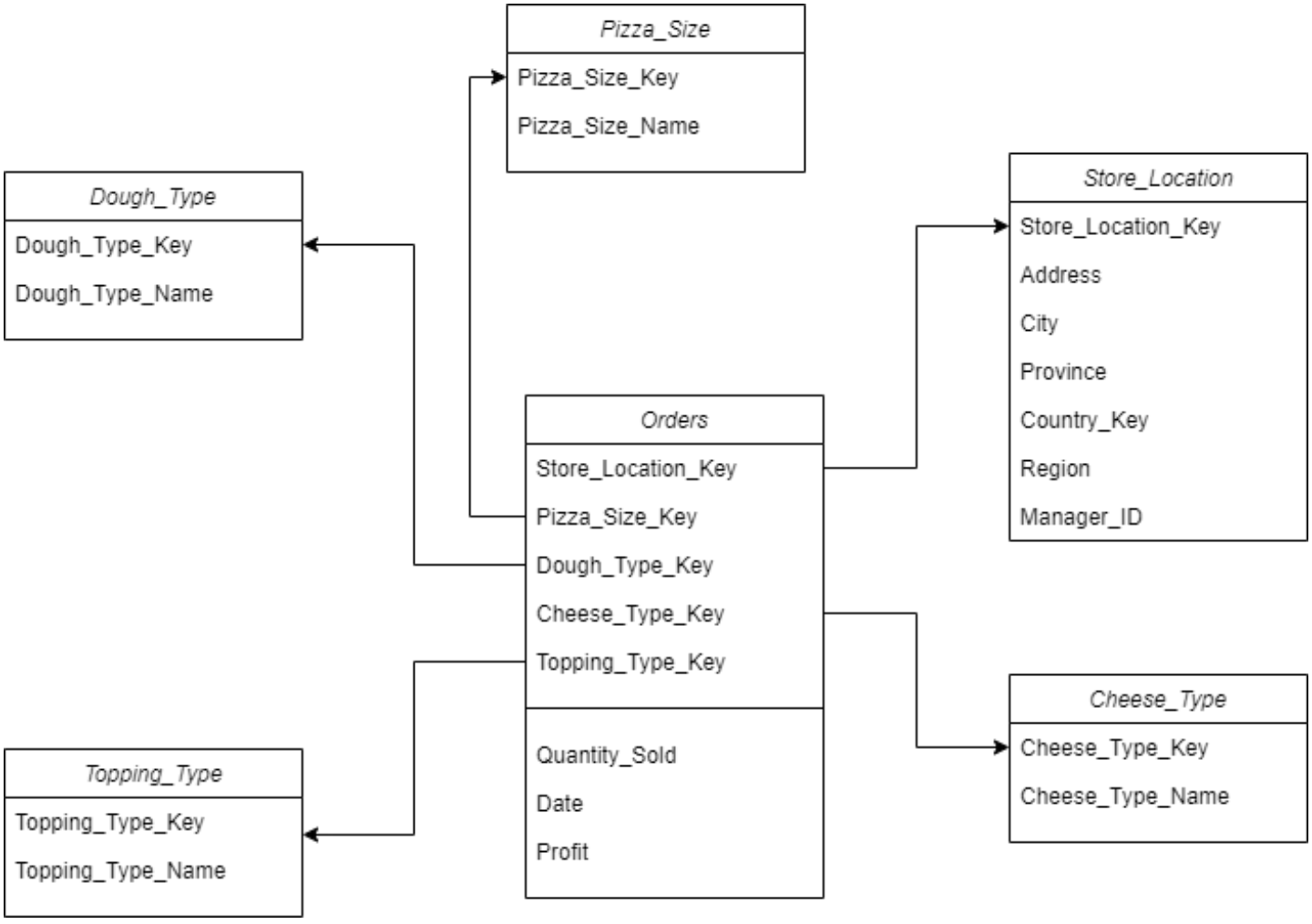
DTI 5126: Fundament

Summer 2021 - Assignment 1

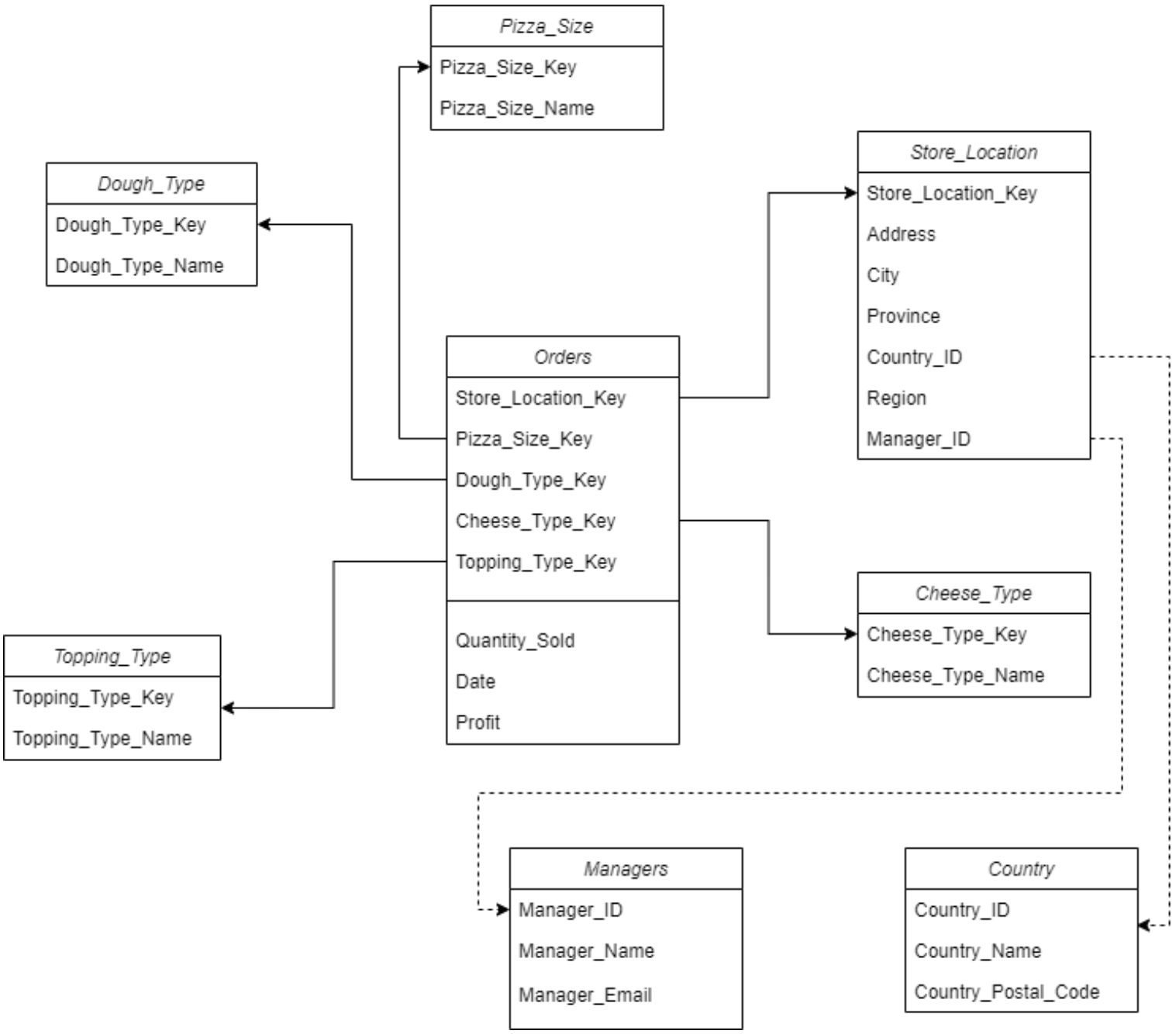
Name: Yomna Jehad Abdelsattar

Part A: Data Warehousing & OLAP

1) a. Star schema:



b. Snowflake schema:



c. Generate a set of sample data stored in csv files for the dimensions and fact table for the snowflake schema in c.

Files: **store_location.csv, Pizza_Size.csv, Dough_Type.csv, Cheese_Type.csv, Topping_Type.csv, orders.csv**

---

2) Using R, read the dimensions files and the profit fact table. Build an OLAP cube for your revenue and show the cells of a subset of the cells

A screenshot of a part of a subset of cells:

```
> # Showing the cells of the cube
> profit_cube
, , Dough_Type_Key = 1, Cheese_Type_Key = 1, Topping_Type_Key = 1, Quantity_Sold = 1, Date = 1

                   Pizza_Size_Key
Store_Location_Key  1  2  3  4  5
                 1 NA NA NA NA NA
                 2 NA NA NA NA NA
                 3 NA NA NA NA NA

, , Dough_Type_Key = 2, Cheese_Type_Key = 1, Topping_Type_Key = 1, Quantity_Sold = 1, Date = 1

                   Pizza_Size_Key
Store_Location_Key  1  2  3  4  5
                 1 NA NA NA NA NA
                 2 NA NA NA NA NA
                 3 NA NA NA NA NA

, , Dough_Type_Key = 3, Cheese_Type_Key = 1, Topping_Type_Key = 1, Quantity_Sold = 1, Date = 1

                   Pizza_Size_Key
Store_Location_Key  1  2  3  4  5
                 1 NA NA NA NA NA
                 2 NA NA NA NA NA
                 3 NA NA NA NA NA

, , Dough_Type_Key = 1, Cheese_Type_Key = 2, Topping_Type_Key = 1, Quantity_Sold = 1, Date = 1

                   Pizza_Size_Key
Store_Location_Key  1  2  3  4  5
                 1 NA NA NA NA NA
                 2 NA NA NA NA NA
                 3 NA NA NA NA NA

, , Dough_Type_Key = 2, Cheese_Type_Key = 2, Topping_Type_Key = 1, Quantity_Sold = 1, Date = 1

                   Pizza_Size_Key
Store_Location_Key  1  2  3  4  5
                 1 NA NA NA NA NA
                 2 NA NA NA NA NA
                 3 NA NA NA NA NA
```

---

3) Drilldown and roll-up operations that would lead to a conclusion about customers preferences

Rollup: Calculate the Profit of each Pizza Size sold in each store and collapse the other dimensions.
We conclude that in store location 1 (Address: 3 Abbas St.), people buy size 5 (xlarge) the most.
In store location 2 (Address: 4 Makram St.), people buy size 5 (xlarge) the most.
In store location 3 (5 Tayaran St.), people buy size 5 (xlarge)the most.
In summary: overall, people buy xlarge pizzas the most anywhere.

Drilldown: Calculate the profit of each combination of Cheese Type and Topping Type in each Store Location to explore which combination is popular in which store.

We find out that in store location 1 (Address: 3 Abbas St.), people buy the combination of Topping 4 (pepperoni) and cheese 2 (cheddar).
In store location 2 (Address: 4 Makram St.), people buy the combination of Topping 4(pepperoni) and cheese 2 (cheddar).
In store location 3 (5 Tayaran St.), people buy the combination of Topping 4(pepperoni) and cheese 2(cheddar).

We conclude that most people buy xlarge pizzas with pepperoni topping and cheddar cheese.

Part B: Data Preparation

1) Import the data set into RStudio and reduce the dataset to only four predictors (age, education, previous, and pdays), and the target, response.

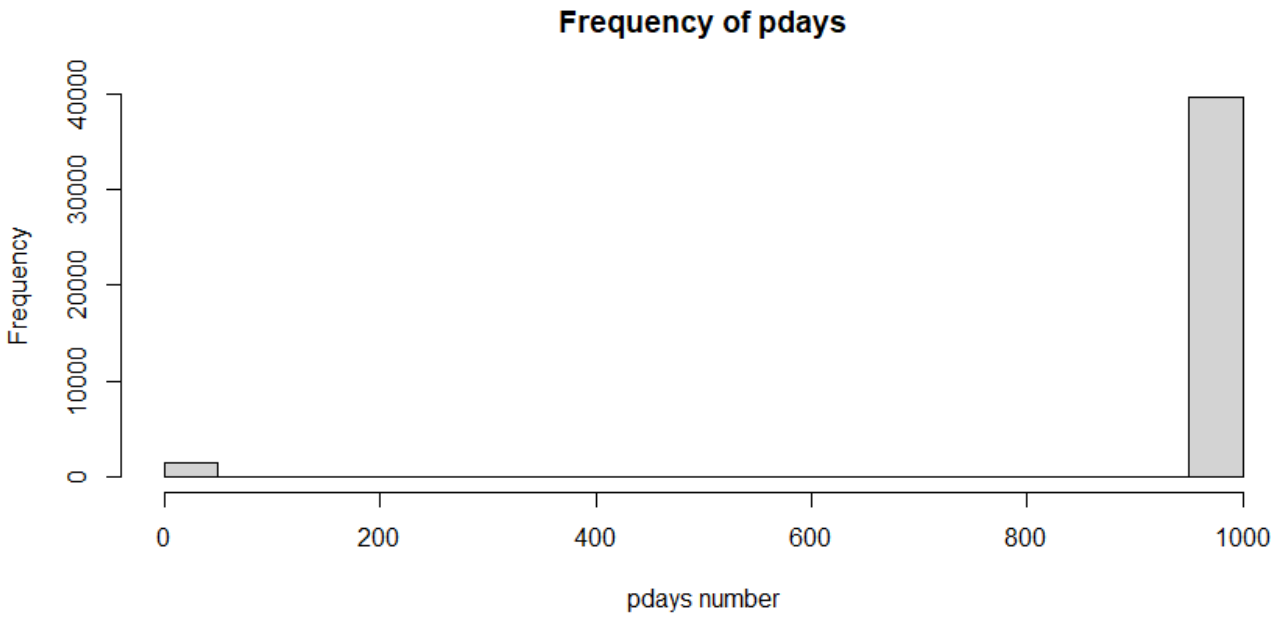
Dataset head screenshot:

```
> head(bnk)
  age   education previous pdays target
1  56    basic.4y        0   999     no
2  57 high.school        0   999     no
3  37 high.school        0   999     no
4  40    basic.6y        0   999     no
5  56 high.school        0   999     no
6  45    basic.9y        0   999     no
>
```

---

2) Change the field value 999 to “NA” to represent missing values.
3) Explain why the field pdays is essentially useless until you handle the 999 code.

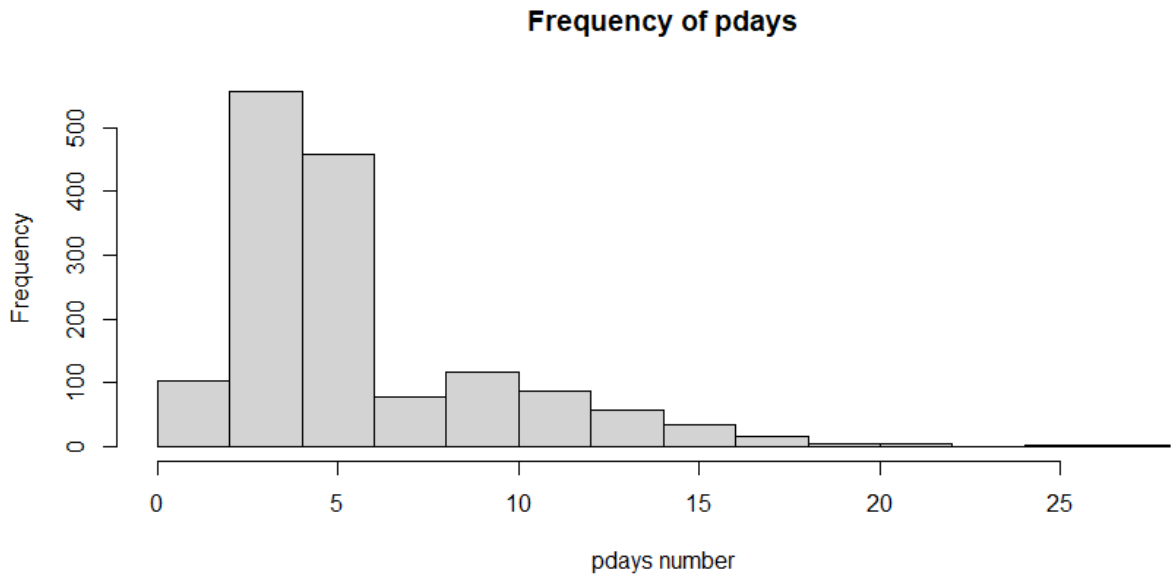
We note that the value 999 exists in the pdays column very commonly, we plot a histogram to investigate



As we can see the highly unbalanced distribution of values (giving 999 which is a very high value instead of a null) makes very important values insignificant, thus we need to properly put a null in the 999 place.

---

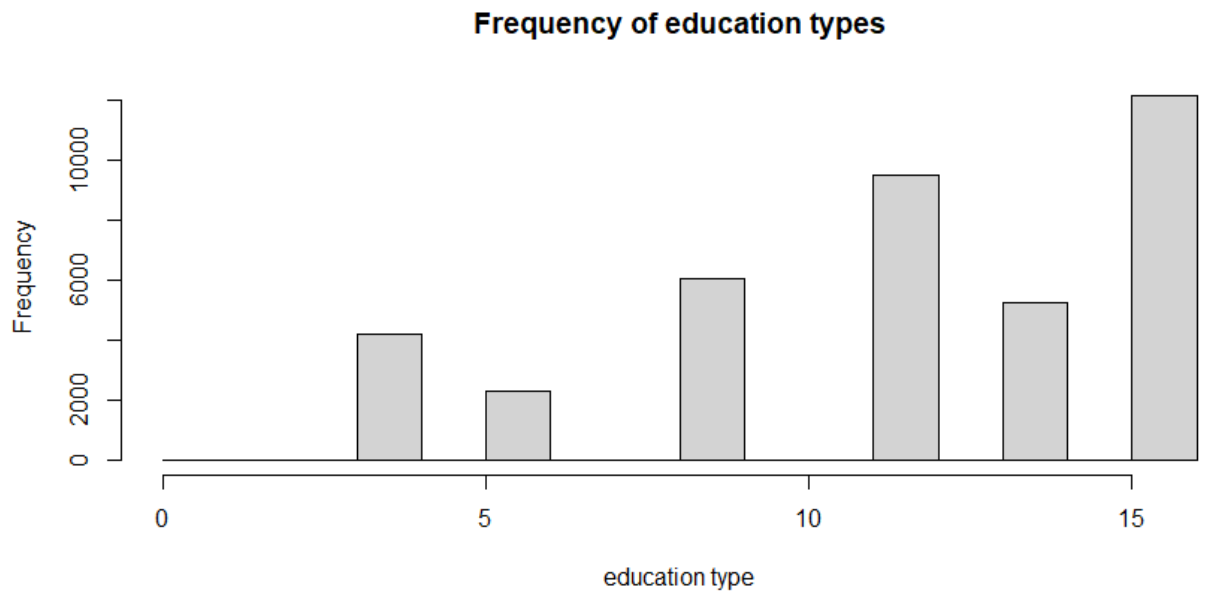
4) After that, we can view the histogram again



Now we can clearly see the important values and distinguish between them.

---

5) Transform the data values of the education field into numeric values using the chart in Table 1.
We plot a histogram for the education field to view the numeric values after the transformation

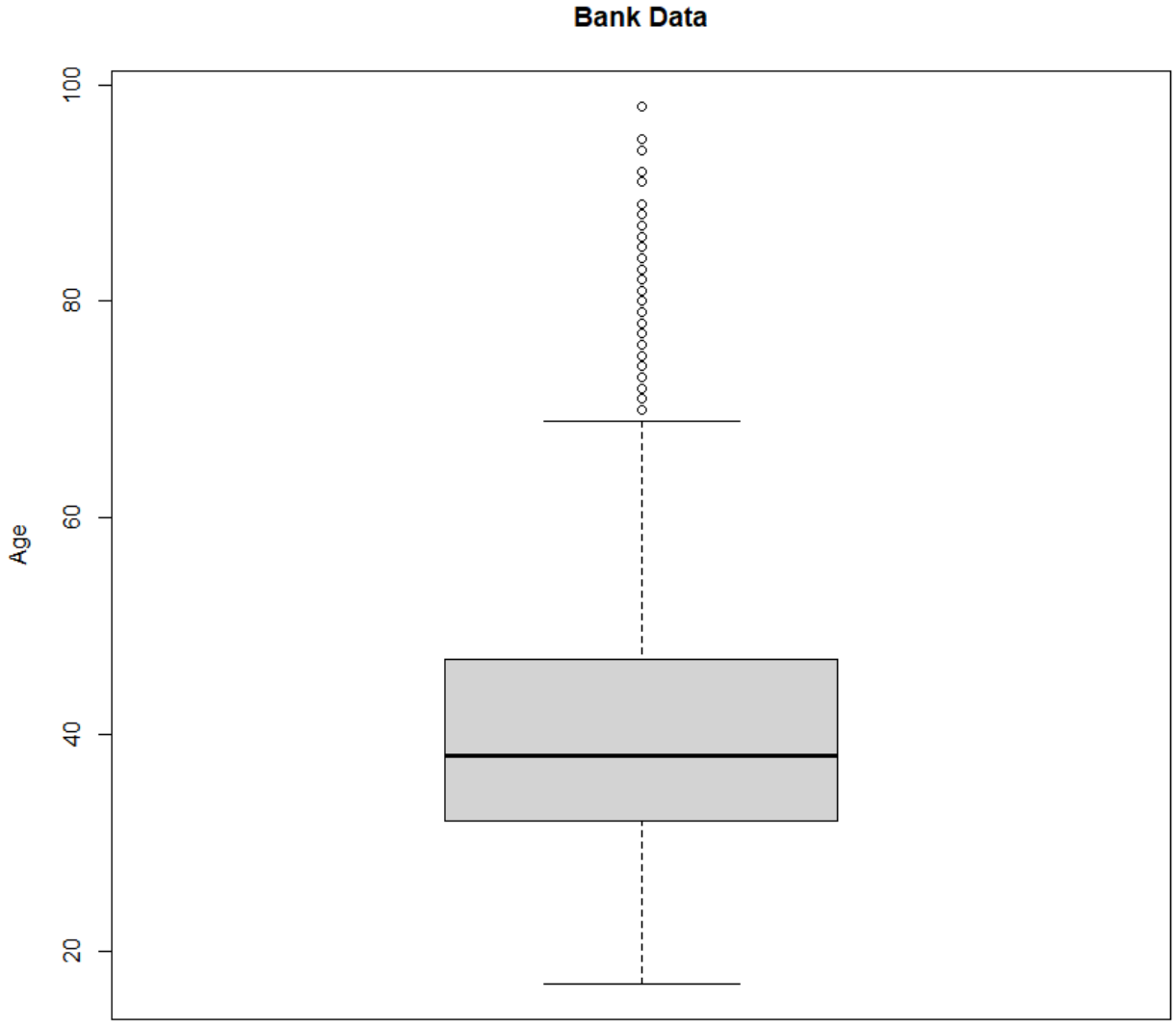


---

6) Compute the mean, median & mode of the age variable. Using a boxplot, give the five number summary of the data. Plot the quantile information.

```
> mfv(bnk$age) #mode
[1] 31
>
> mean(bnk$age) #mean
[1] 40.02406
> median(bnk$age) #median
[1] 38
```

BOXPLOT:

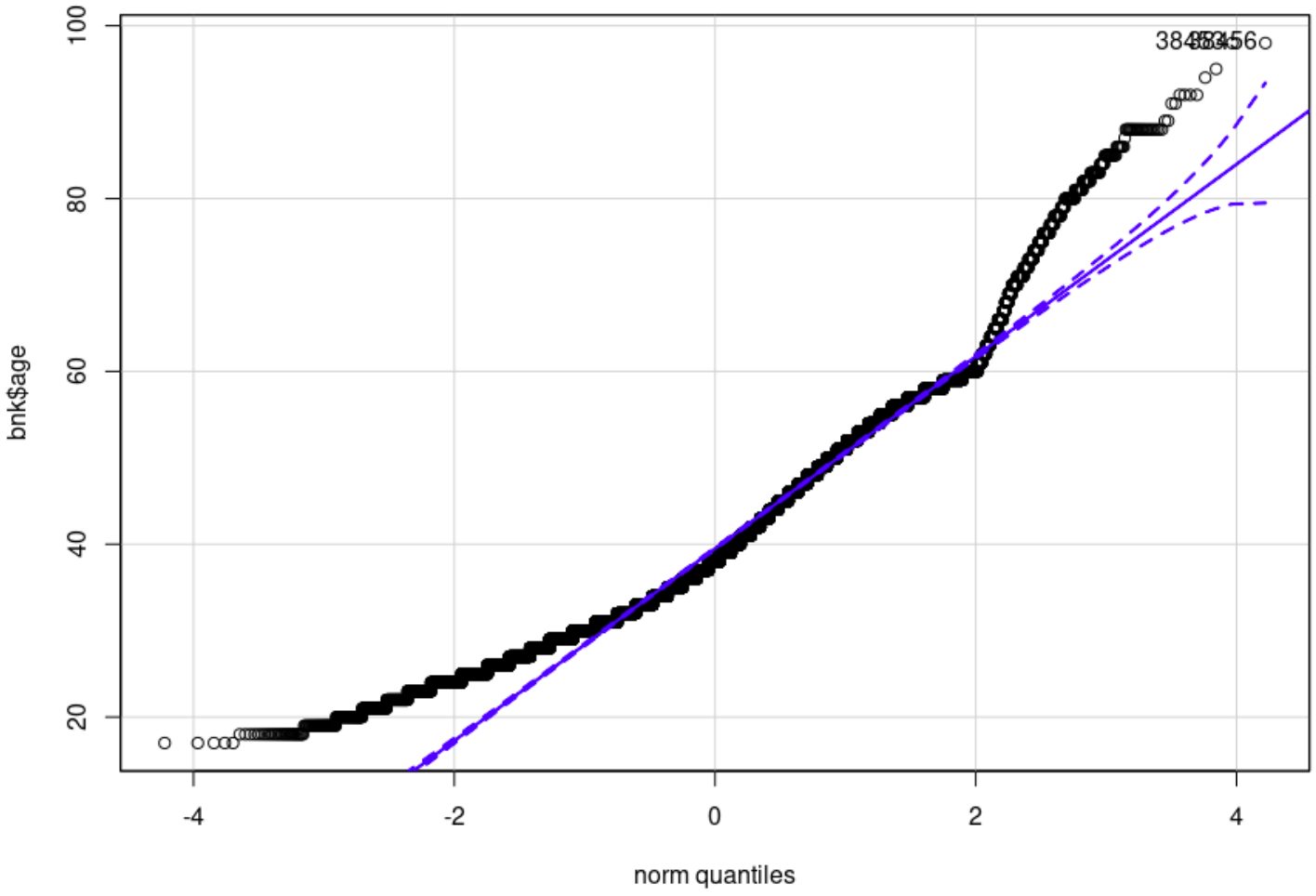

Bank Data
Age
100
80
60
40
20

5 number summary:

```
> #First quartile:
> quantile(bnk$age, 0.25)
25%
 32
>
> #Second quartile or median:
> quantile(bnk$age, 0.5)
50%
 38
>
>
> #Third quartile:
> quantile(bnk$age, 0.75)
75%
 47
>
> max(bnk$age) #max
[1] 98
> min(bnk$age) #min
[1] 17
```

```
> #The interquartile range is the difference between the 75th percentile and 25th percentile
> quantile(bnk$age, 0.75, names = FALSE) - quantile(bnk$age, 0.25, names =FALSE)
[1] 15
>
> #OR
> IQR(bnk$age)
[1] 15
```

qqplot:

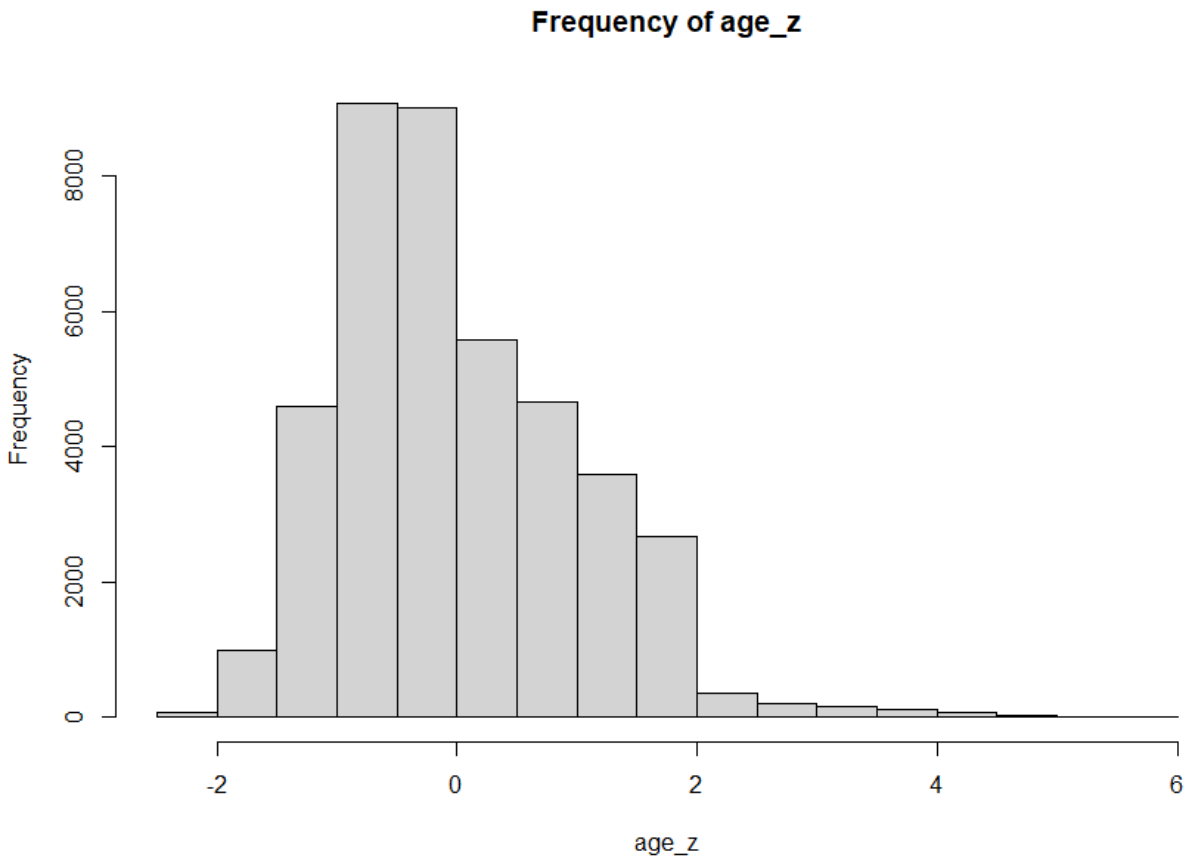


7) Standardize the age variable and save it as a new variable, age_z.

```
attr(,"scaled:center")
[1] 40.02406
attr(,"scaled:scale")
[1] 10.42125
>
```

8) Obtain a listing of all records that are outliers according to the field age_z.

I first plot the histogram to visualize the possible outliers range



Visually detected range: age>3, age <-3

A screenshot of some of outlier records:

```
> bnk_outliers
      age education previous pdays target    age_z
27758  76        16        0    NA     no 3.452171
27781  73        16        1    NA     no 3.164298
27801  88         4        0    NA     no 4.603665
27803  88         4        0    NA    yes 4.603665
27806  88         4        0    NA    yes 4.603665
27809  88         4        0    NA     no 4.603665
27811  88         4        0    NA    yes 4.603665
27812  88         4        0    NA    yes 4.603665
27813  88         4        0    NA     no 4.603665
27814  88         4        0    NA    yes 4.603665
27815  88         4        0    NA     no 4.603665
27816  88         4        0    NA     no 4.603665
27817  88         4        0    NA    yes 4.603665
27818  88         4        0    NA    yes 4.603665
27819  88         4        0    NA    yes 4.603665
27827  95         6        0    NA     no 5.275369
```

NOTE:
Provided in the zip files: the .csv files, the code .R files and the diagrams .PNG files.