IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Eduardo J. Ramos C.
December 2022

Outline

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

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

Introduction

1. Project background and context

   Space X 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 Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

2. Problems you want to find answers

   - 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?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Via SpaceX Rest API	Web Scrapping from Wikipedia

- Perform data wrangling

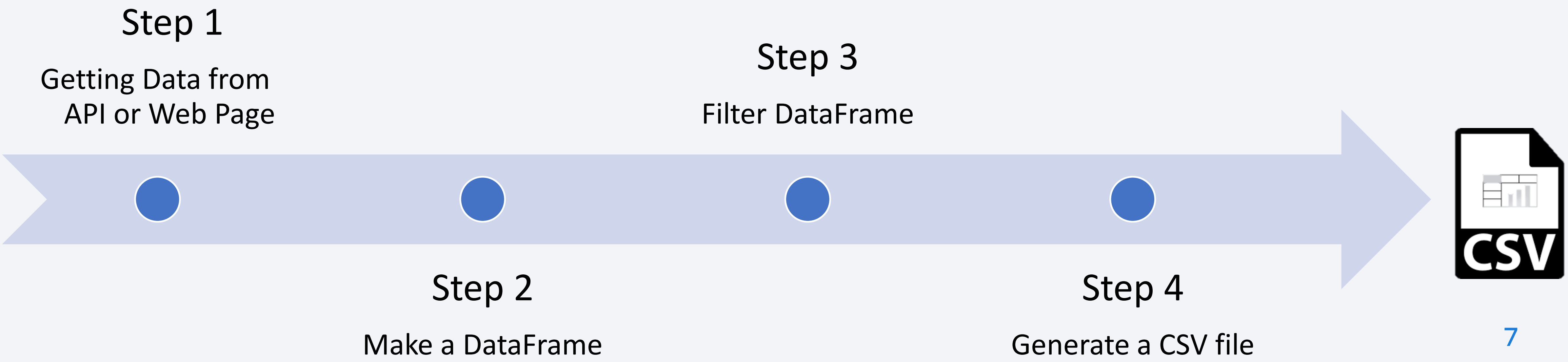
Filtering the data	Dealing with missing values
Using One Hot Encoding to prepare the data to a binary classification	

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Building, tuning and evaluation of classification models to ensure the best results

Data Collection

- Data collection is a process of grouping data from different sources and media in order to filter and obtain information that can be analyzed.
- 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

Requesting rocket launch data from SpaceX API → Decoding the response content using .json() and turning it into a dataframe using .json_normalize() → Requesting needed information about the launches from SpaceX API by applying custom functions → Filtering the dataframe to only include Falcon 9 launches → Creating a dataframe from the dictionary → Constructing data we have obtained into a dictionary → Exporting the data to CSV

In [38]:
```
data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
data_falcon9
```

4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0(
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0(
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0(
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1(
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1(
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1(
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1(
91	88	2020-	Falcon 9	15600.0	VLEO	KSC LC	True	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1(

GitHub URL

Data Collection - Scraping

Requesting Falcon 9 launch data from Wikipedia

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
response = requests.get(static_url)
```

Creating a BeautifulSoup object from the HTML response

```
soup=BeautifulSoup(response.text,"html.parser")
```

Extracting all column names from the HTML table header

```
html_tables=soup.find_all('table')
print(html_tables)
```

```
for columnas in first_launch_table.find_all("th",scope="col"):
    column_names.append(columnas.text.removesuffix("\n"))
```

Collecting the data by parsing HTML tables

```
extracted_row = 0
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
    for rows in table.find_all("tr"):
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
        else:
            flag=False
        row=rows.find_all('td')
        if flag:
            extracted_row += 1
            launch_dict['Flight No.'].append(flight_number)
```

Constructing data we have obtained into a dictionary

Creating a dataframe from the dictionary

```
df=pd.DataFrame(launch_dict)
```

Exporting the data to CSV

```
df.to_csv('spacex_web_scraped_modificado.csv', index=False)
```

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

GitHub URL

Data Wrangling

```
df["LaunchSite"].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Calculate the number of launches on each site

```
df["Orbit"].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

Calculate the number and occurrence of each orbit

```
landing_outcomes=df["Outcome"].value_counts()
```

```
print(landing_outcomes)
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

Calculate the number and occurrence of mission outcome per orbit type

```
df.head(5)
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004

Create a landing outcome label from Outcome column

Export dataset as .cvs

```
df.to_csv("dataset_part_2.csv", index=False)
```

GitHub URL

EDA with Data Visualization

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

- 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

Line Plot
Line plots show trends in data over time (time series)

- Success Rate Yearly Trend

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

- Payload Mass vs Orbit Type

GitHub URL

EDA with SQL

Performed SQL queries:

- *Names of the unique launch sites in the space mission*
- *5 records where launch sites begin with the string 'CCA'.*
- *The total payload mass carried by boosters launched by NASA (CRS).*
- *Average payload mass carried by booster version F9 v1.1.*
- *Date when the first succesful landing outcome in ground pad was acheived.*
- *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.*
- *Names of the booster_versions which have carried the maximum payload mass.*
- *The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.*
- *Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.*

GitHub URL

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.

GitHub URL

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.

GitHub URL

Predictive Analysis (Classification)

Creating a NumPy array from the column "Class" in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

GitHub URL

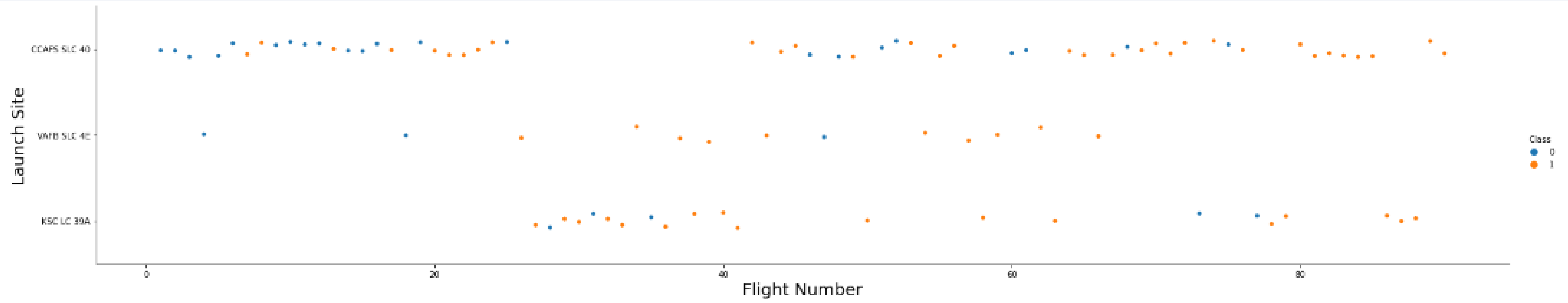
Results

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

Section 2

Insights drawn from EDA

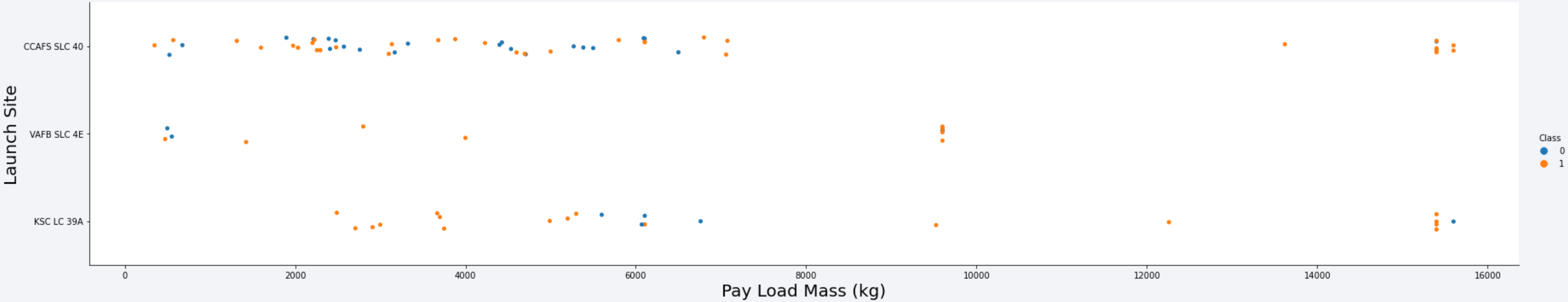
Flight Number vs. Launch Site



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

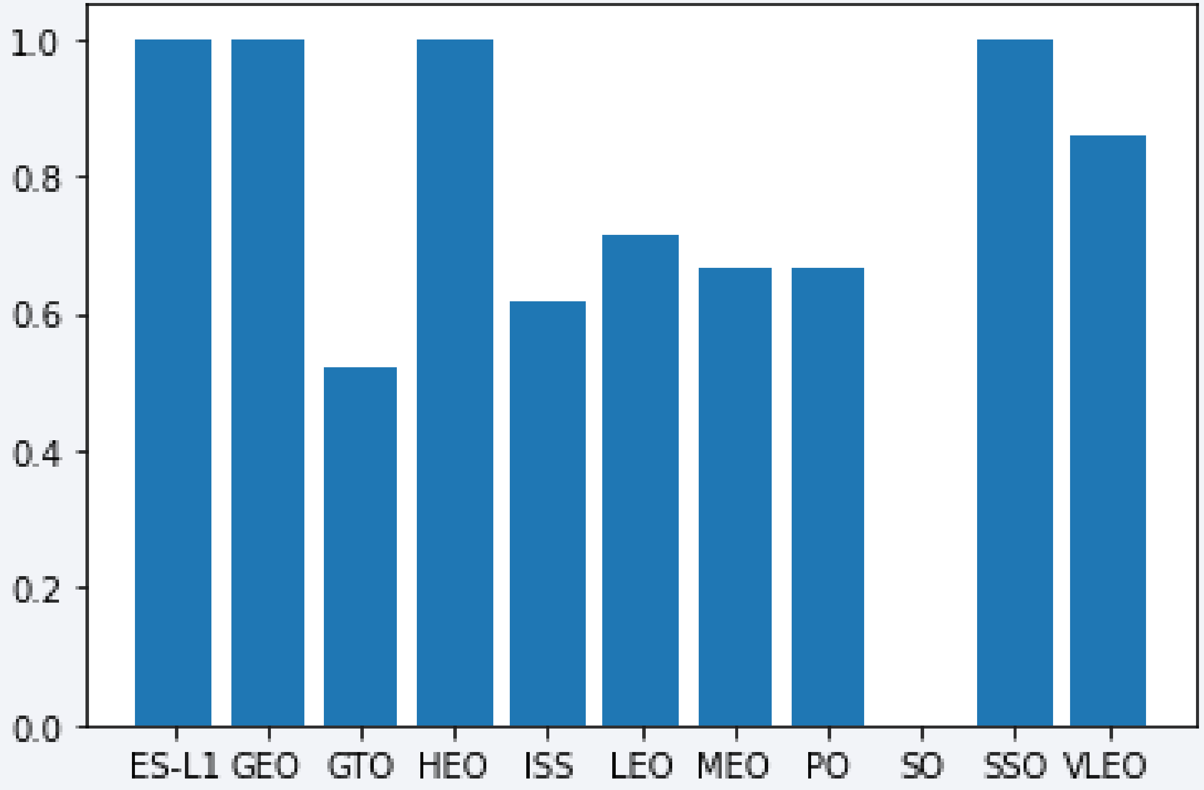


Explanation:

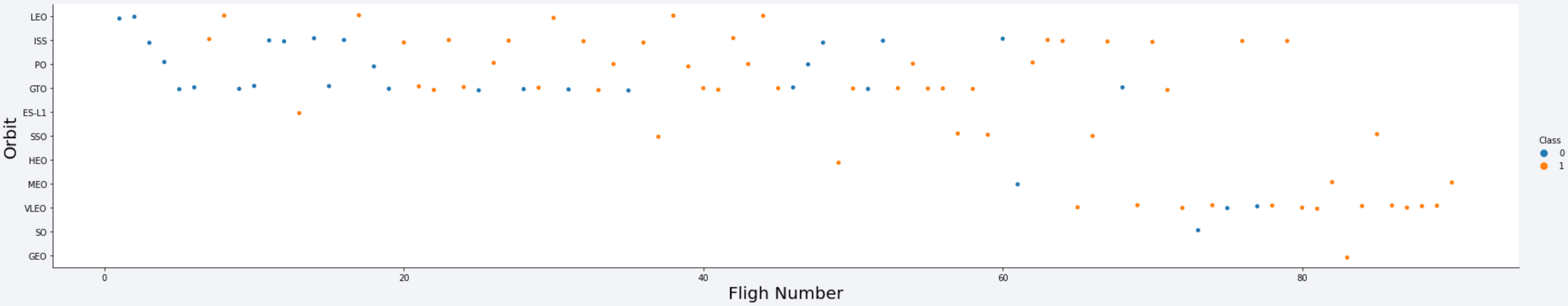
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

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:

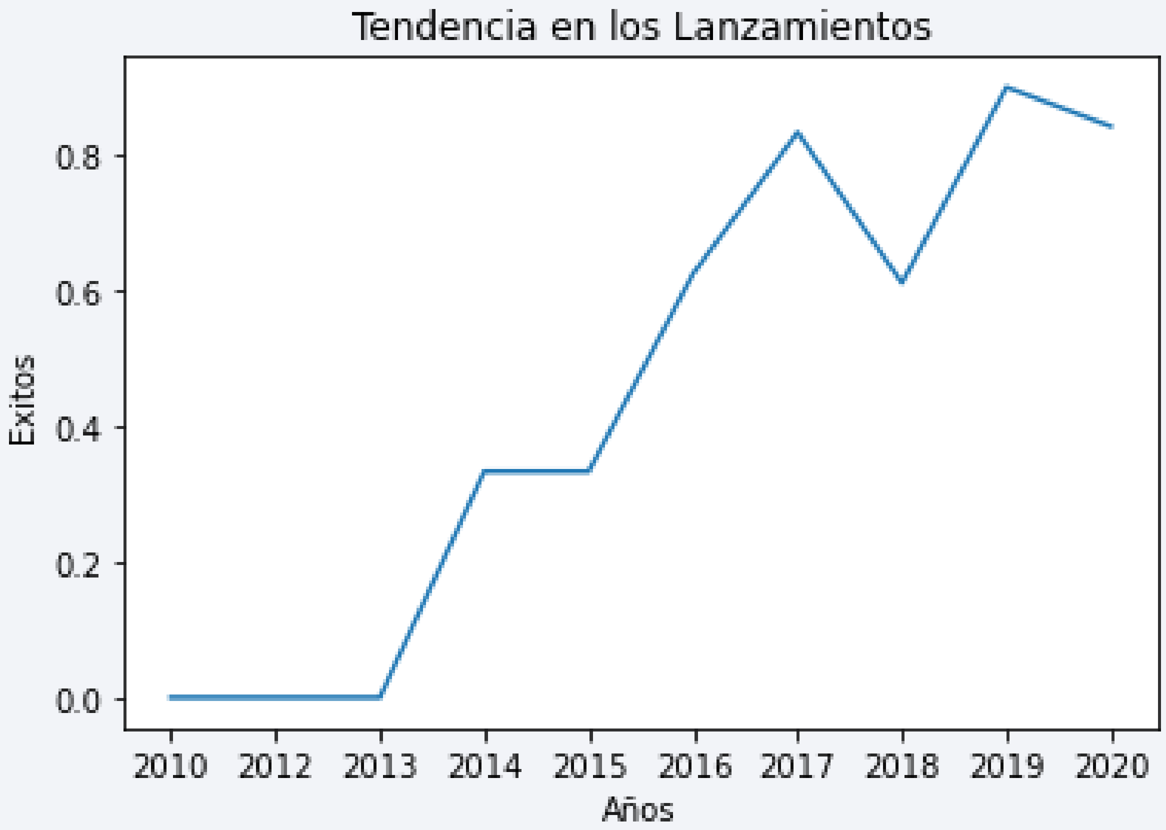
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload 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 DISTINCT "Launch_Site", count("Launch_Site") FROM SPACEXTBL GROUP by "Launch_Site"
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site	count(Launch_Site)
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

Explanation:

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

Launch Site Names Begin with 'CCA'

```
# SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE "CCA%" limit 5   # Funciona en el sqlite browser
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE "CCA%" limit 5;
```

...

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	
1	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	Spa
2	08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of ...	0	LEO (ISS)	NAS
3	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NAS
4	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NAS
5	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NAS

Explanation:

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

Total Payload Mass

```
# SELECT sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Payload like "%crs%"   # Funciona en el sqlite browser
%sql SELECT sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Payload like "%crs%"
```

...

total_payload_mass
45596

Explanation:

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

Average Payload Mass by F9 v1.1

```
# SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE "%F9 v1.1%"   # Funciona en el sqlite browser
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE "%F9 v1.1%"
```

AVG(PAYLOAD_MASS__KG_)
2534.66666666667

Explanation:

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

First Successful Ground Landing Date

```
# select min(Date) as first_successful_landing from SPACEXTBL WHERE "Landing _Outcome" like 'Success (ground pad)';"
%sql select min(Date) as first_successful_landing from SPACEXTBL WHERE "Landing _Outcome" like ='Success (ground pad)';
```

...

first_successful_landing
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

```sql
SELECT "Booster_Version","PAYLOAD_MASS__KG_" from SPACEXTBL WHERE "Landing _Outcome" like \n
'Success (drone ship)' AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000;
```

	Booster_Version	PAYLOAD_MASS__KG_
1	F9 FT B1022	4696
2	F9 FT B1026	4600
3	F9 FT B1021.2	5300
4	F9 FT B1031.2	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

```
%sql select "Mission_Outcome",count(Mission_Outcome) from SPACEXTBL group by "Mission_Outcome"
```

	Mission_Outcome	count(Mission_Outcome)
1	Failure (in flight)	1
2	Success	99
3	Success (payload status unclear)	1

Explanation:

Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql select Booster_Version , PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_=(select max("PAYLOAD_MASS__KG_") \n
                                                                                          from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Explanation:

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

2015 Launch Records

SQL 1 (1) | SQL 1 | SQL 3 | SQL 4

```sql
SELECT Date,substr(Date,4,2) as month,launch_site,"Booster_Version","Landing _Outcome"
from (SELECT Date,substr(Date,4,2) as mes,launch_site,"Booster_Version","PAYLOAD_MASS__KG_","Landing _Outcome"
from SPACEXTBL where "Landing _Outcome" like "%failure (drone ship%") where substr(Date,7,4)="2015"
```

	Date	month	launch_site	Booster_Version	Landing _Outcome
1	10-01-2015	01	CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
2	14-04-2015	04	CCAFS LC-40	F9 v1.1 B1015	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.

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXTBL
      where date between '2010-06-04' and '2017-03-20'
      group by 'landing__outcome' order by count_outcomes desc;
```

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

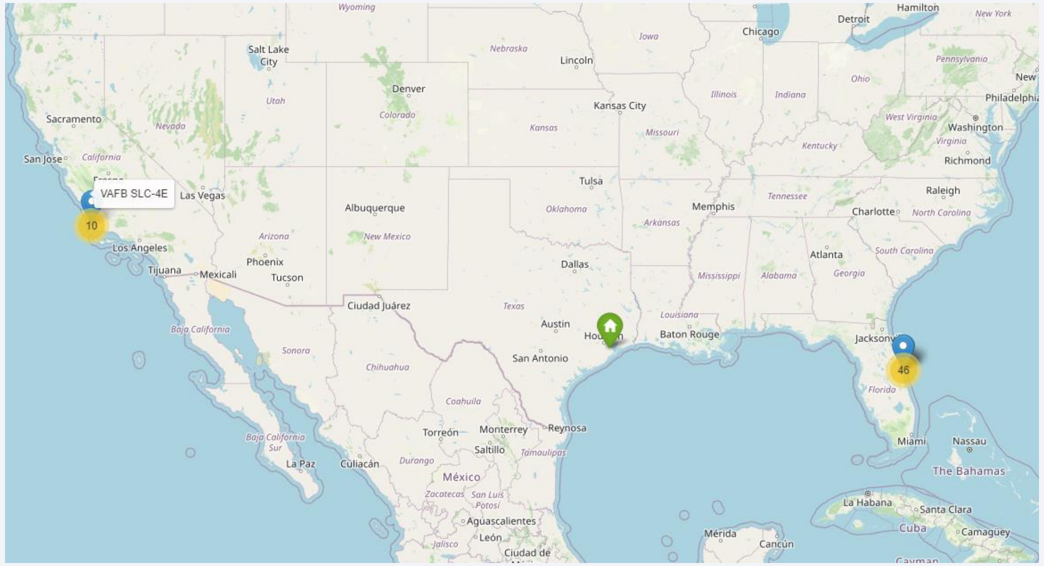
Explanation:

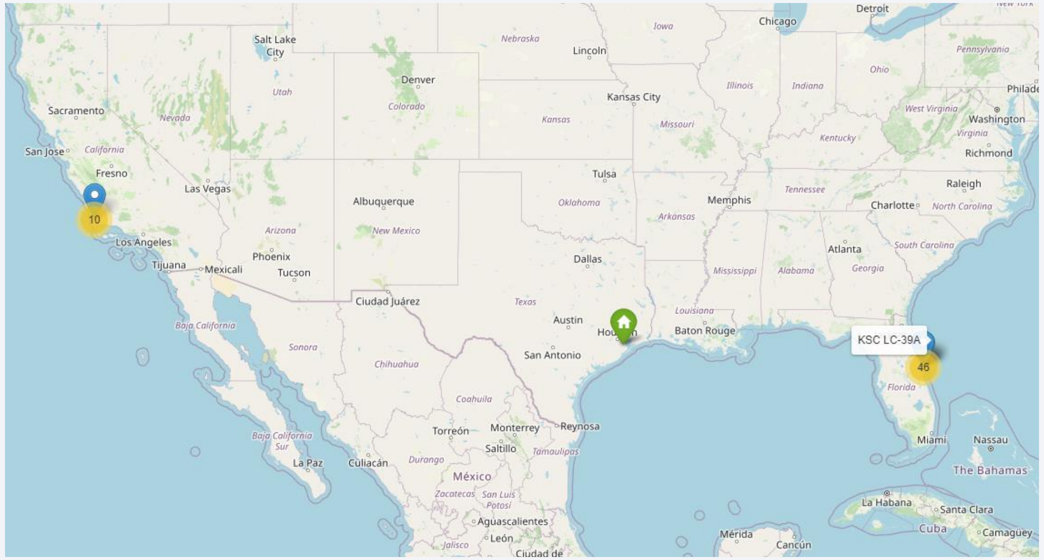
Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

Section 3

Launch Sites Proximities Analysis

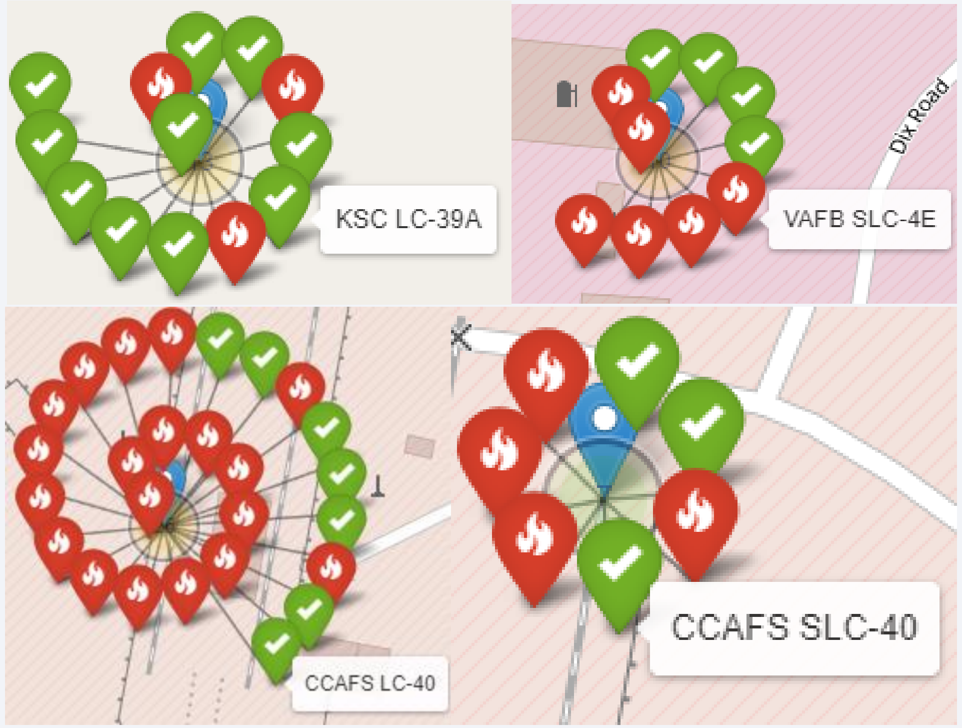
All Launch Sites on Folium Map





The ideal launch sites are near the equator and near coastlines. As can be seen on the map, the three launch sites are as far south as the US geography allows and very close to its coasts.

All launch sites' location markers



Green Marker shows successful launches and Red Marker shows failures.

From these screenshots its easily understandable that KSC LC-39A has the maximum probability of success

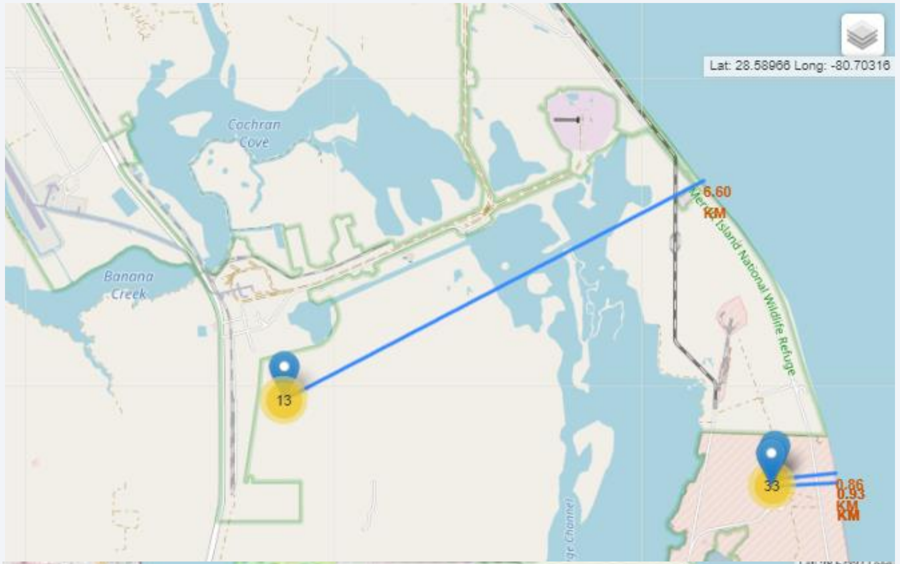
Distance from the launch site KSC LC-39A to its proximities

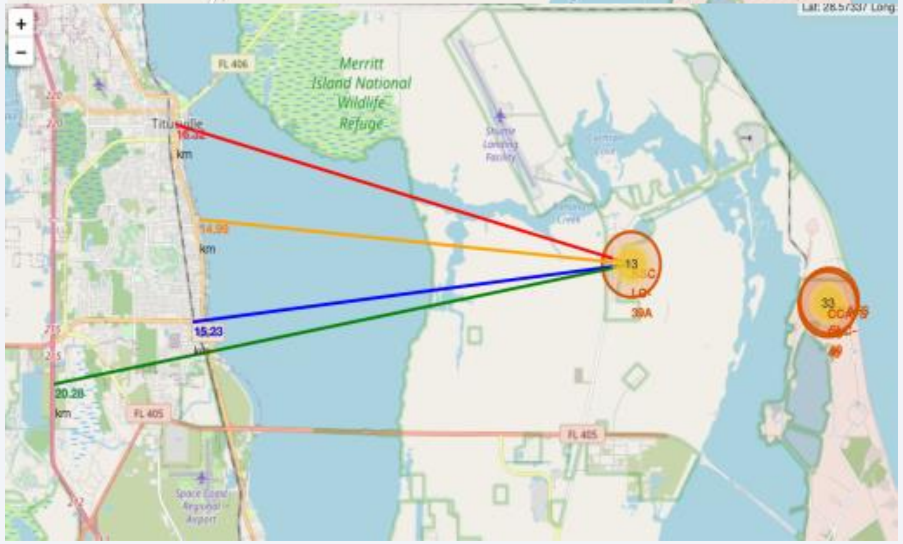
- Are launch sites in close proximity to railways?
    - Yes, distance < 2 Km

- Are launch sites in close proximity to highways?
    - No, distance < 15 km

- Are launch sites in close proximity to coastline?
    - Yes, distance < 5 Km

- Do launch sites keep certain distance away from cities?
    - Yes, distance > 15 km

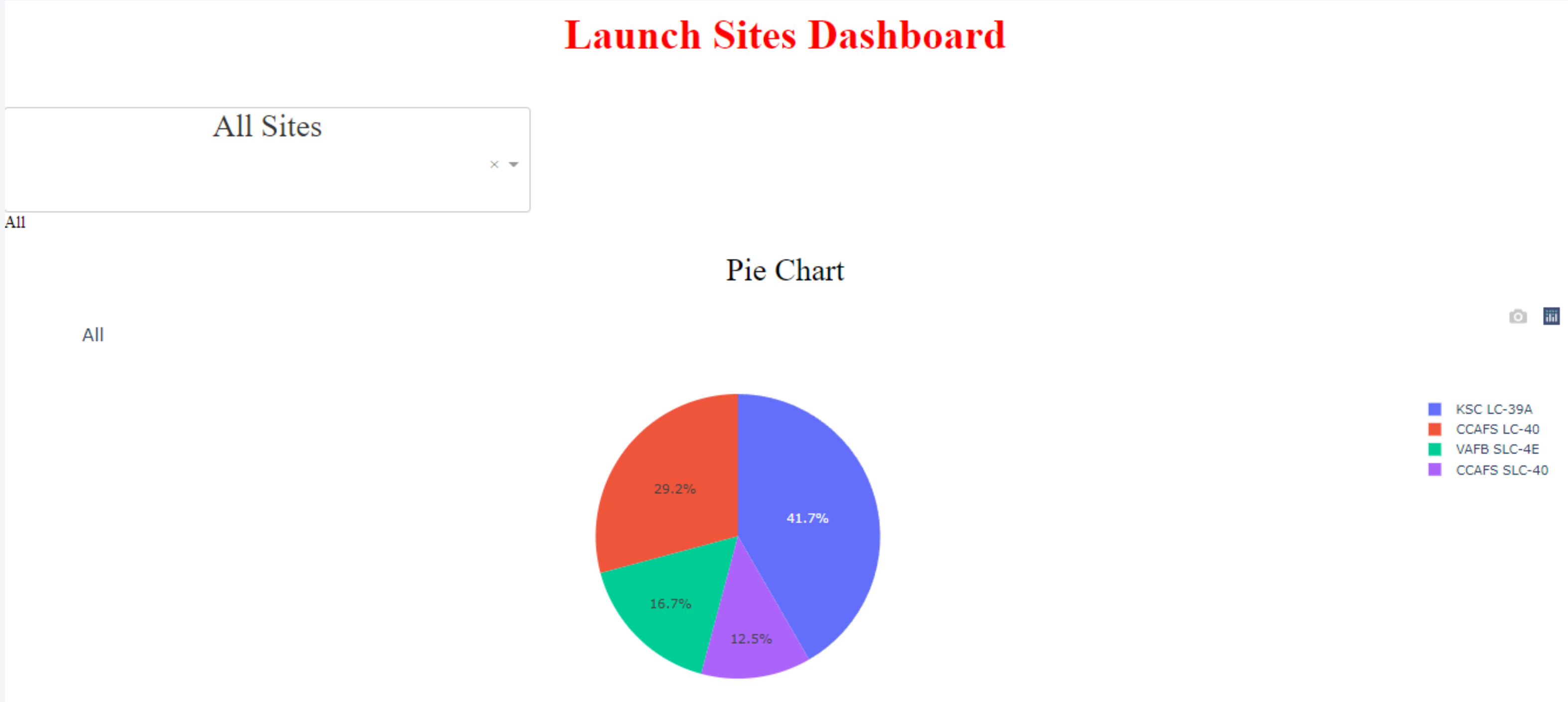




Section 4

Build a Dashboard with Plotly Dash

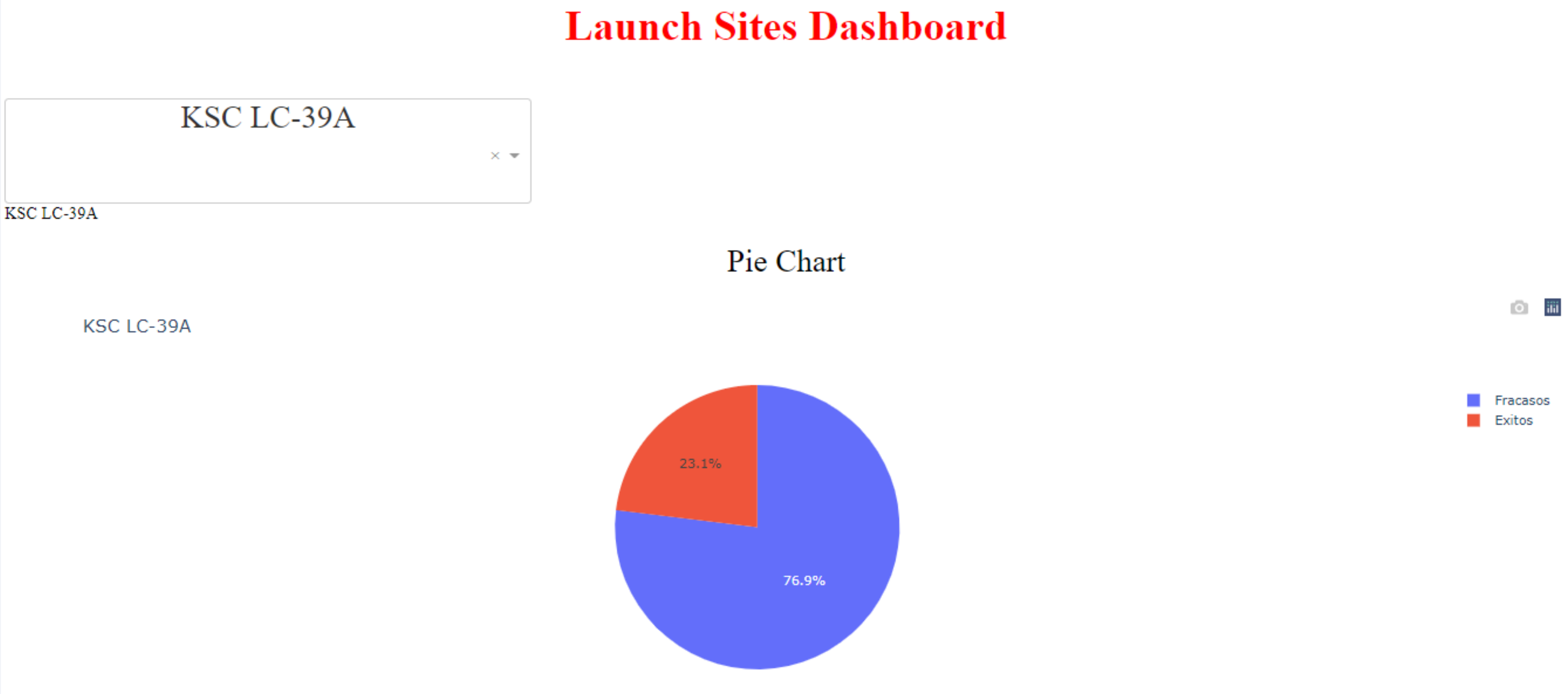
Launch Sites Dashboard



Explanation:

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

Launch site with highest launch success ratio



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 sites

Explanation:

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

Section 5

Predictive Analysis (Classification)

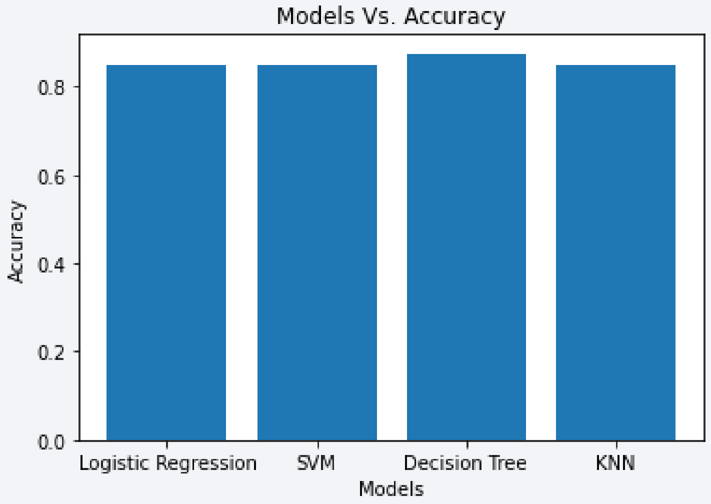
Classification Accuracy

In [76]:

```
model=['Logistic Regression','SVM','Decision Tree','KNN']
accuracy=[logreg_cv.score(X_test,Y_test),svm_cv.score(X_test,Y_test),tree_cv.score(X_test,Y_test),knn_cv.score(X_test,Y_
accuracy_models=[logreg_cv.best_score_,svm_cv.best_score_,tree_cv.best_score_,knn_cv.best_score_]
accuracy=[logreg_cv.score(X_test,Y_test),svm_cv.score(X_test,Y_test),tree_cv.score(X_test,Y_test),knn_cv.score(X_test,Y_
resultados=pd.DataFrame(accuracy_models,model).reset_index().rename(columns={'index':'Model',0:'Accuracy'})
resultados
```

Out[76]:

	Model	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	Decision Tree	0.875000
3	KNN	0.848214



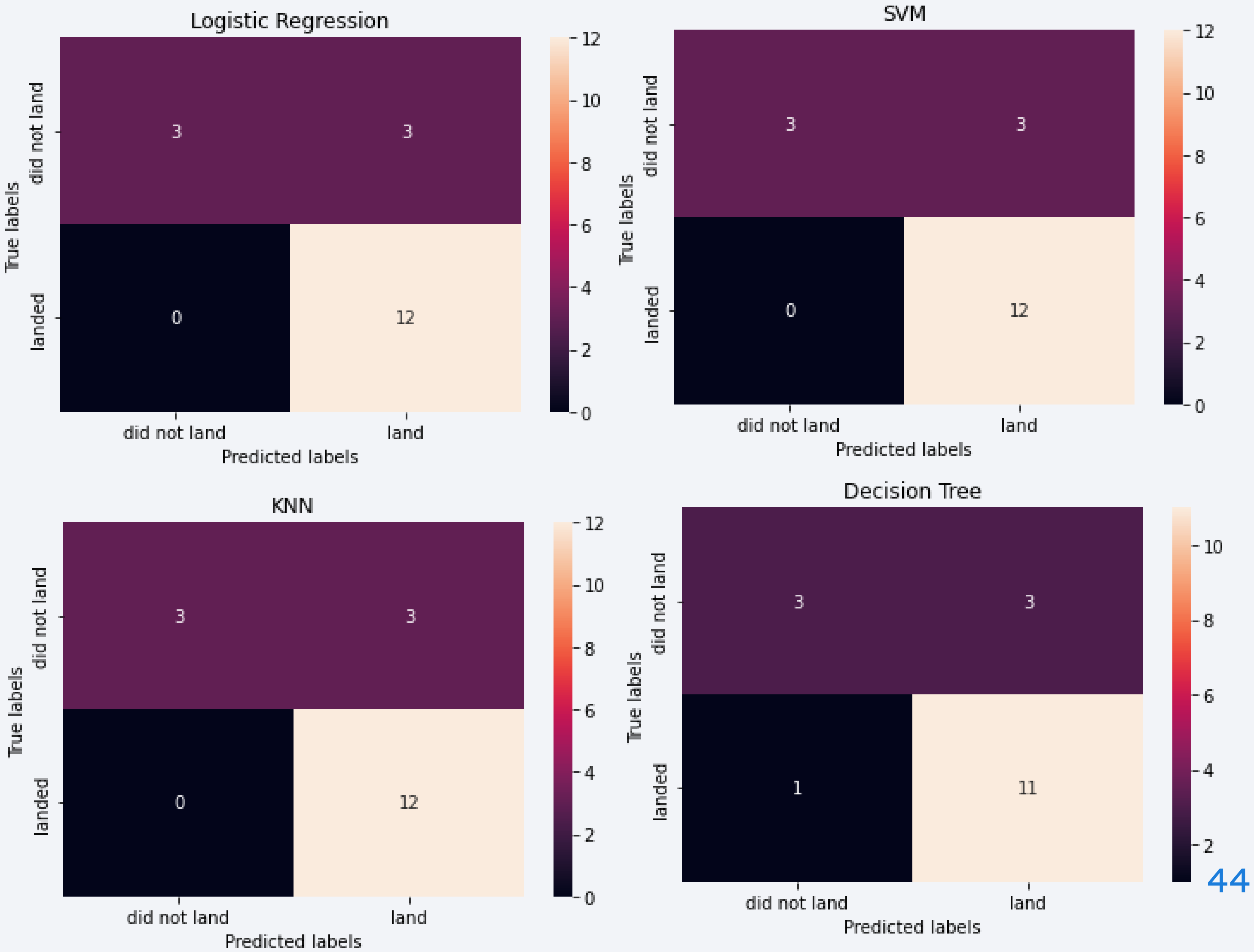
Explanation:

- 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.

Confusion Matrix

Explanation:

- Examining the confusion matrix, we see that Decision Tree can distinguish between the different classes.



Conclusions

- 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.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- 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.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!