HEALTHCARE DATA ANALYSIS AND INSIGHTS

Coding Mentor: Janani Lakshmi Narayanan
Student Name: Parameswari Manthiramoorthi

JULY 24, 2024
ENTRI ELEVATE
Tn-Dsml-May24-Batch4

Table of Contents

1. Introduction

This project aims to unlock actionable insights from extensive healthcare datasets, focusing on patient health profiles, medical histories, and healthcare costs. By employing advanced data analytics techniques, we strive to provide healthcare stakeholders with evidence-based recommendations for improving patient care, optimizing resource allocation, and managing costs effectively. Through rigorous data cleaning, transformation, and analysis, this report aims to contribute to informed decision-making in the healthcare sector, driving towards enhanced clinical outcomes and operational efficiencies.

2. Problem Statement

- **Data Abundance**: The healthcare industry generates vast datasets daily, comprising medical examinations, hospitalization records, and patient profiles.
- **Objectives**: This project aims to analyze these datasets to extract insights into patient health profiles, medical histories, and healthcare costs.
- **Analytical Focus**: The focus is on exploring relationships between various health metrics, identifying trends, and visualizing patterns within the data.
- **Stakeholder Impact**: The ultimate goal is to provide actionable insights that can empower healthcare providers and policymakers to enhance patient care, optimize resource allocation, and manage healthcare costs effectively.
- **Methodology**: The project will employ rigorous data cleaning, transformation, exploration, and analysis techniques to derive meaningful conclusions from the data.
- **Outcome**: By uncovering these insights, the project seeks to contribute to informed decision-making in healthcare, aiming to improve overall healthcare outcomes and operational efficiencies.

3. Dataset Description

- **Source:**
  https://drive.google.com/uc?export=download&id=1zelh7bZrE7F290QtTABHgHYn4B7JDbZO

- **Contents**: It includes three main tables:

  - **Medical Examinations**: Contains detailed records of medical tests, patient health metrics (e.g., BMI, HbA1C), and medical histories (e.g., heart issues, cancer history).

  - **Customer Names**: Provides demographic details such as customer IDs and names, which are split into titles, first names, and last names.

  - **Hospitalization Details**: Includes records of hospital visits, charges, hospital and city tiers, and associated demographic information (e.g., state IDs).

4. Data Cleaning

❖ Objective: Ensure data integrity and consistency for accurate analysis.

Step	Description
1	**4.1 Check for Missing Values** Identify and handle missing values marked with '?' in each column of "Medical Examinations" and "Hospitalization Details" tables. ***FORMULA: =COUNTIF (****column range, "?")* MEDICAL EXAMINATIONS TABLE HOSPITALISATION DETAILS TABLE

Customer ID	0
BMI	0
HBA1C	0
Heart Issues	0
Any Transplants	0
Cancer history	0
NumberOfMajorSurgeries	0
smoker	2
Total number of missing values marked with '?' in each column	2

Customer ID	6
year	2
month	3
date	0
children	0
charges	0
Hospital tier	1
City tier	1
State ID	2
Total number of missing values marked with '?' in each column	15

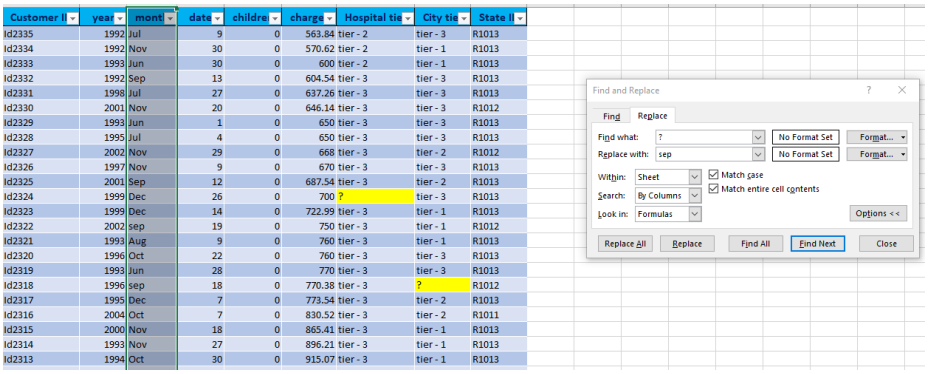
Step	Description
2	

4.2 Fill Missing Values

Impute missing values for ‘month’ with 'Sep' and ‘year’ with the rounded average value.

4.2.1 Fill missing values for ‘month’ column

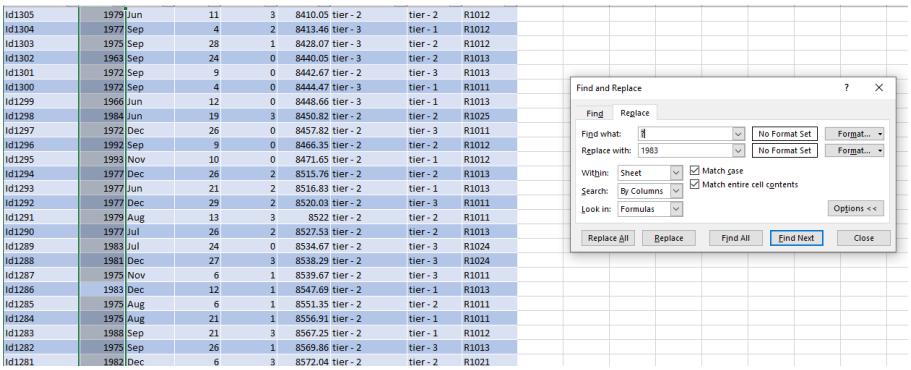
- No of missing values in month column is 3.
- Use Find and Replace (CTRL+H) for '?' to replace with 'Sep'.
- 3 values replaced.



Customer Id	year	month	date	children	charge	Hospital tier	City tier	State Id
Id2335	1992	Jul	9	0	563.84	tier - 2	tier - 3	R1013
Id2334	1992	Nov	30	0	570.62	tier - 2	tier - 1	R1013
Id2333	1993	Jun	30	0	600	tier - 2	tier - 1	R1013
Id2332	1992	Sep	13	0	604.54	tier - 3	tier - 3	R1013
Id2331	1998	Jul	27	0	637.26	tier - 3	tier - 3	R1013
Id2330	2001	Nov	20	0	646.14	tier - 3	tier - 3	R1012
Id2329	1993	Jun	1	0	650	tier - 3	tier - 3	R1013
Id2328	1995	Jul	4	0	650	tier - 3	tier - 3	R1013
Id2327	2002	Nov	29	0	668	tier - 3	tier - 2	R1012
Id2326	1997	Nov	9	0	670	tier - 3	tier - 3	R1013
Id2325	2001	Sep	12	0	687.54	tier - 3	tier - 2	R1013
Id2324	1999	Dec	26	0	700	?	tier - 3	R1013
Id2323	1999	Dec	14	0	722.99	tier - 3	tier - 1	R1013
Id2322	2002	sep	19	0	750	tier - 3	tier - 1	R1012
Id2321	1993	Aug	9	0	760	tier - 3	tier - 1	R1013
Id2320	1996	Oct	22	0	760	tier - 3	tier - 3	R1013
Id2319	1993	Jun	28	0	770	tier - 3	tier - 3	R1013
Id2318	1996	sep	18	0	770.38	tier - 3	?	R1012
Id2317	1995	Dec	7	0	773.54	tier - 3	tier - 2	R1013
Id2316	2004	Oct	7	0	830.52	tier - 3	tier - 2	R1011
Id2315	2000	Nov	18	0	865.41	tier - 3	tier - 1	R1013
Id2314	1993	Nov	27	0	896.21	tier - 3	tier - 1	R1013
Id2313	1994	Oct	30	0	915.07	tier - 3	tier - 1	R1013

4.2.2 Fill missing values for ‘year’ column

- Formula: *=ROUND (AVERAGE (B2:B2344),0)*
- No of missing values in Year column = *2*.
- Use Find and Replace (CTRL+H) for '?' to replace with 1983.
- 2 values Replaced.



Id1305	1979	Jun	11	3	8410.05	tier - 2	tier - 2	R1012
Id1304	1977	Sep	4	2	8413.46	tier - 3	tier - 1	R1012
Id1303	1975	Sep	28	1	8428.07	tier - 3	tier - 2	R1012
Id1302	1963	Sep	24	0	8440.05	tier - 3	tier - 2	R1013
Id1301	1972	Sep	9	0	8442.67	tier - 2	tier - 3	R1013
Id1300	1972	Sep	4	0	8444.47	tier - 3	tier - 1	R1011
Id1299	1966	Jun	12	0	8448.66	tier - 3	tier - 1	R1013
Id1298	1984	Jun	19	3	8450.82	tier - 2	tier - 2	R1025
Id1297	1972	Dec	26	0	8457.82	tier - 2	tier - 3	R1011
Id1296	1992	Sep	9	0	8466.35	tier - 2	tier - 2	R1012
Id1295	1993	Nov	10	0	8471.65	tier - 2	tier - 1	R1012
Id1294	1977	Dec	26	2	8515.76	tier - 2	tier - 2	R1013
Id1293	1977	Jun	21	2	8516.83	tier - 2	tier - 1	R1013
Id1292	1977	Dec	29	2	8520.03	tier - 2	tier - 3	R1011
Id1291	1979	Aug	13	3	8522	tier - 2	tier - 2	R1011
Id1290	1977	Jul	26	2	8527.53	tier - 2	tier - 2	R1013
Id1289	1983	Jul	24	0	8534.67	tier - 2	tier - 3	R1024
Id1288	1981	Dec	27	3	8538.29	tier - 2	tier - 3	R1024
Id1287	1975	Nov	6	1	8539.67	tier - 2	tier - 3	R1011
Id1286	1983	Dec	12	1	8547.69	tier - 2	tier - 1	R1013
Id1285	1975	Aug	6	1	8551.35	tier - 2	tier - 2	R1011
Id1284	1975	Aug	21	1	8556.91	tier - 2	tier - 1	R1011
Id1283	1988	Sep	21	3	8567.25	tier - 2	tier - 1	R1012
Id1282	1975	Sep	26	1	8569.86	tier - 2	tier - 3	R1013
Id1281	1982	Dec	6	3	8572.04	tier - 2	tier - 2	R1021

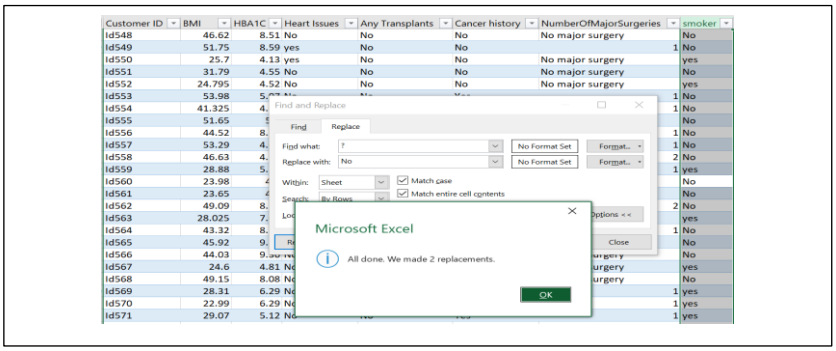
Step	Description
3	

4.3 Determine The Most Frequently Occurring Values

Determine the most frequently occurring values in the ‘smoker’, 'Hospital tier' and 'City tier' columns.

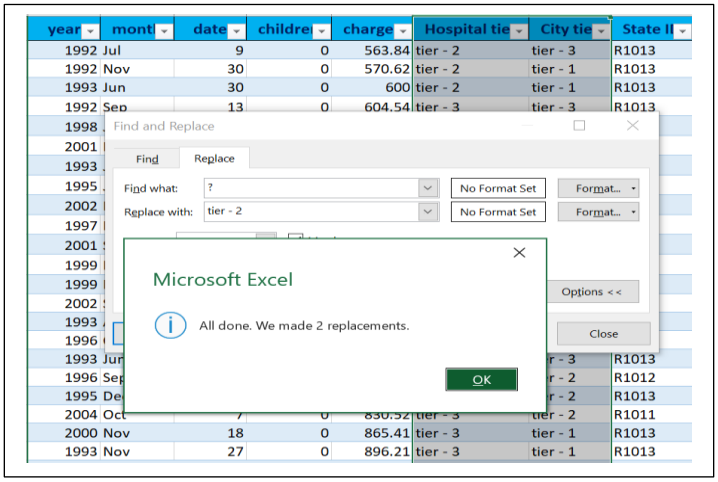
4.3.1 Fill missing values for ‘Smoker’ column

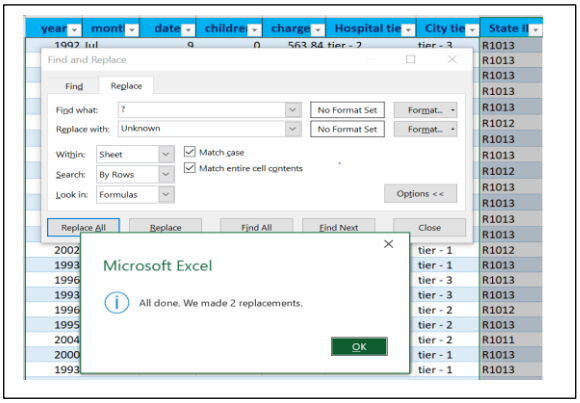
- The most frequently occurring values in the ‘smoker’ = *No*
  - *=INDEX(H2:H2336,MODE(MATCH(H2:H2336,H2:H2336,0)))*



4.3.2 Fill missing values for ‘hospital tier’ column and ‘city tier’ column

- The most frequently occurring values in the 'hospital tier' = tier - 2
- *=INDEX(G2:G2336,MODE(MATCH(G2:G2336,G2:G2336,0)))*
- The most frequently occurring values in the 'city tier' = = tier - 2
- *=INDEX(H2:H2344,MODE(MATCH(H2:H2344,H2:H2344,0)))*

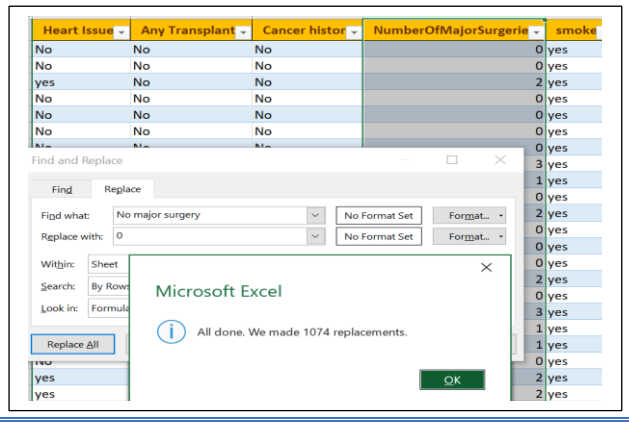


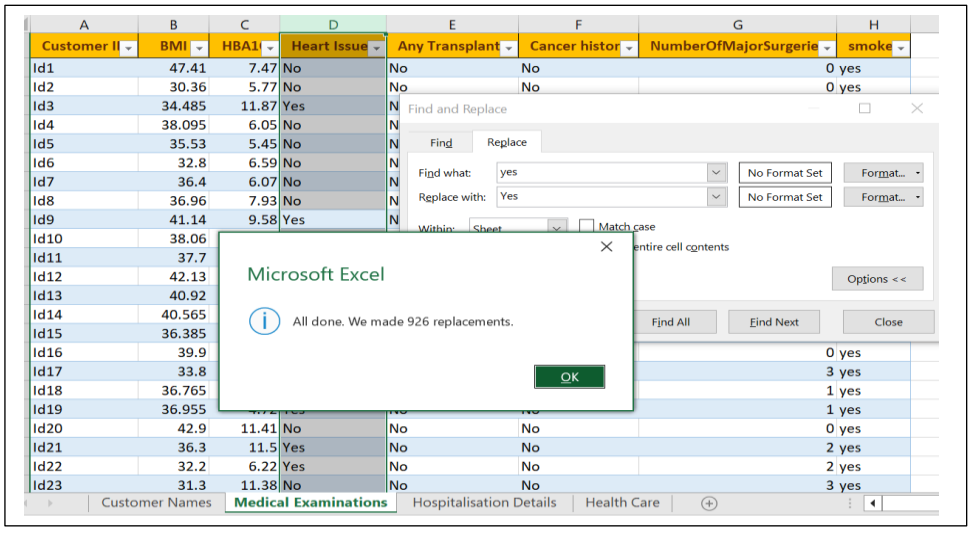
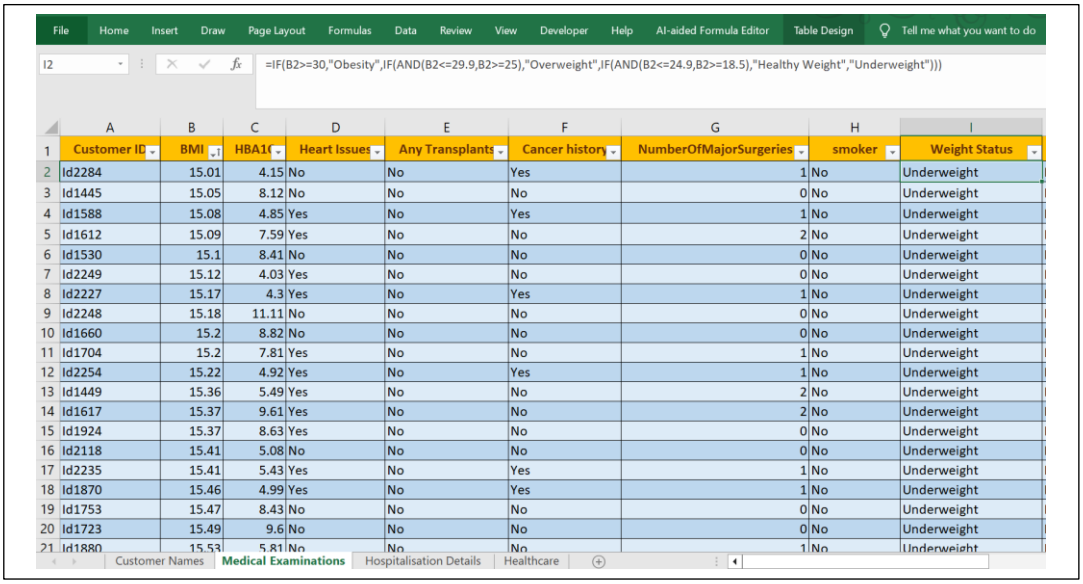
Step	Description
4	### 4.4 Fill missing State ID values Strategically address missing 'State ID' values to maintain data completeness. ❖ No of missing 'State ID' values are =2 ❖ Use Find and Replace (CTRL+H) for '?' to replace with 'Unknown'. ❖ 2 values Replaced. 

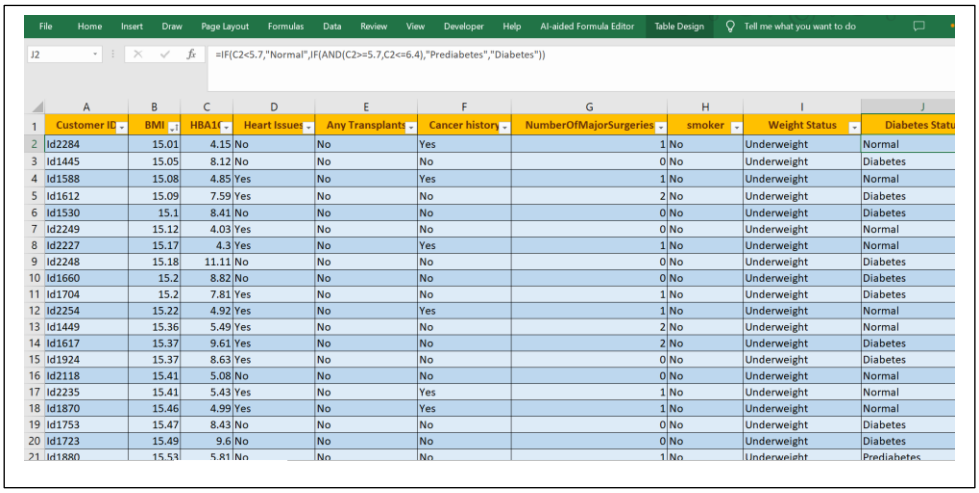
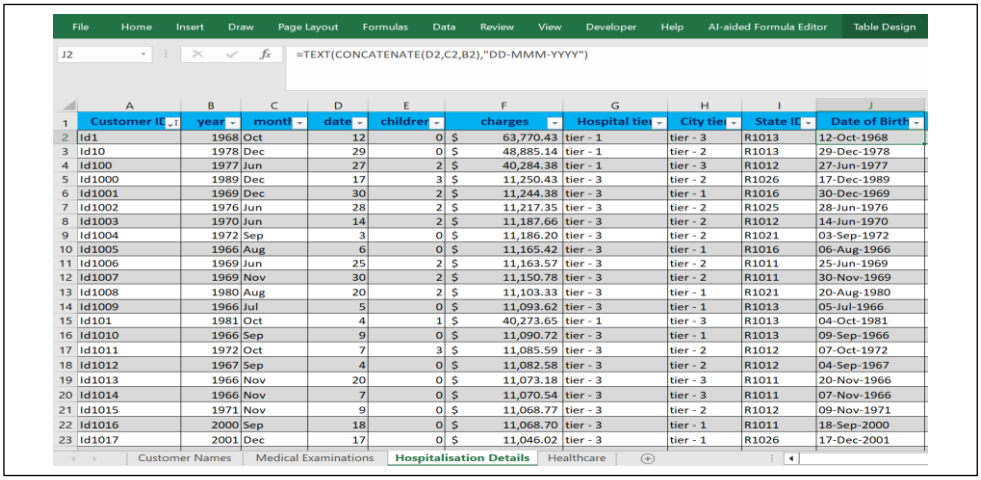
5. Data Transformation

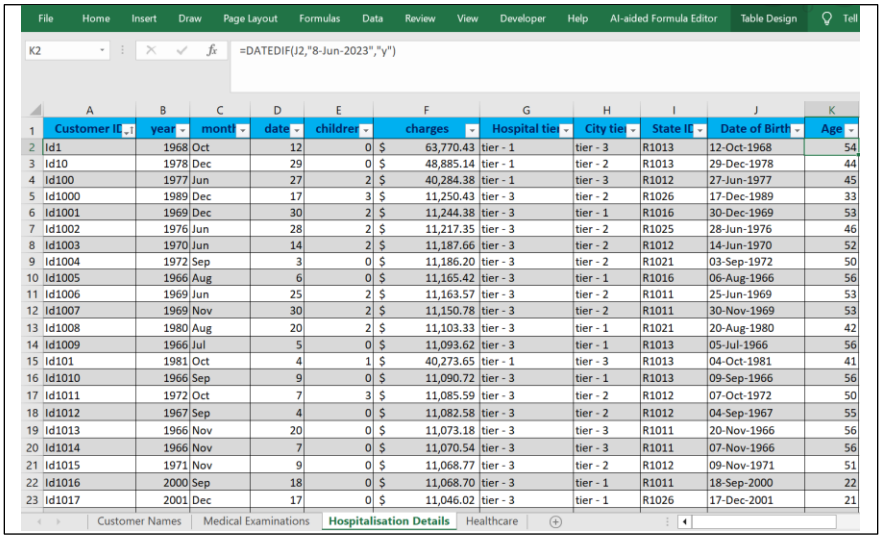
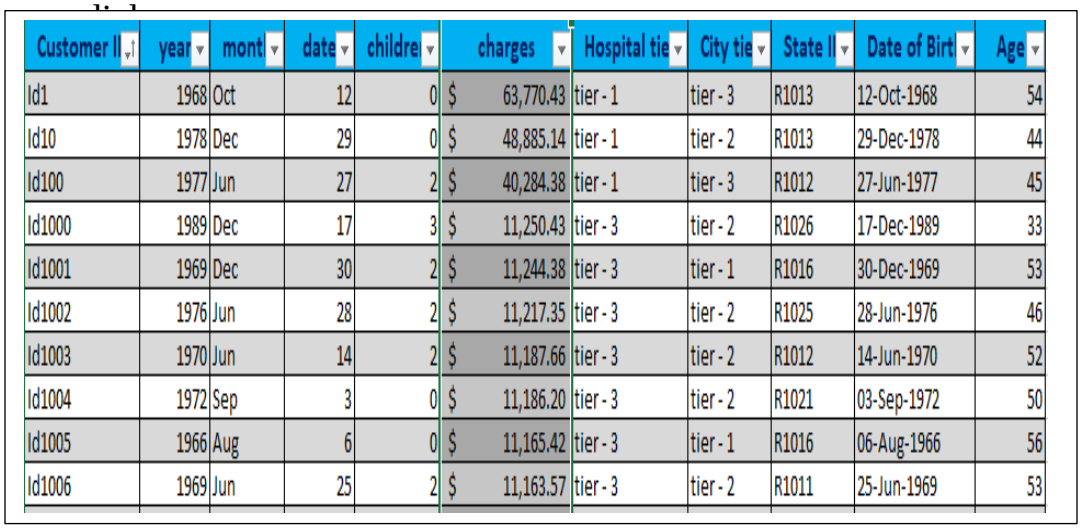
Step	Description
1	**5.1 Split Names** ❖ Split the 'names' column in "Customer Names" table into 'Title', 'First Name', and 'Last Name'. ❖ *Split 'Names' Column: Select the 'Names' column. Go to Data > Text to Columns. Select Delimited > Next. Choose Comma and other as delimiters > Finish.*
2	**5.2 Convert NumberOfMajorSurgeries** ❖ Convert "NumberOfMajorSurgeries" column in "Medical Examinations" table to numerical data. ❖ Use Find and Replace (CTRL+H) for ' No major surgery' to replace with '0'. ❖ Change data format to Number

CustomerId	name	Title	First Name	Last Name
Id1	Hawks, Ms. Kelly	Ms	Kelly	Hawks
Id2	Lehner, Mr. Matthew D	Mr	Matthew D	Lehner
Id3	Lu, Mr. Phil	Mr	Phil	Lu
Id4	Osborne, Ms. Kelsey	Ms	Kelsey	Osborne
Id5	Kadala, Ms. Kristyn	Ms	Kristyn	Kadala
Id6	Baker, Mr. Russell B.	Mr	Russell B	Baker
Id7	Macpherson, Mr. Scott	Mr	Scott	Macpherson
Id8	Hallman, Mr. Stephen	Mr	Stephen	Hallman
Id9	Moran, Mr. Patrick R.	Mr	Patrick R	Moran
Id10	Benner, Ms. Brooke N.	Ms	Brooke N	Benner



Step	Description
3	5.3 Check Inconsistencies ❖ Check for inconsistencies in 'Heart Issues' and 'smoker' columns and propose corrective actions if necessary. 
4	5.4 Categorize Weight Status ❖ Create new columns: “Weight Status” based on BMI values. *=IF(B2>=30,"Obesity”, IF(AND(B2<=29.9, B2>=25),"Overweight”, IF(AND(B2<=24.9,B2>=18.5),"Healthy Weight”, “Underweight")))* 

Step	Description
5	## 5.5 Categorize Diabetes Status Create new columns: “Diabetes Status” based on HbA1C values. *=IF(C2<5.7,"Normal",IF(AND(C2>=5.7,C2<=6.4),"Prediabetes","Diabetes"))* 
6	## 5.6 Merge Date Columns Merge and format date columns in “Hospitalization Details” table. *=TEXT(CONCATENATE(D2,C2,B2),"DD-MMM-YYYY")* 

Step	Description
7	## 5.7 Calculate Age • Calculate the 'Age' using the 'Date of Birth' column and today's date (8th June 2023). *=DATEDIF(J2,"8-Jun-2023","y")* 
8	## 5.8 Format Charges • Format the 'charges' column as currency ($) using the Format Cells 

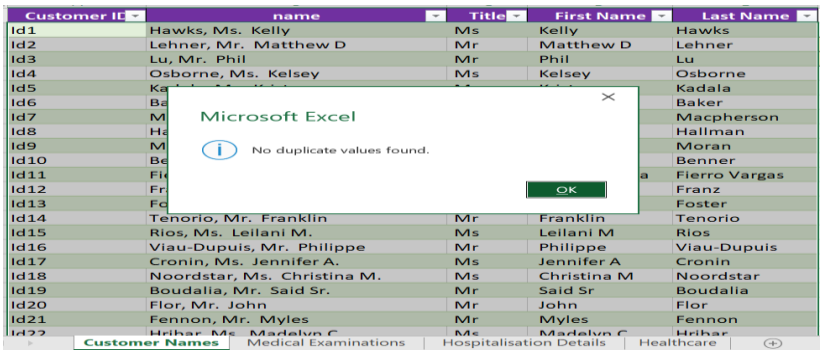
Customer ID	year	month	date	children	charges	Hospital tier	City tier	State ID	Date of Birth	Age
id1	1968	Oct	12	0	$ 63,770.43	tier - 1	tier - 3	R1013	12-Oct-1968	54
id10	1978	Dec	29	0	$ 48,885.14	tier - 1	tier - 2	R1013	29-Dec-1978	44
id100	1977	Jun	27	2	$ 40,284.38	tier - 1	tier - 3	R1012	27-Jun-1977	45
id1000	1989	Dec	17	3	$ 11,250.43	tier - 3	tier - 2	R1026	17-Dec-1989	33
id1001	1969	Dec	30	2	$ 11,244.38	tier - 3	tier - 1	R1016	30-Dec-1969	53
id1002	1976	Jun	28	2	$ 11,217.35	tier - 3	tier - 2	R1025	28-Jun-1976	46
id1003	1970	Jun	14	2	$ 11,187.66	tier - 3	tier - 2	R1012	14-Jun-1970	52
id1004	1972	Sep	3	0	$ 11,186.20	tier - 3	tier - 2	R1021	03-Sep-1972	50
id1005	1966	Aug	6	0	$ 11,165.42	tier - 3	tier - 1	R1016	06-Aug-1966	56
id1006	1969	Jun	25	2	$ 11,163.57	tier - 3	tier - 2	R1011	25-Jun-1969	53

6. Data Exploration

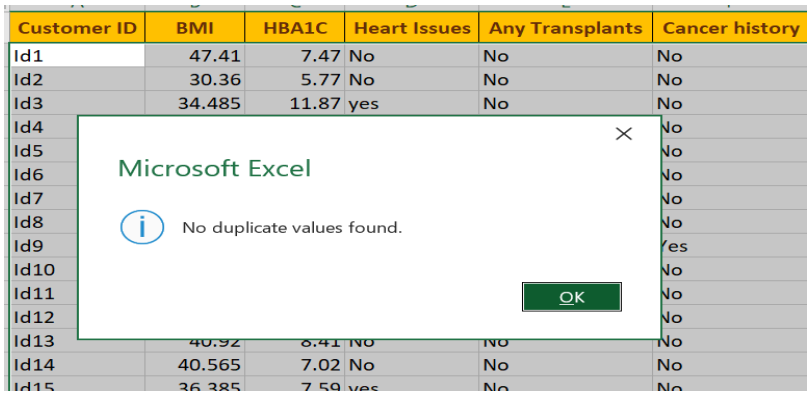
6.1 Customer Names Table

6.1.1 Find duplicate Customer ID

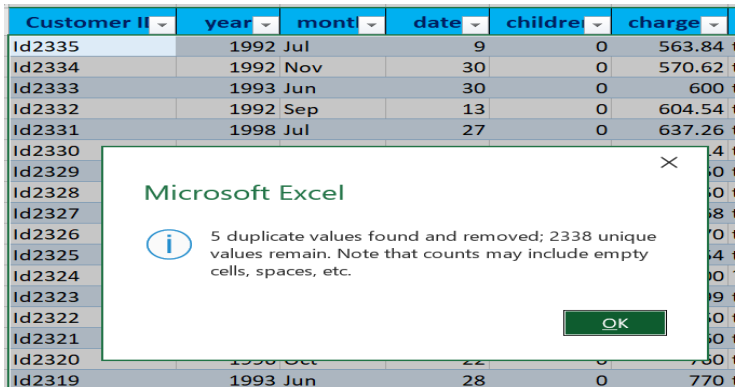
Are there any duplicate Customer IDs in the dataset?



Customer Names Table : Duplicate customer ID = 0



Medical Examinations Table: Duplicate customer ID = 0

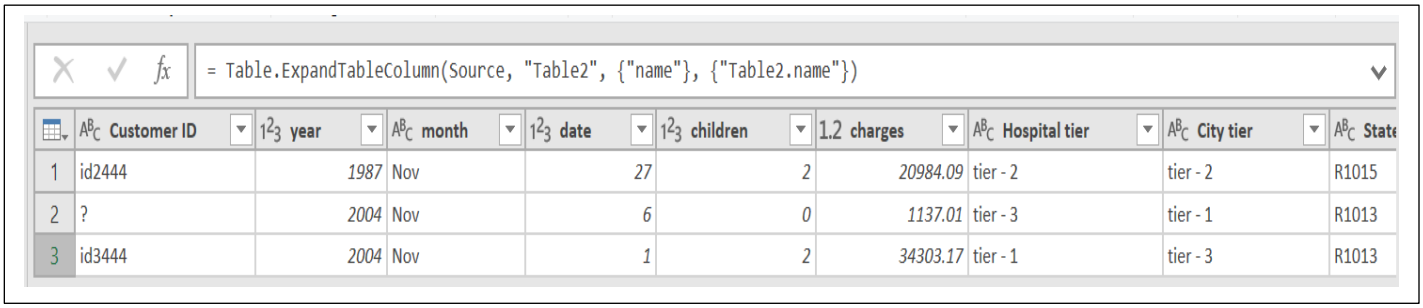


Hospitalisation Details Table: Duplicate customer ID = 5

6.1.2 Find Total No of Customer

How many customers are included in the dataset?

- ❖ Loaded both (Customer Names Table and Hospitalisation Details Table) datasets into Excel using Power Query.
- ❖ Merged the datasets based on a customer ID using Left Anti Join.
- ❖ Identified rows unique to the second dataset (extra rows).
- ❖ Number of extra rows found in the second dataset.



= Table.ExpandTableColumn(Source, "Table2", {"name"}, {"Table2.name"})

	Customer ID	year	month	date	children	charges	Hospital tier	City tier	State
1	id2444	1987	Nov	27	2	20984.09	tier - 2	tier - 2	R1015
2	?	2004	Nov	6	0	1137.01	tier - 3	tier - 1	R1013
3	id3444	2004	Nov	1	2	34303.17	tier - 1	tier - 3	R1013

- ❖ Remove Extra Rows from the data sets because customer name not found.
- ❖ Now Counted total customers = 2335
- ❖ *=COUNTA ($A$2: $A$2336)*

6.2 Medical Examination Table

Question	Methodology	*Formula*	Output
6.2.1 How many customers have a history of cancer?	Filter the Cancer history column for Yes.	*=COUNTIF(F2:F2336,"Yes")*	391
6.2.2 Identify the customer(s) with the highest BMI.	Sort the BMI column from highest to lowest.	*=MAX($B$2:$B$2336)*	55.05
6.2.3 How many customers have Diabetes?	Filter the HBA1C column for values >= 6.5.	*=COUNTIF(J2:J2336,"Diabetes")*	797
6.2.4 How many obese customers have heart issues?	Filter BMI >= 30, then filter Heart Issues for Yes.	*=COUNTIFS(D2:D2336,"Yes",I2:I2336,"Obesity")*	490
6.2.5 Total number of major surgeries performed on customers?	Ensure NumberOfMajorSurgeries is numeric, then sum values.	*=SUM(G2:G2336)*	1579
6.2.6 Percentage of customers who have undergone any transplants?	Filter Any Transplants for Yes, calculate percentage.	*=COUNTIF(E2:E2336,"Yes")/COUNTA(E2:E2336)*100*	6.167
6.2.7 Average HBA1C value of customers who are smokers?	Filter Smoker for Yes, calculate average of HBA1C.	*=AVERAGEIF(H2:H2336,"yes",C2:C2336)*	6.61834
6.2.8 How many customers with heart issues have done transplant?	Filter Heart Issues for Yes, then Any Transplants for Yes.	*=COUNTIFS(D2:D2336,"yes",E2:E2336,"yes")*	19
6.2.9 Average BMI of customers who have done more than 2 major surgeries?	Filter NumberOfMajorSurgeries for > 2, calculate average BMI.	*=AVERAGEIF(G2:G2336,">2",B2:B2336)*	32.97614

6.3 Hospitalization Details Table

6.3.1 Summary Statistics

❖ Calculate all the Summary statistics for the 'charges' column.

Statistic	Description	*Excel Formula*	Summary statistics ($)
Mean (Average)	The arithmetic average of charges.	*=AVERAGE(f2:f2344)*	13559.0679
Median	The middle value of charges.	*=MEDIAN(f2: f2344)*	9634.54
Mode	The most frequently occurring charge amount.	*=MODE.SNGL(f2:f2344)*	650
Standard Deviation	The dispersion or spread of charges.	*=STDEV.S(f2:f2344)*	11922.6584
Variance	The variability of charges (square of Std Dev).	*=VAR.S(f2:f2344)*	142149784
Minimum	The smallest charge amount.	*=MIN(f2:f2344)*	563.84
Maximum	The largest charge amount.	*=MAX(f2:f2344)*	63770.43
Range	The difference between the maximum and minimum charges.	*=MAX(f2:f2344) - MIN(f2:f2344)*	63206.59
First Quartile (Q1)	The value below which 25% of charges fall.	*=QUARTILE.INC(f2:f2344, 1)*	5084.01
Third Quartile (Q3)	The value below which 75% of charges fall.	*=QUARTILE.INC(f2:f2344, 3)*	17029.675
Interquartile Range	The difference between Q3 and Q1 for charges.	*=QUARTILE.INC(f2:f2344, 3) - QUARTILE.INC(f2:f2344, 1)*	11945.665
Count	The number of charge values.	*=COUNT(f2:f2344)*	2343

6.3.2 Median Age and The Most Common Age

Which is the median age and the most common age in the dataset?

- Identified median Age = *39*
- *=MEDIAN (K2: K2344)*
- most common ages *= 18*
- *=MODE.SNGL(K1:K2344)*

6.3.3 Average Hospitalization Charges

Find the average hospitalization charges for customers who are more than 50 years old.

- Calculated average charges for older customers *= $ 17,856.79*
- *=AVERAGEIF(K2:K2344,">50",F2:F2344)*

6.3.4 Total charges across Hospital tiers

- Compare the total charges across different hospital tiers.
- Total charges across tiers = *$ 31,768,896.02*

Hospital tier	Sum of charges
tier - 1	$ 9,310,917.49
tier - 2	$ 15,899,488.89
tier - 3	$ 6,558,489.64
Grand Total	**$ 31,768,896.02**

6.3.5 Highest Average Hospitalization Charges

Which city tier has the highest average hospitalization charges?

- Identified city tier with highest charges = *tier – 3 = $ 14082.56*

Row Labels	Average of charges
tier - 3	$ 14,082.56
tier - 2	$ 13,454.84
tier - 1	$ 13,051.35
Grand Total	**$ 13,542.00**

6.3.6 Average Charges for People

Calculate the average charges for people who have more than 2 children.

Calculated average charges for customers with multiple children *=$ 14217.52*

Row Labels	Average of charges
3	$ 14,500.43
4	$ 13,850.66
5	$ 8,786.04
Grand Total	**$ 14,217.52**

6.3.7 Average Number of Children

- Find the integer average number of children of customers who are less than 40 years old.
- Calculated average number of children for younger customers = *1.1*
- *=AVERAGEIF(K2:K2338,"<40",E2:E2338)*

7. Data Analysis

7.1 Combine Tables

- Combined all three tables using Customer ID as common column.
- Combine the tables using VLOOKUP in a new sheet “Healthcare”.

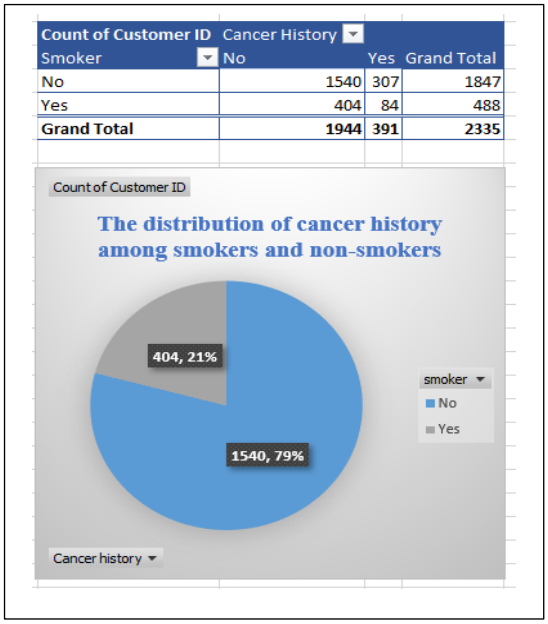
- =VLOOKUP($A2,Table5[#All],4,FALSE)
- =VLOOKUP($A2,Table2[#All],2,FALSE)
- =VLOOKUP($A2,Table2[#All],3,FALSE)
- =VLOOKUP($A2,Table2[#All],4,FALSE)
- =VLOOKUP($A2,Table2[#All],5,FALSE)
- =VLOOKUP($A2,Table2[#All],6,FALSE)
- =VLOOKUP($A2,Table2[#All],7,FALSE)
- =VLOOKUP($A2,Table2[#All],8,FALSE)
- =VLOOKUP($A2,Table2[#All],9,FALSE)
- =VLOOKUP($A2,Table2[#All],10,FALSE)
- =VLOOKUP($A2,Table1[#All],10,FALSE)
- =VLOOKUP($A2,Table1[#All],6,FALSE)
- =VLOOKUP($A2,Table1[#All],7,FALSE)
- =VLOOKUP($A2,Table1[#All],8,FALSE)
- =VLOOKUP($A2,Table1[#All],9,FALSE)
- =VLOOKUP($A2,Table1[#All],11,FALSE)

Customer ID	First Name	BMI	HBA1C	Heart Issues	Any Transplants	Cancer history	NumberOfMajorSurgeries	smoker	Weight Status	Diabetes Status	Date of Birth	charges	Hospital tier	City tier	State ID	Age
Id1	Kelly	47.41	7.47	No	No	No	0	Yes	Obesity	Diabetes	12-Oct-1968	$ 63,770.43	tier - 1	tier - 3	R1013	54
Id2	Matthew D	30.36	5.77	No	No	No	0	Yes	Obesity	Prediabetes	08-Jun-1977	$ 62,592.87	tier - 2	tier - 3	R1013	46
Id3	Phil	34.485	11.87	Yes	No	No	2	Yes	Obesity	Diabetes	11-Sep-1970	$ 60,021.40	tier - 1	tier - 1	R1012	52
Id4	Kelsey	38.095	6.05	No	No	No	0	Yes	Obesity	Prediabetes	06-Jun-1991	$ 58,571.07	tier - 1	tier - 3	R1024	32
Id5	Kristyn	35.53	5.45	No	No	No	0	Yes	Obesity	Normal	19-Jun-1989	$ 55,135.40	tier - 1	tier - 2	R1012	33
Id6	Russell B	32.8	6.59	No	No	No	0	Yes	Obesity	Diabetes	04-Aug-1962	$ 52,590.83	tier - 1	tier - 3	R1011	60
Id7	Scott	36.4	6.07	No	No	No	0	Yes	Obesity	Prediabetes	27-Oct-1994	$ 51,194.56	tier - 1	tier - 3	R1011	28
Id8	Stephen	36.96	7.93	No	No	No	3	Yes	Obesity	Diabetes	27-Jun-1958	$ 49,577.66	tier - 2	tier - 2	R1013	64
Id9	Patrick R	41.14	9.58	Yes	No	Yes	1	Yes	Obesity	Diabetes	04-Sep-1963	$ 48,970.25	tier - 1	tier - 2	R1013	59
Id10	Brooke N	38.06	10.79	No	No	No	0	Yes	Obesity	Diabetes	29-Dec-1978	$ 48,885.14	tier - 1	tier - 2	R1013	44
Id11	Paola Andrea	37.7	5.96	Yes	No	No	2	Yes	Obesity	Prediabetes	22-Jul-1959	$ 48,824.45	tier - 2	tier - 1	R1011	63

7.2 Pivot Tables and Visualizations

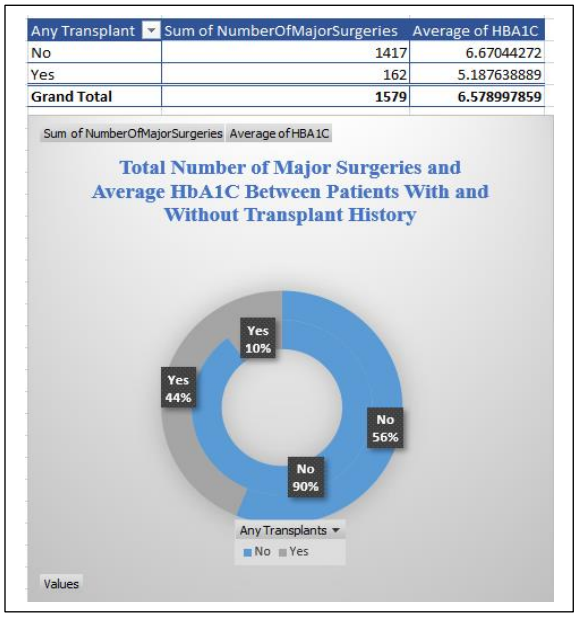
7.2.1 Distribution of Cancer History Among Smokers and Non-Smokers (Pie Chart)

1	Create a Pivot Table
	a) Select your dataset.
	b) Go to the Insert tab and click on PivotTable.
	c) Place the PivotTable in a new worksheet.
	d) Drag Smoker to the Rows area.
	e) Drag Cancer History to the Columns area.
	f) Drag Customer ID to the Values area and set it to Count.
2	Create a Pie/Donut Chart
	a) Select the PivotTable.
	b) Go to the Insert tab.
	c) Click on Pie Chart and select either a Pie or Donut chart.
	d) Format the chart as needed.



7.2.2 Difference in Major Surgeries and Average HbA1C Based on Transplants (Donut Chart)

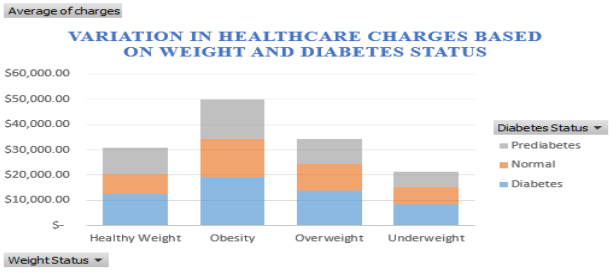
1	Create a Pivot Table
	a) Select your dataset.
	b) Go to the Insert tab and click on PivotTable.
	c) Place the PivotTable in a new worksheet.
	d) Drag Any Transplants to the Rows area.
	e) Drag NumberOfMajorSurgeries to the Values area and set it to Sum.
	f) Drag HBA1C to the Values area again and set it to Average.
2	Create a Pie/Donut Chart
	a) Select the PivotTable.
	b) Go to the Insert tab.
	c) Click on Pie Chart and select either a Pie or Donut chart.
	d) Format the chart as needed.



7.2.3 Variation in Healthcare Charges by Weight and Diabetes Statuses (Column/Bar Chart)

1	Create a Pivot Table
	a) Select your dataset.
	b) Go to the Insert tab and click on PivotTable.
	c) Place the PivotTable in a new worksheet.
	d) Drag Weight Status to the Rows area.
	e) Drag Diabetes Status to the Columns area.
	f) Drag Charges to the Values area and set it to Sum.
2	Create a Column/Bar Chart
	a) Select the PivotTable.
	b) Go to the Insert tab.
	c) Click on Column Chart or Bar Chart.
	d) Format the chart as needed.

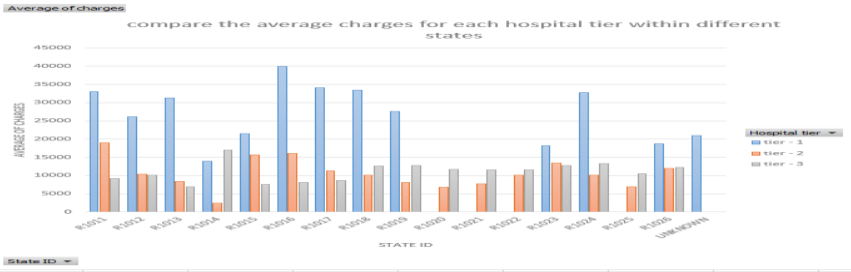
Average of charges	Diabetes Statu			
Weight Status	Diabetes	Normal	Prediabetes	Grand Total
Healthy Weight	$ 12,509.82	$ 7,897.08	$ 10,389.96	$ 9,842.51
Obesity	$ 19,162.19	$ 15,188.44	$ 15,579.10	$ 16,661.46
Overweight	$ 13,920.07	$ 10,380.81	$ 10,008.42	$ 11,420.63
Underweight	$ 8,281.67	$ 7,030.57	$ 6,015.34	$ 7,350.37
Grand Total	**$ 15,973.54**	**$ 12,090.72**	**$ 12,690.99**	**$ 13,529.92**



7.2.4 Average Charges for Each Hospital Tier Within Different States (Column/Bar Chart)

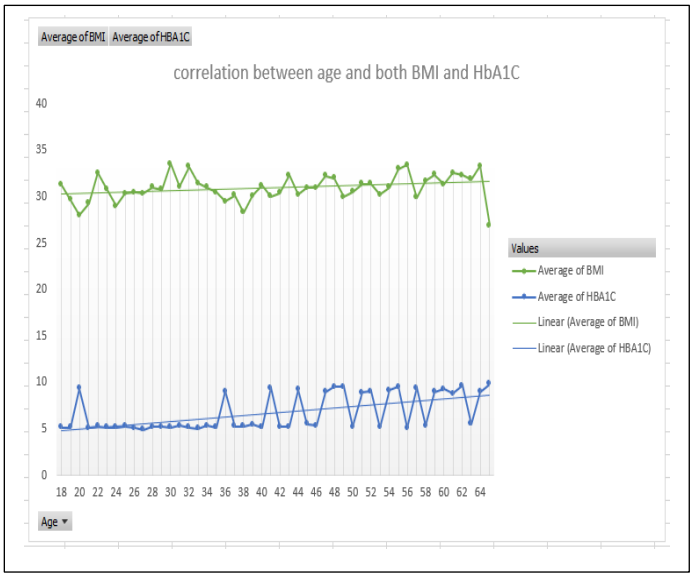
1	Create a Pivot Table
	a) Select your dataset.
	b) Go to the Insert tab and click on PivotTable.
	c) Place the PivotTable in a new worksheet.
	d) Drag State ID to the Rows area.
	e) Drag Hospital Tier to the Columns area.
	f) Drag Charges to the Values area and set it to Average.
2	Create a Column/Bar Chart
	a) Select the PivotTable.
	b) Go to the Insert tab.
	c) Click on Column Chart or Bar Chart.
	d) Format the chart as needed.

Average of charges	Column Labels			
Row Labels	tier - 1	tier - 2	tier - 3	Grand Total
R1011	33081.37379	18997.28069	9250.361786	19466.24611
R1012	26111.45794	10334.27932	10179.94197	12016.47391
R1013	31328.10224	8439.2844	6890.756824	10514.9398
R1014	13891.86	2395.17	16950.485	13478.05692
R1015	21523.52	15675.54333	7570.755	13948.61
R1016	39868.61625	16075.96333	8133.008409	13589.26344
R1017	34070.59571	11365.62188	8667.366154	14806.10778
R1018	33475.82	10139.55333	12570.97	13272.78667
R1019	27621.61	8116.72	12706.976	13633.14346
R1020		6760.543333	11716.45333	9238.498333
R1021		7668.178723	11518.78565	8933.378143
R1022		10102.02667	11637.812	10650.52143
R1023	18261.7575	13457.09458	12785.637	13786.14921
R1024	32732.18714	10174.346	13360.97889	12865.1495
R1025		6886.036842	10510.31381	8788.78225
R1026	18709.798	11919.60786	12296.77973	12489.92107
UNKNOWN	20996.53			20996.53
Grand Total	**30129.19859**	**11865.26877**	**9462.269307**	**13529.91803**



7.2.5 Correlation Between Age and Both BMI and HbA1C (Line/Scatter Plot)

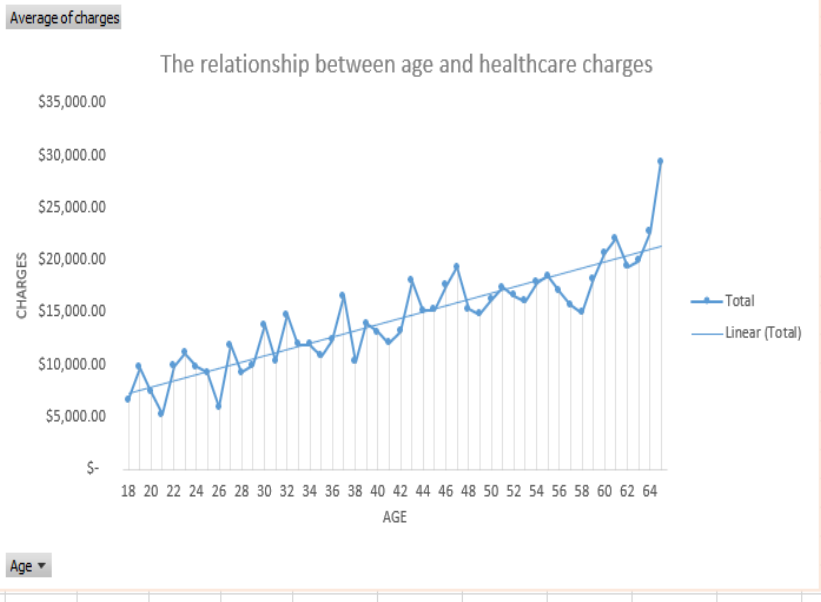
1	Create a Scatter Plot for Age vs. BMI
	a) Select your dataset.
	b) Go to the Insert tab.
	c) Click on Scatter Chart and select Scatter with only Markers.
	d) Select the data range for Age as the X-axis and BMI as the Y-axis.
	e) Format the chart as needed.
	f) Create a Scatter Plot for Age vs. HbA1C
2	Select your dataset.
	a) Go to the Insert tab.
	b) Click on Scatter Chart and select Scatter with only Markers.
	c) Select the data range for Age as the X-axis and HBA1C as the Y-axis.
	d) Format the chart as needed.



7.2.6 Relationship Between Age and Healthcare Charges (Line/Scatter Plot)

1	Create a Scatter Plot
	a) Select your dataset.
	b) Go to the Insert tab.
	c) Click on Scatter Chart and select Scatter with only Markers.
	d) Select the data range for Age as the X-axis and Charges as the Y-axis.
	e) Format the chart as needed.

Age	Average of charges
18	$ 6,611.31
19	$ 9,677.19
20	$ 7,411.45
21	$ 5,214.53
22	$ 9,806.23
23	$ 11,092.93
24	$ 9,765.57
25	$ 9,184.89
26	$ 5,920.59
27	$ 11,846.24
28	$ 9,219.82
29	$ 9,914.80
30	$ 13,787.85
31	$ 10,346.34
32	$ 14,698.09
33	$ 11,891.08
34	$ 11,886.59
35	$ 10,817.26
36	$ 12,290.95
37	$ 16,503.81



8. Conclusion

By following these detailed steps in Excel, you can effectively clean, transform, explore, and analyze healthcare data, leading to valuable insights and informed decision-making. Each step ensures the data is accurate, relevant, and actionable, allowing healthcare stakeholders to optimize patient care, resource allocation, and cost management strategies.