Data Science Capstone Project

GIRIJA O K

17-11-2024

IBM Developer
SKILLS NETWORK

Outline

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

IBM Developer
SKILLS NETWORK

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

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

IBM Developer
SKILLS NETWORK

Introduction

Project Background and Context:

SpaceX has revolutionized space travel by making it more affordable, with its Falcon 9 rocket launches priced at $62 million—significantly cheaper than other providers who charge over $165 million. This cost reduction is largely due to SpaceX's ability to reuse the first stage of the rocket. Predicting whether the first stage will land successfully is critical for estimating the cost of a launch. By using publicly available data and machine learning models, our goal is to predict the likelihood of a successful first stage landing.

Questions to be Answered:

- How do variables like payload mass, launch site, number of flights, and orbits impact the success of the first stage landing?
- Has the rate of successful landings improved over the years?
- Which machine learning algorithm is best for predicting the success of the first stage landing?

IBM Developer
SKILLS NETWORK

Methodology

Data Collection:

- We gathered data using the SpaceX REST API and web scraping from Wikipedia for additional information about launches.

Data Wrangling:

- **Filtering:** We cleaned the data by keeping only relevant entries.
- **Handling Missing Values:** We filled in or removed missing data to ensure accuracy.
- **One-Hot Encoding:** We converted categories (like launch sites) into numbers for our model.

Exploratory Data Analysis (EDA):

- We used visualizations (graphs and charts) to explore patterns in the data.
- SQL queries were used to summarize key statistics.

Interactive Visual Analytics:

- **Folium:** We mapped launch sites to see their locations.
- **Plotly Dash:** We created interactive dashboards for exploring the data and results.

Predictive Analysis:

- We built and tested several classification models to predict the success of the first stage landing, fine-tuning them to get the best results.

IBM Developer SKILLS NETWORK

Methodology

IBM Developer
SKILLS NETWORK

Data collection

We collected data using two methods:

1. **SpaceX REST API**:
   - We used the API to get information such as:
     - Flight number, date, booster version, payload mass, orbit, launch site, outcome, and more.
2. **Wikipedia Web Scraping**:
   - We scraped data from the SpaceX Wikipedia page, which included details like:
     - Flight number, launch site, payload, customer, outcome, and time.

By using both methods, we were able to gather complete data for a more detailed analysis.

Data collection- wrangling, and formatting

SpaceX API

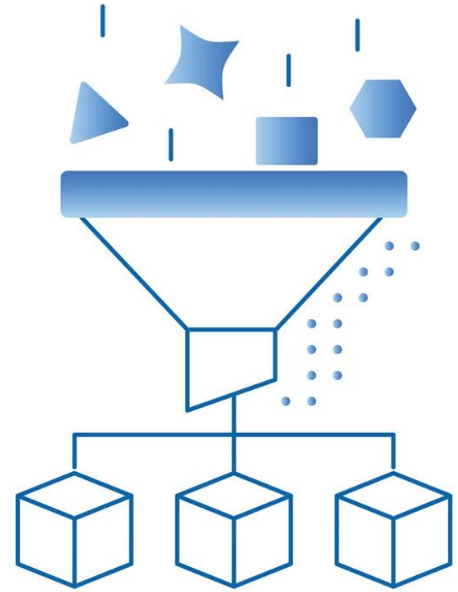
·The API used is https://api.spacexdata.com/v4/rockets/.

·The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.

·Every missing value in the data is replaced the mean the column that the missing value belongs to.

·We end up with 90 rows or instances and 17 columns or features. The picture below shows the first few rows of the data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857



IBM Developer SKILLS NETWORK

Data collection- wrangling, and formatting

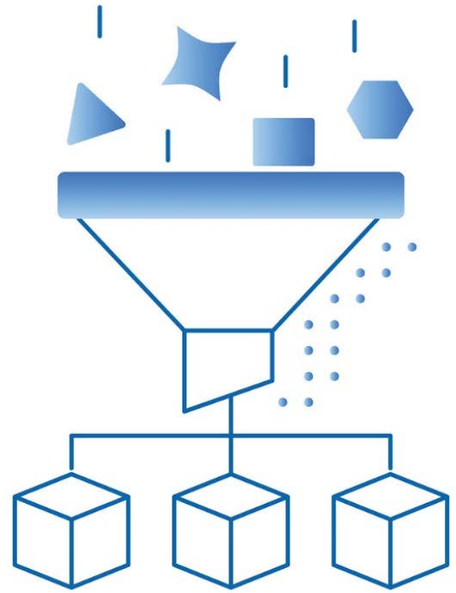
Web scraping

The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

The website contains only the data about Falcon 9 launches.

We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10



Exploratory Data Analysis (EDA)

Pandas and NumPy

·Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:

·The number of launches on each launch site

·The number of occurrence of each orbit

·The number and occurrence of each mission outcome

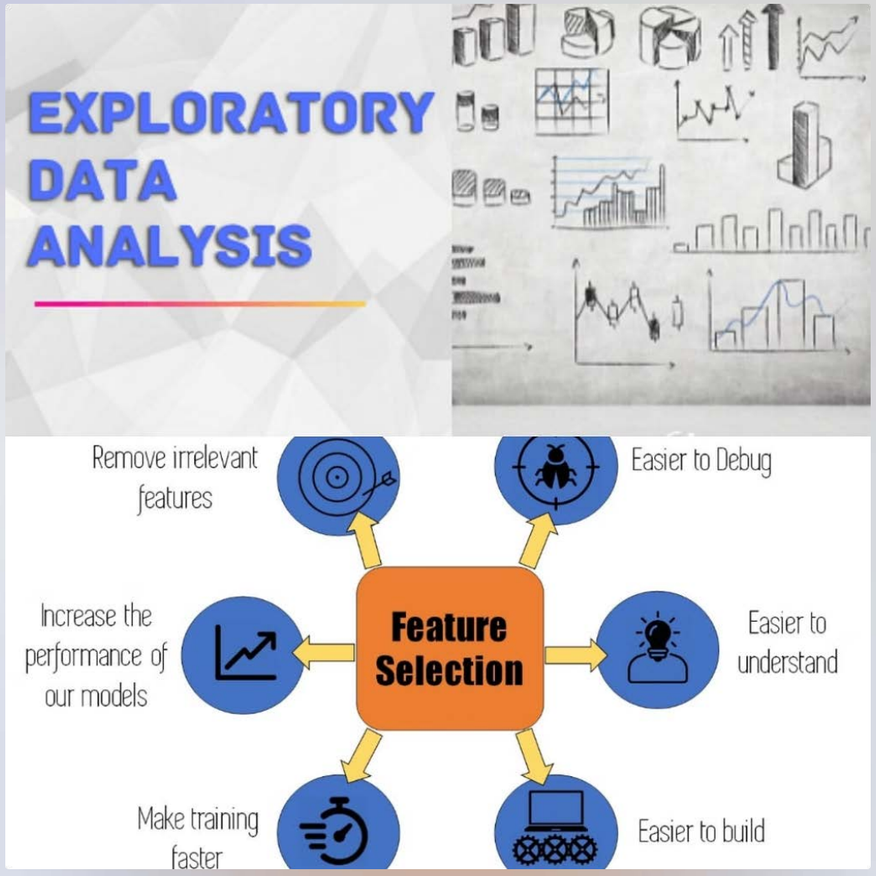
SQL

·The data is queried using SQL to answer several questions about the data such as:

·The names of the unique launch sites in the space mission

·The total payload mass carried by boosters launched by NASA (CRS)

·The average payload mass carried by booster version F9 v1.1



Data Visualization

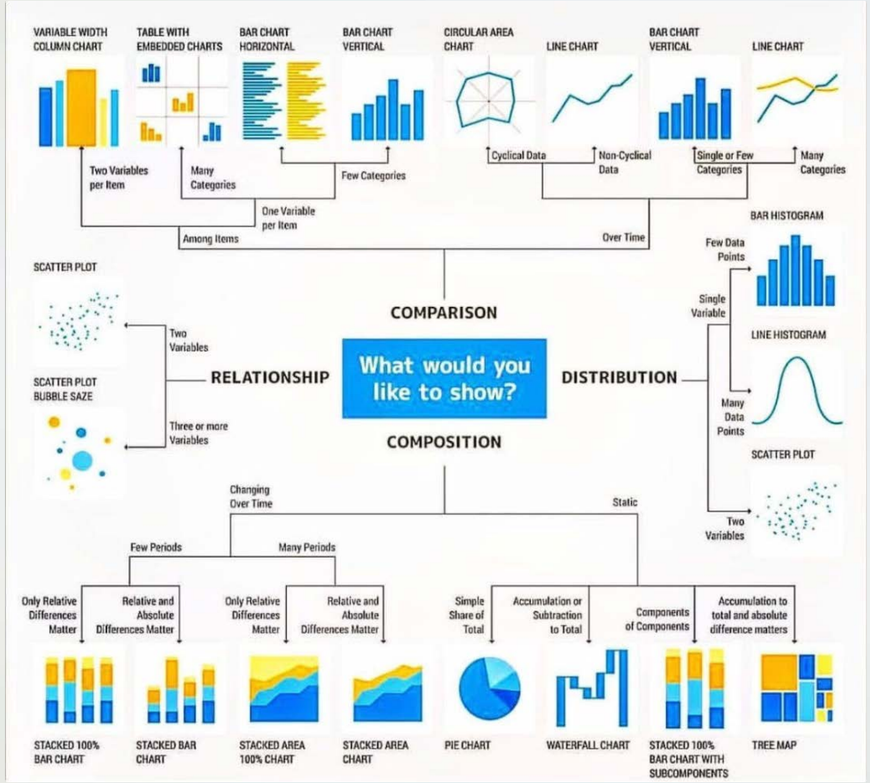
Charts were plotted:

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 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).

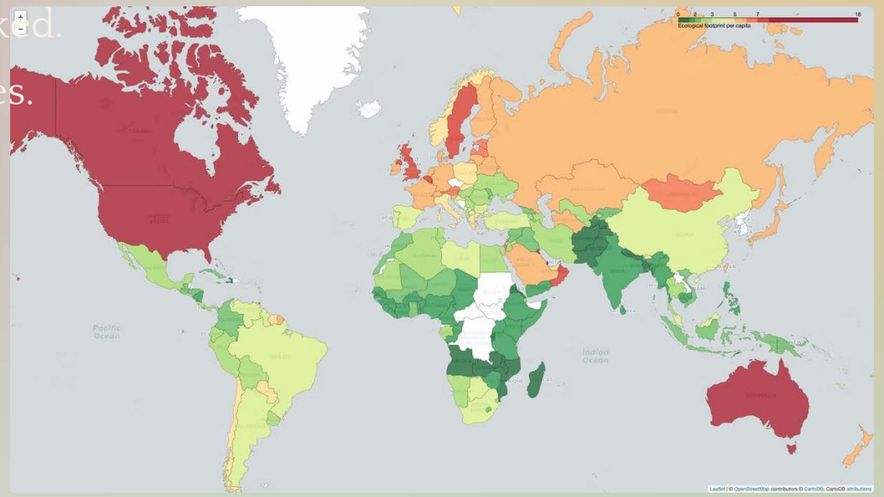


IBM Developer SKILLS NETWORK

Build an interactive map with Folium

1. **Markers for Launch Sites:**
   - Placed markers (dots) on a map for NASA Johnson Space Center and other launch sites.
   - Each marker has a circle around it and shows the name of the site when clicked.
   - These markers also show how close the sites are to the equator and coastlines.
2. **Colored Markers for Launch Outcomes:**
   - Added colored markers for launch outcomes:
     - **Green** for successful launches.
     - **Red** for failed launches.
   - Used clusters to group the markers based on outcomes, helping you see the success rates for each site.
3. **Distances to Nearby Locations:**
   - Drawn colored lines to show how far the Kennedy Space Center is from things like the nearest railway, highway, coastline, and city.

This creates a clear map showing where the launch sites are, how successful their launches have been, and how close they are to nearby features like roads and cities.



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.

Machine Learning Prediction

Functions from the Scikit-learn library are used to create our machine learning models.

The machine learning prediction phase include the following steps:

·Standardizing the data

·Splitting the data into training and test data

·Creating machine learning models, which include:

·Logistic regression

·Support vector machine (SVM)

·Decision tree

·K nearest neighbors (KNN)

·Fit the models on the training set

·Find the best combination of hyperparameters for each model

·Evaluate the models based on their accuracy scores and confusion matrix

RESULTS

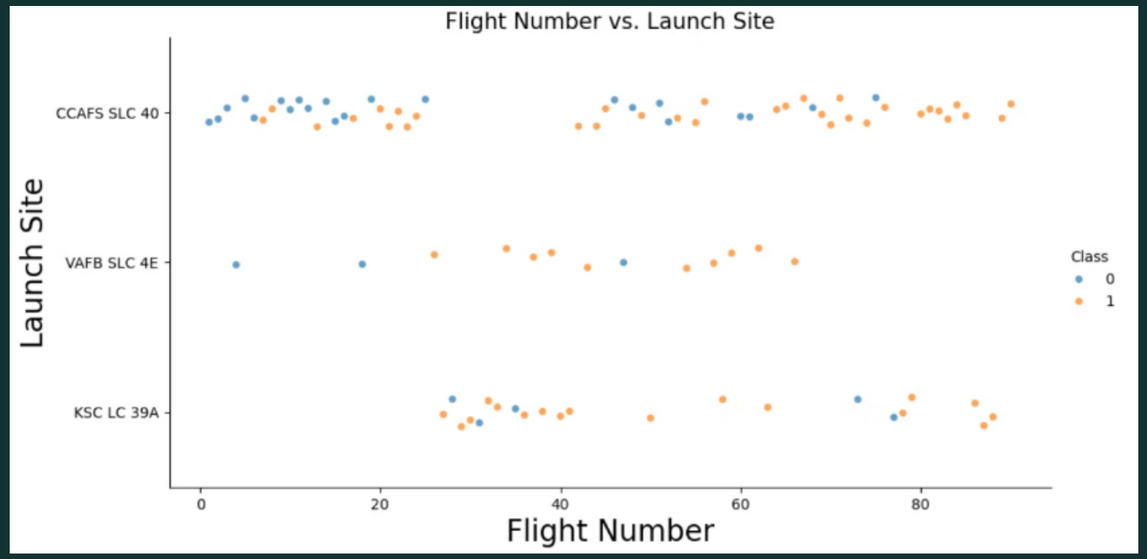
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The results are split into 5 sections:

- SQL (EDA with SQL)
- Matplotlib and Seaborn (EDA with Visualization)
- Folium
- Dash
- Predictive Analysis

EDA with Visualization

The relationship between flight number and launch site

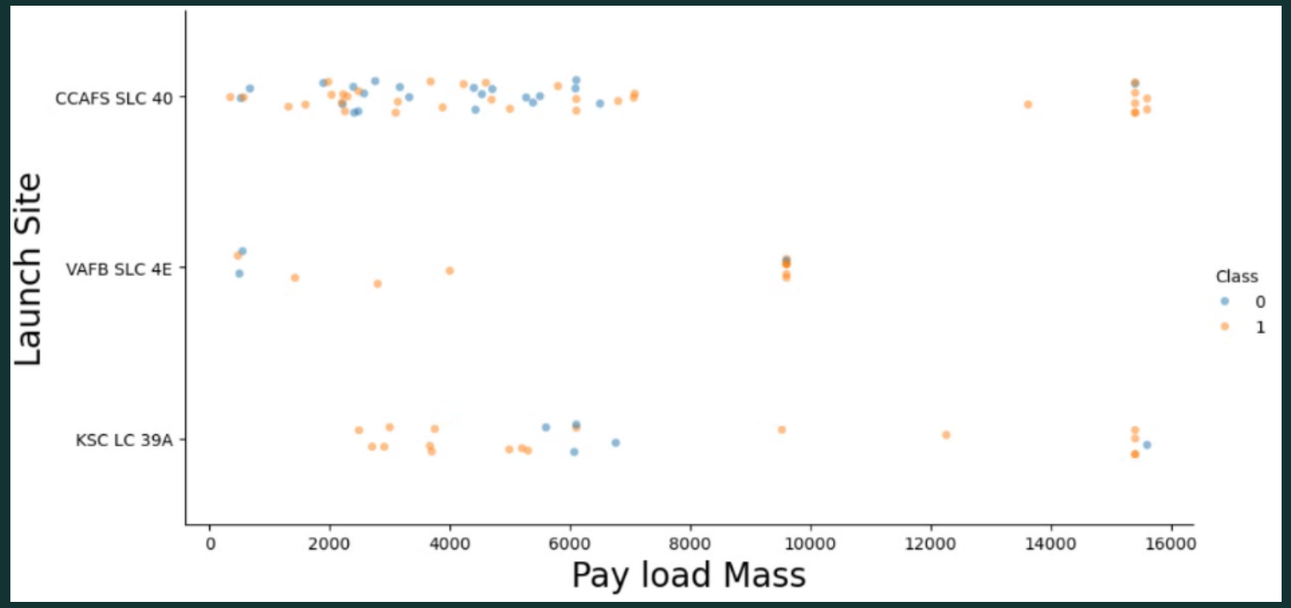


Explanation:

- 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

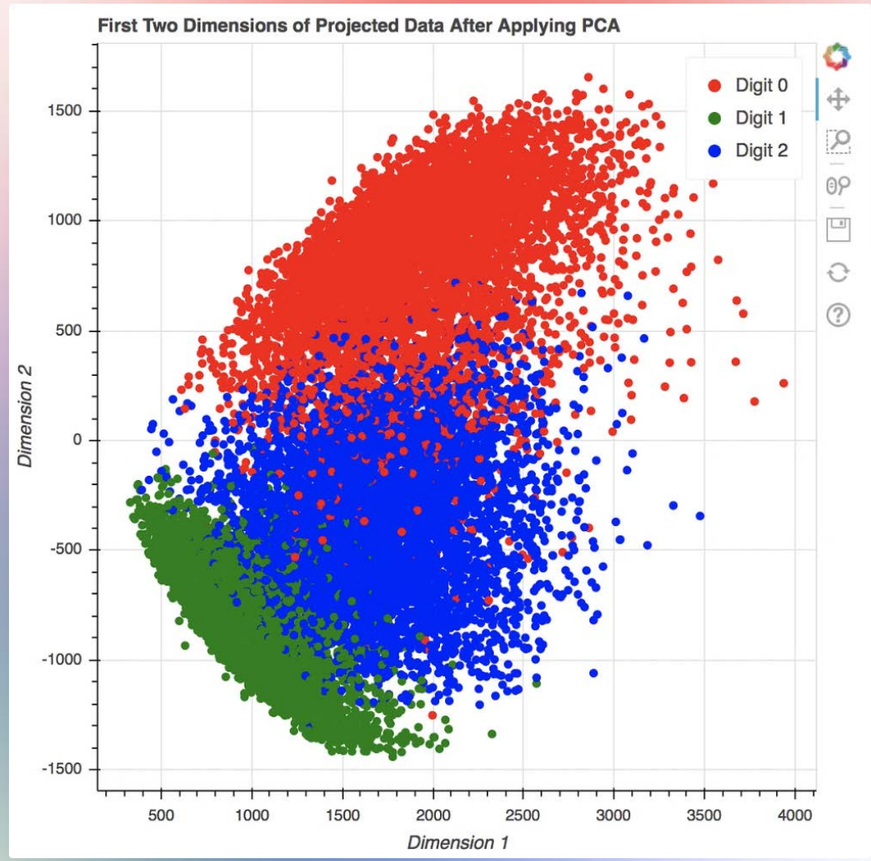
The relationship between payload mass and launch site



Explanation: • For every launch site the higher the payload mass, the higher the success rate.

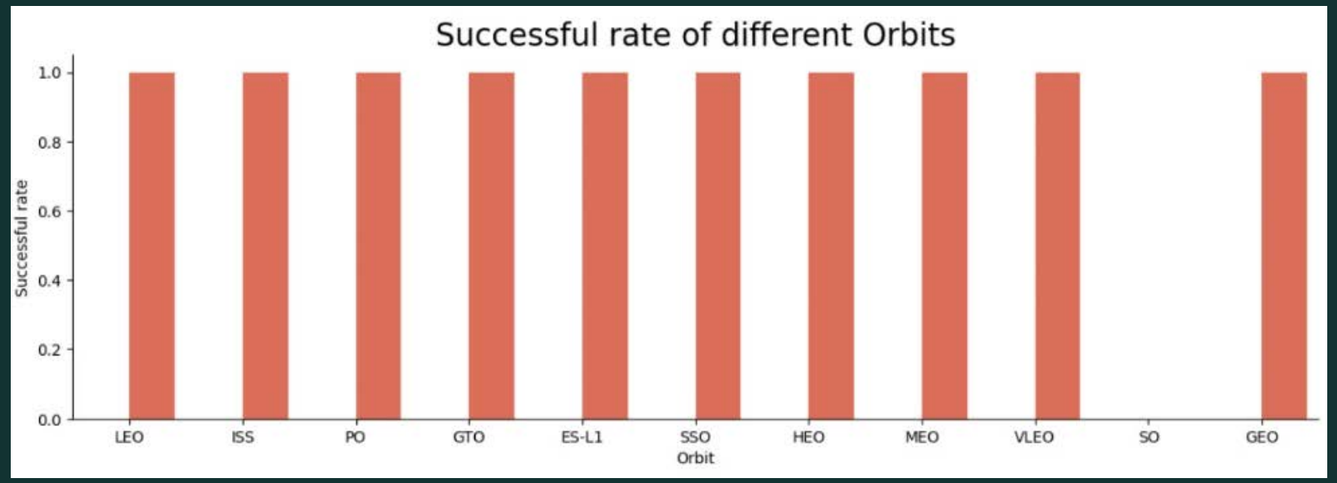
• Most of the launches with payload mass over 7000 kg were successful.

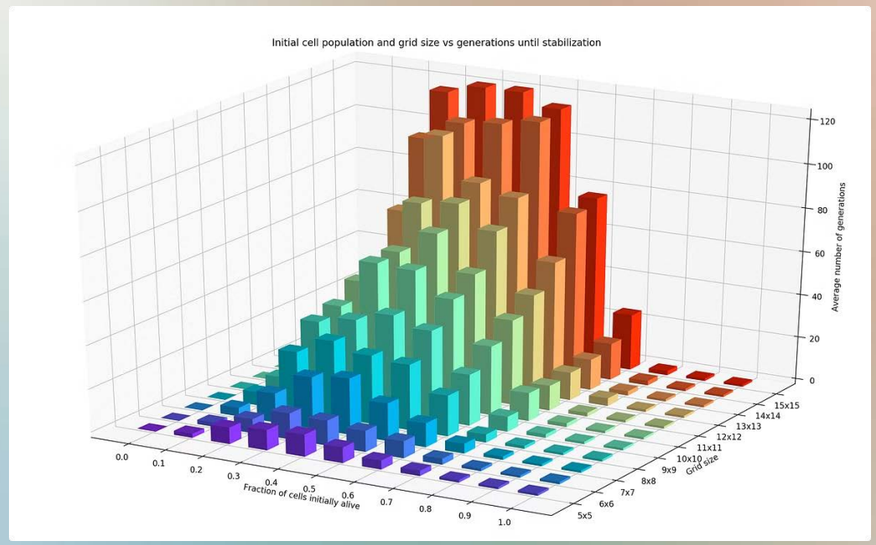
• KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.



EDA with Visualization

The relationship between Success rate and Orbit type



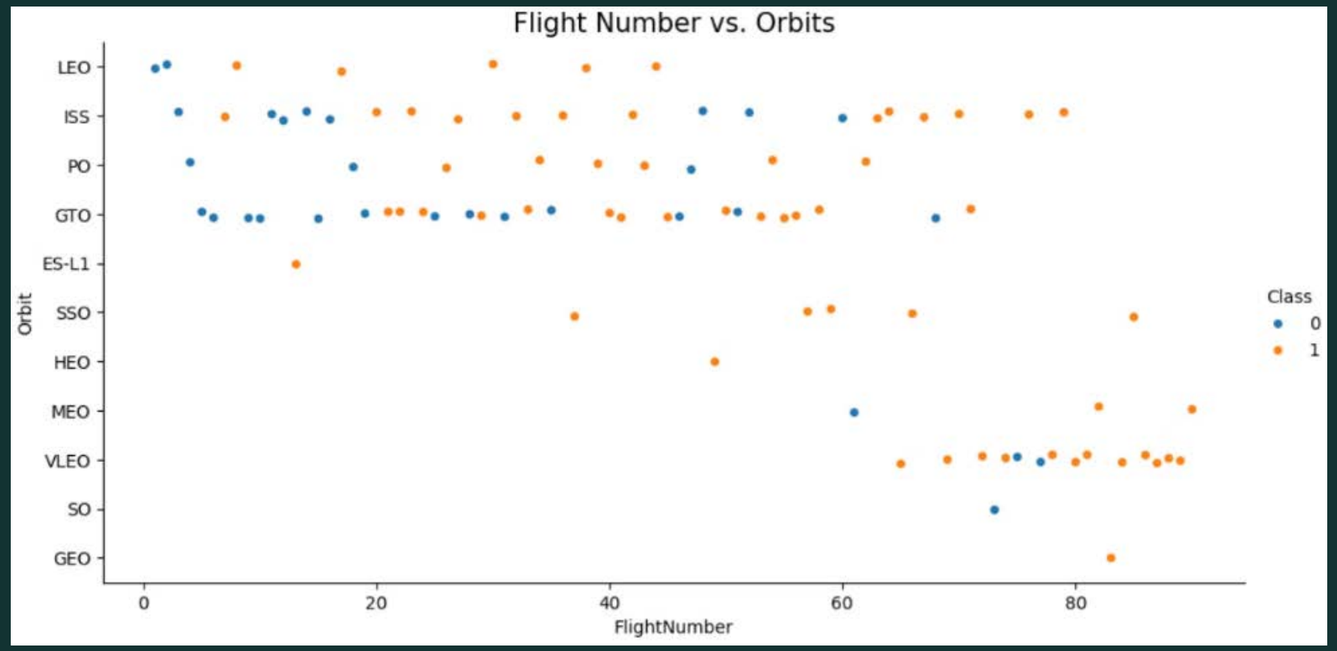


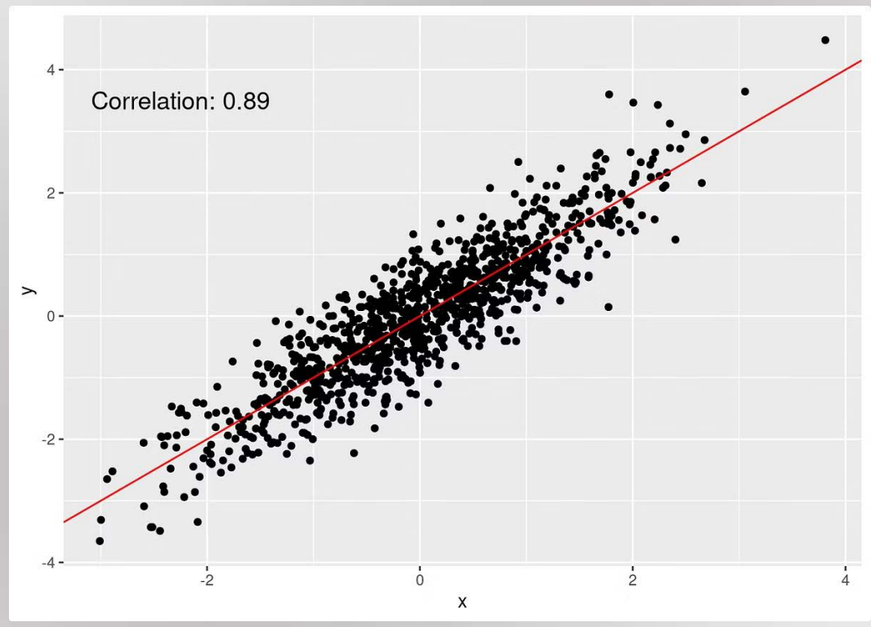
Explanation: • Orbits with 100% success rate:

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

EDA with Visualization

The relationship between Flight Number and Orbit type

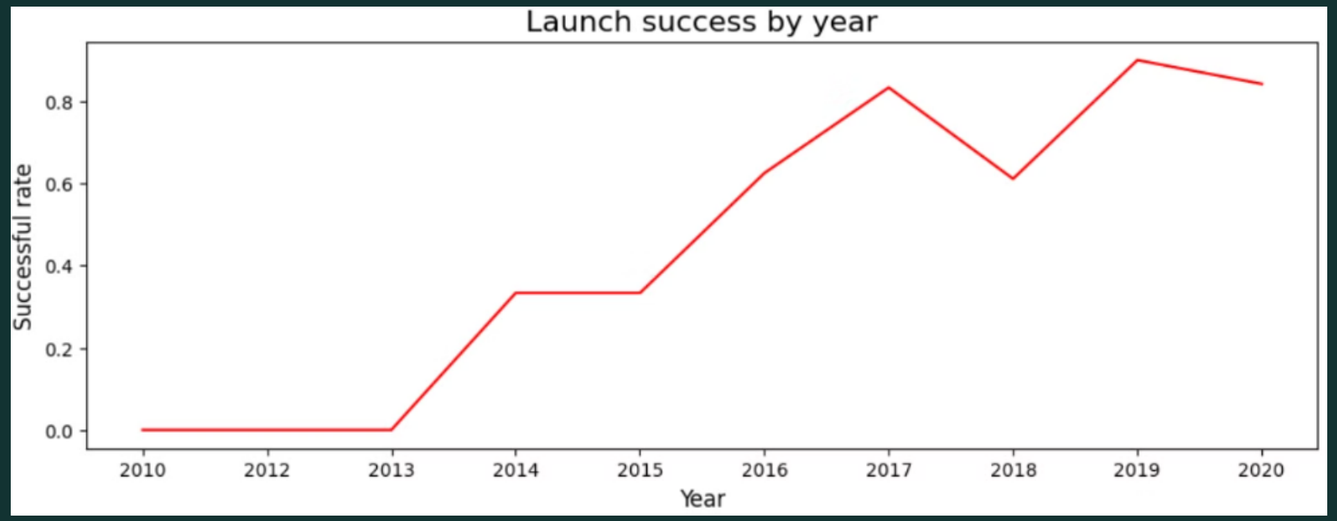




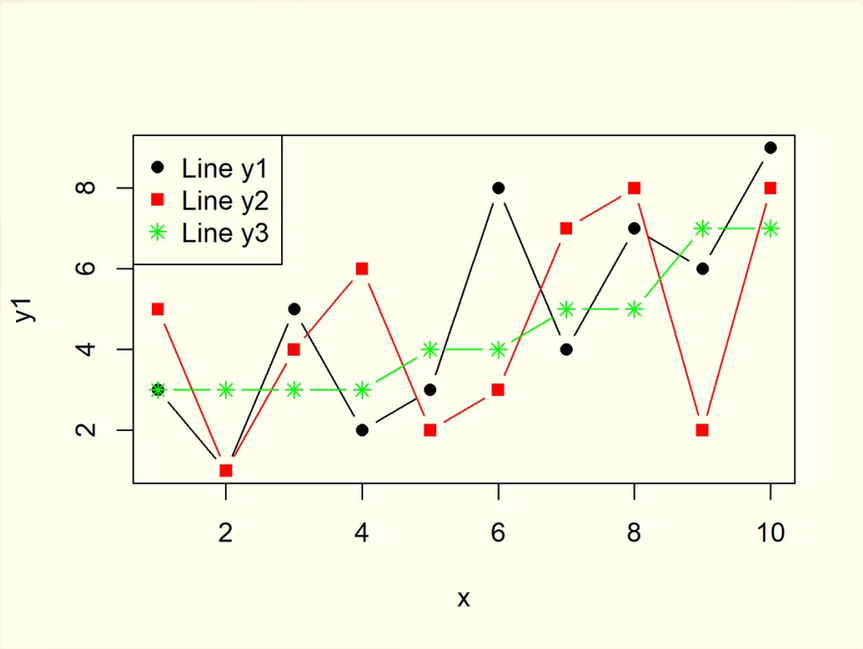
Explanation: • In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

EDA with Visualization

Launch success yearly trend

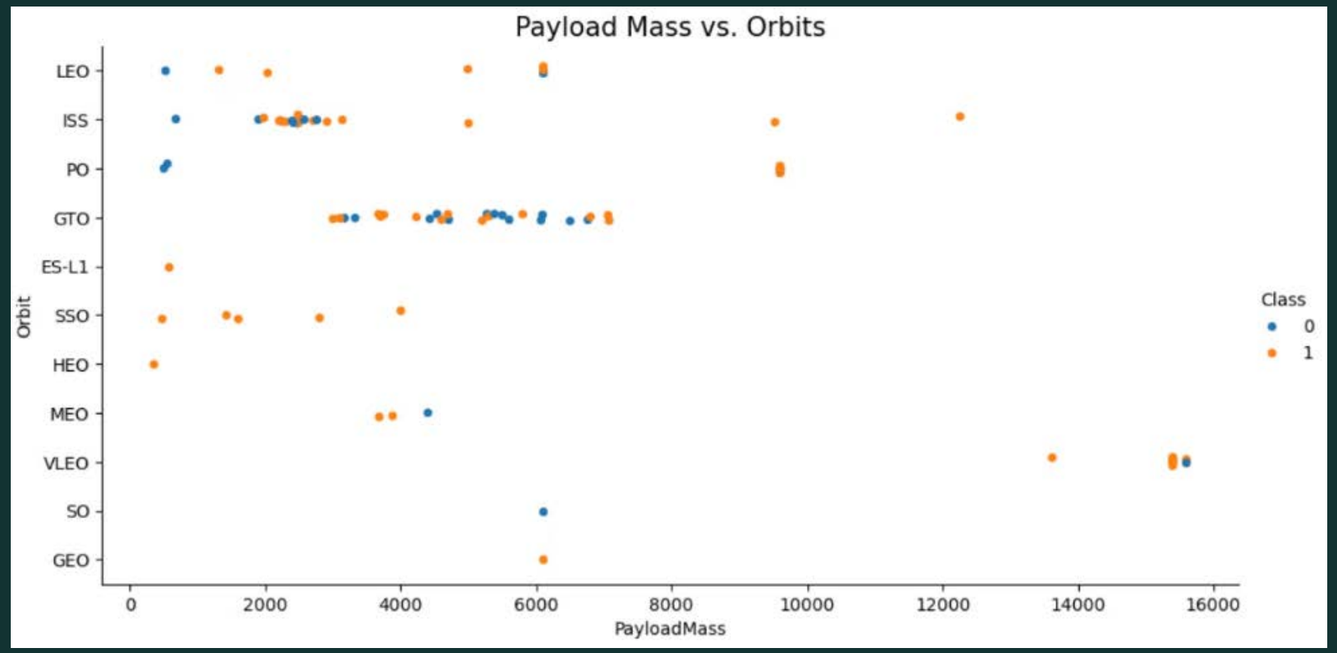


Explanation: • The success rate since 2013 kept increasing till 2020.

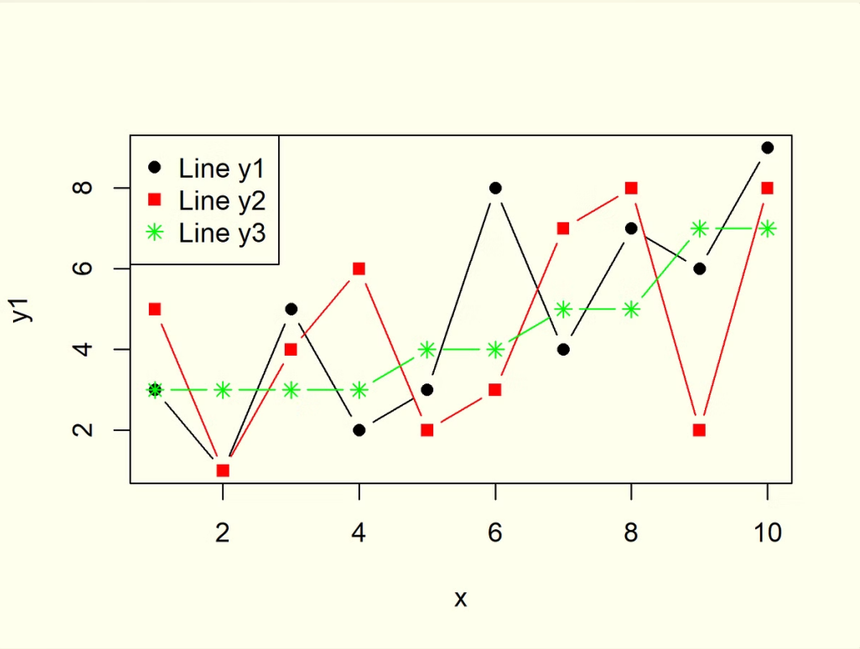


EDA with Visualization

The relationship between Payload Mass and Orbit type



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



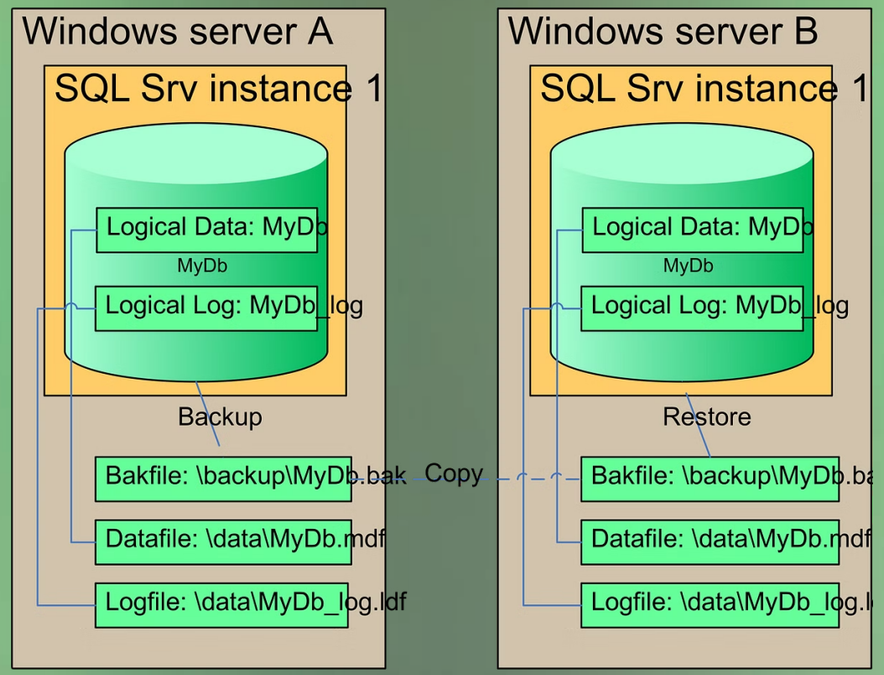
SQL (EDA with SQL)

The names of the unique launch sites in the space mission

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

5 records where launch sites begin with ‘CCA’

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



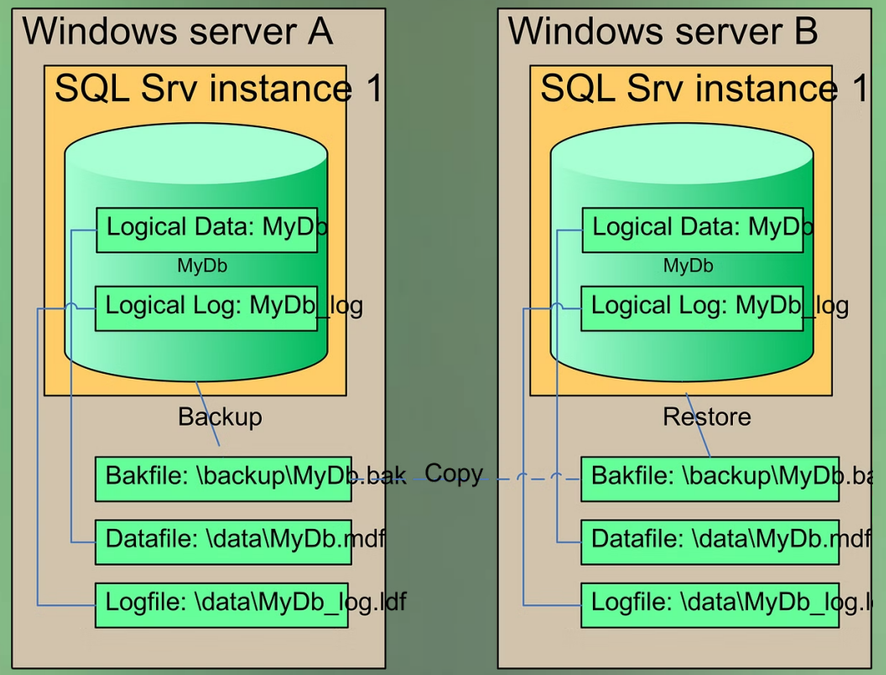
SQL (EDA with SQL)

The names of the booster versions which have carried the maximum payload mass

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

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40



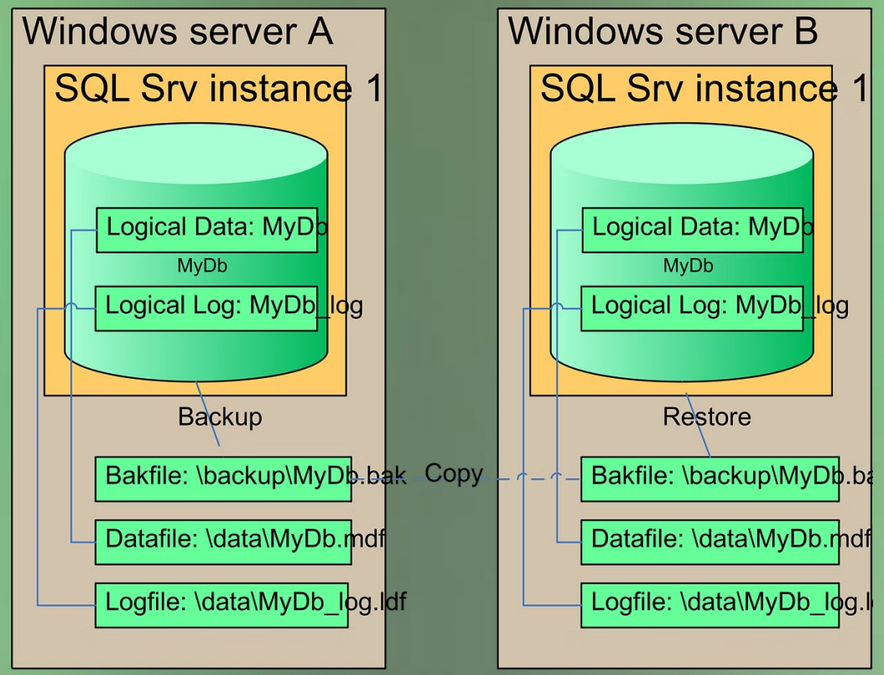
SQL (EDA with SQL)

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

landing__outcome	landing_count
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

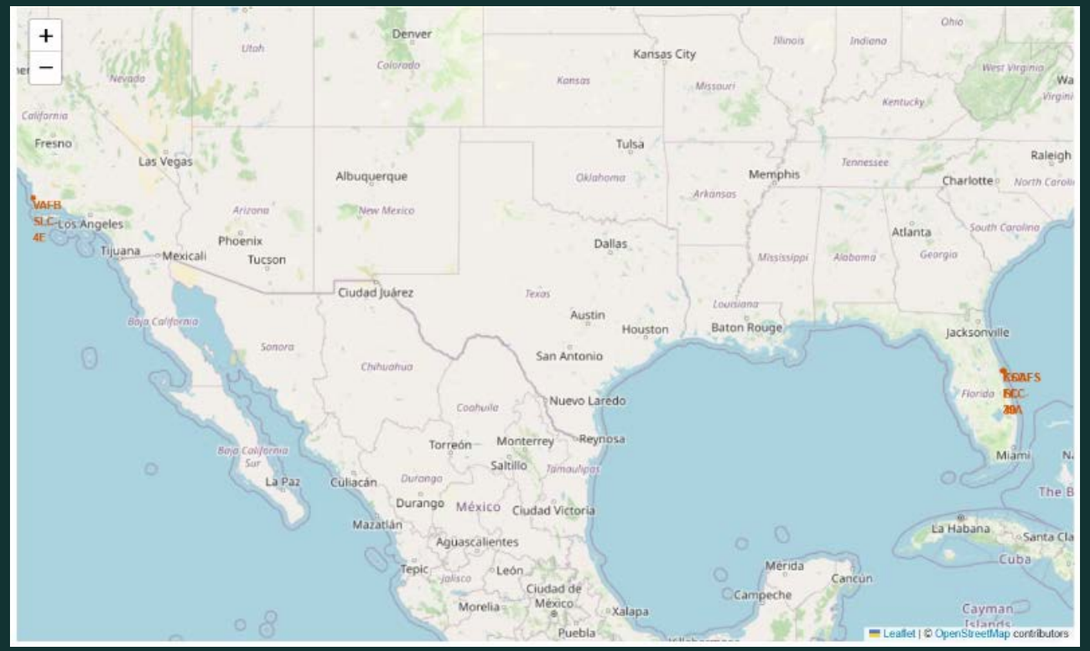
The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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



Folium

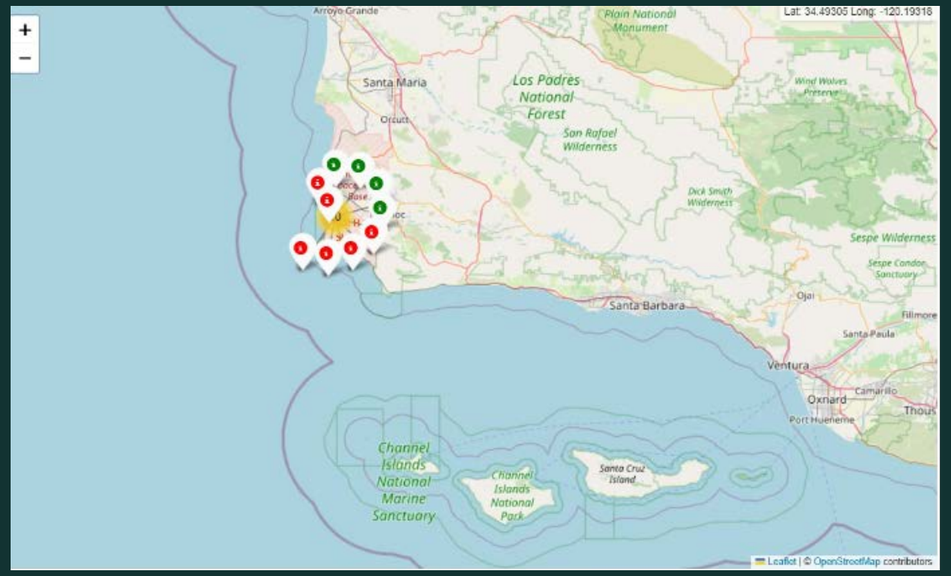
All launch sites on map



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.

Folium

The succeeded launches and failed launches for each site on map

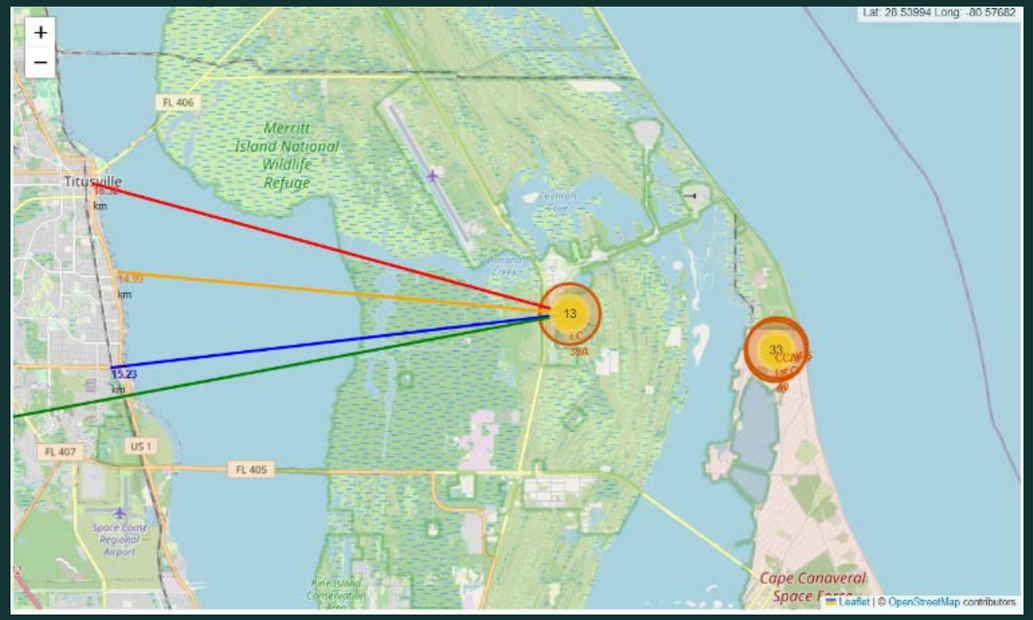


Explanation: · From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- Green Marker = Successful Launch
- Red Marker = Failed Launch
- · Launch Site KSC LC-39A has a very high Success Rate.

Folium

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

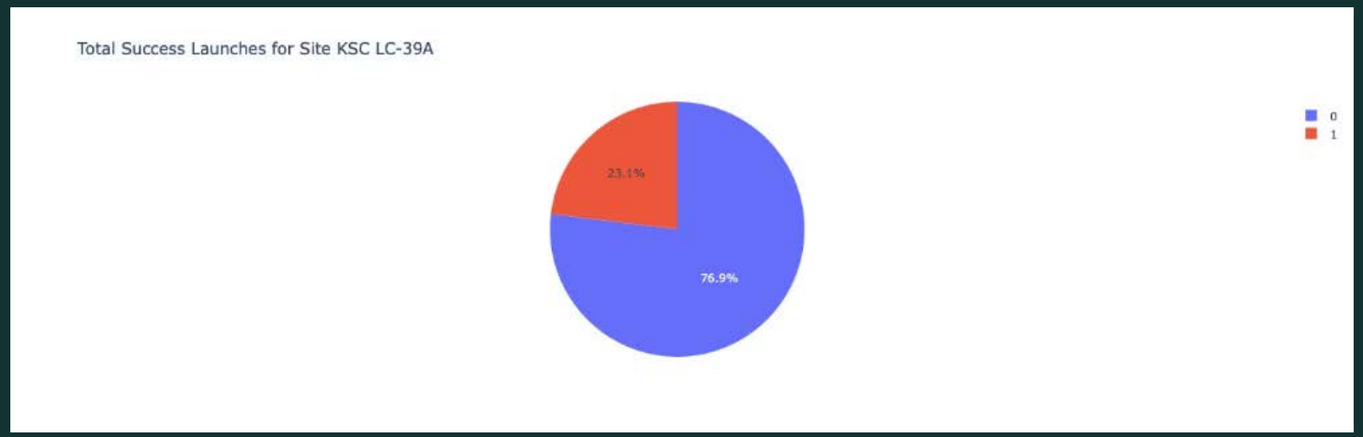


Explanation: • From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)

Build a Dashboard with Plotly Dash

Launch succe



The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.

•0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.

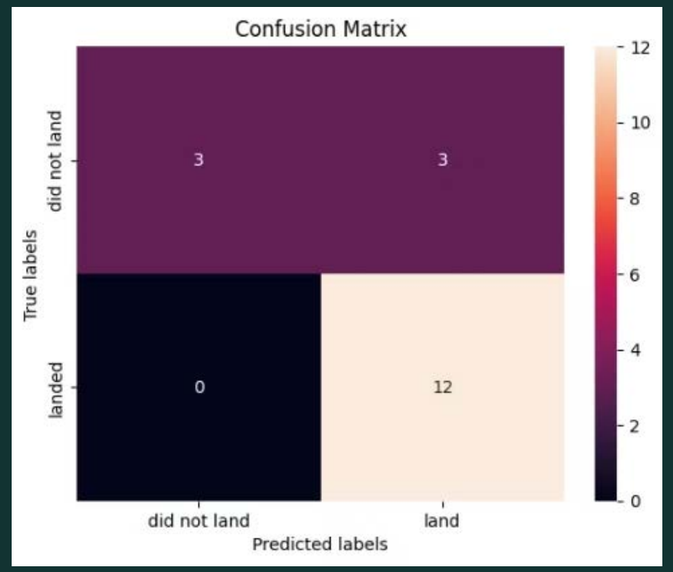
Predictive analysis (Classification)

Logistic regression

•GridSearchCV best score: 0.8464285714285713

•Accuracy score on test set: 0.8333333333333334

Confusion matrix:



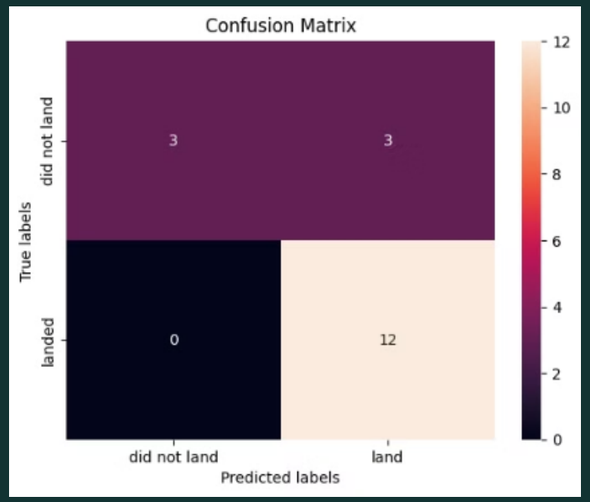
Predictive analysis (Classification)

Support vector machine (SVM)

•GridSearchCV best score: 0.8482142857142856

•Accuracy score on test set: 0.8333333333333334

Confusion matrix:



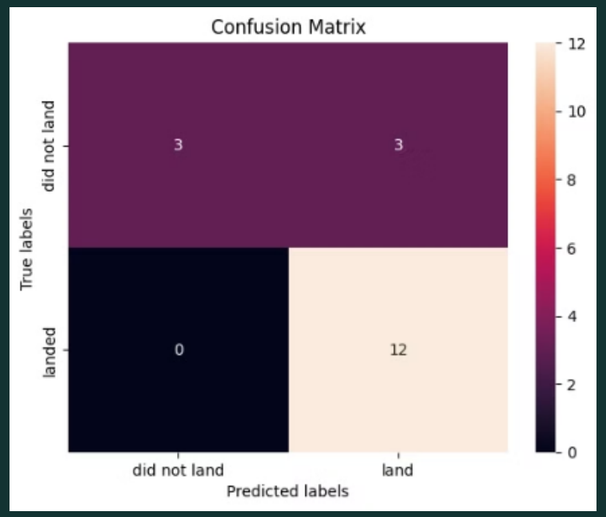
Predictive analysis (Classification)

Predictive analysis (Classification)

K nearest neighbors (KNN)

- GridSearchCV best score: 0.8482142857142858
- Accuracy score on test set: 0.8333333333333334

Confusion matrix:



Predictive analysis (Classification)

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.

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

Scores and Accuracy of the Entire Data Set

DISCUSSION

From the data visualization section, we can see that some features may have correlation with the mission outcome in several ways. For example, with heavy payloads the successful landing or positive landing rate are more for orbit types Polar, LEO and ISS. However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Therefore, each feature may have a certain impact on the final mission outcome. The exact ways of how each of these features impact the mission outcome are difficult to decipher. However, we can use some machine learning algorithms to learn the pattern of the past data and predict whether a mission will be successful or not based on the given features.

Conclusion

- Decision Tree Model is the best algorithm for this dataset.
- 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.