Week-11 Submission

Anandita Misra

2023-10-20

Week-9

Data Story Topic

The topic that I have chosen is **Sockonomics Unveiled**:This investigation explores the impact of missing socks in U.S. laundry between 2000-2019 on household finances, shedding light on how this seemingly trivial, unrelated and very bizarre issue can affect both current income and allocation of future expenses of households in the U.S. We evaluate the decrease in current income and change in allocation of future expenses.

Data Sources curated so far

1- The data sources curated so far are from Kaggle on personal expenditures in the U.S. by State from 1997-2019- https://www.kaggle.com/code/davidbroberts/us-personal-expenditures-1997-2019 (https://www.kaggle.com/code/davidbroberts/us-personal-expenditures-1997-2019). There is also a data source from Data.gov which provides data on expenditures, income and demographic characteristics of consumers in the U.S. It provides it in different forms which I have accessed from here-https://catalog.data.gov/dataset/consumer-expenditure-survey-dbf32 (https://catalog.data.gov/dataset/consumer-expenditure-survey-dbf32) . These will tell me how much of income on average of a typical household in U.S. is allocated to clothing.

2- The third data source is Statista from where I have accessed Average annual expenditure on laundry and cleaning supplies per consumer unit in the United States from 2007 to 2022 https://www.statista.com/statistics/305499/us-expenditure-on-laundry-and-cleaning-supplies/ (https://www.statista.com/statistics/305499/us-expenditure-on-laundry-and-cleaning-supplies/).

3- To see the percentage of socks within the apparel expenditure, I have accessed-https://www.statista.com/topics/965/apparel-market-in-the-us/#topicOverview (https://www.statista.com/topics/965/apparel-market-in-the-us/#topicOverview) from Statista on the different categories of expenditure that make up the U.S. Apparel Market.

Week-10

Title-

Sockonomics Unveiled- how do lost socks from U.S. laundry shape American household finances?

Why is it important-

In accordance to one report by Bureau of Labor Statistics, Americans spent $198 per annum on average, on socks and hosiery in 2019. This highlights a financial burden that lost socks may pose on families. For a thriving economy, the financial stability of households is a crucial component. The phenomenon of the missing socks may seem inconsequential, but when we extrapolate across all households(millions in US), it represents a substantial economic loss.

Secondly, by the emphasis on sustainable development goals (UNEP, 2020) missing socks encapsulates a broader consumer behavior and wastefulness. We have to understand the economic impact in order to stress on the seriousness of the matter on wastefulness. This knowledge is important for more efficient household spending habits which are also sustainable. The World Economic Forum (WEF) in 2016 has shown that “seemingly small, everyday choices made by consumers can collectively have a substantial impact on resource use and waste generation.”

Lastly, the allocation of future expenses depends to a large extent on purchasing decisions made today. Missing socks lead to increased sock purchases in the future as replacement. In the context of economic theory, seemingly trivial factors can have ripple effects throughout the economy. Behavioral economics research, cited by the World Bank, indicates that “present decisions can indeed have far-reaching effects on future expenses, often referred to as the”spillover effect” or “decision cascades”“(World Bank, 2015). Hence, in the larger context of household financial planning, studying the economic impact of missing socks is significant. According to the behavioural economist Thaler in 1999,”presence of”friction” in decision-making processes, can influence consumer choices and have broader economic consequence.”

Specific Columns/Variables

Here is a survey that was conducted- This is the dataset- https://gwern.net/doc/psychology/2019-01-21-eric-socksurvey.csv (https://gwern.net/doc/psychology/2019-01-21-eric-socksurvey.csv) These were the survey questions. The data is based on a survey done with an international audience of sample size n=455. It was conducted on a personalized quiz website. For this dataset, I will be using the question (2), (3) relevant to my data story- (2) QN will be used to estimate how many socks will be washed per week. There is a correlation between number of socks owned and number of socks used. Hence, number of socks used will then be correlated to number of socks washed per week.

(3)QN will be used to find out future expenses on replacement socks. We will see the frequency of missing socks and then, take into account the frequency of the replacement through this survey and find out future expenditure on replacement socks.

Hence, column “Count” and “Frequency” will be used.

Do you have enough pairs of socks? Yes/No

How many pairs of socks do you have? (Numeric)

How often do you buy replacement socks? Monthly Semi-annually Annually Less or never

Who buys your socks? Me Spouse/significant-other Relative Other

Once we take this, we will use the sock loss formula to estimate sock losses in a week -

The Sock Loss Formula given by Samsung-

Sock loss index = (L+C)-(P x A)

Higher the value, the more likely you will lose socks. We will use an adapted version of the formula, and keep everything else constant (using standard value) except number of socks washed in a week: Prob= 0.38+(0.005 x L)+(0.0012 x C)-(0.0159 x P x A)3

Where:

L = Laundry size Calculated by multiplying the number of people in the household (p) with the frequency of washes in a week (f). (here f will be taken at average, p will be taken yearly from this dataset- https://usafacts.org/data/topics/people-society/population-and-demographics/population-data/average-family-size/ (https://usafacts.org/data/topics/people-society/population-and-demographics/population-data/average-family-size/) the column will be “Average number of people in a family”)

C = Washing complexity Calculated by adding how many types of wash (t) households do in a week (darks + whites) and multiplying that by the number of socks washed in a week (s) (here t will be held constant and s will be estimated from the above dataset as mentioned)

P = The positivity towards doing laundry Measured on a scale of 1 to 5 with 1 being 'Strongly dislike doing clothes washing' to 5 which represents 'Strongly enjoy doing clothes washing' (we will take this at average)

A = Degree of Attention Which is the sum how many of these things you do at the start of each wash check pockets, unroll sleeves, turn clothes the right way and unrolling socks"

We will take this information, use the sock loss formula, to find out the number of missing socks by year (since we take the number of people in a household by year), then use the single average price of all types of socks in US from the CPI (generalise it to all years to control for price) and again use the average number of people in a household in US (yearly as above in dataset) to get an estimate of the lost current expenditure in that particular year on missing socks. To find future expenditure on replacement socks, we do the same thing, except also take into account the frequency of replacement socks from the first dataset (qn 3), to find the future expenditure arising from these replacement socks, with each particular year as the reference year for future expenditure from that year onwards.

https://www.kaggle.com/code/davidbroberts/us-personal-expenditures-1997-2019 (https://www.kaggle.com/code/davidbroberts/us-personal-expenditures-1997-2019) We then use the column of "Clothing and footwear" from the above dataset to find out the percentage of expenses in missing socks as part of clothing and footwear in years from 1997-2019 and also, as percentage of total expenditure.

Challenges and Errors:

The main challenge in this investigation to find relevant accurate data from surveys conducted. Collecting data on missing socks and their financial or economic impact is challenging. It is difficult to distinguish the direct impact of missing socks from other factors affecting household finances and amongst laundry-related variables like the types of wash, or the positivity towards wash. We are taking these as constant to avoid confusion. We also have to make estimations or take an average number in order to provide a clear pathway of cause-effect of only lost socks on current and future expenditure and no other variable. We make an estimation of the number of socks washed per week from the number of socks owned and take averages like the average price of all types of socks from CPI- Consumer Price Index.

Furthermore, the dataset used does not specifically include a column for "missing socks." Therefore, it is necessary to use proxy variables or devise a methodology to estimate the financial effect which we did through the sock loss formula.

To ensure the accuracy of results, we need to account for potential errors, such as missing or partial data in the surveys, and we extrapolate the survey information of an international audience into the average price of socks, average number of people in a household in US, the average total expenditure, we have to make an assumption to get a constant average number. We also have to figure out a way to take into account, from missing socks, the frequency of replacement socks in order to find an estimate of future expenditure.

Week-11

Changes from past diary entries-

After a thorough research on finding datasets and data sources for Sockonomics Unveiled, unfortunately, I wasn't able to find detailed datasets, meaning- the datasets I found were either about clothing and apparel and the expenditure on them, and not separate data on socks which is too narrow to find. The dataset on socks based on a survey I found had little information on "missing" socks. The most relevant variables were about the number of socks people have and the frequency of replacement. If I use these variables, to find out the number of socks that go missing in laundry on average in a week, I will have to make a lot of estimations based on the sock loss formula that has been established. Many variables there are not available and we can make estimations or assumptions on the variables, but making so many such estimations does not help with the idea of data exploration and visualization on the topic. The data available is too narrow and less to be able to explore or visualize. The sock loss formula has been calculated by Samsung based on a survey on laundry and socks they conducted- but the survey data has not been made publicly available. That was the biggest drawback and forced me to make many estimations about laundry cycles, laundry size, washing complexity, etc. Hence, the estimations and the variables I chose in the last week's diary entry and the exploration I did was valid but due to so many estimations, it becomes very narrow a topic.

This is a new topic I wish to work on-

Finding the right dog for you: Congrats! You're a pawfect match!

Under this topic, I will delve into the evolving popularity of dog breeds and names over a ten-year period from 2007 to 2017, specifically focusing on data from the United States between 2007 and 2017. The primary objective is to gain valuable insights into the preferences of dog owners during this decade and unravel the dynamics of dog breeds and names for people to make their preferred choice before adopting a dog.

Importance of the Project:

Cultural and Societal Insights: Understanding the changing preferences in dog breeds and names provides essential cultural insights and societal preferences- as societies evolve, so do their preferences, and these preferences reflect broader cultural shifts. Analyzing dog ownership trends can serve as a mirror to societal changes in values, lifestyle choices, and economic factors (maybe the popularity was based on price of raising?).

Companion Animals' Role: Dogs have consistently held a unique place in human society as companion animals- they are a man's best friend. Their popularity and the choices made in selecting a particular breed or name can provide insights into the evolving roles of dogs in households. Say, if a german shepherd is being adopted more, maybe the household prioritises safety by choosing a guard dog.These choices can reflect how dogs transition from working animals to beloved family members, influencing our choices and responsibilities towards them.

Responsible Pet Ownership: The United Nations, through various agencies like the Food and Agriculture Organization (FAO), emphasizes responsible pet ownership. Studying the popularity of dog breeds and names can highlight trends in pet ownership, which can be related to issues like breed-specific legislation, pet health, and animal welfare. It can promote responsible choices regarding dog breeds and emphasize the importance of providing the best care for pets. Many times because of an imperfect (or impawfect) match, dog owners in the US reduce their care on pet health and often times, even give their dogs up for adoption. The importance of finding a perfect match goes beyond aesthetics and comfort, but on compatibility to avoid disappointing results later since dogs develop symptoms of attachment and disattachment veru frequently.

Implications for the Pet Industry: Changes in dog breed and name preferences can have far-reaching implications for the pet industry. From dog food and grooming products to pet accessories and services, the industry relies on understanding consumer preferences. Analyzing these preferences can help pet-related businesses tailor their offerings to meet the evolving demands of dog owners. This is much more relevant nowadays with even student entrepreneurs even at NUS, working on pet products.

Global Perspectives on Pets: The United Nations recognizes the importance of pets and animals in achieving the Sustainable Development Goals (SDGs), such as promoting well-being and responsible consumption. By studying dog ownership trends in the United States, this project can contribute to a broader understanding of global pet ownership dynamics and their impact on sustainability and human well-being.

The datasets I have found are specific to US from 2007 to 2017. I will be taking two years to show how the popularity changed over a decade.

We will be focusing on dog breeds and dog names.

Starting with the 2017 dataset-

I’ll be using the variables of “Color”, “Breed”, “DogName”- with this I will be assessing the popular breed, popular color and the number of dogs with names starting with a particular letter.

I will also be using a combination of the variables “Color” and “Breed” in the sense of the AND binary operator to find out which color and breed combinations are most popular.

I will be doing the same for the 2007 dataset- and the visualizations will show the changes over a decade in the US.

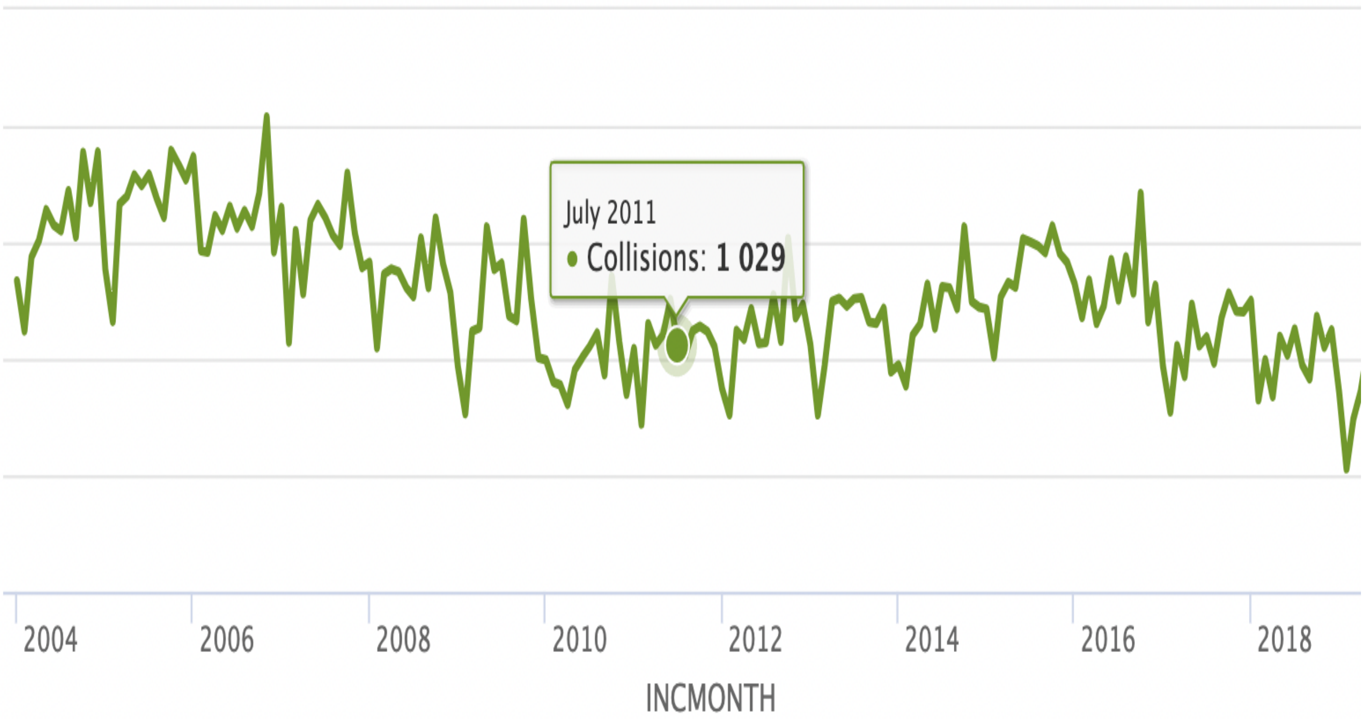
To go deeper, I will also be assessing the variable “LicenseType”. This will help to show preferences for dogs more clearly.

Visualizations to be used-

For Data Visualizations- 1) Popular dog breeds over time- I will be depicting the change from 2007 to 2017 using a line chart and also adding a trendline. Here, it is useful for me to use the datasets for years in between 2007 to 2017 to better depict the time series change. This will be a bit complex since there are different datasets. (I will have to create a new dataset from the existing ones.) But, I will be taking the top 10 breeds of 2017 (latest year available to me) and then assessing the popularity of those breeds in different years from 2007 to 2017. I will be having the “year” on the x-axis and the “Breed” on the y-axis. There would be separate lines for the different breeds and hence, we can clearly compare the trend. (categoric variable) 2) Popular dog colors- I will create two pie charts to display the distribution of dog colors in 2007 and 2017 to see shifts in color preferences. (categoric variable) 3) Color and Breed Combinations- I will use a heatmap to represent the frequency of different color and breed combinations. This will help me identify the most popular color-breed combinations.For this maybe, subpanels through facets can be used. (categoric variable)

Layers of ggplot2 like labels, aesthetic mappings like colour- colour will be utilised to distinguish between different trendlines (colour in visualization 1). I can integrate these graphs in shiny app to actually show these changes by different years- the year will be chosen in the sliderinput. To do this, I can maybe change the graph where the slider input will adjust it by the width of year- number of bins but it would show by divisions of 1 year, 3 years, or 5 years. When the pointer hovers over any point on the trendlines, the exact number can be shown alongwith the year as well. It can be something like this-

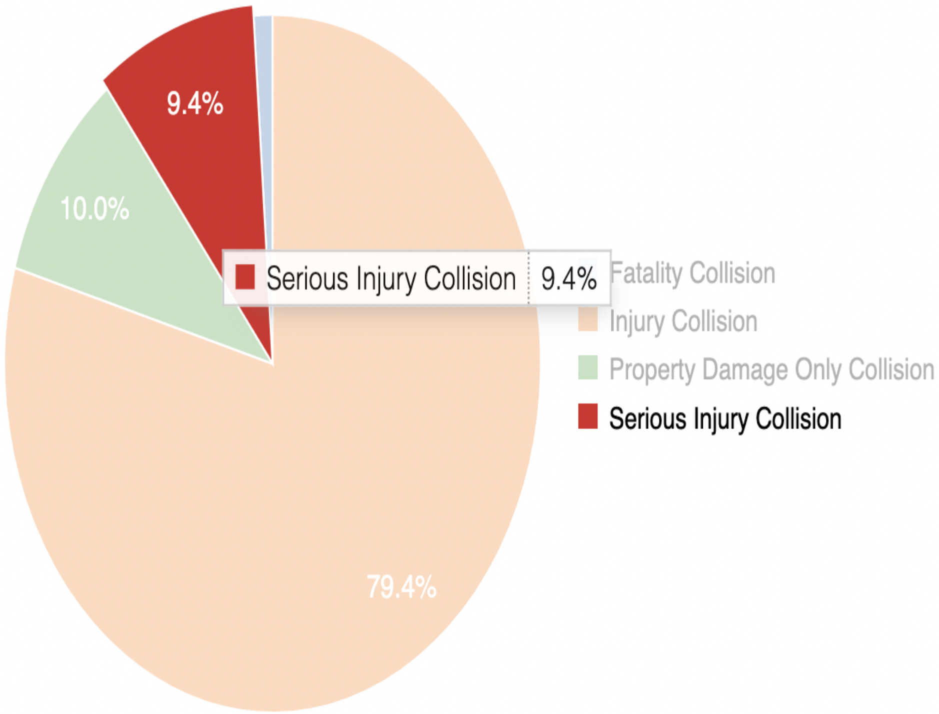
```
knitr::include_graphics("1.png")
```



Insert caption here

Pie chart can be made interactive by adding a brief description when the pointer hovers to the piece of pie- the description can be specific to dog breeds within that colour.

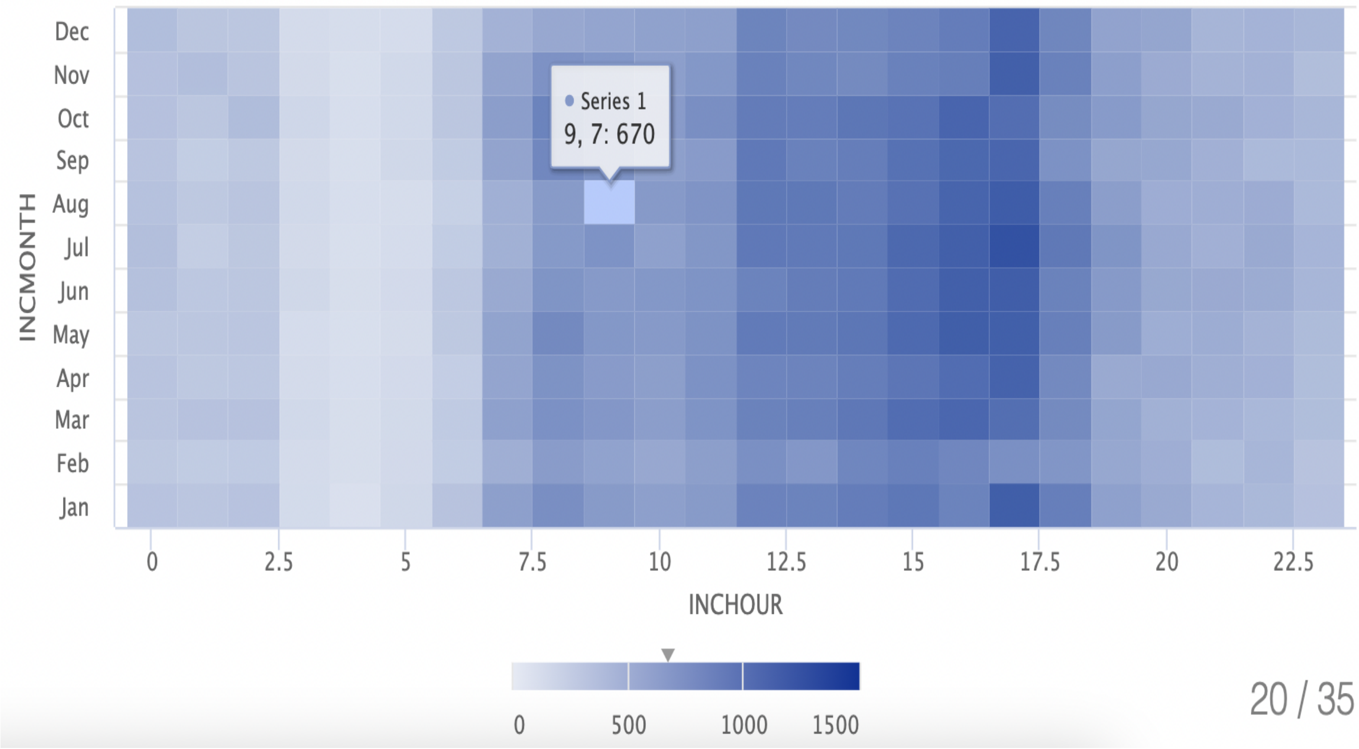
```
knitr::include_graphics("2.png")
```



Insert caption here

Even the heatmap can be made interactive as such using Shiny-

```
knitr::include_graphics("3.png")
```



Insert caption here

I wish to use Shiny to make a complete interactive dashboard which includes popularity on all aspects of colour, breed, License Type, name, so that users can choose any filter or adjust any filter they like and see the frequency of dogs according to that filter. I would also like to use Shiny for a whole comparative analysis between 2007 and 2017.

I would use the sidebar panel to put in the filters there uaing sliderinput- of names of breeds, colours from cold to warm, types. If difficult, I would choose top 10 breeds or top 10 license types to narrow down filters. it will be a difficult undertaking so hopefully, with narrowing down, it will be a bit easier.

Concepts-

1. Choice of right visualization and using ggplot to create the line chart, adding trendlines for separate breeds, pie charts, heatmap - Week-7
2. Using the shiny app and customizing of the sidebar panel for the filters, adjusting of the slider input for the different values or names of breeds, type, colours, etc. All changes to the user interface and also, adding 3 dog images, using and learning from all examples- Week-8
3. Integration of Shiny dashboard with Quarto- Week-8
4. Adding a new tab or page to Quarto Website - Self-taught
5. Choosing a custom template on website for image and text with a divider like on my “Home” tab for the web page- Self-taught
6. Adding icons embedded with links on quarto website- Self-taught
7. Adding a flowchart which can be customised in any way on Quarto wesbite for summary- this will be an important summary aspect for my webpage- Self-taught
8. Making a heatmap- Self taught
9. Making a pie chart- self taught
10. Making a line chart with different lines for different breeds- Self-taught

Challenges faced

- The challenge that I faced mostly was with the integration of shiny app on quarto website. The errors were confusing. For now, I have integrated it. I was able to load the project again and the iframe tag started working since I specified it was html or block in the brackets after the ticks like ```{block}. After that, I am still getting an error due to python. I installed it and jupyter as well, but am not able to figure out where it is being used. Hence, for now, I have added proof of integration along with the basic placeholders/framework for the visualizations and write up.
- I needed to change my topic and it was challenging to assess if data is sufficient or not. I had to take the decision since data was too narrow. Hence, I had to redo the importance of the project, etc and find relevant datasets again.
- It is difficult to find a way to combine different datasets together. I had to create a new csv file for the top 10 breeds of the years from 2007-2017 so that I have a timeseries graph showing the changes in popularity from 2007-2017.
- It was difficult to learn how interactivity can be brought about- I searched for different graphs which were interactive so that I could have an idea of the possible ways to make it interactive- not in terms of how to do it but what features can be added. I have also added these graphs- this is the source- Yollin, B. (n.d.). Interactive charts in shiny. https://byollin.github.io/RInteractiveCharts/#20 (https://byollin.github.io/RInteractiveCharts/#20)
- It is difficult to combine the columns of colour and breed to take both into account together in the heatmap. For that, a new column must be created of colour plus breed and can be shown along years. What we can also do to fix this challenge, is take colour on one axis and breed on another axis on the heatmap- choosing a heatmap helps solve the problem to a large extent instead of getting complex with data.