COMPAS Data Wrangling and Analysis

Dasha Asienga

2024-02-23

The thesis body will have more in-depth descriptions of the data analysis as well as select output and results from this file. This file is intended for general preliminary analysis of the COMPAS data set.

Reading in the Data

```
#read in the data
compas_path <- "/home/dasienga24/Statistics-Senior-Honors-Thesis/Data Sets/COMPAS/compas_data.csv"
compasdata <- read.csv(compas_path)
```

The Data Set

The COMPAS data set has 12076 observations of participants that were evaluated for the risk of recidivism by the COMPAS tool. There are 29 variables of interest as described below:

- `id`: unique person identifier.
- `compas_person_id`: unique COMPAS case identifier.
- `name`: full name.
- `first`: first name.
- `last`: last name.
- `sex`: sex categorized as male or female.
- `race`: race categorized as African-American, Asian, Caucasian, Hispanic, Native American, or Other.
- `age`: numeric age, ranging from 18 to 96.
- `age_cat`: age categorized as Less than 25, 25 - 45, or Greater than 45.
- `marital_status`: marital status categorized as Single, Significant Other, Married, Widowed, Separated, Divorced, or Unknown.
- `custody_status`: custody status categorized as Jail Inmate, Prison Inmate, Pretrial Defendant, Parole, Residential Program, or Probation.
- `juv_fel_count`: number of prior juvenile felonies, ranging from 0 to 20.
- `juv_misd_count`: number of prior juvenile misdemeanors, ranging from 0 to 13.
- `juv_other_count`: number of other prior juvenile offenses, ranging from 0 to 17.
- `priors_count`: number of non-juvenile prior offenses, ranging from 0 to 43.
- `days_b_screening_arrest`: number of days between COMPAS screening and arrest.
- `c_days_from_compas`: the number of days since COMPAS screening.
- `c_charge_degree`: the charge degree according to the appropriate laws.
- `c_charge_desc`: the charge description in words.
- `type_of_assessment`: the type of assessment, in this case, the assessment is 'Risk of Recidivism'.
- `raw_score`: COMPAS tool raw score on risk of recidivism.
- `decile_score`: decile rank on a scale of 1 - 10 based on the COMPAS raw score.
- `score_text`: COMPAS risk of recidivism based on the decile scores and categorized as High, Medium, or Low.

- `is_violent_recid`: categorical variable recording whether a defendant was accused of a violent crime within 2 years (0 = N, 1 = Y).
- `num_vr_cases`: number of times a defendant was accused of a violent crime within 2 years.
- `is_recid`: categorical variable recording whether a defendant was accused of a crime within 2 years (0 = N, 1 = Y).
- `num_r_cases`: number of times a defendant was accused of a crime within 2 years.
- `days_in_jail`: number of days spent in jail.
- `days_in_prison`: number of days spent in prison.

```r
colnames(compasdata)
```

```
##  [1] "id"                      "compas_person_id"
##  [3] "name"                    "first"
##  [5] "last"                    "sex"
##  [7] "race"                    "age"
##  [9] "age_cat"                 "marital_status"
## [11] "custody_status"          "juv_fel_count"
## [13] "juv_misd_count"          "juv_other_count"
## [15] "priors_count"            "days_b_screening_arrest"
## [17] "c_days_from_compas"      "c_charge_degree"
## [19] "c_charge_desc"           "type_of_assessment"
## [21] "raw_score"               "decile_score"
## [23] "score_text"              "is_violent_recid"
## [25] "num_vr_cases"            "is_recid"
## [27] "num_r_cases"             "days_in_jail"
## [29] "days_in_prison"
```

Data Wrangling

Before proceeding with the data analysis, we first need to handle some data anomalies. We'll also only consider COMPAS cases within 30 days of arrest to improve the data quality. This resulted in 9638 total observations.

```r
compasdata <- compasdata %>%
  filter(decile_score > 0 & is_recid != -1 & days_b_screening_arrest >= -30 &
           days_b_screening_arrest <= 30) %>%
  mutate(days_b_screening_arrest = abs(days_b_screening_arrest))

count(compasdata)
```

```
##      n
## 1 9638
```

Next, let's also make sure that there are no duplicate participants.

```r
clean_compasdata <- compasdata[-which(duplicated(compasdata$id)), ]
```

We'll proceed with this data set and 9387 observations total.

Descriptive Statistics

Now that the data is clean, let's generate some descriptive statistics to understand the distribution of the variables in the data set and their relationships with each other.

First, below is a glimpse of the data as described above. Notice that there is a lot of missing data for `num_vr_cases` and `num_r_cases` because that information is only recorded for participants that recommit a crime in the next 2 years.

```
glimpse(clean_compasdata)
```

```
## Rows: 9,387
## Columns: 29
## $ id                      <int> 1, 3, 4, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, ~
## $ compas_person_id        <int> 56418, 51601, 38864, 59301, 61330, 56890, 6199~
## $ name                    <chr> "miguel hernandez", "kevon dixon", "ed philo",~
## $ first                   <chr> "miguel", "kevon", "ed", "marsha", "edward", "~
## $ last                    <chr> "hernandez", "dixon", "philo", "miles", "riddl~
## $ sex                     <chr> "Male", "Male", "Male", "Male", "Male", "Male"~
## $ race                    <chr> "Other", "African-American", "African-American~
## $ age                     <int> 69, 34, 24, 44, 41, 43, 39, 20, 26, 27, 23, 37~
## $ age_cat                 <chr> "Greater than 45", "25 - 45", "Less than 25", ~
## $ marital_status          <chr> "Single", "Single", "Single", "Separated", "Si~
## $ custody_status          <chr> "Jail Inmate", "Jail Inmate", "Jail Inmate", "~
## $ juv_fel_count           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ juv_misd_count          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ juv_other_count         <int> 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ priors_count            <int> 0, 0, 4, 0, 14, 3, 0, 0, 0, 0, 3, 0, 0, 0, 1, ~
## $ days_b_screening_arrest <int> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 20, ~
## $ c_days_from_compas      <int> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 490,~
## $ c_charge_degree         <chr> "(F3)", "(F3)", "(F3)", "(M1)", "(F3)", "(F3)"~
## $ c_charge_desc           <chr> "Aggravated Assault w/Firearm", "Felony Batter~
## $ type_of_assessment      <chr> "Risk of Recidivism", "Risk of Recidivism", "R~
## $ raw_score               <dbl> -2.78, -0.76, -0.66, -1.93, -0.16, -0.72, -1.7~
## $ decile_score            <int> 1, 3, 4, 1, 6, 4, 1, 10, 5, 4, 6, 1, 3, 4, 1, ~
## $ score_text              <chr> "Low", "Low", "Low", "Low", "Medium", "Low", "~
## $ is_violent_recid        <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ num_vr_cases            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ is_recid                <int> 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1~
## $ num_r_cases             <int> NA, 3, 1, NA, 3, NA, NA, NA, NA, NA, 1, NA, NA~
## $ days_in_jail            <dbl> 8, 10, 139, 1, 48, 17, 3, 46, 87, 1, 4, 1, 0, ~
## $ days_in_prison          <dbl> 0, 53, 0, 0, 2130, 0, 0, 3948, 0, 0, 0, 0, 0, ~
```

Next, we will perform some univariate analysis for the variables in the data set before proceeding to conduct some bivariate and multivariate analysis.

Univariate Analysis

Univariate analysis will involve looking at some summary statistics and visualizations of the different variables in the data set.

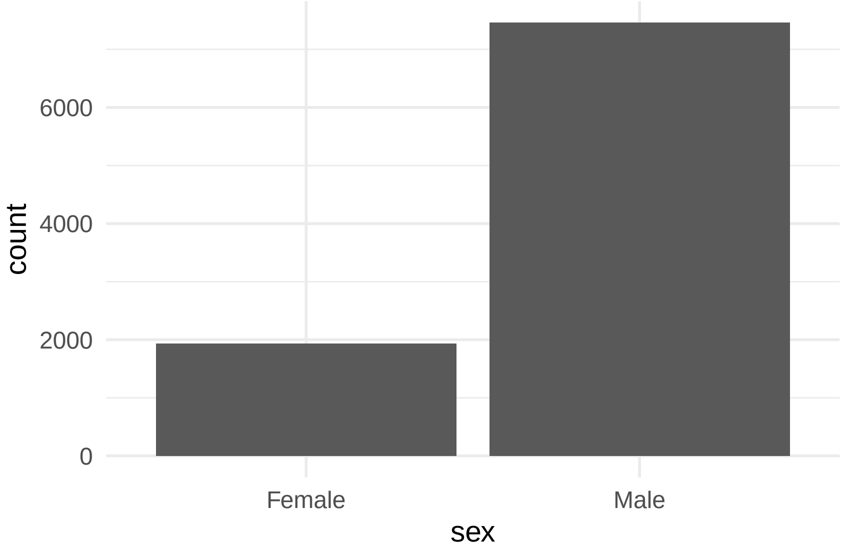
Categorical Variables

There 7457 males and 1930 females in the data set.

```
tally(clean_compasdata$sex)
```

```
## X
## Female   Male
##   1930   7457
```

```
ggplot(data = clean_compasdata, mapping = aes(x = sex)) +
  geom_bar() +
  theme_minimal()
```

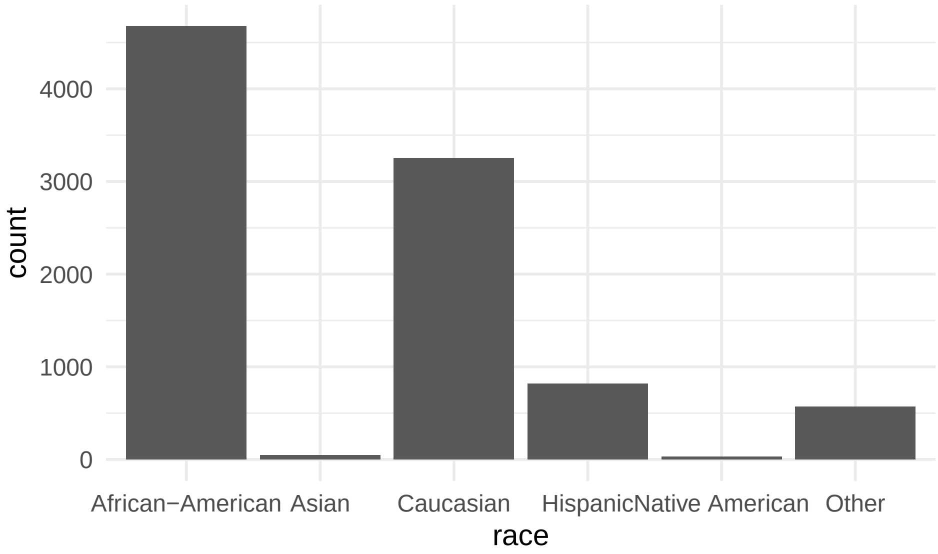


Most of the participants are African-American and Caucasian, with only 27 Native Americans and 48 Asians.

```
tally(clean_compasdata$race)
```

```
## X
## African-American            Asian        Caucasian         Hispanic
##             4674               48             3250              818
##  Native American            Other
##               27              570
```

```
ggplot(data = clean_compasdata, mapping = aes(x = race)) +
  geom_bar() +
  theme_minimal()
```



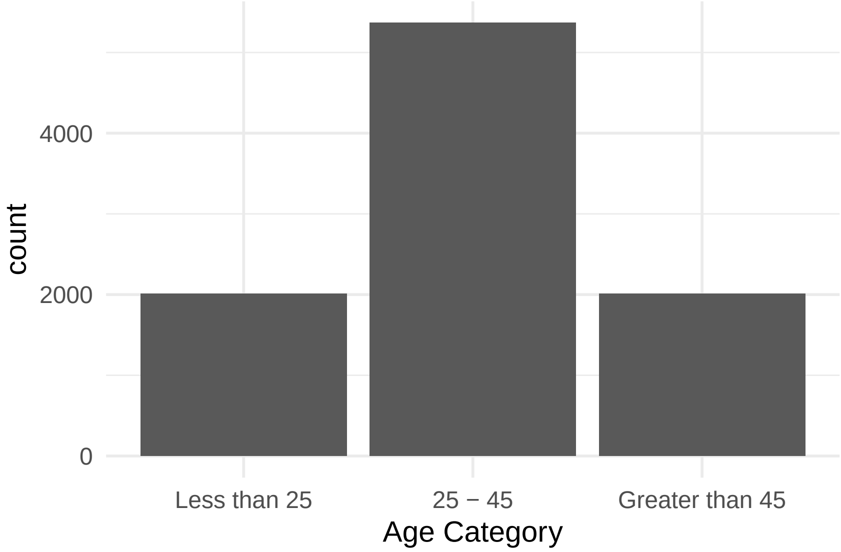
Majority of the participants are between the age of 25 and 45, with about the same number of participants less than 25 and greater than 25.

```
tally(clean_compasdata$age_cat)
```

```
## X
##         25 - 45 Greater than 45    Less than 25
##            5366            2012            2009
```

```
order <- c("Less than 25", "25 - 45", "Greater than 45")

ggplot(data = clean_compasdata, mapping = aes(x = age_cat)) +
  geom_bar() +
  theme_minimal() +
  scale_x_discrete(limits = order) +
  labs(x = "Age Category")
```



Continuous Variables

Bivariate Analysis

Multivariate Analysis

Demographic Group Analysis

Logistic Regression

Seldonian Classification

Results