



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Prince Boateng

[https://github.com/prensboateng/IBM\\_Data\\_Science\\_trepo](https://github.com/prensboateng/IBM_Data_Science_trepo)



# Outline

---

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

# Executive Summary

---

- The following methodologies were used to analyze data:
  - Data Collection using web scraping and SpaceX API;
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
  - Machine Learning Prediction.
- Summary of all results
  - It was possible to collect valuable data from public sources;
  - EDA allowed to identify which features are the best to predict success of launchings;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

# Introduction

---

- The objective is to evaluate the viability of the new company Space Y to compete with Space X.
- 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.

Section 1

# 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\\_Falcon/\\_9/\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_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) using visualization and SQL

# Methodology

---

## Executive Summary

- 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

---

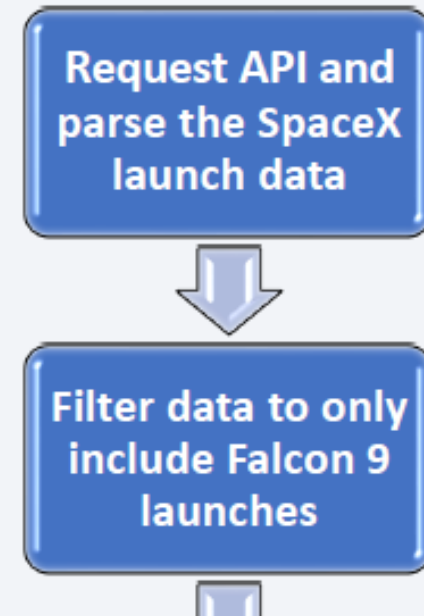
- Data sets were collected from Space X API (<https://api.spacexdata.com/v4/rockets/>) and from Wikipedia ([https://en.wikipedia.org/wiki/List\\_of\\_Falcon\\_9\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)), using web scraping technics.



# Data Collection – SpaceX API

---

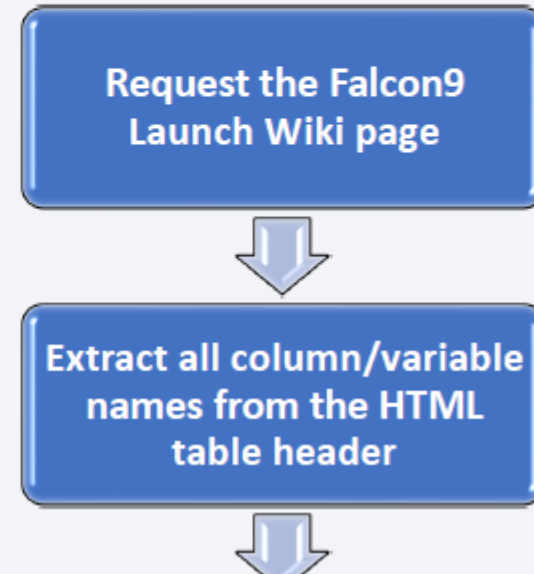
- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.



## Data Collection - Scraping

---

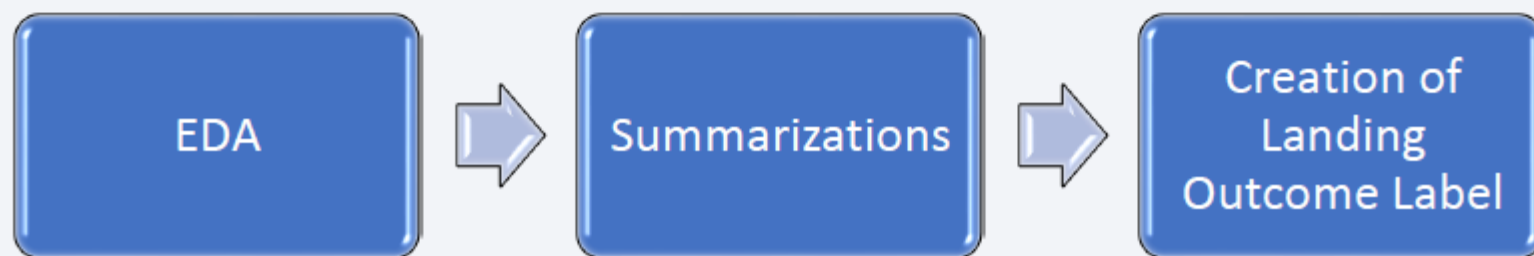
- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.



# 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.

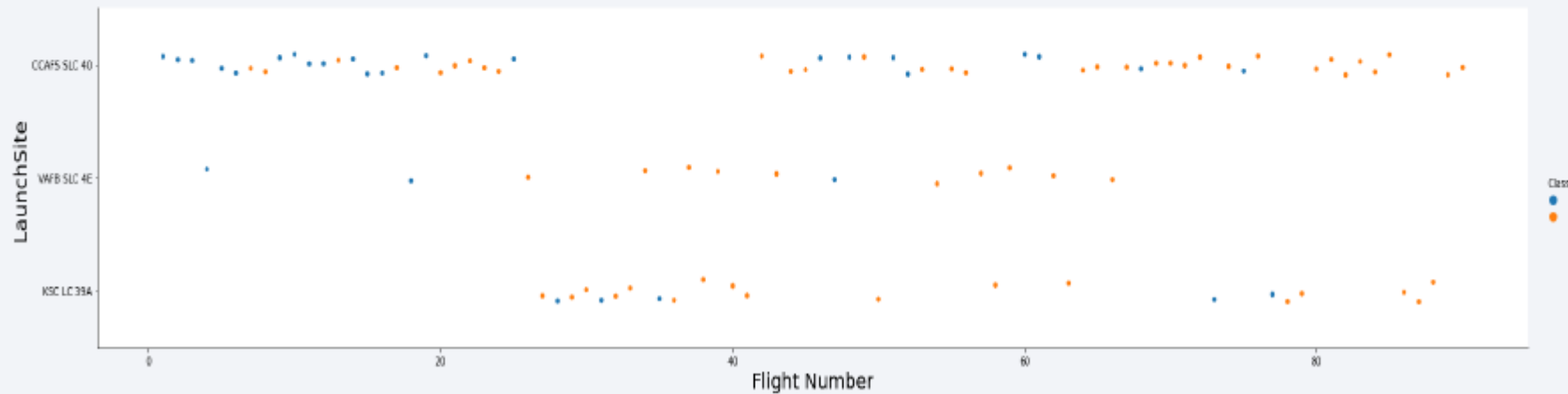




# EDA with Data Visualization

---

- 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 in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

# EDA with SQL

---

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 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

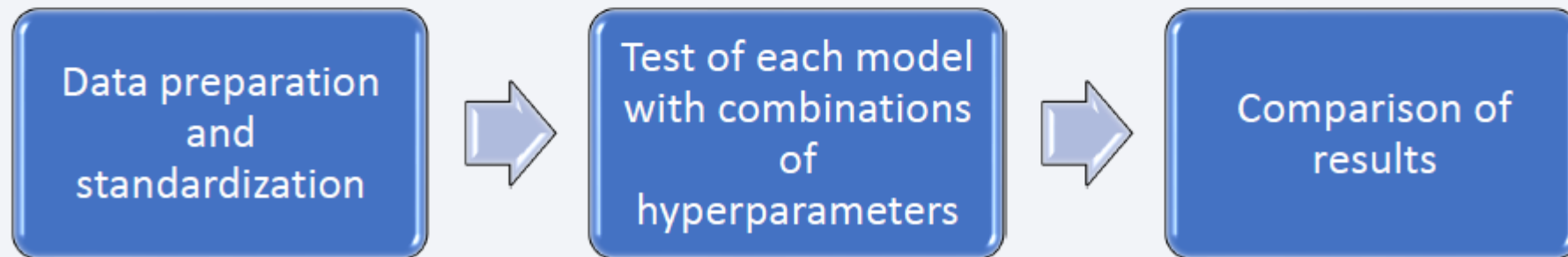
---

- 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.



# Conclusions

---

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

# END OF PRESENTATION

Thank you!

