IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

KONDRATENKO Kirill

20/02/2022

Outline

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

Executive Summary

- The data was collected from 2 sources (SpaceX API and web scraping) which further was cleaned and reshaped in order to be fed into the prediction models. Explorative analysis was performed with visualization tools and SQL queries to reveal important correlations and find the most important features for prediction models. Study with Folium has allowed to better understand the important factors related to the launch sites. Finally, interactive dashboard was constructed to assist in the launch site success rate study as well as to observe the influence of payload on the launch outcome.
- 4 prediction models were built and optimized, however it was impossible to find the best performing one due to the insufficient data set size. The accuracy of 0,83 was achieved.

Introduction

- SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- The main goals of this study include understanding the factors which influence the success rate of Falcon first stage landing for the purpose of building and optimizing a model which predicts successful landing.

Section 1

Methodology

Methodology

Executive Summary

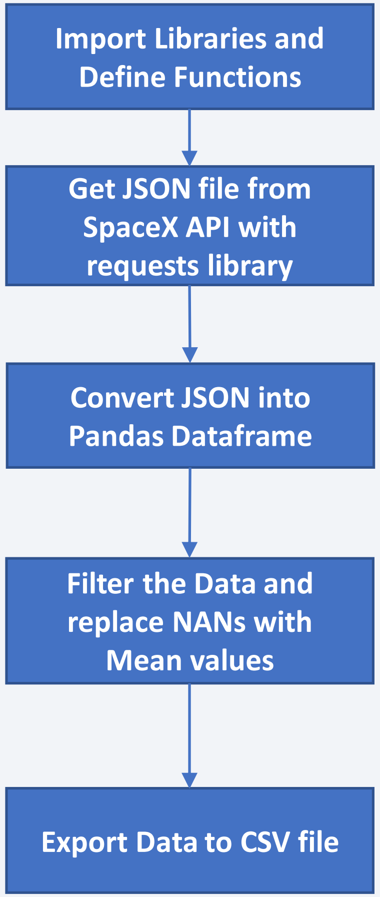
- Data collection methodology:
  - Data collected from 2 sources (SpaceX API and Wikipedia page)
- Data wrangling
  - NaNs replaced with mean values, training labels are determined
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - 4 models were trained and their hyper parameters were tuned

Data Collection

- Data was collected by utilizing 2 approaches:
    1. Request to the SpaceX API
    2. Web Scraping from a Wikipedia page
- The data collection process consists of the following steps:
    - Extract the data in the raw form (as JSON file or HTML)
    - Convert the data to a pandas dataframe format
    - Remove and/or replace NaNs

Data Collection – SpaceX API

- SpaceX Launch data was obtained through a request from SpaceX API
- The data was inserted from JSON file to a Pandas data frame, respective IDs were matched with the actual data items
- Next, the data was filtered to include only Falcon 9 launches and numeric NaN values were replaced by mean. Finally, the data was exported as a CSV file.
- GitHub URL: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-/blob/master/Data%20Collection%20API.ipynb


Import Libraries and Define Functions

Get JSON file from SpaceX API with requests library

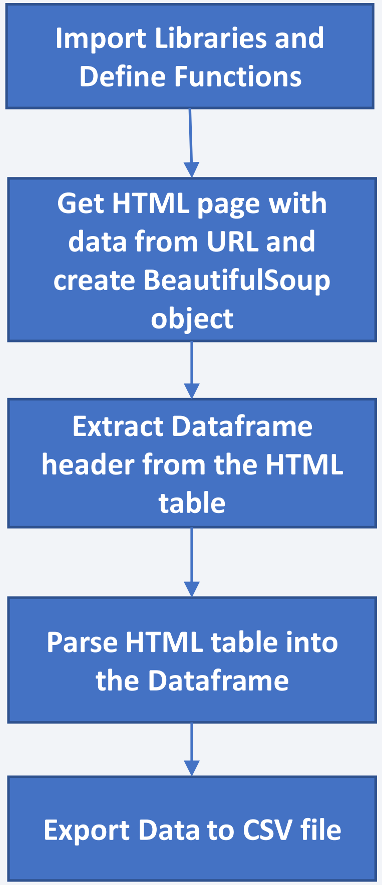
Convert JSON into Pandas Dataframe

Filter the Data and replace NANs with Mean values

Export Data to CSV file

Data Collection - Scraping

- Falcon 9 launch records were extracted from a web page on Wikipedia
- The HTML page was requested and its contents were converted to a BeatifulSoup object
- The dataframe column names were extracted from the HTML table, then the data was parsed into a dictionary (keys correspond to columns). Finally, the dictionary was transformed into a data frame and exported as a CSV file.
- GitHub URL: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



Import Libraries and Define Functions
Get HTML page with data from URL and create BeautifulSoup object
Extract Dataframe header from the HTML table
Parse HTML table into the Dataframe
Export Data to CSV file

Data Wrangling

- The number of launches at each site was calculated, as well as the number of launches per orbit and per outcome.
- The launch data was labelled according to the outcome (0 for negative outcome and 1 for positive).
- The success rate was determined (0.66)
- GitHub URL: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- **Outcome** per **Launch Site** as a function of **Flight Number** was plotted to see if there is a link between the success rate and the launch location.
- **Outcome** per **Launch Site** as a function of **Payload** was plotted to see the correlation between the payload mass and launch location, as well as relationship between the payload and successful outcome.
- **Success Rate** per **Orbit** was plotted to find which orbits have high success rate
- **Outcome** per **Orbit** as a function of **Flight Number** and **Payload** was plotted to explore the possible link between these parameters and successful outcome.
- Finally, yearly **Success** trend was plotted to illustrate the progress.
- GitHub URL: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-/blob/master/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

The following SQL queries were performed:

- Display the names of unique launch sites
- Display 5 records where launch sites begin with 'CCA‘
- Display the total payload mass launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- Display the date of the first successful landing on the ground pad
- Display the names of the boosters successfully landed in drone ship with the payload mass between 4000 and 6000 kg
- Display the total number of successful and failure mission outcomes
- Display the names of the booster versions which have carried the maximum payload mass
- Display the failed outcomes in drone ship, their booster versions, and launch site names in 2015
- Rank the count of landing outcomes between the 2010-06-04 and 2017-03-20 in descending order
- GitHub URL: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

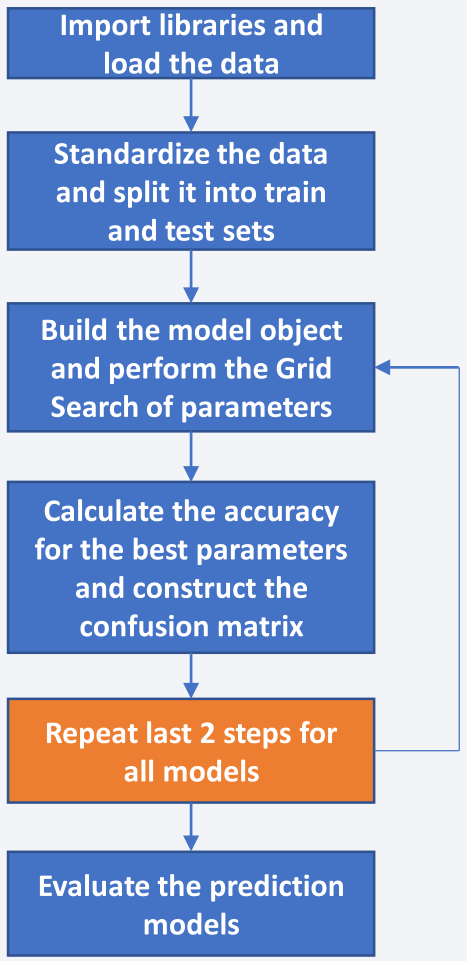
- Launch sites and their names were marked on the map, circular areas were added to highlight the launch site.
- Marker clusters were created with success/fail outcomes for each site.
- Distance was calculated and lines were traced from the launch site to the closest railroad, highway and city to find insight on the launch site location.
- GitHub URL: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-/blob/master/Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Pie chart was constructed to visualize the success rate according to the launch site with a drop-down menu to select the launch site of interest (for both pie and scatter plot).
- Scatter plot of launch outcome as a function of payload was added as well. An interactive slider was added to select the payload range of interest.
- These plots allow to visualize the most relevant information regarding the launch outcome for each launch site.
- GitHub URL: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- The data was standardized and split into training and test sets;
- The model object was created alongside with the hyper parameter dictionary; Grid search was performed to find best model hyper parameters;
- The accuracy was calculated and the confusion matrix was built to evaluate the model;
- The following models were tested: logistic regression, SVM, decision tree classifier and k nearest neighbors
- GitHub URL: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-/blob/master/Machine%20Learning%20Prediction%20.ipynb



Import libraries and load the data

Standardize the data and split it into train and test sets

Build the model object and perform the Grid Search of parameters

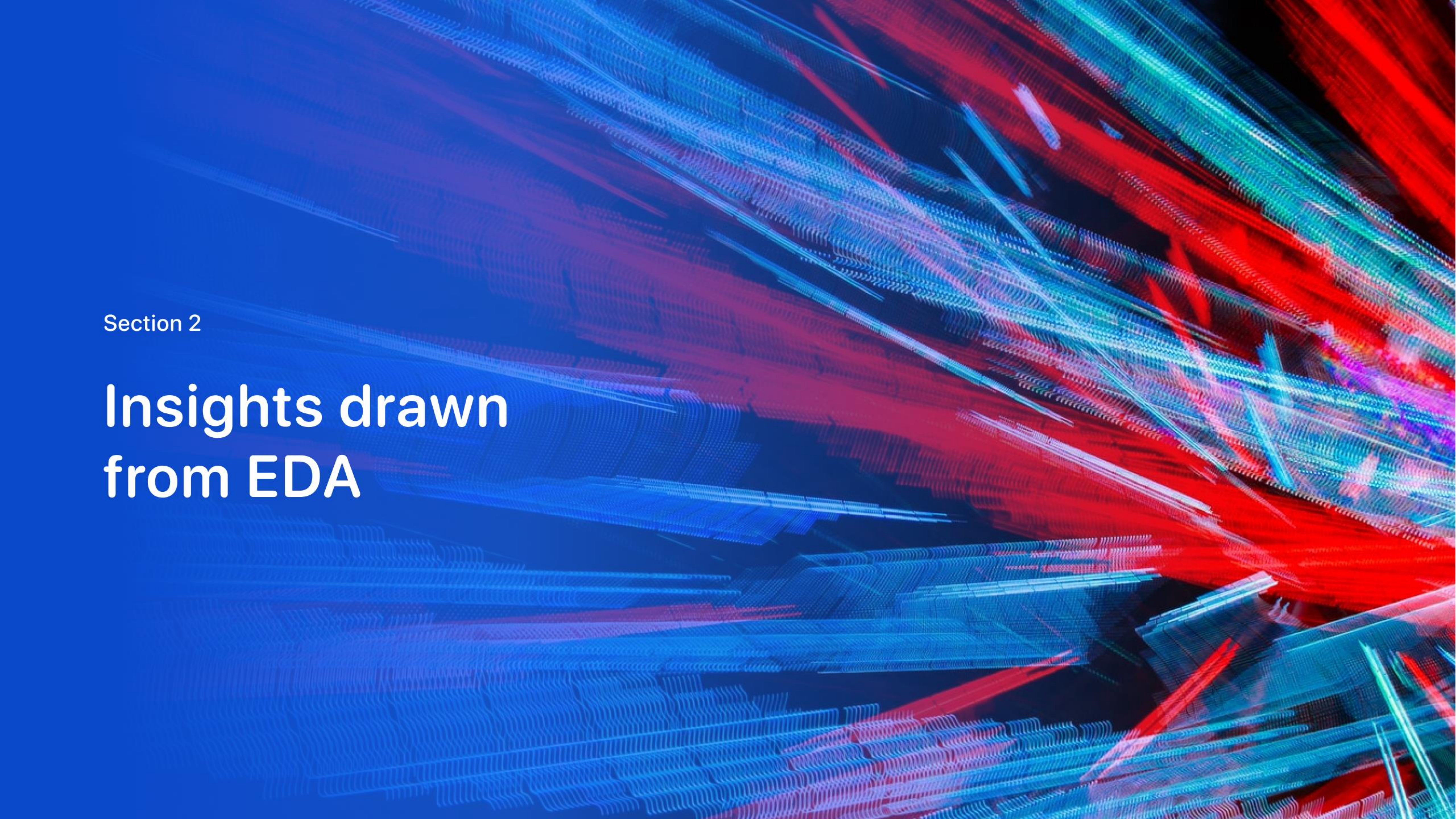
Calculate the accuracy for the best parameters and construct the confusion matrix

Repeat last 2 steps for all models

Evaluate the prediction models

Results

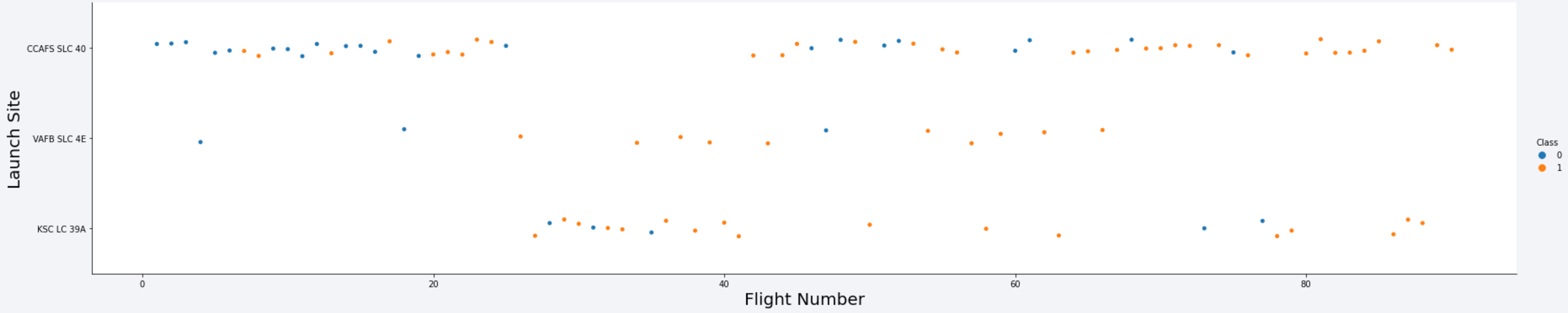
- Exploratory data analysis results
  - Launch success is correlated with Flight Number, Payload Mass, Orbit, Launch Site, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial
- Interactive analytics demo in screenshots
  - Allows a convenient way to correlate launch site and payload to success rate (screenshots are presented in the section 5)
- Predictive analysis results:
  - The following models were tested: logistic regression, SVM, decision tree classifier and k nearest neighbors
  - The accuracy of these models was similar (0,83)



Section 2

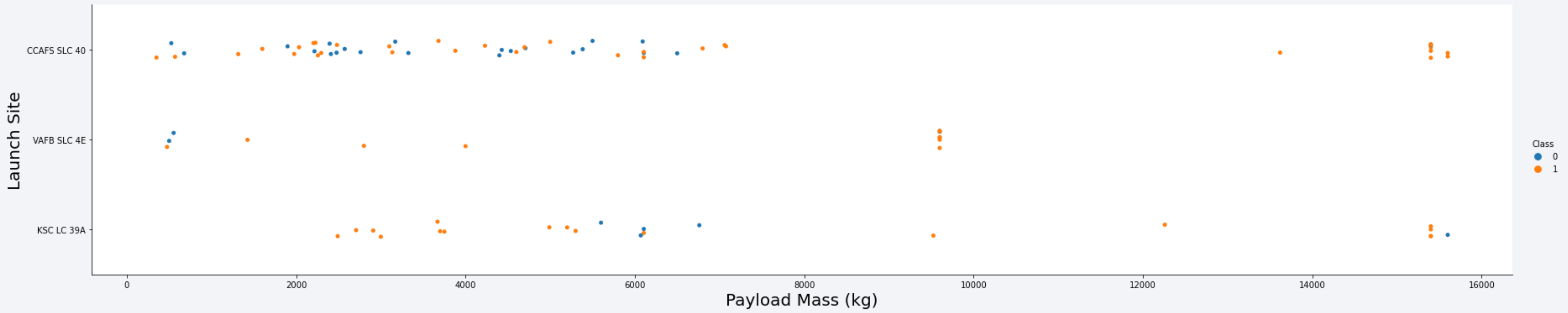
Insights drawn from EDA

Flight Number vs. Launch Site



From this plot, it seems that for 2 launch sites (KSC LC-39A and VAFB SLC-4E) the majority of the outcomes are positive (landing of the 1st stage) regardless of the fligth number. For the third launch site (CCAFS SLC-40), early attempts were mostly unsuccessful, however the success rate has improved later on.

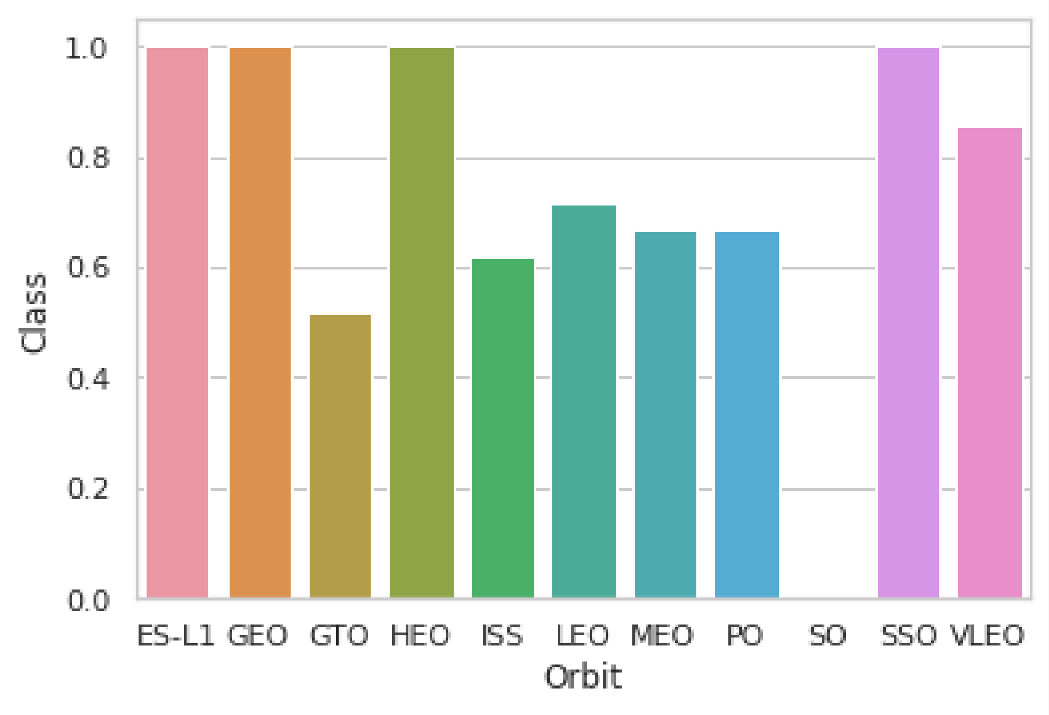
Payload vs. Launch Site



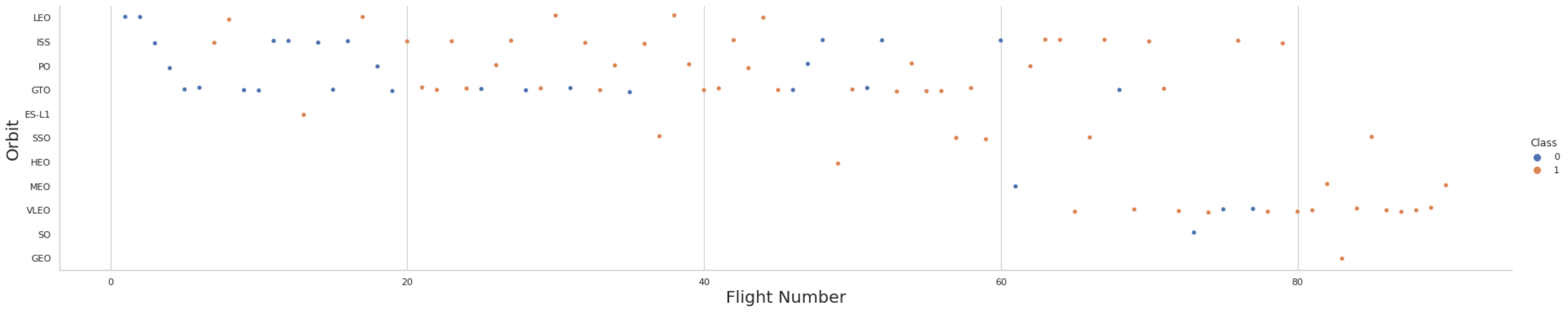
For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000 kg).

Success Rate vs. Orbit Type

- The following orbits have the highest success rate (100%): ES-L1, GEO, HEO and SSO.

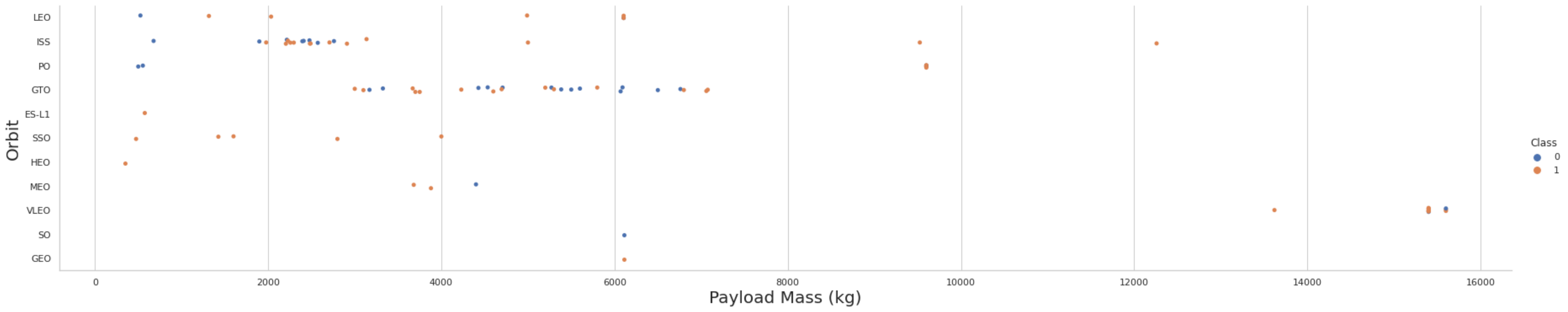


Flight Number vs. Orbit Type



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

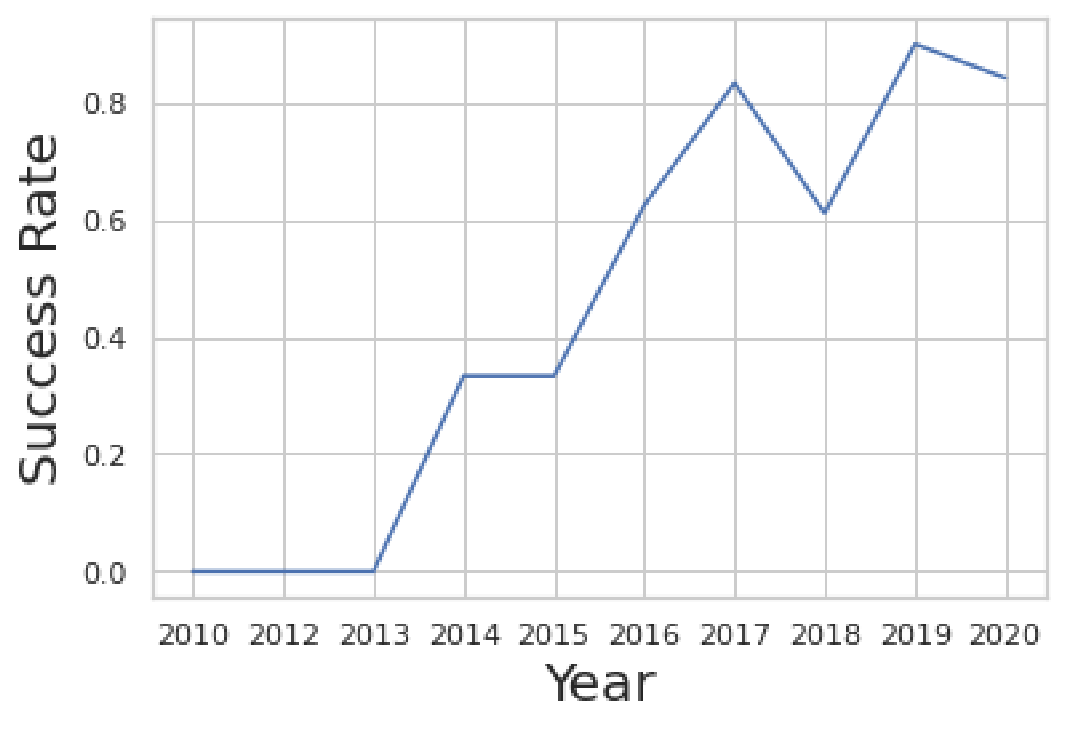
Payload vs. Orbit Type



- Successful landing or positive landing rate with heavy payloads are higher for Polar, LEO and ISS.
- Similar conclusion cannot be reached for the GTO orbit.

Launch Success Yearly Trend

The success rate kept increasing from 2013 till 2020.

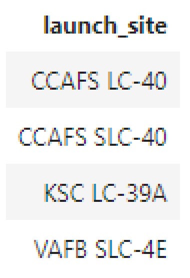


All Launch Site Names

- SQL query “find the names of the unique launch sites”:

```
%sql select distinct LAUNCH_SITE from SWX81833.SPACEXTBL
```

- Query result: there are 4 unique launch sites.



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

Launch Site Names Begin with 'CCA'

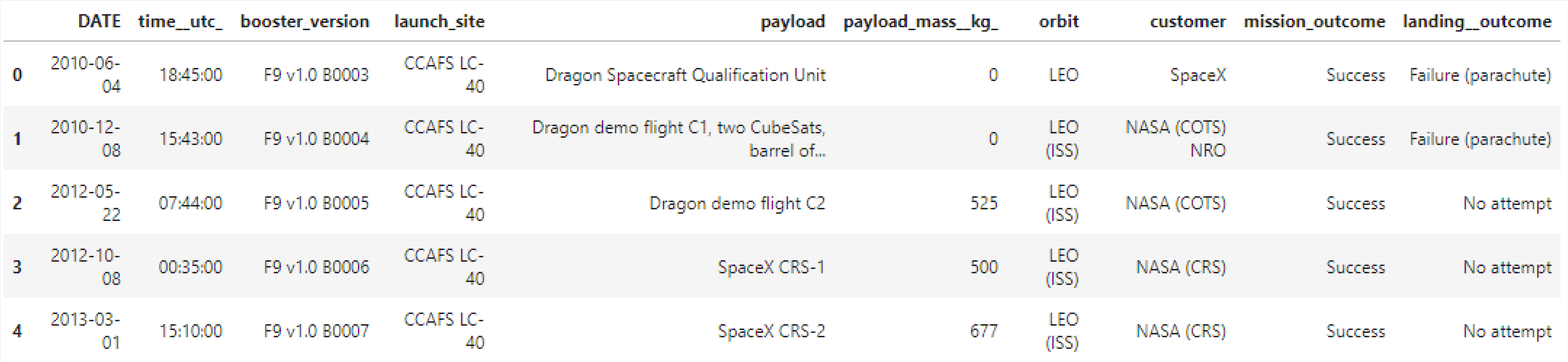
- Find 5 records where launch sites begin with `CCA`

sites = %sql select * from SWX81833.SPACEXTBL where LAUNCH_SITE like 'CCA%'

sites_df = sites.DataFrame()

sites_df.head()

- Query result:



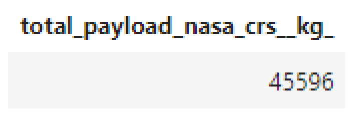
	DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM(CASE WHEN CUSTOMER like 'NASA (CRS)' THEN
payload_mass__kg_ END) as total_payload_NASA_CRS__kg_ from
SWX81833.SPACEXTBL
```

- Query result: total payload for NASA (CRS) is 45596 kg.



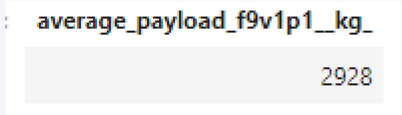
total_payload_nasa_crs__kg_
45596

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(CASE WHEN booster_version like 'F9 v1.1' THEN
payload_mass__kg_ END) as average_payload_F9v1p1__kg_ from
SWX81833.SPACEXTBL
```

- Query result: average payload for F9 v1.1 is 2928 kg.



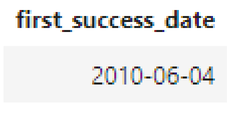
average_payload_f9v1p1__kg_
2928

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql SELECT min(CASE WHEN MISSION_OUTCOME like 'Success%' THEN DATE END) as
first_success_date from SWX81833.SPACEXTBL
```

- Query result: first successful landing happened on 2010-06-04.



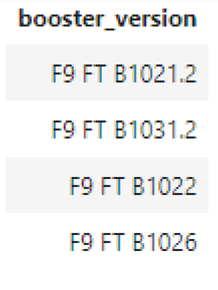
first_success_date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT distinct BOOSTER_VERSION FROM SWX81833.SPACEXTBL WHERE
(PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000 and
LANDING__OUTCOME like 'Success (drone ship)')
```

- Query result:



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

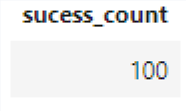
Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

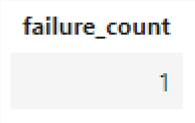
```
%sql SELECT COUNT(*) as Sucess_count from SWX81833.SPACEXTBL WHERE
MISSION_OUTCOME like 'Success%‘
```

```
 %sql SELECT COUNT(*) as Failure_count from SWX81833.SPACEXTBL WHERE
MISSION_OUTCOME like 'Failure%'
```

- Query result:



sucess_count
100



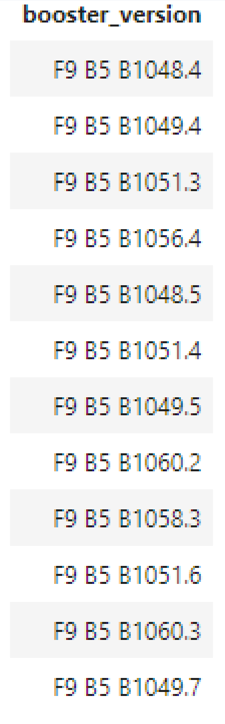
failure_count
1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql SELECT BOOSTER_VERSION FROM
SWX81833.SPACEXTBL WHERE
PAYLOAD_MASS__KG_ = (SELECT
MAX(PAYLOAD_MASS__KG_) FROM
SWX81833.SPACEXTBL)
```

- Query result is presented on the right.



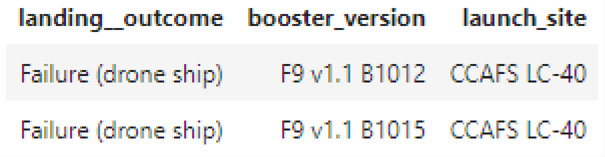
booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

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
SWX81833.SPACEXTBL WHERE (landing__outcome LIKE 'Failure (drone ship)') and
(date LIKE '2015%')
```

- Query result:



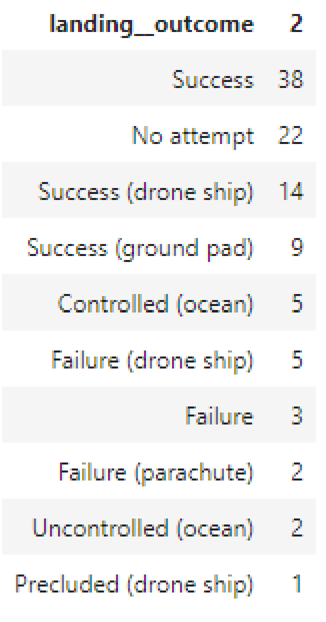
landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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(*) FROM SWX81833.SPACEXTBL GROUP
BY landing__outcome ORDER BY COUNT(*) DESC
```

- Query result:



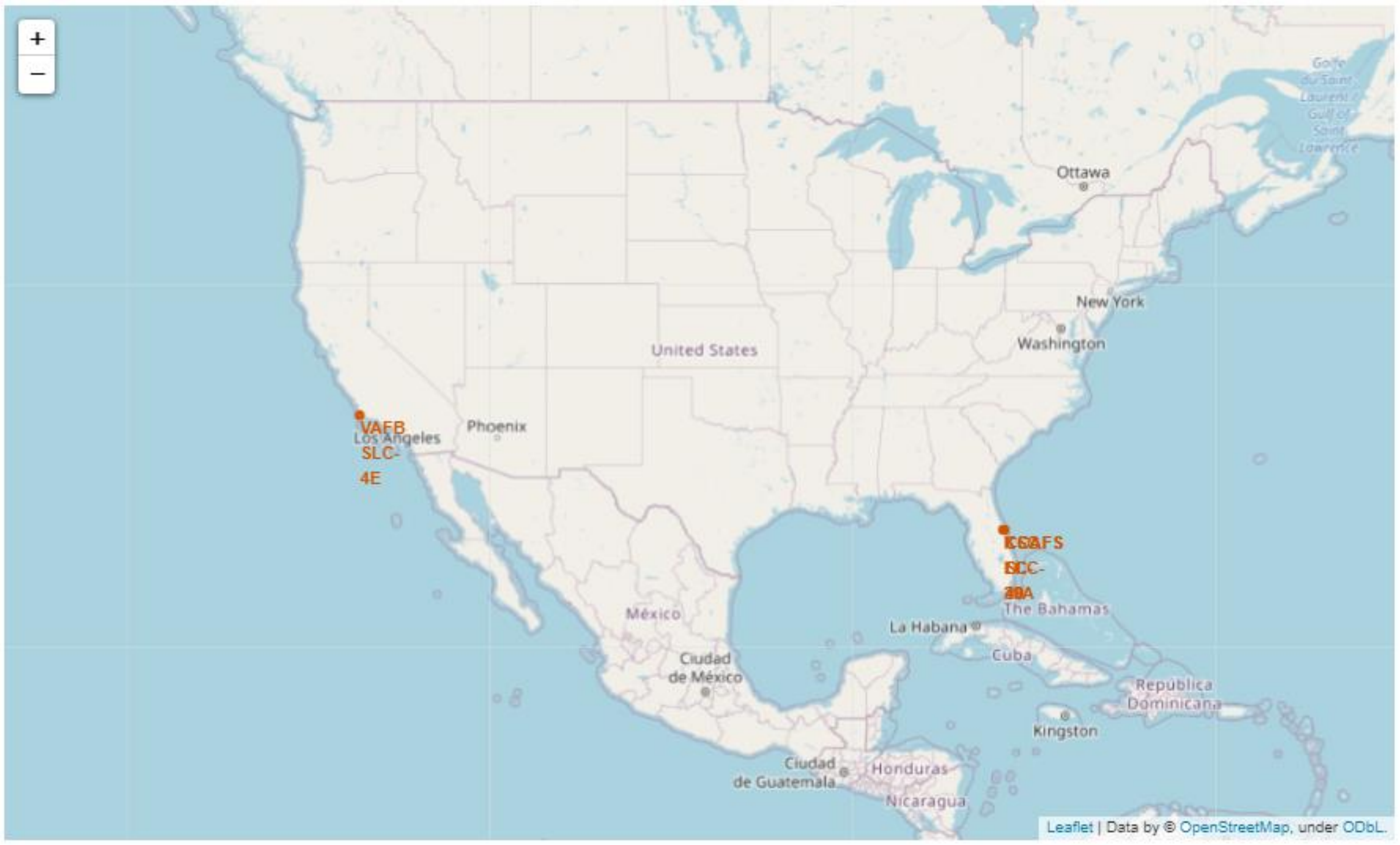
landing__outcome	2
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

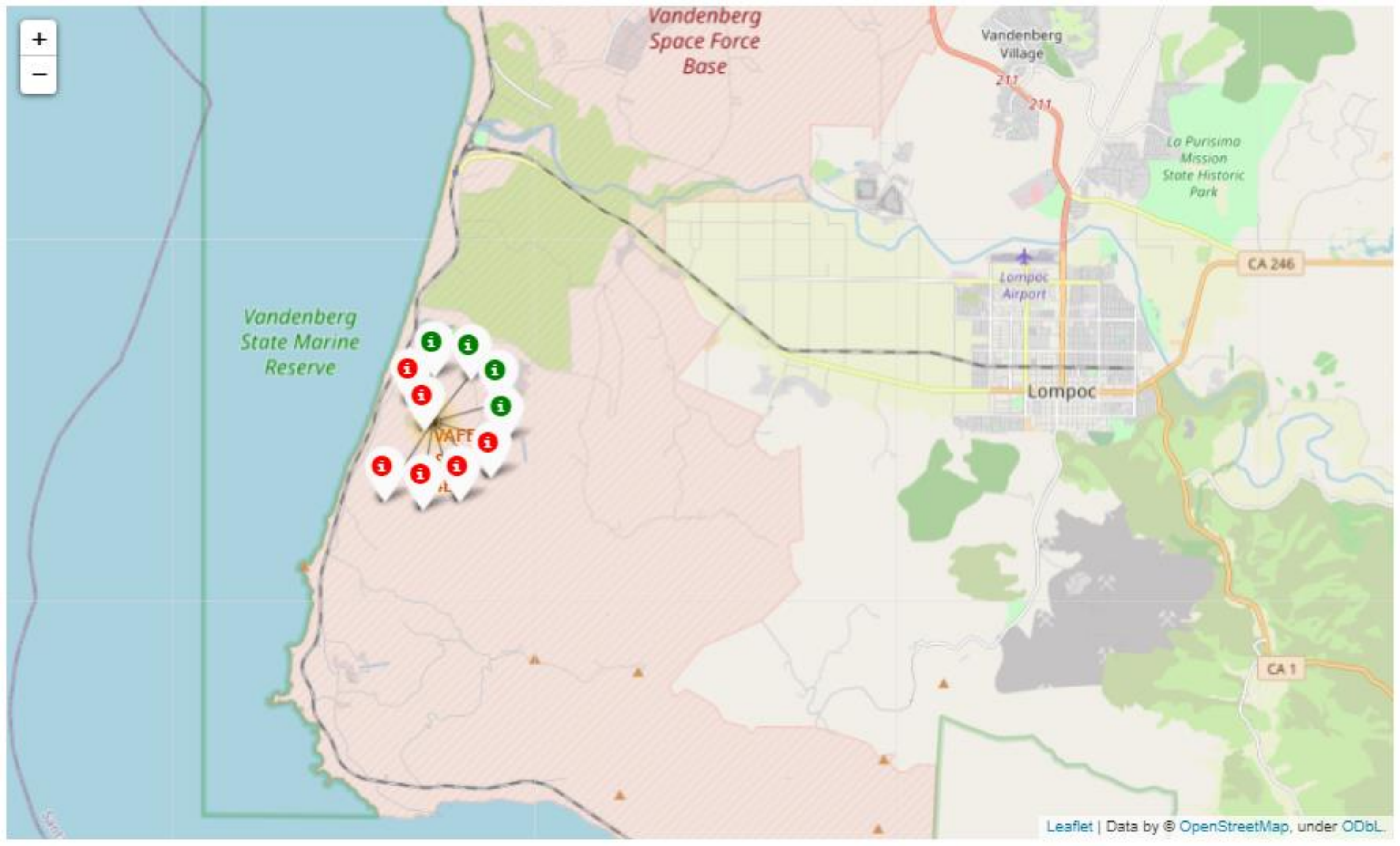
Launch site markers

- Interactive map includes location markers as well as launch site names.



Marker cluster for launch outcomes

- Interactive map demonstrates the color coded markers (green – positive, red – negative) for the launch site VAFB SLC-4E

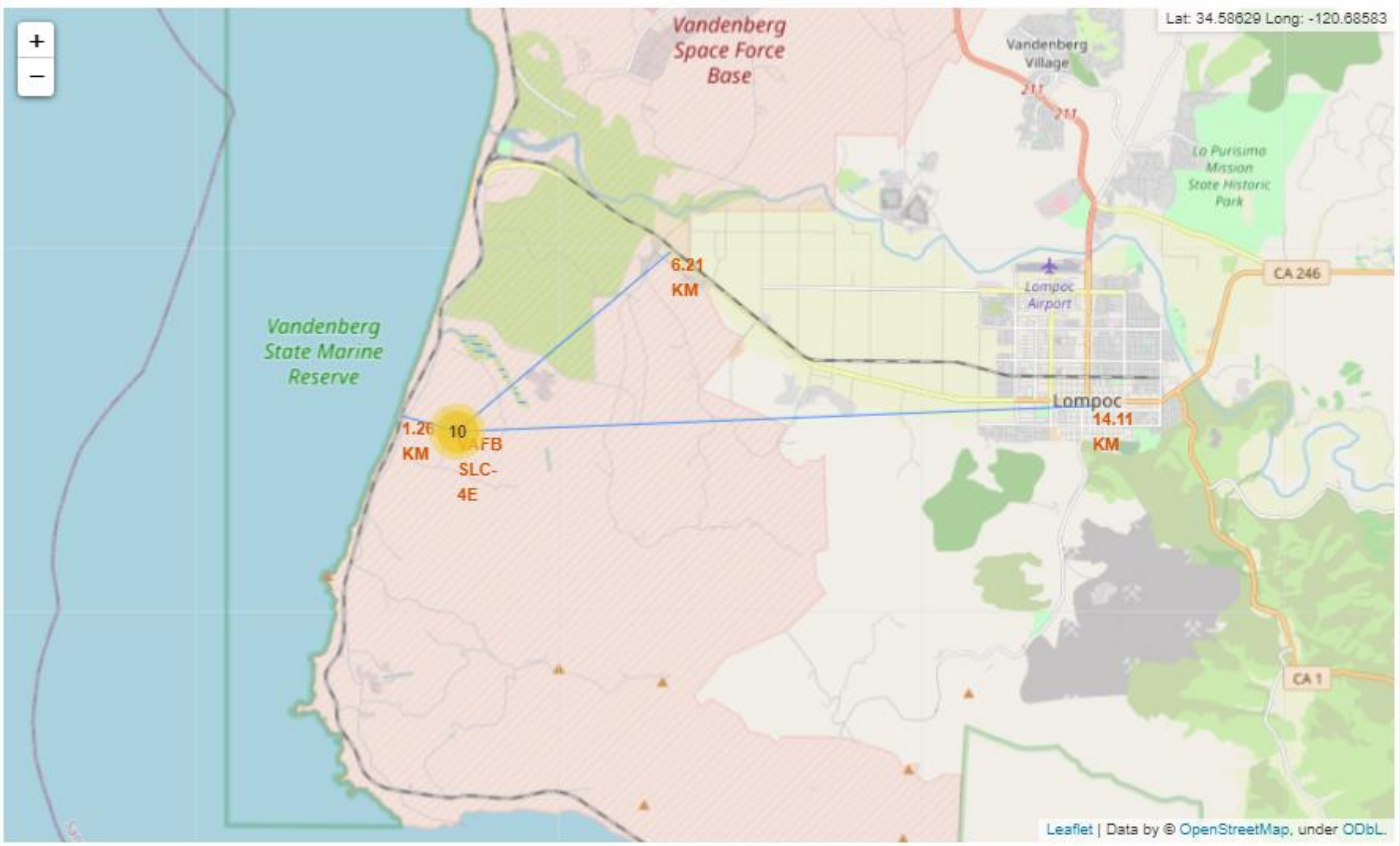


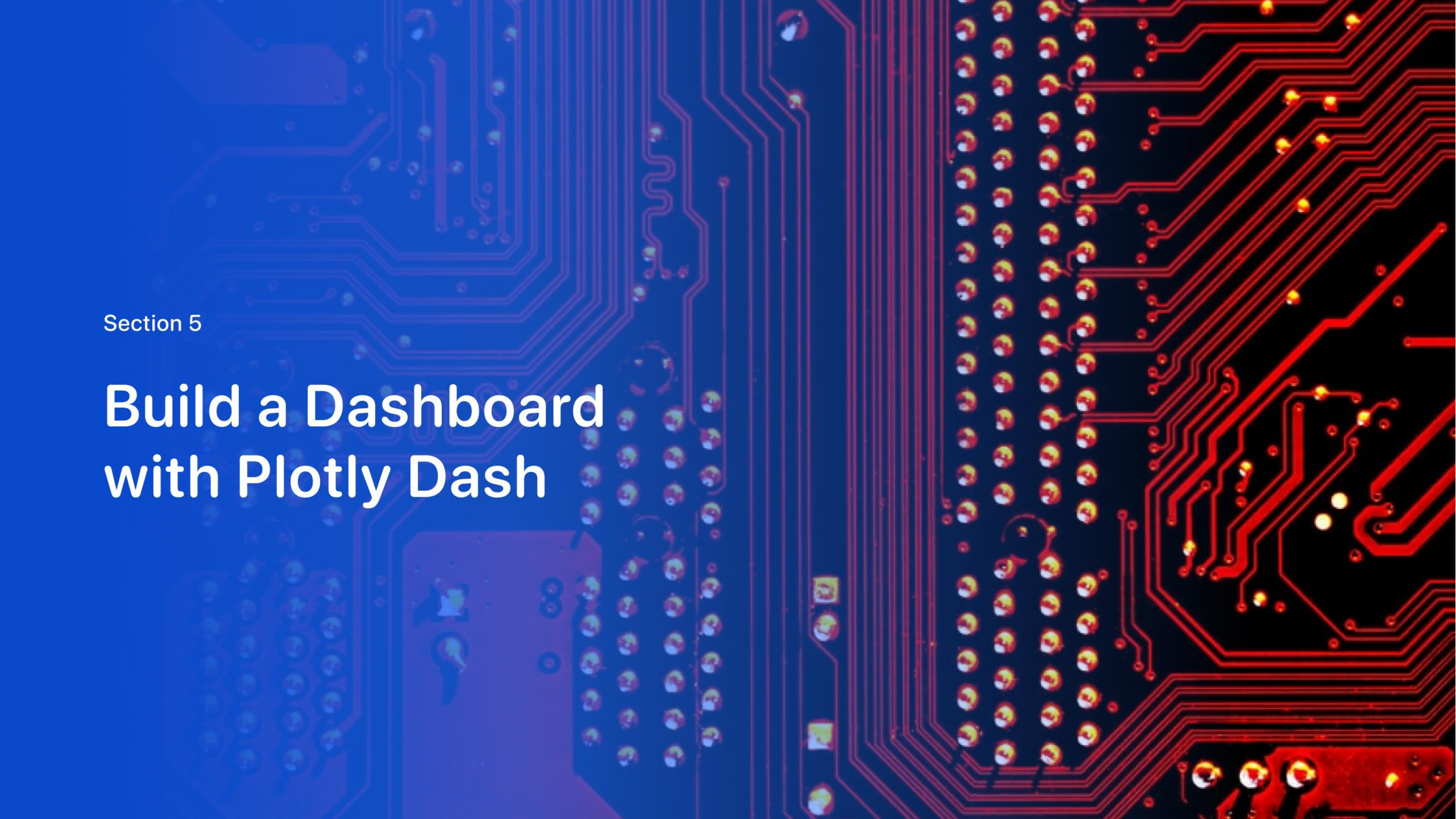
+
−
Vandenberg Space Force Base
Vandenberg Village
211
La Purisima Mission State Historic Park
Lompoc Airport
CA 246
Lompoc
Vandenberg State Marine Reserve
VAFB
CA 1
Leaflet | Data by © OpenStreetMap, under ODbL.

Distance from landmarks to launch site

- Interactive map demonstrates the distances between the launch site VAFB SLC-4E and the surrounding landmarks.

- The launch sites are located close to railroads and highways in order to improve logistics and at a more significant distance from the cities to protect their population.


+
−
Lat: 34.58629 Long: -120.68583
Vandenberg Space Force Base
Vandenberg Village
211
211
La Purisima Mission State Historic Park
6.21 KM
Lompoc Airport
CA 246
Vandenberg State Marine Reserve
1.26 KM
10
VAFB SLC-4E
Lompoc
14.11 KM
CA 1
Leaflet | Data by © OpenStreetMap, under ODbL.

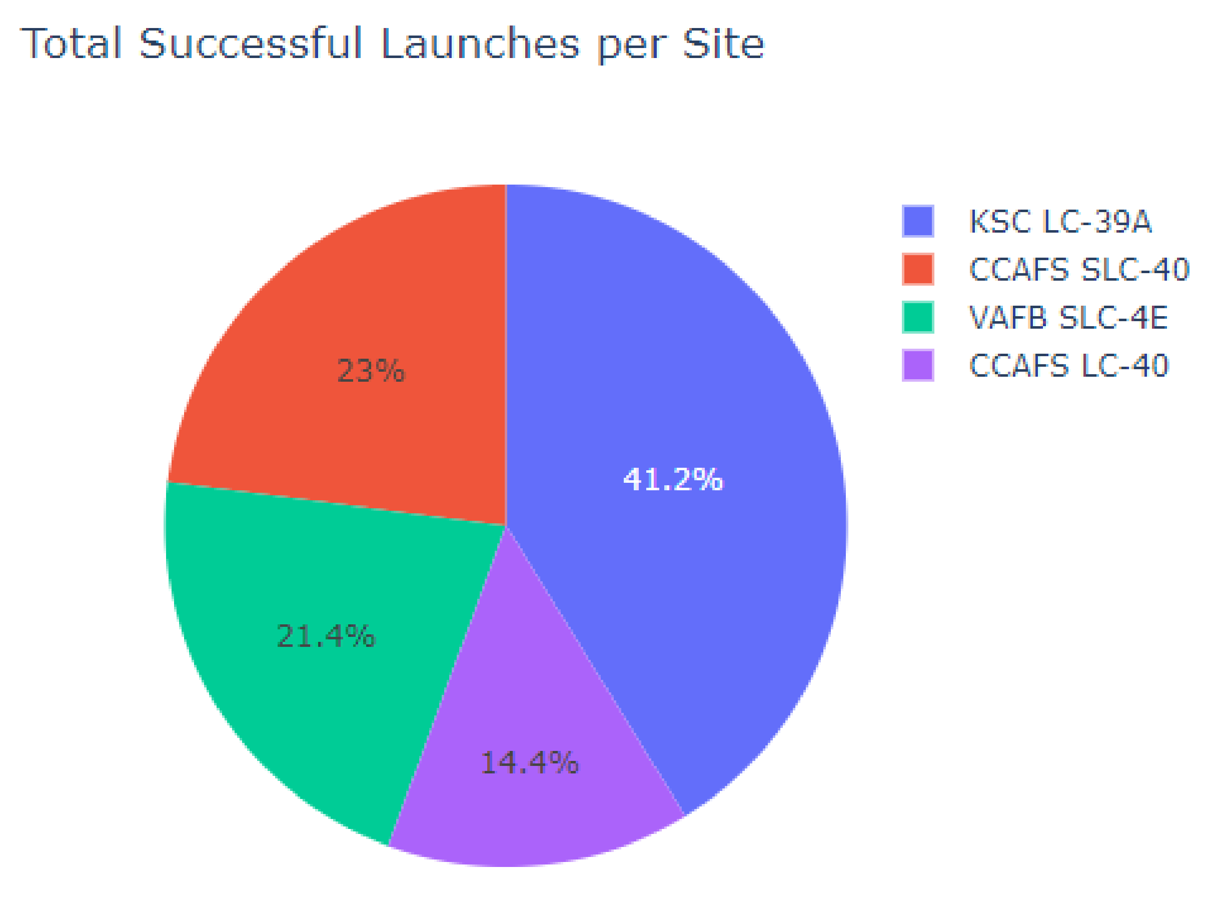


Section 5

Build a Dashboard with Plotly Dash

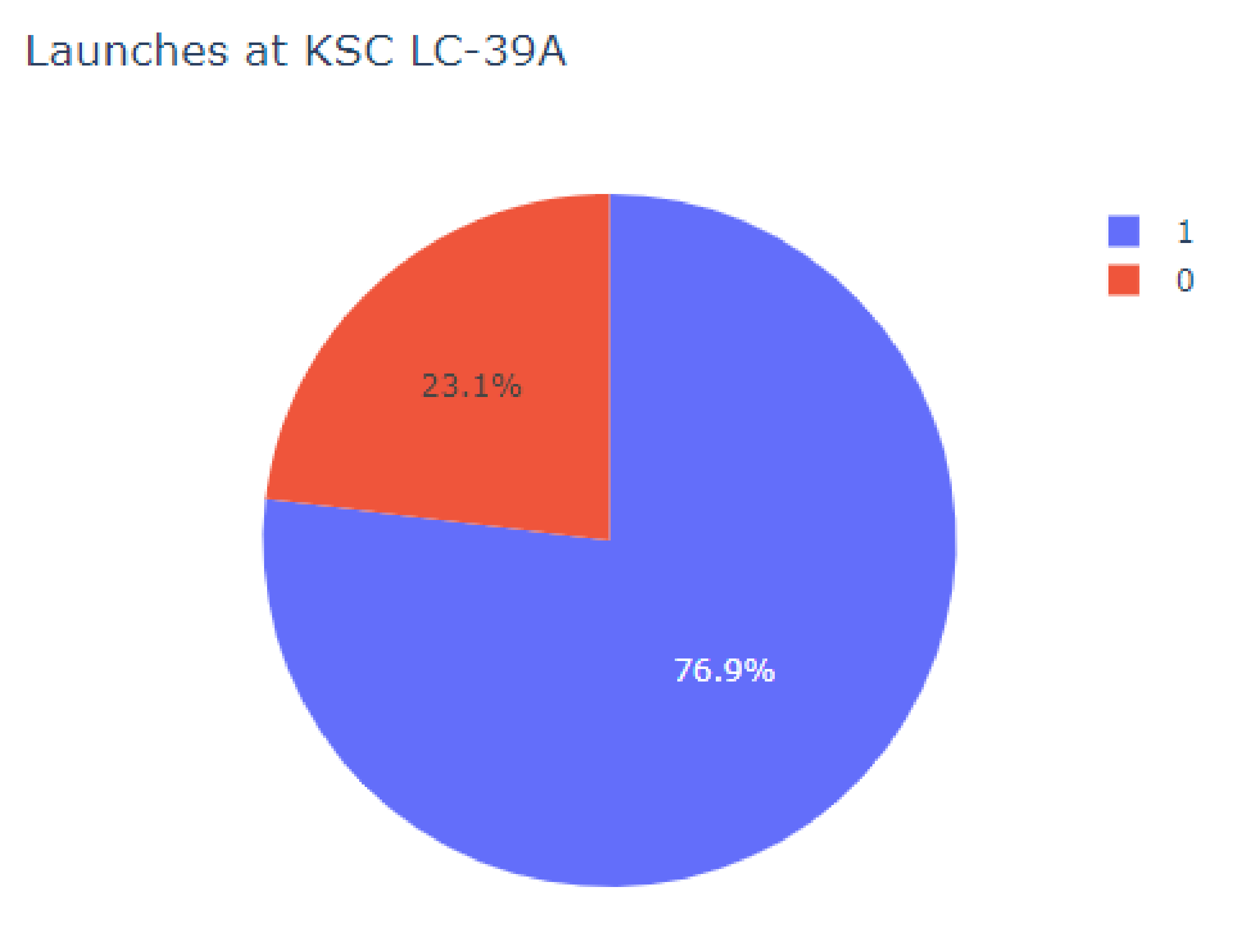
Piechart of success rate for all launch sites

- Out of 4 launch sites, the KSC LC-39A has the highest amount of successful outcomes (41,2%)



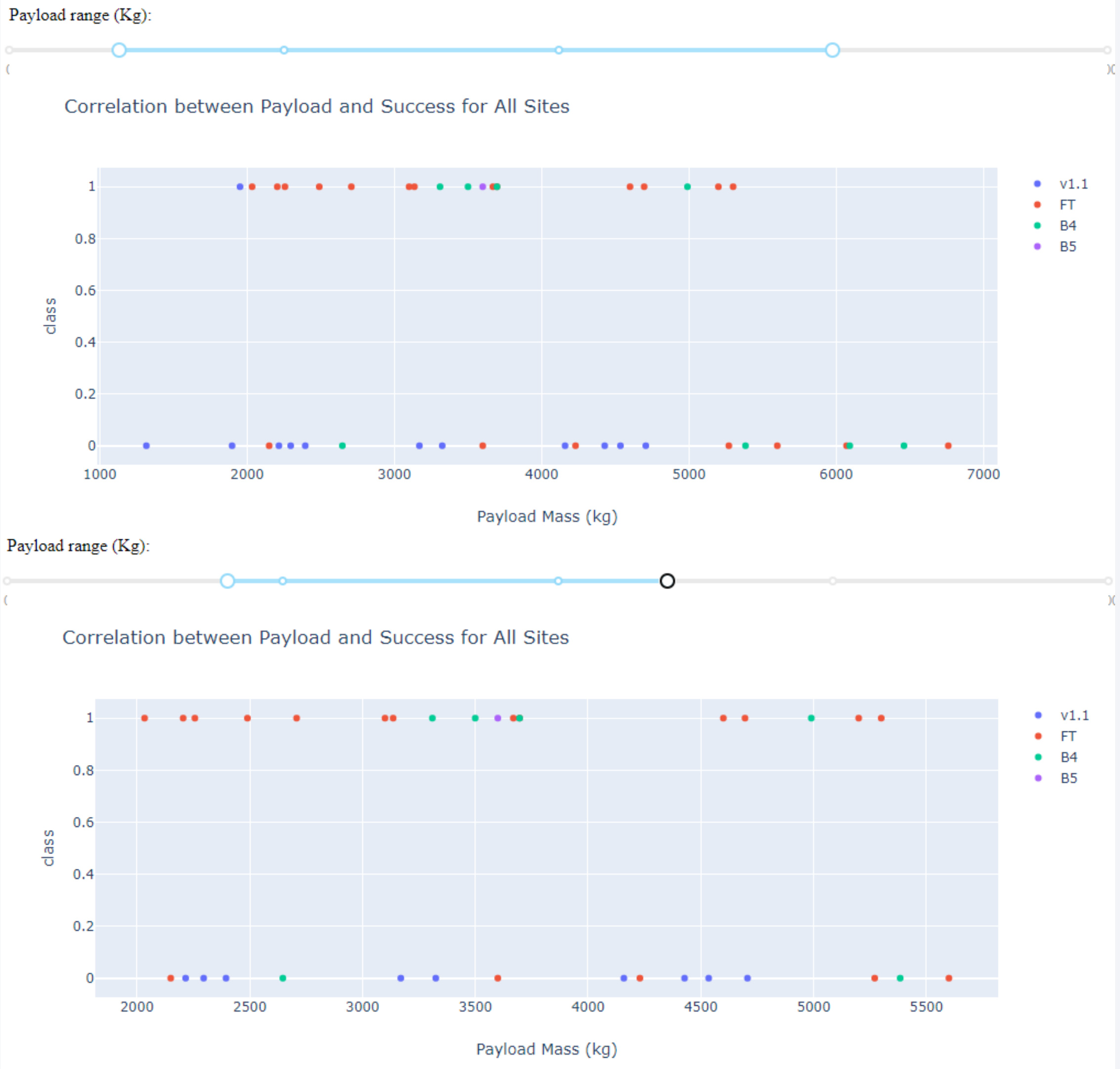
Piechart for the KSC LC-39A launch site

- The KSC LC-39A launch site has the highest success rate (76,9%)



Interactive Payload vs. Launch Outcome scatter plot

- The interactive scatter plot of Payload vs. Launch outcome with adjustable Payload range allows to conclude that the Payload mass ranges between 2000 and 3600 as well as between 4500 and 5400 have highest number of successful outcomes

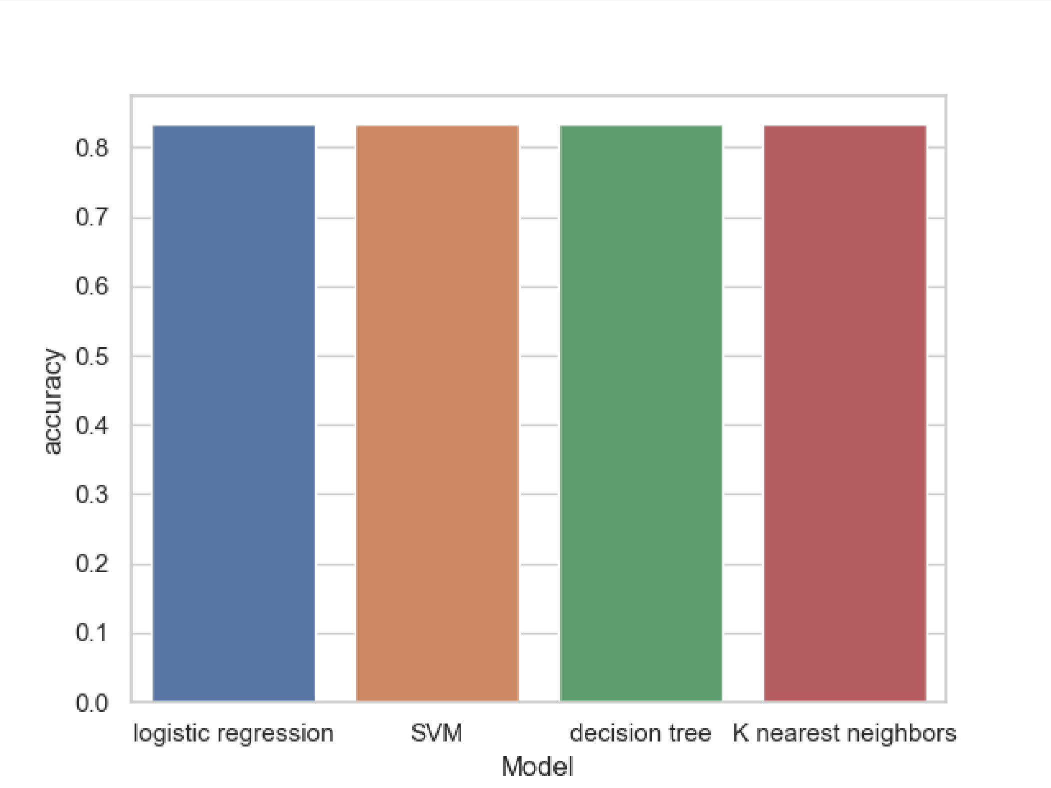


Section 6

Predictive Analysis (Classification)

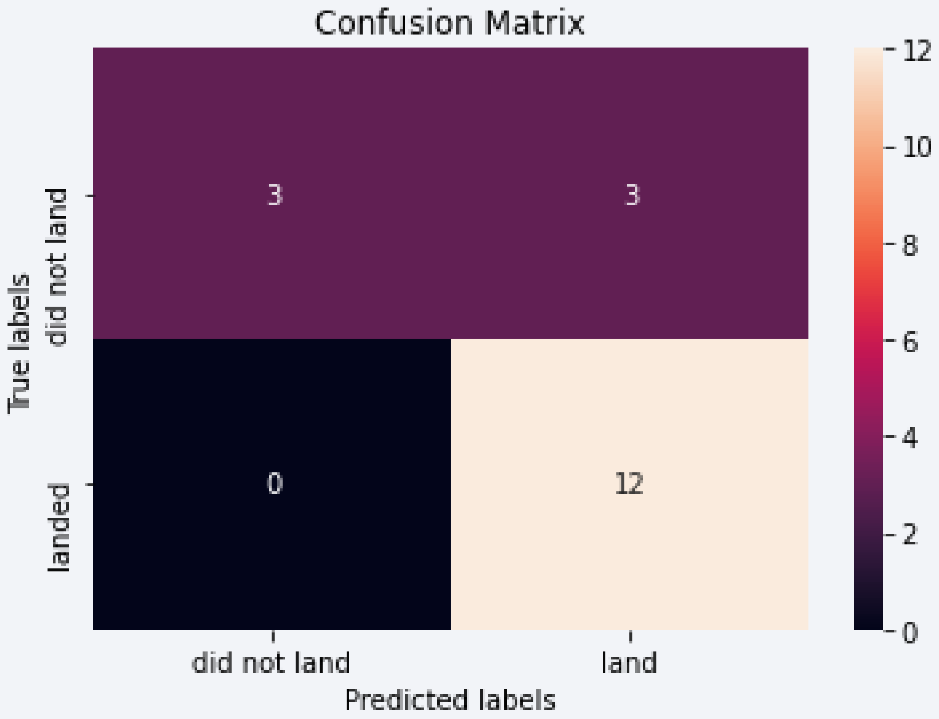
Classification Accuracy

- The model accuracy (4 models presented in the figure) is similar (0,83) due to small data set size (91 flight entry).
- For a larger sample size, the decision tree classifier is expected to perform the best due to the nature of the problem.



Confusion Matrix

- The confusion matrices for the 4 models are identical due to small test size (18 entries).
- All 4 models have problems with false positives



Conclusions

- Exploratory analysis was executed on launch data from 2 sources;
- Important parameters an features correlated to launch success rate were determined;
- Geographical location of launch sites was studied in relation to launch outcome;
- Interactive dashboard was constructed to better understand the underlying relation between launch outcome, launch site and payload
- 4 prediction models for launch outcome classification were built and optimized, accuracy of 0,83 was achieved.

Appendix

- GitHub link to the project repository: https://github.com/kkn1993/Applied-Data-Science-Capstone-IBM-

Thank you!