IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<**Said E.**/ SpaceY Company>
<16.07.22>

Outline

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

https://github.com/sef-ffm/ibm_ds_capstone

Executive Summary

- **Summary of methodologies**
  - We collect data, as much and relevant as possible from various sources (Webscraping, API)
  - After raw data has been collected, we need to improve the quality by performing data wrangling
  - We start exploring the processed data using SQL, visualizations (Folium) and basic statistical analysis
  - We'll drill down by splitting the data into groups defined by categorical variables or factors in the data
  - We build, evaluate, and refine predictive models for discovering more exciting insights
- **Summary of all results**
  - We present the results of our exploratory data analysis
  - We present some interactive analytics and visualisations
  - We build a model to predict the landing of the first stage of the Falcon 9 rocket with a high level of accuracy

https://github.com/sef-ffm/ibm_ds_capstone

Introduction

- **Project background and context**
  - Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each. Much of the savings is because Space X can reuse the first stage of the Falcon 9 rocket.
  - SpaceY is in a bidding contest with SpaceX with the same customers to deliver the payload into orbit. Therefore, if SpaceY we can determine if the first stage will land, it can determine the cost of a launch. This information can be used if SpaceY wants to bid against SpaceX.
  - This goal of the project is to create a machine learning prediction model, to determine if the first stage will land successfully.
- **Problems we want to find answers for**
  - What is the price of each launch – i.e. what determines, whether the first stage of the Falcon 9 rocket will land successfully?
  - Are there any patterns / dependencies in the parameters that have an impact on the landing of the first stage of the Falcon 9 rocket?
  - How do we make a reliable prediction whether a landing of the first stage will be successful, given the different parameters of each launch from public information using the machine learning models?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We have collected the data using various methods, including SpaceX API, webscraping publicly available SpaceX information as well via csv files
- Perform data wrangling
  - We performed some Exploratory Data Analysis in order to find some pattern in the data and to determine training labels for the outcomes (1 meaning the stage landed and 0 meaning it did not land)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We have preprocessed the data, allowing us to standardize our data, and perform a train-test split, We trained the model and performed Grid Search, allowing us to find the hyperparameters that allow a given algorithm to perform best.
  - Using the best hyperparameter values, we determined the model with the best accuracy using the training data. We tested Logistic Regression, Support Vector machines, Decision Tree Classifier and K-nearest neighbors. Finally, we produced a confusion matrix.

https://github.com/sef-ffm/ibm_ds_capstone

Data Collection

- Two different data collection methods were used:
  - **API**: using a get() request to SpaceX API, decoding the response content as a Json using .json() and turning it into a Pandas dataframe using .json_normalize(). Some data cleaning and handling of missing values was performed on the dataset.
  - **Webscraping:** we extracted the Falcon 9 launch records HTML table from Wikipedia using BeautifulSoup. The contents of the table was parsed and converted it into a Pandas data frame.

https://github.com/sef-ffm/ibm_ds_capstone

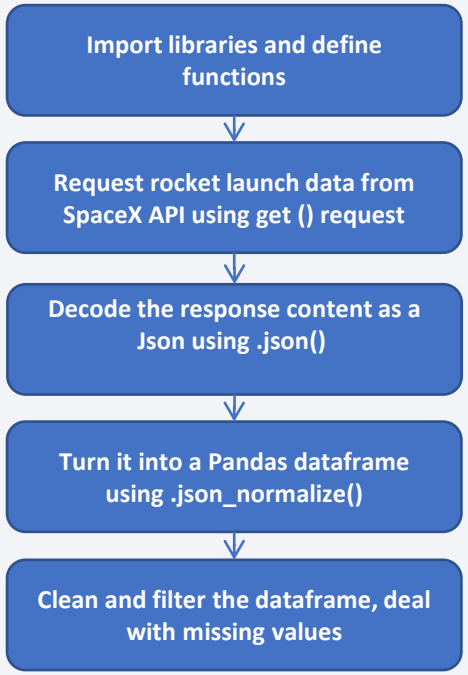
Data Collection – SpaceX API

- We used a get() request to SpaceX API, decoded the response content as a Json using .json() and turned it into a Pandas dataframe using .json_normalize(). Some data cleaning and handling of missing values was performed on the dataset, also using some pre-defined functions.

- Notebook available on Github:

https://github.com/sef-ffm/ibm_ds_capstone/blob/main/Week%201%20-%20jupyter-labs-spacex-data-collection-api.ipynb

Flowchart:



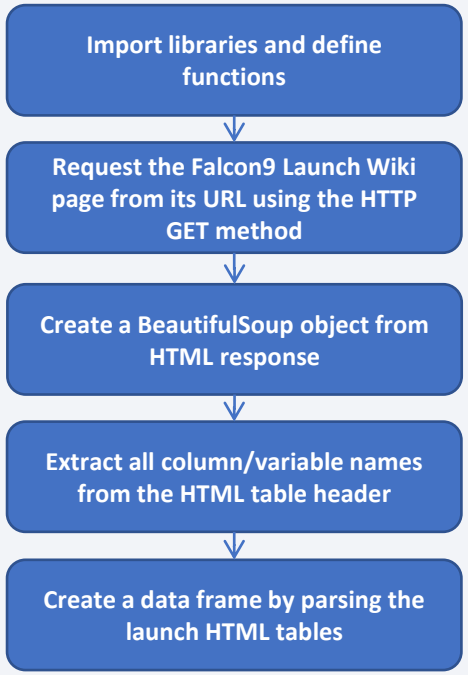
Data Collection - Scraping

- We used Python's BeautifulSoup Package to scrape a Wikipedia article which contains Falcon 9 launch records. The table was parsed (using some functions) and converted into a Pandas dataframe.

- Notebook available on Github:

https://github.com/sef-ffm/ibm_ds_capstone/blob/main/Week%201%20%20-%20jupyter-labs-webscraping.ipynb

Flowchart:



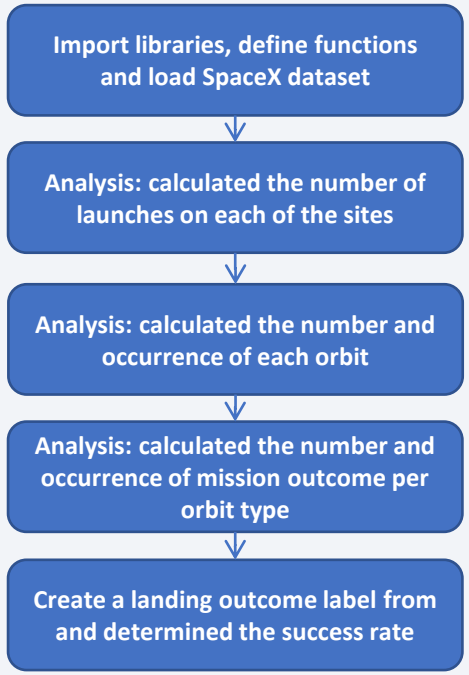
Data Wrangling

- We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models. We converted the landing outcomes into Training Labels with 1 meaning that the booster successfully landed and 0 meaning it was unsuccessful.

- Notebook available on Github:

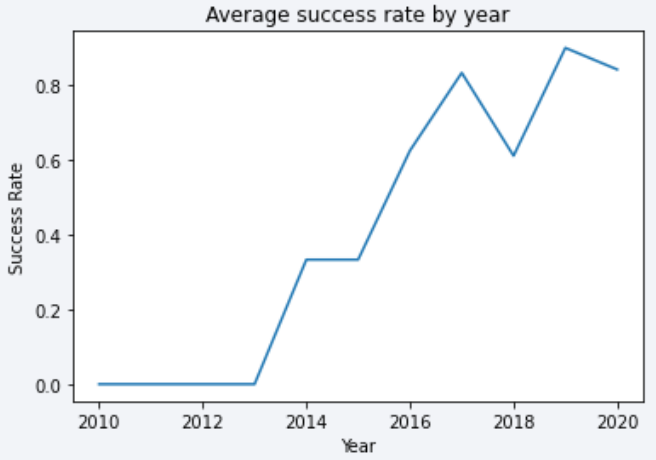
https://github.com/sef-ffm/ibm_ds_capstone/blob/main/Week%201%20-%20labs-jupyter-spacex-Data%20wrangling.ipynb

Flowchart:



EDA with Data Visualization

- We performed exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib. The following charts were plotted in order to understand the interrelationships between success, payload, flight No., launch site and orbit type. Further, we have plotted the avarage success rate by year:
    - Payload mass vs. flight number
    - Flight number vs launch site
    - Payload mass vs launch site
    - Class vs orbit
    - FlightNumber and Orbit type
    - Payload and Orbit type
    - Average success rate by year



- Notebook available on Github:

https://github.com/sef-ffm/ibm_ds_capstone/blob/main/Week%202%20-%20jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- For further EDA, the SpaceX dataset was loaded into a Db2 database, and SQL queries were performed using sqlalchemy, ibm_db_sa and ipython-sql packages to produce the following results:
  - Names of the unique launch sites in the space mission
  - Records where launch site is CCAFS LC-40 or CCAFS SLC-40
  - Total payload mass carried by boosters launched by NASA
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failed mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions, and launch site names in year 2015
  - 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

- Notebook available on Github:

https://github.com/sef-ffm/ibm_ds_capstone/blob/main/Week%202%20-%20jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- We have added circles and markers to the Folium map to mark the launch sites on the map using the site‘s latitute and longitude coordinates.
- We have created marker clusters to show launch outcomes for each site, and to see which sites have high success rates.
- We have added lines to the map in order to show proximities to the important coordinates (e.g. coastline, railway).

- Notebook available on Github:

https://github.com/sef-ffm/ibm_ds_capstone/blob/main/Week%203%20-%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We have built a Plotly dashboard. It provides plots and interactions and lets the user understand the interaction between launch site, payload mass and success rates.
- The dashboard outputs a:
    - Pie chart to show the % of successful launches by launch site (when no particular launch site is selected)
    - Success/failure rates for a selected launch site, given a specified payload range, defferentiating between booster version categories.

- Python file available on Github:

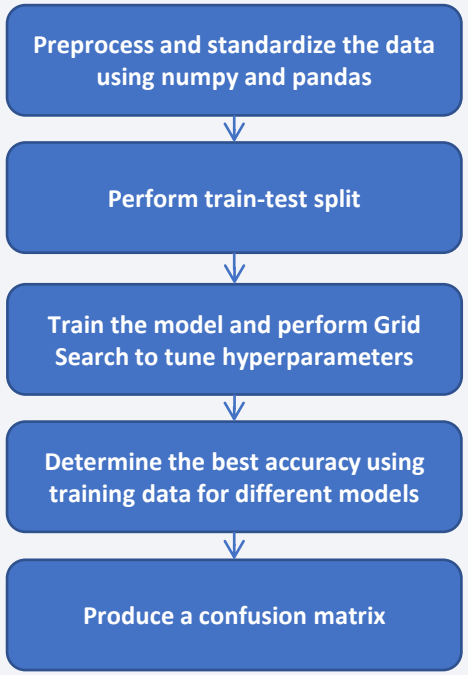
https://github.com/sef-ffm/ibm_ds_capstone/blob/main/Week%203%20-%20Interactive%20Dashboard%20With%20Plotly.py

Predictive Analysis (Classification)

- We have preprocessed the data, allowing us to standardize our data, and perform a train-test split, We trained the model and performed Grid Search, allowing us to find the hyperparameters that allow a given algorithm to perform best.
- Using the best hyperparameter values, we will determine the model with the best accuracy using the training data. We tested Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.
- Finally, we produced a confusion matrix.
- Notebook available on Github:

https://github.com/sef-ffm/ibm_ds_capstone/blob/main/Week%204%20-%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Flowchart:



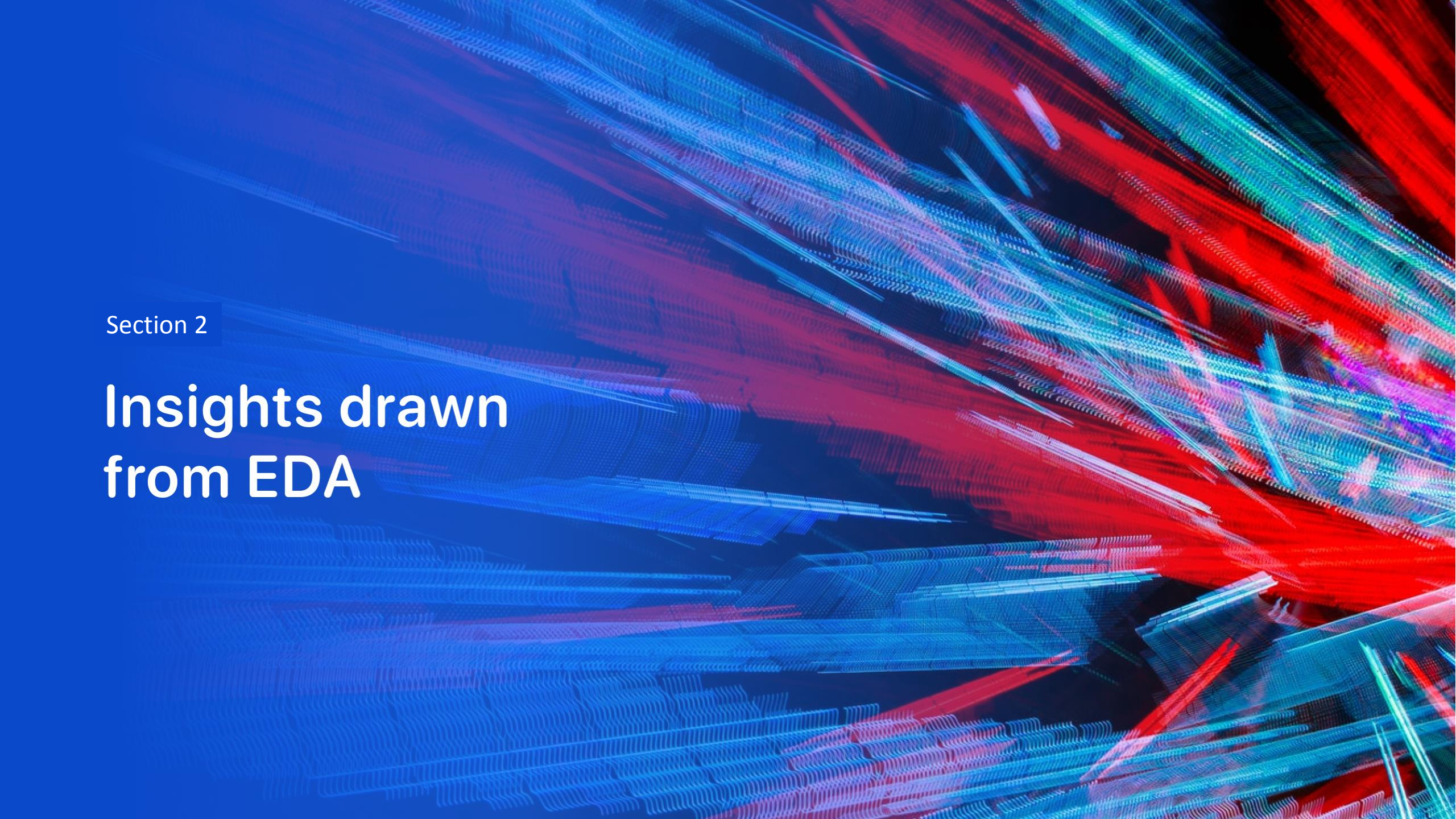
Results

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

https://github.com/sef-ffm/ibm_ds_capstone

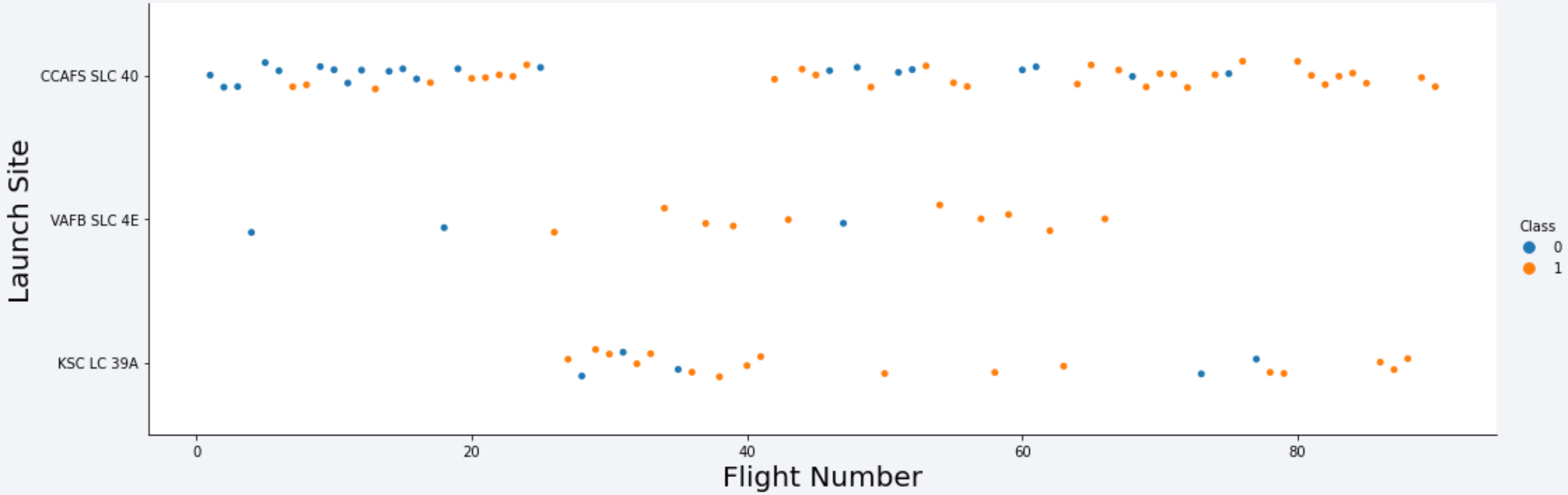
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

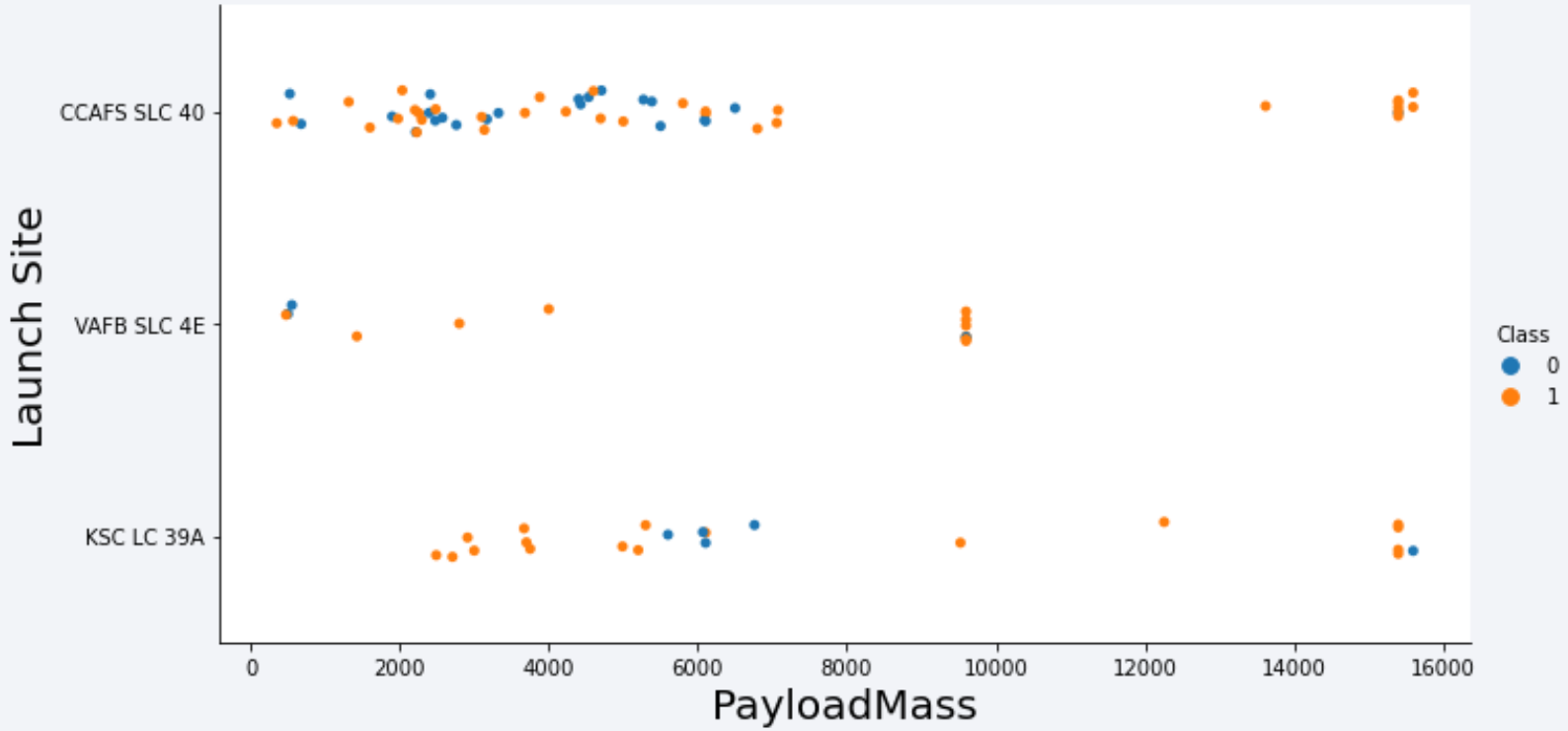
- The plot shows a positive relationship between the flight number and the success rate



https://github.com/sef-ffm/ibm_ds_capstone

Payload vs. Launch Site

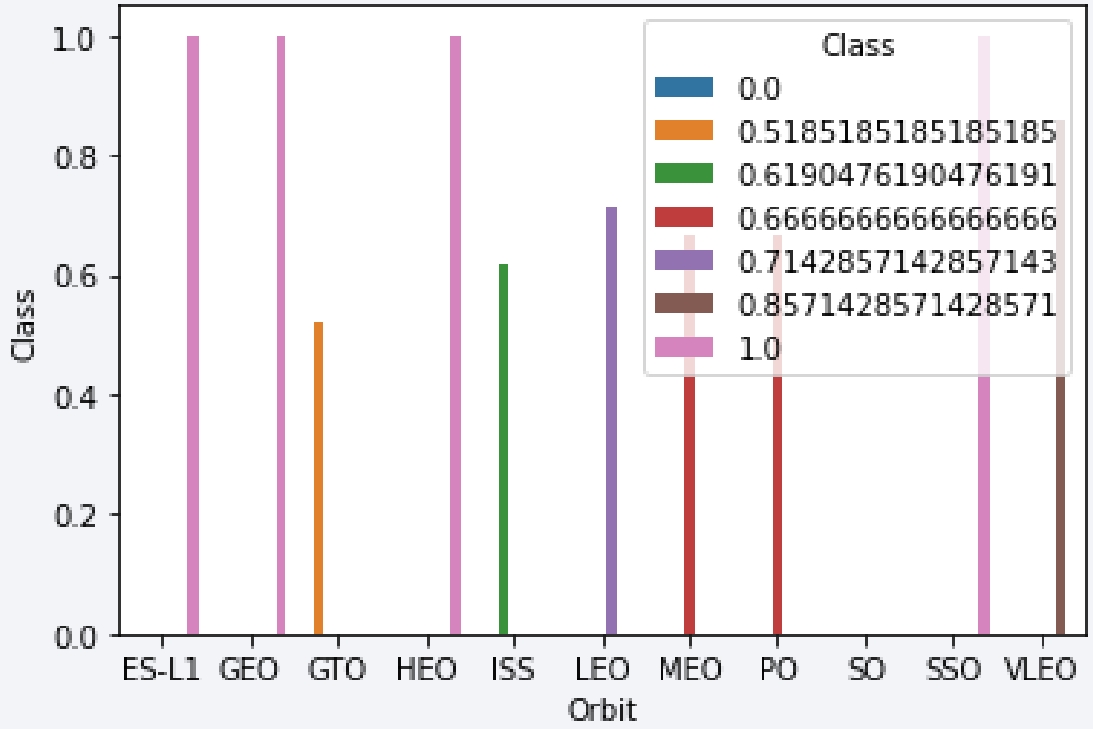
- The plot shows a positive relationship between the flight number the success rates for CCAFS SLC 40 and VAFB SLC 4E



https://github.com/sef-ffm/ibm_ds_capstone

Success Rate vs. Orbit Type

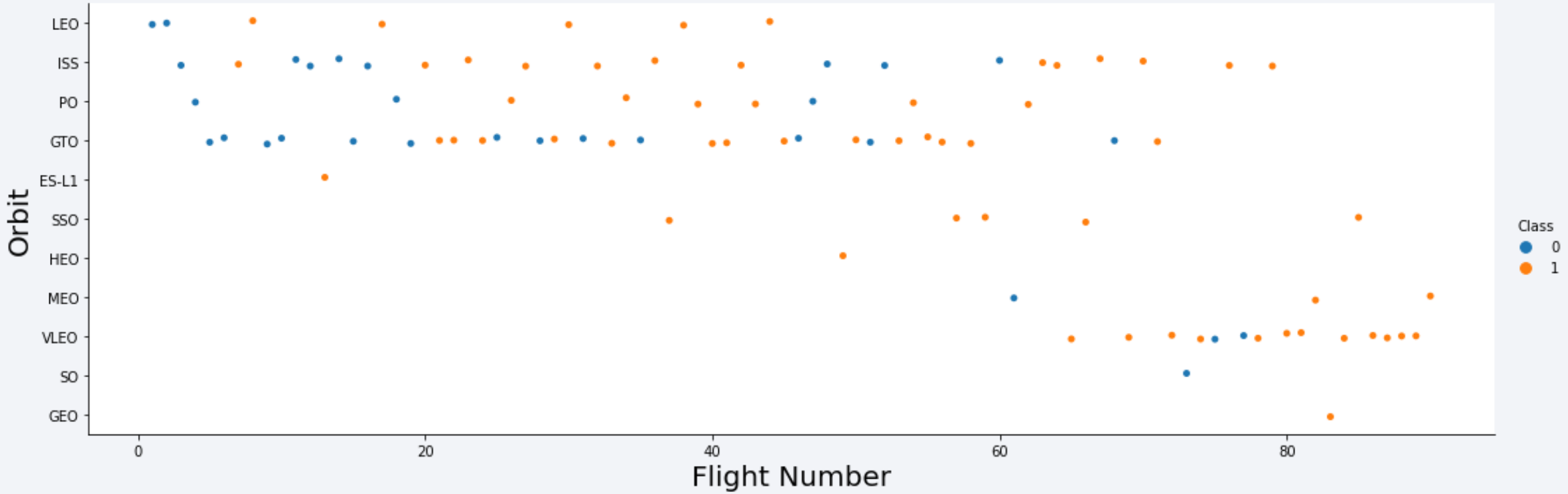
- Some orbits had a higher success rate than the others (e.g. ES-L1, HEO, SSO)



https://github.com/sef-ffm/ibm_ds_capstone

Flight Number vs. Orbit Type

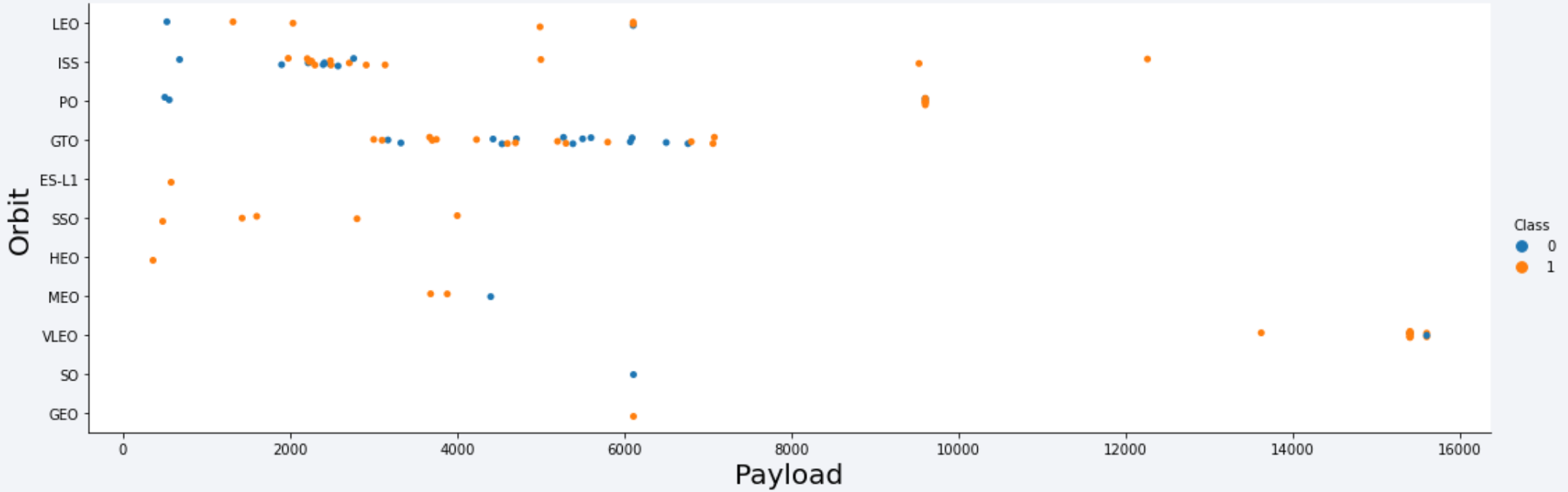
- For certain orbits, heavier payloads have a higher success rate (VLEO, ISS, LEO, PO)



https://github.com/sef-ffm/ibm_ds_capstone

Payload vs. Orbit Type

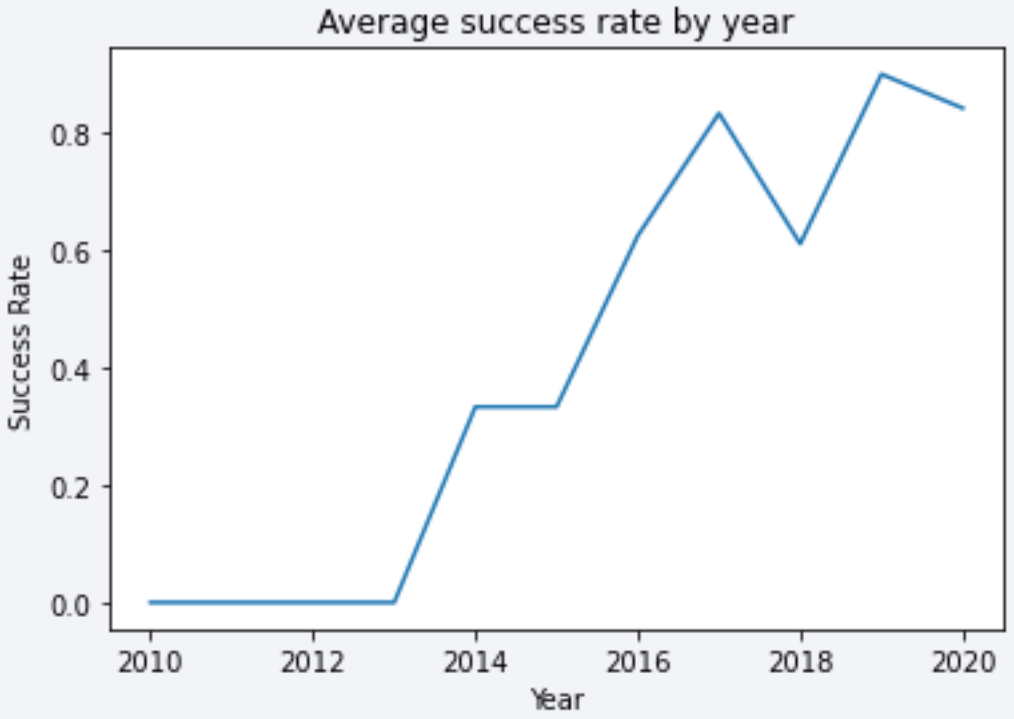
- More successful landings for certain orbits, as payload increases (ISS, PO, LEO)



https://github.com/sef-ffm/ibm_ds_capstone

Launch Success Yearly Trend

- Success rate has been overall increasing since the first launch, with small declines in 2018 and 2020



https://github.com/sef-ffm/ibm_ds_capstone

All Launch Site Names

- The following SQL query was used to get the unique launch sites from the SpaceX data

```
%sql select UNIQUE(LAUNCH_SITE) from SPACEXTBL;
```

- The result:

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

https://github.com/sef-ffm/ibm_ds_capstone

Launch Site Names Begin with 'CCA'

- The following SQL query was used to get 5 records where launch sites begin with the string ‘CCA’

```
%sql SELECT * from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

- The result:

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

https://github.com/sef-ffm/ibm_ds_capstone

Total Payload Mass

- The following SQL query was used to display the total payload mass carried by boosters launched by NASA (CRS)

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

- The result:

payloadmass
45596

Average Payload Mass by F9 v1.1

- The following SQL query was used to display the average payload mass carried by booster version F9 v1.1

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

- The result:

payloadmass
2928

First Successful Ground Landing Date

- The following SQL query was used to list the date when the first successful landing outcome in ground pad was acheived.

```
%sql select min(DATE) from SPACEXTBL where LANDING__OUTCOME = 'Success (ground pad)' ;
```

- The result:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The following SQL query was used to 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 BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

- The result:

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

https://github.com/sef-ffm/ibm_ds_capstone

Total Number of Successful and Failure Mission Outcomes

- The following SQL query was used to list the total number of successful and failure mission outcomes

```
%sql select MISSION_OUTCOME, count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

- The result:

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

https://github.com/sef-ffm/ibm_ds_capstone

Boosters Carried Maximum Payload

- The following SQL query was used to list the names of the booster versions which have carried the maximum payload mass

```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_= (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

- The result:

boosterversion
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

https://github.com/sef-ffm/ibm_ds_capstone

2015 Launch Records

- The following SQL query was used to list the failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

```
%sql select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where LANDING__OUTCOME = 'Failure (drone ship)' and YEAR(DATE) = '2015'
```

- The result:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

https://github.com/sef-ffm/ibm_ds_capstone

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

- The following SQL query was used to 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(LANDING__OUTCOME) as landingoutcomes FROM (select * from spacextbl WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20') group by landing__outcome order by landingoutcomes desc
```

- The result:

landing__outcome	landingoutcomes
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

https://github.com/sef-ffm/ibm_ds_capstone

Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites on a Map

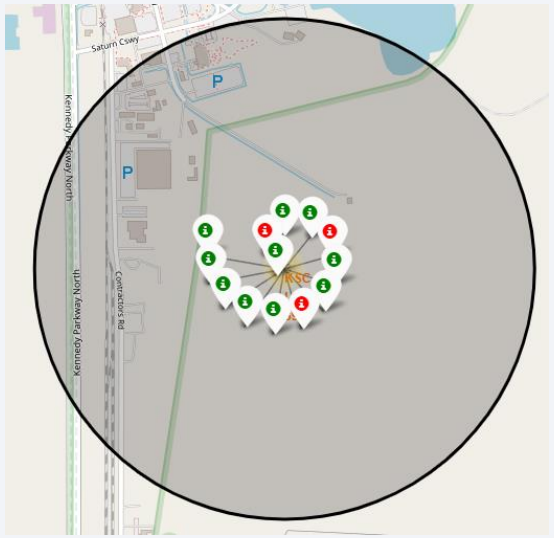
- The following map shows the location of SpaceX launch sites, which are on USA East and West Coast.

https://github.com/sef-ffm/ibm_ds_capstone

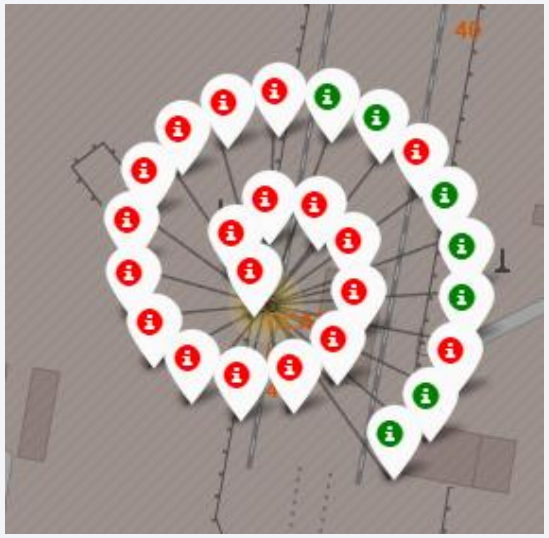
Color-labeled launch outcomes

- The below screenshots show the four launch sites (first three are Florida launch sites, and the latter is California launch site) as well as the outcomes of the launch, whereby green color denotes successful launches and red color denotes unsuccessful launches

KSC LC-39A

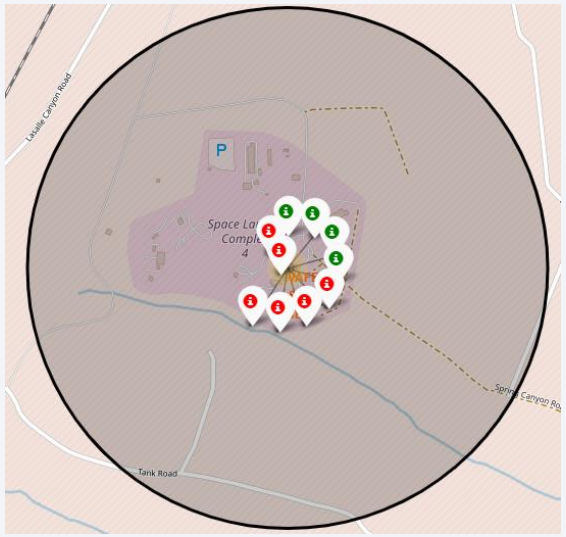


CCAFS LC-40



CCAFS SLC-40

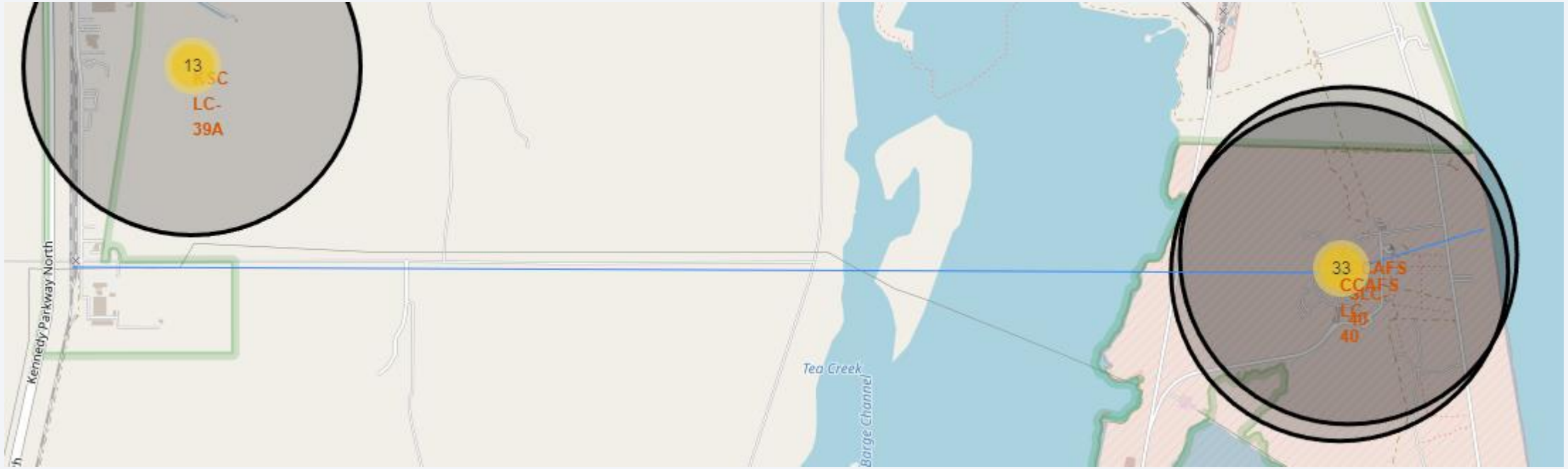
VAFB SLC-4E



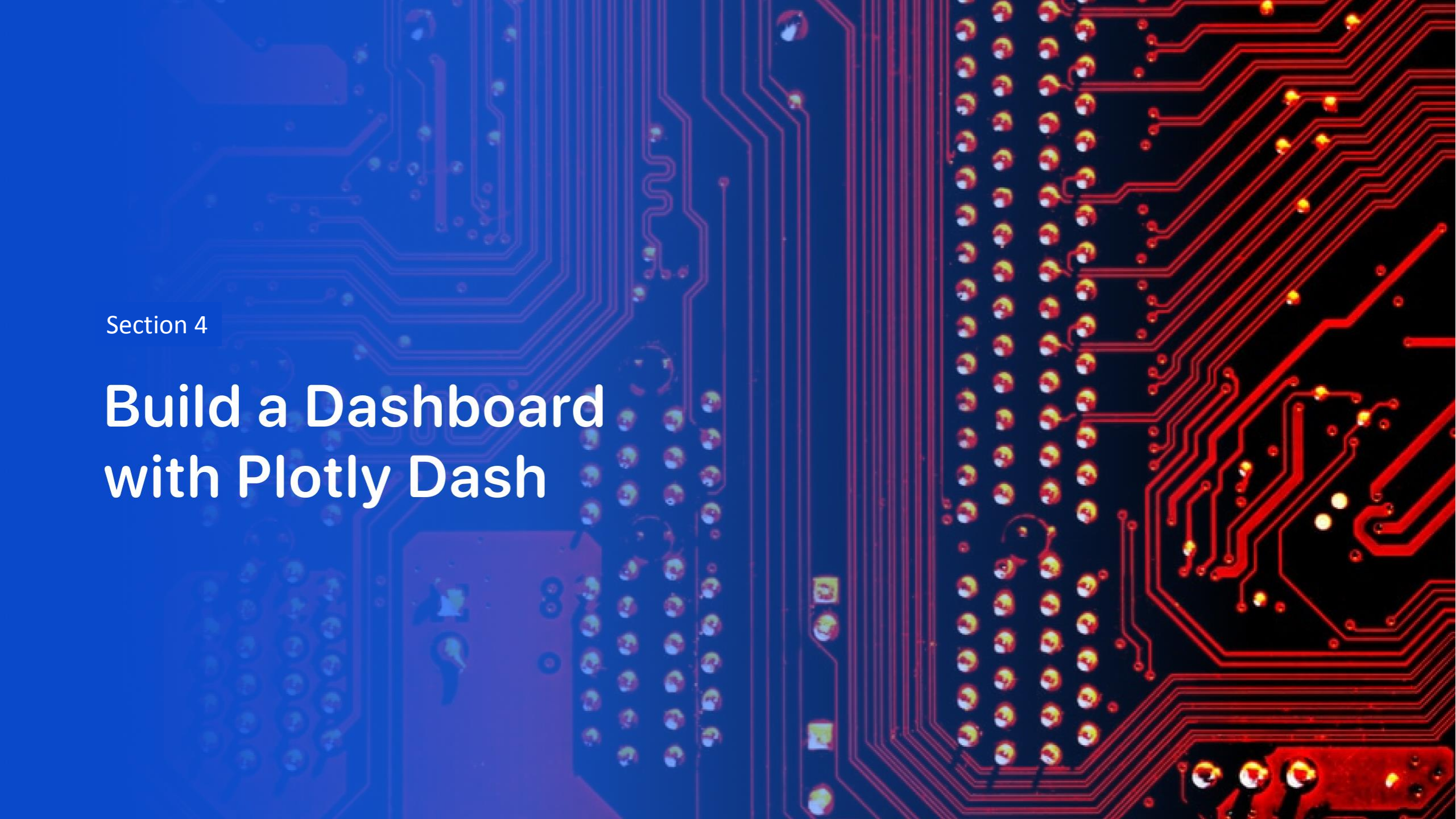
https://github.com/sef-ffm/ibm_ds_capstone

Distance of launch site to landmarks

- The below screenshots show the proximity of the launch site to the nearest landmarks (coastline and railway station), denoted by the blue line.

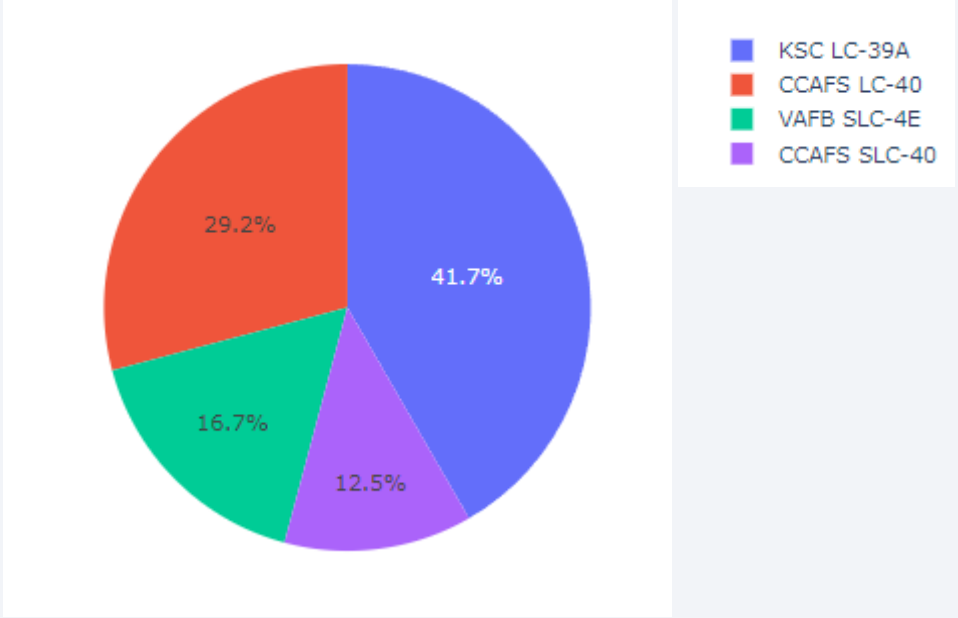


https://github.com/sef-ffm/ibm_ds_capstone


Section 4
Build a Dashboard with Plotly Dash

Success Count for all Launch Sites

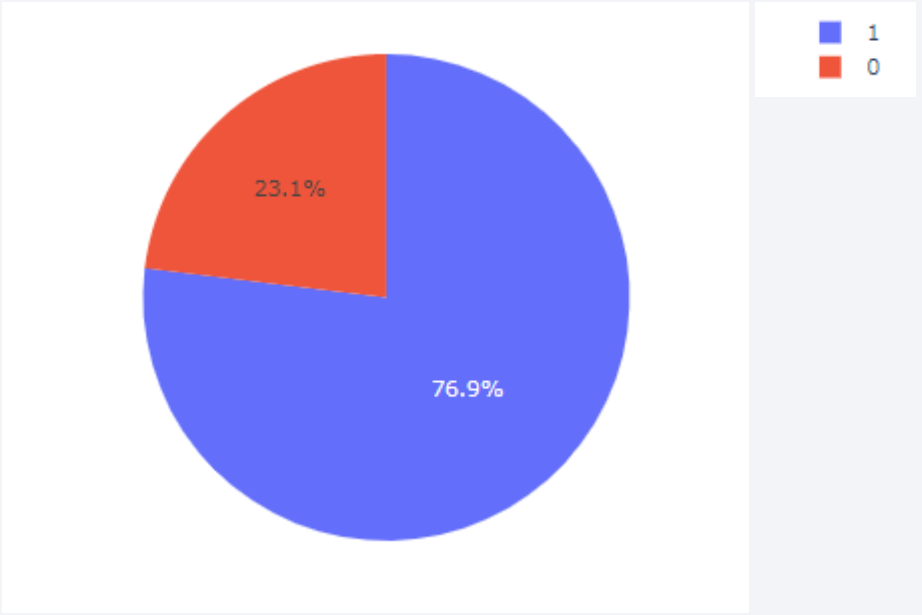
- The pie chart shows that KSC LC-39A accounts for 42% of all successful launches, followed by CCAFS LC-40 (29%).



https://github.com/sef-ffm/ibm_ds_capstone

Launch Site with Highest Success Ratio

- The most successful launch site is KSC LC-39A with at 77% success ratio

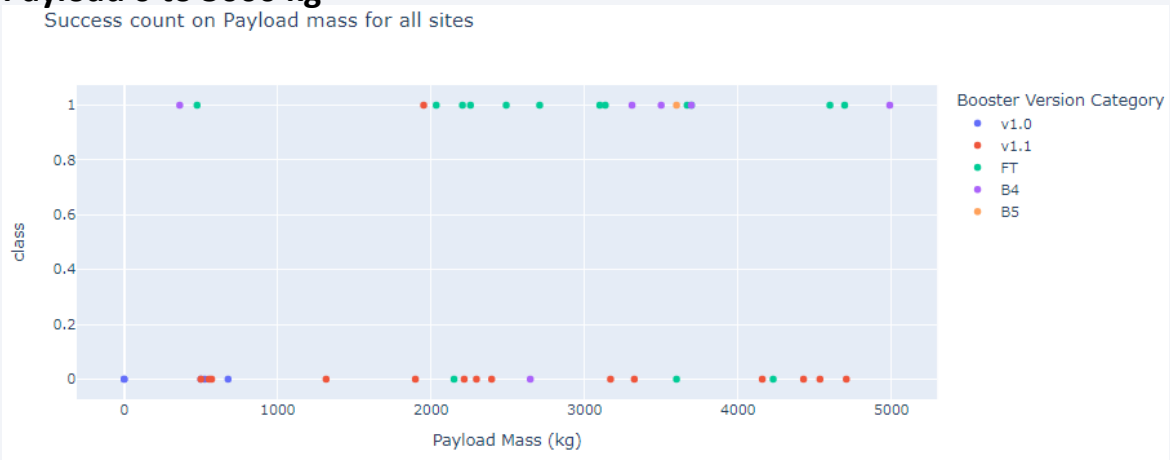


https://github.com/sef-ffm/ibm_ds_capstone

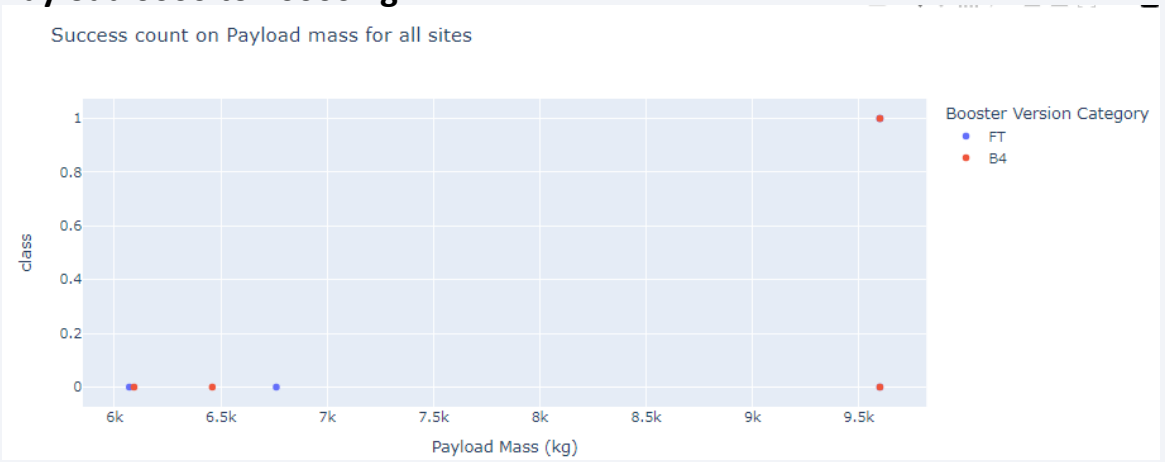
Successes for Different Payloads for all Sites

- The two panels show success rates for two different payload categories (0 to 5000kg and 6000 to 10000kg).
- In general, launches with higher payloads show a higher success ratio.

Payload 0 to 5000 kg



Payload 6000 to 10000 kg



https://github.com/sef-ffm/ibm_ds_capstone

Section 5

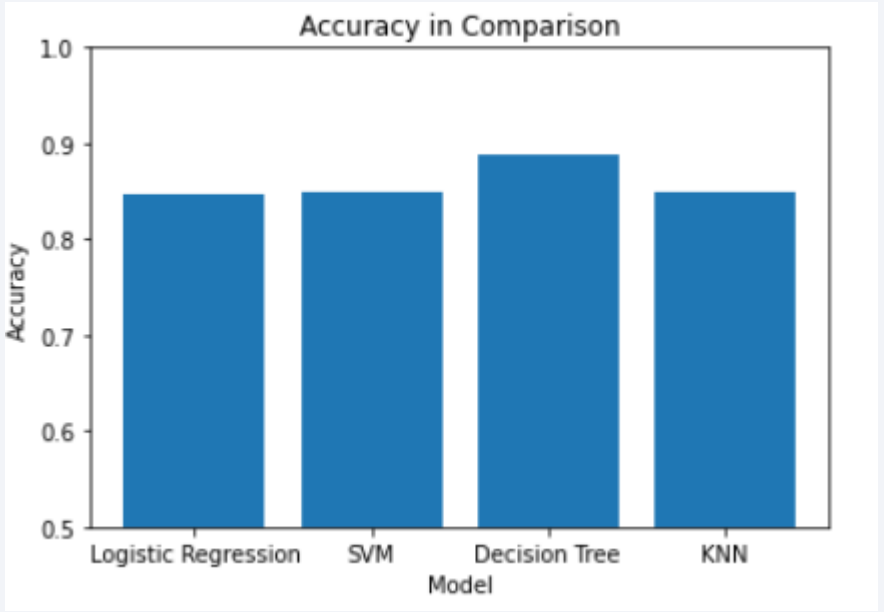
Predictive Analysis (Classification)

https://github.com/sef-ffm/ibm_ds_capstone

Classification Accuracy

- Decision Tree classifier produces the highest classification accuracy:

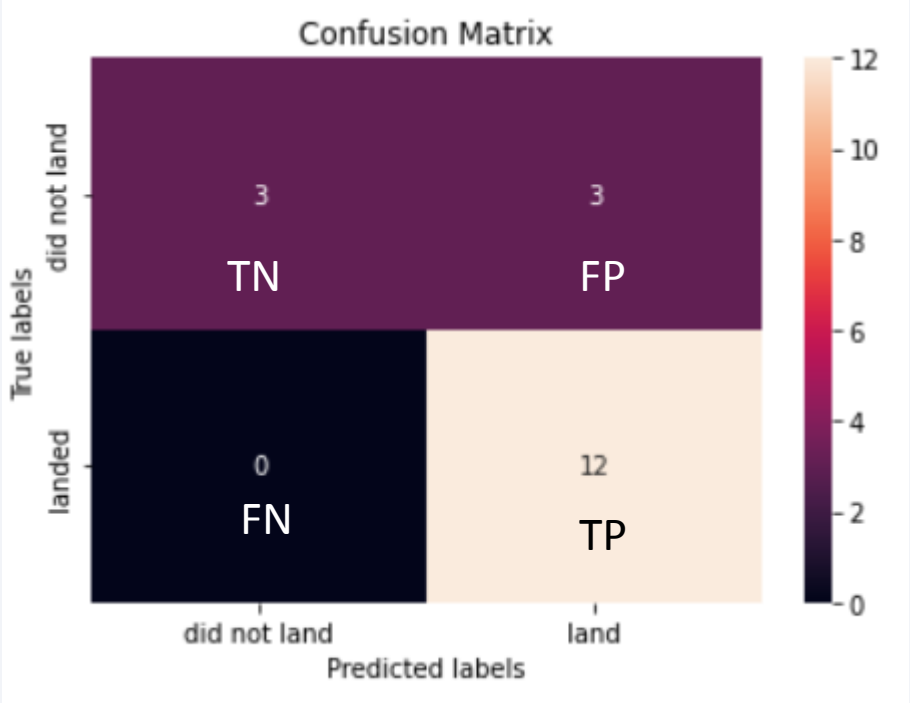
	Accuracy
Logistic Regression	0.846
SVM	0.848
Decision Tree	0.888
KNN	0.848



https://github.com/sef-ffm/ibm_ds_capstone

Confusion Matrix

- The confusion matrix of the decision tree classifier is shown on the right.
- The confusion matrix shows 3 false positives and does not have any false negatives.
- Accuracy is (TP+TN)/(TP + TN + FP + FN) = 15/18 = 83%



https://github.com/sef-ffm/ibm_ds_capstone

Conclusions

- The more flights are launched by SpaceX, the higher the probability of a successful landing, i.e. average success rate increases over time
- Some orbits have a higher success rates (e.g. ES-L1, HEO, SSO)
- There are more successful landings for certain orbits, as payload increases (ISS, PO, LEO, VLEO)
- The most successful launch site is KSC LC-39A with at 77% success ratio
- In general, launches with higher payloads show a higher success ratio.
- Decision tree has the best accuracy in predicting the launch outcome

https://github.com/sef-ffm/ibm_ds_capstone

Thank you!