Winning Space Race with Data Science

SANGATI DAVEEDU

December,07, 2025

Outline

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

Executive Summary

In this Project we are trying to predict the First Stage Landing of Falcon 9 using machine learning and determine the best model for this particular case. If we can determine the whether the first stage will land, then we can calculate the cost of a launch. The methodologies are as follows: Business Understanding, Analytic understanding, Data collection, Data Wrangling, EDA, Visualization, Data Modeling and Model Evaluation. The machine learning algorithms that are used for modeling are : Logistic regression, SVM, Decision Tree, and KNN.

The result is pretty interesting, where all models have similar test score in around 83.4%. But, we conclude that SVM is the best model since the accuracy score in the training is not too far off compared to the test score. This means the model performs pretty well. On the other hand, Decision tree has around 89% training score which means that we have some overfitting going on in there.

Introduction

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. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

So we got ourselves a problem: based on the historical data, can we predict the first stage landing so that we can determine the launch cost?

Section 1

Methodology

Methodology

Executive Summary

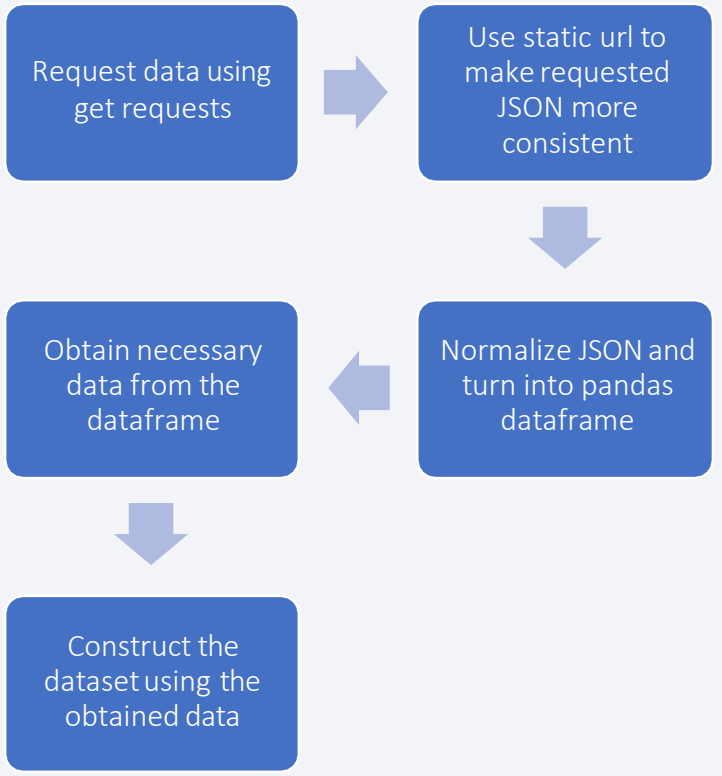
- Data collection methodology:
  - Data was collected using SPACEX REST API calls and Web Scraping
- Perform data wrangling
  - Replacing missing values with mean values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using hyperparameter tuning to tune models and then evaluate them to find the best classification model

Data Collection

- The Data Collection process is consisted of 2 methods: SpaceX API and Scraping.
- Using SpaceX API: starts with creating pipeline consisting of auxiliary functions to construct the dataset. Then we continue with requesting rocket launch data from SpaceX API, use static response to normalize it and then turn it into Pandas DataFrame. Using the previously created functions, we construct the dataset based on the json DataFrame.
- Using Web Scraping: almost the same, but we use beautiful soup to parse the html and extract all the information available in the wikipedia. Then we construct the dataset.

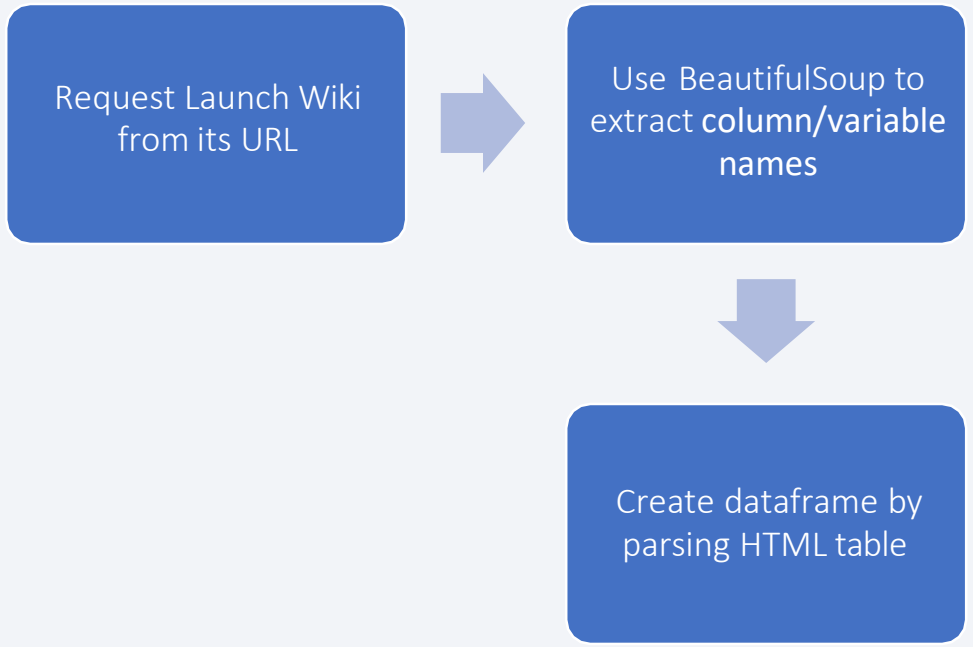
Data Collection - SpaceX API

- The data collection using SpaceX API process can be seen on the following flowchart.
- You can see the full notebook on this github link.



Data Collection - Scraping

- The data collection using web scraping process can be seen on the following flowchart.
- You can see the full notebook on this github link.



Data Wrangling

- We identified that there are some missing values in PayloadMass and LandingPad columns.
- What we did was we replaced the null values in PayloadMass with its mean value.
- We left the LandingPad column as it was.

EDA with Data Visualization

- For the data visualization, the charts that were plotted are as follows:

1. Flight Number vs Launch site : to understand whether launch site and their respective flight number has a correlation with landing success.
2. Payload vs Launch site : to know the correlation between payload mass from each launch site with landing success.
3. Success rate/orbit : to find out whether some orbit are more/less likely to guarantee a landing success.
4. Etc.

- Full notebook on github link here.

EDA with SQL

- Some performed SQL query on this project:

1. Display unique launch sites name
2. Display 5 launch sites beginning with 'CCA'
3. Show total payload mass launched by NASA (CRS)
4. Show average payload mass carried by specific booster
5. Rank the count of landing outcomes
6. Etc

- Full notebook on github link here.

Build an Interactive Map with Folium

- Basically what we did:

1. Mark all launch sites on a map
2. Mark the success/failed launches for each site on the map to know which launch site has higher success rate
3. Calculate the distances between a launch site to its proximities and draw the line to understand what typically launch sites are built near to or far from.

- Github link here.

Predictive Analysis (Classification)

- We use 4 ML algorithms in this project: Logistic Regression, SVM, Decision Tree and KNN
- First we split the dataset into training and testing dataset with test size 0.2
- Then we do hyperparameter tuning using Grid Search to determine the best parameter for each model and the fit them to X_train and y_train.
- Best model is judged based on its accuracy score.
- Full notebook on github here.

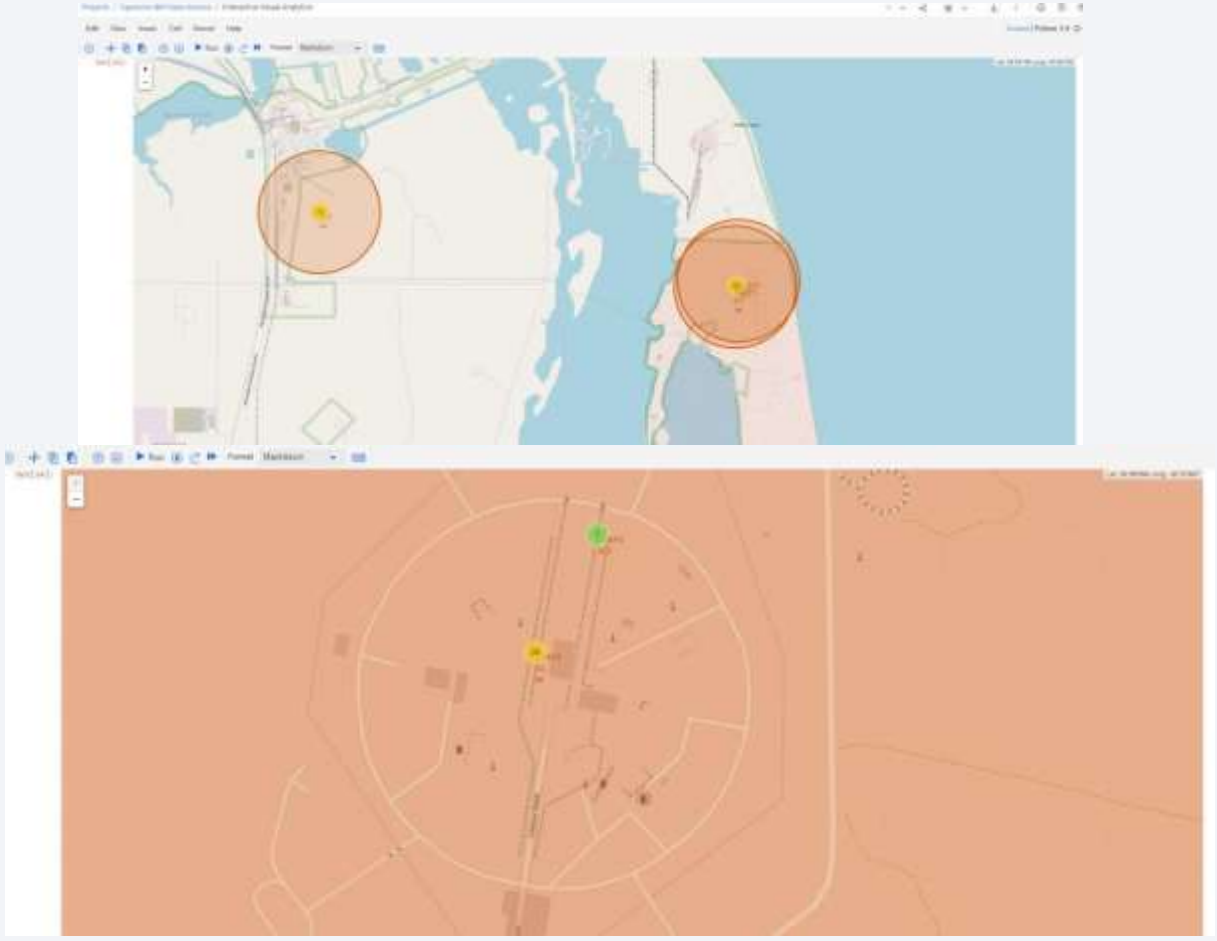
Results

- Exploratory data analysis results

1. CCAFS SLC 40 has the most launch number with 55 launches.
2. The first stage landing success rate is around 66.67%
3. VAFB-SLC launchsite there are no rockets launched for heavypayload mass
4. ES L1, SSO, GEO, and HEO orbits have the highest success rate.
5. In LEO orbit, success are related to number of flights
6. In some orbit the heavier the payload the more likely the landing to succeed.

Results

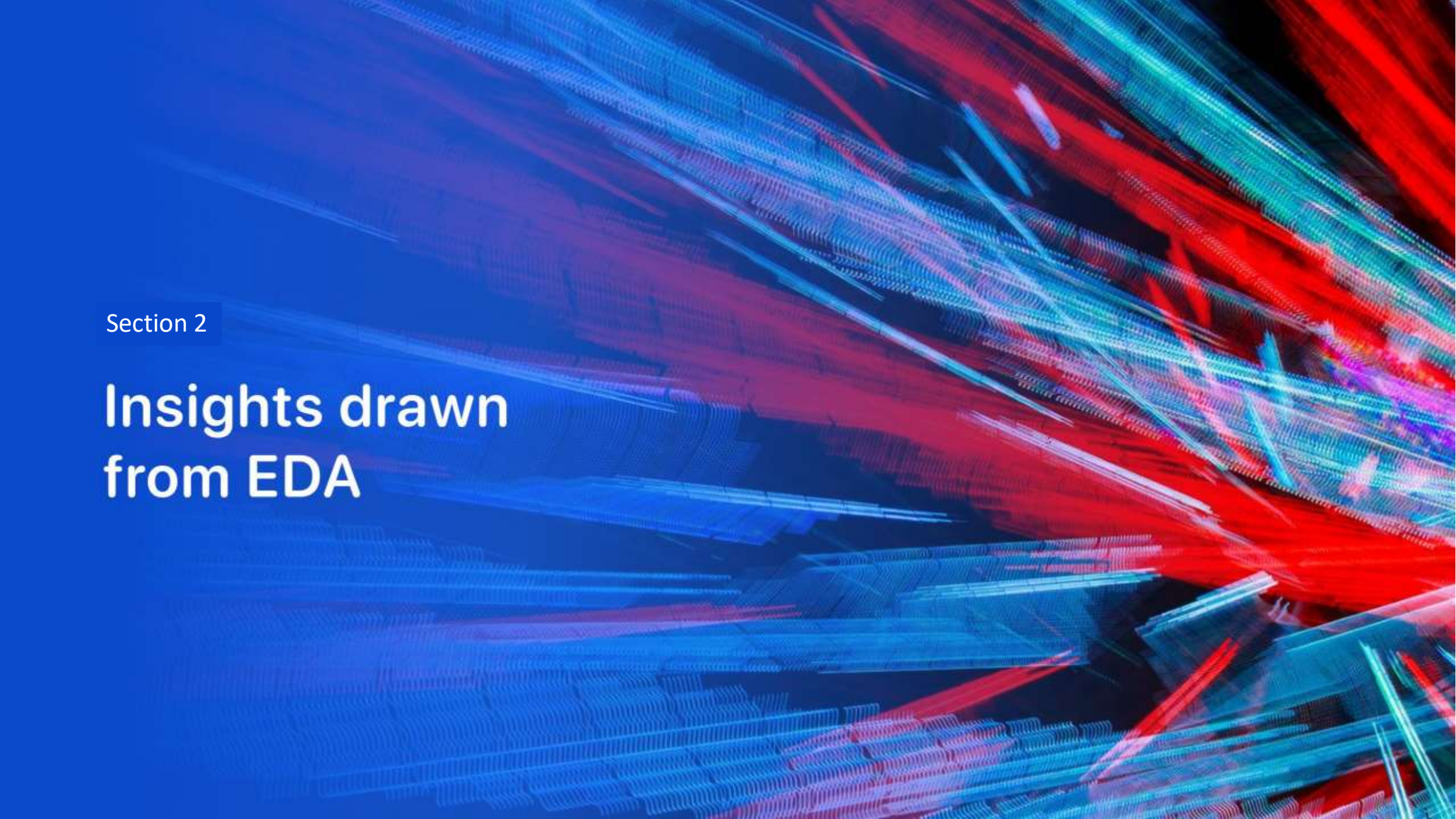
- Some interactive analytics demo in screenshots



Results

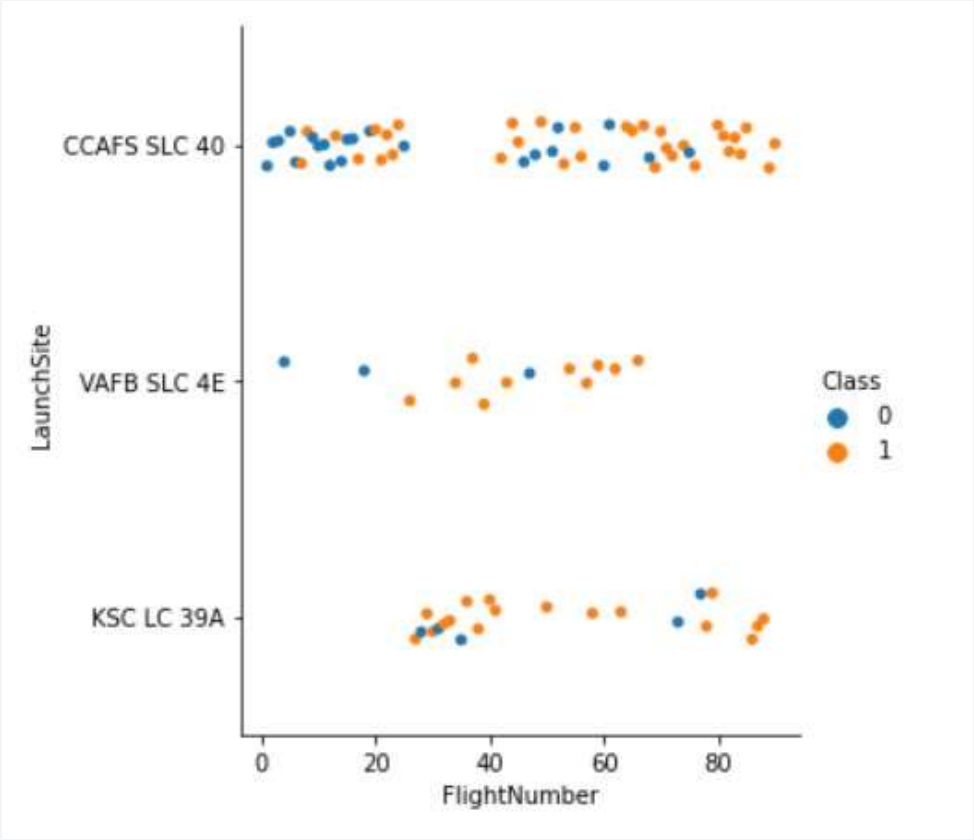
- Predictive analysis results

```
logistic regression
accuracy : 0.8464285714285713
test score : 0.8333333333333334
svm
accuracy : 0.8482142857142856
test score : 0.8333333333333334
decision tree
accuracy : 0.8910714285714286
test score : 0.8333333333333334
KNN
accuracy : 0.8482142857142858
test score : 0.8333333333333334
```


Section 2
Insights drawn from EDA

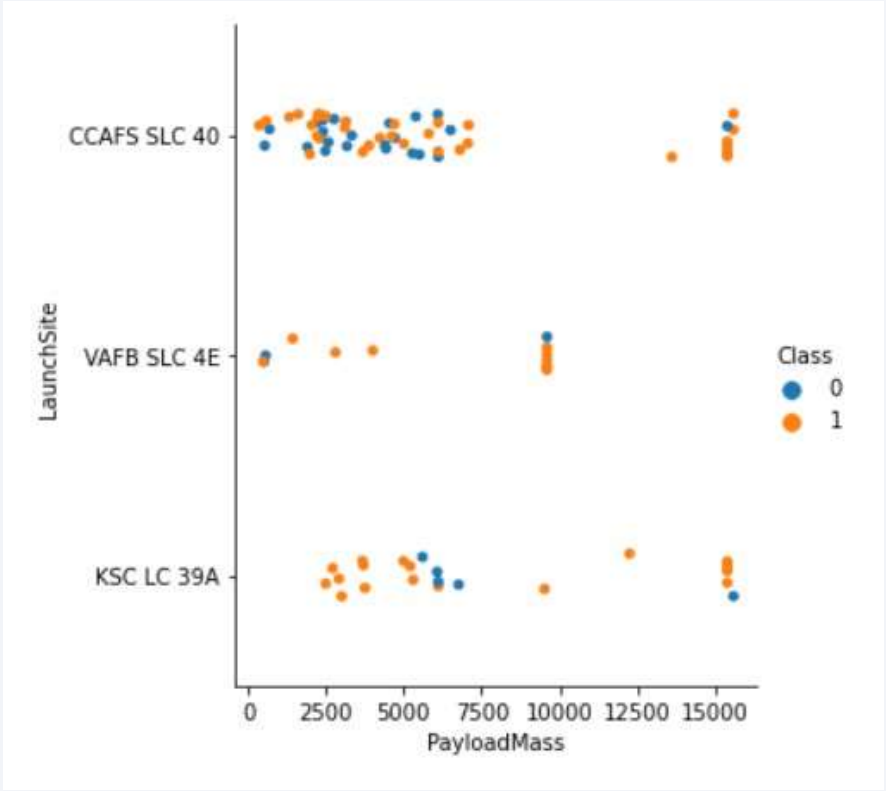
Flight Number vs. Launch Site

- As we can see, it's pretty difficult to get any insights from the graph.
- However, for site VAFB SLC 4E it appears that the higher the flight number the success rate is also higher.
- As for the other 2 launch site it's too difficult to tell.



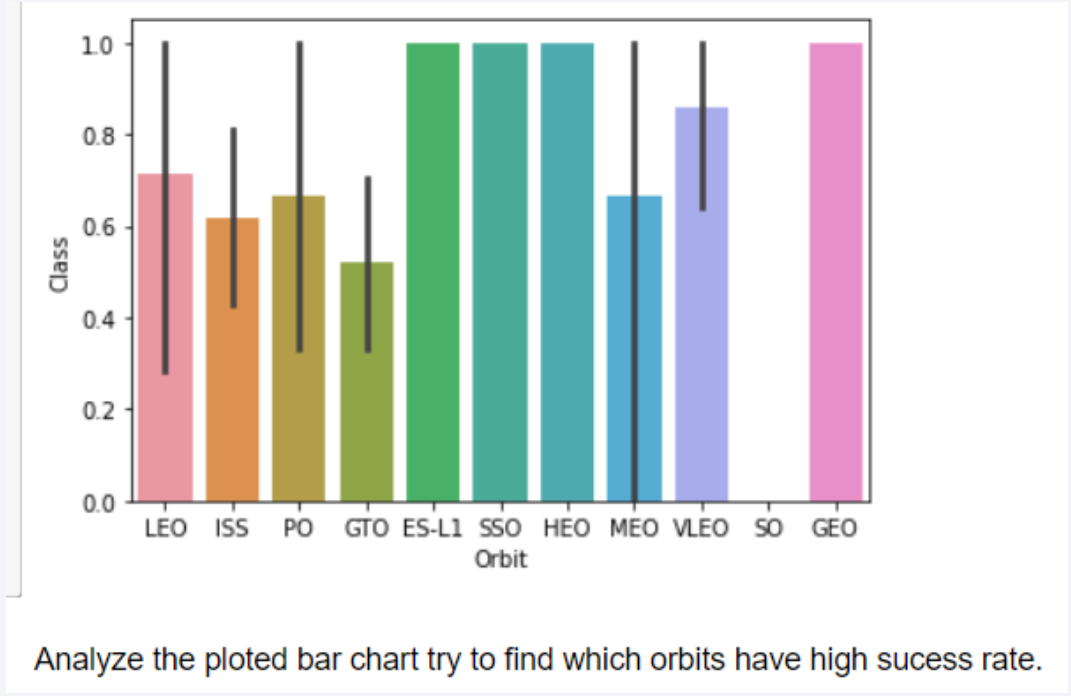
Payload vs. Launch Site

- We can see that launch site VAFB SLC 4E doesn't launch booster with payload >10000
- Again, it’s pretty difficult to tell the correlation of launch site and payload mass with the landing success.



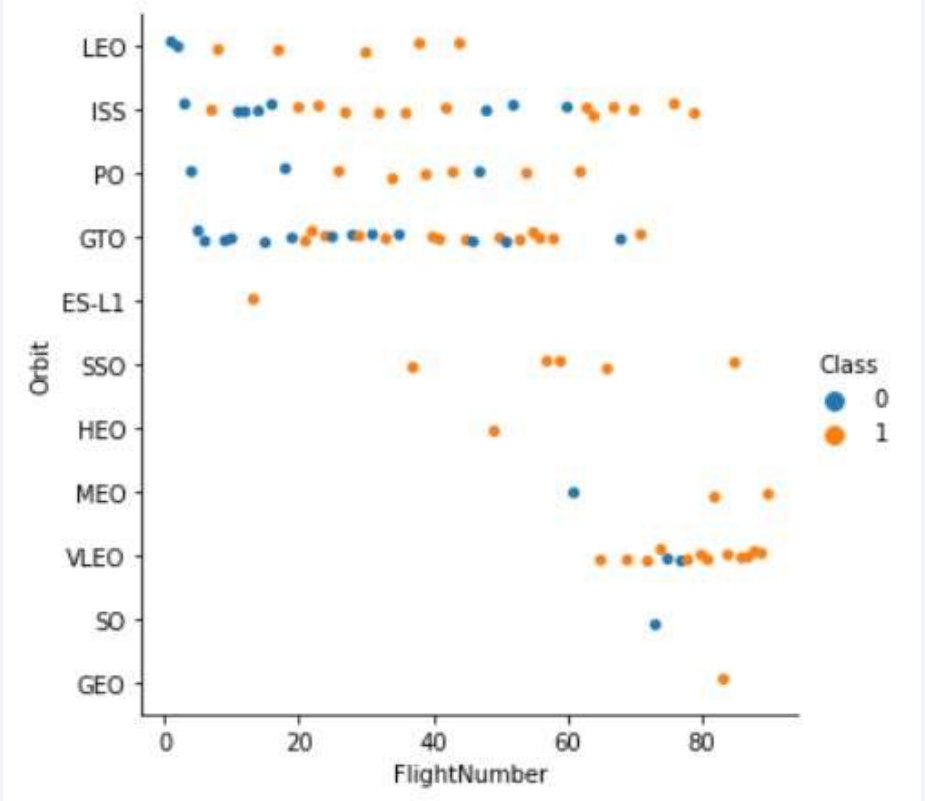
Success Rate vs. Orbit Type

- We can see that 4 orbits ES-L1, SSO, HEO, and GEO have the highest success rate compared to the other orbit type.



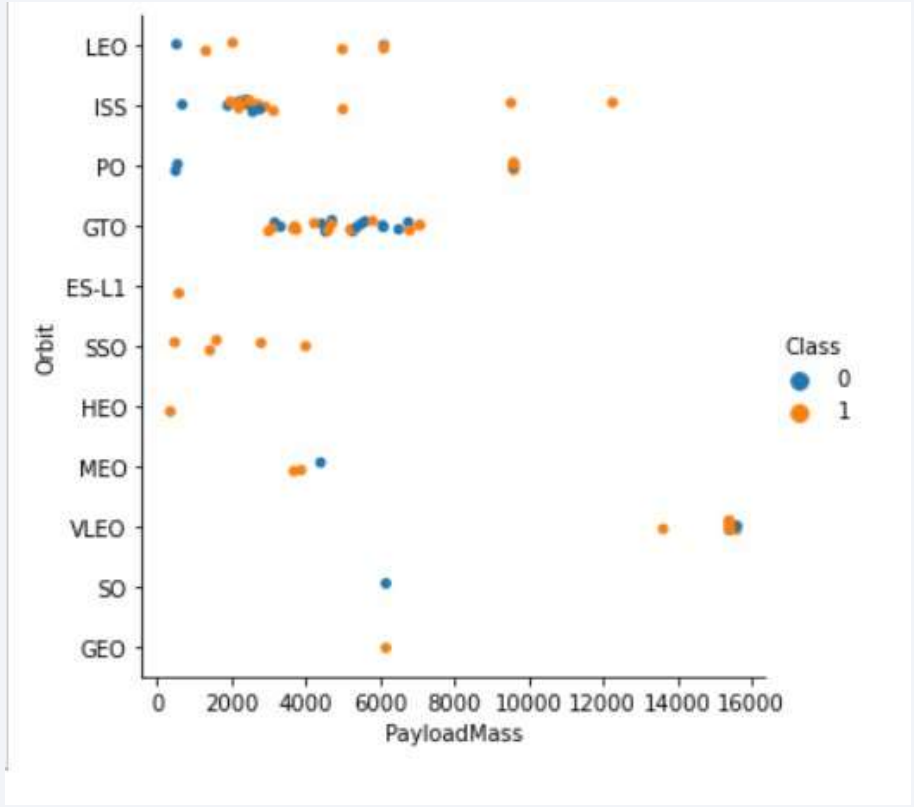
Flight Number vs. Orbit Type

- For LEO orbit, more flight number most likely means more success landing.
- No relationship between the flight number and success in the GTO orbit.



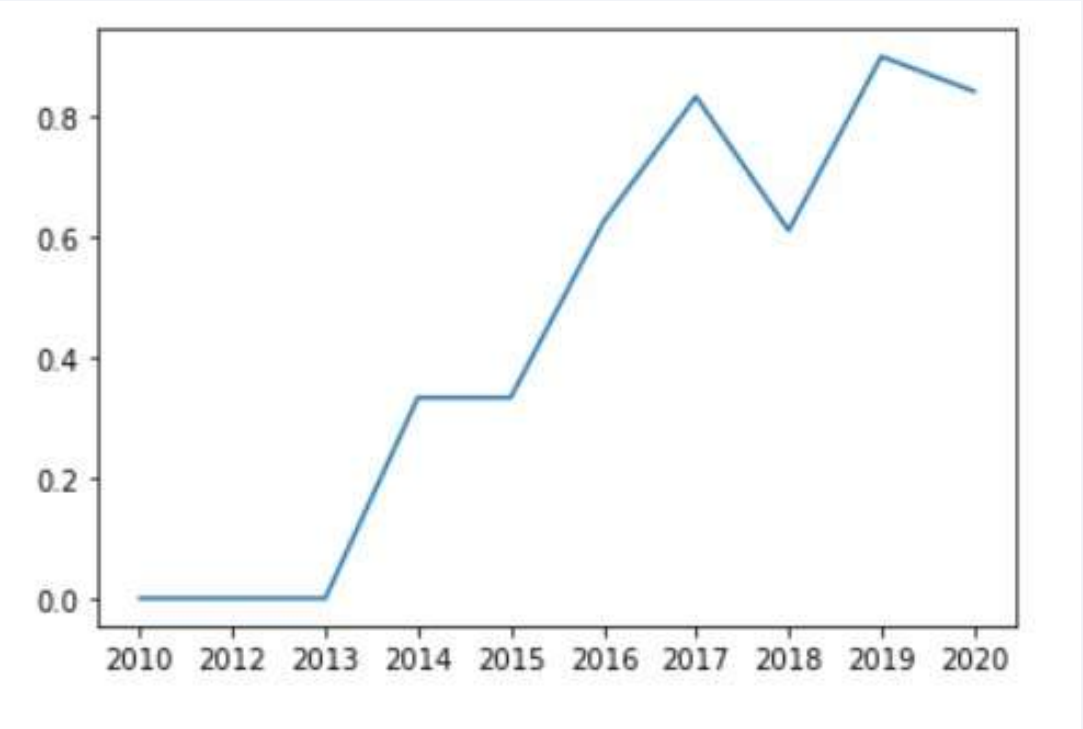
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- From the graph on the right hand side, we can see that from year 2013 to 2020 the success trend are relatively improving.



All Launch Site Names

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

- There are 5 unique launch sites as we can see on the graph on the left hand side.

Launch Site Names Begin with 'CCA'

- Below are 5 records with launch site names begin with 'CCA'

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

```
%%sql
SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET
WHERE customer='NASA (CRS)'
```

```
 * ibm_db_sa://cyy66994:***@fbd88901-ebdb-4a4f-a32e
Done.
```

1
45596

- Total Payload Mass carried by bossters launched by NASA (CRS) is 45596.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET
WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://cyy66994:***@fbd88901-ebdb-4a4f-a3
Done.
```

1
2534

- Average Payload Mass carried by F9 v1.1 booster is 2534.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) FROM SPACEXDATASET
WHERE landing__outcome ='Success (ground pad)'
```

```
 * ibm_db_sa://cyy66994:***@fbd88901-ebdb-4a4f-a
Done.
```

1
2015-12-22

- First successful landing outcome in ground pad was achieved in December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Below are the successful drone ship landing with Payload between 4000 and 6000. As we can see, there are 5 launches and all booster version name starts with 'F9 FT'.

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

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

- The mission outcome success rate is actually really good with only 1 failure and technically 100 success.

Boosters Carried Maximum Payload

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

- These are the list of booster with maximum payload. As we can see, we have 12 booster which name all starts with 'F9 B5'.

2015 Launch Records

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

- We only have 2 failed landing in drone ship in 2015 and both are launched from the same launch site: CCAFS LC-40.

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

landing__outcome	number
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

- On the left is landing outcome rank between 2010-06-04 and 2017-03-20. As we can see, no attempt is the most outcome with 10 times occurrence. While precluded (drone ship) is the least with only 1 occurrence.

Section 3

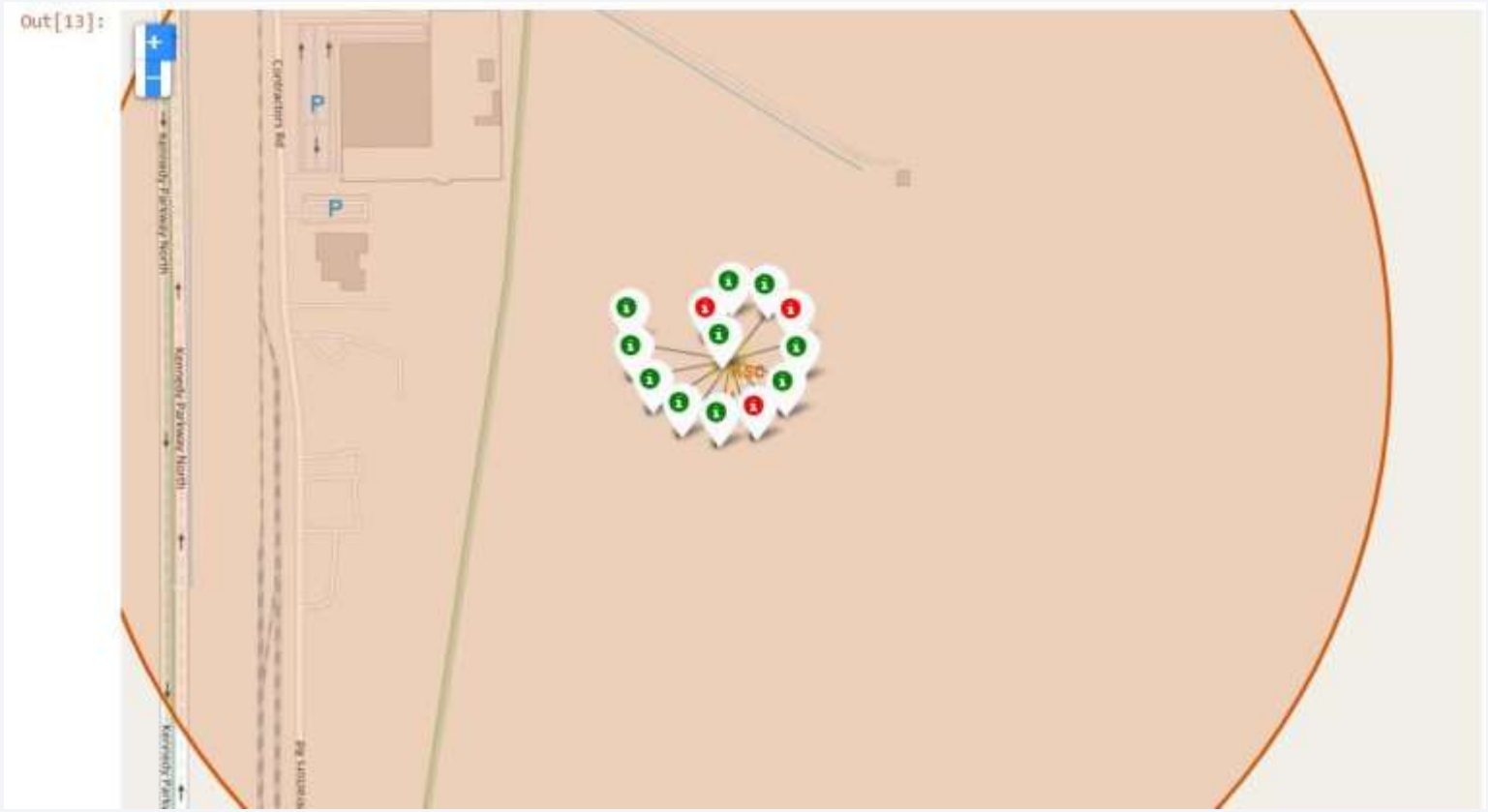
Launch Sites Proximities Analysis

Launch Sites Map

Out[9]:

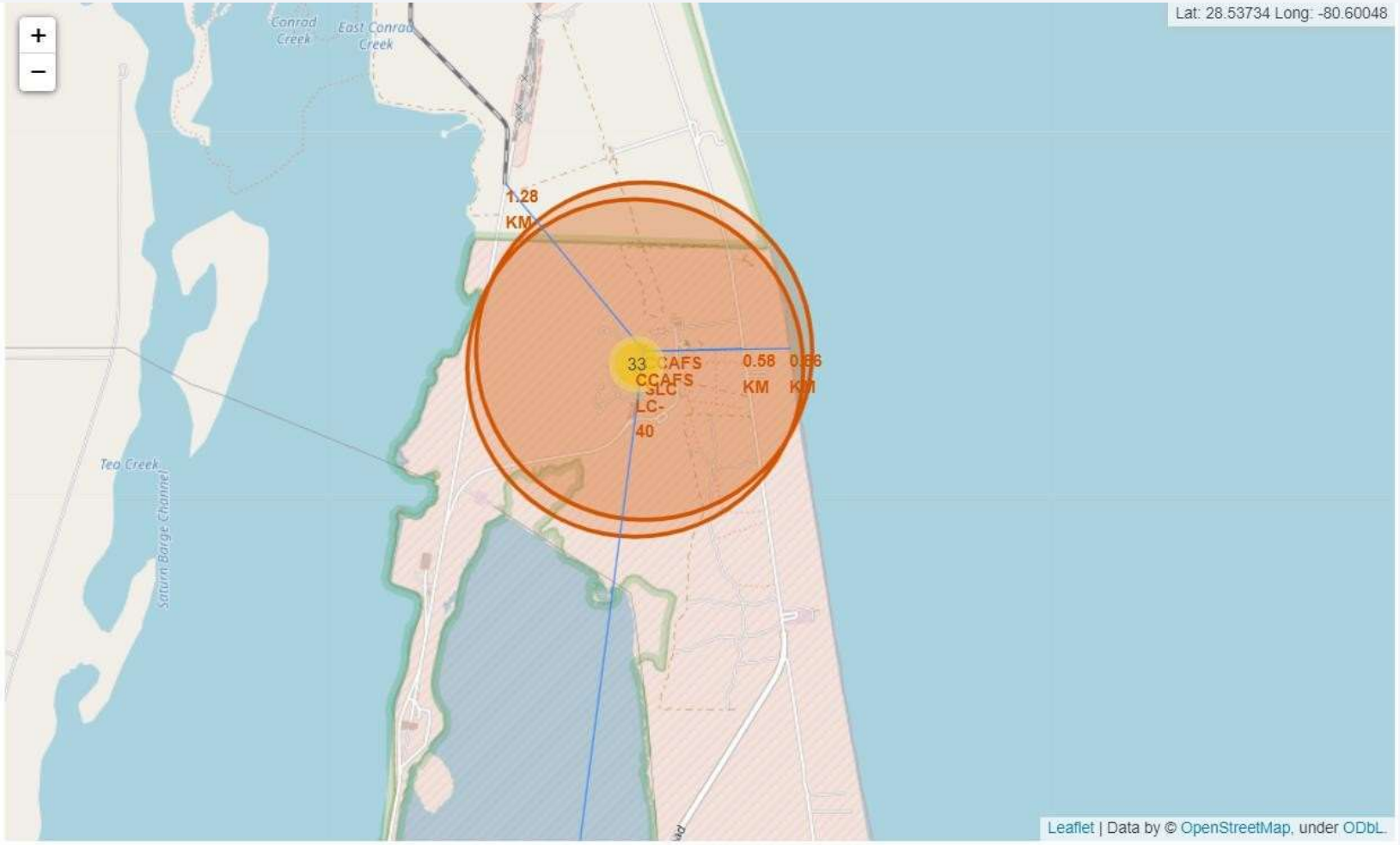
- As we can see, we have 5 launch sites which all located near the coastline.
- 2 of them located on the West coast and the other 3 on the East Coast.

Success/Failed Launch Map



- The map on the left is one of the launch site which is KSC LC - 39A.
- We can see there are red and green markers which represent failed launches and successful launches respectively.
- We can infer that KSC LC -39A has a decent success landing rate, having only failed 3 times in 13 launches.
- We did this visualization analysis for the other launch sites as well.

<Folium Map Screenshot 3>



- The map on the left shows one of the launch sites with its distance to its proximities.
- We can see that the launch site is pretty close to highway (0.58 km), railway (1,28 km), and coastline (0.86 km). This are most likely due to convenience factors in transporting people/resources.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

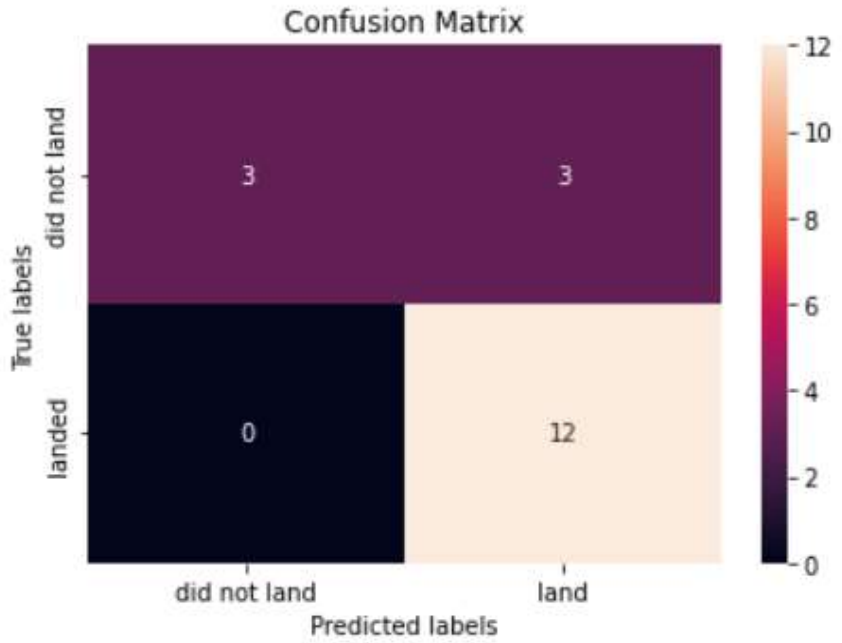
```
logistic regression
accuracy : 0.8464285714285713
test score : 0.8333333333333334
svm
accuracy : 0.8482142857142856
test score : 0.8333333333333334
decision tree
accuracy : 0.8910714285714286
test score : 0.8333333333333334
KNN
accuracy : 0.8482142857142858
test score : 0.8333333333333334
```

- On the left is the accuracy score and test score on each models we used in this classification
- As we can see, the model with the highest accuracy score on training model is decision tree (89%), but the test score is significantly lower (83.34%) which is the same as the other model.
- That means, in decision tree model that we use we have some degree of overfitting.
- Therefore, **the best performing model** is the model with the least difference in accuracy and test score: SVM or KNN.
- However, you can see from the picture KNN edges SVM a little in accuracy score so we will pick KNN as our best model.

Confusion Matrix

We can plot the confusion matrix

```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(y_test,yhat)
```



- The confusion matrix of our KNN model is shown on the left.
- Top left is the True Negative (TN) value, while the top right is the False Positive (FP) value.
- Bottom left is False Negative (FN) value, while bottom right is the True Positive (TP) value.
- Accuracy is calculated using the following formula:
- Accuracy = (TP+TN)/(TP+TN+FP+FN)
- Which in this case, equals to = (12+3)/(12+3+3+0) = **0.83333**

Conclusions

- Decision Tree is the model with the highest accuracy but overfits when tested in the testing dataset.
- KNN is the best performing model for this case.
- In some orbit the heavier the payload the more likely the landing to succeed.
- The landing success rate is currently around 66.67%
- Launch sites are typically close to highway, railway, and coastal line.

Thank you!