IBM DATA SCIENCE CAPSTONE PROJECT – **SPACEX ANALYSIS**

ADEETYA UPADHYAY

29-12-2021

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
  - Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Data Collection with SpaceX API, and Wiki Web Scraping
- Data Wrangling
- Exploratory Data Analysis using:
  - SQL
  - Data Visualization
- Interactive Visual Analytics with Folium and Dashboard creation with Plotly Dash
- Predictive Analysis using Machine Learning:
  - K Nearest Neighbor
  - Decision Tree
  - Support Vector Machine
  - Logistic Regression

INTRODUCTION

Context and Question:

SpaceX 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 SpaceX 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 a client still wishes to consider using Falcon 9 for their launches.

Question: Based on certain factors, how accurately can we determine if a Falcon 9's first stage will land successfully?

Sub-Questions:

- Which factors influence a successful landing?
- To what degree do such factor influence success rate of a landing and what is the consistency of achieving such results?
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate?

Methodology

Procedural Techniques and Purpose of Analysis

METHODOLOGY

- Data collection methodology:
  - SpaceX Rest API
  - (Web Scrapping) from Wikipedia
- Performed data wrangling (Transforming data for Machine Learning)
  - One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- Performed exploratory data analysis (EDA) using visualization and SQL
  - Plotting : Scatter Graphs, Bar Graphs to show relationships between variables to show patterns of data.
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - How to build, tune, evaluate classification models
  - Compared and found the best classification model

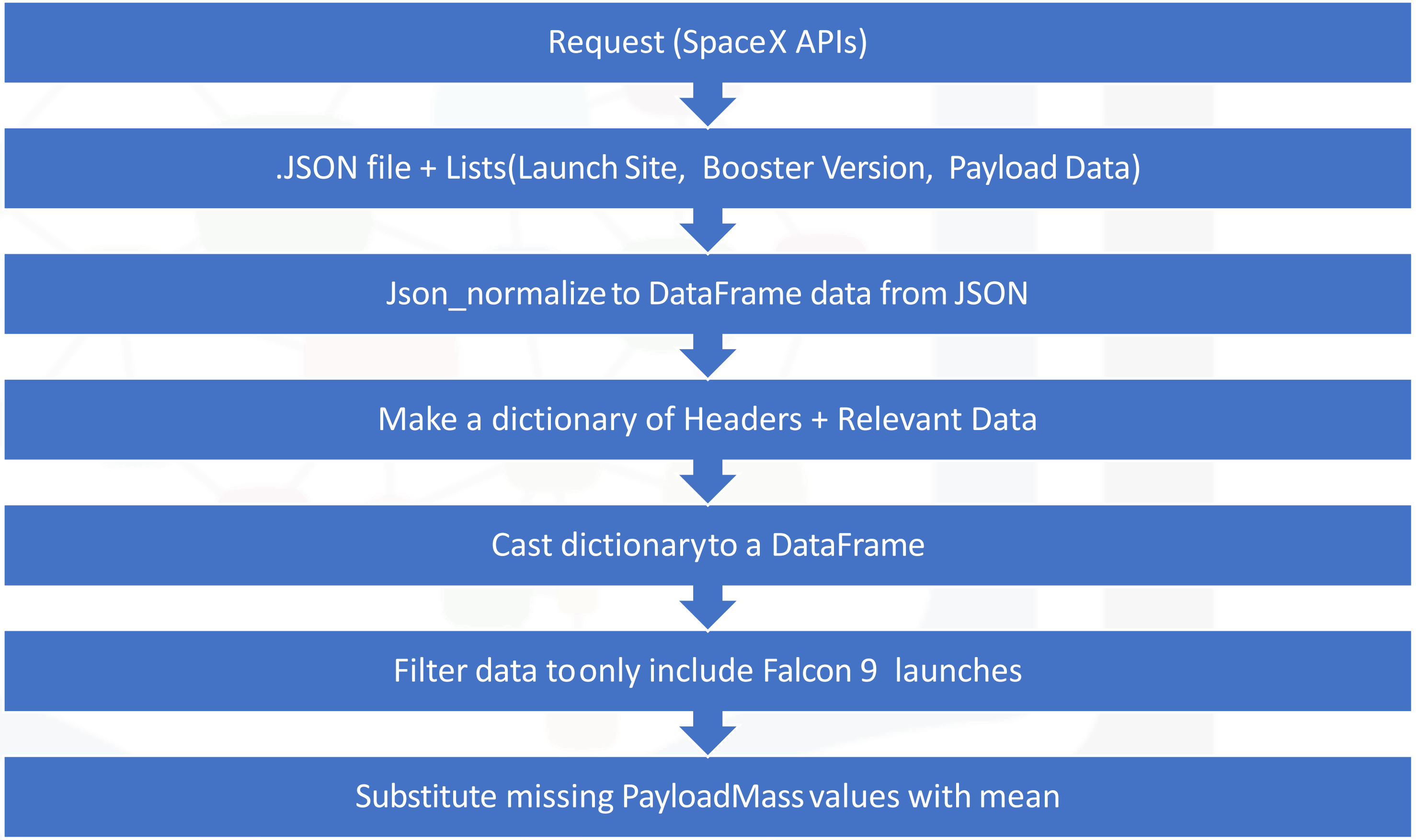
METHODOLOGY

- 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 web-scraping.
- Space X API Data Columns:
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins,Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Web Scraping Data Columns:
    - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

METHODOLOGY

Data Collection – SpaceX API

GitHub File



Request (SpaceX APIs)

.JSON file + Lists(Launch Site, Booster Version, Payload Data)

Json_normalize to DataFrame data from JSON

Make a dictionary of Headers + Relevant Data

Cast dictionaryto a DataFrame

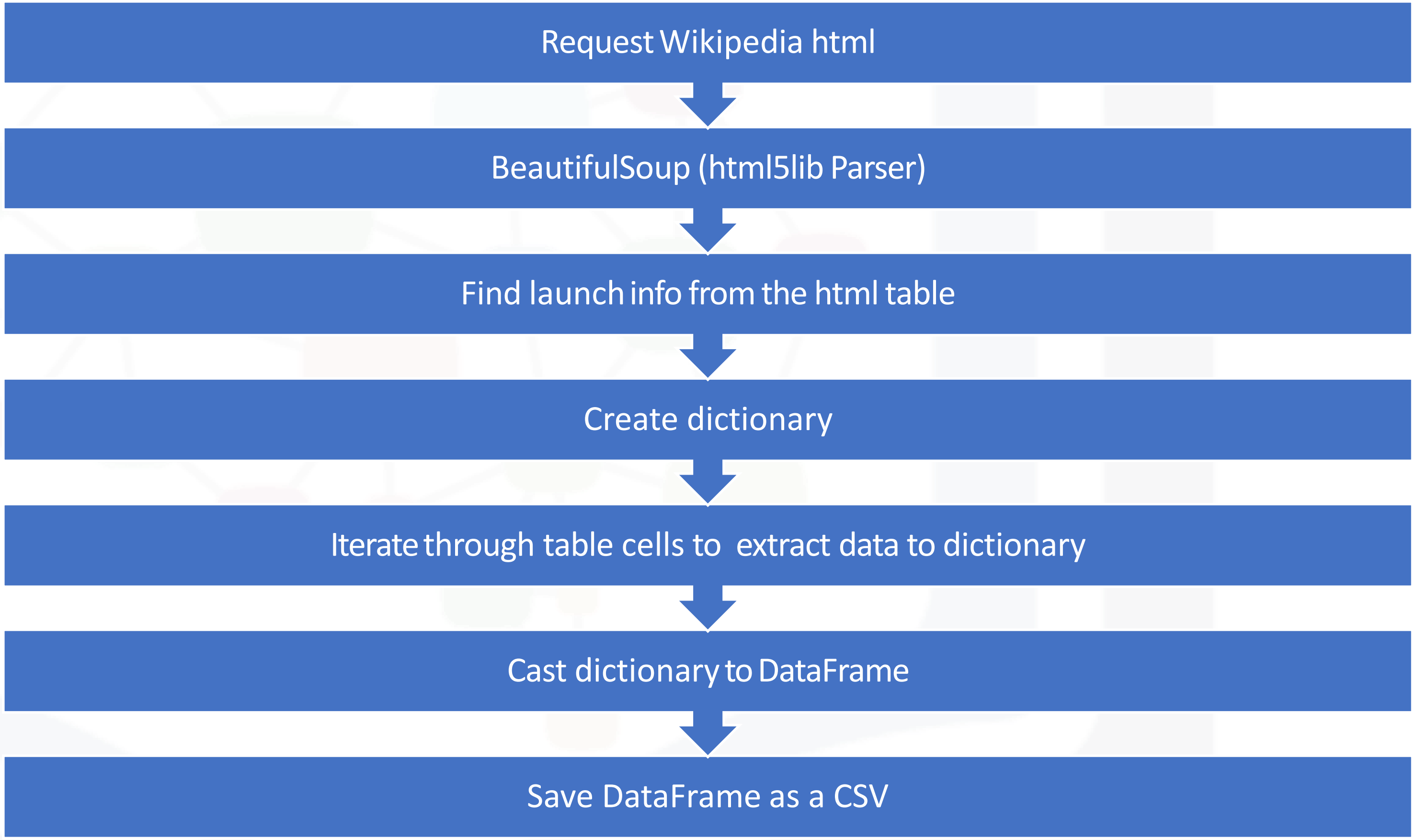
Filter data toonly include Falcon 9 launches

Substitute missing PayloadMass values with mean

METHODOLOGY

Data Collection – Web Scraping

GitHub File



Request Wikipedia html

BeautifulSoup (html5lib Parser)

Find launch info from the html table

Create dictionary

Iterate through table cells to extract data to dictionary

Cast dictionary to DataFrame

Save DataFrame as a CSV

METHODOLOGY

AU | SPACEX

Data Wrangling

GitHub File

- Determine Landing Outcome
  - Create a training label with landing outcomes where successful = 1 & failure = 0.
  - Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’
  - 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

METHODOLOGY

EDA with Data Visualisation

GitHub File

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Plots Used:
  - Scatter plot
  - Line Charts
  - Bar Charts
- 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
- One-hot encoding the data with categorical values so that numerical classification can be performed upon them.

METHODOLOGY

Exploratory Data Analysis with SQL

GitHub File

- 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
- Several functions were used including AVERAGE, COUNT, MAX, MIN and DISTINCT
- Subqueries were made to collect and compare two individual records

METHODOLOGY

Interactive Maps with Folium

GitHub File

- 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 we can visualize successful landings relative to location.
- We also calculated the distance between launch site and proximities to understand the logistical aspect of this process

METHODOLOGY

Dashboard with Plotly Dash

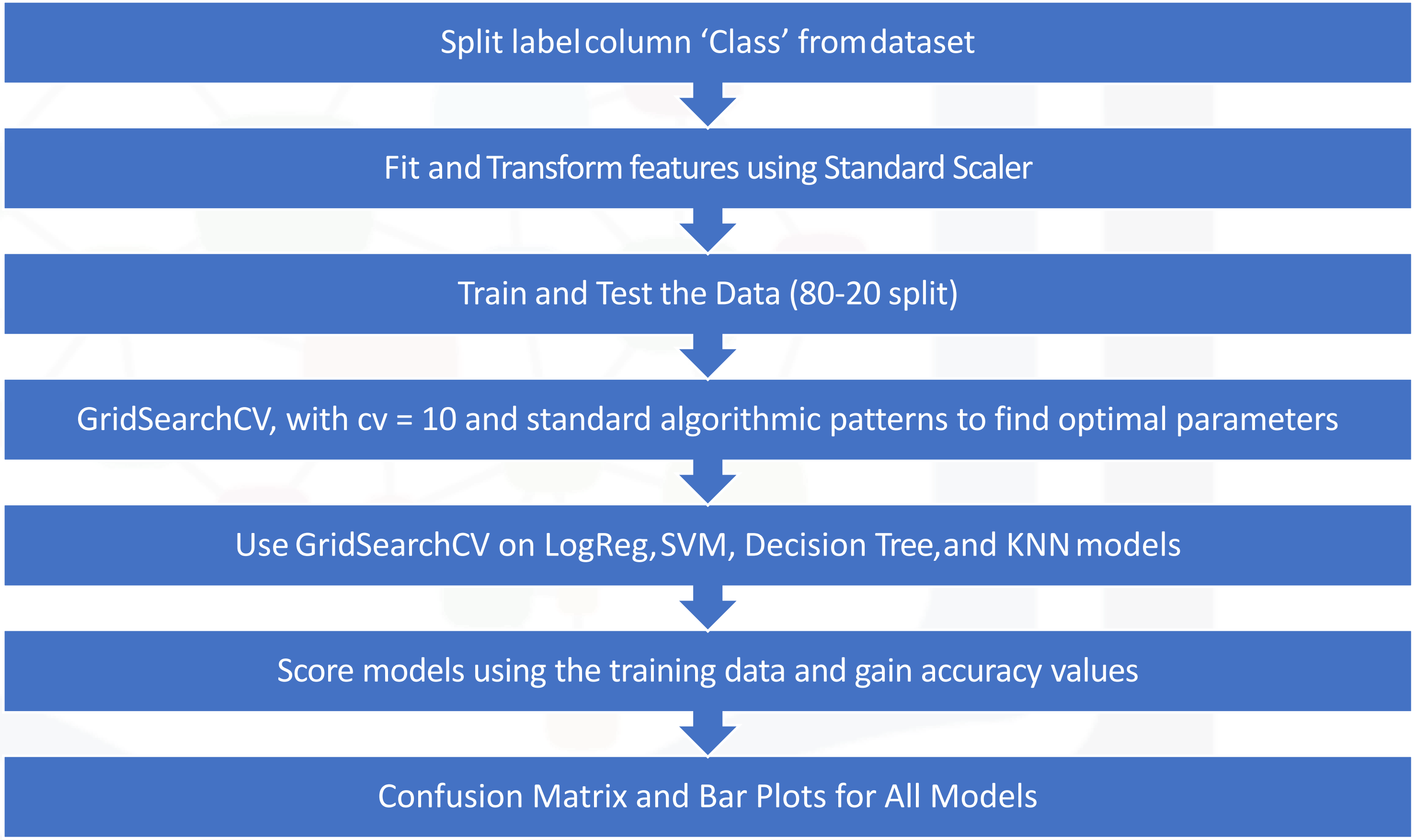
GitHub File

- 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.
- It was built on a dynamic flask server which allowed for a clean GUI platform while running python scripts in the background

METHODOLOGY

Predictive Analysis

GitHub File



Split label column ‘Class’ from dataset

Fit and Transform features using Standard Scaler

Train and Test the Data (80-20 split)

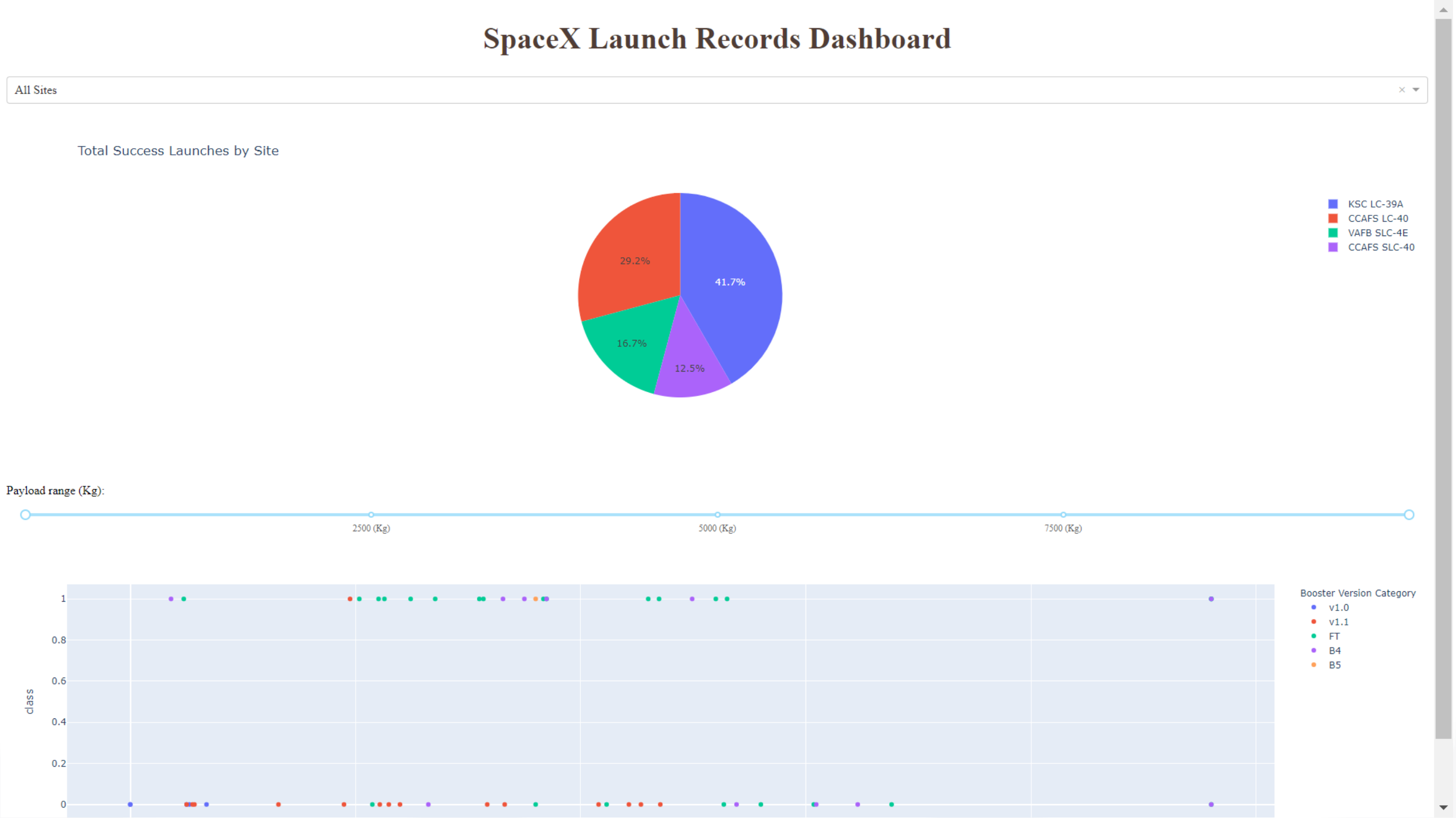
GridSearchCV, with cv = 10 and standard algorithmic patterns to find optimal parameters

Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Score models using the training data and gain accuracy values

Confusion Matrix and Bar Plots for All Models

METHODOLOGY



SpaceX Launch Records Dashboard

All Sites

SpaceX Dashboard with Plotly Dash

GitHub File

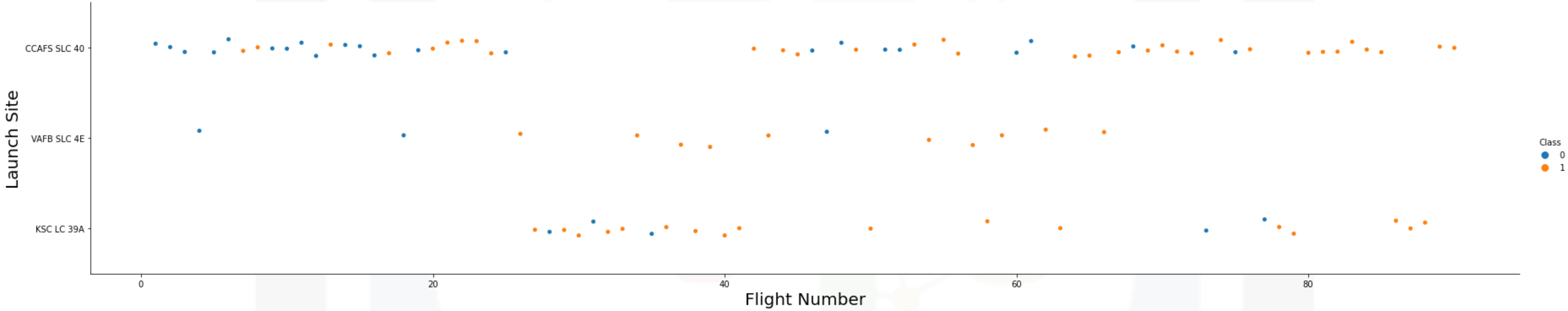
EDA with Visualisation

All Task Results with Analysis

IBM Developer

SKILLS NETWORK

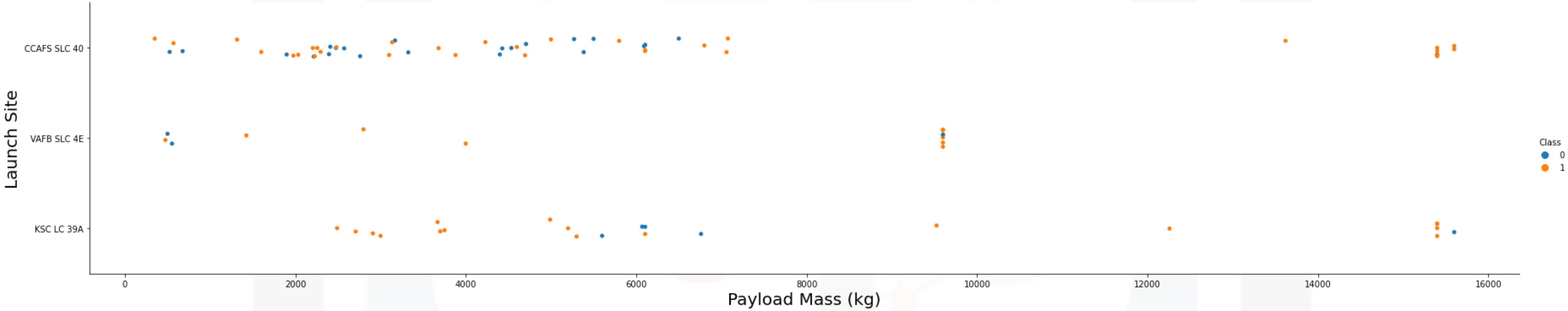
RESULTS – TASK 1



Blue indicates failed launches and orange indicates successful launches. Graph suggests an increase in success rate over time (indicated in Flight Number). CCAFS appears to be the main launch site as it has the most number of take offs.

GitHub File

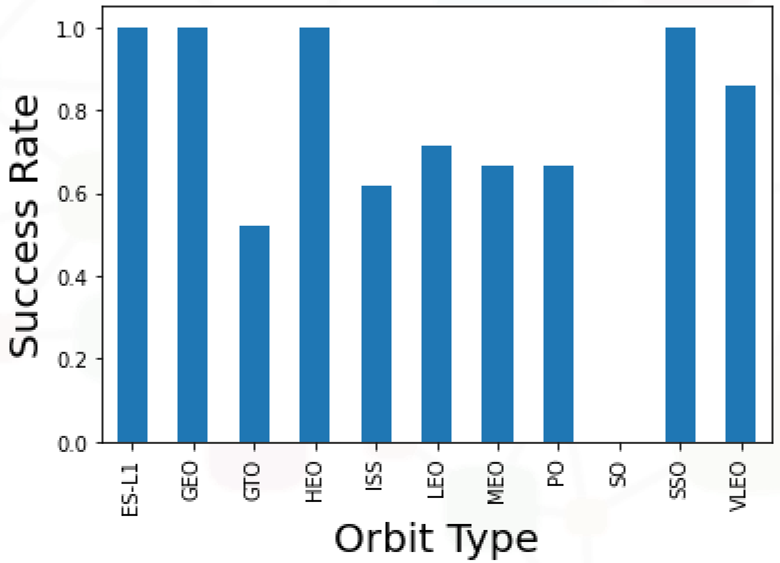
RESULTS – TASK 2



Blue indicates failed launches and orange indicates successful launches. Most of the payload mass is between 0 and 6000 kg and the highest rate of mission failure appears with payload masses of lower weight. Different launch sites too appear to have different payload mass requirements.

GitHub File

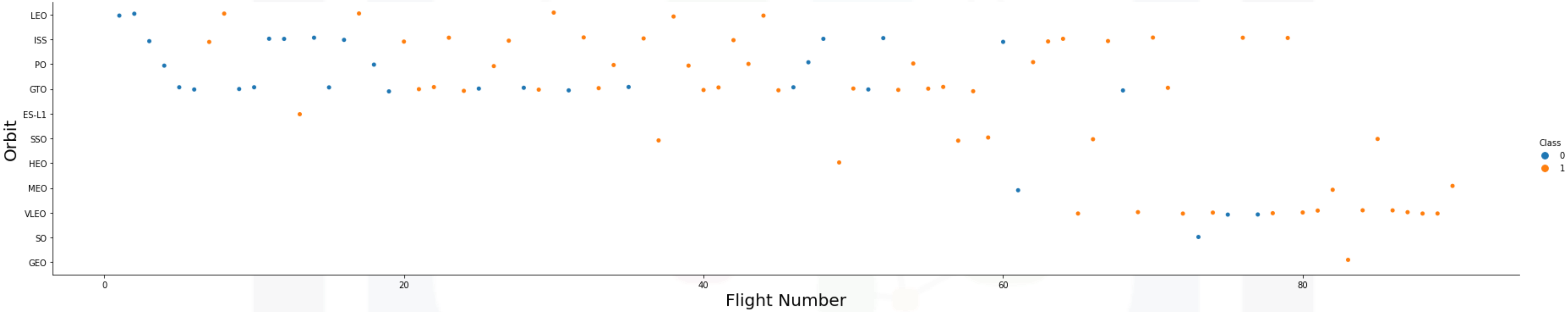
RESULTS – TASK 3



- ES-L1 (1), GEO (1), HEO (1) and SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- A general trend is observable that higher up in the Earth’s orbit, the higher the chances of success rate, however, this could be due to fewer missions

GitHub File

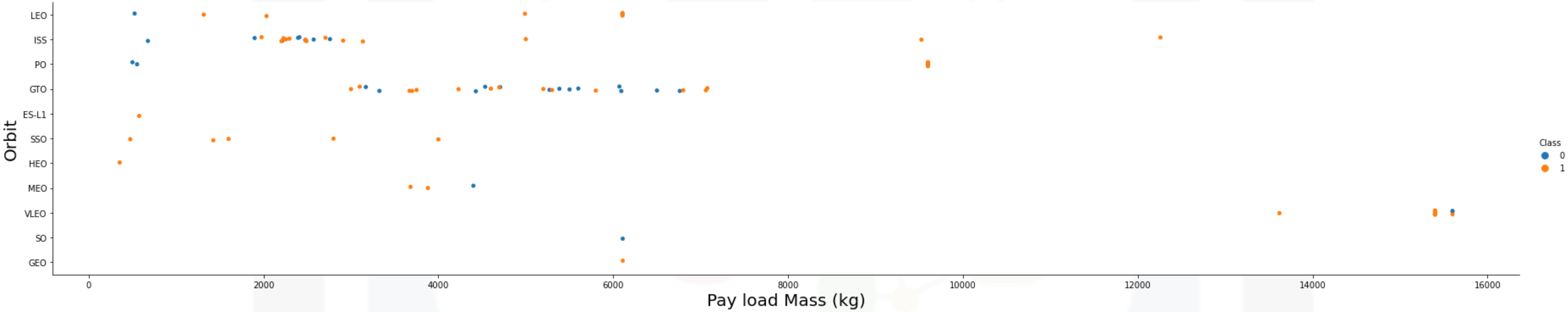
RESULTS – TASK 4



Blue indicates failed launches and orange indicates successful launches. Most of the payload mass is between 0 and 6000 kg and the highest rate of mission failure appears with payload masses of lower weight. Different launch sites too appear to have different payload mass requirements.

GitHub File

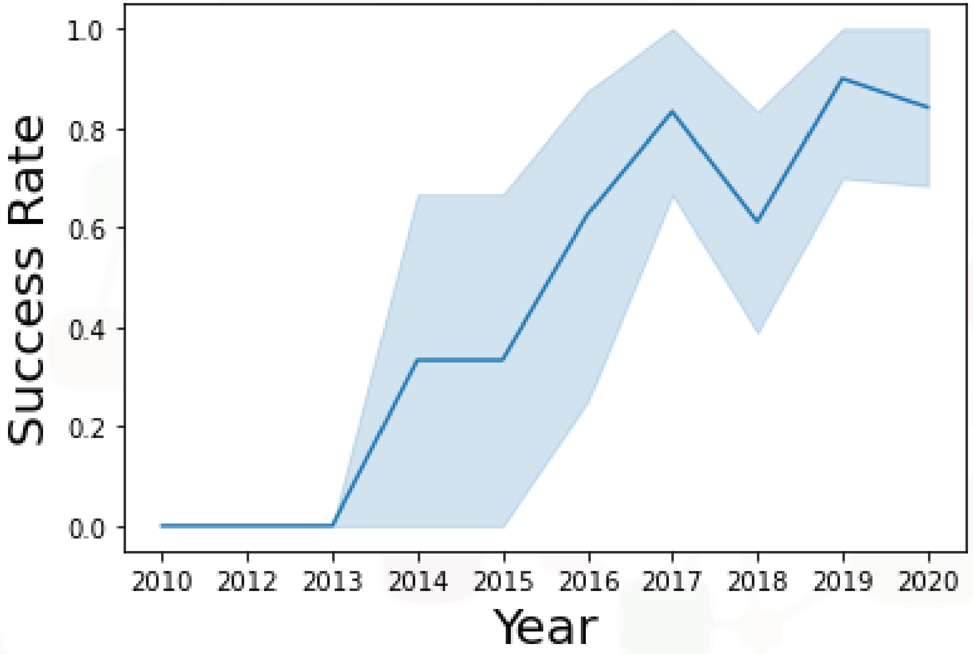
RESULTS – TASK 5



Blue indicates failed launches and orange indicates successful launches. 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. This suggests that higher the orbit, higher the payload.

GitHub File

RESULTS – TASK 6



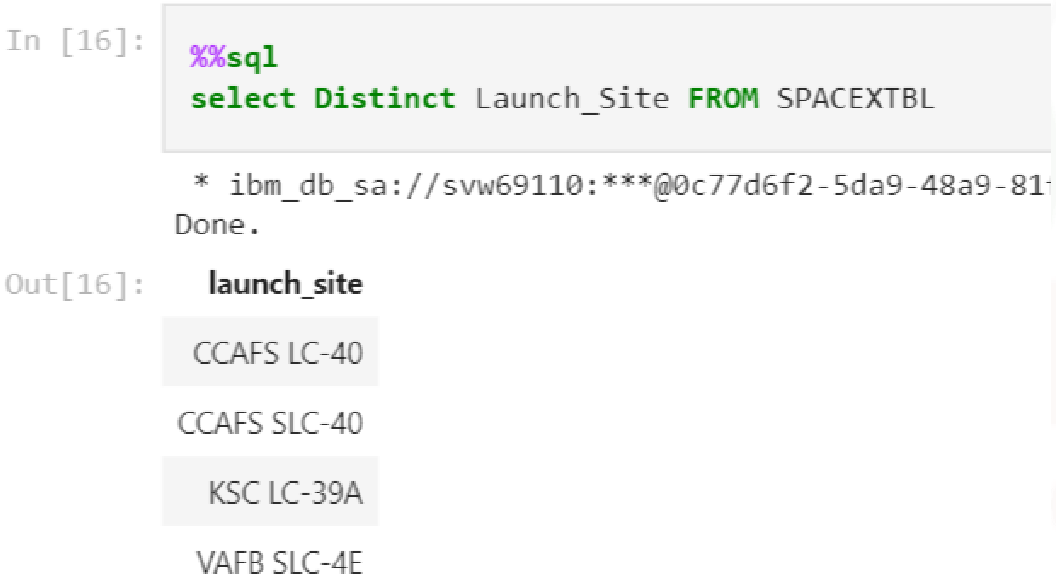
Success generally increases over time since 2013 with a slight dip in 2018. For the first 3 years, there is an indication of complete failure of launches, before a sharp increase. Success in recent years at around 80% and will continue to rise.

GitHub File

EDA with SQL

All Task Results with Analysis

RESULTS – TASK 1



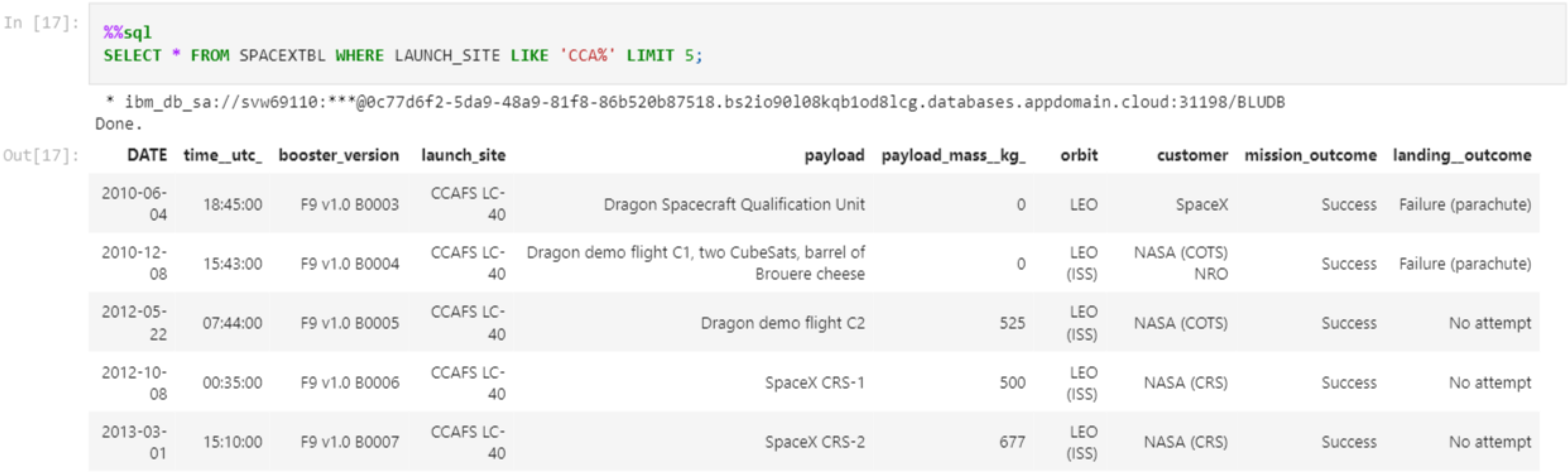
```
In [16]:  %%sql
          select Distinct Launch_Site FROM SPACEXTBL

           * ibm_db_sa://svw69110:***@0c77d6f2-5da9-48a9-81
          Done.

Out[16]:    launch_site
           CCAFS LC-40
          CCAFS SLC-40
            KSC LC-39A
           VAFB SLC-4E
```

- Query unique launch site names from database.

- CCAFS LC-40 was the previous name of CCAFS SLC-40. Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

GitHub File

RESULTS – TASK 2



In [17]:

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

```
 * ibm_db_sa://svw69110:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/BLUDB
Done.
```

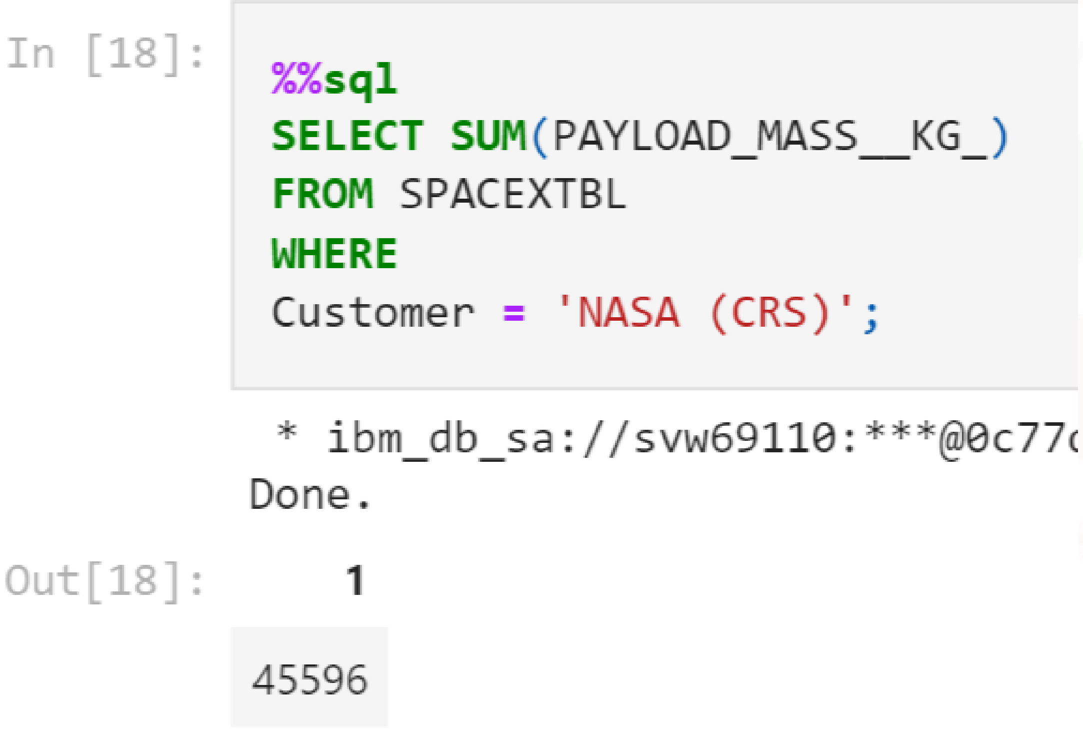
Out[17]:

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 5 Launch sites are all CCAFS LC-40, that begin with CCA

GitHub File

RESULTS – TASK 3

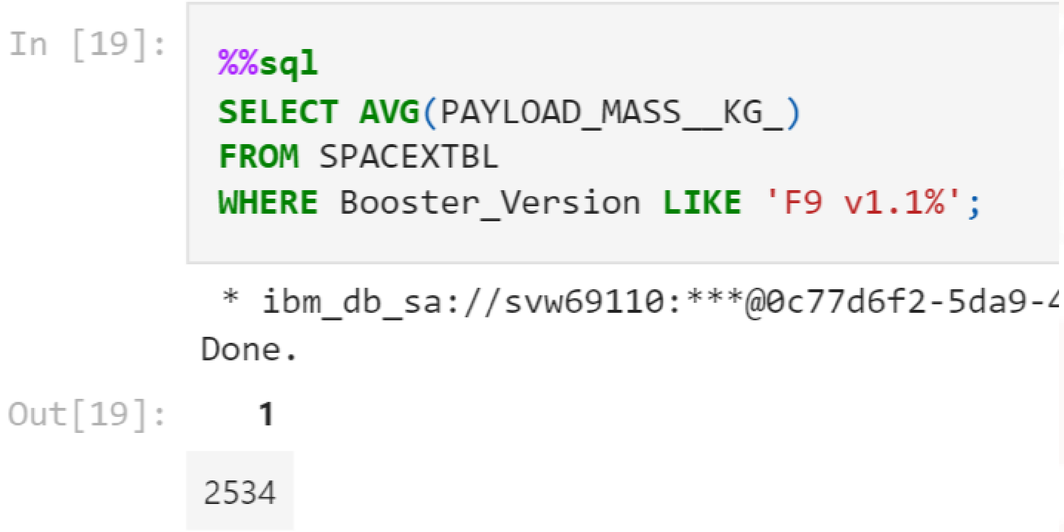


```
In [18]:  %%sql
          SELECT SUM(PAYLOAD_MASS__KG_)
          FROM SPACEXTBL
          WHERE
          Customer = 'NASA (CRS)';

           * ibm_db_sa://svw69110:***@0c77
          Done.

Out[18]:      1
          45596
```

- 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).

GitHub File

RESULTS – TASK 4



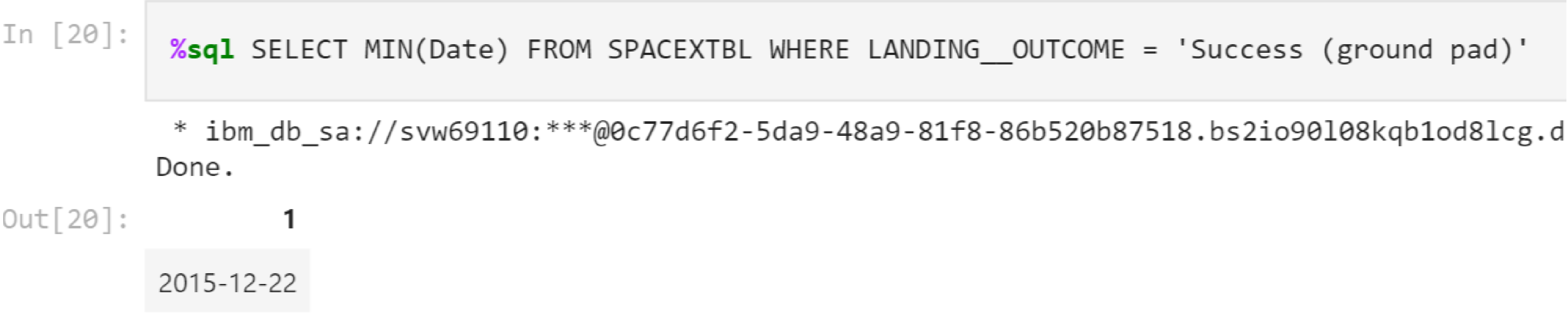
```
In [19]: %%sql
         SELECT AVG(PAYLOAD_MASS__KG_)
         FROM SPACEXTBL
         WHERE Booster_Version LIKE 'F9 v1.1%';
```

```
 * ibm_db_sa://svw69110:***@0c77d6f2-5da9-4
Done.
Out[19]:  1
         2534
```

- 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

GitHub File

RESULTS – TASK 5

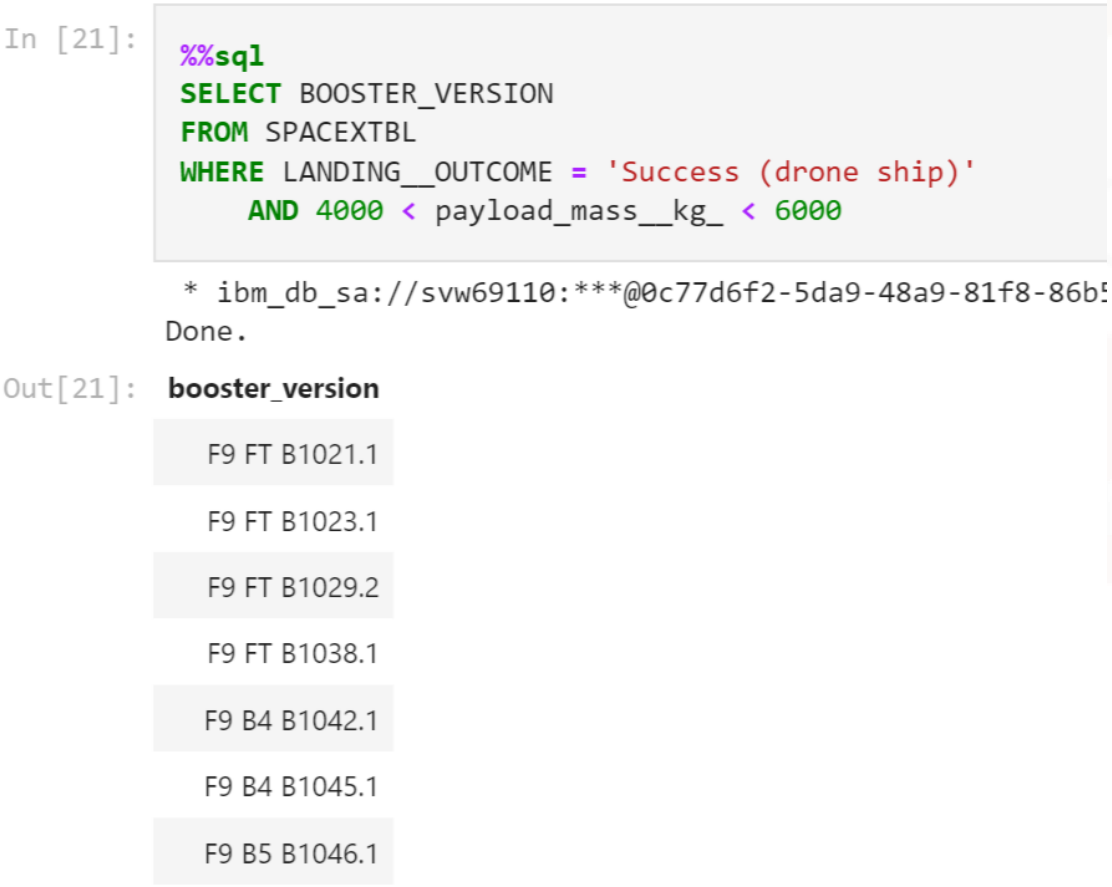


```
In [20]: %sql SELECT MIN(Date) FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)'

          * ibm_db_sa://svw69110:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.d
         Done.

Out[20]:          1
         2015-12-22
```

- 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 started appearing towards the end of 2014.

GitHub File

RESULTS – TASK 6

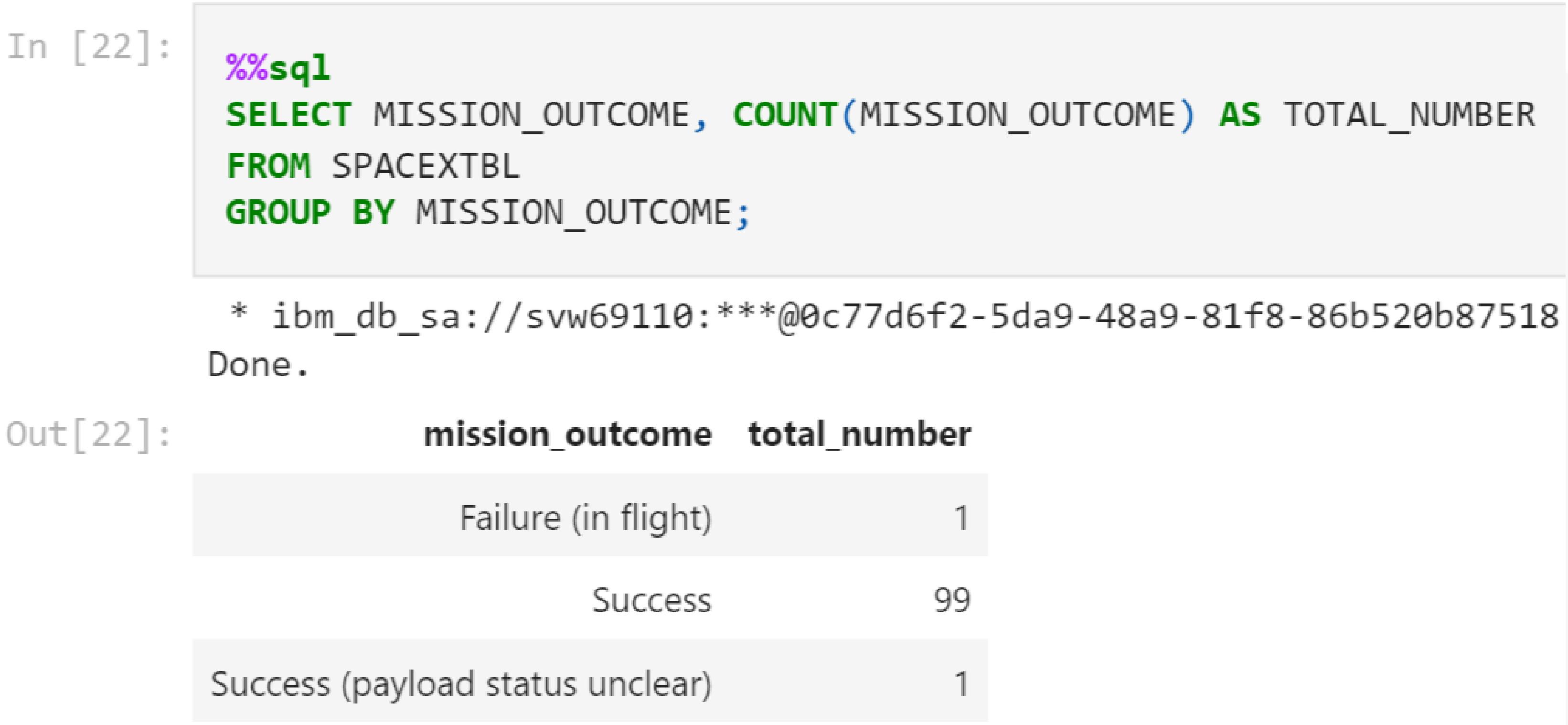


```
In [21]:  %%sql
          SELECT BOOSTER_VERSION
          FROM SPACEXTBL
          WHERE LANDING__OUTCOME = 'Success (drone ship)'
              AND 4000 < payload_mass__kg_ < 6000

           * ibm_db_sa://svw69110:***@0c77d6f2-5da9-48a9-81f8-86b5
          Done.

Out[21]:  booster_version
              F9 FT B1021.1
              F9 FT B1023.1
              F9 FT B1029.2
              F9 FT B1038.1
              F9 B4 B1042.1
              F9 B4 B1045.1
              F9 B5 B1046.1
```

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

GitHub File

RESULTS – TASK 7

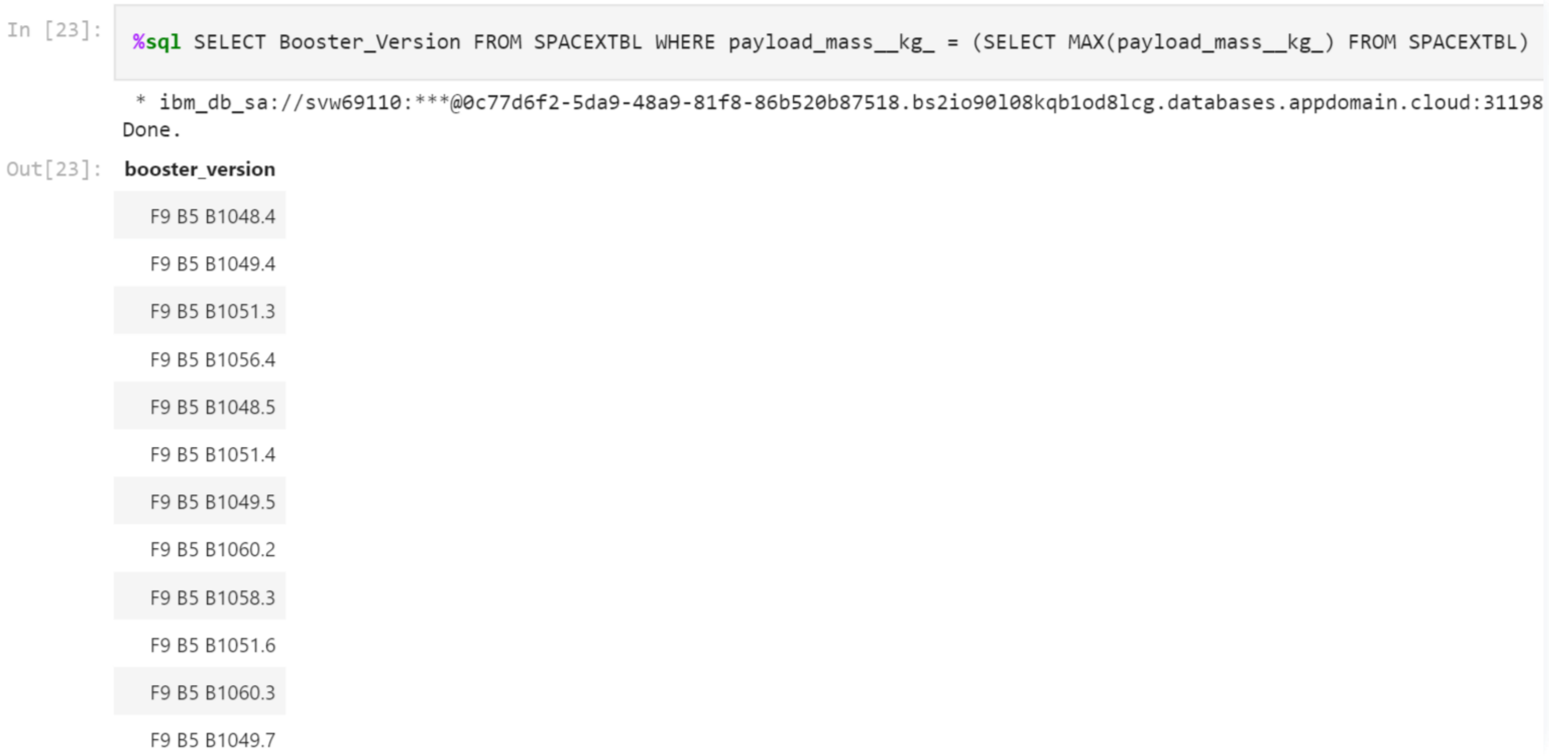


```
In [22]:  %%sql
          SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
          FROM SPACEXTBL
          GROUP BY MISSION_OUTCOME;

           * ibm_db_sa://svw69110:***@0c77d6f2-5da9-48a9-81f8-86b520b87518
          Done.
Out[22]:
```

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

- This query returns a count of each mission outcome. SpaceX appears to achieve its mission outcome nearly 99% of the time. This indicates that once the spacecraft has successfully launched, the success of the mission is extremely high. Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

RESULTS – TASK 8



In [23]:

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

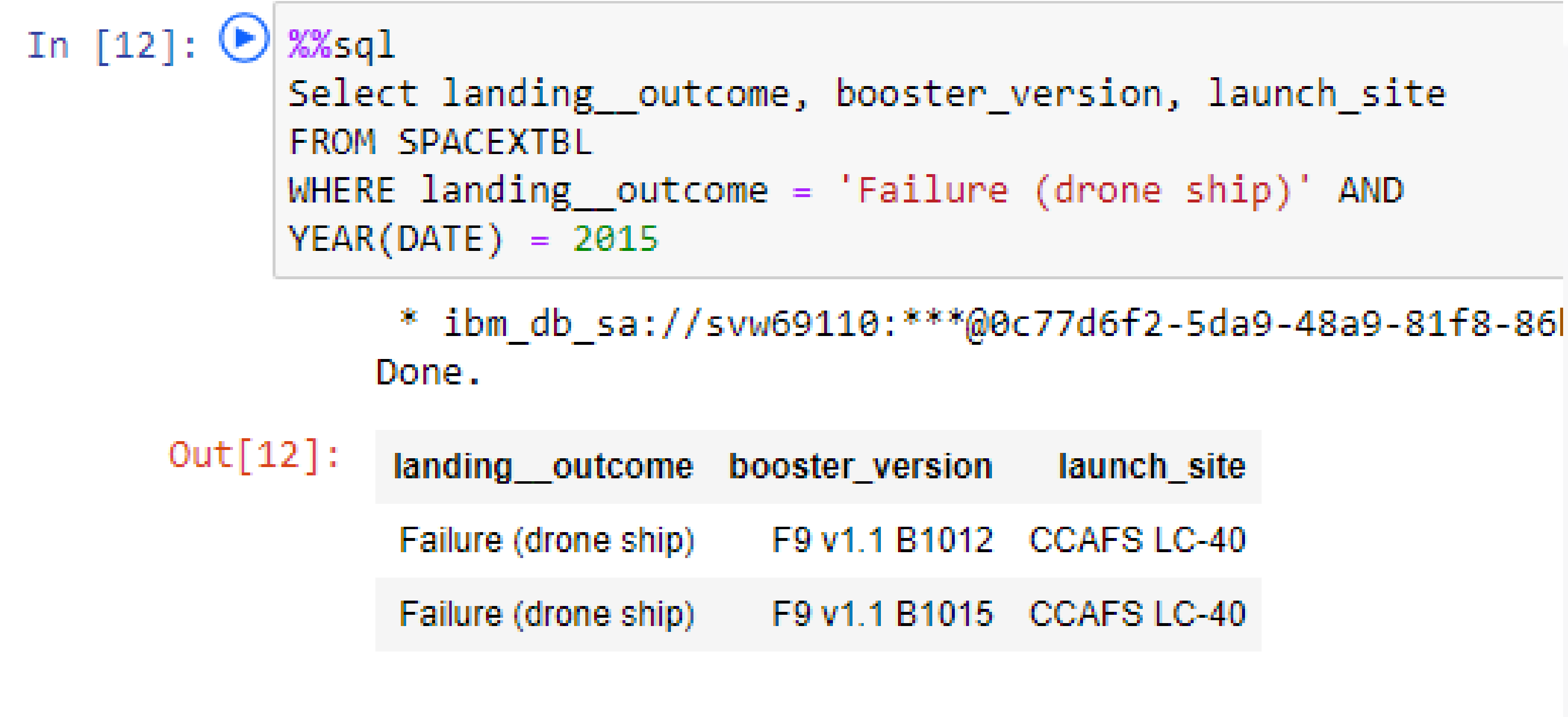
```
 * ibm_db_sa://svw69110:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198
Done.
```

Out[23]:

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

- 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.

RESULTS – TASK 9



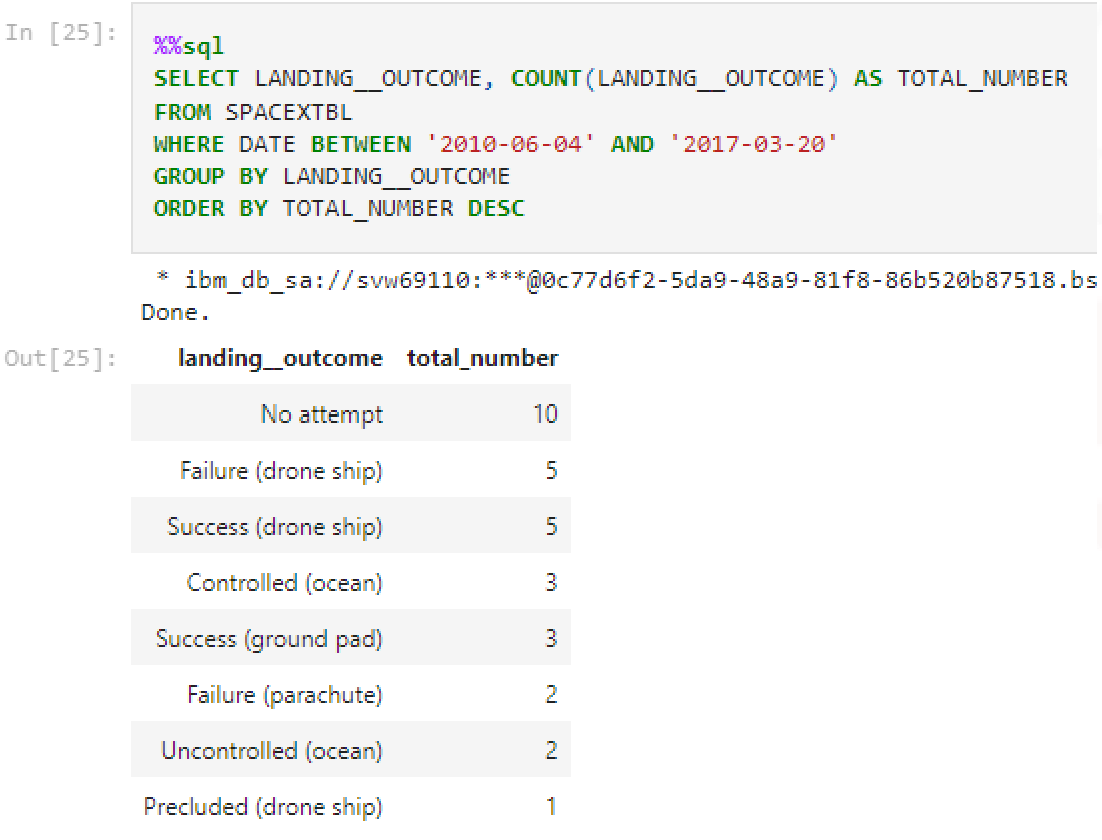
```
In [12]:  %%sql
          Select landing__outcome, booster_version, launch_site
          FROM SPACEXTBL
          WHERE landing__outcome = 'Failure (drone ship)' AND
          YEAR(DATE) = 2015

             * ibm_db_sa://svw69110:***@0c77d6f2-5da9-48a9-81f8-86
            Done.
```

Out[12]:

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

- 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.

RESULTS – TASK 10



In [25]:

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

```
 * ibm_db_sa://svw69110:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs
Done.
```

Out[25]:

landing__outcome	total_number
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

- 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

GitHub File

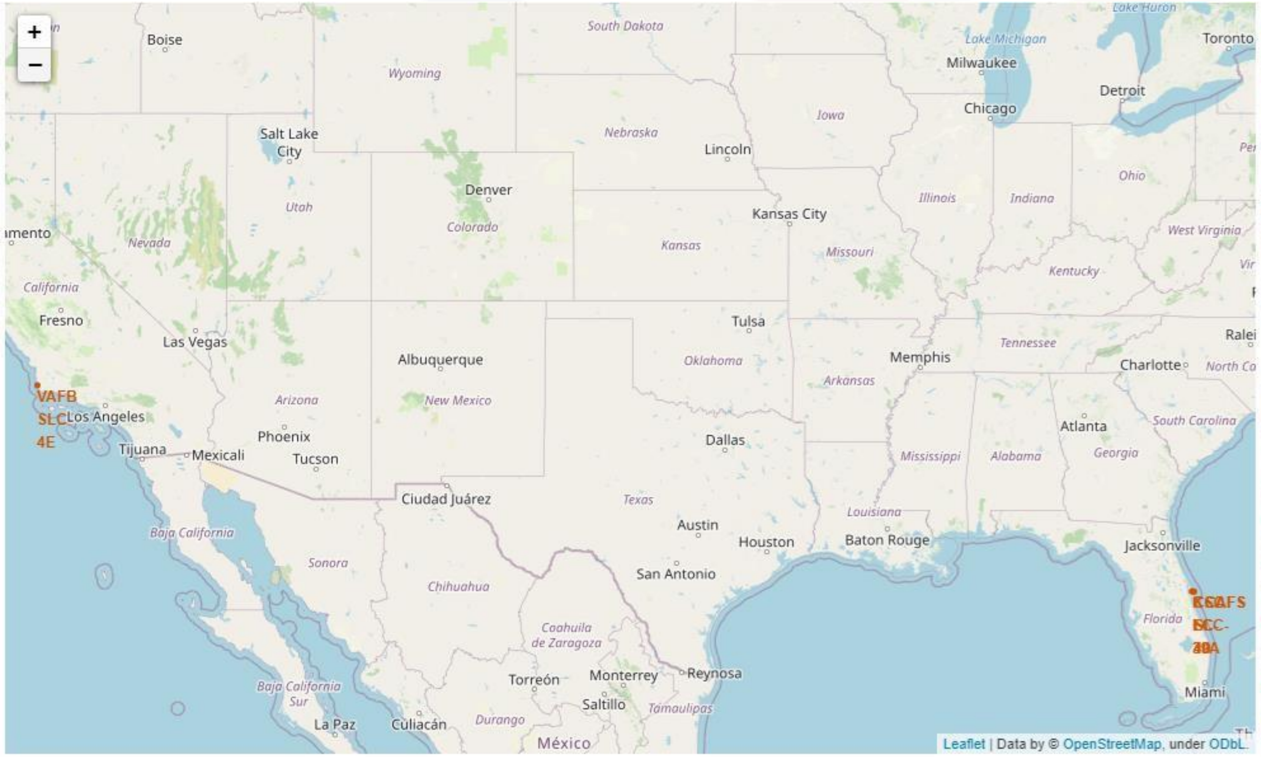
Interactive Maps with Follium

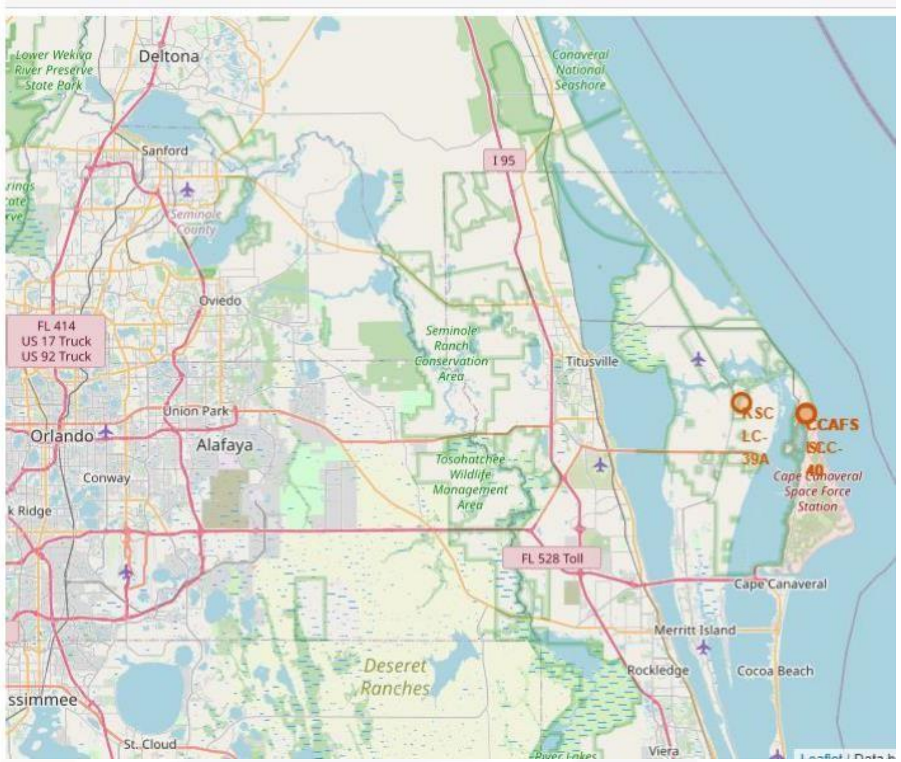
All Task Results with Analysis

IBM Developer

SKILLS NETWORK

RESULTS – TASK 1

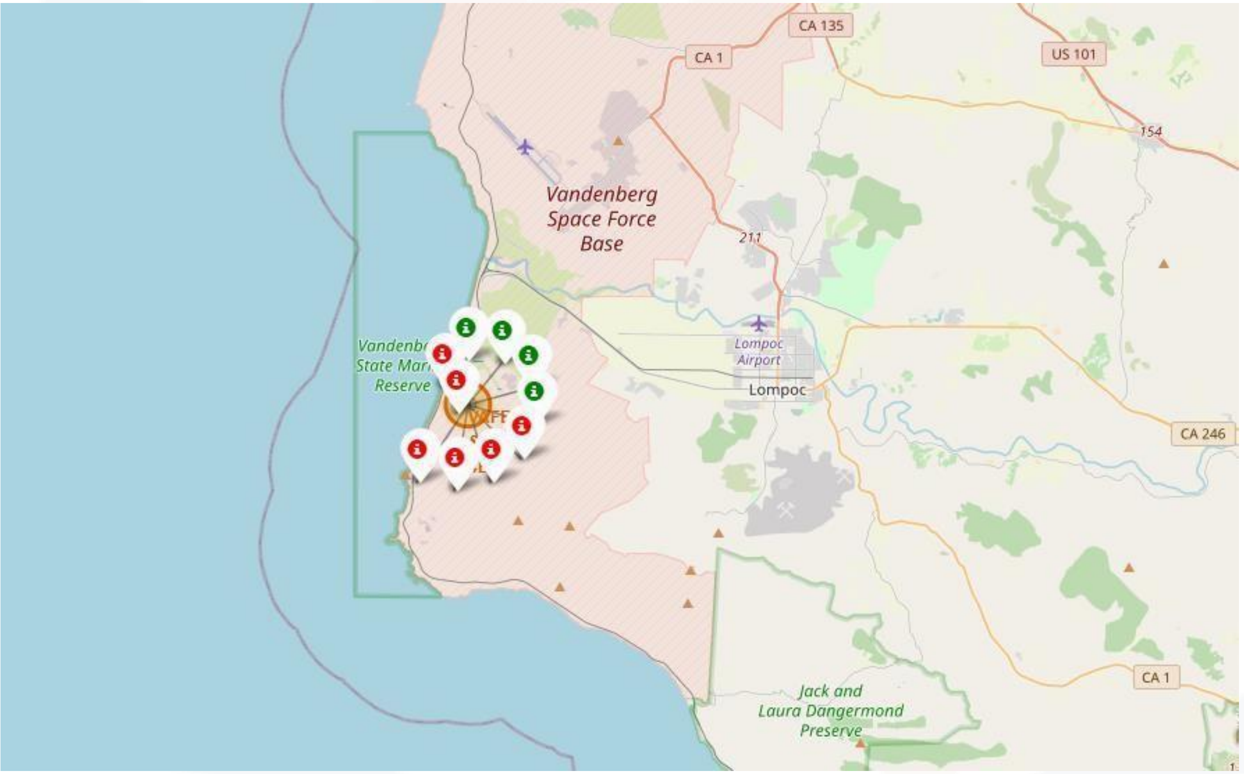




- The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

GitHub File

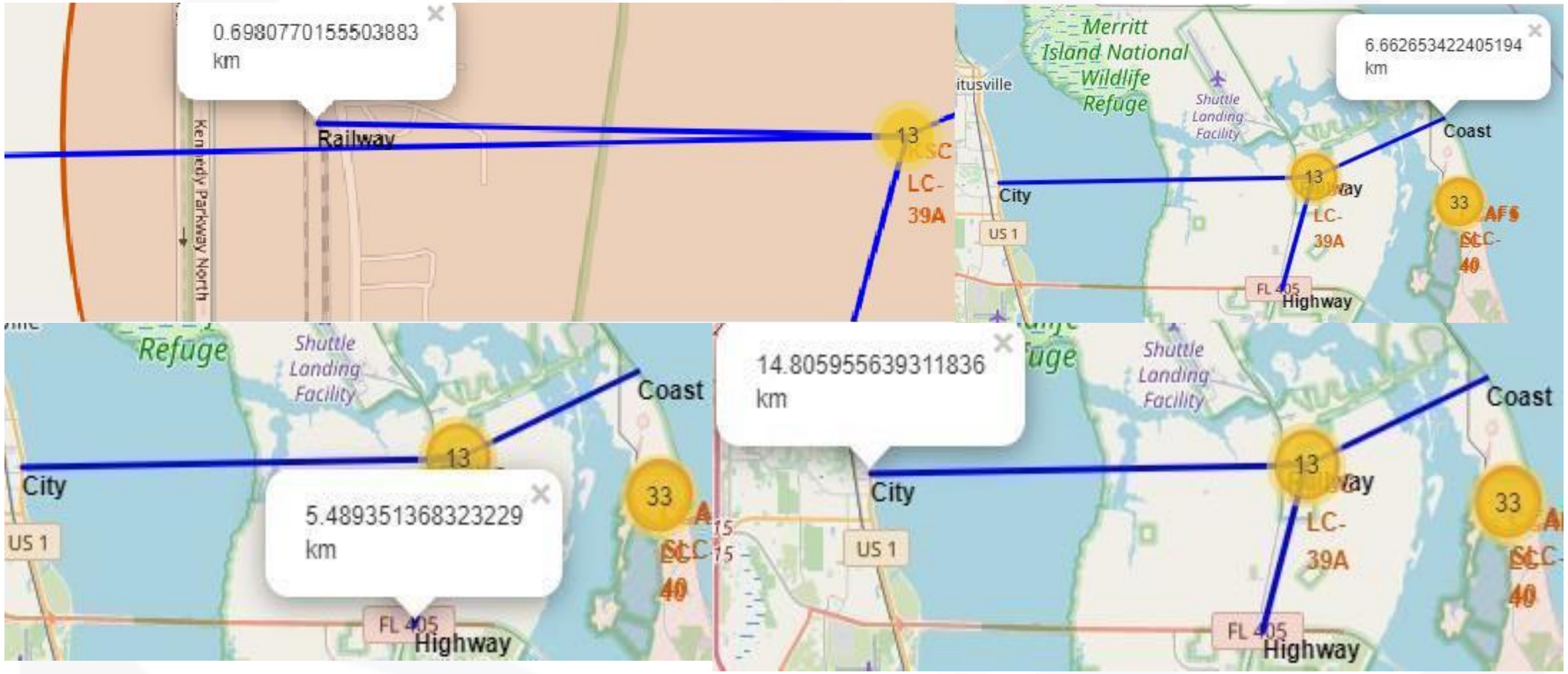
RESULTS – TASK 2


CA 135
CA 1
US 101
154
Vandenberg Space Force Base
211
Lompoc Airport
Lompoc
Vandenberg State Marine Reserve
CA 246
CA 1
Jack and Laura Dangermond Preserve

- 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.

GitHub File

RESULTS – TASK 3



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.

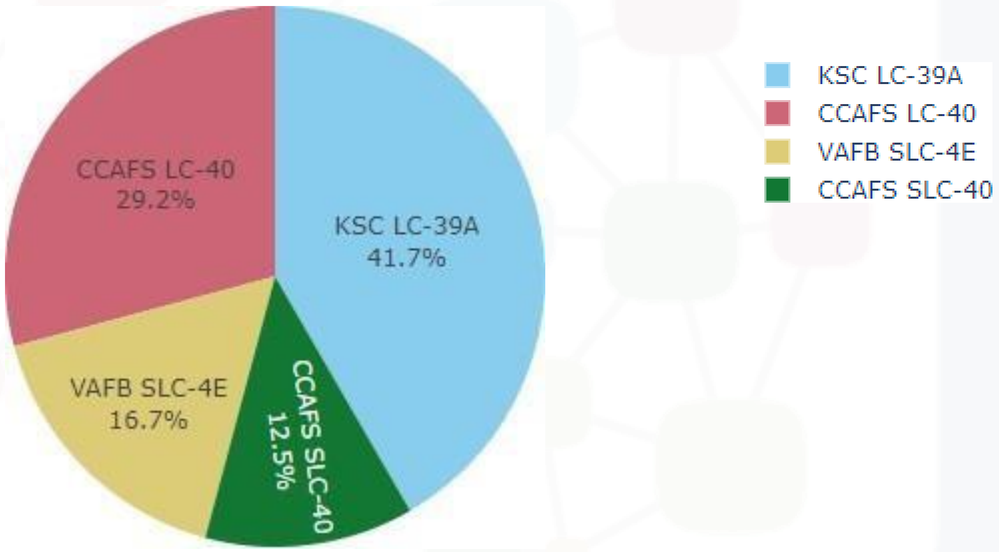
Plotly-Dash Dashboard

All Task Results with Analysis

IBM Developer

SKILLS NETWORK

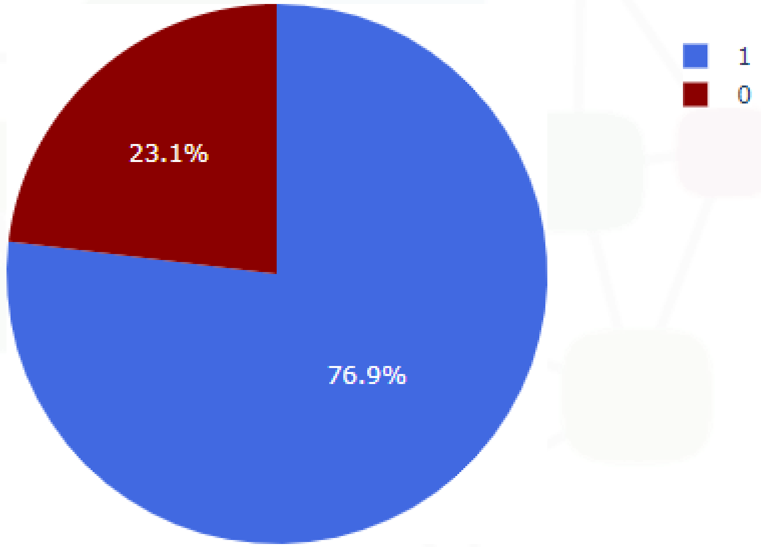
RESULTS – TASK 1



- 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 where 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.

GitHub File

RESULTS – TASK 2

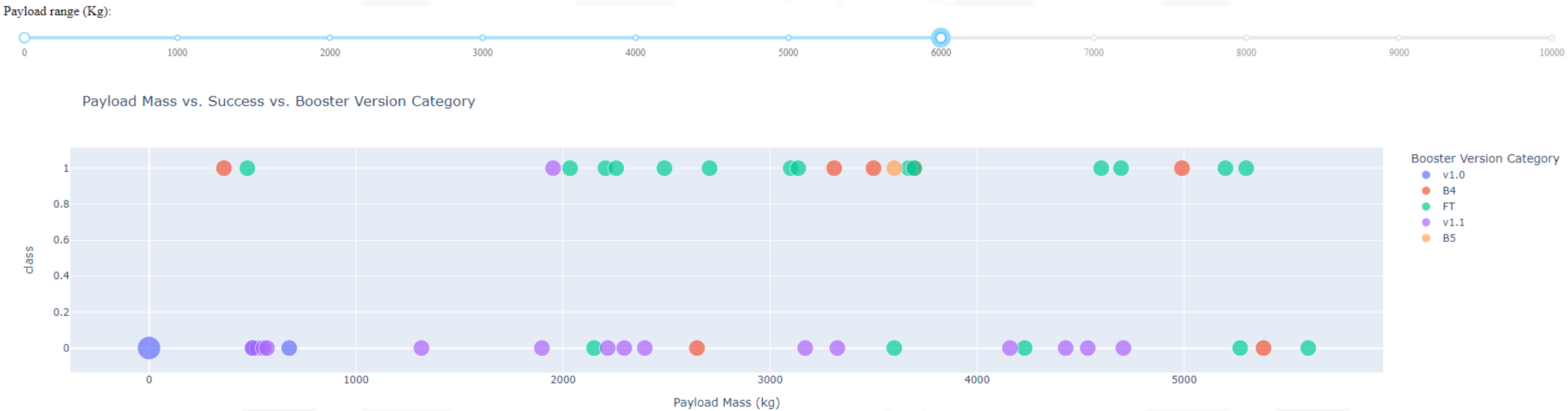


- This is the plot of the highest success rate of launch sites. KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings. (77% and 23% respectively)

[GitHub File]

RESULTS – TASK 3

AU | SPACEX



Payload range (Kg):

Payload Mass vs. Success vs. Booster Version Category

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. The highest number of successful landings are accounted for within the range 2000-4000 kgs

GitHub File

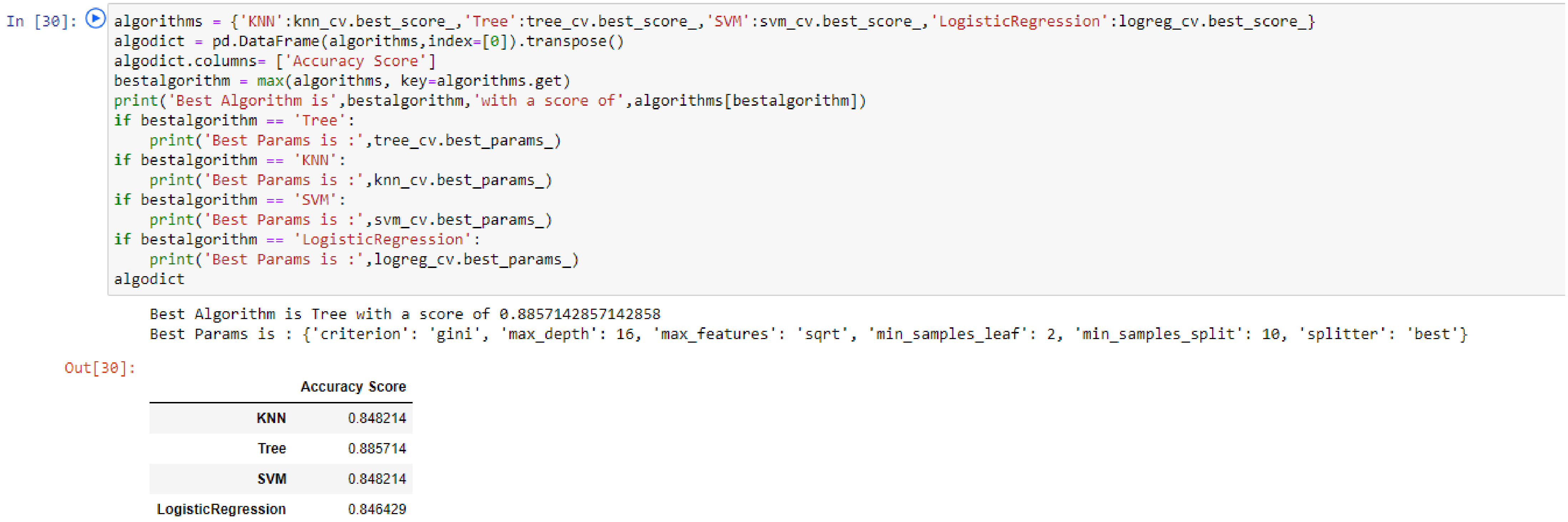
IBM Developer

SKILLS NETWORK

Predictive Analysis (Classification)

All Task Results with Analysis

RESULTS



In [30]:

```
algorithms = {'KNN':knn_cv.best_score_,'Tree':tree_cv.best_score_,'SVM':svm_cv.best_score_,'LogisticRegression':logreg_cv.best_score_}
algodict = pd.DataFrame(algorithms,index=[0]).transpose()
algodict.columns= ['Accuracy Score']
bestalgorithm = max(algorithms, key=algorithms.get)
print('Best Algorithm is',bestalgorithm,'with a score of',algorithms[bestalgorithm])
if bestalgorithm == 'Tree':
    print('Best Params is :',tree_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best Params is :',knn_cv.best_params_)
if bestalgorithm == 'SVM':
    print('Best Params is :',svm_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best Params is :',logreg_cv.best_params_)
algodict
```

```
Best Algorithm is Tree with a score of 0.8857142857142858
Best Params is : {'criterion': 'gini', 'max_depth': 16, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}
```

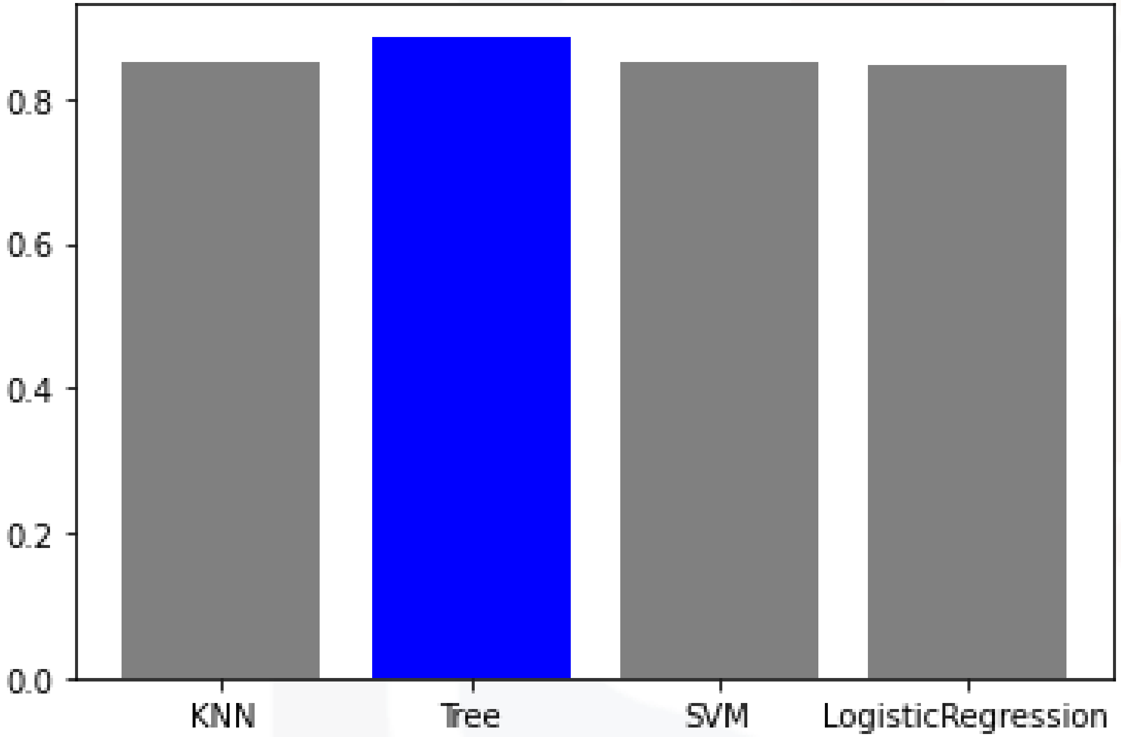
Out[30]:

	Accuracy Score
KNN	0.848214
Tree	0.885714
SVM	0.848214
LogisticRegression	0.846429

Despite having similar values, the Decision Tree Algorithm was the most accurate algorithm with an accuracy of 88.6%, with the specific hyperparameters as shown above.

GitHub File

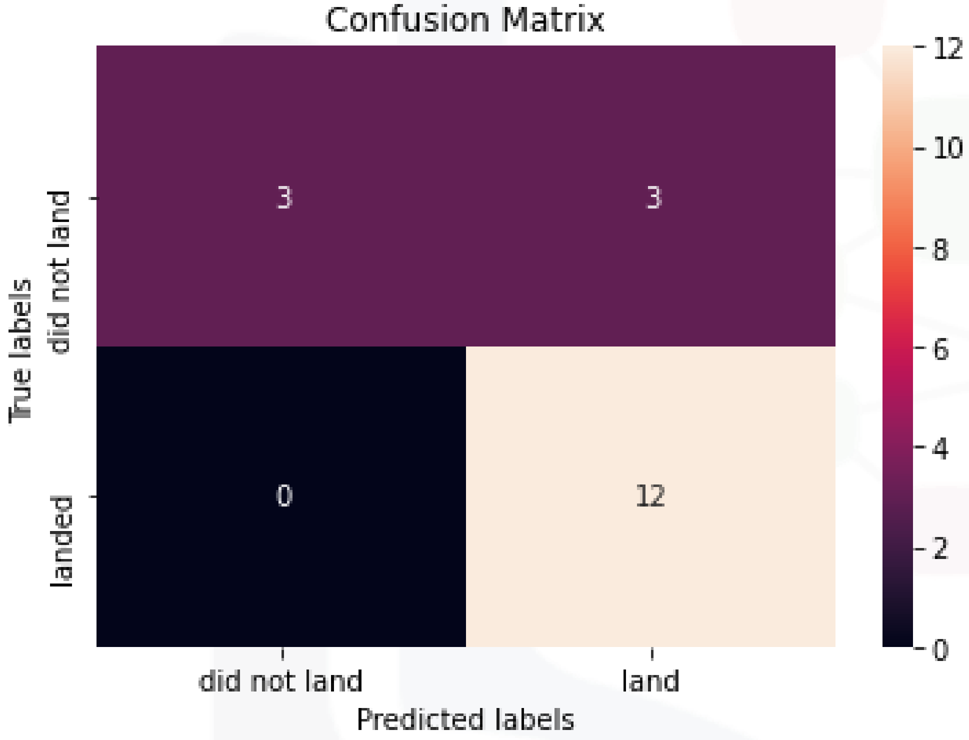
RESULTS



- All models had virtually the same accuracy on the test set at 84.8% accuracy. 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.

GitHub File

RESULTS



- 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-predicted successful landings.

GitHub File

Conclusion and Appendix

All Task Results with Analysis

CONCLUSION

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- We can see that KSC LC-39A had the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate
- Once in Orbit, there is an extremely low chance of mission failure

CONCLUSION

- A higher payload often means that the flight orbital distance will be higher in the atmosphere
- A general trend is observable that higher up in the Earth’s orbit, the higher the chances of success rate, however, this could be due to fewer missions
- More heavy-payload spaceships are launched from CCAFS which also has the highest success rate

APPENDIX

GitHub Repository URL:

https://github.com/AdeetyaU/IBM-Data-Science-Capstone/

Thank you 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

Course Link:

https://www.coursera.org/professional-certificates/ibm-data-science

THANK YOU!

A Presentation Made By Adeetya Upadhyay

IBM Developer

SKILLS NETWORK