



Palestine Technical University – Kadoorie
College of Engineering and Technology
Department of Computer Systems Engineering

ADULT DATASET

By:

Manar Fuqha – 202010224

(Mohammad Ameen) Abohasan – 201913064

Supervisor:

Dr. Anas Melhem

Project submitted in the Data Mining course of the requirements
for the bachelor's degree in computer systems engineering.

Tulkarm, Palestine

Aug, 2023

Part 1

Objective

Applying data analysis, cleaning, and finding associations

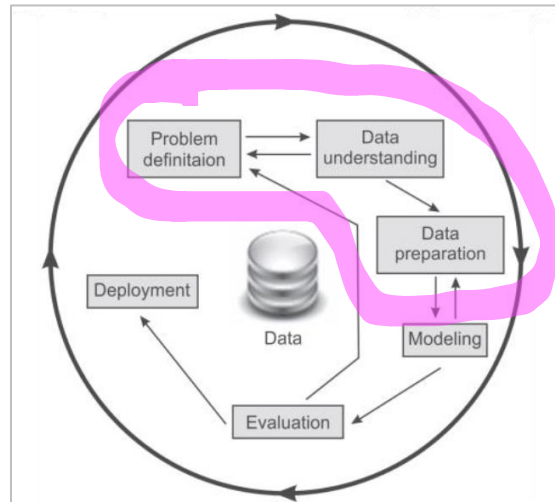






Figure 1 Data mining process

 UC Irvine
Machine Learning
Repository

Search datasets... 

Login 



Adult

Donated on 4/30/1996

Predict whether income exceeds \$50K/yr based on census data. Also known as "Census Income" dataset.

DOWNLOAD

CITE

257 citations

173732 views

<https://archive.ics.uci.edu/dataset/2/adult>

Prediction task is to determine whether a person makes **over 50k** a year.

Has **Missing Values**? Yes.

Listing of attributes:

- 1) **age - numerical**
continuous.
- 2) **workClass - nominal**
Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- 3) **finalWeight - numerical**
continuous.
- 4) **education – nominal**
Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- 5) **educationNum - numerical**
continuous.
- 6) **maritalStatus - nominal**
Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- 7) **occupation - nominal**
Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- 8) **relationship - nominal**
Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- 9) **race - nominal**
White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- 10) **sex - nominal**
Female, Male.
- 11) **capitalGain - numerical**
continuous.
- 12) **capitalLoss - numerical**
continuous.
- 13) **hoursPerWeek - numerical**
continuous.
- 14) **nativeCountry - nominal**
United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.

income - nominal (target variable), >50K, <=50k.

```
> myAdult <- read_delim (
  'C:/Users/moham/Documents/adult.data',
  delim = ",",
  col_names = c (
    "age",
    "workClass",
    "finalWeight",
    "education",
    "educationNum",
    "maritalStatus",
    "occupation",
    "relationship",
    "race",
    "sex",
    "capitalGain",
    "capitalLoss",
    "hoursPerWeek",
    "nativeCountry",
    "income"
  )
)
```

```
R 4.3.1 ~ /
> myAdult <- read_delim (
+   'C:/Users/moham/Documents/adult.data',
+   delim = ",",
+   col_names = c (
+     "age",
+     "workClass",
+     "finalWeight",
+     "education",
+     "educationNum",
+     "maritalStatus",
+     "occupation",
+     "relationship",
+     "race",
+     "sex",
+     "capitalGain",
+     "capitalLoss",
+     "hoursPerWeek",
+     "nativeCountry",
+     "income"
+   )
+ )
Rows: 32561 Columns: 15
— Column specification —
Delimiter: ",",
chr (9): workClass, education, maritalStatus, occupation, relationship, race, sex, nativeCountry, income
dbl (6): age, finalWeight, educationNum, capitalGain, capitalLoss, hoursPerWeek
```

Total number of rows:

```
> nrow(myAdult)
```

```
> nrow(myAdult)
[1] 32561
```

Total number of columns:

```
> ncol(myAdult)
```

```
> ncol(myAdult)
[1] 15
```

	age	workClass	finalWeight	education	educationNum	maritalStatus	occupation	relationship	race	sex	capitalGain	capitalLoss	hoursPerWeek	nativeCountry	income
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
2	50	Self-emp-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	>50K
3	38	Private	215846	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	>50K
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	>50K
5	28	Private	359429	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=20K
6	37	Private	284652	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	>50K
7	49	Private	161087	8th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	>50K
8	52	Self-emp-inc	238642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
9	31	Private	40781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14264	0	50	United-States	>50K
10	42	Private	158449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	\$178	0	40	United-States	>50K
11	37	Private	283464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
12	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	Honolulu	>50K
13	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	>=50K
14	32	Private	200219	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	>=50K
15	40	Private	121722	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	U.S.	>50K
16	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Alaska	Male	0	0	45	Mexico	<=50K
17	25	Self-emp-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
19	38	Private	28807	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
20	43	Self-emp-inc	262175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	40	United-States	<=50K
21	40	Private	189324	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
22	34	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
23	35	Federal-gov	76845	8th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
24	43	Private	117037	11th	7	White	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
25	59	Private	108015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
26	56	Local-gov	216851	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
27	16	Private	188784	<1st-grade	6	Never-married	Craft-repair	Own-child	White	Male	0	0	40	United-States	<=50K

Representation of the missing value as NA instead of "?":

```
> myAdult[myAdult == "?"] <- NA
```

```
> unique(myAdult$age)
```

```
[1] 39 50 38 53 28 37 49 52 31 42 30 23 32 40 34 25 43 54 35 59 56 19 20 45 22 48 21 24 57 44 41 29 18 47 46 36 79 27 67 33 76 17 55 61 70 64 71
[48] 68 66 51 58 26 60 90 75 65 77 62 63 80 72 74 69 73 81 78 88 82 83 84 85 86 87
```

```
> unique(myAdult$workClass)
```

```
[1] "State-gov"      "Self-emp-not-inc" "Private"      "Federal-gov"    "Local-gov"      NA              "Self-emp-inc"
[8] "without-pay"    "Never-worked"
```

```
> unique(myAdult$finalWeight)
```

```
[1] 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 280464 141297 122272 205019 121772 245487 176756
[18] 186824 28887 292175 193524 302146 76845 117037 109015 216851 168294 180211 367260 193366 190709 266015 386940 59951
[35] 311512 242406 197200 544091 84154 265477 507875 88506 172987 94638 289980 337895 144361 128354 101603 271466 32275
[52] 226956 51835 251585 109832 237993 216666 56352 147372 188146 59496 293936 149640 116632 105598 155537 183175 169846
[69] 191681 200681 101509 309974 162298 211678 124744 213921 32214 212759 309634 125927 446839 276515 51618 159937 343591
[86] 346253 268234 202051 54334 410867 249977 286730 212563 117747 226296 115585 191277 202683 171095 249409 124191 198282
[103] 149116 188300 103432 317660 304873 194901 189265 124692 432376 65324 335605 377869 102864 95647 303090 197371 247552
[120] 102632 199915 118853 77143 267989 301606 287828 111697 114937 129305 365739 69621 43323 120985 254202 146195 125933
[137] 56920 163127 34310 81973 66614 232782 316868 196584 105376 185814 175374 108293 181232 174662 186009 198183 163003
[154] 296158 252903 187715 214542 494223 191535 228456 38317 252752 78374 88419 201080 207157 235485 102628 25828 54826
[171] 124953 175325 96062 428030 149624 253814 312956 483777 183930 37274 181344 114580 633742 286370 29054 304030 143129
[188] 135105 99928 109567 155222 159567 523910 120939 130760 197387 99374 56795 138992 32921 397317 170653 259323 254817
[205] 48211 140164 128757 36270 210563 65368 160943 208358 153790 58515 125417 635913 313321 182609 109434 255004 197860
[222] 187656 51744 176681 140359 243313 24215 167687 314209 176796 538583 130408 159732 110978 76714 268700 170525 180138
[239] 115076 115458 347890 196001 273905 119156 179488 203580 236596 183916 207578 153141 112763 390781 171328 27382 259014
[256] 303044 117789 172579 187666 204518 150042 88092 245918 146013 378322 257295 218956 21174 185480 222205 69867 191260
[273] 30653 209109 70377 477983 170924 190174 193787 279472 34918 97688 175413 173960 205759 425161 220531 176609 371987
[290] 193884 200352 127595 220419 231931 248402 111095 57424 157443 278130 169469 146268 153718 217460 238638 303296 173321
[307] 193945 83082 193815 34987 59306 142897 860348 205607 199698 191954 138714 399087 423158 159841 174308 50356 186110
[324] 200381 174309 78383 211601 187728 321171 127921 206565 224563 132178 178686 98545 242606 270942 94235 71195 104112 261192
[341] 94936 296478 119272 85043 293364 241895 36135 151989 101128 156464 117963 192262 111363 329752 372020 95432 161400
[358] 96129 111949 117125 348022 270092 180609 174575 410439 92262 183081 362589 212448 481060 185885 89821 184018 256649
[375] 160323 350845 267404 35633 80914 172927 174319 214955 344991 108699 117312 396099 134152 162028 25429 232392 220098
[392] 301302 277946 98101 196164 115562 96975 137300 86872 132178 416103 108574 288353 227689 166481 445382 110145 317253
[409] 123147 364657 42346 241951 118500 188386 1033222 92440 190762 426017 243867 240283 61777 175024 92003 188401 228528
[426] 133373 255191 204653 222289 287480 107762 202521 204116 29662 116358 208405 284843 117018 81281 340148 363425 45857
[443] 191073 405855 298227 290521 56915 146538 258872 206399 197332 245062 197583 234885 72887 180374 351299 54012 115745
[460] 288825 132601 193374 170070 126708 35598 33983 192776 118551 201965 139883 285020 303990 49401 279196 211870 281432
[477] 161155 197904 111746 170721 70100 193626 271749 189775 401531 286967 164427 91039 347934 371373 32220 187251 178107
[494] 343121 262749 403107 64293 303588 324960 114060 48925 180980 181054 388093 249609 112131 543162 91996 141944 251804
[511] 37070 337587 189346 222216 267044 214635 204226 108116 99146 196232 248344 186035 177905 85812 221172 99183 190387
[528] 202692 109339 108658 197202 101739 231559 207853 190942 102345 41493 190027 210525 133937 237903 163862 201872 84179
[545] 51662 233327 259510 184831 245724 27053 205343 229328 319560 136218 54576 323069 148291 152453 114053 212960 264052
[562] 82804 334273 27337 188436 433665 110663 87490 354351 95469 242718 22463 158156 350162 165532 28738 283635 86646
[579] 195733 69884 199713 181659 340939 197747 34292 156764 25826 103948 137390 105138 39352 168387 267147 99399 214242
[596] 200408 136455 239824 217039 51290 175674 194404 45612 410114 182521 339772 169658 200853 247564 249909 208122 109881
[613] 207824 369027 114117 51048 102388 190483 462440 109351 34383 241832 124187 153614 267556 205469 268090 165039 120451
[630] 154374 103649 35723 262601 226181 175697 248145 289436 75654 199378 160968 188563 55849 195322 402089 78277 158611
[647] 169496 130959 556660 292472 143774 288341 71592 167358 106742 219288 174524 335183 261293 111900 194360 81145 341204
[664] 249362 247019 114746 172146 110457 80077 368700 182556 219420 240817 102726 226267 125457 204021 161141 190290 430828
[681] 59342 136721 149422 86644 195124 167350 113000 140027 262425 316702 335453 202480 203628 118710 189620 475028 110866
[698] 243605 163870 80145 295566 63042 229148 242552 177665 208103 296450 70282 271767 144995 382635 295697 194141 378418
```

```
[715] 214399 125831 271328 50459 162140 177937 111502 299047 223212 118474 352139 173093 181655 332702 51164 234901 131414
[732] 260960 156052 279914 192453 200939 151408 112847 316929 126319 197422 267736 267034 193047 356089 223515 87510 145111
[749] 48093 31757 285854 120064 167381 103408 101460 420537 119411 128272 386773 283268 301526 151790 106252 188557 171114
[766] 327323 244147 280282 116442 282579 51838 73585 226902 279129 146908 196690 49572 237601 169628 36671 231193 192130
[783] 149704 102102 32185 196061 211046 31577 162343 128831 316688 90758 274363 154538 106085 315859 51471 193830 231043
[800] 23780 169879 270333 138768 191571 219941 94113 137510 32607 93208 254440 186556 169871 167159 171871 154411 129227
[817] 110331 34269 174355 680390 233130 165474 257780 194259 280093 177387 28929 105304 499233 180572 321435 86108 198124
[834] 135162 146813 291175 387569 102895 33274 86551 138192 118966 99784 90980 177407 96467 327886 111567 166545 142182
[851] 188798 38563 216284 191547 285335 142712 80945 309055 62339 176186 266855 48087 121313 143437 160724 282753 194636
[868] 153044 411797 117683 376540 72393 270335 96226 95336 258498 149698 205865 155781 406468 177119 144397 372525 164170
[885] 183800 177307 170108 341995 226508 87418 109165 28856 175897 99697 90270 152375 171550 211154 202570 168496 68898
[902] 93235 278924 311020 175878 543028 202027 158926 76860 136063 186648 257509 98155 274198 97083 29825 262153 214738
[919] 138022 91842 373662 162003 52114 241843 375871 186934 176900 21906 132222 143653 78602 465507 196373 293227 241752
[936] 166398 184682 250802 325159 174675 227065 269080 177722 133461 239683 398473 298785 123424 176286 150062 169240 288273
[953] 526968 57066 323573 368825 189721 164966 94954 202046 161538 105252 200153 178326 255957 188693 182977 159929 123207
[970] 284317 184699 154474 318280 254907 349221 335973 126701 122159 187370 124793 192835 290226 112840 89325 33109 82465
[987] 329980 148294 168212 343642 115244 162572 356067 271567 180804 123011 109186 220537 124827 767403
```

[reached getoption("max.print") -- omitted 20648 entries]

> unique(myAdult\$education)

```
> unique(myAdult$education)
[1] "Bachelors" "HS-grad" "11th" "Masters" "9th" "Some-college" "Assoc-acdm" "Assoc-voc" "7th-8th"
[10] "Doctorate" "Prof-school" "5th-6th" "10th" "1st-4th" "Preschool" "12th"
```

> unique(myAdult\$educationNum)

```
> unique(myAdult$educationNum)
[1] 13 9 7 14 5 10 12 11 4 16 15 3 6 2 1 8
```

> unique(myAdult\$maritalStatus)

```
> unique(myAdult$maritalStatus)
[1] "Never-married" "Married-civ-spouse" "Divorced" "Married-spouse-absent" "Separated"
[6] "Married-AF-spouse" "Widowed"
```

> unique(myAdult\$occupation)

```
> unique(myAdult$occupation)
[1] "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Prof-specialty" "Other-service" "Sales" "Craft-repair"
[8] "Transport-moving" "Farming-fishing" "Machine-op-inspct" "Tech-support" NA "Protective-serv" "Armed-Forces"
[15] "Priv-house-serv"
```

> unique(myAdult\$relationship)

```
> unique(myAdult$relationship)
[1] "Not-in-family" "Husband" "Wife" "Own-child" "Unmarried" "Other-relative"
```

> unique(myAdult\$race)

```
> unique(myAdult$race)
[1] "White" "Black" "Asian-Pac-Islander" "Amer-Indian-Eskimo" "Other"
```

> unique(myAdult\$sex)

```
> unique(myAdult$sex)
[1] "Male" "Female"
```

> unique(myAdult\$capitalGain)

```
> unique(myAdult$capitalGain)
[1] 2174 0 14084 5178 5013 2407 14344 15024 7688 34095 4064 4386 7298 1409 3674 1055 3464 2050 2176 594 20051 6849 4101
[24] 1111 8614 3411 2597 25236 4650 9386 2463 3103 10605 2964 3325 2580 3471 4865 99999 6514 1471 2329 2105 2885 25124 10520
[47] 2202 2961 27828 6767 2228 1506 13550 2635 5556 4787 3781 3137 3818 3942 914 401 2829 2977 4934 2062 2354 5455 15020
[70] 1424 3273 22040 4416 3908 10566 991 4931 1086 7430 6497 114 7896 2346 3418 3432 2907 1151 2414 2290 15831 41310 4508
[93] 2538 3456 6418 1848 3887 5721 9562 1455 2036 1831 11678 2936 2993 7443 6360 1797 1173 4687 6723 2009 6097 2653 1639
[116] 18481 7978 2387 5060
```

> unique(myAdult\$capitalLoss)

```
> unique(myAdult$capitalLoss)
[1] 0 2042 1408 1902 1573 1887 1719 1762 1564 2179 1816 1980 1977 1876 1340 2206 1741 1485 2339 2415 1380 1721 2051 2377 1669 2352 1672 653
[29] 2392 1504 2001 1590 1651 1628 1848 1740 2002 1579 2258 1602 419 2547 2174 2205 1726 2444 1138 2238 625 213 1539 880 1668 1092 1594 3004
[57] 2231 1844 810 2824 2559 2057 1974 974 2149 1825 1735 1258 2129 2603 2282 323 4356 2246 1617 1648 2489 3770 1755 3683 2267 2080 2457 155
[85] 3900 2201 1944 2467 2163 2754 2472 1411
```

> unique(myAdult\$hoursPerWeek)

```
> unique(myAdult$hoursPerWeek)
[1] 40 13 16 45 50 80 30 35 60 20 52 44 15 25 38 43 55 48 58 32 70 2 22 56 41 28 36 24 46 42 12 65 1 10 34 75 98 33 54 8 6 64 19 18 72 5 9
[48] 47 37 21 26 14 4 59 7 99 53 39 62 57 78 90 66 11 49 84 3 17 68 27 85 31 51 77 63 23 87 88 73 89 97 94 29 96 67 82 86 91 81 76 92 61 74 95
```

> unique(myAdult\$nativeCountry)

```
> unique(myAdult$nativeCountry)
[1] "United-States" "Cuba" "Jamaica" "India"
[5] NA "Mexico" "South" "Puerto-Rico"
[9] "Honduras" "England" "Canada" "Germany"
[13] "Iran" "Philippines" "Italy" "Poland"
[17] "Columbia" "Cambodia" "Thailand" "Ecuador"
[21] "Laos" "Taiwan" "Haiti" "Portugal"
[25] "Dominican-Republic" "El-Salvador" "France" "Guatemala"
[29] "China" "Japan" "Yugoslavia" "Peru"
[33] "Outlying-US(Guam-USVI-etc)" "Scotland" "Trinidad&Tobago" "Greece"
[37] "Nicaragua" "Vietnam" "Hong" "Ireland"
[41] "Hungary" "Holand-Netherlands"
```

> unique(myAdult\$income)

```
> unique(myAdult$income)
[1] "<=50K" ">50K"
```

Structure of the dataset and the data types of its columns:

> str(myAdult)

```
> str(myAdult)
'spc_tbl_ [32,561 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ age      : num [1:32561] 39 50 38 53 28 37 49 52 31 42 ...
 $ workClass : chr [1:32561] "State-gov" "Self-emp-not-inc" "Private" "Private" ...
 $ finalWeight : num [1:32561] 77516 83311 215646 234721 338409 ...
 $ education : chr [1:32561] "Bachelors" "Bachelors" "HS-grad" "11th" ...
 $ educationNum : num [1:32561] 13 13 9 7 13 14 5 9 14 13 ...
 $ maritalStatus : chr [1:32561] "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
 $ occupation : chr [1:32561] "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Handlers-cleaners" ...
 $ relationship : chr [1:32561] "Not-in-family" "Husband" "Not-in-family" "Husband" ...
 $ race        : chr [1:32561] "white" "white" "white" "Black" ...
 $ sex         : chr [1:32561] "Male" "Male" "Male" "Male" ...
 $ capitalGain : num [1:32561] 2174 0 0 0 0 ...
 $ capitalLoss : num [1:32561] 0 0 0 0 0 0 0 0 0 ...
 $ hoursPerWeek : num [1:32561] 40 13 40 40 40 40 16 45 50 40 ...
 $ nativeCountry : chr [1:32561] "United-States" "United-States" "United-States" "United-States" ...
 $ income       : chr [1:32561] "<=50K" "<=50K" "<=50K" "<=50K" ...
 - attr(*, "spec")=
 .. cols(
 ..   age = col_double(),
 ..   workClass = col_character(),
 ..   finalWeight = col_double(),
 ..   education = col_character(),
 ..   educationNum = col_double(),
 ..   maritalStatus = col_character(),
 ..   occupation = col_character(),
 ..   relationship = col_character(),
 ..   race = col_character(),
 ..   sex = col_character(),
 ..   capitalGain = col_double(),
 ..   capitalLoss = col_double(),
 ..   hoursPerWeek = col_double(),
 ..   nativeCountry = col_character(),
 ..   income = col_character()
 .. )
 - attr(*, "problems")=<externalptr>
```

Total number of missing values:

```
> sum(is.na(myAdult))
```

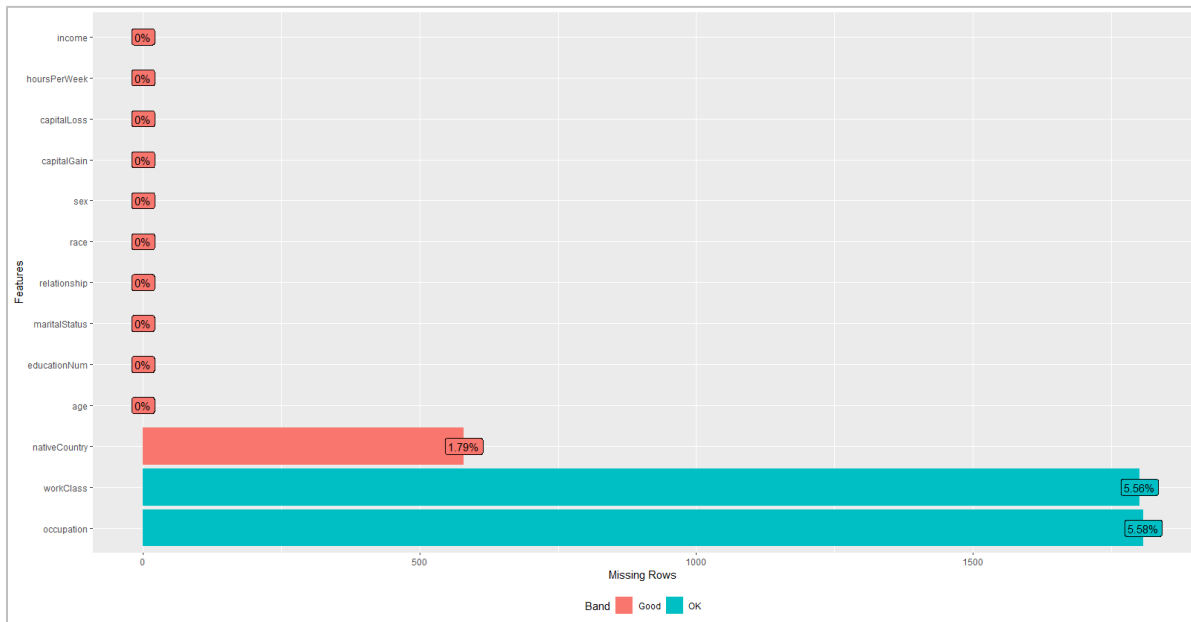
Total number of incomplete cases (rows with at least one missing value):

```
> sum(!complete.cases(myAdult))
```

```
> myAdult[myAdult == "?"] <- NA
> sum(is.na(myAdult))
[1] 4262
> sum(!complete.cases(myAdult))
[1] 2399
```

```
> library("DataExplorer")
```

```
> plot_missing(myAdult)
```



We have three attributes with missing data. no row in the dataset contains more than 25% missing data.

```
> sum(apply(X = is.na(myAdult), MARGIN = 1, FUN = mean) > 0.25)
```

```
> sum(apply(X = is.na(myAdult), MARGIN = 1, FUN = mean) > 0.25)
[1] 0
```

Then, no rows are deleted from the dataset.

Ignore **unnecessary columns** those that have no impact on the income attribute:

1- Ignoring the education attribute because the **educationNum** attribute already represents it as numbers and it is not necessary.

2- Ignoring the **finalWeight** attribute because it is unnecessary and has no impact on the analysis


```
> myAdult <- subset(myAdult, select = c("age", "workClass", "educationNum", "maritalStatus", "occupation", "relationship", "race", "sex", "capitalGain", "capitalLoss", "hoursPerWeek", "nativeCountry", "income"))
```

age	workClass	educationNum	maritalStatus	occupation	relationship	race	sex	capitalGain	capitalLoss	hoursPerWeek	nativeCountry	income	
1	39	State-gov	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
2	50	Self-emp-not-inc	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
3	38	Private	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
4	53	Private	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
5	28	Private	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
6	37	Private	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
7	49	Private	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
8	52	Self-emp-not-inc	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
9	31	Private	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
10	42	Private	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
11	37	Private	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
12	30	State-gov	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
13	23	Private	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
14	32	Private	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
15	40	Private	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	NA	>50K

Showing 1 to 15 of 32,561 entries, 13 total columns

Console

Terminal

Background Jobs

```

R 4.3.1 - ~/
> str(myAdult)
tibble [32,561 x 13] (S3: tbl_df/tbl/data.frame)
 $ age      : num [1:32561] 39 50 38 53 28 37 49 52 31 42 ...
 $ workClass : chr [1:32561] "State-gov" "Self-emp-not-inc" "Private" "Private" ...
 $ educationNum : num [1:32561] 13 13 9 7 13 14 5 9 14 13 ...
 $ maritalStatus: chr [1:32561] "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
 $ occupation  : chr [1:32561] "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Handlers-cleaners" ...
 $ relationship: chr [1:32561] "Not-in-family" "Husband" "Not-in-family" "Husband" ...
 $ race        : chr [1:32561] "White" "White" "White" "Black" ...
 $ sex         : chr [1:32561] "Male" "Male" "Male" "Male" ...
 $ capitalGain : num [1:32561] 2174 0 0 0 0 ...
 $ capitalLoss : num [1:32561] 0 0 0 0 0 0 0 0 ...
 $ hoursPerWeek: num [1:32561] 40 13 40 40 40 40 16 45 50 40 ...
 $ nativeCountry: chr [1:32561] "United-States" "United-States" "United-States" "United-States" ...
 $ income      : chr [1:32561] "<=50K" "<=50K" "<=50K" "<=50K" ...
>

```

```
> install.packages("ggplot2")
> library("ggplot2")
```

```
> install.packages("ggplot2")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/moham/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
also installing the dependencies 'colorspace', 'farver', 'labeling', 'munsell', 'RColorBrewer', 'viridisLite', 'gtable', 'isoband', 'scales'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/colorspace_2.1-0.zip'
Content type 'application/zip' length 2633989 bytes (2.5 MB)
downloaded 2.5 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/farver_2.1.1.zip'
Content type 'application/zip' length 1505042 bytes (1.4 MB)
downloaded 1.4 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/labeling_0.4.2.zip'
Content type 'application/zip' length 62592 bytes (61 KB)
downloaded 61 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/munsell_0.5.0.zip'
Content type 'application/zip' length 244941 bytes (239 KB)
downloaded 239 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/RColorBrewer_1.1-3.zip'
Content type 'application/zip' length 55876 bytes (54 KB)
downloaded 54 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/viridisLite_0.4.2.zip'
Content type 'application/zip' length 1300095 bytes (1.2 MB)
downloaded 1.2 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/gtable_0.3.3.zip'
Content type 'application/zip' length 225470 bytes (220 KB)
downloaded 220 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/isoband_0.2.7.zip'
Content type 'application/zip' length 1968240 bytes (1.9 MB)
downloaded 1.9 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/scales_1.2.1.zip'
Content type 'application/zip' length 614394 bytes (599 KB)
downloaded 599 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/ggplot2_3.4.3.zip'
Content type 'application/zip' length 3329305 bytes (3.2 MB)
downloaded 3.2 MB

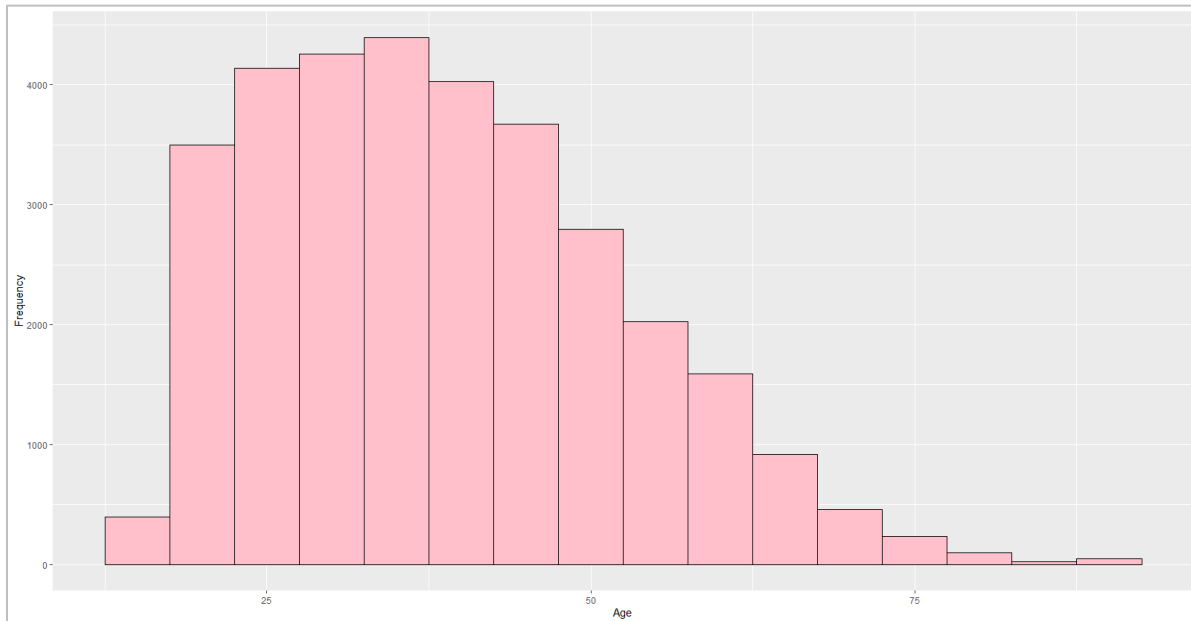
package 'colorspace' successfully unpacked and MD5 sums checked
package 'farver' successfully unpacked and MD5 sums checked
package 'labeling' successfully unpacked and MD5 sums checked
package 'munsell' successfully unpacked and MD5 sums checked
package 'RColorBrewer' successfully unpacked and MD5 sums checked
package 'viridisLite' successfully unpacked and MD5 sums checked
package 'gtable' successfully unpacked and MD5 sums checked
package 'isoband' successfully unpacked and MD5 sums checked
package 'scales' successfully unpacked and MD5 sums checked
package 'ggplot2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:/Users/moham/AppData/Local/Temp/RtmpmytbMq/downloaded_packages
> library("ggplot2")
```

Removing outliers based on a threshold of **5%**.

Age

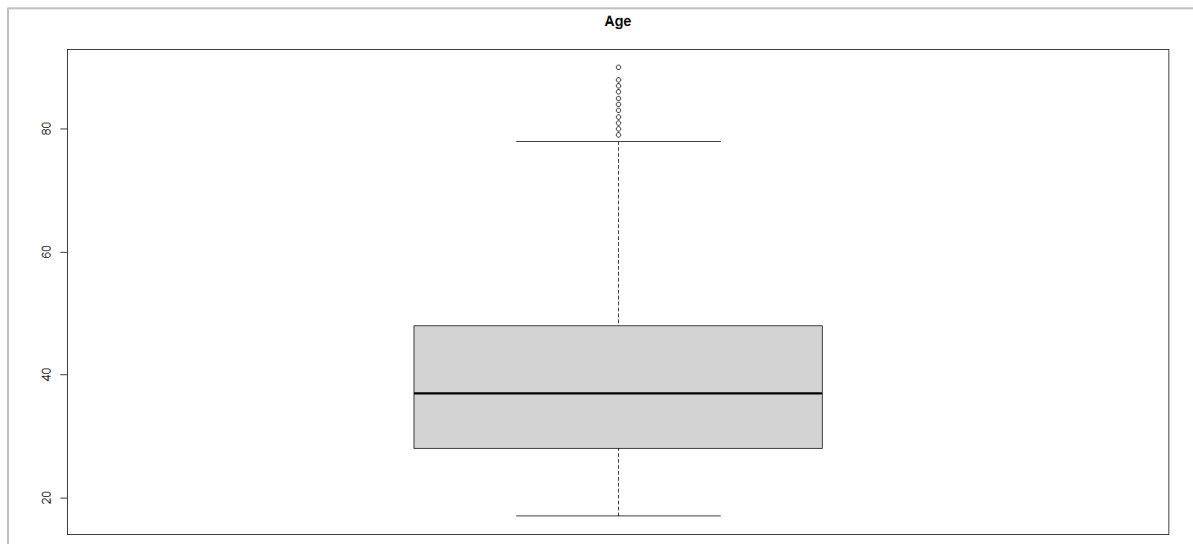
```
> ggplot(data = myAdult, aes(x = age)) + geom_histogram(binwidth = 5,  
fill = "pink", color = "black") + labs(x = "Age", y = "Frequency")
```



```
> qAge <- quantile(myAdult$age, probs = c(0.25,0.75))  
> iqrAge <- IQR(myAdult$age)  
> lowerAge <- qAge[1] - (1.5 * iqrAge)  
> upperAge <- qAge[2] + (1.5 * iqrAge)  
> ageOutliersPercentage <- length(myAdult$age[myAdult$age < lowerAge |  
myAdult$age > upperAge]) / length(myAdult$age) * 100  
> ageOutliersPercentage
```

```
> qAge <- quantile(myAdult$age, probs = c(0.25,0.75))  
> iqrAge <- IQR(myAdult$age)  
> lowerAge <- qAge[1] - (1.5 * iqrAge)  
> upperAge <- qAge[2] + (1.5 * iqrAge)  
> ageOutliersPercentage <- length(myAdult$age[myAdult$age < lowerAge | myAdult$age > upperAge]) / length(myAdult$age) * 100  
> ageOutliersPercentage  
[1] 0.4391757
```

```
> boxplot(myAdult$age, main = "Age")
```



```
> sum(is.na(myAdult$age)) / length(myAdult$age) * 100
```

```
> sum(is.na(myAdult$age)) / length(myAdult$age) * 100  
[1] 0
```

No missing value.

Because the age outlier percentage is less than 5%, we are going to remove outliers:

```
> myAdult <- subset(myAdult, age >= max(summary(myAdult$age)[1], lowerAge) &  
  age <= min(summary(myAdult$age)[6], upperAge))
```

```
> boxplot(myAdult$age, main = "Age after removing outliers")
```

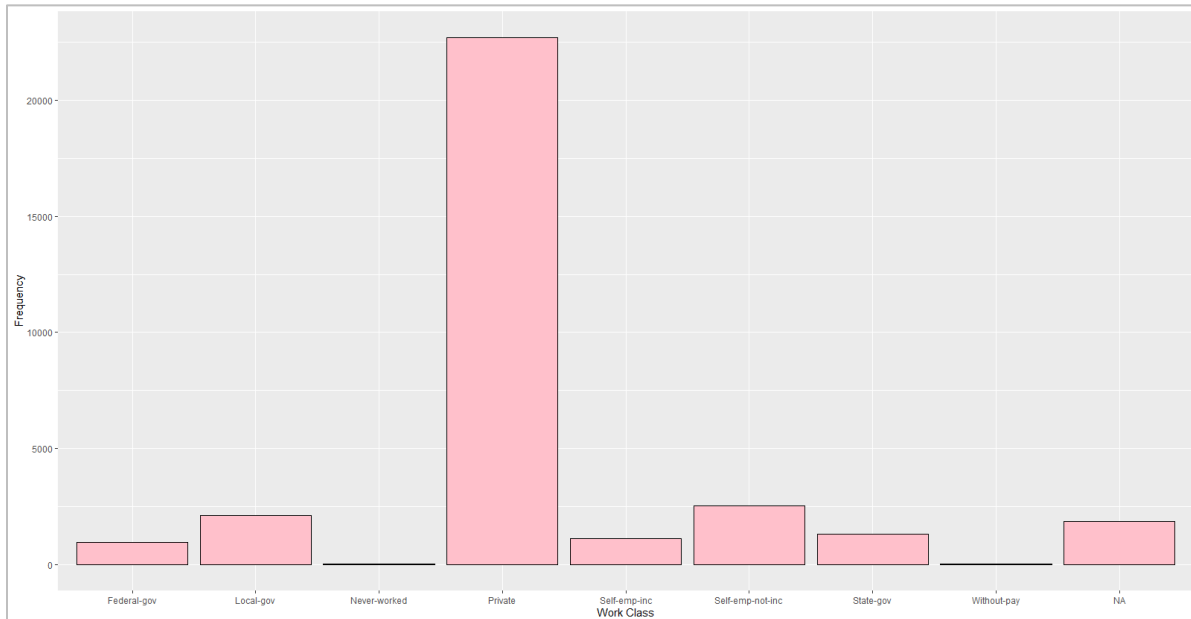


```
> ageOutliersPercentage <- length(myAdult$age[myAdult$age < lowerAge |  
myAdult$age > upperAge]) / length(myAdult$age) * 100  
> ageOutliersPercentage
```

```
> ageOutliersPercentage <- length(myAdult$age[myAdult$age < lowerAge | myAdult$age > upperAge]) / length(myAdult$age) * 100  
> ageOutliersPercentage  
[1] 0
```

Work Class

```
> ggplot(data = myAdult, aes(x = workClass)) + geom_bar(fill = "pink",  
  color = "black") + labs(x = "Work Class", y = "Frequency")
```

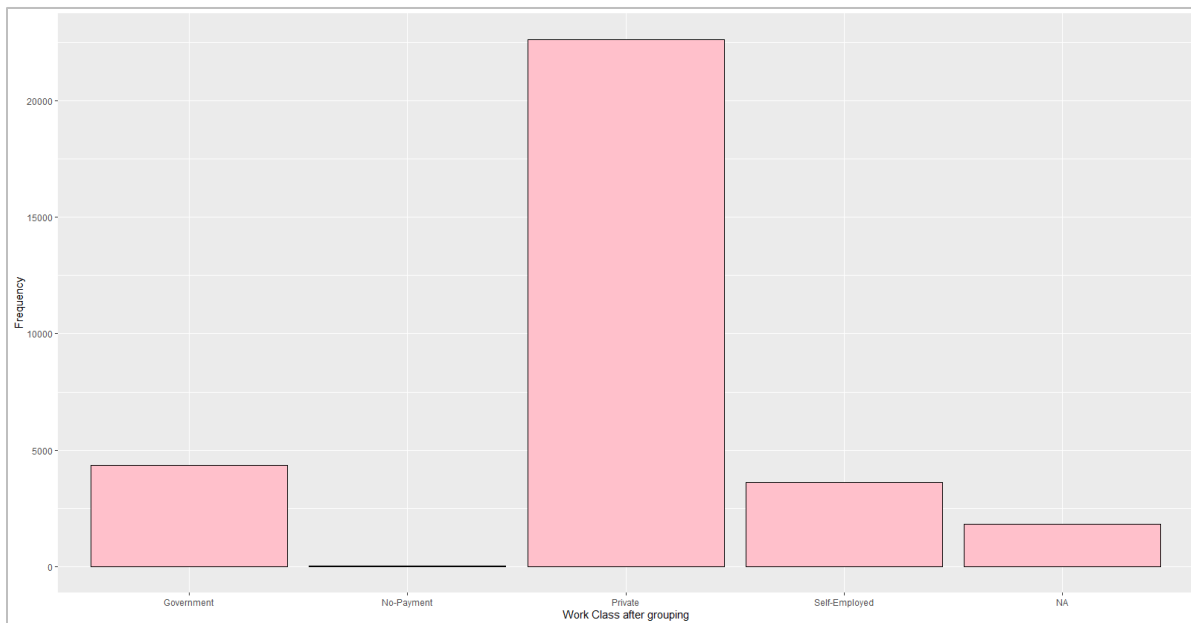


To improve the analysis, we will reduce the number of categories in this attribute by grouping them. We will combine the “Never-worked” and “Without-pay” groups into one group “No-payment”. Those who work in the government, whether “Federal-gov”, “State-gov”, or “Local-gov”, will be grouped into one group “Government”, while those who are “Self-emp-inc”, and “Self-emp-not-inc” will be grouped into a separate group “Self-Employed”.

```
> myAdult$workClass[myAdult$workClass == 'Never-worked' |  
  myAdult$workClass == 'without-pay' ] <- 'No-Payment'  
> myAdult$workClass[myAdult$workClass == 'Federal-gov' |  
  myAdult$workClass == 'State-gov' | myAdult$workClass == 'Local-gov'] <-  
  'Government'  
> myAdult$workClass[myAdult$workClass == 'Self-emp-inc' |  
  myAdult$workClass == 'Self-emp-not-inc'] <- 'Self-Employed'  
> unique(myAdult$workClass)
```

```
> myAdult$workClass[myAdult$workClass == 'Never-worked' | myAdult$workClass == 'without-pay' ] <- 'No-Payment'  
> myAdult$workClass[myAdult$workClass == 'Federal-gov' | myAdult$workClass == 'State-gov' | myAdult$workClass == 'Local-gov'] <- 'Government'  
> myAdult$workClass[myAdult$workClass == 'Self-emp-inc' | myAdult$workClass == 'Self-emp-not-inc'] <- 'Self-Employed'  
> unique(myAdult$workClass)  
[1] "Government" "Self-Employed" "Private" NA "No-Payment"
```

```
> ggplot(data = myAdult, aes(x = workClass)) + geom_bar(fill = "pink",
  color = "black") + labs(x = "Work Class after grouping", y = "Frequency")
```



```
> sum(is.na(myAdult$workClass)) / length(myAdult$workClass) * 100
```

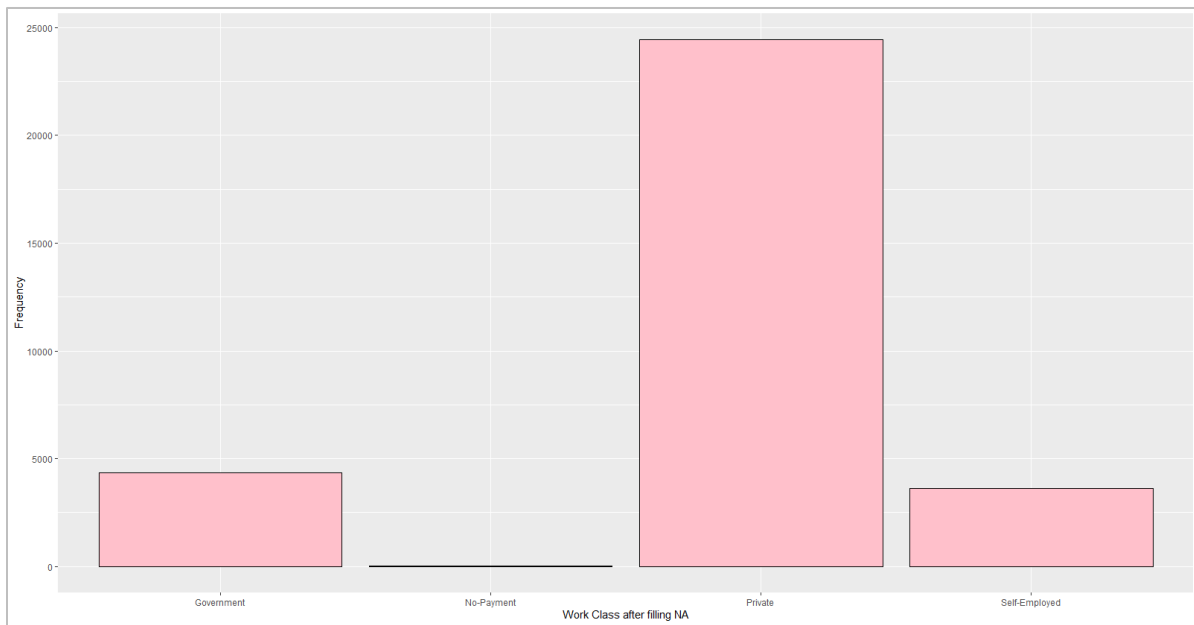
```
> sum(is.na(myAdult$workClass)) / length(myAdult$workClass) * 100
[1] 5.555556
```

We will not delete the column for missing values because the column is important, and we have to substitute the missing values with the most frequently occurring value (mode).

```
> myAdult$workClass[is.na(myAdult$workClass)] <- names(which.max(table(
  myAdult$workClass)))
> sum(is.na(myAdult$workClass)) / length(myAdult$workClass) * 100
```

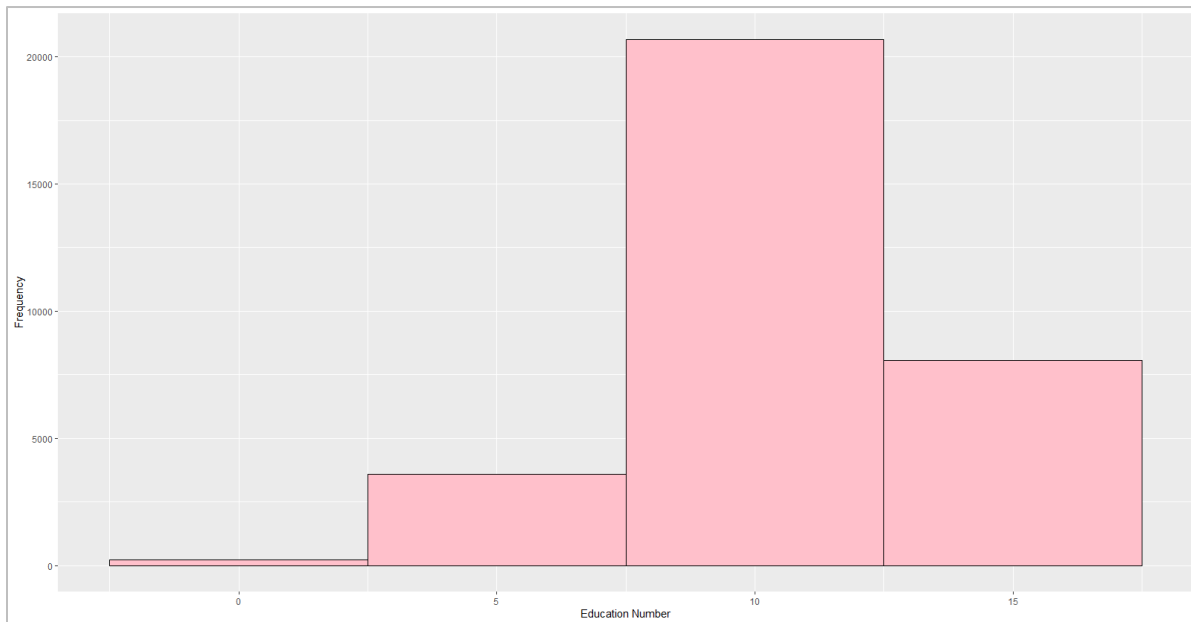
```
> sum(is.na(myAdult$workClass)) / length(myAdult$workClass) * 100
[1] 0
```

```
> ggplot(data = myAdult, aes(x = workClass)) + geom_bar(fill = "pink",  
  color = "black") + labs(x = "Work Class after filling NA", y = "Frequency")
```



Education Number

```
> ggplot(data = myAdult, aes(x = educationNum)) + geom_histogram(binwidth = 5, fill = "pink", color = "black") + labs(x = "Education Number", y = "Frequency")
```



```
> sum(is.na(myAdult$educationNum)) / length(myAdult$educationNum) * 100
```

```
> sum(is.na(myAdult$educationNum)) / length(myAdult$educationNum) * 100  
[1] 0
```

No missing value.

```
> unique(myAdult$educationNum)
```

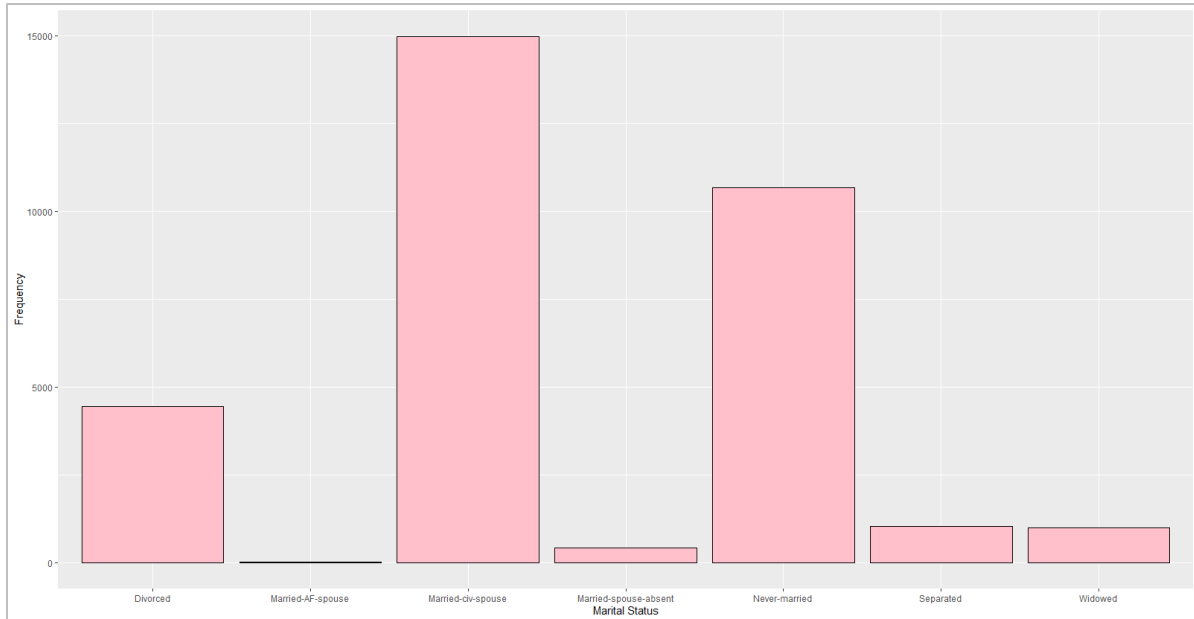
```
> unique(myAdult$educationNum)  
[1] 13 9 7 14 5 10 12 11 4 16 15 3 6 2 1 8
```

```
> myAdult$educationNum <- cut(myAdult$educationNum, breaks = c(0, 8, 10, 12, 14, Inf), labels = c("Elementary", "High School", "Some College", "Associate", "Higher Degree"), right = FALSE)  
> unique(myAdult$educationNum)
```

```
> myAdult$educationNum <- cut(myAdult$educationNum, breaks = c(0, 8, 10, 12, 14, Inf), labels = c("Elementary", "High School", "Some College", "Associate", "Higher Degree"), right = FALSE)  
> unique(myAdult$educationNum)  
[1] Associate High School Elementary Higher Degree Some College  
Levels: Elementary High School Some College Associate Higher Degree
```

Marital Status

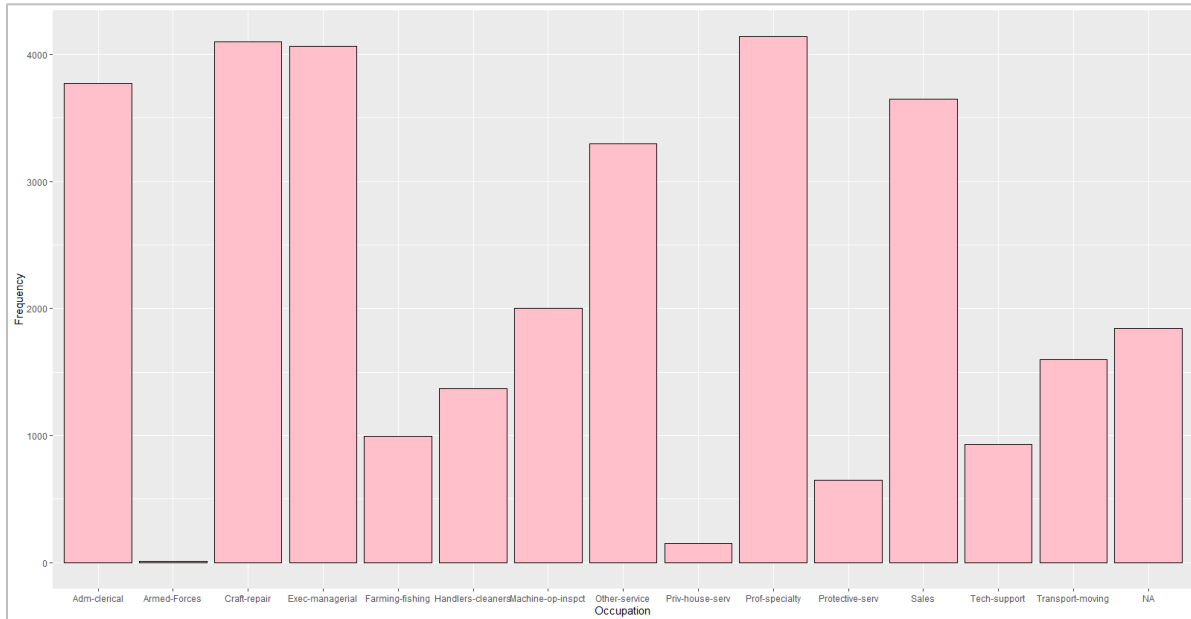
```
> ggplot(data = myAdult, aes(x = maritalStatus)) + geom_bar(fill = "pink",  
  color = "black") + labs(x = "Marital Status", y = "Frequency")
```



Everything is fine (unbiased), and there are no missing values.

Occupation

```
> ggplot(data = myAdult, aes(x = occupation)) + geom_bar(fill = "pink",  
  color = "black") + labs(x = "Occupation", y = "Frequency")
```



```
> sum(is.na(myAdult$occupation)) / length(myAdult$occupation) * 100
```

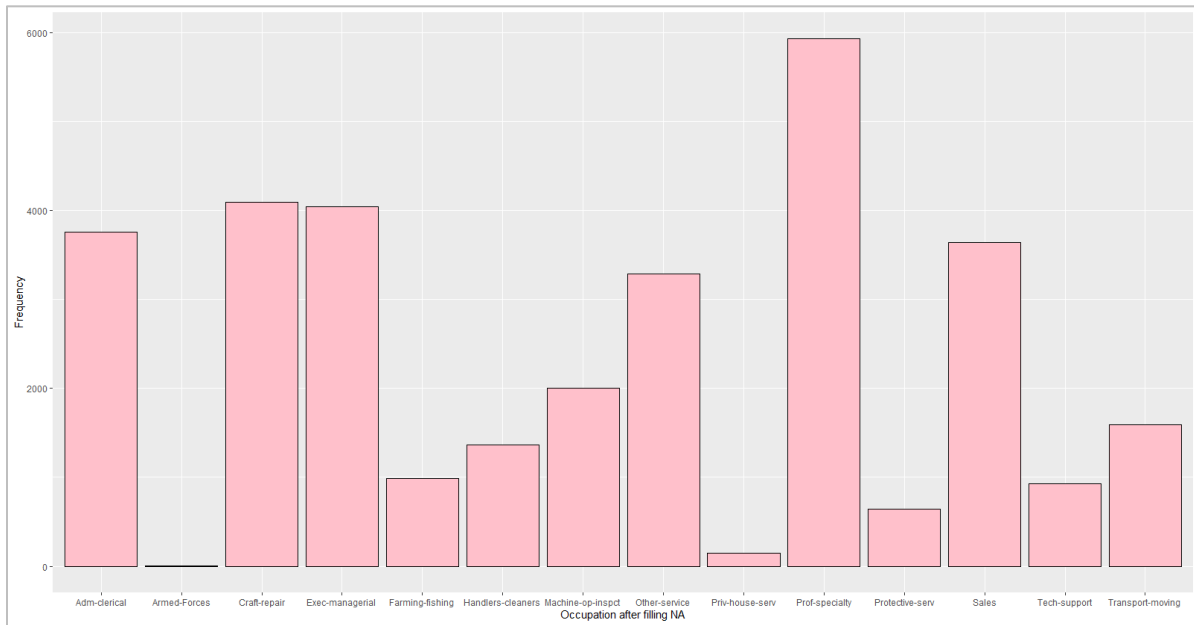
```
> sum(is.na(myAdult$occupation)) / length(myAdult$occupation) * 100  
[1] 5.577148
```

We will not delete the column for missing values because the column is important, and we have to substitute the missing values with the most frequently occurring value (mode).

```
> myAdult$occupation[is.na(myAdult$occupation)] <- names(which.max(table(  
  myAdult$occupation)))  
> sum(is.na(myAdult$occupation)) / length(myAdult$occupation) * 100
```

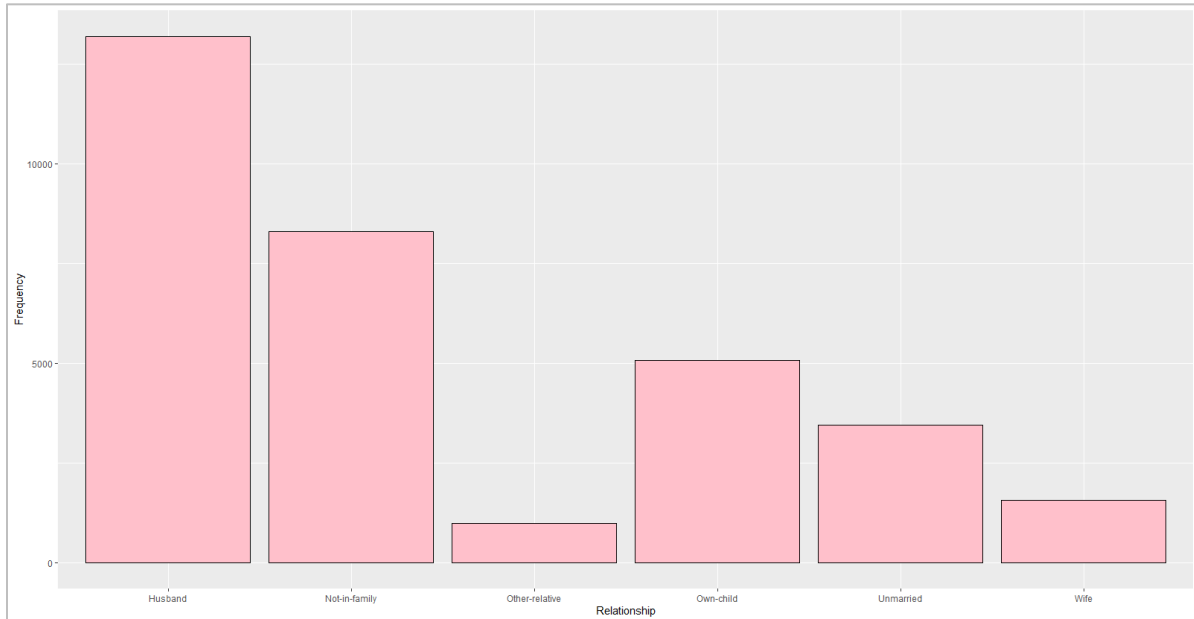
```
> myAdult$ occupation[is.na(myAdult$occupation)] <- names(which.max(table(myAdult$occupation)))  
> sum(is.na(myAdult$occupation)) / length(myAdult$occupation) * 100  
[1] 0
```

```
> ggplot(data = myAdult, aes(x = occupation)) + geom_bar(fill = "pink", color  
= "black") + labs(x = "Occupation after filling NA", y = "Frequency")
```



Relationship

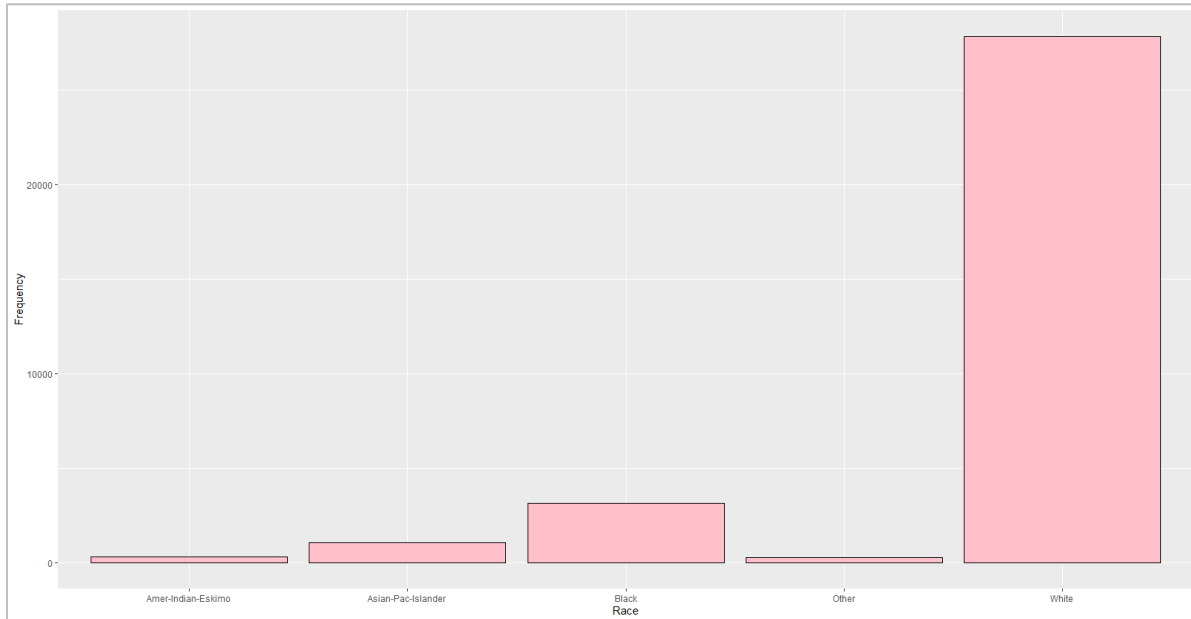
```
> ggplot(data = myAdult, aes(x = relationship)) + geom_bar(fill = "pink",  
  color = "black") + labs(x = "Relationship", y = "Frequency")
```



Everything is fine (unbiased), and there are no missing values.

Race

```
> ggplot(data = myAdult, aes(x = race)) + geom_bar(fill = "pink",  
  color = "black") + labs(x = "Race", y = "Frequency")
```



```
> sum(complete.cases(myAdult$race) & myAdult$race == "white") /  
  length(myAdult$race) * 100
```

```
> sum(complete.cases(myAdult$race) & myAdult$race == "white") / length(myAdult$race) * 100  
[1] 85.40626
```

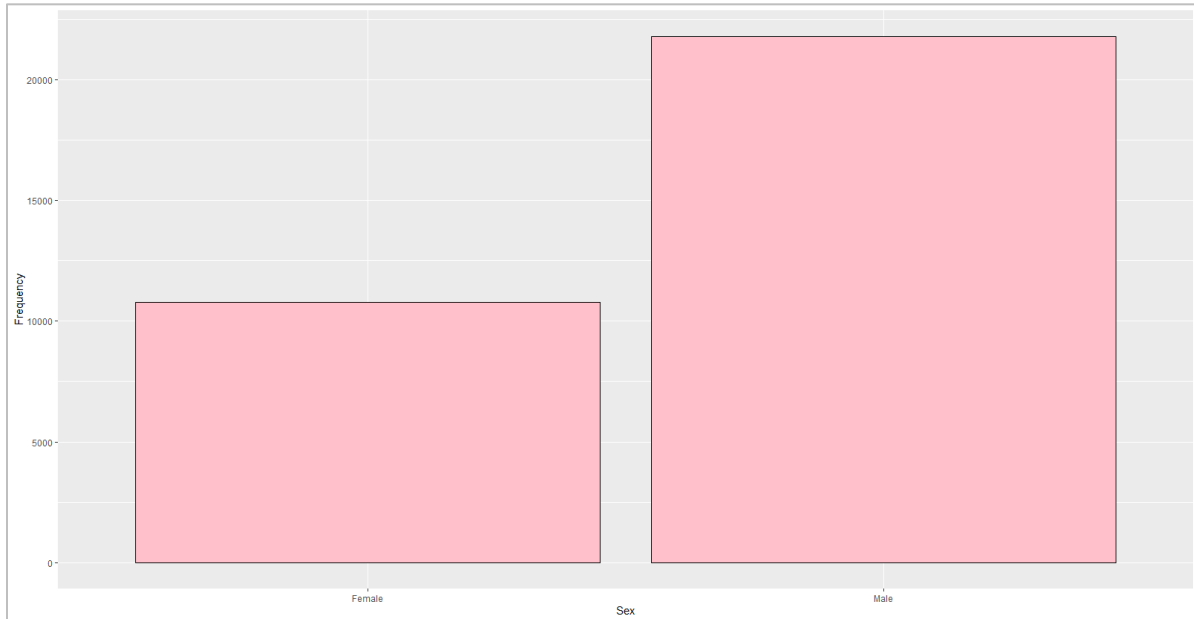
The attribute is heavily biased to the White (85.4%)

We are going to remove it.

```
> myAdult <- subset(myAdult, select = -race)
```

Sex

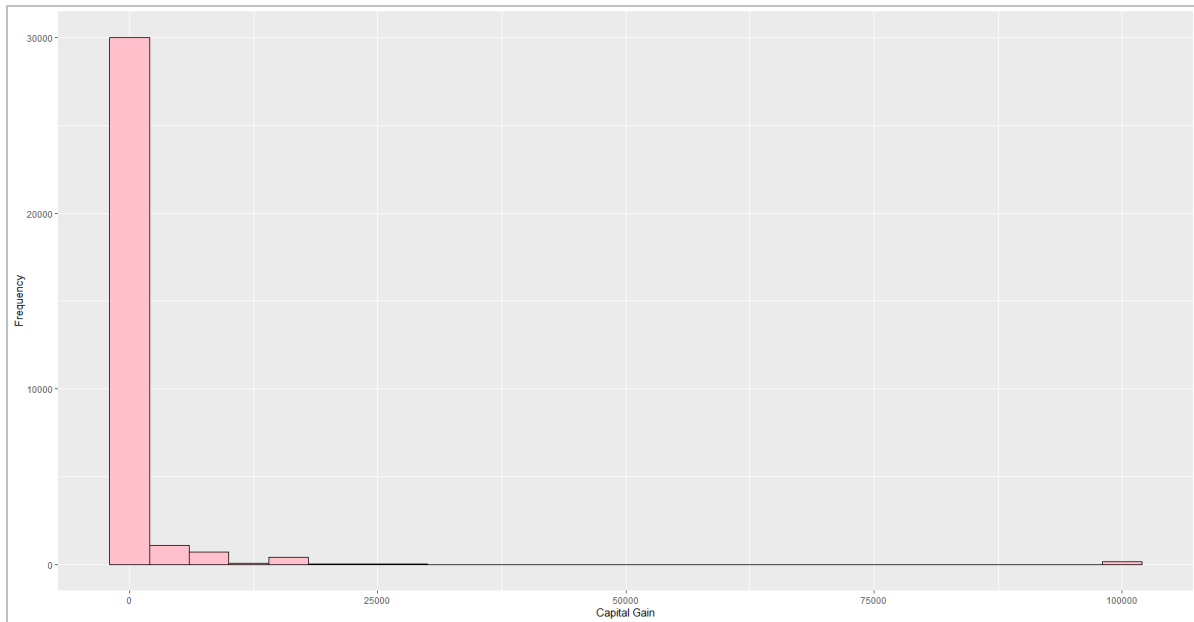
```
> ggplot(data = myAdult, aes(x = sex)) + geom_bar(fill = "pink",  
  color = "black") + labs(x = "Sex", y = "Frequency")
```



Everything is fine (unbiased), and there are no missing values.

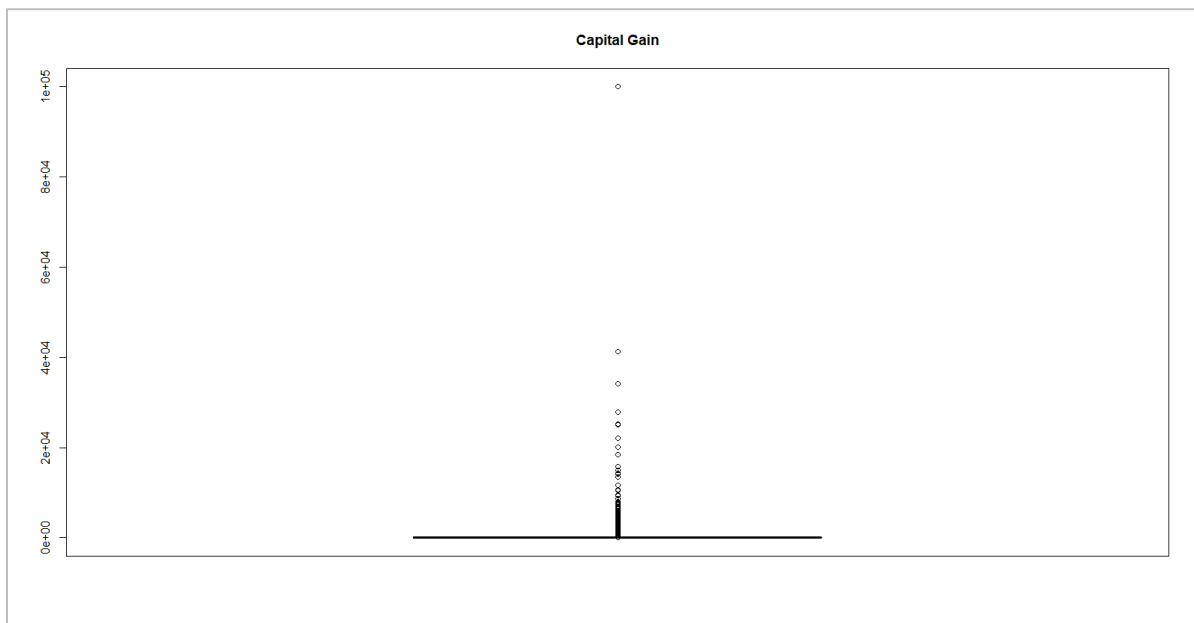
Capital Gain

```
> ggplot(data = myAdult, aes(x = capitalGain)) + geom_histogram(  
  binwidth = 4000, fill = "pink", color = "black") + labs(x = "Capital Gain",  
  y = "Frequency")
```



We will look for outliers in this property and delete them if they are less than 5%. Otherwise, we will smooth (using the binning approach).

```
> boxplot(myAdult$capitalGain, main = "Capital Gain")
```




```

> qCapGain <- quantile(myAdult$capitalGain, probs = c(0.25,0.75))
> iqrCapGain <- IQR(myAdult$capitalGain)
> lowerCapGain <- qCapGain[1] - (1.5 * iqrCapGain)
> upperCapGain <- qCapGain[2] + (1.5 * iqrCapGain)
> capGainOutliersPercentage <- length(myAdult$capitalGain[
  myAdult$capitalGain < lowerCapGain | myAdult$capitalGain > upperCapGain]) /
  length(myAdult$capitalGain) * 100
> capGainOutliersPercentage

```

```

> qcapGain <- quantile(myAdult$capitalGain, probs = c(0.25,0.75))
> iqrCapGain <- IQR(myAdult$capitalGain)
> lowerCapGain <- qcapGain[1] - (1.5 * iqrCapGain)
> upperCapGain <- qcapGain[2] + (1.5 * iqrCapGain)
> capGainOutliersPercentage <- length(myAdult$capitalGain[myAdult$capitalGain < lowerCapGain | myAdult$capitalGain > upperCapGain]) / length(myAdult$capitalGain) * 100
[1] 8.304029

```

Because the percentage is greater than 5%, we will use bin smoothing (equal frequency) to smooth data and decrease the influence of outliers to improve accuracy.

We opted to split it into 15 bins and save them in the gainBins variable. We also constructed two vectors: count to count the number of entries in each bin and sum to preserve the summing of each bin.

```

> gainBins <- ntile(myAdult$capitalGain, n = 15)
> meanGainbins <- replicate(15,0)
> for(i in 1 : 15)
  myAdult$capitalGain[gainBins==i] <- mean(myAdult$capitalGain[gainBins==i])

```

```

> gainBins<-ntile(myAdult$capitalGain, n = 15)
> meanGainbins <- replicate(15,0)
> for(i in 1 : 15)
+   myAdult$capitalGain[gainBins==i] <- mean(myAdult$capitalGain[gainBins==i])

```

After eliminating the outliers, all values became zero, therefore the attribute is no longer required, and we will delete it from our data because it is the same value for all rows.

```

> myAdult <- subset(myAdult, select = -capitalGain)

```

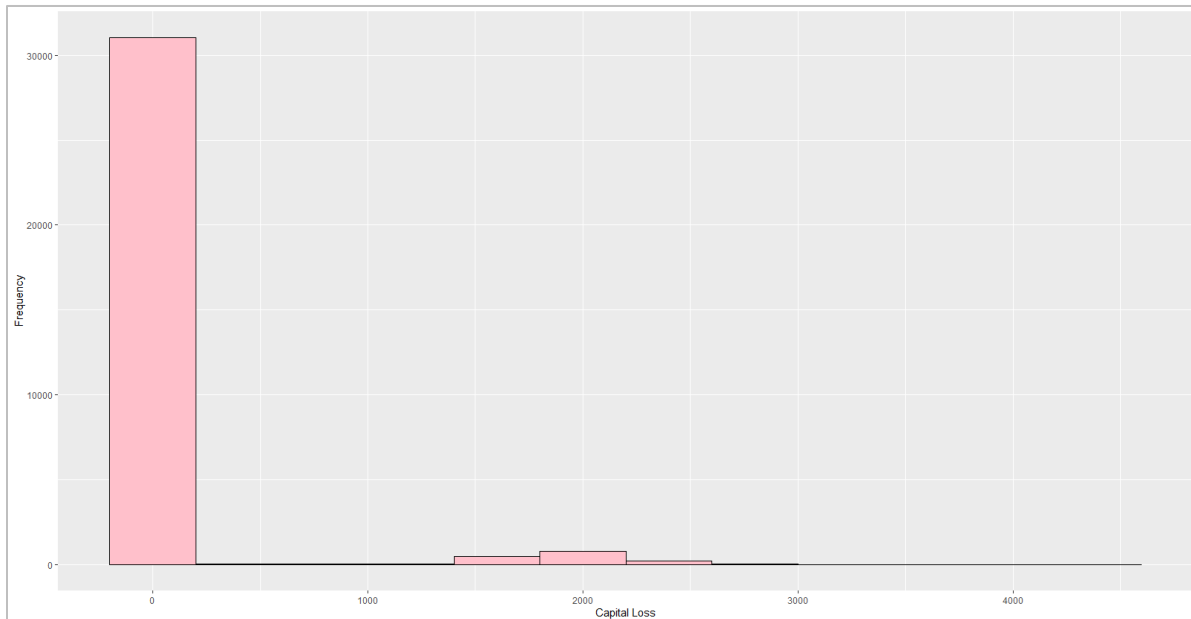
```

> myAdult <- subset(myAdult, select = -capitalGain)

```

Capital Loss

```
> ggplot(data = myAdult, aes(x = capitalLoss)) + geom_histogram(  
  binwidth = 400, fill = "pink", color = "black") + labs(x = "Capital Loss",  
  y = "Frequency")
```



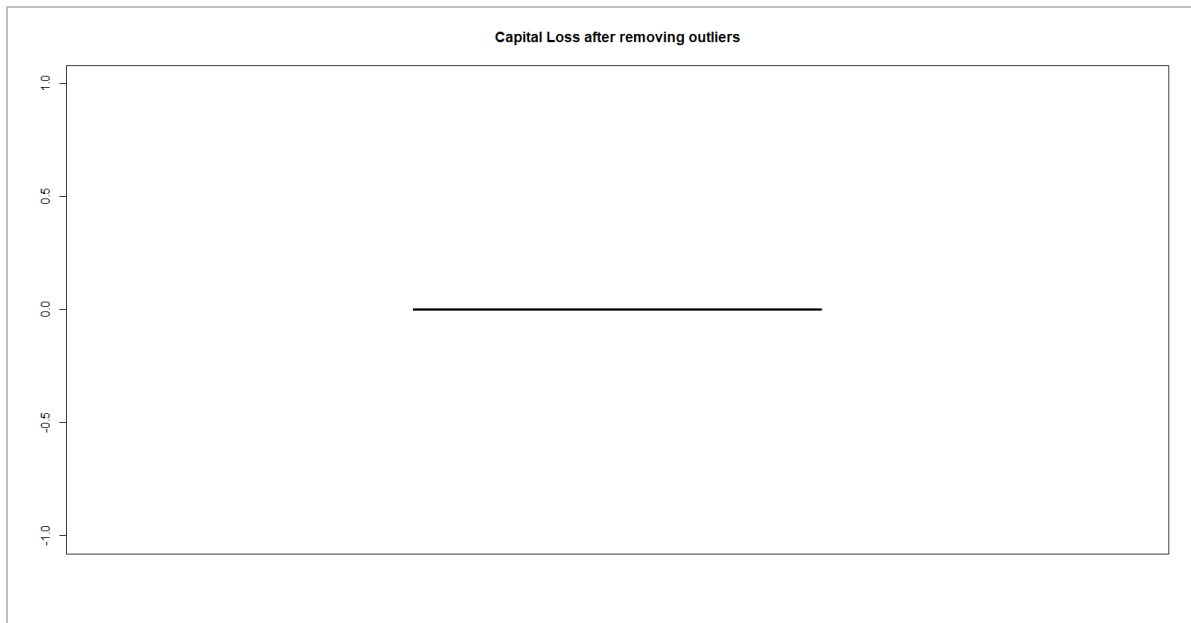
```
> qCapLoss <- quantile(myAdult$capitalLoss, probs = c(0.25,0.75))  
> iqrCapLoss <- IQR(myAdult$capitalLoss)  
> lowerCapLoss <- qCapLoss[1] - (1.5 * iqrCapLoss)  
> upperCapLoss <- qCapLoss[2] + (1.5 * iqrCapLoss)  
> capLossOutliersPercentage
```

```
> qCapLoss <- quantile(myAdult$capitalLoss, probs = c(0.25,0.75))  
> iqrCapLoss <- IQR(myAdult$capitalLoss)  
> lowerCapLoss <- qCapLoss[1] - (1.5 * iqrCapLoss)  
> upperCapLoss <- qCapLoss[2] + (1.5 * iqrCapLoss)  
> capLossOutliersPercentage <- length(myAdult$capitalLoss[myAdult$capitalLoss < lowerCapLoss | myAdult$capitalLoss > upperCapLoss]) / length(myAdult$capitalLoss) * 100  
> capLossOutliersPercentage  
[1] 4.664076
```

Because the capitalLoss outlier percentage is less than 5%, we are going to remove outliers:

```
> myAdult <- subset(myAdult, capitalLoss >=  
  max(summary(myAdult$capitalLoss)[1], lowerCapLoss) & capitalLoss <=  
  min(summary(myAdult$capitalLoss)[6], upperCapLoss))
```

```
> boxplot(myAdult$capitalLoss, main = "Capital Loss after removing outliers")
```



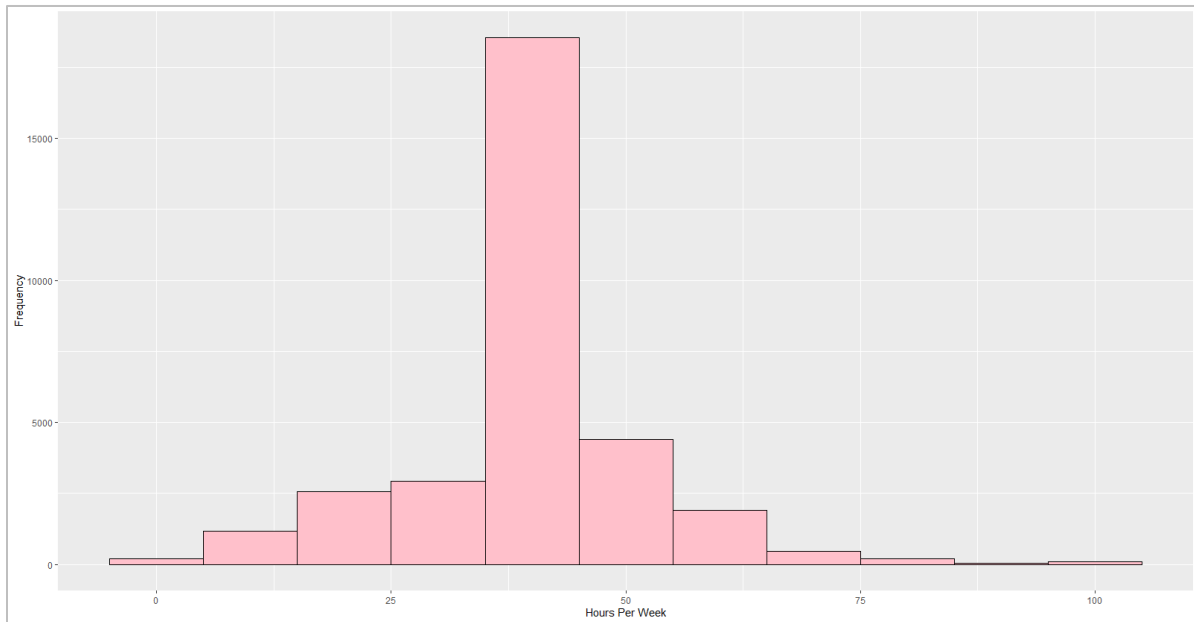
After eliminating the outliers, all values became zero, therefore the attribute is no longer required, and we will delete it from our data because it is the same value for all rows.

```
> myAdult <- subset(myAdult, select = -capitalLoss)
```

```
> myAdult <- subset(myAdult, select = -capitalLoss)
```

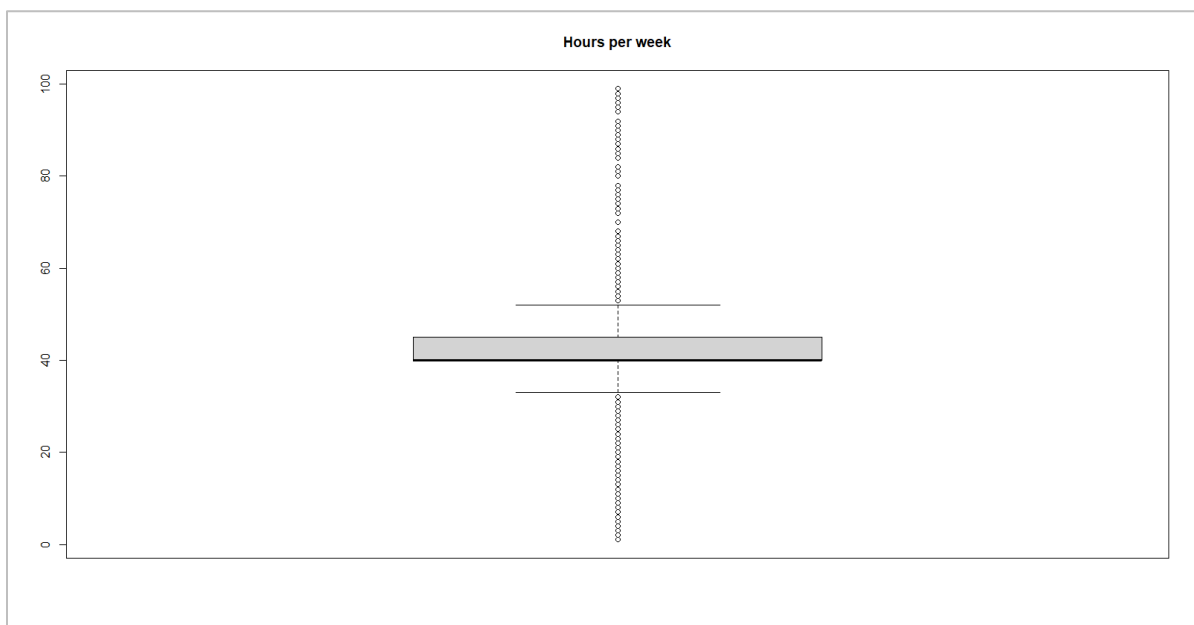
Hours per week

```
> ggplot(data = myAdult, aes(x = hoursPerWeek)) + geom_histogram(  
  binwidth = 10, fill = "pink", color = "black") + labs(x = "Hours Per Week",  
  y = "Frequency")
```



We will look for outliers in this property and delete them if they are less than 5%. Otherwise, we will smooth (using the binning approach).

```
> boxplot(myAdult$hoursPerWeek, main = "Hours per week")
```



```

> qHpw <- quantile(myAdult$hoursPerWeek, probs = c(0.25,0.75))
> iqrHpw <- IQR(myAdult$hoursPerWeek)
> lowerHpw <- qHpw[1] - (1.5 * iqrHpw)
> upperHpw <- qHpw[2] + (1.5 * iqrHpw)
> hpwOutliersPercentage <- length(myAdult$hoursPerWeek[myAdult$hoursPerWeek <
  lowerHpw | myAdult$hoursPerWeek > upperHpw]) / length(myAdult$hoursPerWeek)
  * 100
> hpwOutliersPercentage

```

```

> qHpw <- quantile(myAdult$hoursPerWeek, probs = c(0.25,0.75))
> iqrHpw <- IQR(myAdult$hoursPerWeek)
> lowerHpw <- qHpw[1] - (1.5 * iqrHpw)
> upperHpw <- qHpw[2] + (1.5 * iqrHpw)
> hpwOutliersPercentage <- length(myAdult$hoursPerWeek[myAdult$hoursPerWeek < lowerHpw | myAdult$hoursPerWeek > upperHpw]) / length(myAdult$hoursPerWeek) * 100
> hpwOutliersPercentage
[1] 27.5707

```

Because the percentage is greater than 5%, we will use bin smoothing (equal frequency) to smooth data and decrease the influence of outliers to improve accuracy.

We opted to split it into 15 bins and save them in the gainBins variable. We also constructed two vectors: count to count the number of entries in each bin and sum to preserve the summing of each bin.

```

> gainBins <- ntile(myAdult$hoursPerWeek, n = 15)
> meanGainbins <- replicate(15,0)
> for(i in 1 : 15)
  myAdult$hoursPerWeek[gainBins==i] <-
  mean(myAdult$hoursPerWeek[gainBins==i])

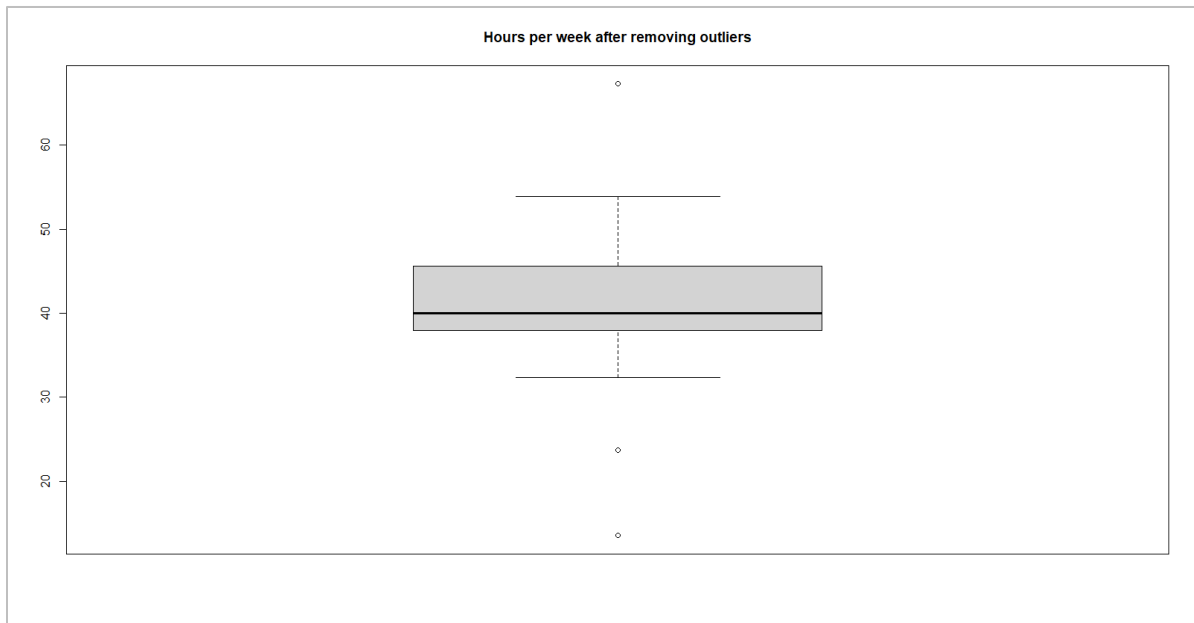
```

```

> gainBins <- ntile(myAdult$hoursPerWeek, n = 15)
> meanGainbins <- replicate(15,0)
> for(i in 1 : 15)
+   myAdult$hoursPerWeek[gainBins==i] <- mean(myAdult$hoursPerWeek[gainBins==i])

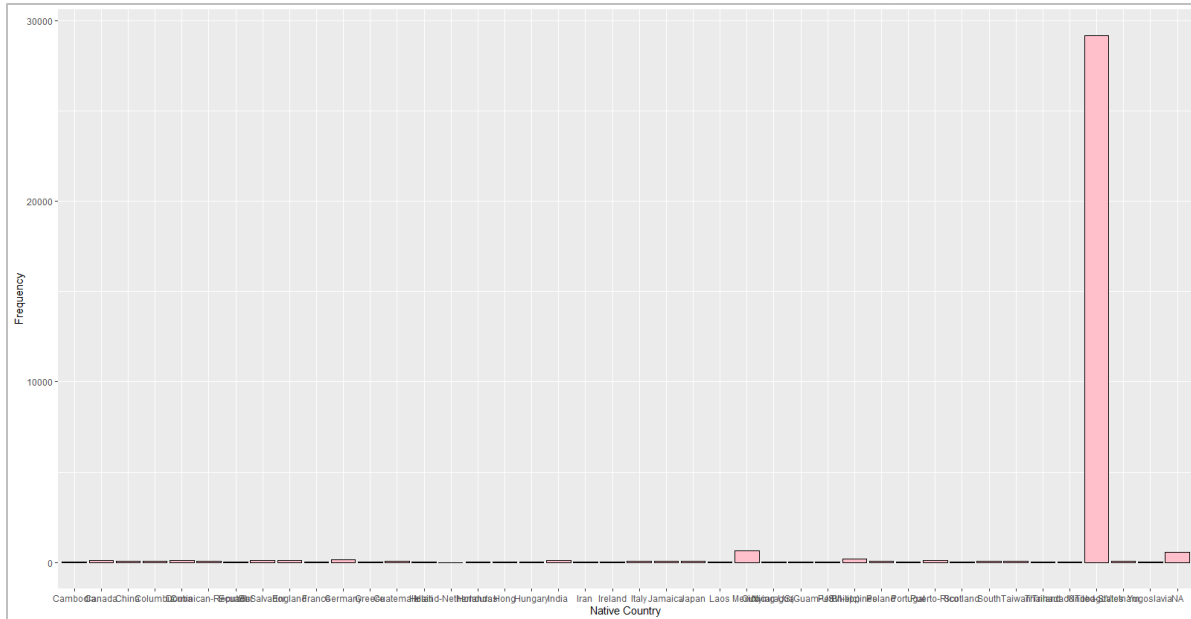
```

```
> boxplot(myAdult$hoursPerWeek,main="Hours per week after removing outliers")
```



Native Country

```
> ggplot(data = myAdult, aes(x = nativeCountry)) + geom_bar(fill = "pink",  
  color = "black") + labs(x = "Native Country", y = "Frequency")
```



```
> sum(complete.cases(myAdult$nativeCountry) & myAdult$nativeCountry == "United-States") / length(myAdult$nativeCountry) * 100
```

```
> sum(complete.cases(myAdult$nativeCountry) & myAdult$nativeCountry == "United-States") / length(myAdult$nativeCountry) * 100  
[1] 89.58912
```

The attribute is heavily biased to the United States (89.5%)

We are going to remove it.

```
> myAdult <- subset(myAdult, select = -nativeCountry)
```

Data Transformation

Normalizing numerical data into range [0,1].

```
> standardize <- function(x, nMin, nMax) { return ((x - min(x)) / (max(x) -  
  min(x)) * (nMax - nMin) + nMin) }  
> myAdult$age <- standardize(myAdult$age, 0, 1)  
> myAdult$hoursPerWeek <- standardize(myAdult$hoursPerWeek, 0, 1)
```


Preparing the income attribute as it's our factor

```
> myAdult$income <- as.factor(myAdult$income)
> class(myAdult$income)
> myAdult$income
```

[illegible]

```
> str(myAdult)
```

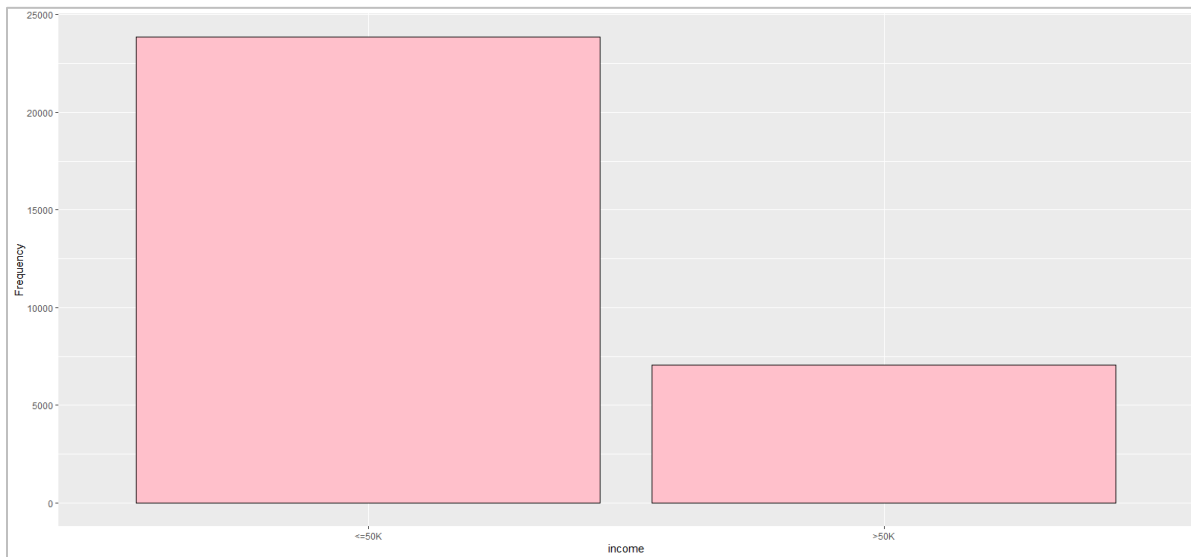
```
> str(myAdult)
tibble [30,906 × 9] (s3: tbl_df/tbl/data.frame)
 $ age      : num [1:30906] 0.361 0.541 0.344 0.59 0.18 ...
 $ workClass : chr [1:30906] "Government" "Self-Employed" "Private" "Private" ...
 $ educationNum : Factor w/ 5 levels "Elementary","High school",...: 4 4 2 1 4 5 1 2 5 4 ...
 $ maritalStatus: chr [1:30906] "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
 $ occupation   : chr [1:30906] "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Handlers-cleaners" ...
 $ relationship : chr [1:30906] "Not-in-family" "Husband" "Not-in-family" "Husband" ...
 $ sex          : chr [1:30906] "Male" "Male" "Male" "Male" ...
 $ hoursPerweek : num [1:30906] 0.454 0 0.454 0.454 0.454 ...
 $ income       : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
```

Adult dataset after cleaning is done

	age	workClass	educationNum	maritalStatus	occupation	relationship	sex	hoursPerWeek	income
1	0.36065574	Government	Associate	Never-married	Adm-clerical	Not-in-family	Male	0.4539759	<=50K
2	0.54098361	Self-Employed	Associate	Married-civ-spouse	Exec-managerial	Husband	Male	0.0000000	<=50K
3	0.34426230	Private	High School	Divorced	Handlers-cleaners	Not-in-family	Male	0.4539759	<=50K
4	0.59016393	Private	Elementary	Married-civ-spouse	Handlers-cleaners	Husband	Male	0.4539759	<=50K
5	0.18032787	Private	Associate	Married-civ-spouse	Prof-specialty	Wife	Female	0.4539759	<=50K
6	0.32786885	Private	Higher Degree	Married-civ-spouse	Exec-managerial	Wife	Female	0.4539759	<=50K
7	0.52459016	Private	Elementary	Married-spouse-absent	Other-service	Not-in-family	Female	0.0000000	<=50K
8	0.57377049	Self-Employed	High School	Married-civ-spouse	Exec-managerial	Husband	Male	0.5131665	>50K
9	0.22950820	Private	Higher Degree	Never-married	Prof-specialty	Not-in-family	Female	0.6756760	>50K
10	0.40983607	Private	Associate	Married-civ-spouse	Exec-managerial	Husband	Male	0.4539759	>50K
11	0.32786885	Private	Some College	Married-civ-spouse	Exec-managerial	Husband	Male	1.0000000	>50K
12	0.21311475	Government	Associate	Married-civ-spouse	Prof-specialty	Husband	Male	0.4539759	>50K
13	0.09836066	Private	Associate	Never-married	Adm-clerical	Own-child	Female	0.1897114	<=50K
14	0.24590164	Private	Associate	Never-married	Sales	Not-in-family	Male	0.6756760	<=50K
15	0.37704918	Private	Some College	Married-civ-spouse	Craft-repair	Husband	Male	0.4539759	>50K
16	0.27868852	Private	Elementary	Married-civ-spouse	Transport-moving	Husband	Male	0.5131665	<=50K
17	0.13114754	Self-Employed	High School	Never-married	Farming-fishing	Own-child	Male	0.3507167	<=50K
18	0.24590164	Private	High School	Never-married	Machine-op-inspct	Unmarried	Male	0.4539759	<=50K
19	0.34426230	Private	Elementary	Married-civ-spouse	Sales	Husband	Male	0.6756760	<=50K
20	0.42622951	Self-Employed	Higher Degree	Divorced	Exec-managerial	Unmarried	Female	0.5131665	>50K
21	0.37704918	Private	Higher Degree	Married-civ-spouse	Prof-specialty	Husband	Male	0.7503825	>50K
22	0.60655738	Private	High School	Separated	Other-service	Unmarried	Female	0.0000000	<=50K
23	0.29508197	Government	Elementary	Married-civ-spouse	Farming-fishing	Husband	Male	0.4539759	<=50K
24	0.68852459	Private	High School	Divorced	Tech-support	Unmarried	Female	0.4539759	<=50K

Showing 1 to 25 of 30,906 entries, 9 total columns

```
> ggplot(data = myAdult, aes(x = income)) + geom_bar(fill = "pink", color = "black") + labs(x = "income", y = "Frequency")
```



```
> table(myAdult$income)
```

```
> table(myAdult$income)
<=50K >50K
23861  7045
```

```
> sum(myAdult$income == '<=50K') / length(myAdult$income) * 100
```

```
> sum(myAdult$income == '<=50K') / length(myAdult$income) * 100
[1] 77.20507
```

our data is **biased** to <=50K class

Part 2

```
> dataframe <- data.frame(workclass = head(myAdult$workClass, n = 100),
+ income = head(myAdult$occupation, n = 100))
> transactions <- as(dataframe, "transactions")
> frequentSet <- eclat(transactions, parameter = list(support = 0.1))
> inspect(frequentSet)
```

```
> library(arules)
> dataframe <- data.frame(workclass = head(myAdult$workClass, n = 100),
+ income = head(myAdult$occupation, n = 100))
> transactions <- as(dataframe, "transactions")
```

```
> frequentSet <- eclat(transactions, parameter = list(support = 0.1))
```

Eclat

parameter specification:

tidLists	support	minlen	maxlen	target	ext
FALSE	0.1	1	10	frequent itemsets	TRUE

algorithmic control:

sparse	sort	verbose
7	-2	TRUE

Absolute minimum support count: 10

create itemset ...

set transactions ... [15 item(s), 100 transaction(s)] done [0.00s].

sorting and recoding items ... [7 item(s)] done [0.00s].

creating bit matrix ... [7 row(s), 100 column(s)] done [0.00s].

writing ... [9 set(s)] done [0.00s].

Creating S4 object ... done [0.00s].

```
> inspect(frequentSet)
```

	items	support	count
[1]	{workclass=Private, income=Sales}	0.11	11
[2]	{workclass=Private, income=Prof-specialty}	0.12	12
[3]	{workclass=Private}	0.73	73
[4]	{income=Prof-specialty}	0.18	18
[5]	{workclass=Government}	0.16	16
[6]	{income=Exec-managerial}	0.15	15
[7]	{income=Sales}	0.12	12
[8]	{workclass=Self-Employed}	0.11	11
[9]	{income=Other-service}	0.10	10

```
> |
```

To find the strong association :-

```
> association_rules <- apriori(transactions, parameter = list(supp = 0.1,
  conf = 0.8))
> inspect(association_rules)
```

```
>
> association_rules <- apriori(transactions, parameter = list(supp = 0.1, conf = 0.8))
Apriori

Parameter specification:
 confidence minval smax arem aval originalSupport maxtime support minlen maxlen target ext
 0.8      0.1    1 none FALSE          TRUE      5    0.1    1    10 rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
 0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 10

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[15 item(s), 100 transaction(s)] done [0.00s].
sorting and recoding items ... [7 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 done [0.01s].
writing ... [1 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
> inspect(association_rules)
   lhs               rhs      support confidence coverage lift    count
[1] {income=Sales} => {workclass=Private} 0.11    0.9166667 0.12    1.255708 11
>
```

The End ..