Data Science Capstone Project

Nguyễn Trí Thanh

https://github.com/TThanh10102002

10/8/2022

OUTLINE

Executive summary

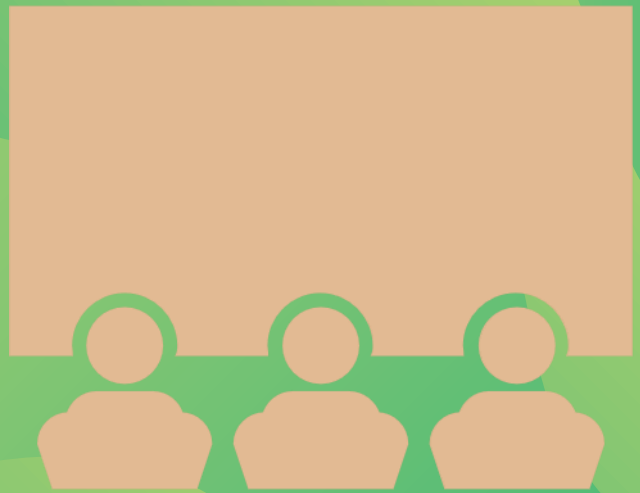
Introduction

Methodology

Results

Conclusion

Appendix



Executive Summary

- Traveling from SpaceX API and SpaceX Wikipedia Webscraping Data Collection to Data Wrangling from the collected data. Exploring data using EDA and SQL, visualization, creating map for SpaceX Launch Site using Folium and Dashboard using dash, html to provide needed information. Gathering relevant columns to be used as features. Changing all categorical variables to binary using one hot encoding. Standardlizing data and applying GridSearchCV to find best parameters for machine learning models. Visualizing accuracy score of all models.
- Using 4 kinds of machine learning models: Logistic Regression, Support Vector Machines, k – Nearest Neighbors and Decision Tree, produced the same accuracy on test dataset with 83,33%. All models over predicted successful landings. More data is needed for better model determination and accuracy.

Introduction

SpaceX Falcon 9

Background

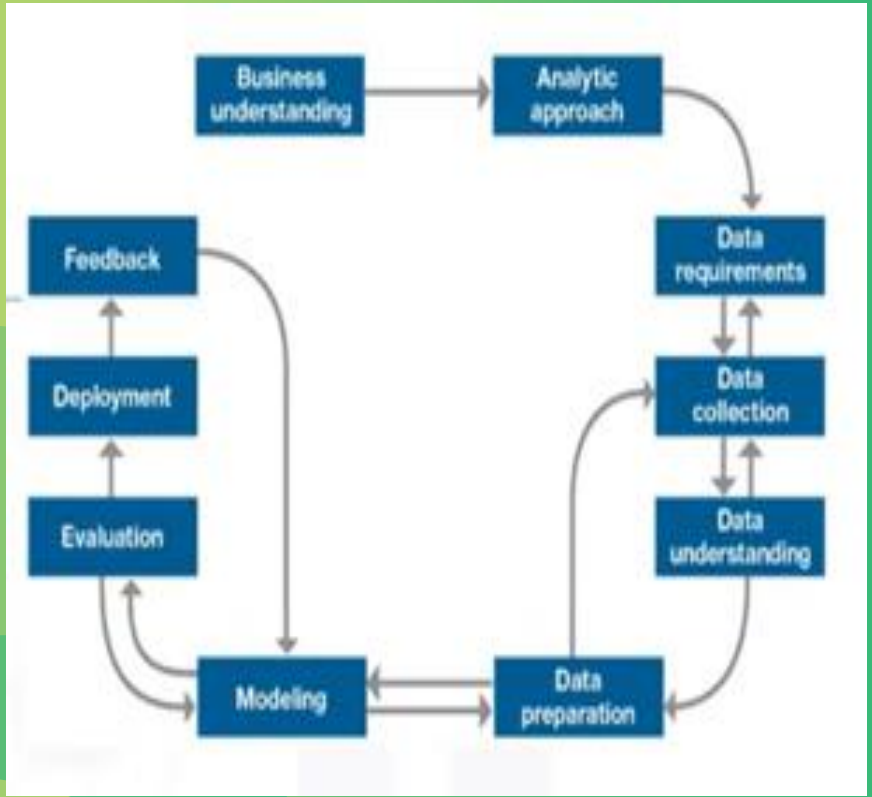
- Commercial Space Age is Here
- Space X has best pricing ($62 million vs. $165 million USD)
- Largely due to ability to recover part of rocket (Stage 1)
- Space Y wants to compete with Space X

Problems

Creating trained machine learning models for prediction of successful Stage 1 recovery

01. Methodology

Report Methodology



Methodology

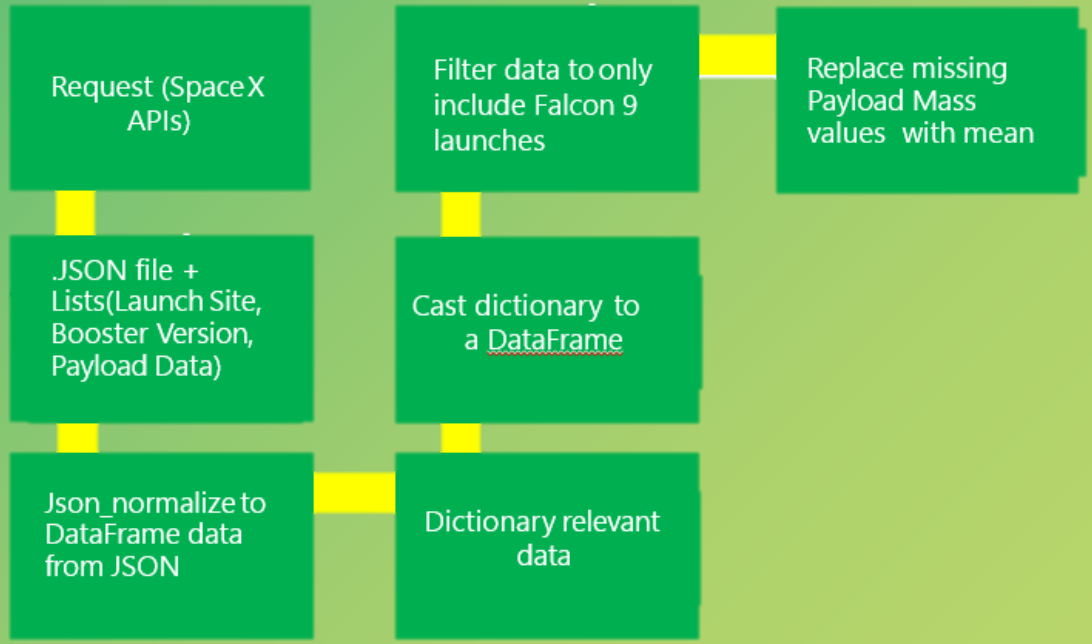
- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV and evaluated using accuracy score.

Data Collection Stage

- Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.
- The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.
- **Space X API Data Features:**
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- **Wikipedia Webscrape Data Features:**
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

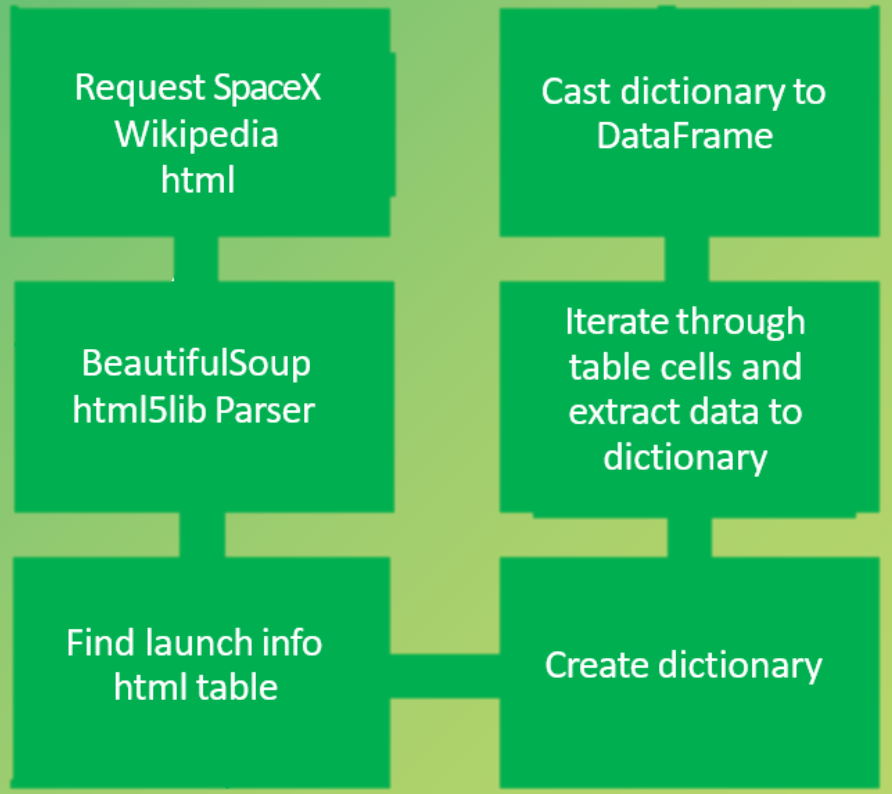
Data Collection with SpaceX API

- Github URL: https://github.com/TThanh10102002/Data-Science-Project/blob/master/Data%20Collection%20API.ipynb



Data Collection with SpaceX Wikipedia HTML

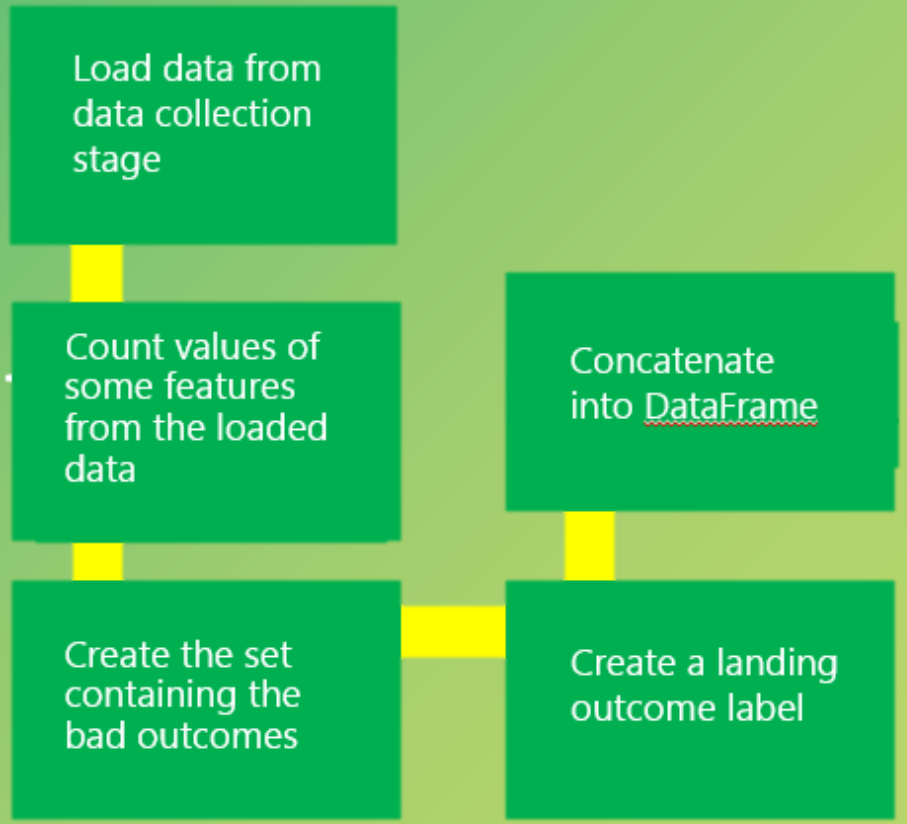
- Github URL: https://github.com/TThanh10102002/Data-Science-Project/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling Stage

- Create a training label with landing outcomes where successful = 1 & failure = 0.
- Outcome column has two components: ‘Landing Outcome’ and ‘Landing Class’
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise. Value Mapping:
- True ASDS, True RTLS, & True Ocean – set to -> 1
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0
- Github URL: https://github.com/TThanh10102002/Data-Science-Project/blob/master/Data%20Wrangling.ipynb

Data Wrangling Flowchart



EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Plots Used:
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
  - Scatter plots, line charts, and bar plots were used to compare correlations between variables to decide if a relationship existed so that they could be used in training the machine learning model.
- Github URL: https://github.com/TThanh10102002/Data-Science-Project/blob/master/EDA%20Visualization.ipynb

EDA with SQL

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes
- Github URL: https://github.com/TThanh10102002/Data-Science-Project/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

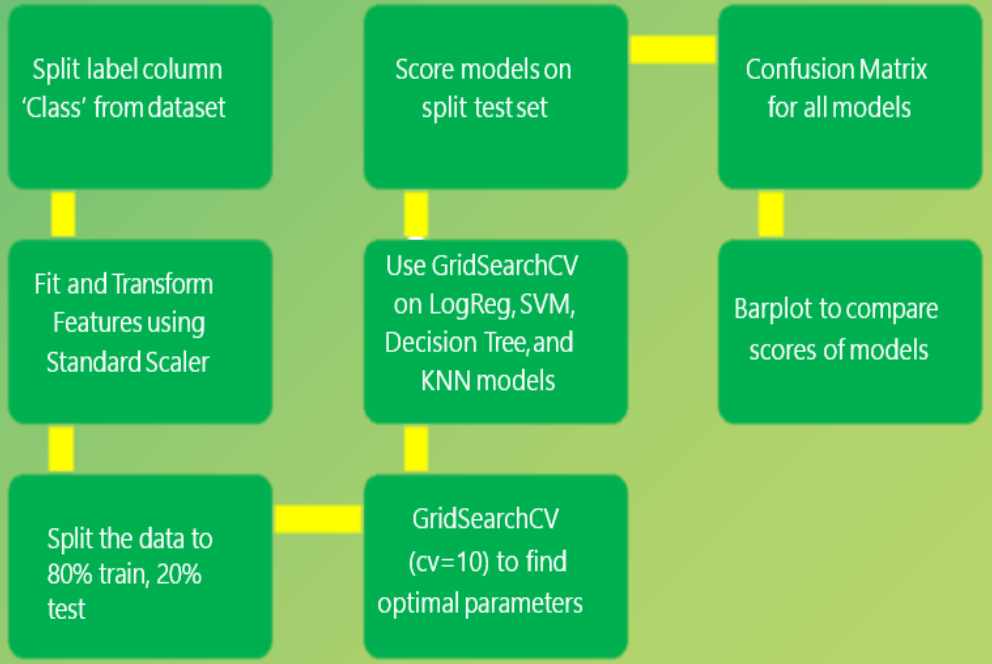
- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.
- Github URL: https://github.com/TThanh10102002/Data-Science-Project/blob/master/Interactive%20Visual%20Analytics.ipynb

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.
- Github URL: https://github.com/TThanh10102002/Data-Science-Project/blob/master/spacex_dash_app.py

Predictive Analysis

Github:
https://github.com/TThanh10102002/Data-Science-Project/blob/master/Machine%20Learning%20Prediction.ipynb



02. Results

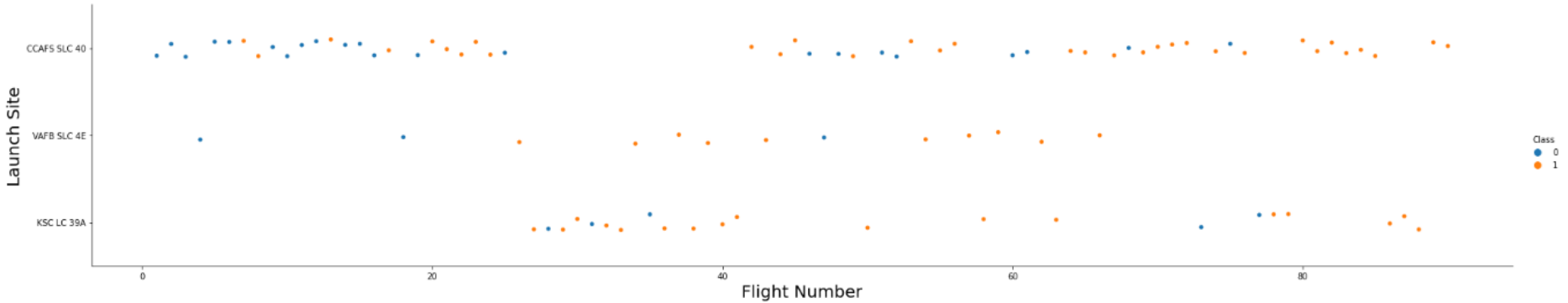
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

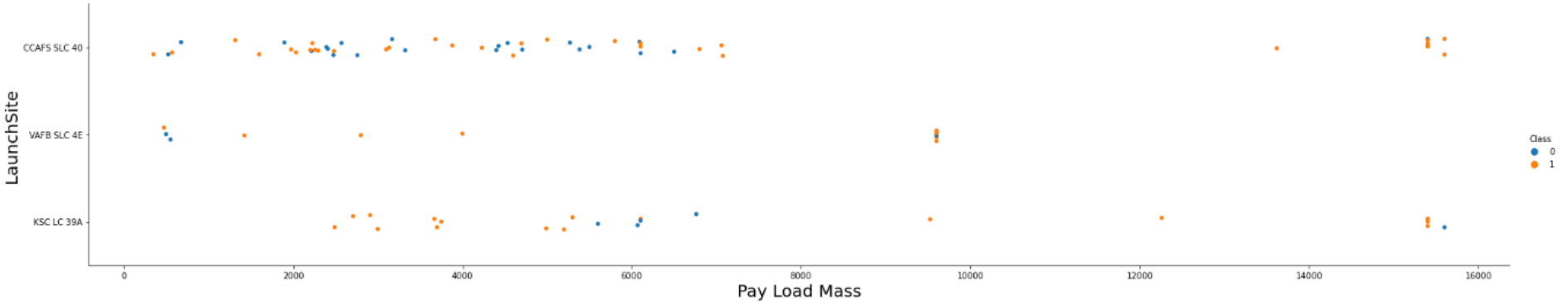
EDA With Visualization

Flight Number vs. Launch Site



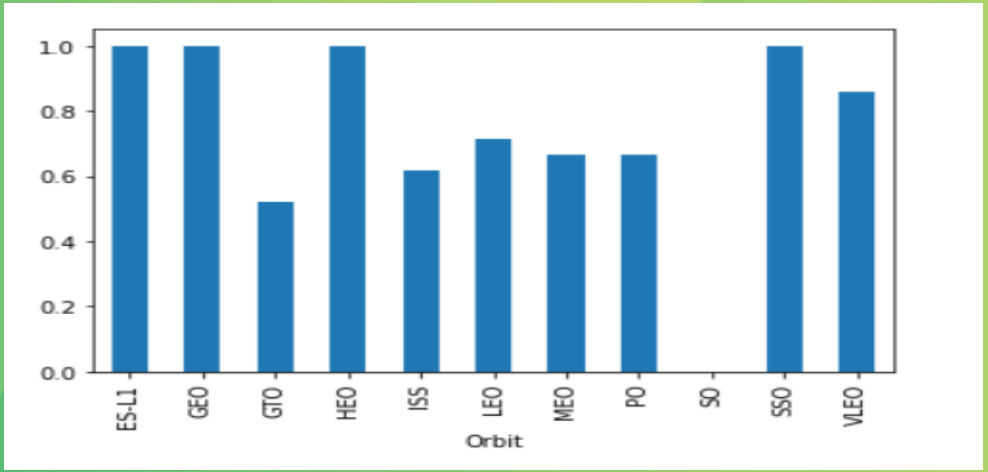
- Orange indicates successful launch, blue indicates failed launch.
- We can see that the successful launch increased over the flight number, the same trend with all 3 launch sites described here.
- There are also some big developments after flight number 20 that we can realize the successful launch overweighted the failed ones.
- Finally, the launch site CCAFS SLC 40 was the most popular launch site among others.

Payload vs. Launch Site



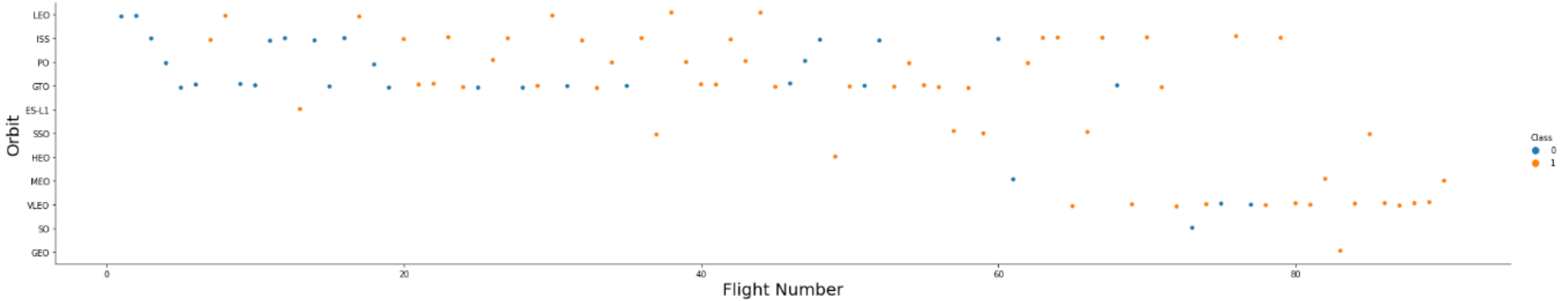
- Orange indicates successful launch, blue indicates failed launch.
- We can see that almost failed launch happened where the launch Payload Mass was below about 8000 kg.
- Launch Site KSC LC 39A performed best when Payload Mass below 6000 and higher than 8000 kg.
- The most successful launch site in Payload test is VAFB SLC 4E. The percentage of launch's success in CCAFS SLC 40 is not stable and hard to say.
- Conclusion: We should use suitable launch site for different payload mass.

Success Rate vs. Orbit Type



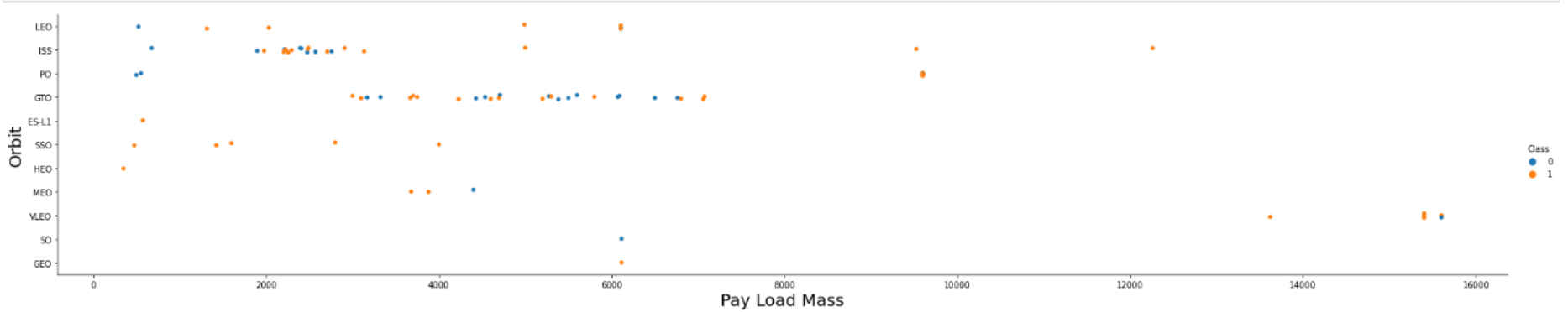
- Success Rate scaled with 0.0 is 0%, 0.2 is 20% and 1.0 is 100%.
- 4 most successful Orbit Type with 100% success rate: ES-L1, GEO, HEO, SSO.
- VLEO had the decent success rate with about 80%.
- The most failed Orbit Type with 0% success rate: SO
- The other orbit types are nearly the same success rate (about 50 - 60%).

Flight Number vs. Orbit Type



- Orange indicates successful launch, blue indicates failed launch.
- Launch Orbit preferences changed over the Flight Number. Almost failed launch occurred below Flight Number 20.
- SpaceX started to experiment with LEO Orbit and closed with MEO Orbit.
- SpaceX seemed to be perform better in lower orbits or Sun – synchronous orbits.

Payload vs. Orbit Type



- Orange indicates successful launch, blue indicates failed launch.
- Most of the experiments happened with Payload Mass below 8000 kg. With the experiments above that threshold, the success rate is nearly 100%.
- Based on the Payload Mass, we can choose the suitable Orbit to assure the success rate of the launch (VLEO for over 12000 kg, SSO for below 6000 kg, etc).

Launch Success Yearly Trend

- Success generally increases over time since 2013 with a huge development from 2015 to 2017 and a slight dip in 2018.
- Success in recent years at over 80%.

EDA With SQL

All Launch Site Names

- Query unique launch site names from database.
- CCAFS SLC-40 and CCAFS LC-40 seemed to represent the same launch site with data entry errors.
- CCAFS LC-40 was the previous name. Likely only 3 unique launch site values: CCAFS SLC-40, KSC LC-39A,VAFB SLC-4E

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXDATASET
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41c
Done.
```

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

Launch Site Names Begin with CCA

First 5 Launch Sites with Name begin as ‘CCA’ from the SpaceX Dataset

```
%sql SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

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

Total Payload Mass

- This query sums the total payload mass in kg where NASA was the customer.
- CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE PAYLOAD LIKE '%CRS%'
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u
Done.
```

1
111268

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass or launches which used booster version F9 v1.1
- Average payload mass of F9 1.1 is on the low end of our payload mass range

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXDATASET WHERE BOOSTER_VERSION LIKE 'F9 v1.1'
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databa
Done.
```

1
2928

First Successful Ground Pad Landing Date

- This query found the the first day that recorded the successful ground pad landing.
- Despite the successful landing on the ground was in 2014, the first successful landing on the ground pad occurred very lately, at 22/12/2015

```
%sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databas
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 kg.

```
%sql SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- This query returns the statements of mission outcomes with their quantities.
- Especially, we can see that 99 launchs were evaluated 'Success' – which seemed to mean SpaceX intended to plan almost the failed launchs, despite only 1 time was real failure. We can also see that there was 1 case the launch succeeded with the payload status unclear.

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QUANTITIES FROM SPACEXDATASET GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

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

Boosters Carried Maximum Payload

- This query returns the booster versions that carried the highest payload mass..
- These booster versions are very similar and all are of the F9 B5 B10xx.x variety.
- This likely indicates payload mass correlates with the booster version that is used.

```
%sql SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET)
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

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

- This query returns the failed landing outcomes in droneship with their landing outcome, booster version and launch site in 2015.
- There were 2 failed ones in 2015, all with Launch site is CCAFS LC-40 and Booster version F9 v1.1 B10xx

```
%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXDATASET WHERE DATE LIKE '2015%' AND LANDING__OUTCOME = 'Failure (drone ship)'
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

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

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

- This query returns a list of landings outcomes and their quantities between 2010-06-04 and 2017-03-20 inclusively.
- Leading of the rank was 'No attempt' outcome with 10 times, twice higher than the following Failure and Success (drone ship) with 5 times equally. The least landing outcome was 'Precluded (drone ship) with only 1 time.

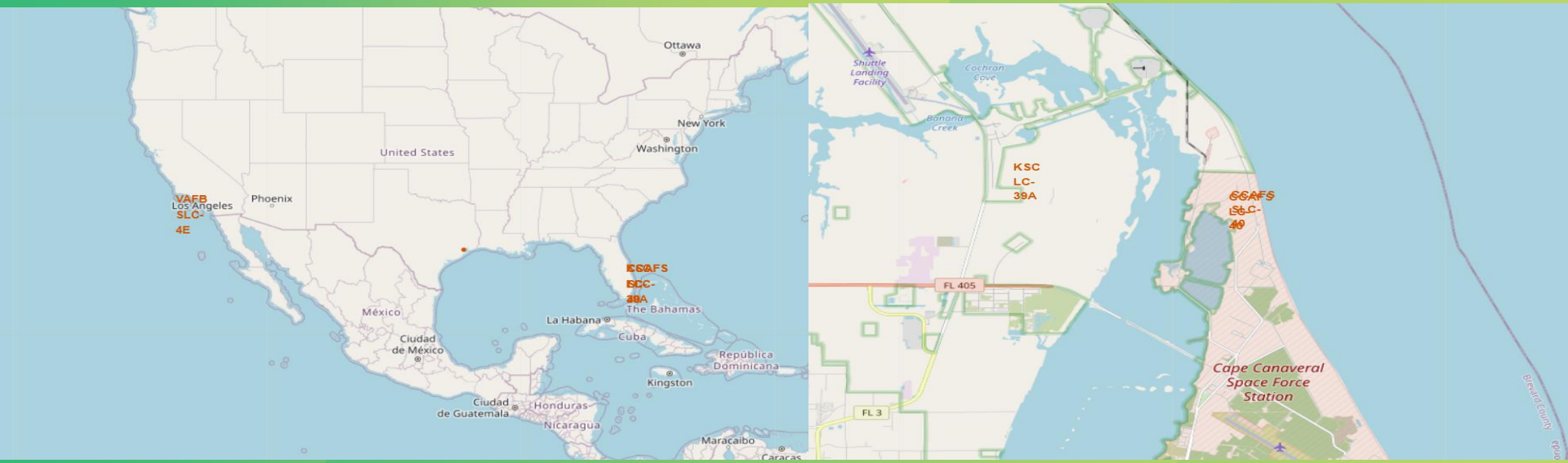
```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS QUANTITIES FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORD
```

```
 * ibm_db_sa://sjm94128:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

landing__outcome	quantities
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

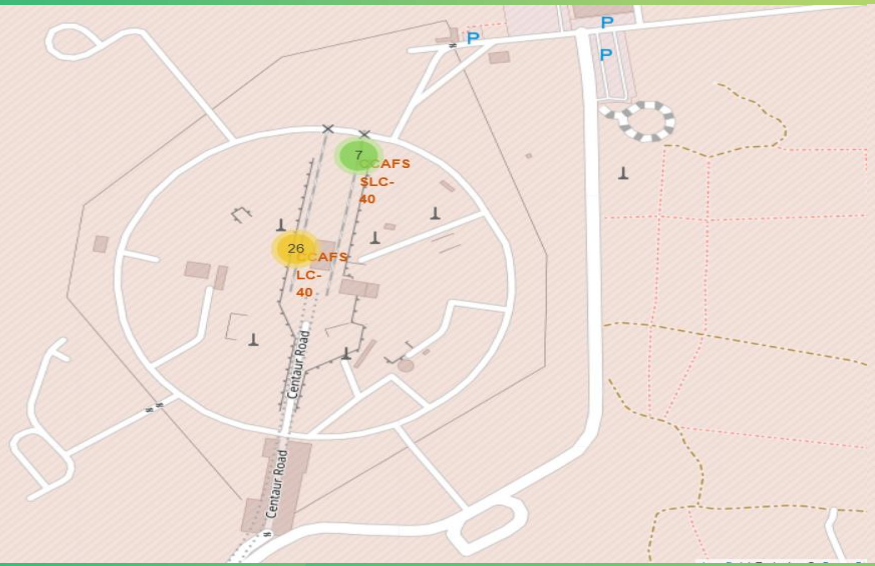
Interactive Map with Folium

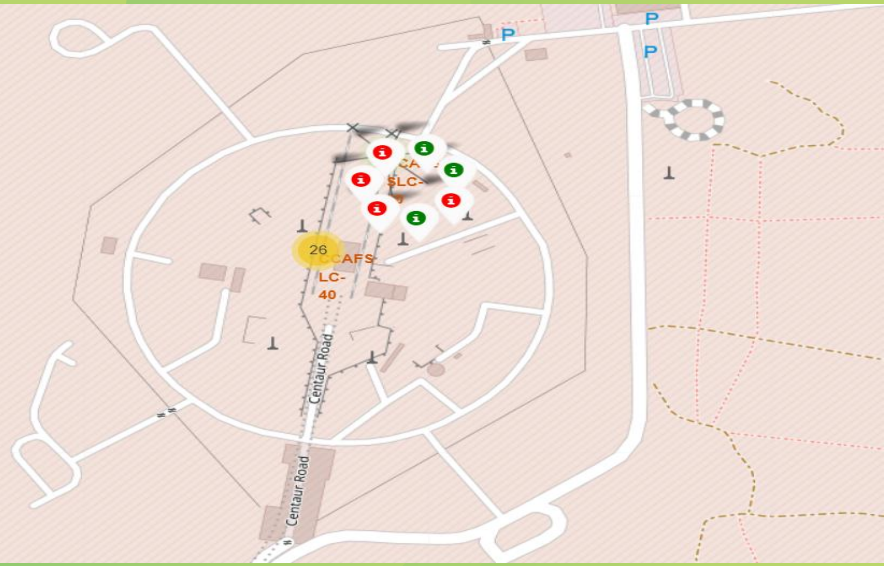
Launch Sites Locations



- The left map show all relative launch sites in US. While the right map show specific launch sites located in Florida (US).

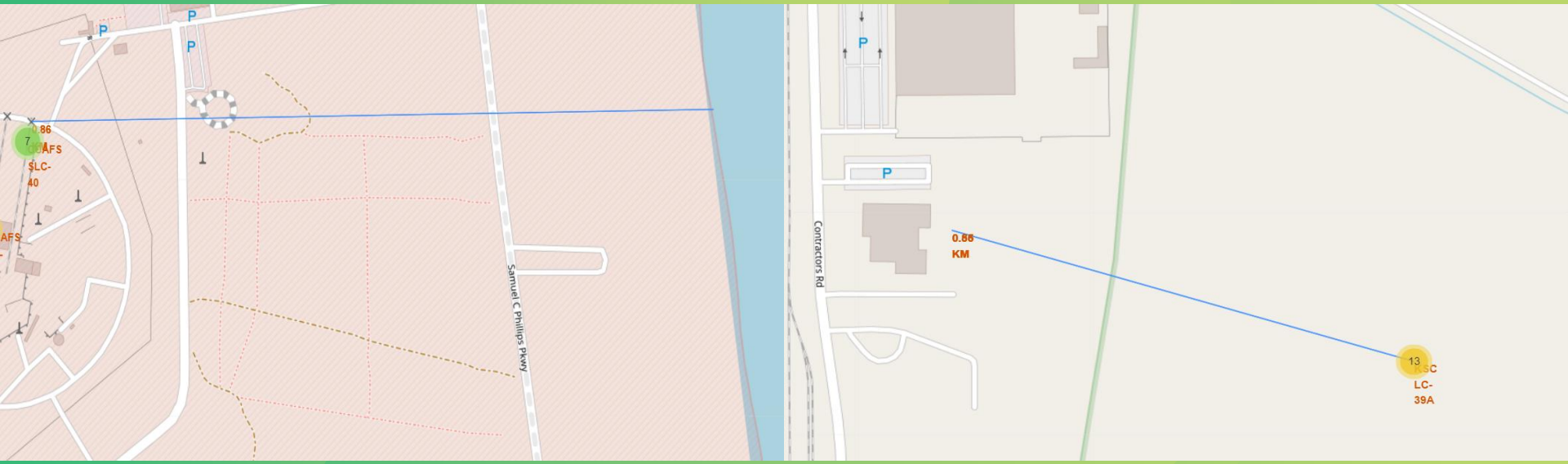
Launch Sites Marker with Colored Cluster





- The left map show some launch sites in Florida at one place. It had been clustered into 2 groups with 7 points and 26 points. The right map show the detailed location of all 7 points in the first cluster.

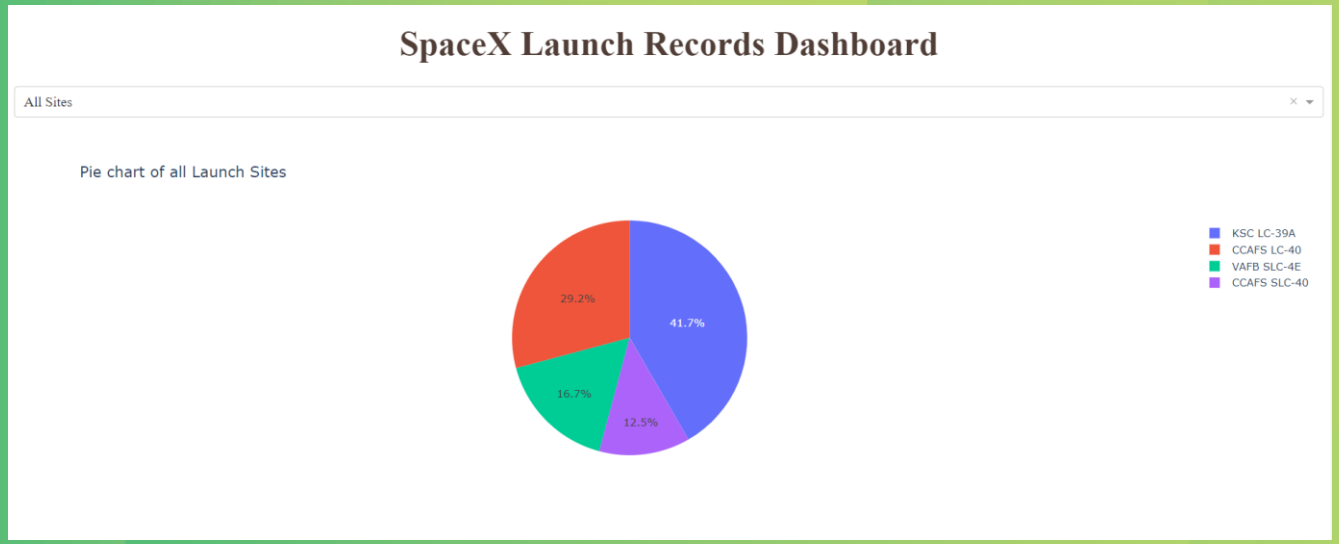
Key Location Proximities



- We can see that launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

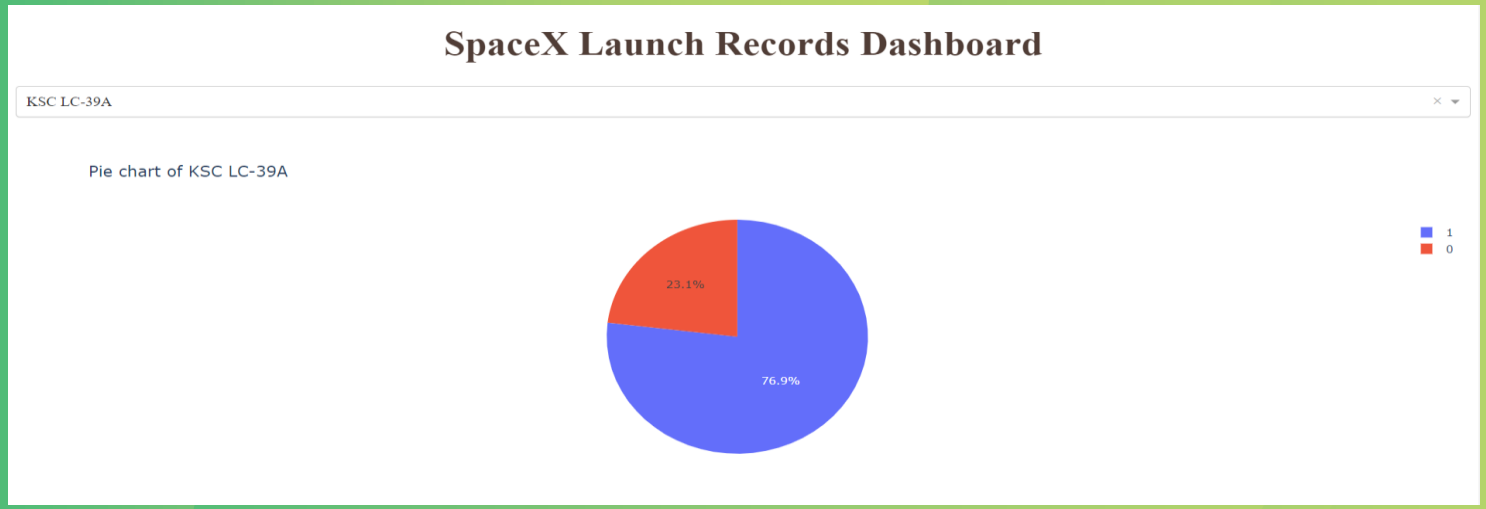
Build a Dashboard with Plotly Dash

Successful Launches across Launch Sites



- This is the distribution of successful landings across all launch sites. CCAFS LC-40 is similar to CCAFS SLC-40 so CCAFS and KSC shared the same amount of successful landings, but a majority of the successful landings where performed with CCAFS LC-40. VAFB had the smallest share of successful landings. This might be smaller sample and increases in difficulty of launching in the west coast.

Highest Success Ratio of All Launch Sites



- As we can see when testing the success ratio of all the launch sites, KSC LC-39A had the highest percentages of success among launch sites with 76.9%. Following was CCAFS LC-40 and CCAFS SLC-40 with the components around 60-70%.

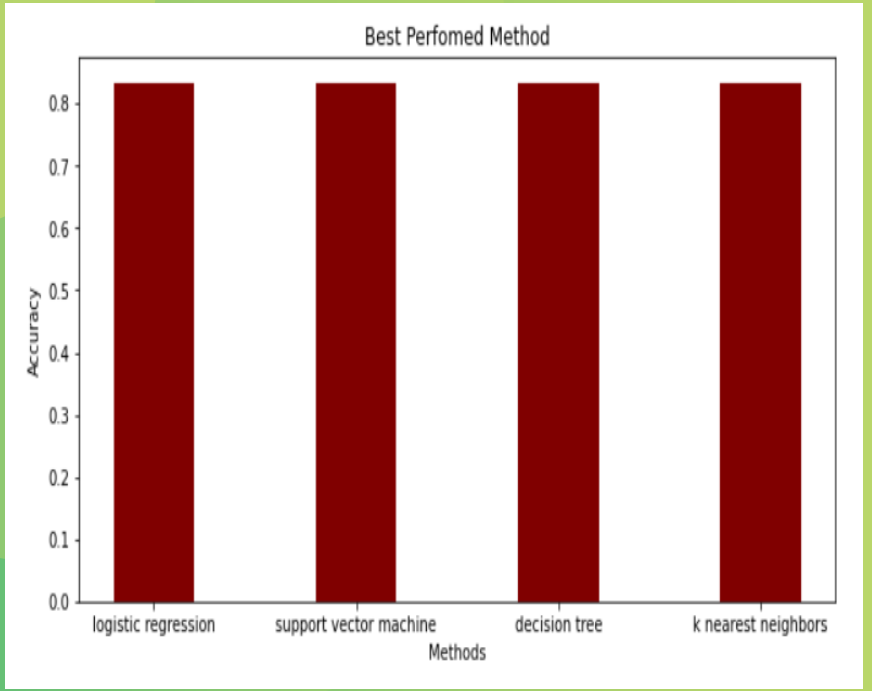
Payload Mass vs. Launch Outcome

- Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Predictive Analysis (Classification)

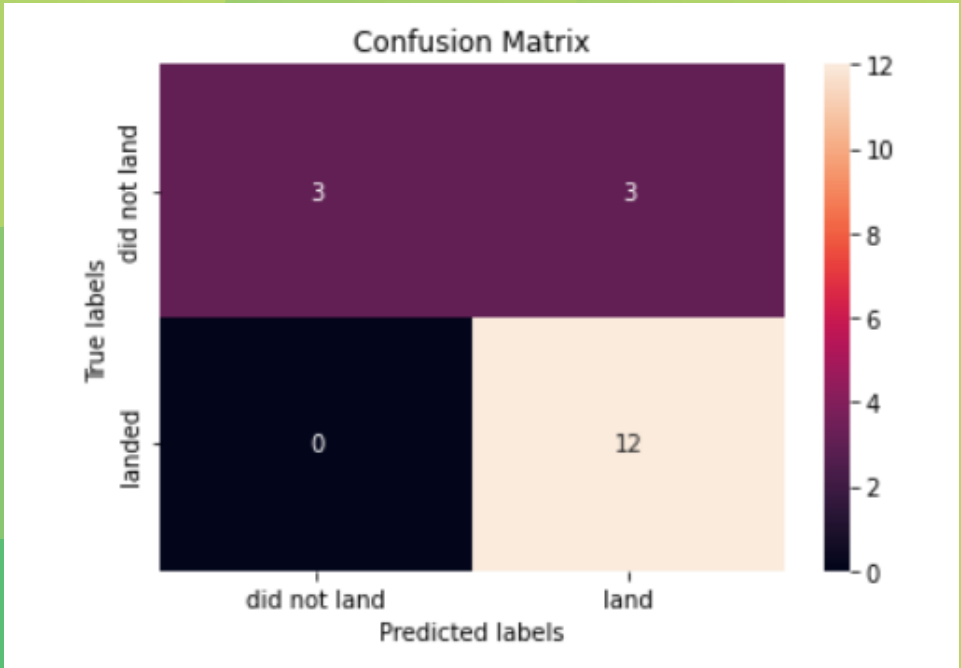
Classification Accuracy

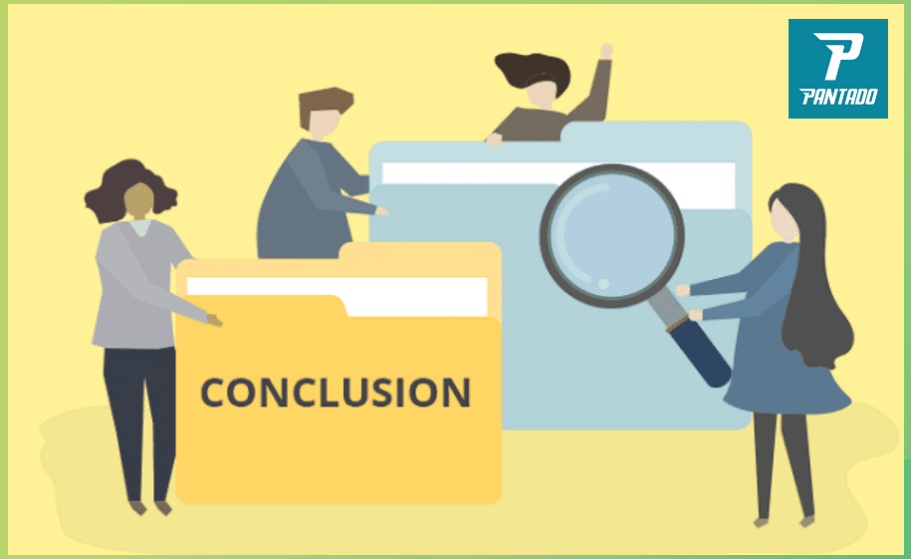
- As we can see from the bar chart on the right hand side, all Machine Learning methods (LogReg, SVM, Decision Tree and KNN) have the same accuracy on test set (about 83,33% accuracy score).
- There is a notice that the sample data for test set was very small (18 data points) so it caused large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.



Confusion Matrix

- Since all models performed the same for the test set, the confusion matrix is the same across all models.
- The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).
- There were no predictions when the true label was landed and the predicted label was did not land (false negatives).
- Our models over predict successful landings.





03. Conclusion

Some main points and insights

Conclusion

- Our task: developing a machine learning model for Space Y who wants to compete against SpaceX.
- The goal of ML model is to predict whether Stage 1 will successfully land to save ~$100 million USD.
- The data was collected from a public SpaceX API and web scraping SpaceX Wikipedia page.
- After collecting the data, I labeled data and stored data into a DB2 SQL database.
- Next step, I created a dashboard for visualization with Plotly Dash and Interactive Map with Folium.
- Further, I created a machine learning model with an accuracy of 83.33% with 4 kinds of algorithms: Logistic Regression, Support Vector Machines, Decision Tree and k – Nearest Neighbors.
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launching to determine whether the launch should be made or not.
- More data should be collected to better determine the best machine learning model and improve accuracy.

Appendix

- GitHub repository url: https://github.com/TThanh10102002/Data-Science-Project/tree/master
- Instructors:
  - **Instructors: Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo**

- Special Thanks to All Instructors:
- https://www.coursera.org/professional-certificates/ibm-data-science?#instructors

Thanks!