ADHD: EXPLORING & PREDICTING RELATIONSHIPS & OUTCOMES

THE EFFECTS OF DEMOGRAPHICS AND BEHAVIORS/HABITS ON ADHD

ADHD AND TOURETTE SYNDROME PREVALENCE

Madeleine Sharp

8.1 Project 2: White Paper

Business Problem:

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common neurodevelopmental disorders (CDC, 2022). It often manifests, is first seen, and is diagnosed in children and persists into and throughout adulthood (CDC, 2022). Given that it is one of the most common neurodevelopmental disorders (CDC, 2022), it behooves the medical community to continually work to diagnose and treat ADHD. Not only that, but because ADHD diagnoses can be tough to nail for all individuals, part of the problem lies in being able to accurately identify ADHD symptoms across the spectrum of those symptoms.

Currently, a discrepancy likely exists between those formally diagnosed with ADHD and those who exhibit potential ADHD symptoms but have not been formally diagnosed (CDC, 2022). In other words, a good portion of individuals that have ADHD may not be medically diagnosed - and this could be for a variety of reasons.

In addition to the complexities of ADHD, complexities also lie in Tourette Syndrome (TS) diagnoses. TS, like ADHD, can only be diagnosed via observation – no blood or other biomarker tests can be conducted to ascertain a diagnosis (TAA, 2022). Interestingly enough, TS is a common co-diagnosis with ADHD (and vice versa), but not always. TS is more poorly understood than even ADHD is, and additional medical and demographic research in these areas is critical.

Background/History:

In previous studies and research, findings have indicated that diagnoses for ADHD do not occur in the same manner between females and males. More specifically, males are more likely to be diagnosed with ADHD than females (12.9% to 5.6%, respectively) (CDC, 2022) some of the reasons for which may be explained by the exhibition of differing symptoms as a result of biology between these two groups, but also perhaps due to other social, socio-economic, or other unassuming reasons (Bitter et al., 2010; Kinman, 2016).

As an example, females have been noted to exhibit more subtle, internalized symptoms (such as inattentiveness) whereas males tend to exhibit more outward, externalized symptoms (such as running, rambunctiousness, and impulsivity). Whether these differences are due to sheer biology between females and males, or whether they are due to the socialization of how females and males should "act" with respect to societal norms for their sex group and gender representation/portrayal, is not fully clear. Either way, the ways in which symptoms manifest for these two groups makes it more difficult for medical practitioners to diagnose females with ADHD (Bitter et al., 2010; Kinman, 2016).

This amounts to fewer females than males being accurately diagnosed, and therefore may result in undiagnosed females encountering a variety of issues later in life down the line; problems in social scenarios, personal relationships, self-esteem, and mental health such as depression, anxiety (social or otherwise), and eating disorders (Bitter et al., 2010; Kinman, 2016). On the flip side, given that males in general (either biologically or from socialization) tend to be more impulsive and greater risk-takers, males may be over-diagnosed with ADHD.

Project Overview & Research Questions:

For this specific project, I wanted to assess ADHD diagnoses and the demographics surrounding that to determine what that landscape looks like. Additionally, I wanted to look at Tourette Syndrome as a comorbidity to ADHD, as well as a variety of factors in relation to those children with ADHD, including overall ADHD information, demographic information, TS information, mental health information, behavioral health information, and overall health information.

I also wanted to explore whether ADHD diagnoses can be predicted based upon the above sets of specific factors.

Based upon that, my main research questions are as follows:

1. What do the overall demographics and health status looks like for children with ADHD?
2. What is the relationship between those children who are diagnosed with ADHD and those diagnosed with TS?
3. Can predictions be made as to an ADHD diagnosis based upon factors such as general health, mental health, demographics, TS diagnosis, etc.?

Hypotheses:

Overall, I hypothesize that more males will have an ADHD diagnosis than females, those who have ADHD will have a higher prevalence of being diagnosed with TS as well, and that ADHD diagnosis (as a dependent variable) will be dependent upon certain demographic or health factors (independent variables).

Data Explanation – Data Preparation & Data Dictionary:

The data selected for use is the National Survey of Children's Health (NSCH) 2019-2020 dataset. This dataset combines two (2) years' worth of survey data and includes features that cover data collection in the areas of child and family health measures, national performance and outcome measures, and subgroups. This is the most recently published dataset from the NSCH, derived from The Child & Adolescent Health Measurement Initiative, published in January of 2022. To access the data, click the link (https://www.childhealthdata.org/dataset/download?rq=10351) and navigate to the DRC Resource section to access the 2019-2020 data.

The sub-measures that comprise this survey include: physical, oral health, and function status, emotional and mental health, health insurance coverage, health care access and quality, community and school activities, family health and activities, neighborhood safety and support, child and family demographics and subgroups, national performance measures, and national outcome measures. Of these subgroups, the ones I am most interested in for completing my project analyses include: overall physical health, emotional and mental health, health care access and quality, demographics, TS prevalence, and ADHD diagnosis outcomes.

The number of observations (n) included in this dataset is 72,210.

The NSCH is a web and mail-based survey from the Census Bureau. The survey includes each of the 50 states of the U.S. plus the District of Columbia. Its aim is to gather data about children's health nationwide. It is weighted to be representative of non-institutionalized children in the U.S., mitigating bias. It also contains a codebook corresponding to the values found for each variable in the dataset, allowing for easy interpretation of the data. These codes/interpretations will be utilized and taken into account during discussion of and work related to data variable encoding.

The dataset includes 72,210 participants ages 3-17 with demographic characteristics, including poverty level of parent(s)/guardian(s), and gender. The dataset is stored in a CSV file with columns being variables and rows being observations, so it is easy to parse and wrangle the data. The time period of which I am assessing the data is 2019 and 2020, since it is the most recent and thus the most applicable to today.

For my project, I wanted to use variables that assessed and were representative of a few sub-areas/sub-topics within this dataset. These areas include variables encapsulating information pertaining to ADHD information, demographic information, Tourette Syndrome information, mental health information, behavioral health information, and overall/general health information.

Given what I wish to assess, and given the background information on ADHD, I have deemed these areas to be the best fit for the scope of my project and exploring my research questions.

The specific variables that make up the above categories, as well as their meanings and descriptions, can be found at the end of this document in Appendix B.

Data Preparation: Selection, Cleaning, and Transformations

Given the data comes in a CSV format and given that the rows are observations while the columns are the variable names/type, this data is already in a "pre-tidied" format. While this dataset can be looked at aggregately, it encompasses individual-level granularity, where rows are individual participants (observations). The unique values seem to be consistent, with only numeric data or string data in each column, so it is relatively clean.

The main issue with the data is that it includes over 800 variables (820, to be exact), and therefore it includes lots of extra and unnecessary information that likely is not useful to the project task/business problem at hand. Therefore, to handle this portion, I narrowed down the dataset to only those variables that were meaningful in the context of this project and my research questions (hand-selected). This primarily consisted of keeping only those variable related to ADHD and specific demographic characteristics.

The finalized set of variables used for this project can be found in Appendix B of this document.

One benefit of working with this dataset is that the variables were essentially already numerically encoded, making it easier for analysis with categorical variables. It is important to note that (per the accompanying data codebook) with each of these variables, 90, 95, 96, and 99 indicate missing data, depending upon which variable is in question. Therefore, as a part of my missing

data cleansing, I needed to identify these “missing value” codes within the dataset, replace them with NaN, and then remove those NaNs (null values).

After the final data variable selections, a variety of data cleaning and transformation steps needed to occur, including renaming the variable column names for readability and better understanding, assessing for and handling missing data, ensuring the data types are appropriate, ensuring the values are encoded properly per the accompanying codebook (and so that analyses can run more smoothly), and checking for any unique variable values to ensure legitimacy. Once these steps were complete, approximately 46,000 rows of data remained with 20 variables, and the final data-frame was ready for use within the next analysis steps.

Methods:

The analysis methods that I utilized for this project included Exploratory Data Analysis (EDA) methods, as well as predictive measures using classification Machine Learning (ML) models.

For my EDA, I focused on descriptive statistics, distributions, and relationships between the variables. This step involved visualizations of the data with graphs and heatmaps.

For my ML efforts, I wanted to see if, based upon the variables within the dataset, predictions could be made as to whether someone would possess a diagnosis of ADHD or not. To assess this, I utilized hyperparameter tuning with a grid search and implemented a few classification-model types within that. These included random forest, logistic regression, and decision tree models. Employing this method allowed me to obtain the “best bang for my buck” by running the models all at once, and then assessing for the best model with the best parameters for my data.

Following that, I reported the model metrics, and also reported the top five features impacting the target variable.

Analysis:

The results of my analysis are below and describe the results from my EDA and my ML classification model outcomes.

Descriptive Statistics:

From the descriptive data output, I observed the general mean, median, and standard deviation of the different numeric variables within the dataset. Each of the numeric values within these variables represents a category rather than a number (variables are not truly quantitative). Rather, for example, if the mean is 1.2 for ADHD_diagnosis, that means that the average leans toward 1, which represents the category "Currently has the ADHD condition".

Overall, because the variable values were encoded, the descriptive statistics mainly indicated that the majority of the dataset included people who did not have ADHD as a condition.

Distributions and Other Visualizations:

A variety of distributions were assessed as part of this project. Please refer to Appendix C at the end of this document to access those visuals, and their respective sub-topic sections.

The visuals include histograms and bar charts of each of the variable sub-areas (demographics, mental health, TS prevalence, etc.).

Overall, these visuals indicate that the majority of the dataset included children not diagnosed with ADHD, and the distribution overall is a population that is primarily white, older than 12 years of age, and pretty evenly split between males and females. More females than males do not have ADHD (as expected).

Additionally, those currently possessing an ADHD diagnosis also had a higher proportion of TS prevalence, and the same can be said for anxiety and depression proportions relative to a current ADHD diagnosis. Those with ADHD don't seem to be bullied much overall.

With respect to other areas of health, most children with ADHD get 1-3 days of at least 60 minutes of exercise, and their health is generally excellent or very good. Behaviorally, those with a current ADHD diagnosis have a higher behavior problem prevalence.

Assessing Relationships (Correlation Heatmap):

In this specific project instance, the variables were already encoded by the NSCH, so they were ready to be used to assess relationships.

Findings indicated that all patients possessed insurance (hence why this variable is grayed out within the correlation heatmap). Strong relationships existed between ADHD severity and medication (-0.93), ADHD diagnosis and medication (-0.94), ADHD diagnosis and severity (0.94), ADHD medication and behavioral treatment (0.91), ADHD severity and ADHD diagnosis, (0.94).

Overall, the strongest relationships occurred between the ADHD variables, some of those being negative correlations and some being positive. As medication and behavioral treatment went up, the ADHD severity and diagnosis went down (in remission).

Predicting ADHD Outcomes/Diagnoses:

For the second analysis portion of my project, I wanted to assess whether a machine learning model could be implemented to predict ADHD diagnosis outcomes based upon the other feature variables.

To complete this project step, I first created dummy variables to account for any multi-collinearity and then selected a Decision Tree machine learning model, as this kind of model tends to be a strong fit for classification data. Please refer to Appendix D for my machine learning confusion matrix and classification report.

While that model performed well with its accuracy (97%), I wanted to improve upon that even further. For the next step, I implemented some hyperparameter tuning, a grid search, and three

different classification model types: Decision Tree, Random Forest, and Logistic Regression. Completing this step would allow me to ascertain if a Decision Tree Model was truly the best fit, or if one of the other two models would be a best fit.

Upon running that model, the results indicated that the best model parameters were in fact those contained within the Decision Tree model, so those results were secondary confirmation of my initial results. This time, the model's accuracy performed at 99% and its recall at 100%. To review the outcome of this process, please also reference Appendix D.

Conclusion:

From the EDA, it is apparent that the large majority of those within the dataset did not have an ADHD diagnosis, but they did have access to necessary health insurance. Additionally, it was more common to see a TS diagnosis in those who also had ADHD, as well as anxiety, depression, and behavioral issues. On a more positive note, children with ADHD did not necessarily experience bullying at greater rates than non-bullying.

From the ML model, we can ascertain that a Decision Tree model was the strongest fit for predicting and ADHD diagnosis based upon the other features present within the dataset. From this, we can conclude that certain "markers" may be predictive of an ADHD diagnosis (especially the five called out).

Assumptions:

From the onset of this project, I had assumptions about what I would find within this data. These assumptions primarily mirror my hypotheses. I thought more males would have an ADHD diagnosis than females, that those with ADHD would have a higher prevalence of being diagnosed with TS as well, and that ADHD diagnosis (as a dependent variable) would be dependent upon certain demographic or health factors (independent variables).

From an outsider's perspective, this project assumes the reader's familiarity with data science techniques, data analysis, and machine learning models.

Limitations:

With respect to the limitations of this project, one item to consider is that as a single individual, I am limited in the amount of work I can complete on this topic, and the scope of which that reaches. I am also limited by the data available for me to analyze, and even then, a smaller subset of that, since I was not able to analyze all data available on this topic for my project and within the project timeline.

Challenges:

No data science project is without its own challenges. I needed to locate a strong, reliable data source that had enough information for me to complete my analyses on, without being overly complicated. Additionally, I needed to know a bit more of the background on ADHD in order to appropriately select the variables from the dataset for use, and that was challenging in its own

right. Lastly, spending enough time on EDA to thoroughly assess relationships as well as selecting the best ML models to run was tough, especially when using primarily categorical variables that were merely encoded numerically (so, still truly categorical).

Future Uses/Additional Applications:

Future use of this dataset, as well as the subsequent findings, may involve additional research in this topic area. Additionally, from an applicability standpoint, the findings I have come across may grant additional support to and for the case regarding ADHD diagnoses, the variables that may impact that, and any other comorbidities such as TS. This may be important for researchers, policy makers, medical leaders, doctors, social scientists, etc.

Recommendations:

In addition to the analyses completed on this dataset, I encourage researchers and data professionals to explore other datasets regarding the same topics. Additionally, outside of the ML models implemented within my project, another potential next step would be to implement SMOTE.

Because the classes are so imbalanced within the dataset, SMOTE methodology can potentially help to create an even stronger ML model.

Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in a dataset in a balanced way. The component works by generating new instances from existing minority cases that are supplied as input.

Another recommendation is to take the five best features selected at the end of the model work and feed those into a new model – using just those features, none of the others from the dataset that were used in the original model. This additional step could also refine the model for improvement.

Implementation Plan:

Following the culmination of this project, my proposed implementation plan would be to conduct additional research, with the inclusion of some SMOTE techniques and model running with even more pared down feature-selection (as discussed above). I think beyond that, the findings showcase that certain factors do play a key role in predicting whether someone has ADHD, and beyond that, a multitude of combined factors do as well. Having this knowledge, providers and caretakers can watch for certain aspects to help make an early diagnosis, if need be, and to be sure that ADHD is truly what is present, versus merely assuming due to gender-related reasons.

Ethical Assessment:

For my specific project, there are a few ethical considerations regarding this data/my project. One of these was ensuring that I had enough data to be representative of the population and to minimize any potentially biased or skewed data. I think this dataset does have some imbalanced class issues, so handling those would probably be important moving forward.

Additionally, I needed to be conscientious of the ways in which the data was collected, as this has the potential to impact outcomes and the meaning of data insights post-analysis. This data is survey data, which means that I am relying of the self-reporting of the individuals and/or their parents for the data within this dataset.

Lastly, it is imperative to consider who the data impacts, and how any findings surrounding the data analysis or insights are applied.

Appendix:

Appendix A:

References:

Bitter, I., Simon, V., Balint, S., Meszaros, A., and Czobar, P. (2010, June). How do different diagnostic criteria, age and gender affect the prevalence of attention deficit hyperactivity disorder in adults? An epidemiological study in a Hungarian community sample. *Eur Arch Psychiatry Clin Neurosci, 260*(4):287-96. doi: 10.1007/s00406-009-0076-3. https://pubmed.ncbi.nlm.nih.gov/19806424/

Centers for Disease Control and Prevention (CDC). (2022). *Data and statistics about ADHD. CDC.* https://www.cdc.gov/ncbddd/adhd/data.html

Centers for Disease Control and Prevention (CDC). (2022). *What is ADHD?* CDC. https://www.cdc.gov/ncbddd/adhd/facts.html

Child and Adolescent Health Measurement Initiative (CAHMI). (2022). *2019-2020 National Survey of Children's Health (2 years combined dataset): SPSS dataset*. Data Resource Center for Child and Adolescent Health supported by Cooperative Agreement U59MC27866 from the U.S. Department of Health and Human Services, Health Resources and Services Administration (HRSA), Maternal and Child Health Bureau (MCHB). childhealthdata.org

Child and Adolescent Health Measurement Initiative (CAHMI) (2022). *2019-2020 National Survey of Children's Health (2 years combined dataset).* SPSS codebook for data users: Child and Family Health Measures, National Performance and Outcome Measures, and Subgroups, Version 1.0. Data Resource Center for Child and Adolescent Health supported by Cooperative Agreement U59MC27866 from the U.S. Department of Health and Human Services, Health Resources and Services Administration (HRSA), Maternal and Child Health Bureau (MCHB). www.childhealthdata.org

Kinman, T. (2016, March 22). Gender differences in ADHD symptoms. Healthline.

https://www.healthline.com/health/adhd/adhd-symptoms-in-girls-and-boys#ADHD-and-Gender-

Ramtekkar, U. P., Reiersen, A. M., Todorov, A. A., Todd, R. D. (2010, March). Sex and age differences in attention-deficit/hyperactivity disorder symptoms and diagnoses: Implications for DSM-V and ICD-11. *J Am Acad Child Adolesc Psychiatry, 49*(3): 217–28.e1-3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101894/

Tourette Association of America. (2022). *What is Tourette?* Tourette Syndrome, an Overview. https://tourette.org/about-tourette/overview/what-is-tourette/

Appendix B:

Dataset Information:

Please see below for more detailed dataset information.

The specific values for each categorical variable used in the dataset, alongside their meanings and descriptions, are below. This is a data dictionary for the potential values that make up the variables.

* HHID - Household ID for the family of the child.
    * Each of these is a unique ID.
* YEAR - The survey year; either 2019 or 2020.
    * 2019 = 2019
    * 2020 = 2020

* age5_1920 - The age of the child; this is split into 5 age category groupings.
    * 0-3 years = 1
    * 4-7 years = 2
    * 8-11 years = 3
    * 12-14 years = 4
    * 15-17 years = 5

* sex_1920 - This is the sex of the child; this is split into two categories.
    * Male = 1
    * Female = 2

* race7_1920 - This is the sex of the child; this is split into seven categories (nH means non-Hispanic).
    * Hispanic = 1
    * White nH = 2
    * Black nH = 3
    * Asian nH = 4
    * American Indian or Alaska Native nH = 5
    * Native Hawaiian and Other Pacific Islander nH = 6
    * Multi-Race nH = 7
    * Other nH = 8

* ADHD_1920 - This is the ADHD diagnosis of the child (ages 3-17). There are three value options for this category.
    * Does not have condition = 1

    * Ever told, but does not currently have condition = 2
    * Currently has condition = 3

* ADHDMed_1920 - This is whether the child is taking ADHD medication (ages 3-17). There are three value options for this category.

    * Currently has condition and taking medication = 1
    * Currently has condition but not taking medication = 2
    * Does not currently have condition = 3

* ADHDBehTreat_1920 - This is whether the child is receiving behavioral treatment for ADHD (ages 3-17). There are three value options for this category.

    * Currently has condition and received behavioral treatment = 1
    * Currently has condition but did not receive behavioral treatment = 2
    * Does not currently have condition = 3

* ADHDSev_1920 - This is the severity of the child's ADHD condition (parent-rated, ages 3-17). Three possible values exist.

    * Does not currently have condition = 1
    * Current condition, rated mild = 2
    * Current condition, rated moderate/severe = 3

* tourette_1920 - This denotes children that have TS (ages 3-17). Three possible values exist.

    * Does not have condition = 1
    * Ever told, but does not currently have condition = 2
    * Currently has condition = 3

* TouretSev_1920 - This is the severity of the child's TS condition (parent-rated, ages 3-17). Three possible values exist.

    * Does not currently have condition = 1
    * Current condition, rated mild = 2
    * Current condition, rated moderate/severe = 3

* bullied_1920 - This denotes children who are bullied, picked on, or excluded by other children during the past 12 months, age 6-17 years. Five possible values exist.

    * Never = 1
    * 1-2 times = 2
    * 1-2 times per month = 3
    * 1-2 times per week = 4
    * Almost every day = 5

* anxiety_1920 - This denotes children that have anxiety (ages 3-17). Three possible values exist.

    * Does not have condition = 1
    * Ever told, but does not currently have condition = 2
    * Currently has condition = 3

* depress_1920 - This denotes children that have depression (ages 3-17). Three possible values exist.

    * Does not have condition = 1
    * Ever told, but does not currently have condition = 2
    * Currently has condition = 3

* behavior_1920 - This denotes children that have behavioral or conduct problems (ages 3-17). Three possible values exist.

    * Does not have condition = 1
    * Ever told, but does not currently have condition = 2
    * Currently has condition = 3

* homework_1920 - This denotes children who do all required homework, from ages 6-17. Four possible values exist.

    * Always = 1
    * Usually = 2
    * Sometimes = 3
    * Never = 4

* argue_1920 - This is the amount a child argues, based on parental input (ages 6-17). Four possible values exist.

    * Always = 1
    * Usually = 2
    * Sometimes = 3
    * Never = 4

* finishes_1920 - This denotes children who finish tasks that they start, from ages 6-17. Four possible values exist.

    * Always = 1
    * Usually = 2
    * Sometimes = 3
    * Never = 4

* ChHlthSt_1920 - This is the child's overall general health status. Three possible values exist.

    * Excellent or very good = 1
    * Good = 2
    * Fair or poor = 3

* CurrIns_1920 - This is the current health insurance status at time of survey. Two possible values exist.

    * Currently insured = 1
    * Currently uninsured = 2

* benefits_1920 - This is the current insurance benefits meeting the child's needs. Three values are possible.

    * Always = 1
    * Usually = 2
    * Sometimes or never = 3

* PhysAct_1920 - This is children who are physically active at least 60 minutes per day, ages 6-17. Four possible values exist.

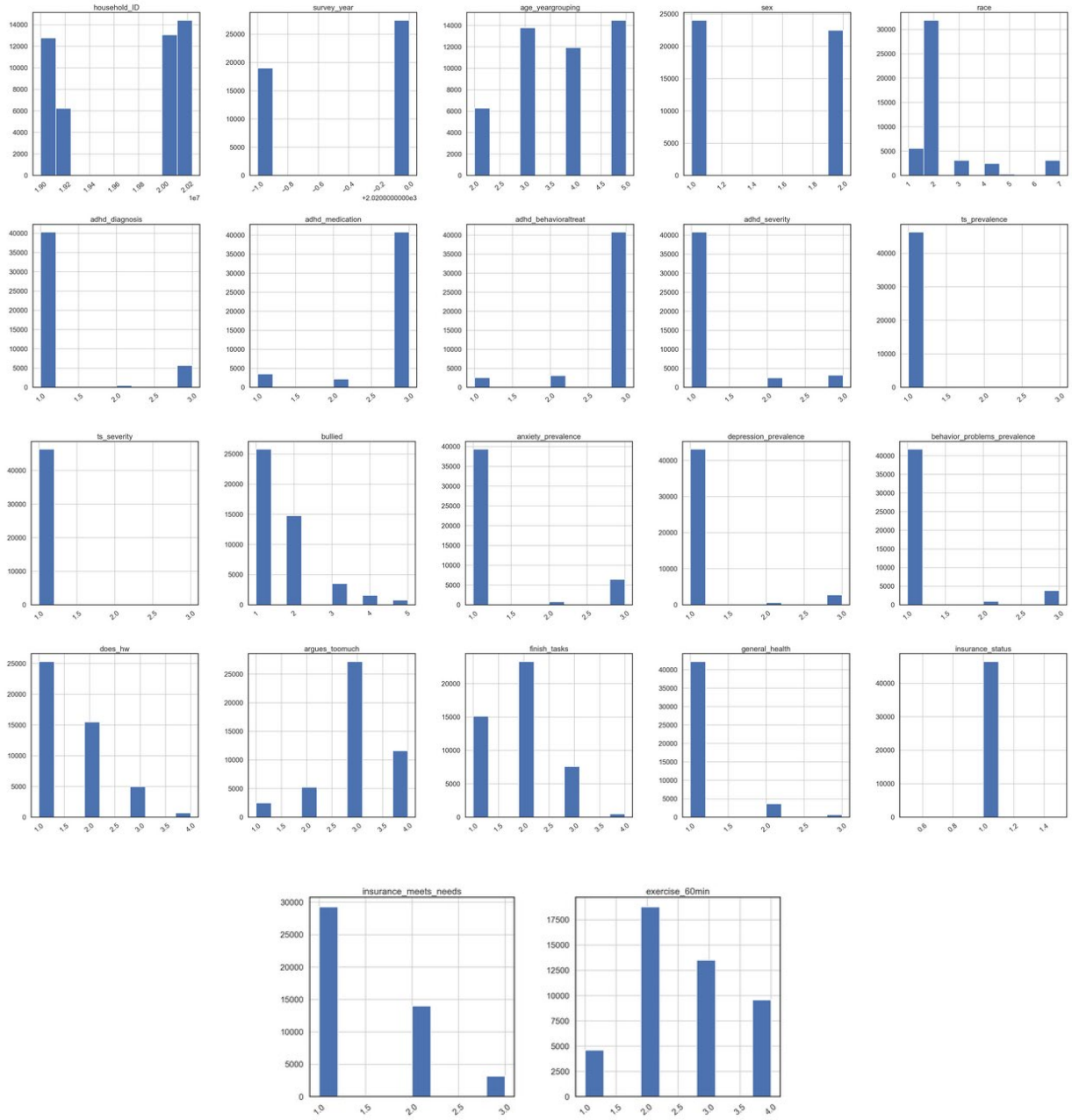
    * 0 days = 1
    * 1-3 days = 2
    * 4-6 days = 3
    * Every day = 4

Appendix C:

Data Analysis – Accompanying Visualizations:

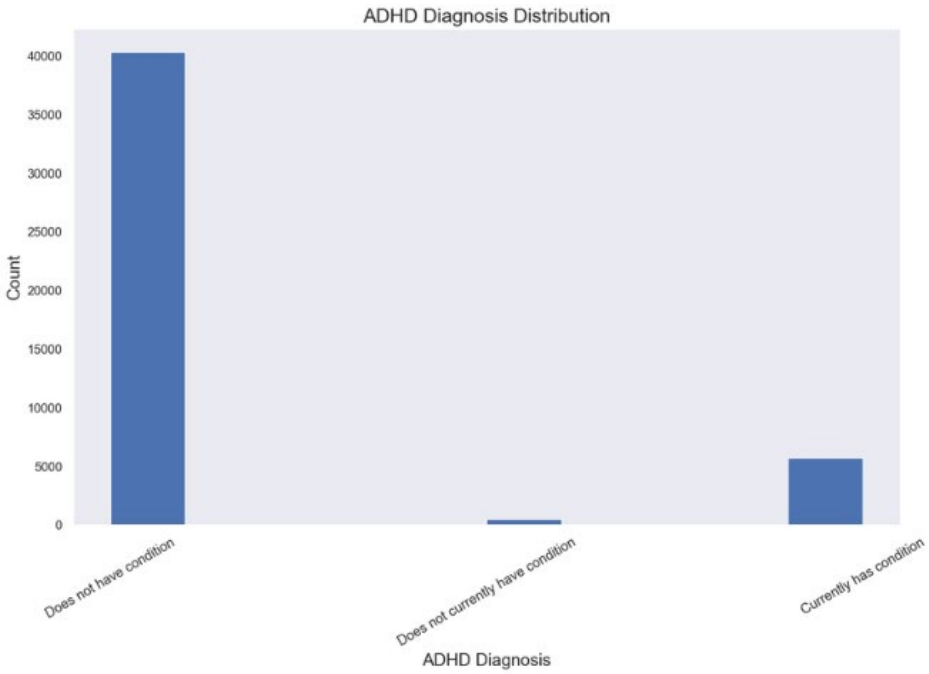
Please see all visualizations below, as well as accompanying descriptions.

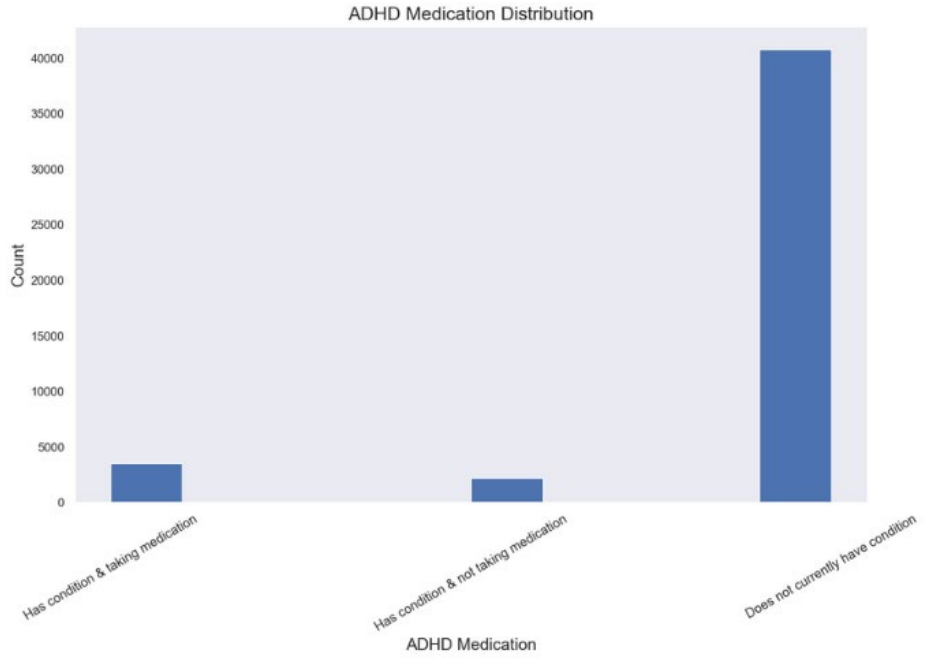
1. Distributions (Histograms) and Bar Charts:

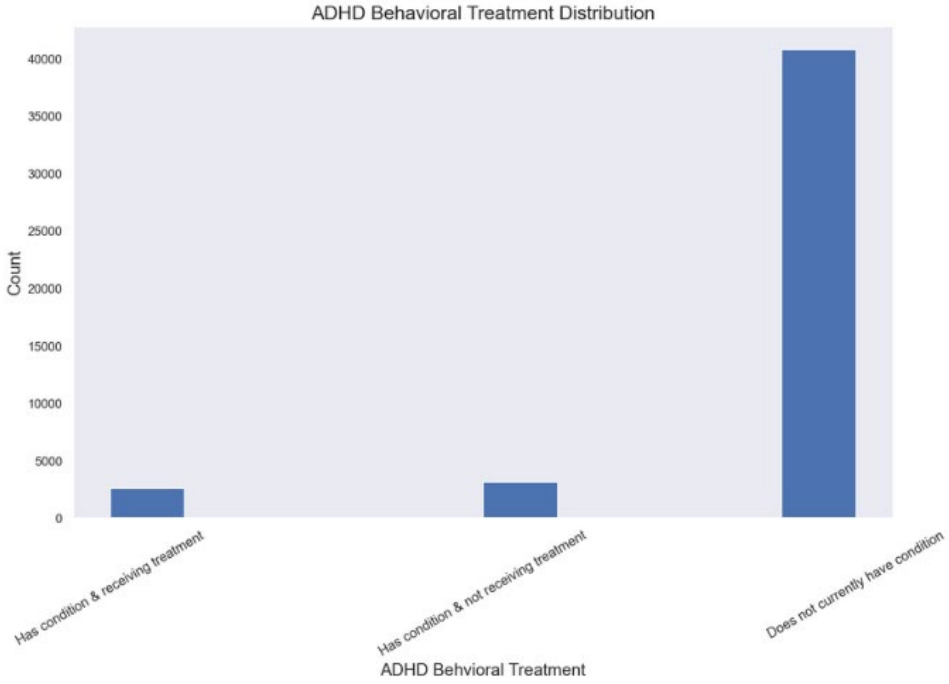


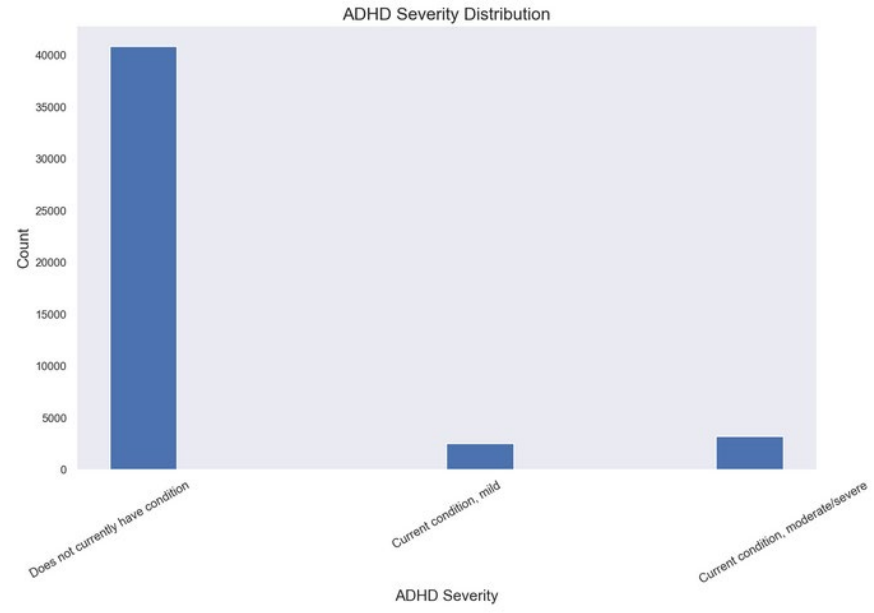
ABOVE: A first overall look at the variable distributions.

ADHD Variable Distributions:

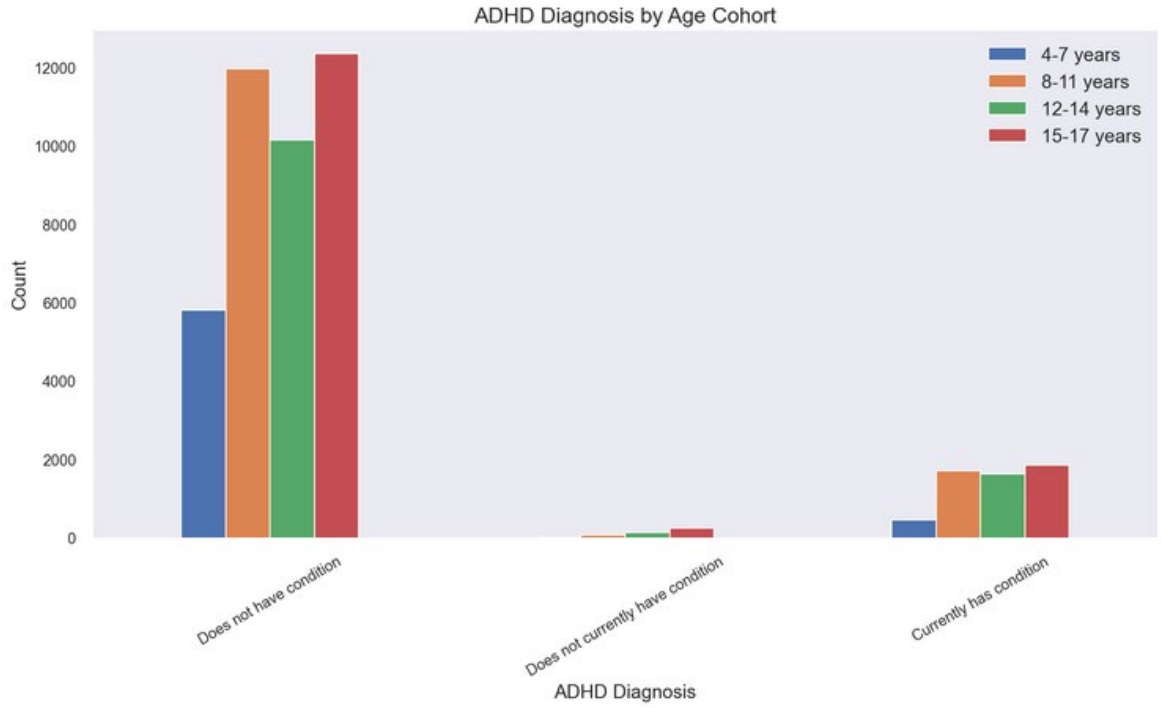


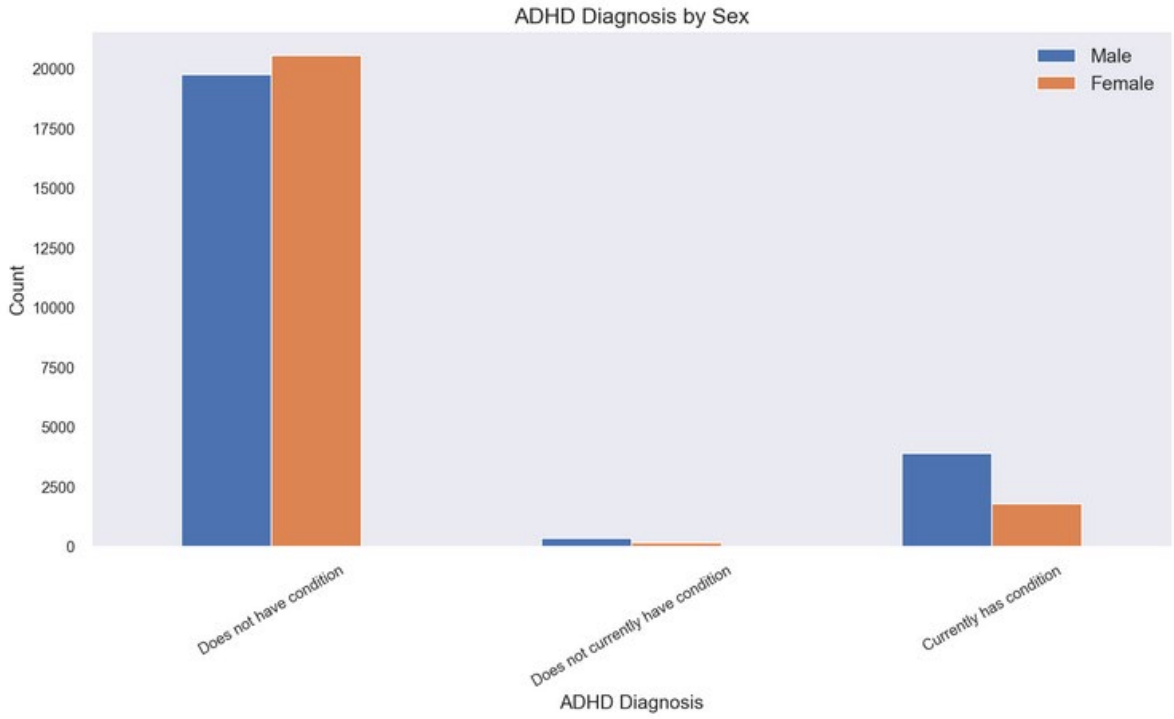


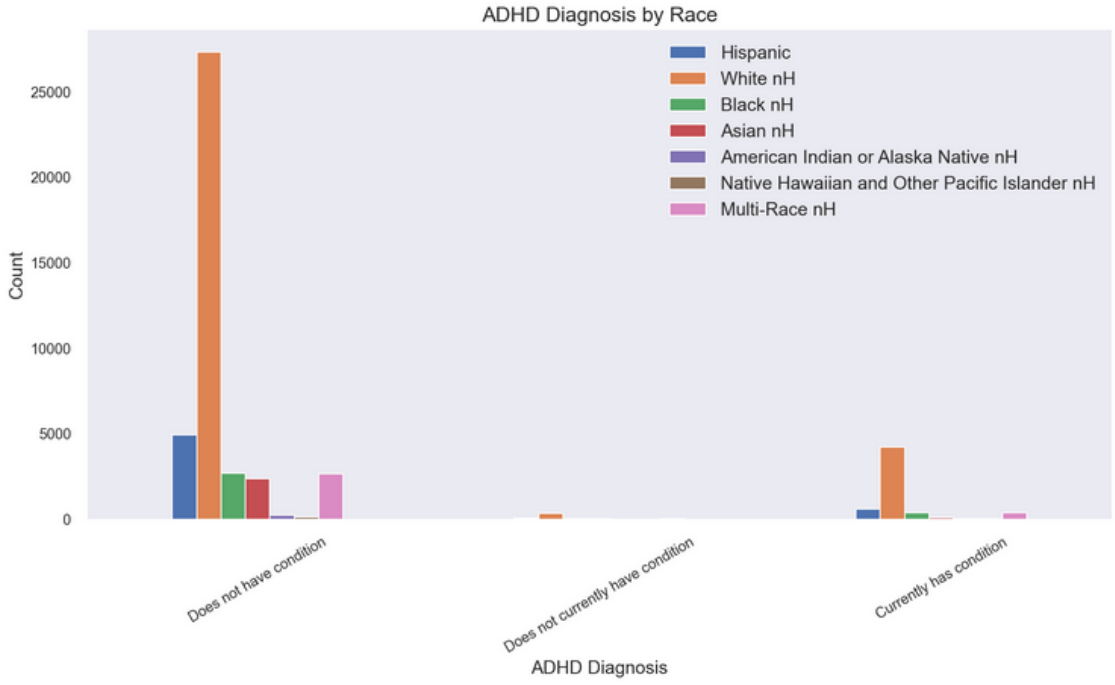

ADHD Behavioral Treatment Distribution
Count
40000
35000
30000
25000
20000
15000
10000
5000
0
Has condition & receiving treatment
Has condition & not receiving treatment
Does not currently have condition
ADHD Behvioral Treatment


ADHD Severity Distribution
Count
40000
35000
30000
25000
20000
15000
10000
5000
0
Does not currently have condition
Current condition, mild
Current condition, moderate/severe
ADHD Severity

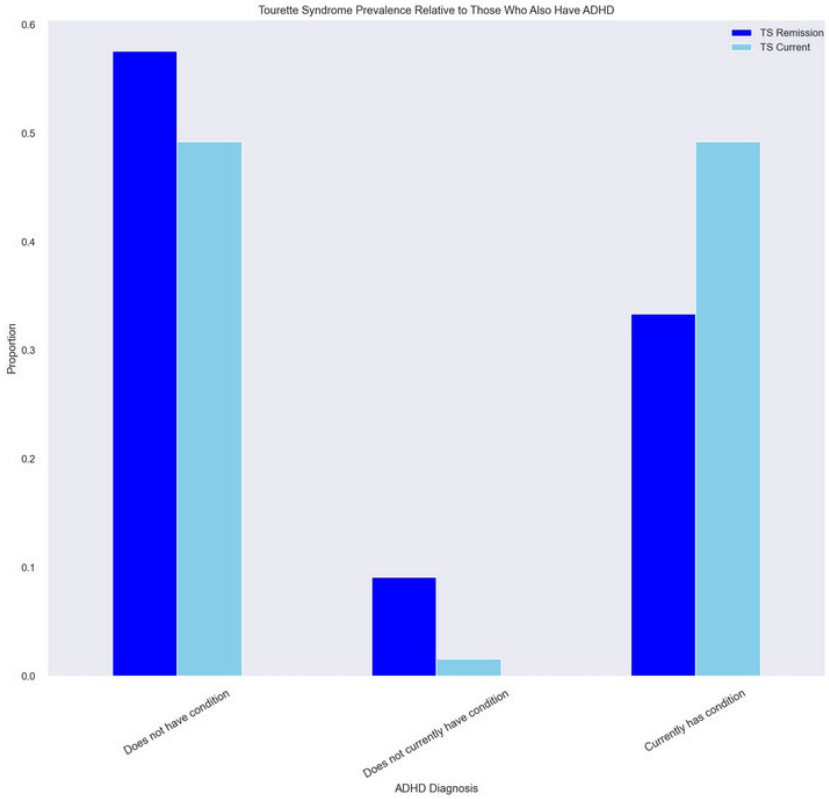
Demographic Variable Distributions:

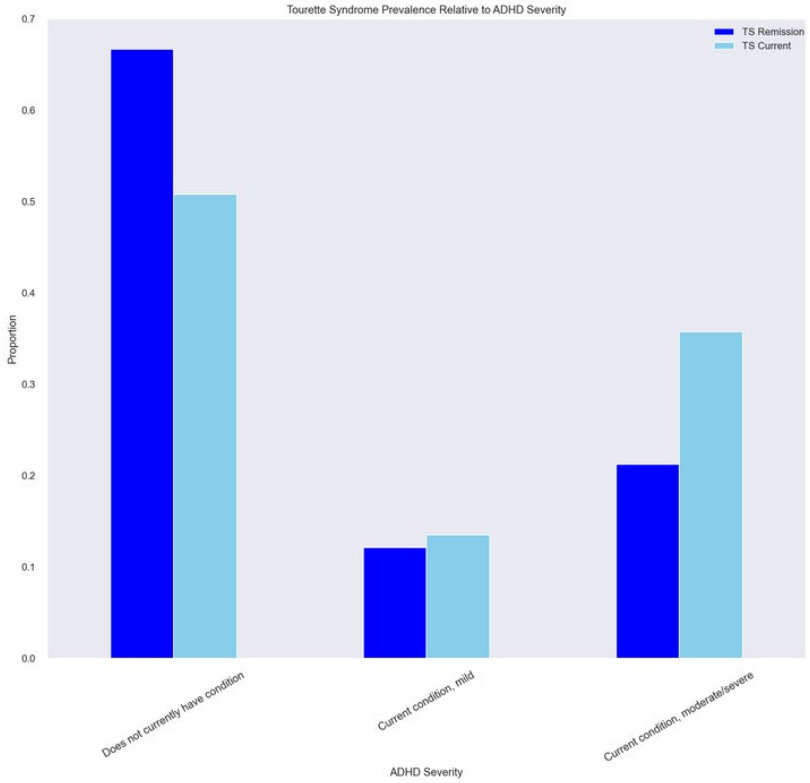




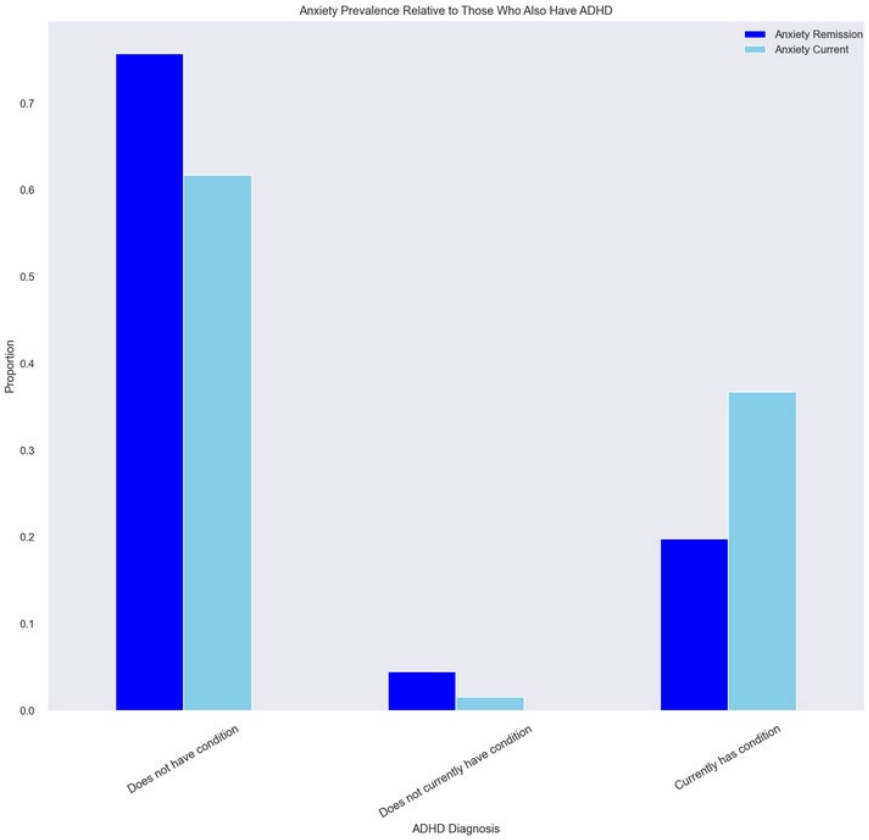

ADHD Diagnosis by Race
Hispanic
White nH
Black nH
Asian nH
American Indian or Alaska Native nH
Native Hawaiian and Other Pacific Islander nH
Multi-Race nH
Count
25000
20000
15000
10000
5000
0
Does not have condition
Does not currently have condition
Currently has condition
ADHD Diagnosis

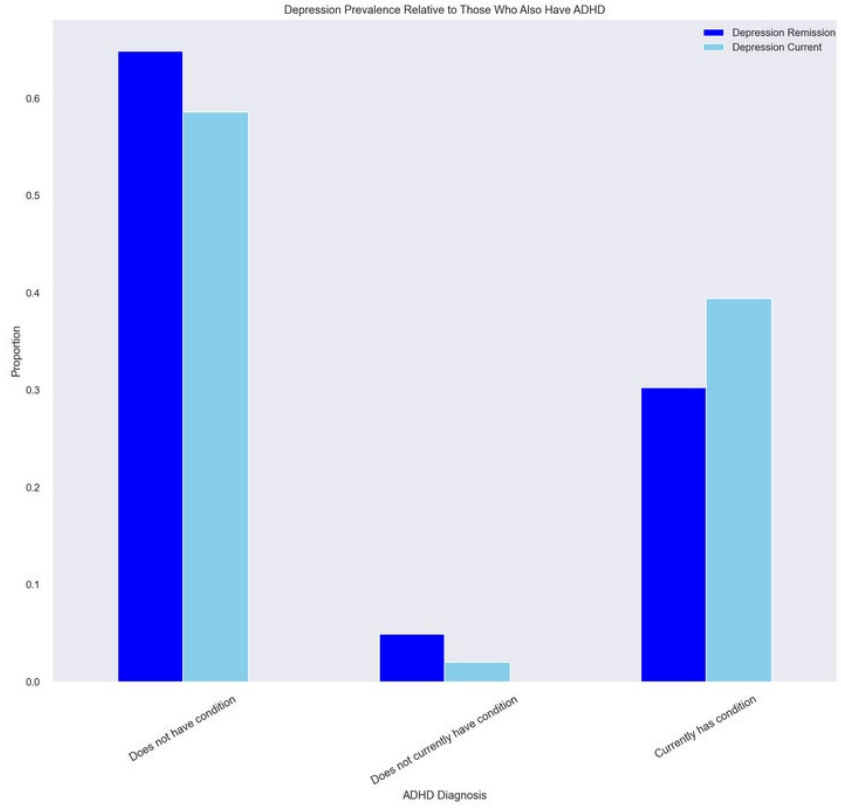
Tourette Syndrome Variable Distributions:

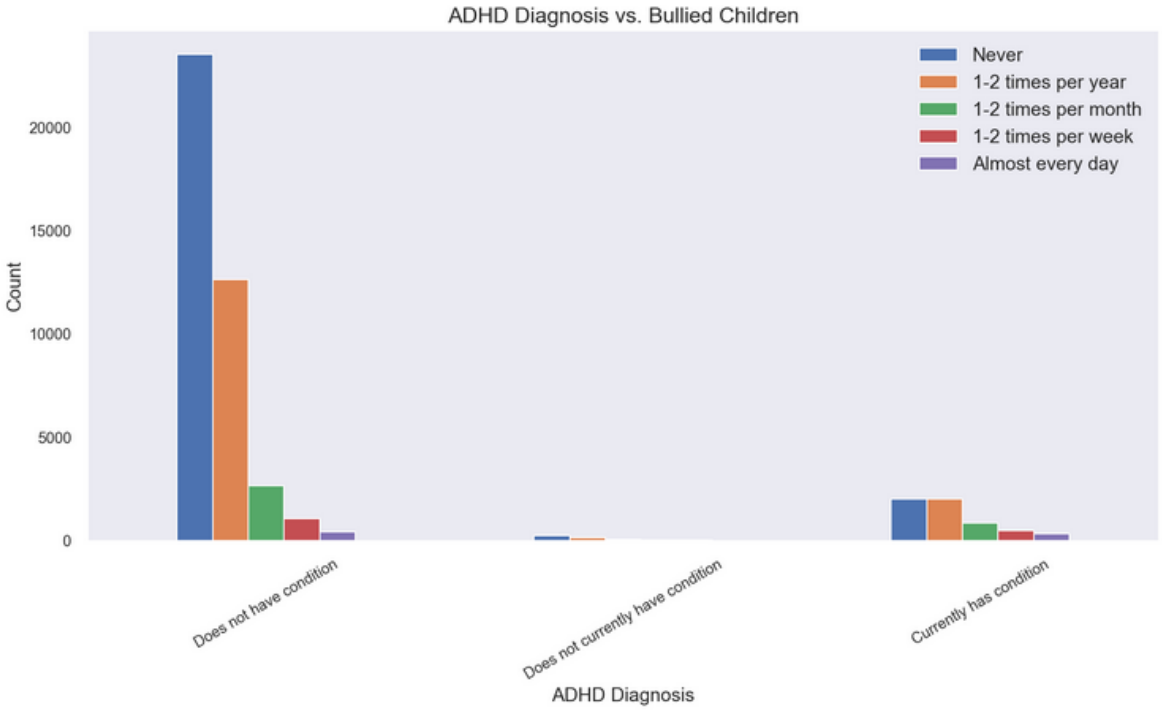




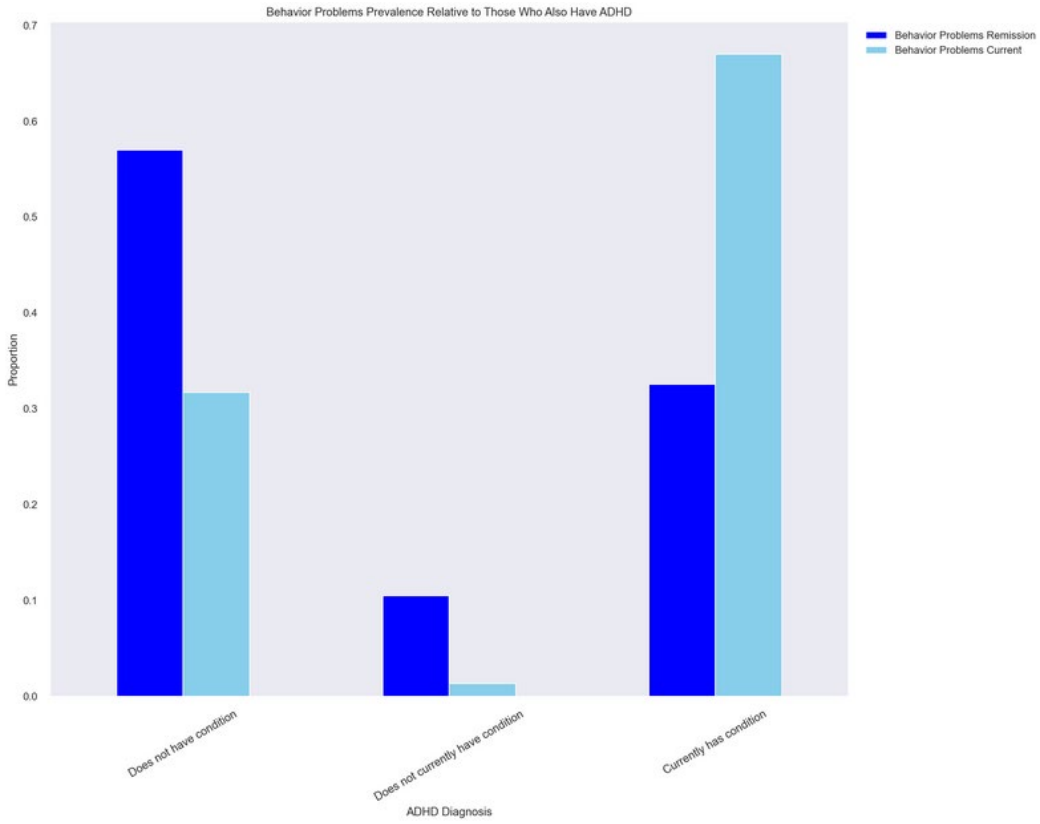
Mental Health Variable Distributions:



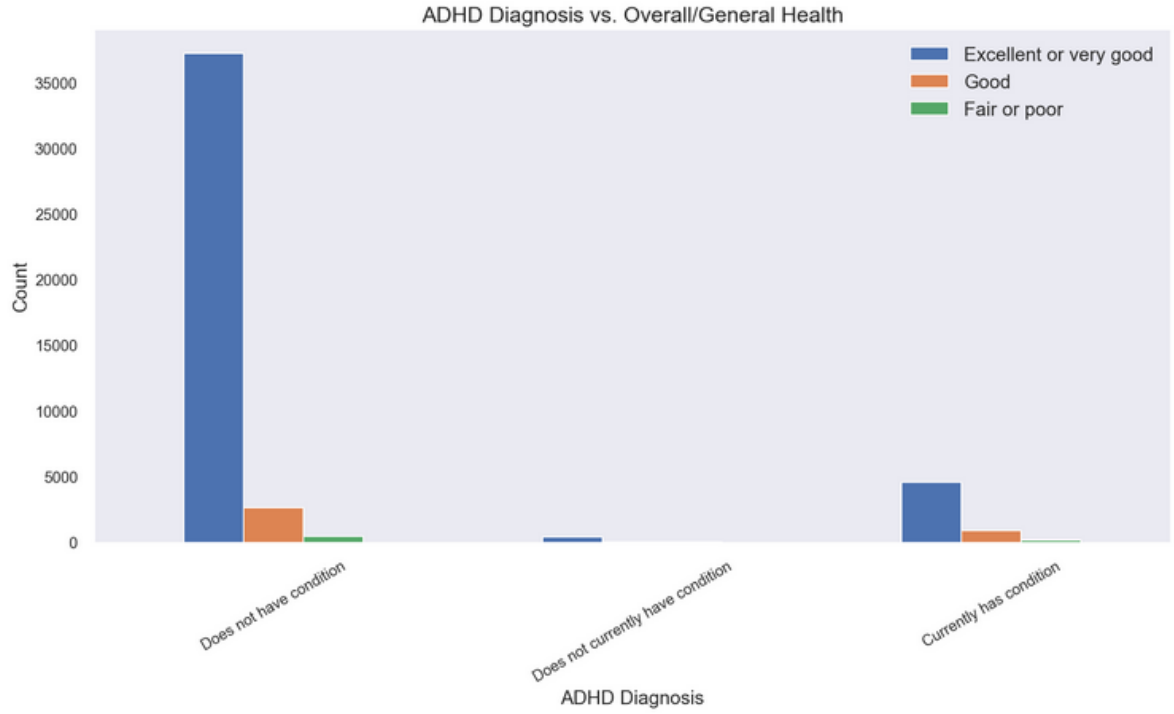




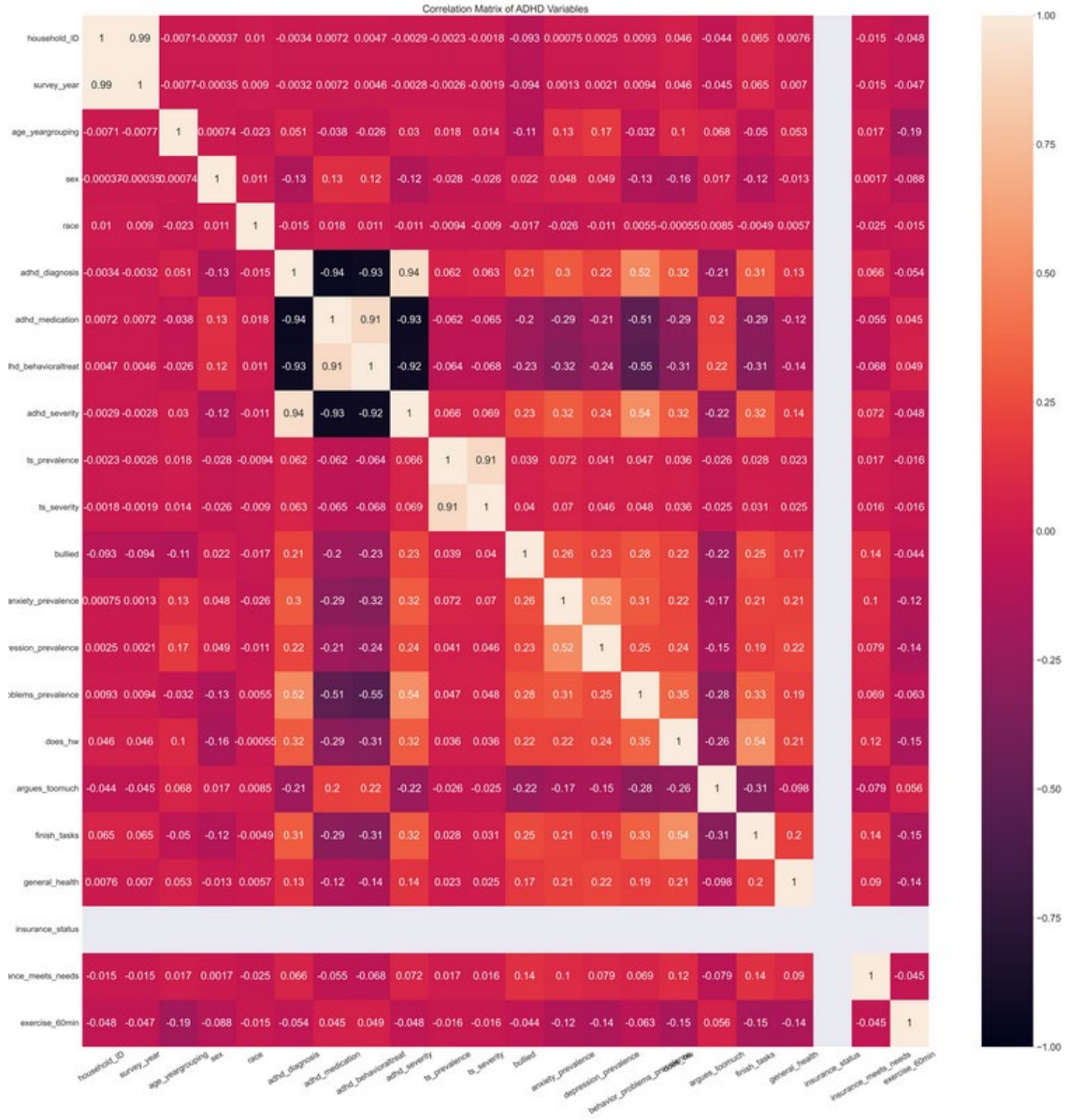
Behavioral Health Variable Distributions:



Overall General Health Variable Distributions:



2. *Correlation Heatmap (Relationships):*



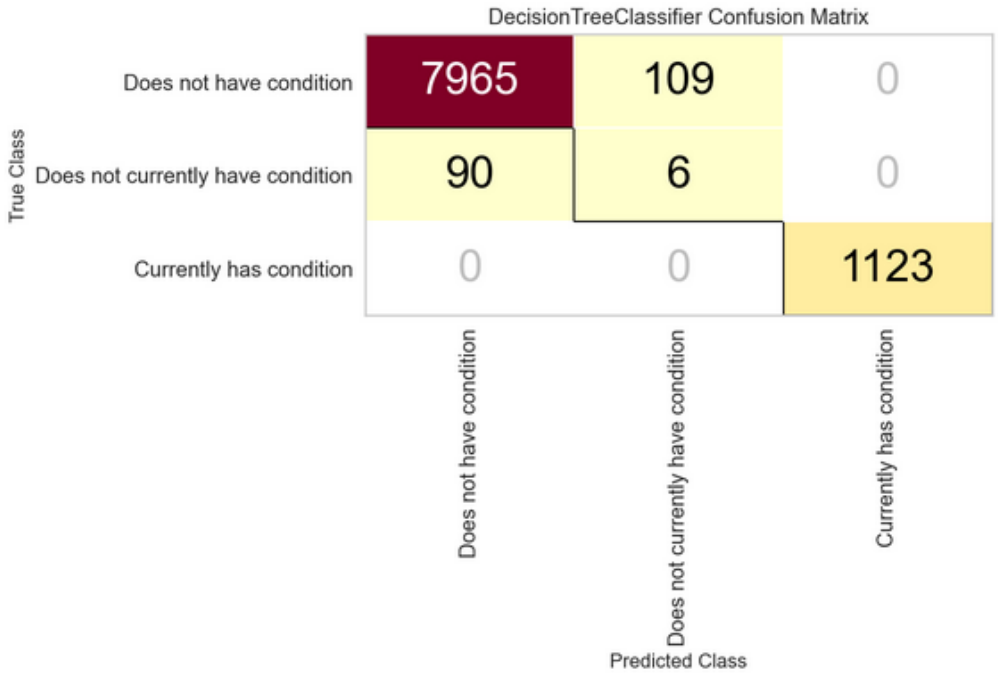
ABOVE: A correlation heatmap of all variables.

Appendix D:

Machine Learning Outputs:

BELOW: Initial Decision Tree Confusion Matrix and Classification Report.

```
Model Accuracy: 97.86%
```



	precision	recall	f1-score	support
1.0	0.99	0.99	0.99	8074
2.0	0.05	0.06	0.06	96
3.0	1.00	1.00	1.00	1123
accuracy			0.98	9293
macro avg	0.68	0.68	0.68	9293
weighted avg	0.98	0.98	0.98	9293

BELOW: Additional model work outputs (hyperparameter tuning and grid search).

```
Wall time: 13min 25s

GridSearchCV(cv=5,
             estimator=Pipeline(steps=[('classifier',
                                        RandomForestClassifier())]),
             n_jobs=-1,
             param_grid=[{'classifier': [RandomForestClassifier()],
                          'classifier__max_features': [1, 2, 3],
                          'classifier__n_estimators': [10, 100, 1000]},
                         {'classifier': [LogisticRegression(class_weight='balanced',
                                                            max_iter=1000,
                                                            solver='saga')],
                          'classifier__C': array([1.00000000e+0...
       5.99484250e+01, 1.66810054e+02, 4.64158883e+02, 1.29154967e+03,
       3.59381366e+03, 1.00000000e+04]),
                          'classifier__penalty': ['l1', 'l2']},
                         {'classifier': [DecisionTreeClassifier(max_depth=5,
                                                                min_samples_split=10,
                                                                random_state=42)],
                          'classifier__class_weight': [None, {0: 1, 1: 5},
                                                       {0: 1, 1: 10},
                                                       {0: 1, 1: 25}],
                          'classifier__max_depth': [5, 10, 25],
                          'classifier__min_samples_split': [2, 5, 10]}])
```

BELOW: Additional model work outputs (best model parameters and Classification Report).

```
{'classifier': DecisionTreeClassifier(max_depth=5, min_samples_split=10, random_state=42),
 'classifier__class_weight': None,
 'classifier__max_depth': 5,
 'classifier__min_samples_split': 10}
```

	precision	recall	f1-score	support
1.0	0.99	1.00	0.99	8074
2.0	0.00	0.00	0.00	96
3.0	1.00	1.00	1.00	1123
accuracy			0.99	9293
macro avg	0.66	0.67	0.66	9293
weighted avg	0.98	0.99	0.98	9293

BELOW: Additional model work outputs (top five independent features impacting target variable of ADHD diagnosis).

	household_ID	adhd_medication	adhd_behavioraltreat	adhd_severity	behavior_problems_prevalence
0	19000167	3	3	1	1
3	19001345	3	3	1	1
7	19001651	3	3	1	1
8	19001670	3	3	1	1
10	19001838	3	3	1	1
...	...	...	...	...	...
72198	20181823	3	3	1	1
72204	20207385	3	3	1	3
72206	20214272	3	3	1	1
72207	20235143	3	3	1	1
72208	20235559	2	2	2	1

46464 rows × 5 columns

Appendix E:

Audience Questions:

1. Did any relationships exist between any of the demographic variables and ADHD? How strong or significant were each of the variables as far as predicting ADHD diagnosis?
    a. From my analyses, there were certainly correlations between the various ADHD variables (diagnosis, medication, severity), but outside of that, not many strong relationships existed according to the correlation heatmap.
    b. As for multi-collinearity, this was handled via the use of dummy variables prior to any model work, so should not have impacted any relationships there during the feeding of the data to the ML model.
    c. I think the greatest impact to any outcomes was the imbalanced variables classes and any potential skew from that. In the future, using a SMOTE methodology with ML techniques might help here.

2. Did a relationship exist between TS and ADHD diagnosis?
    a. Yes, those children with a current ADHD diagnosis were more likely to have a TS current TS diagnosis.

3. How did you determine which variables were most important prior to your analysis? Do you feel there were any you missed that might have been important to include? How do those variables compare to the even more pared-down variables as a result of feature selection?
    a. My determination for the most important variables primarily came from the following:
        i. My literature review of what was most important to explore.
        ii. The variables available to me in the data.
        iii. Those variables that I ascertained might impact an ADHD diagnosis.

4. What was your initial hypothesis, and do you feel the results and outcomes of your project work supports this? Why or why not?
    a. My initial hypothesis consisted of three parts. Overall, I hypothesized that more males would have an ADHD diagnosis than females, that those with ADHD would have a higher prevalence of being diagnosed with TS as well, and that ADHD diagnosis (as a dependent variable) would be dependent upon certain demographic or health factors (independent variables).

5. What do you recommend for those who want to help in this area? What other ways can those who are interested use data to inform others and instill change?
    a. Firstly, I recommend more research, and continued research.
    b. I also recommend additional, robust ML methods.
    c. I think continuing research and education can help pinpoint factors that influence ADHD prior to a diagnosis.

6. Talk a bit more about your ML model methods – how did you ensure that the data was transformed appropriately and then properly fed into those models?
    a. To ensure appropriate data handling, I needed to, of course, properly clean and transform the data, but then I also needed to use dummy variable encoding. Following that, I needed to spend some time figuring out what the best model would be, and then feed that cleaned data into the model. The main outstanding issue was the imbalanced classes, but that did not appear to be an issue too much for the model.

7. Why do you think this topic is important to look into further than the research that already exists out there?
    a. I think more data from a variety of sources would be useful, particularly data that is not just survey data.

8. Do you believe that with additional data and additional research in this topic area will help to pinpoint methodologies for better diagnoses for those with ADHD or TS? Why or why not?
    a. I do – research and analyses always bring to light new insights, and those insights often hold the keys to progression and to ensuring better implementation of findings as far as real-world applicability.

9. Is there any other data that you wish was available that you couldn't find, or that was not in a usable format for you?
    a. Again, I do think having non-survey data would be beneficial, so finding some of that would prove useful and helpful for more robust analyses.

10. What challenges and limitations did you personally face during this project? In addition to that, what do you think are the overall challenges and limitations of your project work itself, and what do you think could be done next to improve upon that?
    a. Overall, I needed to locate a strong, reliable data source that had enough information for me to complete my analyses on, without being overly complicated. Additionally, I needed to know a bit more of the background on ADHD in order to appropriately select the variables from the dataset for use, and that was challenging in its own right. Lastly, spending enough time on EDA to thoroughly assess relationships as well as selecting the best ML models to run was tough, especially when using primarily categorical variables that were merely encoded numerically (so, still truly categorical).

11. Why do you believe the results are what they are? What are resources or references either support or do not support your findings? Why do you think your results are what they are?

a. I believe the results are what they are given that ADHD is more often diagnosed in males, and it also tends to run alongside other neurological or behavioral conditions (per the literature). I also think this is why the ML model was able to predict ADHD diagnosis outcomes so strongly – many of those features used in the ML modeling were likely helpful as predictor variables, and I am certain this is probably true in the real-world as well.