Data Science Capstone Project IBM

RITVIK JOHNSON

05/01/2024

Outline

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

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 all results

- It was possible to collected valuable data from public sources.

- EDA allowed to identify which features are the best to predict success of launchings.

- Machine Learning Prediction showed the best model to predict which characteristics are.

INTRODUCTION

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage

SCISPACE

RESEARCH METHODOLOGY

Everything You need to Know

Methodology

Executive Summary

- Data collection methodology:
    - Data from Space X was obtained from 2 sources:
    - Space X API (‘https://api.spacexdata.com/v4/rockets/’)
    - WebScraping (‘https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches’)
- Perform data wrangling
    - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

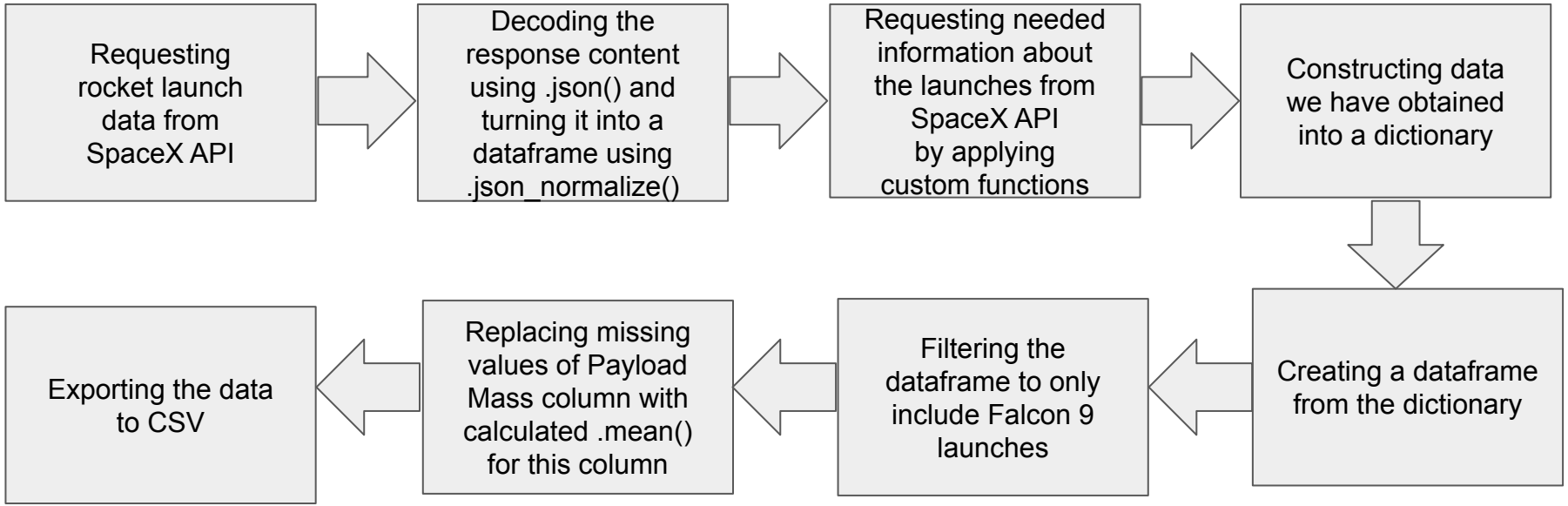
Data Collection

• Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

[25]:

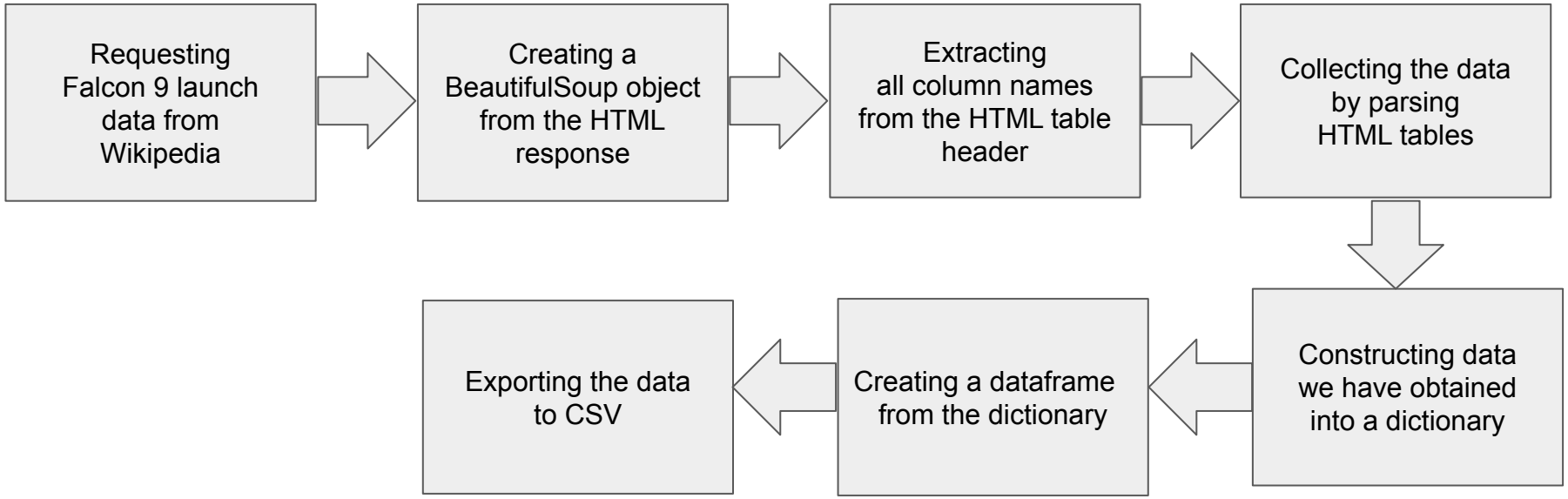
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	Re
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	

Data Collection – SpaceX API



[GitHub URL : Data Collection With API]

Data Collection - Web Scraping



[GitHub URL : Data Collection With Web Scraping]

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

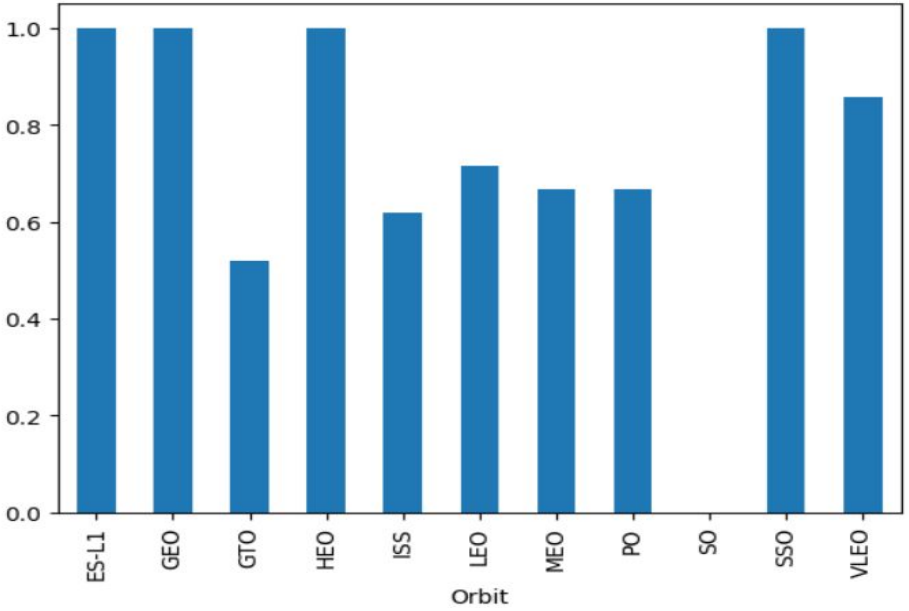
We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

GitHub URL : Data Wrangling

EDA with Data Visualization

Charts were plotted:

- 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



[GitHub URL : EDA with Data Visualization]

EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000
  - Total number of successful and failure mission outcomes;

EDA with SQL

- Names of the booster versions which have carried the maximum payload mass. Use a subquery
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (groundpad)) between the date 2010-06-04 and 2017-03-20, in descending order.

GitHub URL : EDA With SQL

Build an Interactive Visual Analytics 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.

GitHub URL : Build an Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

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

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

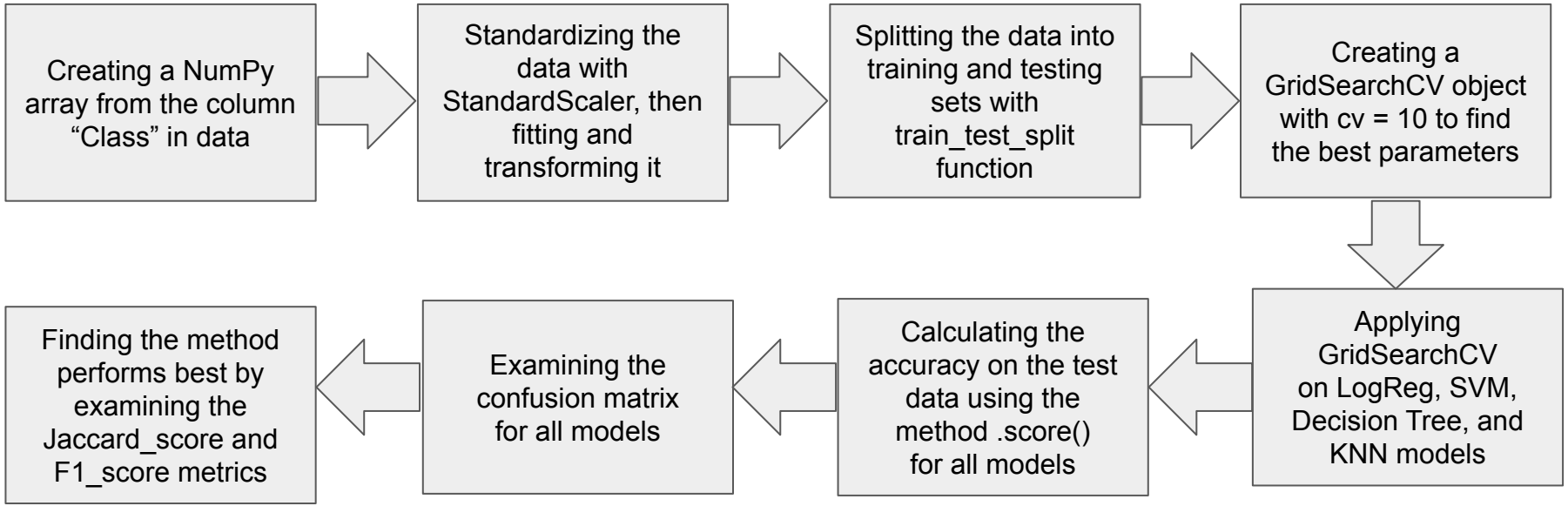
- Added a slider to select Payload range.

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

- Added a scatter chart to show the correlation between Payload and Launch Success.

GitHub URL : Build a Dashboard with Plotly Dash

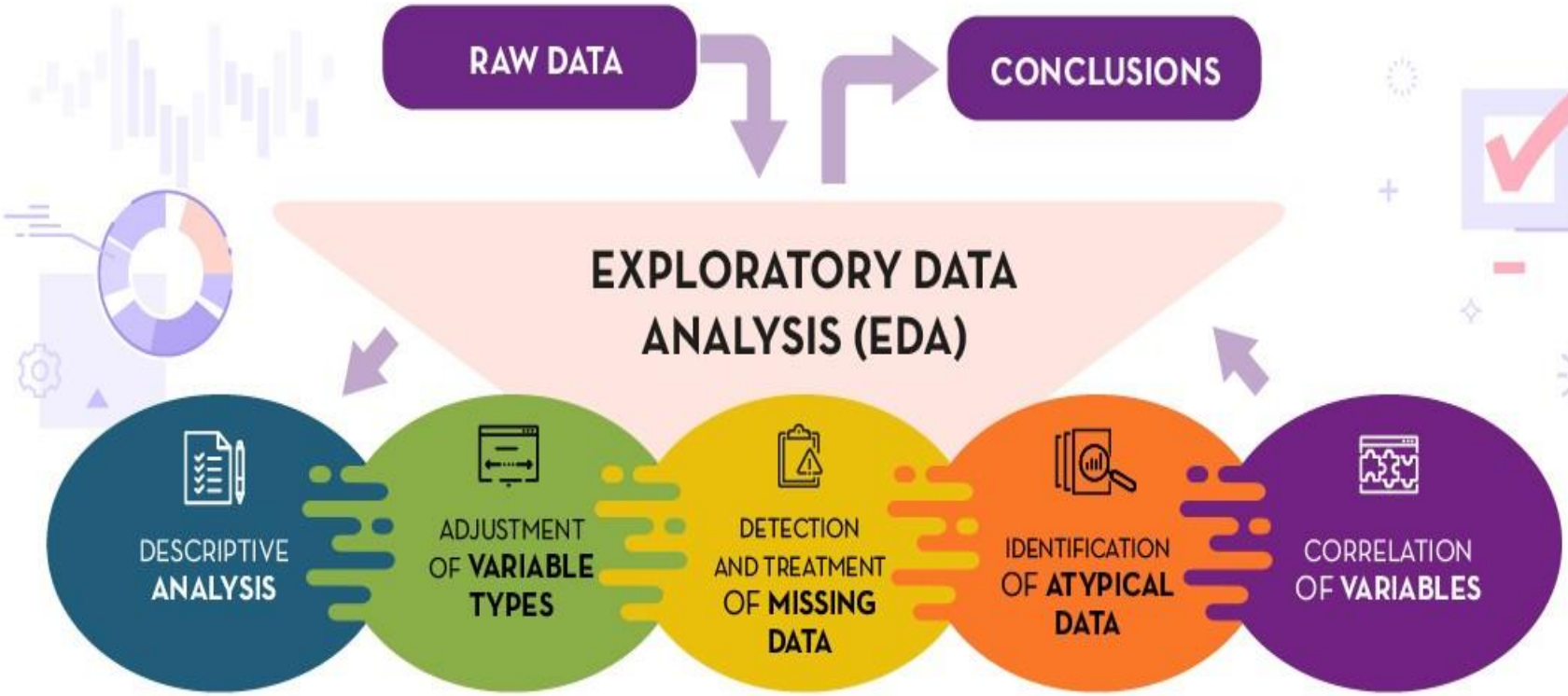
Predictive Analysis (Machine Learning)



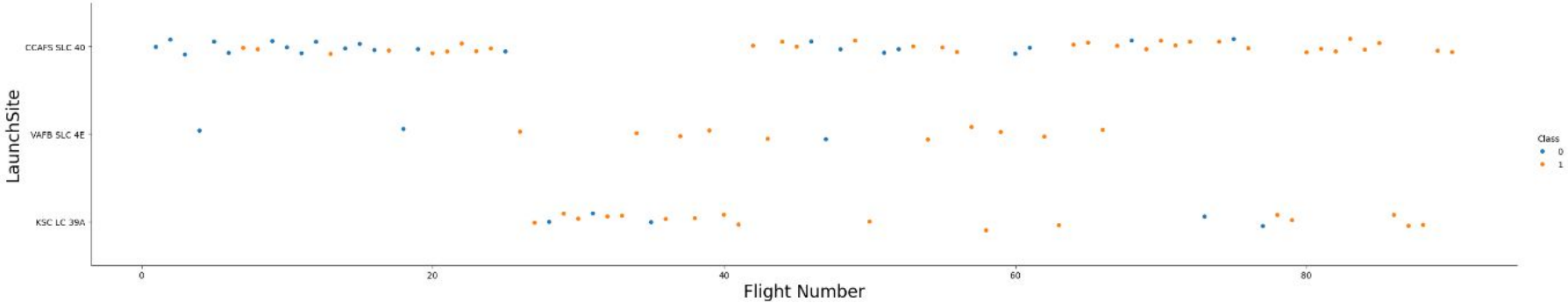
[GitHub URL : Predictive Analysis (Machine Learning)]

Results

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

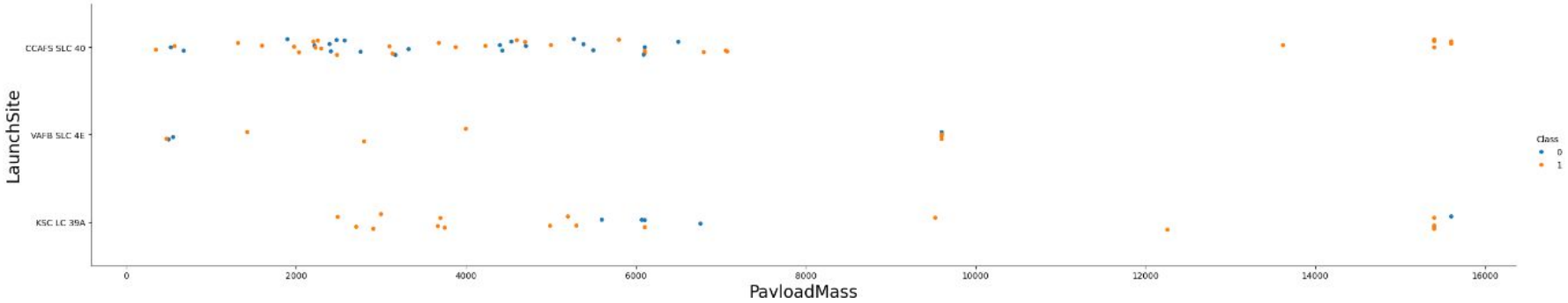

RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.
- In second place VAFB SLC 4E and third place KSC LC 39A.
- It's also possible to see that the general success rate improved over time.

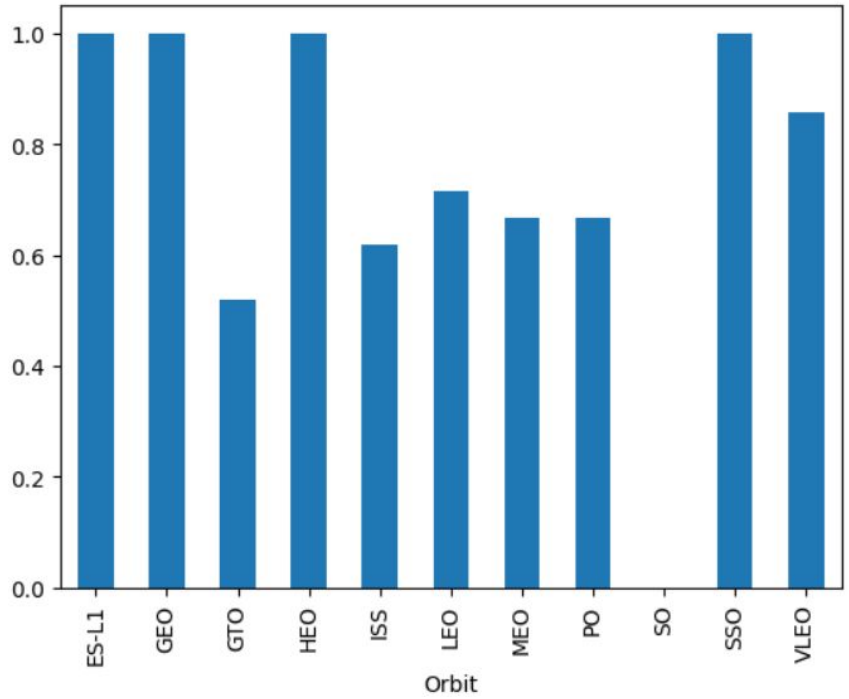
Payload Mass vs. Launch Site



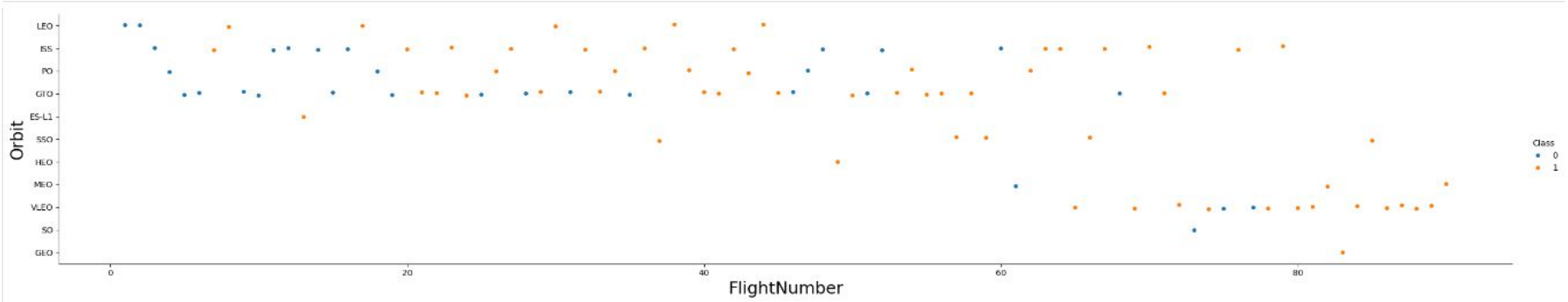
- Payloads over 9,000kg (about the weight of a school bus) have excellent success Rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit

- The biggest success rates happens to orbits:
- ES-L1
- GEO
- HEO
- SSO.
- VLEO (above 80%) and
- LFO (above 70%).

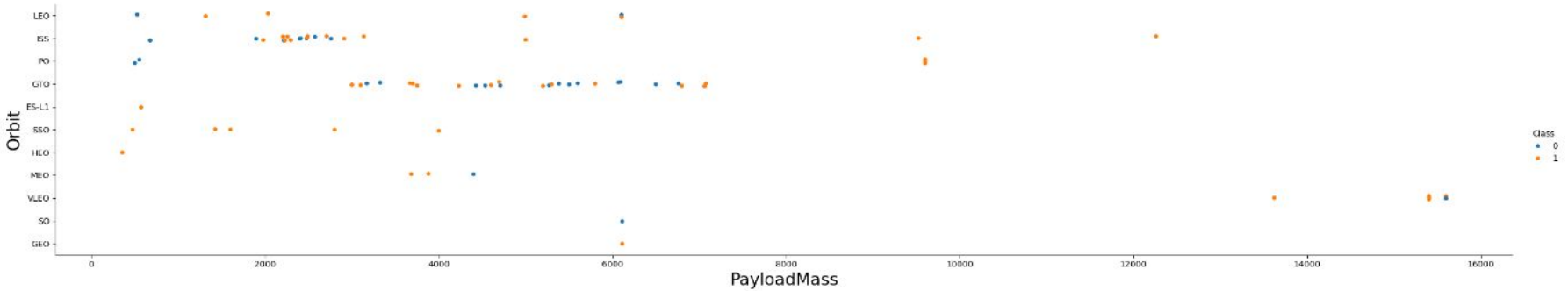


Flight Number vs. Orbit



- Apparently, success rate improved over time to all orbits.
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

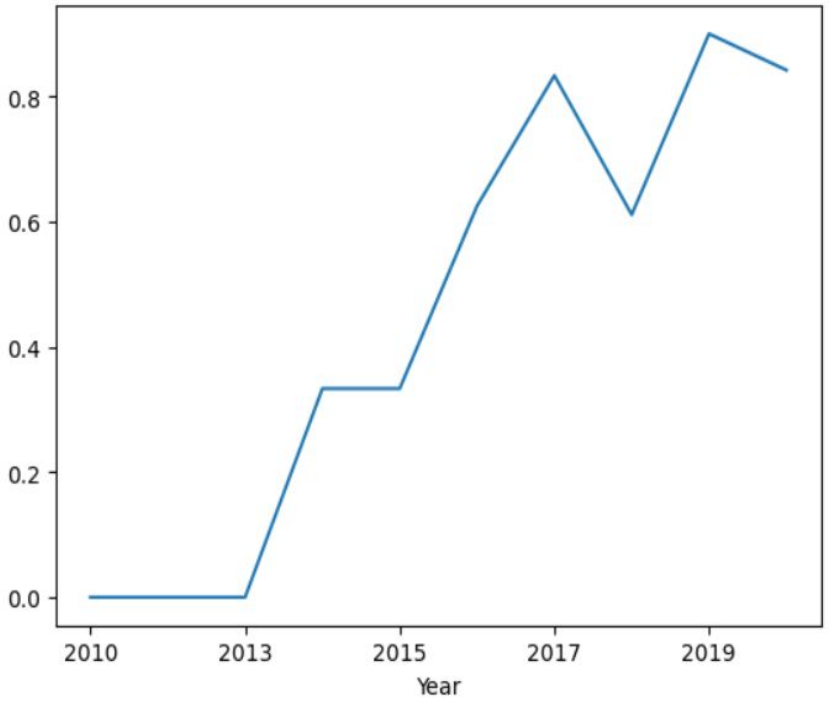
Payload Mass vs. Orbit Type



- Apparently, there is no relation between payload and success rate to orbit GTO.
- ISS orbit has the widest range of payload and a good rate of success.
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020.
- It seems that the first three years were a period of adjusts and improvement of technology.



EDA with SQL

All Launch Site Names

- According to data, there are four launch sites
- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

Task 1

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

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

Task 2

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

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

```
 * sqlite:///my_data1.db
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

Task 3

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

```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD_ LIKE '%CRS%';
```

```
 * sqlite:///my_data1.db
Done.
```

TOTAL_PAYLOAD
111268

- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

Task 4

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

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

AVG_PAYLOAD
2928.4

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

FIRST_SUCCESS_GP
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload Mass

Task 8

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

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VE
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Site Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
sql SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone sh
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The list above has the only two occurrences of 2015 Launch Site.

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

Task 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.

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

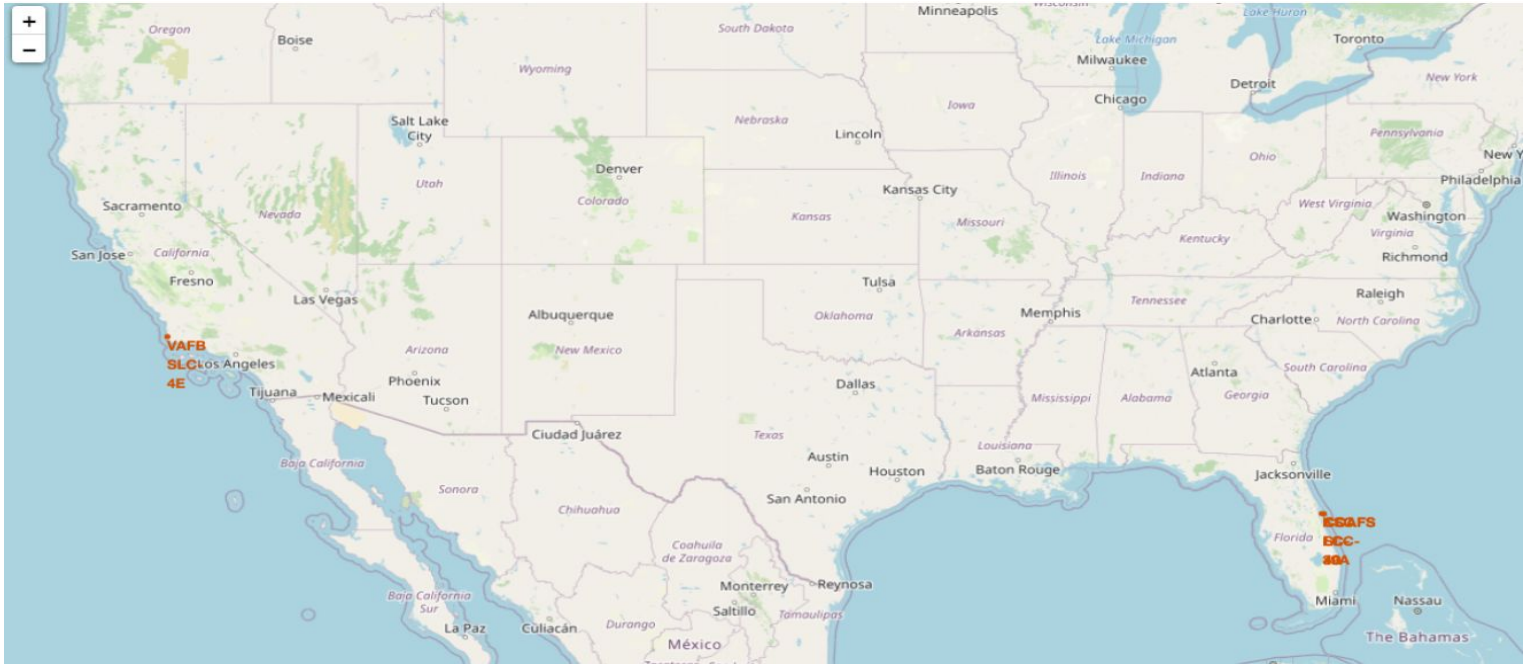
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	QTY
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This view of data alerts us that “No attempt” of 10 must be taken in account.

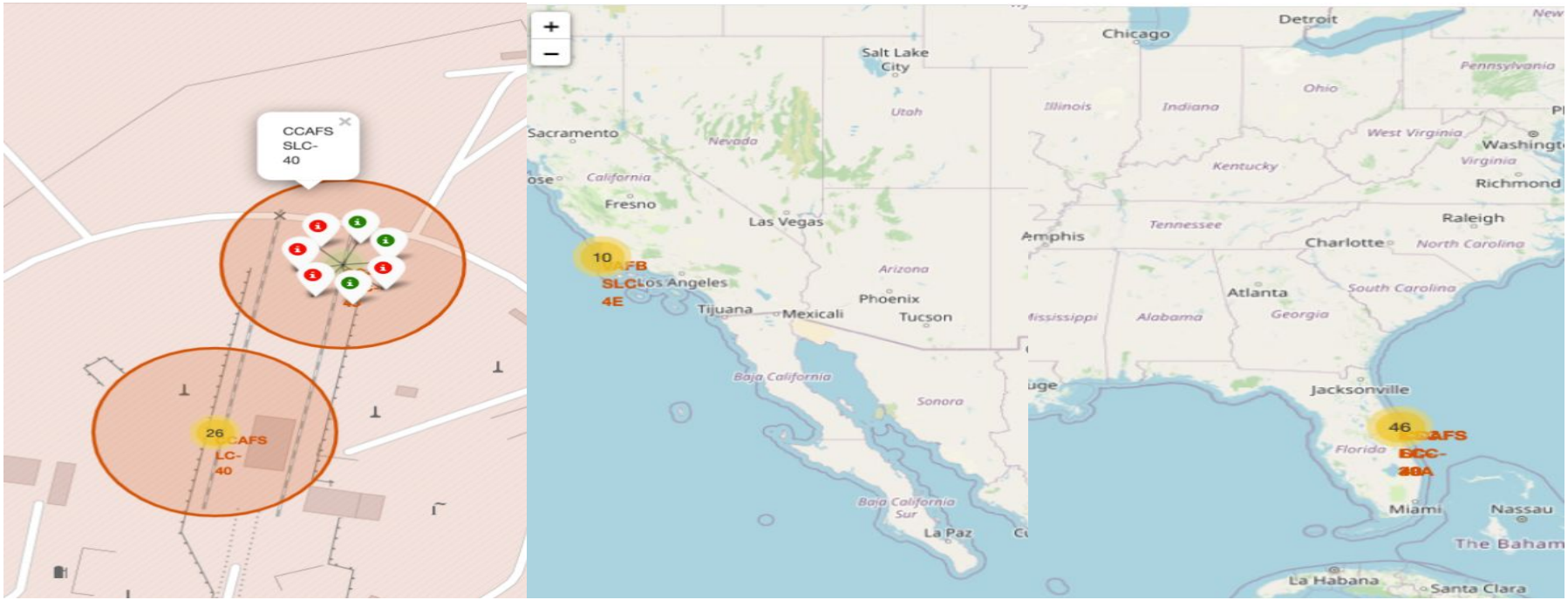
Interactive map with Folium

All launch sites



- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

Success/Failed Launches for each site

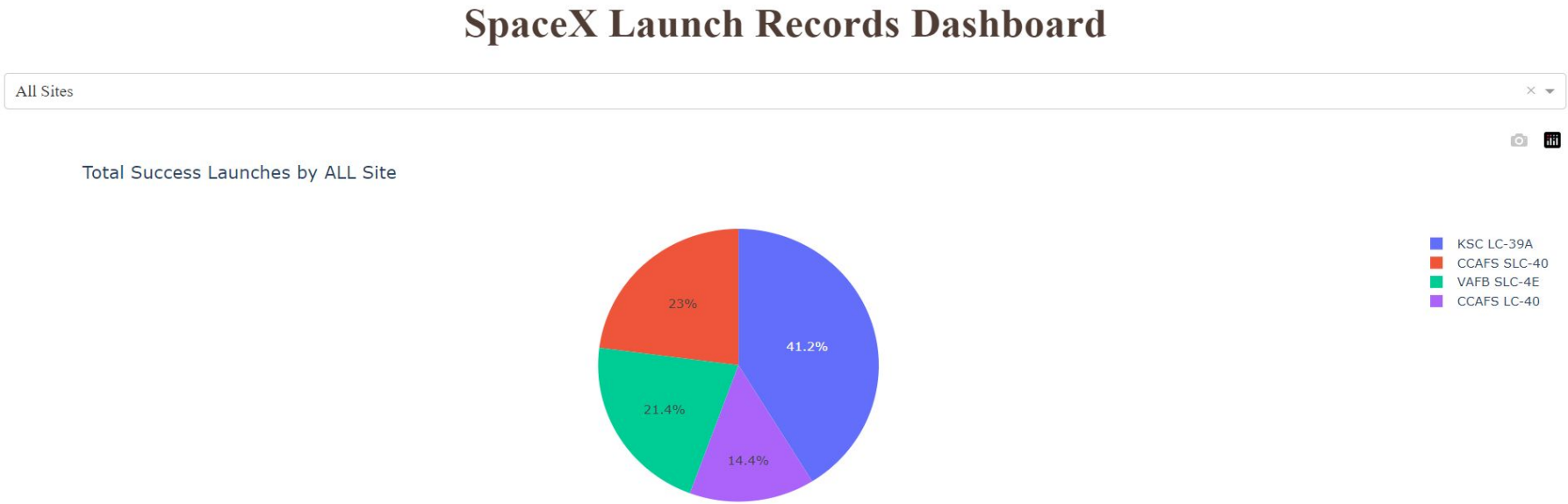


- Green markers indicate successful and red ones indicate failure

Distances between a launch site

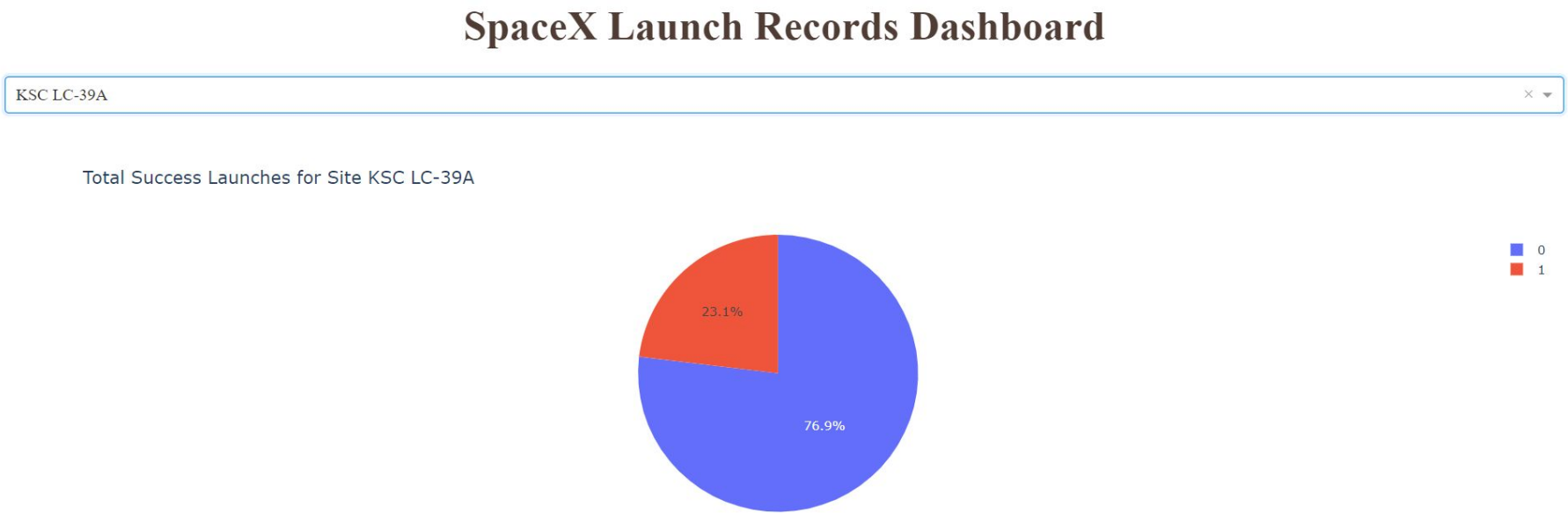
Build a Dashboard with Plotly Dash

Successful Launches count of all Site



- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch Success Ratio for KSC LC-39A



• KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

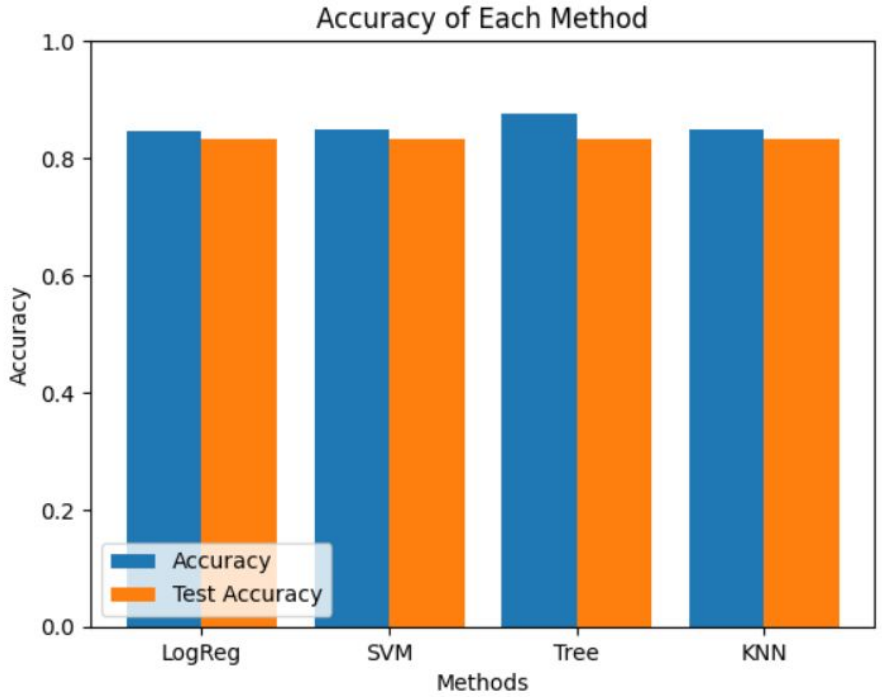
Payload Mass vs. Launch Outcome for All Sites

- The charts show that payloads between 2000 and 6000 kg have the highest success rate.

Predictive analysis (Machine Learning)

Classification Accuracy

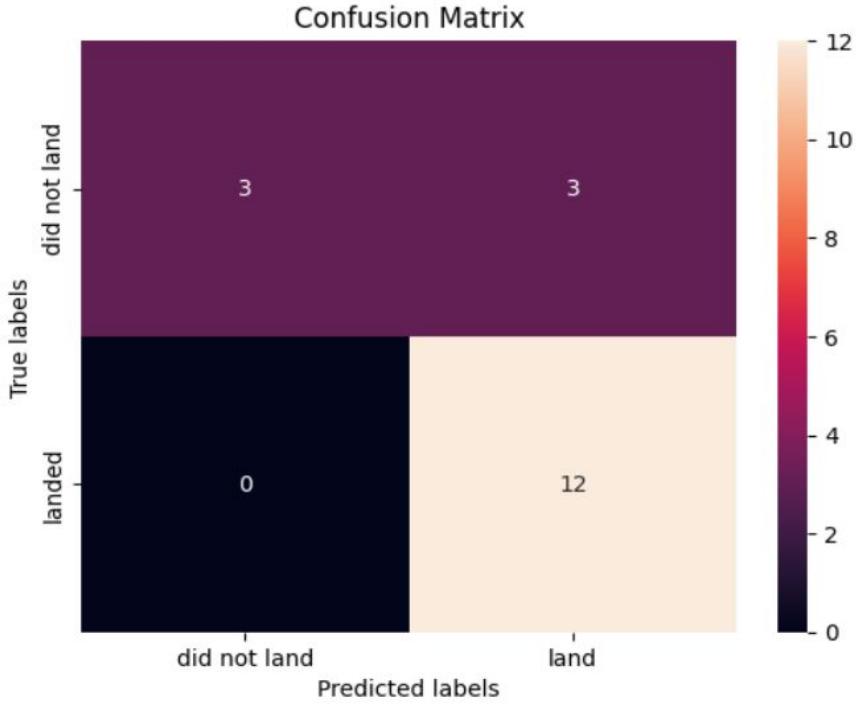
- Four classification models were tested, and their accuracies are plotted beside.
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix of Decision Tree Classifier

Explanation:

• Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Different data sources were analyzed, refining conclusions along the process.
- The best launch site is KSC LC-39A.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Launches above 7,000kg are less risky.
- Most of launch sites are in proximity to the Equator lineand all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.

Appendix

Special Thanks to:

INSTRUCTOR

COURSERA

IBM


THANK YOU!