MEDICAL DATA OVERVIEW

EXPLORING HEALTHCARE DATA: INSIGHTS FROM EXPLORATORY DATA ANALYSIS AND PREDICTIVE MODELING FOR ENHANCED PATIENT CARE

Analysis by Tendai Milicent Jonhasi

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The project focuses on leveraging advanced data analytics to enhance healthcare outcomes by analyzing a comprehensive dataset comprising patient records, medical conditions, procedures, medications, and health observations. Its aim is to empower healthcare organizations with actionable insights derived from data-driven approaches, enabling informed decision-making and fostering a culture of continuous improvement in patient care. The integration of predictive analytics with interactive visualization tools supports evidence-based healthcare strategies aimed at enhancing overall patient well-being and quality of life.

We conducted thorough data exploration to understand the structure and quality of the dataset. This involved identifying missing values, outliers, and inconsistencies which were addressed through data cleaning processes.

Through EDA, we uncovered key insights into disease prevalence, patient demographics, medication usage patterns, and health monitoring trends. This analysis provided valuable context for understanding the healthcare landscape within the dataset.

Our project developed predictive models to estimate Quality Adjusted Life Years (QALY) based on health observations. By leveraging machine learning techniques, we aimed to identify factors influencing patient quality of life and provide actionable insights for enhanced patient care.

We integrated our analytical findings into a Power BI dashboard, enabling healthcare stakeholders to interactively explore patient profiles, visualize health metrics, and monitor key performance indicators (KPIs) related to patient care and outcomes.

Key Findings and Recommendations:

Our analysis revealed significant insights into disease prevalence, medication effectiveness, and patient risk profiling. Based on these findings, we recommend personalized care interventions, optimized resource allocation, and targeted health interventions to improve patient outcomes and healthcare delivery.

In summary, our project empowers healthcare organizations with actionable insights derived from data-driven approaches, enabling informed decision-making and fostering a culture of continuous improvement in patient care and population health management. The integration of predictive analytics with interactive visualization tools supports evidence-based healthcare strategies aimed at enhancing overall patient well-being and quality of life.

PROJECT OVERVIEW

Dataset

This data was generated using Synthea, a synthetic patient generator that models the medical history of synthetic patients. Their mission is to output high-quality synthetic, realistic but not real, patient data and associated health records covering every aspect of healthcare. The resulting data is free from cost, privacy, and security restrictions, enabling research with Health IT data that is otherwise legally or practically unavailable.

1 DATABASE

The Database contains several tables detailed below:

1. Patients: Contains demographic information such as age, gender, and vital dates like birthdate and deathdate.
2. Conditions: Describes various medical conditions diagnosed in patients over time.
3. Medications: Lists medications prescribed to patients along with associated details like reason for prescription.
4. Observations: Captures different health observations including body metrics (BMI, weight, height), allergies, and immunizations.
5. Procedures: Records medical procedures performed on patients with reasons and descriptions.
6. Allergies: Contains all allergies that patients have recorded
7. Careplans, Claims and other medical-related data

2 TOOLS FOR ANALYSIS

1. Data Exploration and Cleansing:
   - MySQL - Used SQL queries to extract and explore data from relational databases
2. Exploratory Data Analysis (EDA):
   - Python (Pandas, NumPy) - Utilized Pandas for data manipulation and preprocessing tasks
   - Visualization Libraries (Matplotlib, Seaborn)
3. Predictive Modeling:
   - Python (Scikit-Learn) - Implemented machine learning algorithms for predictive modeling (e.g., linear regression, decision tree, random forest) to estimate QALY based on health observations.
4. Dashboard Development
   - Power BI - Developed interactive dashboards using Power BI to visualize key insights derived from data exploration
   - Integrated SQL queries and Python scripts within Power BI to dynamically connect to the database and apply predictive models

SCOPE AND METHODOLOGY

Exploratory and Descriptive Analytics

Understanding Disease Prevalence: Identify the most prevalent medical conditions within the patient population to prioritize resource allocation and targeted interventions.

Patient Risk Profiling: Develop profiles of patients at higher risk based on their demographic information, medical history, and health metrics to enable proactive care management.

Health Monitoring Trends: Analyze trends in key health metrics such as BMI, weight, and height to monitor population health and identify areas of concern

Medication Effectiveness: Evaluate the effectiveness of medications in managing specific health conditions and their impact on patient outcomes.

Predictive and Prescriptive Analytics

Quality of Life Assessment: Build predictive models to estimate Quality Adjusted Life Years (QALY) based on health observations and identify factors influencing patient quality of life.

Clinical Decision Support: Provide actionable insights to healthcare providers to support clinical decision-making and enhance personalized care delivery.

Optimizing Healthcare Resource Allocation: Inform resource allocation strategies based on prevalent conditions, patient demographics, and healthcare utilization patterns.

INSIGHTS FROM EXPLORATORY AND DESCRIPTIVE ANALYTICS

While we uncover various insights from the data (see all sql queries), we will focus our attention on the following:

1. Show all prevalent disorders/diseases that have an occurrence of over 30% or 300 in the population that do not have the word Mg in the middle or the end

```
SELECT DISTINCT item AS Prevalent_Diseases FROM all_prevalences
WHERE all_prevalences.occurrences > 300 AND item NOT LIKE '% Mg%';
```

Prevalent_Diseases
Viral Sinusitis (Disorder)
Streptococcal Sore Throat (Disorder)
Acute Viral Pharyngitis (Disorder)
Acute Bronchitis (Disorder)
Normal Pregnancy
Otitis Media
Sprain Of Ankle

2. Retrieve all patients that suffered a stroke and where given Clopidogrel 75 MG Oral Tablet?

```
SELECT medications.patient, medications.description,
medications.reasondescription FROM medications WHERE
medications.reasondescription = "Stroke" AND
medications.description = "Clopidogrel 75 MG Oral Tablet" ORDER
BY medications.patient;
```

patient	description
0bcc9845-c873-492e-96e1-9771ebcbc2df	Clopidogrel 75 MG Oral Tablet
23ffe43f-c4b4-432f-ae2f-3aa21b7d513d	Clopidogrel 75 MG Oral Tablet
299a069f-9235-4558-a636-81e6b6417140	Clopidogrel 75 MG Oral Tablet
2e3420e7-1f2b-4da0-a170-cc7c8ffa8bef	Clopidogrel 75 MG Oral Tablet
357c4996-7893-4a4e-b0e8-8ab16588b282	Clopidogrel 75 MG Oral Tablet
3d392eab-9c13-4201-9f13-e1ca412f6dac	Clopidogrel 75 MG Oral Tablet

INSIGHTS FROM EXPLORATORY AND DESCRIPTIVE ANALYTICS

3. Retrieve all patients with Diabetes or Prediabetes?

*SELECT * FROM conditions WHERE conditions.description = "Diabetes" OR conditions.description = "Prediabetes";*

PATIENT	ENCOUNTER	CODE	DESCRIPTION
96b24072-e1fe-49cd-a22a-6dfb92c3994c	4e7beaee-50c2-4609-8a2b-b32fb3dc5a3b	15777000	Prediabetes
96b24072-e1fe-49cd-a22a-6dfb92c3994c	4e7beaee-50c2-4609-8a2b-b32fb3dc5a3b	44054006	Diabetes
de43eb48-496c-46d4-8c5b-be6125a38c15	febdc129-bd41-4b7c-b1e7-a4ce07931544	44054006	Diabetes
de43eb48-496c-46d4-8c5b-be6125a38c15	febdc129-bd41-4b7c-b1e7-a4ce07931544	15777000	Prediabetes

5. Retrieve all patients who received Acetaminophen and the reason description?

*SELECT * FROM conditions WHERE conditions.description = "Diabetes" OR conditions.description = "Prediabetes";*

patient	description	reasondescription
8772ca86-5853-46f5-9179-e2351feb22a6	Acetaminophen 325 MG / oxyCODONE Hydrochl...	Primary fibromyalgia synd...
5b30f187-8b4e-4484-8c94-18b5cdb2151a	Acetaminophen 325 MG / oxyCODONE Hydrochl...	Primary fibromyalgia synd...
71949668-1c2e-43ae-ab0a-64654608defb	Acetaminophen 160 MG	Acute bronchitis (disorder)
2c884d0f-62a1-4371-becc-36a98cdc4f52	Acetaminophen 160 MG	Acute bronchitis (disorder)
0853f544-fdef-4464-8745-239cccfe94b2	Acetaminophen 160 MG	Acute bronchitis (disorder)

4. What other procedures were done to patients with normal pregnancies excluding 'Standard pregnancy test'?

SELECT DISTINCT procedures.description FROM procedures WHERE procedures.reasondescription = "Normal pregnancy" AND NOT procedures.description = "Standard pregnancy test";

description
Augmentation of labor
Cesarean section
Induced termination of pregnancy
Medical induction of labor
Childbirth
Premature birth of newborn
Episiotomy
Epidural anesthesia
Instrumental delivery

INSIGHTS FROM EXPLORATORY AND DESCRIPTIVE ANALYTICS

6. Link patients to their medications and diagnosis reasons?

SELECT patients.patient, patients.first, patients.last, medications.description AS medication, medications.reasondescription AS diagnosis FROM patients JOIN medications ON medications.patient = patients.patient;

patient	first	last	medication	diagnosis
71949668-1c2e-43ae-ab0a-64654608defb	Elly	Koss	Phenazopyridine hydrochloride 100 MG [Pyridium]	Escherichia coli urinary tract infection
71949668-1c2e-43ae-ab0a-64654608defb	Elly	Koss	NITROFURANTOIN MACROCRYSTALS 50 MG [...	Escherichia coli urinary tract infection
c2caaace-9119-4b2d-a2c3-4040f5a9cf32	Kim	Barr...	Penicillin V Potassium 250 MG	Streptococcal sore throat (disorder)
c2caaace-9119-4b2d-a2c3-4040f5a9cf32	Kim	Barr...	Penicillin V Potassium 250 MG	Streptococcal sore throat (disorder)
c2caaace-9119-4b2d-a2c3-4040f5a9cf32	Kim	Barr...	Acetaminophen 325 MG / oxyCODONE Hydrochl...	NULL
96b24072-e1fe-49cd-a22a-6dfb92c3994c	Jac...	Sha...	Penicillin V Potassium 250 MG	Streptococcal sore throat (disorder)
96b24072-e1fe-49cd-a22a-6dfb92c3994c	Jac...	Sha...	Penicillin V Potassium 500 MG	Streptococcal sore throat (disorder)

8. How many patients are allergic to mould, grass pollen, tree and house dust mite?

SELECT allergies.patient, allergies.description FROM allergies WHERE allergies.description IN ("Allergy to mould", "Allergy to grass pollen", "Allergy to tree pollen", "House dust mite allergy");

patient	description
ab6d8296-d3c7-4fef-9215-40b156db67ac	Allergy to tree pollen

7. Retrieve patient information along with associated procedure details for Asian female patients only?

SELECT patients.patient, patients.race, patients.ethnicity, patients.gender, procedures.description AS procedures, procedures.reasondescription AS diagnosis FROM patients LEFT OUTER JOIN procedures ON patients.patient = procedures.patient WHERE patients.race = "asian" AND patients.gender = "F";

patient	race	ethnicity	gender	procedures	diagnosis
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL
684887d3-1080-4b73-bef9-e78595722f01	asian	chinese	F	Intramuscular injection	NULL

INSIGHTS FROM EXPLORATORY AND DESCRIPTIVE ANALYTICS

9. Calculate average BMI, number of (BMI) readings per patient, and the maximum BMI recorded that exceeds a threshold indicating obesity (BMI > 30)?

```
SELECT observations_cleaned.patient, AVG(CASE WHEN observations_cleaned.description = 'Body Mass Index' THEN observations_cleaned.value END) AS `Avg BMI`, COUNT(CASE WHEN observations_cleaned.description = 'Body Mass Index' THEN observations_cleaned.value END) AS `Number of Readings`, MAX(CASE WHEN observations_cleaned.description = 'Body Mass Index' AND observations_cleaned.value > 30 THEN observations_cleaned.value END) AS `Max Obese BMI`FROM observations_cleanedWHERE observations_cleaned.description = 'Body Mass Index'GROUP BY observations_cleaned.patient;
```

patient	Avg BMI	Number of Readings	Max Obese BMI
401f0b8c-3e96-4de5-a9aa-c9bbe51b8407	51.2399999999995	5	53.65
81979c08-df99-4792-b608-65fcf28f4019	51.735	4	53.57
ecdec4e3-194b-4e49-8baa-f20c8fdef087	51.734	10	53.22
29e87640-97a8-4df1-b19f-c24bf2b52cb7	52.11	7	52.28
0e9db90e-e07c-4269-b96e-d9f9df054dba	51.712	10	51.76
08608e32-0dc5-43a8-ba12-99fbb8c9575c	51.57899999999999	10	51.65
2835c7ab-938a-4cdd-ba36-13486040c3ce	51.301	10	51.61
bea0b538-ce64-4158-9f0f-1b73442413e4	51.26299999999999	10	51.45

INSIGHTS FROM EXPLORATORY AND DESCRIPTIVE ANALYTICS

10. Categorize BMI values into different categories ('Underweight', 'Healthy', 'Overweight', 'Obese') for patients based on their BMI observations?

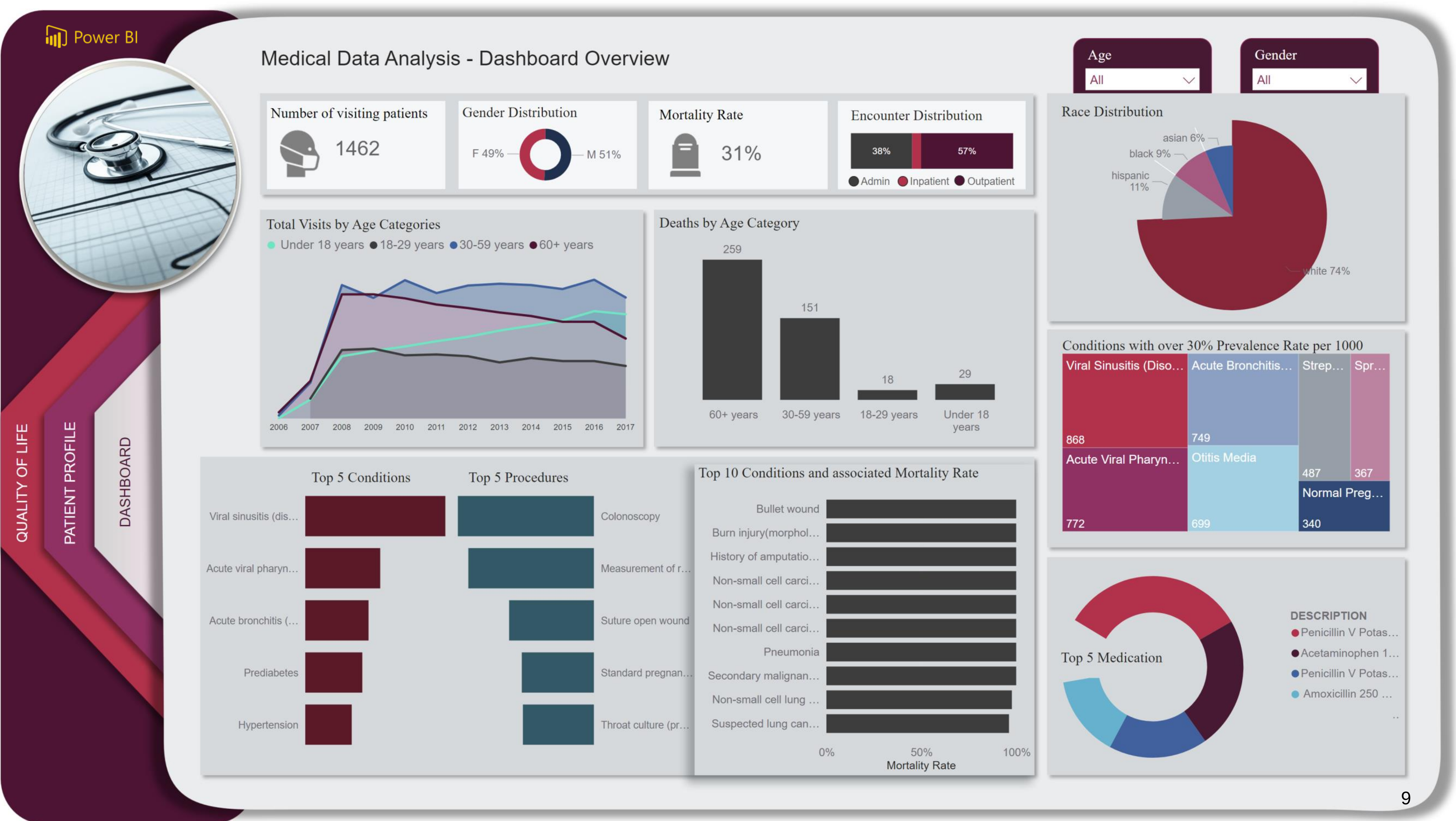
```
SELECT observations_cleaned.patient, AVG(observations_cleaned.value) AS `Avg BMI`, CASE WHEN observations_cleaned.value < 18.5 THEN 'Underweight' WHEN observations_cleaned.value >= 18.5 AND observations_cleaned.value < 25 THEN 'Healthy' WHEN observations_cleaned.value >= 25 AND observations_cleaned.value < 30 THEN 'Overweight' WHEN observations_cleaned.value >= 30 THEN 'Obese' END AS `BMI category`FROM observations_cleanedWHERE observations_cleaned.description = 'Body Mass Index'
```

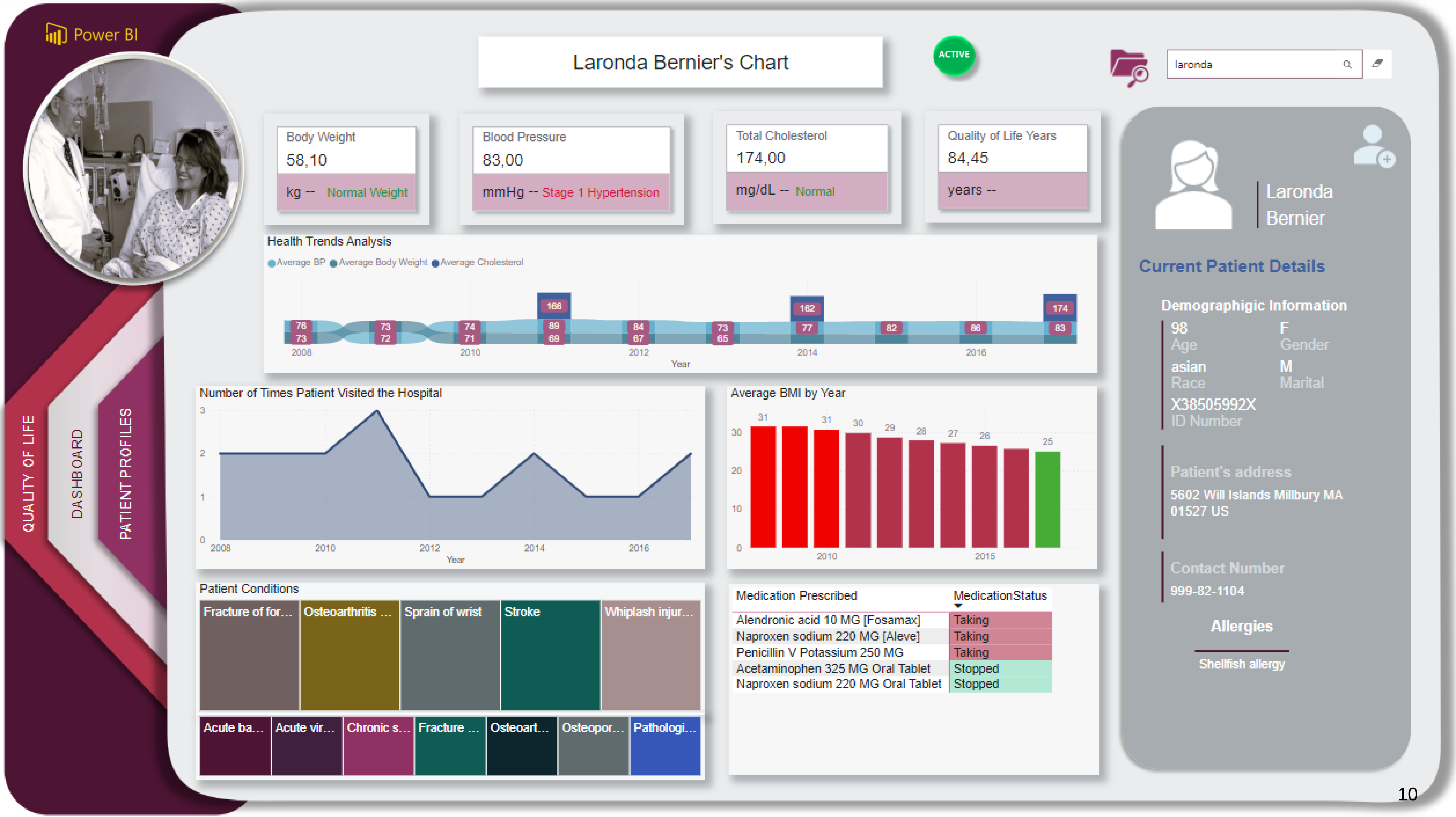
patient	Avg BMI	BMI category
71949668-1c2e-43ae-ab0a-64654608defb	21.505000000000003	Healthy
96b24072-e1fe-49cd-a22a-6dfb92c3994c	37.260909090909095	Obese
79266ca2-b4e3-45e5-af65-7914bd4511a0	33.644	Obese

11.Calculate the age at death for patients who have a recorded death date?

```
SELECT patients.patient, patients.birthdate, patients.deathdate, TIMESTAMPDIFF(YEAR, patients.birthdate, patients.deathdate) AS `age at death`FROM patientsWHERE patients.deathdate IS NOT NULLORDER BY `age at death` DESC;
```

patient	birthdate	deathdate	age at death
093c5b43-9b43-40a4-8cb6-ba2292cba7c0	1919-06-01	2016-03-06	96
76c2608b-8134-4316-a336-7d3a01323488	1919-04-03	2016-01-21	96
2c884d0f-62a1-4371-becc-36a98cdc4f52	1921-02-07	2016-08-22	95
cc74c757-dfb4-4c16-a442-1b73b87d419d	1921-02-26	2016-03-05	95
14d1eeda-429d-46ee-b104-72adfaca22a5	1917-10-24	2013-10-23	95
4ee2c837-e60f-4c54-9fdf-8686bc70760b	1929-04-08	2023-11-11	94
67e2a5d5-eae2-49bf-968c-c48e41de6aa1	1920-09-11	2015-09-09	94


Power BI
Medical Data Analysis - Dashboard Overview
Age
All
Gender
All
Number of visiting patients
1462
Gender Distribution
F 49%
M 51%
Mortality Rate
31%
Encounter Distribution
38%
57%
Admin
Inpatient
Outpatient
Race Distribution
asian 6%
black 9%
hispanic 11%
white 74%
Total Visits by Age Categories
Under 18 years
18-29 years
30-59 years
60+ years
2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Deaths by Age Category
259
151
18
29
60+ years
30-59 years
18-29 years
Under 18 years
Conditions with over 30% Prevalence Rate per 1000
Viral Sinusitis (Diso…
868
Acute Bronchitis…
749
Strep…
487
Spr…
367
Acute Viral Pharyn…
772
Otitis Media
699
Normal Preg…
340
QUALITY OF LIFE
PATIENT PROFILE
DASHBOARD
Top 5 Conditions
Viral sinusitis (dis…
Acute viral pharyn…
Acute bronchitis (…
Prediabetes
Hypertension
Top 5 Procedures
Colonoscopy
Measurement of r…
Suture open wound
Standard pregnan…
Throat culture (pr…
Top 10 Conditions and associated Mortality Rate
Bullet wound
Burn injury(morphol…
History of amputatio…
Non-small cell carci…
Non-small cell carci…
Non-small cell carci…
Pneumonia
Secondary malignan…
Non-small cell lung …
Suspected lung can…
0%
50%
100%
Mortality Rate
Top 5 Medication
DESCRIPTION
Penicillin V Potas…
Acetaminophen 1…
Penicillin V Potas…
Amoxicillin 250 …


Power BI
Laronda Bernier's Chart
ACTIVE
laronda
Body Weight
58,10
kg -- Normal Weight
Blood Pressure
83,00
mmHg -- Stage 1 Hypertension
Total Cholesterol
174,00
mg/dL -- Normal
Quality of Life Years
84,45
years --
Health Trends Analysis
Average BP Average Body Weight Average Cholesterol
Year
Number of Times Patient Visited the Hospital
Year
Average BMI by Year
Patient Conditions
Fracture of for...
Osteoarthritis ...
Sprain of wrist
Stroke
Whiplash injur...
Acute ba...
Acute vir...
Chronic s...
Fracture ...
Osteoart...
Osteopor...
Pathologi...
Medication Prescribed
MedicationStatus
Alendronic acid 10 MG [Fosamax] Taking
Naproxen sodium 220 MG [Aleve] Taking
Penicillin V Potassium 250 MG Taking
Acetaminophen 325 MG Oral Tablet Stopped
Naproxen sodium 220 MG Oral Tablet Stopped
QUALITY OF LIFE
DASHBOARD
PATIENT PROFILES
Laronda Bernier
Current Patient Details
Demographigic Information
98
Age
F
Gender
asian
Race
M
Marital
X38505992X
ID Number
Patient's address
5602 Will Islands Millbury MA 01527 US
Contact Number
999-82-1104
Allergies
Shellfish allergy

QUALITY OF LIFE MODEL PREDICTION

CREATION OF SUB-DATASET

```
SELECT    patient_id,    gender,    marital_status,    age,    bmi_value,    qaly_value,    coronary_heart_disease,    diabetes,
CASE      WHEN age <> 0 THEN qaly_value / age        ELSE NULL  -- Handle division by zero if age is zero    END AS HRQLFROM    (
SELECT          p.patient AS patient_id,          p.gender AS gender,          p.marital AS marital_status,          TIMESTAMPDIFF(YEAR,
p.birthdate, LEAST('2017-12-31', COALESCE(p.deathdate, '2017-12-31'))) AS age,          COALESCE(o1.value, 0) AS bmi_value,
COALESCE(o2.value, 0) AS qaly_value,          CASE WHEN c1.description = 'Coronary Heart Disease' THEN 1 ELSE 0 END AS
coronary_heart_disease,          CASE WHEN c2.description = 'Diabetes' THEN 1 ELSE 0 END AS diabetes       FROM          patients p
LEFT JOIN          observations o1 ON p.patient = o1.patient          AND o1.description = 'Body Mass Index'       LEFT JOIN
observations o2 ON p.patient = o2.patient          AND o2.description = 'Quality adjusted life years'       LEFT JOIN          conditions c1
ON p.patient = c1.patient          AND c1.description = 'Coronary Heart Disease'       LEFT JOIN          conditions c2 ON p.patient =
c2.patient          AND c2.description = 'Diabetes'       WHERE          o1.description IS NOT NULL OR o2.description IS NOT NULL
OR c1.description IS NOT NULL OR c2.description IS NOT NULL    ) AS summary_table;
```

patient_id	gender	marital_status	age	bmi_value	qaly_value	coronary_heart_disease	diabetes	HRQL
00269bb7-e3ab-43a9-9cdf-cdf9b6e3b2b3	F	NULL	10	0	0	0	0	0
00341a88-1cc1-4b39-b0f9-05b0531991a0	F	S	45	42.54	42.66364...	0	0	0.94...
00341a88-1cc1-4b39-b0f9-05b0531991a0	F	S	45	41.45	42.66364...	0	0	0.94...

INSIGHTS FROM PREDICTIVE AND PRESCRIPTIVE ANALYTICS

```
# import data
import pandas
from sklearn import linear_model
df = pandas.read_csv ('QALYs_data.csv')
df
```

	patient_id	gender	marital_status	age	bmi_value	qaly_value	coronary_heart_disease	diabetes	HRQL
0	a1851c06-804e-4f31-9d8f-388cd52d4ad0	F	M	62	37.68	0.000000	1	0	0.000000
1	a1851c06-804e-4f31-9d8f-388cd52d4ad0	F	M	62	38.40	0.000000	1	0	0.000000
2	a1851c06-804e-4f31-9d8f-388cd52d4ad0	F	M	62	38.86	0.000000	1	0	0.000000

```
#convert categorical values to numeric representation
df['gender'] = df['gender'].astype('category')
df['gender'] = df['gender'].cat.codes + 1
#Female is 1 and Male is 2

df['marital_status'] = df['marital_status'].astype('category')
df['marital_status'] = df['marital_status'].cat.codes + 1
#Single is 1 and Married is 2
```

```
#drop duplicates
df.sort_values(by='patient_id', ascending=True,
inplace=True)

df.drop_duplicates(subset='patient_id', keep='last',
inplace=True)
```

patient_id	gender	marital_status	age	bmi_value	qaly_value	coronary_heart_disease	diabetes	HRQL
9b6e3b2b3	1	0	10	0.00	0.000000	0	0	0.000000
0531991a0	1	2	45	38.50	42.663646	0	0	0.948081
607044c99	2	0	3	14.89	2.000000	0	0	0.666667
0ac257cbf4	2	0	14	14.46	14.000000	0	0	1.000000
a93822fa11	1	0	19	20.46	18.764392	0	0	0.987600

```
#fill NULLs with 0

df['HRQL'] = df['HRQL'].fillna(0)
df.isnull().sum()
```

INSIGHTS FROM PREDICTIVE AND PRESCRIPTIVE ANALYTICS

```
#ready for modelling - predicting HRQL with several training features
columns_to_drop = ['patient_id', 'HRQL','qaly_value']
X = df.drop(columns=columns_to_drop)
X
```

gender	marital_status	age	bmi_value	coronary_heart_disease	diabetes
1	0	10	0.00	0	0
1	2	45	38.50	0	0
2	0	3	14.89	0	0

```
#Our label is HRQL
y = df['HRQL']
y
```

```
0.000000
0.948081
0.666667
1.000000
```

```
#split dataset for training and testing
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)

from sklearn.linear_model import LinearRegression
lr = LinearRegression()

lr.fit(X_train, y_train)
```

```
m = lr.coef_
m
```

```
array([ 0.01045343, -0.03734285, -0.00830805,  0.02179931, -0.11165309,
       -0.000845  ])
```

INSIGHTS FROM PREDICTIVE AND PRESCRIPTIVE ANALYTICS

```
#get tools for OLS regression analysis
pip install statsmodels

import pandas as pd
from sklearn.model_selection import train_test_split
import statsmodels.api as sm
import matplotlib.pyplot as plt

# Initialize and fit the OLS model using statsmodels
ols_model = sm.OLS(y_train, X_train).fit()

# Make predictions on test data
y_pred_test = ols_model.predict(X_test)

# Print the summary to check for significance of the coefficients
print(ols_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   HRQL   R-squared:                       0.602
Model:                            OLS   Adj. R-squared:                  0.600
Method:                 Least Squares   F-statistic:                     256.2
Date:                Tue, 21 May 2024   Prob (F-statistic):          2.41e-199
Time:                        17:55:44   Log-Likelihood:                -185.63
No. Observations:                1023   AIC:                             385.3
Df Residuals:                    1016   BIC:                             419.8
Df Model:                           6
Covariance Type:            nonrobust

                              coef    std err          t      P>|t|
-------------------------------------------------------------------
const                       0.6149      0.040     15.372      0.000
gender                      0.0105      0.018      0.569      0.569
marital_status             -0.0373      0.019     -2.006      0.045
age                        -0.0083      0.000    -17.579      0.000
bmi_value                   0.0218      0.001     33.444      0.000
coronary_heart_disease     -0.1117      0.040     -2.813      0.005
diabetes                   -0.0008      0.037     -0.023      0.982
```

```
#Visualize the predicted vs. actual HRQL using scatter plot
plt.scatter(y_test, y_pred_test)
plt.xlabel("Actual HRQL")
plt.ylabel("Predicted HRQL")
plt.title("Linear Regression: Actual vs. Predicted HRQL")
plt.show()
```

Quality of Life Assessment

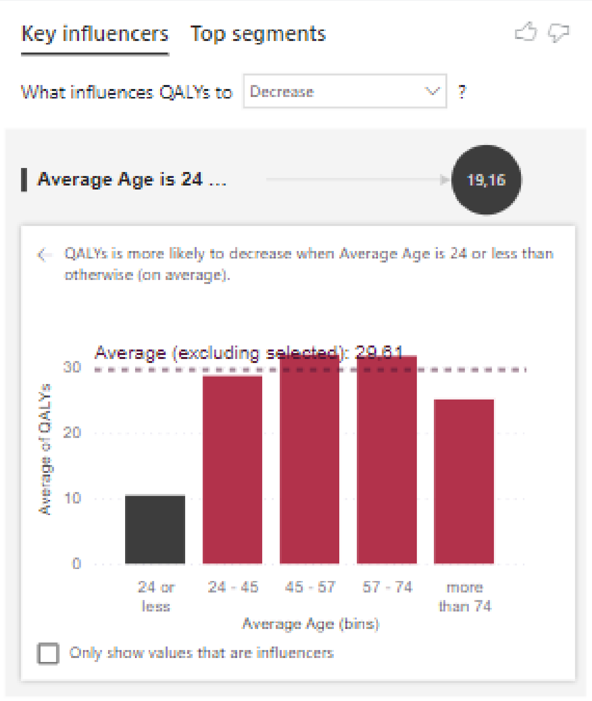
We built a predictive model for quality of life to assess the impact of various lifestyle behaviors on quality-adjusted life years (QALYs), a crucial measure in health economics. Our data has QALYs data, which is a product of the patient's life years and health-related quality of life (HRQL). Leveraging insights from Noto et al. (2021) regarding the significance of physical exercise and health on HRQL, as well as findings from L, we seek to elucidate the impact of these health metrics on QALYs. [1]. Our analysis incorporates intervention studies targeting individuals with obesity [2], cardiovascular disease [3], and diabetes [4], aiming to discern the clarity of the relationship between these conditions and quality of life. We control for age, gender and marital status to isolate the effects of lifestyle behaviors.

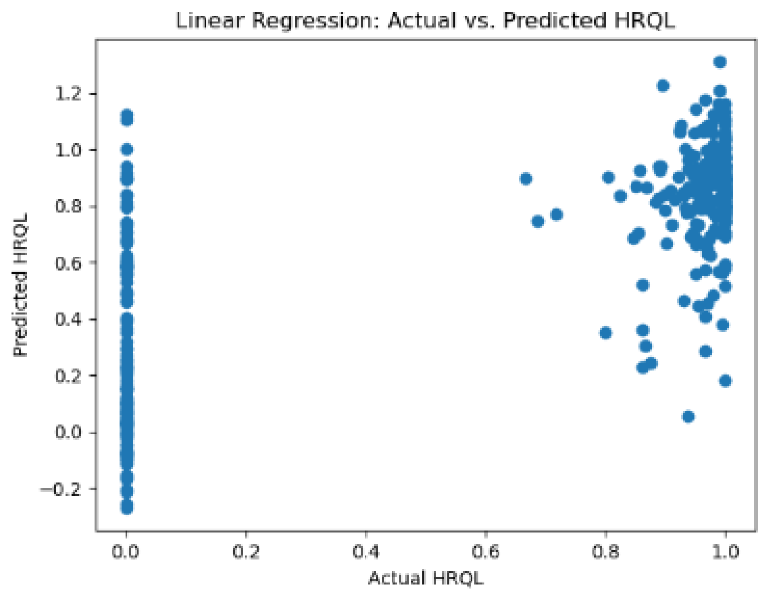
We propose a ordinary least squares regression model where the health-related quality of life is a function of control variables gender, marital status and age, and predictors, including BMI, coronary heart disease (CD) and diabetes (Diab) - both dichotomous variables, with respective coefficients b0, b1, b2, b3, b4, and b5. The standard error (SE) term captures the variability unexplained by the model. This model aims to provide insights into how lifestyle behaviors influence the overall quality of life, as measured by QALYs.

(1) QALYs = Life years X HRQL , where HRQL is between 0 and 1 is utility of a certain health state rating between 0 (being dead) and 1 (perfect health) [5]

The linear regression model is defined as :

(2) $$HRQL = c + b0(gender) + b1(martial\ status) + b2(age) + b3(BMI) + b4(CD) + b5(Diab) + SE$$





r2_score = 0.602. F-statistic = 256.2
b_coeffients = array([0.0105 (p = 0.569), -0.0373 (p = 0.045), -0.0083 (p = 0.0), 0.0218 (p = 0.01), -0.1117 (p = 0.005), -0.0008 (p = 0.982)])

KEY INSIGHTS

Interpretation of the OLS regression results:
1. 60.2% of the variance in the HRQL is explained by the independent variables in the model.
2. A high F-statistic with a very low p-value suggests that the **model is statistically significant**.
3. Gender is not a statistically significant predictor despite some studies suggesting that male gender is associated with better HRQL.
4. **Marital status is statistically significant and a change from single to married status shows a decrease in HRQL**
5. **Age is also statistically significant and each additional year decreases HRQL**, echoing studies that suggest that older individuals may experience lower HRQL compared to younger counterparts.
6. The **BMI coefficient is highly significant** but an increase in BMI increases the HRQL contradicting literature.
7. **Coronary heart disease has an inverse relationship with HRQL and is highly significant.**
8. Having diabetes has a negligible effect on HRQL and has no statistical significance.

Model Limitations:
* does not account for other significant factors affecting HRQL, e.g lifestyle choices and various exogenous and endogenous influences and further categorization .
* the data distribution is not normal and is left-skewed, data preprocessing could improve accuracy and predictive power.

Overall, while the model provides significant insights into the factors affecting HRQL, addressing the identified limitations and incorporating additional relevant variables could enhance its explanatory power and accuracy.

While life years serve as an objective metric, unlike HRQL which is subjective, their influence on QALYs is not uniform across all age groups. Specifically, individuals below the age of 24 and those over 74 years old exhibit a disproportionally negative impact on QALYs. Conversely, HRQL exerts a more pronounced effect on QALYs among very young and elderly populations. This nuanced relationship underscore the importance of considering both life years and HRQL when evaluating overall quality of life and health outcomes across different categories.

RECOMMEDATIONS AND APPLICATIONS

Outcomes from this study can be summarized with the following

CLINICAL DECISION SUPPORT

Data provides insights to healthcare providers for better diagnosis, treatment planning, and patient management

HEALTHCARE RESOURCE ALLOCATION

Data informs resource allocation strategies based on prevalent conditions and patient needs.

IMPACT OF LIFESTYLE BEHAVIOURS ON QUALITY OF LIFE

Predicting factors influencing patient quality of life and provide actionable insights for enhanced patient care

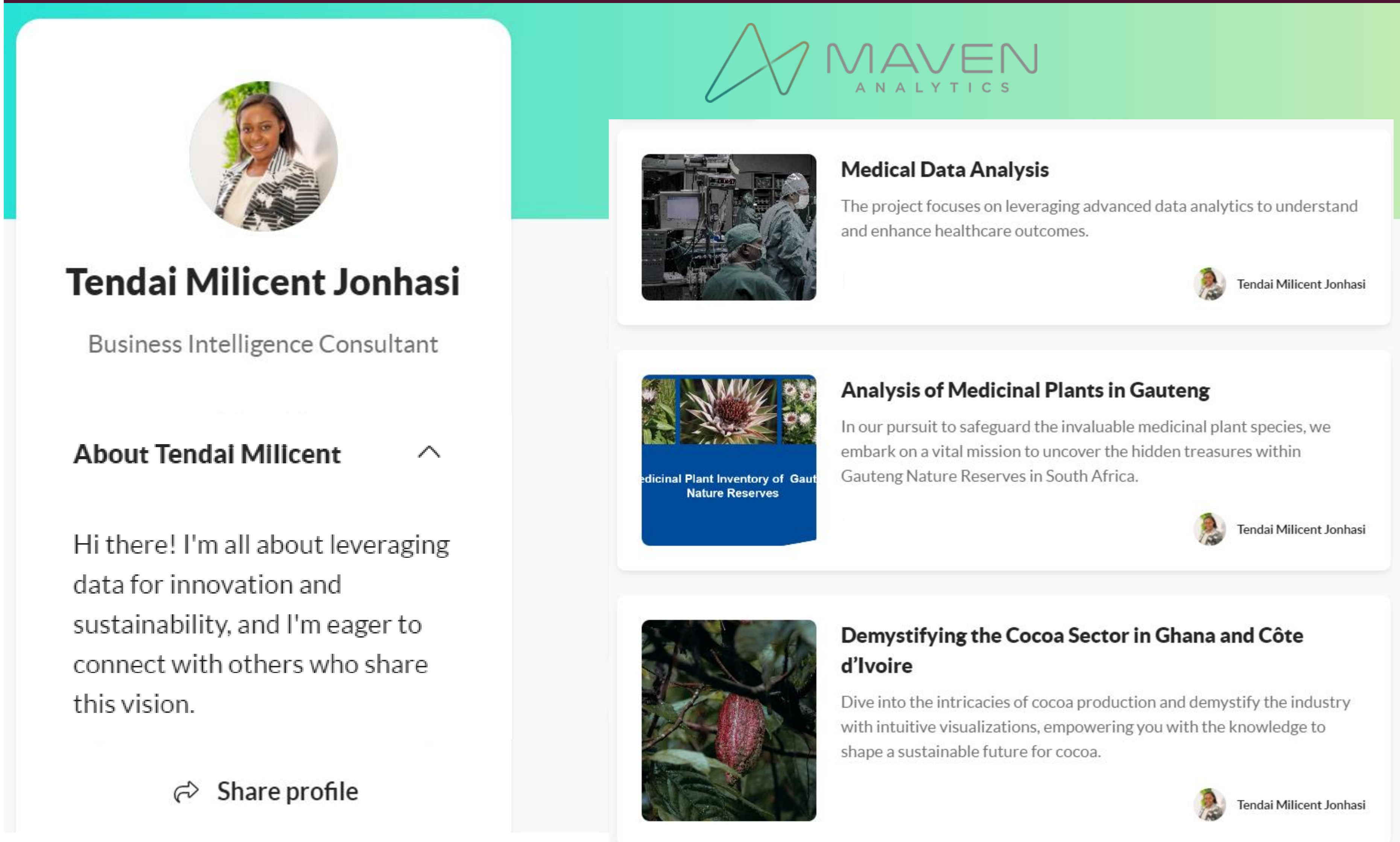
PATIENT-CENTERED CARE

Personalize care plans and interventions based on individual patient profiles and risk factors.

RESEARCH AND POLICY IMPLICATIONS

Contribute to medical research and health policy decisions by highlighting key healthcare trends and challenges

CHECK OUT THE REST OF MY PORTFOLIO!



Interact with Power BI Dashboards on my Maven Analytics Portfolio page

https://mavenanalytics.io/profile/Tendai%20Milicent-Jonhasi/199912327

TENDAI MILICENT JONHASI

tendaimjonhasi@gmail.com

https://www.linkedin.com/in/tendai-milicent-jonhasi/

https://github.com/tendaimjay

https://mavenanalytics.io/profile/Tendai%20Milicent-Jonhasi/199912327