NAME: AVIREDDY NVSRK ROHAN

REG.NO: 19BCE1180

DATASET LINK: https://data.world/informatics-edu/diabetes-prediction

SHORT DESCRIPTION:

Original data came from the Biostatistics program at Vanderbilt
http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets

Several hundred rural African-American patients were included. The diabetes.csv file contains the raw data of all patients, including those with missing data. This can be used for descriptive statistics. The data dictionary to explain the columns can be found here: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/Cdiabetes.html and http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Diabetes.html

```
R.version.string
```

```
'R version 4.1.1 (2021-08-10)'
```

```
library(dplyr)
```

```
dia_data = read.csv("/content/Diabetes_Classification.csv",header=TRUE)
head(dia_data)
```

A da

Patient.number	Cholesterol	Glucose	HDL.Chol	Chol.HDL.ratio	Age	Gender	Height
<int>	<int>	<int>	<int>	<dbl>	<int>	<chr>	<int>

```
str(dia_data)
```

```
'data.frame':   390 obs. of  18 variables:
 $ Patient.number : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Cholesterol    : int  193 146 217 226 164 170 149 164 230 179 ...
 $ Glucose        : int  77 79 75 97 91 69 77 71 112 105 ...
 $ HDL.Chol       : int  49 41 54 70 67 64 49 63 64 60 ...
 $ Chol.HDL.ratio : num  3.9 3.6 4 3.2 2.4 2.7 3 2.6 3.6 3 ...
 $ Age            : int  19 19 20 20 20 20 20 20 20 20 ...
 $ Gender         : chr  "female" "female" "female" "female" ...
 $ Height         : int  61 60 67 64 70 64 62 72 67 58 ...
 $ Weight         : int  119 135 187 114 141 161 115 145 159 170 ...
 $ BMI            : num  22.5 26.4 29.3 19.6 20.2 27.6 21 19.7 24.9 35.5 ...
 $ Systolic.BP    : int  118 108 110 122 122 108 105 108 100 140 ...
 $ Diastolic.BP   : int  70 58 72 64 86 70 82 78 90 100 ...
 $ waist          : int  32 33 40 31 32 37 31 29 31 34 ...
 $ hip            : int  38 40 45 39 39 40 37 36 39 46 ...
 $ Waist.hip.ratio: num  0.84 0.83 0.89 0.79 0.82 0.93 0.84 0.81 0.79 0.74 ...
 $ Diabetes       : chr  "No diabetes" "No diabetes" "No diabetes" "No diabetes" ...
 $ X              : int  6 NA NA NA NA NA NA NA NA NA ...
 $ X.1            : int  6 NA NA NA NA NA NA NA NA NA ...
```

```
dim(dia_data)
```

```
390 · 18
```

DATA CLEANING

```
dia_data=subset (dia_data, select = -X)
```

```
dia_data=subset (dia_data, select = -X.1)
```

```
dia_data=subset (dia_data, select = -Patient.number)
```

```
any(is.na(dia_data))
```

```
FALSE
```

```
head(dia_data,10)
```

A data.frame: 10 × 15

	Cholesterol	Glucose	HDL.Chol	Chol.HDL.ratio	Age	Gender	Height	Weight	BMI
	<int>	<int>	<int>	<dbl>	<int>	<chr>	<int>	<int>	<dbl>
1	193	77	49	3.9	19	female	61	119	22.5
2	146	79	41	3.6	19	female	60	135	26.4
3	217	75	54	4.0	20	female	67	187	29.3
4	226	97	70	3.2	20	female	64	114	19.6
5	164	91	67	2.4	20	female	70	141	20.2
6	170	69	64	2.7	20	female	64	161	27.6
7	149	77	49	3.0	20	female	62	115	21.0
8	164	71	63	2.6	20	male	72	145	19.7
9	230	112	64	3.6	20	male	67	159	24.9
10	179	105	60	3.0	20	female	58	170	35.5

```
library(graphics)
library(ggplot2)
```

descriptive statistics

```
summary(dia_data)
```

```
 Cholesterol       Glucose         HDL.Chol      Chol.HDL.ratio
Min.   : 78.0   Min.   : 48.0   Min.   : 12.00   Min.   : 1.500
1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00   1st Qu.: 3.200
Median :203.0   Median : 90.0   Median : 46.00   Median : 4.200
Mean   :207.2   Mean   :107.3   Mean   : 50.27   Mean   : 4.525
3rd Qu.:229.0   3rd Qu.:107.8   3rd Qu.: 59.00   3rd Qu.: 5.400
Max.   :443.0   Max.   :385.0   Max.   :120.00   Max.   :19.300
     Age           Gender              Height          Weight
Min.   :19.00   Length:390         Min.   :52.00   Min.   : 99.0
1st Qu.:34.00   Class :character   1st Qu.:63.00   1st Qu.:150.2
Median :44.50   Mode  :character   Median :66.00   Median :173.0
Mean   :46.77                      Mean   :65.95   Mean   :177.4
3rd Qu.:60.00                      3rd Qu.:69.00   3rd Qu.:200.0
Max.   :92.00                      Max.   :76.00   Max.   :325.0
     BMI         Systolic.BP     Diastolic.BP        waist
Min.   :15.20   Min.   : 90.0   Min.   : 48.00   Min.   :26.00
1st Qu.:24.10   1st Qu.:122.0   1st Qu.: 75.00   1st Qu.:33.00
```

VISUALIZATION

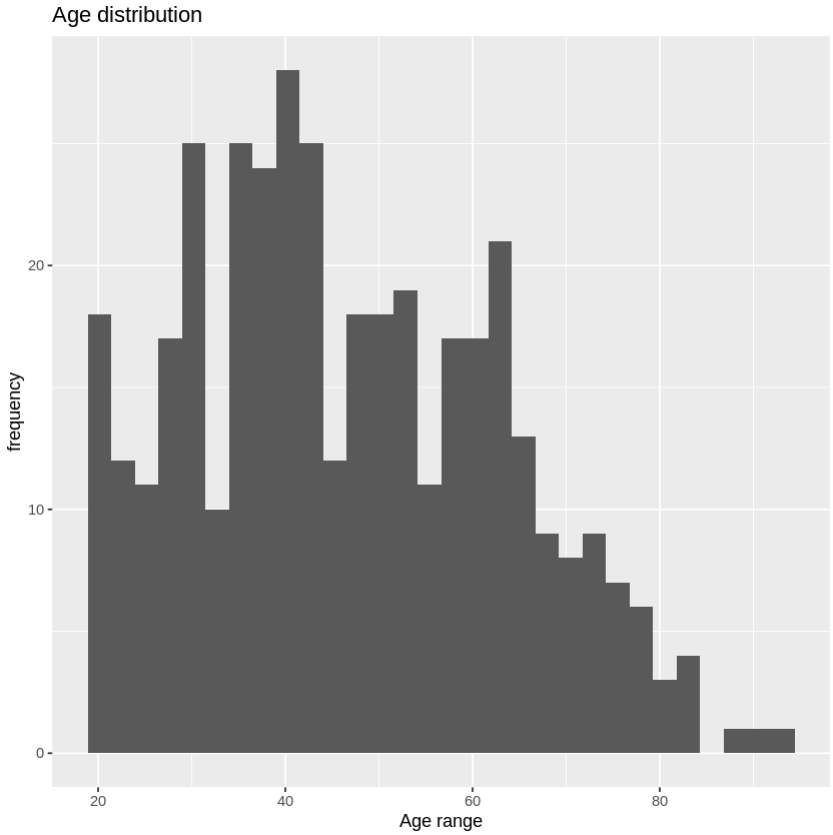
```
Max.   :55.80   Max.   :250.0   Max.   :124.00   Max.   :56.00
```

AGE DISTRIBUTION

```
1st Qu.:39.00   1st Qu.:0.8500   Class :character
```

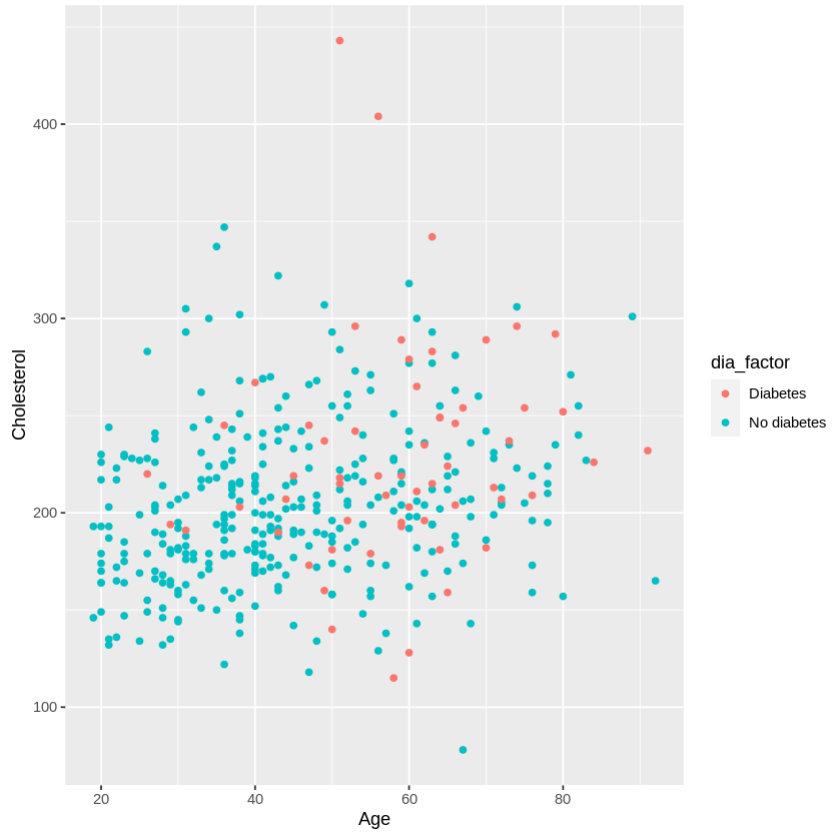
```
ggplot(dia_data,aes(x=Age))+geom_histogram()+labs(title ="Age distribution",x="Age range",y="
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



PLOTTING AGE VS CHOLESTROL

```
dia_factor=factor(dia_data$Diabetes,levels=c("Diabetes","No diabetes"),labels = c("Diabetes",
ggplot(dia_data,aes(x=Age,y=Cholesterol))+geom_point(aes(col=dia_factor))
```



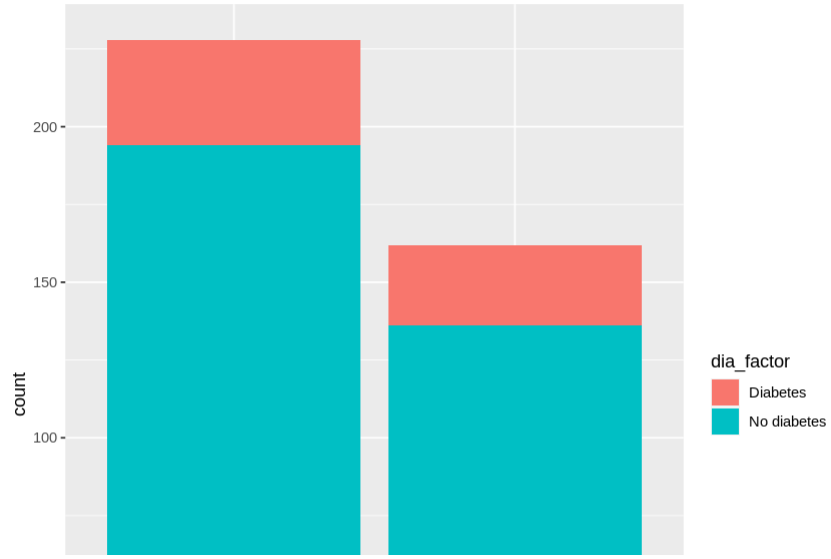
Double-click (or enter) to edit

```
unique(dia_data$Diabetes)
```

'No diabetes' · 'Diabetes'

```
k<-  dia_data%>%
    group_by(Diabetes)%>%
    dplyr::select(Gender)
ggplot(k,aes(x=k$Gender))+geom_bar(aes(fill=dia_factor),position = "stack")
```

```
Adding missing grouping variables: `Diabetes`
```



linear model for cholesterol

```
temp=subset(dia_data,select=-Gender)
temp=subset(temp,select=-Diabetes)
cor(temp)
```

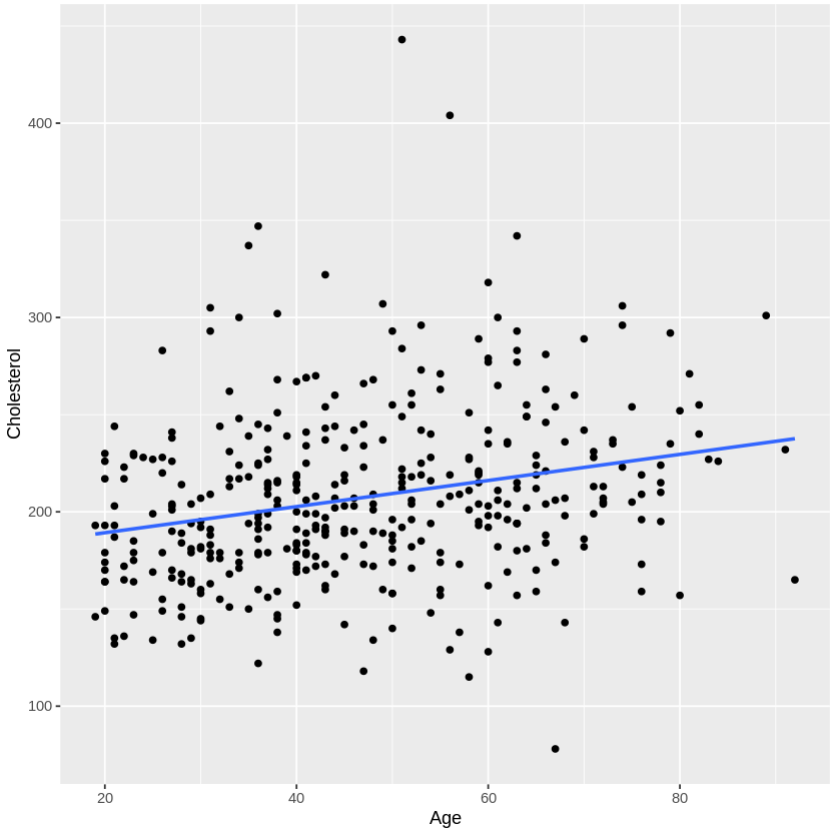
A matri

	Cholesterol	Glucose	HDL.Chol	Chol.HDL.ratio	Age	He
Cholesterol	1.00000000	0.15810208	0.19316170	0.47592687	0.247333470	-0.0636
Glucose	0.15810208	1.00000000	-0.15830196	0.28220951	0.294391967	0.0980
HDL.Chol	0.19316170	-0.15830196	1.00000000	-0.68186750	0.028209718	-0.0872
Chol.HDL.ratio	0.47592687	0.28220951	-0.68186750	1.00000000	0.163200861	0.0811
Age	0.24733347	0.29439197	0.02820972	0.16320086	1.000000000	-0.0822
Height	-0.06360077	0.09805180	-0.08723825	0.08116201	-0.082228781	1.0000
Weight	0.06235863	0.19035786	-0.29188280	0.27881232	-0.056783859	0.2553
BMI	0.09169469	0.12928649	-0.24186039	0.22840692	-0.009163800	-0.2595
Systolic.BP	0.20774144	0.16277716	0.03180658	0.11550522	0.453417229	-0.0407
Diastolic.BP	0.16624130	0.02026227	0.07834183	0.03824208	0.068648733	0.0436
waist	0.13403782	0.22233555	-0.27669673	0.31326175	0.150584800	0.0574
hip	0.09336358	0.13822294	-0.22383721	0.20890202	0.004675448	-0.0959
Waist.hip.ratio	0.09184679	0.18511730	-0.15877658	0.24332911	0.275187519	0.2525

```
library(tidyverse)
```

```
ggplot(dia_data,aes(x=Age,y=Cholesterol))+geom_point()+geom_smooth(method='lm',se=FALSE)
```

```
`geom_smooth()` using formula 'y ~ x'
```



✓ 2s completed at 08:28