HealthConnect Data Analysis Report

1. Introduction

Objective:

HealthConnect aims to utilize data analysis to identify trends in patient health conditions, treatment effectiveness, and geographical distribution. The goal is to provide insights that help tailor healthcare services for more efficient outcomes. This report focuses on analyzing patient demographics, treatment success rates, and common health conditions based on the organization's patient database.

Dataset Overview:

The dataset includes anonymized patient information containing the following variables:

- **Demographic Information**: Age, gender
- **Health Information**:, medical condition, medication, test results

---

2. Data Cleaning and Preprocessing

Steps Taken:

1. **Standardized Name Column**: I applied proper functions to the "Name" column to standardize text formatting and removed any commas present in the names for consistency.
2. **Checked for Blank Spaces**: I checked the dataset for any blank spaces but found none.
3. **Corrected Hospital Column Inconsistencies:** I addressed any inconsistencies in the "Hospital" column to ensure uniform naming conventions.
4. **Rounded Billing Amount:** I rounded up the "Billing Amount" to a whole number and converted its data type into currency format for consistency.
5. **Removed Duplicate Rows**: I removed duplicate entries and ensured that each row in the dataset is unique.
6. **Converted Date Columns**: I converted both the "Date of Admission" and "Discharge Date" columns to date format for consistency and easier analysis.
7. **Checked for Outliers**: I checked for outliers using the Z-score method in the "Billing Amount", "Age", and "Room Number" columns. No significant outliers were detected.
8. **Created new column**: Age Group.

---

3. Data Analysis

Descriptive Statistics:

1. **Patient Demographics Overview**:

- **Age Distribution**: The majority of patients fall between the ages of 20-80, with a peak in the 50-59 range.
- **Gender Distribution**: Approximately 50.02% of patients are female and 49.08% are male.

Gender Distribution of Medical Conditions

a. Arthritis:

- **Females**: 4,251 cases
- **Males**: 4,188 cases
- **Observation**: Slightly more common in females, with a marginal difference of 63 cases.

b. Asthma:

- **Females**: 4,103 cases
- **Males**: 4,169 cases
- **Observation**: Asthma is slightly more common in males, with a difference of 66 cases.

c. Cancer:

- **Females**: 4,166 cases
- **Males**: 4,128 cases
- **Observation**: Cancer appears marginally more prevalent in females, with a difference of 38 cases.

d. Diabetes:

- **Females**: 4,201 cases
- **Males**: 4,183 cases
- **Observation**: Diabetes shows almost equal prevalence, with only 18 more cases in females.

e. Hypertension:

- **Females**: 4,157 cases
- **Males**: 4,162 cases
- **Observation**: Hypertension is nearly equally distributed, with males having 5 more cases.

f. Obesity:

- **Females**: 4,133 cases
- **Males**: 4,159 cases
- **Observation**: Slightly more cases of obesity in males, with a difference of 26 cases.

Total Cases by Gender:

Count of Medical Cond	Medical Conc						
Gender	Arthritis	Asthma	Cancer	Diabetes	Hypertension	Obesity	Grand Total
Female	4251	4103	4166	4201	4157	4133	25011
Male	4188	4169	4128	4183	4162	4159	24989
Grand Total	**8439**	**8272**	**8294**	**8384**	**8319**	**8292**	**50000**

- **Females**: 25,011 total cases
- **Males**: 24,989 total cases
- **Observation**: Overall, females account for slightly more total cases than males, but the difference is negligible (22 cases).

Insights:

- Most conditions show a balanced distribution between genders, with minor variations.
- Arthritis and cancer lean slightly toward females, while asthma and obesity are marginally higher in males.
- Diabetes and hypertension have near-equal gender prevalence.

2. **Common Health Conditions**:
    - The most common diagnoses include:
        - **Arthritis**: 16.9% of the total patients
        - **Asthma**: 16.5% of the total patients
        - **Cancer**: 16.6% of the total patients
        - **Diabetes**: 16.8% of the total patients
        - **Obesity**: 16.6% of the total patients
        - **Hypertension**: 16.6% of the total patients
3. **Treatment Success Rates**:
    - **Penicillin**: Highest success in treating asthma (35%), lowest in treating cancer (31.5%).
    - **Aspirin**: Highest success in treating hypertension (35%), lowest in treating arthritis (32.3%).
    - **Ibuprofen**: Highest success in treating asthma (35.2%), lowest in treating arthritis (31.6%).
    - **Lipitor**: Highest success in treating arthritis (33.9%), lowest in treating hypertension (31.4%).
    - **Paracetamol**: Highest success in treating asthma (34.1%), lowest in treating diabetics (31.8%).

      Key Insights:

- **Treatment Effectiveness**: Certain treatments (e.g., Ibuprofen for asthma) show higher success rates than others.
- **Age Distribution and Medical Condition**

Count of Age group	Age Group								
Medical Conditions	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	Grand Total
Arthritis	261	1179	1287	1222	1287	1188	1255	760	8439
Asthma	244	1194	1214	1205	1207	1235	1228	745	8272
Cancer	242	1217	1241	1198	1228	1218	1202	748	8294
Diabetes	238	1207	1229	1263	1251	1275	1206	715	8384
Hypertension	248	1197	1203	1219	1259	1202	1265	726	8319
Obesity	237	1238	1186	1224	1293	1233	1148	733	8292
Grand Total	**1470**	**7232**	**7360**	**7331**	**7525**	**7351**	**7304**	**4427**	**50000**

- **Age Groups with Highest and Lowest Totals**:

  Highest Total Count: Age group **50-59** with **7525** patients.

  Lowest Total Count: Age group **10-19** with **1470** patients.

1. Arthritis:

- **Peak Age Group**: **30-39** with **1287** patients.
    - This suggests that arthritis begins to significantly affect people in their 30s and continues steadily into older age.
- **Decline**: In the **80-89** age group, the count drops to **760**, potentially due to mortality or a reduction in reporting among the elderly.

2. Asthma:

- **Peak Age Group**: **20-29** with **1194** patients.
    - This indicates asthma is more common among younger adults and may reduce in prevalence or reporting as individuals age.
- **Decline**: Drops to **745** in the **80-89** group, showing a lower prevalence in the older population.

3. Cancer:

- **Peak Age Group**: **50-59** with **1228** patients.
    - Cancer prevalence rises significantly as individuals age, peaking in their 50s and remaining high through the 60s and 70s.
- **Decline**: Falls to **748** in the **80-89** age group, possibly due to lower survival rates or reporting in this age bracket.

4. Diabetes:

- **Peak Age Group**: **60-69** with **1275** patients.
    - Diabetes is more common among older adults, particularly those in their late middle ages and early senior years.
- **Decline**: Drops to **715** in the **80-89** group.

5. Hypertension:

- **Peak Age Group**: **70-79** with **1265** patients.
    - Hypertension is most prevalent among seniors, reaching its highest in the 70-79 range.
- **Decline**: Similar to other conditions, the count reduces to **726** in the **80-89** group.

6. Obesity:

- **Peak Age Group**: **50-59** with **1293** patients.
    - Obesity is most reported in middle-aged adults, possibly due to lifestyle factors.
- **Decline**: Drops to **733** in the **80-89** group, indicating reduced prevalence in older age groups, possibly due to weight loss with aging or under-reporting.

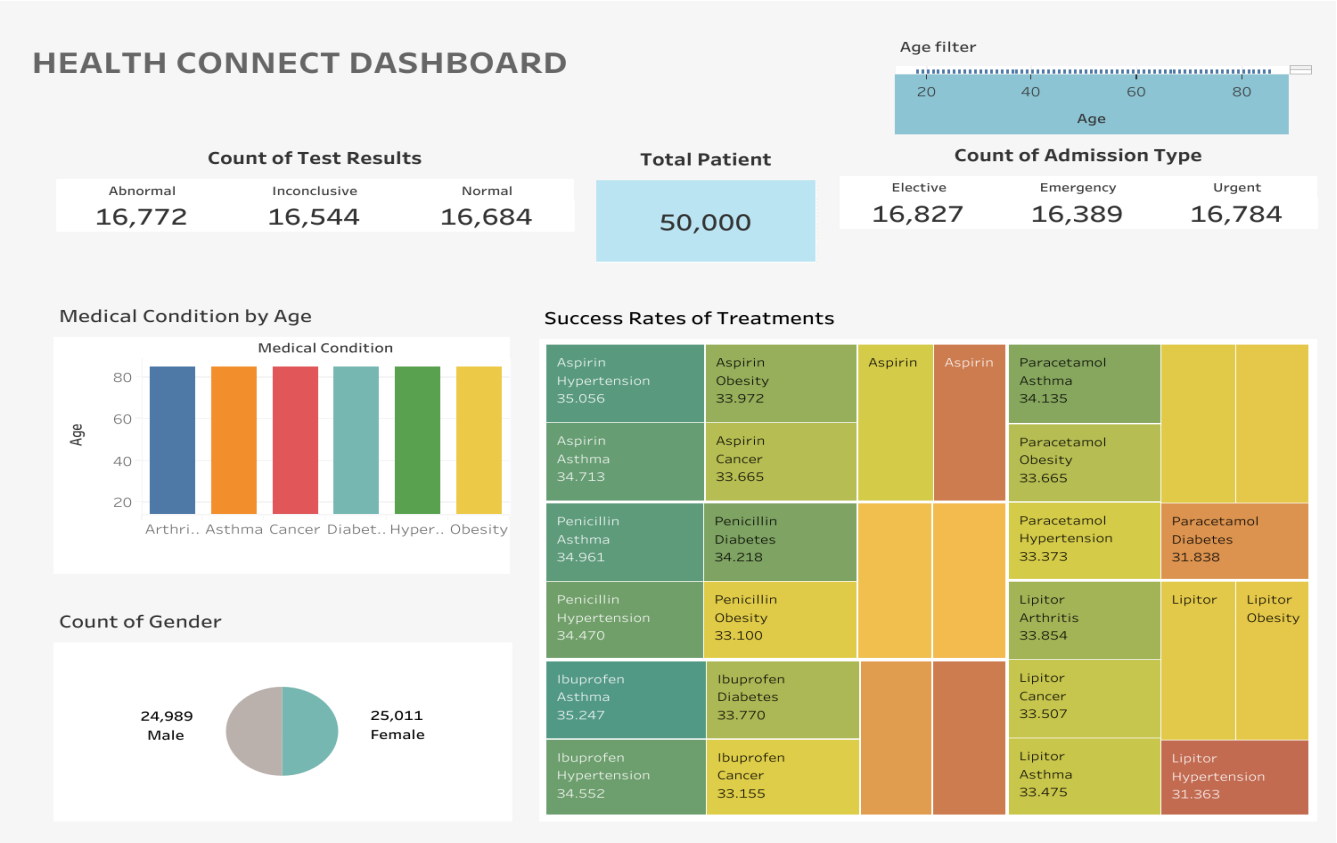
General Observations:

- Across all conditions, the **20-69** age range sees the majority of cases, showing these ages are more prone to chronic health conditions.
- The **80-89** age group consistently has the lowest counts, which may reflect a combination of natural population attrition, lower diagnosis rates, or reduced healthcare engagement at advanced ages.

---

4. Data Visualization

Visualizations Created:



1. **Patient Demographics Distribution**:
    - **Bar Chart**: Showcased the count of medical condition of patient by age.
    - **Pie Chart**: Displayed the gender distribution, confirming a slight female majority in the dataset.
2. **Treatment Success Rates**:
    - **Treemap**: Displayed success rates by Medical Condition and Medication, emphasizing areas for potential improvement.
3. **Key Performance Indicators:**
    - Showing count of admission types and count of test results.

---

5. Advanced Analytics

Predictive Modeling:

Decision Trees: A decision tree model was used to classify patients by their likelihood of Test Result, based on Age, Medical Condition and Medication.

- Age is the most important feature, contributing 65% to the decision tree model's ability to predict patient outcomes. This suggests that the patient's age has a strong relationship with the predicted outcome and is heavily relied upon by the model for decision splits.
- Medical Condition contributes 22% to the decision tree's predictions. While less influential than age, it is still a significant factor in determining patient outcomes. This shows that specific medical conditions have a moderate impact on the outcomes.
- Medication is the least important feature, contributing 13%. This indicates that the type of medication prescribed plays a relatively smaller role in predicting outcomes, at least in the current dataset and model.

---

6. Key Findings

- **Common Health Conditions**: Arthritis, asthma, diabetes, and obesity are the most prevalent conditions, with asthma showing the highest success rate in treatments.
- **Treatment Effectiveness**: Ibuprofen shows the highest success rate for asthma, whereas Lipitor has lowest success rates for Hypertension.
- Most conditions show a balanced distribution between genders, with minor variations.
- Arthritis and cancer lean slightly toward females, while asthma and obesity are marginally higher in males.
- Diabetes and hypertension have near-equal gender prevalence.

---

7. Recommendations

- **Targeted Healthcare Programs**: Focus on improving treatment strategies for conditions like Arthritis, where treatment success rates are lower.