Harpartap Singh
14-09-2024

Outline

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

Summary

Methodologies Used for Data Analysis:

- Data Collection through web scraping and the SpaceX API
- Exploratory Data Analysis (EDA), involving data wrangling, visualization, and interactive visual analytics
- Machine Learning Prediction

Summary of Results:

- Valuable data was successfully collected from public sources.
- EDA helped identify the key features that predict the success of launches.
- Machine Learning Prediction determined the most effective model for identifying critical characteristics to optimize this opportunity using all collected data.

Introduction

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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 SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch

- **Key Questions:**
- What trends and patterns are evident in SpaceX launch data?
- How do different factors (e.g., rocket type, launch site) impact the success rate of launches?
- What are the key indicators of a successful or failed launch?
- How can we predict the success of future launches based on historical data?
- What interactive visualizations can best represent the launch data and support decision-making?
- **Desirable answers:**
- The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
- Where is the best place to make launches.

METHODOLOGY

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_SpaceX_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) by using visualization tools to uncover patterns and insights, and SQL to query and manage the data.

Methodology (Contd.)

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

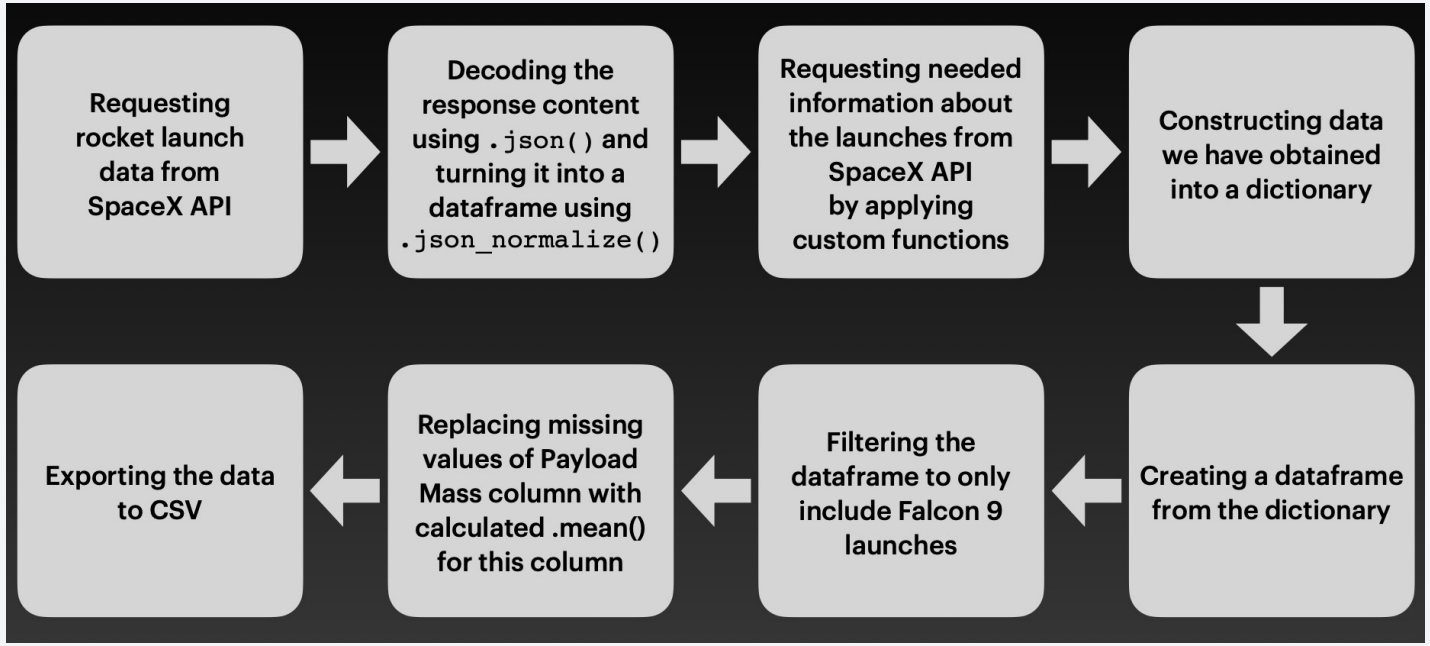
Data Collection

1. Data Sources

- SpaceX Launch Records: Gathered from SpaceX's official datasets and third-party data providers.
- APIs: Used SpaceX API for real-time and historical launch data.
- Supplementary Sources: Accessed additional data on rocket types and mission outcomes from aerospace data repositories.
- 2. Data Collection Steps
- Define Requirements: Identified needed data fields such as launch dates, rocket types, sites, and mission outcomes.
- Public Datasets: Downloaded CSV files or accessed online repositories.
- APIs: Retrieved data via API calls, ensuring accuracy and completeness.

Data Collection - SpaceX API

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.



Data Collection - Scraping

- **Data Source**

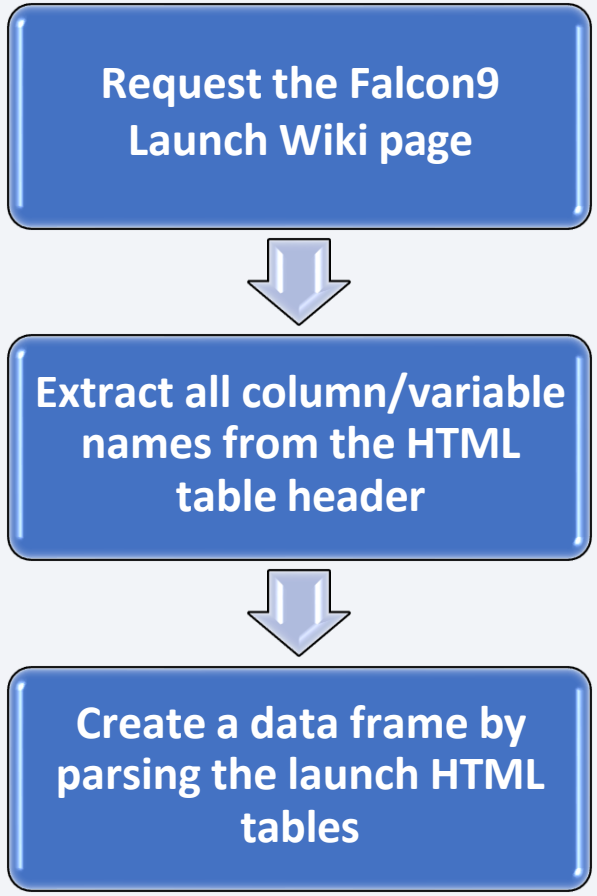
  SpaceX launch data is readily available on Wikipedia, which serves as a comprehensive and up-to-date source of information. This includes details on launch dates, rocket types, payloads, mission outcomes, and more.

- **Data Collection Process**

  The data extraction process follows a structured approach as defined in the flowchart. This ensures that the right data is captured systematically, avoiding manual errors and improving consistency in data collection.

- **Automated Download**

  The information from Wikipedia is downloaded using automated tools or scripts, streamlining the process and reducing the time required to retrieve the data. This automation helps keep the data current and reduces the need for manual intervention.

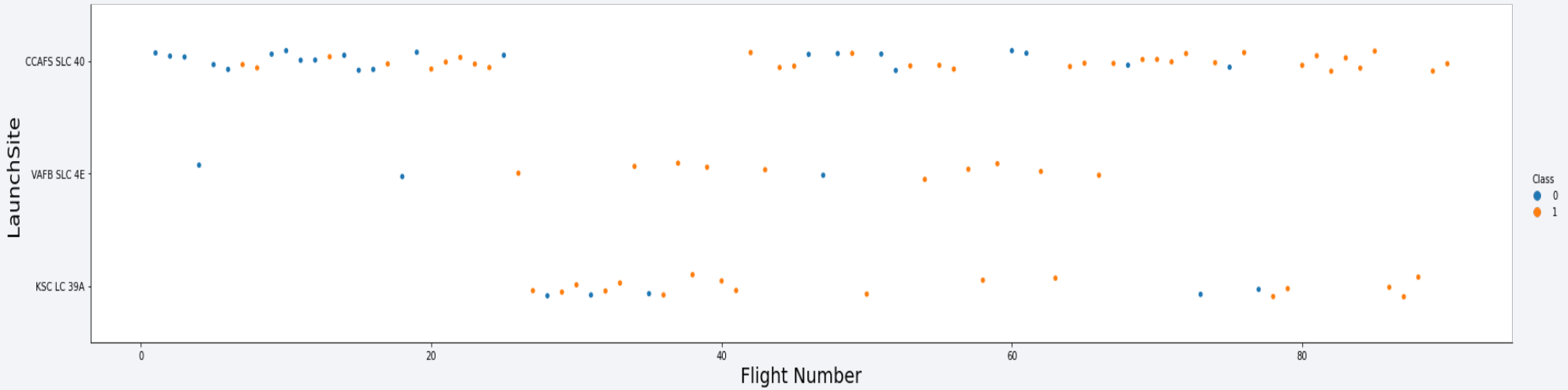


Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



EDA with SQL

- The following SQL queries were performed:
  - Names of the unique launch sites used in space missions.
  - Top 5 launch sites whose names begin with the string 'CCA'.
  - Total payload mass carried by boosters launched by NASA under the Commercial Resupply Services (CRS) program.
  - Average payload mass carried by Falcon 9 booster version 1.1.
  - Date of the first successful landing outcome on a ground pad.
  - Names of boosters with successful drone ship landings and payload masses between 4000 and 6000 kg.
  - Total count of successful and failed mission outcomes.
  - Booster versions that have carried the maximum payload mass.
  - Failed landing outcomes on drone ships in 2015, along with the corresponding booster versions and launch site names.
  - Rank of landing outcomes (e.g., Failure on drone ship, Success on ground pad) between the dates 2010-06-04 and 2017-03-20

Build an Interactive Map with Folium

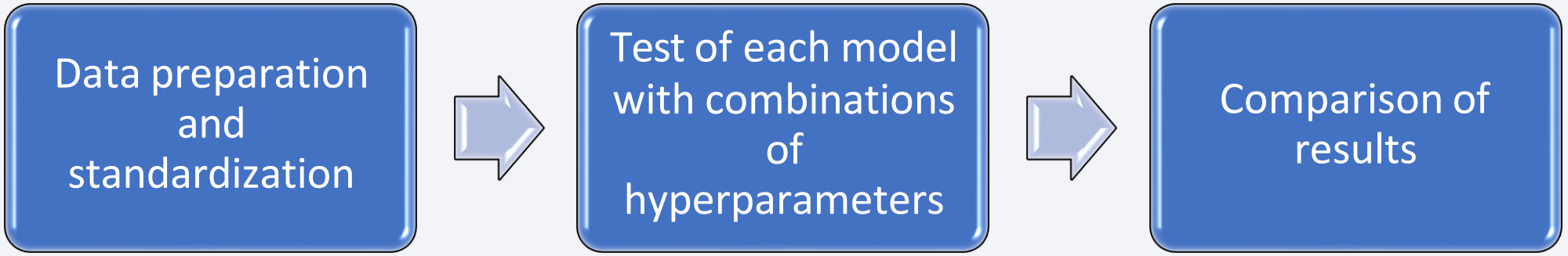
- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
  - Lines are used to indicate distances between two coordinates.

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



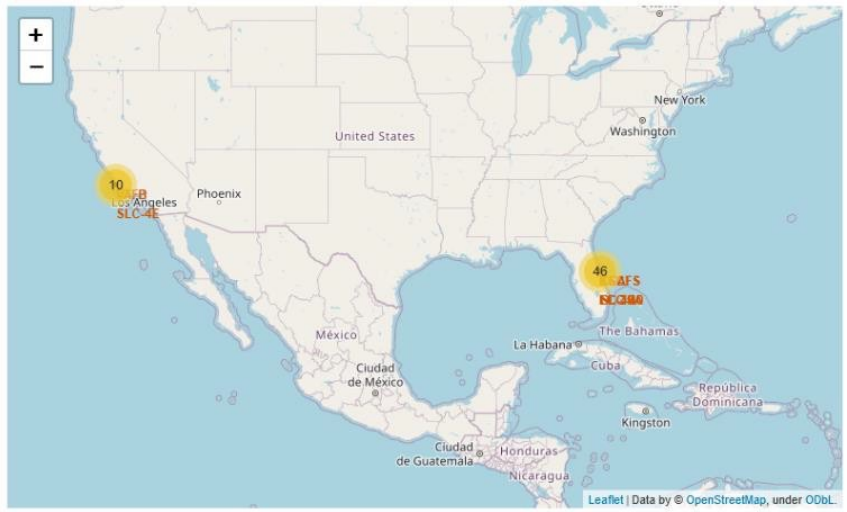
Results

- Exploratory data analysis results:
  - SpaceX uses four different launch sites.
  - The initial launches were conducted for SpaceX itself and NASA.
  - The average payload of the Falcon 9 v1.1 booster is 2,928 kg.
  - The first successful landing outcome occurred in 2015, five years after the first launch.
  - Several Falcon 9 booster versions successfully landed on drone ships with payloads above the average.
  - Nearly 100% of mission outcomes were successful.
  - Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed at landing on drone ships in 2015.
  - The quality of landing outcomes improved over the years..

Results (Contd.)

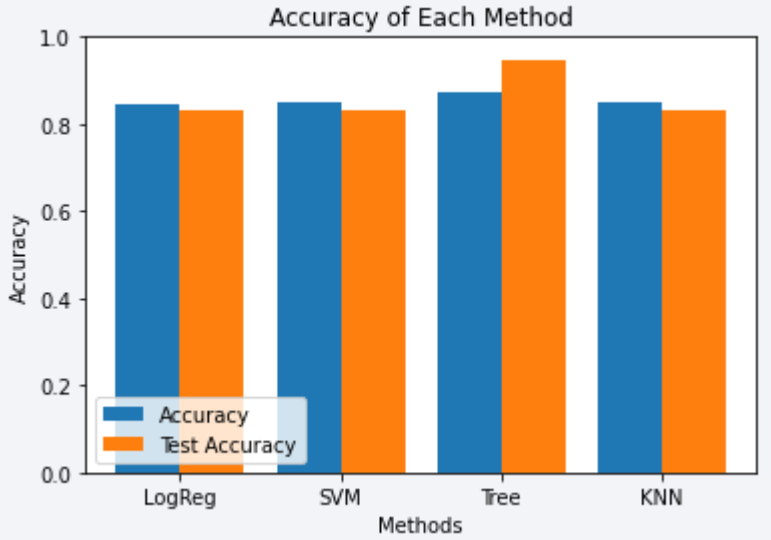
Using interactive analytics, it was possible to identify the following:

- Launch sites are generally situated in safe locations, often near the sea.
- These sites are equipped with strong logistical infrastructure.
- The majority of launches take place at east coast launch sites.



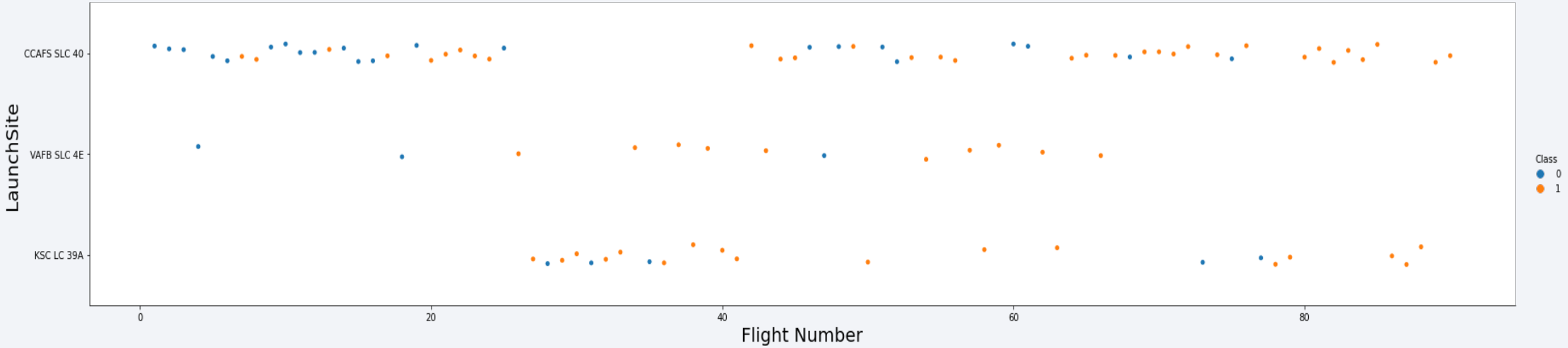
Results (Contd.)

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.



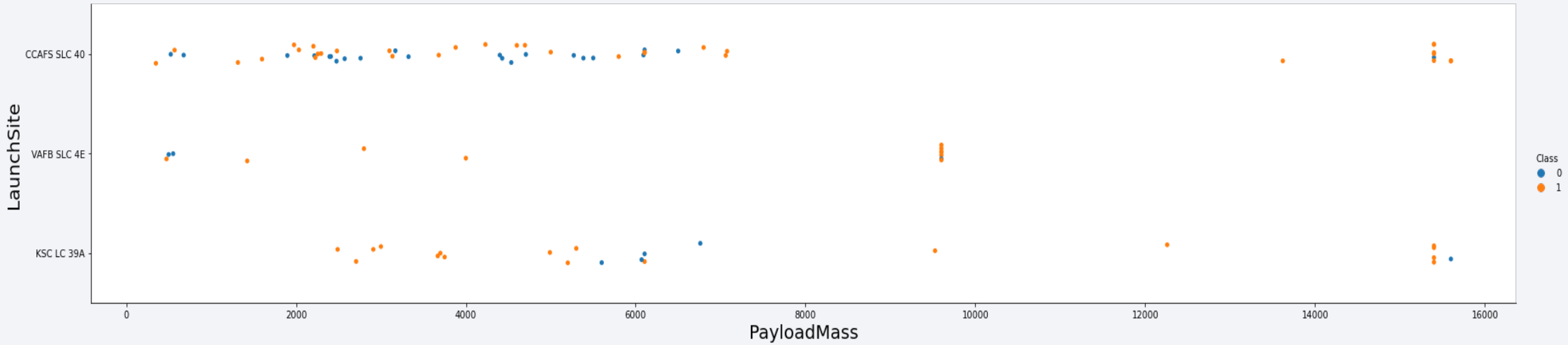
INSIGHTS DRAWN FROM EDA

Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

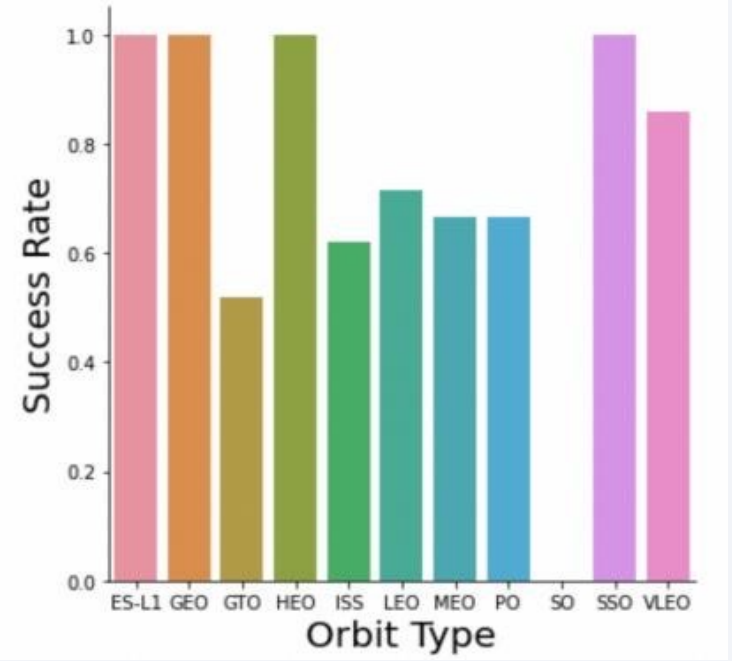
Payload vs. Launch Site



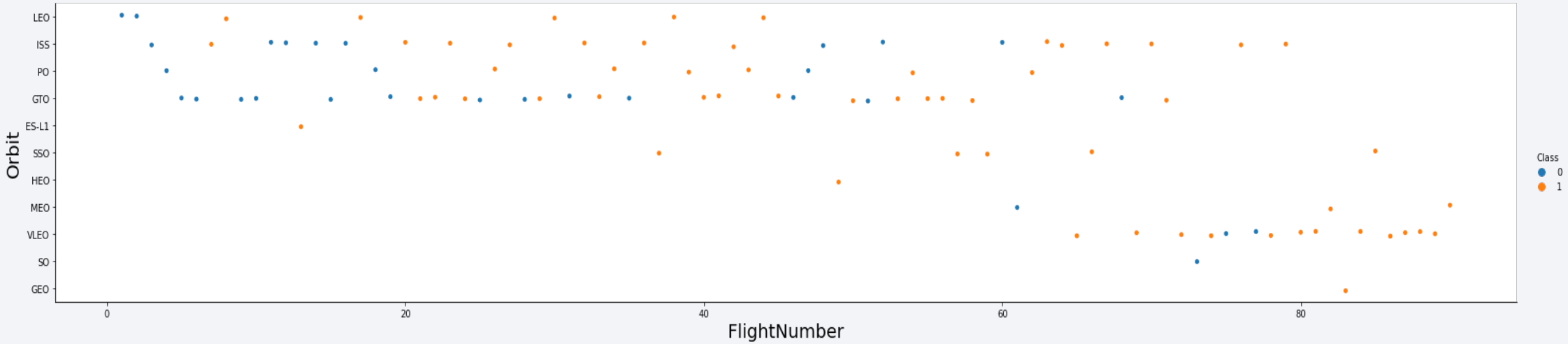
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- 100 % success rates happens to orbits:
  - ES-L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).

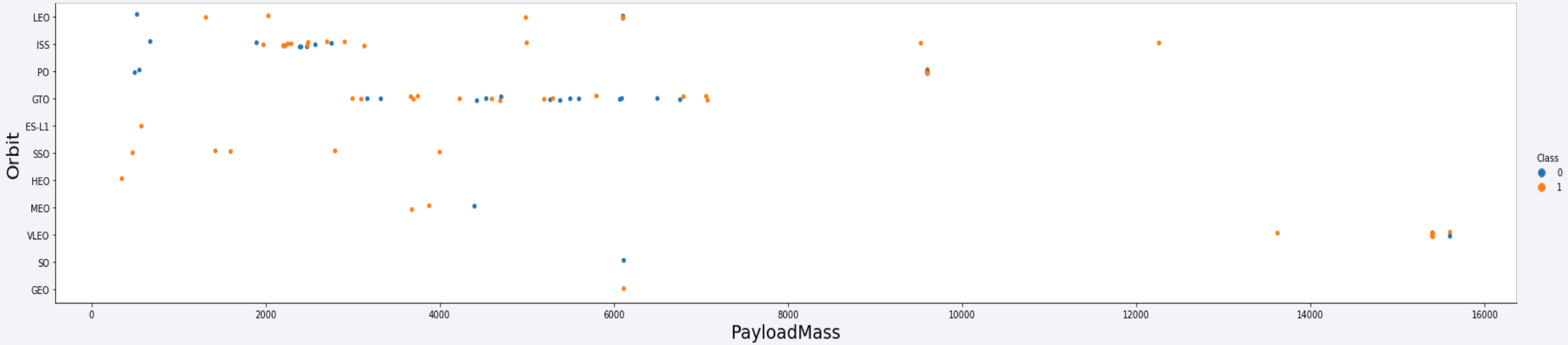


Flight Number vs. Orbit Type



- Apparently, success rate improved over time to all orbits;
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

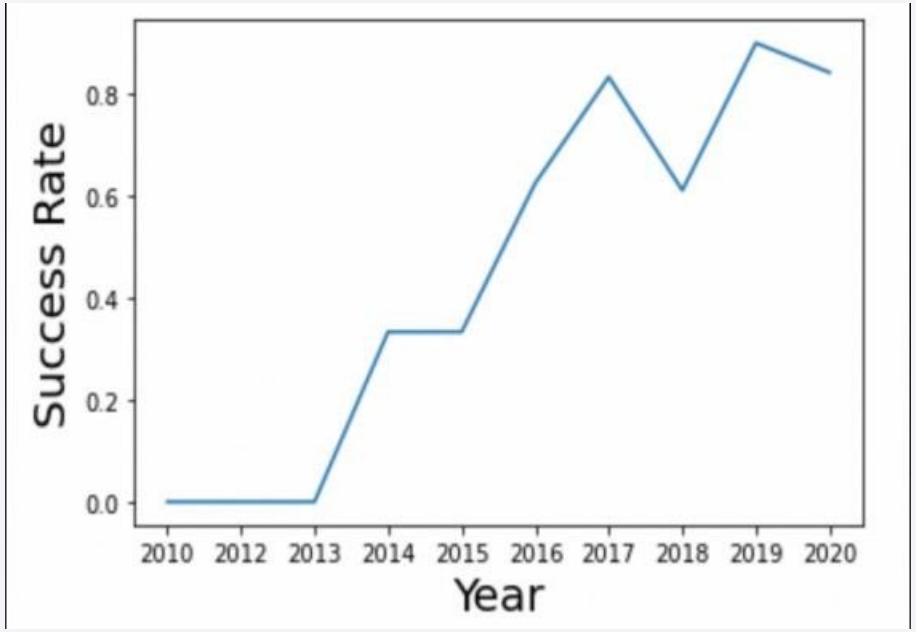
Payload vs. Orbit Type



- Apparently, there is no relation between payload and success rate to orbit GTO;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.



All Launch Site Names

- According to the Space dataset, there are 4 launch sites:

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[4]:
```

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

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

- Displaying 5 records where launch sites begin with the string 'CCA':

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA:

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

total_payload_mass
45596

- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[7]:
```

average_payload_mass
2534

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2015-12-22

- We achieved this by filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

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

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
             where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

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

- The list above has the only two occurrences.

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

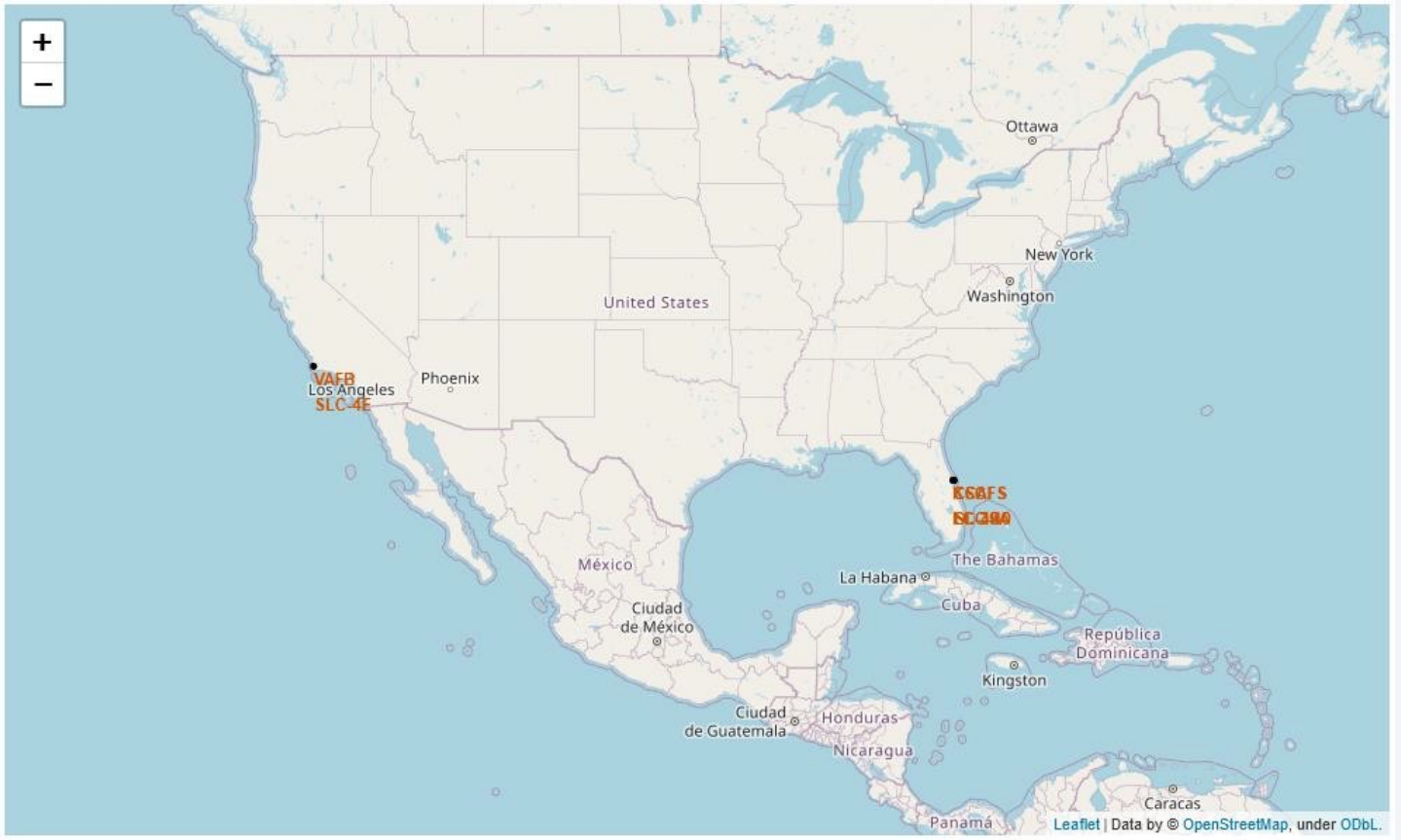
•Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

Landing Outcome	Occurrences
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

LAUNCHSITES PROXIMITY ANALYSIS

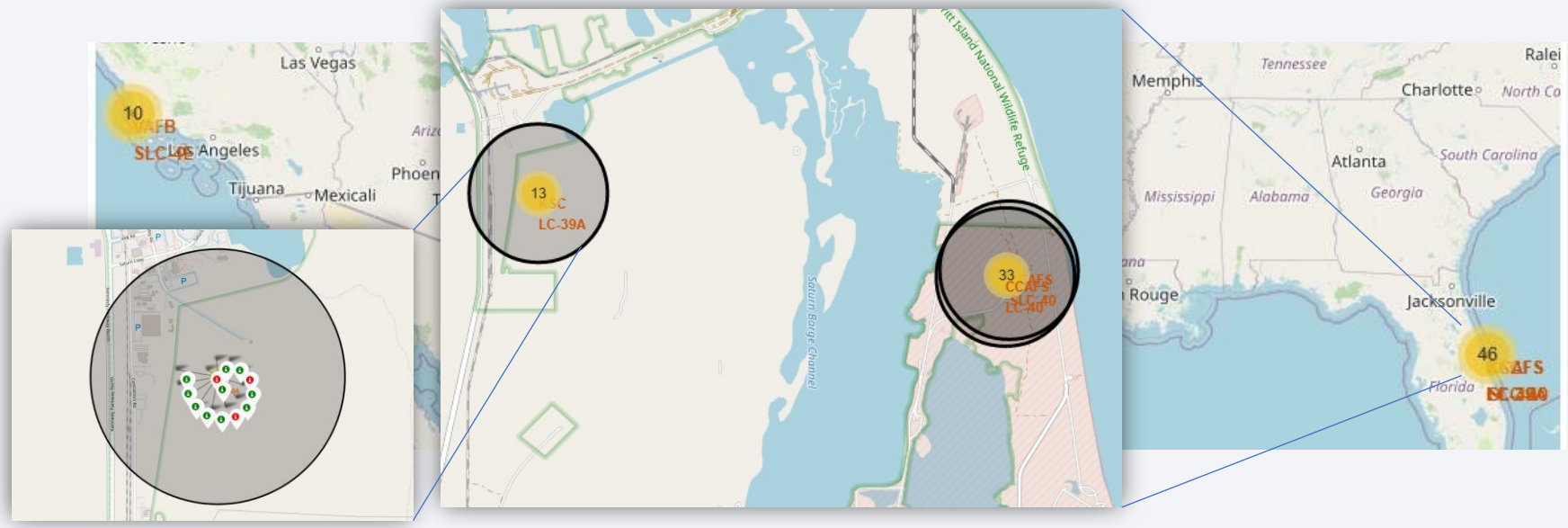
All launch sites

Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.



Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes



- Green markers indicate successful and red ones indicate failure.

Logistics and Safety

Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

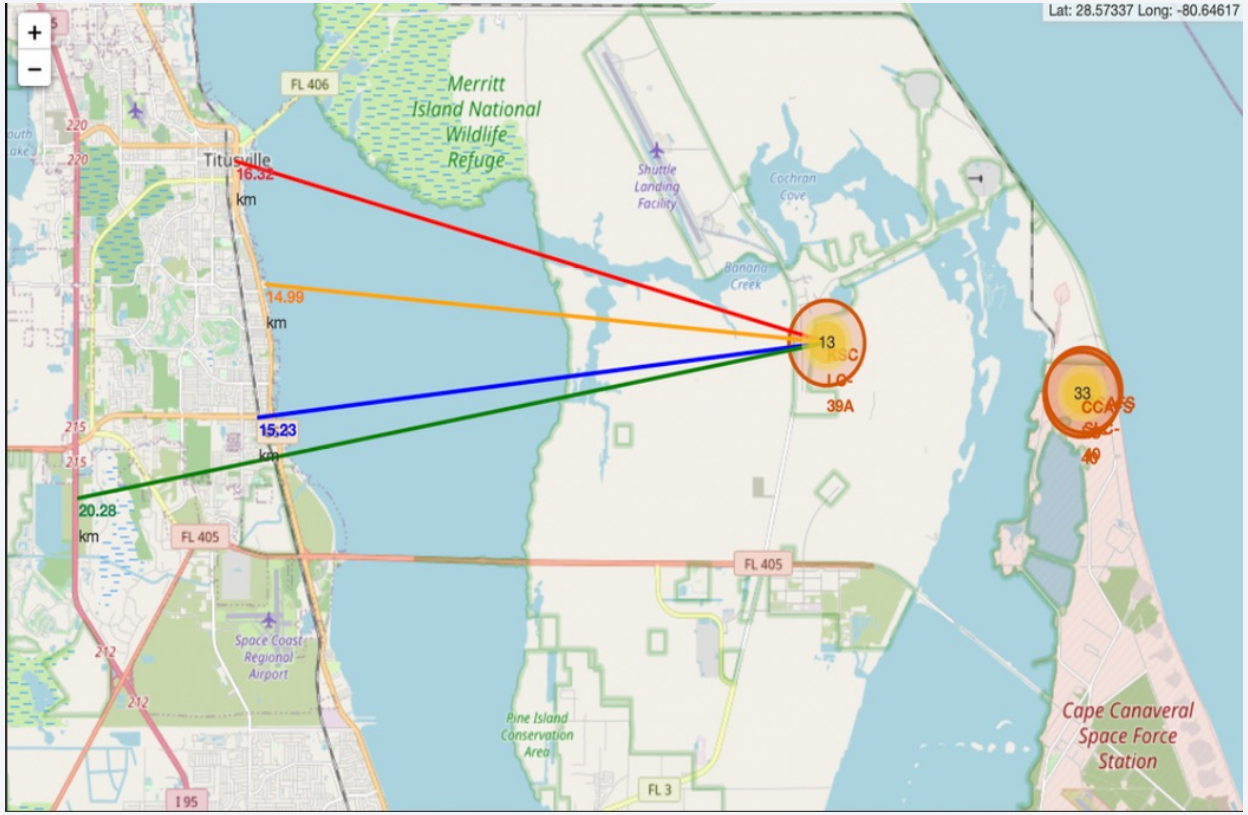
- relative close to railway (15.23 km)

- relative close to highway (20.28 km)

- relative close to coastline (14.99 km)

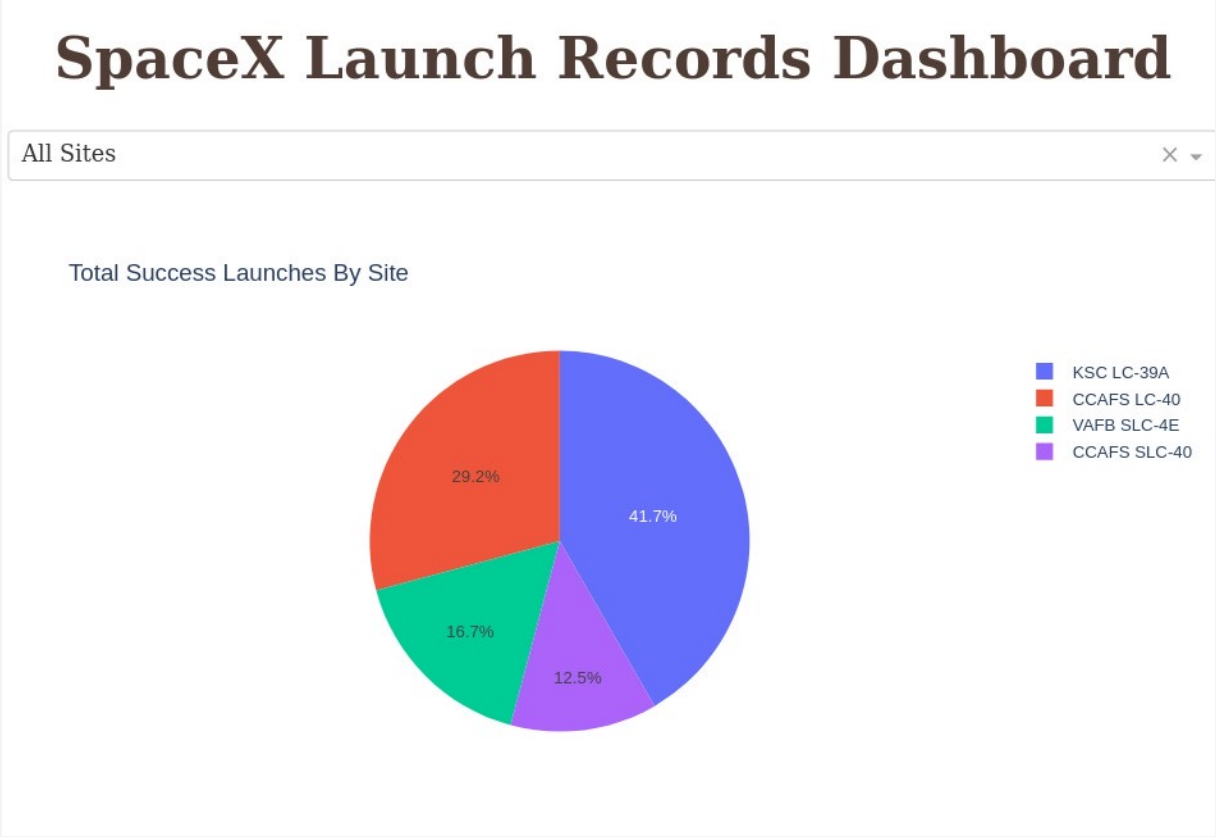
•Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

•Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



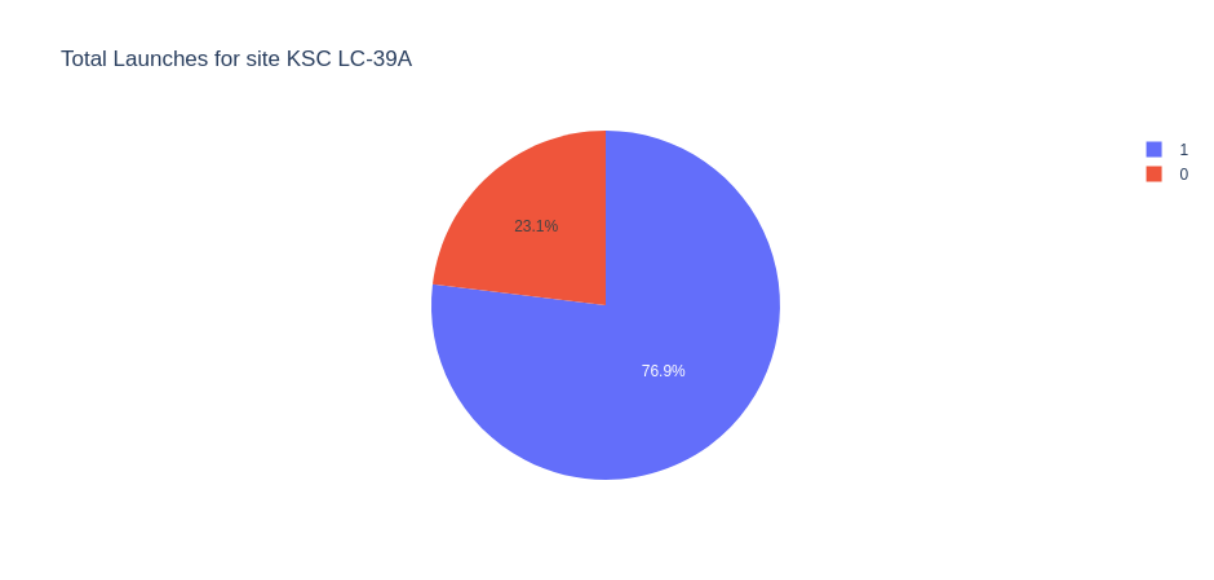
BUILDING A DASHBOARD WITH POTLY DASH

Successful Launches by Site



- The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings

Payload vs. Launch Outcome

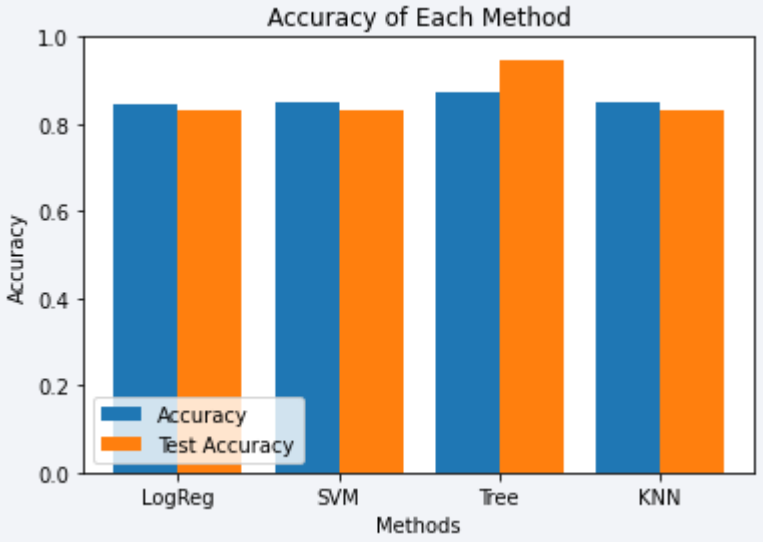
- The charts show that payloads between 2000 & 5500 kg have the highest success rate

PREDICTIVE ANALYSIS (CLASSIFICATION)

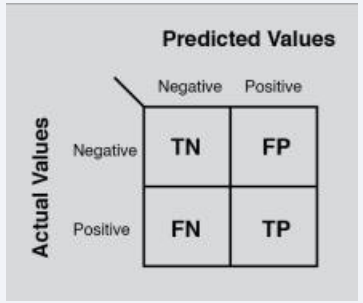
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix of Decision Tree Classifier



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

Conclusions

- SpaceX has achieved substantial improvements in landing technology, with success rates showing notable increases over the years. This advancement is a key factor in the company's growing reliability in space missions.
- The CCAFS SLC-40 launch site stands out as one of the most successful locations for launches, consistently delivering higher success rates compared to other sites.
- The Decision Tree model has proven to be a highly effective tool for predicting the likelihood of successful rocket landings, offering an accuracy rate that can significantly support future missions.
- By utilizing the Decision Tree model, SpaceX can optimize costs by accurately forecasting landing outcomes, leading to more efficient mission planning and better resource allocation.

Appendix

- Data Collection: Data was gathered using the SpaceX API and web scraping from Wikipedia. Key challenges included handling incomplete data and ensuring accuracy in scraped information.

- Model Comparison: Four models—Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors—were compared. The Decision Tree showed the highest accuracy at 87%, making it the best predictor of landing success.

- Feature Importance: The most influential factors for predicting successful landings were launch site, payload mass, and orbit type. These key features were used to improve model accuracy.

Thank you!

Instructor, Coursera ,IBM, Evaluator