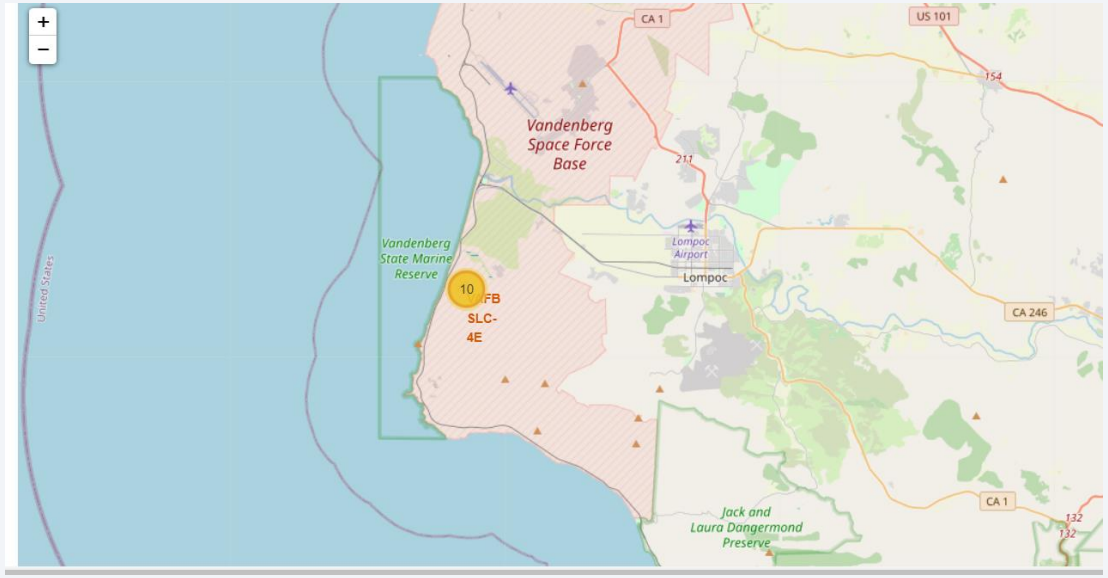
Launch Sites Proximities Analysis

Launch sites global map markers

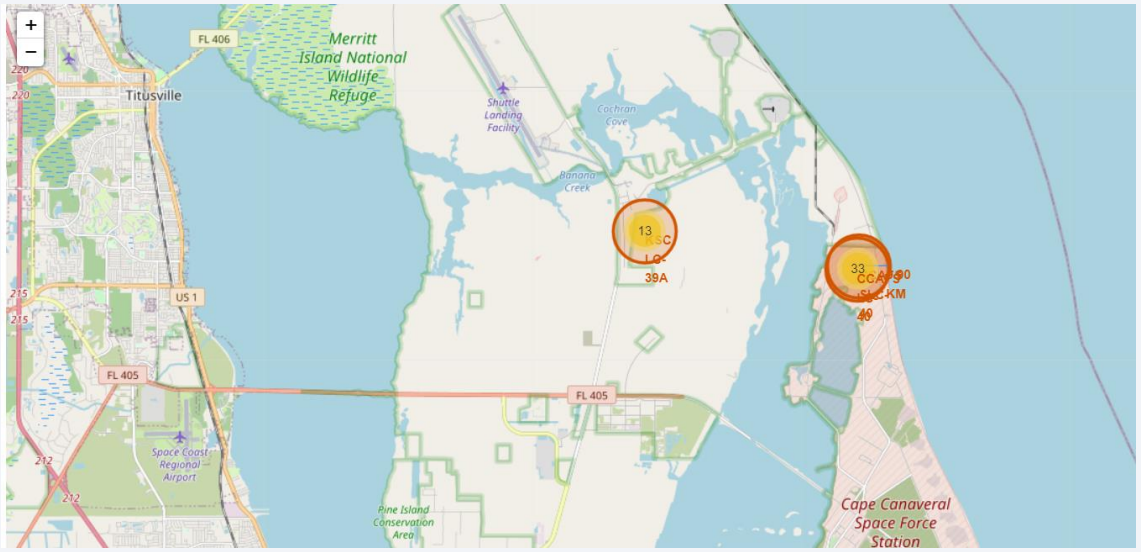
SpaceX launch sites are located around the United States of America coasts. Specifically in Florida and California.



Launch sites

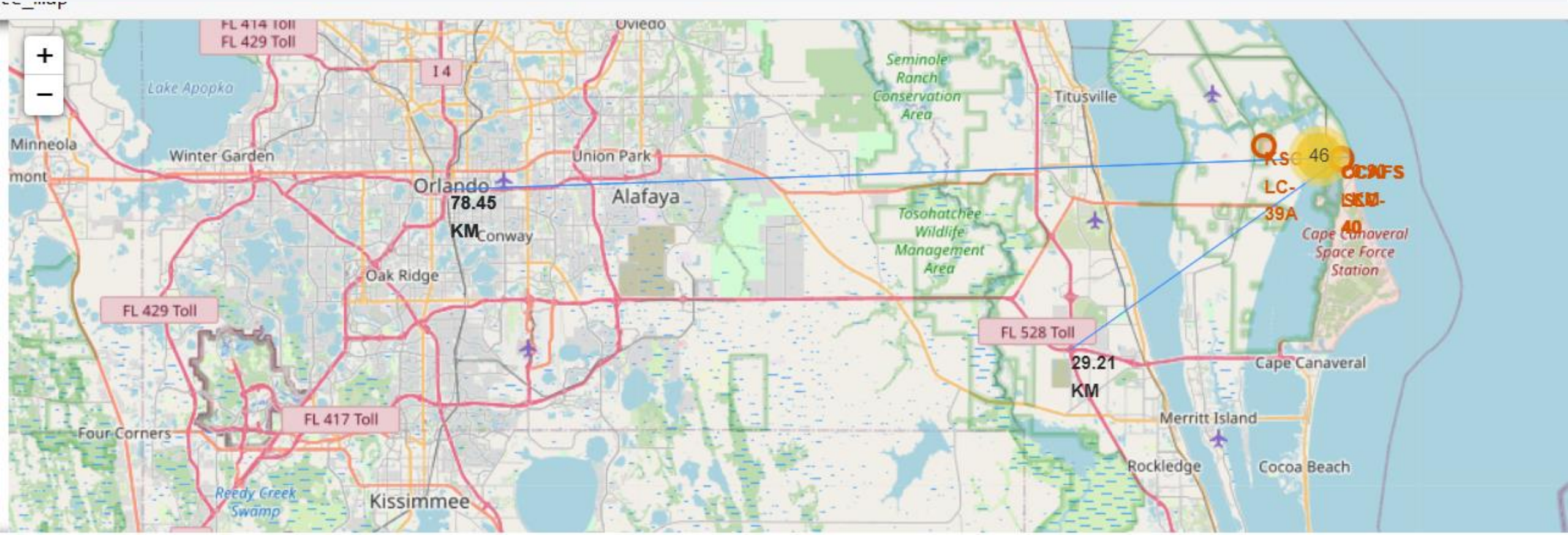


California launch sites



Florida Launch sites

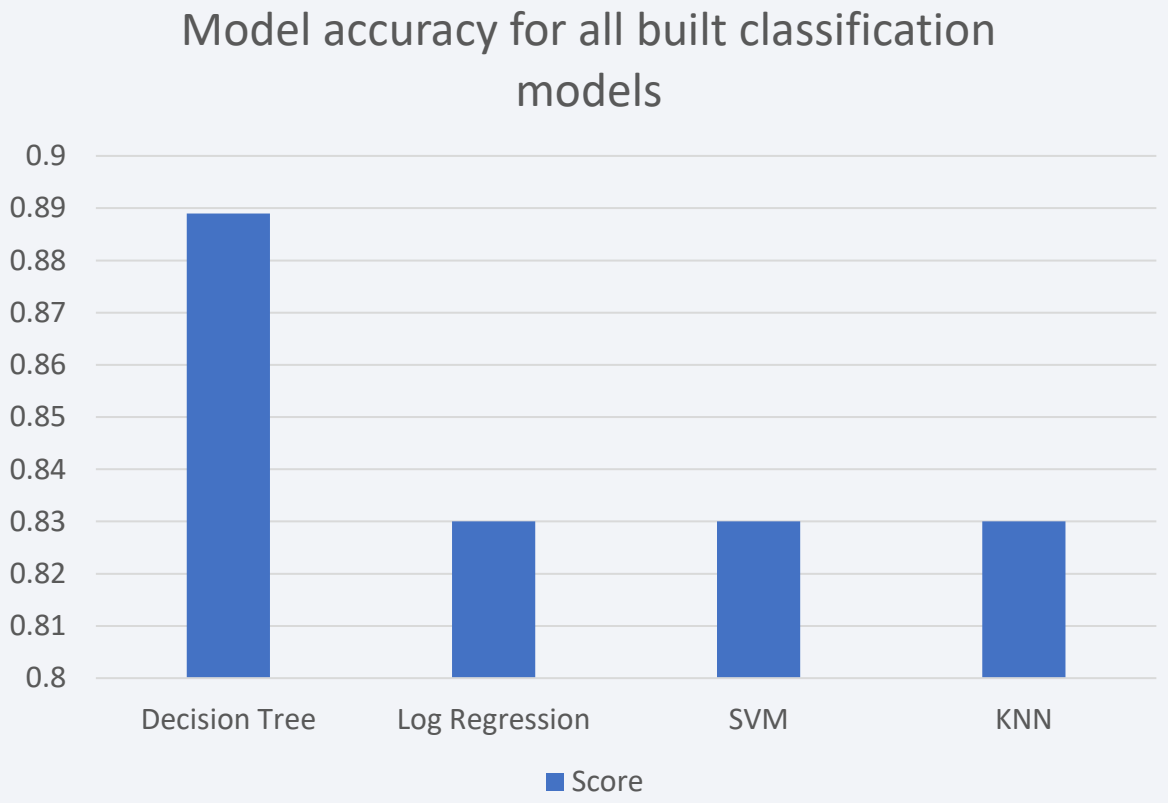
Distance lines to the proximities



Section 6

Predictive Analysis (Classification)

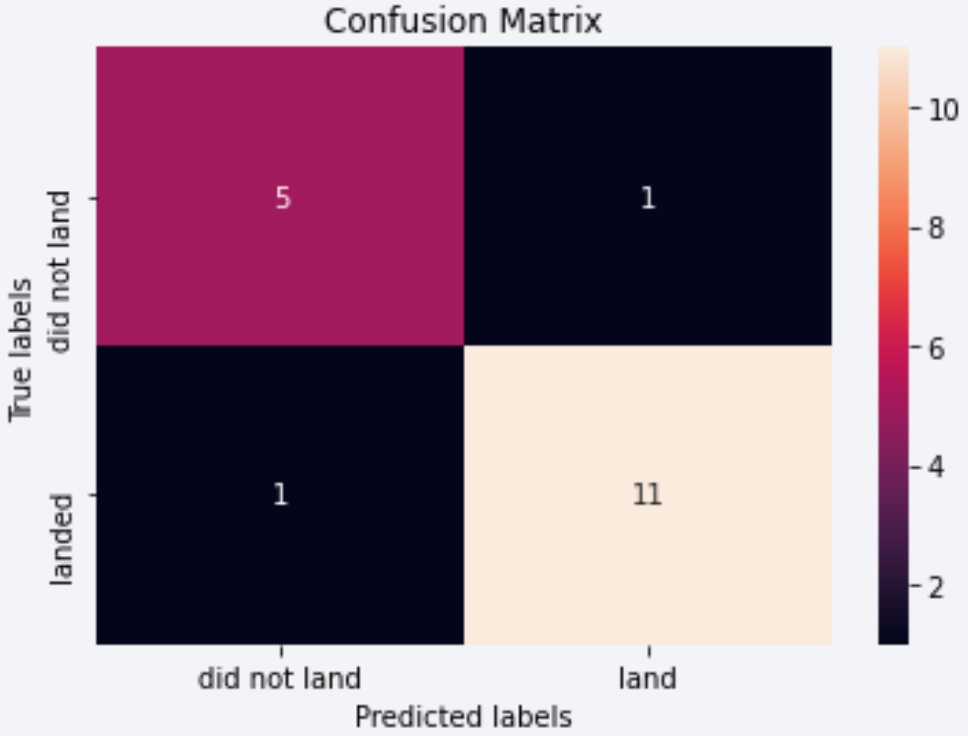
Classification Accuracy



- Decision Tree has the highest classification accuracy on the test data: 0.889

Confusion Matrix

- Confusion matrix of the best performing model: Decision Tree



Examining the confusion matrix, it is possible to see that Decision Tree model can distinguish between the different classes. The main problem are the false positives.

Conclusions

- Success rate increases significantly since 2013 with slightly drop between 2019 and 2020
- Low weighted payloads perform better than the heavier payloads
- CCAFS SLC 40 has the greater number of launches, so the success rate is greater also
- Orbit GEO,HEO,SSO,ES L1 has the best Success Rate
- VLEO orbit shows the higher payload mass, GTO shows payload mass in the interval of 3000 to 7000 kg with a mixture of success and unsuccessful landings
- The Tree Classifier Algorithm is the best for Machine Learning for this dataset

Thank you!