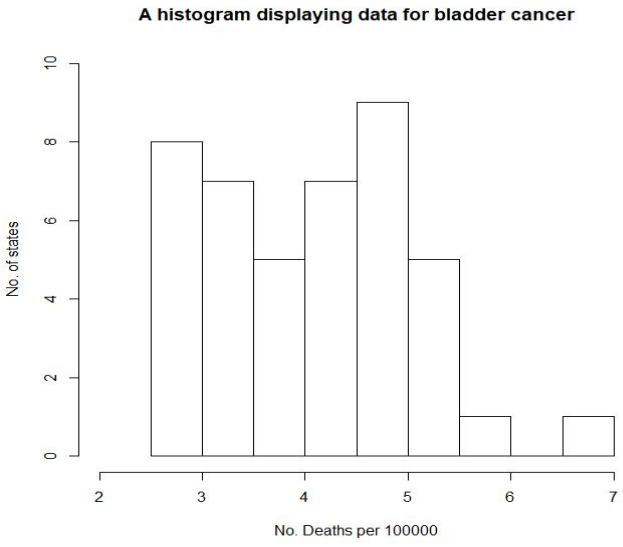
A Statistical Report On The Relationship Between Cigarette Sales And Various Forms Of Cancer

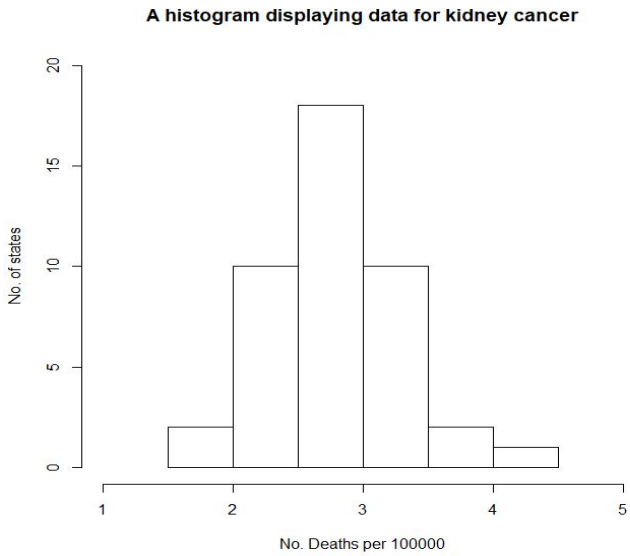
This report aims to show a link between cigarette sales and specific cancer deaths. The exploratory methods used are graphical analyses. Results indicate that increased sales lead to higher deaths from kidney/lung/bladder cancers, with lung cancer being responsible for more deaths in 1960.

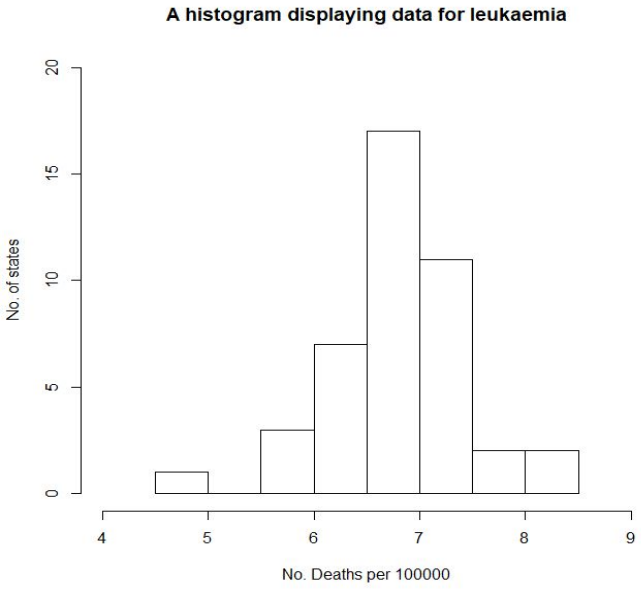
This study investigates the relationship between cigarette sales (per person) and the number of deaths from various forms of cancer. Previous studies have shown that cigarette use is a risk factor for cancer-related diseases, particularly lung cancer, which has a high mortality rate. In this report we aim to determine a link between increasing cigarette sales and instances of death for the following forms of cancer: bladder, lung, kidney and leukaemia. Furthermore we aim to analyse the relative number of deaths from each cancer. The data available are the cigarette sales per person in 43 US states in 1960 along with the number of deaths per 100,000 from the various forms of cancer above.

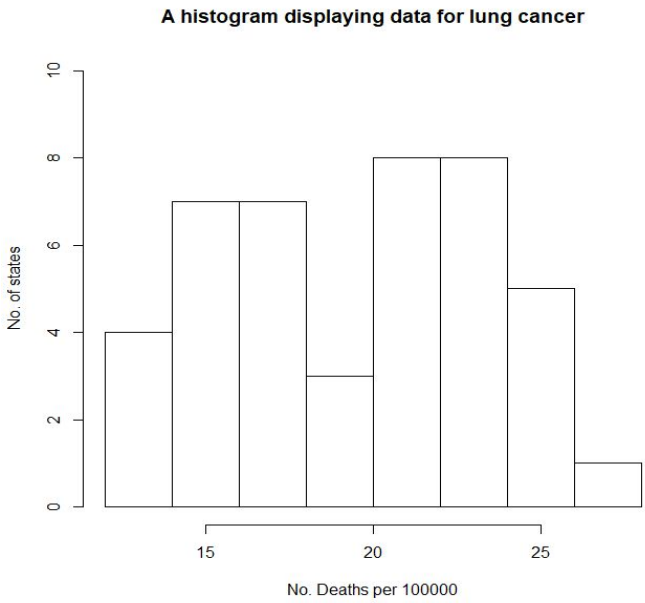
The methods used to display data include histograms, boxplots and scatterplots (presented with regression line). Firstly, histograms were produced for each cancer where number of deaths were plotted against frequency (the number of US states) with 8 equal width class intervals. Limits for these respective axes were determined using the minimum/maximum elements for each cancer type (x-axis) and by the height of the modal class (y-axis). Secondly, the data for each cancer type was presented as boxplots on one set of axes. Here the number of deaths was plotted against type of cancer. Then scatterplots were made between cigarette sales (vertical) and cancer deaths (horizontal). The y-axis had the same limits for each plot (between 15 and 40) whilst x-axis had limits again dependent on maxima/minima for cancers. Regression lines were also added to these plots (aide for viewing correlation). The sample statistics calculated for each variable were the mean and standard deviation.

Histograms for each type of cancer:









Description of distributions

Majority of the data for the number of deaths per 100,000 for bladder cancer lies in the range of approximately 2.5-5.5. There is a gap between 6 and maximum value near 7.

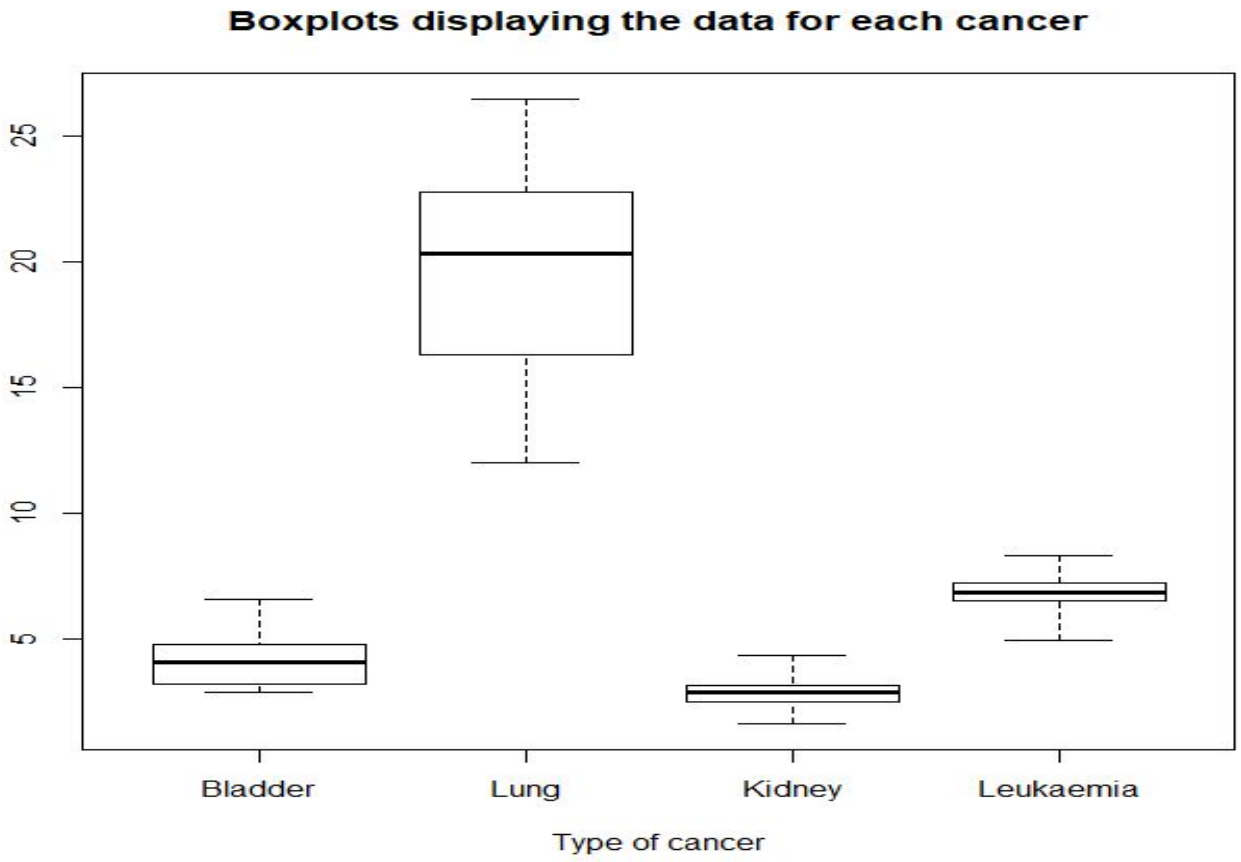
Kidney cancer deaths have a bell-like distribution with few observations below 2 deaths, bulk of the data between 2-4 deaths, 1 observation greater than 4. The modal class at 2.5-3 deaths has large frequency with 17 observations.

The largest proportion of leukaemia data lies in range 6-7.5. Similarly to kidney cancer, around modal class it is almost symmetric with exception to the observation at about 4.5, where there is a gap between this and rest of the data.

The histogram for lung cancer shows that the frequency is quite evenly distributed over the range. There are two modal classes at frequency 8 corresponding to classes between 20 and 24 deaths.

Boxplots

The boxplot shows the data for lungs has the largest range and the greatest number of deaths compared with the other cancers. The median value for lung cancer is 20 deaths, more than double that of the other forms of cancer. The other forms of cancer have data that overlaps, namely between leukaemia and bladder, and bladder and kidney respectively. These types of cancer have median observations under 10 deaths. The interquartile range for bladder is greater than that of kidney/leukaemia showing that it's data is more spread around its median.



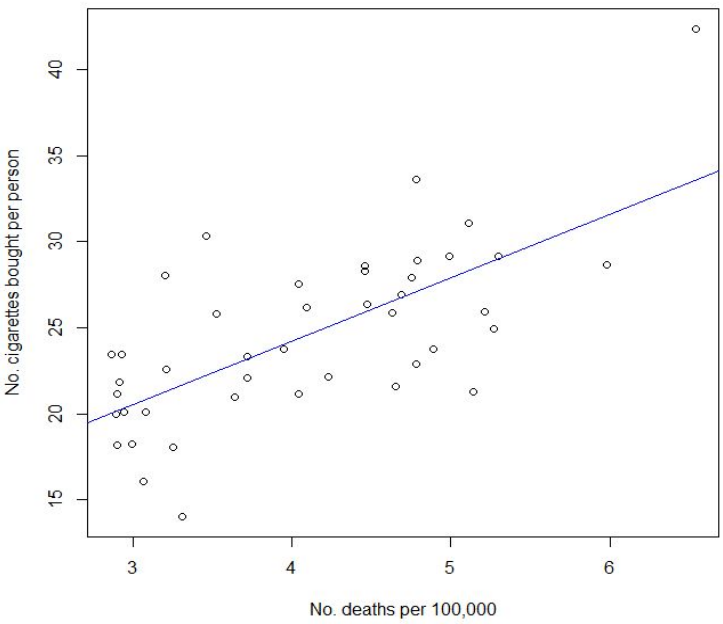
Variable	CIG	BLAD	LUNG	KID	LEUK
Mean	24.553	4.087	19.476	2.787	6.824
S.D	5.090	0.949	4.110	0.523	0.645

Outliers

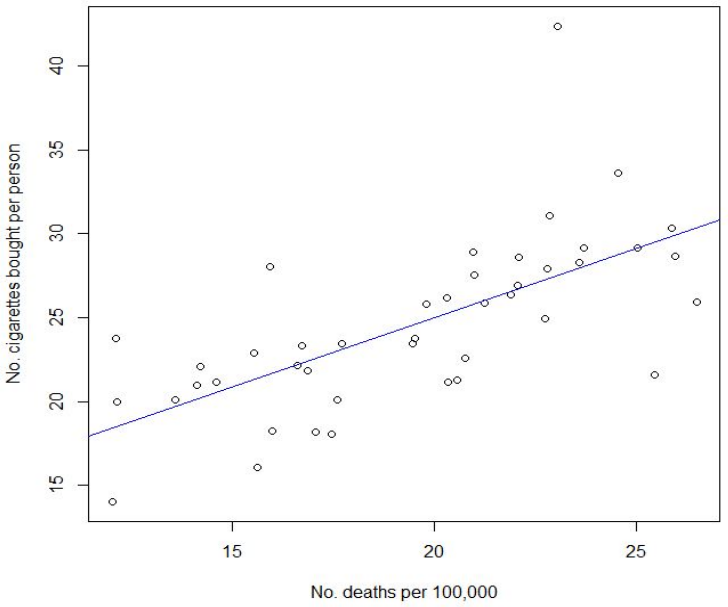
Outlier in Nebraska as no. of bladder cancer deaths was 6.54 (>2s.d from mean), plus cigarettes sold per person in this region was much higher (42.4).
Arkansas had a high extreme for kidney cancer (4.32) and low extreme for leukaemia (4.9).

Scatter plots:

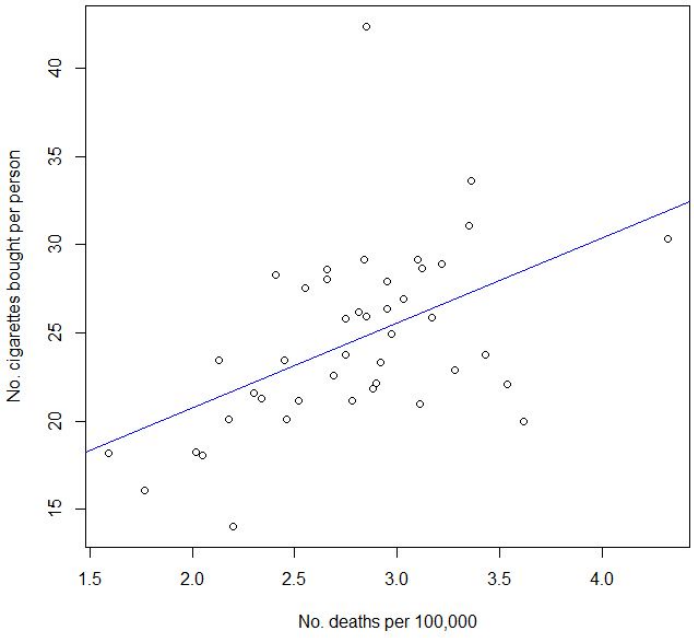
A scatterplot showing the relationship between CIG and BLAD



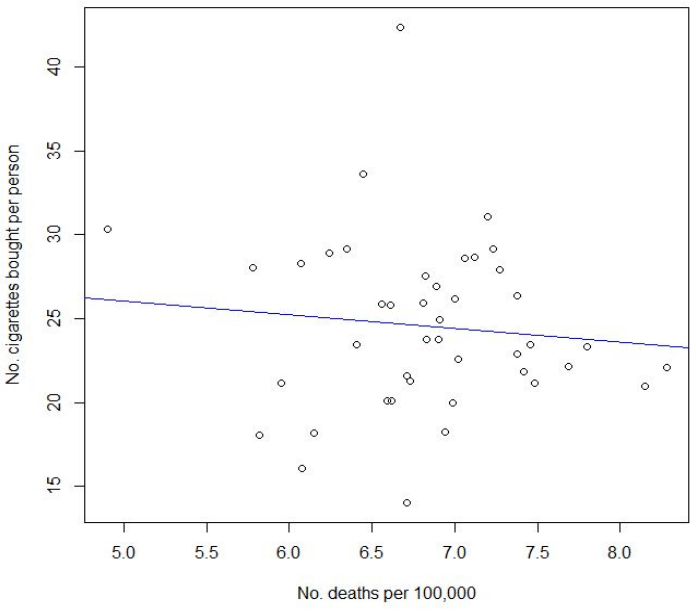
A scatterplot showing the relationship between CIG and LUNG



A scatterplot showing the relationship between CIG and KID



A scatterplot showing the relationship between CIG and LEUK



We conclude that there is a positive correlation between cigarettes sales and deaths for lung, kidney and bladder cancer. Lung cancer has the largest mortality rate. There is insufficient evidence from data to conclude a link between cigarette sales and deaths from leukaemia. Limitations include that we don't know how the average for cigarette sales calculated in each state. Using death from each cancer isn't as reflective of the effect of cigarette sales as number of cancer diagnoses. Data collection could have been extended over more years.

Word count:744