Exploratory Data Analysis - Incidents of Cancer

Jonathan D'Souza

9/15/2018

Introduction

Given a Data Set for cancer incidences for a select group of counties.... this study attempts to explore the relationships between the outcome variable : Annual Incident Count and other key independent variables.

```
cancer<-read.csv("cancer.csv")  #Assumes file in current working directory
names(cancer)
```

```
##  [1] "X"                    "avgAnnCount"          "medIncome"
##  [4] "popEst2015"           "annCOuntpercent"      "povertyPercent"
##  [7] "binnedInc"            "MedianAge"            "MedianAgeMale"
## [10] "MedianAgeFemale"      "Geography"            "AvgHouseholdSize"
## [13] "PercentMarried"       "PctNoHS18_24"         "PctHS18_24"
## [16] "PctSomeCol18_24"      "PctBachDeg18_24"      "PctHS25_Over"
## [19] "PctBachDeg25_Over"    "PctEmployed16_Over"   "PctUnemployed16_Over"
## [22] "PctPrivateCoverage"   "PctEmpPrivCoverage"   "PctPublicCoverage"
## [25] "PctWhite"             "PctBlack"             "PctAsian"
## [28] "PctOtherRace"         "PctMarriedHouseholds" "BirthRate"
## [31] "deathRate"
```

```
nrow(cancer)
```

```
## [1] 3047
```

Annual Incident Count is better expressed as percent of population

```
summary(cancer$avgAnnCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.0    76.0   171.0   606.3   518.0 38150.0
```

```
cancer$AnnCountPercent<-with(cancer,100*avgAnnCount/popEst2015)
```

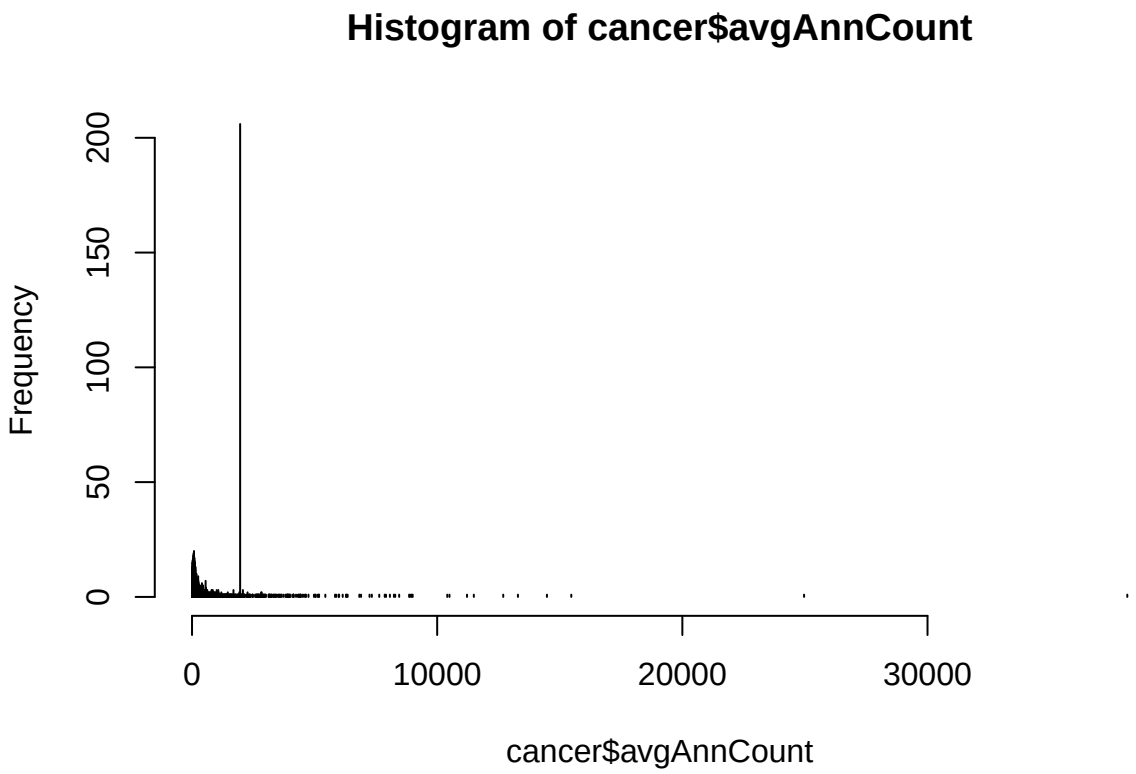
Univariate Analysis of Key Variables

It is clear that the annual count percent has some outliers given that max % >100 (can't be more incidents than the population) Plotting the Avg annual count shows a big spike in values

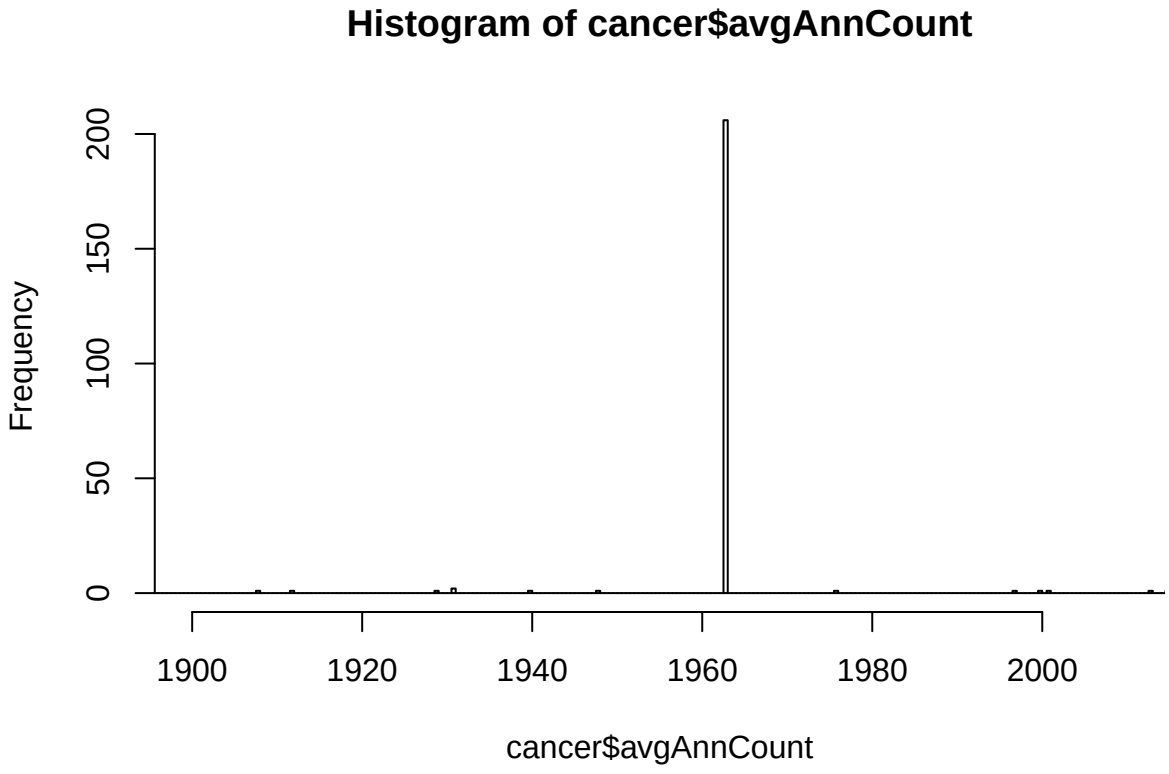
```
summary (cancer$AnnCountPercent)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.09281   0.48020   0.56240   2.32400   0.64870 236.80000
```

```
hist(cancer$avgAnnCount,100000)
```

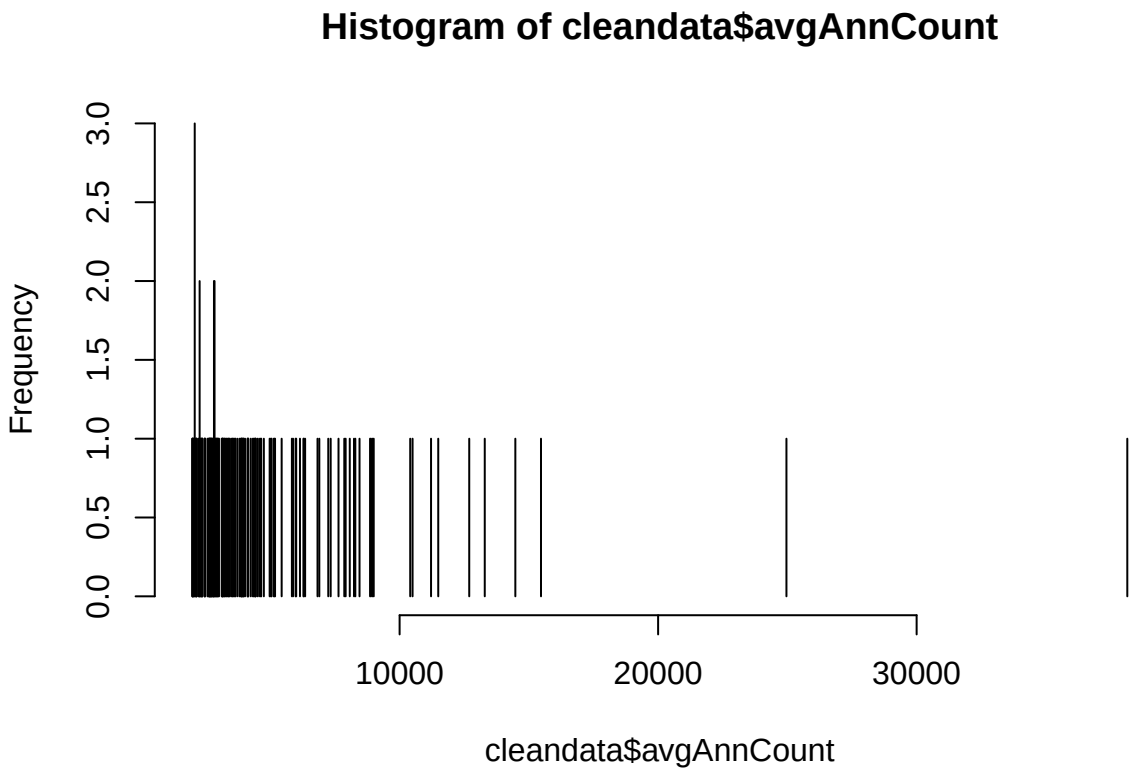



```
#Try with smaller range
hist(cancer$avgAnnCount,100000,xlim=c(1900,2010))
```



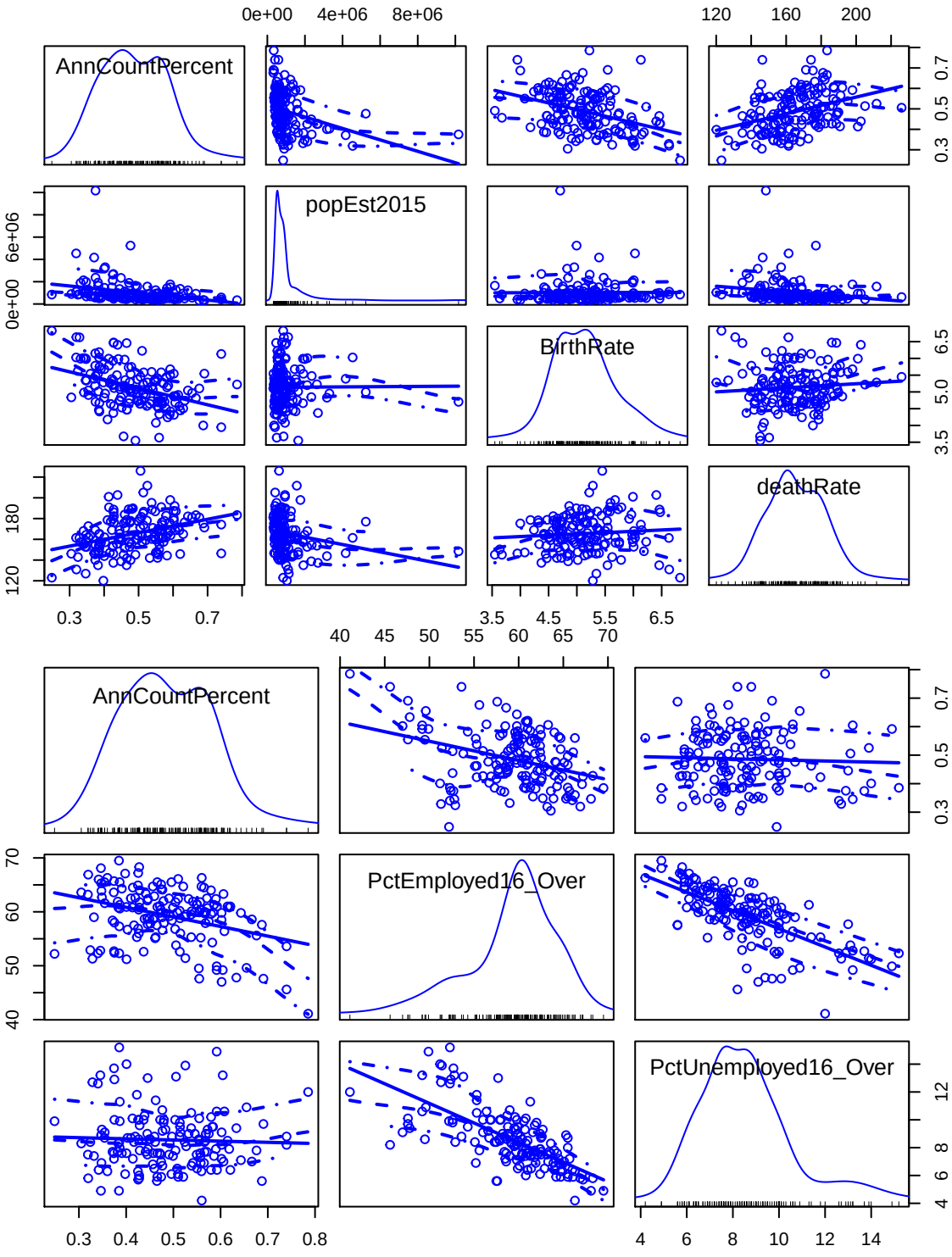

```
#Get these outlier values
```

```r
cleandata<-subset(cancer,avgAnnCount>1970 & avgAnnCount>1960)
hist(cleandata$avgAnnCount,100000)
```



Analysis of Key Relationships

Explore how your outcome variable is related to the other variables in your dataset. Make sure to use visualizations to understand the nature of each bivariate relationship. What transformations can you apply to clarify the relationships you see in the data? Be sure to justify each transformation you use.



Analysis of Secondary Effects (10 pts)

What secondary variables might have confounding effects on the relationships you have identified? Ex- plain how these variables affect your understanding of the data.

Conclusion (20 pts)

Summarize your exploratory analysis. What can you conclude based on your analysis? 2