Data Cleaning and Preprocessing

Sabbir Ahmed Hemo

June 13, 2020

Loading libraries

```
library(tidyverse)
library(janitor)
library(knitr)
```

Loading datasets

```
df <- read_csv("./data/upwork_data_0.csv") %>%
  filter(!is.na(over_50k))

glimpse(df)
```

```
## Rows: 26,934
## Columns: 15
## $ id             <dbl> 12106, 28951, 24570, 16358, 9375, 10738, 20733, 1317...
## $ age            <dbl> 32, 43, 35, 31, 64, 55, 41, 39, 60, 62, -1, 32, 27, ...
## $ workclass      <chr> "Private", "State-gov", "Private", "Private", "Priva...
## $ education      <chr> "HS-grad", "Some-college", "HS-grad", NA, "Some-coll...
## $ education_num  <dbl> 9, 10, 9, 14, 10, 10, 10, 13, 6, 9, 9, 10, 13, 13, 6...
## $ marital_status <chr> "Divorced", "Divorced", "Married-civ-spouse", "Never...
## $ occupation     <chr> "Adm-clerical", "Adm-clerical", "Exec-managerial", "...
## $ relationship   <chr> "Other-relative", "Unmarried", "Wife", "Not-in-famil...
## $ race           <chr> "W hite", "W hite", "White", "Black", "White", "Whit...
## $ sex            <chr> "Female", "Female", "Female", "Male", "Female", "Mal...
## $ capital_gain   <dbl> 0, 0, 0, 0, 10566, 0, 0, 0, 0, 0, 0, 3464, 0, 0, 501...
## $ capital_loss   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ hours_per_week <dbl> 40, 40, 40, 40, 35, 70, 50, 35, 40, 40, 40, 40, 40, ...
## $ native_country <chr> "United-States", "United-States", "United-States", "...
## $ over_50k       <chr> "<=50K", "<=50K", ">50K", "<=50K", "<=50K", ">50K", ...
```

Summary table for numeric variables

```
summary(df %>% select_if(is.numeric)) %>% kable()
```

id	age	education_num	capital_gain	capital_loss	hours_per_week
Min. : 3	Min. :-1.0	Min. : 1.000	Min. :-99999.0	Min. :-2457.00	Min. : 1.00
1st Qu.: 8123	1st Qu.:27.0	1st Qu.: 9.000	1st Qu.: 0.0	1st Qu.: 0.00	1st Qu.: 38.00
Median :16316	Median :36.0	Median : 9.000	Median : 0.0	Median : 0.00	Median : 40.00
Mean :16279	Mean :35.7	Mean : 9.638	Mean : 516.1	Mean : 60.95	Mean : 41.67
3rd Qu.:24435	3rd Qu.:47.0	3rd Qu.:10.000	3rd Qu.: 0.0	3rd Qu.: 0.00	3rd Qu.: 44.00
Max. :32560	Max. :90.0	Max. :16.000	Max. : 99999.0	Max. : 4356.00	Max. :250.00

Cross tabs of categorical variables with dependent variable *over_50K*

```
df %>%
  select_if(is.character) %>%
  map(.x = names(select(., -over_50k)), .f = ~tabyl(df, !!sym(.x), over_50k) %>%
        adorn_percentages() %>%
        adorn_pct_formatting() %>%
        adorn_ns() %>%
        kable())
```

[[1]]

workclass	<=50K	>50K
?	93.7% (1534)	6.3% (104)
Federal-gov	72.9% (548)	27.1% (204)
Local-gov	83.3% (1380)	16.7% (276)
Never-worked	100.0% (6)	0.0% (0)
Private	86.3% (16477)	13.7% (2622)
Self-emp-inc	61.0% (445)	39.0% (284)
Self-emp-not-inc	81.8% (1670)	18.2% (371)
State-gov	87.7% (877)	12.3% (123)
Without-pay	100.0% (13)	0.0% (0)

[[2]]

education	<=50K	>50K
10th	93.8% (750)	6.2% (50)
11th	94.8% (943)	5.2% (52)
12th	92.2% (320)	7.8% (27)
1st-4th	95.9% (142)	4.1% (6)
5th-6th	94.7% (268)	5.3% (15)
7th-8th	94.7% (531)	5.3% (30)
9th	94.5% (399)	5.5% (23)
Assoc-acdm	74.7% (666)	25.3% (226)
Assoc-voc	73.8% (850)	26.2% (301)
Bachelors	94.0% (2621)	6.0% (167)
Doctorate	81.4% (92)	18.6% (21)

education	<=50K	>50K
HS-grad	84.0% (7368)	16.0% (1403)
Masters	90.4% (643)	9.6% (68)
Preschool	100.0% (41)	0.0% (0)
Prof-school	79.3% (134)	20.7% (35)
Some-college	81.0% (4896)	19.0% (1152)
NA	84.9% (2286)	15.1% (408)

[[3]]

marital_status	<=50K	>50K
Divorced	94.4% (3739)	5.6% (222)
Married-AF-spouse	66.7% (12)	33.3% (6)
Married-civ-spouse	68.6% (7656)	31.4% (3500)
Married-spouse-absent	97.1% (362)	2.9% (11)
Never-married	98.2% (9460)	1.8% (174)
Separated	96.9% (879)	3.1% (28)
Widowed	95.1% (842)	4.9% (43)

[[4]]

occupation	<=50K	>50K
?	93.7% (1540)	6.3% (104)
Adm-clerical	89.8% (3031)	10.2% (344)
Armed-Forces	100.0% (8)	0.0% (0)
Craft-repair	78.9% (2950)	21.1% (790)
Exec-managerial	71.8% (1947)	28.2% (766)
Farming-fishing	90.6% (808)	9.4% (84)
Handlers-cleaners	94.5% (1186)	5.5% (69)
Machine-op-inspct	88.9% (1631)	11.1% (203)
Other-service	96.3% (2913)	3.7% (112)
Priv-house-serv	100.0% (139)	0.0% (0)
Prof-specialty	84.4% (2135)	15.6% (395)
Protective-serv	72.4% (401)	27.6% (153)
Sales	83.3% (2486)	16.7% (499)
Tech-support	76.6% (601)	23.4% (184)
Transport-moving	80.7% (1174)	19.3% (281)

[[5]]

relationship	<=50K	>50K
Husband	68.7% (6733)	31.3% (3062)
Not-in-family	95.3% (6891)	4.7% (337)
Other-relative	97.0% (872)	3.0% (27)
Own-child	99.3% (4684)	0.7% (35)
Unmarried	96.5% (3017)	3.5% (109)
Wife	64.5% (753)	35.5% (414)

[[6]]

race	<=50K	>50K
Amer-Indian-Eskimo	88.3% (263)	11.7% (35)
Asian-Pac-Islander	73.4% (708)	26.6% (257)
Black	87.8% (2550)	12.2% (355)
Other	90.7% (235)	9.3% (24)
W hite	94.8% (3183)	5.2% (176)
White	83.6% (16011)	16.4% (3137)

[[7]]

sex	<=50K	>50K
Female	93.7% (8892)	6.3% (594)
Male	80.6% (14058)	19.4% (3390)

[[8]]

native_country	<=50K	>50K
?	85.4% (408)	14.6% (70)
Cambodia	61.1% (11)	38.9% (7)
Canada	79.0% (79)	21.0% (21)
China	72.1% (49)	27.9% (19)
Columbia	100.0% (52)	0.0% (0)
Cuba	84.1% (69)	15.9% (13)
Dominican-Republic	95.5% (63)	4.5% (3)
Ecuador	88.5% (23)	11.5% (3)
El-Salvador	95.6% (87)	4.4% (4)
England	81.5% (53)	18.5% (12)
France	76.2% (16)	23.8% (5)
Germany	82.1% (87)	17.9% (19)
Greece	85.2% (23)	14.8% (4)
Guatemala	93.7% (59)	6.3% (4)
Haiti	90.2% (37)	9.8% (4)
Holand-Netherlands	100.0% (1)	0.0% (0)
Honduras	100.0% (13)	0.0% (0)
Hong	65.0% (13)	35.0% (7)
Hungary	72.7% (8)	27.3% (3)
India	62.9% (56)	37.1% (33)
Iran	78.6% (22)	21.4% (6)
Ireland	91.3% (21)	8.7% (2)
Italy	82.5% (47)	17.5% (10)
Jamaica	87.3% (69)	12.7% (10)
Japan	69.6% (39)	30.4% (17)
Laos	88.2% (15)	11.8% (2)
Mexico	96.6% (572)	3.4% (20)
Nicaragua	93.8% (30)	6.2% (2)
Outlying-US(Guam-USVI-etc)	100.0% (12)	0.0% (0)
Peru	96.4% (27)	3.6% (1)
Philippines	67.8% (122)	32.2% (58)

native_country	<=50K	>50K
Poland	88.7% (47)	11.3% (6)
Portugal	91.4% (32)	8.6% (3)
Puerto-Rico	93.0% (93)	7.0% (7)
Scotland	90.0% (9)	10.0% (1)
South	84.5% (60)	15.5% (11)
Taiwan	68.9% (31)	31.1% (14)
Thailand	82.4% (14)	17.6% (3)
Trinadad&Tobago	88.2% (15)	11.8% (2)
United-States	85.1% (20404)	14.9% (3569)
Vietnam	91.2% (52)	8.8% (5)
Yugoslavia	71.4% (10)	28.6% (4)

Cleaning and recoding variables

```
df_cleaned <- df %>%
  mutate(hours_per_week = replace(
    x = hours_per_week,
    list = hours_per_week > 100,
    values = NA), # removing outliers
         age = replace(x = age,
                       list = age < 18,
                       values = NA), # removing outliers
         relationship = case_when(
           relationship %in% c("Husband", "Wife") ~ "with spouse",
           relationship %in% c("Not-in-family", "Unmarried") ~ relationship,
           TRUE ~ "Without spouse" # Recoding into 4 category
         ),
         occupation = case_when(
           occupation %in% c('?', 'Armed-Forces', 'Farming-fishing',
                             'Handlers-cleaners', 'Other-service',
                             'Priv-house-serv') ~ 'Low Salary Jobs', # Recoding low salary jobs togethe
           TRUE ~ occupation
         ),
         marital_status = case_when(
           marital_status %in% c("Married-civ-spouse",
                                 "Divorced", "Never-married") ~ marital_status,
           marital_status %in% c("Separated",
                                 "Widowed") ~ "Sep or widowed", # Recoding separted and widowed into si
           TRUE ~ "Others"
         ),
         workclass = case_when(
           workclass %in% c("?", "Never-worked",
                            "Without-pay") ~ "Others",
           TRUE ~ workclass
         ),
         native_country = case_when(
           native_country %in% c("Canada", "China",
                                 "Cuba", "India", "Philippines", "United-States") ~ native_country,
           TRUE ~ "Others"
         ),
```

```
        race = str_replace_all(race, " ", "")) %>%
  select(-starts_with("capital"), -education) %>% # Removing variable education, capital_gain and capit
  distinct(id, .keep_all = T) %>%  # Removing all duplicates using the ID
  select(-id)
```

Summary table for numeric variables after cleaning

```
summary(df_cleaned %>% select_if(is.numeric)) %>% kable()
```

age	education_num	hours_per_week
Min. :22.00	Min. : 1.000	Min. : 1.00
1st Qu.:30.00	1st Qu.: 9.000	1st Qu.:38.00
Median :38.00	Median : 9.000	Median :40.00
Mean :40.13	Mean : 9.636	Mean :39.74
3rd Qu.:49.00	3rd Qu.:10.000	3rd Qu.:43.00
Max. :90.00	Max. :16.000	Max. :99.00
NA's :2807	NA	NA's :244

Cross tabs of categorical variables with dependent variable *over_50K* after cleaning

```
df_cleaned %>%
  select_if(is.character) %>%
  map(.x = names(select(., -over_50k)),
      .f = ~tabyl(df_cleaned, !!sym(.x), over_50k) %>%
        adorn_percentages() %>%
        adorn_pct_formatting() %>%
        adorn_ns() %>%
        kable())
```

[[1]]

workclass	<=50K	>50K
Federal-gov	73.0% (537)	27.0% (199)
Local-gov	83.1% (1335)	16.9% (272)
Others	93.7% (1496)	6.3% (101)
Private	86.3% (15967)	13.7% (2542)
Self-emp-inc	61.0% (434)	39.0% (277)
Self-emp-not-inc	81.8% (1624)	18.2% (362)
State-gov	87.6% (848)	12.4% (120)

[[2]]

marital_status	<=50K	>50K
Divorced	94.3% (3621)	5.7% (219)
Married-civ-spouse	68.6% (7436)	31.4% (3399)
Never-married	98.2% (9158)	1.8% (168)
Others	95.5% (361)	4.5% (17)

marital_status	<=50K	>50K
Sep or widowed	96.0% (1665)	4.0% (70)

[[3]]

occupation	<=50K	>50K
Adm-clerical	89.7% (2940)	10.3% (337)
Craft-repair	78.8% (2863)	21.2% (768)
Exec-managerial	71.6% (1883)	28.4% (746)
Low Salary Jobs	94.7% (6395)	5.3% (361)
Machine-op-inspct	88.9% (1584)	11.1% (197)
Prof-specialty	84.3% (2063)	15.7% (385)
Protective-serv	72.1% (387)	27.9% (150)
Sales	83.3% (2407)	16.7% (481)
Tech-support	76.5% (574)	23.5% (176)
Transport-moving	80.8% (1145)	19.2% (272)

[[4]]

relationship	<=50K	>50K
Not-in-family	95.3% (6694)	4.7% (331)
Unmarried	96.5% (2912)	3.5% (105)
with spouse	68.3% (7271)	31.7% (3376)
Without spouse	98.9% (5364)	1.1% (61)

[[5]]

race	<=50K	>50K
Amer-Indian-Eskimo	88.2% (254)	11.8% (34)
Asian-Pac-Islander	73.2% (681)	26.8% (249)
Black	87.7% (2469)	12.3% (347)
Other	91.1% (226)	8.9% (22)
White	85.2% (18611)	14.8% (3221)

[[6]]

sex	<=50K	>50K
Female	93.7% (8608)	6.3% (576)
Male	80.5% (13633)	19.5% (3297)

[[7]]

native_country	<=50K	>50K
Canada	78.4% (76)	21.6% (21)
China	73.1% (49)	26.9% (18)
Cuba	83.8% (67)	16.2% (13)

native_country	<=50K	>50K
India	62.1% (54)	37.9% (33)
Others	89.0% (2109)	11.0% (261)
Philippines	67.4% (116)	32.6% (56)
United-States	85.1% (19770)	14.9% (3471)

The data looks good. New data contains 26114 rows. It's ready for training model using Random Forest.