Correlation between Alcohol Consumption at Venues and Rental Prices

This project delves into the domains of health and to a lesser extent education. The question that it aims to answer is: **Is there a Correlation between Alcohol Purchased in the Last Seven Days and Rental Prices?** The reason for this investigation resides in the number of alcohol related deaths and injuries per year in Australia. Every day approximately 15 people die and 430 people are hospitalised due to alcohol, which comes to 5554 deaths and 157,132 hospitalisations each year.[1] Unfortunately for Australia, these numbers only seem to grow each year and as a result, the results of this study can be used to help target the necessary education to a certain economic bracket and identify possible factors that could lead to alcoholism. Additionally it can help with the positioning of drug busses and fine distribution. For example in areas with higher incomes and higher alcohol usage, increased deduction of points and harsher fines can be implemented for driving under the influence.

The Datasets:

The datasets to be used are the 'LGA Purchased alcohol in the last 7 days' released by VicHealth and 'Quarterly median rents by LGA Dec 2015' released by The Department of Human Services. These datasets can be found at:

- Alcohol Purchase: http://data.aurin.org.au/dataset/vic-govt-vichealth-vichealth-alcohol-purchased-last-7-days-lga.

- Median Rent: http://www.dhs.vic.gov.au/about-the-department/documents-and-resources/research,-data-and-statistics/rental-reports-2015.

The Alcohol dataset includes a Victorian average, lower confidence interval, upper confidence interval, and significantly different to the Victorian average (high, low or none) by LGA. The dataset doesn't include uncertain responses and respected peoples' right to avoid answering. It is believed that the survey was conducted in December 2011. The dataset was in the form of a CSV file.

The median rent dataset includes the median rent per week by LGA per financial quarter dating back to 1999. Additionally it sections off the LGAs into their greater areas and provides the average for the greater areas as well as each LGA. Moreover, it includes the average for the Metro and Non-Metro areas. Furthermore, the dataset includes the number of properties accounted for (Count). For properties with no data a '-' is used instead of a value. Finally, the data mentioned above is included for one, two and three bedroom flats as well as two, three and four bedroom houses. Like the alcohol dataset, the median dataset is in the form of a CSV file.

Planning and Pre-Processing:

Prior to the integration of the data, pre-processing had to occur. In the pre-processing phase, the data for the median rent was examined a couple of choices had to be made. The first choice was whether to use prices for houses or for flats. The second choice was how many bedrooms. In the end two bedroom flats were chosen as they represented a balance between pricing and amount of data taken into account. The remaining tabs for all other combinations of bedroom number and house/flat were then closed on excel. This resulted in 16 years of housing prices

[1] "Alcohol-Related Illnesses Kill 15 Australians Each Day", *ABC News*, 2014 <http://www.abc.net.au/news/2014-07-31/15-australians-die-each-day-from-alcohol-related-illness:-study/5637050> [accessed 15 May 2016].

which had to be reduced to a single quarter. This was made easy as the survey for the alcohol dataset was conducted in December 2011, which was also the date for the final financial quarter for 2011, meaning that there was a match. The remaining data was then cleaned out using the first of many idiosyncratic python scripts which created a new csv file with all other financial quarters deleted and all missing data was replaced with the LGAs' greater area average. The average was used in place of missing data as the LGAs generally had a relatively similar price as the other LGAs in their greater area. Upon inspection of the cleaned script, it was found that the group totals were all forgotten, along with the Melbourne average. This was amended by deleting the data manually. The final course of action to create the clean CSV was ordering the columns in alphabetical order of LGA (to match the alcohol dataset) and saving it as a new CSV file.

The alcohol dataset was a lot easier to clean. This was due to the fact that it was already a rather small dataset and only required the removal of unneeded columns, leaving behind only the LGA names, Victorian Average and the LGA Value. This was also done through an idiosyncratic python script which took the old csv, removed the unnecessary columns, and then outputted the cleaned dataset to a new csv file.

The method used in the pre-processing phase had a few weaknesses. The first being the manual excel editing. While it is a valid method of processing, it is prone to human error and there can be varying results amongst different people. A second limitation was the idiosyncratic scripts. These make the processing of data a lot easier and simpler, while providing a starting point, however they cannot be reused for a dataset other than the one they were created for. This results in a less efficient implementation as lots of the same code was reused but with slight variations.

Integration:

The integration of the data posed interesting challenges. It raised several questions of whether certain parts of the data were even necessary. For example the count in the rental prices dataset was carried over but no use was found for it in the integration phase. It is here where it was realised that certain pieces of information are important in certain datasets but once exported into a new dataset it could seem meaningless. This was also the case with the Victorian Average column, as it was useful to know but it didn't help to have integrated within the final dataset to be used for graphing. As a result, the count column was removed, while the average column was retained for future reference.

Integration was done multiple ways. Firstly, the dataset that replaced missing data with the averages was integrated. This was done by taking the two cleaned datasets from the planning phase and writing a short script that would integrate them. This script involved iterating over the columns in both files and only extracting the relevant columns. It extracted the LGA names and median rent values from the rent dataset, and the Victorian average as well as LGA average from the alcohol dataset.

Upon completing the integration of the datasets, precautionary measures were taken and a few more datasets were created in excel. The first being the creation of a data set which does not take into account missing data. This would remove the LGA entirely. This was done manually in excel. Additionally an average for greater areas was calculated and combined in a single CSV. This was done through a less than efficient idiosyncratic script (as it initially came as an afterthought). The average by Greater areas dataset is limited as it is rather small and may mean that very little can be deduced from it.

The final set of integration came as each greater area received its own dataset with the

values from the initial dataset with the averages replacing missing data. This was done through extensive manual excel editing, however this would have been easier had it been done on python. It was done through excel as it was initially an experiment to see whether the data changed drastically between the normal data and a single greater area.

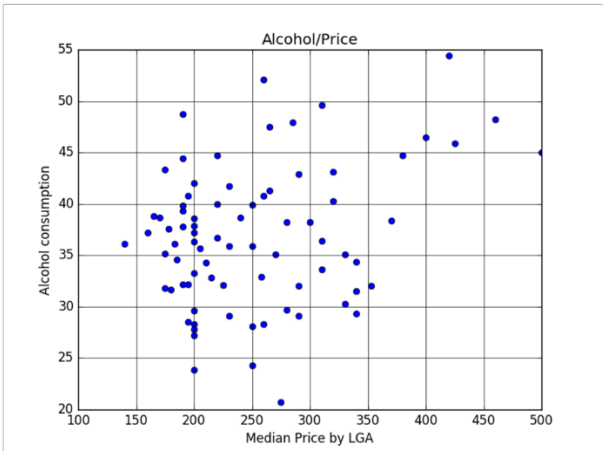
The limitations of the method used in integrating the data is more dangerous than in the pre-processing despite using identical techniques. This is a result of the more widely used excel editing in the integration phase.

Results:

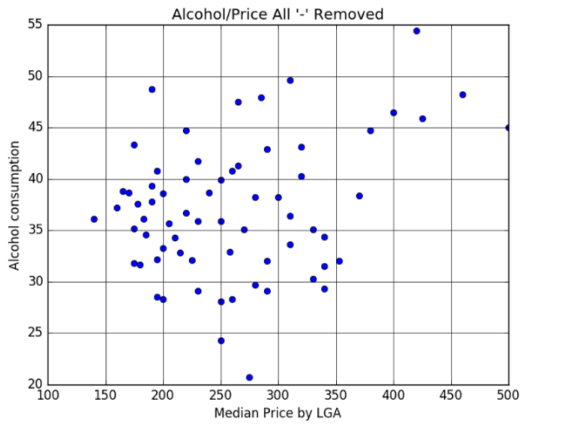
The visualisations were done through python scripts. There was one for the original dataset, one for greater area only and a final one for the remaining datasets including the combined dataset where missing data is removed. The correlation coefficients of the data were calculated in excel using the CORREL function. After the visualisations and correlation coefficients were done, the visualisations were examined for potential outliers. In the event of outliers the LGA in question would be researched for a possible cause.

The first visualisation done was the plot on the left. The plot creates doubt in the premise of the investigation and no reassurance was had when the missing data was removed completely as the correlation coefficient was still around 0.3, which implies a weak positive relationship.

Left: Missing data replaced by greater area Average,



Right: Missing data was removed entirely

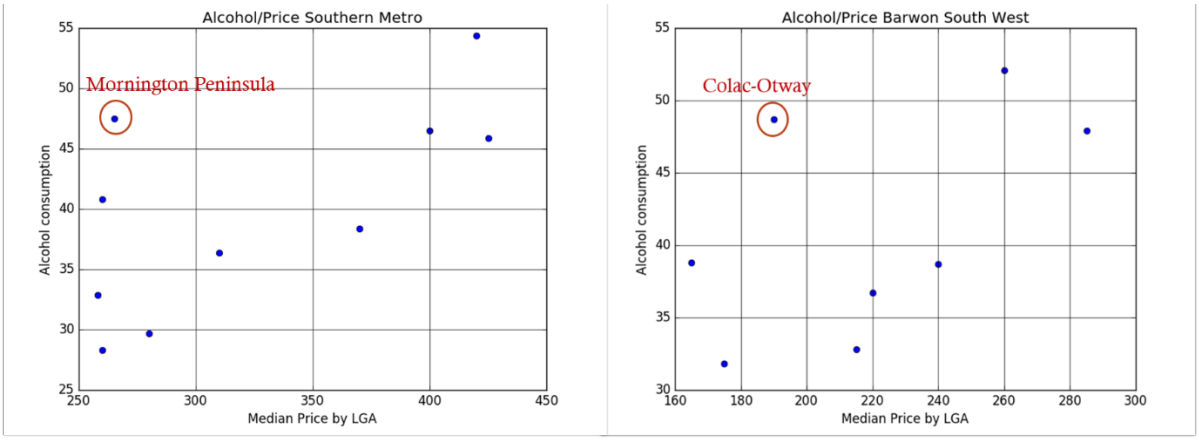


Correlation Table: 0 = no, ±0.3 = weak, ±0.5 = moderate, ±0.7 = strong, ±1 = perfect relationship

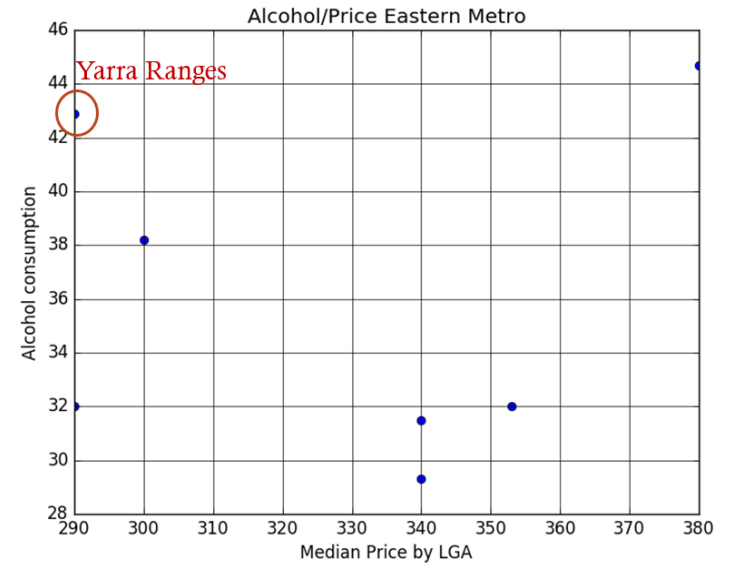
Dataset	***Correlation Coefficient***
Combined(Average replaces missing data)	0.303591
Combined(Missing data removed)	0.310252
Greater Areas Averages	0.116334
Baron South West	0.532366
Eastern Metro	0.03561
Gippsland	-0.38695
Grampians	-0.03507
Hume	-0.02853
Loddon Mallee	0.382051
North Western Metro	0.664713
Southern Metro	0.666823

Correlation between Alcohol Consumption at Venues and Rental Prices

The second set of visualisations was of the Southern Metro area and the Baron South West area. Both areas have an apparent outlier and have a moderate positive relationship. In the case of the Southern Metro Area, the Mornington Peninsula is a popular vacation spot, with many people owning holiday homes in the area. This is significant as the survey was done during a summer month which could have impacted the amount of people in the area as well as the amount of people drinking. As for Barwon South West's Colac-Otway also contains a few popular summer spots as it includes a large portion of the Great Ocean Road including Apollo Bay. Therefore, for the same reasons as the Mornington Peninsula the amount of alcohol consumption may have increased. Should we remove the Mornington Peninsula and Colac-Otway from the dataset, the correlation coefficient for Southern Metro becomes 0.83311 and Baron South West becomes 0.737695. These are a stark contrast to all the other correlation coefficients shown above as they are both strong linear relationships.

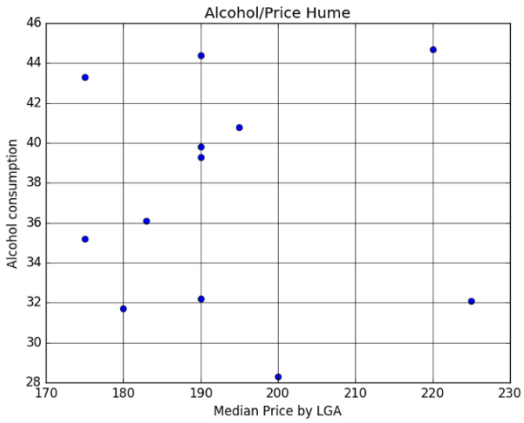


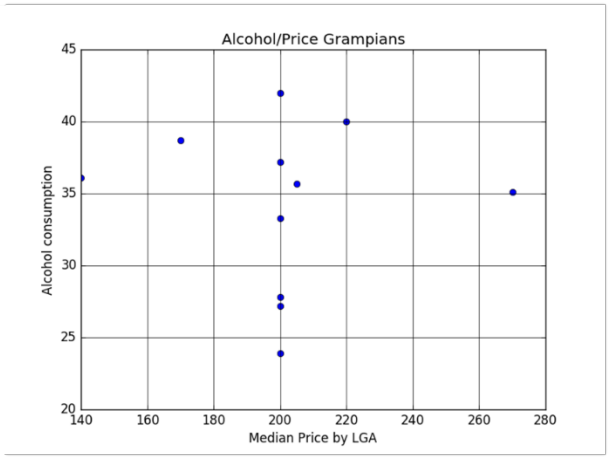
The Eastern Metro region appears at first glance to have no correlation. However upon inspection it is seen that the Yarra Ranges are a part of the Eastern Metro area. The Yarra Ranges are home to many popular wineries and thus is likely to have an effect on the amount of alcohol purchased. Once the Ranges are removed from the equation, the correlation coefficient jumps to 0.36853 which is a weak relationship.



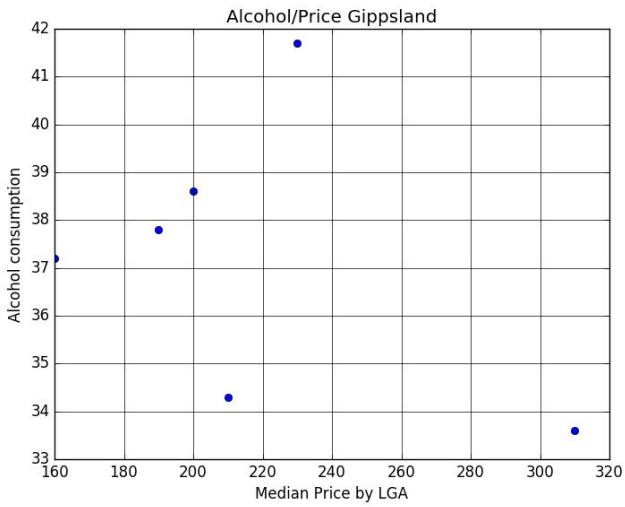
The Grampians and the Hume areas show are insignificantly negative and show no

significant correlation. This is a result of their data being scattered and inconsistent with any trend. The dataset obtained from averaging the greater areas also belongs here as it has an insignificant relation at a 0.116334 correlation coefficient. However the Greater Areas averages is flawed as it doesn't take into account number of surveyed individuals and weighs all greater areas equally.



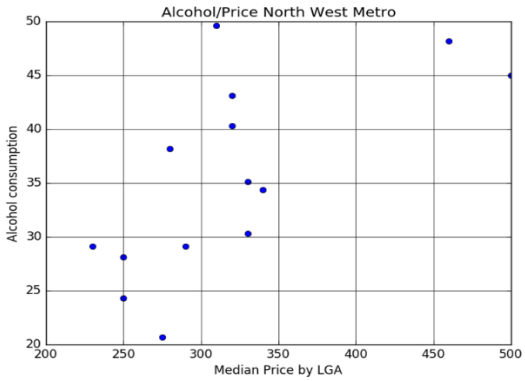


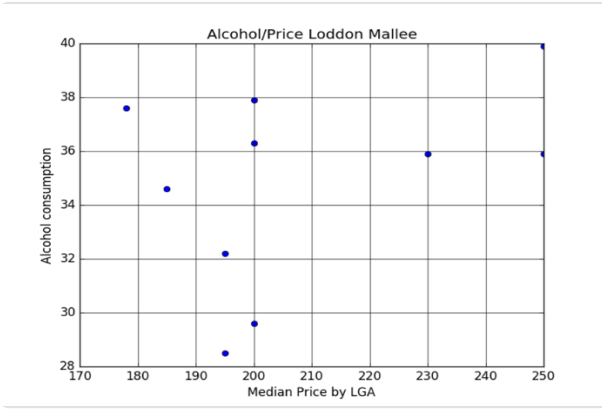
The Gippsland dataset is the only area with a weak negative correlation as it has a correlation coefficient of approximately -0.38. This goes against the trend but this also shows how the different greater areas can vary. It also shows that it is important to look past the initial observations and notice that there may be the opposite problem in a certain area. Additionally, it also appears that the population in Gippsland tends to purchase alcohol more frequently as all but two LGAs are above the average (36.3). This may suggest an underlying issue.



The North West Metro and Loddon Mallee are in line with the trend that has been set up, however they are also the only greater areas that do not have any obvious outliers. North West Metro is more strongly correlated than Loddon Mallee with an arguably moderate positive correlation to Loddon Mallee's weak correlation. However, Loddon Mallee seems to avoid the extremes, unlike North West Metro, where there is both the lowest and the highest alcohol purchasing values.

Correlation between Alcohol Consumption at Venues and Rental Prices





Overall, there is a weak positive correlation between rental prices and alcohol purchase. This changes amongst the Greater Areas, with some having strong positive correlations (once outliers are removed) and others having weak negative correlations. There are only two Greater Areas with no correlation which shows that generally there is a correlation. The results also show that there will generally be a weak positive correlation, meaning that more people with higher rental prices tend to purchase more alcohol than those with lower rental prices.

Limitations in the method do exist. As mentioned earlier the Greater Areas averages have been ignored as the implementation was ignorant of the number of surveys done per greater area and treated all greater areas as equal despite some areas having multiple LGAs with no values. Additionally the inclusion of the LGAs without values and replacing them with the average is also flawed as it may skew the datasets and the correlation co-efficient, meaning that there should be a better method for missing data sets to be put in place.

Value:

The pre-processing, integration and analysis added value compared to the raw data. The pre-processing helped return only the data to be used which assisted the later stages. Without the pre-processing stage looking at the data would mean looking at 384(16 years *4 quarters*6 combinations of home/flat/bedroom) separate pieces of data for the rental prices alone. The alcohol dataset is much more forgiving, however there would be a few extra unnecessary columns. The integration of the data built on the pre-processing and combined both datasets into one readily available dataset with all the important information from the two initial datasets. Furthermore it helped identify the points of exploration to provide more robust visualisations and answers by inciting the idea of checking for correlation within the greater areas.

The visualisations and analysis helped make sense of all the data. It provided a visual version of the dataset which makes it easier to spot outliers and better see trends in the data sets. With the raw data on its own, understanding the entirety of the data in depth would be impossible as it would not provide a tracking medium in the same way that graphs and charts do. Raw data would have resulted in the mental filtration of the data which could have led to cherry picked results and would have greatly increased the risk of an unethical and misguided report. Additionally the visualisations placed both datasets on the same axis showing one in relation to the other which, as a result makes it easier for an uninvolved person to make a fair assessment of the data.

Overall going through the pre-processing, integration and analysis helped create a fair interpretation of the data. This would not have been possible by looking purely at the raw data as it contained data that is unnecessary and irrelevant to the report.

Correlation between Alcohol Consumption at Venues and Rental Prices

Challenges and Reflections:

The biggest challenge faced was making sense of the data in the analysis/visualisation stages. The initial results from the 'combined (Averages replace missing data)' scatter plot was blindsiding as it results appeared to be insignificant at first glance. This challenge was overcome by taking a step back and breaking the data down into Greater Areas which would prove useful in understanding the data.

Another challenge was the exploration of different methods of doing the same thing. This occurred in both the pre-processing and the integration phases. The entire project could have been done using excel, however learning to utilise python scripts to get a similar result would prove difficult yet rewarding. This also raised the issue of the most efficient method. In some cases excel would be more efficient and time saving, other times writing a short script would be much quicker.

A surprisingly large challenge was downloading some of the libraries used onto a personal computer as the libraries would not download or work. However, this was overcome by simply using Jupyter Notebook which proved to be a helpful tool in the visualisation stages. Once the scripts were working on Jupyter they were saved as python files for compatibility reasons.

Should the project be repeated, more time would be spent understanding and learning visualisation skills to create interactive graphs and plots as that was not possible with the current investigation due to a lack of skill in that particular department. Additionally the first point of visualisation would be changed to the greater areas rather than all LGAs at once in order to see the progression of the relationship between the datasets.

Question Resolution:

The results help answer the question as they provided figures and factual mathematical evidence to prove the extent of the relation between rental prices and alcohol purchase. Furthermore, the correlation coefficient found showed the varying degrees of correlation amongst the greater areas. It also showed that there was a weak positive correlation overall.

These results are useful as it shows that in most areas those in a higher socio-economic standard are more likely to drink regularly. This can be used by the government to start initiatives to educate the public on the dangers of alcoholism as well as driving under the influence. Having an LGA and an average rental price helps paint a picture and target the campaigns at those with the highest risk of descending into alcoholism. Additionally it can help with the positioning of drug busses and fine distribution. This creates a tiered system which can help deter those with a larger income from driving under the influence. This could also interest researchers should they want to study areas with different socio-economic backgrounds that drink similar amounts for health consequences.

Finally this can assist those who are looking to open licensed venues as a correspondence between property prices and rental prices makes investments in certain areas more lucrative. This could lead to less failed investments due to ignorance of the topic, which would help both the investor and the government as they would both avoid a case of bankruptcy.

Overall this could help people make wise decisions. From investors looking to open up a licensed venue, to the government tackling the growing problem of alcoholism in Australia, to the everyday person who would be more informed on the risks of the consumption of alcohol on a regular basis.

Correlation between Alcohol Consumption at Venues and Rental Prices

Code:

In total, seven python scripts were written of varying lengths, of which the longest is 117 lines. The total amount of code written is approximately 364 lines worth. However that does include code copied over between files. The major libraries used where Matplotlib's plotly and CSV. Matplotlib was used for visualisations, while CSV was used to read and write new CSV files. All code used was written from scratch. When using Matplotlib however, code was adapted from a stack overflow forum. All code written was written in python as it is a powerful and diverse data wrangling tool.

No code other than python was used. Aside for the adapted section mentioned above, no publicly available code was used either.

Bibliography:

"Alcohol-Related Illnesses Kill 15 Australians Each Day", ABC News, 2014 <http://www.abc.net.au/news/2014-07-31/15-australians-die-each-day-from-alcohol-related-illness:-study/5637050> [accessed 15 May 2016]