

Your Document Title

Author Name

16 April, 2024

Introduction

We as human beings like to divide actions into moral and amoral. Some of these amoral actions, such as murder, theft and fraud, are codified in law and we call those transgressions crimes. It is in our best interest to avoid crime happening. Therefore, we try to punish those that commit them in hopes of it being a deterrent for future repetition. However, there might be other ways in which to minimise crime. In addition, we see that crime isn't happening uniformly everywhere. This leads us to the question: what makes a crime more likely to happen in a region?

Clay[1857], concentrating on crime in England, analysed the following characteristics: Beer-(ale)-house numbers, worship attendance and public school attendance. The argument he provides for the latter two, is both allow for the public access to Christian moral values (although he criticises that the ability to read on its own is not enough). It is to note, the “public schools” mentioned by Clay concentrated mostly on teaching read and writing and were not free but accepted anyone who could effort it. In the case of beer-houses, he argues “the temptation to animal pleasure” corrupts a person. To help his case he tries to show the positive correlation of beer-houses, the negative correlations of public school and worship attendance to crime.

In this paper we want to verify (or deny) these effects of the aforementioned characteristics and try to predict crime rate from them using the same data set. However, before starting it is important to mention, as noticed by Clay[1857], the recorded crime rate may be inaccurate because of systematic reasons since each county handles crime and punishment differently on the executive and juridical levels. This will impact the accuracy of all possible analysis and models.

Exploratory Data Analysis

We begin our analysis with a look at the first few lines of our dataset.

First few data elements:

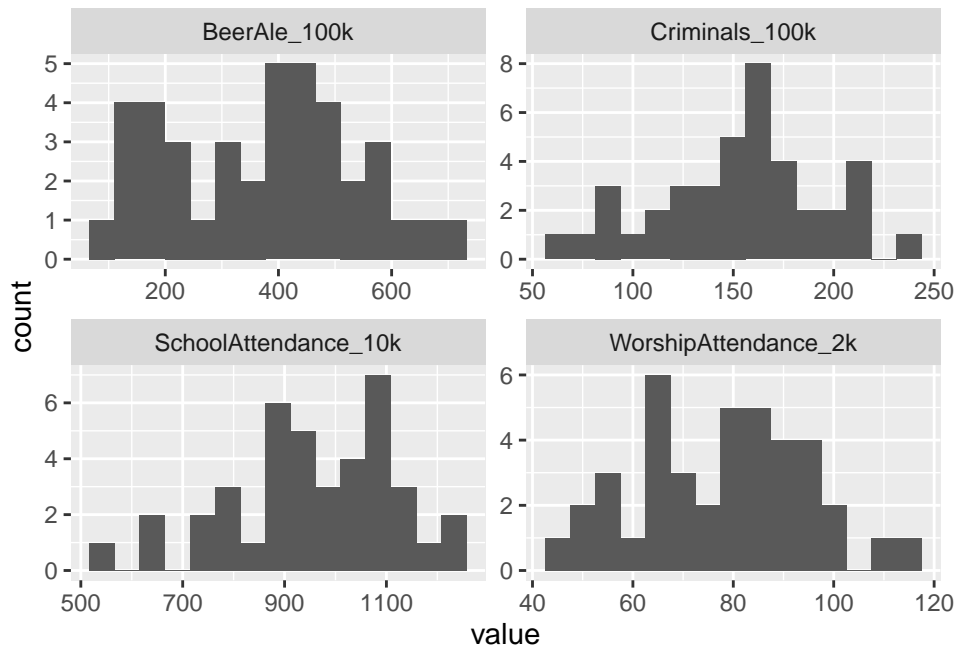
County	RegionName	RegionCode	Criminals_100k	BeerAle_100k	SchoolAttendance_10k	WorshipAttendance_2k
Middlesex	SouthEastern	1	200	541	560	43
Surrey	SouthEastern	1	160	504	630	48
Kent	SouthEastern	1	160	552	790	68
Sussex	SouthEastern	1	147	295	820	67

County	RegionName	RegionCode	Criminals_100k	BeerAle_100k	SchoolAttendance_10k	WorshipAttendance_2k
Hants	SouthEastern	1	178	409	990	79
Berks	SouthEastern	1	205	568	930	69

Data summary:

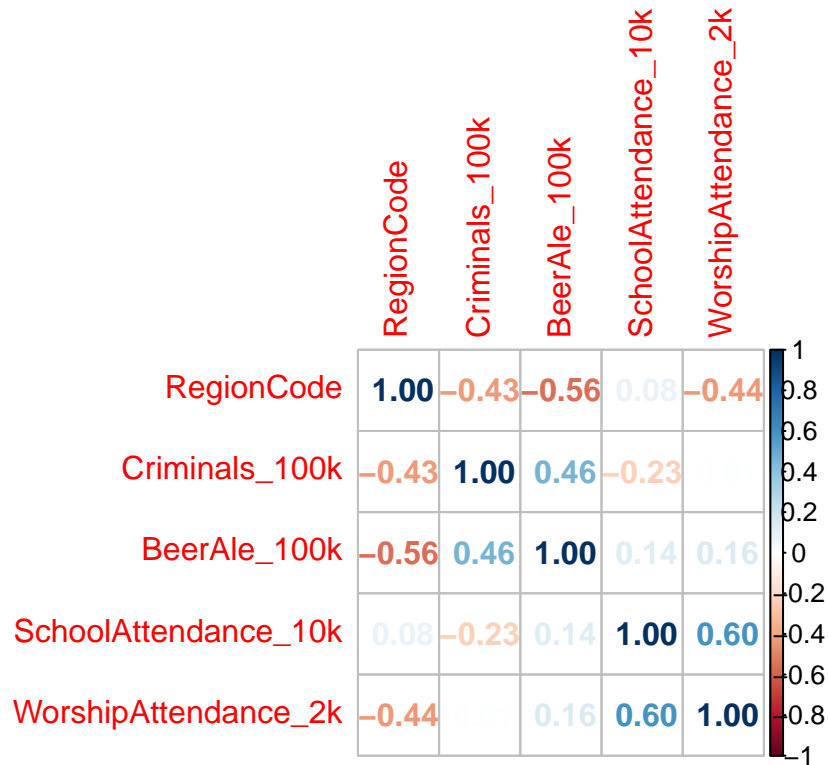
```
##      County      RegionName      RegionCode  Criminals_100k
##  Length:40      Length:40      Min.    :1.00  Min.    : 66.0
##  Class :character  Class :character  1st Qu.:1.00  1st Qu.:127.0
##  Mode  :character  Mode  :character  Median :3.00  Median :157.5
##                                     Mean  :3.45  Mean  :152.9
##                                     3rd Qu.:5.00  3rd Qu.:174.2
##                                     Max.  :8.00  Max.  :241.0
##  BeerAle_100k  SchoolAttendance_10k  WorshipAttendance_2k
##  Min.    : 87.0  Min.    : 560.0  Min.    : 43.0
##  1st Qu.:209.0  1st Qu.: 880.0  1st Qu.: 65.0
##  Median :407.0  Median : 965.0  Median : 79.5
##  Mean   :374.9  Mean   : 957.8  Mean   : 77.5
##  3rd Qu.:490.8  3rd Qu.:1082.5  3rd Qu.: 91.0
##  Max.   :708.0  Max.   :1250.0  Max.   :113.0
```

The dataset contains information about 40 different counties. `County` and `RegionName` are categorical variables, and each of the 8 regions is assigned a `RegionCode`, which is a number from 1 to 8. `Criminals_100k` is the number of criminals per 100,000 inhabitants for any particular county. Similarly, `BeerAle_100k`, `SchoolAttendance_10k` and `WorshipAttendance_2k` are social indicators measured numerically as a proportion of the population. In order to better visualise the numerical data, we plot histograms of each numerical variable, with the exception of the categorical `RegionCode`.

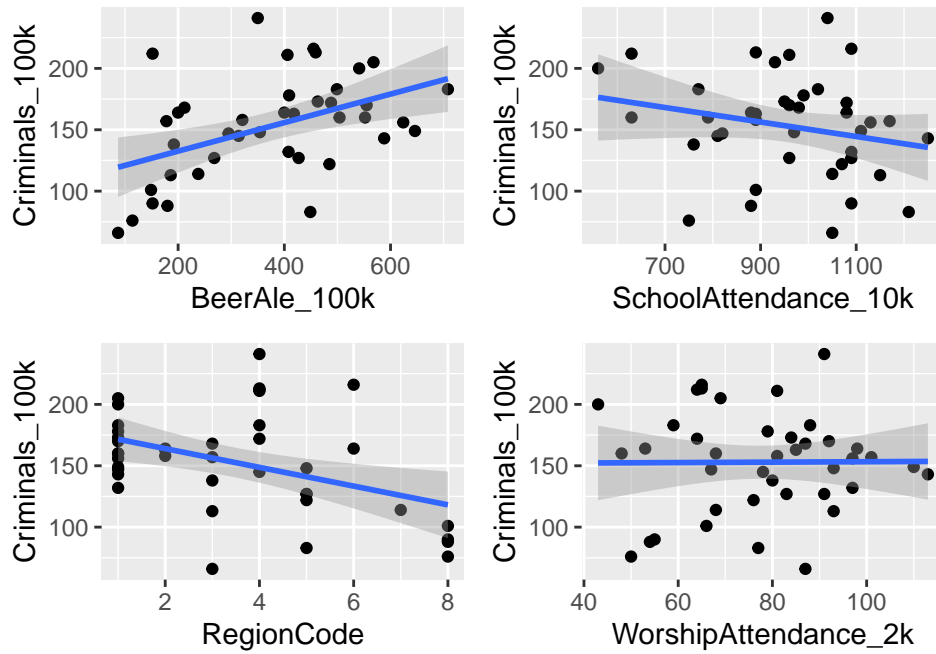


The `BeerAle_100k` variable appears to be bimodal. The histogram suggests that there are two common levels of alcohol consumption within the entire population, one around the 200, and another around 400 per 100,000 population. According to the data summary, `Crimanals_100k` has a mean of 152.9 and a median of 157.5, suggesting the symmetry that is also reflected in the histogram. `SchoolAttendance_10k` is slightly left skewed, while `WorshipAttendance_2k` has a varied but loosely symmetric distribution.

The linear dependence between each pair of numeric variables is expressed in the following correlation matrix:



There appears to be negligible linear dependence between worship attendance and criminality. School attendance is slightly negatively correlated with the prevalence of crime. Criminal behaviour is positively correlated with `BeerAle_100k` with a correlation coefficient of 0.46, suggesting that counties with a more dominant culture of frequenting bars and pubs are also where more crime happens. In order to better visualise these dependencies, we regress `Criminals_100k` on each of these variables.



There appears to be a very clear linear dependence of criminal behaviour on bar attendance, with most datapoints falling within the 95% confidence interval of the regression line. There is high variation of criminality across school attendance and worship attendance.