IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Md Mohidul Haque>
<15.12.2023>

Outline

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

Executive Summary

- Following Methodiologies used to analyzed the data:
  - Web scraping and SpaceX API used to collect data
  - EDA (Exploratory Data Analysis)
  - ML
- Summary of all results
  - Collect data from public sourses
  - Feature selection by EDA
  - Predictive ML model to choose impactful characteristics

Introduction

- Project background:

  Evaluate competeability of a company Space Y against Space X
- Outcomes:
  - Successful landing of the rockets in the first stage is the best way to estimate the total cost for launches
  - Place to make launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data obtained from two public sources.
    1. SpaceX API
    2. WebScraping
- Perform data wrangling
  - landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into train and test set, use different classification algorithms, choose the best one

Data Collection

- Using SpaceX API
- Wikipedia by WebScraping

Data Collection – SpaceX API

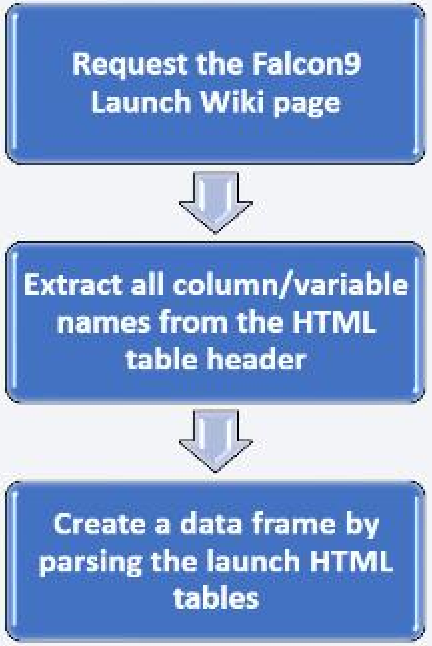
- Public API where data is stored
- GitHub URL

https://github.com/MohidulHaqueTushar/Applied-Data-Science-Capstone/blob/main/1.%20jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

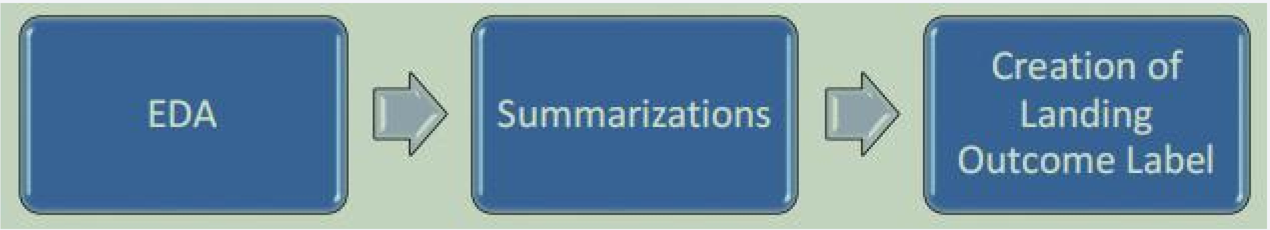
- From Wikipedia
- GitHub URL

https://github.com/MohidulHaqueTushar/Applied-Data-Science-Capstone/blob/main/2.%20jupyter-labs-webscraping.ipynb



Data Wrangling

- Flowcharts

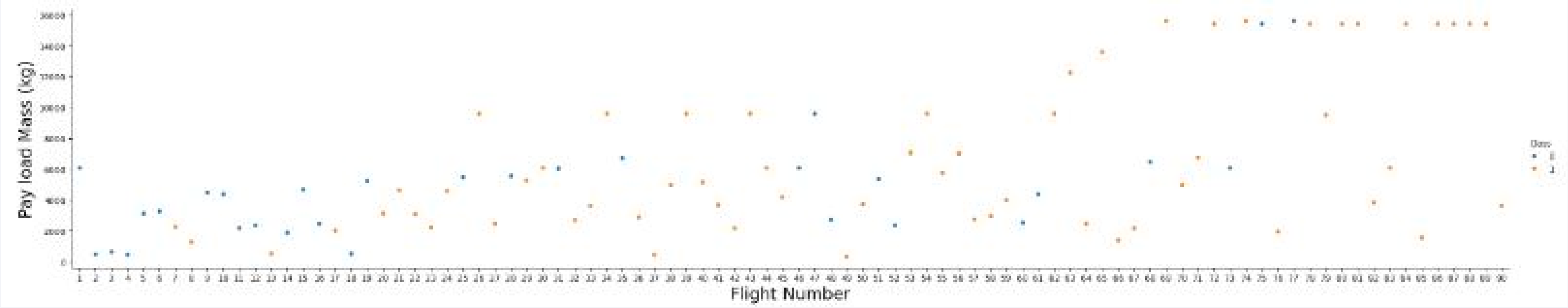


- GitHub URL

https://github.com/MohidulHaqueTushar/Applied-Data-Science-Capstone/blob/main/3.%20labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatterplots and Barplots were used to visualized the relationship between features



- GitHub URL

https://github.com/MohidulHaqueTushar/Applied-Data-Science-Capstone/blob/main/5.%20IBM-DS0321EN-SkillsNetwork_labs_module_2_5.%20jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Performed SQL queries:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months 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.
- GitHub URL: https://github.com/MohidulHaqueTushar/Applied-Data-Science-Capstone/blob/main/4.%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
- GitHub URL

https://github.com/MohidulHaqueTushar/Applied-Data-Science-Capstone/blob/main/6.%20IBM-DS0321EN-SkillsNetwork_labs_module_3_6.%20lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

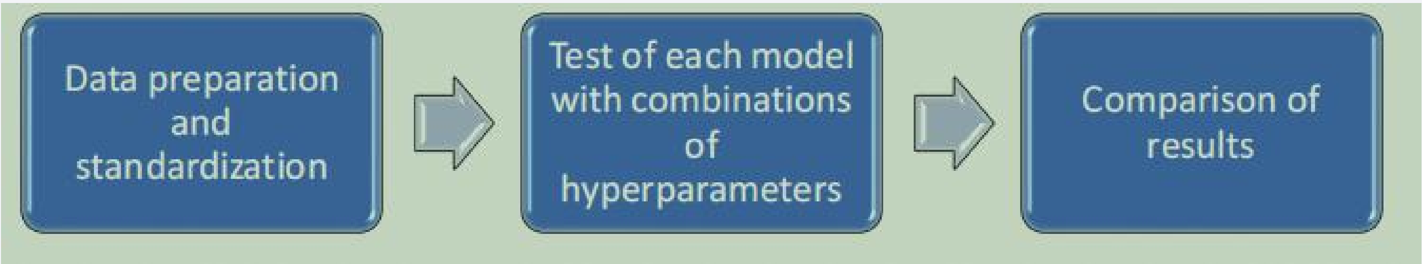
- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- GitHub URL

https://github.com/MohidulHaqueTushar/Applied-Data-Science-Capstone/blob/main/7.%20Interactive%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

- Classification Models:
    - Logistic regression
    - Support Vector Machine
    - Decision Tree
    - K-Nearest Neighbors



- GitHub URL

https://github.com/MohidulHaqueTushar/Applied-Data-Science-Capstone/blob/main/8.%20IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- EDA results:
  - Has 4 different launch sites
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
  - The number of landing outcomes became as better as years passed

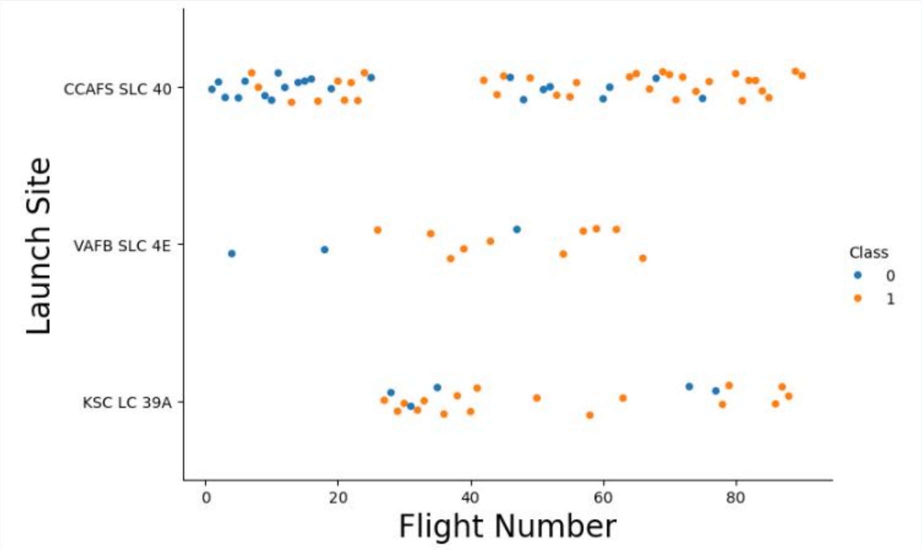
Results

- Launch sites: safe places

Section 2

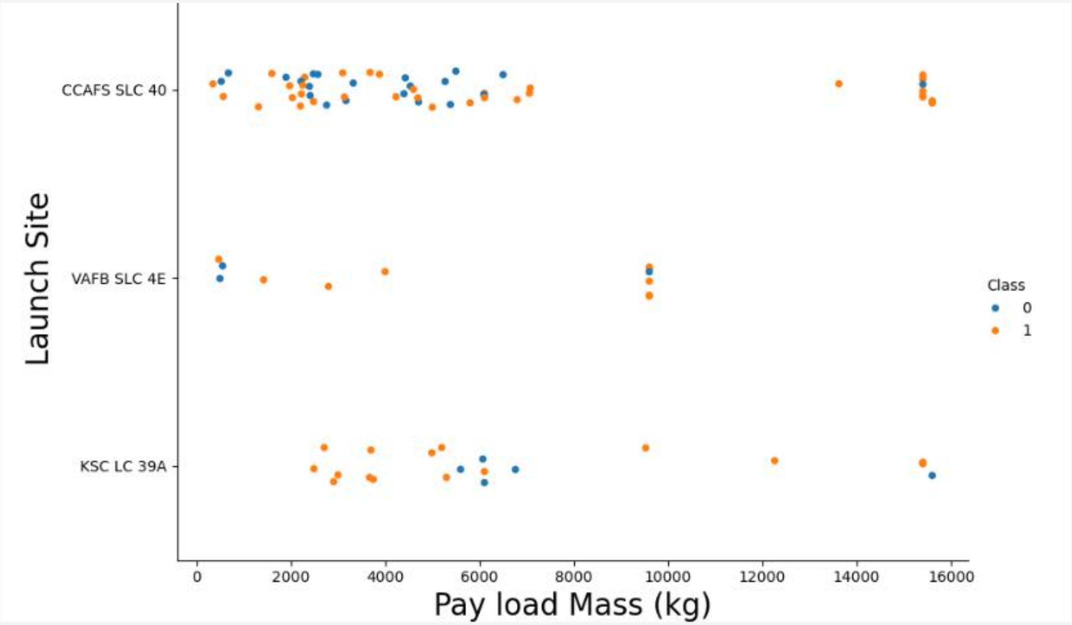
Insights drawn from EDA

Flight Number vs. Launch Site



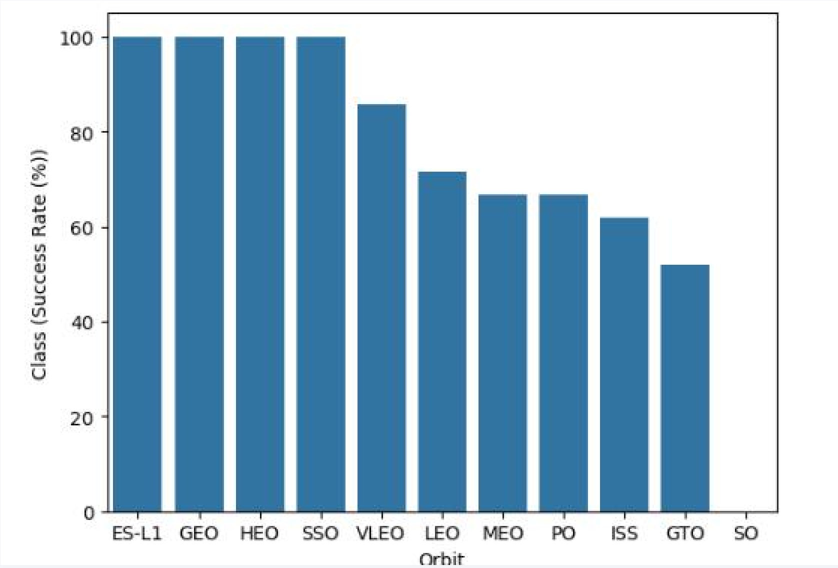
Possible that the general success rate improve over time.

Payload vs. Launch Site



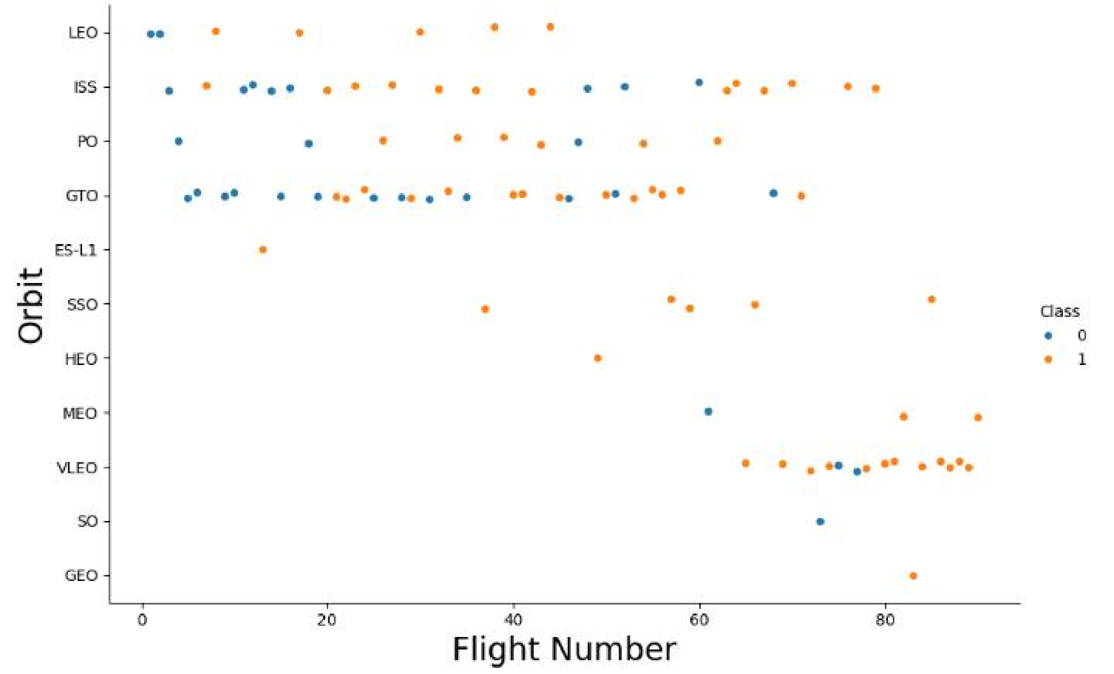
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate

Success Rate vs. Orbit Type



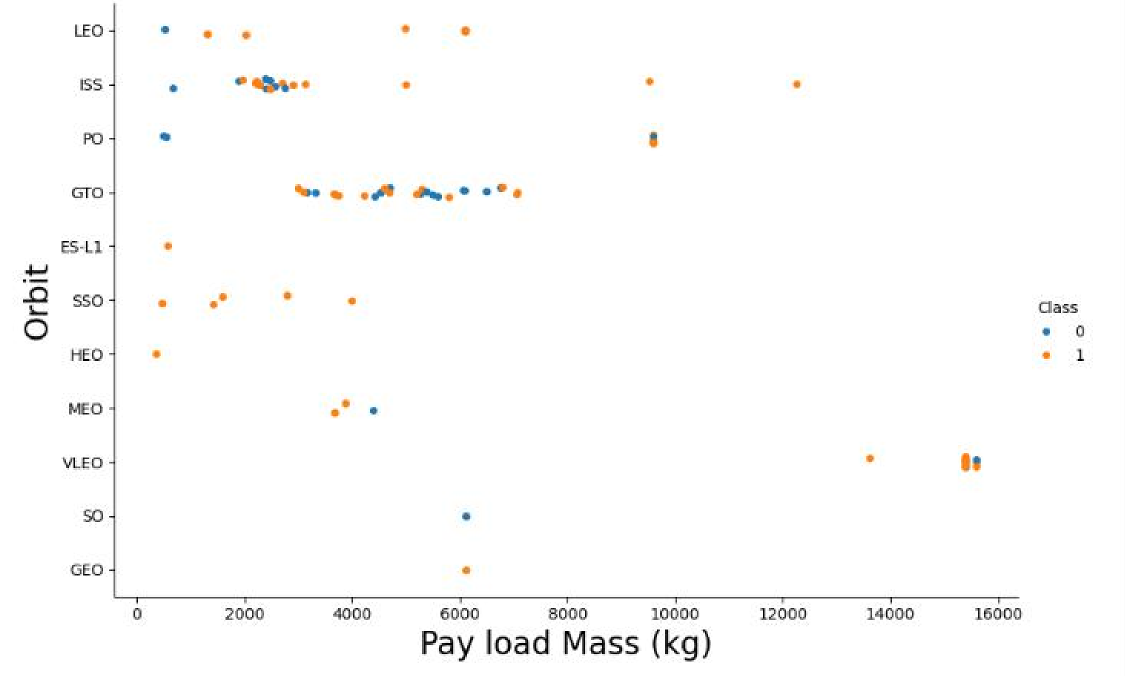
- ES-L1, GEO, HEO, SSO has the same success rate approximately

Flight Number vs. Orbit Type



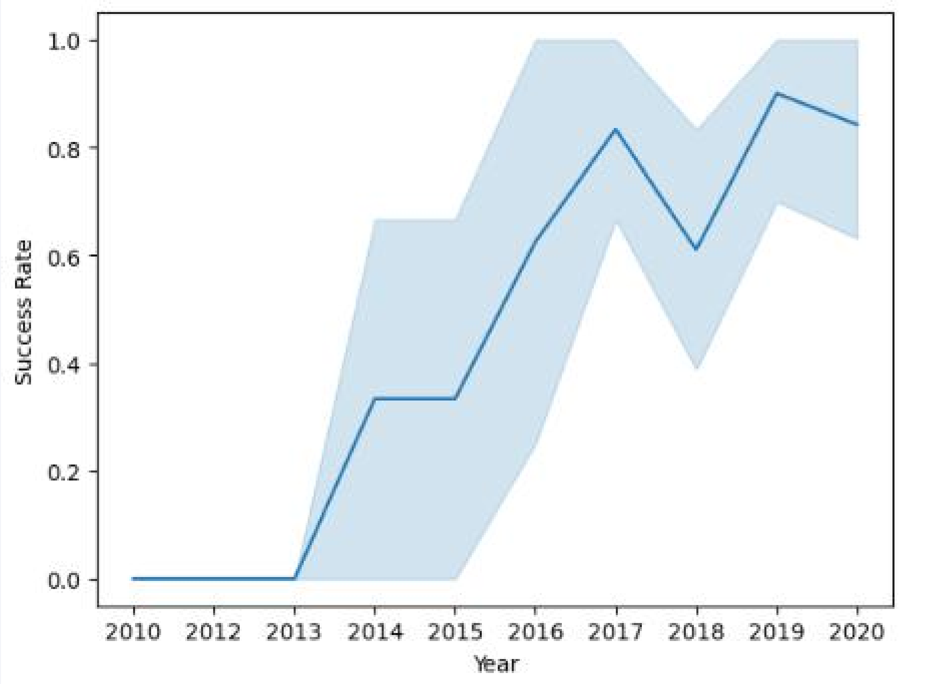
- Success rate improve over time

Payload vs. Orbit Type



- No relation between payload and orbit GTO

Launch Success Yearly Trend



- First three years : no significant improvement

All Launch Site Names

There are four Launch Sites:

Out[54]:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Total payload carried by boosters from NASA

Out[56]:

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

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

Out[57]:

Payload Mass Kgs	Customer	Booster_Version
2928.4	SES	F9 v1.1

First Successful Ground Landing Date

- First successful landing outcome on ground pad

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [58]:
```
%sql SELECT MIN(DATE) FROM 'SPACEXTBL' WHERE "Landing _Outcome" = "Success (ground pad)";
```

```
 * sqlite:///my_data1.db
Done.
```

Out[58]:

MIN(DATE)
None

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT Booster_Version, Payload FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_I
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	Payload

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date,7,4), substr(Date, 4, 2),"Booster_Version", "Launch_Site", Payload, "PAYLOAD_MASS__KG_", "Mission_O
```

```
* sqlite:///my_data1.db
Done.
```

substr(Date,7,4)	substr(Date, 4, 2)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	"Landing _Outcome"

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

Task 10

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.

[64]:

```
%sql SELECT * FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%' AND (Date BETWEEN '04-06-2010' AND '20-03-2017') ORDEI
```

```
 * sqlite:///my_data1.db
Done.
```

[64]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome

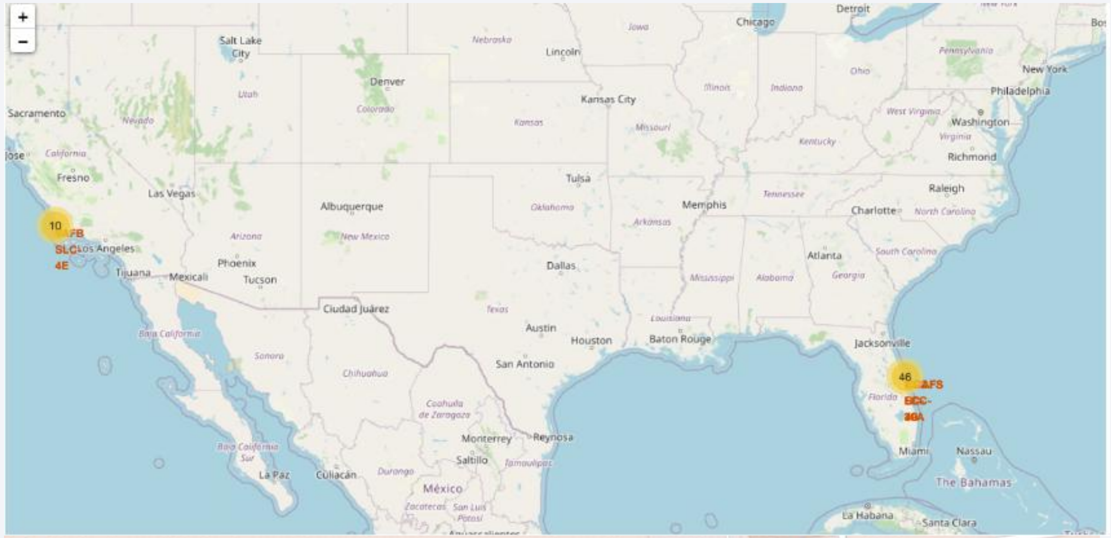
Section 3

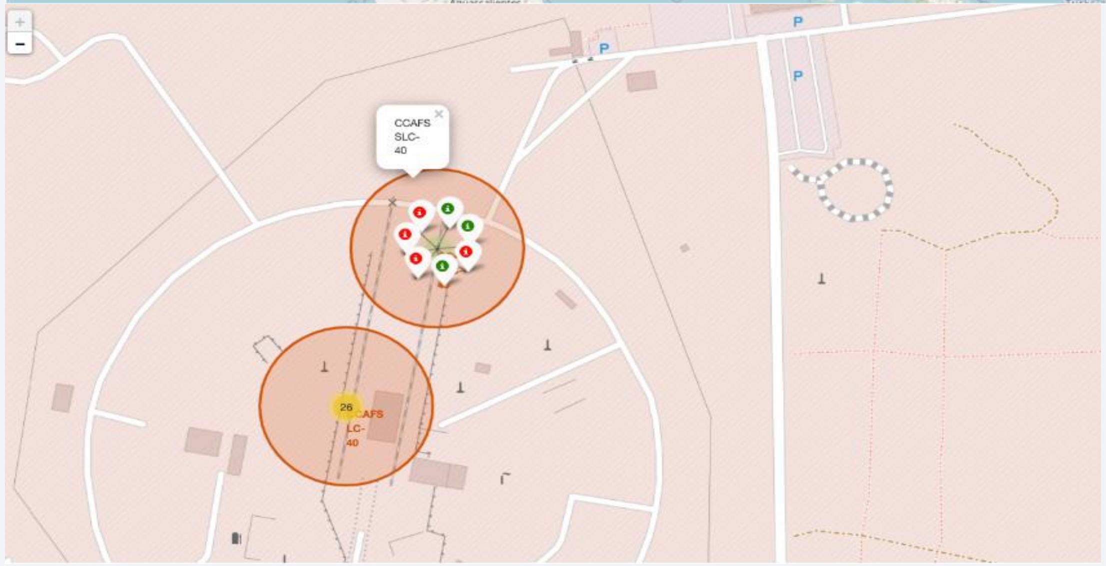
Launch Sites Proximities Analysis

All launch sites

- Launch sites are near sea
- Launch sites are not far from roads and railroads

Success/Failed Launches For Each Site



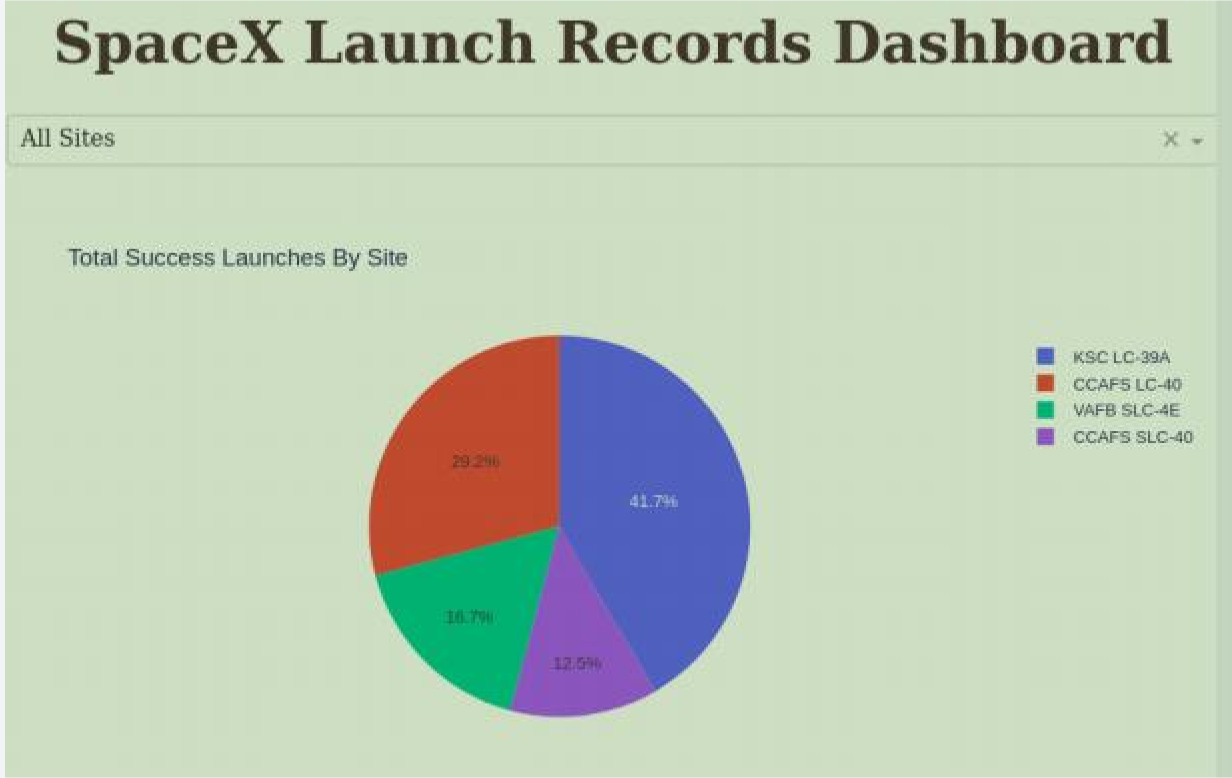


Distances Between a Launch Site to it’s Proximities

Section 4

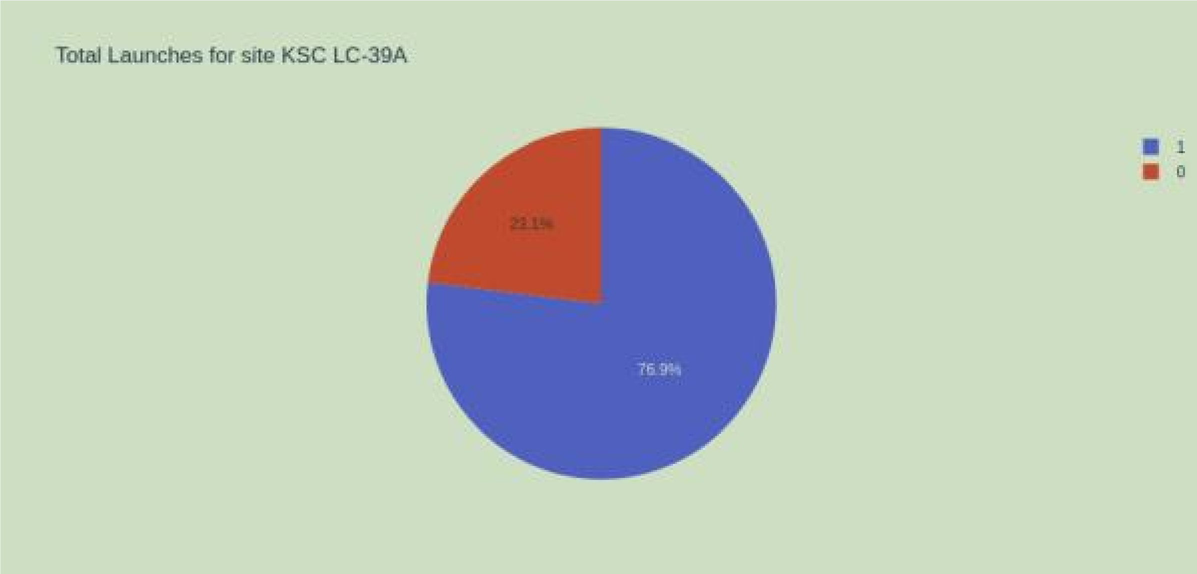
Build a Dashboard with Plotly Dash

Successful Launches by Site



- KSC LC-39A has highest rate of success

Launch Success Ratio for KSC LC-39A



- Around 76.9% launches are successful

Payload vs. Launch Outcome

Section 5

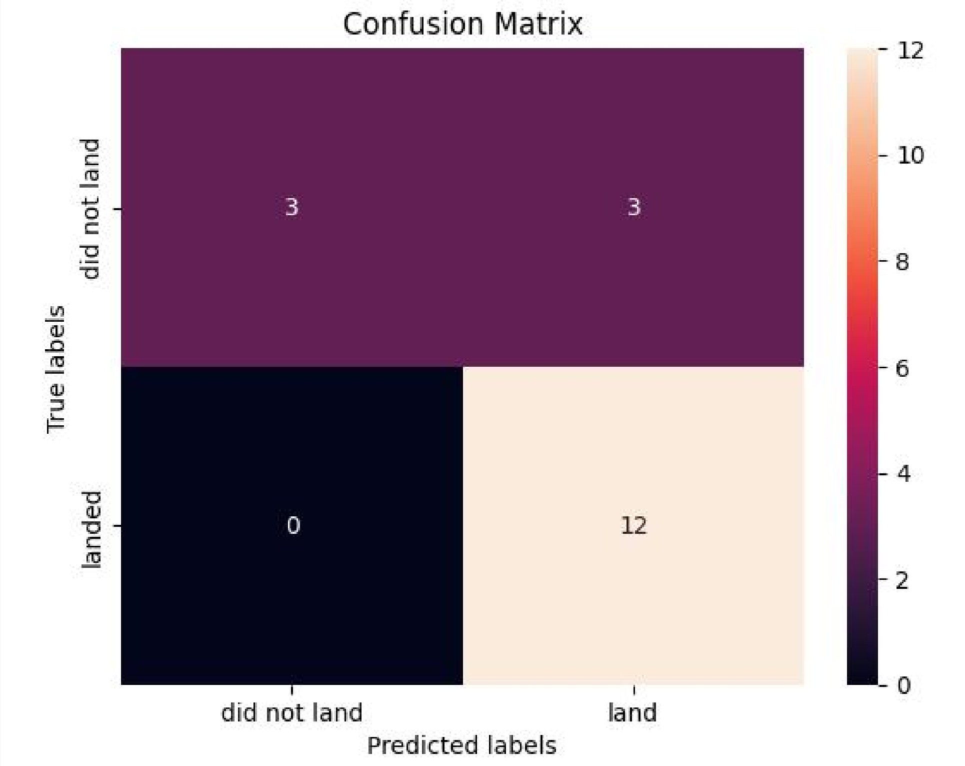
Predictive Analysis (Classification)

Classification Accuracy

	0
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

- It seems all models equally effective for this particular dataset

Confusion Matrix



Conclusions

- Data collected from open sources : SpaceX API, Wikipedia
- KSC LC-39A is the best launch site among four launch sites
- 0-6000 kg launches are risky
- Success rate improves over time among all launch sites
- Lunch Sites are close to sea, not over populated
- All ML algorithms seems to work fine, need more fine tuning to choose perfect model

Appendix

- Python Tutorial : Common Syntex, numpy, matplotlib, pandas

  Github URL: https://github.com/MohidulHaqueTushar/Introduction-of-Python-to-Data-Science.git

Thank you!