Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via API and Web-scrapping separately
  - Data Wrangling
  - Exploratory Data Analysis (EDA) using SQL as well with Data Visualization
  - Interactive Map with Folium and Dashboards with Plotly Dash
  - Predictive Analysis using Machine Learning (ML) Algorithm
- Summary of all results
  - Data collected from public source
  - EDA informed us which features are best to predict the success of the launch
  - ML algorithms informed us the best model to predict ideal conditions for the success.

Introduction

- Project background and context
  - The objective of the project is to evaluate the viability of new company SpaceY to compete with SpaceX
  - In this capstone project, I will take the role of a data scientist working for a new rocket company. Space Y that would like to compete with SpaceX. My job is to determine the price of each launch. I will do this by gathering information about Space X and creating dashboards. I will also determine if SpaceX will reuse the first stage. I will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
  - Best way to estimate the total cost for launches, by predicting first stage of rockets
  - Best launching area for the company.

Section 1

Methodology

Methodology

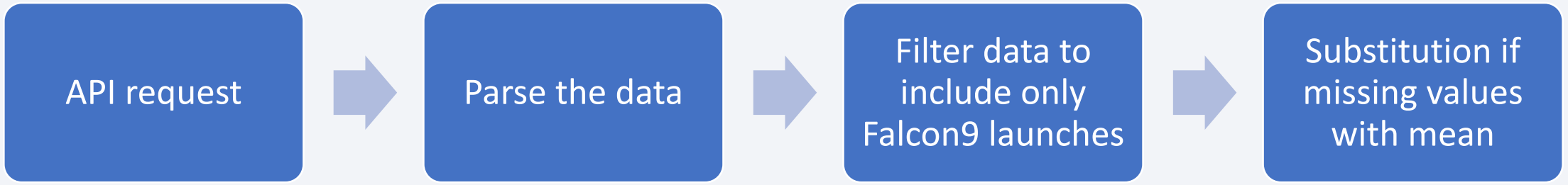
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and by web. Scrapping of Wikipedia page of falcon9 and falcon other vehicles.
  - Perform data wrangling
    - One-hot coding was performed to drop irrelevant columns.
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - build, tune, evaluate classification models like logistic regression, KNN, SVM, Decision Tree Models etc.

Data Collection

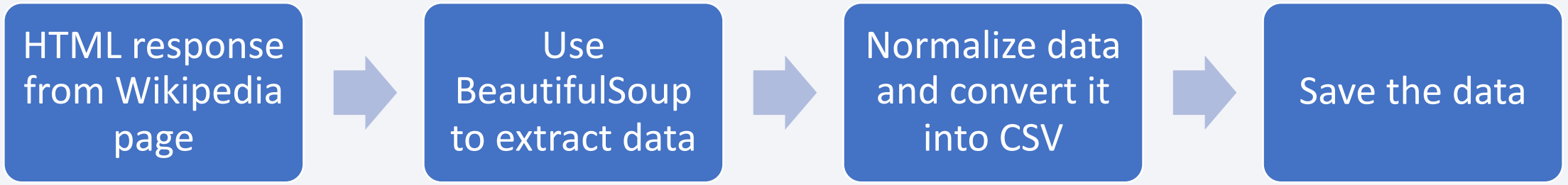
- **The Data**
  - API Method: I make a get request to the SpaceX API. Request and parse the SpaceX launch data using the GET request
    - Source: (https://api.spacexdata.com/v4/rockets/)
  - Webscraping method: Web scrap Falcon 9 launch records with BeautifulSoup: Extract a Falcon 9 launch records HTML table from Wikipedia Parse the table and convert it into a Pandas data frame
    - Source: (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- **Data Wrangling:**
  - Filter the dataframe to only include Falcon 9 launches
  - Dealing with Missing Values- replace nan values with the mean.

Data Collection – SpaceX API



- SpaceX offers public API where data can be collected and used for analysis.
- Steps such as returns such as data in JSON and normalization of data were taken.
- GitHub notebook: API Notebook

Data Collection - Scraping

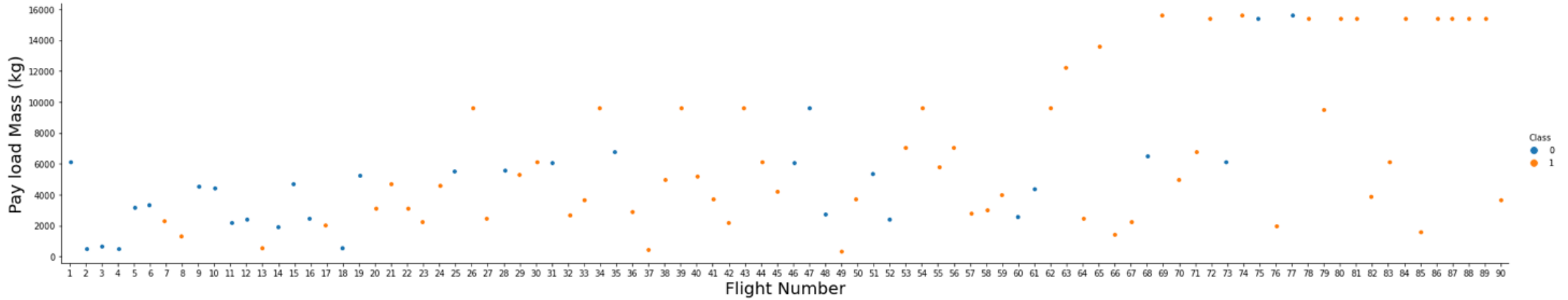


- Wikipedia page is used to obtained the data.
- BeautifulSoup object is created and table is extracted from the webpage.
- Data is normalized and DataFrame is created
- GitHub notebook: Scraping Notebook

Data Wrangling

- Exploratory Data analysis is done.
- Calculations such as number of launches on each site, occurrences on each orbit, success rate by calculating mean of landing outcome column.
- Landing outcome labeled from outcome column is created.
- Finally, the dataset is exported to CSV.
- GitHub notebook: Wrangling Notebook

EDA with Data Visualization



- We visualize the data using scatter graph, bar graph and line graph.
- Scatter Graph: Flight Number vs Payload Mass, Payload Mass vs Launch Site, Flight Number vs Orbit, Payload Mass vs Orbit
- Bar graph: Orbit vs Success rate
- Line graph: Year vs Success rate
- GitHub notebook: Visualization Notebook

EDA with SQL

- The Data table is manually loaded on db2 cloud and SQL queries were done on IBM Watson and is connected to jupyter notebook.
- The following SQL queries were done:
    - Getting the names of unique launch sites and Getting 5 records where launch sites begin with the character string 'CCA'
    - Calculate the total Payload Mass carried by boosters and Getting the date when the first successful landing outcome was achieved.
    - Getting the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - Getting the total number of successful and failure mission and Getting the names of booster versions which have carried maximum payload mass.
    - Getting the failed landing versions, their booster and launch sites for year 2005 and Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20.
- GitHub notebook: SQL Notebook

Build an Interactive Map with Folium

- Map objects such as markers, circles, lines, etc. created and added to a folium map
- Circles indicate highlighted areas around specific coordinates.
- For a successful launch, a green marker is placed and for unsuccessful a red marker is placed around the launch site.
- Diatnce from launch site to various landmarks like coastlines, city, highways and. Railways have been calculated.
- Lines are used to indicate distances between two coordinates.
- GitHub Notebook: Folium Notebook

Build a Dashboard with Plotly Dash

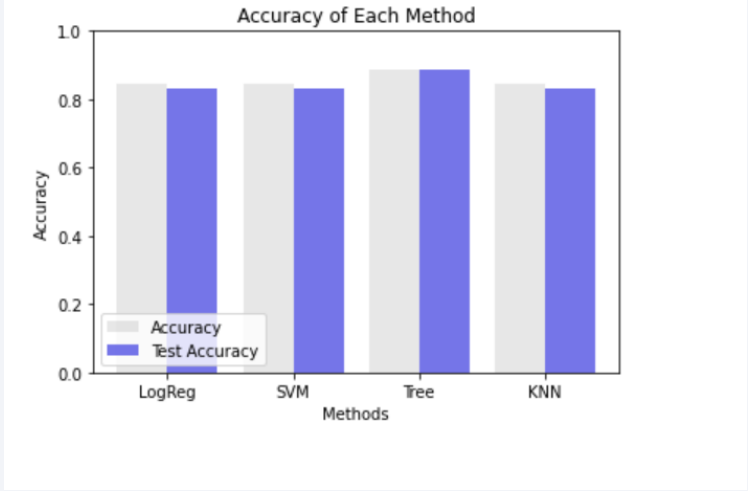
- Graphs and plots such as percentage of launches by site and payload range are used to visualize . The combination allowed to analyze the relation between payloads and launch sites.
- An interactive Dashboard was built using Plotly Dash where user can select launch sites ad payload range ad plot the results.
- Pie Chart is created showing success rate of all launch or for a particular launch site as per the selection of user.
- GitHub Notebook: Dashboard Notebook

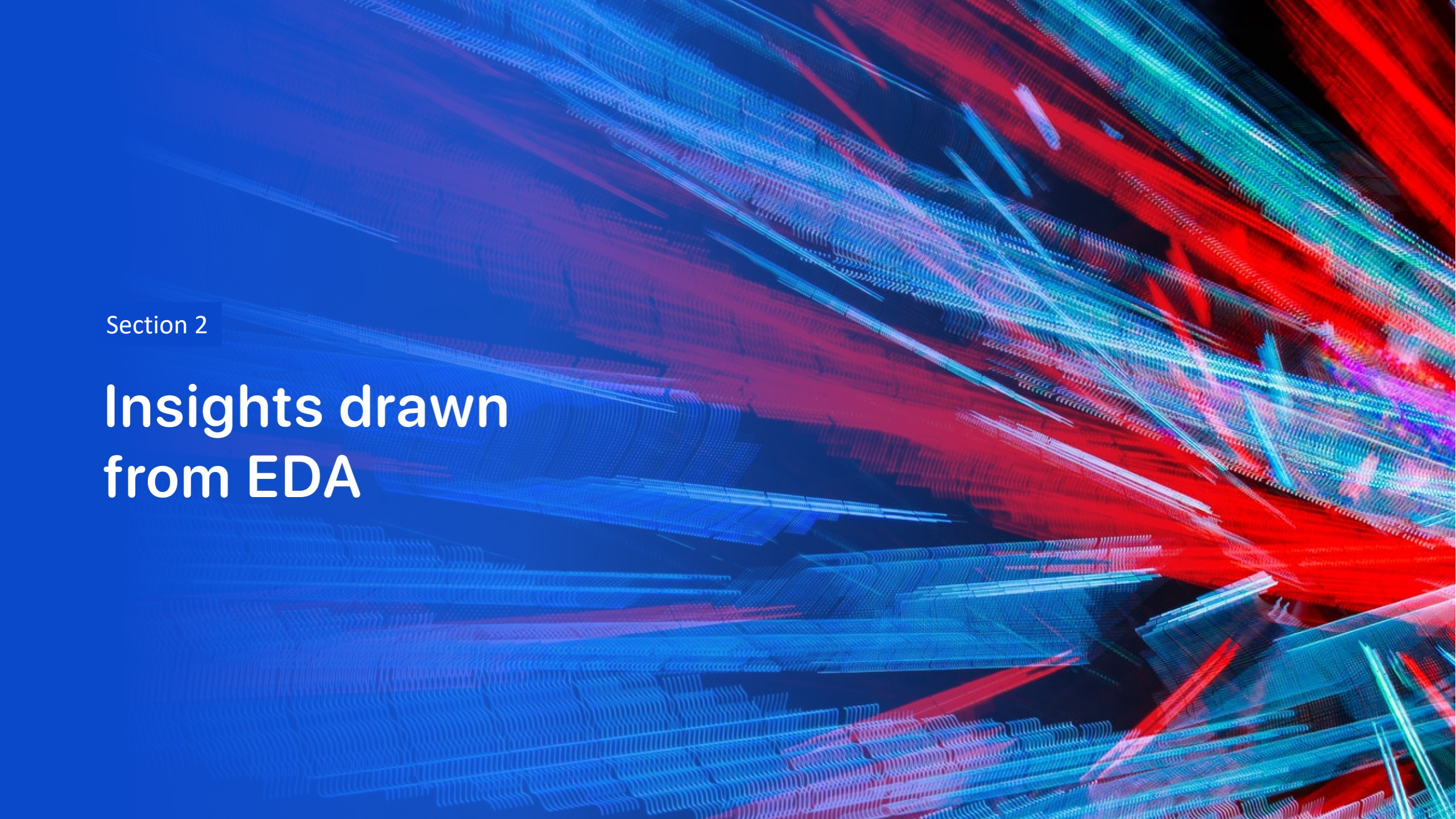
Predictive Analysis (Classification)

- Four classification models were compared: Logistic Regression, Support vector Machine (SVM) decision Tree and K-nearest neighbor (KNN)
- GridSearchCV is used to tune hyper parameters for the 4 types of model used to find the best parameters.
- Score of accuracy is measured for all the models and model with highest score is selected
- GitHub notebook: Prediction Notebook

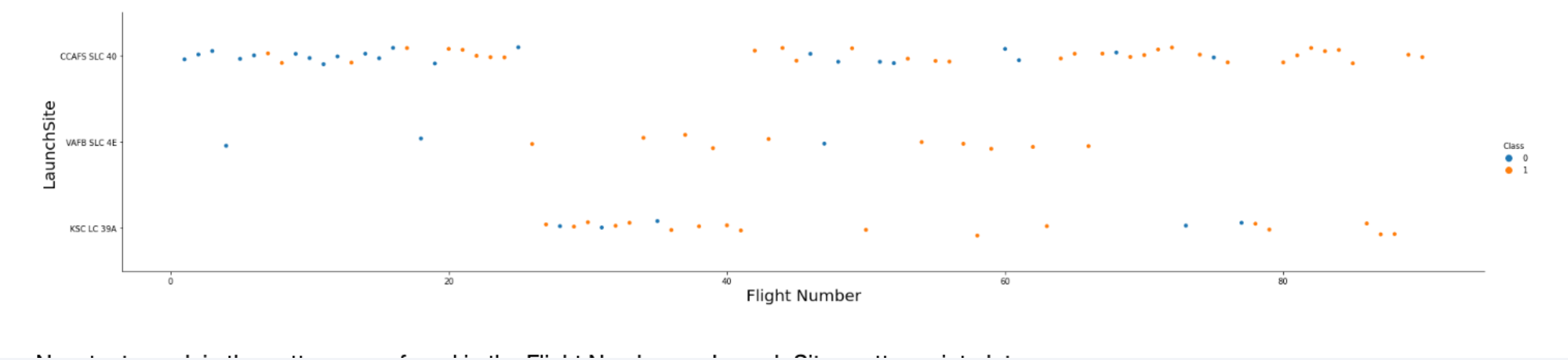
Results

- Exploratory data analysis results
    - SpaceX uses four different launch sites- CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E
    - The average payload of Falcon9 v1.1 booster is 2928 kg
    - The first success landing outcome was in 2015 fiver years after the first launch.
    - As years passed, the number of landing outcomes became better.
- Interactive analytics demo in screenshots-
    - Using Folium, it is identified that launch sites are near sea and have a good logistic infrastructure
    - Most launches happen at East coast launch sites.
- Predictive analysis results-
- The Decision Tree Classifier is the best model to predict successful landings, having accuracy over 88% and accuracy for test data over 88%.



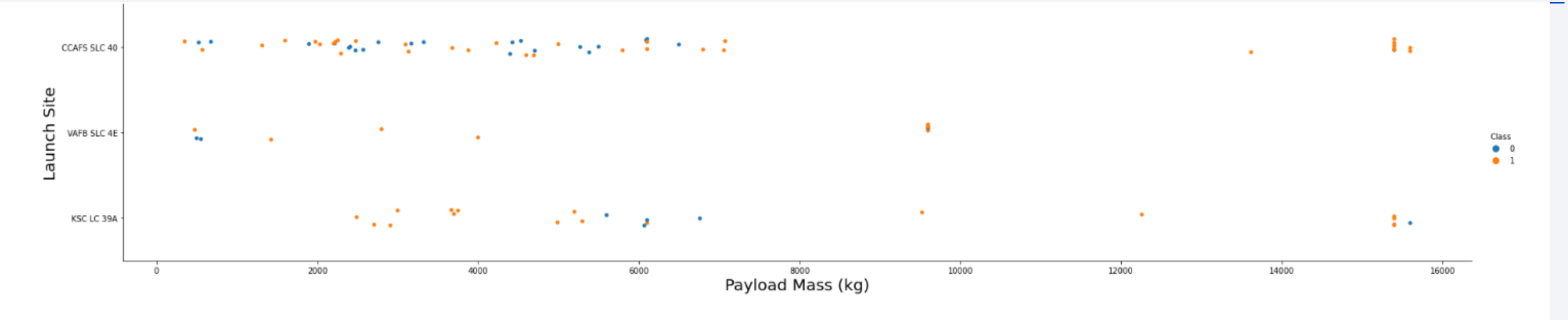

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Scatter plot of Flight Number vs. Launch Site.
  - In the scatter plot, we see that CCAF5 SLC 40 is the launc site where recent launches are successful.
  - VAFB SLC 4E is on second and KSC LC 39A on third.
  - Also, over time we see improvement in success outcome.

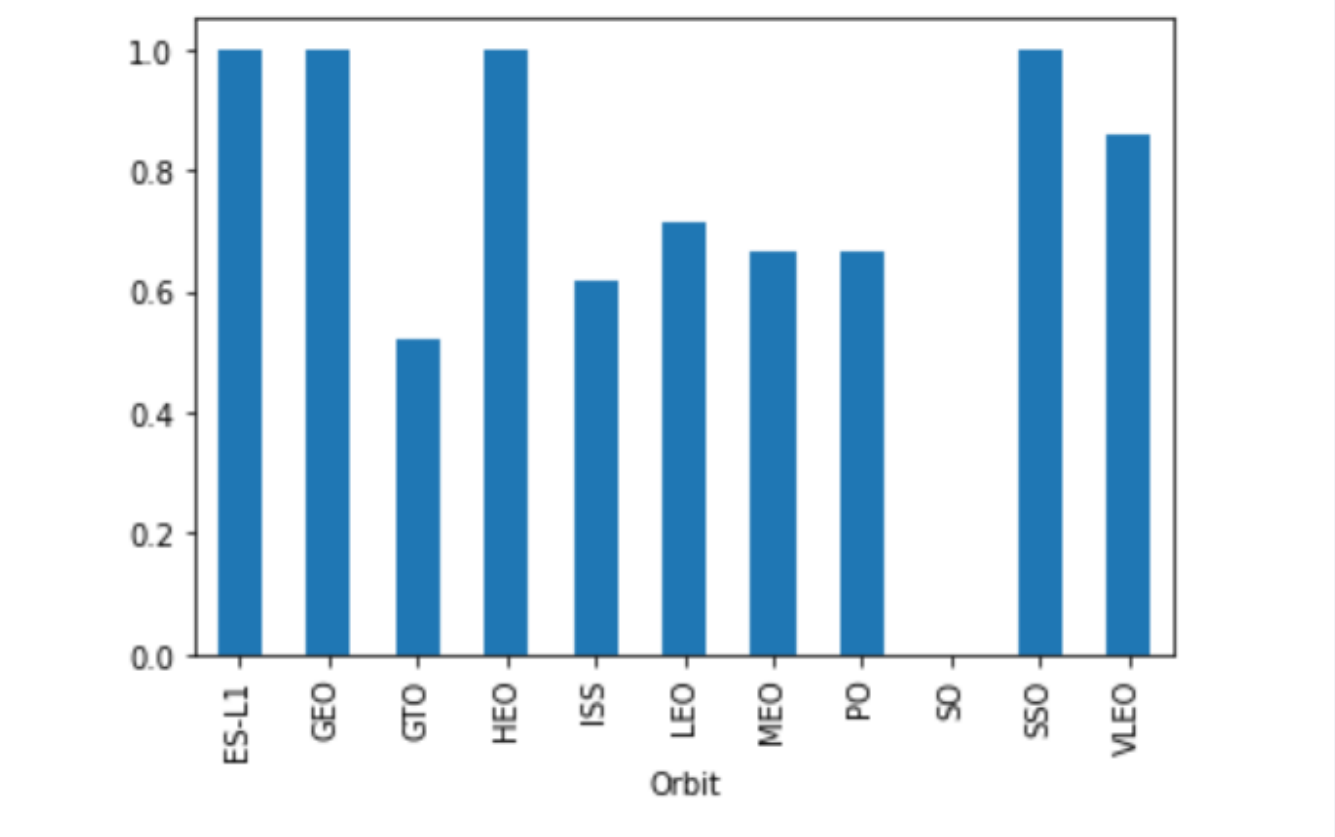
Payload vs. Launch Site



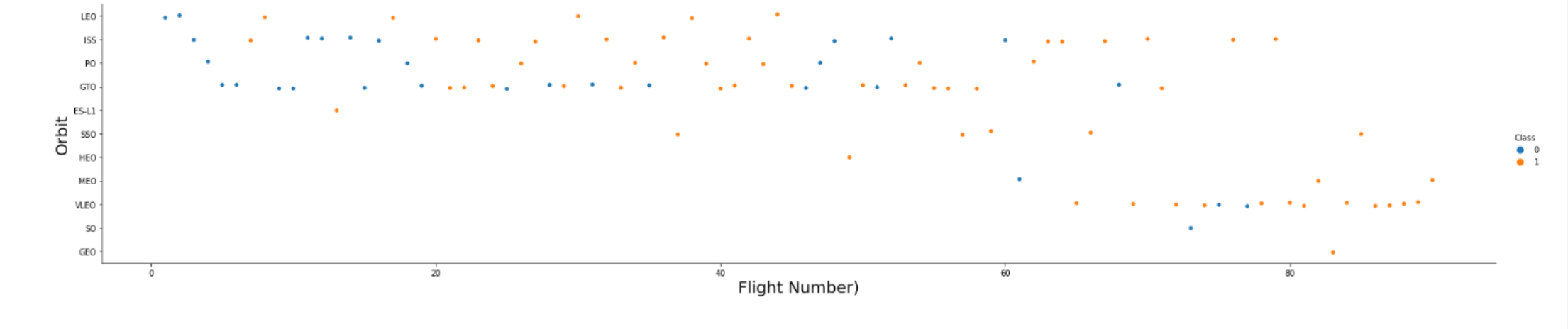
- Scatter plot of Payload vs. Launch Site
  - observing Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
  - Payloads over 9000 kg have excellent success rate
  - We can not come to a conclusion based on the visualization above.

Success Rate vs. Orbit Type

- The success rates of ES-L1, GEO, HEO and SSO are high followed by VLEO and LFO.
- Or in other words, orbits ES-L1, GEO, HEO and SSO have highest success rate.

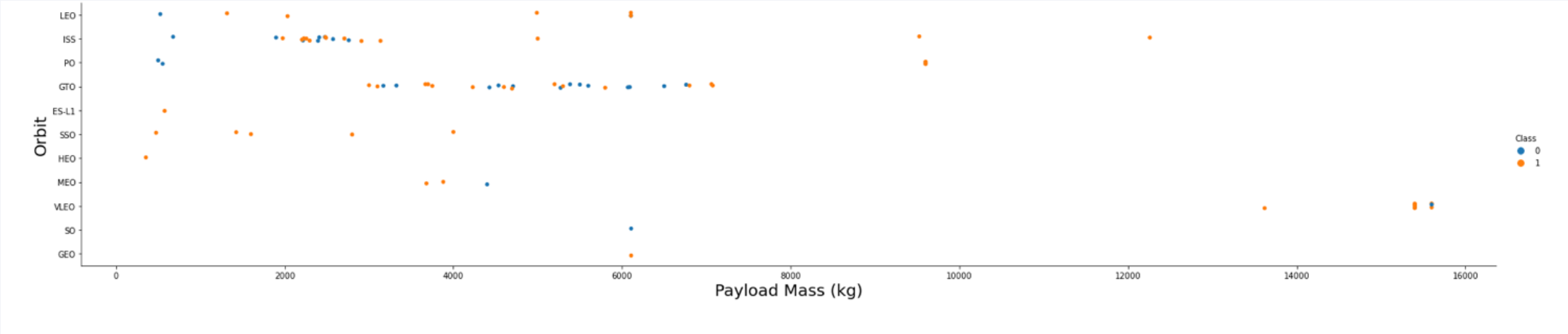


Flight Number vs. Orbit Type



- scatter point of Flight number vs. Orbit type
  - It looks like success rate improved over time in all orbits.
  - Launch frequency in VLEO orbit is increased in recent time.
  - For LEO orbit, the success rate increases with number of flights.

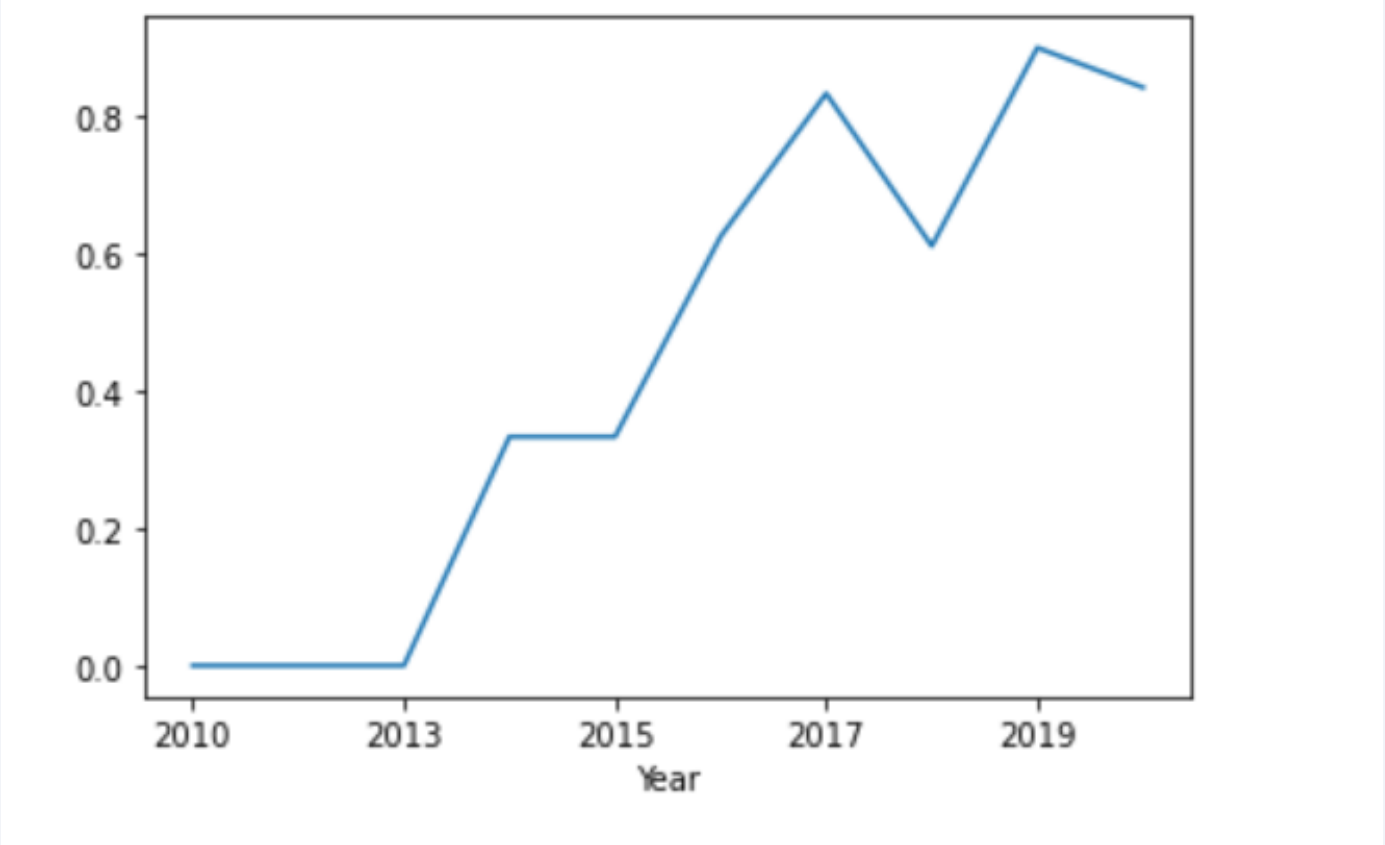
Payload vs. Orbit Type



- Scatter point of payload vs. orbit type
  - There are very few launches on SO and GEO orbits.
  - ISS orbit has the widest range of payload and a good rate of success.
  - Also, higher the payload mass for orbits LEO, ISS and PO, the more the success rate.

Launch Success Yearly Trend

- line chart of yearly average success rate-
  - the success rate since 2013 kept increasing till 2020
  - 2010 to 2013 period of stagnancy suggest the improvement in infrastructure and technology



All Launch Site Names

- Using SQL query, we find 4 unique launch sites.
- They are selected based on the unique occurrence in the data set

```
In [5]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1
 * ibm_db_sa://myg76123:***@55fbc997-9266-4331-afd3-888b05e734c0.bs
ud:31929/bludb
Done.
```

Out[5]:

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

Launch Site Names Begin with 'CCA'

```
[6]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * ibm_db_sa://myg76123:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

ut[6]:

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

- 5 records where launch sites begin with `CCA` are in the above table.
- I used SQL query where clause to get all the launch sites that begin with ‘CCA’ and used limit 5 for results.

Total Payload Mass

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

```
 * ibm_db_sa://myg76123:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.
ud:31929/bludb
Done.
```

total_payload
111268

- total payload carried by boosters from NASA is 111268
- I used SQL query and used where clause to calculate the total payload.

Average Payload Mass by F9 v1.1

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

```
[8]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
 * ibm_db_sa://myg76123:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appd
ud:31929/bludb
Done.
```

:[8]:

avg_payload
2928

- the average payload mass carried by booster version F9 v1.1 is calculated suing SQL query and is 2928 kg. I filtered the data by booster version and calculated the average mass.

First Successful Ground Landing Date

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

Hint:Use min function

```
[9]: %sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
 * ibm_db_sa://myg76123:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.
ud:31929/bludb
Done.
```

t[9]:

first_success_gp
2015-12-22

- I find out the first successful landing outcome on ground pad using the SQL query using min function and where clause, The first successful landing outcome in ground pas was achieved on 22nd December 2015.

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 SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_
```

```
 * ibm_db_sa://myg76123:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.clo
ud:31929/bludb
Done.
```

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

- The above table contains the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- I use SQL query and where clause to arrive at this list.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://myg76123:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

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

- I calculate the total number of successful and failure mission outcomes as given in the above table
- I use SQL query and where clause to arrive at this list. I group mission outcomes an counting records for each group

Boosters Carried Maximum Payload

- The table contains the list the names of the booster which have carried the maximum payload mass
- I use SQL query and where clause to arrive at this list.

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

2015 Launch Records

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

- The list of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are given above.
- I use SQL query and where clause to arrive at this list.
- The list has only two occurrences of failure.

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

- The Ranking 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 is given above.

- I use SQL query and where clause to arrive at this list.
- ‘No attempt’ has the highest quantity of 10

landing__outcome	qty
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 3
Launch Sites
Proximities Analysis

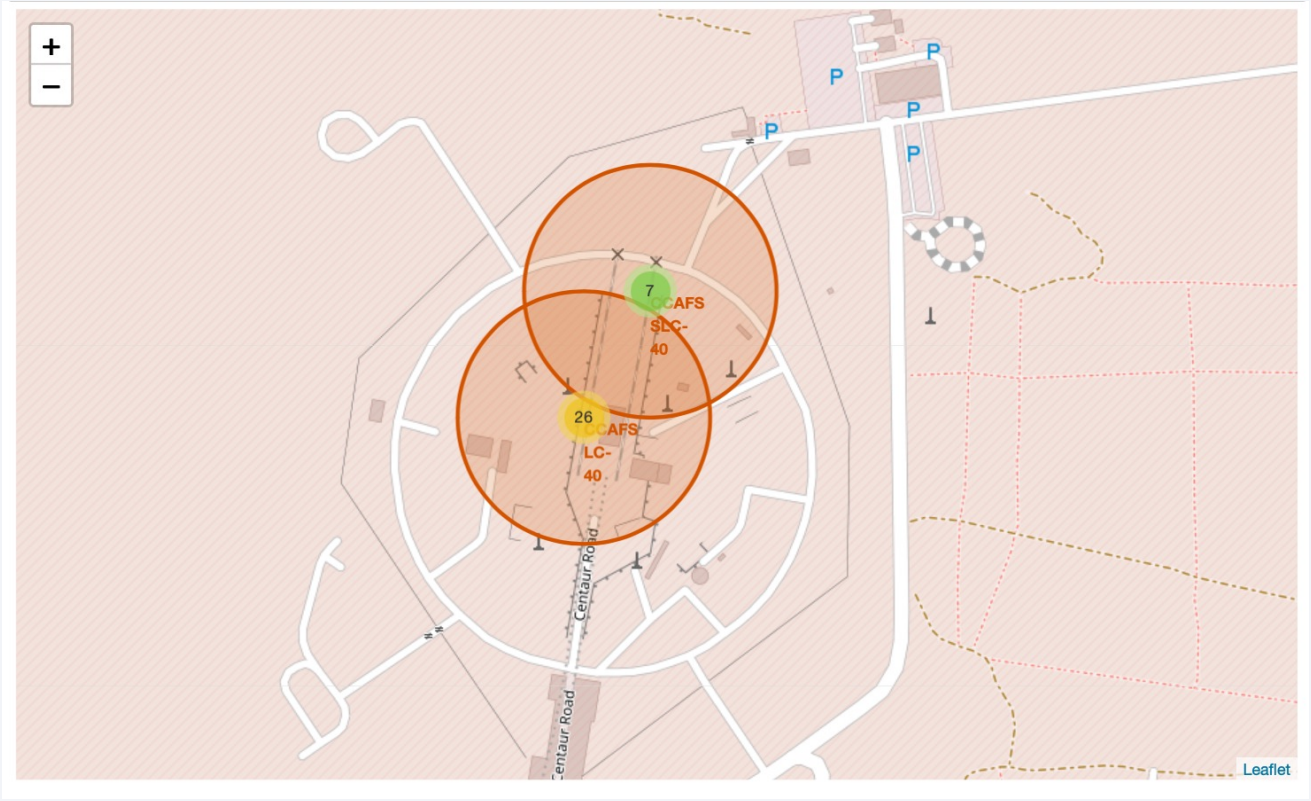
Folium Map: Sites on a global map

- Marking all the launch sites on global map.
- Explain the important elements and findings on the screenshot
- The launch sites are at east and west coast of USA.
- Possibly due to safety reason, launch sites are at coast.

Out[8]:

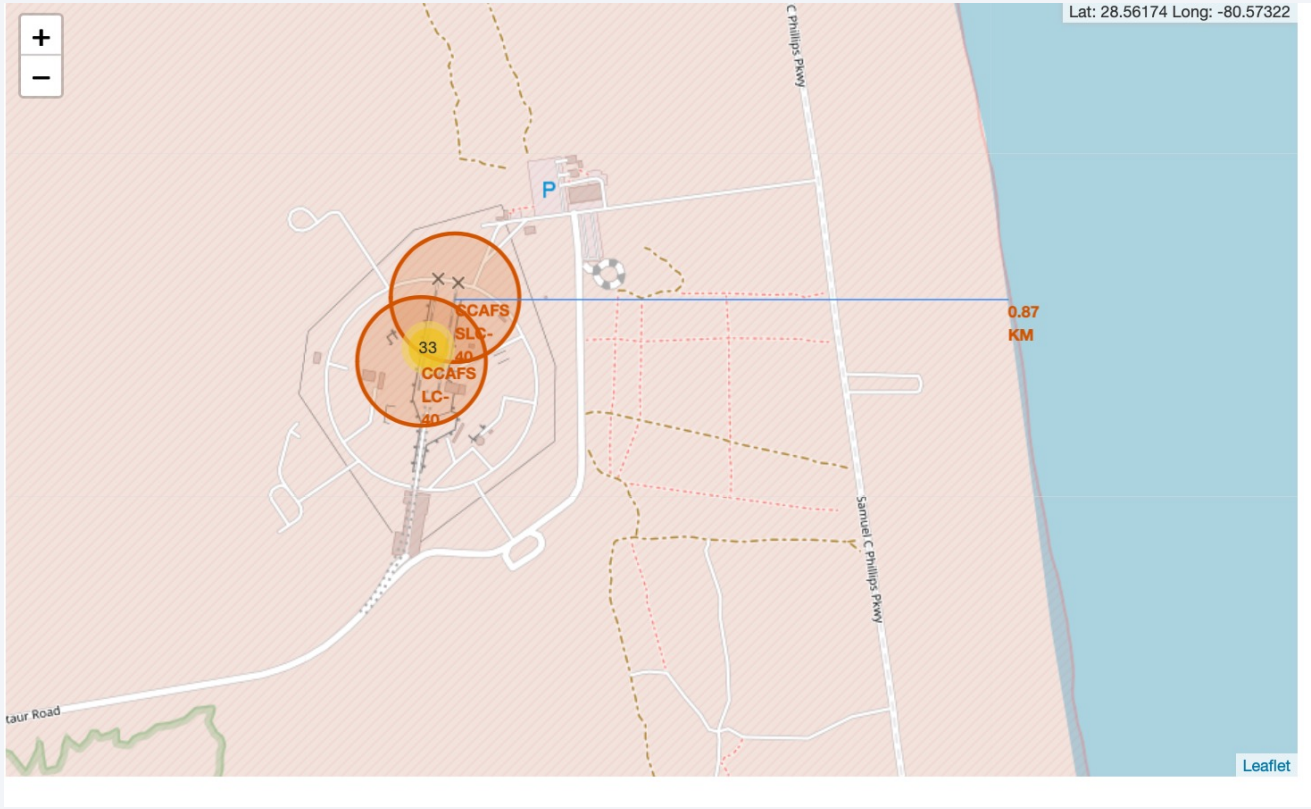
Launch Outcomes by Site

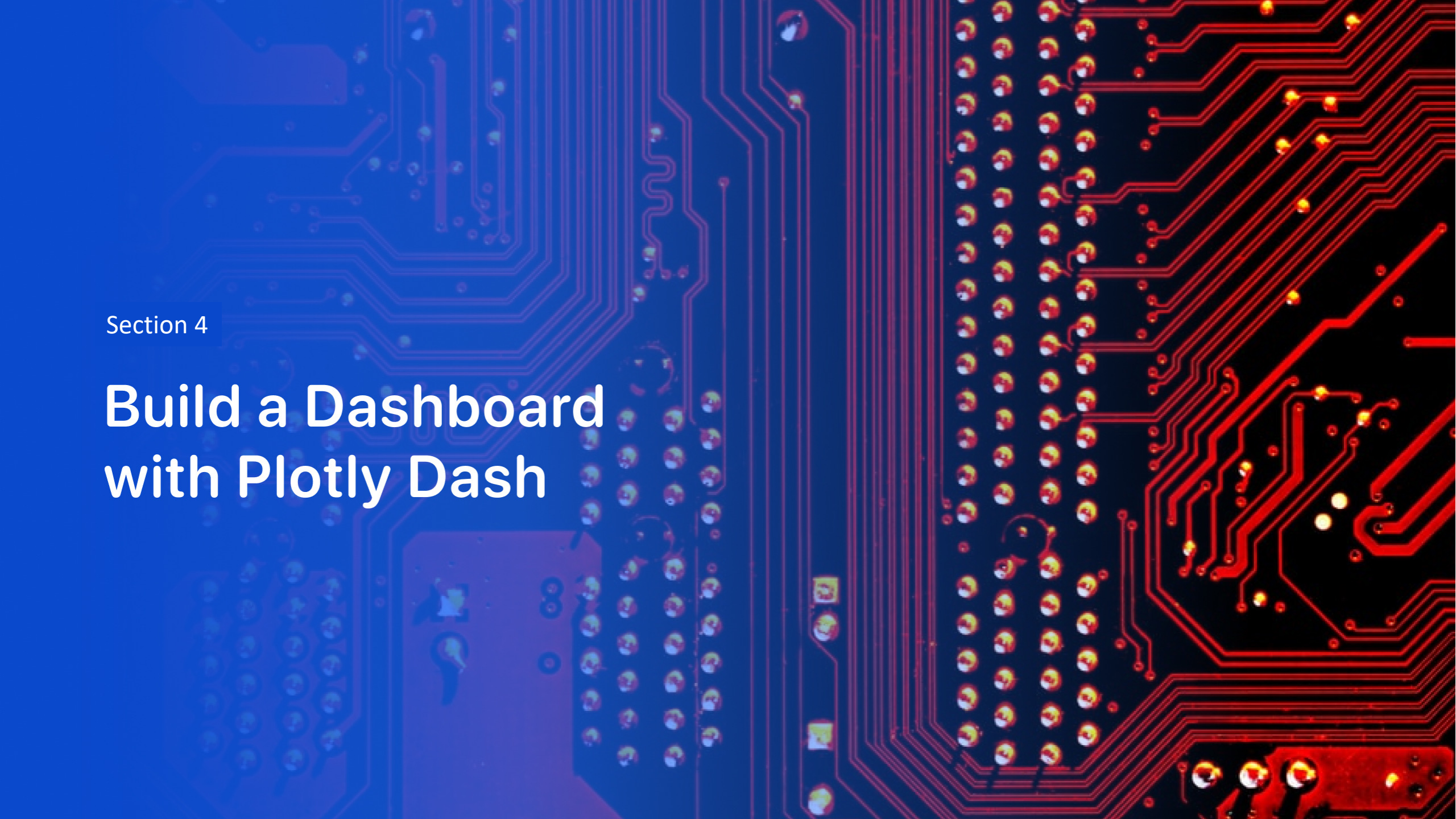
- Color-labeled launch outcomes on the map
- As we can see, at CCAFS SLC-40 site , I have placed a green marker for class =1, i.e. successful outcome.



Launch Site proximities.

- the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed.
- In the screenshot, violet color line indicates the distance of CCAFS SLC-40 site from coastline.




Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

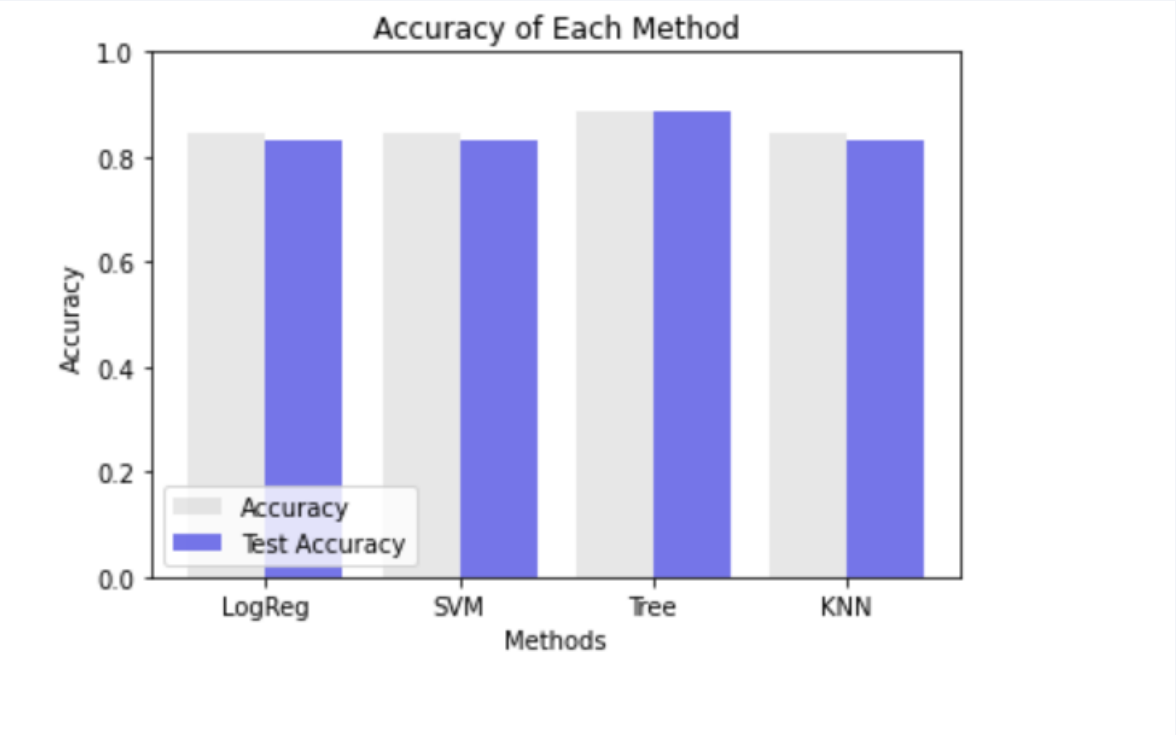
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Visualization of the built model accuracy for all built classification models, in a bar chart is given on the right side

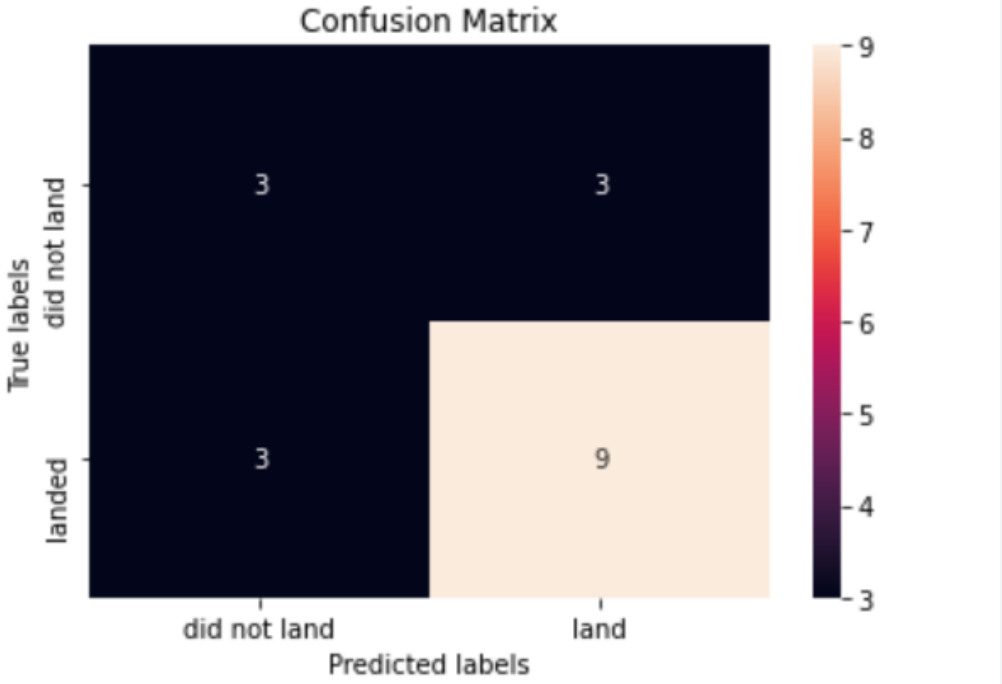
- The model which has the highest classification accuracy is the Decision Tree classifier with accuracy of 87%



Confusion Matrix

- The confusion matrix of the best performing model (which is Decision Tree Classifier) shows high value of true positive and true negative compared to false ones.

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- The success of Launch started to increase from the year 2013 till 2020. That is the success rate started to increase.
- KSC LC-39A has higher success rate compared to rest of the launch sites.
- Lower payloads have higher success rate when compared to higher payloads
- Launch sites are near coastlines
- The decision Tree classifier is the best machine learning algorithm for the given data.

Thank you

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!