IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Miguel Zapico
27/06/2022

Outline

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

Executive Summary

- This projects attempts to predict the successful landing of a SpaceX first stage based on data from previous missions
- Data was collected from the SpaceX API and Wikipedia, and cleaned to get a neat workable dataframe
- Exploratory analysis was done to see trends and patterns in the data, both to relate fields to the outcome and to explore geographic issues
- Four classification models were tried and tested
- Three of those models were identified as giving promising results to predict landing success: Logistic regression, Support vector machine and K-nearest neighbourgh
- Estimated accuracy is 0.83

Introduction

- SpaceX provides a satellite launching service cheaper than competitors thanks to the possibility of landing the first stage rocket and reusing it. This competitive advantage depends on the successful landing and recovery of the first stage, what is not a guaranteed event.
- This project attempts to analyze the relevant factors for the landing outcome and build a machine learning model to predict the success of future missions.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Mission data was collected from the SpaceX API and Wikipedia
- Perform data wrangling
  - The data was transformed to obtain a final dataset with neat variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several classification algorithms were used, including KNN, logistic regression, SVN and decision trees

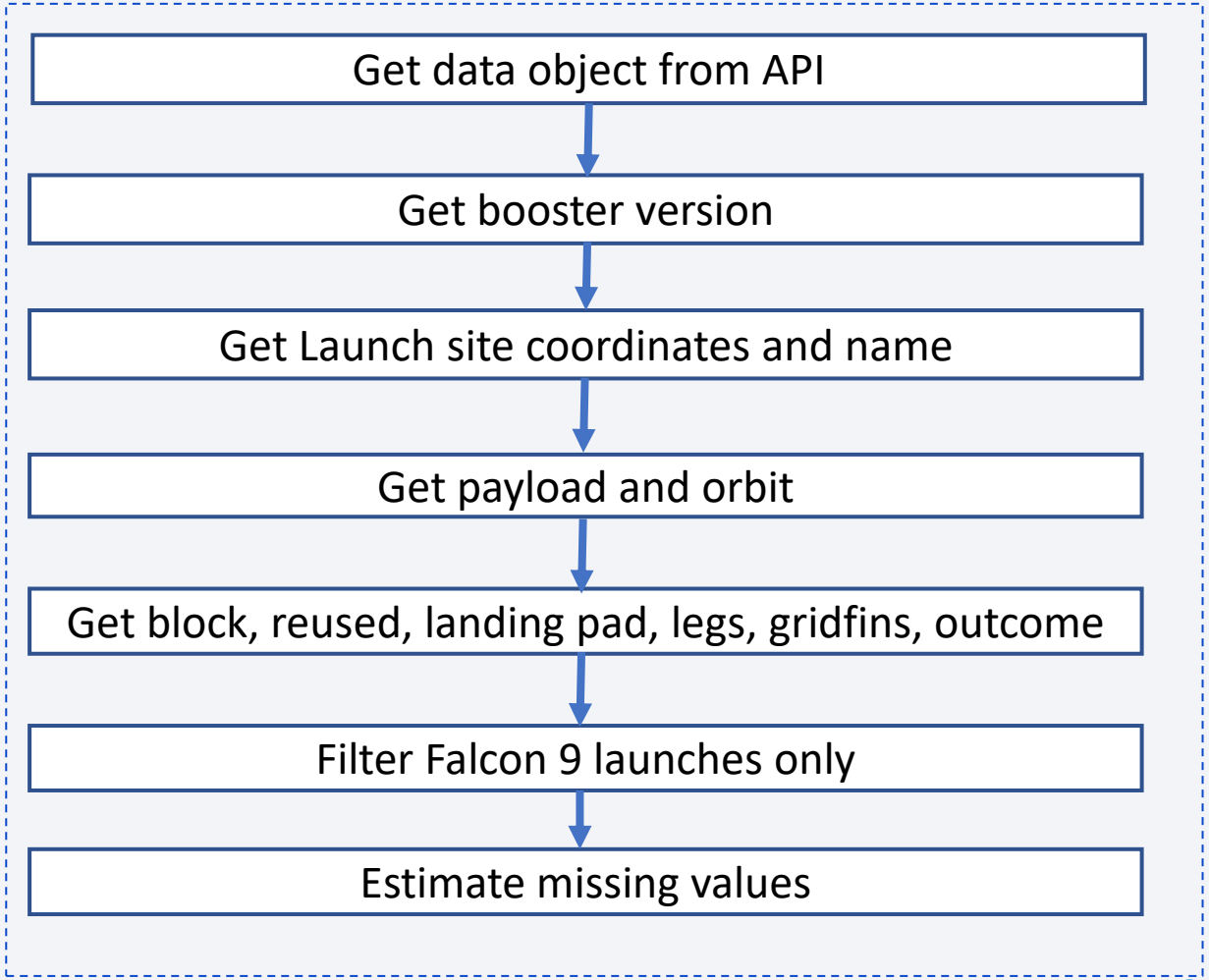
Data Collection

- Data collection was done in two different ways, they are further described in the next slides:
  - Reading directly from the SpaceX API
  - Webscraping from Wikipedia

Data Collection – SpaceX API

- The data object was collected from the API and functions used to compose it into a dataframe

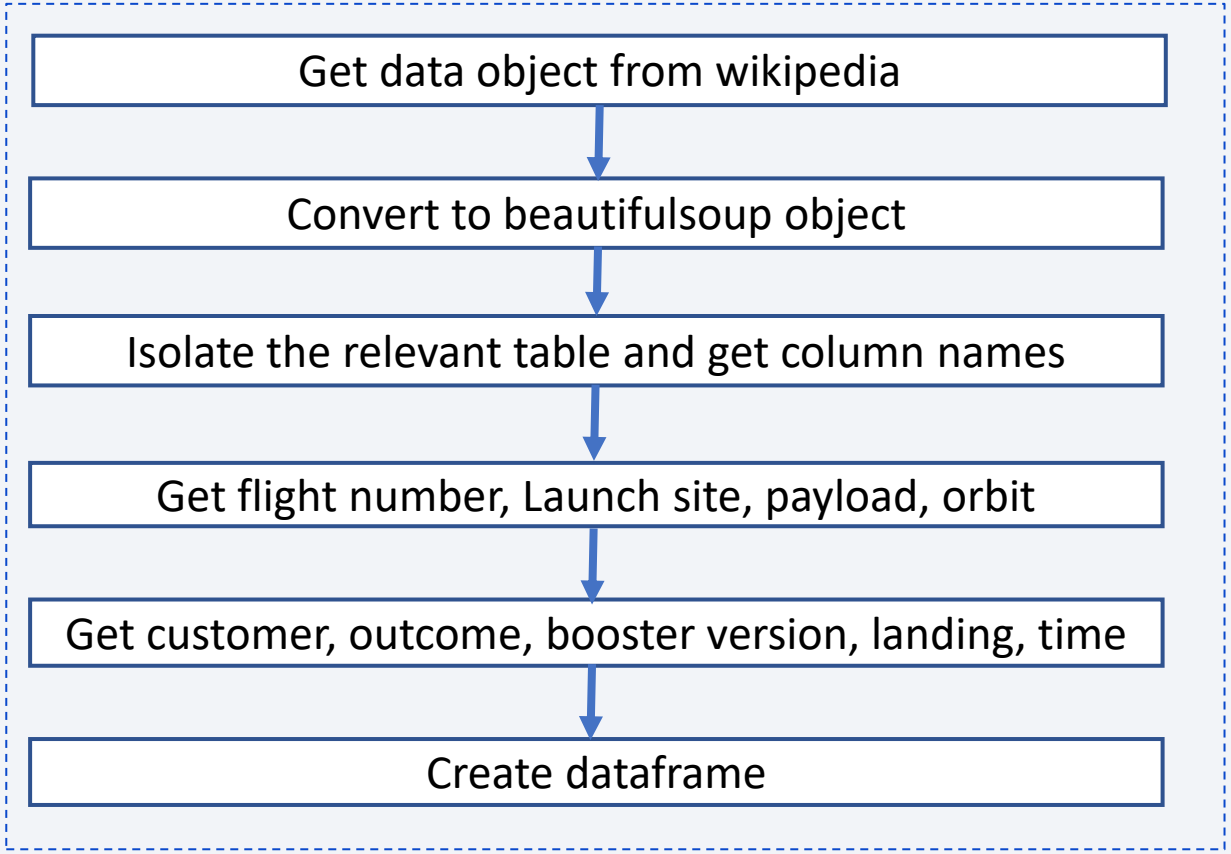
- https://github.com/MZBasingstoke/FinalProjectDataScience/blob/main/mod10_1-spacex-data-collection-api.ipynb



Data Collection - Scraping

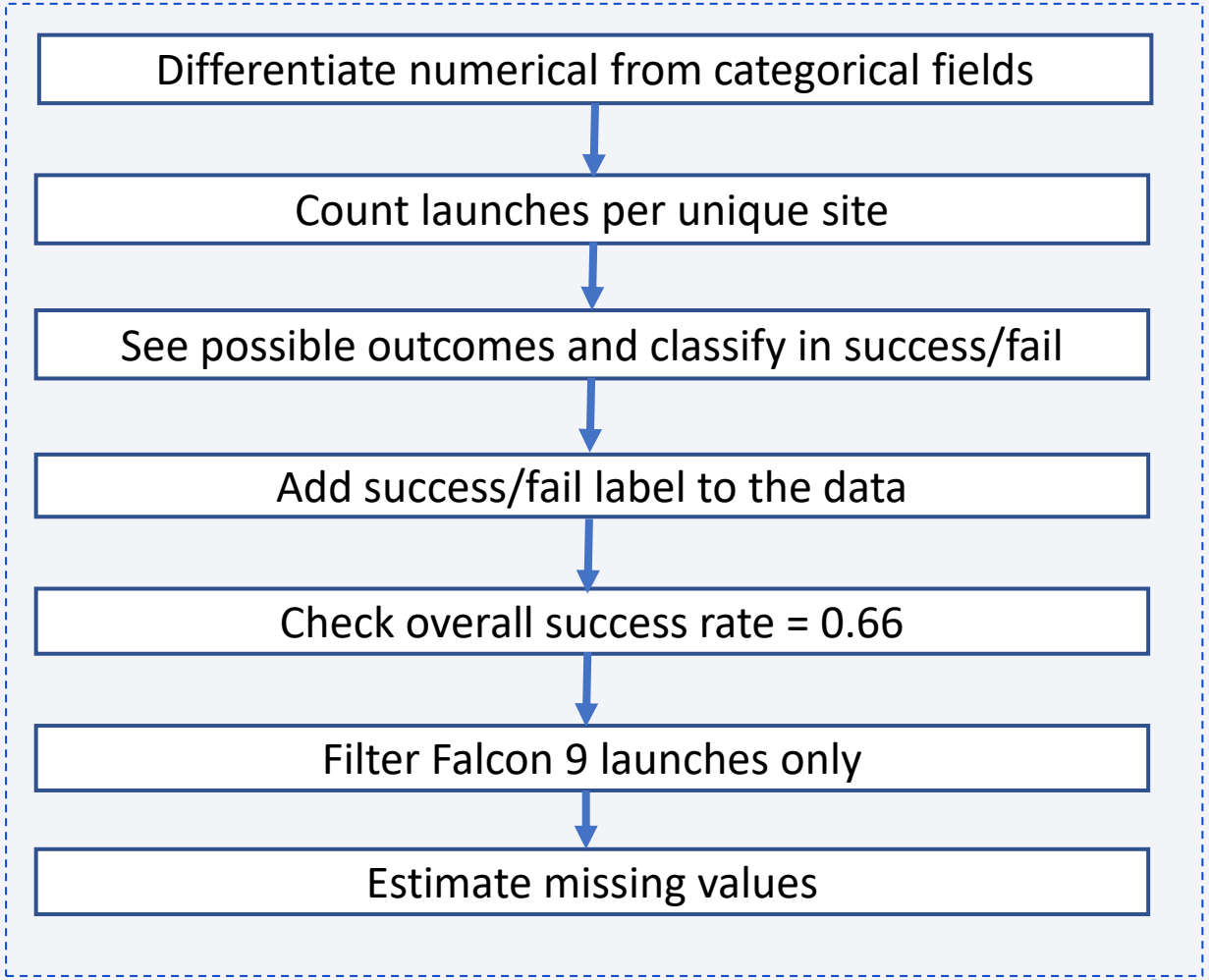
- Data was webscraped from Wikipedia, converted to a beautifulsoup object, and to a dataframe using functions to obtain clean data from the table cells

- https://github.com/MZBasingstoke/FinalProjectDataScience/blob/main/mod10_2_webscraping.ipynb



Data Wrangling

- Data was separated in numerical and categorical, and the unique launching sites were considered as well as their frequency
- The different types of outcomes were analysed and used to classify each launch in either success or failure
- That label was added to the data and will be used as the outcome label for the machine learning models
- https://github.com/MZBasingstoke/FinalProjectDataScience/blob/main/mod10_3_Data%20wrangling.ipynb



EDA with Data Visualization

- To help visualize success patterns, several features were plotted together in scatter plots with a colour code to distinguish between success and failure, yellow = success, blue = failure. The results are shown further down in this presentation
- The notebook with that work can be found at:

https://github.com/MZBasingstoke/FinalProjectDataScience/blob/main/mod10_5_eda-dataviz.ipynb

EDA with SQL

- The data based was interrogated with SQL queries. The queries are shown in the next slides and the results in the results section
- The notebook showing this work can be found at:

https://github.com/MZBasingstoke/FinalProjectDataScience/blob/main/mod10_4_eda-sql.ipynb

EDA with SQL - Queries

- Names of the launch sites
- 5 launches from CCA launch sites
- Total payload carried by boosters launched by NASA CRS
- Average payload carried by booster F9 v1.1
- Date for the first successful landing on a ground pad
- Booster names successful on drone ships with payload between 4000 – 6000 kg
- Total number of successes and failures
- Names of the boosters that carried the maximum load

EDA with SQL - Queries

- Months of the failed landings in drone ships in 2015
- Count of successful landings between 04/06/2010 and 20/03/2017

Build an Interactive Map with Folium

- A marker and circle was added to each launch site and plotted in a map with Folium
- For each site, a graph showing the successes and fail through time was also added.
- Finally, the proximity to railways, highways, coast and the nearest city was also studied
- The results are shown in the results section
- The notebook showing this work can be found at:

https://github.com/MZBasingstoke/FinalProjectDataScience/blob/main/mod10_6_launch_site_location.ipynb

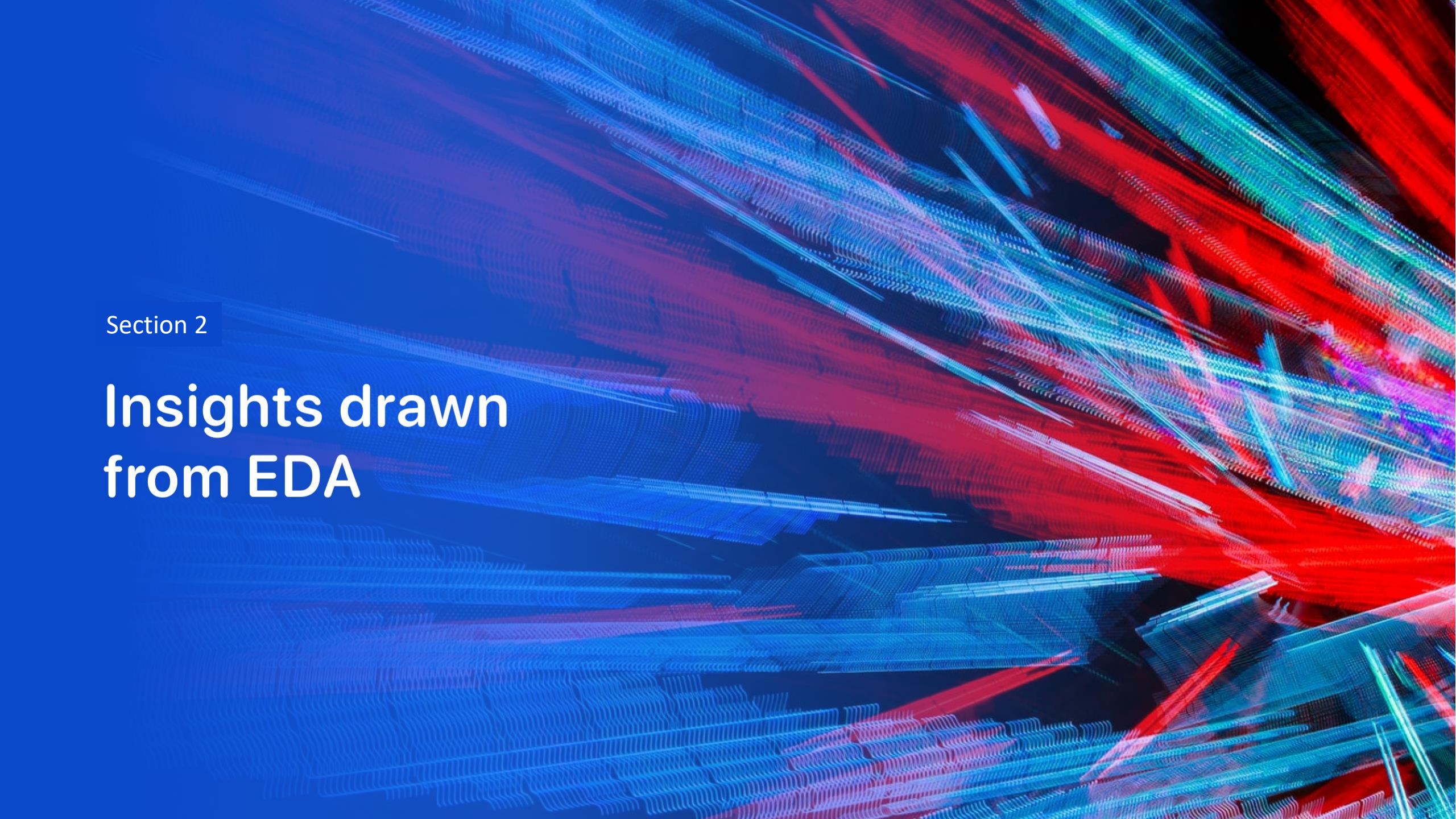
Predictive Analysis (Classification)

- Four classification models were tried and tested to predict the landing outcome of a future mission based on data from the past
- Available instances were separated into train and test sets. The models were trained and the accuracy on the test set reported
- Confusion matrix was plotted and the accuracy scores calculated for the four models. Three models were identified as giving promising results.
- This work can be found at:

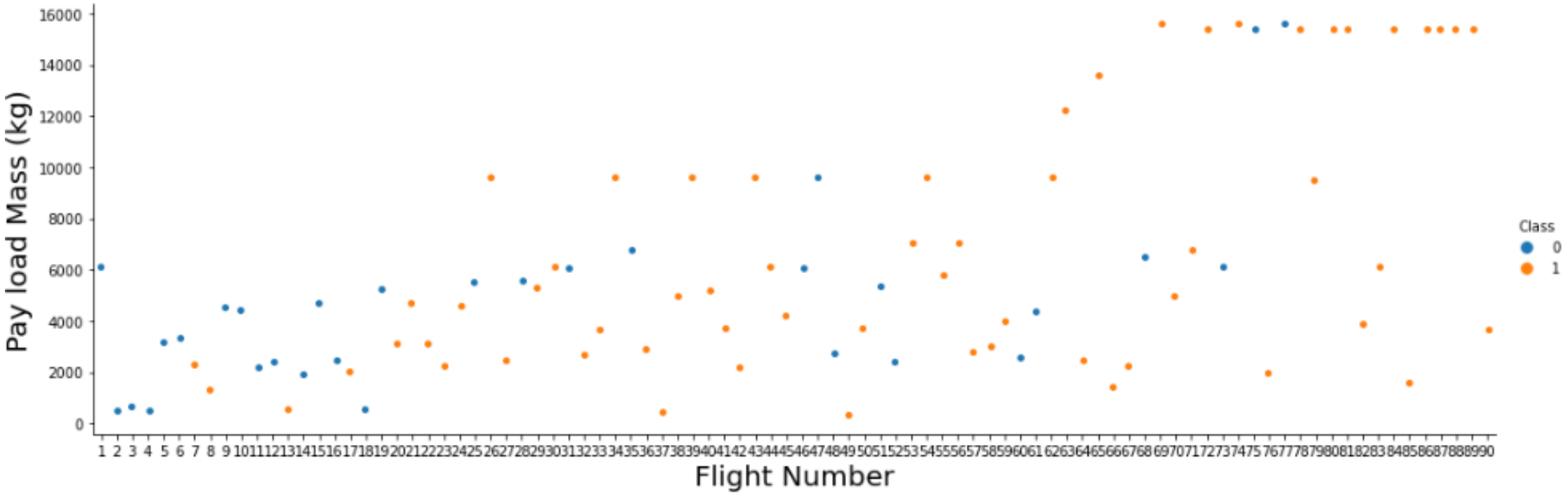
https://github.com/MZBasingstoke/FinalProjectDataScience/blob/main/mod10_7_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

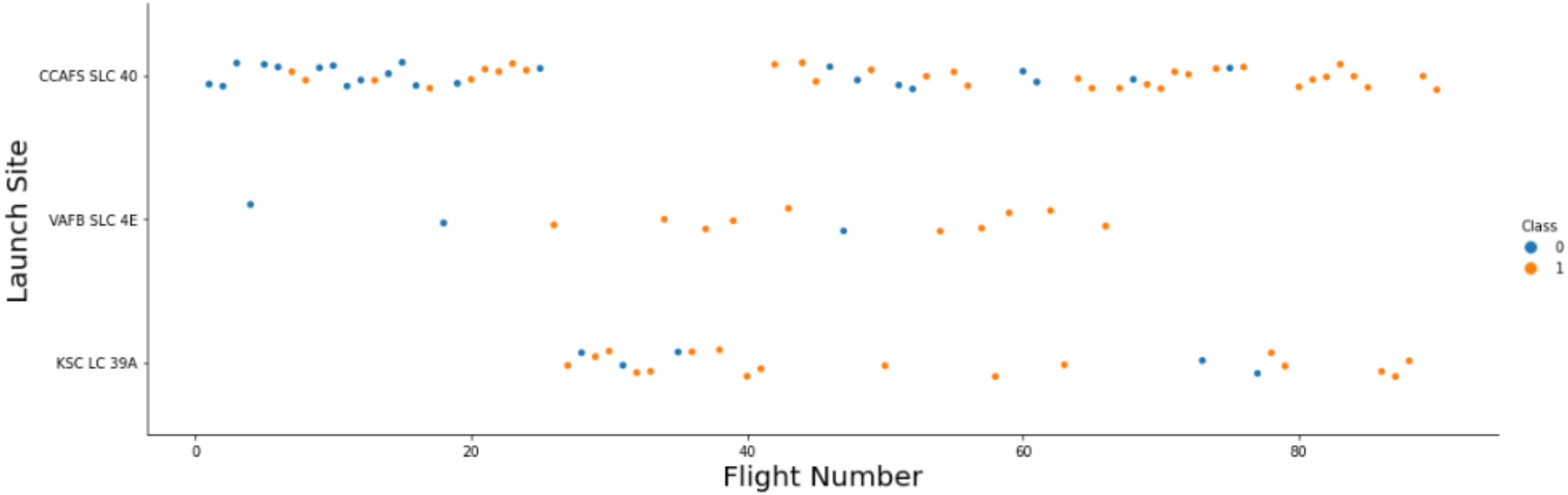

Section 2
Insights drawn from EDA

Payload mass vs Flight number



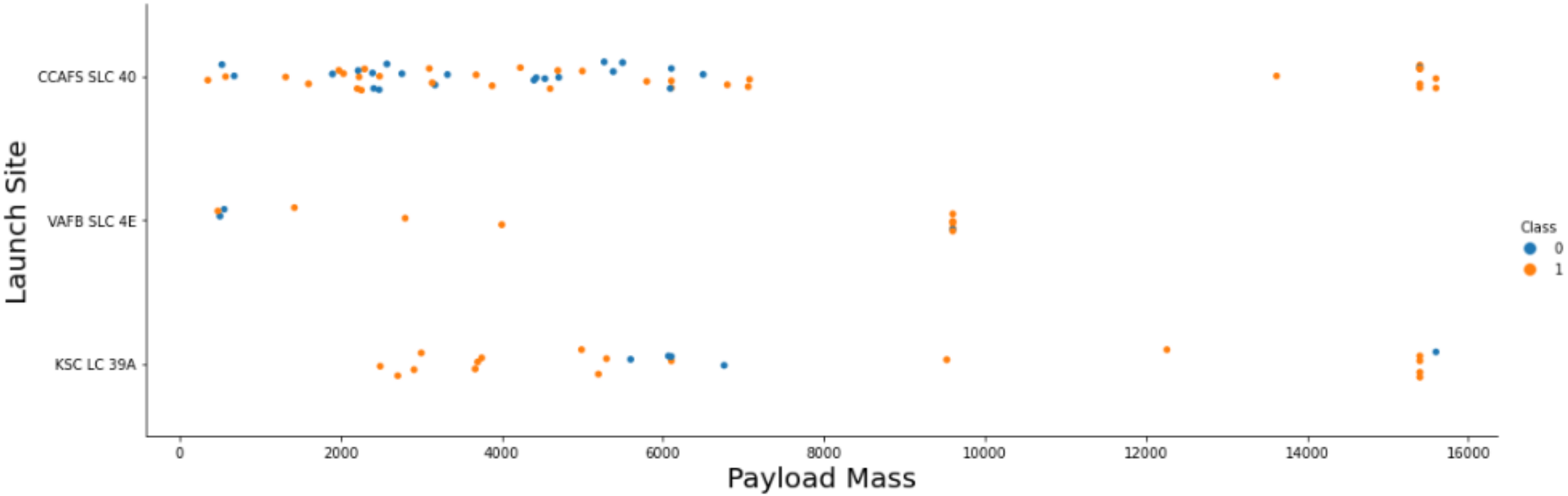
- It is clear that the success rate improved over time, and a higher payload seemed to have a negative effect in the first half of the flights, not so much in the most recent missions

Flight Number vs. Launch Site



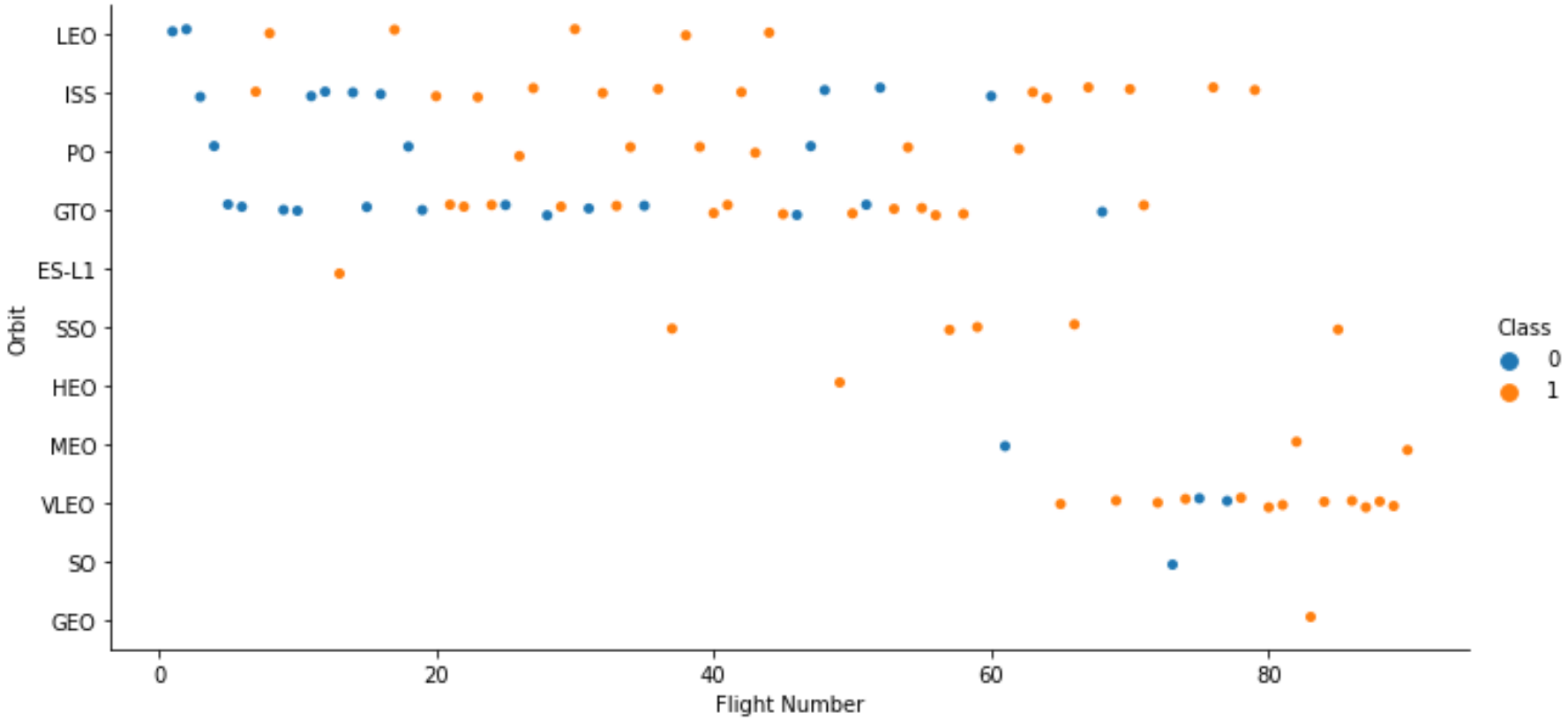
- Success rate from CCAFS LC-40 is lower than the other two sites, but this may be skewed by the fact of that site being used for most of the initial missions where the failure rate was much higher

Payload vs. Launch Site



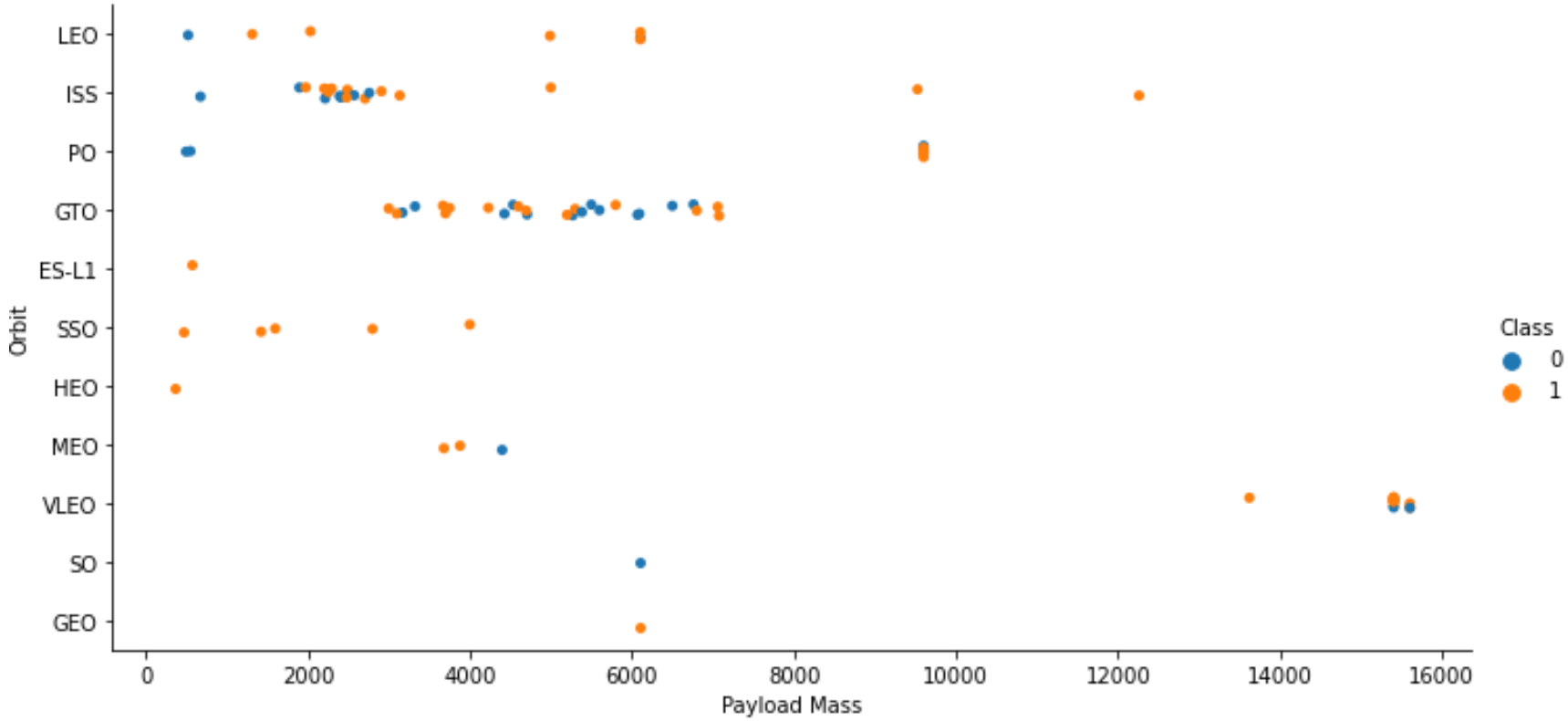
- It is observed that VAFB SLC 4E has not seen any heavy launch

Flight Number vs. Orbit Type



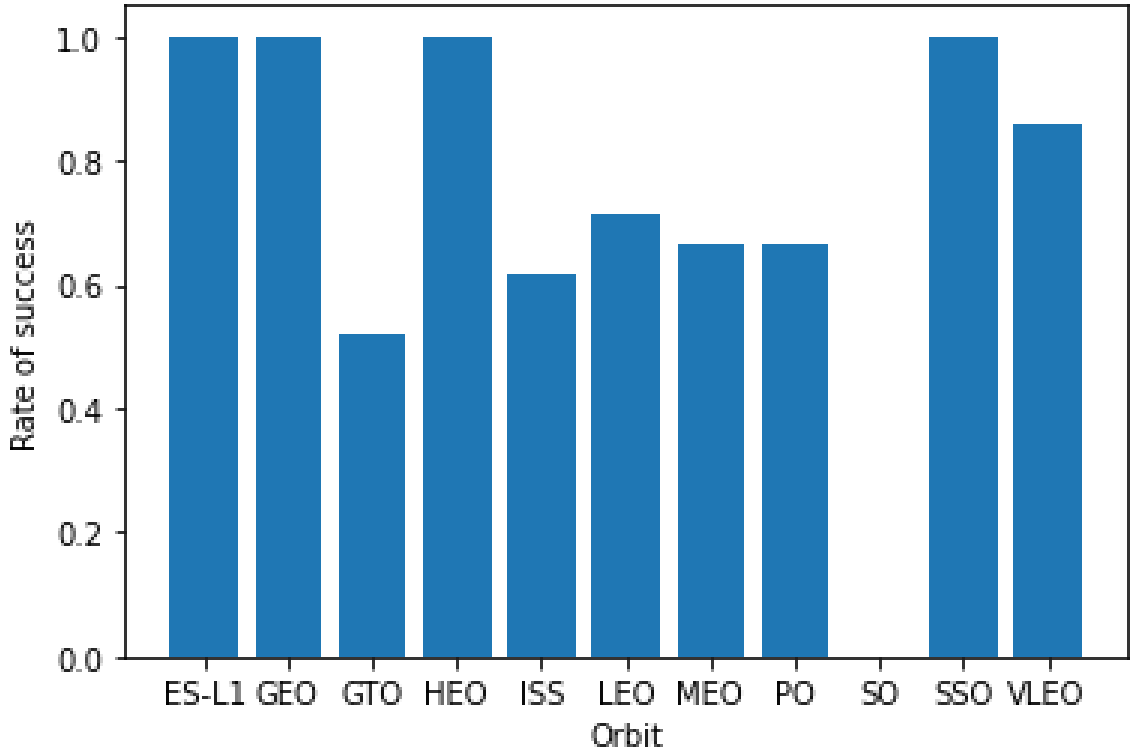
- GTO and ISS orbits seems to have a lower success rate. It is also observed that orbits LEO, ISS, PO and GTO were used from the beginning, but the other orbits have been progressively introduced in the program when the project was more mature. This may explain their apparently higher success

Payload vs. Orbit Type



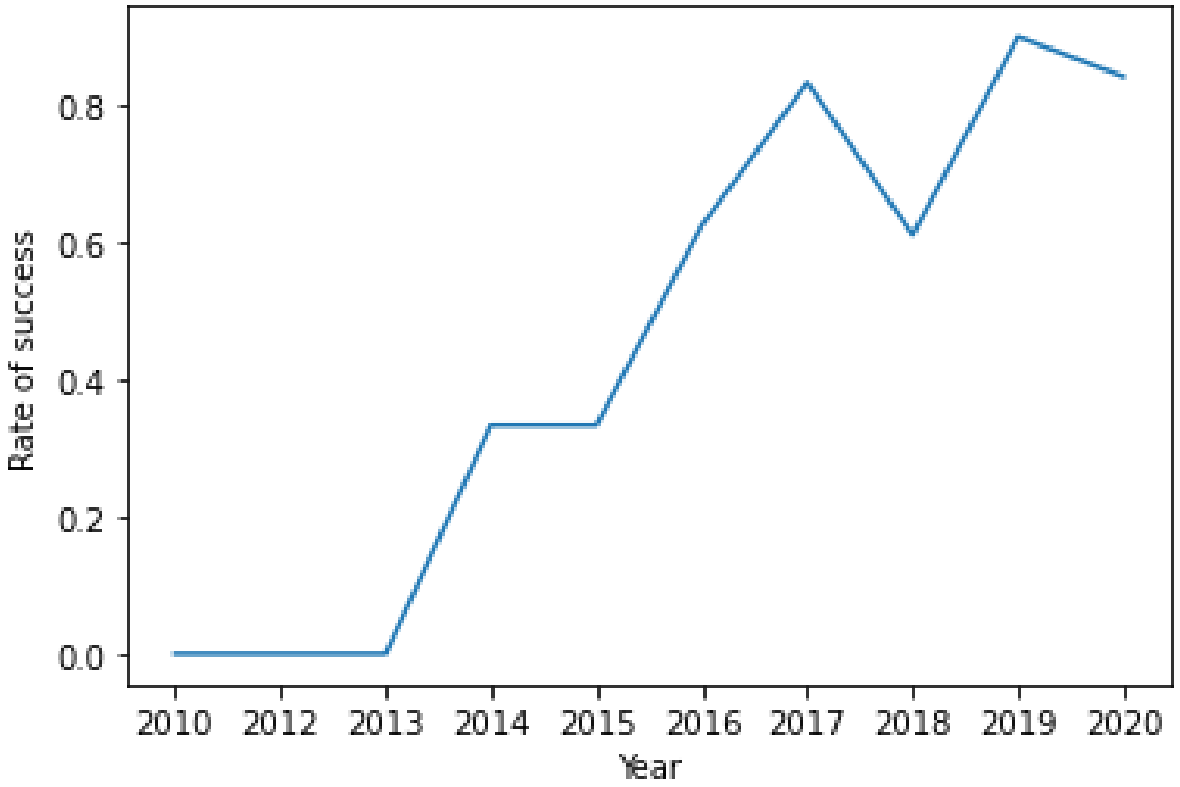
- High payloads were sent only to orbits ISS, VLEO and PO

Success Rate vs. Orbit Type



- 100% success rate for ES-L1, GEO and HEO, as well as 0% for SO are based in a single sample, so it is not indicative of the real probability. SSO however, has a 100% success rate based on multiple samples, and LEO seems slightly better than other commonly used orbits

Launch Success Yearly Trend



- Success rate clearly improved through the years from 0 to over 80%, although there is a drop in 2018

All Launch Site Names

```
1  %sql select distinct Launch_Site from SPACEXTBL
```

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

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

- There are four different launch sites, one in California (VAFB SLC-4E) and three in Florida

Launch Site Names Begin with 'CCA'

```
1 %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- These are the first 5 launches from launch sites beginning with CCA, which are located in the same facility in Florida

Total Payload Mass

```
1 %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer like 'NASA (CRS)%'
```

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

sum(PAYLOAD_MASS__KG_)
48213

- 48213 kg is the total payload for all the missions launched by NASA CRS

Average Payload Mass by F9 v1.1

```
1 %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- Launches with booster version F9 v1.1 carried an average payload of 2535 kg

First Successful Ground Landing Date

```
In [13]:
1 %%sql
2 select min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2)) as date_YYYYMMDD from SPACEXTBL
3 where [Landing _Outcome] like 'Success (ground%'
```

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

Out[13]:

date_YYYYMMDD
20151222

- The first successful landing on a ground pad was accomplished on the 22nd of December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select Booster_Version from SPACEXTBL
where [Landing _Outcome] like 'Success (drone ship)%' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

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

- Those were the booster names used for successful landings on drone ships carrying a payload between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
1 %sql select count (*) from SPACEXTBL where Mission_Outcome like 'Success%'
```

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

count (*)
100

```
1 %sql select count (*) from SPACEXTBL where Mission_Outcome like 'Failure%'
```

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

count (*)
1

- A total of 100 missions were catalogued as Success, this includes cases with payload status unclear
- Only 1 mission was catalogued as a failure. Failure mission outcome should not be confused with landing failure

Boosters Carried Maximum Payload

```
1 %sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ in (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 maximum payload was carried using these booster versions

2015 Launch Records

```
%%sql
select substr(Date,4,2) as Month, [Landing _Outcome], Booster_Version, Launch_Site  from SPACEXTBL
where substr(Date,7,4) = '2015' and [Landing _Outcome] like 'Failure (drone ship)'
```

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

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

- There were two failed landing_outcomes in drone ship in 2015. Their booster versions, and launch site names are detailed in this query

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

```
%%sql
select [Landing _Outcome], count([Landing _Outcome]) from SPACEXTBL where [Landing _Outcome] like 'Success%' and
substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) >= '20100604' and
substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) <= '20170320'
group by [Landing _Outcome] order by count([Landing _Outcome]) desc
```

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

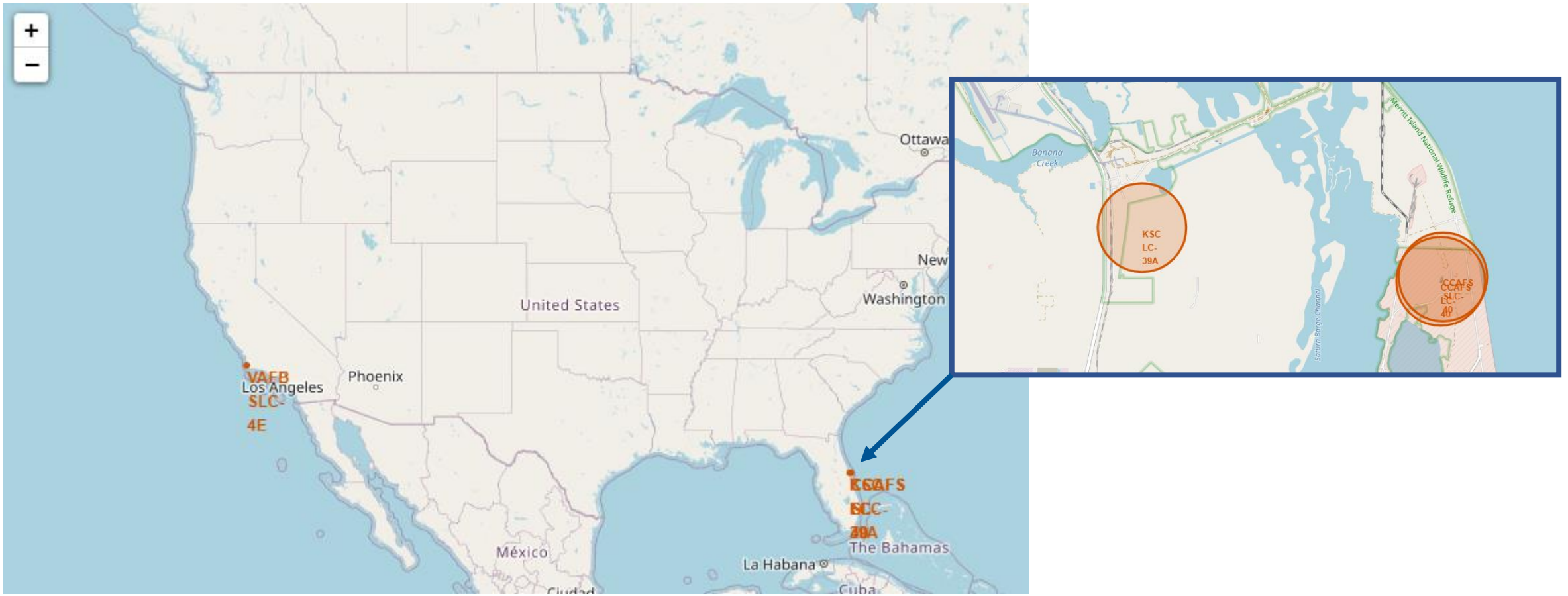
Landing _Outcome	count([Landing _Outcome])
Success (drone ship)	5
Success (ground pad)	3

- Between 2010-06-04 and 2017-03-20, there were 5 failed landings on a drone ship and 3 fail landings on a ground pad

Section 3

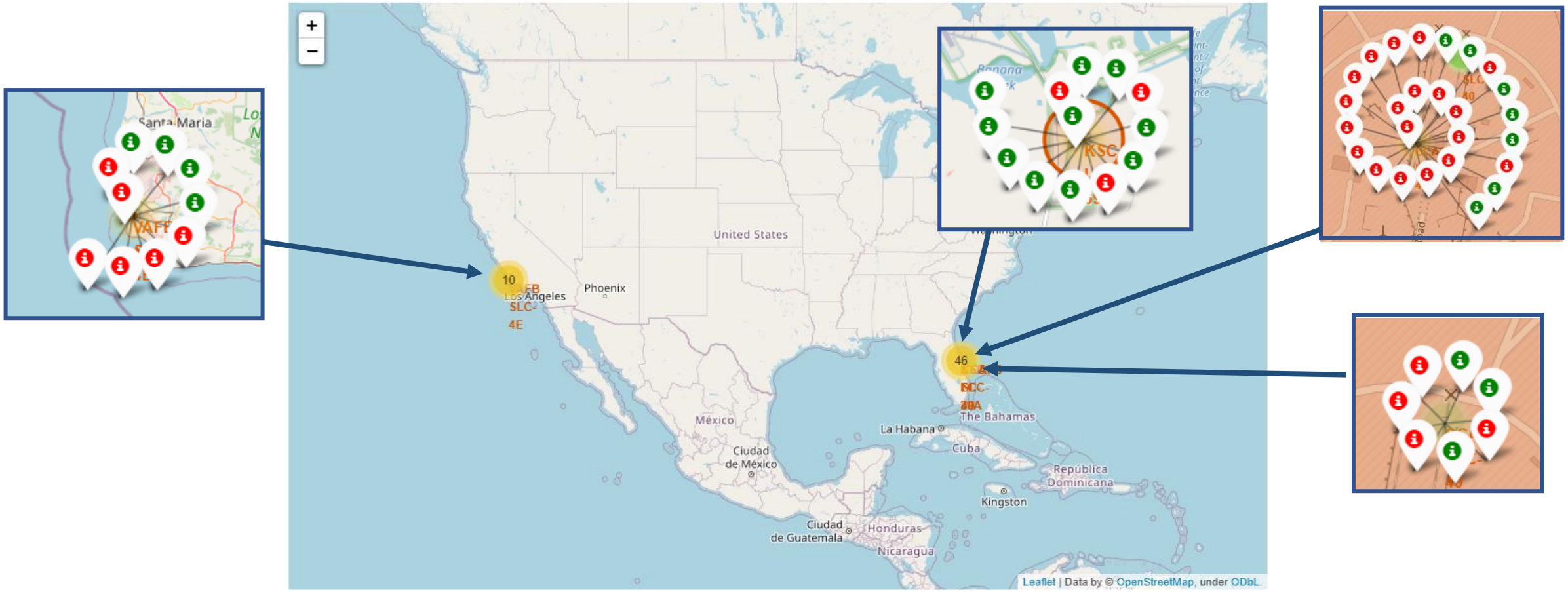
Launch Sites Proximities Analysis

Map – Launch sites



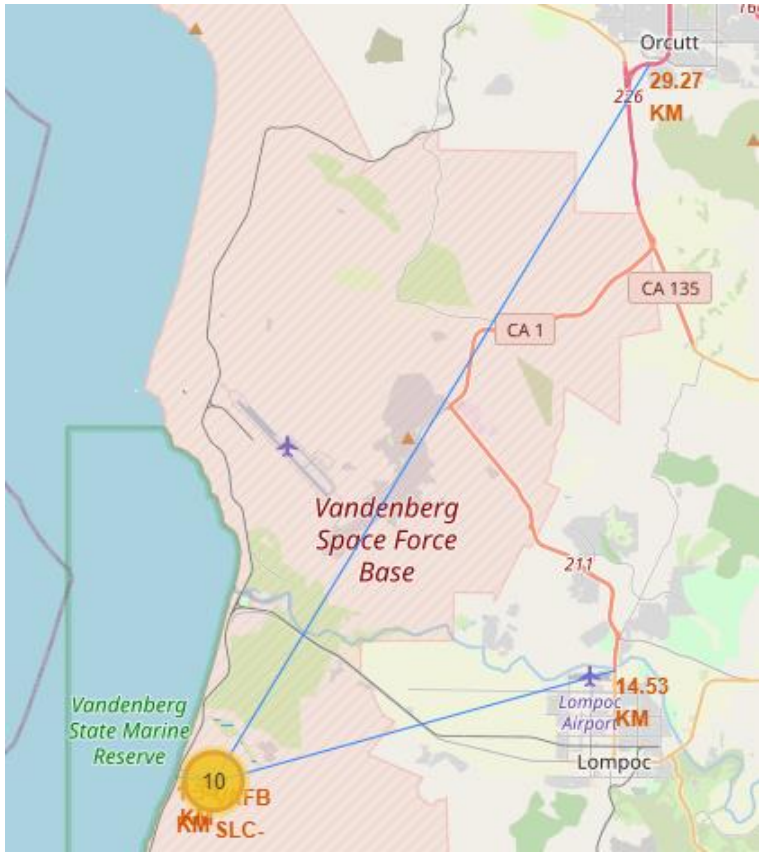
Launch sites are in California and Florida, identified with a marker and orange circle. The detail window show the three different sites in Florida, two of them (CCAFS SLC-40 and CCAFS LC-40) in the same facility. They are all located in the south of the country

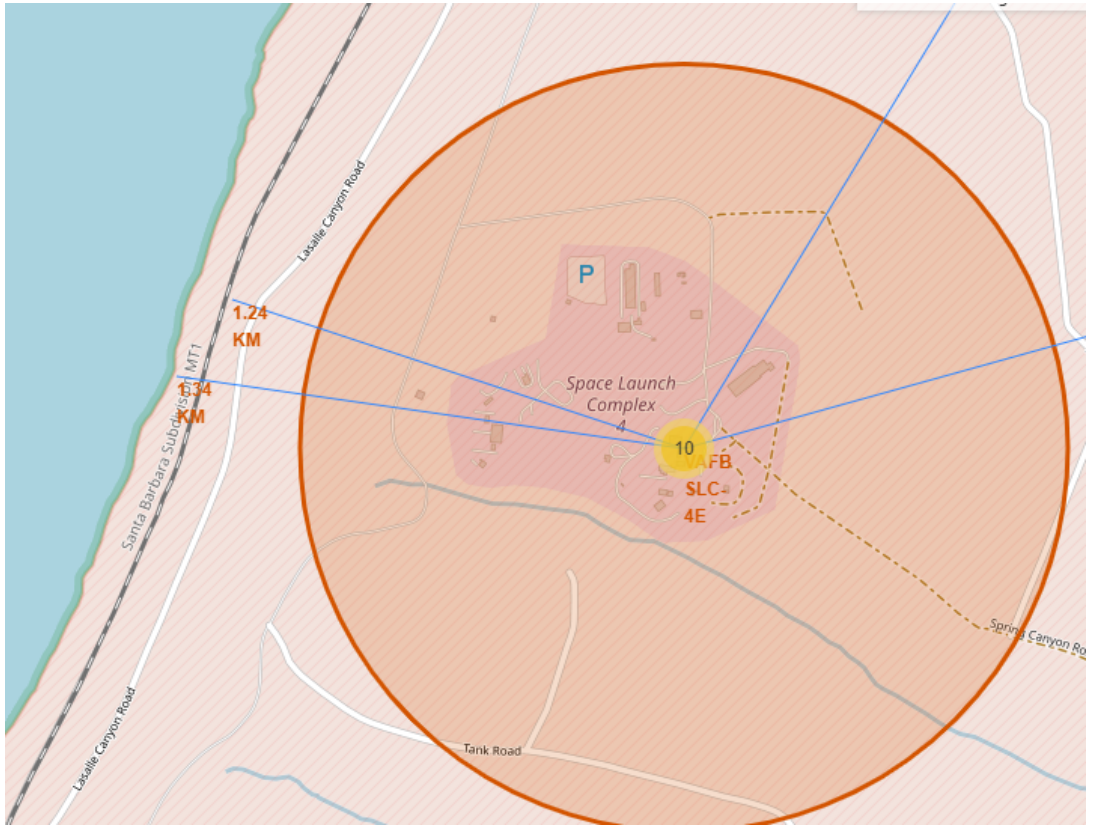
Map – Launch sites success/fail



Success failure points were added to each site, they can be viewed by zooming in and clicking on each site. The spirals show the timing

Map – Launch sites proximity



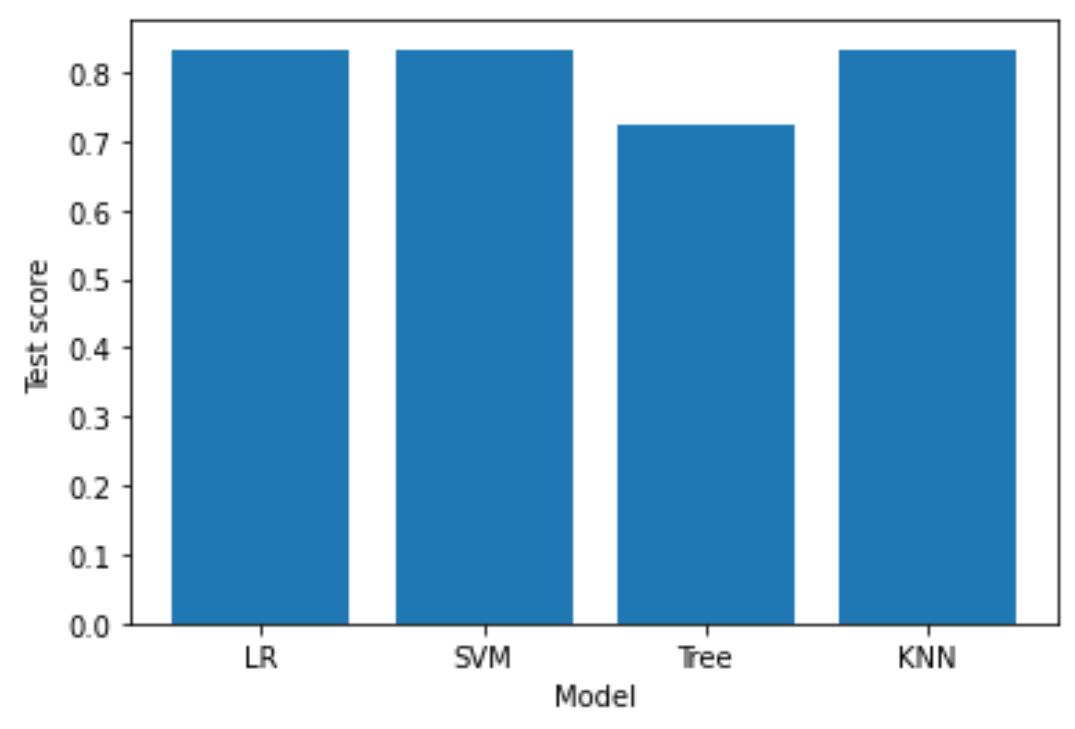


Launching sites (California displayed) are typically very close to a railway line and the coast, while also separated from cities. Proximity to a highway does not seem to be crucial factor, since the Florida sites are very close to one, but not so the California site

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- Accuracy on the test set was similar (0.83) for Logistic regression, Support vector machine and K-nearest neighbourghs.
- Accuracy for the Decision tree algorithm was lower at 0.72

Confusion Matrix

- The three models (LogReg, SVM and KNN sharing the highest performance also share the same confusion matrix
- No false negatives appeared in the test set, only 3 false positives

Conclusions

- Any of the 3 models with the highest accuracy can be used for prediction, there is no preference for any of them according to the score in the test set
- Accuracy is 0.83, which is meaningful, and the F1 score 0.89, making it a good model
- Training accuracy for those 3 models is between 0.84 - 0.85, similar to the test. However, the training accuracy for the decision tree was 0.87 vs 0.72 for the test, indicating overfitting

Appendix

- All the notebooks are available in Github

https://github.com/MZBasingstoke/FinalProjectDataScience

Thank you!