DATA SCIENCE CAPSTONE PROJECT

SPACE X

BY Arishkumar N

OUTLINE

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDI

EXECUTIVE SUMMARY

- SUMMARY OF METHODOLOGIES
  - DATA COLLECTION
  - DATA WRANGLING
  - EXPLORATORY DATA ANALYSIS WITH DATA VISUALIZATION
  - EXPLORATORY DATA ANALYSIS WITH SQL
  - BUILDING AN INTERACTIVE MAP WITH FOLIUM
  - BUILDING A DASHBOARD WITH PLOTLY DASH
  - PREDICTIVE ANALYSIS (CLASSIFICATION)

- SUMMARY OF ALL RESULTS
  - EXPLORATORY DATA ANALYSIS RESULTS
  - INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS
  - PREDICTIVE ANALYSIS RESULTS

INTRODUCTION

SpaceX is the most successful company of the commercial spaceage, making space travel affordable. The company advertises Falcon9 rocket launches on its website, with a cost of 62 million dollars;other providers cost upward of 165 million dollars each, much of thesavings is because SpaceX can reuse the First stage. Therefore, if wecan determine if the First stage will land, we can determine the costof a launch. Based on public information and machine learningmodels, we are going to predict if SpaceX will reuse the First stage

METHODOLOGY

SPACEX IS THE MOST SUCCESSFUL COMPANY OF THE COMMERCIAL SPACEAGE, MAKING SPACE TRAVEL AFFORDABLE. THE COMPANY ADVERTISES FALCON9 ROCKET LAUNCHES ON ITS WEBSITE, WITH A COST OF 62 MILLION DOLLARS;OTHER PROVIDERS COST UPWARD OF 165 MILLION DOLLARS EACH, MUCH OF THESAVINGS IS BECAUSE SPACEX CAN REUSE THE FIRST STAGE. THEREFORE, IF WECAN DETERMINE IF THE FIRST STAGE WILL LAND, WE CAN DETERMINE THE COSTOF A LAUNCH. BASED ON PUBLIC INFORMATION AND MACHINE LEARNINGMODELS, WE ARE GOING TO PREDICT IF SPACEX WILL REUSE THE FIRST STAGE

DATA COLLECTION METHODOLOGY:

USING SPACEX REST API

USING WEB SCRAPPING FROM WIKIPEDIAPERFORMED DATA WRANGLING

FILTERING THE DATA- DEALING WITH MISSING VALUES

USING ONE HOT ENCODING TO PREPARE THE DATA TO A BINARY CLASSIFICATION PERFORMED EXPLORATORY DATA ANALYSIS (EDA)

USING VISUALIZATION AND SQLPERFORMED INTERACTIVE VISUAL ANALYTICS USING FOLIUM AND PLOTLY DASHPERFORMED PREDICTIVE ANALYSIS

USING CLASSIFICATION MODELS

BUILDING, TUNING AND EVALUATION OF CLASSIFICATION MODELS TO ENSURE THE BESTRESULTS

DATA COLLECTION

Data collection process involved a combination of API requests from SpaceX RESTAPI and Web Scraping data from a table in SpaceX's Wikipedia entry.We had to use both of these data collection methods in order to get completeinformation about the launches for a more detailed analysis

Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite,Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount,Serial, Longitude, Latitude

Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launchoutcome, Version Booster, Booster landing, Date, Time

DATA WRANGLING

IN THE DATA SET, THERE ARE SEVERAL DIFFERENT CASES WHERE THEBOOSTER DID NOT LAND SUCCESSFULLY. SOMETIMES A LANDING WASATTEMPTED BUT FAILED DUE TO AN ACCIDENT; FOR EXAMPLE, TRUEOCEAN MEANS THE MISSION OUTCOME WAS SUCCESSFULLY LANDEDTO A SPECI

IC REGION OF THE OCEAN WHILE FALSE OCEAN MEANSTHE MISSION OUTCOME WAS UNSUCCESSFULLY LANDED TO A SPECI

ICREGION OF THE OCEAN. TRUE RTLS MEANS THE MISSION OUTCOMEWAS SUCCESSFULLY LANDED TO A GROUND PAD FALSE RTLS MEANSTHE MISSION OUTCOME WAS UNSUCCESSFULLY LANDED TO A GROUNDPAD.TRUE ASDS MEANS THE MISSION OUTCOME WAS SUCCESSFULLYLANDED ON A DRONE SHIP FALSE ASDS MEANS THE MISSIONOUTCOME WAS UNSUCCESSFULLY LANDED ON A DRONE SHIP.WE MAINLY CONVERT THOSE OUTCOMES INTO TRAINING LABELS WITH“1” MEANS THE BOOSTER SUCCESSFULLY LANDED, “0” MEANS IT WASUNSUCCESSFUL.

EDA WITH DATA VISUALIZATION

CHARTS WERE PLOTTED:
FLIGHT NUMBER VS. PAYLOAD MASS, FLIGHT NUMBER VS. LAUNCH SITE, PAYLOAD MASSVS. LAUNCH SITE, ORBIT TYPE VS. SUCCESS RATE, FLIGHT NUMBER VS. ORBIT TYPE,PAYLOAD MASS VS ORBIT TYPE AND SUCCESS RATE YEARLY TRENDSCATTER PLOTS SHOW THE RELATIONSHIP BETWEEN VARIABLES. IF A RELATIONSHIP EXISTS,THEY COULD BE USED IN MACHINE LEARNING MODEL.BAR CHARTS SHOW COMPARISONS AMONG DISCRETE CATEGORIES. THE GOAL IS TO SHOW THERELATIONSHIP BETWEEN THE SPECI

IC CATEGORIES BEING COMPARED AND A MEASUREDVALUE.LINE CHARTS SHOW TRENDS IN DATA OVER TIME (TIME SERIES).

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

IRST 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 BUTLESS 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 INYEAR 2015•RANKING THE COUNT OF LANDING OUTCOMES (SUCH AS FAILURE (DRONE SHIP) OR SUCCESS (GROUND PAD)) BETWEEN THE DATE2010-06-04 AND 2017-03-20 IN DESCENDING ORDER

BUILD AN INTERACTIVE MAP WITH FOLIUM

MARKERS OF ALL LAUNCH SITES:

- ADDED MARKER WITH CIRCLE, POPUP LABEL AND TEXT LABEL OF NASA JOHNSON SPACE CENTER USINGITS LATITUDE AND LONGITUDE COORDINATES AS A START LOCATION.- ADDED MARKERS WITH CIRCLE, POPUP LABEL AND TEXT LABEL OF ALL LAUNCH SITES USING THEIR LATITUDEAND LONGITUDE COORDINATES TO SHOW THEIR GEOGRAPHICAL LOCATIONS AND PROXIMITY TO EQUATOR ANDCOASTS.

COLOURED MARKERS OF THE LAUNCH OUTCOMES FOR EACH LAUNCH SITE:

- ADDED COLOURED MARKERS OF SUCCESS (GREEN) AND FAILED (RED) LAUNCHES USING MARKER CLUSTER TOIDENTIFY WHICH LAUNCH SITES HAVE RELATIVELY HIGH SUCCESS RATES.

DISTANCES BETWEEN A LAUNCH SITE TO ITS PROXIMITIES:

- ADDED COLOURED LINES TO SHOW DISTANCES BETWEEN THE LAUNCH SITE KSC LC-39A (AS ANEXAMPLE) AND ITS PROXIMITIES LIKE RAILWAY, HIGHWAY, COASTLINE AND CLOSEST CITY.

BUILD A DASHBOARD WITH PLOTLY DASH

LAUNCH SITES DROPDOWN LIST:

- ADDED 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 THESUCCESS 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 DI FF

ERENT BOOSTER VERSIONS:

- ADDED A SCATTER CHART TO SHOW THE CORRELATION BETWEEN PAYLOAD AND LAUNCH SUCCESS.

PREDICTIVE ANALYSIS CLASSIFICATION)

CREATING A NUMPYARRAY FROM THE COLUMN“CLASS” IN DATASTANDARDIZING THEDATA WITHSTANDARDSCALER, THEN

ITTING ANDTRANSFORMING ITSPLITTING THE DATA INTOTRAINING AND TESTINGSETS WITHTRAIN_TEST_SPLITFUNCTIONCREATING AGRIDSEARCHCV OBJECTWITH CV = 10 TO

INDTHE BEST PARAMETERSAPPLYINGGRIDSEARCHCVON LOGREG, SVM,DECISION TREE, ANDKNN MODELSCALCULATING THEACCURACY ON THE TESTDATA USING THEMETHOD .SCORE()FOR ALL MODELSEXAMINING THECONFUSION MATRIXFOR ALL MODELSFINDING THE METHODPERFORMS BEST BYEXAMINING THEJACCARD_SCORE ANDF1_SCORE METRICS
GITHUB URL: MACHINE LEARNING PREDICTION

RESULTS

- EXPLORATORY DATA ANALYSIS RESULTS
- INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS
- PREDICTIVE ANALYSIS RESULTS

FLIGHT NUMBER VS. LAUNCH SITE

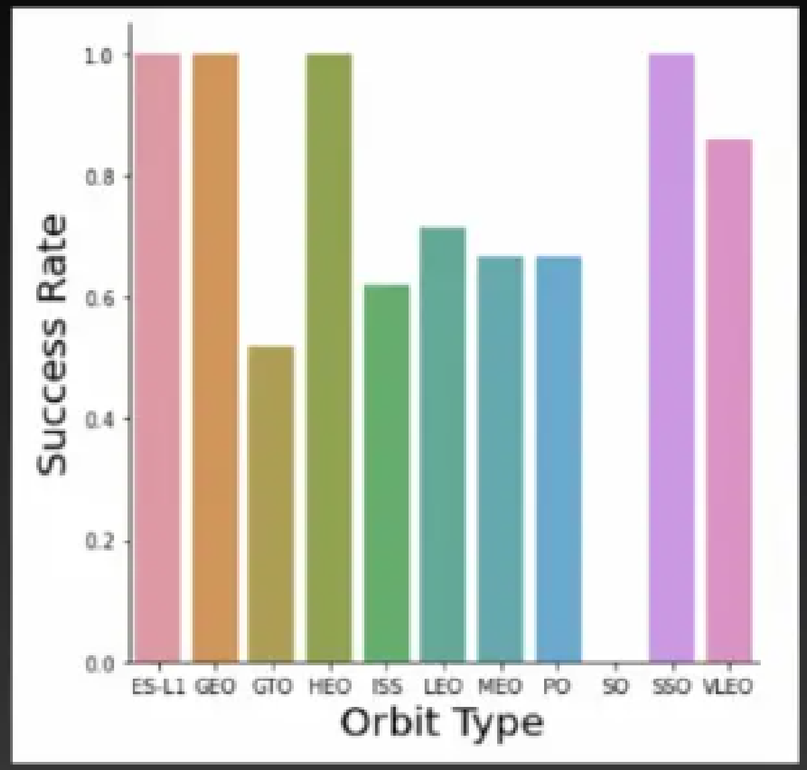
EXPLANATION:

•THE EARLIESTLIGHTS ALL FAILED WHILE THE LATEST LIGHTS ALL SUCCEEDED.•THE CCAFS SLC 40 LAUNCH SITE HAS ABOUT A HALF OF ALL LAUNCHES.•VAFB SLC 4E AND KSC LC 39A HAVE HIGHER SUCCESS RATES.•IT CAN BE ASSUMED THAT EACH NEW LAUNCH HAS A HIGHER RATE OF SUCCESS.

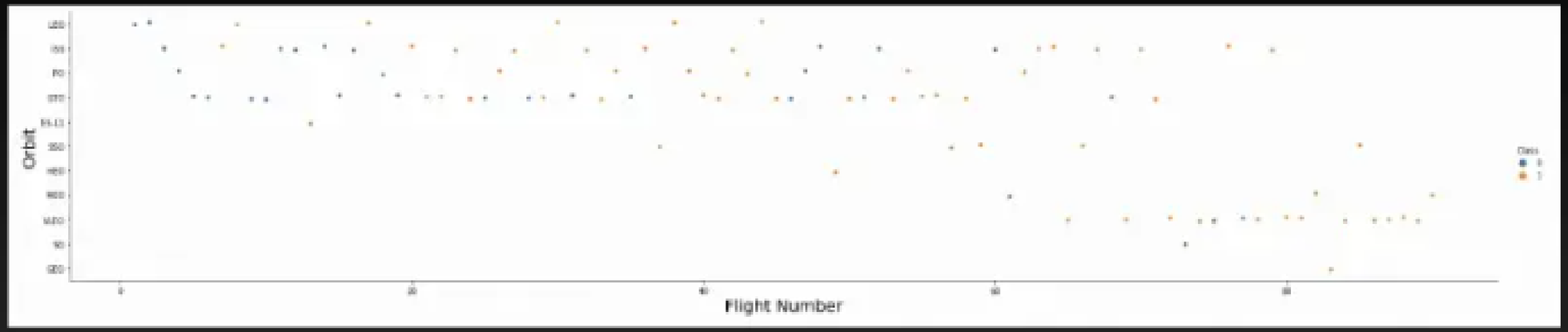
Success rate vs. Orbit type

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



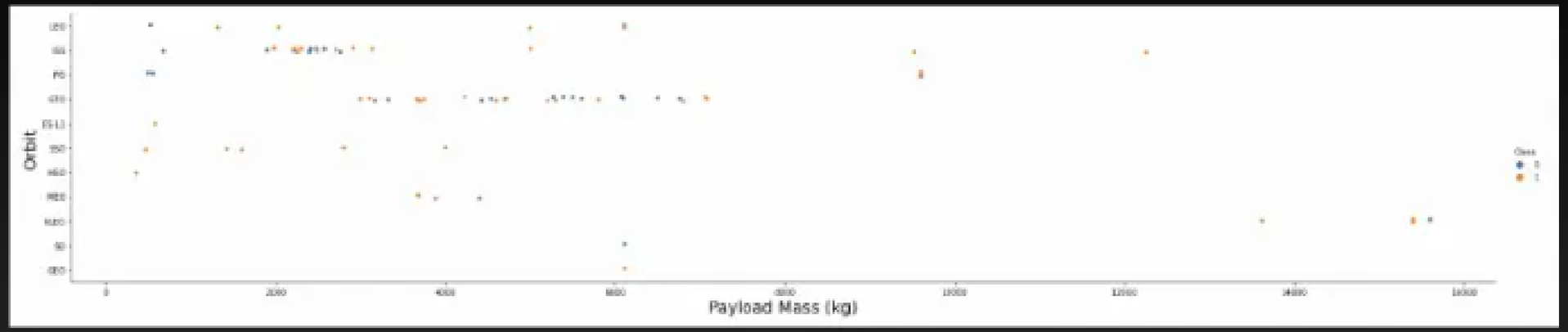
Flight Number vs. Orbit type



Explanation:

- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload Mass vs. Orbit type



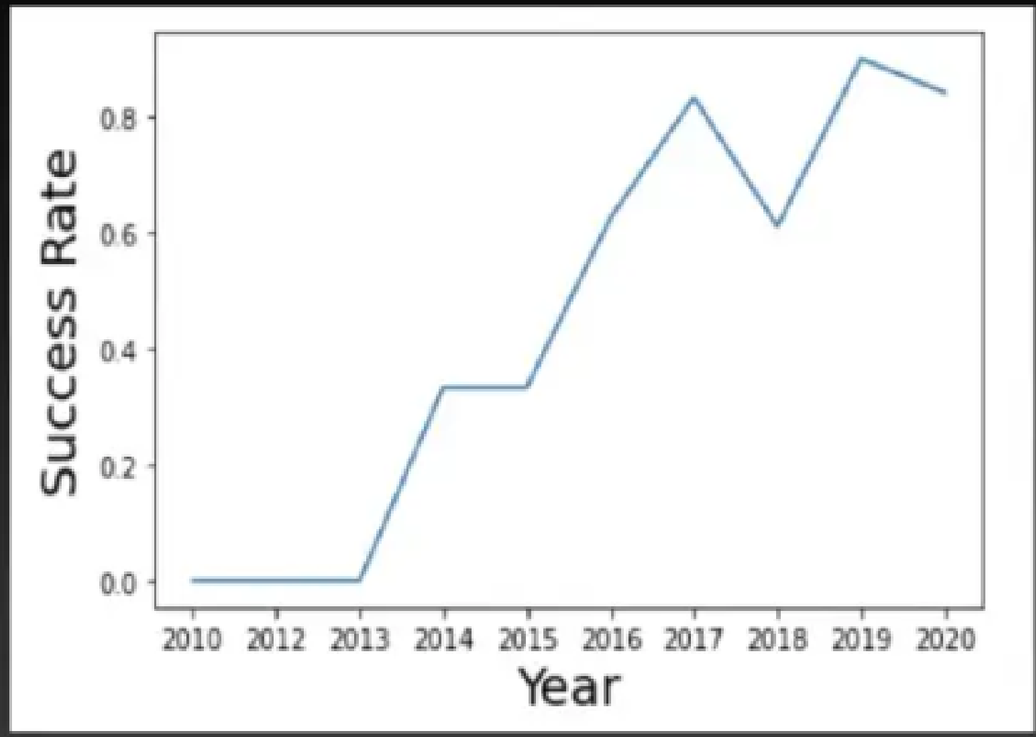
Explanation:

- 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.



EDA WITH SQL

All launch site names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Explanation:

- Displaying the names of the unique launch sites in the space mission.

Launch site names begin with `CCA`

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

Explanation:

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

Total payload mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

Explanation:

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

Average payload mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

Explanation:

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

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

Out[10]:

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

Explanation:

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

2015 launch records

```
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.bs2io90108kqblod8lcg.databases.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)

Explanation:

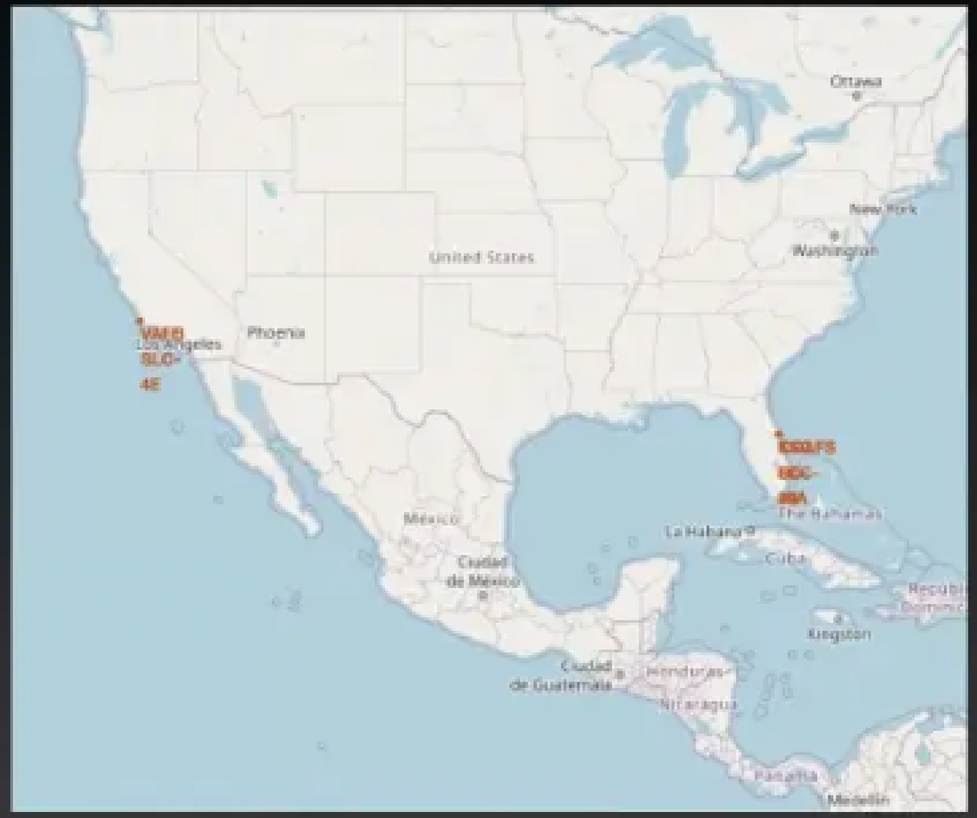
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

Interactive map with Folium

All launch sites' location markers on a global map

Explanation:

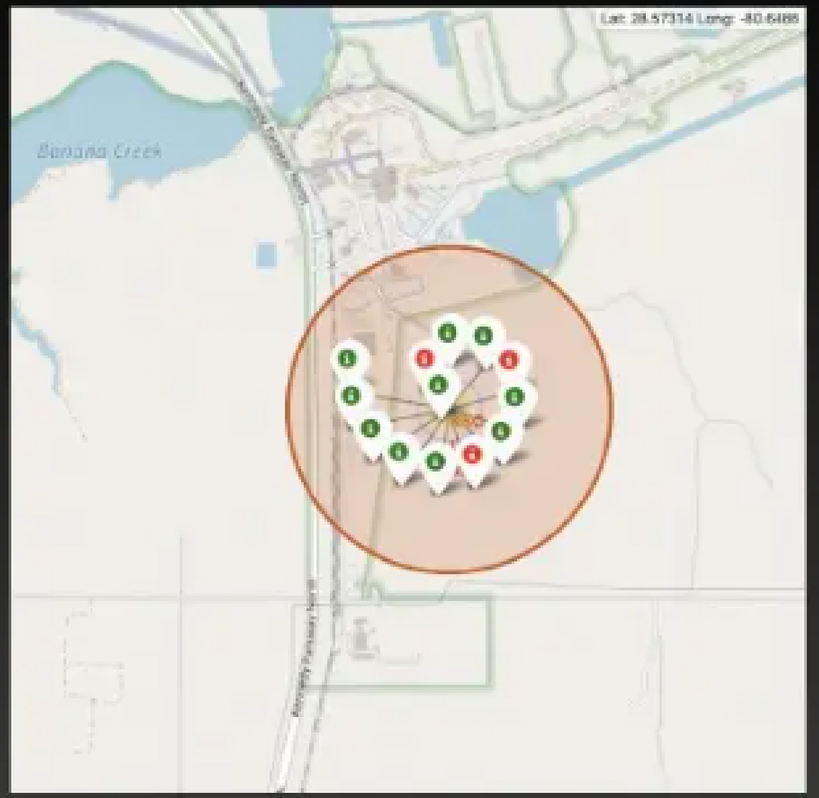
- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



Colour-labeled launch records on the map

Explanation:

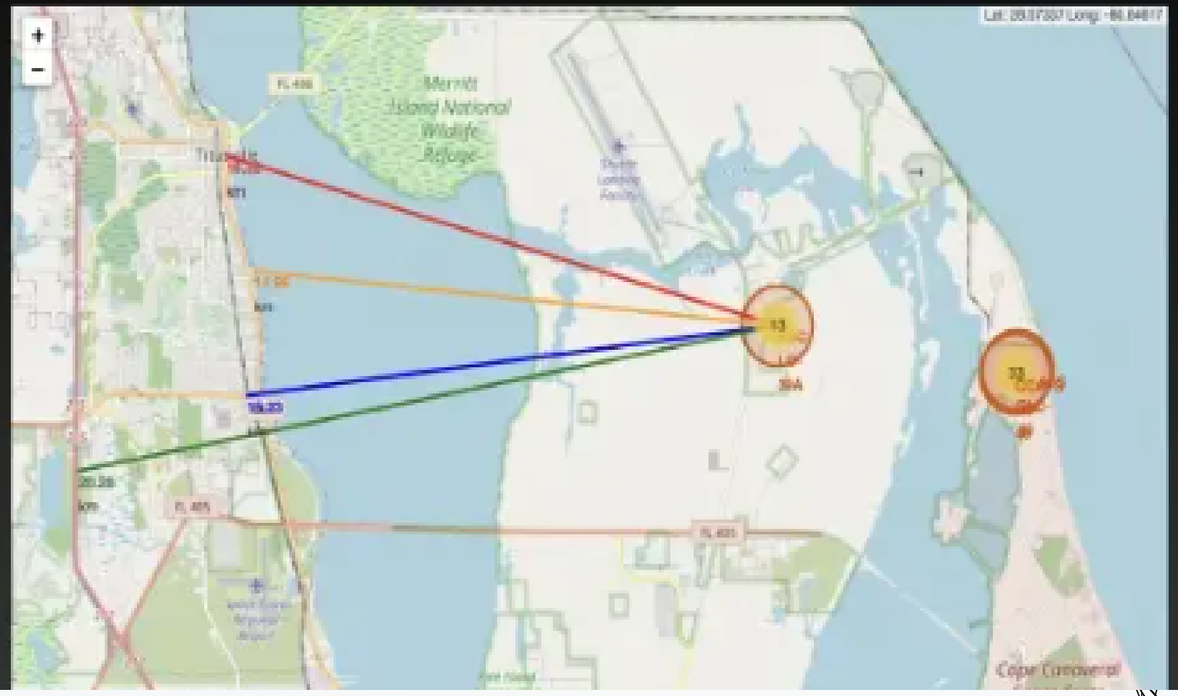
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.



Distance from the launch site KSC LC-39A to its proximities

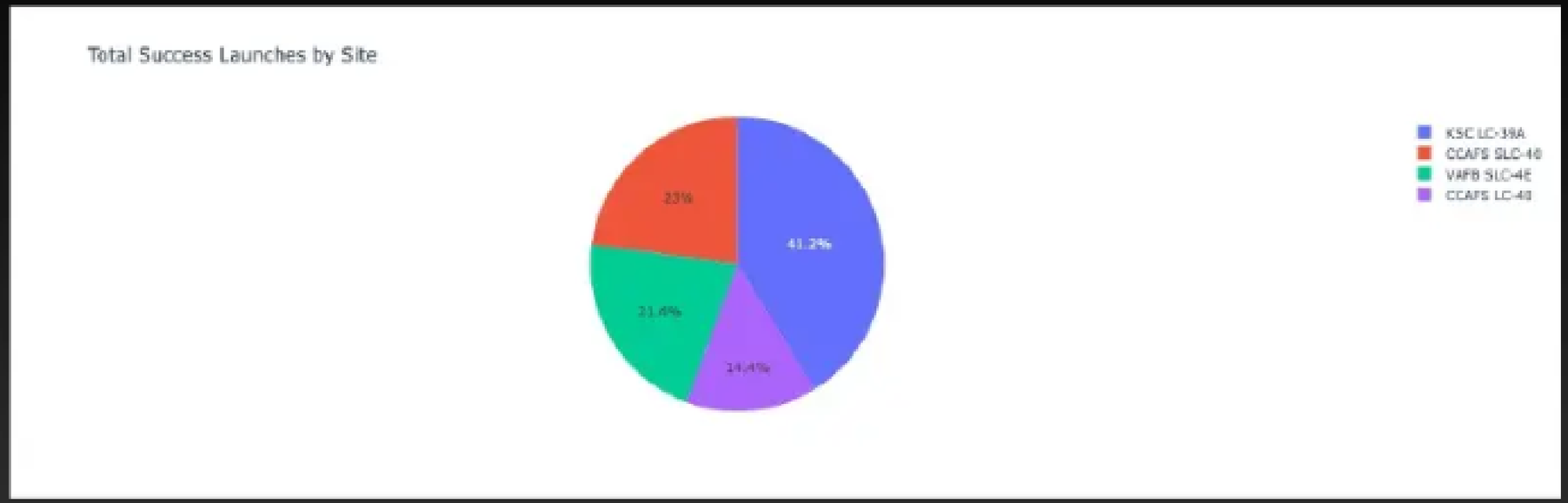
Explanation:

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially



Build a Dashboard with PlotlyDash

Launch success count for all sites



Explanation:

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

Payload Mass vs. Launch Outcome for all sites

Explanation:

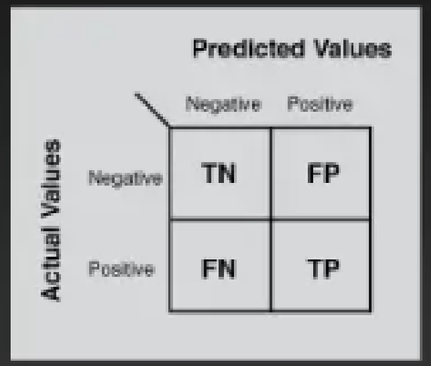
- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

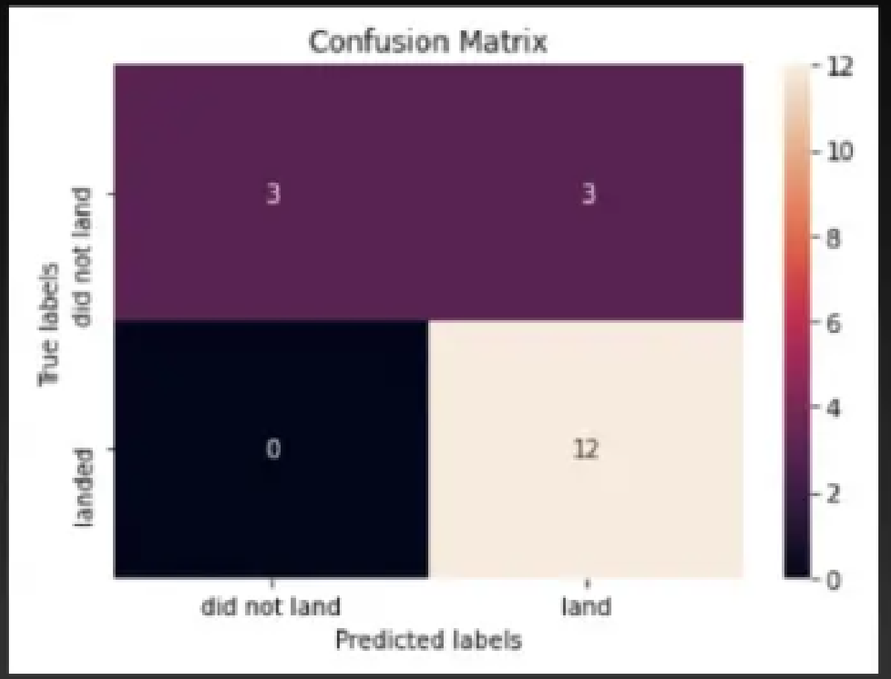
Predictive analysis(Classification)

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

DECISION TREE MODEL IS THE BEST ALGORITHM FOR THIS DATASET .•LAUNCHES WITH A LOW PAYLOAD MASS SHOW BETTER RESULTSTHAN LAUNCHES WITH A LARGER PAYLOAD MASS.

•MOST OF LAUNCH SITES ARE IN PROXIMITY TO THE EQUATOR LINEAND ALL THE SITES ARE IN VERY CLOSE PROXIMITY TO THE COAST.•THE SUCCESS RATE OF LAUNCHES INCREASES OVER THE YEARS.•KSC LC-39A HAS THE HIGHEST SUCCESS RATE OF THE LAUNCHESFROM ALL THE SITES.•ORBITS ES-L1, GEO, HEO AND SSO HAVE 100% SUCCESS RATE.

APPENDIX

SPECIAL THANKS TO:
INSTRUCTORS
COURSERA
IBM