

Data Science Capstone Project

Abhijeet Rojatkar
<https://github.com/abhi1045>



OUTLINE

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

Executive summary

- ✓ Data was gathered from the public SpaceX API and the SpaceX Wikipedia page. The 'class' label column was created to categorize successful landings.
- ✓ SQL, visualization, folium maps, and dashboards were used to investigate data.
- ✓ I compiled a list of relevant columns to use as features. Using a single hot encoding, all category variables were converted to binary.
- ✓ GridSearchCV was used to determine the optimum parameters for machine learning models after standardizing data. Visualize all of the models' accuracy scores.
- ✓ Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
- ✓ All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings.
- ✓ For better model determination and accuracy, more data is needed.

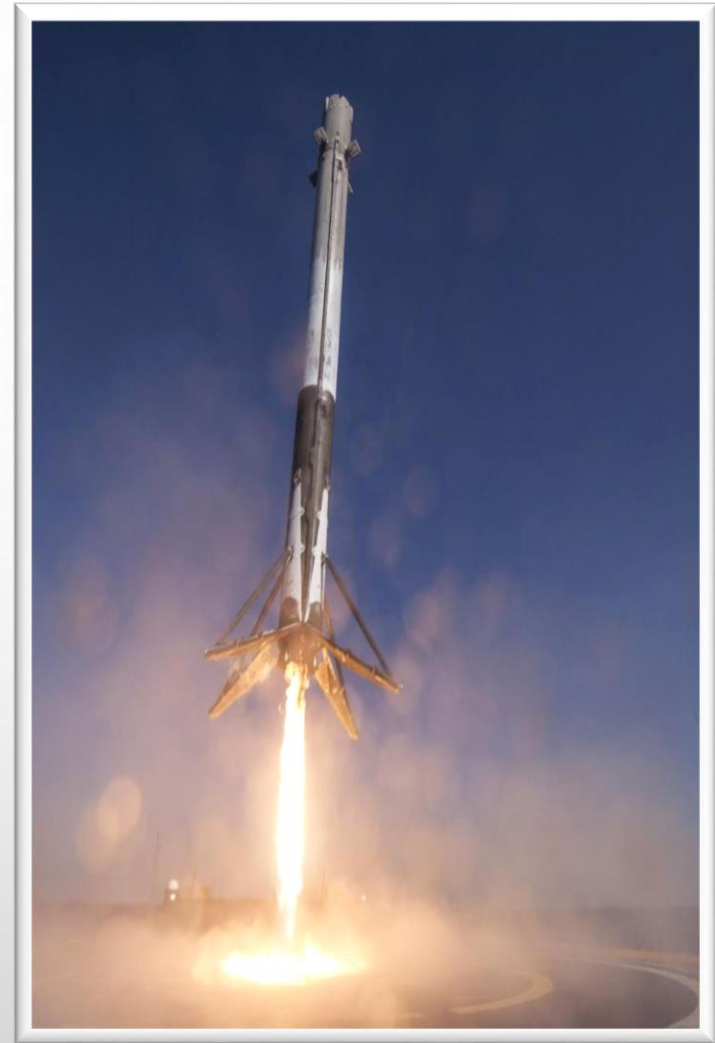
INTRODUCTION

✓ BACKGROUND:

- ✓ The commercial space age has arrived.
- ✓ The most affordable option is SpaceX (\$62 million vs. \$165 million USD).
- ✓ Mostly owing to the ability to recover a portion of the rocket (stage 1).
- ✓ SPACEY aspires to be a rival to SpaceX.

✓ PROBLEM:

- ✓ SPACEY tasks us to train a machine learning model to predict successful stage 1 recovery.



SpaceX Falcon 9 Rocket – The Verge



METHODOLOGY

OVERVIEW OF DATA COLLECTION, DATA WRANGLING, DATA
VISUALIZATION, DASHBOARD AND CLASSIFICATION MODELS

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 like Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors using Gridsearch CV.

DATA COLLECTION OVERVIEW

- ✓ Data collection process involved a combination of API requests from SpaceX public API and web scraping data from a table in SpaceX'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 web scraping.

- ✓ SPACEX API data columns:

- ✓ Flightnumber, date, boosterversion, payloadmass, orbit, launchsite, outcome, flights, gridfins, Reused, legs, landingpad, block, reusedcount, serial, longitude, latitude.

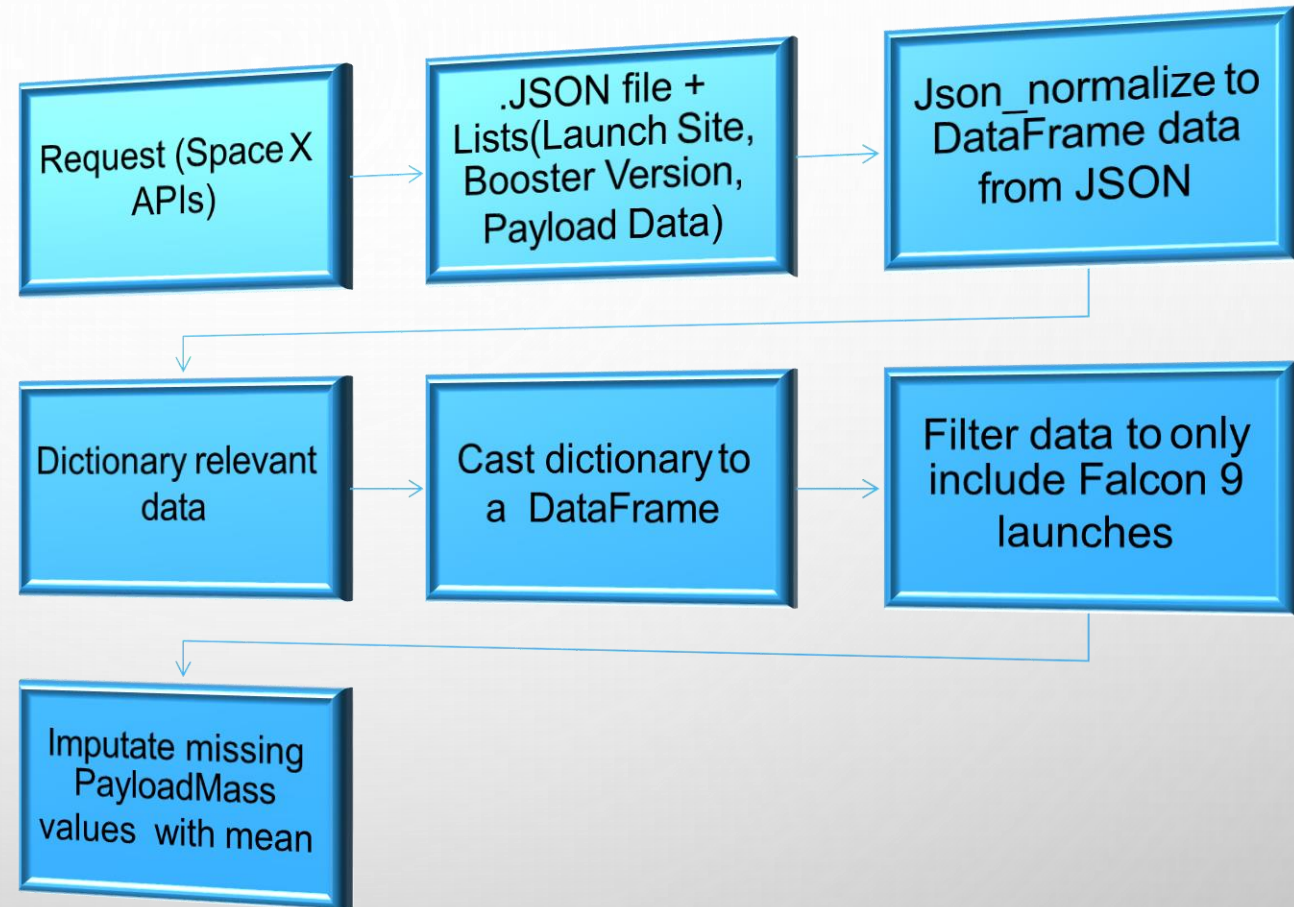
- ✓ Wikipedia Web Scrape data columns:

- ✓ Flight no., Launch site, payload, payloadmass, orbit, customer, launch outcome, version booster, booster landing, date, time.

DATA COLLECTION SPACEX API

[GITHUB URL](#)

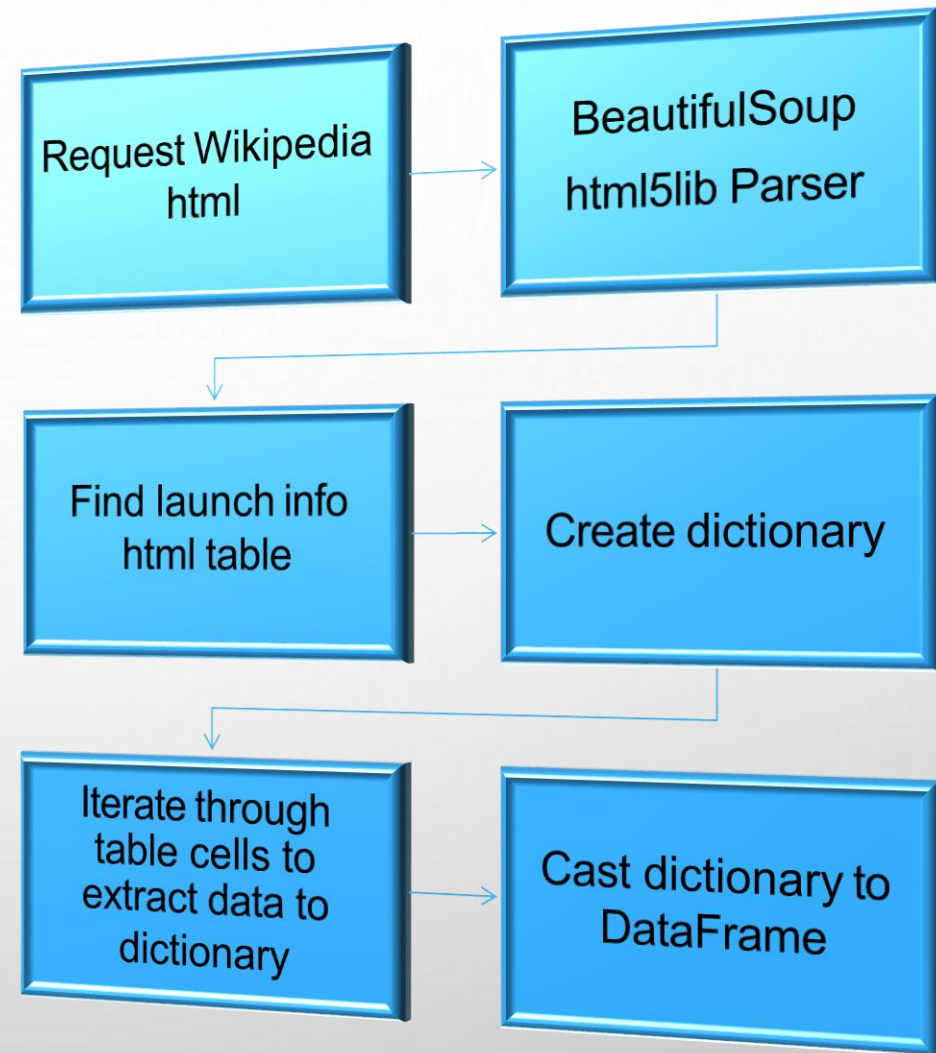
[IBM CLOUD URL](#)



DATA COLLECTION WEB SCRAPPING

[GITHUB URL](#)

[IBM CLOUD URL](#)



DATA WRANGLING

- ✓ Create a training label with a successful landing outcome of 1 and a failure outcome of 0.
- ✓ 'Mission outcome' and 'landing location' are the two components of the outcome column.
- ✓ New training label column 'class' with a value of 1 if 'mission result' is true and 0 otherwise.

✓ Value Mapping:

- ✓ Set to -> 1 for true ASDS, true RTLS, and true ocean.
- ✓ Set to -> 0 for None none, false ASDS, none ASDS, false ocean, false RTLS
- ✓ [GITHUB URL](#)
- ✓ [IBM CLOUD URL](#)

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.
 - ✓ Decide if a relationship exists so that they could be used in training the machine learning model.
 - ✓ Insights from Graph are given below of each graphs.
 - ✓ [GITHUB URL](#)

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](#)

BUILD AN INTERACTIVE MAP WITH FOLIUM

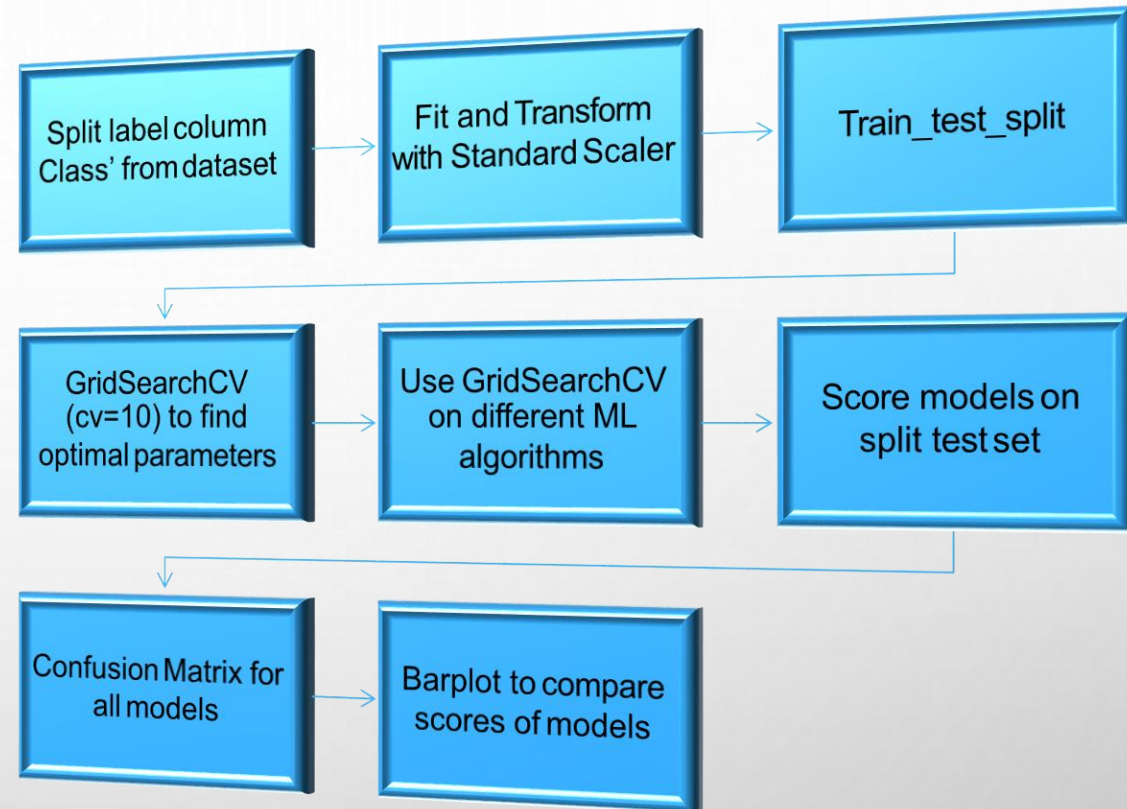
- ✓ Folium maps mark launch sites, successful and unsuccessful landings, and a proximity example to key locations: railway, highway, coast, and city.
- ✓ This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.
- ✓ [GITHUB URL](#)

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](#)

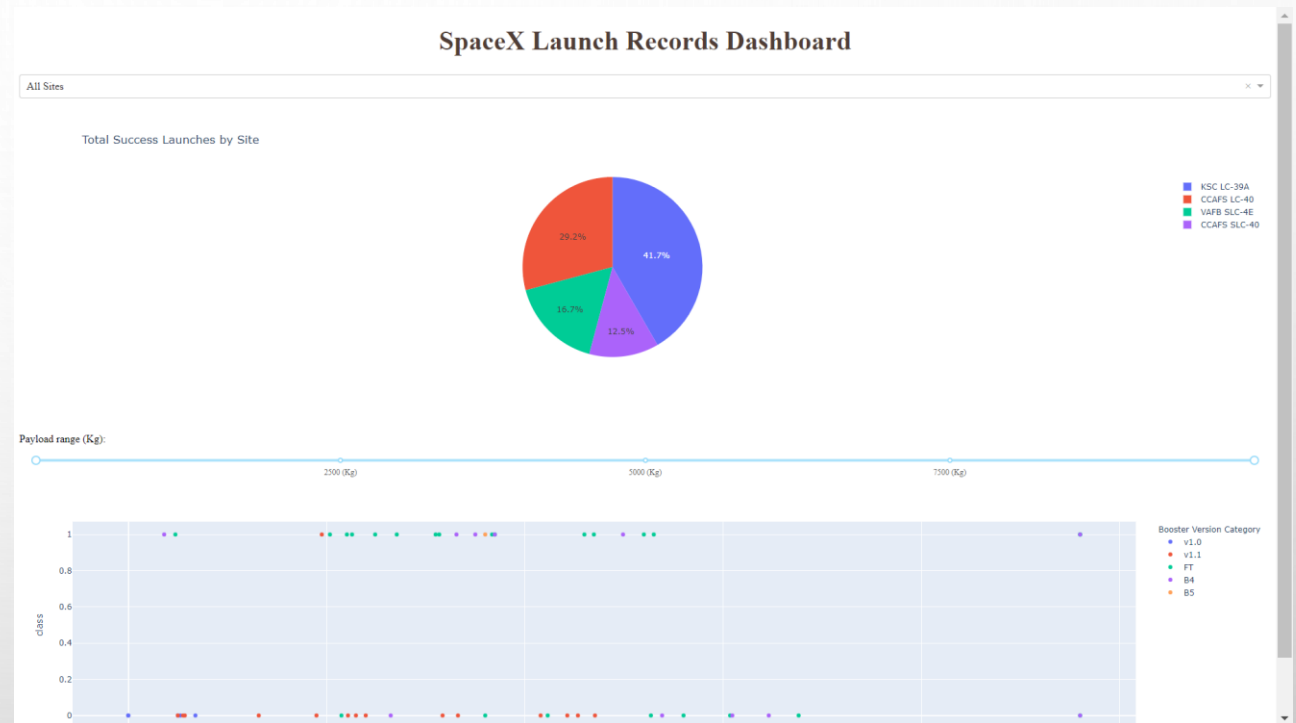
PREDICTIVE ANALYSIS (CLASSIFICATION)


[GITHUB URL](#)



RESULTS

- ✓ This is a preview of the Plotly dashboard.
- ✓ The following slides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium.
- ✓ Finally the results of our model with about **83% accuracy**.





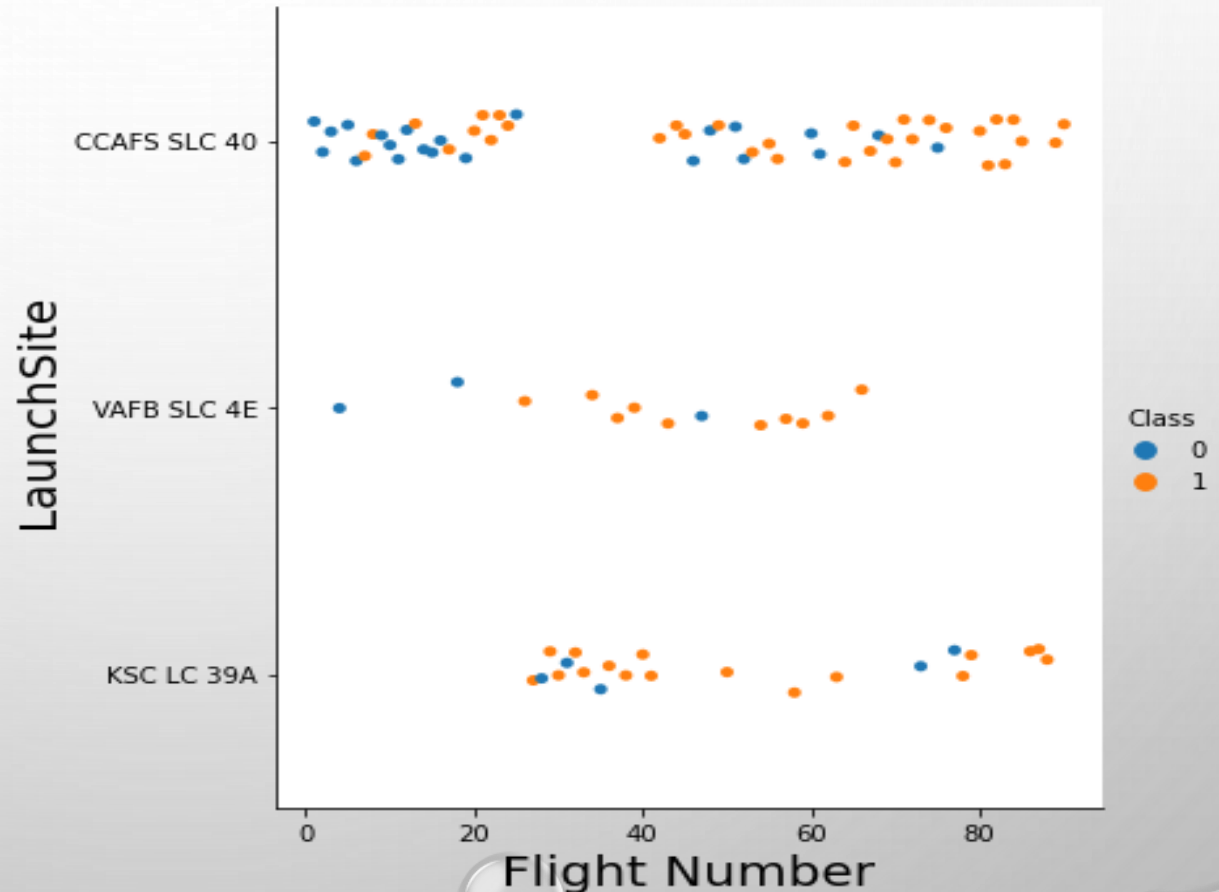
INSIGHTS DRAWN FROM EDA WITH VISUALIZATION

EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS



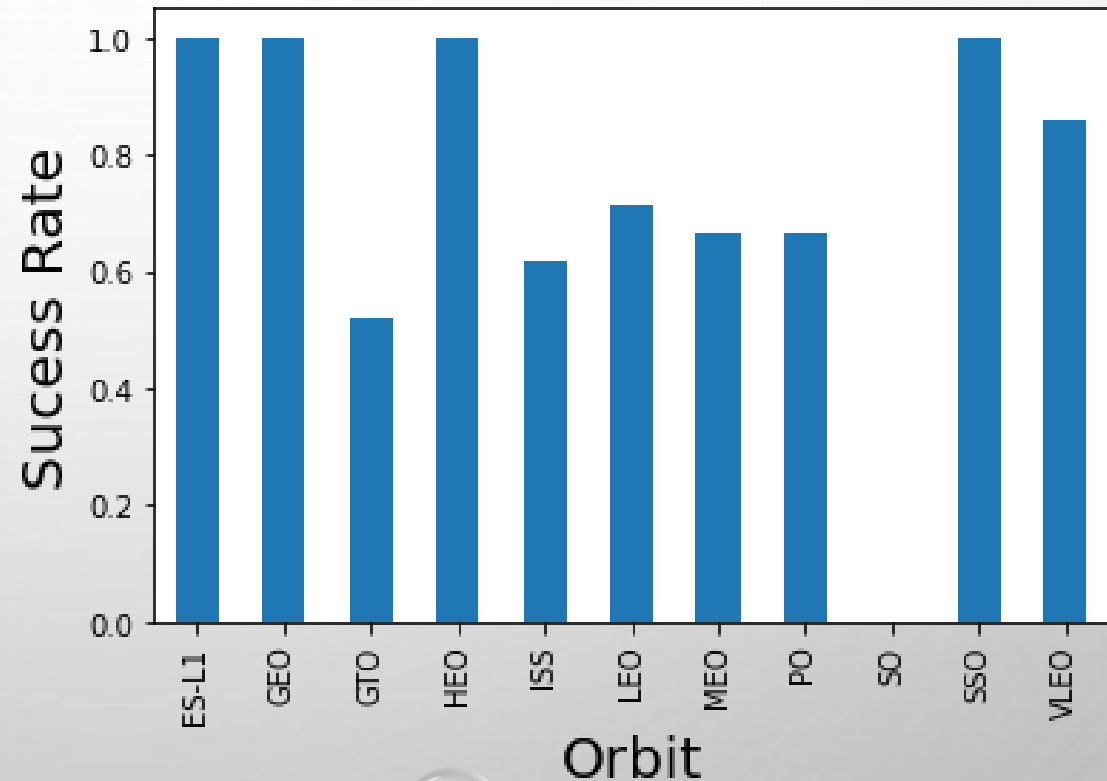
FLIGHT NUMBER VS. LAUNCHSITE

- ✓ Orange indicates successful launch; Blue indicates unsuccessful launch.
- ✓ Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

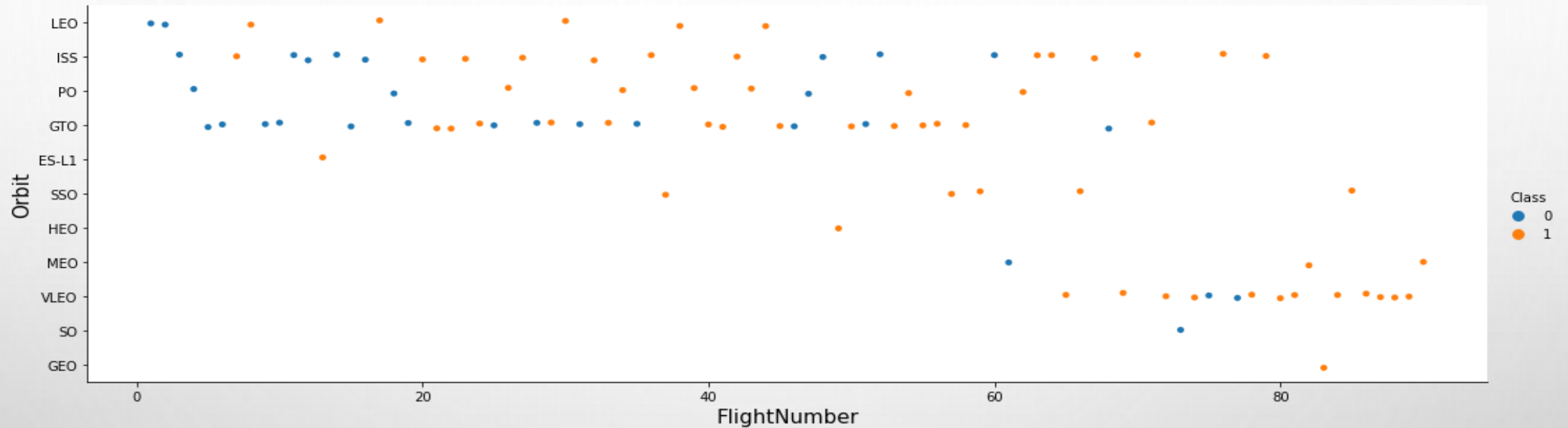


SUCCESS RATE VS. ORBIT TYPE

- ✓ Success Rate Scale with 0 as 0%, 0.6 as 60%
1 as 100%
- ✓ ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)
- ✓ SSO (5) has 100% success rate
- ✓ VLEO (14) has decent success rate and attempts
- ✓ SO (1) has 0% success rate
- ✓ GTO (27) has the around 50% success rate but largest sample.

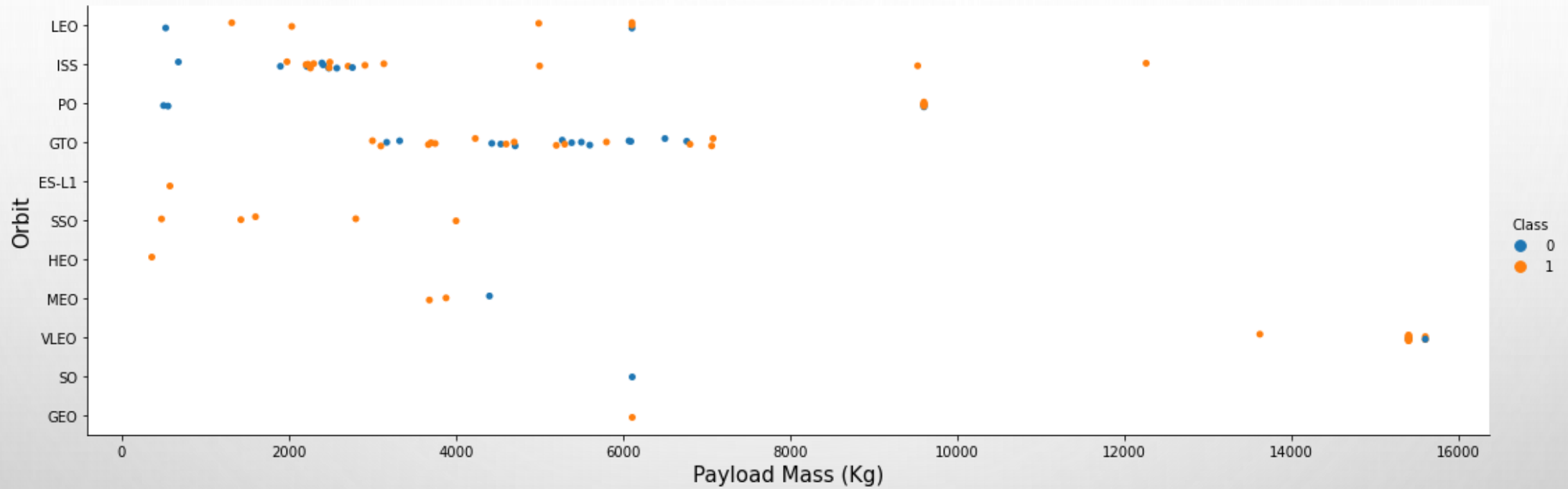


FLIGHT NUMBER VS. ORBIT TYPE



- ✓ Orange indicates successful launch; Blue indicates unsuccessful launch.
- ✓ Launch Orbit preferences changed over Flight Number.
Launch Outcome seems to correlate with this preference.
- ✓ SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun-synchronous orbits.

PAYLOAD MASS VS. ORBIT TYPE



- ✓ Orange indicates successful launch; Blue indicates unsuccessful launch.
- ✓ Payload mass seems to correlate with orbit.
- ✓ LEO and SSO seem to have relatively low payload mass.
- ✓ The other most successful orbit VLEO only has payload mass values in the higher end of the range.

LAUNCH SUCCESS YEARLY TREND

- ✓ Success generally increases over time since 2013 with a slight dip in 2018.
- ✓ Success in recent years at around 80%.





EDA WITH SQL

EXPLORATORY DATA ANALYSIS WITH SQL DB2 INTEGRATED IN
PYTHON WITH SQLACHEMY

ALL LAUNCH SITES NAMES

- ✓ Query unique launch site names from database.
- ✓ CCAFS SLC-40 and CCAFSSLC-40 likely all 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

In [4]:

```
%%sql  
SELECT UNIQUE LAUNCH_SITE  
FROM SPACEXDATASET;  
  
* ibm_db_sa://ftb12020:***@0c77d6f:  
Done.
```

Out[4]:

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

LAUNCHING SITES NAME BEGINNING WITH 'CCA'

```
In [5]: %%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

```
Out[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

- ✓ First five entries in database with Launch Site name beginning with CCA.

TOTAL PAYLOAD MASS (KG) FROM NASA

- ✓ 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_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg

45596

AVERAGE PAYLOAD MASS (KG) 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_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

avg_payload_mass_kg
2928

FIRST SUCCESSFUL GROUND PAD LANDING DATE

- ✓ This query returns the first successful ground pad landing date.
- ✓ First ground pad landing wasn't until the end of 2015.
- ✓ Successful landings in general appear starting 2014.

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success

2015-12-22

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing_outcome = 'Success (drone ship)' AND payload_mass_kg BETWEEN 4001 AND 5999;

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.database
Done.
```

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

- ✓ This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 no inclusively.

TOTAL NUMBER OF EACH MISSION OUTCOME

- ✓ This query returns a count of each mission outcome.
- ✓ SpaceX appears to achieve its mission outcome nearly 99% of the time.
- ✓ This means that most of the landing failures are intended.
- ✓ Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-1
Done.
```

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

BOOSTERS THAT CARRIED MAXIMUM PAYLOAD (KG)

- ✓ This query returns the booster versions that carried the highest payload mass of 15600 kg.
- ✓ 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, PAYLOAD_MASS_KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS_KG_ = (SELECT MAX(PAYLOAD_MASS_KG_) FROM SPACEXDATASET);

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1
Done.
```

booster_version	payload_mass_kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 FAILED DRONES SHIP LANDING RECORDS

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;

* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- ✓ This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.
- ✓ There were two such occurrences.

RANKING COUNT OF SUCCESSFUL LANDINGS BETWEEN 2010-06-04 AND 2017-03-20

- ✓ This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.
- ✓ There are two types of successful landing outcomes: drone ship and ground pad landings.
- ✓ There were 8 successful landings in total during this time period

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

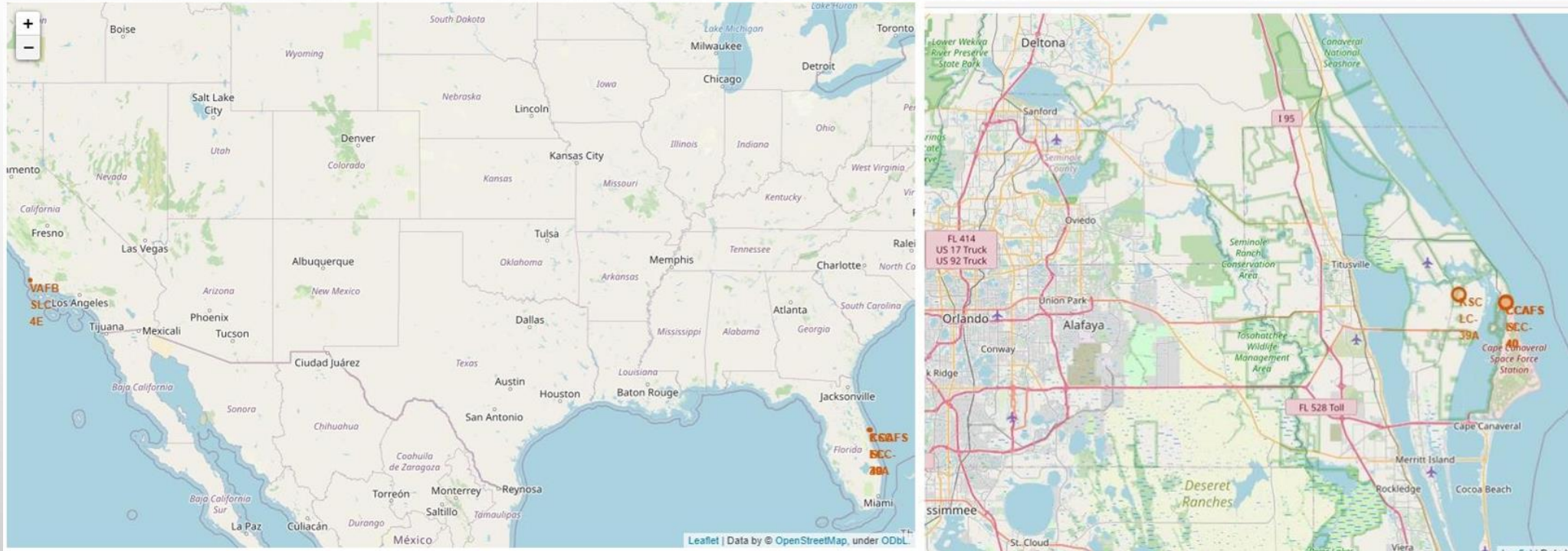
* ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3



INTERACTIVE MAP WITH FOLIUM

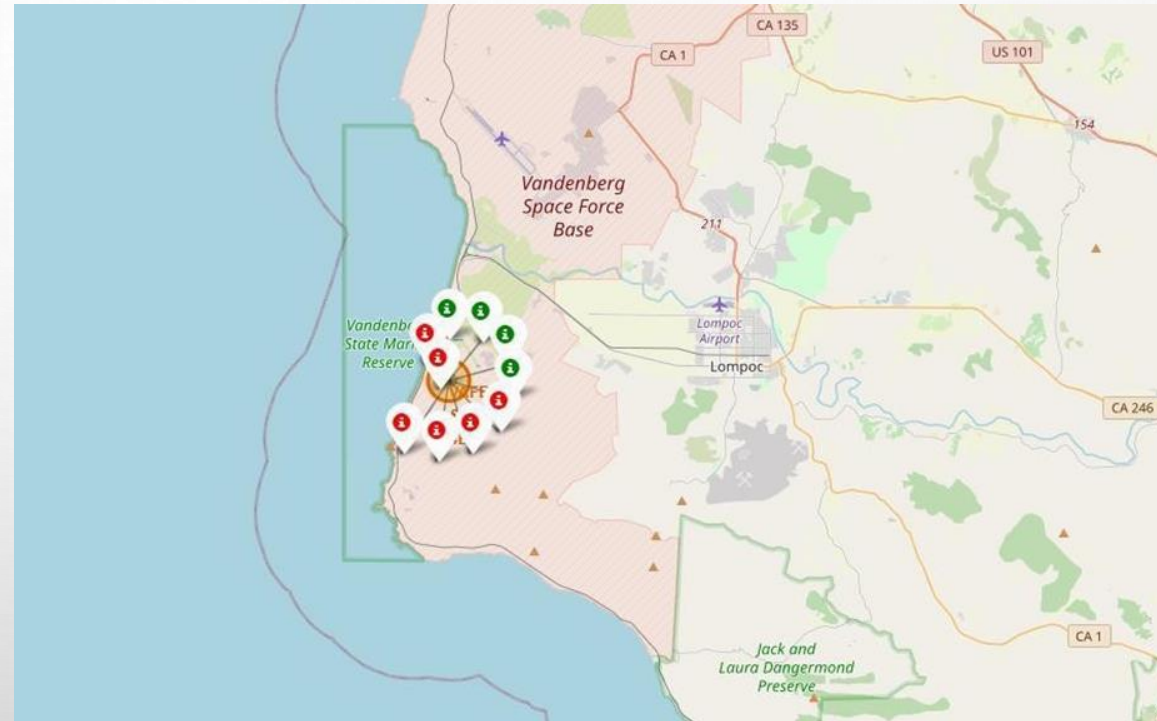
LAUNCH SITE LOCATIONS



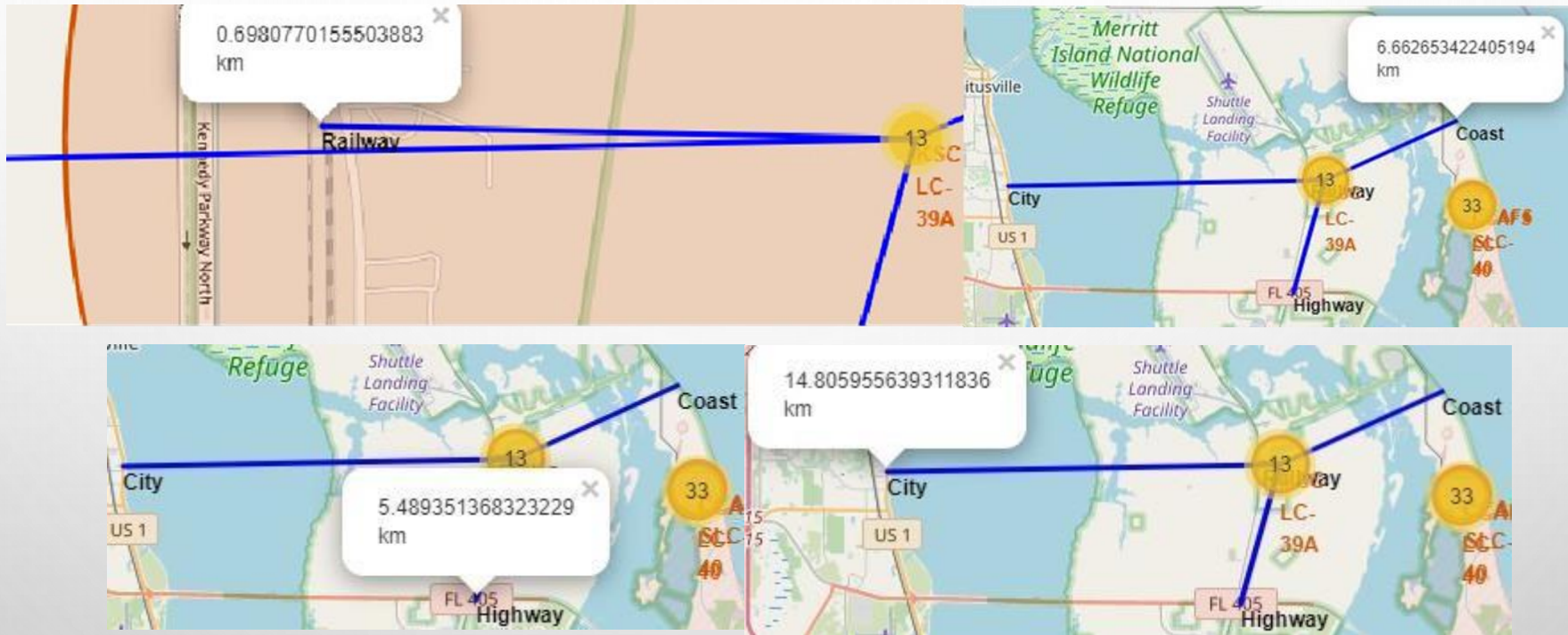
- ✓ The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other.
- ✓ All launch sites are near the ocean.

COLOR CODED LAUNCH MARKERS

- ✓ Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed
- ✓ landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.



KEY LOCATION PROXIMITIES



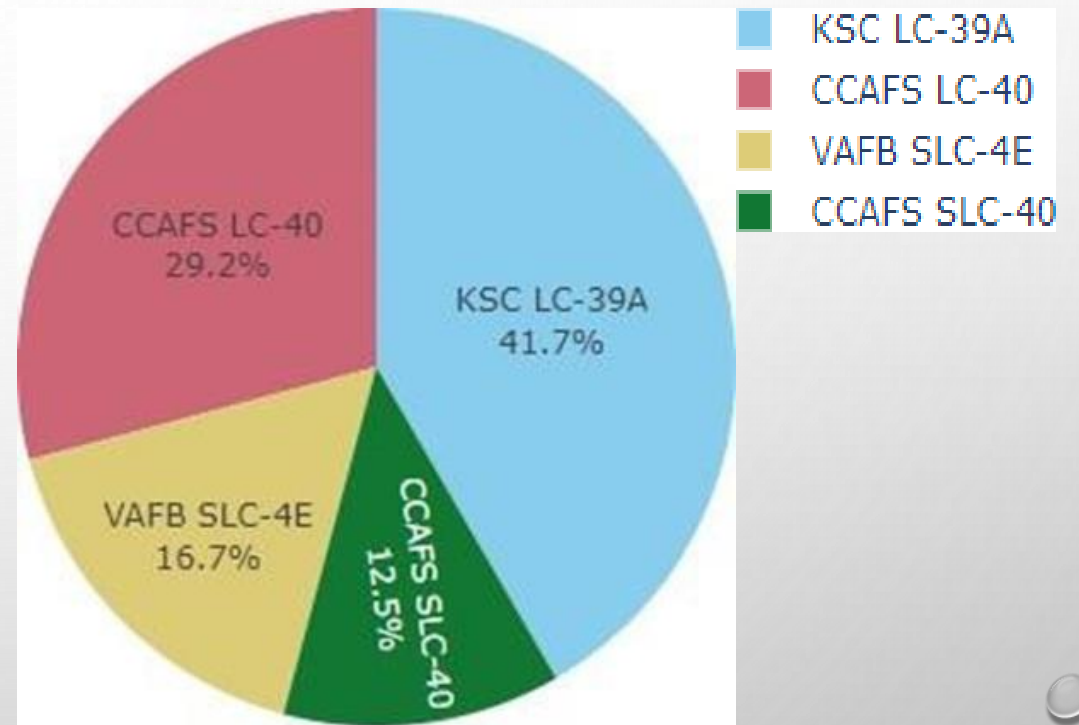
- ✓ Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.



BUILD DASHBOARD WITH PLOTLY DASH

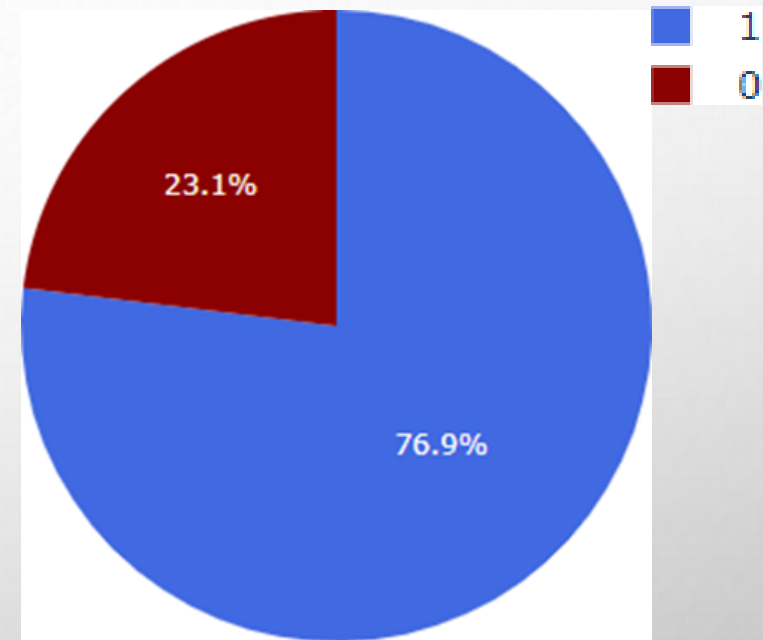
SUCCESSFUL LAUNCHES ACROSS SITES

- ✓ This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings were performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.



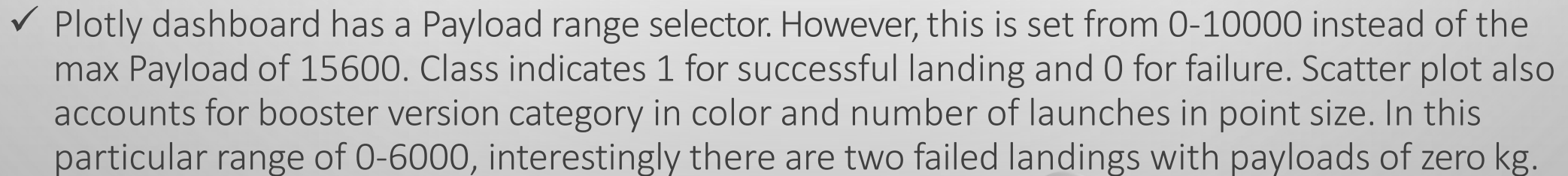
HIGHEST SUCCESS RATE LAUNCH SITE

- ✓ KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.



KSC LC-39A Success Rate (blue=success)

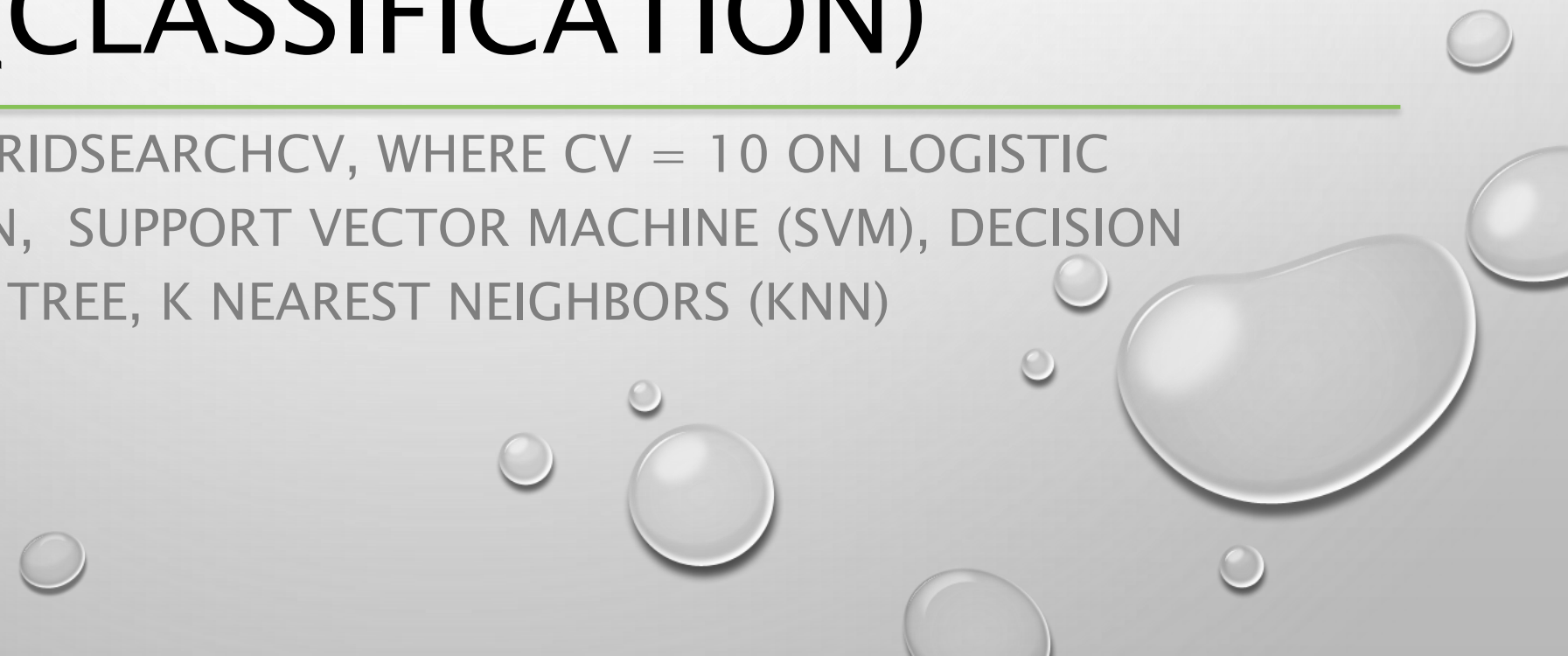
Payload range (Kg):





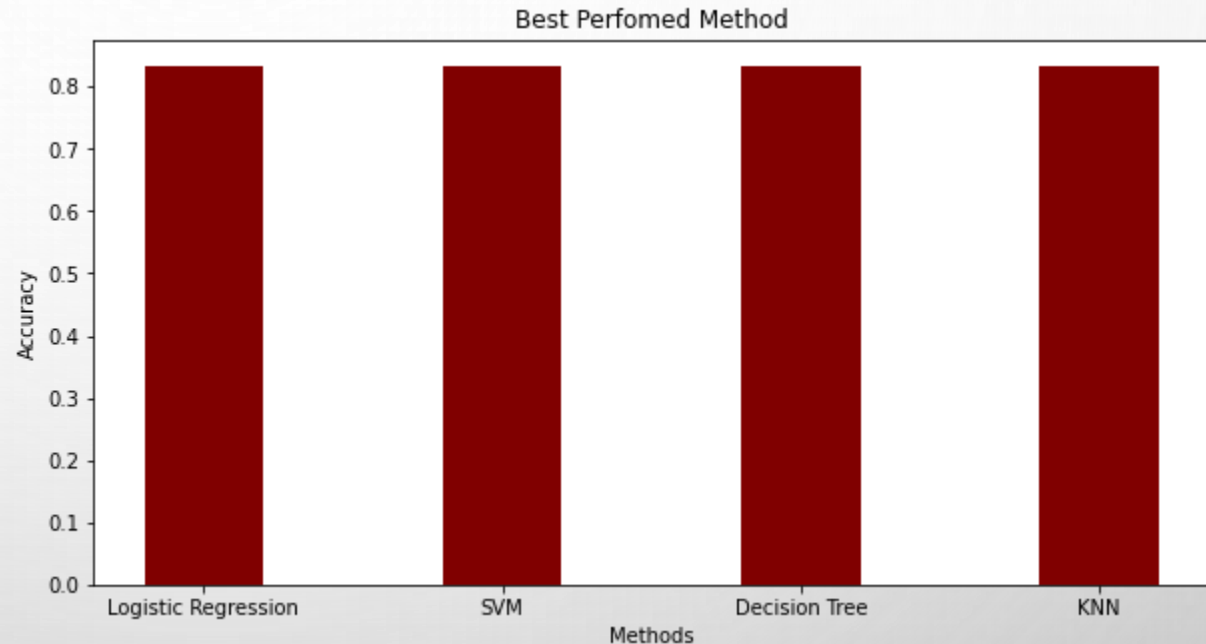
PREDICTIVE ANALYSIS (CLASSIFICATION)

USING GRIDSEARCHCV, WHERE $CV = 10$ ON LOGISTIC
REGRESSION, SUPPORT VECTOR MACHINE (SVM), DECISION
TREE, K NEAREST NEIGHBORS (KNN)



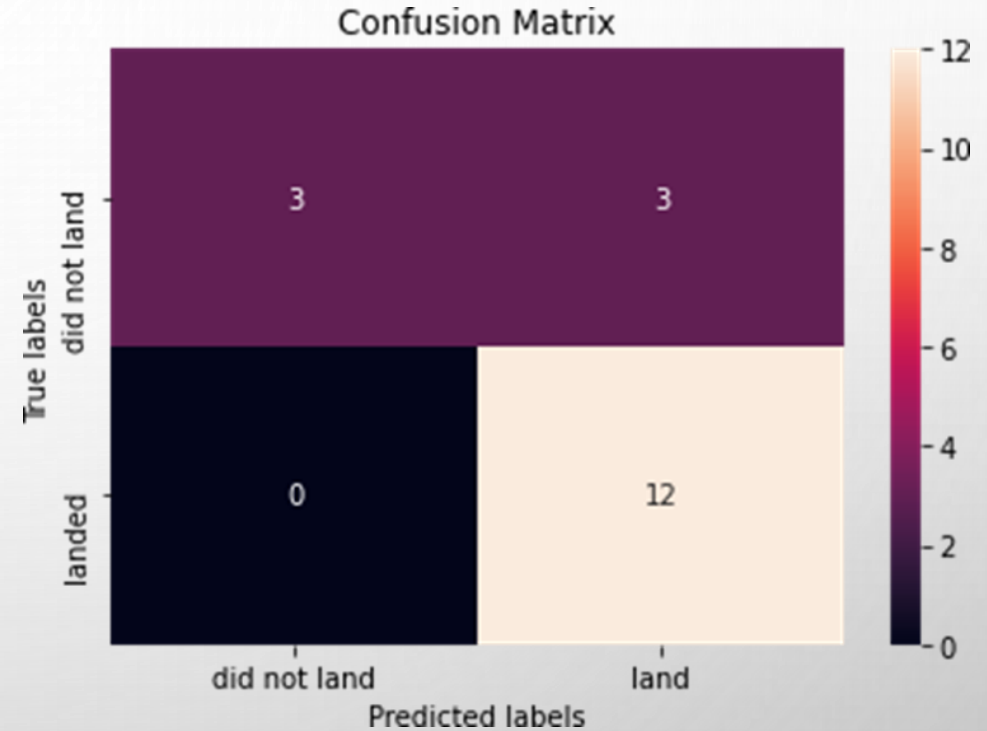
CLASSIFICATION ACCURACY

- ✓ All models had virtually the same accuracy on the test set at 83.33%. It should be noted that test size is small at only sample size of 18.
- ✓ This can cause 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). Our models over predict successful landings.



- ✓ Correct predictions are on a diagonal from top left to bottom right.

CONCLUSION

- ✓ Our task: to develop a machine learning model for space Y who wants to bid against SPACEX.
- ✓ The goal of model is to predict when stage 1 will successfully land to save ~\$100 million USD
- ✓ Used data from a public SPACEX API and web scraping SPACEX Wikipedia page.
- ✓ Created data labels and stored data into a DB2 SQL database.
- ✓ Created a dashboard for visualization.
- ✓ We created a machine learning model with an accuracy of 83%.
- ✓ 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 launch to determine whether the launch should be made or not.
- ✓ If possible more data should be collected to better determine the best machine learning model and improve accuracy.

APPENDIX

✓ GITHUB REPOSITORY URL:


✓ [HTTPS://GITHUB.COM/ABHI1045/IBM_DATA_SCIENCE_PROFESSIONAL_CERTIFICATE_CAPSTONE_PROJECT](https://github.com/ABHI1045/IBM_DATA_SCIENCE_PROFESSIONAL_CERTIFICATE_CAPSTONE_PROJECT)

✓ 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](https://www.coursera.org/professional-certificates/ibm-data-science?#instructors)



THANK YOU!
