Alcohol consumptions: its relation to mortality rate.

DOMAIN: **Health** and **Communities**.

QUESTION:

"Will controlling the amount of alcohol consumption reduces the rate of vehicle collisions and deaths in Victoria? If yes, which type of alcoholic beverage should the government focus on?"

According to Klepova (2018), "In Australia almost one third of all the fatal road collisions are the fault of drink drivers. Yet this ratio it three times higher than in Russia, where people consume three liters more alcohol per capita". However, is alcohol consumption per capita really correlated to vehicle collisions? This study will focus on not only the correlation between the alcohol consumptions and fatal road collisions but also its relation to the alcohol related mortality rate in general. If there is any positive outcomes in any of these relations, the government can use this analysis to address the problem.

DATASET:

In order to answer this question, the first step is to gather relevant datasets related to alcohol consumption and alcohol-related fatalities.
Three main datasets were used in this report:

1. **Apparent Consumption of Alcohol, Australia –** total_alcohol_consumption.csv**.**
   Source: http://stat.data.abs.gov.au//Index.aspx?QueryId=318
   - Dataset contains the amount of alcohol consumption in total and per capita in each year from 2000 to 2016 between 5 different alcoholic beverages in litres. *(See figure 1 for more information)*
2. **Causes of Death, Australia (VICTORIA) –** Underlying_causes_of_death.xls**.**
   Source: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/3303.0Main+Features100002016
   - The original data contains 3 different tables and a "contents" sheet – which contains the index and organization information: Table 3.1 contains the causes of death in Victoria 2016, table 3.2 contains the causes of death in Victoria from 2007 to 2016, and table 3.3 contains the filtered data for each age from under 1 year to 95 years and over in 2016. (*See figure 1 for more information)*
3. **Crashes Last Five Years –** Crashes_Last_Five_Years.csv**.**
   Source: https://vicroadsopendata-vicroadsmaps.opendata.arcgis.com/datasets/crashes-last-five-years
   - The dataset contains the detailed information about fatal and injury crashes in Victoria from 2012 to 2017 and was reported by Victoria Police. *(See figure 1 for more information)*

TYP	TYPE OF VOLUME	MEASURE	BEVT
1	Volume of pure alcohol	Total apparent consumption ('000 litres)	1
2	Volume of pure alcohol	Total apparent consumption ('000 litres)	2

	2007		
Causes of deaths	Males	Females	Totals
Intestinal infectious diseases	11	9	20
Other salmonella infections (A02)	4	3	7

X	Y	OBJECTID	ACCIDENT_NO
145	-37	2693452	T20120013207
144	-37	2693453	T20120013209

Figure 1: Sample Apparent Consumption of Alcohol, Causes of death and Crashes Last Five Years dataset respectively

PRE-PROCESSING:

Since all of the raw datasets contain redundant information that do not help to solve the question, pre-processing step was applied for all datasets in order to select only high value information and also to remove missing values.

total_alcohol_consumption.csv:

Only the following columns: *"Measure"*, *"Beverage Type"*, *"Time"* and *"Value"* were selected because they contain the values that actually help solving the question. For example, *"Measure"* indicates the value which either it is the "Total apparent consumption" or "Per capita apparent consumption". In this report, only "Per capita" data were chosen to be used because: the population in Vic is different each year, the data will be inconsistent if choosing "Total apparent consumption" as a main source to evaluate. "Beverage Type" indicates the beverage type, the data originally comes with the "Total all beverages" measurement, thus this can be exploited to increase time-efficiency. Columns such as "TYP", "Type of Volume", "MEA", etc., are removed since they contain magic number and redundant data. The chosen columns' names were modified by substituting spaces with "_". Since the value is one dimension, box plot then was then implemented to find outliers and does not detect any outliers nor missing values. This data was then filtered to display the result from 2012-2016 to make sure that it is in the same time period with other datasets.

crashes_last_five_year.csv:

Only '*ACCIDENT_DATE', 'DAY_OF_WEEK', 'ALCOHOL_RELATED'*, *'ACCIDENT_TIME', 'INJ_OR_FATAL', 'FATALITY'* were maintained. Only data which are *'ALCOHOL_RELATED'* are chosen. '*ACCIDENT_DATE', 'ACCIDENT_TIME', 'DAY_OF_*WEEK' are the main features that would be used to combine and integrate with other datasets because it indicates time period. *'INJ_OR_FATAL', 'FATALITY',* indicate the total number of injuries and fatalities, fatalities itself. These feature plays as the main role to find the correlation between alcohol consumption and car accidents. All other redundant columns were removed in order to increase time-efficiency. Filtered data was then stored separately in different `DataFrame` objects for calculating and analysis. This made the process of calculating and analysing easier comparing to just modifying a large dataset. In the *"ACCIDENT_DATE",* time format: "hours.minutes.seconds" was then converted to "hours:minutes:seconds" using string replacement in order to make sure that the data will be compatible with `pandas.to_datetime`. The data and then selected only from 2012-2016 in order to match with the other datasets. The resulted filtered data showed no missing values however, there is one suspected outlier when drawing lag plot for *'INJ_OR_FATAL'* **(Figure 2)**. Likewise, when taking the sum of *'INJ_OR_FATAL'*, IQR method found a suspected outlier at 767. However, due to the limit in the number of figures and the data will be incomplete without these 2 numbers. Thus, the suspected outliers were decided to be kept.

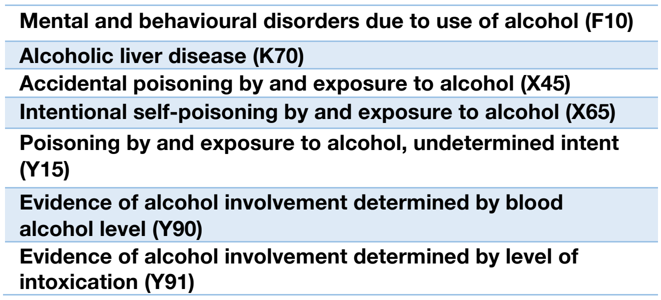
underlying_causes_of_death.xls:

Since the original dataset is in XLS, data transformation was implemented in order to process to further analysis. First, only "Table 3.2" was kept and its decorations were removed using Excels. Likewise, other irrelevant tables were removed by the same method. Then, the data was read using `pandas.read_excel`. Although the data was pre-processed using Excels, there were still limitations that Python could not read the input data correctly. Specifically, columns' names were wrong and displayed as "`Unnamed: X`" – in which `X` is the index of columns; moreover, there were columns with only NULL data located randomly in the dataset. Two functions were then written in order to solve the issue: `remove_missing_data()` and `get_name()`. First, all of the data was passed to `remove_missing_data()`, any columns which contains NULL data was then removed. Seconds, all the columns were passed into `get_name()` to rename as in the original XLS file (*"Males", "Females", "Total"*). Columns *"Males", "Females"* and *"Total"* store the number of male, female, total facilities. Next, causes of deaths were then filtered to select only alcohol related in order to find correlation with the alcohol consumption dataset. And then the filtered data was checked again to make sure that the cause was correctly alcohol involved[1] (**Figure 3)**. Due to the fact that Pandas Excel Reader could not detect the column correctly – in which year were located incorrectly, *Divide and Conquer*[2]was applied in order to re-categorised the data into correct years. To be specific, a number of `DataFrame` objects were created for each year. Data from raw data were then passed into relevant

[1] This might happen because containing the keywords related to alcohol does not mean it was alcoholic causes.

[2] *Divide and Conquer* is a terminology in computing which basically means breaking down the problem into the smaller problems. The solutions for the sub-problems are then combined to get the final solution for the original problem.

`DataFrame` objects. This is somewhat cumbersome but it can ensure that the final result is correct. Afterwards, it was recognized that the *'Total'* contained wrong value. It was recalculated by the following formula: `'Total' = 'Males' + 'Females'`. Then, the total number of alcohol related disease for each year are calculated. IQR and Boxplot were then implemented but detected no outliers. Finally, only data from 2012-2016 were taken in order to make sure all of the data are in the same timeline. This concludes the process of pre-processing.



Mental and behavioural disorders due to use of alcohol (F10)
Alcoholic liver disease (K70)
Accidental poisoning by and exposure to alcohol (X45)
Intentional self-poisoning by and exposure to alcohol (X65)
Poisoning by and exposure to alcohol, undetermined intent (Y15)
Evidence of alcohol involvement determined by blood alcohol level (Y90)
Evidence of alcohol involvement determined by level of intoxication (Y91)

Figure 3: Causes of death that were examined

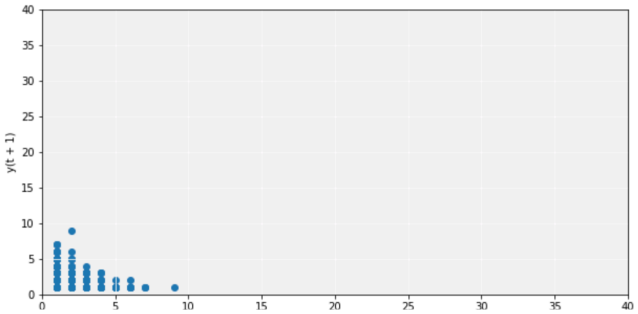


Figure 2: Lag plot for 'INJ_OR_FATAL'

INTEGRATION:

Since the result of pre-processed stage came out pretty successfully, there was not much of the work needed to be done here.

- For *total_alcohol_consumption.csv,* data for total alcohol consumption per capita was then grouped with 'FATALITY' in *crashes_last_five_years.csv* and 'Total' disease in *underlying_causes_of_death.xls*, according to years. After merging, the grouped datasets were normalized in order to carry out correlation analysis.
- For *crashes_last_five_years.csv* feature creation and data integration were implemented. First, since the original data came with just *'INJ_OR_FATAL'*, and *'FATALITY'*, the number for *'INJURY'* is calculable by taking `'INJ_OR_FATAL' – 'FATALITY'`. Second, *'ACCIDENT_DATE'* which contains the date of the accident is merged with 'ACCIDENT_TIME' as 'ACCIDENT_TIME' to facilitate the process of filtering and also makes it easier to compare with other datasets. Because data had been separated into small data frames, all of them are then merged into one large dataset for comparison. The raw data set does not come with any feature counting number of crashes. As a result, a new feature *'Number_of_crashes'* which counts the number of alcohol related crashes for each year was also calculated for each year by counting the number of cases. This is needed because it is also interesting to find out if there is any relation between the alcohol consumption and number of vehicle collisions.
- For *underlying_causes_of_death.xls***,** the pre-processed stage carried out most of the work. There were not too many works that need to be done here excepted from calculating the 'Total' value for each year and then combining all of the divided datasets together. This was necessary because it is the fundamental information to find the correlation with alcohol consumption. This concludes the process integration.

RESULT:

Figure 4 demonstrates the total amount of alcohol consumption (TAC) from 2012 to 2016. It is obvious that there is a slightly decline within 4 years by 0.34 litres from 10.04 to 9.7 with a trough of 9.52 in 2015. According to **Figure 5**, it can be seen that the amount of beer consumption (TB) and the amount of wine consumption (TW) accounted for the highest percentage (40% and 38%). Moreover, when calculating Pearson Correlation between TB and TAC, and between TW and TAC, it shows that both TW, TB and TAC has a high positive liner correlation $r_{TW\text{-}TAC} \approx 0.78$ and $r_{TB\text{-}TAC} \approx 0.96$.

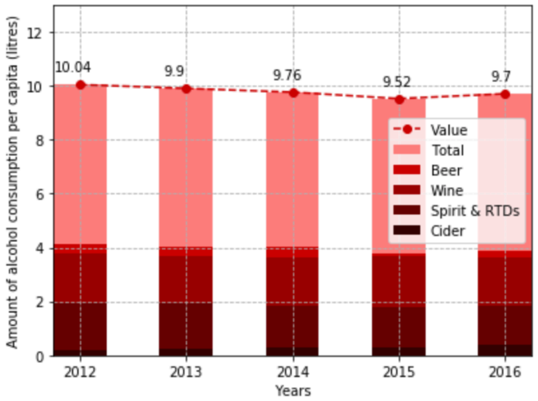


Figure 4: Amount of alcohol consumption per capita from 2012 to 2016

Likewise, the total amount of Spirits and RTDs (TSR) also occupied for 19% and $r_{TSR-TAC} \approx 0.94$. In contrast, total amount of ciders (TC) occupied the less and also has a completely negative linear correlation with TAC $r_{TC-TAC} \approx -0.82$. **Figure 6** shows the heat map of Pearson Correlation between each beverage types. This concludes that if there is correlation between TAC and mortality rate, the beverage type that should be focused on are Beer, Wine, Spirits and RTDs.

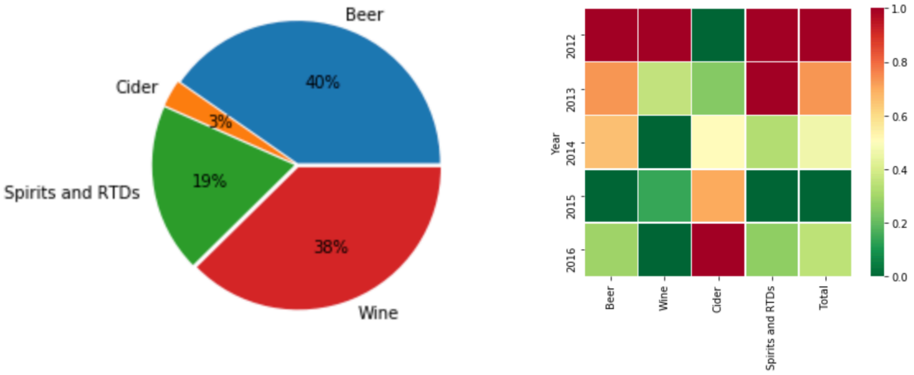


Figure 5,6: Total amount of alcohol consumption from 2012-2016 & Heat map of Pearson Correlation between each beverage type

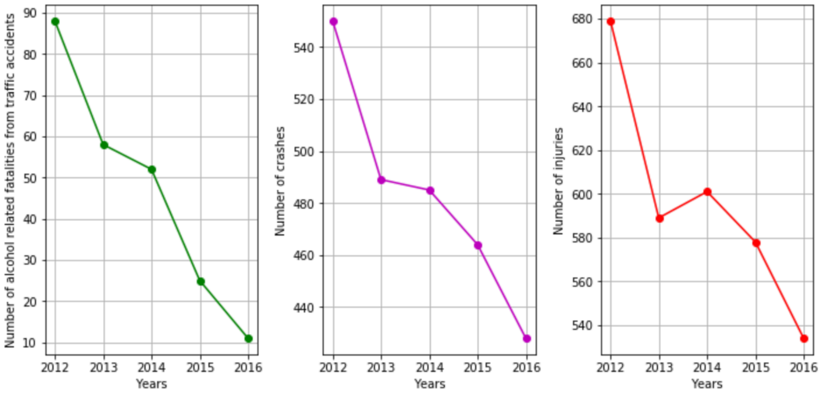


Figure 7: Number of alcohol related fatalities, car collisions and number injuries from 2012-2016.

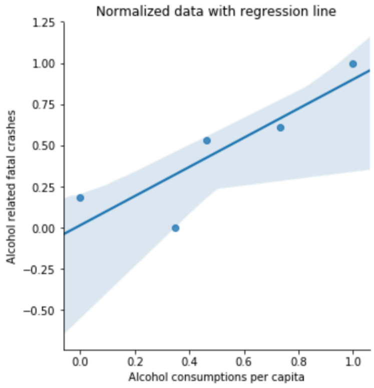


Figure 8: Regression line between alcohol consumption and fatal crashes from 2012-2016

Figure 7 shows that there is a similar fall in three different areas. From 2012 to 2016, the number of fatalities from alcohol, road collisions and total number of injuries dropped dramatically from 88 to 11, 550 to 428, and 679 to 534 respectively. As a result, it is clearly that there is a linear correlation between the 3 features. Which means: if there is a linear correlation between number of fatalities and alcohol consumption. There will be linear correlations between alcohol consumption and two other datasets. **Figure 8** demonstrates the line of best fit when comparing the number of fatalities with the amount of alcohol consumption from 2012 to 2016. Thus, it is clearly seen that there is a positive linear correlation between alcohol consumption and morality rates from traffic accidents. As a result, the total alcohol consumption is also positively linear correlated with number of traffic crashes ($r \approx 0.78$) and number of injuries ($r \approx 0.72$). Likewise, the number of beer, wine, Spirits and RTDs are correlated with the mortalities rate from car accidents with $r=0.9$, $r \approx 0.83$, $r \approx 0.81$ respectively. This concludes that by controlling the amount of alcohol consumption per capita, the amount of vehicle collisions and fatalities from car accidents will be reduced.

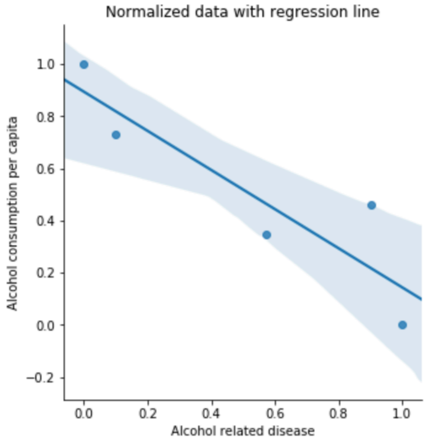


Figure 9: Regression line between alcohol consumption and alcohol related disease

Figure 9 shows the regression line between alcohol consumption and number of alcohol related disease. It is obvious that there is a negative linear correlation between alcohol consumption and alcohol disease by years. As it can be clearly seen from **Figure 10**, in 2012 alcohol consumption accounted for the highest amount, whearas alcohol related disease accounted for the lowest. Similarly, in 2015, while alcohol consumption had the lowest value, total amount of alcohol related disease got the highest value. However, this is understandable because according to **Figure 11**, most of the deaths were alcoholic liver disease (K70), mental and behavioural disorder (F10). These diseases normally resulted from a period of time continuously drinking alcohol. Thus, the relationship between alcohol consumption and alcohol related disease is inconsistance and unconcludable even though there is a negative linear correlation.

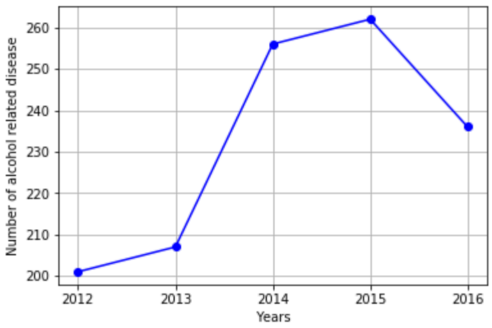


Figure 10: Total amount of alcohol related disease from 2012-2016

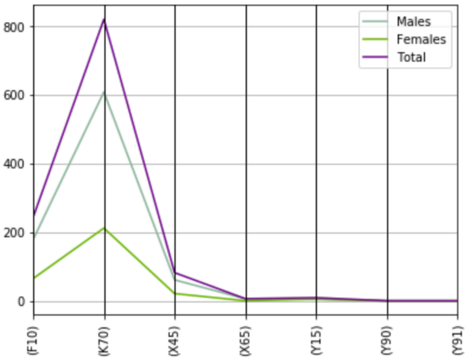


Figure 11: Total number of death by causes from 2012-2016

VALUE

The raw data set contains interesting information however mostly for general, multi-purpose use. If all of the raw data was used without the need of pre-processing, not only run time efficency would decreased but was also impossible for human to interprete and process wrangling based on the the raw data set. Thus, by selecting only necessary and important data, it made the process much faster as well as easier to modify and analyse. While intergration helped to reduce the dimensions of dataset, visualisation made it easier to compare and identify the correlation between different datasets.

CHALLENGE AND REFLECTION

It was quite easy to notice the correlation between alcohol consumption and fatal crashes in the 2 datasets. However when it came to alcohol disease, it was more difficult to find the correlation. K-Mean clustering, Normalized Mutual Information (NMI) was implemented in order to find the correlation but there is no correlation. Time line was extended to find the correlation using both NMI and Pearson correlation method but there is no correlation. Lastly, I tried to calculate NMI and Pearson correlation for each feature in each dataset as well as alcohol correlation with car crashes and disease but both showed uncorrelation. Since there is no correlation, it is really hard to build a learning model and predict future data.

QUESTION RESOLUTION

From the result, it can be conclude that alcohol consumption is highly correlated with fatal crashes, number of car collisions, and car accident injuries. Alcohol consumption can be adjust via controlling the amount of alcoholic products such as wine, beer, Spirits and RTDs. This will help Victoria road & transport department reducing the amount of car accidents.

CODE

Around 300 lines of code was written before wrangling from phase 2. About 529 lines of code was written in Phase 3 mostly for finding correlation and cleaning up the code from phase 2. Finally, this reduced to around 470 lines of code. In total, there is approximately 588 lines of code was written fron scratch. The following library were used: `pandas, matplotlib, numpy, math, sklearn, seaborn, scipy.` Some of the functions were adapted from stackoverflow are cited inside the code file. The details are included in README.txt.

BIBILIOGRAPHY

Klepova, K 2018, 'Too high': 30% of fatal crashes in Australia are due to drink driving, viewed 6th May 2018, <https://www.sbs.com.au/yourlanguage/russian/en/article/2018/01/06/too-high-30-fatal-crashes-australia-are-due-drink-driving>