Medical Data Visualization Project

Project Overview

A comprehensive data analysis and visualization project examining the relationship between cardiovascular disease and various health factors using Python's data science ecosystem.

Problem Statement

To explore and visualize patterns in medical examination data to understand how lifestyle choices, body measurements, and blood markers correlate with cardiovascular disease outcomes.

Dataset

- **Source**: Medical examination data (`medical_examination.csv`)
- **Size**: Patient records with 13 features
- **Target Variable**: Cardiovascular disease presence (binary)

Key Features Analyzed:

- **Objective Features**: Age, height, weight, gender
- **Examination Features**: Blood pressure (systolic/diastolic), cholesterol levels, glucose levels
- **Subjective Features**: Smoking, alcohol intake, physical activity
- **Derived Feature**: BMI-based overweight classification

Technical Implementation

Technologies Used

- **Python Libraries**: pandas, matplotlib, seaborn
- **Data Processing**: Data cleaning, normalization, feature engineering
- **Visualization**: Categorical plots, correlation heatmaps

Key Data Processing Steps

1. **Feature Engineering**: Created BMI-based overweight indicator (BMI > 25)
2. **Data Normalization**: Standardized categorical variables (0 = good, 1 = bad)
3. **Data Cleaning**: Removed physiologically impossible measurements:
   - Invalid blood pressure readings
   - Extreme height/weight values (outside 2.5th-97.5th percentiles)

Visualizations Created

1. Categorical Analysis Plot

- **Purpose**: Compare health factor distributions between patients with/without cardiovascular disease
- **Method**: Seaborn catplot with melted data structure
- **Variables**: Cholesterol, glucose, smoking, alcohol, physical activity, overweight status
- **Layout**: Side-by-side comparison panels for cardio=0 vs cardio=1

2. Correlation Heatmap

- **Purpose**: Identify relationships between all health variables
- **Method**: Seaborn heatmap with masked upper triangle
- **Features**: Clean correlation matrix showing variable interdependencies
- **Design**: Professional color scheme with clear value annotations

Key Insights & Findings

- Visual comparison of health factor prevalence in cardiovascular disease patients
- Correlation patterns between lifestyle choices and health outcomes
- Data quality assessment through outlier identification and removal

Technical Skills Demonstrated

- **Data Manipulation**: Complex pandas operations, data melting/reshaping
- **Statistical Analysis**: Correlation analysis, percentile-based filtering
- **Data Visualization**: Multi-panel plots, heatmaps, categorical visualizations
- **Code Organization**: Modular function-based approach
- **Data Quality**: Comprehensive data cleaning and validation

Project Structure

```
medical_data_visualizer.py
├── Data Import & Preprocessing
├── Feature Engineering (BMI calculation)
├── Data Normalization
├── draw_cat_plot() function
│   ├── Data melting and grouping
│   └── Categorical visualization
└── draw_heat_map() function
    ├── Data cleaning and filtering
    ├── Correlation calculation
    └── Heatmap visualization
```

Business Value

- **Healthcare Analytics**: Provides insights for medical decision-making
- **Risk Assessment**: Visualizes factors associated with cardiovascular disease
- **Data-Driven Medicine**: Supports evidence-based healthcare approaches

Portfolio Highlights

- Demonstrates proficiency in the complete data science workflow
- Shows ability to work with real-world medical data
- Exhibits strong data visualization and statistical analysis skills
- Proves capability in handling messy data and implementing quality controls

Future Enhancements

- Machine learning model development for disease prediction
- Interactive dashboard creation
- Additional statistical testing and hypothesis validation
- Integration with larger healthcare datasets