COVID-19 and Factors Relating to Healthcare Attrition

Kim Bui & Camille Eastvold

https://github.com/kimberly-bui/Wrangling-Final-Project

1. Introduction

Since 2019, the coronavirus has been infecting the world. Years later the U.S. is still dealing with the symptoms of high turnover and poor retention rates across the healthcare industry. Is the job market still in recovery?

The COVID-19 pandemic has had profound and lasting effects on the healthcare sector, particularly in the United States. Healthcare institutions continue to grapple with issues such as high turnover, reduced job satisfaction, and an increased workload for remaining staff. Nurses, as a cornerstone of the healthcare workforce, have been disproportionately affected.

This analysis finds trends in job satisfaction, income, and overtime hours by comparing data from 2018 (pre-pandemic) and 2022 (post-pandemic). By identifying key factors contributing to healthcare attrition, this research aims to provide actionable insights for improving retention and addressing burnout among healthcare workers. In this project, we are using 2022 Healthcare Employee Attrition[1] dataset to find emerging trends after the pandemic to see the changes for nurses using national report data, we obtained from 2018 National Sample Survey of Registered Nurses Nursing Solutions Inc (NSI)[2].

2. Data

This project uses two primary sources of data: both are survey results from 2022 Healthcare Employee Attrition dataset from Kaggle and 2018 National Survey for Nursing Solutions Inc.

2.1 2018 Survey Data

We collected data from the National Sample Survey of Registered Nurses Nursing Solutions Inc, found in the National Library of Medicine.

The overall findings of the survey were presented in table 3 of this report. Through web scraping script, *00_webscrape.ipynb* we were able to gather all contents of the table.

[1] https://www.kaggle.com/datasets/jpmiller/employee-attrition-for-healthcare?resource=download

[2] https://pmc.ncbi.nlm.nih.gov/articles/PMC10742910/#healthcare-11-03173-t001

The table contained metrics from original 2018 survey data and simulated data calculated by an algorithm. After importing the scraped data into a data frame, we dropped unnecessary columns including the simulated algorithm data. We kept the real survey results in 2018. In the report response data, we gathered results from 43,937 Registered Nurses and transformed the results into a simple, easily interpreted table. We transposed the data frame to replicate the table in the original report. For our last steps before merging the data frames, we converted all data types for each attribute to be float. Then saved the cleaned data into *clean_webscraped.csv.*

2.2 2022 Survey Data

The second data source we will use in this project is Employee Attrition for Healthcare dataset from Kaggle, derived from 2022 NSI National Health Care Retention Report. This data shows survey data from 272 hospitals across 32 states. This survey covers 589,901 healthcare workers, and 166,087 Registered Nurses. We created a data frame by importing the *raw_2022.csv* into *01_kaggle.ipynb* that can be found in our project folder. There was not much cleaning needed for this dataset. However, we did drop columns containing factors not recorded in the 2018 data and changed all data types to float. After cleaning and transforming we converted the data frame into *clean_kaggle.csv* file.

2.3 Combining 2018 and 2022 Data

To ensure seamless integration, during the cleaning stages of each source, we confirmed each data frame had the same data types and variables. A description of each variable is listed in *Table 1, 2 and 3.*

The format was crucial in order for us to be able to merge the datasets. The data we web scraped was a table that took averages/count from all the survey results. By pulling key factors and calculating the averages/count from our Kaggle dataset, we were able to create identical tables. For this analysis we decided to use the merged data frame to see the overall differences between 2018 and 2022 attrition trends.

Once both data frames were transposed and calculated into similar tables, we renamed all the columns by adding the year the responses were pulled from, for example: overtime_2018 and overtime_2022. Then created a union merger between both datasets, this combined and showed all columns/rows for both datasets. In the *merge_df*, there is data comparing factors contributing to attrition rates, before and after COVID-19. Lastly, once we merged, we rounded each data point to two decimal places and downloaded the results into *merged.csv* file.

Table 1: Data Dictionary from Kaggle

Column	Type	Source	Description
Age	Numeric	Kaggle	Age of Employee.
Attrition	Text	Kaggle	Whether the employee has left the organization.
Gender	Text	Kaggle	Gender of the employee.
Job_Satisfaction	Numeric	Kaggle	Satisfaction with the job.
Marital_Status	Text	Kaggle	Marital status of the employee.
Income	Numeric	Kaggle	Yearly income of the employee.
OverTime	Text	Kaggle	Whether the employee works overtime.

Table 2: Data Dictionary from webscrape

Column	Type	Source	Description
age	Numeric	Survey	Age of the respondent
job_dissatisfied	Numeric	Survey	Number of respondents dissatisfied with their job
job_satisfied	Numeric	Survey	Number of respondents satisfied with their job
female	Numeric	Survey	Number of female respondents
male	Numeric	Survey	Number of male respondents
married	Numeric	Survey	Number of married respondents
married	Numeric	Survey	Number of married respondents

Table 3: Data Dictionary from Merge

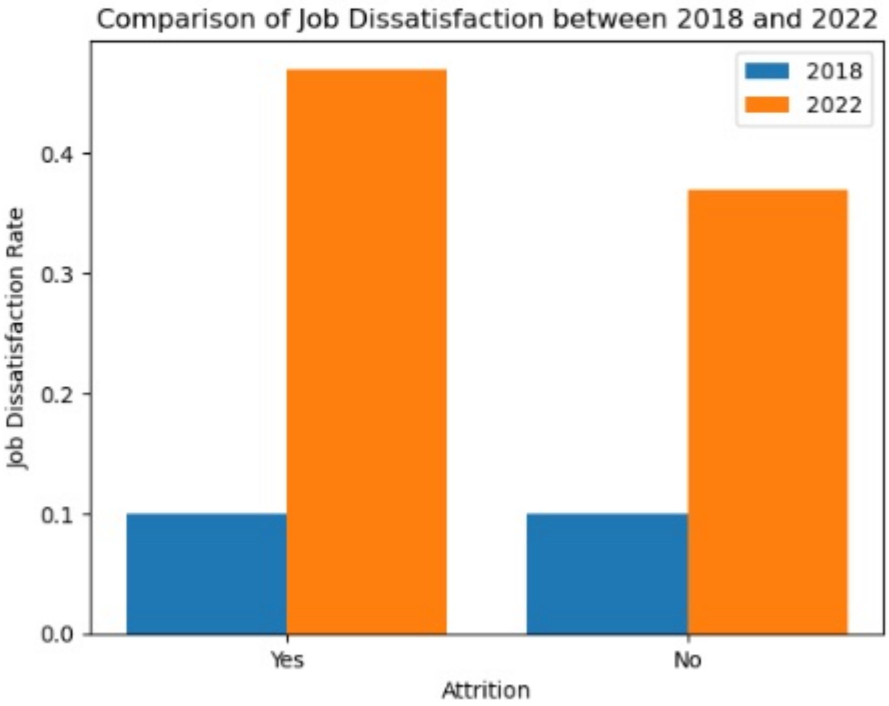
Column	Type	Source	Description
Attrition	Text	Both	Category showing if the respondents left their job = Yes or stayed working = No.
job_satisfied_2018	Numeric	Both	Number of respondents satisfied with their job in 2018
female_2018	Numeric	Both	Number of female respondents in 2018
male_2018	Numeric	Both	Number of male respondents in 2018
married_2018	Numeric	Both	Number of married respondents in 2018
single_2018	Numeric	Both	Number of single respondents in 2018
overtime_2018	Numeric	Both	Number of respondents working overtime in 2018
individual_income_2018	Numeric	Both	Individual income of the respondents in 2018
job_dissatisfied_2022	Numeric	Both	Number of respondents dissatisfied with their job in 2022
job_satisfied_2022	Numeric	Both	Number of respondents satisfied with their job in 2022
female_2022	Numeric	Both	Number of female respondents in 2022
male_2022	Numeric	Both	Number of male respondents in 2022
married_2022	Numeric	Both	Number of married respondents in 2022
single_2022	Numeric	Both	Number of single respondents in 2022

Column	Type	Source	Description
Attrition	Text	Both	Category showing if the respondents left their job = Yes or stayed working = No.
overtime_2022	Numeric	Both	Number of respondents working overtime in 2022
individual_income_2022	Numeric	Both	Individual income of the respondents in 2022

3. Analysis

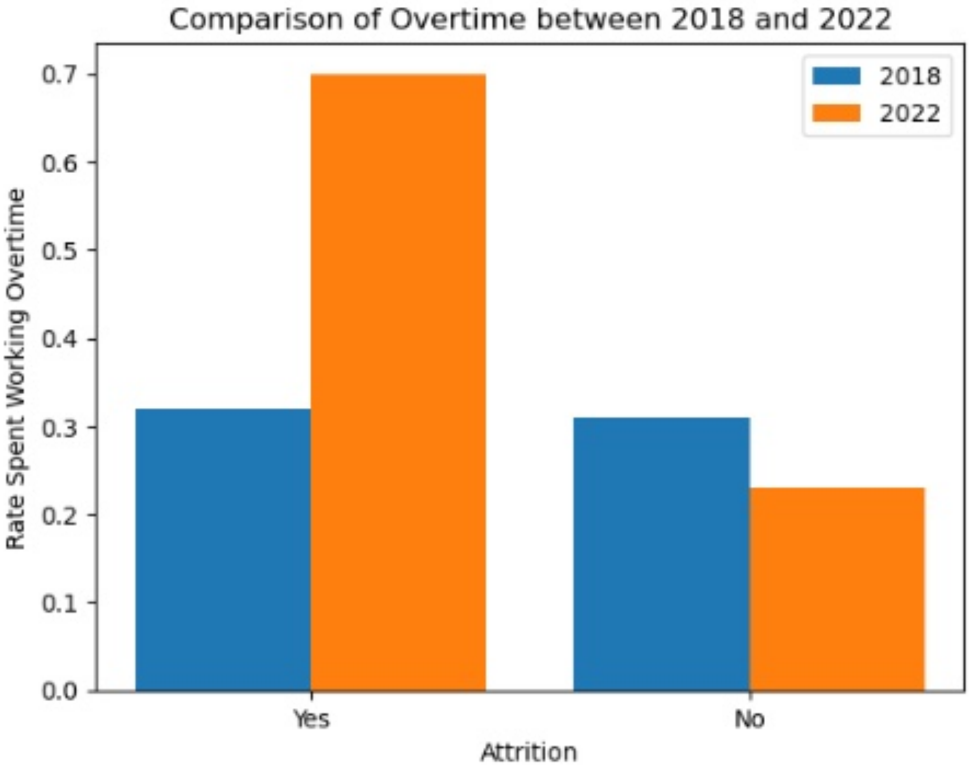
This project aims to analyze the effects COVID-19 had on employment in healthcare with job satisfaction, environment ratings, pay and other ratings. We want to discover reasons for quitting and ways healthcare industries can improve. Our research analysis includes:

Has job satisfaction for nurses changed since COVID?



According to our dataset, you can see how many people were dissatisfied with their job after COVID compared to before in 2018. Rates have gone from nearly 10% dissatisfaction to nearly 50% of nurses who took the survey. This resulted in those who were dissatisfied leaving their jobs.

With high nurse turnover, has working overtime contributed to attrition rates?



In the chart above you can see how those who left their jobs were working more overtime in 2022. In 2018, the average time spent working overtime was less than half. Since COVID, the overtime rates have increased.

In our research we discovered the healthcare industry is facing a period called the ‘Great Resignation’ according to the 2022 NSI National Health Care Retention & RN Staffing Report. This is due to the high turnover rates due to COVID.

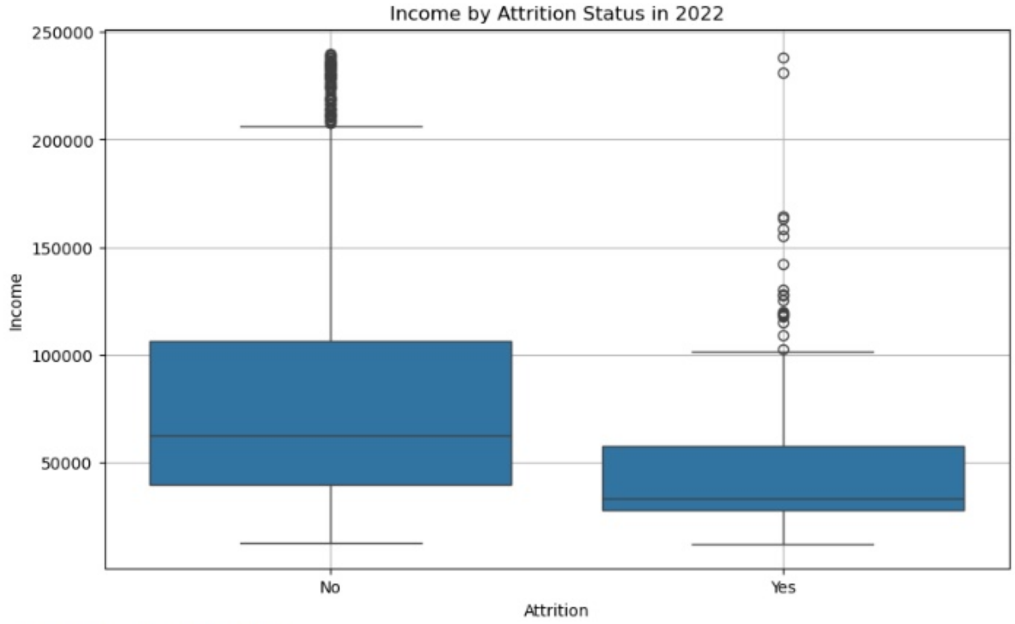
Post-pandemic job satisfaction dropped due to factors such as:

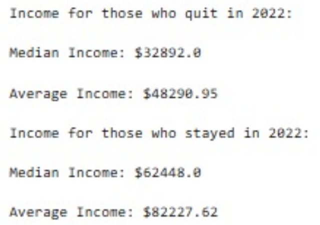
- Increased workload and stress during the pandemic.
- Emotional and physical toll on nurses.

- Persistent staffing shortages.

Here you can see evidence of that, the more shortages, the more others have to work overtime which can lead to turnover.

Does income factor into attrition?

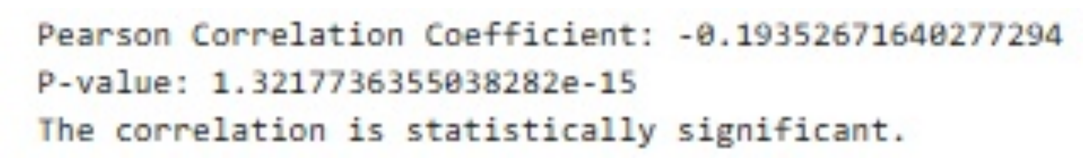




```
Income for those who quit in 2022:

Median Income: $32892.0

Average Income: $48290.95

Income for those who stayed in 2022:

Median Income: $62448.0

Average Income: $82227.62
```

Diving deeper into post pandemic data, we can see significant differences in income between those who left their jobs and those who stayed. As we've learned COVID has put extreme tolls on healthcare workers. Low job satisfaction and increased workload can cause those who do not feel they are being compensated fairly to leave. In our findings between both parties, there is over a $20,000 dollar difference between income.

Is there a direct correlation between income and attrition?

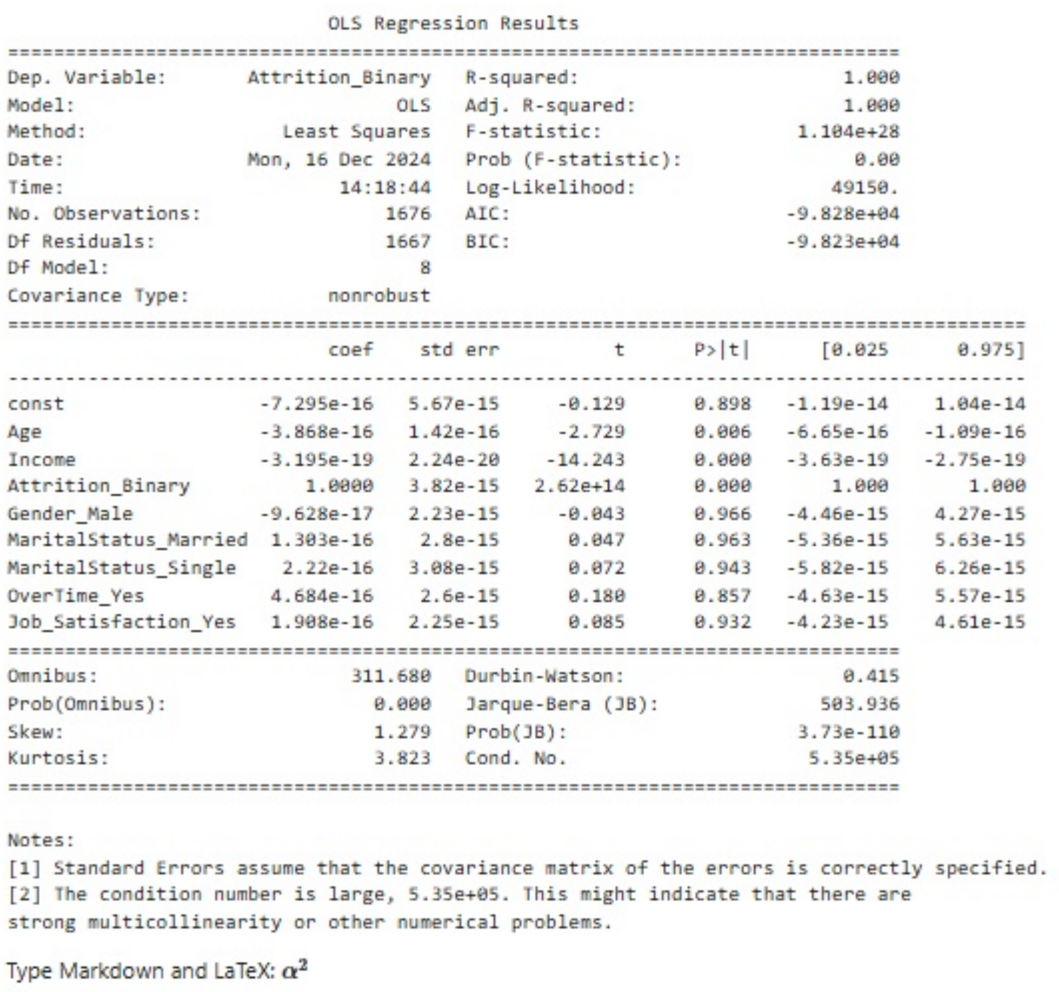


```
Pearson Correlation Coefficient: -0.19352671640277294
P-value: 1.3217736355038282e-15
The correlation is statistically significant.
```

For this analysis we used Pearson Correlation Coefficient test to check for a direct relationship between income and attrition. This test shows there is a statistically significant weak correlation between Attrition and Income in the year 2022. Since our p value of 1.3217736355038282e-15 was less than .05, that means it is significant enough to reject null hypothesis that there is no relationship. However, the correlation value of -0.19352671640277294, indicated a weak negative correlation. In conclusion, this means income may have an impact on whether a nurse quits or not, but it is not the dominating factor. Perhaps it is a combination of factors.

Which factors contribute to high attrition rates?

To answer this supervised regression problem, we decided to run a linear regression model analysis. Our target variable is Attrition and the remaing variables are the features we are testing. In order to run a linear regression model we must make our target variable numeric. Using one hot encoding to transform attrtion category to numeric values of 0 and 1, the model tested for any correlation among any of the variables listed.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Attrition_Binary   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.104e+28
Date:                Mon, 16 Dec 2024   Prob (F-statistic):               0.00
Time:                        14:18:44   Log-Likelihood:                 49150.
No. Observations:                1676   AIC:                        -9.828e+04
Df Residuals:                    1667   BIC:                        -9.823e+04
Df Model:                           8
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                 -7.295e-16   5.67e-15     -0.129      0.898   -1.19e-14    1.04e-14
Age                   -3.868e-16   1.42e-16     -2.729      0.006   -6.65e-16   -1.09e-16
Income                -3.195e-19   2.24e-20    -14.243      0.000   -3.63e-19   -2.75e-19
Attrition_Binary          1.0000   3.82e-15   2.62e+14      0.000       1.000       1.000
Gender_Male           -9.628e-17   2.23e-15     -0.043      0.966   -4.46e-15    4.27e-15
MaritalStatus_Married  1.303e-16    2.8e-15      0.047      0.963   -5.36e-15    5.63e-15
MaritalStatus_Single    2.22e-16   3.08e-15      0.072      0.943   -5.82e-15    6.26e-15
OverTime_Yes           4.684e-16    2.6e-15      0.180      0.857   -4.63e-15    5.57e-15
Job_Satisfaction_Yes   1.908e-16   2.25e-15      0.085      0.932   -4.23e-15    4.61e-15
==============================================================================
Omnibus:                      311.680   Durbin-Watson:                   0.415
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              503.936
Skew:                           1.279   Prob(JB):                    3.73e-110
Kurtosis:                       3.823   Cond. No.                     5.35e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.35e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Type Markdown and LaTeX: α^2

The regression results indicate higher income slightly reduces likelihood of attrition. Younger nurses are slightly more likely to leave. Other factors are not significant. This could be due to the overfitting model. The R squared value is 1 which means it is overfitting because there is no unexplained variability.

4. Conclusion

This study investigated the impact of COVID-19 on healthcare worker job satisfaction, income, and overtime hours, ultimately affecting employee attrition rates. By analyzing data from 2018 (pre-pandemic) and 2022 (post-pandemic), we identified key trends and potential contributing factors.

Key Findings:

- **Job Satisfaction:** Job satisfaction in the healthcare sector declined significantly post-pandemic. This decrease is likely due to factors such as increased workload, stress, and staffing shortages.
- **Income and Job Satisfaction:** While income remains a factor in job satisfaction, there is still a weak correlation. This suggests other factors, like work-life balance, have become more significant.
- **Overtime Hours:** Overtime hours increased considerably in 2022 compared to 2018. This could be attributed to burnout, staffing shortages, or changes in workload distribution.

Implications:

- The decline in job satisfaction highlights the need for healthcare institutions to prioritize employee well-being. Initiatives promoting work-life balance, stress management, and adequate staffing can help improve employee retention.
- While income may not be the sole driver of job satisfaction, it remains important. Fair compensation strategies that acknowledge the demanding nature of healthcare work are crucial.
- The decrease in overtime hours may be a positive sign for employee well-being, but it also suggests potential staffing challenges. Strategies to address workload distribution and attract new nurses are essential.

Limitations:

- This study relied on publicly available datasets, which may have limitations or biases.
- Correlation does not imply causation. Further research is needed to determine the specific causal factors influencing job satisfaction and attrition.
- The sample sizes vary between two datasets. There were more responses in 2022, giving more room for variance.

Future Research:

- Analyzing data on specific job roles within healthcare can provide more targeted insights.
- Qualitative studies exploring the experiences of healthcare workers could provide deeper understanding of their concerns and motivations.
- Longitudinal studies can track trends over time and assess the effectiveness of interventions aimed at improving employee retention.

By addressing the challenges identified in this study, healthcare institutions can create a more supportive work environment and improve employee retention. This, in turn, will benefit the quality of patient care and the overall health of the healthcare system.