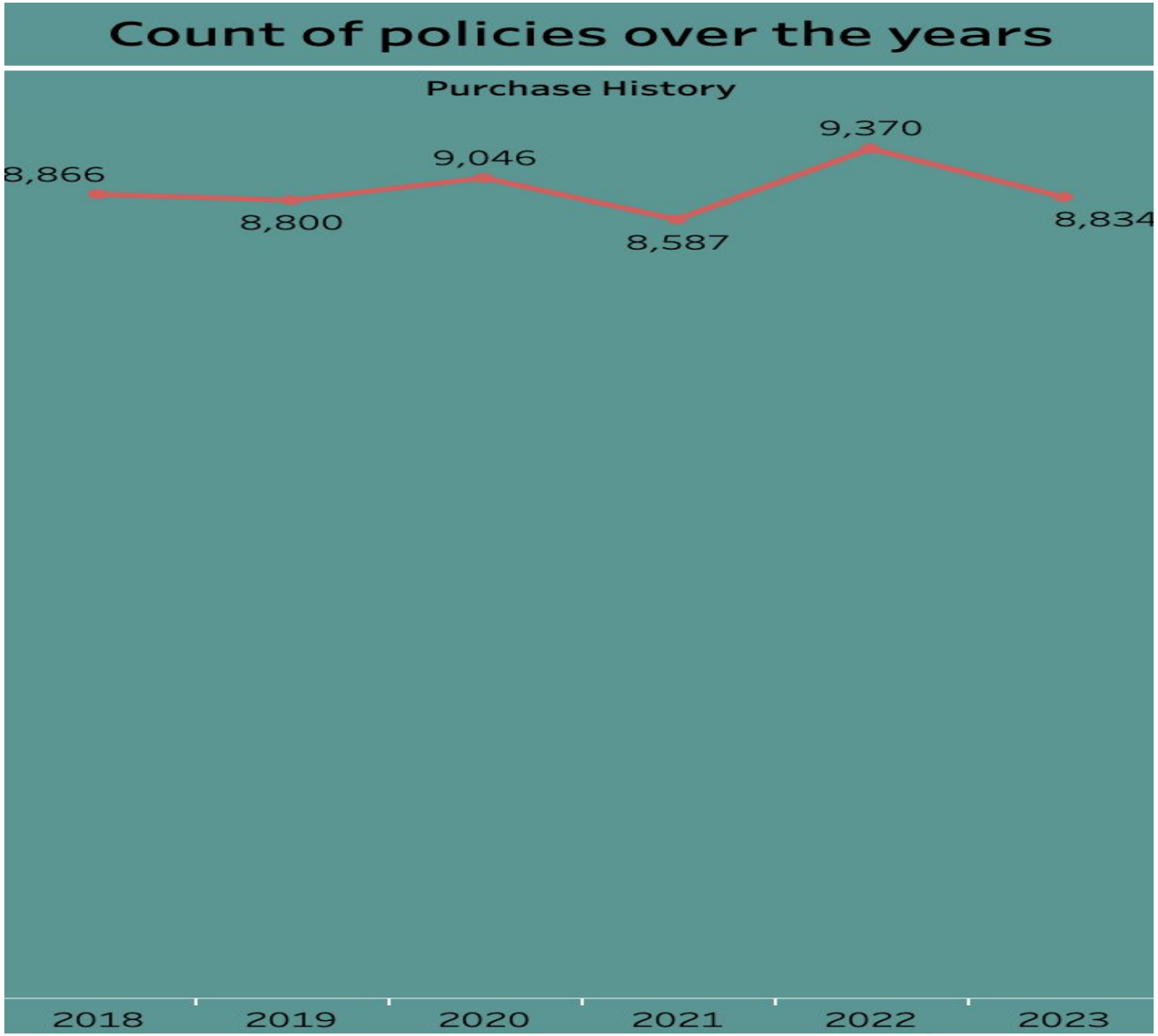
Insurance policy data analysis using Tableau ,Excel and R

Bala Mira G

EDA



The line graph tracks the number of insurance policies from 2018 to 2023.

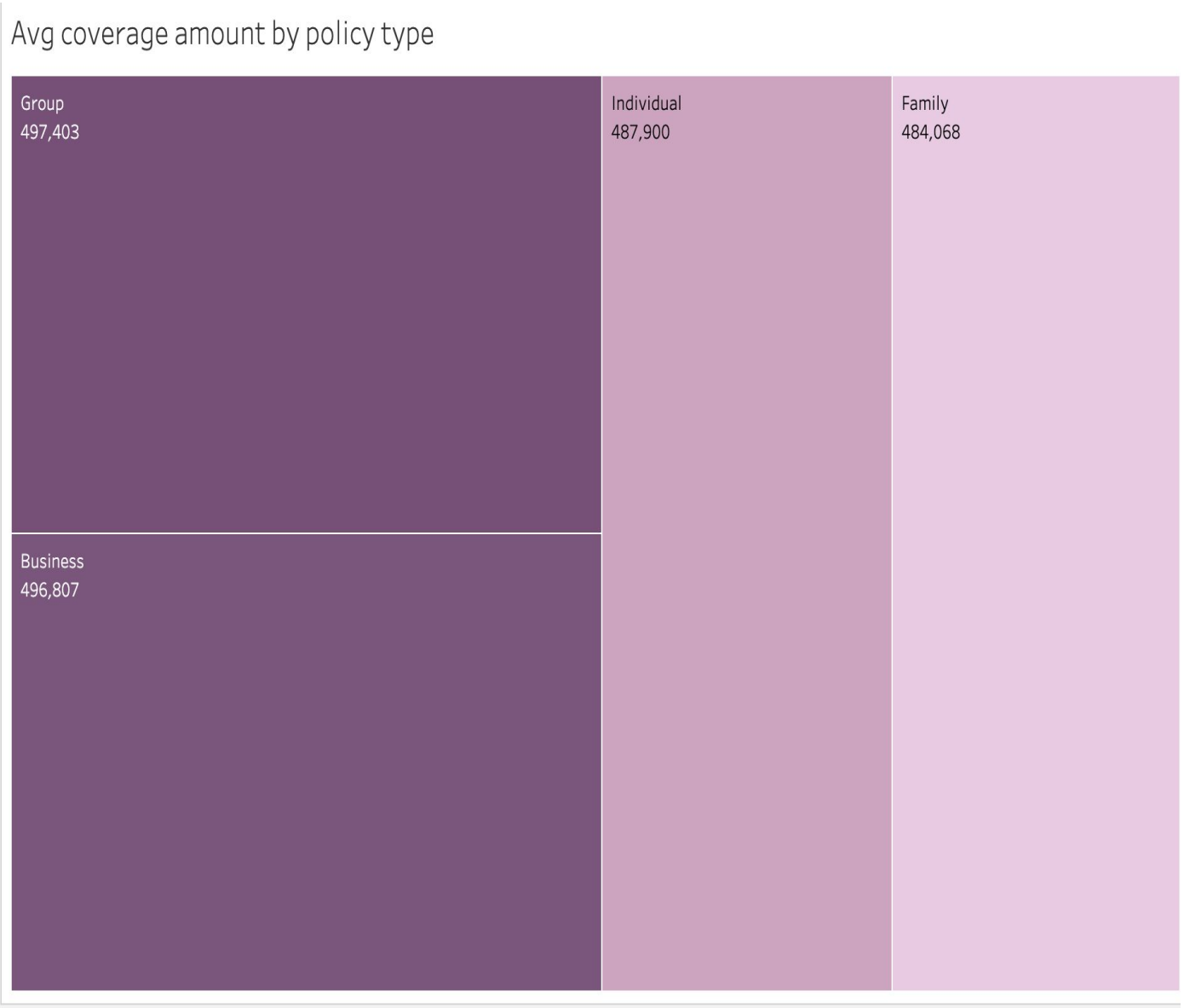
- In 2018, there were 8,866 policies.
- The count increased to 9,046 in 2020.
- It peaked at 9,370 in 2022.
- However, in 2021, there was a sharp decline to 8,587 policies.
- By 2022, it slightly rebounded to 9370, but then decreased again to 8,834 in 2023.

Overall, the graph shows fluctuations in policy counts over the years, with a notable peak in 2020 and 2022 followed by some variability.

EDA

- **Nurses** may have limited time for insurance-related matters due to demanding work schedules. Their focus is primarily on patient care. Nursing salaries, especially for entry-level positions, may be low.. This could affect their ability to invest in additional insurance policies.Insurance companies can work to develop more customised insurance products for them.
- **Teachers** often benefit from group insurance policies provided by educational institutions. These policies cover health, life, and disability.
- Stability and Planning: Teachers value stability and long-term planning. Insurance policies provide financial security for themselves and their families.
- **Entrepreneurs** often take personal responsibility for their insurance needs. They may have policies for business continuity, liability coverage, and personal protection.
- They also recognize the importance of safeguarding their ventures. Insurance policies mitigate risks related to business operations, property, and employees.

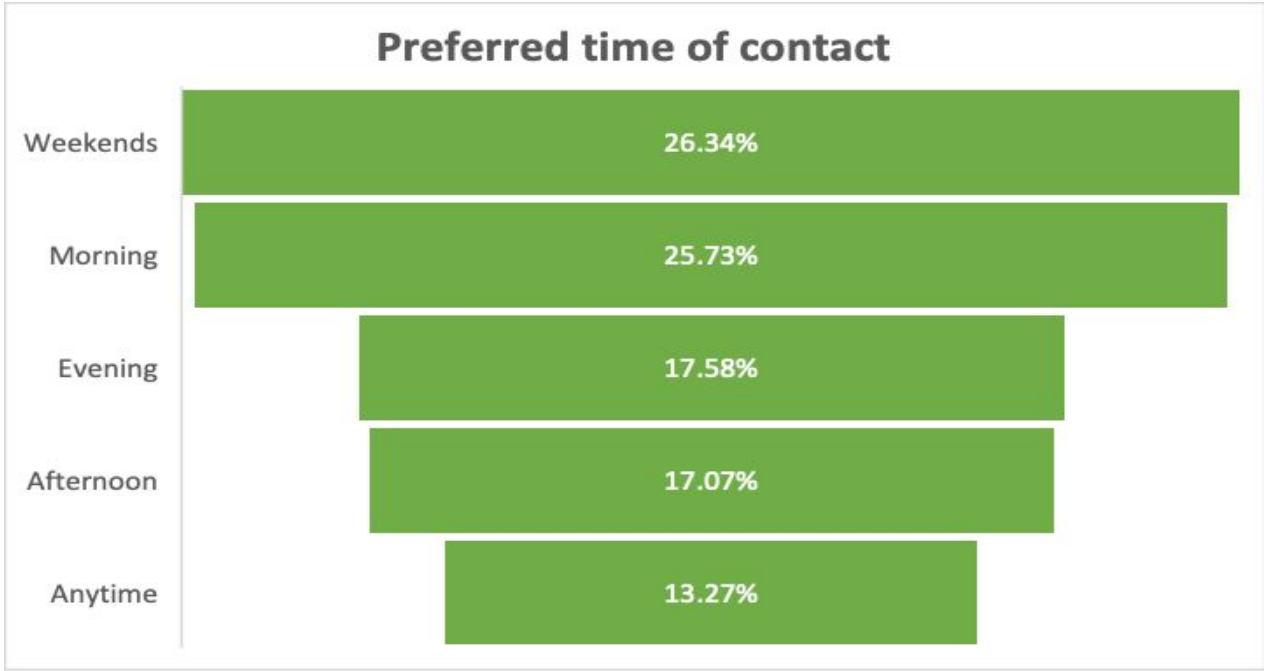
EDA



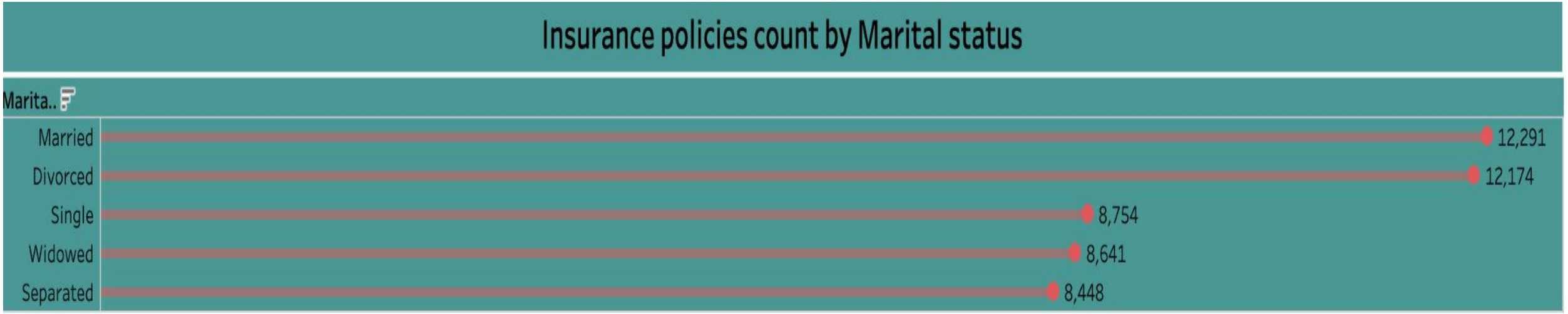
- The heatmap depicts the average coverage amount by policy type
- **Group Policies:** The darkest shade (purple) corresponds to an average coverage amount of 497,403.
- **Individual Policies:** Slightly lighter shade (still purple) with an average coverage amount of 487,900.
- **Family Policies:** The lightest shade (pale purple) indicates an average coverage amount of 484,068.
- Group policies tend to have the highest coverage, followed closely by Business policies, while Family and Individual policies have slightly lower average coverage amounts.

EDA

- The Weekends nd Morning bubble stand out, suggesting the most preferred times for customer contact.
- Other time slots (Morning, Afternoon, Evening, and Anytime) have smaller bubbles, indicating less frequent preferences.
- Contacting the customers at the right time can help in driving more business

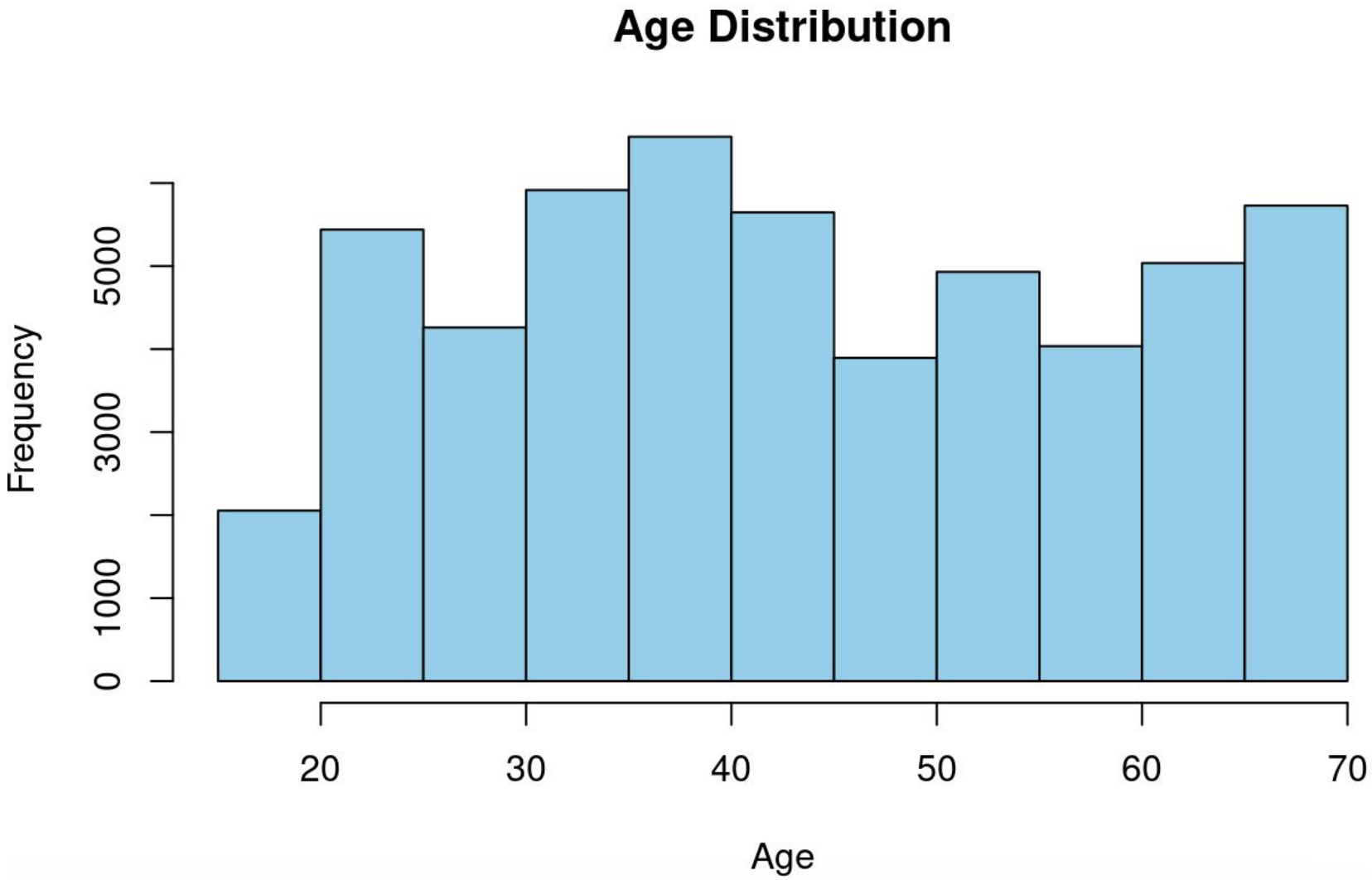


EDA



- Married: Holds the highest number of policies (12,291).Typically, married couples share financial responsibilities and long-term planning. They may invest in joint policies for family protection, home, and health.
- Divorced: Follows closely with a similar count (12,174).Divorcees may become solely responsible for their own well-being and financial security. This can lead to a focus on insurance policies.
- Single: Has fewer policies (8,754).They often prioritize personal needs. Their policies may pertain to health, life coverage, and asset protection.
- Widowed and Separated hold the lowest count (8,641 and 8448).
- In summary, marital status significantly influences insurance policy ownership and the Insurers should tailor marketing to specific life stages (e.g., newlyweds, divorcees).

EDA



- The histogram depicts the age distribution of the customers.
- The x-axis corresponds to age, divided into intervals of 10 years each.
- The y-axis represents frequency, indicating how many individuals fall into each age group.
- The graph does not follow a normal distribution.
- The frequency distribution is not symmetric.
- There are noticeable variations in frequency across different age groups.
- The highest frequency occurs in the age group from around mid-twenties to mid-thirties.

Hypothesis testing
Influence of Income Level on the Coverage amount

```
Call:
lm(formula = `Coverage Amount` ~ `Income Level`, data = policy1)

Residuals:
    Min      1Q  Median      3Q     Max
-450321 -243016  -15422  246355  515645

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.035e+05  2.865e+03  175.71  < 2e-16 ***
`Income Level` -1.320e-01  3.166e-02   -4.17 3.05e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 268400 on 53501 degrees of freedom
Multiple R-squared:  0.0003249, Adjusted R-squared:  0.0003062
F-statistic: 17.39 on 1 and 53501 DF,  p-value: 3.054e-05
```

- The coefficient estimate for Income Level is -0.132.
- The p-value associated with this coefficient is 3.054e-05, which is less than 0.05, indicating that the coefficient is statistically significant.
- Since the coefficient estimate is negative (-0.132) and statistically significant, we can conclude that as income level increases, the coverage amount tends to decrease.
- Therefore, the relationship between income and coverage amount is negative.
- **However, the low R-squared values suggest that the model does not explain much of the variability in coverage amounts.**
- **Further investigation or additional predictors may be needed to improve the model’s performance.**

Real life scenario:

- The above relationship is observed since insurance companies often use a concept called risk pooling.
- In this method, individuals with similar risk profiles are grouped together, and their premiums are pooled to cover potential losses. Higher-income individuals often have fewer dependents or financial obligations, leading to lower coverage needs.
- Conversely, lower-income individuals may have more financial responsibilities, such as mortgage payments, education expenses, or supporting dependents, leading to higher coverage needs.

Hypothesis testing
Influence of Age on the Income Level (Correlation)

- The company wanted to determine if there is any form of association between age and the income level of the insurance holder
- We perform the correlation analysis to find out the same.
- The correlation coefficient is approximately -0.0034.
- The value suggests a very weak or negligible linear relationship between age and income level. In other words, there is almost no linear association between age and income level in our dataset.
- In this case, the correlation coefficient being close to zero indicates that age and income level are not strongly related in a linear manner in our data.
- Correlation doesn't always imply causation. We test this hypothesis by further performing more statistical tests.

```
> # Calculate correlation coefficient
> correlation <- cor(policy1$Age, policy1$`Income Level`)
>
> # Print the correlation coefficient
> print(correlation)
[1] -0.003446592
```

Hypothesis testing
Influence of Age on the Income Level (Regression analysis)

```
> model11 <- lm(`Income Level` ~ Age, data = policy1)
>
> # Summarize the regression model
> summary(model11)

Call:
lm(formula = `Income Level` ~ Age, data = policy1)

Residuals:
   Min     1Q Median     3Q    Max
-62978 -31207  -2048  33220  67439

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 83138.094    490.148 169.618   <2e-16 ***
Age            -8.377     10.508  -0.797    0.425
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36650 on 53501 degrees of freedom
Multiple R-squared:  1.188e-05, Adjusted R-squared:  -6.812e-06
F-statistic: 0.6355 on 1 and 53501 DF,  p-value: 0.4253
```

- The company wanted to determine if there is any form of association between Income level and age.
- **Coefficients:** The coefficient for the Age predictor variable is estimated to be -8.377. This suggests that for each increase in age, the income level is estimated to decrease by $8.377. But the coefficient is not statistically significant, as indicated by the associated p-value of 0.425.
- **Intercept:** The intercept of the regression model is estimated to be $83,138.094. This represents the estimated income level when the age is zero, which may not be meaningful in this context.
- **Multiple R-squared:** This value (1.188e-05) represents the proportion of variance in the dependent variable (Income Level) that is explained by the independent variable (Age). In this case, the multiple R-squared value is very close to zero, indicating that the model explains almost none of the variance in the income level.
- **F-statistic:** This value (0.6355) tests the overall significance of the model. Since the p-value associated with the F-statistic is 0.4253, which is greater than the typical significance level of 0.05, we fail to reject the null hypothesis that all coefficients in the model are equal to zero.
- Overall, based on this analysis, there is no evidence to suggest that age is a significant predictor of income level in the dataset.

Hypothesis testing
Welch 2 sample t test between average income of male and female

```
> t_test_result <- t.test(`Income Level` ~ Gender, data = policy1)
>
> # Print t-test results
> print(t_test_result)

        Welch Two Sample t-test

data:  Income Level by Gender
t = -2.281, df = 53391, p-value = 0.02255
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 -1344.081  -101.742
sample estimates:
mean in group Female   mean in group Male
            82396.77             83119.68
```

- The company wanted to observe if there is any statistical difference between the average income level of male and female.
- We perform the Welch 2 sample t test to analyze further.
- p-value: The probability of observing the data if the null hypothesis (that there is no difference in means between genders) is true.
- In this case, the p-value is 0.02255, which is less than the conventional significance level of 0.05.
- The p value suggests that there is evidence to reject the null hypothesis, indicating that there is a statistically significant difference in average income between genders.
- On the contrary, the alternative hypothesis: specifies that the true difference in means between the Female and Male groups is not equal to zero.
- Overall, based on the results of the Welch Two Sample t-test, there is evidence to suggest that there is a statistically significant difference in average income between genders.

Hypothesis testing
ANOVA, Welch 2 sample t test and Mann Whitney U test

```
> anova_result <- aov(`Income Level` ~ Occupation, data = policy1)
>
> # Print ANOVA table
> summary(anova_result)
               Df    Sum Sq   Mean Sq F value   Pr(>F)
Occupation      8 3.518e+10 4.398e+09   3.275 0.000972 ***
Residuals   53494 7.183e+13 1.343e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> # Perform Wilcoxon Rank Sum Test
> wilcox.test(Age ~ Gender, data = policy1)

        Wilcoxon rank sum test with continuity correction

data:  Age by Gender
W = 368952770, p-value = 1.637e-10
alternative hypothesis: true location shift is not equal to 0
```

```
      Welch Two Sample t-test

data:  Age by Gender
t = 6.4862, df = 53186, p-value = 8.879e-11
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 0.590502 1.101918
sample estimates:
mean in group Female   mean in group Male
         44.57587           43.72966
```

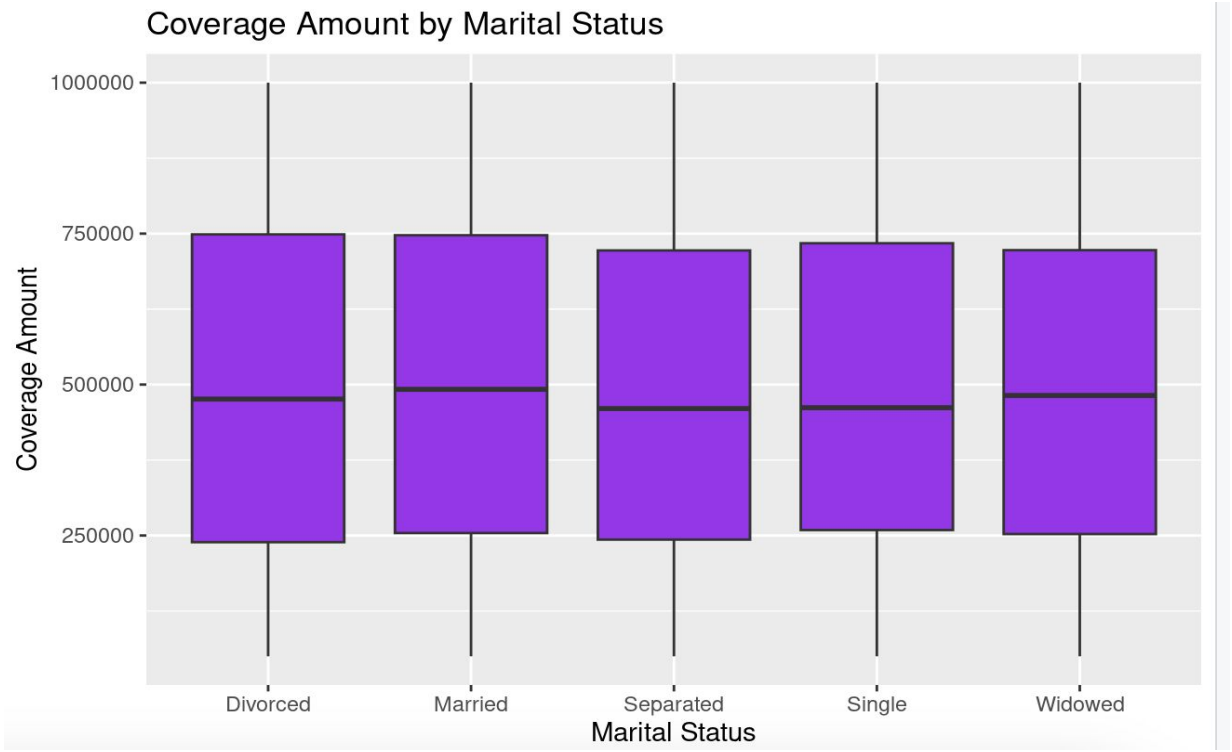
- The first test (ANOVA) suggests that there is a statistically significant association between occupation and the outcome variable being analyzed since the p value is less than 0.05.
- The second result(Welch 2 sample t test) suggests that there is a true difference between the average age of Male an Female insurance holders since the p value is less than 0.05.
- The third test which is the Wilcoxon rank sum test is a suitable statistical method for analyzing age data to investigate potential differences between males and females in terms of median age.
- The result from the test gives a P value which is very small and suggests that we have a strong evidence to reject the null hypothesis.
- From all the statistical tests that we performed, we reject the null hypothesis and establish the alternate hypothesis.

Hypothesis testing
Chi Squared test between Marital status and Coverage amount

We have performed the ANOVA test using R to assess the association between marital status and coverage amount.

- The p-value (Pr(>F)) associated with Marital Status is 0.000232, which is less than the typical significance level of 0.05.
- Therefore, we reject the null hypothesis and conclude that there is a significant difference in coverage amount across different marital status categories.
- In simpler terms, this means that marital status has a statistically significant impact on coverage amount.

```
> # Perform ANOVA test
> anova_result <- aov(`Coverage Amount` ~ `Marital Status`, data = policy1)
>
> # Summarize the ANOVA results
> summary(anova_result)
                  Df    Sum Sq   Mean Sq F value   Pr(>F)
`Marital Status`     4 1.562e+12 3.905e+11   5.422 0.000232 ***
Residuals        53498 3.853e+15 7.202e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```



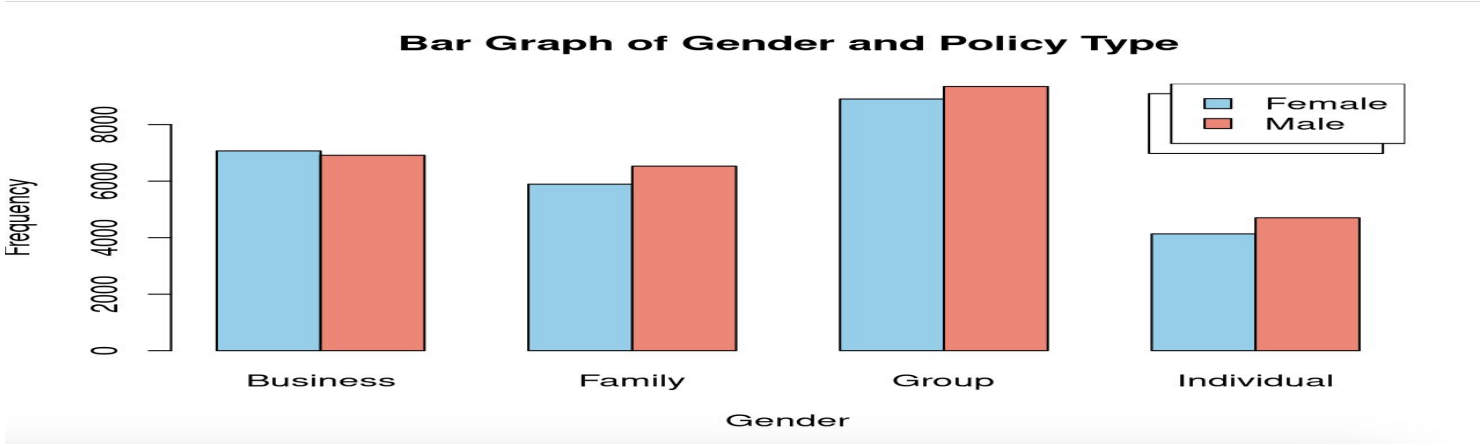
Hypothesis testing
Association between Income Level and Geographical information & Association between Gender and Policy type

```
> model3 <- lm(`Income Level` ~ `Geographic Information`, data = policy1)
> summary(model3)
```

- Result: After performing a linear regression between Income Level and Geographical information , the model revealed that certain geographic locations have a significant impact on income level.
- For example, Gujarat shows a significant positive effect on income level, while Sikkim shows a significant negative effect.

```
> contingency_table <- table(policy1$`Policy Type`, policy1$Gender)
>
> # Perform a chi-squared test
> chi_square_result <- chisq.test(contingency_table)
>
> # Display the test result
> print(chi_square_result)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 40.131, df = 3, p-value = 9.994e-09
```

- The result of Pearson's chi-squared test indicates a statistically significant association between policy type and gender (X-squared = 40.131, df = 3, p-value < 0.001).
- Therefore, we can reject the null hypothesis and conclude that there is an association between policy type and gender in our dataset.
- We can further analyze this hypothesis with a bar graph plotted using R studio.



Association between Preferred language of communication and age

- The Kruskal-Wallis test results indicate that there is a significant association between preferred language and age (p-value < 0.05).
- This suggests that the preferred language of communication varies significantly across different age groups in our dataset.

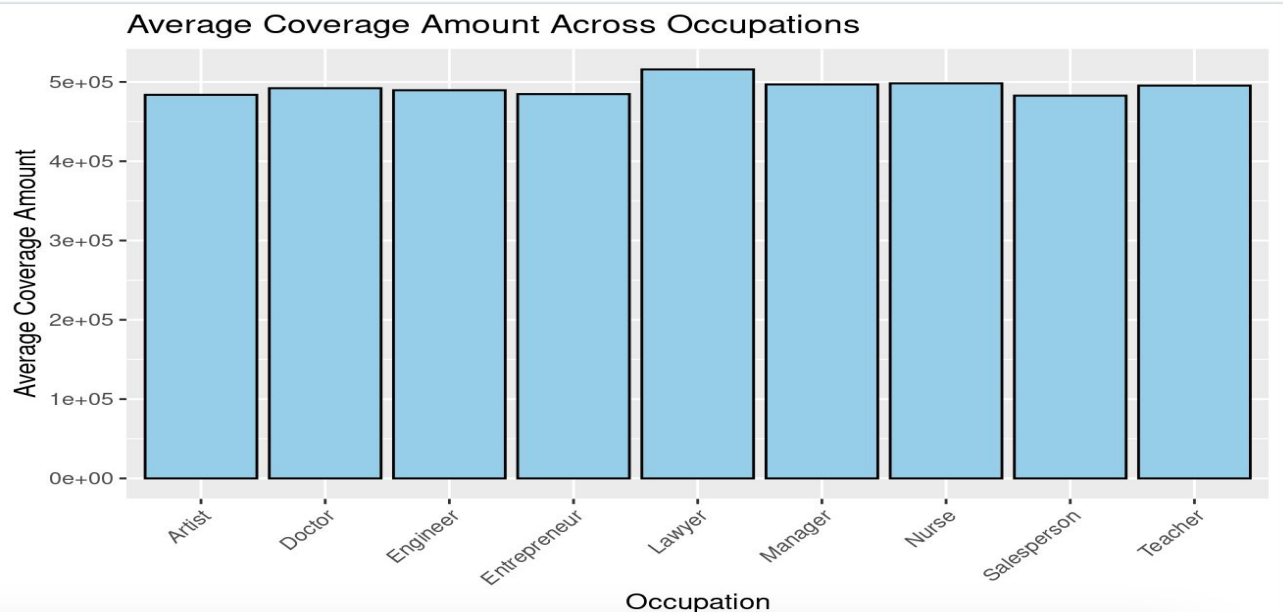
```
> result <- kruskal.test(`Preferred Language` ~ Age, data = policy1)
>
> # Print the result
> print(result)

        Kruskal-Wallis rank sum test

data:  Preferred Language by Age
Kruskal-Wallis chi-squared = 116.1, df = 52, p-value = 8.499e-07
```

```
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 56.0825, df = 4, p-value = 0


                             Comparison of x by group
                                   (Bonferroni)
Col Mean-|
Row Mean |    English     French     German   Mandarin
---------+-------------------------------------------
  French |   4.366328
         |    0.0001*
         |
  German |   5.571763   1.245685
         |    0.0000*     1.0000
         |
Mandarin |   6.821782   3.129767   2.088288
         |    0.0000*    0.0087*     0.1839
         |
 Spanish |   2.207123  -1.730825  -2.816855  -4.334480
         |     0.1365     0.4174    0.0242*    0.0001*

alpha = 0.05
Reject Ho if p <= alpha/2
```

Average coverage amount and occupation

The Kruskal-Wallis test we performed indicates that there is a statistically significant difference in the average coverage amount across different occupations.

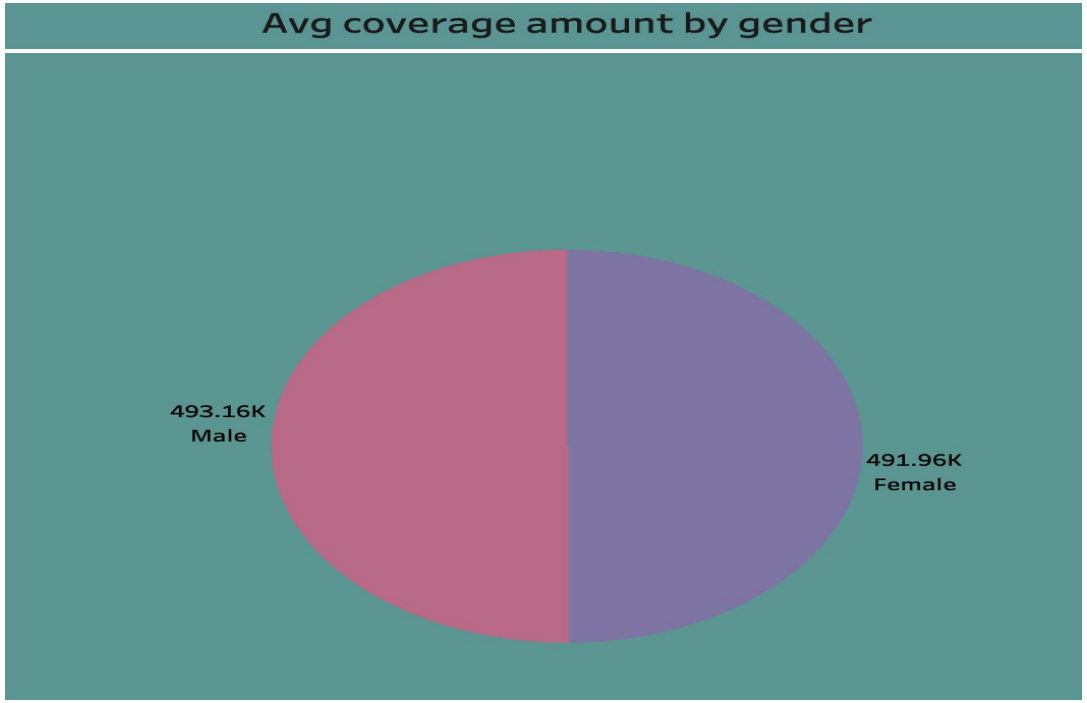
- **Interpretation:**
  - Kruskal-Wallis chi-squared: This value represents the test statistic.
  - p-value: This is the probability of obtaining the observed data if the null hypothesis (i.e., no difference between groups) were true. In this case, the p-value is very small (8.904e-13), indicating strong evidence against the null hypothesis.
- Since the p-value is less than the typical significance level of 0.05, we can conclude that there is a significant difference in the average coverage amount across different occupations.
- In other words, we have enough evidence to suggest that at least one occupation has a different average coverage amount compared to the others.




```
> result_kruskal <- kruskal.test(`Coverage Amount` ~ Occupation, data = policy2)
> print(result_kruskal)

        Kruskal-Wallis rank sum test

data:  Coverage Amount by Occupation
Kruskal-Wallis chi-squared = 73.718, df = 8, p-value = 8.904e-13
```

Average coverage amount and Gender

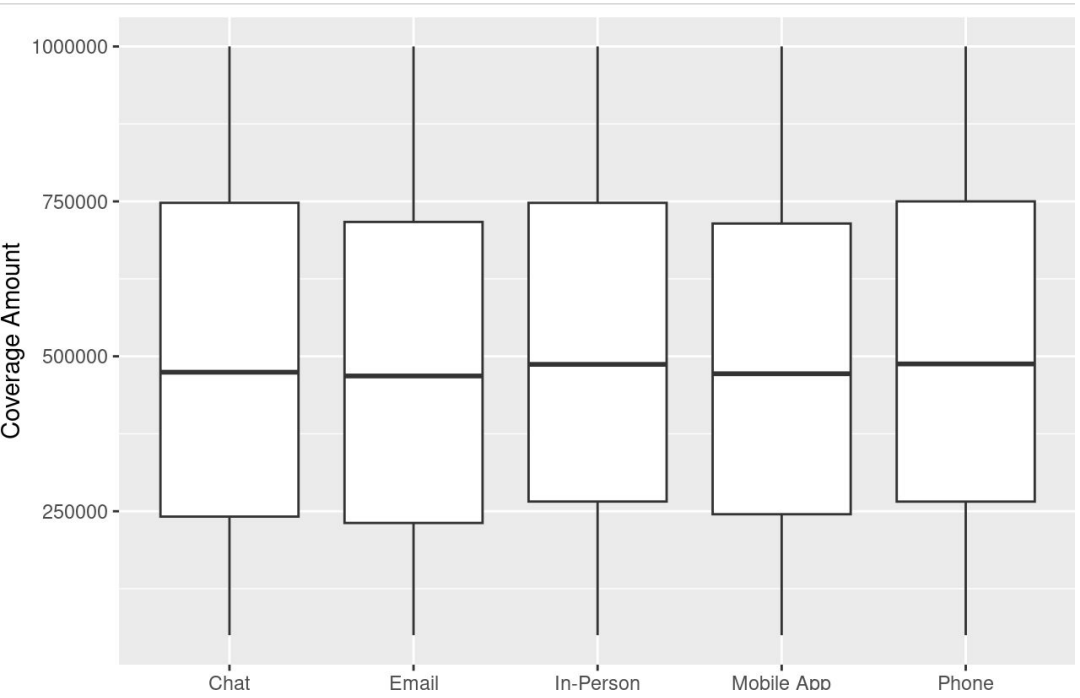
- The Kruskal-Wallis rank sum test we performed on the coverage amount by gender indicates that there is no statistically significant difference in the average coverage amount between different genders.
- **Interpretation:**
- p-value: This is the probability of obtaining the observed data if the null hypothesis (i.e., no difference between groups) were true.
- In this case, the p-value is relatively high (0.6746) indicating weak evidence against the null hypothesis.
- Since the p-value is greater than the typical significance level of 0.05, we fail to reject the null hypothesis.
- Therefore, based on this test, there is no evidence to suggest that there is a significant difference in the average coverage amount between different genders.




```
> result_kruskal <- kruskal.test(`Coverage Amount` ~ Gender, data = policy2)
> print(result_kruskal)

        Kruskal-Wallis rank sum test

data:  Coverage Amount by Gender
Kruskal-Wallis chi-squared = 0.17629, df = 1, p-value = 0.6746
```

Association between mode of Interactions with customer service and the coverage amount

- The business wanted to check if the mode of interaction with customers such as (phone calls, emails, chat support, etc.) can influence the overall customer service experience.
- For example, if there is any possibility of customers who have had positive experiences with customer service opting to have higher coverage amounts due to increased satisfaction and trust with the insurance provider.
- The null hypothesis states that there is no relation between mode of interaction and coverage amount .However, the p value (which is less than 0.05) suggests that we have enough evidence to reject the null hypothesis and suggest the alternate hypothesis.




```
> ggplot(policy1, aes(x = `Interactions with Customer Service`, y = `Coverage Amount`)) +
+       geom_boxplot() +
+      labs(x = "Mode of Interaction", y = "Coverage Amount")
> anova_model <- aov(`Coverage Amount` ~ `Interactions with Customer Service`, data = policy1)
> summary(anova_model)
                                    Df    Sum Sq   Mean Sq F value   Pr(>F)
`Interactions with Customer Service`     4 2.646e+12 6.616e+11   9.189 2.04e-07 ***
Residuals                           53498 3.852e+15 7.200e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Key Takeaways and Recommendations

- Salespersons form the highest number of policyholders while the nurses form the lowest.
- Group policy types has the highest number of takers followed bt Business policy type.
- Married individuals have the highest number of policies followed by Divorcees.
- The preferred time to contact for most of the customers is either during Morning or the weekends.
- There are noticeable variations in frequency across different age groups of people who have opted to have insurance policies.
- Our statistical test proves that there is an association between mode of service(email,chat and phone) and the coverage amount.
- There is no statistically significant difference in the average coverage amount between different genders.
- Avg coverage amount varies across different professions.
- The preferred language of communication varies significantly across different age groups. Insurance companies can cater to the language needs to suit the customers' preferences.
- The results from linear regression performed between Income Level and Geographical information , the model revealed that certain geographic locations have a significant impact on income level.
- From our dataset, we did not find any strong association to check for the influence of age on the income level.
- Based on the summary of the Welch Two Sample t-test, we have enough evidence to suggest that there is a statistically significant difference in average income between genders.
- Though the strength of the relationship is weak, we were able to deduce the relationship on the Influence of Income Level on the Coverage amount.