DSC 520 Final Paper

Allison Forte

2022-05-22

Does meat consumptions around the world correlate with life expectancy at birth, happiness, or other health statistics?

Research has shown that the high consumption of meat throughout the world is unsustainable for our planet's future. While we know that meat consumption is not a sustainable dietary choice, what else does it influence? Overall, are countries that consume less meat happier? Healthier? Do citizens of countries that eat less meat live longer?

Perhaps this information can help sway the general public and strengthen the argument that consuming less meat is the way of the future. We are already seeing more meat alternatives in grocery stores and a trend toward meatless Mondays and other dietary shifts. Perhaps additional reasons for consuming less meat will help those who are still sticking to a heavily meat diet reconsider.

By analyzing several data sets that provide information about meat consumption by country and happiness, life expectancy, and other health indicators, we will be able to narrow in on a conclusion about meat consumption and how it relates to happiness and overall health using data that can be verified. We could use this analysis to predict how countries' happiness and health might change if their meat consumption did as well.

Research questions to address

1. Is there a correlation between meat consumption and happiness by country?
2. Is the difference in life expectancy in countries with lower meat consumption compared to those with higher meat consumption statistically significant?
3. Does the type of meat influence these relationships?
4. What additional health indicators correlate with meat consumption?
5. Which countries are tending down overall in meat consumption per year? Are these countries trending up in happiness and life expectancy?
6. Across the world is the consumption of certain types of meat on the rise or decline?

By looking closely at the listed questions, the relationship between meat consumption and other variables will be quantified and determined to be statistically significant or not. While performing this analysis, we must be aware of confounding variables that are hiding the true source of the correlation. We will be able to draw initial conclusions from this analysis but additional verification will be required to explore and validate the relationships further.

In order to address these questions, we will combine data from all the data sets and filter out data for countries that do not have information on all variables. We will need to ensure that each data set includes country name or code and use one of those to join the data sets ensuring that they are formatted properly so we can see the meat consumption and other variables for the same year together. Formatting the data properly will ensure our final results are as meaningful as possible.

Datasets that will be used

1. Worldwide Meat Consumption
    - Source: Kaggle.com
    - Description: This data set is a CVS file with 5 columns: location, subject (type of meat), time, measure, and value. This data includes statistics from 1990 through 2026 as it was refreshed in 2018 to include projections. It includes 28 members of the EU along with other countries.
2. World Health Statistics 2020|Complete|Geo-Analysis
    - Source: Kaggle.com
    - Description: This data set has 39 different files and includes 186 columns detailing health statistics throughout the world. Statistics are available since 2000 for 184 coountries.
3. World Happiness Report up to 2022
    - Source: Kaggle.com
    - Description: This data set includes happiness scores from 2015 to 2022 for 158 countries. There are 12 columns in the data set including the country, region, happiness score and rank, life expectancy, and GDP per capita. This data was collected from the Gallup World Survey.
4. ISO Country Codes - Global
    - Source: Kaggle.com
    - Description: This data set contains country codes for 246 countries. This data can be used to ensure all used data sets can be linked as some sort by country coode and others sort by country name. This data set has 5 columns.

Packages that will be used

- jsonlite: to generate and parse json data
- ggplot2: for data visualization
- dplyr: for working with data frames
- pastecs: for data analysis
- pander:
- knitr: for report generation
- xlsx: for reading data
- plyr: for splitting and combining data
- ppcor: for calculating correlations
- tidyverse: contains several other packages including ggplot2 and dplyr which are already listed

Other packages I do not expect to need: DBI and RSQLite: I will not need to connect to a database management system.

More necessary packages may become apparent as the analysis is performed.

Visuals (plots and tables) that will illustrate the results of the research questions

- Several scatter plots will illustrate relationships between variables in my data sets
- Tables can be used to neatly summarize the key statistics and findings from the research

Additional information that will be required

- To complete this project, I first must be able to extract the correct data from the data sets. With the knowledge from the course this should be possible.
- To compare the variables in question I will be able to create scatter plots and run correlation tests. This should be possible with the information from this course.
- Determining which variables are confounding the main variables in question will still be a challenge.
- Understanding how valid the findings are based on how many years of data I can use would be helpful for speaking to the validity of the results.

Milestone 2

Data preparation and cleansing

Final data set

- To prepare the data for analysis it will be combined into one data set. The data will be loaded in to multiple data frames and then specific columns from each data frame will be pulled into one main data frame. The country_codes_df will serve as the base data frame.
- Data will be aligned based on country name or country code. The need to have both country code and country name to match by was the reason country_codes_df was chosen as the base data frame.
- The final data frame has 49 columns but still has rows that have NA values in some columns

```
summary(country_codes_df)
```

```
##  English.short.name.lower.case Alpha.2.code       Alpha.3.code
##  Length:246                    Length:246         Length:246
##  Class :character              Class :character   Class :character
##  Mode  :character              Mode  :character   Mode  :character
##
##
##
##
##   Numeric.code    ISO.3166.2        happiness_2015  happiness_2016
##  Min.   :  4.0   Length:246         Min.   :2.839   Min.   :2.905
##  1st Qu.:215.0   Class :character   1st Qu.:4.518   1st Qu.:4.404
##  Median :429.0   Mode  :character   Median :5.253   Median :5.389
##  Mean   :431.8                      Mean   :5.406   Mean   :5.422
##  3rd Qu.:650.5                      3rd Qu.:6.300   3rd Qu.:6.361
##  Max.   :894.0                      Max.   :7.587   Max.   :7.526
##                                     NA's   :99      NA's   :101
##  happiness_2017  happiness_2018  happiness_2019  happiness_2020
##  Min.   :2.693   Min.   :2.905   Min.   :3.083   Min.   :2.567
##  1st Qu.:4.514   1st Qu.:4.446   1st Qu.:4.530   1st Qu.:4.689
##  Median :5.293   Median :5.378   Median :5.380   Median :5.541
##  Mean   :5.378   Mean   :5.405   Mean   :5.437   Mean   :5.492
##  3rd Qu.:6.168   3rd Qu.:6.272   3rd Qu.:6.205   3rd Qu.:6.250
##  Max.   :7.537   Max.   :7.632   Max.   :7.769   Max.   :7.809
##  NA's   :101     NA's   :102     NA's   :102     NA's   :104
```

```
##  happiness_2021  happiness_2022     life_expectancy_birth_2015
##  Min.   :2.523   Length:246         Min.   :47.67
##  1st Qu.:4.832   Class :character   1st Qu.:65.09
##  Median :5.545   Mode  :character   Median :72.81
##  Mean   :5.534                      Mean   :71.63
##  3rd Qu.:6.282                      3rd Qu.:77.26
##  Max.   :7.842                      Max.   :83.62
##  NA's   :107                        NA's   :84
##  life_expectancy_birth_2019 cancer_30_70_probability_2015
##  Min.   :50.75              Min.   : 8.60
##  1st Qu.:66.64              1st Qu.:14.80
##  Median :73.92              Median :19.20
##  Mean   :72.66              Mean   :19.05
##  3rd Qu.:77.84              3rd Qu.:22.95
##  Max.   :84.26              Max.   :30.80
##  NA's   :84                 NA's   :84
##  cancer_30_70_probability_2016 beef_consumption_2015 beef_consumption_2016
##  Min.   : 8.40                 Min.   :  0.00685     Min.   :  0.00668
##  1st Qu.:14.55                 1st Qu.:  3.27150     1st Qu.:  3.37594
##  Median :18.90                 Median :  8.46238     Median :  8.30800
##  Mean   :18.83                 Mean   : 16.17743     Mean   : 16.05939
##  3rd Qu.:22.90                 3rd Qu.: 16.47186     3rd Qu.: 17.30942
##  Max.   :30.60                 Max.   :175.24360     Max.   :171.29216
##  NA's   :84                    NA's   :203           NA's   :203
##  beef_consumption_2017 beef_consumption_2018 beef_consumption_2019
##  Min.   :  0.00651     Min.   :  0.0062      Min.   :  0.00594
##  1st Qu.:  3.38919     1st Qu.:  3.3919      1st Qu.:  3.38663
##  Median :  8.34366     Median :  8.3728      Median :  8.40554
##  Mean   : 16.39000     Mean   : 16.3836      Mean   : 16.39351
##  3rd Qu.: 17.66352     3rd Qu.: 17.4379      3rd Qu.: 17.38872
##  Max.   :176.02033     Max.   :176.1413      Max.   :177.33365
##  NA's   :203           NA's   :203           NA's   :203
##  beef_consumption_2020 beef_consumption_2021 beef_consumption_2022
##  Min.   :  0.00574     Min.   :  0.00569     Min.   :  0.00563
##  1st Qu.:  3.38734     1st Qu.:  3.39500     1st Qu.:  3.40466
##  Median :  8.46453     Median :  8.54118     Median :  8.61421
##  Mean   : 16.43174     Mean   : 16.49126     Mean   : 16.54266
##  3rd Qu.: 17.42767     3rd Qu.: 17.56228     3rd Qu.: 17.72547
##  Max.   :178.00687     Max.   :178.66266     Max.   :179.29767
##  NA's   :203           NA's   :203           NA's   :203
##  pig_consumption_2015 pig_consumption_2016 pig_consumption_2017
##  Min.   :  0.0000     Min.   :  0.0000     Min.   :  0.0000
##  1st Qu.:  0.5224     1st Qu.:  0.5247     1st Qu.:  0.5329
##  Median :  4.9328     Median :  4.6972     Median :  4.7056
##  Mean   : 17.4027     Mean   : 17.6348     Mean   : 17.5235
##  3rd Qu.: 17.5812     3rd Qu.: 17.3792     3rd Qu.: 16.9571
##  Max.   :249.2984     Max.   :251.2984     Max.   :251.9765
##  NA's   :203          NA's   :203          NA's   :203
##  pig_consumption_2018 pig_consumption_2019 pig_consumption_2020
##  Min.   :  0.0000     Min.   :  0.0000     Min.   :  0.0000
##  1st Qu.:  0.5369     1st Qu.:  0.5416     1st Qu.:  0.5487
##  Median :  4.6350     Median :  4.6414     Median :  4.6912
##  Mean   : 17.5992     Mean   : 17.7373     Mean   : 17.8289
##  3rd Qu.: 16.6826     3rd Qu.: 17.3958     3rd Qu.: 17.3176
```

```
##  Max.   :253.0434     Max.   :254.1102     Max.   :255.1771
##  NA's   :203          NA's   :203          NA's   :203
##  pig_consumption_2021 pig_consumption_2022 poultry_consumption_2015
##  Min.   :  0.0000     Min.   :  0.0000     Min.   :  0.00097
##  1st Qu.:  0.5539     1st Qu.:  0.5567     1st Qu.:  6.17696
##  Median :  4.7329     Median :  4.7266     Median : 15.89952
##  Mean   : 17.9100     Mean   : 17.9698     Mean   : 23.91733
##  3rd Qu.: 17.1095     3rd Qu.: 16.8586     3rd Qu.: 34.43544
##  Max.   :256.2439     Max.   :257.3107     Max.   :142.89600
##  NA's   :203          NA's   :203          NA's   :203
##  poultry_consumption_2016 poultry_consumption_2017 poultry_consumption_2018
##  Min.   :  0.00095        Min.   :  0.00098        Min.   :  0.00099
##  1st Qu.:  6.16977        1st Qu.:  6.21131        1st Qu.:  6.28922
##  Median : 16.35275        Median : 16.33332        Median : 16.36511
##  Mean   : 24.13623        Mean   : 24.29991        Mean   : 24.42863
##  3rd Qu.: 36.16305        3rd Qu.: 36.19965        3rd Qu.: 36.13008
##  Max.   :140.89600        Max.   :141.50000        Max.   :143.00000
##  NA's   :203              NA's   :203              NA's   :203
##  poultry_consumption_2019 poultry_consumption_2020 poultry_consumption_2021
##  Min.   :  0.00102        Min.   :  0.00104        Min.   :  0.00107
##  1st Qu.:  6.36243        1st Qu.:  6.40288        1st Qu.:  6.44180
##  Median : 16.45150        Median : 16.51179        Median : 16.57060
##  Mean   : 24.61494        Mean   : 24.78235        Mean   : 24.94406
##  3rd Qu.: 36.17674        3rd Qu.: 36.28151        3rd Qu.: 36.38753
##  Max.   :144.50000        Max.   :146.00000        Max.   :147.50000
##  NA's   :203              NA's   :203              NA's   :203
##  poultry_consumption_2022 sheep_consumption_2015 sheep_consumption_2016
##  Min.   :  0.00107        Min.   : 0.00907       Min.   : 0.00881
##  1st Qu.:  6.49634        1st Qu.: 0.45825       1st Qu.: 0.45286
##  Median : 16.65124        Median : 1.15876       Median : 1.14466
##  Mean   : 25.12440        Mean   : 2.74647       Mean   : 2.65198
##  3rd Qu.: 36.50915        3rd Qu.: 3.02828       3rd Qu.: 2.66089
##  Max.   :149.00000        Max.   :27.25500       Max.   :27.14442
##  NA's   :203              NA's   :203            NA's   :203
##  sheep_consumption_2017 sheep_consumption_2018 sheep_consumption_2019
##  Min.   : 0.0087        Min.   : 0.00862       Min.   : 0.00869
##  1st Qu.: 0.4497        1st Qu.: 0.44916       1st Qu.: 0.44818
##  Median : 1.1717        Median : 1.17765       Median : 1.18919
##  Mean   : 2.7078        Mean   : 2.78029       Mean   : 2.76186
##  3rd Qu.: 3.0314        3rd Qu.: 3.06789       3rd Qu.: 3.09490
##  Max.   :27.3500        Max.   :27.65000       Max.   :27.95000
##  NA's   :203            NA's   :203            NA's   :203
##  sheep_consumption_2020 sheep_consumption_2021 sheep_consumption_2022
##  Min.   : 0.00871       Min.   : 0.0087        Min.   : 0.00866
##  1st Qu.: 0.44694       1st Qu.: 0.4461        1st Qu.: 0.44533
##  Median : 1.19430       Median : 1.1982        Median : 1.19634
##  Mean   : 2.77773       Mean   : 2.8060        Mean   : 2.82332
##  3rd Qu.: 3.16244       3rd Qu.: 3.1985        3rd Qu.: 3.23738
##  Max.   :28.25000       Max.   :28.5500        Max.   :28.85000
##  NA's   :203            NA's   :203            NA's   :203
```

```
head(country_codes_df)
```

```
##   English.short.name.lower.case Alpha.2.code Alpha.3.code Numeric.code
```

```
## 1                     Zimbabwe           ZW          ZWE          716
## 2                       Zambia           ZM          ZMB          894
## 3                        Yemen           YE          YEM          887
## 4               Western Sahara           EH          ESH          732
## 5            Wallis and Futuna           WF          WLF          876
## 6         Virgin Islands, U.S.           VI          VIR          850
##      ISO.3166.2 happiness_2015 happiness_2016 happiness_2017 happiness_2018
## 1 ISO 3166-2:ZW          4.610          4.193          3.875          3.692
## 2 ISO 3166-2:ZM          5.129          4.795          4.514          4.377
## 3 ISO 3166-2:YE          4.077          3.724          3.593          3.355
## 4 ISO 3166-2:EH             NA             NA             NA             NA
## 5 ISO 3166-2:WF             NA             NA             NA             NA
## 6 ISO 3166-2:VI             NA             NA             NA             NA
##   happiness_2019 happiness_2020 happiness_2021 happiness_2022
## 1          3.663         3.2992          3.145          2,995
## 2          4.107         3.7594          4.073          3,760
## 3          3.380         3.5274          3.658           <NA>
## 4             NA             NA             NA           <NA>
## 5             NA             NA             NA           <NA>
## 6             NA             NA             NA           <NA>
##   life_expectancy_birth_2015 life_expectancy_birth_2019
## 1                      58.48                      60.68
## 2                      60.50                      62.45
## 3                      67.47                      66.63
## 4                         NA                         NA
## 5                         NA                         NA
## 6                         NA                         NA
##   cancer_30_70_probability_2015 cancer_30_70_probability_2016
## 1                          19.4                          19.3
## 2                          18.0                          17.9
## 3                          30.7                          30.6
## 4                            NA                            NA
## 5                            NA                            NA
## 6                            NA                            NA
##   beef_consumption_2015 beef_consumption_2016 beef_consumption_2017
## 1                    NA                    NA                    NA
## 2              9.153843              9.044507              9.043238
## 3                    NA                    NA                    NA
## 4                    NA                    NA                    NA
## 5                    NA                    NA                    NA
## 6                    NA                    NA                    NA
##   beef_consumption_2018 beef_consumption_2019 beef_consumption_2020
## 1                    NA                    NA                    NA
## 2               8.99359              8.925343              8.881105
## 3                    NA                    NA                    NA
## 4                    NA                    NA                    NA
## 5                    NA                    NA                    NA
## 6                    NA                    NA                    NA
##   beef_consumption_2021 beef_consumption_2022 pig_consumption_2015
## 1                    NA                    NA                   NA
## 2              8.947339              9.095192             1.539622
## 3                    NA                    NA                   NA
## 4                    NA                    NA                   NA
## 5                    NA                    NA                   NA
```

```
## 6                    NA                    NA                    NA
##   pig_consumption_2016 pig_consumption_2017 pig_consumption_2018
## 1                   NA                   NA                   NA
## 2             1.493061               1.4599              1.46208
## 3                   NA                   NA                   NA
## 4                   NA                   NA                   NA
## 5                   NA                   NA                   NA
## 6                   NA                   NA                   NA
##   pig_consumption_2019 pig_consumption_2020 pig_consumption_2021
## 1                   NA                   NA                   NA
## 2             1.477567              1.49734             1.517664
## 3                   NA                   NA                   NA
## 4                   NA                   NA                   NA
## 5                   NA                   NA                   NA
## 6                   NA                   NA                   NA
##   pig_consumption_2022 poultry_consumption_2015 poultry_consumption_2016
## 1                   NA                      NA                      NA
## 2              1.53685                2.714077                2.631999
## 3                   NA                      NA                      NA
## 4                   NA                      NA                      NA
## 5                   NA                      NA                      NA
## 6                   NA                      NA                      NA
##   poultry_consumption_2017 poultry_consumption_2018 poultry_consumption_2019
## 1                       NA                       NA                       NA
## 2                 2.613563                   2.5996                 2.602177
## 3                       NA                       NA                       NA
## 4                       NA                       NA                       NA
## 5                       NA                       NA                       NA
## 6                       NA                       NA                       NA
##   poultry_consumption_2020 poultry_consumption_2021 poultry_consumption_2022
## 1                       NA                       NA                       NA
## 2                  2.61267                 2.626448                   2.6433
## 3                       NA                       NA                       NA
## 4                       NA                       NA                       NA
## 5                       NA                       NA                       NA
## 6                       NA                       NA                       NA
##   sheep_consumption_2015 sheep_consumption_2016 sheep_consumption_2017
## 1                     NA                     NA                     NA
## 2              0.5428155              0.5263998              0.5216671
## 3                     NA                     NA                     NA
## 4                     NA                     NA                     NA
## 5                     NA                     NA                     NA
## 6                     NA                     NA                     NA
##   sheep_consumption_2018 sheep_consumption_2019 sheep_consumption_2020
## 1                     NA                     NA                     NA
## 2              0.5217832              0.5248867              0.5303748
## 3                     NA                     NA                     NA
## 4                     NA                     NA                     NA
## 5                     NA                     NA                     NA
## 6                     NA                     NA                     NA
##   sheep_consumption_2021 sheep_consumption_2022
## 1                     NA                     NA
## 2              0.5354517              0.5406727
## 3                     NA                     NA
```

```
## 4                   NA                   NA
## 5                   NA                   NA
## 6                   NA                   NA
```

What do you not know how to do right now that you need to learn to import and cleanup your dataset?

- I need to understand how to transform the happiness_2022 column as it is a character instead of an integer. The data transformations I tried have not worked and resulted in every value becoming NA.
- I need to be more confident with how to clear lines of data that have NA values in key places so they do not interfere with the results.
- While I was able to extract the data I needed form the data frames I suspect there is a cleaner method I could have used that would not require as many intermediary steps.

Discuss how you plan to uncover new information in the data that is not self-evident.

- In order to uncover new information, I will analyze the relationship between different variables. Creating plots and then linear regression models will help to quantify the strength of any relationships that exist.
- Combining multiple data sets will allow relationships to be explored that have not been looked at in past studies.

What are different ways you could look at this data to answer the questions you want to answer?

- Looking at changes in variables over time and seeing the trend will allow us to see if the trends correlate with other trends. It would be interesting to quantify the delta between various time frames and see if change is speeding up or slowing down. In particular, the change in happiness year-over-year can be compared to the change in meat consumption year-over-year and then it can also be compared over a 5 year span.

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

- Multiple data frames have been combined for this analysis. Because we are looking to compare happiness and meat consumption we combined the 2 data frames to make analysis easier. An alternative method would have been to leave the data in their own data frames and compare frame to frame. For the purpose of creating regression models and plots, having the data in one data frame will make this analysis simpler.
- If needed, new variables can be created to show the change over time (example: change in happiness from 2015-2020 and change in meat consumption from 2015-2020) to allow for simple comparison.
- When building regression models I may divide the data into 2 sets so one can be used for training and one for testing.
- A variable for total_meat_consumption per year may need to be created by totaling each type of meat.

How could you summarize your data to answer key questions?

- I will be able to summarize by data by looking at the overall average in meat consumption for each type and the overall average for happiness. By looking at the summary statistics for each year we can see if there are overall trends. Then looking at the same relationship on a country level we can see if the relationship can be verified.

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

- Scatter plots comparing different variables will be most beneficial to start. Using years across the bottom and creating several plots showing happiness over time and meat comsumption over time will help. Over laying these graphs will allow us to compare the changes over time.
- A scatter plot plotting meat consumption by type to happiness could be interesting but looking at total meat consumption compared to happiness would likely be more telling.

What do you not know how to do right now that you need to learn to answer your questions?

- At this time I think I will be able to complete the planned analysis based on knowledge we have acquired throughout this course.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

- At this point I think it could be interesting to use machine learning to be able to predict the happiness score for theoretical changes in meat consumption. If a model can be created, the original data frame for meat consumption could be used for predicting future happiness levels as it contains forecasted meat consumption for the same countries.

Some additional questions you may want to consider asking yourself as you work through this section of the project:

- What features could you filter on?
  - Filtering on region might be an interesting view that could highlight differences by continent or region in addition to any other relationships uncovered.
- How could arranging your data in different ways help?
  - Having the data primarily arranged by country allows us to see all the variables for each country together but sorting the data by year instead of country could highlight some other relationnships.
- Can you reduce your data by selecting only certain variables?

  - Pulling in only the most important variables from the initial data sets has kept the final data frame relatively simple. Filtering out the NA values will simplify the data further.

- Could creating new variables add new insights?
  - Total meat consumption (the sum of the different meat types) could be helpful.

- Could summary statistics at different categorical levels tell you more?
  - Using summary statistics for decade level changes or regional averages could be an interesting view of the data.

- How can you incorporate the pipe (%>%) operator to make your code more efficient?
  - I could likely use the pipe operator when combining my data frames to simplify my code.