IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Marco Leti>
<05 October 2021>

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- We, at SpaceY, will predict if the Falcon 9 first stage will land successfully through the use of Data Science. If we can determine if the first stage will land, we can determine the cost of a launch

Section 1

Methodology

Methodology

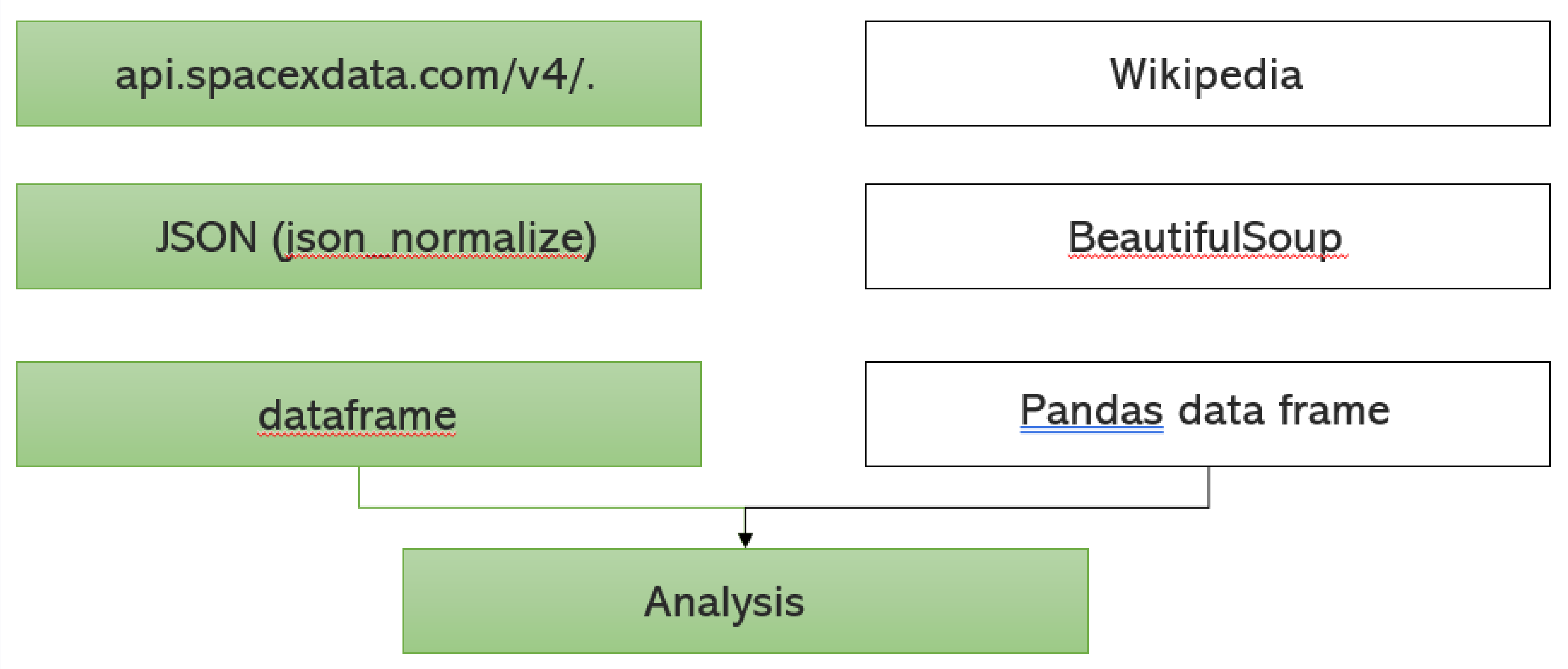
Executive Summary

- Data collection methodology:
  - Request to the SpaceX API and web scraping Wikipedia page
- Perform data wrangling
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome per orbit type
  - Create a landing outcome label from Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Usage of Logistic regression, SVM, Tree and KNN models

Data Collection

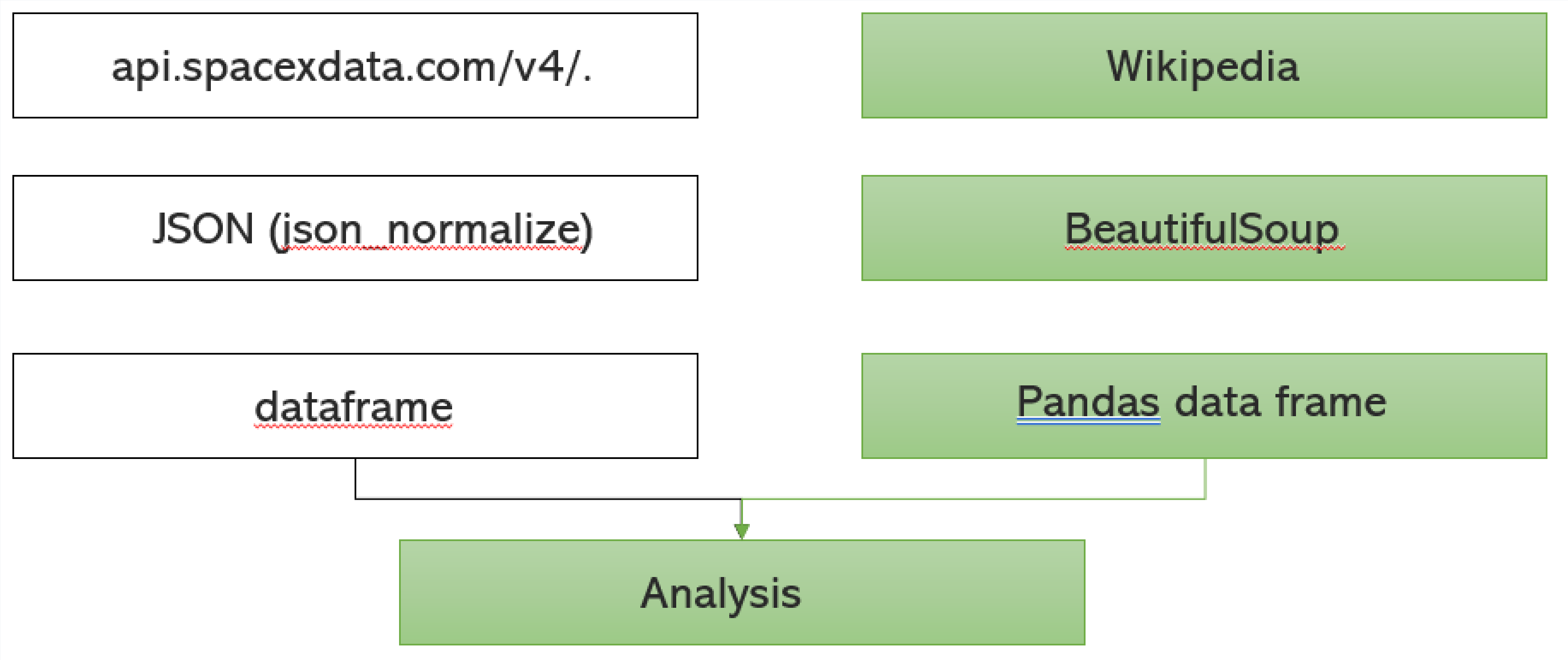
- A request to the SpaceX API was performed. The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/. We have the different end points, for example: /capsules and /cores We will be working with the endpoint api.spacexdata.com/v4/launches/past.
- To convert this JSON to a dataframe, we will use the json_normalize function.
- Another source for obtaining Falcon 9 Launch data is web scraping related Wiki pages using the Python BeautifulSoup package
- Then we parse the data from those tables and convert them into a Pandas data frame for further visualization and analysis.

Data Collection – SpaceX API



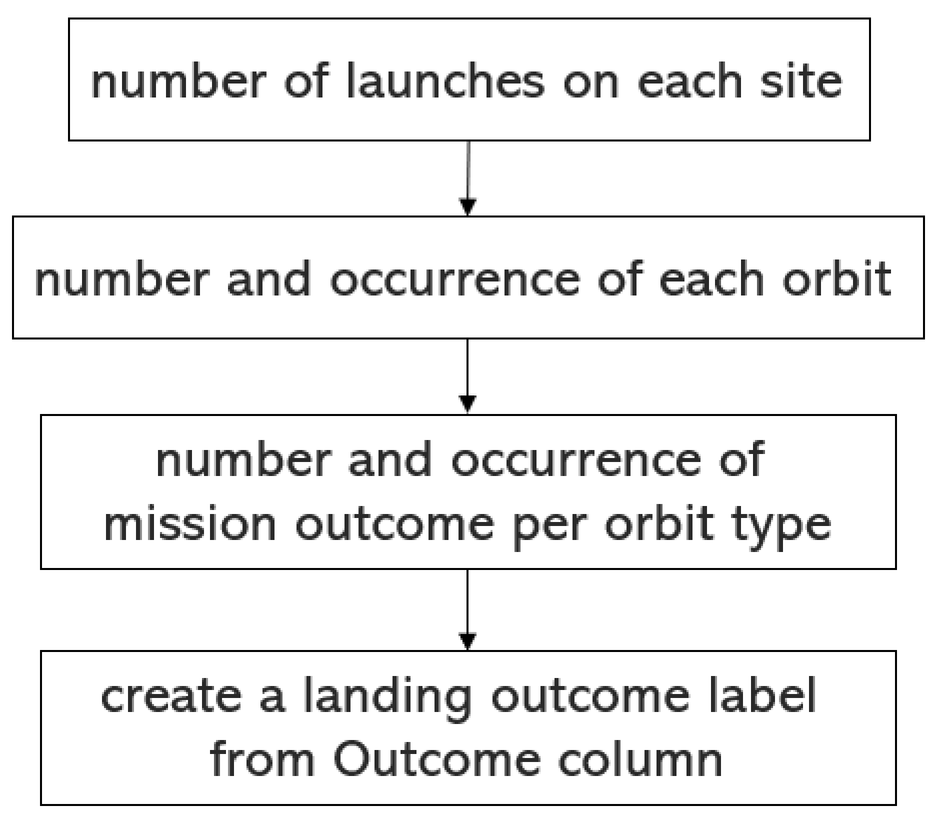
Source: https://github.com/MarcoLeti/IBMProfessionalDS/blob/main/1.1_Data_Collection.ipynb

Data Collection - Scraping



Source: https://github.com/MarcoLeti/IBMProfessionalDS/blob/main/1.1_Data_Collection.ipynb

Data Wrangling



Source: https://github.com/MarcoLeti/IBMProfessionalDS/blob/main/1.2_DataWrangling.ipynb

EDA with Data Visualization

- Scatterplot to Visualize the relationship between Flight Number and Launch Site
- Scatterplot Visualize the relationship between Payload and Launch Site
- Bar Chart to Visualize the relationship between success rate of each orbit type
- Scatterplot to Visualize the relationship between FlightNumber and Orbit type
- Scatterplot to Visualize the relationship between Payload and Orbit type
- Line chart to Visualize the launch success yearly trend

Source: https://github.com/MarcoLeti/IBMProfessionalDS/blob/main/2.2_EDAWithPython.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an Interactive Map with Folium

Markers:

- All launch sites on a map were marked. We added to answer to these questions:
  - Are all launch sites in proximity to the Equator line?
  - Are all launch sites in very close proximity to the coast?
- Success/failed launches for each site on the map. We added to answer to these questions:
  - to easily identify which launch sites have relatively high success rate

Source: https://github.com/MarcoLeti/IBMProfessionalDS/blob/main/3.1_VisualAnalytics.ipynb

Build a Dashboard with Plotly Dash

- Pie Chart with success rates of specific sites or total will allow to visualize if there is a correlation between place and success or failure
- Scatter plot showing the relation between payload mass and class specific for every booster version, will allow to identify if there is any correlation between success and payload for a specific site or for the general case

Source: https://github.com/MarcoLeti/IBMProfessionalDS/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We built 4 different classification models:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- Hyperparameter tuning done with GridSearchCV function with cv of 10
- Confusion matrix and accuracy scores calculated to find the best model

Source: https://github.com/MarcoLeti/IBMProfessionalDS/blob/main/4.1_PredictiveAnalytics.ipynb

Results

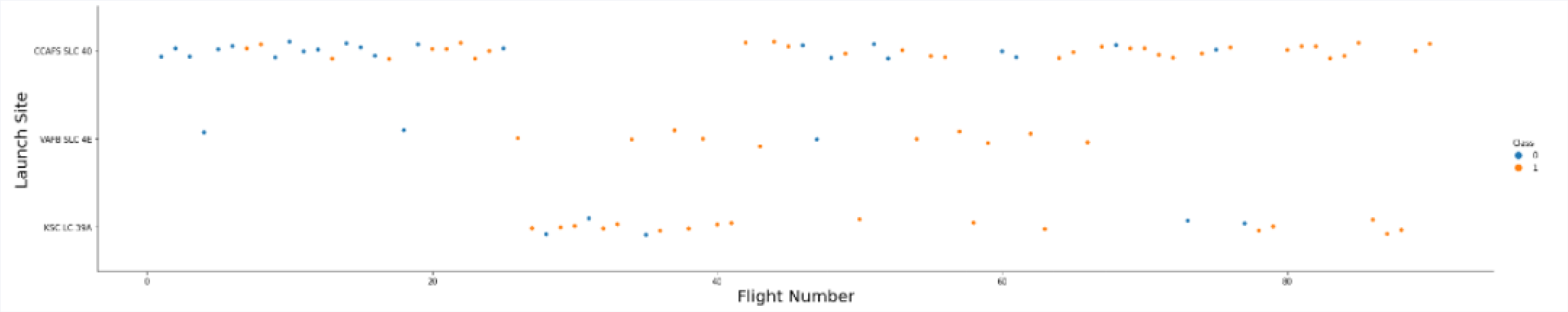
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

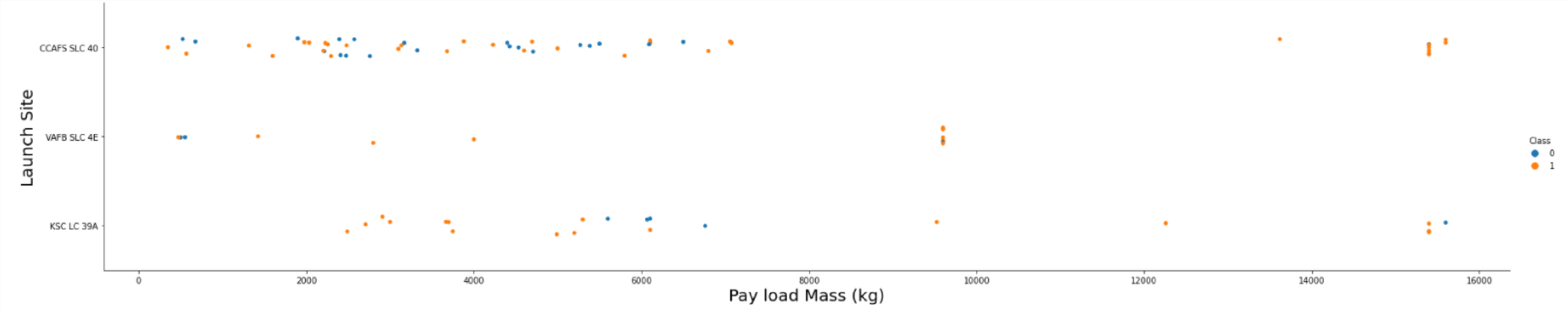
- Flight Number vs. Launch Site



- Positive launch outcomes seem to be related with high numbe

Payload vs. Launch Site

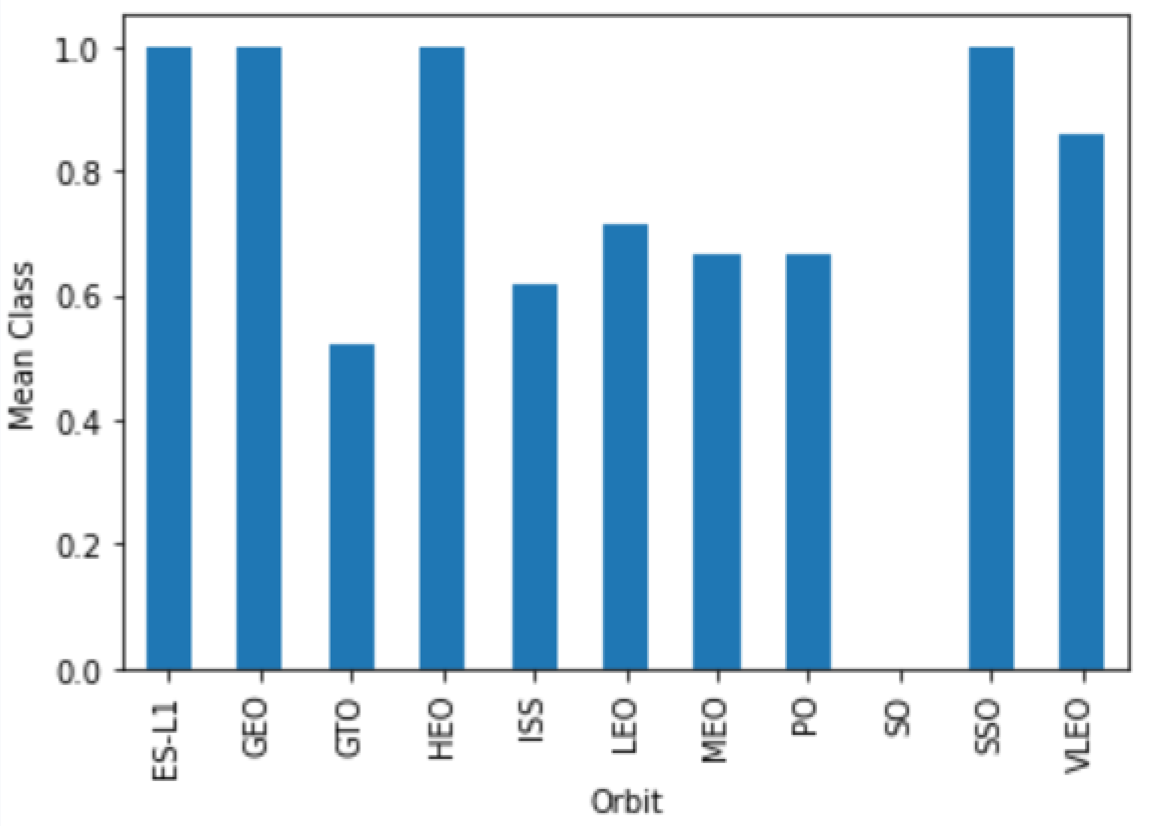
- Payload vs. Launch Site



- Some measures of payload mass seem to be launched only on certain sites.

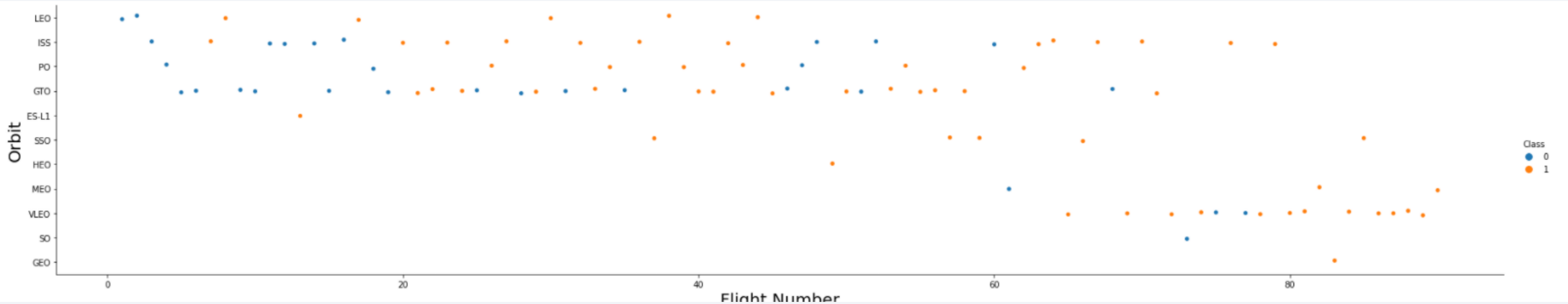
Success Rate vs. Orbit Type

- Some Orbit Types have very high success rates, other have very low.
- It seems that a successful launch is dependent on the orbit type.



Flight Number vs. Orbit Type

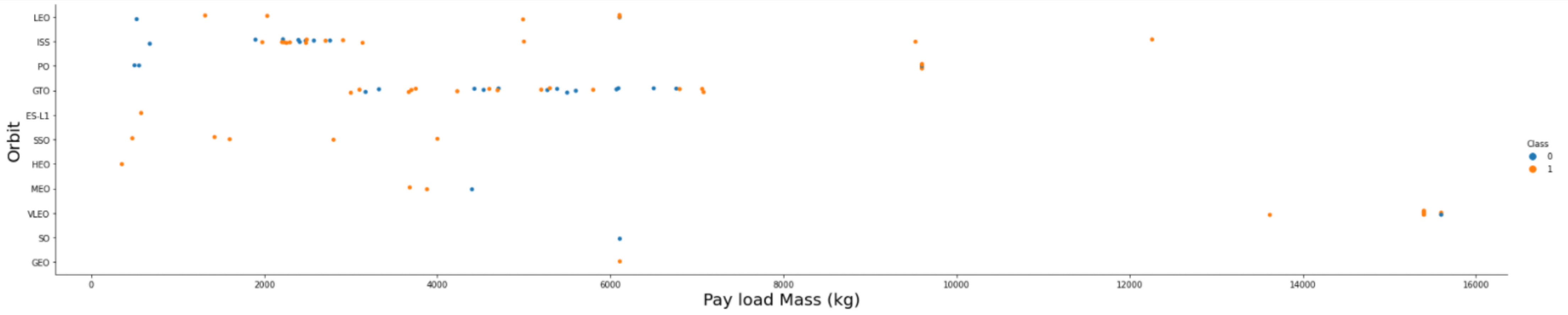
- Flight number vs. Orbit type



- 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

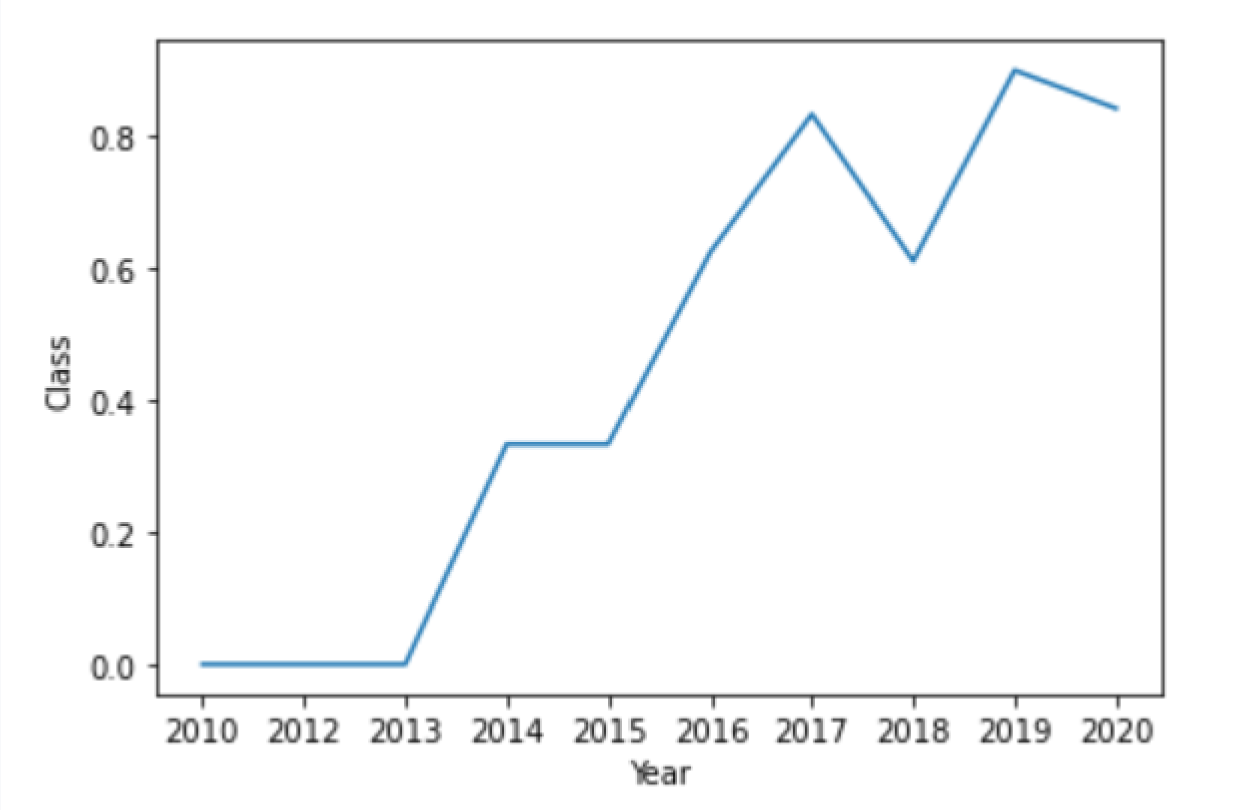
Payload vs. Orbit Type

- payload vs. orbit type



- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits

Launch Success Yearly Trend



- Success rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

```
%sql SELECT DISTINCT launch_site FROM SPACEXTBL
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8f
Done.
```

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`

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

```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/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

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql SELECT booster_version, SUM(payload_mass__kg_) AS total_payload_mass_kg FROM SPACEXTBL GROUP BY booster_version;
```

```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/
Done.
```

booster_version	total_payload_mass_kg
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092
F9 B4 B1045.1	362

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(payload_mass__kg_) AS average_payload_mass_f9v11_kg FROM SPACEXTBL WHERE booster_version LIKE 'F9 v1.1%';
```

```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUD
Done.
```

average_payload_mass_f9v11_kg
2534

First Successful Ground Landing Date

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

```
%sql SELECT MIN(date) AS first_successful_landing FROM SPACEXTBL WHERE mission_outcome = 'Success';
```

```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appd
Done.
```

first_successful_landing
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 SPACEXTBL WHERE mission_outcome = 'Success' AND payload_mass__kg_ > 4000 AND payload_mass__kg_
< 6000;
```

```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

booster_version
F9 B4 B1040.2
F9 B4 B1040.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5 B1058.2
F9 B5B1054
F9 B5B1060.1
F9 B5B1062.1
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1032.2
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1032.1
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome,COUNT(*) AS total_mission_outcome FROM SPACEXTBL GROUP BY mission_outcome;
```

```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdoma
Done.
```

mission_outcome	total_mission_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

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

```
%sql SELECT DISTINCT booster_version FROM SPACEXTBL WHERE booster_version IN (SELECT booster_version FROM SPACEXTBL ORDER BY payload_mass
__kg_ DESC LIMIT 1);
```

```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

booster_version
F9 B5 B1048.4

2015 Launch Records

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

```
%sql SELECT DISTINCT booster_version, launch_site FROM SPACEXTBL where landing__outcome = 'Failure (drone ship)' AND date LIKE '%2015%';
```

```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
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 SPACEXTBL WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY COUNT(*) DESC;
```

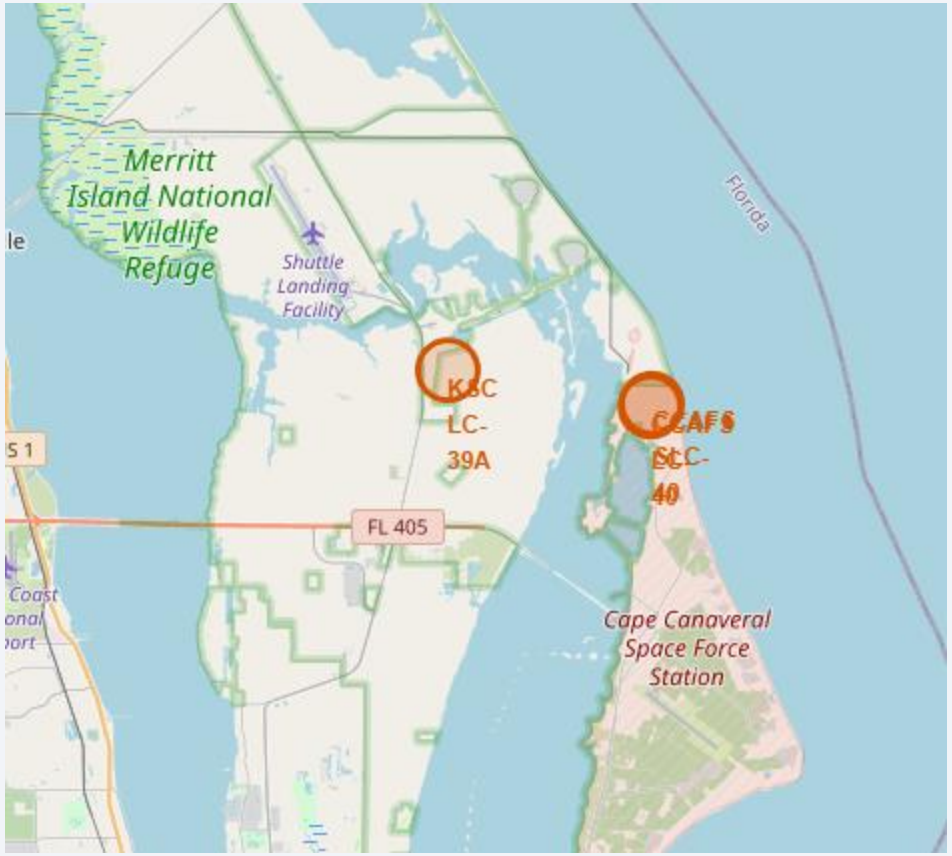
```
 * ibm_db_sa://qcg78281:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/BLUDB
Done.
```

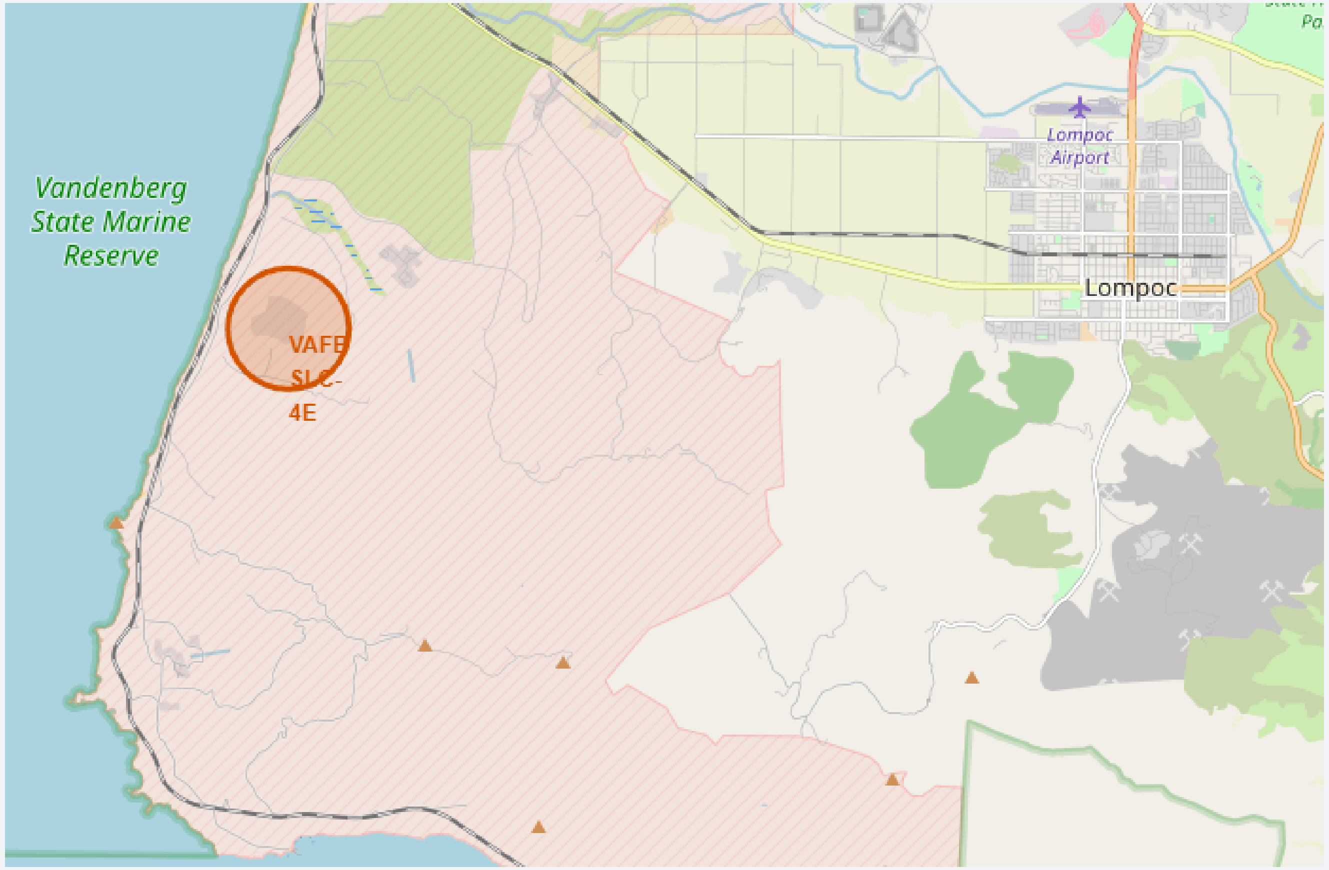
landing__outcome	**2**
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

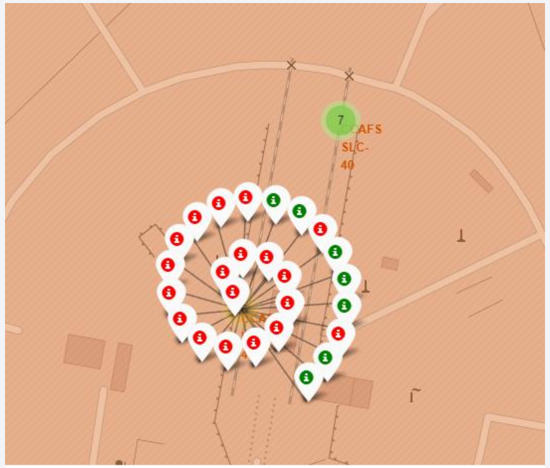
Launch sites locations



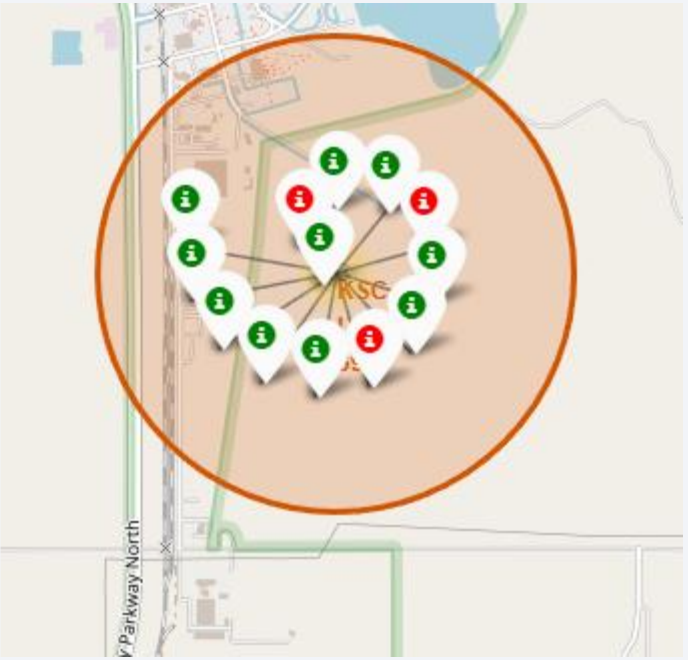


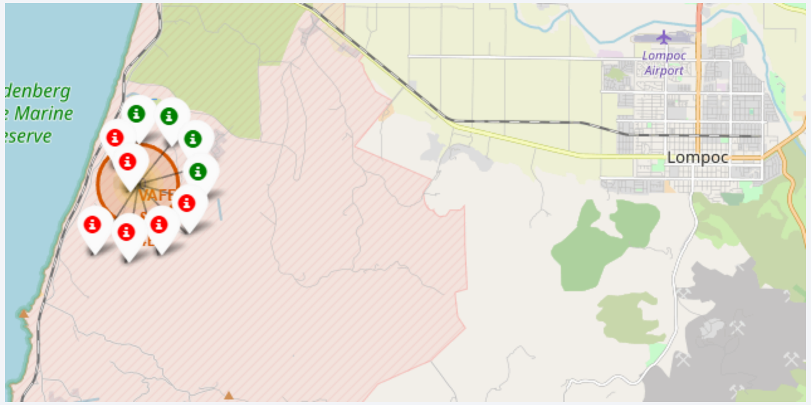
- Launch sites in proximity to the Equator line
- Launch sites in very close proximity to the coast

Success/failed launches for each site



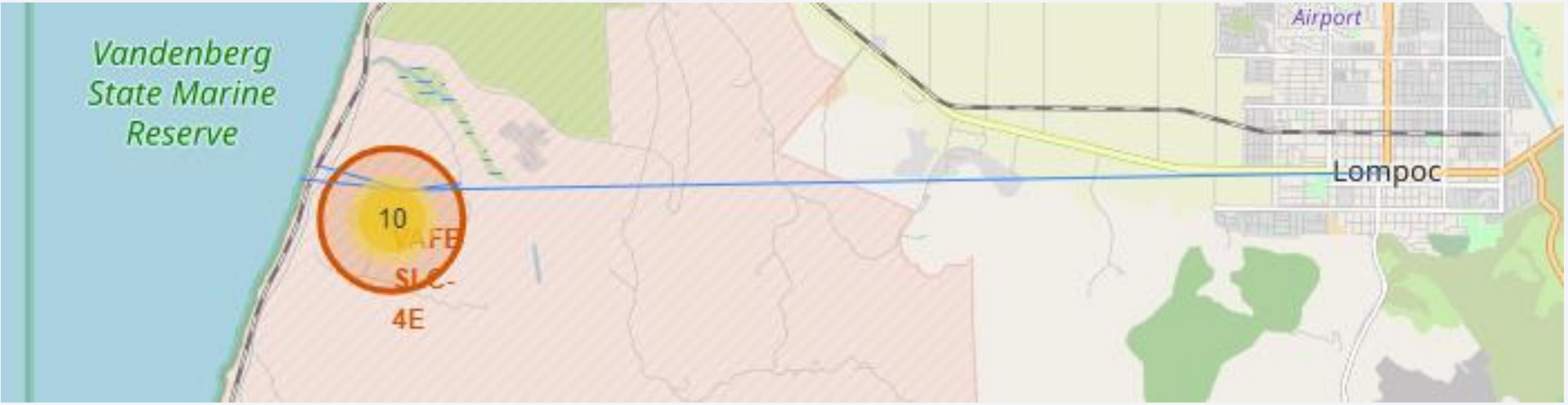
- Bad rates sites





- Good rates sites

Launch Sites and Proximities

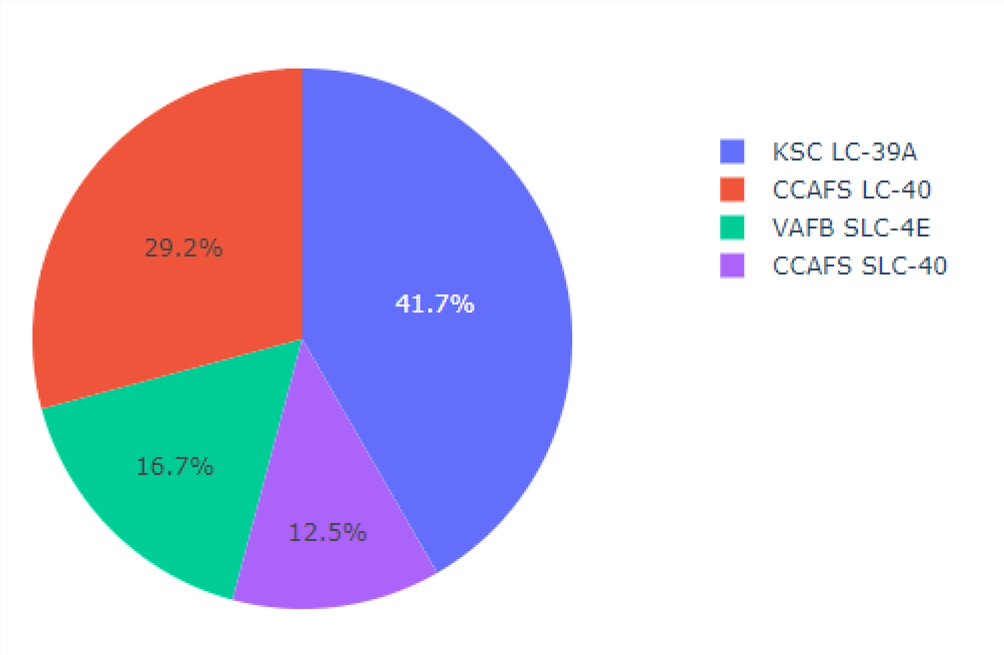


- All sites are close to the coast and far from cities. Some of them are close to the highway and railway

Section 5

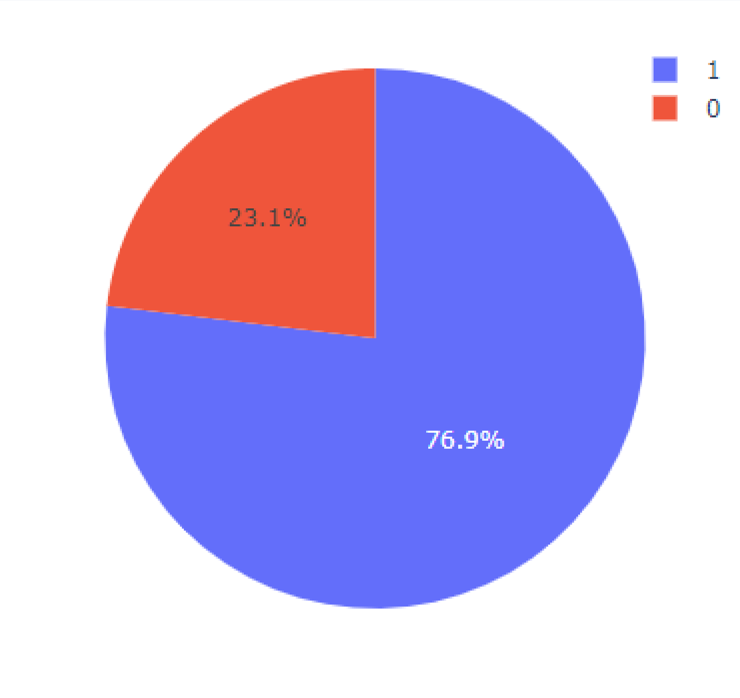
Build a Dashboard with Plotly Dash

Success count for all sites



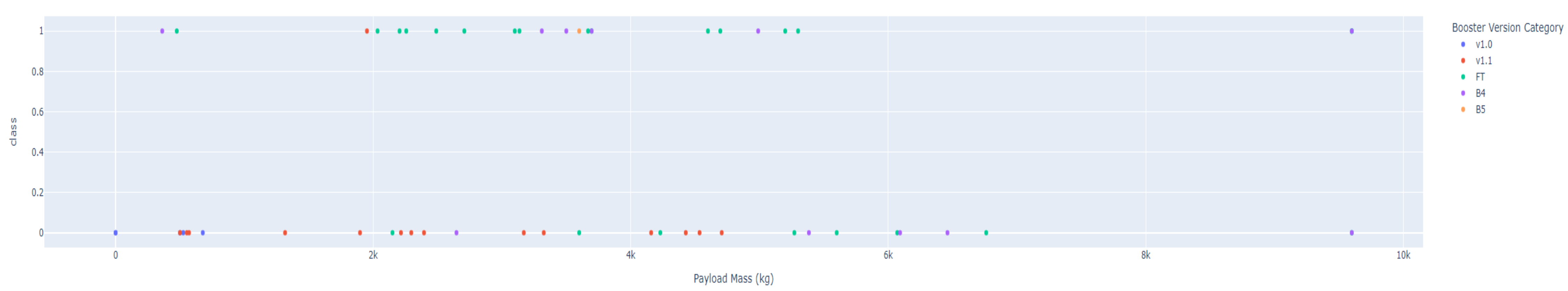
- The majority of the successful launches have been done on the site KSC LC-39A

Highest launch success



- For the site KSC LC-39A it is possible to see that the success rate is very high: 76.9%

Payload vs. Launch Outcome



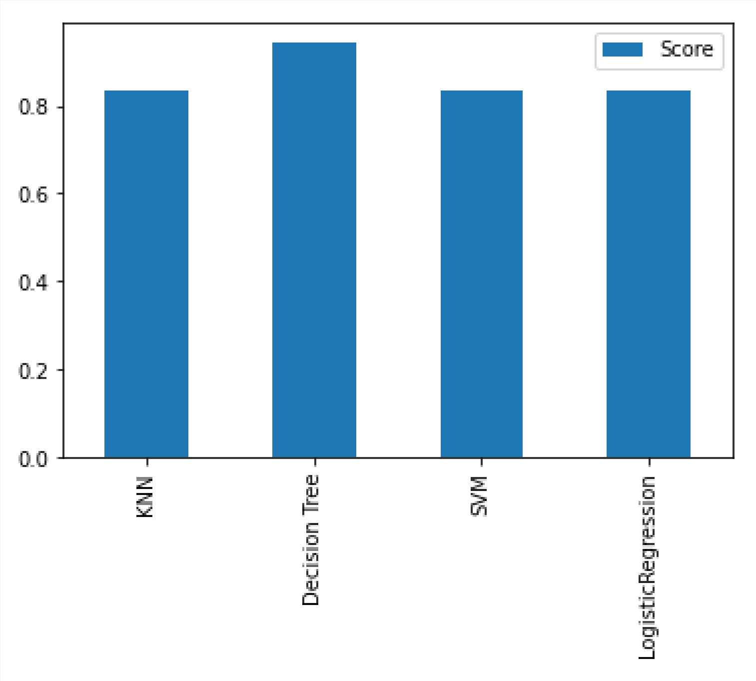
- We can clearly see from the plot that the booster version FT has higher success rate when the payload is between 2k and 6k.
- The booster v1.1 instead has a bad success rate

Section 6

Predictive Analysis (Classification)

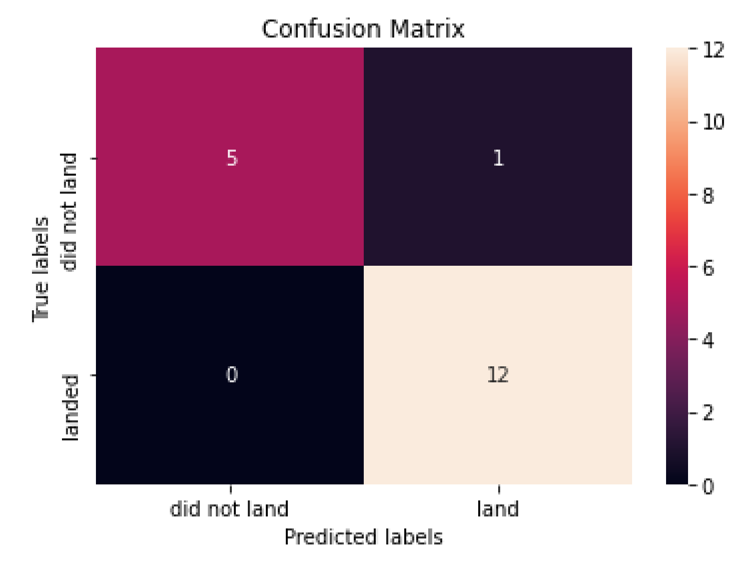
Classification Accuracy

- The highest accuracy is given by the decision tree model.
- The other models perform more or less the same



Confusion Matrix

- 5 records on the test set which did not land were identified correctly by the model
- 1 record which did not land was identified incorrectly as landed by the model
- 12 landed cases were all identified correctly



Conclusions

- Decision tree model performed the best
- We can use this model to predict with a high confidence if the launch will land successfully
- 0.94% of success of the model on the test set is high rate which shows that we did not overfit the training set.
- We are ready to open our SpaceY company ☺

Appendix

Original data set preview

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Thank you!