**Observations and Recommendations**

**Data Exploration:**

* All feature names and data types are correct.
* The size of the data is 1337 \* 7.
* The data set has 4 numerical features and 3 categorical features.
* The dataset doesn’t contain any missing values

**Exploratory Data Analysis:**

* The BMI and charges features have outliers.
* Outliers are treated by using 1.5 IQR.
* If we are treating outliers in charges with 1.5 IQR around 300 data points will be removed.
* If we apply boxcox transformation, the data values will be changed.
* As charges is the target variable, outliers were not treated because, the data will be transformed in box cox.

**The Observations of EDA:**

1. This data set includes all age groups. As age is increasing the medical expenses are also increasing.
2. The patient who has BMI more than 25.0 and less than 50.0 has more medical expenses in the data set. But technically those who are having 50 BMI can be considered as higher risk patients. But in our data the patients those who have 50 BMI might be athletes with fit muscle weight or there may be any reason, those doesn’t have much medical expenses. This could be the reason, we are finding outliers in this feature. The patients who have BMI less than 18.5 also have higher medical expenses because of underweight.
3. The dependents of patient, which is categorised by “children” feature doesn’t impact the target variable “charges”. The patients who do not have children have more medical expenses than the patient who have 5 children. So, it doesn’t have any correlation with charges.
4. Charges is our target variable of our dataset. This feature has outliers. If we apply any outlier treatment technique the data will get impacted. So the outlier treatment was not done on charges.
5. Gender(sex) has an impact on charges. The number of males and females are almost same but female patients have higher medical expenses than male.
6. If we compare the health costs of minimum and maximum values of smoker and non-smokers, the minimum value of the smoker is starting at almost closer to the maximum value of non-smokers. By this we can estimate that smokers will have higher medical expenses than the non-smokers. Even though non-smokers have outliers, their medical charges are not too expensive comparative to the smoker.
7. The southeast region has more medical expenses comparatively than other regions. The remaining three regions have almost similar medical expenses. The higher age with high BMI patients are from southeast region who have higher medical expenses. It indicates that southeast region patients tend to have more health issues. Further, southeast region has more smoker patients than the other regions.

**Recommendations:**

1. This data set includes information on patients from all age groups and their BMI provides valuable insights that health care insurance company can consider as an important features for assessing risk profiles and determining appropriate coverage premiums. As age is increasing the BMI is also increasing parallelly.
2. By analysing BMI information, insurer can better assess the health risks and provide customized coverage premiums. Insurer should consider BMI less than 18.5 as high-risk category as they do have higher medical expenses.
3. Insurer can introduce new strategies to fix premium for female members of health insurance.
4. Insurer can introduce interested premiums in southeast region which has more high-risk patients at higher age, higher count of smokers and higher BMI.
5. Instead of charging higher premiums to non-smokers who are more in count, Insurer should concentrate on smokers who are considered as higher risk patients as those who have more medical expenses even though the count of smokers is less.