Data Cleaning Assessment

Robert Kraemer
COMP 4448

Goal

- Prepare and explore US Census data for later use in identifying characteristics associated with a person making more or less than $50,000 per year

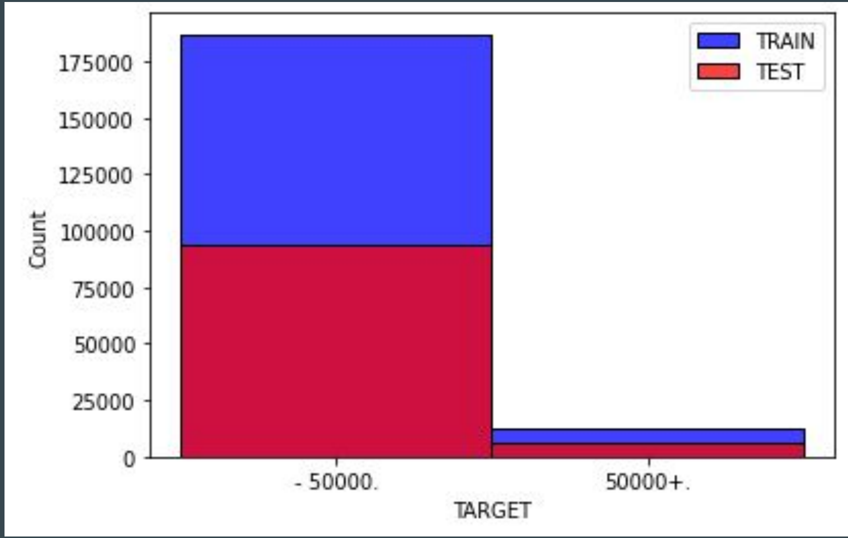
AAGE	ACLSWKR	ADTIND	ADTOCC	AHGA	AHRSPAY	AHSCOL	AMARITL	AMJIND	AMJOCC	...	PEFNTVTY	PEMNTVTY	PENATVTY
47	Not in universe	0	0	High school graduate	0	Not in universe	Married-civilian spouse present	Not in universe or children	Not in universe	...	United-States	United-States	United-States
25	State government	47	28	High school graduate	0	Not in universe	Married-civilian spouse present	Public administration	Protective services	...	United-States	United-States	United-States
9	Not in universe	0	0	Children	0	Not in universe	Never married	Not in universe or children	Not in universe	...	Cuba	Mexico	United-States
6	Not in universe	0	0	Children	0	Not in universe	Never married	Not in universe or children	Not in universe	...	United-States	United-States	United-States
23	Not in universe	0	0	Bachelors degree(BA AB BS)	0	Not in universe	Never married	Not in universe or children	Not in universe	...	United-States	United-States	United-States
42	Private	29	2	High school graduate	2000	Not in universe	Married-civilian spouse present	Transportation	Executive admin and managerial	...	United-States	United-States	United-States
8	Not in universe	0	0	Children	0	Not in universe	Never married	Not in universe or children	Not in universe	...	United-States	?	United-States
68	Private	33	35	Some college but no degree	450	Not in universe	Widowed	Retail trade	Precision production craft & repair	...	United-States	United-States	United-States

PRCITSHP	SEOTR	VETQVA	VETYN	WKSWORK	YEAR	TARGET
Native-Born in the United States	0	Not in universe	2	0	95	- 50000.
Native-Born in the United States	0	Not in universe	2	52	94	- 50000.
Native-Born in the United States	0	Not in universe	0	0	95	- 50000.
Native-Born in the United States	0	Not in universe	0	0	95	- 50000.
Native-Born in the United States	2	Not in universe	2	16	95	- 50000.
Native-Born in the United States	0	Not in universe	2	52	94	50000+.
Native-Born in the United States	0	Not in universe	0	0	95	- 50000.
Native-Born in the United States	2	Not in universe	2	25	95	- 50000.

Description of the Data

- Input Variables:
  - Continuous
    - Age, Wage per hour, Capital gains, Capital losses, Dividends from stocks, Num persons worked for employer, Weeks worked in year
  - Nominal
    - Class of worker, Detailed industry recode, Detailed occupation recode, Education, Enroll in edu inst last wk, Marital stat, Major industry code, Major Occupation code, Race, Hispanic origin, Sex, Member of labor union, Reason for unemployment, Full or part time employment stat, Tax filer stat, Region of previous residence, State of previous residence, Household and family stat, Household summary in household, Migration code-change in msa, Migration code-change in reg, Migration cod-move within reg, Live in this house 1 year ago, Migration prev res in sunbelt, Family members under 18, Country of birth father, Country of birth mother, Country of birth self, Citizenship, Own business or self employed, Fill inc questionnaire for veteran's admin, Veterans benefits, year

- Output Variable:
  - Target
    - \- 50,000.
    - 50,000+.



Data Cleaning

- Process
  - Clean text values, normalize the text; remove white spaces
  - Inspect unique values to all variables to ensure values are not in the data
  - Convert missing values to np.NaN
  - Make certain variable types are set correctly for each variable
  - Handle missing values and drop variables with that are missing more than 30% of the data
  - Drop/ignore instance weight, as indicated in census_income_metadata.txt

```
def convert_missing_values(df):
    identifiers = ['?','NA','nan','Do not know','Not in universe','Not identifiable',
                   'Not in universe or children','Not in universe under 1 year old']
    df.replace(to_replace=identifiers, value=np.NaN, inplace=True)

def cols_to_int(df):
    cols = ['AAGE','AHRSPAY','CAPGAIN','CAPLOSS','DIVVAL','NOEMP','WKSWORK']
    df[cols] = df[cols].astype('int')

def cols_to_category(df):
    cols = ['ACLSWKR','ADTIND','ADTOCC','AHGA','AHSCOL','AMARITL','AMJIND','AMJOCC','ARACE','AREORGN',
            'ASEX','AUNMEM','AUNTYPE','AWKSTAT','FILESTAT','GRINREG','GRINST','HHDFMX','HHDREL','MIGMTR1',
            'MIGMTR3','MIGMTR4','MIGSAME','MIGSUN','PARENT','PEFNTVTY','PEMNTVTY','PENATVTY','PRCITSHP','SEOTR',
            'VETQVA', 'VETYN', 'TARGET']
    df[cols] = df[cols].astype('category')

def ignore_marsupwt(df):
    try:
        cols = ['MARSUPWT']
        df.drop(cols, inplace=True, axis=1)
    except:
        print('MARSUPWT has already been dropped')

def drop_variables_missing_gte_30(df_list):
    try:
        cols = df_list[0].columns[df_list[0].isna().sum()>.3*len(df_list[0])]
        for df in df_list:
            df.drop(cols, inplace=True, axis=1)
    except:
        print('Columns have already been dropped')

def clean_data(dataframe_list):
    for dataframe in dataframe_list:
        convert_missing_values(dataframe)
        cols_to_int(dataframe)
        cols_to_category(dataframe)
        ignore_marsupwt(dataframe)
    drop_variables_missing_gte_30(dataframe_list)
```

Data Cleaning

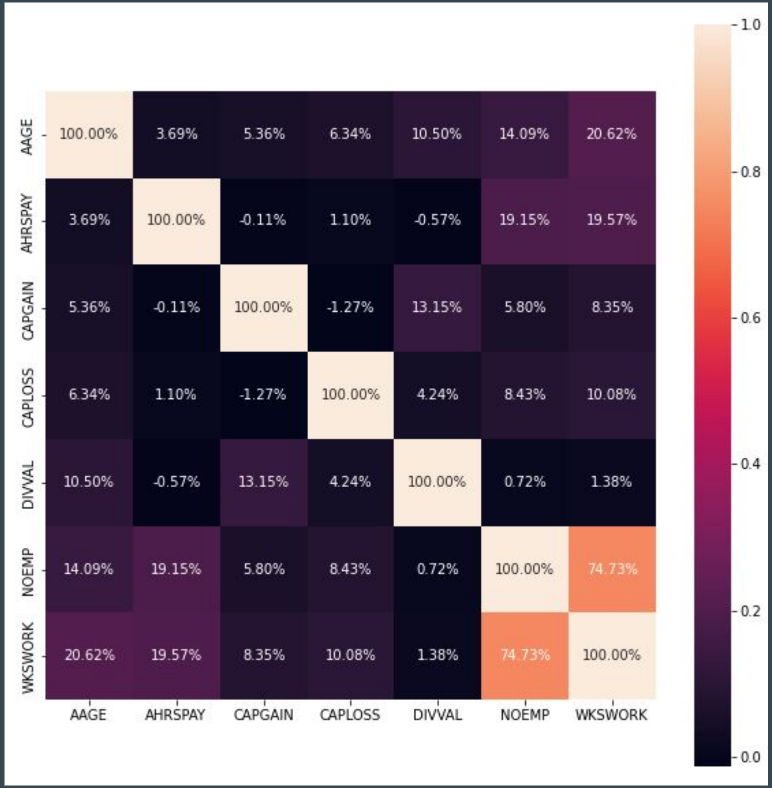
- Summary
  - Missing Values (converted to np.NaN):
    - '?', 'NA','nan', 'Do not know', 'Not in universe', 'Not identifiable', 'Not in universe or children', 'Not in universe under 1 year old'
  - 25 input variables are missing less than 30% of the data
  - 28% continuous, 72% nominal

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 199523 entries, 0 to 199522
Data columns (total 42 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   AAGE      199523 non-null  int64
 1   ACLSWKR   199523 non-null  object
 2   ADTIND    199523 non-null  int64
 3   ADTOCC    199523 non-null  int64
 4   AHGA      199523 non-null  object
 5   AHRSPAY   199523 non-null  int64
 6   AHSCOL    199523 non-null  object
 7   AMARITL   199523 non-null  object
 8   AMJIND    199523 non-null  object
 9   AMJOCC    199523 non-null  object
 10  ARACE     199523 non-null  object
 11  AREORGN   198649 non-null  object
 12  ASEX      199523 non-null  object
 13  AUNMEM    199523 non-null  object
 14  AUNTYPE   199523 non-null  object
 15  AWKSTAT   199523 non-null  object
 16  CAPGAIN   199523 non-null  int64
 17  CAPLOSS   199523 non-null  int64
 18  DIVVAL    199523 non-null  int64
 19  FILESTAT  199523 non-null  object
 20  GRINREG   199523 non-null  object
 21  GRINST    199523 non-null  object
 22  HHDFMX    199523 non-null  object
 23  HHDREL    199523 non-null  object
 24  MARSUPWT  199523 non-null  float64
 25  MIGMTR1   199523 non-null  object
 26  MIGMTR3   199523 non-null  object
 27  MIGMTR4   199523 non-null  object
 28  MIGSAME   199523 non-null  object
 29  MIGSUN    199523 non-null  object
 30  NOEMP     199523 non-null  int64
 31  PARENT    199523 non-null  object
 32  PEFNTVTY  199523 non-null  object
 33  PEMNTVTY  199523 non-null  object
 34  PENATVTY  199523 non-null  object
 35  PRCITSHP  199523 non-null  object
 36  SEOTR     199523 non-null  int64
 37  VETQVA    199523 non-null  object
 38  VETYN     199523 non-null  int64
 39  WKSWORK   199523 non-null  int64
 40  YEAR      199523 non-null  int64
 41  TARGET    199523 non-null  object
dtypes: float64(1), int64(12), object(29)
memory usage: 63.9+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 199523 entries, 0 to 199522
Data columns (total 26 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   AAGE      199523 non-null  int64
 1   ADTIND    199523 non-null  category
 2   ADTOCC    199523 non-null  category
 3   AHGA      199523 non-null  category
 4   AHRSPAY   199523 non-null  int64
 5   AMARITL   199523 non-null  category
 6   ARACE     199523 non-null  category
 7   AREORGN   198343 non-null  category
 8   ASEX      199523 non-null  category
 9   AWKSTAT   199523 non-null  category
 10  CAPGAIN   199523 non-null  int64
 11  CAPLOSS   199523 non-null  int64
 12  DIVVAL    199523 non-null  int64
 13  FILESTAT  199523 non-null  category
 14  HHDFMX    199523 non-null  category
 15  HHDREL    199523 non-null  category
 16  NOEMP     199523 non-null  int64
 17  PEFNTVTY  192810 non-null  category
 18  PEMNTVTY  193404 non-null  category
 19  PENATVTY  196130 non-null  category
 20  PRCITSHP  199523 non-null  category
 21  SEOTR     199523 non-null  category
 22  VETYN     199523 non-null  category
 23  WKSWORK   199523 non-null  int64
 24  YEAR      199523 non-null  category
 25  TARGET    199523 non-null  category
dtypes: category(19), int64(7)
memory usage: 14.3 MB
```

Exploratory Analysis

- Observe the distribution of the output (TARGET) variable (slide 3)
- Descriptive statistics for numerical values (slide 7)
- Observe correlation between continuous variables
- Explore distributions of continuous variables of TRAIN and TEST datasets



Exploratory Analysis

- We see that the age distribution looks fairly normal, with some outliers, while several of the other variables are trending towards two peaks.

	AAGE	AHRSPAY	CAPGAIN	CAPLOSS	DIVVAL	NOEMP	WKSWORK
mean	34.494199	55.426908	4.347190e+02	37.313788	1.975295e+02	1.956180	23.174897
median	33.000000	0.000000	0.000000e+00	0.000000	0.000000e+00	1.000000	8.000000
min	0.000000	0.000000	0.000000e+00	0.000000	0.000000e+00	0.000000	0.000000
max	90.000000	9999.000000	9.999900e+04	4608.000000	9.999900e+04	6.000000	52.000000
var	497.776045	75568.060368	2.206680e+07	73927.667758	3.936905e+06	5.593819	595.920755
std	22.310895	274.896454	4.697531e+03	271.896428	1.984164e+03	2.365126	24.411488
skew	0.373290	8.935097	1.899082e+01	7.632565	2.778650e+01	0.751561	0.210169
kurtosis	-0.732824	NaN	NaN	NaN	NaN	NaN	NaN