IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Beatrice Porcu
April 2025

Outline

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

Executive Summary

- Summary of methodologies
  - DATA RETRIEVAL
    - API requests to SpaceX data API https://api.spacexdata.com/
    - Web scraping via BeautifulSoup of Wikipedia data on Launch https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - EDA: Pandas, Numpy, SQL
  - Visualization: Seaborn, Matplotlib
  - Interactive visual analytics: Folium, Dash
  - ML Prediction: ScikitLearn
- Summary of all results
  - Exploratory Data analysis
  - Machine Learning Predictions

Introduction

- Project prompt:

> *In this capstone, you will take the role of a data scientist working for a new rocket company. Space Y that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk. Your job is to determine the price of each launch. You will do this by gathering information about Space X and creating dashboards for your team. You will also determine if SpaceX will reuse the first stage. Instead of using rocket science to determine if the first stage will land successfully, you will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.*

- Problems you want to find answers
  - What are the key indicators of a successful launch?
  - Is it possible to predict only via ML if a launch will be successful or not?
  - What is the best model to predict accurately launch outcomes?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data is collected from the Web, either via API request or Web Scraping via BeautifulSoup, and subsequently formatted and organized in Dataframes functional to further the analysis
- Perform data wrangling
  - Missing data is handled, while the categorical Outcome column is transformed in 0 for bad outcome and 1 for successful launches
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Different models are built, tuned and compared to obtain the most precise predictions

Data Collection

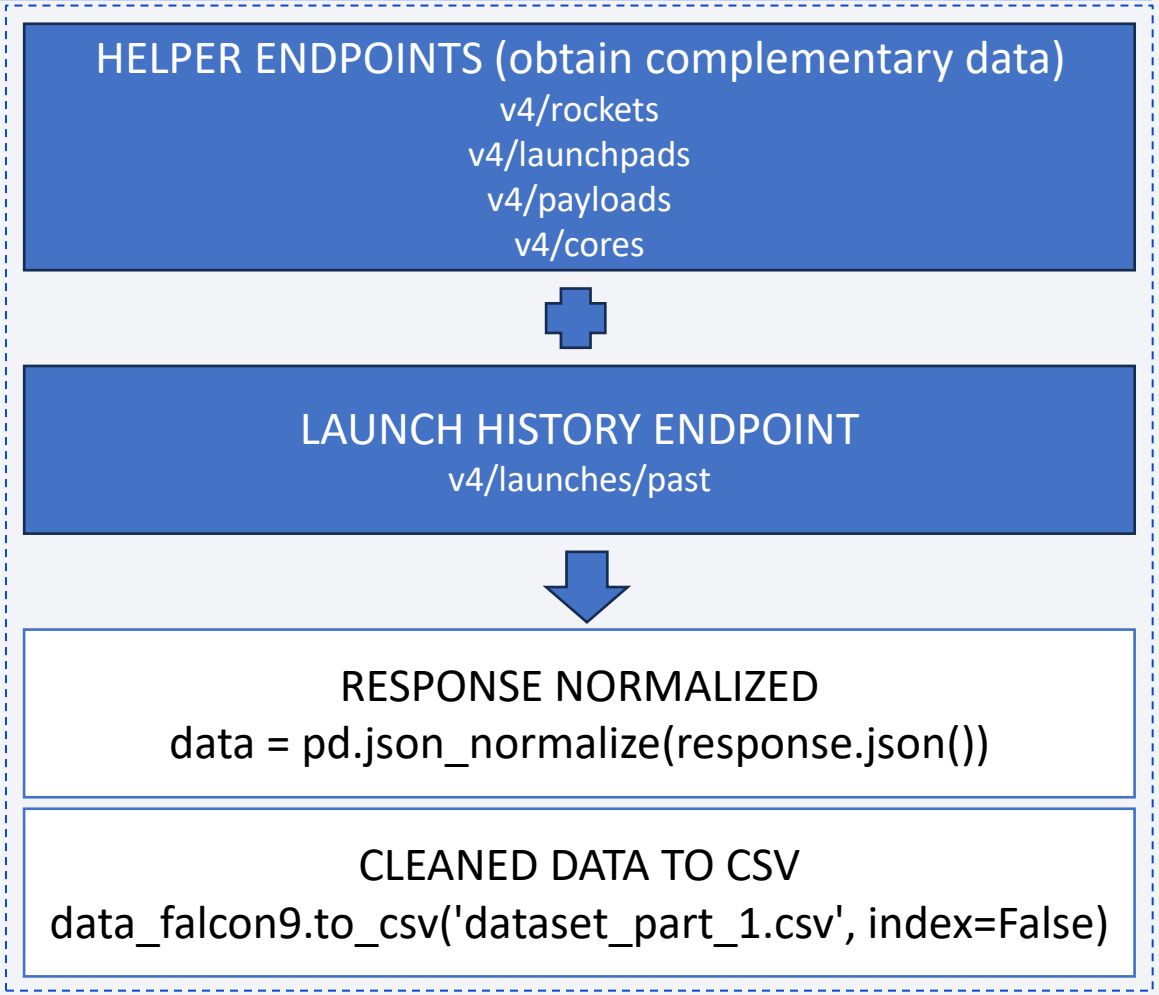
Data sets were collected on the web via:

- API requests to SpaceX data API https://api.spacexdata.com/
  - Extracted data: ['Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome']
- Web scraping via BeautifulSoup of Wikipedia data on Launch https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - Extracted data: BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Collection – SpaceX API

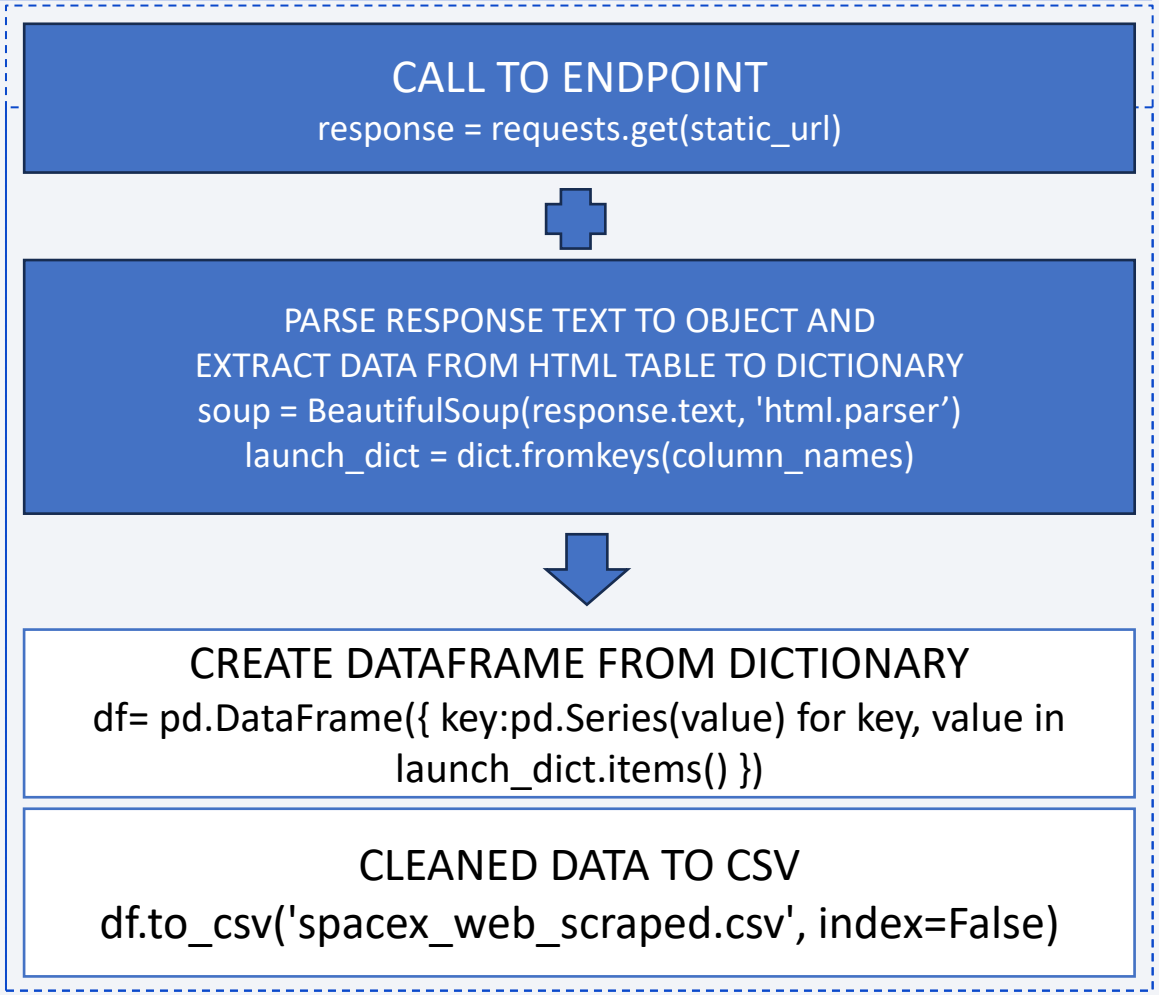
- Data is retrieved by calling a series of API endpoints to obtain different data about both launches and Falcon version used from the SpaceX public API.
- The data is then normalized and formatted via *json_normalize()* to obtain a json response and combined in a single Dataframe.

- Link to notebook



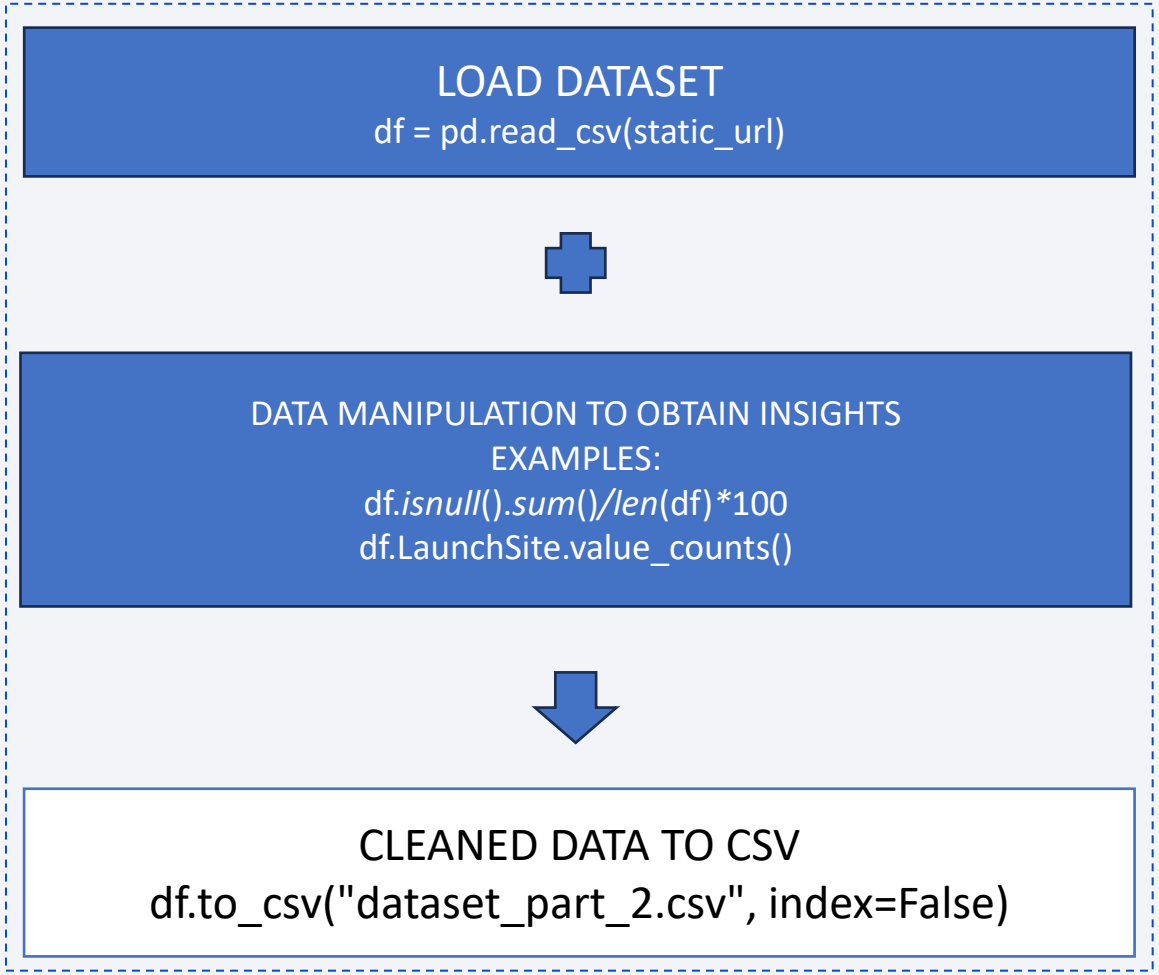
Data Collection - Scraping

- Data is scraped by tables from *List of Falcon 9 and Falcon Heavy launches* Wikipage updated on 9th June 2021
- The page is scraped via the BeautifulSoup python library
- Link to notebook



Data Wrangling

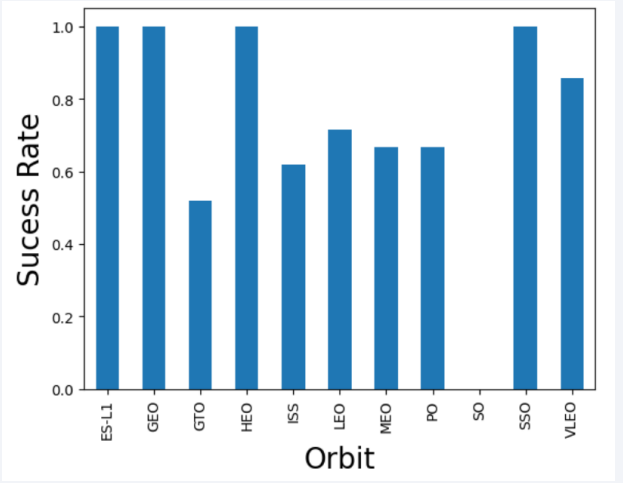
- Data is analyzed via different metrics to obtain insights useful to guide the model preparation:
  - Number of launches on each site
  - Number and occurrence of each orbit
  - Mission outcome of the orbits
  - Creation of *Class* column for outcome results
- Insights:
  - The most used launchpad is the CCAFS SLC 40 (Cape Canaveral Space Launch Complex 40)
  - The most used orbit is the GTO
  - The success rate outcome is close to 67%
- Link to notebook

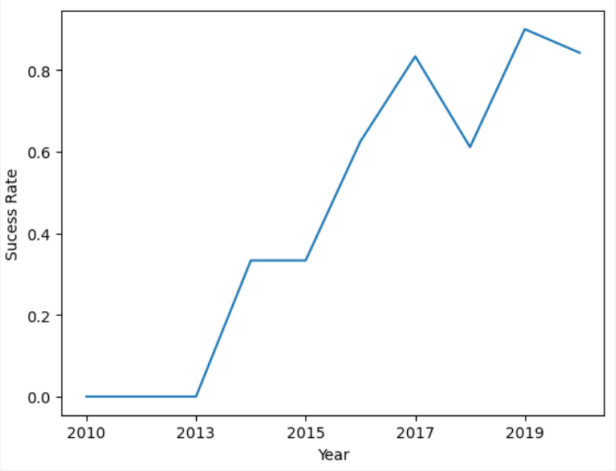


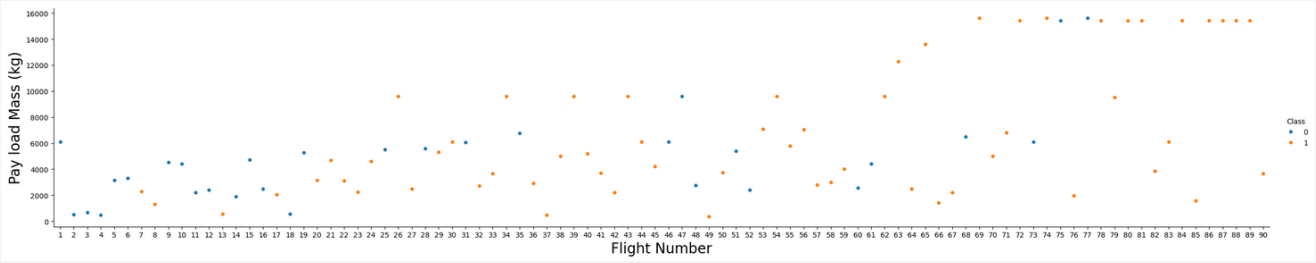
EDA with Data Visualization

- Different combinations of data variables are represented with scatter plots, bar charts and line charts, to get some insights on how the variables relate to one another and if they influence the launch outcome or not.

- Link to notebook







EDA with SQL

- After establishing a connection to the Database via the sqlite3 connector, data is queried to find relevant information
- Link to notebook

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

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome="Success (drone ship)" and PAYLOAD_MASS__KG_ between 4001 and 5999
```

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Unique launch sites in the space mission

```
%sql SELECT distinct(Launch_Site) from SPACEXTBL
```

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTBL where Launch_Site like "CCA%" limit 5
```

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = "NASA (CRS)"
```

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = "F9 v1.1"
```

List the date when the first succesful landing outcome in ground pad was acheived.

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome ="Success (ground pad)"
```

EDA with SQL - continued

List the total number of successful and failure mission outcomes

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

Rank 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.

```
%%sql
  SELECT Landing_Outcome, count(*) as count from SPACEXTBL
  where Date between '2010-06-04' and '2017-03-20'
  group by Landing_Outcome
  order by count desc
```

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

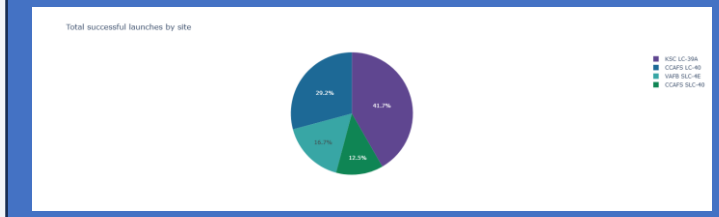
```
%%sql
  select
    CASE substr(Date, 6, 2)
      WHEN '01' THEN 'January'
      WHEN '02' THEN 'February'
      WHEN '03' THEN 'March'
      WHEN '04' THEN 'April'
      WHEN '05' THEN 'May'
      WHEN '06' THEN 'June'
      WHEN '07' THEN 'July'
      WHEN '08' THEN 'August'
      WHEN '09' THEN 'September'
      WHEN '10' THEN 'October'
      WHEN '11' THEN 'November'
      WHEN '12' THEN 'December'
    END
    AS Month_Name, Booster_Version, Launch_Site
    FROM SPACEXTBL
    where Landing_Outcome = "Failure (drone ship)"
      and substr(Date,0,5)='2015'
```

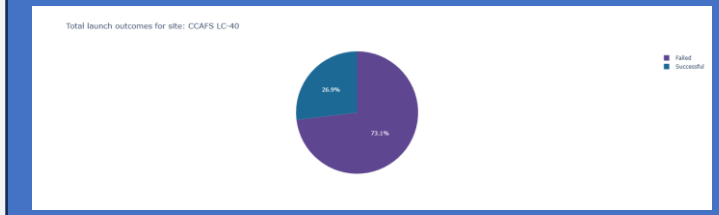
Build an Interactive Map with Folium

- Data from launches and locations is plotted with Folium maps, in order to gain insights about the possible correlation between location of launch site and launch outcome (e.g. does a launch site near water perform better?)
- The analysis is done in three phases:
    1. Mark all launch sites on a map
    2. Mark the success/failed launches for each site on the map
    3. Calculate the distances between a launch site to its proximities (railways, highways, the sea, other cities)
- The maps created help find geographical patterns to link to other correlations gathered from the preceding analysis, in order to establish the best location for a successful launch outcome
- Link to notebook

Build a Dashboard with Plotly Dash

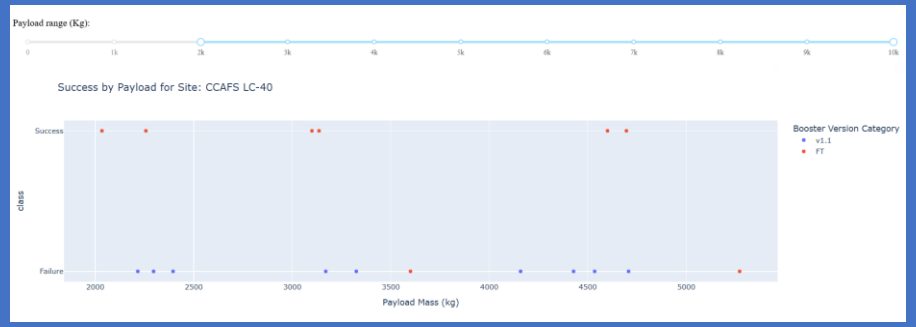
- The Plotly dashboard gives answer to the questions:
  - *Which site has the largest successful launches?*
  - *Which site has the highest launch success rate?*
  - *Which payload range(s) has the highest launch success rate?*
  - *Which payload range(s) has the lowest launch success rate?*
  - *Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?*
- Link to file





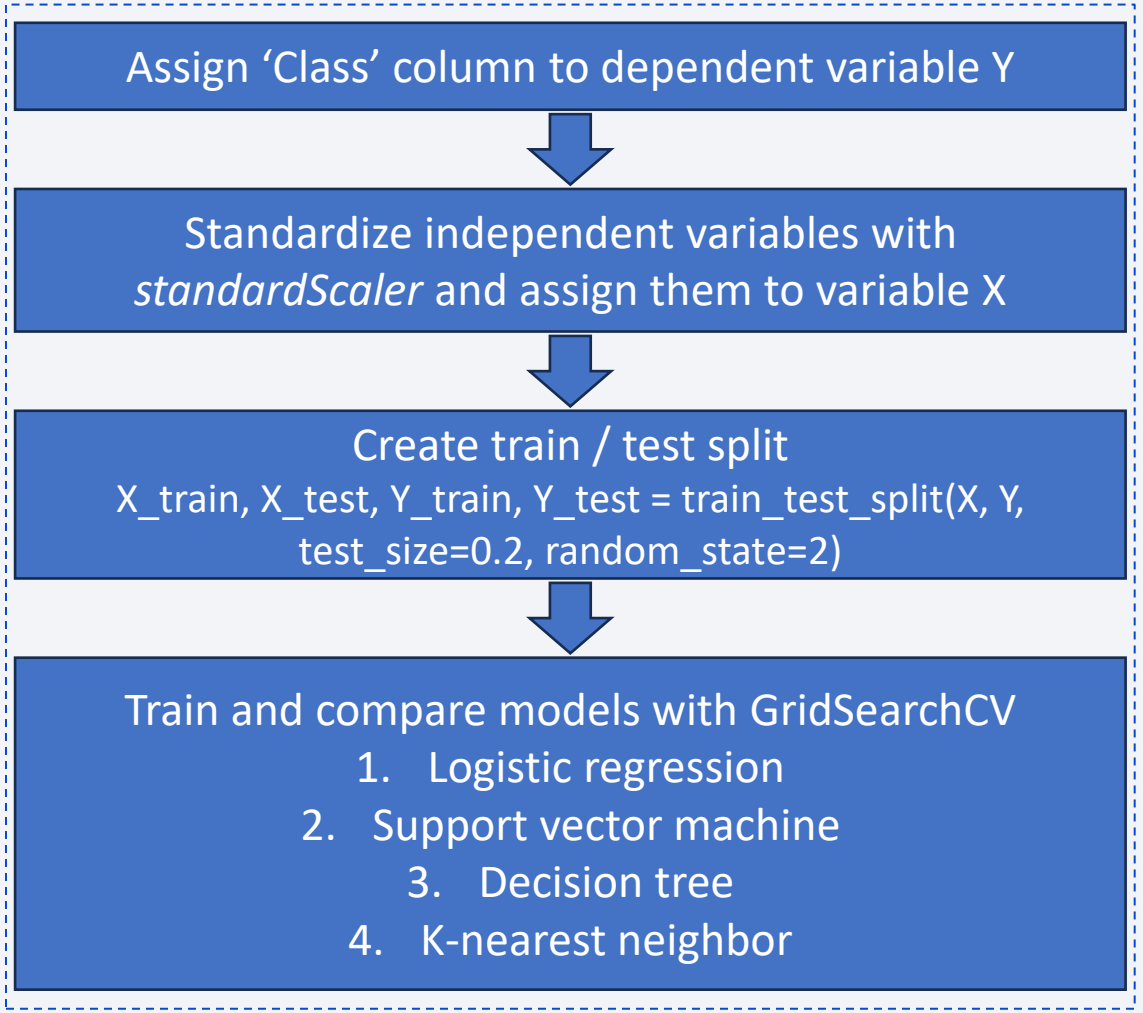
First, a pie chart shows the site with the highest number of successful launches, or alternatively the ratio of successful/failed launches for each site

A scatter plot then shows how the payload range, selected with a slider, influences the launch outcome both for all sites or for the site previously selected



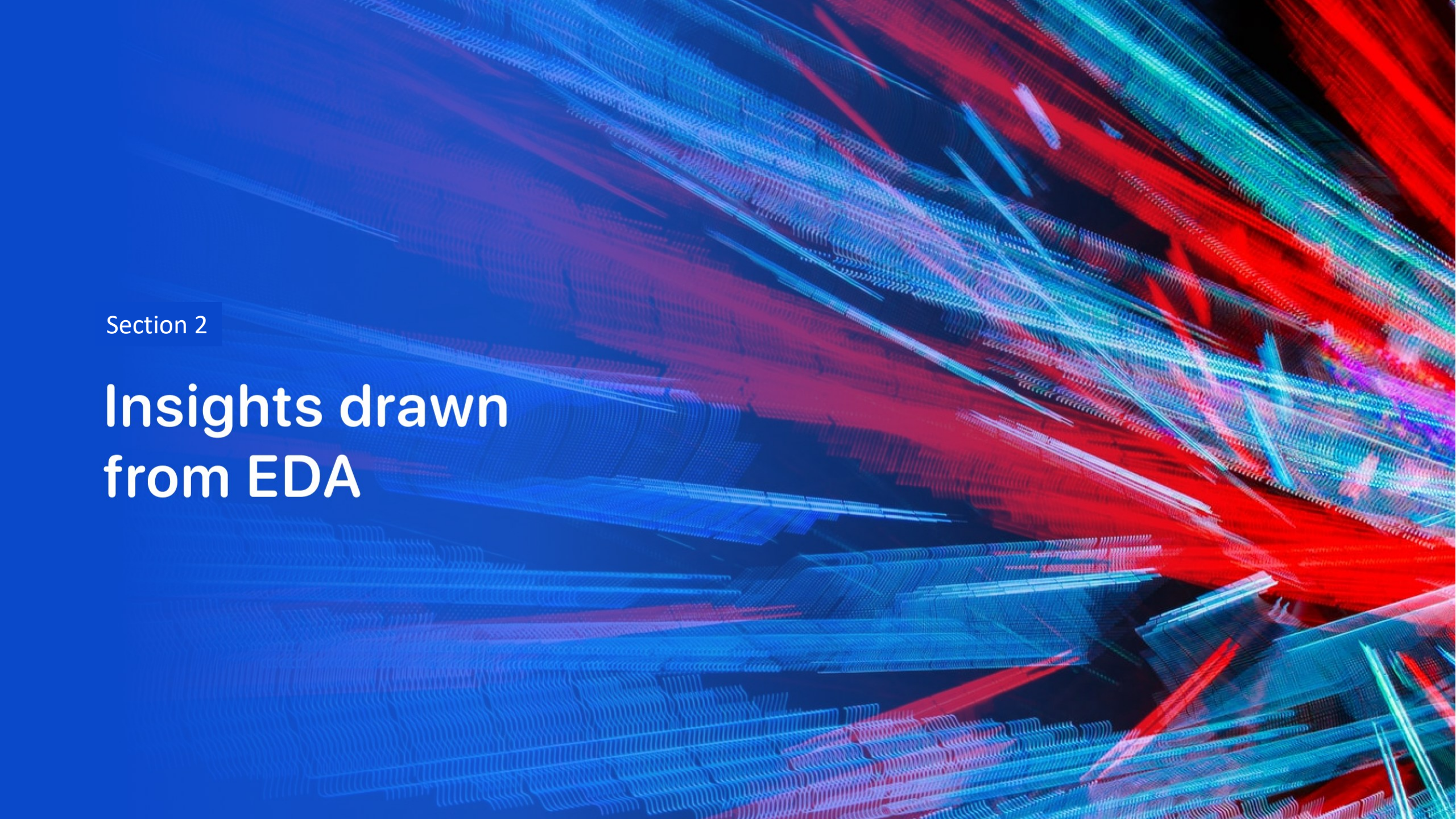
Predictive Analysis (Classification)

- Data is prepared and split into training and test sets in order to train and compare different models
- For each model a confusion matrix is plotted and compared to the ones of other models
- Accuracy score is also determined for every model
- Link to notebook

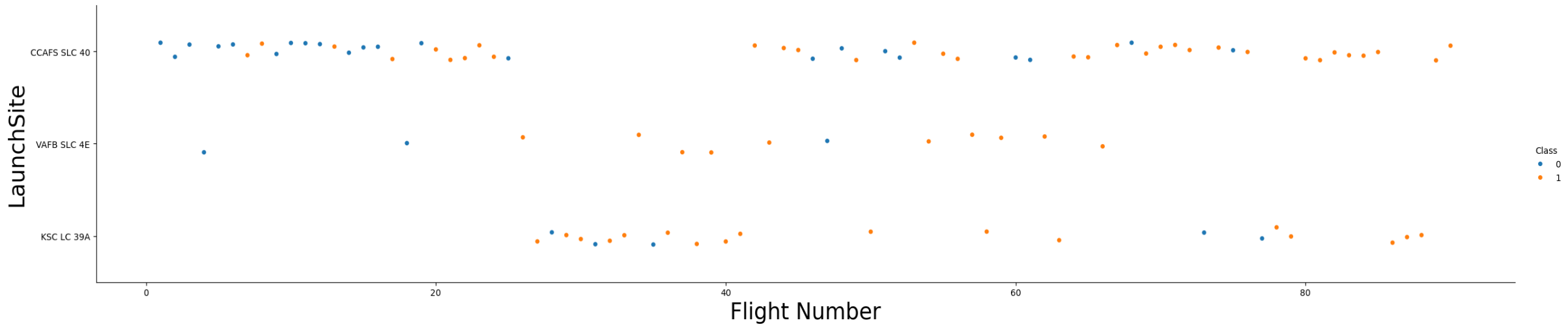


Results

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

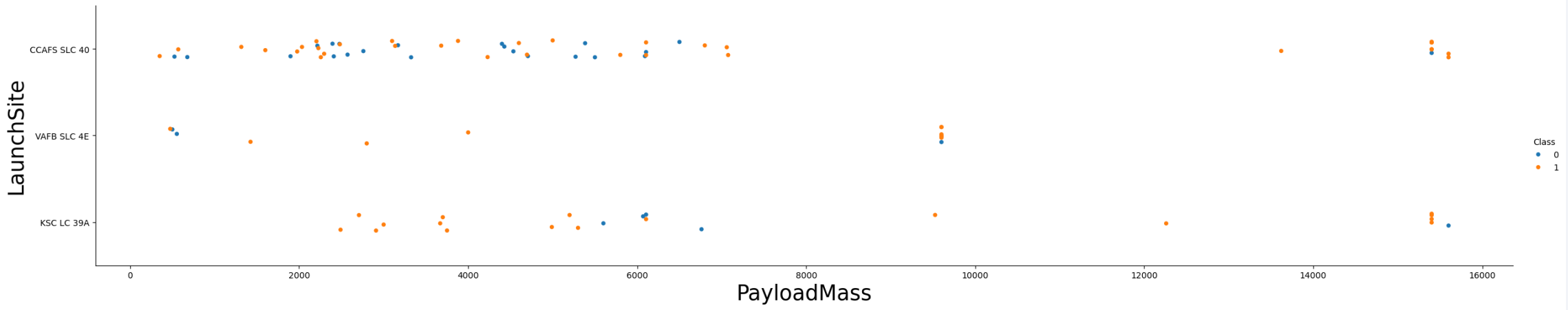

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- As the scatter plot shows, as number of launches increases, so does the success rate.
- As time passes, success rate increases for all launch sites
- KSC LC 39A is the launch site with highest success rate, but CCAPS SLC 40 has the most launches

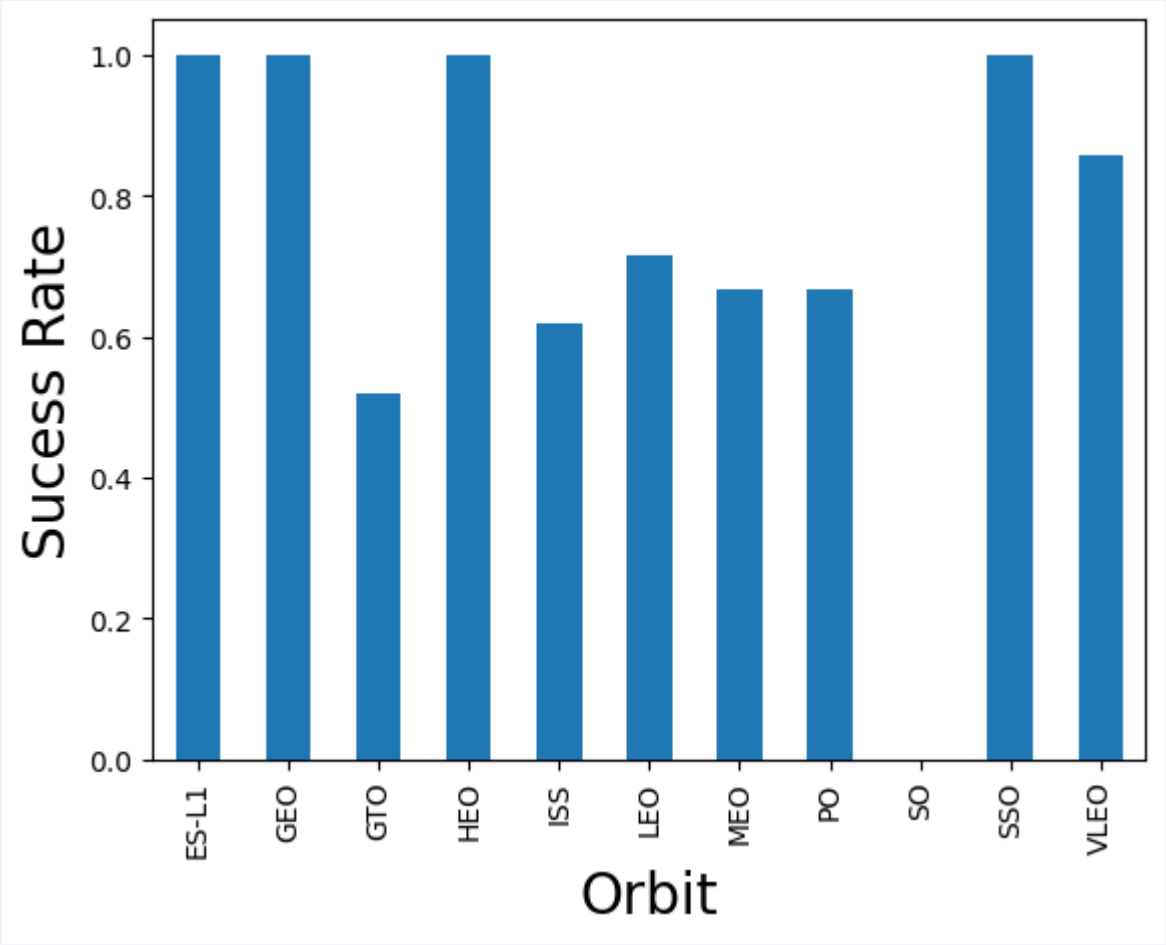
Payload vs. Launch Site



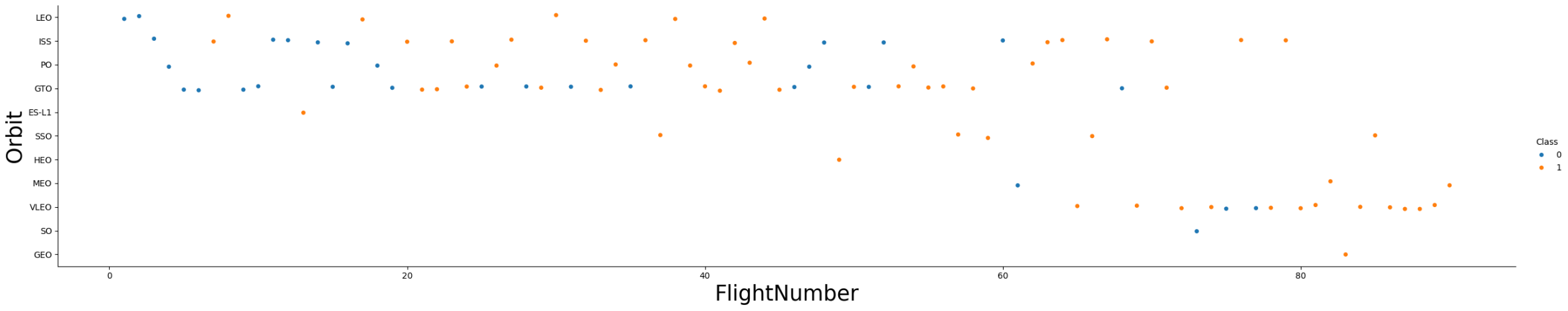
- Although is not particularly clear from this plot, it seems that for higher payloads the success rate increases
- KSC LC 39A has an almost totally successful rate for smaller payloads

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO have all a 100% success rate
- SO has a 0% success rate
- The remaining orbits have varying success rates, but all mostly over 50%



Flight Number vs. Orbit Type



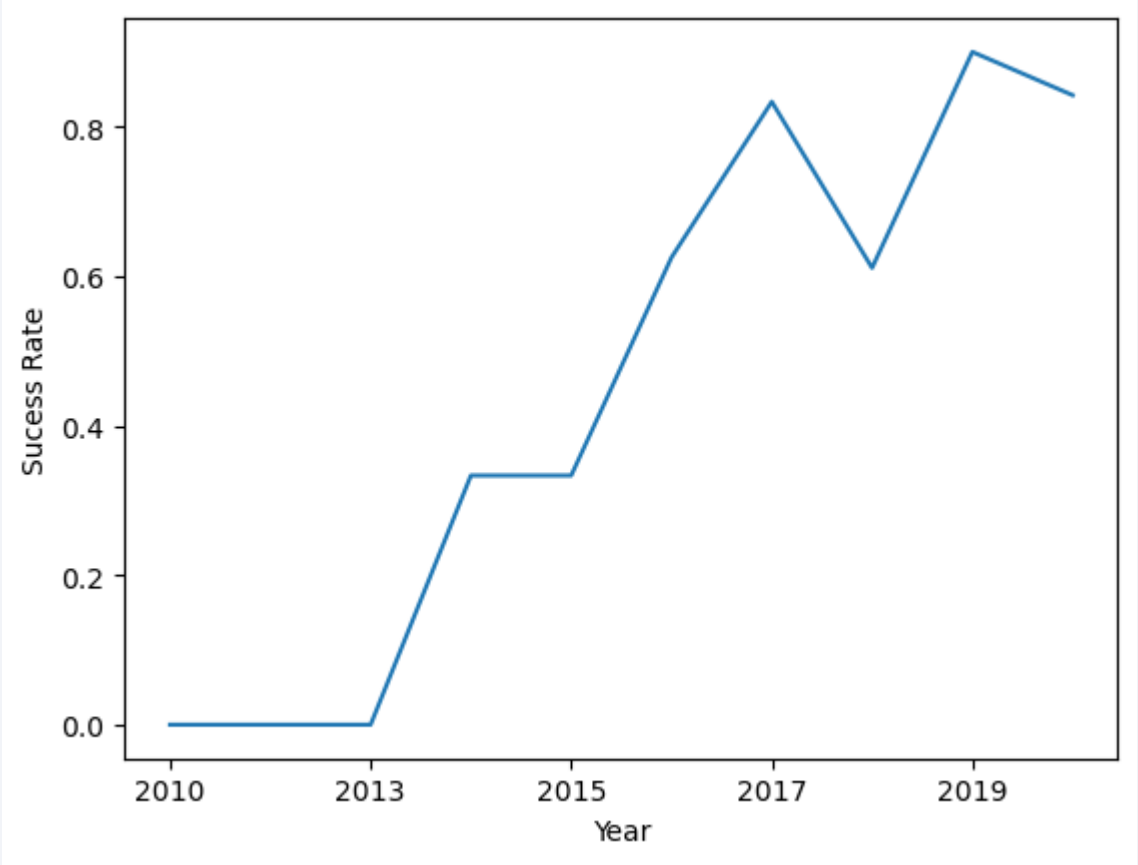
- Except for the LEO orbit, it does not seem to exist a strong relation between orbit and number of flights

Payload vs. Orbit Type

- Heavy payloads seem to perform better on Polar, LEO and ISS orbits
- For GTO and other orbits the distinction is unclear

Launch Success Yearly Trend

- The success rate drastically increased in more recent years
- Although there's a downward tendency around 2018, the trend is still going upwards with a slight depression around 2020



All Launch Site Names

```
%sql SELECT distinct(Launch_Site) from SPACEXTBL
```

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

- Data is filtered to retrieve all unique values in the launch site column, removing all duplicates

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL 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 query retrieves the first 5 records where launch sites begins with `CCA`. The *like* keyword and the % wildcard help retrieve all relevant results without searching for the entire string

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_)
            from SPACEXTBL
            where Customer = "NASA (CRS)"
```

sum(PAYLOAD_MASS__KG_)
45596

- The query calculates through the function *sum()* the total payload carried by boosters from NASA (CRS)

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_)
        from SPACEXTBL
        where Booster_Version = "F9 v1.1"
```

AVG(PAYLOAD_MASS__KG_)
2928.4

- The query calculates the average (*avg()*) payload mass carried by booster version F9 v1.1. the filtering in the *where* clause is the alternative version to the *like* keyword used in previous queries, where only part of the string to use as filter was used; in this case, the = sign requires the entire string to be used as filter to be present, otherwise the results will not be those expected.

First Successful Ground Landing Date

```
%sql select min(Date) from SPACEXTBL where Landing_Outcome ="Success (ground pad)"
```

min(Date)
2015-12-22

- The query returns the date of the first successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version
        from SPACEXTBL
        where Landing_Outcome="Success (drone ship)"
        and PAYLOAD_MASS__KG_ between 4001 and 5999
```

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

- The query returns the list of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. To be noted the chaining of *where* conditions and the use of *between* to filter a range on a column of numerical data

Total Number of Successful and Failure Mission Outcomes

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

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

- The query calculates the total number of successful and failure mission outcomes.
- The double entry for the *Success* results is caused by some differences in formatting of the column data. It would be ideal before proceeding to correct these kind of 'errors' and decide the most appropriate procedure to standardize the data.
- ALTERNATE QUERY: Instead of just group by mission outcome, an alternative way would be to group and filter using wildcard '*Failure%*' and '*Success%*': in this way it would be solved the problem of different formatting, but it would lose the granularity of results where other info beyond success or failure are given. In this case it should not be a problem, given that in the end the outcome of the success column would be one of only two values

Boosters Carried Maximum Payload

```
%sql SELECT BOOSTER_VERSION
        FROM SPACEXTBL
        WHERE PAYLOAD_MASS__KG_ =
            (SELECT MAX(PAYLOAD_MASS__KG_)
            FROM SPACEXTBL)
```

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

- The query lists the names of the booster which have carried the maximum payload mass

2015 Launch Records

Month_Name	Booster_Version	Launch_Site
January	F9 v1.1 B1012	CCAFS LC-40
April	F9 v1.1 B1015	CCAFS LC-40

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Given that SQLLite does not support monthnames, a switch case and *substr(Date, 6,2)* has been used to insert them in the result outcome.

```
%%sql
  select
    CASE substr(Date, 6, 2)
      WHEN '01' THEN 'January'
      WHEN '02' THEN 'February'
      WHEN '03' THEN 'March'
      WHEN '04' THEN 'April'
      WHEN '05' THEN 'May'
      WHEN '06' THEN 'June'
      WHEN '07' THEN 'July'
      WHEN '08' THEN 'August'
      WHEN '09' THEN 'September'
      WHEN '10' THEN 'October'
      WHEN '11' THEN 'November'
      WHEN '12' THEN 'December'
    END
    AS Month_Name, Booster_Version, Launch_Site
    FROM SPACEXTBL
    where Landing_Outcome = "Failure (drone ship)"
      and substr(Date,0,5)='2015'
```

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

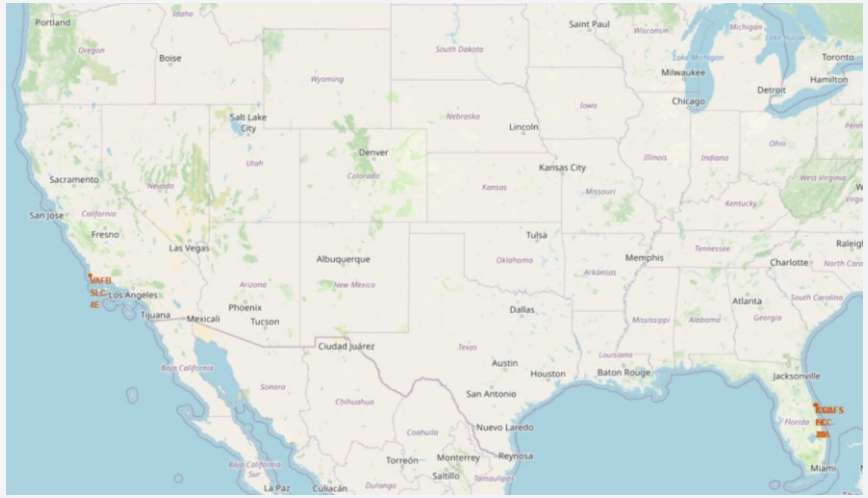
```
%%sql
   SELECT Landing_Outcome, count(*) as count from SPACEXTBL
   where Date between '2010-06-04' and '2017-03-20'
   group by Landing_Outcome
   order by count desc
```

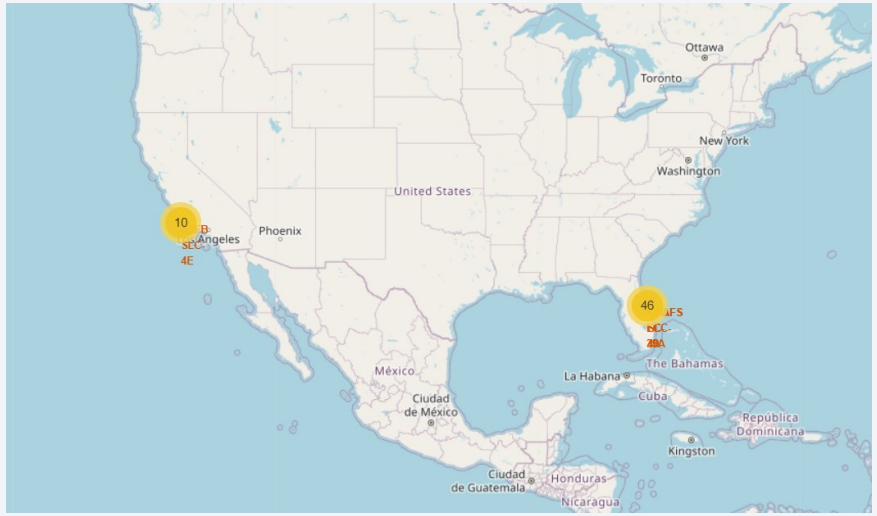
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- The query result is 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
- Again, the *between* keyword is used to filter a range of dates


Section 3
Launch Sites
Proximities Analysis

Launch sites locations



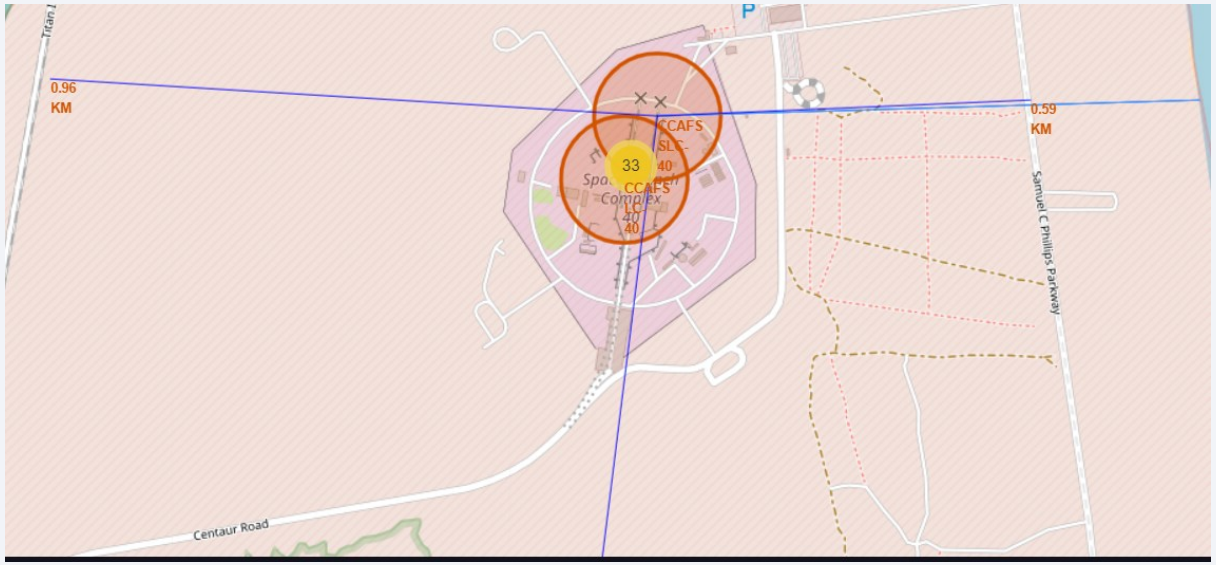


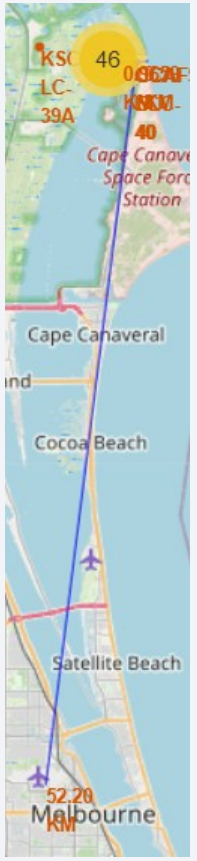
- All launch sites are situated on US coasts, near the equator line.
- 10 of the total sites are situated on the East coast, while the remaining 46 are on the West Coast

Launch outcomes per launch site

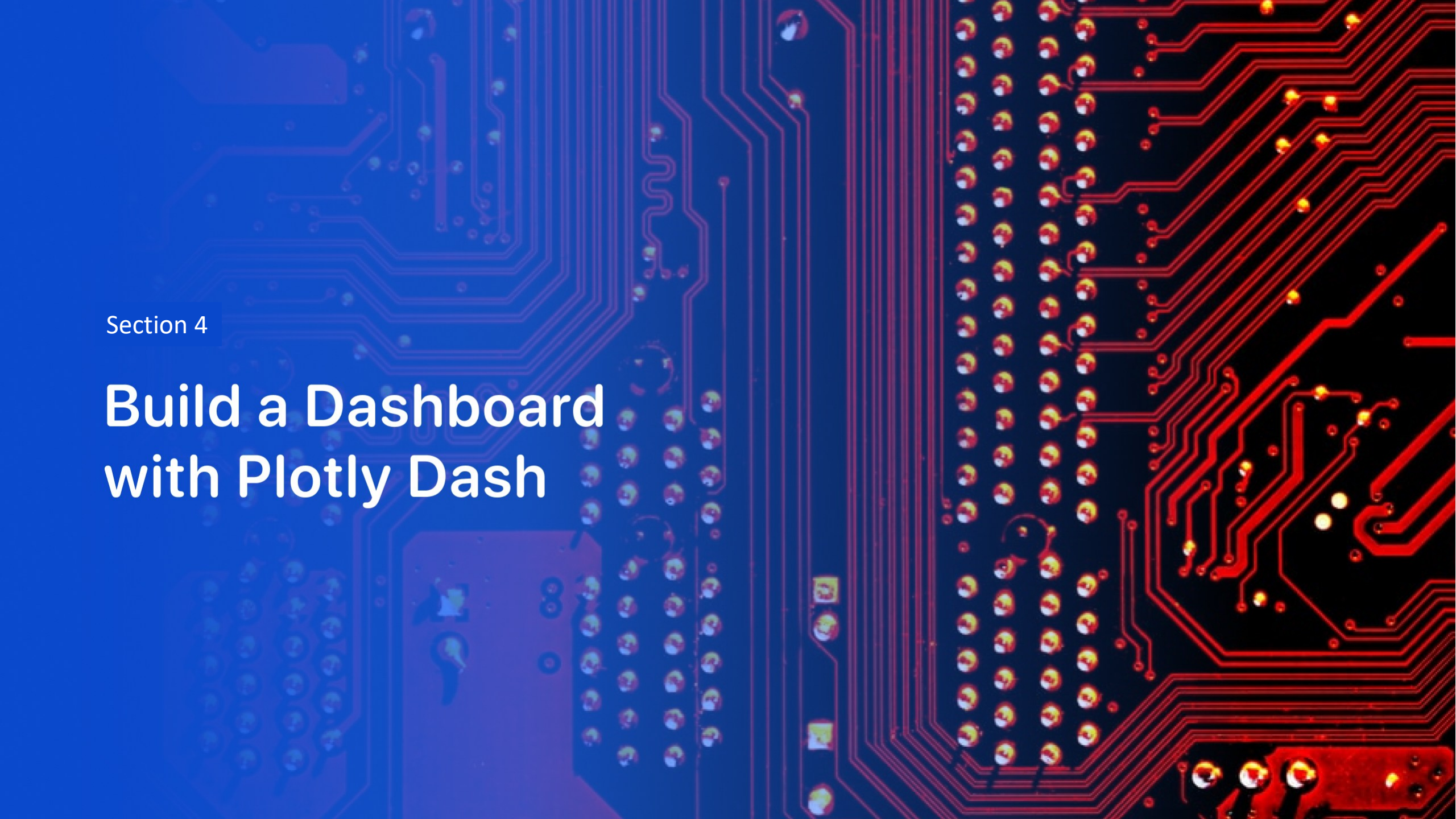
- Zooming in on the landmarks additional data is visible
- Each landmark has been highlighted with marks to indicate all launches registered in the dataset, indicating with green the launches with a successful outcome and red for failures

Distance between launch sites and other landmarks

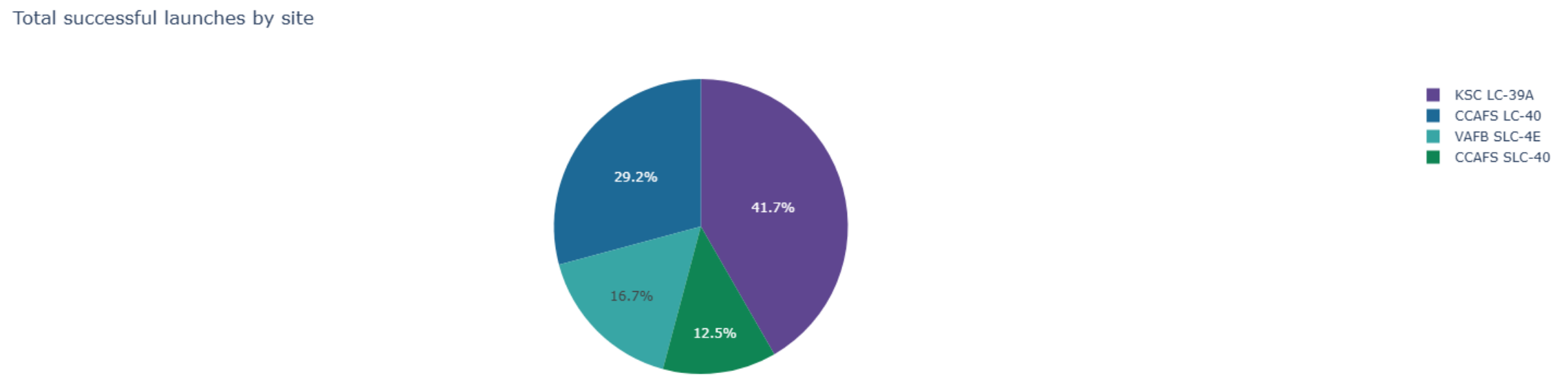




- An example launch site has also been correlated to its proximities such as railway, highway, coastline, and nearest city, with distance calculated and displayed
- By these findings it can be determined that while sites are quite close to other landmarks, their vicinity to the ocean helps decrease the possibility of debris from launches can cause damages to people and cities

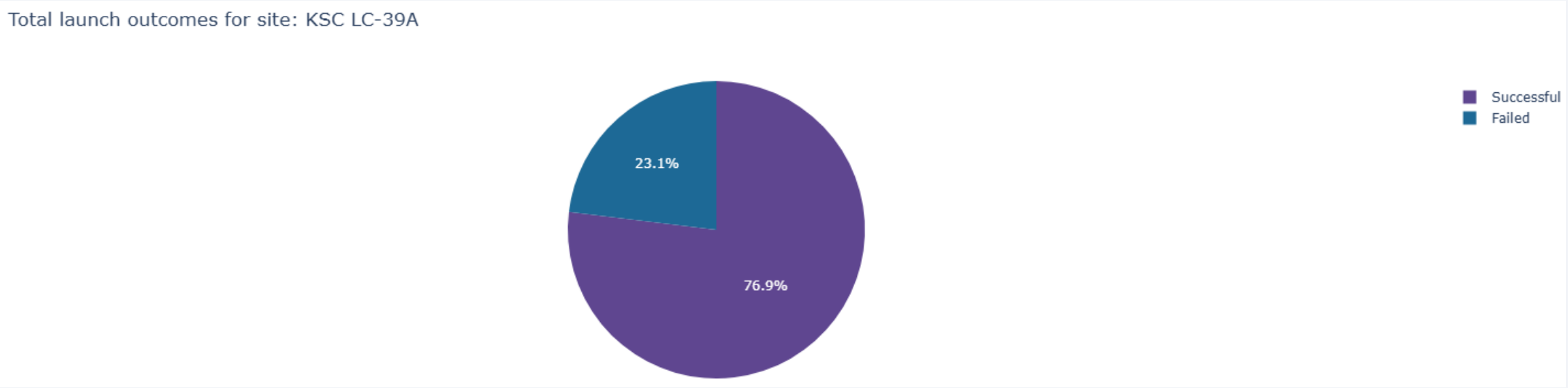

Section 4
Build a Dashboard with Plotly Dash

Total successful launches by site – All sites



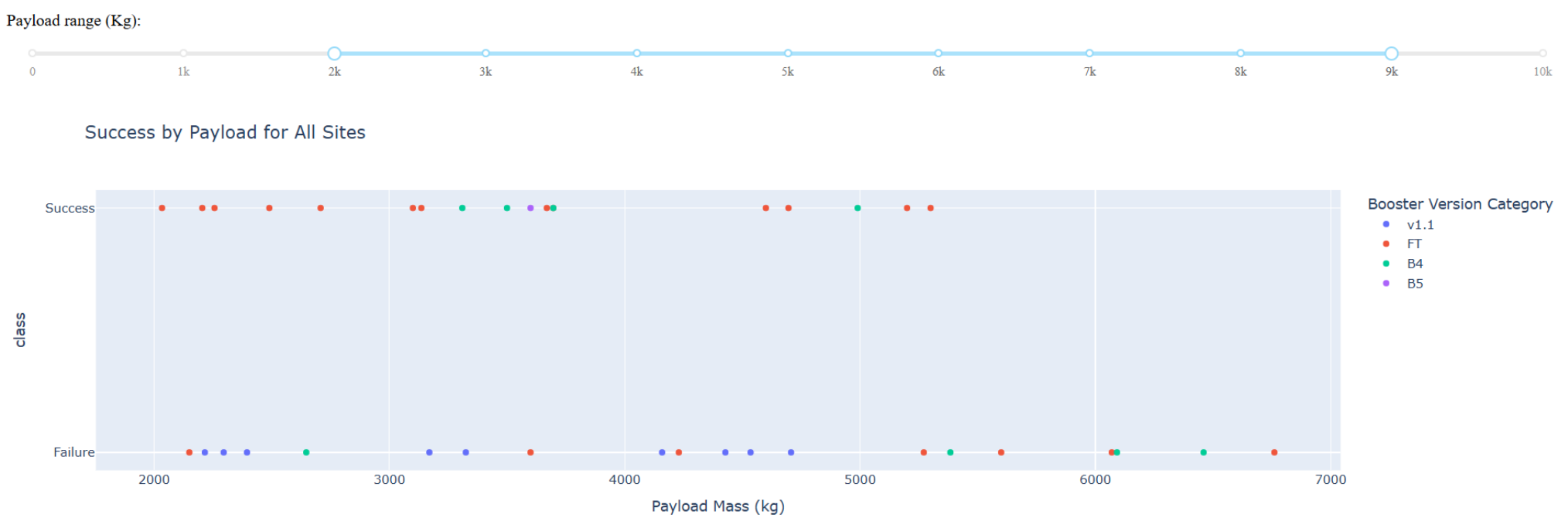
- The pie graphs shows clearly that almost half the successful launches come from the the launch site KSC LC-39 A (41.7%), followed by CCAFS LC-40, VAFB SLC-4E and finally CCAFS SLC-40

Total successful launches by site – Most successful site



- Filtering the data by the site with most successful outcomes from the previous graph, it shows that the success rate is almost at 77%, so more than 7 out of 10 successful launches

Correlation Payload mass – Successful launches



- Selecting different payload ranges for all sites combined, it's visible that smaller payloads (about 2000-5000 kgs) have higher success rates

Section 5

Predictive Analysis (Classification)

Classification Accuracy

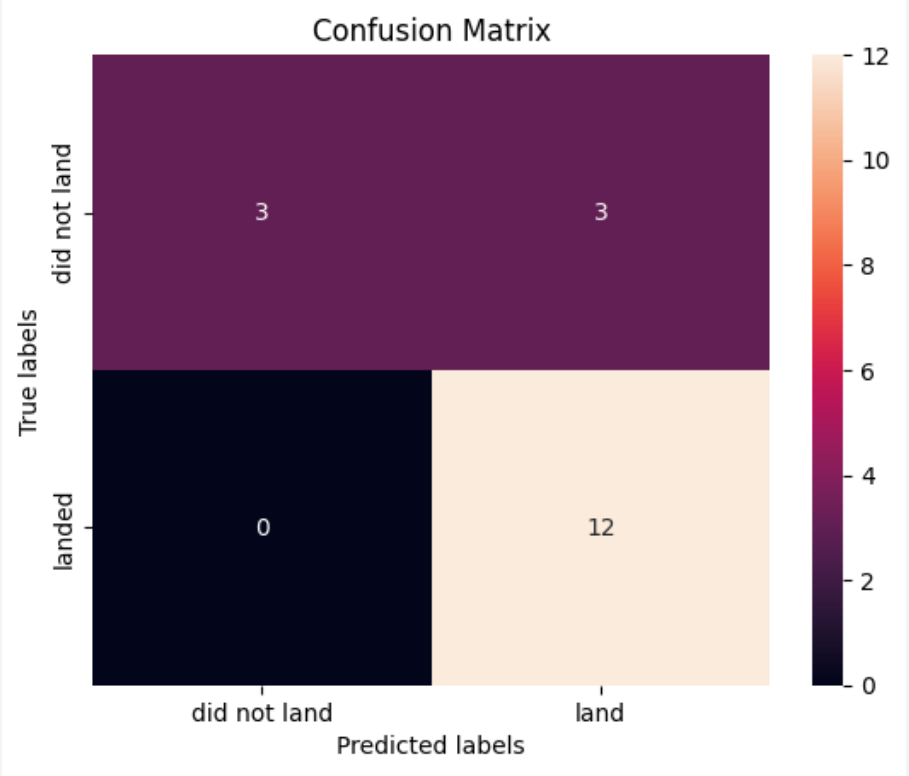
- Data is prepared and split into training and test sets in order to train and compare different models
- 4 models have been trained to compare:
  - Logistic regression
  - Support vector machine
  - K-Nearest neighbor
  - Decision tree

	ML Method	Accuracy Score (%)	Best score
0	Support Vector Machine	83.333333	0.848214
1	Logistic Regression	83.333333	0.846429
2	K Nearest Neighbour	83.333333	0.848214
3	Decision Tree	83.333333	0.876786

- For each model a confusion matrix is plotted and compared to the ones of other models, obtaining equivalent results
- Accuracy score is also determined for every model
- As shown in the screenshot, all models have equivalent performances, with a slightly better result for the Decision tree model; this equivalence in results is most probably due to the small sample of the dependent variable Y data (only 18 entries)

Confusion Matrix

- As previously established, the Decision tree model performed slightly better than the other ones, but the value of false positives is still quite high, and it should be taken into consideration when implementing future decisions



Conclusions

- The technology advances influence the success rate, which is increasing in recent years
- Lighter payloads (2000 – 5000 kgs) seem to have better launch outcomes; heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.
- ES-L1, GEO, HEO, SSO orbits have all a 100% success rate
- Almost half the successful launches come from the launch site KSC LC-39 A (41.7%), followed by CCAFS LC-40, VAFB SLC-4E and finally CCAFS SLC-40
- While sites are quite close to other landmarks, their vicinity to the ocean helps decrease the possibility of debris from launches can cause damages to people and cities
- All models have equivalent performances, with a slightly better result for the Decision tree model; this equivalence in results is most probably due to the small sample of the dependent variable Y data (only 18 entries)
- The Tree Classifier Model performed slightly better among those tested, but the small sample size of the dependent variable cannot allow to reduce the number of false positives any further

Appendix

- Full Github repo Link: https://github.com/beatrice-porcu/IBM-Data-Science-Capstone-Project

Thank you!