Exploratory Data Analysis With Statistical Analysis "MaxGet Insurance"

EDA

- Data Scientists are focused to normalize and analyze data statistically, categorically using plots to find potential benefits for an organization.
- Exploratory Data Analysis popularly known as EDA is a process of performing some initial investigations on the dataset to discover the structure and the content of the given dataset.
- It is often known as Data Profiling. It is an unavoidable step in the entire journey of data analysis right from the business understanding part to the deployment of the models created.
- EDA is where we get the basic understanding of the data in hand which then helps us in the further process of Data Cleaning & Data Preparation.
- Analyze Database in visualize way that it shows the correlation and impact on each other.
- Understand Database/variables. Find missing values.
- Find outliers to get better analytics results.
- Find what are the different variables in database that influence the ask.
- Find what are the factors affect the ask.
- Fine what could be plausible reason for that.
- Find how the data is distributed.

MaxGet Insurance Background & Goal

- MaxGet is an Insurance Company.
- Planning to launch new plan for their existing and new prospects.
- Before they launch their new plan, they would like to understand statistically, what will be performance of the New Plan.

MaxGet Insurance Company's Goal

- Investigate what would be performance of New Plan based on existing beneficiary database

Data Scientist Approach

- We are looking into and analyzing beneficiaries' information like gender, children, region, smokers etc.
- That means, our focus will be on in comparison of
  - Beneficiaries' Gender vs Claims made them
  - Beneficiaries are Smoker/Non-Smoker vs Claims made by them
  - Beneficiaries' Children counts vs Claims made by them
  - Beneficiaries' BMI vs Claims made them
  - Beneficiaries' Region, Smoking Habits, BMI impact of Gender etc

Data Scientist ASSUMPTIONs

- Considering we don't have direct interaction with Business team or Product Managers, so based on what we understand, we have built this analysis.
    - The Health Insurance Customer's data is a simple random sample from the population data.
    - Age wise dataset excluding those who are above 64 considering they are covered by Government
    - Gender wise excluding those who are not in Male or Female categories
    - BMI = Square (Individual's Weight (in kilogram) / Individual's Height (in sq meter).
    - An ideal BMI is within range of 18.5 to 24.9.

Data Information

Variable	Description
AGE	This is an integer indicating the age of beneficiary
GENDER	This is Policy holder's gender, male or female
BMI	This is the Body Mass Index which provides a sense of how over or under weight
KIDS	This is an integer indicating the number of children / dependents covered by the insurance plan
SMOKER	This is yes or no depending on whether the insured regularly smokes tobacco.
REGION	This is the beneficiary's place of residence in the U.S. divided into four geographic regions like North, South, East, West.
CLAIMS	Individual Medical cost billed by health insurance.

Data Information

;1338 Observations with 7 Variables (columns).

4 Integers and 3 Object Variables with storage of 73.2+kb.

Converting 3 Object to Categorical variable, storage reduced to 46.2kb.

Gender, Smoker, Region are are Categorical variables.

Age, BMI, Kids, Claims are are integer variables.

Kids is integer variable but for detailed study, converted into Categorical during the analysis.

Data Information

Observations On Data

- age
  - Beneficiaries age ranging from 18 to 64 with an average of 39.
  - Shows all population as adult considering all 18 and above.

- bmi
  - Beneficiaries BMI ranges from 16 to 53.
  - Up to 25% beneficiaries are within healthy BMI range of 18 to 25 (tentative/close by).
  - Nearly 26% to 75% beneficiaries BMI ranges OUTSIDE of healthy BMI zone.
  - Considering Max BMI is 53 and up to 75% are within 35 BMI,
    - Few beneficiaries are in danger zone of BMI (health), obese.

- kids
  - Beneficiaries number of kids ranging from none (0) to 5.
  - Up to 25% beneficiaries do not have Kids for sure.
  - Up to 50% beneficiaries have 0 or 1 kid.
  - Up to 75%, 2 or less kids; however, there are beneficiaries with 4 and 5 kids as well.

- claims
  - Claims are highly skewed positive/right side where MEAN 13270.
  - Median (less than MEAN) is just 9382 whereas Max is 63770.
  - That means most people would require basic Medicare and only few suffer from diseases which cost more.

Data Information

- gender
  - Gender data indicates there are two unique genders in beneficiaries, almost equally distributed.
  - Out of 1338 beneficiaries, 676 are MALE and remailing 662 are FEMALEs.

- smoker
  - 1064 beneficiaries do not smoke, which is nearly 80% population.

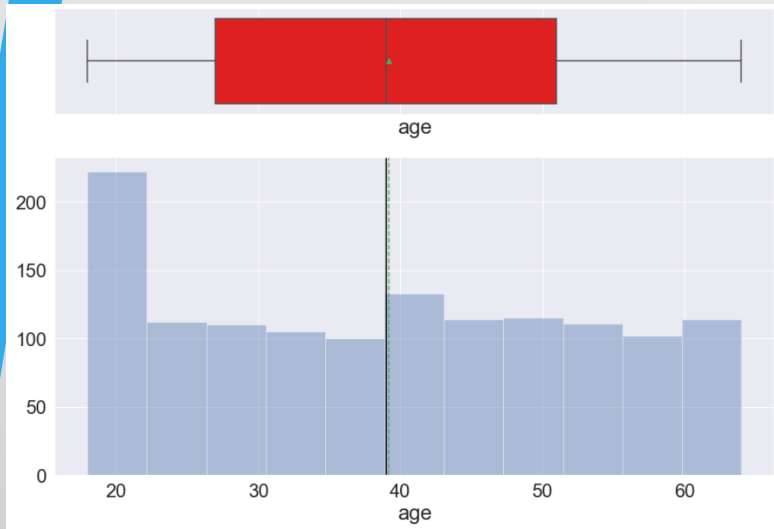
- region
  - Beneficiaries are spread into FOUR categories.
  - South is region of major beneficiaries' population.

Python Libraries Used

- numpy
  - Numerical Python (numpy) is used for working with multi-dimensional arrays.
- Pandas
  - Pandas is used for data manipulation and analysis. It offers structures for data manipulations.
- matplotlib.pyplot
  - Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
  - matplotlib.pyplot is a state-based interface to matplotlib. Pyplot is mainly intended for interactive plots.
- %matplotlib inline
  - Displays output inline. IPython kernel has the ability to display plots by executing code.
- Seaborn –
  - Seaborn is a Python data visualization library built on top of Matplotlib.
  - Seaborn contains a number of patterns and plots for data visualization.
- scipy.stats
  - Used for Scientific and Technical Computing
- sklearn.preprocessing / LabelEncoder
  - As Label Encoding in Python is part of data preprocessing, hence we will take an help of preprocessing module from sklearn package
- Copy
  - A copy is sometimes needed so one can change one copy without changing the other.
  - In Python, there are two ways to create copies : In order to make these copy, we use copy module.
  - We use copy module for shallow and deep copy operations.

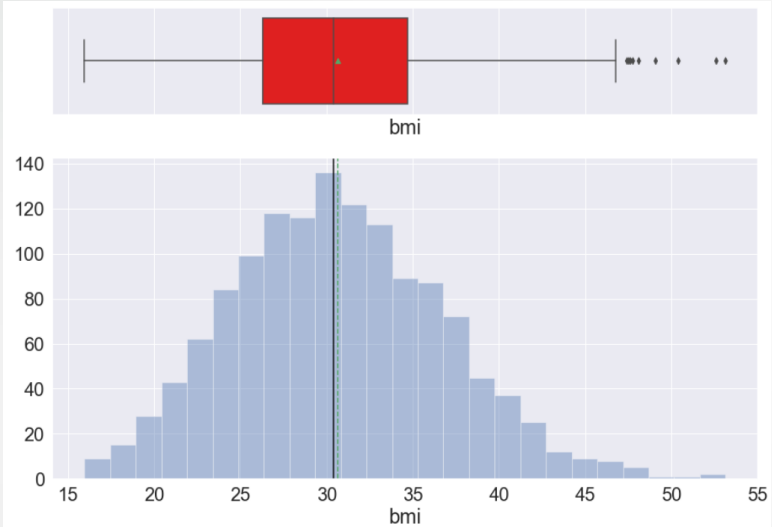
Univariate Analysis
MaxGet Insurance

EDA – age, bmi and kids



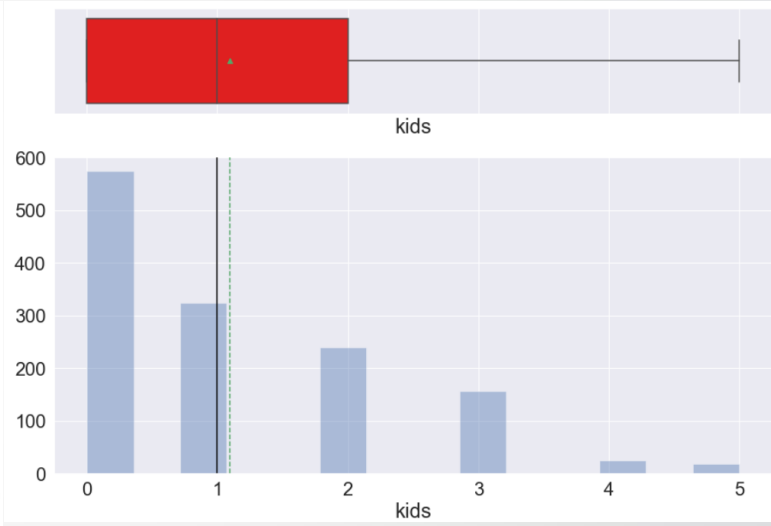
Beneficiaries – Age

- Age seems uniformly distributed, with both MEAN and MEDIAN around 40.
- 50% beneficiaries are less than 40 years of age. There are no outliers in AGE of beneficiaries.
- Complete dataset apparently of Adult population.



Beneficiaries – BMI

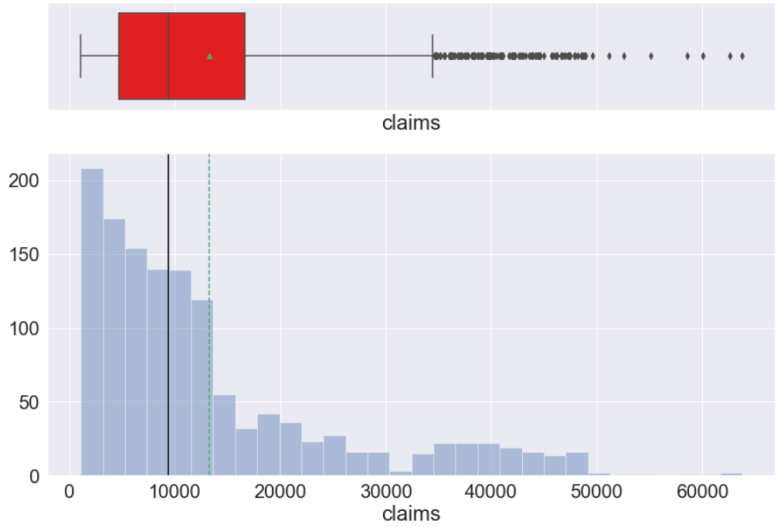
- BMI Looks fairly normal distribution. Although, there are few Outliers.
- Fitness perspective,
  - Looks like nearly 800 beneficiaries in the range of bmi 25 to 35. BMI above 25 is not healthy sign.
  - There are few beneficiaries above 35 BMI as well which means they are very much in Danger with respect to their obese-ness.



Beneficiaries –Kids

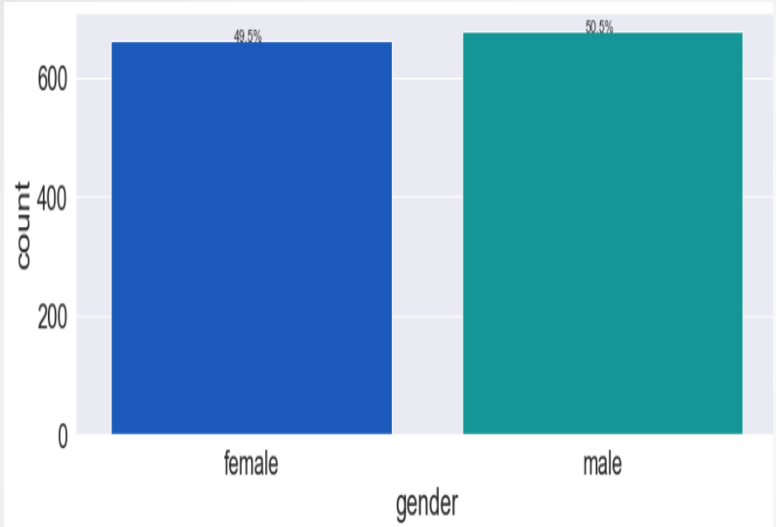
- Number of Kids has Right skewed distribution considering MEAN > MEDIAN.
- Plot suggests nearly 600 beneficiaries have 0 kid, 300+ have 1, 200+ have 2, 150+ have 3, very few have 4 and 5 kids.
- The plot suggest we should convert kids to Categorical Variable for further analysis.

EDA – Claims, Gender and Kids



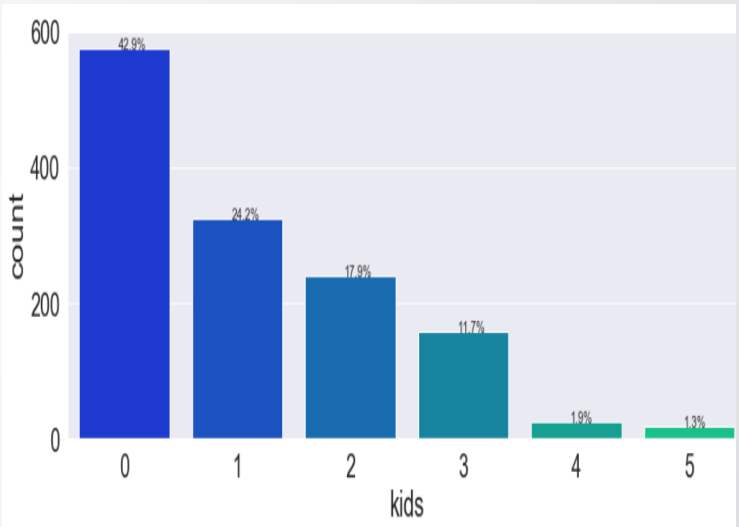
Beneficiaries – Claims

- Claims data are highly skewed, right side considering MEAN > MEDIAN.
- Nearly 700 beneficiaries Claims less then / equal to MEDIAN value; that mean (disciplined/controlled) claims.
- It's showing lots of Outliers towards higher end indicating that some people spend higher for their medicals.



Beneficiaries – gender

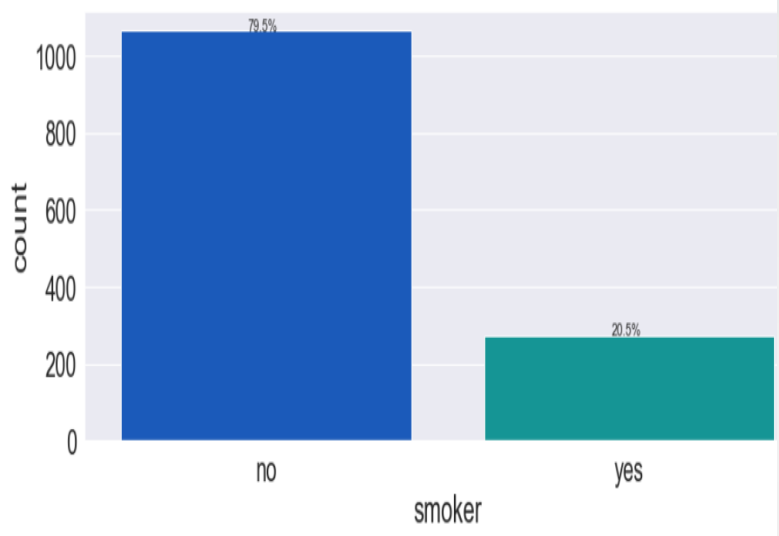
- The distribution of observation across genders is fairly same as we saw earlier and, in this plot, as well.



Beneficiaries – Kids

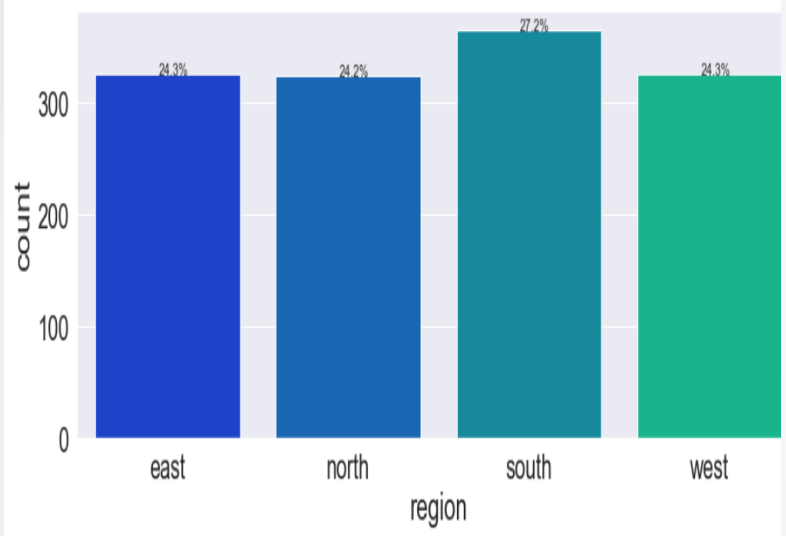
- Kids column converted into categorical for detailed analysis.
- Nearly 43% beneficiaries do not have any kids.
- Nearly 24% beneficiaries do have 1 kid.
- Nearly 18% beneficiaries have 2 & 12% have 3 kids.
- Less than 2% each beneficiaries 4 or 5 kids.

EDA – Smoker, Region



Beneficiaries – Smoker

- As mentioned earlier, nearly 80% beneficiaries are Non-Smoker. 20% are smoker.
- Further analysis can be done to see how this 20% smoker affects the insurance claims.



Beneficiaries – Region

- South is favorite region; maximum beneficiaries are from this area - 27+%
- Other three region have 24+% beneficiaries.
- Overall fair distribution but nearly 3% difference with South is statistically significant. We will explore that later.

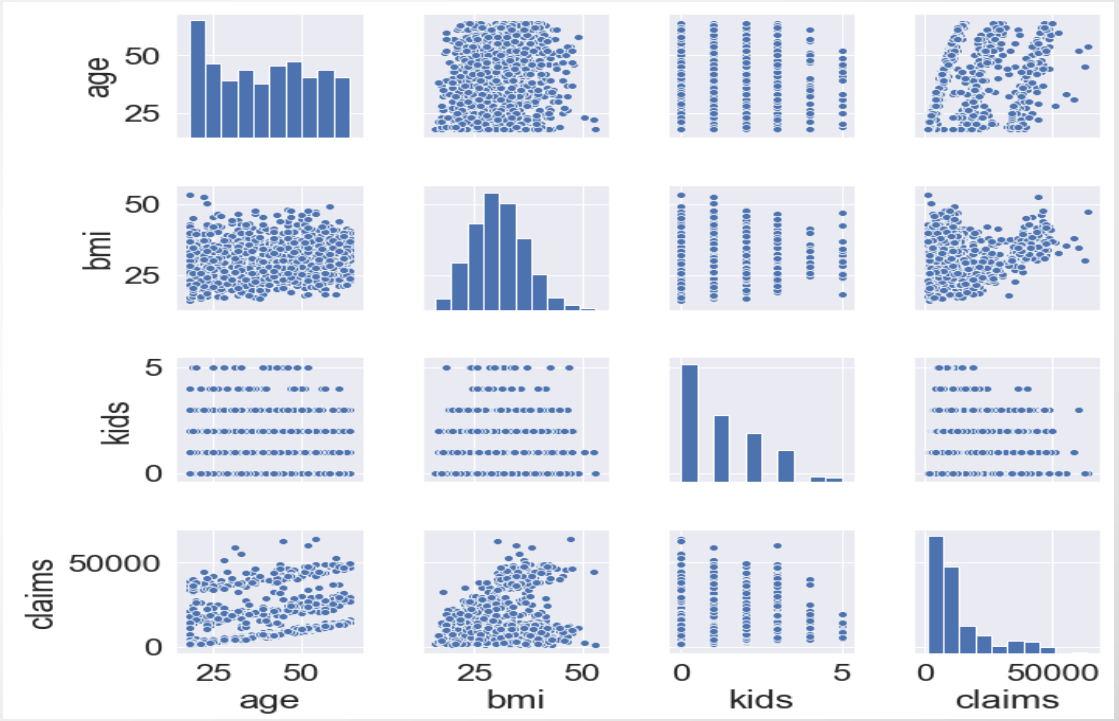
Bivariate Analysis
GetMax Insurance

Bivariate Analysis - Heatmap

- Correlation between all the variables is Positive but not so high that we don't fine much impact.
- The Correlation value is 1 means there is its linear trend between the two variables.
- When the Correlation value is close to 1 the correlation is the more positive.
  - Means one variable increases so does other.
- **Age/BMI is 0.11 which is most close to 1. So, we may see some impact here.**
- The Correlation value is close to ZERO means there is no linear trend between the two variables.
- When the Correlation value is closer to -1, the correlation is negative.
  - Means, instead of both increasing one variable will decrease as the other increases.

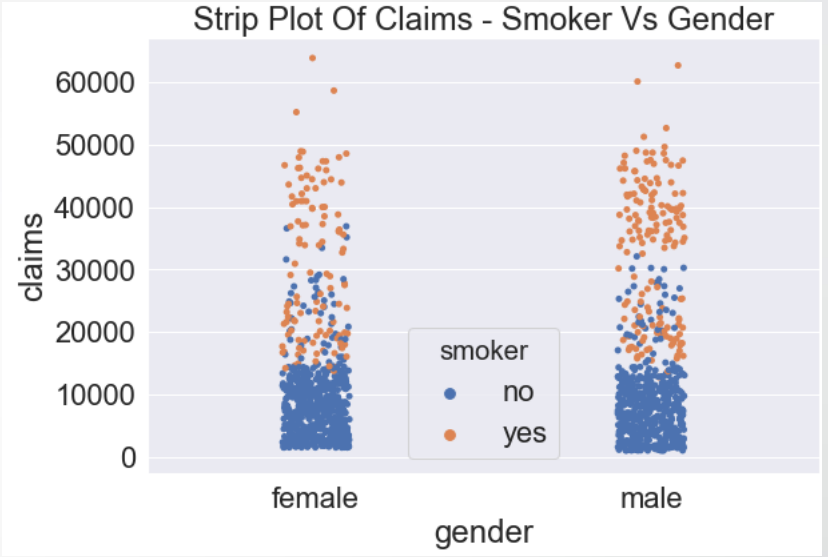
Bivariate Analysis - Pairplot

- Age/Claims -
  - There is an interesting pattern between Age and Claims.
  - Seems slightly, as Age increase, Claims value is increasing.
  - It is possible that for the same ailment, older people are charged more than the younger ones.
- BMI/Claims -
  - As per observation, more the BMI, higher the Claims. BMI above 25 having more Claims and higher Claims as well.
- Kids/Claims -
  - Strange that beneficiaries with zero kids have claims up to Max claim price. And
  - Beneficiaries with 4 and 5 kids claims are limited.
- Age/Kids & BMI/Kids –
  - Not much impact to observe.
- Age/BMI –
  - It's observed that even at Young age, BMI showing very high like obsessed for few.

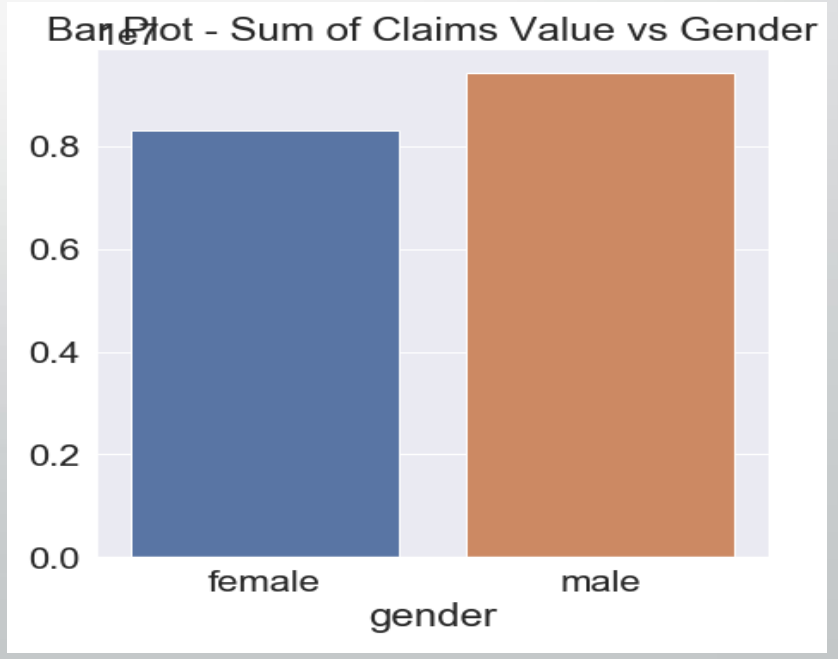


Bivariate Analysis – Strip & Bar Plot

- Strip Plot of Claims – Smoker vs Gender
  - Strip Plot clearly shows that Smoker has strong impact on Claims Value but there is not impact due to Gender.



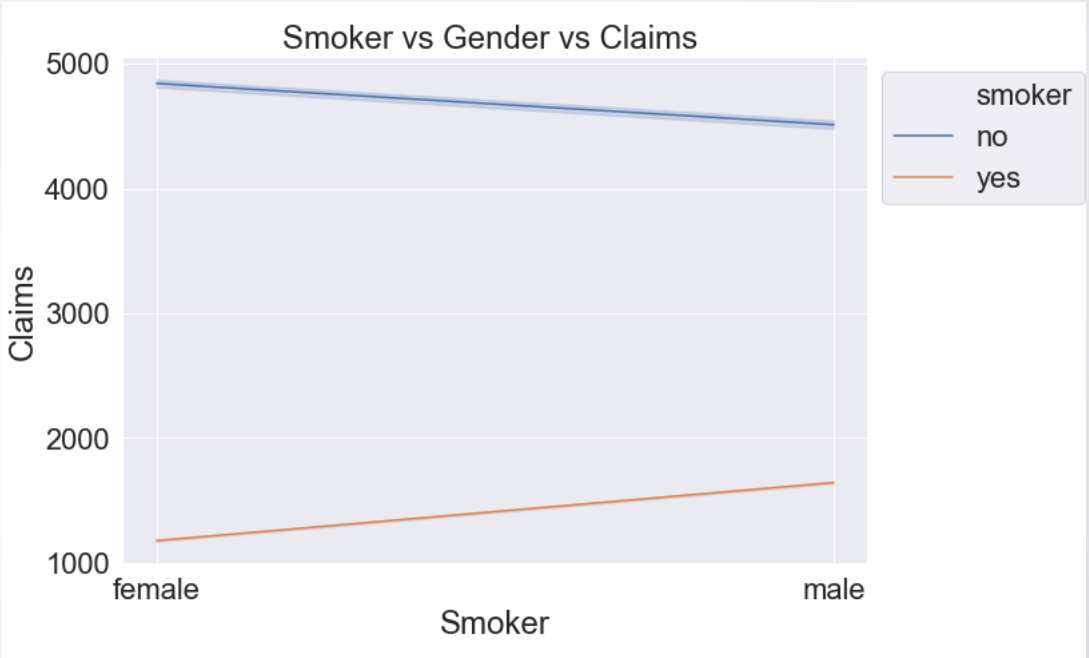
- BAR Plot **of Sum Of Claims** by Gender
  - Sum Of Claims are apparent that MALEs have higher claims in Total compared to Females.



Multivariate Analysis Financial Institute

Multivariate Analysis

- Line Plot of Sum Of Claims vs Smoker and Gender :
  - Line plot of Smoker Female vs Smoker Male vs Claims clearly shows
  - Male or Female, when not smoking, claims lessor compared to those who are Smoking.

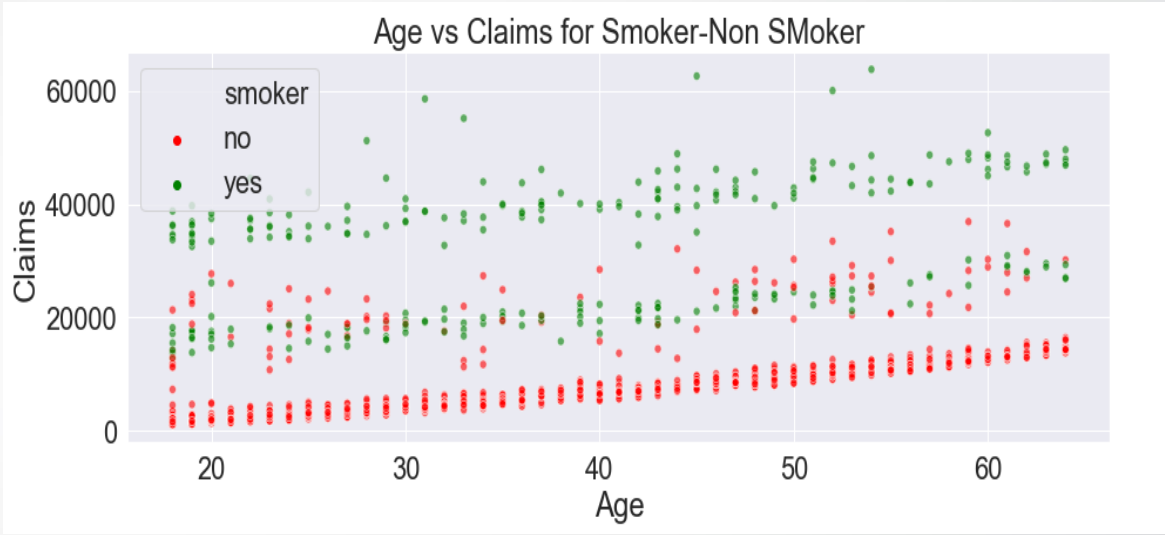


Statistical Analysis

:2:

Prove (disprove) that
Medical Claims made by the people who smoke is greater than those who don't.

- Visually the difference between charges of Smokers and Non-Smokers is apparent.
- The non-smokers have much lower claims compared to the smokers.



- T-Test : While doing the T-Test it is found that
  - Reject the NULL Hypothesis that the Mean Claims of Smokers Is Less Than or Equal To Non-Smokers.
  - So, it is proved that Medical Claims made by the people who smoke is greater than those who don't.

- T-Test Calculation :
  - Ho = "MEAN (Average) Claims of Smokers is Less Than or Equal To Non-Smokers"
  - Ha = "MEAN (Average) Claims of Smokers is Greater than Non-Smokers"
  - sx = np.array(df[df.smoker == "yes"].claims) *#Selecting Charges Corresponding To Smokers As An Array*
  - sy = np.array(df[df.smoker == "no"].claims) *# Selecting Charges Corresponding To Non-Smokers As An Array*
  - t, p_value = stats.ttest_ind(sx,sy) # Performing Individual T Test
  - print(t,p_value) ***# 4.135717921088609 3e-283***
  - print("Tstat :",t,"p_value :",p_value/2) # Since it is One Tailed Test
  - p_value < 0.05 ***# Reject the NULL Hypothesis that the Mean Claims of Smokers Is Less Than or Equal To Non-Smokers***

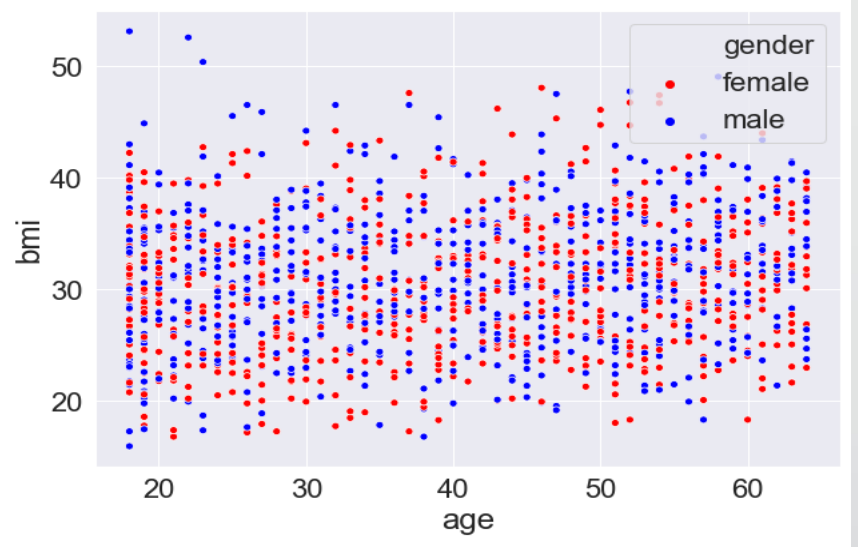
:3:

Prove (disapprove) with Statistical Evidence that the BMI of Females is different from that of Males.

- Mean of BMI of Female : 30.377749244713023
- Mean of BMI of Male : 30.943128698224832

- BMI of Females is 30.38, lessors than Males 30.94.
  - But it is not that MAJOR difference.

- Scatter plot shows,
  - There seems to be no apparent relationship between Gender and BMI.



- T-Test : While doing the T-Test it is found that
  - Failed to Reject the Null Hypothesis that the BMI of Females is same as that of BMI Males.
  - So, it is proved that BMI of Females is same as BMI of Males.

- T-Test Calculation :
  - Ho = "Mean BMI of Females is same as that of males"
  - Ha = "Mean BMI of Females is different from males"
  - x = np.array(df[df.gender == "male"].bmi) *# Selecting Male BMI Values as an array to x*
  - y = np.array(df[df.gender == "female"].bmi) *# Selecting Female BMI Values as an array to y*
  - t, p_value = stats.ttest_ind(x ,y) *# Performing an independent Test*
  - print("p_Value = ",p_value) ***# 0.08997637178984932***
  - p_value > 0.05: ***# Failed to Reject the Null Hypothesis that the BMI Females is same as that of BMI Males"***

:4:

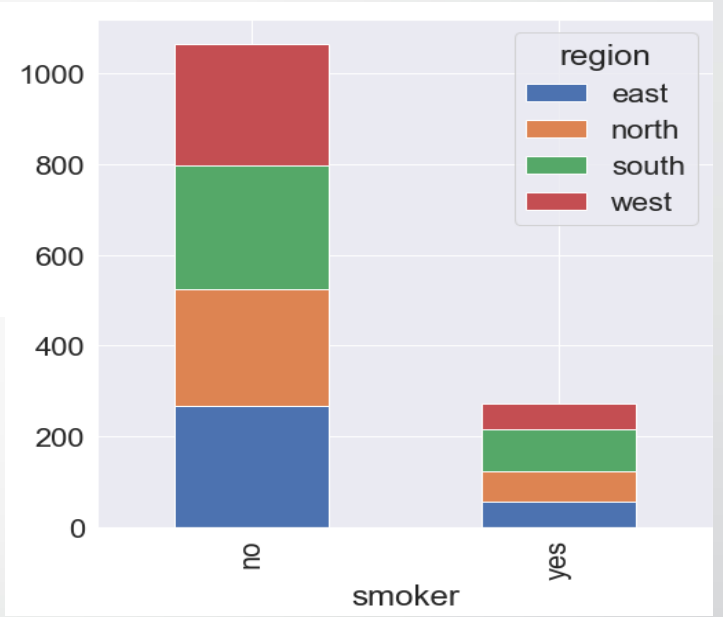
Is the proportion of smokers different across different regions?

- Apparently, only South region has slightly higher Smoker and Non-Smoker beneficiaries, 273.

- For other 3 regions looks close by numbers, 257/267/267.

- Chi-Sq Test :
- Check if Smoker vs Non-Smoker are different in different regions (kind of trend of regions)
  - As analyzed earlier, thru Crosstab, there is no statistical difference in proportion of Smokers across regions.
  - There is no trend line for Smoking habits in different regions; it all same/similar.

smoker / region	no	yes
east	267	58
north	257	67
south	273	91
west	267	58



- Test Calculation :
  - Ho = "Proportion Of Smokers Not Different Across Regions"
  - Ha = "Proportion of Smokers is Different Across Regions"
  - crosstab = pd.crosstab(df['smoker'],df['region']) *# Creating Contingency Table of Smoker vs Region*
  - chi, p_value, dof, expected = stats.chi2_contingency(crosstab)
  - print(p_value) ***# 0.06171954839170547***
  - p_value > 0.05: ***# Failed To Reject the Null Hypothesis that Proportion Of Smokers is Different Across Regions***

:5:

Is the MEAN BMI of Women with no children, once child and two children the same ?

- Females BMI With 0 Kids 30.36 // Females BMI With 1 Kid 30.05
- Females BMI With 2 Kids 30.65 // Females BMI With 3 Kids 30.44
- Females BMI With 4 Kids 31.94 // Females BMI With 5 Kids 30.62

- Females BMI with different kids (0,1,2,3,5) has no difference overall.
- Females BMI with 5 kids is slightly higher than others.

- Analysis of Variance Test (ANOVA) :
  - Statistical Analysis suggest that there is no impact on BMI due to number of kids.

- ANOVA Test Calculation :
  - Ho = "Females BMI With different kids is same/No of kids doesn't impact BMI"
  - Ha = "Females BMI with different kids is different/No of kids impact BMI"
  - zero = wd[wd.kids==0]['bmi'] *# Females BMI With 0 Kids **30.36***
  - one = wd[wd.kids==1]['bmi'] *# Females BMI With 1 Kid **30.05***
  - two = wd[wd.kids==2]['bmi'] *# Females BMI With 2 Kids **30.65***
  - f_stat, p_value = stats.f_oneway(zero,one,two)
  - print(p_value) ***# 0.7158579926754841***
  - p_value > 0.05:
    - ***Failed To Reject the Null Hypothesis that Females BMI With different kids is same as the p_value (0.716) > 0.05***

Conclusions

- At the end of Analysis of 1338 observations of MaxGet Insurance firm with 7 variables like Age, Gender, BMI, Kids, Claims, Region; overall conclusion is
  - No BIG impact due to Number of Kids and Region on Claims submitted by beneficiaries.
    - However, Higher BMI has Higher value claims in correlation at some scale.
  - Higher Age beneficiaries have high BMI in correlation at some scale.
  - All the beneficiaries are adult only, up to age of 64 starting from 18.
  - Major beneficiaries have no kids compare to 1,2,3,4,5 Kids beneficiaries.
  - Smokers are impacting a lot on Claims Value and counts.
  - Gender and BMI has no correlation, but Age and BMI has tiny correlation.
  - Females BMI overall has no impact of how much kids they have.
- Claims value is mostly impacted because of beneficiaries smoking habits.
  - Followed by higher BMI (obese) does creates an impact on Claims value.

Recommendations

- MaxGet Insurance company should focus on
  - Claims value to reduce heavily by helping beneficiaries controlling and quitting their Smoking habits
  - Claims value to reduce by helping beneficiaries controlling their BMI
    - How :
      - Creating awareness program on Smoking is Injurious to Health
      - Creating awareness programs on Weight Management.
      - Creating awareness programs on how Diet, Exercise helps controlling BMI.
      - Creating few Meditational programs, Running Events etc