Big Data Deliverable 2

Team :

Khyaati Bhumireddy (801312911)

Darshini Chalumuri (801308863)

Swathi Geerlakunta (801310428)

Isaiah Thomas (801060357)

Chris Kelly (800159813)

GitHub Link :

https://github.com/ithomasdev/ITCS6100_group6

Data Understanding and Preparation :

1. Firstly , we have loaded the dataset into Amazon Simple Storage Service (S3) primarily by creating a new bucket by name bigdataprojectgrpou6 and loaded the dataset.
2. Then, created a book in the Sagemaker and started exploratory data analysis.
3. Used QuickSight's drag-and-drop interface to create charts, tables, and visualizations to summarize the key features and insights from the EDA.

In [ ]:
```
import numpy as np
import pandas as pd
import os
```

Importing Dataset

In [2]:
```
df=pd.read_csv(r"athletes.csv")
```

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 423006 entries, 0 to 423005
Data columns (total 27 columns):
 #   Column      Non-Null Count   Dtype  
---  ------      --------------   -----  
 0   athlete_id  423003 non-null  float64
 1   name        331110 non-null  object 
 2   region      251262 non-null  object 
 3   team        155160 non-null  object 
 4   affiliate   241916 non-null  object 
 5   gender      331110 non-null  object 
 6   age         331110 non-null  float64
 7   height      159869 non-null  float64
 8   weight      229890 non-null  float64
```

```
 9   fran        55426 non-null   float64
 10  helen       30279 non-null   float64
 11  grace       40745 non-null   float64
 12  filthy50    19359 non-null   float64
 13  fgonebad    29738 non-null   float64
 14  run400      22246 non-null   float64
 15  run5k       36097 non-null   float64
 16  candj       104435 non-null  float64
 17  snatch      97280 non-null   float64
 18  deadlift    115323 non-null  float64
 19  backsq      110517 non-null  float64
 20  pullups     50608 non-null   float64
 21  eat         93932 non-null   object
 22  train       105831 non-null  object
 23  background  98945 non-null   object
 24  experience  104936 non-null  object
 25  schedule    97875 non-null   object
 26  howlong     109206 non-null  object
dtypes: float64(16), object(11)
memory usage: 87.1+ MB
```

In [ ]:
```
Data Preparation
```

In [4]:
```
df.columns
```

Out[4]:
```
Index(['athlete_id', 'name', 'region', 'team', 'affiliate', 'gender', 'ag
e',
       'height', 'weight', 'fran', 'helen', 'grace', 'filthy50', 'fgoneba
d',
       'run400', 'run5k', 'candj', 'snatch', 'deadlift', 'backsq', 'pullup
s',
       'eat', 'train', 'background', 'experience', 'schedule', 'howlong'],
      dtype='object')
```

In [5]:
```
df.describe()
```

Out[5]:

	athlete_id	age	height	weight	fran	
count	423003.000000	331110.000000	1.598690e+05	229890.000000	5.542600e+04	3.0279
mean	292748.166538	32.516750	1.206217e+02	170.896137	9.886691e+02	1.2079
std	184969.660327	7.730671	2.097995e+04	58.379799	7.200430e+04	6.8240
min	82.000000	13.000000	0.000000e+00	1.000000	1.000000e+00	1.0000
25%	135091.500000	27.000000	6.600000e+01	145.000000	2.150000e+02	5.2500
50%	275839.000000	31.000000	6.900000e+01	170.000000	2.900000e+02	5.9500
75%	473188.000000	37.000000	7.200000e+01	192.000000	3.920000e+02	6.9400
max	633083.000000	125.000000	8.388607e+06	20175.000000	8.388607e+06	8.3886

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
athlete_id         3
name           91896
region        171744
team          267846
affiliate     181090
gender         91896
age            91896
height        263137
weight        193116
fran          367580
helen         392727
grace         382261
filthy50      403647
fgonebad      393268
run400        400760
run5k         386909
candj         318571
snatch        325726
deadlift      307683
backsq        312489
pullups       372398
eat           329074
train         317175
background    324061
experience    318070
schedule      325131
howlong       313800
dtype: int64
```

In [7]:
```
df.drop('filthy50', axis=1, inplace=True)
```

In [8]:
```
df.drop('run400', axis=1, inplace=True)
```

In [9]:
```
df.shape
```

Out[9]:
```
(423006, 25)
```

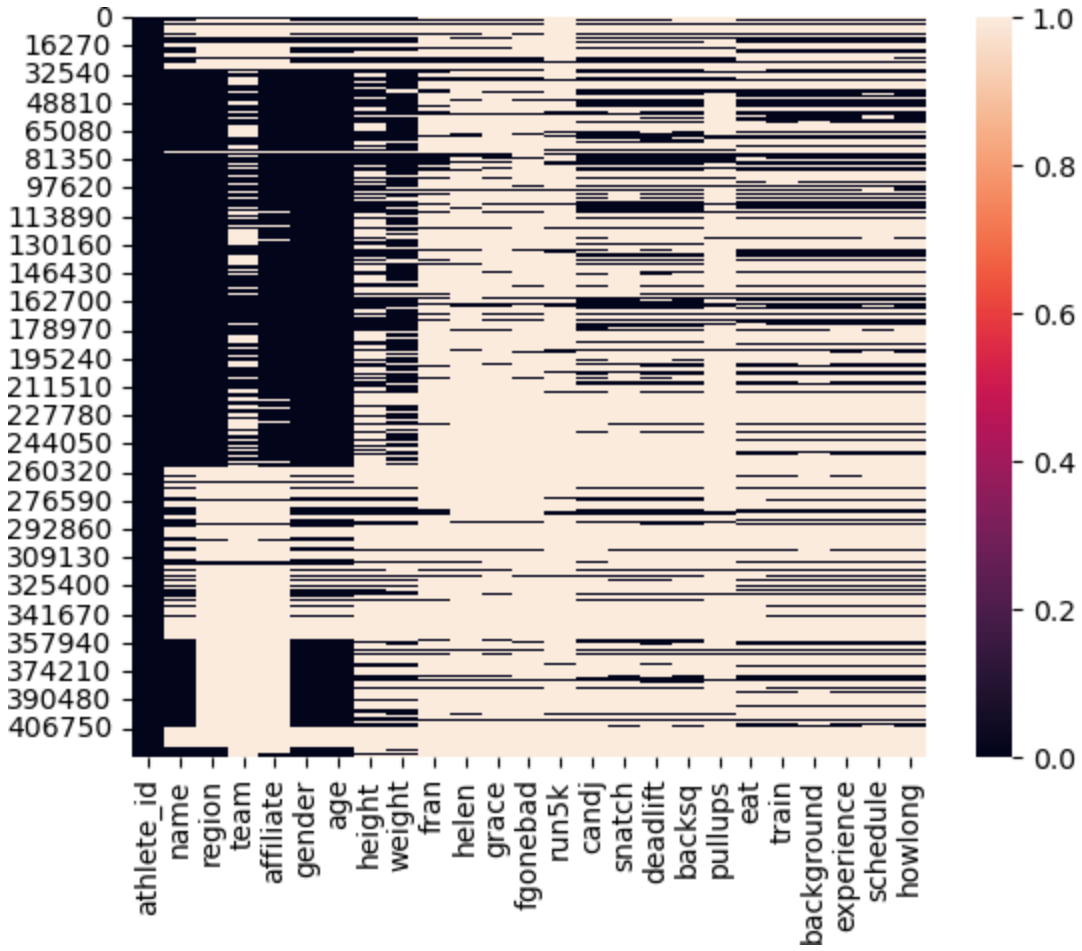
In [10]:
```
df.isnull().sum()
```

Out[10]:
```
athlete_id         3
name           91896
region        171744
team          267846
affiliate     181090
gender         91896
age            91896
height        263137
weight        193116
fran          367580
helen         392727
grace         382261
fgonebad      393268
run5k         386909
candj         318571
snatch        325726
deadlift      307683
backsq        312489
pullups       372398
eat           329074
train         317175
background    324061
experience    318070
schedule      325131
howlong       313800
dtype: int64
```

Generating Heatmap

In [11]:
```
import seaborn as sns
sns.heatmap(df.isnull())
```

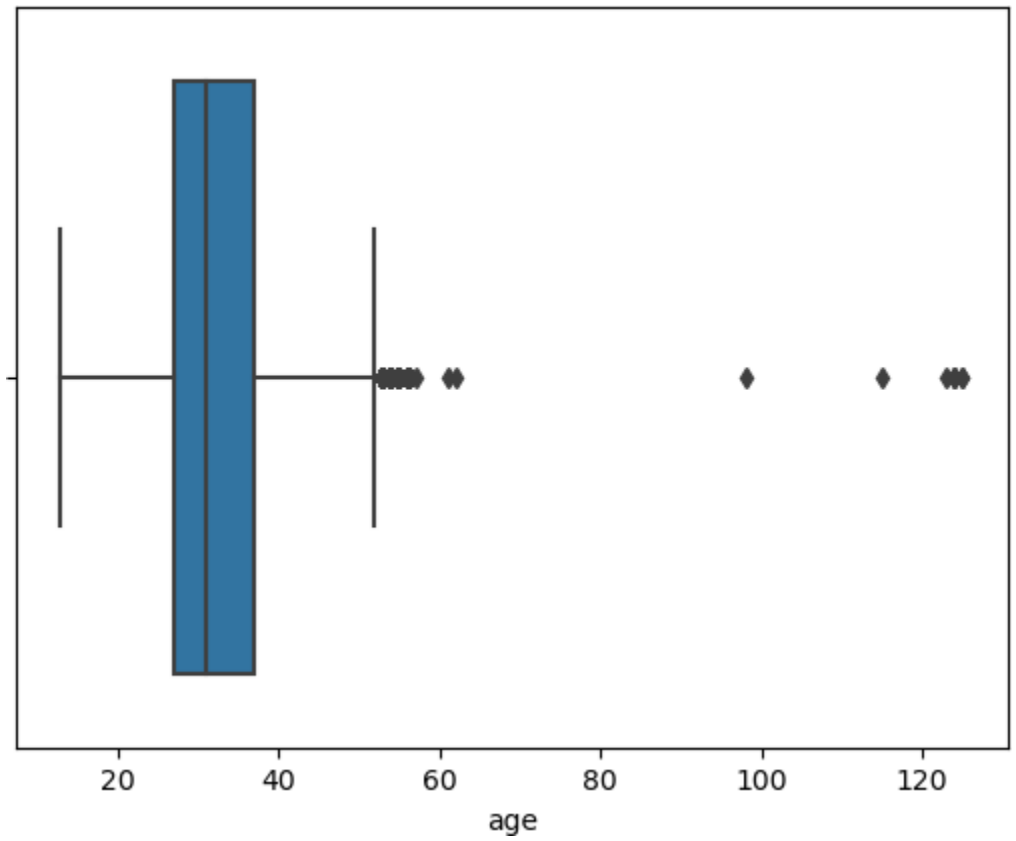
Out[11]:
```
<AxesSubplot: >
```



In [ ]:
```
Creating Box Plot
```

In [13]:
```
sns.boxplot(x=df['age'])
```

Out[13]:
```
<AxesSubplot: xlabel='age'>
```



In [ ]:
```
Creating Histogram
```

In [17]:
```
import pandas as pd


# Check for NaN values in the column
if  df['pullups'].isna().any():
    # Replace NaN values with 0
    df['pullups'] = df['pullups'].fillna(0)

# Convert the column to an integer type
df['pullups'] = df['pullups'].astype(int)
```

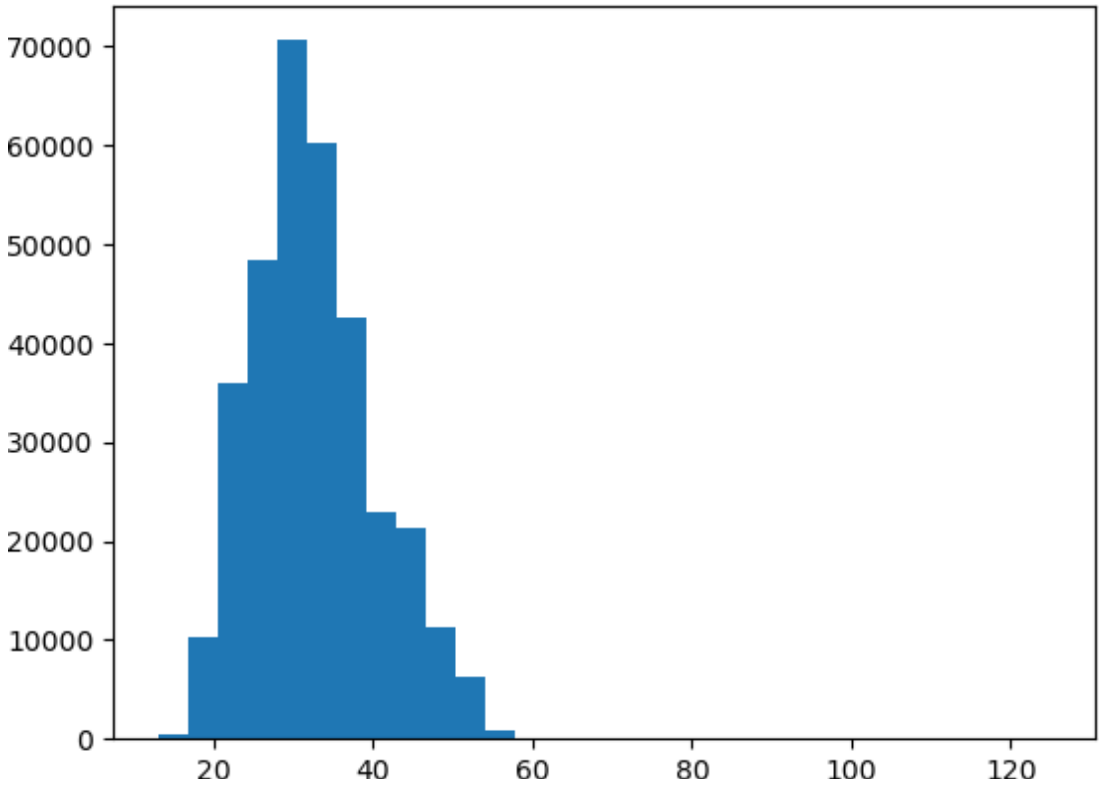
In [20]:
```
import pandas as pd


# Check for NaN values in the column
if  df['backsq'].isna().any():
    # Replace NaN values with 0
    df['backsq'] = df['backsq'].fillna(0)

# Convert the column to an integer type
df['backsq'] = df['backsq'].astype(int)
```

In [28]:
```
plt.hist(df['age'], bins=30)
```

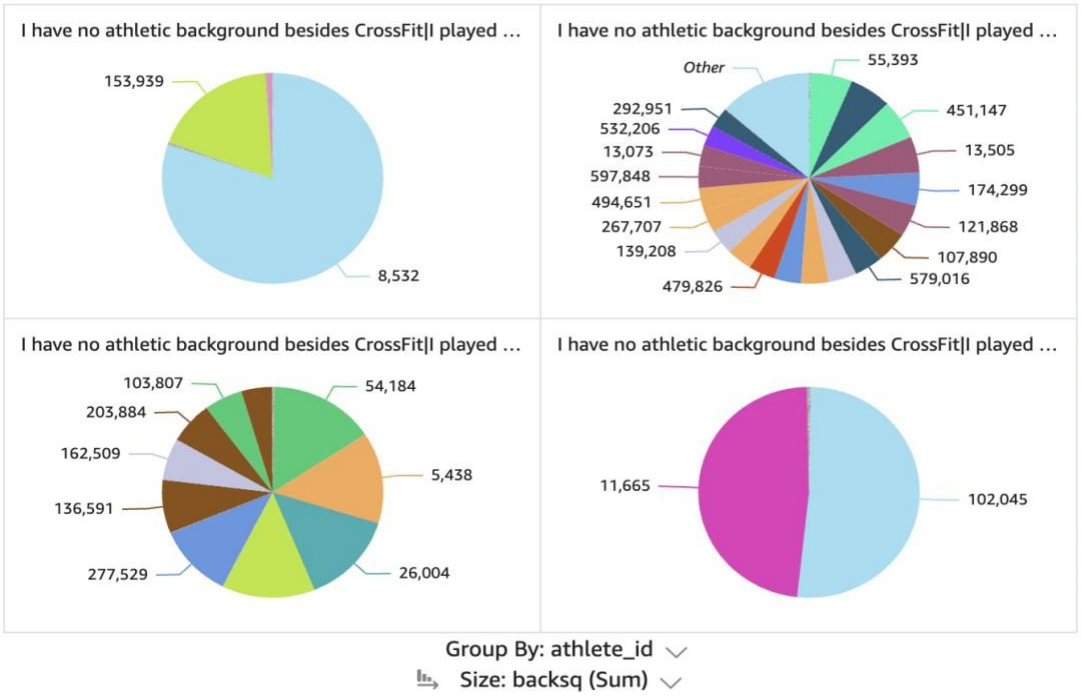
Out[28]:
```
(array([4.8100e+02, 1.0281e+04, 3.6021e+04, 4.8468e+04, 7.0575e+04,
        6.0269e+04, 4.2606e+04, 2.2886e+04, 2.1265e+04, 1.1215e+04,
        6.1840e+03, 8.4900e+02, 1.0000e+00, 1.0000e+00, 0.0000e+00,
        0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00,
        0.0000e+00, 0.0000e+00, 1.0000e+00, 0.0000e+00, 0.0000e+00,
        0.0000e+00, 0.0000e+00, 1.0000e+00, 0.0000e+00, 6.0000e+00]),
 array([ 13.        ,  16.73333333,  20.46666667,  24.2       ,
         27.93333333,  31.66666667,  35.4       ,  39.13333333,
         42.86666667,  46.6       ,  50.33333333,  54.06666667,
         57.8       ,  61.53333333,  65.26666667,  69.        ,
         72.73333333,  76.46666667,  80.2       ,  83.93333333,
         87.66666667,  91.4       ,  95.13333333,  98.86666667,
        102.6       , 106.33333333, 110.06666667, 113.8       ,
        117.53333333, 121.26666667, 125.        ]),
 <BarContainer object of 30 artists>)
```



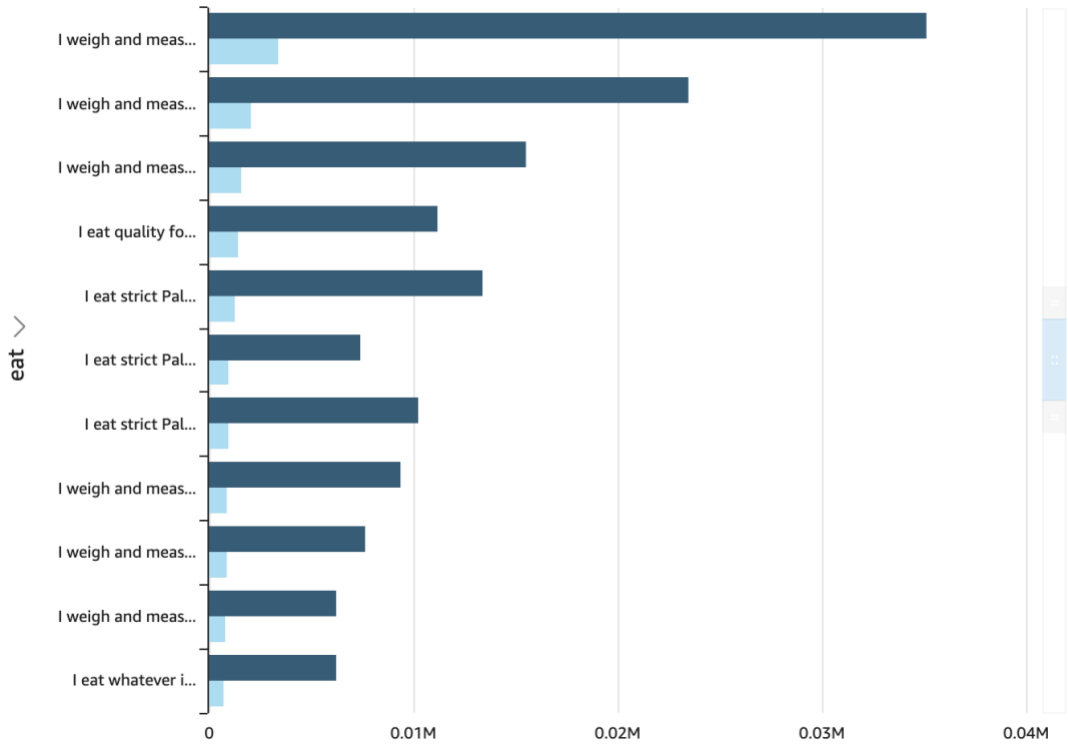
Dashboard

Sum of Backsq by Athlete_id and Background

SHOWING BOTTOM 20 IN BACKGROUND AND TOP 20 IN ATHLETE_ID

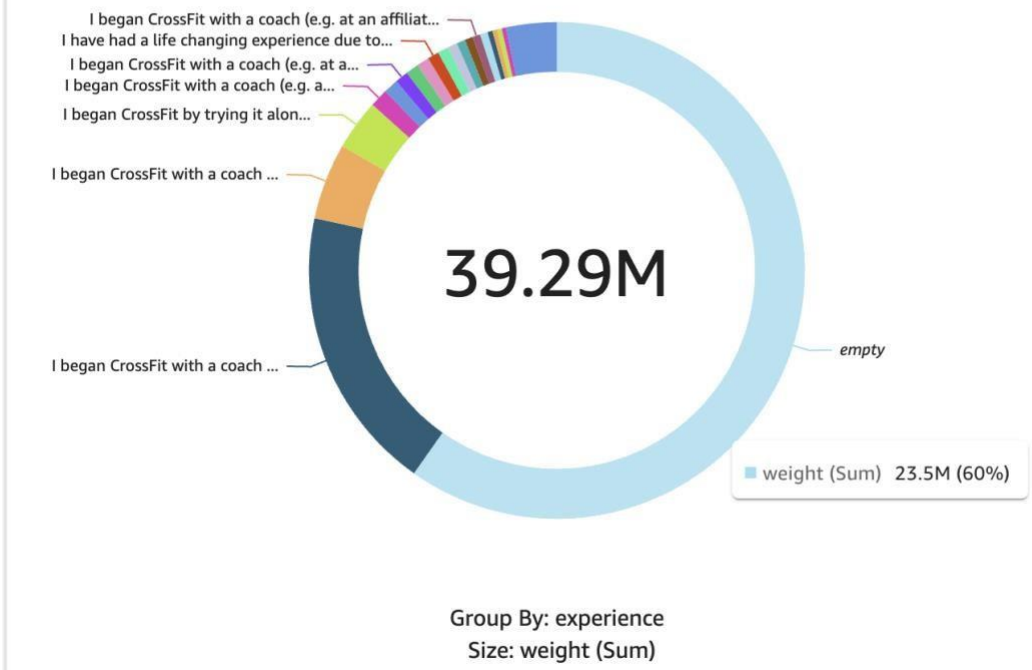


Sum of Age and Sum of Deadlift by Eat



Sum of Weight by Experience

SHOWING TOP 20 IN EXPERIENCE

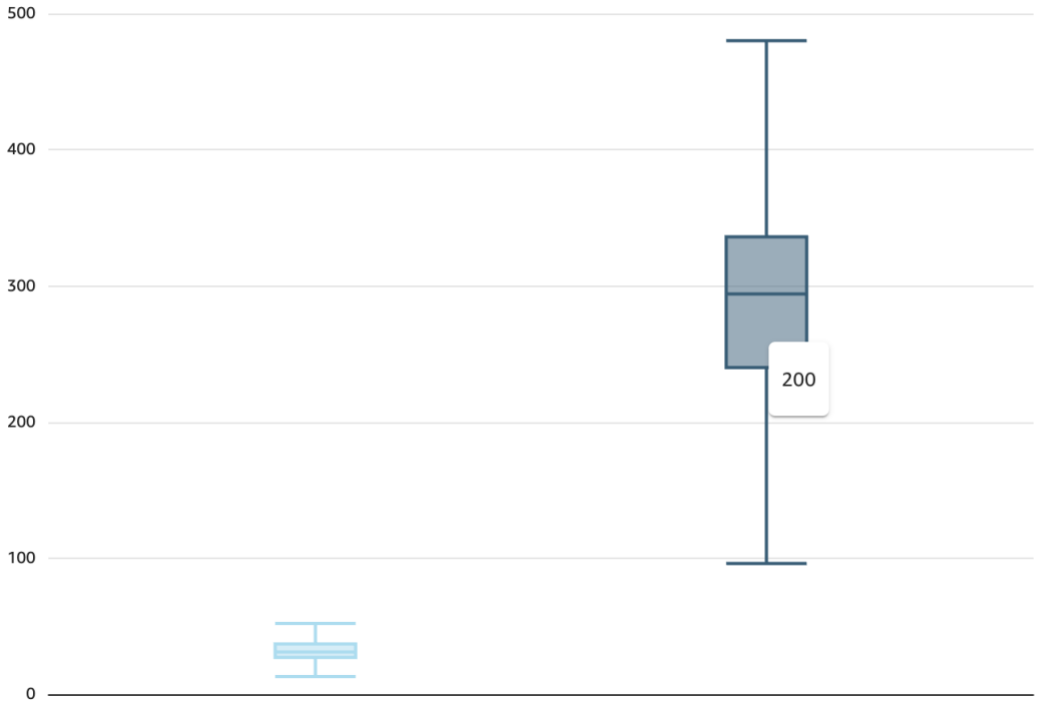


Experience

- empty
- I began Cro...
- I began Cro...
- I began Cro...
- I began Cro...
- I began Cro...
- I began Cro...
- I began Cro...
- I began Cro...
- I have had ...
- I began Cro...
- I began Cro...
- I began Cro...
- I began Cro...
- I began Cro...
- I began Cro...

Group By: experience
Size: weight (Sum)

Sum of Age and Fgonebad



Legend

- age
- fgonebad