Final Project

Justin 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)

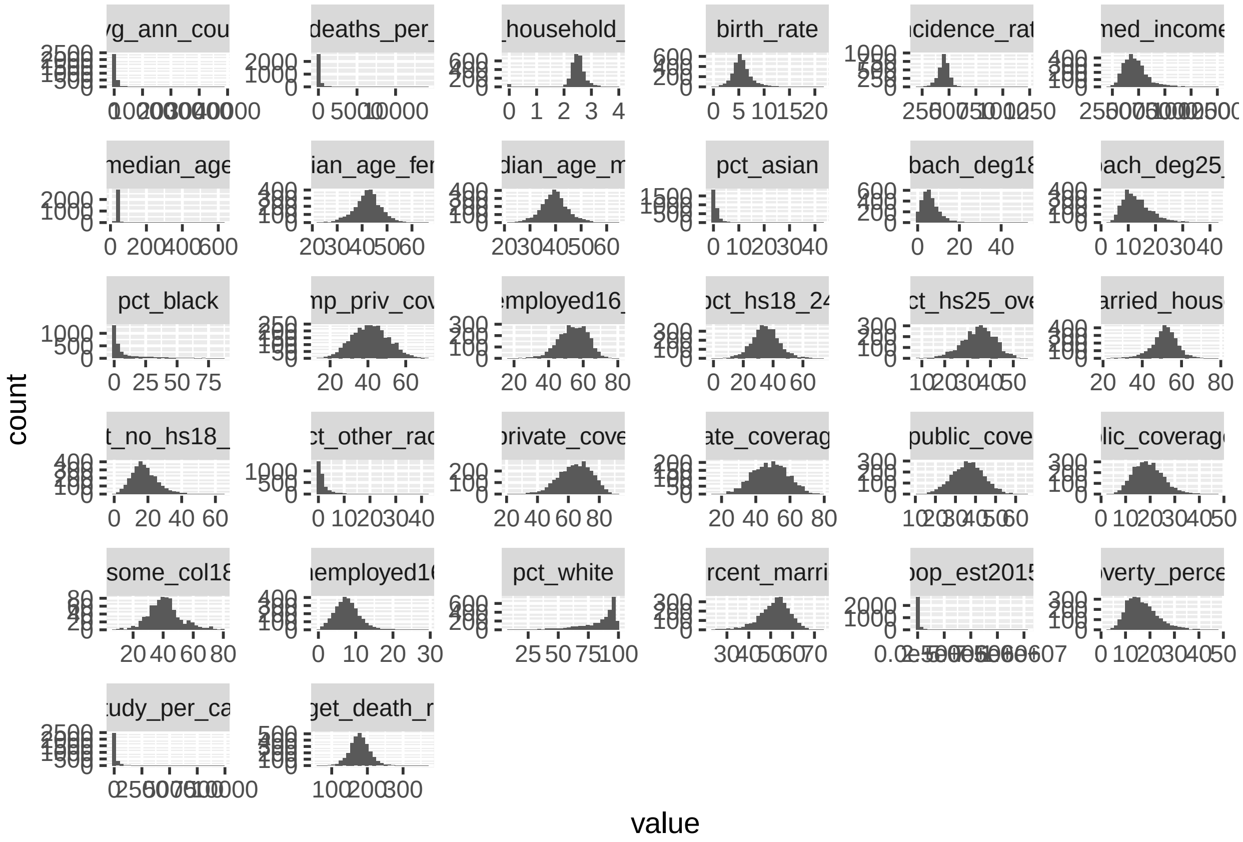
- **pct_other_race:** percent of county residents who identify in a category which is not White, Black, or Asian (b)

- **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_fe
  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
```

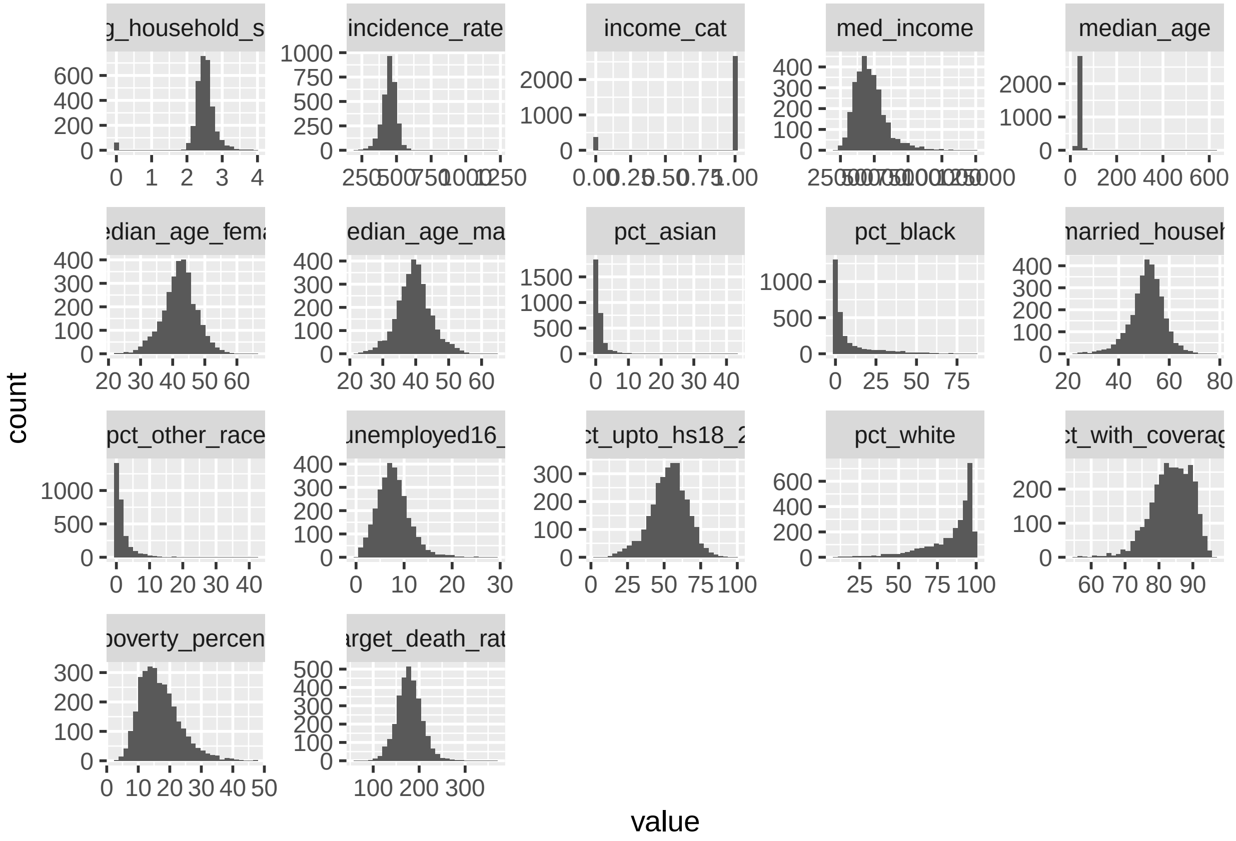
Check correlation and distribution:

```
cor(cancer_county) %>%
  knitr::kable()
```

	target_death_rate	incidence_rate	med_income	poverty_percent	median_age	median
target_death_rate	1.0000000	0.4494317	-0.4286149	0.4293890	0.0043751	
incidence_rate	0.4494317	1.0000000	-0.0010362	0.0090463	0.0180892	
med_income	-0.4286149	-0.0010362	1.0000000	-0.7889652	-0.0132877	
poverty_percent	0.4293890	0.0090463	-0.7889652	1.0000000	-0.0292800	
median_age	0.0043751	0.0180892	-0.0132877	-0.0292800	1.0000000	

	target_death_rate	incidence_rate	med_income	poverty_percent	median_age	median
median_age_male	-0.0219294	-0.0147332	-0.0916626	-0.2140010	0.1291195	
median_age_female	0.0120484	-0.0091056	-0.1532784	-0.1481635	0.1246784	
avg_household_size	-0.0369053	-0.1184000	0.1120653	0.0743076	-0.0319441	
pct_unemployed16_over	0.3784124	0.0999795	-0.4531077	0.6551481	0.0185904	
pct_white	-0.1774000	-0.0145098	0.1672254	-0.5094328	0.0350094	
pct_black	0.2570236	0.1134890	-0.2702316	0.5115297	-0.0171732	
pct_asian	-0.1863311	-0.0081234	0.4258442	-0.1572887	-0.0384239	
pct_other_race	-0.1898936	-0.2087483	0.0836349	0.0470959	-0.0302765	
pct_married_households	-0.2933253	-0.1521763	0.4460829	-0.6049528	0.0145036	
pct_upto_hs18_24	0.2443042	-0.0929669	-0.3212077	0.2517431	0.0401926	
pct_with_coverage	-0.2292798	0.2302489	0.5566583	-0.6516658	0.0049621	
income_cat	-0.3030288	0.0110839	0.4765990	-0.6344122	0.0103377	

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



The discriptive statistics:

```
state_summary = function(x){
mean = mean(x)
max = max(x)
min = min(x)
median = median(x)
var = var(x)
sd = sd(x)
sample_size=length(x)-sum(is.na(x))
tibble(mean, max, min, median, var, sd, sample_size)
}

df_target_death_rate <-state_summary(cancer_county$target_death_rate)
df_incidence_rate <-state_summary(cancer_county$incidence_rate)
df_med_income <-state_summary(cancer_county$med_income)
df_poverty_percent<-state_summary(cancer_county$poverty_percent)
df_median_age<-state_summary(cancer_county$median_age)
df_median_agemale<-state_summary(cancer_county$median_age_male)
df_median_agefemale<-state_summary(cancer_county$median_age_female)
df_avg_household_size<-state_summary(cancer_county$avg_household_size)
df_pct_unemployed16_over<-state_summary(cancer_county$pct_unemployed16_over)
df_pct_white<-state_summary(cancer_county$pct_white)
df_pct_black<-state_summary(cancer_county$pct_black)
df_pct_asian<-state_summary(cancer_county$pct_asian)
df_pct_other_race<-state_summary(cancer_county$pct_other_race)
df_pct_married_households<-state_summary(cancer_county$pct_married_households)
df_pct_upto_hs18_24<-state_summary(cancer_county$pct_upto_hs18_24)
df_pct_with_coverage<-state_summary(cancer_county$pct_with_coverage)

state_des <- bind_rows(df_target_death_rate,
                       df_incidence_rate,
                       df_med_income,
                       df_poverty_percent,
                       df_median_age,
                       df_median_agemale,
                       df_median_agefemale,
                       df_avg_household_size,
                       df_pct_unemployed16_over,
                       df_pct_white,
                       df_pct_black,
                       df_pct_asian,
                       df_pct_other_race,
                       df_pct_married_households,
                       df_pct_upto_hs18_24,
                       df_pct_with_coverage)
variable<- c("target_death_rate", "incidence_rate","med_income", "poverty_percent", "median_age", "medi

state_wholedes <- cbind(variable, state_des)

knitr::kable(state_wholedes)
```

variable	mean	max	min	median	var	sd
target_death_rate	178.664063	362.80000	59.70000	1.781000e+02	7.701464e+02	2.775151e+01
incidence_rate	448.268586	1206.90000	201.30000	4.535494e+02	2.976874e+03	5.456073e+01

variable	mean	max	min	median	var	sd
med_income	47063.281917	125635.00000	22640.00000	4.520700e+04	1.449638e+08	1.204009e+04
poverty_percent	16.878175	47.40000	3.20000	1.590000e+01	4.107639e+01	6.409087e+00
median_age	45.272333	624.00000	22.30000	4.100000e+01	2.052496e+03	4.530448e+01
median_agemale	39.570725	64.70000	22.40000	3.960000e+01	2.731125e+01	5.226017e+00
median_agefemale	42.145323	65.70000	22.30000	4.240000e+01	2.801425e+01	5.292849e+00
avg_household_size	2.479662	3.97000	0.02210	2.500000e+00	1.841906e-01	4.291744e-01
pct_unemployed16_over	7.852412	29.40000	0.40000	7.600000e+00	1.191886e+01	3.452371e+00
pct_white	83.645286	100.00000	10.19916	9.005977e+01	2.683052e+02	1.638003e+01
pct_black	9.107978	85.94780	0.00000	2.247576e+00	2.112528e+02	1.453454e+01
pct_asian	1.253965	42.61942	0.00000	5.498117e-01	6.813543e+00	2.610276e+00
pct_other_race	1.983523	41.93025	0.00000	8.261852e-01	1.237428e+01	3.517710e+00
pct_married_households	51.243872	78.07540	22.99249	5.166994e+01	4.320188e+01	6.572814e+00
pct_upto_hs18_24	53.226518	100.00000	4.80000	5.390000e+01	1.601814e+02	1.265628e+01
pct_with_coverage	83.595011	95.70000	54.60000	8.400000e+01	3.536646e+01	5.946971e+00

Model building:

```r
# building full model
full_model <- lm(target_death_rate ~., data = cancer_county)
summary(full_model)
```

```
##
## Call:
## lm(formula = target_death_rate ~ ., data = cancer_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -132.811  -11.710   -0.008   11.850  129.454
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.030e+02  1.320e+01   7.808 7.96e-15 ***
## incidence_rate          2.195e-01  7.485e-03  29.328  < 2e-16 ***
## med_income             -4.878e-04  6.762e-05  -7.215 6.81e-13 ***
## poverty_percent         3.101e-01  1.559e-01   1.989 0.046756 *
## median_age             -4.321e-03  8.324e-03  -0.519 0.603763
## median_age_male        -2.500e-01  2.105e-01  -1.188 0.235102
## median_age_female      -1.204e-01  2.086e-01  -0.577 0.563802
## avg_household_size      5.976e-01  1.004e+00   0.595 0.551774
## pct_unemployed16_over   8.448e-01  1.524e-01   5.544 3.22e-08 ***
## pct_white              -5.202e-03  5.868e-02  -0.089 0.929369
## pct_black              -1.164e-02  5.655e-02  -0.206 0.836891
## pct_asian              -1.954e-01  1.873e-01  -1.043 0.296831
## pct_other_race         -9.198e-01  1.235e-01  -7.446 1.24e-13 ***
## pct_married_households -9.767e-02  8.927e-02  -1.094 0.273994
## pct_upto_hs18_24        3.827e-01  3.795e-02  10.085  < 2e-16 ***
## pct_with_coverage      -7.372e-02  1.022e-01  -0.721 0.470747
## income_cat             -5.252e+00  1.503e+00  -3.493 0.000484 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 20.6 on 3030 degrees of freedom
## Multiple R-squared:  0.4521, Adjusted R-squared:  0.4492
## F-statistic: 156.3 on 16 and 3030 DF,  p-value: < 2.2e-16
```

```
# Using the stepwise
stepwise_model <- stepAIC(full_model, direction = "both", trace = FALSE)
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + med_income +
##     poverty_percent + median_age_male + pct_unemployed16_over +
##     pct_other_race + pct_upto_hs18_24 + income_cat, data = cancer_county)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -132.047  -11.853   -0.066   11.894  129.669
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.298e+01  8.056e+00  11.541  < 2e-16 ***
## incidence_rate         2.187e-01  7.094e-03  30.829  < 2e-16 ***
## med_income            -5.160e-04  5.771e-05  -8.941  < 2e-16 ***
## poverty_percent        3.621e-01  1.403e-01   2.582 0.009882 **
## median_age_male       -3.904e-01  8.713e-02  -4.480 7.74e-06 ***
## pct_unemployed16_over  8.733e-01  1.464e-01   5.965 2.73e-09 ***
## pct_other_race        -8.969e-01  1.141e-01  -7.861 5.26e-15 ***
## pct_upto_hs18_24       3.894e-01  3.262e-02  11.937  < 2e-16 ***
## income_cat            -5.176e+00  1.493e+00  -3.466 0.000535 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.58 on 3038 degrees of freedom
## Multiple R-squared:  0.4514, Adjusted R-squared:   0.45
## F-statistic: 312.5 on 8 and 3038 DF,  p-value: < 2.2e-16
```

```
vif(stepwise_model)
```

```
##        incidence_rate            med_income       poverty_percent
##              1.077115              3.471605              5.810092
##       median_age_male pct_unemployed16_over        pct_other_race
##              1.491046              1.837179              1.158462
##      pct_upto_hs18_24            income_cat
##              1.225470              1.742843
```

```
# Cp and AIC and Adjusted R2

model_dig <- glance(stepwise_model) %>%
  as.data.frame() %>%
  dplyr::select(adj.r.squared, sigma, p.value, AIC, BIC) %>%
  rename(RES = sigma) %>%
  mutate(cp = ols_mallows_cp(stepwise_model, full_model))

model_dig
```

```
##   adj.r.squared     RES p.value      AIC      BIC       cp
## 1      0.4499715 20.5816       0 27088.68 27148.9 4.689769
```