IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Dmytro Kainara>
<02.10.2021>

Outline

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

Executive Summary

Summary of methodologies

The research was carried out in the Jupyter notebook environment using Python and SQL. he following libraries were used to perform tasks: Pandas, NumPy, scikit-learn, Matplotlib, seaborn, etc. Machine learning methods were used to achieve the objective (Logistic Regression, Support Vector Machine, k-nearest neighbors, Decision Tree Classifier)

Summary of all results

The study found a correlation between successful launches and the conditions for launching SpaceX rockets. The most effective prediction models have shown are support vector machine and k-nearest neighbors. The accuracy of the models are 0.78

Introduction

Project background and context

The official cost of launching a Falcon 9 rocket is $62 million. However, if the first stage is lost, the cost of the launch could be $162 million. So if you can predict the probability of the first stage of a missile being lost, you can calculate the full cost of a launch.

Problems of the research

- Substantiate the relationship between launch conditions and successful missile launches
- Develop an effective model for predicting successful/unsuccessful missile launch as a function of launch conditions

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Description of data collection
- Process of data wrangling
  - Description of data processing
- Process of exploratory data analysis (EDA) using visualization and SQL
- Process of interactive visual analytics using Folium and Plotly Dash
- Process of predictive analysis using classification models
  - Description of building, tuning, evaluating classification models

Data Collection

The API from the official website of SpaceX was used for data collection.

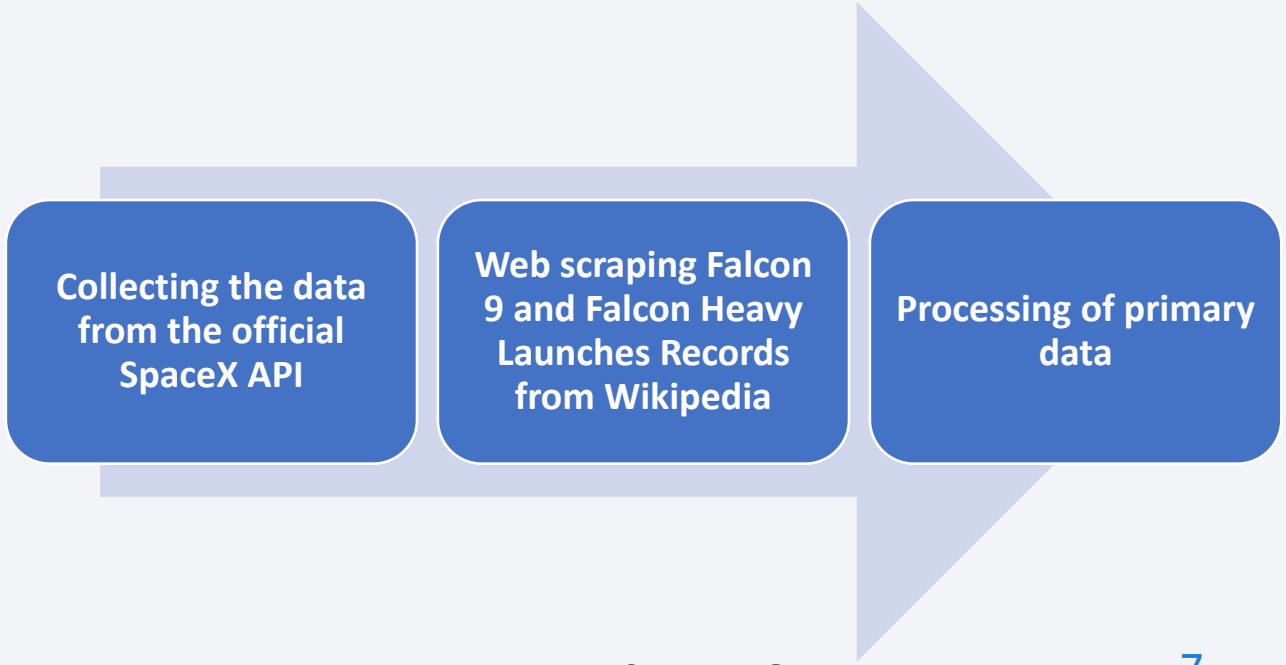
Also, a parsing of data from the website wikipedia.org was used.

The sources on data collection

is available at links:

Link1

Link2



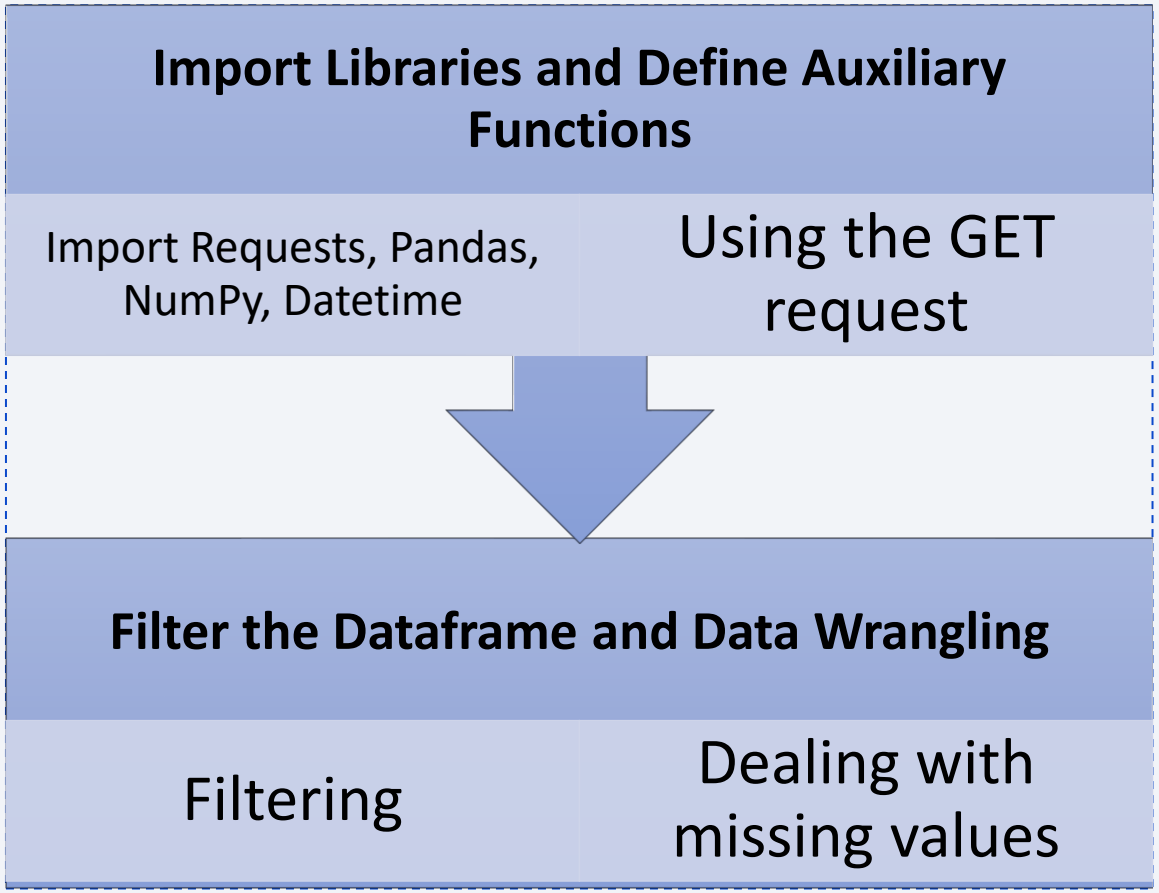
The process of Data Collection

Data Collection – SpaceX API

The aim of data collection stage are collection and processing of primary information on conditions of launch of SpaceX rockets

The official SpaceX API was used to collect data on launch conditions. Pandas, requests, numpy, datetime libraries were used to realize the goal.

Details of the data collection process is available at links: GitHub Link

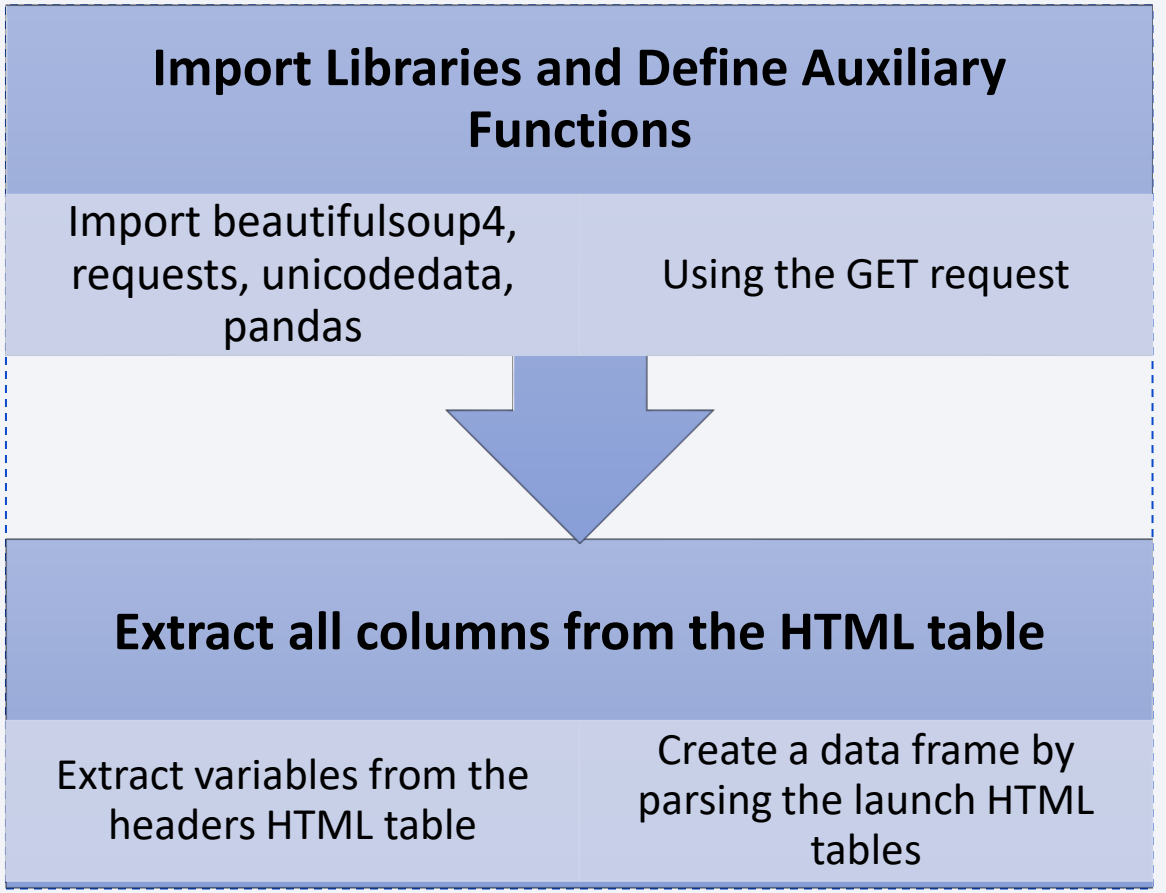


Data Collection - Scraping

The aim of Scraping stage is collection and processing of primary information from Wikipedia

Wikipedia contains information about launch site, orbit and launch results.

Details of the data collection process is available at links: GitHub Link



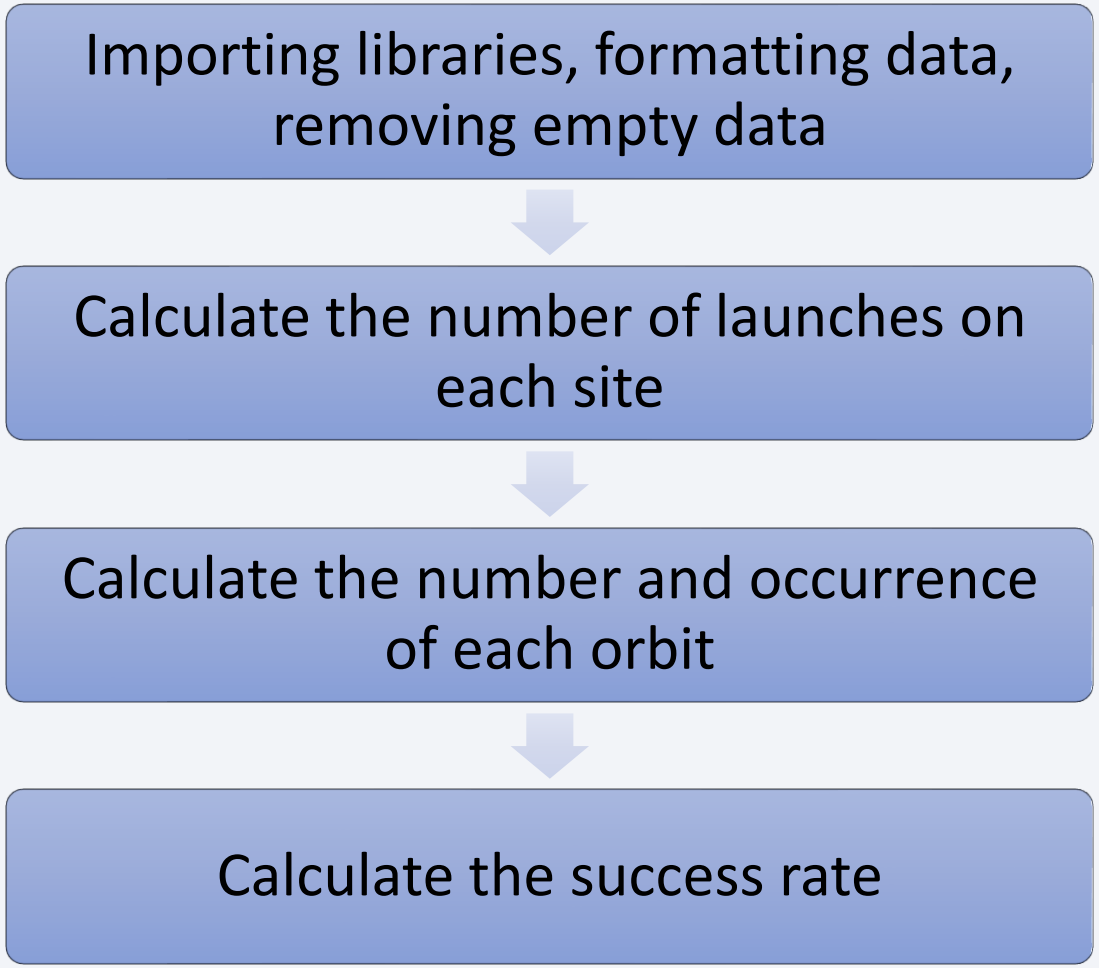
Data Wrangling

The goal of Data wrangling is to analyze SpaceX's basic missile launch indicators

First of all the necessary libraries were imported, empty values were excluded and formatting to a common standard.

The number of launches from different sites was further calculated and the orbits of the launches were analyzed. At that stage, it was determined that the success of missile launches was 0.67.

Details of the data wrangling process is available at links: GitHub Link



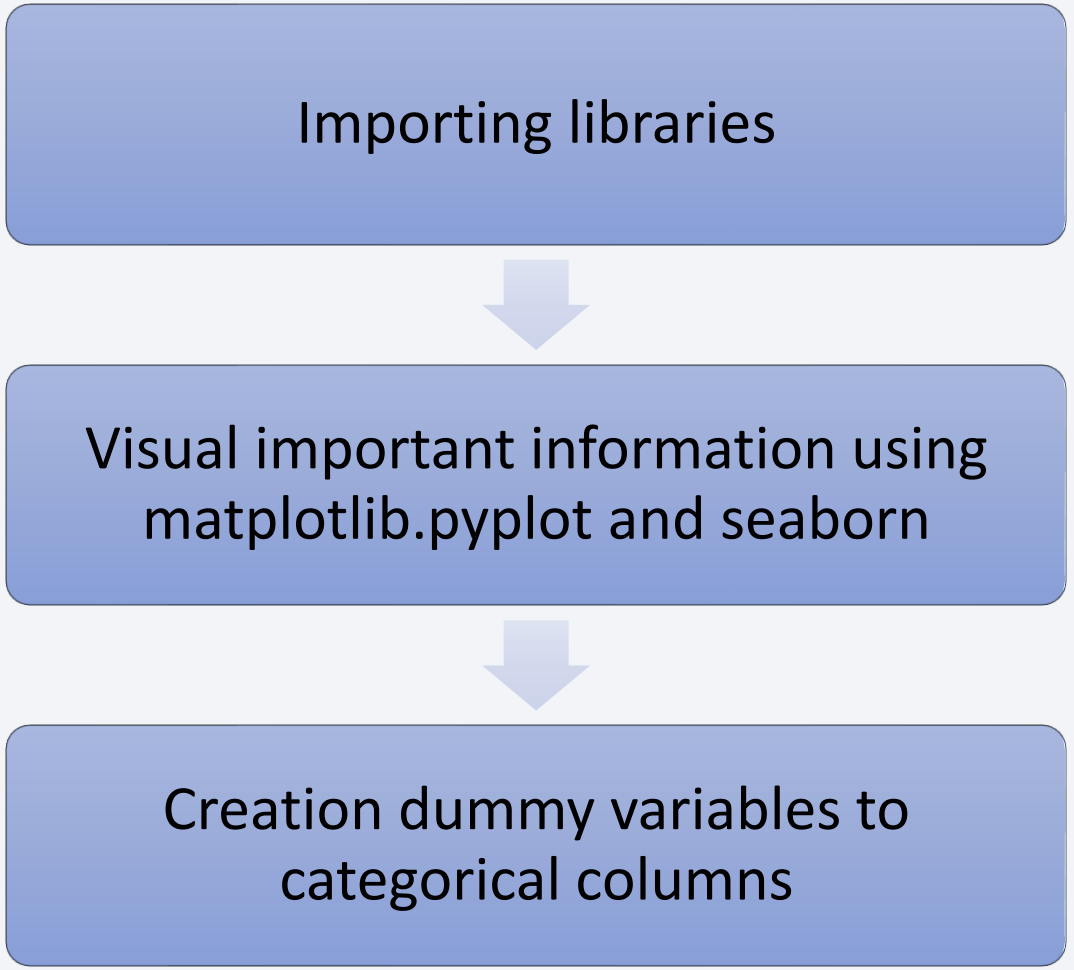
EDA with Data Visualization

The goal of Data Visualization is the visual analyze SpaceX's basic missile launch indicators

For data visualization, matplotlib.pyplot and seaborn libraries were used. After the basic graphics were created, there were dummy variables to categorical columns.

It were visualized the relationship between Flight Number and Launch Site, the relationship between Payload and Launch Site, the relationship between success rate of each orbit type, the relationship between FlightNumber and Orbit type, the relationship between Payload and Orbit type, the launch success yearly trend. The analysis makes it possible to assess the effectiveness and characteristics of missile launches in a dynamic manner, taking into account the conditions of the launches.

Details of the data visualization process is available at links: GitHub Link



EDA with SQL

The goal of EDA with SQL stage is creation and analysis SQL database.

Results of SQL queries:

- The number of launch sites used is: CCAFS LC-40 – 26, CCAFS SLC-40 – 34, KSC LC-39A – 25, VAFB SLC-4E - 16
- The total payload mass carried by boosters launched by NASA is 45596 kg
- The average payload mass carried by booster version F9 v1.1 is 2534 kg
- The date when the first successful landing outcome in ground pad was achieved is 2015 – 12 – 22
- There are four boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- There are 100 successful and 1 failure mission outcomes
- F9 B5 is carried the maximum payload mass

Details of the EDA with SQL is available at links: GitHub Link

Build an Interactive Map with Folium

The goal of building an interactive map with folium stage is analysis launch sites used by SpaceX

In order to achieve this objective, the following measures have been taken:

- Mark all launch sites on a map (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E)
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

Details of the Interactive Map with Foliumis available at links:
GitHub Link

Build a Dashboard with Plotly Dash

The goal of building a Dashboard with Plotly Dash stage is creating a visual tool for online data analysis.

- It was success-pie-chart and scatter chart to show the correlation between payload and launch success added to Plotly Dash.
- Success-pie-chart and scatter chart can show the main information about dynamics of launches

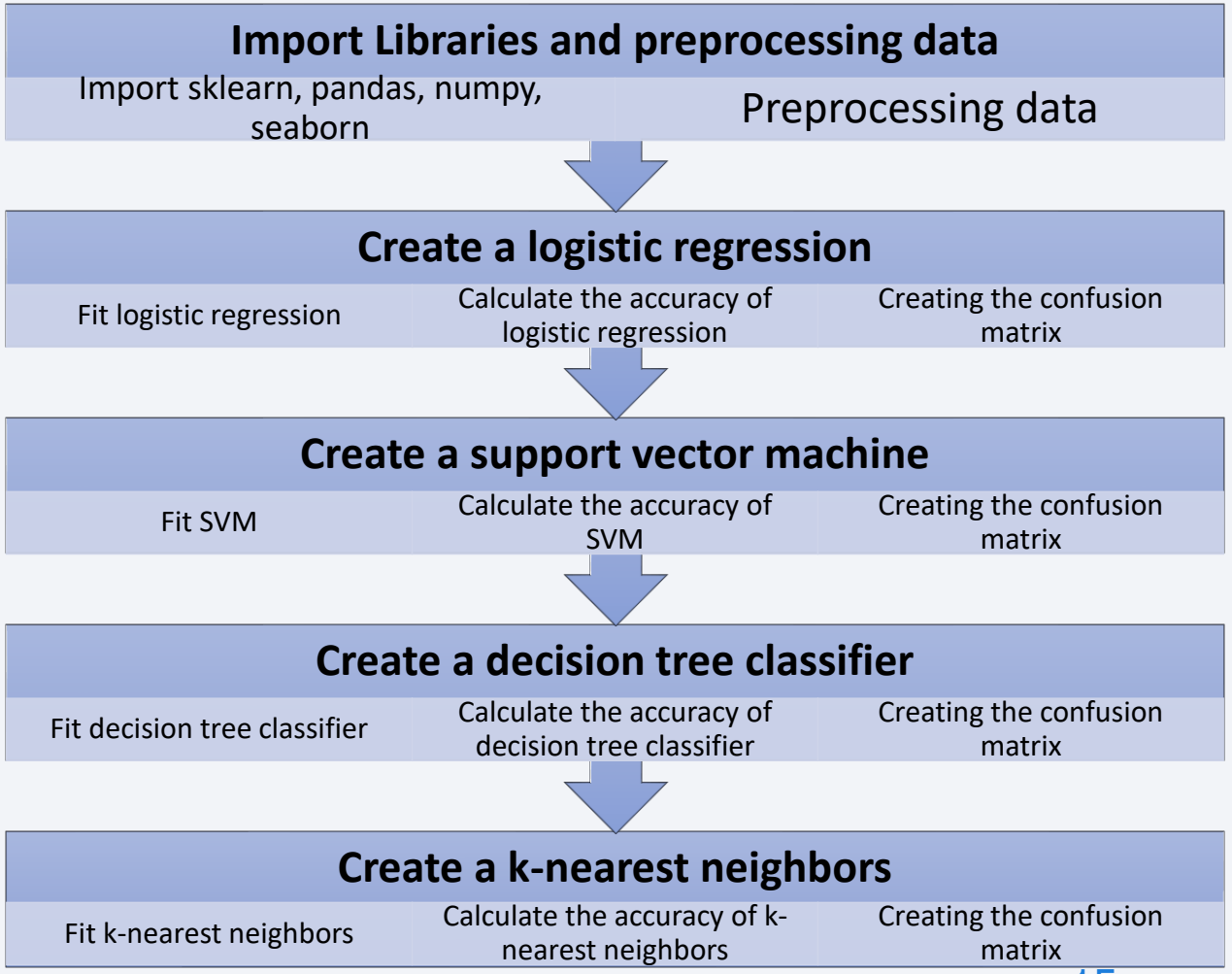
Details of the Dashboard with Plotly Dash stage available at links: GitHub Link

Predictive Analysis (Classification)

The goal of Predictive Analysis stage is creating a model to predict successful and unsuccessful launches

Four machine learning algorithms were used to create a predictive model. As a result, it was found that SVM and k-nearest neighbors performed the most efficiently

Details of the Predictive Analysis stage available at links: GitHub Link

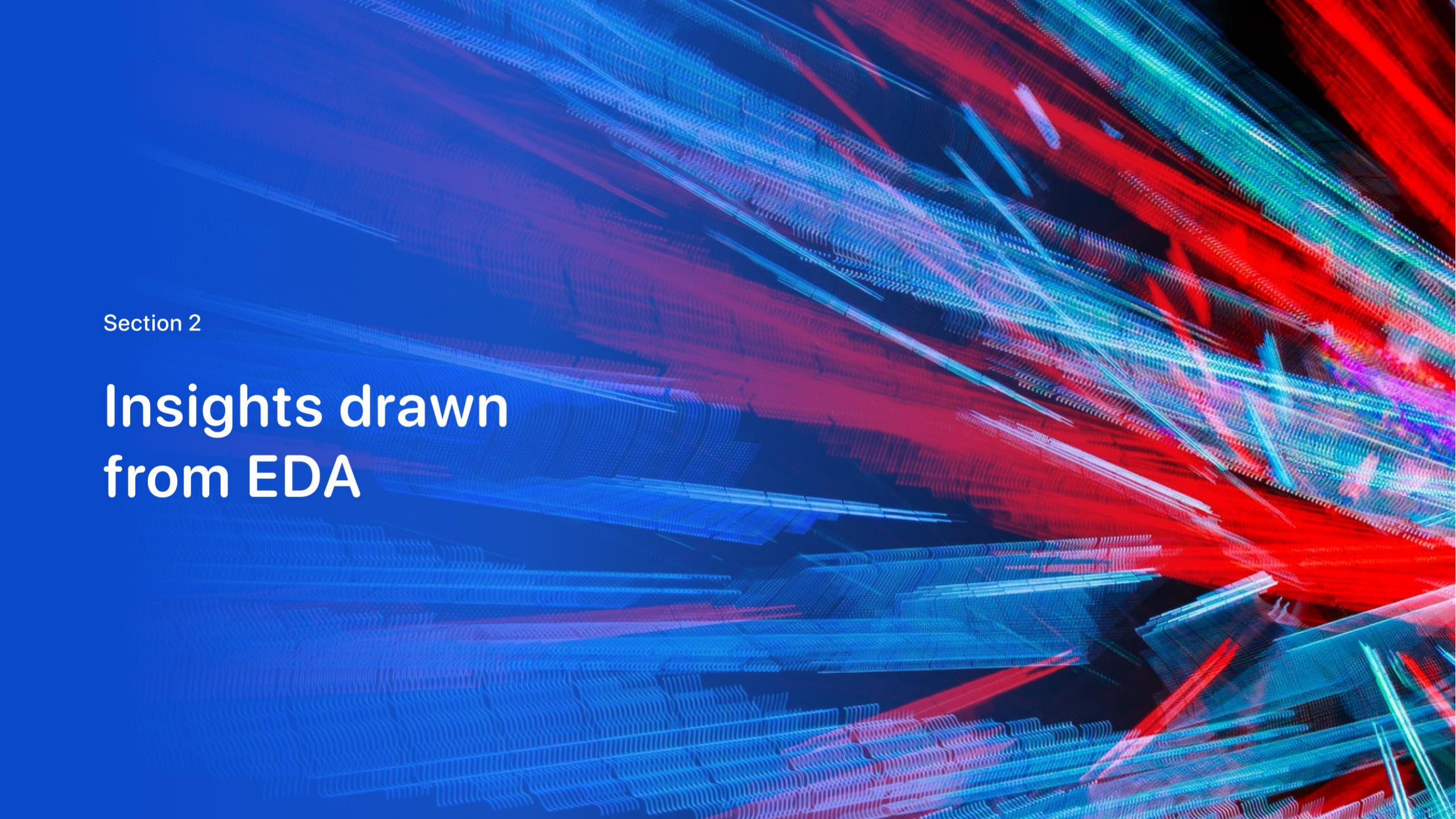


Results

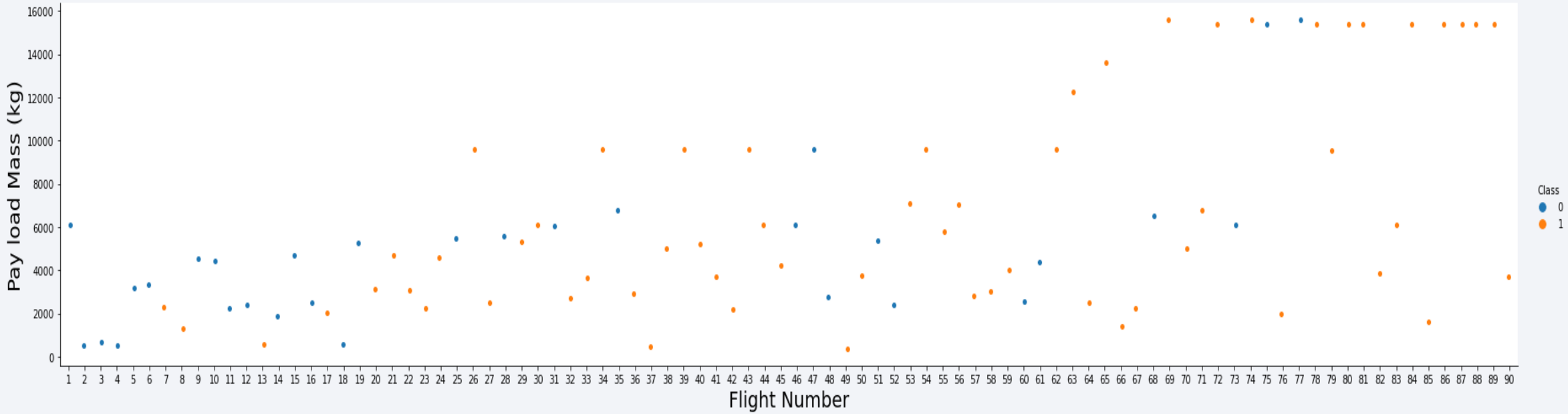
- The relationship between launch conditions and launch results has been determined.
- Interactive analytics was created for quick analysis of information about launches
- As a result of the research two equivalent models of machine learning based on SVM and k-nearest neighbors were built. The accuracy of models are 0.77.

Section 2

Insights drawn from EDA

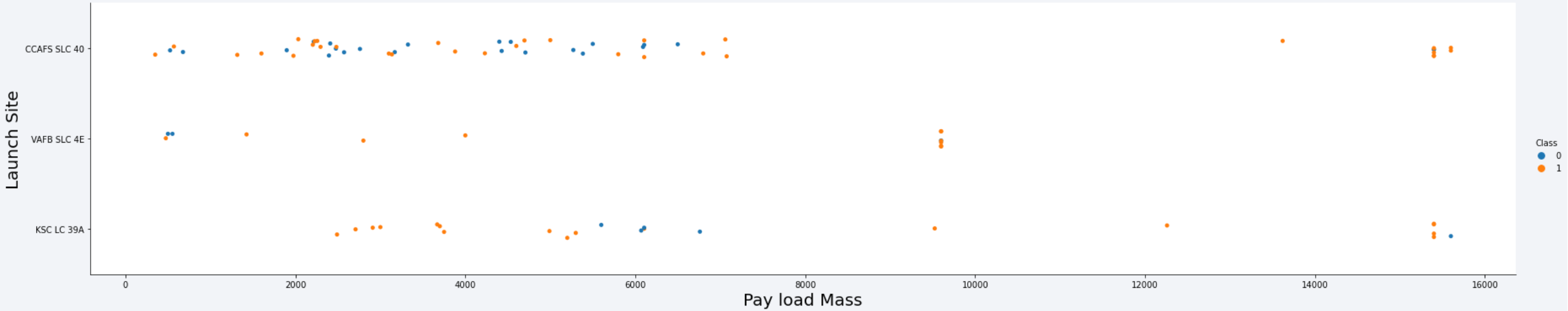


Flight Number vs. Launch Site



We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.

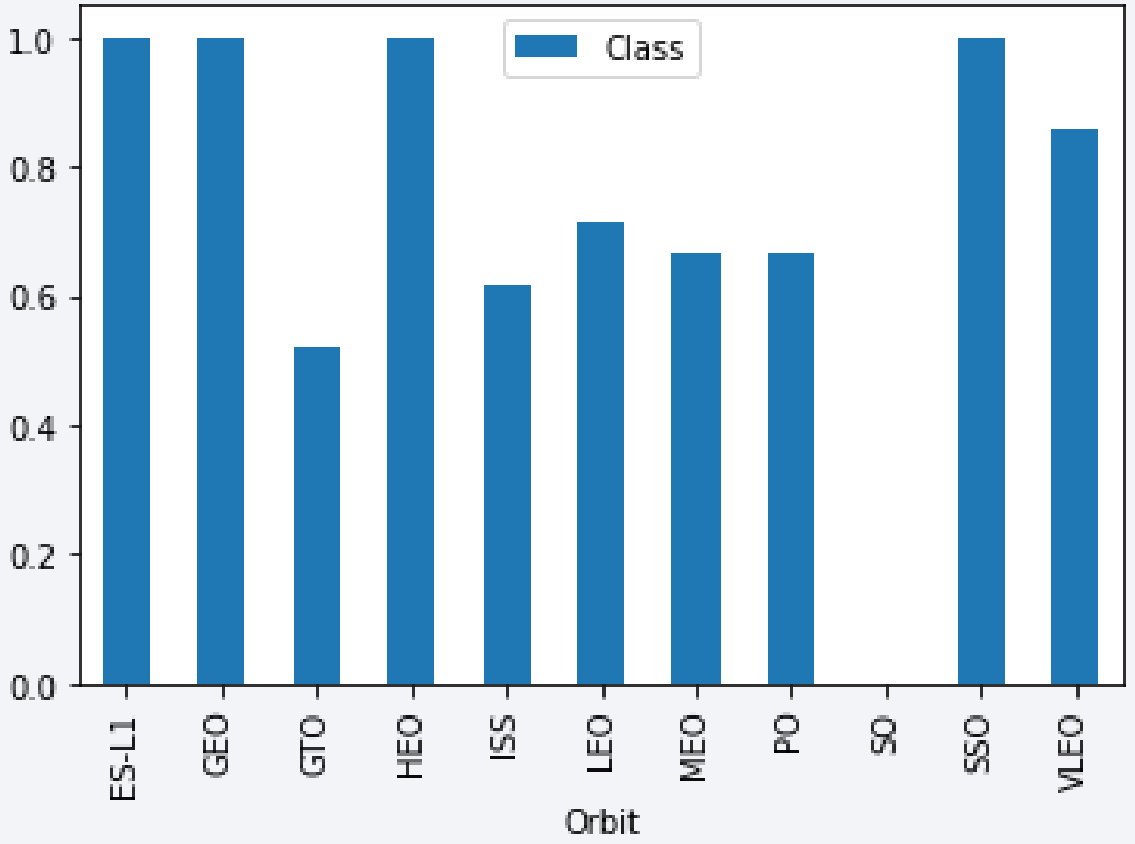
Payload vs. Launch Site



We can made the following conclusions:

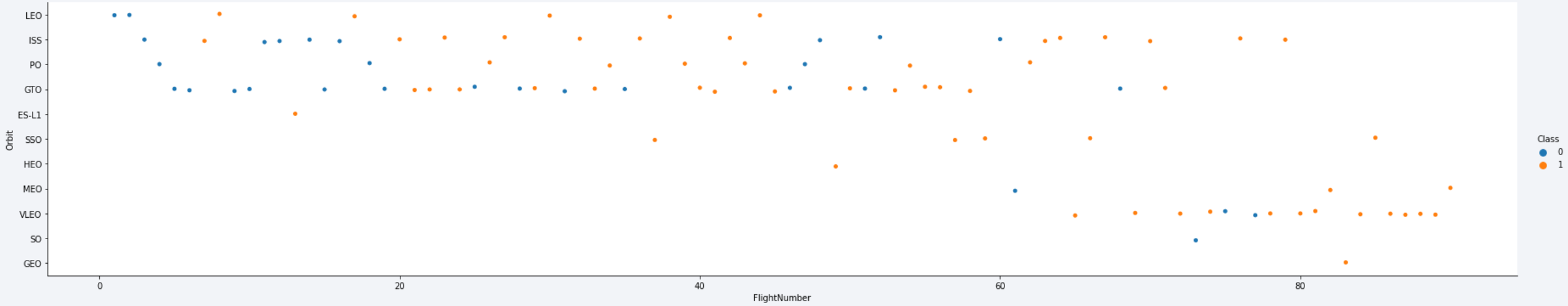
1. It`s a correlation between launch site "CCAFS SLC 40", pay load mass and unsuccessful launches: the most often unsuccessful launches accounted for CCAFS SLC 40 with a pay load mass from 500 to 6000.
2. Rocket launches from VAFB SLC 4E and KSC LC 39A demonstrate a high level of reliability
3. Rocket launches with pay load mass from 8000 to 16000 are succesful in most cases

Success Rate vs. Orbit Type



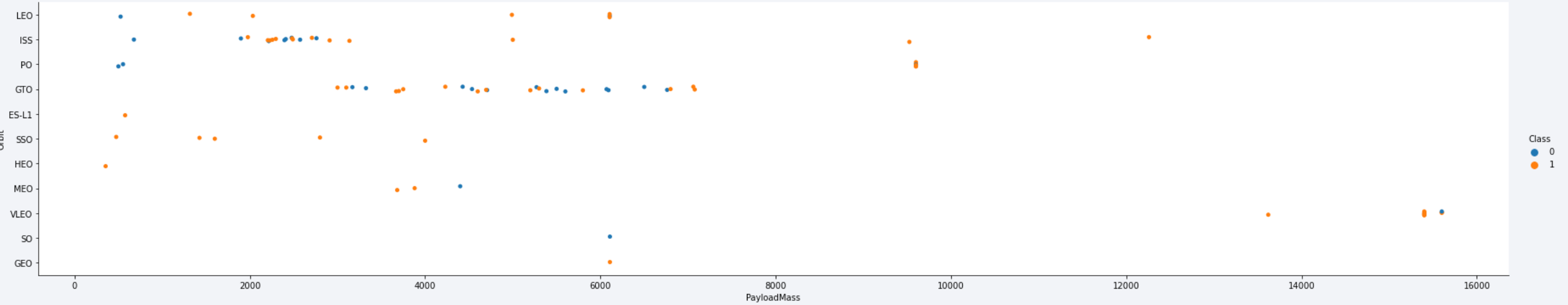
The most sucessful orbit are: ES-L1, GEO, HEO and SSO

Flight Number vs. Orbit Type



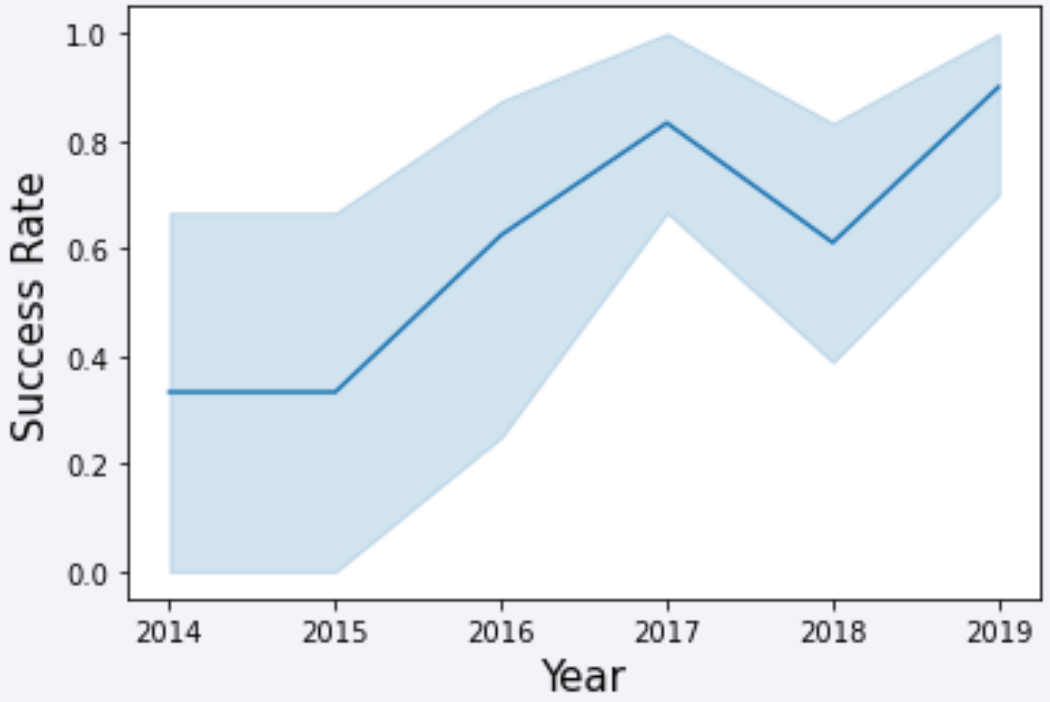
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



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

All Launch Site Names

```
%sql SELECT launch_site, COUNT (launch_site) FROM SPACEX GROUP BY launch_site
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

launch_site	2
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

IT WAS DETERMINED THAT 4 SPACE LAUNCH SITES WERE USED

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEX WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

As a result of executing the SQL query, the first 5 starts with `CCA` were displayed

Total Payload Mass

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

```
%sql SELECT SUM (payload_mass__kg_) FROM SPACEX WHERE customer LIKE 'NASA (CRS)'
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
45596

The total payload mass carried by boosters launched by NASA is 45596 kg

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG (payload_mass__kg_) FROM SPACEX WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
2534

The average payload mass carried by booster version F9 v1.1 is 2534 kg

First Successful Ground Landing Date

```
%sql SELECT MIN (DATE) FROM SPACEX WHERE landing__outcome LIKE 'Success (ground pad)'
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
2015-12-22

The date when the first successful landing outcome in ground pad was achieved is 2015 – 12 – 22

Successful Drone Ship Landing with Payload between 4000 and 6000

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 SPACEX WHERE landing__outcome LIKE 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

There are four boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 (F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2)

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT(mission_outcome) FROM SPACEX WHERE mission_outcome LIKE '%Success%'
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
100

```
%sql SELECT COUNT(mission_outcome) FROM SPACEX WHERE mission_outcome LIKE '%Failure%'
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

1
1

There are 100 successful and 1 failure mission outcomes

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT booster_version FROM SPACEX WHERE payload_mass__kg_ LIKE (SELECT MAX (payload_mass__kg_) FROM SPACEX)
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

There are 12 Boosters Carried Maximum Payload

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT landing__outcome, booster_version, launch_site       FROM SPACEX WHERE YEAR (DATE) LIKE '2015'
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
No attempt	F9 v1.1 B1014	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
No attempt	F9 v1.1 B1016	CCAFS LC-40
Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
Success (ground pad)	F9 FT B1019	CCAFS LC-40

- There 7 ailed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
[108]: %sql SELECT * FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY landing__outcome DESC
```

```
 * ibm_db_sa://mrk09262:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

[108]:

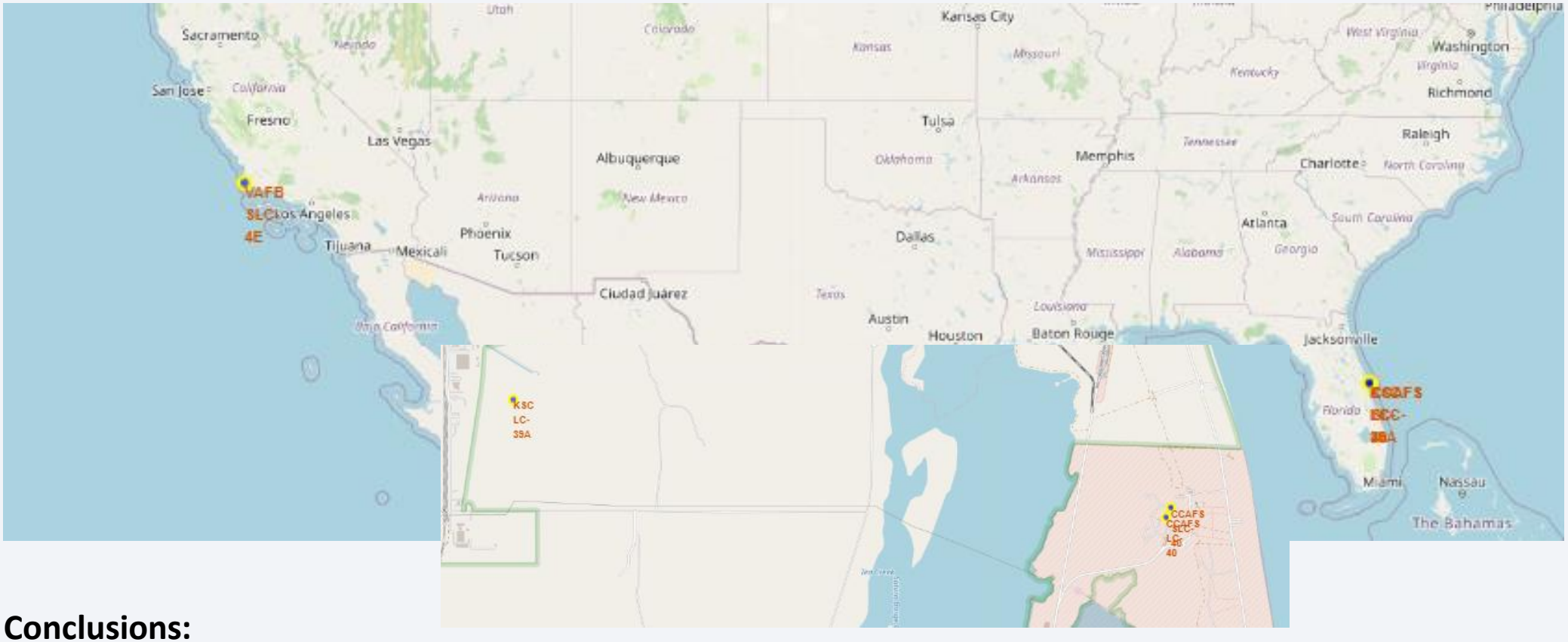
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2013-09-29	16:00:00	F9 v1.1 B1003	VAFB SLC-4E	CASSIOPE	500	Polar LEO	MDA	Success	Uncontrolled (ocean)
2014-09-21	05:52:00	F9 v1.1 B1010	CCAFS LC-40	SpaceX CRS-4	2216	LEO (ISS)	NASA (CRS)	Success	Uncontrolled (ocean)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)

There is screenshot of the output about rank landing outcomes between 2010-06-04 and 2017-03-20

Section 4

Launch Sites Proximities Analysis

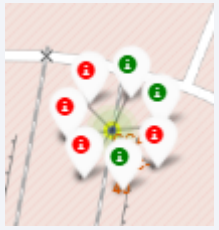
Launch sites on a map



Conclusions:

1. All launch sites is situated in the USA far from the equator line. It can be assumed, that saving finances on logistics is a more efficient solution, than saving on launches from Equator line
2. All launch sites is situated very close proximity to the coast

The success/failed launches



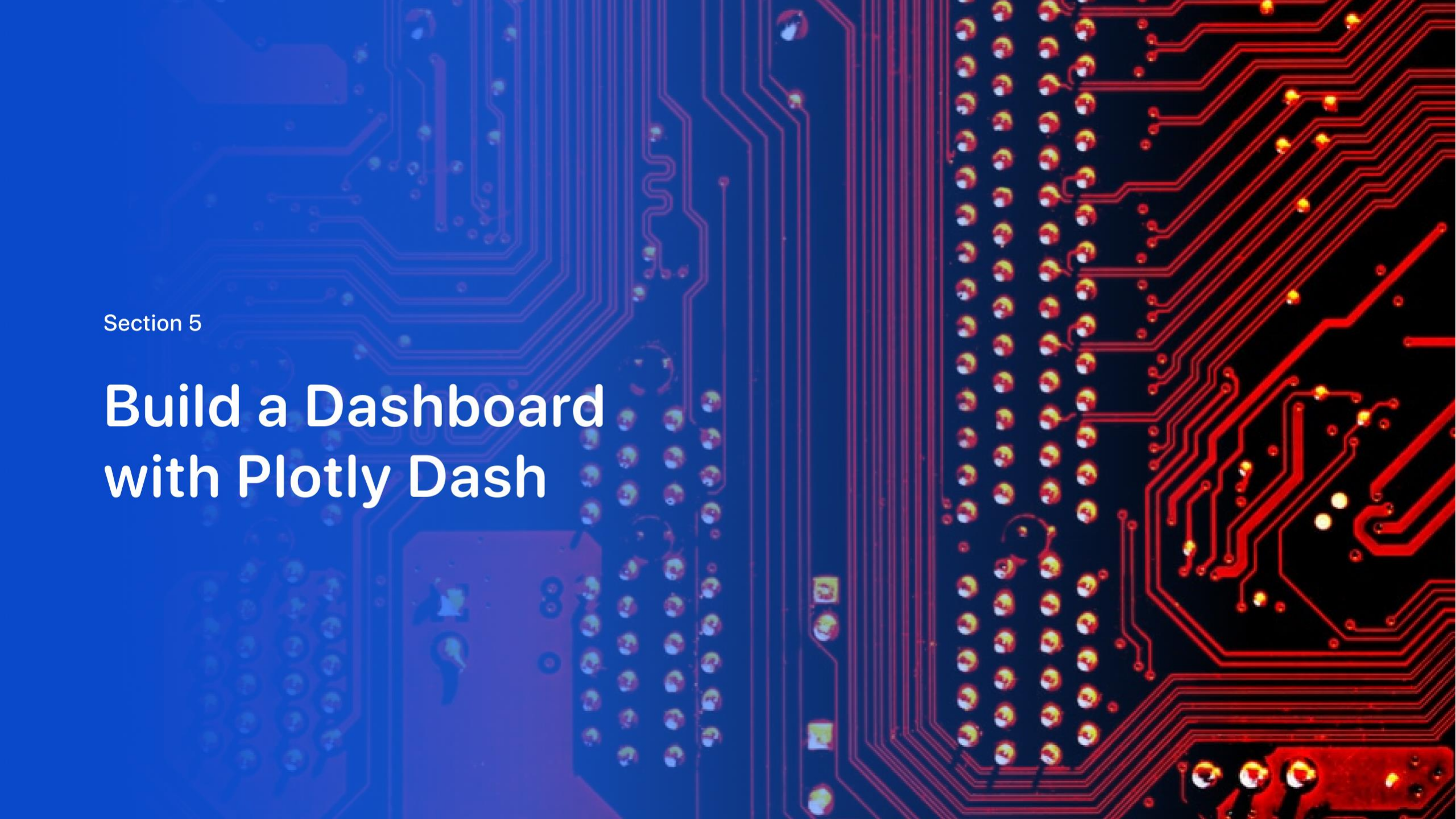
The most popular launch site is the CCAFS LC-40 with 26 launches. 46 rockets were launched on the east coast of the United States and on the west 10

The distances between a launch site to its proximities

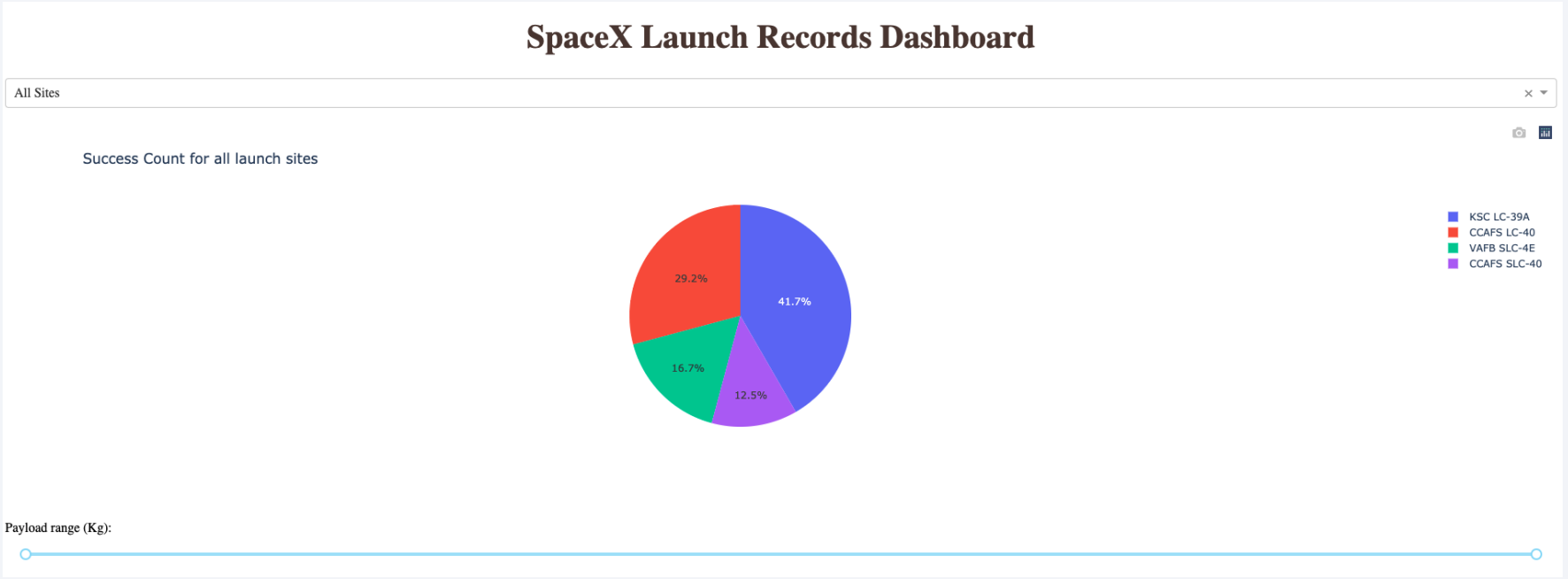
- It was revealed that After analyzing the location of the CCAFS SLC-40, CCAFS LC-40, KSC LC-39A in close proximity to railways, highways, coastline and from cities.
- VAFB SLC 4E far from away city (13.76 km from Lompoc)

Section 5

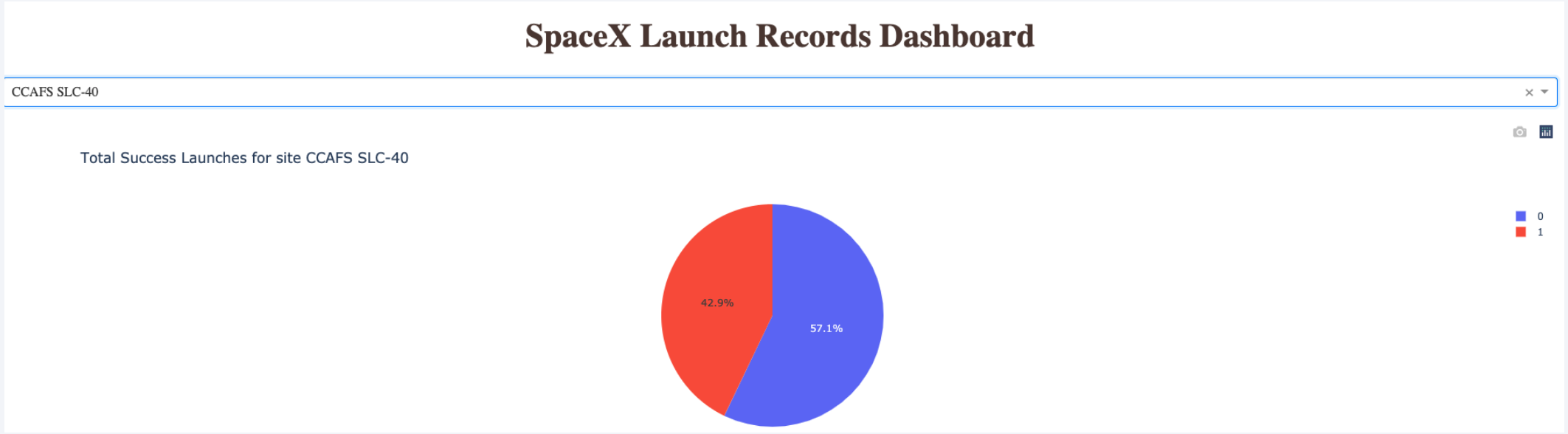
Build a Dashboard with Plotly Dash



Pie chart count of successful launch sites

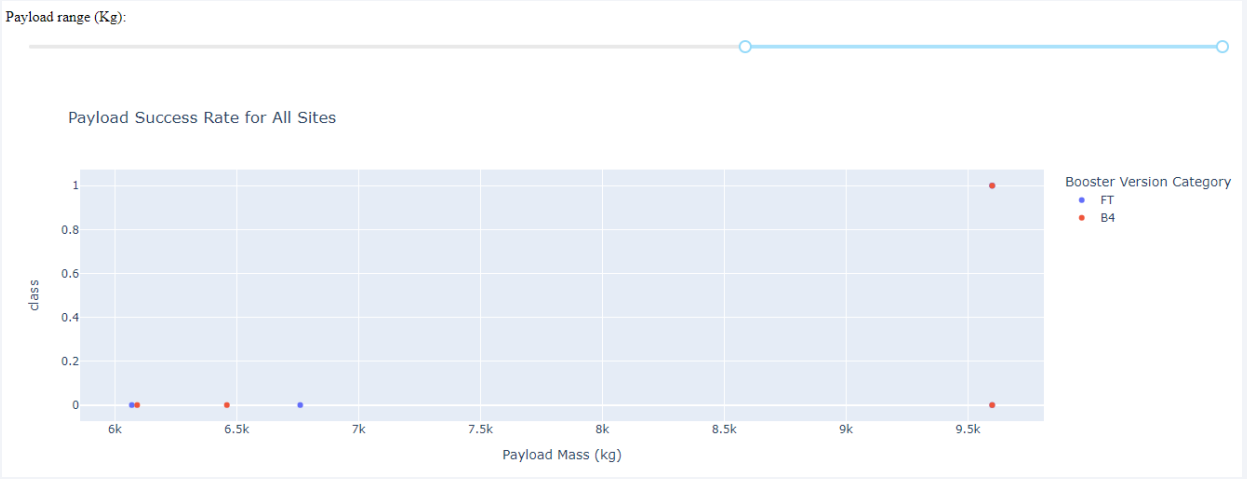


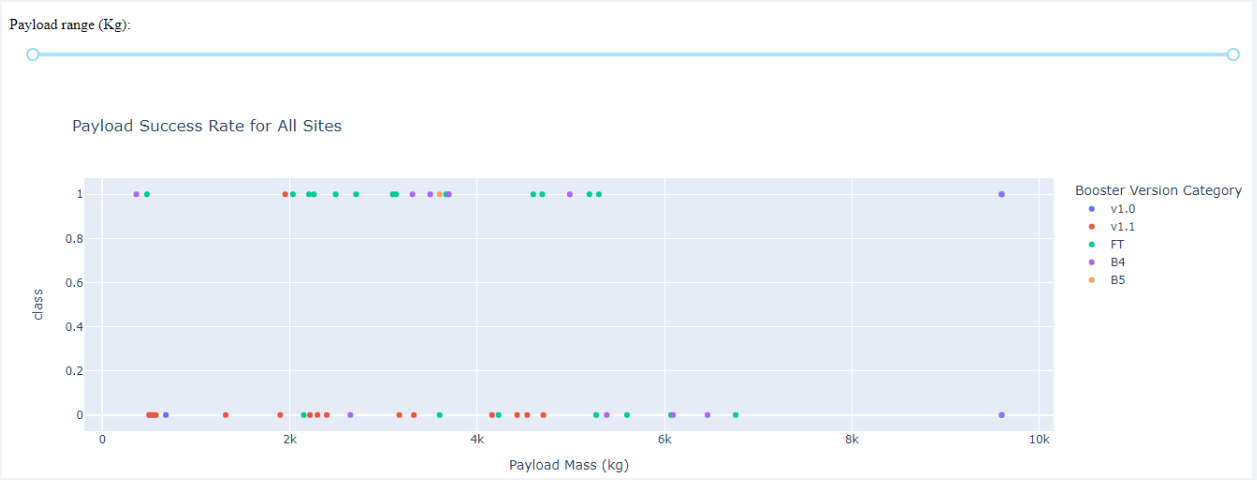
Interactive pie chart



Payload Success Rate for All Sites

Booster FT, B4, B5 have the highest success rates. At the same time, B4 is used for heavy loads.



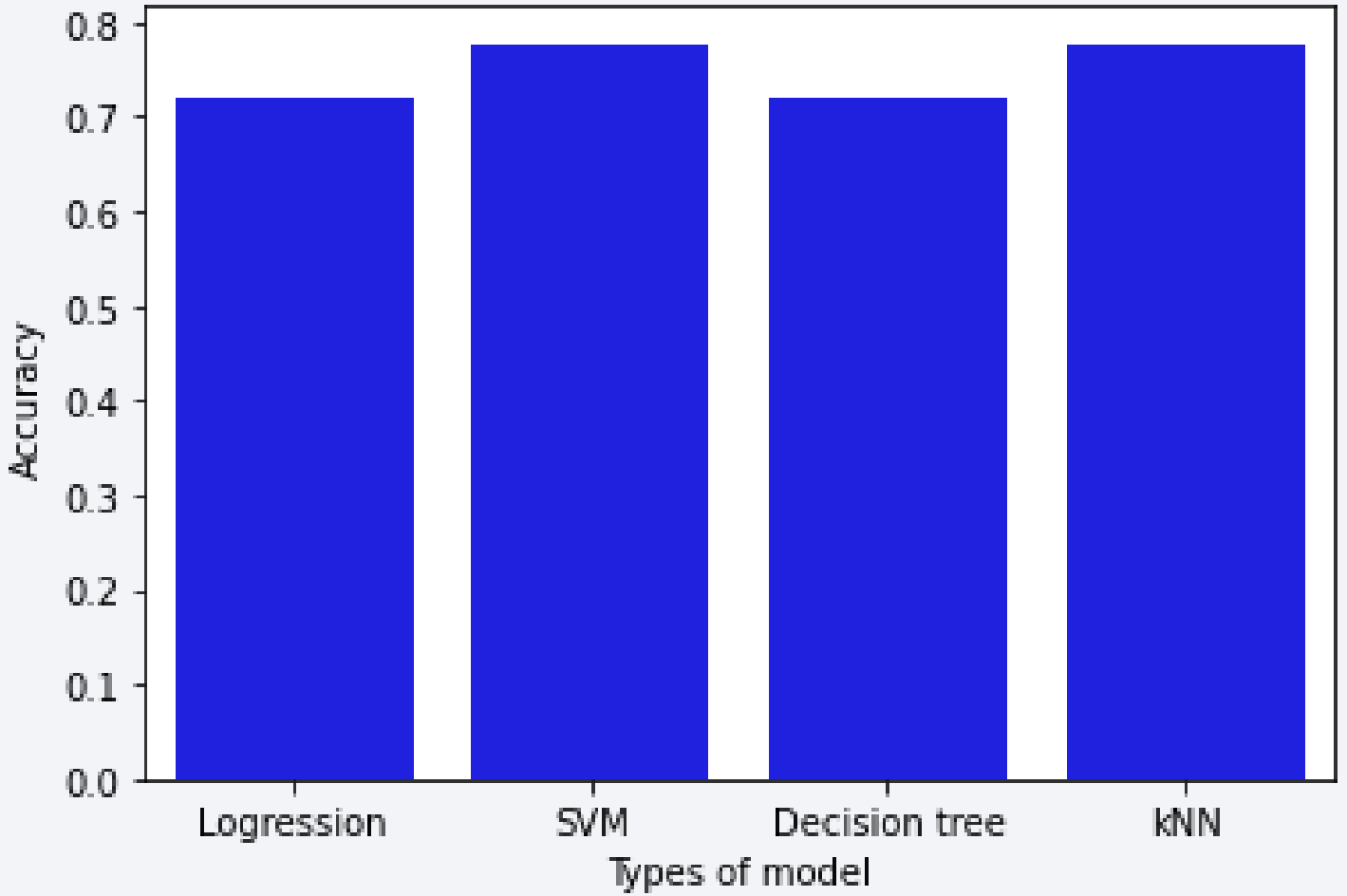


Section 6

Predictive Analysis (Classification)

Classification Accuracy

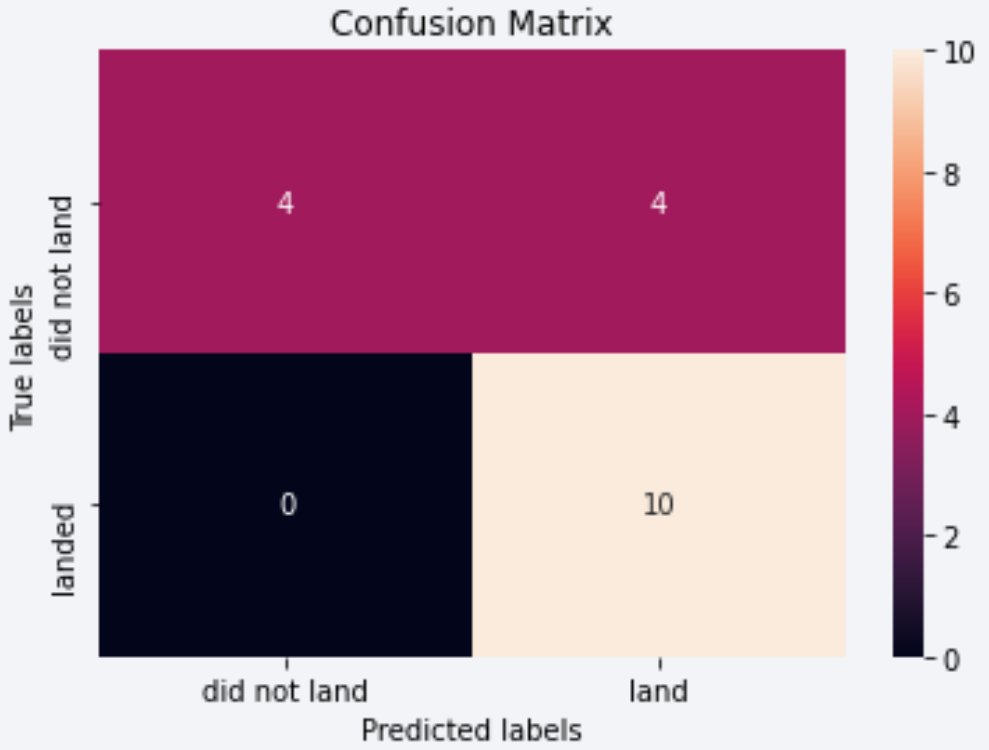
The SVM and k-nearest are equal. The accuracy of models are 0.77.



Confusion Matrix

kNN

SVM



The confusion matrix are equal for kNN and SVM. We see that the major problem is false positives.

Conclusions

- The main data sources were open sources and official data.
- Analysis of the launch infrastructure showed that all launch sites are located directly in proximity to the coast and have a well-developed infrastructure. Point 3
- The constructed predictive models kNN and SVM have a high Accuracy level. However, the main problem is is false positives results

Appendix

- Github repository: https://github.com/KainaraDm/Applied-Data-Science-Capstone

Thank you!