

Yarram Gnaneswar reddy

21471A05P6



2

•

E

x

e

c

ut

i

v

e

S

u

mm

a

r

y

•

Introduction

•

Methodology

•

Results

•

Conclusion

•

Appendix

O

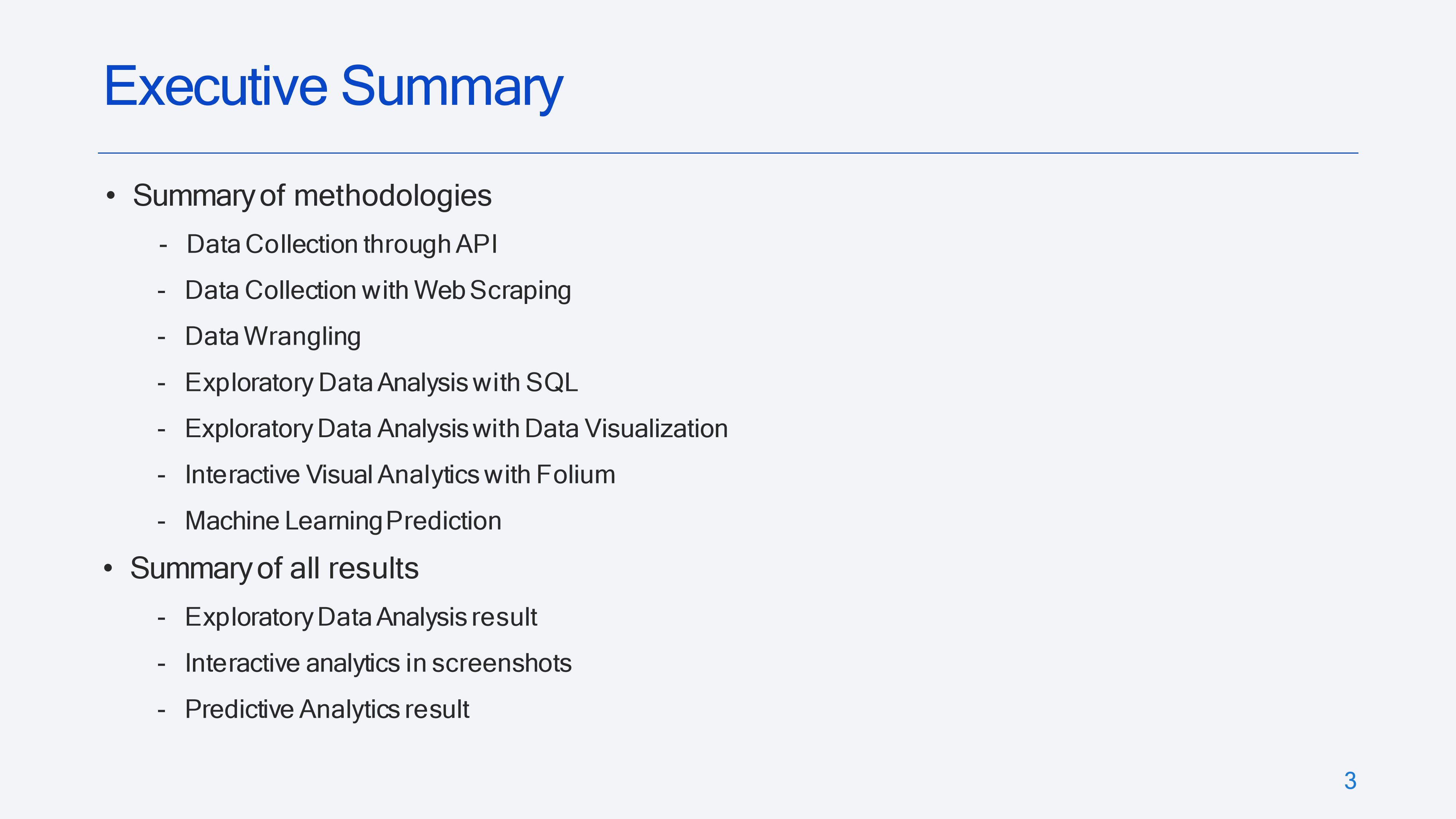
u

t

li

n

e





Introduction

•

P

r

o

j

e

c

t

ba

ck

g

r

oun

d

an

d

c

ont

e

x

t

Space

X

advertises

Falcon

9

rocket

launches

on

its

website

with

a

cost

of

62

million

dollars

;

other

providers

cost

upward

of

165

million

dollars

each,

much

of

the

savings

is

because

Space

X

can

reuse

the

first

stage

.

Therefore,

if

we

can

determine

if

the

first

stage

will

land,

we

can

determine

the

cost

of

a

launch

.

This

information

can

be

used

if

an

alternate

company

wants

to

bid

against

space

X

for

a

rocket

launch

.

This

goal

of

the

project

is

to

create

a

machine

learning

pipeline

to

predict

if

the

first

stage

will

land

successfully

.

•

Problems

you

want

to

find

answers

-

What

factors

determine

if

the

rocket

will

land

successfully?

-

The

interaction

amongst

various

features

that

determine

the

success

rate

of

a

successful

landing.

-

What

operating

conditions

needs

to

be

in

place

to

ensure

a

successful

landing

program

.

4



5



E

x

e

c

u

t

i

v

e

S

u

mm

a

r

y

•

D

at

a

c

o

ll

e

c

t

i

o

n

m

e

t

hodo

l

og

y

:

•

Data

was

collected

using

SpaceX

API

and

web

scraping

from

Wikipedia.

•

Perform

data

wrangling

•

One

-

hot

encoding

was

applied

to

categorical

features

•

Perform

exploratory

data

analysis

EDA

)

(

using

visualization

and

SQL

•

Perform

interactive

visual

analytics

using

Folium

and

Plotly

Dash

•

Perform

predictive

analysis

using

classification

models

•

How

to

build,

tune,

evaluate

classification

models

M

e

t

hodo

l

o

g

y



•

The

data

was

collected

using

various

methods

-

Data

collection

was

done

using

get

request

to

the

SpaceX

API.

-

Next,

we

decoded

the

response

content

as

a

Json

using

.json()

function

call

and

turn

it

into

a

pandas

dataframe

using

.json\_normalize().

-

We

then

cleaned

the data,

checked

for

missing

values

and

fill in

missing

values

where

necessary.

-

In

addition, we performed web

scraping

from

Wikipedia

for

Falcon

9

launch

records with

BeautifulSoup.

-

The

objective

was

to

extract

the

launch

records

as

HTML

table,

parse

the

table

and

convert

it

to

a

pandas

dataframe

for

future

analysis.

Data

Collection



•

We

used

the get

request

to the

SpaceX

API

to

collect

data,

clean

t

h

e

r

eque

s

t

ed

dat

a

an

d

d

i

d

s

o

m

e

basic

data

wrangling

and

formatting.

•

The

link

to

the

notebook

is

https://github.com/chuksoo/IBM

-

Data

-

Science

-

Capstone

-

SpaceX/blob/main/Data%20Collect

ion%20API.ipynb.

Data

Collection

–

SpaceX

API



•

W

e

app

li

ed

we

b

sc

r

app

i

n

g

t

o

webscrap

Falcon

9

launch

records

with

BeautifulSoup

•

We

parsed

the

table

and

converted

it

into

a

pandas

dataframe.

•

The

link

to

the

notebook

is

https://github.com/chuksoo/IBM

-

Data

-

Science

-

Capstone

-

SpaceX/blob/main/Data%20Collect

ion%20with%20Web%20Scraping

.ipynb.

Data

Collection

-

Scraping



•

We

performed

exploratory

data

analysis

an

d

de

t

e

r

m

i

ne

d

t

he

t

r

a

i

n

i

n

g

l

abe

l

s

.

•

We

calculated

the

number

of

launches

at

each

site,

and the

number

and

occurrence

of

each

orbits

•

W

e

c

r

eat

ed

l

and

i

n

g

ou

t

c

o

m

e

l

abe

l

f

r

o

m

out

c

o

m

e

c

o

l

u

m

n

an

d

e

x

po

r

t

e

d

t

he

r

e

s

u

l

t

s

to

csv.

•

The

link

to

the

notebook

is

https://github.com/chuksoo/IBM

-

Data

-

Science

-

Capstone

-

SpaceX/blob/main/Data%20Wrangling.ip

ynb.

Data

Wrangling



•

W

e

e

x

p

l

o

r

e

d

t

he

dat

a

b

y

v

i

s

ua

li

z

i

ng

t

h

e

r

e

l

a

t

i

on

s

h

i

p

be

tw

ee

n

f

l

i

gh

t

nu

m

be

r

an

d

l

aun

c

h

S

i

te,

pa

y

l

oa

d

and

l

aun

c

h

s

i

t

e,

s

u

cc

e

s

s

r

at

e

of

ea

c

h

o

r

b

i

t

type, flight

number

and

orbit type,

the

launch

success

yearly

trend.

EDA

with

Data

Visualization

•

The

link

to

the

notebook

is

https://github.com/chuksoo/IBM

-

Data

-

Science

-

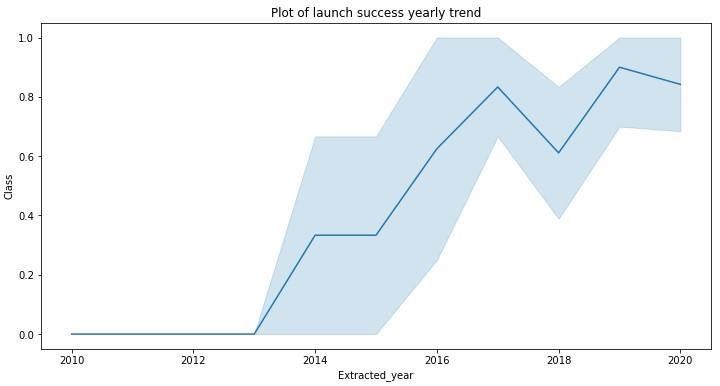
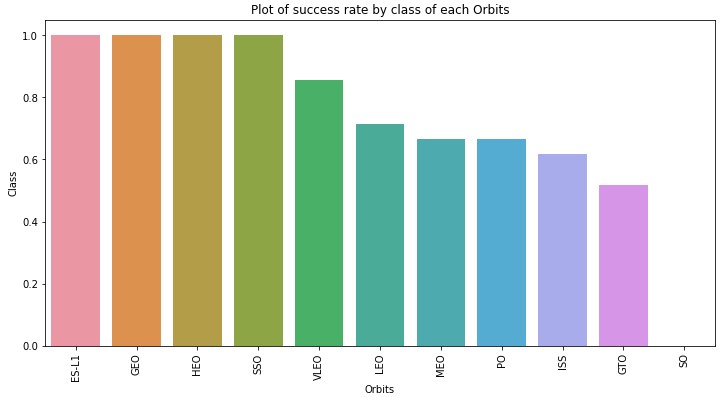
Capstone

-

SpaceX/blob/main/EDA%20with%20D

ata%20Visualization.ipynb

11



•

We

loaded

the

SpaceX

dataset

into

a

PostgreSQL

database

without

leaving

the

jupyter

notebook.

•

We

applied

EDA

with

SQL

to get

insight

from

the

data.

We

wrote

queries

to

find

out

for

instance:

-

The

names

of

unique

launch

sites

in

the

space

mission.

-

The

total

payload

mass

carried

by

boosters

launched

by

NASA

)

(

CRS

-

The

average

payload

mass

carried

by

booster

version

F9

v1.1

-

The

total

number

of

successful

and

failure

mission

outcomes

-

The

failed

landing

outcomes

in

drone

ship,

their

booster

version

and

launch

site

names.

•

The

link

to

the

notebook

is

https://github.com/chuksoo/IBM

-

Data

-

Science

-

Capstone

-

SpaceX/blob/main/EDA%20with%20SQL.ipynb

EDA

with

SQL



•

We

marked

all

launch

sites,

and added

map

objects

such

as

markers, circles,

lines

to

mark

the

success

or

failure

of

launches

for

each

site

on

the

folium

map.

•

We

assigned

the

feature

launch

outcomes

(

failure or

success)

to

class

0

and

1

.i.e.,

0

for

failure,

and

1

for

success.

•

Using

the

color

-

labeled

marker

clusters,

we

identified

which

launch

sites

have

r

e

l

a

t

i

v

e

l

y

h

i

g

h

s

u

cc

e

ss

r

at

e

.

•

We

calculated

the

distances

between

a

launch

site

to

its

proximities.

We

answered

some

question

for

instance:

-

A

r

e

l

aun

c

h

s

i

t

e

s

n

e

a

r

r

a

il

w

a

y

s

,

h

i

g

h

w

a

y

s

an

d

c

o

a

s

t

li

n

e

s.

-

Do

launch

sites

keep

certain

distance

away

from

cities.

Build

an

Interactive

Map

with

Folium



•

We

built

an

interactive

dashboard

with

Plotly

dash

•

We

plotted

pie

charts

showing

the

total

launches

by

a

certain

sites

•

We

plotted

scatter

graph

showing

the

relationship

with

Outcome

and

Payload

Mass

(

Kg

)

for

the

different

booster

version.

•

The

link

to

the

notebook

is

https://github.com/chuksoo/IBM

-

Data

-

Science

-

Capstone

-

SpaceX/blob/main/app.py

Build

a

Dashboard

with

Plotly

Dash



•

We

loaded

the data

using

numpy

and

pandas,

transformed

the

data,

split

our

data

into

training

and

testing.

•

We

built

different

machine

learning

models

and

tune

different

hyperparameters

using

GridSearchCV.

•

We

used

accuracy

as

the

metric

for our

model,

improved

the

model

using

feature

engineering

and

algorithm

tuning.

•

We

found

the

best

performing

classification

model.

•

The

link

to

the

notebook

is

https://github.com/chuksoo/IBM

-

Data

-

Science

-

Capstone

-

SpaceX/blob/main/Machine%20Learning%20Prediction.ipynb

P

r

ed

i

c

t

i

v

e

A

na

l

y

s

i

s

(

C

l

a

ss

i

f

i

c

a

t

i

on

)



•

E

x

p

l

o

r

at

o

r

y

da

t

a

ana

l

ys

i

s

r

e

s

u

l

ts

•

Interactive

analytics

demo

in

screenshots

•

P

r

ed

i

c

t

i

v

e

ana

l

ys

is

r

e

s

u

l

ts

R

e

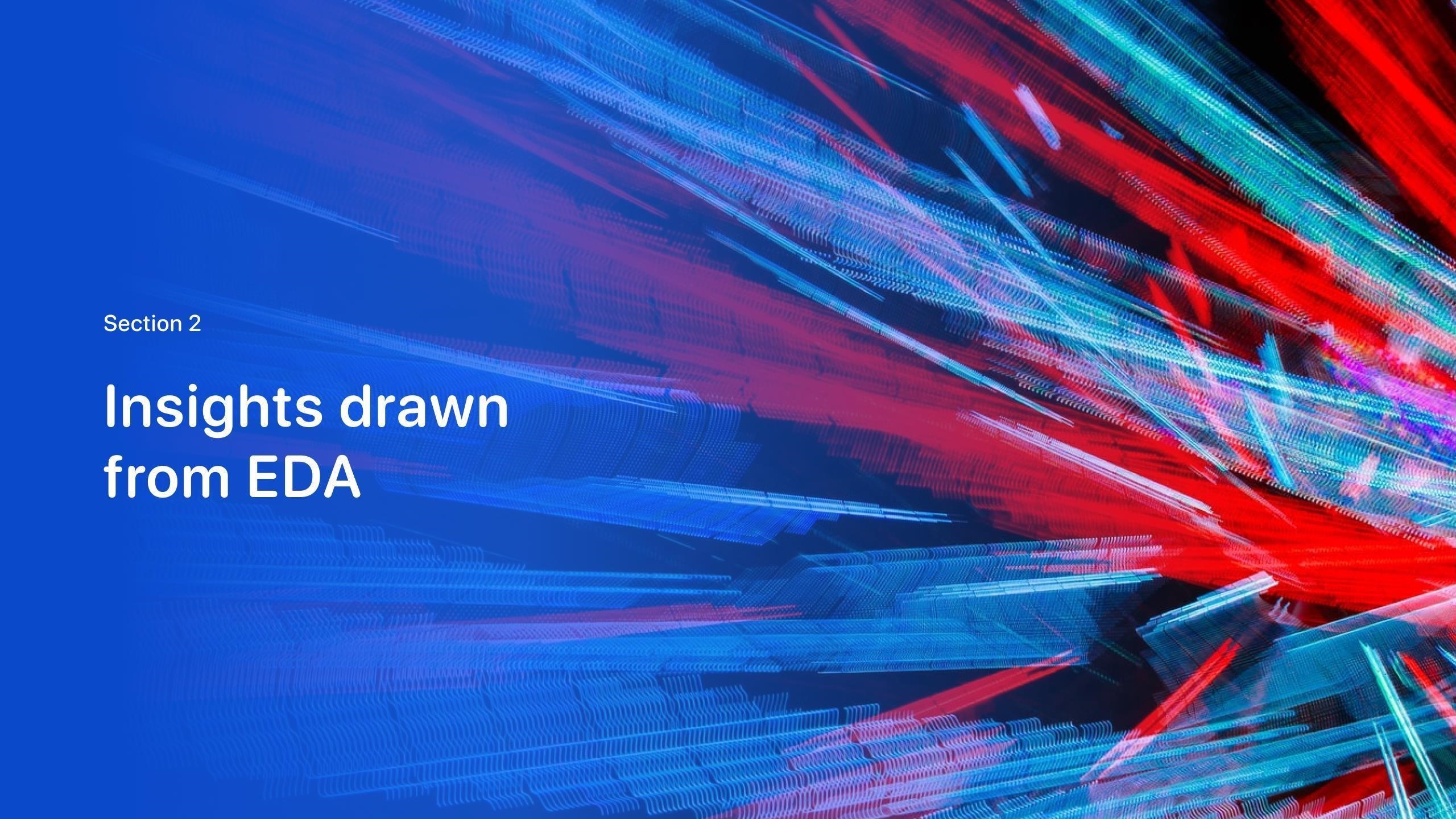
s

u

l

t

s





•

From

the

plot,

we

found that the

larger

the

flight

amount

at a

launch

site,

the

greater

the

success

rate

at

a

launch

site.

F

l

i

gh

t

N

u

m

be

r

vs

.

Laun

c

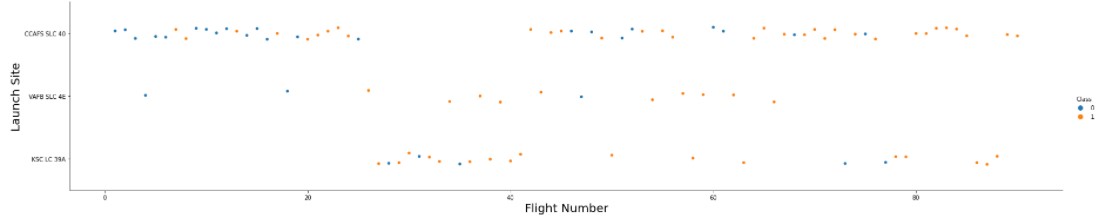
h

S

i

t

e



Payloadvs.



LaunchSite

19

# Success Rate vs. Orbit Type

• From the plot, we can see that ES-L1,GEO,HEO,SSO, VLEO had the most success rate.



20



•

The

plot

below

shows

the

Flight

Number

vs.

Orbit

type.

We

observe

that

in

the

LEO

orbit,

success

is

related

to the

number

of

flights

whereas

in

the

GTO

orbit,

there

is

no

relationship

between

flight

number

and

the

orbit.

F

l

i

gh

t

N

u

m

be

r

vs

.

O

r

b

i

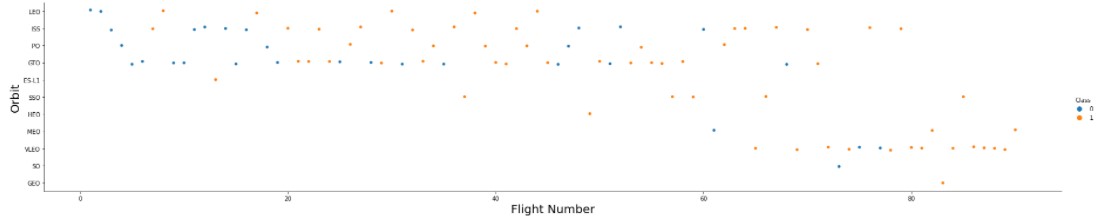
t

T

y

p

e



21



•

We

can observe

that

with

heavy

payloads,

the

successful

landing are

more

for

PO,

LEO

and

ISS

orbits.

P

a

y

l

oad

vs

.

O

r

b

i

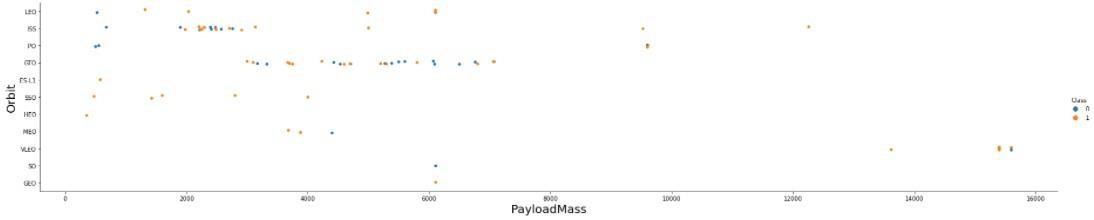
t

T

y

p

e

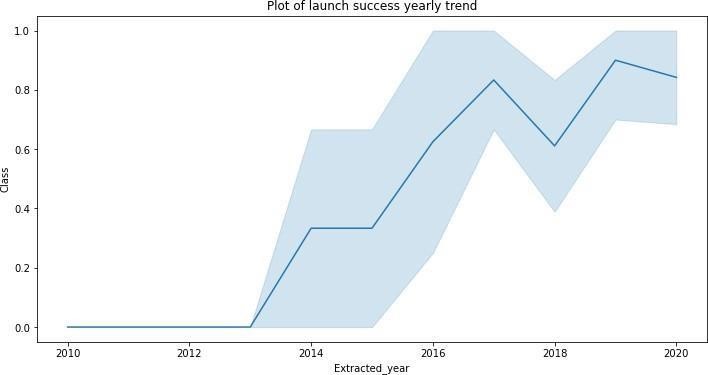


22

# Launch SuccessYearly Trend

• From the plot, we can

observe that successrate since 2013 kept on increasing till 2020.



23

# All Launch Site Names

• Weused the key word **DISTINCT** to show only unique launch sites from the SpaceX data.

24



•

We

used

the

query

above

to

display

5

records

where

launch

sites

begin

with

`CCA`

Launch

Site

Names

Begin

with

'CCA'



25



•

We

calculated

the

total

payload

carried by

boosters

from

NASA

as

45596

using

the

query

below

T

o

t

a

l

P

a

y

l

oa

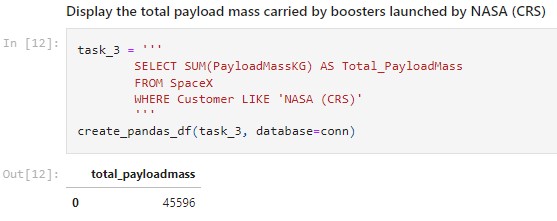
d

M

a

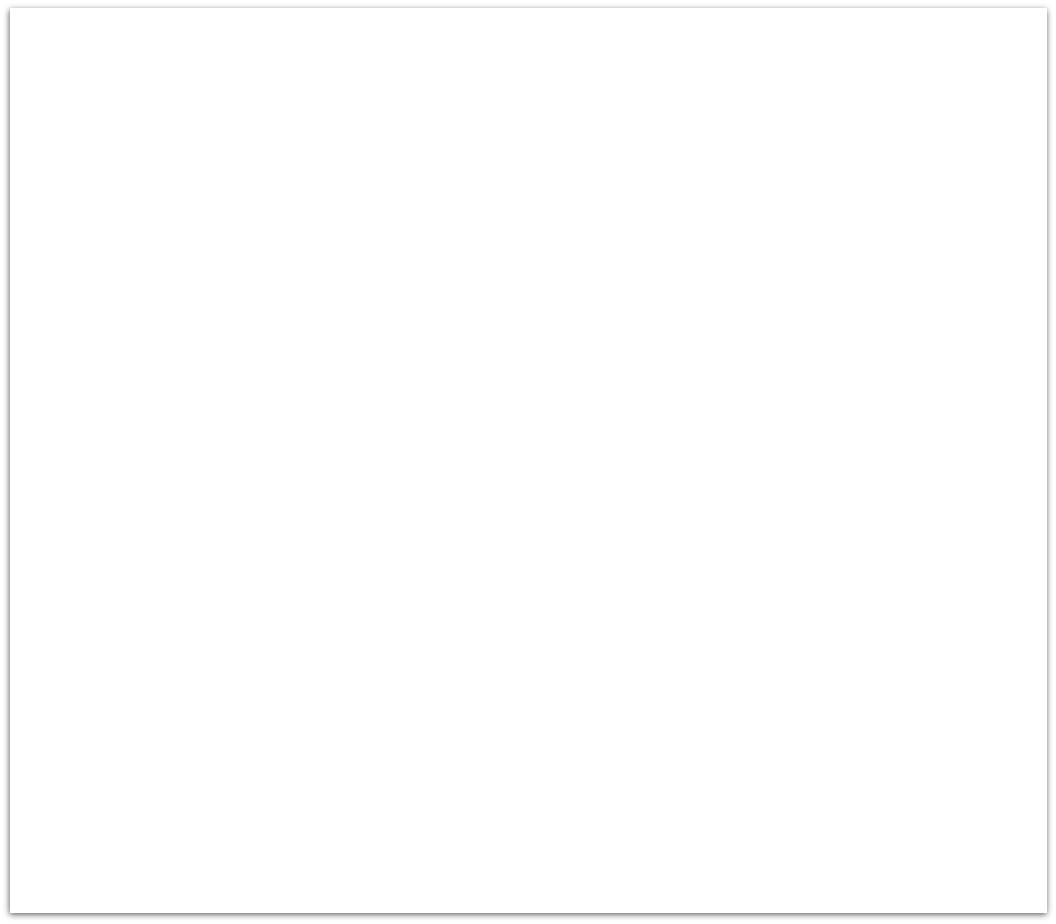
s

s



26

# Average Payload Mass by F9 v1.1



• Wecalculated the average payloadmasscarriedby boosterversionF9 v1.1 as

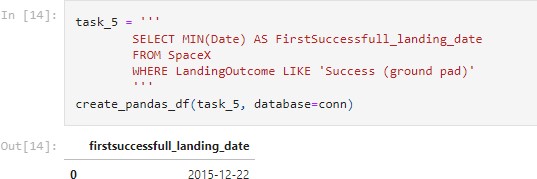
2928.4

## First SuccessfulGround Landing Date

* We observedthat the dates of the

first successful landing outcome on ground pad was 22ndDecember

2015



* We usedthe **WHERE**clauseto filter for boosterswhichhave successfullylandedondrone ship and appliedthe **AND** condition to determine successfullanding with payload massgreaterthan 4000 but less than6000

Successful

Drone

Ship

Landing

with

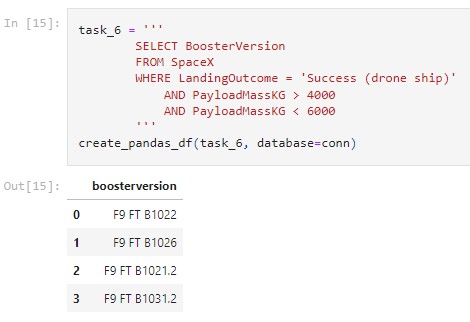
Payload

between

4000

and

6000



Total

Number

of

Successful

and

Failure

Mission

Outcomes

•

W

e

u

s

ed

w

il

d

c

ard

li

k

e

‘%

’

t

o

f

il

t

e

r

f

o

r

**W**

**H**

**E**

**R**

**E**

M

i

ss

i

on

O

u

tc

o

m

e

w

a

s

a

s

u

c

c

e

ss

or

a

f

a

il

ure.



Boosters

Carried

Maximum

Payload

•

W

e

de

t

e

r

m

i

n

e

d

t

h

e

bo

o

st

e

r

t

ha

t

ha

v

e

c

ar

r

i

e

d

t

h

e

m

a

x

i

m

u

m

p

a

y

l

o

a

d

u

s

i

n

g

a

s

ubq

u

e

r

y

i

n

t

h

e

**W**

**H**

**E**

**R**

**E**

c

l

au

s

e

a

n

d

t

h

e

**M**

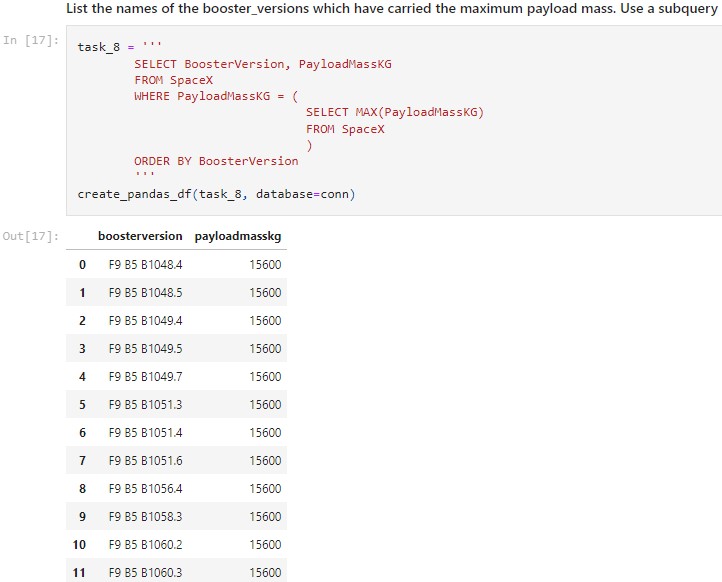
**A**

**X**

**(**

**)**

function.



32

•

We

used

a

combinations

of the

**WHERE**

clause,

**LIKE**

,

**AND**

,

and

**BETWEEN**

conditions

to

filter

for failed landing

outcomes

in

drone

ship, their

booster

v

e

r

s

i

on

s

,

an

d

l

aun

c

h

s

i

t

e

na

m

e

s

f

o

r

y

ea

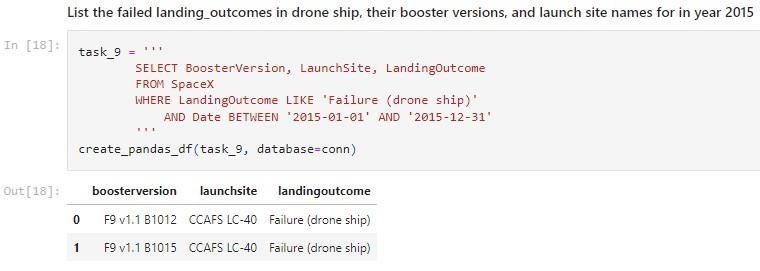
r

2015

2015

Launch

Records



and the **COUNT**of landing

Ran

k

Land

i

n

g

O

u

t

c

o

m

e

s

B

e

t

wee

n

2010

-

06

-

04

and

2017

-

03

-

20

•

W

e

s

e

l

e

ct

e

d

Land

i

n

g

ou

tc

o

m

e

s

ou

tc

o

m

e

s

f

ro

m

t

he

da

t

a

and

u

s

ed

t

he

**W**

**H**

**E**

**R**

**E**

c

l

au

s

e

t

o

f

il

t

er

f

or

l

a

n

d

i

ng

ou

t

c

o

m

e

s

**B**

**E**

**T**

**W**

**EE**

**N**

2010

-

06

-

04

to

2010

-

03

-

20.

•

W

e

ap

p

li

ed

t

h

e

**G**

**R**

**O**

**U**

**P**

**B**

**Y**

c

l

au

s

e

to

group

the

landing

outcomes

and

t

h

e

**O**

**RD**

**E**

**R**

**B**

**Y**

c

l

au

s

e

t

o

o

u

tc

o

m

e

i

n

de

sc

end

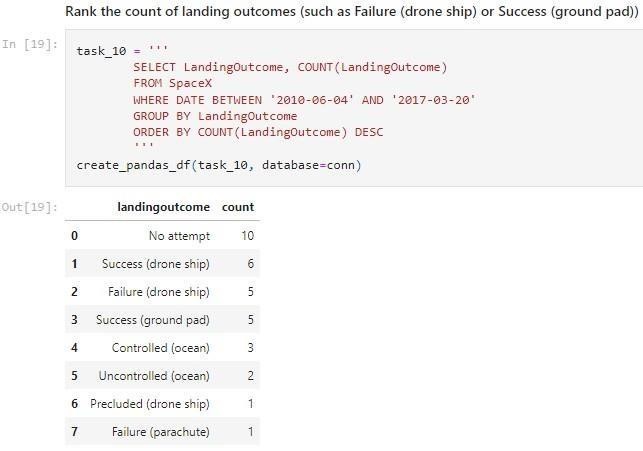
i

n

g

order.

33



order the grouped landing





All

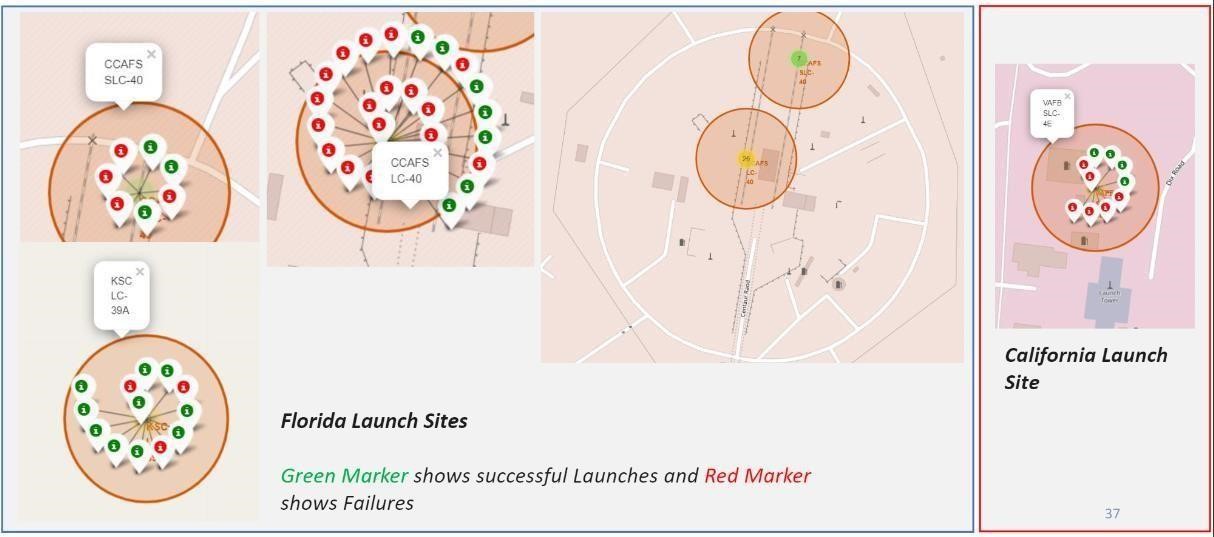
launch

sites

global

map

markers



Markers

showing

launch

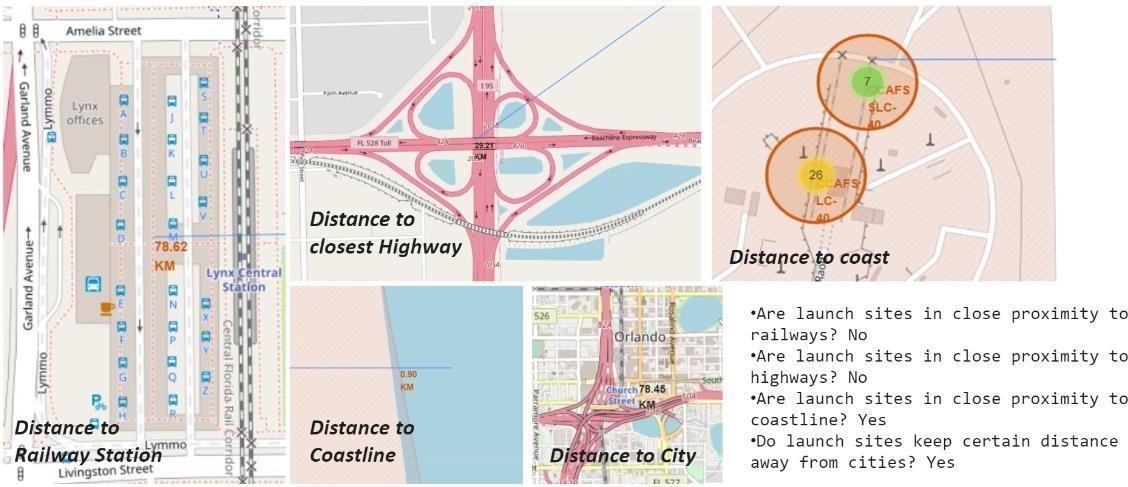
sites

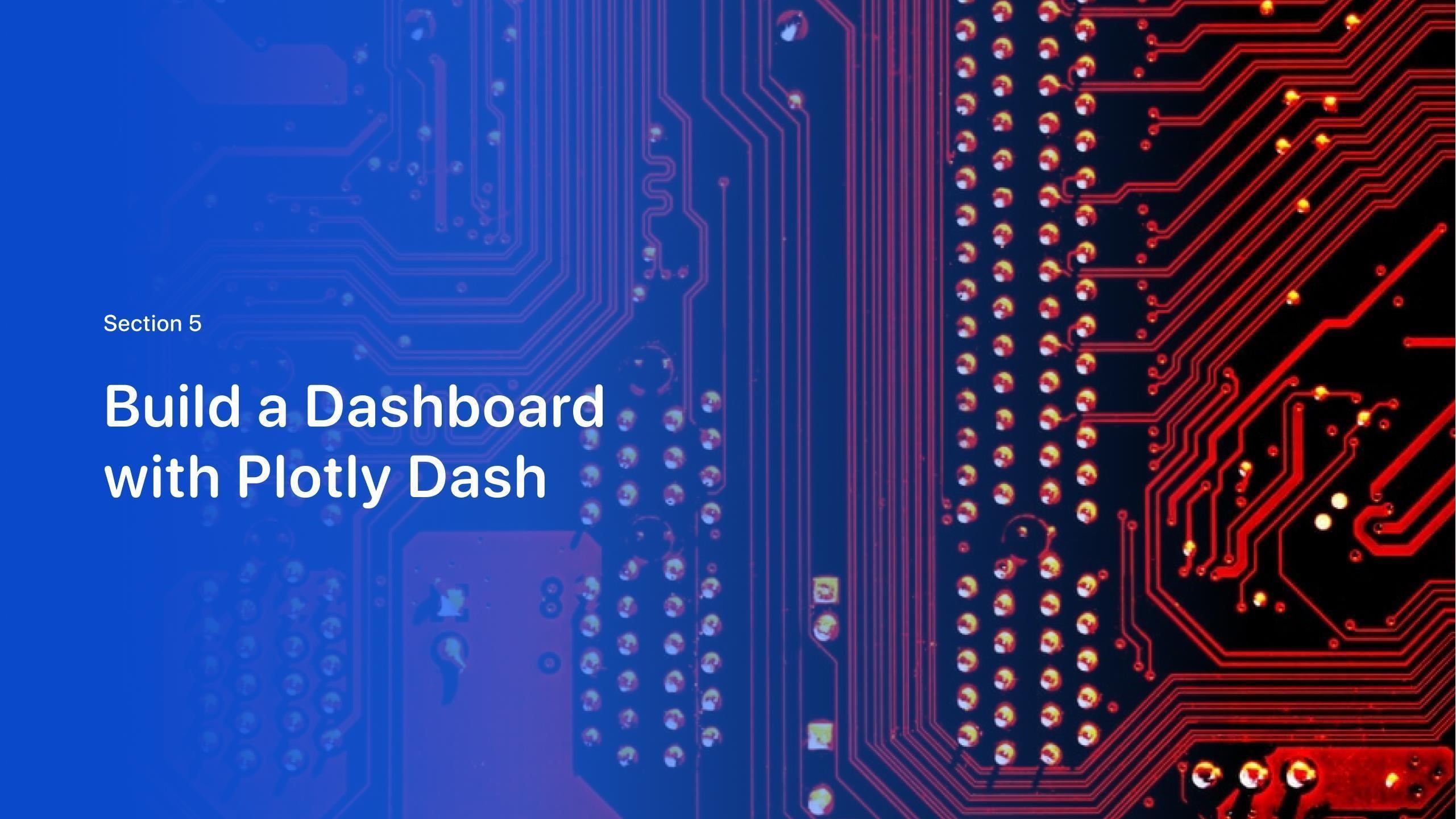
with

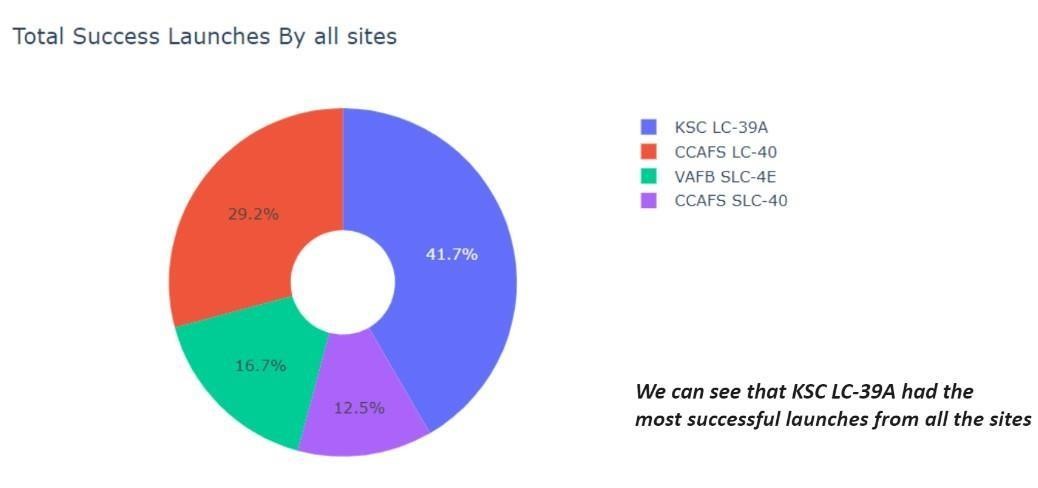
color

labels

# Launch Site distance to landmarks







39

Pie

chart

showing

the

success

percentage

achieved

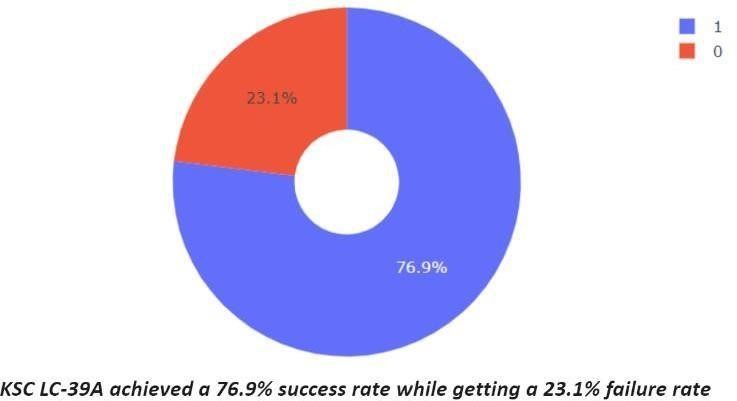
by

each

launch

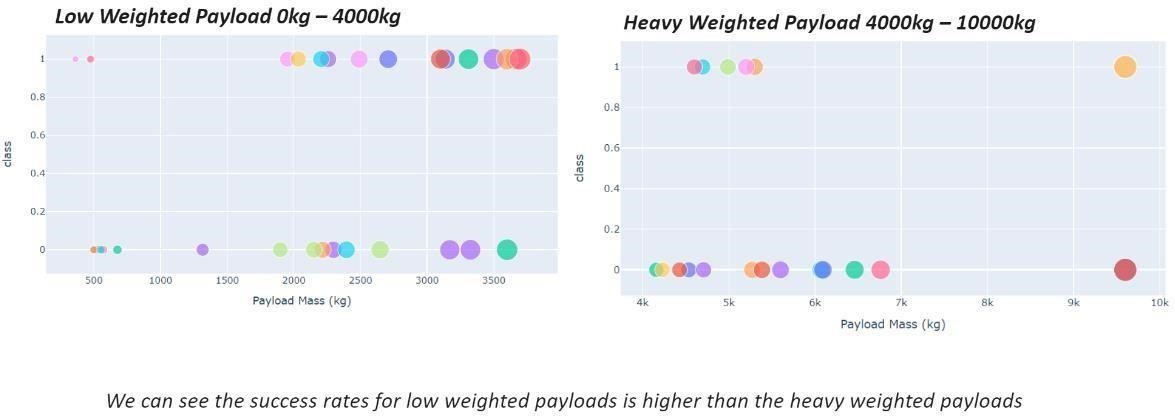
site

Pie chart showing the Launch site with the highest launch success ratio



40

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider



41





Classification

Accuracy

•

The

decision

tree

classifier

is

the

model

with

the

highest

classification

accuracy



43



•

T

h

e

c

on

f

u

s

i

o

n

m

a

t

r

i

x

f

o

r

t

h

e

de

c

i

s

i

o

n

t

r

ee

classifier

shows

that

the

classifier

can

distinguish

between

the

different

classes.

T

h

e

m

a

j

o

r

p

r

ob

l

e

m

i

s

t

h

e

f

a

l

s

e

po

s

i

t

i

v

e

s

.

i

.e.,

un

s

u

cc

e

ss

f

u

l

l

and

i

n

g

m

a

r

k

e

d

a

s

s

u

cc

e

ss

f

u

l

landing

by

the

classifier.

C

on

f

u

s

i

o

n

M

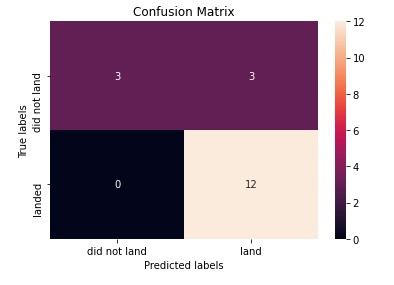
a

t

r

i

x



44



We

can conclude

that:

•

The

larger

the

flight

amount

at

a

launch

site,

the

greater

the

success

rate

at

a

launch

site.

•

Launch

success

rate

started

to

increase

in

2013

till

2020.

•

Orbits

ES

-

L1,

GEO,

HEO,

SSO,

VLEO

had

the

most

success

rate.

•

KSC

LC

-

39

A

had

the

most

successful

launches

of

any

sites.

•

The

Decision

tree

classifier

is

the

best

machine

learning

algorithm

for

this

task.

45

Conclusions

