PROJECT DOCUMENTATION

DATE: 22/12/2023

PROJECT TITLE: Health Insurance Analysis

OBJECTIVE: The goal of this data analysis project is to gain insights into the factors influencing health insurance charges. We aim to identify patterns, correlations, and trends that can help in understanding the key drivers of health insurance costs.

INTRODUCTION:

In an era where healthcare costs continue to be a significant concern, understanding the underlying factors that contribute to these expenses is crucial for informed decision-making and resource allocation. This data analysis project aims to shed light on the intricacies of healthcare charges by leveraging the power of Power BI and delving into a comprehensive dataset. The dataset encompasses diverse demographic and lifestyle factors, including age, sex, BMI, number of children, smoking status, region, and healthcare charges.

DATASET DESCRIPTION:

Age: age of the primary beneficiary

Sex: insurance contractor gender, female, male

BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

Children: Number of children covered by health insurance / Number of dependents

Smoker: Smoking

Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

Charges: Individual medical costs billed by health insurance

TOOLS

1. **Microsoft Excel:** Data Exploration, Data Cleaning and Preparation, Data Visualization, and Basic Analysis.
2. **Microsoft Power Query:** Data Transformation, and Data Integration.
3. **Microsoft Word:** Documentation and Reporting.
4. **Microsoft Power BI:** Data Visualization and dashboard.
5. **Microsoft PowerPoint:** Presentation of Findings.
6. **Microsoft Teams:** Collaboration and Sharing Reports.

METHODOLOGY

- Data Exploration
- Data Cleaning and Preparation
- Data Visualization and Dashboard
- Insights and Recommendations
- Conclusion

DATA EXPLORATION:

- Dataset was loaded into PowerBI.
- Examine the number of rows and columns in the dataset to understand its size.
  Number of rows: 1339
  Number of columns: 7
- Review the data types of each column (e.g., numeric, categorical) to ensure they are assigned correctly.
- Calculate basic summary statistics (mean, median, standard deviation, min, max) for numerical columns to understand the central tendency and variability of the data.

Summary statistics of age column was calculated using DAX Functions:

1. Mean: 39.19

```
Mean(age) = AVERAGE(insurance[age])
```

2. Median: 39

```
Median(age) = MEDIAN(insurance[age]
```

3. Mode: 18

```
Mode (Age) = MINX(TOPN(1, ADDCOLUMNS(VALUES(insurance[age]), "percent",
CALCULATE(COUNT(insurance[age]))),[percent],DESC), insurance[age])
```

4. Range: 46

```
RANGE(AGE) = MAX(insurance[age]) - MIN(insurance[age])
```

5. Q1: 26

```
Q1(AGE) = PERCENTILE.EXC(insurance[age], 0.25)
```

6. Q2: 39

```
Q2(AGE) = PERCENTILE.EXC(insurance[age], 0.5)
```

7. Q3: 51

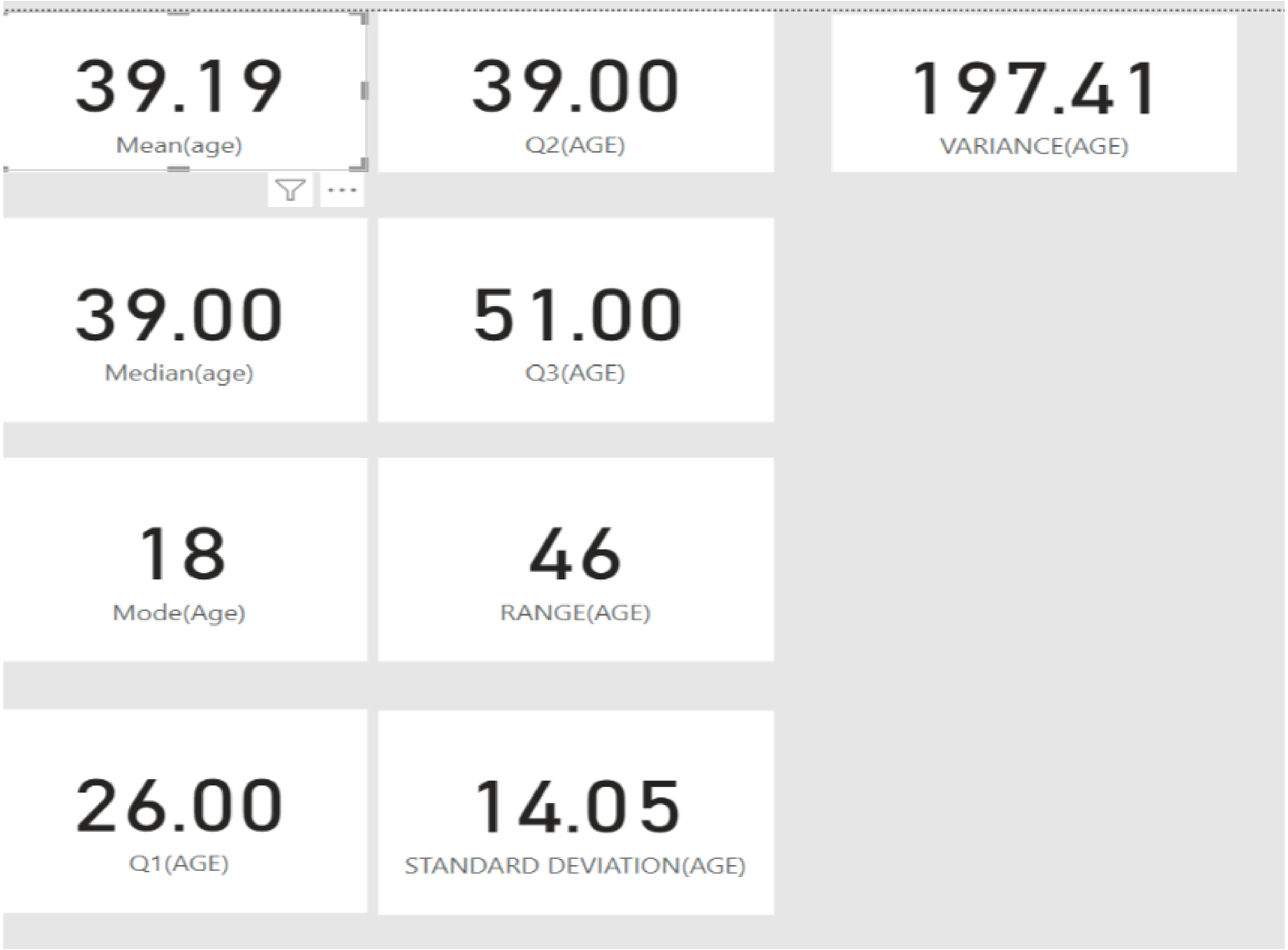
```
Q3(AGE) = PERCENTILE.EXC(insurance[age], 0.75)
```

8. Standard deviation: 14.05

```
STANDARD DEVIATION(AGE) = STDEV.P(insurance[age])
```

9. Variance: 197.41

```
VARIANCE(AGE) = VAR.P(insurance[age])
```



39.19
Mean(age)

39.00
Q2(AGE)

197.41
VARIANCE(AGE)

39.00
Median(age)

51.00
Q3(AGE)

18
Mode(Age)

46
RANGE(AGE)

26.00
Q1(AGE)

14.05
STANDARD DEVIATION(AGE)

Summary statistics of BMI column was calculated using functions:

1. Mean: 30.66

```
Mean(BMI) = AVERAGE(insurance[bmi])
```

2. Median: 30.40

```
Median(BMI) = MEDIAN(insurance[bmi])
```

3. Mode: 32.30

```
Mode(BMI) = MINX(TOPN(1, ADDCOLUMNS(VALUES(insurance[bmi]), "percent",
CALCULATE(COUNT(insurance[bmi]))),[percent],DESC), insurance[bmi])
```

4. Range: 37.17
5. Q1: 26.29

```
Mode(BMI) = MINX(TOPN(1, ADDCOLUMNS(VALUES(insurance[bmi]), "percent",
CALCULATE(COUNT(insurance[bmi]))),[percent],DESC), insurance[bmi])
```

6. Q2: 30.40

```
Q2(BMI) = PERCENTILE.EXC(insurance[bmi], 0.5)
```

7. Q3: 34.70

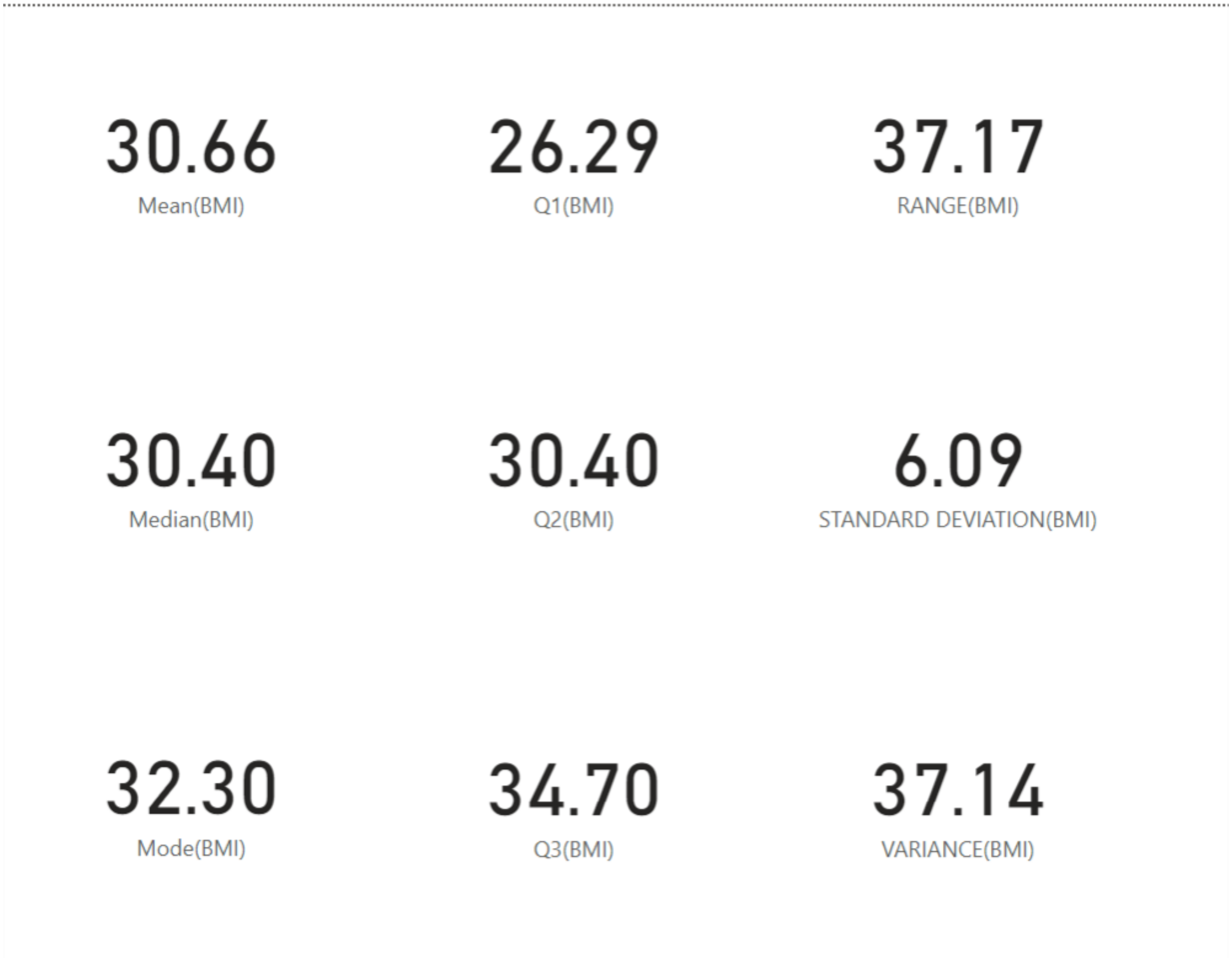
```
Q3(BMI) = PERCENTILE.EXC(insurance[bmi], 0.75)
```

8. Standard deviation: 6.09

```
STANDARD DEVIATION(BMI) = STDEV.P(insurance[bmi])
```

9. Variance: 37.14

```
VARIANCE(BMI) = VAR.P(insurance[bmi])
```



30.66
Mean(BMI)

26.29
Q1(BMI)

37.17
RANGE(BMI)

30.40
Median(BMI)

30.40
Q2(BMI)

6.09
STANDARD DEVIATION(BMI)

32.30
Mode(BMI)

34.70
Q3(BMI)

37.14
VARIANCE(BMI)

Summary statistics of charges column:

1. Mean: $13270

```
MEAN(CHARGES) = AVERAGE(insurance[charges])
```

2. Median: $9390

```
MEDIAN(CHARGES) = MEDIAN(insurance[charges])
```

3. Mode: $1640

```
Mode(CHARGES) = MINX(TOPN(1, ADDCOLUMNS(VALUES(insurance[charges]), "percent",
CALCULATE(COUNT(insurance[charges]))),[percent],DESC), insurance[charges])
```

4. Range: $62650

```
RANGE(CHARGES) = MAX(insurance[charges]) - MIN(insurance[charges])
```

5. Q1: $4740

```
Q1(CHARGES) = PERCENTILE.EXC(insurance[charges], 0.25)
```

6. Q2: $9390

```
Q2(CHARGES) = PERCENTILE.EXC(insurance[charges], 0.5)
```

7. Q3: $16780

```
Q3(CHARGES) = PERCENTILE.EXC(insurance[charges], 0.75)
```

8. Standard deviation: $12100

```
STANDARD DEVIATION(CHARGES) = STDEV.P(insurance[charges])
```

9. Variance: $146.44M

```
VARIANCE(CHARGES) = VAR.P(insurance[charges])
```

Summary statistics of children column was calculated using measures:

1. Mean: 1.09

```
Mean(CHILDREN) = AVERAGE(insurance[children])
```

2. Median: 1

```
Median(CHILDREN) = MEDIAN(insurance[children])
```

3. Mode: 0

```
Mode(children) = MINX(TOPN(1, ADDCOLUMNS(VALUES(insurance[children]), "percent",
CALCULATE(COUNT(insurance[children]))),[percent],DESC), insurance[children])
```

4. Range: 5

```
RANGE(CHILDREN) = MAX(insurance[children]) - MIN((insurance[children]))
```

5. Q1: 0

```
Q1(CHILDREN) = PERCENTILE.EXC(insurance[children], 0.25)
```

6. Q2: 1

```
Q2(CHILDREN) = PERCENTILE.EXC(insurance[children], 0.5)
```

7. Q3: 2

```
Q3(CHILDREN) = PERCENTILE.EXC(insurance[children], 0.75)
```

8. Standard deviation: 1.2

```
STANDARD DEVIATION(CHILDREN) = STDEV.P(insurance[children])
```

9. Variance: 1.45

```
VARIANCE(CHILDREN) = VAR.P(insurance[children])
```

- Count the unique values in categorical columns to understand the diversity of the data and identify potential issues.

Count of unique values in sex column:
Male: 676
Female: 663

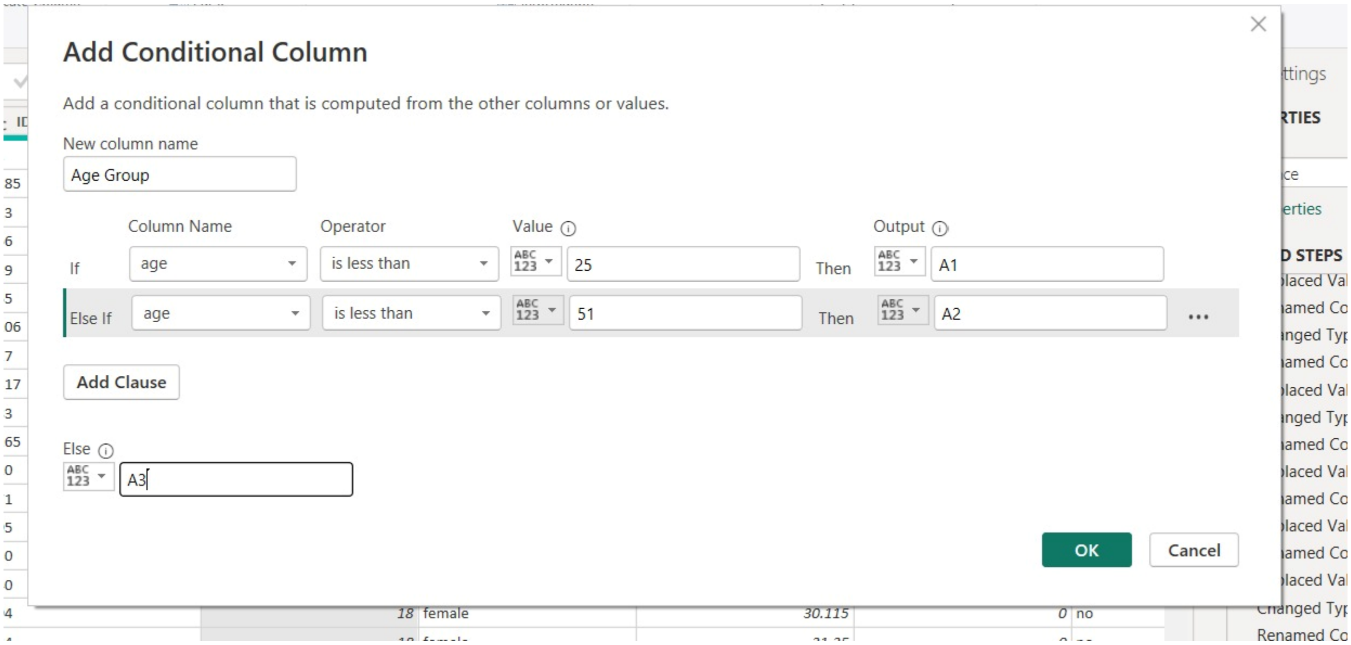
Count of unique values in smoker column:
Smokers: 274
Non-smokers: 1065

Count of unique values in region column:

Southwest: 326
Southeast: 364
Northwest: 325
Northeast: 324

➔**histograms or box plots** to visualize the distribution and outliers.

We added two extra columns to help us with visualizing: 1) ID Column (Indexing Column)
2) Age Group Column (Conditional Column)



Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name: Age Group

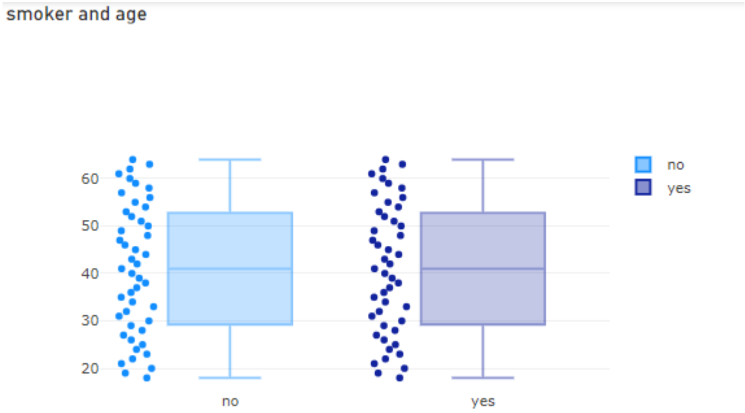
	Column Name	Operator	Value		Output
If	age	is less than	25	Then	A1
Else If	age	is less than	51	Then	A2

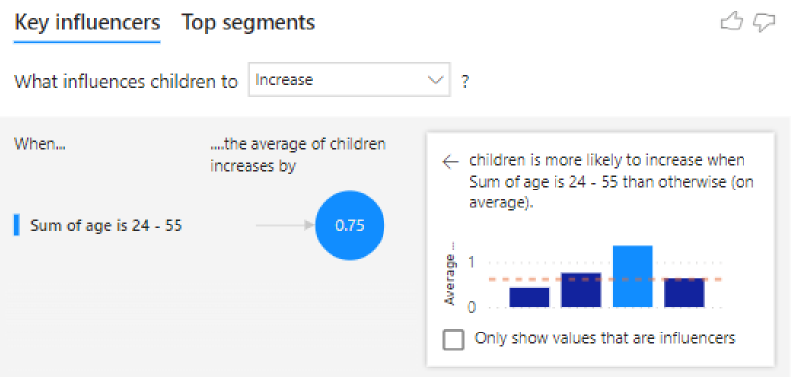
Add Clause

Else: A3

OK Cancel

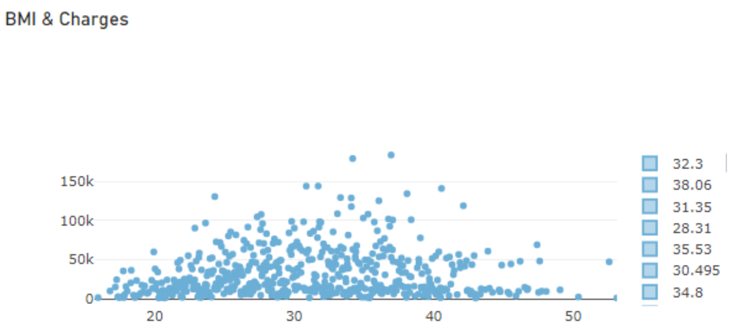
- <u>Histograms and box plots of Age</u>

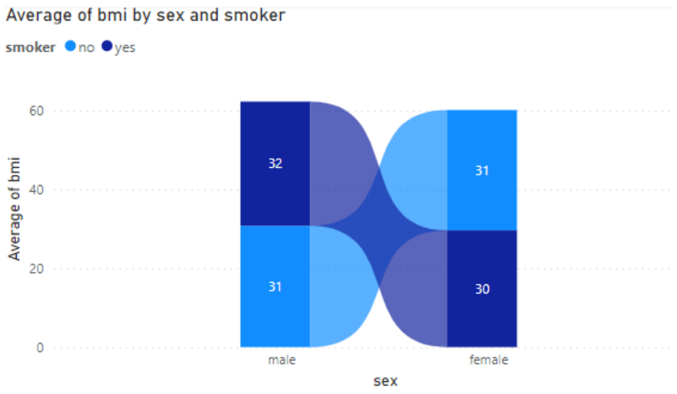



Key influencers
Top segments
What influences children to
Increase
?
When...
....the average of children increases by
Sum of age is 24 - 55
0.75
Average ...
1
0
Only show values that are influencers

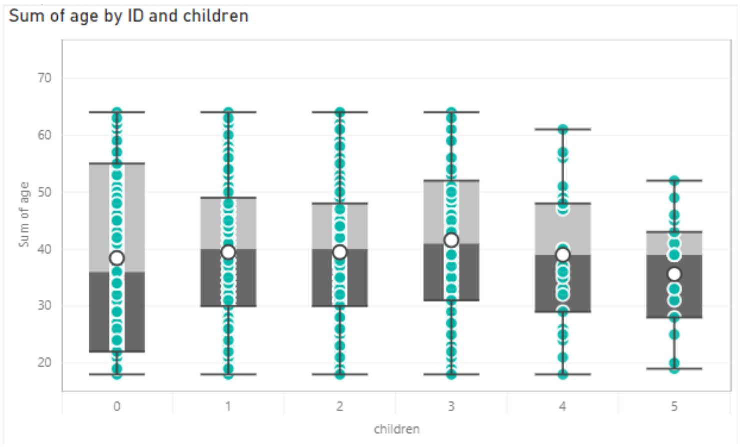
children is more likely to increase when Sum of age is 24 - 55 than otherwise (on average).

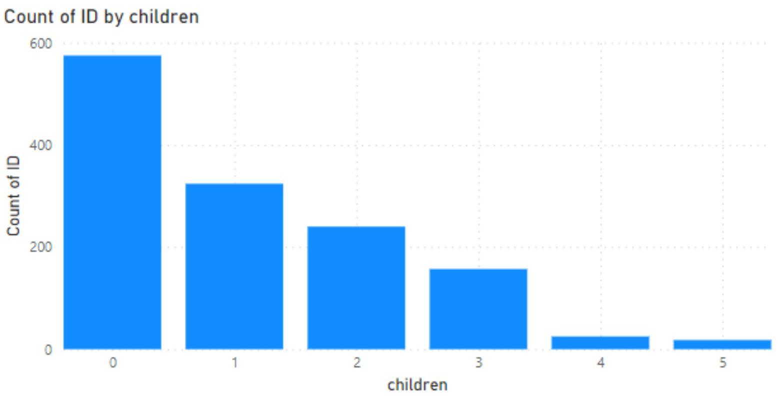
- Histograms and box plots of BMI



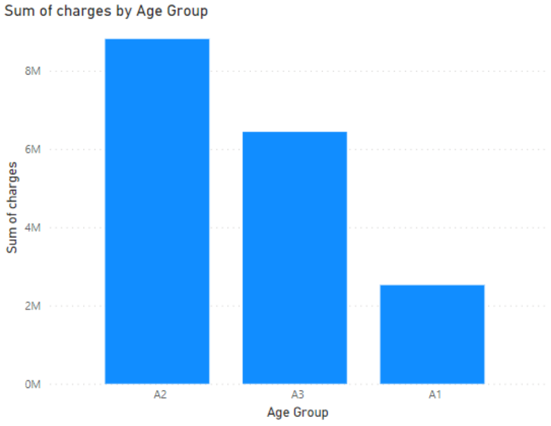


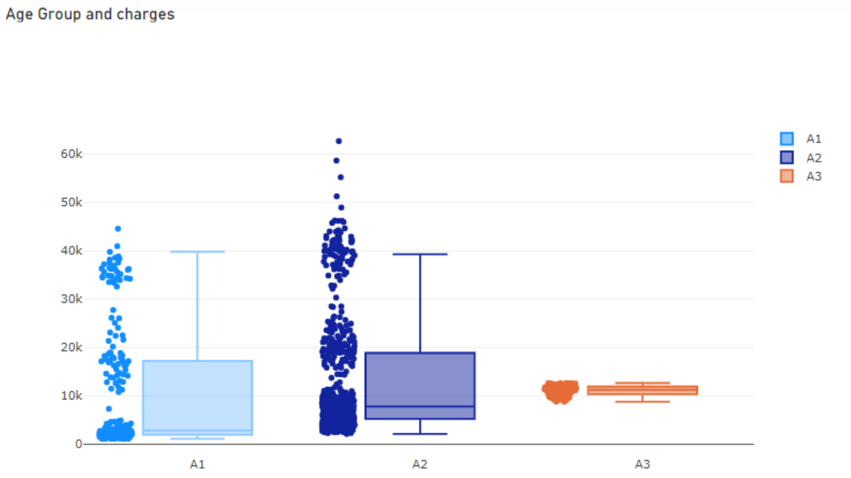
- Histograms and box plots of children





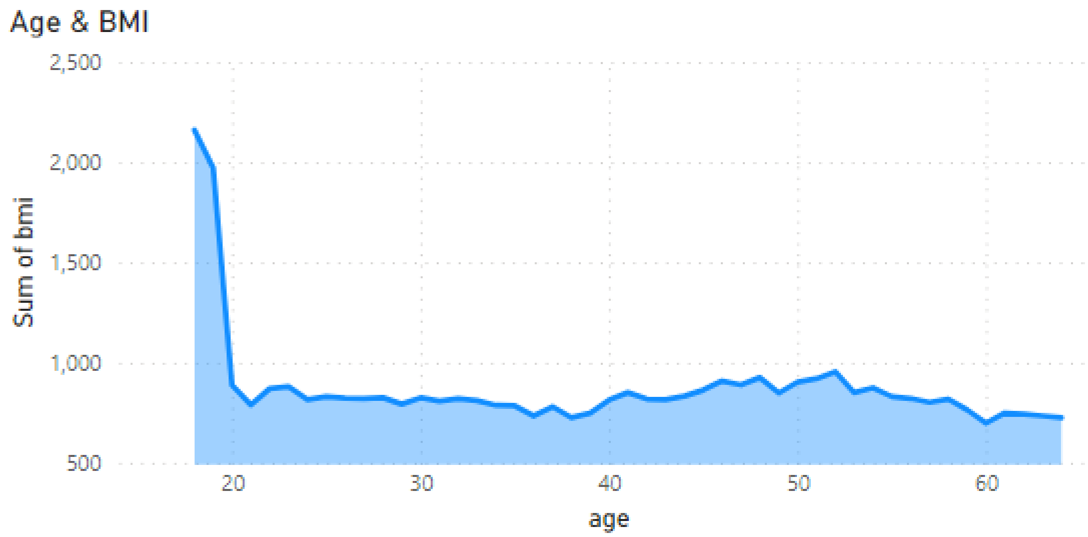
- Histograms and box plots of charges



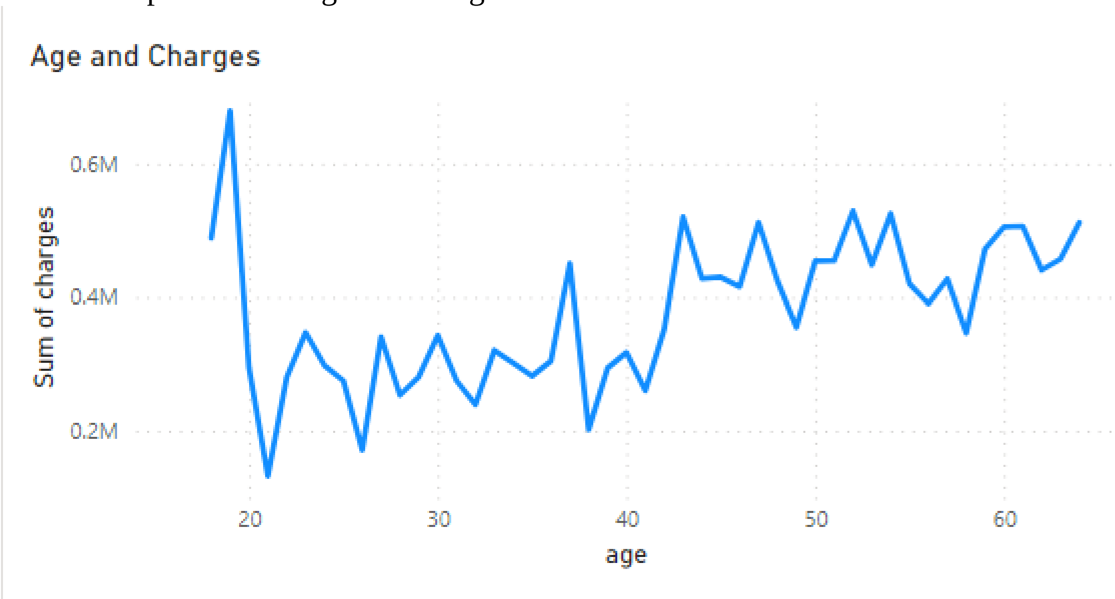


➔Use **scatter plots, correlation matrices, or pair plots** to investigate relationships between numerical variables.

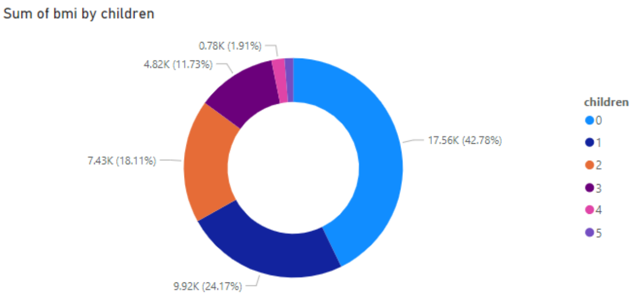
Relationships between Age and BMI:



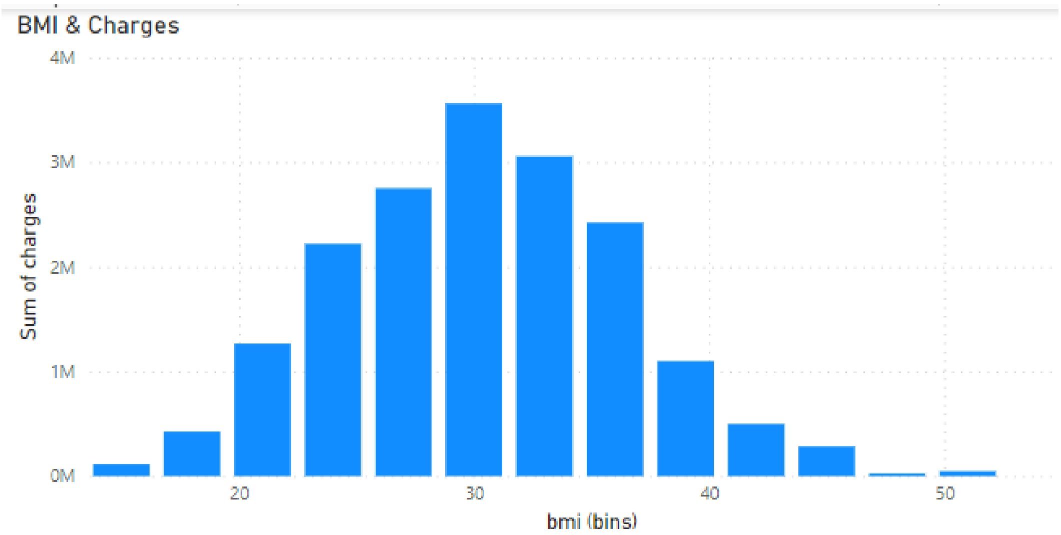
Relationships between Age and charges:



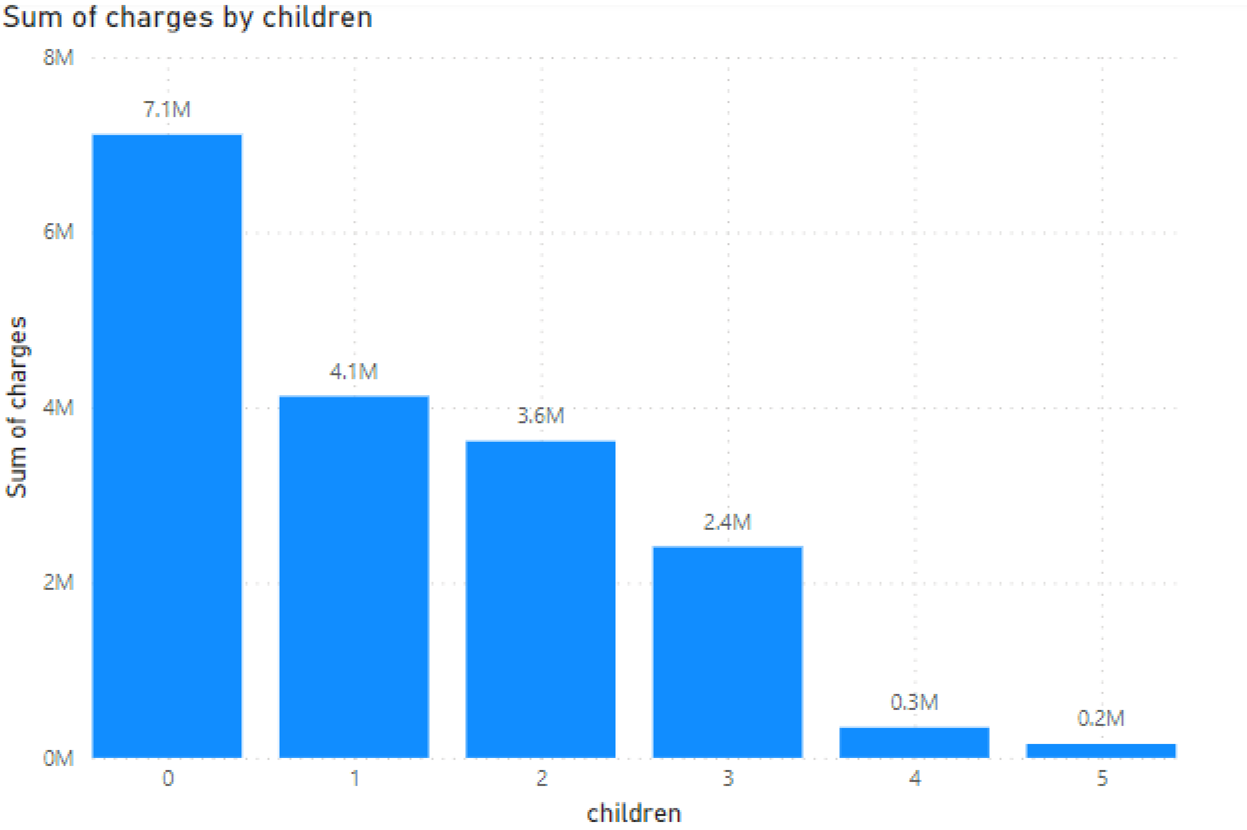
Relationships between BMI and children:



Relationships between BMI and charges:



Relationships between children and charges:



DATA CLEANING:

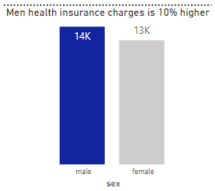
- No missing or null values

DATA ANALYSIS:

1. A new table was created for the measures.

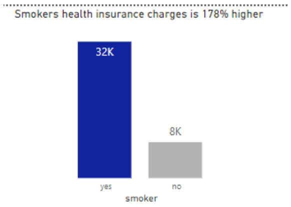
2. Write a DAX measure for average charges, in the insurance table.

```
avg_charges =
CALCULATE(
   AVERAGE('insurance'[charges])
)
```

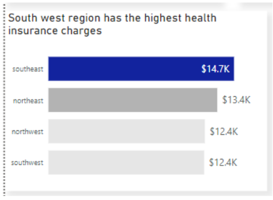


3. Write a DAX measure to calculate the percentage difference, in the insurance table, for the smoker column and visualize.

```
Smoker_%Difference =
VAR AvgChargesSmokers = CALCULATE(AVERAGE('insurance'[charges]), 'insurance'[smoker] = "yes")
VAR AvgChargesNonSmokers = CALCULATE(AVERAGE('insurance'[charges]), 'insurance'[smoker] = "no")
RETURN
    DIVIDE(AvgChargesSmokers - AvgChargesNonSmokers, [Avg_charges], 0)
```



4. Visualize average charges per region.



5. Created a new column Sexes = IF('insurance'[sex] = "female", "Female","Male")
6. Created a new column Smokers = IF('insurance'[smoker] = "no", "Non-smokers","Smokers")

7. Write the DAX formula for an age bracket column. The age bracket column will group the age column into 18-25, 26-35, and 36-45, 46-55, 56-65 in the insurance table, for the age column and visualize.

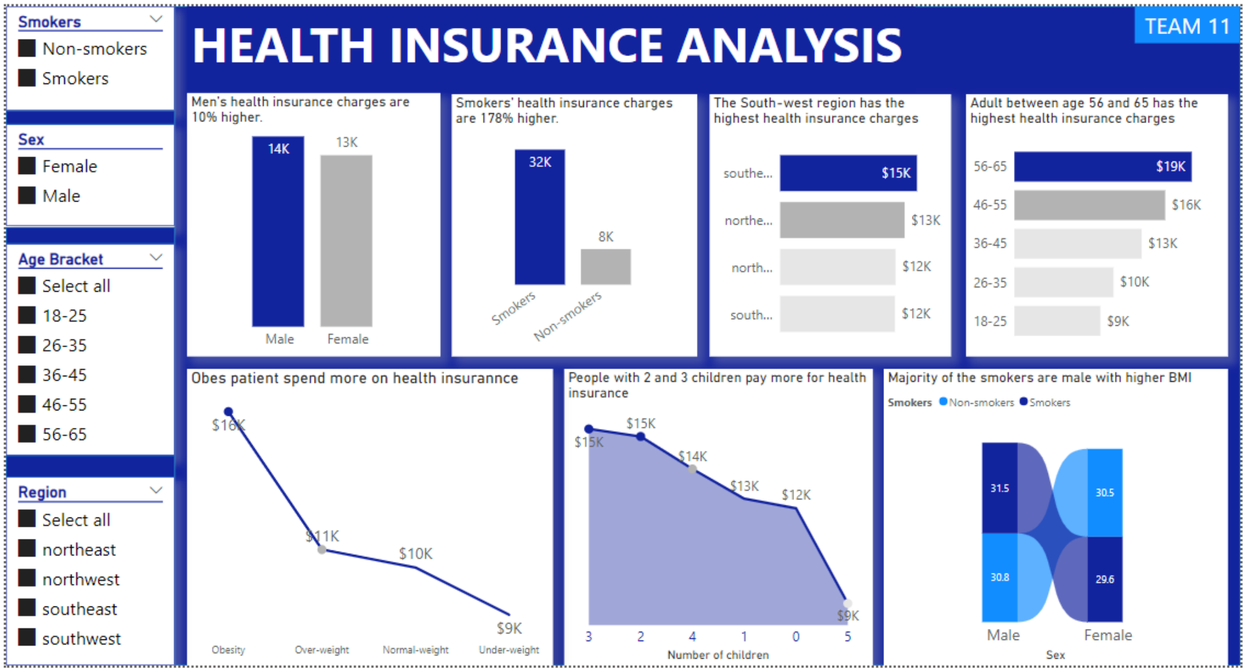
```
AgeBracket =

SWITCH(
    TRUE(),
    'insurance'[age] >= 18 && 'insurance'[age] <= 25, "18-25",
    'insurance'[age] >= 26 && 'insurance'[age] <= 35, "26-35",
    'insurance'[age] >= 36 && 'insurance'[age] <= 45, "36-45",
    'insurance'[age] >= 46 && 'insurance'[age] <= 55, "46-55",
    'insurance'[age] >= 56 && 'insurance'[age] <= 65, "56-65",
    "Other"
)
```

8. BMI classification according to the National Institute of Health:

- Severely underweight - BMI less than 16.5kg/m^2
- Underweight - BMI under 18.5 kg/m^2
- Normal weight - BMI greater than or equal to 18.5 to 24.9 kg/m^2
- Overweight – BMI greater than or equal to 25 to 29.9 kg/m^2
- Obesity – BMI greater than or equal to 30 kg/m^2

```
BMI Class =
SWITCH(
    TRUE(),
    'insurance'[bmi] <= 18.4, "Under-weight",
    'insurance'[bmi] >= 18.5 && 'insurance'[bmi] <= 24.99, "Normal-weight",
    'insurance'[bmi] >= 25 && 'insurance'[bmi] <= 29.99, "Over-weight",
    'insurance'[bmi] >= 30 && 'insurance'[bmi] <= 54, "Obesity",
    "Other"
)
```

DATA VISUALIZATION



HEALTH INSURANCE ANALYSIS

Men's health insurance charges are 10% higher.

Smokers' health insurance charges are 178% higher.

The South-west region has the highest health insurance charges

Adult between age 56 and 65 has the highest health insurance charges

Obes patient spend more on health insurannce

People with 2 and 3 children pay more for health insurance

Majority of the smokers are male with higher BMI

INSIGHTS

- Men's health insurance charges are 10% higher.
- Smokers' health insurance charges are 178% higher.
- The South-west region has the highest health insurance charges.
- Obese patient spends more on health insurance.
- People with 2 and 3 children pay more for health insurance.
- Majority of the smokers are male with higher BMI.
- Most of the health insurance customers have no children.

RECOMMENDATIONS

- **Targeted Marketing:**
  - Consider targeted marketing strategies for men, as they tend to have higher health insurance charges. This could involve tailoring advertising or promotional campaigns to appeal specifically to male demographics.
- **Smoking Cessation Programs:**
  - Develop and promote smoking cessation programs or health initiatives to reduce smoking rates among the insured population. This could potentially lead to lower health insurance charges for both individuals and the insurance provider.
- **Regional Pricing Strategies:**

    - Investigate the reasons behind the higher health insurance charges in the South-west region. Explore whether regional health factors, healthcare infrastructure, or other variables contribute to the increased costs. Adjust pricing strategies accordingly.
- **Wellness Programs for Obesity:**
    - Implement wellness programs targeting obesity. Providing resources and incentives for policyholders to adopt healthier lifestyles may lead to lower health insurance charges over time.
- **Family Planning Education:**
    - Develop educational programs or materials on family planning to inform individuals with 2 or 3 children about potential health insurance cost implications. This could include promoting family planning services or policies that address the needs of families with multiple children.
- **Smoking and BMI Interventions:**
    - Target interventions towards male smokers with higher BMI. Implementing programs that address both smoking cessation and weight management may have a positive impact on health outcomes and insurance charges.
- **Customized Insurance Plans:**
    - Consider offering customized insurance plans that cater to the specific needs and characteristics of different customer segments. This could involve creating plans tailored for non-smokers, individuals with no children, or those with specific health concerns.
- **Customer Education:**
    - Provide educational materials or workshops to inform customers about the factors influencing health insurance charges. This transparency can build trust and help customers make informed decisions about their insurance coverage.
- **Data Collection and Analysis:**
    - Continuously gather and analyze data to monitor trends and identify new insights. Regularly updating the analysis will enable the insurance provider to adapt strategies based on evolving customer behaviors and market conditions.
- **Customer Surveys:**
    - Conduct customer surveys to gather feedback on insurance preferences, challenges, and expectations. This information can be valuable in refining existing insurance plans and developing new offerings that better meet customer needs.