Axis Insurance

Case Study

Sven Meydell

Objectives

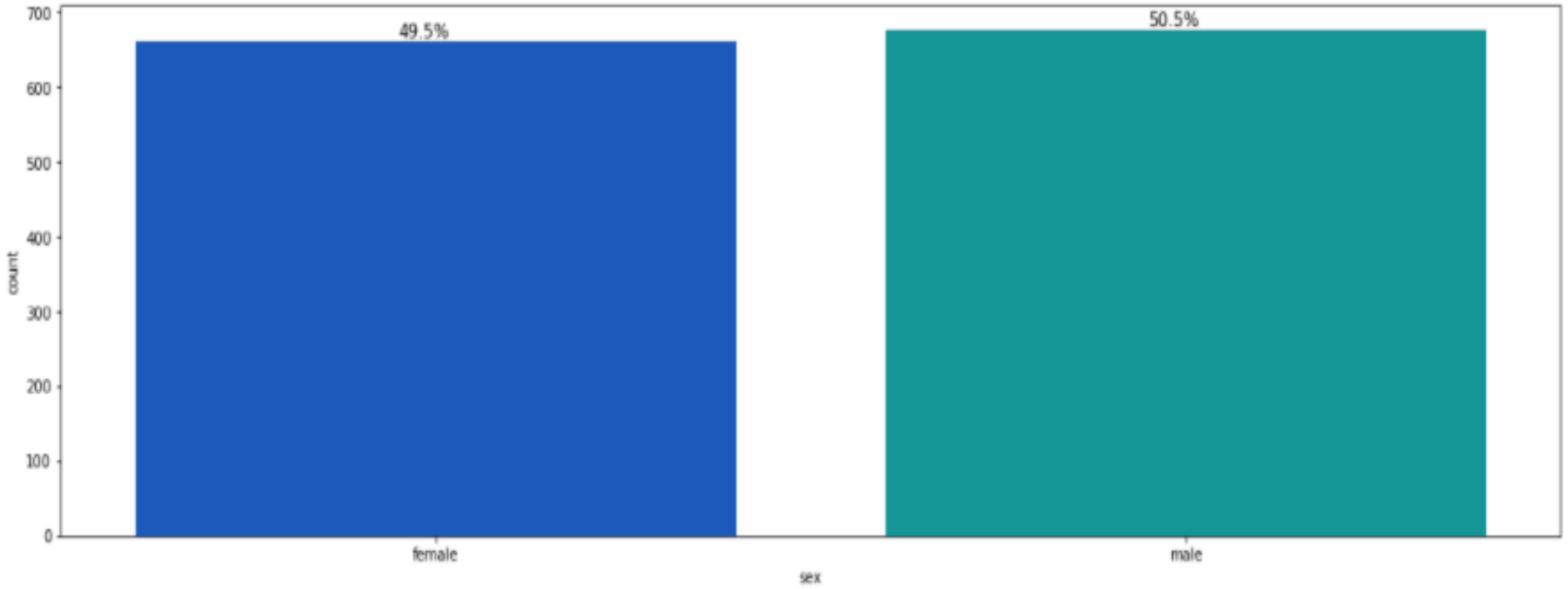
- Perform Exploratory Data Analysis on dataset
- Prove/disprove:
  - Medical claims made by smokers are greater than those made by non-smokers
  - BMI of females is different from that of males
- Determine if proportion of smokers is significantly different across different regions
- Determine (statistically) whether or not the BMIs are equal for females with:
  - No children
  - One child
  - Two children

(Level of Significance to be Used: 0.05)

Data Provided

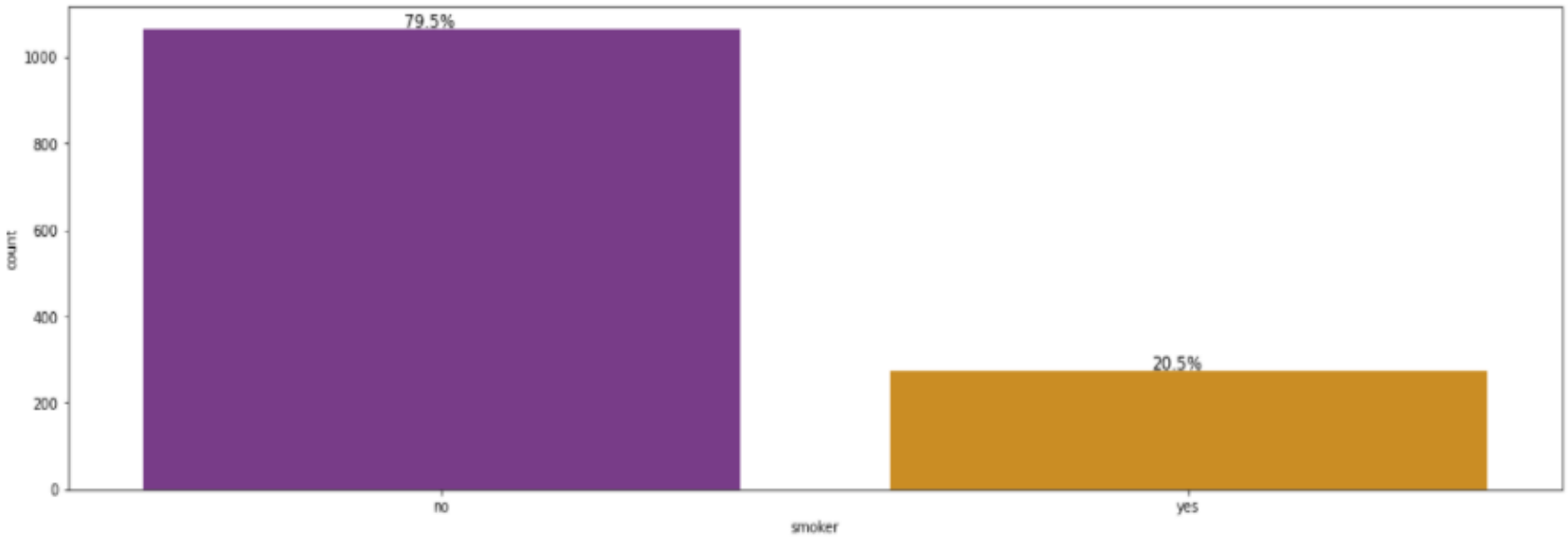
- **Age** - of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government)
- **Sex** – gender of beneficiary, (male/female)
- **BMI** - body mass index (how over or underweight a person is relative to their height)
  - An ideal BMI is within the range of 18.5 to 24.9
- **Children** - the number of children/dependents covered by the insurance plan
- **Smoker** - if the insured regularly smokes tobacco (yes/no)
- **Region** - place of residence (USA) of beneficiary
  - 4 geographic regions: Northeast, Southeast, Southwest, or Northwest
- **Charges** - medical costs (for the individual) billed to health insurance

EDA – Gender



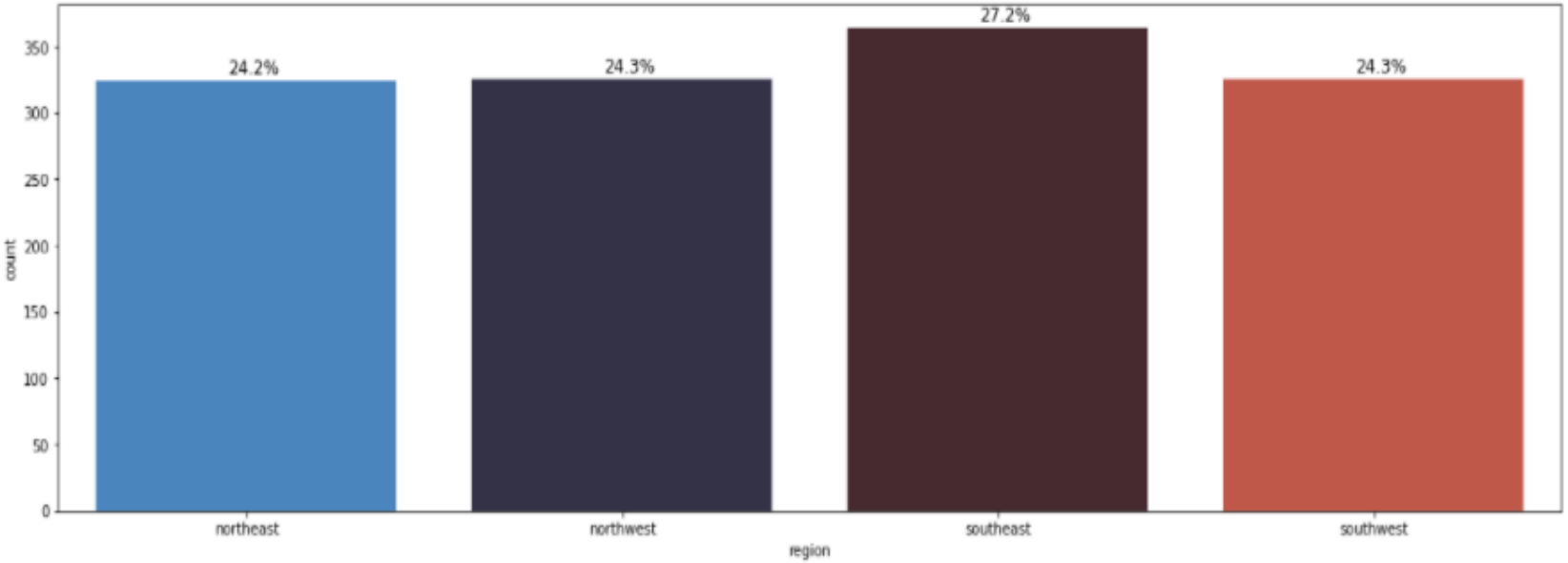
- The **gender split is roughly 50/50,** with slightly more males in the sample

EDA – Smoker/Non-Smoker



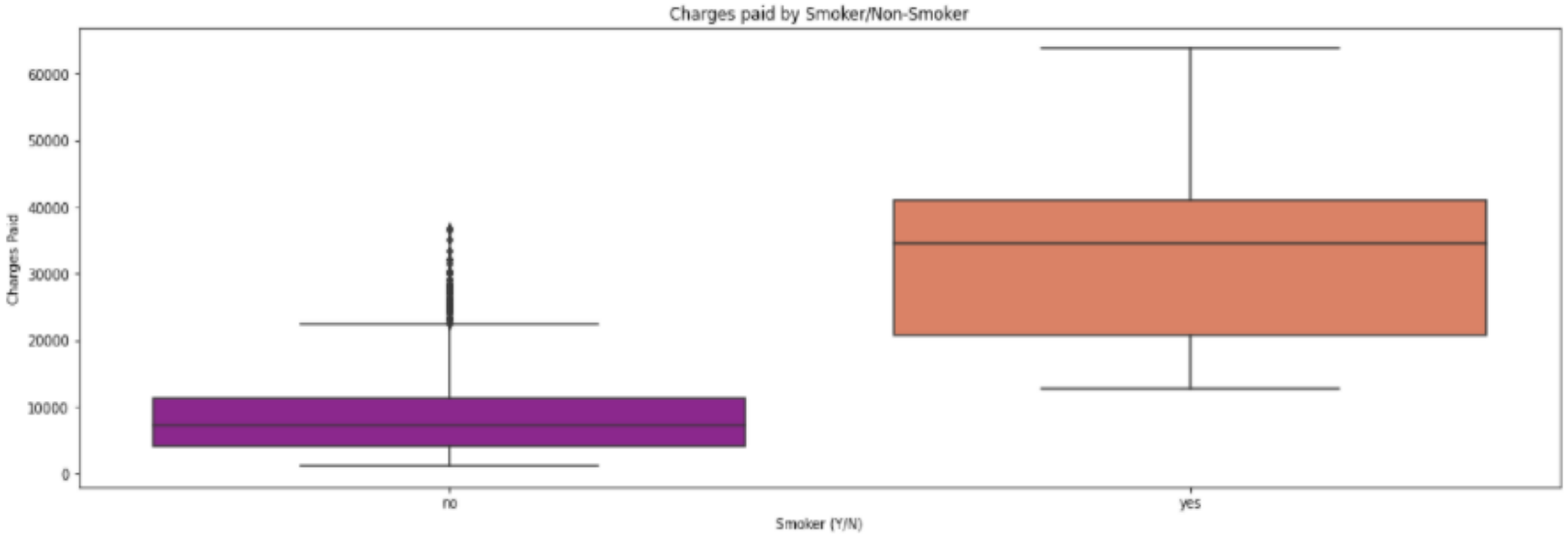
- The sample set is primarily composed of non-smokers
  - About **20%** of the beneficiaries are **smokers**

EDA – Regions



- The sample set is relatively **evenly split across the 4 regions**
- The Southeast shows up slightly more often than the other regions

EDA – Charges by Smoker/Non-Smoker



- On a total claim basis, **smokers only make up around 20% of the sample data but account for almost 50% of total charges**
- Average claims by smokers appear to be nearly 4x the amount of average claims made by non-smokers ($32k/$8.4k) and are nearly 2.5x the average claims within the sample ($32k/$13.3k)
- **Medical claims amounts appear to be disproportionally higher for smokers than by non-smokers**

Statistical Analysis: Charges by Smoker/Non-Smoker

Statistical Analysis - 0.05 Level of Significance

Split of Claims by Smokers vs. Non-Smokers

Stating the Null and Alternative Hypothesis

Let μ_1, μ_2 be the mean claims paid to beneficiaries who are **Smokers** vs. **Non-Smokers**.

Null Hypothesis

$$H_0 : \mu_1 <= \mu_2$$

Alternate hypothesis

$$H_a : \mu_1 > \mu_2$$

Results

Test: **2 Sample t-Test (Greater)**

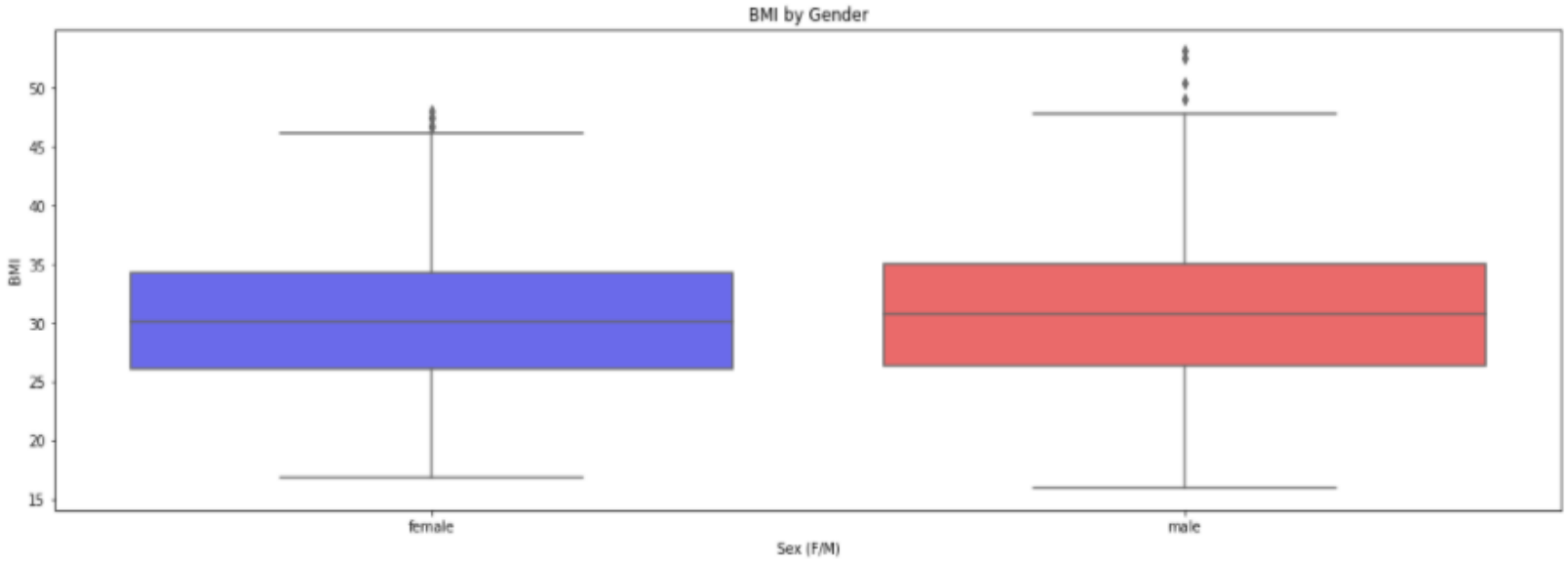
P-Value: 2.94473222335849e-103

Accept/Reject Null: Reject

Findings

- As the **p-value is significantly lower** than the significance level of 0.05, we therefore **reject the null hypothesis**
- There is **enough statistical significance** that medical claims made by smokers are more than those made by non-smokers

EDA – BMI by Gender



- Initial observations show the **BMIs for both genders are relatively equal**
- The means/standard deviations/variances are very similar between the genders

Statistical Analysis: BMI by Gender

Statistical Analysis - 0.05 Level of Significance

BMI of Females vs. Males

Stating the Null and Alternative Hypothesis

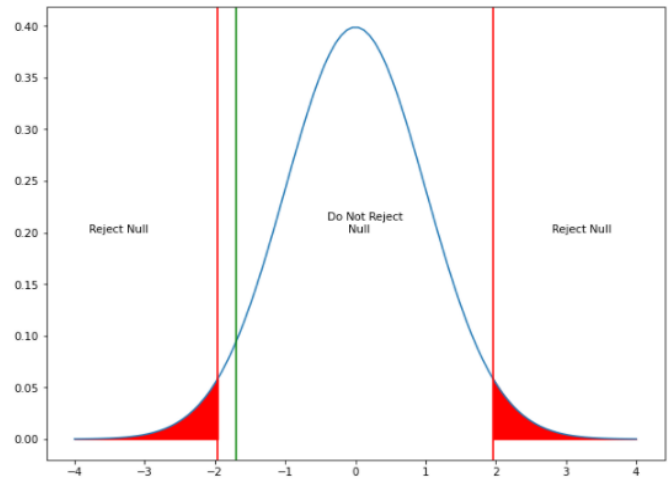
Let μ_1, μ_2 be the mean BMI scores of **Female** and **Male** beneficiaries.

Null Hypothesis

$$H_0 : \mu_1 = \mu_2$$

Alternate Hypothesis

$$H_a : \mu_1 \neq \mu_2$$



The Test Statistic (-1.7) **lies outside the rejection regions**, we fail to reject the null hypothesis

Results

Test: **2 Sample t-Test (Two-Sided)**

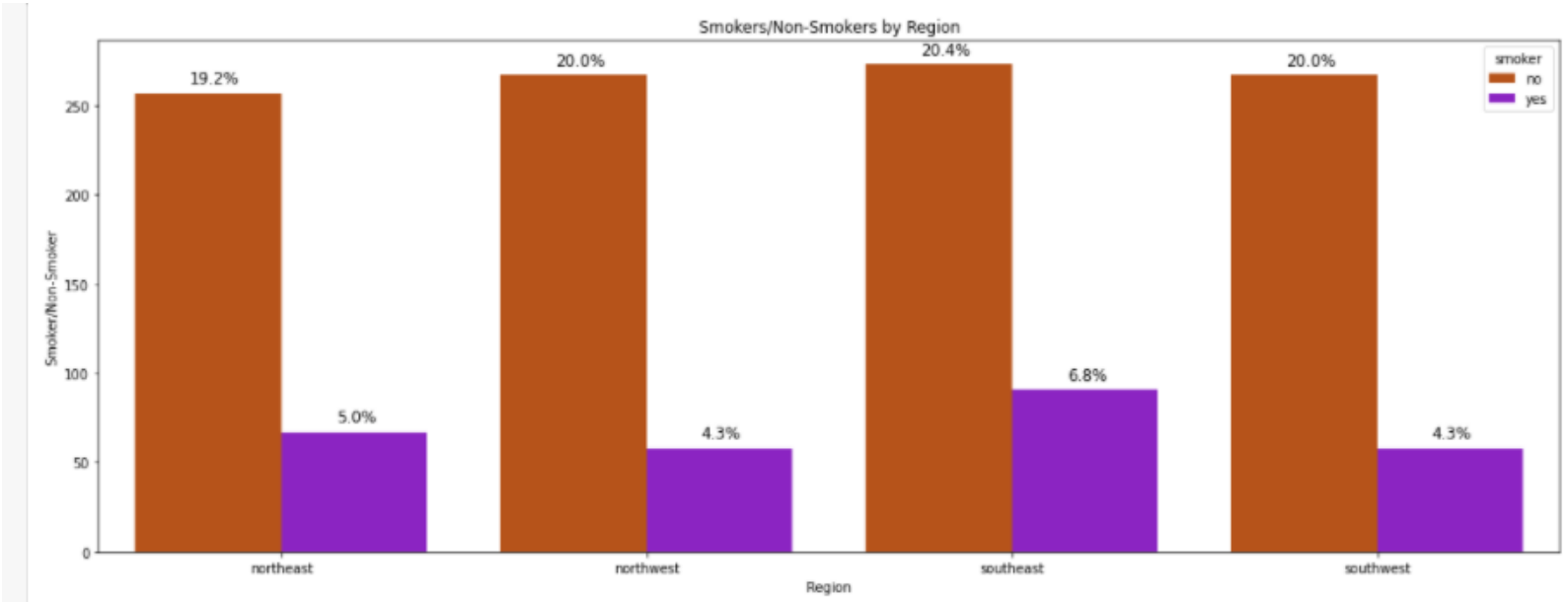
P-Value: 0.08992

Accept/Reject Null: Fail to Reject

Findings

- As the **p-value is significantly higher** than the significance level of 0.05, we therefore **fail to reject the null hypothesis**
- There is **not enough statistical significance** to state that the BMI of females is different to that of males

EDA – Smokers by Region



- The **Southeast appears to have a slightly higher proportion of smokers**, followed by the Northeast
- The proportion of non-smokers is relatively equal in all regions except the Northeast, which is slightly lower

Statistical Analysis: Smokers by Region

Statistical Analysis - 0.05 Level of Significance

Proportions of Smokers by Region

Stating the Null and Alternative Hypothesis

For Regions within the USA

Null Hypothesis

> H_0 : Smoking preference is independent of Region

Alternate Hypothesis

> H_a : Smoking preference is dependent on Region.

Results

Test: **Chi-Square Test (Independence)**

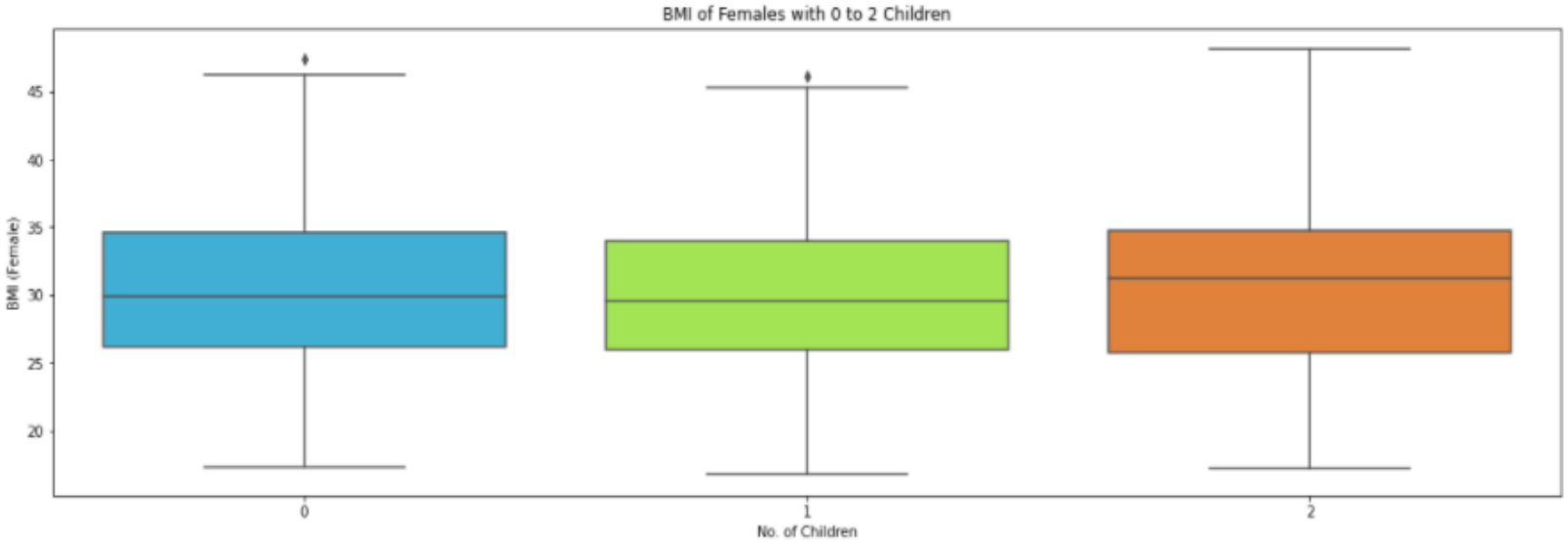
P-Value: 0.06171

Accept/Reject Null: Fail to Reject

Findings

- As the **p-value is significantly higher** than the significance level of 0.05, we therefore **fail to reject the null hypothesis**
- There is **not enough statistical significance** to state that smoking preference is dependent on region

EDA – BMI of Females (0-2 Children)



- **BMI scores are relatively equal amongst the three categories**
- The median BMI for females with 2 children is slightly higher than for those with 0 to 1 children

Statistical Analysis: (Shapiro-Wilk's Test) BMI of Females with 0-2 Children

Statistical Analysis - 0.05 Level of Significance

Shapiro-Wilk's Test

Null Hypothesis

> H_0 : Female BMI scores follow a normal distribution

Alternate Hypothesis

> H_a : Female BMI scores do not follow a normal distribution

Results

Test: **Shapiro-Wilk's Test**

P-Value: 0.00354

Accept/Reject Null: Reject

Findings

- As the **p-value is significantly higher** than the significance level of 0.05, we therefore **reject the null hypothesis** that the response follows the normal distribution

Statistical Analysis: (Levene's Test) BMI of Females with 0-2 Children

Statistical Analysis - 0.05 Level of Significance

Levene's Test

Null Hypothesis

> H_0: All the population variances are equal

Alternate Hypothesis

> H_a: At least one variance is different from the rest

Results

Test: **Levene's Test**

P-Value: 0.38994

Accept/Reject Null: Fail to Reject

Findings

- As the **p-value is significantly higher** than the significance level of 0.05, we therefore **fail to reject the null hypothesis** that the population variances are equal

Statistical Analysis: (ANOVA Test) BMI of Females with 0-2 Children

Statistical Analysis - 0.05 Level of Significance

ANOVA Test

Null Hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Alternate Hypothesis

H_a : At Least One of the Female Groupings (0/1/2 Children) is Different from the Rest

Results

Test: **ANOVA Test**

P-Value: 0.71585

Accept/Reject Null: Fail to Reject

Findings

- As the **p-value is significantly higher** than the significance level of 0.05, we therefore **fail to reject the null hypothesis** that all three of the female BMI groupings are equal

Statistical Analysis: (Tukey HSD) BMI of Females with 0-2 Children

Multiple Comparison Test (Tukey HSD)

Null Hypothesis

$$H_0 : \mu_1 = \mu_2 \text{ and } \mu_1 = \mu_3 \text{ and } \mu_2 = \mu_3$$

Alternate Hypothesis

$$H_a : \mu_1 \neq \mu_2 \text{ or } \mu_1 \neq \mu_3 \text{ or } \mu_2 \neq \mu_3$$

Results

Test: **Tukey HSD**

Multiple Comparison of Means - Tukey HSD, FWER=0.05

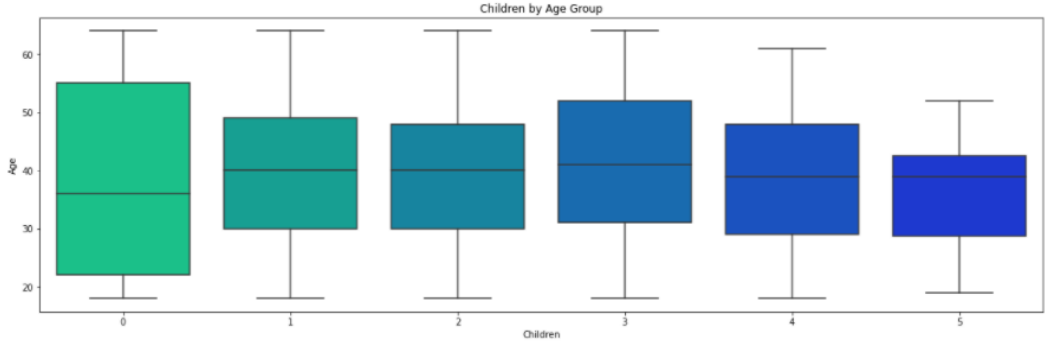
group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.3089	0.8494	-1.7186	1.1008	False
0	2	0.2883	0.8942	-1.2636	1.8402	False
1	2	0.5971	0.6797	-1.1323	2.3265	False

Findings

- As expected, based on findings from the ANOVA test above, the p-values (p-adj) are **higher than the level of significance of 0.05 for all three groupings**
- As expected, based on findings from the ANOVA test above, the p-values (p-adj) are **higher than the level of significance of 0.05 for all three groupings**
- The p-values are relatively similar for the BMIs of females with 0 and 1 child (0.85 and 0.89 respectively), but **substantially lower for females with 2 children (0.68)**

Recommendations

- The insurance company should plan for claims from smokers to be higher than non-smokers, both in frequency and overall claim amounts
- Although catering to families is very important, a large portion of the beneficiaries sampled had no children and could provide opportunities for growth
  - This same group sampled had the largest age range (mid-20s to mid-50s)



- There should be a focus on selling more insurance to younger individuals (in the 18 to 40 range)
  - These individuals make up at least 50% of the customer base and spend far less on average, excluding outliers, in claims proportionate to age