NAME : Hariprasad K K

REG NO : *19BCE7079*

COURSE : *DATA_ANALYTICS*

DATE : *05-11-2021*

1)Data cleaning
On the “findata” dataset, perform data cleaning using the Janitor package.
On the “mental-health-in-tech-2016_20161114 "dataset, perform data cleaning
operations related to mislabeled variables, changing faulty data types, and identifying
duplicated and distinct values.

CODE:

```
data<-read.csv("findata.csv") data
clean<-clean_names(my_data) clean
colnames(clean) tabyl(clean, county_state)
clean %>% tabyl(county_state) %>% adorn_pct_formatting(digits
=2,affix_sign=TRUE) clean %>% get_dupes(county_state)
```

```
> data<-read.csv("findata.csv")
> data
```

	County..State	Type	Value
1	Autauga, AL	Temp	66.0
2	Baldwin, AL	Temp	68.7
3	Barbour, AL	Temp	66.8
4	Bibb, AL	Temp	64.4
5	Blount, AL	Temp	63.2
6	Bullock, AL	Temp	66.3
7	Butler, AL	Temp	66.8
8	Calhoun, AL	Temp	63.4
9	Chambers, AL	Temp	63.8
10	Cherokee, AL	Temp	62.7
11	Chilton, AL	Temp	64.4
12	Choctaw, AL	Temp	66.4
13	Clarke, AL	Temp	66.9
14	Clay, AL	Temp	62.9
15	Cleburne, AL	Temp	62.6
16	Coffee, AL	Temp	67.5
17	Colbert, AL	Temp	62.0
18	Conecuh, AL	Temp	67.9
19	Coosa, AL	Temp	64.1
20	Covington, AL	Temp	67.7
21	Crenshaw, AL	Temp	66.4
22	Cullman, AL	Temp	62.7
23	Dale, AL	Temp	67.8
24	Dallas, AL	Temp	66.3
25	DeKalb, AL	Temp	61.1
26	Elmore, AL	Temp	65.8
27	Escambia, AL	Temp	68.3
28	Etowah, AL	Temp	63.2
29	Fayette, AL	Temp	63.6
30	Franklin, AL	Temp	61.8
31	Geneva, AL	Temp	68.5
32	Greene, AL	Temp	65.2
33	Hale, AL	Temp	65.2
34	Henry, AL	Temp	67.8
35	Houston, AL	Temp	68.8
36	Jackson, AL	Temp	61.1
37	Jefferson, AL	Temp	64.2
38	Lamar, AL	Temp	63.5

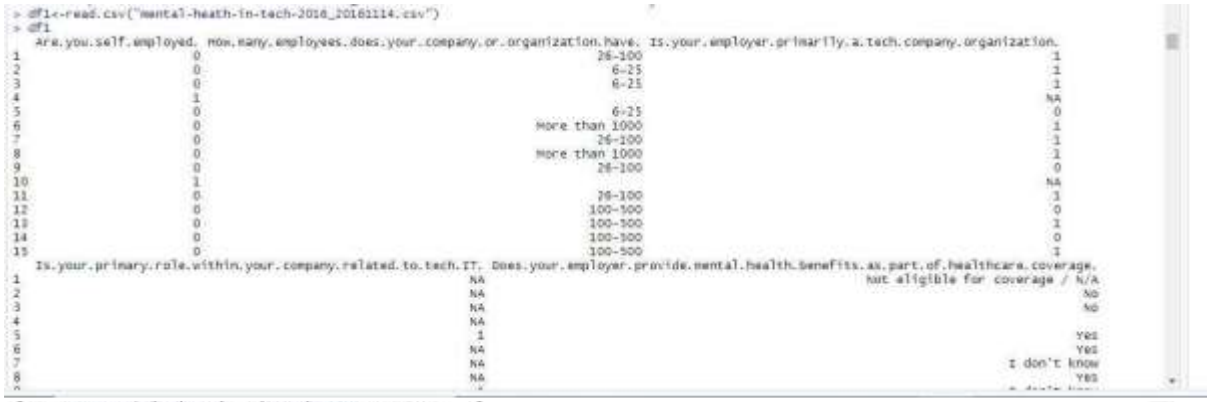
```
> clean<-clean_names(data)
> colnames(clean)
[1] "county_state" "type"         "value"
>
```

```
> tabyl(clean, county_state)
                    county_state n       percent
                 Abbeville, SC 5 3.365644e-04
                    Acadia, LA 5 3.365644e-04
                  Accomack, VA 5 3.365644e-04
                       Ada, ID 5 3.365644e-04
                     Adair, IA 4 2.692515e-04
                     Adair, KY 5 3.365644e-04
                     Adair, MO 5 3.365644e-04
                     Adair, OK 4 2.692515e-04
                     Adams, CO 5 3.365644e-04
                     Adams, IA 4 2.692515e-04
                     Adams, ID 5 3.365644e-04
                     Adams, IL 5 3.365644e-04
                     Adams, IN 5 3.365644e-04
                     Adams, MS 5 3.365644e-04
                     Adams, ND 4 2.692515e-04
                     Adams, NE 4 2.692515e-04
                     Adams, OH 5 3.365644e-04
                     Adams, PA 5 3.365644e-04
                     Adams, WA 5 3.365644e-04
                     Adams, WI 5 3.365644e-04
                   Addison, VT 5 3.365644e-04
                     Aiken, SC 5 3.365644e-04
                    Aitkin, MN 5 3.365644e-04
                   Alachua, FL 5 3.365644e-04
                  Alamance, NC 5 3.365644e-04
                   Alameda, CA 5 3.365644e-04
                   Alamosa, CO 5 3.365644e-04
                    Albany, NY 5 3.365644e-04
                    Albany, WY 5 3.365644e-04
> clean %>% tabyl(county_state) %>% adorn_pct_formatting(digits =2,affix_sign=TRUE)
                    county_state n percent
                 Abbeville, SC 5   0.03%
                    Acadia, LA 5   0.03%
                  Accomack, VA 5   0.03%
                       Ada, ID 5   0.03%
                     Adair, IA 4   0.03%
                     Adair, KY 5   0.03%
                     Adair, MO 5   0.03%
                     Adair, OK 4   0.03%
                     Adams, CO 5   0.03%
                     Adams, IA 4   0.03%
                     Adams, ID 5   0.03%
                     Adams, IL 5   0.03%
                     Adams, IN 5   0.03%
                     Adams, MS 5   0.03%
                     Adams, ND 4   0.03%
                     Adams, NE 4   0.03%
                     Adams, OH 5   0.03%
                     Adams, PA 5   0.03%
                     Adams, WA 5   0.03%
                     Adams, WI 5   0.03%
                   Addison, VT 5   0.03%
                     Aiken, SC 5   0.03%
                    Aitkin, MN 5   0.03%
                   Alachua, FL 5   0.03%
```

```
> clean %>% get_dupes(county_state)
                county_state dupe_count       type         value
1                Abbeville, SC          5       Temp  6.350000e+01
2                Abbeville, SC          5      Crime  5.118567e+02
3                Abbeville, SC          5   GradRate  8.800000e-01
4                Abbeville, SC          5  Political  2.804640e-01
5                Abbeville, SC          5 HousPrices  2.090000e+05
6                   Acadia, LA          5       Temp  6.870000e+01
7                   Acadia, LA          5      Crime  1.639766e+02
8                   Acadia, LA          5   GradRate  9.100000e-01
9                   Acadia, LA          5  Political  5.667969e-01
10                  Acadia, LA          5 HousPrices  1.450000e+05
11                Accomack, VA          5       Temp  5.960000e+01
12                Accomack, VA          5      Crime  1.900567e+02
13                Accomack, VA          5   GradRate  9.000000e-01
14                Accomack, VA          5  Political  1.165128e-01
15                Accomack, VA          5 HousPrices  1.899500e+05
16                     Ada, ID          5       Temp  5.160000e+01
17                     Ada, ID          5      Crime  2.065004e+02
18                     Ada, ID          5   GradRate  7.700000e-01
19                     Ada, ID          5  Political  9.237539e-02
20                     Ada, ID          5 HousPrices  3.799000e+05
21                   Adair, IA          4       Temp  4.710000e+01
22                   Adair, IA          4      Crime  6.691649e+01
23                   Adair, IA          4   GradRate  9.700000e-01
24                   Adair, IA          4  Political  3.484650e-01
25                   Adair, KY          5       Temp  5.870000e+01
26                   Adair, KY          5      Crime  9.609225e+01
27                   Adair, KY          5   GradRate  9.700000e-01
28                   Adair, KY          5  Political  6.454512e-01
29                   Adair, KY          5 HousPrices  1.499000e+05
30                   Adair, MO          5       Temp  5.110000e+01
31                   Adair, MO          5      Crime  2.150790e+02
32                   Adair, MO          5   GradRate  9.100000e-01
33                   Adair, MO          5  Political  2.474086e-01
34                   Adair, MO          5 HousPrices  1.199500e+05
```

Dataset2

df1<-read.csv("mental-heath-in-tech-2016_20161114.csv") df1
file.info("mental-heath-in-tech-2016_20161114.csv")$size str(df1)
summary(df1)
df1 <- df1 %>% rename(employees =
How.many.employees.does.your.company.or.organization.have.)
colnames(df1)
typeof(df1$employees)
df1$employees <- as.factor(df1$employees)
typeof(df1$employees)
df1 <- df1[!duplicated(df1$ID_Column_Name),]
df1 <- df1 %>% distinct(ID_Column_Name, .keep_all = TRUE)



```
> df1<-read.csv("mental-heath-in-tech-2016_20161114.csv")
> df1
   Are.you.self.employed. How.many.employees.does.your.company.or.organization.have. Is.your.employer.primarily.a.tech.company.organization.
1                       0                                                    26-100                                                  1
2                       0                                                     6-25                                                   1
3                       0                                                     6-25                                                   1
4                       1                                                                                                            NA
5                       0                                                     6-25                                                   0
6                       0                                                More than 1000                                              1
7                       0                                                    26-100                                                  1
8                       0                                                More than 1000                                              1
9                       0                                                    26-100                                                  0
10                      1                                                                                                            NA
11                      0                                                    26-100                                                  1
12                      0                                                   100-500                                                  0
13                      0                                                   100-500                                                  1
14                      0                                                   100-500                                                  0
15                      0                                                   100-500                                                  1
   Is.your.primary.role.within.your.company.related.to.tech.IT. Does.your.employer.provide.mental.health.benefits.as.part.of.healthcare.coverage.
1                                                            NA                                                 Not eligible for coverage / N/A
2                                                            NA                                                                             No
3                                                            NA                                                                             No
4                                                            NA                                                                               
5                                                             1                                                                            Yes
6                                                            NA                                                                            Yes
7                                                            NA                                                              I don't know
8                                                            NA                                                                            Yes
```

```
> file.info("mental-health-in-tech-2016_20161114.csv")$size
[1] 1104203
> str(df1)
'data.frame':   1433 obs. of  63 variables:
 $ Are.you.self.employed.
                                : int  0 0 0 1 0 0 0 0 0 1 ...
 $ How.many.employees.does.your.company.or.organization.have.
                                : Factor w/ 7 levels "","1-5","100-500",..: 4 6 6 1 6 7 4 7 4 1 ...
 $ Is.your.employer.primarily.a.tech.company.organization.
                                : int  1 1 1 NA 0 1 1 1 0 NA ...
 $ Is.your.primary.role.within.your.company.related.to.tech.IT.
                                : int  NA NA NA NA 1 NA NA NA 1 NA ...
 $ Does.your.employer.provide.mental.health.benefits.as.part.of.healthcare.coverage.
                                : Factor w/ 5 levels "","I don't know",..: 4 3 3 1 5 5 2 5 2 1 ...
 $ Do.you.know.the.options.for.mental.health.care.available.under.your.employer.provided.coverage.
                                : Factor w/ 5 levels "","I am not sure",..: 3 1 1 1 5 2 4 5 4 1 ...
 $ Has.your.employer.ever.formally.discussed.mental.health..for.example..as.part.of.a.wellness.campaign.or.other.official.communication..
                                : Factor w/ 4 levels "","I don't know",..: 3 4 3 1 3 3 3 3 3 1 ...
 $ Does.your.employer.offer.resources.to.learn.more.about.mental.health.concerns.and.options.for.seeking.help.
                                : Factor w/ 4 levels "","I don't know",..: 3 4 3 1 3 4 3 4 3 1 ...
 $ Is.your.anonymity.protected.if.you.choose.to.take.advantage.of.mental.health.or.substance.abuse.treatment.resources.provided.by.your.employer.
                                : Factor w/ 4 levels "","I don't know",..: 2 4 2 1 3 4 2 4 2 1 ...
 $ If.a.mental.health.issue.prompted.you.to.request.a.medical.leave.from.work..asking.for.that.leave.would.be.
                                : Factor w/ 7 levels "","I don't know",..: 7 5 3 1 3 5 5 7 6 1 ...
 $ Do.you.think.that.discussing.a.mental.health.disorder.with.your.employer.would.have.negative.consequences.
                                : Factor w/ 4 levels "","Maybe","No",..: 3 3 2 1 4 4 3 3 4 1 ...
 $ Do.you.think.that.discussing.a.physical.health.issue.with.your.employer.would.have.negative.consequences.
                                : Factor w/ 4 levels "","Maybe","No",..: 3 3 3 1 2 4 3 3 4 1 ...
```

```
> summary(df1)
 Are.you.self.employed. How.many.employees.does.your.company.or.organization.have. Is.your.employer.primarily.a.tech.company.organization.
 Min.   :0.0000                 :287                                                 Min.   :0.0000
 1st Qu.:0.0000         1-5     : 60                                                 1st Qu.:1.0000
 Median :0.0000         100-500 :248                                                 Median :1.0000
 Mean   :0.2003         26-100  :292                                                 Mean   :0.7705
 3rd Qu.:0.0000         500-1000 : 80                                                3rd Qu.:1.0000
 Max.   :1.0000         6-25    :210                                                 Max.   :1.0000
                        More than 1000:256                                           NA's   :287
 Is.your.primary.role.within.your.company.related.to.tech.IT. Does.your.employer.provide.mental.health.benefits.as.part.of.healthcare.coverage.
 Min.   :0.000                                                                                :287
 1st Qu.:1.000                                                I don't know                    :319
 Median :1.000                                                No                              :213
 Mean   :0.943                                                Not eligible for coverage / N/A: 83
 3rd Qu.:1.000                                                Yes                             :531
 Max.   :1.000
 NA's   :1170
 Do.you.know.the.options.for.mental.health.care.available.under.your.employer.provided.coverage.
                :287
 I am not sure:352
 N/A            :133
 No             :354
```

```
> df1 <- df1 %>% rename(employees = How.many.employees.does.your.company.or.organization.have.)
> colnames(df1)
 [1] "Are.you.self.employed."

 [2] "employees"

 [3] "Is.your.employer.primarily.a.tech.company.organization."

 [4] "Is.your.primary.role.within.your.company.related.to.tech.IT."

 [5] "Does.your.employer.provide.mental.health.benefits.as.part.of.healthcare.coverage."

 [6] "Do.you.know.the.options.for.mental.health.care.available.under.your.employer.provided.coverage."

 [7] "Has.your.employer.ever.formally.discussed.mental.health..for.example..as.part.of.a.wellness.campaign.or.other.official.communication.."

 [8] "Does.your.employer.offer.resources.to.learn.more.about.mental.health.concerns.and.options.for.seeking.help."

 [9] "Is.your.anonymity.protected.if.you.choose.to.take.advantage.of.mental.health.or.substance.abuse.treatment.resources.provided.by.your.employer."

[10] "If.a.mental.health.issue.prompted.you.to.request.a.medical.leave.from.work..asking.for.that.leave.would.be."

[11] "Do.you.think.that.discussing.a.mental.health.disorder.with.your.employer.would.have.negative.consequences."

[12] "Do.you.think.that.discussing.a.physical.health.issue.with.your.employer.would.have.negative.consequences."
```

```
> typeof(df1$employees)
[1] "integer"
> df1$employees <- as.factor(df1$employees)
> typeof(df1$employees)
[1] "integer"
```

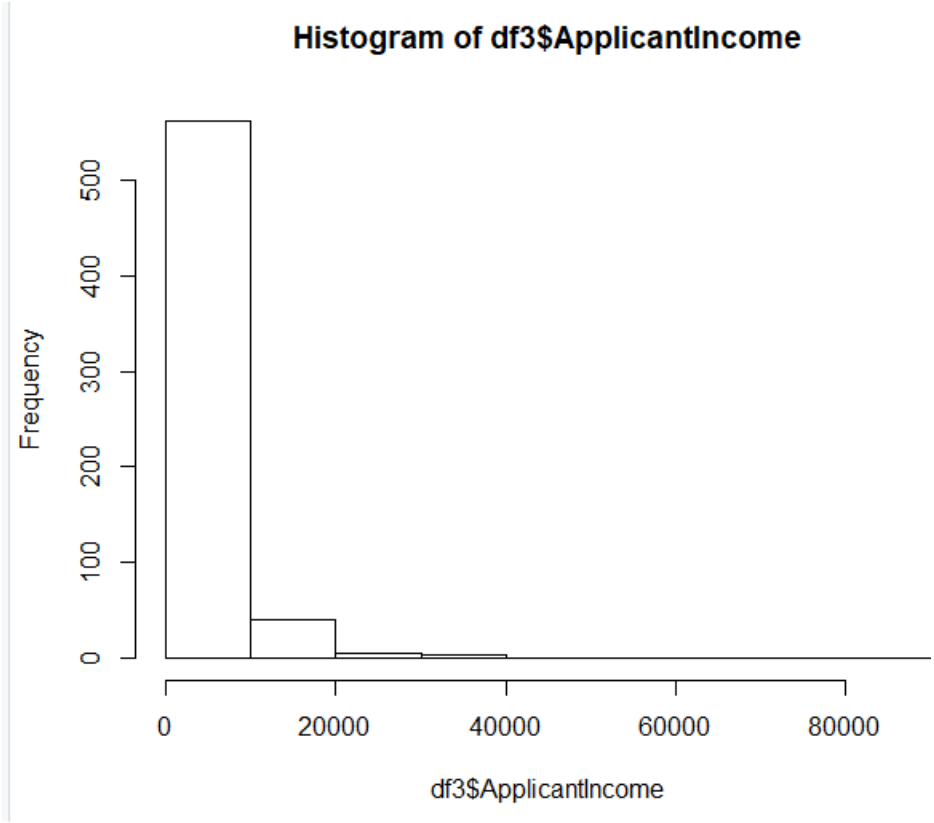
DATASET3

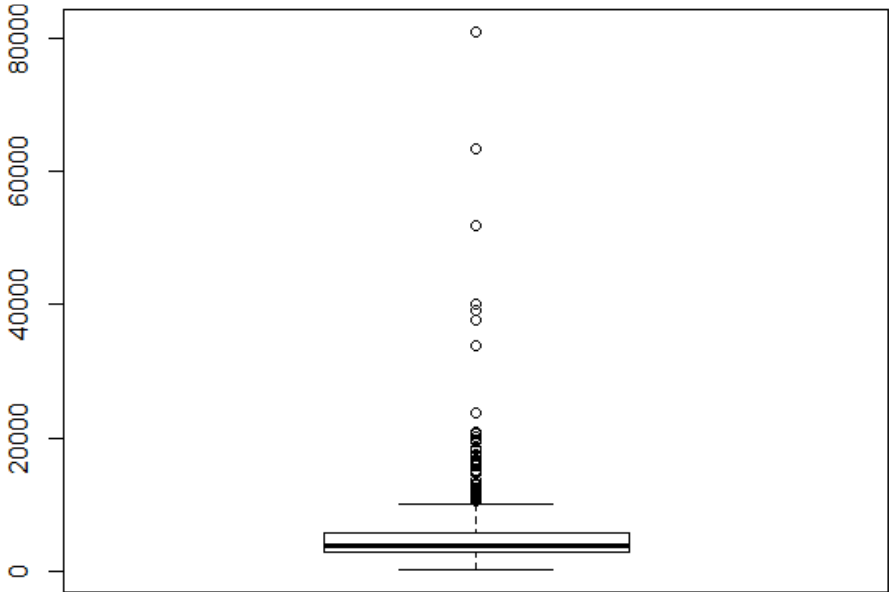
df3<-read.csv("loan.csv
") df3
class(df3)

```
dim(df3)
summary(df
3)
hist(df3$ApplicantIncome)
boxplot(df3$ApplicantIncome)
df3$Gender<-str_trim(df3$Gend
er)
df3$Loan_Status<-str_replace(df3$Loan_Status,"Y",
"Yes") any(is.na(df3))
sum(is.na(df3$LoanAmoun
t)) na.omit(df3)
df3[is.na(df3)]<- 0
df3$Dependents[is.na(df3$Dependents)]<-0
df3$ApplicantIncome[is.na(df3$ApplicantIncome)]<-median(df3$ApplicantIncome,na
.rm = "TRUE")
data1<-unite(data = df3,col = Married,Education,Gender)
```

```
> df3<-read.csv("loan.csv")
> df3
    Loan_ID Gender Married Dependents     Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History
1  LP001002   Male      No          0      Graduate            No             5849                0          NA               360              1
2  LP001003   Male     Yes          1      Graduate            No             4583             1508         128               360              1
3  LP001005   Male     Yes          0      Graduate           Yes             3000                0          66               360              1
4  LP001006   Male     Yes          0  Not Graduate            No             2583             2358         120               360              1
5  LP001008   Male      No          0      Graduate            No             6000                0         141               360              1
6  LP001011   Male     Yes          2      Graduate           Yes             5417             4196         267               360              1
7  LP001013   Male     Yes          0  Not Graduate            No             2333             1516          95               360              1
8  LP001014   Male     Yes         3+      Graduate            No             3036             2504         158               360              0
9  LP001018   Male     Yes          2      Graduate            No             4006             1526         168               360              1
10 LP001020   Male     Yes          1      Graduate            No            12841            10968         349               360              1
11 LP001024   Male     Yes          2      Graduate            No             3200              700          70               360              1
12 LP001027   Male     Yes          2      Graduate                            2500             1840         109               360              1
13 LP001028   Male     Yes          2      Graduate            No             3073             8106         200               360              1
14 LP001029   Male      No          0      Graduate            No             1853             2840         114               360              1
15 LP001030   Male     Yes          2      Graduate            No             1299             1086          17               120              1
16 LP001032   Male      No          0      Graduate            No             4950                0         125               360              1
17 LP001034   Male      No          1  Not Graduate            No             3596                0         100               240             NA
18 LP001036 Female      No          0      Graduate            No             3510                0          76               360              0
19 LP001038   Male     Yes          0  Not Graduate            No             4887                0         133               360              1
20 LP001041   Male     Yes          0      Graduate                            2600             3500         115                NA              1
21 LP001043   Male     Yes          0  Not Graduate            No             7660                0         104               360              0
22 LP001046   Male     Yes          1      Graduate            No             5955             5625         315               360              1
23 LP001047   Male     Yes          0  Not Graduate            No             2600             1911         116               360              0
24 LP001050             Yes          2  Not Graduate            No             3365             1917         112               360              0
25 LP001052   Male     Yes          1      Graduate                            3717             2925         151               360             NA
26 LP001066   Male     Yes          0      Graduate           Yes             9560                0         191               360              1
```

```
> class(df3)
[1] "data.frame"
> dim(df3)
[1] 614  13
> summary(df3)
     Loan_ID       Gender    Married   Dependents        Education   Self_Employed ApplicantIncome CoapplicantIncome   LoanAmount
 LP001002:  1          : 13      :  3   : 15      Graduate    :480      : 32       Min.   :  150   Min.   :    0   Min.   :  9.0
 LP001003:  1   Female:112   No :213   0 :345      Not Graduate:134   No :500       1st Qu.: 2878   1st Qu.:    0   1st Qu.:100.0
 LP001005:  1   Male  :489   Yes:398   1 :102                         Yes: 82       Median : 3812   Median : 1188   Median :128.0
 LP001006:  1                         2 :101                                       Mean   : 5403   Mean   : 1621   Mean   :146.4
 LP001008:  1                         3+: 51                                       3rd Qu.: 5795   3rd Qu.: 2297   3rd Qu.:168.0
 LP001011:  1                                                                      Max.   :81000   Max.   :41667   Max.   :700.0
 (Other) :608                                                                                                      NA's   :22
 Loan_Amount_Term Credit_History     Property_Area Loan_Status
 Min.   : 12      Min.   :0.0000   Rural    :179   N:192
 1st Qu.:360      1st Qu.:1.0000   Semiurban:233   Y:422
 Median :360      Median :1.0000   Urban    :202
 Mean   :342      Mean   :0.8422
 3rd Qu.:360      3rd Qu.:1.0000
 Max.   :480      Max.   :1.0000
 NA's   :14       NA's   :50
> hist(df3$ApplicantIncome)
>
```


Histogram of df3$ApplicantIncome
Frequency
0
100
200
300
400
500
0
20000
40000
60000
80000
df3$ApplicantIncome


0
20000
40000
60000
80000

```
> any(is.na(df3))
[1] TRUE
> sum(is.na(df3$LoanAmount))
[1] 22
> na.omit(df3)
    Loan_ID Gender Married Dependents     Education Self_Employed ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History
2  LP001003   Male     Yes          1      Graduate            No            4583             1508         128              360              1
3  LP001005   Male     Yes          0      Graduate           Yes            3000                0          66              360              1
4  LP001006   Male     Yes          0  Not Graduate            No            2583             2358         120              360              1
5  LP001008   Male      No          0      Graduate            No            6000                0         141              360              1
6  LP001011   Male     Yes          2      Graduate           Yes            5417             4196         267              360              1
7  LP001013   Male     Yes          0  Not Graduate            No            2333             1516          95              360              1
8  LP001014   Male     Yes         3+      Graduate            No            3036             2504         158              360              0
9  LP001018   Male     Yes          2      Graduate            No            4006             1526         168              360              1
10 LP001020   Male     Yes          1      Graduate            No           12841            10968         349              360              1
11 LP001024   Male     Yes          2      Graduate            No            3200              700          70              360              1
12 LP001027   Male     Yes          2      Graduate                          2500             1840         109              360              1
13 LP001028   Male     Yes          2      Graduate            No            3073             8106         200              360              1
14 LP001029   Male      No          0      Graduate            No            1853             2840         114              360              1
15 LP001030   Male     Yes          2      Graduate            No            1299             1086          17              120              1
16 LP001032   Male      No          0      Graduate            No            4950                0         125              360              1
18 LP001036 Female      No          0      Graduate            No            3510                0          76              360              0
19 LP001038   Male     Yes          0  Not Graduate            No            4887                0         133              360              1
21 LP001043   Male     Yes          0  Not Graduate            No            7660                0         104              360              0
22 LP001046   Male     Yes          1      Graduate            No            5955             5625         315              360              1
23 LP001047   Male     Yes          0  Not Graduate            No            2600             1911         116              360              0
24 LP001050             Yes          2  Not Graduate            No            3365             1917         112              360              0
26 LP001066   Male     Yes          0      Graduate           Yes            9560                0         191              360              1
27 LP001068   Male     Yes          0      Graduate            No            2799             2253         122              360              1
28 LP001073   Male     Yes          2  Not Graduate            No            4226             1040         110              360              1
```

2) Data imputation

On the in-built "IRIS" dataset, perform data imputation using

i) MICE

CODE:

ii) iris
iii) data<- iris
iv) summary(iris)
v) iris.mis <- prodNA(iris, noNA = 0.1)
vi) summary(iris.mis)
vii) iris.mis <- subset(iris.mis, select = -c(Species))
viii) md.pattern(iris.mis)
ix) par("mar")
x) par(mar=c(1,1,1,1
xi)))
xii) imputed_Data <- mice(iris.mis, m=5, maxit = 50, method = 'pmm', seed = 500)
xiii) summary(imputed_Data)
xiv) mice(data = iris.mis, m = 5, method = "pmm", maxit = 50, seed =
xv) 500) imputed_DataimpSepal.Width
xvi) completeData <- complete(imputed_Data,2)
xvii) fit <- with(data = iris.mis, exp = lm(Sepal.Width ~ Sepal.Length +
xviii) Petal.Width)) combine <- pool(fit)
xix) summary(combine)

OUTPUT:

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
```

```
> data<- iris
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

```
> iris.mis <- prodNA(iris, noNA = 0.1)
> summary(iris.mis)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :49
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:45
 Median :5.800   Median :3.000   Median :4.400   Median :1.300   virginica :44
 Mean   :5.843   Mean   :3.056   Mean   :3.749   Mean   :1.153   NA's      :12
 3rd Qu.:6.400   3rd Qu.:3.400   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.700   Max.   :2.500
 NA's   :12      NA's   :18      NA's   :16      NA's   :17
> iris.mis <- subset(iris.mis, select = -c(Species))
> summary(iris.mis)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.400   Median :1.300
 Mean   :5.843   Mean   :3.056   Mean   :3.749   Mean   :1.153
 3rd Qu.:6.400   3rd Qu.:3.400   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.700   Max.   :2.500
 NA's   :12      NA's   :18      NA's   :16      NA's   :17
>
```

```
> md.pattern(iris.mis)
    Sepal.Length Petal.Length Petal.Width Sepal.Width
96             1            1           1           1  0
14             1            1           1           0  1
13             1            1           0           1  1
2              1            1           0           0  2
9              1            0           1           1  1
2              1            0           1           0  2
2              1            0           0           1  2
9              0            1           1           1  1
3              0            0           1           1  2
              12           16          17          18 63
> par("mar")
[1] 5.1 4.1 4.1 2.1
> par(mar=c(1,1,1,1))
```

```
> imputed_Data <- mice(iris.mis, m=5, maxit = 50, method = 'pmm', seed = 500)

 iter imp variable
  1   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   4  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   5  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   4  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   5  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   4  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   5  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  4   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  4   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  4   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
```

```
> summary(imputed_Data)
Class: mids
Number of multiple imputations:  5
Imputation methods:
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       "pmm"        "pmm"        "pmm"        "pmm"
PredictorMatrix:
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length            0           1            1           1
Sepal.Width             1           0            1           1
Petal.Length            1           1            0           1
Petal.Width             1           1            1           0
> mice(data = iris.mis, m = 5, method = "pmm", maxit = 50, seed = 500)

 iter imp variable
  1   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   4  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  1   5  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   3  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   4  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  2   5  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   1  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
  3   2  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
```

```
> imputed_Data$imp$Sepal.Width
      1   2   3   4   5
27  3.4 3.4 3.8 3.8 3.8
31  2.8 3.2 3.2 3.4 3.0
34  3.4 4.0 3.9 2.6 3.4
36  3.4 3.5 3.4 3.5 3.5
38  3.0 3.0 3.0 3.1 3.2
39  3.2 3.4 2.9 2.8 2.3
43  3.2 3.0 3.0 3.0 2.9
51  3.2 3.0 3.5 2.8 3.4
63  3.8 3.0 3.8 3.6 3.1
67  2.7 2.3 2.2 2.7 2.4
80  3.0 2.3 3.0 2.9 3.0
90  2.8 3.0 2.5 2.5 2.7
111 3.0 2.3 2.7 3.1 3.0
112 2.8 3.4 3.1 3.0 2.9
133 3.0 2.9 2.8 2.4 3.0
139 3.0 3.1 2.8 2.5 2.9
145 2.9 2.5 3.0 3.0 3.0
146 3.8 2.3 3.2 3.0 2.5
> completeData <- complete(imputed_Data,2)
>
```

ii) Amelia package

```
data("iris")
iris.mis <- prodNA(iris, noNA =
0.1) summary(iris.mis)
amelia_fit <- amelia(iris.mis, m=5, parallel = "multicore", noms
="Species") amelia_fit$imputations[[1]]
amelia_fit$imputations[[2]]
```

```
amelia_fit$imputations[[3]]
amelia_fit$imputations[[4]]
amelia_fit$imputations[[5]]
amelia_fit$imputations[[5]]$Sepal.Le
ngth
    write.amelia(amelia_fit, file.stem = "imputed_data_set")
```

output:

```
> data("iris")
> #seed 10% missing values
> iris.mis <- prodNA(iris, noNA = 0.1)
> summary(iris.mis)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :48
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.500   1st Qu.:0.300   versicolor:41
 Median :5.700   Median :3.000   Median :4.400   Median :1.300   virginica :49
 Mean   :5.828   Mean   :3.054   Mean   :3.826   Mean   :1.209   NA's      :12
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.200   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
 NA's   :13      NA's   :17      NA's   :17      NA's   :16
> #specify columns and run amelia
```

3)Data conversion

On the “loan” dataset, perform the data imputation and data conversion techniques.

```
dat <- read_csv("loan.csv")
glimpse(dat) summary(dat)
summary(dat$Gender)
sapply(dat, function(x) sum(is.na(x)))
dat$Gender[is.na(dat$Gender)] <- median(dat$Gender, na.rm = TRUE)
sapply(dat, function(x) sum(is.na(x))) table(dat$Gender)
dat$Married[is.na(dat$Married)] <- "Yes" table(dat$Married)
    skewness(dat$Dependents)
```

output:

```
> dat <- read_csv("loan.csv")

-- Column specification --------------------------------------------------------
cols(
  Loan_ID = col_character(),
  Gender = col_character(),
  Married = col_character(),
  Dependents = col_character(),
  Education = col_character(),
  Self_Employed = col_character(),
  ApplicantIncome = col_double(),
  CoapplicantIncome = col_double(),
  LoanAmount = col_double(),
  Loan_Amount_Term = col_double(),
  Credit_History = col_double(),
  Property_Area = col_character(),
  Loan_Status = col_character()
)

> |
```

```
> glimpse(dat)
Rows: 614
Columns: 13
$ Loan_ID           <chr> "LP001002", "LP001003", "LP001005", "LP001006", "LP001008", "LP001011", "LP001013", "LP001014", "LP001018", "LP001020", "~
$ Gender            <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "Male", "~
$ Married           <chr> "No", "Yes", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", "Yes", "No", "No", "No", "~
$ Dependents        <chr> "0", "1", "0", "0", "0", "2", "0", "3+", "2", "1", "2", "2", "2", "0", "2", "0", "1", "0", "0", "0", "0", "1", "0", "2", "~
$ Education         <chr> "Graduate", "Graduate", "Graduate", "Not Graduate", "Graduate", "Graduate", "Not Graduate", "Graduate", "Graduate", "Grad~
$ Self_Employed     <chr> "No", "No", "Yes", "No", "No", "Yes", "No", "No", "No", "No", "No", NA, "No", "No", "No", "No", "No", "No", "No", NA, "No~
$ ApplicantIncome   <dbl> 5849, 4583, 3000, 2583, 6000, 5417, 2333, 3036, 4006, 12841, 3200, 2500, 3073, 1853, 1299, 4950, 3596, 3510, 4887, 2600, ~
$ CoapplicantIncome <dbl> 0, 1508, 0, 2358, 0, 4196, 1516, 2504, 1526, 10968, 700, 1840, 8106, 2840, 1086, 0, 0, 0, 0, 0, 3500, 0, 5625, 1911, 1917, 2~
$ LoanAmount        <dbl> NA, 128, 66, 120, 141, 267, 95, 158, 168, 349, 70, 109, 200, 114, 17, 125, 100, 76, 133, 115, 104, 315, 116, 112, 151, 191~
$ Loan_Amount_Term  <dbl> 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 120, 360, 240, 360, 360, NA, 360, 360, 360, 360, 36~
$ Credit_History    <dbl> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, NA, 0, 1, 1, 0, 1, 0, 0, NA, 1, 1, 1, 1, 1, NA, 1, 1, 1, 1, 1, 0, 1, 1, 1~
$ Property_Area     <chr> "Urban", "Rural", "Urban", "Urban", "Urban", "Urban", "Urban", "Semiurban", "Urban", "Semiurban", "Urban", "Urban", "Urba~
$ Loan_Status       <chr> "Y", "N", "Y", "Y", "Y", "Y", "Y", "N", "Y", "N", "Y", "Y", "Y", "N", "Y", "Y", "Y", "N", "N", "Y", "N", "Y", "N", "N", "~
> summary(dat)
   Loan_ID            Gender            Married          Dependents          Education         Self_Employed      ApplicantIncome coapplicantIncome
 Length:614         Length:614         Length:614         Length:614         Length:614         Length:614         Min.   :  150   Min.   :    0
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 2878   1st Qu.:    0
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 3812   Median : 1188
                                                                                                                  Mean   : 5403   Mean   : 1621
                                                                                                                  3rd Qu.: 5795   3rd Qu.: 2297
                                                                                                                  Max.   :81000   Max.   :41667

   LoanAmount     Loan_Amount_Term Credit_History   Property_Area      Loan_Status
 Min.   :  9.0   Min.   : 12      Min.   :0.0000   Length:614         Length:614
 1st Qu.:100.0   1st Qu.:360      1st Qu.:1.0000   Class :character   Class :character
 Median :128.0   Median :360      Median :1.0000   Mode  :character   Mode  :character
 Mean   :146.4   Mean   :342      Mean   :0.8422
 3rd Qu.:168.0   3rd Qu.:360      3rd Qu.:1.0000
 Max.   :700.0   Max.   :480      Max.   :1.0000
 NA's   :22      NA's   :14       NA's   :50
>
```

```
> skewness(dat$Dependents)
[1] NA
> |
```