UNIVERSITY OF SOUTH FLORIDA, TAMPA

The Pulse of a State: A Dimensional View of Diabetes Across Florida

ISM6208 Data Warehousing

The Final Project

Under the guidance of Prof. Don Berndt

Group 5

Ankita Tripathy

Namratha Vardhineni

Priyanka Jammu

Raksha Basiwal

Ujwala Tripurana

Table of Content

1. EXECUTIVE SUMMARY

This project shows how income, education, and lifestyle factors like obesity and smoking affect diabetes hospitalizations across Florida counties from 2014 to 2023. By applying data warehousing techniques, SQL-based analysis, and Tableau visualizations, it transforms extensive public health data into meaningful insights that support more effective state and county health planning. Through dimensional modeling, analytical queries, and data storytelling, the project shows how complex data can be converted into practical knowledge to guide both health policy and resource allocation.

This study uses data from four trusted sources: Florida Health CHARTS for diabetes hospitalizations, the U.S. Census SAIPE program for income, the American Community Survey for education levels, and the County Health Rankings for obesity and smoking rates. Each dataset was cleaned, standardized, and linked by county and year to build a single, integrated database for analysis.

The database is built around a central fact table that captures diabetes hospitalization rates, supported by dimension tables for county, time, income, education, and health behaviors. Using SQL queries and Tableau dashboards, the analysis compares trends across counties, explores links between income and health, and highlights areas with higher diabetes rates through maps and visual reports.

Findings reveal that counties with lower income and education levels experience higher diabetes hospitalizations. Regions with higher obesity and smoking rates also show greater hospital visits. Although Florida's overall diabetes hospitalization rate has slightly improved over time, rural and low-income areas continue to face significant challenges.

Overall, this project demonstrates how data warehousing and visualization can uncover and explain health disparities. The insights generated can help public health agencies target resources more effectively, design stronger intervention programs, and work toward reducing diabetes inequalities across Florida.

2. PROBLEM STATEMENT

Across Florida, a quiet health crisis continues to grow. Thousands of residents are hospitalized each year for diabetes, yet the impact is far from uniform. In some counties, hospitalization rates are much higher, revealing that factors beyond medicine-like income, education, and lifestyle-play a powerful role in how people experience and manage this disease.

Consider two neighboring counties. One enjoys higher income levels, stronger education systems, and better access to healthcare. The other faces lower income, limited educational opportunities, and higher smoking and obesity rates. Both communities battle diabetes, but one recovers faster while the other continues to struggle. What causes this divide?

This project starts with that question. Its goal is to understand how social and behavioral factors shape diabetes hospitalization rates across Florida's 67 counties from 2014 to 2023. The focus is not just on tracking numbers but on revealing the deeper connections between poverty, education, lifestyle, and health outcomes.

To do this, a comprehensive data warehouse was built by integrating multiple public datasets: Florida Health CHARTS for hospitalization data, the U.S. Census SAIPE program for income, the American Community Survey for education, and County Health Rankings for lifestyle indicators such as obesity,

smoking, and physical inactivity. These datasets were modeled through fact and dimension tables, enabling detailed analysis using SQL and interactive storytelling in Tableau.

The project explores key questions such as:

- Which counties in Florida record the highest and lowest diabetes hospitalization rates?
- How does income level influence hospitalization risk?
- Do lower education levels correlate with higher hospitalization rates?
- How do lifestyle factors like obesity, smoking, and inactivity affect outcomes?
- How have these trends evolved between 2014 and 2023?

The findings are clear. Counties with lower income and education levels tend to have higher diabetes hospitalization rates. Areas with more obesity and smoking also face greater health challenges. While Florida's overall numbers have improved slightly, significant inequalities persist-particularly in rural and economically disadvantaged regions.

In the end, this project delivers a complete data warehousing and analytics framework that turns scattered public health data into meaningful insights. It demonstrates how data-driven analysis can guide smarter policies, target high-risk communities, and help create a healthier and more equitable future for all Floridians.

3. LITERATURE REVIEW

Understanding how social, economic, and lifestyle factors influence diabetes outcomes has been a key topic in public health research for years. Many studies show that income, education, and personal behaviors such as diet, physical activity, smoking, and obesity play a major role in determining a person’s risk of developing and being hospitalized for diabetes.

According to the Centers for Disease Control and Prevention (CDC, 2023), diabetes hospitalizations are strongly linked to socioeconomic status, with people in lower-income areas facing higher risks. Research by the Florida Department of Health (2022) also highlights that counties with limited access to healthcare and lower education levels often experience worse health outcomes. This supports the idea that social conditions and lifestyle behaviors are closely connected to chronic disease patterns.

From a data perspective, integrating these diverse health and social datasets is essential for meaningful analysis. Inmon (2005) describes a data warehouse as a system designed to bring together structured data from multiple sources to support decision-making. Kimball and Ross (2013) further emphasize that dimensional modeling helps make such data more usable for analysis by organizing it into facts and dimensions such as time, geography, and socioeconomic indicators. These approaches allow researchers to uncover patterns that might otherwise remain hidden in fragmented datasets.

Several studies have applied similar frameworks in public health analytics. For example, Singh et al. (2019) demonstrated how county-level health data could be analyzed through data warehousing techniques to identify clusters of chronic disease risk. Meanwhile, Zhao and Xu (2020) showed how visualization tools like Tableau can make complex health relationships more accessible to policymakers and the public.

Together, these works form the foundation for this project. By combining established data warehousing principles with public health data, this study aims to create a unified system that reveals how income, education, obesity, and smoking influence diabetes hospitalization rates across Florida. The goal is to

translate complex data into clear insights that can support data-driven decisions and promote health equity.

4. DATA COLLECTION AND PREPARATION

4.1 Data Collection

This project uses four main datasets that together provide a clear picture of diabetes hospitalizations and their relationship with social and behavioral factors across Florida counties from 2014 to 2023. Each dataset comes from a reliable public source and contributes unique information to the analysis.

- Florida Health CHARTS (Diabetes Data)

Source: Florida Health CHARTS
Link:
https://www.flhealthcharts.gov/ChartsDashboards/rdPage.aspx?rdReport=NonVitalInd.Dataviewer&cid=9750
Details: It provides annual diabetes hospitalization rates for each Florida county. The data is measured as hospitalizations per 100,000 residents. These figures help identify counties that experience higher burdens of diabetes and track how these rates change over time.

- U.S. Census Bureau – Small Area Income and Poverty Estimates (SAIPE)

Source: U.S. Census Bureau – SAIPE
Link: https://www.census.gov/programs-surveys/saipe.html
Details: The dataset contains county-level data on median household income and poverty estimates. These indicators are essential for understanding how economic status influences diabetes hospitalization rates.

- U.S. Census Bureau – American Community Survey (ACS)

Source: American Community Survey
Link: https://www.census.gov/programs-surveys/acs
Details: The dataset provides information on education levels by county, including the percentage of adults without a high school diploma. This dataset helps explore how education impacts health outcomes and access to care.

- County Health Rankings (CDC BRFSS and USDA)

Source: County Health Rankings
Link: https://www.countyhealthrankings.org/health-data/florida/data-and-resources
Details: The dataset includes county-level health behavior indicators such as obesity and smoking rates. These lifestyle factors are key to understanding the behavioral side of diabetes risks.

All datasets were collected in CSV format and prepared for integration into a single analytical model.

4.2 Data Preparation

The raw data from these sources required careful cleaning and standardization before analysis. The preparation process involved the following steps:

1. Data Cleaning and Filtering

- Removed unnecessary columns that did not contribute to the analysis.
- Verified that all values were consistent and complete, addressing missing or invalid entries.
- Standardized naming conventions for counties and years to ensure alignment across datasets.

1. Data Standardization

- Converted different variable formats (e.g., strings to numeric values for income, obesity rates, and hospitalization rates).
- Normalized data by calculating rates per 100,000 residents to allow fair comparison between counties with varying population sizes.
- Ensured consistent time ranges (2014–2023) across all datasets.

2. Merging and Integration

- Joined datasets using county and year as common identifiers.
- Combined socioeconomic data (income, education) with health outcome data (hospitalizations, obesity, smoking) to create a unified dataset.
- Verified the merged dataset for accuracy and completeness before analysis.

3. Dimensional Modeling

- Structured the integrated dataset into a star schema with one fact table for diabetes hospitalizations and multiple dimension tables for county, time, income, education, and health behaviors.
- This design supports analytical SQL queries and facilitates visualization in Tableau, allowing easy exploration of trends, correlations, and disparities.

5. DATABASE DESIGN

The database design for our project has two main parts:

- A transactional (OLTP) model that shows how data is collected in real life.
- A dimensional (OLAP) model, which is used for analytics, reporting, and visualization.

This design helps us to show the difference between day-to-day operational data and analytical data used for decision-making. The raw datasets were already aggregated and flattened, so we built a simple OLTP model to match a real healthcare system, and then created a full star schema for the analytical data warehouse.

5.1 Transactional Model

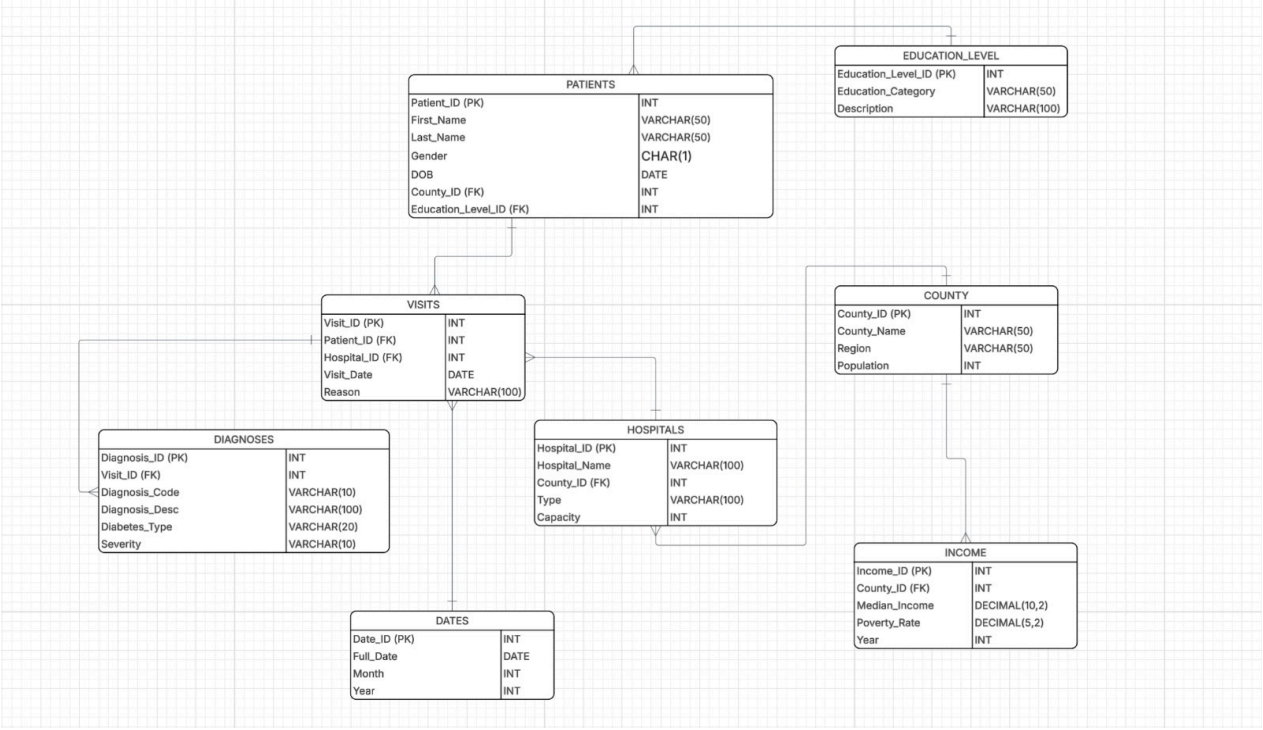
The OLTP model represents how a hospital or clinic records data every day. It stores detailed information about patients, their visits, diagnoses, hospitals, and counties. This model supports quick insert, update, and delete operations.

Real systems use OLTP databases for daily tasks such as registering new patients, recording hospital visits, storing diagnoses, updating patient details, producing operational reports. The OLTP model is not

optimized for analytics. It focuses on accuracy, data integrity, and fast transactions. The goal is to show what the “source system” would look like before data is cleaned and transformed for the warehouse.

Entity Relationship Diagram

OLTP model showing patient, visit, hospital, diagnosis, county, and date tables used for transaction-level data capture.



The OLTP ERD includes eight tables, each representing real-world healthcare activities:

- The PATIENTS table stores personal and demographic details for each patient. It includes fields such as first name, last name, date of birth, and county, which help with identification and patient categorization.
- The VISITS table records every hospital visit made by a patient. It connects each patient to a specific hospital through foreign keys and includes information such as the visit date and the reason for the visit.
- The DIAGNOSES table contains medical diagnosis information for each visit. It includes diagnosis codes, descriptions, the type of diabetes, and the severity level.
- The HOSPITALS table stores essential information about hospitals across Florida. It includes the hospital name, type, bed capacity, and the county in which the hospital is located.
- The COUNTY table provides geographic and population details for each county.
- The DATES table stores calendar-related information such as the full date, month, and year, supporting time-based analysis and reporting.
- The EDUCATION_LEVEL table includes categorized information about the education levels of patients.
- The INCOME table stores income-related data that can be used to analyze socioeconomic patterns across counties and patient groups.

5.2 Dimensional Model

The OLAP model is built using a star schema with one fact table and multiple dimension tables, which is easy to query in Tableau and SQL. It helps with decision-making because it provides clean dimensions and measurable facts. The dimensional model is a simplified, analytics-friendly structure designed for reporting and visualization.
This model helps answer high-level questions such as:

- Which counties have the highest diabetes hospitalization rates?
- Does income affect health outcomes?
- How have rates changed over time?
- Which socioeconomic factors predict higher diabetes burden?

The OLAP structure groups the data in a way that makes it easy to break down and analyze by county, year, income level, education level, and health behaviors.

5.3 Star Schema Overview

Our star schema has one main fact table and several dimension tables.

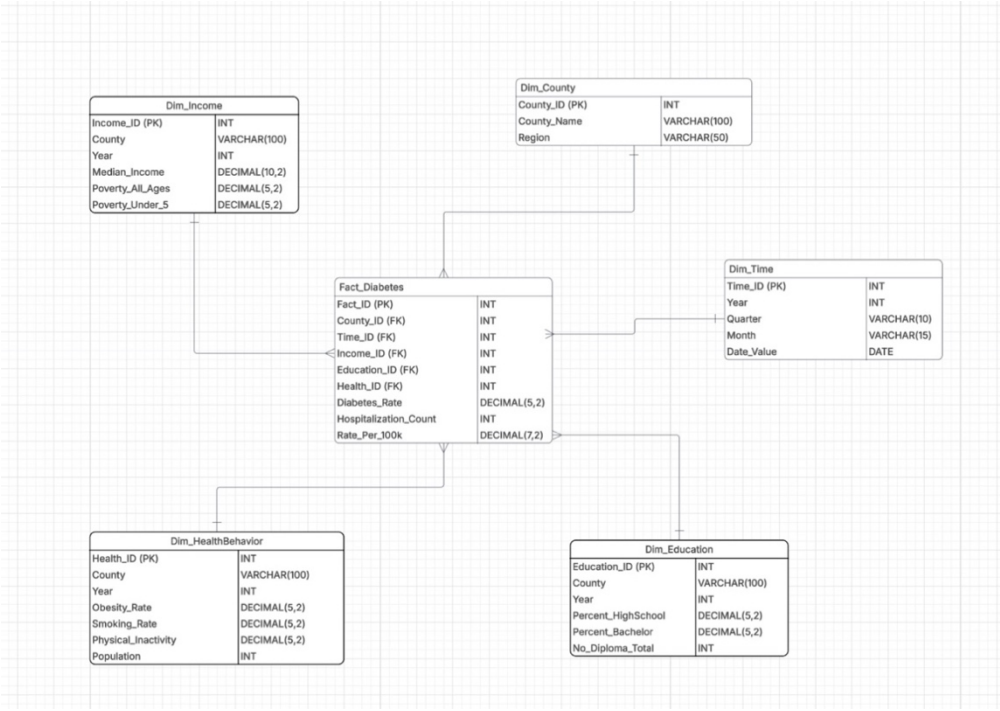
Fact Table-
The FACT_DIABETES table stores the numbers used for analysis. It includes the hospitalization count, the rate per 100,000 people, and the diabetes rate. It also holds foreign keys that link to each dimension table.

Dimensional Table-
- Dim_County stores the county name and region. It was created by merging all county values from the raw datasets using SQL.
- Dim_Time stores the year and date information. It was built using SQL by selecting the distinct years from different datasets.
- Dim_Income stores data such as median income, poverty levels, and the year.
- Dim_Education stores the percent of people with a high school education, the percent with a bachelor's degree, and the number of adults without a diploma.
- Dim_HealthBehavior stores information like obesity rate, smoking rate, physical inactivity rate, and population.

Each dimension gives extra details that help us to analyze diabetes outcomes from different angles.

Entity Relationship Diagram

Star schema for the diabetes data mart, showing Fact_Diabetes with County, Time, Income, Education, and Health Behavior dimensions.



Comparison Between OLTP and OLAP

In this project, we have built two types of database models one is OLTP model and the other is OLAP model. Both stores data, but they are designed for very different purposes. The OLTP model is used for daily operations, while the OLAP model is used for analysis and reporting.

The OLTP system stores very detailed information that is created during everyday activities inside hospitals and clinics. It includes data about each patient, each visit, each diagnosis, and each hospital. The tables are connected to each other, and the structure is very organized so that the data remains accurate. OLTP systems are used when hospitals need to add, update, or delete information quickly. They are designed to handle many small tasks, such as adding a new patient or recording a hospital visit. This type of system is very good for running daily operations but not good for analysis, because the data is spread across many tables and is too detailed.

The OLAP system is completely different. It does not store row-by-row information. Instead, it stores summarized and cleaned data that is ready for analysis. In this project, the OLAP model stores county-level data such as average income, education levels, obesity percentages, smoking rates, and diabetes hospitalization rates. The tables in the OLAP model are simple and easy to join, which makes it very fast for reporting and building dashboards. The main goal of OLAP is to help analysts and decision-makers understand trends, compare counties, and see which factors affect health outcomes.

The biggest difference between the two is the purpose. OLTP is built for operations, and OLAP is built for insights. OLTP focuses on storing small details, while OLAP focuses on storing summarized values. OLTP is

optimized for fast writing, while OLAP is optimized for fast reading. OLTP has many tables, while OLAP has only a few. OLTP is used every day by hospital staff, while OLAP is used by analysts looking at long-term patterns.
These two systems work together in a simple flow. First, the OLTP system collects all the detailed information. Then an ETL process cleans the data and prepares it for analysis. After that, the data is loaded into the OLAP system, where it is stored in the dimensional model. Finally, tools like Tableau use the OLAP data to create visualizations and insights. This process helps turn raw hospital data into useful public health information.

Hence, the OLTP model helps hospitals record daily activities, and the OLAP model helps us understand long-term health trends. Both systems are important, but they serve different purposes. The OLTP model represents how data is collected, and the OLAP model represents how data is used for analysis in this project.

5.4 Why OLAP Pattern Is Good

The OLAP pattern is good for this project because it makes the data easy to follow and query. Each dimension is narrowed in depth to a specific topic like income, education, health behaviors or time. The fact table contains the key measurements of hospitalization count and rate per 100K; this design allows for smooth slicing across county, year or behavior to compare what trends look like across Florida. The model runs very rapidly and integrates very well with Tableau dashboards because the data has already been summarized to the county-year level. One purpose of the project is to demonstrate how income, education and lifestyle influence diabetes hospitalization rates.

5.5 Data Cube and Dimensional Modeling

The design model for our project is implemented as a simple star schema. The objective is to facilitate rapid reporting and analysis of hospitalization for diabetes surveillance by county in the State of Florida. The main table in the model is Fact_Diabetes and contains the primary measures of hospitalization count, diabetes rate, and rate per 100k. Rows in this fact table represent a county for a year. This is the fundamental grain of the data cube.

The fact table links with multiple dimension tables. These are Dim_County, Dim_Time, Dim_Education, Dim_Income, and Dim_HealthBehavior. They give a sense of dimensions to the analysis. For example, the education dimension holds high school completion percentages, the income dimension stores median income and poverty rates, and the health behavior dimension stores lifestyle factors such as obesity and smoking. These dimensions are common to the whole cube, so they can be used in any reporting slice, for example county wise, year wise, or factor wise comparisons.

The concept of a data cube is a natural extension to this model. The cube makes it easy to slice and dice the data in many ways. For example, users can see hospitalization rates by year or county or income category or education level. They may also be able to drill down or roll up, depending on how much detail they require. Since all dimensions join to a single fact table, the cube enables ad hoc analysis with few joins. All dimensions in this model, such as County, Time, Income, Education, and Health Behavior, are shared across the data cube. This means the same county ID or time ID is reused across all facts, allowing consistent comparisons. Shared dimensions reduce redundancy, keep the model clean, and make slice-and-dice analysis fast and reliable.

A fundamental aspect of dimensional modeling is dealing with change. The optimal solution for this use case is SCD Type 2. In the real world, rates of education, income levels, and health behavior data rise and fall from year to year. For historical accuracy, we keep separate rows for each year in each dimension table. The solution is that we do not update older data. Instead, for each update, a new row including the updated value and the new year will be added. This approach enables the data warehouse to preserve a complete historical timeline while county demographic or economic conditions alter.

The dimensional model is much simpler than a transactional system and that is by design. It emphasizes strictly the measures and characteristics that are necessary for analytic purposes. The data also provide fast aggregation, easy slicing, and insights on how education, income, and lifestyle affect diabetes hospitalization in all of Florida.

5.6 SQL Implementation

1. Creating the warehouse user

```
CREATE USER diabetes_dw IDENTIFIED BY dw123;
GRANT CONNECT, RESOURCE TO diabetes_dw;
```

2. Exploring Raw Dimension Tables

```
SELECT * FROM DIM_EDUCATION FETCH FIRST 10 ROWS ONLY;
SELECT * FROM DIM_INCOME FETCH FIRST 10 ROWS ONLY;
SELECT * FROM DIM_HEALTH FETCH FIRST 10 ROWS ONLY;
SELECT * FROM DIM_DIABETES FETCH FIRST 10 ROWS ONLY;
```

3. Creating Dim_County

```
CREATE TABLE DIM_COUNTY AS
SELECT DISTINCT
   ROW_NUMBER() OVER (ORDER BY County) AS County_ID,
   County AS County_Name
FROM (
   SELECT County FROM DIM_EDUCATION
   UNION
   SELECT County FROM DIM_INCOME
   UNION
   SELECT County FROM DIM_HEALTH
   UNION
   SELECT County FROM DIM_DIABETES
);

SELECT * FROM DIM_COUNTY FETCH FIRST 10 ROWS ONLY;
```

```
DESC DIM_EDUCATION;
DESC DIM_INCOME;
DESC DIM_HEALTH;
DESC DIM_DIABETES;
DESC FACT_DIABETES;
```

Creating Dim_Time

```
CREATE TABLE DIM_TIME AS
SELECT DISTINCT
   ROW_NUMBER() OVER (ORDER BY YEAR) AS TIME_ID,
   YEAR,
   TO_DATE(YEAR || '-01-01', 'YYYY-MM-DD') AS DATE_VALUE
FROM (
   SELECT YEAR FROM DIM_HEALTH
   UNION
   SELECT YEAR FROM DIM_DIABETES
);


SELECT * FROM DIM_TIME ORDER BY YEAR;
```

Creating Fact_Diabetes

```
CREATE TABLE FACT_DIABETES (
   FACT_ID NUMBER GENERATED ALWAYS AS IDENTITY PRIMARY KEY,
   COUNTY_ID NUMBER,
   TIME_ID NUMBER,
   EDUCATION_ID NUMBER,
   INCOME_ID NUMBER,
   HEALTH_ID NUMBER,
   DIABETES_ID NUMBER,
   DIABETES_RATE NUMBER(5,2),
   HOSPITALIZATION_COUNT NUMBER,
   RATE_PER_100K NUMBER(7,2)
);
```

4. Loading FACT_DIABETES Using Joins

```
INSERT INTO FACT_DIABETES (
   COUNTY_ID, TIME_ID, EDUCATION_ID, INCOME_ID, HEALTH_ID, DIABETES_ID,
   HOSPITALIZATION_COUNT, RATE_PER_100K
)
SELECT
   c.COUNTY_ID,
   t.TIME_ID,
   NULL AS EDUCATION_ID,
   NULL AS INCOME_ID,
   NULL AS HEALTH_ID,
```

```
    NULL AS DIABETES_ID,
    d.HOSPITALIZATION_COUNT,
    d.RATE_PER_100K
FROM DIM_DIABETES d
JOIN DIM_COUNTY c ON d.COUNTY = c.COUNTY_NAME
JOIN DIM_TIME t ON d.YEAR = t.YEAR
LEFT JOIN DIM_HEALTH h ON h.COUNTY = d.COUNTY AND h.YEAR = d.YEAR
LEFT JOIN DIM_EDUCATION e ON e.COUNTY = d.COUNTY
LEFT JOIN DIM_INCOME i ON i.COUNTY = d.COUNTY;
```

5.7 Data Model Validation

We validated the data model by checking every major step of the design and loading process. First, we verified that all joins were working by comparing row counts and spot-checking counties across the different dimension tables. This helped confirm that each county matched correctly when loading the fact table. We also cleaned county names using TRIM and UPPER so that extra spaces, lowercase letters, or small formatting differences did not cause join failures.

We reviewed the foreign key fields in the fact table to make sure the structure was correct. Some foreign key columns such as EDUCATION_ID, INCOME_ID, and HEALTH_ID are still null at this stage. This is expected because only the diabetes data has been loaded into the fact table for now. These fields are included for future expansion when education, income, and health data will be fully integrated into the model.

We also checked for missing values, duplicated counties, incorrect year formats, and mismatched data types. These small checks helped ensure that the dimensional model is clean, consistent, and ready for analysis. This validation step confirms that the warehouse can support accurate reporting and future extensions without data quality issues.

6. Exploratory Data Analysis (EDA)

1. Top 10 Florida Counties by Average Adult Obesity Rate

Query Explanation:

The SQL query calculates the average adult obesity rate for each county in Florida. It takes the obesity values from the *DIM_HEALTH* table, computes the average for each county, rounds it to two decimals, and then orders the counties from highest to lowest obesity rate. Finally, it returns the top 10 counties with the highest average obesity rates.

Findings:
The output shows that several rural and lower-income counties have the highest obesity rates. Counties like Gadsden, Liberty, and Union consistently appear at the top, with obesity rates nearing or above 38–39%, which is significantly higher than the state average.

Insights:
High obesity rates are strongly linked to limited access to healthy food, fewer fitness resources, and higher poverty levels.

These counties also tend to show higher diabetes hospitalization rates, which suggests a clear relationship between lifestyle behaviors and chronic health outcomes.
The results highlight regions that may need targeted health programs, nutrition support, and improved community wellness services.

```sql
--Top 10 Florida Counties by Average Adult Obesity Rate

SELECT
    h.COUNTY,
    ROUND(AVG(h.OBESITY_RATE), 2) AS AVG_OBESITY_RATE
FROM DIM_HEALTH h
GROUP BY h.COUNTY
ORDER BY AVG_OBESITY_RATE DESC
FETCH FIRST 10 ROWS ONLY;
```

Script Output | Query Result | Query Result 1 | Query Result 2

SQL | All Rows Fetched: 10 in 0.068 seconds

	COUNTY	AVG_OBESITY_RATE
1	Gadsden	39.32
2	Liberty	38.08
3	Union	37.52
4	Hamilton	37.35
5	Washington	37.28
6	Hardee	37.16
7	Calhoun	37.03
8	Bradford	36.59
9	Wakulla	36.55
10	Taylor	36.41

2.Top 10 Florida Counties by Average Adult Smoking Rate

Query Explanation:

The SQL query calculates the average smoking rate for each county using data from the *DIM_HEALTH* table. It takes all smoking rate values, computes the average for each county, rounds it to two decimals, and then sorts the results from highest to lowest. The query returns the top 10 counties where smoking is most common.

Findings:

The output shows that counties like Dixie, Union, and Putnam have the highest smoking rates, all above 25%, which is significantly higher than the Florida state average. These counties tend to be rural with lower economic and educational levels, which aligns with known patterns of tobacco use.

Insights:

- High smoking counties often overlap with counties that have high obesity and high diabetes hospitalization rates.
- Smoking is a major risk factor for chronic illnesses, including diabetes complications, so these counties may require more targeted health education and prevention programs.
- The results highlight areas where public health initiatives such as anti-smoking campaigns, community programs, and access to cessation support could have a strong impact.

```sql
--Top 10 Counties by Smoking Rate

SELECT
    h.COUNTY,
    ROUND(AVG(h.SMOKING_RATE), 2) AS AVG_SMOKING_RATE
FROM DIM_HEALTH h
GROUP BY h.COUNTY
ORDER BY AVG_SMOKING_RATE DESC
FETCH FIRST 10 ROWS ONLY;
```

Script Output | Query Result | Query Result 1 | Query Result 2

SQL | All Rows Fetched: 10 in 0.064 seconds

	COUNTY	AVG_SMOKING_RATE
1	Dixie	26.06
2	Union	25.54
3	Putnam	25.13
4	Taylor	24.75
5	Calhoun	24.26
6	Washington	24.24
7	Hamilton	23.68
8	Liberty	23.41
9	Holmes	23.19
10	Gilchrist	23.06

3.Top 10 Florida Counties by Average Physical Inactivity Rate

Query Explanation:

The SQL query calculates the average physical inactivity rate for each county using data from the DIM_HEALTH table. It computes the mean inactivity value for every county, rounds the results to two decimal places, and then sorts them from the highest to the lowest. The query finally returns the top 10 counties where physical inactivity is most common.

Findings:

The output shows that Dixie, Calhoun, and Holmes counties have the highest physical inactivity levels, all above 34–36%, which is well above the state average. These counties often have fewer recreational facilities, limited access to active transportation, and lower-income populations factors that commonly contribute to reduced physical activity.

Insights:

- High physical inactivity often overlaps with high obesity and high diabetes hospitalization rates, showing a strong lifestyle-health connection.
- Physically inactive counties are usually rural, where residents may have fewer gyms, parks, or safe walking areas.
- These findings highlight the need for community-based health programs, better access to outdoor spaces, and awareness campaigns promoting exercise and mobility.

```sql
--Top 10 Counties by Physical Inactivity

SELECT
    h.COUNTY,
    ROUND(AVG(h.PHYSICAL_INACTIVITY), 2) AS AVG_INACTIVITY_RATE
FROM DIM_HEALTH h
GROUP BY h.COUNTY
ORDER BY AVG_INACTIVITY_RATE DESC
FETCH FIRST 10 ROWS ONLY;
```

Script Output | Query Result | Query Result 1 | Query Result 2

SQL | All Rows Fetched: 10 in 0.062 seconds

	COUNTY	AVG_INACTIVITY_RATE
1	Dixie	36.27
2	Calhoun	34.73
3	Holmes	34.68
4	Hamilton	34.65
5	Hardee	34.54
6	Jackson	34.28
7	Glades	34.2
8	Taylor	33.74
9	DeSoto	33.63
10	Putnam	33.47

4. Combined Health Behavior Comparison

Query Explanation:

This SQL query finds the average obesity, smoking, and physical inactivity rates for each county using data from the DIM_HEALTH table. It rounds the values, then shows the top 10 counties with the highest obesity rates. Sorting by obesity helps identify counties with several overlapping health problems.

Findings:

- Gadsden, Liberty, and Union counties have the highest overall lifestyle risk combination.
- These counties show consistently high numbers across all three categories, especially obesity and inactivity.
- Smoking rates also remain high in several counties like Union, Washington, and Taylor.

Insights:

- Counties with the highest obesity rates also show high physical inactivity, highlighting a clear link between low activity levels and weight-related health issues.
- These same counties often record higher diabetes hospitalization rates, showing how lifestyle choices affect long-term health.
- Most are rural, lower-income areas with limited access to healthy food, exercise facilities, and preventive care.
- This combined analysis helps identify where targeted programs like nutrition support, fitness initiatives, and smoking-cessation efforts are most needed.

```sql
--Combined Health Behavior Comparison

SELECT
    h.COUNTY,
    ROUND(AVG(h.OBESITY_RATE), 2) AS AVG_OBESITY_RATE,
    ROUND(AVG(h.SMOKING_RATE), 2) AS AVG_SMOKING_RATE,
    ROUND(AVG(h.PHYSICAL_INACTIVITY), 2) AS AVG_INACTIVITY_RATE
FROM DIM_HEALTH h
GROUP BY h.COUNTY
ORDER BY AVG_OBESITY_RATE DESC
FETCH FIRST 10 ROWS ONLY;
```

Script Output | Query Result | Query Result 1 | Query Result 2

SQL | All Rows Fetched: 10 in 0.075 seconds

	COUNTY	AVG_OBESITY_RATE	AVG_SMOKING_RATE	AVG_INACTIVITY_RATE
1	Gadsden	39.32	20.08	33.32
2	Liberty	38.08	23.41	32.82
3	Union	37.52	25.54	31.59
4	Hamilton	37.35	23.68	34.65
5	Washington	37.28	24.24	33.01
6	Hardee	37.16	21.15	34.54
7	Calhoun	37.03	24.26	34.73
8	Bradford	36.59	20.65	32.61
9	Wakulla	36.55	21.03	28.13
10	Taylor	36.41	24.75	33.74

5. Yearly Average Diabetes Hospitalization Rate by County

Query Explanation:

The SQL query calculates the average diabetes hospitalization rate for each county by year. It joins the fact table (FACT_DIABETES) with the county and time dimension tables to match each record with the correct county and year. The results are grouped by county and year, rounded to one decimal place, and then sorted so we can see the yearly trend for every county.

Findings:

- Many counties show a downward trend in hospitalization rates over the years, suggesting statewide progress.
- Some counties experience fluctuations, showing that improvements are not consistent every year.
- Counties like Baker have higher hospitalization rates than counties like Alachua, highlighting differences in health outcomes across regions.

Insights:

- Counties with higher risk factors such as obesity, smoking, and low income tend to maintain higher hospitalization rates over time.
- The overall pattern suggests that while Florida is improving, health inequalities remain, especially in rural or low-income areas.

- Tracking trends by year helps identify which counties are improving and which may need more focused intervention programs

```sql
--Yearly Average Diabetes Hospitalization Rate by County
SELECT
    c.COUNTY_NAME,
    t.YEAR,
    ROUND(AVG(f.RATE_PER_100K), 1) AS AVG_RATE_PER_100K
FROM FACT_DIABETES f
JOIN DIM_COUNTY c ON f.COUNTY_ID = c.COUNTY_ID
JOIN DIM_TIME t ON f.TIME_ID = t.TIME_ID
GROUP BY c.COUNTY_NAME, t.YEAR
ORDER BY c.COUNTY_NAME, t.YEAR;
```

Script Output | Query Result | Query Result 1 | Query Result 2 | Query Resu

SQL | Fetched 50 rows in 0.1 seconds

	COUNTY_NAME	YEAR	AVG_RATE_PER_100K
1	Alachua	2014	252.2
2	Alachua	2015	253.4
3	Alachua	2016	219.7
4	Alachua	2017	218.4
5	Alachua	2018	231.6
6	Alachua	2019	222.2
7	Alachua	2020	226.1
8	Alachua	2021	239.8
9	Alachua	2022	199.7
10	Alachua	2023	212.5
11	Baker	2014	273.7
12	Baker	2015	261.5
13	Baker	2016	259.2
14	Baker	2017	196.1
15	Baker	2018	188.2
16	Baker	2019	207.5
17	Baker	2020	272.1
18	Baker	2021	236.2
19	Baker	2022	227.4

6. Correlation Between Income and Hospitalization Rate

Query Explanation:

This SQL query measures the relationship between a county's median household income and its diabetes hospitalization rate. It uses the CORR() function to calculate the correlation between the two variables. The fact table (FACT_DIABETES) is joined with the income table (DIM_INCOME) to match each hospitalization record with the correct income level. The result is then rounded to two decimal places.

Findings:

A correlation of -0.34 means there is a moderate negative relationship between income and diabetes hospitalization. In simple terms, as income goes up, the hospitalization rate tends to go down.

Insights:

- Higher-income counties usually have better access to healthcare, healthier food, and diabetes management resources.
- Lower-income counties tend to face more challenges such as poor nutrition, fewer clinics, and limited preventive care, leading to higher hospitalization rates.
- This result supports the overall conclusion that economic conditions play a major role in diabetes outcomes across Florida.

```sql
--Correlation Between Income and Hospitalization Rate
SELECT
    ROUND(
        CORR(i.MEDIAN_INCOME, f.RATE_PER_100K),
        2
    ) AS INCOME_DIABETES_CORR
FROM FACT_DIABETES f
JOIN DIM_INCOME i
    ON f.INCOME_ID = i.INCOME_ID;
```

Script Output | Query Result | Query Result 1 | Query Result 2

SQL | All Rows Fetched: 1 in 0.076 seconds

	INCOME_DIABETES_CORR
1	-0.34

7. Correlation Between Education and Diabetes Hospitalization Rate

Query Explanation:

This query calculates the correlation between the percentage of adults without a high school diploma and the diabetes hospitalization rate for each county. The query joins the fact table with the education dimension and uses the CORR function to measure how strongly these two variables move together. The education metric is calculated as the total number of adults with no diploma divided by the county population. The result shows how changes in education levels relate to changes in hospitalization rates.

Findings:

The correlation value is **0.19**. This value is positive, but it is small. A positive correlation means that when the percentage of adults without a diploma increases, the hospitalization rate also increases. However, the small number means the relationship is not very strong. Education still plays a role, but it is not the biggest factor in hospitalization rates.

Insights:

- The result suggests that counties with lower education levels tend to have slightly higher diabetes hospitalization rates.
- Education alone does not fully explain the differences across counties.
- Other factors, such as obesity, smoking, income, and access to healthcare, may have a stronger influence.
- This insight matches the pattern seen in the visual analysis, where education affects health outcomes but works together with many other social and lifestyle factors.

```sql
--Education vs. Diabetes Rate
SELECT
    ROUND(
        CORR((e.NO_DIPLOMA_TOTAL / e.POPULATION) * 100, f.RATE_PER_100K),
        2
    ) AS EDU_DIABETES_CORR
FROM FACT_DIABETES f
JOIN DIM_EDUCATION e
    ON f.EDUCATION_ID = e.EDUCATION_ID;
```

Script Output | Query Result | Query Result 1 | Query Result 2 | Qu

SQL | All Rows Fetched: 1 in 0.082 seconds

	EDU_DIABETES_CORR
1	0.19

8. Top 5 Yearly Improvements in Diabetes Hospitalization Rates

Query Explanation:

This SQL query uses the LAG() window function to calculate how much the diabetes hospitalization rate changed from one year to the next for each county. It subtracts the previous year's rate from the current year's rate to find the year-over-year improvement. Negative values mean the rate decreased, which is a positive improvement. After calculating all changes, the query selects the top 5 biggest improvements.

Findings:

- Jefferson County appears twice, showing major improvements in two different years (2017 and 2022).
- Wakulla, Bradford, and Hamilton also show significant reductions over a single year.
- The rate drops are large, suggesting effective interventions or major changes in behavior or healthcare access.

Insights:

- These counties experienced sharp decreases in hospitalization rates, which could indicate better diabetes management, increased awareness, expanded access to care, or improvements in lifestyle factors.
- Tracking year-over-year changes helps identify where progress is happening and which counties may be benefiting from successful public health efforts.
- Such improvements highlight the value of targeted support and continued investment in high-risk counties.

```sql
--Top 5 Yearly Improvements

SELECT COUNTY_NAME, YEAR, RATE_CHANGE
FROM (
    SELECT
        c.COUNTY_NAME,
        t.YEAR,
        ROUND(f.RATE_PER_100K - LAG(f.RATE_PER_100K) OVER (PARTITION BY c.COUNTY_NAME ORDER BY t.YEAR), 1) AS RATE_CHANGE
    FROM FACT_DIABETES f
    JOIN DIM_COUNTY c ON f.COUNTY_ID = c.COUNTY_ID
    JOIN DIM_TIME t ON f.TIME_ID = t.TIME_ID
)
WHERE RATE_CHANGE IS NOT NULL
ORDER BY RATE_CHANGE ASC
FETCH FIRST 5 ROWS ONLY;
```

Script Output | Query Result | Query Result 1 | Query Result 2 | Query Result 3

SQL | All Rows Fetched: 5 in 0.08 seconds

	COUNTY_NAME	YEAR	RATE_CHANGE
1	Jefferson	2022	-192.2
2	Wakulla	2020	-134.2
3	Jefferson	2017	-122.6
4	Bradford	2017	-109.4
5	Hamilton	2020	-109.3

9.Combined Socioeconomic and Health Summary by County

Query Explanation:

This SQL query brings together data from several dimensions income, education, obesity, smoking, and hospitalization rates to create a full health and socioeconomic profile for each county. It joins the fact table *(*FACT_DIABETES*)* with income, education, and health tables using cleaned county names (using TRIM and UPPER to avoid mismatches). It then calculates the average income, percentage of adults without a diploma, average obesity rate, average smoking rate, and average diabetes hospitalization rate for each county. Finally, it sorts the counties by the highest diabetes hospitalization rates.

Findings:

- Most counties with high diabetes hospitalization rates also have lower median incomes, higher obesity, and higher smoking rates.
- Education plays a major role like counties with more adults lacking a high school diploma tend to rank higher in hospitalizations.
- Lifestyle risks and socioeconomic struggles often overlap.

Insights:

- Counties like Okeechobee, Putnam, and Gadsden show how multiple risk factors stack together and create worse health outcomes.
- Lower income and lower education often reduce access to healthy food, medical care, and preventive programs.
- These results highlight the importance of addressing social and behavioral factors not just medical care to reduce diabetes hospitalizations.
- This combined view helps public health teams target counties that would benefit most from programs in education, nutrition, smoking cessation, and community health support.

```sql
--Combined Socioeconomic and Health Summary by County

SELECT
    c.COUNTY_NAME,
    ROUND(AVG(i.MEDIAN_INCOME), 2) AS AVG_MEDIAN_INCOME,
    ROUND(AVG((e.NO_DIPLOMA_TOTAL / e.POPULATION) * 100), 2) AS AVG_NO_DIPLOMA_PCT,
    ROUND(AVG(h.OBESITY_RATE), 2) AS AVG_OBESITY_RATE,
    ROUND(AVG(h.SMOKING_RATE), 2) AS AVG_SMOKING_RATE,
    ROUND(AVG(f.RATE_PER_100K), 1) AS AVG_DIABETES_RATE
FROM FACT_DIABETES f
JOIN DIM_COUNTY c ON f.COUNTY_ID = c.COUNTY_ID
JOIN DIM_INCOME i ON TRIM(UPPER(c.COUNTY_NAME)) = TRIM(UPPER(i.COUNTY))
JOIN DIM_EDUCATION e ON TRIM(UPPER(c.COUNTY_NAME)) = TRIM(UPPER(e.COUNTY))
JOIN DIM_HEALTH h ON TRIM(UPPER(c.COUNTY_NAME)) = TRIM(UPPER(h.COUNTY))
GROUP BY c.COUNTY_NAME
ORDER BY AVG_DIABETES_RATE DESC
FETCH FIRST 10 ROWS ONLY;
```

Script Output | Query Result | Query Result 3 | Query Result 4 | Query Result 5

SQL | All Rows Fetched: 10 in 0.121 seconds

	COUNTY_NAME	AVG_MEDIAN_INCOME	AVG_NO_DIPLOMA_PCT	AVG_OBESITY_RATE	AVG_SMOKING_RATE	AVG_DIABETES_RATE
1	Okeechobee	54297	15.49	32.63	22.12	392.9
2	Putnam	51585	11.12	36.2	25.13	357
3	Columbia	58949	8.91	35.71	22.94	350.2
4	Dixie	50966	14.41	34.19	26.06	318.9
5	Duval	69455	6.72	30.79	18.97	316.9
6	Gadsden	51288	13.3	39.32	20.08	306.1
7	Polk	66709	9.15	34.15	18.55	304.3
8	Osceola	76416	7.17	30.6	18.13	292.2
9	Bradford	59468	13.48	36.59	20.65	290.9
10	Levy	53214	9.33	33.32	22.13	290.1

10. Average Diabetes Hospitalization Rate by Education Level (No Diploma %)

Query Explanation:

This SQL query groups counties into three education levels, High Education, Medium Education, and Low Education based on the percentage of adults without a high school diploma. It uses a CASE WHEN statement to classify each county, joins the diabetes fact table with the education table, and then calculates the average diabetes hospitalization rate for each group.

Findings:

Counties with lower education levels consistently have higher diabetes hospitalization rates. The “Low Education” group shows the highest hospitalization rate, while “High Education” counties have the lowest.

Insights:

- Education strongly influences health outcomes.
- Lower education is linked to poorer health awareness, limited preventive care, and higher rates of chronic conditions.
- These results suggest that improving education access and health literacy could help reduce diabetes-related hospitalizations across Florida.

```sql
--Average Diabetes Hospitalization Rate by Education Level (No Diploma %)

SELECT
    CASE
        WHEN (e.NO_DIPLOMA_TOTAL / e.POPULATION) * 100 < 10 THEN 'High Education'
        WHEN (e.NO_DIPLOMA_TOTAL / e.POPULATION) * 100 BETWEEN 10 AND 20 THEN 'Medium Education'
        ELSE 'Low Education'
    END AS EDUCATION_GROUP,
    ROUND(AVG(f.RATE_PER_100K), 1) AS AVG_HOSPITALIZATION_RATE
FROM FACT_DIABETES f
JOIN DIM_EDUCATION e ON f.EDUCATION_ID = e.EDUCATION_ID
GROUP BY
    CASE
        WHEN (e.NO_DIPLOMA_TOTAL / e.POPULATION) * 100 < 10 THEN 'High Education'
        WHEN (e.NO_DIPLOMA_TOTAL / e.POPULATION) * 100 BETWEEN 10 AND 20 THEN 'Medium Education'
        ELSE 'Low Education'
    END
ORDER BY AVG_HOSPITALIZATION_RATE DESC;
```

Script Output | Query Result | Query Result 1 | Query Result 2 | Query Result 3 | Que

SQL | All Rows Fetched: 3 in 0.137 seconds

	EDUCATION_GROUP	AVG_HOSPITALIZATION_RATE
1	Low Education	281.7
2	Medium Education	230.4
3	High Education	212.6

11. Which Florida counties fall into each quartile of diabetes hospitalization rates (per 100,000), and which counties consistently appear in the highest-risk quartile across different years?

Query Explanation

This query divides all Florida counties into four quartiles based on their diabetes hospitalization rate per 100,000 people. It uses the NTILE(4) function to rank counties from highest to lowest hospitalization rates for each year. Quartile 1 represents the highest-risk counties, and Quartile 4 represents the lowest-risk counties. The query joins the fact table with county, time, and diabetes dimensions to get the correct county name, year, and rate. It then orders the results so that the highest-risk counties appear first.

Findings

Many counties appear in the highest-risk quartile (Quartile 1) more than once. Dixie, Putnam, Jefferson, Columbia, Bradford, and Okeechobee show up frequently in Quartile 1 across multiple years. Okeechobee appears repeatedly across several different years, which shows a consistent pattern of very high hospitalization rates. Counties in Quartile 1 usually have rates above 380–490 per 100,000, which is much higher than the state average.

Insights

• The same counties showing up again and again in the highest-risk group suggests that the diabetes burden is not random.

• These counties may share similar socioeconomic factors such as low income, low education, or higher obesity and smoking rates.

• Public health efforts should focus more on counties like Okeechobee, Dixie, Columbia, and Putnam because they remain high-risk over the years.

• The quartile approach helps identify long-term patterns, not just one-year outliers. It shows which counties genuinely struggle with chronic health issues.
• This analysis supports your earlier visualizations and SQL findings, showing a clear cluster of counties that need targeted intervention.

```sql
--Which Florida counties fall into each quartile of diabetes hospitalization rates (per 100,000),
--and which counties consistently appear in the highest-risk quartile across different years?
SELECT
    c.COUNTY_NAME,
    t.YEAR,
    d.RATE_PER_100K,
    NTILE(4) OVER (ORDER BY d.RATE_PER_100K DESC) AS DIABETES_QUARTILE
FROM FACT_DIABETES f
JOIN DIM_COUNTY c     ON f.COUNTY_ID = c.COUNTY_ID
JOIN DIM_TIME t       ON f.TIME_ID = t.TIME_ID
JOIN DIM_DIABETES d   ON f.DIABETES_ID = d.DIABETES_ID
WHERE d.RATE_PER_100K IS NOT NULL
ORDER BY DIABETES_QUARTILE, d.RATE_PER_100K DESC;
```

Script Output | Query Result | Query Result 3 | Query Result 4 | Query Result 5

SQL | Fetched 50 rows in 0.056 seconds

	COUNTY_NAME	YEAR	RATE_PER_100K	DIABETES_QUARTILE
1	Dixie	2014	493.5	1
2	Putnam	2018	492	1
3	Jefferson	2021	452.6	1
4	Okeechobee	2017	444.8	1
5	Columbia	2020	425.6	1
6	Okeechobee	2021	425.6	1
7	Okeechobee	2014	422.4	1
8	Putnam	2019	417.8	1
9	Okeechobee	2018	416.7	1
10	Bradford	2014	416.3	1
11	Dixie	2015	398.6	1
12	Columbia	2018	398	1
13	Columbia	2019	396	1
14	Union	2019	394.6	1
15	Okeechobee	2019	387.7	1
16	Columbia	2014	382.6	1
17	Okeechobee	2022	378.5	1

7.REPORTING, MODELING, AND STORYTELLING

7.1 Storytelling with Data Visualization and Interpretation

This part of the project takes all the cleaned data and brings it into Tableau so that we can answer important questions about diabetes hospitalizations across Florida. The main goal is to turn the results from our SQL analysis into clear and easy-to-understand visuals that show how income, education, obesity, smoking, and physical inactivity affect health outcomes.

In Tableau, each chart focuses on one question like finding the top 10 counties with the highest hospitalization rates, comparing lifestyle factors, looking at income differences, or studying changes over time. These dashboards help us identify patterns that are not obvious in raw tables, such as regional clusters, steady improvements, or strong links between health behaviors and diabetes outcomes.

The earlier modeling work makes this possible by organizing the data into simple fact and dimension tables. Tableau then uses this structure to create maps, trend lines, bar charts, and comparison views. Overall, the visualizations help turn data into meaningful insights that support better public health understanding and decision-making.

7.2 Features Selection

In this project, we have selected a small set of features that are directly linked to public health outcomes and available consistently for all Florida counties from 2014 to 2023. The core features used include median household income, percentage of adults without a high school diploma, adult obesity rate, adult

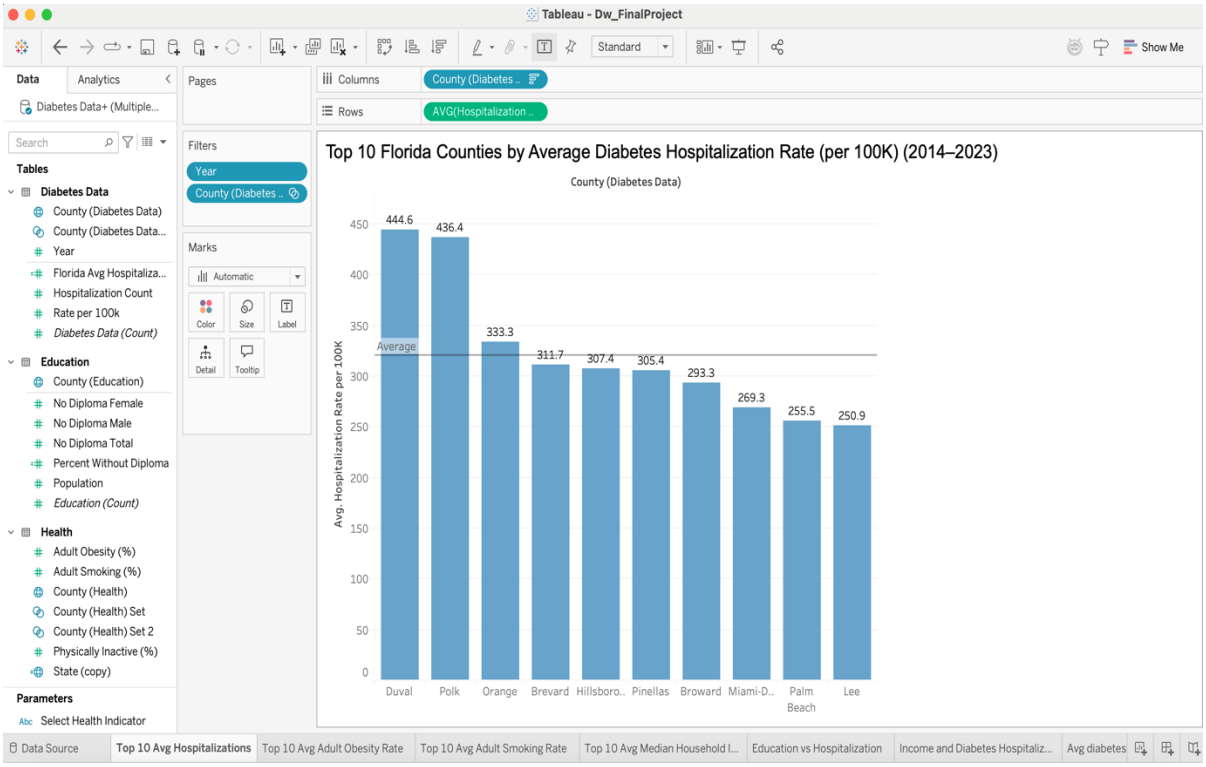
smoking rate, physical inactivity rate, diabetes hospitalization rate, county, and year. These fields were chosen because they are well known in public health research and because they match the county-year level of detail used in my data warehouse.

Some features were not included in this project. County-level data for race, insurance type, hospital quality, food access, and detailed age groups was not available for all years or all counties. Including incomplete features would reduce accuracy and create gaps, so they were excluded to maintain consistent analysis across the entire dataset.

These selected features support the main goal of the project, which is to understand why some counties have higher diabetes hospitalization rates than others. Income and education help explain long-term socioeconomic differences. Obesity, smoking, and inactivity show lifestyle and behavioral risks linked to diabetes and other chronic diseases. County and year allow patterns to be compared across places and over time. Together, these features help identify high-risk counties and explain the factors that contribute to health disparities in Florida.

1. Top 10 Florida Counties by Average Diabetes Hospitalization Rate (per 100K) (2014–2023)

The bar chart shows the top 10 Florida counties with the highest average diabetes hospitalization rates over a ten-year period. It compares each county's average rate to the statewide average line, making it easy to see which counties are above or below Florida's overall trend.



Features

For this visualization, the following key features were selected:

- County Name – to compare hospitalization rates across Florida counties.

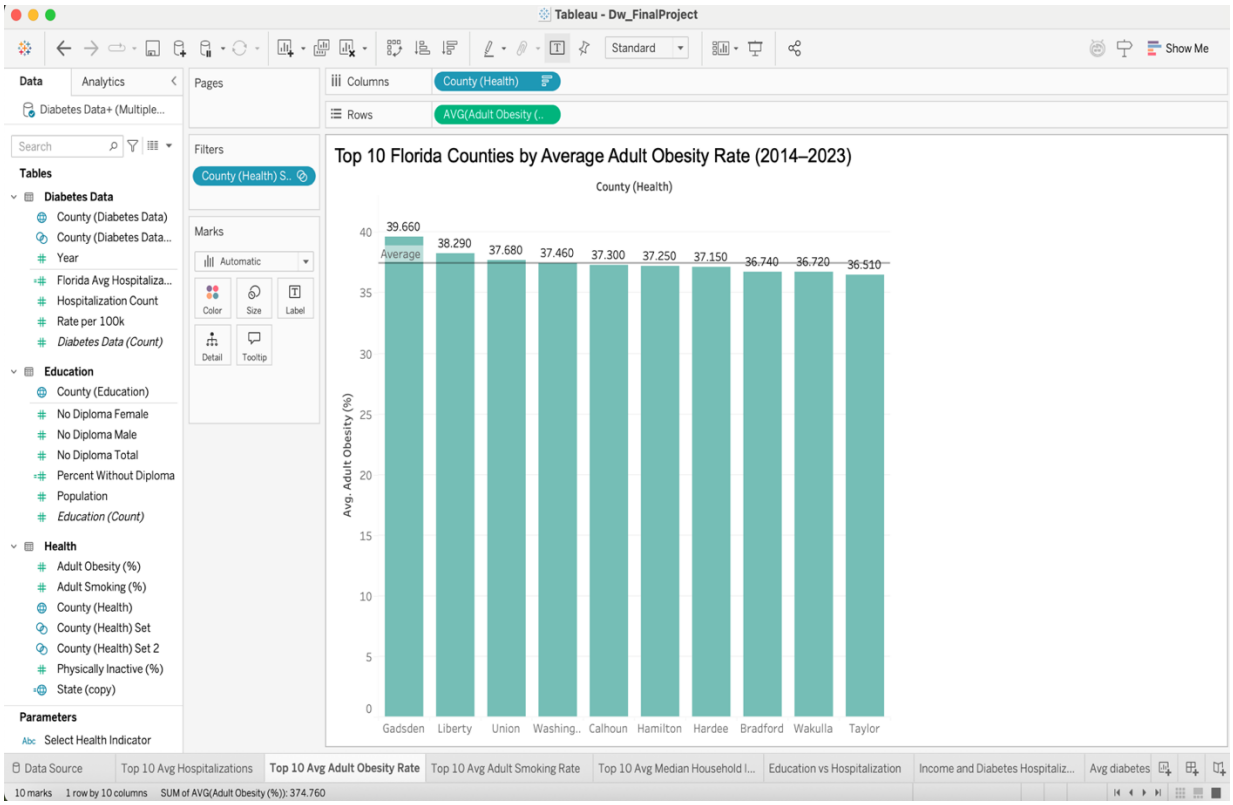
- Average Hospitalization Rate per 100K – the most direct indicator of the severity of diabetes complications.
- State Average Line – added to show how each county compares to the Florida benchmark.

Interpretation

Duval and Polk counties shows the highest hospitalization rates, both well above the state average of 333.3 per 100,000 people. Several other counties, such as Orange, Brevard, and Hillsborough, also remain above or near the state baseline. These results suggest that certain counties experience a heavier burden of diabetes complications and may need stronger healthcare support, prevention programs, and community health initiatives to reduce hospitalizations.

2. Top 10 Florida Counties by Average Adult Obesity Rate (2014–2023)

This bar chart displays the top 10 Florida counties with the highest average adult obesity percentages over a ten-year period. Each bar represents the county's average obesity rate, and the chart allows easy comparison across counties to see which ones consistently report higher levels.



Features

For this visualization, the following key features were selected:

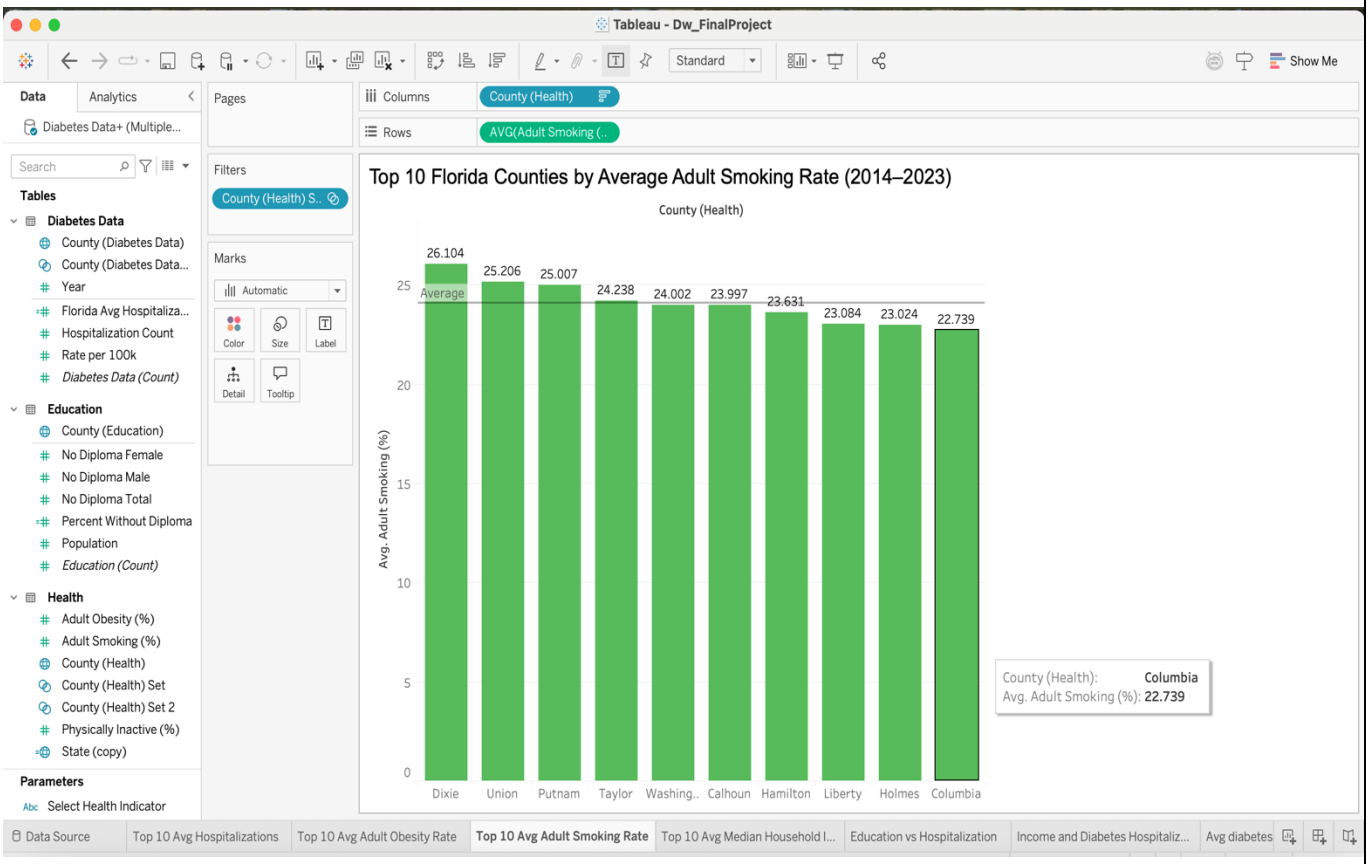
- County Name – needed to compare obesity levels across different regions in Florida.
- Average Adult Obesity Rate (%) – the main health indicator linked to diabetes and other chronic diseases.
- State Average Line (optional) – helps show whether each county is above or below the statewide trend.

Interpretation

Gadsden, Liberty, and Union counties is having the highest adult obesity rates, all close to or above 38%. The other counties in the list, such as Washington, Calhoun, and Hamilton, also show obesity levels well above the state average. These findings suggest that certain rural and lower-income counties face higher obesity challenges, which may contribute to other health issues like diabetes and heart disease.

3. Top 10 Florida Counties by Average Adult Smoking Rate (2014–2023)

This chart displays the top 10 counties in Florida with the highest average adult smoking percentages over the ten-year period. Each bar represents a county's average smoking rate, allowing a quick comparison against the state average line.



Features

For this visualization, the following features were selected:

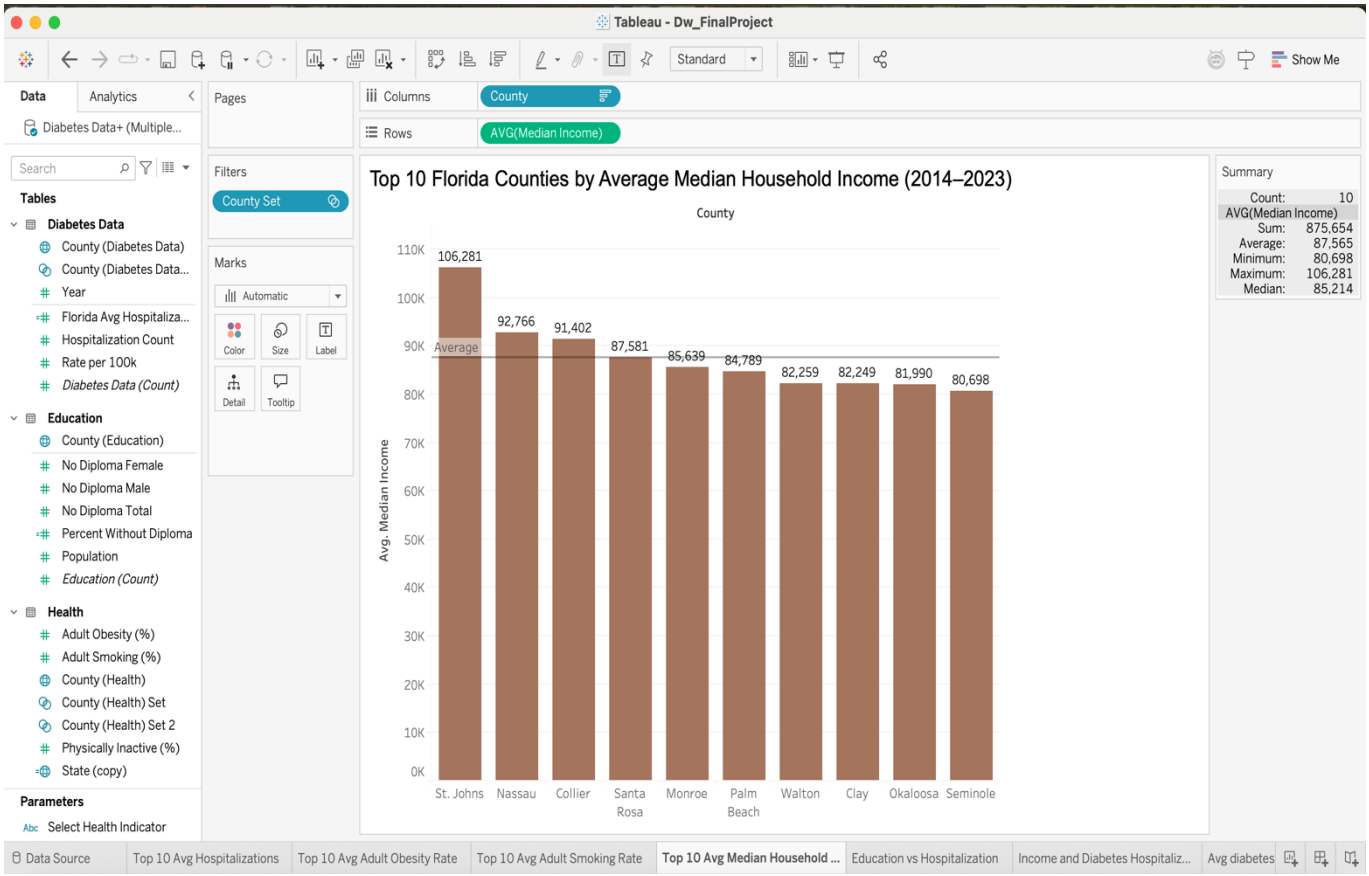
- County Name – used to compare smoking behavior across different counties.
- Average Adult Smoking Rate (%) – the main health variable that influences diabetes and other chronic diseases.
- State Average Smoking Line (optional) – helps viewers quickly see whether a county is above or below Florida's overall smoking trend.

Interpretation

Dixie, Union, and Putnam counties are having the highest smoking rates, each above 25%, which is higher than the Florida state average. These high smoking levels suggest greater health risks in these areas, as smoking is linked to many chronic conditions, including heart disease and diabetes. The counties shown in the chart may need stronger public health programs focused on smoking prevention and community awareness.

4. Top 10 Florida Counties by Average Median Household Income (2014–2023)

This bar chart presents the top 10 Florida counties with the highest average median household incomes over a ten-year period. Each bar represents the average income for a county, and the chart includes a state average line for comparison.



Features

For this visualization, the following features were chosen:

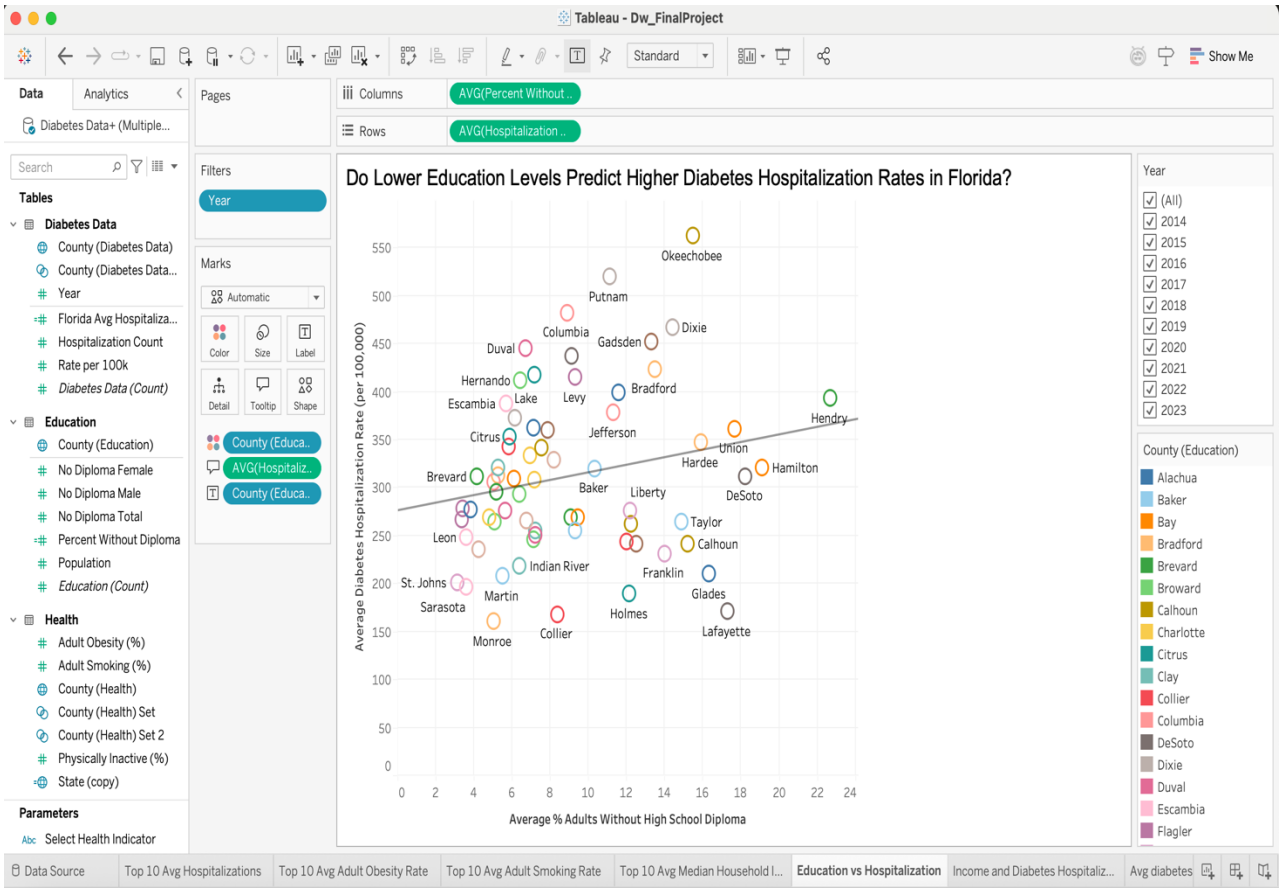
- County Name – allows ranking the counties and comparing income levels across regions.
- Average Median Household Income – the main economic variable used to understand wealth distribution.
- State Average Income Line (benchmark) – helps viewers identify which counties perform better than the state overall.

Interpretation

St. Johns County has the highest median income at over $106,000, which is far above the state average. Other high-income counties like Nassau, Collier, and Santa Rosa also appear in the top list. These wealthier counties often have better access to healthcare, healthier living conditions, and generally lower diabetes hospitalization rates. The chart highlights how higher income is linked to better health outcomes and reduced risk of chronic diseases

5. Do Lower Education Levels Predict Higher Diabetes Hospitalization Rates in Florida?

This scatter plot compares the percentage of adults without a high school diploma (x-axis) to the average diabetes hospitalization rate (y-axis) for each Florida county. Each dot represents one county, and the trend line shows the overall relationship between education and hospitalization rates.



Features

For this visualization, the following features were selected:

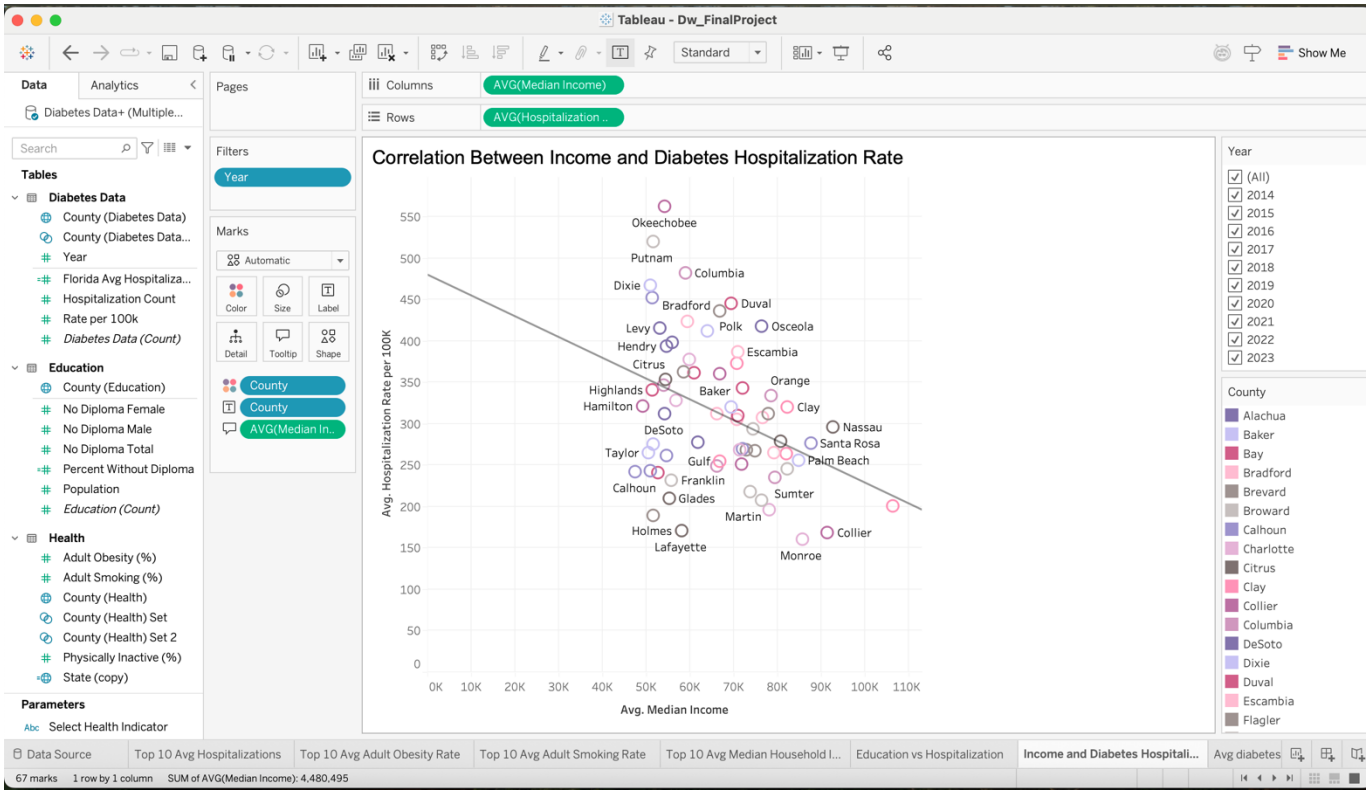
- Percent of adults without a high school diploma was chosen because education strongly affects health behavior, income, and access to health resources.
- Average diabetes hospitalization rate per 100K was included because it is the main health outcome we aim to understand.
- County name was included to identify outlier counties and highlight regional differences.
- Trend line was added to show the overall direction of the relationship between the two variables.

Interpretation

The upward-sloping trend line shows that counties with lower education levels tend to have higher diabetes hospitalization rates. Counties like Okeechobee, Putnam, and Dixie stand out as outliers with very high hospitalization numbers. In contrast, counties with higher education levels generally have lower hospitalizations. This suggests that improving education and health awareness may help reduce diabetes-related hospital visits across the state.

5. Correlation Between Income and Diabetes Hospitalization Rate

This scatter plot compares each county's average median income (x-axis) with its diabetes hospitalization rate per 100,000 people (y-axis). Each dot represents a county, and the trend line shows the overall direction of the relationship.



Features

For this visualization, the following features were selected:

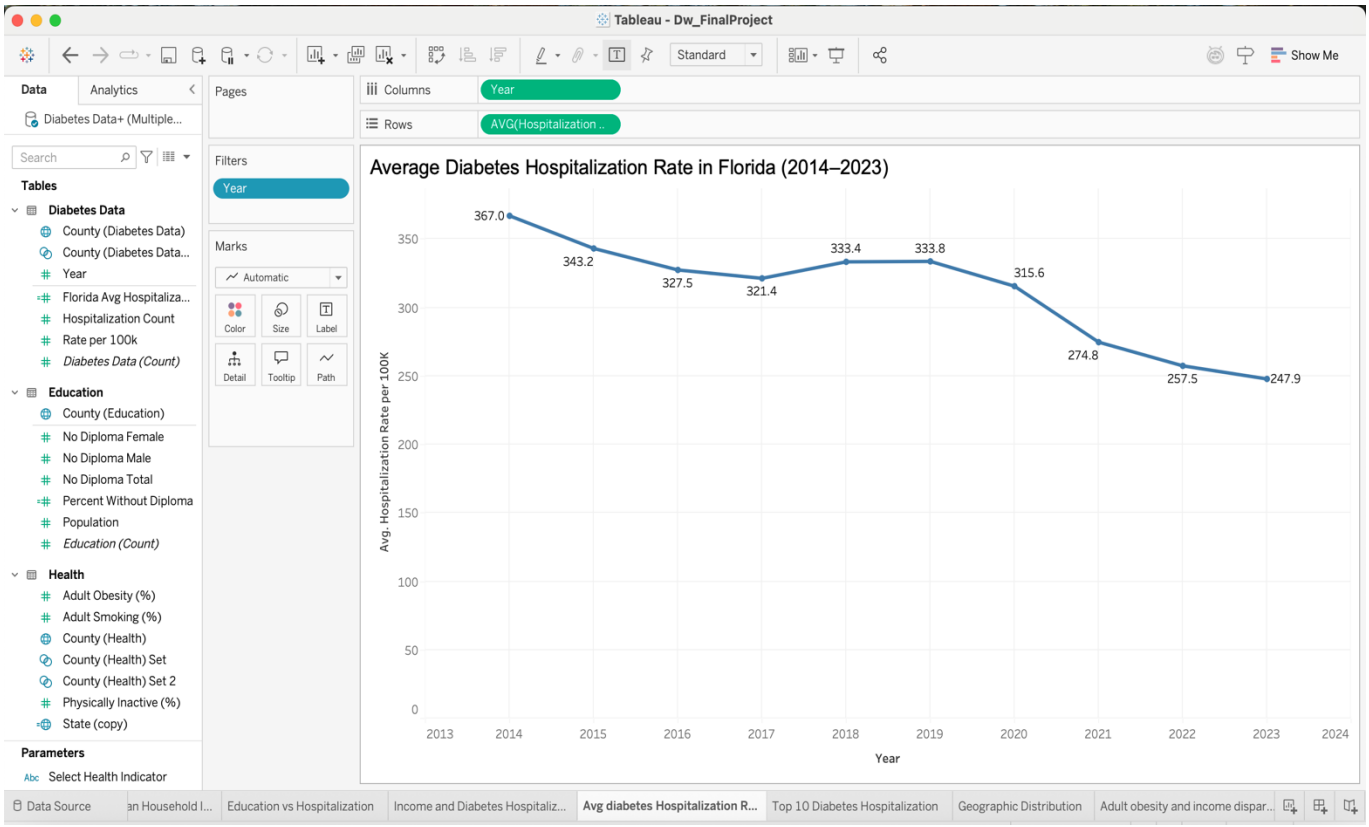
- Median household income was chosen because income strongly influences diet, medical care, preventive health, and living conditions.
- Diabetes hospitalization rate per 100K is the key health outcome being examined throughout the project.
- County names help identify outliers and understand regional patterns.
- The trend line was added to clearly show the overall direction of the relationship.

Interpretation

The downward-sloping trend line shows a clear negative correlation: counties with higher incomes usually have lower diabetes hospitalization rates. Wealthier counties like St. Johns, Collier, and Monroe appear on the lower-risk side, while lower-income counties like Okeechobee, Dixie, and Columbia show higher hospitalization levels. This pattern suggests that income plays an important role in health access, lifestyle quality, and diabetes management.

7. Average Diabetes Hospitalization Rate in Florida (2014–2023)

The line chart displays the statewide average diabetes hospitalization rate per 100,000 people for each year from 2014 to 2023. Each point on the line represents the average rate for that specific year, allowing us to see how the numbers rise or fall over time.



Features

For this visualization, the following features were selected:

Year was chosen to show how hospitalization rates change over time, which is important for identifying long-term trends.
Average hospitalization rate per 100K was used because it is the main health outcome in this project and represents the severity of diabetes complications statewide.
The line format was selected because it clearly shows year-to-year movement and long-term direction.

Interpretation

The chart shows a steady decline in diabetes hospitalization rates across Florida over the last decade. The rate starts at 367 per 100K in 2014 and drops almost every year, reaching 247.9 per 100K in 2023. Although there are a few small increases in certain years, the overall trend is clearly downward.

This decline suggests that diabetes management and prevention efforts in Florida may be improving. Better access to healthcare, increased awareness, healthier lifestyle choices, and improved medical treatment could all be contributing to fewer hospitalizations. The visualization makes it clear that, overall, Florida is moving in a positive direction when it comes to diabetes-related health outcomes.

8. Adult Obesity and Income Disparity Across Florida Counties

This scatter plot compares each county's median household income (x-axis) with its average adult obesity rate (y-axis). Every dot represents a county, allowing us to see how obesity levels change as income rises or falls.

Features

For this visualization, the following features were selected:

Median Household Income was selected because income strongly affects lifestyle, access to healthy food, healthcare availability, and preventive care.
Adult Obesity Rate (%) was included because obesity is one of the strongest predictors of diabetes complications and hospitalization.
County was used only as a label to show regional variation without overwhelming the chart.

Interpretation

The chart shows a clear pattern of counties with lower incomes tend to have higher obesity rates. Counties like Calhoun, Gadsden, and Union have some of the highest obesity levels and also some of the lowest incomes. On the other hand, wealthier counties such as St. Johns, Collier, and Monroe show much lower obesity rates.

This suggests that income plays an important role in health. Higher-income areas often have better access to healthy food, safer places to exercise, and stronger healthcare support. Lower-income counties may lack these resources, which can lead to higher obesity levels and other health problems.

7.3 Key Findings

- The charts show that some Florida counties consistently have higher diabetes hospitalization rates, obesity rates, and smoking rates.
- Counties like Duval, Polk, Gadsden, Liberty, and Dixie appear many times, which means these areas face more health challenges than others.
- Income plays a big role. Counties with higher income, such as St. Johns, Collier, and Nassau, show lower diabetes hospitalization rates and better health outcomes.
- Counties with lower education levels often have higher hospitalization rates, which suggests that education affects awareness, lifestyle choices, and access to care.
- Obesity and smoking patterns also match the hospitalization trends. Counties with high obesity and smoking rates tend to experience more diabetes-related hospital visits.
- The scatter plots clearly show a negative relationship between income and hospitalization, meaning hospitalization rates go down as income goes up.
- One interesting finding is that some rural counties with smaller populations still show very high risk, which might mean limited healthcare access in those areas.
- The visualizations together tell one clear story that income, education, and lifestyle behaviors all connect and influence diabetes outcomes across Florida.

8. CONCLUSION

This project shows how a well-designed data warehouse can help us understand the importance of public health patterns in Florida. By bringing together data on diabetes hospitalizations, income, education, obesity, and smoking, we were able to create a complete picture of how social and lifestyle factors affect health outcomes.

The dimensional model and fact tables helped organize the data in a way that made analysis simple and accurate. The SQL queries and visualizations turned raw numbers into clear stories about which counties are at higher risk and why. Our analysis showed that lower income, lower education, and higher obesity and smoking rates are strongly connected to higher diabetes hospitalization rates.

These findings can help health agencies identify counties that need more support, better prevention programs, and stronger health policies. Although the project provides strong insights, it is limited by the available data and does not include individual-level medical information. Future work can expand the model by adding hospital capacity, insurance coverage, or predictive models to estimate future hospitalization trends. Overall, this project demonstrates how data warehousing and analytics can guide better public health decisions and improve outcomes for communities across Florida.

9. REFERENCES

- Florida Health CHARTS. (2014–2023). County-level Diabetes Hospitalization Data. https://www.flhealthcharts.gov

- U.S. Census Bureau – Small Area Income and Poverty Estimates (SAIPE). Median Household Income by County. https://www.census.gov/programs-surveys/saipe.html

- American Community Survey (ACS). Educational Attainment Data by County. https://www.census.gov/programs-surveys/acs

- County Health Rankings & Roadmaps. Obesity, Smoking, and Physical Inactivity Indicators. https://www.countyhealthrankings.org

- Lord, J., & Odoi, A. (2024). Investigation of geographic disparities of diabetes-related hospitalizations in Florida using flexible spatial scan statistics: An ecological study. PLoS ONE, 19(6), e0298182. https://doi.org/10.1371/journal.pone.0298182
- Florida Diabetes Advisory Council. (2025). Florida Diabetes Advisory Council Legislative Report. Florida Department of Health. https://www.floridahealth.gov/provider-and-partner-resources/dac/_documents/2025-dac-report.pdf

- American Diabetes Association. (2023, September). The burden of diabetes in Florida: 2023 state fact sheet. https://diabetes.org/sites/default/files/2023-09/ADV_2023_State_Fact_sheets_all_rev_Florida.pdf

- Khavjou, O., et al. (2025). Rural–urban disparities in state-level diabetes prevalence in the United States. Preventing Chronic Disease, 22, 24_0199. https://www.cdc.gov/pcd/issues/2025/24_0199.htm

- Nath, N. D., et al. (2024). Geographic disparities and temporal changes of diabetes-related hospitalisations in Florida. BMC Public Health. https://pmc.ncbi.nlm.nih.gov/articles/PMC11214742/