Determining the price of a SpaceX Falcon 9 rocket launch

Ramiro Vasquez Di Zeo

August 14th, 2023

IBM Developer
SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
  - Data collection
  - Data wrangling
  - EDA and interactive visual analytics
  - Predictive analysis
- Results
  - EDA with visualization
  - EDA with SQL
  - Interactive map with FOLIUM
  - Plotly Dash Dashboard
  - Predictive analysis (classification)
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars.
  - Falcon 9 rocket is made up of three stages: Stage One, Stage Two and The Fairings.
- Other providers cost upwards of 165 million dollars each launch.
  - Much of the savings is because SpaceX can reuse the first stage.
- Sometimes the first stage does not land. Sometimes it will crash. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.
- If we can determine if the first stage will land, we can determine the cost of a launch.

INTRODUCTION

- We want to determine the price of each launch, we will do this by gathering information about Space X and creating dashboards.
- We will also determine if SpaceX will reuse the first stage.
- Instead of using rocket science to determine if the first stage will land successfully, we will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

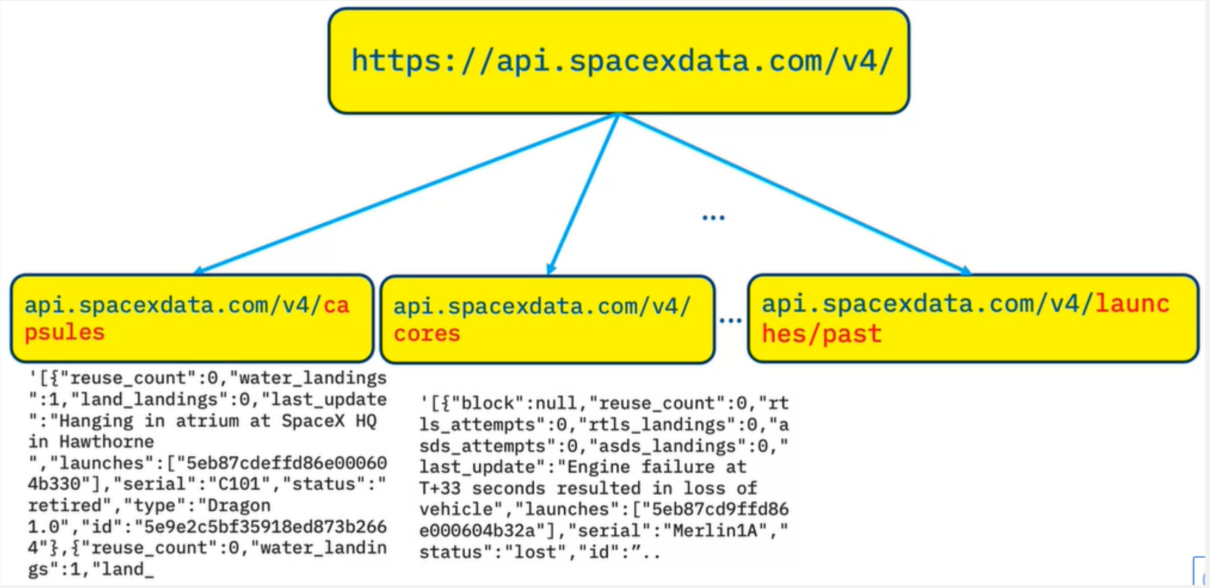
METHODOLOGY

- We will be working with SpaceX launch data that is gathered from an API, specifically the SpaceX REST API.
- We will build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully.
- Using the best hyperparameter values, we will determine the model with the best accuracy using the test data. We will test:
    - Logistic Regression
    - Support Vector machines
    - Decision Tree Classifier
    - K-nearest neighbors

DATA COLLECTION METHODOLOGY

REST API

We will be working with SpaceX launch data that is gathered from an API, ***specifically the SpaceX REST API.*** This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.



DATA COLLECTION METHODOLOGY

Web scraping

Another popular data source for obtaining Falcon 9 Launch data is ***web scraping*** related Wiki pages. We will be using the ***Python BeautifulSoup package*** to web scrape some HTML tables that contain valuable Falcon 9 launch records.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003

DATASET REVIEW

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.561857
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1006	-80.577366	28.561857
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1007	-80.577366	28.561857
8	9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1008	-80.577366	28.561857
9	10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1011	-80.577366	28.561857

DATA WRANGLING METHODOLOGY

We will perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

In the data set, there are several different cases where the booster did not land successfully, We will mainly convert those outcomes into Training Labels with '1' means the booster successfully landed '0' means it was unsuccessful, defined by the feature **'Class'.**

	Class	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite
0	0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40
1	0	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40
2	0	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40
3	0	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E
4	0	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40
5	0	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40
6	1	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40
7	1	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

We will relate different characteristics of the dataset through the visualization of graphs and EDA to find their relationships and patterns.

We will observe the behavior of:

- Flight Number vs. Payload Mass
- Flight Number vs Launch Site
- Launch Site vs Payload Mass
- Success rate and orbit type
- Flight Number vs Orbit type
- Payload vs Orbit type
- Launch success yearly trend

We will mark on a map the success/failed launches for each launch site and the proximity of these sites to a coastline, railway, highway and city.

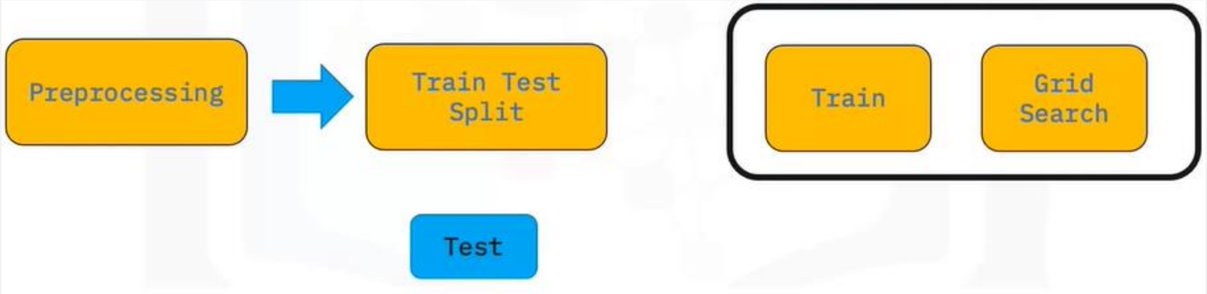
PREDICTIVE ANALYSIS METHODOLOGY

Classification

We will build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully. This will include:

- Preprocessing, allowing us to standardize our data.
- Train_test_split, allowing us to split our data into training and testing data.

We will train the model and perform Grid Search, allowing us to find the hyperparameters that allow a given algorithm to perform best.



PREDICTIVE ANALYSIS METHODOLOGY

Using the best hyperparameter values, we will determine the model with the best accuracy using the test data. We will test:

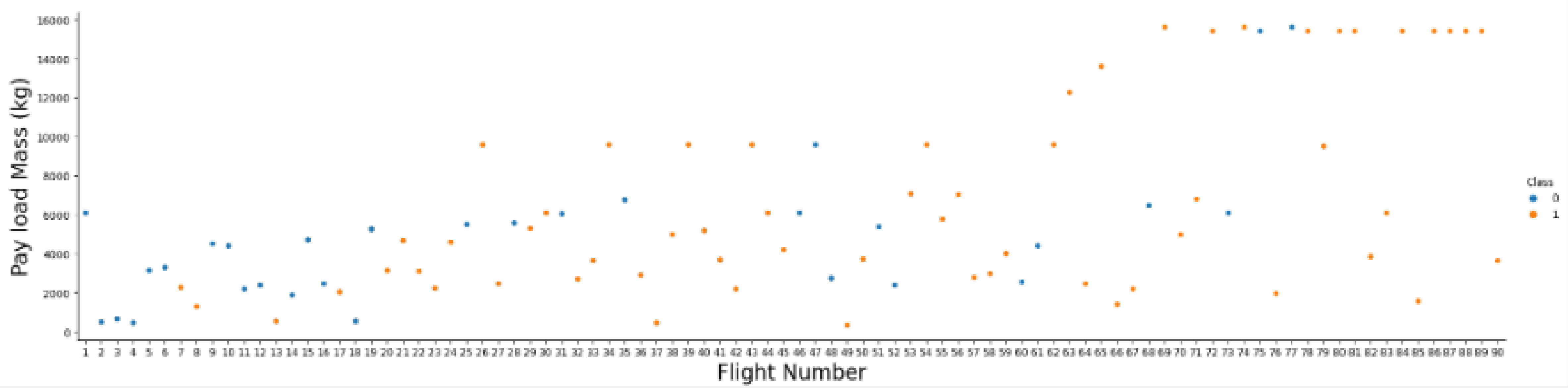
- Logistic Regression
- Support Vector machines,
- Decision Tree Classifier,
- K-nearest neighbors.

Finally, we will output the confusion matrix.

Section 2
Results

EDA WITH VISUALIZATION RESULTS

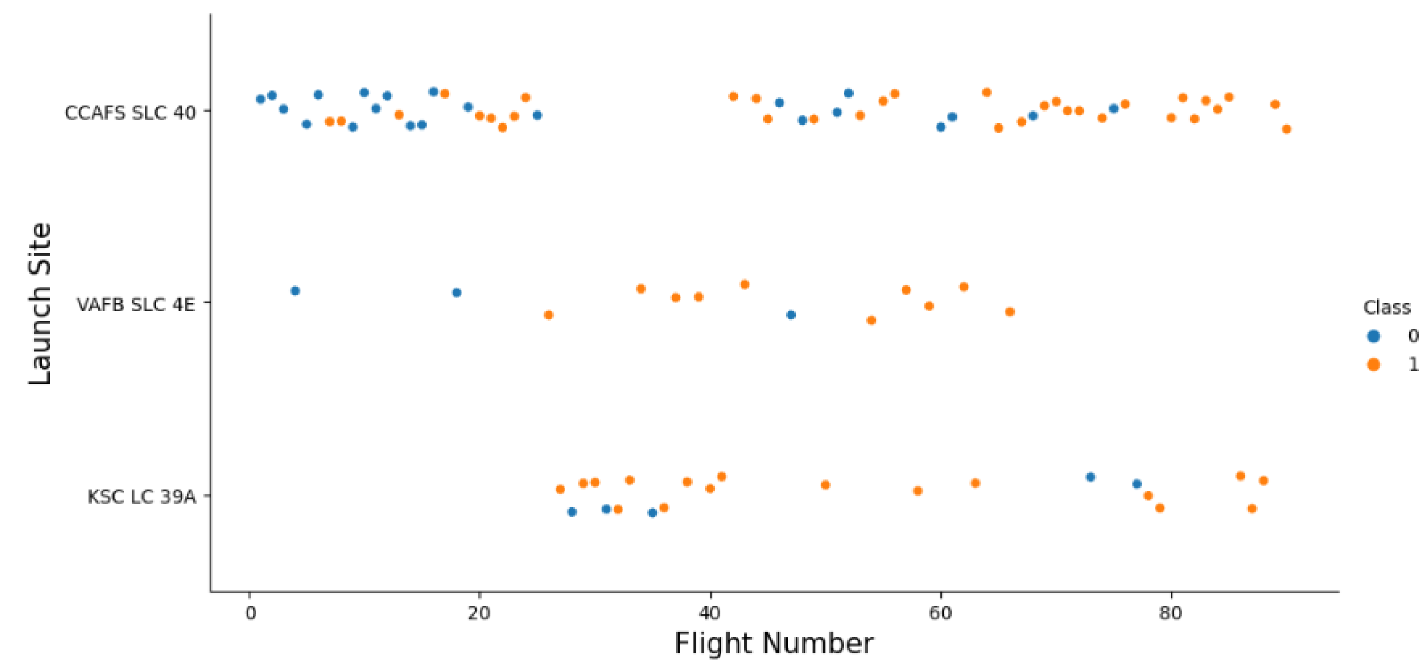
Flight Number vs. Payload Mass (blue mark = no landing, yellow mark = landing)



We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

EDA WITH VISUALIZATION RESULTS

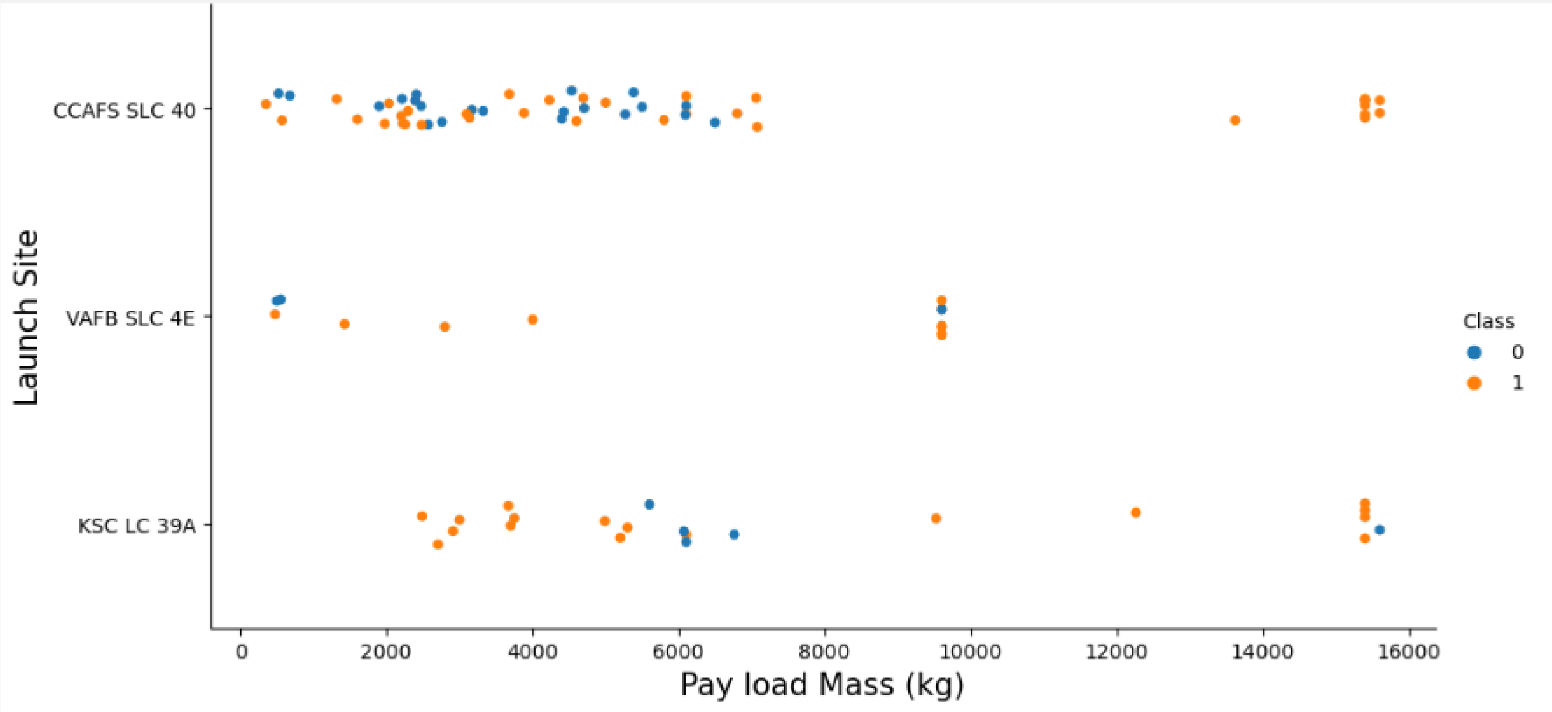
Flight Number vs Launch Site (blue mark = no landing, yellow mark = landing)



- There are too many launches from CCAFS SLC 40 that have not landed.
- There are few launches from VAFB SLC 4E but most have landed successfully.
- Most launches from KSC LC 39A have landed successfully.

EDA WITH VISUALIZATION RESULTS

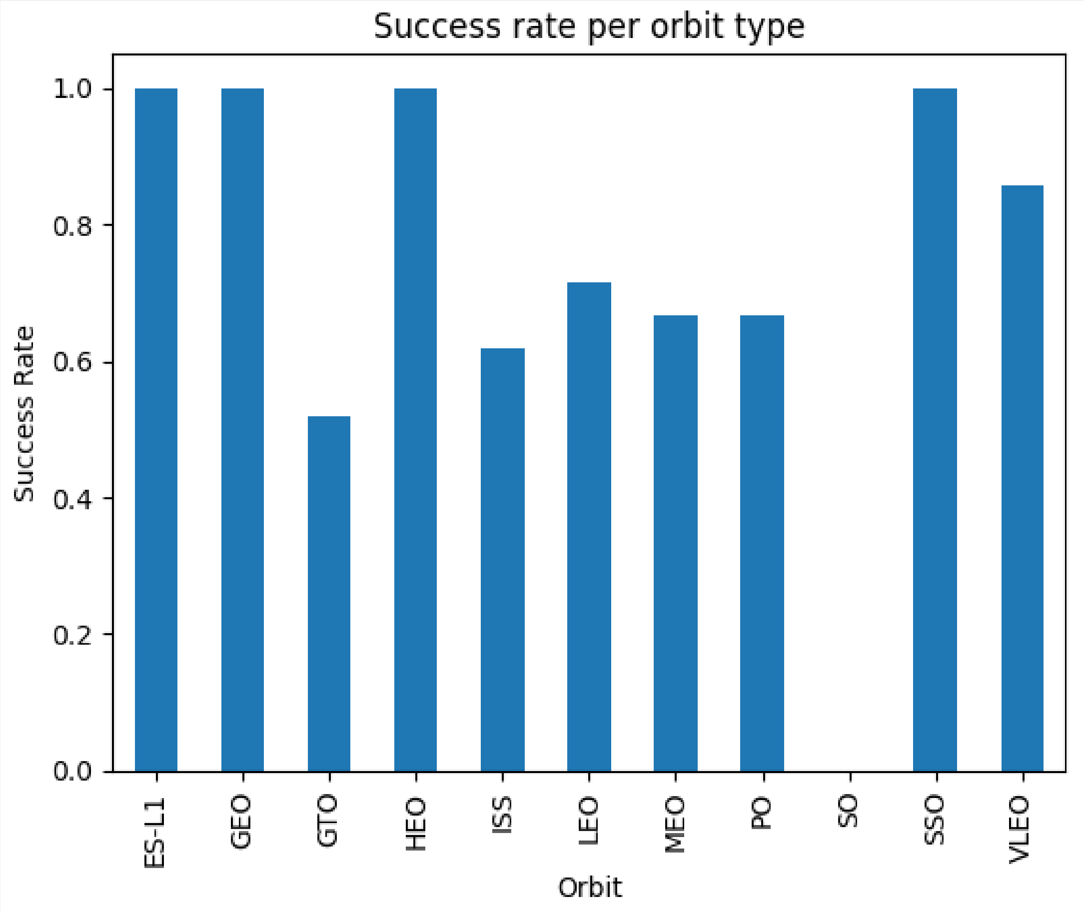
Launch Site vs Payload Mass (blue mark = no landing, yellow mark = landing)



We will find for the VAFB-SLC launch site there are no rockets launched for heavypayload mass(greater than 10000).

EDA WITH VISUALIZATION RESULTS

Success rate of each orbit type

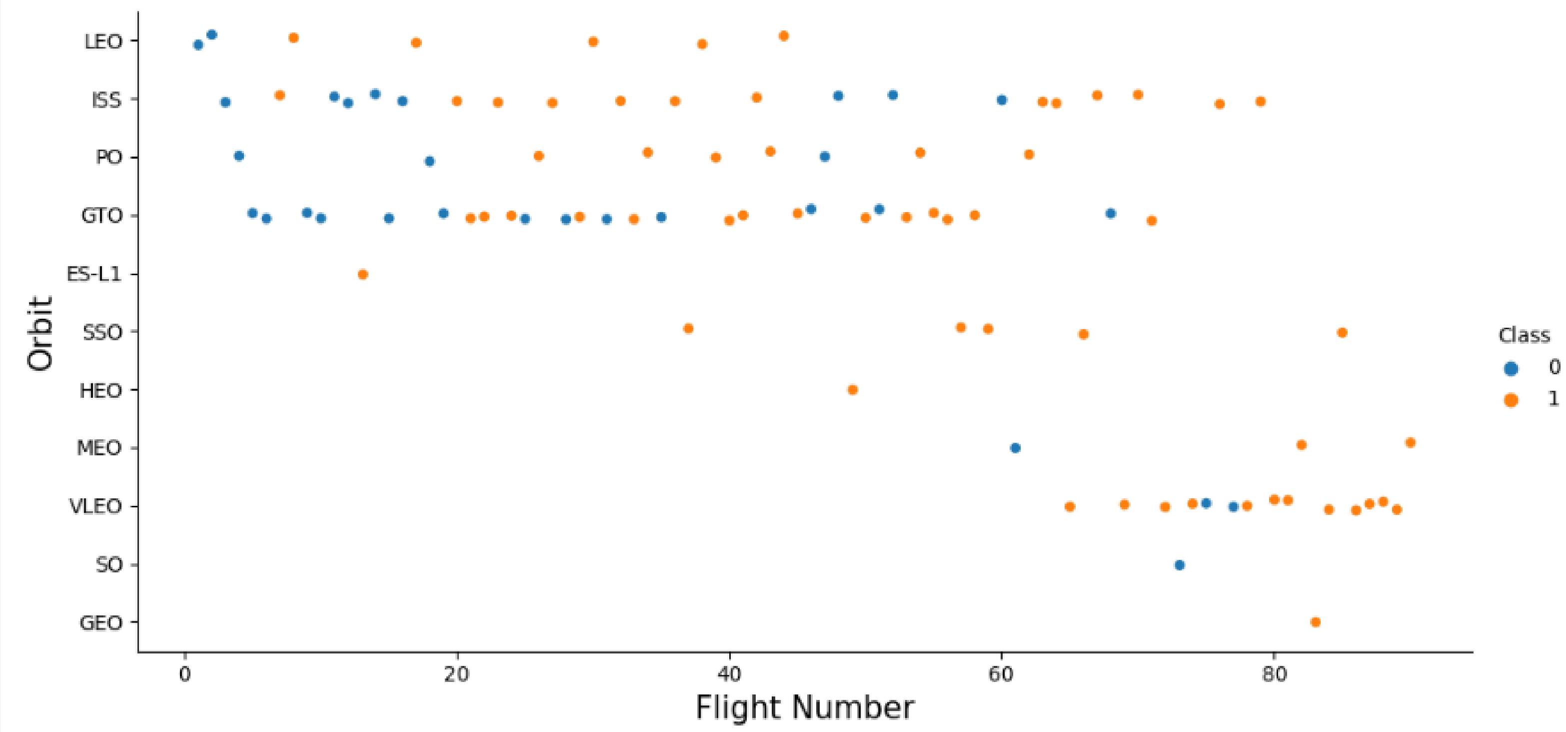


- The orbits with high success rate are:
  - ES-L1
  - GEO
  - HEO
  - SSO

- The success rate variate from 0 for no landing to 1 for successful landing

EDA WITH VISUALIZATION RESULTS

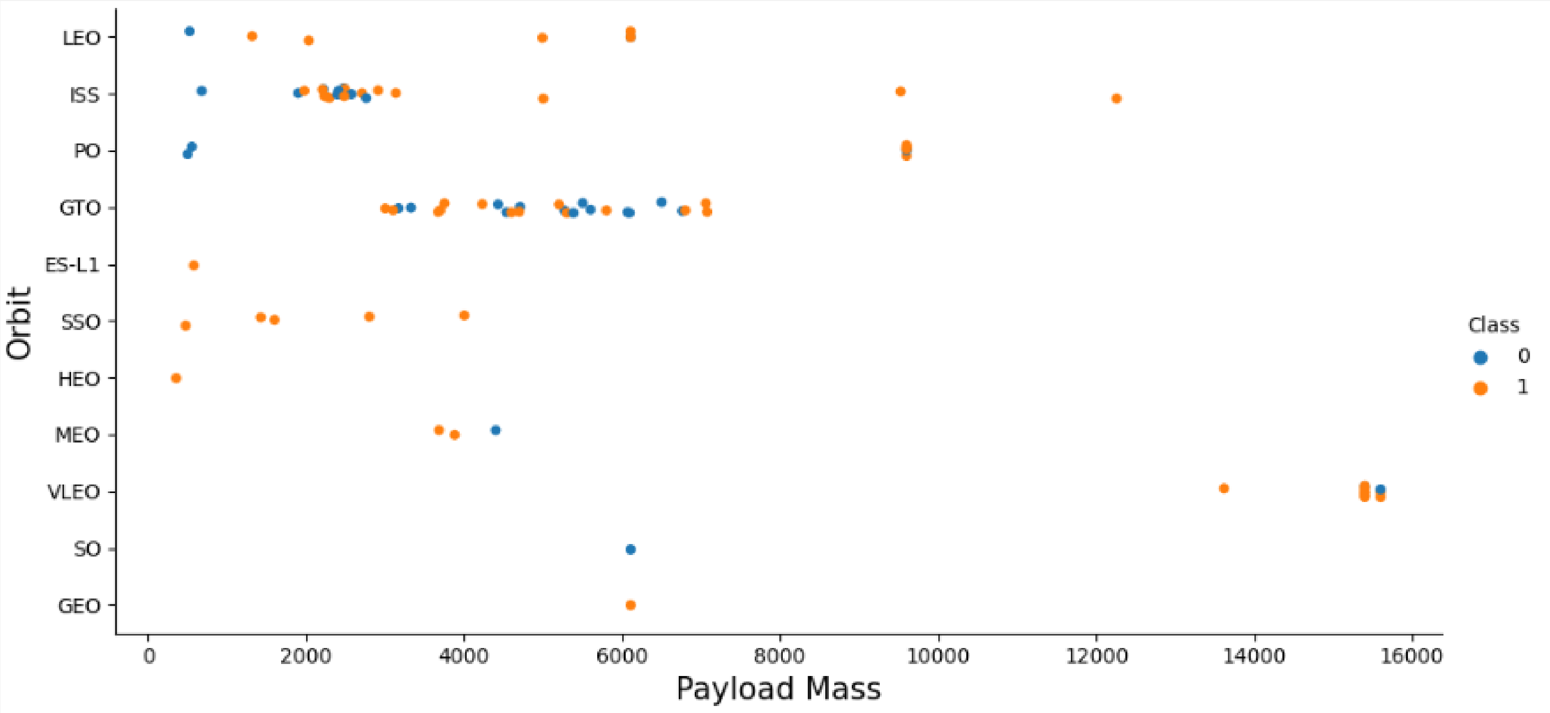
Flight Number vs Orbit type (blue mark = no landing, yellow mark = landing)



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 RESULTS

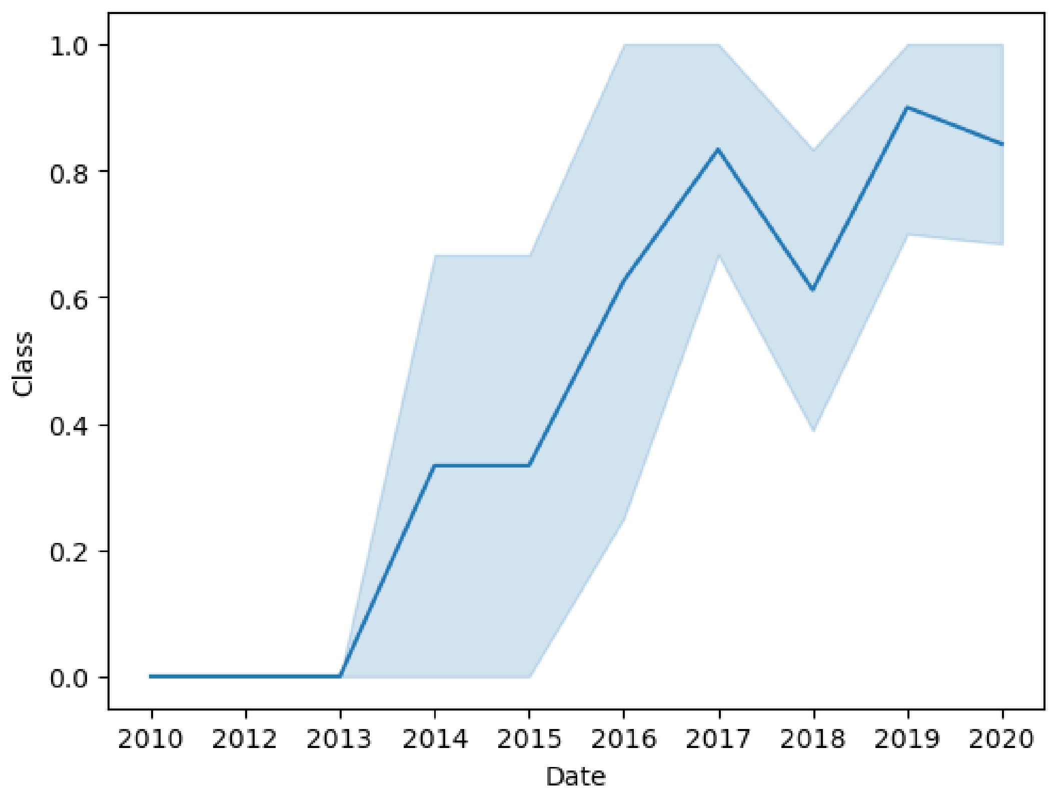
Payload vs Orbit type (blue mark = no landing, yellow mark = landing)



- With heavy payloads the successful landing or positive landing rate are more for 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.

EDA WITH VISUALIZATION RESULTS

Launch success yearly trend



We can observe that the sucess rate since 2013 kept increasing till 2020

EDA WITH SQL RESULTS

Through the exploratory data analysis with SQL we were able to obtain:

- The total payload mass carried by boosters launched by NASA (CRS): 48.213Kg
- The total number of successful and failure mission outcomes.
- From 2019 onward they have been launched rockets with the heaviest payload 15.600 Kg, obtaining a significantly high success rate.

Mission_Outcome	Landing_Outcome	TOTAL
Success	Controlled (ocean)	5
Success	Failure	3
Success	Failure (drone ship)	5
Success	Failure (parachute)	2
Success	No attempt	21
Success	No attempt	1
Failure (in flight)	Precluded (drone ship)	1
Success	Success	38
Success	Success (drone ship)	14
Success	Success (ground pad)	9
Success	Uncontrolled (ocean)	2

EDA WITH SQL RESULTS

Through the exploratory data analysis with SQL we were able to obtain:

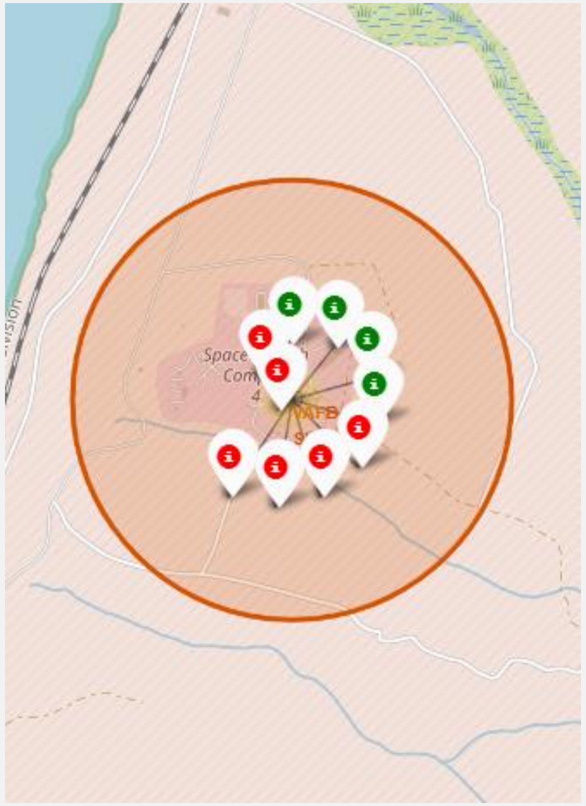
- The date when the first successful landing outcome in ground pad was achieved: 2015-12-22
- The launches that have landed on drone ship between 2010 to 2017 have all been failed, while those that landed on a ground pad have all been successful.

INTERACTIVE MAP WITH FOLIUM RESULTS

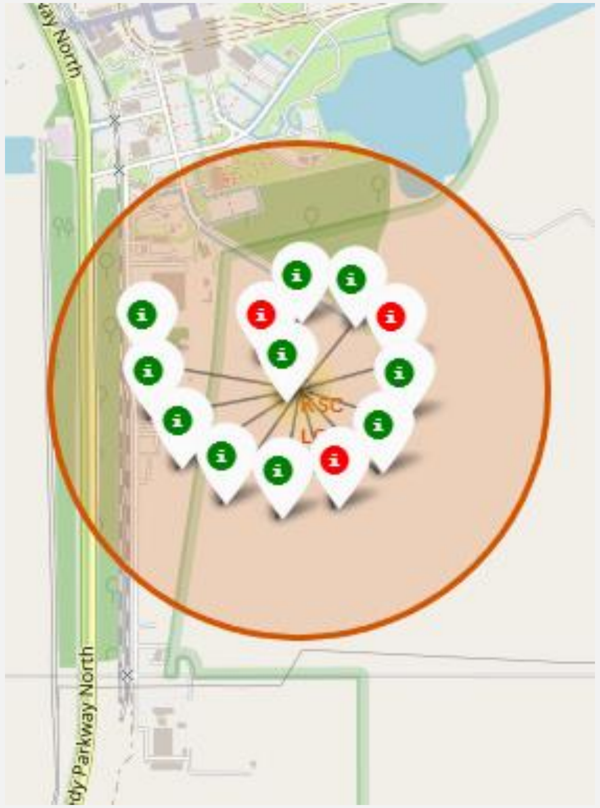
Marking the success/failed launches for each site on the map

INTERACTIVE MAP WITH FOLIUM RESULTS

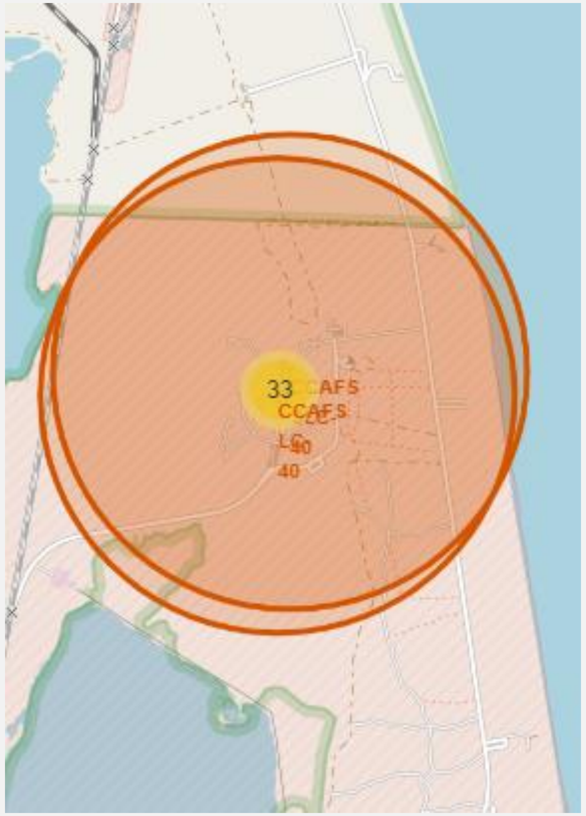
Marking the success/failed launches for each site on the map



VAFB SLC-4E



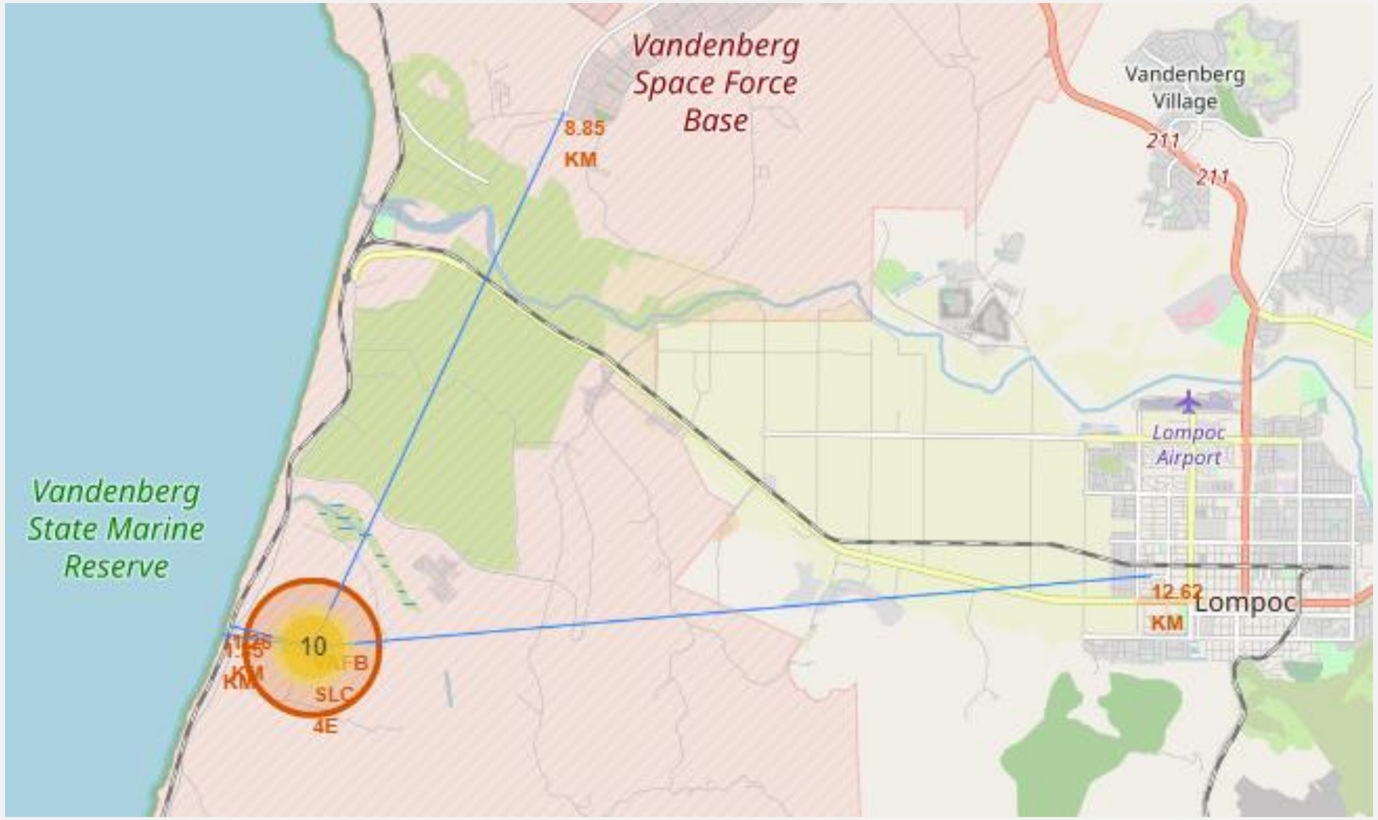
KSC LC-39A



CCAFS LC-40, CCAFS SLC-40

INTERACTIVE MAP WITH FOLIUM RESULTS

Calculating the proximity of each site to a coastline, railway, highway and city



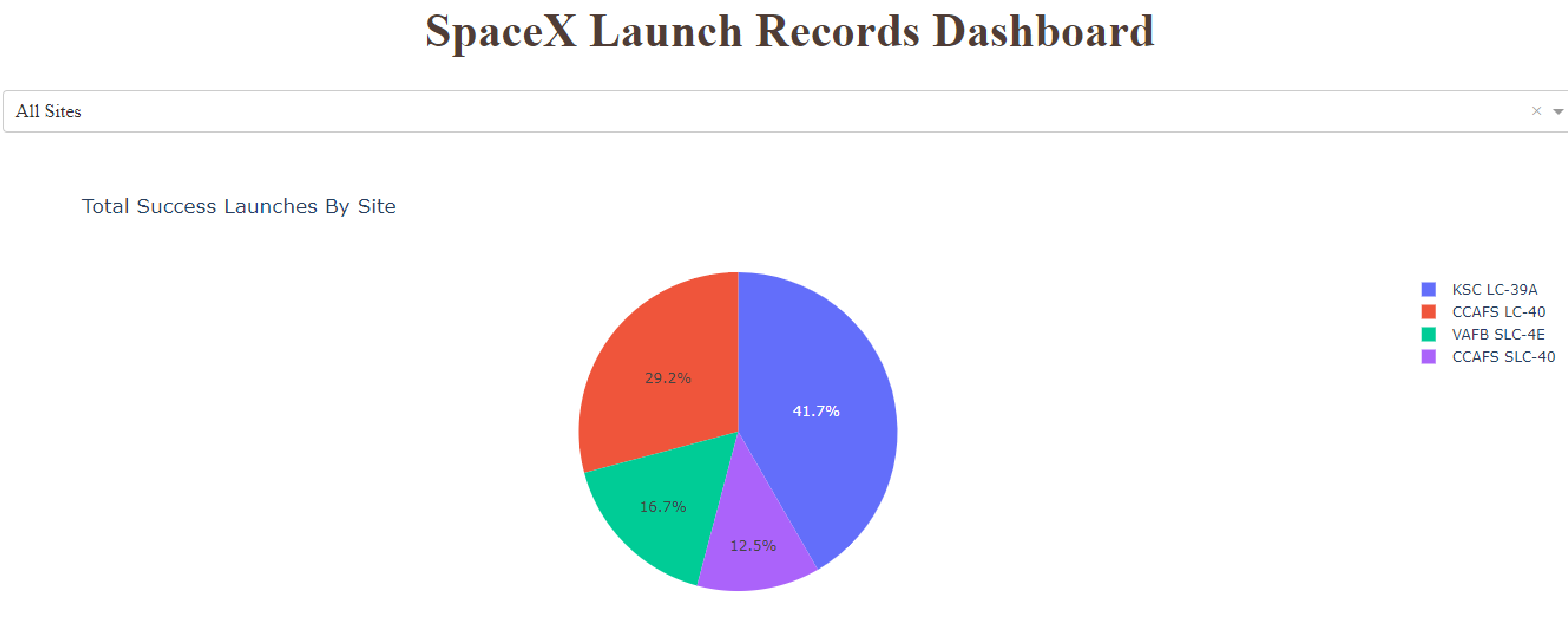
VAFB SLC-4E

INTERACTIVE MAP WITH FOLIUM RESULTS

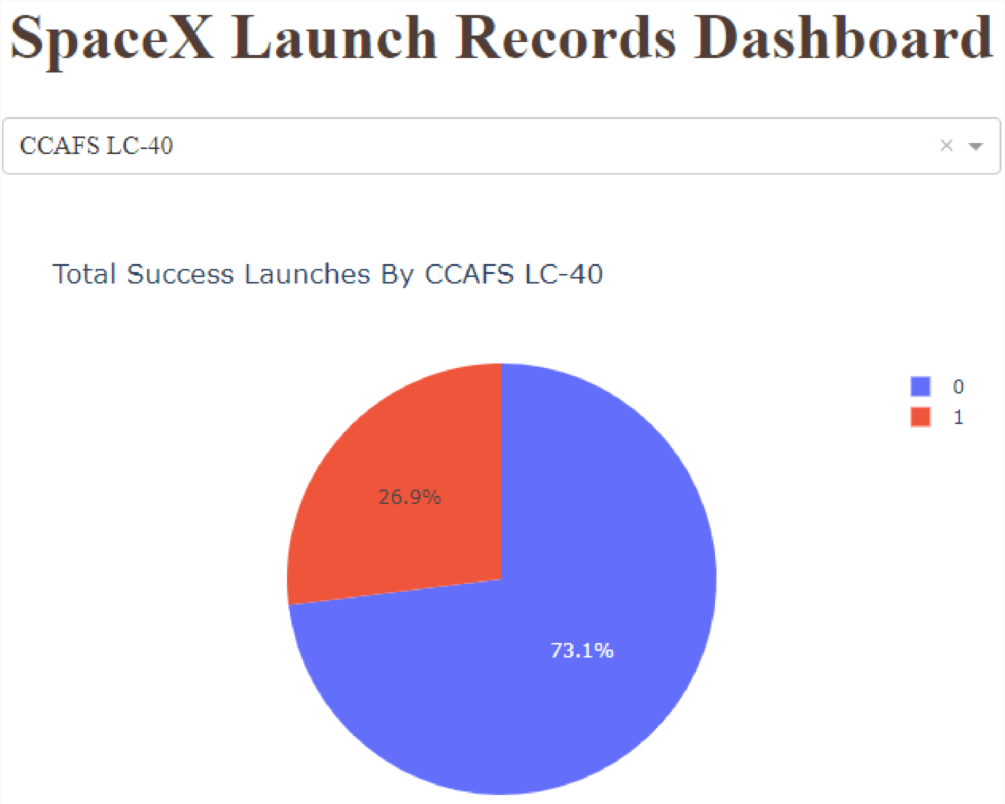
Calculating the proximity of each site to a coastline, railway, highway and city

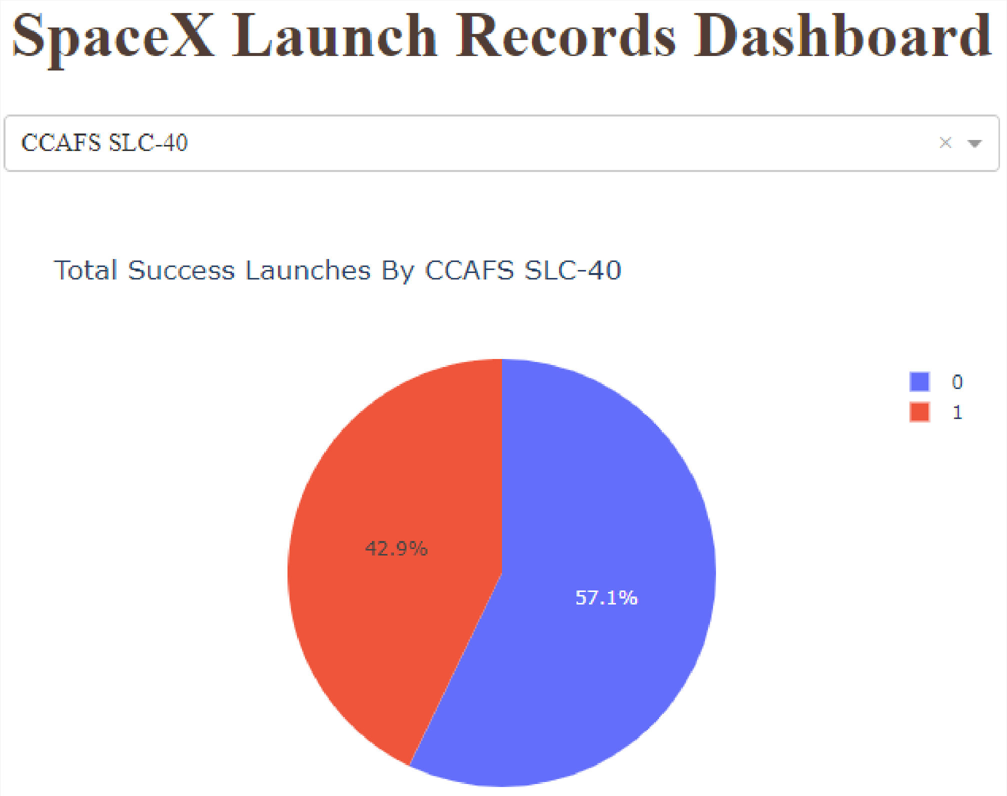
- The closest launch sites to a coastline are: CCAFS LC-40 and CCAFS SLC-40
- All launch sites have an average proximity to a railway of 1,15 km
- VAFB SLC-4E have a longest distance to a highway (8,85 km)
- The launch sites farthest from city are: CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A with a distance of 51,5 km. VAFB SLC-4E is closer to a city with 12,6 km.
- KSC LC 39A had the most successful launches of any sites.

PLOTLY DASH DASHBOARD RESULTS

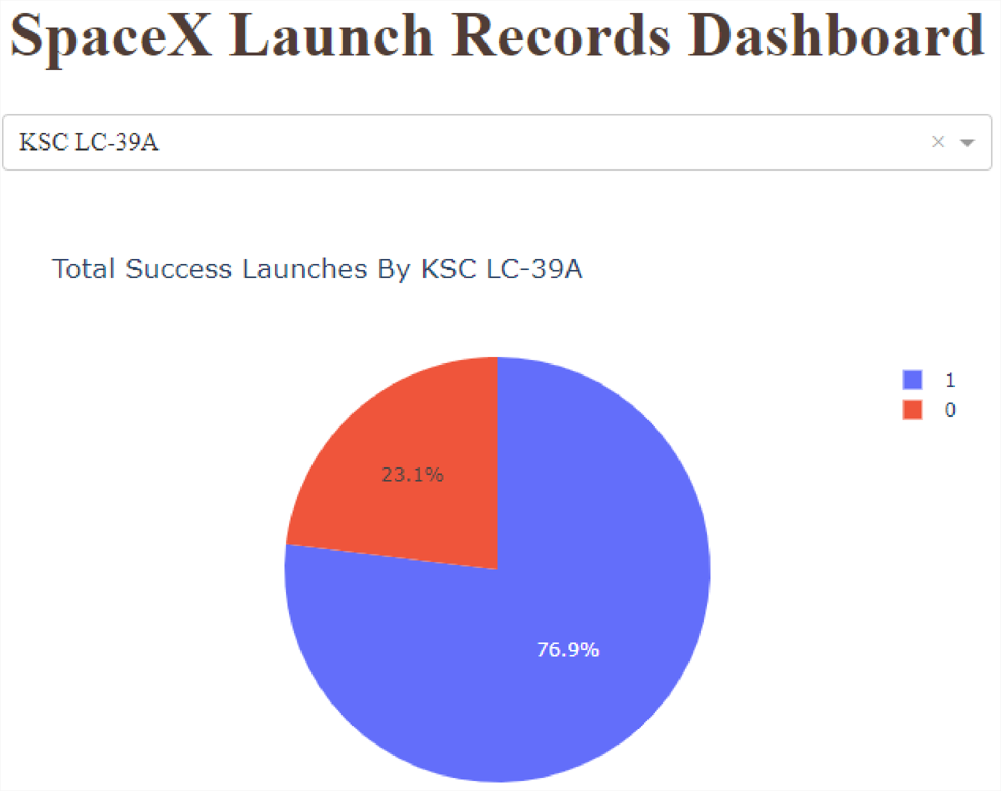


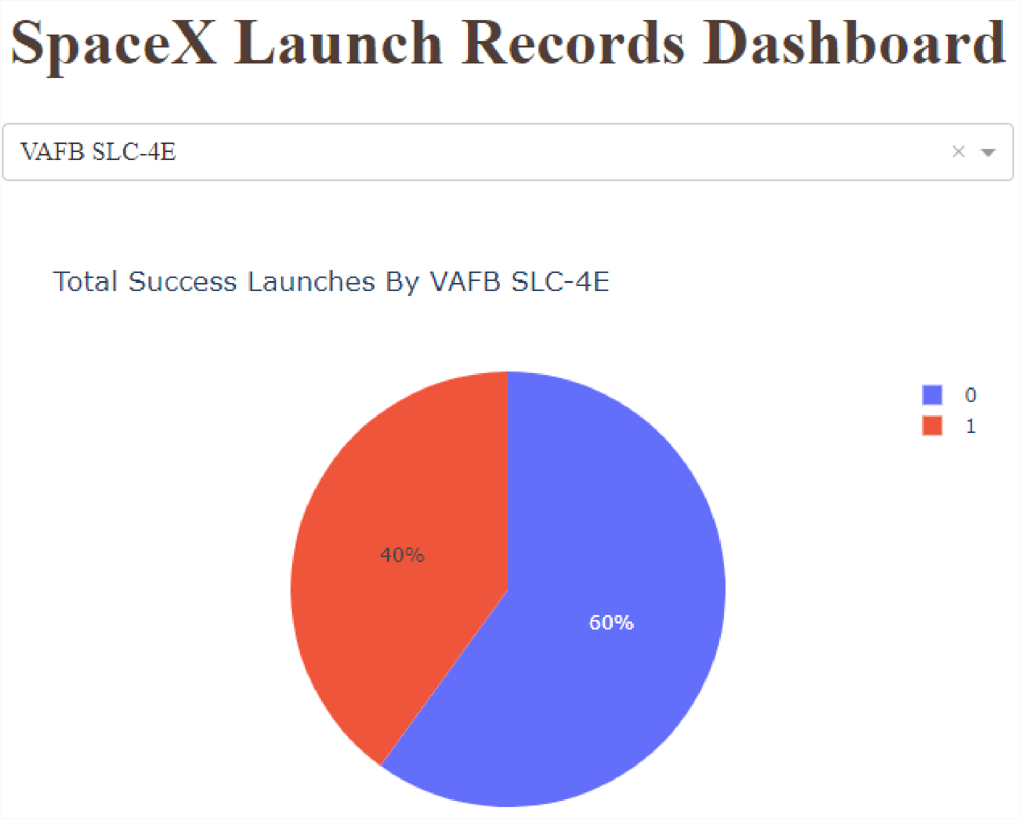
PLOTLY DASH DASHBOARD RESULTS





PLOTLY DASH DASHBOARD RESULTS





PLOTLY DASH DASHBOARD RESULTS

PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS

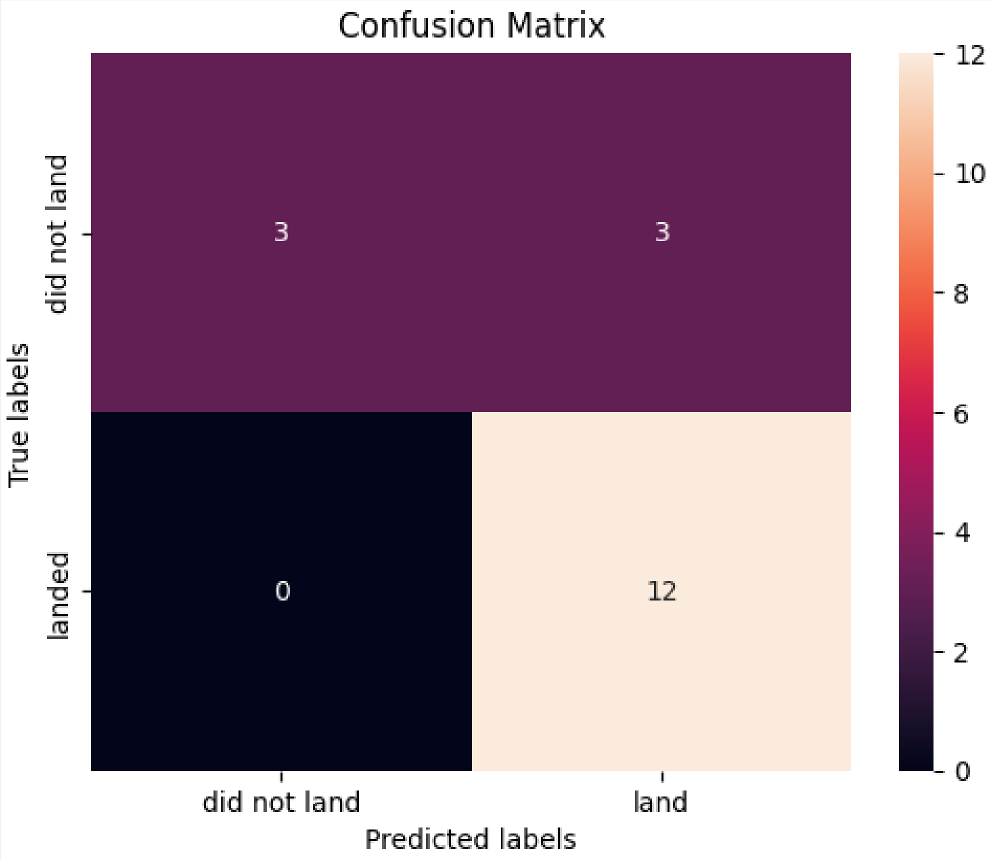
Logistic Regression Model

After defining and training our Logistic Regression Model the best parameters and the accuracy of the model on training data are:

tuned hpyerparameters: (best parameters) {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}

accuracy: 0.8464285714285713

The accuracy on the test data is: 0.8333333333333334



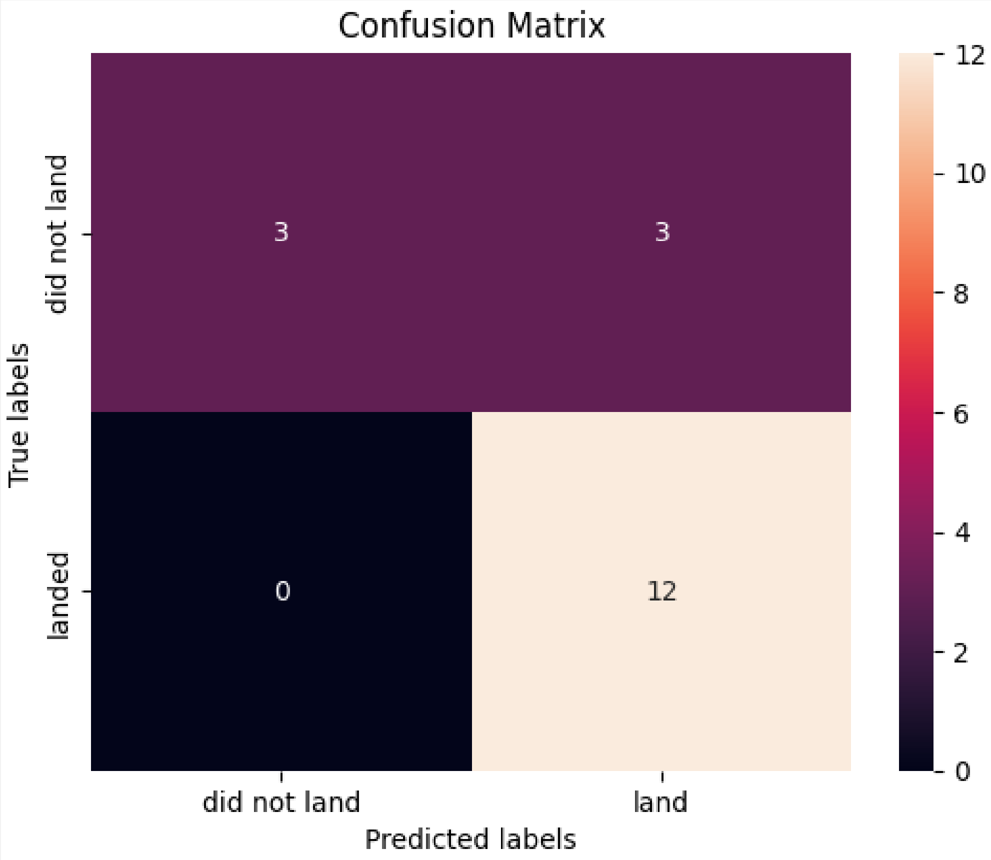
PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS

Support Vector Machine - SVM

After defining and training our Support Vector Machine Model the best parameters and the accuracy of the model on training data are:

tuned hpyerparameters: (best parameters) {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
accuracy: 0.8482142857142856

The accuracy on the test data is: 0.8333333333333334



PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS

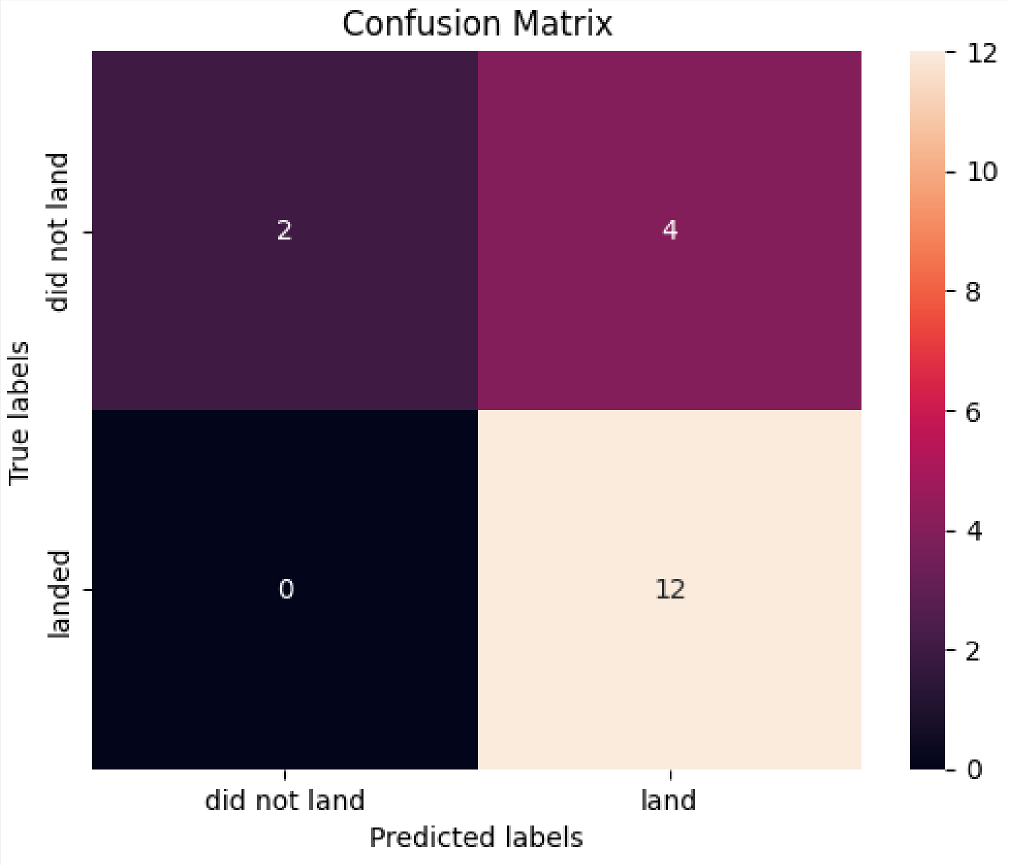
Decision Tree Classifier

After defining and training our Decision Tree Classifier the best parameters and the accuracy of the model on training data are:

tuned hpyerparameters: (best parameters) {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'random'}

accuracy: 0.8625

The accuracy on the test data is: 0.7777777777777778



PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS

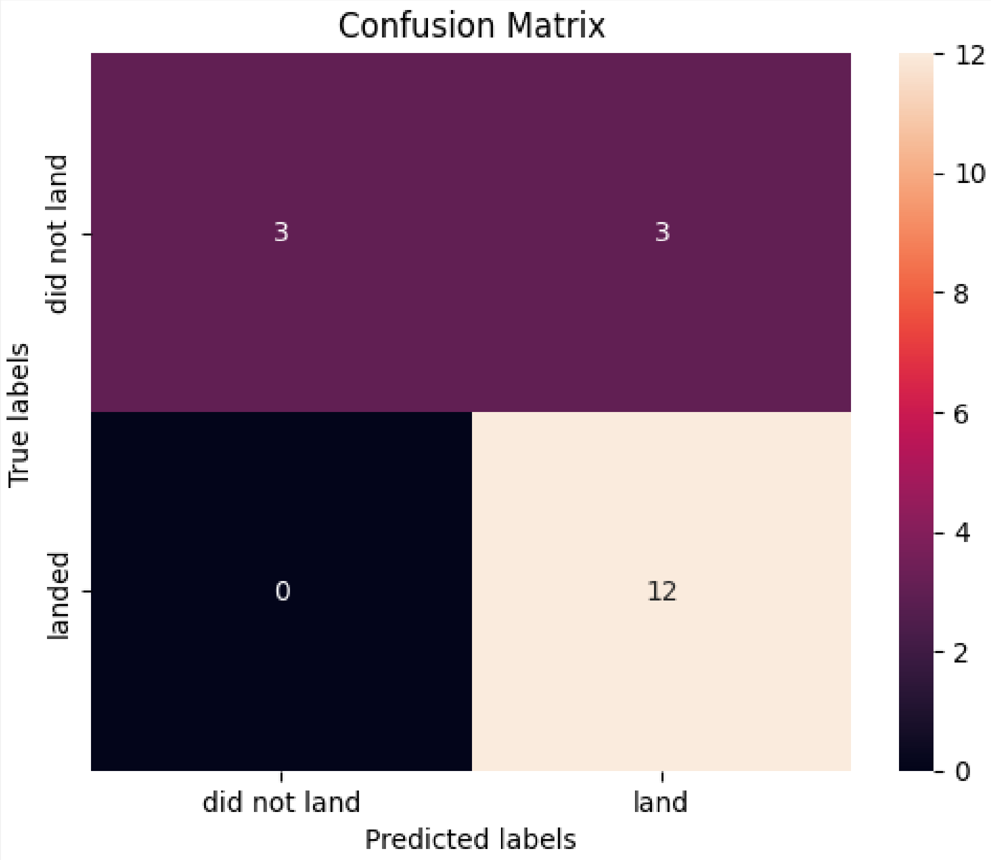
K Nearest Neighbors

After defining and training our K Nearest Neighbors Model the best parameters and the accuracy of the model on training data are:

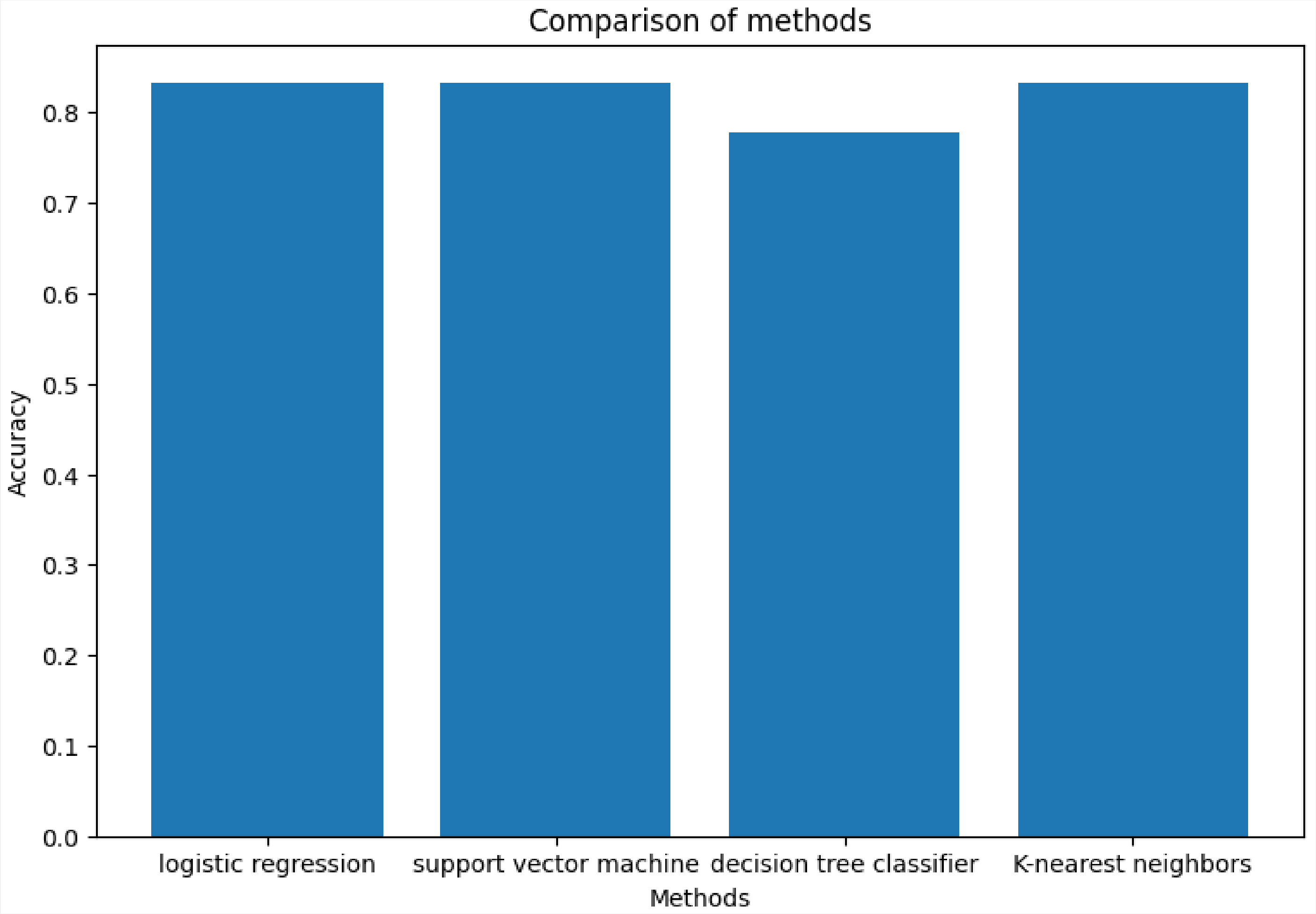
tuned hpyerparameters: (best parameters) {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}

accuracy: 0.8482142857142858

The accuracy on the test data is: 0.8333333333333334



PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS



Section 3

Conclusion

CONCLUSION

- Logistic Regression, SVM and K Nearest Neighbors models has the higher accuracy on test data: 0.8334, any of these models can correctly predict the landing of Stage 1 of the Falcon 9 rocket.
- Decision Tree Classifier has a lower accuracy rate, due to the purpose of this analysis it is advised not to use this model to predict.

APPENDIX

- The Jupyter NoteBooks about the analysis are in this GitHub repository:
  [rvdizeo/Determining_the_price_of_a_SpaceX_Falcon_9_rocket_launch (github.com)]