IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yassine Rodani
13 November 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection and wrangling
  - EDA with SQL and with data visualization
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Predictive analysis results

Introduction

- Project background and context

- In this project, using the SpaceX data, we will determine whether the first stage of a SpaceX Falcon9 rocket will land successfully or not, thereby determining the effective cost of a launch.

- Problems you want to find answers

- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API and Web Scrapping from SpaceX Wikipedia page
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

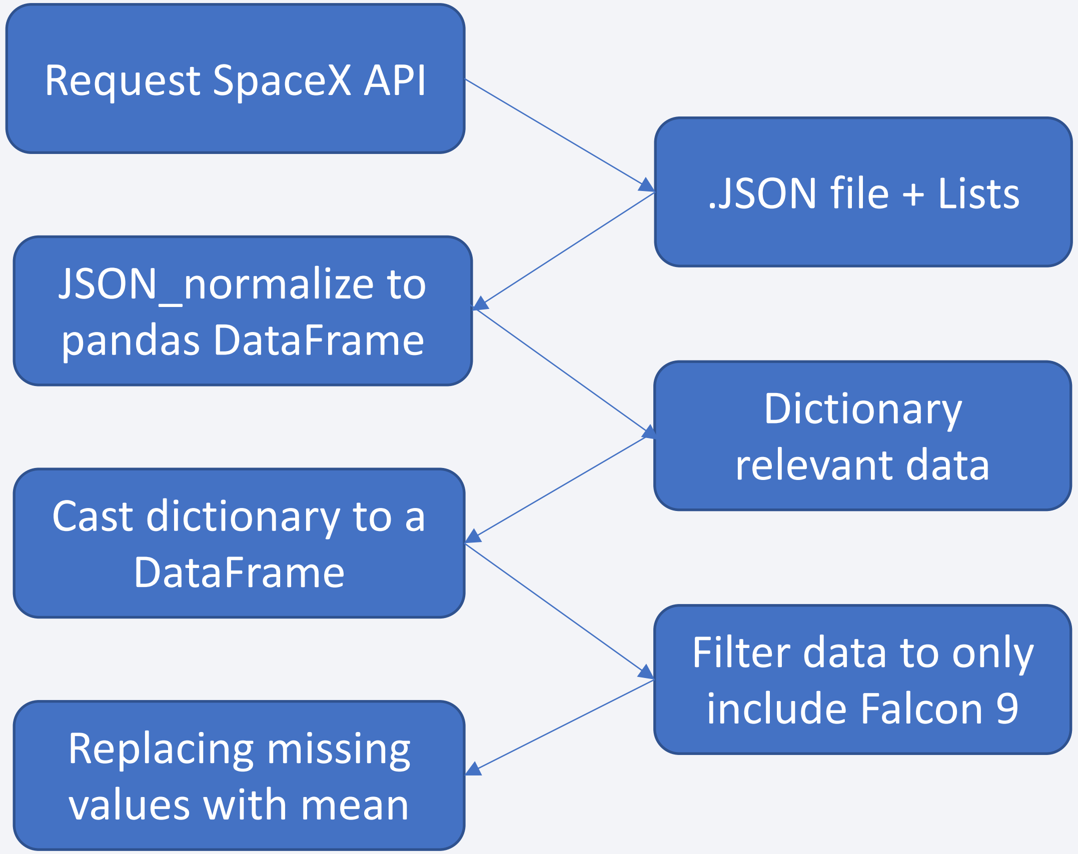
Data Collection

- Data collection process involved a combination of :
  - API requests from Space X public API
    - This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - Web scrapping from SpaceX Wikipedia page

- The next slide will show the sequence of procecessing the data from SpaceX public API, and the other one after will show sequence of processing the data from webscrapping.

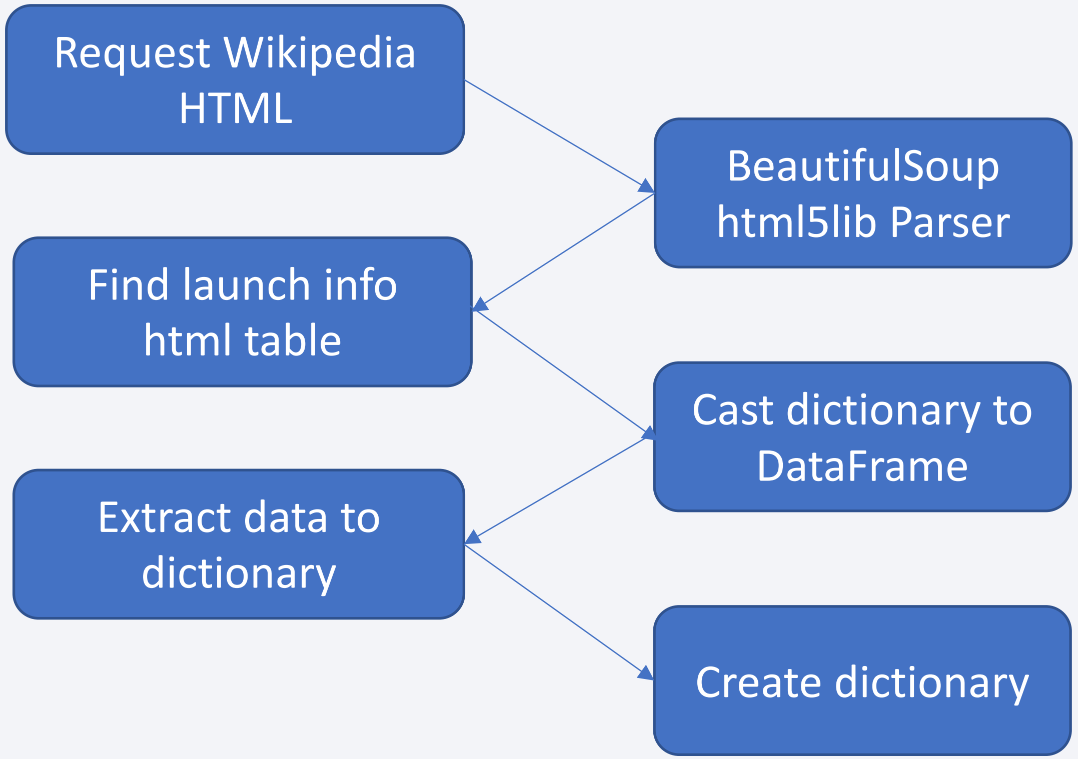
Data Collection – SpaceX API

- https://github.com/yassine-rd/IBM-Data-Science-Professional-Certification-Capstone/blob/master/Data_Collection.ipynb



Data Collection - Scraping

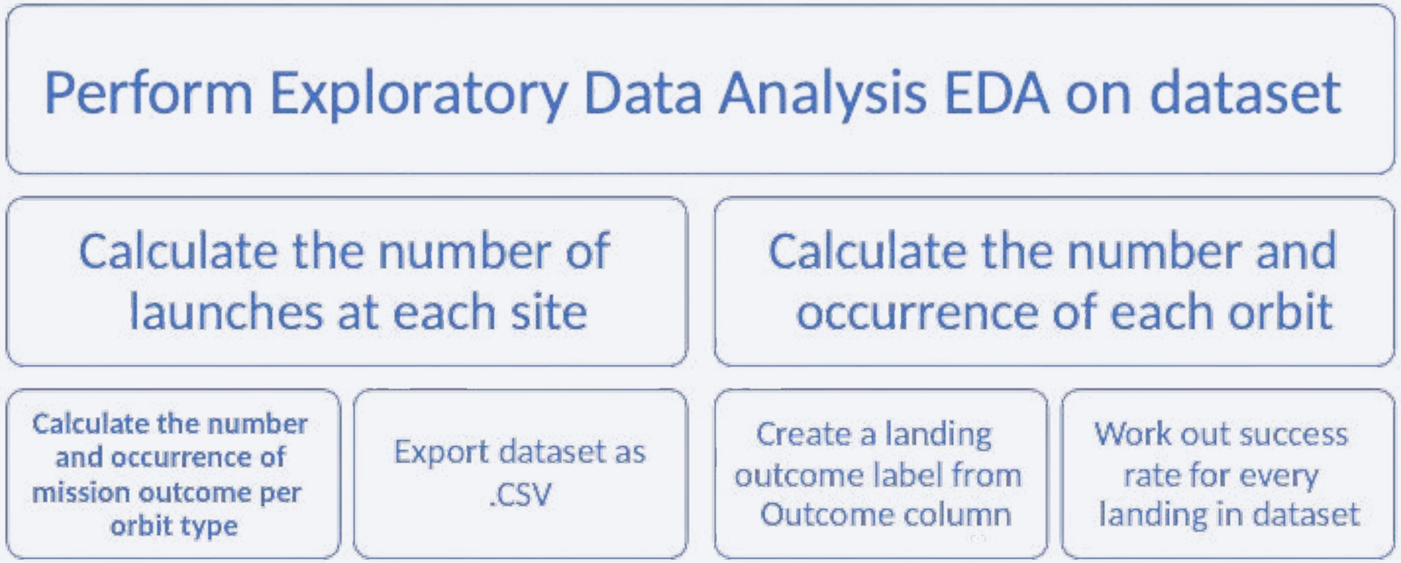
- https://github.com/yassine-rd/IBM-Data-Science-Professional-Certification-Capstone/blob/master/Data_Collection_with_Web_Scraping_lab.ipynb



Data Wrangling

https://github.com/yassine-rd/IBM-Data-Science-Professional-Certification-Capstone/blob/master/Data_Wrangling.ipynb

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, 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, etc..
- We mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.

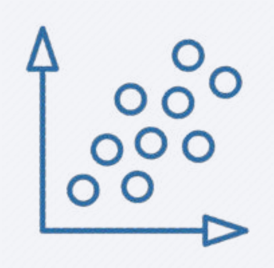


EDA with Data Visualization

https://github.com/yassine-rd/IBM-Data-Science-Professional-Certification-Capstone/blob/master/Exploratory_Analysis_Using_Pandas_and_Matplotlib.ipynb

- Charts plotted :
  - Flight Number VS. Payload Mass
  - Flight Number VS. Launch Site
  - Payload VS. Launch Site
  - Orbit VS. Flight Number Payload VS. Orbit Type Orbit VS. Payload Mass
  - Mean VS. Orb
  - Success Rate VS. Year

Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exist so that they could be used in training the ML model



EDA with SQL

https://github.com/yassine-rd/IBM-Data-Science-Professional-Certification-Capstone/blob/master/Exploratory_Analysis_Using_SQL.ipynb

- Loaded data set into IBM DB2 Databae
- Queried using SQL python integration
- Queries were made to get better understanding of the dataset
- Queried
  - Information about launch site names
  - Missions outcomes
  - Various pay load sizes of customers and booster versions
  - Landing outcomes

Build an Interactive Map with Folium

https://github.com/yassine-rd/IBM-Data-Science-Professional-Certification-Capstone/blob/master/Interactive_Visual_Analytics_and_Dashboard.ipynb

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()

Predictive Analysis (Classification)

https://github.com/yassine-rd/IBM-Data-Science-Professional-Certification-Capstone/blob/master/Machine_Learning_Prediction.ipynb

- Building Model
    - Load our dataset into NumPy and Pandas
    - Transform Data
    - Split our data into training and test data sets
    - Check how many test samples we have
    - Decide which type of machine learning algorithms we want to use
    - Set our parameters and algorithms to GridSearchCV
    - Fit our datasets into the GridSearchCV objects and train our dataset.

- Evaluating Model
    - Feature Engineering
    - Algorithm Tuning

- Evaluating Model
    - Check accuracy for each model
    - Get tuned hyperparameters for each type of algorithms
    - Plot Confusion Matrix

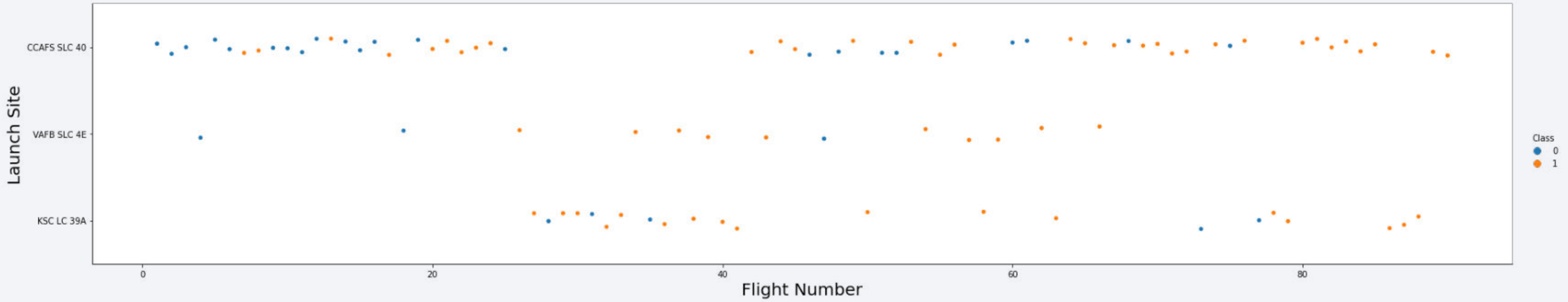
Results

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

Section 2

Insights drawn from EDA

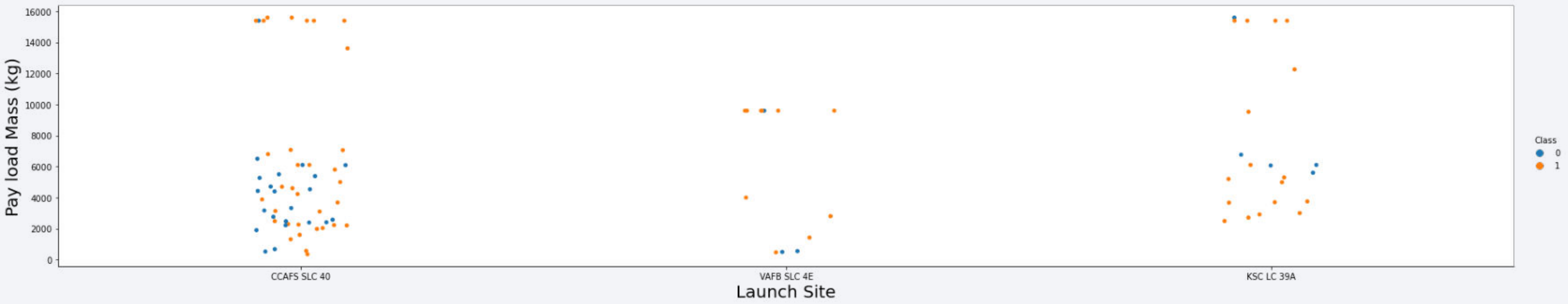
Flight Number vs. Launch Site



- Orange indicate ssuccessful launch
- Blue indicates unsuccessful launch

Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appearsto be the main launch site as it has the most volume.

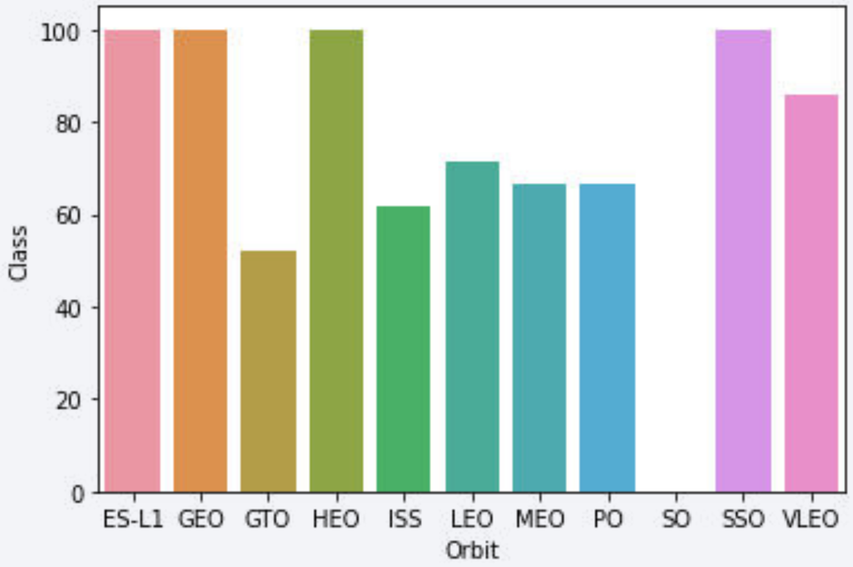
Payload vs. Launch Site



- Orange indicate ssuccessful launch
- Blue indicates unsuccessful launch

Payloadmass appearsto fallmostly between 0-7000 kg. Differentlaunch sites also seem to use different payloadmass.

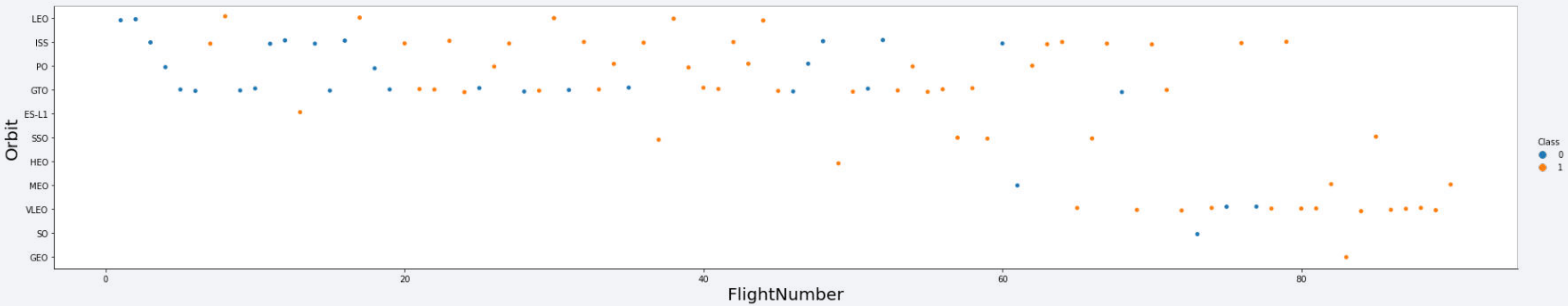
Success Rate vs. Orbit Type



Success Rate Scale with %

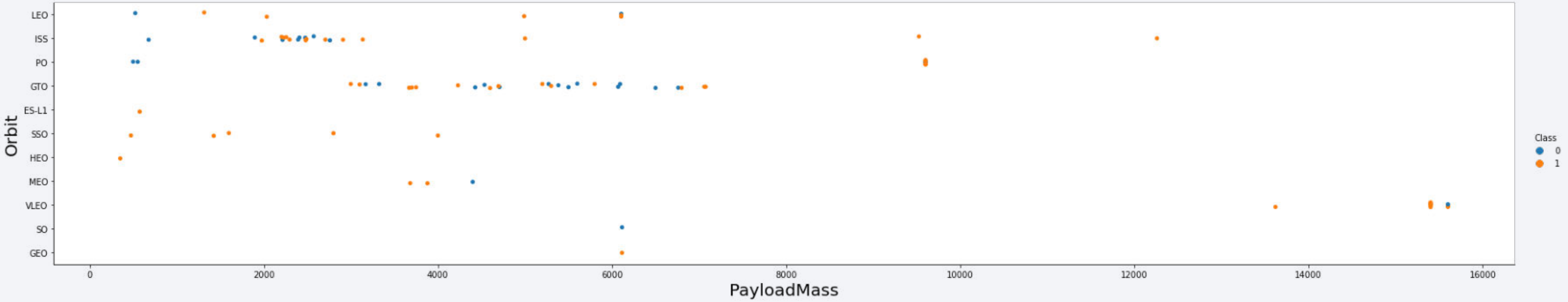
- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis)
- SSO (5) has 100%success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

Flight Number vs. Orbit Type



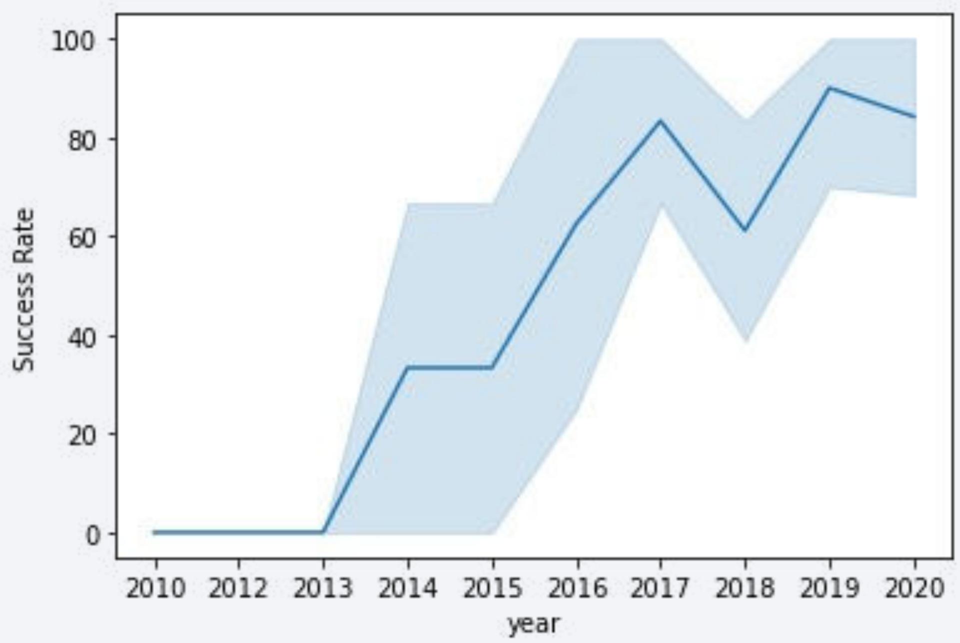
- Launch Orbit preferences changed over FlightNumber.
- Launch Outcome seems to correlatewith this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appearsto performbetterin lower orbits or Sun-synchronousorbits

Payload vs. Orbit Type



- Payloadmassseems to correlatewith orbit
- LEOand SSO seem to haverelatively lowpayload mass
- The other mostsuccessful orbit VLEO only has payload mass valuesin the higher end of the range

Launch Success Yearly Trend



- 95%confidence interval (light blue shading)

- Success generally increases over time since 2013 with a slight dip in 2018
- Successin recent years at around 80%

All Launch Site Names

Task 1

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

[13]
```
%sql select distinct launch_site from SPACE_X_TABLE
```
Python

```
* ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

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

- Query unique launch site names from database.
- CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors.
- CCAFS LC-40 was the previous name.
- Likely only 3 unique launch_site values: CCAFS SLC40

Launch Site Names Begin with 'CCA'

Task 2

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

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

[14] Python

```
* ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

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

First five entries in databasewith Launch Site name beginning with CCA.

Total Payload Mass

Task 3

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

```
[15] %sql select sum(payload_mass__kg_) as sum from SPACE_X_TABLE where customer like 'NASA (CRS)'
```

Python

```
 * ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

SUM
45596

- This query sumsthe totalpayload mass in kg where NASA was the customer.
- CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql select avg(payload_mass__kg_) as average from SPACE_X_TABLE where booster_version like 'F9 v1.1%'
```

[16] Python

```
* ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

average
2534

- This query calculatesthe average payloadmass or launches which used booster versionF9 v1.1
- Average payload mass of F9 1.1 is on the low endof our payloadmassrange

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
%sql select min(date) as date from SPACE_X_TABLE where mission_outcome like 'Success'
```

[17] Python

```
 * ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

DATE
2010-06-04

- This query list thr first successful ground landing date

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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 SPACE_X_TABLE where (mission_outcome like 'Success') and
(payload_mass__kg_ between 4000 and 6000) and (landing__outcome like 'Success (drone ship)')
```

[19] Python

```
* ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-
3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

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

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as count from
SPACE_X_TABLE group by mission_outcome order by mission_outcome
```

[20] Python

```
* ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-
3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

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

- This query returns a count ofeach mission outcome.
- SpaceX appears to achieve its mission outcome nearly 99% ofthe time.
- This means that most ofthe landing failures are intended.
- Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters Carried Maximum Payload

Change Presentation

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

```
%sql select booster_version from SPACE_X_TABLE
where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACE_X_TABLE)
```

[22] Python

```
* ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

booster_version
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

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar and all are of the F9 B5 B10xx.x variety.
- This likely indicatespayload mass correlates with the booster versionthatis used.

2015 Launch Records

Task 9

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

```
%sql select MONTHNAME(DATE) as month, landing__outcome, booster_version, launch_site
from SPACE_X_TABLE where date like '2015%' and landing__outcome like 'Failure (drone ship)'
```

[23] Python

```
* ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

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

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

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

Task 10

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 count from SPACE_X_TABLE
where date >= '2010-06-04' and date <= '2017-03-20' group by landing__outcome order by count desc
```

[24] Python

```
* ibm_db_sa://mrd04677:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- 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

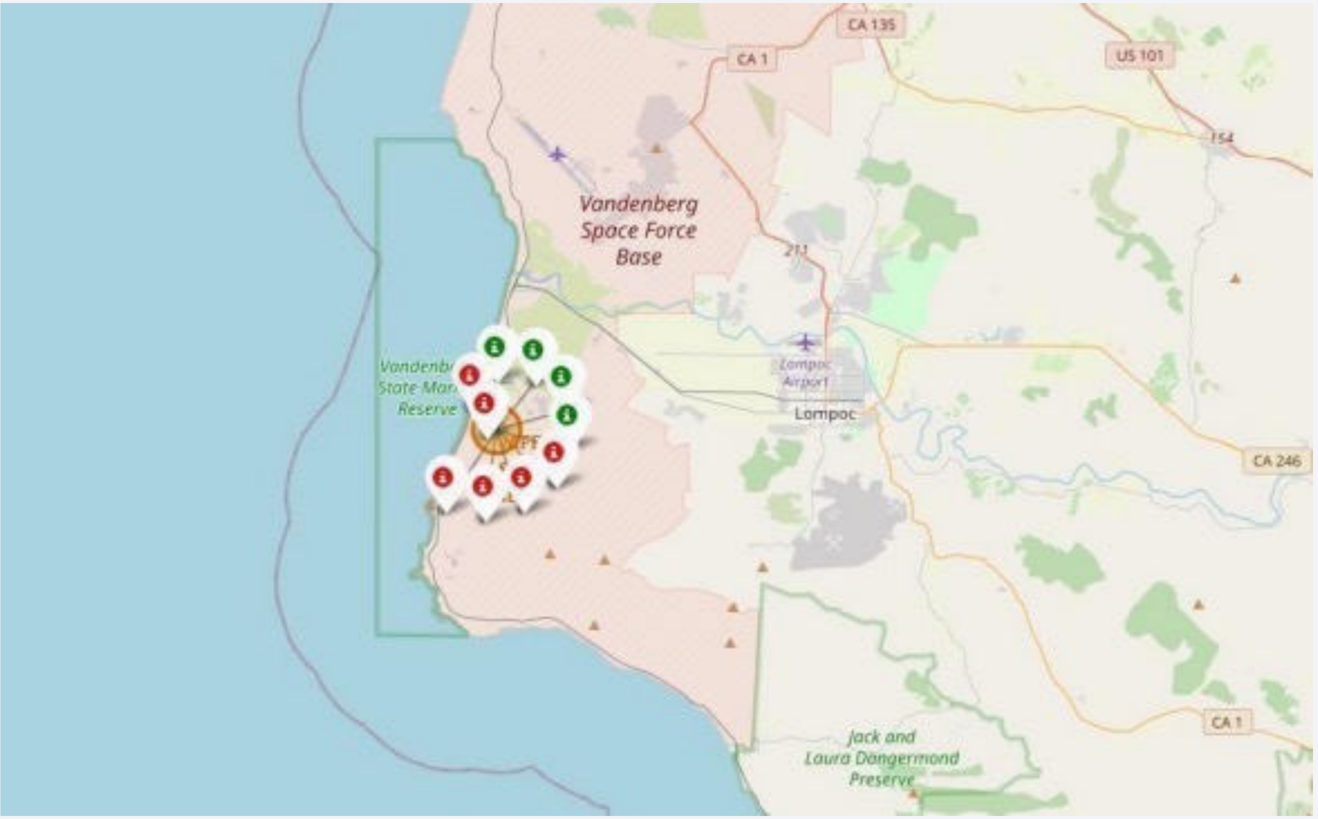
Section 3

Launch Sites Proximities Analysis

Launch Site Locations

- The map shows all launch sites relative US map.

Color-Coded Launch Markers

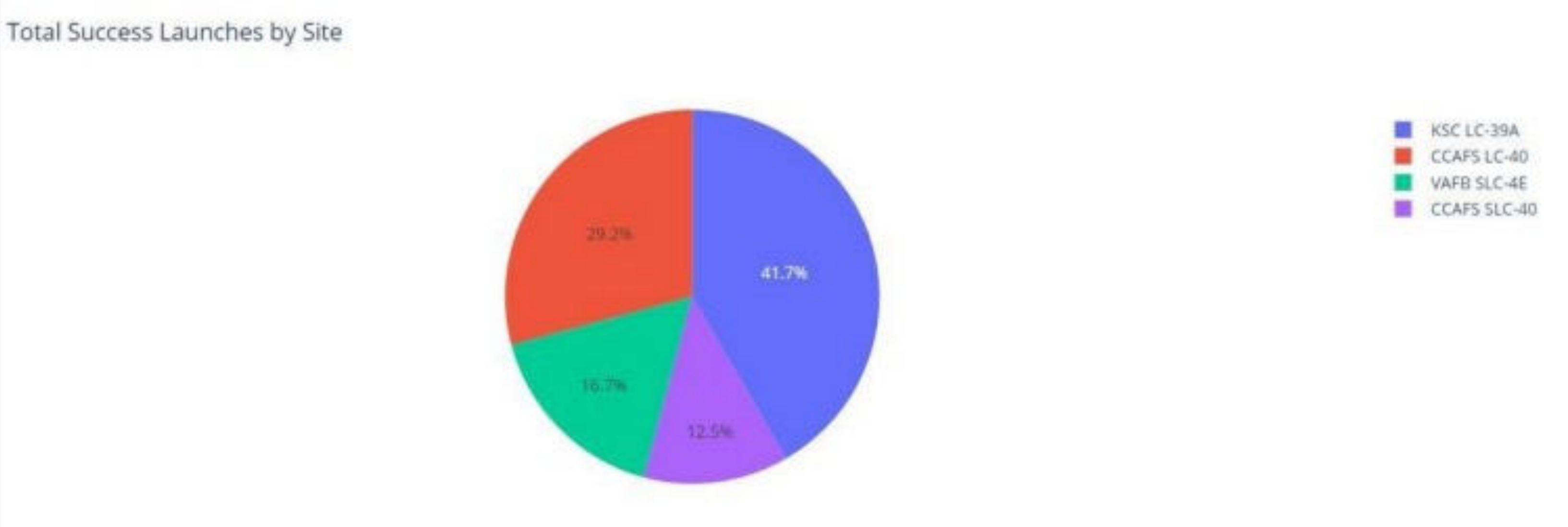


- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this exampleVAFB SLC-4E shows 4 successful landings and 6 failed landings.

Section 4

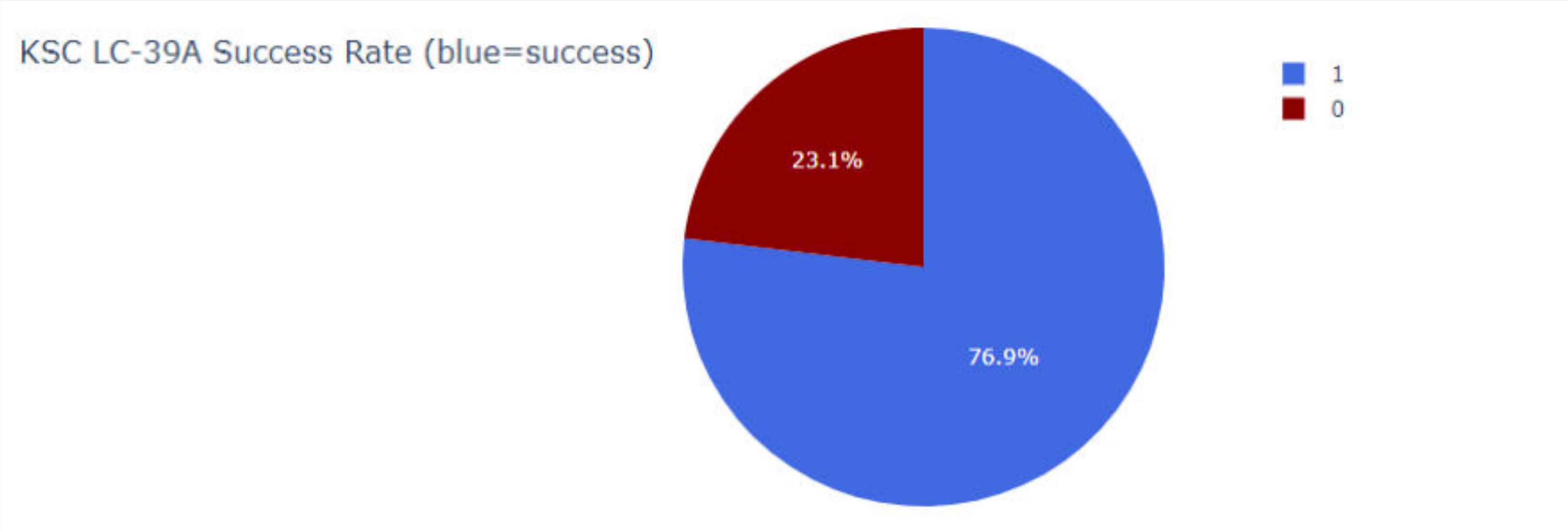
Build a Dashboard with Plotly Dash

Successful LaunchesAcrossLaunchSites



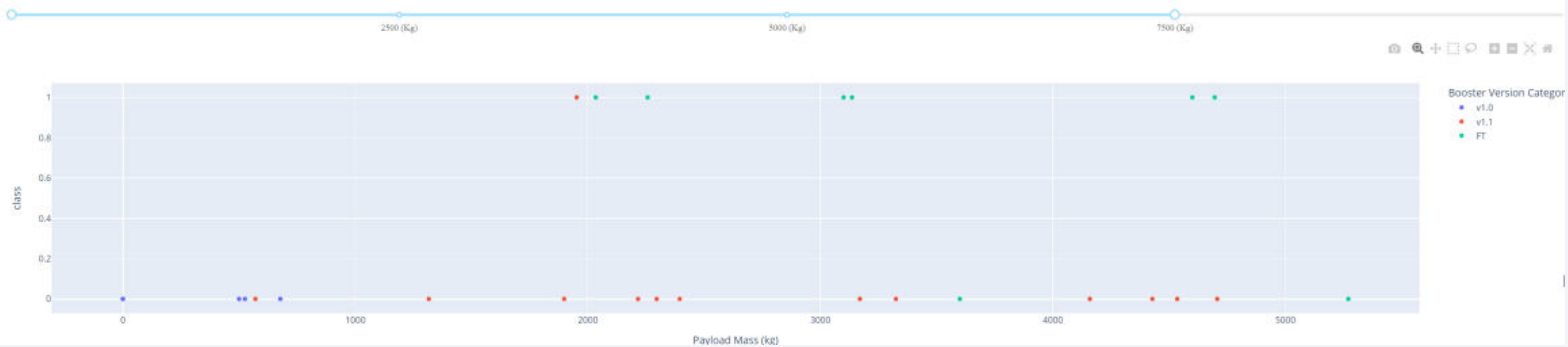
- This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This maybe due to smaller sample and increase in difficulty of launching in the west coast.

Highest Success RateLaunchSite



- KSC LC-39A has the highestsuccess rate with 10 successful landings and 3 failed landings.

PayloadMass vs. Successvs. Booster Version Category

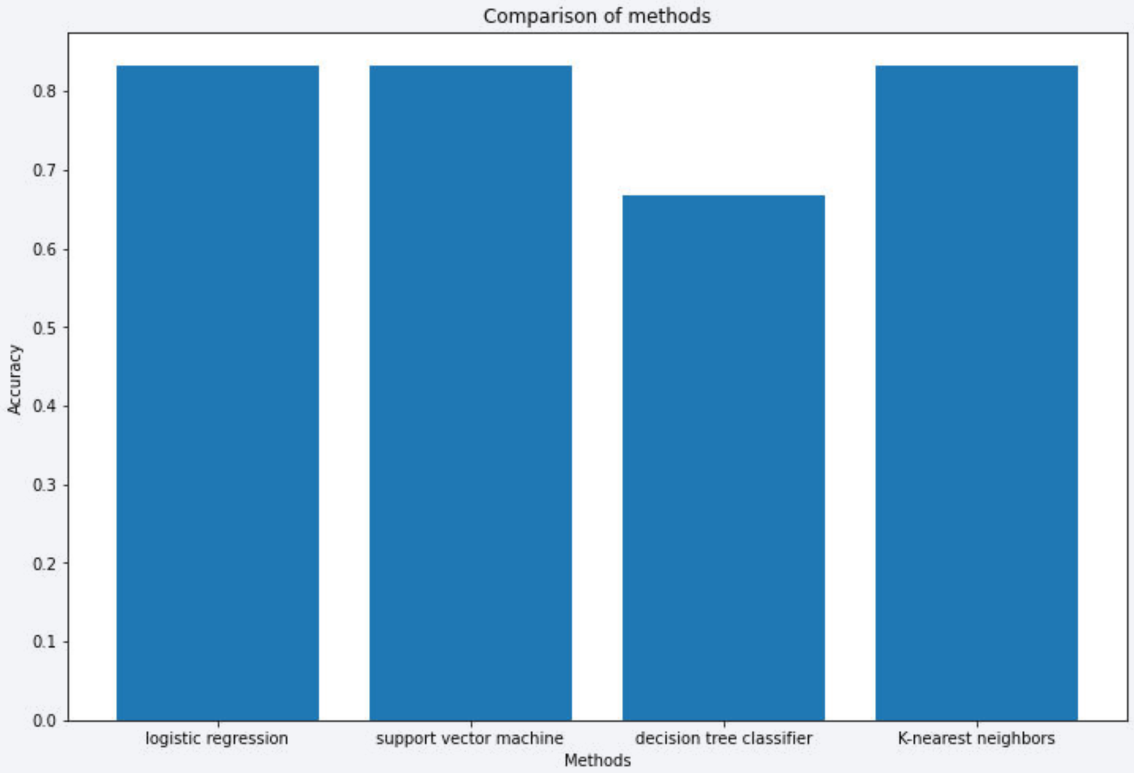


- Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-7500,interestinglythere are two failed landings with payloads ofzero kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- The models had virtually the same accuracy on the test set at 83.33%accuracy, except the decision tree classifier with 72,23 %.

- This can cause large variance in accuracy results,such as those in Decision TreeClassifier model in repeated runs. We likely need more data to determine the best model.

Confusion Matrix

- Since all models performed the same forthe test set, the confusionmatrix isthe same across all models.
- The models predicted 12 successful landings when the true label wassuccessful landing.
- The models predicted 3 unsuccessful landingswhen the true label was unsuccessful landing.
- The models predicted 3 successful landingswhen the true label was unsuccessful landings(false positives).

Conclusions

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- • The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- We can see that KSC LC-39A had the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Appendix

- Github repository url :
  - https://github.com/yassine-rd/IBM-Data-Science-Professional-Certification-Capstone

- Special Thanks to All Instructors
  - https://www.coursera.org/professional-certificates/ibm-data-science

Thank you!