Research Project DSC520

Nicholas De Santos

2023-05-10

Step 1: Introduction and Planning

Introduction:

My main topic for this research project will be on focused on the detection of ASD (Autism Spectrum Disorder) in young adults/adults. It's known that as you get older it typically gets more difficult to get diagnosed with ASD. Experts attribute this to the fact that as children with undiagnosed ASD get older they develop better masking techniques in order to "fit in" with "neuro-typical" individuals. My research will cover current methods used to attempt to evaluate current ASD diagnostic tests as well as investigate whether those methods can be improved.

Research questions:

Are there certain characteristics that are more typical in individuals with ASD? How accurate are current diagnostic tests for ASD? How can we estimate the number of undiagnosed individuals with ASD?

Approach:

My approach for these research questions is to look at some data avaiable for popular ASD diagnostic tests. For example one test that I have been seeing is the Autism Quotient test (AQ test). I'm interested in this test in particular because many of the inputs are numerical but categorical variables are also included in many datasets to see if certain mixtures of parameters could help predict a person's diagnosis. Including more categorical data can also help determine if there are differences inc ertain groups of people getting diagnosed or not (perhaps a bias).

How your approach addresses (fully or partially) the problem:

As for the accuracy of current diagnostic test I hope to find data that includes both accurate diagnosis as well as resutls from the test to compare. As for the question of estimating the number of diagnosed people with ASD I will primarily rely on good trends data for that. I believe google trends data is useful considering people that search for "autism quiz" or "am I autistic?" can shine a light on how many people think they could psosibly be autistic and we can cmpare that number to reported diagnosis.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows):

- https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults?resource=download (https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults?resource=download)
- https://psyteachr.github.io/ug1-practical/aq-data-and-recap.html (https://psyteachr.github.io/ug1-practical/aq-data-and-recap.html)
- https://www.cdc.gov/ncbddd/autism/data.html (https://www.cdc.gov/ncbddd/autism/data.html)
- Also planning on looking into https://trends.google.com/home (https://trends.google.com/home) data for more insight

Required Packages:

I mostly plan on working with base R and of course ggplot2 for visuals. I'm sure I will also need dplyr in order to organize whatever datasets that I find online. Most of the data that I have found thus far is excel data so I know I will also need readxl.

Plots and Table Needs:

For tabels, one possible table I would need would be the data explaining the AQ question results. The question answeres need to be converted to numerical values in order to be analysed. As for plots I believe categorical plots would be needed mostly because the result we're looking at is binary (either they are diagnosed or not). We could see a comparison of what groups are more likely to be diagnosed with ASD.

Questions for future steps:

As of right now I don't have too many questions about the research project. Right now I'm mainly focused on exploring this research topic. I was planning on doing a different research topic but I actually just recently was diagnosed with ASD. And it just had me thinking if there can be better ways to diagnose people. Preferably earlier in their life of course but seeing that many people go undiagnosed past their childhood years, I want to see if there are better ways to "tell", if you know what I mean.

Step 2: Working With Data

How to import and clean my data

```
library(readxl)

#Reading in data: AQ Data
setwd("E:/Academics/Masters/DSC520/Research Project/archive")
AQ_data <- read.csv("autism_screening.csv")

#Reading in data: Autism By State
setwd("E:/Academics/Masters/DSC520/Research Project")
disability_data <- read.csv("Disability_by_state.csv")

#Reading in data: Google Trends Data
disorder_data <- read_excel("asd_ocd_adhd.xlsx")
symptoms_data <- read_excel("symptoms_asd_adhd_ocd.xlsx")
```

As far as cleaning data, I looked into the AQ data and the Autism by state data and there doesn't seem to be any missing values or strange data entries. I can always look into it using R features but I believe I have a lot of work to do with fixing missing data with the Google Trends data. The data that is missing is date data so I haven't decided if I just want to get rid of data since there's no way to find it or just guess based on the dates around the missing data entries. Since the exact data isn't necessarily important I could also just work with months and years instead.

What does the final data set look like?

As far as the AQ data set, it's pretty final. The disability data set needs to be filtered out to just include information for autism since that is the topic of my research. And lastly the Google trends data will be all filled in with dates.

What are different ways you could look at this data?

As far as the AQ data you could split up the data by different categories/demographics and analyze it based on different subgroups or just the entire sample population as a whole. For the disability data, we will just be looking at the prevalence of autism by state. Compared to the Google trends data that will most likely be an analysis over time.

How could you summarize your data to answer key questions?

I do plan on having some sort of accuracy table based on the results of the AQ test as well as possibly creating a model that details what parameters in the data set could be most significant in predicting whether a person is diagnosed with ASD or not.

What types of plots and tables will help you to illustrate the findings to your questions?

The Google trends data is mainly the data that I want to plot over time. It will sort of help me illustrate the prevalence of ASD along with other disorders based on Google searches and if there's possibly any correlation between different disorder that could assist in diagnosis.

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

I think it would be interesting to incorperate machine learning into my research. I could attempt to made a model (especially for the AQ test results) that predicts whether or not a person is diagnosed with ASD.

Questions for future steps

As of right now I don’t think I have any questions for future steps. I’m just really excited to dive further into this research.