Cancer EDA_Tim_Mike_Craig_W203_4

Tim, Mike, Craig, Wei

1/21/2018

Introduction

This report presents an initial exporatory data analysis identifying the key features associated with cancer rates and deaths based on geographic location in the form of a county in the United States. The goal is to use the findings from the analysis to develop strategies to improve future cancer outcomes. We are a team of data scientists motivated to in promote understanding the societal factors that impact mortality rates of cancer among various communities in the United States. We are grateful for the grant awarded to us by a health government agency to complete this study.

Research Question

Our task was to answer these two key research quesitons:

1. What are the key county level characteristics associated with mortality rates from cancer?
2. Are there trends in county level characteristics that can be identified to inform social intervention to decrease cancer mortality rates?

About the Data

```r
df <- read.csv("cancer.csv")
class(df)  # dimensions are 3047 rows x 30 columns
## [1] "data.frame"
dim(df)  # dimensions of the data set
## [1] 3047   30
```

```r
# sum((df$MedianAge>300), na.rm=TRUE) # identifies 30 observations median age
# greater than 300
df$MedianAge_gr100 = 0  #creates a new column to save the dubious data
df$MedianAge_gr100[df$MedianAge > 300] <- df$MedianAge[df$MedianAge > 300]  # saves the dubious data
df$MedianAge[df$MedianAge > 300] <- NA  #update the Median Age
```

Analysis of Key Relationships

Table below show below shows the correlation coeffiencients for r

```r
x <- names(df)[!names(df) %in% c("deathRate", "X")]

vars <- c()
cors <- c()

for (i in x) {
    if (class(df[, i]) == "numeric") {
        vars <- c(vars, i)
        cors <- c(cors, round(cor(df[, i], df$deathRate, use = "complete.obs"), digits = 4))
    }
}

cor_df <- data.frame(vars, cors, row.names = NULL)
cor_df <- cor_df[order(abs(cor_df$cors), decreasing = T), ]

kable(cor_df, col.names = c("Variable", "Correlation Coefficient"), row.names = F,
    caption = "Correlation Coefficients of Indepent Variables to Death Rate")
```

Table 1: Correlation Coefficients of Indepent Variables to Death Rate

Variable	Correlation Coefficient
PctBachDeg25_Over	-0.4855
povertyPercent	0.4294
PctEmployed16_Over	-0.4120
PctHS25_Over	0.4046
PctPublicCoverage	0.4046
PctPrivateCoverage	-0.3861
PctUnemployed16_Over	0.3784
PctMarriedHouseholds	-0.2933
PctBachDeg18_24	-0.2878
PctEmpPrivCoverage	-0.2674
PercentMarried	-0.2668
PctHS18_24	0.2620
PctBlack	0.2570
PctOtherRace	-0.1899
PctSomeCol18_24	-0.1887
PctAsian	-0.1863
PctWhite	-0.1774
avgAnnCount	-0.1435
PctNoHS18_24	0.0885
BirthRate	-0.0874
AvgHouseholdSize	-0.0369
MedianAgeMale	-0.0219
MedianAgeFemale	0.0120
MedianAge_gr100	0.0050
MedianAge	-0.0043

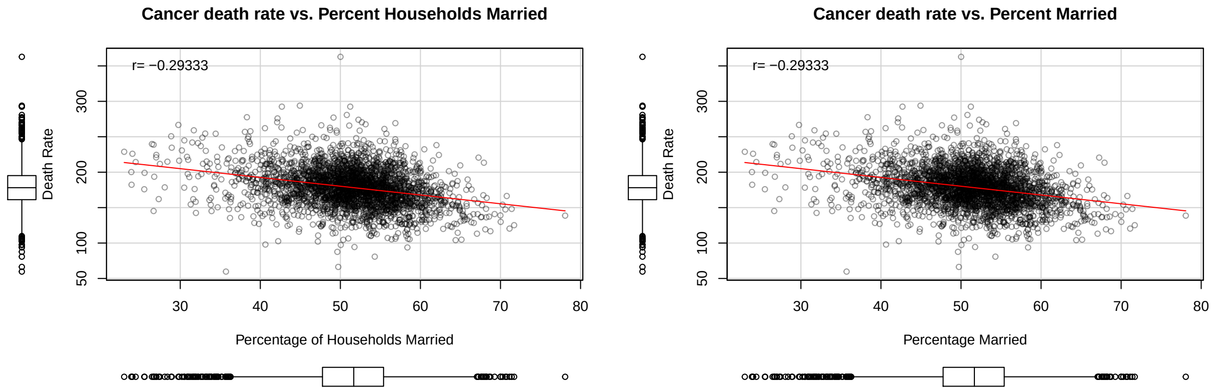
MIKE TO ADD HIS SECTION HERE

Marriage vs. Death Rate

Marriage seems to have a slight negative relationship with cancer death rates. The dataset had two variables to explore: *PctMarried* and *PercentMarriedHouseholds*. Unsurprisingly, these variables have correlations very close to each other, with *PctMarried* having -0.293 and *PctMarriedHouseholds* having -0.293. This may suggest that partner support leads to better outcomes with cancer mortality and would be worth investigating further.

```
par(mfrow = c(1, 2))
r_cor = round(cor(df$deathRate, df$PctMarriedHouseholds, use = "complete.obs"), 5)
scatterplot(df$PctMarriedHouseholds, df$deathRate, xlab = "Percentage of Households Married",
    ylab = "Death Rate", main = "Cancer death rate vs. Percent Households Married",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))

r_cor = round(cor(df$deathRate, df$PctMarried, use = "complete.obs"), 5)
scatterplot(df$PctMarried, df$deathRate, xlab = "Percentage Married", ylab = "Death Rate",
    main = "Cancer death rate vs. Percent Married", legend("topleft", bty = "n",
        legend = paste("r=", r_cor)), col = c("red", "green", rgb(0, 0, 0, 100, maxColorValue = 255)))
```



Race vs. Death Rate

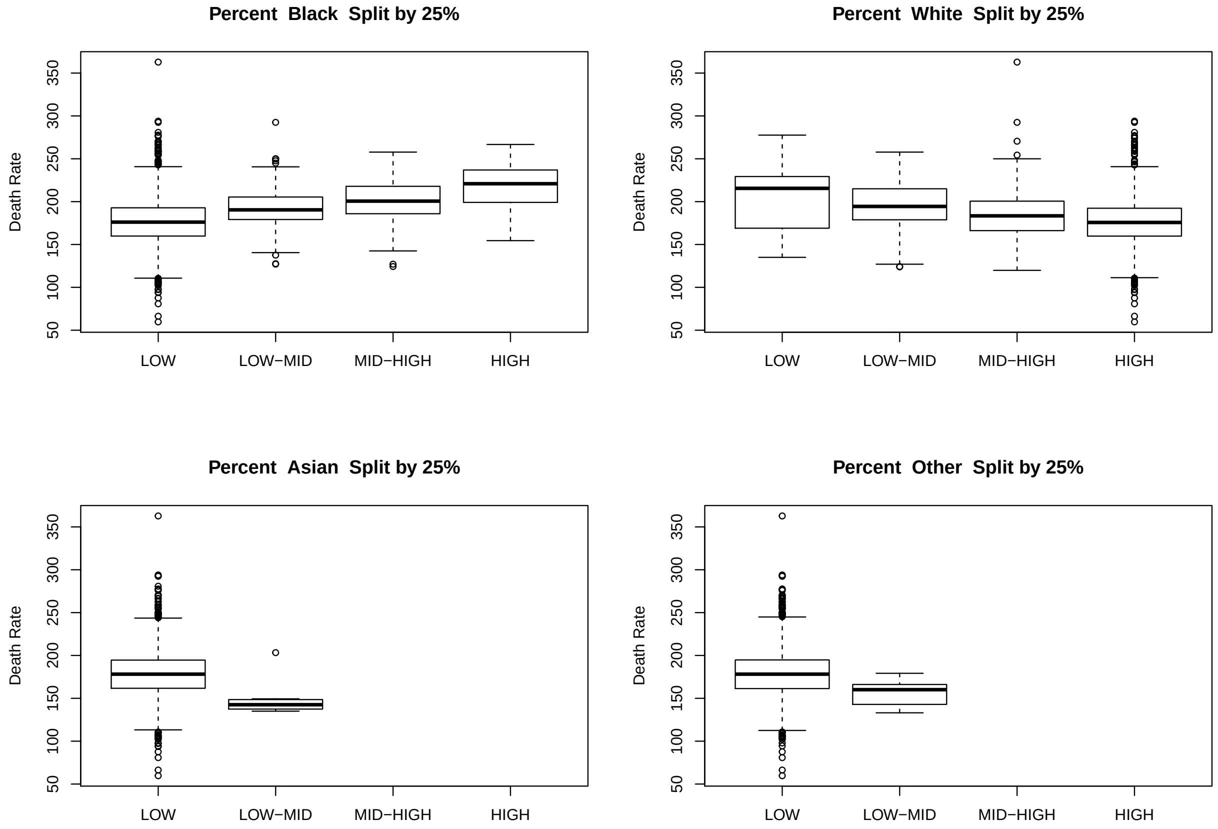
Race has some interesting interactions with Death Rate in the fact that relationships differ with the prominence of a single race in a county. All races except for Black have a negative correlation, though a weak one, with the death rate. Conversely, the Black has a positive correlation that is stronger than the other correlations. We can gain the most insight looking at white and black communities, because generally these two are more prominent. In our sample, there is not enough data to gain any valuable insight into prominent Asian communities or communities of other races.

The race percentage variables have signficant skew to their distributions, so to better represent this data, have chosen to display them in box plots with binned percentages of each quartile. This allows us to examine how the mean changes through each quartile and better see the relationship as the concentration of each race in a county grows.

```
par(mfrow = c(2, 2))
bplot_race <- function(race, racec) {
    race_bin = cut(race, breaks = c(0, 25, 50, 75, 100), label = c("LOW", "LOW-MID",
        "MID-HIGH", "HIGH"))

    boxplot(df$deathRate ~ race_bin, data = df, main = paste("Percent ", racec, " Split by 25%"),
        ylab = "Death Rate")
}

bplot_race(df$PctBlack, "Black")
bplot_race(df$PctWhite, "White")
bplot_race(df$PctAsian, "Asian")
bplot_race(df$PctOther, "Other")
```



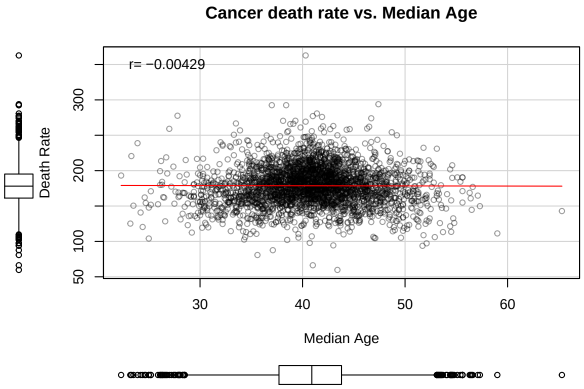
Age vs. Death Rate

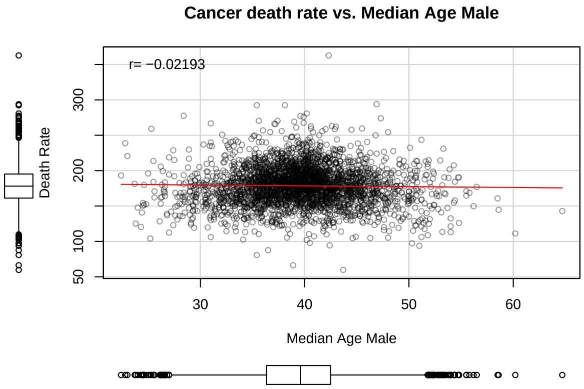
The interesting result about age compared to death rate is that the correlation is rather weak. One guess that as a community ages, so do the ages. Our sample contradicts this assumption, with the Median Age having a correlation of only -0.00429. We see similar values in Median Age of Males (-0.02193) and Median Age of Females (0.01205).

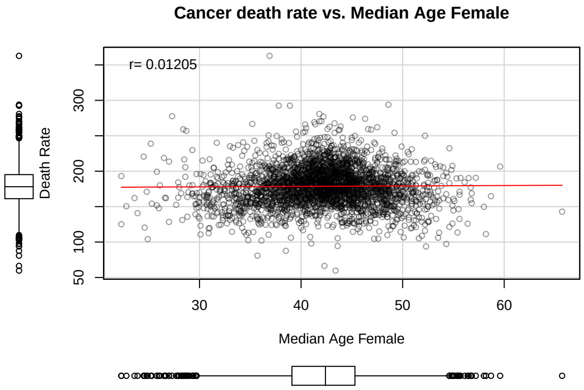
```r
par(mfrow = c(1, 3))
r_cor = round(cor(df$deathRate, df$MedianAge, use = "complete.obs"), 5)
scatterplot(df$MedianAge, df$deathRate, xlab = "Median Age", ylab = "Death Rate",
    main = "Cancer death rate vs. Median Age", legend("topleft", bty = "n", legend = paste("r=",
        r_cor)), col = c("red", "green", rgb(0, 0, 0, 100, maxColorValue = 255)))

r_cor = round(cor(df$deathRate, df$MedianAgeMale, use = "complete.obs"), 5)
scatterplot(df$MedianAgeMale, df$deathRate, xlab = "Median Age Male", ylab = "Death Rate",
    main = "Cancer death rate vs. Median Age Male", legend("topleft", bty = "n",
        legend = paste("r=", r_cor)), col = c("red", "green", rgb(0, 0, 0, 100, maxColorValue = 255)))

r_cor = round(cor(df$deathRate, df$MedianAgeFemale, use = "complete.obs"), 5)
scatterplot(df$MedianAgeFemale, df$deathRate, xlab = "Median Age Female", ylab = "Death Rate",
    main = "Cancer death rate vs. Median Age Female", legend("topleft", bty = "n",
        legend = paste("r=", r_cor)), col = c("red", "green", rgb(0, 0, 0, 100, maxColorValue = 255)))
```







END MIKE SECITON