Potential Confounders

GVPUF - County

Potential Covariates

We evaluated what percentage of each of our covariates of interest were missing:

```
tbl <- medicare %>%
  select(
    county, average_age, percent_male, `percent_non-hispanic_white`,
    percent_african_american, percent_hispanic,
    percent_eligible_for_medicaid
  ) %>%
  summarize_all(.funs = function(x)
    return(paste0(round(sum(is.na(x))/length(x), 3)*100, "%"))) %>%
  t() %>% data.frame()

names(tbl) <- c("% of Missing values")
kableExtra::kable(tbl)
```

	% of Missing values
county	0%
average_age	0%
percent_male	0.1%
percent_non-hispanic_white	42.2%
percent_african_american	42.2%
percent_hispanic	42.2%
percent_eligible_for_medicaid	0.5%

We investigated the pattern of missingness:

```
medicare %>%
  mutate(
    p_white_missing = is.na(`percent_non-hispanic_white`),
    p_aa_missing = is.na(percent_african_american),
    p_h_missing = is.na(percent_hispanic)
  ) %>%
  group_by(year, p_white_missing, p_aa_missing, p_h_missing) %>%
  summarize(cnt = n())
```

```
## `summarise()` regrouping output by 'year', 'p_white_missing', 'p_aa_missing' (override with `.groups
```

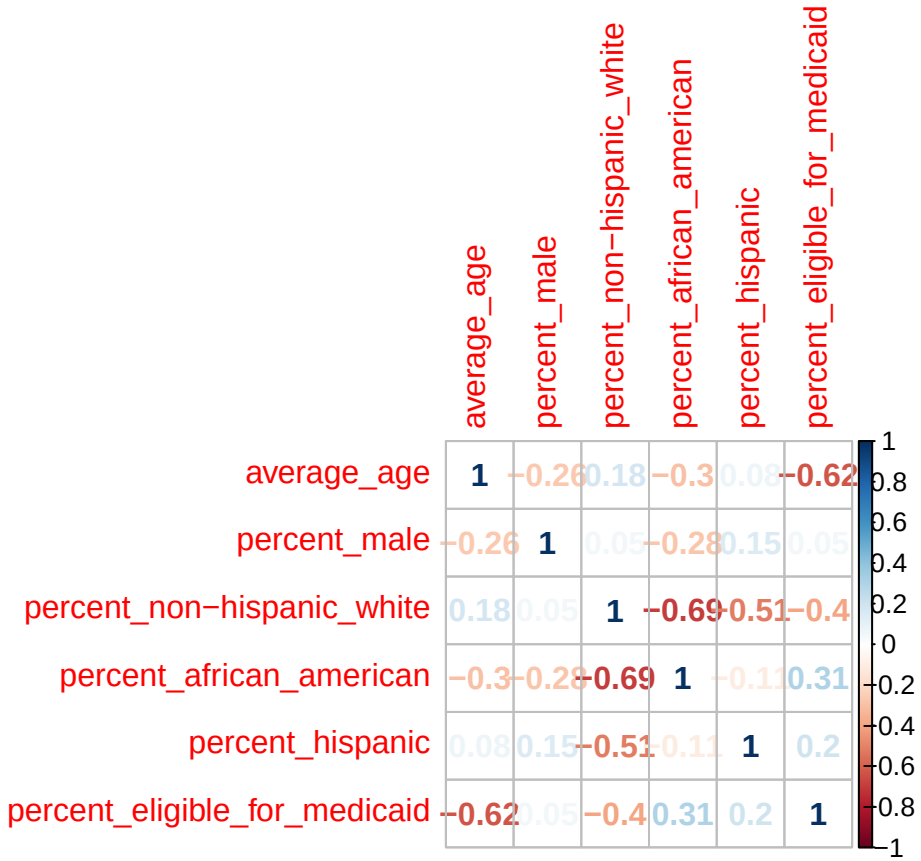
```
## # A tibble: 24 x 5
## # Groups:   year, p_white_missing, p_aa_missing [24]
##     year p_white_missing p_aa_missing p_h_missing   cnt
##    <dbl> <lgl>           <lgl>        <lgl>       <int>
##  1  2007 FALSE           FALSE        FALSE        1651
##  2  2007 TRUE            TRUE         TRUE         1494
##  3  2008 FALSE           FALSE        FALSE        1664
##  4  2008 TRUE            TRUE         TRUE         1483
##  5  2009 FALSE           FALSE        FALSE        1703
```

```
##  6  2009 TRUE            TRUE         TRUE         1442
##  7  2010 FALSE           FALSE        FALSE        1741
##  8  2010 TRUE            TRUE         TRUE         1404
##  9  2011 FALSE           FALSE        FALSE        1791
## 10  2011 TRUE            TRUE         TRUE         1354
## # ... with 14 more rows
```

It appears that the same counties do not report these metrics. Given the high percentage of missingness in these covariates, I've decided to drop them from my potential covariates list.

Check correlaton in the covariates Complete cases with race variables

```
medicare %>%
  select(
    average_age, percent_male, `percent_non-hispanic_white`,
    percent_african_american, percent_hispanic,
    percent_eligible_for_medicaid
  ) %>% na.omit() %>%
  cor() %>% corrplot::corrplot(method = "number")
```

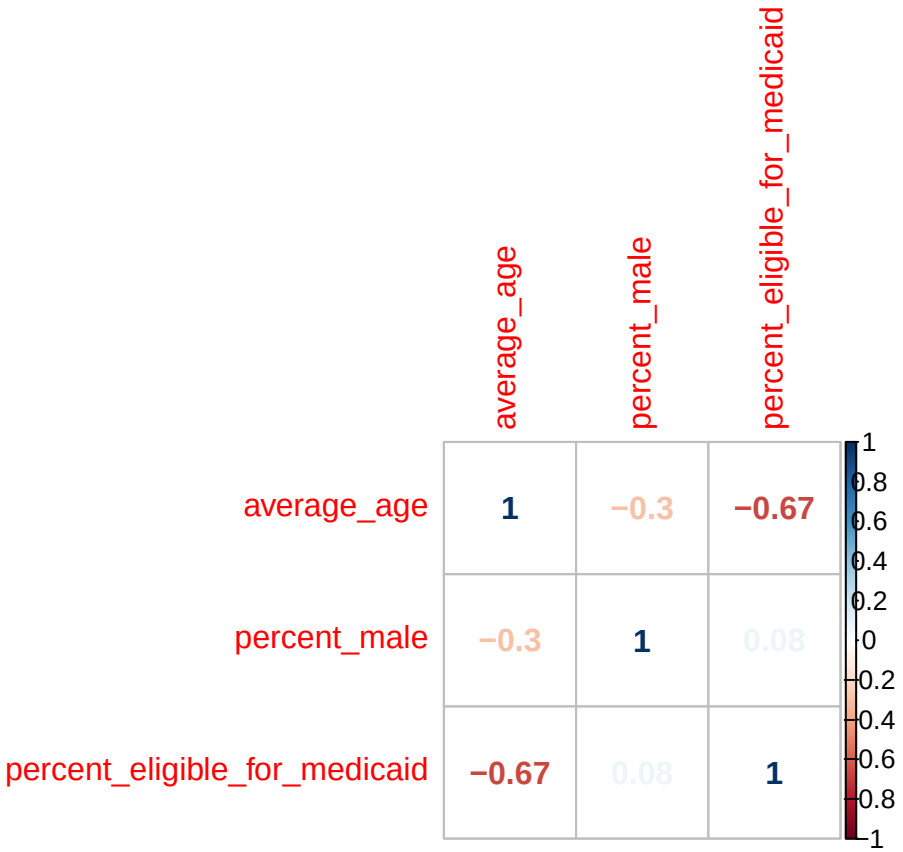


There is a strong correlation between percent African American and percent non-Hispanic white.

Complete cases without race variables

```
medicare %>%
  select(
    average_age, percent_male, percent_eligible_for_medicaid
  ) %>% na.omit() %>%
  cor() %>%
```

```
corrplot::corrplot(method = "number")
```



Without the race variables, which have a high amount of missingness, there is a high correlation between percent eligible for Medicaid and average age.

GVPUF - HRR

Potential Covariates

We evaluated what percentage of each of our covariates of interest were missing:

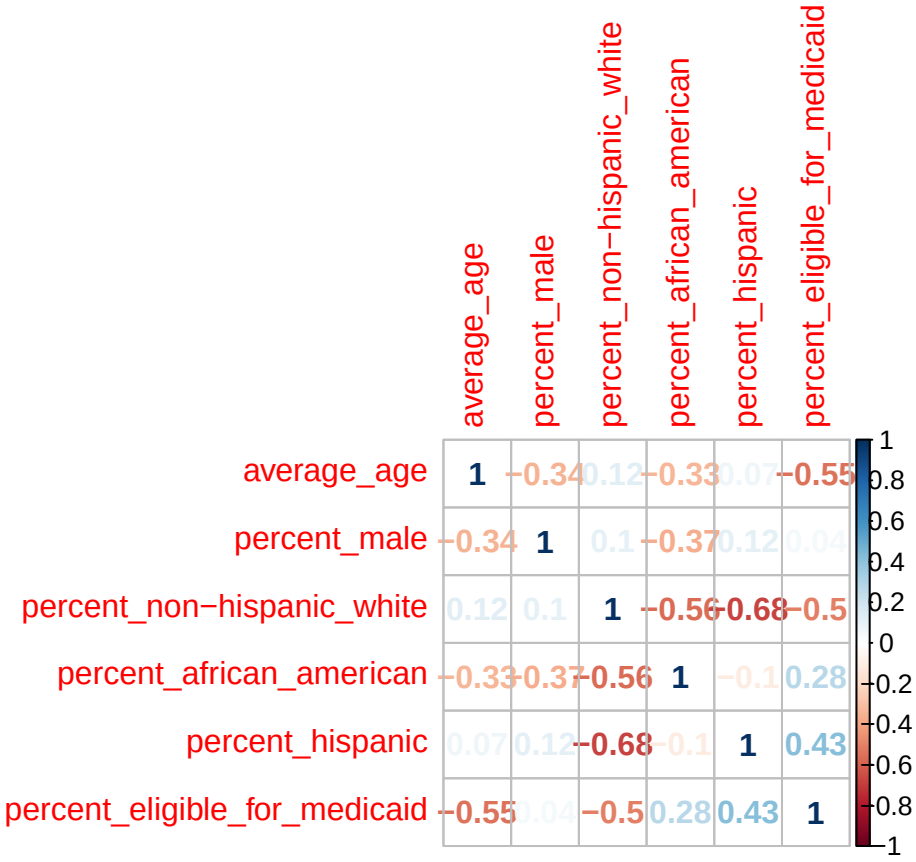
```
tbl <- hrr %>%
  select(
    hrr, average_age, percent_male, `percent_non-hispanic_white`, percent_african_american, percent_hisp
  ) %>%
  summarize_all(.funs = function(x)
    return(paste0(round(sum(is.na(x))/length(x), 3)*100, "%"))) %>%
  t() %>% data.frame()

names(tbl) <- c("% of Missing values")
kableExtra::kable(tbl)
```

	% of Missing values
hrr	0%
average_age	0.6%
percent_male	0.6%
percent_non-hispanic_white	0.6%
percent_african_american	0.6%
percent_hispanic	0.6%
percent_eligible_for_medicaid	0.6%

There is a very small amount of missingness across this data.

```
hrr %>%
  select(
    average_age, percent_male, `percent_non-hispanic_white`,
    percent_african_american, percent_hispanic,
    percent_eligible_for_medicaid
  ) %>% na.omit() %>%
  cor() %>% corrplot::corrplot(method = "number")
```



Check correlaton in the covariates

There is a high amount of correlation between percent Hispanic and percnet non-Hispanic white.