## DATA SCIENCE CAPSTONE PROJECT

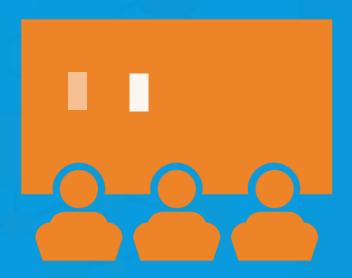


Lissa Stinson
February 19, 2021



## 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





#### 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

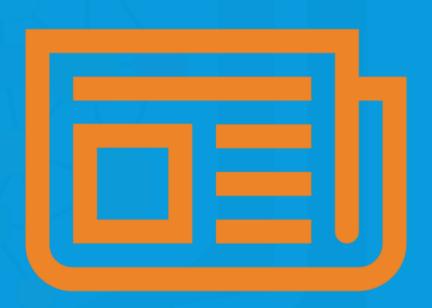
- How do variable influences such as payload mass, launch site, and orbits affect the success?
- What is the effect of each relationship of variables on outcomes?
- What are the conditions which will aid SpaceX to achieve the best results?





## **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 sets were gathered.

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

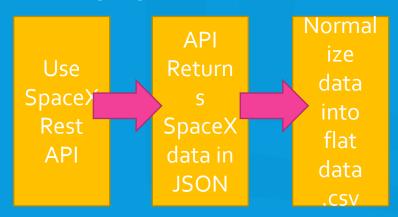
The API provided data related to launches, including information about the rocket uses, payload delivered, launch specs, landing specs, and landing outcomes.

The goal was to predict whether SpaceX will attempt to land a rockets.

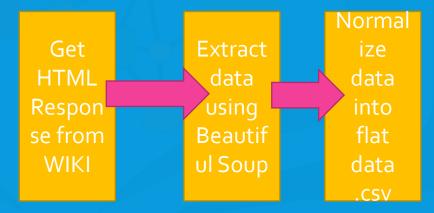
The SpaceX REST API URL starts with api.spacexdata.com/V4/

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

#### **DATA WRANGLING**



#### **WEB SCRAPING**







## DATA COLLECTION – REST API

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)

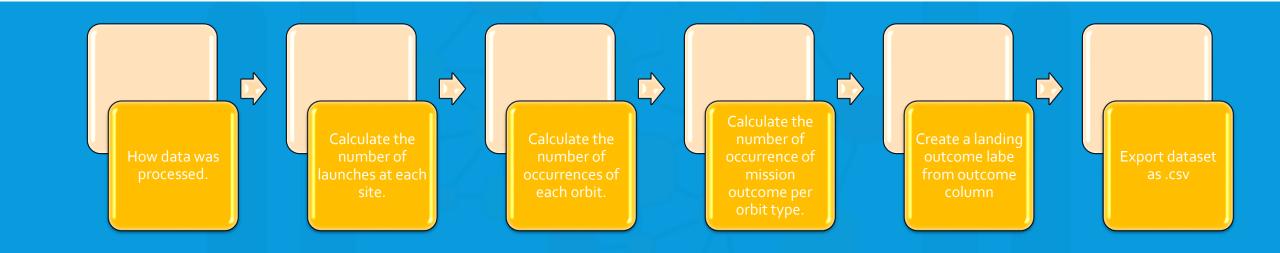
GitHub Lab

Data Collection API - Lab 1





### **DATA WRANGLING**



GitHub Lab

Data Wrangling – Lab 2





### **EDA WITH SQL**

- Performed SQL queries:
- Displaying 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.

#### GitHub Lab

EDA with SQL – Lab 3





#### 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).

GitHub Lab

EDA with Visualization — Lab 4





# BUILD AN INTERACTIVE MAP WITH FOLIUM

#### Git Hub Lab

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 coordinates to show their geographical locations and proximity to Equator and coasts.
  - Coloured markers of the launch outcomes for each Launch Site:
    - Added coloured markers of success (Green) and failed (Red) launches using Maker Cluster to identify which launch sites have relatively high success rates.
  - Distances between a launch site to its proximities.





#### 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

Check how many test samples there were.

data sets.

Decide which type of matching learning algorithms to use.

Set the parameters and algorithms to GrdSearchCV



#### **Evaluating Model**

Check accuracy for each model.

Get tuned hyperparameters for each type of algorithms

Plot confusion matrix.



#### **Improving Model**

Feature Engineering Algorithm tuning.



#### Finding the best performing classification model

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





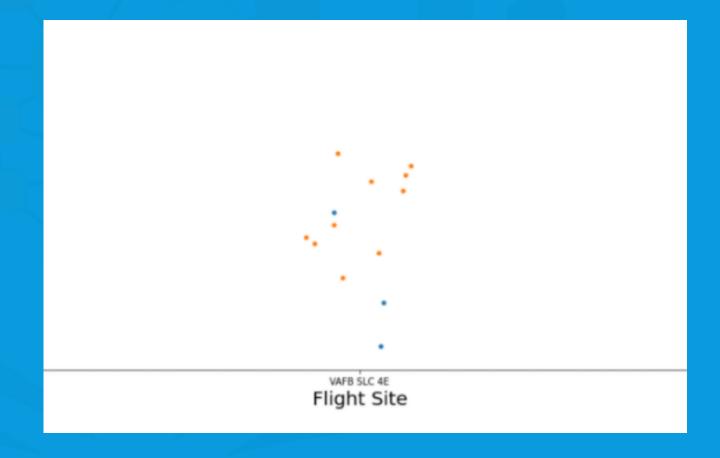
### RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS.
PREDICTIVE ANALYSIS RESULTS.

## EDA WITH VISUALIZATIONS

## FLIGHT NUMBER VS. FLIGHT SITE

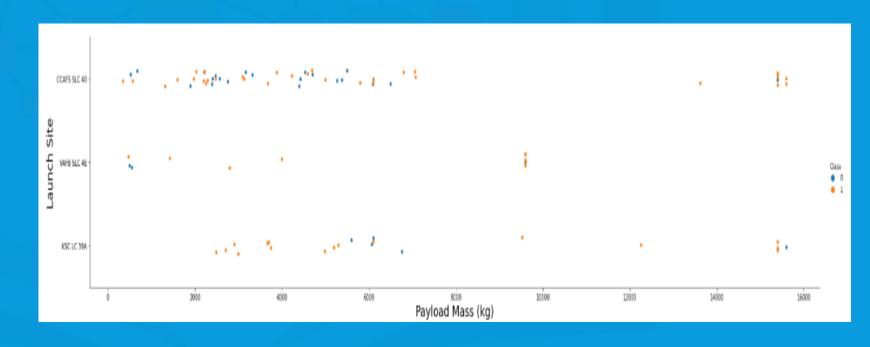
The more flights at a launch site the greater the success rate at a launch site.





## PAYLOAD VS LAUNCH SITE

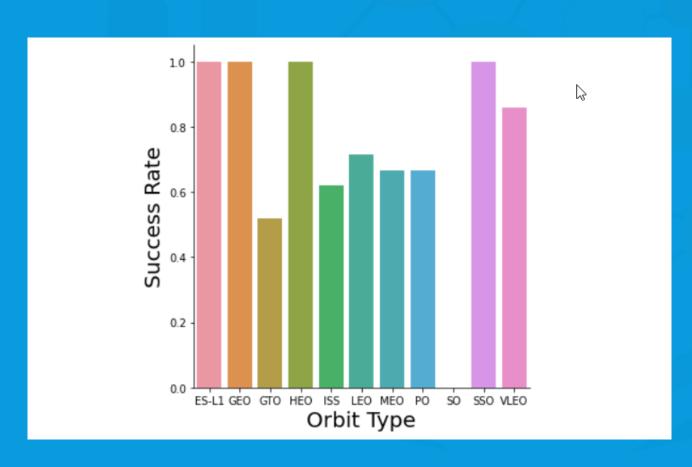
The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocker. There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependent on Pay Load Mass for a success launch.







### SUCCESS RATE VS ORBIT TYPE



#### Explanation:

Orbits with 100% success rate:

ES-L1, GEO, HEO, SSO

Orbits with o% success rate:

SO

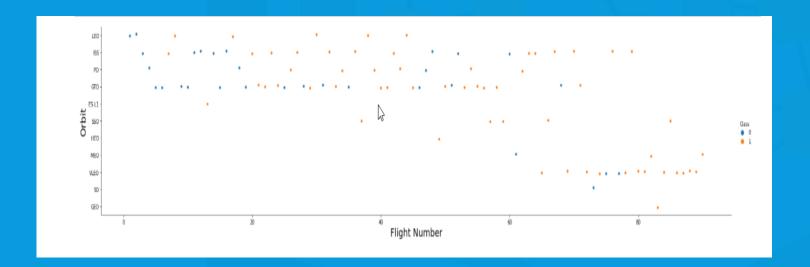
Orbits with success rate between 50% and 85%

GTO, ISS, LEO, MEO, PO



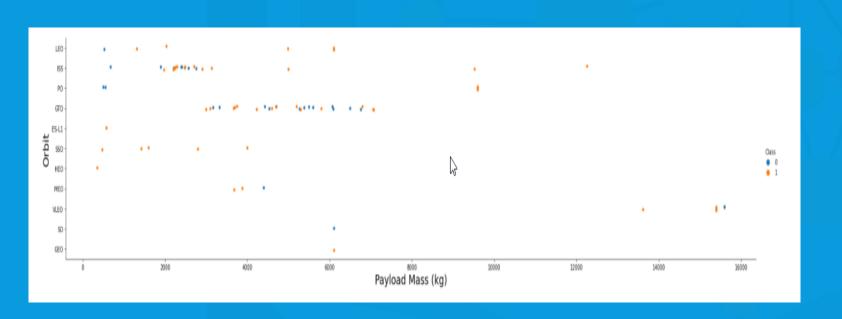


### FLIGHT NUMBER VS ORBITTYPE



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

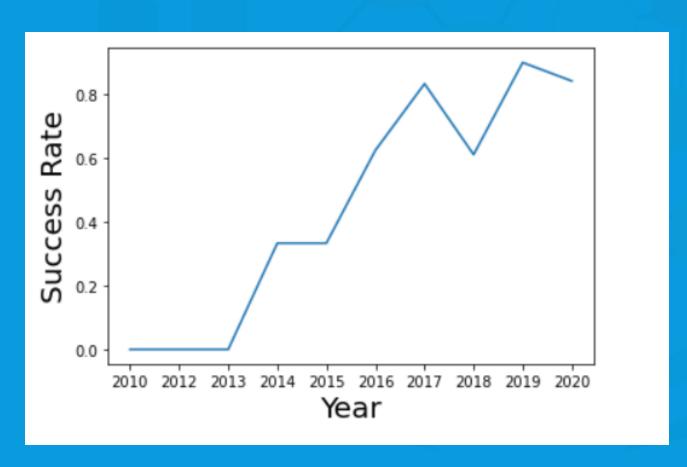
### PAYLOAD MASS VS ORBIT TYPE



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



### LAUNCH SUCCESS YEARLY TREND



The success rate since 2013 kept increasing till 2020.

IBM **Developer** 



## EDA WITH SQL

#### ALL LAUNCH SITE NAMES

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

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



### LAUNCH SITE NAMES BEGIN WITH CCA

| In [24]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;  * ibm_db_sa://pkn91707:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb Done.  Out[24]: Date DD- MM-YYYY time_utc_ booster_version launch_site payload payload_mass_kg_ orbit customer mission_outcome landing  2010-04-06 18:45:00 F9 v1.0 B0003 CCAFS LC- 40 Dragon Spacecraft Qualification Unit 0 LEO SpaceX Success Failure | 3_outcome     |
|---|---------------|
| Done.  Out [24]:  Date DD-  MM-YYYY  time_utc_ booster_version launch_site  payload payload_mass_kg_ orbit customer mission_outcome landing  2010-04-06 18:45:00 F9 v1.0 80003 CCAFS LC-  Dragon Spacecraft Qualification Unit 0 LFO SpaceX Success Failure   | g_outcome     |
| MM-YYYY time_utc_ booster_version launcn_site payload payload_mass_kg_ orbit customer mission_outcome landing   | g_outcome     |
| 2010-04-06 18:45:00 E9 v1.0 80003 Dragon Spacecraft Qualification Unit 0 LEO SpaceX Success Failure   |               |
|   | (parachute)   |
| 2010-08-12 15:43:00 F9 v1.0 B0004 CCAFS LC- Dragon demo flight C1, two CubeSats, 0 LEO NASA (COTS) Success Failure  | e (parachute) |
| 2012-08-10 00:35:00 F9 v1.0 B0006 CCAFS LC- SpaceX CRS-1 500 LEO NASA (CRS) Success   | No attempt    |
| 2013-01-03 15:10:00 F9 v1.0 B0007 CCAFS LC-<br>40 SpaceX CRS-2 677 LEO (ISS) NASA (CRS) Success   | No attempt    |
| 2013-03-12 22:41:00 F9 v1.1 CCAFS LC-<br>40 SES-8 3170 GTO SES Success  | No attempt    |

"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.





#### TOTAL PAYLOAD MASS

Using function SUM the total in the PAYLOAD\_MASS\_KG and the WHERE clause filters the dataset to only perform calculations on customer NASA

```
In [6]: %scx 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
```





## AVERAGE PAYLOAD MASS BY 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

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.



#### FIRST SUCCESSFUL GROUND LANDING DATE

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

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing_outcome = 'Success (ground pad)';

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb Done.

Out[8]: first_successful_landing

2015-12-22
```





## 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.

## TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES.

Listing the total number of successful and failure mission outcomes.

In [10]: %sql select mission\_outcome, count(\*) as total\_number from SPACEXDATASET group by mission\_outcome;

\* ibm\_db\_sa://wzf08322:\*\*\*@0c77d6f2-5d{\2-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb Done.

Out[10]:

| mission_outcome                  | total_number |  |
|----------------------------------|--------------|--|
| Failure (in flight)              | 1            |  |
| Success                          | 99           |  |
| Success (payload status unclear) | 1            |  |

#### **BOOTERS CARRIED MAXIMUM PAYLOAD**

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

```
In [30]: %sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

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

Out[30]:

#### booster\_version

F9 B5 B1048.4

F9 B5 B1049.4

F9 B5 B1049.5

F9 B5 B1060.2

F9 B5 B1058.3

## 2015 LAUNCH RECORDS

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the month 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.bs2io90108kqb1od8lcg.datagases.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) |





#### RANK SUCCESS COUNT BETWEEN 2016-06-04 AND 2017-03-20

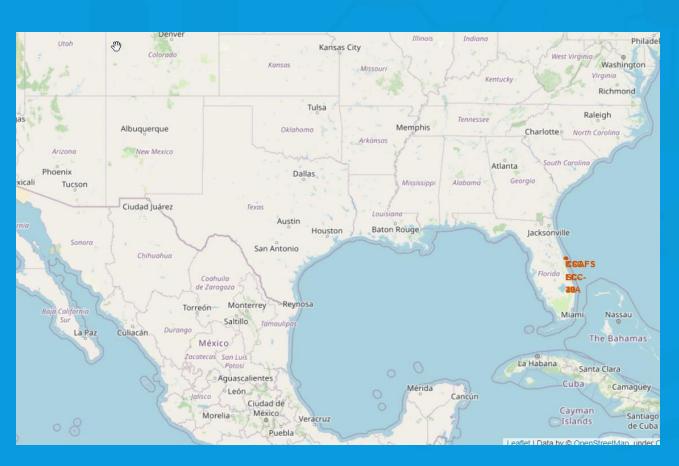
Ranking the count of landing outcomes such as Failure = drone ship or Success - ground pad between the date 2010-06-04 and 2017-03-20 in desc order.





## 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 Earth.

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

# COLOUR LABELED LAUNCH RECORDS ON THE MAP



#### **Explanation:**

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

Green maker + successful launch

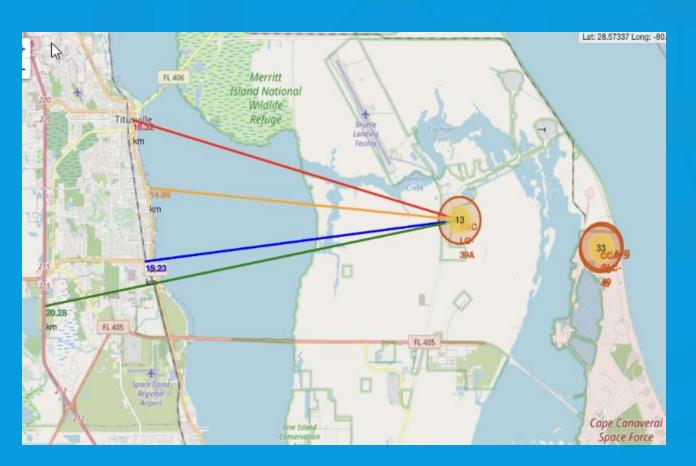
Red marker + failed launch

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 closest to railway
Relative closest to highway
Relative closest to coastline

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

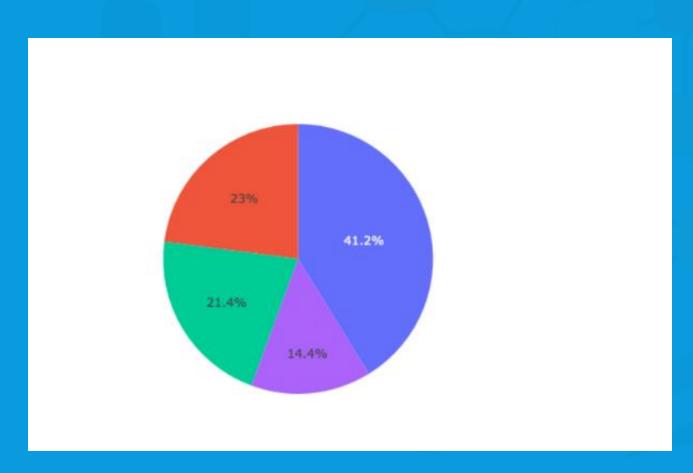
Failed rocket with its high speed can cover distances like 15-20 km in few seconds. IT could potentially dangerous to populated areas.





## BUILD A DASHBOARD WITH PLOTLY DASH

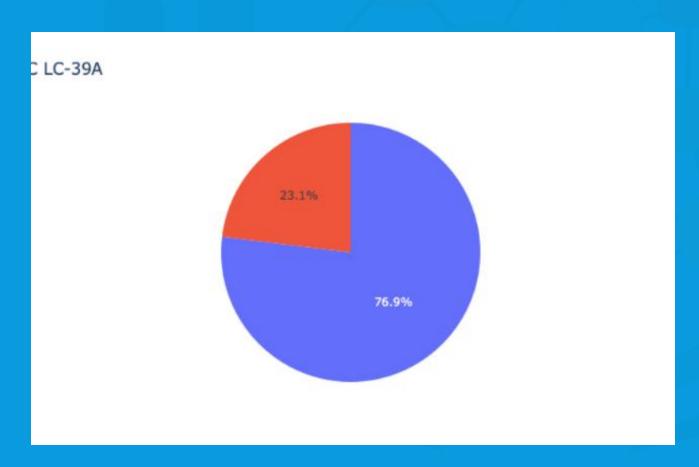
## LAUNCH SUCCESS COUNT FOR ALL SITES



#### **Explanation:**

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

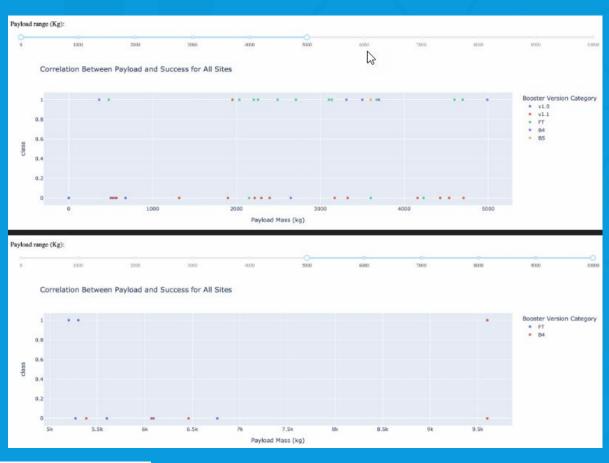
#### LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATION



#### **Explanation:**

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

# 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 highest accuracy.

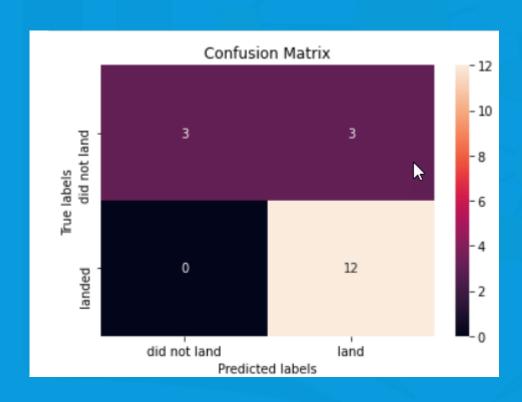
|               | LogReg   | SVM      | Tree     | KNN      |
|---------------|----------|----------|----------|----------|
| Jaccard_Score | 0.800000 | 0.800000 | 0.800000 | 0.800000 |
| F1_Score      | 0.888889 | 0.888889 | 0.888889 | 0.888889 |
| Accuracy      | 0.833333 | 0.833333 | 0.833333 | 0.833333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

|               | 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

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes, WE see that the major problem is false positives.





#### CONCLUSION



- The decision tree model is the best algorithm for this dataset.
- The low weighted payloads perform better than 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

