

DAX and Basic Visualizations

Venkata Reddy Konasani

Contents

- DAX formulas
- Basic Visualizations
- Working with tables
- Working with measures
- Univariate Analysis

Stroke Case study

Step 5 - Univariate Analysis

Categorical Variables Exploration

General_Details_table

- ☐ Σ age
- ☐ ☒ Age_high_risk
- ☐ ever_married
- ☐ gender
- ☐ ☒ High_Risk_factor2
- ☐ id
- ☐ Residence_type
- ☐ work_type_Cleaned1

Risk_Factors_Table

- ☐ Σ avg_glucose_level
- ☐ Σ bmi
- ☐ Σ heart_disease
- ☐ Σ hypertension
- ☐ Patient_id
- ☐ smoking_status
- ☐ Σ stroke

Create Frequency Tables for categorical Variables

Age_high_risk	Count of id
Low Risk Age	3418
High Risk Age	1692
Total	5110

ever_married	Count of id
Yes	3336
No	1753
NA	21
Total	5110

Residence_type	Count of id
Urban	2593
Rural	2511
NA	6
Total	5110

smoking_status	Count of id
never smoked	1889
Unknown	1552
formerly smoked	882
smokes	787
Total	5110

gender	Count of id
Female	2985
Male	2113
NA	11
Other	1
Total	5110

High_Risk_factor2	Count of id
Low Risk2	4949
High Risk2	161
Total	5110

work_type_Cleaned1	Count of id
Private	2921
Self-employed	818
children	687
Govt_job	656
Never_worked	22
NA	6
Total	5110

Display Percentages along with counts

Percent_of_total =
`COUNT(General_Details_table[id])/COUNTROWS(ALL(General_Details_table[id]))`

Age_high_risk	Count of id
Low Risk Age	3418
High Risk Age	1692
Total	5110

ALL() is used to count rows beyond the selected context

Display Percentages along with counts

Age_high_risk	Count of id	Percent_of_total
Low Risk Age	3418	66.89%
High Risk Age	1692	33.11%
Total	5110	100.00%

Result will be a decimal value. Click on the measure and change the format

Final Result

Null values here can be cleaned

ever_married	Count of id	Percent_of_total
NA	21	0.41%
No	1753	34.31%
Yes	3336	65.28%
Total	5110	100.00%

Residence_type	Count of id	Percent_of_total
NA	6	0.12%
Rural	2511	49.14%
Urban	2593	50.74%
Total	5110	100.00%

smoking_status	Count of id	Percent_of_total
NA	12	0.23%
formerly smoked	882	17.26%
never smoked	1889	36.97%
smokes	787	15.40%
Unknown	1540	30.14%
Total	5110	100.00%

gender	Count of id	Percent_of_total
Female	2985	58.41%
Male	2113	41.35%
NA	11	0.22%
Other	1	0.02%
Total	5110	100.00%

Work_type_Clean	Count of id	Percent_of_total
children	687	13.44%
Govt_job	656	12.84%
NA	6	0.12%
Never_worked	22	0.43%
Private	2921	57.16%
Self-employed	818	16.01%
Total	5110	100.00%

Risk_Factor1	Count of id	Percent_of_total
High Risk Age	1692	33.11%
Low Risk Age	3418	66.89%
Total	5110	100.00%

Risk_Factor2	Count of id	Percent_of_total
High Risk2	161	3.15%
Low Risk	4949	96.85%
Total	5110	100.00%

Further Cleaning Smoking Status

×

Replace Values

Replace one value with another in the selected columns.

Value To Find

Replace With

> Advanced options

OKCancel

Final Result

Age_high_risk	Count of id	Percent_of_total
Low Risk Age	3418	66.89%
High Risk Age	1692	33.11%
Total	5110	100.00%

Residence_type	Count of id	Percent_of_total
Urban	2593	50.74%
Rural	2511	49.14%
NA	6	0.12%
Total	5110	100.00%

gender	Count of id	Percent_of_total
Female	2985	58.41%
Male	2113	41.35%
NA	11	0.22%
Other	1	0.02%
Total	5110	100.00%

ever_married	Count of id	Percent_of_total
Yes	3336	65.28%
No	1753	34.31%
NA	21	0.41%
Total	5110	100.00%

smoking_status	Count of id	Percent_of_total
never smoked	1889	36.97%
Unknown	1552	30.37%
formerly smoked	882	17.26%
smokes	787	15.40%
Total	5110	100.00%

High_Risk_factor2	Count of id	Percent_of_total
Low Risk2	4949	96.85%
High Risk2	161	3.15%
Total	5110	100.00%

work_type_Cleaned1	Count of id	Percent_of_total
Private	2921	57.16%
Self-employed	818	16.01%
children	687	13.44%
Govt_job	656	12.84%
Never_worked	22	0.43%
NA	6	0.12%
Total	5110	100.00%

Interpretation

- Observe the percentage of each category
- Make a comment on top2 or top3 maximum frequency and minimum frequency items.
- For example, observe whether the data contain equal percentage of Male and Female population?

Discrete Variables

General_Details_table

☐ Σ age

☐ \mathbb{E}_{fx} Age_high_risk

☐ ever_married

☐ gender

☐ \mathbb{E}_{fx} High_Risk_factor2

☐ id

☐ Residence_type

☐ work_type_Cleaned1

Risk_Factors_Table

☐ Σ avg_glucose_level

☐ Σ bmi

☐ Σ heart_disease

☐ Σ hypertension

☐ Patient_id

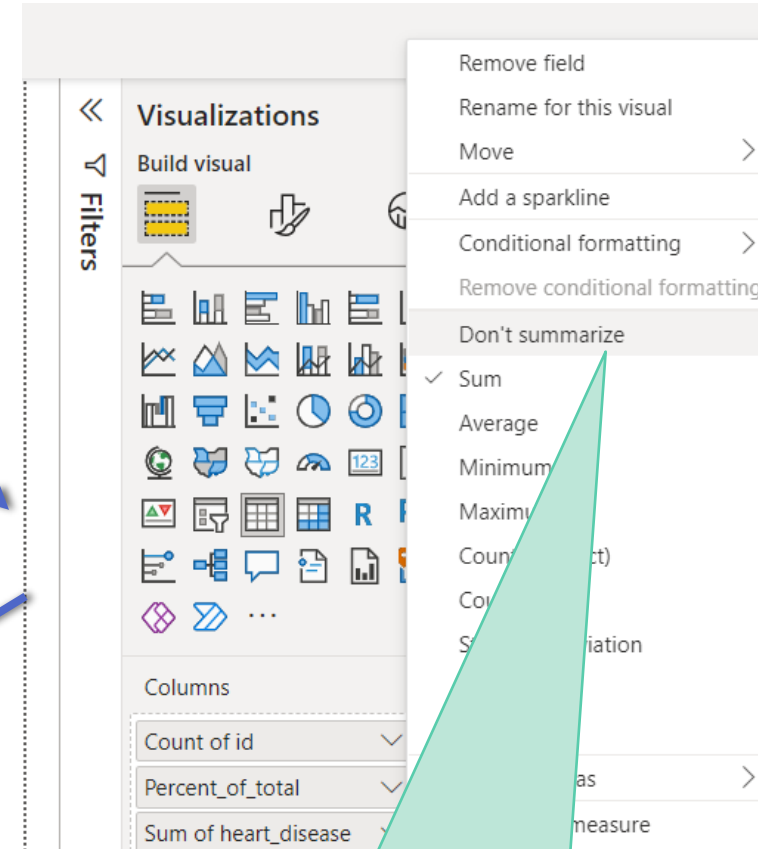
☐ smoking_status

☐ Σ stroke

Discrete Variables – Exploration

Count of id	Percent_of_total	Sum of heart_disease
5110	100.00%	275

heart_disease	Count of id	Percent_of_total
	12	0.23%
0	4823	94.38%
1	275	5.38%
Total	5110	100.00%



Use this “Don’t summarize option”

Discrete Variables – Exploration

heart_disease	Count of id	Percent_of_total
0	4823	94.38%
1	275	5.38%
	12	0.23%
Total	5110	100.00%

stroke	Count of id	Percent_of_total
0	4861	95.13%
1	249	4.87%
Total	5110	100.00%

Count of id	Percent_of_total	hypertension
4601	90.04%	0
497	9.73%	1
12	0.23%	
5110	100.00%	

- Make a note of these missing values. We will handle them later

Continuous Variables

General_Details_table

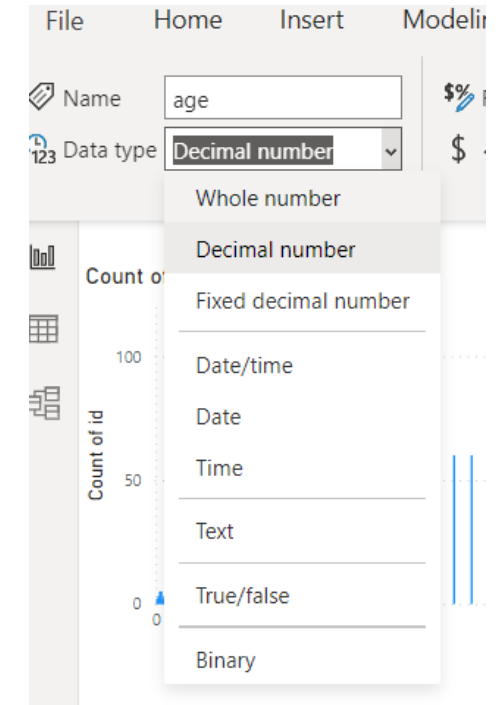
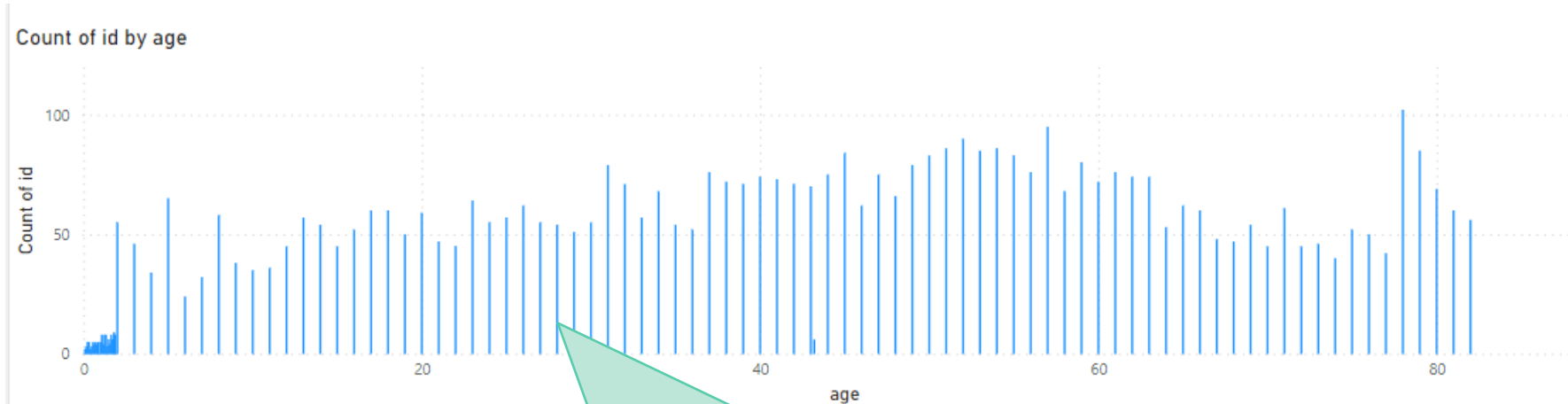
- ☐ \sum age
- ☐ \sum Age_high_risk
- ☐ ever_married
- ☐ gender
- ☐ \sum High_Risk_factor2
- ☐ id
- ☐ Residence_type
- ☐ work_type_Cleaned1

Risk_Factors_Table

- ☐ \sum avg_glucose_level
- ☐ \sum bmi
- ☐ \sum heart_disease
- ☐ \sum hypertension
- ☐ Patient_id
- ☐ smoking_status
- ☐ \sum stroke

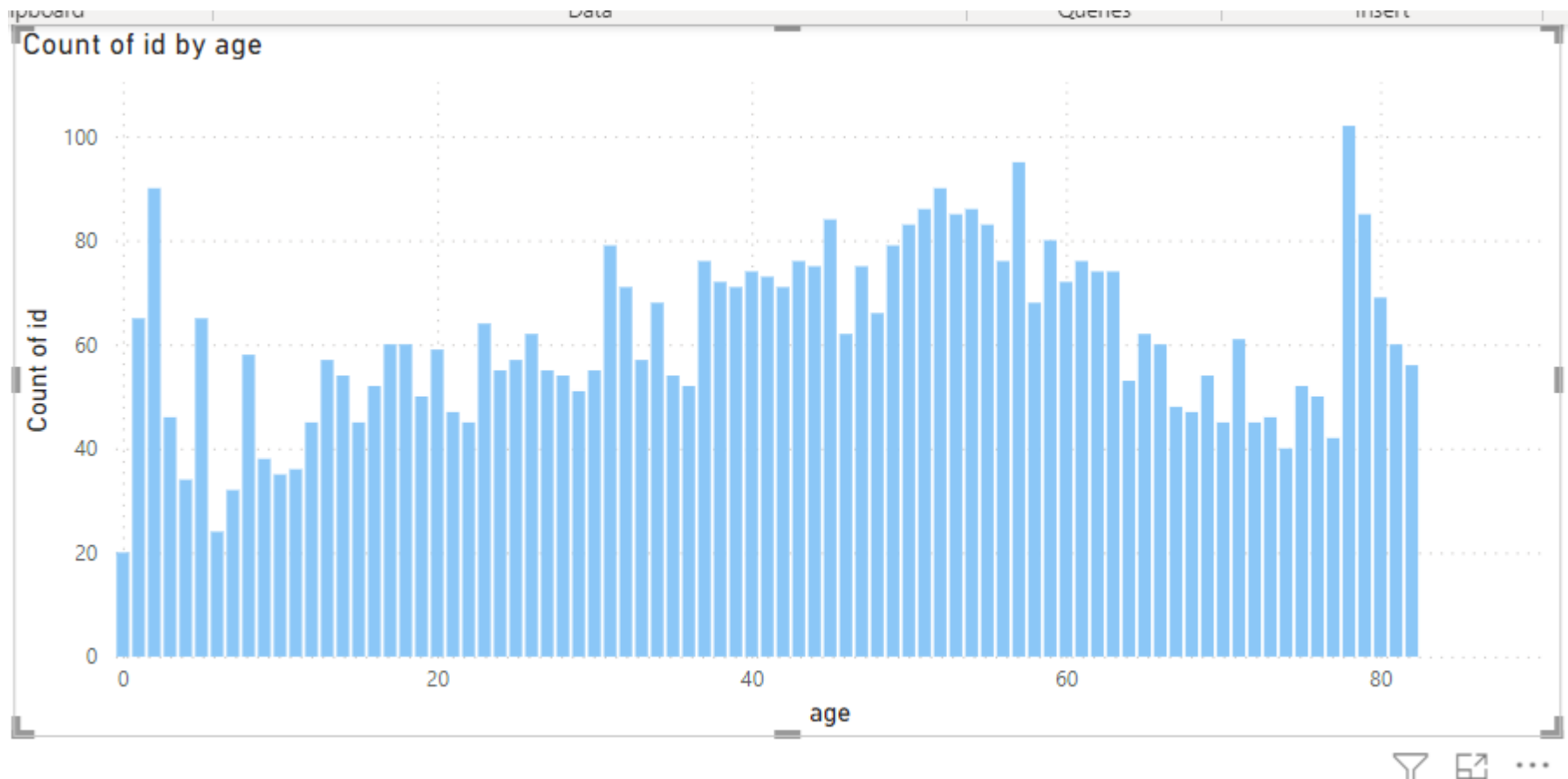
Distribution Charts

- Convert Age into an integer then draw the clustered column chart
- Clustered column chart >> age on X- axis >> Count of id on y-axis



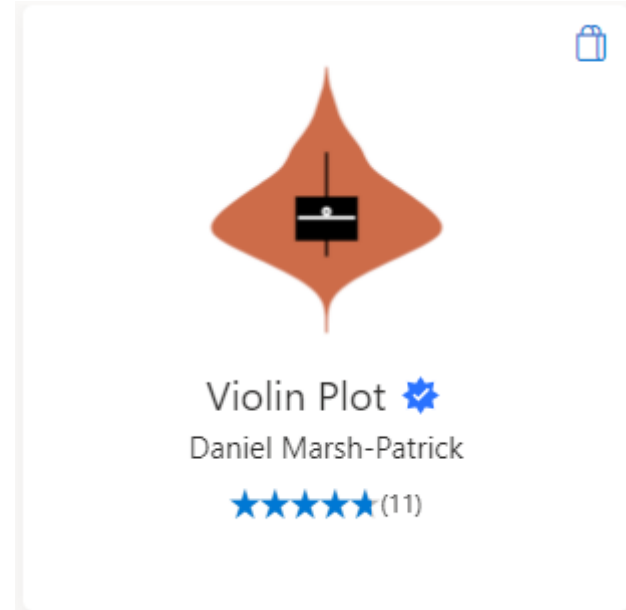
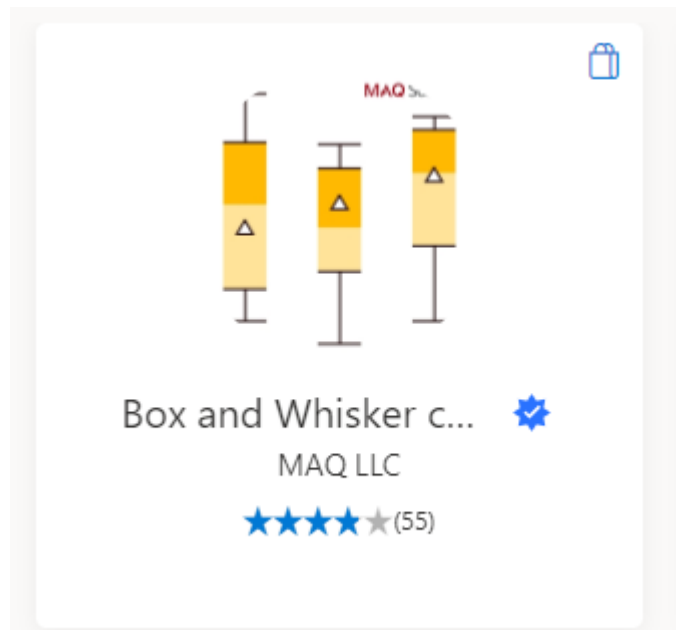
Distribution Charts

- Convert Age into an integer then draw the clustered column chart
- Clustered column chart >> age on X- axis >> Count of id on y-axis



Box Plot and Violin Chart

- Histogram, Box Plot and Violin plot almost tell the same story.
- Box Plot and Violin plots need to be downloaded using “Get more visuals” option



Get more visuals - option

If more visuals doesn't work, then you import the visual from a file

The screenshot shows the 'Get more visuals' dialog box in Power BI. The dialog has a title bar with a close button (X). Below the title bar, there are four options: 'Import a visual from a file', 'Remove a visual', 'Restore default visuals', and 'Get more visuals'. The 'Get more visuals' option is selected, and it points to a list of visualizations. The list is titled 'AppSource visuals' and includes a search bar and a 'Sort by: Popularity' dropdown. The visualizations are displayed in a grid, each with a thumbnail, a title, a developer name, and a star rating. The visualizations include: 'Box and Whisker c...' by MAQ LLC (56 stars), 'Play Axis (Dynamic ...' by mprozil (42 stars), 'Inforiver Premium ...' by xViz LLC (10 stars), 'Inforiver Charts' by xViz LLC (2 stars), 'Drill Down Timeline...' by ZoomCharts (10 stars), 'Dynamic KPI Card ...' by Entech SPA AG, 'Text Filter' by Microsoft Corporation, 'Chiclet Slicer' by Microsoft Corporation, 'Timeline Slicer' by Microsoft Corporation, and 'Gantt' by Microsoft Corporation.

Get more visuals

Import a visual from a file

Remove a visual

Restore default visuals

AppSource visuals

Search

Sort by: Popularity

Box and Whisker c...
MAQ LLC
★★★★★ (56)

Play Axis (Dynamic ...
mprozil
★★★★★ (42)

Inforiver Premium ...
xViz LLC
★★★★★ (10)

Inforiver Charts
xViz LLC
★★★★★ (2)

Drill Down Timeline...
ZoomCharts
★★★★★ (10)

Dynamic KPI Card ...
Entech SPA AG

Text Filter
Microsoft Corporation

Chiclet Slicer
Microsoft Corporation

Timeline Slicer
Microsoft Corporation

Gantt
Microsoft Corporation

Box Plot options & Result

Id on Axis and age on Value

Axis

id

Axis category I

Add data fields here

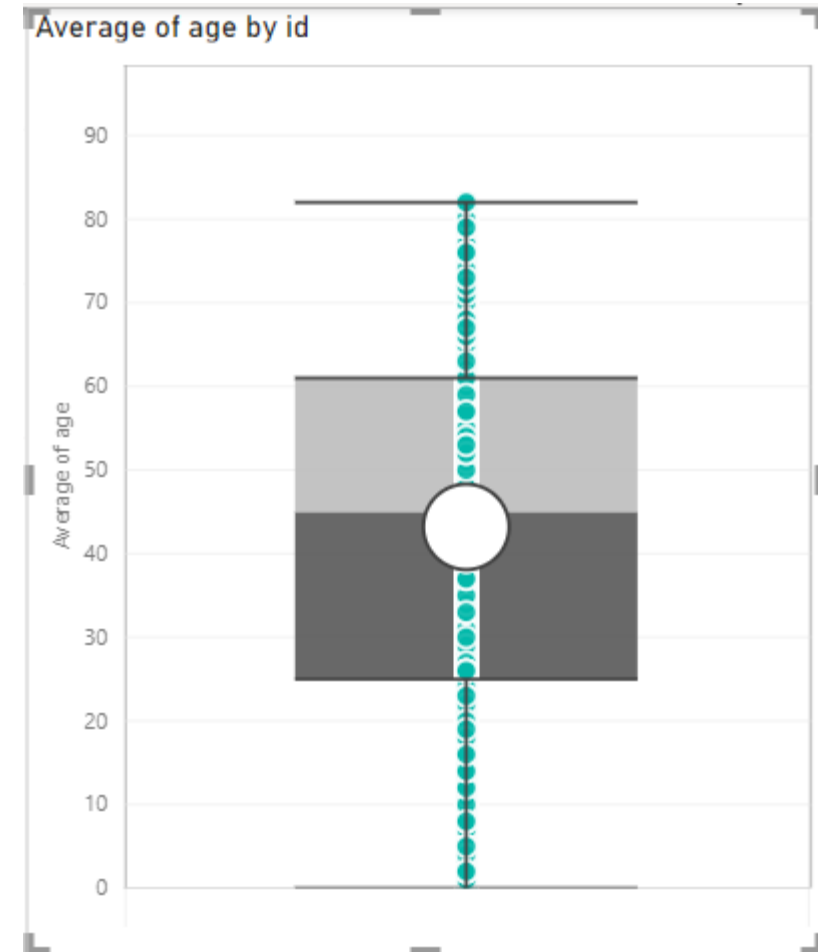
Axis category II

Add data fields here

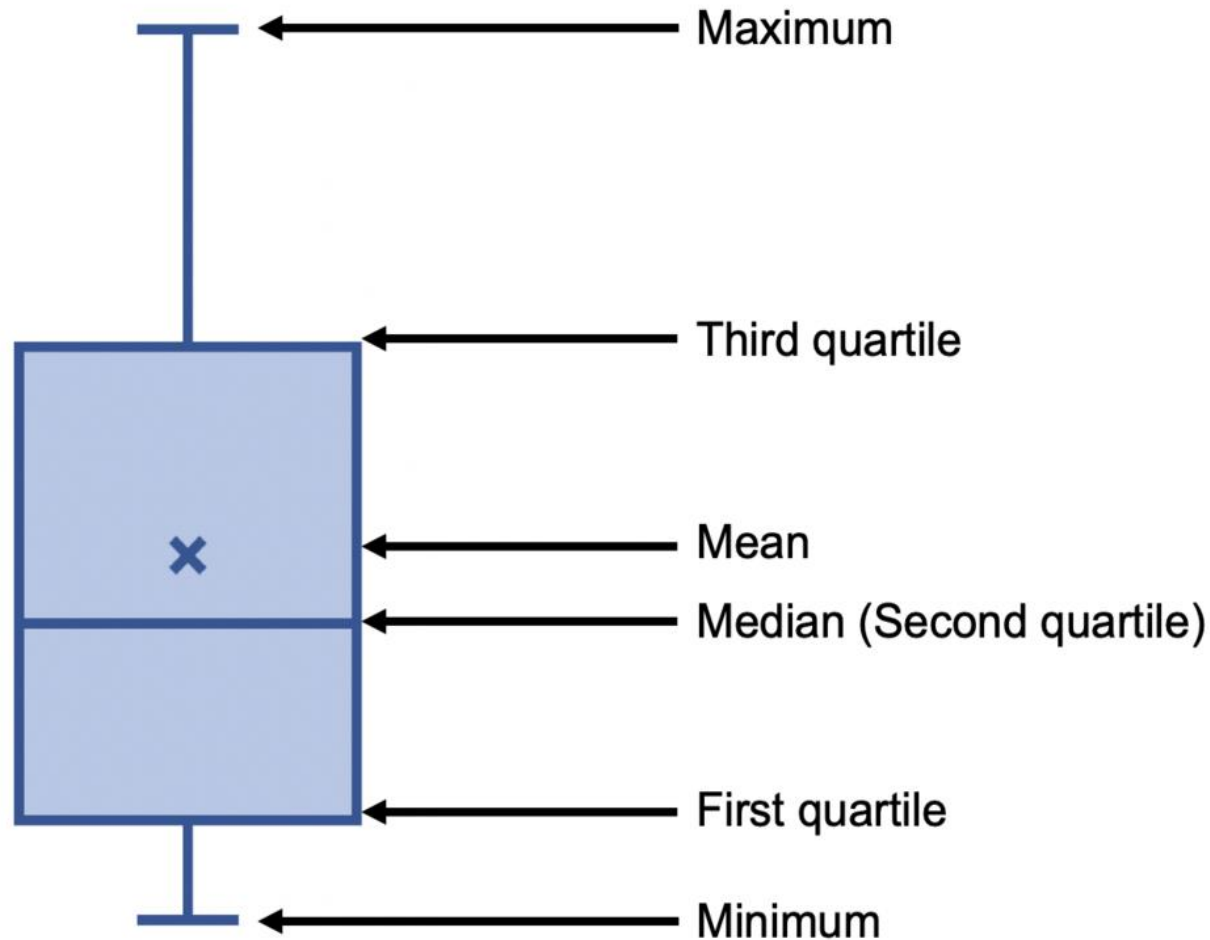
Value

Average of age

Dots size



How to interpret a box plot?



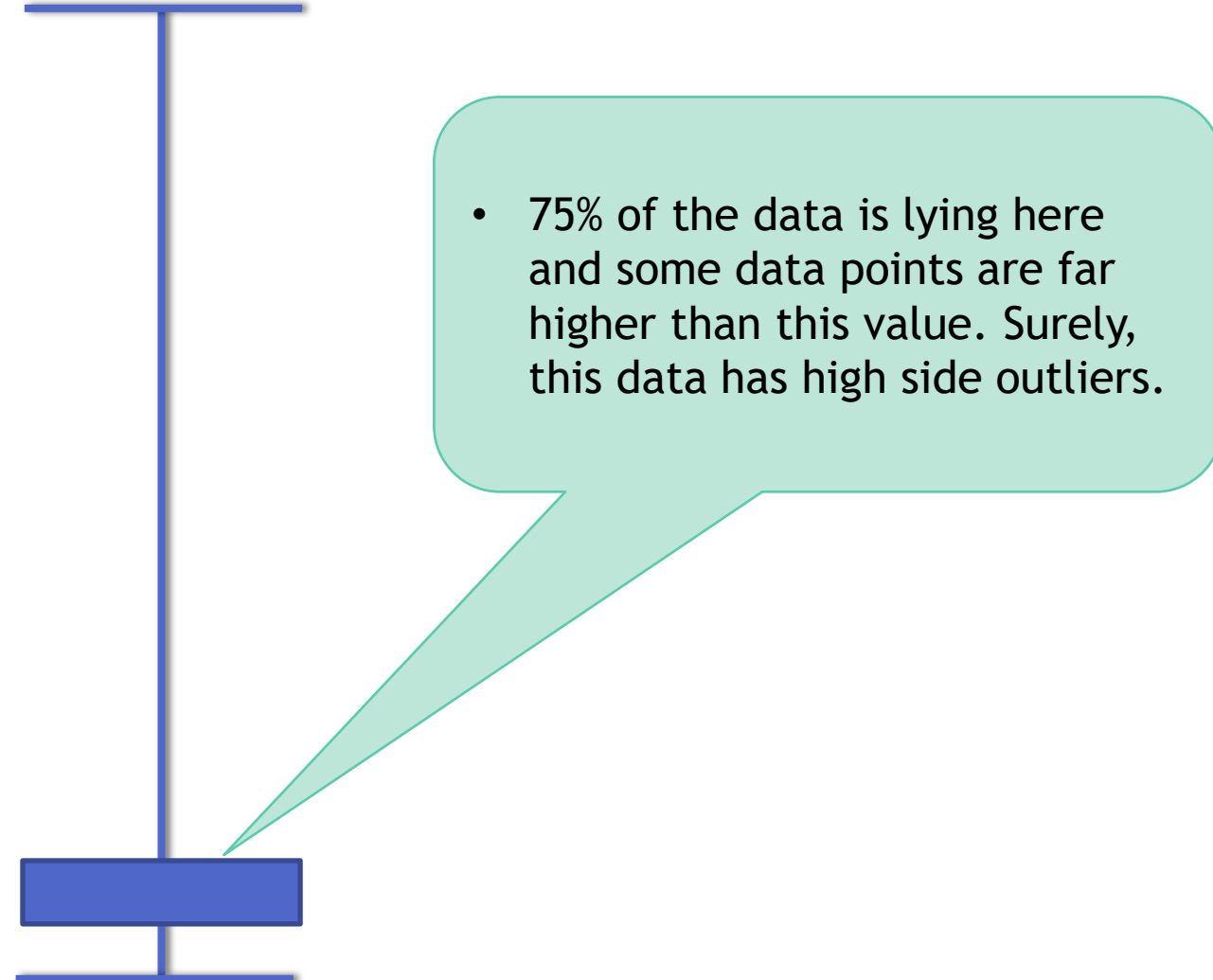
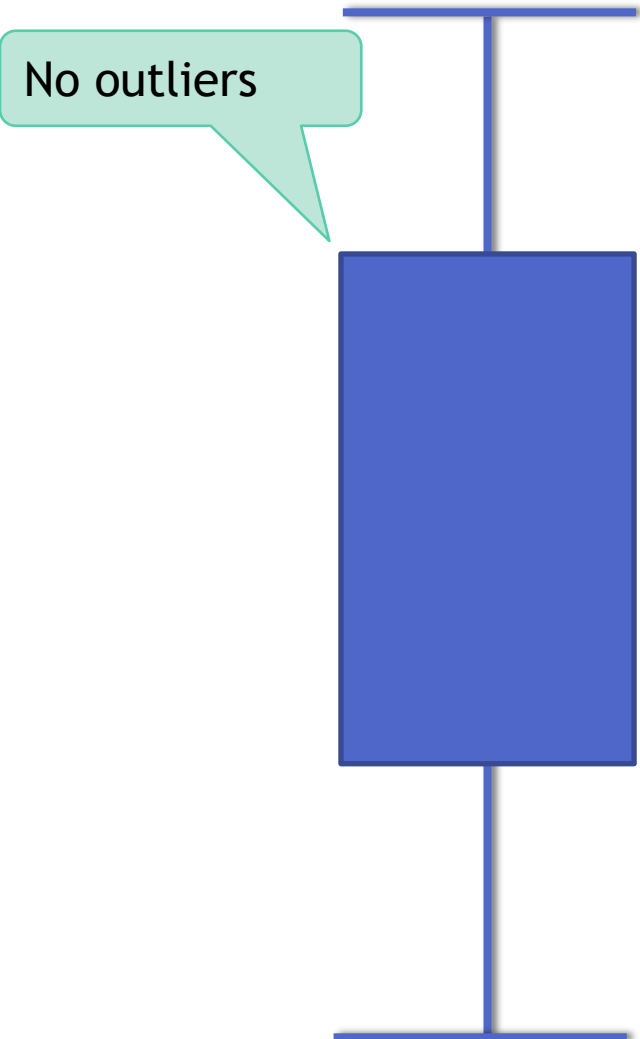
- 25% of the data lies below the first quartile value.
- 50% of the data lies below the second quartile value.
- 75% of the data lies below third quartile value.

How to interpret a box plot?

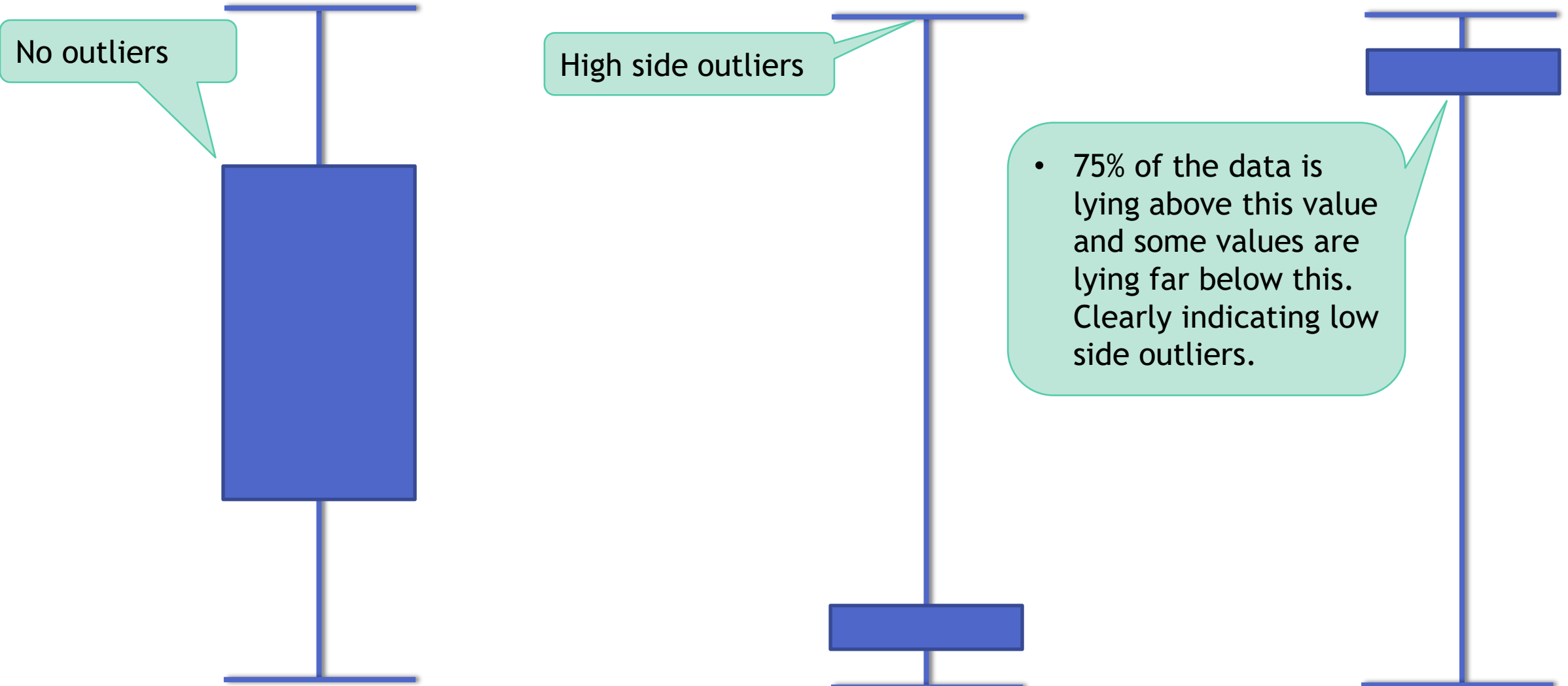


- If the box is well distributed like this, then there are no outliers in this variable

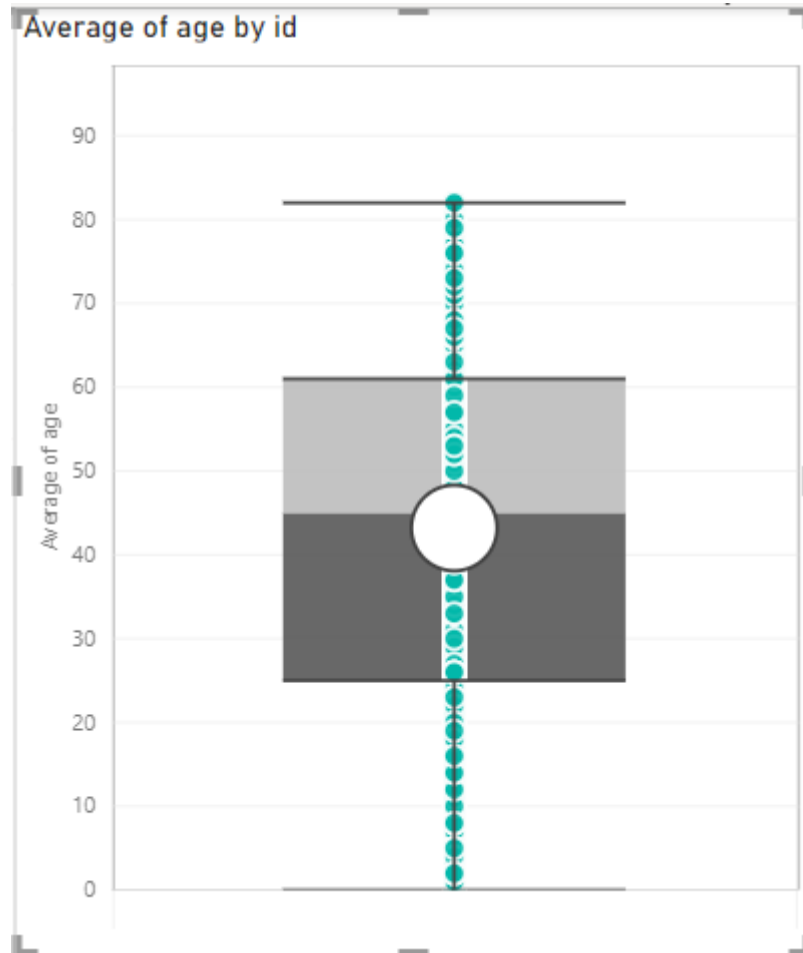
How to interpret a box plot?



How to interpret a box plot?



How to interpret a box plot?



Median Type	Inclusive
Whisker Type	Min/Max
Mean	43.20
Quartile 1	25.00
Median	45.00
Quartile 3	61.00
Maximum	82.00
Minimum	0.00
IQR	36.00
Upper Whisker	82.00
Lower Whisker	0.00

In the age variable

- 25% of the data lies below the first quartile value - $Q1=25$
- 50% of the data lies below the second quartile value- $Q2=45$
- 75% of the data lies below third quartile value. - 61
- There is a fair distribution, not many outliers are observed.

Violin Plot options & Result

Sampling

id

Measure Data

Average of age

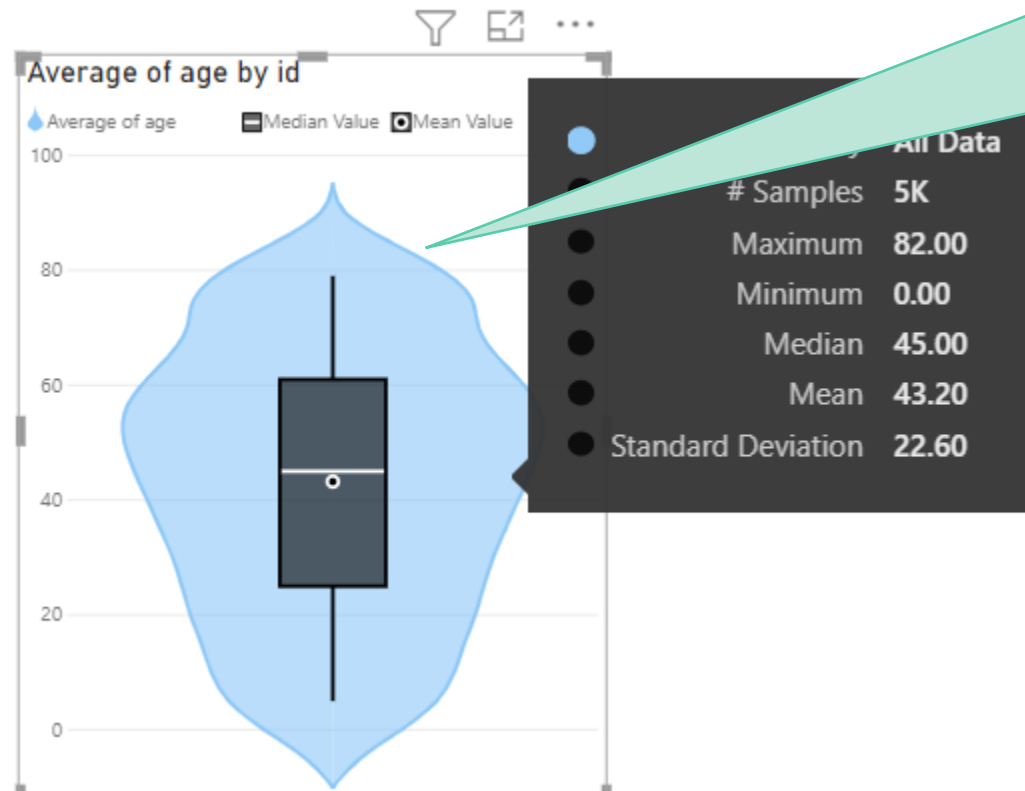
Category

Add data fields here

Drill through

Cross-report ☐ Off

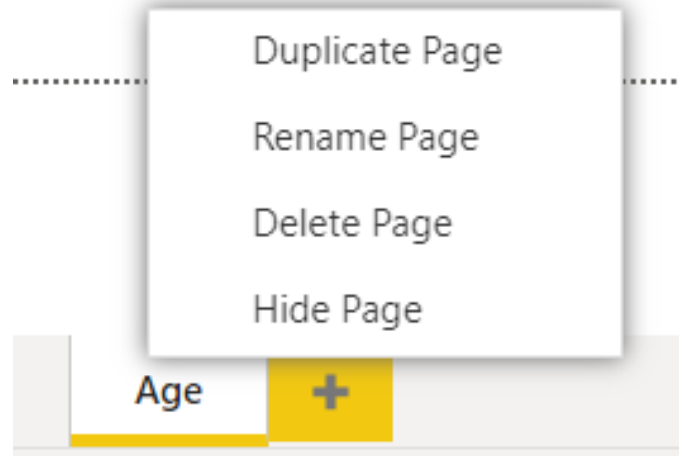
Keep all filters ☒ On



- Interpretation of violin plot is same as interpretation of box plot.
- A violin plot contains a box plot inside it.

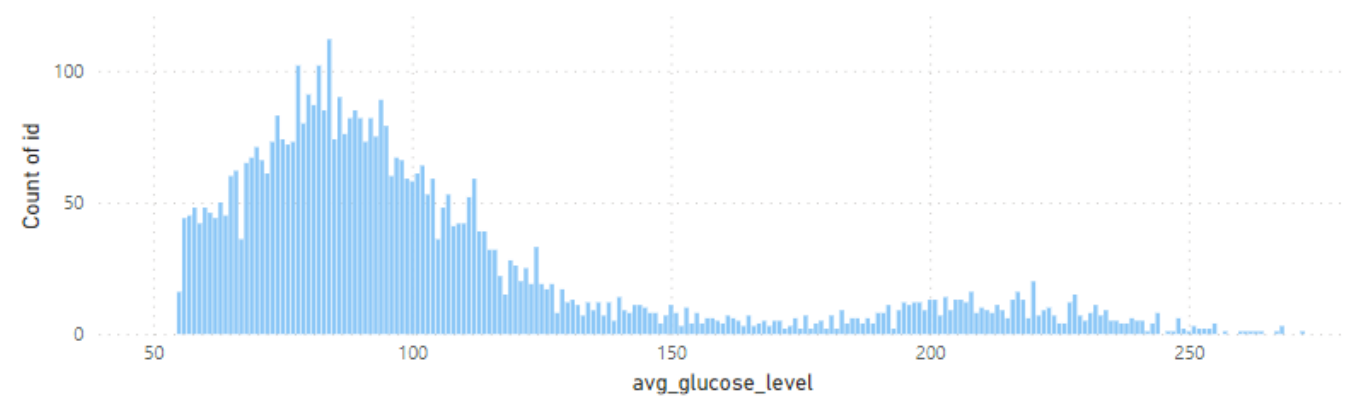
Avg Glucose Level variable

- Repeat the same graphs for this variable.
- Use Duplicate page option to save time.

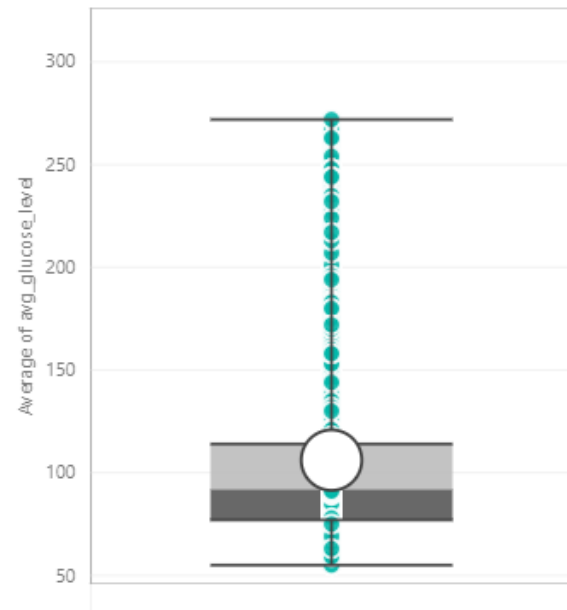


Avg Glucose Level variable

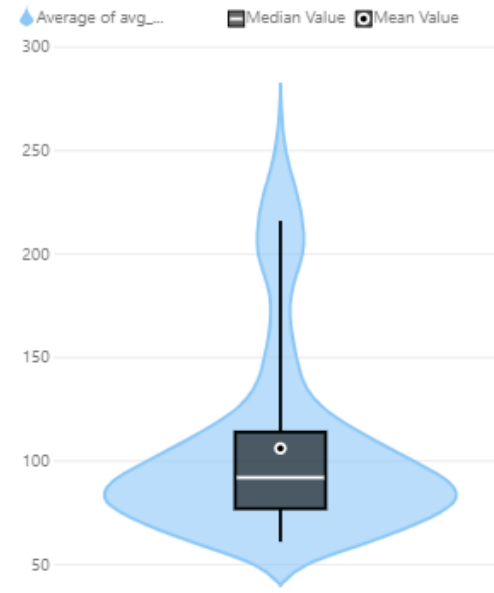
Count of id by avg_glucose_level



Average of avg_glucose_level by id



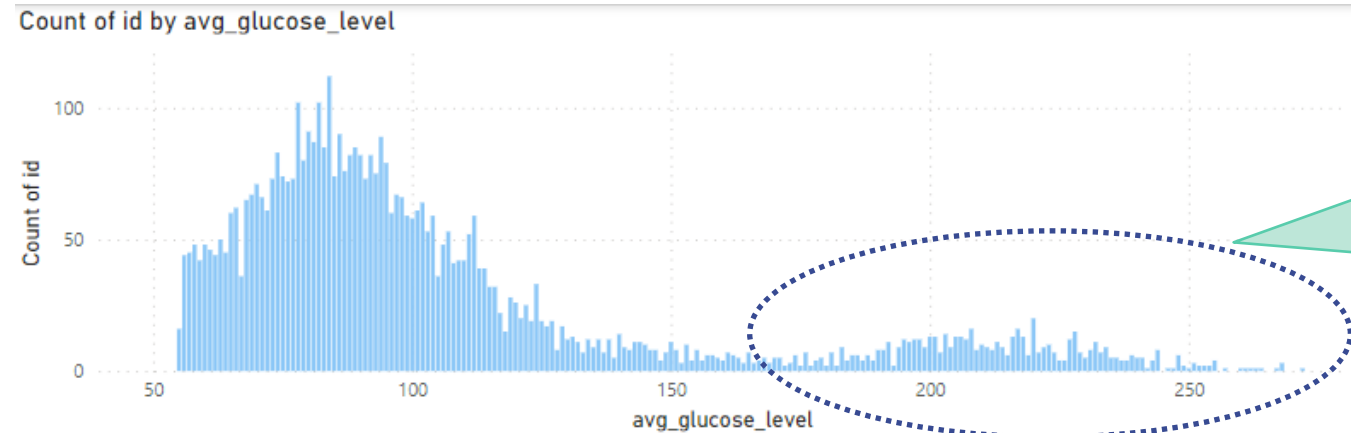
Average of avg_glucose_level by id



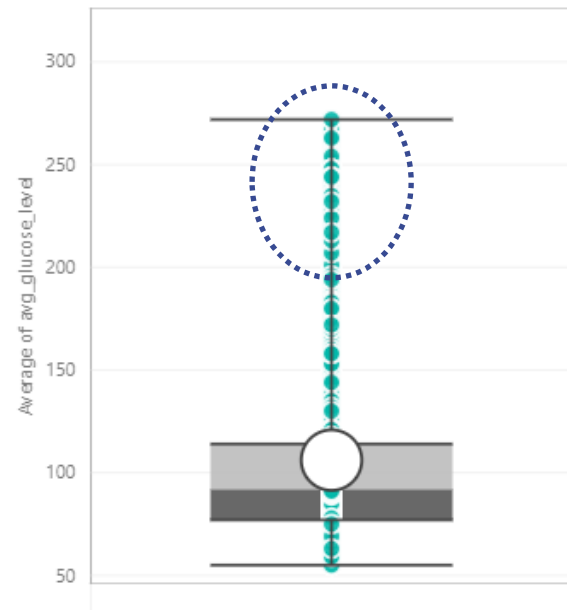
Median Type	Inclusive
Whisker Type	Min/Max
Mean	106.12
Quartile 1	77.00
Median	92.00
Quartile 3	114.00
Maximum	272.00
Minimum	55.00
IQR	37.00
Upper Whisker	272.00
Lower Whisker	55.00

Avg Glucose Level variable

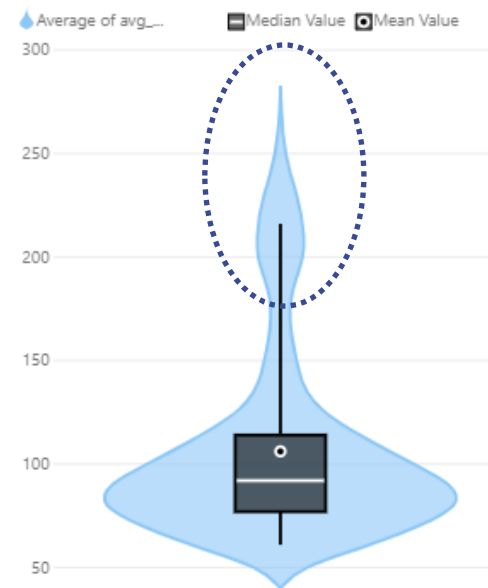
- All three graphs clearly indicating outliers
- We need NOT create all three graphs, Box plot is sufficient to detect the outliers in the data



Average of avg_glucose_level by id

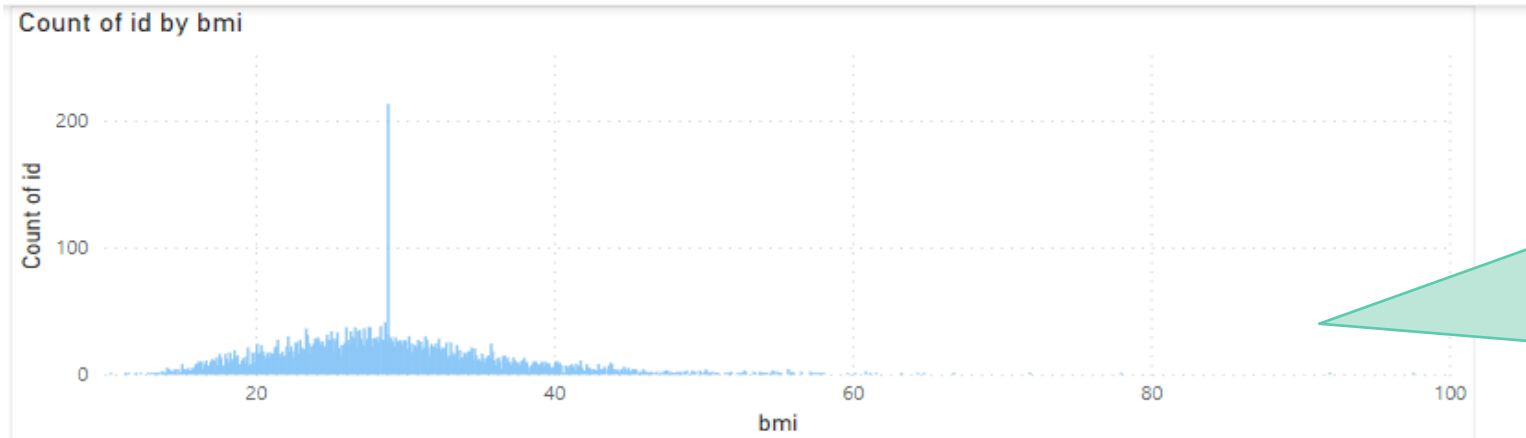


Average of avg_glucose_level by id



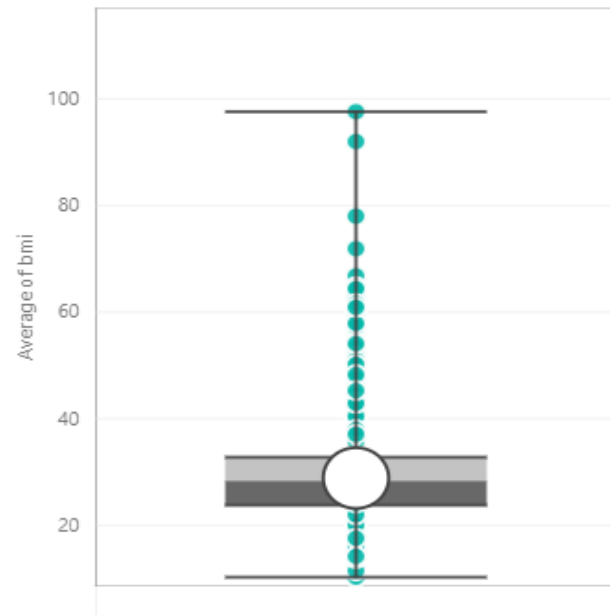
Median Type	Inclusive
Whisker Type	Min/Max
Mean	106.12
Quartile 1	77.00
Median	92.00
Quartile 3	114.00
Maximum	272.00
Minimum	55.00
IQR	37.00
Upper Whisker	272.00
Lower Whisker	55.00

bmi

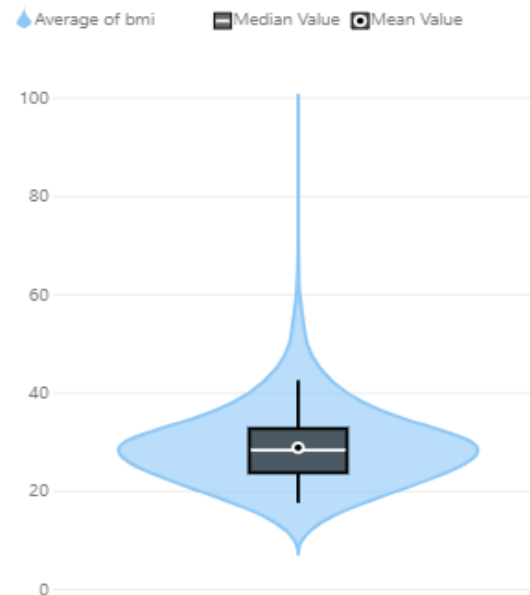


- BMI also has outliers.
- 75% of the people have BMI < 32. In some cases BMI is near to 97

Average of bmi by id



Average of bmi by id



Median Type	Inclusive
Whisker Type	Min/Max
Mean	28.90
Quartile 1	23.80
Median	28.40
Quartile 3	32.80
Maximum	97.60
Minimum	10.30
IQR	9.00
Upper Whisker	97.60
Lower Whisker	10.30

Stroke Case study

Step 6 – Bi-Variate Analysis

Bi-Variate Analysis

- Analyze every variable with respect to target.
- Identify the influential variables on the target.

Categorical Variables Exploration

General_Details_table

- ☐ Σ age
- ☐ ☒ Age_high_risk
- ☐ ever_married
- ☐ gender
- ☐ ☒ High_Risk_factor2
- ☐ id
- ☐ Residence_type
- ☐ work_type_Cleaned1

Risk_Factors_Table

- ☐ Σ avg_glucose_level
- ☐ Σ bmi
- ☐ Σ heart_disease
- ☐ Σ hypertension
- ☐ Patient_id
- ☐ smoking_status
- ☐ Σ stroke

Stroke_rate
4.87%

stroke	Count of id	Percent_of_total
0	4861	95.13%
1	249	4.87%
Total	5110	100.00%

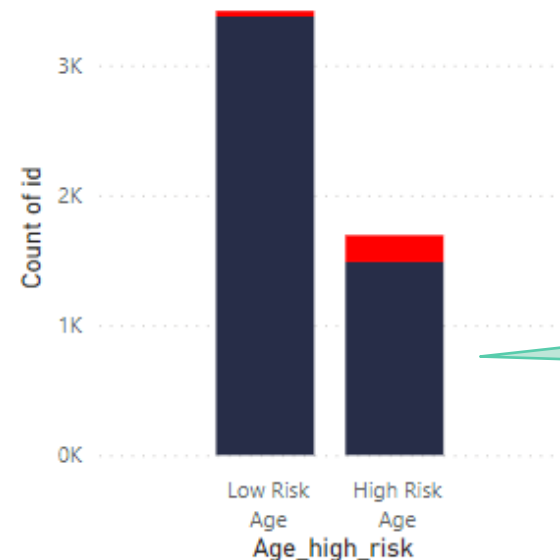
Relation between Stroke and High age risk factor

stroke	High Risk Age	Low Risk Age	Total
0	30.57%	69.43%	100.00%
1	82.73%	17.27%	100.00%
Total	33.11%	66.89%	100.00%

- Matrix >> Stroke on Rows >> Age_high_risk on columns>>Count of id on values. >> Click on Values >> Show value as >> percentage of row.

Count of id by Age_high_risk and stroke

stroke ● 0 ● 1

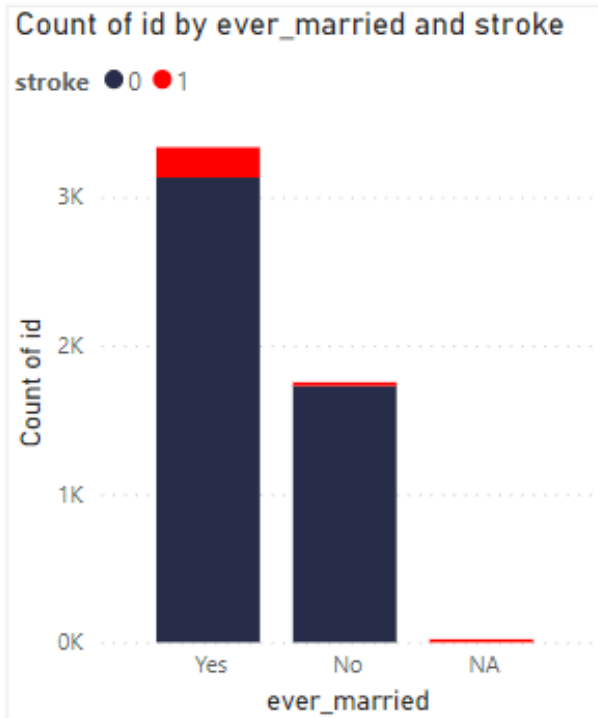


- Out of all the people who experienced stroke, we observed 82.73% are from high age risk factor. Whereas only 33% belong to high age group.

- Stacked column chart >> Age_high_risk on X-axis >> Count of id on Y axis >> Legend >> Stroke.
- Format >> Columns >> Select a category>>More Colours

Stroke vs. Marital Status

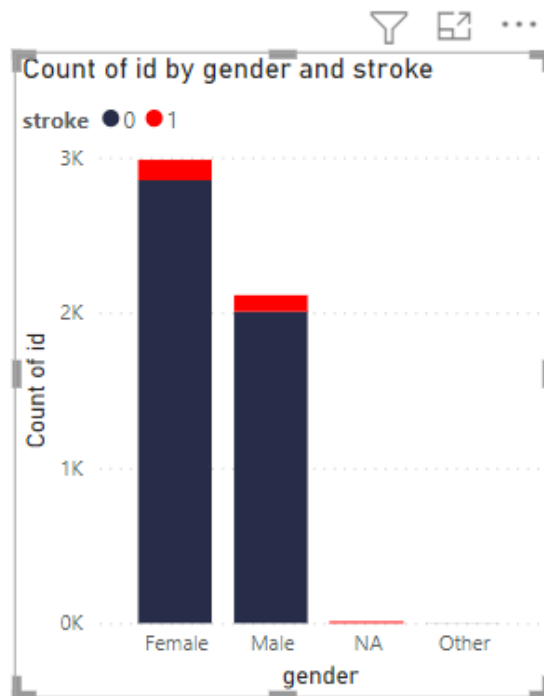
stroke	NA	No	Yes	Total
0		35.55%	64.45%	100.00%
1	8.43%	10.04%	81.53%	100.00%
Total	0.41%	34.31%	65.28%	100.00%



- Unmarried people have chances of stroke.
- This is a deceptive result. We need to be careful, usually unmarried people have less age, hence stroke cases percentage is less

Stroke vs. Gender

stroke	Female	Male	NA	Other	Total
0	58.69%	41.29%		0.02%	100.00%
1	53.01%	42.57%	4.42%		100.00%
Total	58.41%	41.35%	0.22%	0.02%	100.00%



- Male and Female have the same probably of getting stroke.

Categorical vs target

- There is one more way of performing the bi-variate analysis.
- We can show the stroke rate by each category

Stroke rate in each category

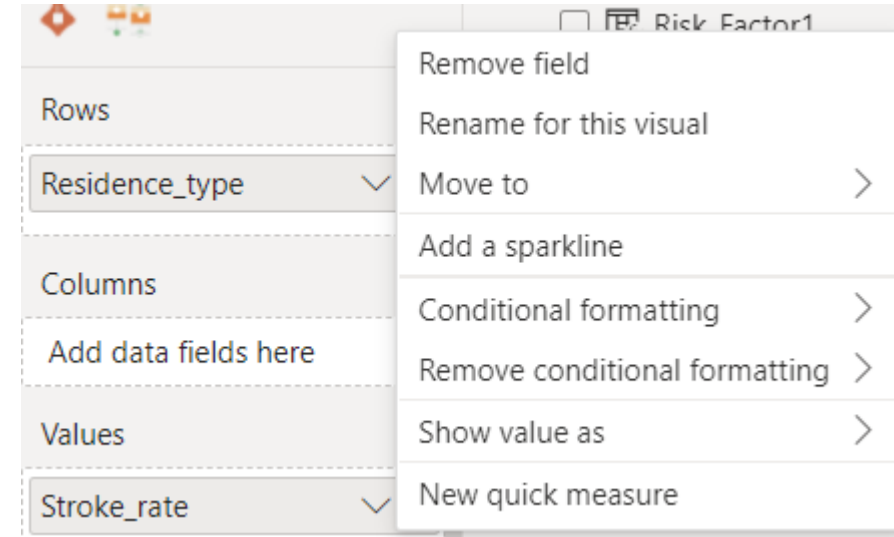
Age_high_risk	Stroke_rate
High Risk Age	12.17%
Low Risk Age	1.26%
Total	4.87%

ever_married	Stroke_rate
NA	100.00%
No	1.43%
Yes	6.09%
Total	4.87%

Residence_type	Stroke_rate
NA	100.00%
Rural	4.42%
Urban	5.09%
Total	4.87%

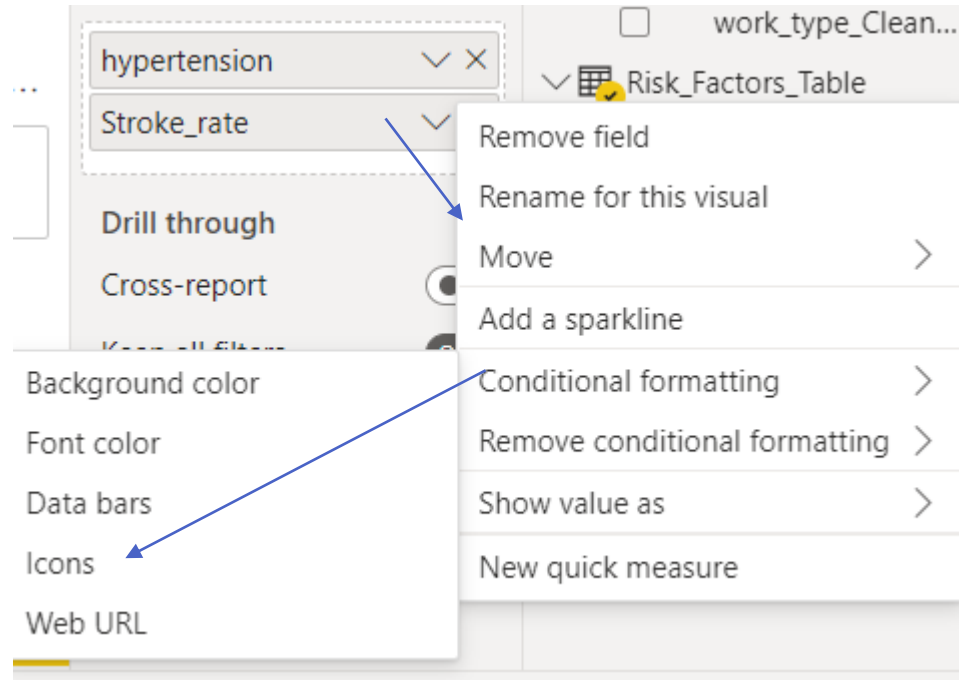
work_type_Cleaned1	Stroke_rate
children	0.29%
Govt_job	4.88%
NA	100.00%
Never_worked	0.00%
Private	4.96%
Self-employed	7.82%
Total	4.87%

smoking_status	Stroke_rate
formerly smoked	7.94%
never smoked	4.76%
smokes	5.34%
Unknown	3.03%
Total	4.87%



- How to perform Conditional Formatting ?
- Click on the stroke rate filed >> Conditional formatting >> Icons

Conditional formatting



Stroke rate in each category

Age_high_risk	Stroke_rate
High Risk Age	12.17%
Low Risk Age	1.26%
Total	4.87%

ever_married	Stroke_rate
NA	100.00%
No	1.43%
Yes	6.09%
Total	4.87%

gender	Stroke_rate
Female	4.42%
Male	5.02%
NA	100.00%
Other	0.00%
Total	4.87%

High_Risk_factor2	Stroke_rate
High Risk2	16.77%
Low Risk2	4.49%
Total	4.87%

Residence_type	Stroke_rate
NA	100.00%
Rural	4.42%
Urban	5.09%
Total	4.87%

work_type_Cleaned1	Stroke_rate
children	0.29%
Govt_job	4.88%
NA	100.00%
Never_worked	0.00%
Private	4.96%
Self-employed	7.82%
Total	4.87%

heart_disease	Stroke_rate
	0.00%
0	4.19%
1	17.09%
Total	4.87%

hypertension	Stroke_rate
	0.00%
0	3.98%
1	13.28%
Total	4.87%

smoking_status	Stroke_rate
formerly smoked	7.94%
never smoked	4.76%
smokes	5.34%
Unknown	3.03%
Total	4.87%

Some categories show really high stroke rate.

Continuous Variables vs. Target

General_Details_table

- ☐ \sum age
- ☐ \mathbb{E}_x Age_high_risk
- ☐ ever_married
- ☐ gender
- ☐ \mathbb{E}_x High_Risk_factor2
- ☐ id
- ☐ Residence_type
- ☐ work_type_Cleaned1

Risk_Factors_Table

- ☐ \sum avg_glucose_level
- ☐ \sum bmi
- ☐ \sum heart_disease
- ☐ \sum hypertension
- ☐ Patient_id
- ☐ smoking_status
- ☐ \sum stroke

Age vs Stroke

- Does Age have an impact on Stroke?

stroke	Average of age
0	41.97
1	67.25
Total	43.20

- We observed average age is for people who experienced stroke.

Continuous Variables vs. Target

stroke	Average of age
0	41.97
1	67.25
Total	43.20

stroke	Average of avg_glucose_level
0	104.59
1	132.54
Total	105.96

stroke	Average of bmi
0	28.83
1	30.22
Total	28.90

- Avg Glucose level seem to have some difference.

- BMI is almost same in stroke and non-stroke cases

Next Step

Step 7 – Multi-Variate Analysis
