Exploratory Data Analysis

```
rm(list = ls())
require(readr)
```

```
## Loading required package: readr
```

```
require(tidyr)
```

```
## Loading required package: tidyr
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(knitr)
```

```
## Loading required package: knitr
```

```
data <- read.csv("data/processed/data.csv")

# We remove these columns because we want to be predicting the turnout
# rate solely from county demographics
data <- data %>% select(-c(X, State, County, fips))
head(data)
```

```
##   frac_coll_plus2010 foreign_share2010 med_hhinc2016 poor_share2010
## 1         0.22199036       0.020154603      54052.80      0.1059177
## 2         0.26071036       0.037591625      52003.09      0.1229422
## 3         0.13349621       0.028143950      33114.85      0.2506308
## 4         0.09924053       0.006859188      39846.45      0.1268499
## 5         0.12633450       0.047343444      46361.12      0.1331379
## 6         0.10972187       0.013493270      31304.78      0.2804486
##   share_white2010 share_black2010 share_hisp2010 share_asian2010
## 1       0.7724616      0.18134174     0.02400542    0.0078302799
## 2       0.8350479      0.09752284     0.04384824    0.0059535136
## 3       0.4675311      0.47190151     0.05051535    0.0036882064
## 4       0.7502073      0.22282349     0.01771765    0.0007418721
## 5       0.8888734      0.01500297     0.08070200    0.0018735955
## 6       0.2191680      0.70221734     0.07119296    0.0017932489
##   gsmn_math_g3_2013 rent_twobed2015 singleparent_share2010 traveltime15_2010
## 1          2.759864        739.3654              0.2833759         0.2041625
```

```
## 2          2.792510        816.8452              0.2778664         0.2753262
## 3          1.600009        527.2908              0.4680706         0.3760492
## 4          1.531674        604.2776              0.3201363         0.2526830
## 5          2.815403        567.6959              0.2589052         0.1943438
## 6          1.039439        266.0000              0.5778636         0.3921350
##     emp2000 ln_wage_growth_hs_grad popdensity2010 ann_avg_job_growth_2004_2013
## 1 0.6095865            -0.06331379       91.80268                   0.010145103
## 2 0.5770263             0.03009291      114.64751                   0.012950056
## 3 0.4532710             0.18936642       31.02921                  -0.020755908
## 4 0.4942406            -0.02007263       36.80634                  -0.004644653
## 5 0.5778096             0.09646260       88.90219                  -0.008120399
## 6 0.3746639             0.36383346       17.52395                   0.026254078
##   job_density_2013 turnout.rate
## 1        40.719135    0.6618366
## 2        50.085987    0.6529056
## 3         9.230672    0.5402712
## 4        12.875392    0.5456975
## 5        36.175354    0.6419098
## 6         6.954023    0.5908043
```

We have no categorical variables. For each of our continuous variables, we summarize the number of missing values, the mean, median, standard deviation, and interquartile range.

```r
predictors <- names(data)[names(data) != "turnout.rate"]
summary_table <- data.frame()

for (predictor in predictors) {
  column <- data[[predictor]]
  num_non_missing <- sum(!is.na(column))
  num_missing <- sum(is.na(column))
  mean_var <- mean(column, na.rm = TRUE)
  median_var <- median(column, na.rm = TRUE)
  sd_var <- sd(column, na.rm = TRUE)
  iqr_var <- IQR(column, na.rm = TRUE)

  summary_table <- rbind(summary_table, data.frame(
    Variable = predictor,
    Non_Missing = num_non_missing,
    Missing = num_missing,
    Mean = round(mean_var, 2),
    Median = round(median_var, 2),
    SD = round(sd_var, 2),
    IQR = round(iqr_var, 2)
  ))
}

kable(summary_table)
```

Variable	Non_Missing	Missing	Mean	Median	SD	IQR
frac_coll_plus2010	3220	43	0.19	0.17	0.09	0.09
foreign_share2010	3142	121	0.04	0.02	0.06	0.04
med_hhinc2016	3219	44	48259.87	46718.22	14039.43	15020.15
poor_share2010	3220	43	0.16	0.15	0.08	0.08
share_white2010	3220	43	0.76	0.85	0.23	0.29

Variable	Non_Missing	Missing	Mean	Median	SD	IQR
share_black2010	3220	43	0.09	0.02	0.14	0.10
share_hisp2010	3220	43	0.10	0.03	0.19	0.07
share_asian2010	3199	64	0.01	0.00	0.02	0.01
gsmn_math_g3_2013	3069	194	3.21	3.24	0.78	0.98
rent_twobed2015	3143	120	684.90	637.72	208.21	196.55
singleparent_share2010	3219	44	0.31	0.30	0.09	0.11
traveltime15_2010	3220	43	0.40	0.38	0.14	0.19
emp2000	3142	121	0.57	0.58	0.08	0.10
ln_wage_growth_hs_grad	2535	728	0.08	0.07	0.14	0.13
popdensity2010	3219	44	286.81	46.87	1772.48	111.93
ann_avg_job_growth_2004_2013	3214	49	0.00	0.00	0.02	0.02
job_density_2013	3217	46	129.48	19.23	855.52	48.72

We also check that our hypothesis that the turnout rate can be predicted from county demographics is reasonable by fitting a linear regression model.

```
lm_model <- lm(turnout.rate ~ ., data = data)
summary(lm_model)
```

```
##
## Call:
## lm(formula = turnout.rate ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69711 -0.04665 -0.00173  0.04503  1.63048
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.217e-01  4.487e-02  13.855  < 2e-16 ***
## frac_coll_plus2010            2.512e-01  3.239e-02   7.754 1.32e-14 ***
## foreign_share2010             3.695e-01  5.967e-02   6.193 6.97e-10 ***
## med_hhinc2016                 4.473e-07  3.182e-07   1.406   0.1600
## poor_share2010               -5.593e-01  5.331e-02 -10.490  < 2e-16 ***
## share_white2010               4.769e-03  3.098e-02   0.154   0.8777
## share_black2010               2.795e-02  3.008e-02   0.929   0.3530
## share_hisp2010               -1.667e-01  3.487e-02  -4.780 1.86e-06 ***
## share_asian2010              -1.274e-01  1.112e-01  -1.145   0.2523
## gsmn_math_g3_2013            -1.736e-03  2.841e-03  -0.611   0.5413
## rent_twobed2015              -3.942e-05  1.800e-05  -2.190   0.0286 *
## singleparent_share2010       -6.541e-02  3.207e-02  -2.039   0.0415 *
## traveltime15_2010            -1.317e-03  1.527e-02  -0.086   0.9313
## emp2000                       1.639e-01  3.530e-02   4.642 3.64e-06 ***
## ln_wage_growth_hs_grad       -3.243e-02  1.351e-02  -2.401   0.0164 *
## popdensity2010               -2.486e-06  6.192e-06  -0.402   0.6880
## ann_avg_job_growth_2004_2013 -6.911e-01  1.357e-01  -5.091 3.84e-07 ***
## job_density_2013             -2.688e-06  1.274e-05  -0.211   0.8329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08167 on 2320 degrees of freedom
##   (925 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.3897, Adjusted R-squared:  0.3852
## F-statistic: 87.14 on 17 and 2320 DF,  p-value: < 2.2e-16
```