**Healthcare Data Analysis Case Study: Exploratory Data Analysis**

Author:

Joanne St.Remy

**Table of Contents:**

1. Introduction
   1. Project Goal
   2. Background on the Dataset
2. Demographic Factors and Diabetes
   1. Diabetes Prevalence by Country of Birth
   2. Diabetes Prevalence by Gender
   3. Diabetes Prevalence by Race
   4. Diabetes Prevalence Adjusted for Sample Size
   5. Age of Diabetes Diagnosis by Race
   6. Age of Diabetes Diagnosis by Country of Origin
3. Employment Status and Diabetes
4. Diabetes by Highest Education Level
5. Diabetes by Region
   1. Number of Individuals with Diabetes by Region
   2. Average Diabetes Diagnosis Age by Region
6. Diabetes and High Cholesterol
7. Impact of Diabetes on Happiness Level
8. Conclusion
   1. Key Insights
   2. Recommendations
   3. Ongoing Research and Updates

**Healthcare Data Analysis Case Study: Exploratory Data Analysis**

**Project Goal:** The goal of this project was to clean, analyze, and visualize data from the "MEPS HC-228 2021 Full Year Population Characteristics" dataset. The project aimed to identify trends between diabetes and certain key demographic characteristics, focusing on age, sex, race, ethnicity, employment status, education, and country of birth.

**Background on the Dataset:**

The dataset is a public use file containing information collected on a nationally representative sample of the civilian noninstitutionalized population of the United States for the calendar year 2021. It includes data from multiple rounds of the Medical Expenditure Panel Survey (MEPS). The dataset contains variables related to survey administration, demographics, person-level conditions, health status, disability days, quality of care, health care delays due to COVID-19, COVID-19 vaccinations, Social Determinants of Health (SDOH), employment, and health insurance.

**Key Findings from the Exploratory Data Analysis:**

**1. Demographic Factors and Diabetes:**

* The dataset included variables related to demographics such as age, sex, race, ethnicity, employment status, education, and country of birth.
* The goal was to investigate the correlation between demographic factors, especially race, and the occurrence of diabetes.
* A hypothesis was proposed that certain racial groups may have a higher prevalence of diabetes.
* The dataset was filtered to select key variables of interest.

**2. Diabetes Prevalence by Country of Birth:**

* Among the individuals with diabetes in the dataset, approximately ***79.26%*** were born in the USA, suggesting a higher chance of being diagnosed with diabetes for US-born individuals.
* A small number of individuals (7) refused to answer the survey question about their country of birth.

**3. Diabetes Prevalence by Gender:**

* Among the individuals with diabetes, there were 1,518 men and 1,747 women.
* The average age of diabetes diagnosis was 48 for men and 47 for women, indicating no significant disparity between genders.
* -The average age of diabetes diagnosis for the entire sample was 48.

|  |  |
| --- | --- |
| **Diabetes by Gender** | |
| **Men** | **Women** |
| 1,518 | 1,747 |
| *46%* | *54%* |

**4. Diabetes Prevalence by Race:**

* The dataset provided information on different racial groups: Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, and Non-Hispanic Other.
* The counts of individuals with diabetes were as follows:
* Hispanic: 703
* Non-Hispanic White: 1,629
* Non-Hispanic Black: 654
* Non-Hispanic Asian: 143
* Non-Hispanic Other: 136
* Visualizations were created to display the diabetes prevalence for each racial group, shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Diabetes Prevalence by Racial Group** | | | | |
| **White Only** | **Hispanic Only** | **Black Only** | **Asian Only** | **Other Reported Race** |
| 1629 | 703 | 654 | 143 | 136 |

A graph of a number of people

Description automatically generated

5. Diabetes Prevalence Adjusted for Sample Size:

* To account for differences in the sample sizes of each racial group, the counts of individuals with diabetes were divided by the total count for each race.
* The resulting percentages were as follows:
* Hispanic: 10%
* Non-Hispanic White: 11%
* Non-Hispanic Black: 16%
* Non-Hispanic Asian: 10%
* Non-Hispanic Other: 13%
* These percentages provide a more accurate comparison of diabetes prevalence among different racial groups.

Here's a breakdown of the insight from these findings:

1. **Non-Hispanic Black:** This group has the highest diabetes prevalence at 16%. It indicates that a significant portion of the individuals in the Non-Hispanic Black racial group have been diagnosed with diabetes.
2. **Non-Hispanic Other:** The diabetes prevalence in this group is 13%, which is the second highest among the racial groups considered. It suggests a relatively higher occurrence of diabetes within the Non-Hispanic Other racial category.
3. **Non-Hispanic White:** The diabetes prevalence for this group is 11%. It is the third highest among the racial groups and indicates that a considerable proportion of Non-Hispanic White individuals in the dataset have been diagnosed with diabetes.
4. **Hispanic:** The diabetes prevalence for this group is 10%, which is lower than the other three groups mentioned above. It suggests that a smaller percentage of Hispanic individuals in the dataset have been diagnosed with diabetes.
5. **Non-Hispanic Asian:** Similar to the Hispanic group, the diabetes prevalence for Non-Hispanic Asians is also 10%. This indicates a relatively lower diabetes prevalence within the Non-Hispanic Asian racial category.

The adjusted percentages allow for a fairer comparison of diabetes rates among different racial groups because they take into account the differences in sample sizes. This means that the percentages are based on the proportion of individuals with diabetes within each racial group relative to the total number of individuals in that group.

Overall, these findings shed light on the variation in diabetes prevalence across racial groups in the dataset. It highlights the importance of understanding healthcare disparities and considering cultural and genetic factors that may influence the risk of diabetes among different populations. Further research and analysis could delve deeper into the underlying reasons for these differences and help inform targeted interventions to improve healthcare outcomes for diverse communities.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Diabetes Prevalence Adjusted for Sample Size** | | | | |
|  |
| **White Only** | **Hispanic Only** | **Black Only** | **Asian Only** | **Other Reported Race** |  |
| 11% | 10% | 16% | 10% | 13% |  |

A graph of diabetes prevalence by racial groups

Description automatically generated

**6. Age of Diabetes Diagnosis by Race:**

* The average age of diabetes diagnosis varied among racial groups.
* White and Asian individuals tend to receive a diabetes diagnosis at a later age compared to Black, Hispanic, and other racial groups.
* Hispanic and other racial groups tend to receive a diabetes diagnosis around the age of 43-44. This finding suggests that cultural foods may have a lesser influence on diabetes development when compared to the potential impact of a US diet. It implies that non-US born Hispanics, who are more inclined to cherish their cultural cuisine rather than indulge in American dishes, may demonstrate a potentially lower diabetes risk.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Diabetes Diagnosis Age by Racial Group** | | | | |
| **Racial Group** | **Average Age** | **Minimum Age** | **Maximum Age** | **Median Age** |  |
| **Hispanic** | **44** | 0 | 80 | 46 |  |
| **White** | **52** | 1 | 85 | 54 |  |
| **Black** | **48** | 5 | 85 | 50 |  |
| **Asian** | **50** | 5 | 74 | 55 |  |
| **Other** | **43** | 5 | 70 | 43 |  |

***Age of Diabetes Diagnosis Varies Across Racial Groups: White and Asian individuals tend to receive diagnoses later, while Hispanic and other racial groups are diagnosed around age 43-44. Cultural foods may play a role, suggesting potentially lower diabetes risk among non-US born Hispanics.***

**Diabetes Diagnosis Age by Racial Group**

A blue bar graph with numbers

Description automatically generated

**7. Age of Diabetes Diagnosis by Country of Origin:**

* Among US-born Hispanics, the average age of diabetes diagnosis was 44.
* Among Hispanics not born in the USA, the average age of diabetes diagnosis was 49.
* A graph of a number of people

  Description automatically generatedThis suggests that US-born Hispanics receive a diabetes diagnosis at an earlier age compared to Hispanics not born in the USA.

|  |  |
| --- | --- |
| **Diabetes by Country of Birth: Hispanic Origins** | |
| **Average Diagnosis Age** | |
| **US Born Hispanics** | **Non-US Born Hispanics** |
| **44** | **49** |

**8. Employment Status and Diabetes:**

* Individuals who have worked that reported having diabetes: 1321
* Individuals who have never worked that reported having diabetes: 457
* These results may be caused by other variables such as age, as the survey contains data from children who are unable to work. So, this may only suggest that working adults are more likely to have a diabetes diagnosis which isn’t particularly surprising.
* Further analysis can be conducted to explore the correlation between employment status and diabetes diagnosis.

**9. Diabetes by Highest Education Level**

|  |  |
| --- | --- |
| **Highest Degree** | **# of Individuals with Diabetes** |
| **High School Diploma** | 1332 |
| **No Degree** | 652 |
| **Bachelor's Degree** | 443 |
| **Other Degree** | 332 |
| **Master's Degree** | 202 |
| **GED** | 187 |
| **Doctorate Degree** | 66 |

**A blue cubes with white text

Description automatically generated**

1. **Diabetes Prevalence and Education Level:** The data showcases a clear association between education level and diabetes prevalence. As the education level increases, the number of individuals with diabetes tends to decrease.
2. **Lower Education Levels and Higher Diabetes Prevalence:** Individuals with lower education levels, such as those with High School Diploma, No Degree, and GED, have relatively higher diabetes prevalence. This suggests that lower educational attainment might be linked to an increased risk of diabetes.
3. **Higher Education Levels and Lower Diabetes Prevalence:** On the other hand, individuals with higher education levels, specifically Master's Degree and Doctorate Degree holders, exhibit the lowest diabetes prevalence. This indicates a potential protective effect of higher education against diabetes risk.
4. **Education and Health Behaviors:** Education can influence health behaviors and lifestyle choices. Individuals with higher education levels might be more aware of healthy practices, leading to a lower risk of diabetes.
5. **Socioeconomic Factors:** Education is often associated with socioeconomic status. Higher education levels might provide individuals with better access to resources, including healthcare, which can contribute to better diabetes management and prevention.
6. **Targeted Interventions:** The findings suggest that targeted interventions focusing on diabetes prevention and management could be beneficial for individuals with lower education levels. Such interventions might include educational programs on diabetes risk factors, healthy lifestyle promotion, and access to healthcare services.
7. **Further Research:** While the data provides valuable insights, additional research is needed to understand the underlying reasons behind the observed trends. Exploring factors such as dietary habits, physical activity, genetics, and access to healthcare among different education levels could provide a more comprehensive understanding of the relationship between education and diabetes prevalence.

The data’s insights indicate a significant association between education level and diabetes prevalence. Individuals with higher education levels tend to have lower diabetes prevalence, while those with lower education levels show higher prevalence rates. Understanding these associations can guide healthcare professionals and policymakers in implementing targeted interventions to reduce diabetes risk and improve overall health outcomes for diverse populations.

**10. Diabetes by Region**

**Number of Individuals with Diabetes by Region**

**A close-up of a cube

Description automatically generated**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Region** | **South** | **West** | **Midwest** | **Northeast** |
| **# of Individuals With Diabetes** | 1353 | 722 | 642 | 487 |

**Average Diabetes Diagnosis Age by Region**

**A close-up of a cube

Description automatically generated**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Region** | **Midwest** | **West** | **Northeast** | **South** |
| **Average Age of Diagnosis** | 49 | 48 | 48 | 47 |

**Number of Individuals with Diabetes by Region:**

1. **South:** The South region has the highest number of individuals with diabetes (1353), indicating a potentially higher prevalence of diabetes in this area.
2. **West:** The West region follows the South with 722 individuals diagnosed with diabetes, suggesting a significant diabetes burden in this region as well.
3. **Midwest:** The Midwest region reports 642 individuals with diabetes, indicating a considerable diabetes prevalence in this part of the country.
4. **Northeast:** The Northeast region has the lowest number of individuals with diabetes (487) among the four regions, indicating a relatively lower prevalence of diabetes compared to the other regions.
5. **Regional Disparities:** The data shows regional disparities in diabetes prevalence, highlighting the need for targeted healthcare interventions and public health efforts in regions with higher diabetes burdens.

**Average Diabetes Diagnosis Age by Region:**

1. **Midwest:** Individuals in the Midwest region tend to receive a diabetes diagnosis at an average age of 49, which is the highest among the four regions.
2. **West:** The West region follows closely with an average age of diabetes diagnosis at 48, suggesting a similar trend to the Midwest.
3. **Northeast:** The Northeast region also reports an average age of 48 for diabetes diagnosis, aligning with the West and Midwest regions.
4. **South:** The South region has the lowest average age of diabetes diagnosis, with individuals being diagnosed at an average age of 47.

**Age and Diabetes Diagnosis:** The data reveals regional variations in the average age of diabetes diagnosis, suggesting potential differences in diabetes management and healthcare practices across regions.

Regional Health Factors: The differences in diabetes prevalence and average diagnosis age among regions might be influenced by various factors, including lifestyle choices, genetics, access to healthcare, and public health initiatives.

**Targeted Interventions:** Understanding regional disparities in diabetes prevalence and diagnosis age can aid in developing targeted interventions to address specific health challenges faced by each region.

The data provides insights into the diabetes prevalence and average diagnosis age across different regions. The South region has the highest number of individuals with diabetes, while the Midwest reports the highest average age of diagnosis. Addressing regional disparities and understanding factors contributing to diabetes prevalence and age of diagnosis can guide healthcare professionals and policymakers in developing effective strategies for diabetes prevention, management, and equitable healthcare access across diverse regions. Further research and targeted interventions are essential to improve diabetes outcomes and promote overall health and well-being in each region.

**11. Diabetes and High Cholesterol**

|  |  |  |
| --- | --- | --- |
| **% of Sample That Reported High Cholesterol** | **Reported High Cholesterol** | |
| **Race** | **No** | **Yes** |
| **Asian Only** | 1% | 3% |
| **Black Only** | 6% | 14% |
| **Hispanic** | 6% | 15% |
| **Other or Multiple** | 1% | 3% |
| **White Only** | 12% | 38% |
| **High Cholesterol Reporting by Race:**  Asian Only: 3% Yes, 1% No Black Only: 14% Yes, 6% No Hispanic: 15% Yes, 6% No Other or Multiple: 3% Yes, 1% No White Only: 38% Yes, 12% No | | |
|  |
|  |
|  |

**A graph of high cholesterol

Description automatically generated**

**High Cholesterol Reporting by Race:**

1. **Asian Only:** Asian individuals in the sample have a relatively low percentage (3%) reporting high cholesterol, while 1% reported no high cholesterol.
2. **Black Only:** Black individuals show a higher percentage (14%) reporting high cholesterol compared to Asians, with 6% reporting no high cholesterol.
3. **Hispanic Only:** Hispanic individuals also exhibit a higher percentage (15%) reporting high cholesterol, similar to the Black population, and 6% reported no high cholesterol.
4. **Other or Multiple Race Reported:** Individuals with multiple racial backgrounds have a low percentage (3%) reporting high cholesterol, and 1% reported no high cholesterol.
5. White Only: White individuals have the highest percentage (38%) reporting high cholesterol, indicating a higher prevalence of high cholesterol in this racial group. Additionally, 12% reported no high cholesterol.

**Cholesterol Levels and Racial Variations:** The table illustrates differences in high cholesterol reporting across racial groups, with higher percentages reported among Black, Hispanic, and White individuals compared to Asian and Other/Multiple racial groups.

**Health Disparities:** Variations in cholesterol reporting might be linked to health disparities, lifestyle choices, genetic factors, and access to healthcare services among different racial populations.

**Risk Factors and Interventions:** Understanding the relationship between race and high cholesterol reporting can help identify at-risk populations and develop targeted interventions for cholesterol management and cardiovascular health within each racial group.

**Promoting Heart Health:** Given the higher prevalence of high cholesterol among some racial groups, targeted educational programs on heart health and healthy lifestyle choices could be beneficial in reducing cholesterol-related health risks.

**12. Impact of Diabetes on Happiness Level**

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of Individuals** | **Level of Happiness** | | |
| **Race** | **No Problem** | **Some Problem** | **Very Big Problem** |
| **Asian Only** | 100% |  |  |
| **Black Only** | 25% | 25% | 50% |
| **Hispanic** | 90% | 10% |  |
| **Other or Multiple** | 100% |  |  |
| **White Only** | 60% | 40% |  |
| **Happiness Levels Across Racial Groups:  Asian Only: 100% No Problem (High happiness level) Black Only: 25% Some Problem, 50% Very Big Problem (Significant challenges with happiness) Hispanic: 90% No Problem, 10% Some Problem (Generally high happiness level) Other or Multiple: 100% No Problem (Positive happiness level) White Only: 60% No Problem, 40% Some Problem (Mixed happiness experiences) Insights show varying happiness levels among racial groups, with notable challenges for Black and White individuals.** | | | |

A diagram of a variety of levels

Description automatically generated

The table displays the number of individuals from different racial groups and their levels of happiness categorized into three options: "No Problem," "Some Problem," and "Very Big Problem."

1. **Asian Only:** All individuals in the Asian Only group reported having "No Problem" with happiness, indicating a high level of happiness among this racial group.

2. **Black Only:** Among the individuals in the Black Only group, 25% reported "Some Problem" with happiness, and 50% reported a "Very Big Problem," indicating that a substantial proportion of this group is experiencing challenges with happiness.

3. **Hispanic:** The majority (90%) of individuals in the Hispanic group reported having "No Problem" with happiness, while 10% reported "Some Problem." This suggests that the overall level of happiness in this group is relatively high.

4. **Other or Multiple:** Similar to the Asian Only group, all individuals in the Other or Multiple group reported "No Problem" with happiness, indicating a positive level of happiness.

5. **White Only:** Among individuals in the White Only group, 60% reported "No Problem" with happiness, while 40% reported "Some Problem." This suggests that a significant portion of this group is experiencing some challenges with happiness.

**Analysis:**

The table provides insights into the levels of happiness among different racial groups. It is noteworthy that the majority of individuals in the Asian Only, Hispanic, and Other or Multiple groups reported "No Problem" with happiness, indicating higher levels of contentment within these groups.

However, there are concerns with happiness levels among individuals in the Black Only and White Only groups. The Black Only group has a notable percentage reporting a "Very Big Problem" with happiness, indicating a need for further exploration into the factors affecting happiness in this group. The White Only group also has a considerable proportion reporting "Some Problem," which could warrant attention to address their challenges with happiness. This may also be due to the sample having more white respondents than other racial groups, so it is important to also take this into account.

It is important to consider that various socio-cultural factors, life experiences, and access to resources may influence the reported levels of happiness among different racial groups. Further research and analysis may be required to delve deeper into the underlying reasons behind these findings and to identify potential interventions to improve overall well-being and happiness across all racial groups.

**Recommendations:**

1. **Diabetes Awareness and Education Campaigns:** Implement targeted awareness and education campaigns on diabetes prevention, management, and risk factors, especially for racial groups with higher diabetes prevalence. These campaigns can focus on lifestyle modifications, healthy eating habits, physical activity, and regular health check-ups.
2. **Cultural Sensitivity in Healthcare:** Enhance cultural sensitivity and diversity training for healthcare providers to address healthcare disparities and provide better care for patients from diverse racial backgrounds. Understanding cultural influences on health behaviors can help tailor healthcare services to meet the needs of specific communities.
3. **Diabetes Screening and Early Detection:** Encourage regular diabetes screening and early detection initiatives, particularly for racial groups with higher diabetes prevalence. Early diagnosis can lead to better management and improved health outcomes.
4. **Targeted Interventions for High Cholesterol:** Develop interventions targeting high cholesterol in racial groups with elevated percentages of reported cases. These interventions can include educational programs on heart-healthy diets, physical activity, and cholesterol management.
5. **Promote Higher Education and Health Literacy:** Enhance efforts to promote higher education and health literacy, as higher education levels appear to correlate with lower diabetes prevalence. Implementing programs that improve health literacy can empower individuals to make informed health decisions.
6. **Regional Health Initiatives:** Develop region-specific health initiatives to address regional disparities in diabetes prevalence and average age of diagnosis. These initiatives should consider cultural, environmental, and socio-economic factors influencing health outcomes in each region.
7. **Mental Health Support for Vulnerable Groups:** Given the varying levels of happiness across racial groups, provide mental health support and resources, especially for those reporting significant challenges with happiness, such as the Black Only group.
8. **Collaboration and Further Research:** Encourage collaboration among researchers, healthcare professionals, and policymakers to conduct further research on healthcare disparities and factors affecting health outcomes across racial groups. This collaborative effort can lead to data-driven policy decisions and targeted interventions.

**Conclusion:**

The exploratory data analysis of the "MEPS HC-228 2021 Full Year Population Characteristics" dataset provides valuable insights into the relationship between diabetes and key demographic factors among diverse racial groups in the United States. The findings shed light on healthcare disparities, regional variations, and the impact of education and culture on diabetes prevalence.

The data reveals that Non-Hispanic Black individuals have the highest diabetes prevalence, while Non-Hispanic White and Asian individuals tend to receive diabetes diagnoses at a later age. Hispanics and individuals from other racial groups show diabetes prevalence rates between those of Non-Hispanic Black and Non-Hispanic White/Asian groups.

Education plays a significant role, as lower educational attainment is associated with higher diabetes prevalence, while higher education levels appear to be protective against diabetes risk. Targeted interventions focusing on diabetes prevention, management, and lifestyle modifications can be beneficial for individuals with lower education levels.

The analysis also highlights regional disparities in diabetes prevalence and age of diagnosis, indicating the need for targeted public health efforts in specific areas. Additionally, exploring the relationship between diabetes and happiness levels across racial groups emphasizes the importance of addressing mental health aspects in healthcare interventions.

To improve healthcare outcomes and reduce diabetes risk for diverse communities, it is essential to implement targeted interventions based on the unique needs and challenges faced by each racial group. These interventions should consider cultural influences, socioeconomic factors, and health disparities.

While the findings from this exploratory analysis provide valuable insights, further research is needed to delve deeper into the underlying reasons behind the observed trends. Combining data analysis with qualitative research and considering additional variables like lifestyle choices, genetic factors, and access to healthcare can lead to more comprehensive understandings of the relationship between demographics and diabetes prevalence.

Overall, this case study underscores the significance of data-driven approaches in understanding healthcare disparities and guiding evidence-based interventions. Collaboration among researchers, healthcare professionals, policymakers, and community stakeholders is vital in implementing effective strategies to improve diabetes management, reduce disparities, and promote overall health and well-being for all individuals, regardless of their racial background.