IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Konstantin Budrin
06.10.2023

Outline

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

Executive Summary

- In this research I use several data analysis and four different predictive models to analyze the public data of SpaceX launch to found some insights about their launch technologies. I use web-scrapping and data wrangling to get data from their public API and process it for using with statistical methods and predictive models. During the research I use different methods to understand more about their launches, landings and cargos deployment on different orbits.
- At the end of my research, I've created a predictive model which can predict the 1st stage landing result with accuracy of 83.3%. Not a fascinating result, but it's also not very bad. Also I've got a lot of useful information about payload mass, booster types and launch results.

Section 1

Methodology

Introduction

In modern world the space technologies is a key to the future. Telekom satellites, orbital observation, space telescopes like Hubble and James Webb – they all our eyes in spaces an even on Earth. We need to create more safe and predictable methods to bring useful cargos on orbits and may be even on other planets. To do that we need to analyze existing space launches experience to solve several questions:

- Which landing pads are most valid to launch spaceships?
- What type of boosters are most safety for different type of orbit?
- Which payload mass limitations are optimal for launches?
- Can we predict the result of landing for 1st stage based on all of this parameters?

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected from official SpaceX site
- Perform data wrangling
    - Data was processed to transform most useful fields to numeric values and exclude or replace anomalies and missing values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

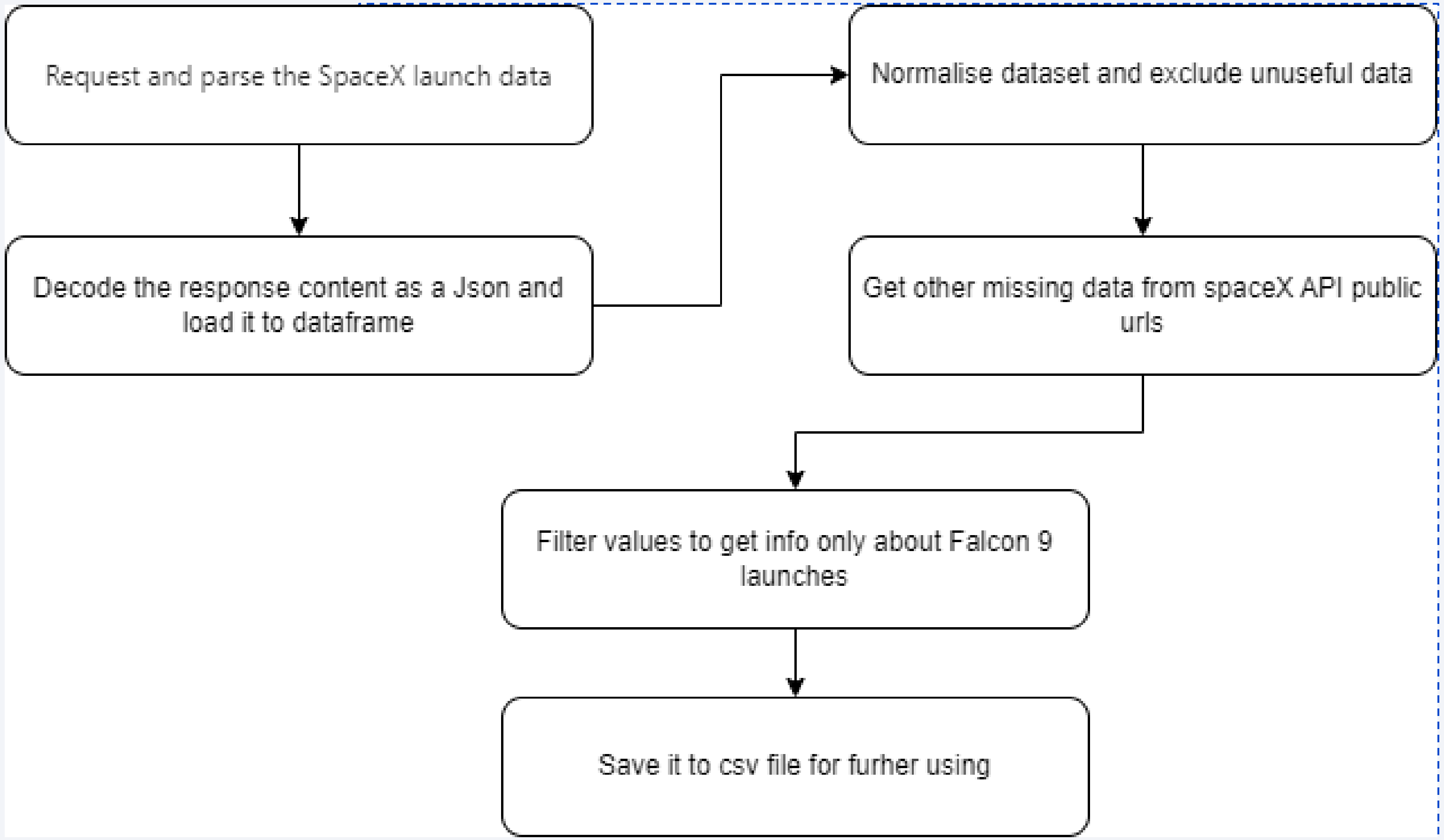
Data Collection

All data sets downloaded from official SpaceX website and their public API

Data Collection – SpaceX API

- Here is the flow chart of Data collection from SpaceX public API

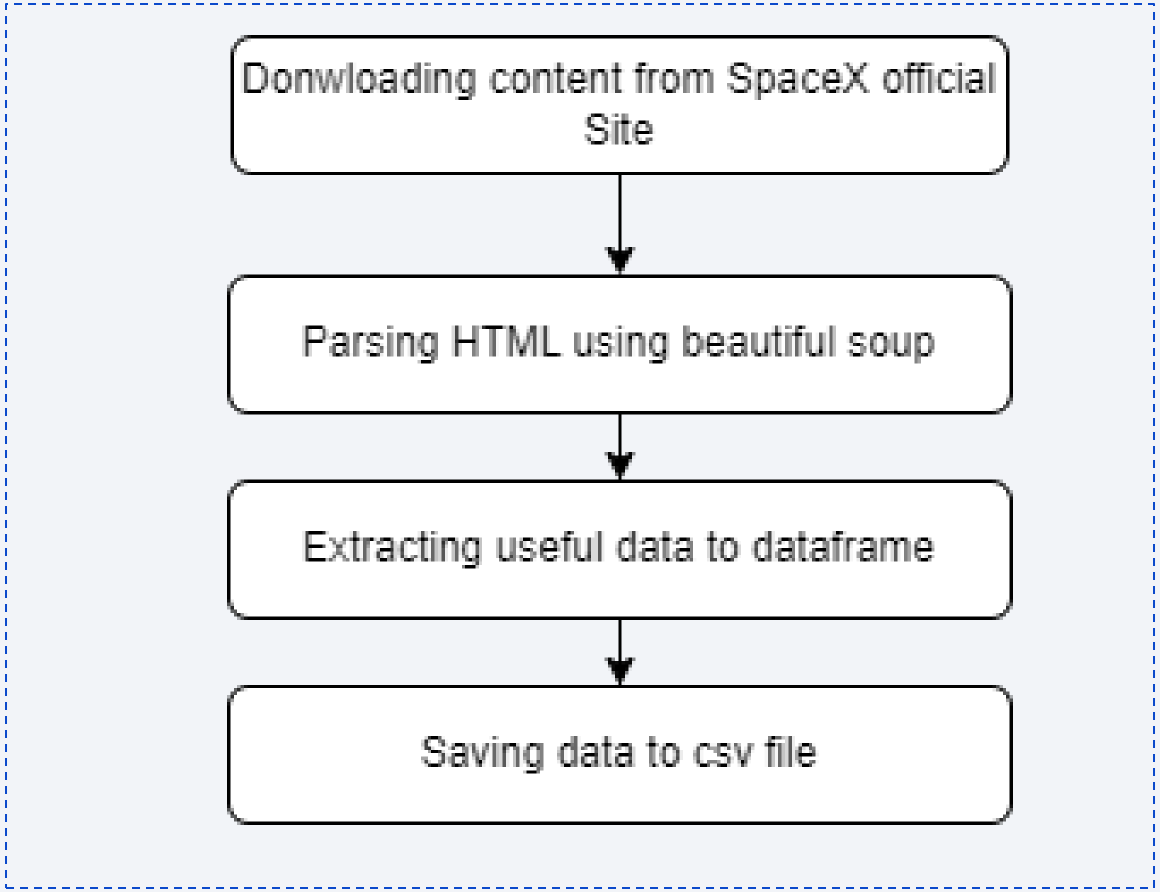
- Github



Data Collection - Scraping

To extract and process data from official SpaceX website we use beautifulsoup python library.

Github



Data Wrangling

After reading the csv file, I've analyzed it to found missing data and replace it. Also I've added a "class" column to dataframe which contains the "Class" of landing outcome based on landing outcome value: 0 for failure landing and 1 for successful landing

Github

EDA with Data Visualization

- I've plotted several charts including:
- Cat plots to show statistics of launches per launch site and payload mass
- Scatter plots to visualize the same in more efficient way
- Bar plot to show success rate on each orbit
- Scatter plot to show payload mass vs orbit statistics and another one to show flight number vs orbit statistics
- Line plot to show changes of success rate per year from 2010 to 2021

Github

EDA with SQL

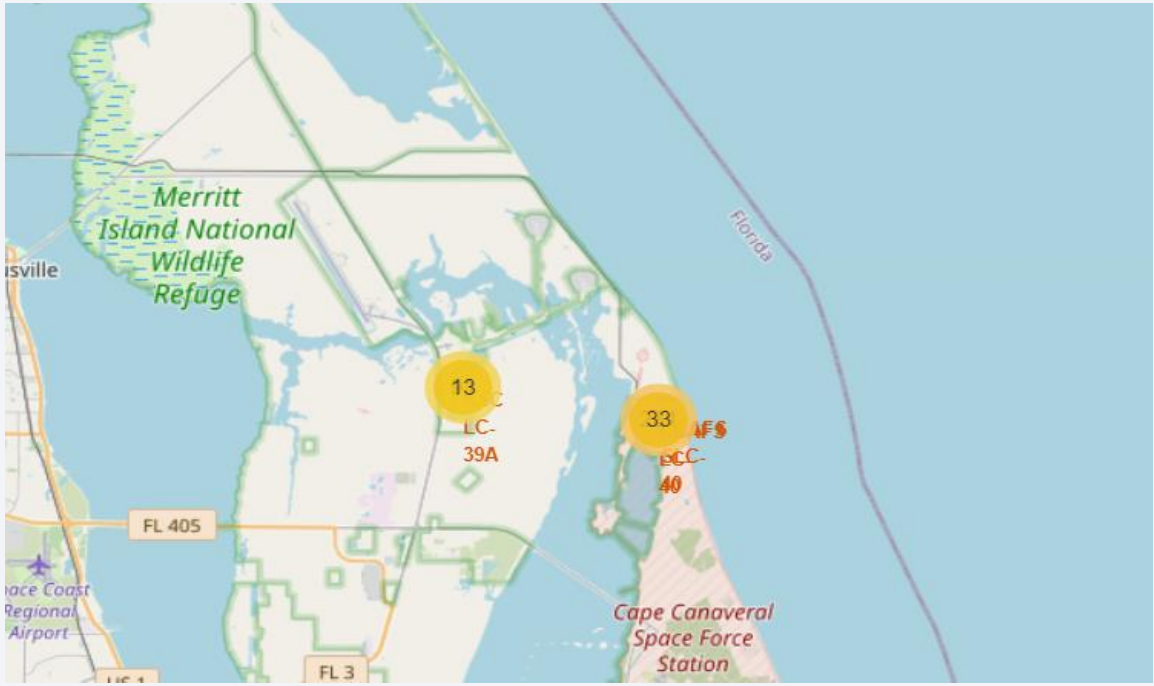
To analyze data I used this sql techniques:

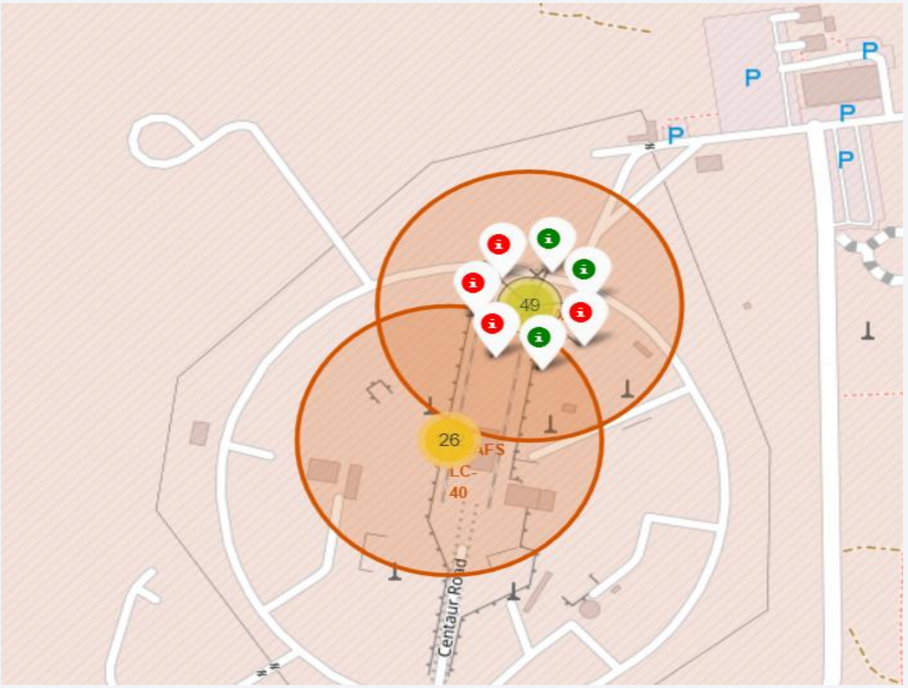
- "Create" query to create the data table and fill it with date from dataframe
- Several different "select" queries with parameters to see lists of booster versions and launch sites
- Several selects with aggerate functions (sum, avg, count) to see some aggregate data about successful/unsuccessful outcomes, average payload mass, etc
- Several "group by" selects to see info about total mission outcomes grouped by success type and landing type.

Github

Build an Interactive Map with Folium

- On a map I've created several markers to visualize launches from different launch sites and a line to show distance between one of the launch site and coastline. You can see full version of document on Github





Build a Dashboard with Plotly Dash

- To better understand statistics of successful/unsuccessful launches per site I've implemented a pie chart diagram which can display total score of success launches for all sites or successful/unsuccessful ratio for each of selected launch sites.
- Also I've implemented a scatter diagram which shows the score of launches by booster version and payload mass, for all sites, or one selected site.
- You can check source on Github

Predictive Analysis (Classification)

I've evaluated grid search on several models to perform best accuracy on test data:

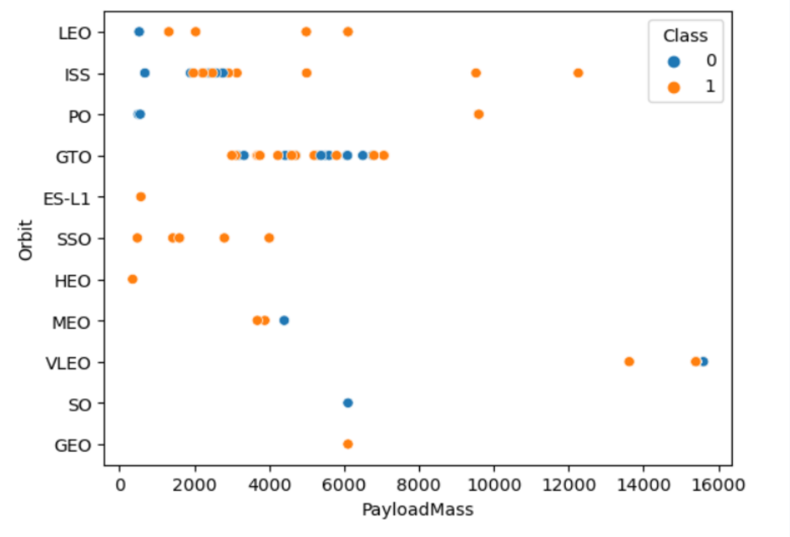
- Logistic regression
- Support Vector Machine
- Decision tree
- K Nearest Neighbours

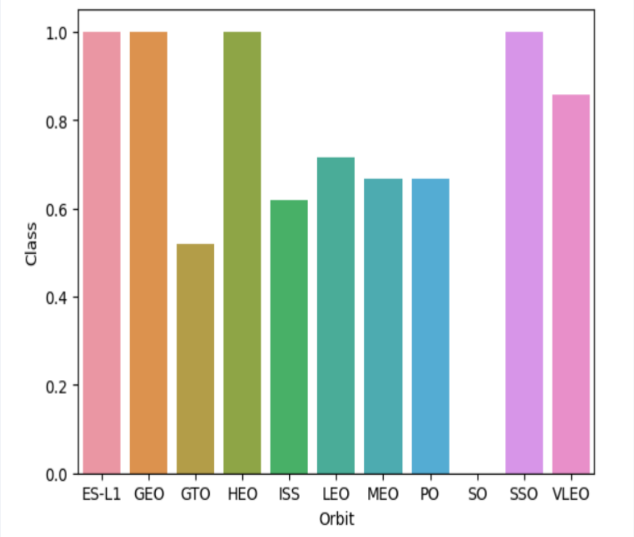
As a result of evaluation I've found that Logistic Regression, SVM and KNN methods have the same accuracy in test data – 83%.

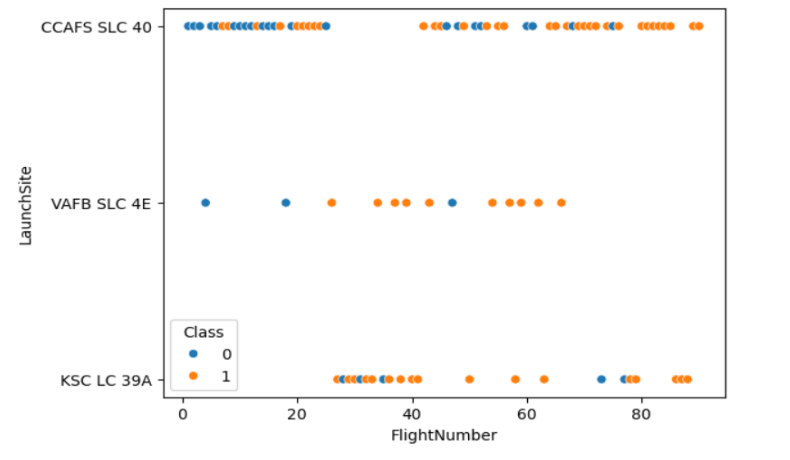
Github

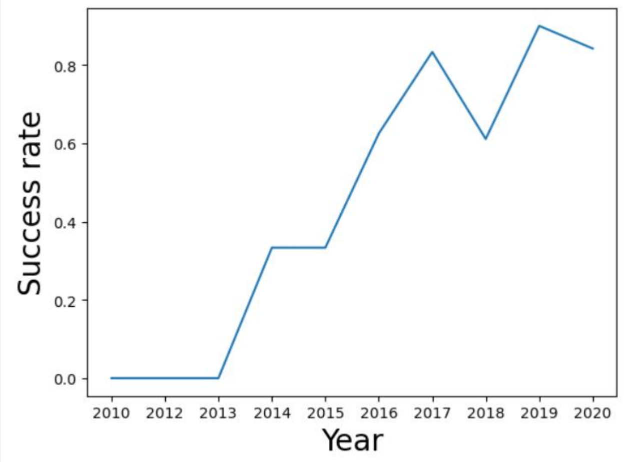
Results

- I've found that most of cargos are bringing on a GTO orbit. Payload Mass are between 2000 kg and 8000 kg
- Most successful launches are going to GEO, HEO, SSO and ES-L1 orbit (100% success score).
- Most failure launches are launching to GTO orbit (~50% success score)
- Total success rate mostly growing from year by year.
- Predictive analysis are quite good to predict success landing (100% score on test data) but not as good on prediction of failures (50% score on test data). More details on slides 41-44







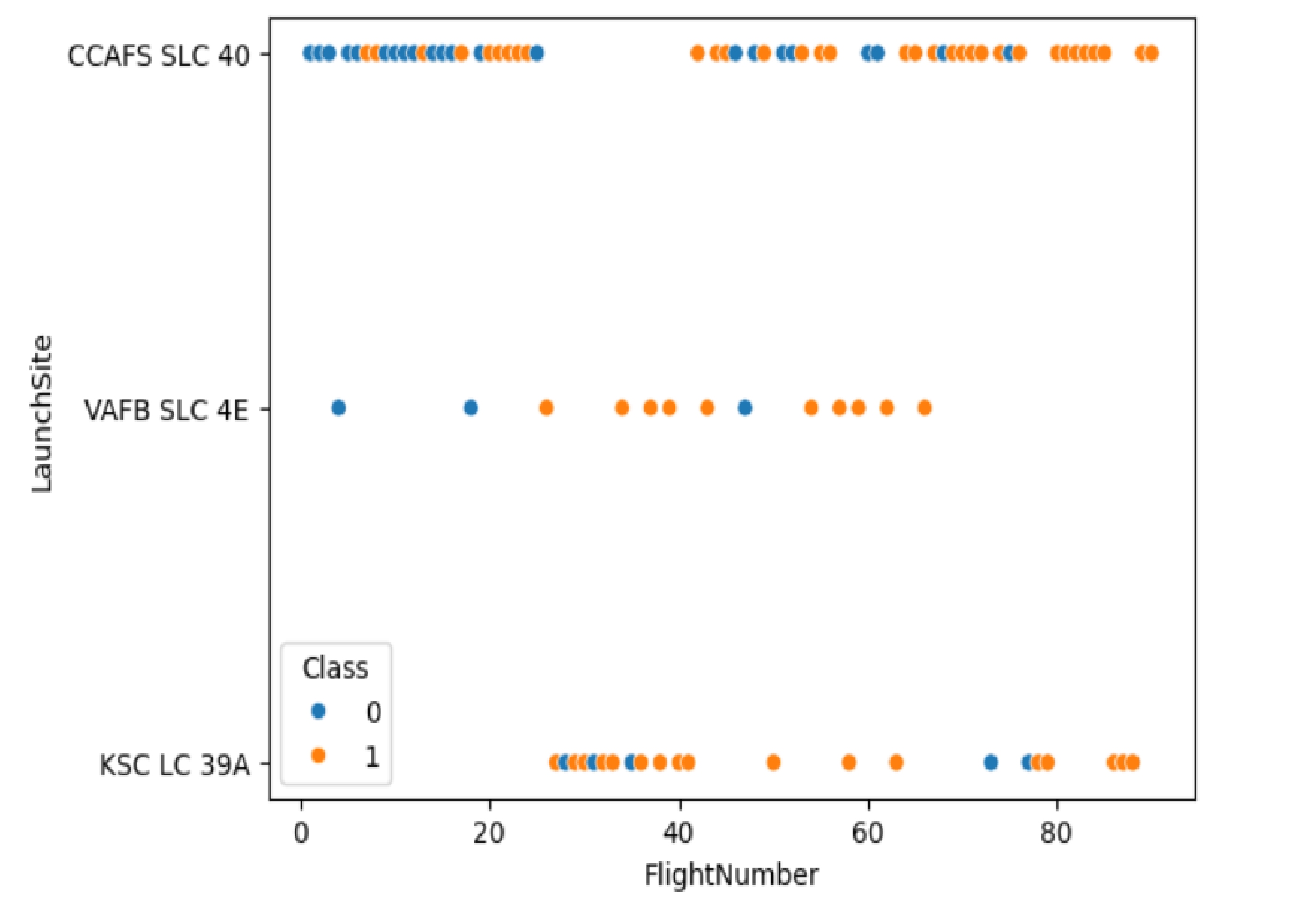


Section 2

Insights drawn from EDA

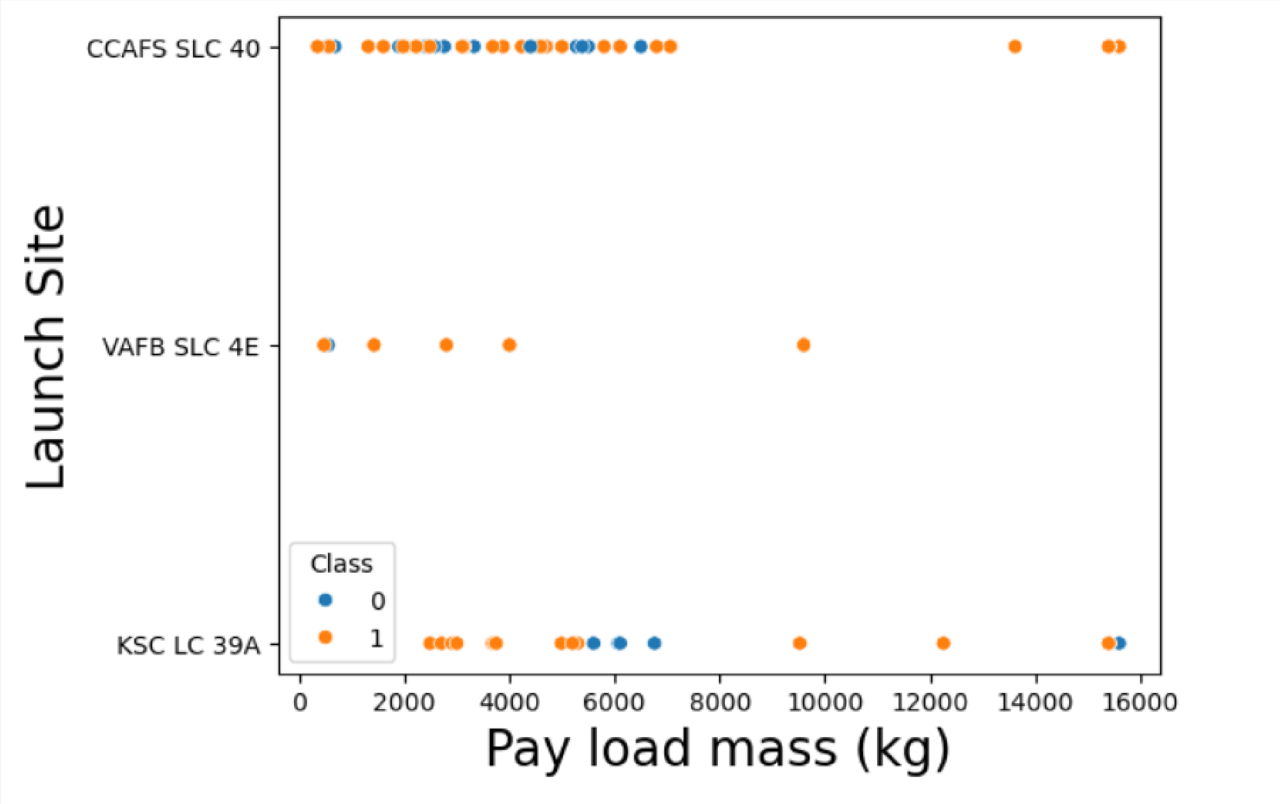
Flight Number vs. Launch Site

- On a screenshot you can see a scatterplot of flight numbers vs launch site. Class means successful/unsuccessful launch. You can see that CCAFS SLC 40 is the most using Launch Site



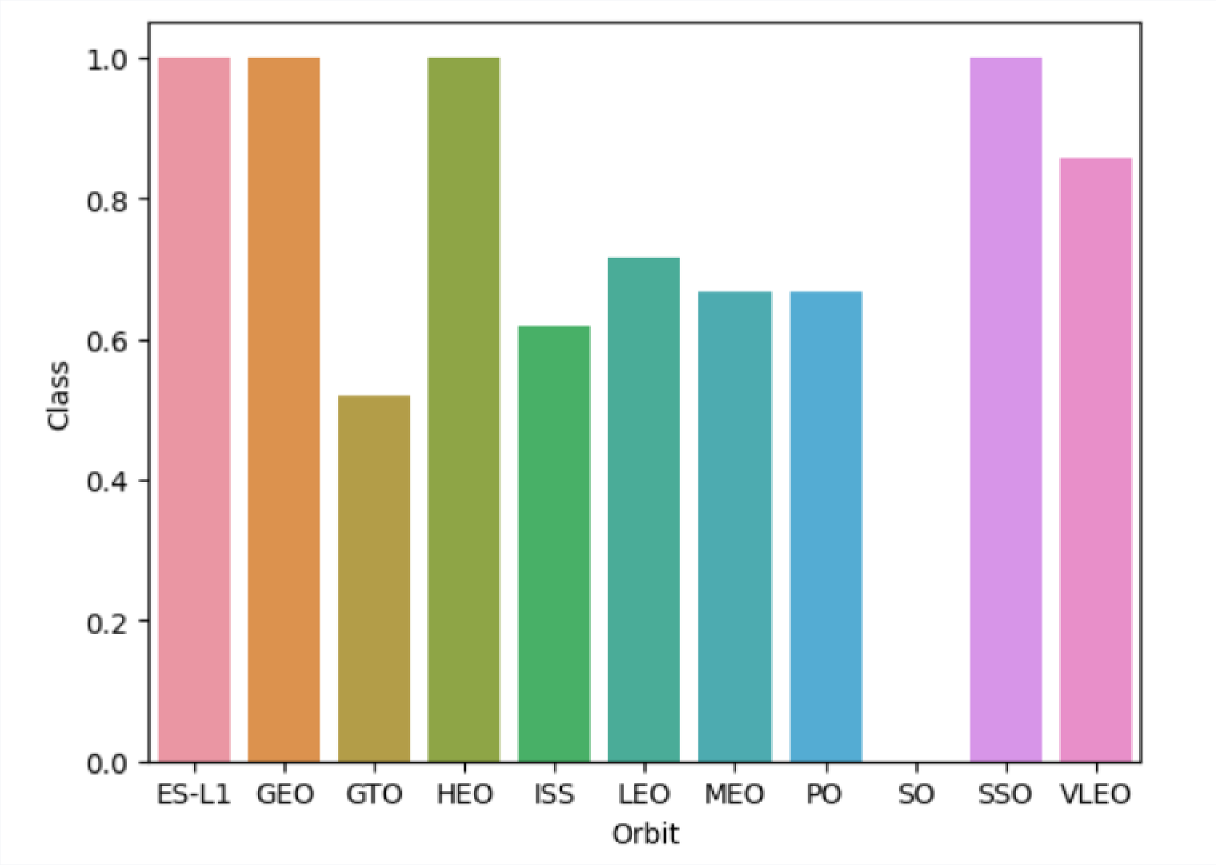
Payload vs. Launch Site

- Here you can see a scatter of payload mass vs launch site. You can see that CCAFS SLC 40 and KSC LC 39A are using to launch most heavy cargos.



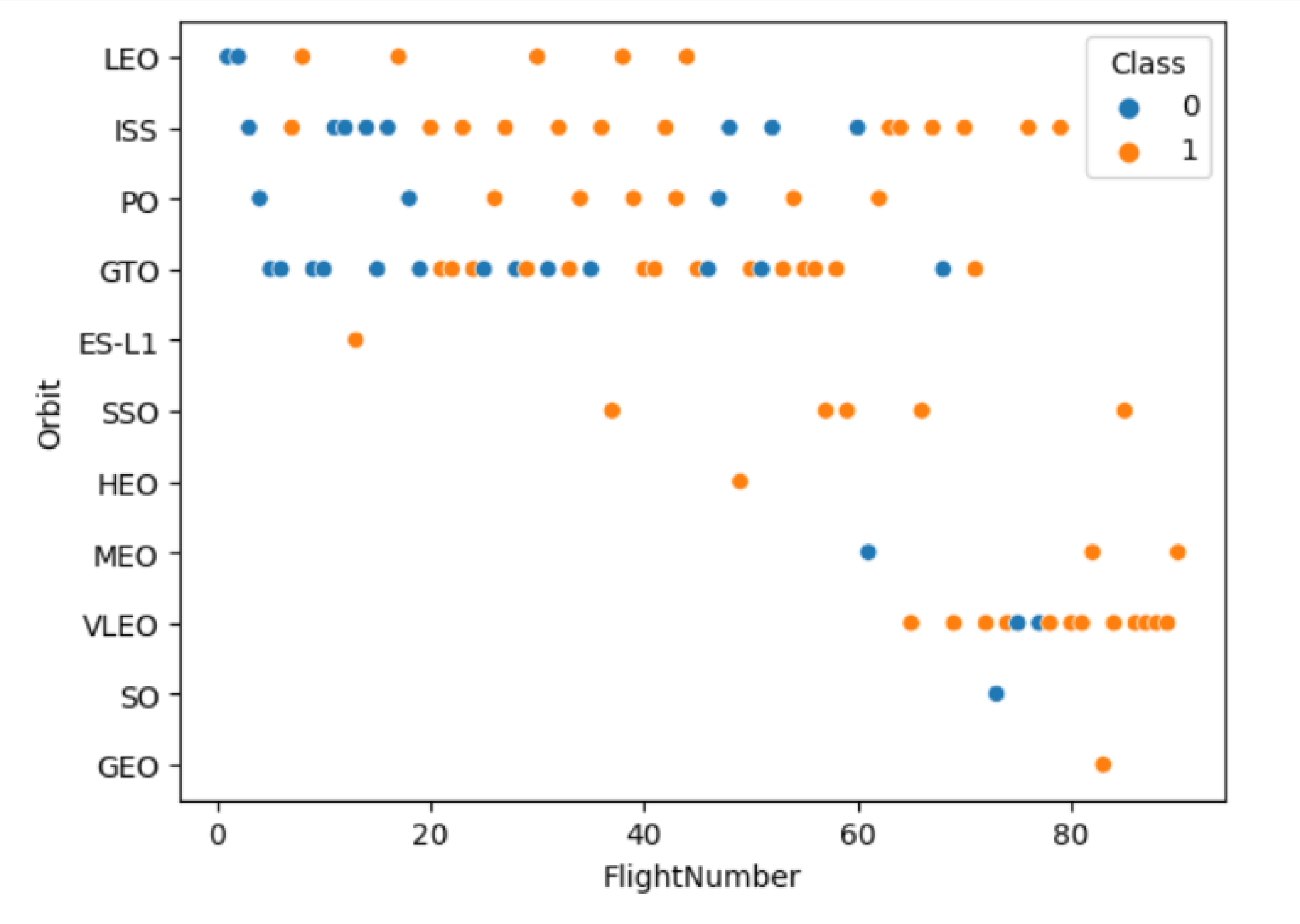
Success Rate vs. Orbit Type

Here you can see the bar chart of successful launch rates for each type of orbit.



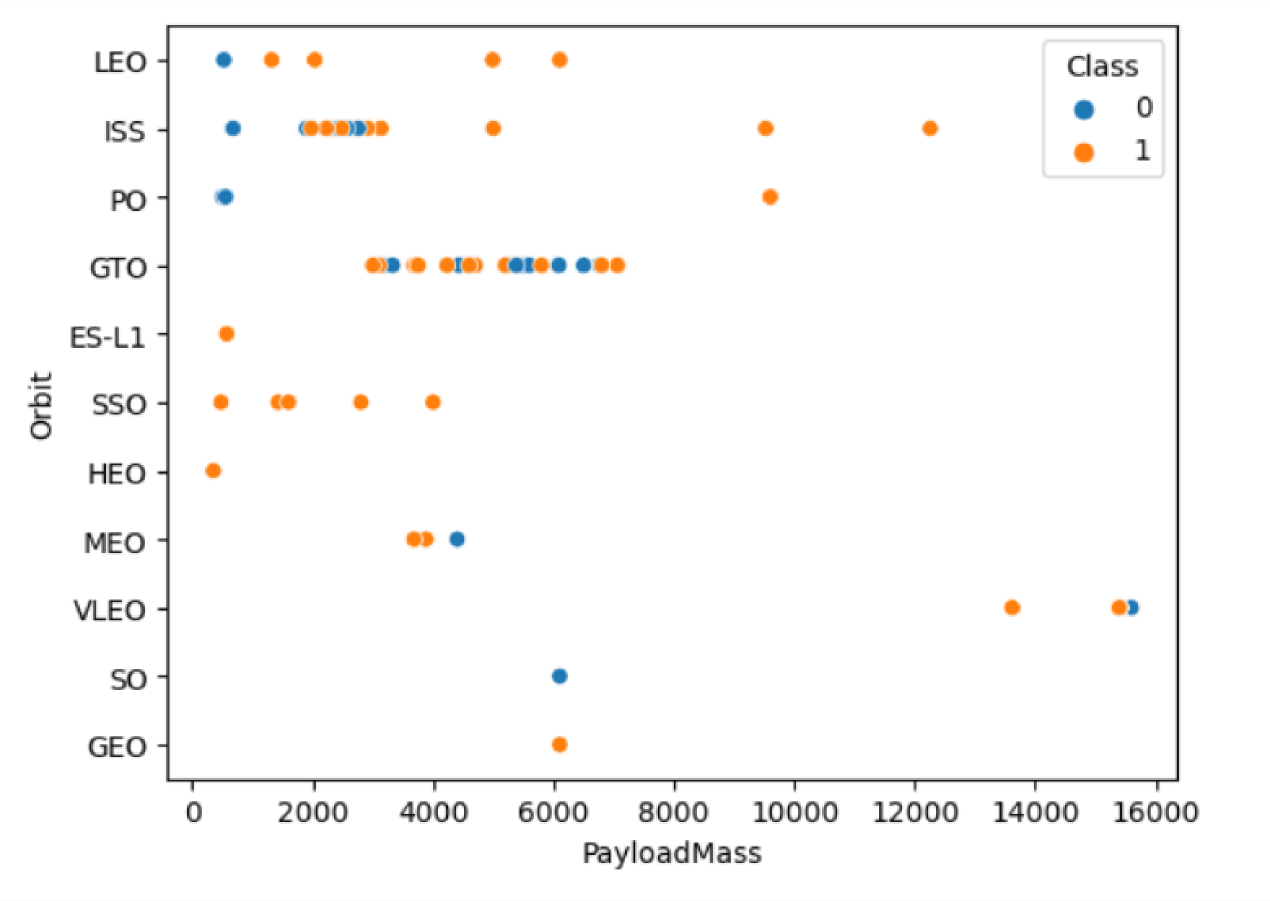
Flight Number vs. Orbit Type

Here you can see a plot which shows the orbit vs flight number statistics. We can see that ISS (international space station orbit), GTO (Geostationary transfer orbit) and LEO (Low Earth orbit) are most often "targets" for SpaceX launches.



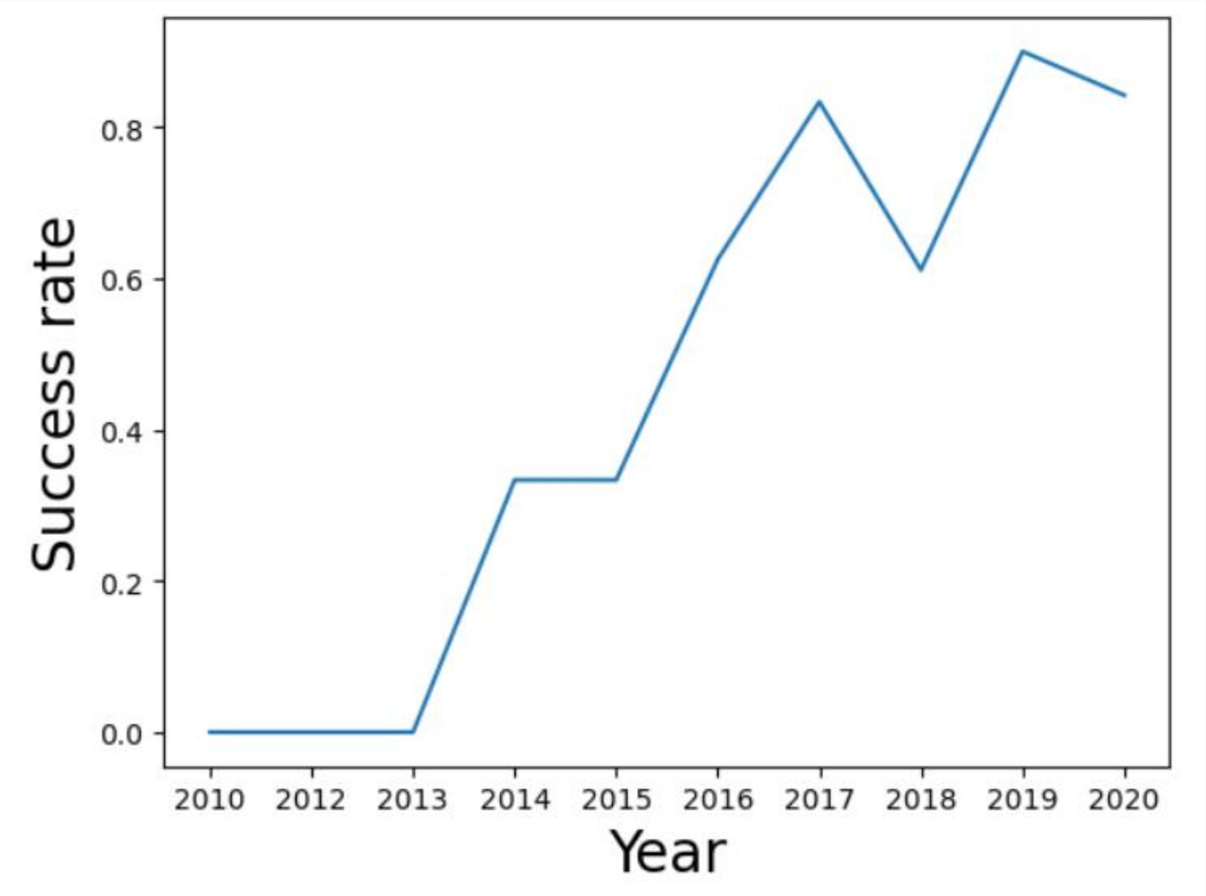
Payload vs. Orbit Type

The "PayloadMass vs Orbit" plot shows that GTO is more "loaded" than any other orbit.



Launch Success Yearly Trend

On this chart you can see that launch successes rate mostly grows from year to year. Just as expected.



All Launch Site Names

Some of SpaceX Launch Sites:

- CCAFS SLC-40 or earlier CCAFS LC-40 - Cape Canaveral Space Launch Complex 40 is a launch pad for rockets located at the north end of Cape Canaveral Space Force Station, Florida. After 2007, the US Air Force leased the complex to SpaceX to launch the Falcon 9 rocket.
- VAFB SLC-4E - Vandenberg Space Force Base Space Launch Complex 4E is a launch and landing site at Vandenberg Space Force Base, California, U.S. It has two pads, both of which are used by SpaceX for Falcon 9, one for launch operations, and other as Landing Zone 4 (LZ-4) for SpaceX landings.
- KSC LC-39A - Kennedy Space Center Launch Complex 39A is the first of Launch Complex 39's three launch pads, located at NASA's Kennedy Space Center in Merritt Island, Florida. The pad, along with Launch Complex 39B, was first designed to accommodate the Saturn V launch vehicle. Typically used to launch NASA's crewed spaceflight missions since the late 1960s, the pad was leased by SpaceX and has been modified to support their launch vehicles.

Launch Site Names Begin with 'CCA'

First five records where launch sites begin with `CCA` are record of launches from CCA LC40. Obviously it's a only launch site which name starts with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total payload mass carried by NASA boosters is 45596

Average Payload Mass by F9 v1.1

Average Payload Mass by F9 v 1.1 is 2534.67 kg

First Successful Ground Landing Date

First Successful Ground Landing Date is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Here is the list of boosters:

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

Total Number of Successful and Failure Mission Outcomes

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

Boosters Carried Maximum Payload

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

- This is the list of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

month	Mission_Outcome	Booster_Version	Launch_Site
10	Success	F9 v1.1 B1012	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40

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

- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

total	Landing_Outcome
21	No attempt
14	Success (drone ship)
9	Success (ground pad)
5	Failure (drone ship)
5	Controlled (ocean)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 3

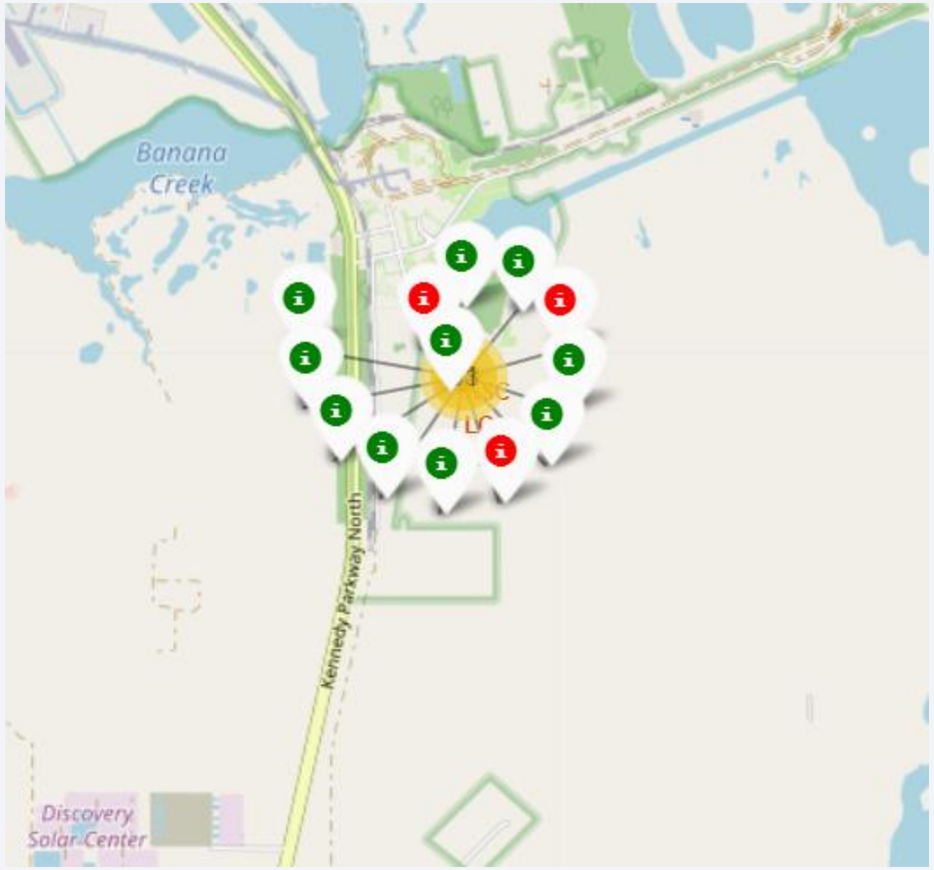
Launch Sites Proximities Analysis

Launch Sites on Map

- As we can see, all sites are located on the coast and as far south as possible. This is not surprising because the closer to the equator, the less fuel is required to start.

Launch markers

- Here on a screenshot you can see red and green launch markers around the KSC LC-39A launch site. Green ones means successful launches, while red ones means failure.

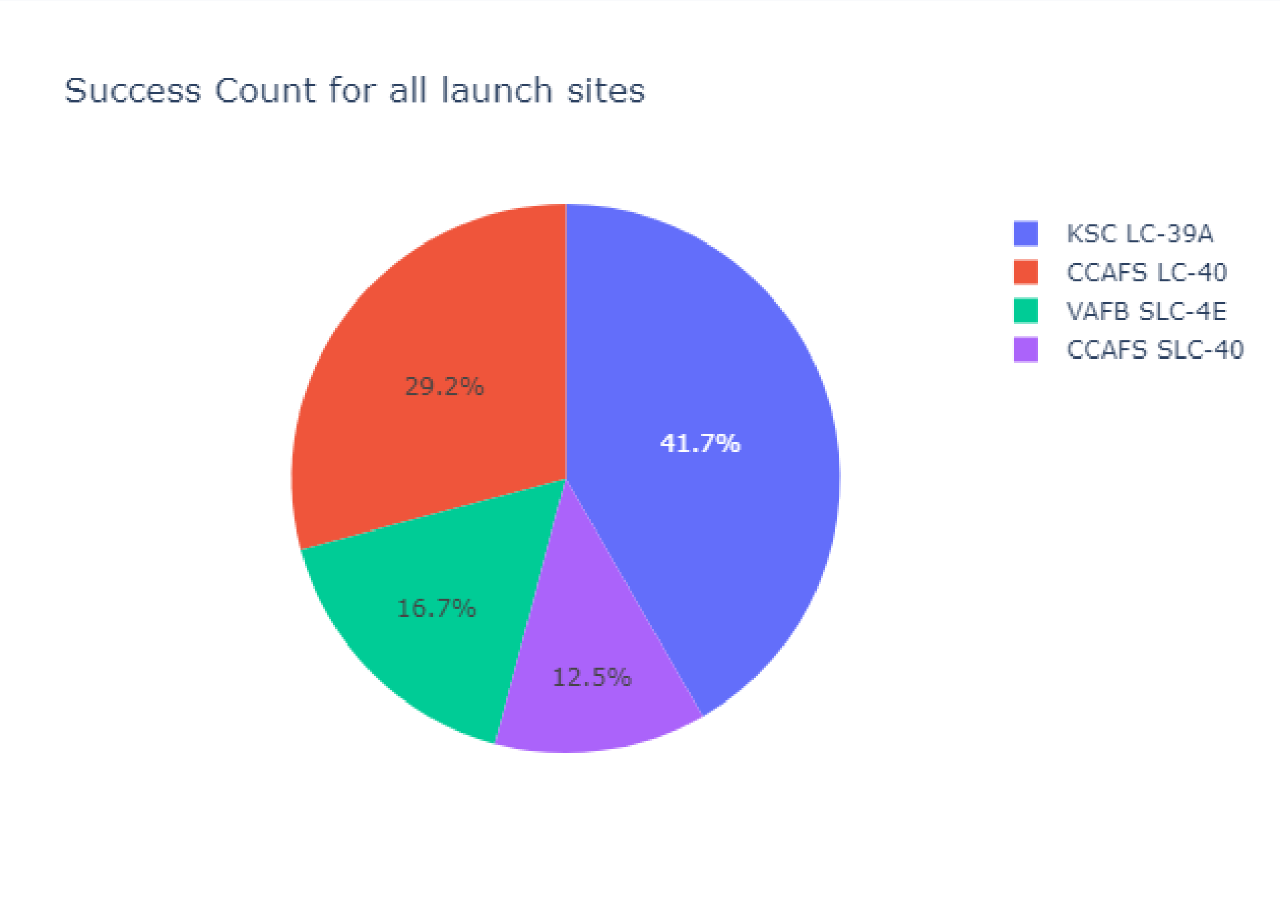


Section 4

Build a Dashboard with Plotly Dash

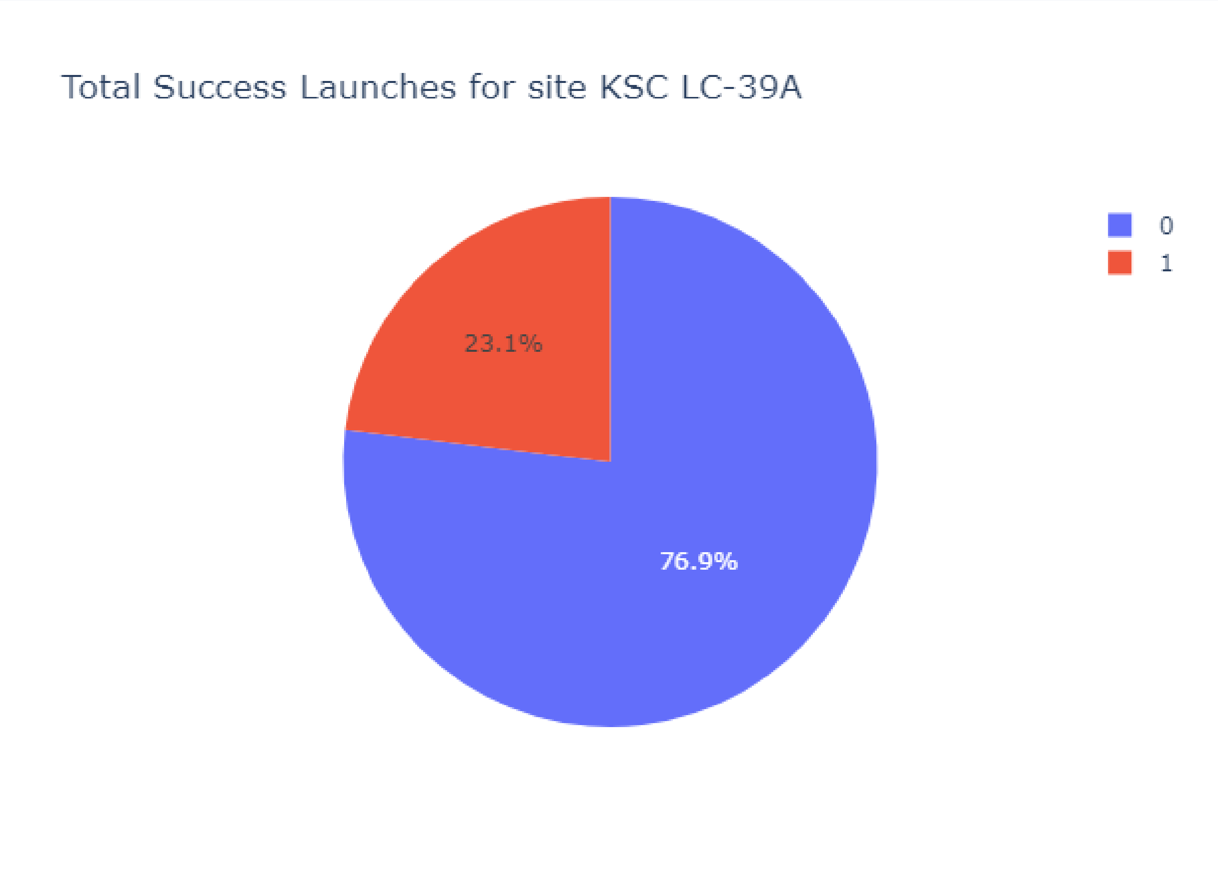
Success Count for all launch sites

- Here is the pie chart which displays a distribution of success launches between launch sites. Each sector displays percents of successful launches from one launch site in total success launches count.



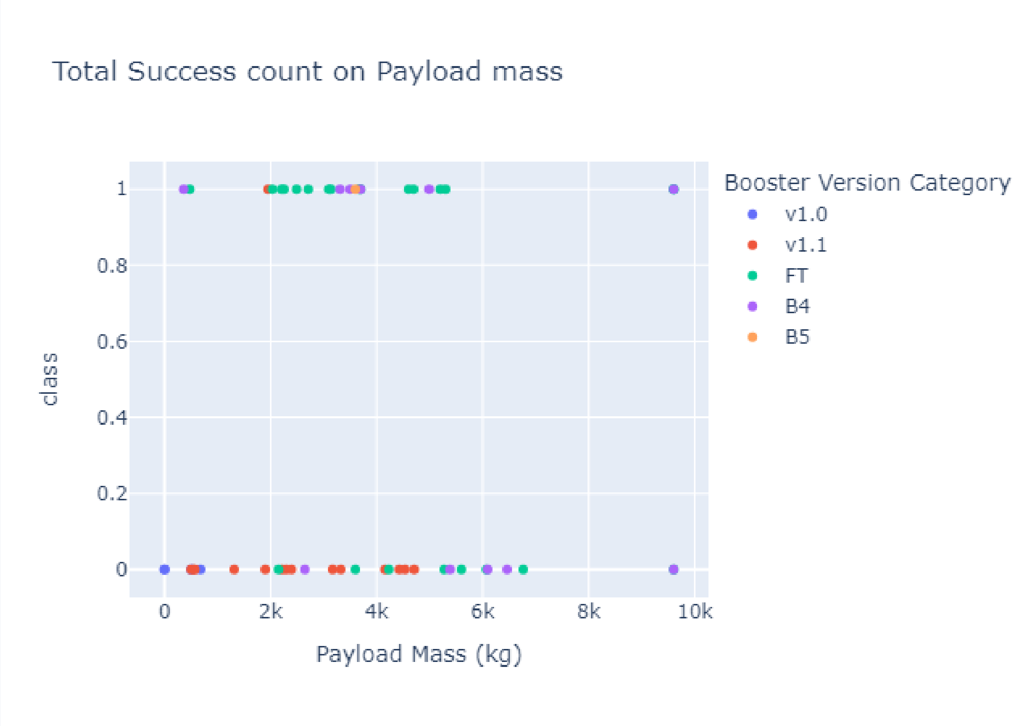
Total Success launches for KSC LC-39A

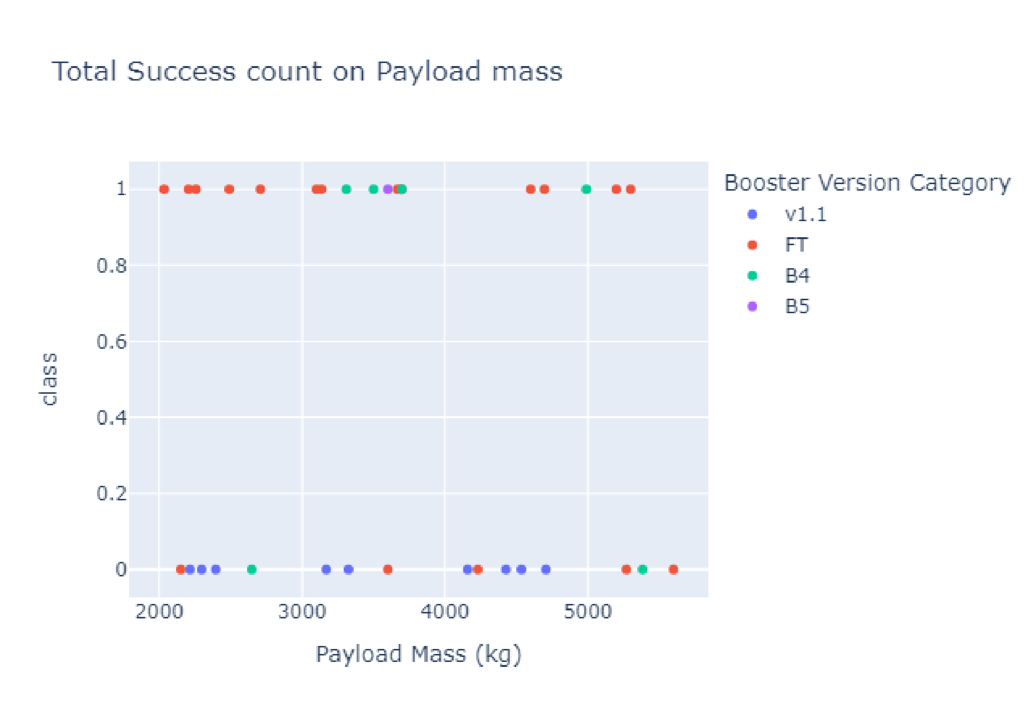
- As we can see, KSC LC-39A has the most successful launches, but it also has the worst success/failure ratio (here is "0" value means failure, when "1" means success)



Payload and Launch Outcome

- Here is the charts of Launch Outcome depending on Payload mass. As before, class 1 means successful launch, when 0 means failure. You can see that most launches brings payload between 2000 kg and 6000 kg. Also we can see that launches with payload mass between 3000 kg and 4000 kg are usually more successful than others.



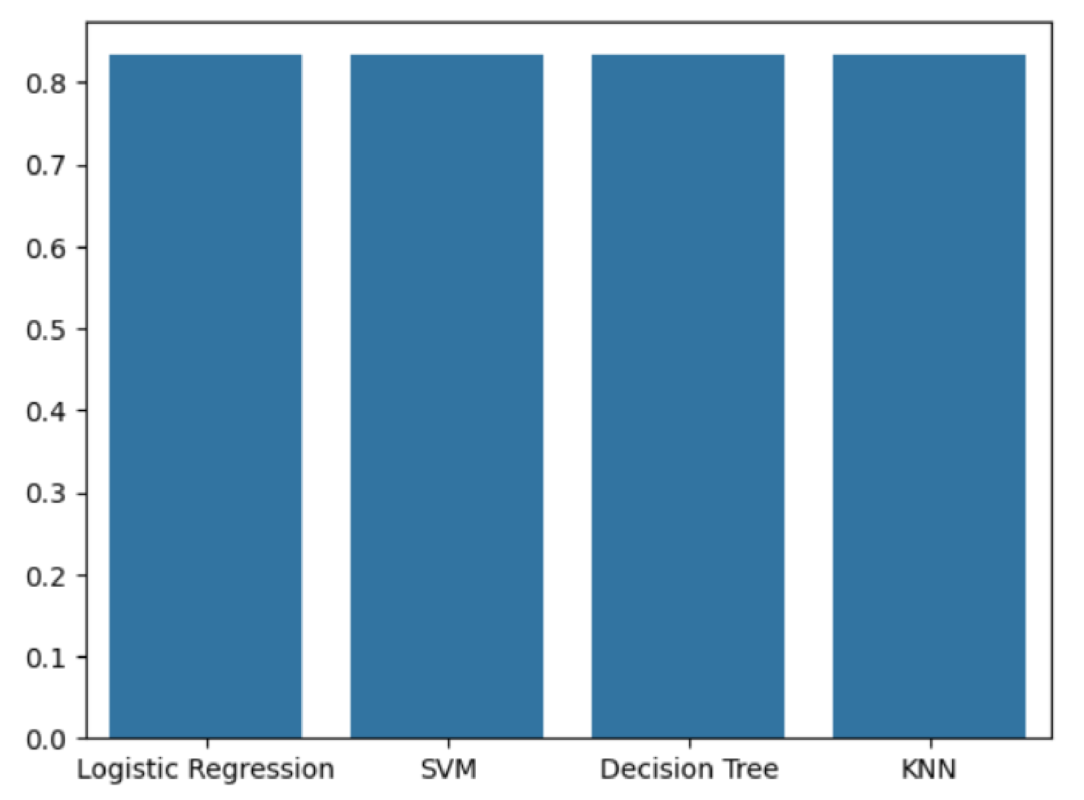


Section 5

Predictive Analysis (Classification)

Classification Accuracy

Here you can see the bar chart of testing accuracy for four different classification models: logistic regression, support vector machine, decision tree and k-nearest neighbors model. As you can see, the accuracy score of all four models after training are completely the same – 83.3%

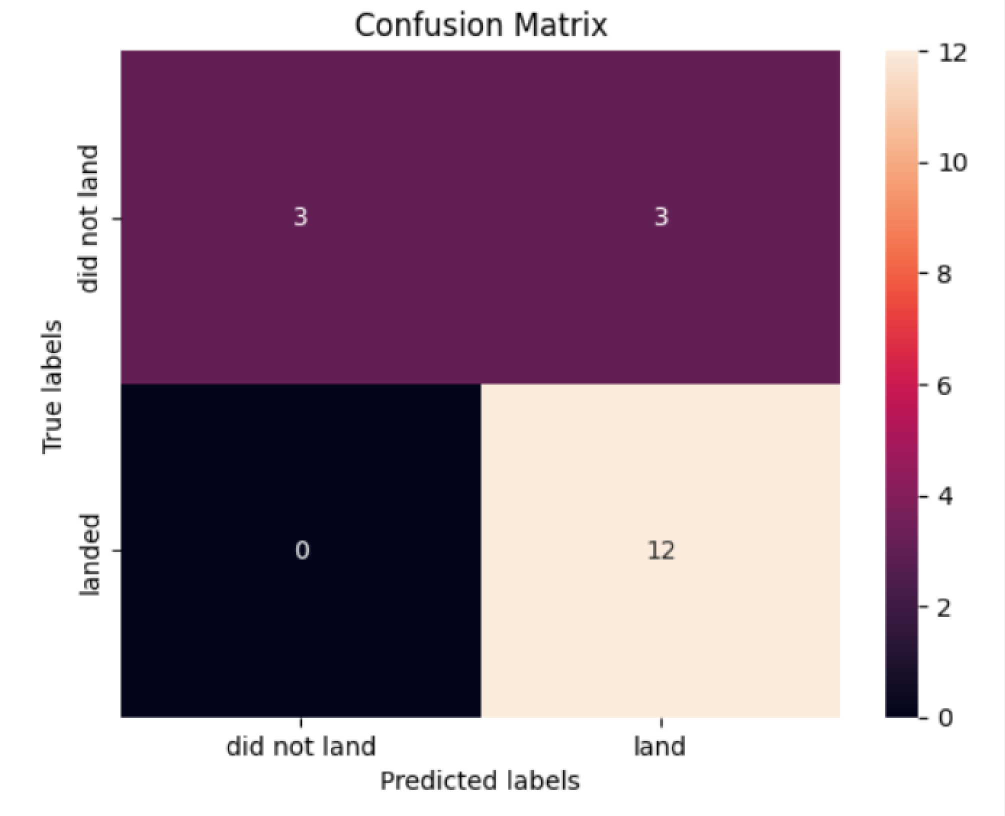


Conclusions

- Using sql data analysis methods I've found important values about SpaceX launches, such as average payload mass, first successful launch date and others.
- With visual analysis we can see that the number of successful launches is mostly growing year by year
- We also can say that payload mass usually ranged between 2000 kg and 6000 kg, and most successful launches has payload mass between 4000 kg and 5000 kg
- We also can predict an outcome landing result using one of four predictive models with accuracy of 83.3% which is not a 100% but still quite high

Confusion Matrix

This is the confusion matrix for all of four classification models we mean above. They all have the same accuracy and they also have the same confusion matrix. You can see, that models can predict successful landing with 100%-accuracy, but have only 50%-accuracy on fail landing prediction.



Appendix

- All files of that project can be find in my github

Developed for IBM Data Science Certification program

Thank you!