Project Survey II - Predicting Unemployment

Name: Ivan Ivanov Ivanov
Matrikelnummer: 3000000

Introduction
Pre processing the data
Descriptive Analysis
Models: Predicting Unemployment with Survey Data
Conclusions

Introduction

In this project I will be analyzing the data set of a survey conducted between 1984 - 1985 in Bangladesh. The survey was carried out by the Institute of Development Studies Bangladesh (IDSB) and the data set was made available on the website of the Harvard Dataverse. The survey was conducted in 1984 and 1985 and the data set contains information about the respondents' demographic and socio-economic characteristics, their employment status, and their attitudes towards various issues related to gender and development. The data set also includes information on the respondents' access to various resources, such as land, education, and healthcare, and their participation in various activities, such as agriculture, trade, and social networks. The data set is a valuable resource for researchers and policymakers interested in understanding the socio-economic dynamics of rural Bangladesh and the factors that influence gender equality and development in the region.

Pre processing the data

First I'm loading the data into the notebook. For this I'm using the read_dataset function of pandas.

```
df = pd.read_dataset('survey_data', '.')
df.head(5)
df.info()
df.describe()
```

```
   year  ...  wmarr
0  1984  ...    NaN
1  1984  ...    NaN
2  1984  ...    NaN
3  1984  ...    NaN
4  1984  ...    NaN

[5 rows x 1363]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5503 entries, 0 to 5502
Columns: 1363 entries, year to wmarr
dtypes: int64(1363)
memory usage: 57.2 MB
            year  ...  ...
count  5503.0000  ...  ...
mean   1984.5000  ...  ...
std       0.5000  ...  ...
min    1984.0000  ...  ...
25%    1984.0000  ...  ...
50%    1984.5000  ...  ...
75%    1985.0000  ...  ...
max    1985.0000  ...  ...

[8 rows x 1363]
```

Since I'm using the pandas library I can use the .info() and .describe() functions to get a first impression of the data. Since the data set has 1363 columns, I'm only printing the first 5 rows of the data set.

```
df.shape
```

```
(5503, 1363)
```

```
df.isnull().sum()
```

```
year       0
hhid       0
hhmem      0
hhsex      0
hhage      0
          ...
hhwt       0
hhsize     0
hhsex2     0
hhage2     0
wmarr      0
dtype: int64
```

```
df.dtypes
```

```
year       int64
hhid       int64
hhmem      int64
hhsex      int64
hhage      int64
           ...
hhwt       int64
hhsize     int64
hhsex2     int64
hhage2     int64
wmarr      int64
dtype: int64
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1)
```

```
      ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
0     ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
1     ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
2     ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
3     ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
4     ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
...   ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
5498  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
5499  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
5500  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
5501  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
5502  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...

[5503 rows x 1353]
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).isnull().sum()
```

```
year       0
hhid       0
hhmem      0
hhsex      0
hhage      0
          ...
hhwt       0
hhsize     0
hhsex2     0
hhage2     0
wmarr      0
dtype: int64
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).dtypes
```

```
year       int64
hhid       int64
hhmem      int64
hhsex      int64
hhage      int64
           ...
hhwt       int64
hhsize     int64
hhsex2     int64
hhage2     int64
wmarr      int64
dtype: int64
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).shape
```

```
(5503, 1353)
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5503 entries, 0 to 5502
Columns: 1353 entries, ... to ...
dtypes: int64(1353)
memory usage: 57.1 MB
None
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).describe()
```

```
       ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
count  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
mean   ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
std    ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
min    ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
25%    ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
50%    ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
75%    ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
max    ...  ...  ...  ...  ...  ...  ...  ...  ...  ...

[8 rows x 1353]
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).head(5)
```

```
   ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
0  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
1  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
2  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
3  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...
4  ...  ...  ...  ...  ...  ...  ...  ...  ...  ...

[5 rows x 1353]
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).isnull.sum()
```

```
AttributeError: 'DataFrame' object has no attribute 'isnull'
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).isnull().sum()
```

```
...    0
...    0
...    0
...    0
...    0
      ...
...    0
...    0
...    0
...    0
...    0
dtype: int64
```

```
df.drop(['year', 'hhid', 'hhmem', 'hhsex', 'hhsex2', 'hhage', 'hhage2', 'hhwt', 'hhsize', 'wmarr'], axis=1).dtypes
```

```
...    int64
...    int64
...    int64
...    int64
...    int64
       ...
...    int64
...    int64
...    int64
...    int64
...    
```