ALCOHOL AND LIFE EXPECTANCY

An analysis between life expectancy and alcohol consumption

Table of contents

- Selected Topic
- Data source
- Data exploration
- Analysis
- Technologies, languages, tools
- Results
- Recommendations for Future Analysis
- Conclusion

Life expectancy and alcohol consumption

It is common belief that alcohol consumption has grown over the last years, even though its adverse consequences on human health.

- Is this true?
- Has alcohol consumption increased over time? If so, does it have impact on life expectancy?
- Is there a difference between developed and developing countries?

Data source

We searched for data on Kaggle, a platform for data scientists and machine learning engineers to share, discover, and analyze datasets.

We found 2 particular datasets to work on.

The dataset, titled "Relationship between Alcohol Consumption and Life Expectancy" contains two csv files: "Life-expectancy-verbose" and "Drinks".

We also worked with a dataset called “Life Expectancy WHO”, which has information provided by the World Health Organization and the World Bank.

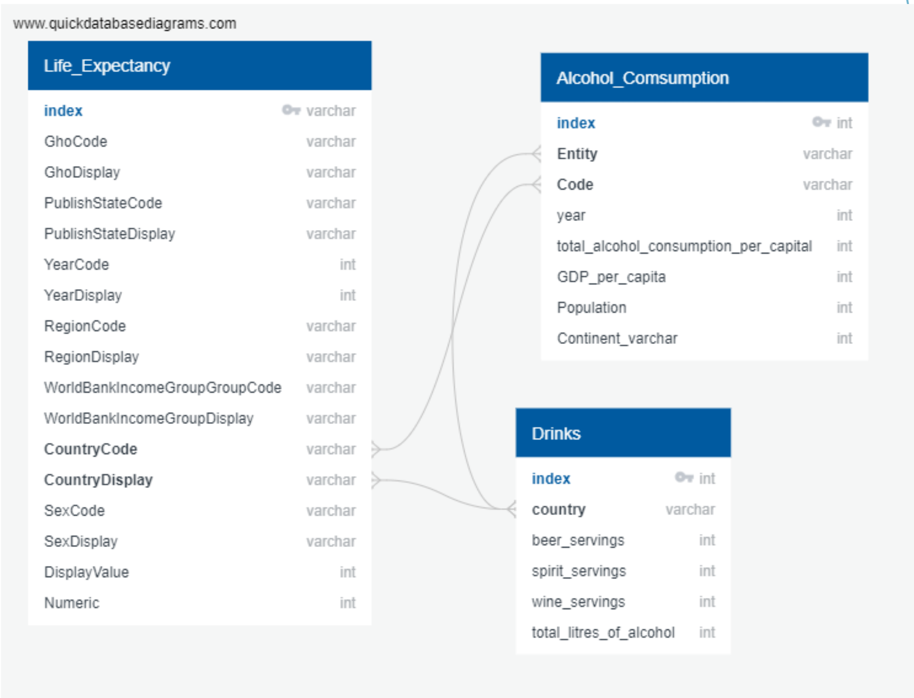
Data Exploration

- Dataset *Relationship between Alcohol Consumption and Life Expectancy*
  - Year variable is not as clean as in the previous dataset and requires to clear rows with erroneous year data. Furthermore, many values are missing.
  - Drinks table has fewer variables and is set to one specific time period.
  - Variables used in this dataset are measured in servings and are classified in beer servings, spirit servings and wine servings.
  - This dataset doesn't include a country code, so it will be necessary to join the country variable with the other datasets.
- Dataset *Life Expectancy WHO*
  - It includes information of life expectancy provided by the World Health Organization. Life expectancy is measured by 3 different variables: Life expectancy at birth (in years) Life expectancy at age 60 (in years) Healthy life expectancy at birth (in years).
  - Additionally, this dataset contains a geographic classification of each country, grouping them in Africa, Americas, Eastern Mediterranean, Europe, SouthEast Asia and Western Pacific.
  - Finally, the Life Expectancy dataset classifies each country into economic groups, according to World Bank standards.
  - Note that dataset is well-organized and has no missing values as for country and year information. Economic classification is null in data of year 2013.

Analysis

- We used the datasets to see the impact and to observe the alcohol consumption over time
- We processed the data using Pandas, dropped null values, duplicates and merged the life expectancy data with the alcohol consumption by country to see if the alcohol consumption have increased over time, and finally get. the data ready for the machine learning model.
- With these two tables we created a database in PostgreSQL (PGadmin).
- In a Jupyter Notebook, we preprocessed the data and merged tables to develop our machine learning models.

ERD



Analysis

- Based on the data of developed countries, we can see that alcohol consumption did not consistently increase over time across all countries. Some countries saw an increase in alcohol consumption from 2000 to 2015, such as Austria, Croatia, Lithuania, and Portugal. Other countries, such as Australia and Japan, saw a decrease in alcohol consumption. Some countries had missing data for 2015, such as Cyprus, Latvia, Luxembourg, Malta, Netherlands, Poland, Romania, Singapore, Slovenia, Spain, Switzerland, and the United States.

- We also found that there seems to be a negative correlation between alcohol consumption and life expectancy by country. Countries with higher alcohol consumption tend to have lower life expectancy, while countries with lower alcohol consumption tend to have higher life expectancy. However, correlation does not necessarily imply causation, and other factors could also be influencing life expectancy in these countries.

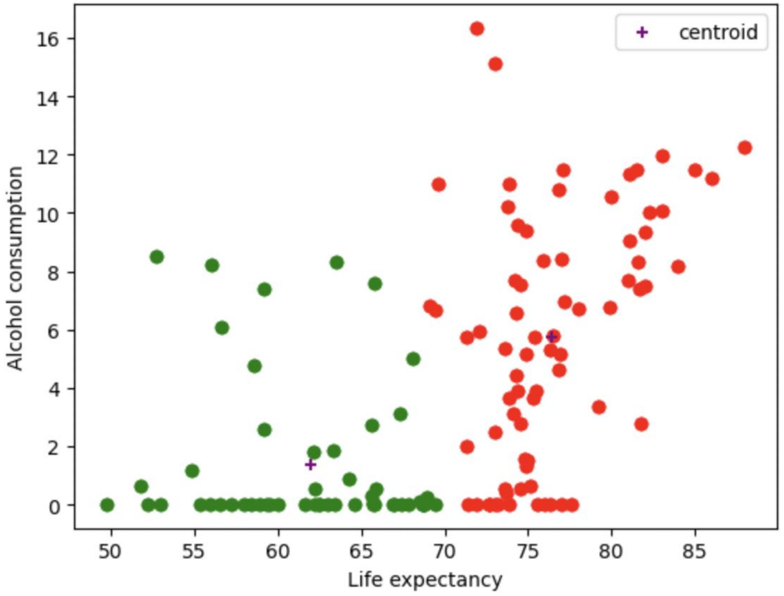
Analysis | Dashboard

- https://public.tableau.com/views/LifeExpectancyvsAlcoholConsumption-LifeExpectancyvsAlcoholConsumption/VerboseDatabase?:language=es-ES&publish=yes&:display_count=n&:origin=viz_share_link

- https://public.tableau.com/views/LifeExpectancyvsAlcoholConsumption-LifeExpectancyvsAlcoholConsumption2/WHODatabase?:language=es-ES&publish=yes&:display_count=n&:origin=viz_share_link

Analysis | Machine Learning

- We started by merging two datasets in a whole database with the purpose of explaining the correlation between life expectancy vs alcohol consumption and other sociodemographic variables that have in big or small proportion relationship to the variable we are trying to explain.
- We created our "*y*" variable as a result of a cluster between people with high life expectancy and low life expectancy by their alcohol consumption.



Analysis | Machine Learning

We trained the (Decision tree and Neural Network) model with the following “*x*” variables:

- 'percentage_expenditure',
- 'BMI',
- 'Total_expenditure',
- 'GDP',
- 'Population',
- 'Income_composition_of_resources',
- 'RegionDisplay',
- 'WorldBankIncomeGroupDisplay,
- 'Status_Developed',

Analysis | Machine Learning | Models

Decision Tree

we got an 0.9375 accuracy score,
so we conclude our model is signifcative.

Confusion Matrix

	predicted 0	predicted 1
actual 0	25	4
actual 1	0	35

Analysis | Machine Learning | Models

Neural Networks

In this AI model we put as an input 2 hidden layers with 8 and 5 neurons respectively and a sigmoid activation function.

Then we compile the model with the following criteria:

- loss="binary_crossentropy"
- optimizer="adam"
- metrics=["accuracy"]

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 8)                 152
_________________________________________________________________
dense_1 (Dense)              (None, 5)                 45
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 6
=================================================================
Total params: 203
Trainable params: 203
Non-trainable params: 0
_________________________________________________________________
```

Then we fitted the model with 100 epochs and finally we got our model loss and model accuracy as follows:

```
64/1 - 0s - loss: 0.7037 - accuracy: 0.5469
Loss: 0.6876508295536041, Accuracy: 0.546875
```

We conclude our NN model is not that accurate given that we exclude from the original table strongly correlated variables with the "***y***" variable such as (Life_expectancy, Alcohol and Adult_Mortality).

Analysis | Machine Learning | Models

Multiple Linear Regression

We created a new table with fewer variables.

Dependent variable "***y***" is : "Life_expectancy_" (49-69 years) and (70-88 years).

Independent variables "x_i" variables are:

- "WorldBankIncomeGroupDisplay" : Socioeconomic status (Low, Medium, High Income)
- "Alcohol" : Alcohol consumption in liters (lt)
- "Status" : Country economic status (Developed or Developing)
- "SexDisplay" : Gender (Male or Female)

We scaled, trained and fitted the variables in order to get the predictions of the "y" variable (life expectancy).

Actual value	Predicted value
73.8	75.472281
78.0	79.443702
72.7	73.576370
74.5	67.473555
76.3	73.938197

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:        Life_expectancy_   R-squared (uncentered):                  0.958
Model:                             OLS   Adj. R-squared (uncentered):             0.957
Method:                  Least Squares   F-statistic:                             1377.
Date:                 Thu, 16 Mar 2023   Prob (F-statistic):                  4.17e-166
Time:                         11:12:29   Log-Likelihood:                        -1017.1
No. Observations:                  248   AIC:                                     2042.
Df Residuals:                      244   BIC:                                     2056.
Df Model:                            4
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            42.7729      3.433     12.458      0.000      36.010      49.536
x2            30.7583      4.845      6.349      0.000      21.215      40.301
x3            44.5339      1.463     30.445      0.000      41.653      47.415
x4             5.7527      1.829      3.146      0.002       2.151       9.354
==============================================================================
Omnibus:                       12.191   Durbin-Watson:                   0.921
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               12.996
Skew:                          -0.542   Prob(JB):                      0.00151
Kurtosis:                       2.715   Cond. No.                         6.98
==============================================================================
```

Technologies, languages and tools

Data process	We processed the data using Pandas Library with Python in a Jupyter Notebook.
Database	We have our datasets in a database we created with PgAdmin and connected to Jupyter Notebook.
Machine Learning	We built our Machine Learning Model using Python in a Jupyter Notebook.
Visualization	We used Tableau to create our graphics and dashboard to show the results.

Results

- For developed countries alcohol consumption did not consistently increase over time across all countries.
- There's a negative correlation between alcohol consumption and life expectancy by country.
- Countries with higher alcohol consumption tend to have lower life expectancy, while countries with lower alcohol consumption tend to have higher life expectancy.
- However, correlation does not necessarily imply causation, and other factors could also be influencing life expectancy in these countries.

Results

- In regards to developing countries there does appear to be a slight increase in alcohol consumption worldwide from 2000 to 2015. However, it's important to note that this trend may not be significant or consistent across all countries, and that there may be other factors affecting alcohol consumption levels in different regions.

- It's also worth noting that while some countries have experienced an increase in alcohol consumption over time, others have actually seen a decrease. Therefore, it's important to look at the data for each individual country in order to fully understand trends in alcohol consumption.

Recommendations for Future Analysis

In order to achieve better results in future analyses, it is advisable to:

- Gather information from more reliable sources, with larger datasets and cleaner data.
- Select a larger set of socioeconomic variables, as well as health variables and, if possible, specific cultural and demographic variables, such as population, religion, among others.

Conclusion

- If we had more time, we would have searched for more information, preferably in official international websites, such as Centers for Disease Control and Prevention (from the USA) or the European Centre for Disease Prevention and Control.

- Additionally, we would’ve tried to improve the precision of the Machine Learning models.

Team members

- Michele Beltrán
- José Antonio Cerda
- Laura Escamilla