Final Project – EDA & Preliminary Results

Claire Yang, Saman De Silva, & Hans Elasri

2025-11-24

```
# loading in cleaned data
dat <- read_csv("data/cleaned-data.csv")
```

```
## New names:
## Rows: 23041 Columns: 28
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (28): ...1, person, family, community, county, province, age, gender, ur...
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...1‘
```

```
glimpse(dat)
```

```
## Rows: 23,041
## Columns: 28
## $ ...1                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ person               <dbl> 100051501, 100051502, 110043107, 110147102, 10016060~
## $ family               <dbl> 100051, 100051, 100051, 100125, 100160, 100160, 1004~
## $ community            <dbl> 624942, 624942, 624942, -9, 800637, 800637, -9, -9, ~
## $ county               <dbl> 45, 45, 45, 3622, 189, 189, 52, 52, 52, 363, 363, 48~
## $ province             <dbl> 11, 11, 11, 44, 12, 12, 13, 13, 13, 13, 13, 13, 13, ~
## $ age                  <dbl> 53, 56, 28, 40, 33, 31, 35, 10, 33, 32, 34, 34, 34, ~
## $ gender               <dbl> 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1~
## $ urban                <dbl> 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1~
## $ vet                  <dbl> -8, -8, -8, -8, -8, -8, -8, -8, -8, -8, -8, -8, -8, ~
## $ edu_yrs              <dbl> 12, 12, 15, -9, 16, 16, 9, 3, -9, 16, 16, 12, 15, 6,~
## $ lang                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ read_books           <dbl> 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1~
## $ hours_tv             <dbl> 3.0, 3.0, 2.0, 5.0, 4.0, 5.0, 7.0, 7.0, 14.0, 3.5, 1~
## $ avg_sleep            <dbl> -8.0, 8.0, 7.0, 7.0, -8.0, 8.0, 8.5, 9.5, 8.0, -8.0,~
## $ smoked               <dbl> 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1~
## $ coverage          <dbl> 2, 2, 2, 4, 2, 2, 5, -8, 5, 2, 2, 5, 6, 5, 5, 5, 78,~
## $ govt_rating       <dbl> 2, 1, 3, 4, 2, 3, 4, -8, 2, 2, 2, 2, 2, 2, 3, 3, 2, ~
## $ welfare_housing   <dbl> -8, -8, -8, -8, -8, -8, -8, -8, -8, -8, -8, -8, -8, ~
## $ ‘self-health‘     <dbl> 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 2, 3, 3, 5, 2, 3, 2~
## $ wechat            <dbl> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1~
## $ ‘2020_employment‘ <dbl> -8, -8, -8, -8, -8, -8, -8, -8, 40201, -8, -8, -8, -~
## $ reason_unemployed <dbl> -8, -8, -8, -8, -8, -8, -8, -8, 2, -8, -8, -8, 2, -8~
## $ children          <dbl> 1, 1, 1, 1, 1, 1, 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 2~
## $ married           <dbl> 2, 2, 1, 1, 2, 2, 2, -8, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ income_satis      <dbl> 4, 4, 4, 3, 4, 4, 2, -8, -8, 4, 4, 4, 4, 5, 2, 2, -8~
## $ income_rel        <dbl> 1, 4, 3, 3, 3, 3, 2, 79, 2, 3, 3, 3, 3, 3, 3, 3, 3, ~
## $ status_rel        <dbl> 1, 5, 3, 3, 2, 3, 3, 3, 2, 3, 3, 2, 3, 4, 3, 3, 3, 3~
```

```r
head(dat)
```

```
## # A tibble: 6 x 28
##    ...1 person family community county province   age gender urban   vet edu_yrs
##   <dbl>  <dbl>  <dbl>     <dbl>  <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     1 1.00e8 100051    624942     45       11    53      0     1    -8      12
## 2     2 1.00e8 100051    624942     45       11    56      1     1    -8      12
## 3     3 1.10e8 100051    624942     45       11    28      1     1    -8      15
## 4     4 1.10e8 100125        -9   3622       44    40      0     1    -8      -9
## 5     5 1.00e8 100160    800637    189       12    33      1     1    -8      16
## 6     6 1.20e8 100160    800637    189       12    31      0     1    -8      16
## # i 17 more variables: lang <dbl>, read_books <dbl>, hours_tv <dbl>,
## #   avg_sleep <dbl>, smoked <dbl>, coverage <dbl>, govt_rating <dbl>,
## #   welfare_housing <dbl>, ‘self-health‘ <dbl>, wechat <dbl>,
## #   ‘2020_employment‘ <dbl>, reason_unemployed <dbl>, children <dbl>,
## #   married <dbl>, income_satis <dbl>, income_rel <dbl>, status_rel <dbl>
```

1. Background & Primary Research Questions

asdf asdf asdf

2. Data Sources & Structure

asdf asdf asdf

3. **EDA Plots**

4. **Variance Decomposition**

5. **Data Trends**

6. **Mathmatical Model**

7. **Initial Findings**

8. **Next Steps**