Tyler Burleson
CSCI-4047-901
Exercise 4

1)

a.

ID	gender	age	income	tax (15%)
1	Women	21	147168	22075.2
2	Female	29	119595	17939.25
3	Female	56	87770	13165.5
4	NA	21	54259	8138.85
5	Male	28	NA	160230
6	Woman	0	128326	19248.9
7	Female	-1	NA	NA
8	Female	24	NA	11820.6
9	Female	38	149473	22420.95
10	Man	48	113663	1136630
11	Woman	33	96649	14497.35
12	Male	55	NA	9776.25
13	Man	52	64944	9741.6
14	Male	47	85300	NA
15	Men	45	80091	12013.65
16	Women	25	56418	8462.7
17	Female	45	54189	8128.35
18	Male	29	67489	10123.35
19	Male	48	137978	20696.7
20	Female	-1	89991	NA
21	Men	55	103803	15570.45
22	Female	20	59504	8925.6
23	Men	-1	-727	22500
24	Female	45	111069	16660.35
25	Male	52	239	10409.1
26	Male	26	116941	17541.15
27	Female	22	63018	252072
28	NA	21	76019	11402.85
29	Women	43	NA	16849.95
30	Woman	32	52020	7803
31	Male	59	75292	11293.8
32	NA	23	148850	22327.5
33	Male	-1	NA	19040.25
34	Male	60	148765	NA

b. **733 ~ 73%**

```
# Check what data is "complete" and remove NA records
data_complete <- na.omit(data)
# 733 records / 1000 records ~73%
data_complete
```

Data	
corule	List of 2
data	1000 obs. of 5 variables
data_complete	733 obs. of 5 variables
newdata	1000 obs. of 5 variables

2) **42.4%**

Tyler Burleson
CSCI-4047-901
Exercise 4

3)

```
# create correction rules
cr <- correctionRules(expression(if(!is.na(age) & age <18) age <- NA,
                                 if(!is.na(income) & income <1) income <- NA,
                                 if(is.na(income) & !is.na(tax..15..)) income <- (tax..15.. / 0.15),
                                 if(!is.na(tax..15..) & (is.na(income) | tax..15.. != (income * 0.15))) tax..15.. <- NA,
                                 if(!is.na(gender) & gender =="Man") gender <- "Male",
                                 if(!is.na(gender) & gender =="Men") gender <- "Male",
                                 if(!is.na(gender) & gender =="Woman") gender <- "Female",
                                 if(!is.na(gender) & gender =="Women") gender <- "Female",
                                 if(is.na(tax..15..) & !is.na(income)) tax..15.. <- (income * 0.15)))

# apply rules
corule <- correctWithRules(cr,data)
newdata <- corule$corrected
newdata
```

This replaces gender attributes to be either male or female. This also replaces incorrect Tax and Income with the correct information if possible, otherwise they are set to NA

4)

```
> print(gender_counts)

Female   Male   <NA>
   455    457     88
> summary(newdata)
       ID            gender              age            income          tax..15..
 Min.   :   1.0   Length:1000        Min.   :18.00   Min.   : -47460   Min.   : -7119
 1st Qu.: 250.8   Class :character   1st Qu.:28.00   1st Qu.:  70420   1st Qu.: 10563
 Median : 500.5   Mode  :character   Median :40.00   Median :  97496   Median : 14624
 Mean   : 500.5                      Mean   :38.96   Mean   : 111360   Mean   : 16704
 3rd Qu.: 750.2                      3rd Qu.:50.00   3rd Qu.: 124448   3rd Qu.: 18667
 Max.   :1000.0                      Max.   :60.00   Max.   :5200320   Max.   :780048
                                     NA's   :98      NA's   :10        NA's   :10
```

Before Summary

```
> sum(is.na(data_imputation))
[1] 0
> summary(data_imputation)
       ID            gender              age            income          tax..15..
 Min.   :   1.0   Length:1000        Min.   :18.00   Min.   : -47460   Min.   : -7119
 1st Qu.: 250.8   Class :character   1st Qu.:28.00   1st Qu.:  70432   1st Qu.: 10565
 Median : 500.5   Mode  :character   Median :39.00   Median :  97888   Median : 14639
 Mean   : 500.5                      Mean   :38.81   Mean   : 111384   Mean   : 16692
 3rd Qu.: 750.2                      3rd Qu.:49.00   3rd Qu.: 124797   3rd Qu.: 18678
 Max.   :1000.0                      Max.   :60.00   Max.   :5200320   Max.   :780048
   ID_imp         gender_imp       age_imp         income_imp      tax..15.._imp
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1000      FALSE:912       FALSE:902       FALSE:990       FALSE:990
                 TRUE :88        TRUE :98        TRUE :10        TRUE :10
```

After Summary