Investigation: The Effect of Internet on Health.

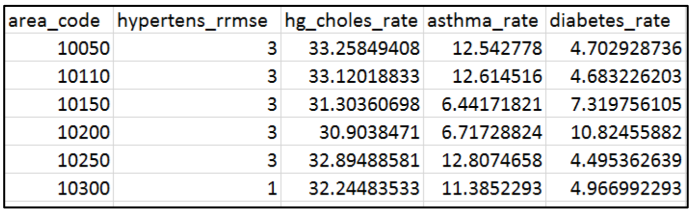
Domain of Project: Telecommunications Infrastructure and Health

Question

"Does internet quality correlate with the risk of developing an illness?"
This question could be of interest to internet providers, telecommunications infrastructure, hospitals and general practitioners; as internet is becoming ever present in this digital world. A correlation like this could highlight the importance of internet regulation or improvement in Victoria for a greater outlook on health.

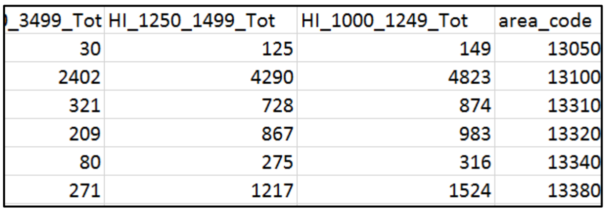
Datasets

Originally I used the LGA Technological Readiness 2011[1] dataset and the LGA11 Self Assessed Health – Modelled Estimate from 2011 to 2013[2]. After some research I've also decided to use the LGA11 Chronic Disease[3]. (All health-related datasets surveyed by Torrens University, and the Technological Readiness Dataset was surveyed by the RAI organisation). A complimentary dataset, LGA11 Total Household Income B28[4] from the Australian Bureau of Statistics will be used in conjunction with these datasets.



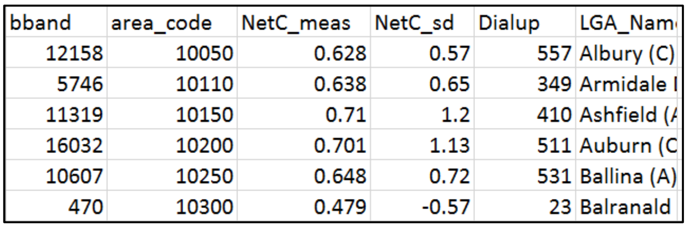
area_code	hypertens_rrmse	hg_choles_rate	asthma_rate	diabetes_rate
10050	3	33.25849408	12.542778	4.702928736
10110	3	33.12018833	12.614516	4.683226203
10150	3	31.30360698	6.44171821	7.319756105
10200	3	30.9038471	6.71728824	10.82455882
10250	3	32.89488581	12.8074658	4.495362639
10300	1	32.24483533	11.3852293	4.966992293

Figure 2 – Sample Chronic Illness Data



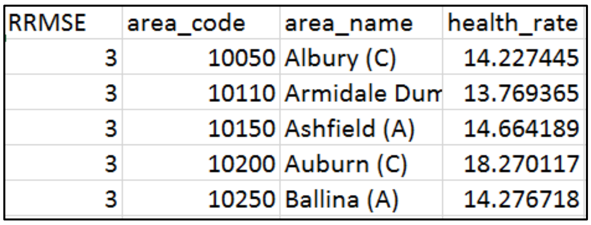
_3499_Tot	HI_1250_1499_Tot	HI_1000_1249_Tot	area_code
30	125	149	13050
2402	4290	4823	13100
321	728	874	13310
209	867	983	13320
80	275	316	13340
271	1217	1524	13380

Figure 2 – Sample Income Data



bband	area_code	NetC_meas	NetC_sd	Dialup	LGA_Nam
12158	10050	0.628	0.57	557	Albury (C)
5746	10110	0.638	0.65	349	Armidale I
11319	10150	0.71	1.2	410	Ashfield (A
16032	10200	0.701	1.13	511	Auburn (C
10607	10250	0.648	0.72	531	Ballina (A)
470	10300	0.479	-0.57	23	Balranald

Figure 3 – Sample Technological Readiness Data



RRMSE	area_code	area_name	health_rate
3	10050	Albury (C)	14.227445
3	10110	Armidale Dum	13.769365
3	10150	Ashfield (A)	14.664189
3	10200	Auburn (C)	18.270117
3	10250	Ballina (A)	14.276718

Figure 4 – Sample Health Rate Data

The LGA Technological Readiness 2011 Dataset contains information of an LGA's 'technological readiness' rating. This dataset includes information regarding percentage of connected households and type of an internet connection per household in an LGA (representing how well the LGA is performing in comparison to a national 'average Australian region') in 2011. Most important for this study, the dataset contains percentage of Broadband connections vs percentage of Dialup connections. This will prove critical to the analysis process.

The LGA11 Self Assessed Health – Modelled Estimate from 2011 to 2013 contains information regarding the percentage of the rate per 100 in an LGA that are self-assessed poor health. This provides a general overview over the health quality of an LGA, by assessing the rate of peoples in an LGA that are 'unhealthy'.

The LGA11 Chronic Disease dataset contains information regarding the types of chronic disease present in the population of an LGA. This includes Arthritis, Asthma, Circulatory, COPD, Diabetes, High Cholesterol, Hypertension, Mental/Behavioural Problems, Musculoskeletal and Respiratory problems.

[1] https://data.aurin.org.au/dataset/rai-ria-ria-2011-technology-indicators-lga2011
[2] https://data.aurin.org.au/dataset/tua-phidu-lga11-selfassessedhealth-modelledestimate-lga2011
[3] https://data.aurin.org.au/dataset/tua-phidu-lga11-chronicdisease-modelledestimate-lga2011
[4] https://data.aurin.org.au/dataset/au-govt-abs-b28-aust-lga-lga2011

The LGA11 Total Household Income B28 records the number of houses with certain weekly income brackets, from Negative income, 1-199, 200-299, 300-399, 400-599, 600-799, 800-999, 1000-1249, 1250-1499, 1500-1999, 2000-2499, 2500-2999, 3000-3499, 3500-3999, to 4000+.

Pre-Processing

Pre-processing was done in a jupyter notebook, utilizing the pandas library. Before the CSV files were loaded into their respective dataframes, data cleaning was done.

- All datasets had a space between the attributes and commas in the CSV files, so these were removed.
- Some of the datasets had very long and complex attribute titles, so these were edited in a text editor to a short form version. For example, 'adm_rsp_all_2_rate_7_11_6_12' in the Hospital Admissions dataset became rsp_rate – the rate of admissions for respiratory problems. This was also similarly done in the chronic health, and self-assessed health dataset.

Data reduction was then done:

- All records that contained a null value were pruned, as these values could skew results.
- In the Hospital Admissions dataset, any attribute regarding pregnancy rates, caesarean sections and hysterectomies were removed from the dataframe as these admittance rates were not a good indicator of the general wellbeing of an LGA.
- All attributes involving counts that were not normalised (via rate or standard deviation) were removed. LGA codes that were not in common to all datasets or not in Victoria were also removed.
- I decided to leave in any outliers, as these could be analysed and bring forward interesting points.
- Any record with a non-accurate error was removed[5].

Visualisations were done using Matplotlib, seaborn and pandas. Initially I planned to use only scatter plots throughout the project to show correlations in the data. However, I found it difficult to show pearson coefficients efficiently – thus I decided to use heatmaps in addition to the scatter plots. It was also far easier to create and show correlation lines compared to manually calculating these via matplotlib.

Integration

Data transformation was done via Python. All CSV files were read into dataframes and linked accordingly by their LGA code. A function was made to take input dataframes and link all the necessary features according to the area code. All values were already numerical data so no conversion was required. This resulted in a large dataframe containing all the variables I needed to draw my results from. I initially had several problems linking the area_codes together, as my function was appending several rows for each dataframe entry, resulting in many 'NaN' values. To overcome this, I created a function check_row() which takes the row entry and the dataframe to be added, and checks for the area_code. Once the area code is found, and it is verified to not be a null record (not in Victoria or doesn't have any null values), it is passed into the function make_row() which makes the new row with all the attributes together. This streamlines the process even if I add more data sets to test with. Percentage of Dialup had to be calculated simply by taking the counts of the number of Dialup Connections divided by Total Connections, so it could be compared later against the already existing percentage of broadband connections (bband_meas). This was also the same for income, as each bracket was converted into a percentage of the total income count.

First, I did an overall analysis of internet connection type vs chronic health illness. I made a heatmap of each different type of internet connection and its correlation with the type of illness (Figure 4), labelling the ticks respectively. Then, looking at the strongest correlations, I made several scatter plots. Later, I also restricted data depending on the variable I wanted to look at; namely, household income, before recreating the aforementioned heatmaps.

Results

Looking at the correlation matrix (**Figure 4**), it's evident that there is a string correlation between certain illnesses and certain internet connections. On the y axis, the percentage of type of connection in an LGA is correlated against type of illness (rate per 100) in the LGA. The strongest correlations for most illnesses are

[5] With a RRMSE less than 3 (1 being most erroneous).

often linked with the highest quality internet (looking at the 'Broadband Connections' axes), where the Broadband standard is significantly faster than Dialup connections[6]. Oddly, the illnesses of Arthritis, COPD (Chronic Obstructive Pulmonary Disease) and Musculoskeletal origin appear to have a strong negative correlation with internet connection quality (**Figures 5 - 8**). Musculoskeletal illnesses are sometimes attributed to time spent using a computer[7] (Wærsted, Hanvold, & Veiersted, 2017). However, Arthritis and COPD however are usually caused by other external sources[8] (ZocDoc, 2017). This perhaps suggests that a faster internet connection could correlate with less time spent overall on a computer as content is found faster for the user, thus less time is spent waiting for a connection. In addition to this, users could be more willing to search for answers regarding health related issues as waiting for a slow connection is no longer a barrier behind their internet usage. Most significant about the entries in the matrix is the difference between the Dialup connection correlation and the Broadband connection

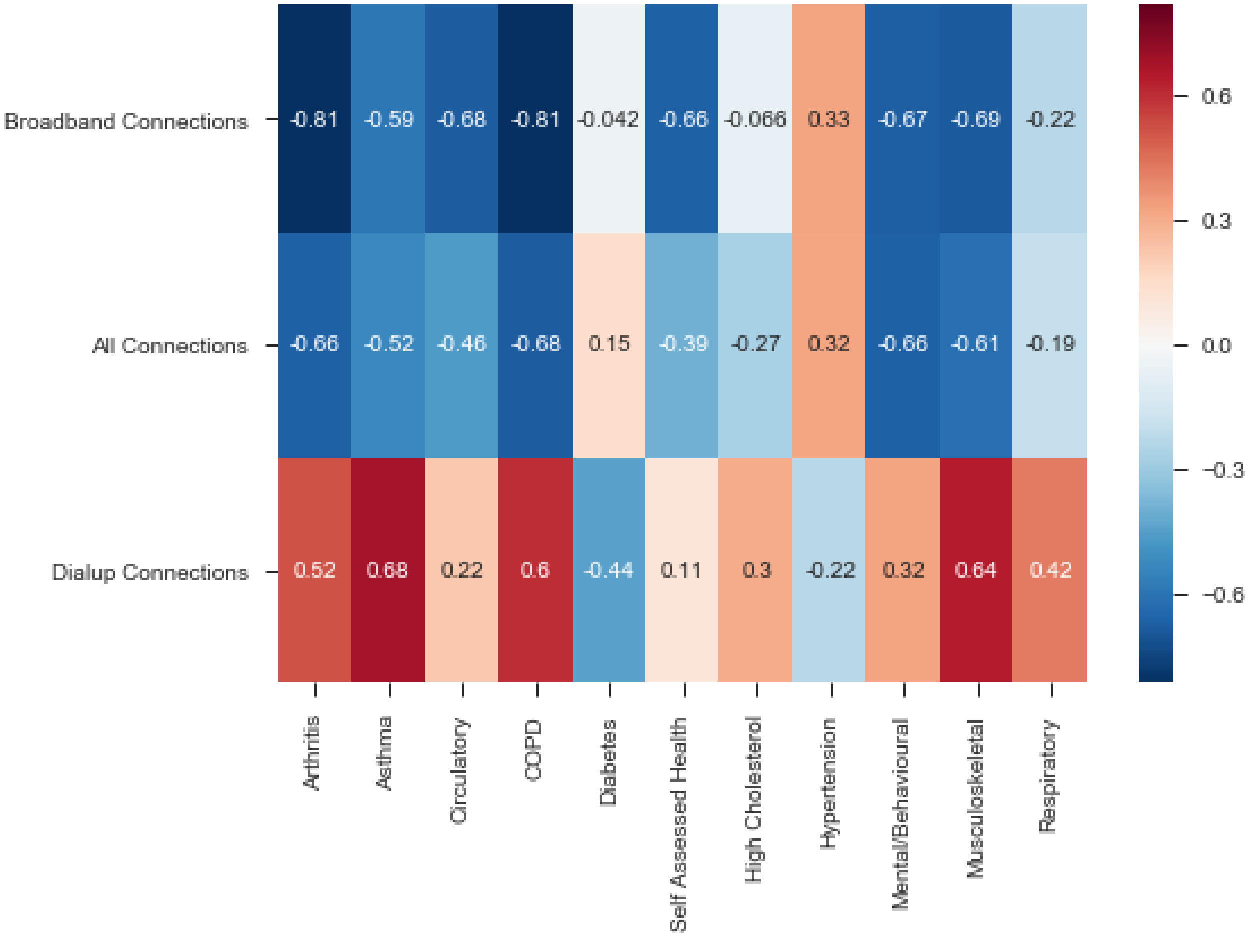


Figure 4 – Percentage of Type of Internet Connection vs Chronic Illness

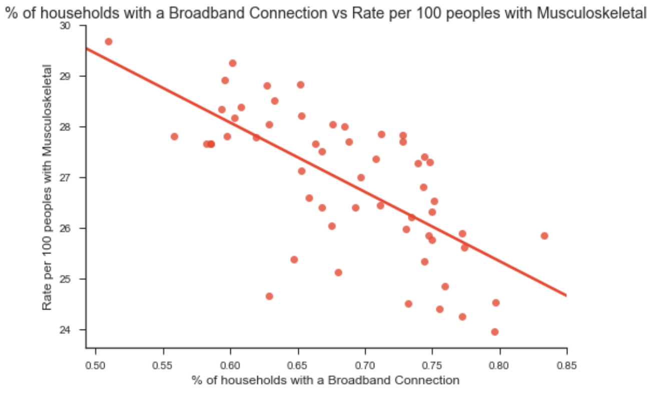


Figure 5 – r = -0.6862

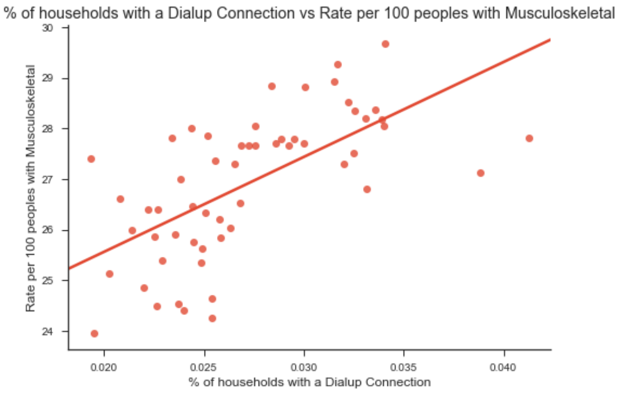


Figure 6 – r = 0.6398

[6] https://en.wikipedia.org/wiki/Dial-up_Internet_access#Replacement_by_broadband
[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2874766/
[8] https://www.zocdoc.com/answers/12150/can-you-get-arthritis-from-computer-use

correlation. The broadband connection correlations are stronger in every case of illness, suggesting that the quality of internet, not just volume, does have an effect on the general wellbeing of an LGA.

To ensure this wasn't just a case of higher income deciding the result of better healthcare, I restricted the records to the LGAs where 20% of households had greater weekly income than 2500[9] and plotted another heatmap (**Figure 9**); that is, some of the most wealthiest LGAs.

As shown, specifically with Arthritis, Circulatory problems, COPD, Diabetes, Self-Assessed health, and Musculoskeletal problems, this holds true. Although this is a rough approximation, it shows that quality of internet does correlate with overall wellbeing in an LGA. This suggests that, while perhaps income is the largest contributor to wellbeing, internet connections can contribute to the overall wellbeing of an LGA regardless of income. Unfortunately, as a limitation of the data found online, there is no way to fully explore the types of connections, as no dataset containing LGA codes and type of connection as outlined in "Type of Access Connection[10]" by the Australian Bureau of Statistics, currently exists. There is also no dataset which quantifies income vs internet connection per household (as in, there is no certain way to validify whether wealthy users have dialup or not). These datasets would allow this study to be taken further and allow for more flexibility in the overall model.

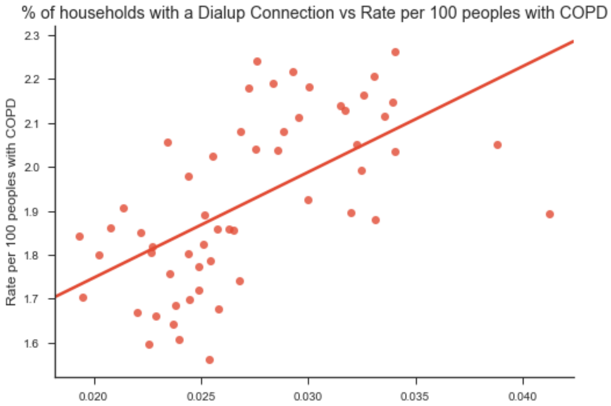


Figure 8 – r = 0.5987

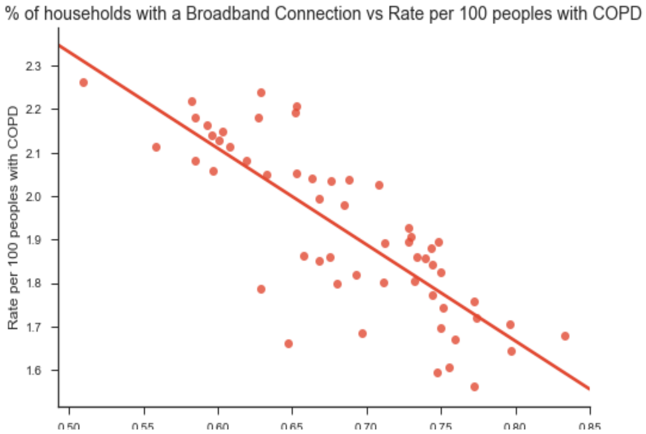


Figure 7 – r = -0.8117

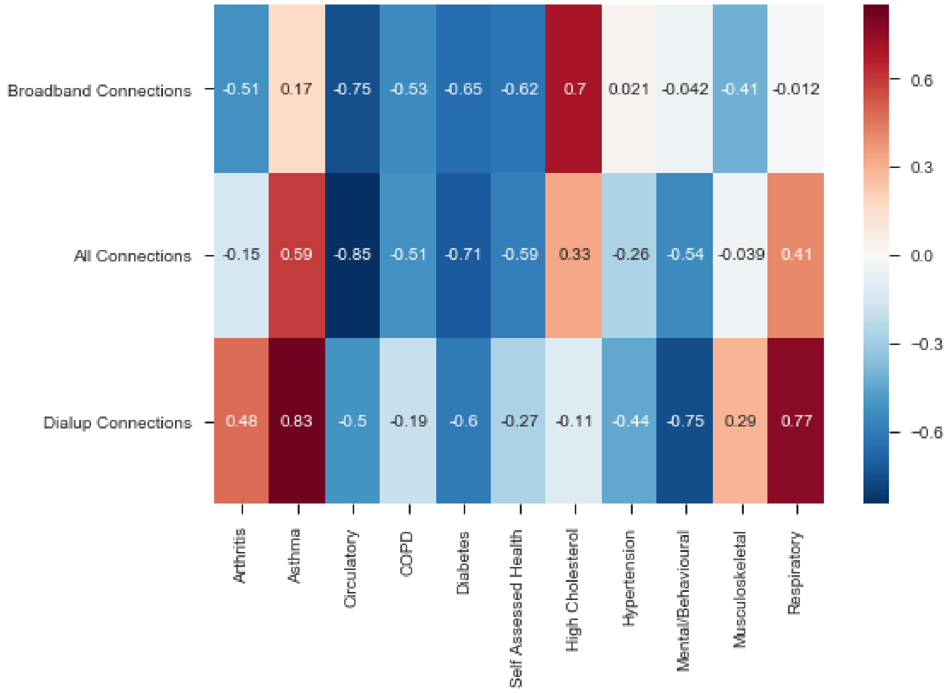


Figure 9 – Heatmap of Percentage of Type of Connection vs Chronic Illness of Wealthy LGAs

[9] Relatively high, as the income brackets for households spanned from 1-199 to 4000+; this was the top 4 brackets

[10] http://www.abs.gov.au/ausstats/abs@.nsf/mf/8153.0

Value

By itself, the data does not grant an insight into the correlation between internet connection and health. It is difficult to draw these conclusions from numbers alone, as internet quality and health are two very different metrics which are difficult to compare outright. With the visualisations, it is easy to see that there is a correlation between types of internet connections and general health. Explanations also help regarding the type of connection and the type of illness, which would require research in the reader's own time in order to effectively draw results from the raw data.

Challenges and Reflections

Working with many datasets presented a challenge compared to my initial investigation. With so many different dimensions to discuss it was difficult to go into each specific health illness and it's correlation value. It was also difficult to draw conclusions regarding income and internet usage, as there was no way to validate from the datasets that the wealthier users were also the same broadband and dialup users. I also initially intended to create a geographical heatmap via shapefiles, however, this was difficult to show as the area of Victoria was so vast. Data processing also proved a challenge, as working with 4 datasets meant there were many exceptions in my functions and some correlations were very weak. This could be improved upon by simplifying the number of datasets, or having access to a dataset with the variables discussed at the end of the results section (more types of connections, or a quantified income + broadband connection household LGA dataset).

Question Resolution

My analysis of the datasets and the correlation between them do suggest a correlation between internet quality and general overall wellbeing inside an LGA. It could be of interest to internet infrastructure providers, as well as civil development, as internet can be considered as a vital addition to the typical amenities.

Code

I wrote all Python from scratch, except for the importation of CSV files, which came from my original investigation. I used Seaborn, Pandas and Matplotlib as my primary Python libraries; Seaborn to generate the heatmaps, Pandas for the dataframe structure and accompanying functions/methods, and Matplotlib as the primary library to plot results.

Bibliography

Australian Buearu of Statistics. (2017, May 7). *Type of Internet Connections.* Retrieved from Australian Buearu of Statistics: http://www.abs.gov.au/ausstats/abs@.nsf/mf/8153.0

LGA Technological Readiness. (2017, April 20). Retrieved from AURIN: https://data.aurin.org.au/dataset/rai-ria-ria-2011-technology-indicators-lga2011

LGA11 Chronic Disease. (2017, April 28). Retrieved from AURIN: https://data.aurin.org.au/dataset/tua-phidu-lga11-chronicdisease-modelledestimate-lga2011

LGA11 Total Household Income B28. (2017, April 28). Retrieved from AURIN: https://data.aurin.org.au/dataset/au-govt-abs-b28-aust-lga-lga2011

Wærsted, M., Hanvold, T. N., & Veiersted, K. B. (2017, May 8). *Computer work and musculoskeletal disorders of the neck and upper extremity: A systematic review.* Retrieved from US National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2874766/

Wikipedia.org. (2017, May 8). *Dialup Internet Access.* Retrieved from Wikipedia.org: https://en.wikipedia.org/wiki/Dial-up_Internet_access#Replacement_by_broadband

ZocDoc. (2017, May 9). *Can you get arthritis from computer use?* Retrieved from zocdoc.com: https://www.zocdoc.com/answers/12150/can-you-get-arthritis-from-computer-use