Data Science Capstone Project

Lissa Stinson

19-02-2022

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- ▶ Summary of Methodologies
  - ▶ Data Collection
  - ▶ Data Wrangling
  - ▶ Exploratory Data Analysis with SQL
  - ▶ Exploratory Data Analysis with Visualizations
  - ▶ Interactive Visualizations
  - ▶ Predictive Analytics

- ▶ Summary of Results
  - ▶ Exploratory Data Analysis results
  - ▶ Interactive Analytics demos in screenshot
  - ▶ Predictive analysis results

IBM Developer

SKILLS NETWORK

INTRODUCTION

- Project Background
  - The project's aim is to predict if Falcon 9 first stage will land successfully. SpaceX advertises Falcon rocket launches on its website with a cost of 62 million dollars. Other providers cost upwards of 165 million dollars. SpaceX provides the savings by reusing the first stage. If we can determine if the first stage will land, we can determine the cost of the launch. This information can be used by competing companies to bid against SpaceX for a rocket launch.

- Questions to be answered
  - How do variable influences such as payload mass, launch site, and orbits affect the success?
  - What is the effect of each relationship of variables on outcomes?
  - What are the conditions which will aid SpaceX to achieve the best results?

METHODOLOGY

- Data Collection Methodology
    - SpaceX Rest API
    - Web scrapping from Wiki
- Performed Data Wrangling (Transformed data for machine learning)
    - One Hot Encoding data fields for Machine learning and dropping irrelevant columns.
- Performed Exploration Data Analysis with visualization.
    - Plotting: Scatter and bar graphs to show relationship between variables and to show patterns of data.
- Performed Interactive Visual Analytics
    - Using Folium and Plotly Dash visualizations.
- Performed Predictive Analysis using Classification models
    - Build, tune, and evaluate classification models

Data Collection

The data set were gathered.

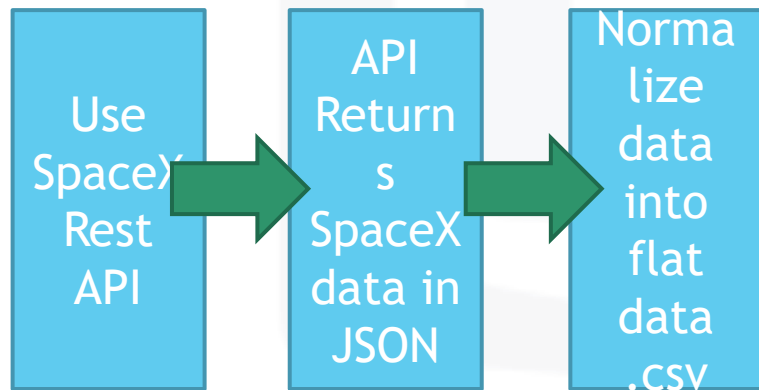- I worked with SpaceX launch data that was gathered from SpaceX REST API.

The API provided data related to launches, including information about the rocket uses, payload delivered, launch specs, landing specs, and landing outcomes.

The goal was to predict whether SpaceX will attempt to land a rockets.
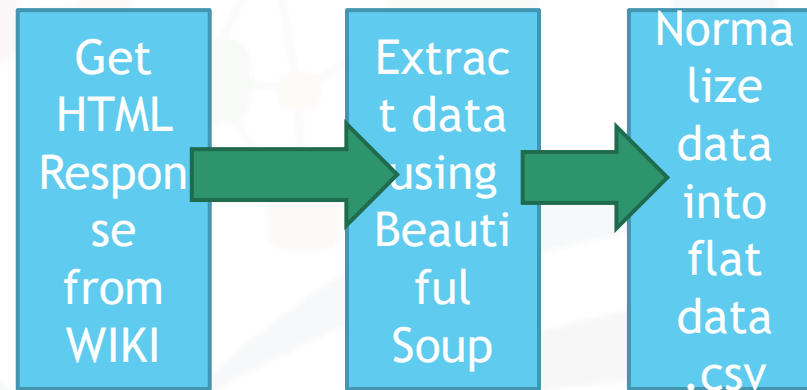
The SpaceX REST API URL starts with api.spacexdata.com/V4/

In addition, the other popular data source for obtaining Falcon 9 launch data is webs-craping Wikipedia using Beautiful Soup.

DATA WRANGLING   WEB SCRAPING

| Use SpaceX Rest API | → | API Returns SpaceX data in JSON | → | Normalize data into flat data .csv | | Get HTML Response from WIKI | → | Extract data using Beautiful Soup | → | Normalize data into flat data .csv |

Data Collection – REST API

Data Collection API – Lab 1
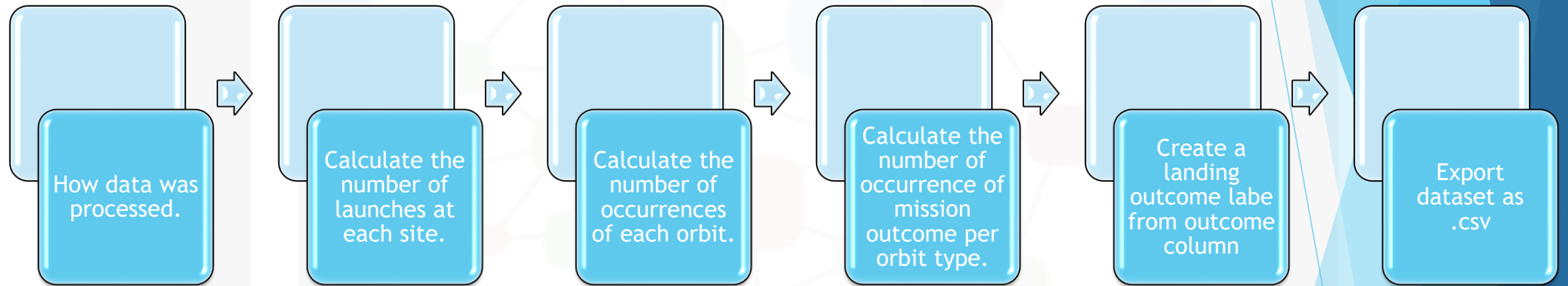
Getting response from API

Converting Response to a .json file.

Apply custom functions to clean data.

Assign a list to dictionary then a data frame.

Filter the dataframe and export to a flat file (.csv)

IBM Developer

SKILLS NETWORK

Data Wrangling

| How data was processed. | → | Calculate the number of launches at each site. | → | Calculate the number of occurrences of each orbit. | → | Calculate the number of occurrence of mission outcome per orbit type. | → | Create a landing outcome labe from outcome column | → | Export dataset as .csv |

https://github.com/nosnitsl/IBM-Data-Science-Capstone.git

Data Wrangling – Lab 2

IBM **Dev**oper

SKILLS NETWORK

EDA with SQL

▶ Performed SQL queries:

▶ Dsplaying the names of the unique launch sites in the space mission.

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

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

▶ Displaying average payload mass carried by booster version F9 v1.1.

▶ Listing the date when the first successful landing outcome in ground pad was achieved.

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

▶ Listing the total number of successful and failure mission outcomes.

▶ Listing the names of the booster versions which have carried the maximum payload mass.

▶ Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015'

▶ 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.

. https://github.com/nosnitsl/IBM-Data-Science-Capstone.git

EDA with SQL – Lab 3

IBM Developer

SKILLS NETWORK

EDA and interactive Visual Methodology

- Charts were plotted.
  - Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload Mass vs Launch Site, Orbit Type vs Success Rate, Flight Number vs Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

  - Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

  - Bar chart show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measure value.

  - Line charts show trends in data over time (time series).

https://github.com/nosnitsl/IBM-Data-Science-Capstone.git

EDA with Visualization – Lab 4

IBM Developer

SKILLS NETWORK

Build an Interactive Map with Folium

https://github.com/nosnitsL/IBM-Data-Science-Capstone.git

Interactive Visualizations – Lab 5

▶ Markers of all Launch Sites:

▶ Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

▶ Add Markers with Circle Popup Label and Text Label of all Launch Sits using their latitude and longitude coordicates to show their geographical locations andproximimity to Equator and coasts.

▶ Coloured markers of the launch outcomes for each Launch Site:

▶ Addec coloured markers of success (Green) and failed (Red) launches using Maker Cluster to identify which launch sites have relatively high success rates.

▶ Distnances between a launch site to its proximities.

IBM Developer

SKILLS NETWORK

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Add a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs Success Rate for the different Booster Versions.

Predictive Analysis Methodology

Building Model

Load the dataset into Numpy and Pandas

Transform data

Split the data into training and test data sets.

Check how many test samples there were.

Decide which type of maching learning aglorithms t ouse.

Set the paramaeters and alogrimiths to GrdSearchCV

Evaluating Model

Check accuracy for each model.

Get tuned hyperparamaters for each type of algorithms

Plot confusion matrix.

Improving Model

Feature Engineering
Algorithm tuning.

Finding the best performing claissfication model

The model with the best accuracy score winds the best performing model

Results

Exploratory data analysis results.
Interactive analytics demo in screenshots.
Predictive analysis results.

EDA with Visualizations

The more flights at a launch site the greater the success rate at a launch site.

VAFB SLC 4E

Flight Site

Flight Number vs. Flight Site

Payload vs Launch Site

The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocker. There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dpependant on Pay Load Mass for a success launch.
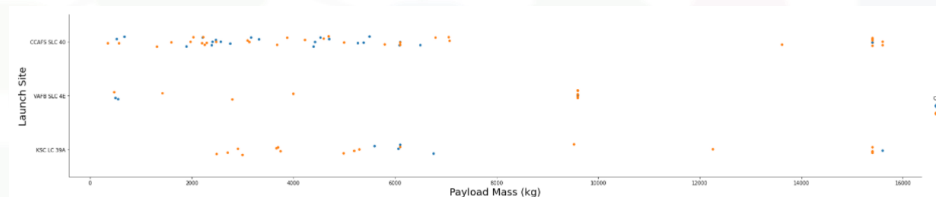
Success Rate vs Orbit Type

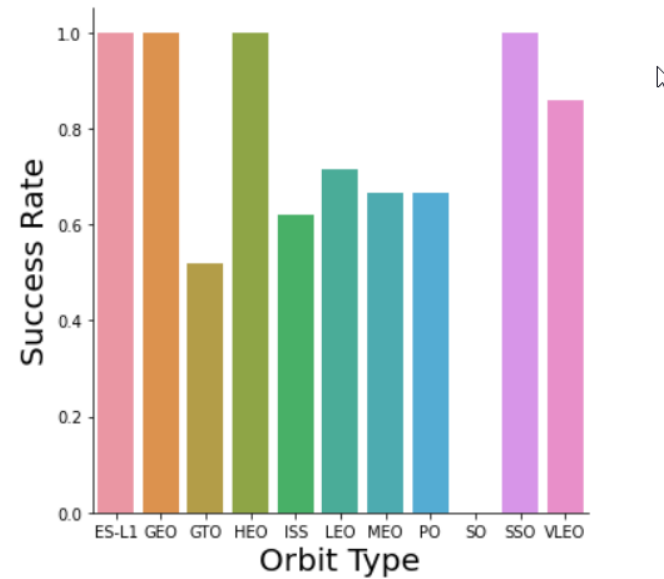Explanation:

Orbits with 100% success rate:
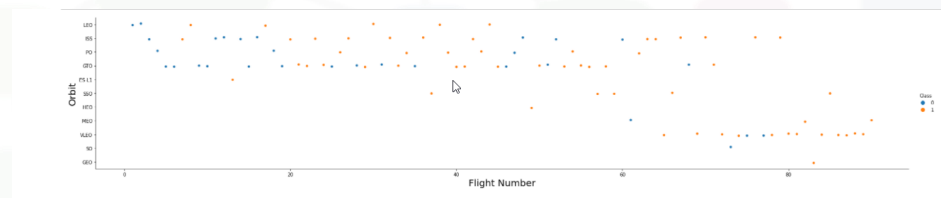
ES-L1, GEO, HEO, SSO

Orbits with 0% success rate:

SO

Orbits with success rate between 50% and 85%

GTO, ISS,LEO,MEO,PO

Flight Number vs Orbit Type

In the visualization you will see that in the LEO orbit the Success appears related to the number of gflights. In addition there seems to be no relaitopnship between flight number when in GTO orbit.

Payload Mass vs Orbit Type

Heavy payloads have a negative influence on
GTO orbits and positive on GTO and Polar LEO
(ISS) orbits.
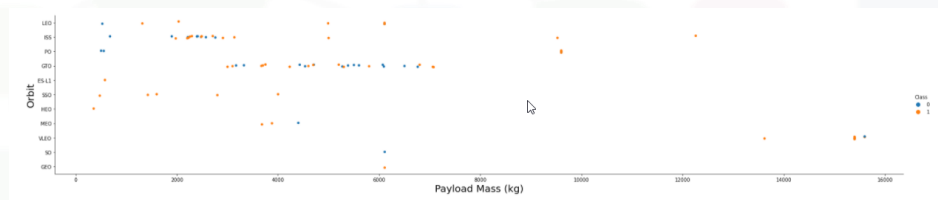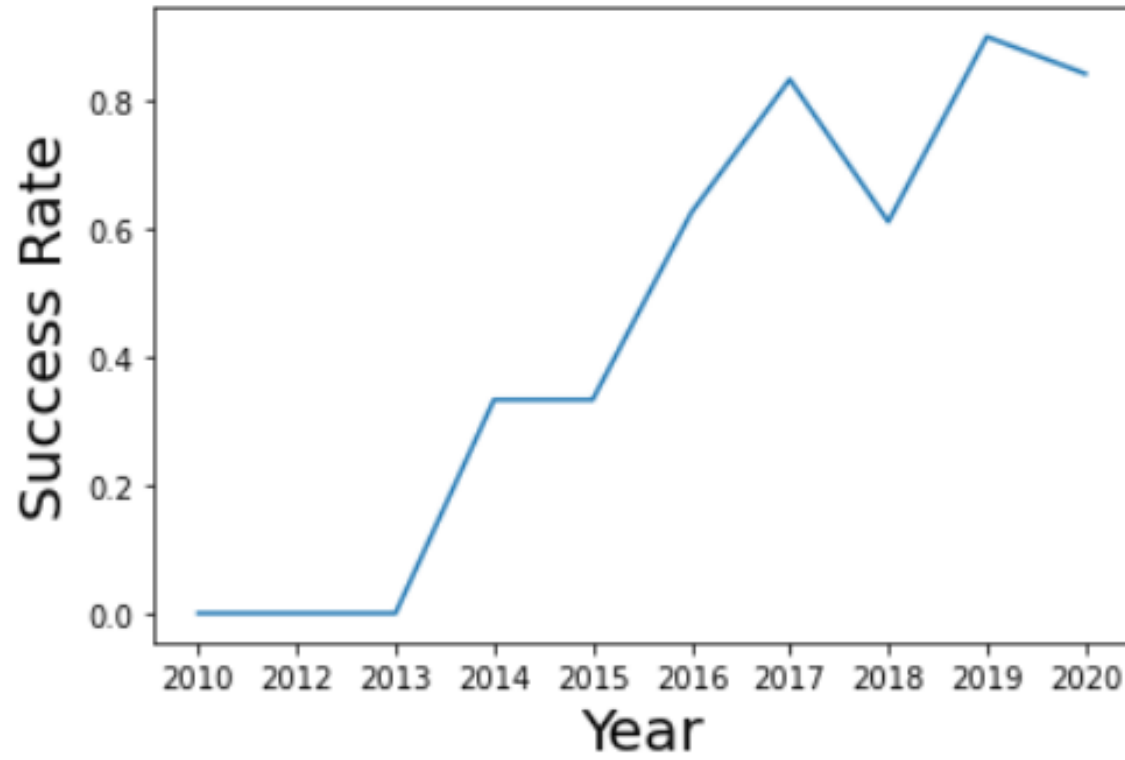
Launch Success Yearly Trend

Explanation:

The success rate since 2013 kept increasing till 2020.

IBM **Developer**

SKILLS NETWORK

EDA with SQL

All Launch Site Names

Descriptions: Using the work distinct in the query will pull the unique values for the launch site column from the table SPACEXTBL.


```
In [16]:  %sql select distinct launch_site from SPACEXTBL;

           * ibm_db_sa://pkn91707:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
          Done.

Out[16]:  launch_site
          CCAFS LC-40
          CCAFS SLC-40
          KSC LC-39A
          VAFB SLC-4E
```

Out[24]:

Date DD-MM-YYYY	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Launch Site Names Begin with CCA

Description: Using keyword "Limit 5" in the query will fetch 5 records from table SPACEX., condition LIKE keyword with wild card "CCA%". The percentage in the end suggest that the launch site name must start with CCA.

IBM Developer

SKILLS NETWORK

```
as total_payload_mass from SPACEXTBL where custome
```

```
1b-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg
```

Total Payload Mass

Description: Using function SUM the total in the PAYLOAD_MASS_KG and the WHERE clause filters the dataset to only perform calculations on customer NASA

Average Payload Mass by F9 v1.1

Description: Using the function AVG works out the average in the column PAYLOAD_MASS_KG.

The WHERE clause filters the dataset to only perform calculations on Booster_Version f9 v1.1.


```
In [26]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL  where booster_version like '%F9 v1.1%';

         * ibm_db_sa://pkn91707:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
         Done.

Out[26]:  average_payload_mass
          3226
```

First successful ground landing date

Description: Listing the date when the first successful landing outcome in ground pad was achieved.

Successful drone ship landing with payload between 4000 and 6000

Listing the name of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

```
In [28]: %sql select booster_version from SPACEXTBL  where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 a
nd 6000;

 * ibm_db_sa://pkn91707:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[28]:

booster_version
F9 FT B1022
F9 FT B1031.2

```
come, count(*) as total_number from SPACEXTBL  group by mission_outcome;
```

```
:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.ap
```

	total_number
	44
clear)	1

Total Number of Successful and Failure mission outcomes.

Explanation: Listing the total number of successful and failure mission outcomes.

```
er_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_)
```

```
n91707:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.app
```

Booters carried maximum payload

Explanation:

Listing the names of the booster versions which have carried the maximum payload mass.

```
onth, date, booster_version, launch_site, landing__o
ailure (drone ship)' and year(date)=2015;
```

`6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg`

unch_site	landing__outcome
CAFS LC-40	Failure (drone ship)
CAFS LC-40	Failure (drone ship)

2015 Launch Records

Explanation:

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the month in year 2015.

# Rank success count between 2016-06-04 and 2017-03-20

Explanation:

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 desc order.

```
In [13]:  %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
              where date between '2010-06-04' and '2017-03-20'
              group by landing__outcome
              order by count_outcomes desc;
```

  * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.

Out[13]:

landing__outcome	count_outcomes
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

IBM Developer

SKILLS NETWORK

Interactive Map with Folium
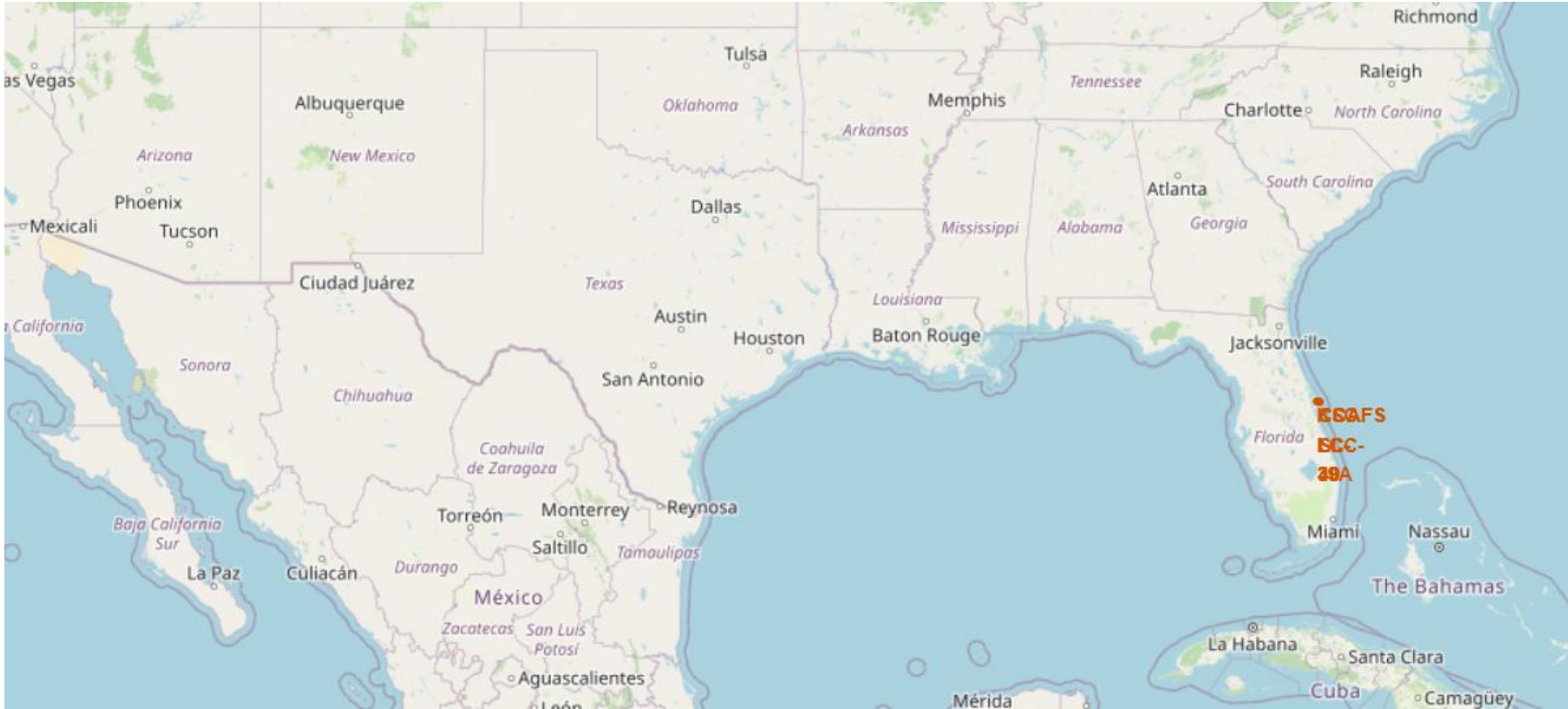
All launch sites location markers on a global map.

Explanation

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 Earch.

All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it mimizes the risk of hanving any debris dropping or exploding near people.

IBM Developer

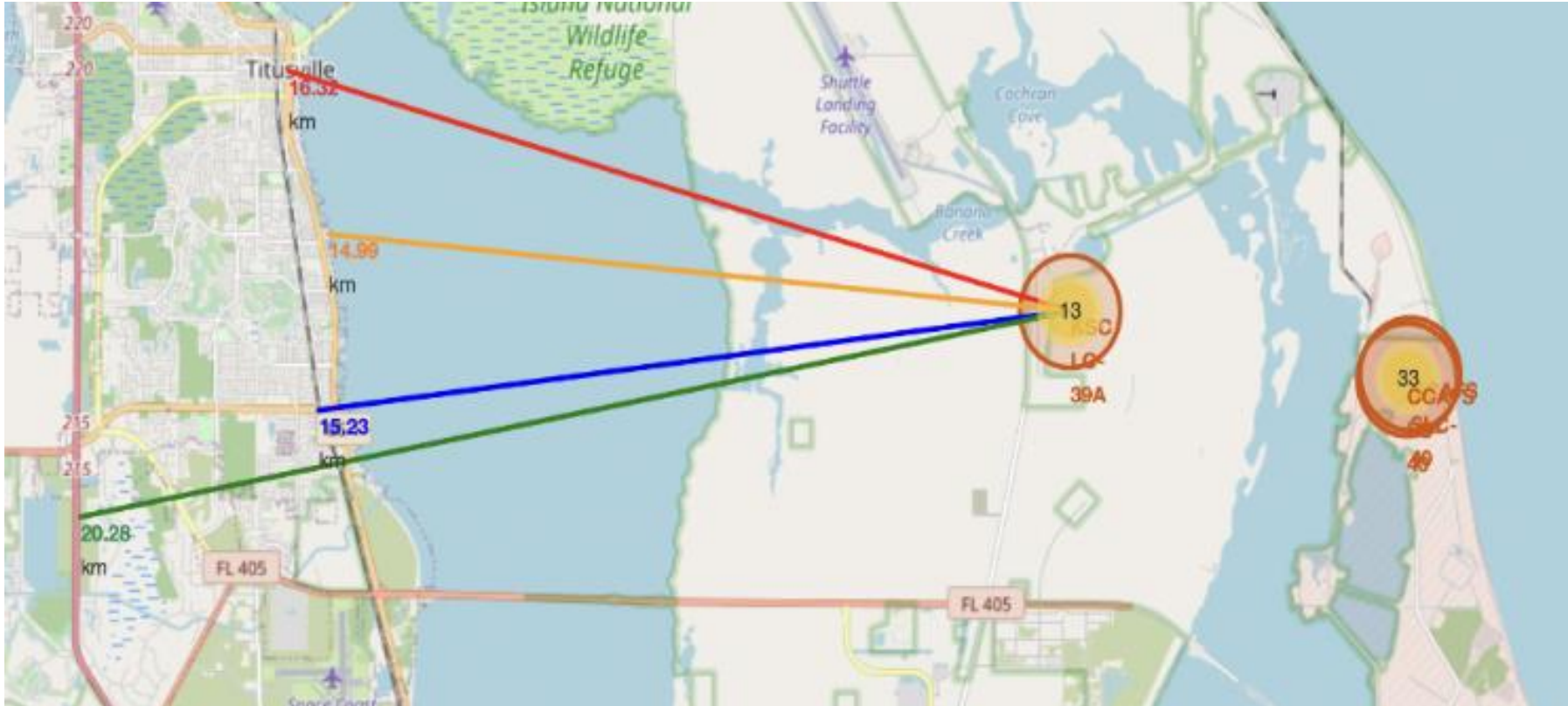**SKILLS NETWORK**

Colour labeled launch records on the map

Xplanation:

From the colour labeled markers we should be able to easily identify which launch sites have relatively high success rates.

Green maker + cuccessful launch

Red marker + failed launch

Launch site KSC LC-39A has a very high success rate.

Distance from the launch site

Explanation:

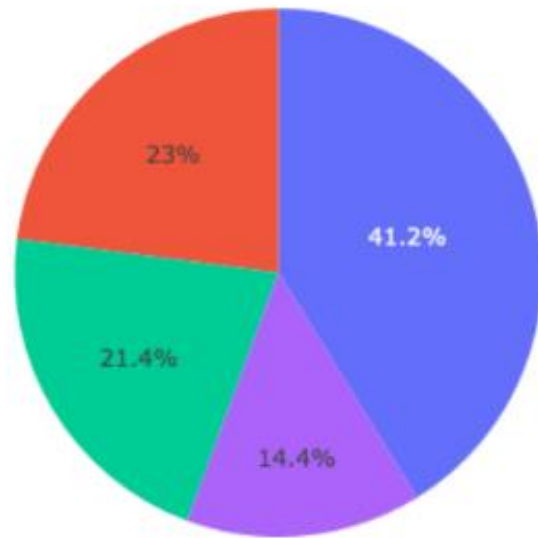From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

- Relative clost to railway
- Relative close to highway
- Relative close to coastline

In other launch site the launch is relative close to its closet city Titusville (16.32 km).

A failed rocket with the maximum reached cross distances like 15.23 km, 14.99 km, 16.32 km, wouldn't than dangerous to populated areas.
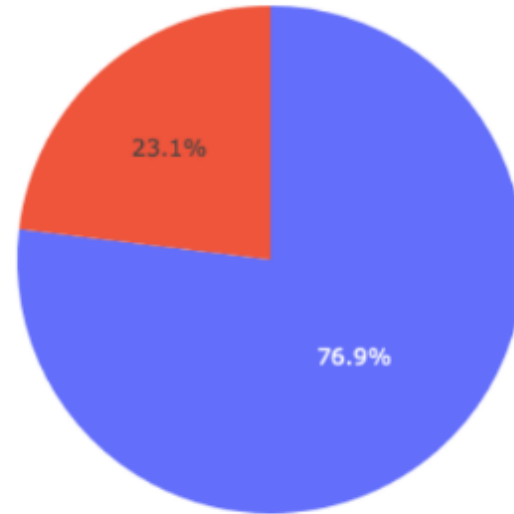
Build a Dashboard with Plotly Dash

Launch Success Count for all Sites

Explanation:

The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

IBM Developer

SKILLS NETWORK

23.1%

76.9%

Launch site with highest launch success ration

Explanation:

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

IBM Developer

SKILLS NETWORK

Payload Mass vs Launch Outcome for all Sites

Explanation:

The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Predictive Analysis (Classification)

Classification Accuracy

▶ Explanation:

  ▶ Based on the score of the Test set, we can not confirm which method performs best.

  ▶ Same Test Set scores may be due to the small test sample size (18 samples) , Therefore, we tested all methods based on the whole dataset.

  ▶ The scores of the whole Dataset confirm that the best model is the Decision Tree model. This model has not only higher scores, but also the hightes accuracy.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation

Examing the confusion matrix, we see that logistic regression can distinguish between the different clasees, WE see that the major problem is false positivies.

Confusion Matrix
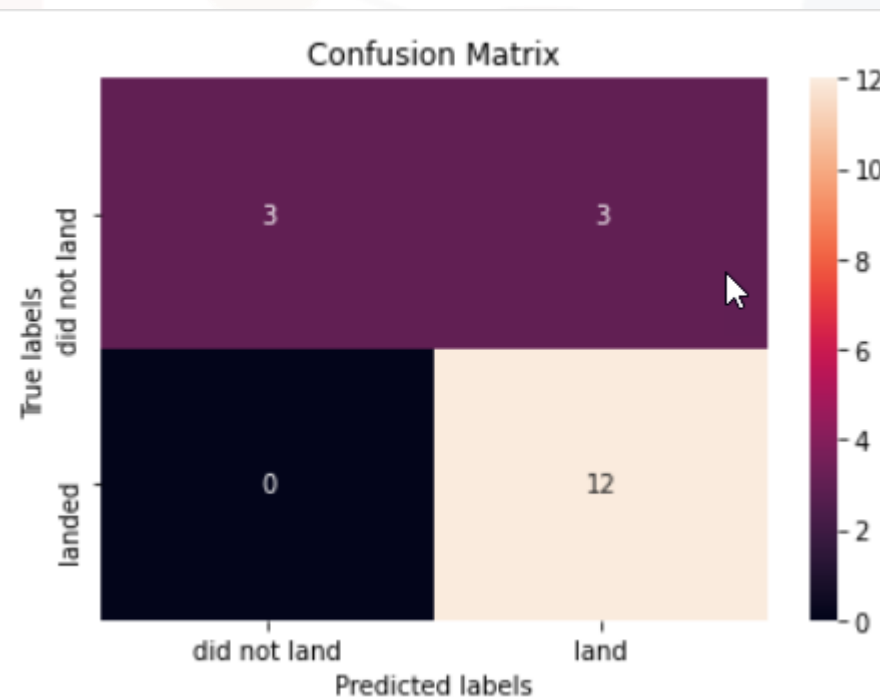
CONCLUSION

▶ The decision tree model is the best algorithm for this dataset.

▶ The low weigheted payloads perform bettere han heavier payloads.

▶ Success rates for SpaceX launches are directly proportional to time and eventually launches will be perfected

▶ KSC LC-39A had the most successful launches

▶ Orbits ES-L1, GEO,HEO, and SSO have 100% success rates

APPENDIX

- IBM Coursera