Capstone Project

Insurance Premium Data

By Kwek Ming Sheng

Case Study – Insurance Premium Data

Background information

The insurance industry is driven by the simple fact that the capital spent by the insurance company in response to beneficiary claims should not exceed customer premium. The higher the difference between the approved customer claims and total premium received, equates to higher profits. Generally the more data we have on a customer, the better we are able to understand their needs, as well as assess the risks involved in insuring them.

Problem Statement

The insurance company aims to remain profitable. We seek to identify factors that influence the premiums charged to customers and determine factors that results in the highest premium charged so that the insurance company can increase the premiums collected. Provide recommendations that would increase the premiums collected by the insurance company via our analysis on the relevant factors.

- We need to analyze the factors influencing the premium prices charged to customers.
- We find out the main customers and differentiate the highest and lowest premium paying customers.

Data

- Data is sourced from https://www.kaggle.com/simranjain17/insurance
- Data consists a total of 1338 rows and 7 columns.

Exploratory Data Analysis (1/14)

- Before data cleaning.
- Import the required packages.

```
import pandas as pd
import numpy as np
import os
```

- Check and change directory / import csv file.

```
pwd
```

```
'C:\\Users\\Ming Sheng'
```

```
os.getcwd()
```

```
'C:\\Users\\Ming Sheng'
```

```
os.chdir("C:\\Users\\Ming Sheng\\Desktop\\Ming Sheng\\Data Analytics\\Course\\Project")
```

```
df_orders = pd.read_csv('insurance.csv', encoding = 'utf-8')
df_orders
```

	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

Exploratory Data Analysis (2/14)

- Check data types and identify presence of missing data.

```
df_orders.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

```
df_orders.isna().any()
```

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

```
df_orders.isna().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Exploratory Data Analysis (3/14)

- Perform data cleaning.
- Converted bmi column in df_orders dataframe to 1 decimal place.
- Converted charges column in df_orders dataframe to 2 decimal places.

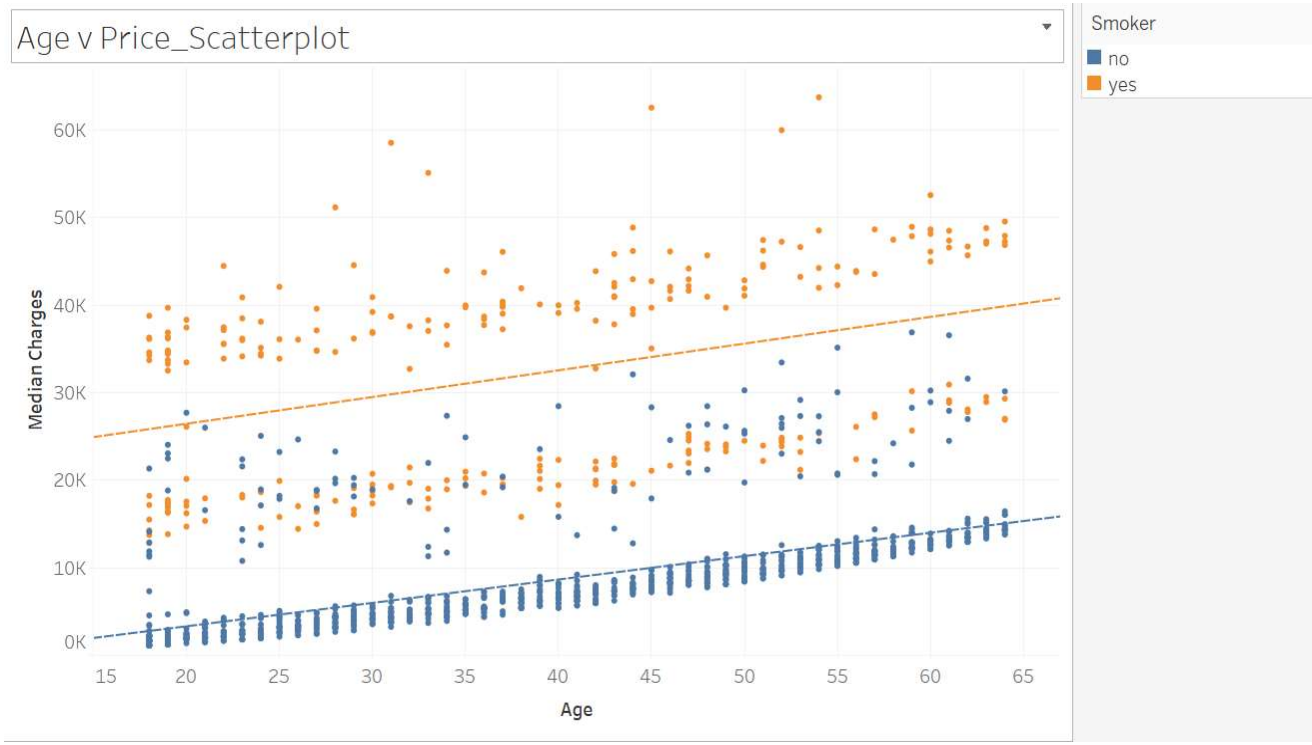
```
df_orders.round({"bmi":1,"charges":2})
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86
...	...	...	...	...	...	...	...
1333	50	male	31.0	3	no	northwest	10600.55
1334	18	female	31.9	0	no	northeast	2205.98
1335	18	female	36.8	0	no	southeast	1629.83
1336	21	female	25.8	0	no	southwest	2007.94
1337	61	female	29.1	0	yes	northwest	29141.36

1338 rows × 7 columns

Exploratory Data Analysis (4/14) - Age

- Perform exploratory data analysis.
- Based on my analysis, the median insurance charges has a positive correlation with the age of the customers.
- The median charges is higher for smokers compared to non-smokers.



Exploratory Data Analysis (5/14) - Smokers

- Based on my analysis, the median insurance charges for smokers amounts to $65,035, is higher for smokers compared to non-smoker (i.e. total median charges $14,625).

- Although the number of smokers is approximately 5 times lower than non-smokers. The median charges for smokers is approximately 5 times higher than non-smokers.


```
df_dec.groupby("smoker").size()
```

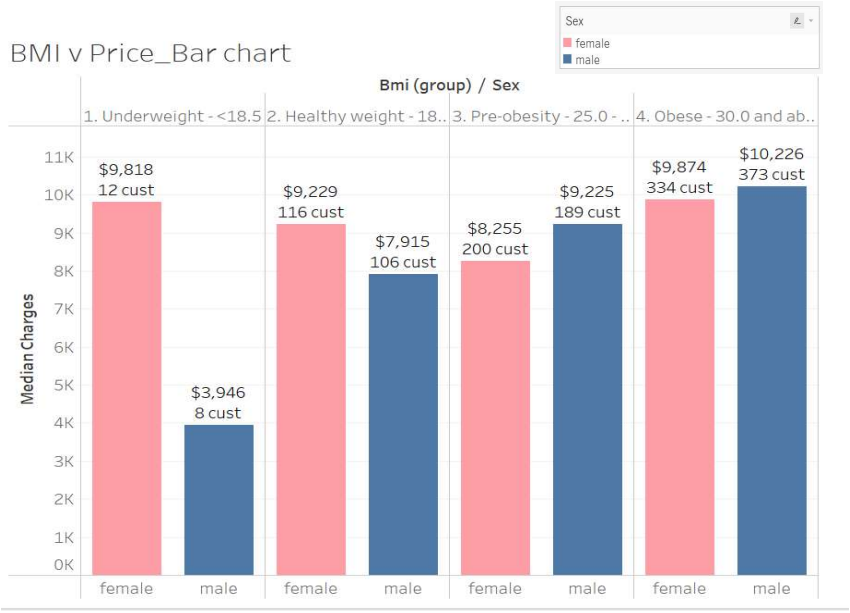
```
smoker
no      1064
yes      274
dtype: int64
```

Exploratory Data Analysis (6/14) - BMI

- Based on my observation, the underweight female smokers between the age 30 to 39 has a larger median charges than the female smokers in the healthy and pre-obesity BMI category.

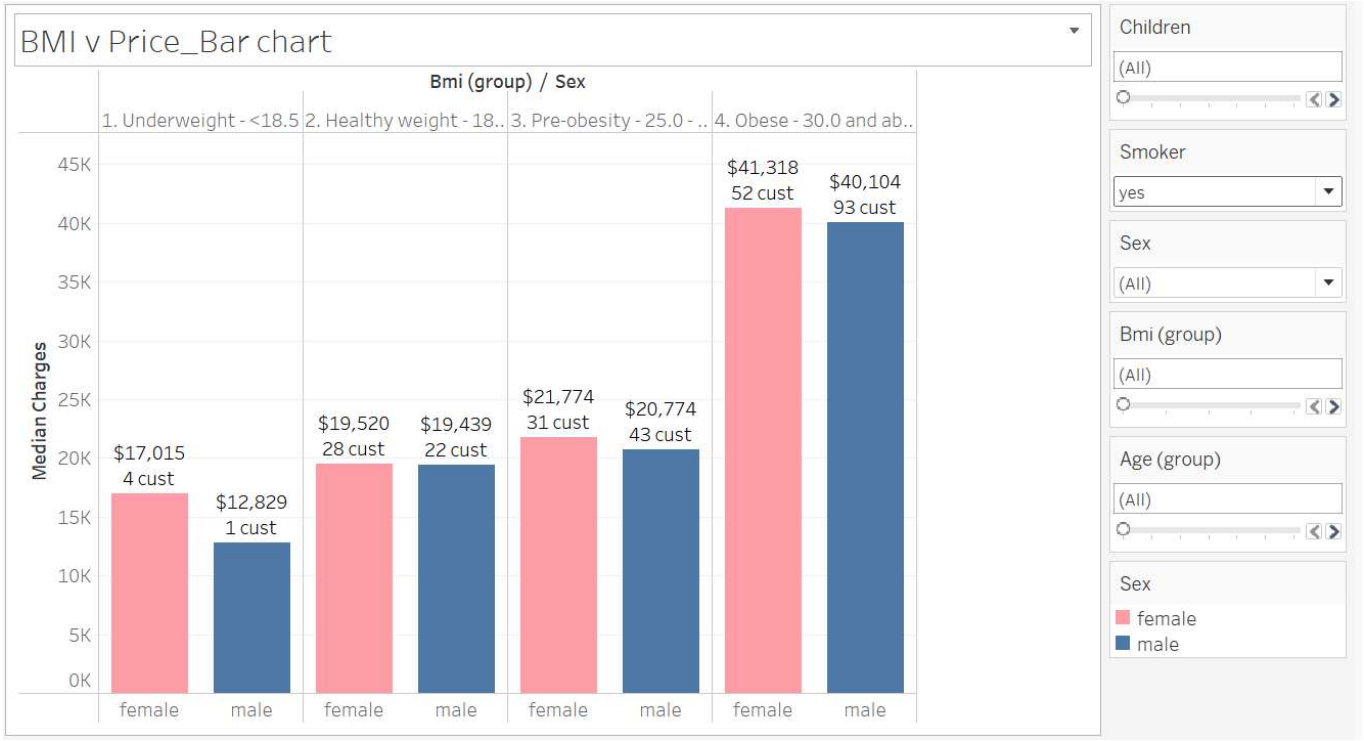
Exploratory Data Analysis (7/14) - BMI

- Based on the diagram below, the median insurance charges is <u>generally similar across all BMI groups</u>.



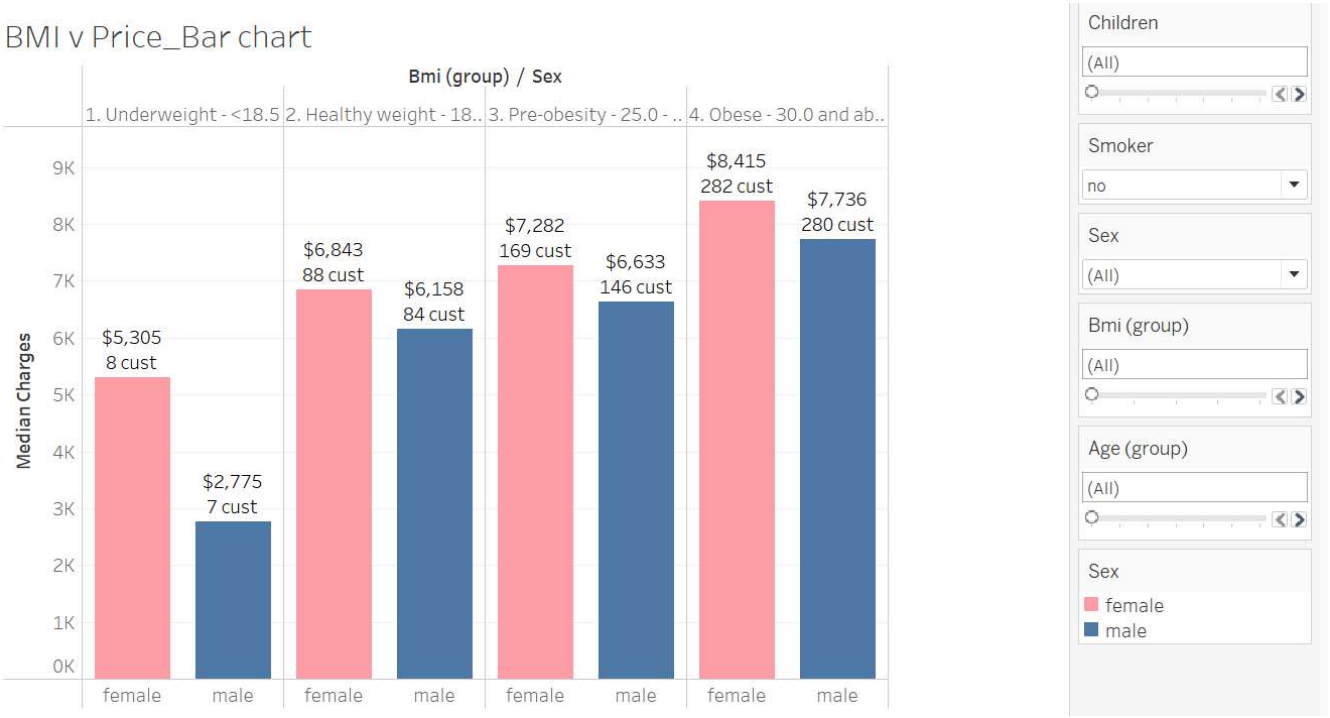
- The median charges for underweight female customers and male customers is <u>$9,818 and $3,946 respectively</u>. The difference arose from <u>the number of female customers (i.e. 4 customers) that smoke compared to male customers (i.e. 1 customer)</u>.

Exploratory Data Analysis (8/14) - BMI



- However, the median insurance charges of the <u>obese customers who smoke are significantly higher than smokers in other BMI categories and non-smokers.</u>

Exploratory Data Analysis (9/14) - BMI

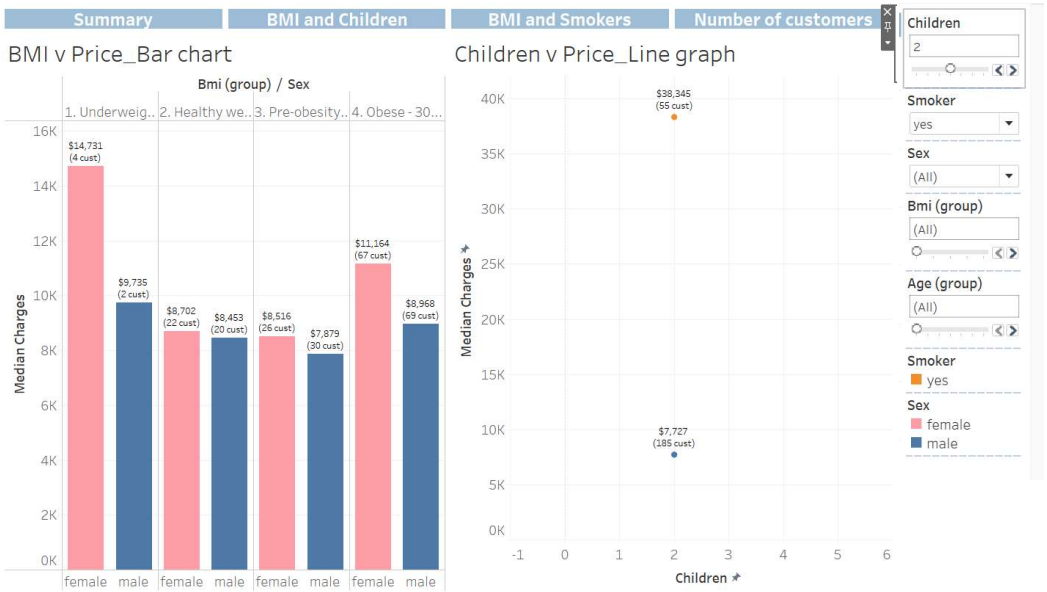


- For non-smoking customers, I observed a <u>positive correlation</u> between the median insurance charges to the non-smoking customers' BMI.

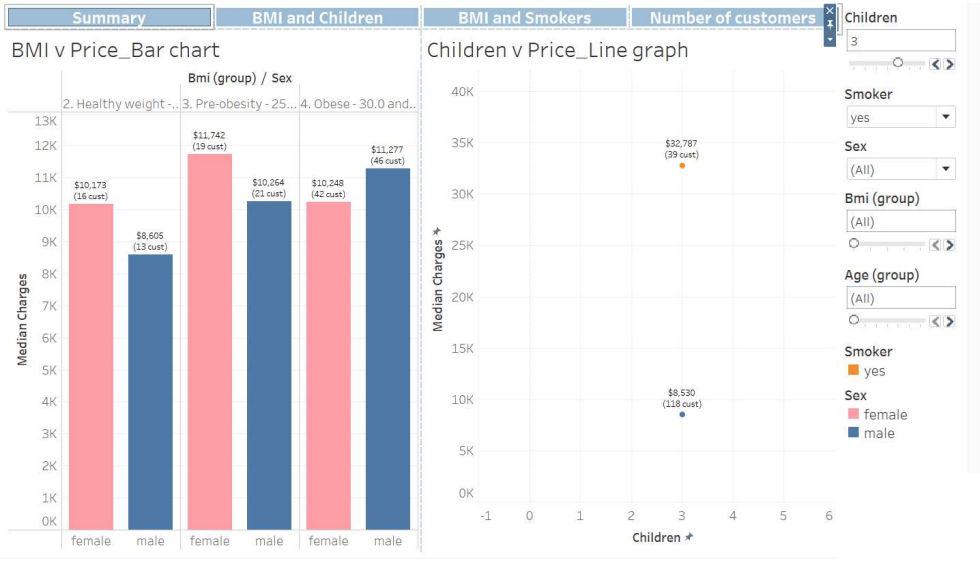
Exploratory Data Analysis (10/14) – Children

- The customers who smoke has a higher median insurance charges than the non-smokers.
- The median charges for customers who smoke has a negative correlation to number of children. Upon further analysis, there are fewer customers under the obese category with 3 or more children. Refer to slides 14 to 15.

Exploratory Data Analysis (11/14) – Children



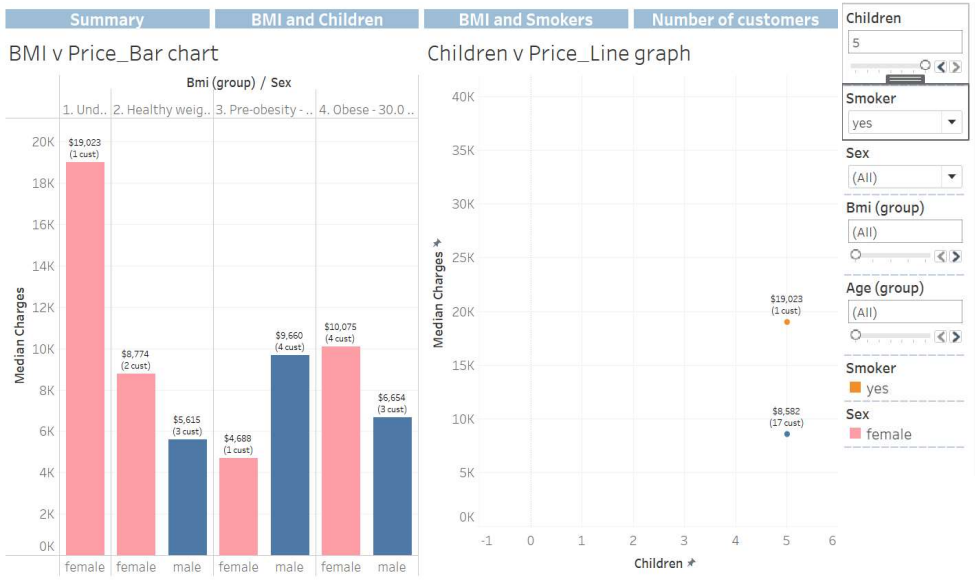
Number of children : 2



Number of children : 3

Exploratory Data Analysis (12/14) – Children

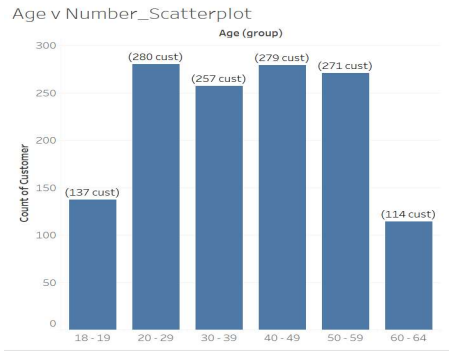
Number of children : 4



Number of children : 5

Exploratory Data Analysis (13/14) – Main customers

- Based on the data, the main customers falls between the ages 20 to 59, but have an average age of 39 years old.




```
df_dec.describe()
```

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663303	1.094918	13270.422280
std	14.049960	6.098257	1.205493	12110.011259
min	18.000000	16.000000	0.000000	1121.870000
25%	27.000000	26.300000	0.000000	4740.287500
50%	39.000000	30.400000	1.000000	9382.030000
75%	51.000000	34.700000	2.000000	16639.915000
max	64.000000	53.100000	5.000000	63770.430000

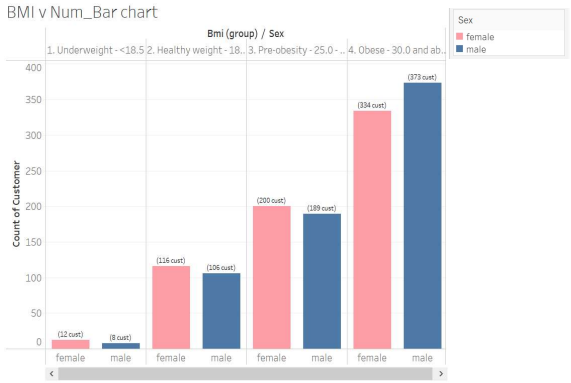
```
df_dec.groupby("sex").size()
```

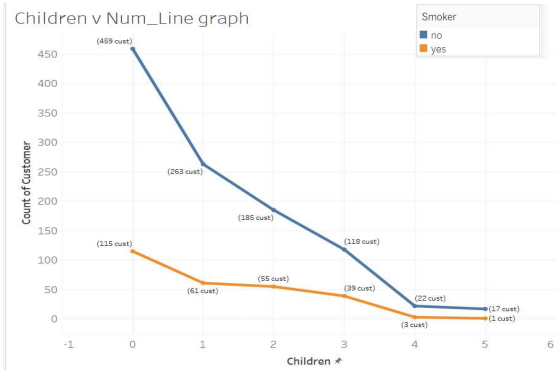
```
sex
female    662
male      676
dtype: int64
```

```
df_dec.groupby("region").size()
```

```
region
northeast    324
northwest    325
southeast    364
southwest    325
dtype: int64
```

- The majority of customers are obese, non-smokers and have no child.






```
df_dec.groupby("smoker").size()
```

```
smoker
no      1064
yes      274
dtype: int64
```

Exploratory Data Analysis (14/14) – Top 5 customers

```
df_sort.head()
```

	age	sex	bmi	children	smoker	region	charges
940	18	male	23.2	0	no	southeast	1121.87
808	18	male	30.1	0	no	southeast	1131.51
1244	18	male	33.3	0	no	southeast	1135.94
663	18	male	33.7	0	no	southeast	1136.40
22	18	male	34.1	0	no	southeast	1137.01

Top 5 customers charged with the lowest premium.

```
df_sort.tail()
```

	age	sex	bmi	children	smoker	region	charges
819	33	female	35.5	0	yes	northwest	55135.40
577	31	female	38.1	1	yes	northeast	58571.07
1230	52	male	34.5	3	yes	northwest	60021.40
1300	45	male	30.4	0	yes	southeast	62592.87
543	54	female	47.4	0	yes	southeast	63770.43

Top 5 customers charged with the highest premium.

Tableau Dashboard

Tableau link: Data analysis project - Insurance premiums charged | Tableau Public

Strategy : Key Recommendations (1/2)

- The factors with affecting the premium charged are mainly the BMI, age and smokers. The most significant factor influencing the premium charged relates to the customers being a smoker.

- It is recommended to consider having a quit-smoking campaign, to ensure customers have a healthier lifestyle. Given that profits are derived from the difference between premium charges and claims made by customers, this will reduce the likelihood of insurance company incurring insurance pay-outs.

- From slide 9, there is a significant difference in the insurance charged for the underweight female smokers between the age 30 to 39. We can infer that the current age group are likely to be working adults with a higher purchasing power. I recommend carrying out surveys to understand the specific age group's customer's needs rather than relying on the customer's profile. Thereafter, perform marketing campaigns (i.e. road shows / advertisements) to target the relevant age group.

Strategy : Key Recommendations (2/2)

Who are the main customers?

- From slide 16, we observed that the majority of the customers are obese, non-smokers, with no child and are between ages 20 to 59.

Who are the top 5 customers who paid the lowest and highest premiums?

- From slide 17, we observed that the top 5 customers who paid the most are mainly obese, smokers, with 3 children or less and are between ages 30 to 49.
- Based on the above analysis on premium charged, I would recommend that the company focus on coming up with varied products tailored for the above customers between ages 20 to 49 who are obese and smoke.
- The company can introduce a loyalty programme for customers ages 20 to 29. Given that this age group is generally younger, the company can continue carry out financial planning and sell insurance products for these customers. There are more sales opportunities as the financial needs of the customer will change as they progress into their parental or aging life stages.
- Nevertheless, the management should continue to consider the amount of claims made to determine the profitability of this age group.

Conclusion

- The main factor which influence the premium prices are whether the customers smoke. Evidently, this indicates those who smoke have the highest health risk.
- With data analysis, we identified that the majority of the customers are obese, non-smokers, with no child and are between ages 20 to 59. While top 5 customers who paid the most are mainly obese, smokers, with 3 children or less and are between ages 30 to 49.
- In conclusion, the customers who pays the highest premium are those who are obese and smoke. However, the data on insurance claims needs to be taken into consideration in order to determine the profitability of customers profile.

For Further exploration

- Analyze the data on insurance claims needs to be taken into consideration in order to determine the profitability of customers profile.

- Further examine the data related to specific health insurance products and the correlations with customer’s purchase.

THANK YOU!