IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rodrigo de Alvarenga Mattos
December 30, 2021

Outline

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

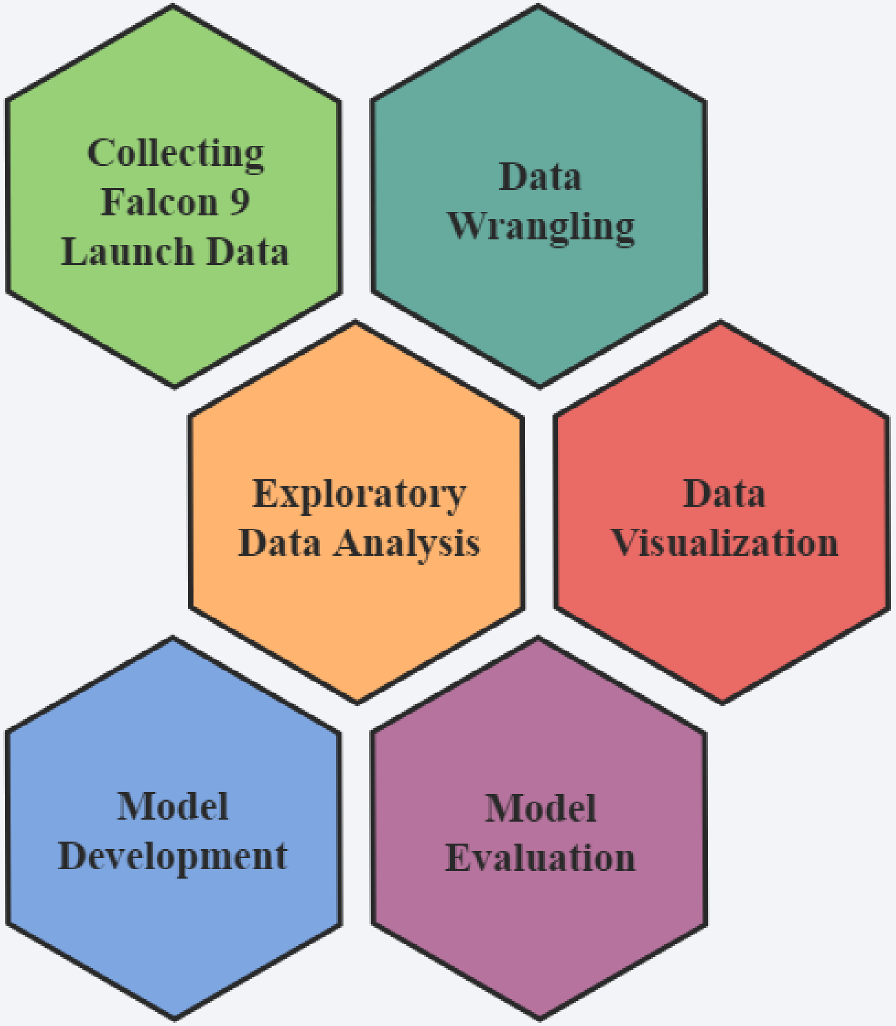
Executive Summary

Summary of Methodologies

The diagram outlines the phases of the Data Science Methodology applied in this project. All the data sets used were collected from public sources.

Summary of All Results

- ✓ Exploratory Data Analysis Results
- ✓ Interactive Analytics Dashboard
- ✓ Predictive Analysis Results



Introduction

- **Project Background and Context**

SpaceX is holding a remarkable position in the commercial space industry as a result of the low cost of launching a rocket. According to the company, the standard payment plan for a Falcon 9 launch is $62 million; while the cost of other providers exceeds $165 million. Much of the savings is because SpaceX can reuse the rocket's first stage.

- **Problems Analyzed**

✓ What determines whether a rocket will land successfully?

✓ Which variables have the greatest influence on the landing success rate?

✓ Can we estimate the cost of a launch by predicting whether the first stage will land?

Section 1

Methodology

Methodology

- Data collection methodology

  Data source: SpaceX REST API – HTTP Requests using Python Requests Library.

  Data source: Wikipedia Article – Web scraping with Beautiful Soup Library.

- Perform data wrangling

  Dealing with missing values, one hot encoding, type cast, transform and cleaning features.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

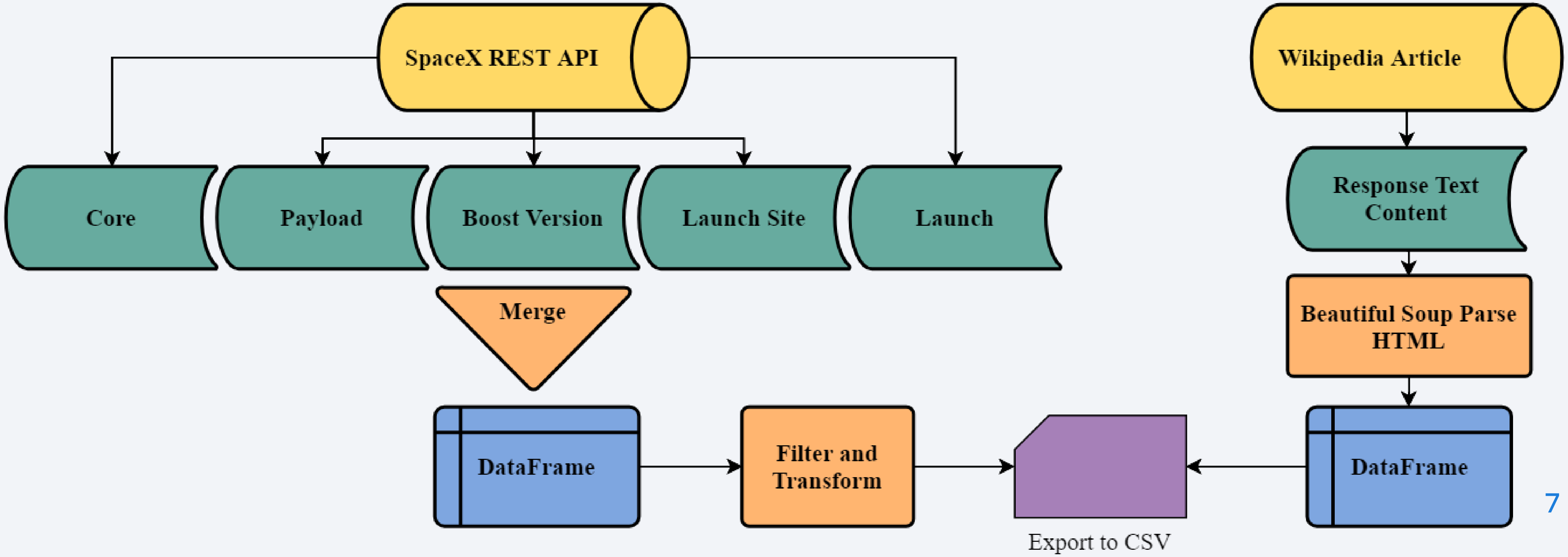
  Data standardization and data set splitting.

  Hyperparameter tuning using Scikit-Learn's GridSearchCV.

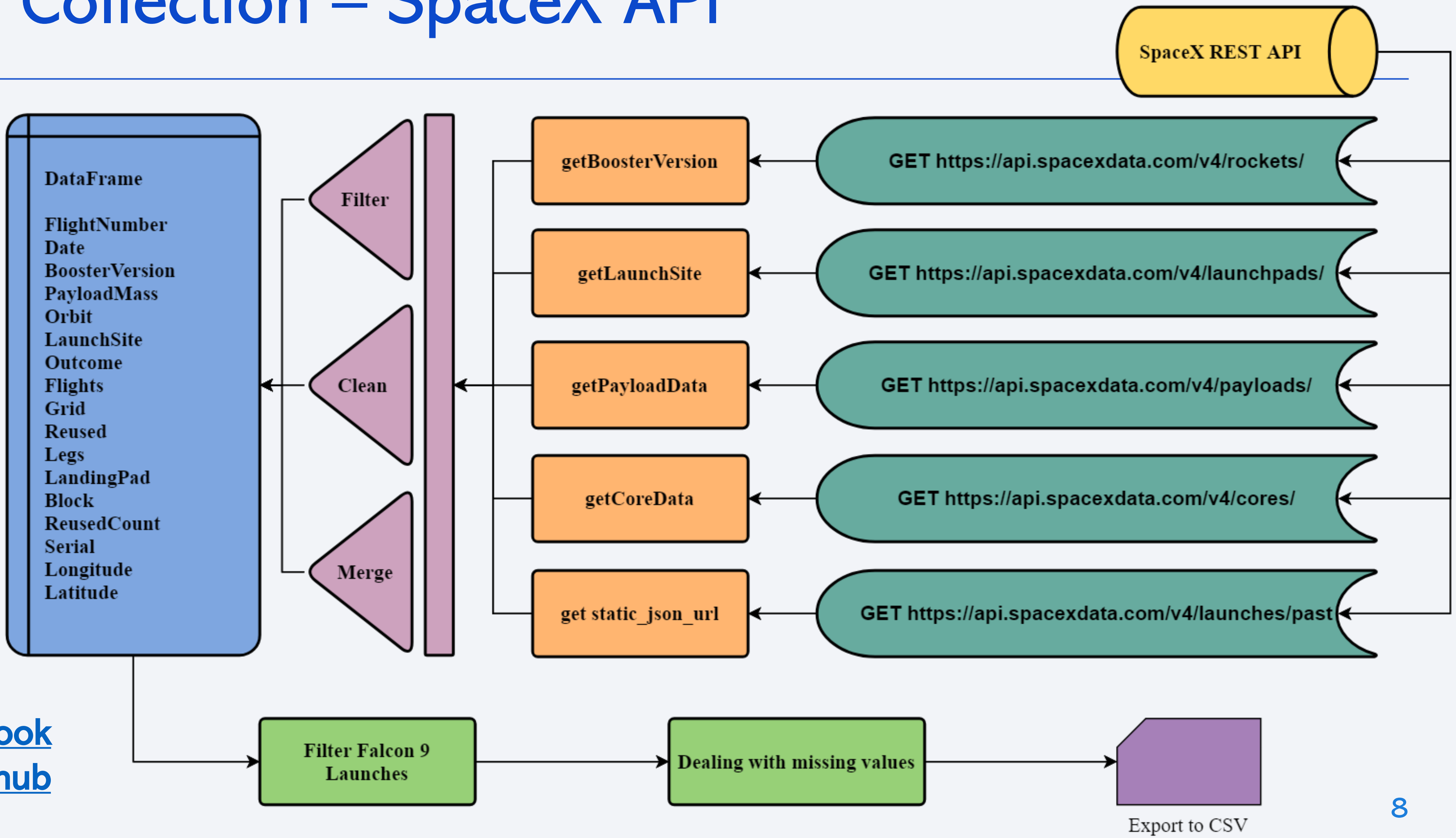
  Evaluation of Logistic Regression, Decision Tree, KNN and SVM classifiers algorithms.

Data Collection

- The data was collected from SpaceX public **REST API**, as shown in the flowchart below.
- The second process shows how data from Wikipedia was collected through **web scraping**.

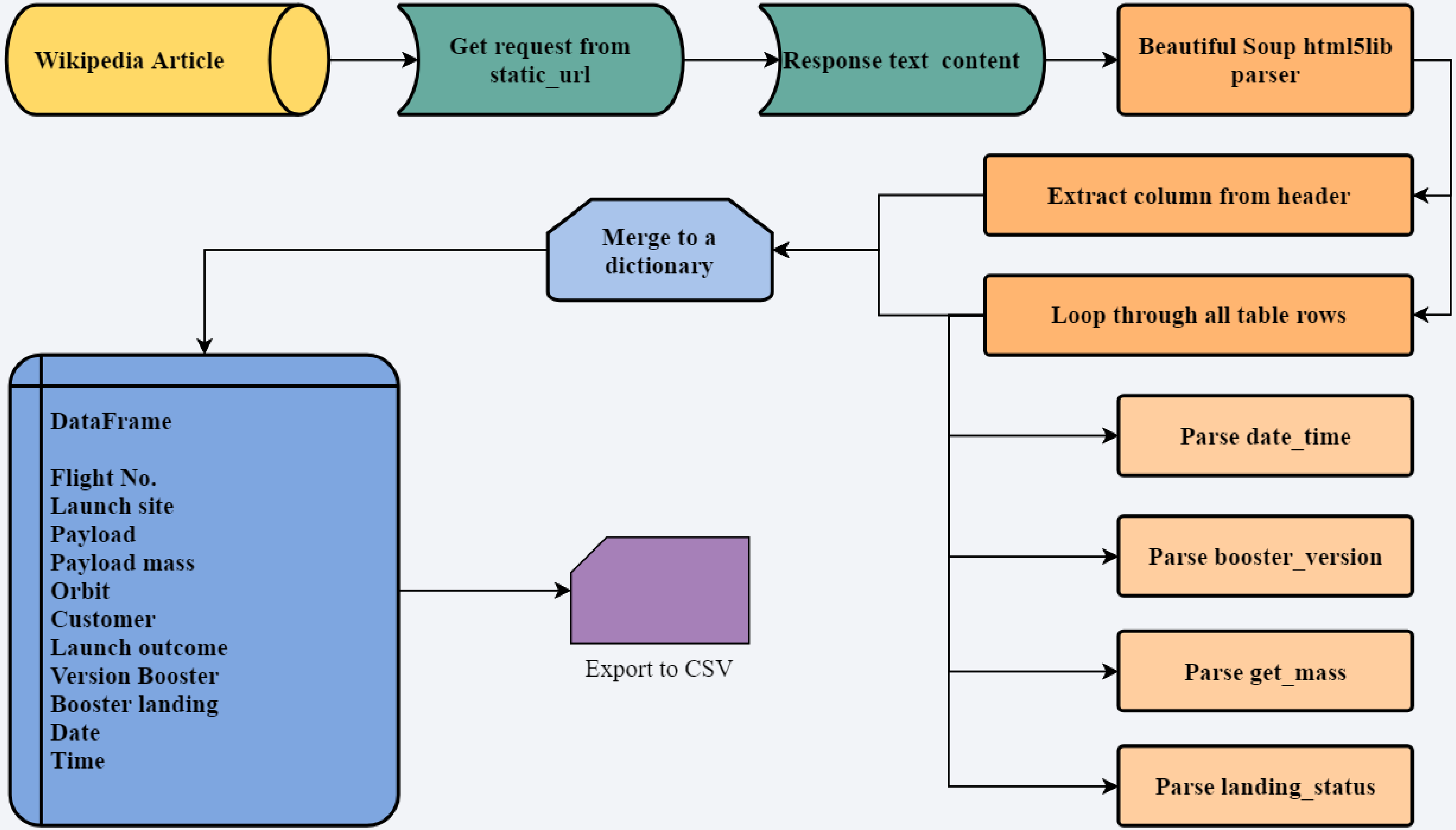


Data Collection – SpaceX API



[Notebook on Github]()

Data Collection – Web Scraping



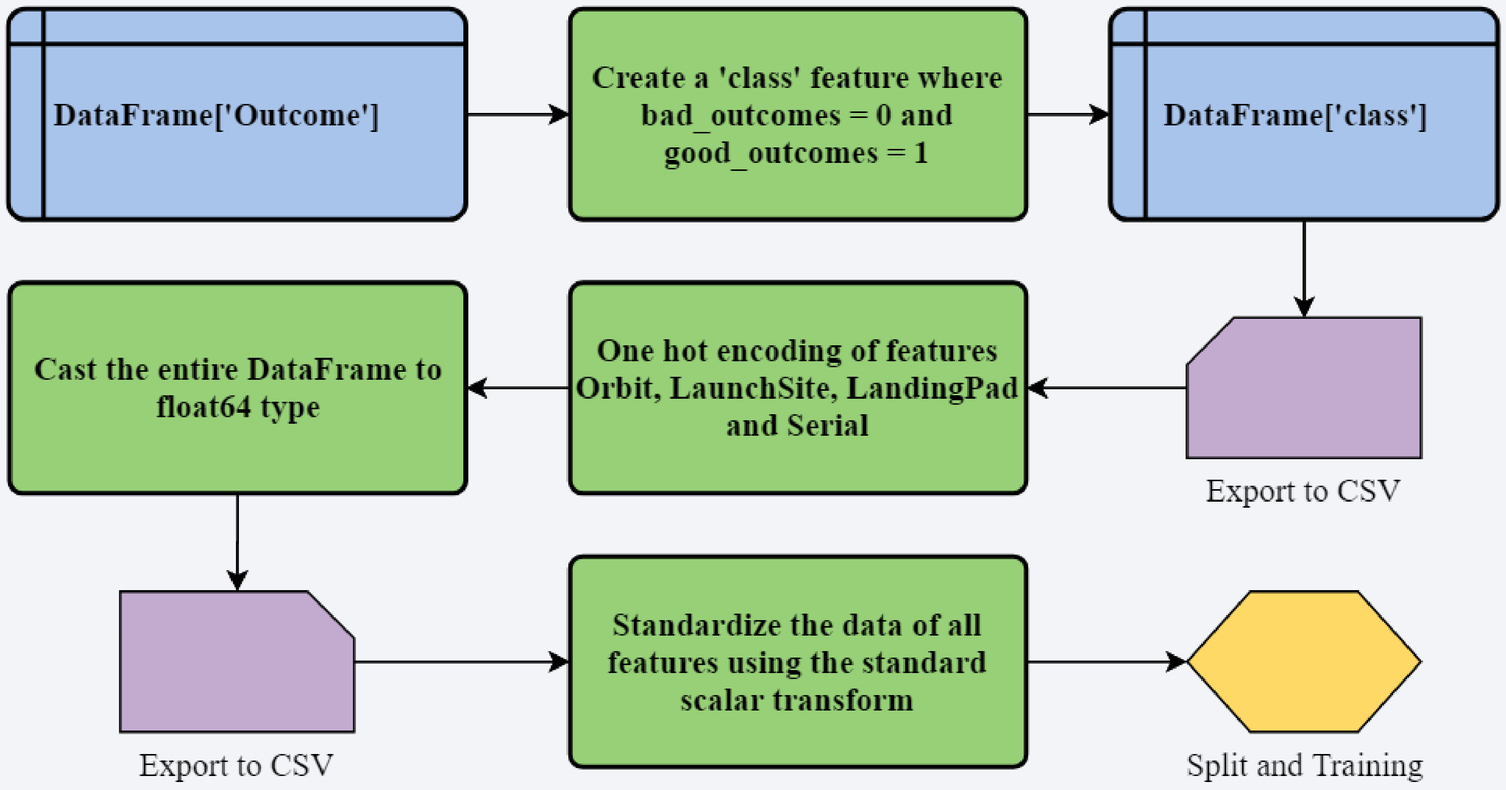
[Notebook on Github]

Data Wrangling

- ✓ The first step was to create the 'class' feature from 'Outcome' as shown in the Notebook.

- ✓ Than we performed one hot encoding of features and the type cast of the entire DataFrame to float as demonstrated in the Notebook.

- ✓ Finally, the data of all features was standardized using a transform as presented in the Notebook.



EDA with Data Visualization

➢ Scatter Chart

Flight Number vs Payload Mass

Flight Number vs Launch Site

Payload Mass vs Launch Site

Orbit vs Flight Number

Payload Mass vs Orbit Type

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. Scatter plots are used to observe relationships between variables.

[Notebook on Github]

➢ Bar Chart

Orbit vs Success Rate

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

➢ Line Chart

Success Rate vs Year

A line chart is a visual comparison of how two variables— shown on the x- and y- axes —are related or vary with each other. It shows related information by drawing a continuous line between all the points on a grid.

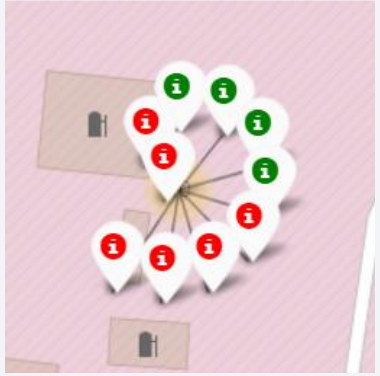
EDA with SQL

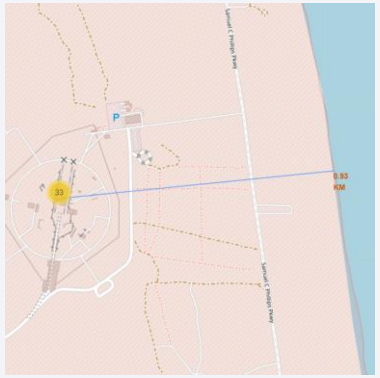
Using SQL queries to answer questions about the data set.

[Notebook on Github]

- ✓ Display the names of the unique launch sites in the space mission.
- ✓ Display 5 records where launch sites begin with the string 'CCA'.
- ✓ Display the total payload mass carried by boosters launched by NASA (CRS).
- ✓ Display average payload mass carried by booster version F9 v1.1.
- ✓ List the date when the first successful landing outcome in ground pad was achieved.
- ✓ List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- ✓ List the total number of successful and failure mission outcomes.
- ✓ List the names of the booster_versions which have carried the maximum payload mass.
- ✓ List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- ✓ Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium





- ✓ We marked NASA and all launch sites (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E) on a **Map** using a **Circle** with a **Popup** and a **Marker** object to analyze their locations in relation to each other.

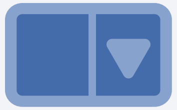
- ✓ Then we added the launch outcomes for each site to identify which location have high success rates. For each SpaceX launch we created a **Marker** with a green **Icon** for success and red **Icon** for failure. The markers where added to a **MarkerCluster** object.

- ✓ Finally, we used the **MousePosition** object to get the coordinates of several points of interests (such as a railway, highway, city and coastline). The distance from the point to the launch site was calculated, a **Polyline** was draw and added to the **Map**.

[Notebook on Github] [Notebook on NbViewer] * GitHub blocks folium content

Build a Dashboard with Plotly Dash

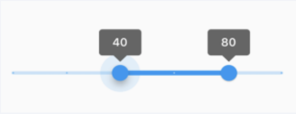
The dashboard was build using the Dash framework and deployed on Google Cloud Platform. You can review the live application here and the source code on Github.



The first interactive component is a **Dropdown** list, where you can select all sites or any specific launch site to plot the graphs for.

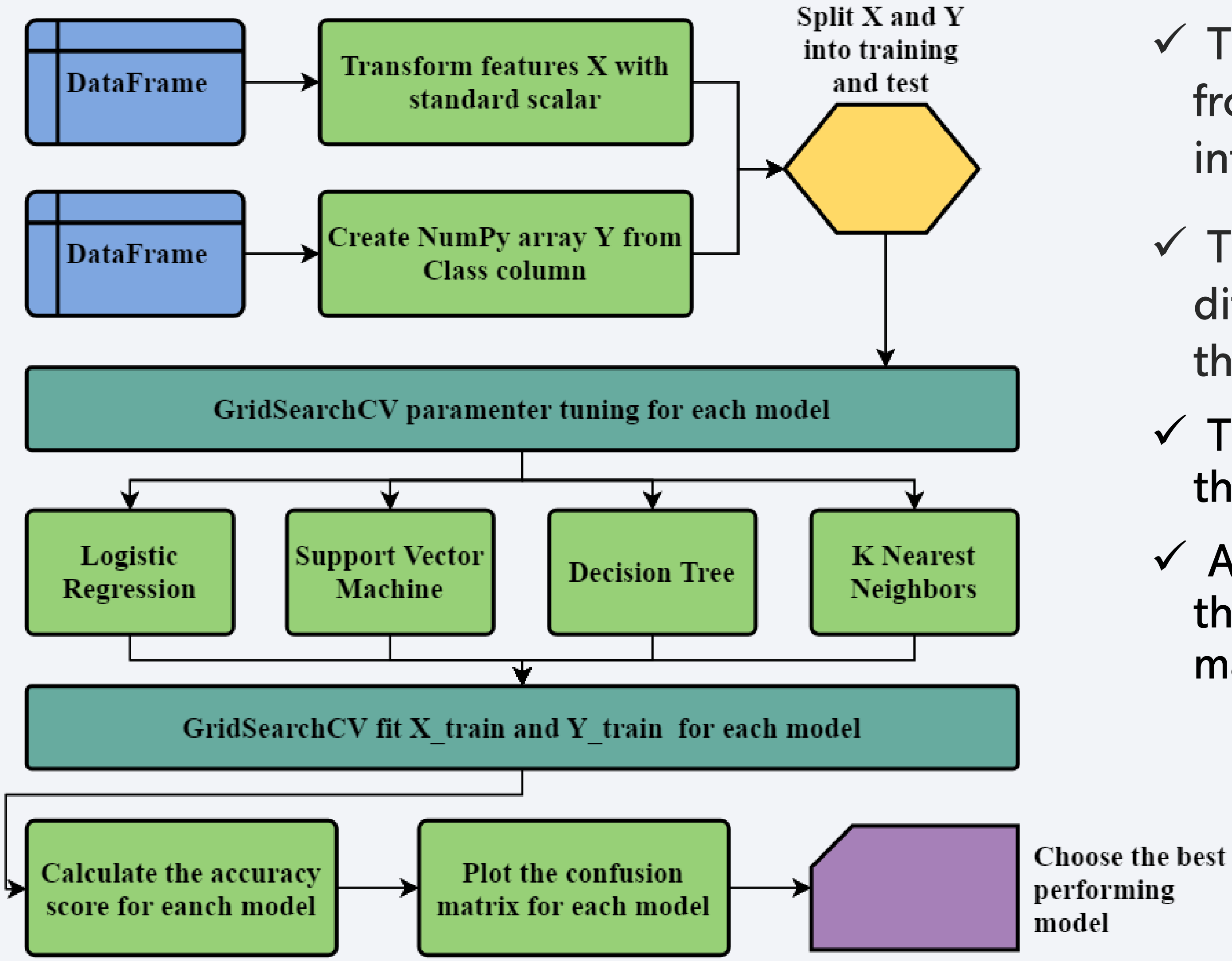
A **Pie Chart** is shown according to the selected option. Each slice on this graph represents the total of **successful launches per site** or the success / failure ratio.

Then a **Scatter Plot** is displayed, where the **launch outcome** in relation to the **payload mass** can be analyzed. Each point on this graph is colored according to the **booster version** label on the right side.



Finally, you can switch the **RangeSlider** on top of the Scatter Plot to analyze the results for different payload mass values.

Predictive Analysis (Classification)



- ✓ The data sets were loaded from csv files, transformed and split into training and test.
- ✓ The model was built applying four different classification algorithms from the Scikit-learn library.
- ✓ The GridSearchCV was used to select the best parameters for each model.
- ✓ All algorithms were evaluated using the accuracy score and the confusion matrix.

[Notebook on Github]

Results

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

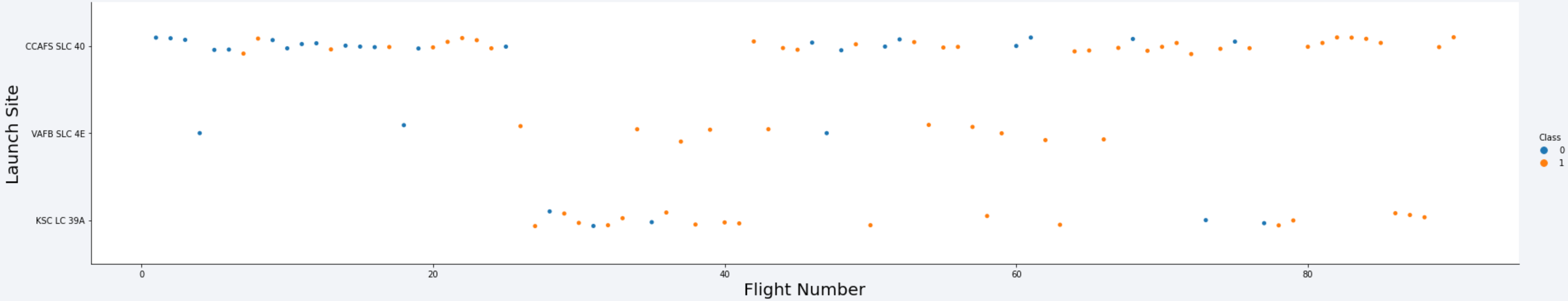
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

According to the Scatter Plot below, it is possible to notice that the success rate of a launch increased with time.

Furthermore, it is evident that both CCAFS SLC 40 and KSC LC 39A accumulate more launches and have similar success rates.

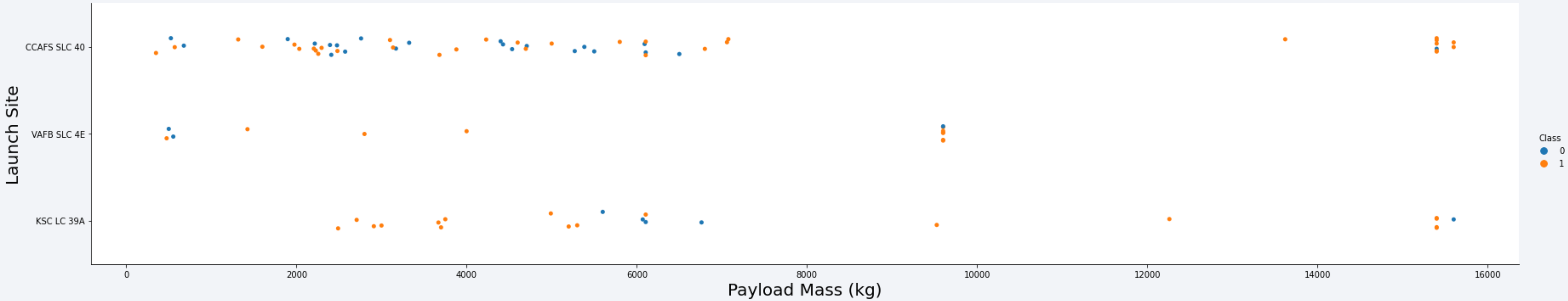


[Notebook on Github]

Payload vs. Launch Site

The graph shows that the success rate for all launch sites are higher when the payload mass is greater than 8000 kg. However, this does not mean that mass is the only variable that establishes a causality.

The majority of successful launches were carrying a payload mass between 2000 kg and 6000 kg.

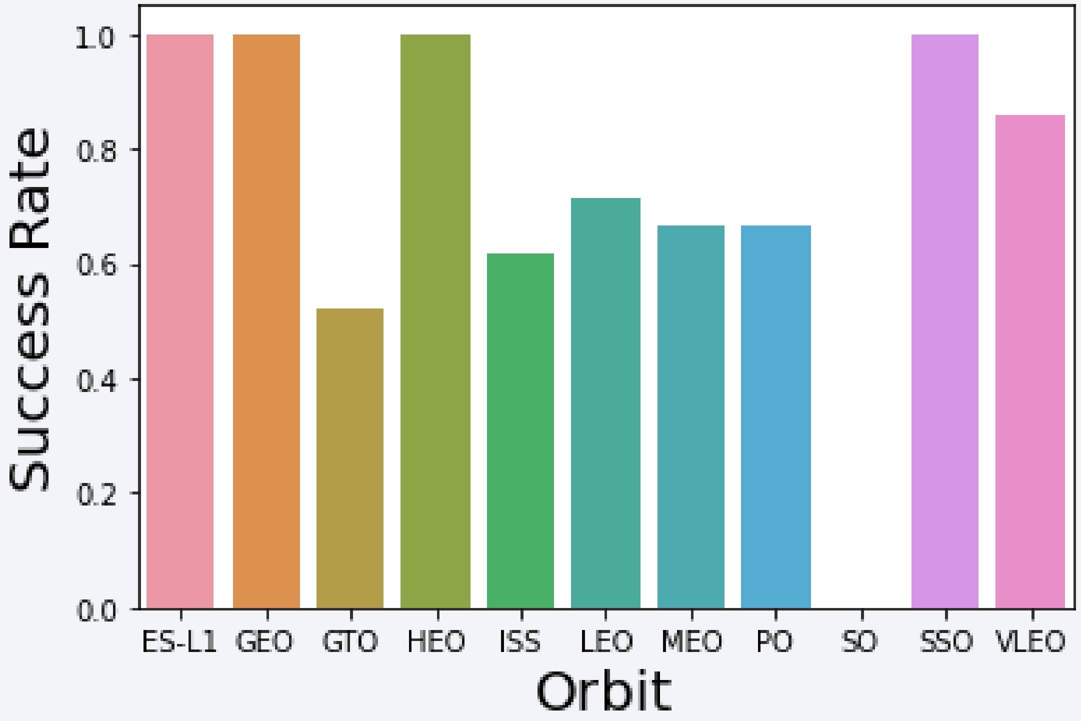


[Notebook on Github]

Success Rate vs. Orbit Type

The Bar Chart shows that the maximum success rate was settled by the ES-L1, GEO, HEO and SSO orbits.

The GTO orbit hit the lower success mark, while the others score around 70% in average.

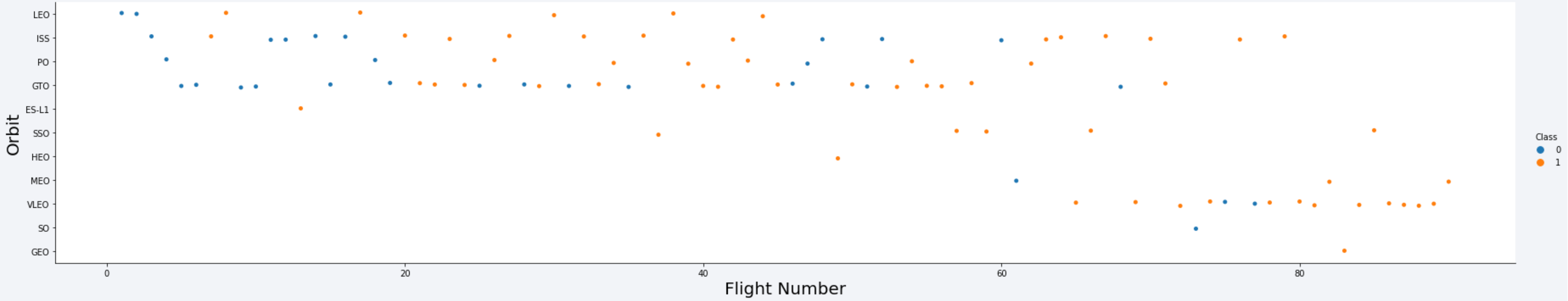


[Notebook on Github]

Flight Number vs. Orbit Type

In this graph it is possible to notice that the SSO orbit hit the 100% success rate with five launches. All the others have only one flight.

In addition, it is clear that most of the launches were destined to the ISS, PO, GTO and VLEO orbits. These four orbits together have approximately 65% average success rate.

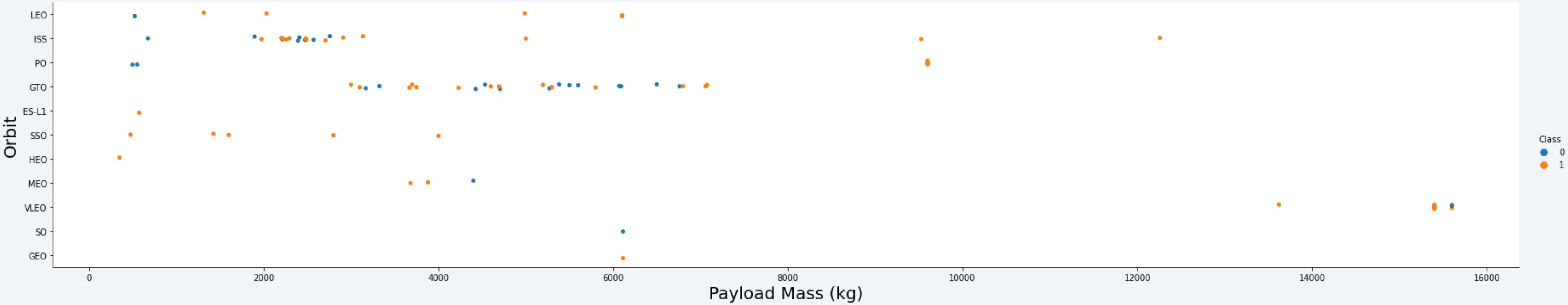


[Notebook on Github]

Payload vs. Orbit Type

The analysis of the graph allows us to observe that the success rate is near 100% for ISS, PO and VLEO orbits when the payload mass exceeds 8000 kg.

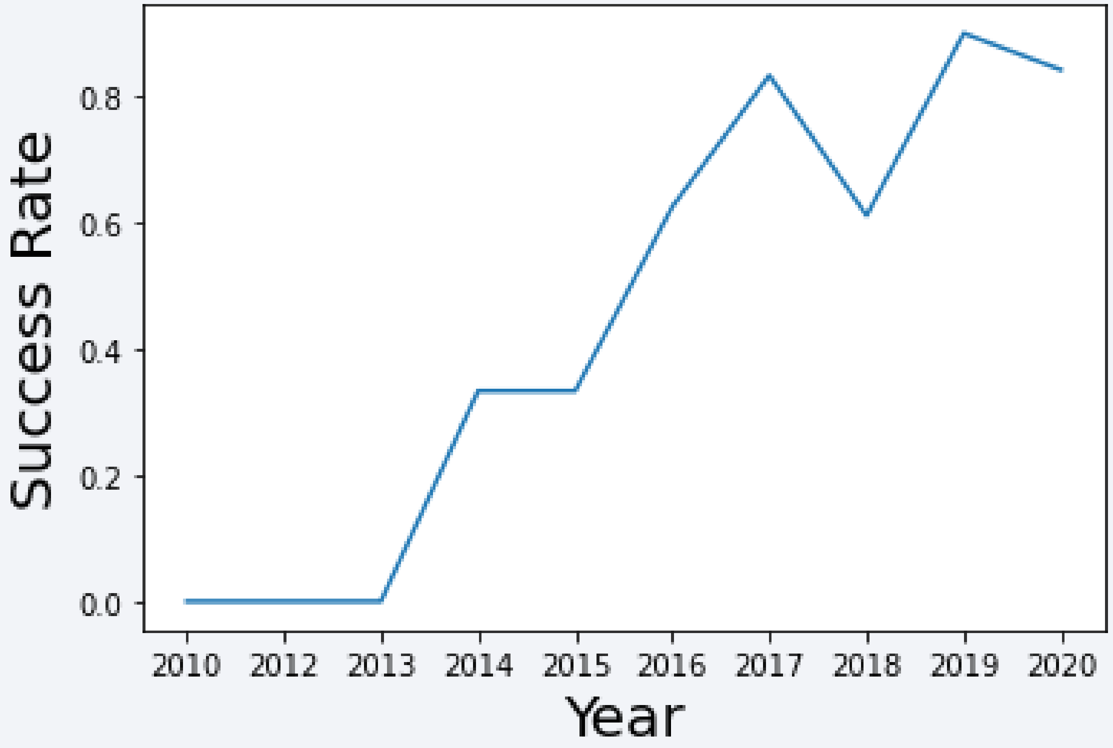
Most successful launches were carrying a payload between 2000 kg and 6000 kg. In the graph you notice a greater concentration of points in the ISS and GTO orbits.



Notebook on Github

Launch Success Yearly Trend

The success rate of Falcon 9 launches has grown continuously over the period 2013-2020.



[Notebook on Github]

All Launch Site Names

The **DISTINCT** statement was used to return only unique values from the launch_site column.

```
select distinct LAUNCH_SITE from SPACEXTBL;
```

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

Notebook on Github

Launch Site Names Begin with 'CCA'

The query statement uses the **wildcard %** after CCA meaning that only the values beginning with CCA will match the condition of the **WHERE** clause. The number of rows returned will be **limit**ed by 5.

```
select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' LIMIT 5;
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Notebook on Github

Total Payload Mass

The query statement uses the built-in function **SUM** to calculate to total payload mass for the customer NASA (CRS) filtered by the **WHERE** clause.

```
select sum(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS from SPACEXTBL
    where CUSTOMER = 'NASA (CRS)';
```

total_payload_mass
45596

[Notebook on Github]

Average Payload Mass by F9 v1.1

The query statement uses the built-in function **AVG** to calculate the average payload mass for flights with version F9 v1.1 of the booster, filtered by the **WHERE** clause.

```
select avg(PAYLOAD_MASS__KG_) as AVERAGE_PAYLOAD_MASS from SPACEXTBL
    where BOOSTER_VERSION = 'F9 v1.1';
```

average_payload_mass
2928

Notebook on Github

First Successful Ground Landing Date

The query statement uses the **wildcard %** after Success, meaning that only the values beginning with Success will match the condition of the **WHERE** clause. The built-in function **MIN** is used to return only the lowest date value, which represents the first occurrence of a success.

```
select min(DATE) as FIRST_SUCCESSFUL_LANDING from SPACEXTBL
     where LANDING__OUTCOME like 'Success%';
```

first_successful_landing
2015-12-22

[Notebook on Github]

Successful Drone Ship Landing with Payload between 4000 and 6000

The **DISTINCT** statement was used to return only unique values from the booster_version column. The **WHERE** clause was used to filter records that fulfill the following condition: boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

```
select distinct BOOSTER_VERSION from SPACEXTBL
    where LANDING__OUTCOME = 'Success (drone ship)' and
    PAYLOAD_MASS__KG_ < 6000 and PAYLOAD_MASS__KG_ > 4000;
```

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

Notebook on Github

Total Number of Successful and Failure Mission Outcomes

The **GROUP BY** statement groups rows that have the same values for the mission outcome. Then the built-in function **COUNT** is used to return the number of rows for each category (or group) of mission outcome.

```
select MISSION_OUTCOME, count(MISSION_OUTCOME) as MISSION_COUNT
    from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

[Notebook on Github]

Boosters Carried Maximum Payload

This query uses a **subquery** to set the condition for the **WHERE** clause. The built-in function **MAX** is used to find the maximum value of the payload mass. Then the value returned by the subquery is used to filter the values of the booster version.

```
select BOOSTER_VERSION, PAYLOAD_MASS__KG_
    from SPACEXTBL
    where PAYLOAD_MASS__KG_=
        (select max(PAYLOAD_MASS__KG_)
        from SPACEXTBL);
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

The **WHERE** clause in this statement is used to filter by two conditions. The first is a failed landing outcome in drone ship. The second uses the built-in function **YEAR** to match only the dates where the year is 2015.

```
select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE, DATE
    from SPACEXTBL
    where LANDING__OUTCOME = 'Failure (drone ship)' and year(DATE) = 2015;
```

landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

[Notebook on Github]

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

This query uses the **group by** statement and the built-in function **COUNT** to calculate the number of landing outcomes. The **WHERE** clause is used to filter by date and the **ORDER BY** keyword sorts the result set in descending order.

```
select LANDING__OUTCOME,
count(LANDING__OUTCOME) as LANDING_COUNT
from SPACEXTBL
where DATE between '2010-06-04' and '2017-03-20'
    GROUP BY LANDING__OUTCOME
    ORDER BY LANDING_COUNT DESC;
```

landing__outcome	landing_count
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

[Notebook on Github]

Section 4

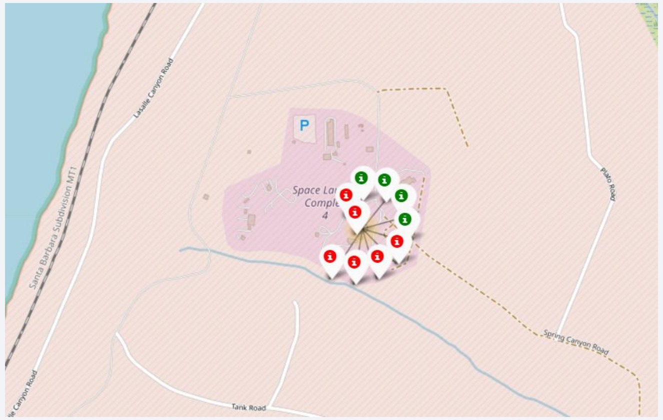
Launch Sites Proximities Analysis

SpaceX Launch Sites Locations

- ✓ All launch sites are located in the United States of America.
- ✓ Their latitude is just above the Tropic of Cancer.
- ✓ All four space complexes are based on the coastline of California and Florida.

[Notebook on NbViewer]

Marks of Success and Failure Launches

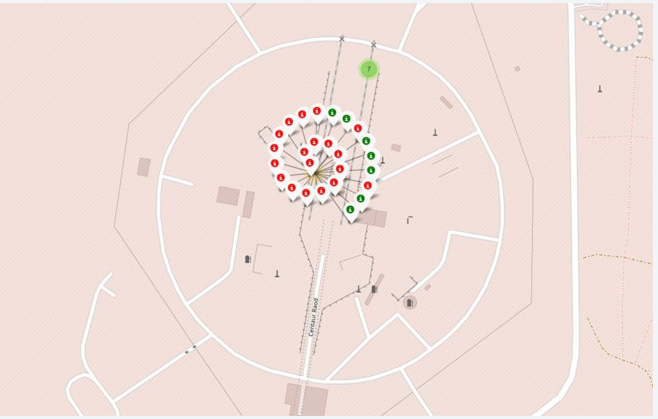


CA VAFB SLC-4E

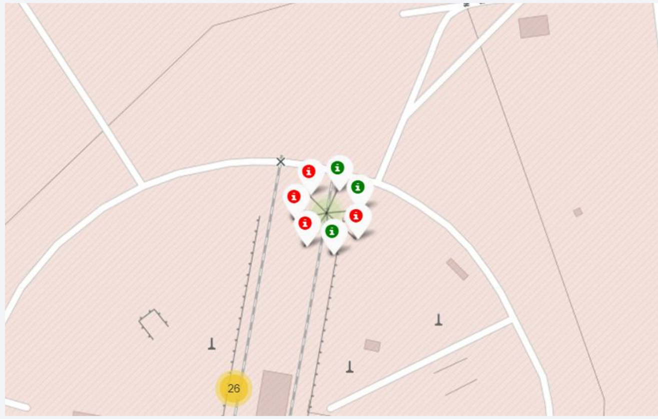
The marks on the map indicate that launch success rate is higher for the Kennedy Space Center (KSC LC-39A).

Success Launch Failure Launch

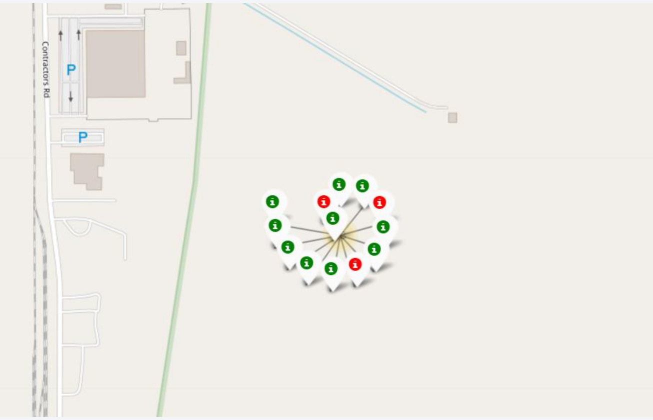
[Notebook on NbViewer]



FL CCAFS LC-40



FL CCAFS SLC-40



FL KSC LC-39A

Distance from the Launch Complex CCAFS LC-40

0.93 Km Coastline

The Cape Canaveral Space Launch Complex is extremely close to the shore and is located at a reasonable distance from the nearest city and the nearest highway.

However, it is relatively far from the Sand Lake Road train station.

[Notebook on NbViewer]

17.9 Km City of Cape Canaveral

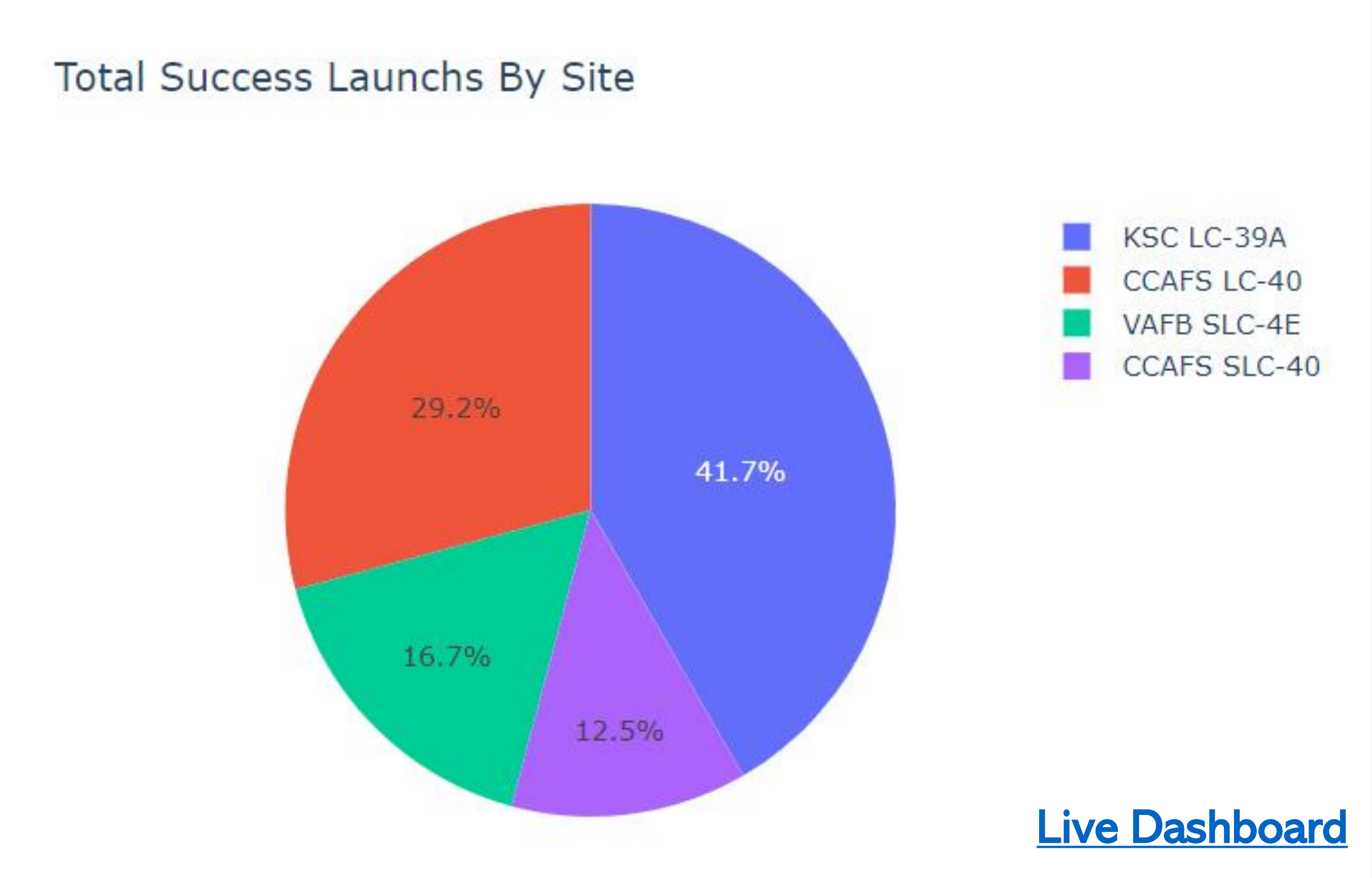
21.07 Km US Highway 1

78.13 Km Railway

Section 5

Build a Dashboard with Plotly Dash

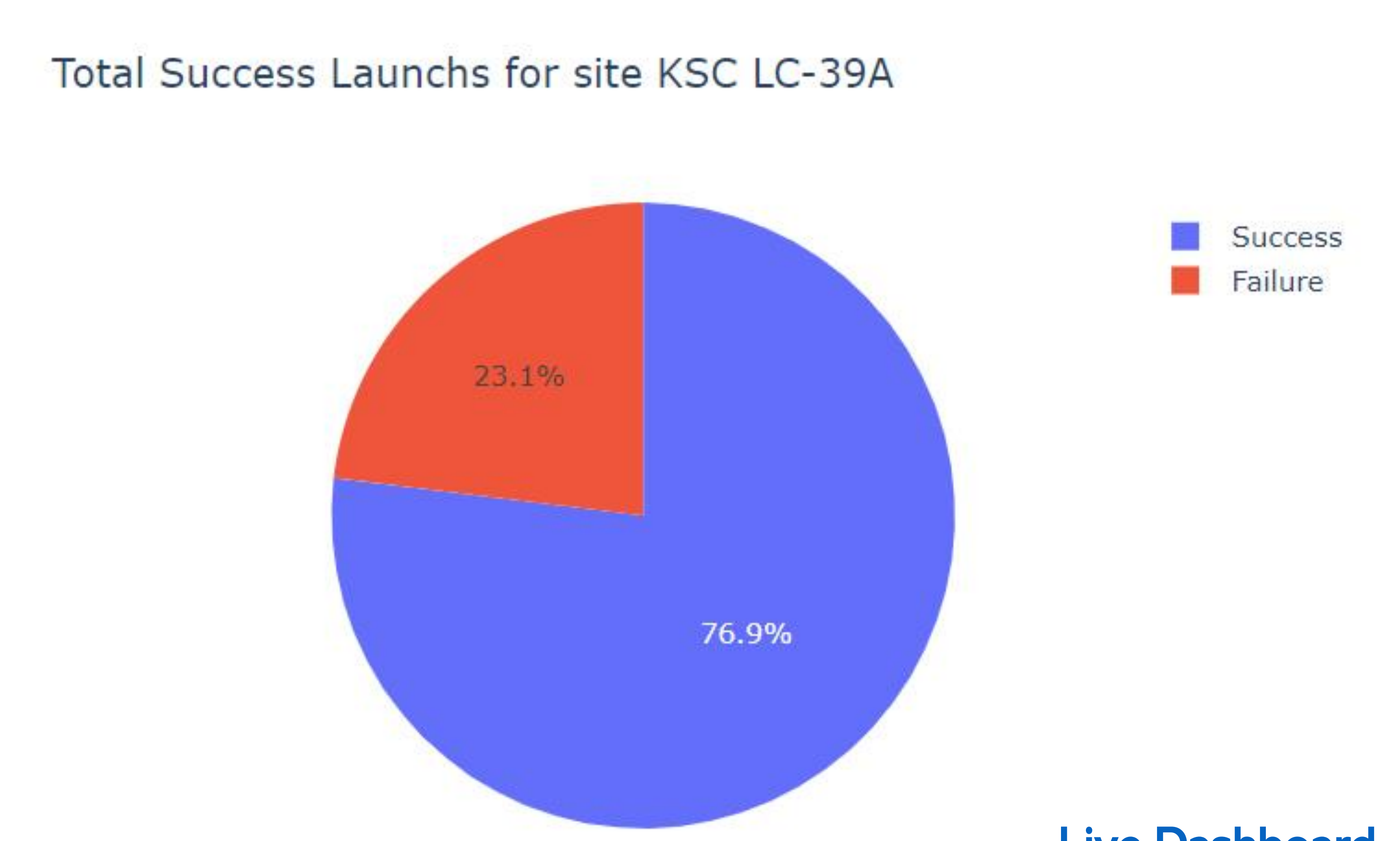
Total Success Launches by Site



Live Dashboard

- ✓ The pie chart validates previous information, gained from the map markers, that the KSC launch complex has the highest success rate.
- ✓ It also indicates a trend, when compared to the number of markers on the map, of a linear relationship between success rate and the number of flights per site.

Total Success Launches for Kennedy Space Center



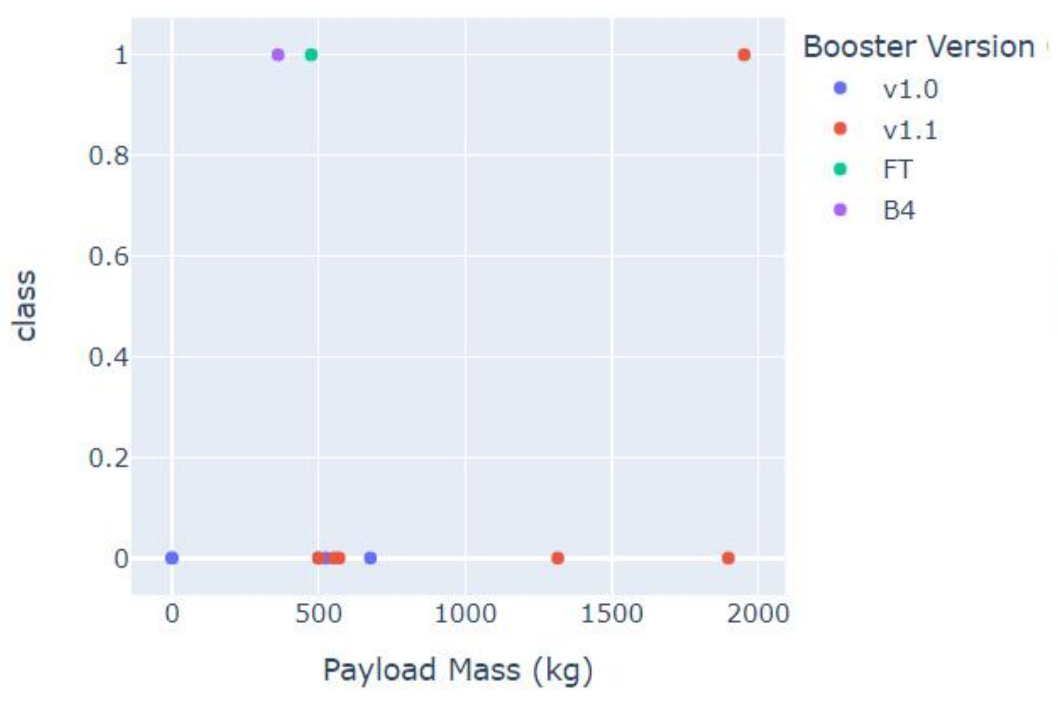
Live Dashboard

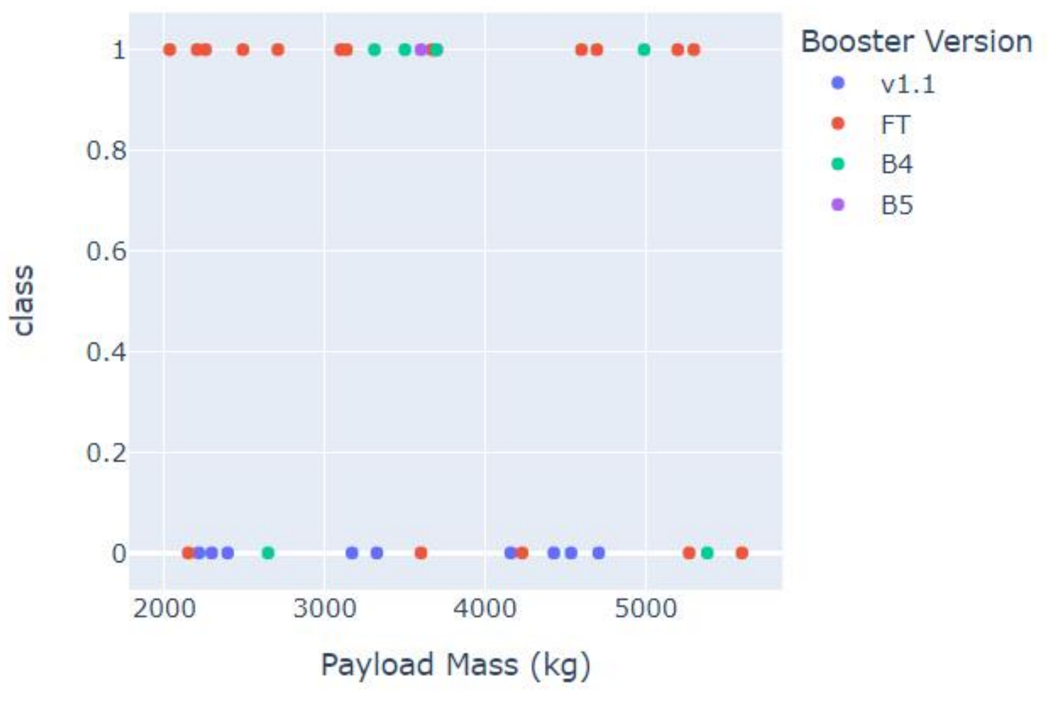
- ✓ Despite having the highest number of successful launches and the highest success rate, there is no clear evidence that the location is the only reason for the outcome.

Launch Outcome by Payload Mass

- ✓ In the first range of payload mass, lower than 2000 kg, there is a higher failure rate and the booster version v1.1 is the most used.
- ✓ In the second range of payload mass, between 2000-6000 kg, the success rate is approximately 52% and the booster version FT is the most used and successful.

[Live Dashboard]

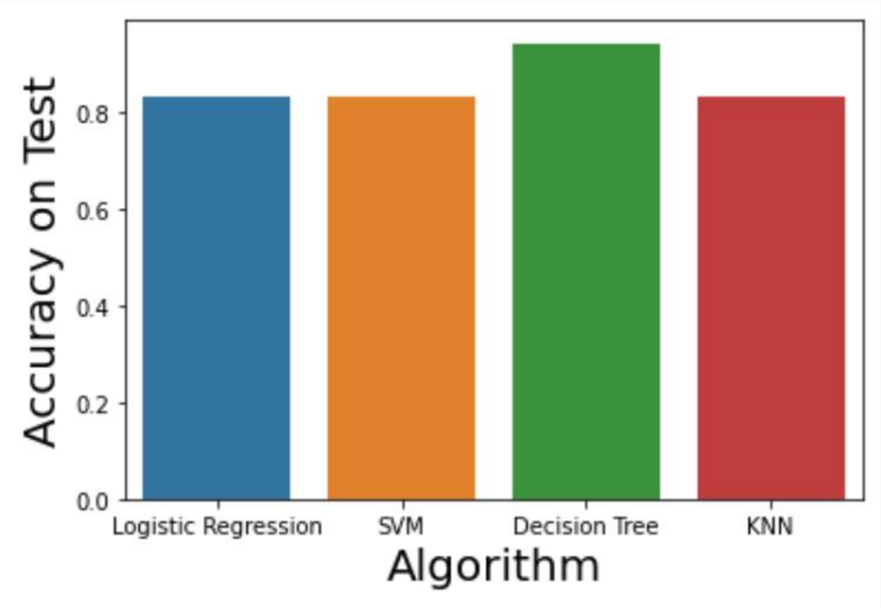




Section 6

Predictive Analysis (Classification)

Classification Accuracy



- ✓ The accuracy scores indicate that the performance improvement brought by the GridSearchCV resulted in identical values for Logistic Regression, SVM and KNN algorithms.
- ✓ Finally, we have a clear winner that performed better on both training and test scores. The **Decision Tree** algorithm satisfies the requirements of this project.

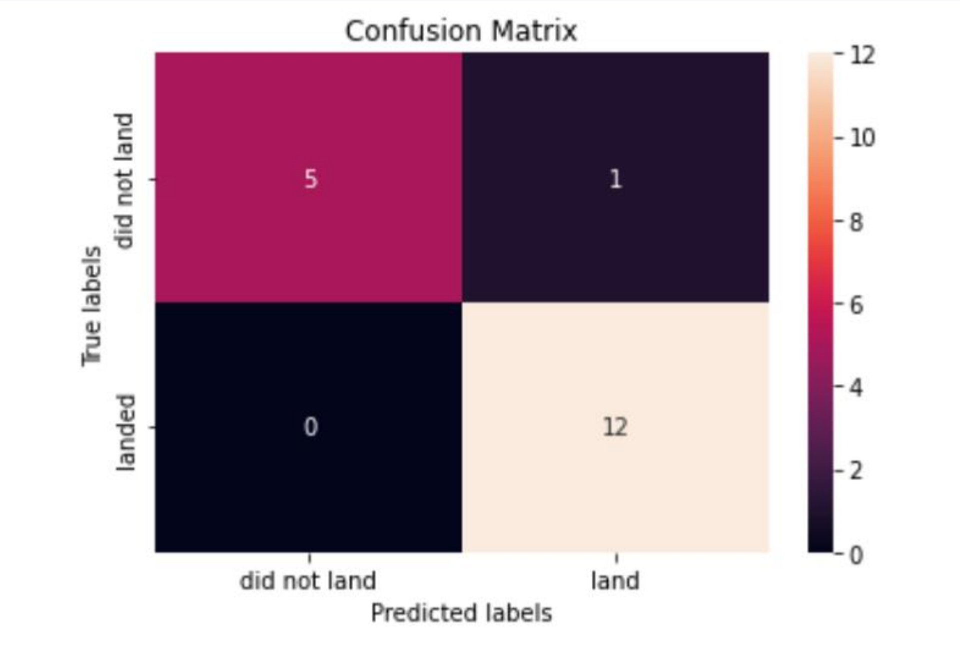
	Algorithm	Best Score on Training	Accuracy on Test	Confusion Matrix	Performance
0	Logistic Regression	0.846429	0.833333	0-3-3-12	High
1	SVM	0.848214	0.833333	0-3-3-12	High
2	Decision Tree	0.891071	0.944444	0-5-1-12	Better
3	KNN	0.848214	0.833333	0-3-3-12	High

Notebook on Github

Confusion Matrix for the Decision Tree

- ✓ The confusion matrix shows that, given a test set of 18 records, we got 17 correct predictions and only 1 incorrect.
- ✓ The incorrect prediction is a false positive, also known as a type I error, false alarm or overestimation.

[Notebook on Github]



True negative Predicted negative Actual negative	**False positive** Predicted positive Actual negative
False negative Predicted negative Actual positive	**True positive** Predicted positive Actual positive

Conclusions

What determines whether a rocket will land successfully?

- ✓ The most important factor is time. The success rate of Falcon 9 launches has grown continuously over the period 2013-2020.

Which variables have the greatest influence on the landing success rate?

- ✓ The Kennedy Space Center (KSC LC-39A) has the highest success rate of 76.9%.
- ✓ A payload mass between 2000 kg and 6000 kg is most likely to give a positive outcome.
- ✓ Most of the launches were destined to the ISS, PO, GTO and VLEO. These four orbits together have approximately 65% average success rate.
- ✓ The new booster versions contributed to a significant improvement in the success rate.

Can we estimate the cost of a launch by predicting whether the first stage will land?

- ✓ According to the predictive analysis results it is possible, with a satisfactory confidence margin, to estimate costs based on launch data.

Appendix – Resources, Tools and Services

❖ Python Libraries

Matplotlib

Seaborn

Pandas

NumPy

Plotly Dash

Scikit-learn

❖ Tools

Visual Studio Code

MS PowerPoint

Adobe Photoshop

❖ Services

IBM Cloud

GitHub

Google App Engine

Thank you!