IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SMHZad
04.10.2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context
    - In this project we want to evaluate the viability of the new company Space Y to compete with Space X, the most successful company of the commercial space age, making space travel affordable.
- Problems you want to find answers
    - How some variables such as payload mass, launch site, and orbits affect the success of the first stage landing
    - Does the success rate increase over the year?
    - Where is the best place to make launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data, Dealing with missing values, One Hot Encoding for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to find the best result

Data Collection

Data collected from

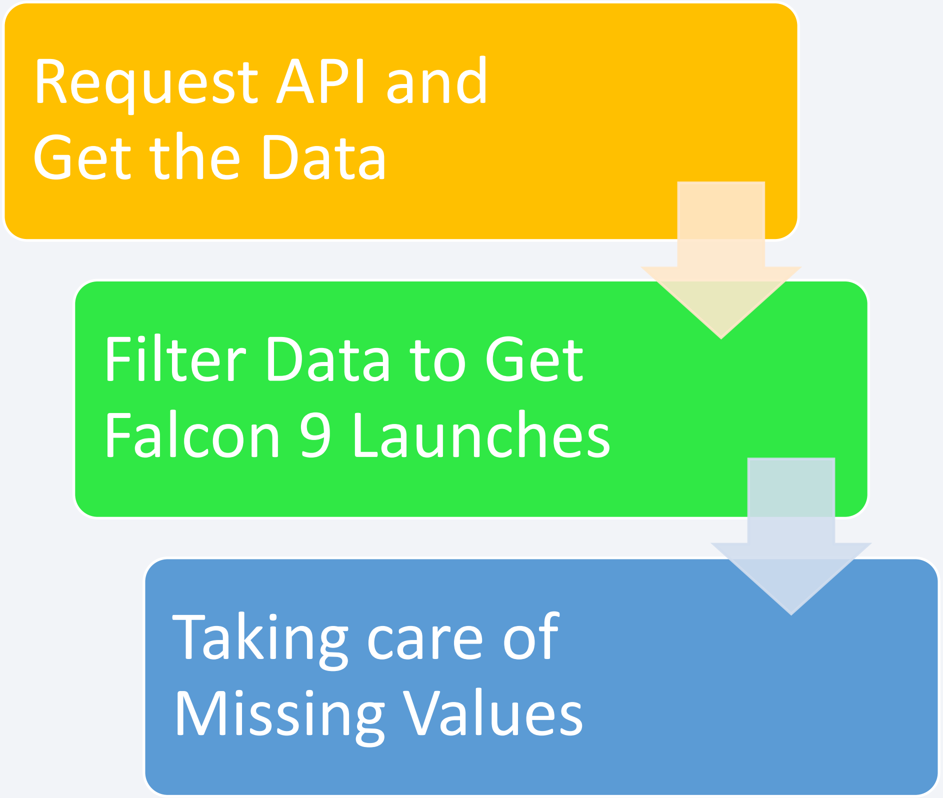
- Space X REST API
  - https://api.spacexdata.com/v4/rockets/

- Wikipedia
  - https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- SpaceX offers a public API. The process is as follows:

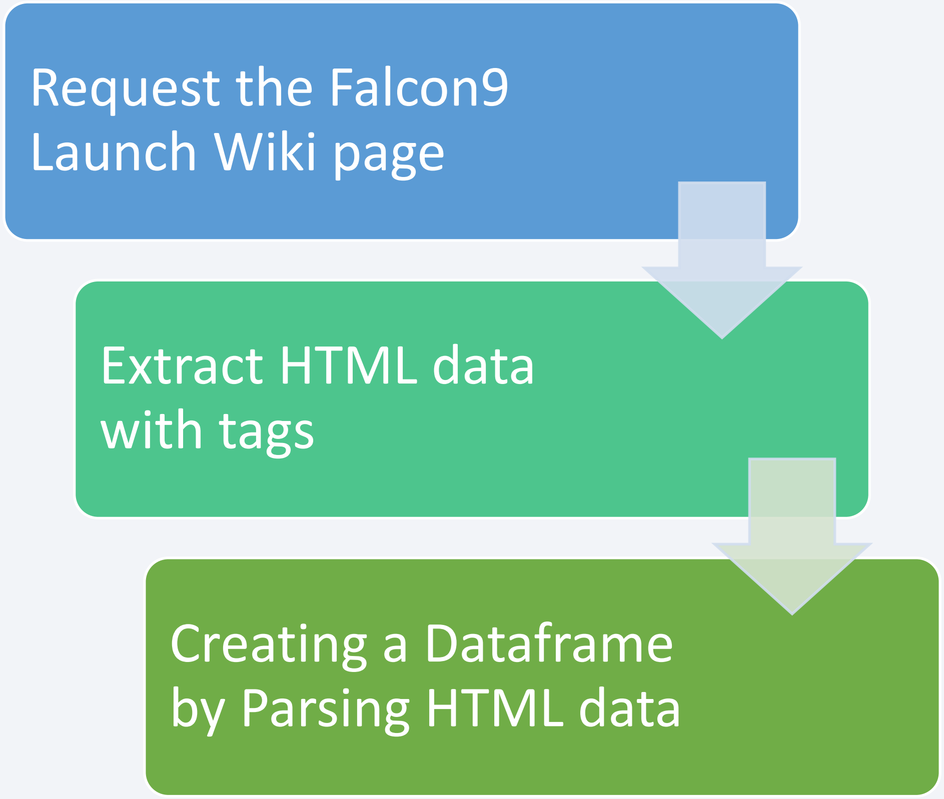
- Source code:
  - https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/01%20-%20Data%20Collection%20API.ipynb



Data Collection - Scraping

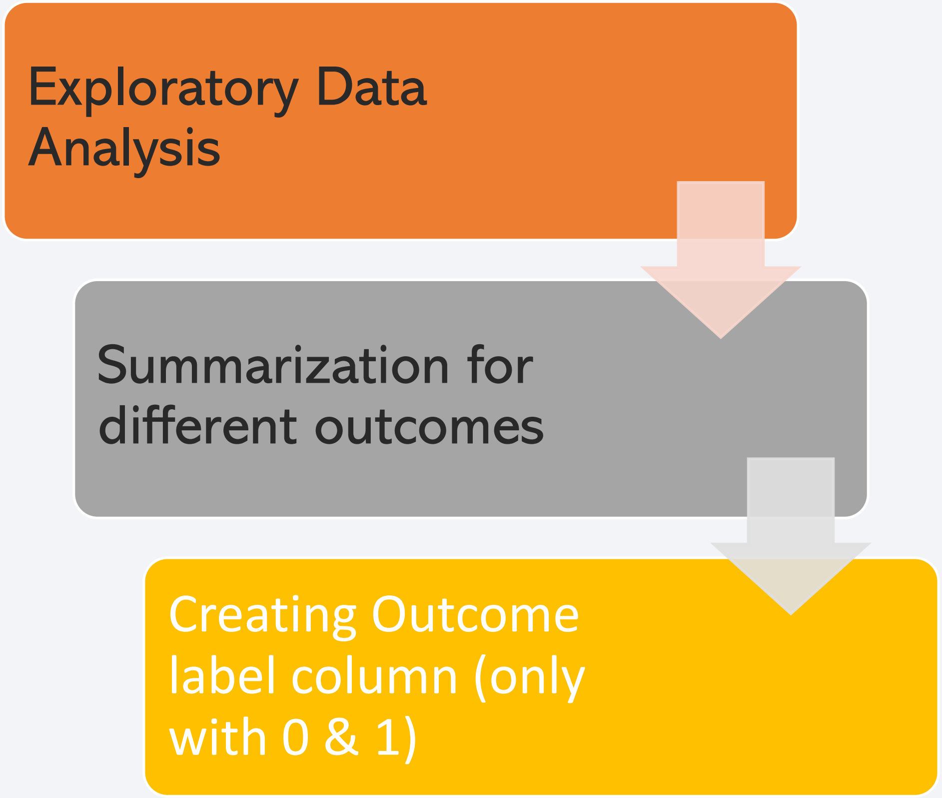
- Wikipedia offers a list of Falcons. The process is as follows:

- Source code:
  - https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/02%20-%20Data%20Collection%20with%20Web%20Scraping.ipynb



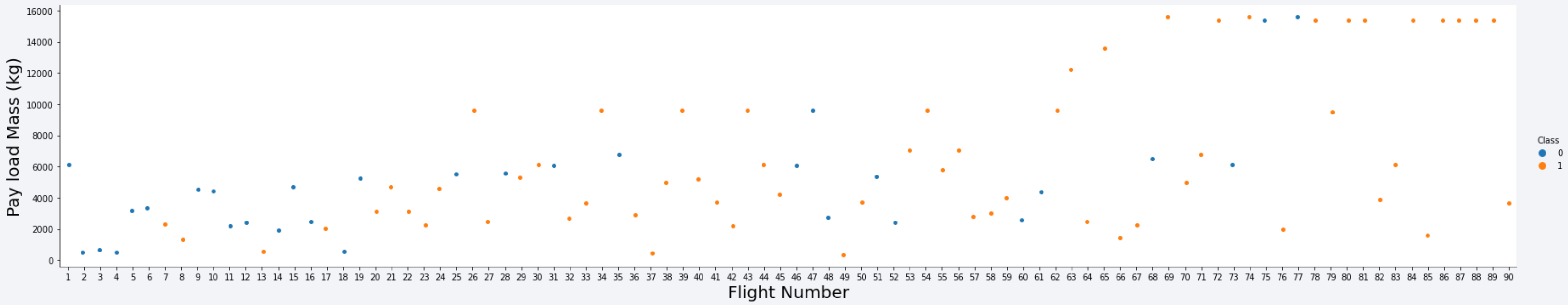
Data Wrangling

- Exploratory Data Analysis (EDA) was performed to learn more about dataset.
- Next, data was summarized for launches per site, occurrences of each orbit and occurrences of mission outcome per orbit.
- Finally, Outcome column is created for landing outcome label.
- Source code:
    - https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/03%20-%20Data%20Wrangling.ipynb



EDA with Data Visualization

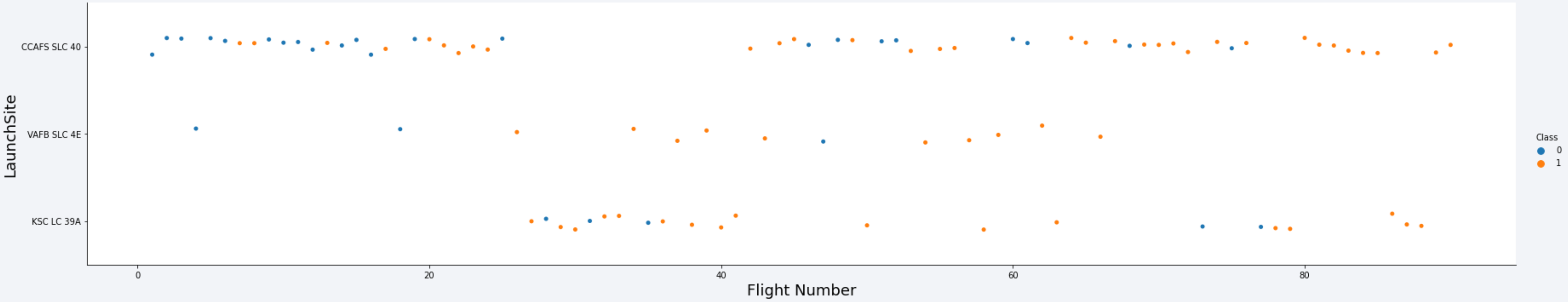
- To visualize data, scatterplots and bar plots were created between pair of features:
  - Payload Mass per Flight Number



- Source: https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/05%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with Data Visualization

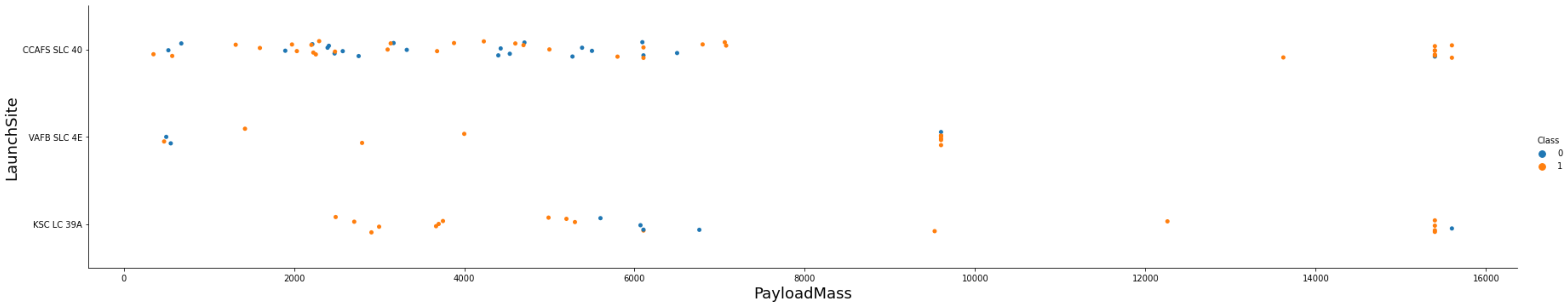
- To visualize data, scatterplots and bar plots were created between pair of features:
  - Launch Site vs. Flight Number



- Source: https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/05%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with Data Visualization

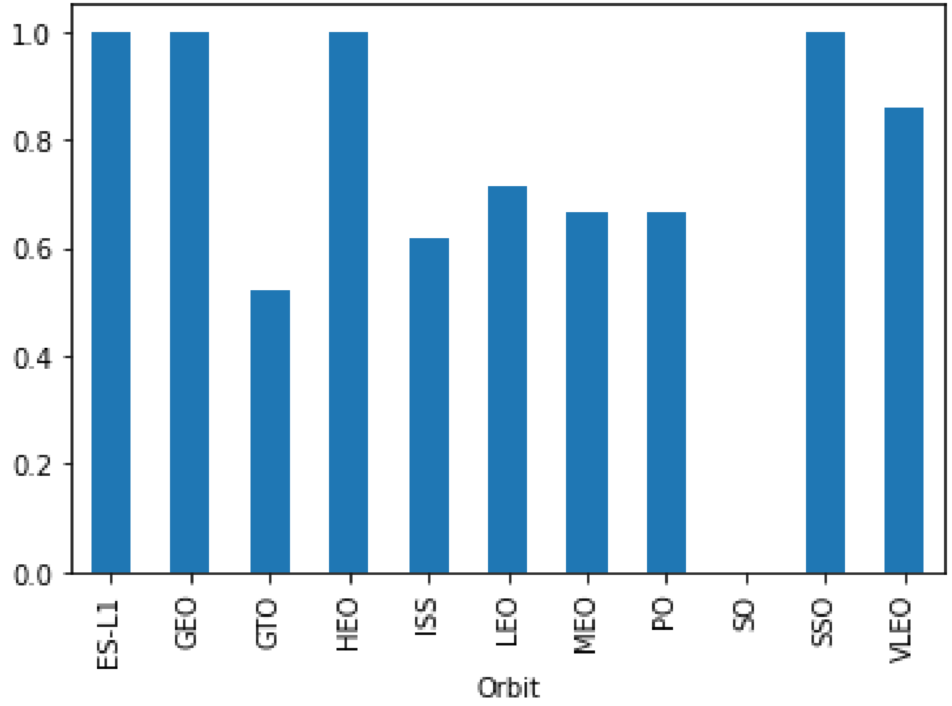
- To visualize data, scatterplots and bar plots were created between pair of features:
  - Launch Site vs. Payload Mass



- Source: https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/05%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with Data Visualization

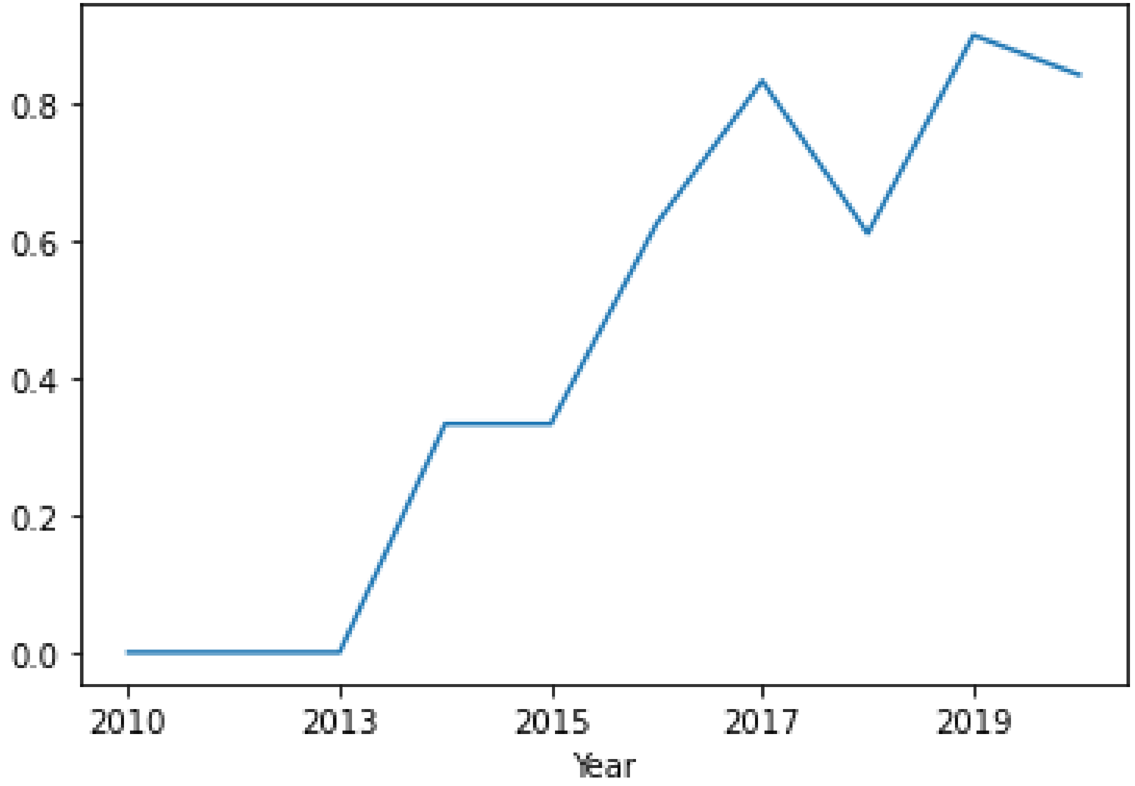
- To visualize data, scatterplots and bar plots were created between pair of features:
  - Success rate of Orbit classes



  - Source: https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/05%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with Data Visualization

- To visualize data, scatterplots and bar plots were created between pair of features:
  - Success rate over years



- Source: https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/05%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- SQL queries performed are:
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begin with “CCA”
  - Total payload mass carried by boosters launched by “NASA (CRS)”
  - Average payload mass carried by booster version F9 v1.1
  - Date of first successful landing outcome in ground pad
  - Names of the successful boosters with payload mass of 4000 to 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions carring the maximum payload mass
  - Failed landing outcomes information in year 2015
  - Landing outcomes between dates 2010-06-04 and 2017-03-20.
- Source: https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/04%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- To make maps more self-explanatory, markers, circles, lines and marker clusters were added to Folium Maps:
    - Markers to show points like launch sites
    - Circles for areas around specific coordinates, like NASA Johnson Space Center
    - Marker clusters to groups of events in coordinates, like launches in a launch site
    - Lines to measure distances between two coordinates

Build a Dashboard with Plotly Dash

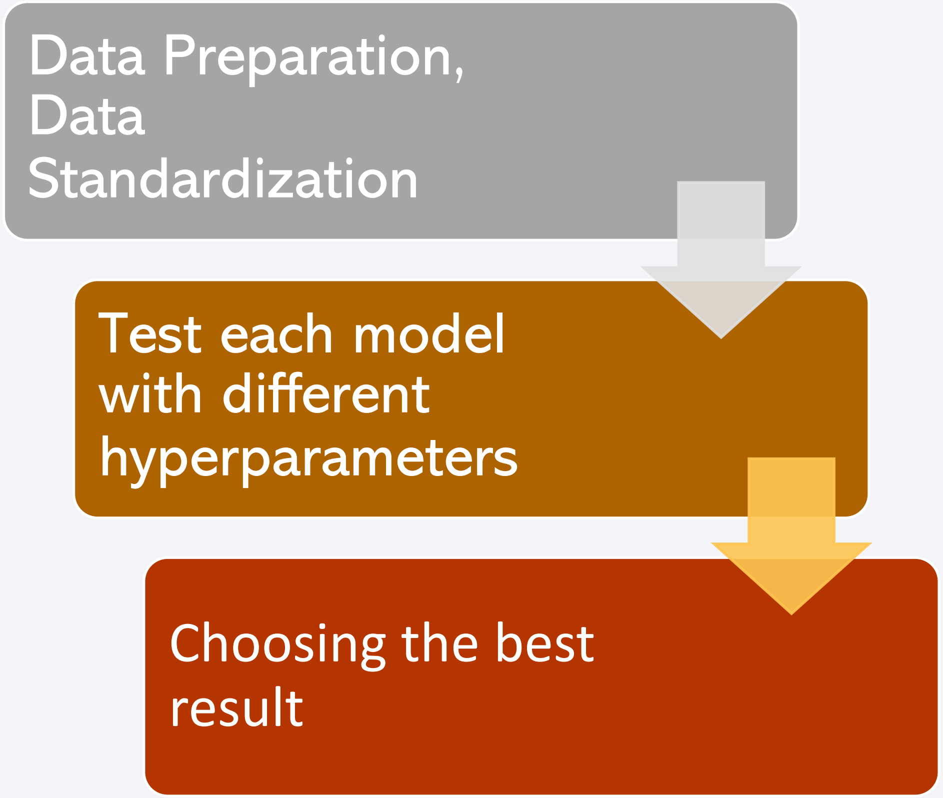
- To create an interactive dashboard, Plotly Dash is used to create these charts:
  - Percentage of launches by site
  - Payload range

- The dashboard was suitable to analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- Source: https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/06%20-%20spacex_dash_app.py

Predictive Analysis (Classification)

- To choose the best classification model, 4 models were implemented:
  - Logistic Regression
  - Support Vector Machines (SVM)
  - Decision Tree
  - K Nearest Neighbors (KNN)
- The process was as the flowchart:



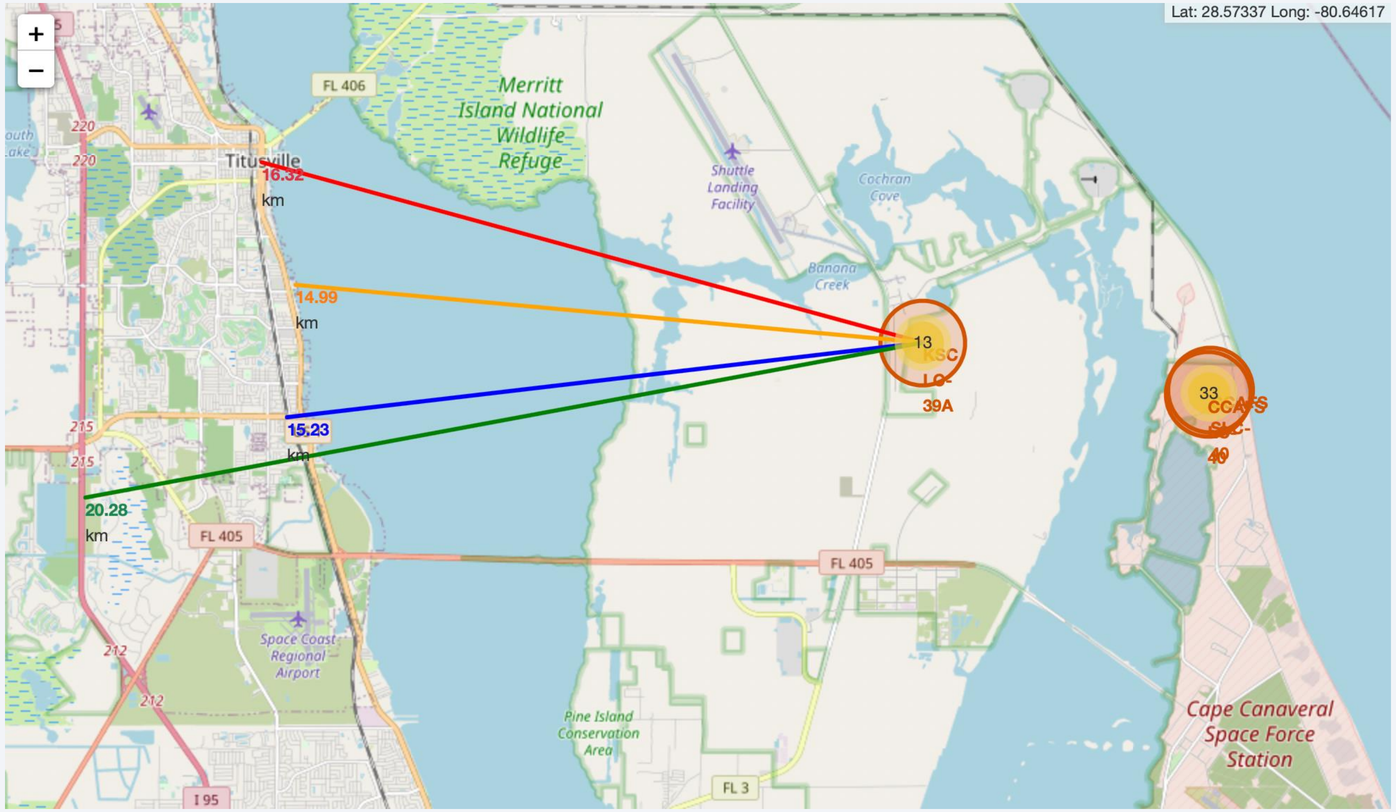
- Source: https://github.com/smhzad/Applied-Data-Science-Capstone/blob/main/07%20-%20Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites.
  - The first launches were done to Space X and NASA
  - The average payload of F9 v1.1 booster is 2,928 kg
  - The first success landing happened in 2015 fiver year after start.
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
  - Almost 100% of mission outcomes were successful
  - Two booster versions failed in 2015: F9 v1.1 B1012 & F9 v1.1 B1015;
  - The number of landing outcomes improves over years

Results

- We can identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around. Most launches happens at east cost sites (33) over west coast (13).



Results

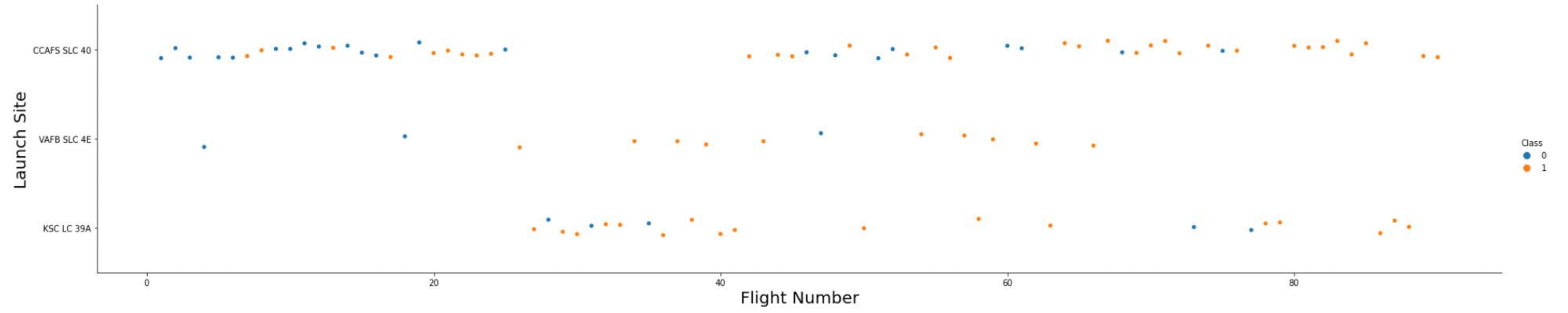
- Most scores are almost the same, but SVM has better score for whole dataset.

	LogReg	SVM	Tree	KNN
Jaccard_Score_whole_dataset	0.833333	0.845070	0.835821	0.819444
Jaccard_Score_train_data	0.800000	0.800000	0.600000	0.800000
F1_Score_whole_dataset	0.909091	0.916031	0.910569	0.900763
F1_Score_train_data	0.888889	0.888889	0.750000	0.888889
Accuracy_whole_dataset	0.866667	0.877778	0.877778	0.855556
Accuracy_train_data	0.833333	0.833333	0.666667	0.833333

Section 2

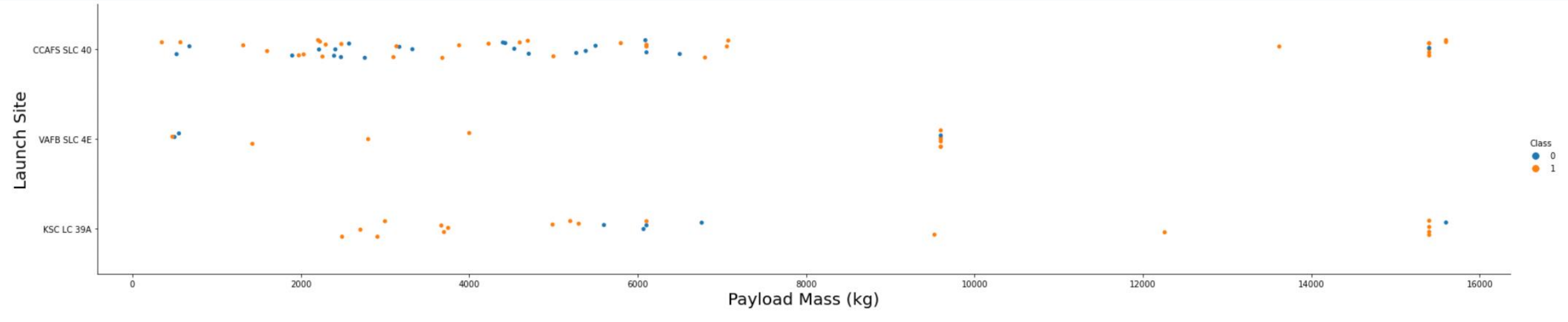
Insights drawn from EDA

Flight Number vs. Launch Site



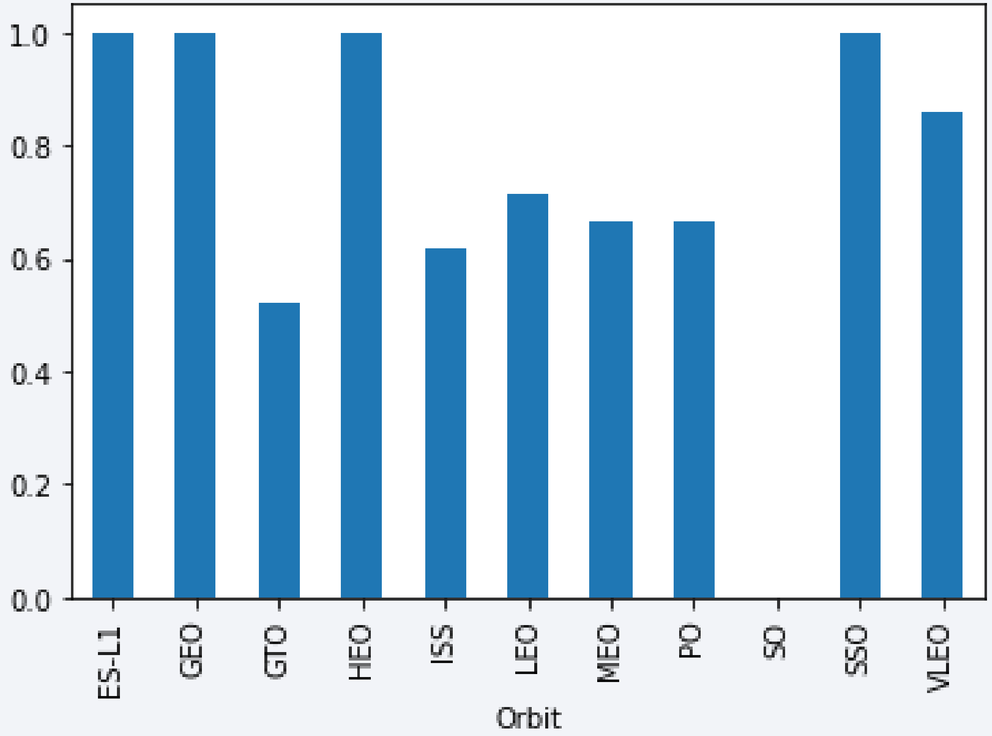
- Explanation:
  - The earliest flights all failed while the latest flights all succeeded.
  - The CCAFS SLC 40 launch site has about a half of all launches.
  - VAFB SLC 4E and KSC LC 39A have higher success rates.

Payload vs. Launch Site



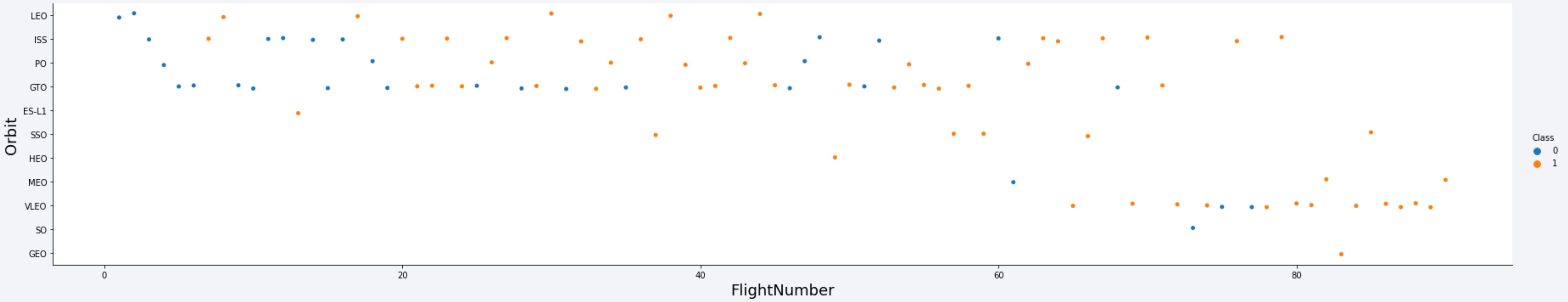
- Explanation:
  - For every launch site the higher the payload mass, the higher the success rate.
  - Most of the launches with payload mass over 7000 kg were successful.

Success Rate vs. Orbit Type



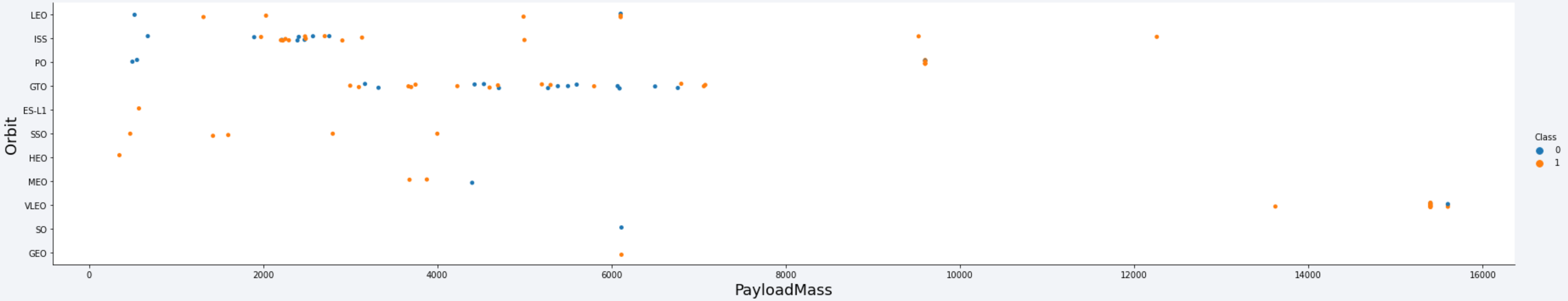
- Explanation:
  - SO orbit type: 0% success
  - SSO, HEO, ES-L1, GEO: 100% success

Flight Number vs. Orbit Type



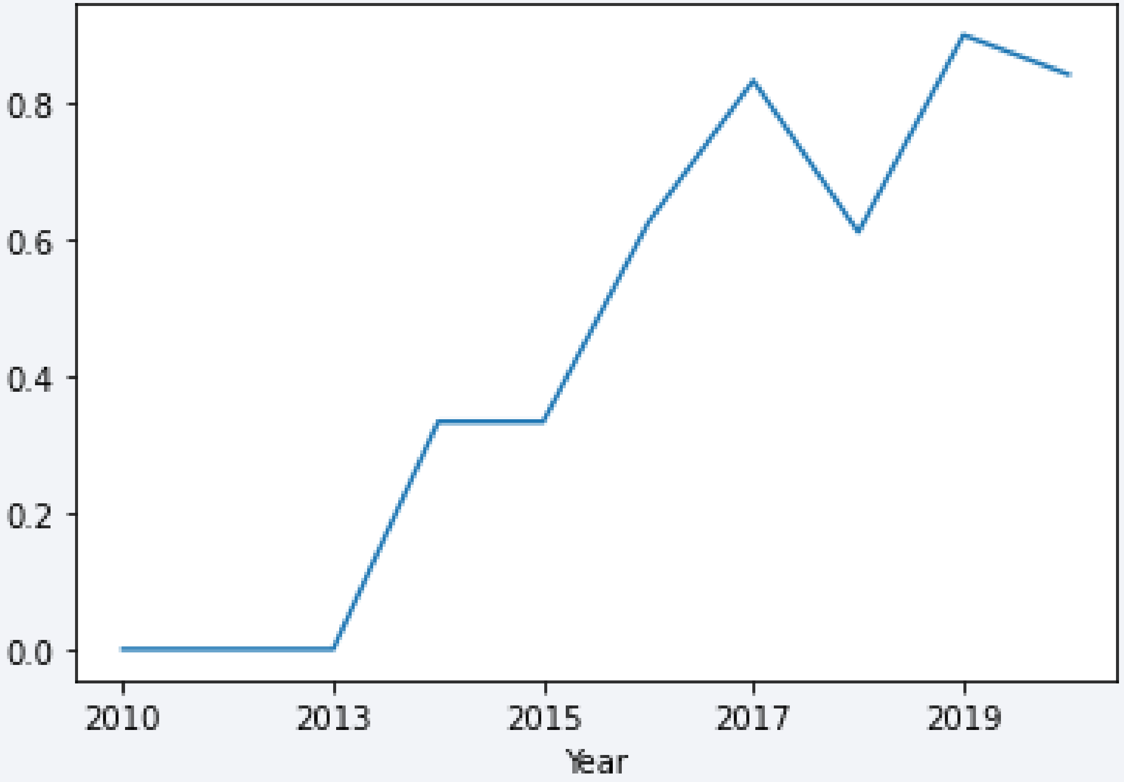
- Explanation:
  - It seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- Explanation:
  - Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend



- Explanation:
  - Success rate increases after 2013 almost continuously.

All Launch Site Names

```
%%sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2i
b?authSource=admin&replicaSet=replset
Done.
```

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

- Explanation:
  - Showing launch sites

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * FROM SPACEX
WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb?authSource=admin&replicaSet=replset
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

- Explanation:
  - Showing launch sites begins with ‘CCA’

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX
WHERE CUSTOMER='NASA (CRS)';
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82df
b?authSource=admin&replicaSet=replset
Done.
```

1
45596

- Explanation:
  - Showing Total Payload Mass

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX
WHERE BOOSTER_VERSION='F9 v1.1';
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82
b?authSource=admin&replicaSet=replset
Done.
```

]:

1
2928

- Explanation:
  - Showing Average Payload Mass by F9 v1.1

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) FROM SPACEX
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82df-4(
b?authSource=admin&replicaSet=replset
Done.
```

1
2015-12-22

- Explanation:
    - Showing First Successful Ground Landing Date

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEX
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.
b?authSource=admin&replicaSet=replset
Done.
```

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

- Explanation:
  - Showing Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEX
WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
ORDER BY BOOSTER_VERSION;
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82d
plicaSet=replset
Done.
```

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

- Explanation:
  - Showing Boosters Carried Maximum Payload

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEX
GROUP BY MISSION_OUTCOME
ORDER BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82df-40d1b13897
plicaSet=replset
Done.
```

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

- Explanation:
  - Showing Total Number of Successful and Failure Mission Outcomes

2015 Launch Records

```
%%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX
WHERE
LANDING__OUTCOME = 'Failure (drone ship)'
AND
DATE_PART('YEAR', DATE) = 2015;
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82df-40d1b
b?authSource=admin&replicaSet=replset
Done.
```

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

- Explanation:
  - Showing 2015 Launch Records

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

```
%%sql
SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEX
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY QTY DESC;
```

```
 * ibm_db_sa://twx04938:***@764264db-9824-4b7c-82df-40d1
plicaSet=replset
Done.
```

landing__outcome	qty
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

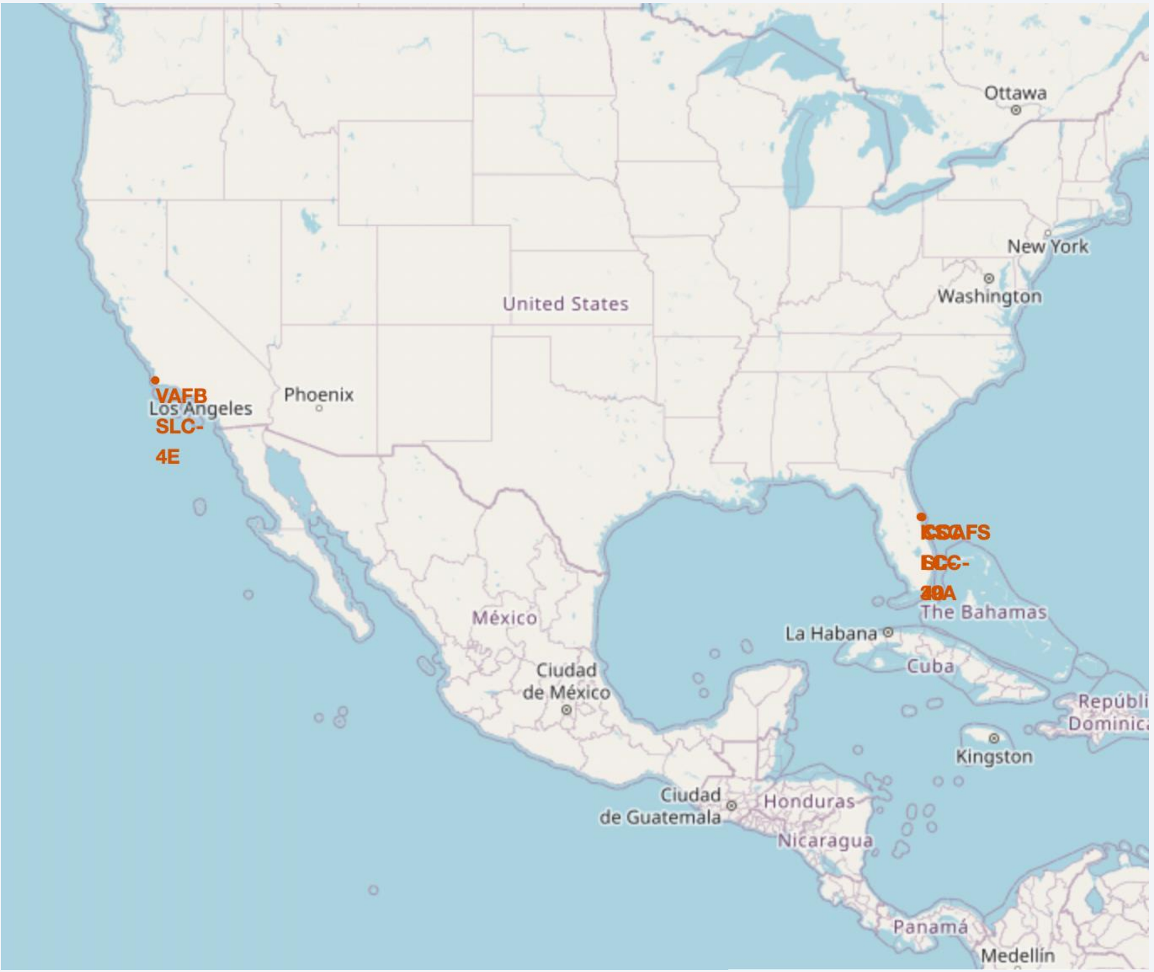
- Explanation:
  - Showing Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

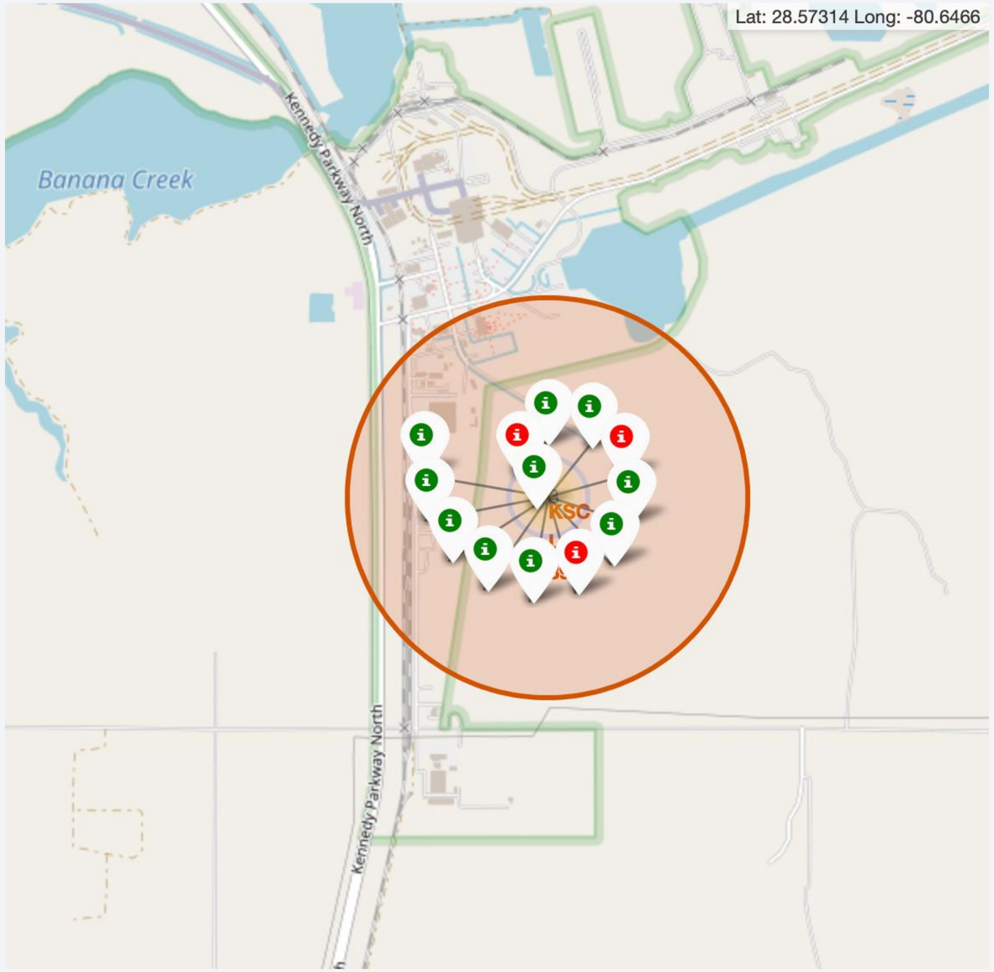
All launch sites' location markers on a global map

- Explanation:
    - Most launch sites are in east coast.
    - Sites are near populated areas and is potentially dangerous.



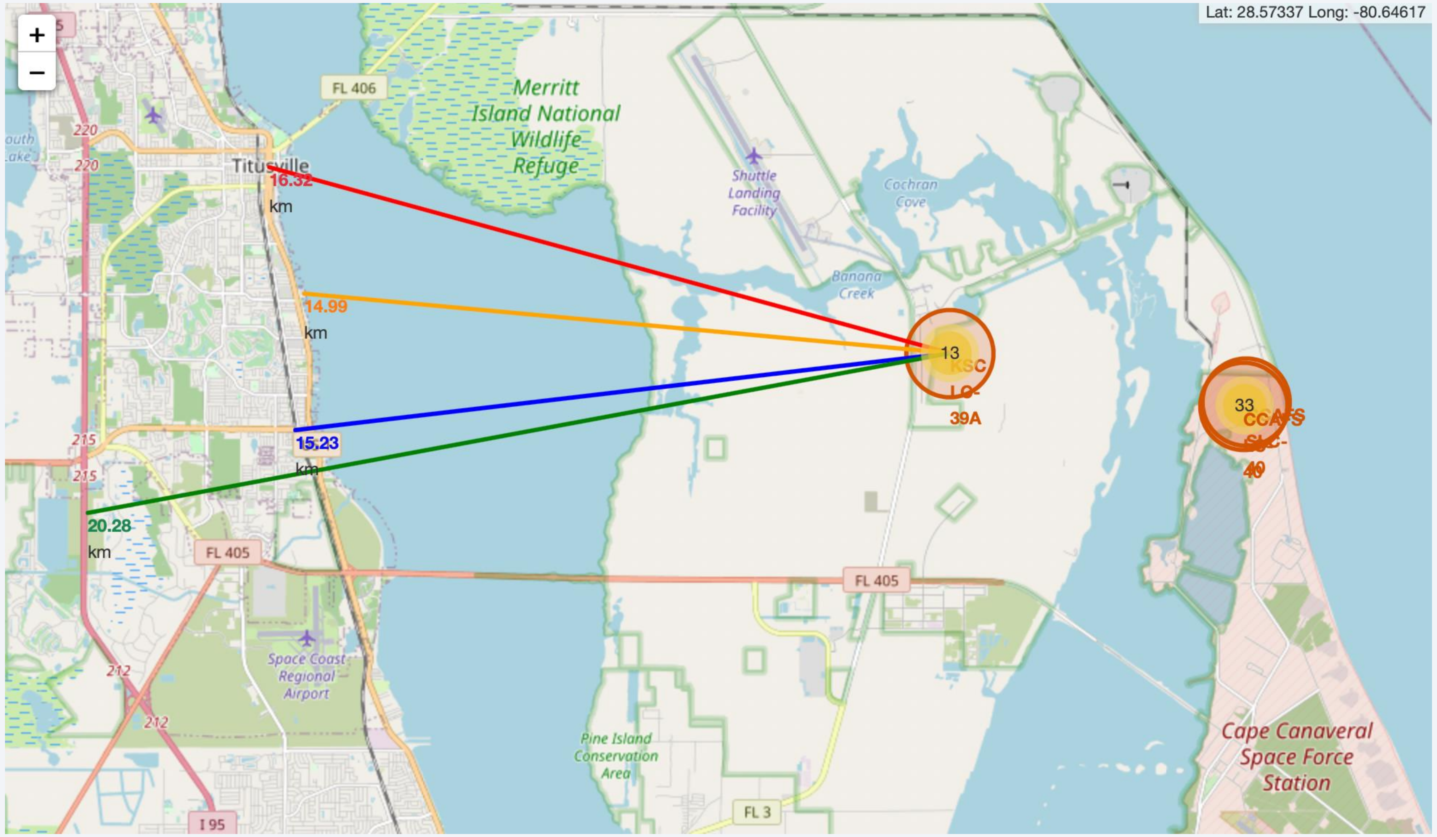
Color-labeled launch records on the map

- Explanation:
  - Green is successful, red in not.
  - Launch Site KSC LC-39A has a very high Success rate



Distance from the launch site KSC LC-39A to its proximities

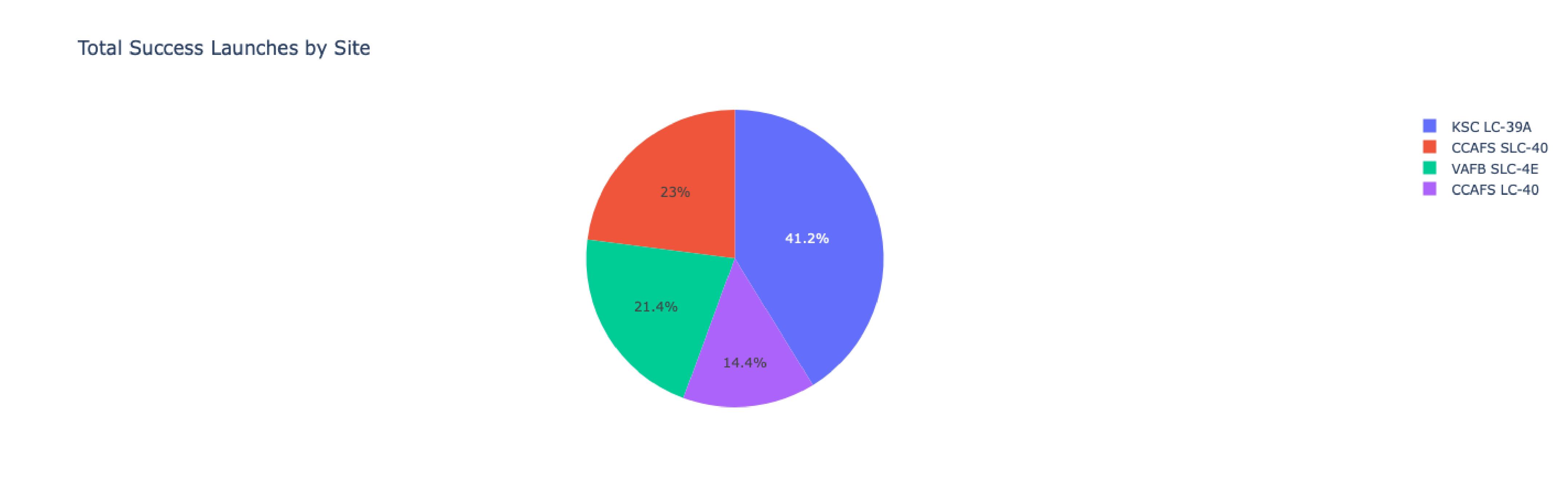
- Explanation:
  - KSC LC-39A is close to railroad, nearest city and coastline.
  - Failure in 15 to 20km range can cause a catastrophy.



Section 4

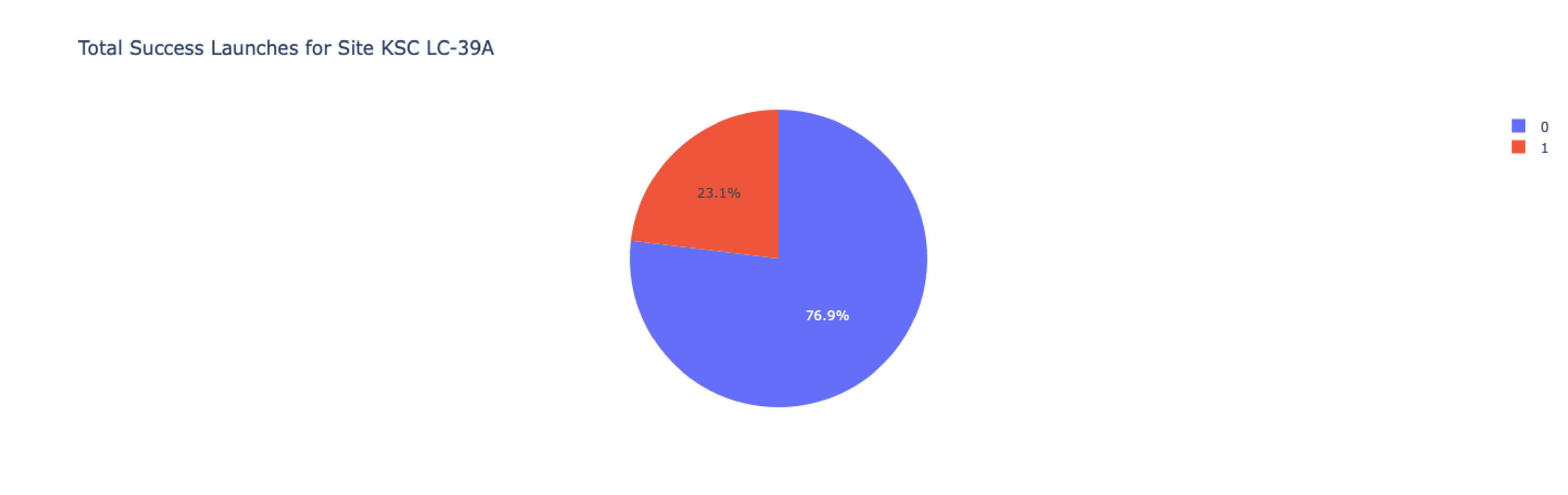
Build a Dashboard with Plotly Dash

Launch success count for all sites



- Explanation:
  - KSC LC-39A has the most successful launches

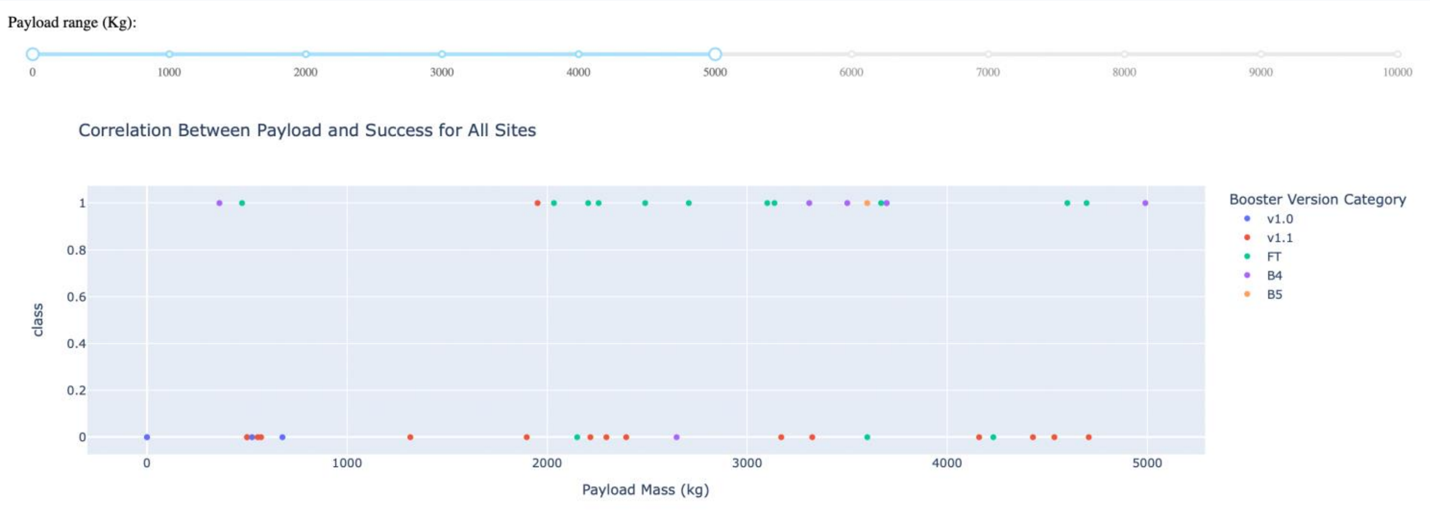
Launch site with highest launch success ratio

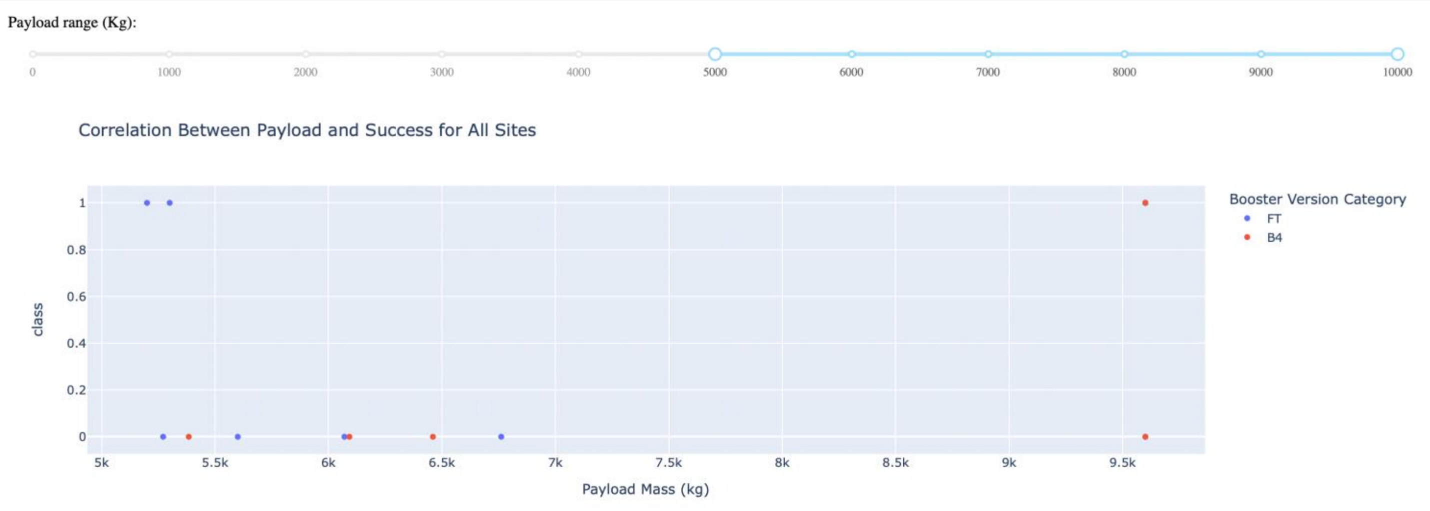


- Explanation:
  - KSC LC-39A has the highest launch success rate: 10 successful, 3 failed

Payload Mass vs. Launch Outcome for all sites

- Explanation:
  - Payloads between 2000 and 5500 kg have the highest success rate





Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Most scores are almost the same, but SVM has better score for whole dataset.

	LogReg	SVM	Tree	KNN
Jaccard_Score_whole_dataset	0.833333	0.845070	0.835821	0.819444
Jaccard_Score_train_data	0.800000	0.800000	0.600000	0.800000
F1_Score_whole_dataset	0.909091	0.916031	0.910569	0.900763
F1_Score_train_data	0.888889	0.888889	0.750000	0.888889
Accuracy_whole_dataset	0.866667	0.877778	0.877778	0.855556
Accuracy_train_data	0.833333	0.833333	0.666667	0.833333

Confusion Matrix

- Confusion matrix of SVM proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

1. The best launch site is KSC LC 39A;
2. Launches above 7,000kg are less risky
3. Success rate increases over time.
4. SVM Classifier can be used to predict successful landings.

Appendix

- To overcome CSV upload to IBM DB2, the following preprocessing has been done on the data: (changing dahs to slash)

```
[ ] import pandas as pd
```

```
df=pd.read_csv("Spacex.csv")
df.head()
```

```
df['Date']= df['Date'].replace('-','/',regex=True)
df.head()
```

```
[ ] df.to_csv("Spacex.csv", index=False, header=True)
```

Thank you!