IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alex Chan
17 July 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- ## Summary of methodologies

The objective of this project is to find a suitable predictive model that can predict the outcome of Falcon 9 First stage landing. Data used for this project is extracted from https://api.spacexdata.com/v4/payloads/.

Four predictive models were built using train-test split (80/20) dataset ; namely

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. KNN

- ## Summary of all results

There are a total of 18 test samples in the test set. Based on GridSearch hyperparameter tuning method, the decision tree model returned the highest accuracy score of 0.8892857 using the training dataset.

1. Logistic Regression – Accuracy Score : 0.84643
2. SVM – Accuracy Score: 0.84821
3. KNN - Accuracy Score: 0.84821

However all four models returned the same accuracy score of 0.83333 with a False Positive of 3, True Negative of 3 and True Positive of 12.

Introduction

- Project background and context

  SpaceX's advertised cost of rocket launches has been in general less expensive than its peers in the industry. This is mainly due to the re-use of first stage. Our objective is to identify the information and/or parameters available in the public domain relating to past SpaceX launches, which can help predict the outcome of the first stage landing. The ability to determine the first stage landing will then allow us to determine the cost of each launch, thus providing strategic information to client Space Y in their sales strategy.

- Problems you want to find answers

  1. Where are the launch sites located and what are the characteristics of these launch site locations? (e.g. coastal area, proximity to logistics and transportation)
  2. Are there any relationship between launch sites and number of successful landing?
  3. What are the other features or factors that could influence the likelihood of a successful landing? (Does launching a rocket closer from the equator and at the eastern coast increases the probability of a successful launch?)

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
- Data collection via SpaceX REST API
- Data collection using Web Scrapping method from Wikipedia
- Perform data wrangling
- Checking for null or missing entries in columns
- Classifying mission outcomes (successful vs unsuccessful) as categorical feature using one hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Gridsearch hyperparameter tuning
    - Confusion Matrix model assessment

Data Collection

1. **Data collection via SpaceX REST API**

i. Data is collected via the open source SpaceX REST API, using HTTP get request to **https://api.spacexdata.com/v4**.

ii. The response is then converted into a json file using .json() and converted into dataframe using .json_normalize()

iii. We then extract selected columns ('rocket', 'payloads', 'launchpad', 'cores’) from the dataframe and use these columns to pull additional information from the SpaceX API.

2. **Data collection via Web Scraping in Wikipedia**

i. Data is collected via Wikipage dated 09th June 2021, using HTTP get request.

ii. The response is then converted into BeautifulSoup object, in order to iterate through the HTML tables to extract the necessary information.

iii. A dataframe is then created by parsing through the HTML tables.

Data Collection – SpaceX API

Obtain response from API

Convert response to Json file and create data

Apply the customized functions to populate data with additional information

Create dictionary with data and create new data frame with the dictionary

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
data = pd.json_normalize(response.json())
```

```
getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

Data Collection – SpaceX API

Create data frame and return only results for falcon 9

```
df = pd.DataFrame.from_dict(launch_dict)
data_falcon9 = df[df['BoosterVersion']!='Falcon 1']
```

Check for missing or null values and replace with mean

```
data_falcon9.isnull().sum()
mean_PayloadMass = data_falcon9['PayloadMass'].mean()
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].replace(np.nan,mean_PayloadMass)
```

Export final data as csv file

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

link to GITHUB WK1-API

Data Collection - Scraping

Obtain response from HTML

Create BeautifulSoup object

Obtain column names from target table

Create dictionary

Fill dictionary by parsing launch tables and create a data frame from dictionary

Link to GitHub

```
response = requests.get(static_url)
```

```
soup = BeautifulSoup(response.text, 'html.parser')
```

```
first_launch_table = html_tables[2]
column_names = []

# Apply find_all() function with `th` element on first_launch_table
# Iterate each th element and apply the provided extract_column_from_header(
# Append the Non-empty column name (`if name is not None and len(name) > 0`)
for th in first_launch_table.find_all('th'):
    name = extract_column_from_header(th)
    if name is not None and len(name) > 0 :
        column_names.append(name)
```

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

```
df=pd.DataFrame(launch_dict)
```

Data Wrangling

1. Establish the count and types of landing outcome

```
# landing_outcomes = values on Outcome column

landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
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
```

2. Re-group landing outcomes as good vs bad outcome

```
for i,outcome in enumerate(landing_outcomes.keys()):
    print(i,outcome)
```

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

We create a set of outcomes where the second stage did not land successfully:

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

```
{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
```

3 Creating a landing outcome label using one hot-encoding ; 0 – bad outcome , 1 – good outcome

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise

landing_class=[]
for outcome in df.Outcome:
    if outcome in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)

print(landing_class)
```

4. Adding the new column ' Class' into Space X dataframe and saving as CSV file

```
df['Class']=landing_class
df[['Class']].head(8)
```

```
df.to_csv("dataset_part_2.csv", index=False)
```

Link to GitHub

EDA with Data Visualization

Scatter Plots are used to determine the relationship between variables. In this case it shows the relationship between outcome of the launch and the flight variables for the following variables
-Payload Mass vs Flight Number
-Launch Site vs Flight Number
-Launch Site vs PayLoad Mass
-Flight Number vs Orbit
-PayloadMass vs Orbit

Bar Chart is used to plot scatter plots to determine the frequency counts for different categorical variable. In this case bar chart is used to show the success rate comparison between Orbit type.

Line chart is used to show the trend data. In this case the line chart how Space X Rocket Success Rate changes over time .

Link to GitHub

EDA with SQL

We performed SQL queries to extract the following information from the database SPACEXTBL

i. Extract the unique/distinct names of launch sites
ii. Display 5 records where launch sites begin with the string 'CCA’
iii. Extract and calculate the total payload mass carried by boosters launched by NASA (CRS)
iv. Calculate and show the average payload mass carried by booster version F9 v1.1
v. Extract the date when the first successful landing outcome in ground pad was achieved
vi. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
vii. List the total number of successful and failure mission outcomes
viii. Using subquery to extract and list the names of the booster_versions which have carried the maximum payload mass.
ix. Extract the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
x. Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Link to GitHub

Build an Interactive Map with Folium

- We used **folium. Circle** with radius 100 to circle the locations of the 4 launch sites in the map in red. To facilitate the identification of the circled locations, folium. **Marker** is used to mark the circled locations with names of the launch site.

- Next we created a MarkerCluster object to create cluster marker at each of the four launch site locations in the map. The purpose of the marker cluster is to display the total number of successful vs unsuccessful launch at each launch site. Folium marker is then used to mark each successful launch in green and unsuccessful launch in red at each launch site.

- The MousePosition is then added to the map so as to return the coordinate for a mouse over a point in time map. The coordinates allow us to retrieve the exact locations of highway, railroad and nearest city from a particular launch site.
- Markers are then plotted at the specific coordinates of the mentioned infrastructure and distance marker created to show the distance between the infrastructure and the launch site in question.
- PolyLine is drawn to provide a visual in terms of the horizontal distance measured between the launch site and the infrastructures (railway, highway, nearest city).

Link to GitHub

Build a Dashboard with Plotly Dash

- The following items were added into the interactive dashbpoard with Plotly Dash

1. **DropDown list** that allows the user to select a particular launch site or all launch sites at once.
2. **Pie Chart** to summarize the total count of (%) of successful launch and failures for a selected launch site or for all launch sites.
3. **Range Slider** that allows the user to select the specific range for PayLoad Mass.
4. **Scatter Plot/Chart** that shows the relationship between the Payload Mass for the user defined range VS the successful launch.

Link to GitHub

Predictive Analysis (Classification)

1. Data Preparation

- Data was first scaled/normalized using preprocessing.StandardScaler() so that all columns have a common scale which otherwise may distort/ influence the machine learning outcome.
- We then perform data splitting to train/test split of 80-20. The purpose of train-test split is to avoid the issue of over-fitting in the machine learning models.

2. Model Building

- Four different models were built using the training data set ,namely Logistic Regression, Decision Tree, SVM and K Nearest Neighbour models.
- The hyperparameters of these models are tuned/optimized using the GridSearchCV method.

3. Model Evaluation

- The tuned models are then tested with the test set and then assessed for its accurancy using the accuracy score method and confusion matrix .

4. Model selection

- Models are then compared using the test accuracy score and the model with the highest accuracy score selected.

Link to GitHub

Results

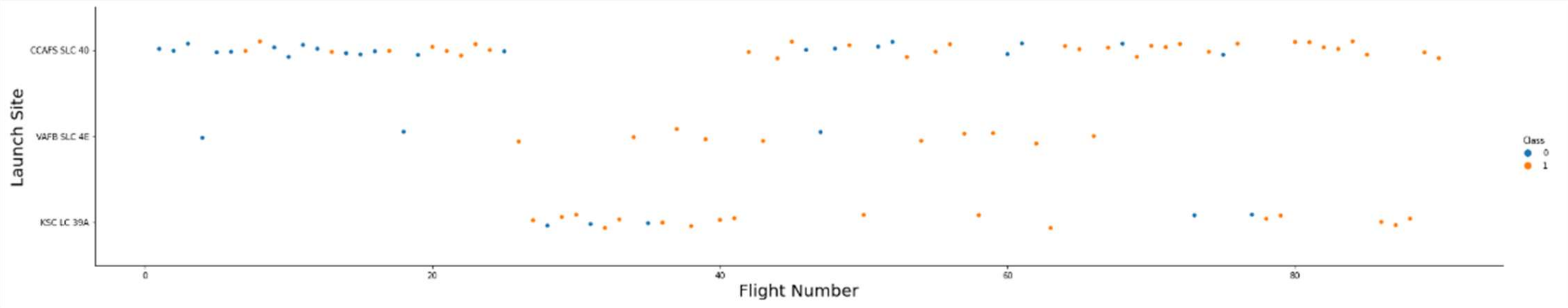
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

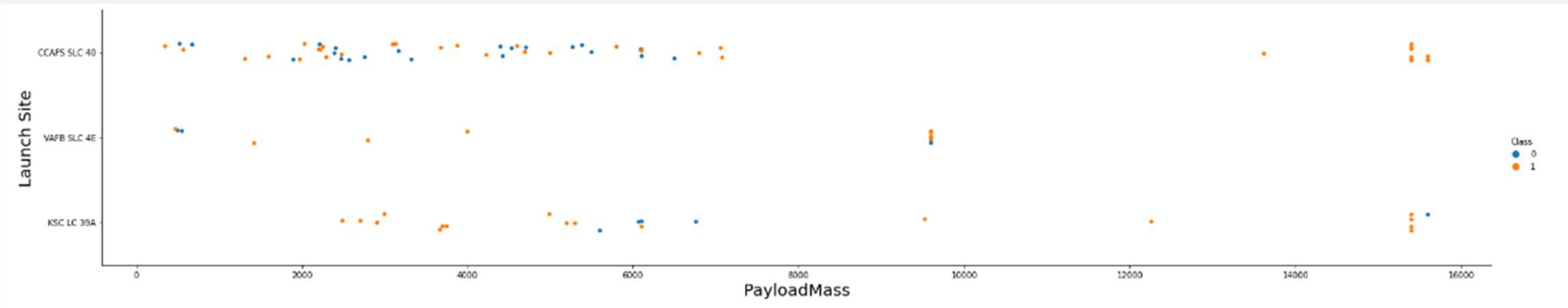
Flight Number vs. Launch Site

- scatter plot of Flight Number vs. Launch Site



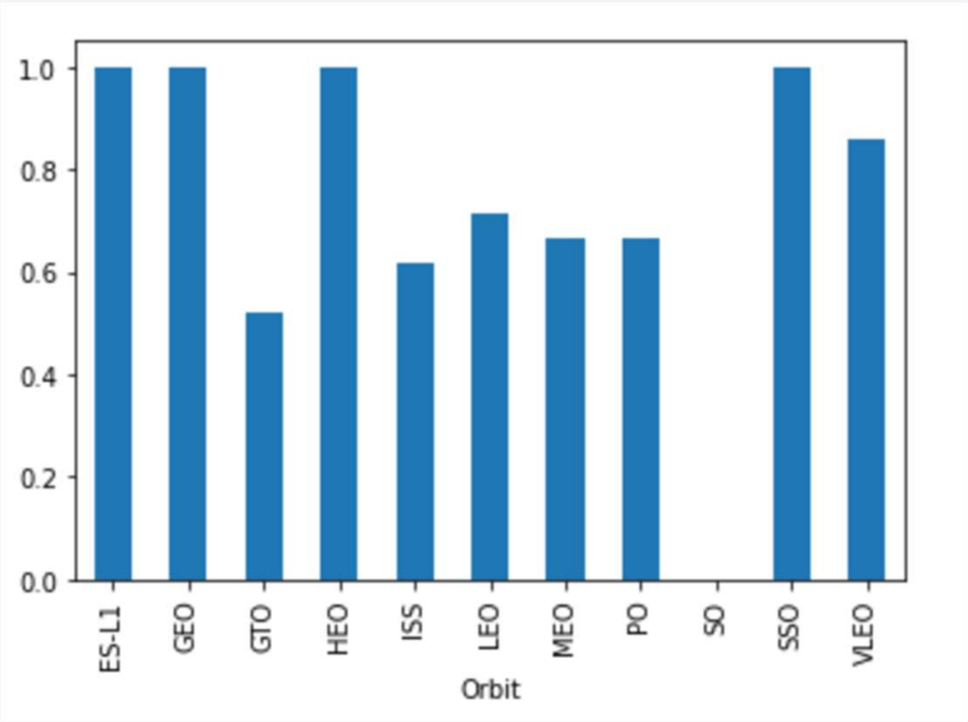
From the scatter plot we can see that the higher the flight number at each launch site, the more likely the launch being successful.

Payload vs. Launch Site



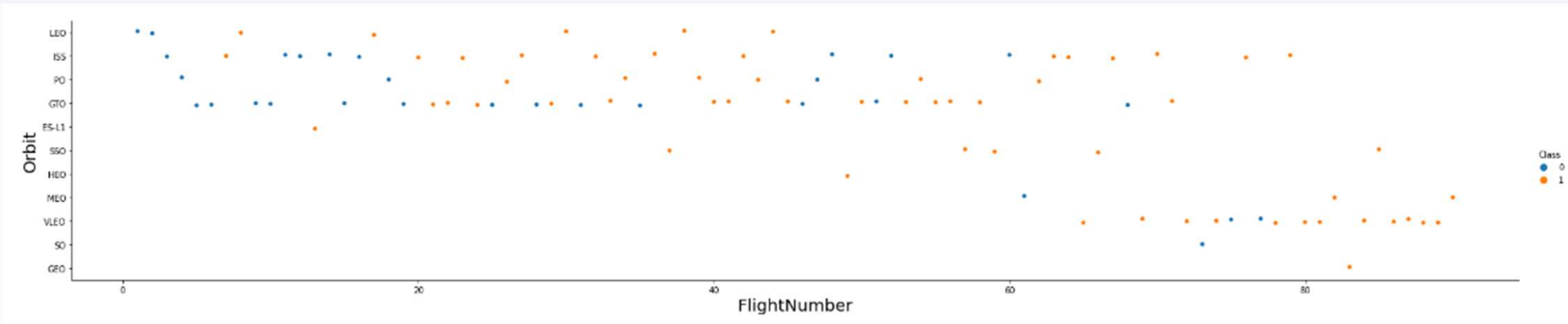
CCAPS SLC 40 has the highest no. of launches with smaller payload mass. The results suggests that the lower the payloadmass, the more likely the launch is successful, as evident in KSC LC 39A launch site.

Success Rate vs. Orbit Type



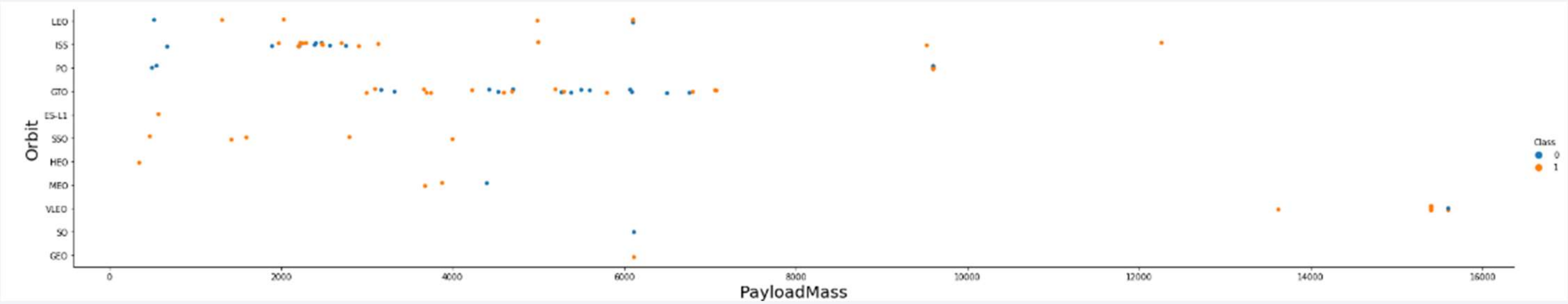
High success rate can be observed for orbit typr ES-L1, GEO , HEO and SSO.

Flight Number vs. Orbit Type



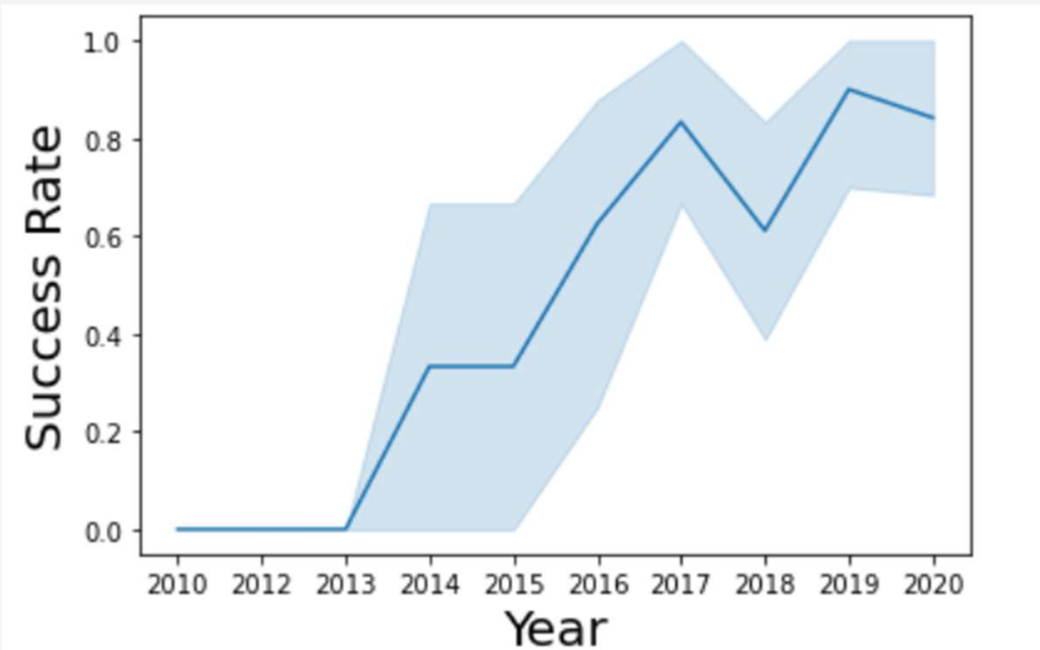
For LEO Orbit the higher the flight number the more likely the successful outcome of the launch. However there appears to be no relationship between the flight number and successful outcome for the case of GTO orbit.

Payload vs. Orbit Type



Successful or positive landing appears to have relationship with heavier payload for Polar, LEO and ISS. However the relationship between payload and successful landing is not evident/clear in the case of GTO.

Launch Success Yearly Trend



From the line chart we can see evidently that the no. of successful launches increased over the years in generally towards 2020. However there is a sharp decrease in success rate between 2017-2018, before the success rate climbed again between 2018 – 2019.

All Launch Site Names

SQL query to extract all unique launch site names

```
%sql select distinct (Launch_Site) from SPACEXTBL;
```

The SQL query selects all distinct launch site names from the column 'Launch_Site' in the SPACEXTBL table.

Query results

Out[7]:

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

There are four distinct launch sites in the table

Launch Site Names Begin with 'CCA'

<u>SQL code to extract 5 record where launch sites begin with ‘CCA’</u>

```
%sql select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5 ;
```

The SQL query selects all columns from SPACEXTBL table where launch site name begins with CCA, limited to first 5 results/rows.

<u>Query results</u>

Out[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

SQL code to aggregate payload_mass_kg column belonging to customer NASA (CRS)

```
In [9]:
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL WHERE Customer = "NASA (CRS)";
```

The SQL query sums up all entries in Payload Mass KG column for rows where Customer is NASA (CRS)

Query results

Out[9]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

<u>SQL query to calculate the average payload mass carried by booster version F9 v1.1</u>

In [10]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL WHERE "Booster_Version" = "F9 v1.1";
```

The SQL query calculates the average of the payload mass for rows where Booster Version is F9 v1.1

Query results

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
In [11]:  %%sql
          SELECT Min(Date)
          FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (ground pad)";
```

The SQL query filters the rows where landing outcome is 'Success (ground pad) and then selects from the Date column the minimum date (smallest date)

Query results

Out[11]:

Min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

In [12]:
```
%%sql
select "Booster_Version"
from SPACEXTBL
where "Landing _Outcome" = "Success (drone ship)" and "PAYLOAD_MASS__KG_" < 6000 and "PAYLOAD_MASS
```

The SQL query filters the rows where landing outcome is ‘Success (drone ship) and payload mass between 4000 and 6000, then returns all results from Booster_Version column.

Query results

Out[12]:

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

Total Number of Successful and Failure Mission Outcomes

In [13]:
```
%%sql
SELECT distinct ("Mission_Outcome"), count(Mission_Outcome)
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

The SQL query first group all records in the SPACEXTBL by Mission_Outcome and then count the no. of records for each distinct Mission_Outcome in the Mission Outcome column.

Query results

Out[13]:

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

Boosters Carried Maximum Payload

In [14]:
```
%%sql
SELECT "Booster_Version","PAYLOAD_MASS__KG_"
FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

The SQL query runs a sub-query to return the maximun payload value from the table SPACEXTBL, and then returns the Booster_Version and Payload Mass result from the table where payload mass is the maximum value.

Query results

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

```
In [15]:
%%sql
SELECT substr(Date, 4,2) as "month_names","Booster_Version", "Launch_Site"
FROM SPACEXTBL
WHERE substr(Date,7,4)='2015' AND "Landing _Outcome" = "Failure (drone ship)"  ;
```

The SQL query first filters all records where the year of the Date column is 2015 using the substring extraction method and landing outcome is 'Failure (drone ship)'. From the filtered results the query then extract and return the month in the Date column (using sub-string extraction method) of the filtered rows as month_names , and the corresponding Booster_vesion and "Launch Site" from the Booster Version and Launch Site columns.

Query results

Out[15]:

month_names	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

In [19]:

```
%%sql

SELECT "LANDING _OUTCOME", COUNT(*) AS Number
FROM SPACEXTBL WHERE "LANDING _OUTCOME" LIKE '%Success%' AND DATE > '04-06-2010'
AND DATE < '20-03-2017'
GROUP BY "LANDING _OUTCOME" ORDER BY COUNT(*) DESC;
```

The SQL query first filter all records where landing outcome contains ‘success’ description and Date falls between the mentioned dates above. From the filtered rows it then group the rows by ‘landing outcome’ , then performs a count of the landing outcomes and then order the results in descending order.

Query results

Out[19]:

Landing _Outcome	Number
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

Launch Site locations in the USA

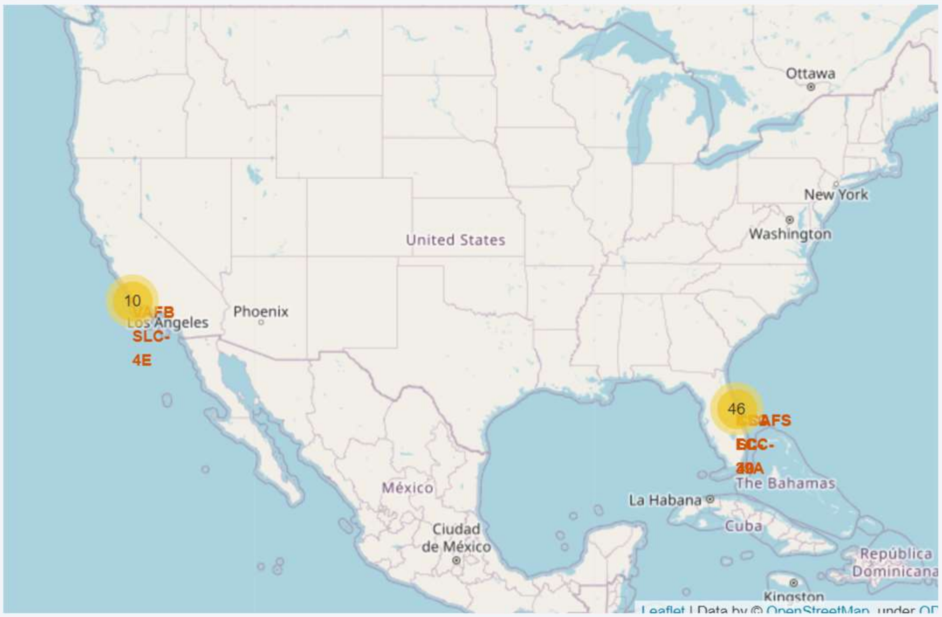


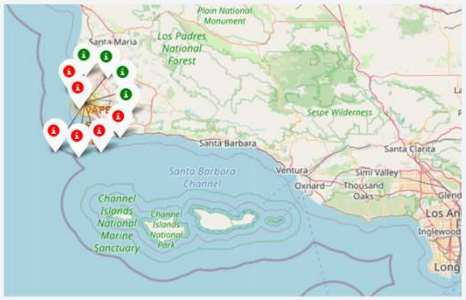
Figure 1.1

Key Findings

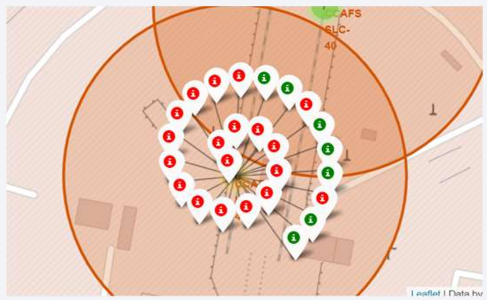
1. Launch Sites are located along the coastal line in both East and Western part of the USA.
2. There appears to be more launches conducted on the eastern coast of the USA than the west coast. This confirms the understanding that launching rockets towards the east tap into the earth rotational speed and offer additional boost in the initial phase of the launch.
3. Both launch sites are located closer to the equator which confirms the understanding that land is moving faster at the equator than any other place on the surface of the Earth, hence the rocket initial launch counteracts any gravitational pull from the earth.

Success/Failed launches at launch sites

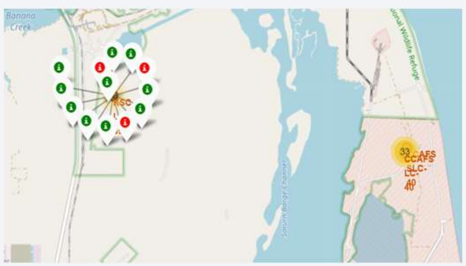
WEST COAST –VAFB SLC-4E



EAST COAST –CCAFS LC 40



EAST COAST –KSC LC 39A



EAST COAST –CCAFS SLC 40

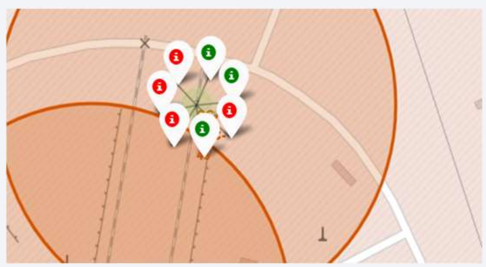


Figure 1.2

Findings

1. Based on reasoning in the previous slide we will expect more successful launches to be recorded in the Eastern coast than western coast.
2. However based on the figure 1.2 in the left, the successful launches recorded (marked in green)(with the exception of KSC LC39A) seem to suggest however that other factors have to be investigated further for low count of successful launches recorded in **CCAFS LC 40 and CCAFS SLC 40 .**

Proximity of infrastructures to launch sites

Distance from launch site to railway ,coastline and highway

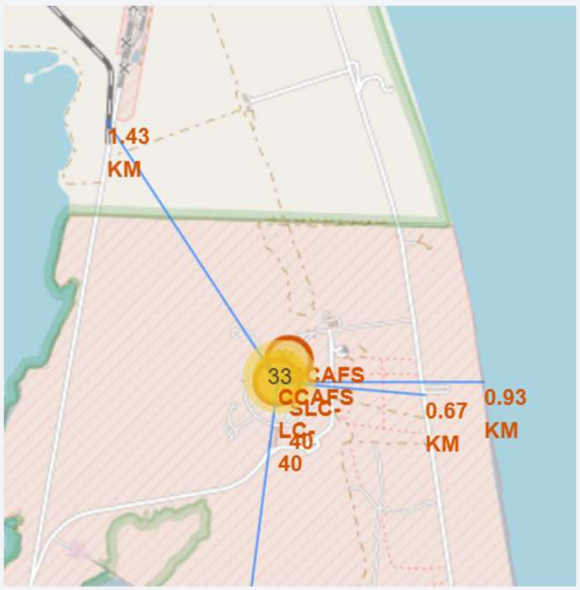


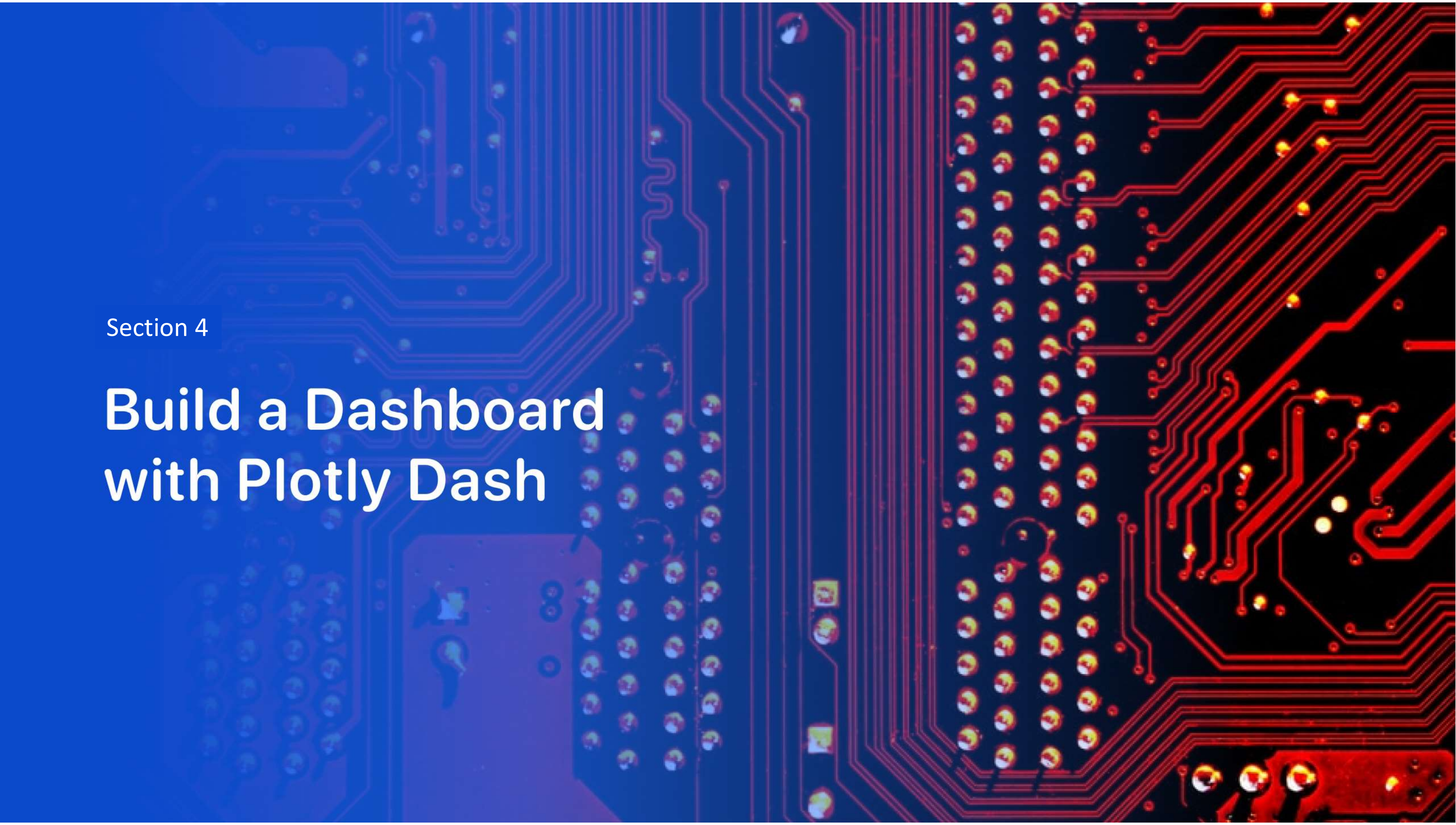
Figure 1.3

Distance from launch site to the nearest city.

Figure 1.4

From figure 1.3, we confirm that launch sites in general are close to access routes (1.43km from railway to CCAFS SLC-40 ,0.67 km from highway and 0.93km from coastline) . Proximity to such access routes facilitate the transportation of the rockets and other logistics cargos to the launch sites.

Figure 1.4 confirms that launch sites in general are far away from the city area (18.07km from Cape Canaveral), so as to minimize any noise pollution and also for safety considerations.



Section 4

Build a Dashboard with Plotly Dash

Total successful launch(%) for all sites

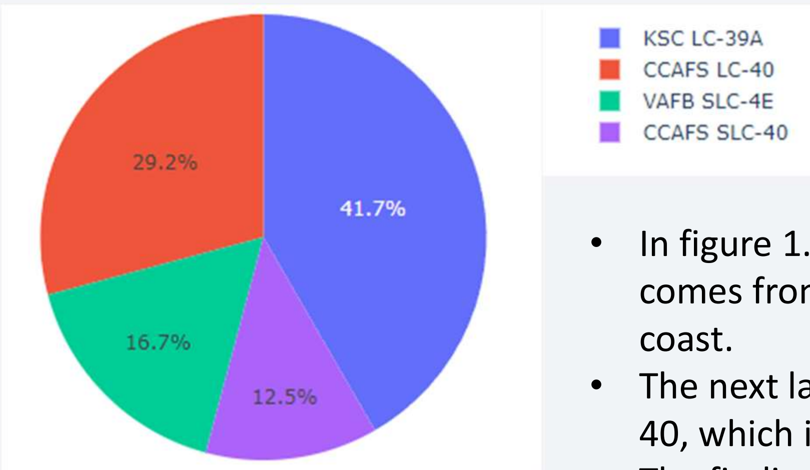
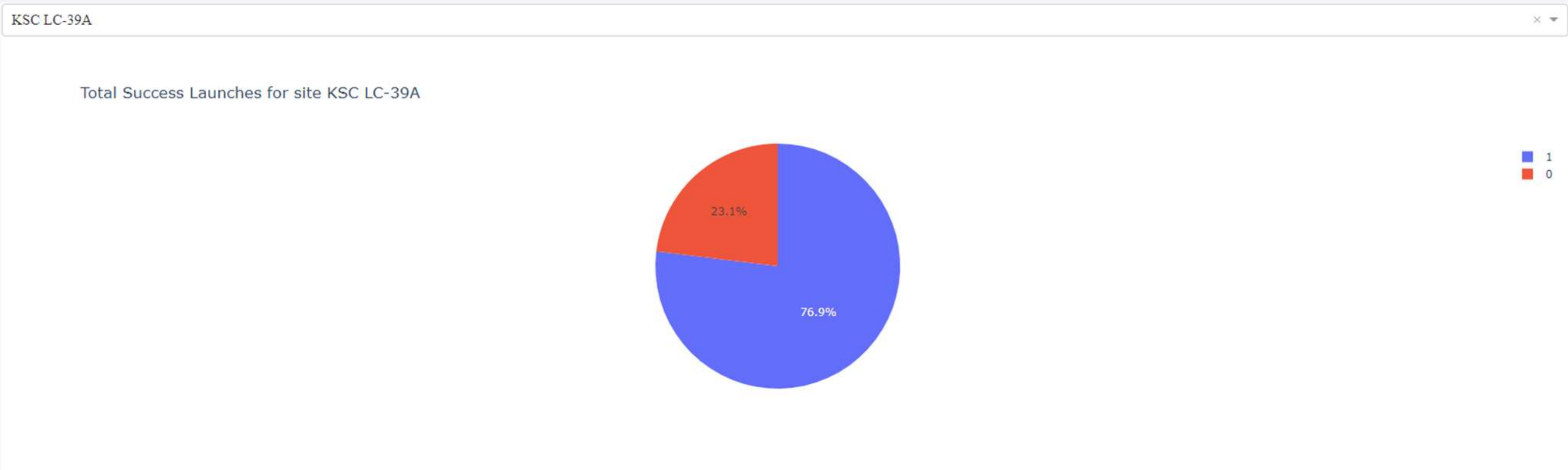


Figure 1.5

- In figure 1.5 we can see that the highest count of successful launch comes from KSC LC-39A which is a launch site located in Eastern coast.
- The next launch site with the highest successful launch is CCAFS LC-40, which is also located in the eastern coast.
- The findings above offers some evidence and support that earth rotational speed towards the east and surface travel speed near the equator contributes to success of the rocket launches.

Split of launch outcomes (%) for KSC LC-39A

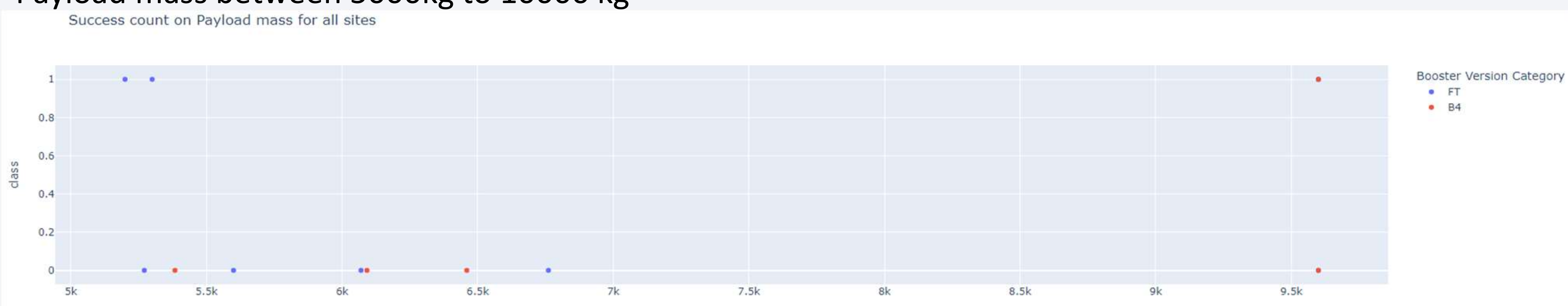


KSC LC-39A with the highest successful launch amongst all four launch sites achieved 76.90% success rate with 23.10% failure rate.

Relationship between payload mass and launch outcome

Payload mass between 0 to 5000 kg

Payload mass between 5000kg to 10000 kg

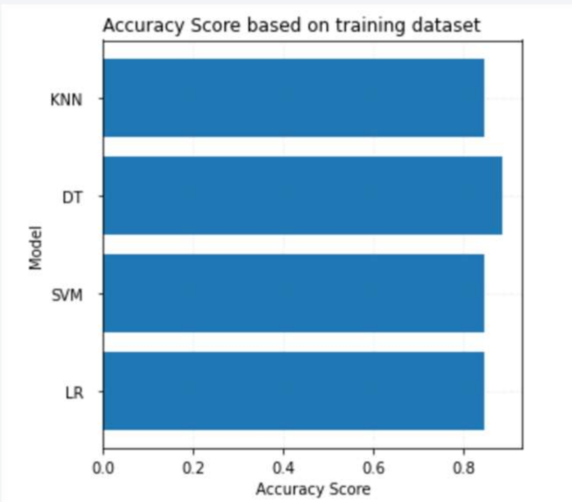


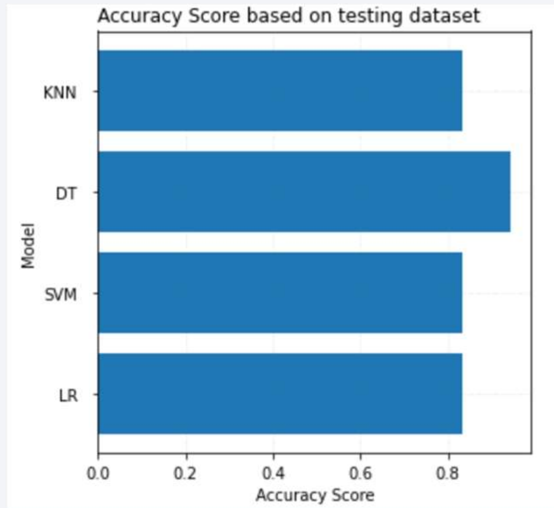
Evidently from the above counts payload mass lower than 5000 kg has a higher count of success launch than those with payloads greater than 5000kg. This confirms that the lower the payload mass, the higher the successful launch outcome

Section 5

Predictive Analysis (Classification)

Classification Accuracy

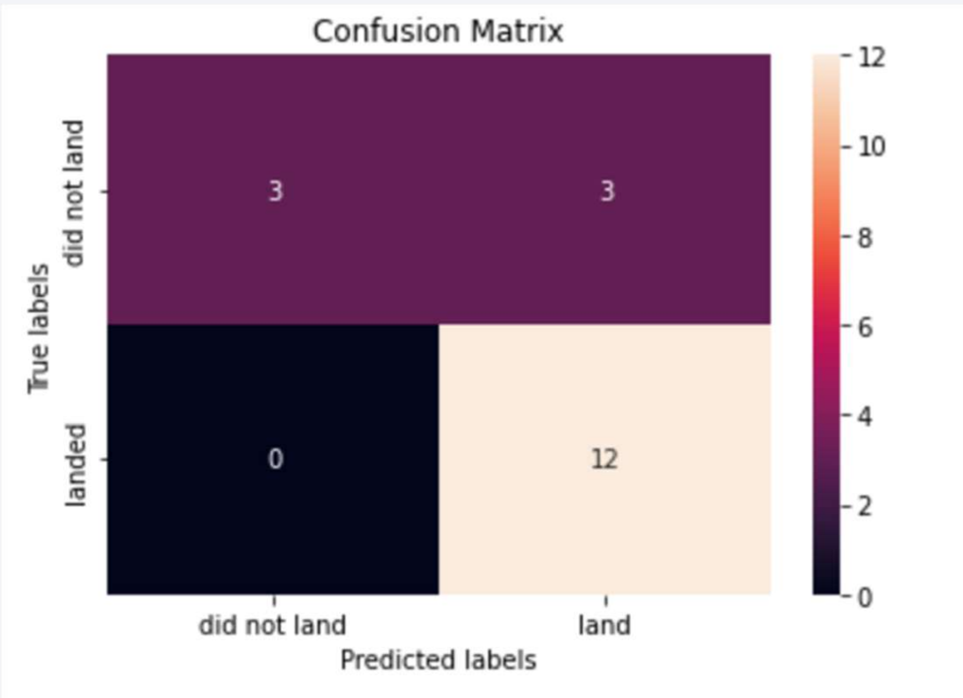




	model	test_Score
0	LR	0.833333
1	SVM	0.833333
2	DT	0.944444
3	KNN	0.833333

Based on training dataset our decision tree model returned the highest accuracy of 0.8875 amongst all four models. Likewise on the test dataset our DT returned the highest accuracy of 0.94444. On

Confusion Matrix



The total no. of test dataset is 18. As per the confusion matrix for Decision Tree Model,
False Positive = 3
True Positive = 12
False Negative = 0
True Negative = 3

Our concern here is with the false positive of 3.

Conclusions

Based on our results in this presentation we conclude the following;

1. the higher the flight number at each launch site, the more likely the launch being successful.
2. the lower the payloadmass, the more likely the launch is successful, as evident in KSC LC 39A launch site
3. Orbit type ES-L1, GEO , HEO and SSO tend to exhbit higher High success rate.
4. In general launch sites in the eastern coast recorded ranked the highest in terms of successful launches. This lend support to the generalized theory that the lower gravitational pull closer to the equator and higher surface speed and earth velocity in the East direction facilitates rocket launches in the East.
5. Decision Tree model with the highest accuracy score both using training and test dataset is the preferred machine learning model for predicting the outcome of a rocket launch in this particular project.
6. Launch sites are generally located near to infrastructures that facilitates logistics arrangement. They are also far away from nearby cities to minimize any noise pollution and for safety reasons.

Thank you!