# Youth Risk Behavior Surveillance System (YRBSS) Analysis Report: Middle School Risk Trends

---

Analyst: Philip Aklorbortu

Date: August 28, 2025

Table of contents

# Executive Summary

This report analyzes Youth Risk Behavior Surveillance System (YRBSS) data on U.S. middle school students, focusing on trends in Physical Activity, Mental Health, Substance Use, and Violence. Key findings show that physical inactivity is more prevalent among females in 7th and 8th grade, mental health risks—especially among females—are rising, and about 9% of students have tried or currently use cigarettes, e-cigarettes, or marijuana. Bullying affects over 40% of students, while weapon carrying has declined but rose slightly in 2017. Males report higher involvement in fights and weapon carrying, while Native Hawaiian or Other Pacific Islander students show the highest inactivity levels. The report recommends expanding Physical Activity programs, improving Mental Health support, launching anti-vaping campaigns, and strengthening digital safety education to reduce risks and improve student well-being.

# 1. Introduction

The Youth Risk Behavior Surveillance System (YRBSS) monitors health-risk behaviors among youth that contribute to the leading causes of death and disability. This project analyzes YRBSS data for middle school students to explore trends in risky behaviors, Physical Activity, Mental Health, Substance Use, and Violence. The data was sourced from the CDC's public datasets and analyzed using Google BigQuery.

# 2. Project Objectives

- Identify key risk behavior trends over the past decade.
- Analyze behavioral differences by Sex, Race, and Grade.
- Determine which risky behaviors are increasing most.
- Support data-driven decision-making in youth health programs.

# 3. Dataset Overview

The dataset includes responses from middle school students across the United States. Each record captures demographic details (sex, race, grade), survey year, and responses to health-related questions categorized into topics such as Violence, Mental Health, Substance Use, and Physical Activity.

| Attribute | Description |
| --- | --- |

| | |
|---|---|
| Data source | Centers for Disease Control and Prevention (CDC) |
| Data format | CSV |
| Number of records | 751K rows |
| Number of variables | 35 columns |
| Period covered | 1991-2017 |

## Key Variables

| **Column Name** | **Description** | **Data Type** |
|---|---|---|
| YEAR | Year of survey completion | Number |
| LocationAbbr | Location abbreviation | Text |
| LocationDesc | Location description | Text |
| DataSource | Abbreviation of data source | Text |
| Topic | Topic (Physical activity, alcohol/drug use, etc.) | Text |
| Subtopic | Subtopic(dependent on the topic) | Text |
| ShortQuestionText | Shortened text of full question | Text |
| Greater_Risk_Question | Greater risk version of the question | Text |
| Description | Additional details regarding the question | Text |
| Data_Value_Symbol | Description of data (e.g. %, $, etc.) | Text |
| Data_Value_Type | Type of data (percentage, count, dollar, etc.) | Text |
| Greater_Risk_Data_Value | Numerical data value pertaining to the greater risk question | Number |

| | | |
|---|---|---|
| Greater_Risk_Data_Value_Footnote_Symbol | Symbol that would be used to flag footnotes for the greater risk question data value | Text |
| Greater_Risk_Data_Value_Footnote | Footnotes for the greater risk question data value | Text |
| Greater_Risk_Low_Confidence_Limit | Low confidence interval limit for the greater risk data value | Number |
| Greater_Risk_High_Confidence_Limit | High confidence interval limit for the greater risk data value | Number |
| Lesser_Risk_Question | Lesser risk version of the question | Text |
| Lesser_Risk_Data_Value | Numerical data value pertaining to the lower risk question | Number |
| Lesser_Risk_Data_Value_Footnote_Symbol | Symbol that would be used to flag footnotes for the lower risk question data value | Text |
| Lesser_Risk_Data_Value_Footnote | Footnotes for the lower risk question data value | Text |
| Lesser_Risk_Low_Confidence_Limit | Low confidence interval limit for the lower risk data value | Number |
| Lesser_Risk_High_Confidence_Limit | High confidence interval limit for the lower risk data value | Number |
| Sample_Size | Sample size for the question | Number |
| Sex | Gender of the respondent | Text |
| Race | Race of the respondent | Text |
| Grade | Grade of the respondent | Text |
| GeoLocation | Latitude & Longitude to be provided for formatting GeoLocation or Geocode in the order (latitude, longitude) | Location |
| TopicId | Lookup identifier value for | Text |

| | | |
|---|---|---|
| | Topic; can be used for filtering the dataset | |
| SubTopicID | Lookup identifier value for subtopic; can be used for filtering the dataset | Text |
| QuestionCode | Lookup identifier value for question; can be used for filtering the dataset | Text |
| LocationId | Lookup identifier value for location; can be used for filtering the dataset | Text |
| StratID1 | Lookup identifier value for sex; can be used for filtering the dataset | Text |
| StratID2 | Lookup identifier value for race; can be used for filtering the dataset | Text |
| StratID3 | Lookup identifier value for grade; can be used for filtering the dataset | Text |
| StratificationType | Indicator for type of location; State, Territory, Locality, Other | Text |

# 4. Methodology

The dataset was uploaded into Google BigQuery. SQL queries extracted trends, calculated averages, and filtered questions by risk categories. Behaviors were grouped by topics such as Mental Health, Violence, Substance Use, and Physical Activity. Aggregations such as averages, counts, and percentages were used to summarize the data across years and demographics.

# 5. Key Questions & Findings

**Getting Familiar With the Dataset**

Before running queries, an initial review of rows and columns was performed to understand the dataset and facilitate filtering. This included retrieving distinct values for Topic, Subtopic, QuestionCode, Sex, Race, Grade, and Location to build the foundation for deeper analysis.

**Getting an overview of the dataset**

This first query selects all the columns and limits the rows to the first 10

```sql
--Overview of dataset
SELECT *
FROM `yrbss-1.yrbss_dataset.yrbss_table`
LIMIT 10;
```

**Finding how many topics and its ID**

This query outputs distinct Topic and its corresponding TopicID

```sql
--Finding how many topics and its ID
SELECT DISTINCT Topic, TopicID
FROM `yrbss-1.yrbss_dataset.yrbss_table`
```

**Finding subtopic and its ID**

```sql
--Finding subtopic and its ID
SELECT DISTINCT Subtopic, SubTopicID
FROM `yrbss-1.yrbss_dataset.yrbss_table`
```

**Finding ShortQuestionText and its QuestionCode**

```
--Finding ShortQuestionText and its QuestionCode
SELECT DISTINCT ShortQuestionText, QuestionCode
FROM `yrbss-1.yrbss_dataset.yrbss_table`
```

**Finding Sex and its StratID1**

```
--Finding Sex and its SratID1
SELECT DISTINCT Sex, StratID1
FROM `yrbss-1.yrbss_dataset.yrbss_table`

```

**Finding Race and its StratID2**

```
--Finding Race and its StratID2
SELECT DISTINCT Race, StratID2
FROM `yrbss-1.yrbss_dataset.yrbss_table`
```

**Finding Grade and its StratID3**

```
--Finding Grade and its StratID3
SELECT DISTINCT Grade, StratID3
FROM `yrbss-1.yrbss_dataset.yrbss_table`
```

**Finding LocationDesc and its LocationID**

```
--Finding LocationDesc and its LocationID
SELECT LocationDesc, LocationId
FROM `yrbss-1.yrbss_dataset.yrbss_table`
```

With the filtered values and dataset overview established, we can now proceed to query for insights.

## 5.1 Physical Activity Trends

1. **Which grades or genders are least active on average?**

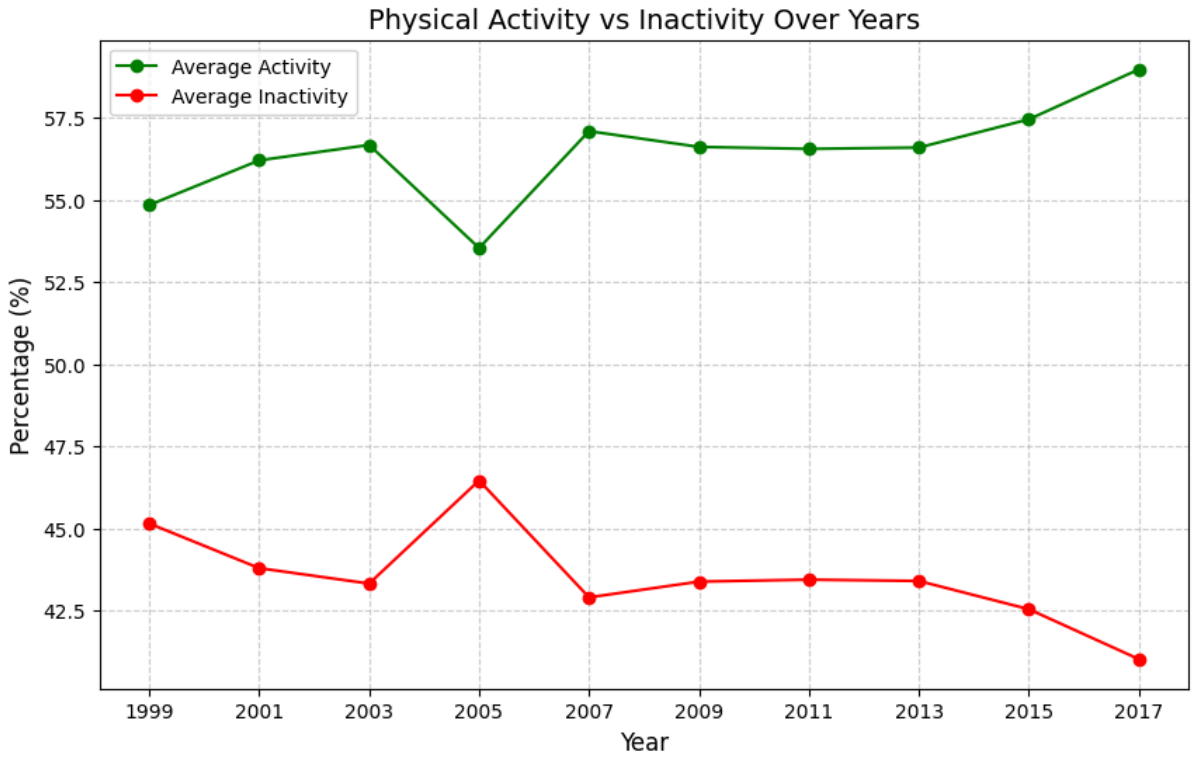
```sql
--Which grades or genders are least active on average?
SELECT
  Grade, Sex,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_inactivity
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE TopicId = 'C06'
  AND Greater_Risk_Data_Value IS NOT NULL
GROUP BY Grade, Sex
ORDER BY avg_inactivity DESC;

```

The query selected Grade and Sex, while creating a new column (avg_inactivity) as the average of Greater_Risk_Data_Value rounded to two decimals. We used Greater_Risk_Data_Value because it better reflects negative outcomes in this dataset.

2. **Has physical activity increased or declined over the years?**

```sql
--Has physical activity increased or declined over the years?
SELECT
  year,
  ROUND(AVG(Lesser_Risk_Data_Value), 2) AS avg_activity,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_inactivity
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE TopicId = 'C06'
  AND Greater_Risk_Data_Value IS NOT NULL
  AND Lesser_Risk_Data_Value IS NOT NULL
GROUP BY year
ORDER BY year;

```

The query results show that average activity increased over the years, while inactivity declined. Results were ordered in descending order, as higher Greater_Risk_Data_Values indicate greater inactivity. The results are graphically represented below.




**3. Which state or region reports the lowest average activity levels?**

```sql
--Which state or region reports the lowest average activity levels?
SELECT LocationDesc,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_inactivity
FROM yrbss-1.yrbss_dataset.yrbss_table
WHERE TopicId = 'C06' AND
  Greater_Risk_Data_Value IS NOT NULL
GROUP BY LocationDesc
ORDER BY avg_inactivity DESC;
```

We narrow down our column selection to LocationDesc and our own new column which calculates the rounded average of the Greater_Risk_Data_Value under the label avg_inactivity. We maintain the previous filtering clauses since they all pertain to the same subject. Grouping is done by LocationDesc and ordering in DESC. The ordering is done in DESC due to the implication of the Greater_Risk_Data_Value. The higher the value, the less active the student is. In other words, the higher the value, the more inactive the student is.

## 5.2 Mental Health

1. **What percentage of students report feeling sad or hopeless?**

```sql
--What percentage of students report feeling sad or hopeless?
SELECT
   ROUND(AVG(Greater_Risk_Data_Value), 2) AS percent_of_hopeless_or_sad
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE QuestionCode IN ('M12', 'M13', 'M14', 'M15', 'M16')
   AND Greater_Risk_Data_Value IS NOT NULL;
```

Finding the rounded average of the Greater_Risk_Data_Value represents the total number of students who feel hopeless or sad. The QuestionCode filters out the questions to only those that suggest the feeling of hopelessness. Indicators such as suicidal intent (M15) and bullying (M12) suggest that the student may be experiencing sadness or hopelessness. However, classification of sadness or hopelessness depends on interpretation, as several questions may fit.

2. **Is there a correlation between mental health indicators and gender?**

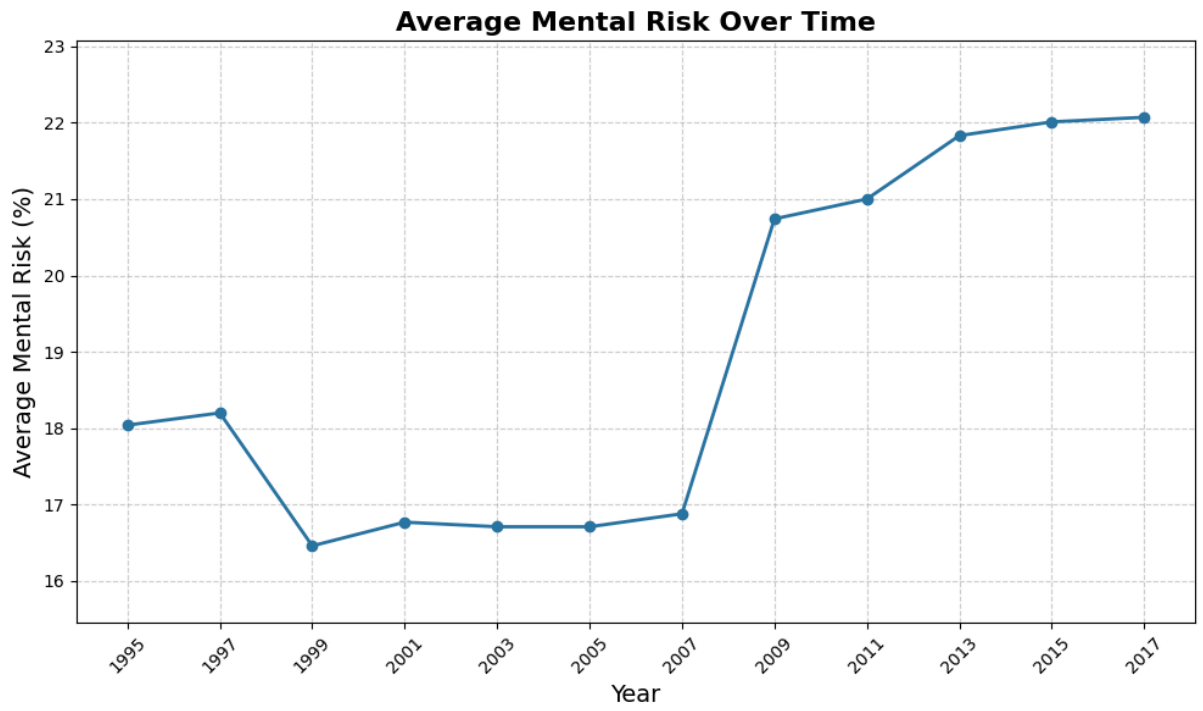
```sql
--Is there a correlation between mental health indicators and gender?
SELECT
  CORR(
    CASE
      WHEN sex = 'Male' THEN 0
      WHEN sex = 'Female' THEN 1
    END,
    Greater_Risk_Data_Value
  ) AS correlation
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE
  QuestionCode IN ('M12', 'M13', 'M14', 'M15', 'M16')
  AND Greater_Risk_Data_Value IS NOT NULL
  AND sex IN ('Male', 'Female');

```

This query outputs a correlation coefficient of 0.34. There is a moderate positive relationship between gender (coded as 0 = male, 1 = female) and the mental health risk value. This means, the sex variable moves from 0 → 1 (male → female), the Greater_Risk_Data_Value tends to increase. In plain language, females tend to report higher levels of mental health risk compared to males.

3. **Are mental health concerns rising or declining over the years?**

```sql
--Are mental health concerns rising or declining over the years?
SELECT
  YEAR,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_mental_risk
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE Greater_Risk_Data_Value IS NOT NULL AND
  QuestionCode IN ('M12', 'M13', 'M14', 'M15', 'M16')
GROUP BY YEAR
ORDER BY YEAR;
```

Comparing the avg_mental_risk across the years helps identify overall trends. This can easily be interpreted by visualizing the output on a line graph as shown below. It can be inferred that over the years there has been a rise in mental health risk, peaking in 2017 after a trough between 1999 and 2007.



## 5.3 Substance Use

1. **What proportion of students have ever tried or currently use cigarettes, e-cigarettes, or marijuana?**

```sql
--What percentage of students have ever tried or currently use cigarettes,
--e-cigarettes, or marijuana?
SELECT
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS percentage_tried_substance
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE QuestionCode IN
  ('M28', 'M17', 'M19', 'QNTB2', 'M18', 'QNTB4',
  'QNFRCGR', 'M21', 'M23','M25', 'QNDAYCGR', 'QNTB3', 'M22',
  'QNFRSKL', 'QNDAYSKL', 'QNFREVP', 'QNDAYEVP')
  AND Greater_Risk_Data_Value IS NOT NULL;

```

The query showed that 8.91% of students have ever tried or currently used cigarettes, e-cigarettes, or marijuana.

2. **Which grade level is most likely to report trying a substance?**

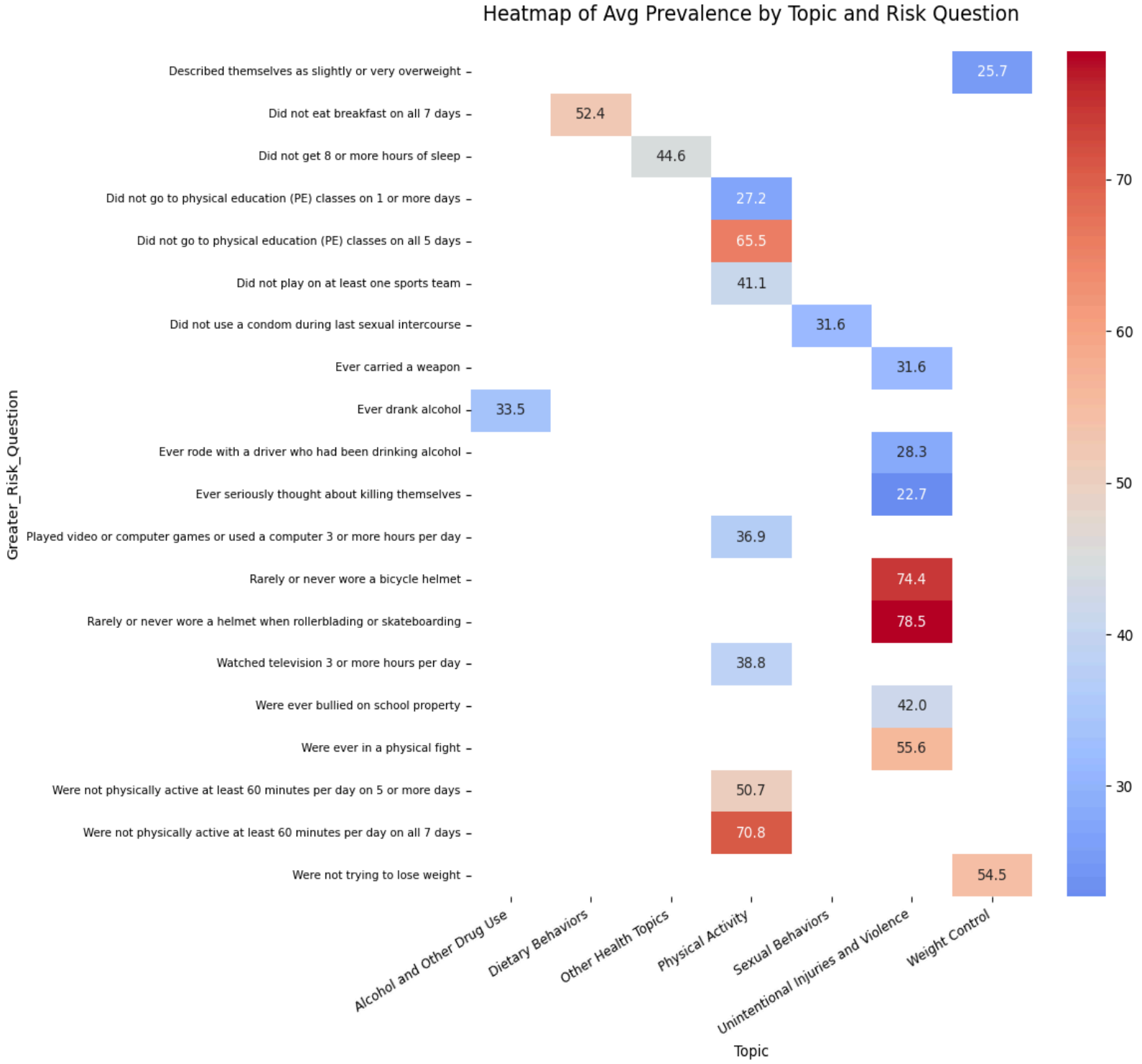
```sql
--Which grade level is most likely to report trying a substance?
SELECT
  Grade,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS percentage_tried_substance
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE QuestionCode IN
  ('M28', 'M17', 'M19', 'QNTB2', 'M18', 'QNTB4',
  'QNFRCGR', 'M21', 'M23','M25', 'QNDAYCGR', 'QNTB3', 'M22',
  'QNFRSKL', 'QNDAYSKL', 'QNFREVP', 'QNDAYEVP')
  AND Greater_Risk_Data_Value IS NOT NULL
GROUP BY Grade
ORDER BY percentage_tried_substance DESC;
```

The query showed that 8th, 7th and 6th grade have 11.82%, 8.45% and 5.39% of students respectively who have tried substances.

3. **Which risky behaviors co-occur with substance use?**

```sql
--Which risky behaviors co-occur with substance use?
WITH substance_codes AS (
  SELECT 'M28' AS code UNION ALL SELECT 'M17' UNION ALL SELECT 'M19'
  UNION ALL SELECT 'QNTB2' UNION ALL SELECT 'M18' UNION ALL SELECT 'QNTB4'
  UNION ALL SELECT 'QNFRCGR' UNION ALL SELECT 'M21' UNION ALL SELECT 'M23'
  UNION ALL SELECT 'M25' UNION ALL SELECT 'QNDAYCGR' UNION ALL SELECT 'QNTB3'
  UNION ALL SELECT 'M22' UNION ALL SELECT 'QNFRSKL' UNION ALL SELECT 'QNDAYSKL'
  UNION ALL SELECT 'QNFREVP' UNION ALL SELECT 'QNDAYEVP'
)

SELECT
  Topic,
  Greater_Risk_Question,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_prevalence
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE Greater_Risk_Data_Value IS NOT NULL
  AND CONCAT(CAST(YEAR AS STRING), '-', LocationDesc, '-', Sex, '-', Race, '-', Grade) IN (
    SELECT DISTINCT
      CONCAT(CAST(YEAR AS STRING), '-', LocationDesc, '-', Sex, '-', Race, '-', Grade)
    FROM `yrbss-1.yrbss_dataset.yrbss_table`
    WHERE QuestionCode IN (SELECT code FROM substance_codes)
      AND Greater_Risk_Data_Value IS NOT NULL
  )
  AND QuestionCode NOT IN (SELECT code FROM substance_codes)
GROUP BY Topic, Greater_Risk_Question
ORDER BY avg_prevalence DESC
LIMIT 20;
```

The analysis identified risky behaviors most often reported alongside substance use. Results highlighted overlaps with Violence, Mental Health issues, and other risky habits. Below is a heatmap to visualize the co-occurrence query


Heatmap of Avg Prevalence by Topic and Risk Question
Described themselves as slightly or very overweight
Did not eat breakfast on all 7 days
Did not get 8 or more hours of sleep
Did not go to physical education (PE) classes on 1 or more days
Did not go to physical education (PE) classes on all 5 days
Did not play on at least one sports team
Did not use a condom during last sexual intercourse
Ever carried a weapon
Ever drank alcohol
Ever rode with a driver who had been drinking alcohol
Ever seriously thought about killing themselves
Played video or computer games or used a computer 3 or more hours per day
Rarely or never wore a bicycle helmet
Rarely or never wore a helmet when rollerblading or skateboarding
Watched television 3 or more hours per day
Were ever bullied on school property
Were ever in a physical fight
Were not physically active at least 60 minutes per day on 5 or more days
Were not physically active at least 60 minutes per day on all 7 days
Were not trying to lose weight
25.7
52.4
44.6
27.2
65.5
41.1
31.6
31.6
33.5
28.3
22.7
36.9
74.4
78.5
38.8
42.0
55.6
50.7
70.8
54.5
Greater_Risk_Question
Alcohol and Other Drug Use
Dietary Behaviors
Other Health Topics
Physical Activity
Sexual Behaviors
Unintentional Injuries and Violence
Weight Control
Topic
70
60
50
40
30

## 5.4 Violence & Bullying

1. **How many students report being bullied at school or online?**

```sql
--How many students report being bullied at school or online?
SELECT
  ShortQuestionText,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS num_bullied_students
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE QuestionCode IN ('M12', 'M13')
  AND Greater_Risk_Data_Value IS NOT NULL
GROUP BY ShortQuestionText
ORDER BY num_bullied_students DESC;
```

42.04% of students are bullied in school whereas 20.09% of students report being bullied online

2. **Which groups (by race or sex) are most affected by violence?**

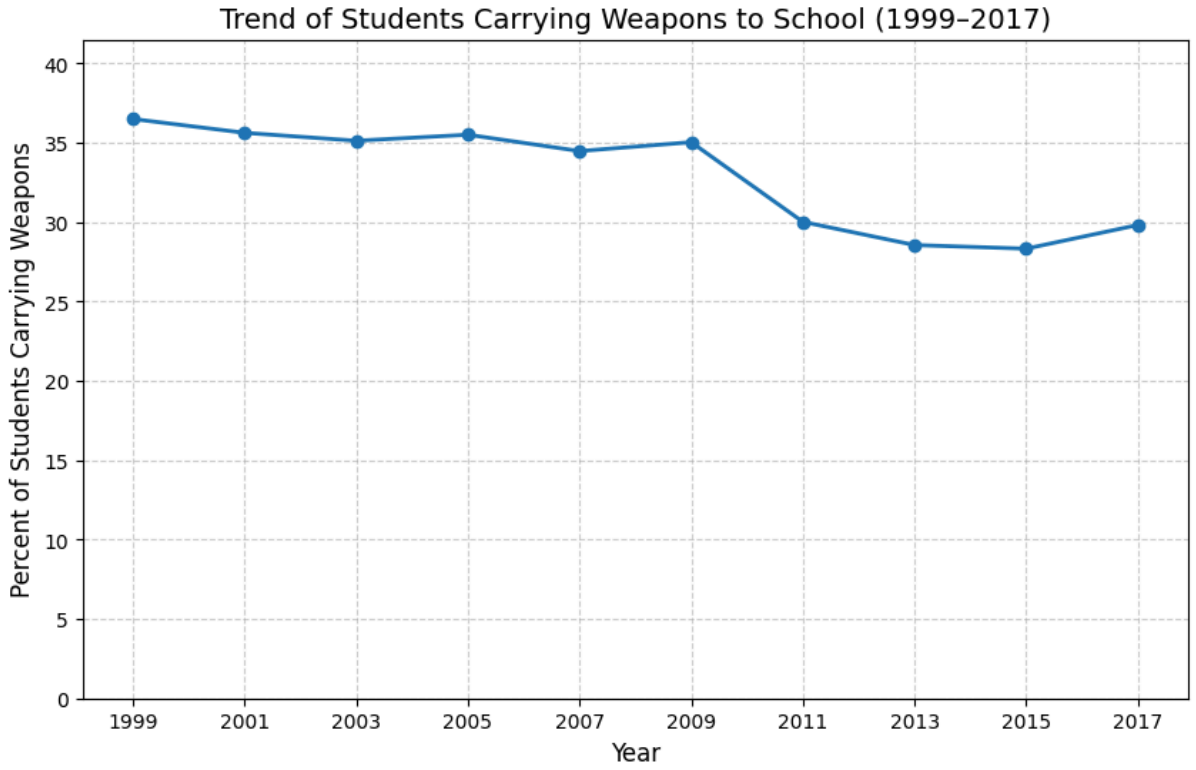
```sql
--Which groups (by race or sex) are most affected by violence?
SELECT
  Race,
  Sex,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_violence_risk
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE
  Greater_Risk_Data_Value IS NOT NULL AND
  QuestionCode IN ('M10', 'M11', 'M12', 'M13',
    'M35', 'M46')
GROUP BY Race, Sex
ORDER BY avg_violence_risk DESC;
```

Native Hawaiian or Other Pacific Islander, White and Black or African American students reported the highest average violence risk (39.71%, 38.76%, and 38.40%, respectively). Males were more frequently represented among these groups.

### 3. Has the rate of carrying weapons to school changed over time?

```sql
--Has the rate of carrying weapons to school changed over time?
SELECT
  YEAR,
  ROUND(AVG(Greater_Risk_Data_Value) , 2) AS percent_weapon_carrying_students
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE
  QuestionCode = 'M10'
GROUP BY YEAR
ORDER BY YEAR;
```

Below is a line graph visualization for the above query. Weapon carrying showed a downward trend from 1999 to 2015 (reaching 28.33%). A slight increase occurred in 2017 (29.82%), marking a 5.26% rise from 2015.



## 5.5 Demographic-Based Insights

1. **Do males engage more in physical fights or weapon carrying?**

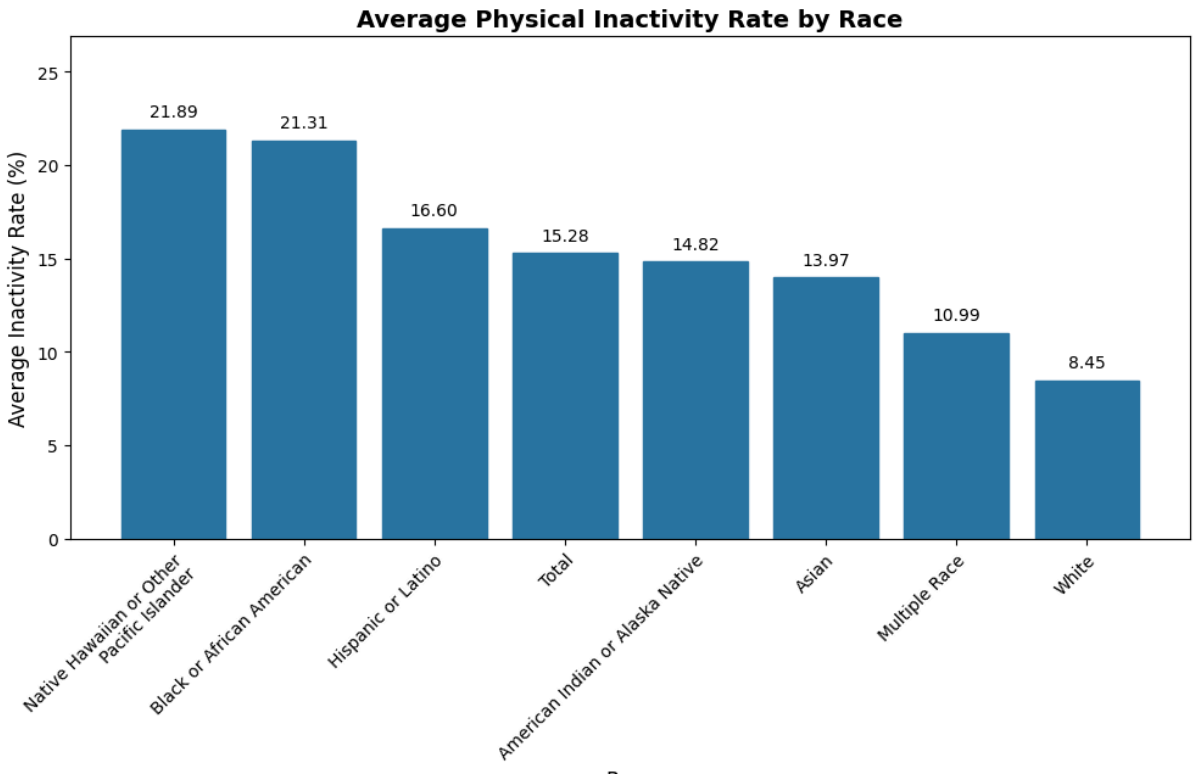
```sql
--Do males engage more in physical fights or weapon carrying?
SELECT
  Sex,
  ShortQuestionText,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_num_of_students
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE QuestionCode IN ('M11', 'M10') AND
  Greater_Risk_Data_Value IS NOT NULL
GROUP BY Sex, ShortQuestionText
ORDER BY ShortQuestionText, avg_num_of_students DESC;
```

Males engage more in physical fights and weapon carrying than females. On average, 67.57% of males reported involvement in physical fights compared to 43.51% of females. Similarly, males are more involved in weapon carrying; averaging 43.91% compared to females who average 19.73%

2. **Which racial group reports the highest level of physical inactivity?**

```sql
--Which racial group reports the highest level of physical inactivity?
SELECT
  Race,
  ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_inactive_rate
FROM `yrbss-1.yrbss_dataset.yrbss_table`
WHERE Greater_Risk_Data_Value IS NOT NULL
  AND QuestionCode = 'QNPA0DAY'
GROUP BY Race
ORDER BY avg_inactive_rate DESC;
```

The bar graph below shows that Native Hawaiian or Other Pacific Islander students report the highest inactivity, while White students report the lowest (most active).

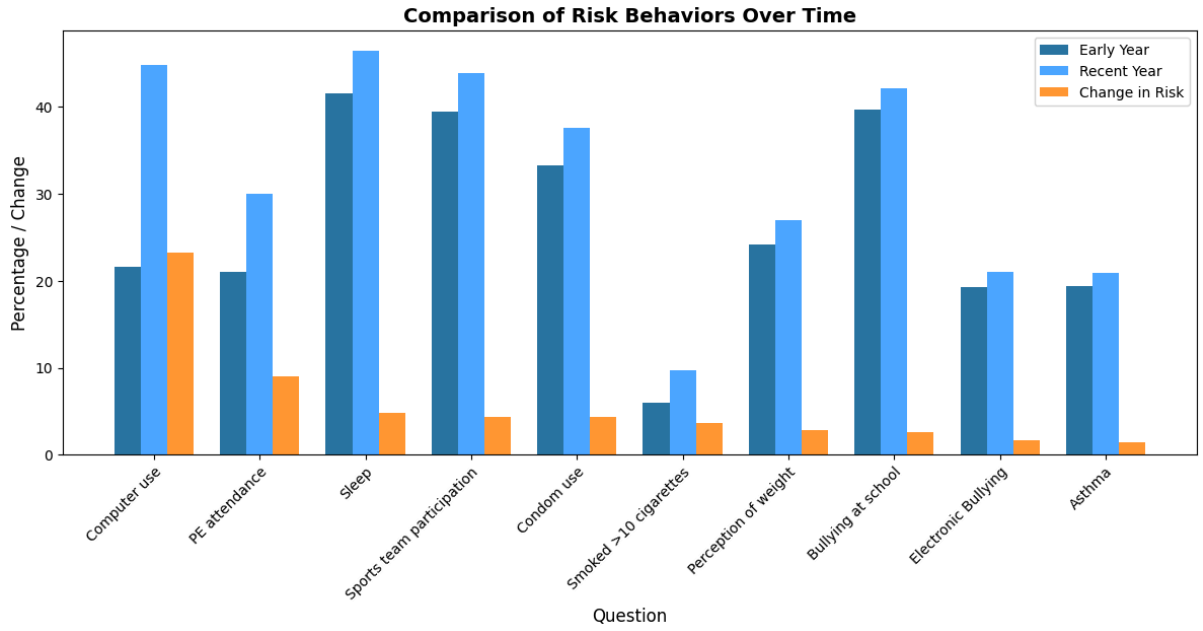

Average Physical Inactivity Rate by Race
Average Inactivity Rate (%)
0
5
10
15
20
25
21.89
21.31
16.60
15.28
14.82
13.97
10.99
8.45
Native Hawaiian or Other Pacific Islander
Black or African American
Hispanic or Latino
Total
American Indian or Alaska Native
Asian
Multiple Race
White

## 5.6 Trend & Comparative Insights

1. **Which risky behavior has increased over the past decade?**

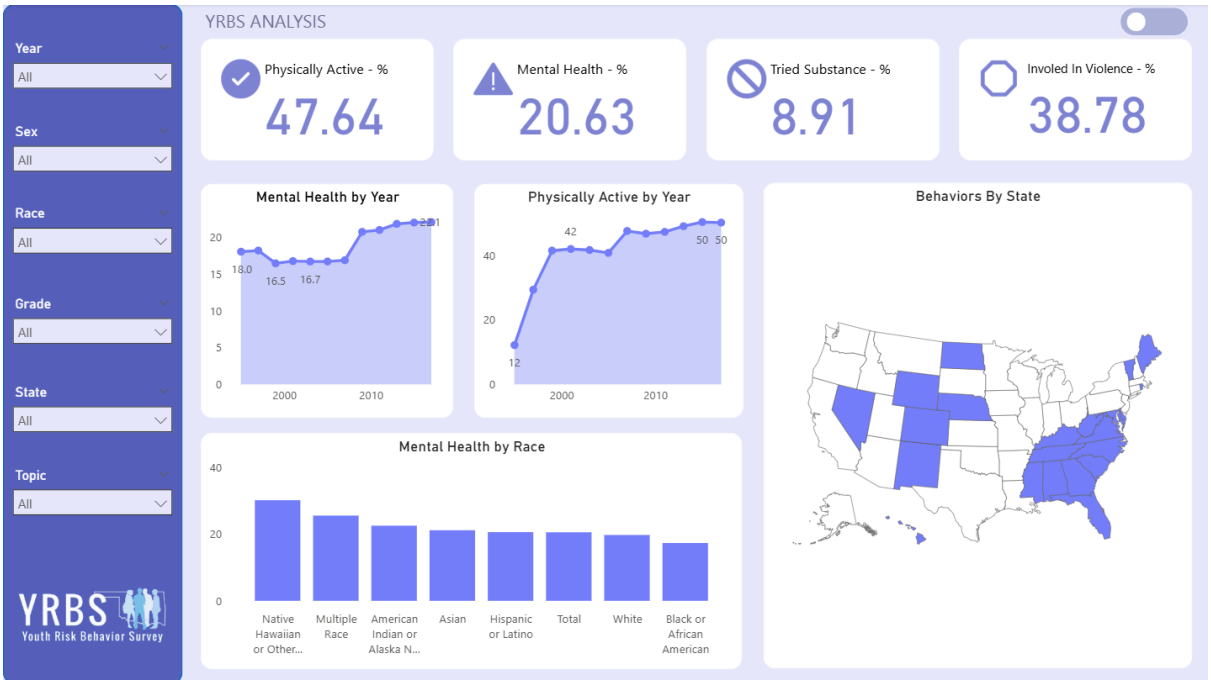
```sql
-- Which risky behavior has increased the most over the past years?
WITH risky_behaviors AS (
  SELECT
    QuestionCode,
    ShortQuestionText,
    YEAR,
    ROUND(AVG(Greater_Risk_Data_Value), 2) AS avg_risk_value
  FROM `yrbss-1.yrbss_dataset.yrbss_table`
  WHERE
    Greater_Risk_Data_Value IS NOT NULL
    AND QuestionCode IN (
      'M7', 'M9', 'M10', 'M11', 'M12', 'M13',
      'M14', 'M15', 'M16', 'M17',
      'M18', 'M19', 'M20', 'M21', 'M22', 'M23', 'QNFRSCIG', 'M24',
      'M25', 'M26', 'M27', 'M28', 'M29', 'M30', 'M31', 'M32', 'M33', 'M34',
      'M35', 'M36', 'M37', 'M38', 'M39',
      'M40', 'QNSODA', 'M41', 'M42', 'M43', 'M44',
      'QNPA0DAY', 'M45', 'M46', 'M47', 'M48', 'M49', 'M50', 'M51'
    )
  GROUP BY QuestionCode, ShortQuestionText, YEAR
),
trend_comparison AS (
  SELECT
    rb.QuestionCode,
    rb.ShortQuestionText,
    MIN(CASE WHEN YEAR = (SELECT MIN(YEAR)
                          FROM risky_behaviors rb2
                          WHERE rb2.QuestionCode = rb.QuestionCode)
             THEN avg_risk_value END) AS early_year_value,
    MAX(CASE WHEN YEAR = (SELECT MAX(YEAR)
                          FROM risky_behaviors rb2
                          WHERE rb2.QuestionCode = rb.QuestionCode)
             THEN avg_risk_value END) AS recent_year_value
  FROM risky_behaviors rb
  GROUP BY rb.QuestionCode, rb.ShortQuestionText
)
SELECT
  QuestionCode,
  ShortQuestionText,
  early_year_value,
  recent_year_value,
  ROUND(recent_year_value - early_year_value, 2) AS change_in_risk
FROM trend_comparison
WHERE early_year_value IS NOT NULL AND recent_year_value IS NOT NULL
ORDER BY change_in_risk DESC
LIMIT 10;

```

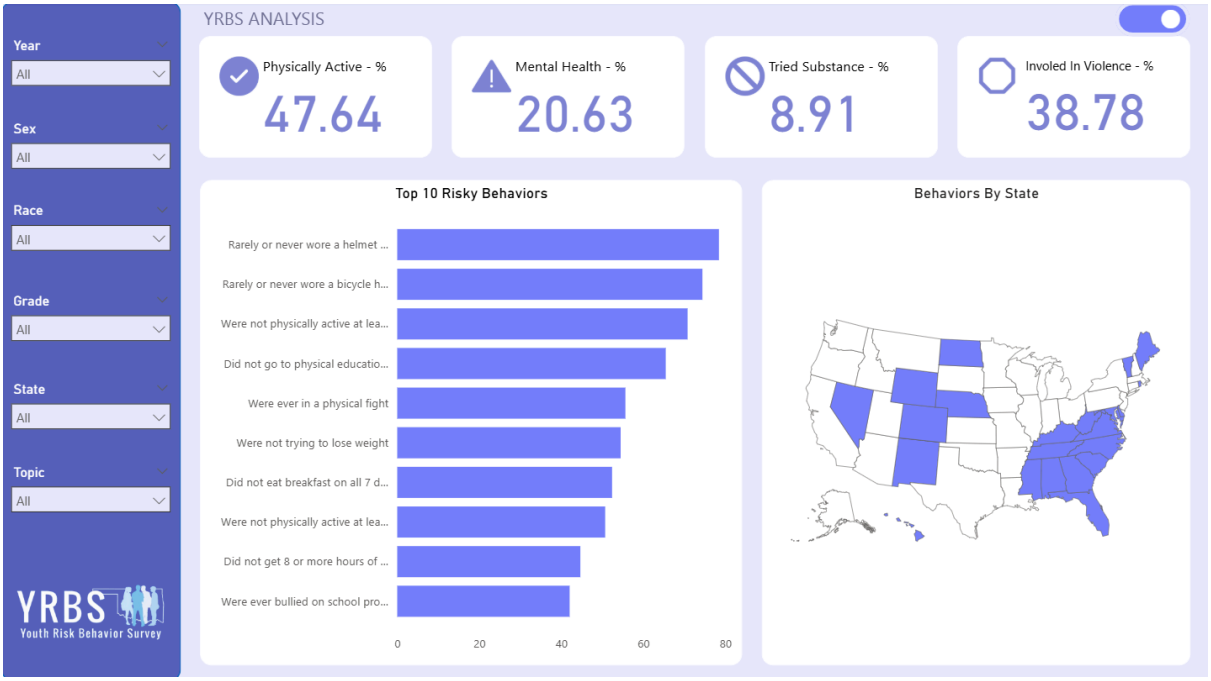
The above query evaluates temporal trends in risky behaviors by comparing the average reported risky behaviour across time. For each behavior (identified by QuestionCode and ShortQuestionText), it computes the mean Greater_Risk_Data_Value by year. Using conditional aggregation, the query extracts values from the earliest and most recent years available for each behavior. The difference between these two time points quantifies the change in prevalence. Finally, the results are ordered by the magnitude of this difference in descending order, and the top 10 behaviors with the largest increases are returned. The bar graph below visualizes the results.




# 6. Visualization

This section contains the Power BI dashboard created with the dataset.






# 7. Conclusion

The analysis reveals growing concerns around Mental Health, Substance Use, and Physical Inactivity among middle school students. These findings underscore the need for targeted interventions that focus on student well-being, with emphasis on Physical Education, Mental Health support, and anti-bullying measures.

# 8. Recommendations

- Implement school programs promoting daily physical activity with particular focus on females in 7th and 8th grades.
- Increase access to mental health counseling, especially for female students.
- Launch anti-vaping campaigns targeting early teens, particularly 8th-grade males.

- Support digital safety education to prevent cyberbullying against students, with particular attention to Native Hawaiian or Other Pacific Islander, White, and Black or African American males.