Winning Space Race with Data Science

S. Evans
08.03.2023

Outline

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

Executive Summary

Summary of methodologies:

- Collection of data from public SpaceX API and by scrapping SpaceX Wikipedia page using Beautiful soup library for it.
- Creation of column ‘Class’ as binary dummy variable for categorical landing outcome.
- Performing EDA (Exploratory Data Analysis) with SQL using the concepts of magic sql.
- Exploring queries using distinct, like, group by as well as aggregate functions and nested queries.
- Performing Data Analysis using Pandas and Matplotlib as well as seaborn.
- Create scatter-, line- and bar-plots for Visualization.
- Creation of a folium map with markers for the launch sites, marker cluster for launch success as well as poly lines. and marks to show distances of landmarks on the map.
- Build up of a Dashboard application with plotly dash containing a drop-down input for launch site and a slider for payload mass. Show results in pie-chart and scatterplot.
- Standardizing the data set.
- creating prediction models, perform a grid search for optimum parameters, train and fit the models, Determination of model accuracy by method score.

Executive Summary

Summary of all results:

The success rate of the landing outcome of ‘stage one’ depends on:

- launch site
- payload
- orbit type
- launch year,
- landing facility
- booster version

4 prediction models which were built, trained and run on a data set considering all the influencing issues. All models have the same accuracy of 83.3%.

The models predict a landing success rate of 66,7%.

Introduction

Business understanding

For the cost calculation of commercial space-transport it is important to know if the ‘first stage’ can be reused. Which means it has to be brought back without crashing. We support SpaceY by supplying answers to the following questions:

- What is the probability for the reuse of the ‘first stage’?
- What are the important factors influencing the landing outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection with API and Web Scraping using BeautifulSoup
- Perform data wrangling
  - Dealing with missing values, creation of dummy variable for categorical landing outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - 4 different models were parameter-optimized with grid search, trained and accuracy testes with score method.

Data Collection

2 types of data collection were used in this project:

Data collection with API:	Data collection with web scraping of related Wiki pages
The data collection from SpaceX includes information about: • the rocket used • payload delivered • launch specifications • landing specifications • landing outcome	The data collection from SpaceX includes information about: • Booster type • Launchpad type • payload mass • core

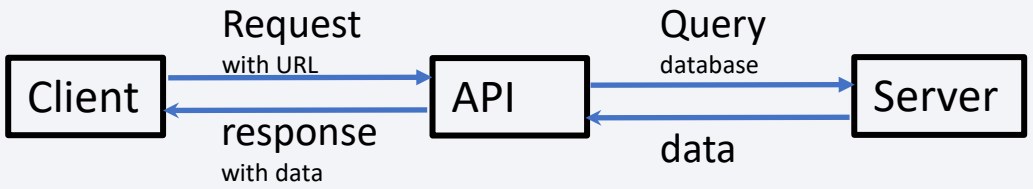
Data Collection – SpaceX API

The data collection process with API includes:

- use of URL to target a specific endpoint of the API
- perform a get request using the requests library
- result can be viewed by calling the .json() method.
- response will be in the form of a JSON
- convert the JSON to a data frame using the json normalize-function
- Using the API again targeting another endpoint
- Chose columns, store data in lists and create data set
- Skip unwanted data

https://github.com/sse2bue/Applied_Data_Science_Capstone/blob/main/1_jupyter-labs-spacex-data-collection-api.ipynb

Flow chart of API request



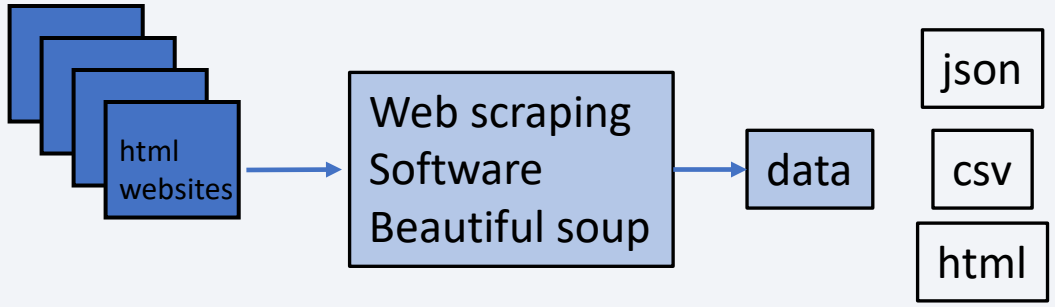
Data Frame:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	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

Data Collection - Scraping

The data collection process web scraping includes:

- using the Python BeautifulSoup package to web scrape some HTML
- parse the data from those tables and convert them into a Pandas data frame



[14]:

	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	Sirius XM	Success\n	F9 v1.0B0003.1	Success	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NaN	NaN	F9 v1.0B0004.1	NaN	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NaN	NaN	F9 v1.0B0005.1	NaN	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NaN	NaN	F9 v1.0B0006.1	NaN	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NaN	NaN	F9 v1.0B0007.1	NaN	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	NaN	NaN	F9 B5B1051.10	NaN	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	NaN	NaN	F9 B5B1058.8	NaN	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	NaN	NaN	F9 B5B1063.2	NaN	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NaN	NaN	F9 B5B1067.1	NaN	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	NaN	NaN	F9 B5	NaN	6 June 2021	04:26

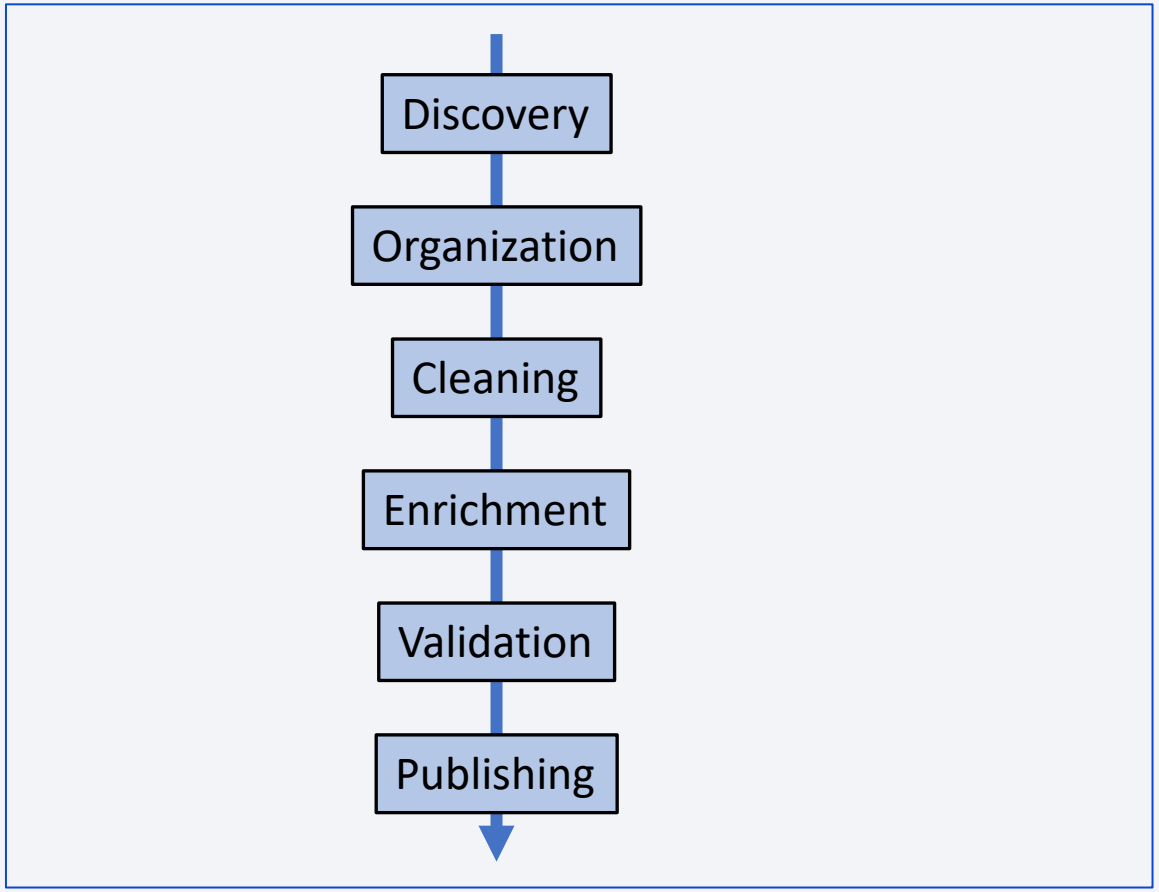
121 rows × 11 columns

https://github.com/sse2bue/Applied_Data_Science_Capstone/blob/main/2_jupyter-labs-webscraping.ipynb

Data Wrangling

The data wrangling includes:

- Identify missing values
- Deal with 0 values, here in payload mass
- Create dummy variables for categorical values (class for landing outcome)



https://github.com/sse2bue/Applied_Data_Science_Capstone/blob/main/3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Overview of charts and purpose

Chart	Purpose
• Scatter plot of Flight Number vs Launch Site overlaying class	• To see the relationship of flight numbers and launch sites as well as the success
• Scatter plot of Payload mass vs Launch Site overlaying class	• To see the relationship of payload mass and launch sites as well as the success
• Bar plot of mean of Class per orbit type	• To compare the success rates of the different orbits
• Scatter plot of Flight Number vs Orbit overlaying class	• To see if the success is related to the flight number or the orbit as well as the success
• Scatter plot of Payload mass vs Orbit overlaying class	• To see if the success is related to the payload mass or the orbit as well as the success
• Line plot of launch success rate vs year	• To see how the success rate changes as time goes on

https://github.com/sse2bue/Applied_Data_Science_Capstone/blob/main/5_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Overview of queries

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters having success in drone ship and payload mass between 4000 and 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List records which display month names, failure landing outcomes in drone ship, booster versions and launch site for the months 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

https://github.com/sse2bue/Applied_Data_Science_Capstone/blob/main/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Overview of map objects and purpose

Chart	Purpose
• Folium marker and circle • Folium circle and color-labeled markers in marker cluster • Add mouse position • Add a distance marker • Draw poly lines	• To mark launch sites on the map with a circle and show the name • To show number of launches and zoom in to see how many were successful or unsuccessful • To get input data for distance calculation • To show the calculated distance between 2 selected points • To show a connecting line between 2 chosen distance points

https://github.com/sse2bue/Applied_Data_Science_Capstone/blob/main/6_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

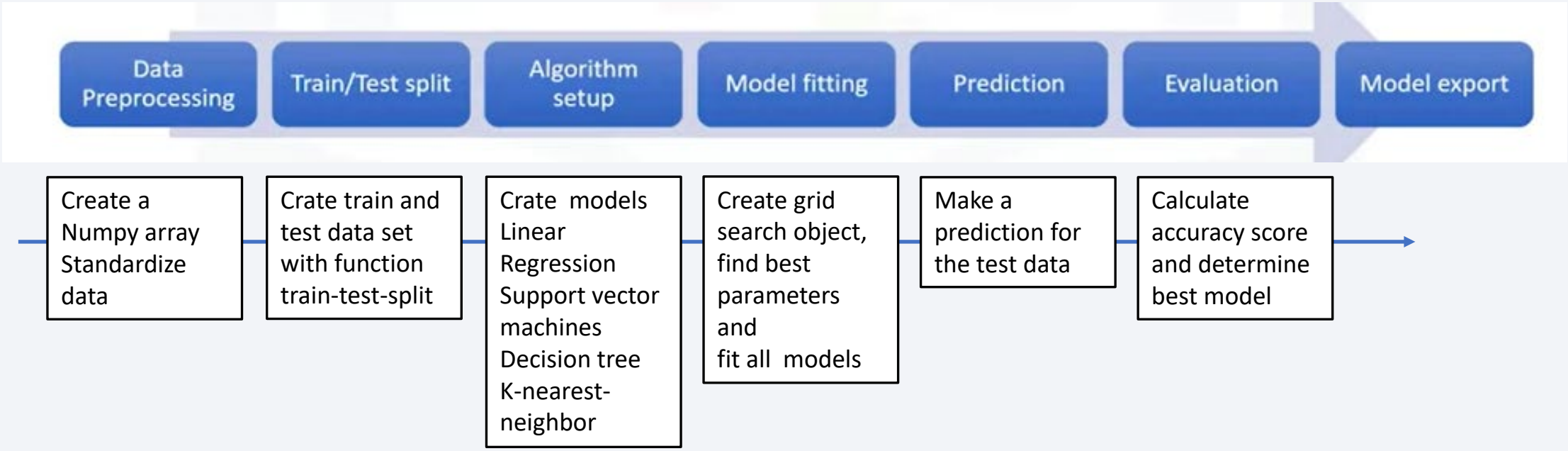
Overview of plots and interactions

Plots/graphs and interactions	Purpose
• Drop down menu for launch sites	• To select every single launch site or all sites
• Pie chart showing success rate of landing	• To show the success rate of the landings for every single launch site or all sites
• Slider for payload	• To select a payload range
• Scatter plot of landing success for various booster types over the payload	• To show the success for various booster types over the payload

https://github.com/sse2bue/Applied_Data_Science_Capstone/blob/main/7_spacex_dash_app.py

Predictive Analysis (Classification)

Model development process



4 models were built: logistical regression, vector support machine, decision tree and k-nearest neighbor.

https://github.com/sse2bue/Applied_Data_Science_Capstone/blob/main/8_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

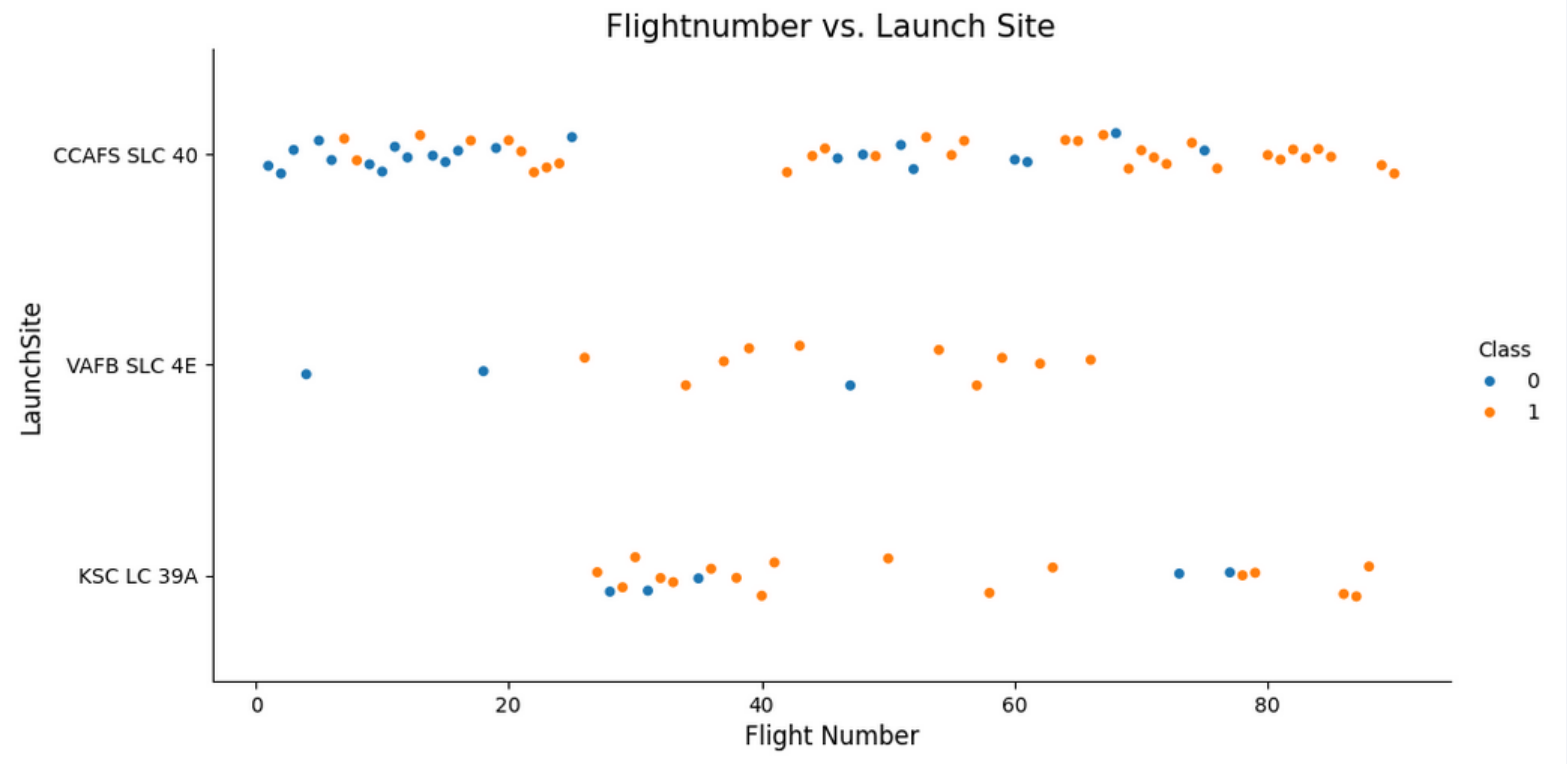
Results

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

Section 2

Insights drawn from EDA

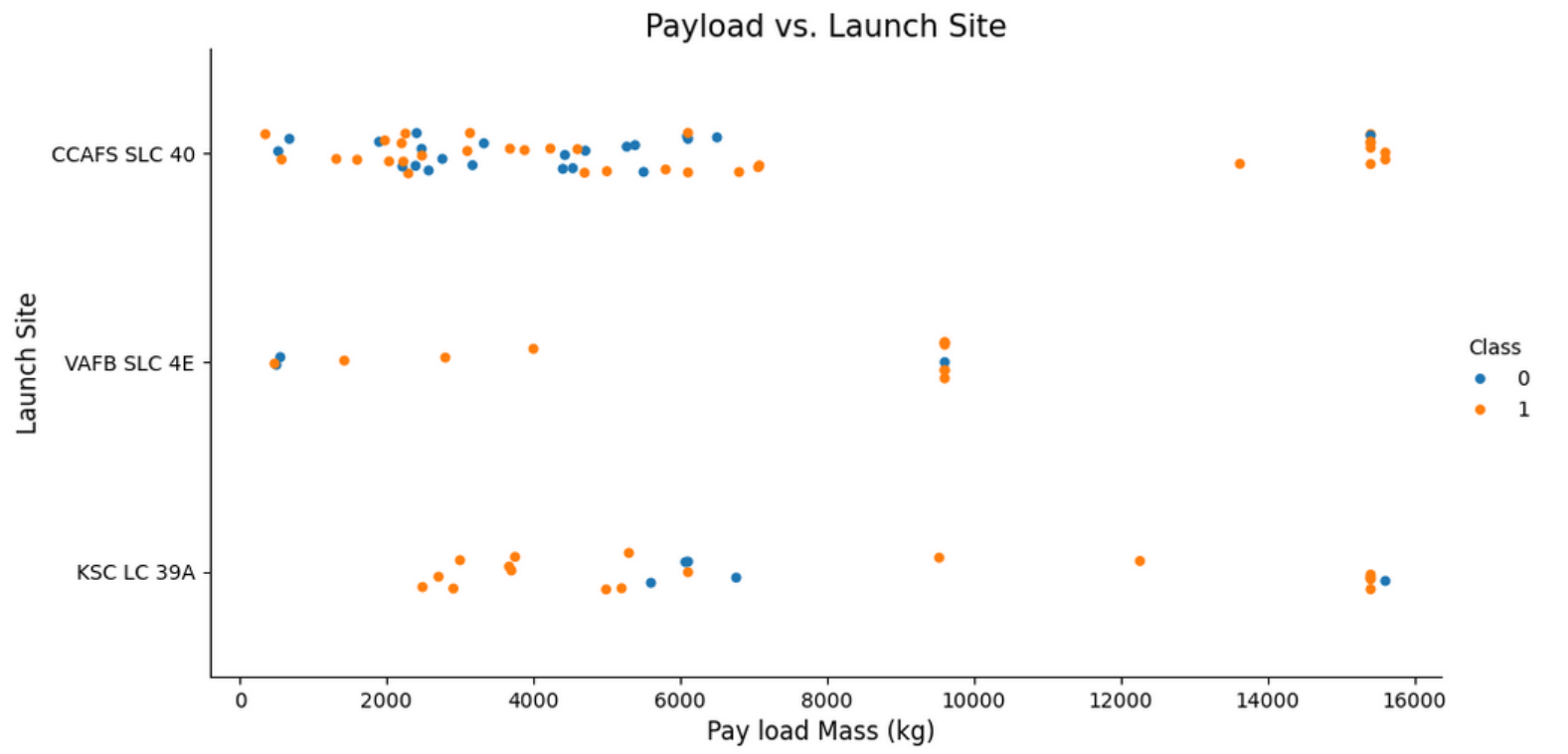
Flight Number vs. Launch Site



There are 3 launch sites:

- **CCAFS SLC 40:**
  has the most launches,
  at the beginning the success rate is low
- **VAFB SLC 4E:**
  has not so many launches but a high success rate
- **KSC LC 39A:**
  has a good success rate

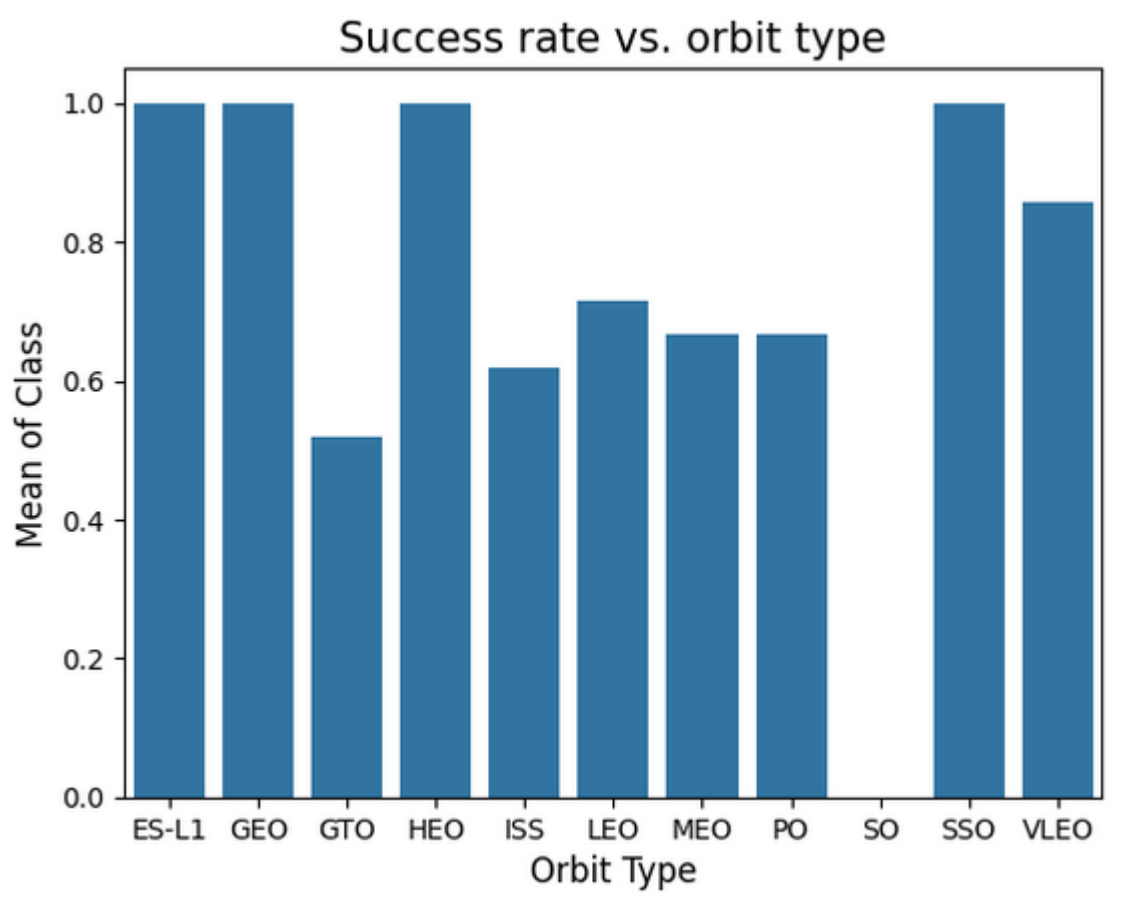
Payload vs. Launch Site



There are 3 launch sites:

- **CCAFS SLC 40:**
  has the most launches
- **VAFB SLC 4E:**
  has no payloads above 10000kg
- **KSC LC 39A:**
  has a lot of successful launches

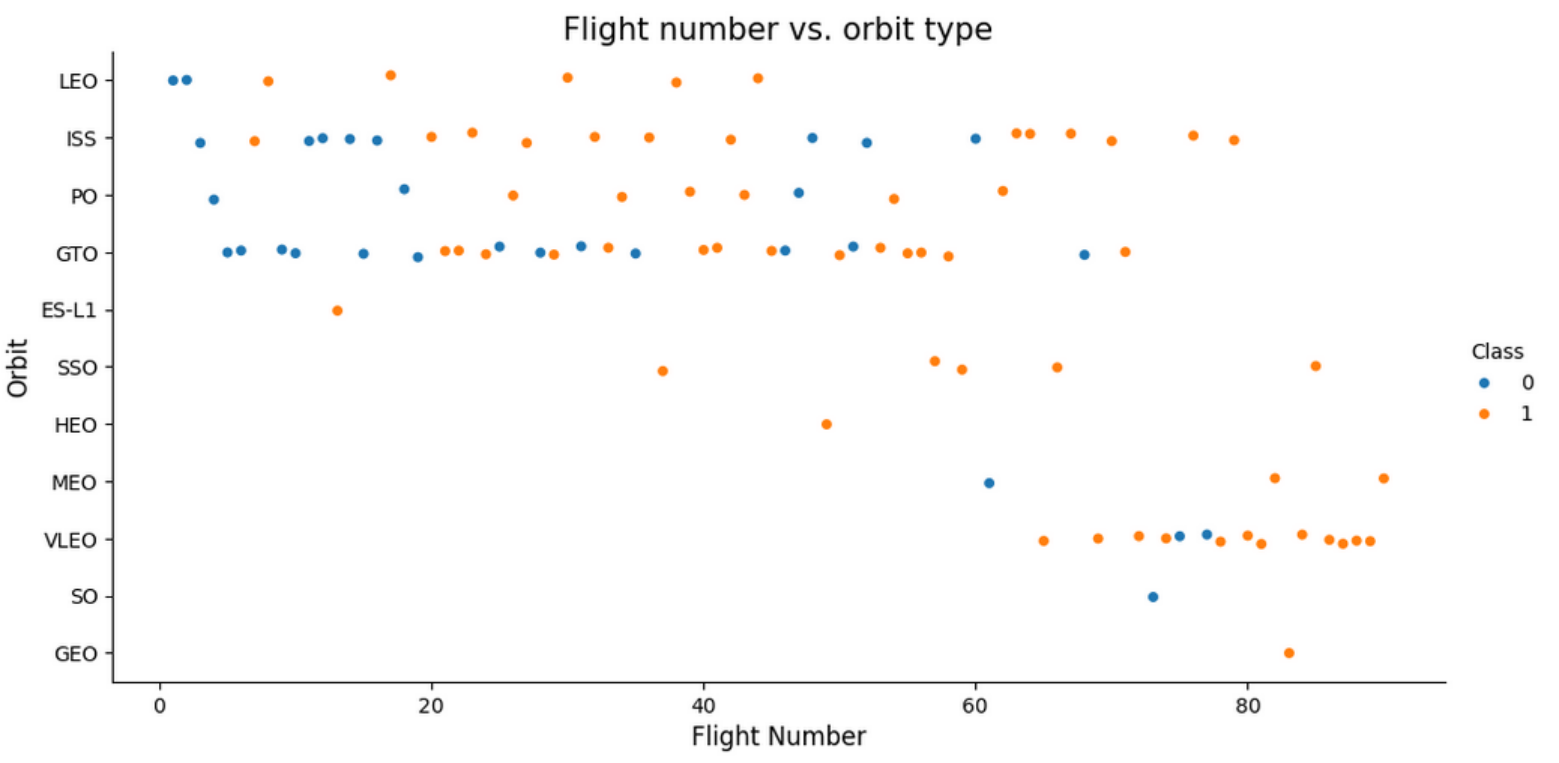
Success Rate vs. Orbit Type



There are 11 orbit types

- **100% success rate for:**
  ES-L1, GEO,HEO and SSO
- **0 successful landing:**
  SO

Flight Number vs. Orbit Type

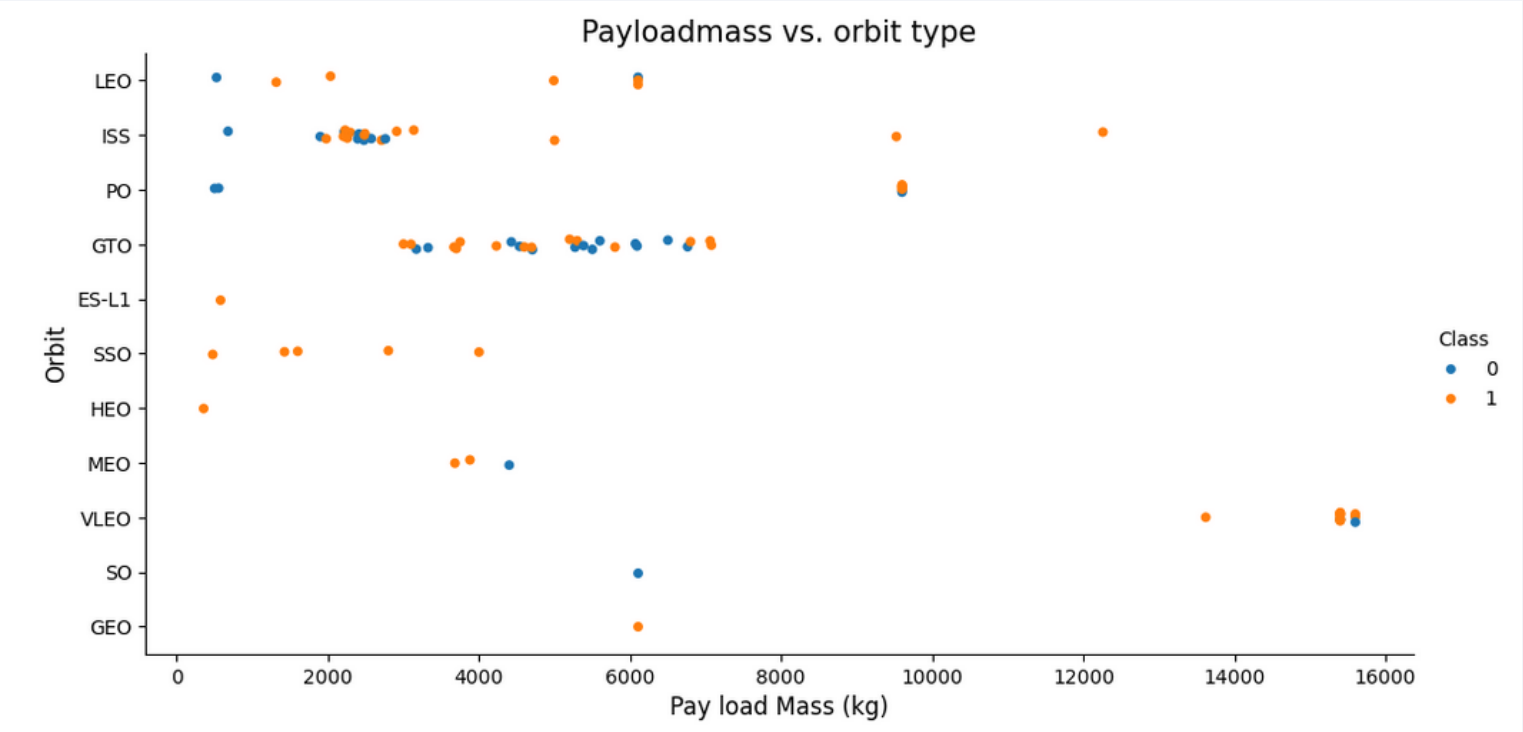


- **Relation ship of flight number and success for orbits**

- **Strong :**
  LEO, MEO
- **Medium:**
  ISS,PO,VLEO
- **none:**
  GTO
  SSO
- **Not enough data**
  ES-L1,HEO,SO, GEO

For orbits LEO and MEO the success seams to be related to the flight numbers. For Orbit GTO there seems no relationship from the data. For ES-L1, HEO,SO, GEO orbits there is not enough data to say something.

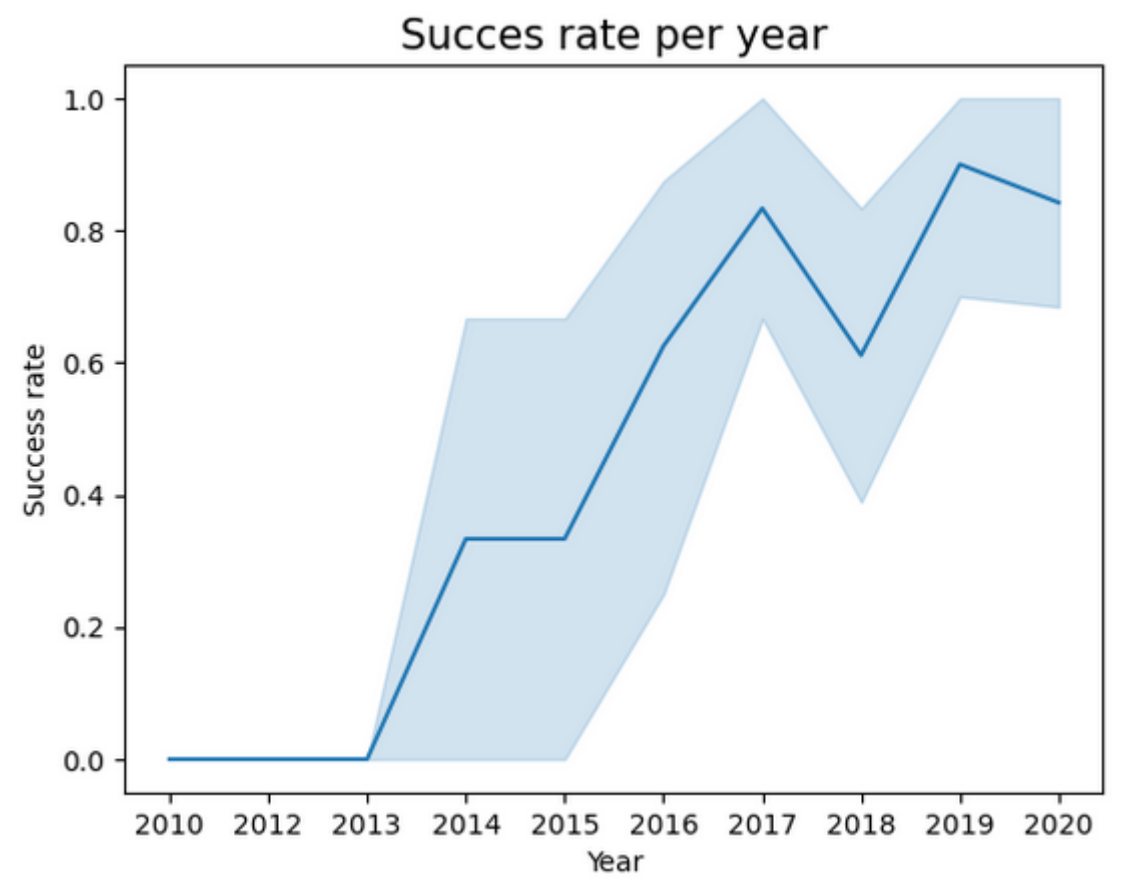
Payload vs. Orbit Type



- **ISS orbit:**
  Is most used for payloads 2000-3000kg.
- **GTO orbit:**
  Is most used for payloads 3000-7000kg.

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.

Launch Success Yearly Trend



- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations
- the success rate since 2013 kept increasing till 2020

All Launch Site Names

With sql selection of distinct (unique) Launch Sites, 4 different names were found in the data set (spaceX).

```
%sql SELECT Distinct Launch_Site FROM SPACEXTABLE
```

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

Launch Site Names Begin with 'CCA'

With sql selection search of string patterns with ‘like’ was done and the result limited to 5.

```
%sql Select* from SPACEXTABLE WHERE Launch_Site Like 'CCA%' LIMIT 5;
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

The first 5 launches of CCA sites took place in site CCAFS LC-40 between 2010 and 2013. The first was from SpaceX the other 4 from NASA.

Total Payload Mass

With the built-in-function select SUM the sum of all payload mass where the customer was NASA was calculated. NASA was searched for in the customer column by search of string patterns.

```
%sql Select SUM (PAYLOAD_MASS__KG_) as 'Total payload mass carried by boosters launched by NASA (CRS)' from \
(Select PAYLOAD_MASS__KG_ from SPACEXTABLE WHERE customer Like '%NASA (CRS)%');
```

Total payload mass carried by boosters launched by NASA (CRS)
48213

The total payload mass of all launches by NASA is 48213 kg.

Average Payload Mass by F9 v1.1

With the built-in-function select AVG the average payload mass of the booster Version ‘F9 v1.1’ was calculated. ‘F9 v1.1’ was searched for in the ‘Customer’ column by search of string patterns.

```
%sql Select AVG (PAYLOAD_MASS__KG_) as'Average payload mass carried by booster version F9 v1.1' from \
(Select PAYLOAD_MASS__KG_ from SPACEXTABLE WHERE Booster_Version Like '%F9 v1.1%');
```

Average payload mass carried by booster version F9 v1.1
2534.6666666666665

The total payload mass of all launches by F9 v1.1 is 2534.67 kg.

First Successful Ground Landing Date

With the built-in-function select MIN the first date of ‘success (ground pad)’ was searched. ‘success (ground pad)’ was searched for in the ‘Landing_Outcome’ column by of string patterns. The date was named ‘First successful landing’.

```
%sql Select MIN(Date) as 'First succesful landing' from SPACEXTABLE WHERE Landing_Outcome Like '%Success (ground pad)%'
```

First succesful landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

With the function ‘select’, 2 string pattern searches were combined with and.

```
%sql Select Booster_Version from SPACEXTABLE WHERE Landing_Outcome Like '%Success (drone ship)%'\
and Booster_Version IN(select Booster_Version from SPACEXTABLE WHERE PAYLOAD_MASS__KG_ Between 4000 and 6000);
```

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

Booster Versions with a payload mass between 4000 and 6000kg.

Total Number of Successful and Failure Mission Outcomes

With count the numbers of ‘Mission_Outcome’ = ‘Success’ was counted and with union the same procedure was added for ‘Failure’.

```
%sql Select Mission_Outcome, COUNT (Mission_Outcome) from SPACEXTABLE where Mission_Outcome = 'Success'\
union Select Mission_Outcome, COUNT (Mission_Outcome)AS COUNT from SPACEXTABLE where Mission_Outcome LIKE '%Failure%'
```

Mission_Outcome	COUNT (Mission_Outcome)
Failure (in flight)	1
Success	98

Of 99 flights 98 had a successful mission outcome.

Boosters Carried Maximum Payload

With a subquery all the ‘Booster_Versions’ from SPACEXTABLE were selected, where the payload mass was maximal.

```
%sql Select Distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select Max(PAYLOAD_MASS__KG_) from SPACEXTABLE) order by Booster_Version;
```

List of Booster Versions carrying maximal 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

2015 Launch Records

List of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015. Find day and year in date using substr(Date,…).

```
%sql Select substr(Date, 6, 2) as Month, Landing_Outcome,Booster_version, Launch_Site, Date from SPACEXTBL \
where Landing_outcome = 'Failure (drone ship)' and substr(Date,1,4)='2015'
```

Month	Landing_Outcome	Booster_Version	Launch_Site	Date
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

In 2015 there have been 2 failed landing outcomes on a drone ship. One in January and one in April.

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

The ‘Landing_Outcome’ was counted and grouped by ‘Landing_Outcome-Count’.

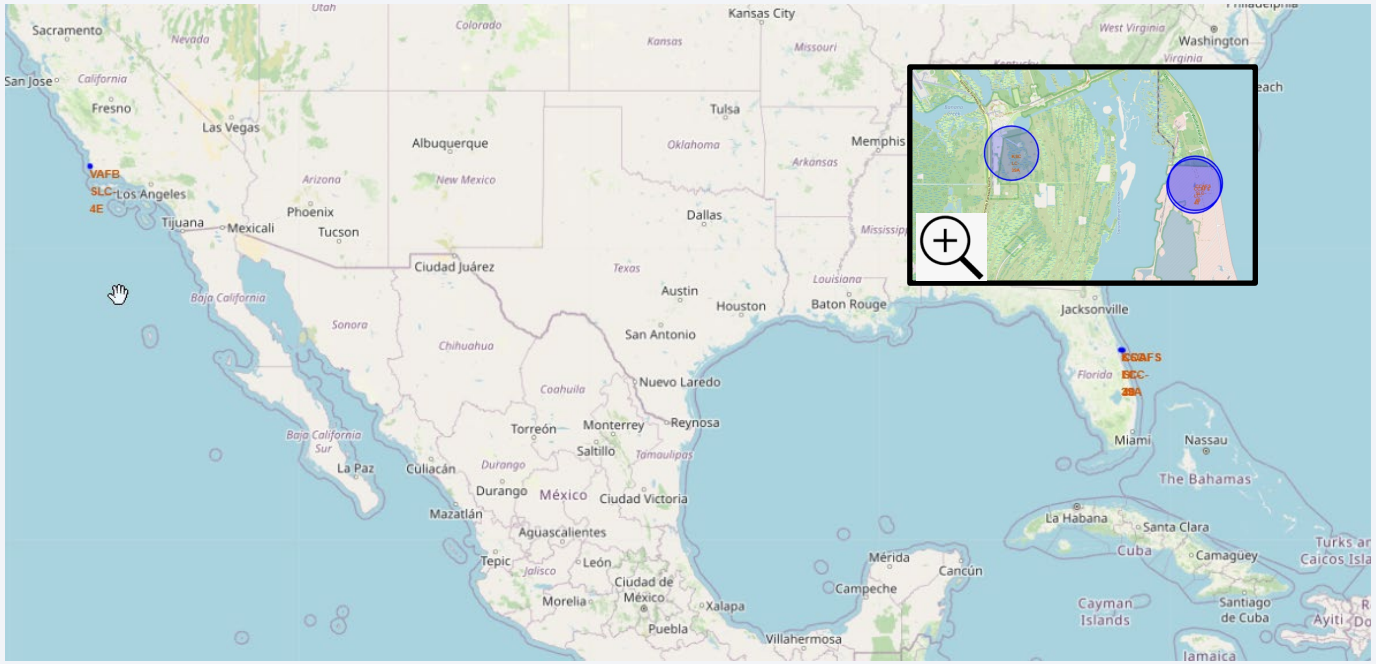
```
%sql SELECT Landing_Outcome, count(Landing_Outcome) as Landing_Outcome_Count, date FROM SPACEXTABLE\
where substr(Date,1,4)||substr(Date,6,2) || substr(Date,9,2) between'20100604'and '20170320'\
Group by Landing_Outcome order by Landing_Outcome_Count desc;
```

Landing_Outcome	Landing_Outcome_Count	Date
No attempt	10	2012-05-22
Success (drone ship)	5	2016-04-08
Failure (drone ship)	5	2015-01-10
Success (ground pad)	3	2015-12-22
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

- In 10 cases no landing was attempted.
- Most landing attempts were made on a drone ship. 5 out of 11 were successful.
- All 3 landing attempts on a ground pad were successful.
- Both landing attempts with parachute were unsuccessful.

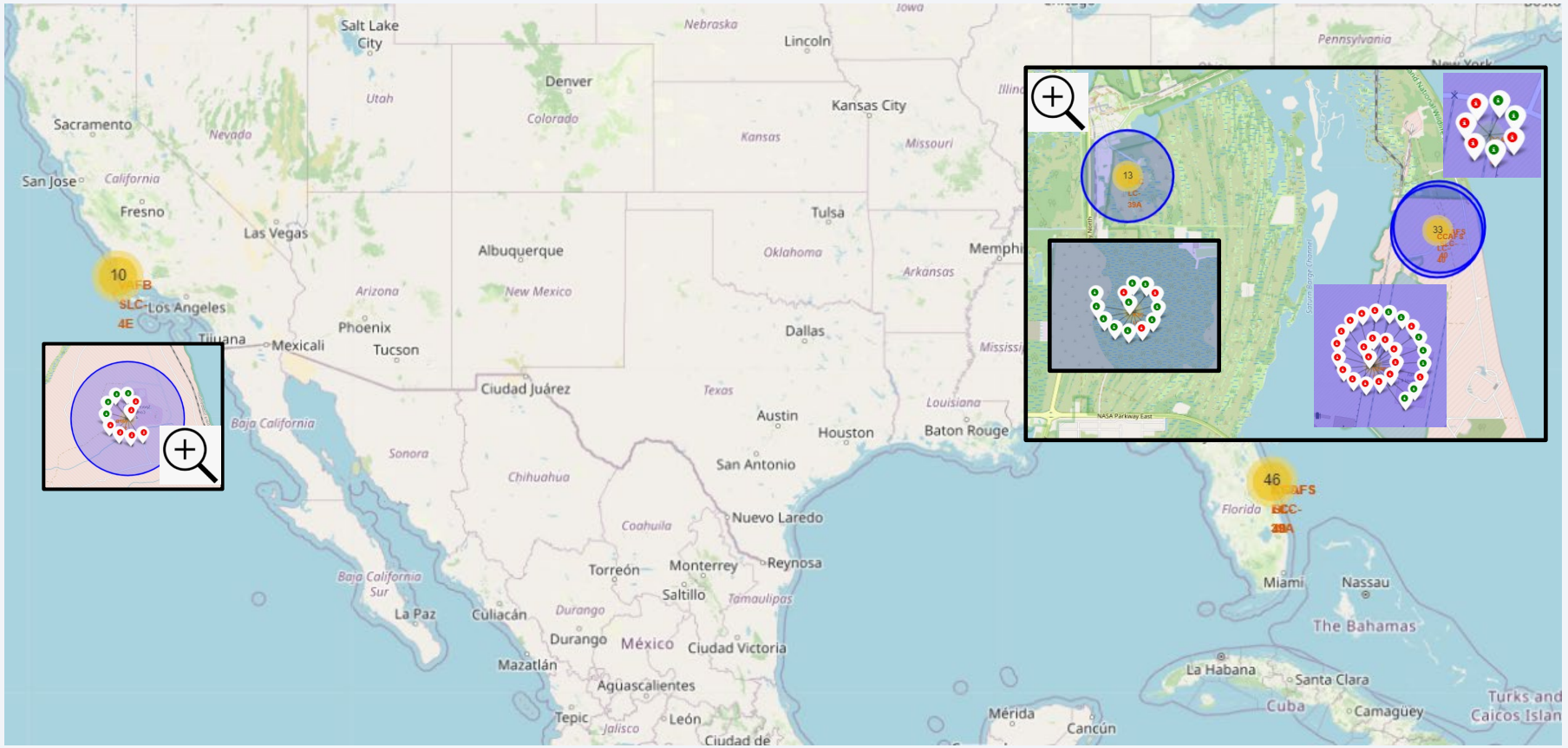

Section 3
Launch Sites
Proximities Analysis

Overview of SpaceX launch site locations



- There is one in California
- There are 3 in Florida
- All locations are by the coast
- All locations are south near the equator.
- Florida is the bigger place with more facilities

Overview of launches success per launch site



Landing outcome

Launch Site	Succesful launch	Unsuccessful launch
VAFB SLC-4E	4	6
KSC LC-39A	10	3
CCAFS LC-40	3	4
CCAFS SLC-40	7	19

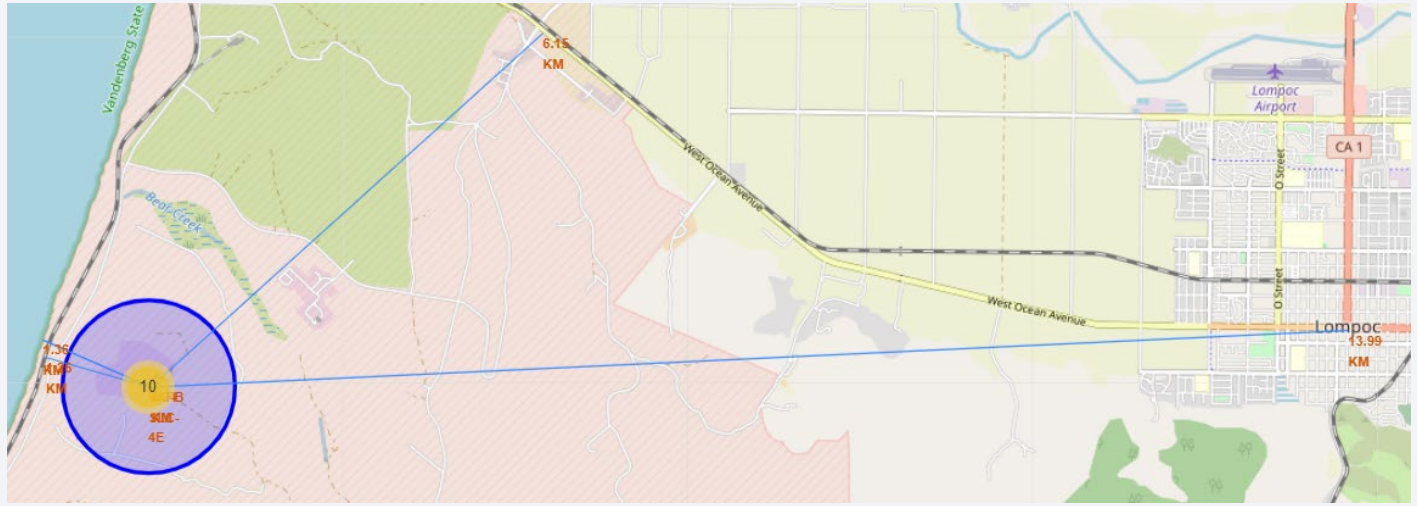
- CCAFS SLC-40 has the most launches
- KSC LC-39A has the most successful launches

Overview of data and distances per launch site

Overview of all SpaceX Launch sites

Launch Site	Lat	Long	Dist. Railway	Dist. Highway	Dist. Coastline	Dist. City	success probability
CCAFS LC-40	28.562302	-80.577356	1.259853	6.145998	1.358214	13.988559	0.269231
CCAFS SLC-40	28.563197	-80.576820	0.693568	0.975611	6.378817	16.264998	0.428571
KSC LC-39A	28.573255	-80.646895	1.219842	0.589605	0.871293	23.165479	0.769231
VAFB SLC-4E	34.632834	-120.610745	1.260355	0.656109	0.924816	23.139344	0.400000

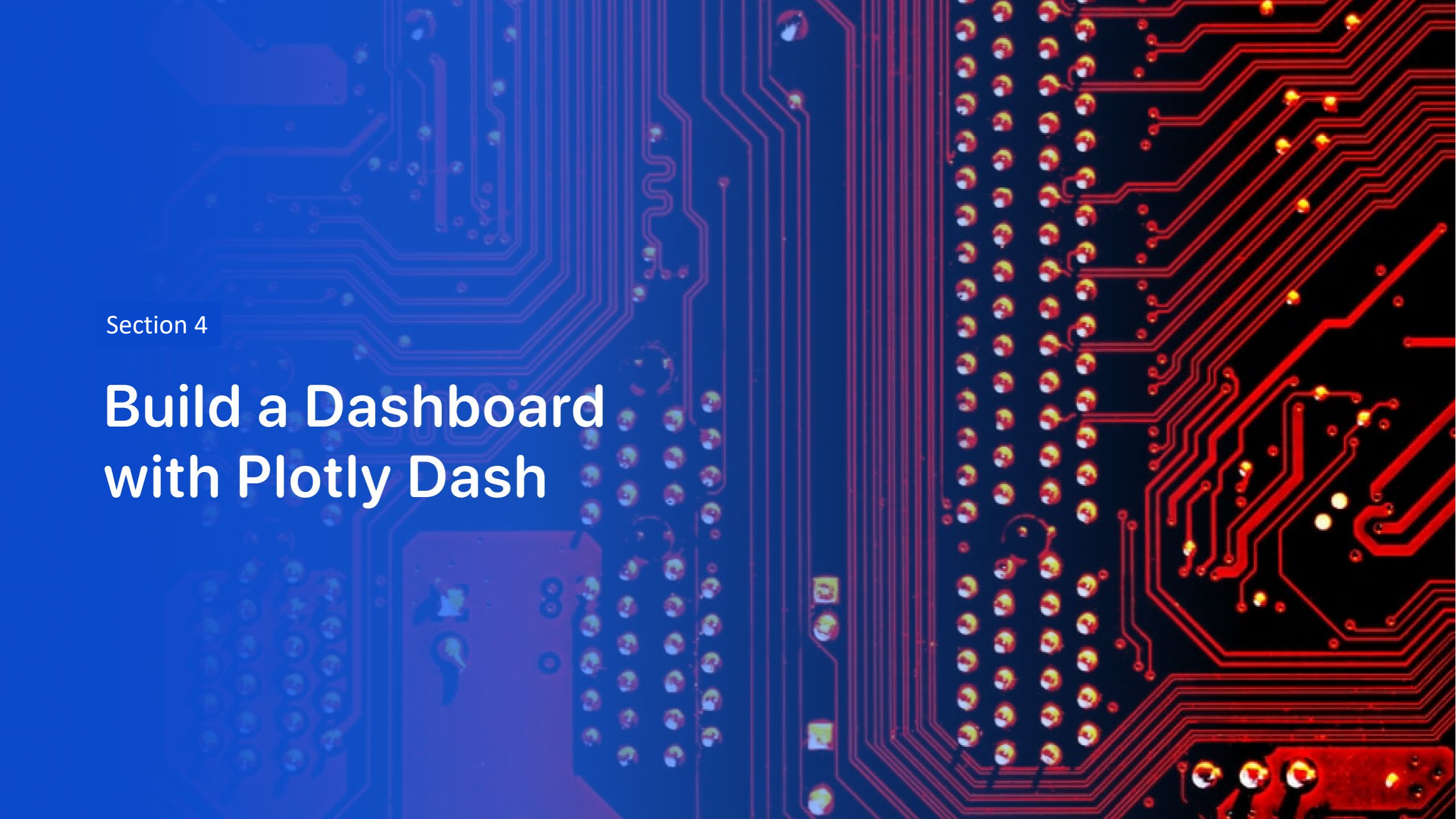
Example of map



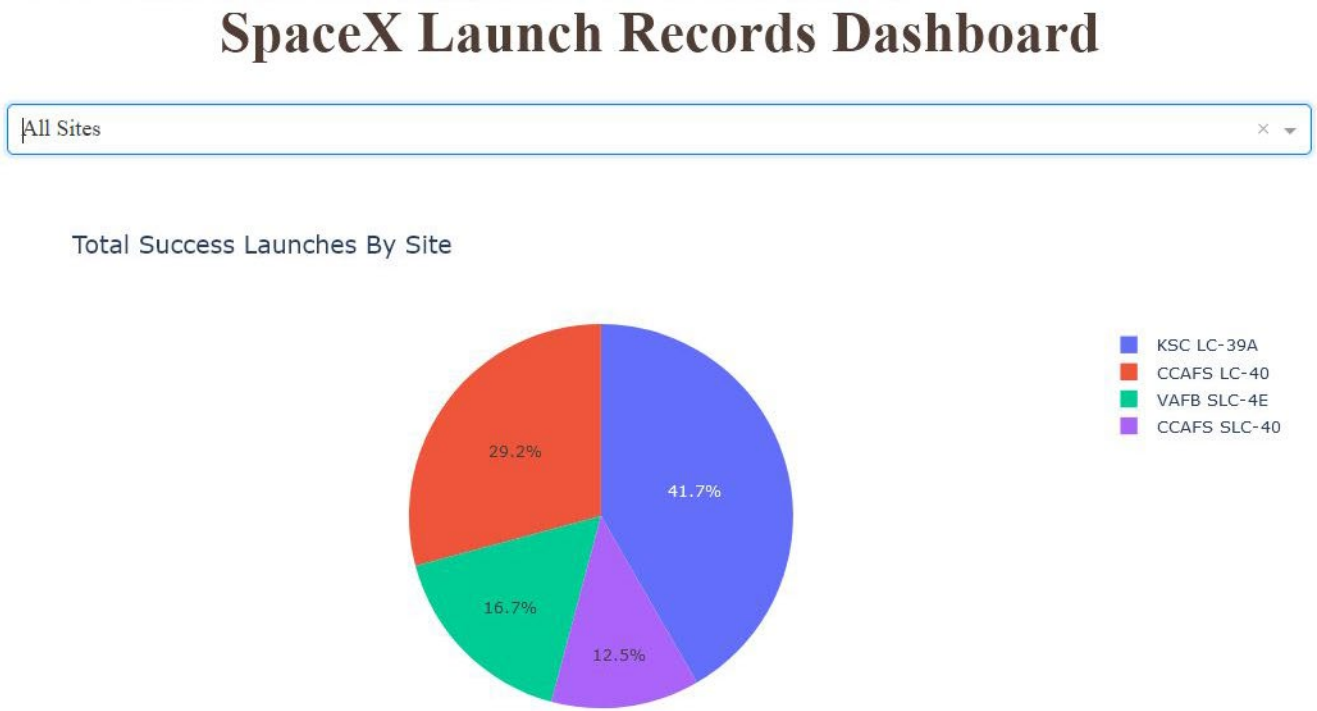
All locations are near a

- Railway
- Highway
- Coastline
- City

This does not influence the success rate.

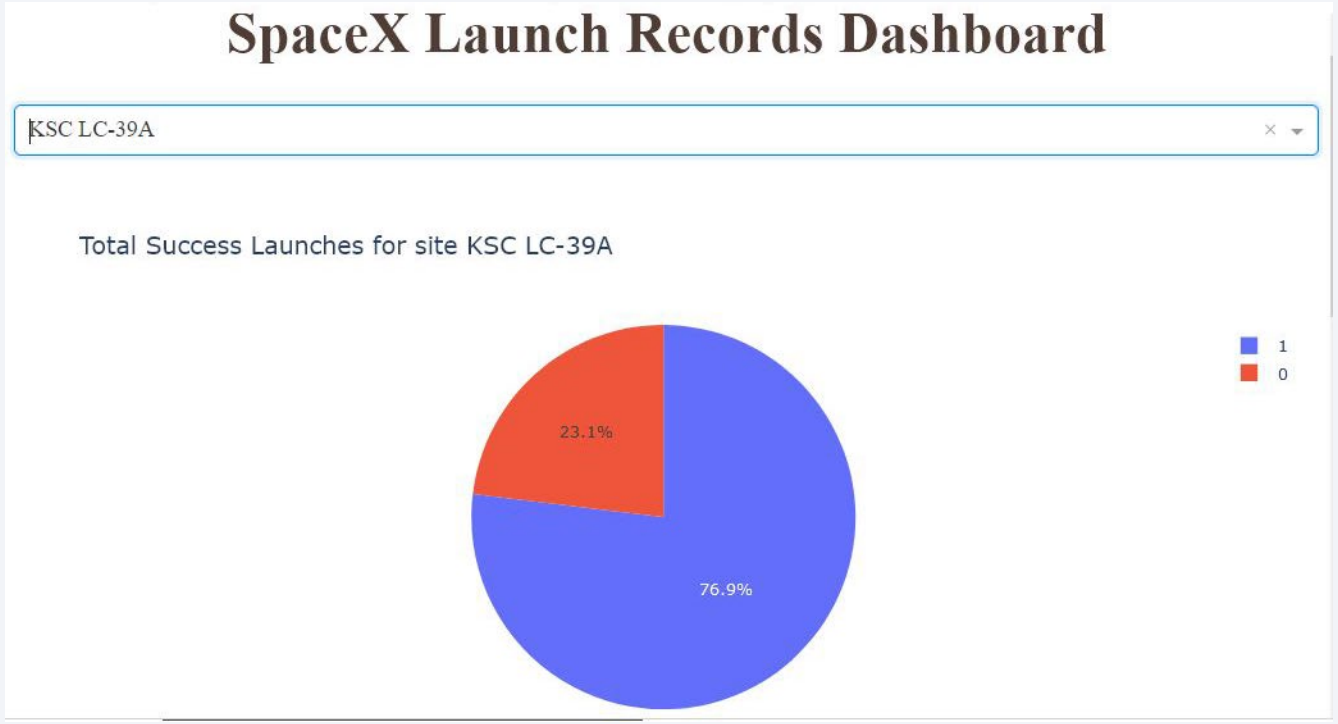

Section 4
Build a Dashboard with Plotly Dash

Success rate of SpaceX launch site locations



- The **highest launch success** rate is 42% by KSC LC-39A
- The **lowest launch success** rate is 13% by CCAFS SLC-40

Success rate of launch site KSC LC-39A



- The **highest launch success** rate is 42% by KSC LC-39A

- Successful launches 77%
- Unsuccessful launches 23%

Launch success rate per payload section

Overview successes per payload section and booster version

Payload/ booster version	0-2500 kg	2500-5000 kg	5000-7500 kg	7500-10000 kg
V1.0				
V1.1	1			
FT	5	6	2	3
B4	1	4		1
B5		1		

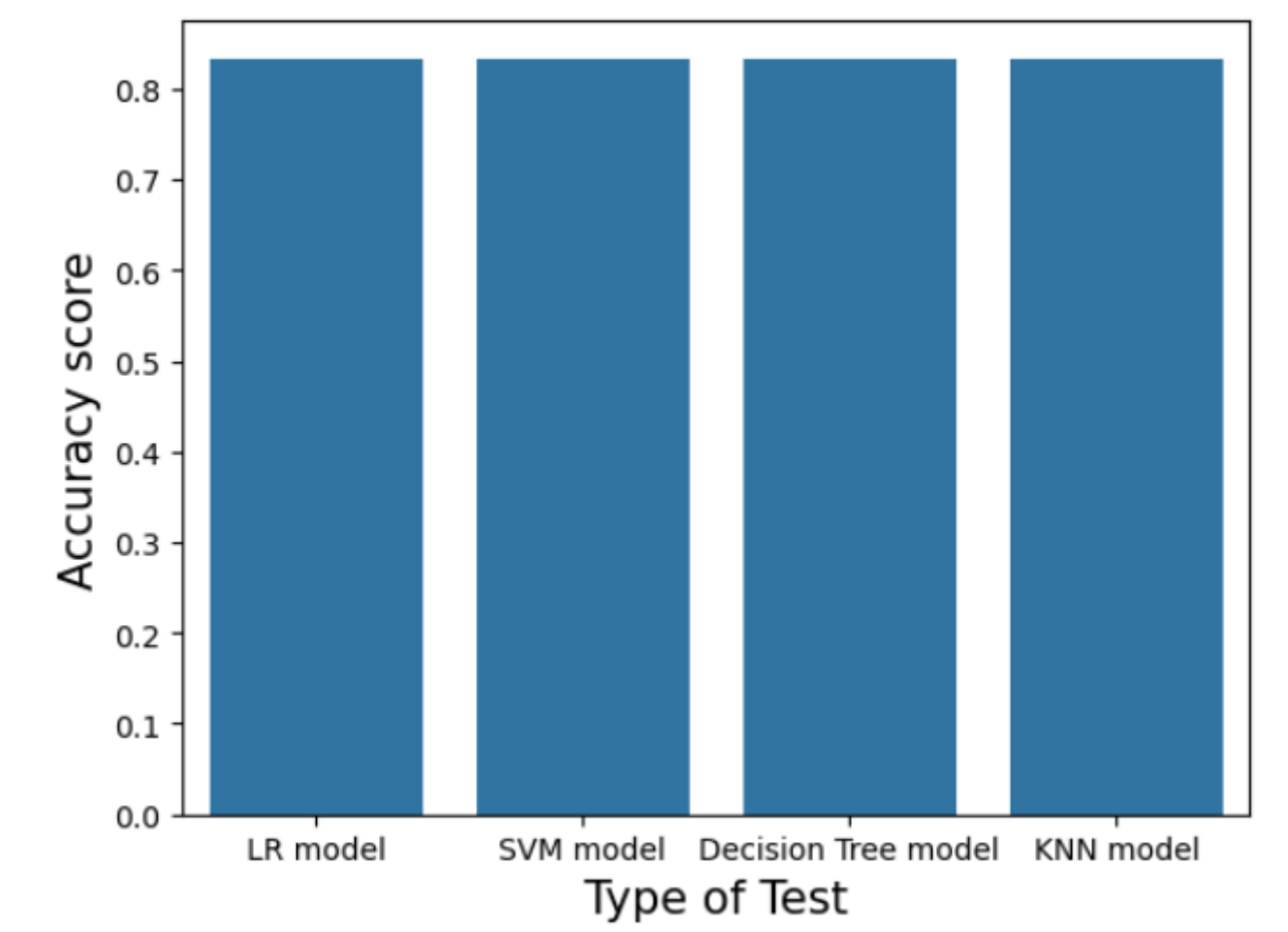
- Launches with 2500-5000 kg payload have a high success rate.
- Launches with booster version FT have a high success rate.

For more details per payload section see Appendix p.49.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

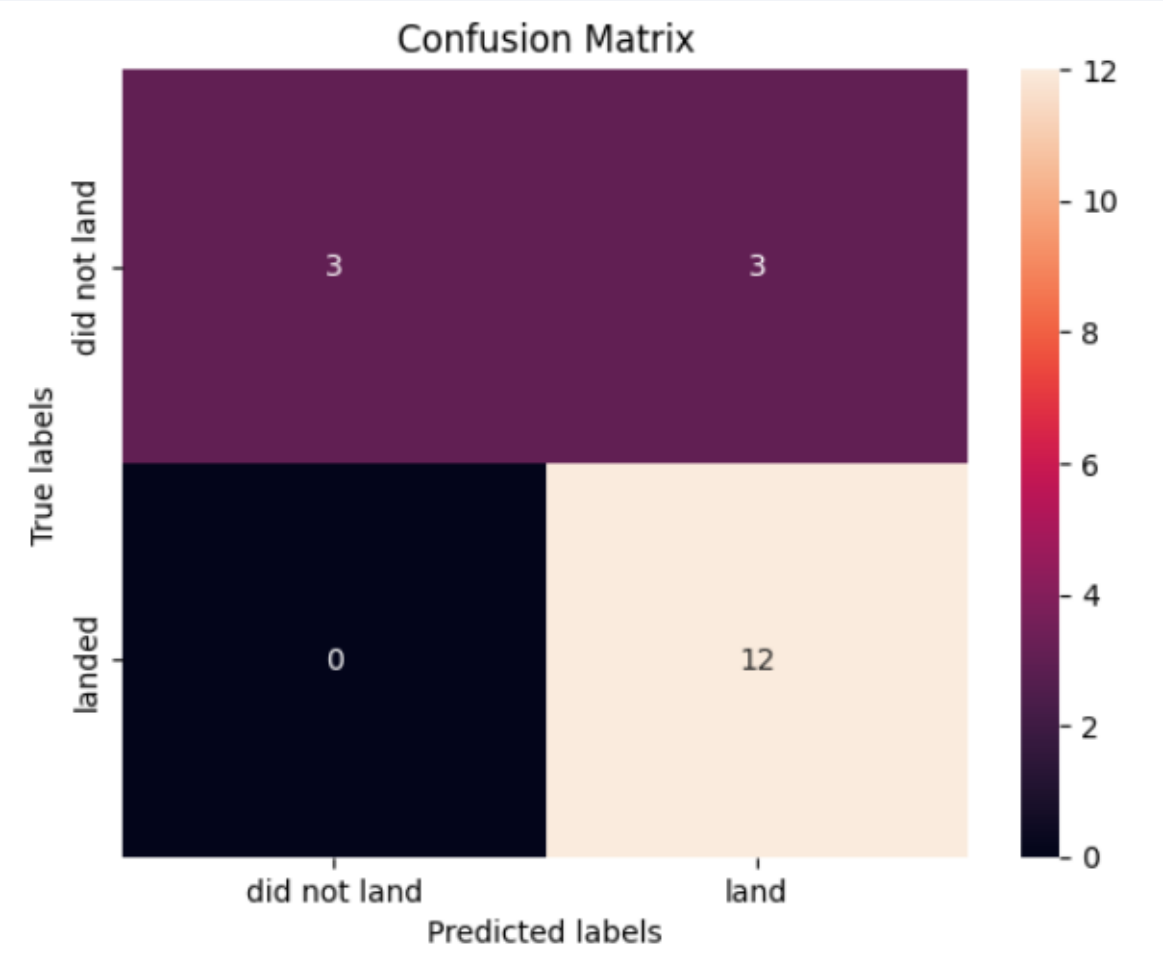


4 different model were built up:

- Logistic Regression model
- Support vector machine model
- Decision Tree model
- K-nearest-neighbor model

All 4 models have the same accuracy score of 0.8333.

Confusion Matrix



All models performed in the same way and have the same accuracy. Therefore, the Confusion Matrix looks the same for all 4 models.

The models predicted for the 18 test cases that 3 first stages did not land and 12 landed successfully, which is correct.

However, it predicted that 3 did land, when they did not.

Conclusions

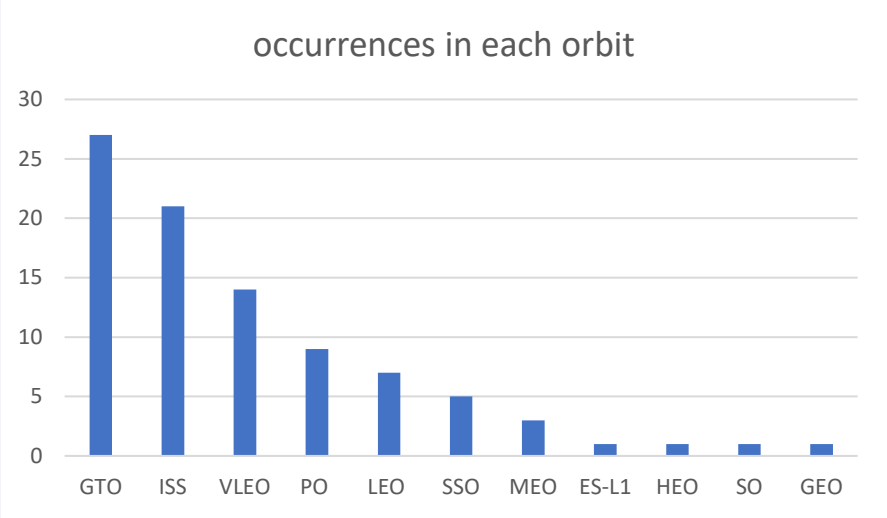
The success rate of the landing outcome of ‘stage one’ depends on:

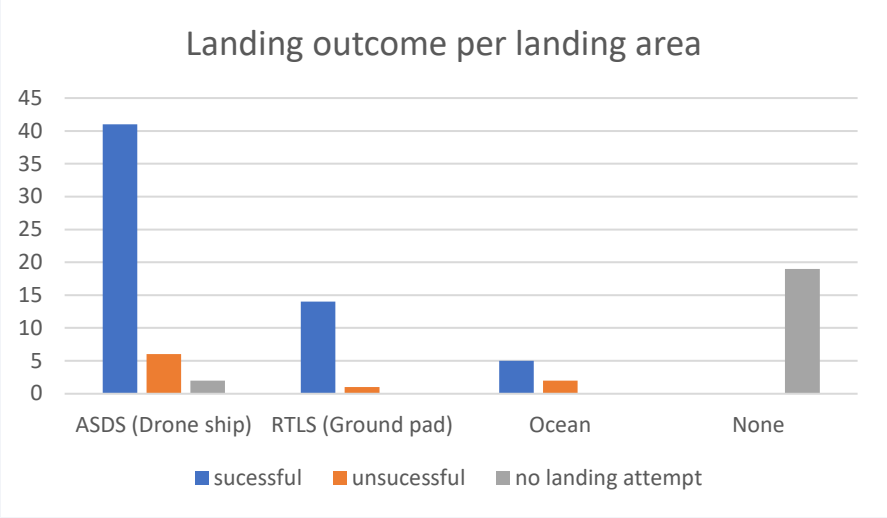
- The launch site, with KSC LC-39 having the highest success rate of all launch sites.
- The payload, with payloads between 2500-5000kg having the highest success rate
- The orbit type, with ES-L1,GEO,HEO and SSO having the highest success rate
- The time, the success rate kept increasing since 2013
- The landing facility, with most attempts carried out on a drone ship with about 50% of landings being successful.
- The booster version, with FT being the highest number of successful outcomes.

- The 4 prediction models which were built, have the same accuracy of 83.3%. With this they predict a landing success rate of 66,7%.

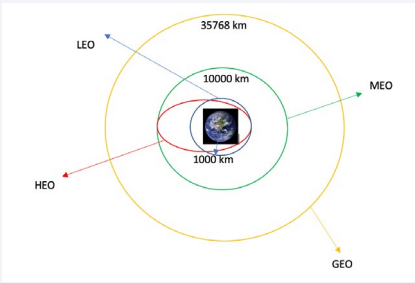
- If the success rate depends on the geographical features of the launch site could not be concluded from this data as all launch sites are located quite south near the equator, near a coastline, near a railway and a highway.

Appendix





- GTO is the most popular orbit

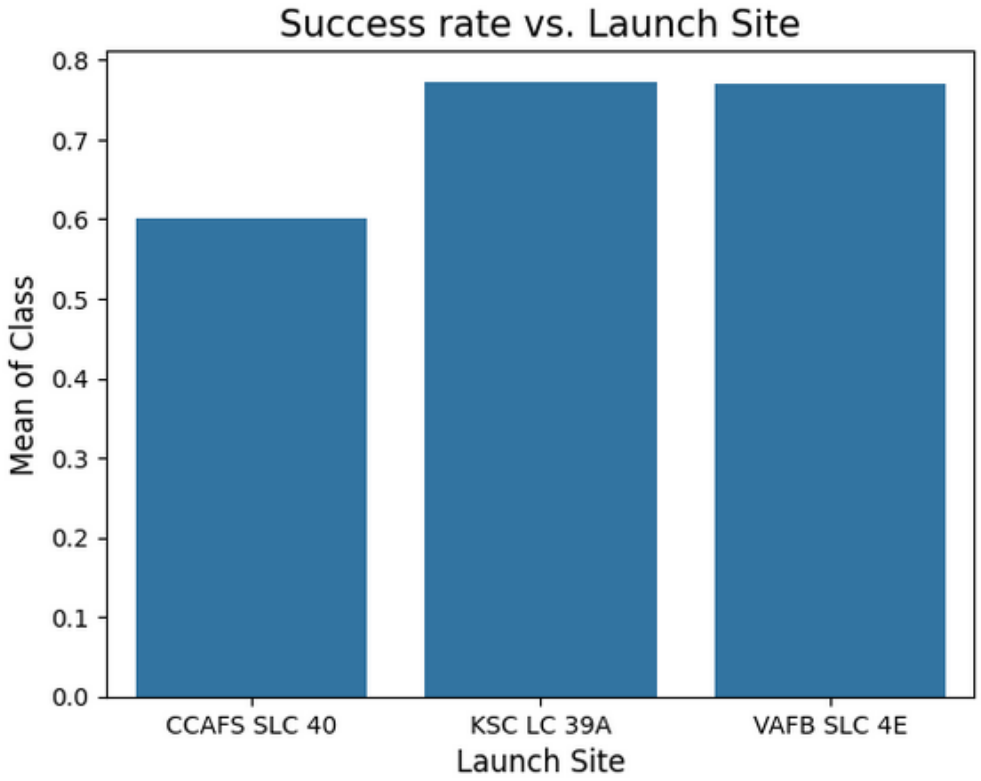


- Landing attempts on drone ships are the most popular

Appendix

Launch Site statistic

	LaunchSite	Class
0	CCAFS SLC 40	0.600000
1	KSC LC 39A	0.772727
2	VAFB SLC 4E	0.769231



There are 3 launch sites:

- **KSC LC 39A:**
  has the highest success rate

LaunchSite	Flights
CCAFS SLC 40	97
KSC LC 39A	44
VAFB SLC 4E	20

Appendix

Launch success rate per payload section

Mind in payload section 7500-10000kg the Chart is not very good, raw data has to be consulted.

Thank you!