COMPAS Data Wrangling and Analysis

Dasha Asienga

2024-02-25

Contents

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 defendants 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.

```
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 defendants.

```
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 defendants 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.

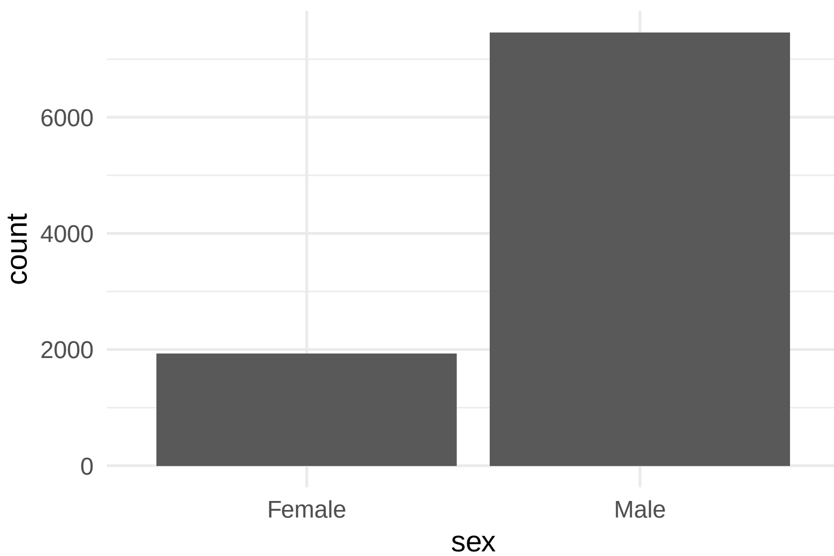
Categorical Variables

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()
```

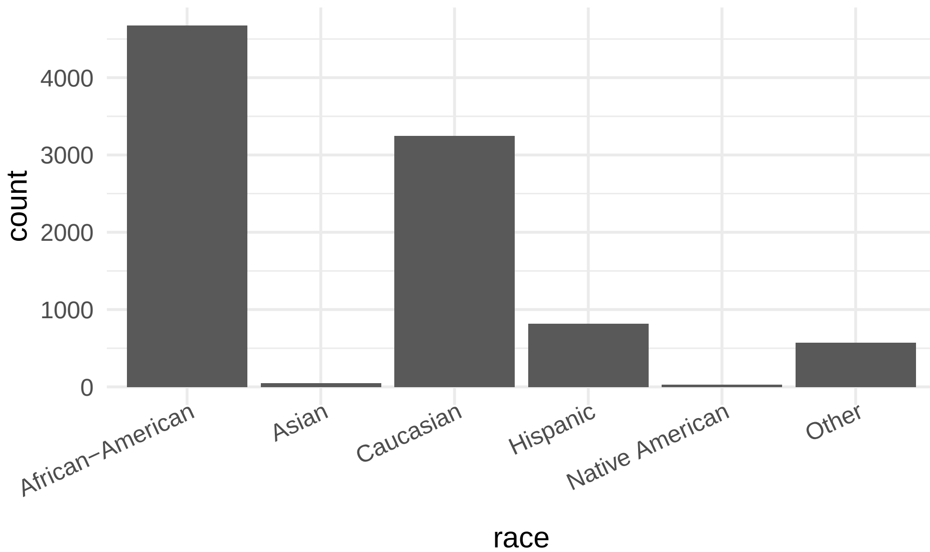


Most of the defendants are African-American and Caucasian, with only 27 Native Americans and 48 Asians.

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

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

```r
ggplot(data = clean_compasdata, mapping = aes(x = race)) +
  geom_bar() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1))
```

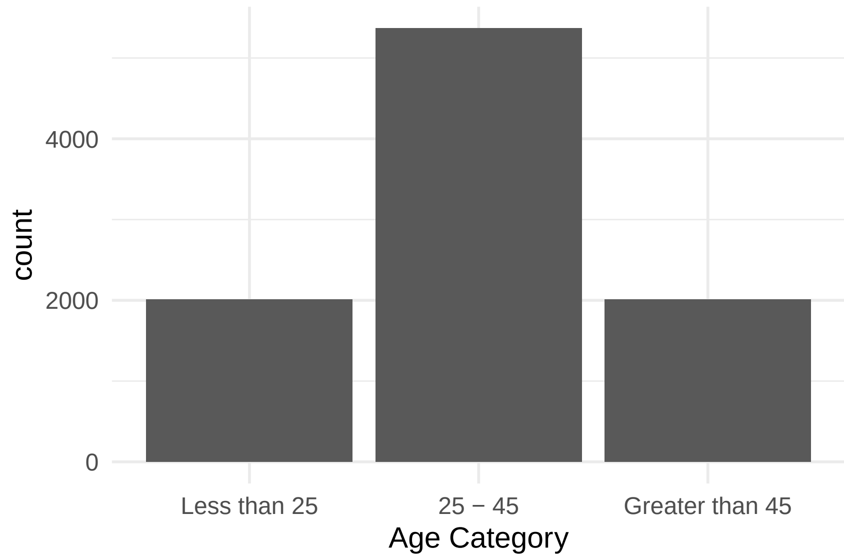


Majority of the defendants are between the age of 25 and 45, with about the same number of defendants less than 25 and greater than 25.

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

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

```r
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")
```

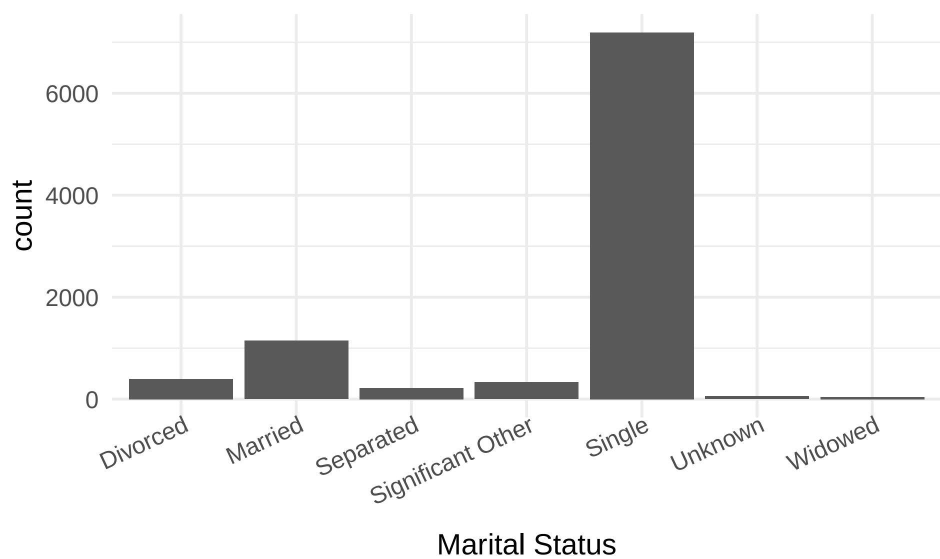


Most of the defendants are single, followed by married.

```
tally(clean_compasdata$marital_status)
```

```
## X
##          Divorced           Married         Separated Significant Other
##               398              1145               219               333
##            Single           Unknown           Widowed
##              7195                57                40
```

```
ggplot(data = clean_compasdata, mapping = aes(x = marital_status)) +
  geom_bar() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1)) +
  labs(x = "Marital Status")
```

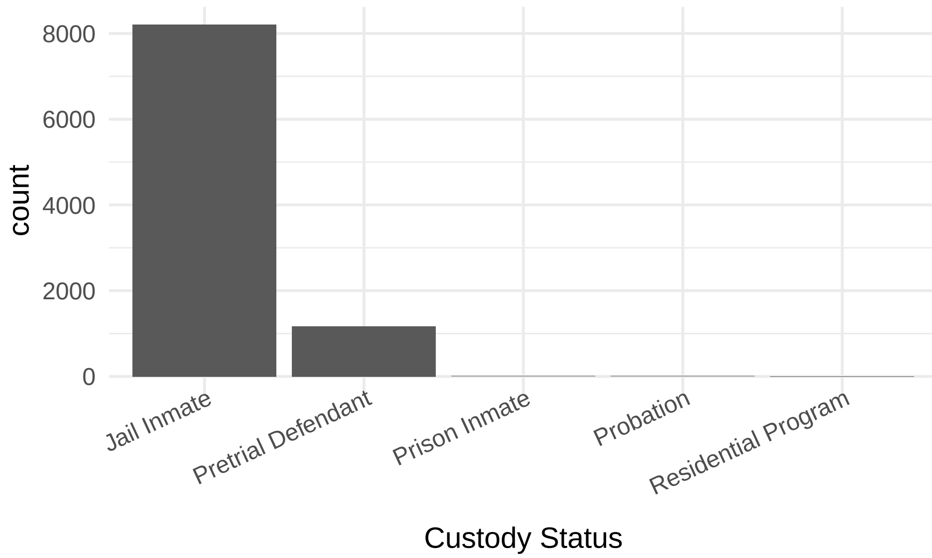


Most of the defendants are jail inmates, with only a handful of prison inmates, probationers, and defendants of the residential program.

```r
tally(clean_compasdata$custody_status)
```

```
## X
##         Jail Inmate  Pretrial Defendant       Prison Inmate           Probation
##                8208                1170                   4                   3
## Residential Program
##                   2
```

```r
ggplot(data = clean_compasdata, mapping = aes(x = custody_status)) +
  geom_bar() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1)) +
  labs(x = "Custody Status")
```



As a data check, all the assessments are for risk of recidivism.

```r
tally(clean_compasdata$type_of_assessment)
```

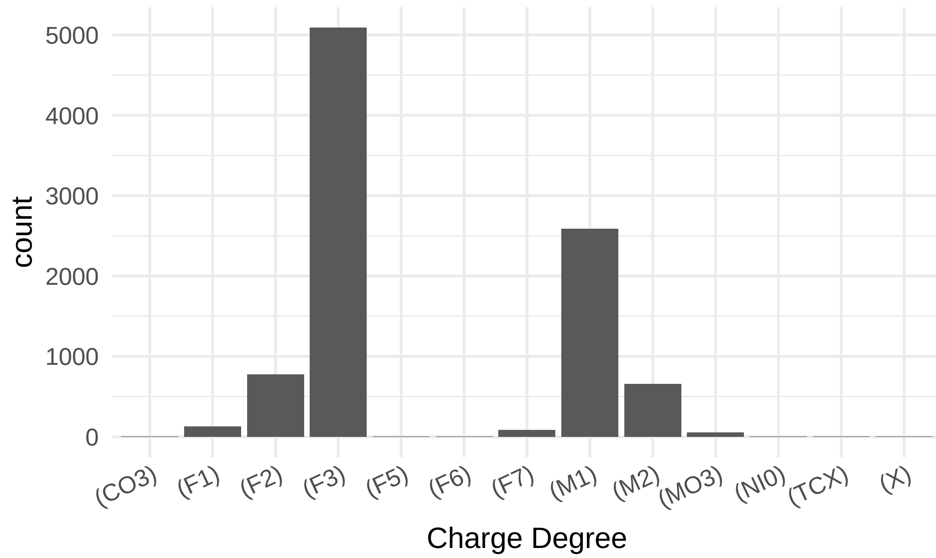
```
## X
## Risk of Recidivism
##               9387
```

There are 13 different charge degrees present in the data set. Most defendants were charged with (F3), which are felonies of the third degree. These are the least serious felonies in Florida and typically include crimes like breaking and entering, collecting and keeping stolen property, fraud, and petty theft. Many other defendants were also charged with (M1), which are a first-degree misdemeanors and can be punished by up to one year in jail. These include simple battery, disorderly conduct, DUI, indecent exposure, marijuana possession, shoplifting, prostitution, and vandalism, among others.

```r
tally(clean_compasdata$c_charge_degree)
```

```
## X
## (CO3)  (F1)  (F2)  (F3)  (F5)  (F6)  (F7)  (M1)  (M2) (MO3) (NI0) (TCX)   (X)
##     1   129   774  5091     5     3    85  2584   658    51     4     1     1
```

```r
ggplot(data = clean_compasdata, mapping = aes(x = c_charge_degree)) +
  geom_bar() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1)) +
  labs(x = "Charge Degree")
```

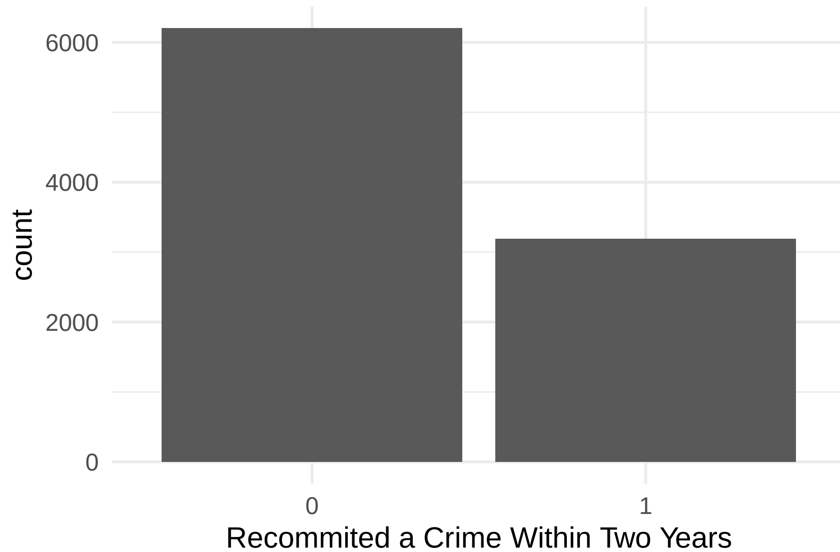


About two-thirds of the defendants did not recommit a crime within two years, while one-thirds did. This is our response variable and is indicative of class imbalance, which can affect the performance of machine learning classification algorithms. This is important to keep in mind when assessing model performance later on.

```r
tally(clean_compasdata$is_recid)
```

```
## X
##    0    1
## 6199 3188
```

```r
ggplot(data = clean_compasdata, mapping = aes(x = as.factor(is_recid))) +
  geom_bar() +
  theme_minimal() +
  labs(x = "Recommited a Crime Within Two Years")
```



Only 745 defendants recommitted a violent crime.

```
tally(clean_compasdata$is_violent_recid)
```

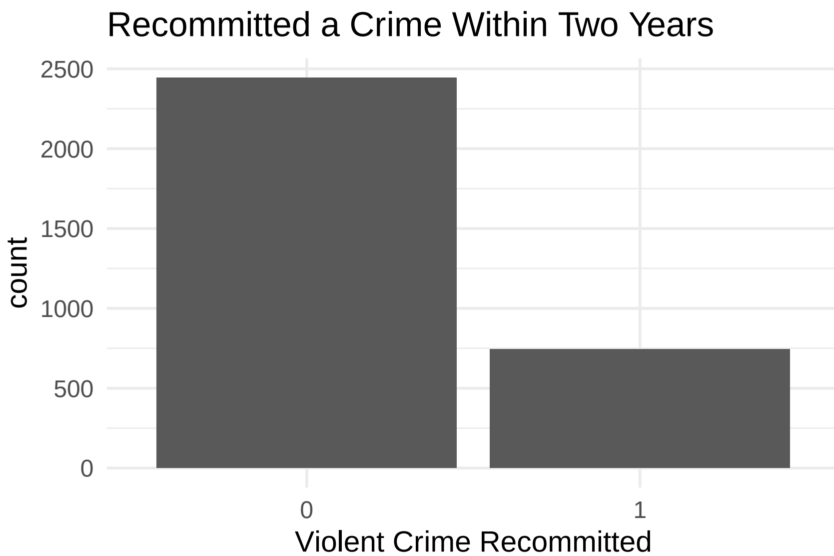
```
## X
##    0    1
## 8642  745
```

Out of the 3188 who recommitted a crime, 2443 re-committed a non-violent crime,

```
tally(clean_compasdata[clean_compasdata$is_recid == 1, ]$is_violent_recid,
      margins = TRUE)
```

```
## X
##     0     1 Total
##  2443   745  3188
```

```
ggplot(data = clean_compasdata[clean_compasdata$is_recid == 1, ],
       mapping = aes(x = as.factor(is_violent_recid))) +
  geom_bar() +
  theme_minimal() +
  labs(x = "Violent Crime Recommitted",
       title = "Recommitted a Crime Within Two Years")
```

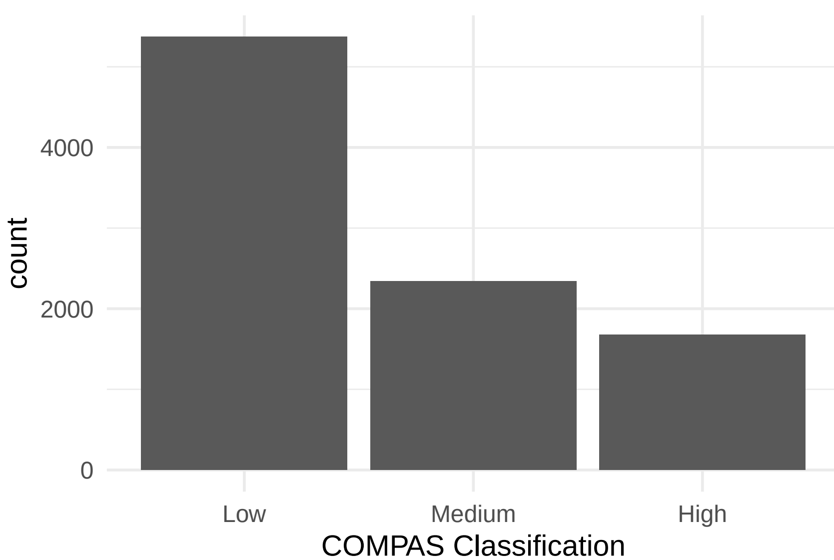


Finally, the COMPAS tool classified more than half of the defendants as low risk. In particular, 5370 were classified as low risk and 1677 as high risk, with the remaining 2340 as medium risk. This is expected since most of the defendants did not recommit a crime within the two year time window.

```
tally(clean_compasdata$score_text)
```

```
## X
##   High    Low Medium
##   1677   5370   2340
```

```
order <- c("Low", "Medium", "High")

ggplot(data = clean_compasdata, mapping = aes(x = score_text)) +
  geom_bar() +
  theme_minimal() +
  scale_x_discrete(limits = order) +
  labs(x = "COMPAS Classification")
```



This wraps up our univariate analysis of the categorical variables. Next, let's examine the univariate distribution of the continuous variables.

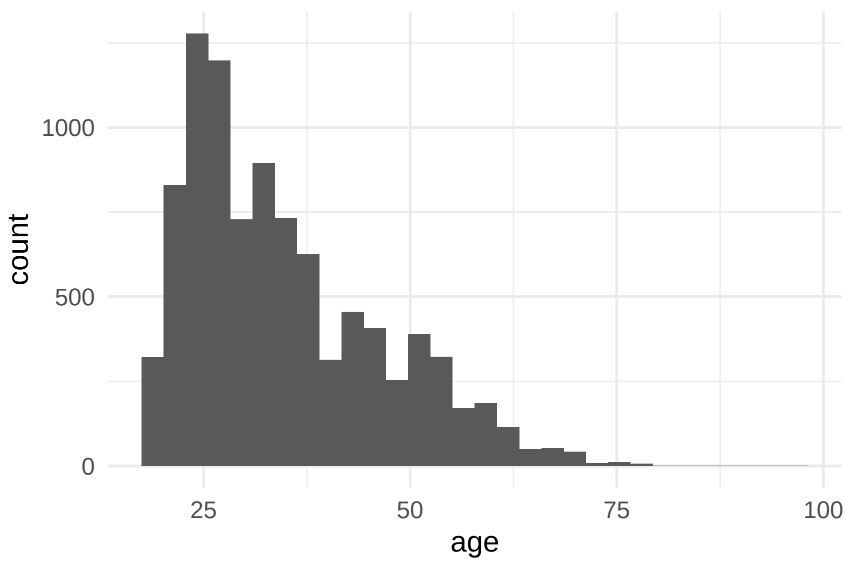
Continuous Variables

The age of the defendants ranges from 18 to 96 with a mean of 34 and a median of 32. There is no missing data. There's a right-skew in the distribution because of the few really old defendants.

```
favstats(clean_compasdata$age)
```

```
##  min Q1 median Q3 max     mean       sd    n missing
##   18 25     32 42  96 34.75413 11.80854 9387       0
```

```
ggplot(data = clean_compasdata, mapping = aes(x = age)) +
  geom_histogram() +
  theme_minimal()
```



Most of the defendants had no juvenile felony accounts. The maximum juvenile felony count is 20. There is not enough variation in this variable.

```
favstats(clean_compasdata$juv_fel_count)
```

```
##  min Q1 median Q3 max       mean        sd    n missing
##    0  0      0  0  20 0.05837861 0.4518127 9387       0
```

Similarly, most defendants had no juvenile misdemeanor counts, which are less serious crimes than felonies. The maximum was 13, but there is not enough variation in this variable.

```
favstats(clean_compasdata$juv_misd_count)
```

```
##  min Q1 median Q3 max      mean        sd    n missing
##    0  0      0  0  13 0.0787259 0.4640061 9387       0
```

Similarly, most defendants had no other juvenile counts, excluding misdemeanors and felonies. The maximum was 11, but there is not enough variation in this variable.

```
favstats(clean_compasdata$juv_other_count)
```

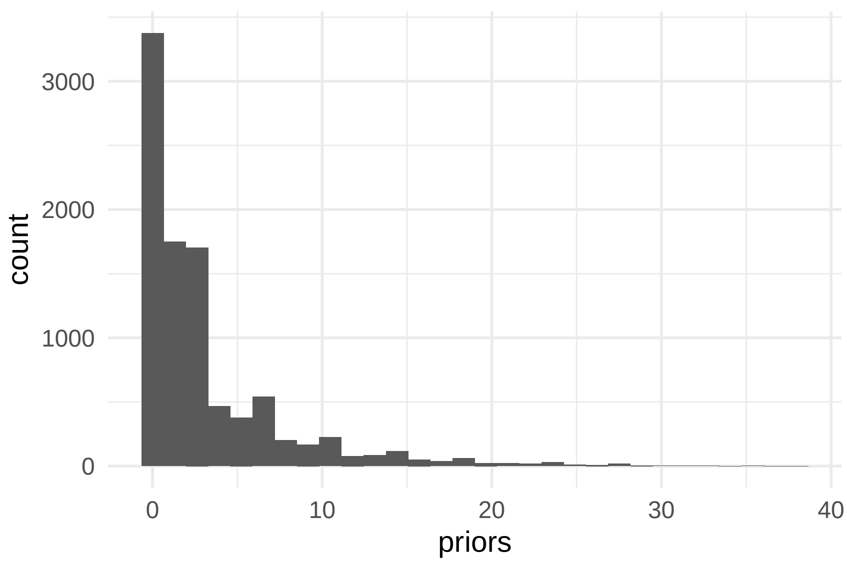
```
##  min Q1 median Q3 max       mean        sd    n missing
##    0  0      0  0  11 0.09917972 0.4683305 9387       0
```

There is slightly more variation in the `priors_count` variable which records the number of non-juvenile prior offenses for each defendant. It ranges from 0 to 38, with a median of 1 and a mean of 3.02, indicating a right skew as visualized in the histogram below. There is no missing data and the standard deviation is 4.586, suggesting that this may be a more informative variable when modeling.

```
favstats(clean_compasdata$priors_count)
```

```
##  min Q1 median Q3 max     mean       sd    n missing
##    0  0      1  4  38 3.023863 4.586441 9387       0
```

```
ggplot(data = clean_compasdata, mapping = aes(x = priors_count)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "priors")
```



The `days_b_screening_arrest` variable indicates how many days passed between arrest and COMPAS screening. It may not be indicative of recidivism, however. We will evaluate this when performing bivariate analysis.

```
favstats(clean_compasdata$days_b_screening_arrest)
```

```
##  min Q1 median Q3 max     mean      sd    n missing
##    0  1      1  1  30 2.140194 4.89312 9387       0
```

The interpretation of this variable is not clear – it seems to indicate the number of days since COMPAS screening to date. We will not include this in the analysis.

```
favstats(clean_compasdata$c_days_from_compas)
```

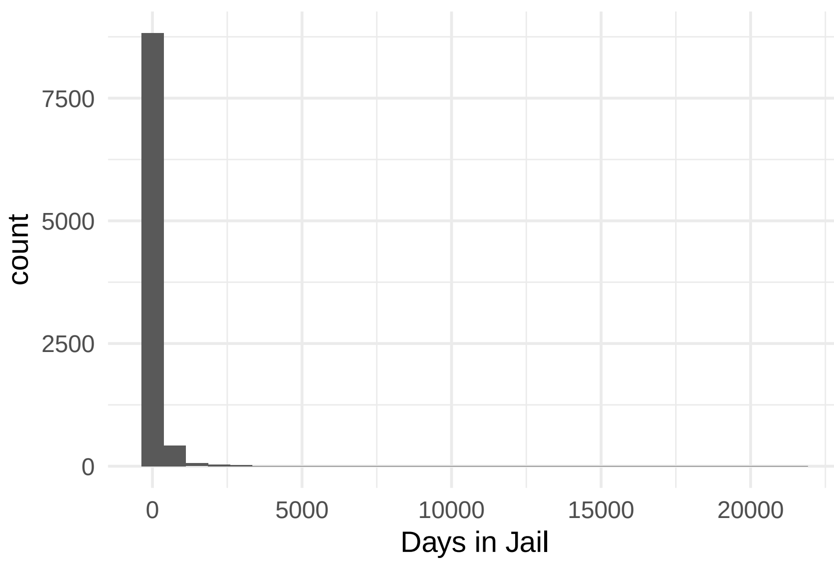
```
##  min Q1 median Q3  max     mean       sd    n missing
##    0  1      1  1 9485 24.92436 263.4065 9387       0
```

The number of days spent in jail ranges from 0 to 21540, with a median of 4 days and a mean of 100 days. This variable is extremely right skewed, as visualized in the histogram. The standard deviation is also 393, indicating a lot of variation that may potentially be useful for predicting the risk of recidivism.

```
favstats(clean_compasdata$days_in_jail)
```

```
##  min Q1 median Q3   max     mean       sd    n missing
##    0  1      4 60 21540 100.1712 393.2173 9387       0
```

```
ggplot(data = clean_compasdata, mapping = aes(x = days_in_jail)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "Days in Jail")
```



The days spent in prison is not as variable as the days spent in jail. The minimum 0 and the maximum is 190739. This skews the mean to 784.7951, but the median is 0. The distinction between jail and prison is still unclear.

```
favstats(clean_compasdata$days_in_prison)
```

```
##  min Q1 median Q3    max     mean       sd    n missing
##    0  0      0  0 190739 784.7951 3473.352 9387       0
```

The number of crimes recommitted by the defendants who re-committed a crime within two years ranges from 1 to 55, with a median of 1 and a mean of 1.73.

```
favstats(clean_compasdata$num_r_cases)
```

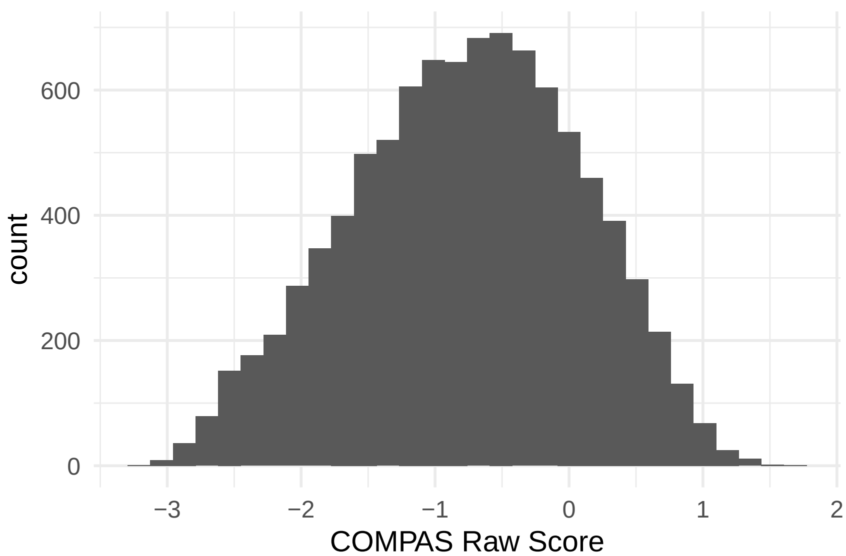
```
##  min Q1 median Q3 max     mean       sd    n missing
##    1  1      1  2  55 1.736512 1.629916 3188    6199
```

Finally, the COMPAS tool outputs a raw score for each defendant. The raw score ranges from -3.21 to 1.69 with a median of -0.74 and a mean of -0.78. The distribution of the raw scores is visualized on the histogram below. The distribution is unimodal and symmetric with a slight left skew.

```
favstats(clean_compasdata$raw_score)
```

```
##    min    Q1 median    Q3  max       mean       sd    n missing
##  -3.21 -1.38  -0.74 -0.15 1.69 -0.7763417 0.856942 9387       0
```

```
ggplot(data = clean_compasdata, mapping = aes(x = raw_score)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "COMPAS Raw Score")
```

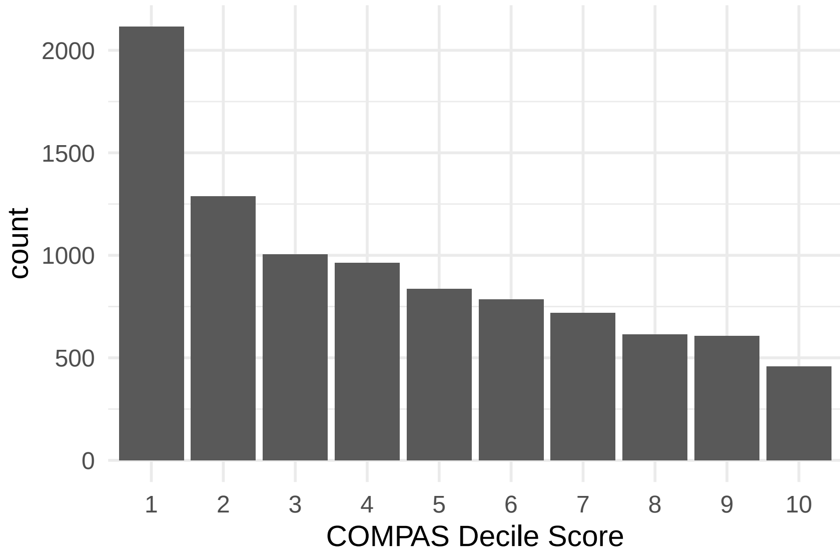


The raw scores are then converted into decile scores that determine the predicted risk of recidivism. The decile scores range from 1 to 10 with a median of 4 and a mean of 4.3. The histogram displays the distribution of the decile scores – it makes me wonder how, or whether, the decile scores are computed from the raw scores.

```
favstats(clean_compasdata$decile_score)
```

```
##  min Q1 median Q3 max     mean       sd    n missing
##    1  2      4  7  10 4.305849 2.849011 9387       0
```

```r
ggplot(data = clean_compasdata, mapping = aes(x = as.factor(decile_score))) +
  geom_bar() +
  theme_minimal() +
  labs(x = "COMPAS Decile Score")
```



Note that the decile scores are mapped to 'low', 'medium', and 'high' risk as detailed in the table below.

```r
clean_compasdata %>%
  dplyr::select(decile_score, score_text) %>%
  filter(score_text != 'N/A') %>%
  rename("Risk" = score_text) %>%
  group_by(Risk) %>%
  summarise("Min" = min(decile_score),
            "Max" = max(decile_score)) %>%
  arrange(Min) %>%
  kable(booktabs = TRUE)
```

Risk	Min	Max
Low	1	4
Medium	5	7
High	8	10

This concludes our univariate analysis of the variables in the COMPAS data set. Next, we will look at some of the bivariate relationships.

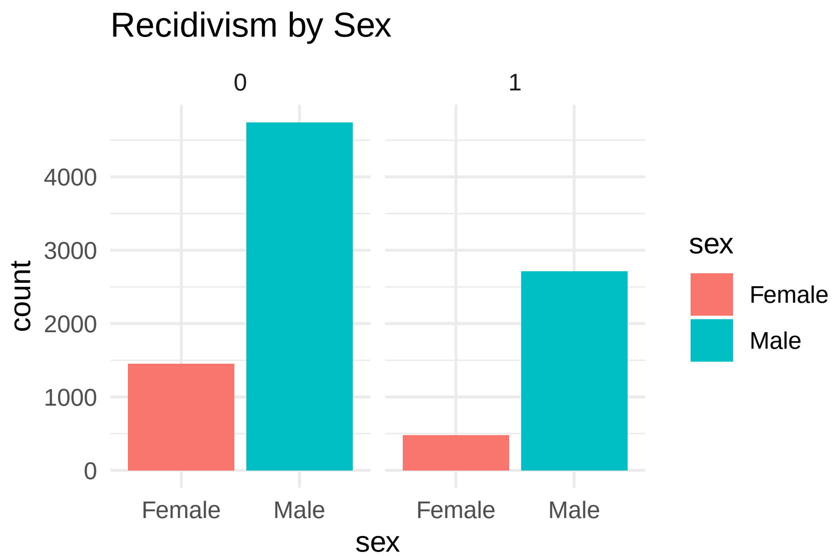
Bivariate Analysis

In this section, we will explore the relationships between our variables and the response variable, `is_recid`, which records whether or not a defendant recommitted a crime within 2 years.

Categorical Variables

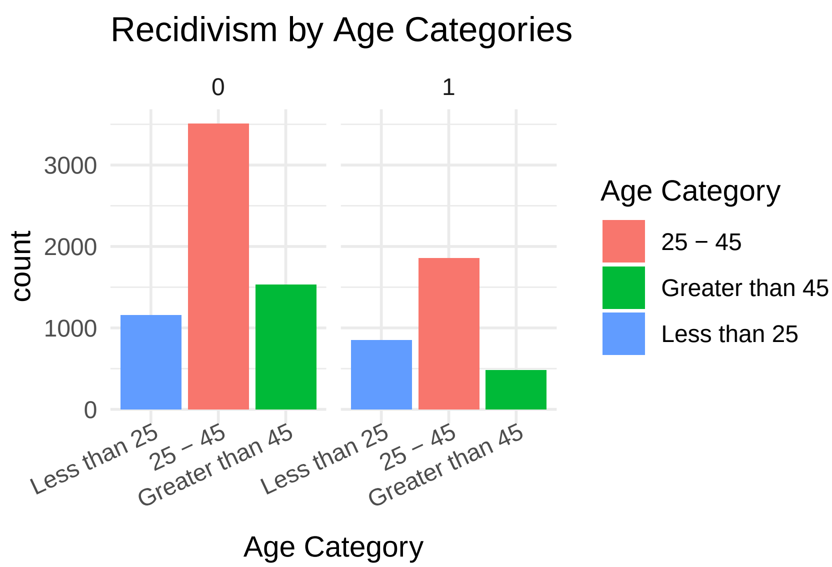
It doesn't appear as though there is much evident relationship between sex and recidivism.

```r
ggplot(data = clean_compasdata, mapping = aes(x = sex, fill = sex)) +
  geom_bar() +
  theme_minimal() +
  facet_wrap(~is_recid) +
  labs(title = "Recidivism by Sex")
```



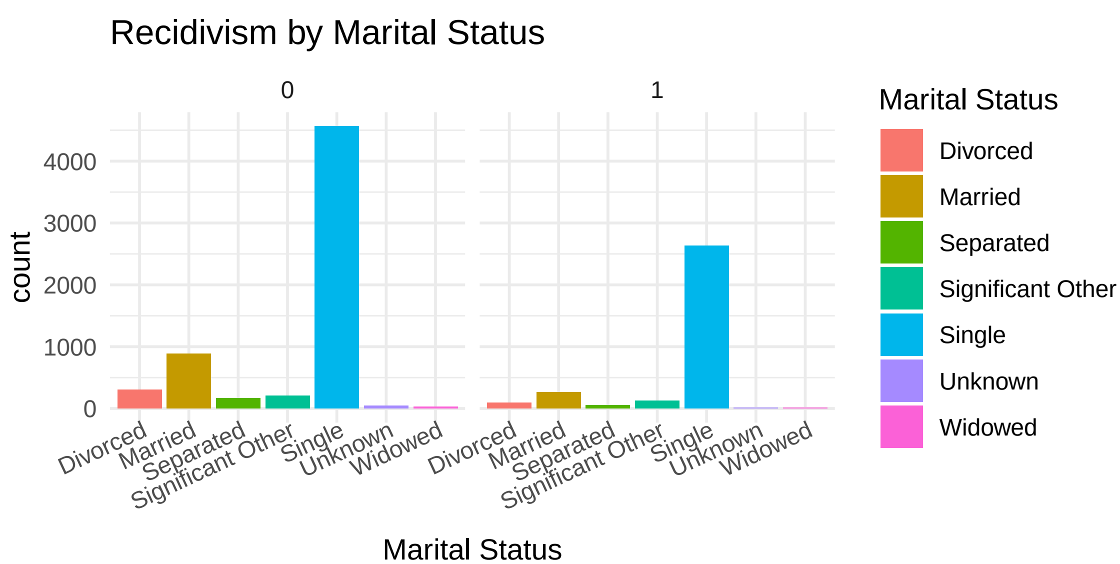
Among defendants who do not recidivate, there are more defendants that are aged 45 in comparison to those less than 25. However, among those that recidivated, there are more defendants that are less than 25 in comparison to those that are greater than 45. This indicates that age may hold some valuable information regarding a defendant's likelihood of recidivism.

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

ggplot(data = clean_compasdata,
       mapping = aes(x = age_cat, fill = age_cat)) +
  geom_bar() +
  theme_minimal() +
  facet_wrap(~is_recid) +
  labs(title = "Recidivism by Age Categories",
       x = "Age Category",
       fill = "Age Category") +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1)) +
  scale_x_discrete(limits = order)
```



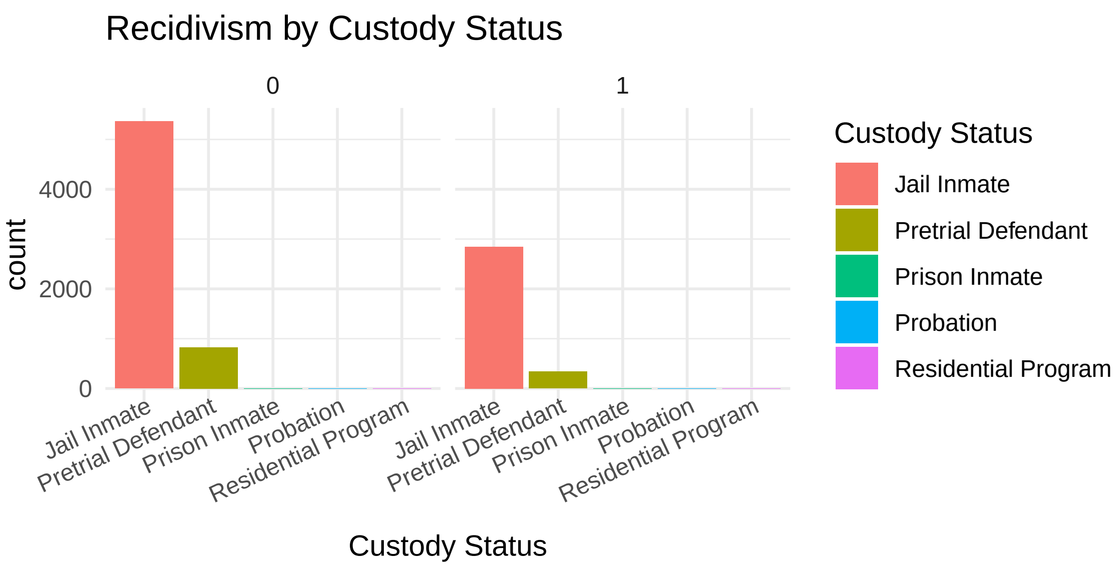
It doesn't appear as though there is much relationship between recidivism and marital status.

```
ggplot(data = clean_compasdata,
       mapping = aes(x = marital_status, fill = marital_status)) +
  geom_bar() +
  theme_minimal() +
  facet_wrap(~is_recid) +
  labs(title = "Recidivism by Marital Status",
       x = "Marital Status",
       fill = "Marital Status") +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1))
```



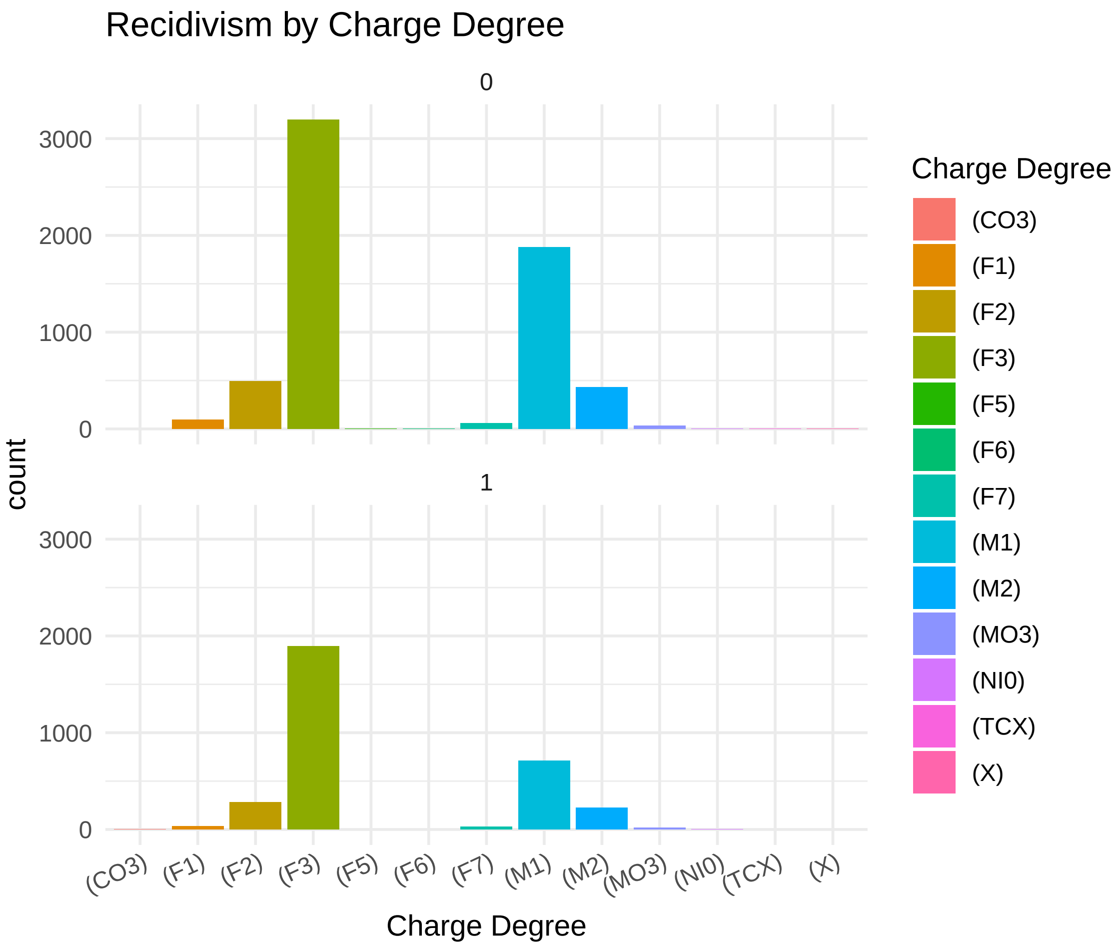
It doesn't appear as though there is much relationship between recidivism and custody status.

```r
ggplot(data = clean_compasdata,
       mapping = aes(x = custody_status, fill = custody_status)) +
  geom_bar() +
  theme_minimal() +
  facet_wrap(~is_recid) +
  labs(title = "Recidivism by Custody Status",
       x = "Custody Status",
       fill = "Custody Status") +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1))
```



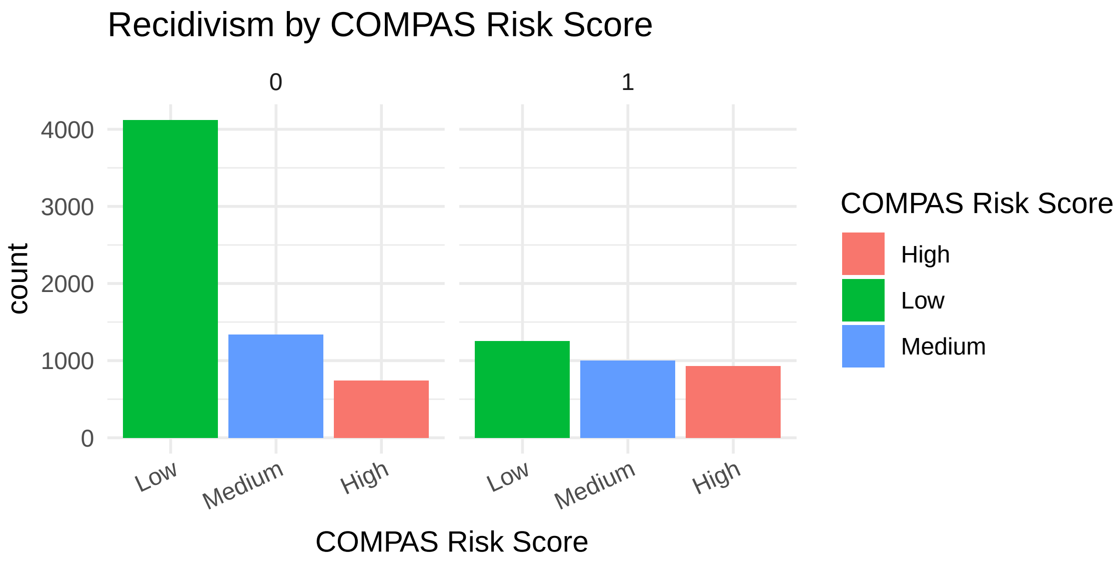
It doesn't appear as though there is much relationship between recidivism and charge degree.

```r
ggplot(data = clean_compasdata,
       mapping = aes(x = c_charge_degree, fill = c_charge_degree)) +
  geom_bar() +
  theme_minimal() +
  facet_wrap(~is_recid, ncol = 1) +
  labs(title = "Recidivism by Charge Degree",
       x = "Charge Degree",
       fill = "Charge Degree") +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1))
```



However, it appears as though the COMPAS tool classifies defendants who recommit a crime as almost as equally risky of recidivism – there is no significant distinction between 'low', 'medium', and 'high' risk for these defendants. For the defendants that don't recommit a crime, most are predicted as 'low' risk, followed by 'medium', and then 'high' risk. Note, however, that this variable will not be included as a predictor in the model as the purpose of this analysis is to assess COMPAS performance, or more generally, standard ML approaches, in comparison to the Seldonian framework.

```
order <- c("Low", "Medium", "High")

ggplot(data = clean_compasdata,
       mapping = aes(x = score_text, fill = score_text)) +
  geom_bar() +
  theme_minimal() +
  facet_wrap(~is_recid) +
  labs(title = "Recidivism by COMPAS Risk Score",
       x = "COMPAS Risk Score",
       fill = "COMPAS Risk Score") +
  theme(axis.text.x = element_text(angle = 25, vjust = 1.2, hjust=1)) +
  scale_x_discrete(limits = order)
```



Next, let's perform a similar analysis for the continuous variables.

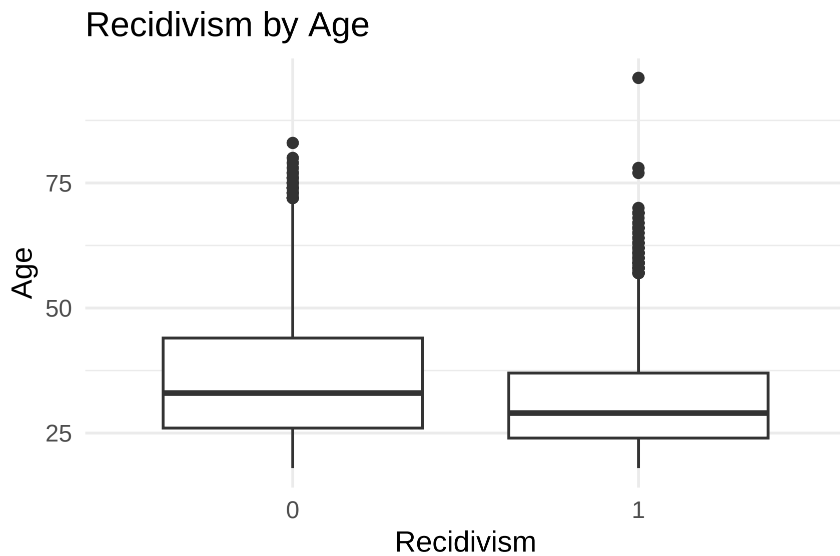
Continous Variables

There is a difference in the mean and median ages for defendants who recommit a crime within two years versus though who don't. Those who recidivate tend to be younger than those who don't, indicating that this will be a useful variable in the model. This is in line with intuition from society.

```
favstats(data = clean_compasdata, age ~ is_recid)
```

```
##   is_recid min Q1 median Q3 max     mean       sd    n missing
## 1        0  18 26     33 44  83 36.04646 12.17525 6199       0
## 2        1  18 24     29 37  96 32.24122 10.62147 3188       0
```

```
ggplot(data = clean_compasdata,
       mapping = aes(x = as.factor(is_recid), y = age)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Recidivism by Age",
       x = "Recidivism",
       y = "Age")
```



There is not much distributional difference in juvenile felony counts for defendants who recidivate versus those who don't.

```
favstats(data = clean_compasdata, juv_fel_count ~ is_recid)
```

```
##   is_recid min Q1 median Q3 max       mean        sd    n missing
## 1        0   0  0      0  0  13 0.03645749 0.3297061 6199       0
## 2        1   0  0      0  0  20 0.10100376 0.6221204 3188       0
```

There is not much distributional difference in juvenile misdemeanor counts for defendants who recidivate versus those who don't.

```
favstats(data = clean_compasdata, juv_misd_count ~ is_recid)
```

```
##   is_recid min Q1 median Q3 max       mean        sd    n missing
## 1        0   0  0      0  0  12 0.04387804 0.3356002 6199       0
## 2        1   0  0      0  0  13 0.14648683 0.6388211 3188       0
```

There is not much distributional difference in juvenile offenses for defendants who recidivate versus those who don't.

```
favstats(data = clean_compasdata, juv_other_count ~ is_recid)
```

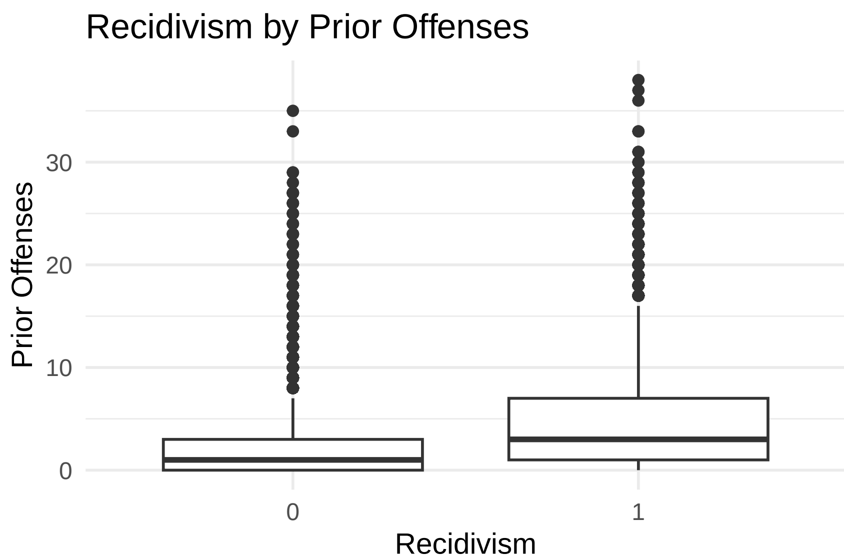
```
##   is_recid min Q1 median Q3 max       mean        sd    n missing
## 1        0   0  0      0  0  11 0.06355864 0.3966132 6199       0
## 2        1   0  0      0  0   9 0.16844417 0.5768642 3188       0
```

There is some distributional difference in non-juvenile prior offenses for defendants who recidivate versus those who don't, as is indicated by the different means and medians. Those who recommit a crime within two years tend to have more prior offenses. This will be a useful variable to include in the models.

```r
favstats(data = clean_compasdata, priors_count ~ is_recid)
```

```
##   is_recid min Q1 median Q3 max     mean       sd    n missing
## 1        0   0  0      1  3  35 2.157283 3.684641 6199       0
## 2        1   0  1      3  7  38 4.708908 5.589893 3188       0
```

```r
ggplot(data = clean_compasdata,
       mapping = aes(x = as.factor(is_recid), y = priors_count)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Recidivism by Prior Offenses",
       x = "Recidivism",
       y = "Prior Offenses")
```



There doesn't appear to be any distributional difference in days between COMPAS screening and arrest for defendants who recidivate versus those who don't. This will not be a useful variable for modeling.

```r
favstats(data = clean_compasdata, days_b_screening_arrest ~ is_recid)
```

```
##   is_recid min Q1 median Q3 max     mean       sd    n missing
## 1        0   0  1      1  1  30 2.149218 4.859912 6199       0
## 2        1   0  1      1  1  30 2.122647 4.957777 3188       0
```

While the means differ because of the right-skew nature of the data, there doesn't appear to be much distributional difference in days since COMPAS screening for defendants who recidivate versus those who don't. This will not be a useful variable for modeling.

```r
favstats(data = clean_compasdata, c_days_from_compas ~ is_recid)
```

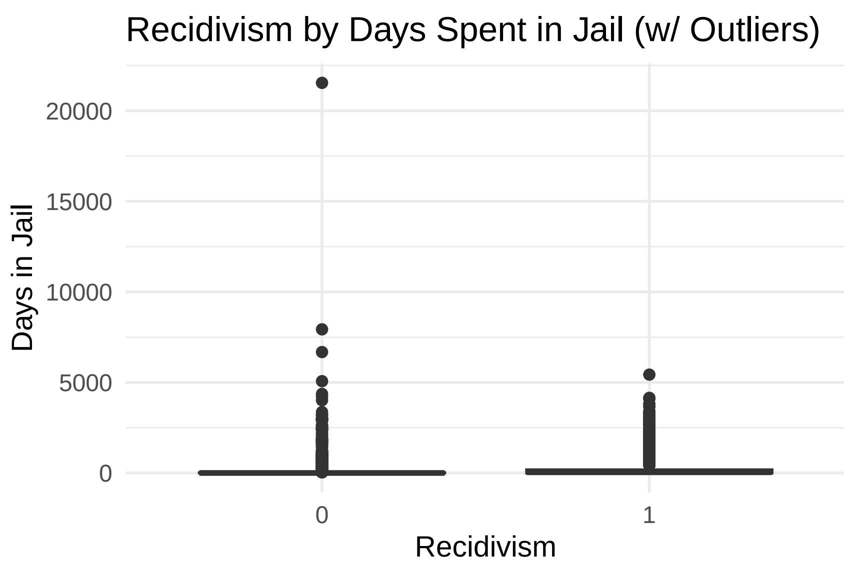
```
##   is_recid min Q1 median Q3  max     mean       sd    n missing
## 1        0   0  1      1  1 9485 31.21052 305.1847 6199       0
## 2        1   0  1      1  1 5450 12.70107 151.5946 3188       0
```

There is an evident difference in the distribution of the number of days spent in jail for participants who recommit a crime within two years versus those who don't. The mean, median, and max value differ significantly, indicating variation that may be useful in modeling. It's hard to visualize the boxplots with all the outliers, so the second boxplot trims the y-axis to better visualize this relationship.

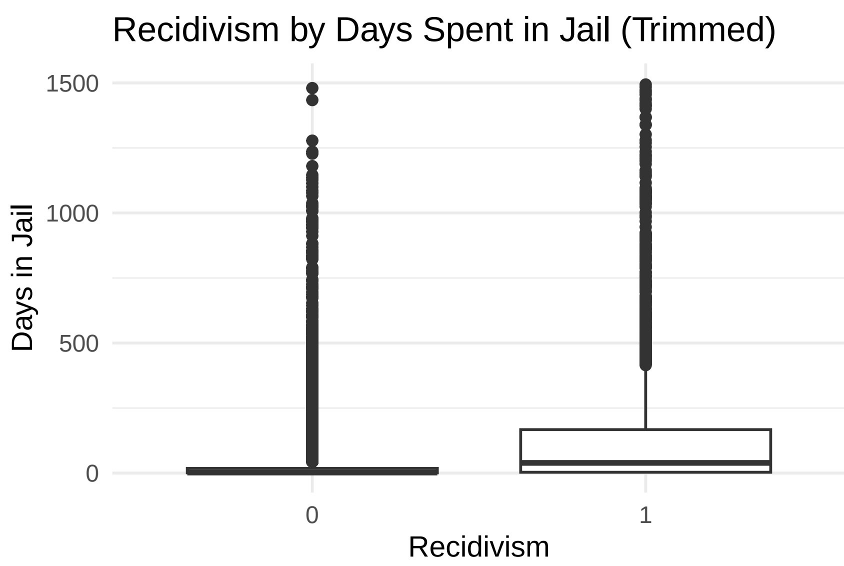
```
favstats(data = clean_compasdata, days_in_jail ~ is_recid)
```

```
##   is_recid min Q1 median     Q3   max      mean       sd    n missing
## 1        0   0  1      2  19.00 21540  65.86562 395.1086 6199       0
## 2        1   0  3     41 178.25  5432 166.87767 380.8262 3188       0
```

```
ggplot(data = clean_compasdata,
       mapping = aes(x = as.factor(is_recid), y = days_in_jail)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Recidivism by Days Spent in Jail (w/ Outliers)",
       x = "Recidivism",
       y = "Days in Jail")
```




```
ggplot(data = clean_compasdata,
       mapping = aes(x = as.factor(is_recid), y = days_in_jail)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Recidivism by Days Spent in Jail (Trimmed)",
       x = "Recidivism",
       y = "Days in Jail") +
  ylim(0,1500)
```



There doesn't appear to be much distributional difference, rather than the effects of extreme right skews, in days spent in prison for defendants who recidivate versus those who don't. The difference between jail and prison is still not clear, so this will not be a useful variable for modeling.

```
favstats(data = clean_compasdata, days_in_prison ~ is_recid)
```

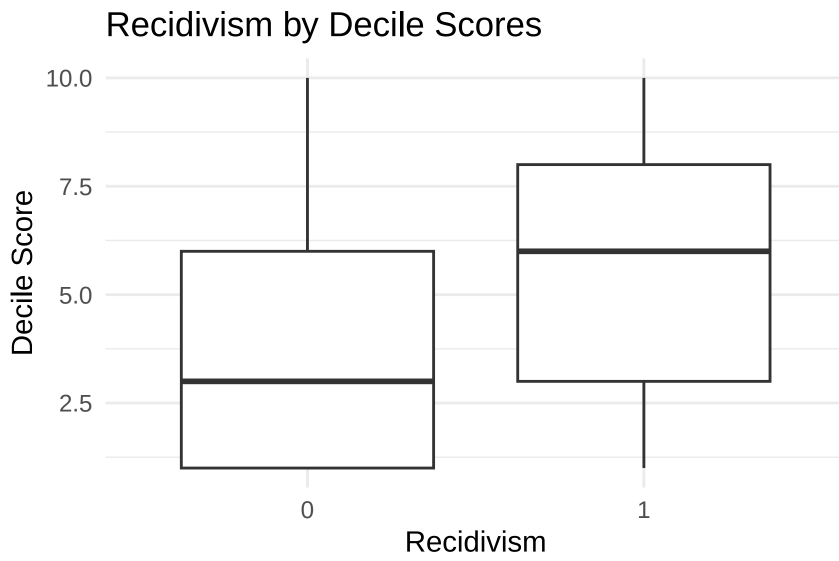
```
##   is_recid min Q1 median     Q3    max      mean       sd    n missing
## 1        0   0  0      0   0.00 190739  441.0958 3017.073 6199       0
## 2        1   0  0      0 223.75  67056 1453.1114 4141.345 3188       0
```

Finally, let's assess the COMPAS decile scores. The median and mean decile scores differ for defendants who recommit a crime within 2 years versus those who don't. The median score for those who don't is 3, which is mapped to low risk. The median score for those who do is 6, which is mapped to medium risk. This indicates that the COMPAS tool has some predictive accuracy. However, the range of scores is the same for both defendants who recidivate versus those who do not, suggesting that the tool is not entirely accurate in its predictions.

```
favstats(data = clean_compasdata, decile_score ~ is_recid)
```

```
##   is_recid min Q1 median Q3 max     mean       sd    n missing
## 1        0   1  1      3  6  10 3.694467 2.667049 6199       0
## 2        1   1  3      6  8  10 5.494668 2.816137 3188       0
```

```
ggplot(data = clean_compasdata,
       mapping = aes(x = as.factor(is_recid), y = decile_score)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Recidivism by Decile Scores",
       x = "Recidivism",
       y = "Decile Score")
```



This wraps up our analysis of the bivariate relationships between the continuous variables in the data set and the response variable: `is_recid`.

Multivariate Analysis

Based on the univariate and bivariate analysis, the 8 most informative predictive variables for modeling will be:

- `sex`
- `age`
- `age category`
- `marital status`
- `custody status`
- `prior offenses`
- `charge degree`
- `days in jail`

We will also include `race` in the modeling data set as our demographic variable, though it will not be included in the models themselves. Finally, `is_recid`, the response variable, will also be selected in the data set.

For further analysis, we will also include the COMPAS decile scores to assess which of these variables may have been used to model the COMPAS risk assessment tool.

Now, let's create a new data set with these 11 variables. Below is a glimpse of the data set.

```
compas_final <- clean_compasdata %>%
  dplyr::select(c(race, sex, age, age_cat, marital_status,
                  custody_status, priors_count, c_charge_degree,
                  days_in_jail, decile_score, is_recid))

glimpse(compas_final)
```

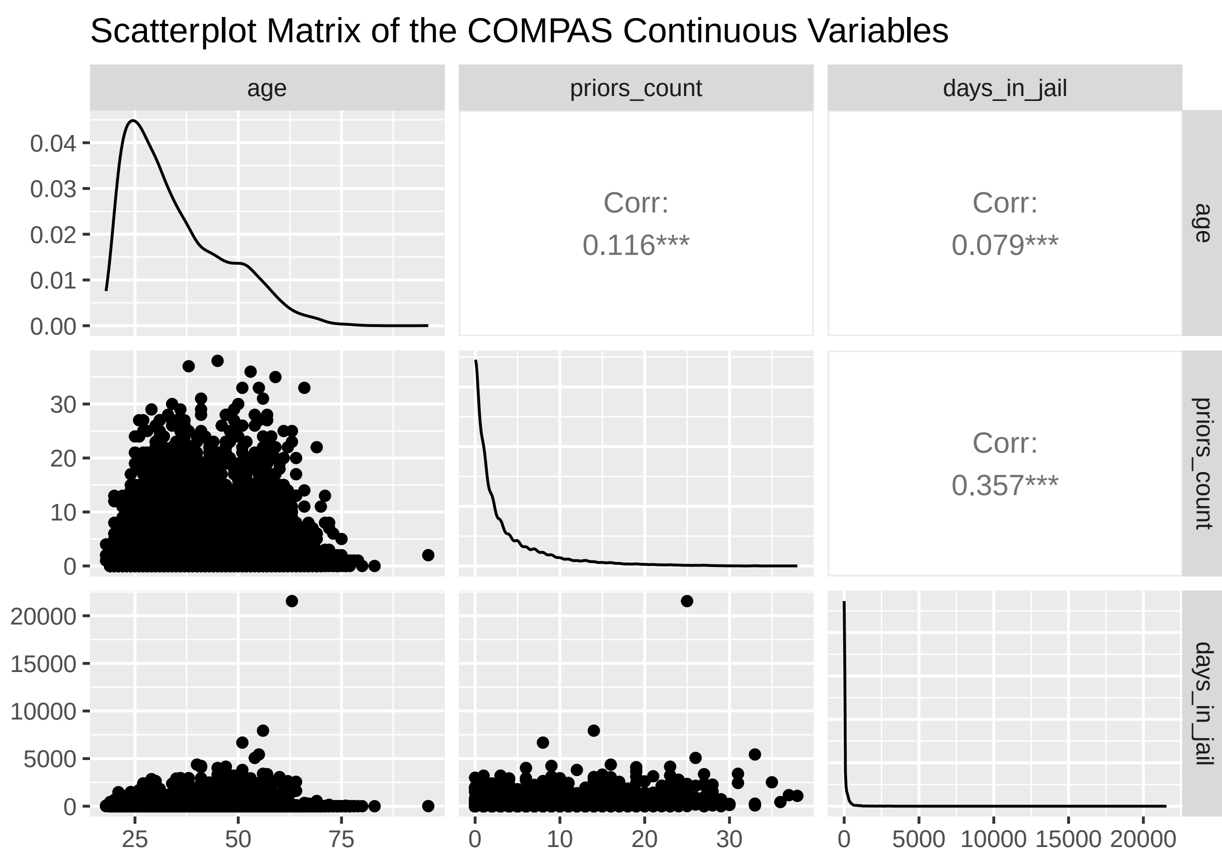
```
## Rows: 9,387
## Columns: 11
## $ race            <chr> "Other", "African-American", "African-American", "Othe~
## $ sex             <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Femal~
## $ age             <int> 69, 34, 24, 44, 41, 43, 39, 20, 26, 27, 23, 37, 22, 41~
```

```
## $ age_cat          <chr> "Greater than 45", "25 - 45", "Less than 25", "25 - 45~
## $ marital_status   <chr> "Single", "Single", "Single", "Separated", "Single", "~
## $ custody_status   <chr> "Jail Inmate", "Jail Inmate", "Jail Inmate", "Jail Inm~
## $ priors_count     <int> 0, 0, 4, 0, 14, 3, 0, 0, 0, 0, 3, 0, 0, 0, 1, 7, 0, 3,~
## $ c_charge_degree <chr> "(F3)", "(F3)", "(F3)", "(M1)", "(F3)", "(F3)", "(M1)"~
## $ days_in_jail     <dbl> 8, 10, 139, 1, 48, 17, 3, 46, 87, 1, 4, 1, 0, 1, 183, ~
## $ decile_score     <int> 1, 3, 4, 1, 6, 4, 1, 10, 5, 4, 6, 1, 3, 4, 1, 3, 1, 10~
## $ is_recid         <int> 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, ~
```

Scatterplot Matrix

First, a scatterplot matrix with just the 3 continuous predictive variables in the final data set, `age`, `prior offenses`, and `days in jail`, will help to elucidate the covariate relationships between the variables. All the variables have moderate to weak correlations, with the strongest correlation of 0.357 being between the number of prior offenses and the number of days spent in jail. There are no concerns for multicollinearity.

```r
ggpairs(data = compas_final,
        columns = c("age", "priors_count", "days_in_jail"),
        title = "Scatterplot Matrix of the COMPAS Continuous Variables")
```



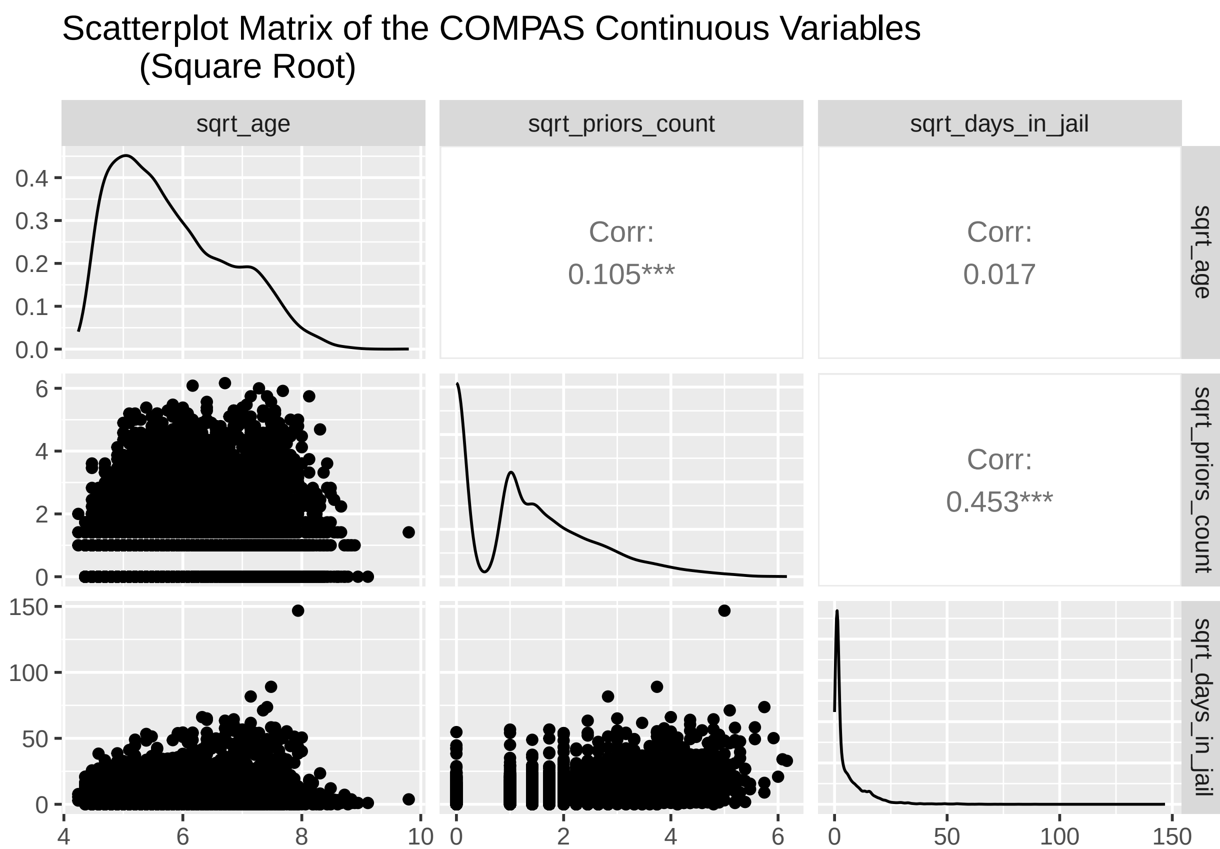
As observed, the variables have a significant right skew and the relationship is non-linear. Let's explore what effect different transformations may have on the covariate relationships. The log transformation resulted in many non-finite values, so we will look at a square root transformation instead.

```r
compas_final <- compas_final %>%
  mutate(sqrt_age = sqrt(age),
```

```
        sqrt_priors_count = sqrt(priors_count),
        sqrt_days_in_jail = sqrt(days_in_jail))
```

While this transformation strengthened the relationship between the number of prior offenses and the number of days spent in jail, it weakened the other correlations. Next, we'll assess how much this affects the relationship with the COMPAS decile scores.

```
ggpairs(data = compas_final,
        columns = c("sqrt_age", "sqrt_priors_count", "sqrt_days_in_jail"),
        title = "Scatterplot Matrix of the COMPAS Continuous Variables
        (Square Root)")
```



Pearson's Correlation Matrix

Decile scores has a moderate relationship with age, prior offenses, and number of days spent in jail – this suggests that these variables may indeed be useful for modeling recidivism.

```
mycordata1 <- compas_final %>%
  dplyr::rename("Age" = age,
         "Priors" = priors_count,
         "Days in Jail" = days_in_jail,
         "Decile Scores" = decile_score) %>%
  dplyr::select("Age", "Priors", "Days in Jail", "Decile Scores")
```

```r
cor(mycordata1) %>%
  kable(digits = 2,
        booktabs = TRUE)
```

	Age	Priors	Days in Jail	Decile Scores
Age	1.00	0.12	0.08	-0.39
Priors	0.12	1.00	0.36	0.45
Days in Jail	0.08	0.36	1.00	0.25
Decile Scores	-0.39	0.45	0.25	1.00

The square root transformation of the predictor variables actually strengthens the relationship with decile scores – this suggests that the square root transformation of these variables may be better for modeling recidivism.

```r
mycordata2 <- compas_final %>%
  dplyr::rename("Square Root Age" = sqrt_age,
         "Priors" = sqrt_priors_count,
         "Days in Jail" = sqrt_days_in_jail,
         "Decile Scores" = decile_score) %>%
  dplyr::select("Square Root Age", "Priors", "Days in Jail", "Decile Scores")

cor(mycordata2) %>%
  kable(digits = 2,
        booktabs = TRUE)
```

	Square Root Age	Priors	Days in Jail	Decile Scores
Square Root Age	1.00	0.11	0.02	-0.39
Priors	0.11	1.00	0.45	0.46
Days in Jail	0.02	0.45	1.00	0.45
Decile Scores	-0.39	0.46	0.45	1.00

Spearman's Correlation Matrix

Spearman's Correlation is better at capturing non-linear relationships. Using Spearman correlations reveals stronger correlations between the variables and the COMPAS decile scores.

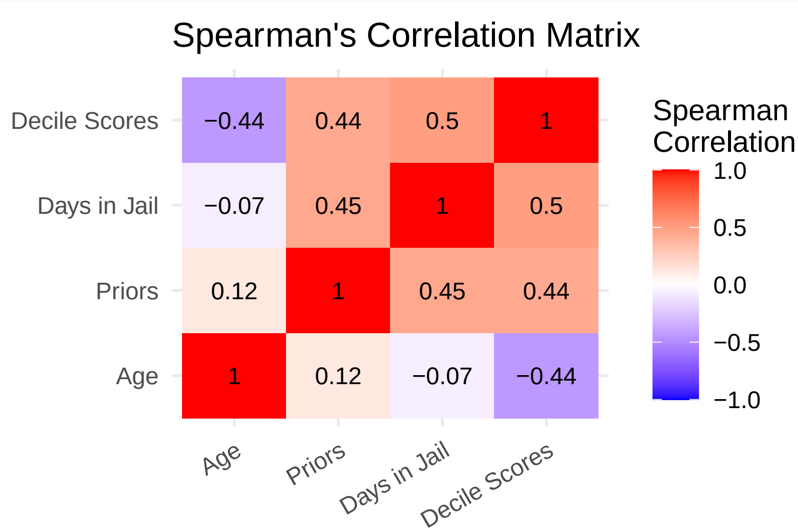
There is no observable difference when calculating Spearman's correlation with the square-root transformed variables versus the original variables.

```r
mycordata3 <- compas_final %>%
  dplyr::rename("Age" = age,
         "Priors" = priors_count,
         "Days in Jail" = days_in_jail,
         "Decile Scores" = decile_score) %>%
  dplyr::select("Age", "Priors", "Days in Jail", "Decile Scores")

cor(mycordata3, method = "spearman") %>%
  kable(digits = 2,
        booktabs = TRUE)
```

	Age	Priors	Days in Jail	Decile Scores
Age	1.00	0.12	-0.07	-0.44
Priors	0.12	1.00	0.45	0.44
Days in Jail	-0.07	0.45	1.00	0.50
Decile Scores	-0.44	0.44	0.50	1.00

Finally, let's visualize these correlations.

```
mycors <- round(cor(mycordata3, method = "spearman"),2)
mycorplot <- melt(mycors)

ggplot(data = mycorplot, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  labs(x = "",
       y = "",
       title = "Spearman's Correlation Matrix") +
  scale_fill_gradient2(
    low = "blue",
    high = "red",
    mid = "white",
    midpoint = 0,
    limit = c(-1, 1),
    space = "Lab",
    name = "Spearman\nCorrelation"
  ) +
  geom_text(aes(Var2, Var1, label = value),
            color = "black",
            size = 3) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 30, vjust = 1, hjust = 1))
```



Now that we have a thorough understanding of the make-up of the data set, we will perform a demographic analysis next to get a better understanding of the racial discrepancies that may be present before, finally, proceeding with the recidivism risk modeling.

Demographic Group Analysis

analyze variables by race.

Logistic Regression

run logistic regression and analyze results.

Seldonian Classification

run seldonian framework on the data set and analyze results.

Results

- remember to look at visuals/ tables from chap 4 and add as appropriate