A Linear Regression, Sentiment, and Natural Language Processing Analysis of Dental Attendance Disparity Factors

Aleeza Mughal, Plainedge High School, Massapequa, NY

E-mail: aleezamughal2@gmail.com

Abstract

A lack of dental visits has been found to predict both oral diseases (cavities and gingivitis) and complex diseases including Diabetes, Alzheimer's Disease, Periodontitis, and cardiovascular disease. I used linear regressions, sentiment analysis, and natural language processing to identify factors associated with decreased participation in dental visits among Americans. Additionally, I created two different word clouds that presented a visual representation of public Tweets showcasing how individuals feel about going to the dentist. The common themes in the word clouds included affordability and appointments. Furthermore, I conducted multiple linear regressions to determine the effect of various independent variables on the likelihood of individuals delaying personal and family care and the out-of-pocket expenditures for self and family. The most influential factors that contributed to the results of these multiple linear regressions included possessing insurance, age, gender, marital status, and earnings. Importantly, the socioeconomic factors of Americans indicate that many people avoid the dentist due to the cost of dental procedures.

Oral health and hygiene have been implicated in diseases elsewhere in the body including Alzheimer's Disease, Cardiovascular Disease, and Diabetes (Lockhart, 2012; Poudel, 2018; Preshaw, 2011). These diseases kill more than 906,146 Americans each year ("2021 Alzheimer's", 2021; "Statistics About," 2019) and although the causes and progression of these illnesses vary, a disparity in dental visits appears to be a significant predictor of illnesses since there is a significant decrease in patient attendance ("Disparities in," 2021). Given that large percentages of surveyed participants do not see a dentist, results suggest Americans may be overlooking oral health as a significant preventative measure for serious diseases (Michas, 2022). The purpose of my study was to explore which factors have the most significant effects on Americans to delay dental care by examining the variables of age, marital status, earnings, race, and political party.

Socioeconomics

Dentists generally recommend an annual dental visit to check for healthy bacteria, oral cancers, anomalies, and periodontitis (Kay, 1999). However, the socioeconomic status of families may reinforce disparities in dental attendance. A survey conducted by the Egyptian Pediatric Association Gazette identified that the dental health of children is influenced by the oral hygiene behavior of children, which is derived from the socioeconomic status of their family (Aslan Ceylan et al., 2018). Families of a higher socioeconomic status reported better oral hygiene habits including tooth brushing, decreased sugar consumption, regular dental examinations, and adequate fluoride supplementation which led to fewer caries of the teeth of patients. According to Statista, disparities in attendance existed even in 2019, with over ⅓ of the American population having not seen a dentist in the past 12 months - a statistic that decreased by an additional 3.2% in 2022. When examining ethnicity and race as factors of attendance for the same years, all ethnicities and race groups showed decreased attendance in 2020 (Elflein, 2022). Among the ethnicities with the highest level of attendance were Non-Hispanic Asians (64.3%) and Non-Hispanic Whites (66.6%) in 2020. As a factor of gender, women (65.8%) reported more visits to the dentist compared to men (59.6%) in 2020 (Elflein, 2022).

Dental visits can be expensive and when procedures are needed to be performed such as cavity fillings, wisdom teeth extractions, and braces, the cost can increase drastically. Dental exams can cost from $0 to $550 depending on insurance coverage; individuals who do not have insurance have to pay out-of-pocket for the routine exam which can be from $80 to $120 ("How much", 2021). Insurance coverage

may restrict which procedures are covered due to being case-dependent. Many dental insurance companies will cover a percentage of preventative, basic, and major restorative care procedures. Almost half of insured Americans are unhappy with their insurance policy (Black, 2021). The average American considers certain criteria when visiting the dentist such as noticing if the dental clinic is in-network with their insurance provider to avoid major expenses. A survey found that 82% of adults with insurance coverage visited the dentist at least once during 2021, but this procreation was 55% percent among adults with no insurance. This gap was slightly lower when comparing children (Elfein, 2021).

COVID-19

In addition to decreased dental care as a result of demographic variables, many offices around the world had to limit and eventually stop any non-emergency procedures in 2020. So, while pandemic appointments increased by 84% compared to the pre-pandemic period, the number of canceled visits increased from 15% to 50% (Migas, 2022). Canceling and postponing dental visits negatively impacted overall physical health while also increasing healthcare costs and decreasing patient visit quality (Ceylan, 2022; Migas, 2022).

Factors that may have existed pre-pandemic also impact dental attendance, including dentophobia which affects approximately 36% of the population. This anxiety can have serious effects on the health of individuals as it presents a barrier for dental visits (Beaton et al., 2014). Many people feel afraid to go to the dentist (dentophobia) which can negatively affect their health. Additionally, due to COVID-19 affecting the health of many Americans, dentophobia of patients and the avoidance of dental appointments may have increased from the fear of catching COVID while at the dentist's office. The exposure of one's mouth and airways to a potentially deadly infection may also affect one's comfort with dental visits (Tofangchiha, 2022).

As the world begins to recover from the COVID-19 pandemic, Americans are being encouraged to visit more healthcare facilities, such as the dentist, to address the treatment they deferred as a result of the pandemic. This paper will explore and quantify the factors that are impeding dental visits by integrating linear regression modeling, natural language processing, and word cloud visualization. I wanted to quantify and qualify the effects of my data, which was unique to my study since I analyzed my data and two different nuanced methods. Quantifying the variables through the regression modeling of survey data in terms of beta coefficients and effect sizes, allowed to understand the value of insurance and finances for patients. Qualifying the values through natural language processing demonstrates the diction analysis by analyzing word cloud frequencies and understanding themes that are present. This study is unique as it explores many factors contributing to dental attendance of various racial groups in America.

Hypotheses

H_1: Individuals who do not have insurance will be more likely to delay care for self and family than individuals with insurance.

H_2: The words associated with cost and insurance would be prominent in the dental-related word clouds.

H_3: The demographic factors will be considered in this study due to the impact they may carry on the results.

Methods

Materials for Descriptive and Regression Analysis in SPSS

Survey on Dental Care and Dental Insurance. Data were drawn from the *2016 Commonwealth Fund Health Insurance Survey* distributed from July 12, 2016 to November 20, 2016 by the Princeton Survey Research Associates International. Data took the form of phone surveys of 6005 adult participants drawn from various states across the US. The participants were surveyed by phone and asked to share information about their health care, insurance coverage, and spending (Appendix A).

Statistical Package for the Social Sciences (v. 25).

I recoded data with values for "chose not to answer" or "did not remember" to a "system missing" value and I created bar graphs to visualize the trends for each of the survey questions. I calculated the median for the out-of-pocket costs, preferring this measure to the mean because of the mean's potential inaccuracy as the result of outliers. If the person reported spending no money out-of-pocket, it was recorded as a "system missing" value to avoid skewing the data.

Materials for Natural Language Processing (NLP)

Dental Tweet Datasets

Two Twitter datasets were used. The first was created by the researcher using 1000 tweets drawn from the Twitter API. The second was drawn from a Kaggle data titled: *Tweets Related to Dental Care Affordability and Dental and Opioid Use.* The tweets were chosen based on searching the keywords "dentistry" and "dentist" on Twitter because they related to the topic.

Python Packages for Twitter Scraping (Tweepy).

In order to conduct Twitter scraping I used Python packages: Pandas, Numpy, Matplotlib, and Pyplot. The Python packages NLTK, OS, word cloud for diction analysis, and word cloud visualization.

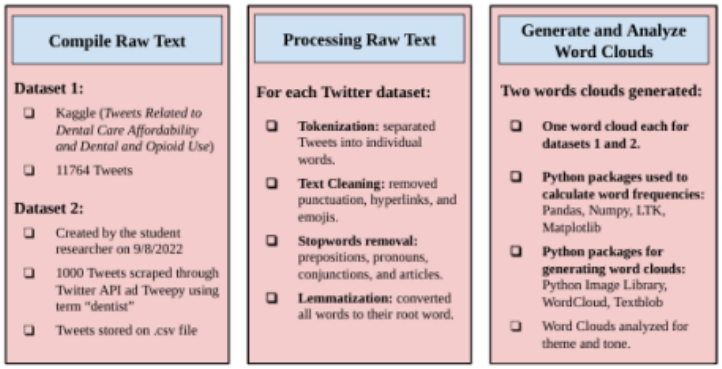
Regression Analyses Using SPSS

I conducted inferential analysis using multiple linear regression modeling to explore the relationship between the independent variables of age, marital status, earnings, race, political party, and possession of insurance and the dependent variables of delaying personal care, delaying family care, out-of-pocket expenditures for self, and out-of-pocket expenditures for family. A table summarizing these results is located in Appendix B.

Natural Language Processing Using Python Analyses Using SPSS

Natural Language Processing was used to collect data for the creation of a word cloud through Twitter Sentiment Analysis (Figure 1).

Figure 1. Procedure for Word Cloud Analysis



Importing Packages

Using Jupyter Notebook in Python, I first imported Pandas and Numpy in order to conduct statistical analyses of diction related to dental care affordability. I imported Textblob to process the Twitter data and visualized the data using a word cloud.

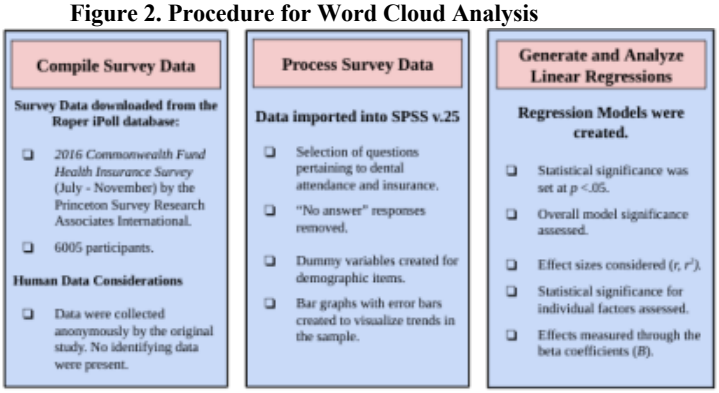
Data Processing

The tweets from Kaggle and the tweets I scraped were loaded into a notebook using the pd_readcsv command. Each tweet was composed of a string. I performed segmentations, removed stopwords (words such as, are, the), tokenized the words, stemmed the words, and conducted lemmatization (the gender of words: run, running, runs, ran are all the same). I also removed emojis, usernames, and retweets (RT).

Word Cloud Visualization and Analysis

Using the Python package word cloud, I created two word clouds for the dental tweets based on the frequency of the remaining words. The size of the words appeared as a function of their frequency in the tweets dataset – consequently, the larger the word, the more frequently it appeared in the dataset. I interpreted the word cloud by identifying themes among the words.

Figure 2. Procedure for Word Cloud Analysis



Results

Twitter Sentiment Analysis

Word Cloud A

Word Cloud A was generated from Python which displays a visual representation of the thoughts of individuals from a Twitter database on their opinion of the dentist in the United Kingdom, with 11764 Tweets (Figure 3). There are many words that are represented that convey the thematic elements of dental affordability. The larger words in Word Cloud A include "afford," "dentist," and "go."

Word Cloud B

Word Cloud B was generated from Python which displays a visual representation of the thoughts of individuals from a Twitter database on their opinion of the dentist in the United States of America, with 1000 Tweets (Figure 3). There are many words that are represented that convey the thematic elements of dental appointments. In Word Cloud B, the largest words are "dentist," "dental," and "teeth."

Figure 1. Method for Interpreting a Python Word Cloud

Private Health Insurance Coverage

I ran a Linear Regression model to determine if there was a relationship between the independent variables (age, marital status, earnings, race, political party) and the dependent variable (whether the individual was personally covered by private health insurance offered through an employer or

union). The overall model was statistically significant ($p < .05$). The independent factors that had a p-value of less than .05 were as follows:

- **Age.** Older individuals were less likely to have private health insurance ($B = -.19$).
- **Marital Status.** Married individuals were more likely to have private health insurance ($B = .10$).
- **Earnings.** Individuals who earned less than 35k were less likely to be covered by private health insurance ($B = .37$).
- **Race.** The only identified race that was statistically significant was that of Native Americans suggesting that Native Americans are less likely to have private health insurance ($B = -.03$).
- **Political Party.** Republicans were more likely to have private health insurance than Democrats ($B = .08$).

Delaying Personal Care

The overall model was statistically significant ($p < .05$). The r-value was .23 and the r^2 value was .05. The following independent variables were statistically significant ($p < .05$):

- **Private Personal Insurance.** Individuals who had personal insurance were less likely to delay personal care ($B = -.18$).
- **Age.** Older individuals were less likely to delay care ($B = -.14$).
- **Marital Status.** Married individuals were slightly more likely to delay care ($B = .01$).
- **Earnings.** Individuals earning less than 35k were more likely to delay care ($B = .07$).
- **Race.** Only Native Americans showed a slightly higher likelihood of delaying personal care ($B = .01$).
- **Political Party.** Democrats also showed a slightly higher likelihood of delaying personal care ($B = .01$).

Delaying Family Care

The overall model was statistically significant ($p < .05$). The r-value was .15 and the r^2 value was .02. The following independent variables were statistically significant ($p < .05$):

- **Private Family Insurance.** Individuals who had family insurance were less likely to delay family care ($B = -.10$).
- **Age.** Older individuals were less likely to delay care ($B = -.59$).
- **Marital Status.** Married individuals were more likely to delay care ($B = .05$).
- **Earnings.** Individuals earning less than 35k were more likely to delay care ($B = .08$).

Out-of-Pocket Expenditures for Self

The overall model was statistically significant ($p < .05$). The r-value was .18 and the r^2 value was .03. The following independent variables were statistically significant ($p < .05$):

- **Private Health Insurance.** Individuals who were personally covered by health insurance were less likely to pay out-of-pocket spending ($B = -.07$).
- **Age.** Older individuals were slightly more likely to pay out-of-pocket ($B = .03$).
- **Marital Status.** A married individual was less likely to pay out-of-pocket ($B = -.02$).
- **Earnings.** An individual who earned less than 35k was less likely to pay out-of-pocket which may be due to not earning enough ($B = -.05$).
- **Race.** All of the races had positive $B_{(constant)}$ which shows that they were more likely to pay out-of-pocket ($B = -.01$).

Out-of-Pocket Expenditures for Family

The overall model was statistically significant ($p < .05$). The r-value was .10 and the r^2 value was .01. The following independent variables were statistically significant ($p < .05$):

- **Private Family Insurance.** Individuals who had insurance to cover their child's procedure were less likely to pay out-of-pocket because they were spending less ($B = -.06$).
- **Age.** Older individuals were less likely to pay out-of pocket for their child's procedure ($B = -.01$).
- **Marital Status.** Married individuals were more likely to pay out-of-pocket for their child's procedure ($B = .08$).
- **Earnings.** Individuals that earned less than 35k were less likely to pay out-of-pocket ($B = -.05$).
- **Political Party.** The Republican and Independent party identifiers were slightly more likely to pay out-of-pocket ($B = .01$ and $B = .01$) than Democrats ($B = -.02$).

Discussion

This study highlighted the disparities in dental attendance. The results showcased that dental attendance across many demographic groups was shaped by their access to dental insurance. The individuals most likely to have health insurance included those who were younger, married, earning more than $35,000, and Republican. Americans who were more likely to delay personal health care included those who were uninsured, younger, married, earning less than $35,000, and Democrat. Individuals who were most likely to delay family care included those with no family insurance, younger, married, earning less than $35,000, and did not identify with a political party. Individuals most likely to pay out-of-pocket expenditures were those who were not covered by insurance, older, unmarried, earning less than $35,000, and did not identify with a political party. Those who were willing to pay out-of-pocket expenditures for their family included those who were uninsured, younger, married, earning less than $35,000, and a Republican or Independent party identifier. These results regarding insurance are expected, as those who don't have insurance are essentially forced to pay out-of-pocket.

The Gallup poll indicates that women had a higher dental attendance compared to men (Bushak, 2014). Additionally, much like the findings of my study, Whites were found in the Gallup poll to be more likely to visit the dentist than Blacks or

Hispanics. In 2019, the percentage of dental attendance of Whites was 68.3% while Blacks were 61.1% and Hispanics were 58.6%. In 2020, the percentage of dental attendance of Whites was 66.6% while Blacks were 56.8% and Hispanics were 55.3% (Elfein, 2022). This showcases that dental attendance has not improved much throughout the years from various races and ethnicities.

According to the Gallup poll, "Seventy percent of Whites and Asians visited in 2013, compared to the 55% of Blacks and Hispanics" (Bushak, 2014). This is similar to results from 2019 and 2020 as Whites and Asians had 68.3% and 70.1% of attendance in 2019 and 66.6% and 64.3% of attendance in 2020. The percentage of Blacks and Hispanics had 61.1% and 58.6% of attendance in 2019 and 56.8% and 55.3% of attendance in 2020 (Elfein, 2022). These results may be explained by the income level of these groups as a higher income would lead to a more likely chance to visit the dentist due to being able to afford the treatment.

References

2021 Alzheimer's disease facts and figures. (2021). *The Journal of the Alzheimer's Association, 17*(3), 327 – 406. https://doi.org/10.1002/alz.12328

Aslan Ceylan, J., Aslan, Y., & Ozcelik, A. O. (2022). The effects of socioeconomic status, oral and dental health practices, and nutritional status on dental health in 12-year-old school children. *Egyptian Pediatric Association Gazette*, *70*(13). https://doi.org/10.1186/s43054-022-00104-3

Beaton, L., Freeman, R., & Humphris, G. (2014). Why are people afraid of the dentist? Observations and explanations. *International Journal of the Kuwait University Health Science Centre, 23*(4), 295 – 301. https://doi.org/10.1159/000357223

Black, M. L. (2021, November 29). *Nearly half of insured Americans skip dental visits, procedures due to cost.* ValuePenguin. https://www.valuepenguin.com/dental-survey#:~:text=When%20looking%20at%20Amer%20icans%20without,can%20have%20unintended%20health%20consequences

Bushak, L. (2014, April 29). *Oral health isn't much of Americans' concern, poll finds: One-third didn't see the dentist last year.* Medical Daily. https://www.medicaldaily.com/oral-health-isnt-much-americans-concern-poll-finds-one-t hird-didnt-see-dentist-last-year-279468

Chemweno, J. (2021, July 27). *The U.S. healthcare system is broken: A national perspective.* Managed Healthcare Executive. https://www.managedhealthcareexecutive.com/view/the-u-s-healthcare-system-is-broken-a-national-perspective

Disparities in oral health. (2021, February 5). Centers for Disease Control and Prevention. https://www.cdc.gov/oralhealth/oral_health_disparities/index.htm

Dentistry Workers and Employers. (n.d.). United States Department of Labor. https://www.osha.gov/coronavirus/control-prevention/dentistry

Dentophobia (fear of dentists): Causes, symptoms & treatments. (2022). Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/22594-dentophobia-fear-of-dentists

Elflein, J. (2022, May 31). *Percentage of U.S. adults 18 to 64 years who stated they have had a dental visit in the past 12 months in 2019 and 2020, by race and ethnicity.* Statista. https://www.statista.com/statistics/1310635/adults-who-had-a-dental-visit-in-the-past-12- months-by-race-and-ethnicity/

Elflein, J. (2022, May 31). *Percentage of U.S. adults 18 to 64 years who stated they have had a dental visit in the past 12 months in 2019 and 2020, by sex.* Statista. https://www.statista.com/statistics/1310620/adults-who-had-a-dental-visit-in-the-past-12- months-by-sex/

Heart disease facts. (2022, October 14). Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm

How much does a dentist checkup cost? (2021, May 7). The Floss by Opencare. https://www.opencare.com/blog/how-much-does-a-dentist-checkup-cost/

Hung, M., Lipsky, M. S., Moffat, R., Lauren, E., Hon, E. S., Park, J., Gill, G., Xu, J., Peralta, L., Cheever, J., Prince, D., Barton, T., Bayliss, N., Boyack, W., & Licari, F. W. (2020). Health and dental care expenditures in the United States from 1996 to 2016. *PLOS One, 15*(6). https://doi.org/10.1371/journal.pone.0234459

Izzy. (2017, November 26). *Tweets related to dental care affordability and dental and opioid use.* Kaggle. https://www.kaggle.com/datasets/izzykayu/tweets

Kay, E. J. (1999). How often should we go to the dentist? *The BMJ, 319*(7204), 204 – 205, https://doi.org/10.1136/bmj.319.7204.204

Kojima, A., Nakano, K., Wada, K., Takahashi, H., Katayama, K., Yoneda, M., Higurashi, T., Nomura,

R., Hokamura, K., Muranaka, Y., Matsuhashi, N., Umemura, K., Kamisaki, Y., Nakajima, A., & Ooshima, T. (2012). Infection of specific strains of Streptococcus mutans, oral bacteria, confers a risk of Ulcerative Colitis. *Scientific Reports.* https://doi.org/10.1038/srep00332

Lockhart, P. B., Bolger, A. F., Papapanou, P. N., Osinbowale, O., Trevisan, M., Levison, M. E., Taubert, K. A., Newburger, J. W., Gornik, H. L., Gewitz, M. H., Wilson, W. R., Smith, S. C., Jr, & Baddour, L. M. (2012). Periodontal Disease and Atherosclerotic Vascular Disease: Does the evidence support an independent association? *Circulation, 125*(20), 2520 – 2544. https://doi.org/10.1161/CIR.0b013e31825719f3

Michas, F. (2022, September 15). *Dental visit in past year adults U.S. 1997-2019.* Statista. https://www.statista.com/statistics/187892/persons-with-a-dental-visit-in-the-past-year-in -the-us-since-1997/

Michaud, P.C., Goldman, D., Lakdawalla, D., Gailey, A., & Zheng, Y. (2011). Differences in health between Americans and Western Europeans: Effects on longevity and public finance. *Social Science & Medicine, 73*(2), 254 – 263. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383030/

Migas, K., Marczak, M., Kozłowski, R., Kot, A., Wysocka, A., & Sierocka, A. (2022). Impact of the COVID-19 pandemic on the dental preferences of patients in the private sector. *International Journal of Environmental Research and Public Health, 19*(4). https://doi.org/10.3390/ijerph19042183

Natural language processing (part 3): Exploratory data analysis & word clouds in Python. (2019, January 5). YouTube. https://www.youtube.com/watch?v=VraAbgAoYSk&t=606s.

Preshaw, P. M., Alba, A. L., Herrera, D., Jepsen, S., Konstantinidis, A., Makrilakis, K., & Taylor, R. (2012). Periodontitis and diabetes: A two-way relationship. *Diabetologia*, *55*(1), 21 – 31. https://doi.org/10.1007/s00125-011-2342-y

Statistics about diabetes. (n.d.). American Diabetes Association. https://diabetes.org/about-us/statistics/about-diabetes#:~:text=Deaths,a%20total%20of%20282%2C801%20certificates

Study reveals how too much fluoride causes defects in tooth enamel. (2020, February 18). New York University News. https://www.nyu.edu/about/news-publications/news/2020/february/fluorosis.html#:~:text=The%20researchers%20found%20that%20exposing,many%20functions%2C%20including%20storing%20calcium.

The Common Wealth Fund. (2016). 2016 commonwealth fund health insurance survey. *Princeton Survey Research Associates International.* https://doi.org/10.25940/ROPER-31115604

Tofangchiha, M., Lin, C.Y., Scheerman, J. F. M., Broström, A., Ahonen, H., Griffiths, M. D., Tadakamadla, S. K., & Pakpour, A. H. (2022, June 27). *Associations between fear of Covid-19, dental anxiety, and psychological distress among Iranian adolescents.* Nature News. https://www.nature.com/articles/s41405-022-00112-w

Word cloud using python. (2020, November 17). AskPython. https://www.askpython.com/python/examples/word-cloud-using-python