IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rudolf Horvat
22.1.2023

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

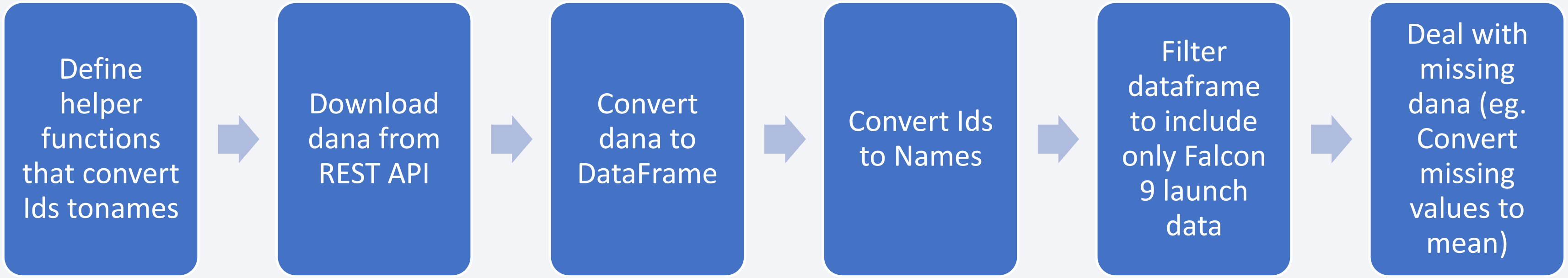
- Data collection methodology:
  - Spacex has published launch related data on their website
  - The data is made available in form of a JSON file from the following public URL https://api.spacexdata.com/v4/launches/past
- Perform data wrangling
  - Missing PlayloadMass dana was replaced with the mean value
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four classification models where fitted, the best model was selected based on crossvalidation and hyperparameter tuning

Data Collection

- Data in the form of a JSON file was downloaded from the spacex
  - REST API at https://api.spacexdata.com/v4/launches/past
    - The data was processed as described in slide Data Collection – SpaceX API
  - Wikipedia website
    - The data was processed as described in slide Data Collection - Scraping

Data Collection – SpaceX API

- GitHub URL: [capstone/data-collection.ipynb at master · rhorvatgm/capstone (github.com)]



Data Collection - Scraping

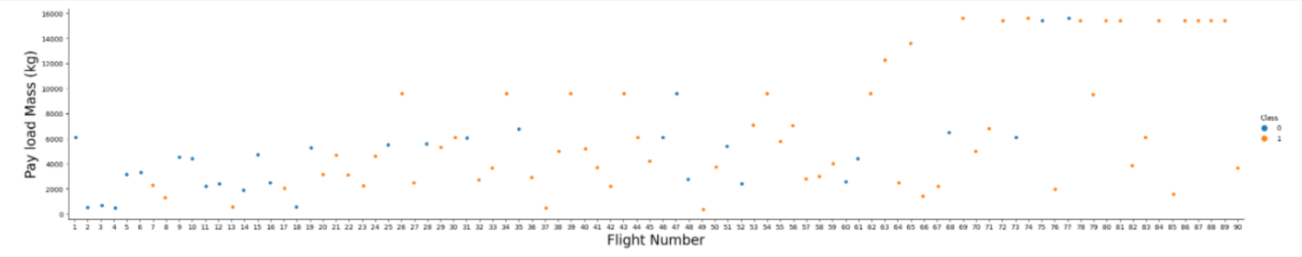
- GitHub URL: [capstone/webscraping.ipynb at master · rhorvatgm/capstone (github.com)]

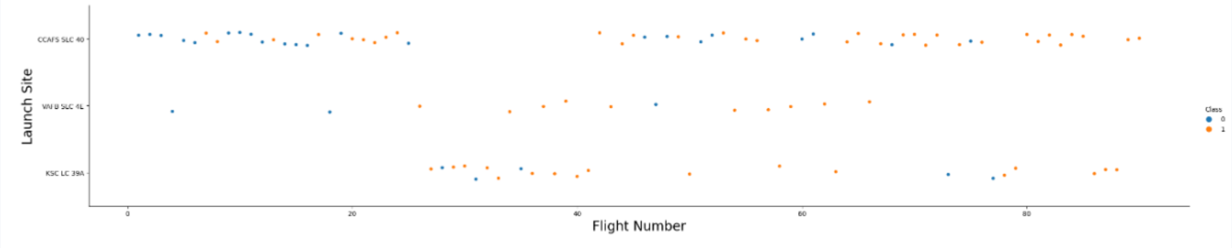
Data Wrangling

- Describe how data were processed
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

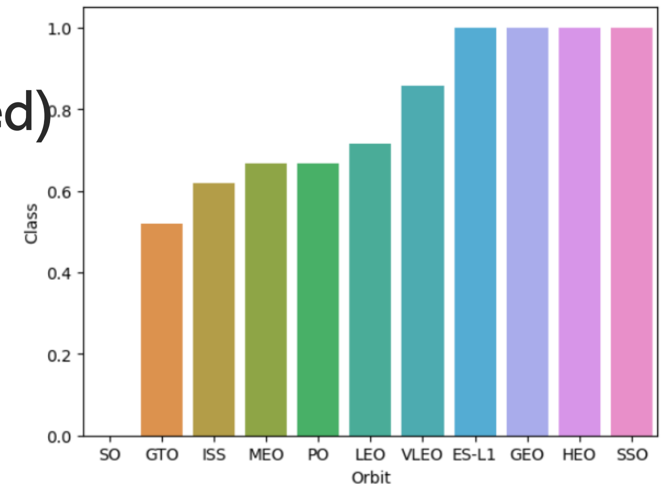
- GitHub URL: [capstone/eda-visualization.ipynb at master · rhorvatgm/capstone (github.com)]()
- Several charts were plotted
  - Flight no vs playload – to analyze how playload was increased and how success rate (indicated by datapoint color) depended on flight no and playload
  - Flight no vs Launch site
  - Launch site vs Payload mass

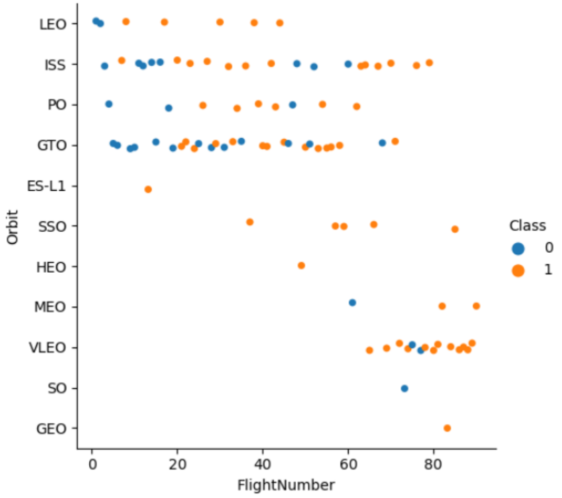


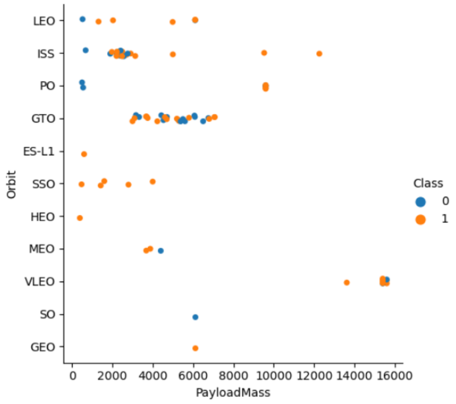


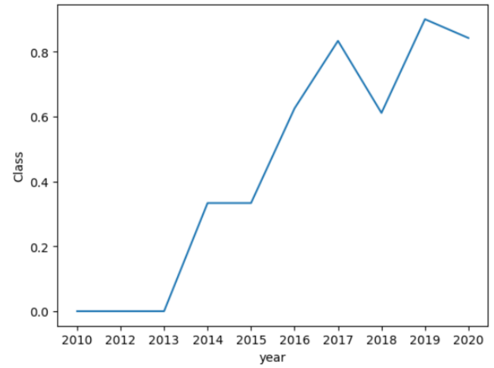
EDA with Data Visualization (continued)

- Several charts were plotted (continued)
  - Orbit vs Success rate
  - Flight no vs Orbit
  - Payload mass vs Orbit
  - Success rate over time









EDA with SQL

- Github URL: [capstone/eda-sql.ipynb at master · rhorvatgm/capstone (github.com)]
- Following SQL queries were performed:
  - Unique launch sites
  - 5 records where launch sites begin with the string 'CCA'
  - total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - date when the first successful landing outcome in ground pad was acheived.
  - 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 failure mission outcomes
  - names of the booster_versions which have carried the maximum payload mass
  - he failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - 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

Build an Interactive Map with Folium

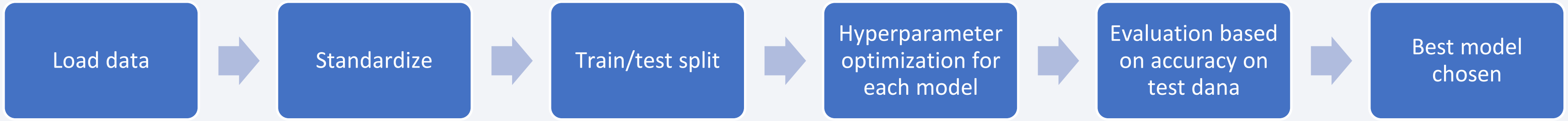
- GitHub URL: [capstone/interactive_visual.ipynb at master · rhorvatgm/capstone (github.com)]
- The following dana was visualized in a form of an interactive map:
  - Launch sites
  - Number of launches per size along with color indication of the landing outcome. Markers were clustered because the same launch site was used for several launches
  - The distances to proximities (railroad, roads, coastline) were calculated and displayed

Build a Dashboard with Plotly Dash

- Github URL: [capstone/spacex_dash_app.py at master · rhorvatgm/capstone (github.com)]
- An interactive dashboard application was created. The use is able to select the launch site and filter the playload mass range
- The results are accordingly updated and displayed in a
  - Piechart format (success rate) and
  - Scatter plot (playload mass vs landing outcome)

Predictive Analysis (Classification)

- First the data was normalized
- Four classification models were chosen
  - Logistic regression, support vector machine, decision tree and k nearest neighbours
- The dana was split into training and test dana
- Hyper parameters where chosen using Gridsearch and crossvalidation with 10 folds
- The accuracy was tested using the best hyperparameters on the test data

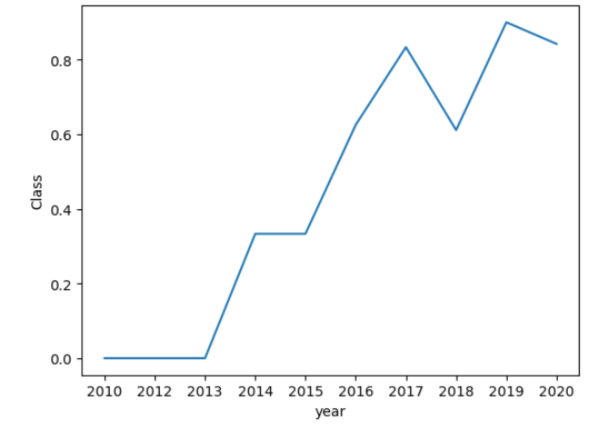


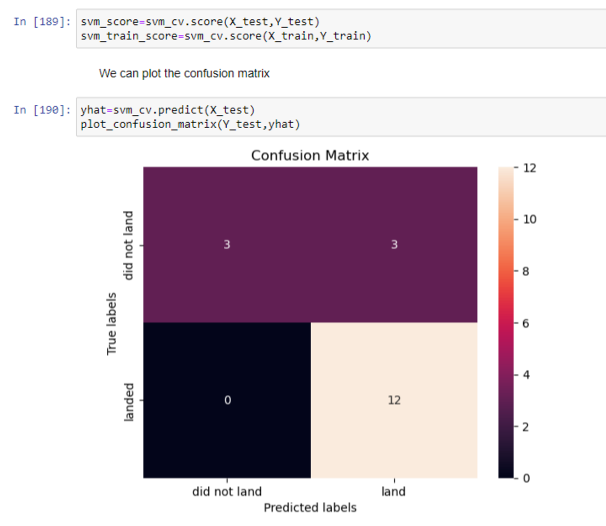
Results

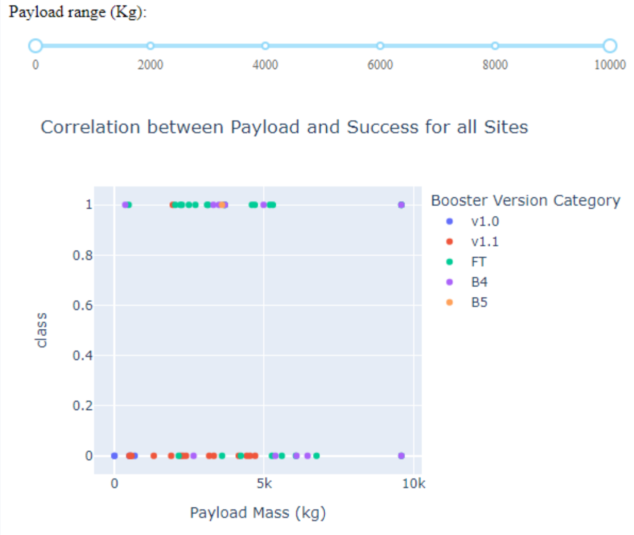
- Exploratory data analysis results
  - Successful landings improved over time

- Success rate variwes over playload mass

- We can predict the landing outcome with a Support vector machine and 83,3% accuracy





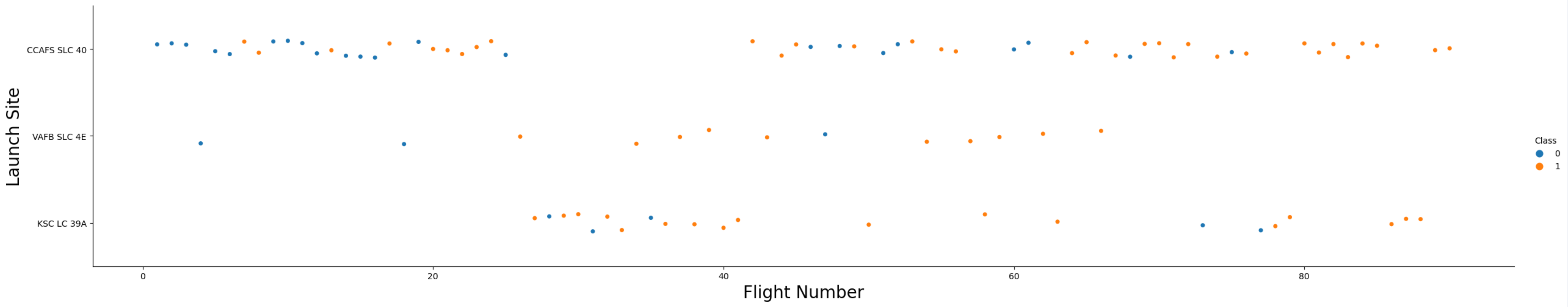


Section 2

Insights drawn from EDA

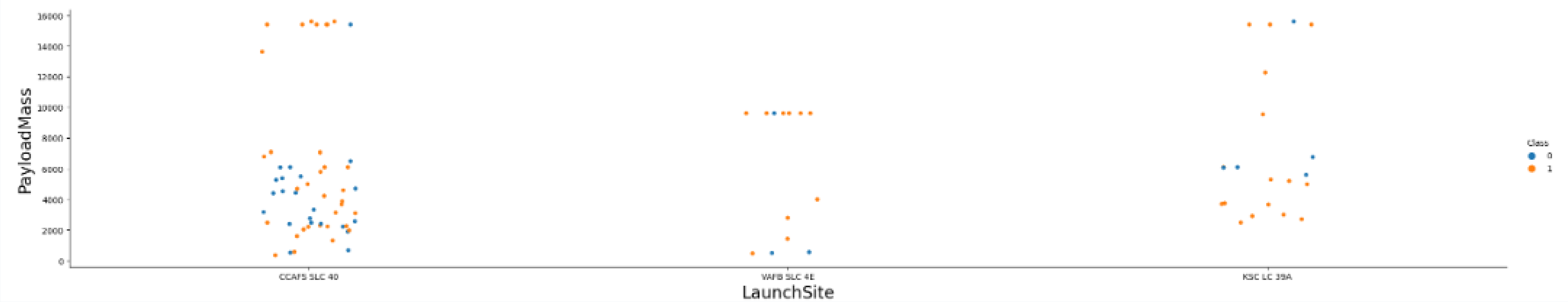
Flight Number vs. Launch Site

- For the first 25 flights mainly the launch site CCAFS SLC 40 was used
- The use of launch site KSC LC 39A was intensified between launch attempts 27 to 41
- Overall, the launch site CCAFS SLC 40 was the most used one

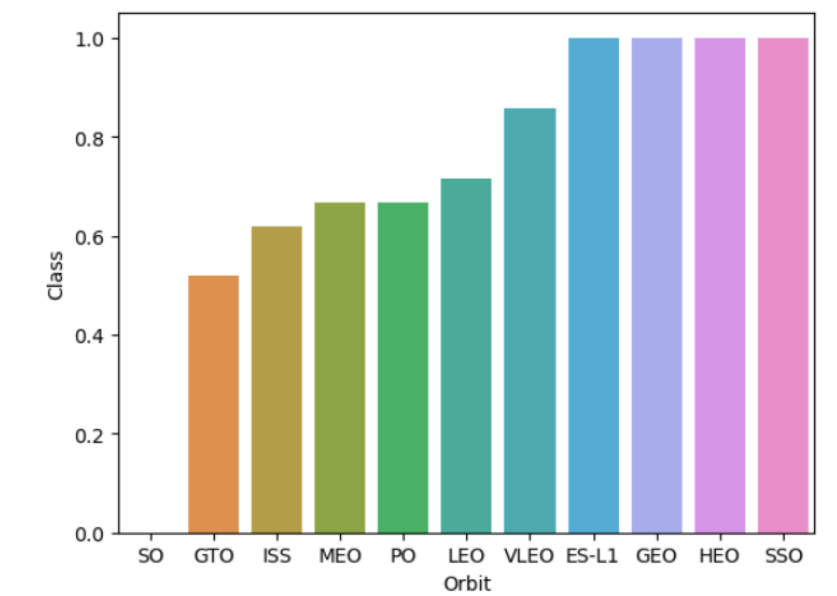


Payload vs. Launch Site

- Most of the low payload (<= 8k kg) launches where conducted from launch site CCAFS SLC 40
- Launch site VAFB SLC 4E was not used for high payload, bud intesnivelly used for payloads aroud 10k kg
- Most of the high payloads (>=14k kg) were launched from CCAFS SLC 40

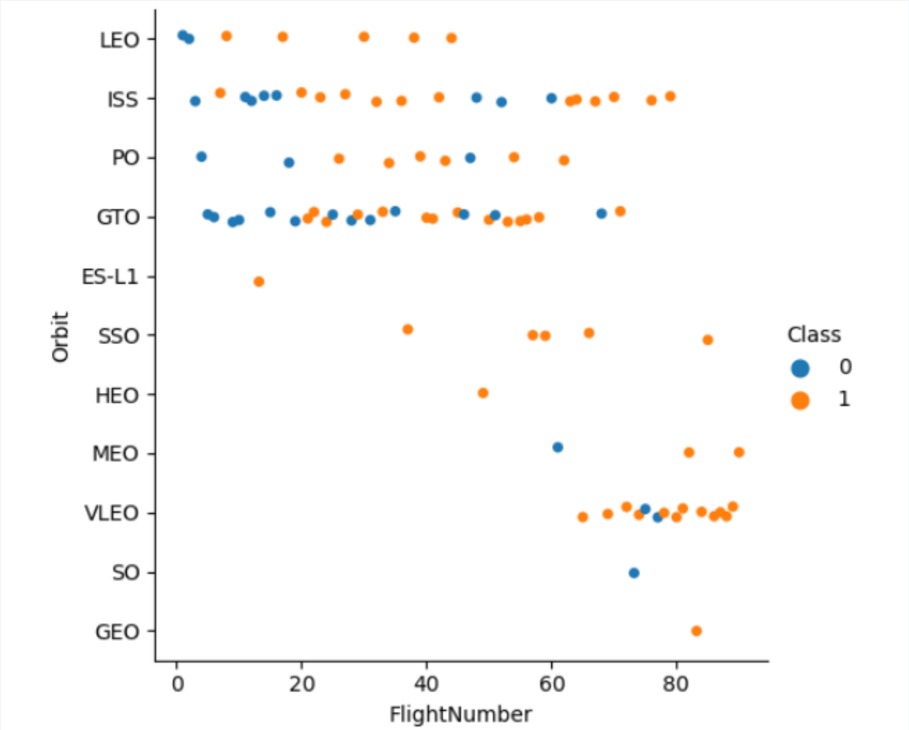


Success Rate vs. Orbit Type



- Success rate varies over different orbits
- There were no successfull landings after a launch attempt to orbit SO
- Successfull landing rates for orbits GTO, ISS, MEO, PO, LEO, VLEO stepwise rise from 50% to 85%
- The landings of launches to orbits ES-L1, GEO, HEO and SSO were 100% successfull

Flight Number vs. Orbit Type

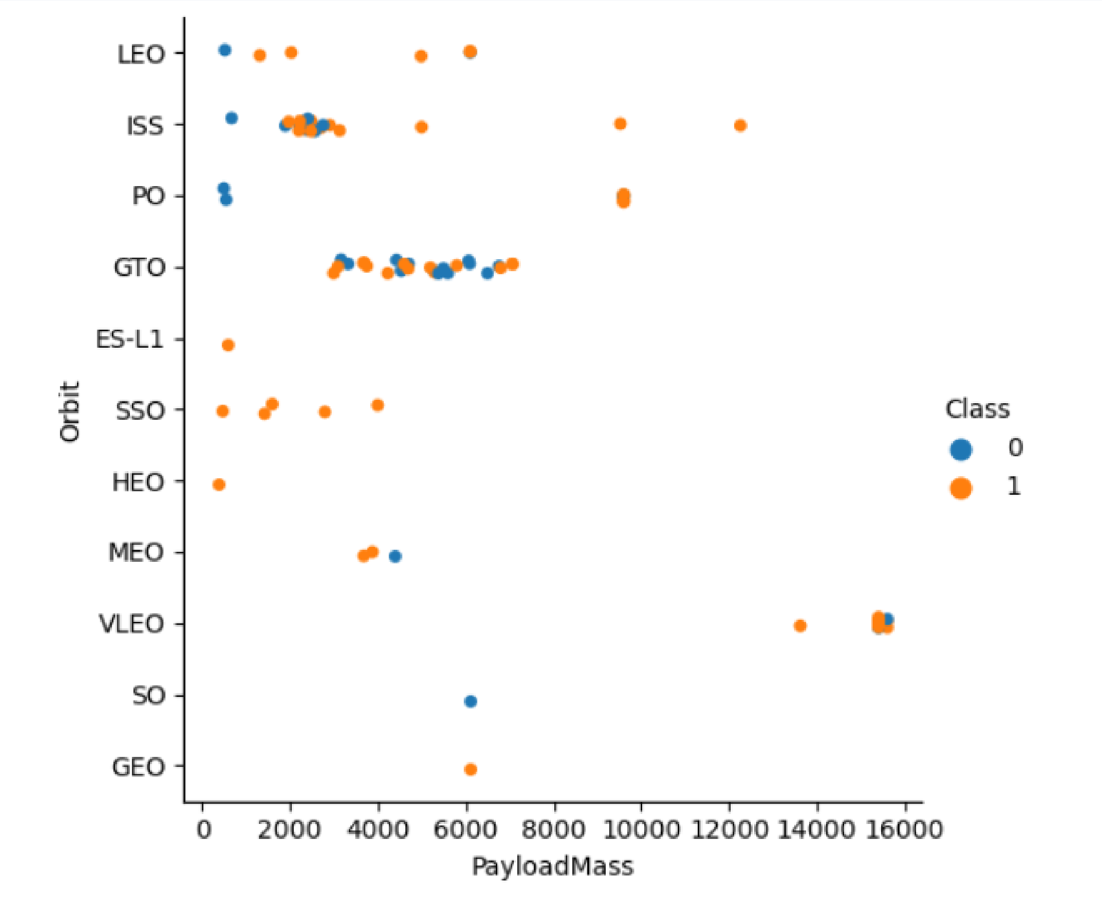


- The most targeted orbits in the first 70 flights were LEO, ISS, PO and GTO
- Later on VLEO was the most targeted orbit
- The success rate of launches targeting SSO is 100%
- Overall success rate improved over time

Payload vs. Orbit Type

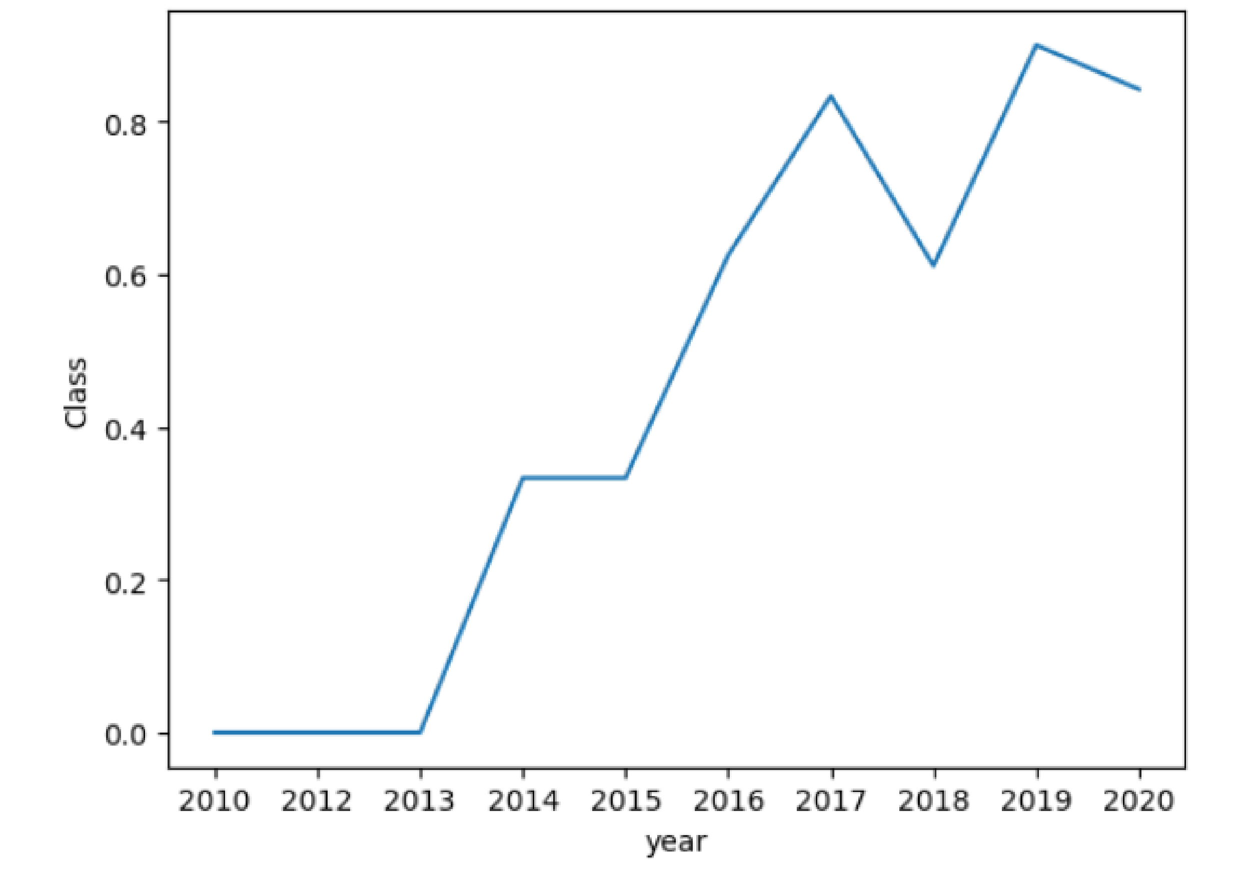
- Show a scatter point of payload vs. orbit type

- Show the screenshot of the scatter plot with explanations



Launch Success Yearly Trend

- The success rate generally rises over time
- In 2019-2020 it's over 80%



All Launch Site Names

- Four different launch site names were present in the datase
- Three of them (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A) in Florida, one (VAFB SLC-4E) in California

```
In [6]: %sql select distinct launch_site from spacex

 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-99
Done.
```

Out[6]:

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

Launch Site Names Begin with 'CCA'

- Here are 5 records where launch sites begin with `CCA’
- All 5 attempts were conducted to the low Earth orbit

In [9]:

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

```
 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[9]:

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

Total Payload Mass

- The total payload carried by boosters from NASA was 45596 kg

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

```
In [22]: %sql select sum(payload_mass__kg_) as total_payload from spacex where customer='NASA (CRS)'

 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
Out[22]:
```

total_payload
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2928 kg

```
In [27]: %sql select avg(payload_mass__kg_) from spacex where booster_version='F9 v1.1'

 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.

Out[27]:   1
         2928
```

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was dec 22 2015

```
In [34]: %sql select min(DATE) from spacex where landing__outcome='Success (ground pad)'

 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.

Out[34]:
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [38]:
```
%sql select distinct booster_version from spacex where landing__outcome='Success (drone ship)' and payload_mass__kg_>4000 and payload_mass__kg_<6000
```

```
 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[38]:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes was 100
- There was only one unsuccessfull mission outcome

```
In [44]: %sql select sum(case when mission_outcome like '%Success%' then 1 else 0 end) as success,sum(case when mission_outcome like '%Success%' then 0 else 1
```

```
 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[44]:

success	failure
100	1

Boosters Carried Maximum Payload

- Here is a list of the boosters which have carried the maximum payload mass

In [48]:
```
%sql select distinct booster_version from spacex where payload_mass__kg_=(select max(payload_mass__kg_) from spacex )
```

```
 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[48]:

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

- Here is a list of failed landing_outcomes in drone ship, their booster version, and launch site names in year 2015

```
In [13]: %sql select booster_version, launch_site from spacex where landing__outcome ='Failure (drone ship)' and year(DATE)=2015

 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.

Out[13]:
```

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

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

- Here are the landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order of incidence

```
In [14]: %sql select landing__outcome,count(*) from spacex where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome order by 2 desc

 * ibm_db_sa://gxp69073:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Out[14]:

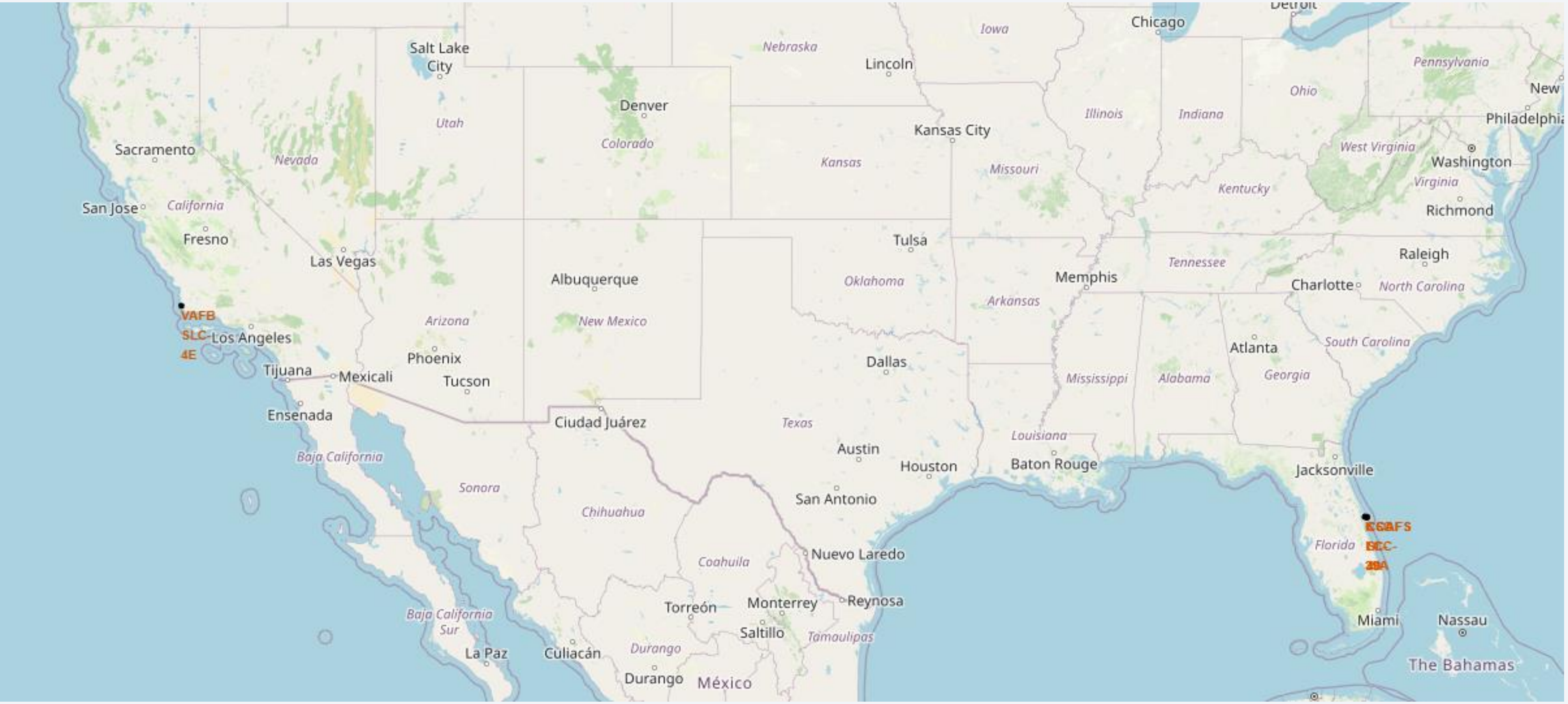
landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

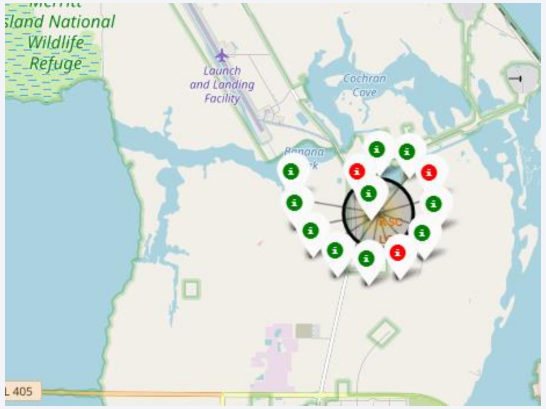
SpaceX launch sites

- There are 4 launch sites, one in California and three in Florida



Outcomes maped by launch sites

California VAFB SLC-4E: 4 sucesses out of 10



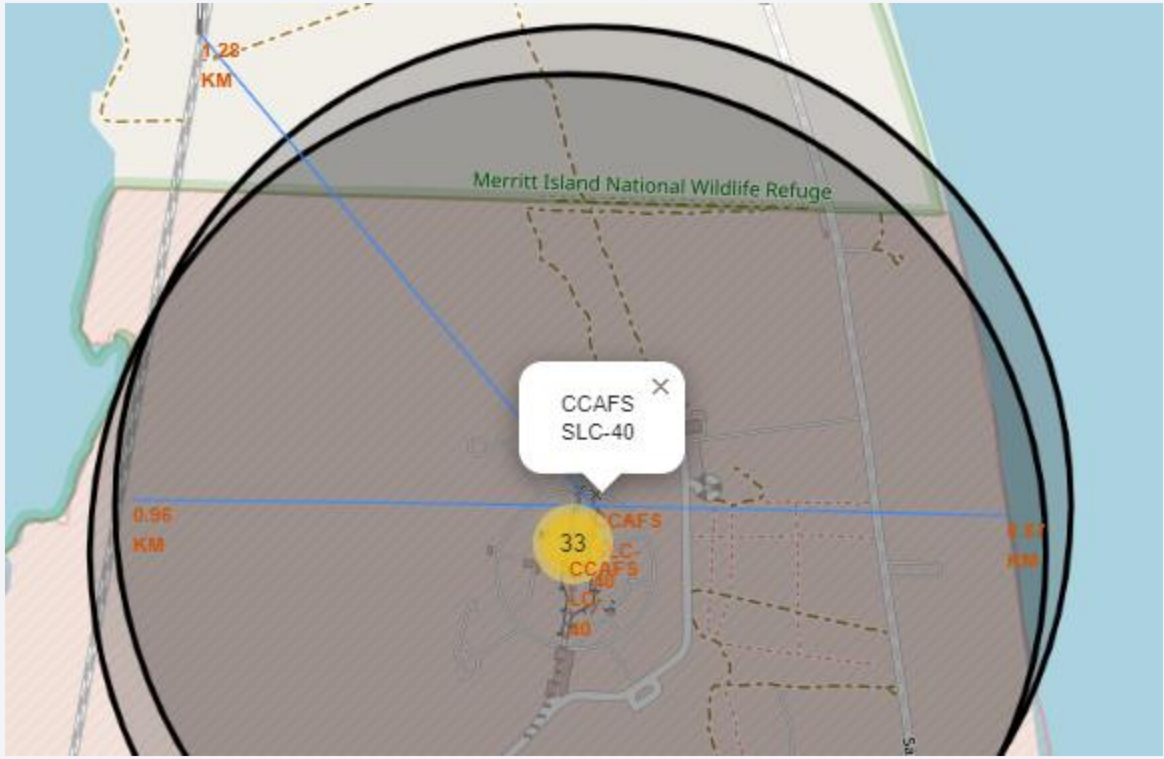
Florida KSC LC-39A: 11 sucesses out of 13 – best success rate

Florida KSC LC-39A: 3 sucesses out of 7

Florida CCAFS LC-40: 7 sucesses out of 26, most used launch site

Launch site CCAFS SLC-40 air distance to proximities

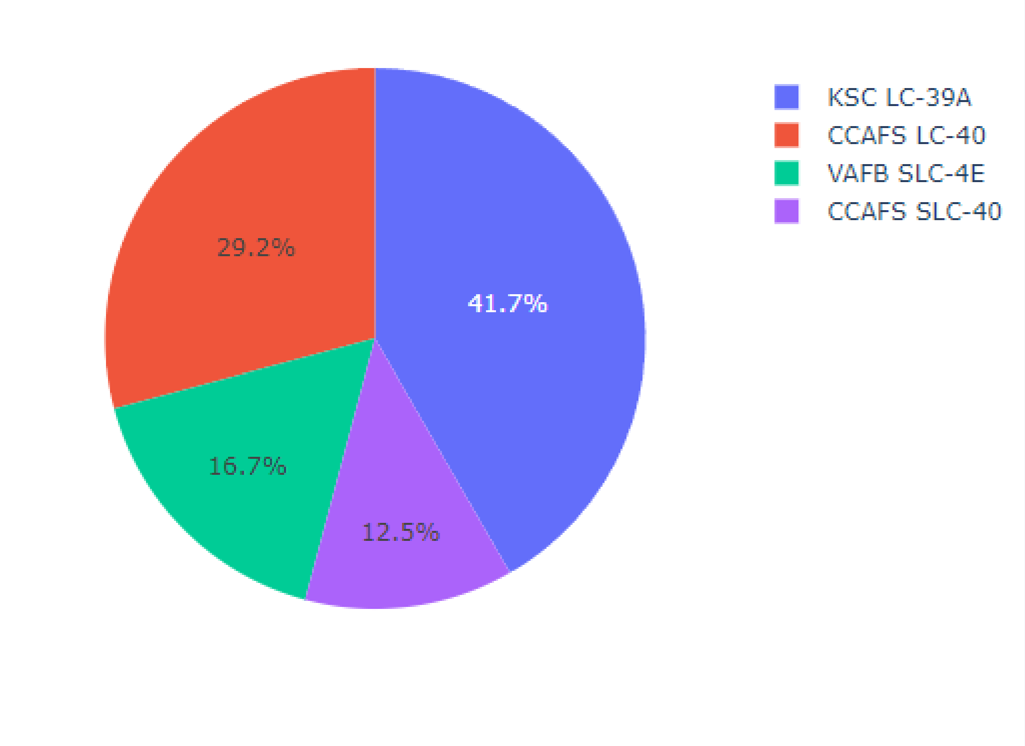
- The air distance to the coastline is aprox. 0.87 km
- The air distance to the NASA railroad is aprox. 1.28 km
- The air distance to TITAN III Road is aprox. 0.96 km



Section 4

Build a Dashboard with Plotly Dash

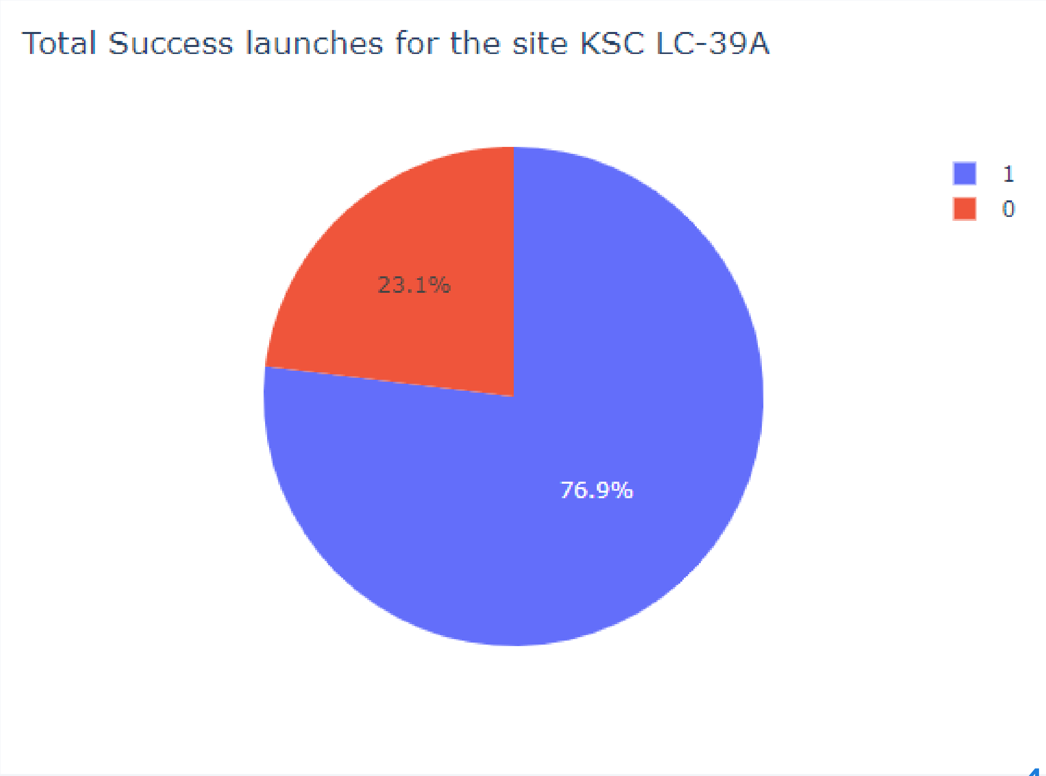
Successfull launches per site



- Most of the successfull launches were conducted from KSC LC-39A and CCAFS LC-40.
- Together these sites account for more then 70% of the successfull attemts.

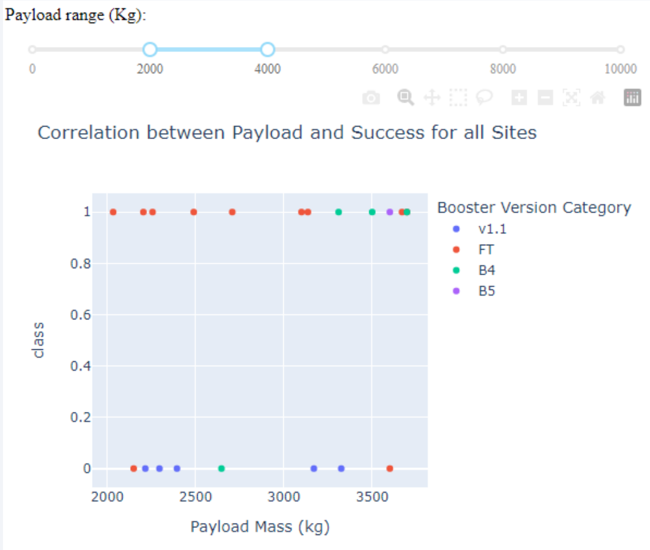
Launch site KSC LC-39A Success rate

- 76,9% of the launches from site KSC LC-39A have successfully landed



Correlation between Payload and Success

- The success rate varies over different payload ranges
- Most of the attempts with payload between 5k and 8k kg have not finished with sucessfull landings
- Payloads between 2k and 4k kg have above average success rate

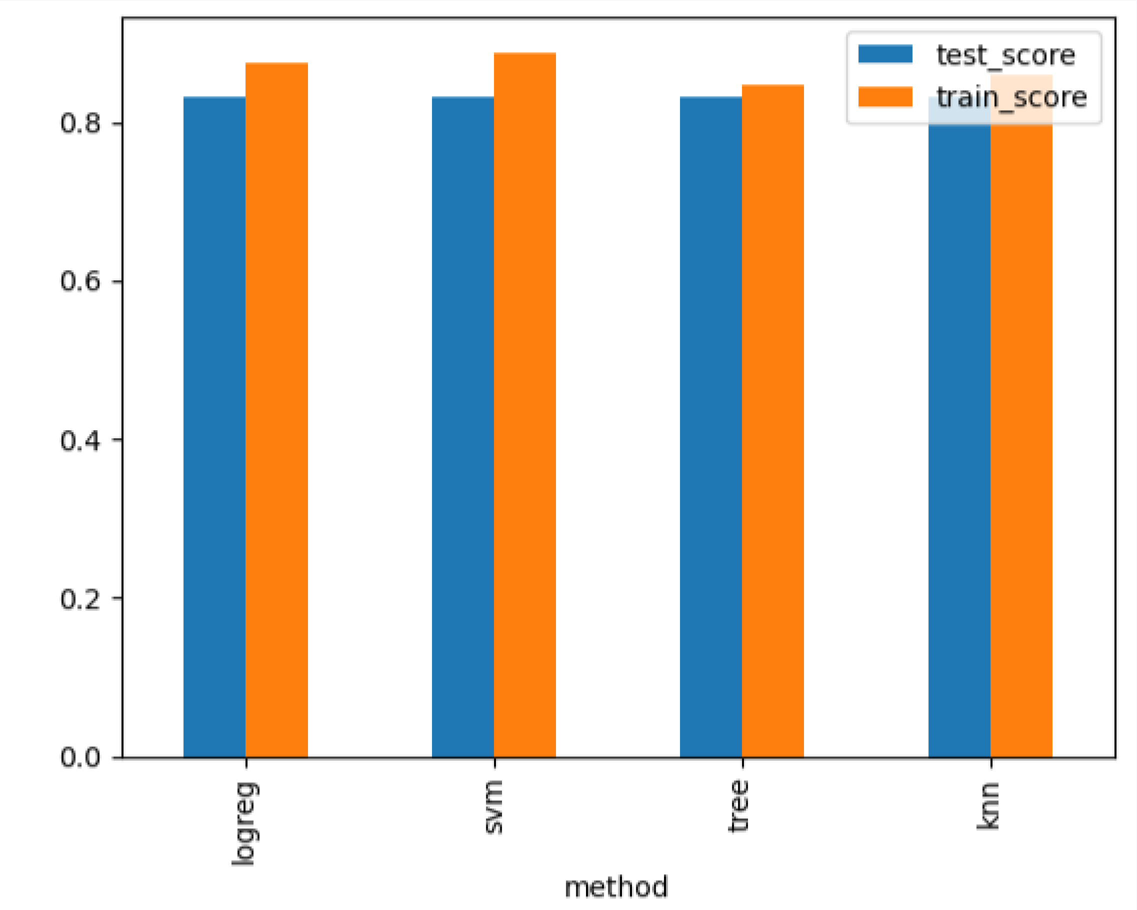


Section 5

Predictive Analysis (Classification)

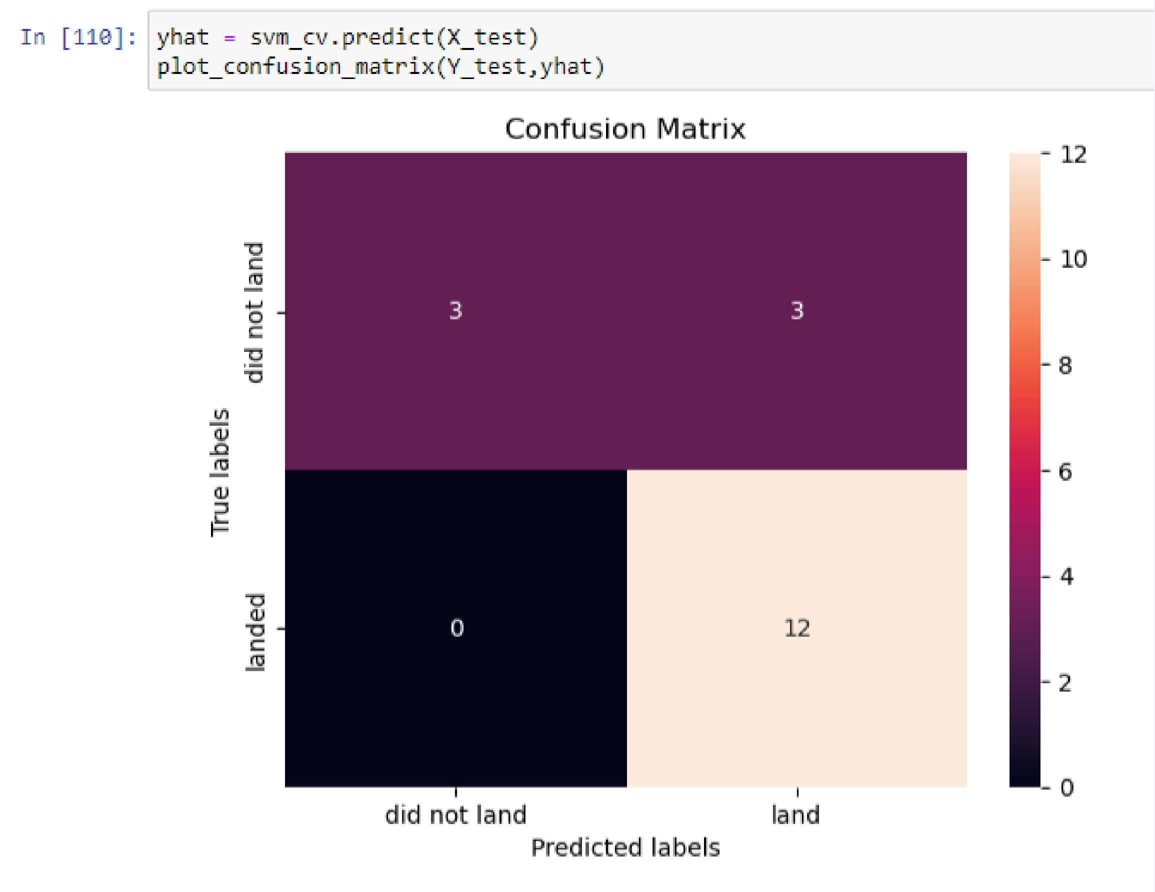
Classification Accuracy

- The support vector machine model had the highest accurary on the training data, but
- all models had the same accuracy (83.3%) on the test data



Confusion Matrix

- The overall accuracy of the model on the training set is 83,3%
-



Conclusions

- All models preformed equally on the test dana, even more
- All confusion matrixes were equal
- The accuracy was 83,3%
- There were no false negative predictions, but there
- Were false positive predictions

Appendix

- Here's the code for the SVM model

```
In [186]: parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'),
                         'C': np.logspace(-3, 3, 5),
                         'gamma':np.logspace(-3, 3, 5)}
          svm = SVC()
```

```
In [187]: svm_cv=GridSearchCV(svm,parameters,cv=10)
          svm_cv.fit(X_train,Y_train)
```

```
Out[187]:  ▸ GridSearchCV
           ▸ estimator: SVC
                ▸ SVC
```

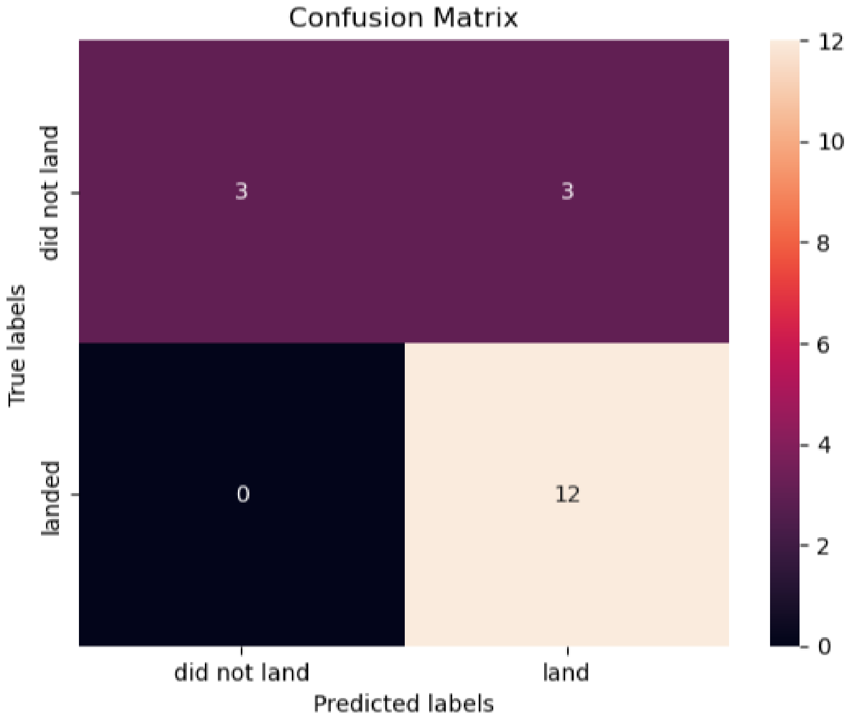
```
In [188]: print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
          print("accuracy :",svm_cv.best_score_)

            tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
            accuracy : 0.8482142857142856
```

Appendix (continued)

```
In [189]: svm_score=svm_cv.score(X_test,Y_test)
          svm_train_score=svm_cv.score(X_train,Y_train)
```

We can plot the confusion matrix

```
In [190]: yhat=svm_cv.predict(X_test)
          plot_confusion_matrix(Y_test,yhat)
```

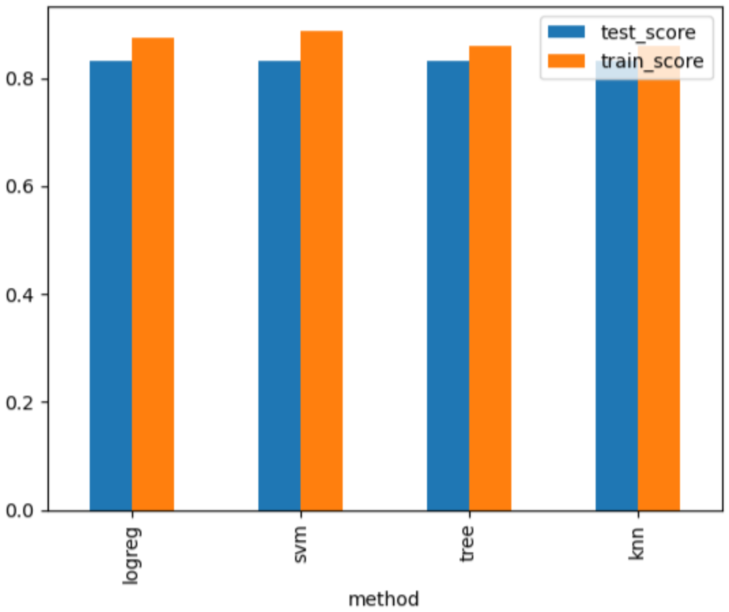


Appendix (continued)

- Model comparison

```
In [208]: scores={'method':['logreg','svm','tree','knn'],'test_score':[logreg_score,svm_score, tree_score,knn_score],'train_score':[logreg_train_score,svm_train_score, tree_train_score,knn_train_score]};
rez=pd.DataFrame(scores);
rez.set_index("method",inplace=True);
rez.plot.bar()
```

```
Out[208]: <AxesSubplot:xlabel='method'>
```



Thank you!