
# Winning Space Race with Data Science

Salah Y. Al-Kafrawi

2021/9/9

# Outline

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

# Executive Summary

- SpaceX advertises Falcon9 rocket launches with a cost of 62 million dollars, whereas other providers charge up to 165 million dollars. It managed to spare over 100 million dollars mainly due to its capability to reuse the bottom part of the rocket, i.e., the first stage reenters the atmosphere and lands on earth.

- If an alternate company wanted to bid against SpaceX for a rocket launch, this information would be critical in cost prediction analysis. In short, predicting whether the first stage will land or not, will help us in forecasting the launch cost. In this project, two data sets about SpaceX Falcon9 have been collected, wrangled, visualized, and analyzed via classification machine learning algorithms, to determine the attributes that affect the rocket first stage landing.

# Introduction

- The human marathon into space started over 60 years ago when the Soviet Union launched their first satellite Sputnik, on October 4, 1957. Since then, governments have been spending billions of dollars on experiments to compete each other on reaching the outer space. In December 2010, SpaceX managed to be the first private company to return a spacecraft from low-earth orbit. This accomplishment helped it to magnificently reduce launches costs.

- We want to identify the categorical/continuous attributes that impact the possibility of a successful landing of the first stage in Falcon9. Questions such as 'does destination orbit affect the landing stage?' 'What about payload mass or launch site location?' Will be our main concern to answer in this analysis.

Section 1

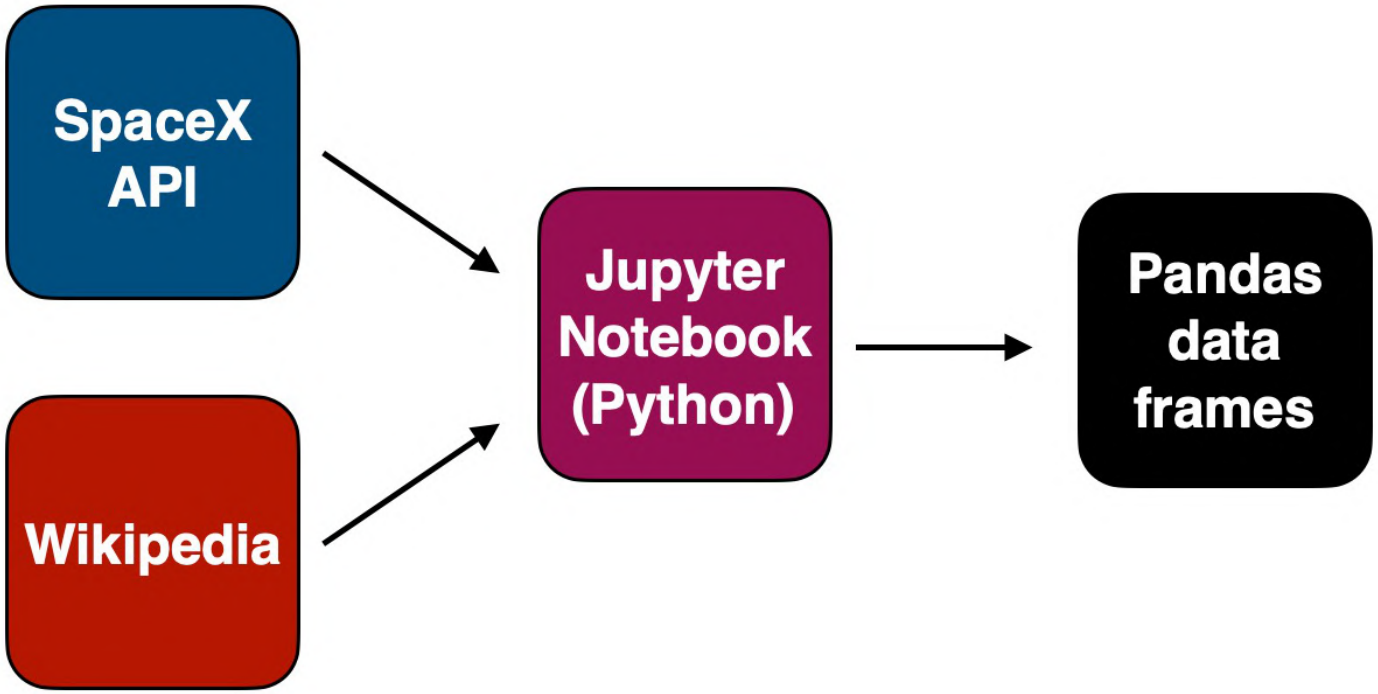
# Methodology

# Methodology Steps

- Executive Summary – Python language, Jupyter Notebook Environment, IBM Cloud
- Data Collection Methodology – SpaceX APIs and Internet Webscraping
- Perform Data Wrangling – Unrelated columns have been removed. One hot encoding was applied.
- Perform Exploratory Data Analysis – (EDA) using visualization and SQL
- Perform Interactive Visual Analytics – Using Folium and Plotly Dash
- Perform Predictive Analysis – Using classification models, optimal parameters were decided via GridSearchCV.

# Data Collection sources

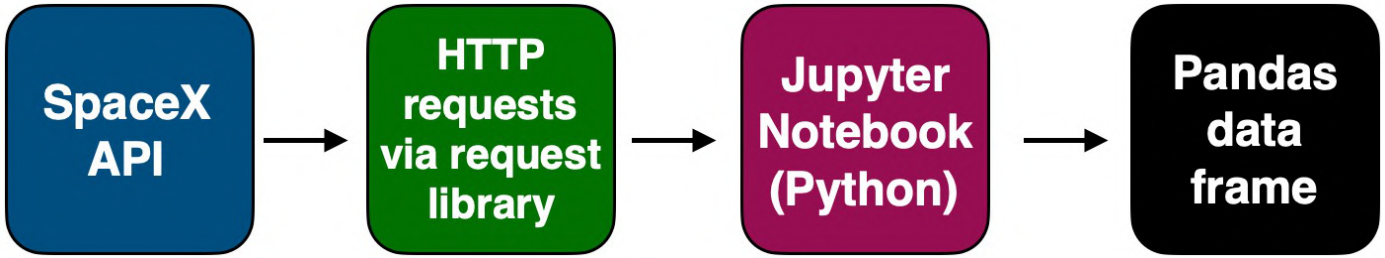
We explored two sources for data collection, SpaceX API, and Falcon9 record on Wikipedia. We took intriguing attributes such as booster name, launchpad location, payload mass, orbit name, landing outcome, number of cores, core serial, core reused count, whether grid fins were used, etc.



# Data Collection – SpaceX API

- First, we requested and parsed the data using GET request. Then, we filtered the data to include only Falcon9 launches. Finally, we dealt with the missing values (Replaced with mean values).

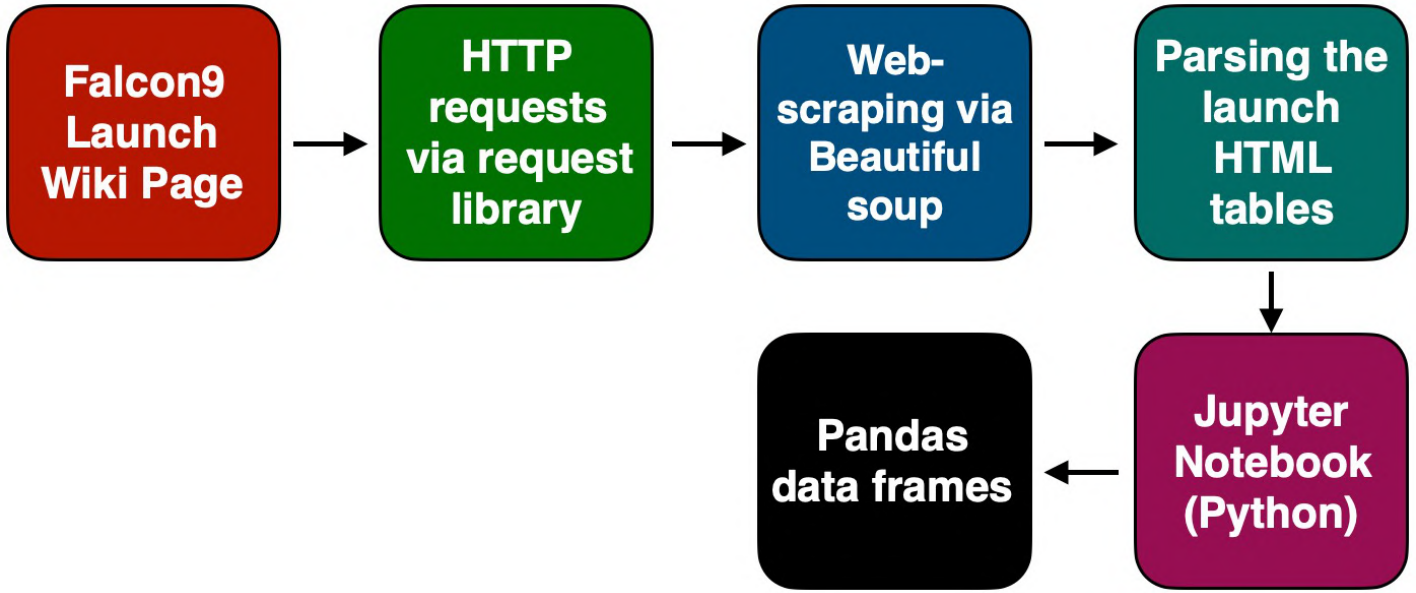
- GitHub URL of the completed SpaceX API calls notebook (Click Here)



# Data Collection – Scraping

- At first, we requested the Falcon9 Launch Wiki page. After that, we used Beautiful Soap library to extract all column/variable names from the HTML table header. Eventually, we created a data frame by parsing the HTML tables.

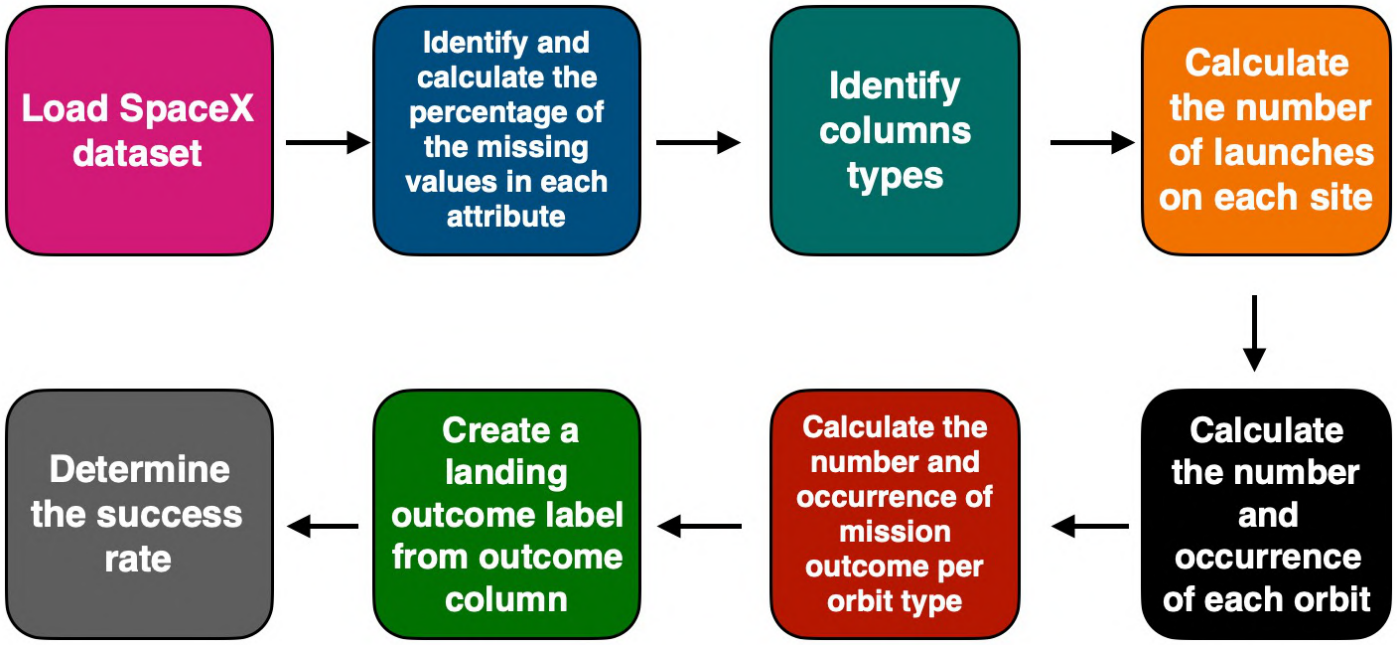
- GitHub URL of the completed SpaceX Webscraping notebook (Click Here)



# Data Wrangling

- We performed Exploratory Data Analysis (EDA) to find patterns in the data, so we can determine what would be the labels for the training supervised models.

- GitHub URL of the completed Data Wrangling notebook (Click Here)



# EDA with SQL

- A list of the SQL inquiries used in the EDA process.

- GitHub URL of the completed EDA with SQL notebook (Click Here)

- **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).**
- **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. Use a subquery.**
- **List the failed landing outcomes in drone ship, their booster versions, and launch site names for the in year 2015.**
- **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.**

# EDA with Data Visualization

- We performed EDA and Feature Engineering using Pandas, Matplotlib, and Seaborn. The list of visualization are as follows,

- GitHub URL of the completed Visualization notebook (Click Here)

- **Scatter plot [(sns.catplot)] between Flight Number (the continuous launch attempts), and Payload Mass to see how would they affect the launch outcome. hue = 'class' i.e., class = 1 implies successful landing, and 0 implies failed landing.**
- **In the same manner, scatter plot between different attributes have been visualized with hue = 'class'. More of these visualizations will be discussed in detail in the EDA section.**

- **Bar chart [plot(kind='bar')] that relates between success rate and orbit, in order to check if there are any relationship between success rate and orbit type.**

- **Line plot [plot(kind = 'line')] that relates success rate and year in order to determine if SpaceX is making a progress or not in successful landing.**

# Build an Interactive Map with Folium

- Folium library was used in EDA to find what factors were considered by SpaceX when selecting a launch site location.

- GitHub URL of the completed Folium maps notebook (Click Here)

- **We drew a circle marker on each of the four sites, which are designated for Falcon9 launching. These locations are as follows,**

  **1 – CCAFS LC-40**
  **2 – CCAFS SLC-40**
  **3 – KSC LC-39A**
  **4 – VAFB SLC-4E**

- **We marked the success/failed launches for each site on the map using [MarkerCluster] object.**

- **We calculated the distances between a launch site (CCAFS LC-40) to its proximities i.e., the closest city, coastline, and highway. Then we drew lines between it to these proximities via [PolyLine] object.**

# Build a Dashboard with Plotly Dash

- An interactive dashboard has been devolved. The dashboard encompasses two figures as follows,

- GitHub URL of the completed Dashboard notebook (Click Here)

- **An interactive Pie Chart that represents a success rate for each of the four sites, namely,**

  **1 – CCAFS LC-40**
  **2 – CCAFS SLC-40**
  **3 – KSC LC-39A**
  **4 – VAFB SLC-4E**
  **5 – All Sites**

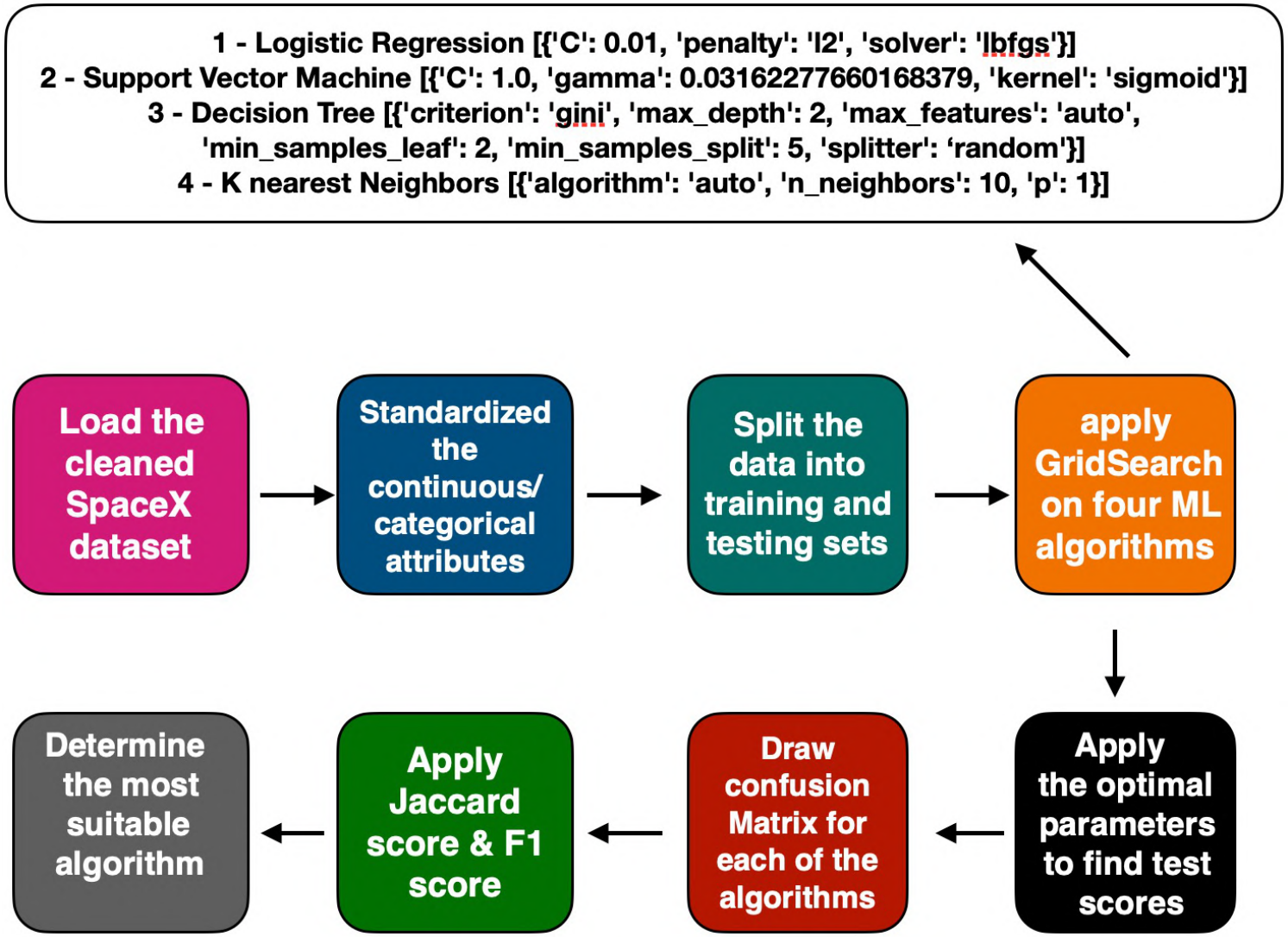
- **An interactive slider to show the outcomes of several payload versions based on their mass.**

# Predictive Analysis (Classification)

- We standardized the cleaned data. Then, we split it into 80% training set and 20% testing set. GridSearchCV was used to find the optimal hyperparameter for each of the following classification algorithms,

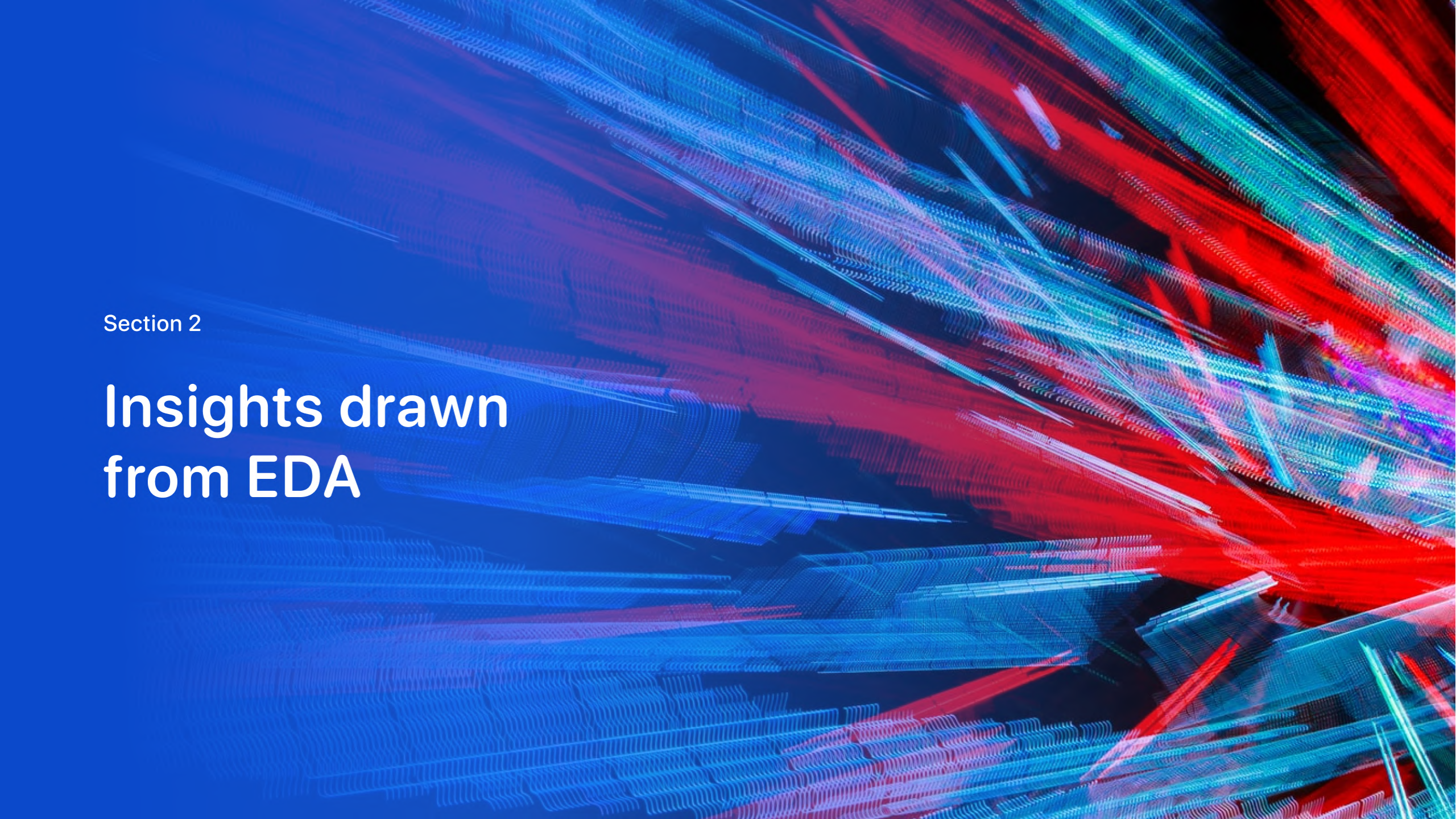
1. Logistic Regression (LR)
2. Support Vector Machine (SVM)
3. Decision Tree (DT)
4. K nearest neighbors (KNN)

- GitHub URL of the completed ML Classification notebook (Click Here)

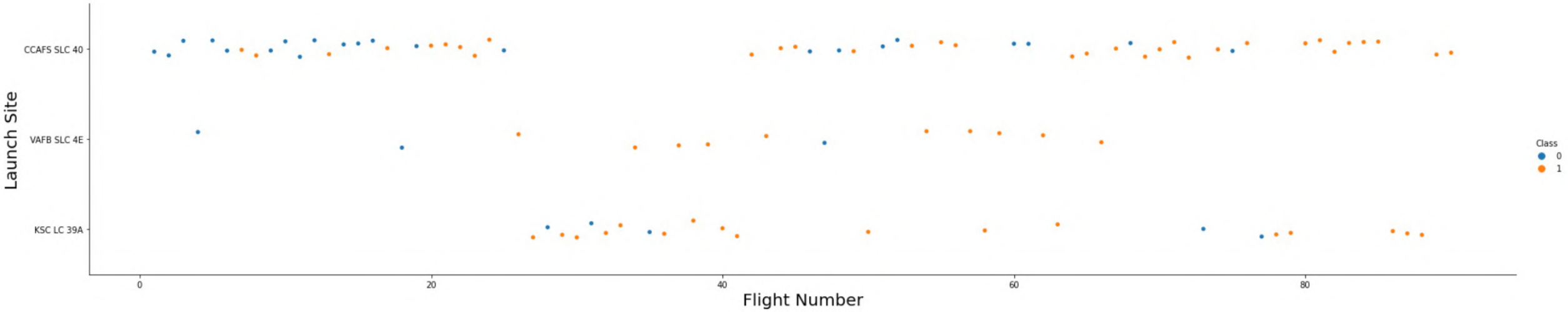


# Results

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

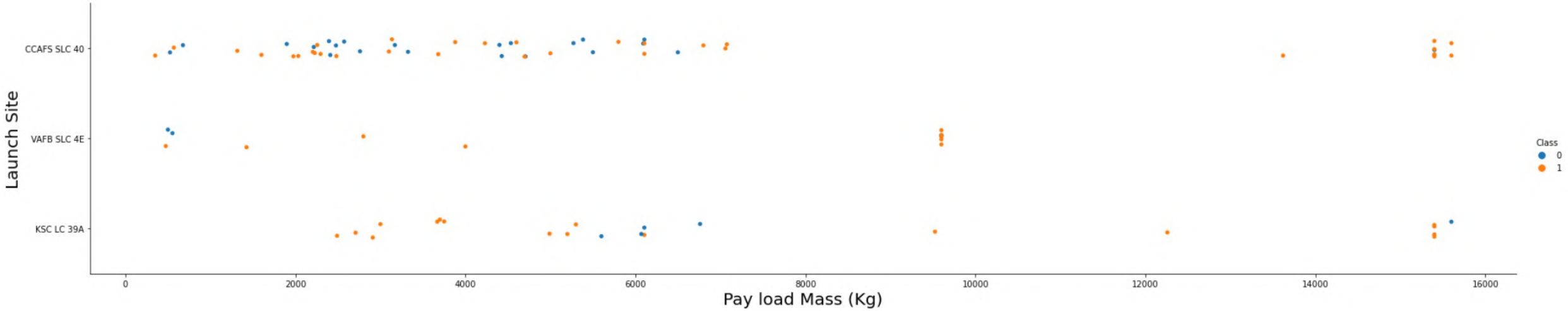

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site




- To begin, in our EDA, class attribute simply represents whether the first stage landed successfully or not. We compared between number of flights and class value, if we look closely at the right-hand side of the figure, most of it is dotted with orange dots that has a value of 1, which implies a successful landing. Thus, we can confidently say that the more trials have been made, the higher the success rate would be in future missions. Furthermore, An interesting insight can be noticed that most of the trials have been made in CCAFS SLC 40 launch site, thus we should give attention to that location.

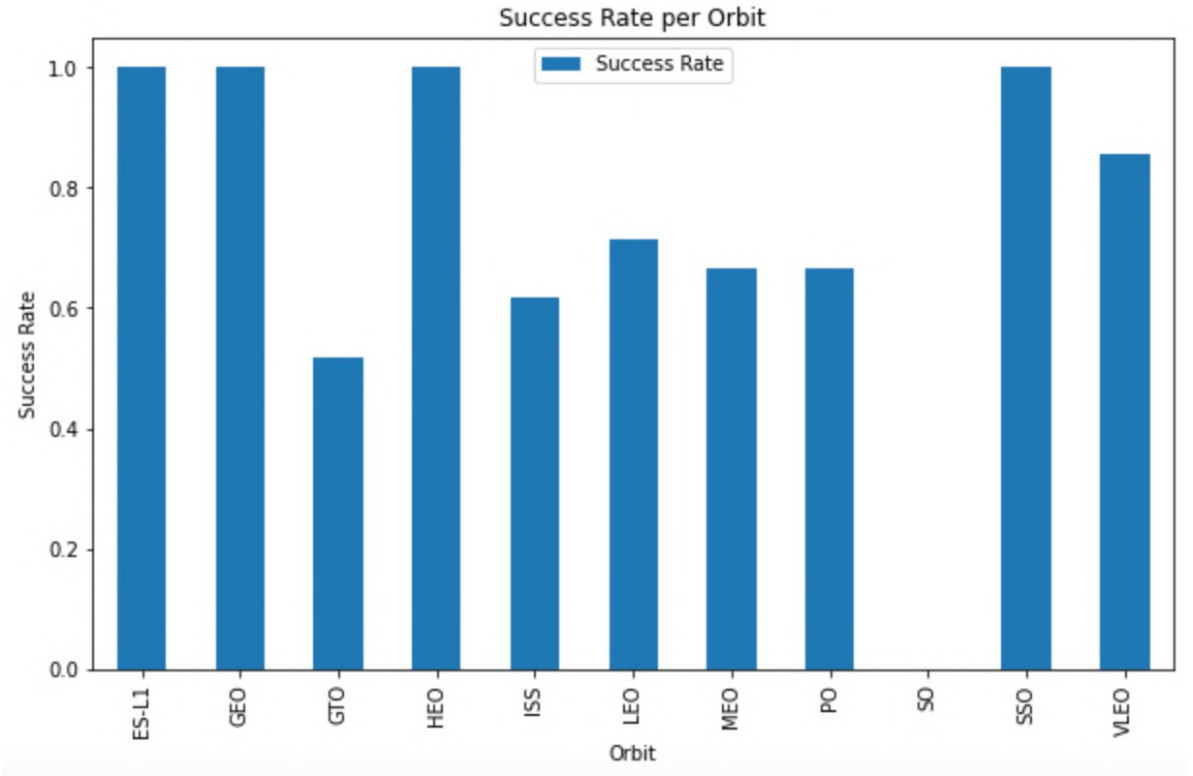
# Payload Mass vs. Launch Site




- Most landing experiments were made in the range of 0-7000 Kg payload mass, this implies that as the payload mass increases, it is less likely that the first stage will return to earth.

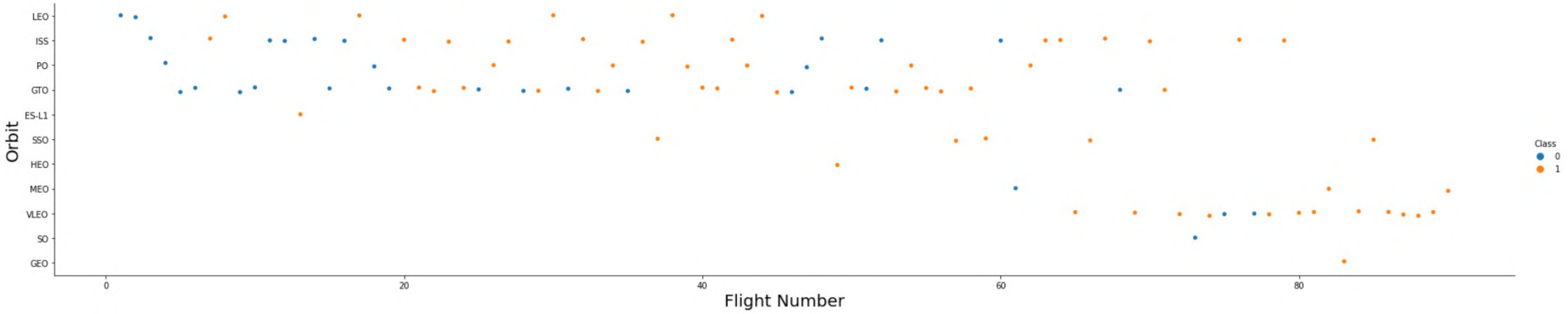
# Success Rate vs. Orbit Type

| Orbit | Success Rate |
|---|---|
| ES-L1 | 1.000000 |
| GEO | 1.000000 |
| GTO | 0.518519 |
| HEO | 1.000000 |
| ISS | 0.619048 |
| LEO | 0.714286 |
| MEO | 0.666667 |
| PO | 0.666667 |
| SO | 0.000000 |
| SSO | 1.000000 |
| VLEO | 0.857143 |



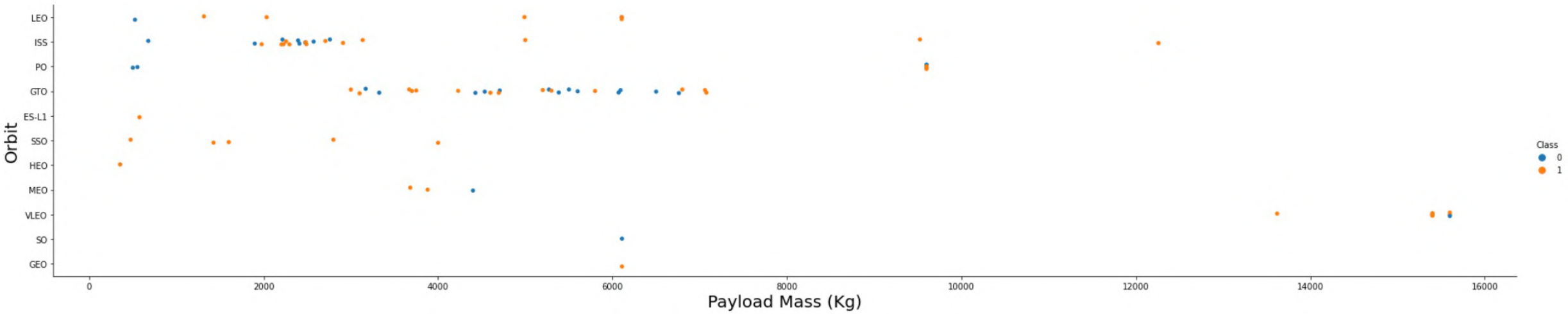

- Success rate per orbit is used in the predictive analysis process, for instance, if the rocket is heading toward SSO orbit, the first stage is more likely to land on earth successful than a first stage of a rocket that is heading towards MEO orbit.

# Flight Number vs. Orbit Type




- Looking closely at the scatter plot, for LEO orbit, it seems to us that as flight number increases, the success rate increases. In contrast, the GTO orbit success rate is not affected by number of flights. No obvious trend can be drawn from this visualization.

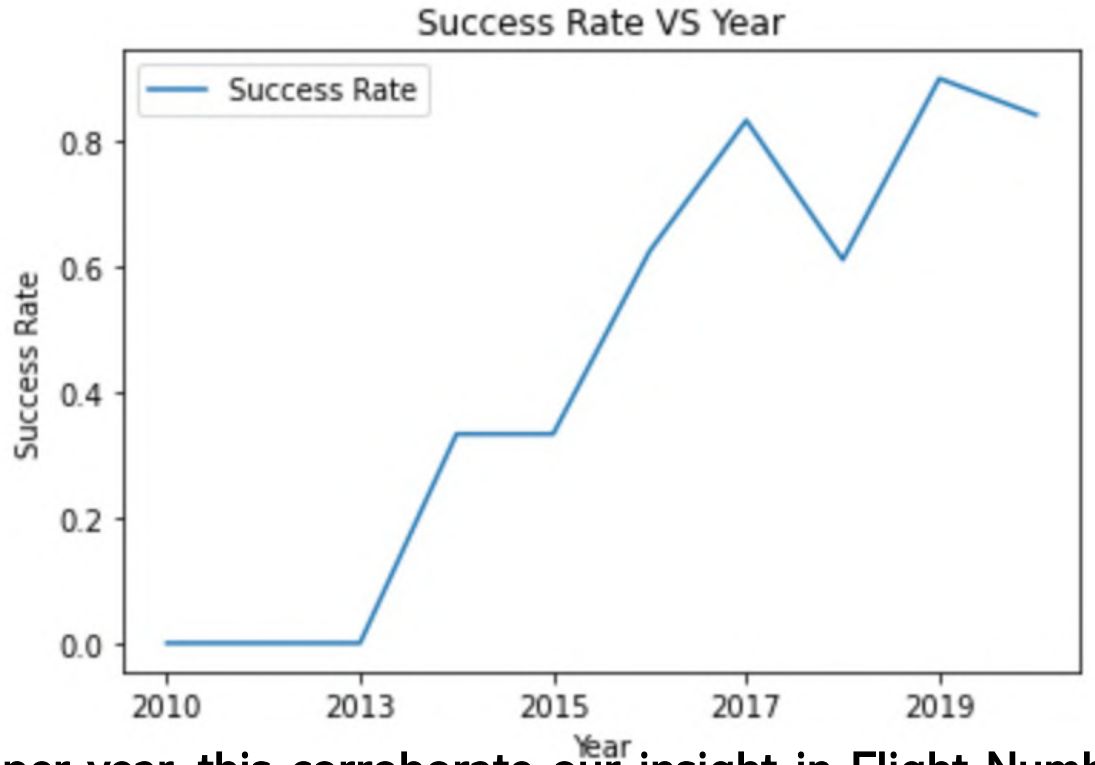
# Payload Mass vs. Orbit Type




- Here we observe the relationship between the orbit and payload mass. Looking closely at GTO orbit, we interpret that mass has a negative impact. On the other hand, the mass has a positive impact on ISS and LEO orbits. Therefore, as the aforementioned slide, no obvious trend between Payload mass and orbit type.

# Launch Success Yearly Trend

| Year | Success Rate |
|---|---|
| 2010 | 0.000000 |
| 2012 | 0.000000 |
| 2013 | 0.000000 |
| 2014 | 0.333333 |
| 2015 | 0.333333 |
| 2016 | 0.625000 |
| 2017 | 0.833333 |
| 2018 | 0.611111 |
| 2019 | 0.900000 |
| 2020 | 0.842105 |




- Yearly trend shows an overall increase per year, this corroborate our insight in Flight Number vs. Launch Site in slide [18], as the number of attempts increases, the possibility of a successful landing also increases.

# SQL - All Launch Sites Names

***Display the names of the unique launch sites in the space mission***

```
%sql select distinct launch_site from SPACEXDATA
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9
Done.
```

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

# SQL - Launch Sites Names Begin with 'CCA'

***Display 5 records where launch sites begin with the string 'CCA'***

```
%sql select  * from SPACEXDATA where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1oc
Done.
```

| DATE | time__utc_ | booster_version | launch_site | payload | payl |
|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 |
| 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 |
| 2012-05-22 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 |
| 2012-10-08 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 |

# SQL - Total Payload Mass Carried - NASA (CRS)

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

```
%sql select sum(payload_mass__kg_) from SPACEXDATA where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lc
Done.
```

| 1 |
|---|
| 45596 |

# SQL - Average Payload Mass by F9 v1.1

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

```
%sql select avg(payload_mass__kg_) as avg_mass from SPACEXDATA where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.ap
Done.
```

| avg_mass |
| --- |
| 2928 |

# SQL - First Successful Ground Landing Date

***List the date when the first succesful landing outcome in ground pad was acheived.***

*Hint:Use min function*

```
%sql select min(DATE) as first_date from SPACEXDATA  where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdo
Done.
```

| first_date |
| --- |
| 2015-12-22 |

# SQL - 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 distinct booster_version from SPACEXDATA\
where (landing__outcome = 'Success (drone ship)' and payload_mass__kg_ >4000 and payload_mass__kg_ <6000)
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.
Done.
```

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

# SQL - Total Number of Successful and Failure Mission Outcomes

- It is worth mentioning that mission outcome is not the same as landing outcome. The main purpose of the mission varies based on the costumer, for example, NASA wants to send a satellite into space. Unlike landing outcome (Class), which implies whether the first stage landed successfully or not.

***List the total number of successful and failure mission outcomes***

```
Successful = %sql select count(*) from SPACEXDATA where mission_outcome like 'Success%'
Failure = %sql select count(*) from SPACEXDATA where mission_outcome like 'Failure%'
Total = %sql select count(*) from SPACEXDATA

print ('Total successful mission outcomes is: ')
print(Successful)
print ('Total Faliure mission outcomes is: ')
print(Failure)
print ('Total mission outcomes is: ')
print(Total)
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.da
Done.
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.da
Done.
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.da
Done.
Total successful mission outcomes is:
+-----+
|  1  |
+-----+
| 100 |
+-----+
Total Faliure mission outcomes is:
+---+
| 1 |
+---+
| 1 |
+---+
Total mission outcomes is:
+-----+
|  1  |
+-----+
| 101 |
+-----+
```

# SQL - Boosters Carried Maximum Payload

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

```
%sql select distinct booster_version from SPACEXDATA\
where payload_mass__kg_ in (select max(payload_mass__kg_) from SPACEXDATA)
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde(
Done.
```

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

# SQL - 2015 Failed Drone Ship Launch Records

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

```
%sql select distinct booster_version, launch_site from SPACEXDATA \
where (landing__outcome = 'Failure (drone ship)' and year(DATE) = 2015)
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.datal
Done.
```

| booster_version | launch_site |
|---|---|
| F9 v1.1 B1012 | CCAFS LC-40 |
| F9 v1.1 B1015 | CCAFS LC-40 |

# SQL - 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(*) as total_number_of_outcomes from SPACEXDATA\
where DATE BETWEEN '2010-06-04' and  '2017-03-20'\
group by landing__outcome\
order by count(landing__outcome) desc
```

```
 * ibm_db_sa://mjh22310:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bluc
Done.
```

| landing__outcome | total_number_of_outcomes |
|---|---|
| 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

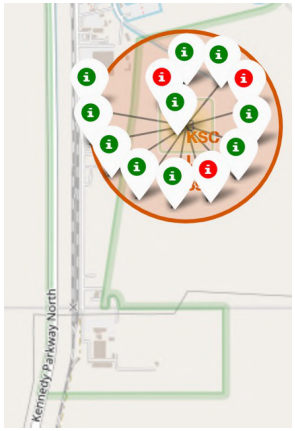
# Launch Site Locations Map

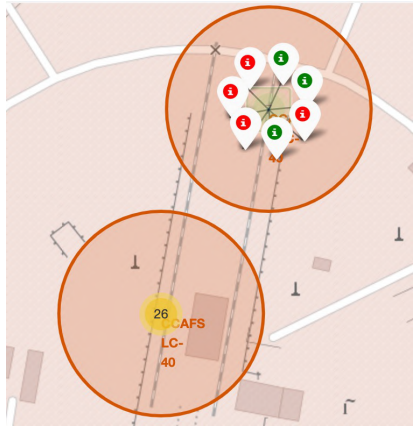

- CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A are in Florida State. Whereas VAFB SLC-4E in California State.
- All site locations are near the coast and Equator line, SpaceX focuses on locations that are close to water and the zeroth latitude.

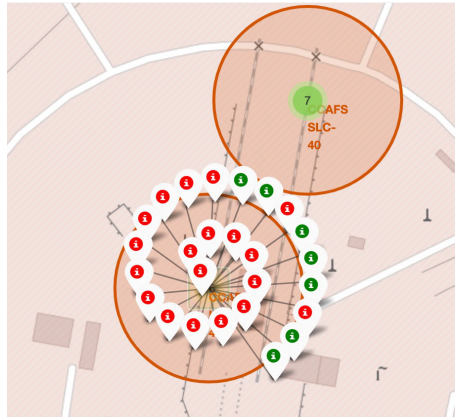


# Landing Outcome Per Site

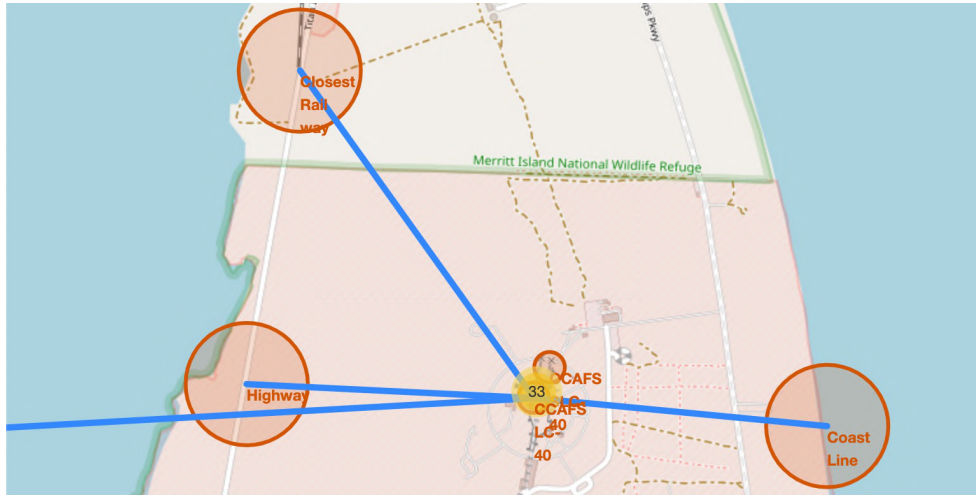
- We assigned each site with two kinds of markers, green marker for successful landing, and red in failure situations.
- Markers give a fast sense of which site is used more frequently and their success rate.
- The highest Success rate site is KSC LC-39A with 0.77%, and CCAFS LC-40 is the lowest with 0.27% success rate.

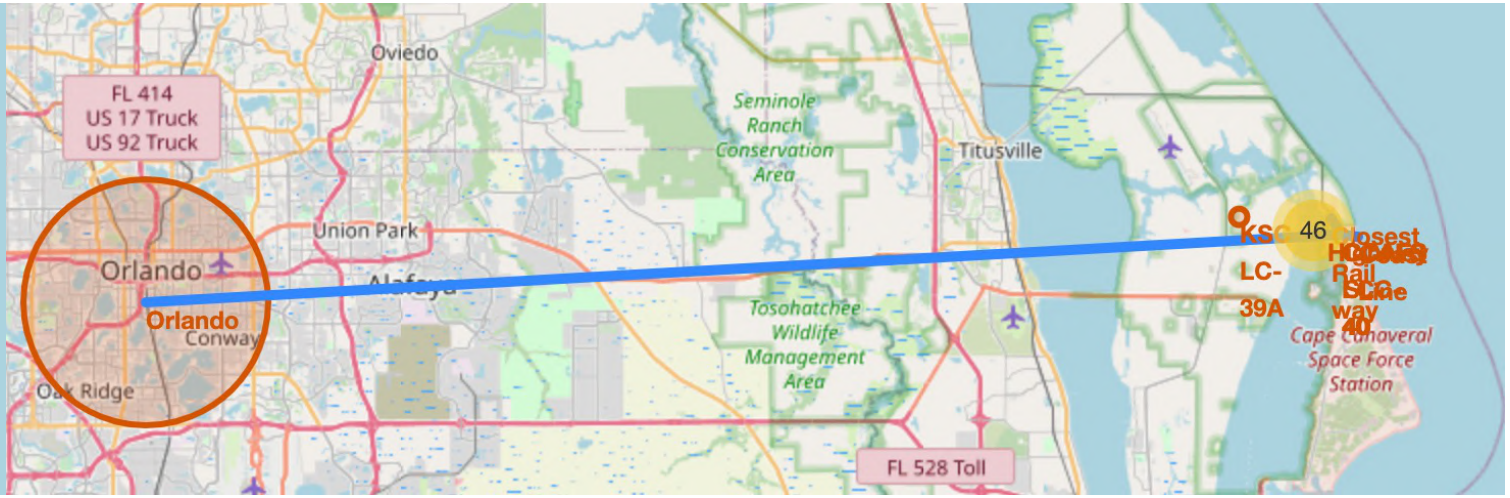




# Closest Proximities to CCAFS LC-40

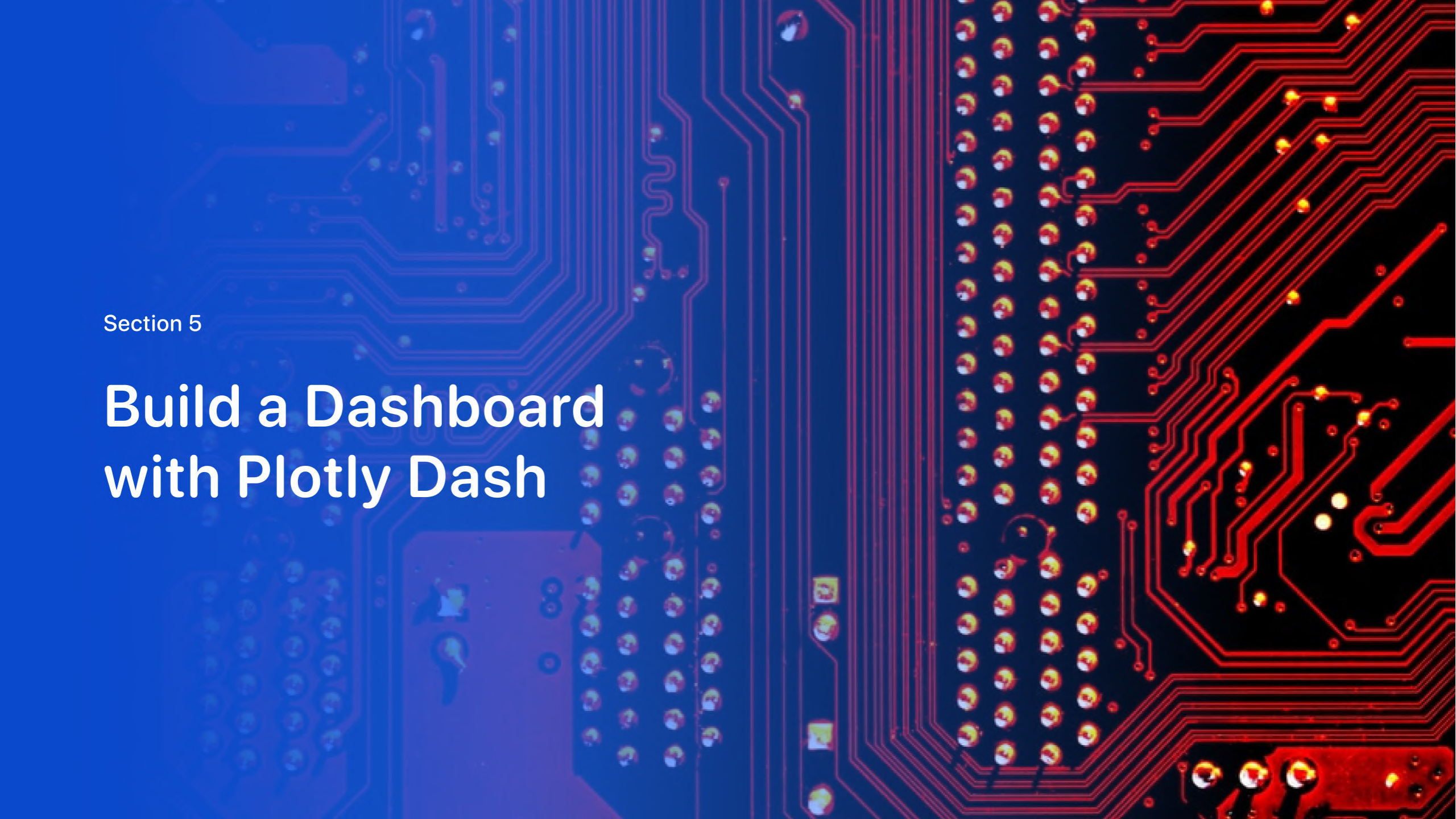


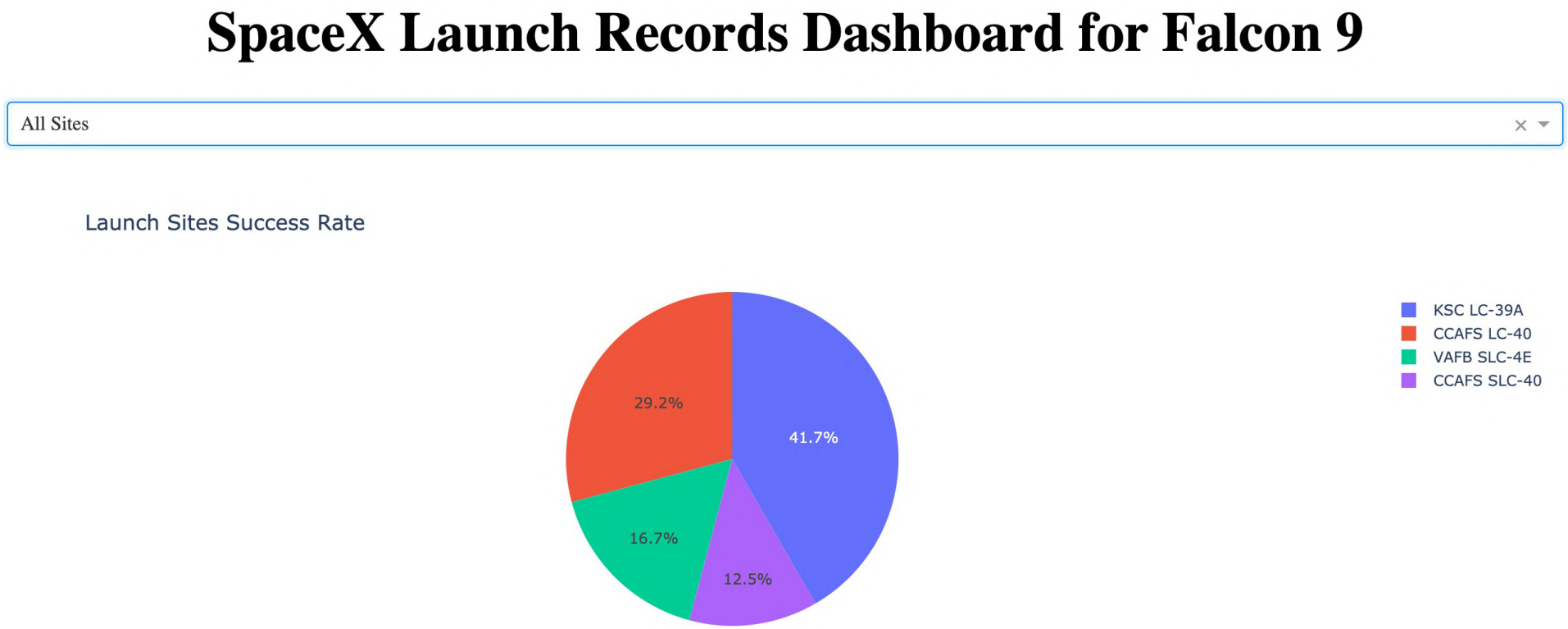

- Looking closely at CCAFS LC-40 site, we noticed that SpaceX focuses on keeping the launching site near to a coastline, railway, and highway. Compared to Orlando City, SpaceX launching sites are very far from inhabitant cities. The same principles are followed in the other stations.

- Orlando City Distance ≈ 78.8 Km, Coastline Distance ≈ 0.97 Km, Highway Distance ≈ 0.95Km

- From the first glance, being close to railways and highways would help a lot in transportation cost-reduction. Being close to a coastline and away from cities adds great safety value to rocket experiments.

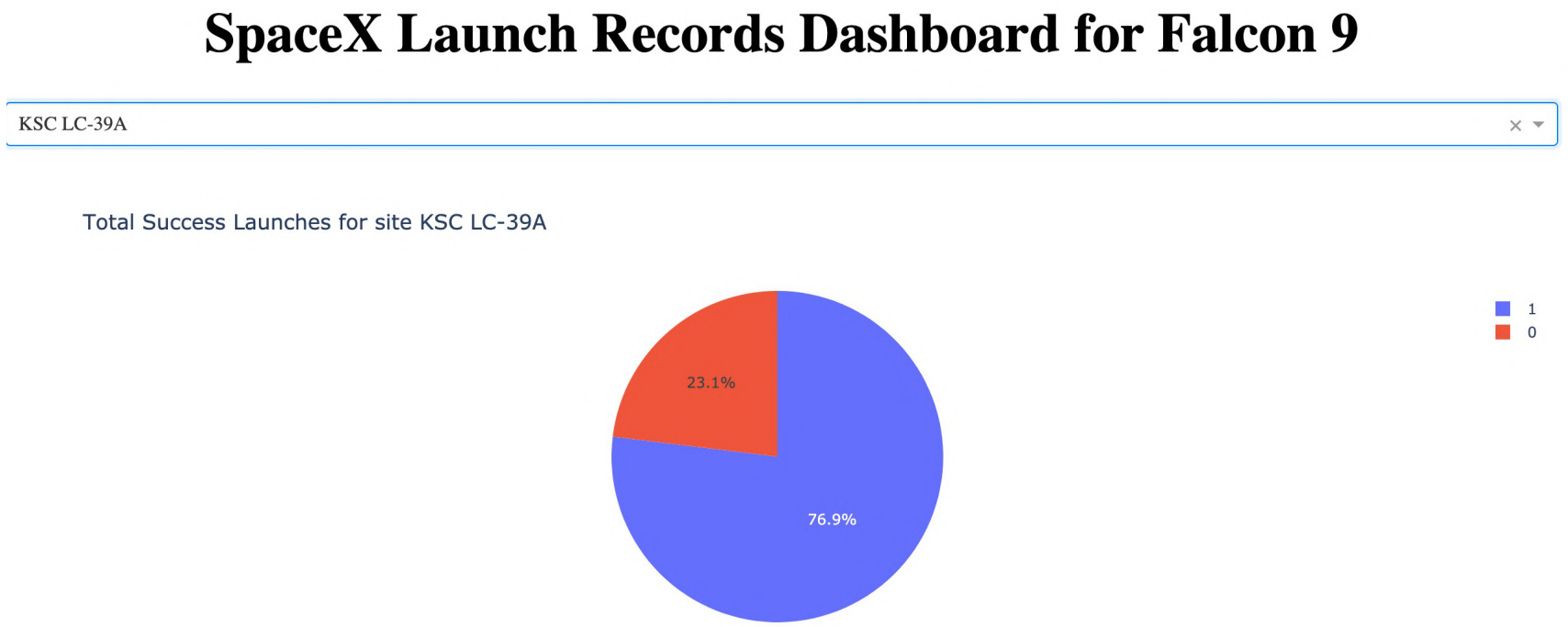

Section 5
Build a Dashboard with Plotly Dash

# Dashboard Upper Part –Success Rate for All Sites



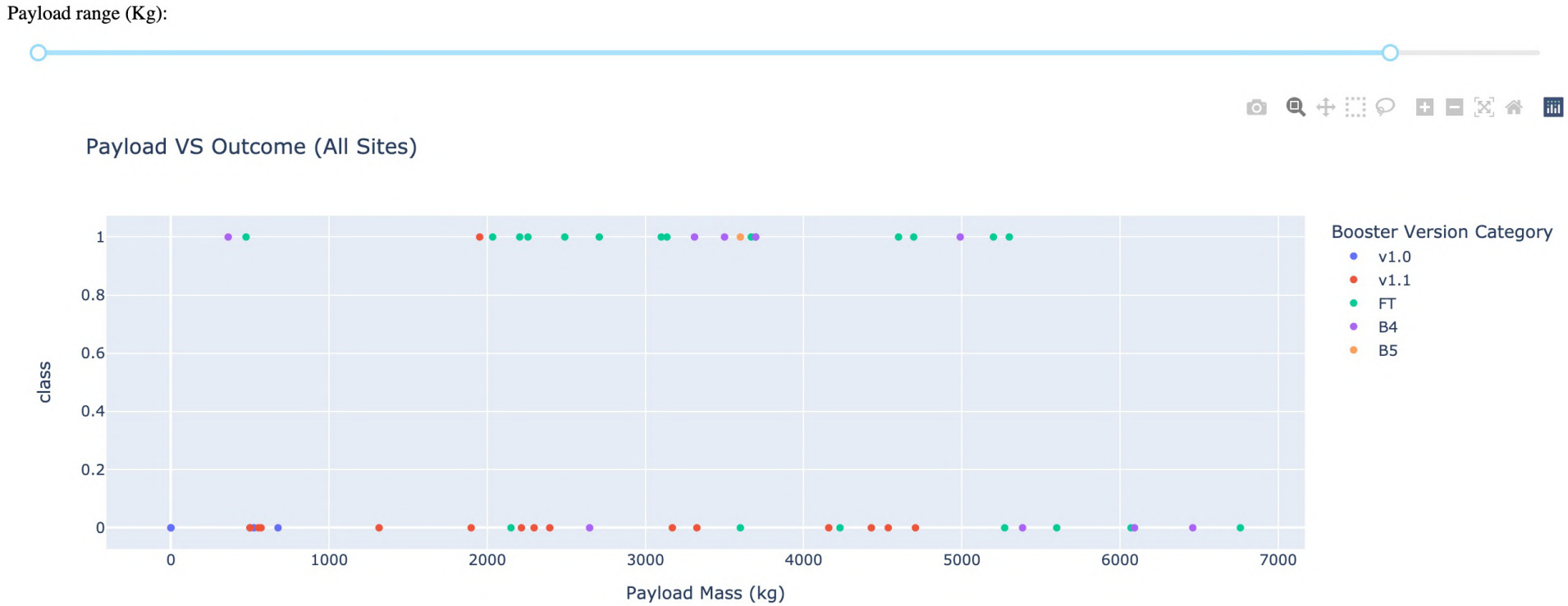

- This Pie Chart compares the success rate of each station to the total number of attempts in all of them. For instance, KSC LC – 39A is the highest with 41.7% success rate compared to the total number of attempts within this specific site and other sites.

# Dashboard Upper Part – Success Rate Per Site




- This Pie Chart compares the success rate of a specific station to the total number of attempts in that particular station only. We observe that success rate here is ≈ 77%, which is equal to our finding previously by using Folium interactive map marking in slide 36.

# Payload vs. Launch Outcome Scatter Plot Per Booster Version



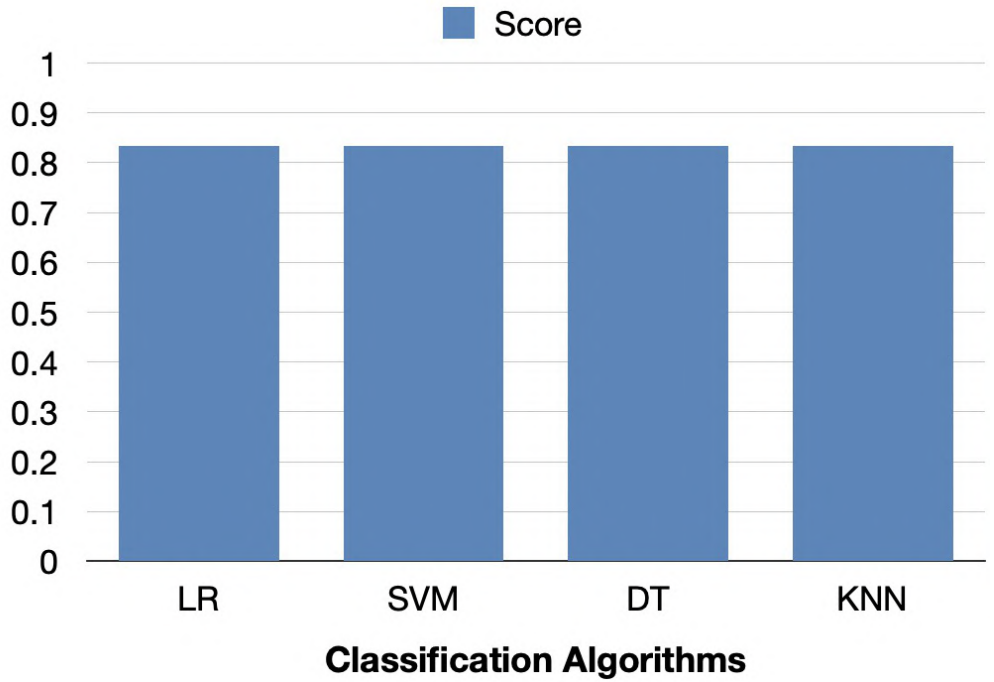

- This interactive slider shows the launch outcome according to two attributes, mass and booster version. An interesting insight we can observe is that success rate per booster version is different for different mass range, for example, FT version in 2000-4000 Kg is likely to land successfully, and vice versa for beyond that.

Section 6

# Predictive Analysis (Classification)

# Classification Accuracy

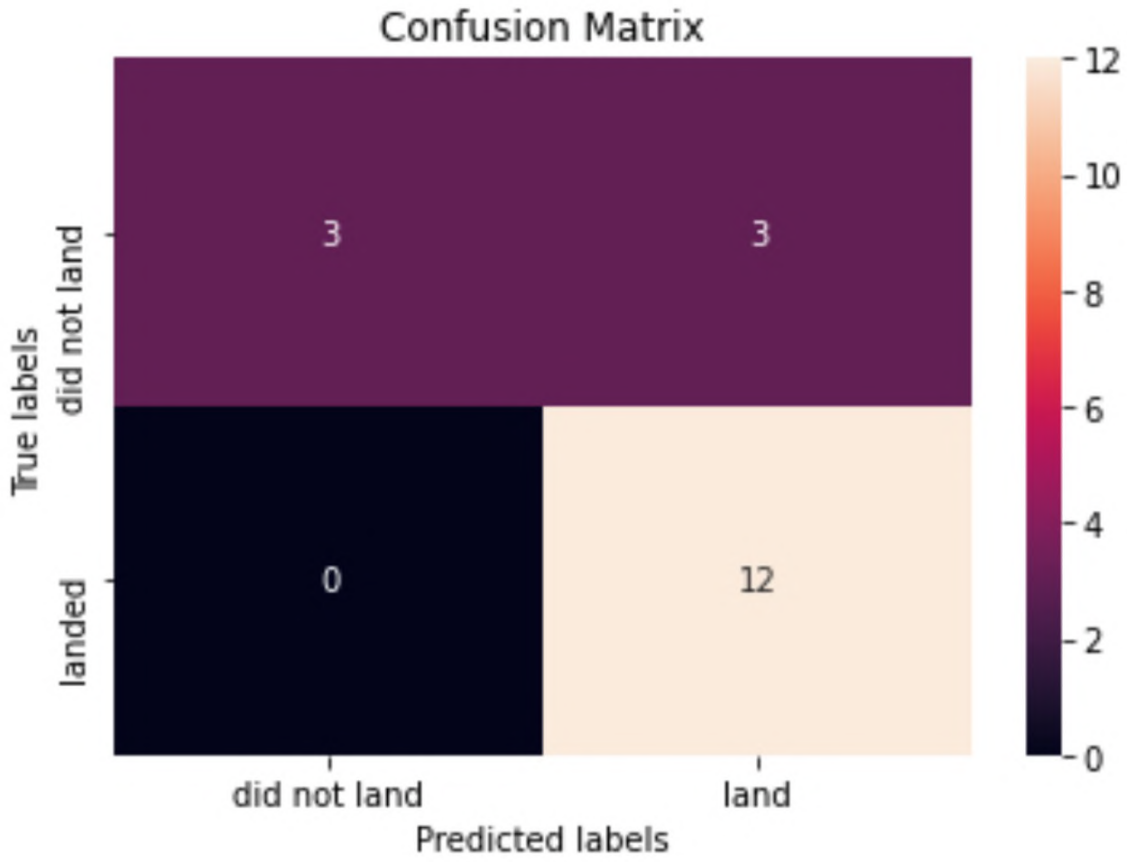
| Classification Algorithm | Score |
|---|---|
| LR | 0.833 |
| SVM | 0.833 |
| DT | 0.833 |
| KNN | 0.833 |




- The data has been split into two categories,
  - A. training set with 72 rows (80% of the data) and 83 attributes.
  - B. Testing set with 18 rows (20% of the data) and 83 attributes.
- An interesting finding in this predictive analysis is that we found the score for each of the used algorithm is identical. In each case, we used GridSearchCV.fit(X_train,Y_train).score(X_test,Y_test).

# Confusion Matrix

- This confusion Matrix is the output of all the used algorithms. This confirms our conclusion that these algorithms are practically the same for this data set.
- The first row represents the real failed experiments. As we can calculate, out of 18 experiments, 6 of them did not land successfully. Our classifier predicted 3 of them accurately, whereas the other 3 are labeled as land, which is incorrect.
- On the other hand, the second row represents the real successful landing experiments. As we can calculate, out of 18 experiments, 12 of them did land successfully. Our classifier predicted all them accurately.



# Conclusions

- In space rocket launching, a successful first stage landing, which enables us to reuse the same stage, reduces enormous amount of cost.
- A wide range of attributes affects the possibility of a successful landing. In our classifier, 83 attributes were taken into consideration. This classifier is crucial in anticipating the cost of future space journeys.
- SpaceX Falcon9 launch sites were all close to a highway, railway, and coastline proximities, which helped in transportation cost-reduction, but this insight requires further investigation.
- SpaceX success rate increased with years, KSC LC – 39A site is the highest in success rate.
- Orbit and booster version affect success rate.

# Appendix

- SpaceX API URL ”Click Here"
- SpaceX Static Wikipedia URL ”Click Here"
- SpaceX data used in ML training “Click Here”


10 Courses

- What is Data Science?
- Tools for Data Science
- Data Science Methodology
- Python for Data Science, AI & Development
- Python Project for Data Science
- Databases and SQL for Data Science with Python
- Data Analysis with Python
- Data Visualization with Python
- Machine Learning with Python
- Applied Data Science Capstone

Sep 8, 2021

## SALAH ALDDEEN YACOUB A. AL KAFRAWI

has successfully completed the online, non-credit Professional Certificate

# IBM Data Science

In this Professional Certificate learners developed and honed hands-on skills in Data Science and Machine Learning. Learners started with an orientation of Data Science and its Methodology, became familiar and used a variety of data science tools, learned Python and SQL, performed Data Visualization and Analysis, and created Machine Learning models. In the process they completed several labs and assignments on the cloud including a Capstone Project at the end to apply and demonstrate their knowledge and skills.

Rav Ahuja
AI & Data Science
Program Director
IBM Skills Network

The online specialization named in this certificate may draw on material from courses taught on-campus, but the included courses are not equivalent to on-campus courses. Participation in this online specialization does not constitute enrollment at this university. This certificate does not confer a University grade, course credit or degree, and it does not verify the identity of the learner.

Verify this certificate at:
coursera.org/verify/professional-cert/AF4SFD9W6WN4

# Thank you!