COUSERA DATA ANALYSE

SPACEX DATA

OUTLINE

- Executive summary
- Introduction
- Methodology
- Result
- 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)

Introduction

PROJECT BACKGROUND AND CONTEXT

- 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.

- 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 of classification models to ensure the best results

Data collection

DATA COLLECTION PROCESS INVOLVED A COMBINATION OF API REQUESTS FROM SPACEX REST API AND WEB SCRAPING DATA FROM A TABLE IN SPACEX'S WIKIPEDIA ENTRY.

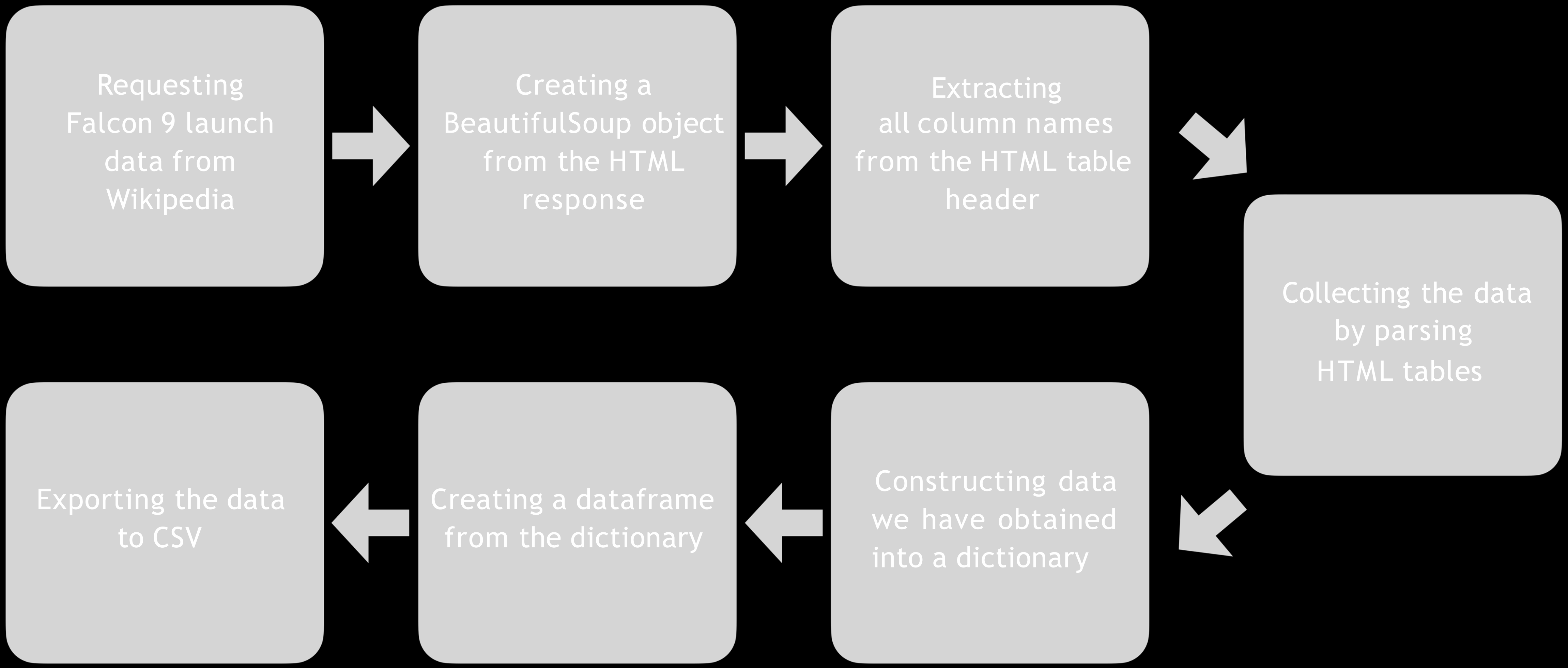
We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

- Data Columns are obtained by using SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data collection – SpaceX API

1. Requesting rocket launch data from SpaceX API
2. Decoding the response content using .json() and turning it into a dataframe using `.json_normalize()`
3. Requesting needed information about the launches from SpaceX API by applying custom functions
4. Constructing data we have obtained into a dictionary
5. Creating a dataframe from the dictionary
6. Filtering the dataframe to only include Falcon 9 launches
7. Replacing missing values of Payload Mass column with calculated .mean() for this column
8. Exporting the data to CSV

Data collection – 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.

EDA with data visualisation

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 and Success Rate Yearly Trend

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

Line charts 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

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.

Dashboard with Plotly and 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.

Predictive analysis (Classification)

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

RESULTS

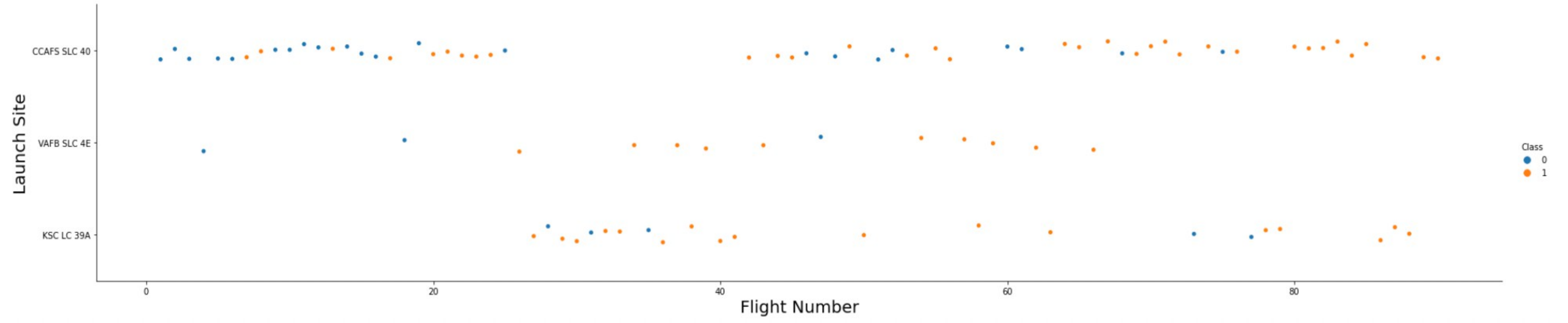
Exploratory data analysis results

Interactive analytics demo in

screenshots

Predictive analysis results

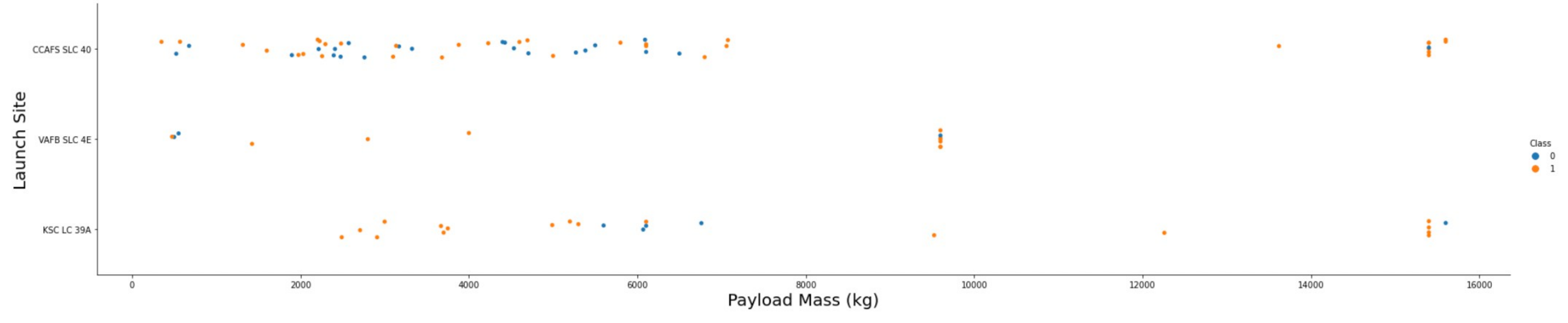
FLIGHT NUMBER VS. LAUNCH SITE



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.

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.

SUCCESS RATE VS. ORBIT TYPE

Explanation:

- Orbits with 100% success rate:

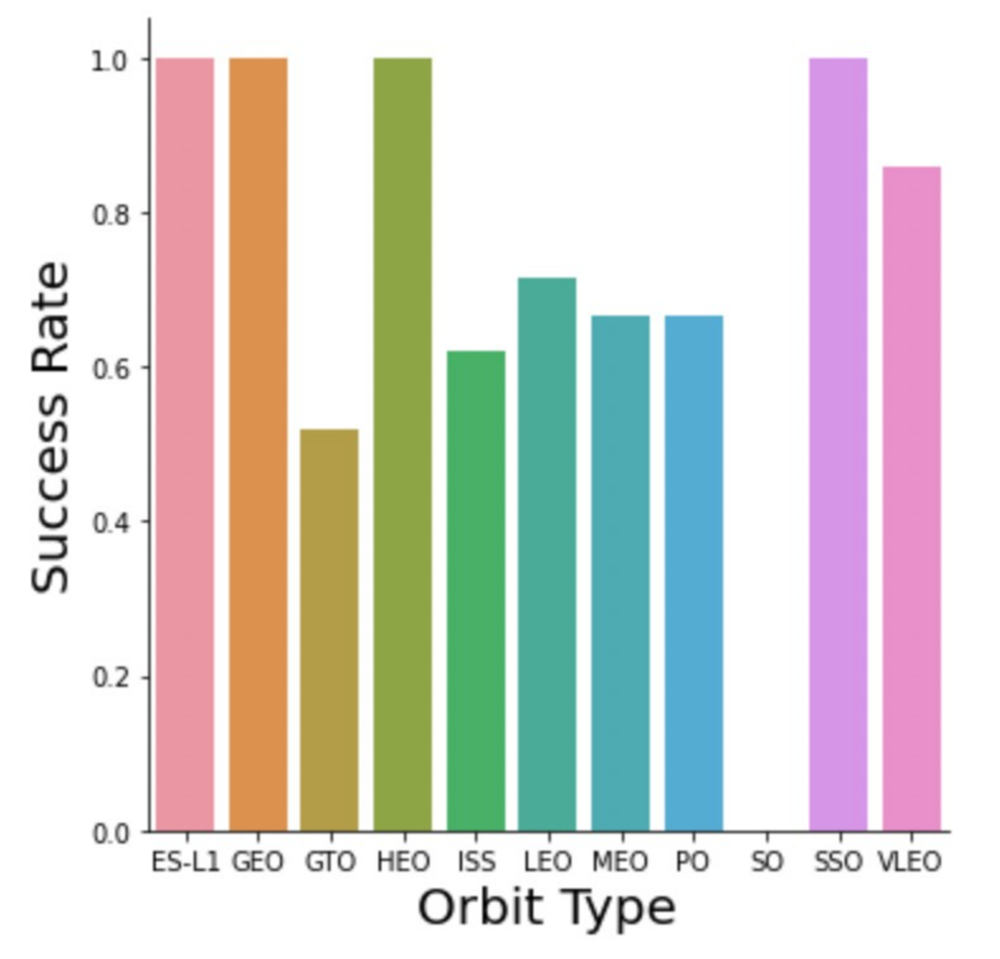
  -ES-L1, GEO, HEO, SSO

- Orbits with 0% success rate:

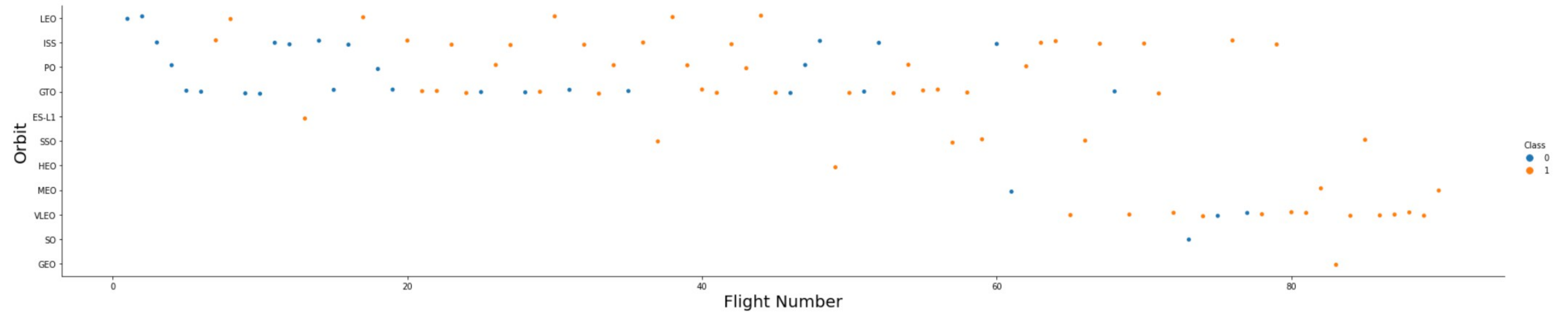
  -SO

- Orbits with success rate between 50% and 85%:

  -GTO, ISS, LEO, MEO, PO



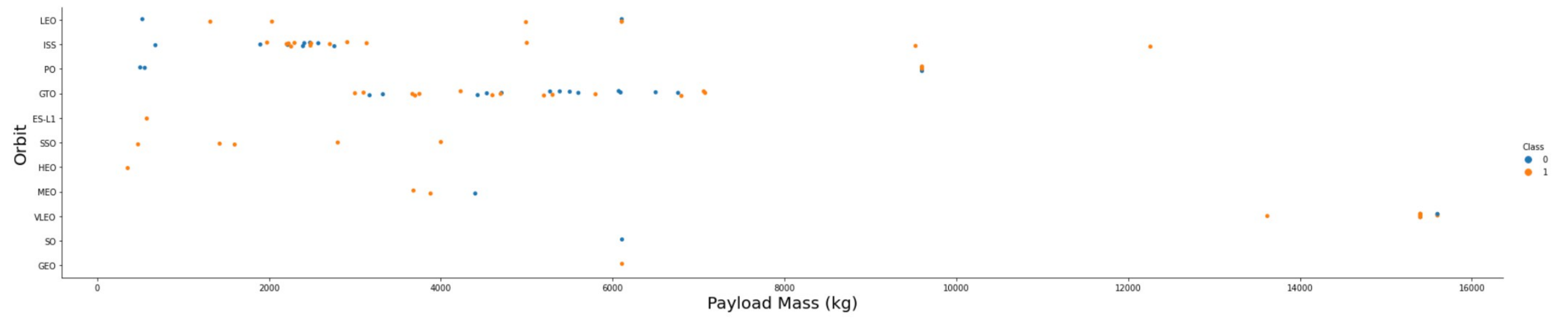
FLIGHT NUMBER VS. ORBIT TYPE



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.

PAYLOAD MASS VS. ORBIT TYPE



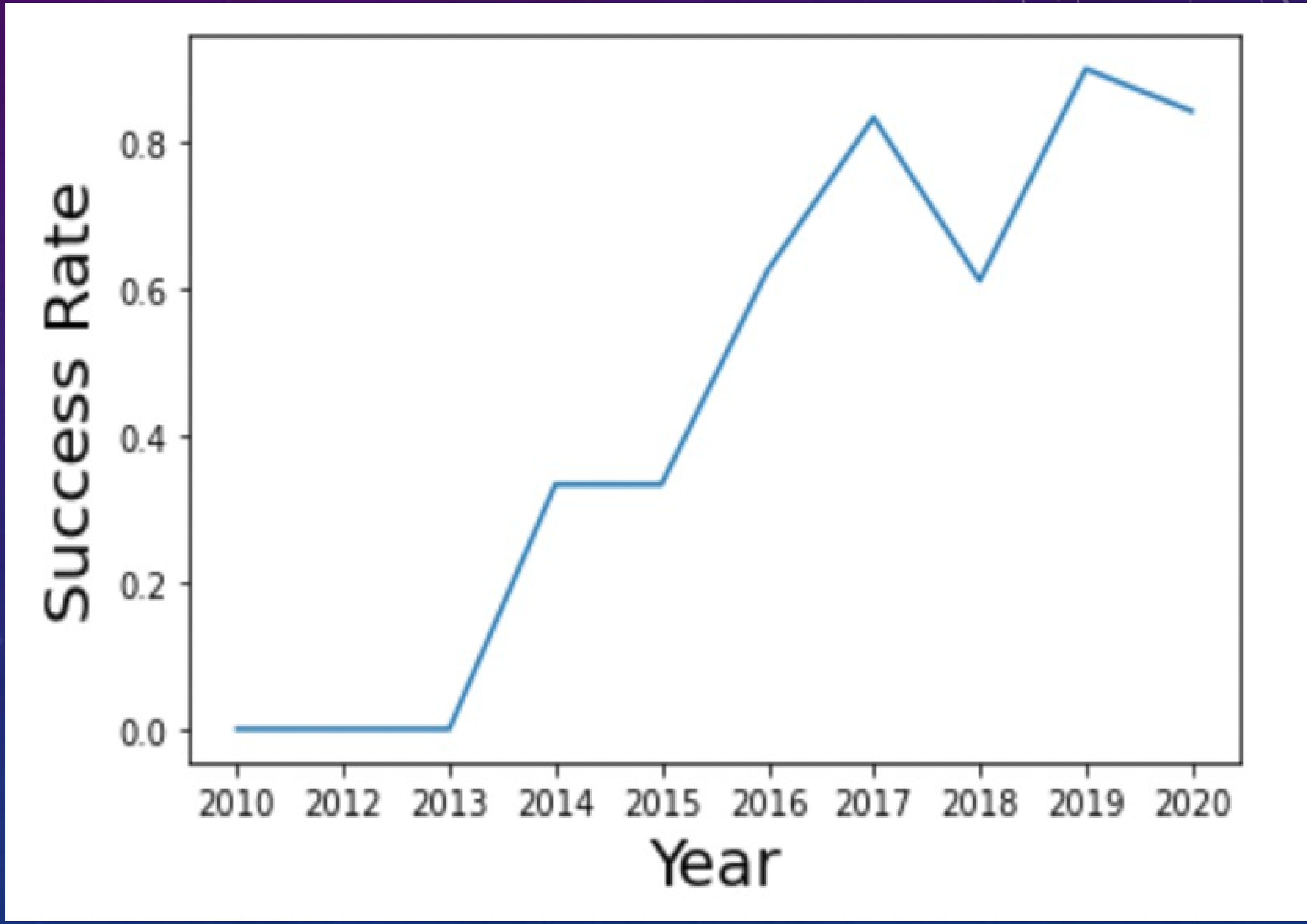
Explanation:

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

LAUNCH SUCCESS YEARLY TREND

Explanation:

- The success rate since 2013 kept increasing till 2020.



ALL LAUNCH SITE NAMES

```
%sql Select Distinc
 * sqlite:///my_data
Done.
Out[13]:
```

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

Explanation:

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

LAUNCH SITE NAMES BEGIN WITH `CCA`

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	0:35:00	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)

Explanation:

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

TOTAL PAYLOAD MASS

Done.

Out[21]:

sum(PAYLOAD_MASS__KG_)
45596

Explanation:

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

AVERAGE PAYLOAD MASS BY F9 V1.1

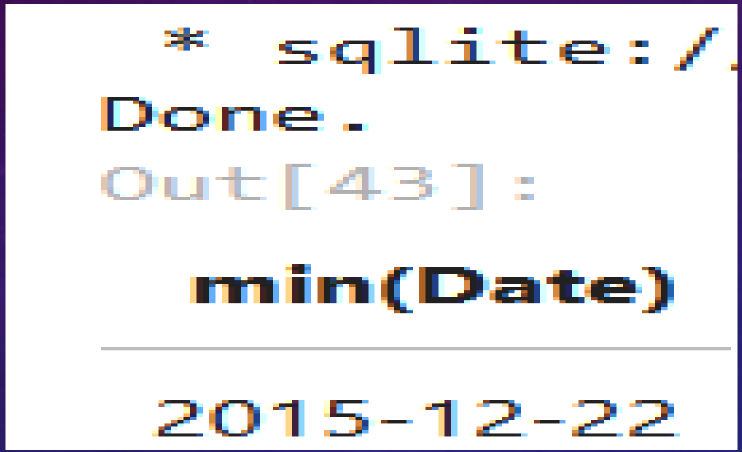
```
%sql select avg(PAYLOAD_MASS__KG_
 * sqlite:///my_data1.db
Done.
Out[28]:
```

avg(PAYLOAD_MASS__KG_)
2928.4

Explanation:

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

FIRST SUCCESSFUL GROUND LANDING DATE



```
 * sqlite:/
Done.
Out[43]:
```

min(Date)
2015-12-22

Explanation:

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

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Explanation:

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

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

```
 * sqlite:///my_data1.db
Done.
Out[62]:
```

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

Explanation:

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

BOOSTERS CARRIED MAXIMUM PAYLOAD

```
 * sqlite:///my_data1.db
Done.
Out[64]:
```

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.

2015 LAUNCH RECORDS

Out[80]:

substr(Date,6,2)	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

Explanation:

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

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

Landing_Outcome	outcome_count
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

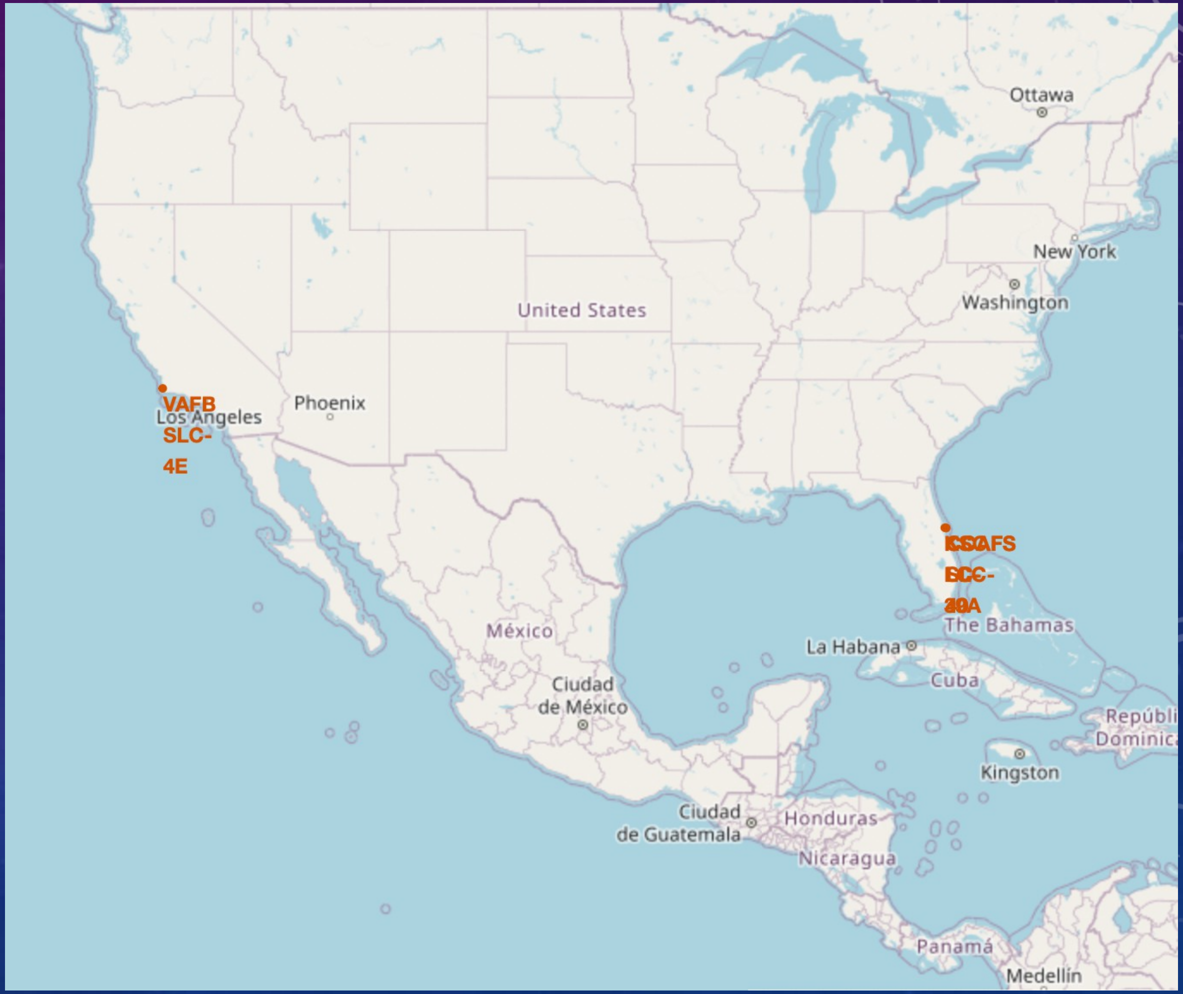
Explanation:

- 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.

ALL LAUNCH SITES' LOCATION MARKERS ON A GLOBAL M

Explanation:

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



COLOUR-LABELED LAUNCH RECORDS ON THE MAP

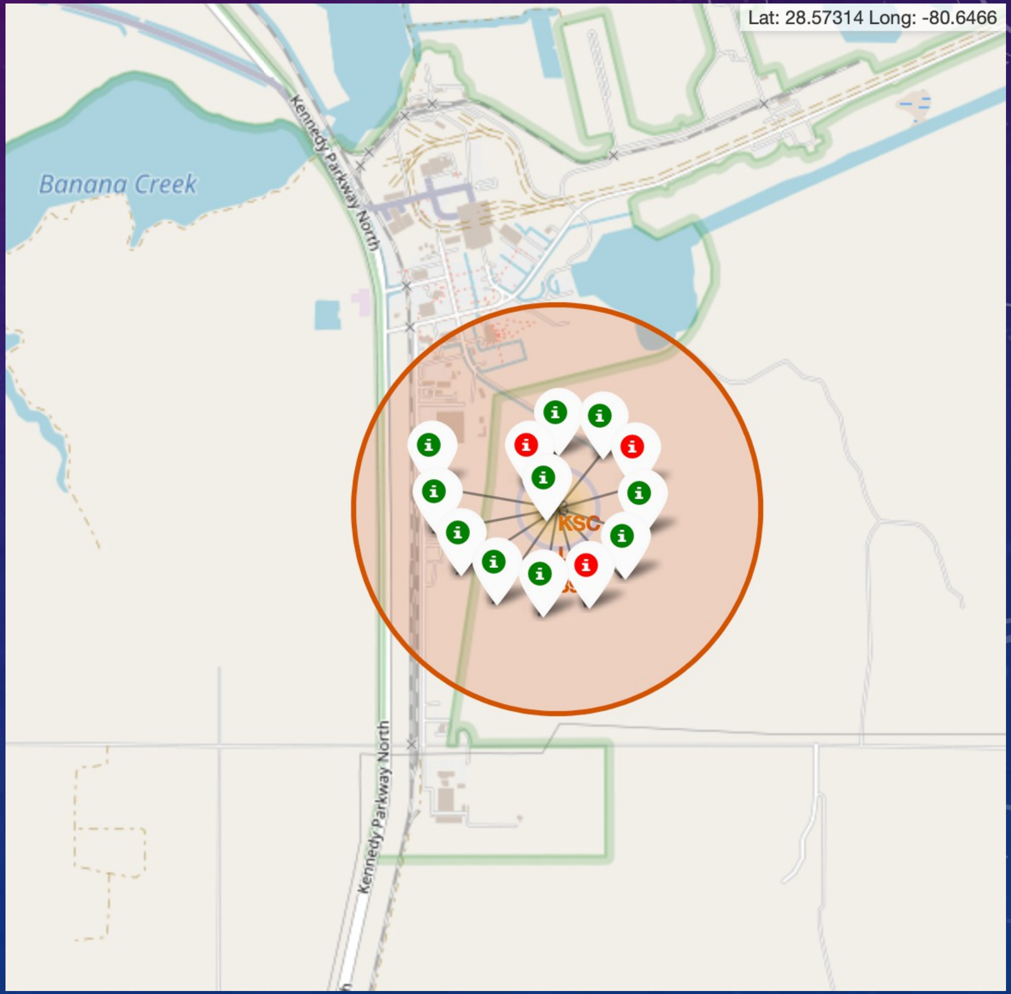
Explanation:

- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

  -Green Marker = Successful Launch

  -Red Marker = Failed Launch

- Launch Site KSC LC-39A has a very high Success Rate.



DISTANCE FROM THE LAUNCH SITE KSCLC-39A TO ITS PROXIMITIES

Explanation:

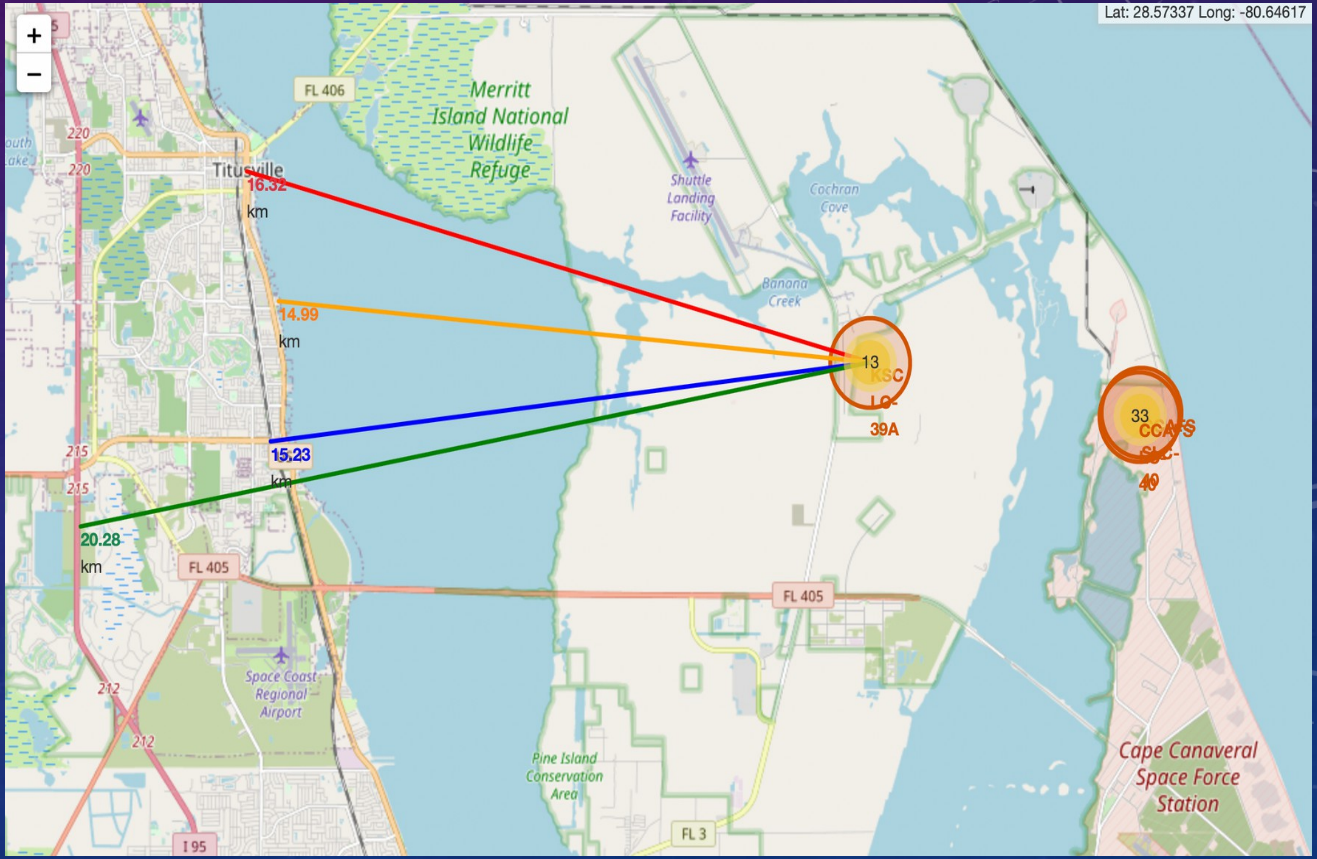
- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

  -relative close to railway (15.23 km)

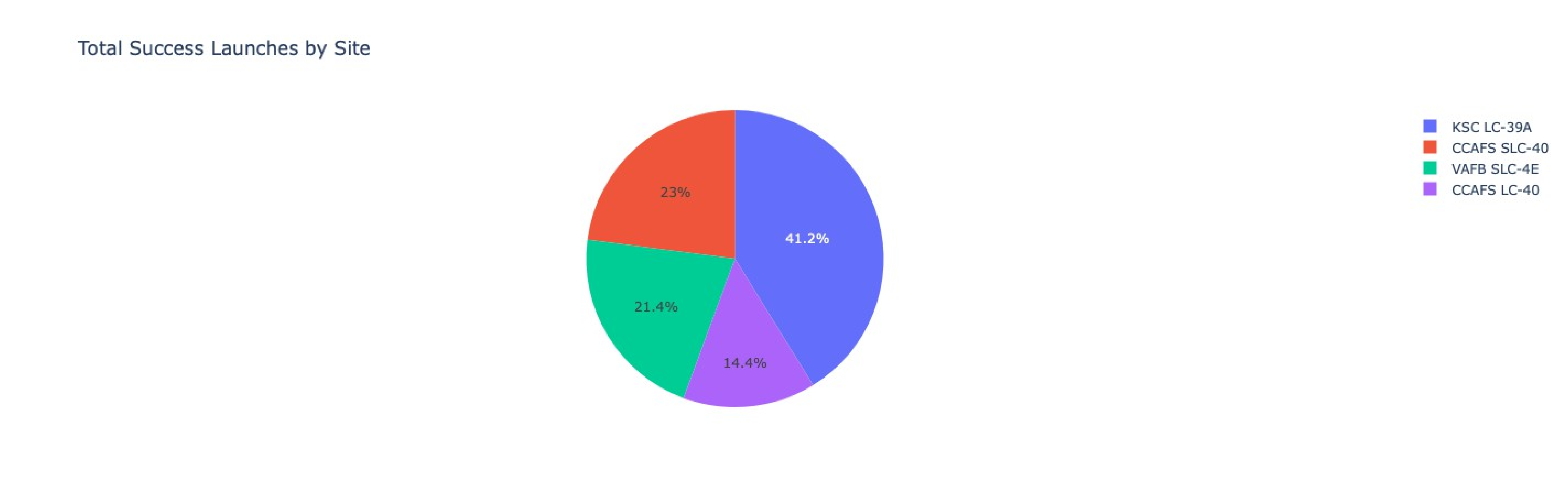
  -relative close to highway (20.28 km)

  -relative close to coastline (14.99 km)

- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



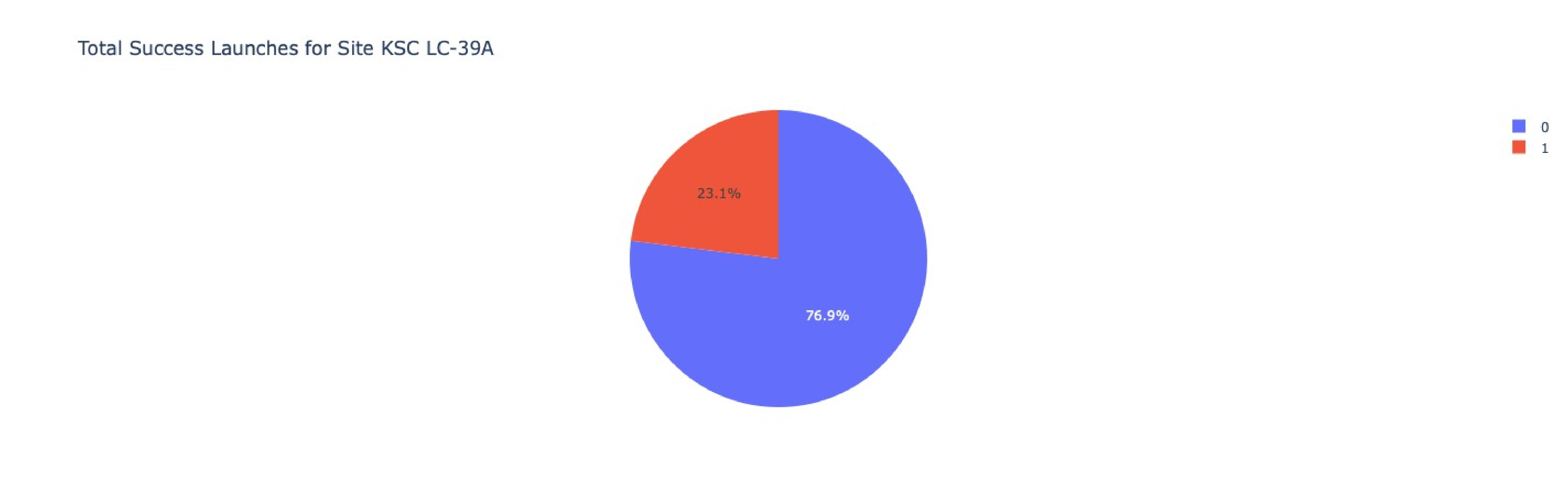
LAUNCH SUCCESS COUNT FOR ALL SITES



Explanation:

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

LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RAT



Explanation:

- 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 SIT

Explanation:

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

CLASSIFICATION ACCURACY

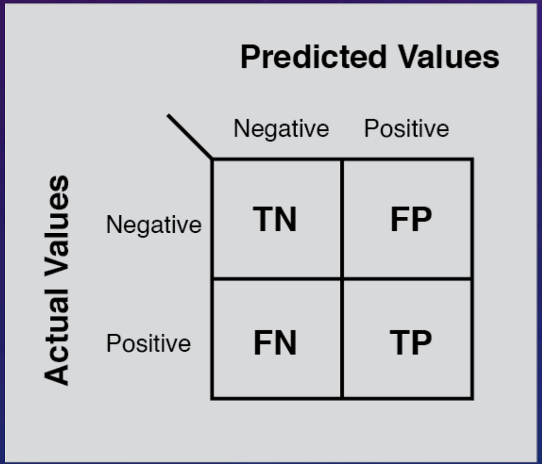
Explanation:

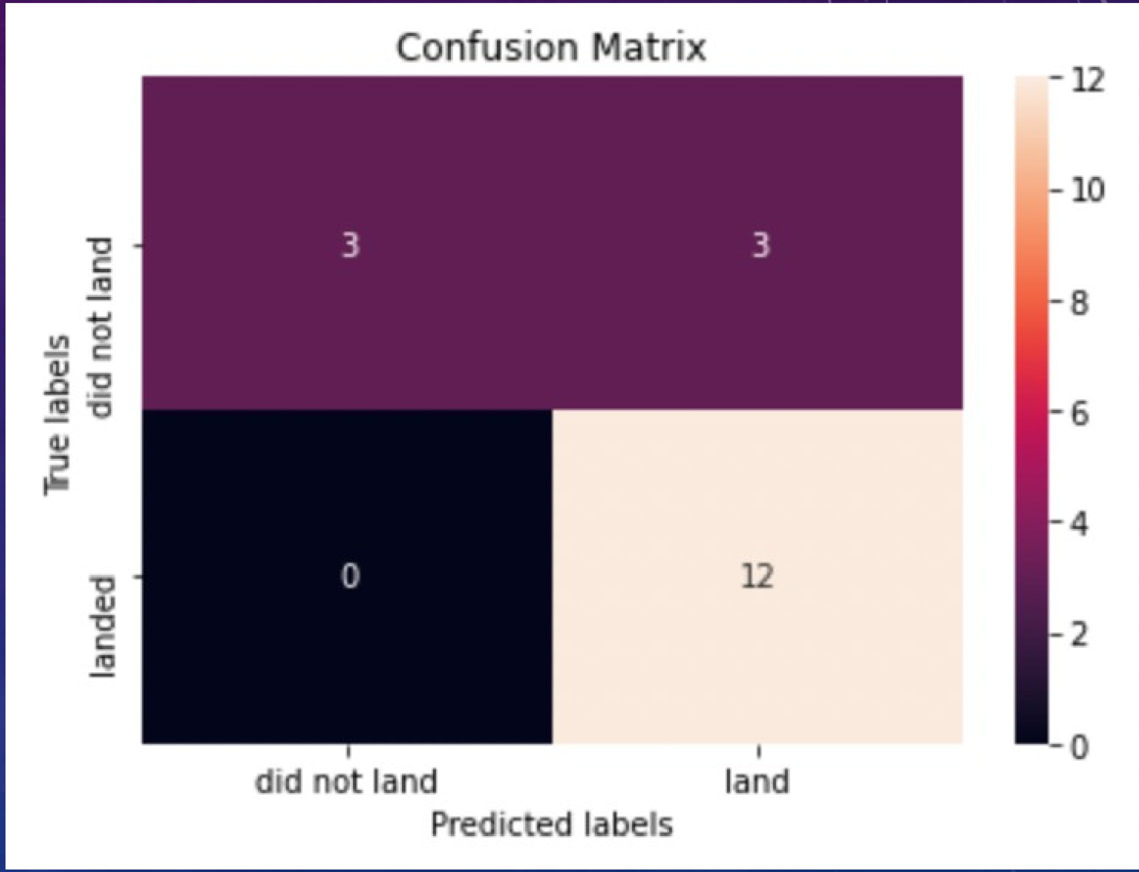
- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.
- Score = 83%

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.





Conclusion

- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.