Capstone Project in Microsoft Excel

Healthcare Data Analysis and Insights 2024

by

Keerthana.N

Table of contents

Project Title: Healthcare Data Analysis and Insights

Problem Statement

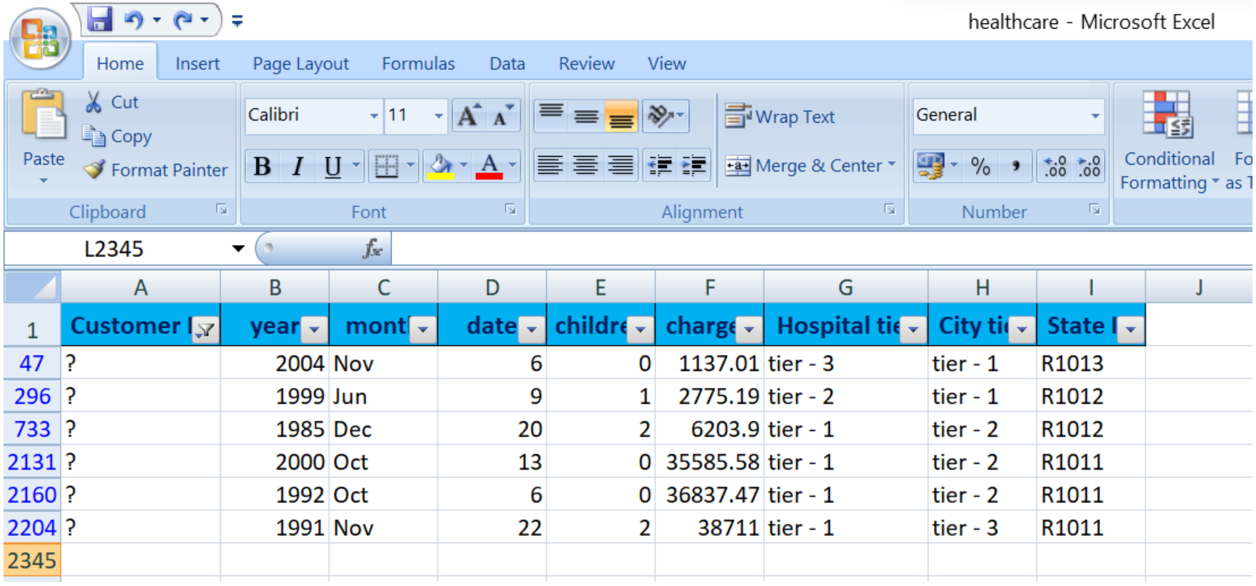
The healthcare industry generates vast amounts of data daily, providing valuable insights for healthcare providers and policymakers to improve patient care, allocate resources effectively, and manage healthcare costs. This project aims to analyze a comprehensive healthcare dataset comprising medical examinations, hospitalization details, and customer profiles to extract insights into patient health profiles, medical histories, and healthcare costs. By exploring relationships between various health metrics, identifying trends, and visualizing key patterns, we aim to deliver actionable insights to healthcare stakeholders for informed decision-making through rigorous data cleaning, transformation, exploration, and analysis.

Problem solving

Data cleaning
Data transformation
Data exploration
Data analysis

Data Cleaning:

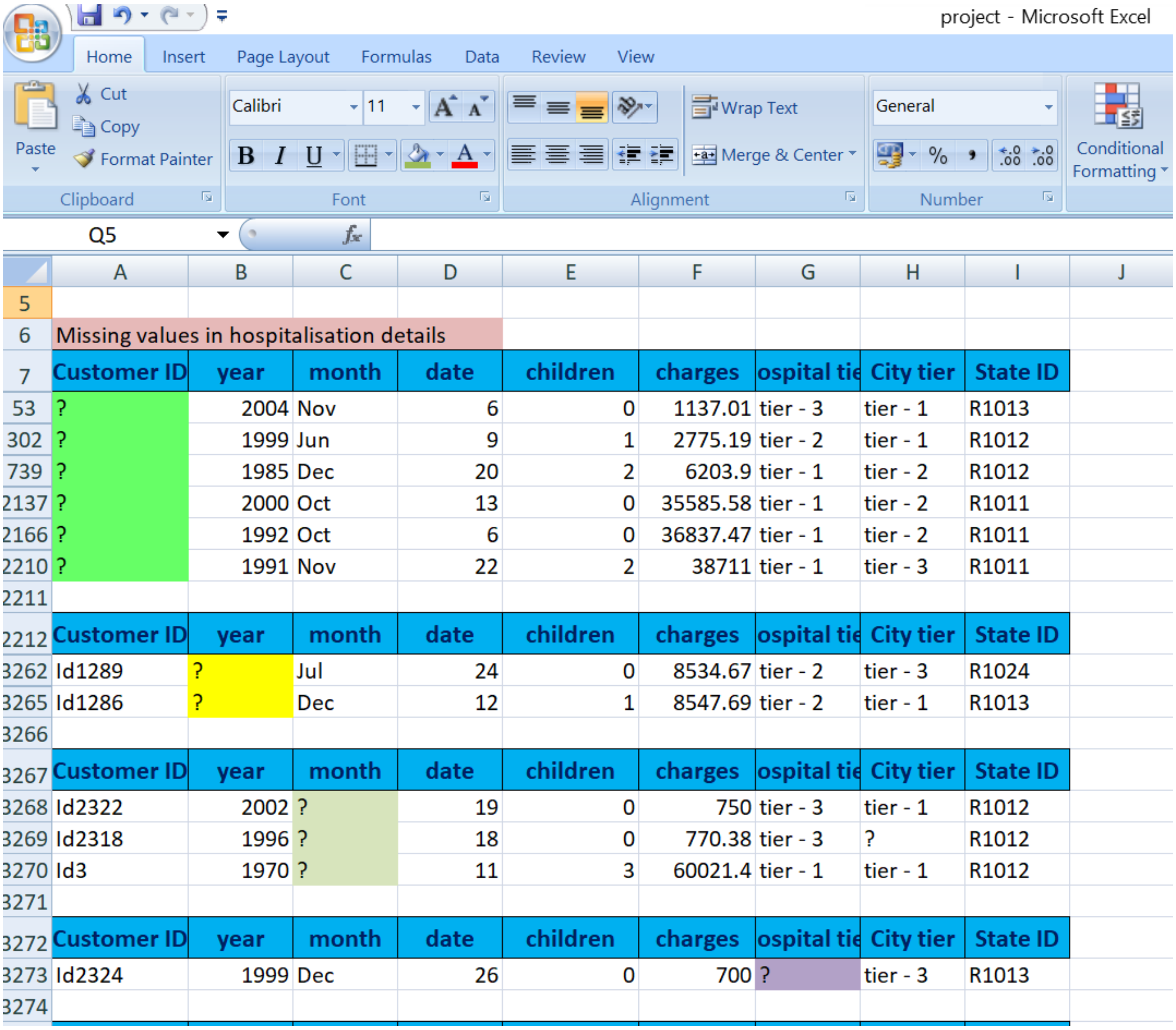
The number of missing values marked with '?' in each column of the “Medical Examinations” Table and "Hospitalization Details" Table.



	A	B	C	D	E	F	G	H	I	J
1	Customer I	year	mont	date	childre	charge	Hospital tie	City ti	State I	
47	?	2004	Nov	6	0	1137.01	tier - 3	tier - 1	R1013	
296	?	1999	Jun	9	1	2775.19	tier - 2	tier - 1	R1012	
733	?	1985	Dec	20	2	6203.9	tier - 1	tier - 2	R1012	
2131	?	2000	Oct	13	0	35585.58	tier - 1	tier - 2	R1011	
2160	?	1992	Oct	6	0	36837.47	tier - 1	tier - 2	R1011	
2204	?	1991	Nov	22	2	38711	tier - 1	tier - 3	R1011	
2345										

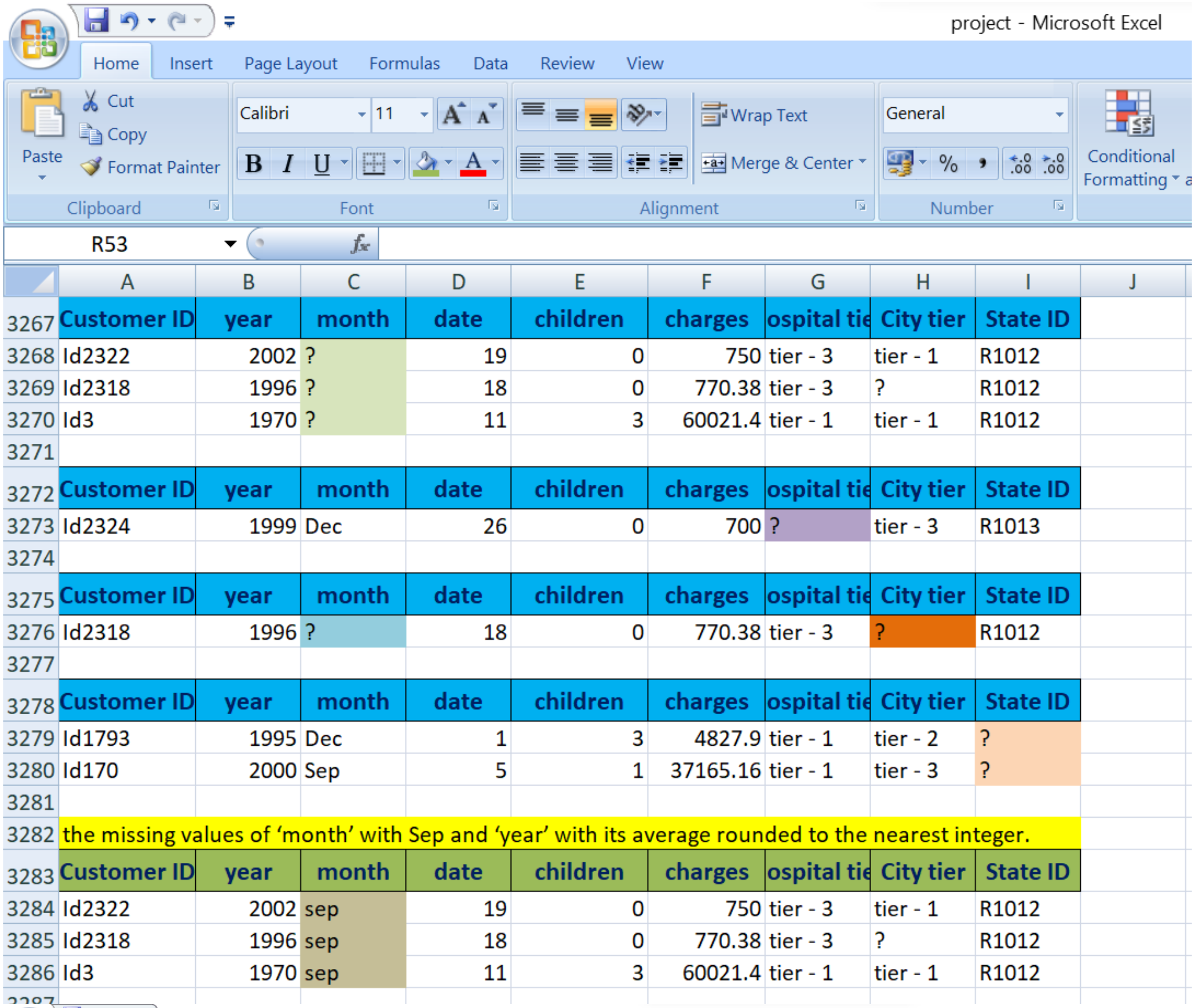
The number of missing values are six.

Fill in the missing values of ‘month’ with Sep and ‘year’ with its average rounded to the nearest integer.



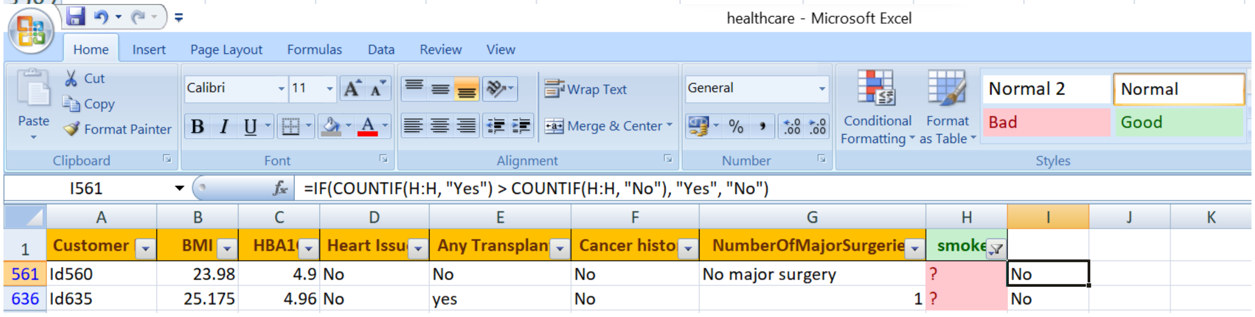
	A	B	C	D	E	F	G	H	I	J
5										
6	Missing values in hospitalisation details									
7	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
53	?	2004	Nov	6	0	1137.01	tier - 3	tier - 1	R1013	
302	?	1999	Jun	9	1	2775.19	tier - 2	tier - 1	R1012	
739	?	1985	Dec	20	2	6203.9	tier - 1	tier - 2	R1012	
2137	?	2000	Oct	13	0	35585.58	tier - 1	tier - 2	R1011	
2166	?	1992	Oct	6	0	36837.47	tier - 1	tier - 2	R1011	
2210	?	1991	Nov	22	2	38711	tier - 1	tier - 3	R1011	
2211										
2212	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
3262	Id1289	?	Jul	24	0	8534.67	tier - 2	tier - 3	R1024	
3265	Id1286	?	Dec	12	1	8547.69	tier - 2	tier - 1	R1013	
3266										
3267	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
3268	Id2322	2002	?	19	0	750	tier - 3	tier - 1	R1012	
3269	Id2318	1996	?	18	0	770.38	tier - 3	?	R1012	
3270	Id3	1970	?	11	3	60021.4	tier - 1	tier - 1	R1012	
3271										
3272	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
3273	Id2324	1999	Dec	26	0	700	?	tier - 3	R1013	
3274										

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



R53

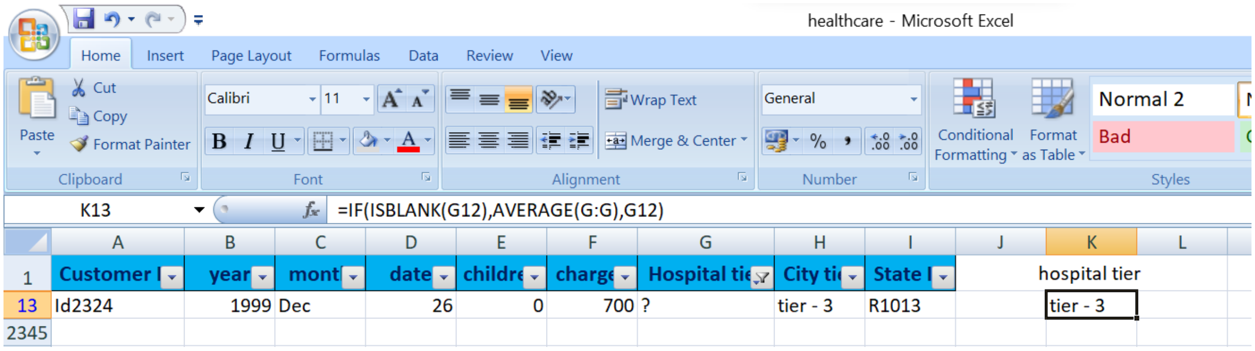
	A	B	C	D	E	F	G	H	I	J
3267	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
3268	Id2322	2002	?	19	0	750	tier - 3	tier - 1	R1012	
3269	Id2318	1996	?	18	0	770.38	tier - 3	?	R1012	
3270	Id3	1970	?	11	3	60021.4	tier - 1	tier - 1	R1012	
3271										
3272	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
3273	Id2324	1999	Dec	26	0	700	?	tier - 3	R1013	
3274										
3275	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
3276	Id2318	1996	?	18	0	770.38	tier - 3	?	R1012	
3277										
3278	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
3279	Id1793	1995	Dec	1	3	4827.9	tier - 1	tier - 2	?	
3280	Id170	2000	Sep	5	1	37165.16	tier - 1	tier - 3	?	
3281										
3282	the missing values of ‘month’ with Sep and ‘year’ with its average rounded to the nearest integer.									
3283	Customer ID	year	month	date	children	charges	ospital tie	City tier	State ID	
3284	Id2322	2002	sep	19	0	750	tier - 3	tier - 1	R1012	
3285	Id2318	1996	sep	18	0	770.38	tier - 3	?	R1012	
3286	Id3	1970	sep	11	3	60021.4	tier - 1	tier - 1	R1012	



I561 =IF(COUNTIF(H:H, "Yes") > COUNTIF(H:H, "No"), "Yes", "No")

	A	B	C	D	E	F	G	H	I	J	K
1	Customer	BMI	HBA1(	Heart Issu	Any Transplan	Cancer histo	NumberOfMajorSurgerie	smoke			
561	Id560	23.98	4.9	No	No	No	No major surgery	?	No		
636	Id635	25.175	4.96	No	yes	No	1	?	No		

If any 'State ID' values are missing, consider filling them with 'Unknown' or using another appropriate strategy.

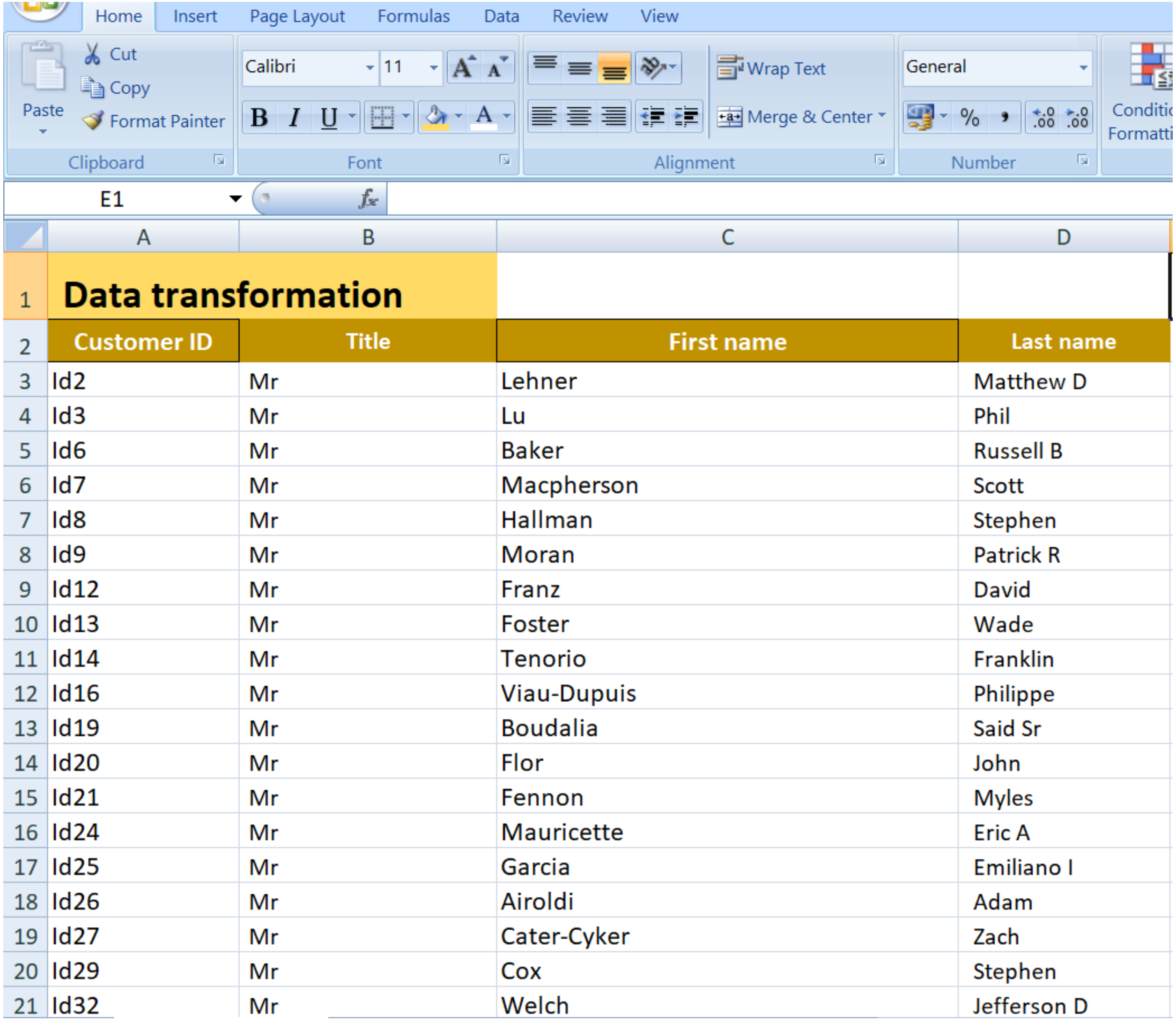


K13 =IF(ISBLANK(G12),AVERAGE(G:G),G12)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Customer I	year	mont	date	childre	charge	Hospital tie	City ti	State I		hospital tier	
13	Id2324	1999	Dec	26	0	700	?	tier - 3	R1013		tier - 3	
2345												

Data Transformation

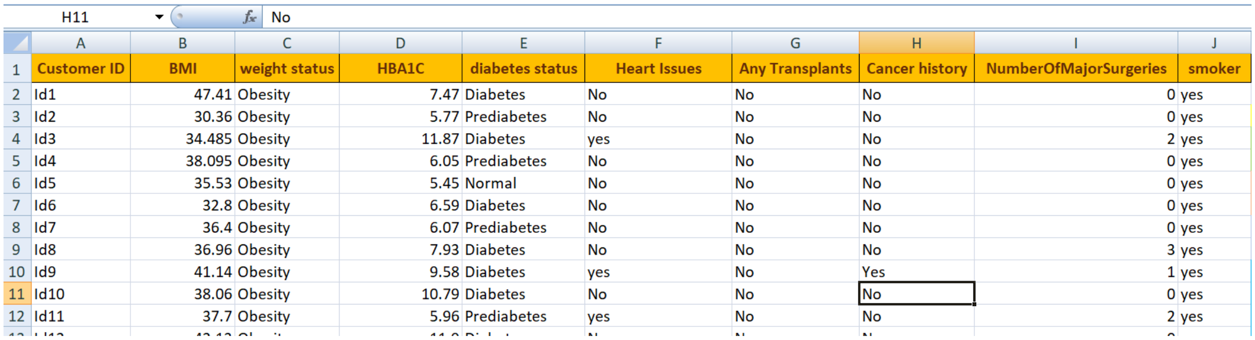
Split the 'names' column in the "Customer Names" Table into 3 meaningful columns: 'Title', 'First Name', and 'Last Name'



E1

	A	B	C	D
1	Data transformation			
2	Customer ID	Title	First name	Last name
3	Id2	Mr	Lehner	Matthew D
4	Id3	Mr	Lu	Phil
5	Id6	Mr	Baker	Russell B
6	Id7	Mr	Macpherson	Scott
7	Id8	Mr	Hallman	Stephen
8	Id9	Mr	Moran	Patrick R
9	Id12	Mr	Franz	David
10	Id13	Mr	Foster	Wade
11	Id14	Mr	Tenorio	Franklin
12	Id16	Mr	Viau-Dupuis	Philippe
13	Id19	Mr	Boudalia	Said Sr
14	Id20	Mr	Flor	John
15	Id21	Mr	Fennon	Myles
16	Id24	Mr	Mauricette	Eric A
17	Id25	Mr	Garcia	Emiliano I
18	Id26	Mr	Airoldi	Adam
19	Id27	Mr	Cater-Cyker	Zach
20	Id29	Mr	Cox	Stephen
21	Id32	Mr	Welch	Jefferson D

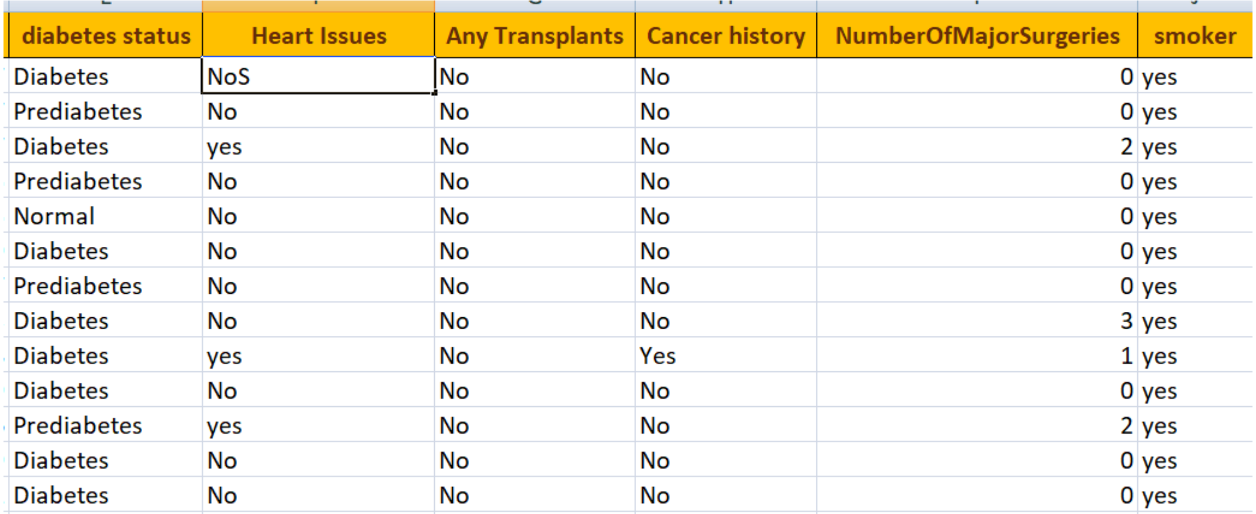
Convert the "NumberOfMajorSurgeries" column in the “Medical Examinations” Table to numerical data by replacing non-numeric characters with meaningful numerical value.



H11 fx No

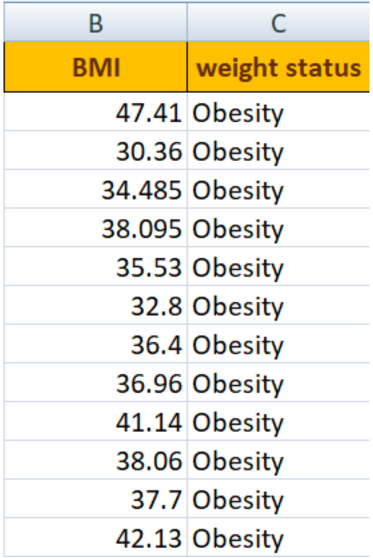
	A	B	C	D	E	F	G	H	I	J
1	Customer ID	BMI	weight status	HBA1C	diabetes status	Heart Issues	Any Transplants	Cancer history	NumberOfMajorSurgeries	smoker
2	Id1	47.41	Obesity	7.47	Diabetes	No	No	No	0	yes
3	Id2	30.36	Obesity	5.77	Prediabetes	No	No	No	0	yes
4	Id3	34.485	Obesity	11.87	Diabetes	yes	No	No	2	yes
5	Id4	38.095	Obesity	6.05	Prediabetes	No	No	No	0	yes
6	Id5	35.53	Obesity	5.45	Normal	No	No	No	0	yes
7	Id6	32.8	Obesity	6.59	Diabetes	No	No	No	0	yes
8	Id7	36.4	Obesity	6.07	Prediabetes	No	No	No	0	yes
9	Id8	36.96	Obesity	7.93	Diabetes	No	No	No	3	yes
10	Id9	41.14	Obesity	9.58	Diabetes	yes	No	Yes	1	yes
11	Id10	38.06	Obesity	10.79	Diabetes	No	No	No	0	yes
12	Id11	37.7	Obesity	5.96	Prediabetes	yes	No	No	2	yes

Check for inconsistencies in the 'Heart Issues' and 'smoker' columns and propose corrective actions if necessary.



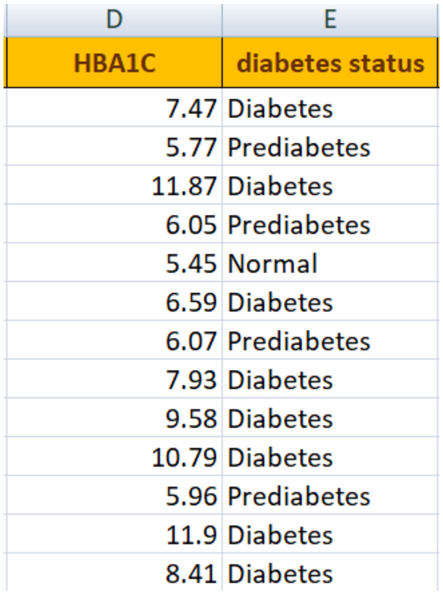
diabetes status	Heart Issues	Any Transplants	Cancer history	NumberOfMajorSurgeries	smoker
Diabetes	NoS	No	No	0	yes
Prediabetes	No	No	No	0	yes
Diabetes	yes	No	No	2	yes
Prediabetes	No	No	No	0	yes
Normal	No	No	No	0	yes
Diabetes	No	No	No	0	yes
Prediabetes	No	No	No	0	yes
Diabetes	No	No	No	3	yes
Diabetes	yes	No	Yes	1	yes
Diabetes	No	No	No	0	yes
Prediabetes	yes	No	No	2	yes
Diabetes	No	No	No	0	yes
Diabetes	No	No	No	0	yes

Create a new column named “Weight Status” that categorizes BMI into different categories



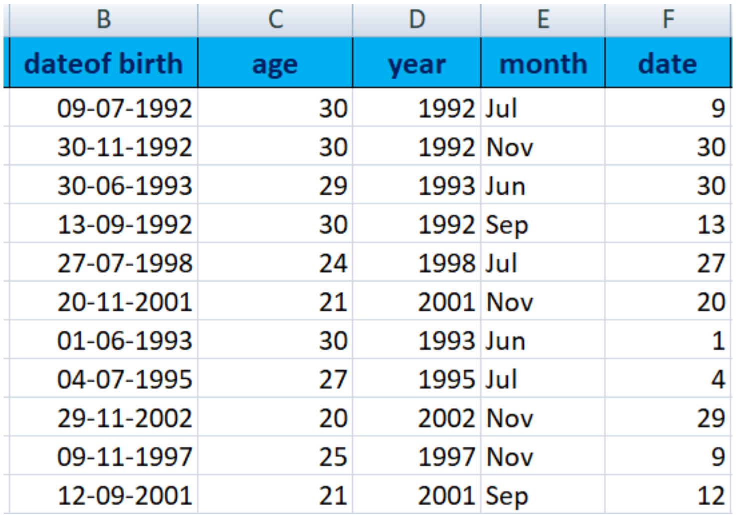
B	C
BMI	weight status
47.41	Obesity
30.36	Obesity
34.485	Obesity
38.095	Obesity
35.53	Obesity
32.8	Obesity
36.4	Obesity
36.96	Obesity
41.14	Obesity
38.06	Obesity
37.7	Obesity
42.13	Obesity

Create a new column named “Diabetes Status” and fill it as per the information given below:



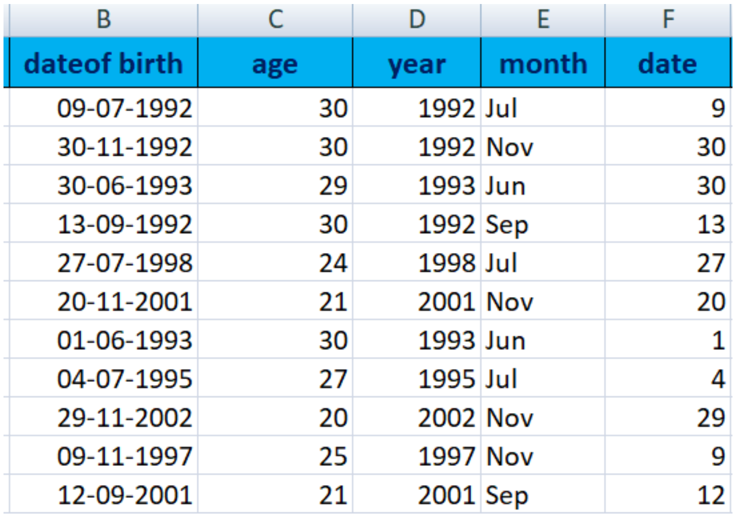
D	E
HBA1C	diabetes status
7.47	Diabetes
5.77	Prediabetes
11.87	Diabetes
6.05	Prediabetes
5.45	Normal
6.59	Diabetes
6.07	Prediabetes
7.93	Diabetes
9.58	Diabetes
10.79	Diabetes
5.96	Prediabetes
11.9	Diabetes
8.41	Diabetes

Merge ‘year’, ‘month’ and ‘date’ columns in the “Hospitalization Details” Table into one column named ‘Date of Birth’ and format it in ‘DD-MMM-YYYY’ custom format.



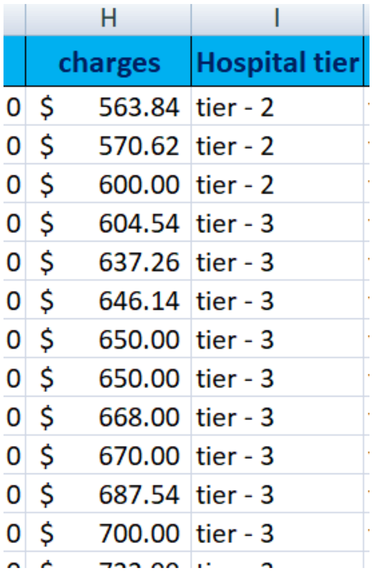
B	C	D	E	F
dateof birth	age	year	month	date
09-07-1992	30	1992	Jul	9
30-11-1992	30	1992	Nov	30
30-06-1993	29	1993	Jun	30
13-09-1992	30	1992	Sep	13
27-07-1998	24	1998	Jul	27
20-11-2001	21	2001	Nov	20
01-06-1993	30	1993	Jun	1
04-07-1995	27	1995	Jul	4
29-11-2002	20	2002	Nov	29
09-11-1997	25	1997	Nov	9
12-09-2001	21	2001	Sep	12

The ‘Age’ of each customer based on their ‘Date of Birth’ and the date of collection of the dataset, which is 8 th June 2023.



B	C	D	E	F
dateof birth	age	year	month	date
09-07-1992	30	1992	Jul	9
30-11-1992	30	1992	Nov	30
30-06-1993	29	1993	Jun	30
13-09-1992	30	1992	Sep	13
27-07-1998	24	1998	Jul	27
20-11-2001	21	2001	Nov	20
01-06-1993	30	1993	Jun	1
04-07-1995	27	1995	Jul	4
29-11-2002	20	2002	Nov	29
09-11-1997	25	1997	Nov	9
12-09-2001	21	2001	Sep	12

Format ‘charges’ column as currency ($).

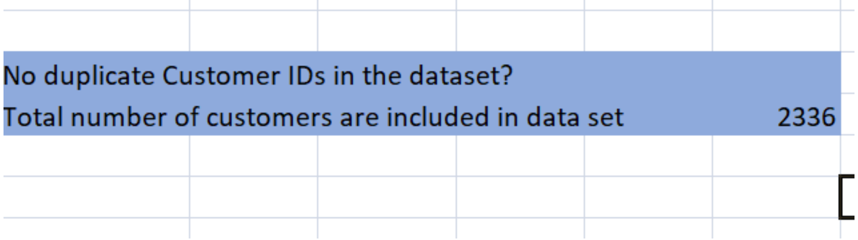


	H	I
	charges	Hospital tier
0	$ 563.84	tier - 2
0	$ 570.62	tier - 2
0	$ 600.00	tier - 2
0	$ 604.54	tier - 3
0	$ 637.26	tier - 3
0	$ 646.14	tier - 3
0	$ 650.00	tier - 3
0	$ 650.00	tier - 3
0	$ 668.00	tier - 3
0	$ 670.00	tier - 3
0	$ 687.54	tier - 3
0	$ 700.00	tier - 3

Data Exploration

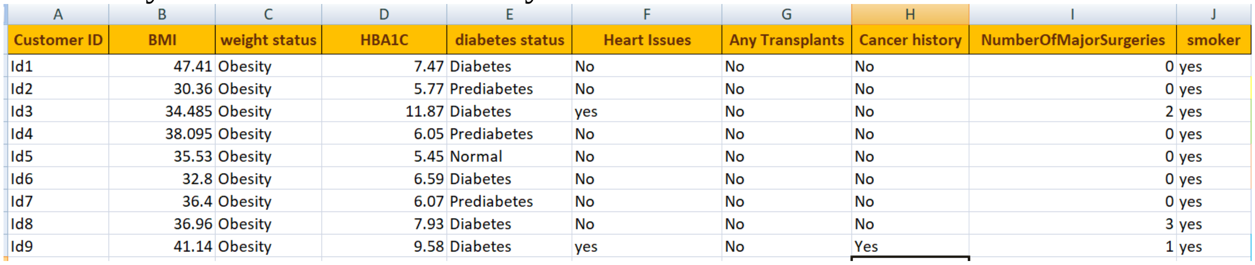
Medical Examination

Are there any duplicate Customer IDs in the dataset? If yes, how many?
How many customers are included in the dataset.



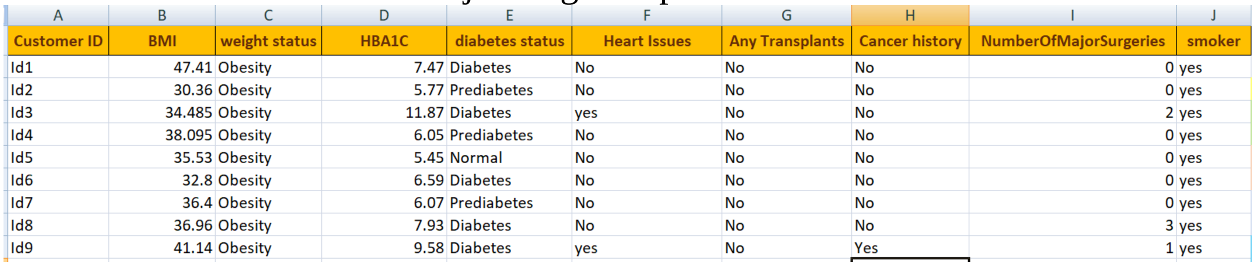
No duplicate Customer IDs in the dataset?
Total number of customers are included in data set 2336

How many customers have a history of cancer?



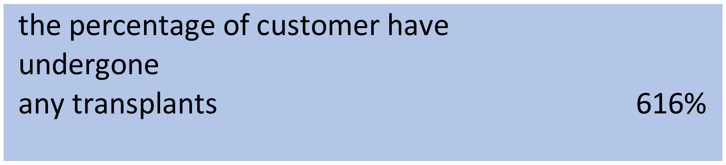
A	B	C	D	E	F	G	H	I	J
Customer ID	BMI	weight status	HBA1C	diabetes status	Heart Issues	Any Transplants	Cancer history	NumberOfMajorSurgeries	smoker
Id1	47.41	Obesity	7.47	Diabetes	No	No	No	0	yes
Id2	30.36	Obesity	5.77	Prediabetes	No	No	No	0	yes
Id3	34.485	Obesity	11.87	Diabetes	yes	No	No	2	yes
Id4	38.095	Obesity	6.05	Prediabetes	No	No	No	0	yes
Id5	35.53	Obesity	5.45	Normal	No	No	No	0	yes
Id6	32.8	Obesity	6.59	Diabetes	No	No	No	0	yes
Id7	36.4	Obesity	6.07	Prediabetes	No	No	No	0	yes
Id8	36.96	Obesity	7.93	Diabetes	No	No	No	3	yes
Id9	41.14	Obesity	9.58	Diabetes	yes	No	Yes	1	yes

What is the total number of major surgeries performed on customers?



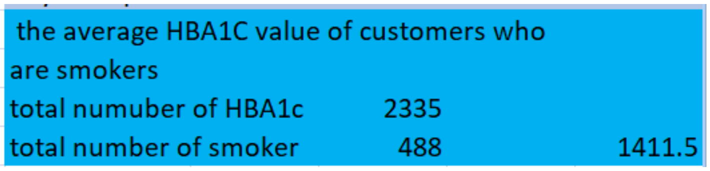
A	B	C	D	E	F	G	H	I	J
Customer ID	BMI	weight status	HBA1C	diabetes status	Heart Issues	Any Transplants	Cancer history	NumberOfMajorSurgeries	smoker
Id1	47.41	Obesity	7.47	Diabetes	No	No	No	0	yes
Id2	30.36	Obesity	5.77	Prediabetes	No	No	No	0	yes
Id3	34.485	Obesity	11.87	Diabetes	yes	No	No	2	yes
Id4	38.095	Obesity	6.05	Prediabetes	No	No	No	0	yes
Id5	35.53	Obesity	5.45	Normal	No	No	No	0	yes
Id6	32.8	Obesity	6.59	Diabetes	No	No	No	0	yes
Id7	36.4	Obesity	6.07	Prediabetes	No	No	No	0	yes
Id8	36.96	Obesity	7.93	Diabetes	No	No	No	3	yes
Id9	41.14	Obesity	9.58	Diabetes	yes	No	Yes	1	yes

Calculate the percentage of customers who have undergone any transplants.



the percentage of customer have undergone any transplants	616%

Find the average HBA1C value of customers who are smokers

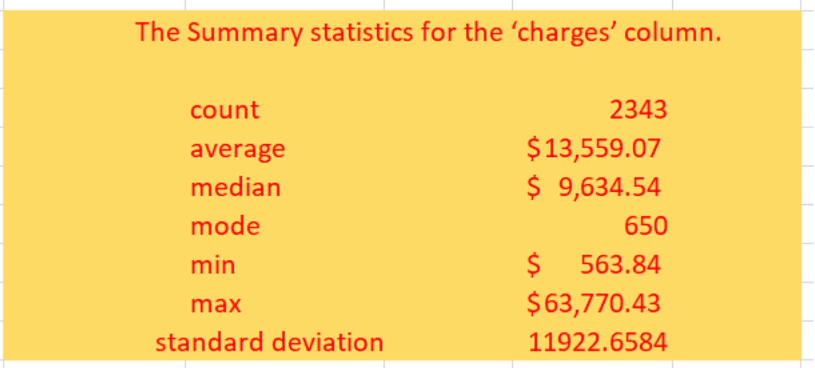


the average HBA1C value of customers who are smokers

total numuber of HBA1c	2335	
total number of smoker	488	1411.5

Hospitalization details

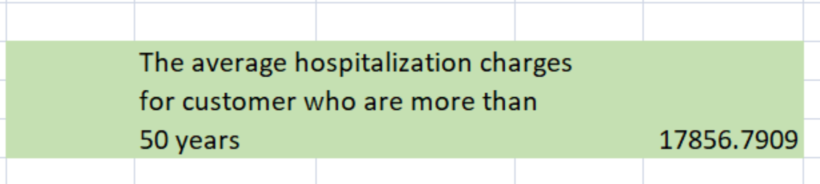
Calculate all the Summary statistics for the ‘charges’ column



The Summary statistics for the ‘charges’ column.

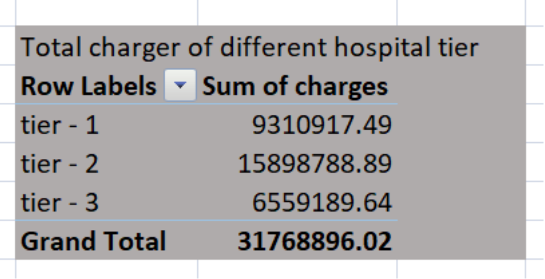
count	2343
average	$13,559.07
median	$ 9,634.54
mode	650
min	$ 563.84
max	$63,770.43
standard deviation	11922.6584

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



The average hospitalization charges for customer who are more than 50 years	17856.7909

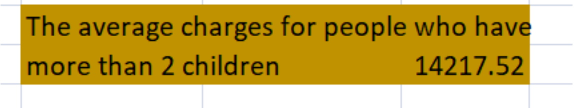
The total charges across different hospital tiers.



Total charger of different hospital tier

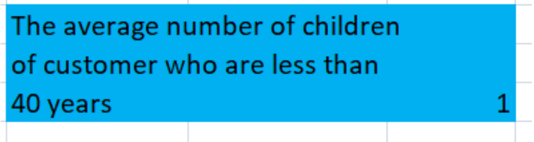
Row Labels	Sum of charges
tier - 1	9310917.49
tier - 2	15898788.89
tier - 3	6559189.64
Grand Total	**31768896.02**

The average charges for people who have more than 2 children



The average charges for people who have more than 2 children	14217.52

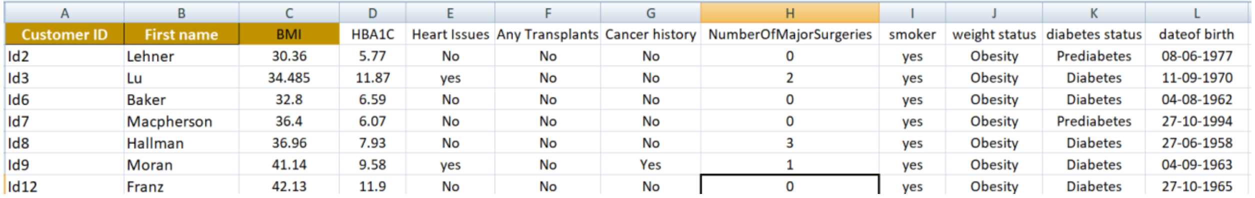
The integer average number of children of customers who are less than 40 years old.



The average number of children of customer who are less than 40 years	1

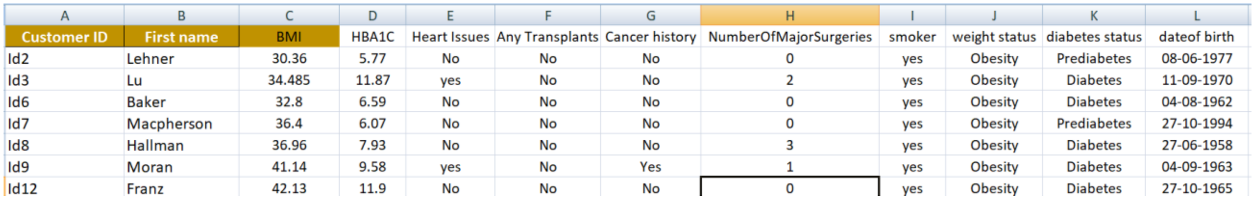
Data Analysis

Create a new sheet named “Healthcare", combine all three tables into one, using Customer ID as the common column, utilizing VLOOKUP.

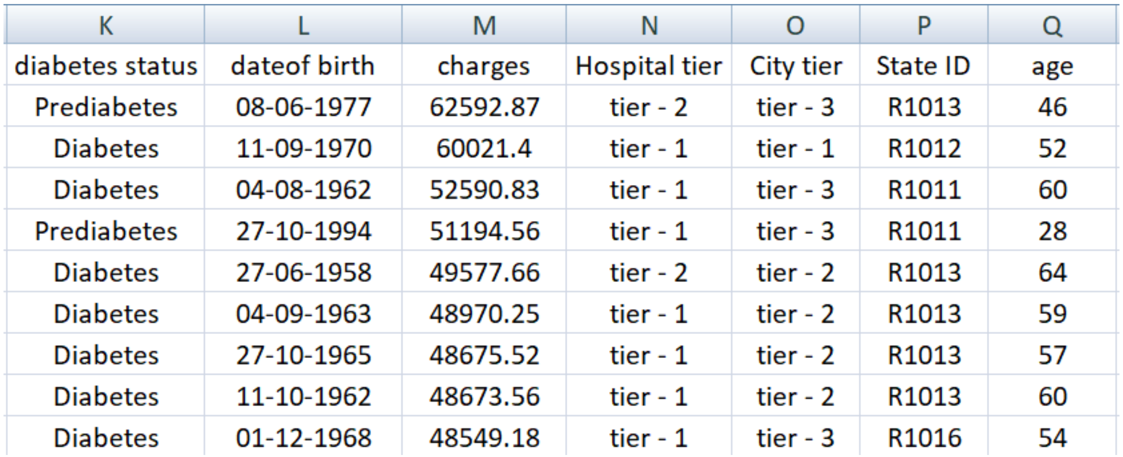


A	B	C	D	E	F	G	H	I	J	K	L
Customer ID	First name	BMI	HBA1C	Heart Issues	Any Transplants	Cancer history	NumberOfMajorSurgeries	smoker	weight status	diabetes status	dateof birth
Id2	Lehner	30.36	5.77	No	No	No	0	yes	Obesity	Prediabetes	08-06-1977
Id3	Lu	34.485	11.87	yes	No	No	2	yes	Obesity	Diabetes	11-09-1970
Id6	Baker	32.8	6.59	No	No	No	0	yes	Obesity	Diabetes	04-08-1962
Id7	Macpherson	36.4	6.07	No	No	No	0	yes	Obesity	Prediabetes	27-10-1994
Id8	Hallman	36.96	7.93	No	No	No	3	yes	Obesity	Diabetes	27-06-1958
Id9	Moran	41.14	9.58	yes	No	Yes	1	yes	Obesity	Diabetes	04-09-1963
Id12	Franz	42.13	11.9	No	No	No	0	yes	Obesity	Diabetes	27-10-1965

Retain the following necessary columns: 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.



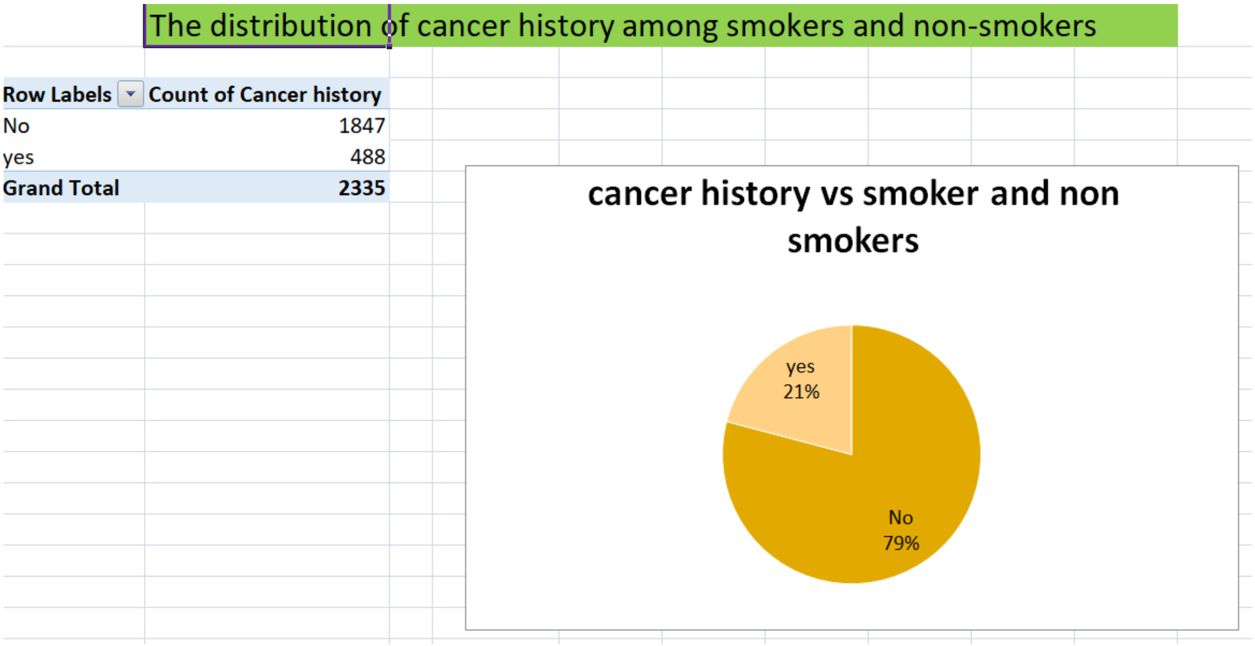
A	B	C	D	E	F	G	H	I	J	K	L
Customer ID	First name	BMI	HBA1C	Heart Issues	Any Transplants	Cancer history	NumberOfMajorSurgeries	smoker	weight status	diabetes status	dateof birth
Id2	Lehner	30.36	5.77	No	No	No	0	yes	Obesity	Prediabetes	08-06-1977
Id3	Lu	34.485	11.87	yes	No	No	2	yes	Obesity	Diabetes	11-09-1970
Id6	Baker	32.8	6.59	No	No	No	0	yes	Obesity	Diabetes	04-08-1962
Id7	Macpherson	36.4	6.07	No	No	No	0	yes	Obesity	Prediabetes	27-10-1994
Id8	Hallman	36.96	7.93	No	No	No	3	yes	Obesity	Diabetes	27-06-1958
Id9	Moran	41.14	9.58	yes	No	Yes	1	yes	Obesity	Diabetes	04-09-1963
Id12	Franz	42.13	11.9	No	No	No	0	yes	Obesity	Diabetes	27-10-1965



K	L	M	N	O	P	Q
diabetes status	dateof birth	charges	Hospital tier	City tier	State ID	age
Prediabetes	08-06-1977	62592.87	tier - 2	tier - 3	R1013	46
Diabetes	11-09-1970	60021.4	tier - 1	tier - 1	R1012	52
Diabetes	04-08-1962	52590.83	tier - 1	tier - 3	R1011	60
Prediabetes	27-10-1994	51194.56	tier - 1	tier - 3	R1011	28
Diabetes	27-06-1958	49577.66	tier - 2	tier - 2	R1013	64
Diabetes	04-09-1963	48970.25	tier - 1	tier - 2	R1013	59
Diabetes	27-10-1965	48675.52	tier - 1	tier - 2	R1013	57
Diabetes	11-10-1962	48673.56	tier - 1	tier - 2	R1013	60
Diabetes	01-12-1968	48549.18	tier - 1	tier - 3	R1016	54

Analysis using Pie/Donut Chart

The distribution of cancer history among smokers and non-smokers

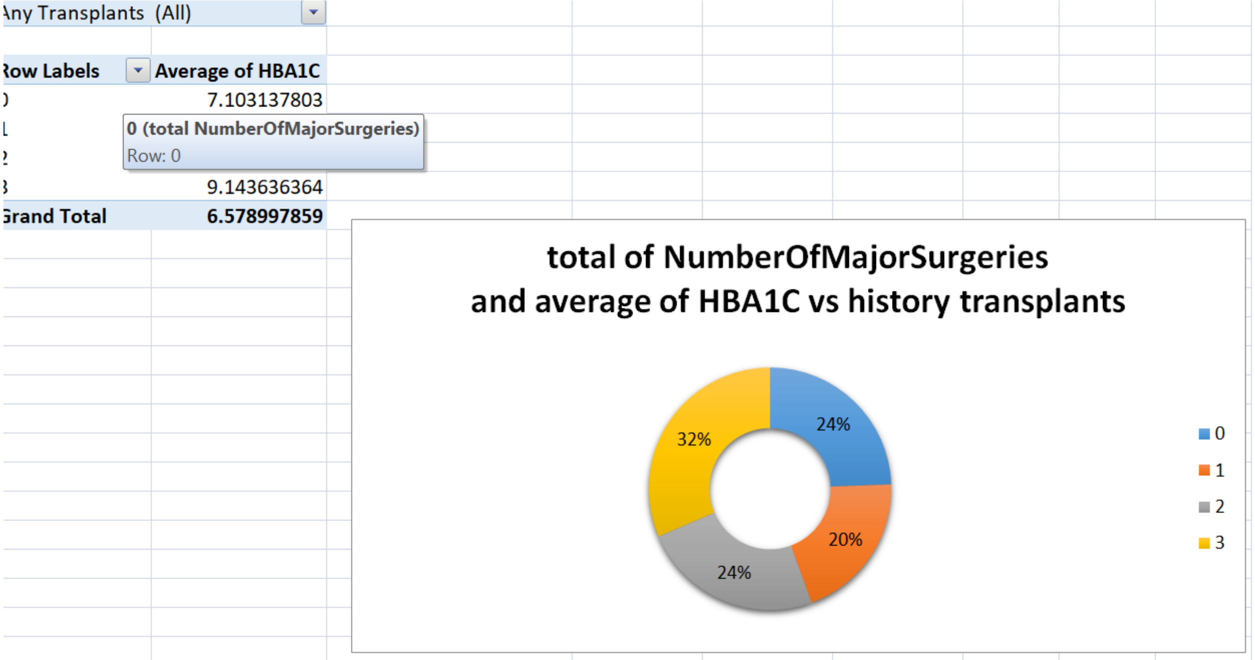


The distribution of cancer history among smokers and non-smokers

Row Labels	Count of Cancer history
No	1847
yes	488
Grand Total	2335

cancer history vs smoker and non smokers

The total number of major surgeries and average HbA1C differ between patients with and without a history of transplants



Any Transplants (All)

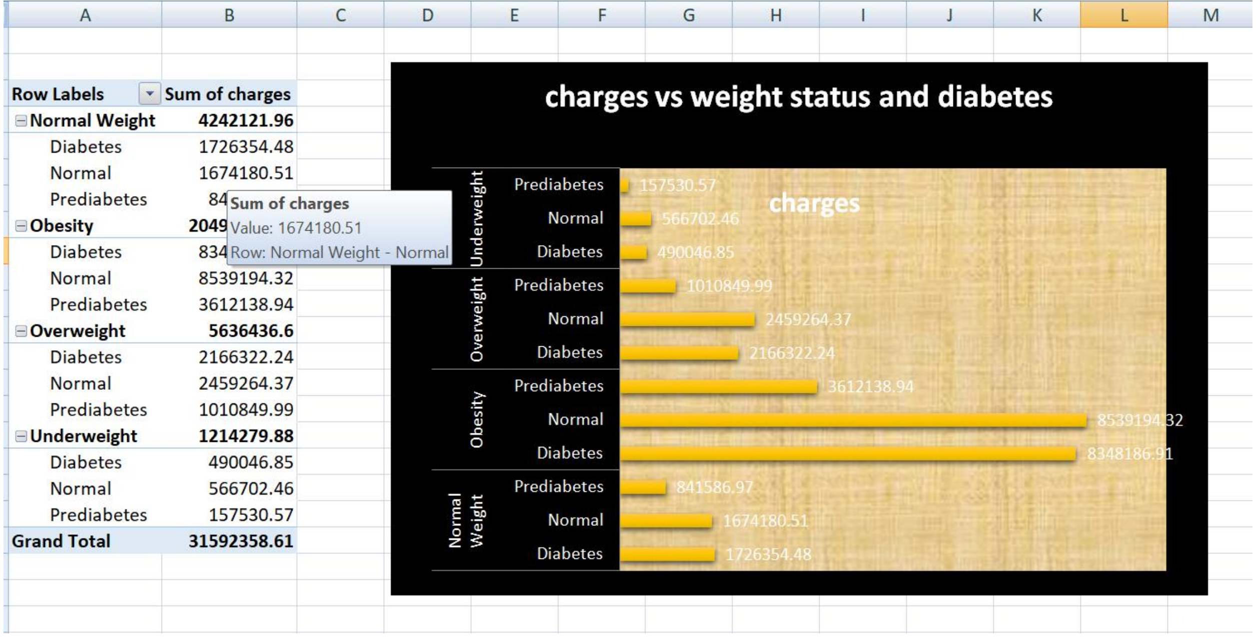
Row Labels	Average of HBA1C
0	7.103137803
1	
2	
3	9.143636364
Grand Total	6.578997859

0 (total NumberOfMajorSurgeries)
Row: 0

total of NumberOfMajorSurgeries and average of HBA1C vs history transplants

Analysis using Column/Bar Chart

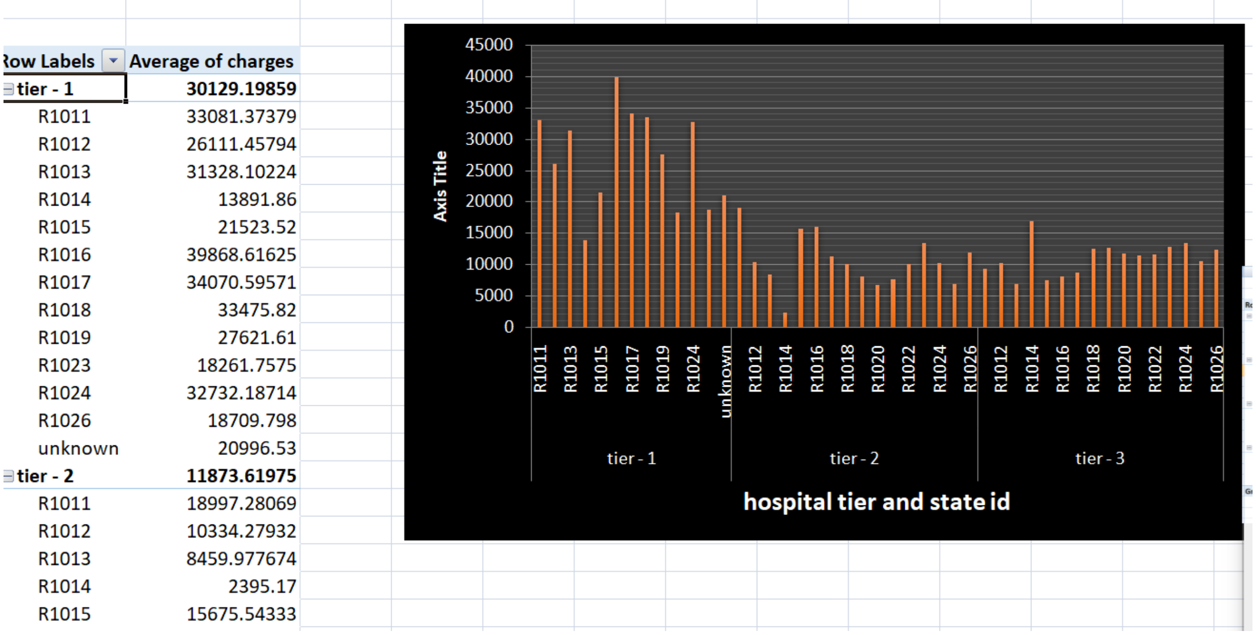
Healthcare charges vary based on different weight statuses and diabetes statuses



Row Labels	Sum of charges
Normal Weight	**4242121.96**
Diabetes	1726354.48
Normal	1674180.51
Prediabetes	84
Obesity	**2049**
Diabetes	834
Normal	8539194.32
Prediabetes	3612138.94
Overweight	**5636436.6**
Diabetes	2166322.24
Normal	2459264.37
Prediabetes	1010849.99
Underweight	**1214279.88**
Diabetes	490046.85
Normal	566702.46
Prediabetes	157530.57
Grand Total	**31592358.61**

Sum of charges
Value: 1674180.51
Row: Normal Weight - Normal

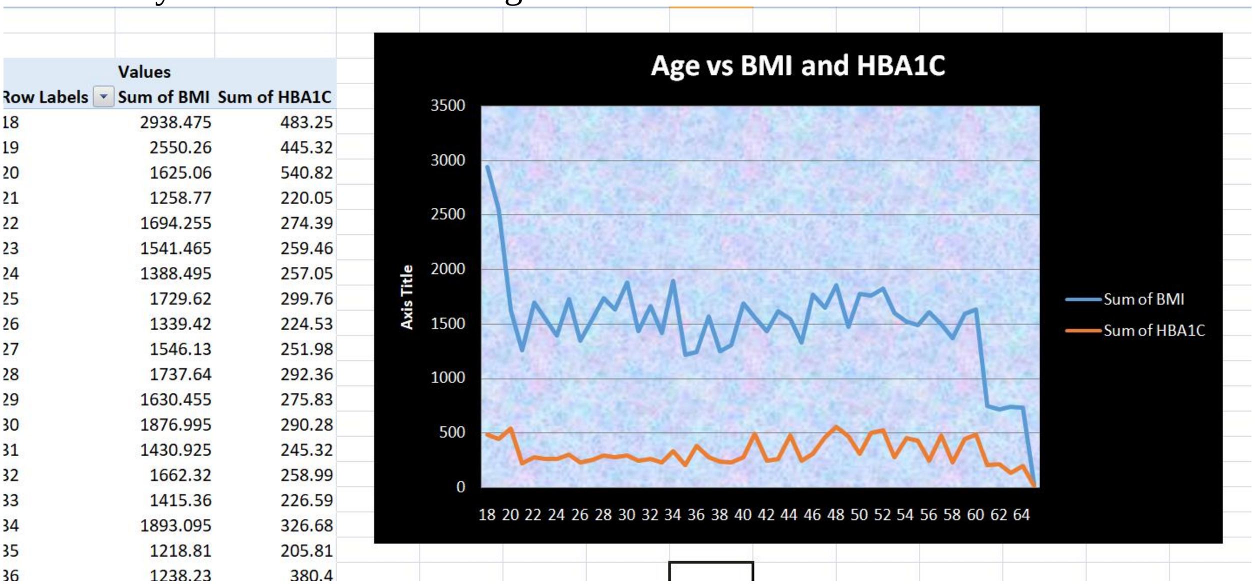
The average charges for each hospital tier within different states



Row Labels	Average of charges
tier - 1	**30129.19859**
R1011	33081.37379
R1012	26111.45794
R1013	31328.10224
R1014	13891.86
R1015	21523.52
R1016	39868.61625
R1017	34070.59571
R1018	33475.82
R1019	27621.61
R1023	18261.7575
R1024	32732.18714
R1026	18709.798
unknown	20996.53
tier - 2	**11873.61975**
R1011	18997.28069
R1012	10334.27932
R1013	8459.977674
R1014	2395.17
R1015	15675.54333

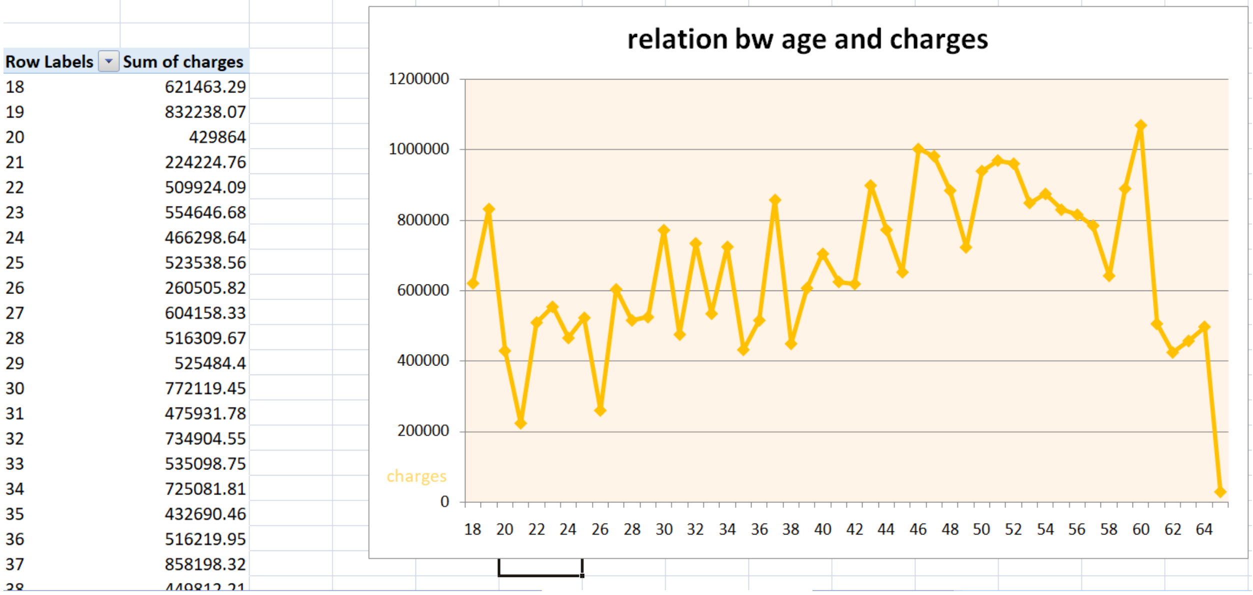
Analysis using Line/Scatter Plot

Is there any correlation between age and both BMI and HbA1C in the dataset



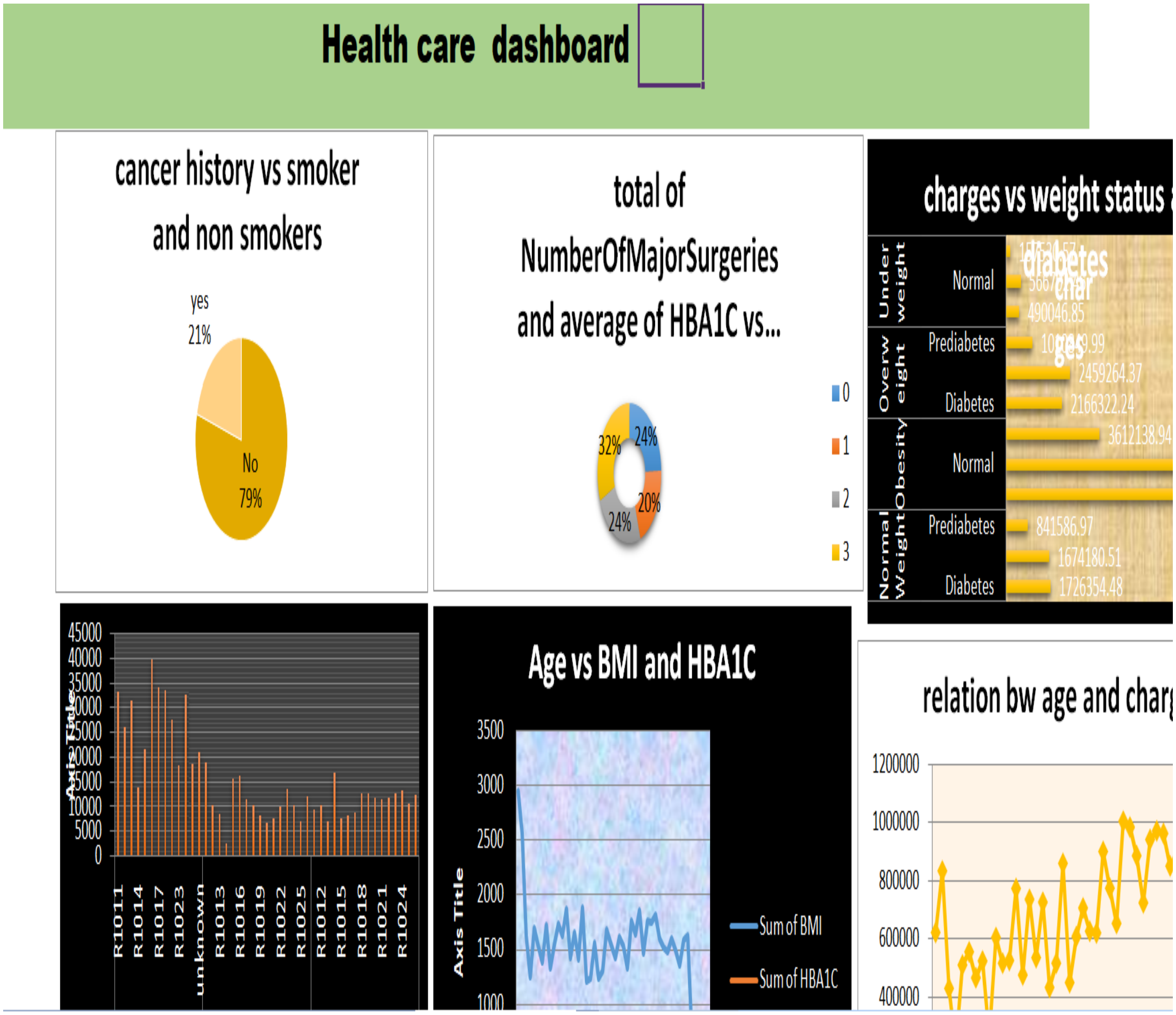
	Values	
Row Labels	Sum of BMI	Sum of HBA1C
18	2938.475	483.25
19	2550.26	445.32
20	1625.06	540.82
21	1258.77	220.05
22	1694.255	274.39
23	1541.465	259.46
24	1388.495	257.05
25	1729.62	299.76
26	1339.42	224.53
27	1546.13	251.98
28	1737.64	292.36
29	1630.455	275.83
30	1876.995	290.28
31	1430.925	245.32
32	1662.32	258.99
33	1415.36	226.59
34	1893.095	326.68
35	1218.81	205.81
36	1238.23	380.4

Explore the relationship between age and healthcare charges



Row Labels	Sum of charges
18	621463.29
19	832238.07
20	429864
21	224224.76
22	509924.09
23	554646.68
24	466298.64
25	523538.56
26	260505.82
27	604158.33
28	516309.67
29	525484.4
30	772119.45
31	475931.78
32	734904.55
33	535098.75
34	725081.81
35	432690.46
36	516219.95
37	858198.32
38	449812.21

Dash Board


Health care dashboard
cancer history vs smoker and non smokers
yes 21%
No 79%
total of NumberOfMajorSurgeries and average of HBA1C vs...
24%
20%
24%
32%
0
1
2
3
charges vs weight status
Under weight
Normal
Overw eight
Prediabetes
Diabetes
Obesity
Normal
Normal Weight
Prediabetes
Diabetes
490046.85
2459264.37
2166322.24
3612138.94
841586.97
1674180.51
1726354.48
diabetes char ges
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
Axis Title
R1011
R1014
R1017
R1023
unknown
R1013
R1016
R1019
R1022
R1025
R1012
R1015
R1018
R1021
R1024
Age vs BMI and HBA1C
3500
3000
2500
2000
1500
1000
Axis Title
Sum of BMI
Sum of HBA1C
relation bw age and charg
1200000
1000000
800000
600000
400000
