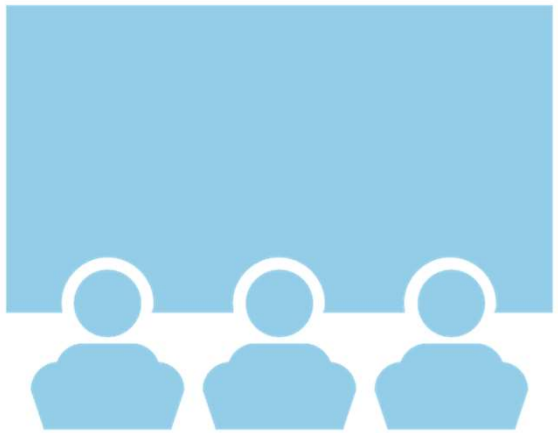
A Prediction model for Forcasting if Space Rockets can be Reused

Data Science Capstone Project: ‘Better Launch Rockets that Come Back’

<Doğan ONAY>

<24.08.2021>

Outline



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

Executive Summary

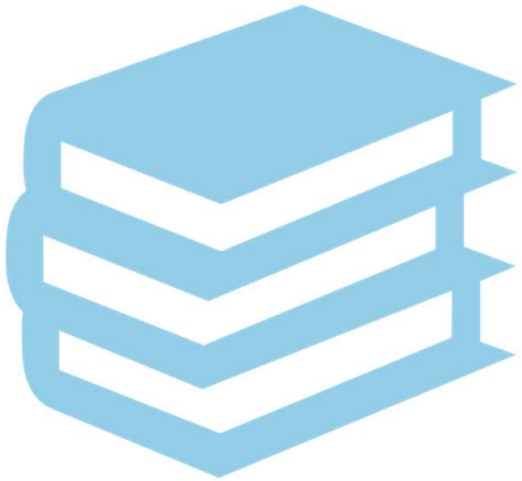
Summary of methodologies

- Coolect data with SpaceX Rest API and web scraping on related Wikipedia page
- Explore data with plots and graphs (scatter plot, bar chart, line chart)
- Explore data with sql queries.
- Create folium maps with coordinates of launch sites.
- Create Dash to Show correlation of variables interactively
- Use classification model with best accuracy to predict if the landing outcome will be successful

Summary of all results

- SpaceX company is using three different launch facility(officially 4 but 2 of them practically at same place)
- Maximum payload mass carried is 15600 kg and used one type of booster
- Launches carried out with reused rockets have %86 success rate
- Launches carrying 6800kg and more have %88 success rate
- Launches to some orbits are more often
- Launch sites have common features(close to ocean, railways; far from city center)
- Logistic Regression, SVM and Knn models are equally good for the data

Introduction



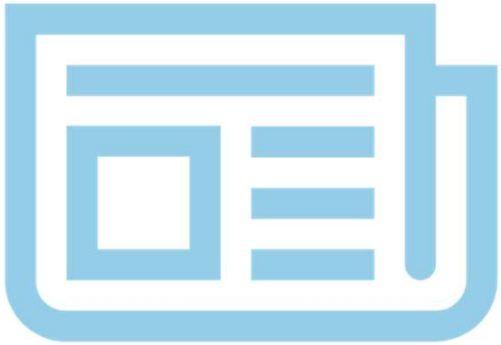
Backround

- World is in the commercial space age now and there are some companies working in this field. What they do is varied with space travel, sub-orbital space flights, satellite constelletion and so on. At the time being SpaceX company seems ahead amoung them with relatively inexpensive rocket launches.
- As a new rocket launch company named SpaceY that would like to compete with SpaceX. Gathering information about SpaceX and source of its success is crucial.
- Since the commercial success of SpaceX is stem from the reuse of first stage of rockets.

Problem

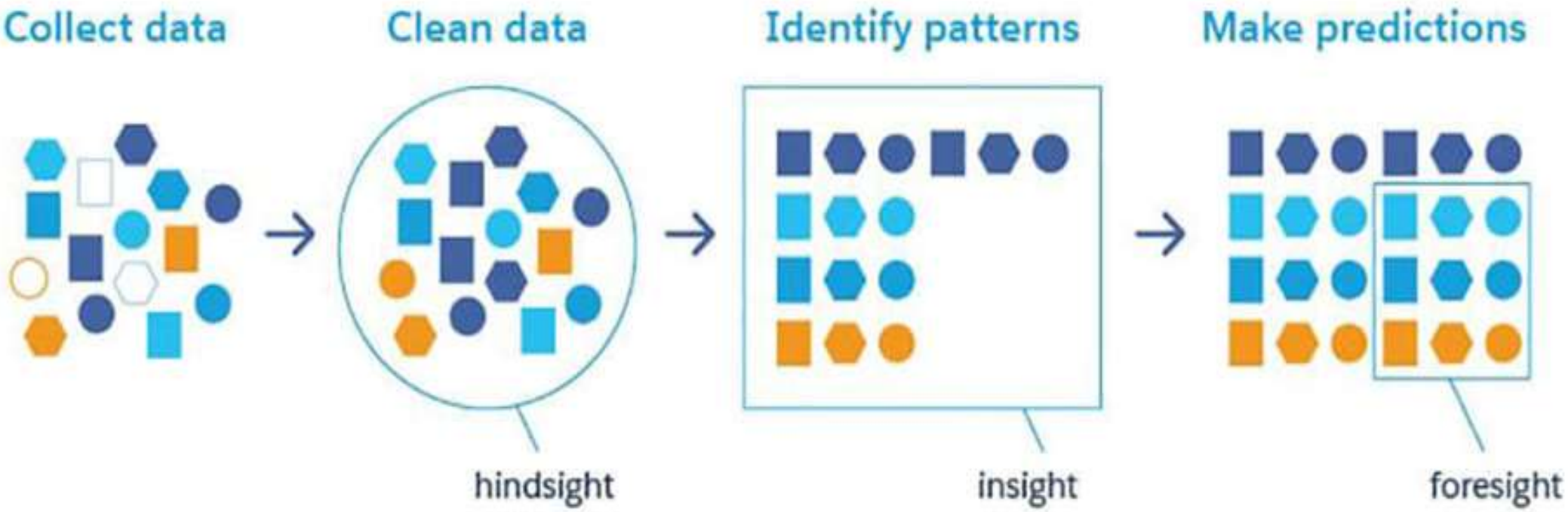
- We should be able to predict if the first stage of rocket land successfully ?

Methodology



- Data collection methodology:
  - Rest API, Web Scraping
- Perform data wrangling
  - Landing Outcomes are reduced into success and failure
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Methodology



SOURCE: amadeus.com

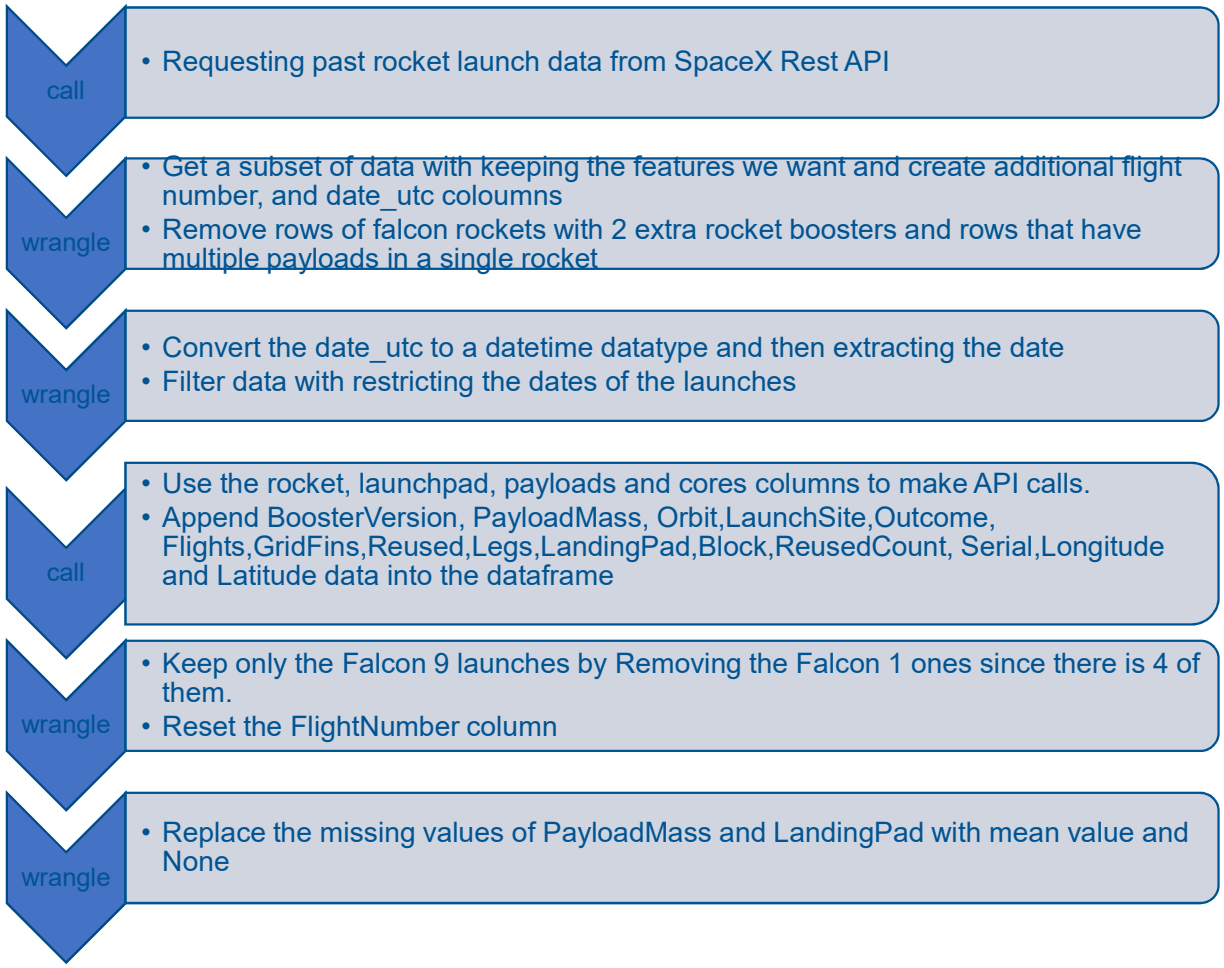
Data collection

Collection of data about SpaceX company's rocket launches

- Make SpaceX Rest API calls to collect data
- Wep scrape on List of Falcon 9 and Falcon Heavy launches Wikipage
- Create Class column showing the landing outcome of each launch

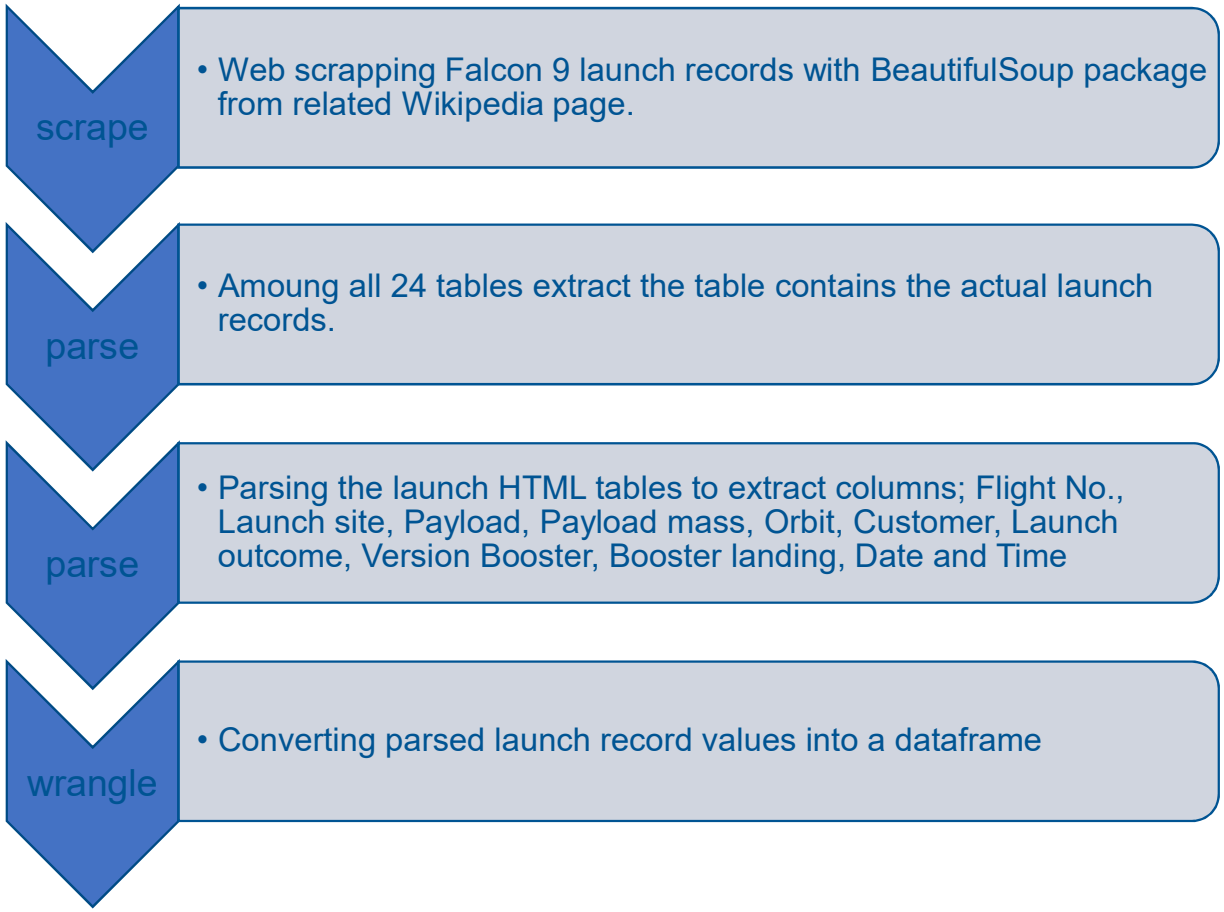
Data collection – SpaceX API

https://github.com/Dogan-87/IBM_Data_Science_Capstone_Project/blob/master/Week1-1-spacex-data-collection-api.ipynb



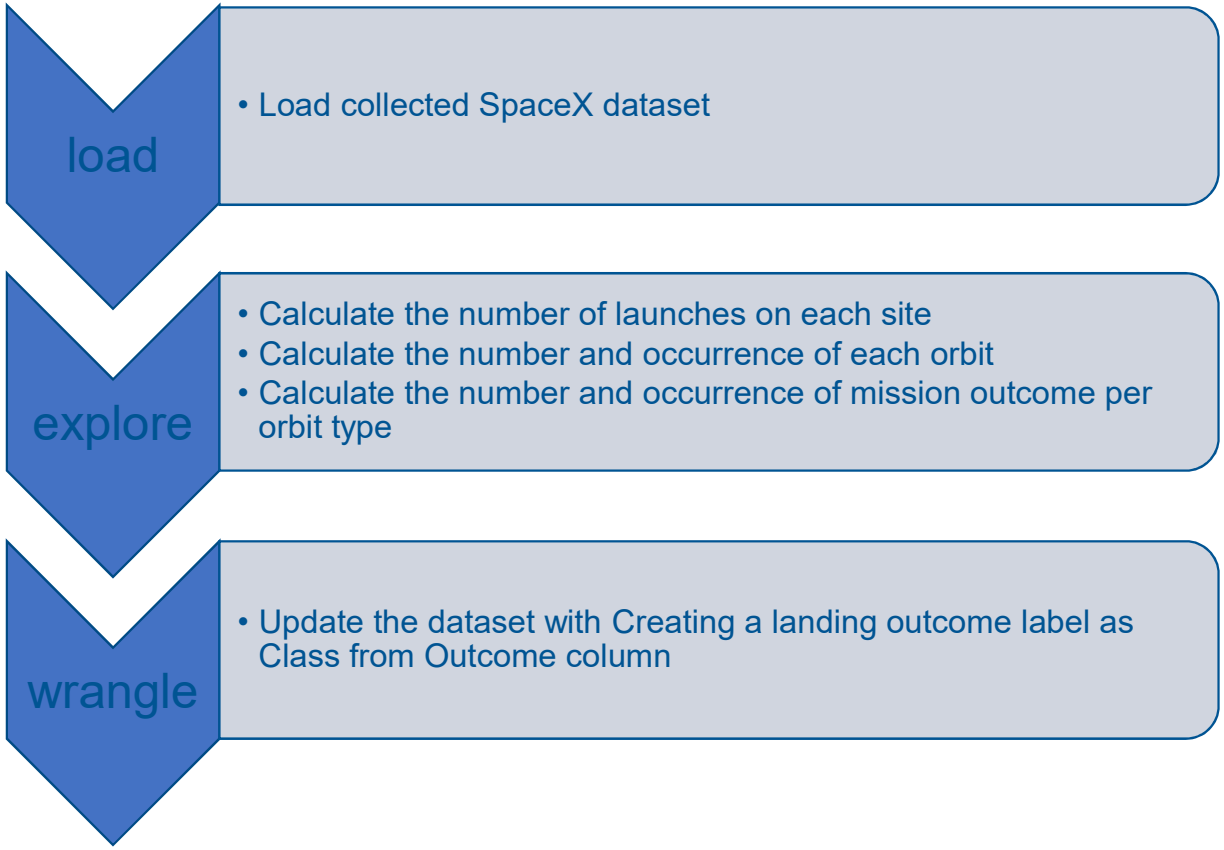
Data collection – Web scraping

https://github.com/Dogan-87/IBM_Data_Science_Capstone_Project/blob/master/Week1-2-webscraping.ipynb



Data wrangling

https://github.com/Dogan-87/IBM_Data_Science_Capstone_Project/blob/master/Week1-3-spacex-Data%20wrangling.ipynb



EDA with data visualization

Created Graphs:

- Scatter plot of Flight Number vs. Launch Site
  - To visualize the relation launch sites with landing success
- Scatter plot of Payload vs. Launch Site
  - To visualize how the landing success of launch sites changes with payload mass
- Barchart for the success rate of each orbit type
  - To visualize the success rate of each orbit
- Scatter point of Flight number vs. Orbit type
  - To visualize how the Orbit types of launches and success state changes with increasing flight numbers
- Scatter point of payload vs. orbit type
  - To visualize how the landing success of orbit types changes with payload mass
- Line chart of yearly average success rate
  - To visualize how the average success rate of launches change in time

Create dummy variables to categorical columns

https://github.com/Dogan-87/IBM_Data_Science_Capstone_Project/blob/master/Week2-1-eda-dataviz.ipynb

EDA with SQL

Query For;

- All launch site names
- Launch site names begin with `CCA`
- Total payload mass NASA (CRS)
- Average payload mass by F9 v1.1
- First successful ground landing date
- Successful drone ship landing with payload between 4000 and 6000
- Total number of successful and failure mission outcomes
- Boosters carried maximum payload
- 2015 launch records
- Rank success count between 2010-06-04 and 2017-03-20

https://github.com/Dogan-87/IBM_Data_Science_Capstone_Project/blob/master/Week2-2-eda-sql-coursera.ipynb

Build an interactive map with Folium

Created Map Objects:

- Circle()
  - To add a highlighted circle area with a text label on launch sites
- Marker()
  - To show each launch site name with icon property
- Marker Cluster()
  - To show multiple launches from the same launch sites with success state.
- Polyline()
  - To show the distance of specific coordinates to choosen launch site.

https://github.com/Dogan87/IBM_Data_Science_Capstone_Project/blob/master/Week3-1-launch-site-location.ipynb

Build a Dashboard with Plotly Dash

Created Plots/Graphs:

- Pie Charts
  - To show the total successful launches count for all sites
  - To Show the Success vs. Failed counts for each sites
- RangeSlider
  - To select payload range
- Scatter Charts
  - To show the correlation between payload and launch success

https://github.com/Dogan87/IBM_Data_Science_Capstone_Project/blob/master/Week3-2-Dash-SpaceX.ipynb

Predictive analysis (Classification)

https://github.com/Dogan-87/IBM_Data_Science_Capstone_Project/blob/master/Week4-1-SpaceX-Machine%20Learning%20Prediction-Part_5.ipynb

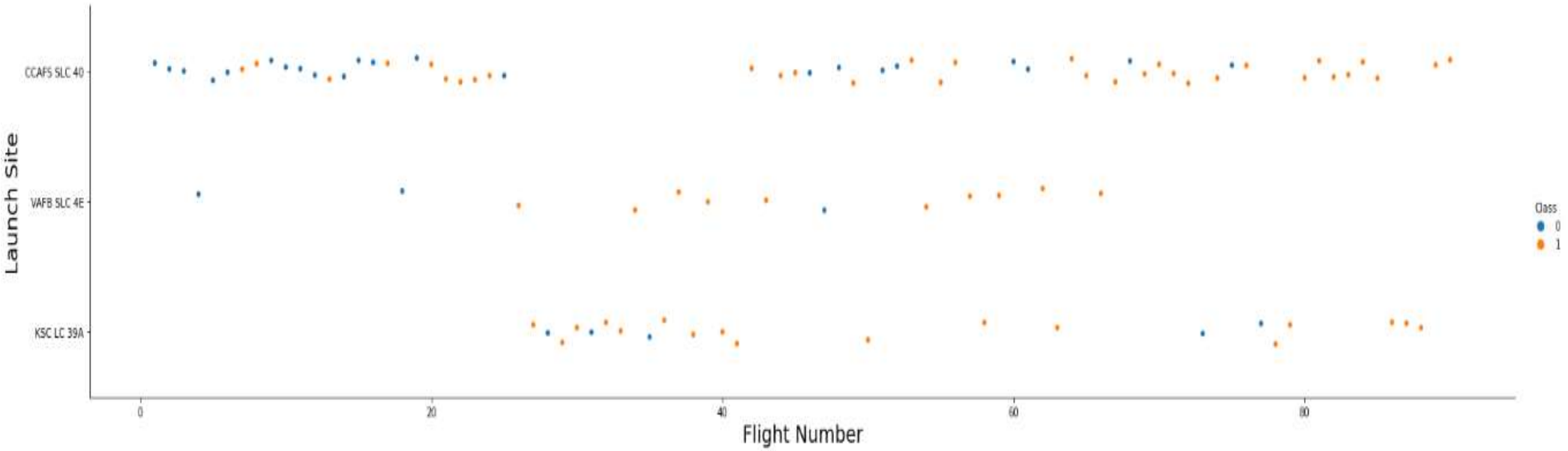
- Create X and Y data as variables
- Standardize the variables
- Apply train_test_split metod
- Fit train sets into a Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbor models
- Find best parameters for each model
- Calculate model accuracies on test set
- Plot Confusion Matrixes
- Compare the methods
- Plot Bar chart to İllustrate the performance of the models
- Use the model with best performance to find out how the accuracy changes when specific test set is used to see if hypotesis is true.

Results

- Exploratory data analysis results
  - SpaceX company is using three different launch facility(officially 4 but 2 of them practically at same place)
  - Maximum payload mass carried is 15600 kg and used one type of booster
  - Launches carried out with reused rockets have %86 success rate
  - Launches carrying 6800kg and more have %88 success rate
  - Launches to some orbits are more often
- Interactive visual analytics results
  - Launch sites of SpaceX company spread two sides of the US.
  - When founding Launch sites to foud it is important to be close the ocean and railways and put some distance to city centre.
- Predictive analysis results
  - Logistic Regression, SVM and Knn models provide the same accuracy rate which is %83
  - Decision Tree model gives %66 accuracy

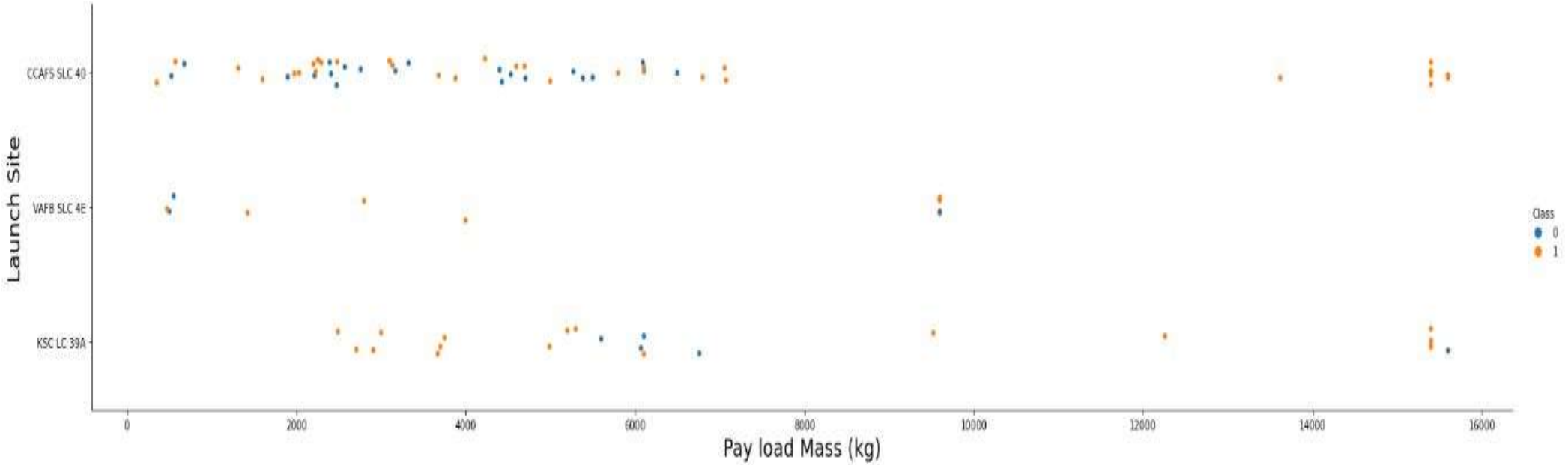
EDA with Visualization

Flight Number vs. Launch Site



- CCAFS SLC 40 launch site has the highest number of launches (55) nearly two times of others.
- On the other hand CCAFS SLC 40 has lowest success rate of %66 while other launch sites share the %77 success rate
- Between 25-42 launches CCAFS SLC 40 launch site not used
- After 65th launch launch site VAFB SLC 4E not used

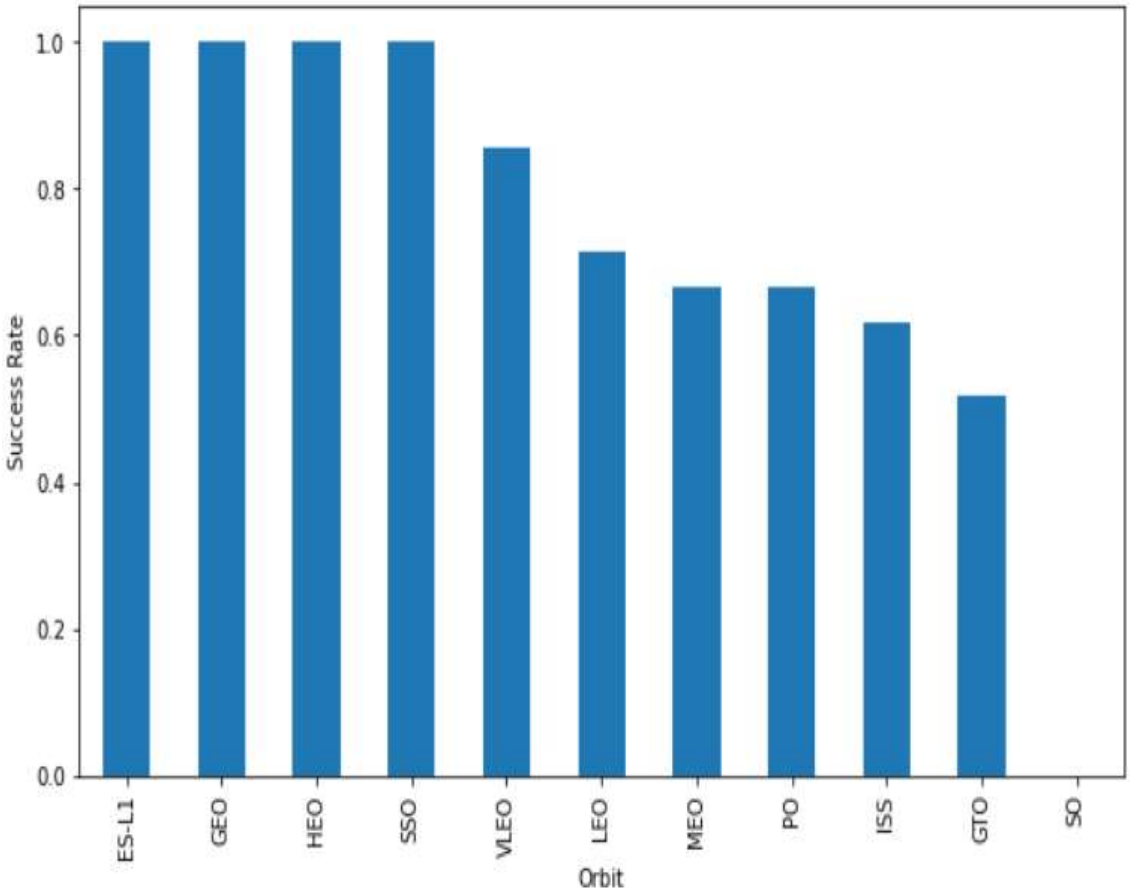
Payload vs. Launch Site



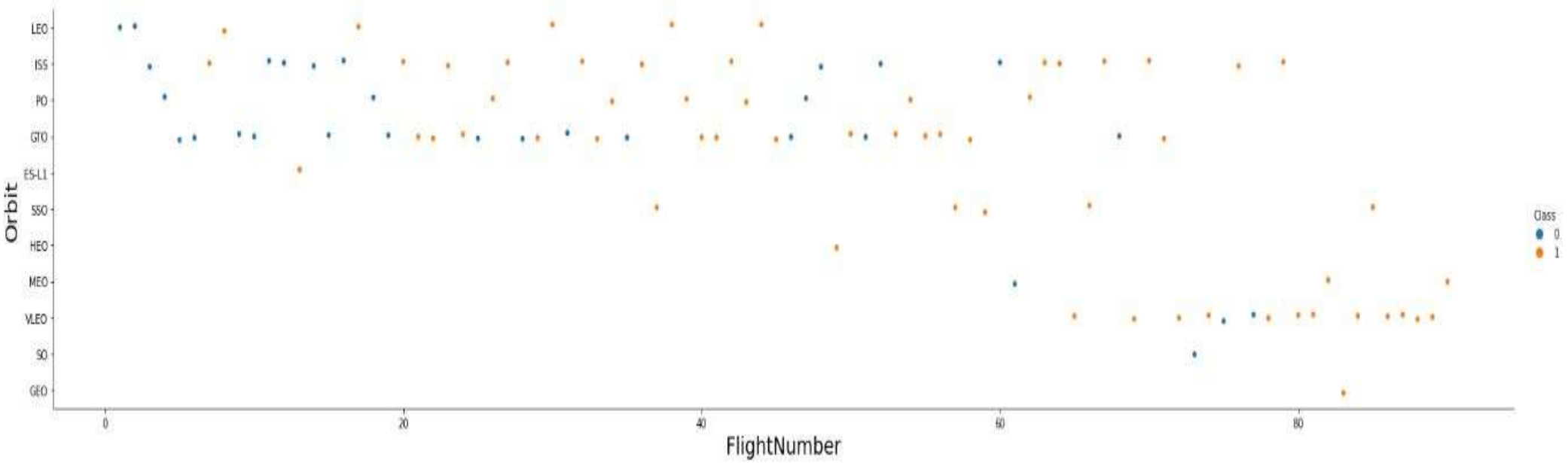
- 2/3 of all launches are under 6800kg of Payload Mass
- CCAFS SLC 40 launch site has the widthest range of payload Mass
- KSC LC 39A launch site has the lowest range of payload Mass
- Launches with over 6800kg of Payload Mass has %88 success rate
- Launches with under 6800kg of Payload Mass has %44 success rate

Success rate vs. Orbit type

- ES-L1, GEO, HEO and SSO orbits have %100 success rate
- Since there is only one launch to each of ES-L1, GEO and HEO orbits their success rates are not creditible
- On the other hand SSO with 5 and VLEO with 14 launches have %100 and %86 success rates
- SO orbit has %0 success rate with only 1 launch. Not enough data to evaluate it like ES-L1, GEO and HEO orbits

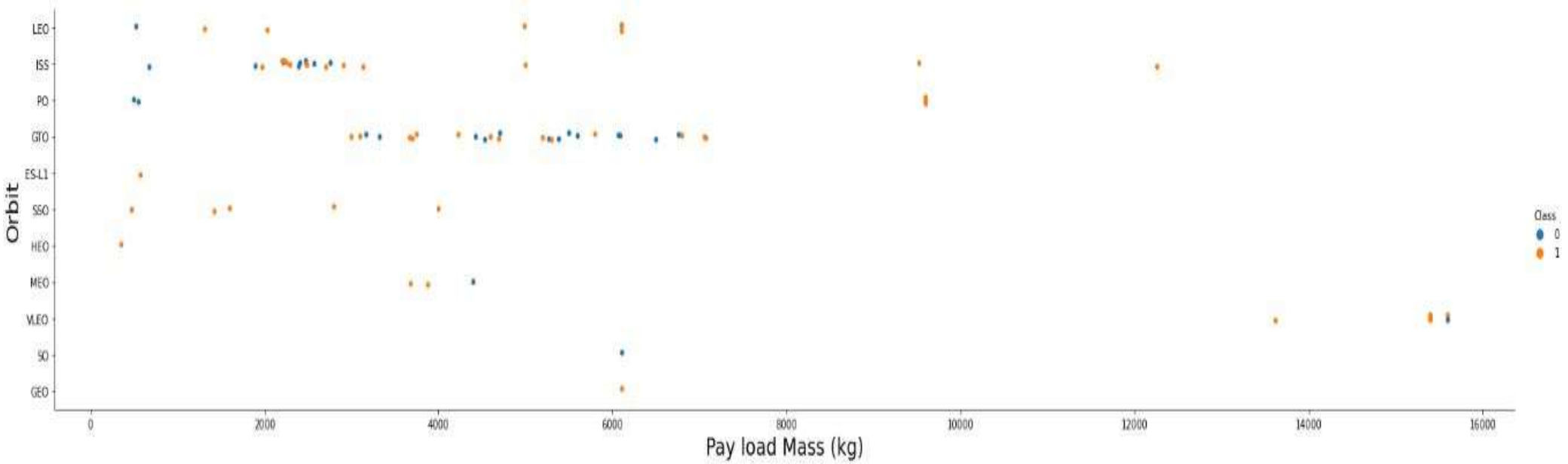


Flight Number vs. Orbit type



- GTO orbit has highest number of launches
- LEO, ISS, PO and GTO orbits relativly dense launch traffic
- Only after 64th launch VLEO orbit launches are started.
- ES-L1, SSO, HEO, MEO, SO and GEO orbit launches occurs rarely.

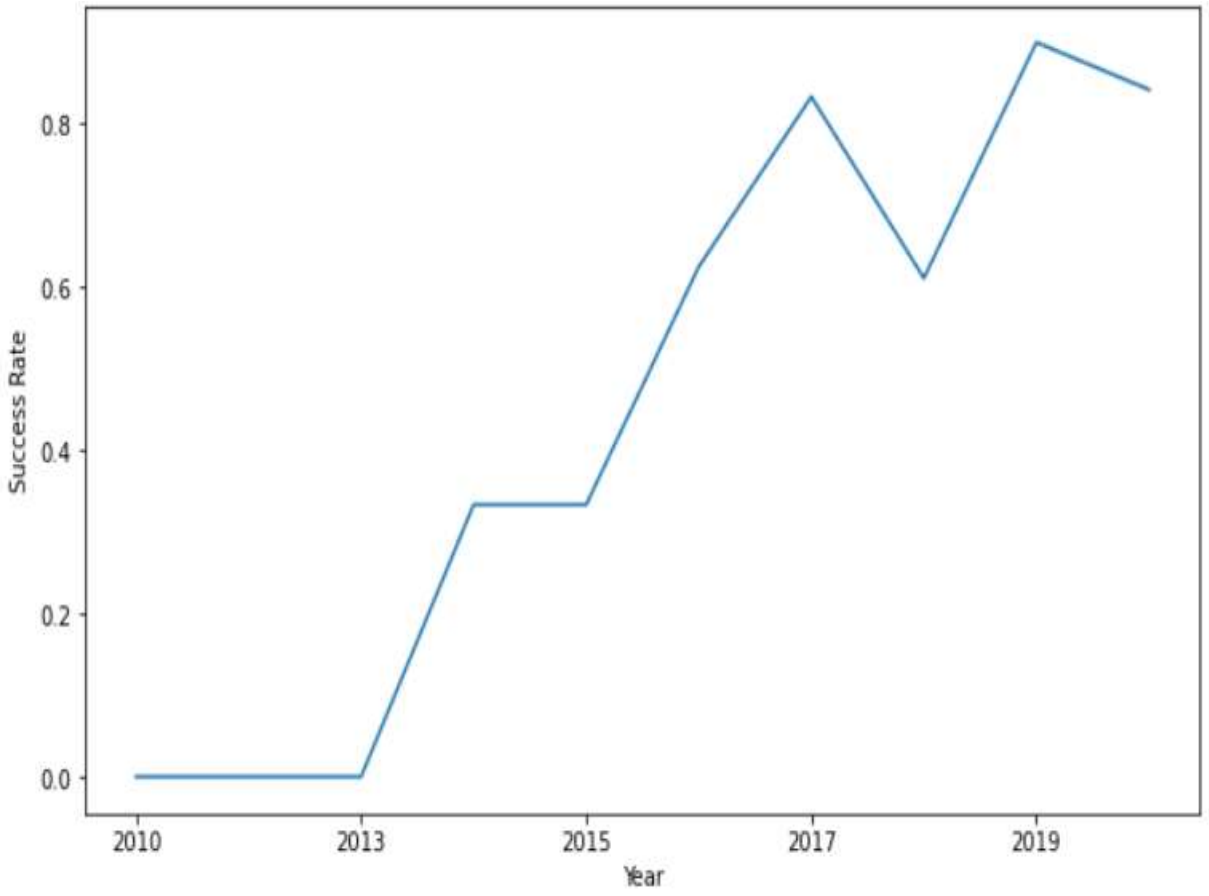
Payload vs. Orbit type



- VLEO orbit have launches with highest Payload Masses
- PO orbit launches have only two payload masses of 500kg (all fail) and 9600 kg (%83 success rate)
- SEO, GEO, HEO, ES-L1 and MEO have relatively narrow payload mass range
- Orbit called ISS has widest payload mass range

Launch success yearly trend

- There is a rising trend of success rate according to years.
- Between the years 2010-2013 success rate is zero and number of launches changes from 0 to 3.
- After 2013 success rate increase sharply and reaches %90 in 2019.
- In the periods of 2017-2018 and 2019-2020 success rate is decreasing



EDA with SQL

All launch site names

%sql SELECT DISTINCT launch_site from SPACEXTBL;

- Names of the unique launch sites in the space mission are as in the figure
- There are 5 of them but it seems different typing styles causes that.

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

Launch site names begin with `CCA`

%sql SELECT launch_site, COUNT(launch_site) AS number_of_launches from SPACEXTBL where launch_site LIKE 'CCA%' GROUP BY launch_site;

- Since there is typing difference on launch site Cape Canaveral Space Launch Complex 40 it is needed to gather them all.
- There are three different typing for the same launch site.
- Total number of launches for the Cape Canaveral Space Launch Complex 40 is 60.

launch_site	number_of_launches
CCAFS LC-40	26
CCAFS SLC-40	33
CCAFSSLC-40	1

Total payload mass NASA (CRS)

%sql SELECT customer, SUM(payload_mass__kg_) AS total_payload from SPACEXTBL GROUP BY customer HAVING customer='NASA (CRS)';

customer	total_payload
NASA (CRS)	45596

- Total payload mass carried by boosters launched by NASA (CRS) as in the output above.

Average payload mass by F9 v1.1

%sql SELECT booster_version, AVG(payload_mass__kg_) AS average_payload from SPACEXTBL GROUP BY booster_version HAVING booster_version='F9 v1.1';

booster_version	average_payload
F9 v1.1	2928

- Average payload mass carried by booster version F9 v.1.1 is 2928 kg.

First successful ground landing date

%sql SELECT MIN(DATE) AS first_success_date, landing__outcome from SPACEXTBL GROUP BY landing__outcome order by first_success_date;

- In the table of First dates of landing outcomes 2015-12-22 shows date of first succesfull landing on the ground.
- Until 2015 it seems the landing plans didn't cover reuse of boosters.

first_date	landing__outcome
2010-06-04	Failure (parachute)
2012-05-22	No attempt
2013-09-29	Uncontrolled (ocean)
2014-04-18	Controlled (ocean)
2015-01-10	Failure (drone ship)
2015-06-28	Precluded (drone ship)
2015-12-22	Success (ground pad)
2016-04-08	Success (drone ship)
2018-07-22	Success
2018-12-05	Failure

Successful drone ship landing with payload between 4000 and 6000

%sql SELECT DATE, booster_version, landing__outcome, payload_mass__kg_ from SPACEXTBL WHERE landing__outcome='Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000 ;

DATE	booster_version	landing__outcome	payload_mass__kg_
2016-05-06	F9 FT B1022	Success (drone ship)	4696
2016-08-14	F9 FT B1026	Success (drone ship)	4600
2017-03-30	F9 FT B1021.2	Success (drone ship)	5300
2017-10-11	F9 FT B1031.2	Success (drone ship)	5200

- As it seems in the output figure there are 4 successful drone ship landing with payload between 4000 and 6000 for 2 year period.

Total number of successful and failure mission outcomes

%sql SELECT mission_outcome, COUNT(mission_outcome) AS number_outcomes from SPACEXTBL GROUP BY mission_outcome;

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

- According to data 100/101 of launches missions are succesfully completed.
- Mission outcome and landing outcome are not the same.
- Landing could intentionaly be planned as fail according to cost, customer and etc.

Boosters carried maximum payload

```
%sql SELECT booster_version, payload_mass__kg_ from SPACEXTBL WHERE payload_mass__kg_=(SELECT MAX(payload_mass__kg_) FROM SPACEXTBL) ORDER BY booster_version ;
```

- Maximum payload mass carried in launches is 15600 kg
- F9 B5 booster with sub-versions listed in the output table is capable of carrying the maximum weight of recorded launches.

booster_version	payload_mass__kg_
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

%sql SELECT DATE, MONTHNAME(DATE) AS month, landing__outcome, booster_version, launch_site from SPACEXTBL WHERE YEAR(DATE) = 2015 AND landing__outcome='Failure (drone ship)‘

DATE	MONTH	landing__outcome	booster_version	launch_site
2015-01-10	January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The outcome above shows failure landing_outcomes in drone ship in the year of 2015.
- There are two failure within three mounths.

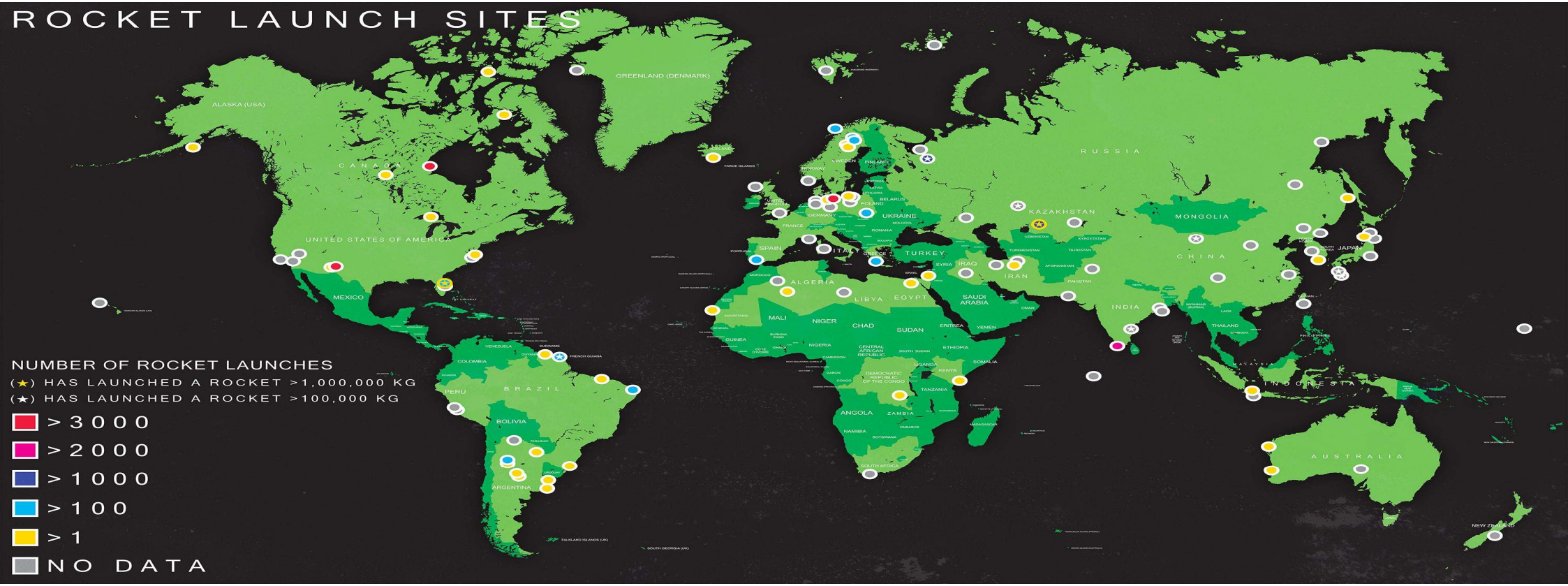
Rank success count between 2010-06-04 and 2017-03-20

```
%sql SELECT landing__outcome, COUNT(landing__outcome) AS count_of_success from SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND landing__outcome='Success' OR landing__outcome='Success (drone ship)' OR landing__outcome='Success (ground pad)' GROUP BY landing__outcome
```

landing__outcome	count_of_success
Success (drone ship)	14
Success (ground pad)	9

- There are 23 successful landing outcome between 2010-06-04 and 2017-03-20
- There are 2 different successful outcome these are drone ship and ground pad
- Since we know that first dates of each outcome type. We could get the same result with changing the starting date with 2015-12-22.

Interactive map with Folium



Source:mapsontheweb.zoom

All Launch Sites in Folium map

- In the above map all launch sites of SpaceX company namely CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E are shown

Success state of Launch Sites

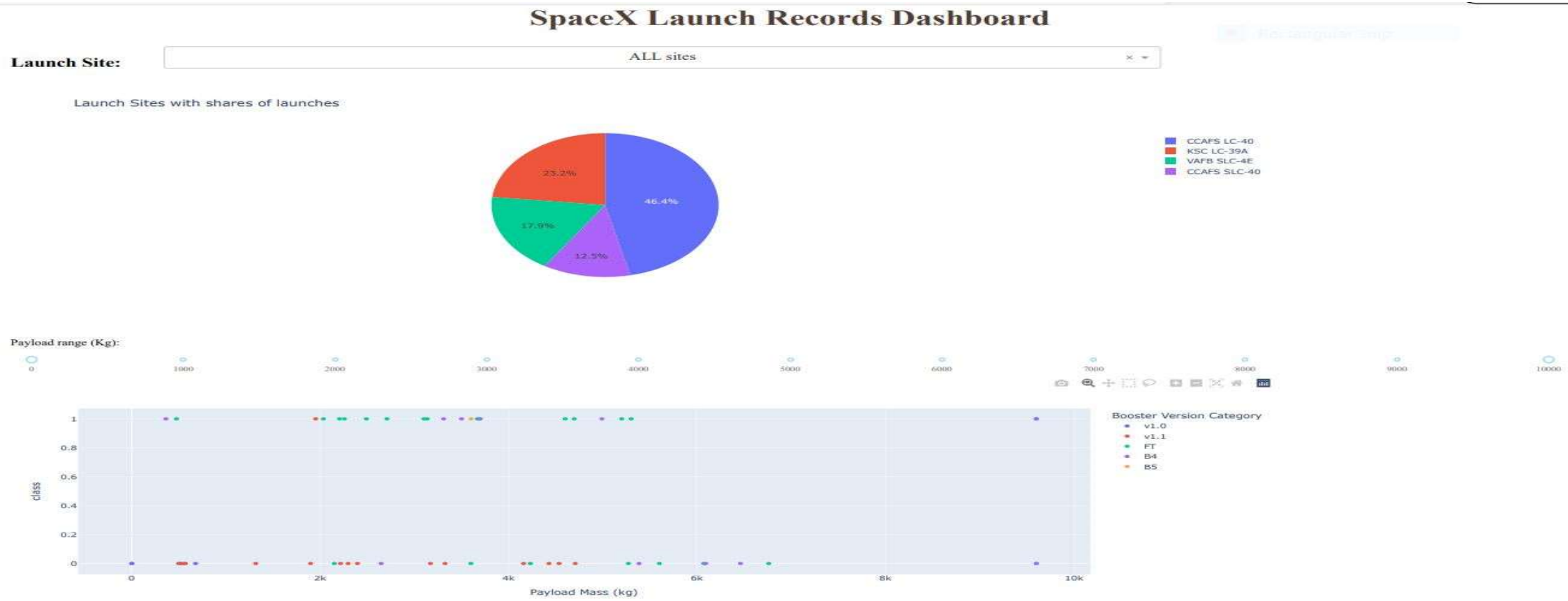
- By using cluster mark the success/failed launches for each site are shown on the map

Launch Site proximities

It can be inferred from the illustration that launch sites are;

- found distant from the cities (23 km away)
- placed close to railways (1.33 km)
- proximite to ocean (1.03 km)

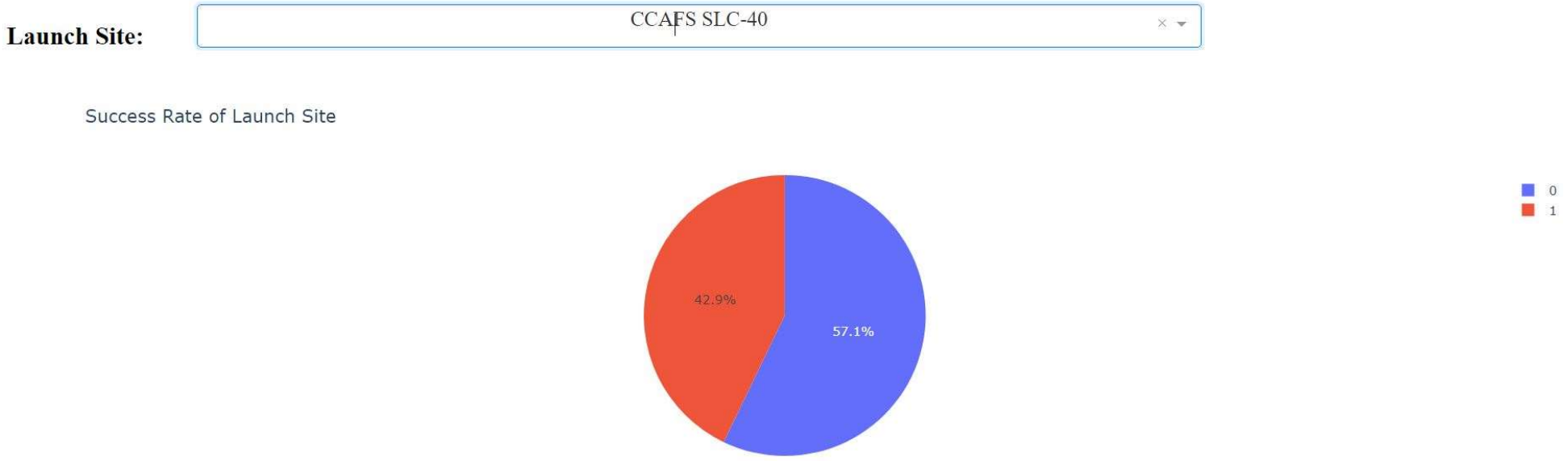
Build a Dashboard with Plotly Dash



SpaceX rocket launch distribution by sites

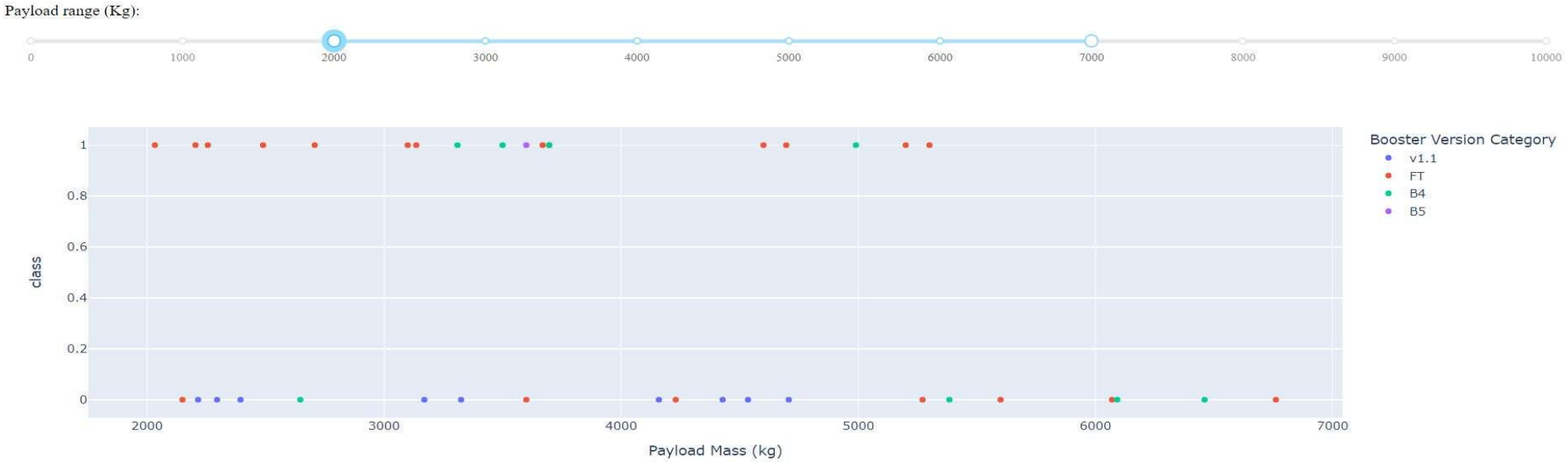
- %46.4 of the total launches are performed from CCAFS LC-40
- For only %12 of the launches CCAFS SLC-40 are used

Launch Site Success Ratio



- CCAFS SLC-40 with %42.9 score has the highest success ratio
- VAFB SLC-4E is coming right after it with %40 success ratio
- Launch Site KSC LC 39A has the lowest success ratio of %23.1

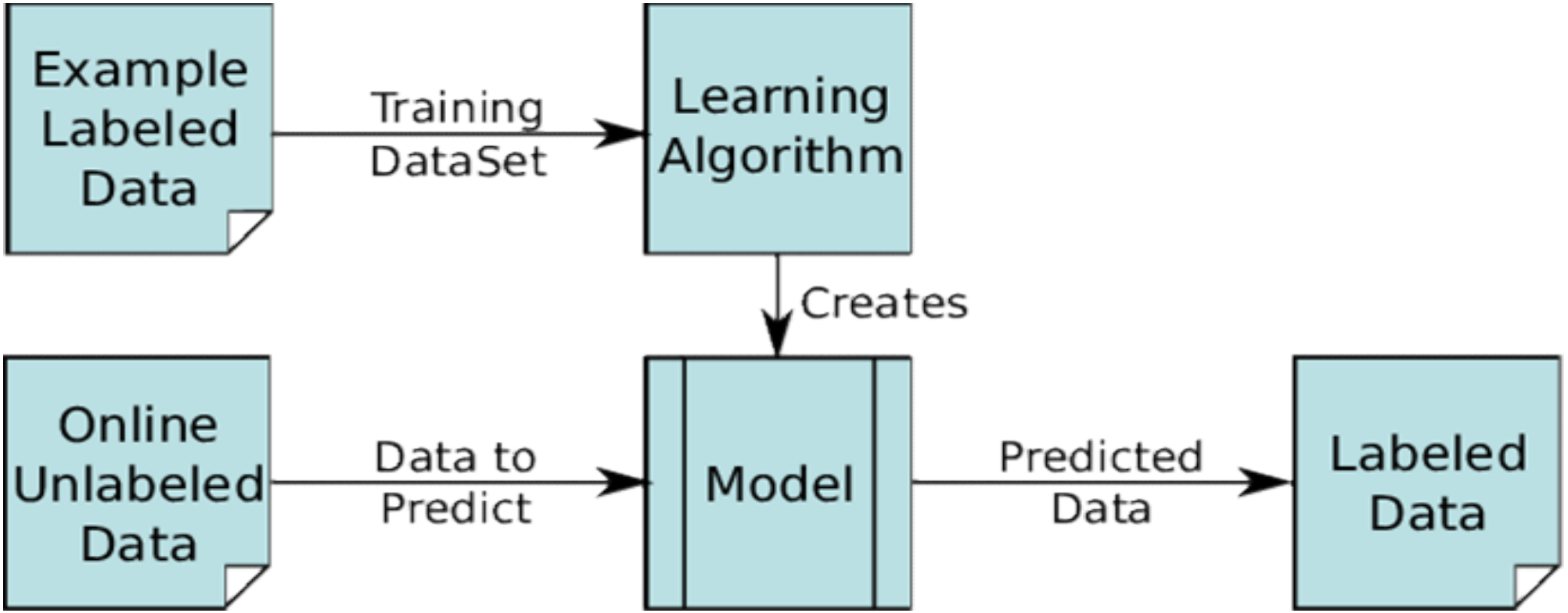
Payload vs.Launch Outcome for all Sites



For the payload mass between 2000 and 7000kg

- Booster version FT is used for over %60 of all successful launches
- All launches using Booster version v1.1 are end up with failure
- Booster version B4 has %50 success rate

Predictive analysis (Classification)

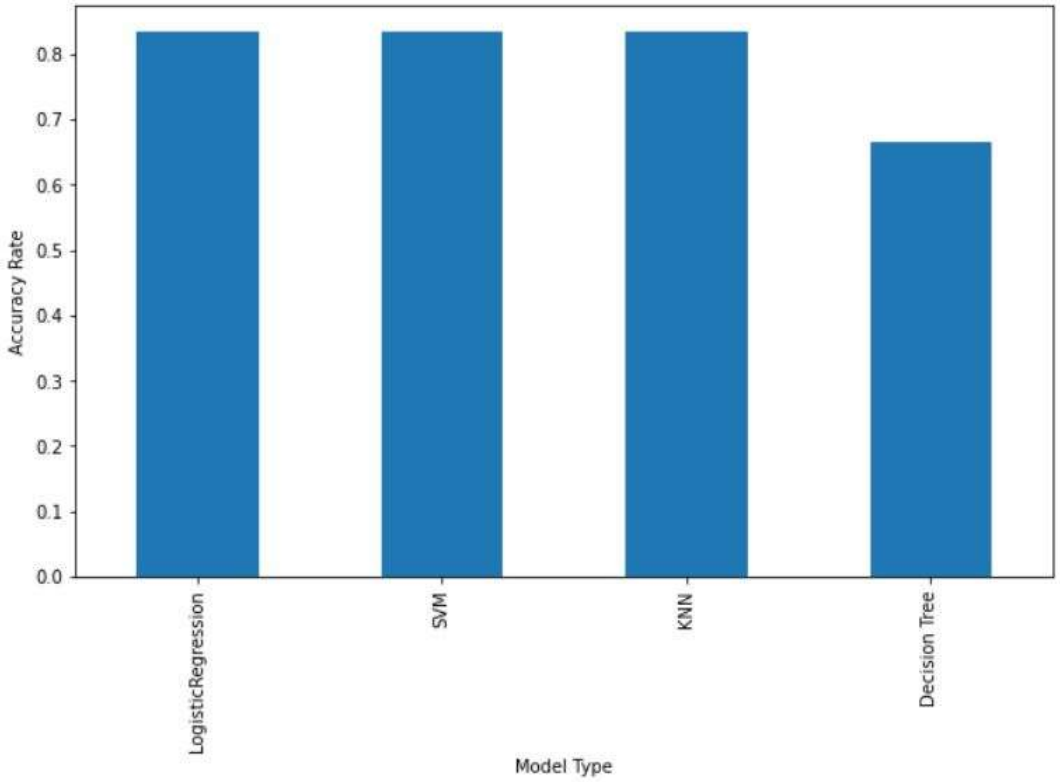


source:researchgate.net

Classification Accuracy

For the prediction models we use our data's 0.8 for train 0.2 for test.

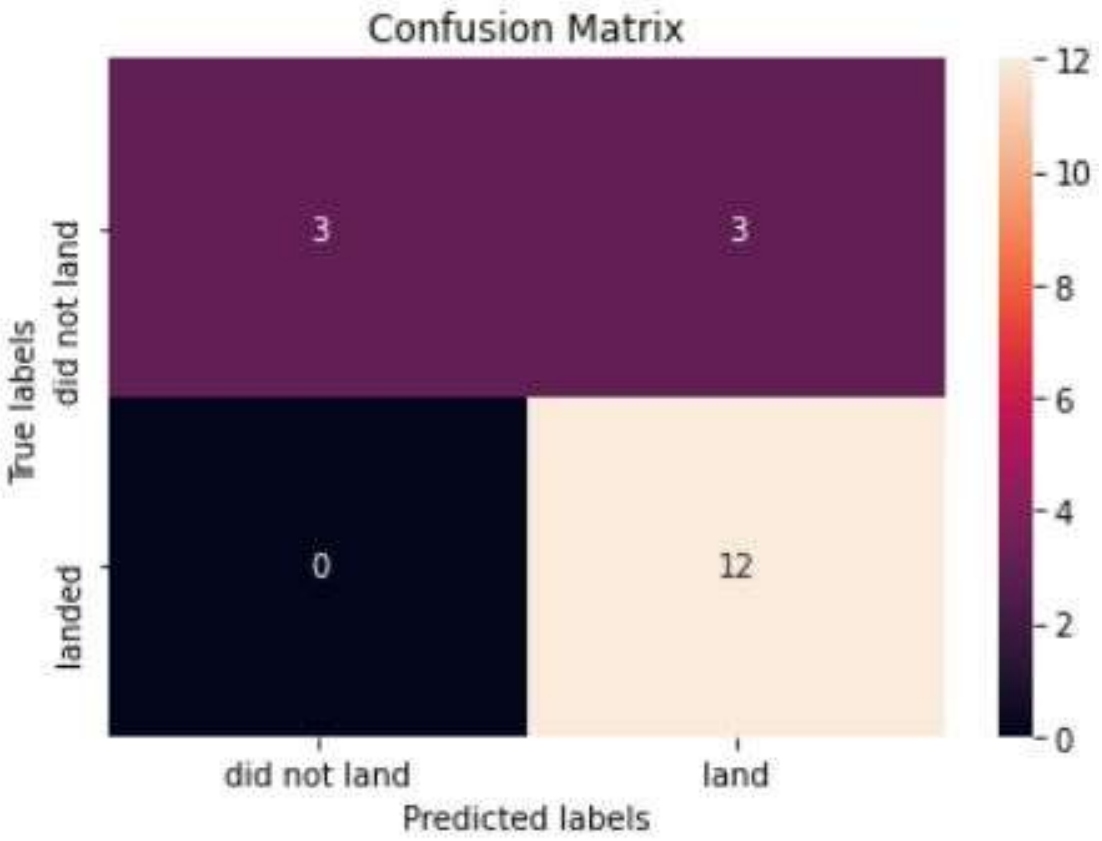
- Logistic Regression, SVM and Knn models provide %83 prediction rate for test data
- Decision Tree gives lower prediction rate with %66
- In the model of decision tree because of the parameters variety. Test accuracy could change greatly that is not the issue for other models.



Confusion Matrix

The matrix is belong to svm prediction model. There are 18 launch in test dataset.

- According to matrix the model predicted all 12 successful landings true
- Again 3 failed landings predicted correctly
- Our model predicted 3 failed landings as successful.
- Test accuracy is %83



CONCLUSION

https://github.com/Dogan-87/IBM_Data_Science_Capstone_Project/blob/master/Week4-1-SpaceX-Machine%20Learning%20Prediction-Part_5.ipynb

- Launches with over 6800kg of payload mass (26 launches) have 0.88 success rate
- launches using formerly successful rockets (60 launches) have 0.86 success rate
- Launches carry out with new rockets have lower success rate.
- In all launches over 6800kg of payload mass formerly successful rockets are used.
- When we use our models to predict above mentioned dataset of specific payload mass
  - Logistic Regression, SVM and Knn models successfully predict %85 of the data
  - The ratio of true prediction of the Decision Tree model is only %11
- It seems there is greater chance to predict the outcomes of launches with payload mass over 6800 kg.

APPENDIX

Algorithm	Train Accuracy Score (.accuracy_score())	Test Accuracy Score (.score())	GridSearchCV(.best_score_)	Yhat(.predict())
LogisticRegression	0.875000	0.833333	0.846429	[1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1]
SVM	0.888889	0.833333	0.848214	[1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1]
Decision Tree	0.819444	0.666667	0.889286	[1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0]
KNN	0.861111	0.833333	0.860952	[1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1]

- Report table created by using all data with 0.2 test size.

Algorithm	Train Accuracy Score (.accuracy_score())	Test Accuracy Score (.score())	GridSearchCV(.best_score_)	Yhat(.predict())
LogisticRegression	0.875000	0.846154	0.846429	[1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
SVM	0.888889	0.846154	0.848214	[1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Decision Tree	0.819444	0.115385	0.889286	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
KNN	0.861111	0.846154	0.860952	[1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

- Report table created by using same train set and data set fulfilling payload mass>=6800kg as test size.
- In decision tree model best parameters seems not suitable to predict test set accuratly.
- Using more convenient train set can lead to a boost in prediction rates.

APPENDIX

Confusion Matrix of specific payload test set data

- The matrix is belong to svm prediction model. There are 26 launch in test dataset.
- According to matrix the model predicted all 22 successful landings true
- Our model predicted 1 successful landings as failed.
- Our model predicted 3 failed landings as successful.
- Test accuracy is %85