IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nur Farhana Binti Ahmad
06/10/2021

Outline

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

Executive Summary

- Summary of methodologies

  1. Data collection methodology:
  2. The data was collected with an API and Web Scraping
  3. Perform data wrangling
  4. The data was processed by performing data analysis on the attributes
  5. Perform exploratory data analysis (EDA) using visualization and SQL
  6. Perform interactive visual analytics using Folium and Plotly Dash
  7. Perform predictive analysis using classification models
  8. Building, tuning, evaluating classification models

Introduction

- Project background and context
  - The project main objective is to determine the price of each launch. This can be obtained by gathering information about Space X and creating dashboards based on the Launch Records. Besides, determining if SpaceX will reuse the first stage. Instead of using rocket science to determine if the first stage will land successfully, a machine learning model is trained and public information to predict if SpaceX will reuse the first stage will be used. The main attributes of interest are Flight Number, Date, Booster version, Payload mass Orbit, and Launch Site.

- Problems you want to find answers

  1. What is the correlation between the Landing of the first stage and the cost of a launch?
  2. Which site has the largest successful launches?
  3. Which site has the highest launch success rate?
  4. Which payload range(s) has the highest launch success rate?
  5. Which payload range(s) has the lowest launch success rate?
  6. Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

Section 1

Methodology

Methodology

Executive Summary

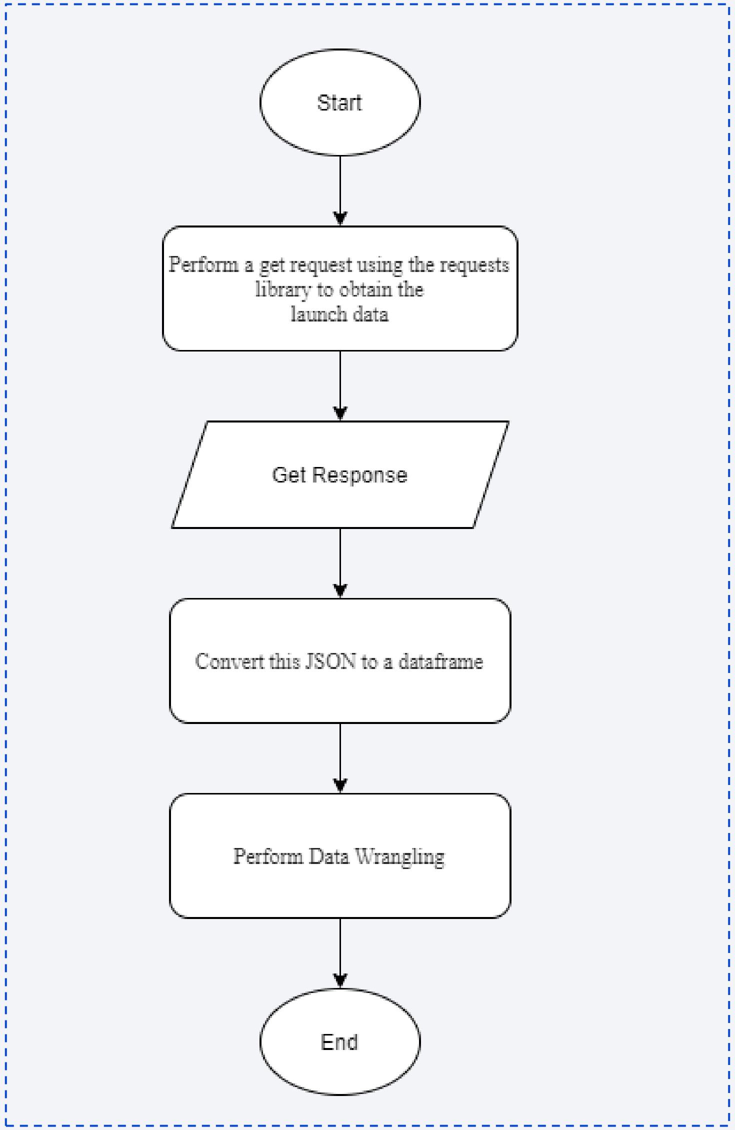
- Data collection methodology:
    - The data was collected with an API and Web Scraping
- Perform data wrangling
    - The data was processed by performing data analysis on the attributes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Building, tuning, evaluating classification models

Data Collection

- The SpaceX launch data that is collected from an API, specifically the SpaceX REST API. This API will give data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome. The goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.

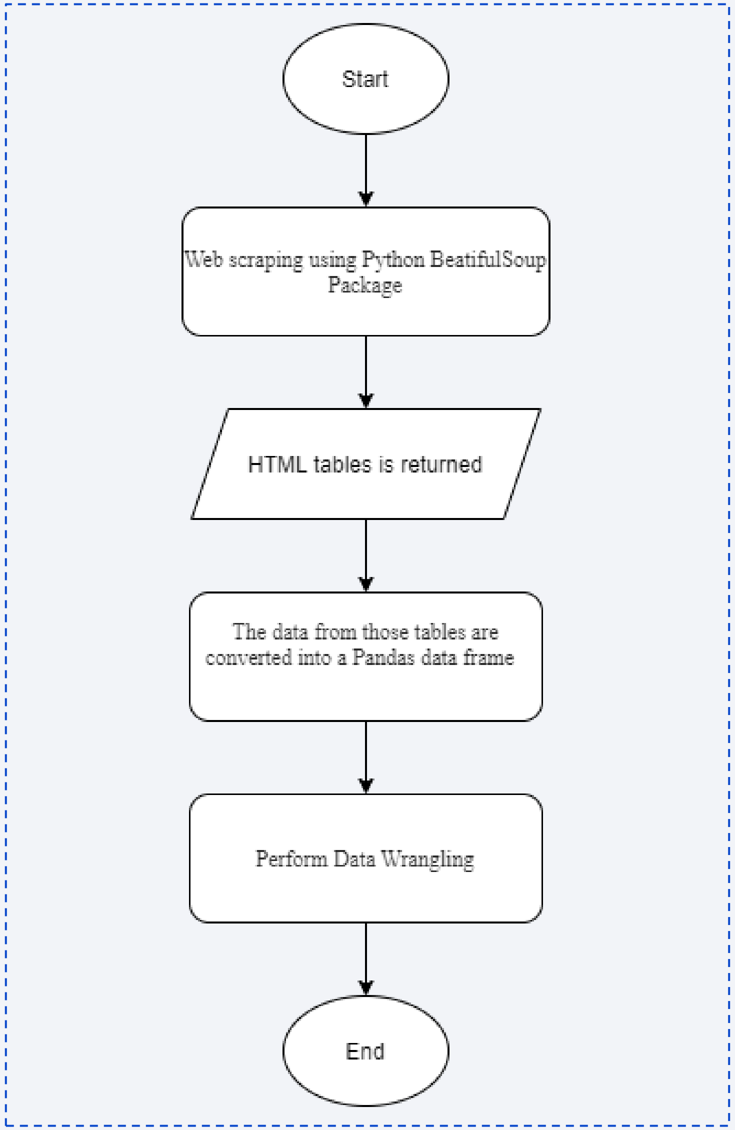
Data Collection – SpaceX API

- Data collection process:
  1. Perform a get request using the requests library to obtain the launch data
  2. Response is in the form of a JSON
  3. Convert this JSON to a dataframe
  4. Transform this raw data into a clean dataset
- Link to Github notebook (https://github.com/nananurfarhana42/IBM-Applied-Data-Science-Capstone/blob/master/Collecting%20the%20data.ipynb)



Data Collection – Scraping

- Data collection process:
    1. Python BeautifulSoup package is used to web scrape HTML tables
    2. The data from those tables are converted into a Pandas data frame for further visualization and analysis
    3. The raw dataset is transformed and cleaned
- Link to Github notebook
  https://github.com/nananurfarhana42/IBM-Applied-Data-Science-Capstone/blob/master/Data%20Collection%20with%20Web%20Scarping.ipynb



EDA with Data Visualization

- **Charts Plotted**

1. Relationship between Flight Number and Launch Site (Scatter Plot)
2. Relationship between Payload and Launch Site (Scatter Plot)
3. Relationship between success rate of each orbit type (Bar Chart)
4. Relationship between FlightNumber and Orbit type (Scatter Plot)
5. Relationship between Payload and Orbit type (Scatter Plot)
6. Launch success yearly trend (Area Chart)

- **Link to Github notebook** (https://github.com/nananurfarhana42/IBM-Applied-Data-Science-Capstone/blob/master/EDA%20with%20Visualization%20lab.ipynb)

EDA with SQL

- SQL queries performed

    1. Display the names of the unique launch sites in the space mission
    2. Display 5 records where launch sites begin with the string 'CCA'
    3. Display the total payload mass carried by boosters launched by NASA (CRS)
    4. Display average payload mass carried by booster version F9 v1.1
    5. List the date when the first successful landing outcome in ground pad was acheived.
    6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    7. List the total number of successful and failure mission outcomes
    8. List the names of the booster_versions which have carried the maximum payload mass
    9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
    10. 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 of the completed EDA with SQL notebook, as an external reference and peer-review purpose (https://github.com/nananurfarhana42/IBM-Applied-Data-Science-Capstone/blob/master/EDA%20with%20SQL%20Lab.ipynb)

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

    1. Add each site's location on a map using site's latitude and longitude coordinates
    2. Create and add folium.Circle and folium.Marker for each launch site on the site map
    3. Mark the success/failed launches for each site on the map
    4. Calculate the distances between a launch site to its proximities

- GitHub URL of the completed interactive map with Folium map, as an external reference and peer-review purpose (https://github.com/nananurfarhana42/IBM-Applied-Data-Science-Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb)

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard
    1. Add a Launch Site Drop-down Input Component
    2. Add a callback function to render success-pie-chart based on selected site dropdown
    3. Add a Range Slider to Select Payload
    4. Add a callback function to render the success-payload-scatter-chart scatter plot
- GitHub URL of the completed Plotly Dash lab, as an external reference and peer-review purpose (https://github.com/nananurfarhana42/IBM-Applied-Data-Science-Capstone/blob/master/spacex_dash_app.py)

Predictive Analysis (Classification)

- Model development process
  1. Perform exploratory Data Analysis and determine Training Labels
     - Create a column for the class
     - Standardize the data
     - Split into training data and test data
  2. Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
     - Find the method performs best using test data
- GitHub URL of the completed predictive analysis lab, as an external reference and peer-review purpose (https://github.com/nananurfarhana42/IBM-Applied-Data-Science-Capstone/blob/master/Machine%20Learning%20Prediction.ipynb)

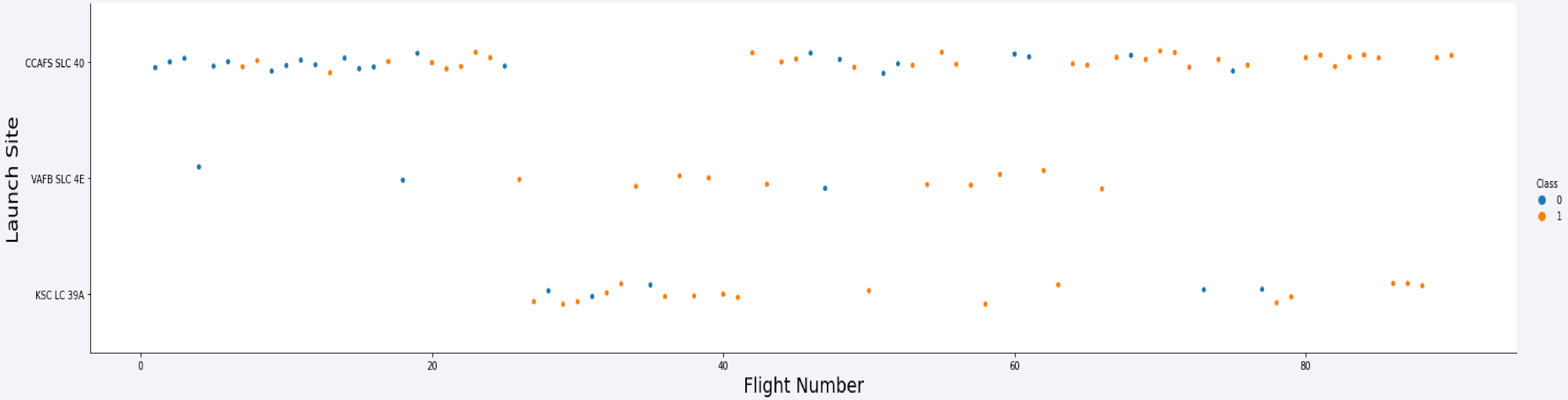
Results

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

Section 2

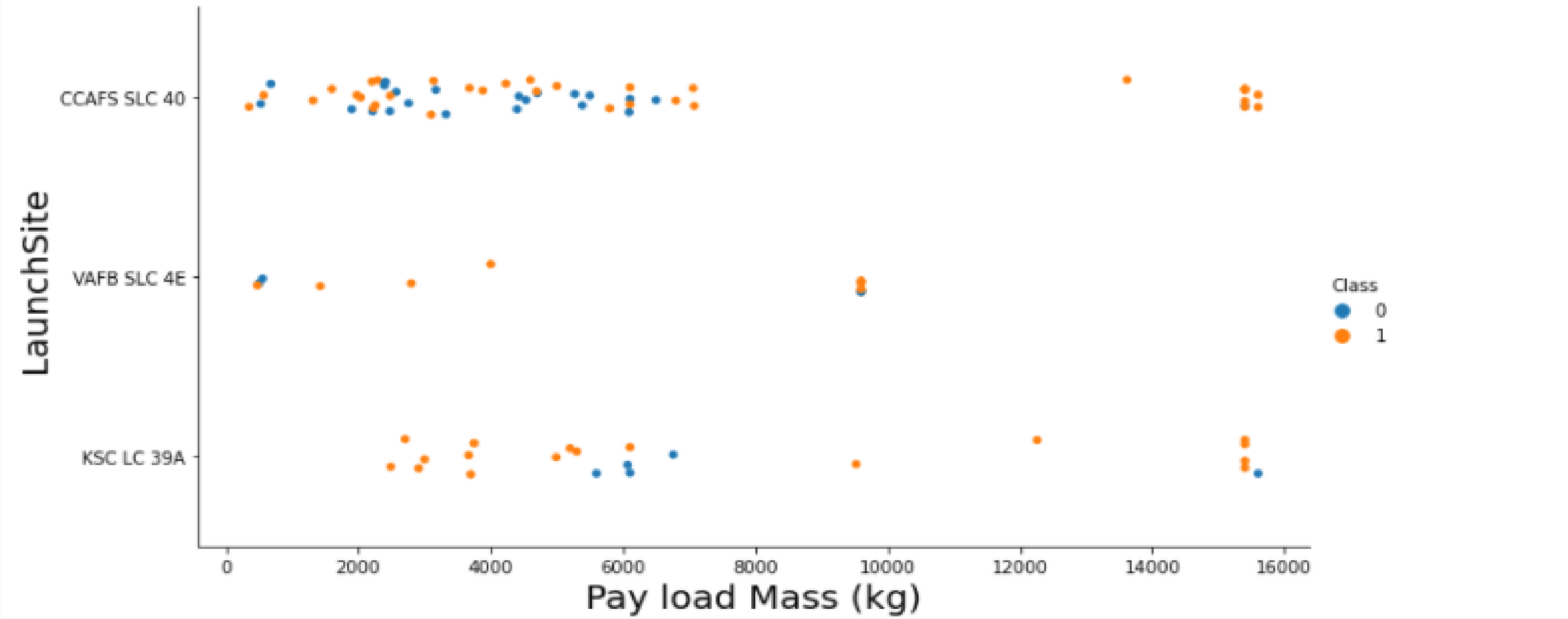
Insights drawn from EDA

Flight Number vs. Launch Site



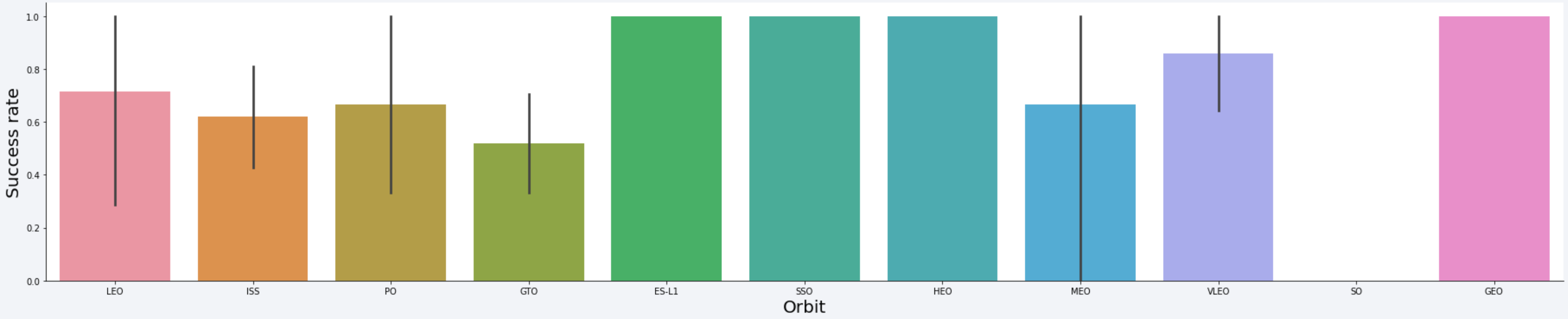
- Each Launch Site has different Flight Number with different success rates.

Payload vs. Launch Site



- CCAFS SLC Launch Site has the highest successful landing
- Increasing the Pay load Mass (kg) will not affect the success or failed landing.

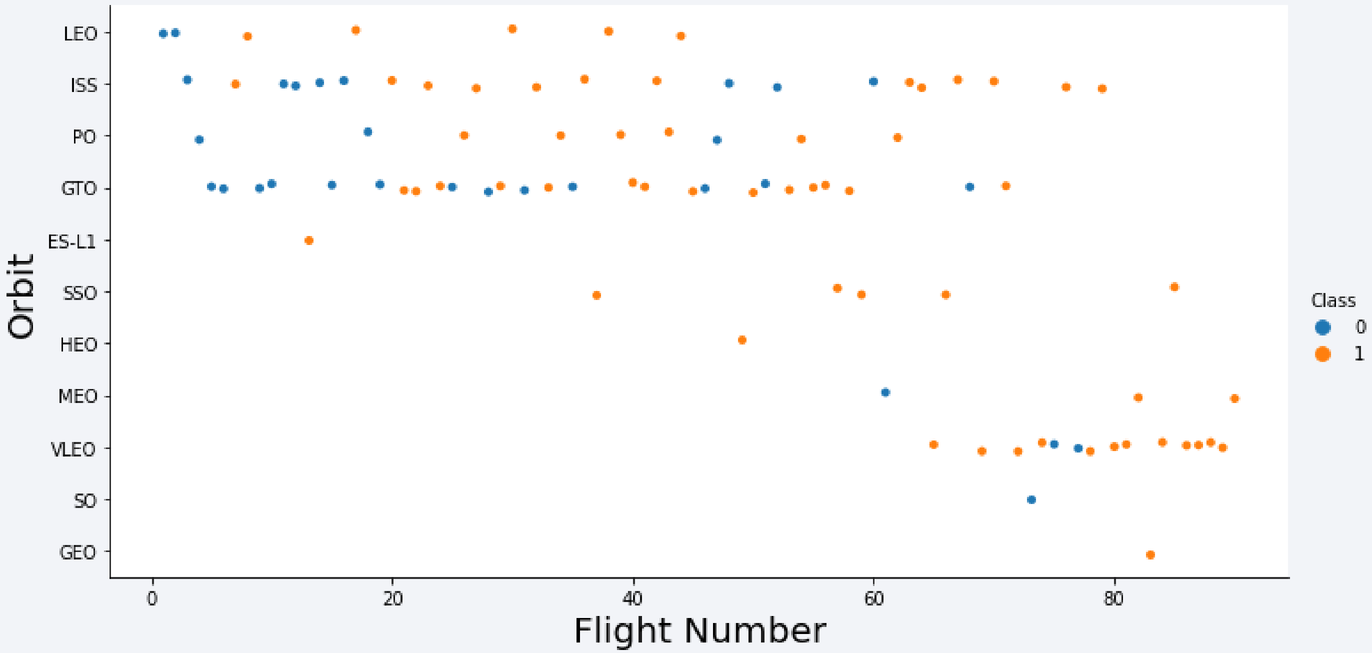
Success Rate vs. Orbit Type



- ES-L1, SSO, HEO, and GEO orbit have the highest success rate with similar values.

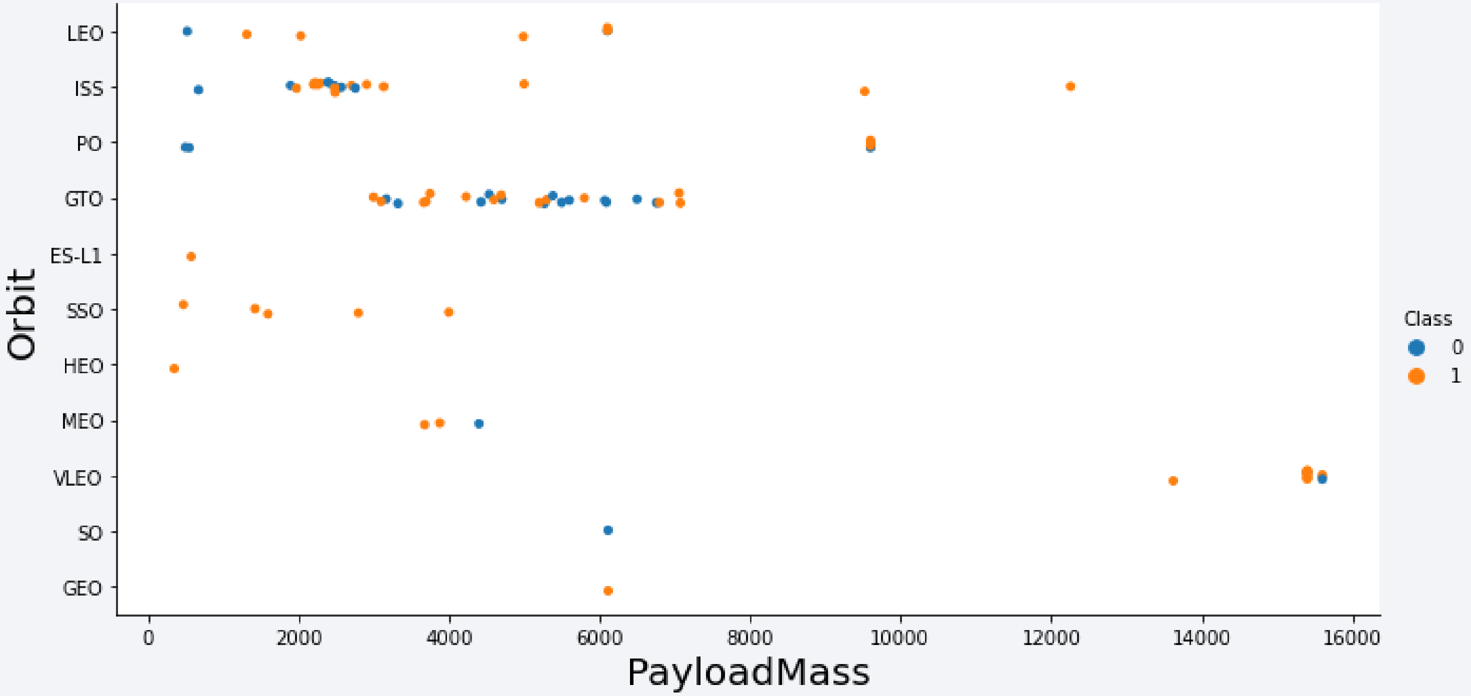
Flight Number vs. Orbit Type

- Different flight number with different orbit type has different success rate of landing.



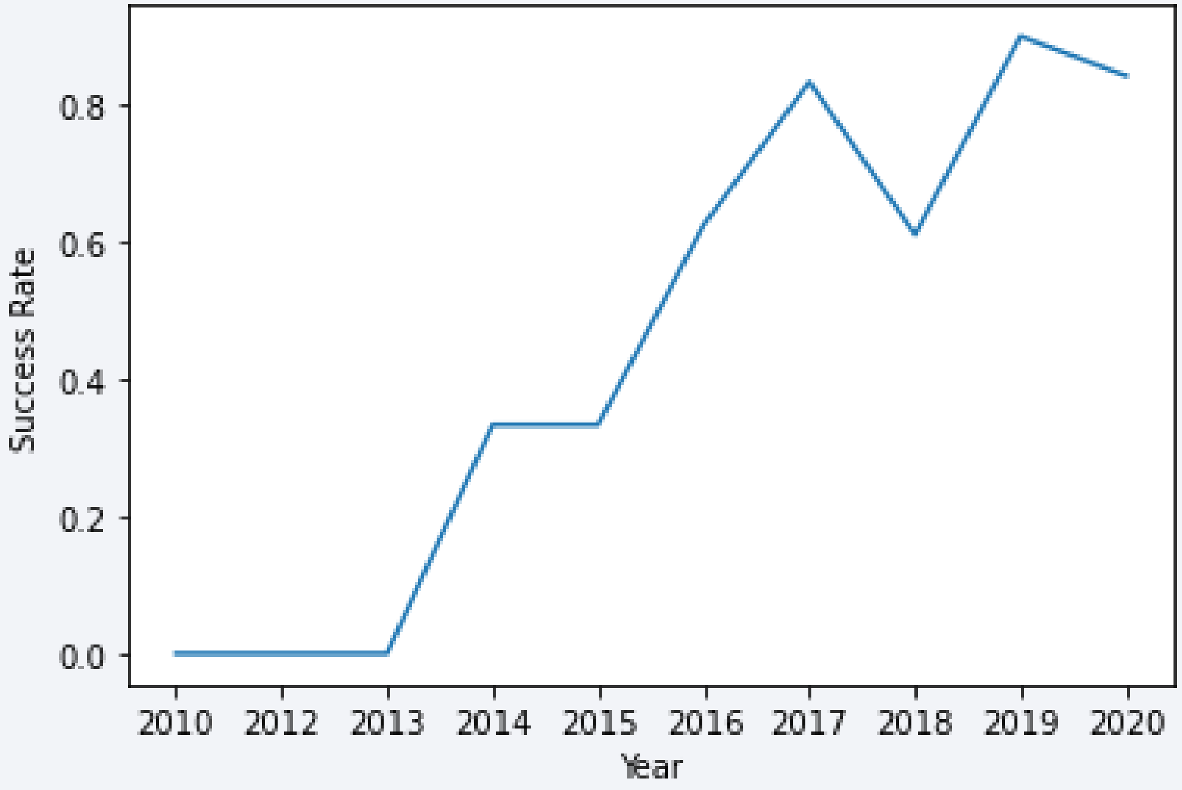
Payload vs. Orbit Type

- The successful landing have Payload Mass (kg) between 2000 to 8000



Launch Success Yearly Trend

- The highest success rate is in 2019
- The lowest success rate is in 2010-2013 with consistent values



All Launch Site Names

- The unique launch sites in the space mission are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

```
%sql SELECT distinct(LAUNCH_SITE) from SPACEXTBL
```

```
 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

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

 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/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

- Calculate the total payload carried by boosters from NASA

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

1
2928

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is in 2015-12-22

```
%sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS_
_KG_ < 6000
```

```
 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/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

- The total number of successful and failure mission outcomes are 100

```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'
```

```
 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

1
100

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are shown below

```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015

```
%sql select monthname(date), landing__outcome, booster_version, launch_site from SPACEXTBL where year(date) = 2015
```

```
 * ibm_db_sa://lyz40877:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

1	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
February	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
March	No attempt	F9 v1.1 B1014	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
April	No attempt	F9 v1.1 B1016	CCAFS LC-40
June	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
December	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

- 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 are shown as in the diagram

```
%sql select * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

```
 * ibm_db_sa://rzq20066:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/BLUDB
Done.
```

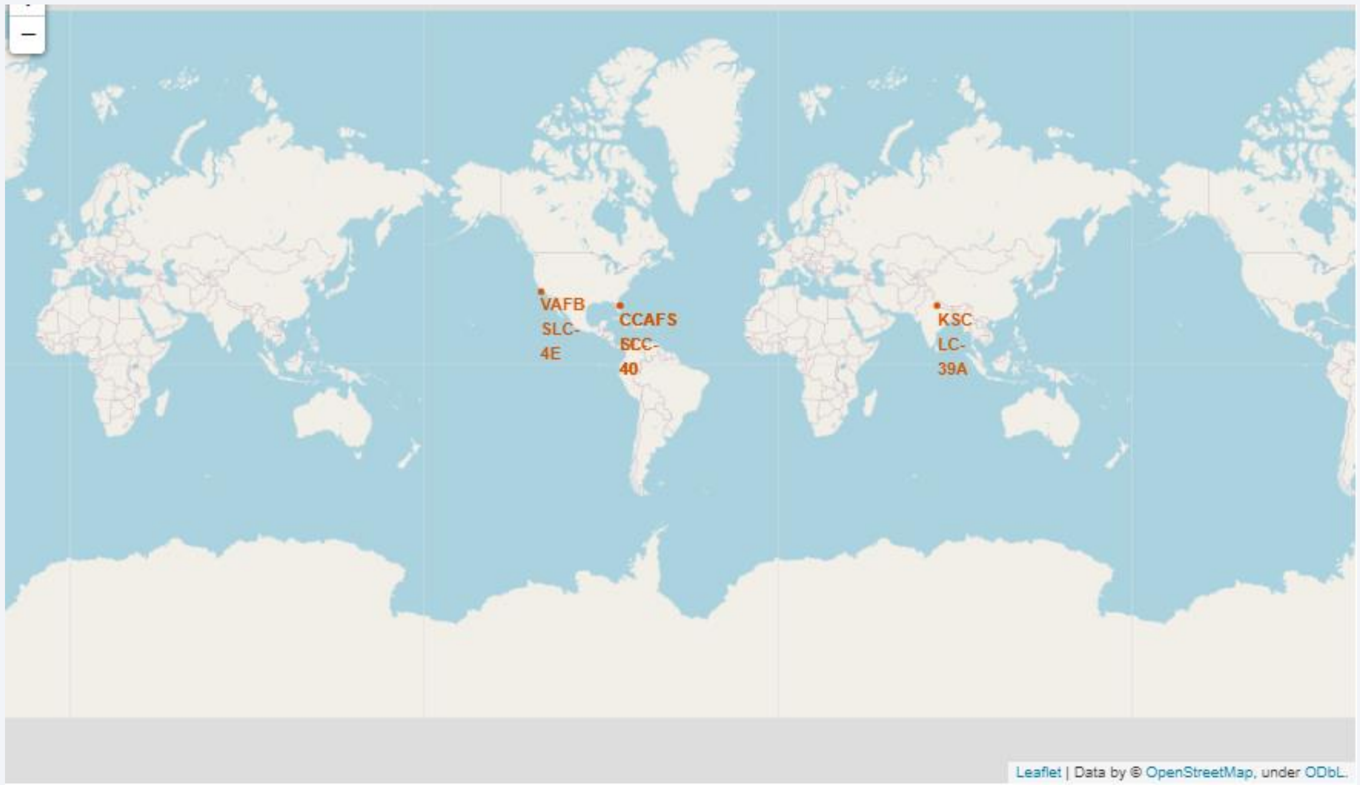
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Section 4

Launch Sites Proximities Analysis

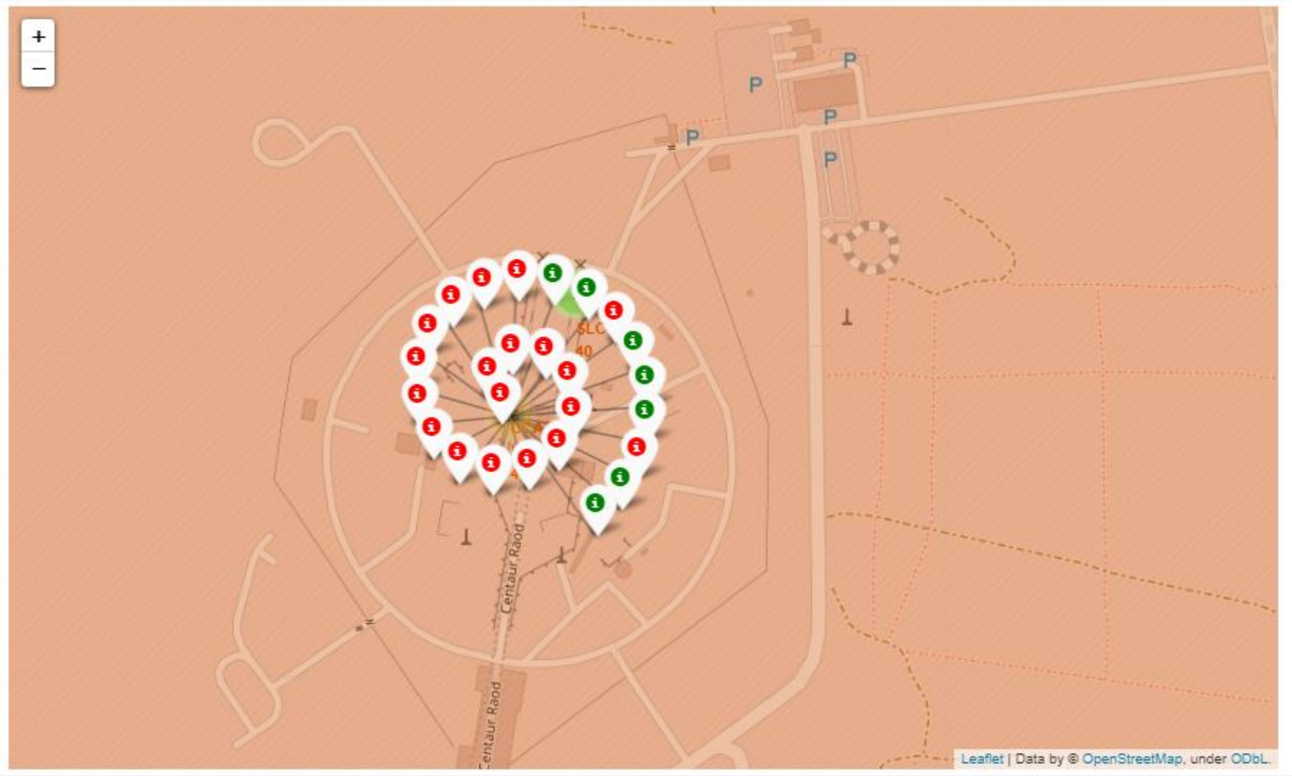
Mark all launch sites on a map

- All launch sites' location markers on the global map include CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.
- The location of each launch sites are given below:
  - CCAFS LC-40: Meritt Island National Wildlife Refuge
  - CCAFS SLC-40: Meritt Island National Wildlife Refuge
  - KSC LC-39A: Dudhwa National Park
  - VAFB SLC-4E Vandenberg State Marine Reserve



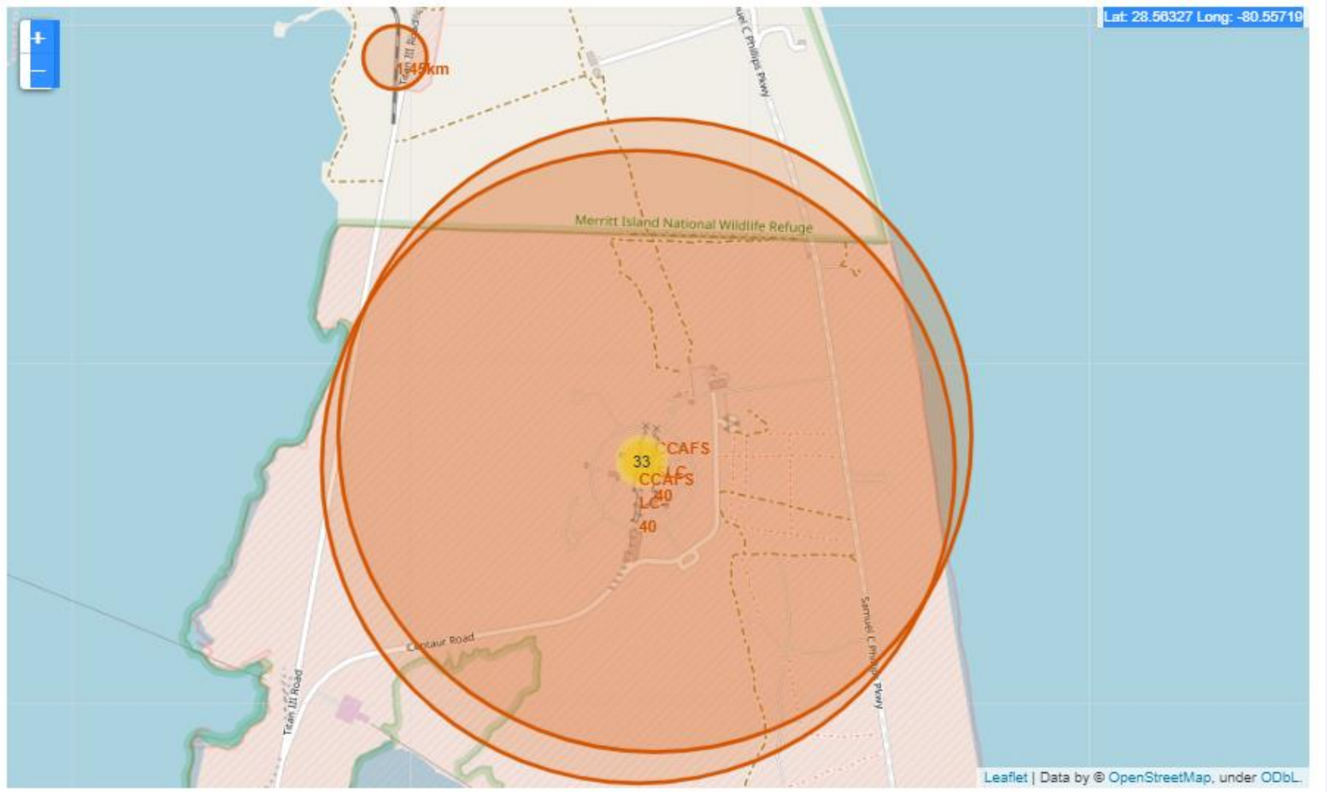
Mark the success/failed launches for each site on the map

- The color-labeled launch outcomes on the map shows whether the space has successful landing with red color indicates successful landing and green color indicates failed landing.
- From the color-labeled markers in marker clusters, it enables us to easily identify which launch sites have relatively high success rates.



Calculate the distances between a launch site to its proximities

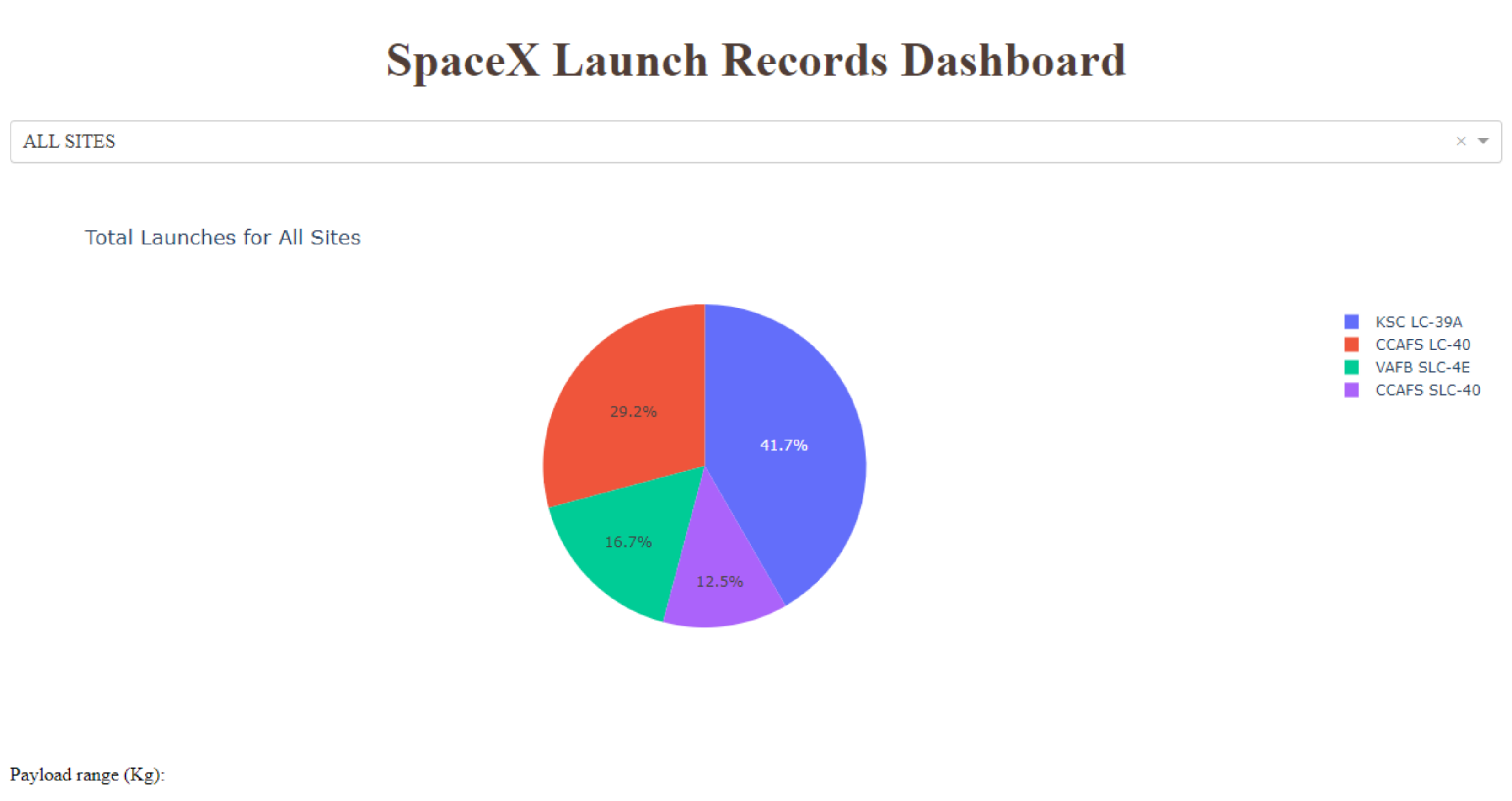
- The generated folium map show the selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- The plot distance lines to the proximities are given below:
  - CCAFS LC-40: 1.45KM
  - CCAFS SLC-40: 1.45KM



Section 5

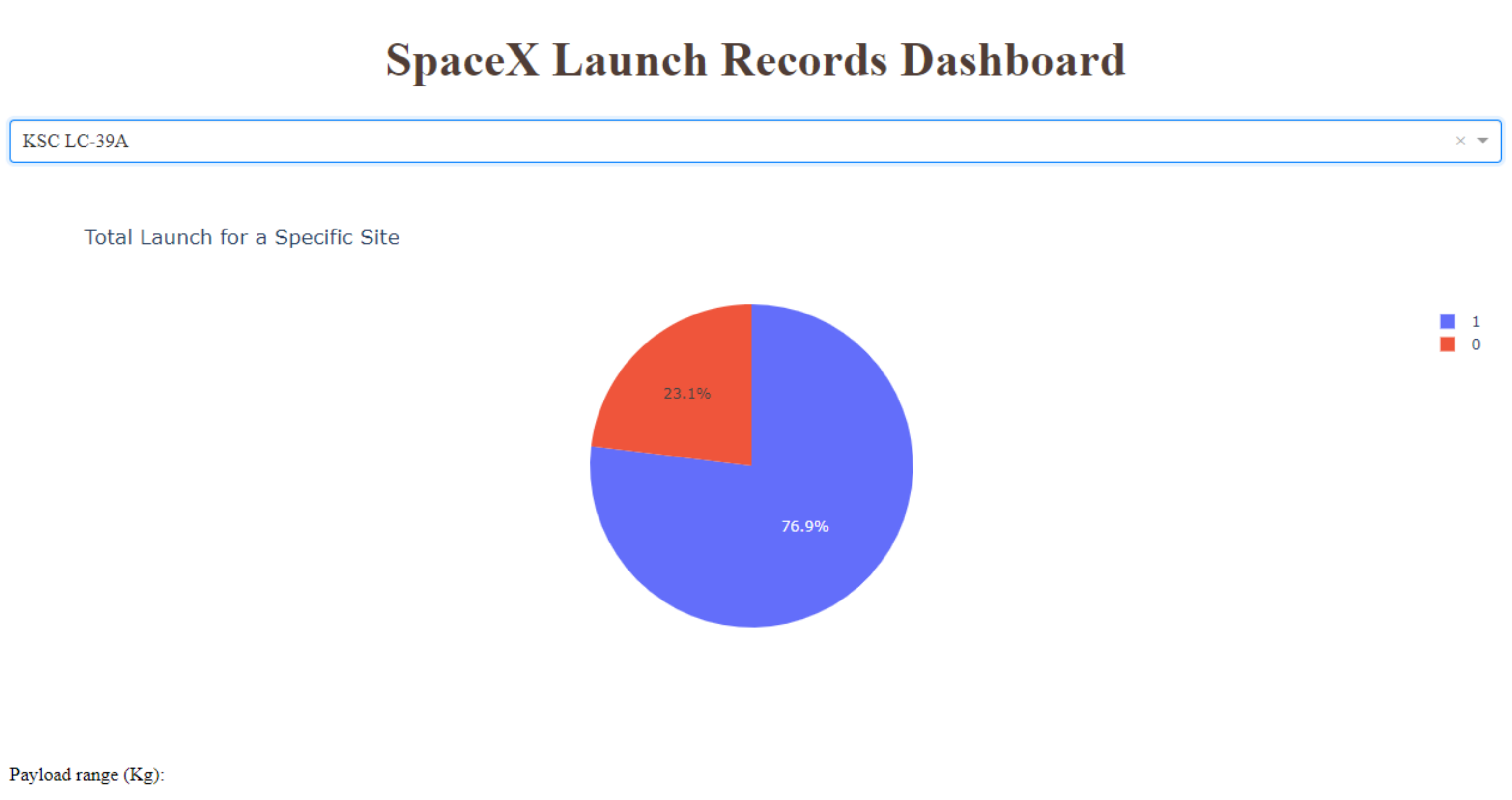
Build a Dashboard with Plotly Dash

Total Launches for All Sites



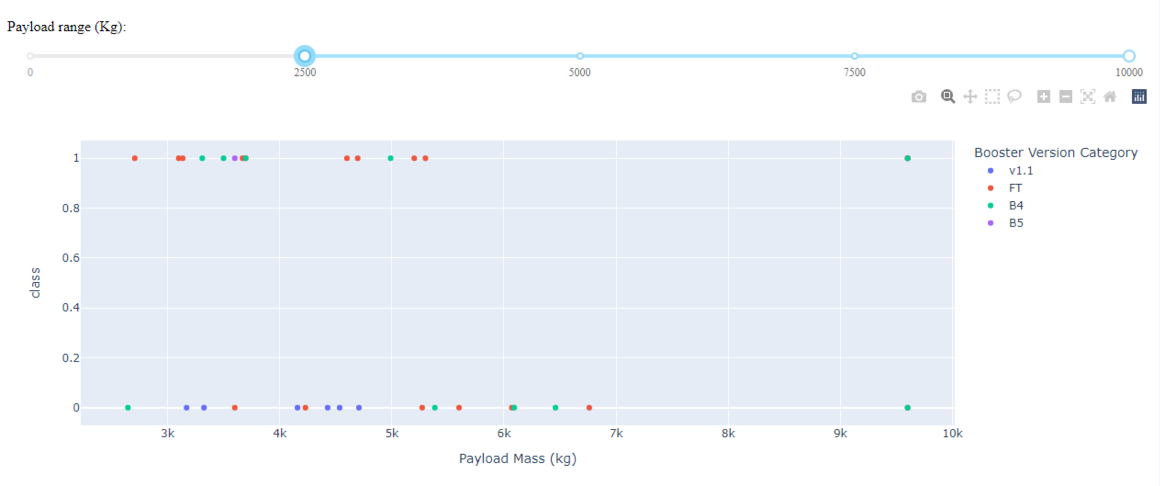
- The KSC LC-39A launch site has the highest launch success rate with 41.7% success.
- The CCAFS SLC-40 launch site has the lowest launch success rate with 12.5% success.

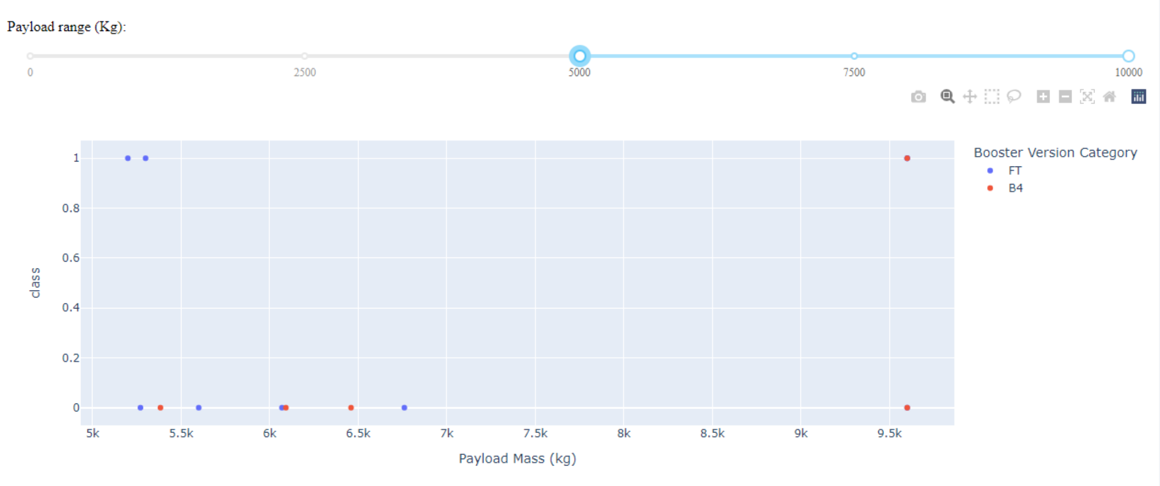
Launch Site with Highest Launch Success Ratio



- KSC LC-39A launch site with the highest launch success ratio has 76.9% success ration, and 23.1% failed rate.

Payload vs. Launch Outcome





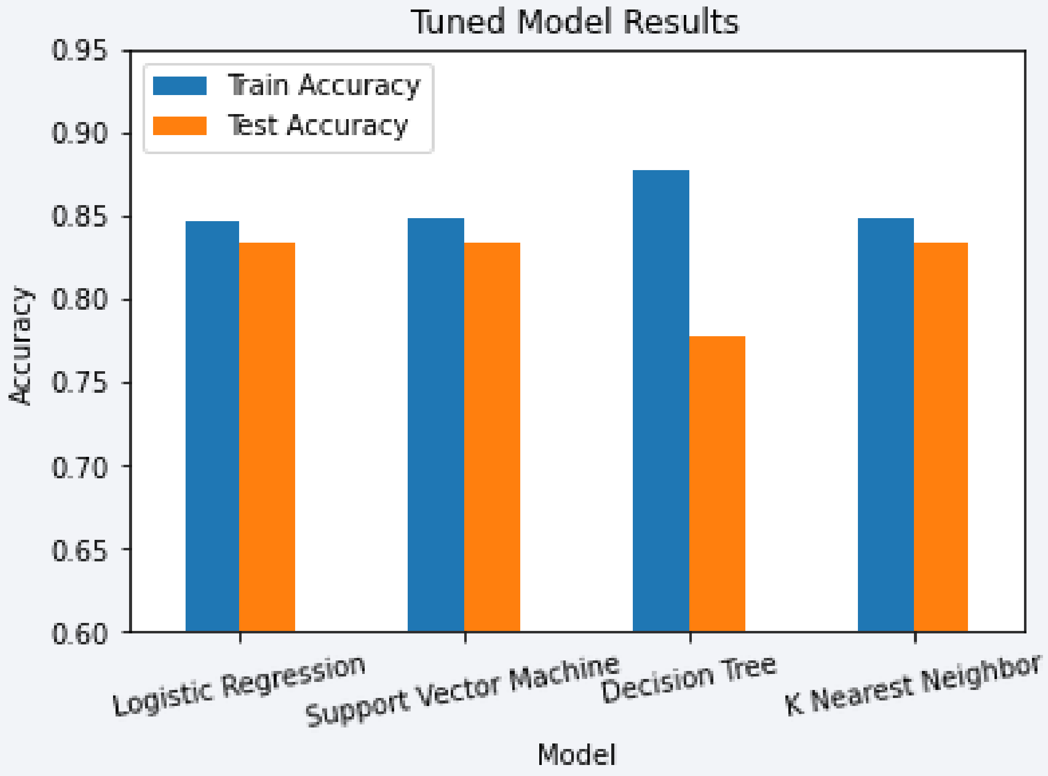
- Payload range of 2000-5000 with booster version of FT and B4 has the largest success rate.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

- From the graph, it can be seen that Logistic Regression, Support Vector Machine, and K Nearest Neighbor have the highest test accuracy.
- Decision tree model has the highest training accuracy.



Confusion Matrix

- Logistic Regression and Support Vector Machine maintain their unbroken lead in prediction accuracy, but Support Vector Machine has the highest training accuracy yet, so this model is preferred.
- However, that the training time is different for the various models, and this could be a factor in a production environment.

```
SUPPORT VECTOR MACHINE
tuned hpyerparameters :(best parameters)  {'C': 31.622776601683793, 'gamma': 31.622776601683793, 'kernel': 'rbf'}
train accuracy : 0.8819642857142856
test accuracy: 0.8333333333333334
Wall time: 2min 30s
```


```
LOGISTIC REGRESSION
tuned hpyerparameters :(best parameters)  {'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}
train accuracy : 0.832142857142857
test accuracy: 0.8333333333333334
Wall time: 1.27 s
```

Conclusions

- Support Vector Machine model is the best performing model
- The KSC LC-39A launch site has the largest successful launches
- The CCAFS SLC-40 launch site has the lowest successful launches
- Payload range of 2000-5000 with booster version of FT and B4 has the largest success rate

Appendix

- Github Repository URL: https://github.com/nananurfarhana42/IBM-Applied-Data-Science-Capstone

Thank you!