DATA SCIENCE FINAL PROJECT

SPACEX

YU LIN YEH /2024.03.06

IBM

OUTLINE

1. Executive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusion

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis result

Introduction

Project background and context

SpaceX is leading the pack in the era of commercial space travel by making it cheaper to explore space. They advertise **Falcon 9 rocket** launches on their website for $62 million each, which is a lot less than the $165 million price tag for launches by other companies.
A big reason for this cost difference is that SpaceX can reuse the first stage of their rockets. So, if we can figure out whether they'll be able to land and reuse that first stage, we can estimate how much a launch will cost. Using public info and fancy computer models, we're going to try and predict if SpaceX will indeed reuse the first stage.

Introduction

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case

Methodology

Data collection methodology

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

- Building, tuning and evaluation to find the best results

Methodology-Data Collection

Rocket launch data: SpaceX API

https://api.spacexdata.com/v4

1. Decoding the response content: .json()
   Turning into dataframe: .json_normalize()
2. Applying functions to get needed information about the launches
3. Constructing data into dictionary and creating dataframe
4. Filtering the dataframe: only Falcon 9 launches
5. Replacing missing values of Payload Mass column
6. Calculating .mean() for Payload Mass column
7. Exporting data to CSV

Falcon 9 launch data: Wikipedia

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

1. Creating a BeautifulSoup object from the HTML response
2. Extract all column/variable names from the HTML table header
3. Create a data frame by parsing the launch HTML tables
4. Constructing data into dictionary and creating dataframe
5. Exporting data to CSV

Methodology-Data Wrangling

There are several different cases where the booster did not land successfully. We try to convert those outcomes into Training Labels with “1” and “0” to show whether the booster successfully landed or not.

Case	Description	Training Label
True Ocean	The mission outcome successfully landed in a specific region of the ocean.	1
False Ocean	The mission outcome unsuccessfully landed in a specific region of the ocean due to accident.	0
True RTLS	The mission outcome successfully landed on a ground pad.	1
False RTLS	The mission outcome unsuccessfully landed on a ground pad due to accident.	0
True ASDS	The mission outcome successfully landed on a drone ship.	1
False ASDS	The mission outcome unsuccessfully landed on a drone ship due to accident.	0

Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting data to CSV

Methodology-EDA with data visualization

Charts were plotted:

Relationship
Flight Number vs. Payload Mass
Flight Number vs. Launch Site
Payload Mass vs. Launch Site
Orbit Type vs. Success Rate
Flight Number vs. Orbit Type
Payload Mass vs. Orbit Type
Success Rate Yearly Trend

Using different ways to plot:

Type	Description
Scatter plots	Display the relationship between variables. If a relationship exists, it can be used in ML models.
Bar charts	Show comparisons among discrete categories. They illustrate the relationship between specific categories and measured values.
Line charts	Show trends in data over time (time series)

Methodology-EDA with SQL

Performed SQL queries:

01
Displaying

1. Names of the unique launch sites in the space mission
2. 5 records where launch sites begin with the string ‘CCA’
3. The total payload mass carried by boosters launched by NASA (CRS)
4. Average payload mass carried by booster version F9 v1.1

02
Listing

1. Date when the first successful landing outcome in ground pad was achieved
2. Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
3. Total number of successful and failure mission outcomes
4. Names of the booster versions which have carried the maximum payload mass
5. Failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

03
Ranking

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

Methodology-Build an interactive map with Folium

Markers of all Launch Sites

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts

Coloured Markers of the launch outcomes for each Launch Site

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates

Distances between a Launch Site to its proximities

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

Methodology-Build a Dashboard with Plotly Dash

Launch Sites Dropdown List

Using dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site)

Using pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions

Using scatter chart to show the correlation between Payload and Launch Success

Slider of Payload Mass Range

Added a slider to select Payload range.

Methodology-Predictive analysis (Classification)

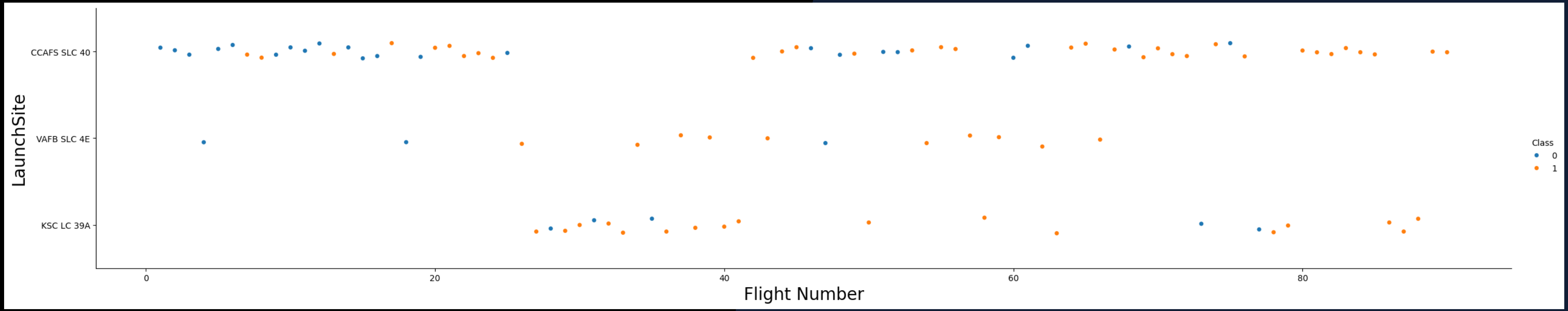
1. Creating a NumPy array from the column “Class” in data
2. Standardizing the data with StandardScaler, then fit and transform it
3. Splitting the data into training and testing sets: **train_test_split( )**
4. Creating a GridSearchCV object with cv = 10 to find the best parameters
5. GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
6. **.score() :** Calculating the accuracy on the test data
7. **Confusion matrix** for all models
8. Finding the method performs best by **Jaccard_score** and **F1_score** metrics

Results

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

Results- EDA with Visualization

Flight Number vs. Launch Site

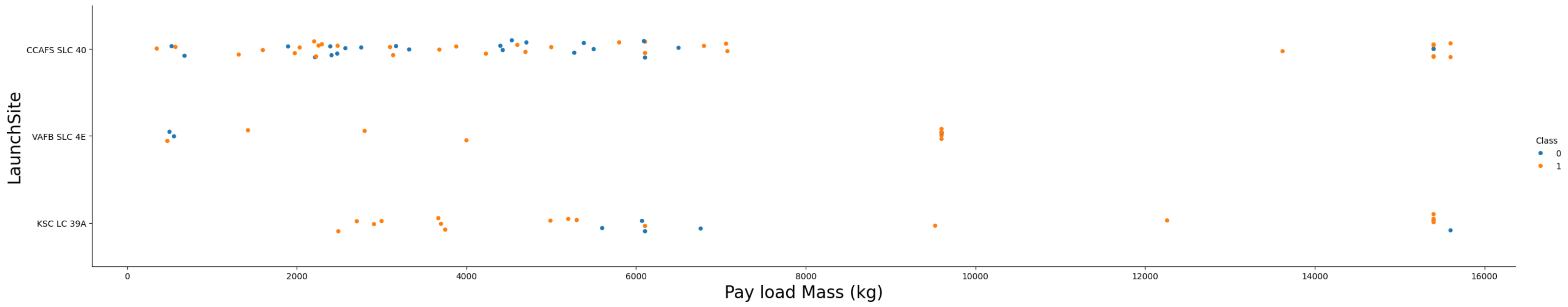


Explanation:

- First 6 flights all failed while the last 13 flights all succeeded.
- Over half of all launches were launched at CCAFS SLC 40 launch site.
- Success rate of KSC LC 39A is a bit higher than VAFB SLC 4E.
- CCAFS SLC 40 has the lowest success rate.
- It can be assumed that each new launch has a higher rate of success.

Results- EDA with Visualization

Payload Mass vs. Launch Site

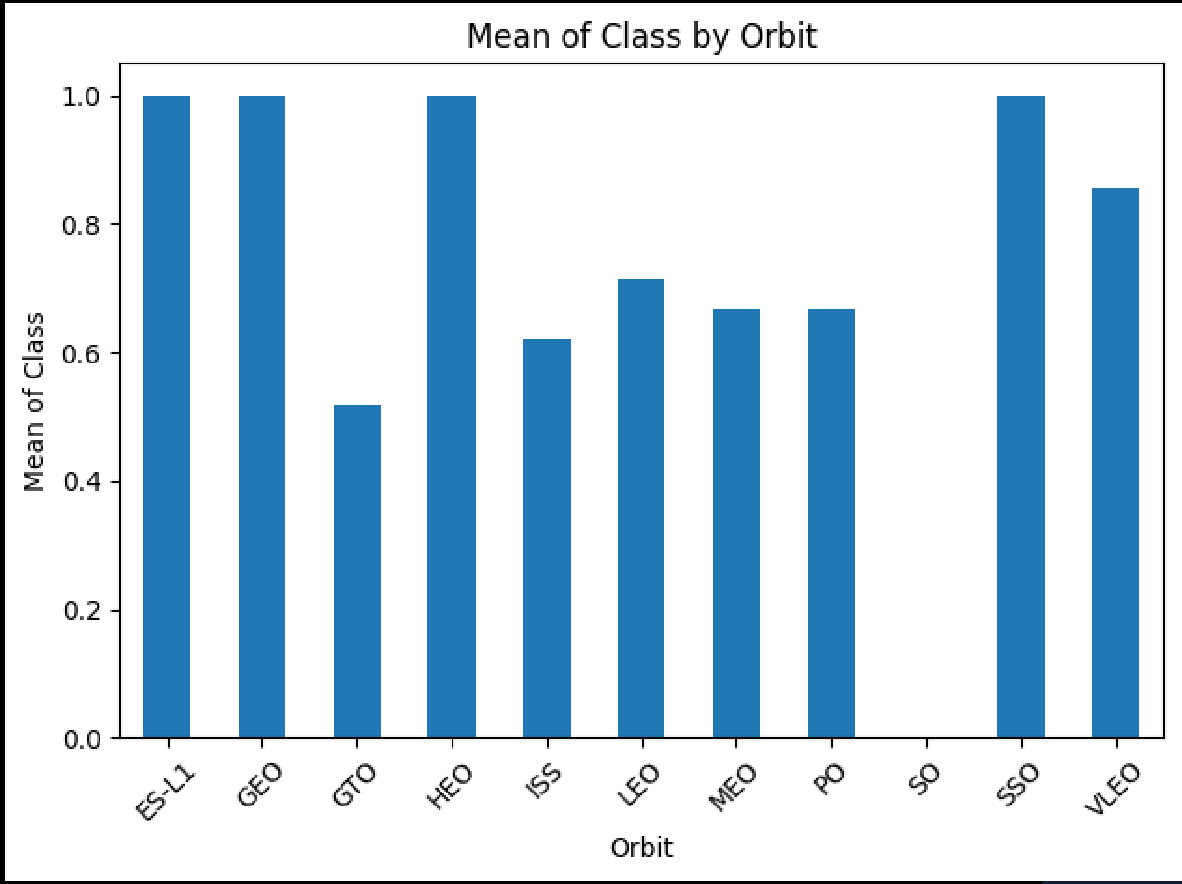


Explanation:

- Every launch site: the higher the payload mass, the higher the success rate.
- Successful launches with payload mass over 8000 kg were much more than successful launches with payload mass under 8000 kg .
- The majority of failures in launches originating from KSC LC 39A were concentrated in the payload mass range of 5000-7000 kg.

Results- EDA with Visualization

Success rate vs. Orbit type

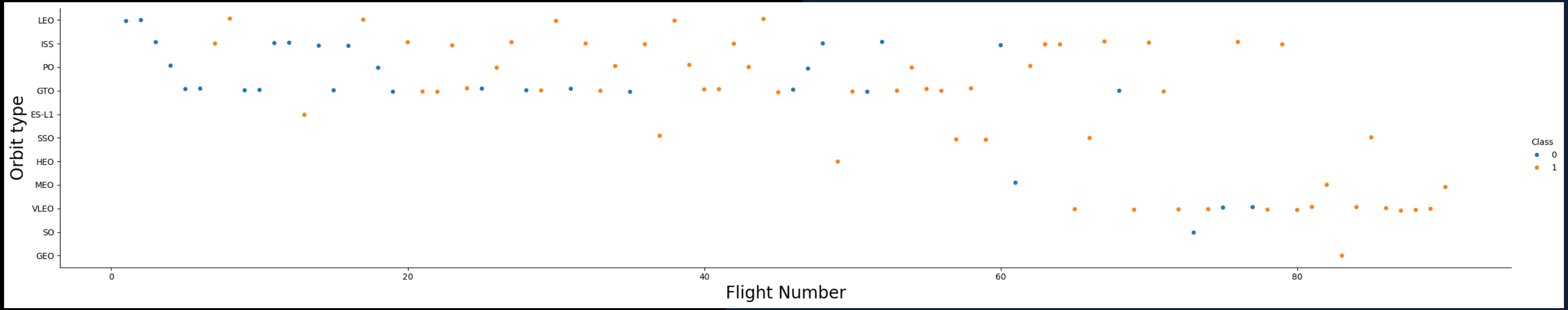


Explanation:

Success rate	Orbit
0%	SO
40-80%	GTO、ISS、LEO、MEO、PO
80-100%	VLEO
100%	ES-L1、GEO、HEO、SSO

Results- EDA with Visualization

Flight Number vs. Orbit type

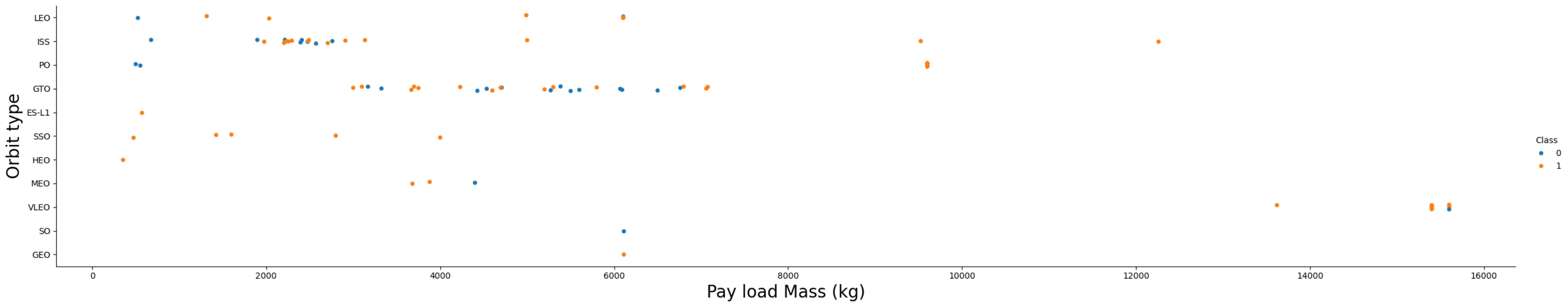


Explanation:

- VLEO and ISS orbit has been more used in recently launches; on the other hand, there seems to be rare flights in GTO and LEO orbit.
- There is something happened in GTO orbit that made launches failed after successful flight.

Results- EDA with Visualization

Payload Mass vs. Orbit type

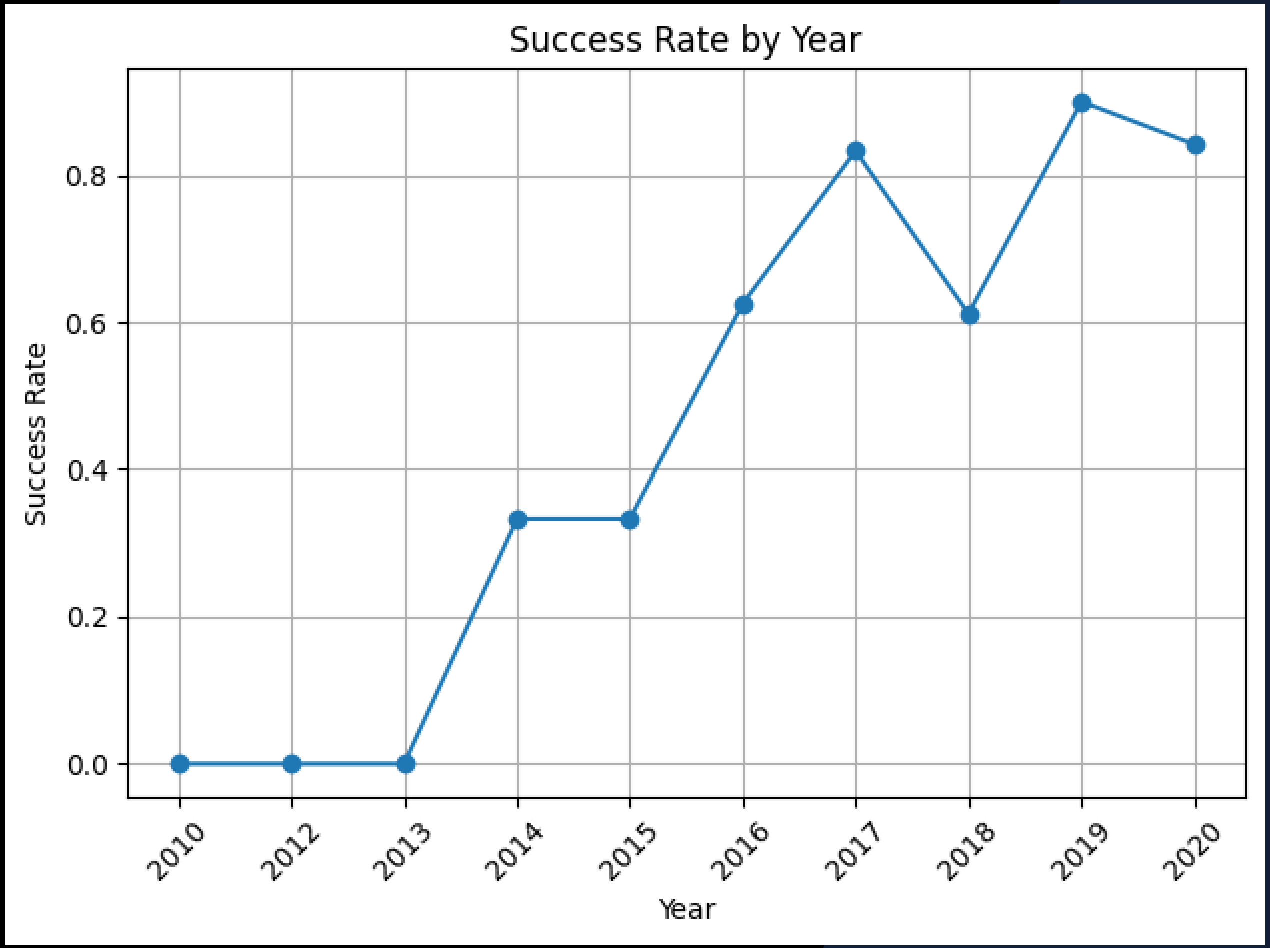


Explanation:

- Launches on SSO orbit were concentrated in the payload mass range of under 4000 kg.
- Launches on LEO orbit were concentrated in the payload mass range of under 7000 kg.
- Launches on GTO orbit were concentrated in the payload mass range of under 8000 kg.
- Heavy payloads have positive influence on GTO and ISS orbits.

Results- EDA with Visualization

Launch success yearly trend



Explanation:

The success rate kept increasing since 2013. Only drop a bit at 2018.

Results- EDA with SQL

```
%sql select distinct launch_site from SPACEXDATASET;
```

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b
Done.

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

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

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases
Done.

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cl
Done.

total_payload_mass
45596

Display the total payload mass carried by boosters launched by NASA (CRS)

Results- EDA with SQL

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%'
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/
Done.
```

average_payload_mass
2534

Display average payload mass carried by booster version F9 v1.1

```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/blu
Done.
```

first_successful_landing
2015-12-22

List the date when the first succesful landing outcome in ground pad was acheived.

```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ betwee
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

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

Results- EDA with SQL

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdo
one.
```

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

List the total number of successful and failure mission outcomes

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET)
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/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

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
    where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3
Done.
```

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

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

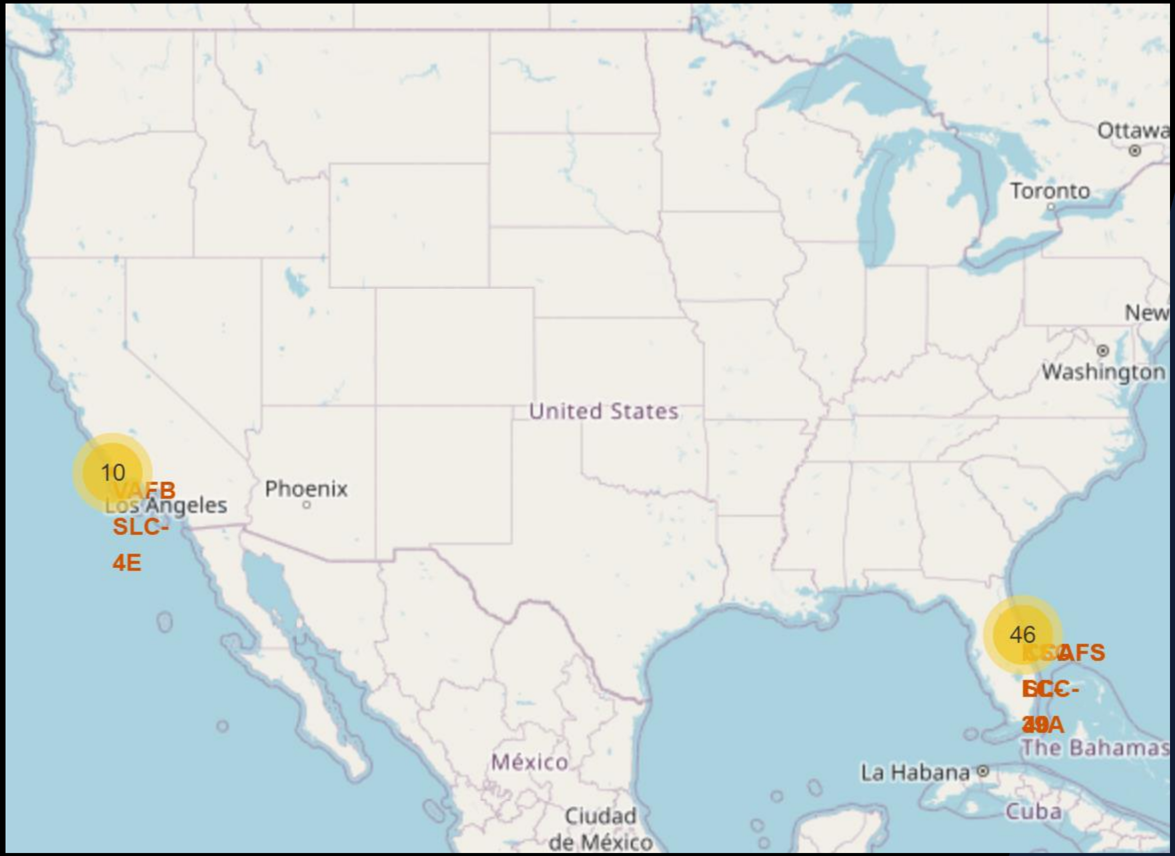
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
    where date between '2010-06-04' and '2017-03-20'
    group by landing__outcome
    order by count_outcomes desc;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1o
Done.
```

landing__outcome	count_outcomes
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

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

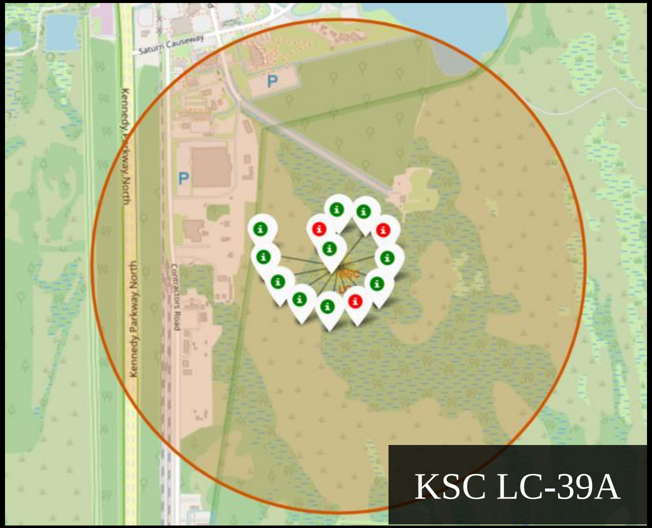
Results- Interactive map with Folium

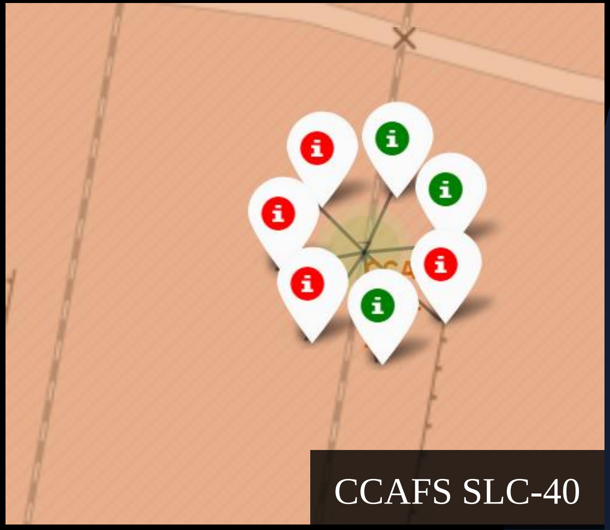


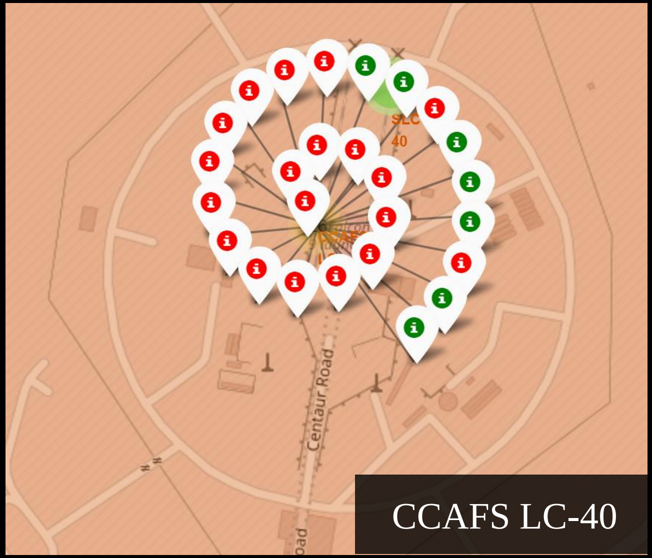
Explanation:

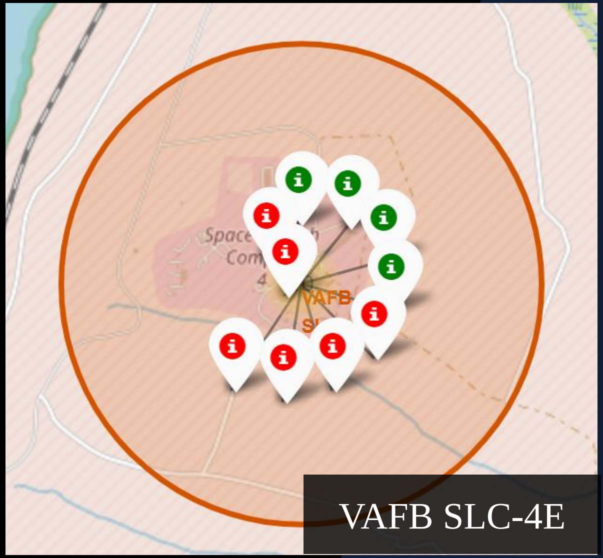
- Most launch sites are located near the equator because the surface speed at the equator is faster than elsewhere. This means that if a rocket is launched from the equator, it will continue to orbit the Earth at the same speed and maintain enough velocity to enter orbit.
- Additionally, all launch sites are situated close to the coast to minimize the risk of rocket debris falling or exploding near populated areas during launch.

Results- Interactive map with Folium









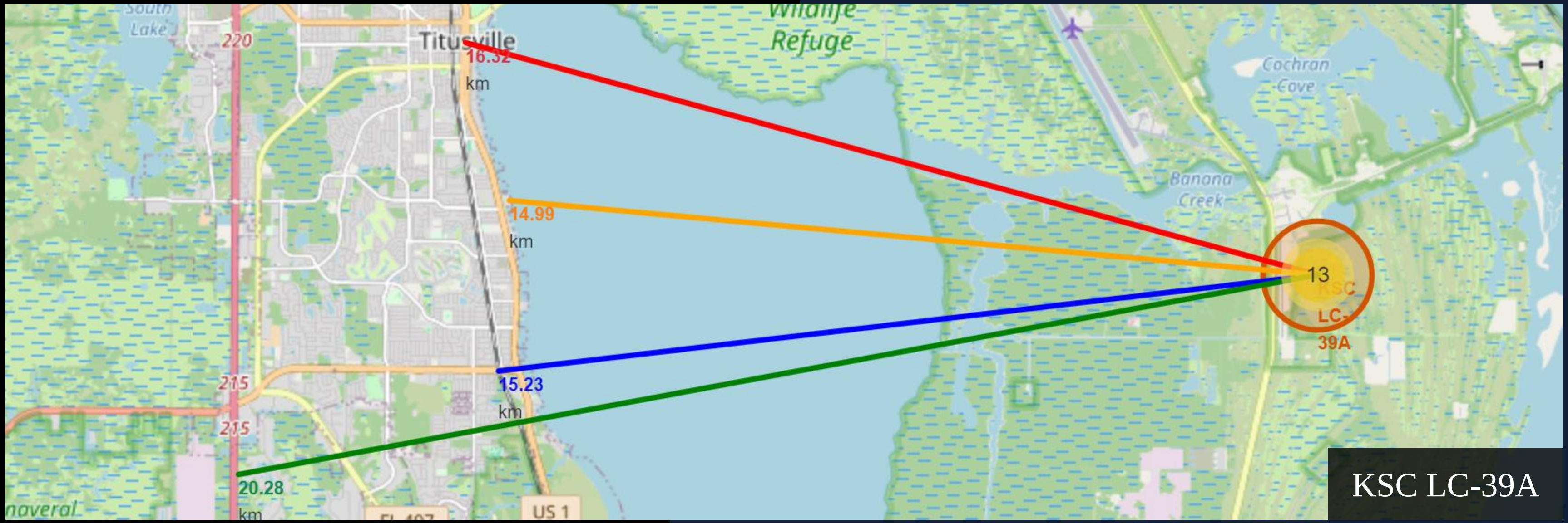
Explanation:

- Using colour-labeled markers to identify launch sites with relatively high success rates.

  Green Marker = Successful Launch

  Red Marker = Failed Launch
- KSC LC-39A has the best success Rate

Results- Interactive map with Folium

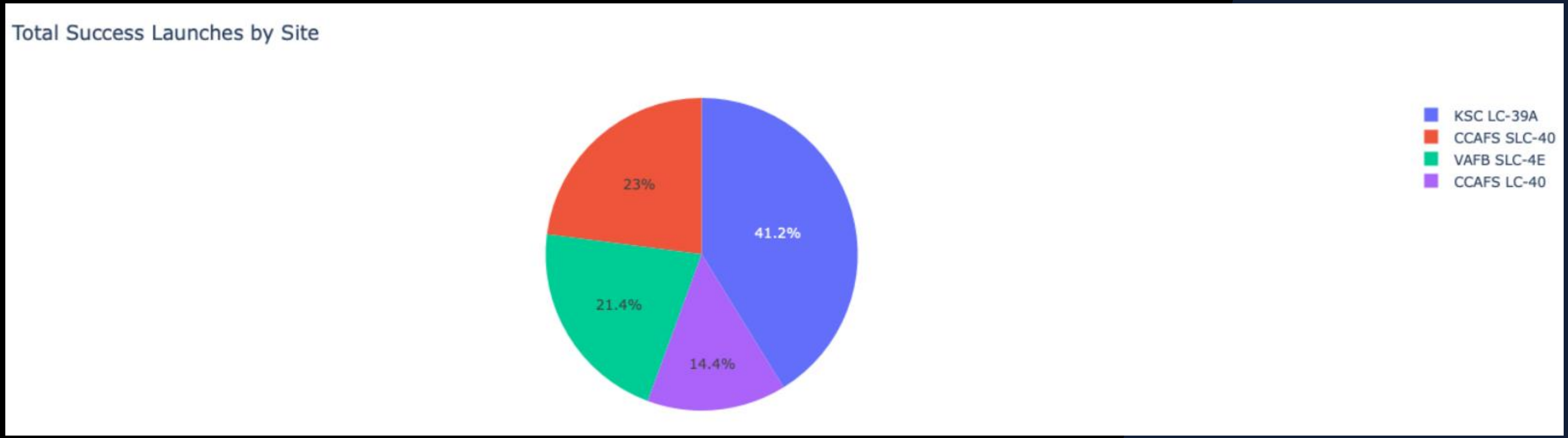


Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly see the reasonable distance to railway (15.23 km), highway (20.28 km) , coastline (14.99 km), and closest city Titusville (16.32 km).
- A malfunctioning rocket, traveling at high velocity, has the capability to traverse distances of 15-20 kilometers within mere seconds. This poses a potential threat to inhabited regions.

Results- Build a Dashboard with Plotly Dash

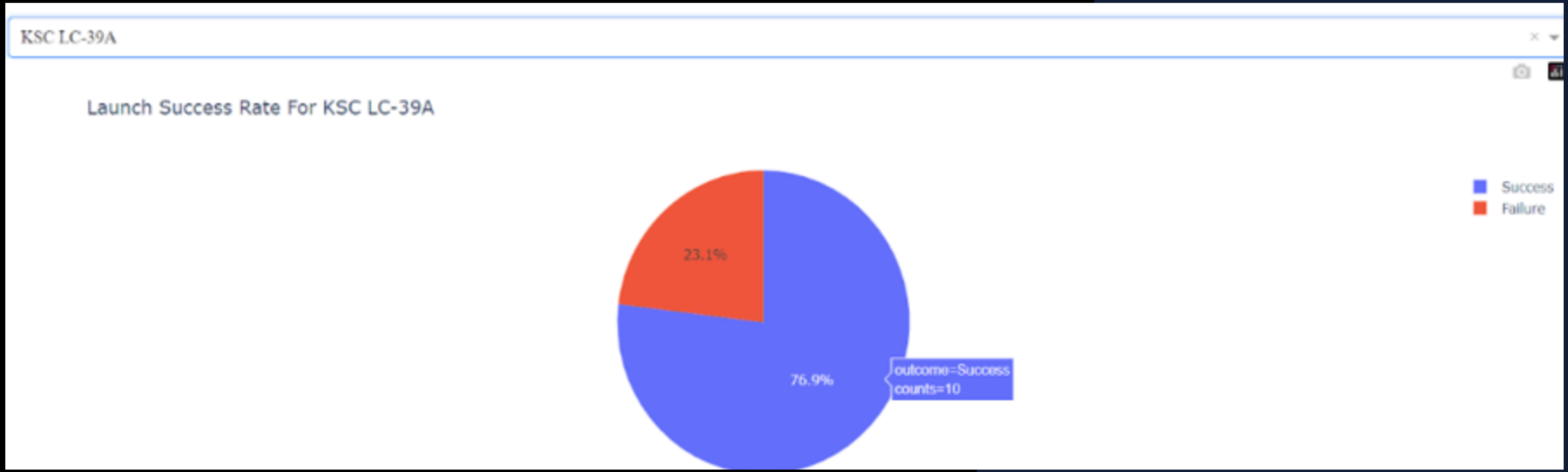
Launch success count for all sites



Explanation:

- KSC LC-39A has the most successful launches
- The success launches amounts of CCAFS SLC-40 and VAFB SLC-4E are closed

Launch site with highest launch success ratio

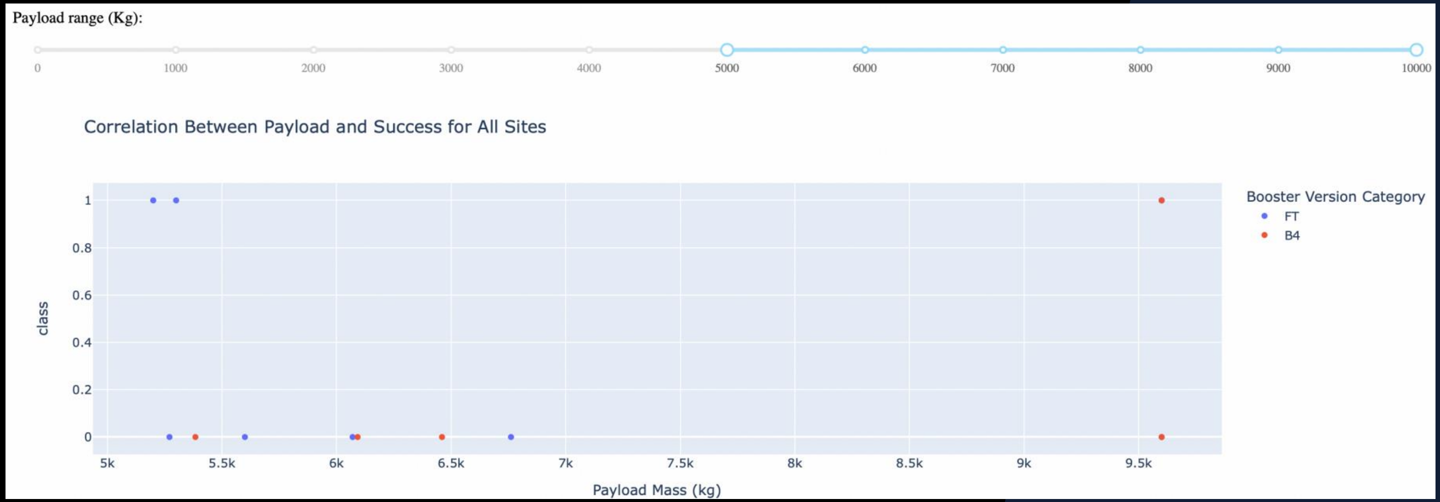


Explanation:

- The launch success rate of KSC LC-39A was very high and closed to 77%

Results- Build a Dashboard with Plotly Dash

Payload Mass vs. Launch Outcome for all sites



Explanation:

- Payloads between 3000-4000kg has the highest success rate, while payloads between 5000-7000kg has the lowest success rate
- Booster version FT in the payload between 2000-4000kg has the highest success rate

Predictive analysis (Classification)

Classification Accuracy

Test Set - Scores and Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Entire Data- Scores and Accuracy

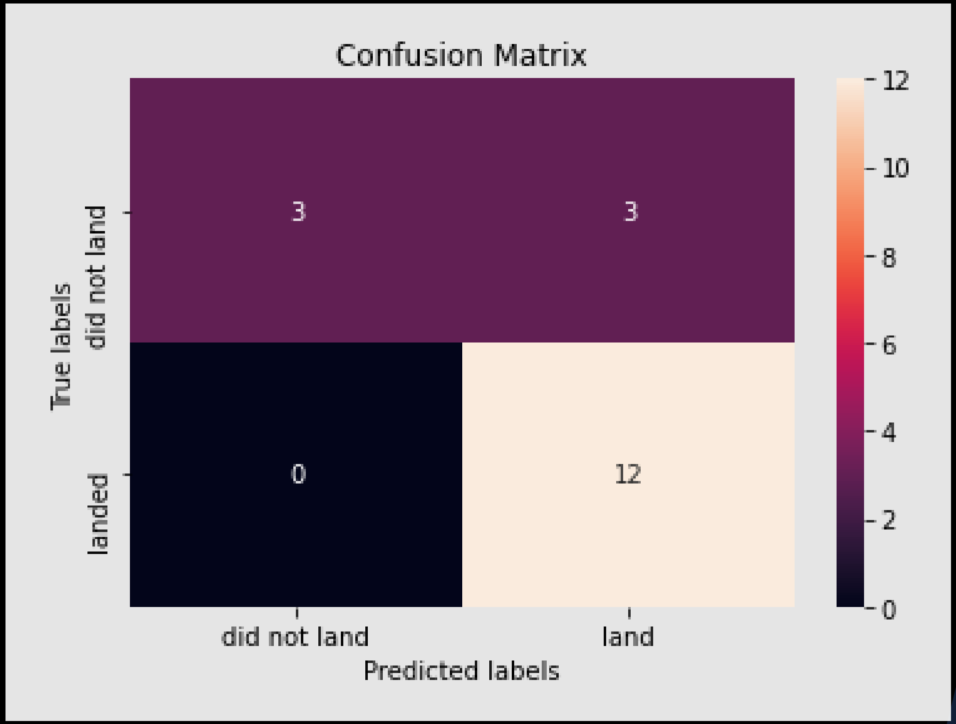
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Explanation:

- Because the scores of those methods on Test Set are same, we can not tell the difference of every method’s performance.
- When the data size is too small, it may cause the result that we can’t easily compare the performance of each method.

- When we extend to the whole dataset, the scores confirm that the best model is the Decision Tree Model.
- This model has not only higher scores, but also the highest accuracy

Predictive analysis (Classification)
Confusion Matrix



Explanation:

- Predicted success result(TP) is 80% and False positive is 20%
- Through using advancing tech and new algorithm, the results for true negatives and false positives may become more accurate

Conclusion

- KSC LC-39A has the highest success rate of the launches from all the sites.
- Every launch site: the higher the payload mass, the higher the success rate.
- The success rate kept increasing since 2013. Only drop a bit at 2018.

  - It is possible that the experience had a great impact on success.
- The GTO orbit has a very large variance of success and failure and should be avoided until you find out the real reason.
- Due to the greater proximity to the earth operating at the observation points, the VLEO orbit has been more used in recent launches.
- Most of launch sites are in proximity to the Equator line, and all the sites are in very close proximity to the coast.
- Reasonable distance of KSC LC-39A to railway, highway, coastline, and closest city Titusville are all between 15-20 km.
- Decision Tree Model is the best algorithm for the whole dataset.

Special Thanks

IBM Data Science Professional Certificate Course

IBM

COURSERA

Github:
IBM_Applied_Data_Science_Capstone_SpaceX

Thank You

YU LIN YEH /2024.03.06