



Data Science Capstone Project

Prepared by :
BOUGHZALA Ghailene

Outline

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

Executive Summary

Project's steps & 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)

Evaluation of results

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

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.

- Problematics :
 - 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

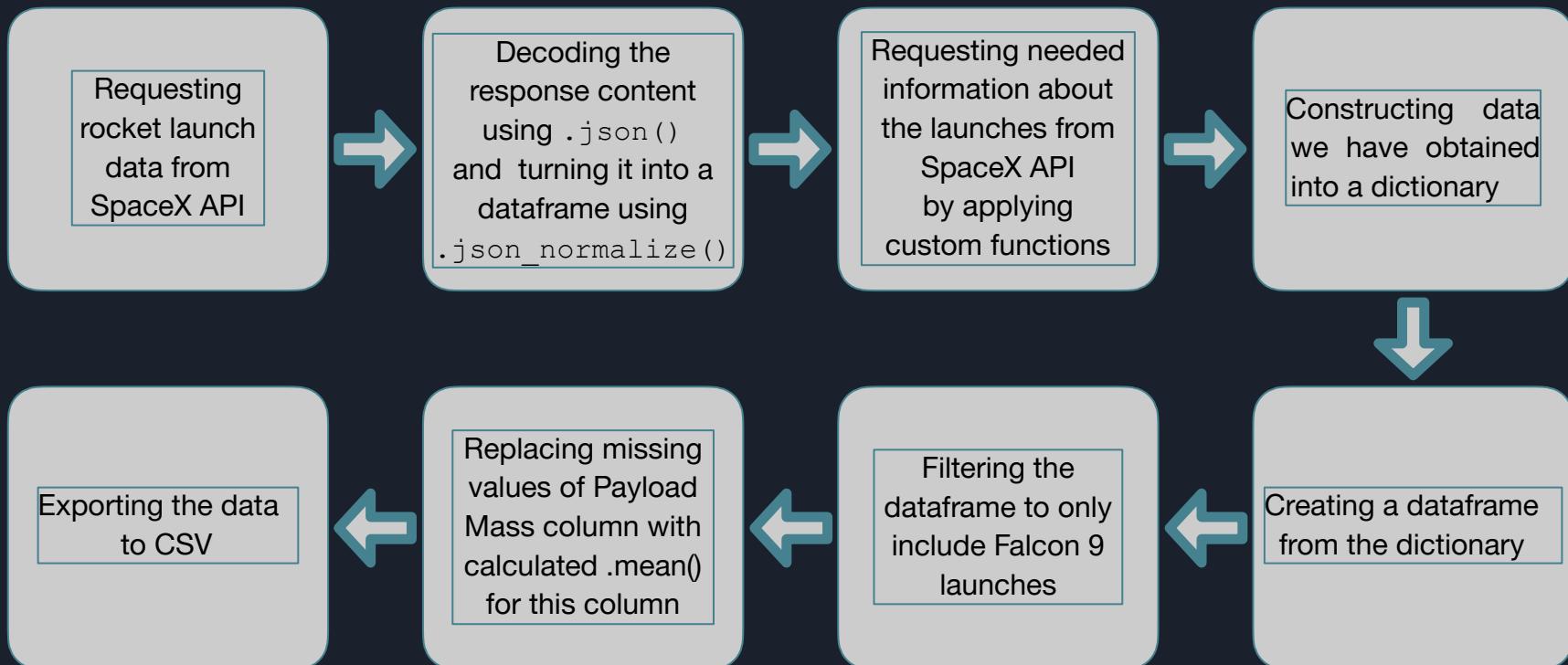
Data collection methodology:

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

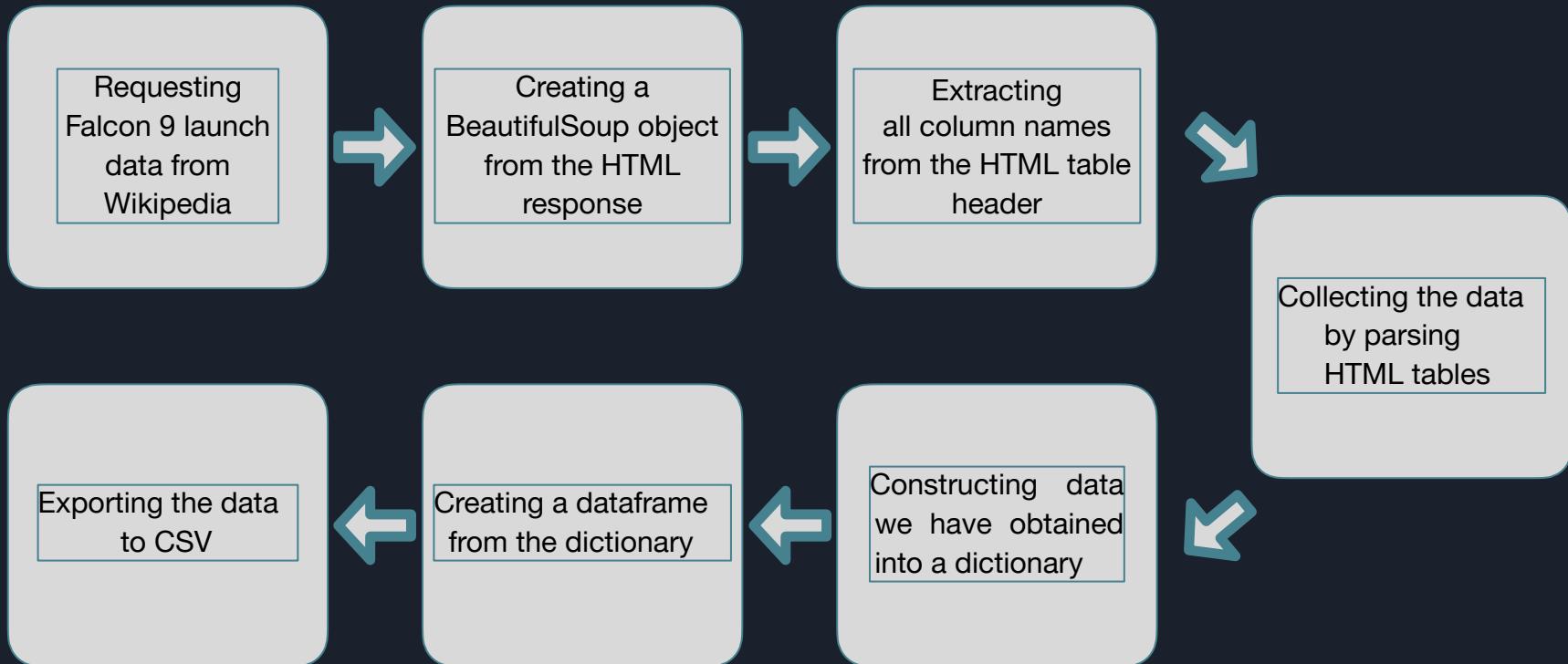
Data wrangling

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

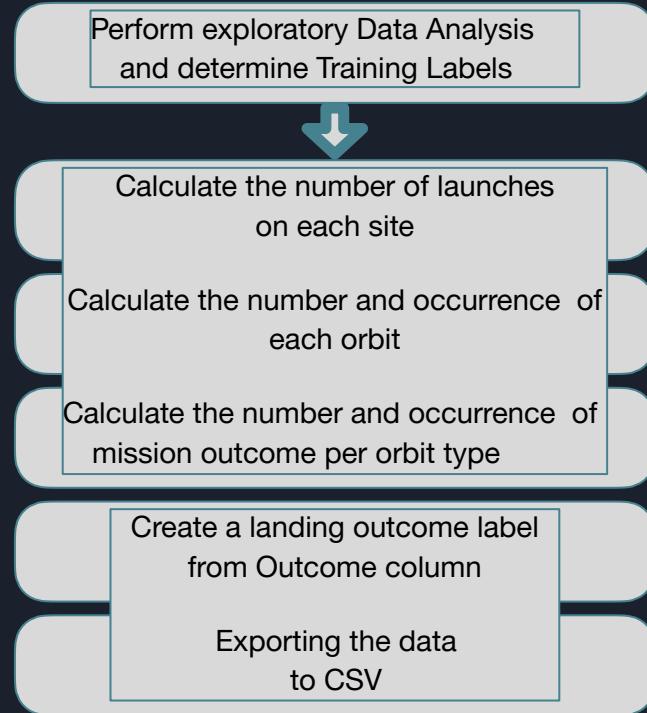
Data collection – SpaceX API



Data collection – Web scraping



Data wrangling



EDA with data visualization

The charts were successfully implemented :

- 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
-
- Scatter plots to determine the relationship between variables.
 - Bar charts to make comparisons among discrete categories. T\$
 - Line charts to show trends in data over time (time series).

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- 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

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.

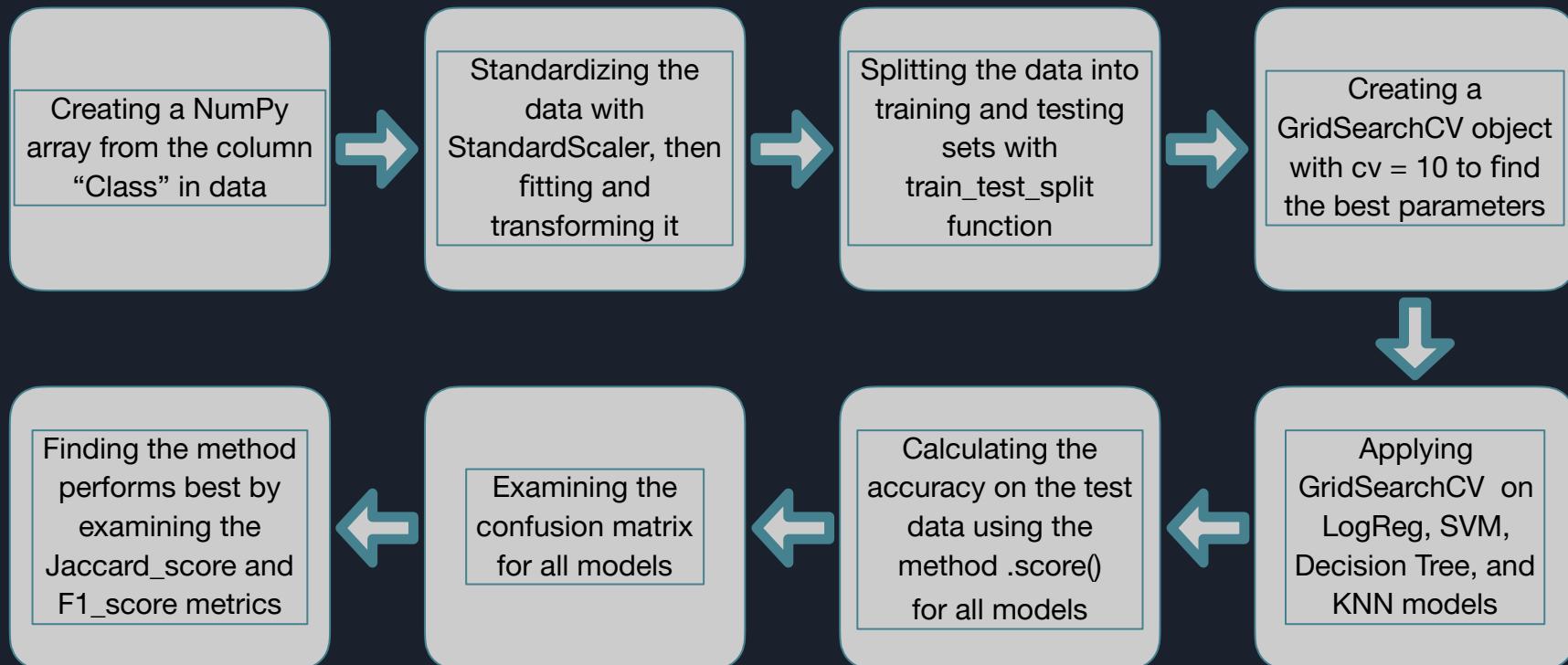
Coloured Markers of the launch outcomes for each Launch Site:

- 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.

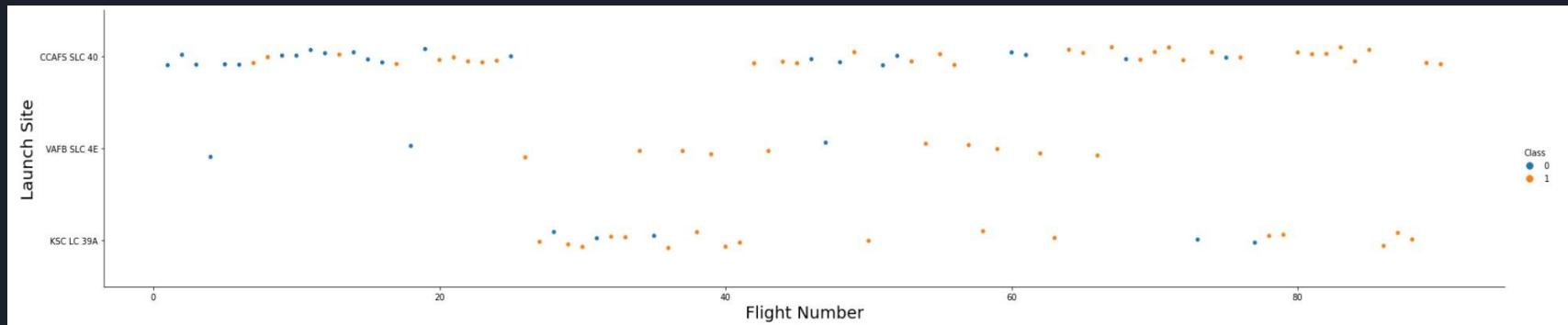
Predictive analysis (Classification)



Evaluation of results

EDA with Visualization

Flight Number vs. Launch Site

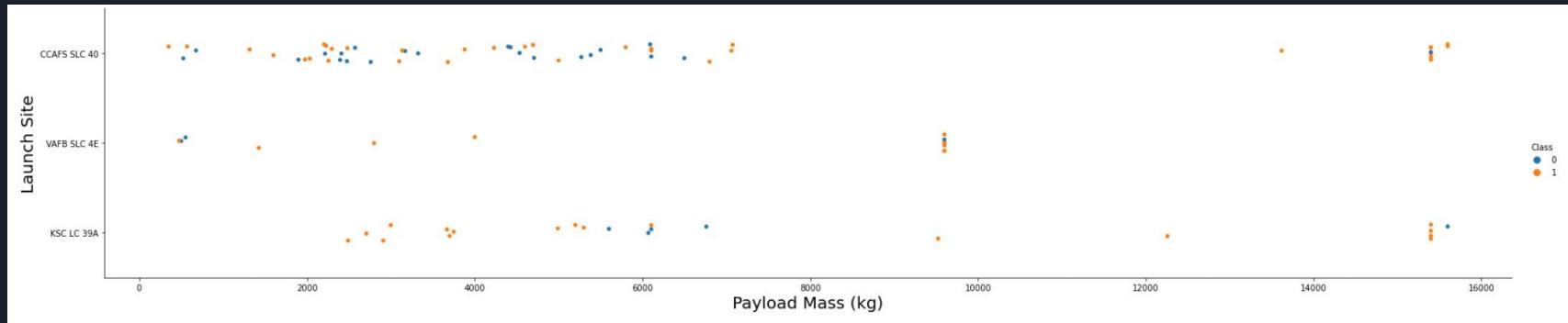


Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

EDA with Visualization

Payload vs. Launch Site



Explanation:

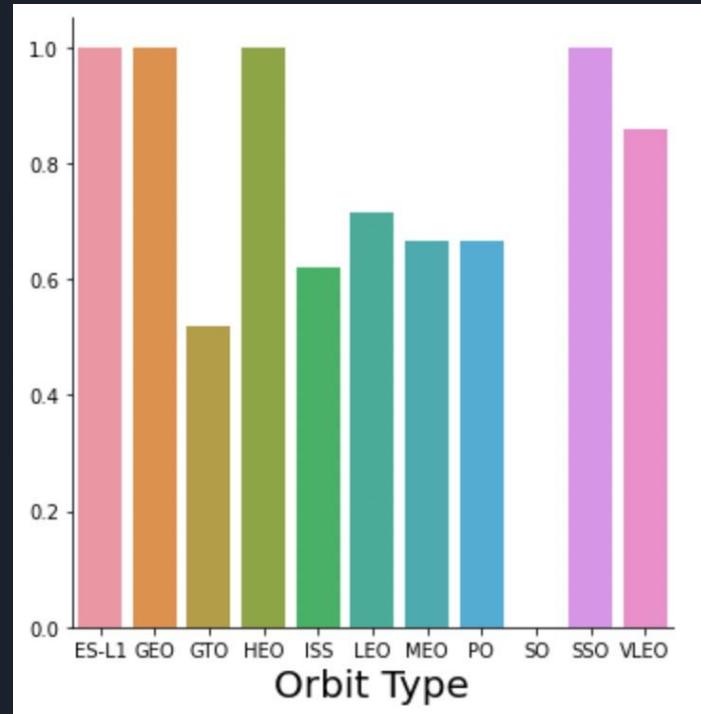
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

EDA with Visualization

Success rate vs. Orbit type

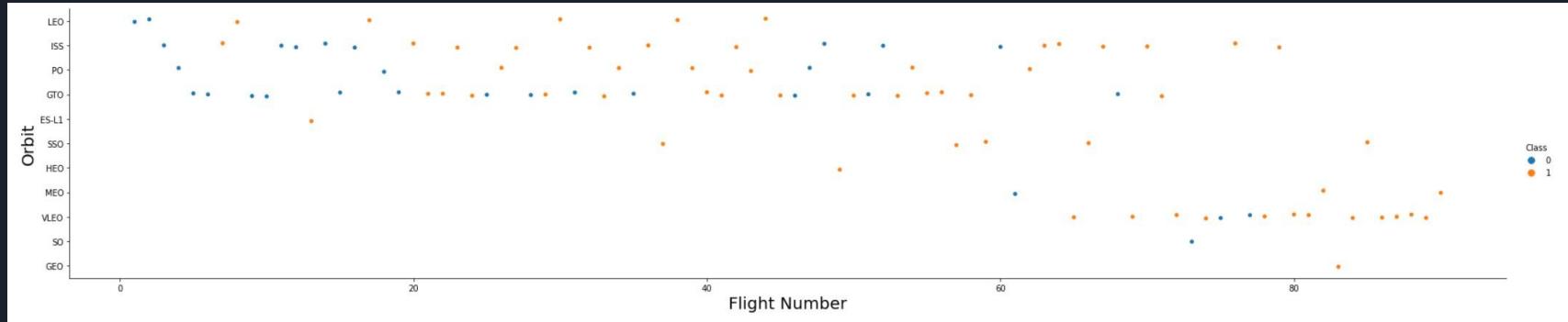
Explanation:

- Orbit types with 100% success rate:
 - ES-L1, GEO, HEO, SSO
- Orbit types with 0% success rate:
 - SO
- Orbit types with success rate between 50% and 85%:
 - GTO, ISS, LEO, MEO, PO



EDA with Visualization

Flight Number vs. Orbit type

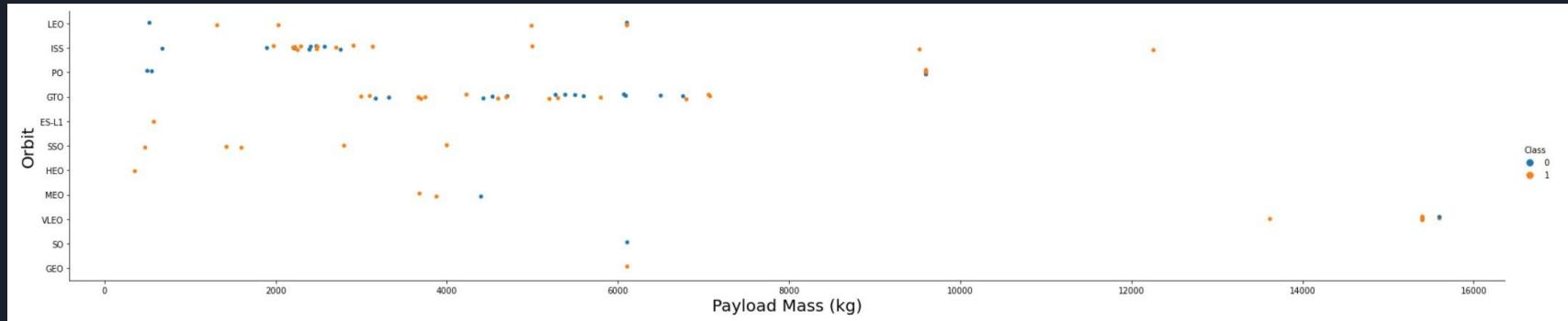


Explanation:

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

EDA with Visualization

Payload Mass vs. Orbit type



Explanation:

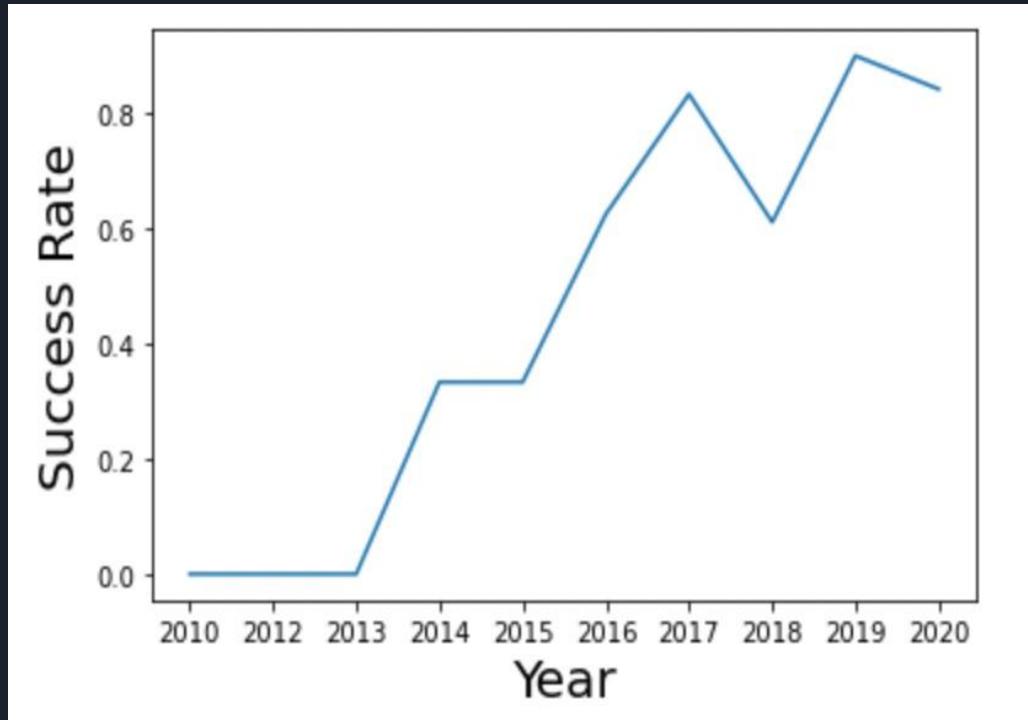
- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

EDA with Visualization

Launch success yearly trend

Explanation:

- The success rate since 2013 kept increasing till 2020.



EDA with SQL

EDA with SQL

All launch site names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bludb  
Done.
```

Out[4]:

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

Explanation:

- Displaying the names of the unique launch sites in the space mission.

EDA with SQL

Launch site names begin with 'CCA'

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

Explanation:

- Displaying 5 records where launch sites begin with the string 'CCA'.

EDA with SQL

Total payload mass

```
In [6]: %sql select sum(payload_mass_kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';  
* ibm_db_sa://wzsf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bludb  
Done.
```

```
Out[6]: total_payload_mass  
45596
```

Explanation:

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

EDA with SQL

Average payload mass by F9 v1.1

average_payload_mass
2534

Explanation:

- Displaying average payload mass carried by booster version F9 v1.1.

EDA with SQL

First successful ground landing date

Done!

first_successful_landing

2015-12-22

Explanation:

- Listing the date when the first successful landing outcome in ground pad was achieved.

EDA with SQL

Successful drone ship landing with payload
between 4000 and 6000

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

Explanation:

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

EDA with SQL

Total number of successful and failure mission outcomes

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

Explanation:

- Listing the total number of successful and failure mission outcomes.

EDA with SQL

Boosters carried maximum payload

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

Explanation:

- Listing the names of the booster versions which have carried the maximum payload mass.

EDA with SQL

2015 launch records

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)

Explanation:

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

EDA with SQL

Rank success count between 2010-06-04 and 2017-03-20

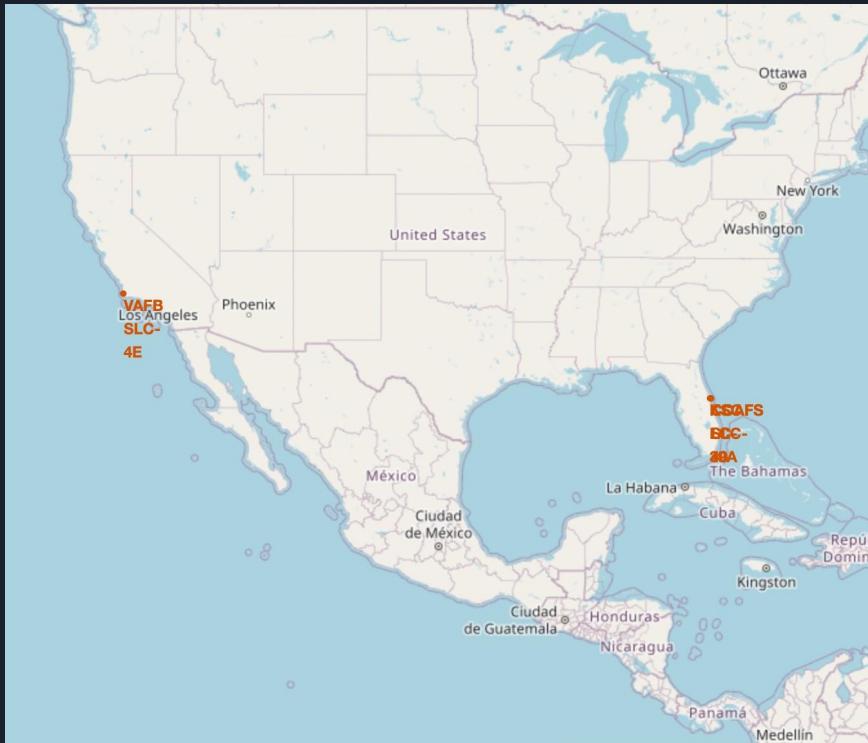
landing_outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Explanation:

- 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.

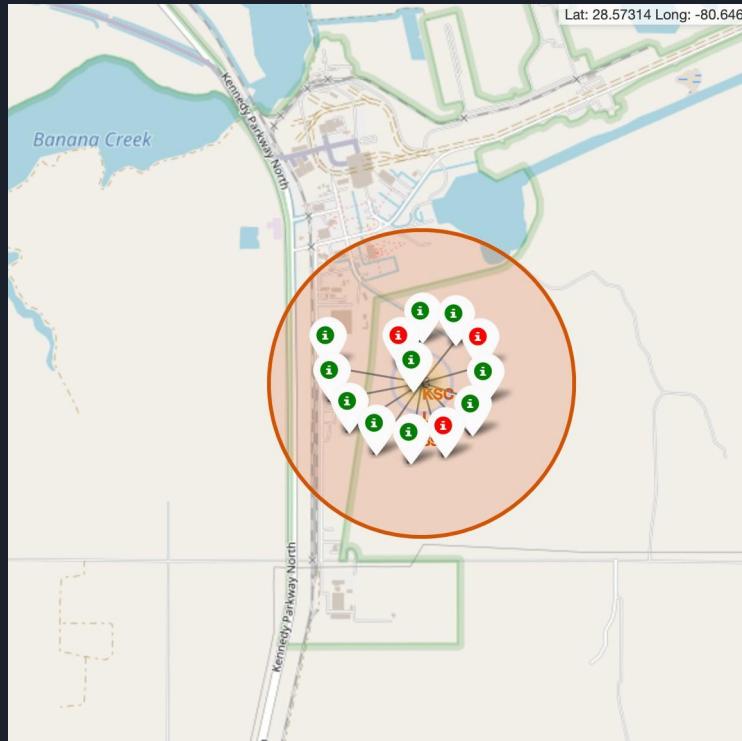
Interactive map with Folium

All launch sites' location markers on a global map



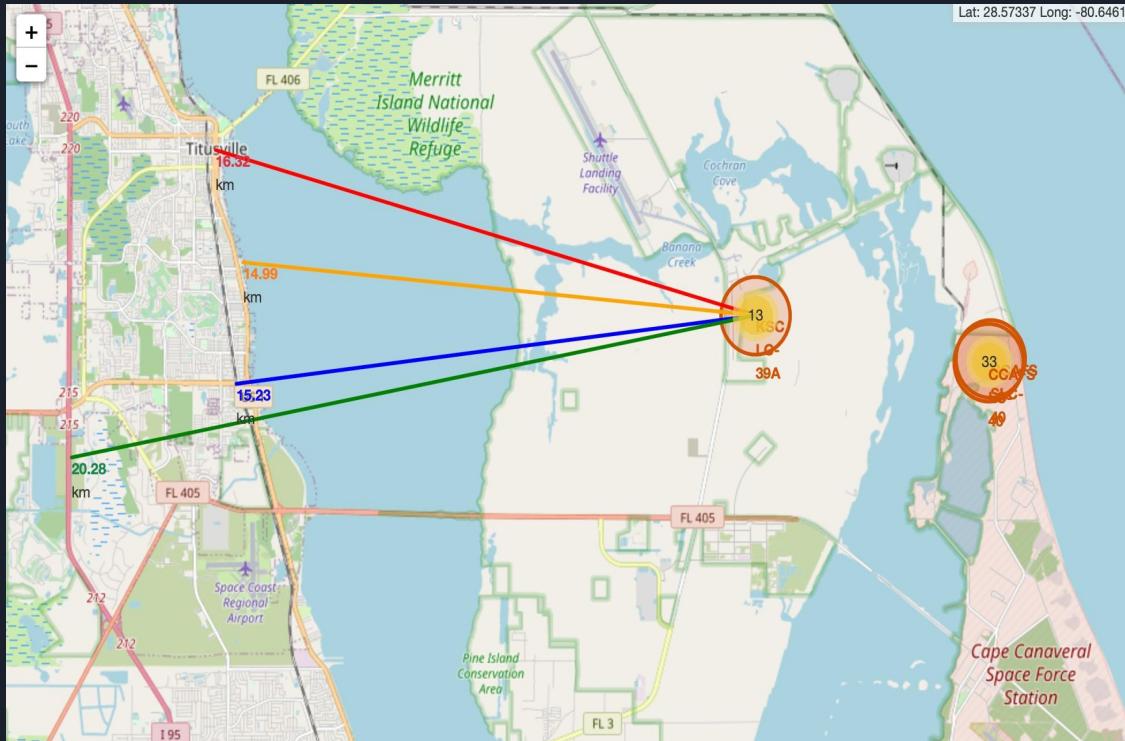
Interactive map with Folium

Colour-labeled launch records on the map



Interactive map with Folium

Distance from the launch site KSC LC-39A to its proximities

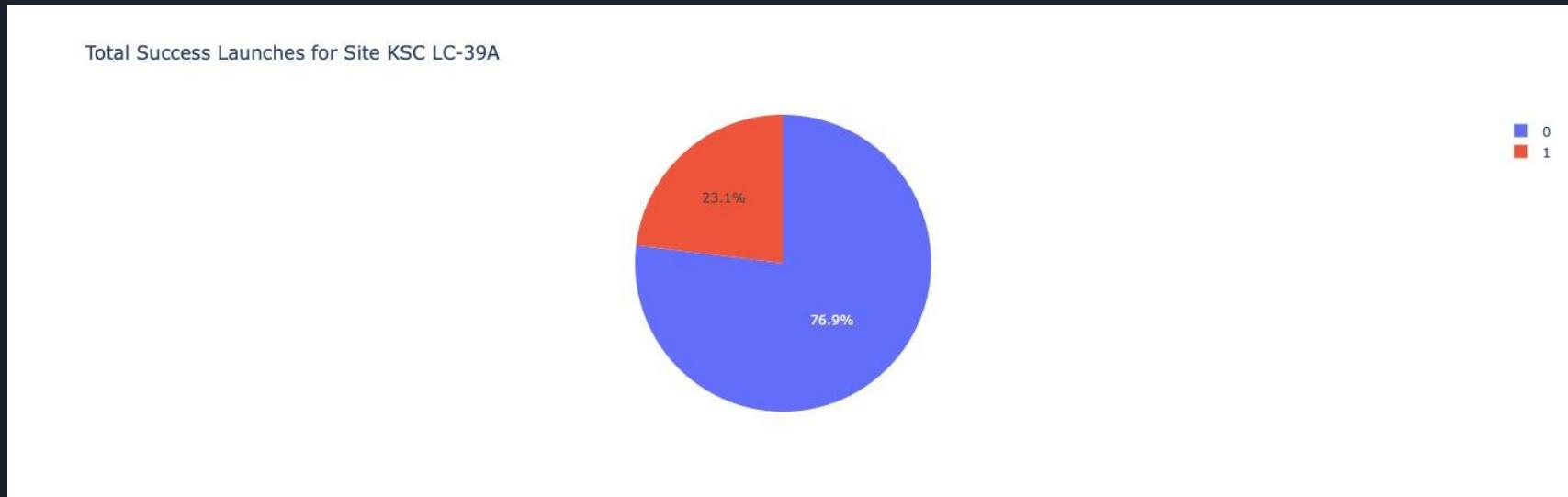


Launch success count for all sites

Total Success Launches by Site



Launch site with highest launch success ratio



Payload Mass vs. Launch Outcome for all sites



Classification Accuracy

Explanation:

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test Set

	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

Scores and Accuracy of the Entire Data Set

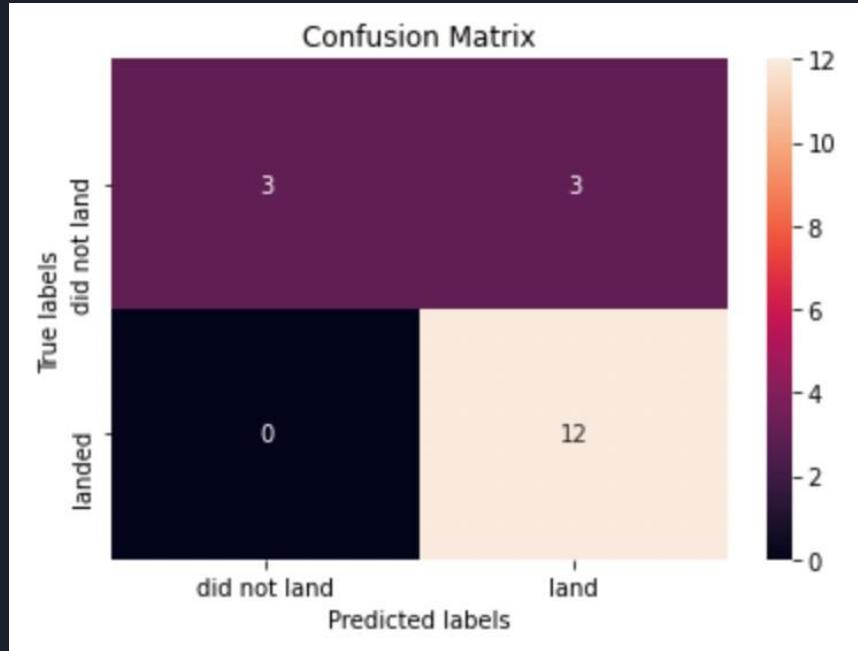
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

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.

		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP



Conclusion

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy
-

Appendix

Special Thanks to:

Instructors Coursera IBM