Final Project

Justine Hsie, Bingyu Sun, Eleanor Zhang, Annie Yu

12/15/2018

Data Import

```
cancer_raw =
  read_csv("./data/Cancer_Registry.csv") %>%
  janitor::clean_names() %>%
  dplyr::select(target_death_rate, geography, everything()) %>%
  separate(geography, into = c("county", "state"), sep = ",")
```

Data varibale dictionary:

- **target_death_rate:** mean per capita (100,000) cancer mortalities (a)
- **avg_ann_count:** mean number of reported cases of cancer diagnosed annually (a)
- **avg_deaths_per_year:** mean number of reported mortalities due to cancer (a)
- **incidence_rate:** mean per capita (100,000) cancer diagnoses (a)
- **med_income:** median income per county (b)
- **pop_est2015:** population of county (b)
- **poverty_percent:** percent of population in poverty (b)
- **study_per_cap** per capita number of cancer-related clinical trials per county (a)
- **binned_inc:** median income per capita binned by decile (b)
- **median_age:** median age of county residents (b)
- **median_age_male:** median age of male county residents (b)
- **median_age_female:** median age of female county residents (b)
- **geography:** county name (b)
- **avg_household_size:** mean household size of county (b)
- **percent_married:** percent of county residents who are married (b)
- **pct_no_hs18_24:** percent of county residents ages 18-24 highest education attained: less than high school (b)
- **pct_hs18_24:** percent of county residents ages 18-24 highest education attained: high school diploma (b)
- **pct_some_col18_24:** percent of county residents ages 18-24 highest education attained: some college (b)
- **pct_bach_deg18_24:** percent of county residents ages 18-24 highest education attained: bachelor's degree (b)
- **pct_hs25_over:** percent of county residents ages 25 and over highest education attained: high school diploma (b)
- **pct_bach_deg25_over:** percent of county residents ages 25 and over highest education attained: bachelor's degree (b)
- **pct_employed16_over:** percent of county residents ages 16 and over employed (b)
- **pct_unemployed16_over:** percent of county residents ages 16 and over unemployed (b)
- **pct_private_coverage:** percent of county residents with private health coverage (b)
- **pct_private_coverage_alone:** percent of county residents with private health coverage alone (no public assistance) (b)

- **pct_emp_priv_coverage:** percent of county residents with employee-provided private health coverage (b)

- **pct_public_coverage:** percent of county residents with government-provided health coverage (b)
- **pct_public_coverage_alone:** percent of county residents with government-provided health coverage alone (b)

- **pct_white:** percent of county residents who identify as White (b)

- **pct_black:** percent of county residents who identify as Black (b)
- **pct_asian:** percent of county residents who identify as Asian (b)

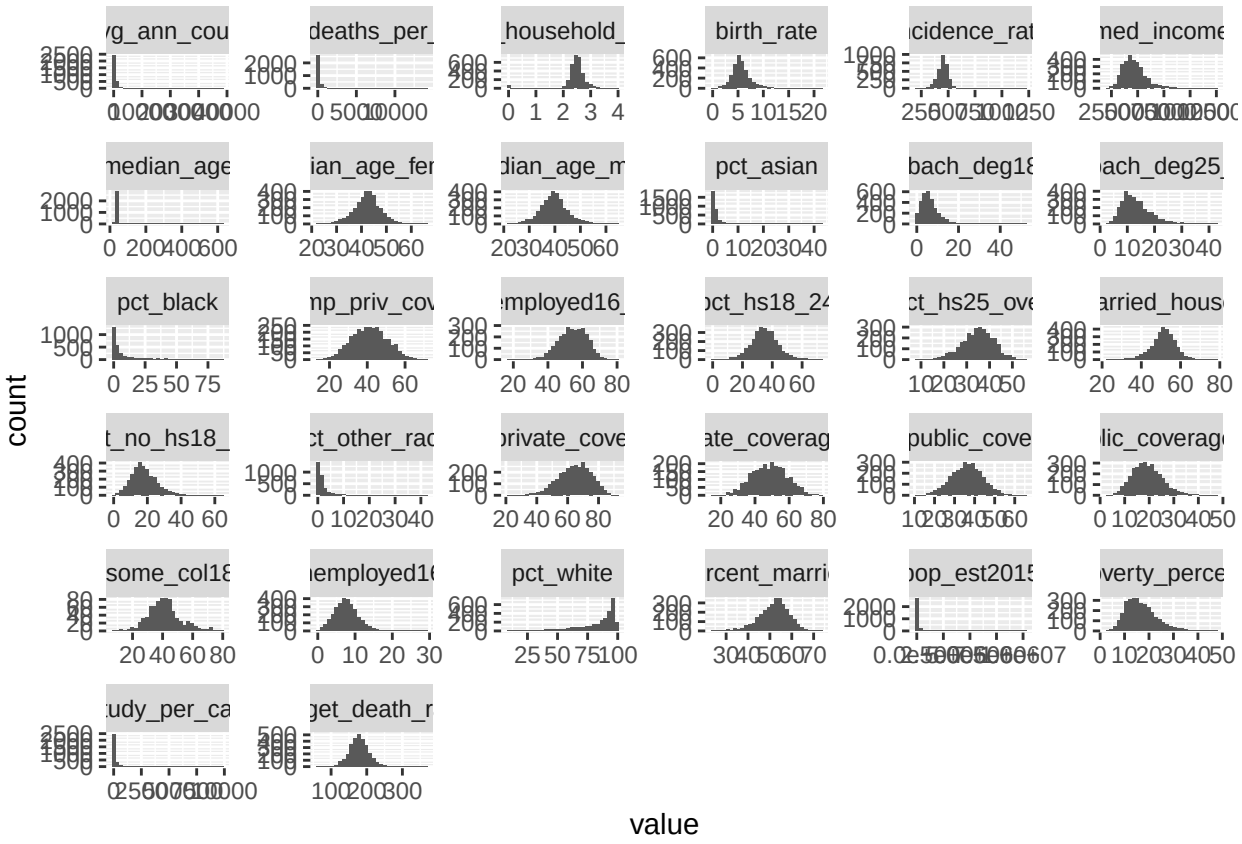
- **pct_other_race:** percent of county residents who identify in a category which is not White, Black, or Asian (b)

- **pct_married_households:** percent of married households (b)
- **birth_rate:** number of live births relative to number of women in county (b)

Look at the distribution of all varibales:

```
cancer_raw %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(bins = 30)
```

```
## Warning: Removed 3046 rows containing non-finite values (stat_bin).
```



Choose variables:

```
cancer_county =
  cancer_raw %>%
  dplyr::select(target_death_rate, incidence_rate, med_income, poverty_percent, median_age:median_age_f
  dplyr::select(-pct_hs25_over, -pct_bach_deg25_over, -pct_employed16_over, -percent_married) %>%
  mutate(pct_upto_hs18_24 = pct_no_hs18_24 + pct_hs18_24,
         pct_above_hs18_24 = 100 - pct_upto_hs18_24,
         pct_with_coverage = pct_private_coverage + pct_public_coverage_alone,
         income_cat = ifelse(med_income < 35000, 0, 1)) %>%
  dplyr::select(-(pct_no_hs18_24:pct_bach_deg18_24), -pct_above_hs18_24, -(pct_private_coverage:pct_publ
  na.omit
```