Insurance Premium Data

Beatrice Kay

Case Study of Insurance Premium Data

Problem Statement:

You work for an insurance company as a data analyst and the agent operations department has requested for a review on all health insurance policy holders. The objective is to provide all insurance agents with a better understanding of their clients background to better serve their needs. It will also reveal insights on who their main client group is and if there are new market trends or potential niches to tap into to gain more customers.

They have provided you with a list of questions to aid your analysis:

- Who is charged a higher premium?
- Who is charged a lower premium?
- Who are the main customers?

Data

- Data is taken from https://www.kaggle.com/simranjain17/insurance
- No information on when or where the data is collected from
- Data is based on health insurance premium charged to each policyholder (assume per year) based on age, gender, BMI, smoking habits and region
- There are a total of 1338 entries in the dataset

Exploratory Data Analysis (1/2)

- Raw data:

```
csv_file = 'insurance.csv'
insurance = pd.read_csv(csv_file)
insurance.head()
```

	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

```
insurance.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
```

```
insurance.isna().any()
```

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

- No missing values observed

Exploratory Data Analysis (1/2)

- charges converted to INT to remove decimal places

```
insurance["charges"] = insurance["charges"].astype(int)
insurance.head()
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884
1	18	male	33.770	1	no	southeast	1725
2	28	male	33.000	3	no	southeast	4449
3	33	male	22.705	0	no	northwest	21984
4	32	male	28.880	0	no	northwest	3866

- Check if min and max values of numerical values are valid and within appropriate range
  - **Age:** 18 – 64 (no zero values)
  - **BMI:** 15.96 – 53.13 (no zero values)

```
insurance.describe()
```

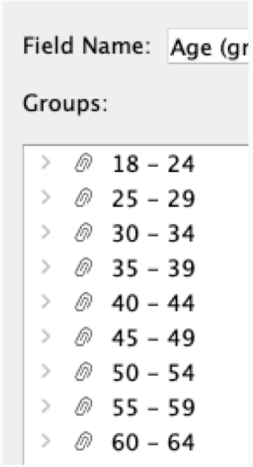
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13269.928999
std	14.049960	6.098187	1.205493	12110.012755
min	18.000000	15.960000	0.000000	1121.000000
25%	27.000000	26.296250	0.000000	4740.000000
50%	39.000000	30.400000	1.000000	9381.500000
75%	51.000000	34.693750	2.000000	16639.250000
max	64.000000	53.130000	5.000000	63770.000000

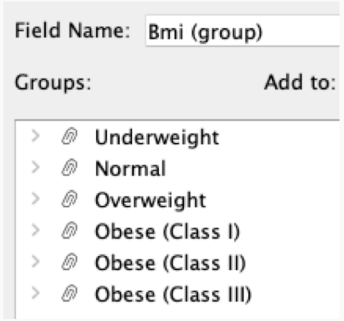
Exploratory Data Analysis (2/2)

- Insurance premium charges vary based on a few factors:

Data	Data Type
Age	Continuous
BMI	Continuous
Gender	Discrete (Male/Female)
Children	Discrete (0, 1, 2, 3, 4, 5)
Smoking	Discrete (Yes/No)
Region	Discrete (NE, NW, SE, SW)

- Groups created in Tableau to further segment data for age and bmi:





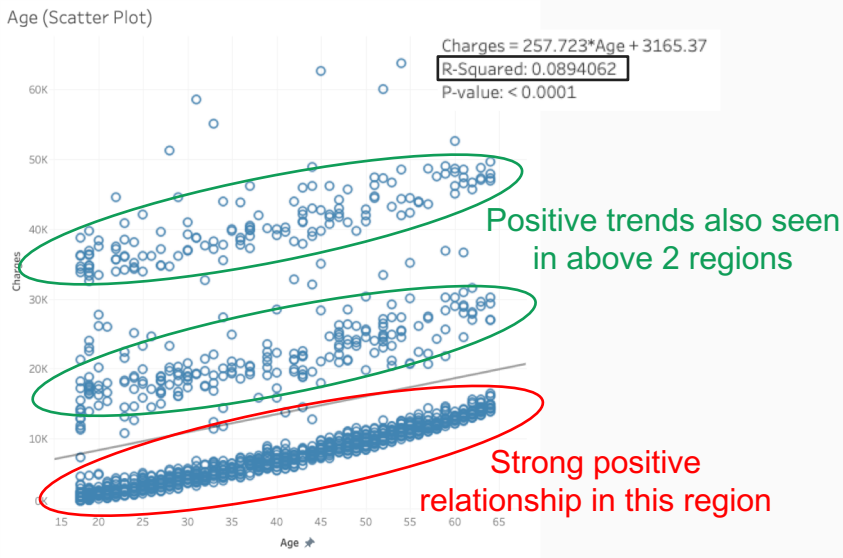
Classification	BMI (kg/m^2)	Risk of comorbidities
Underweight	<18.5	Low (but risk of other clinical problems increased)
Normal range	18.5–24.9	Average
Overweight (preobese)	25.0–29.9	Mildly increased
Obese	≥30.0	
Class 1	30.0–34.9	Moderate
Class 11	35.0–39.9	Severe
Class 111	≥40.0	Very severe

smartcademy

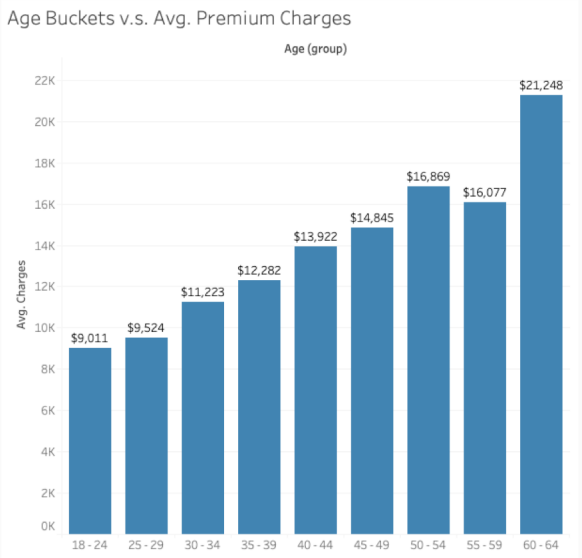
Exploratory Data Analysis (2/2) | Continuous Variables

(1) Age v.s. Premium Charges

- Premium charges generally increase with age
- Large amount of outliers weakens overall correlation coefficient. However, if data is further split into separate regions with their own best fit lines, individual positive trends can be observed. This could be differentiating between the different tiers of health insurance, i.e. higher premiums with higher coverage. Or individuals with underlying health conditions or other risk factors that would require a higher premium



Scatter plot of age and premium charge



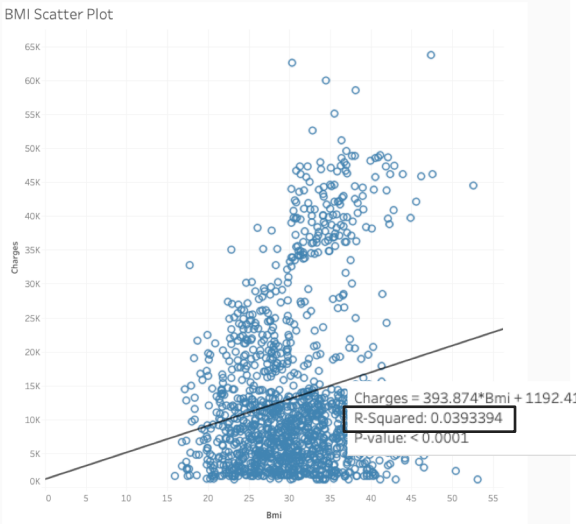
Bar graph of age bucket and avg premium charge

smartcademy

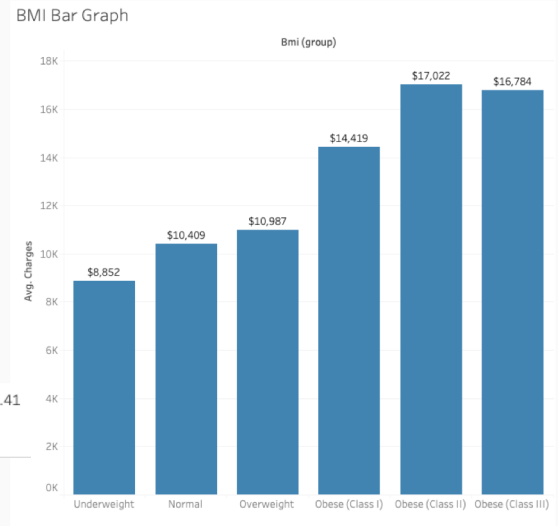
Exploratory Data Analysis (2/2) | Continuous Variables

(2) BMI v.s. Premium Charges

- Relationship between BMI and premium charge is weak as correlation coefficient is close to 0.
- Insurance companies do not solely rely on BMI as a health indicator as you could have a high BMI and not have any health complications. However, most individuals with high BMI typically have weight-related health issues. Individuals could have been charged higher premiums because of: (1) underlying health issues unrelated to BMI, (2) weight-related health issues related to BMI, (3) or that they voluntarily opted for a higher coverage plan



Scatter plot of BMI and premium charge



Bar graph of BMI and avg premium charge

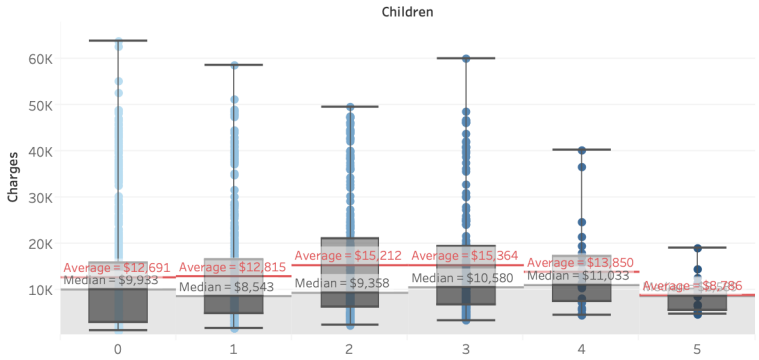
Classification	BMI (kg/m^2)	Risk of comorbidities
Underweight	<18.5	Low (but risk of other clinical problems increased)
Normal range	18.5–24.9	Average
Overweight (preobese)	25.0–29.9	Mildly increased
Obese	≥30.0	
Class 1	30.0–34.9	Moderate
Class 11	35.0–39.9	Severe
Class 111	≥40.0	Very severe

smartcademy

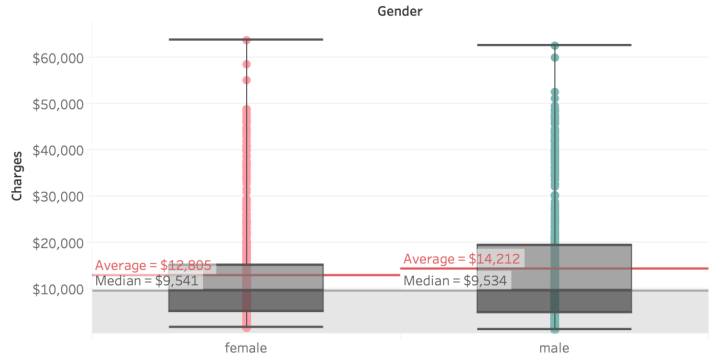
Exploratory Data Analysis (2/2) | Discrete Variables

Smoking (2)

Children (2)



Gender (2)



Region (2)

Majority of the box plots (except for smoking) show a right skewed distribution as the **average (mean) is much higher than the median**. In this case, using the **median** is more appropriate. For standardization purposes, the median is chosen to be shown instead for all plots.

smartcademy

Exploratory Data Analysis (2/2) | Discrete Variables

(3) Smoking v.s. Premium Charges

- Smokers are charged a higher premium

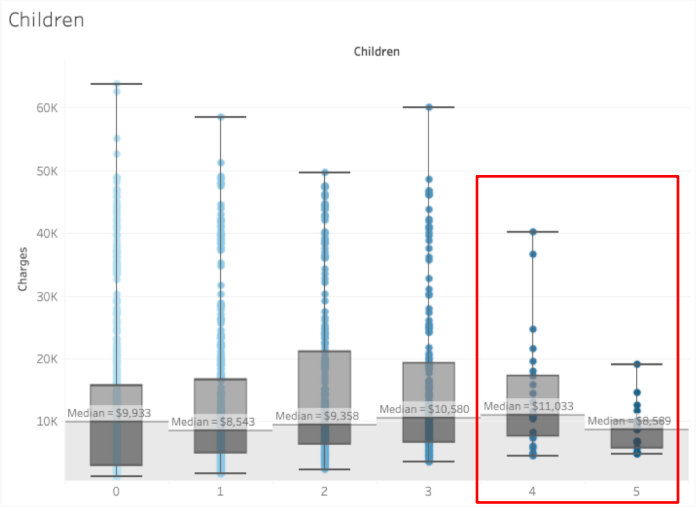
Box plot of smoking and premium charge

smartcademy

Exploratory Data Analysis (2/2) | Discrete Variables

(4) Children v.s. Premium Charges

- Increasing trend in premium charges up to 3 or 4 children but decreases for the 5th child.
- Most insurance plans charge for a maximum of 3 children (below 20 years old) on a family plan. This means that a family of 4 children pay around the same amount as a family of 3 children.
- Financial assistance is also available for larger families to help premium charges in the form of “premium tax credits”. This could explain why families with 5 children or more are charged lesser.
- However, the sample size for 4 – 5 children is also significantly lesser as compared to 0 – 3 children. The collected data could be unrepresentative of the actual premium charges.



Sample Size

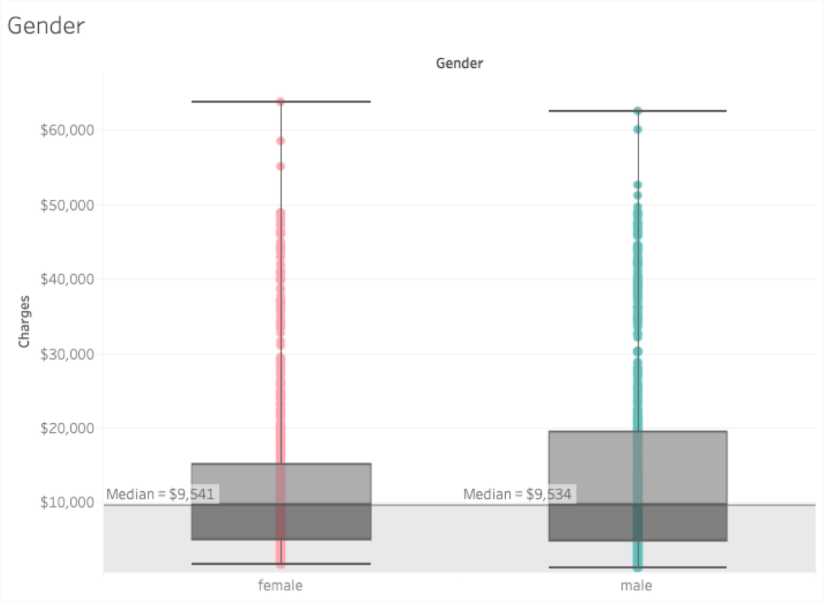
Children	
0	573
1	324
2	240
3	157
4	25
5	18

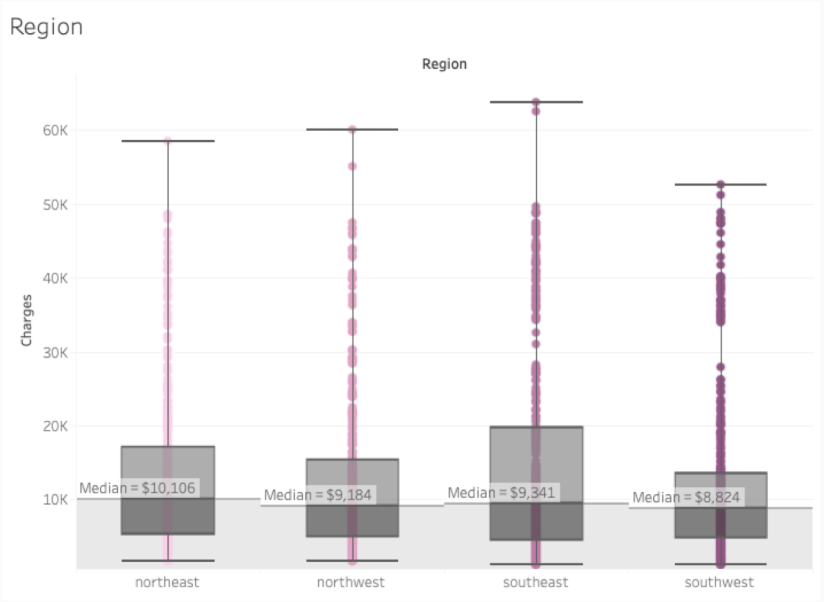
smartcademy

Exploratory Data Analysis (2/2) | Discrete Variables

(5) Gender and (6) Region v.s. Premium Charges

- No significant effect on premium charge





smartcademy

Dashboard

https://public.tableau.com/views/InsurancePremiumData_16561529673790/InsurancePremiumData?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Strategy: Key Recommendations

- **Who is charged a higher premium?**
  - Older individuals as chances of health problems increases with age
  - Smokers as they are more likely to develop health concerns in the future
  - Individuals with 3 – 4 children as their insurance plan includes extra coverage

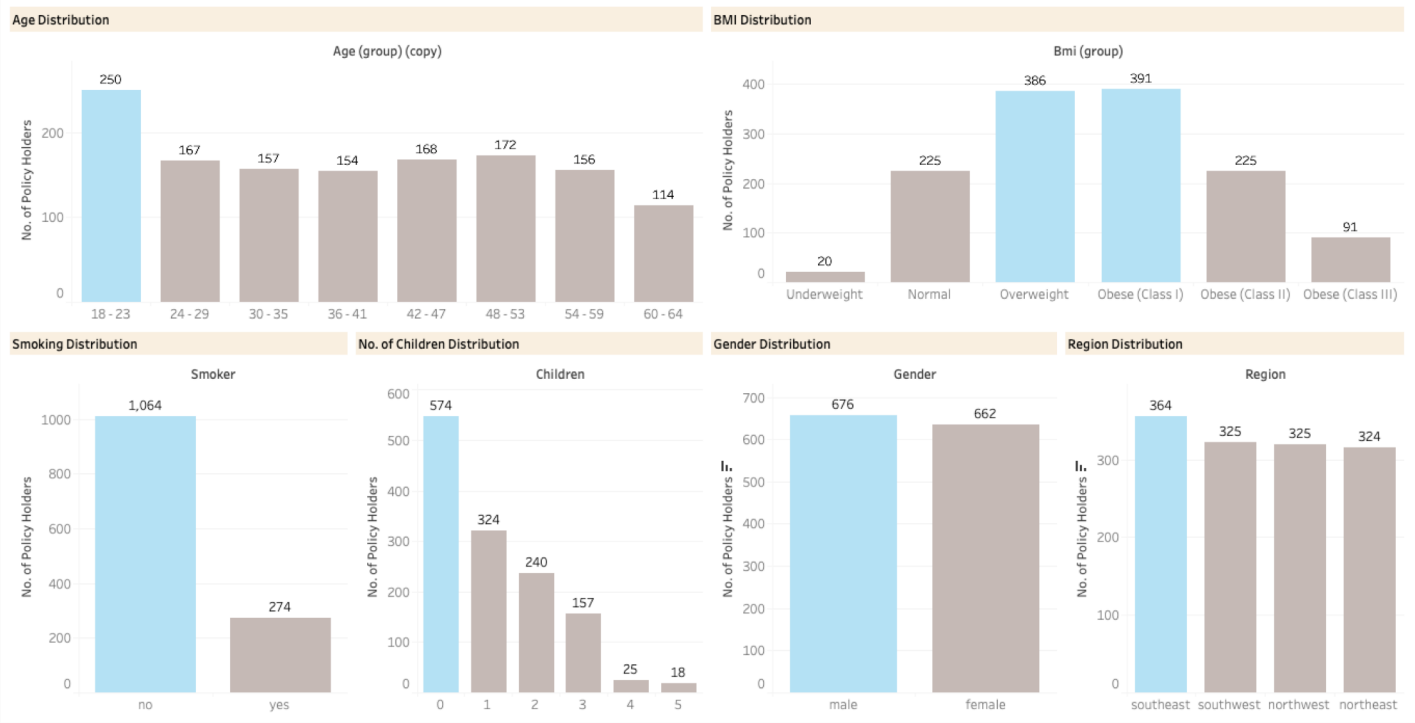
- **Who is charged a lower premium?**
  - Younger individuals due to lower health risks
  - Non smokers as they are less prone to health risks
  - Individuals with less than 2 or no children. Individuals with large families may also receive some form of financial aid to help pay part of their premium charges

- Weak correlation observed between BMI and premium charges. Individuals could have been charged lower/higher premiums based on other factors not revealed in the data set

- Gender and Region do not have a significant impact on premium charge

Strategy: Key Recommendations

- **Who are the main customers?**



- **18 - 23 years old**
- **Overweight and obese**
- Non-smokers
- **No children**
- No clear distinction for gender and region

smartcademy

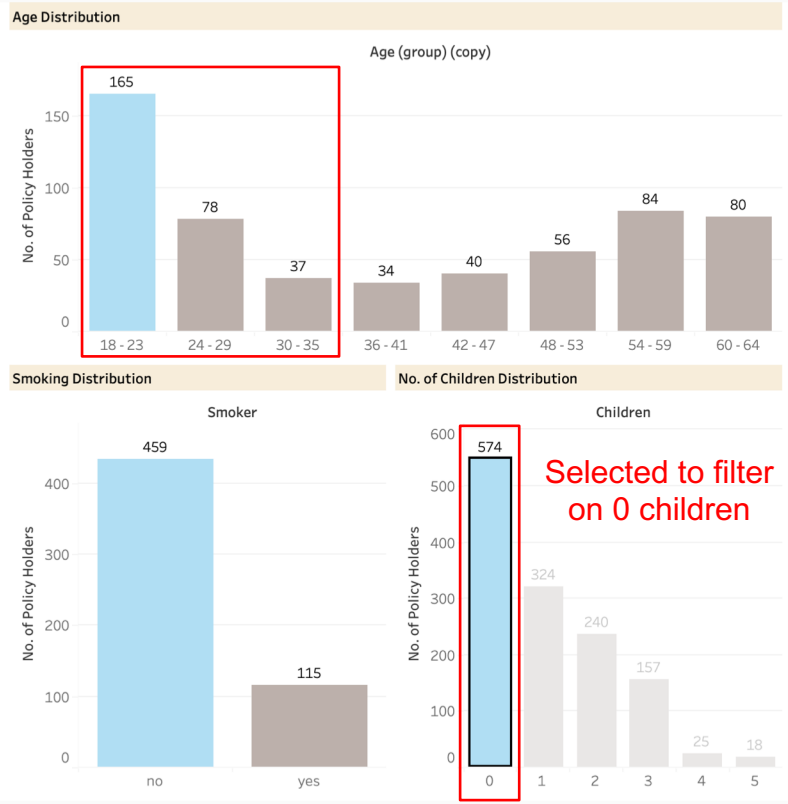
Strategy: Key Recommendations

- **Recommendation for main customers:**
  - Main age group of 18 – 23 years old suggests that individuals are buying health insurance plans at a younger age. This could be because the younger generation is more financially literate and understand the benefits of buying insurance early. It is generally understood that when you are younger, you are less likely to have pre-existing health conditions. This allows you to "lock-in" your premium at a lower price. Any health conditions that arise before your policy is active will be excluded from the plan coverage or incur higher premium charges.
    - Target potential new niche market of Gen Z/university graduates

  - Bulk of customers are overweight or obese. This could indicate that these customers are at a higher risk of health complications and are more likely to make more policy claims. However, BMI should not be the only health indicator to determine premium charge, so more information about individual or family medical history has to be properly documented and considered along with BMI to determine the appropriate premium charge
    - Ensure that all client profiles are up to date with medical history and background. Review past claims history for overweight/obese clients to determine if more weightage should be placed on BMI to determine premium charge

Strategy: Key Recommendations

- **Recommendation for main customers:**
  - Most customers do not have children and majority are still within the 18 – 23 age group. This could be because they are not married or ready to start a family.
    - To monitor this group of customers closely as some could potentially get married and have children. Conduct yearly/bi-yearly policy reviews with them to reassess their needs for life insurance or health insurance for their family members.
    - This can also be extended to customers within the 24 – 29 or 30 – 35 age group.



smartcademy

Conclusion

- Older, smoking individuals with 3 – 4 children are charged higher premiums because they are at higher risk of health complications and also require a higher coverage
- Main customers are 18 – 23 years old who do not smoke and have no children but are in the overweight/obese weight class.
- Data set is too small with only 1338 entries and no information is provided on sampling method which could be biased
- Data is lacking other factors that could significantly affect premium charge, for example: pre-existing medical conditions, family medical history and choice of profession.
- It should also include coverage type as more expensive premiums tend to offer more comprehensive coverage. This will allow the data set to be segmented further for more accurate analysis
- Deeper statistical analysis like hypothesis testing and multi linear regression can be conducted to quantify the relative contribution of each independent factor against premium charges

THANK YOU

smartcademy