COMP20003 Elements of Data Processing

Assignment 2 Report

Is there a relation between weather conditions and the consumption of goods of various industries in Victoria?

Group 39

Brandon Vincent Liongosari 1094413

Dominic Henley 1186484

Evelyn Putri 1118759

Ivan Adinata 1163497

I. Overview

A. Background

Victoria is renowned for having its wide-ranging climate conditions, with February being its hottest month and rapidly falling to its coldest month in July (Climate of Victoria, 2021). This raises the need to produce an insightful, data-driven analysis on the relationship between weather and consumption of goods in hopes of supporting the sustainability and economy in Victoria. Thus, we plan to investigate the question:

Is there a relation between weather conditions and the consumption of goods of various industries in Victoria?

B. Importance

B1. Business Owners, Investors

This investigation provides data assisting businesses in different industries in making future business judgments such as seasonal employment/unemployment, cut in costs, and increase budgeting in response to weather conditions.

B2. Government

This investigation allows the government, especially tax officers, to analyse industries that thrive or suffer on specific weather conditions. Knowledge of this relationship may potentially assist decisions in provision of government interventions (e.g. Subsidies, Taxation) for industries during a specific weather condition.

B3. Activists and the environment

This research has the potential to identify industries which are highly affected by different weather conditions, allowing early prevention of industries/goods that have the potential to harm the environment.

II. Dataset and Feature Selection

The research question requires variables measuring weather conditions and consumption of goods.

A. Weather Conditions

Measurements of weather conditions are obtained from Australian Government Bureau of Meteorology on http://www.bom.gov.au/climate/data/stations/. Melbourne Tullamarine Airport was chosen as the weather station to represent Victoria as it serves as the primary airport in Victoria, containing data ranging back to 1970.

A1. Minimum and Maximum Temperature

Monthly mean minimum and maximum temperature are two of the variables chosen to represent changes in weather conditions as they are common indicators of weather. These data are in the format of csv, found in:

- http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=36&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=086282
- http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=38&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=086282

A2. Monthly Total Rainfall

Monthly total rainfall is another variable chosen to represent different climate conditions that could affect sales. It shows the amount of rain and wetness during the month. This dataset is measured by the Melbourne Airport station in the form of csv, found in:

- http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=139&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=086282

B. Measure of Consumption of Goods

Data of consumption of goods is extracted from the Australian Bureau of Statistics of Retail Trade in Australia with data from April 1982 to March 2021 available in:

- https://www.abs.gov.au/statistics/industry/retail-and-wholesale-trade/retail-trade-australia/latest-release#data-download.

The datasets used for this analysis is titled “Retail turnover, state by industry subgroup, seasonally adjusted”. This dataset allows a diverse range of subgroups for analysis and aids in removing external factors (seasonal fluctuations). Retail turnover is defined as the value of consumer purchases on retail in a specified period (MB-International, 2019), which reflects the amount of consumer spending in an industry. Moreover, the term seasonally adjusted refers to the method of smoothing out a time series by minimizing the effect of predictable seasonal fluctuations (Majaski, 2021).

This is in the form of XLS, which contains industry subgroups in all states of Australia:

1. Supermarket and grocery stores
2. Liquor retailing
3. Other specialised food retailing
4. Food retailing
5. Furniture, floor coverings, houseware and textile goods retailing
6. Electrical and electronic goods retailing
7. Hardware, building and garden supplies retailing
8. Household goods retailing
9. Clothing retailing
10. Footwear and other personal accessory retailing
11. Clothing, footwear and personal accessory retailing
12. Department stores
13. Newspaper and book retailing
14. Other recreational goods retailing
15. Pharmaceutical, cosmetic and toiletry goods retailing
16. Other retailing n.e.c.
17. Other retailing
18. Cafes, restaurants and catering services
19. Takeaway food services
20. Cafes, restaurants and takeaway food services
21. Total (Industry)

C. Inflation

This analysis weighs in and adjusts inflation according to the base year. This information is in the form of csv, taken from:

- https://www.in2013dollars.com/Australia-inflation#citation

III. Preprocessing and Data Wrangling

The sales and weather dataset are merged by month and year (e.g. May 2017). The time frame of the analysis starts from January 1982 to December 2020, as 1982 and 2020 is the first and last year where both datasets have complete data of the full year.

The XLS sales dataset contains three different sheets. The second sheet containing the information required was saved as a csv format as it contained the information required.

A. Drop Columns

The sales dataset was first changed by dropping columns of months between its start month (April 1982) and the chosen start month (January 1983).

The weather dataset's columns were also dropped between its start month (July 1970 to January 1983) to follow the month range.

B. Filtering Data

The sales dataset contains data for all states in Australia, hence the function: `.filter(regex='Victoria')` was used in filtering all Victorian related data.

C. Changing Data Types and Format

The weather dataset is transposed to follow the format of the sales dataset. After transposing, the data frame is still in an incompatible format, hence requiring iteration to be appended to a new dataframe in the format of columns by month and year. The new weather data frame is then appended to the sales data frame.

Furthermore, the sales dataset required type conversion from object to float to enable calculation.

D. Removing Outliers

Removing the outliers of all the variables uses stats.zscore() from the library scipy, where outliers are determined when z index is bigger than 3. (CTSPedia: CTSpedia.OutLier, 2021)

E. Addition of Inflation Adjustment

The addition of inflation adjustment requires the sales data frame to contain information about its year, hence the dataframe is appended the year and the price of 100 dollars for that year. Given the price of 100 dollars, the sales in that year is adjusted by the formula:

$$Adjusted\ Sales = Original\ sales * \frac{\text{Value of 100 dollars in base year}}{\text{Value of 100 dollars in current year}}$$

(Webster, 2021)

This allows sales comparison to be adjusted by inflation, removing an external factor in determining the changes of retail turnover.

F. Scatterplot, Correlation Coefficient, Normalized Mutual Information Score

From the dataframe produced, a scatterplot between the industries in 2.B and measures of temperatures are created using `matplotlib.pyplot`. A Pearson correlation coefficient (PCC) and a normalized mutual information (NMI) score between these variables are

calculated using the functions `.corr()` from the pandas library and `normalized_mutual_info_score()` from the sklearn.metrics.cluster library. The calculation of mutual information uses Sturges' rule of [1 + log2n] to determine the amount of bins (Bahret et al., 2021)

IV. Key Findings

Figure 1. Pearson Correlation Coefficient and Normalized Mutual Information of different weather conditions against different industries seasonally adjusted

	Max temp		Min temp		Rainfall	
Industry	**NMI**	**Corr**	**NMI**	**Corr**	**NMI**	**Corr**
Supermarket and grocery stores	0.053	0.125*	0.036	0.070	0.053	-0.069
Liquor retailing	0.049	0.122	0.048	0.068	0.038	-0.043
Other specialised food retailing	0.049	0.104	0.042	0.054	0.037	-0.058
Food retailing	0.058	0.124	0.047	0.070	0.057	-0.066
Furniture, floor coverings, houseware and textile goods retailing	0.040**	0.100	0.040	0.040	0.048	-0.082
Electrical and electronic goods retailing	0.052	0.102	0.055*	0.050	0.053	-0.092
Hardware, building and garden supplies retailing	0.047	0.012	0.034	0.071*	0.057*	-0.032
Household goods retailing	0.050	0.120	0.045	0.061	0.044	-0.077
Clothing retailing	0.047	0.077	0.038	0.033	0.042	0.025*
Footwear and other personal accessory retailing	0.051	0.101	0.048	0.047	0.036**	-0.087
Clothing, footwear and personal accessory retailing	0.042	0.099	0.042	0.044	0.047	-0.020
Department stores	0.035	0.026	0.049	-0.006	0.037	-0.067
Newspaper and book retailing	0.044	-0.116**	0.045	-0.094**	0.052	0.022
Other recreational goods retailing	0.046	0.037	0.045	0.007	0.056	-0.148**
Pharmaceutical, cosmetic and toiletry goods retailing	0.040**	0.112	0.036	0.064	0.049	-0.043
Other retailing n.e.c.	0.058	0.119	0.040	0.065	0.055	-0.058
Other retailing	0.058	0.112	0.048	0.057	0.050	-0.070
Cafes, restaurants and catering services	0.058	0.121	0.045	0.068	0.044	-0.071
Takeaway food services	0.044	0.094	0.033**	0.060	0.046	0.019
Cafes, restaurants and takeaway food services	0.060*	0.118	0.055*	0.068	0.044	-0.046
Total (Industry)	0.053	0.121	0.044	0.065	0.050	-0.064

NMI and Pearson Correlation Coefficient are rounded by three significant figures.
* = Highest Value
** = Lowest Value

A. Maximum temperature

In Figure 1, the maximum temperature data ranges from 0.040-0.060 for the NMI)score and -0.116-0.125 for PCC. With small values of NMI scores, it indicates that there is a small reduction in uncertainty between the variables. However, the diverse range in the correlation indicate that there are differences in terms of strength and positivity in the relationship between higher temperature and sales of different industries. The industry "Supermarket and Grocery Stores" has the highest PCC with 0.125 indicating a weakly positive relationship. The industry "Newspaper and book retailing" has the lowest PCC with -0.116 indicating a weakly negative relationship between the variables. Most industries including but not limited to "Cafes, Restaurants, etc", "Households retailing", "Liquor retailing" are positively correlated with maximum temperature.

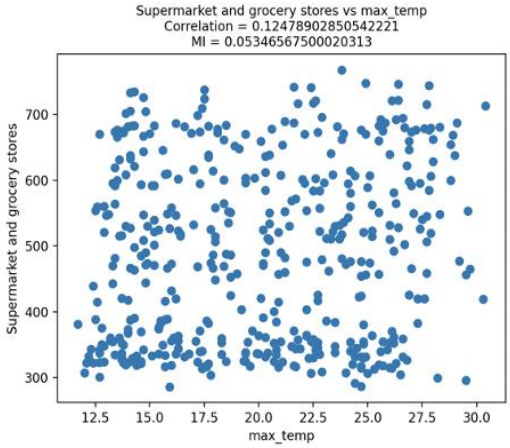

Figure 2.1. Supermarket Grocery Store Industry against maximum temperature

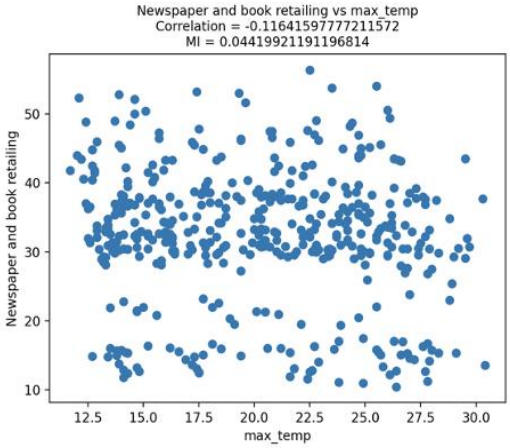

Figure 2.2. Newspaper and book retailing industry against maximum temperature

B. Minimum temperature

The data for PCC value between minimum temperature and sales of different industries ranges from -0.094 ("Newspaper and book retailing" industry) to 0.071 ("Hardware, building and garden supplies retailing" industry). Most industries are positively correlated with minimum temperature. Similar to the analysis on maximum temperature, the NMI ranges from 0.033 and 0.055 which shows very little variability between industries.

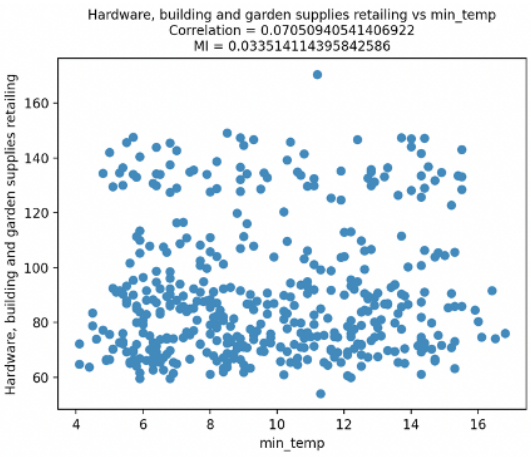

Figure 3.1. Hardware, building and garden supply retailing industry against minimum temperature

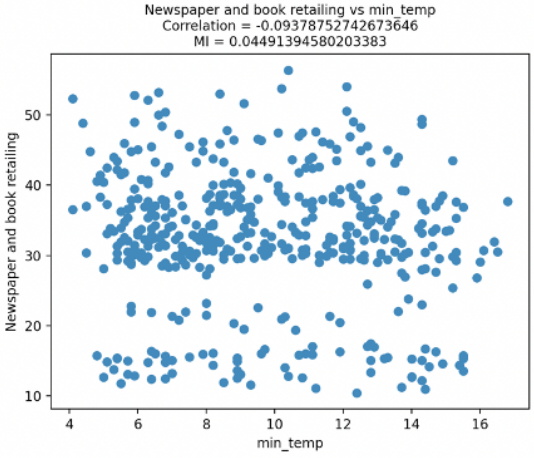

Figure 3.2. Newspaper and book retailing industry against minimum temperature

C. Rainfall

The PCC values for rainfall shows values with most industries having a weak negative correlation ranging from -0.148 ("Other recreational goods retailing" industry) to 0.025 ("Clothing retailing" industry). The correlation values for rainfall differs from the data of the temperature as most of the data is negatively correlated. The NMI for rainfall is similar to the temperature, ranging from 0.036 to 0.057, showing a small reduction in uncertainty between variables.

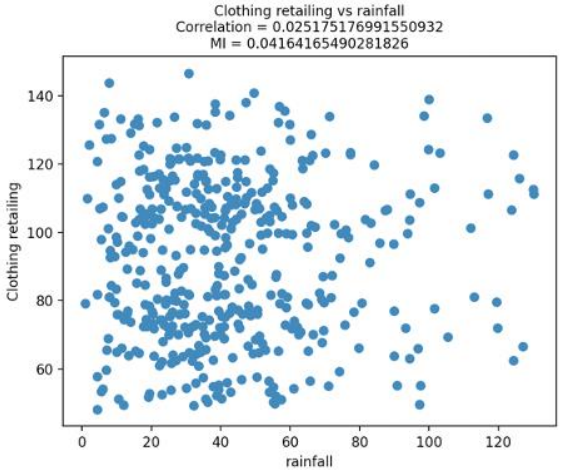

Figure 4.1. Clothing retailing industry against rainfall

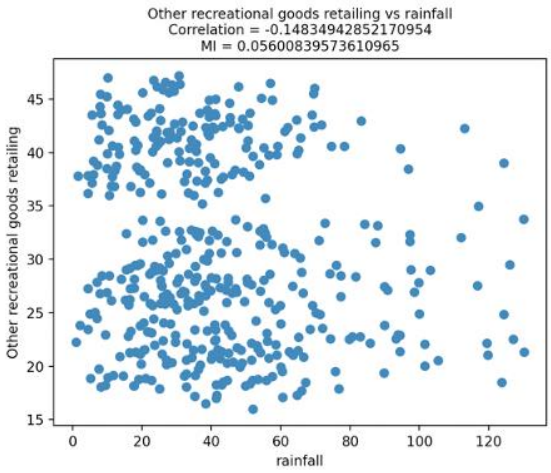

Figure 4.2. Other recreational goods retailing industry against rainfall

V. Conclusion

This research concludes that different industries may have a diverse type of linear relationship with changes in weather conditions. For the measure of maximum and minimum temperatures, it can be deduced that most industries may have a positive weak linear relationship with increasing temperature. The industry "Newspaper and book retailing" is consistent with these variables to be the most negatively correlated industry with increasing temperature. While this analysis shows small correlation, further analysis is required to determine whether there is causality between the variables.

For the analysis between rainfall and sales, there is a varying linear relationship between industries. By PCC, we can infer that there may be a negatively weak linear relationship between more rainfall and most industries. However, some industries such as "Newspaper and book retailing" may have a weak positive linear relationship with more rainfall.

Nevertheless, different industries have a varying linear relationship with different weather conditions, as seen in Figure 1. The measure of maximum and minimum temperatures shows a similar linear relationship whereas monthly rainfalls show a more negative relationship for most industries.

VI. Evaluation

A. Limitations: External Factors

This analysis takes into account external factors affecting sales, such as inflation between years and seasonal hype, adjusting them accordingly. However, this analysis did not consider other external factors such as:

- Discounts, mass promotions
- Incentives by government in specific industries
- Conditions in the economy (e.g. loss of purchasing power)
- Natural disasters

B. Limitations: Weather Station

This analysis uses Melbourne Airport weather station as its source of climate data. Thus, the data gathered might not fully represent the whole State of Victoria.

C. Suggestions for future studies

For future investigation, removing more external factors and using a more diverse choice of weather conditions such as sunlight or average wind speed would allow a more thorough analysis on the effect of weather on sales. Moreover, a more specific time frame (daily/weekly reports) might show a deeper understanding of the effect of the two variables.

Words: 1500

References

Australian Bureau of Statistics. 2021. *Retail Trade, Australia, March 2021*. [online] Available at: <https://www.abs.gov.au/statistics/industry/retail-and-wholesale-trade/retail-trade-australia/latest-release#data-download> [Accessed 15 May 2021].

Bahret, A., Williams, A., Kleyner, A., Meixner, A., Hart, A., Norton, A., Carlson, C., Jackson, C., Stapelmann, C., Craggs, D., Raheja, D., Deeney, D., Lehr, D., Plucknette, D., Schenkelberg, F., Williams, G., Tabasso, G., Hutchins, G., Kovacevic, J., Reyes-Picknell, J., Anderson, J., Paschkewitz, J., Switzer, K., Stewart, K., Gray, K., Warrington, L., Konrad, M., Sondalini, M., Regan, N., Parendo, P., Sage, P., Harkins, R., Allen, R., Latino, R., Kalwarowsky, R., Chan, R., Turcott, S., Wachs, S., Rodgers, T., Syed, U., Know, D., Reliability, S., Podcast, R., Design, Q., Disrupted, M., Podcast, P., Matters, R., Podcast, T., Series, A., Work, A., Notes, C., Career, o., Culture, A., Leadership, E., 2000s, M., Improvement, P., Reliability, o., Management, A., Reliability, C., Asset, C., CMMS, E., RCM, E., Reliability, M., Maintenance, P., Blitz®, R., Project, R., Blog, T., RCA, T., Reliability, o., Reliability, A., Reliability, A., Ridge, A., Topics, M., Insights, R., Technology, R., Safety, o., Insights, C., Applications, E., Techniques, o., Analytics, B., NPD, E., Durability, I., FMEA, I., Concepts, I., 3, T., Academy, T., Reflections, R., DRAFT, R., Authors, A., Resources, F., Publications, F., Education, H., Integration, R., Groups, S., Knowledge, 1., course, R., Course, S., Course, R., Assessment, M., course, P., course, R., course, R., Course, 5., Course, 5., Course, C., Events, U., Listing, C., Webinars, U., Calendar, W., Home, M., Harkins, R., Fiedeldey, M. and Fiedeldey, M., 2021. *Sturge's Rule; A Method for Selecting the Number of Bins in a Histogram*. [online] Accendo Reliability. Available at: <https://accendoreliability.com/sturges-rule-method-selecting-number-bins-histogram/> [Accessed 15 May 2021].

Bom.gov.au. 2021. *Mean Maximum Temperature - 086282 - Bureau of Meteorology*. [online] Available at: <http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=36&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=086282> [Accessed 15 May 2021].

Bom.gov.au. 2021. *Mean Minimum Temperature - 086282 - Bureau of Meteorology*. [online] Available at: <http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=38&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=086282> [Accessed 15 May 2021].

Bom.gov.au. 2021. *Monthly Rainfall - 086282 - Bureau of Meteorology*. [online] Available at: <http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=139&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=086282> [Accessed 15 May 2021].

Ctspedia.org. 2021. *CTSPedia: CTSpedia.OutLier*. [online] Available at: <https://www.ctspedia.org/do/view/CTSpedia/OutLier#:~:text=Any%20z%2Dscore%20greater%20than,standard%20deviations%20from%20the%20mean> [Accessed 15 May 2021].

Majaski, C., 2021. *What Is a Seasonal Adjustment?* [online] Investopedia. Available at: <https://www.investopedia.com/terms/s/seasonal-adjustment.asp> [Accessed 15 May 2021].

MB-International, 2019. *Retail Turnover / Globally consistent and comparable*. [online] MBI english. Available at: <https://www.mbi-geodata.com/en/purchasing-power/retail-turnover/#:~:text=Retail%20Turnover%20measures%20the%20turnover%20of%20local%20retail%20trade%20at,the%20consumers'%20place%20of%20expenditure.&text=Therefore%20showing%20the%20retail%20turnover,retail%20of%20a%20given%20area.> [Accessed 15 May 2021].

Weather-climate.com. 2021. *Climate of Victoria*. [online] Available at: <http://www.weather-climate.com/victoria.html> [Accessed 15 May 2021].

Webster, I., 2021. *Australia Inflation Calculator: AUD from 1923 to 2021*. [online] In2013dollars.com. Available at: <https://www.in2013dollars.com/Australia-inflation#citation> [Accessed 15 May 2021].