PROJECT DOCUMENTATION

DATE: 20/12/2023

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:

1. Age: age of the primary beneficiary
2. Sex: insurance contractor gender, female, male
3. 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
4. Children: Number of children covered by health insurance / Number of dependents
5. Smoker: Smoking
6. Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
7. 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 Teams: Collaboration, and Sharing Reports.

METHODOLOGY

- Data Cleaning
- Data Exploration
- Data Preprocessing
- Data Visualization and Dashboard
- Insights and Recommendations

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.

DATA CLEANING

- No missing, and null values.
- No empty cells
- No outliers

DATA PREPROCESSING WITH POWER QUERY

- Rename all the columns to sentence case.
- Converted the sex column to upper case.
- Converted the children column to text data type.
- Round up the Charges column.
- Replace the yes with “smokers”, and no with “non-smoker”
- Create a new column for the age bracket
- Create a new column for the BMI
- Replace the values in the region column. The compound words now have a separator and each word is capitalized.
- Created a new column for the age group. The age bracket column was grouped into 56-65, 46-55, 36-45, 26-35, 18-25, and in the insurance table, for the age column.

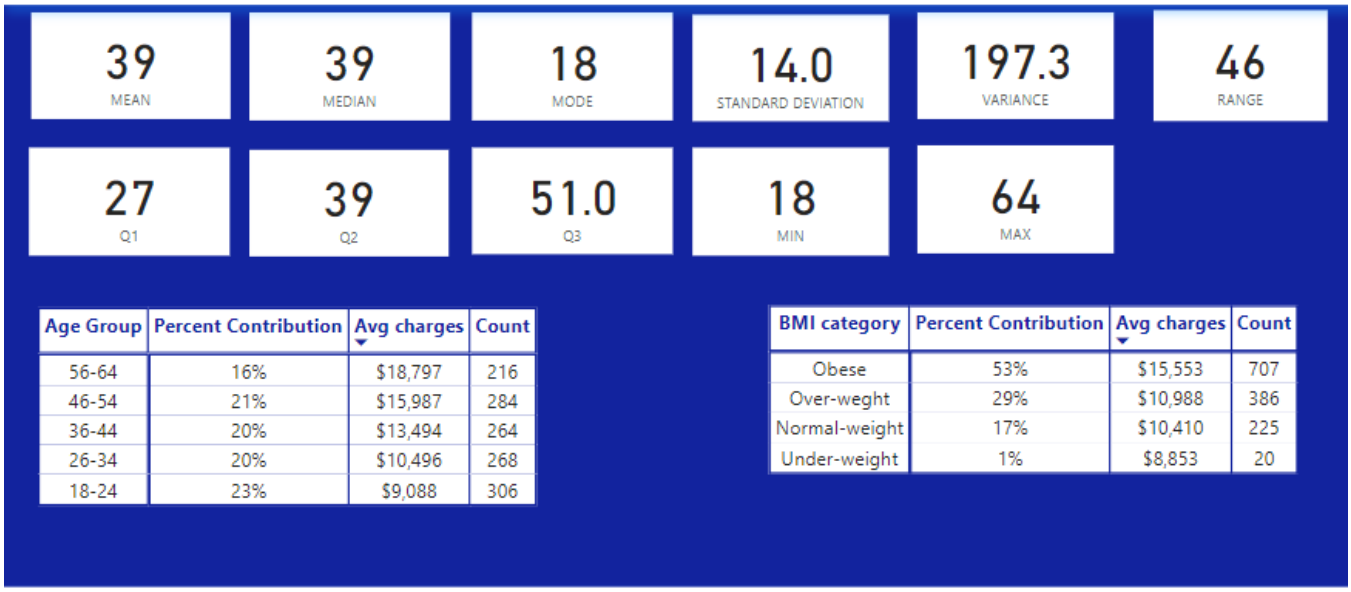
= Table.AddColumn("Age Group", each if [Age] >= 56 then "56-65" else if [Age] >= 46 then "46-55" else if [Age] >= 36 then "36-45" else if [Age] >= 26 then "26-35" else "18-25")

- Round the BMI column to 2 decimal places.
- Created a new column for the BMI group.
- Obesity – BMI greater than or equal to 30 kg/m^2
- Overweight – BMI between 25 to 29.9 kg/m^2
- Normal weight – BMI between 18.5 to 24.9 kg/m^2
- Underweight - BMI under 18.5 kg/m^2
- Severely underweight - BMI less than 16.5kg/m^2

= Table.AddColumn("BMI category", each if [BMI] >= 30 then "Obesity" else if [BMI] >= 25 then "Over-weght" else if [BMI] >= 18.5 then "Normal-weight" else "Under-weight")

Note: While creating an age group using a conditional column in Power Query, start with the highest value and end it with the least value, or else, you will not get the desired result. For instance, I had to start with obesity >= 30 as my first condition.

SUMMARY STATISTICS OF THE AGE VARIABLE



Age Group	Percent Contribution	Avg charges	Count
56-64	16%	$18,797	216
46-54	21%	$15,987	284
36-44	20%	$13,494	264
26-34	20%	$10,496	268
18-24	23%	$9,088	306

BMI category	Percent Contribution	Avg charges	Count
Obese	53%	$15,553	707
Over-weght	29%	$10,988	386
Normal-weight	17%	$10,410	225
Under-weight	1%	$8,853	20

SUMMARY STATISTICS OF THE BMI VARIABLE

SUMMARY STATISTICS OF AGE COLUMN USING DAX:

1. Mean(age) = AVERAGE(insurance[age])
2. Median(age) = MEDIAN(insurance[age]
3. Mode (Age) = MINX(TOPN(1, ADDCOLUMNS(VALUES(insurance[age]), "percent", CALCULATE(COUNT(insurance[age]))),[percent],DESC), insurance[age])
4. RANGE(AGE) = MAX(insurance[age]) - MIN(insurance[age])
5. Q1(AGE) = PERCENTILE.EXC(insurance[age], 0.25)
6. Q2(AGE) = PERCENTILE.EXC(insurance[age], 0.5)
7. Q3(AGE) = PERCENTILE.EXC(insurance[age], 0.75)
8. STANDARD DEVIATION(AGE) = STDEV.P(insurance[age])
9. VARIANCE(AGE) = VAR.P(insurance[age])
10. MAX(AGE) = MAX(insurance[Age])

11. MIN(AGE) = MIN(insurance[Age])

SUMMARY STATISTICS OF BMI COLUMN USING DAX:

1. Mean(BMI) = AVERAGE(insurance[bmi])
2. Median(BMI) = MEDIAN(insurance[bmi])
3. Mode(BMI) = MINX(TOPN(1, ADDCOLUMNS(VALUES(insurance[bmi]), "percent", CALCULATE(COUNT(insurance[bmi]))),[percent],DESC), insurance[bmi])
4. RANGE(AGE) = MAX(insurance[bmi]) - MIN(insurance[bmi])
5. Q1(AGE) = PERCENTILE.EXC(insurance[bmi], 0.25)
6. Q2(BMI) = PERCENTILE.EXC(insurance[bmi], 0.5)
7. Q3(BMI) = PERCENTILE.EXC(insurance[bmi], 0.75)
8. STANDARD DEVIATION(BMI) = STDEV.P(insurance[bmi])
9. VARIANCE(BMI) = VAR.P(insurance[bmi])
10. MIN(BMI) = MIN(insurance[BMI])
11. MAX(BMI) = MAX(insurance[BMI])

SUMMARY STATISTICS OF THE NUMBER OF CHILDREN VARIABLE

1	1	0	1.5	1.2	5
MEAN	MEDIAN	MODE	VARIANCE	STANDARD DEVIATION	RANGE

0.00	1	2	0	5
Q1	Q2	Q3	MIN	MAX

Children	Percent Contribution	Avg charges	Count of Children
0	43%	$12,366	574
1	24%	$12,732	324
2	18%	$15,074	240
3	12%	$15,356	157
4	2%	$13,851	25
5	1%	$8,787	18

SUMMARY STATISTICS OF THE CHARGES VARIABLE

13K	9383	1138	12K	147M	63K
MEAN	MEDIAN	Mode	STANDARD DEVIATION	VARIANCE	RANGE

4734	9383	17K	$1.12K	$63.8K
Q1	Q2	Q3	MIN	MAX

SUMMARY STATISTICS OF CHARGES COLUMN USING DAX:

1. MEAN(CHARGES) = AVERAGE(insurance[charges])
2. MEDIAN(CHARGES) = MEDIAN(insurance[charges])
3. Mode(CHARGES) = MINX(TOPN(1, DDCOLUMNS(VALUES(insurance[charges]), "percent", CALCULATE(COUNT(insurance[charges]))),[percent],DESC), insurance[charges])
4. RANGE(CHARGES) = MAX(insurance[charges]) - MIN(insurance[charges])
5. Q1(CHARGES) = PERCENTILE.EXC(insurance[charges], 0.25)
6. Q2(CHARGES) = PERCENTILE.EXC(insurance[charges], 0.5)
7. Q3(CHARGES) = PERCENTILE.EXC(insurance[charges], 0.75)
8. STANDARD DEVIATION(CHARGES) = STDEV.P(insurance[charges])
9. VARIANCE(CHARGES) = VAR.P(insurance[charges])

10. MIN(CHARGES) = MIN((insurance[Charges]))
11. MAX(CHARGES) = MAX((insurance[Charges]))

SUMMARY STATISTICS OF CHILDREN COLUMN USING DAX:

1. Mean(CHILDREN) = AVERAGE(insurance[children])
2. Median(CHILDREN) = MEDIAN(insurance[children])
3. Mode(children) = MINX(TOPN(1, ADDCOLUMNS(VALUES(insurance[children]), "percent", CALCULATE(COUNT(insurance[children]))),[percent],DESC), insurance[children])
4. RANGE(CHILDREN) = MAX(insurance[children]) - MIN((insurance[children]))
5. Q1(CHILDREN) = PERCENTILE.EXC(insurance[children], 0.25)
6. Q2(CHILDREN) = PERCENTILE.EXC(insurance[children], 0.5)
7. Q3(CHILDREN) = PERCENTILE.EXC(insurance[children], 0.75)
8. STANDARD DEVIATION(CHILDREN) = STDEV.P(insurance[children])
9. VARIANCE(CHILDREN) = VAR.P(insurance[children])
10. MIN(CHILDREN) = MIN((insurance[Children]))
11. MAX(CHILDREN) = MAX(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

OBSERVATIONS

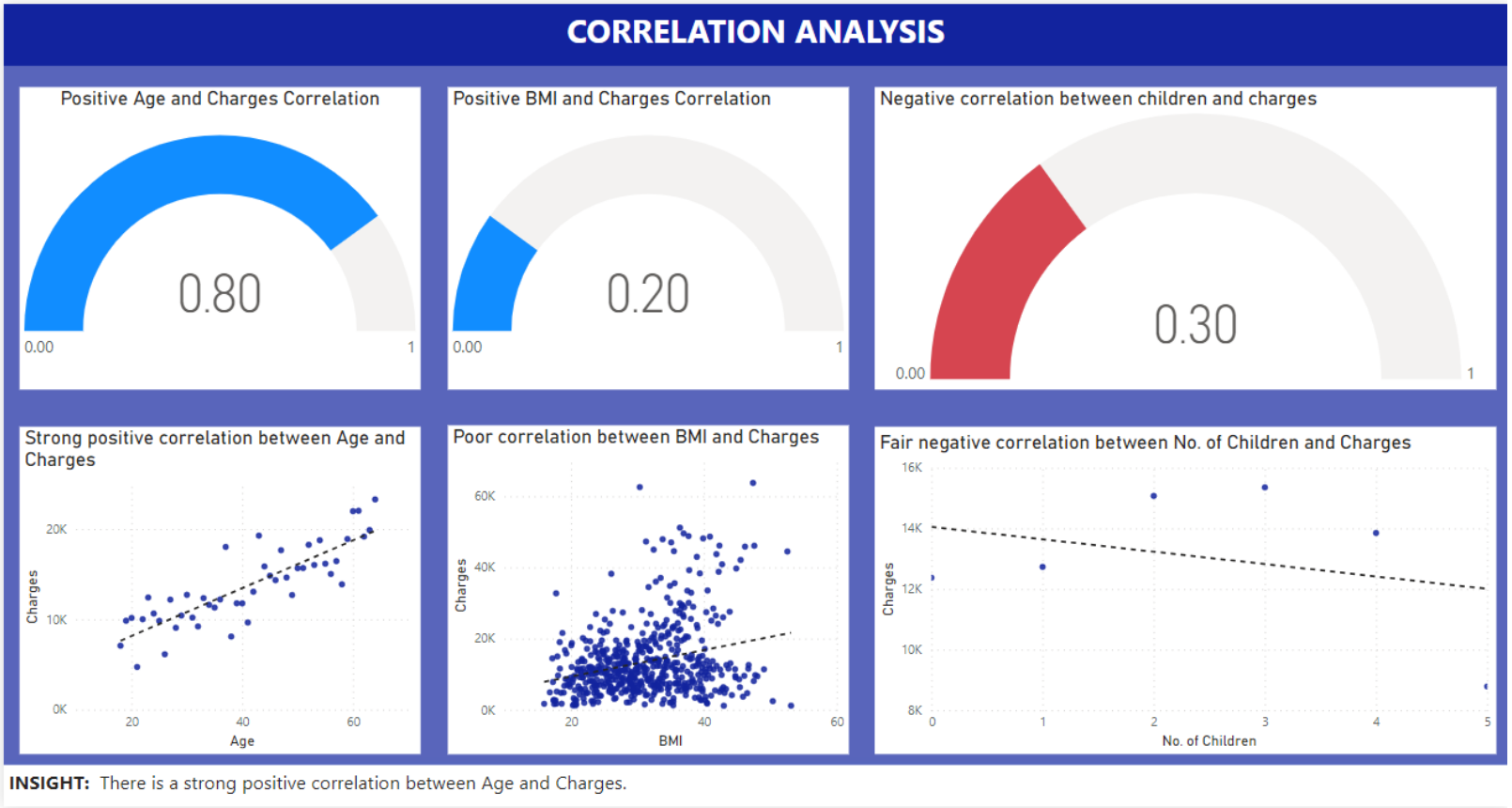
1, About 23% of the customer base falls within the age range of 18 to 25 years old, The lower fees could be a contributing factor. As age increases, so do the charges.

2. The highest charges are applied to approximately 53% of the customer base, who are classified as obese

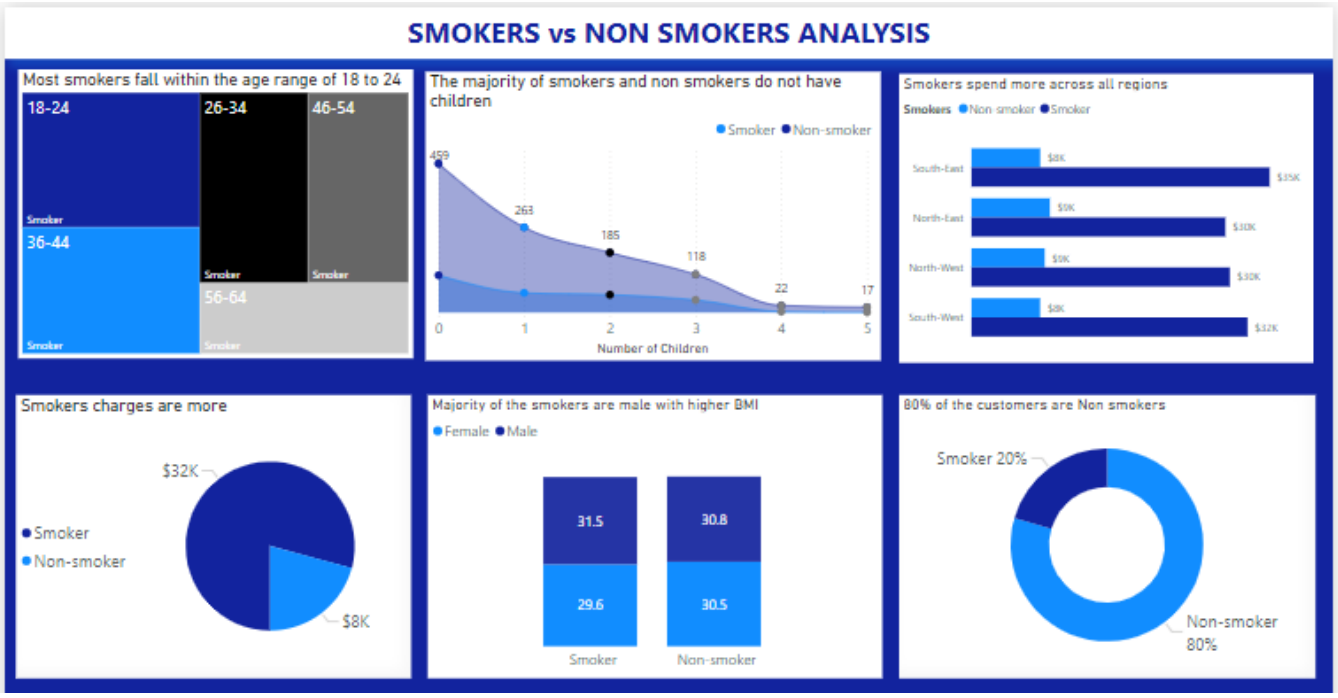
3. The absence of children is characteristic of about 43% of the customer base, leading to increased charges for this demographic.

The average charge for health insurance is $13,271.

CORRELATION ANALYSIS



SMOKER ANALYSIS



Measures to calculate the count of smokers and percentage difference.

1. SmokerCount = CALCULATE(COUNT(insurance[Smokers]), FILTER(insurance, insurance[Smokers]="Smoker"))
2. Smoker_%Difference =

```
VAR AvgChargesSmokers = CALCULATE(AVERAGE('insurance'[Charges]), 'insurance'[Smokers] = "yes")
VAR AvgChargesNonSmokers = CALCULATE(AVERAGE('insurance'[Charges]), 'insurance'[Smokers] = "no")
RETURN
    DIVIDE(AvgChargesSmokers - AvgChargesNonSmokers, [Avg_charges], 0)
```

INSIGHT

A significant portion of smokers, approximately falling between ages 18 to 24, suggests that early adulthood is a critical period for addressing
smoking habits and promoting health awareness.

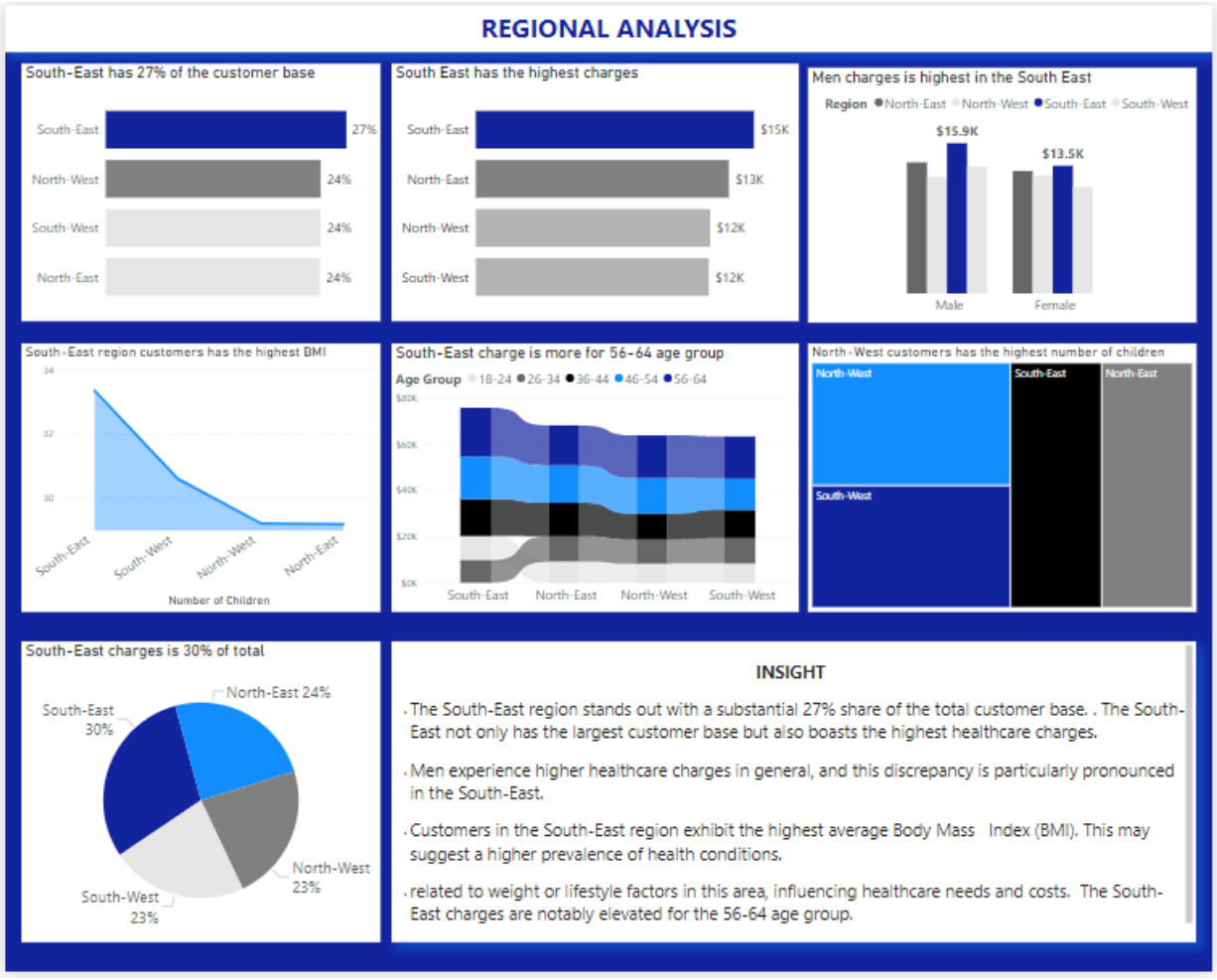
The prevailing absence of children among both smokers and non-smokers implies that family size may not be a primary driver of smoking behavior
or health insurance decisions for this dataset.

Smokers consistently incur higher charges across all regions, highlighting the potential financial burden associated with smoking-related health issues.

The majority of smokers being male emphasizes the gendered nature of smoking habits. Tailoring health interventions to address male-specific
health concerns and smoking cessation programs may be particularly impactful.

The absence of children is characteristic of about 43% of the customer base, leading to increased charges for this demographic.

REGIONAL ANALYSIS



1. South-East Average = CALCULATE(AVERAGE(insurance[Charges]), insurance[Region]="South-East")
2. South-East percent count = DIVIDE(CALCULATE(COUNT(insurance[Region]), insurance[Region]="South-East"), COUNT(insurance[Region]))
3. South-East Percentage of total charges = DIVIDE(CALCULATE(SUM(insurance[Charges]), insurance[Region]="South-East"), SUM(insurance[Charges]))
4. South-East total charges = CALCULATE(SUM(insurance[Charges]), insurance[Region]="South-East")

INSIGHT

A significant portion of smokers, approximately falling between ages 18 to 24, suggests that early adulthood is a critical period for addressing
smoking habits and promoting health awareness.

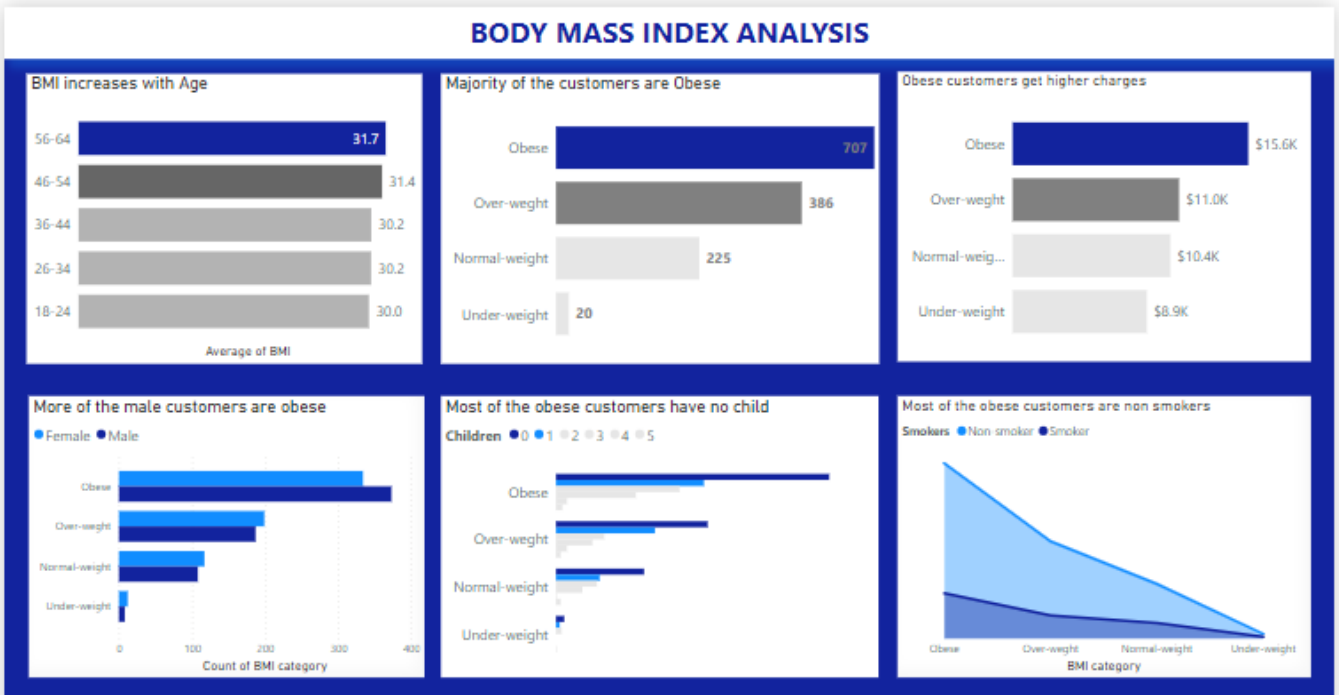
The prevailing absence of children among both smokers and non-smokers implies that family size may not be a primary driver of smoking behavior
or health insurance decisions for this dataset.

Smokers consistently incur higher charges across all regions, highlighting the potential financial burden associated with smoking-related health issues.

The majority of smokers being male emphasizes the gendered nature of smoking habits. Tailoring health interventions to address male-specific
health concerns and smoking cessation programs may be particularly impactful.

The absence of children is characteristic of about 43% of the customer base, leading to increased charges for this demographic.

BMI ANALYSIS



DAX MEASURES

1. 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.

Age group =

```
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"
```

)

2. 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 between 18.5 to 24.9 kg/m^2
- Overweight – BMI between 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 ANALYSIS:

A new table was created for the measures.

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

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

2. Write a DAX measure to calculate the percentage contribution of each categorical variable. This compares the percentage count of each variable to the total.

```
PercentContribution = DIVIDE(COUNT(insurance[Age]), 1339) /*The result is the same if irrespective of the column I use. I decided to default it to age column*/
```

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

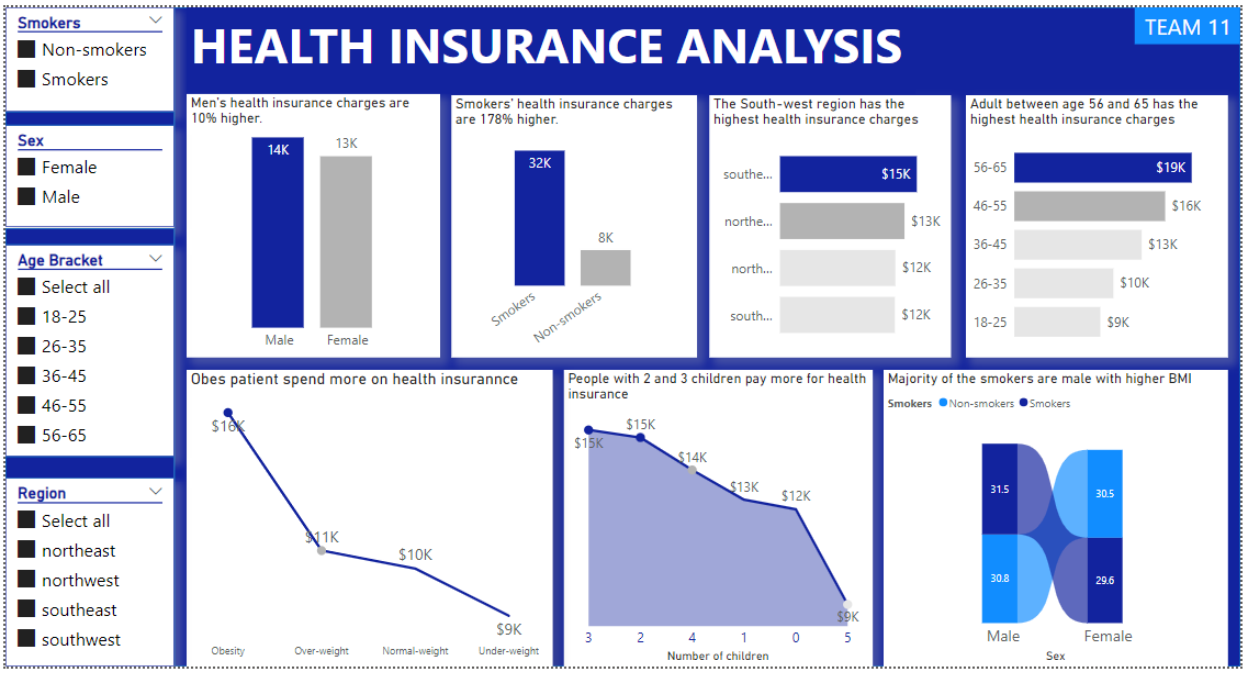
```
Smoker_%Difference =
VAR AvgChargesSmokers = CALCULATE(AVERAGE('insurance'[charges]), 'insurance'[smoker] = "yes")
VAR AvgChargesNonSmokers = CALCULATE(AVERAGE('insurance'[charges]), 'insurance'[smoker] = "no")
RETURN
   DIVIDE(AvgChargesSmokers - AvgChargesNonSmokers, [Avg_charges], 0)
```

DASHBOARD SIZE:
Width: 1200px
Height: 1500px

Colours:

Deep blue: #12239E
Light blue: #118DFF
Deep gray: #B3B3B3
Light gray: #E6E6E6

DASHBOARD



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.
- A significant portion of the customer base for the health insurance company falls within the age range of 18 to 25 years old.

RECOMMENDATIONS

1. **Develop insurance plans:**
   Develop insurance plans that cater specifically to the healthcare needs and preferences of individuals in the 18 to 25 age group. This might include coverage for preventive care, mental health services, and options for fitness and wellness programs.

2. **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.

3. **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.

4. **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.

5. **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.

6. **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.

7. **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.

8. **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.

9. **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.

10. **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.

11. **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.