Winning Space Race with Data Science

Błażej Kowalski
28 July 2025

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
  - Creating an interactive map using Folium
  - Creating a interactive dashboard using Plotly Dash
  - Predictive analysis
- Summary of all results:
  - Exploratory Data Analysis results
  - Interactive analytics demo screenshots
  - Predictive analysis results

Introduction

- Project background and context

  SpaceX is considered as the most successful company, which provides commercial flights into space, both manned and unmanned missions. The company provides much cheaper services than the competition (62 milion dollars comparing to 165 and more per launch). This cost reduction comes mostly thanks to reuse of the first stage booster. Determining if the booster landing will be successful may help in estimation of the launch cost. Basing on the public information and machine learning, the possibility of reuse of the first stage will be predicted.

- Problems you want to find answers

  - Impact of specific variables, such as payload mass, launch site or orbit, on success of landong of the first tage booster.
  - Is there any growing trend of the numer of successful landings over years.
  - What classification algorithm has the best performace in this case

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Wikipedia Web Scrapping
- Perform data wrangling
  - Data Filtering to keep only Falcon 9 data
  - Dealing with missing data – for missing payload mass, average mass was used
  - One-Hot Encoding to prepare data for 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, testing and fine-tuning classification models to ensure the final one gives best possible results

Data Collection

Data collection proces was a combination of API requests from SpaceX REST API and web scraping data from Wikipedia, using BeautifulSoup API.

Both methods have been used to obtain complete dataset, which will be useful for future steps and prediction.

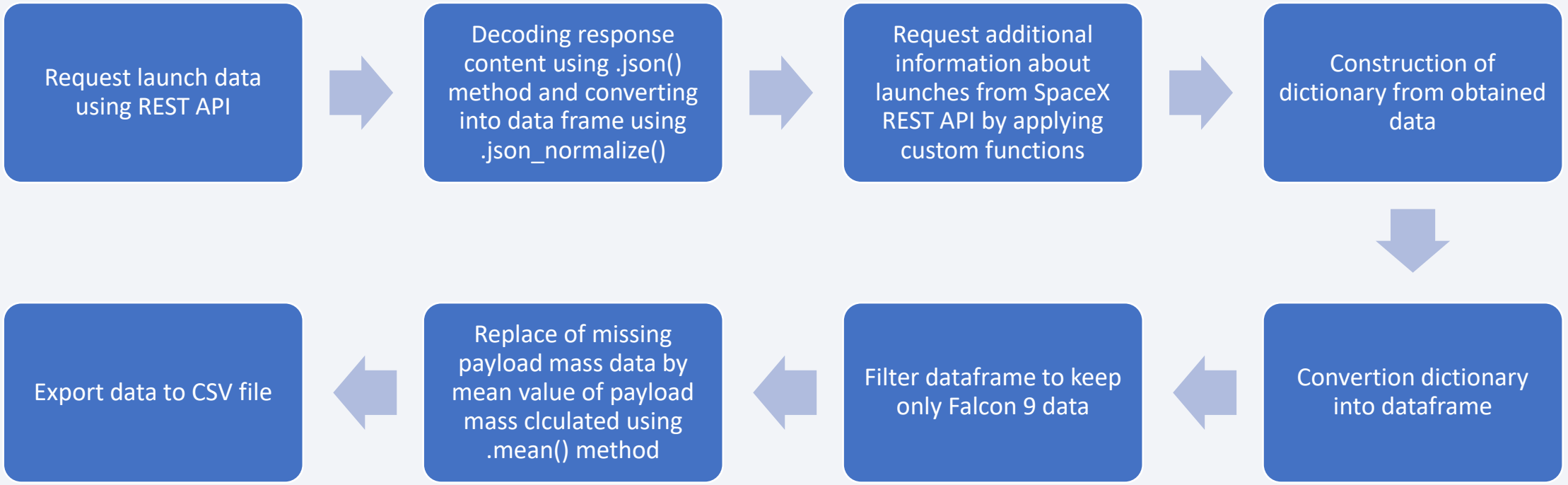
Features obtained from SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Features obtained from Web Scraping:

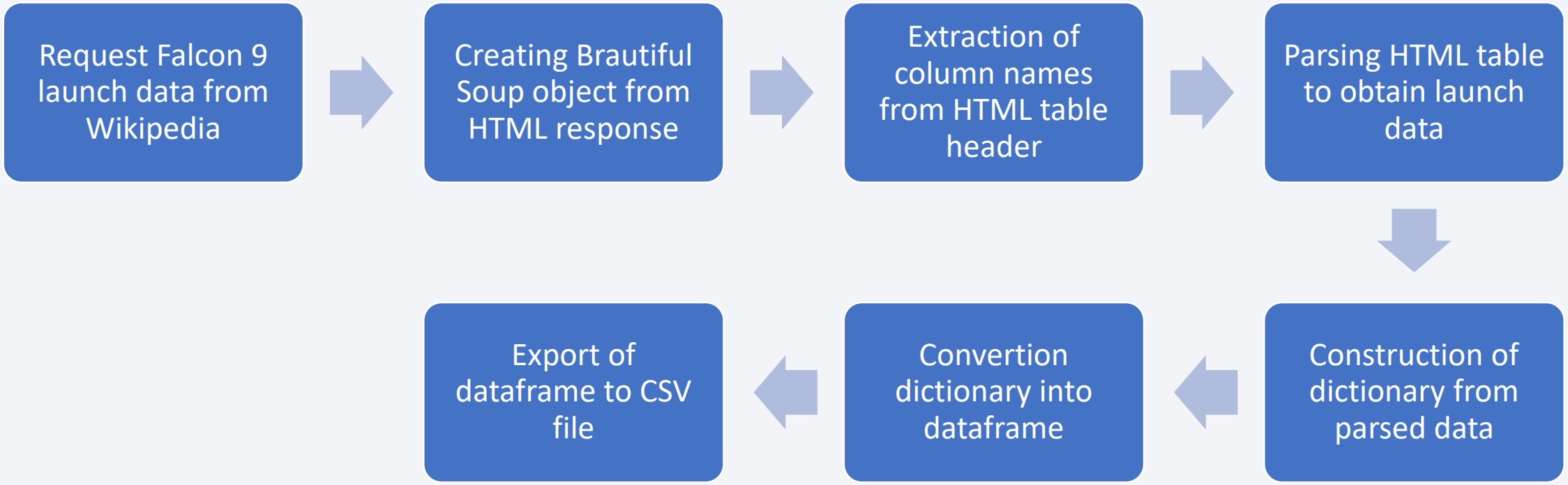
Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub: Data Collection API

Data Collection - Scraping

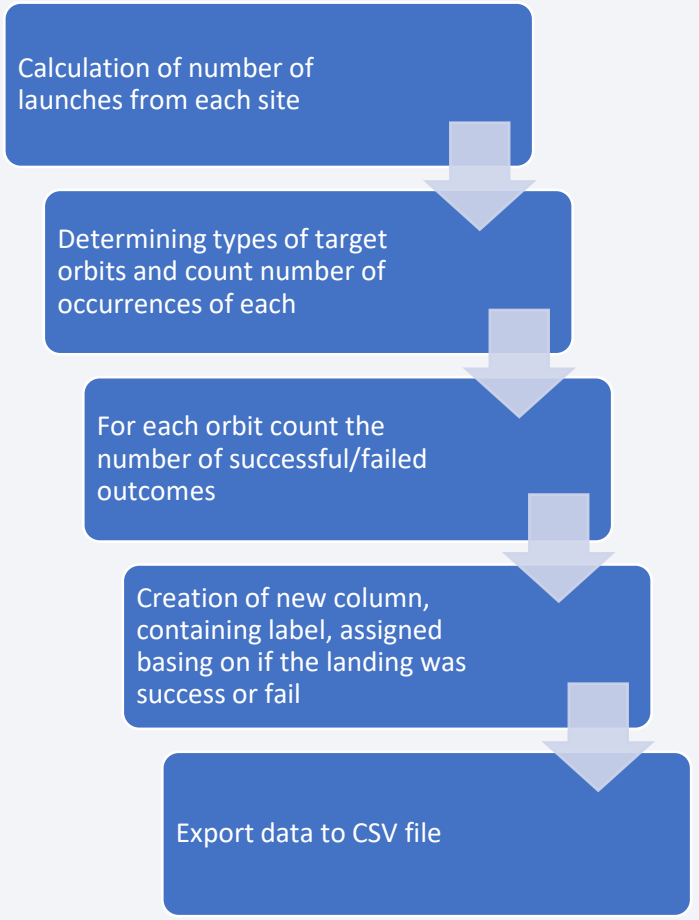


GitHub: Web Scraping

Data Wrangling

There are several outcomes of the landing proces. Not every booster was supposed to land, some of tchem were destroyed on purpose. In this case „landing” in specific area was considered as success and flagged as *True*, while unsuccessful one was flagged as *False*.

Since determining if mission was successful is the main goal, the label stored in new column, has been assigned to each landing – 1 for successful outcome and 0 in case of failure.



EDA with Data Visualization

Charts plotted:

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload Mass vs Launch Site
- Orbit vs Success Rate
- Flight Number vs Orbit
- Payload Mass vs Orbit
- Success Rate Yearly Trend

There were used 3 types of charts:

- Scatter plots to check if there is any relations between variables and if it can be impactful on machine learning model
- Bar plots to allow comparison between discrete categories
- Line plot to show trend over the time.

GitHub: EDA with Data Visualization

EDA with SQL

Performed SQL queries

- Display all unique launch sites
- Display 5 records, where launch site was starting with ‚CCA'
- Display total payload mass carried by boosters by NASA (CRS)
- Display average payload mass carried by boster version F9 v1.1
- Display date of first successful landing on the ground pad
- Display names of the boosters, carrying payload of 4000 to 6000kg, which sucsessfuly landed on the drone ship
- Display counts of specific mission outcomes
- List all booster versions, that carried maximum payload mass
- List failed landings on drone ship in year 2015, including such information as month of landing, booster version and launch site.
- Ranking all of the landing outcomes in descending order, between 2010-06-04 and 2017-03-20.

GitHub: EDA with SQL

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Marker with circle, popup and text label of NASA Space Center in Huston. Longitude and Latitude of this facility was used as starting coordinates
- Markers with circle, popup and text label of all launch sites using their coordinates to show their location and distance to Equator or coastlines

Coloured Markers of launch outcomes for each Launch Site:

- Each launch was presented as separate marker, which color indicates if the outcome of the mission was success (green) or failure (red). Since there were multiple launches from the same location, marker cluster was used. This operation helped to identify launch sites with relatively high success ratio.

Distances between Launch Site and its proximities:

- Added lines to present distance between launch site and coastline, railway, highway or city.

GitHub: Interactive Map with Folium

Build a Dashboard with Plotly Dash

Dropdown list of launch sites:

- Dropdown list allowing to chose specific launch site or see data combined from all sites

Pie chart presenting success/fail ratio:

- Pie chart showing success/fail ratio for selected launch site. In case of data from all launch sites, pie chart presents count of successful outcomes for each launch site

Payload mass slider:

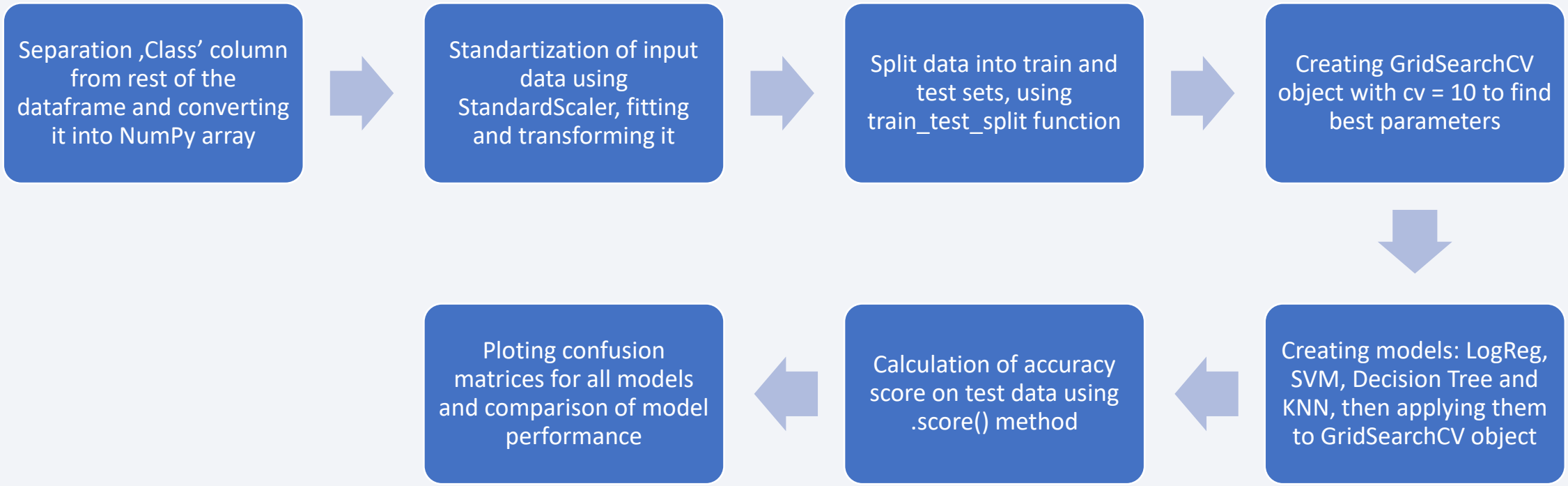
- Slider that allows to select payload mass range

Scatter plot of Payload Mass vs Success Rate for different booster version:

- Scatter plot showing correlation between payload mass, booster version and mission outcome.

GitHub: Dashboard with Plotly Dash

Predictive Analysis (Classification)



GitHub: Predictive Analysis (Classification)

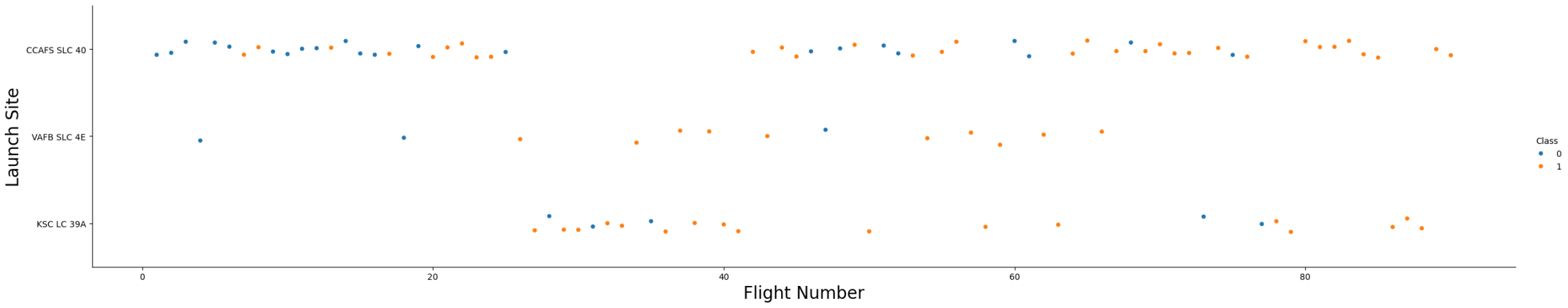
Results

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

Section 2

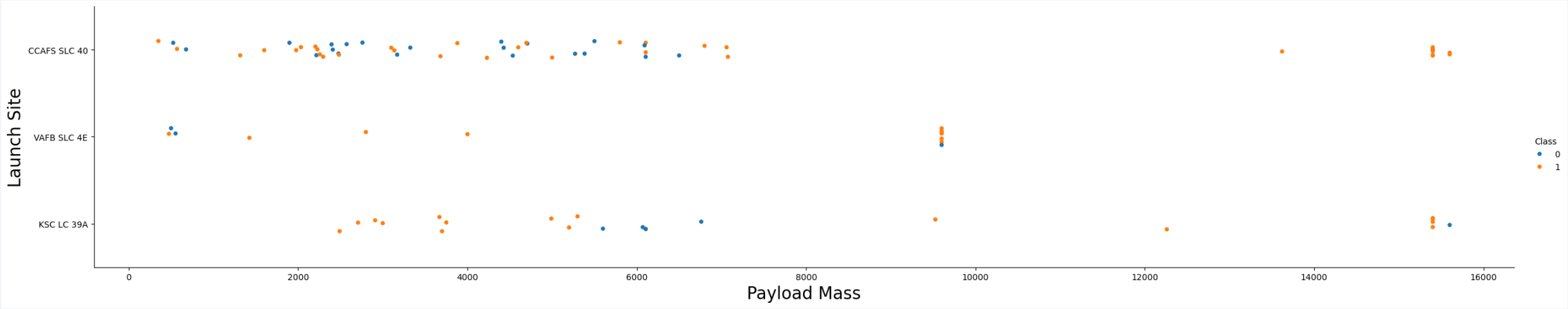
Insights drawn from EDA

Flight Number vs. Launch Site



- Success rate increases with time
- Most launches took place from CCAFS SLC 40
- VAFS SLC 4E and KSC LC 39A have greater success rate comparing to CCAFS SLC 40, yet number of launches from those sites is lower than from the later one.

Payload vs. Launch Site



- Most launches with heavy loads were successful
- There was no launch from VAFB SLC 4E with load exceeding 10000kg
- KSC LC 39A has great success rate for payloads lighter than 5500kg
- Most launches with payload exceeding 7000kg were successful

Success Rate vs. Orbit Type

Orbits with 100% success rate:

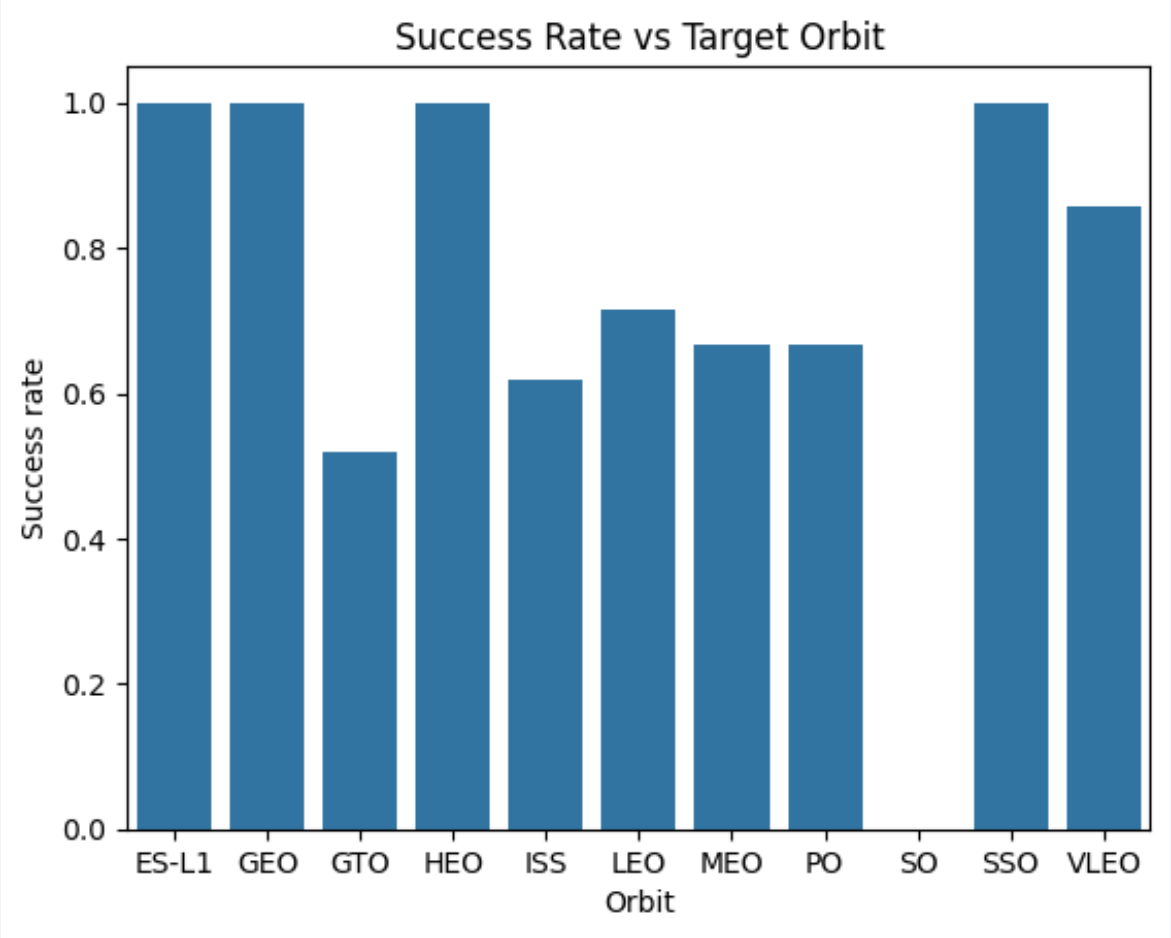
– ES-L1, GEO, HEO, SSO

Orbits with 0% success rate:

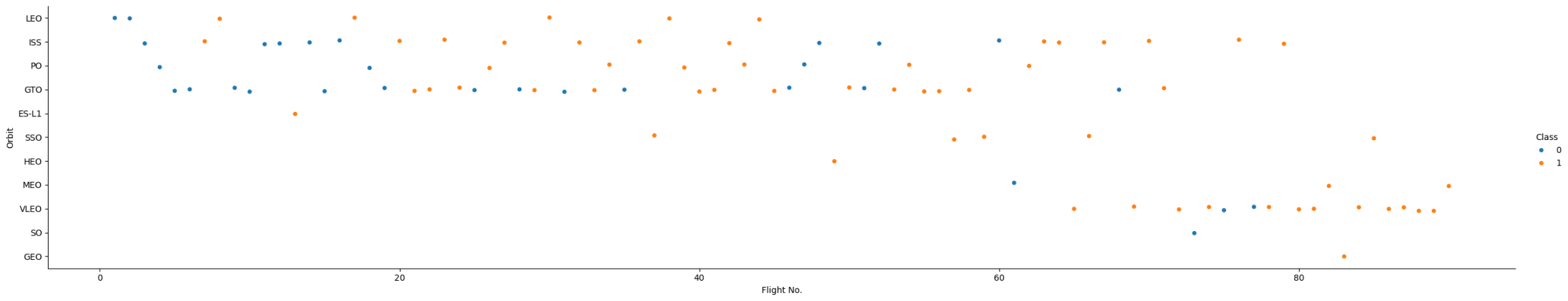
– SO

Orbits with success rate above 60%:

– ISS, LEO, MEO, PO, VLEO

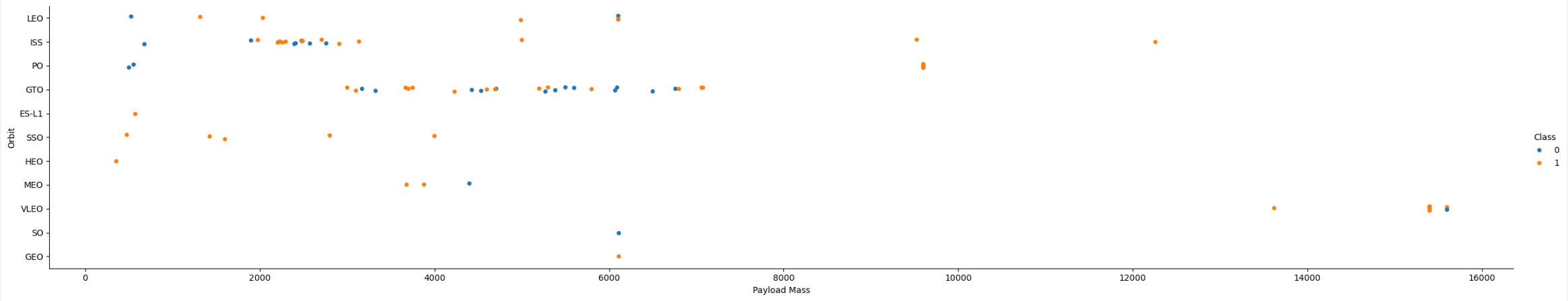


Flight Number vs. Orbit Type



- After 2 failed missions, LEO seems to be quite favorable target. The correlation can be noticed.
- For GTO and ISS, there is no correlation between success and flight number.
- There was only one mission to SO, which is reason, why success rate is 0%

Payload vs. Orbit Type

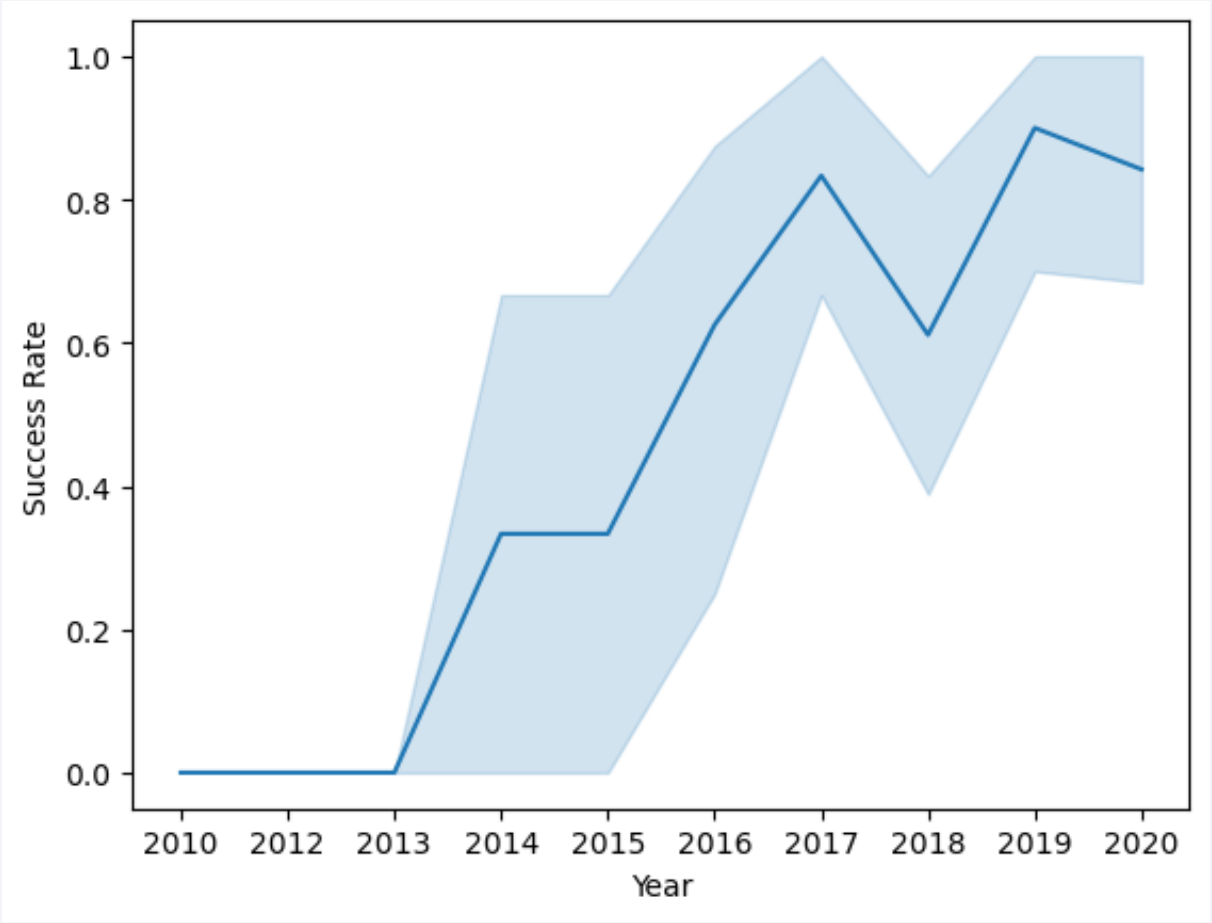


- The heaviest payloads are delivered to VLEO and are mostly successful.
- Delivery of heavy loads for ISS and polar orbit are successful missions

Launch Success Yearly Trend

The success rate keeps growing between 2013 and 2020.

There is notable drop of success rate in 2018, yet it does not impact the growing trend over the years.



All Launch Site Names

```
In [10]: %sql select distinct Launch_Site from SPACEXTBL
```

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

Out[10]:

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

There are 4 unique names in Launch_Site column.

Launch Site Names Begin with 'CCA'

```
In [12]: %sql select * from SPACEXTBL where Launch_Site like "CCA%" limit 5
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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 shows first 5 results for missions, which were launched from launch site, which name starts wit ‚CAA'.

Total Payload Mass

```
In [15]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = "NASA (CRS)"
 * sqlite:///my_data1.db
Done.
Out[15]: sum(PAYLOAD_MASS__KG_)
         45596
```

This query shows the total payload mass carried by SpaceX for NASA (CRS).

Average Payload Mass by F9 v1.1

```
In [17]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like "F9 v1.1%"
 * sqlite:///my_data1.db
Done.
Out[17]: avg(PAYLOAD_MASS__KG_)
         2534.6666666666665
```

Query displays average payload mass caried by F9 v1.1 booster. The average payload mass was about 2535kg.

First Successful Ground Landing Date

```
In [21]: %sql select min(Date) from SPACEXTBL where landing_outcome = "Success (ground pad)"
 * sqlite:///my_data1.db
Done.
Out[21]: min(Date)
         2015-12-22
```

Listing first successful landing on the ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [22]:
```
%%sql
select booster_version from SPACEXTBL where landing_outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and
```

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

Out[22]:

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

This query lists all bossers, that managed to land successfully on drone ship after carrying payload, which weight ranged between 4000 and 6000kg.

Total Number of Successful and Failure Mission Outcomes

```
In [27]: %sql select mission_outcome, count(mission_outcome) from SPACEXTBL group by mission_outcome
 * sqlite:///my_data1.db
Done.
```

Out[27]:

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

Listing total number of successful or failure mission outcomes.

Boosters Carried Maximum Payload

```
In [28]: %%sql
         select booster_version from SPACEXTBL where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

Out[28]:

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

List of all booster versions, that carried maximum payload.

2015 Launch Records

In [29]:
```
%%sql
select substr(Date, 6, 2) as 'Month',landing_outcome, booster_version, launch_site
from SPACEXTBL
where landing_outcome = 'Failure (drone ship)'
and substr(date, 0, 5) = '2015'
```

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

Out[29]:

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

List of failed landings on drone ship in 2015.

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

In [13]:
```
%%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;
```

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

Out[13]:

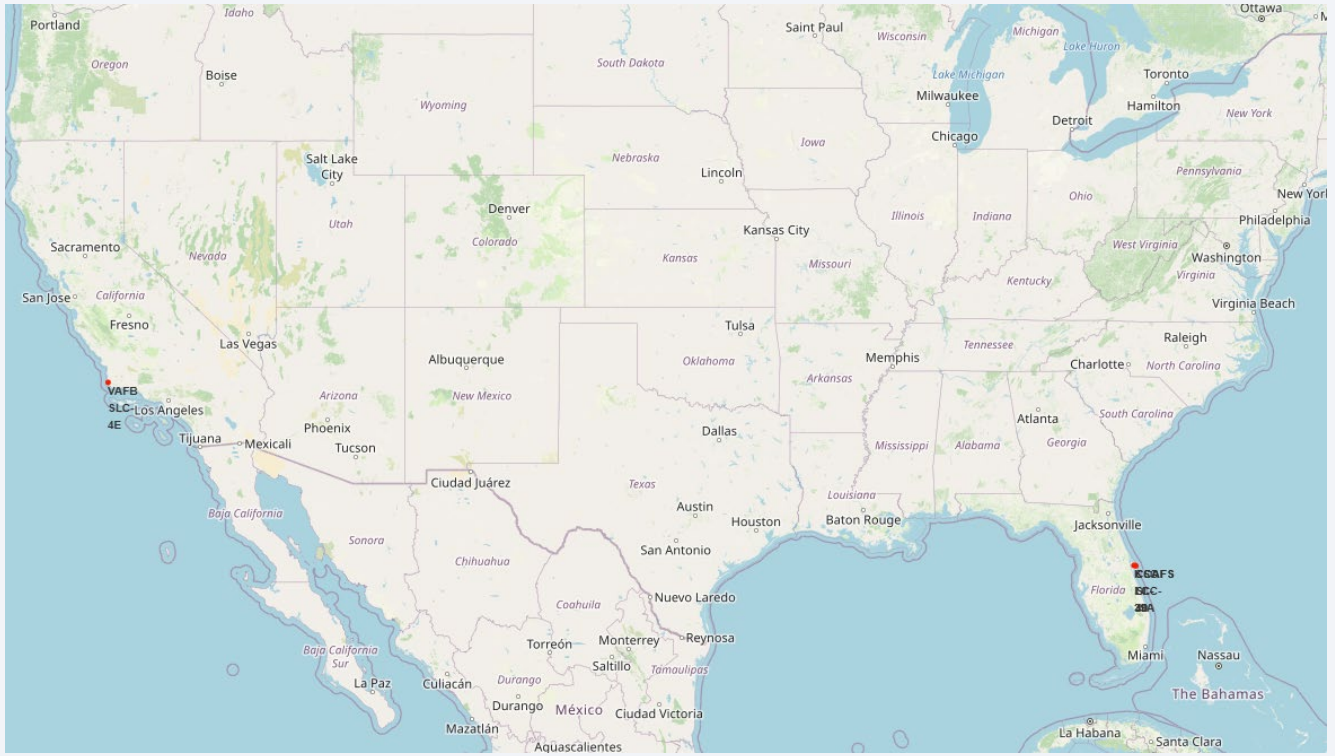
Landing_Outcome	Count_
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Ranking of frequency of mission outcomes in descending order between 2010-06-04 and 2017-03-20.


Section 3
Launch Sites
Proximities Analysis

Launch Sites' Locations

- All launch sites aim to be as close to the Equator as it is possible. On the Equator velocity of the Earth's Surface is the highest and equals 1670km/h. Object launched from proximity of Equator already has this velocity, which is beneficial in terms of fuel burnt.
- The shown launch sites are located close to the coastline to reduce risk of debris falling on the ground and causing casualties in case of malfunction or accident

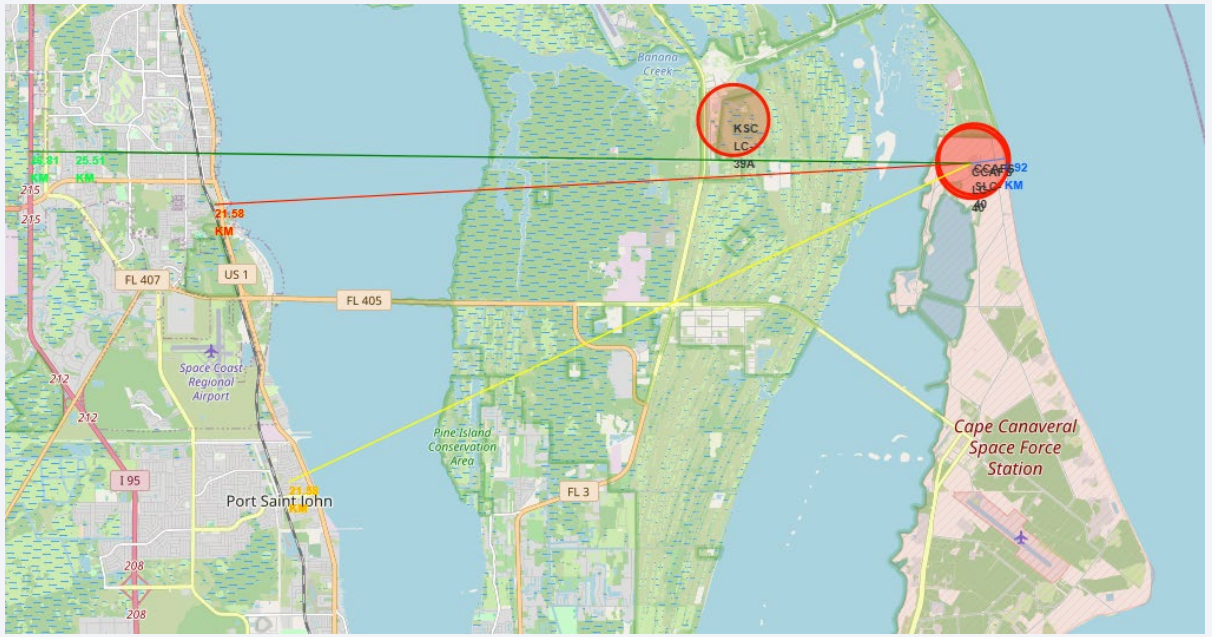


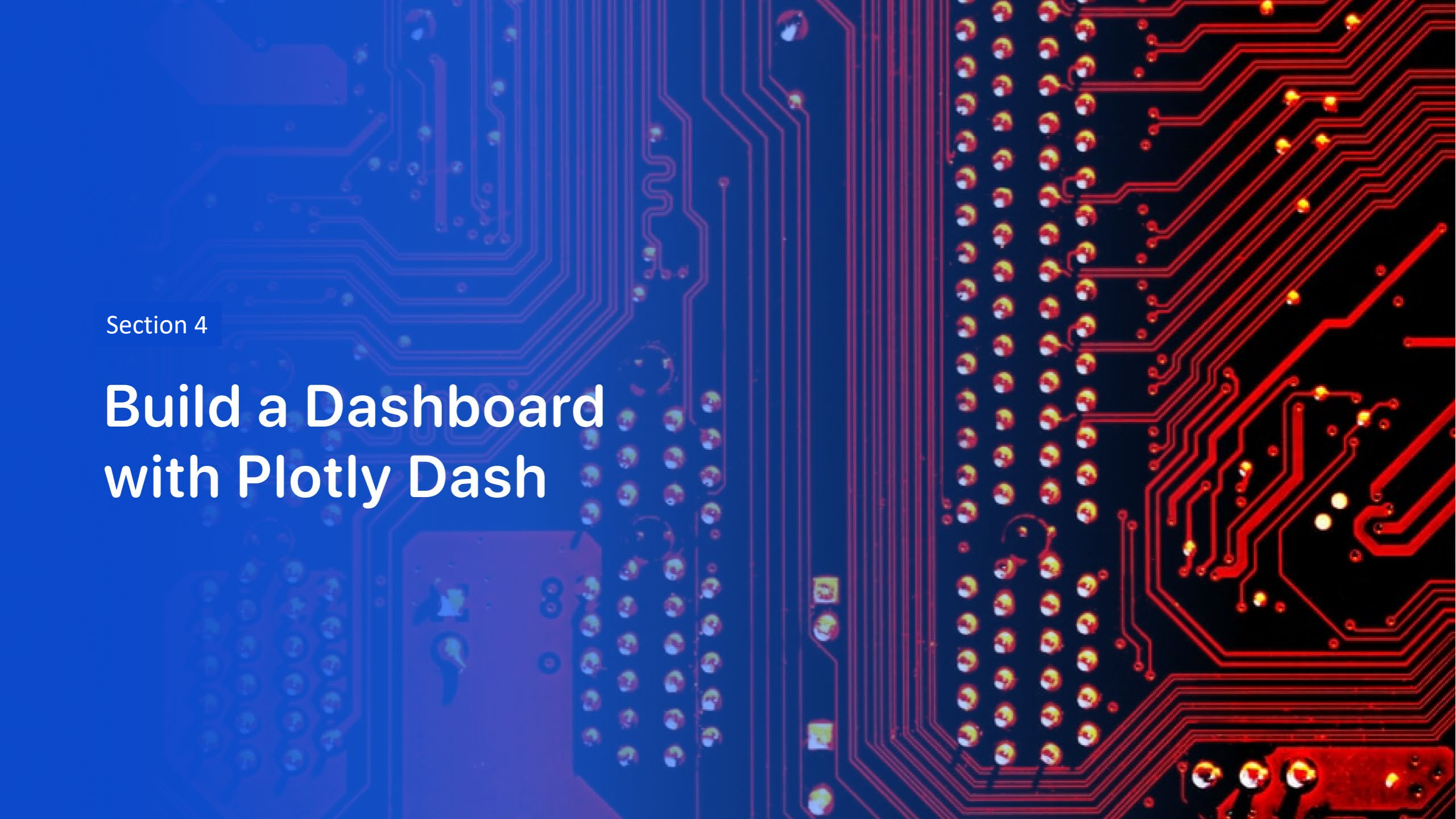
Mission Outcomes Presented on Map

- Markers with color label allow to present mission outcome directly on the map:
  - Green = success
  - Red = Fail
- To keep map readable, marker cluster has been used, since many markers are assigned to the same coordinates.
- KSC LC-39A seem to have very positive success ratio.

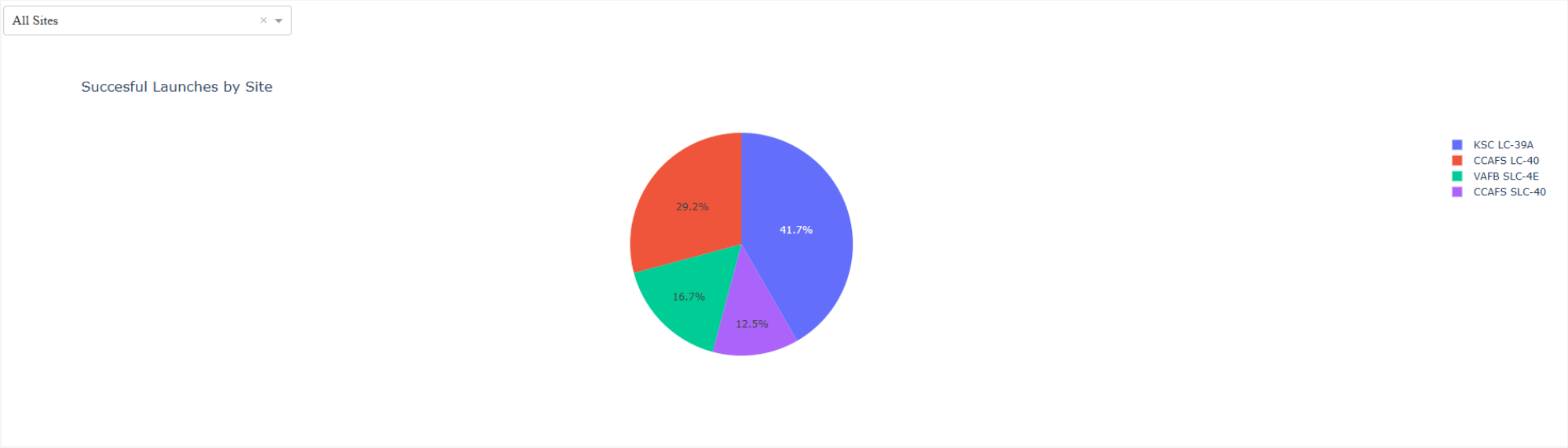
Launch Site Proximities CCAF-LC40

- Launch sites require good communication to provide possibility to deliver payloads. Hence proximity of highways and/or railways is beneficial.
- Presence of human settlements may be also dangerous for people living there. However launch sites cannot be placed in deserted places due to the need of ground staff.



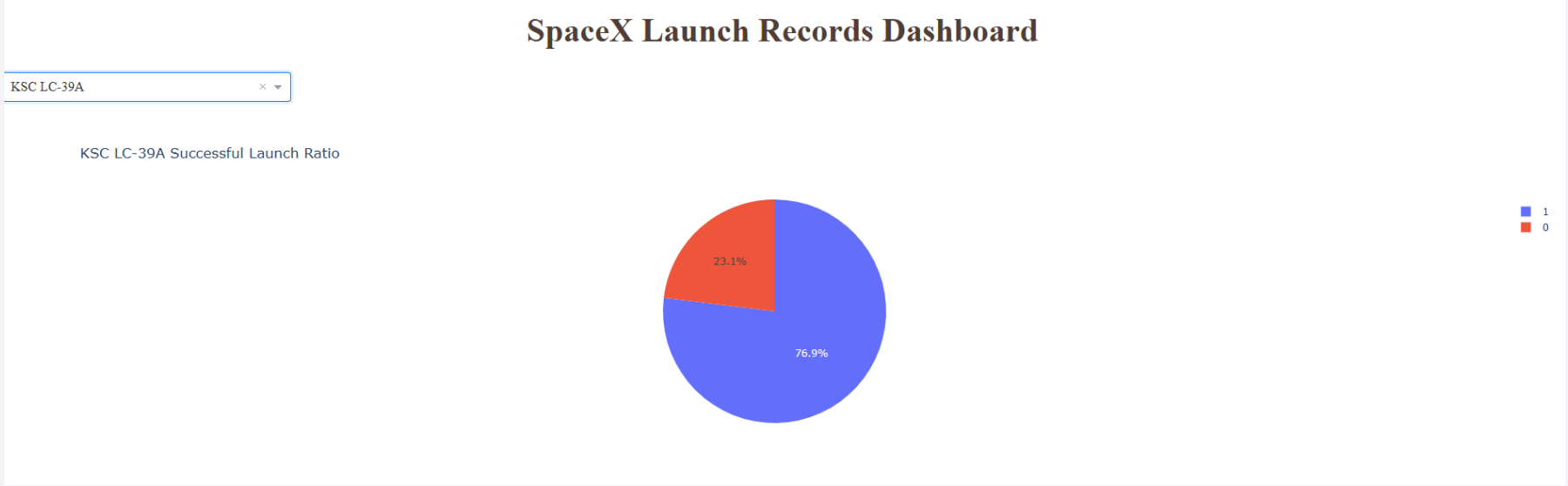

Section 4
Build a Dashboard with Plotly Dash

Launch Site Share in Succesfull Launches



Presented pie chart shows share of each launch site in successful launches.

Success Rate for Launch Site with Highest Success Rate



Success ratio for KSC-LC39A site, which has the highest success rate amongst all launch sites

<Dashboard Screenshot 3>

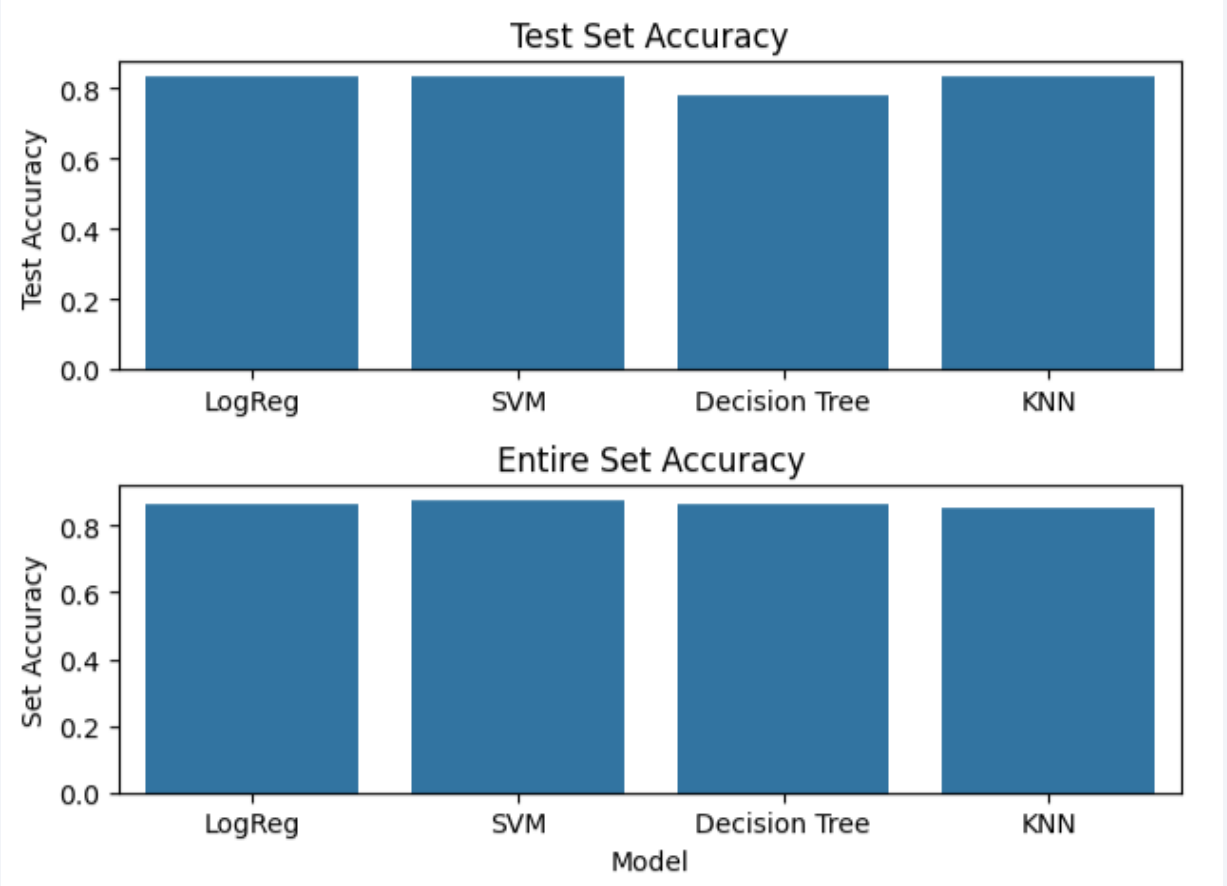
- Majority of payload mass ranged between 2500 kg and 6000kg
- Mission outcomes for the heaviest payloads (>6000kg) are rare – there is only one such case.
- For light payloads (<4000kg) the success rate is distributed in similar proportions or even with some advantage of successful outcomes

Section 5

Predictive Analysis (Classification)

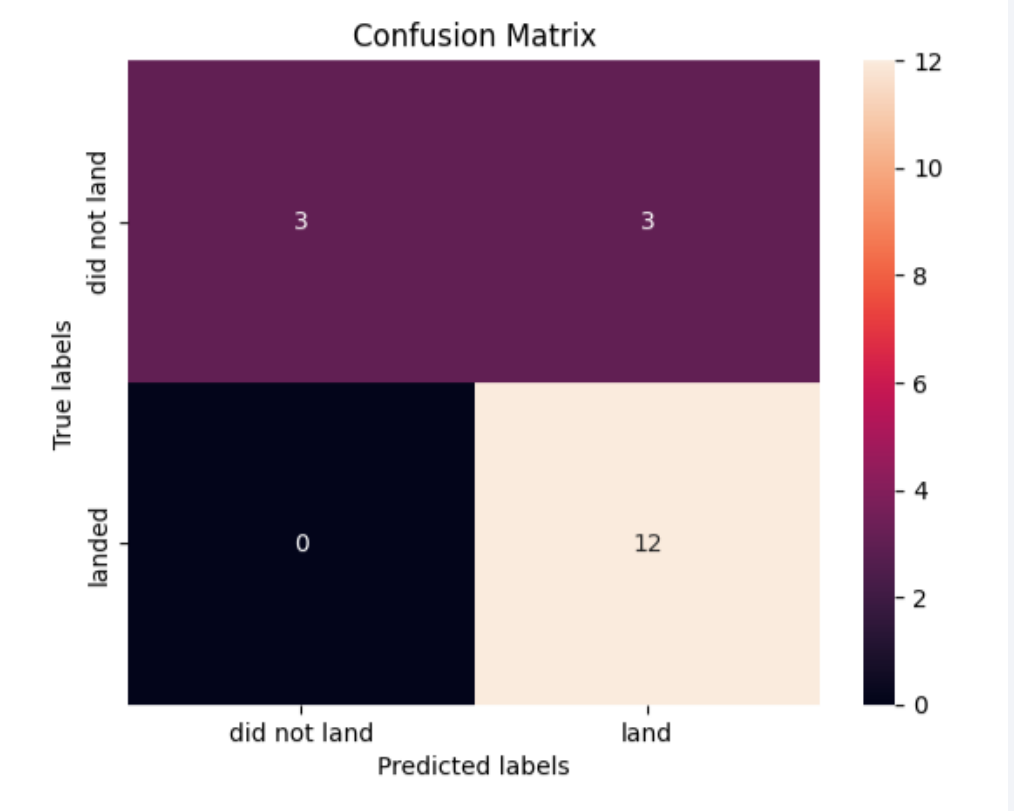
Classification Accuracy

- For test set all models perform well and have the same accuracy (83.33%), except Decision Tree, which has slightly worse accuracy (77.78%). This may be caused by very small test set of 18 sampes
- Accuracy for entire set is similar for all models, with best performance of SVM (87.78%), followed by Decision Tree and Logistic Regression (86.67%). KNN has worst accuracy of 85.56%



Confusion Matrix

- Confusion matrices for all models are the same. Models seem to tend to classify some part of results as false positive.



Conclusions

- Launches with low payload have greater success rate than the heavier payloads
- Launch sites aim to be as close to Equator as possible. Also location well comunicated and close to the coastline are the most beneficial ones
- There is growing trend in missions' successful outcome rate
- The best success to failure ratio is achieved on KSC LC-39A launch site
- ES-L1, GEO, HEO and SSO orbits have 100% succes rate
- SVM seem to have best accuracy amongst all tested models

Thank you!