IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zander Rana

06/01/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Data Collection
- Data Wrangling
- EDA For Data Understanding and Preprocessing
- Visual Analytics for Data Understanding
- Predictive Analysis
- Conclusion

Executive Summary

- Methodologies:
  - **Business Understanding**: Finding out why some SpaceX launches fail, and others succeed.
  - **Analytic Approach**: We will use data from SpaceX involving rocket launches and determine whether a given launch succeeds or not .
  - **Data Requirement**: We believe collecting data about previous launches and seeing features that affect whether the launch succeeds is necessary for our problem.
  - **Data Collection**: We will collect information from the SpaceX via API about the past launches that took place and the data describing the launches such as payload, location, booster, orbit, etc..
  - **Data Understanding**: Using descriptive statistics and visualization to see if our SpaceX data is valid for context.
  - **Data Preprocessing**: Preprocessing and cleaning raw data via data wrangling.
  - **Modeling**: Identifying and using a classification model to predict whether a launch will succeed based on its feature data.
  - **Evaluation**: Testing the efficiency of the model.
  - **Deploy**: Generate report and deploy.
  - **Feedback**: Get any feedback and insight from product to make further improvements.
- Based on our results, we identified features that heavily impacted launch's results. These results will be explained down the line.

Introduction

- SpaceX has advertised that their Falcon 9 rocket launches cost a total of 62 million dollars.
  - This is cheaper compared to other providers cost whose cost are around 165 million dollars each
- Much of the savings is because SpaceX can reuse the first stage.
- We must determine whether these Falcon 9 launches succeed in order to identify the costs of these launches, advocating for more efficient rocket launches in the future.

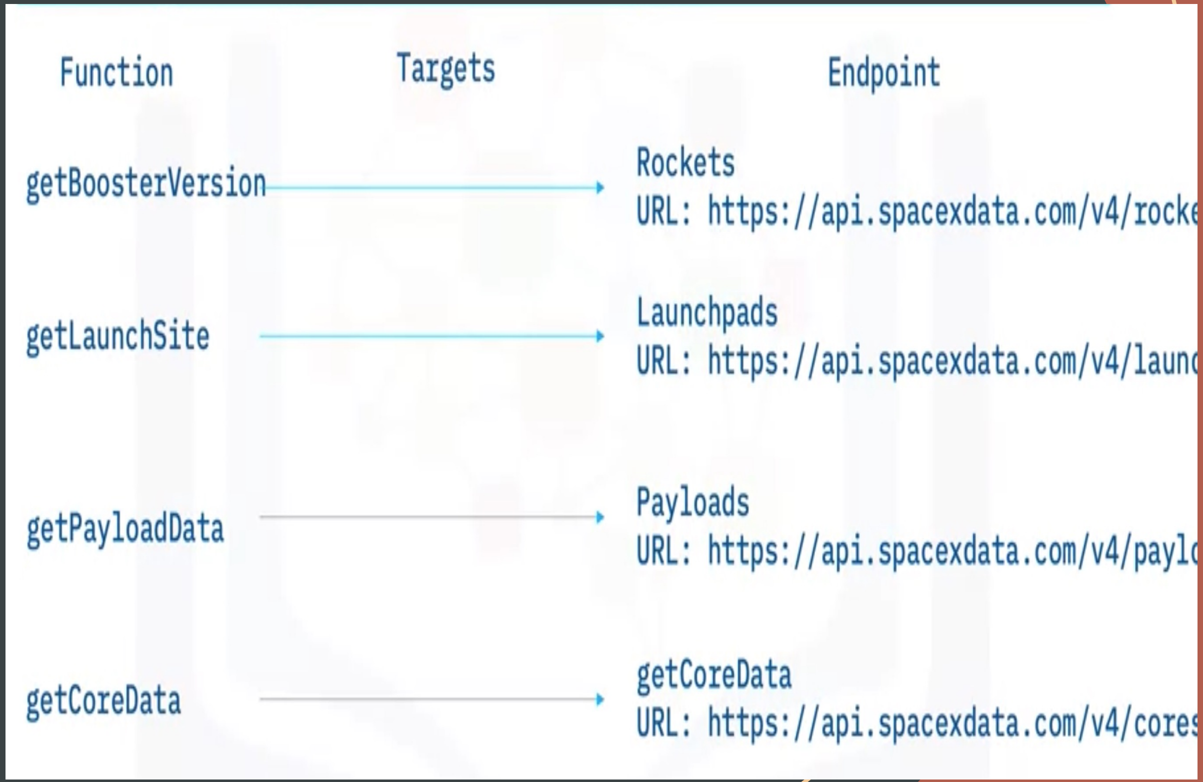
Methodology

As stated before, in the Executive Summary, the main methodologies involving data binning for these problem are:

- **Data Collection**: We will collect information from the SpaceX via API about the past launches that took place and the data describing the launches such as payload, location, booster, orbit, etc..
- **Data Understanding**: Using EDA, that is descriptive statistics and visualization, to see if our SpaceX data is valid for context.
  - Using visual analytics for the data suing Folium and Plotly/Dash.
- **Data Preprocessing**: Preprocessing and cleaning raw data via Data Wrangling.
- **Modeling**: Identifying and using a classification model to predict whether a launch will succeed based on its feature data.
- **Evaluation**: Testing the efficiency of the model.

DATA COLLECTION

- We collected data of past rocket launches from the SpaceX using the SpaceX REST API.
- We issued a HTTP GET request of the SpaceX URL in JSON format and split the content in order to extract data from each of the features necessary for the project.
- The features that we collected are:
  - The booster type of the rocket launch.
  - The launch site and its location.
  - The payload.
  - The orbit that the rockets aim towards. This is a big factor since different orbits vary in distances from the Earth's atmosphere and vary in orbit.
  - Core components including the number of flights, grid fins, landing pad, and most importantly, the outcome.



Following Functions Used To Issue GET Request And Extract Feature Data

- getBoosterVersion() does request and extracts data for the Booster variable.
- getLaunchSite() does request and extracts data for the Launch Site, Latitude, and Longitude variables.
- getPayloadData() does request and extracts data for Payload variable.
- getCoreData() does request and extracts data for the core variables including launch pad and outcome.

- After collecting data from all of the launces from the SpaceX REST API, we filtered our data to only the Falcon 9 launches based on the booster version since that’s the focus of the project.

- Before the filter:

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitu
1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.7431
2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.7431
4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.7431
5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.7431
6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.5773

- After the filter:

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	S
1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B
2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B
3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	E
4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	E
5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	E

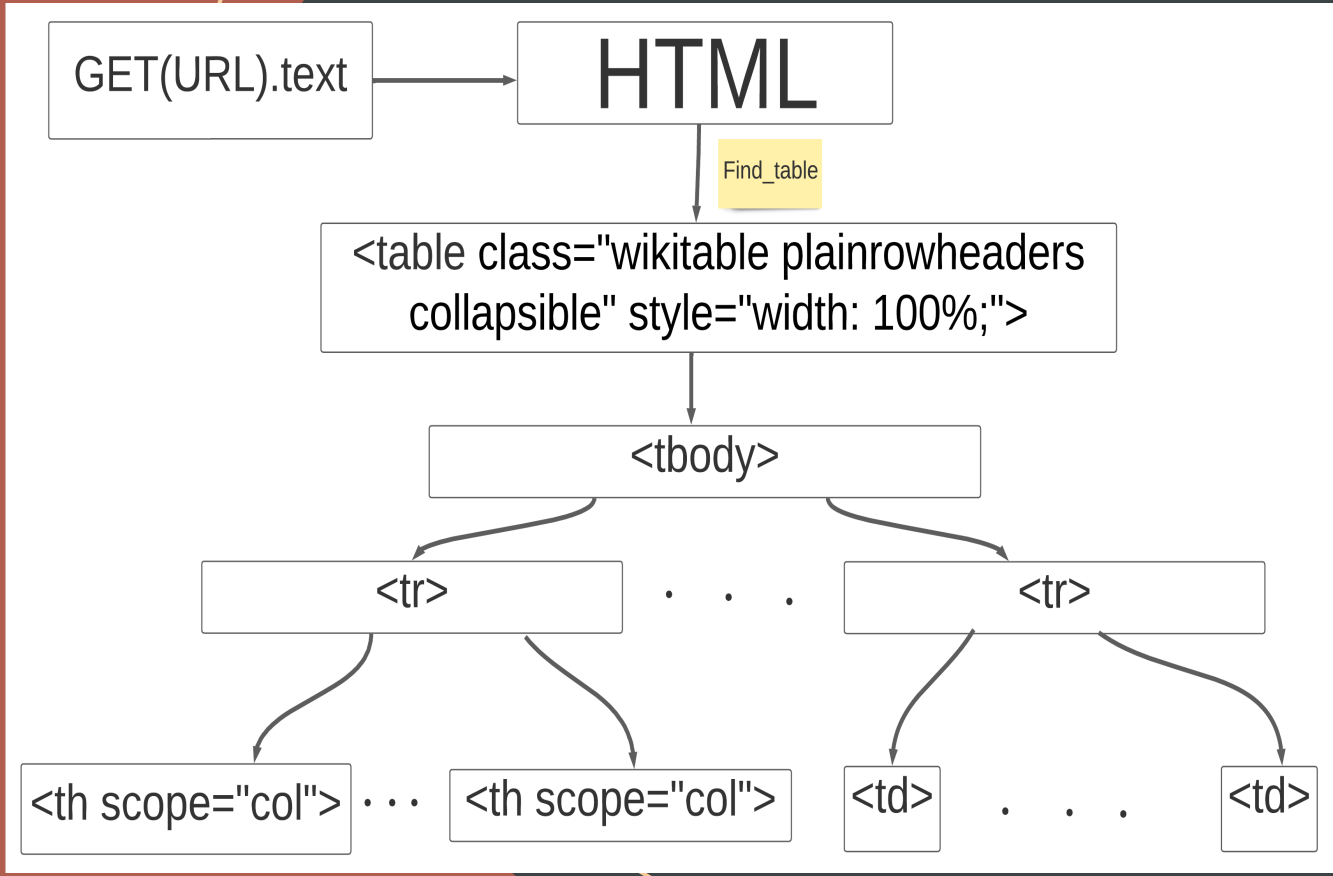
Data Collection Notebook

- To find out more about how we collected data via REST API in JSON format, visit the following GitHub link:
  - https://github.com/zanderRana/IBM_APPLIED_CAPSTONE_SPACEX_LAUNCHES/blob/master/SPACEX%20LAUNCH%20DATA%20COLLECTION.ipynb

Could we have done another way besides JSON?

- Yes, another method that we could use and have tested for the Data Collection process is web scraping using BeautifulSoap from BS4.
- Like the last process, we issue an HTTP GET request for a Wikipedia page containing information about SpaceX rocket launches.
- But instead of extracting data responses using JSON, we use BeautifulSoap to extract HTML tag elements from the response.
- By browsing through the **<table>** elements, we looked for the data necessary for the problem and collected information from the **<thead>** elements for columns and the **<tbody>** elements for the records themselves.

- Here is a chart depicting the BeautifulSoap Tree.
- We get a response from the HTTP request in the form of HTML and use the built-in find('table')) function to find the <table> elements.
- Our goal table contains data about Falcon 9 launches, so we find that specific table.
- Based on the chart, we construct our table:
  - The column names are extracted from the table based on the <th> elements that contain the names. These elements are stored within the first <tr> element of the table's body aka <tbody>. These names are added to our DataFrame.
  - The remaining <tr> elements contain the data record values corresponding to each of the table column names of course. We collect those row values for our DataFrame.



Here is the resulting DataFrame after collecting text data from the HTML tags:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

121 rows × 11 columns

Data Collection With Web Scraping Notebook

- To find out more about how we collected data via REST API and web scraping with BeautifulSoap, visit the following link:
  - https://github.com/zanderRana/IBM_APPLIED_CAPSTONE_SPACEX_LAUNCHES/blob/master/SPACEX%20LAUNCH%20DATA%20COLLECTION%20WITH%20WEB%20SCRAPING.ipynb

DATA WRANGLING

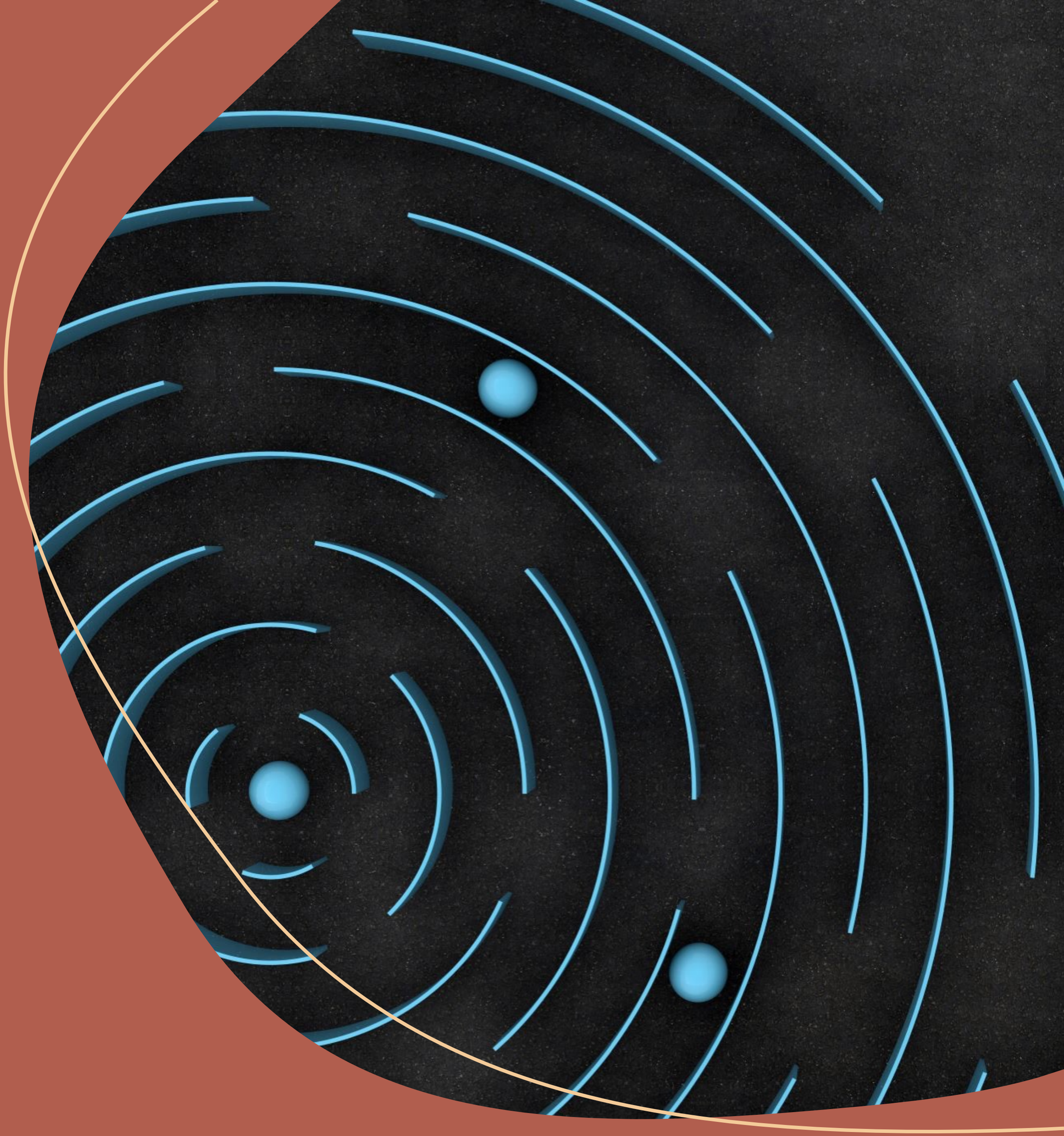
Missing values

- We noticed in our table that we have several missing values.
- Here is a table depicting the amount of NAN values present in our collected table.
- The LandingPad variable has too many missing values to the point where we thought it wasn't necessary to use the variable.
- The PayloadMass variable on the other hand only has a few missing values.
- We replaced the missing values in the payloadMass variable with the average payload mass.

FlightNumber	0
Date	0
BoosterVersion	0
PayloadMass	5
Orbit	0
LaunchSite	0
Outcome	0
Flights	0
GridFins	0
Reused	0
Legs	0
LandingPad	26
Block	0
ReusedCount	0
Serial	0
Longitude	0
Latitude	0
dtype: int64	

Exploratory Data Analysis for Data Understanding

- In the first steps of our data analysis, we wanted to see the frequency of the different launch sites, the orbits in our data, and the different launch outcomes.
- CCAFS SLC 40 is by far the most common launch site with 55 records.
- The most common orbit that rocket launches aim towards is the GTO orbit with 27 records. This orbit is located 22,236 miles above the Earth.
- Th most common outcome is True ASDS, with 41 records. This meant that the rocket successfully landed to a drone ship.



Frequency tables

```
GTO       27
ISS       21
VLEO      14
PO         9
LEO        7
SSO        5
MEO        3
ES-L1      1
HEO        1
SO         1
GEO        1
Name: Orbit, dtype: int64
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Focusing on Outcome

- With the different number of outcomes in our data, we wanted to simplify it to just seeing if a launch succeeded or not.
- We thus create a new column to add to our table, named 'Class'.
- This column states whether a launch succeeded or failed. A value of 1 meant the launch succeeded and a value of 0 meant the launch failed.
- To create this column based on the outcomes, we labeled the launch records as failed launches (class=0) if their outcome was any of the following:
  - False ASDS, False Ocean, False RTLS, None ASDS, and None None
- If the launch record had an outcome other than the ones specified as failed outcomes, the launch record would be labeled a successful launch (class=1).
- After creating and adding the class column, we calculated its mean to find an average success rate of 66.67%.

Here is the updated DataFrame with the class column

BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

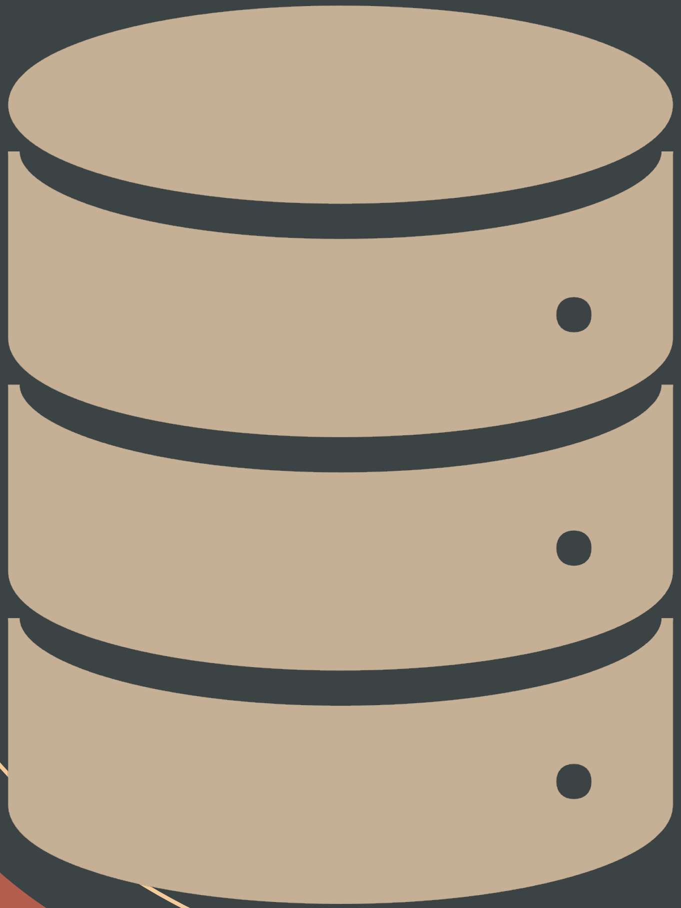
Data Wrangling Notebook

- To find out more about how we performed Data Wrangling and analyzed the frequency of the launch data, visit the following link (QUICK NOTE: The missing value replacement is shown in the last notebook):
  - https://github.com/zanderRana/IBM_APPLIED_CAPSTONE_SPACEX_LAUNCHES/blob/master/SPACEX%20DATA%20WRANGLING.ipynb

EDA For Data Understanding and Data Preprocessing

EDA With SQL

- For the following pages, we will be using SQL to query our data and draw more insights.
- I have used SQL to query launch data based on time of launch, booster versions, ranges of payload, and the types of landing outcomes of course.



All Launch Site Names

- Although shown before, here are the different launch sites displayed as a query result. As stated before, CCAFS SLC 40 is the most common launch site .

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

Launch site names beginning with CCA

- Here are 5 records of launches that involve launch sites associated with CCA. Notice how most of the launches have a successful mission outcome.

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

Total Payload Mass

- Here is a query result of the total payload mass of boosters carried by NASA. The total mass is 48,213 KG.

total_mass
48213

Average Payload Mass by F9 v1.1

- Here is a query result of the average payload mass of launches with the booster F9 V1.1. The average payload is 2,928 KG.

avg_payload
2928

First Successful Ground Landing Date

- Here is a query result of the date when the first successful landing outcome in ground pad was achieved.
- The first successful landing was on December 22, 2015.

1
2015-12-22

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

Successful Drone Ship Landing Between 4000 and 6000

- Here are the boosters that have succeeded in a drone landing and has a payload between 4000 and 6000

Total Number Of Successful and Failure Mission Outcomes

- Here are the frequency counts of the successful and failure mission outcomes. Almost all of the mission outcomes were a success with only one failure.

mission_outcome	amount
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Here are the boosters that carried the highest payload. Most of them are of the F9 B5 variant

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Here are the failed landing outcomes involving a drone ship in 2015.

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank landing Outcomes Between 2010-06-04 And 2017-03-20

- Here are the ranks of the landing outcomes in descending order.
- It seems that were a lot of failures involving a parachute and a lot outcomes are precluded.

landing_outcome	RANK
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

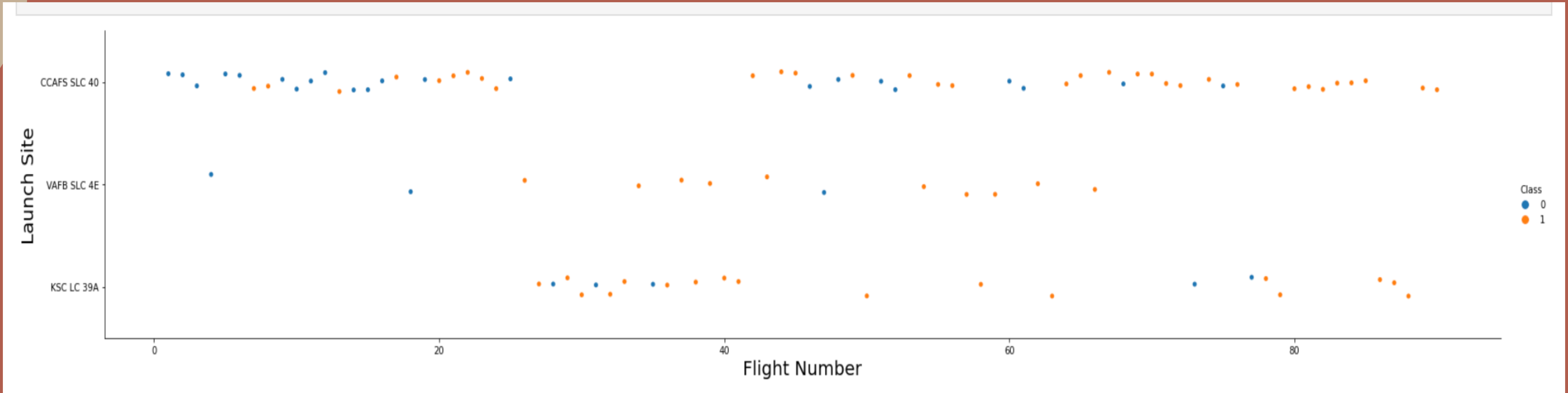
SQL Notebook

- To see the SQL query themselves, visit the following link:
  - https://github.com/zanderRana/IBM_APPLIED_CAPSTONE_SPACEX_LAUNCHES/blob/master/SPACEX%20EDA%20WITH%20SQL.ipynb

EDA With Data Visualization

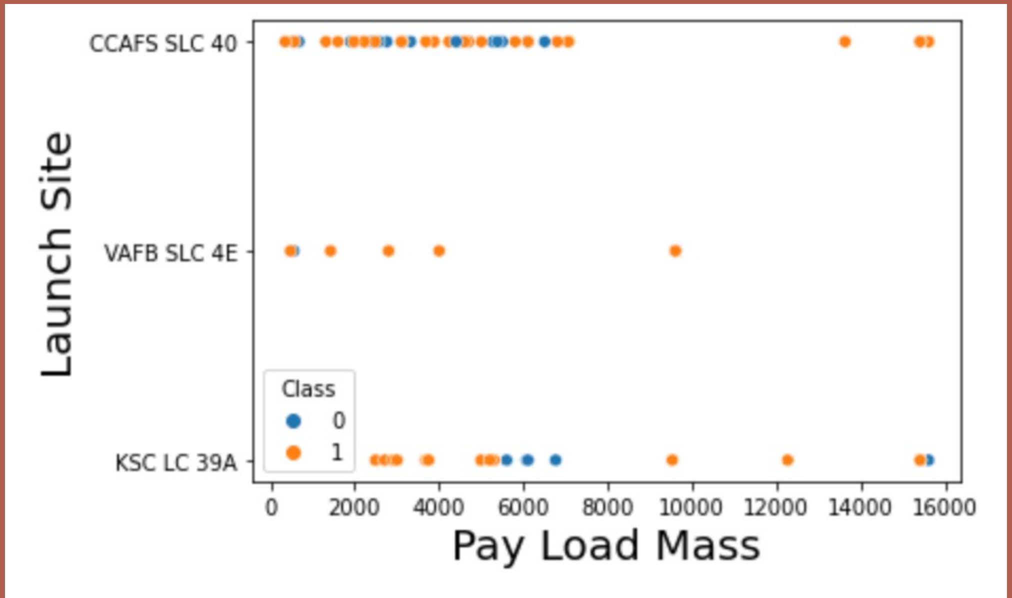
- For the following pages, we will be graphing our data to identify relationships and patterns that heavily impact the landing outcome of a launch.
- I have used catplots, scatterplots, bar graphs, and line graphs to see the distribution of data based on features and Identifying patterns in the data that affect whether launches have failed or succeeded

Flight Number Vs Launch Site



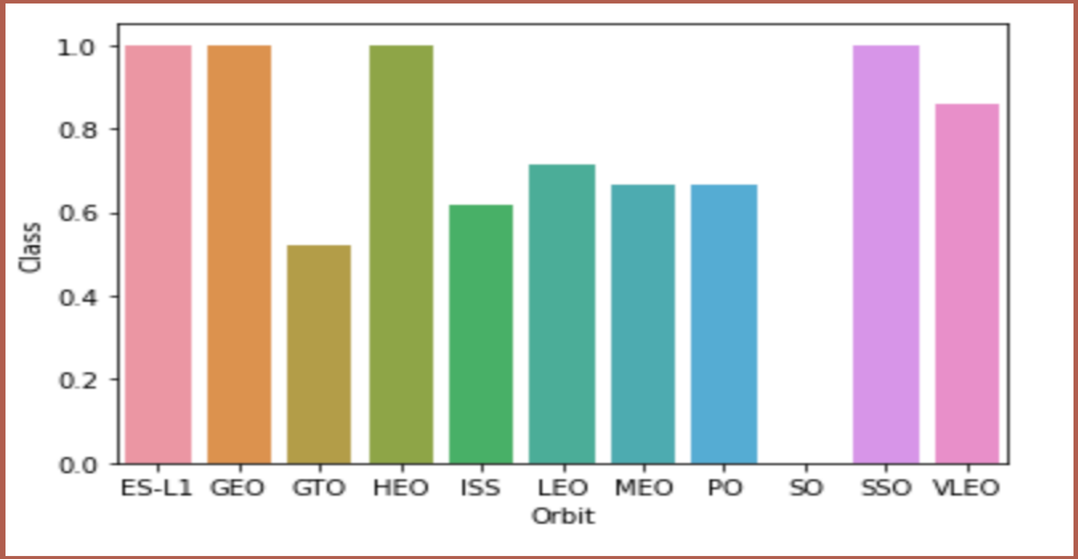
- Here is a catplot of the relationship between Flight Numbers and the launch site.
- You can see that the success rate of the launches landing heavily differs based on the launch site. Some sites progressively get more succeeding launches over time. This will be important when we determine whether a launch succeeds or not.
- VAFB SLC 4E and KSC LC 39A have high success rates.
- CCAFS SLC 40 starts out with majority of launches being failures, but over time the launch site

Payload vs Launch Site



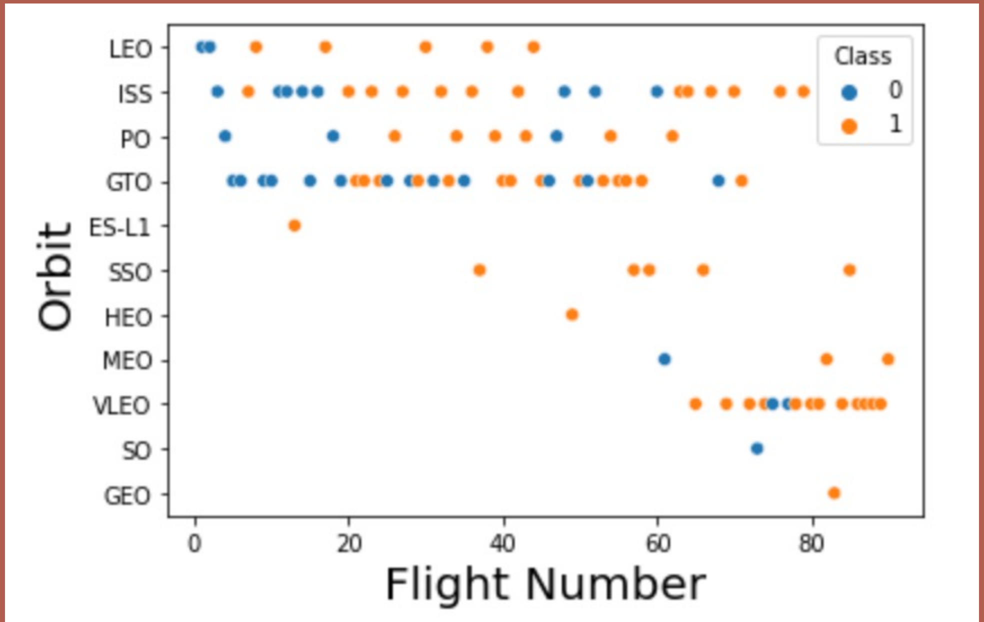
- Here is the scatterplot of the relationship between Pay load mass and Launch Site.
- Once again, the success rate varies by launch site.
- While VAFB SLC 4E has no failure, we can see that the other two launch sites sitck to a certain payload range where most of the launches are between 2000 and 8000.
- For CCAFS SLC 40, having a high payload really benefitted its launches where the extremely high payloads all had successful launches. KSC LC 39A had one failure despite high payload.

Success Rate Vs Orbit Type



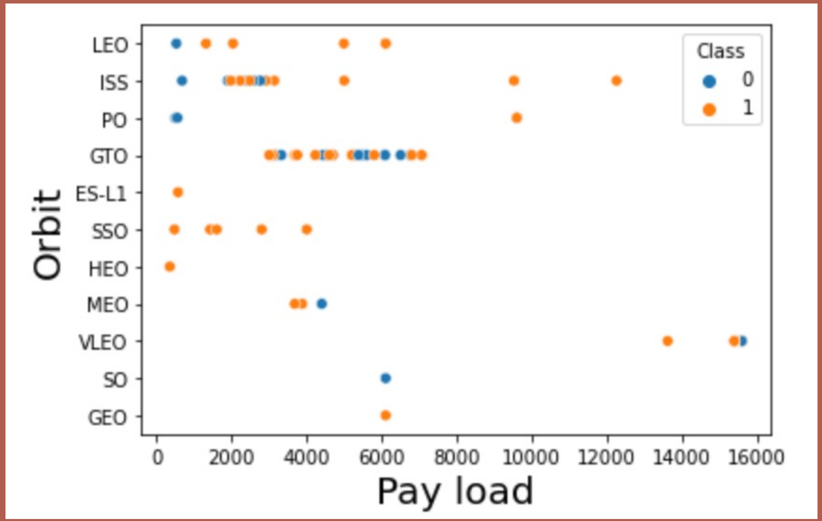
- Here is the bar graph of the relationship between Orbit and the Success Rate.
- It seems that the orbits ES-L1, GEO, HEO, SSO , and VLEO have very high, near perfect success rates.
  - These orbits vary in orbital shape where some are elliptical while others are circular. Most importantly, these orbits also vary in distance from Earth. So, it doesn't seem that orbit shape or distance impact the success rate but instead its the orbit's function itself

Flight Number Vs Orbit Type



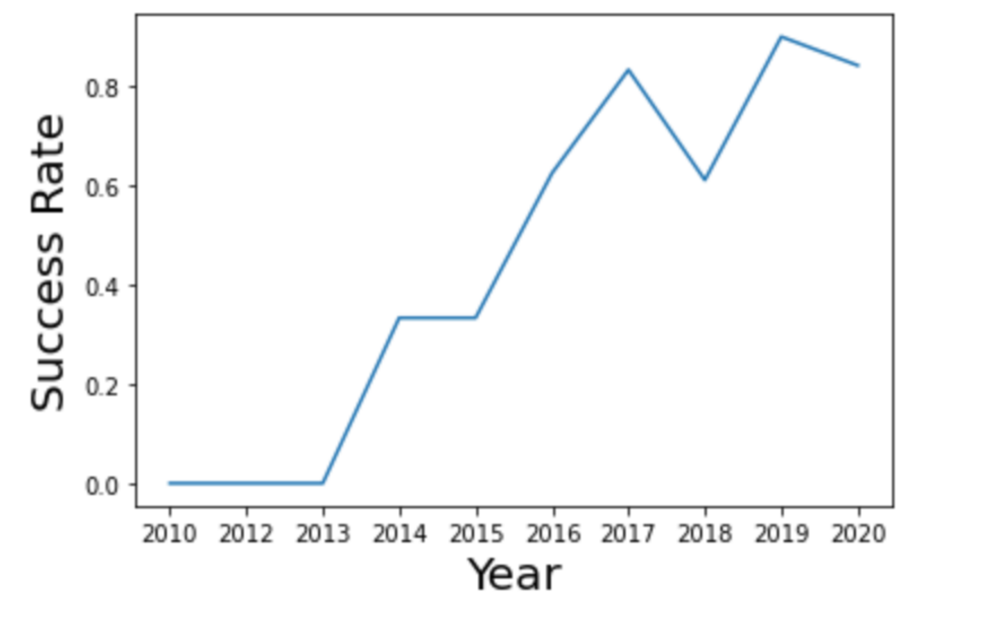
- Here is the scatter plot showing the relationship between flight number and orbit.
- As stated before, the success rates vary by orbit.
- Some orbits have more successful launches over time like LEO and VLEO, but other orbits are inconsistent with the launch outcomes regardless of time and how many launches are aimed to them.

Payload vs Orbit Type



- Here is a scatter plot showing the
- Unlike the launch sites, most of the orbits don't lie within a certain payload range. GTO and SSO are the only orbits that lie within ranges (3000,7000) and (0,4000) respectively.
- Like the launch sites however, some orbits payload doesn't matter for successful launches and thus no relationship, but for others the higher payloads result in them having successful launches, ISS being a big example.

Launch Success Yearly Trend



- Here is a line plot showing the success rate over the years.
- As you can see, ever since 2013, the success rate has been consistently getting higher until 2020 where it dropped lower than 2019, but still very high relative to previous years.

Results and Insights

- Based on the results, Orbits and launch sites themselves can impact the success rate of the launch.
- Some orbits and some launch sites were able to have more successful launches based on number of launches over time and the amount the payload where the higher the payload, the higher the chance of success.
- But this same idea is not present in other orbits and launch sites as well where the data relationships are more inconsistent.
- This is why on top of just determining the outcome by just the orbit and launch site, we need to be on the lookout for special launch sites like CCAFS SLC 40 and special orbits like ISS since they had more successful launches by using higher payloads and doing more flights.

Preprocessing the categorical variables

- Since Machine Learning models can't process qualitative data and knowing that launch site and orbits are important features based on our previous insights, we used the pandas get_dummies function to create dummy variables for all of the categorical variables in our table.
- Our table went from having 12 columns to know having 80 columns.
- Now every value is converted to float, and we have fully processed dataset for our models.
- The results are in the next slide

New DataFrame After Generating Dummies

	FlightNumber	PayloadMass	Flights	GridFins	Reused	Legs	Block	ReusedCount	Orbit_ES-L1	Orbit_GEO	...	Serial_B1048	Serial_B1049	Serial_B1050	Serial_B10
0	1.0	6104.959412	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
1	2.0	525.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
2	3.0	677.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
3	4.0	500.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
4	5.0	3170.000000	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
85	86.0	15400.000000	2.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0	0
86	87.0	15400.000000	3.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0	0
87	88.0	15400.000000	6.0	1.0	1.0	1.0	5.0	5.0	0.0	0.0	...	0.0	0.0	0.0	1
88	89.0	15400.000000	3.0	1.0	1.0	1.0	5.0	2.0	0.0	0.0	...	0.0	0.0	0.0	0
89	90.0	3681.000000	1.0	1.0	0.0	1.0	5.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0

90 rows × 80 columns

Data Visualization And Preprocessing Notebook

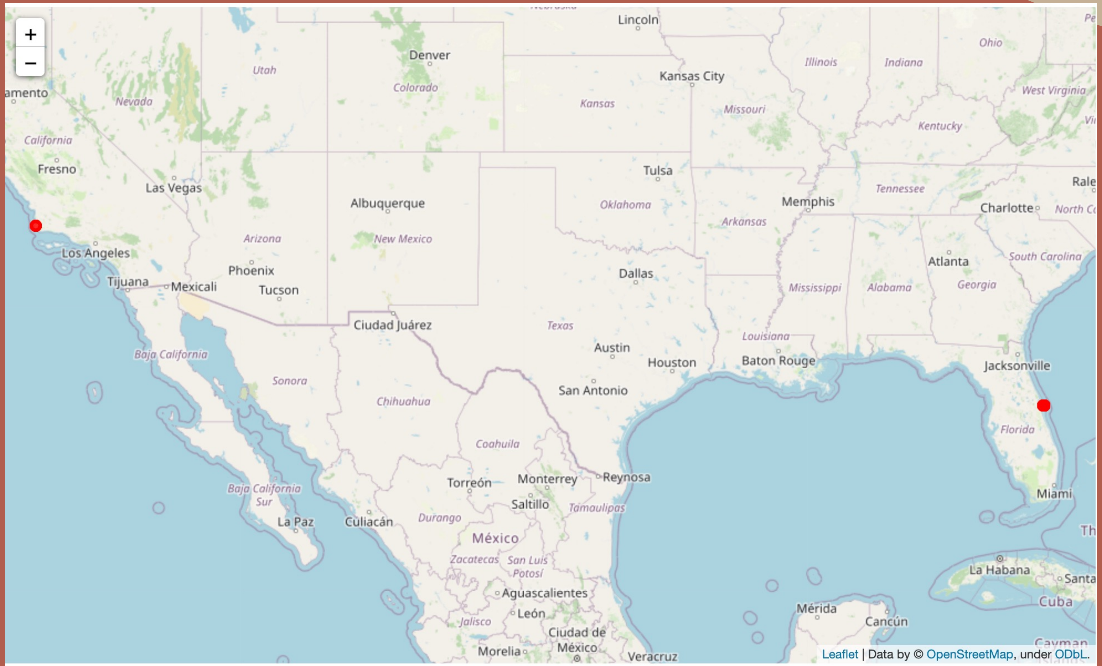
- To find out more about how we implemented graphs for visualization and preprocessed categorical data, see the following link:
  - https://github.com/zanderRana/IBM_APPLIED_CAPSTONE_SPACEX_LAUNCHES/blob/master/SPACEX%20EDA%20VISUALIZATION.ipynb

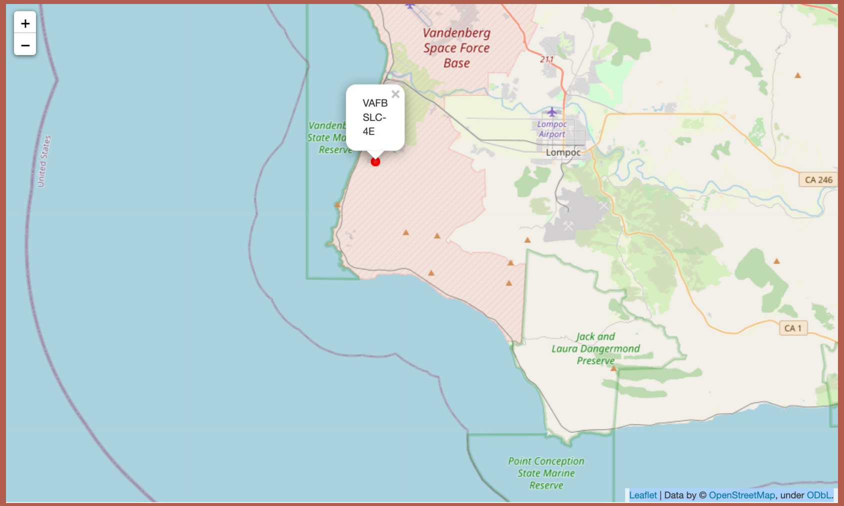
Visual Analytics For Data Understanding

- We used Folium to create maps visioning where the launch sites took place along with their outcome for more visual understanding.
  - I used circle markers, regular markers with popups and icons for description, polylines, and marker_clusters grouping close markers in my maps.
- We also used plotly and dash to build an interactive dashboard for users to see the distributions of launch sites along with their outcomes and payload mass for more visual understanding.

Launch Location Markers For the MAP

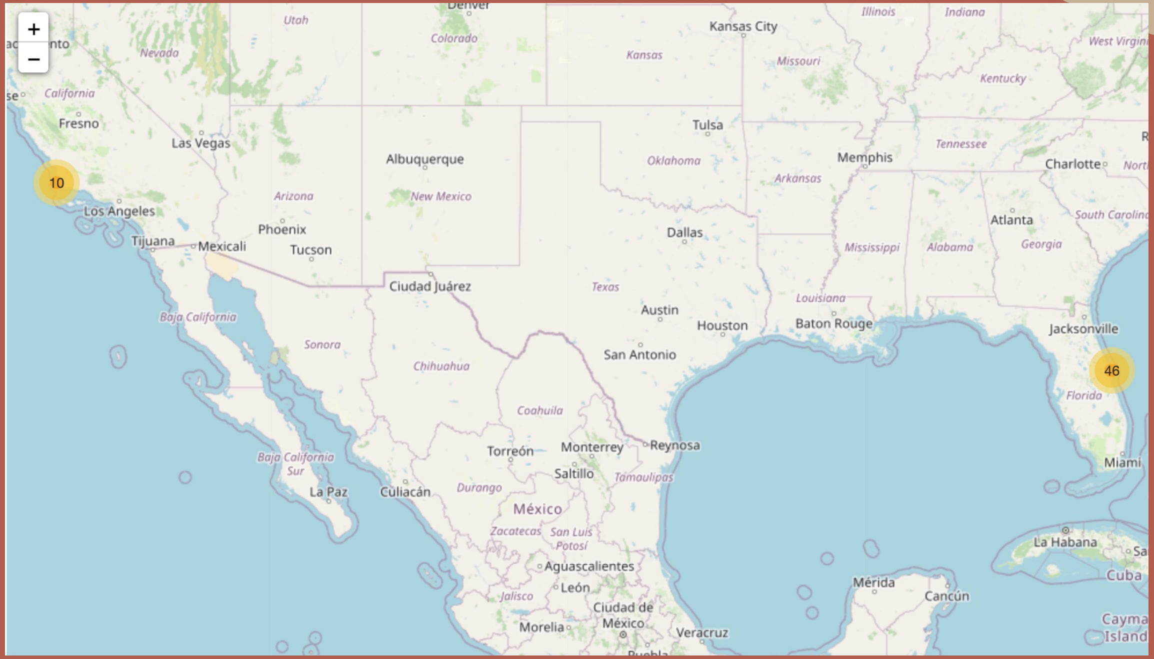
- Here are the location of all the launch sites, which are mainly in three specific spots, two in Florida and one in California

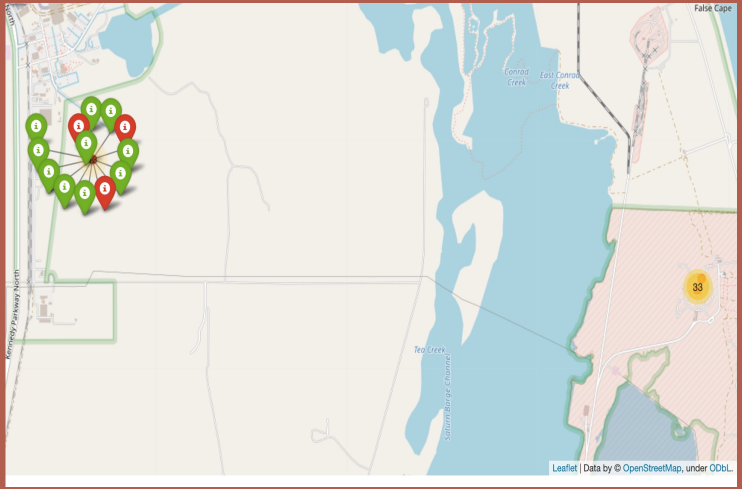




Marker clusters and outcomes

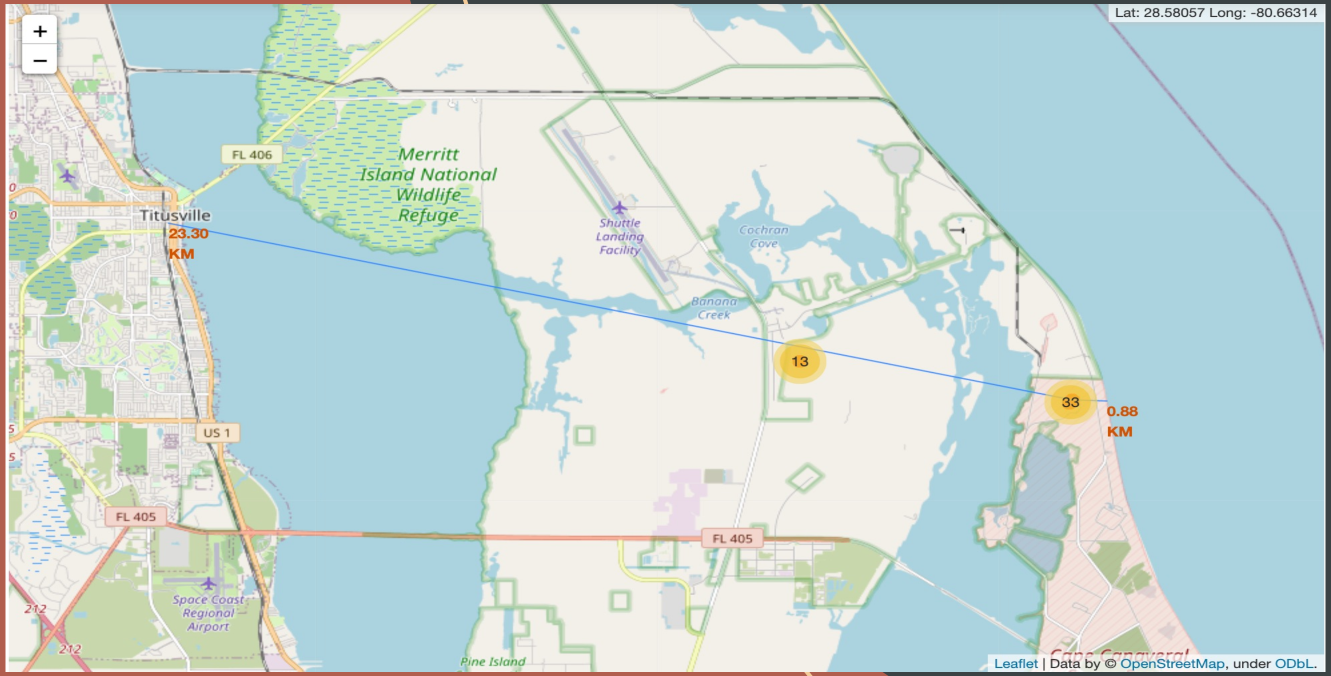
- Here are the cluster markers.
- As you can see in the Florida launch sites, the left launch site (KSC LC-39A) has more successful launches than the launches from the right launch site (CCAFS LC-40 and CCAFS SLC-40 and)

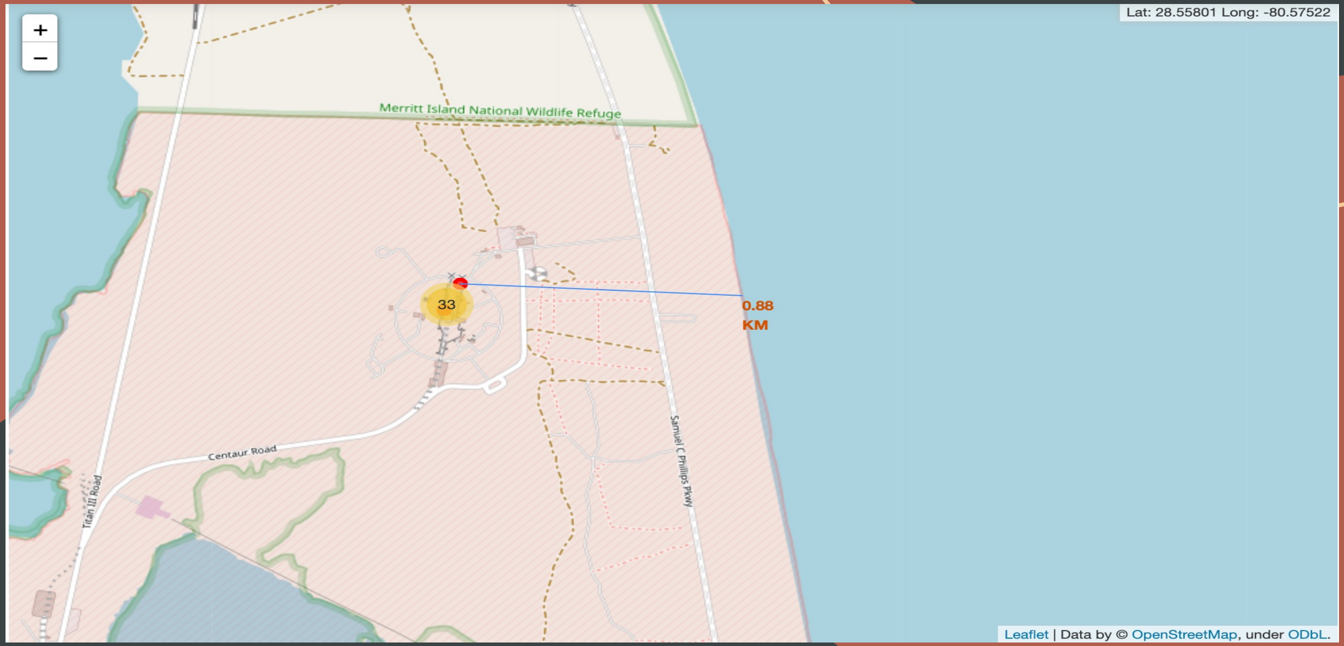




Line Markers in Map

- Here is a line marker showing the closest coastline and closest city to CCAFS SLC-40

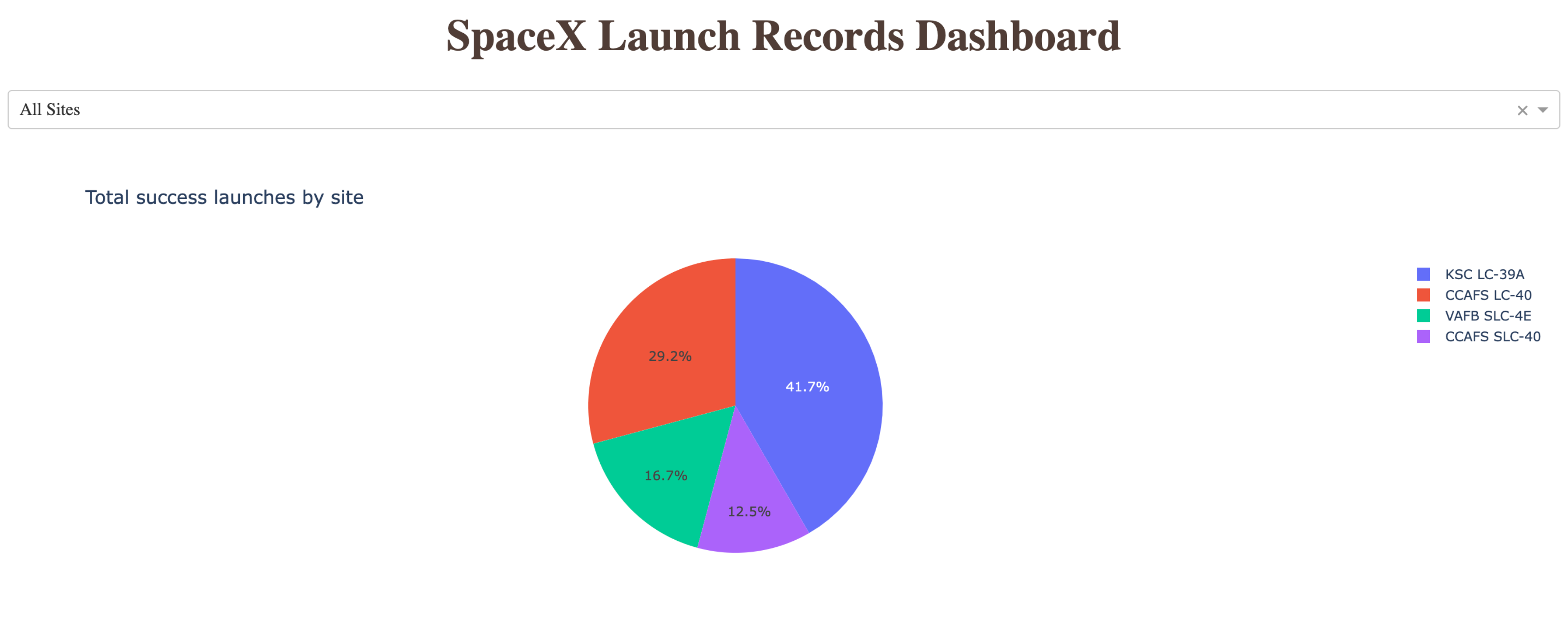




Folium Map Notebook

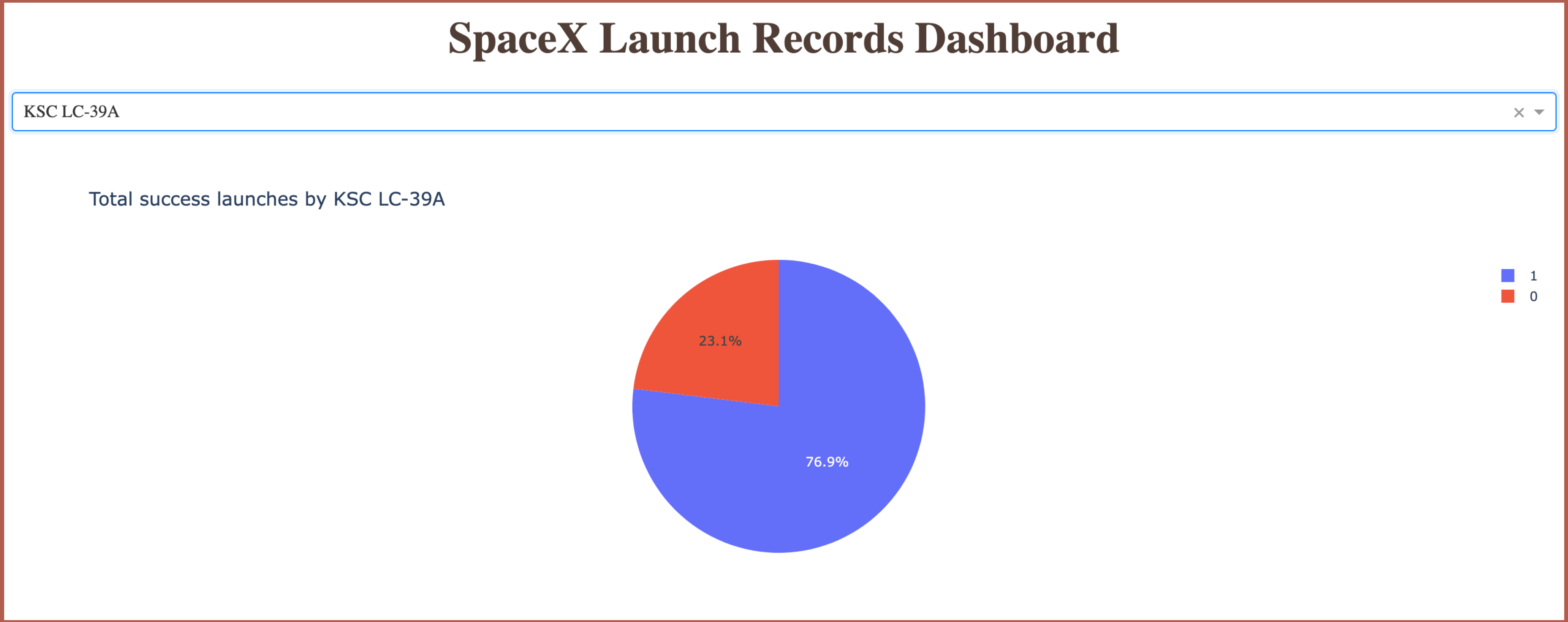
- To find out more about how the maps are implemented, see the following link (GITHUB for some reason cannot show the maps, so I apologize for the inconvenience).
  - https://github.com/zanderRana/IBM_APPLIED_CAPSTONE_SPACEX_LAUNCHES/blob/master/SPACEX%20MAP%20FOLIUM.ipynb

INTERACTIVE DASHBOARD PIE CHART FOR ALL SITES



- Here is the Pie chart for all sites in the dashboard. As you can see, KSC LC-39A has the most successful launches (this was shown earlier in the maps and graphs)

INTERACTIVE DASHBOARD PIE CHART FOR KSC LC-39A



- Here is the Pie chart for the site with the highest success rate, KSC LC-39A, and you can see how the success rate is with 76.9% of the launches being successful.

INTERACTIVE DASHBOARD SCATTER PLOT WITH RANGE LINE

- Here is the Scatter plot showing the relationship between boosters and the payload.
- As you can see, there seems to be no relationship between successful outcomes and the booster type.
- Most of the boosters with successful launches lie in the range between 2k and 4k and yet there are more failure launches in that range too. Some.
- Higher payloads barely yield more successful landings.

Interactive Dashboard Code

- To see how this dashboard was implemented, visit the following link:
  - https://github.com/zanderRana/IBM_APPLIED_CAPSTONE_SPACEX_LAUNCHES/blob/master/SPACEX%20INTERACTIVE%20DASHBOARD.py

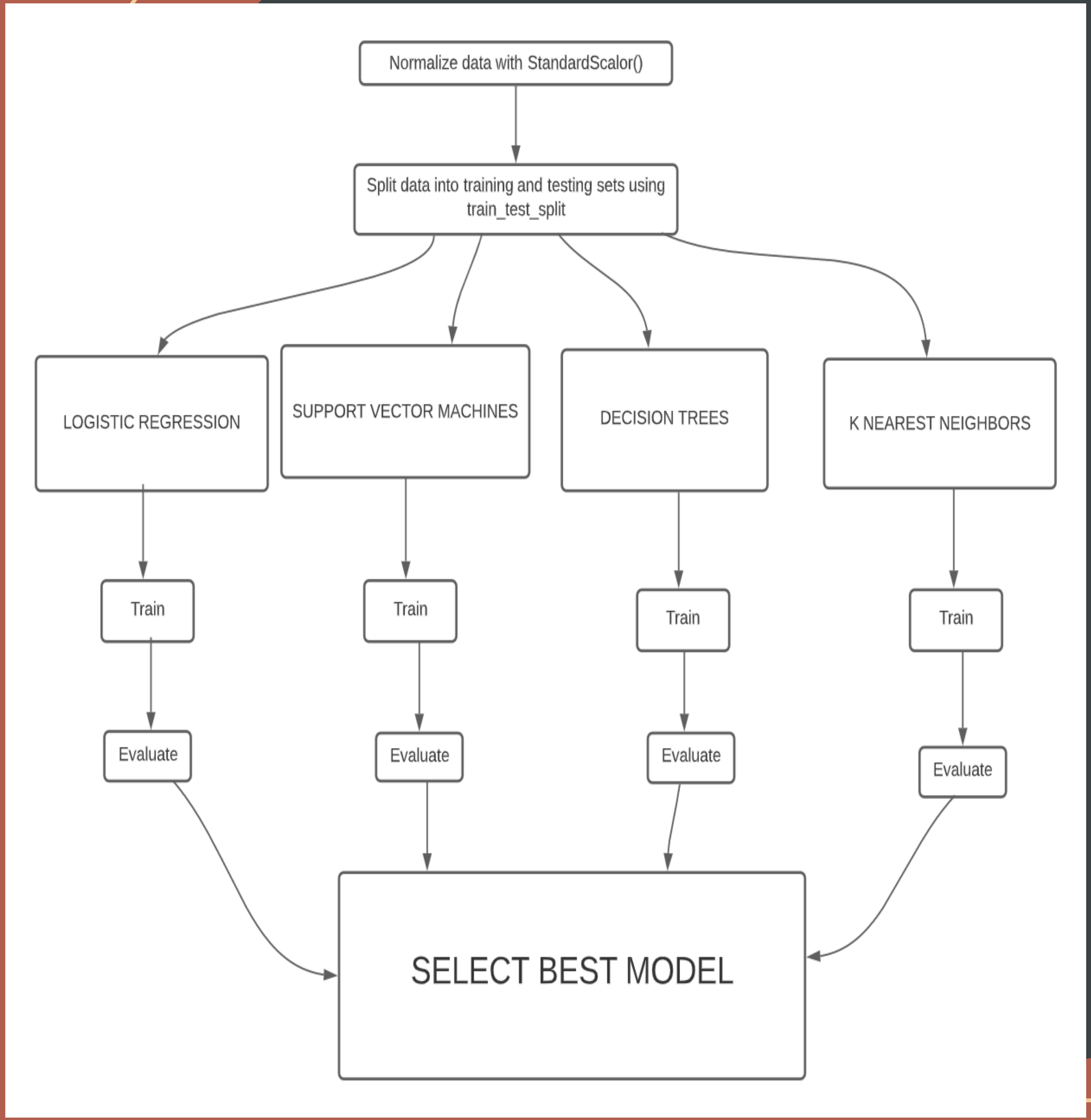
Predictive Analysis

Machine Learning models

- One of the last steps in the whole methodology: model prediction and evaluation.
- Since we are solving a yes/no problem regarding if a launch will succeed, we are using a classification model.
- Using our preprocessed data, we tested out different classification models to predict whether a launch succeeded given all the features discussed before including launch site, orbit, payload, flight number, booster, etc.

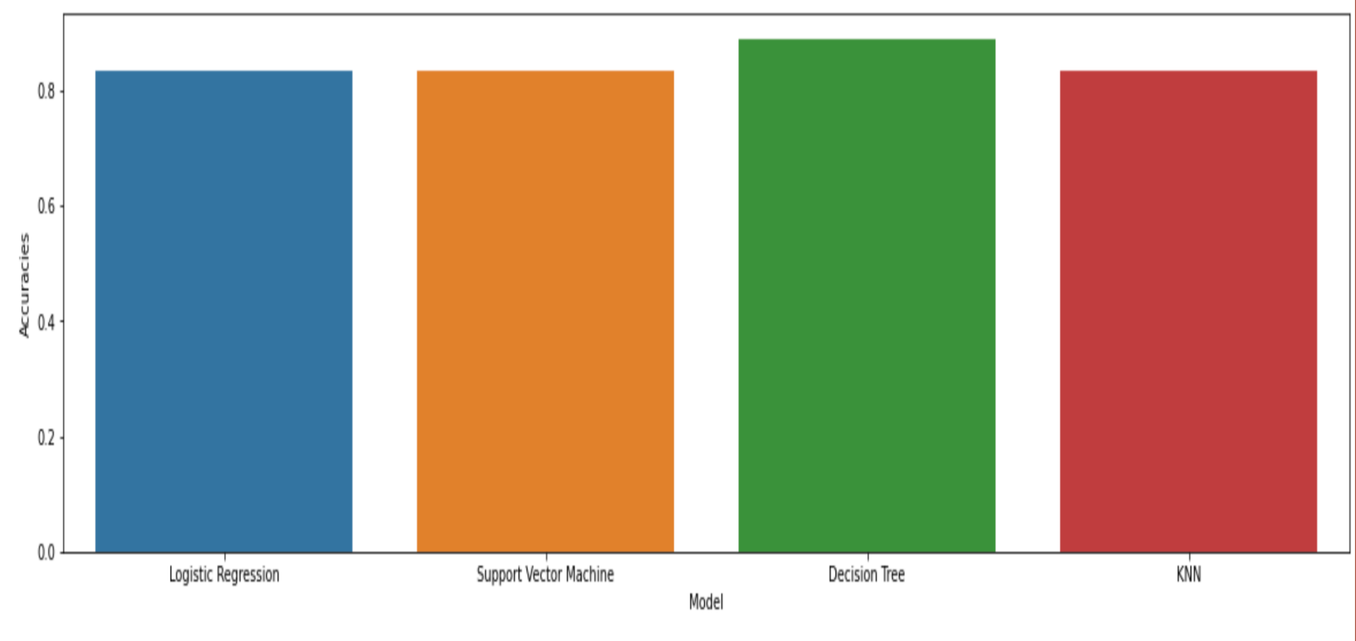
Process

- As you can see from the flow chart, we first normalize the data then split the data into our training and testing sets.
- We then train and evaluate each of the 4 classification models: LogReg, SVM, DT, and KNN. We train them using GridSearchCV to find the best hyperparameters associated with each of the models for maximum efficiency.
- Then we select the most efficient model with its associated parameters.



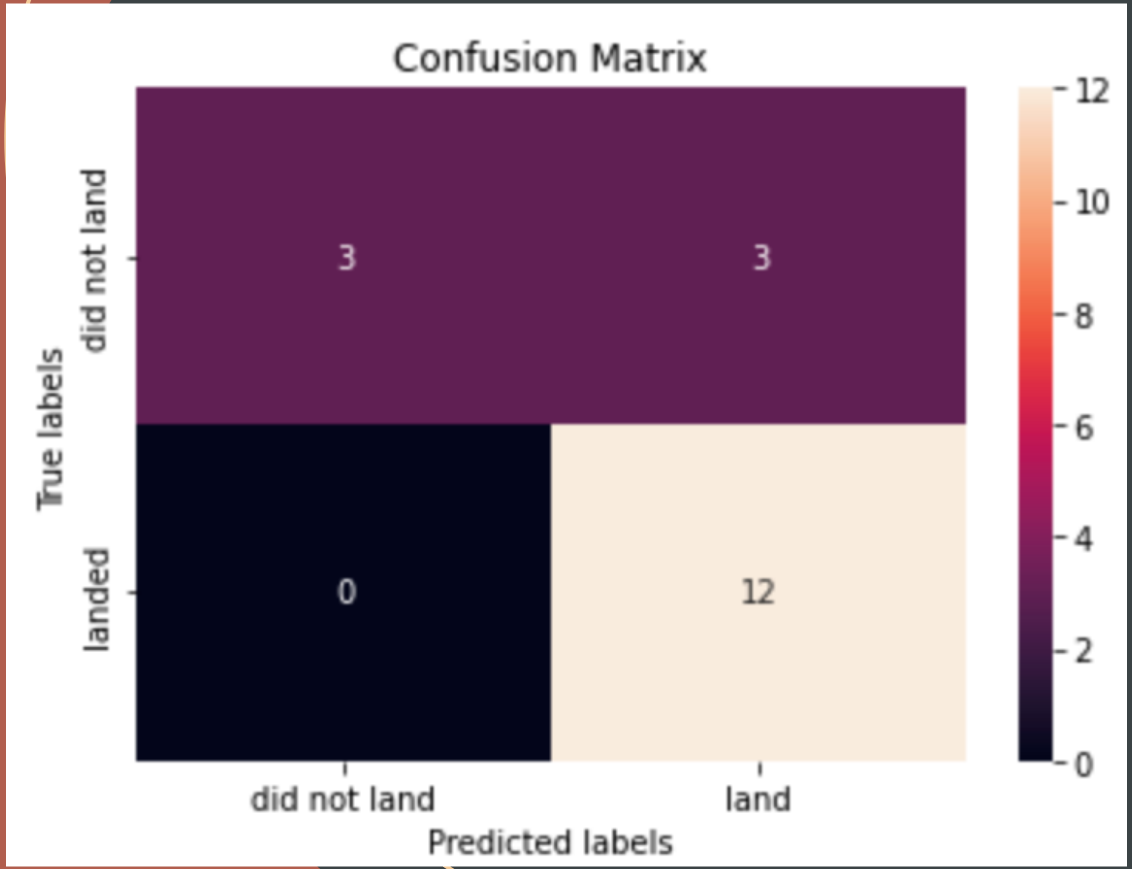
Best Model

- After testing and evaluating our models with R-Squared and the confusion matrix, most of the models had similar results.
- But one model slightly stood out from the rest, Decision Trees.
- We found that the Decision tree model with a max depth of 16 is the highest performing model.
- Here is a graph showing the models and their accuracies



Confusion Matrix

- Although Decision Tree had the highest score accuracy of the four models, all the models had the exact same Confusion Matrix.
- All four models had only 3 false positives.



Results and Insights

- The Decision Tree Model is the most ideal model for this problem. There are certain features that have little to no impact on the landing outcome such as the booster while other features like the Orbit and launch site had a bigger impact.
- Furthermore, some of the values for each of the features were heavily impacted by other features as we saw in previous studies such as how CCAFS SLC 40 had more successful outcomes with higher payload and flight numbers or how the orbit ISS had more successful launches with higher payload.
- Decision Trees can address both key terms since it prioritizes certain features over the others and follows a logical flow based on the conditions.

Machine Learning Prediction Notebook

- To find out more about how we trained our models and evaluated them, see the following link:
  - https://github.com/zanderRana/IBM_APPLIED_CAPSTONE_SPACEX_LAUNCHES/blob/master/SPACEX%20MACHINE%20LEARNING%20PREDICTIONS.ipynb

Conclusion

- Based on the data we have gathered, analyzed, and modeled, if we want to have more successful launches, we must consider the functionality of the different orbits where each orbit may or may not benefit from higher payloads and flight numbers. The same logic applies to other features especially for each of the different launch sites.
- Utilizing each of the orbits and launch sites with the desired payload and features examined in this study will allow the SpaceX rockets to be more efficient than they are right now even with their recent upward trend.
- This study can help companies like SpaceX ensure that they can have the most efficient rockets with the minimum cost possible when utilizing the different types of features that we have closely analyzed and proved in our data.
- The Decision Tree model is ensured to accurately compute the chance of success using the features we trained it with.

THANK YOU