Winning Space Race with Data Science

Sandeep Mittal
14-Aug-2022

Outline

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

Executive Summary

Executive Summary

- Methodology
  - Data was collected from various sources. The primary sources were the SpaceX API and Falcon 9 Wiki site. Data was then filtered, cleaned, null values tackled and standardized. After that, exploratory data analysis was carried out. Data visualization was done. A number of charts and graphs were used to find out the relationships between various features. Interactive maps were created to get a sense of where the launch sites were located, what was near them, and the success/failure depicted on the map. Database queries were written to understand the data and get a sense of trends in the data. Classification exercise was carried out to develop a model which can predict whether the first stage of the rocket will land successfully or not based on various input features.

Executive Summary

- Key results are:
  - Success rate has gone up with experience as the time has passed.
  - Maximum number of successful launches have been for Payload in the range 2000 kg to 4000 kg.
  - FT booster version has the highest success rate across the whole range of Payload mass.
  - Based on the analysis exercise, we can conclude that sites KSC LC-39A and CCAFS LC-40 have the most number of successful launches of Falcon 9. These sites also have the best success rate for the launches from them.
- We can use decision tree classification model to predict whether the first stage will land successfully or not.

Introduction

Introduction

- Project background and context

  The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX. SpaceX's accomplishments include sending spacecraft to the International Space Station, Starlink, a satellite internet constellation providing satellite Internet access, and sending manned missions to space. One reason SpaceX can do this is the rocket launches are relatively inexpensive.

  Space Y, a company that would like to compete with SpaceX founded by billionaire industrialist Allon Musk. In order to compete, SpaceY would like to determine if SpaceX will reuse the first stage. This will help in determining the price of each launch.

Introduction

- Problems we want to find answers for

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes, the first stage does not land. Sometimes it will crash. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.

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

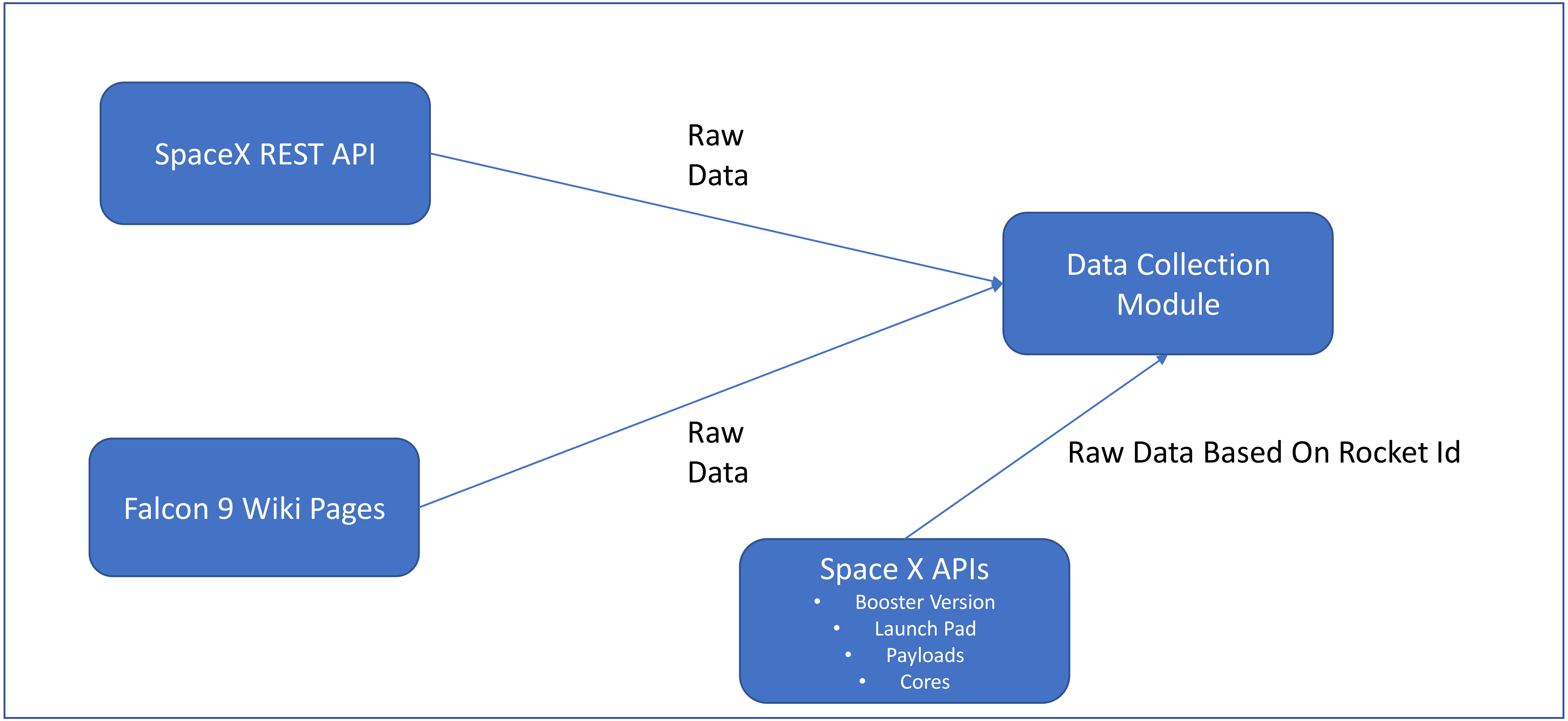
Data Collection

Data Collection

In this project, we worked with SpaceX launch data that is gathered from an API, specifically the SpaceX REST API available at https://github.com/r-spacex/SpaceX-API. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

We also be performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches available at https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches.

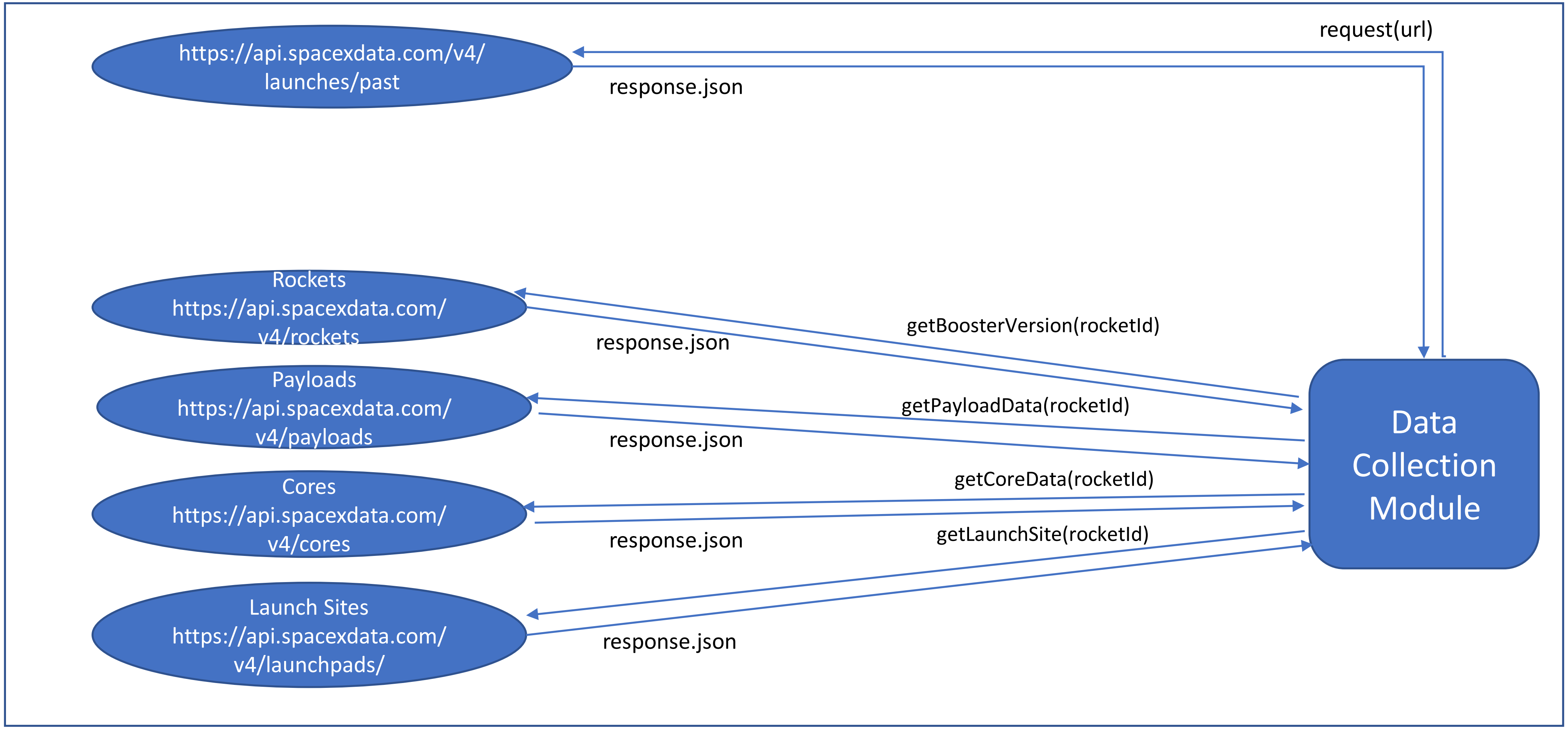
Data Collection



Data Collection – SpaceX API

- SpaceX launch data is available via REST API. The endpoint for the SpaceX historical data is https://api.spacexdata.com/v4/launches/past.
- An HTTP call is made to this endpoint and JSON data is received in return.
- JSON data is then converted into pandas dataframe for further processing.
- In some of the columns, like rocket, we have an identification number, not actual data. This means we will need to use the API again targeting another endpoint to gather specific data for each ID number. These end points are for Booster, Launchpad, payload, and Core. Following are the end points for various kinds of information:
  - https://api.spacexdata.com/v4/rockets/ (Booster info)
  - https://api.spacexdata.com/v4/launchpads/ (Launch site info)
  - https://api.spacexdata.com/v4/payloads (Payload mass and Orbit information)
  - https://api.spacexdata.com/v4/cores (Technical info and Landing Outcome data)
- Data is received in JSON format from these endpoints, which is further converted into pandas dataframes and merged with the main dataframe created earlier.

Data Collection – SpaceX API Calls



Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook:

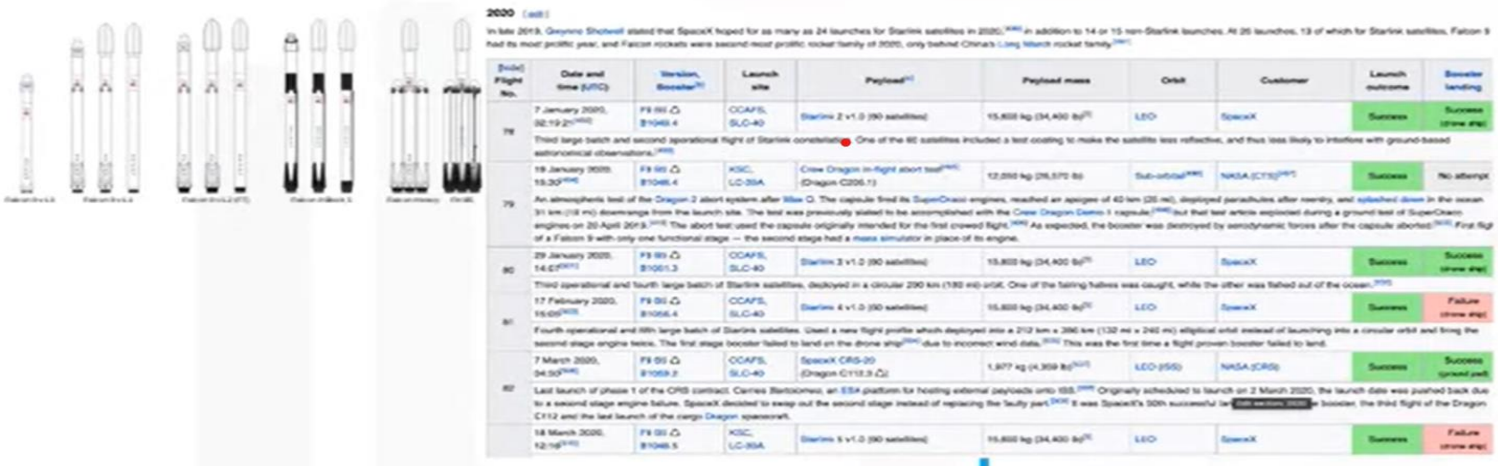
https://github.com/sandeepmittal70/my-ibm-ds-cert/blob/main/SpaceX%20Data%20Collection.ipynb

Data Collection - Scraping

- Data is further collected from Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches available at https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- We used Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records. Then we parsed the data from those tables and converted them into a Pandas data frame for futher analysis.

Data Collection - Scraping

Web scraping Falcon 9 Launch records



Web scraping with BeautifulSoup

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Data Collection - Scraping

- GitHub URL of the completed web scraping notebook:

https://github.com/sandeepmittal70/my-ibm-ds-cert/blob/main/SpaceX%20Web%20Scraping.ipynb

Data Wrangling

Data Wrangling

- After collecting the data from SpaceX API and Web scraping the Wiki site, the data was made available in a pandas dataframe.
- The dataframe contained data for Falcon 9 and other rockets as well. The dataframe was first filtered to extract data from Falcon 9 only.
- Not all gathered data is perfect. We may end up with data that contains NULL values. We must sometimes deal with these null values in order to make the dataset viable for analysis.

Data Wrangling

- We can see below the number of Null values in data in various columns:

```
FlightNumber       0
Date               0
BoosterVersion     0
PayloadMass        5
Orbit              0
LaunchSite         0
Outcome            0
Flights            0
GridFins           0
Reused             0
Legs               0
LandingPad        26
Block              0
ReusedCount        0
Serial             0
Longitude          0
Latitude           0
```

Data Wrangling

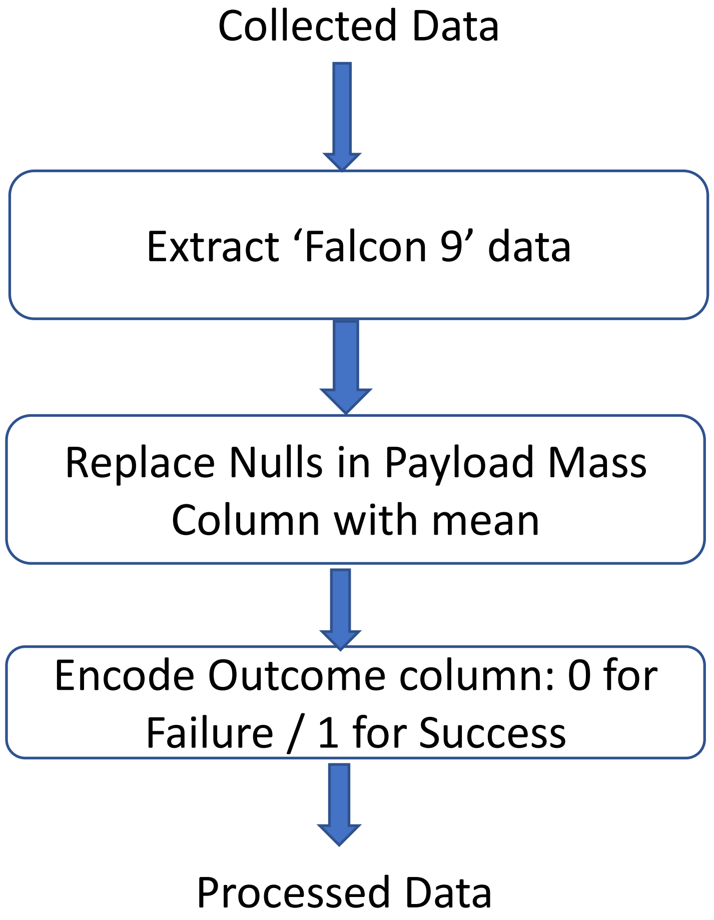
- We see that columns ‘LandingPad’ and ‘PayloadMass’ have some Null values in them.
- Dealing with Null values
  - We replaced nulls in PayloadMass column with the mean of the existing data. This is a reasonable approximation.
  - The LandingPad column was left alone. Null values represented when landing pads were not used. This was dealt with using one hot encoding later on.

Data Wrangling

- Created a landing outcome label from Outcome column. This variable represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully. Existing values were encoded as zeros and ones as follows:

```
True ASDS   - 1
True RTLS   - 1
True Ocean  - 1
False Ocean - 0
None None   - 0
None ASDS   - 0
False RTLS  - 0
False ASDS  - 0
```

Data Wrangling



Data Wrangling

- GitHub URL of the completed data wrangling related notebook

https://github.com/sandeepmittal70/my-ibm-ds-cert/blob/main/SpaceX%20Data%20wrangling.ipynb

EDA with Data Visualization

EDA with Data Visualization

- Discussed below is the list of charts which were plotted and why they were plotted
  - Categorical plot - Flight Number vs Payload mass

    This shows how the Flight Number (indicating the continuous launch attempts.) and Payload variables affect the launch outcome.

  - Categorical plot - Flight Number vs Launch Site

    This chart show the success rates of flights from each of the launch sites.

  - Categorical plot - Payload Vs. Launch Site

    This plot shows us if there is any relationship between launch sites and their payload mass. For example, how heavy a payload has a launch site handled in the past.

EDA with Data Visualization

- Bar chart - Orbit type and Success rate.

  This chart helps us visualize which orbits have high success rate.
- Categorical plot – Flight Number and Orbit type

  This chart helps us visualize how the success rate of an orbit type is related to the number of flights.
- Categorical plot – Payload vs Orbit type

  This plot helps us visualize the success rate of an orbit type based on the size of the payload. Do heavy payloads have higher success rate in particular orbit type than othes?
- Line plot - Launch success rate yearly trend

  This plot shows how the average success rate changes over the years. Has it improved or has it deteriorated? What is the trend over the years?

EDA with Data Visualization

- GitHub URL of the completed EDA with data visualization notebook

https://github.com/sandeepmittal70/my-ibm-ds-cert/blob/main/SpaceX-labs-eda-dataviz.ipynb

EDA with SQL

EDA with SQL

- Following SQL queries were performed
  - Display the names of the unique launch sites in the space mission
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Total Payload carried by Falcon 9 over the years
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

EDA with SQL

- 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 records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

EDA with SQL

- GitHub URL of the completed EDA with SQL notebook

https://github.com/sandeepmittal70/my-ibm-ds-cert/blob/main/SpaceX-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Build an Interactive Map with Folium

- Map objects added to the Folium maps
  - Circle around the coordinates of the launch site show the launch sites on the map
  - Text label marker giving the name of the launch site
  - Popup text which is displayed in the circle around the launch site coordinates when clicked
  - Red (failed launch) and Green (successful launch) icon markers at each of the launch sites. These markers were added to a marker cluster. Marker clusters allow multiple markers for the same coordinates. From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

Build an Interactive Map with Folium

- Added a MousePosition on the map to get coordinate of a point on the map as pointed by the mouse pointer. As such, while you are exploring the map, you can easily find the coordinates of any points of interests (such as railway) in the proximity of a launch site.
- Add polyline to a location in the proximity of a launch site along with a distance marker giving the distance between the launch site and the proximity location in kilometers.

Build an Interactive Map with Folium

- GitHub URL of the completed interactive map with Folium map

https://github.com/sandeepmittal70/my-ibm-ds-cert/blob/main/SpaceX_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard
  - A pie chart showing total successful launches for all sites. We can easily see pictorially which site has the largest successful launches.
  - For each site, a pie chart is drawn to depict the percentage of successful launches. We can compare the percentages of various sites and figure out which site has the maximum success rate.
  - For the pie chart functionality, a drop-down list was added which allowed the user to select ‘All sites’ or a specific launch site.
  - Correlation between Payload mass and launch success for all sites and each site separately. This is a categorical scatter plot. It helps us answer questions like ‘which payload range has the highest or the lowest success rate’ or ‘which booster version has the highest launch success rate’.
  - A range slider was added to select the payload range for which the scatter plot was drawn to show the correlation between payload and launch success rate.

Build a Dashboard with Plotly Dash

- GitHub URL of the completed Plotly Dash lab

https://github.com/sandeepmittal70/my-ibm-ds-cert/blob/main/spacex_dash_app_my-solution.py

Predictive Analysis (Classification)

Predictive Analysis (Classification)

- In order to build a predictive model, first, the features were standardized. Standardization of a dataset is a common requirement for many machine learning estimators. After standardization, individual features look like normally distributed data (Gaussian or Normally distributed with 0 mean and unit variance). This makes the features to be in the same order of magnitude to each other so that a single feature does not dominate the estimator and make the effect of other features insignificant.
- In order to get a good estimator from the data, the data was split into a training set (80% of data) and test set (20% of data).

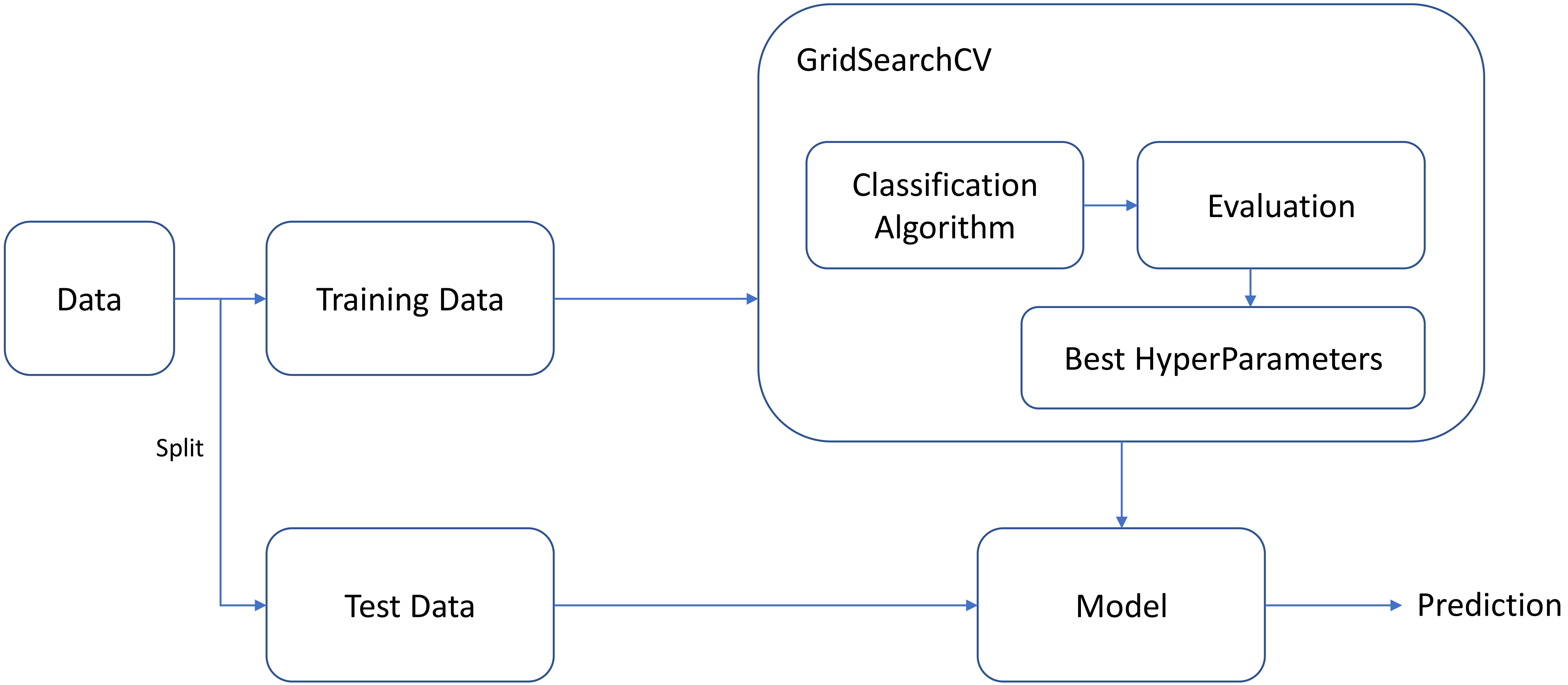
Predictive Analysis (Classification)

- For each classification algorithm - Logistic Regression, SVM, Decision Tree and K-Nearest Neighbours – following steps were carried out:
  - The classification estimator for the classifier was created.
  - Different values of hyperparameters were passed to GridSearchCV which found the best hyperparameter values using 10-fold cross validation.
  - Then predictions were made using the test data with this best model and the result compared with the actual result.
  - An accuracy score was calculated for the predictions on the test data.
- The classifier with the best accuracy score was chosen as the preferred classifier for making predictions.

Predictive Analysis (Classification)

- A few words regarding cross-validation
  - In this approach, called k-fold CV, the training set is split into k smaller sets. For each of the k "folds", a model is trained using of the folds as training data. The resulting model is validated on the remaining part of the data (called test data) and accuracy calculated. The performance measure reported by k-fold cross-validation is then the average of the values computed in the loop.
- GridSearchCV –
  - GridSearchCV does exhaustive search over specified parameter values for an estimator using cross-validation. The parameters of the estimator are optimized by cross-validated grid-search. The parameters selected are those that maximize the score of the left-out data.

Predictive Analysis (Classification)



Predictive Analysis (Classification)

- GitHub URL of the completed predictive analysis lab

https://github.com/sandeepmittal70/my-ibm-ds-cert/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

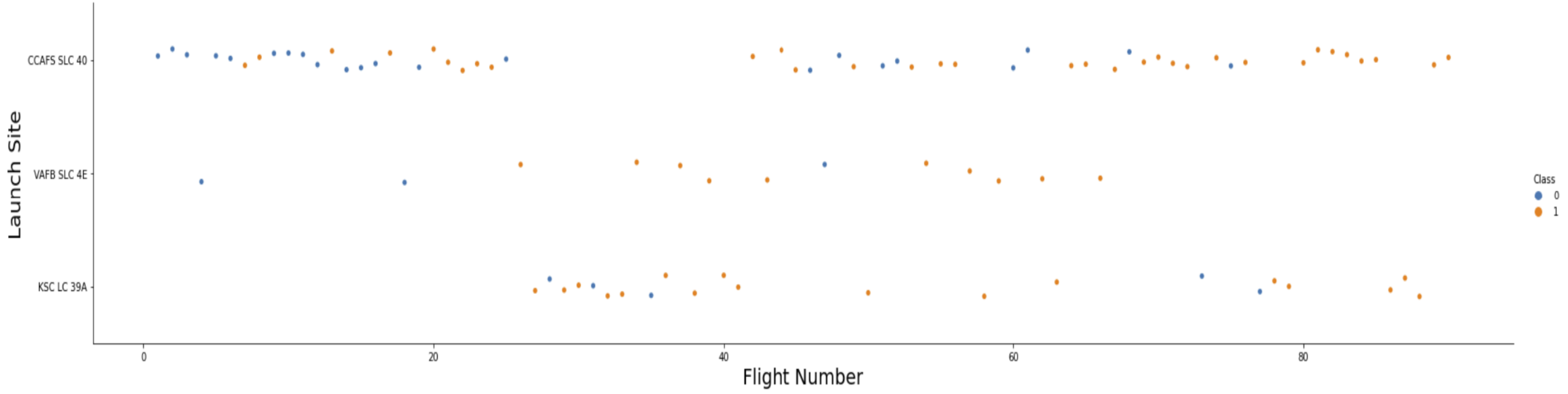
Results

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

Section 2

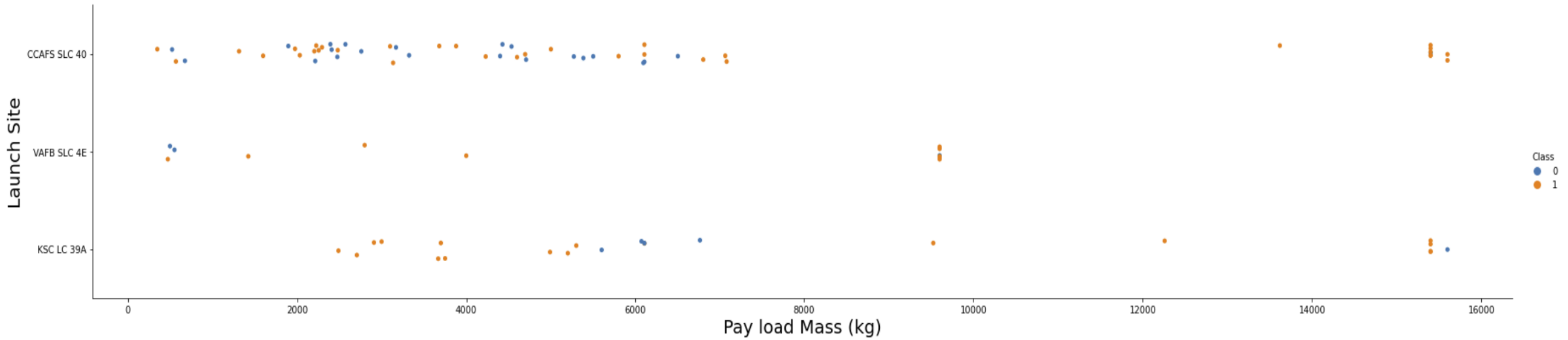
Insights drawn from EDA

Flight Number vs. Launch Site



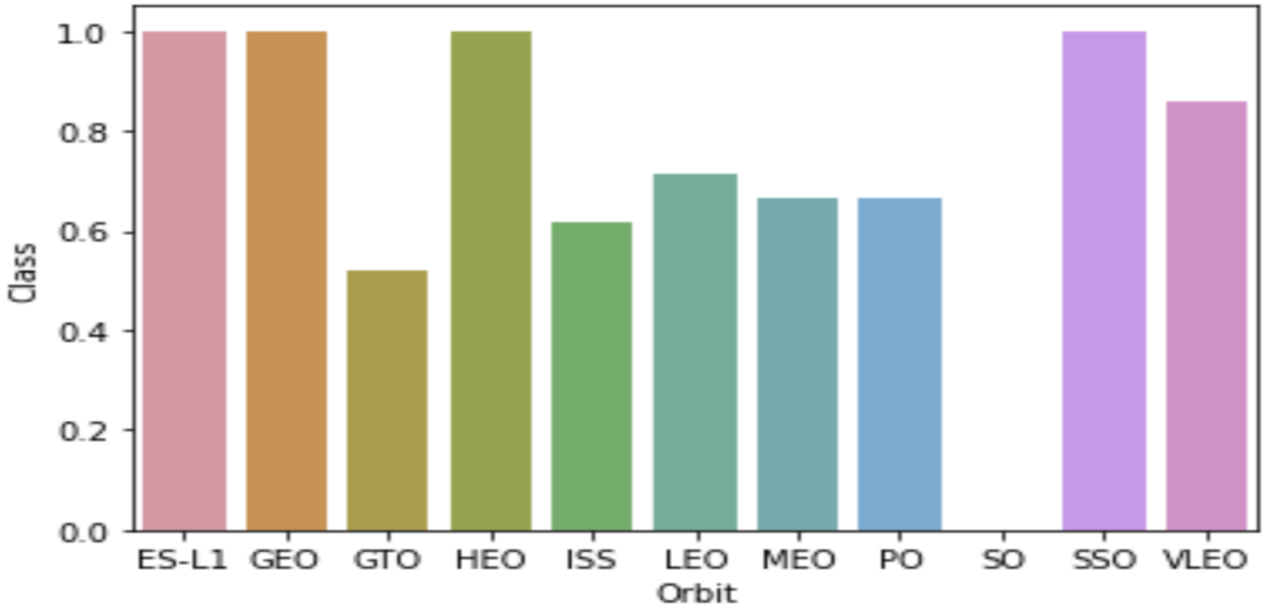
- Flight number represents flight numbers as time passes by. With each flight launch, the launch site gains experience.
- As we can see from the scatter plot, for all the launch sites, as the time passes, i.e., as the flight number increases, the number of successful landings (orange dots) increases. Experience counts.

Payload vs. Launch Site



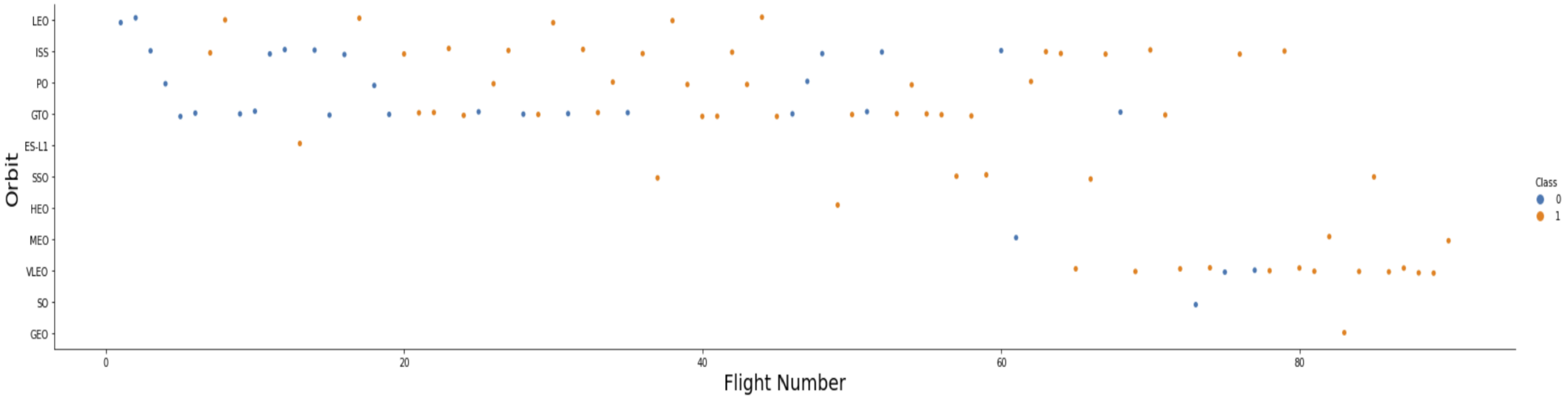
- As the payload mass increases, the number of successful landings also increases, although the launch sites 'KSC LC 39A' and 'VAFB-SLC ' have good success rate at lower payload as well.
- For payload mass 10000 kg and above, the success rate is nearly 100% for all remaining all the sites barring 'VAFB-SLC', where there have been no launches of payload 10000 kg or more.

Success Rate vs. Orbit Type



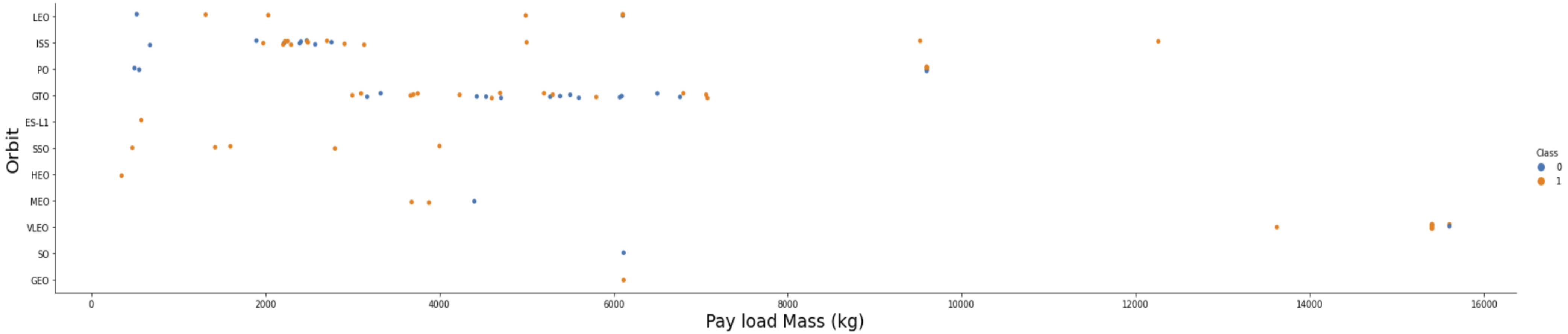
- From the chart we can see that for the orbit type 'ES-L1', 'GEO', 'HEO' and 'SSO', the success rate is 100%.
- Orbit type 'SO' does not have any successes.

Flight Number vs. Orbit Type



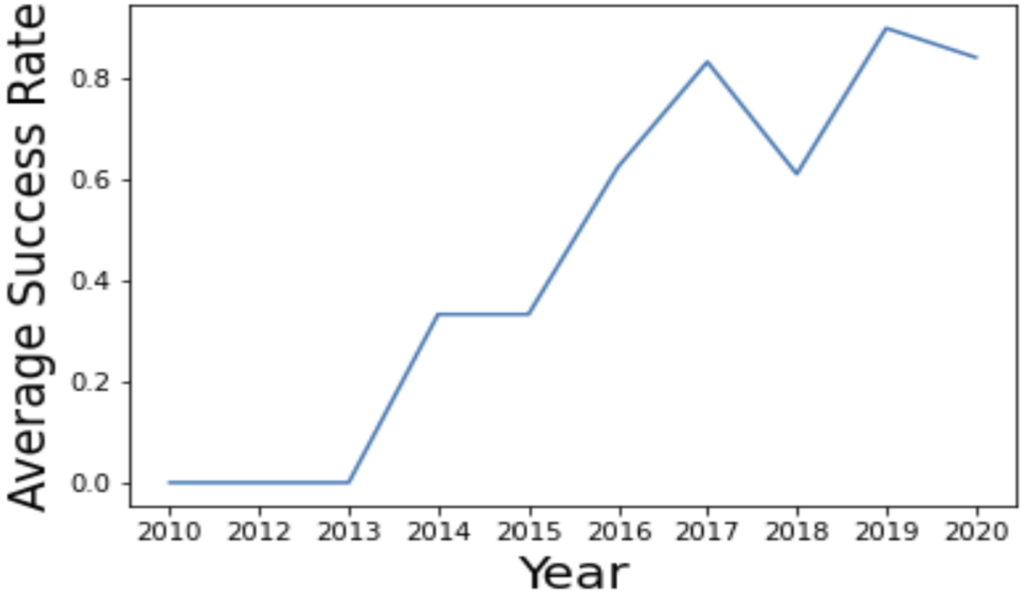
- For LEO orbit the Success appears related to the number of flights
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing rate are more for LEO and ISS.
- SSO orbit has good success rate even at lower payloads.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- Till 2013, the success rate was 0. However, the sucess rate since 2013 picked up and kept increasing till 2020.

All Launch Site Names

- Result

Launch Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

- Explanation

There are four launch sites in United States which the company SpaceX has used for Falcon 9 launches since 2010 till 2020.

Launch Site Names Begin with 'CCA'

- Result

  On the next slide

- Explanation

  The query lists the names of launch sites which begin with ‘CCA’. The retrieved records are randomly selected. There is no guarantee of which launch sites may be included. In our results, records for only one site are displayed, namely, CCAFS LC-40.

Launch Site Names Begin with 'CCA'

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

Total Payload Mass

- Result

Total Number of Flights	Total Payload Carried by boosters launched by NASA (CRS)	Average Payload
20	45596	2279.8

- Explanation

  Although the total payload seems impressive at 45,596 kg, but per flight payload is only 2,280 kg which is far lighter than the overall average of 6,138 kg (not shown in the results here). NASA seems to be launching rockets with light payloads.

Average Payload Mass by F9 v1.1

- Result

Average Payload Carried by Booster Version F9 v1.1
2928.4

- Explanation

  The average payload carried by booster version 'F9 v1.1' is far lighter than the overall average of 6,138 kg (not shown in the results here). It seems that this booster version is meant for lighter payloads.

First Successful Ground Landing Date

- Result

Landing Date	Landing Outcome
2015-12-22	Success (ground pad)

- Explanation

  It took 5 years since 2010 for Falcon 9 to achieve first successful landing on ground. This also happens to be the first successful overall, whether on ground, drone or ocean.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Result

Booster	Payload Mass
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- Explanation

Four successful landings have been for payload between 4000 and 6000 kg out of a total of 24 landings (successful or failed on any surface) in this payload mass range.

Total Number of Successful and Failure Mission Outcomes

- Result

Mission Outcomes	Total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Explanation

Mission outcome results are great – 100 out of 101 successful outcomes.

Boosters Carried Maximum Payload

- Result
- Explanation

  12 times the boosters carried the maximum payload mass out of a total of 101 missions.

Booster Version	Payload Mass
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- Result

Year	Month	Landing Outcome	Booster Version	Launch Site
2015	Jan	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	Apr	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Explanation

These were the first two landing failures on a drone ship since 2010.

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

- Result

Rank	Landing Outcome	Number of Successful Landings
1	Success (drone ship)	5
2	Success (ground pad)	3

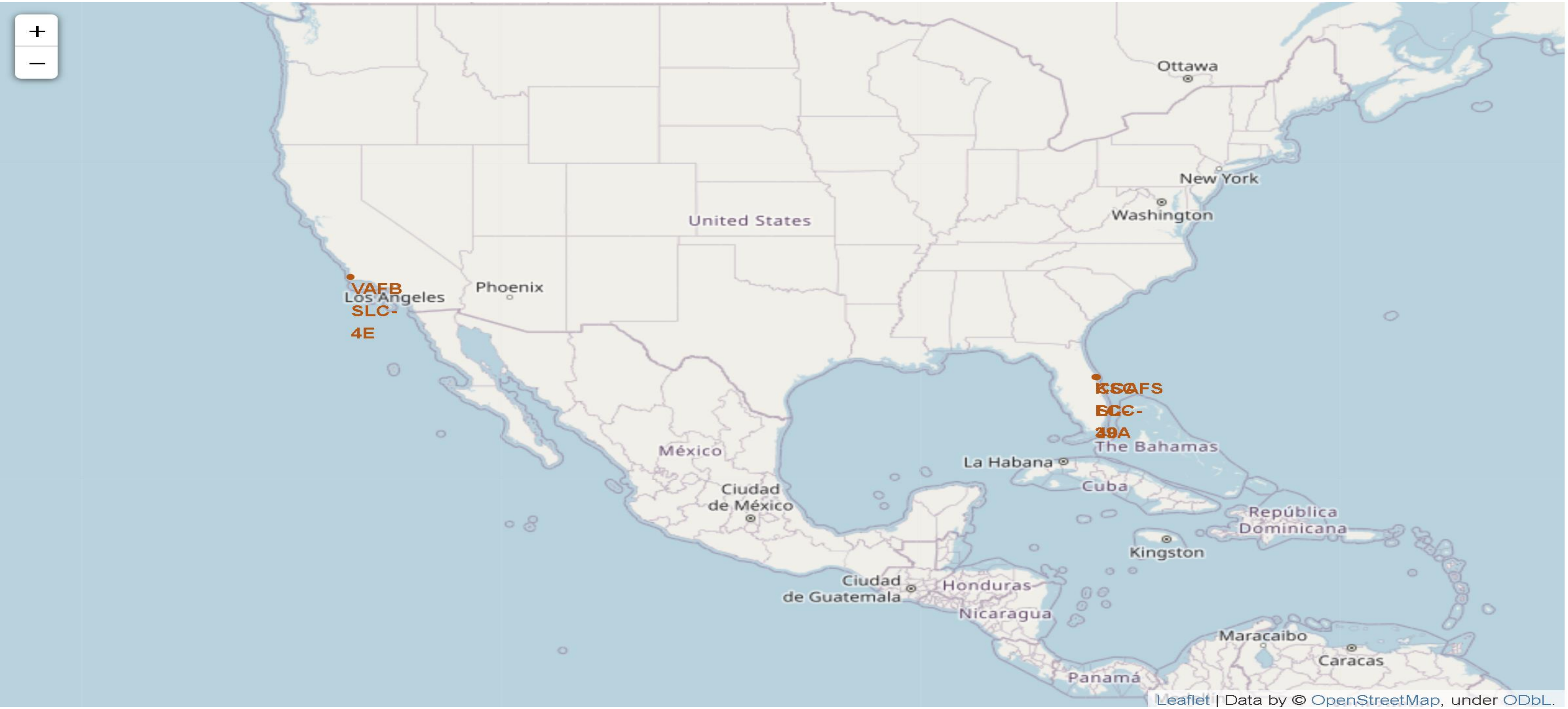
- Explanation

There were a total of 8 successful landings between 2010-06-04 and 2017-03-20, out of which 5 were on drone ship and 3 on ground pad.

Section 3

Launch Sites Proximities Analysis

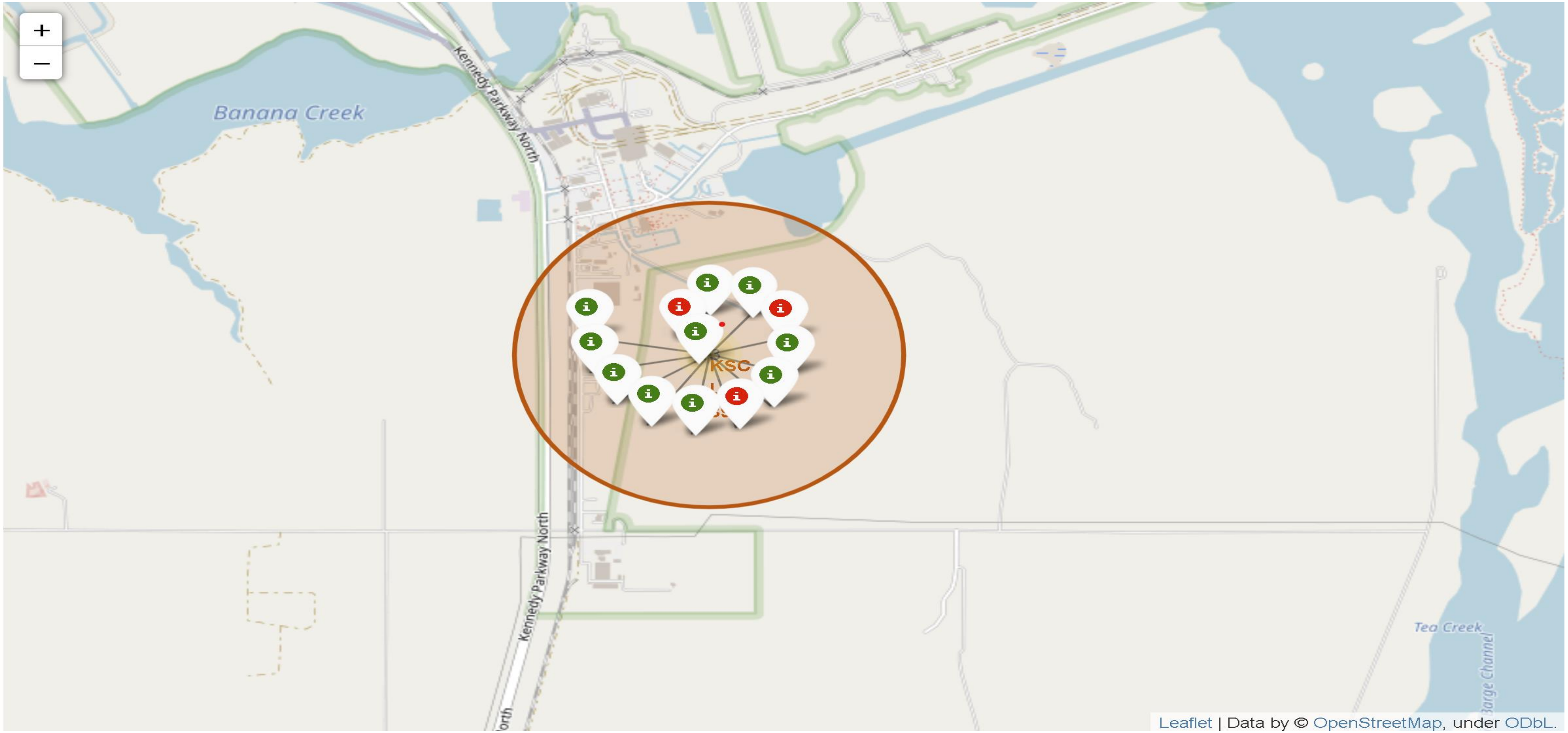
Location of SpaceX’s Falcon 9 Launch Sites



Location of SpaceX’s Falcon 9 Launch Sites

- Findings
  - There are four launch sites of SpaceX’s Falcon 9 rocket. All four are in United States – 3 in Florida and 1 in California. All the sites are near the coastline.
- Folium’s circle element and text icon have been used to mark the launch site location.

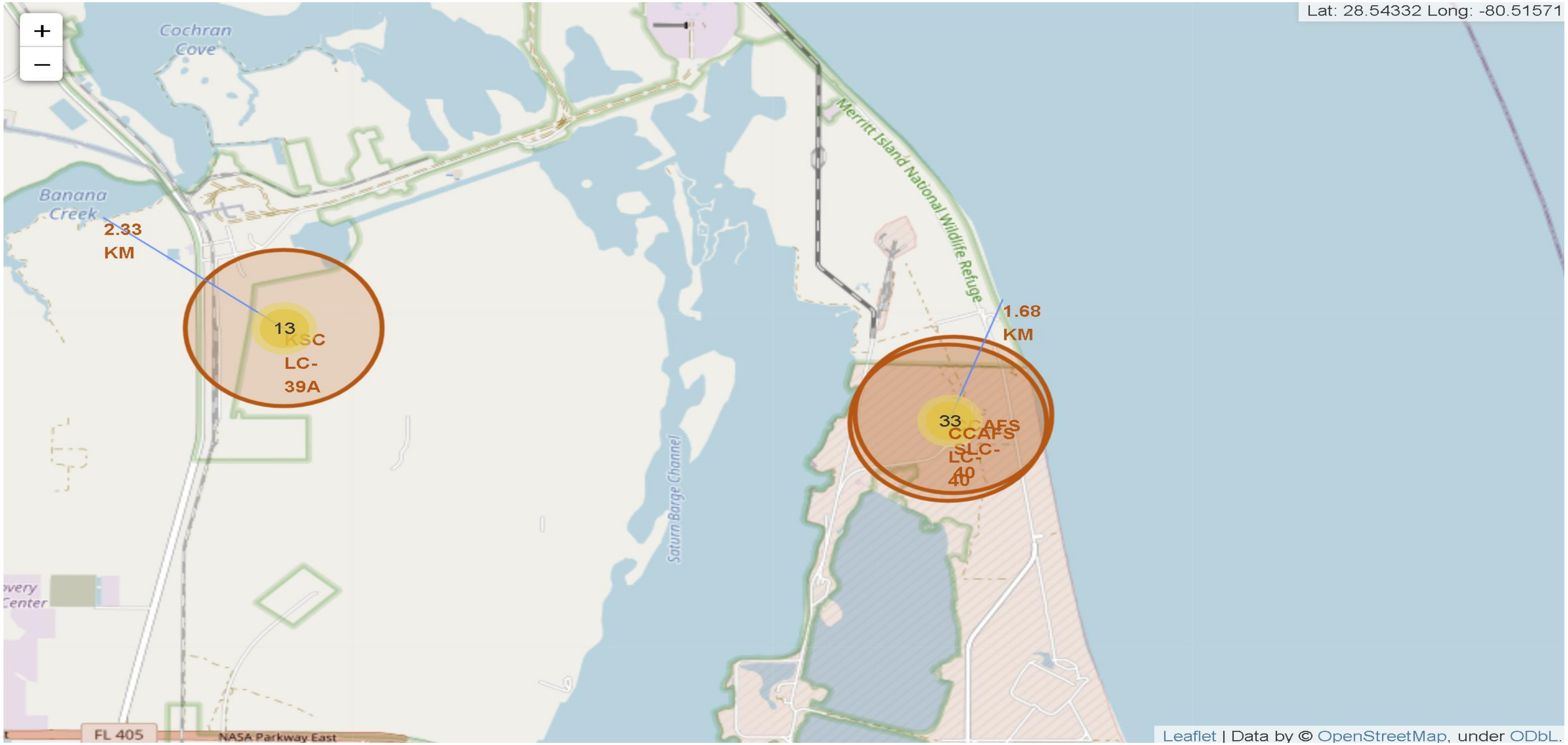
Color-labelled Launch Outcomes



Color-labelled Launch Outcomes

- Findings
  - Launch outcomes have been superimposed on top of the launch site location on the map. Outcomes are color-coded – red for a failure and green for success.
  - The launch site shown the image on the previous slide is KSC LC-39A.
  - We can see that there are more successful outcomes than failures on this launch site.
- Folium's marker cluster has been used to bunch all the outcome markers as the coordinates are the same for all the outcome markers for a particular launch site.

Distance of a Launch Site to a Landmark in Proximity



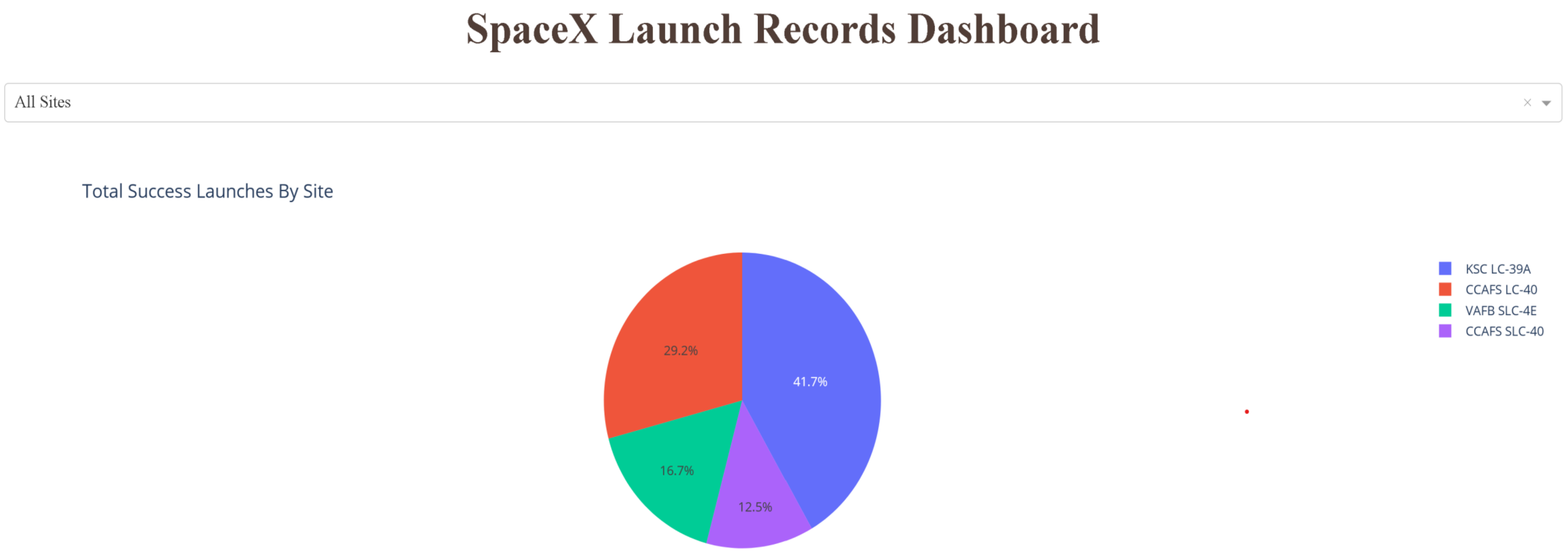
Distance of a Launch Site to a Landmark in Proximity

- Findings
  - The map shown in the previous slide shows the distance of the site
    - CCAFS LC-40 to coastline (1.68 Km)
    - KSC LC-39A to Banana Creek (2.33 Km)
- A poly line connecting the launch site to the proximity and a text marker showing the distance was added to the map.

Section 4

Build a Dashboard with Plotly Dash

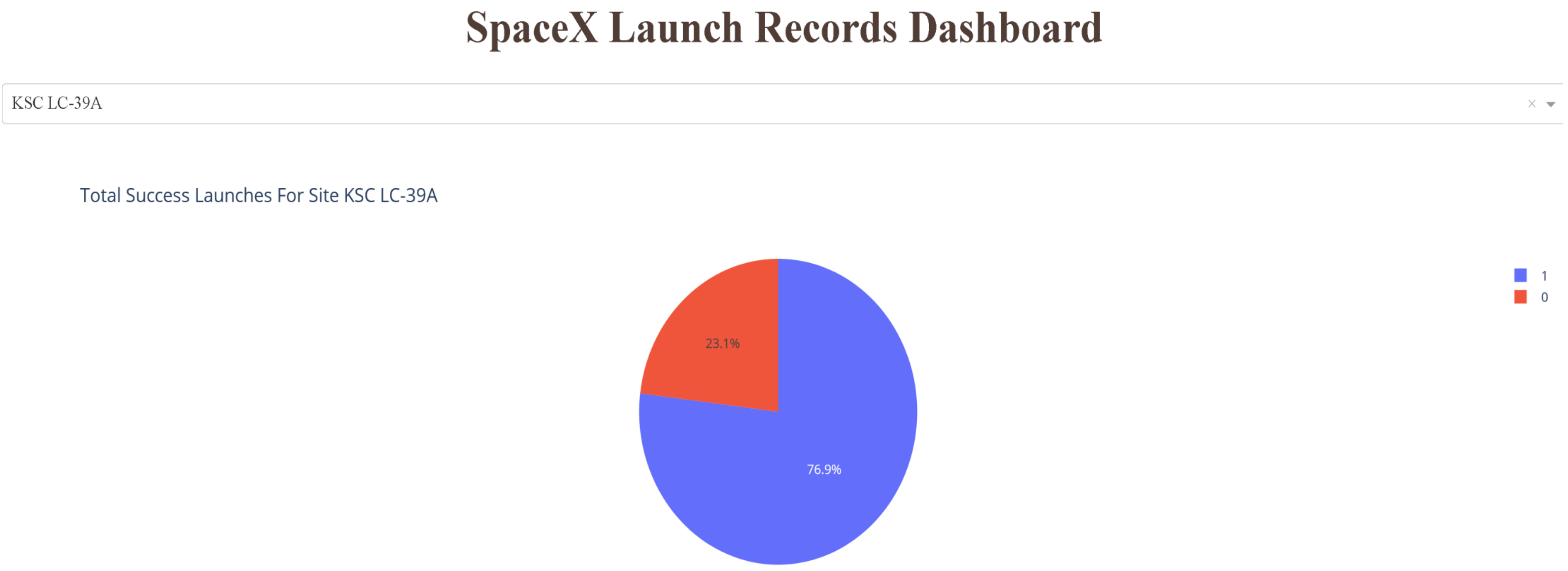
Launch Success Count For All Sites



Launch Success Count For All Sites

- Findings
  - We can see among all the launch sites, KSC LC-39A has the maximum number of launches at 41.7% of the total launches.
  - At the other extreme, CCAFS SLC-40 launch site has the minimum number of launches at 12.5% of the total.

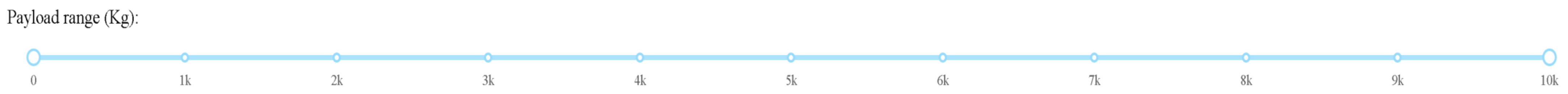
Launch Site With The Highest Success Ratio

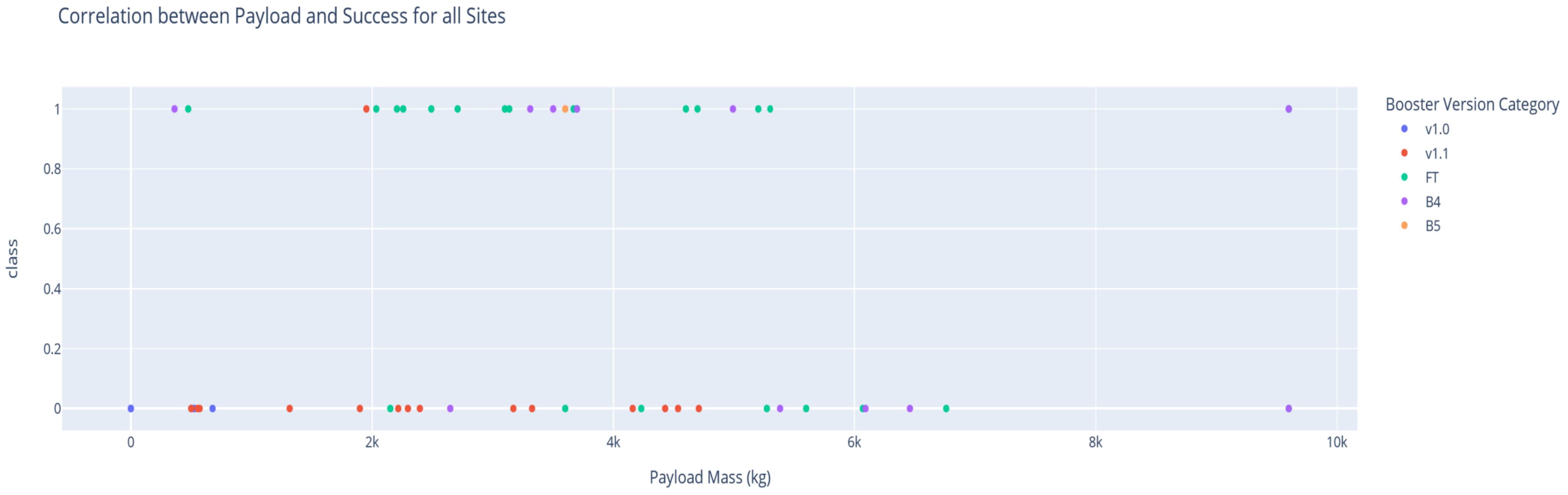


Launch Site With The Highest Success Ratio

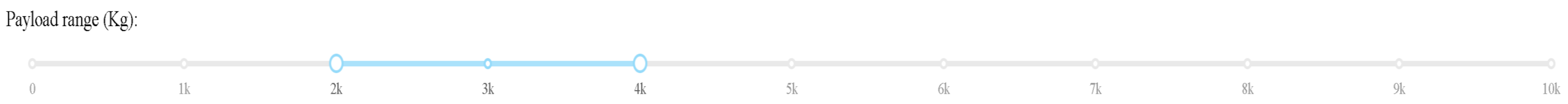
- Findings
  - Site KSC LC-39A has the highest success ratio at 76.9%. This site has the maximum number of launches out of all the launch site as well.
  - Next best is the site CCAFS LC-40 at 73.1%.

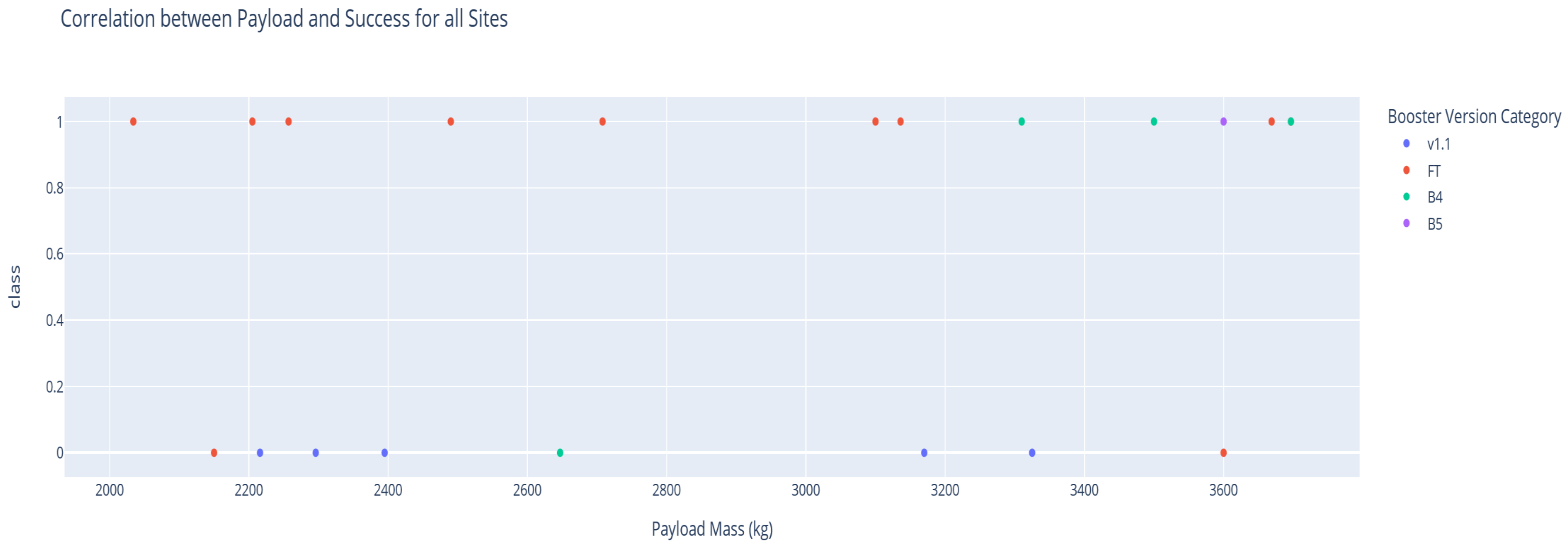
Payload vs Launch Outcome – Maximum Range





Payload vs Launch Outcome – 2k to 4k





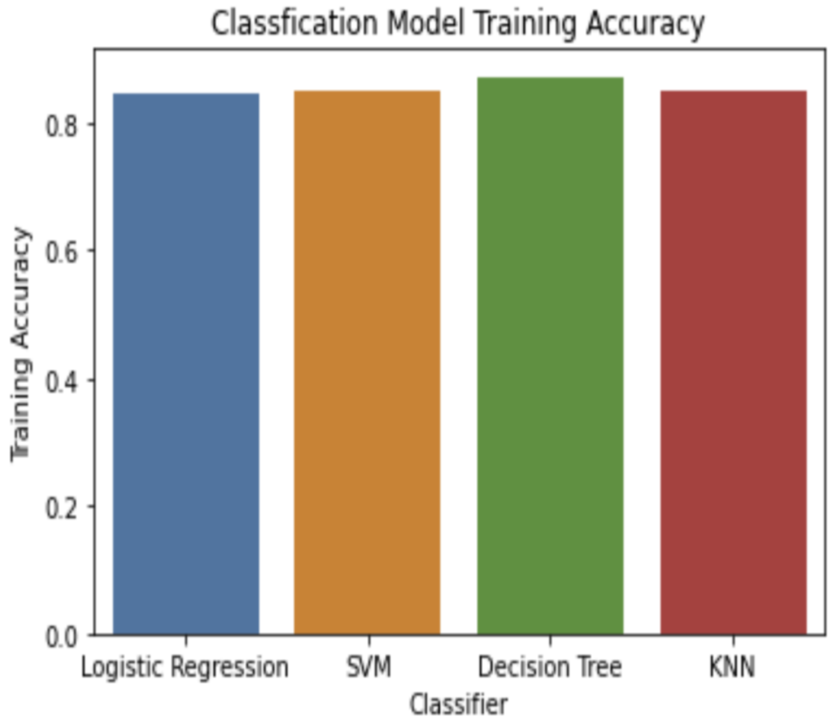
Payload vs Launch Outcome

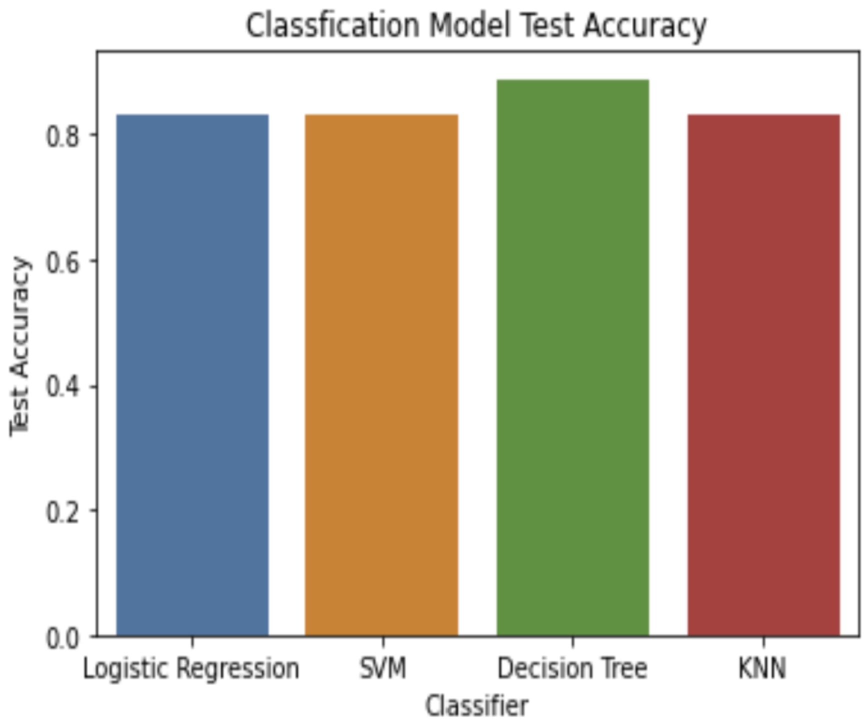
- Findings
  - From the chart ‘Payload vs Launch Outcome – Maximum Range’, we can see that the most number of successful launches are for the Payload mass range of 2000 kg to 4000 kg.
  - From the chart ‘Payload vs Launch Outcome – 2k to 4k’, we can see that for the 2k to 4k payload mass range, Booster version ‘FT’ has the highest success. From the first chart, we can see that ‘FT’ booster version has the highest success in the maximum range of 0 to 10k as well.

Section 5

Predictive Analysis (Classification)

Classification Accuracy





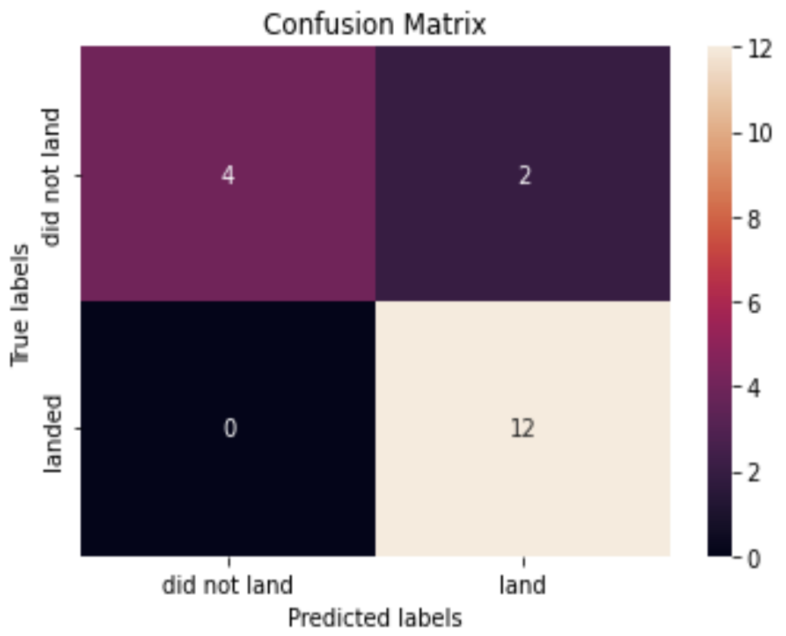
Classification Accuracy

	Classifier	Training Accuracy	Test Accuracy
0	Logistic Regression	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Decision Tree	0.871429	0.888889
3	KNN	0.848214	0.833333

- Decision Tree classifier has the highest accuracy on the test data – 0.89. It also has the highest training set accuracy of 0.87.

Confusion Matrix of Decision Tree Classifier

- Here is the confusion matrix of the decision tree classifier which had the highest accuracy score among all the classifiers.
- We can see that it has classified most of the outcomes correctly – both negative and positive.
- Only 2 outcomes whose true value was ‘did not land’ were misclassified as ‘landed’.



Conclusions

Conclusions

- Key conclusions are:
  - Success rate has gone up with experience as the time has passed.
  - Maximum number of successful launches have been for Payload in the range 2000 kg to 4000 kg.
  - FT booster version has the highest success rate across the whole range of Payload mass.
  - Based on the analysis exercise, we can conclude that sites KSC LC-39A and CCAFS LC-40 have the most number of successful launches – 41.7% and 29.2% of the total launches for Falcon 9. These sites also have the best success rate for the launches from them. So, if we are evaluating where to launch from in order to have the best chance of first stage coming back successfully, these are the sites aim for.
  - Decision tree classification model is the best prediction model to predict whether the first stage will land successfully or not.

Appendix

- Appendix A – Data Description
- Appendix B – SQL Queries used in Exploratory Data Analysis

Appendix A

Data Description

Data Description

- Data set contains data for Falcon 1 and Falcon 9. We will focus our analysis on Falcon 9 only. Data will be filtered accordingly.
- The data contains several Space X launch facilities:
  - Cape Canaveral Space Launch Complex 40 VAFB SLC 4E ,
  - Vandenberg Air Force Base Space Launch Complex 4E (SLC-4E),
  - Kennedy Space Center Launch Complex 39A KSC LC 39A .
- Data contains launch data from year 2010 till 2020.

Data Description

- There are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident. Various value of mission outcome are as follows:
    - ‘True Ocean’ means the mission outcome was successfully landed to a specific region of the ocean while
    - ‘False Ocean’ means the mission outcome was unsuccessfully landed to a specific region of the ocean.
    - ‘True RTLS’ means the mission outcome was successfully landed to a ground pad
    - ‘False RTLS’ means the mission outcome was unsuccessfully landed to a ground pad.
    - ‘True ASDS’ means the mission outcome was successfully landed on a drone ship
    - ‘False ASDS’ means the mission outcome was unsuccessfully landed on a drone ship.

Data Description

- Each launch aims to achieve a dedicated orbit. Here are some common orbit types:
  - LEO: Low Earth orbit (LEO)is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects in outer space are in LEO [1].
  - VLEO: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation[2].
  - GTO A geosynchronous orbit is a high Earth orbit that allows satellites to match Earth's rotation. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot for monitoring weather, communications and surveillance. Because the satellite orbits at the same speed that the Earth is turning, the satellite seems to stay in place over a single longitude, though it may drift north to south," NASA wrote on its Earth Observatory website [3] .

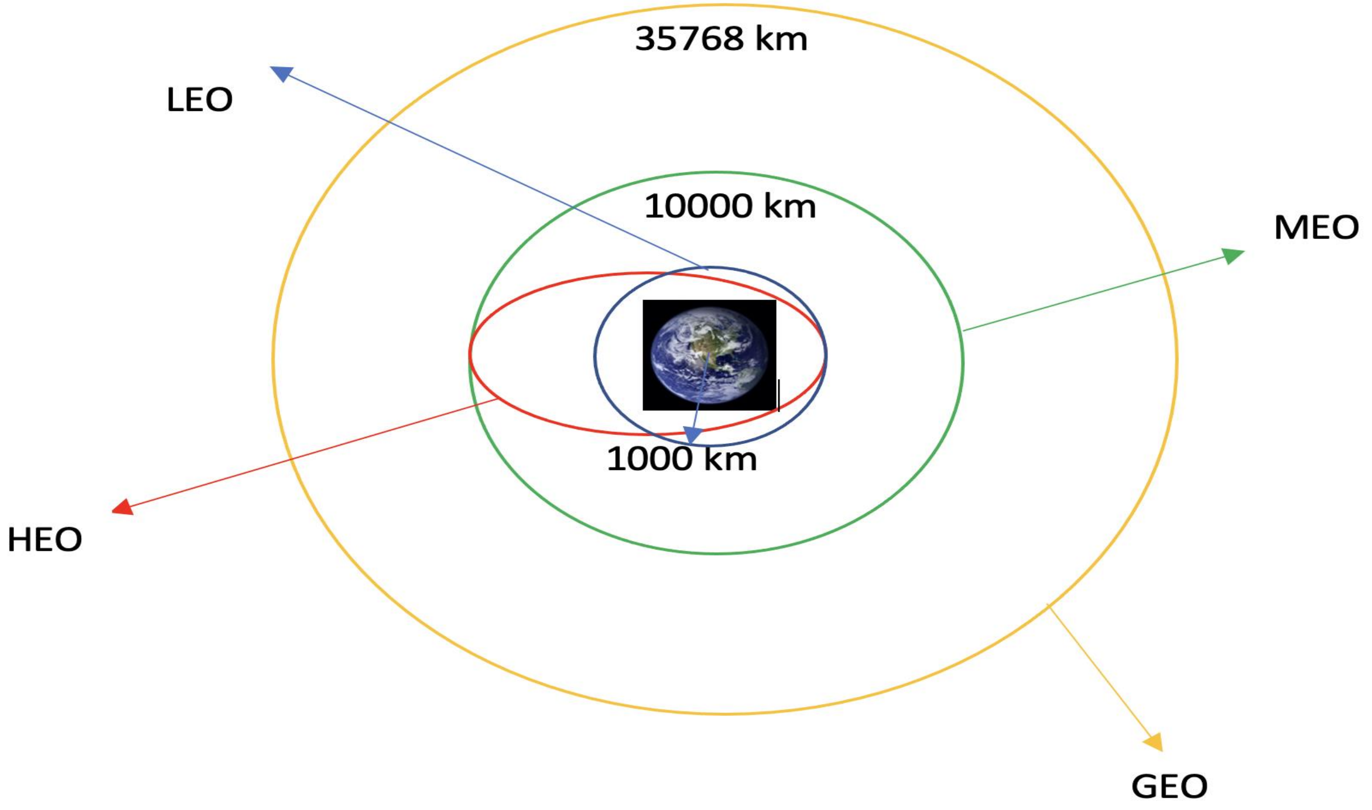
Data Description

- SSO (or SO): It is a Sun-synchronous orbit also called a heliosynchronous orbit is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time [4] .
- ES-L1 :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth [5] .
- HEO A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth [6].
- ISS A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada) [7]

Data Description

- MEO Geocentric orbits ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi). Also known as an intermediate circular orbit. These are "most commonly at 20,200 kilometers (12,600 mi), or 20,650 kilometers (12,830 mi), with an orbital period of 12 hours [8]
- HEO Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 mi) [9]
- GEO It is a circular geosynchronous orbit 35,786 kilometres (22,236 miles) above Earth's equator and following the direction of Earth's rotation [10]
- PO It is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usually a planet such as the Earth [11]

Data Description - Orbit Types



Data Description

- Some other key features are
    - Grid Fins: these help with landing Reused,
    - Legs: used in landing Landing pad,
    - Block,
    - Reused count,
    - Serial,
    - Longitude and latitude of launch

Appendix B

SQL Queries used in Exploratory Data Analysis

SQL Queries used in Exploratory Data Analysis

- All Launch Site Names

```
select distinct launch_site as "Launch Site“
from spacextbl
```

- Launch Site Names Begin with 'CCA'

```
select * from spacextbl
where launch_site like 'CCA%’
limit 5
```

SQL Queries used in Exploratory Data Analysis

- Total Payload Mass

```
select sum(payload_mass__kg_) as "Total Payload“,
          count(*) as “Total Number of Flights”
from spacextbl
where customer = 'NASA (CRS)’
```

- Average Payload Mass by F9 v1.1

```
select avg(payload_mass__kg_) as "Average Payload"
from spacextbl
where booster_version = 'F9 v1.1’
```

SQL Queries used in Exploratory Data Analysis

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

```
select substr(date, 7, 4)||'-'||substr(date, 4, 2)||'-'||substr(date, 1, 2) as "Landing Date",
        "Landing _Outcome" as "Landing Outcome“
from spacextbl
where "Landing _Outcome" = 'Success (ground pad)’
order by 1 limit 1
```

- Successful Drone Ship Landing with Payload between 4000 and 6000

```
select booster_version as "Booster",
        payload_mass__kg_ as "Payload Mass“
from spacextbl
where "Landing _Outcome" = 'Success (drone ship)’
  and payload_mass__kg_ between 4000 and 6000
```

SQL Queries used in Exploratory Data Analysis

- Total Number of Successful and Failure Mission Outcomes

```
select trim(mission_outcome) as "Mission Outcomes",
        count(*) as "Total“
from spacextbl
group by trim(mission_outcome)
```

- Boosters Carried Maximum Payload

```
select booster_version as "Booster Version",
        payload_mass__kg_ as "Payload Mass“
from spacextbl
where payload_mass__kg_ =
   (select max(payload_mass__kg_) from spacextbl)
order by 1
```

SQL Queries used in Exploratory Data Analysis

- 2015 Launch Records

```
select  substr(date, 7, 4) Year,
        case substr(date, 4, 2)
           when '01' then 'Jan'  when '02' then 'Feb' when '03' then 'Mar'
           when '04' then 'Apr' when '05' then 'May' when '06' then 'Jun'
           when '07' then 'Jul' when '08' then 'Aug' when '09' then 'Sep'
           when '10' then 'Oct' when '11' then 'Nov' when '12' then 'Dec'
           else 'Invalid Month'
        end as Month,
        "Landing _Outcome" as "Landing Outcome",
        booster_version as "Booster Version",
        launch_site as "Launch Site"
from spacextbl
where "Landing _Outcome" = 'Failure (drone ship)'
  and substr(date, 7, 4) = '2015'
```

SQL Queries used in Exploratory Data Analysis

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

```
select  rank() over (order by count(*) desc) Rank,
        "Landing _Outcome" as "Landing Outcome",
        count(*) "Number of Successful Landings“
from  spacextbl
where "Landing _Outcome" like '%Success%’
  and substr(date, 7, 4)||'-'||substr(date, 4, 2)
      ||'-'||substr(date, 1, 2)
      between '2010-06-04' and '2017-03-20’
group by "Landing _Outcome“
```

Thank you!