# Project Report: Exploring Health Inequalities in Heart Disease Risk

\*\*Date:\*\* [Insert Today’s Date]

## 1. Defined Project Title (High-Impact Focus)

\*\*“Uncovering Health Inequalities: How Access to Care and Lifestyle Choices Shape Heart Disease Risk”\*\*

Why this works:  
- \*\*“Uncovering”\*\* signals a \*\*deep, investigative\*\* approach.  
- Highlights \*\*Access to Care\*\* and \*\*Lifestyle Choices\*\*—two \*\*policy-relevant\*\* and \*\*actionable\*\* factors.  
- Positions you as someone who thinks \*\*beyond basic prediction models\*\* and \*\*analyzes real-world inequities\*\*.

## 2. Analysis Framework

We will examine how \*\*healthcare access\*\* and \*\*lifestyle inequalities\*\* contribute to \*\*disparities in heart disease risk\*\*.  
  
\*\*Two Core Analytical Pillars:\*\*  
  
✅ \*\*Access to Care Inequalities\*\*  
- Whether \*\*lack of healthcare access\*\* increases heart disease risk.  
- Impact of \*\*delayed diagnosis\*\* due to \*\*cost barriers\*\*.  
  
✅ \*\*Lifestyle Inequalities\*\*  
- How \*\*smoking\*\*, \*\*alcohol consumption\*\*, and \*\*physical inactivity\*\* amplify risk—especially in vulnerable groups.  
- Investigate whether \*\*healthy behaviors\*\* are \*\*unequally distributed\*\* across socioeconomic groups.

## 3. Objective of the Project

\*\*Primary Goal:\*\*  
To analyze the effect of \*\*healthcare access\*\* and \*\*lifestyle factors\*\* on \*\*heart disease risk\*\* and quantify how \*\*inequalities\*\* in these domains affect patient outcomes.  
  
\*\*Secondary Goals:\*\*  
1. Identify \*\*which factors\*\* (e.g., access to care vs. lifestyle) contribute the most to \*\*heart disease disparities\*\*.  
2. Provide \*\*data-driven recommendations\*\* to improve \*\*preventive care\*\* and \*\*reduce health inequalities\*\*.  
3. Showcase \*\*advanced data analytics\*\* and \*\*health inequality modeling\*\* on your resume.

## 4. Methods, Skills, and Tools

✅ \*\*Methods:\*\*  
- \*\*Exploratory Data Analysis (EDA)\*\* – Identify patterns and trends in access and lifestyle variables.  
- \*\*Statistical Testing\*\* – Perform \*\*chi-square\*\* and \*\*logistic regression\*\* to quantify inequalities.  
- \*\*Predictive Modeling\*\* – Build a \*\*logistic regression model\*\* to evaluate \*\*risk\*\* based on these factors.  
- \*\*Inequality Index Creation\*\* – Quantify the degree of disparity across population groups.  
  
✅ \*\*Skills:\*\*  
- Data Cleaning & Feature Engineering  
- Statistical Analysis (Chi-Square, Correlation)  
- Predictive Modeling (Logistic Regression)  
- Data Visualization (Bar Plots, Correlation Heatmaps)  
- Storytelling with Data  
  
✅ \*\*Tools:\*\*  
- \*\*RStudio\*\* – For data processing, visualization, and modeling.  
- \*\*Excel (optional)\*\* – Quick visual sanity checks.  
- \*\*GitHub\*\* – To showcase code and project documentation.

## 5. Breakdown of Steps and Substeps

\*\*Phase 1: Project Setup & Data Review (1 day)\*\*  
- Define target variables (Heart Disease, Access, Lifestyle).  
- Confirm data integrity (check for missing values, distribution issues).  
  
\*\*Phase 2: Access to Care Analysis (3 days)\*\*  
- Compare \*\*heart disease risk\*\* for patients with/without \*\*healthcare access\*\*.  
- Statistical Test: \*\*Chi-Square\*\* (to assess significance).  
- Visual: \*\*Bar chart\*\* of heart disease prevalence by access status.  
  
\*\*Phase 3: Lifestyle Inequalities Analysis (3 days)\*\*  
- Evaluate how lifestyle factors impact heart disease:  
 - Physical Activity (PhysActivity)  
 - Smoking (Smoker)  
 - Heavy Alcohol Consumption (HvyAlcoholConsump)  
- Statistical Tests: Correlation and \*\*Logistic Regression\*\*.  
- Visual: Compare \*\*risk factors\*\* across groups (e.g., active vs. inactive).  
  
\*\*Phase 4: Model Building & Inequality Index (4 days)\*\*  
- Build a \*\*logistic regression model\*\* using both \*\*access\*\* and \*\*lifestyle\*\* factors.  
- Create an \*\*Inequality Impact Index\*\* (e.g., % risk increase due to lack of access).  
  
\*\*Phase 5: Interpretation & Reporting (2 days)\*\*  
- Highlight \*\*policy insights\*\* from your findings.  
- Prepare a \*\*resume-ready summary\*\* and \*\*presentation slides\*\*.

## 6. Estimated Timeline

|  |  |  |
| --- | --- | --- |
| Phase | Time Commitment | Target Completion Date |
| Project Setup & Data Review | 1 day | [Insert Date + 1 Day] |
| Access to Care Analysis | 3 days | [Insert Date + 4 Days] |
| Lifestyle Inequalities Analysis | 3 days | [Insert Date + 7 Days] |
| Model Building & Inequality Index | 4 days | [Insert Date + 11 Days] |
| Interpretation & Reporting | 2 days | [Insert Date + 13 Days] |
| \*\*Total Time\*\* | \*\*~2 weeks\*\* | \*\*Resume-Ready by Mid-April\*\* |

## 7. Total Time Commitment

⏳ \*\*~2 Weeks\*\* (10-13 days of work)  
- Aim to \*\*complete by mid-April\*\* to showcase on your resume while continuing your job hunt

03.25.2025

**1. Data Cleaning Progress**

You’ve successfully completed the following data cleaning tasks:

|  |  |  |
| --- | --- | --- |
| **Column** | **Issue** | **Status** |
| HeartDiseaseorAttack | Unexpected value (0.1032...) | ✅ Fixed (0 or 1 only) |
| Disease.percentage | Unexpected value (0.1032...) | ✅ Fixed (0 or 1 only) |
| **Other Columns** | Checked for missing values (NAs) | ✅ No missing values |

✅ **Outcome:** The dataset is now **clean**—all erroneous entries are fixed, and there are **no missing values** across columns.

**Quick Data Check Summary**

|  |  |
| --- | --- |
| **Variable** | **✅ Fixed Output** |
| **AnyHealthcare** | 0 (12,391), 1 (241,289) |
| **NoDocbcCost** | 0 (232,354), 1 (21,326) |
| **PhysActivity** | 0 (61,270), 1 (192,410) |
| **Smoker** | 0 (146,680), 1 (107,000) |
| **HvyAlcoholConsump** | 0 (239,730), 1 (13,950) |
| **Fruits** | 0 (88,933), 1 (164,747) |
| **Veggies** | 0 (47,148), 1 (206,532) |
| **BMI.capped** | Valid BMI range (18–50) |
| **Income** | Valid range (1–8) |

2. Exploratory data analysis:

Goal- To understand the **distribution** of key variables and **detect patterns** in heart disease risk related to **access to care** and **lifestyle factors**.

**📝 Summary of Findings: Investigating the Heart Disease-Anomaly in Heavy Alcohol Consumers**

**1. The Anomaly**

We observed that **non-drinkers** have a **higher percentage of heart disease (9.54%)** compared to **heavy drinkers (6.08%)**, which contradicts the expected pattern where heavy alcohol consumption is typically associated with increased health risks.

**2. Key Insights from Our Investigation**

✅ **(a) Age Distribution Discrepancy**

* **Mean Age of Non-Drinkers**: 8.12 (encoded value)
* **Mean Age of Heavy Drinkers**: 7.59 (encoded value)

📊 **Interpretation**:  
These unrealistic age values indicate that the **age variable** is encoded and not yet properly decoded into actual age groups. This discrepancy suggests that the **heavy drinker group** might be **younger** on average—potentially reducing their heart disease risk.

✅ **(b) High Blood Pressure (BP) Status**

* **Non-drinkers with High BP**: 41.1%
* **Heavy drinkers with High BP**: 42.7%

📊 **Interpretation**:  
While heavy drinkers show **slightly higher hypertension rates**, the difference is small (~1.6%), meaning **hypertension alone does not explain** the lower heart disease prevalence in heavy drinkers.

✅ **(c) Possible Data Quality or Sampling Bias**

1. **Healthier Drinker Bias**: It is possible that individuals categorized as **heavy drinkers** in this dataset may be **younger and healthier**, which can reduce heart disease risk.
2. **Survivorship Bias**: Individuals with severe health complications from heavy drinking **may not be included** in the sample due to poor health or death.
3. **Underreporting**: Heavy drinkers might **underreport** both their alcohol consumption and underlying health conditions.

**📌 Preliminary Explanation of the Anomaly**

The lower heart disease rate among heavy drinkers appears to be **partially explained by the following**:

1. **Age Confounding**: The heavy drinking group is likely **younger**, reducing their heart disease risk.
2. **Healthier Drinker Bias**: Healthier individuals may tolerate alcohol better, skewing the data.
3. **Data Encoding Errors**: Misinterpreted age categories may mask the true risk.

**🔍 Next Steps for a Deeper Analysis**

1. **Decode the Age Variable** – Ensure accurate age representation for a proper comparison.
2. **Check Additional Risk Factors** – Explore differences in **diabetes**, **smoking**, and **cholesterol** across groups.
3. **Logistic Regression** – Build a model to **isolate the effect** of alcohol consumption while controlling for other factors.

Would you like to proceed with age decoding or move toward advanced statistical modeling?

**Current Progress vs. Planned Timeline:**

|  |  |  |
| --- | --- | --- |
| **Task** | **Planned Date** | **Actual Status** |
| ✅ Data Cleaning | Completed by Mar 22 | ✅ Finished on time |
| ✅ Lifestyle Inequalities (EDA) | Mar 25 – Mar 30 | 🟡 75% complete (ahead of schedule) |
| 🔜 Access to Care Analysis | Mar 31 – Apr 3 | Starting next (ahead of time) |
| 🔜 Statistical Modeling | Apr 4 – Apr 8 | Yet to start |
| 🔜 Final Report Draft | Apr 9 – Apr 12 | Targeting Apr 1 start (ahead) |

04.07.2025,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Step** | **Task** | **Status** | **Completion Date** | **Notes** |
| 1 | Define objectives & research focus | ✅ Complete | March 25, 2025 | Focus areas: Lifestyle + Access to Care |
| 2 | Data cleaning & preprocessing | ✅ Complete | March 28, 2025 | Cleaned data, BMI & age categories added |
| 3 | Exploratory Data Analysis (EDA) | ✅ Complete | March 30, 2025 | Used dplyr, ggplot2 for summary & visuals |
| 4 | Analyze Lifestyle Inequalities | ✅ Complete | March 31, 2025 | Explored smoking, alcohol, physical activity, etc. |
| 5 | Analyze Access to Care Inequalities | ✅ Complete | April 6, 2025 | Plots generated and saved with readable axis labels |
| 6 | Derive insights & patterns | 🔄 In Progress | -- | You’ll begin interpreting the trends next |
| 7 | Draft final report | ⏳ Not Started | -- | Structure ready — we can begin this now |
| 8 | Optional modeling | ❌ Not Started | -- | Can be added for portfolio enhancement |
| 9 | Final presentation/resume-ready summary | ⏳ Not Started | -- | To be created once insights/report are ready |

**📈 Insights & Interpretation**

**1. AnyHealthcare**

* **X-axis categories**:  
  0 = No, 1 = Yes
* **Observation**:  
  Individuals without any healthcare coverage (0) show a **higher rate of heart disease** compared to those with coverage.
* **Insight**:  
  **Lack of health insurance is associated with a greater risk of heart disease**, likely due to reduced access to preventive care.

**2. NoDocbcCost**

* **X-axis categories**:  
  0 = No (could see a doctor), 1 = Yes (couldn’t afford doctor visit)
* **Observation**:  
  People who reported cost as a barrier (1) have a **higher heart disease rate**.
* **Insight**:  
  **Cost-related barriers are strongly associated with poorer heart health**, underlining the role of financial access.

**3. CholCheck**

* **X-axis categories**:  
  0 = No, 1 = Yes
* **Observation**:  
  Those who did not check cholesterol (0) had slightly **higher heart disease rates**.
* **Insight**:  
  **Routine cholesterol screening is linked with reduced risk**, potentially due to early intervention.

**4. Education**

* **X-axis categories**:  
  1 = Never attended school, 2 = Grades 1–8, 3 = Grades 9–11,  
  4 = Grade 12 or GED, 5 = Some college/AA, 6 = College graduate
* **Observation**:  
  Heart disease rates **decrease with higher education levels**.
* **Insight**:  
  **Education strongly influences health awareness, lifestyle choices, and healthcare access**, thus impacting heart health.

**5. Income**

* **X-axis categories**:  
  1 = <$10,000, 2 = $10,000–15,000, … up to 8 = $75,000+
* **Observation**:  
  Individuals in the **lowest income groups (1–3)** have significantly **higher heart disease rates** than those in groups 6–8.
* **Insight**:  
  **Socioeconomic status is a critical determinant**. Lower income is linked to higher disease risk due to lifestyle, stress, and limited access to care.

**✅ Summary of Step 6:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Risk Pattern** | **Key Takeaway** |
| AnyHealthcare | No insurance → higher risk | Insurance improves preventive care access |
| NoDocbcCost | Cost barrier → higher risk | Affordability affects treatment-seeking |
| CholCheck | No screening → slightly higher risk | Preventive checks reduce risk |
| Education | Lower education → higher risk | Awareness & literacy matter |
| Income | Lower income → higher risk | Socioeconomic support is key |

**🔹 Smoker vs Non-Smoker**

* **Insight:** Smokers show a consistently higher percentage of heart disease (~13%) compared to non-smokers (~9%).
* **Interpretation:** Smoking significantly raises cardiovascular risk — consistent with known medical literature.
* **Plot Placement:** Lifestyle Inequalities > Subsection: Smoking Status

**🔹 Physical Activity**

* **Insight:** Physically inactive individuals had a heart disease rate of ~12.5%, while active individuals were closer to 8.5%.
* **Interpretation:** Lack of exercise is a strong risk factor, possibly linked to obesity, poor circulation, or associated habits (like sedentary jobs or comorbidities).
* **Plot Placement:** Lifestyle Inequalities > Subsection: Physical Activity

**🔹 Alcohol Consumption (HvyAlcoholConsump)**

* **Insight:** Heavy alcohol consumers (defined by dataset's threshold) had a marginally higher rate of heart disease.
* **Interpretation:** While not as dramatic as smoking or exercise, the data suggests a trend where excessive alcohol may correlate with other unhealthy behaviors.
* **Plot Placement:** Lifestyle Inequalities > Subsection: Alcohol Consumption

**🔹 Diet: Fruits and Vegetables**

* **Insight:** People who consume fruits and vegetables regularly show slightly lower risk (~8.2% vs ~10.4%).
* **Interpretation:** Diet appears to be a preventive factor — but likely plays a more significant role in long-term outcomes or when combined with other behaviors.
* **Plot Placement:** Lifestyle Inequalities > Subsection: Diet Quality

**🔹 Access to Care: AnyHealthcare, NoDocbcCost**

* **Insight:** People without healthcare access or those who avoided seeing a doctor due to cost had ~12%+ risk, vs ~9% among those with access.
* **Interpretation:** Access to preventive care and early intervention could play a major role in heart disease outcomes.
* **Plot Placement:** Access Inequalities > Subsection: Healthcare Access

**🔹 Education & Income**

* **Insight:** Lower income and education levels are strongly correlated with higher heart disease percentages.
* **Interpretation:** Social determinants of health like education and income show clear disparity — possibly reflecting awareness, lifestyle, and access.
* **Plot Placement:** Access Inequalities > Subsection: Socioeconomic Factors