

Case Study 3

Insurance Premium

Case Study – Insurance Premium

Problem Statement

- In 2022, Singapore's life expectancy is around 83 years old. **Obesity has been rising** in recent years, leading to **a higher risk of health problems** and shortened life expectancy
- Ideally, a **beneficiary claim should not exceed the customer premium claim**. In reality, due to obesity, **obese policyholder's claims had exceeded their** premium
- Obese policyholders have **higher needs** and **demands for medical care**
- **More claims** lead to **lesser profit**

Data

- Data is taken from <https://www.kaggle.com/datasets/simranjain17/insurance> (insurance.csv)
- Age from 18 to 64 years old
- BMI from 15.96 to 53.13
- Premium charges from \$1.12k to \$63.8k
- There is a total of 1,338 policyholders

- Steps taken to collect and consolidate the data

```
import pandas as pd
insurance_csv = 'insurance.csv'
ins = pd.read_csv(insurance_csv)
```

- This is how the data appears before any cleaning

ins

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

Exploratory Data Analysis (1/2)

- Check for duplicated rows and remove if any
- Check for correct data type and any missing values
- Check for any negatives values
- Rename charges column to premium_charges

```
# Check for any duplicated row  
ins.loc[ins.duplicated(), :]
```

	age	sex	bmi	children	smoker	region	charges
581	19	male	30.59	0	no	northwest	1639.5631

```
# Remove duplicate and keep original  
ins.drop_duplicates(inplace=True)
```

```
ins = ins.rename(columns={"charges": "premium_charges"})
```

```
((ins['bmi'] < 0) & (ins['charges'] < 0)).values.any()  
#check if there is any negative values  
False
```

```
ins.dtypes
```

```
age          int64  
sex          object  
bmi          float64  
children     int64  
smoker       object  
region       object  
charges      float64  
dtype: object
```

```
ins.isna().any()
```

```
age          False  
sex          False  
bmi          False  
children     False  
smoker       False  
region       False  
charges      False  
dtype: bool
```

Exploratory Data Analysis (1/2)

- age_group and bmi_group was created from bmi and age columns
- bmi and premium_charges was converted to 2 decimal place
- After data is cleaned, saved to insurance_cleaned.xlsx format

	age	age_group	sex	bmi	bmi_group	smoker	children	region	premium_charges
0	19	Young Adults	female	27.90	Overweight	yes	0	southwest	16,884.92
1	18	Young Adults	male	33.77	Obese	no	1	southeast	1,725.55
2	28	Young Adults	male	33.00	Obese	no	3	southeast	4,449.46
3	33	Young Adults	male	22.70	Healthy	no	0	northwest	21,984.47
4	32	Young Adults	male	28.88	Overweight	no	0	northwest	3,866.86
...
1333	50	Middle Age	male	30.97	Obese	no	3	northwest	10,600.55
1334	18	Young Adults	female	31.92	Obese	no	0	northeast	2,205.98
1335	18	Young Adults	female	36.85	Obese	no	0	southeast	1,629.83
1336	21	Young Adults	female	25.80	Overweight	no	0	southwest	2,007.94
1337	61	Seniors	female	29.07	Overweight	yes	0	northwest	29,141.36

```
def age_group(x):  
    if x >= 18 and x <=35:  
        return 'Young Adults'  
    elif x >= 36 and x <=55:  
        return "Middle Age"  
    else:  
        return "Seniors"
```

```
ins['age_group'] = ins['age'].apply(age_group)
```

```
def bmi_group(x):  
    if x < 18.5:  
        return 'Underweight'  
    elif x > 18.5 and x <=24.9:  
        return "Healthy"  
    elif x > 25.0 and x <=29.9:  
        return "Overweight"  
    else:  
        return "Obese"
```

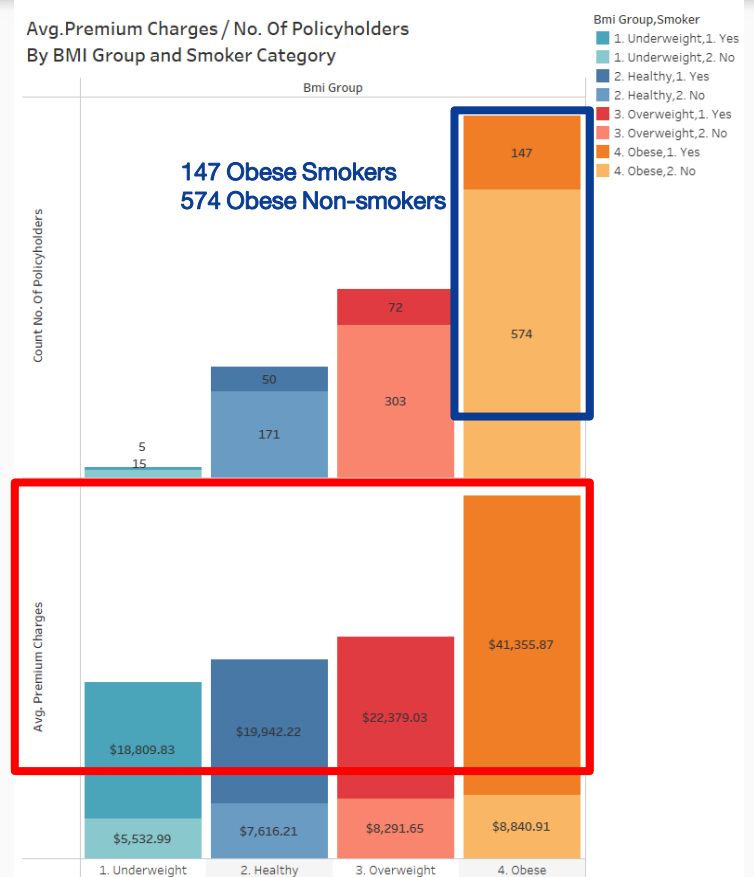
```
ins['bmi_group'] = ins['bmi'].apply(bmi_group)
```

```
ins['bmi'] = ins['bmi'].map(lambda x: round(x, 2))  
ins['premium_charges'] = ins['premium_charges'].map(lambda x: round(x, 2))  
ins.head()
```

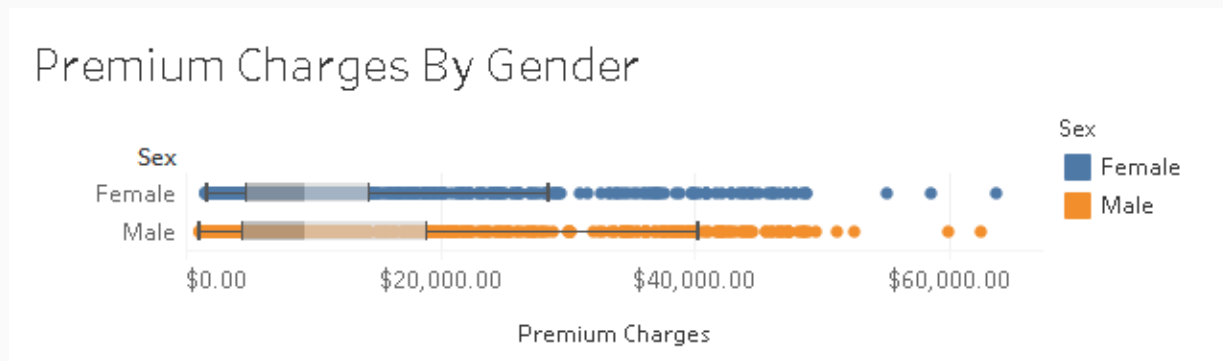
```
ins.to_excel("insurance_cleaned.xlsx")
```

Exploratory Data Analysis (2/2)

- Average Premium Charges / No. Of Policyholders plotted against BMI Group and Smoker Category
- There is a total of 721 obese policyholders
- As BMI increases, charges increase as well



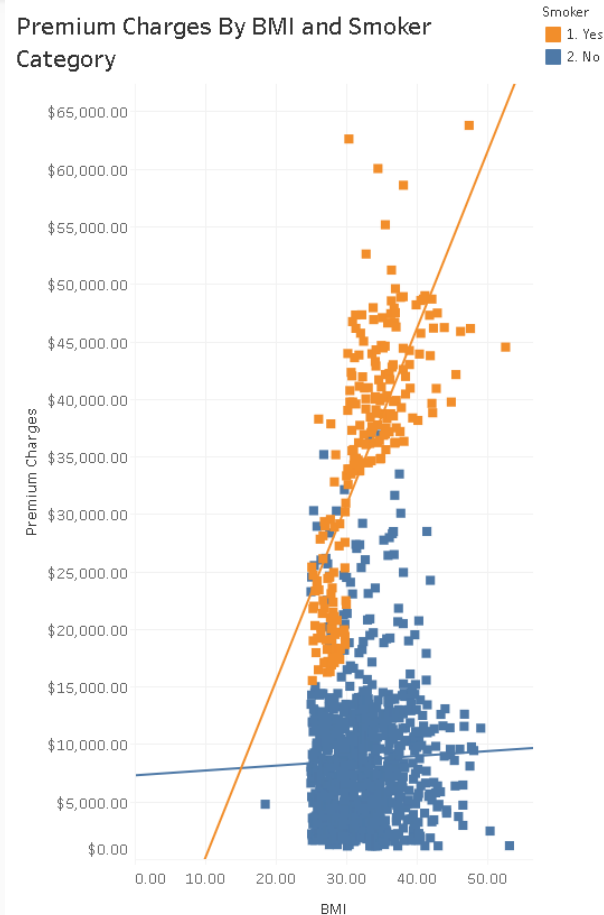
Exploratory Data Analysis (2/2)



- Premium Charges plotted against Gender
- Males are charged more (as compared to Females)

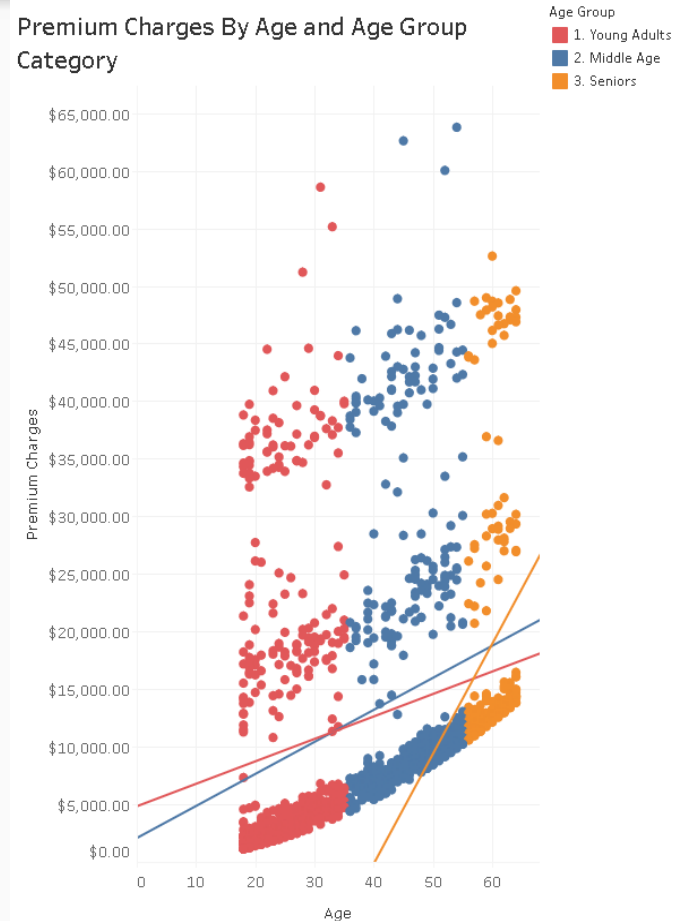
Exploratory Data Analysis (2/2)

- Premium Charges plotted against BMI and Smoker category
- Charges increases with BMI
- Smokers are charged more (as compared to non-smokers)



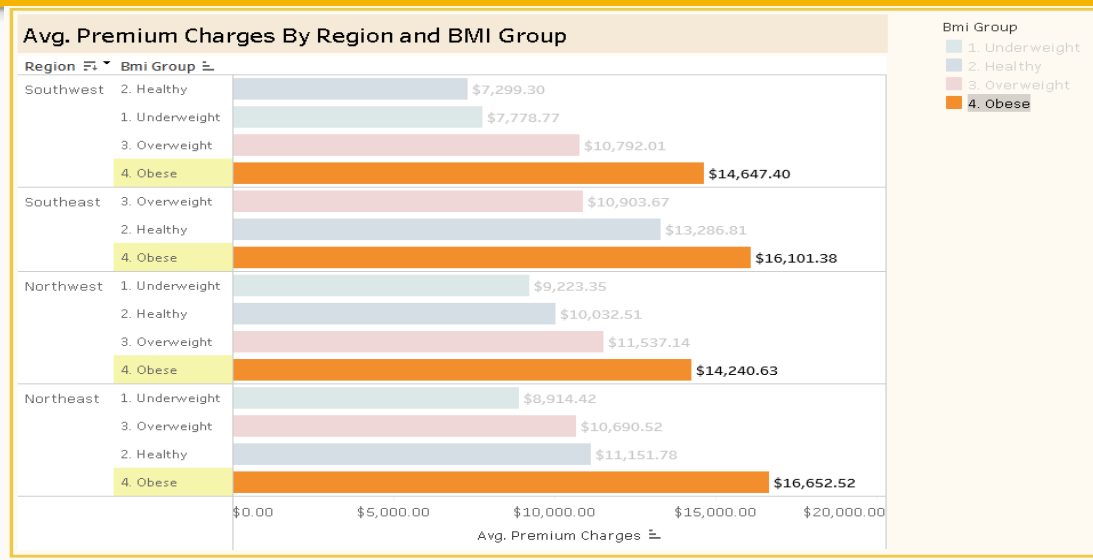
Exploratory Data Analysis (2/2)

- Premium Charges plotted against Age and Age Group
- Charges increases with Age
- Seniors are charged more (as compared to young adults and middle aged)



<https://public.tableau.com/app/profile/kaijie.ng/viz/CaseStudyInsurance/InsurancePremiumsDashboard?publish=yes>

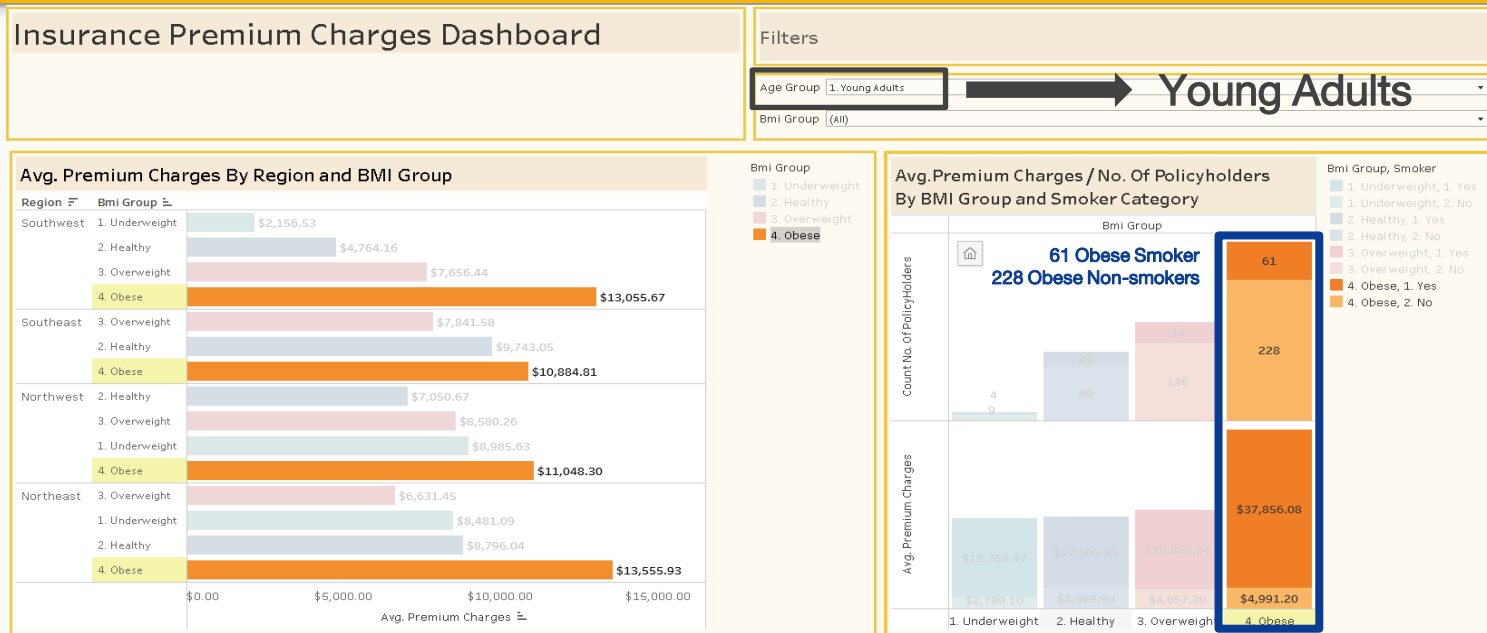
Strategy: Key Recommendations



Who are the main customers?

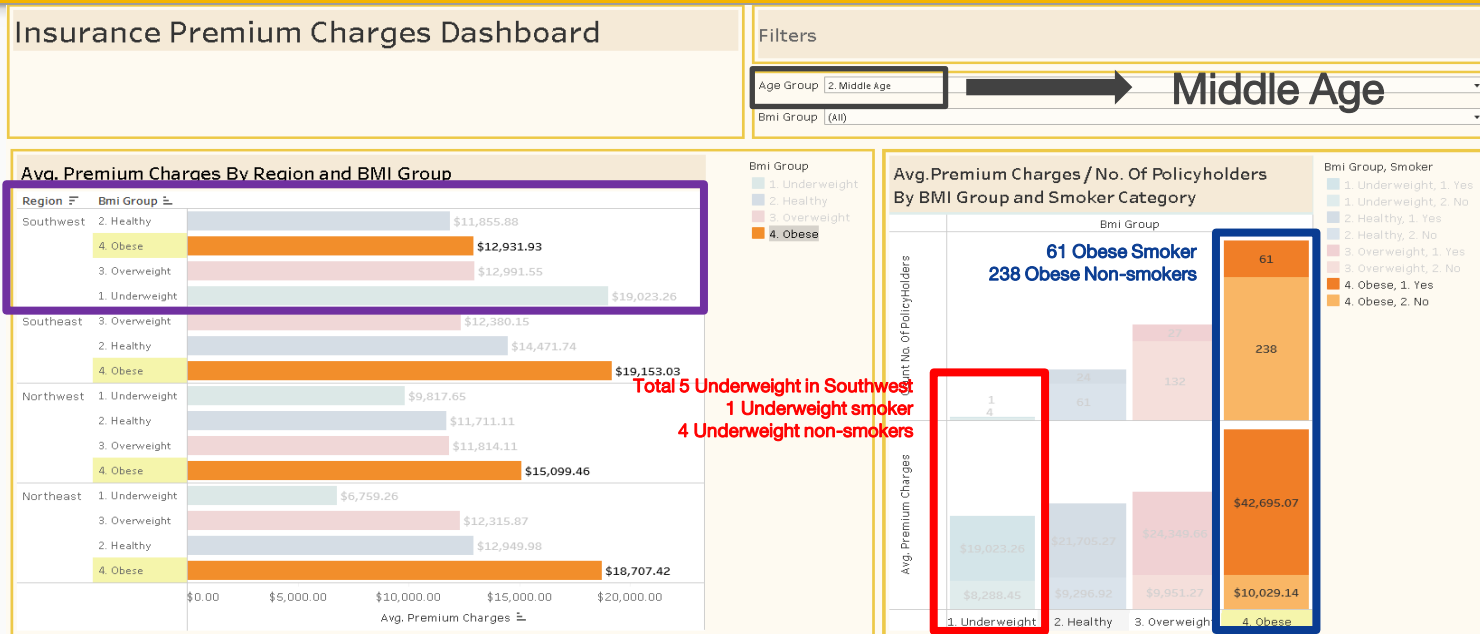
- Obese policyholders paid the highest Average Premium Charges across all different regions
- We will further analyse on the 3 different types of Age Group (Young Adults, Middle Age & Seniors)

Strategy: Key Recommendations



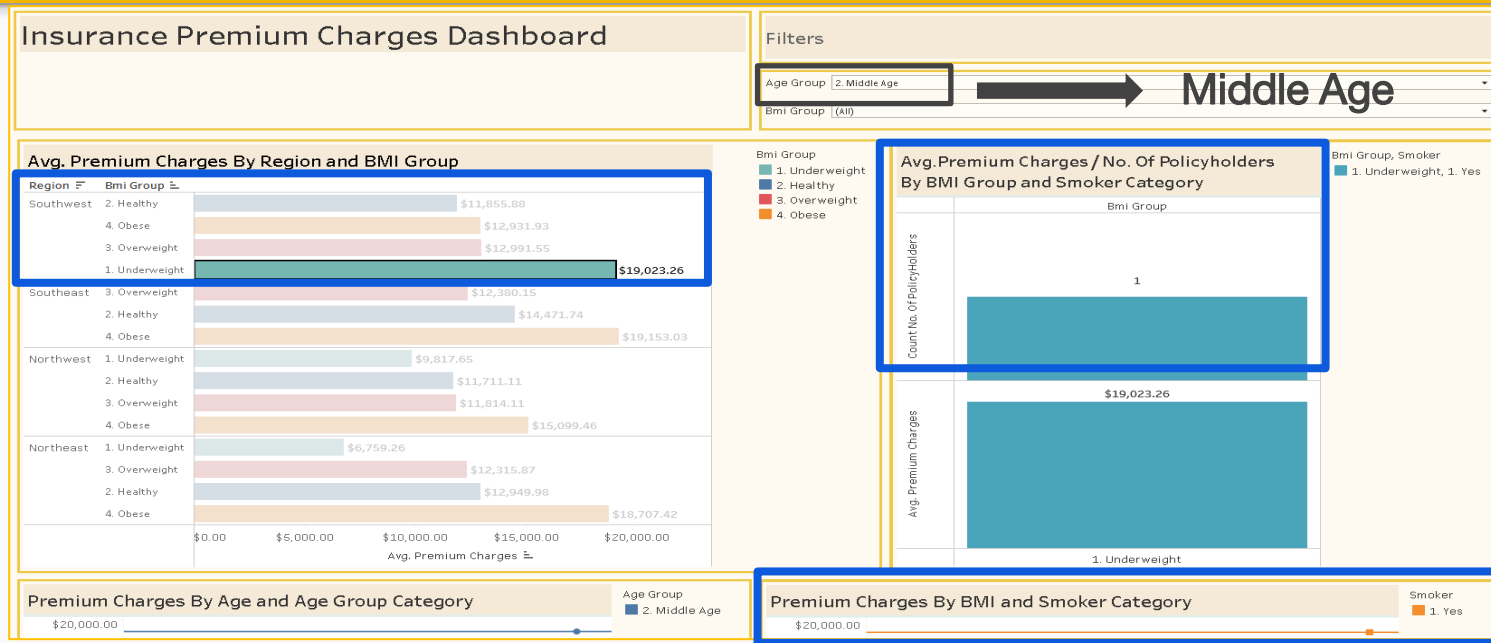
- Obese policyholders paid the highest Average Premium Charges across all different regions
- There is a total of 61 obese smokers and 228 obese non-smokers policyholders in Young Adults group

Strategy: Key Recommendations



- Obese policyholders paid the highest Average Premium Charges across all different regions except Southwest Underweight policyholders
- There is only a total of 5 underweight policyholders in Middle Age group
- There is a total of 61 obese smokers and 238 non-smokers policyholders in Middle Age group

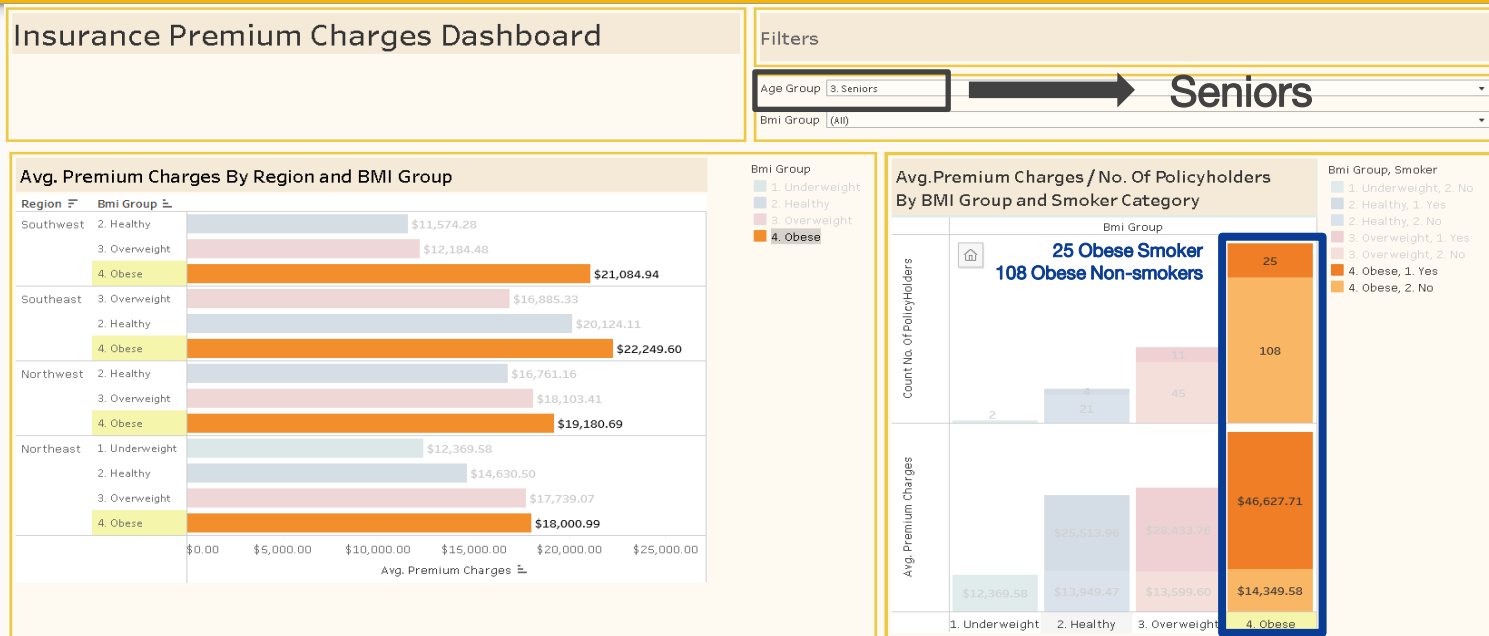
Strategy: Key Recommendations



Why Middle Age group in Southwest Underweight group paid the highest Average Premium Charges ?

- There is only 1 policyholder in Middle Age group in Southwest that is underweight
- Policyholder is a smoker therefore charged at a higher premium

Strategy: Key Recommendations



- Obese policyholders paid the highest Average Premium Charges across all different regions
- There is a total of 25 obese smokers and 108 non-smokers policyholders in Seniors group

Strategy: Key Recommendations

To reduce obese people from claiming, the company can help them to lose weight while maintaining customer retention rate at the same time

- The recommendation is to collaborate with government to help promote health awareness
- Study customer's lifestyle habits, and reward them with food/entertainment vouchers if they hit a certain quota of exercising

Strategy: Key Recommendations

To attract new customers, the company can implement a new Insurance Enhance Benefit Package (IEBP), mainly targeting obese people

- Rate of return (ROR) for normal BMI 5%
- IEBP ROR 3% only, but if their weight hits optimal weight, the package will auto-enhance to 7% for a certain period

Conclusion

- Data analysis has identified **obese policyholders paid the highest across all different regions** as main customers, but there could be changes with more data in the future
- To **reduce more claims** from the obese group, the company can **collaborate with government to help promote health awareness**
- **Study customers' lifestyle habits** and **reward them with food/entertainment vouchers** if they hit a certain exercise quota. This helps in **customer retention**
- To **attract new customers**, the company can implement a new Insurance Enhance Benefit Package **targeting obese customers**
- However, **competitors** might also implement IEBP **at an even lower price**. The company has to consider these to **determine profit or loss**
- **Surveys** needed for customers' preferred **vouchers** and **government health agency collaboration**

Conclusion

What have you learnt?

- I have learnt new technical skills like python, tableau and sql
- Identify different sources of data and basic data quality standards
- Apply the right visualisation for different types of charts

What were the main challenges in applying data analytics skills to this project?

- To identify and research problem statement and recommendations
- Limited data columns from excel

THANK YOU