FINAL PROJECT

Abarna Sivaprakasam

College of Professional Studies, Northeastern University

ALY6010: Probability Theory and Introductory Statistics

Dr. Mimoza Dimodugno

December 11, 2023

Table of Contents

INTRODUCTION

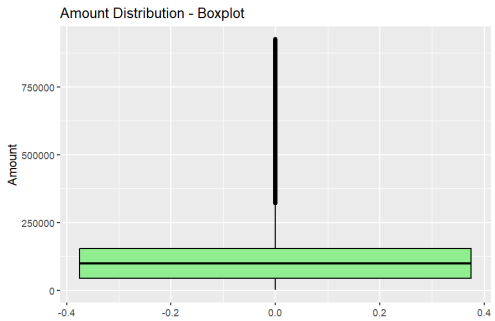
This report uses inferential statistics and exploratory data analysis (EDA) to provide an in-depth analysis of a US medical malpractice dataset. The dataset comprises various information such as the age of the claimant, the amount of the claim payment, the severity rating to the patient(1-9), whether the claimant took a private attorney or not, the claimant's marital status, type of medical insurance that they enrolled in and the specialty of the physician.

EXPLORATORY DATA ANALYSIS

1. About the dataset:
    - Title: Medical Malpractice
    - Source: Kaggle
    - Link: https://www.kaggle.com/datasets/willianoliveiragibin/mendical-insurance-and-malpratice
    - Summary: The dataset contains 79,210 rows and 8 columns. This report delves into data patterns and correlations to understand better healthcare factors, patient safety, legal effects, and economic impacts.
2. Data Cleaning:
    - Missing Values: There were no missing values in this dataset.

```
> sum(is.na(data))
[1] 0
```

    - Outliers: Outliers in claim amount were identified.

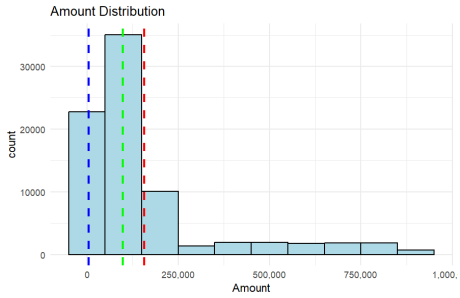


3. Summary of the data:

```
     Amount          Severity         Age        Private.Attorney Marital.Status
 Min.   :  1576   Min.   :1.0   Min.   : 0.0   Min.   :0.0000   Min.   :0.00
 1st Qu.: 43670   1st Qu.:3.0   1st Qu.:28.0   1st Qu.:0.0000   1st Qu.:1.00
 Median : 98131   Median :4.0   Median :43.0   Median :1.0000   Median :2.00
 Mean   :157485   Mean   :4.8   Mean   :42.7   Mean   :0.6609   Mean   :1.89
 3rd Qu.:154675   3rd Qu.:7.0   3rd Qu.:58.0   3rd Qu.:1.0000   3rd Qu.:2.00
 Max.   :926411   Max.   :9.0   Max.   :87.0   Max.   :1.0000   Max.   :4.00
  Specialty          Insurance           Gender
 Length:79210       Length:79210       Length:79210
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

4. Descriptive Statistics:

```
                vars     n      mean        sd median   trimmed      mad  min
Amount             1 79210 157484.55 193135.09  98131 110706.60 81858.79 1576
Severity           2 79210      4.80      2.08      4      4.54     1.48    1
Age                3 79210     42.70     19.81     43     42.91    22.24    0
Private.Attorney   4 79210      0.66      0.47      1      0.70     0.00    0
Marital.Status     5 79210      1.89      1.00      2      1.80     0.00    0
Specialty*         6 79210      7.81      4.84      7      7.38     2.97    1
Insurance*         7 79210      2.98      1.03      3      3.07     1.48    1
Gender*            8 79210      1.40      0.49      1      1.37     0.00    1
                   max  range  skew kurtosis     se
Amount          926411 924835  2.25     4.30 686.23
Severity             9      8  0.78    -0.63   0.01
Age                 87     87 -0.07    -0.79   0.07
Private.Attorney     1      1 -0.68    -1.54   0.00
Marital.Status       4      4  0.71     0.38   0.00
Specialty*          20     19  0.71    -0.04   0.02
Insurance*           5      4 -0.58    -0.35   0.00
Gender*              2      1  0.42    -1.82   0.00
```



The average claim amount is $157,484, with a broad range (SD = $193,135). Severity ratings vary from 1 to 9 with an average of 4.8. The claimants' average age is 42.7 years, with a moderate dispersion (SD = 19.81). Private attorneys are involved in 66% of cases, with different distributions by marital status, physician specialty, insurance type, and gender. Skewness and kurtosis levels indicate moderate deviations from normality.

5. Cross Tabulations:

- Cross Tabulation between Marital status and private attorney:

  Marital.Status is represented by five categories: 0, 1, 2, 3, and 4 which represent Married, widowed, separated, divorced, and single respectively.

  Private.Attorney has two categories: 0 (No private attorney representation) and 1 (Private attorney representation).

```
       0     1
0    797  3035
1   8362 14440
2  11968 29252
3      0   994
4   5734  4628
```

- Cross-tabulation between Marital status and gender
- Cross-tabulation between marital status and severity

	Female	Male
0	1133	2699
1	18772	4030
2	23146	18074
3	994	0
4	3725	6637

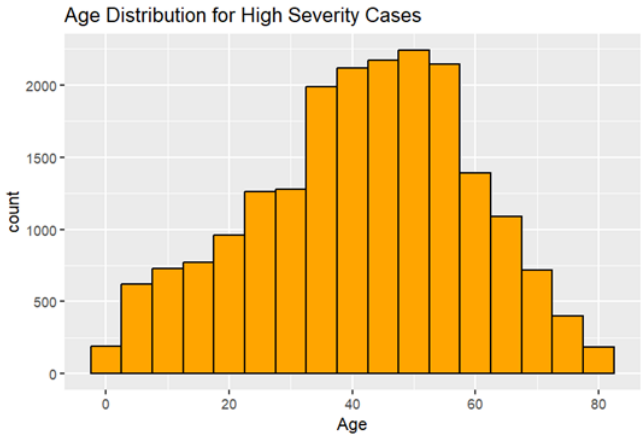
	1	2	3	4	5	6	7	8	9
0	17	37	610	652	248	277	862	510	619
1	220	492	8378	4293	3349	1149	2028	883	2010
2	338	660	13825	8396	5091	1663	5208	1837	4202
3	28	59	276	193	137	78	22	52	149
4	62	92	5162	2175	790	208	753	345	775

Individuals with separated marital status frequently hire private attorneys, although divorcees never do. Widowed females outnumber males, and severity levels differ depending on marital status, with widowed persons having more severe cases. Patterns appear in terms of specializations, such as anesthesiology cases being prominent in severity levels 3, 4, 7, and 9.

6. <u>Subset analysis:</u>

Subsets were created for categorical, discrete, and continuous variables.

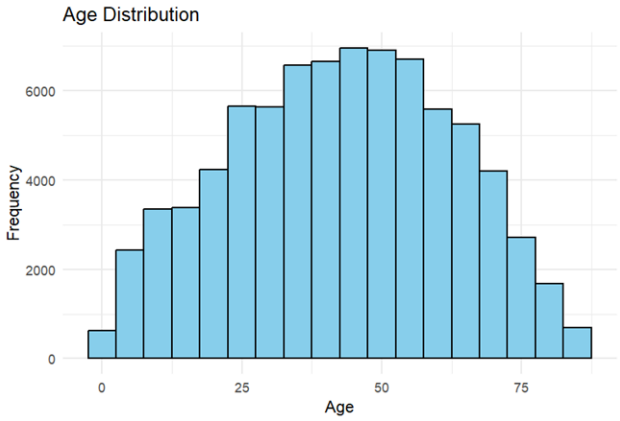
- Categorical Variable:



It shows only the values where the Severity variable is greater than or equal to 7. It filters out cases with severity levels 1 to 6 and keeps only those with severity levels 7, 8, and 9.

From the graph, we can infer that People around the age of 50 have high severity levels.

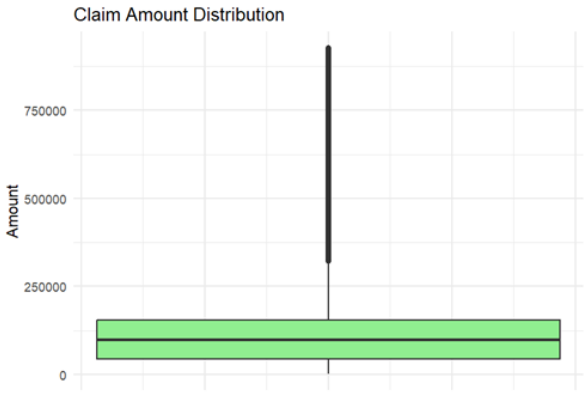
- Discrete Variable



There are more people around the age of 49.

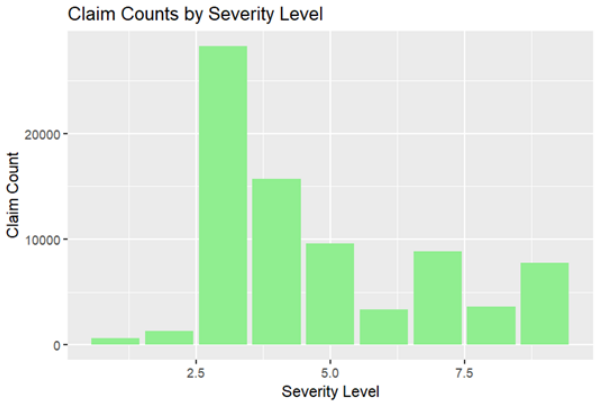
- Continuous Variable

Amount is the continuous variable.



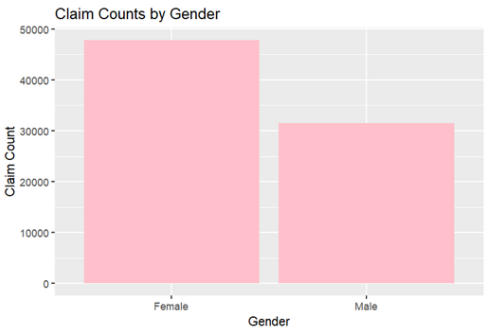
7. <u>Visualizations</u>

- Claim Count by Severity Level



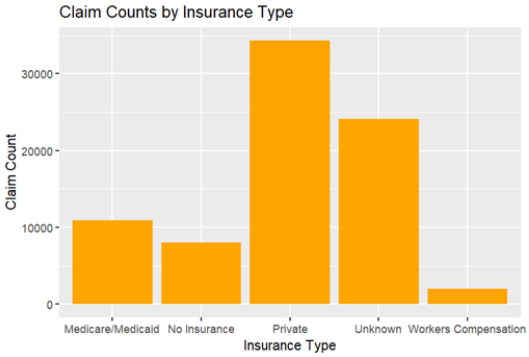
There are more medical malpractice claims with a severity rating of 3.

- Claim count by gender



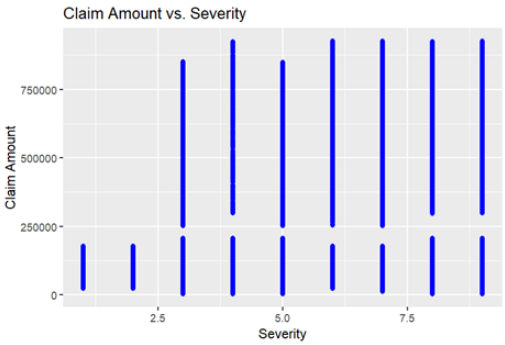
There are more medical malpractices associated with the females.

- Claim count by Insurance



There are more malpractice claims associated with private insurance companies.

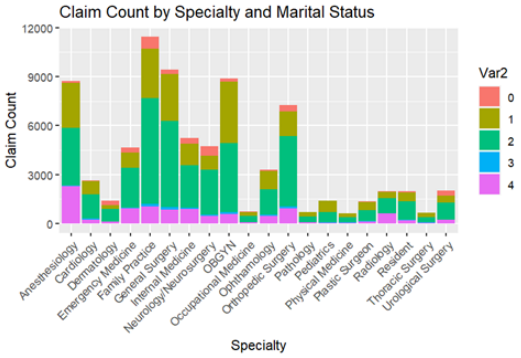
- Claim amount and severity



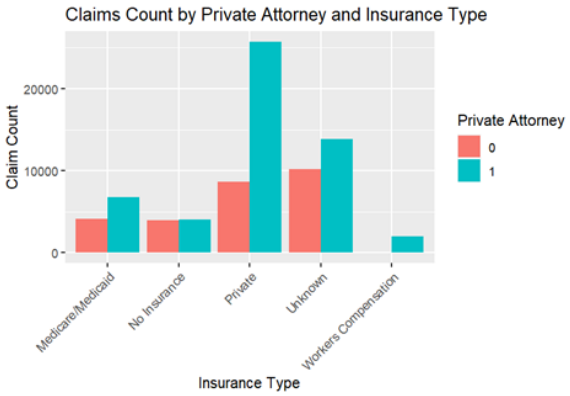
We can infer that severity levels 4,6,7,8, and 9 are almost the same amount.

- Claim amount by Specialty and Marital Status

From this graph, we can infer that there are more medical malpractice claims to marital status 4 in Anesthesiology, and Marital status 0 and 2 in Family Practices

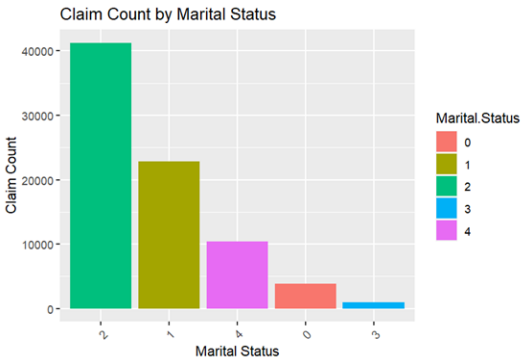


- Private attorney and Insurance type



There are more people with private attorneys.

- .Claim count by Marital status:



There are more people from marital status 2 associated with medical malpractices.

INFERENTIAL STATISTICS AND HYPOTHESIS TESTING

Question 1: *Is there enough data to compare claim amounts between private attorney cases and all instances?*

Reason: The initial EDA revealed a distribution of claim amounts, bringing the question of whether there is enough data to compare claim numbers specifically involving private attorney services.

- **Null Hypothesis** - The mean claim amount presented by the private attorney is equal to the overall mean claim amount in the dataset.
- **Alternate Hypothesis** - The mean claim amount presented by the private attorney is not equal to the overall mean claim amount in the dataset.

Code explanation:

```
# Question 1: Is there sufficient evidence to conclude that the average claim amounts for cases represented by privat

# Null Hypothesis (H0): The mean claim amount for cases with private attorneys is equal to the overall mean claim amc
# Alternative Hypothesis (H1): The mean claim amount for cases with private attorneys is different from the overall m
one_sample_t_test_result <- t.test(data$Amount, mu = mean(data$Amount))

# Display null and alternative hypotheses
cat("\nNull Hypothesis: The mean claim amount for cases with private attorneys is equal to the overall mean claim amc
cat("Alternative Hypothesis: The mean claim amount for cases with private attorneys is different from the overall mea
# Display t-test results
cat("\nOne-Sample t-test Results for Claim Amounts:\n")
print(one_sample_t_test_result)
# Interpret the results
cat("\nInterpretation:\n")
if (one_sample_t_test_result$p.value < 0.05) {
  cat("Reject the null hypothesis. There is sufficient evidence that the average claim amounts for cases with private
} else {
  cat("Fail to reject the null hypothesis. No sufficient evidence of a difference in average claim amounts.\n")
}
```

In this case, we used t.test to do a one-sample t-test on the claim amount. We utilized the if statement in the interpretation section to see if the p-value was less than 0.05. If it is true, there is enough evidence to reject the null hypothesis, showing that there is a claim amount difference between the private attorney and the total dataset. Otherwise, it shows that it rejects the null 3 hypothesis, which states that the mean claim amount is the same for private attorneys and the total data set.

Output:

```
Null Hypothesis: The mean claim amount for cases with private attorneys is equal to the overall mean claim amount.
> cat("Alternative Hypothesis: The mean claim amount for cases with private attorneys is different from the overall mean cl
aim amount.\n")
Alternative Hypothesis: The mean claim amount for cases with private attorneys is different from the overall mean claim amc
unt.
> # Display t-test results
> cat("\nOne-Sample t-test Results for Claim Amounts:\n")

One-Sample t-test Results for Claim Amounts:
> print(one_sample_t_test_result)

        One Sample t-test

data:  data$Amount
t = 0, df = 79209, p-value = 1
alternative hypothesis: true mean is not equal to 157484.6
95 percent confidence interval:
 156139.5 158829.6
sample estimates:
mean of x
 157484.6

> # Interpret the results
> cat("\nInterpretation:\n")

Interpretation:
> if (one_sample_t_test_result$p.value < 0.05) {
+   cat("Reject the null hypothesis. There is sufficient evidence that the average claim amounts for cases with private att
orneys are different from the overall average claim amount.\n")
+ } else {
+   cat("Fail to reject the null hypothesis. No sufficient evidence of a difference in average claim amounts.\n")
+ }
Fail to reject the null hypothesis. No sufficient evidence of a difference in average claim amounts.
```

The t-test statistic, calculated as 0 with 79209 degrees of freedom, yields a p-value of 1. Comparing this p-value to the common significance level of 0.05 and considering the alternative hypothesis that the true mean is not equal to 157,484.6 (the overall mean claim amount), the interpretation leads to the conclusion of "Fail to reject the null hypothesis." Therefore, there is insufficient evidence to assert a difference in average claim amounts between cases with private attorneys and the overall average claim amount.

Question 2: *Does the distribution of claim severity differ significantly among physician specialties?*

Reason: The EDA emphasized discrepancies in claim severity between physician specialties, motivating an investigation into the statistical significance of these differences.

- **Null Hypothesis -** There is no significant claim severity distribution across the different specialties.
- **Alternate Hypothesis -** There is significant claim severity distribution across the different physician specialties.

Code Explanation:

```r
# Question 2: Does the distribution of claim severity differ significantly among physician specialties?

# Null Hypothesis (H0): There is no significant difference in claim severity distribution across physician specialtie
# Alternative Hypothesis (H1): Claim severity distribution varies significantly among different physician specialties

# Perform Kruskal-Wallis test
kruskal_result <- kruskal.test(Severity ~ Specialty, data = data)

# Display null and alternative hypotheses
cat("\nNull Hypothesis: There is no significant difference in claim severity distribution across physician specialtie
cat("Alternative Hypothesis: Claim severity distribution varies significantly among different physician specialties.\

# Display Kruskal-Wallis test results
cat("\nKruskal-Wallis Test Results for Claim Severity by Physician Specialty:\n")
print(kruskal_result)

# Interpret the results
cat("\nInterpretation:\n")
if (kruskal_result$p.value < 0.05) {
  cat("Reject the null hypothesis. There is sufficient evidence that claim severity distribution differs significantl
} else {
  cat("Fail to reject the null hypothesis. No sufficient evidence of a difference in claim severity distribution.\n")
}
```

Here, kruskal.test was employed. To compare the distribution of two or more independent groups, apply this non-parametric test. Here, it examines whether different physical disciplines have different severity distributions. A test statistic and a p-value are provided by the Kruskal-Wallis test findings.

To conclude, the test's p-value is compared to the significance level, which is typically 0.05. It determines whether the p-value is smaller than the standard significance level of 0.05 in the interpretation section. The null hypothesis is rejected if the p-value is less than 0.05, providing sufficient evidence of a substantial variation in the distribution of claim severity among medical specializations.

Result:

```
Null Hypothesis: There is no significant difference in claim severity distribution across physician specialties.
> cat("Alternative Hypothesis: Claim severity distribution varies significantly among different physician specialties.\n")
Alternative Hypothesis: Claim severity distribution varies significantly among different physician specialties.
>
> # Display Kruskal-Wallis test results
> cat("\nKruskal-Wallis Test Results for Claim Severity by Physician Specialty:\n")

Kruskal-Wallis Test Results for Claim Severity by Physician Specialty:
> print(kruskal_result)

        Kruskal-Wallis rank sum test

data:  Severity by Specialty
Kruskal-Wallis chi-squared = 5437.9, df = 19, p-value < 2.2e-16

>
> # Interpret the results
> cat("\nInterpretation:\n")

Interpretation:
> if (kruskal_result$p.value < 0.05) {
+   cat("Reject the null hypothesis. There is sufficient evidence that claim severity distribution differs significantly am
ong physician specialties.\n")
+ } else {
+   cat("Fail to reject the null hypothesis. No sufficient evidence of a difference in claim severity distribution.\n")
+ }
Reject the null hypothesis. There is sufficient evidence that claim severity distribution differs significantly among physi
cian specialties.
```

Interpretation

Results of the Kruskal-Wallis test: 5437.9 is the chi-squared value. There is more evidence to reject the null hypothesis the higher the chi-squared score.

Degrees of freedom (df) = 19. The number of groups being compared determines the degrees of freedom in the Kruskal-Wallis test (p-value < 2.2e-16; very low p-value). The incredibly small p-value indicates compelling evidence opposing the null hypothesis.

Interpretation: We reject the null hypothesis since the p-value is less than the generally accepted significance level of 0.05.

There is enough data to conclude that there are significant variations in the distribution of claim severity between medical disciplines.

Question 3: *Is there a significant difference in severity between males and females?*

Reason: The EDA revealed gender differences, prompting an examination into whether these differences extend to claim severity.

- **Null Hypothesis (H0):** The mean claim severity for male claimants is equal to the mean claim severity for female claimants.
- **Alternative Hypothesis (H1):** The mean claim severity for male claimants is not equal to the mean claim severity for female claimants.

Code explanation:

```r
#Question 3: Is there a significant difference in claim severity between male and female claimants?
# Null Hypothesis (H0): The mean claim severity for male claimants is equal to the mean claim severity for female cla
# Alternative Hypothesis (H1): The mean claim severity for male claimants is different from the mean claim severity f

# Perform Two-sample t-test
t_test_gender_result <- t.test(data$Severity ~ data$Gender)

# Display null and alternative hypotheses
cat("\nNull Hypothesis: The mean claim severity for male claimants is equal to the mean claim severity for female cla
cat("Alternative Hypothesis: The mean claim severity for male claimants is different from the mean claim severity for

# Display Two-sample t-test results
cat("\nTwo-sample T-test Results for Claim Severity between Male and Female Claimants:\n")
print(t_test_gender_result)

# Interpret the results
cat("\nInterpretation:\n")
if (t_test_gender_result$p.value < 0.05) {
  cat("Reject the null hypothesis. There is sufficient evidence that the mean claim severity differs between male and
} else {
  cat("Fail to reject the null hypothesis. No sufficient evidence of a difference in mean claim severity between male
}
```

By using t.test on the severity column to compare males and females.

The code checks whether the p-value is less than 0.05. If the p-value is less than 0.05, it indicates sufficient evidence to reject the null hypothesis indicating The mean claim severity for male claimants is not equal to the mean claim severity for female claimants. Otherwise, The mean claim severity for male claimants is equal to the mean claim severity for female claimants.

Result:

```
Null Hypothesis: The mean claim severity for male claimants is equal to the mean claim severity for female claimants.
> cat("Alternative Hypothesis: The mean claim severity for male claimants is different from the mean claim severity for fem
ale claimants.\n")
Alternative Hypothesis: The mean claim severity for male claimants is different from the mean claim severity for female cla
imants.
>
> # Display Two-sample t-test results
> cat("\nTwo-sample T-test Results for Claim Severity between Male and Female Claimants:\n")

Two-sample T-test Results for Claim Severity between Male and Female Claimants:
> print(t_test_gender_result)

        Welch Two Sample t-test

data:  data$Severity by data$Gender
t = -7.9091, df = 67307, p-value = 2.631e-15
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 -0.14923856 -0.08996137
sample estimates:
mean in group Female   mean in group Male
            4.752188             4.871788

>
> # Interpret the results
> cat("\nInterpretation:\n")

Interpretation:
> if (t_test_gender_result$p.value < 0.05) {
+   cat("Reject the null hypothesis. There is sufficient evidence that the mean claim severity differs between male and fem
ale claimants.\n")
+ } else {
+   cat("Fail to reject the null hypothesis. No sufficient evidence of a difference in mean claim severity between male and
female claimants.\n")
+ }
Reject the null hypothesis. There is sufficient evidence that the mean claim severity differs between male and female claim
ants.
```

The t-test results show a significant difference in mean claim severity between males and females. We reject the null hypothesis with a t-statistic of -7.9091 and an extremely low p-value of 2.631e-15. This implies that the severity of medical malpractice claims varies on average between male and female claimants, with females having lower mean claim severity than males.

Question 4: *Is there a correlation between the claimant's age and the severity?*

Reason: Observing the age distribution in the EDA raised the possibility of a relationship between age and claim severity.

- **Null Hypothesis -** There is no correlation between the claimant’s age and severity.
- **Alternate Hypothesis** - There is a correlation between the claimant’s age and severity.

Code explanation:

```r
# Question 3: Is there a statistically significant correlation between the age of the claimant and the sev

# Null Hypothesis (H0): There is no correlation between the age of the claimant and claim severity.
# Alternative Hypothesis (H1): There is a significant correlation between age and claim severity.

# Perform Pearson correlation test (assuming 'Age' and 'Severity' are the columns for age and severity)
correlation_result <- cor.test(data$Age, data$Severity)

# Display null and alternative hypotheses
cat("\nNull Hypothesis: There is no correlation between the age of the claimant and claim severity.\n")
cat("Alternative Hypothesis: There is a significant correlation between age and claim severity.\n")

# Display Pearson correlation test results
cat("\nPearson Correlation Test Results for Age and Claim Severity:\n")
print(correlation_result)

# Interpret the results
cat("\nInterpretation:\n")
if (correlation_result$p.value < 0.05) {
  cat("Reject the null hypothesis. There is sufficient evidence of a statistically significant correlation
} else {
  cat("Fail to reject the null hypothesis. No sufficient evidence of a significant correlation between age
}
```

The Pearson correlation test is performed using the corr.test() function, which determines whether the p-value is less than the standard significance level of 0.05. A p-value of less than 0.05 indicates that there is enough evidence to reject the null hypothesis, showing a statistically significant link. If the p-value is 0.05 or above, it means that there is insufficient evidence to reject the null hypothesis.

Result:

```
Null Hypothesis: There is no correlation between the age of the claimant and claim severity.
> cat("Alternative Hypothesis: There is a significant correlation between age and claim severity.\n")
Alternative Hypothesis: There is a significant correlation between age and claim severity.
>
> # Display Pearson correlation test results
> cat("\nPearson Correlation Test Results for Age and Claim Severity:\n")

Pearson Correlation Test Results for Age and Claim Severity:
> print(correlation_result)

        Pearson's product-moment correlation

data:  data$Age and data$Severity
t = -15.186, df = 79208, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06081960 -0.04693202
sample estimates:
       cor
-0.05387842

>
> # Interpret the results
> cat("\nInterpretation:\n")

Interpretation:
> if (correlation_result$p.value < 0.05) {
+   cat("Reject the null hypothesis. There is sufficient evidence of a statistically significant correlation between the ag
e of the claimant and the severity of the claim.\n")
+ } else {
+   cat("Fail to reject the null hypothesis. No sufficient evidence of a significant correlation between age and claim seve
rity.\n")
+ }
Reject the null hypothesis. There is sufficient evidence of a statistically significant correlation between the age of the
claimant and the severity of the claim.
>
```

Pearson Correlation Test findings:

Pearson's correlation coefficient (cor) is roughly -0.0539 in the correlation test findings. This negative number indicates that there is a weak negative linear relationship between the claimant's age and the severity of the claim.

The t-statistic is -15.186, and there are 79208 degrees of freedom (df).

The extremely low p-value (< 2.2e-16) indicates strong evidence against the null hypothesis.

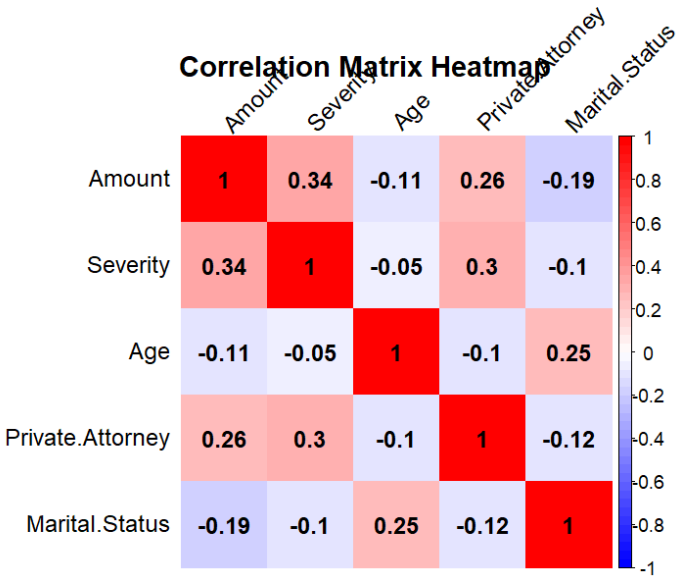
Interpretation:

Because the p-value is less than 0.05, there is enough evidence to reject the null hypothesis.

As a result, it is possible to conclude that there is a statistically significant negative link between the claimant's age and the severity of the claim.

CORRELATION MATRIX

The correlation matrix represents the level and direction of the correlations between variables. The variables Amount and Severity have a moderate positive connection (0.34), indicating that bigger claim amounts are related to higher severity levels. Age has a small negative association (-0.11) with Amount, indicating that claim amounts drop with advancing age. Private Attorney representation has a somewhat positive association (0.26) with Amount, meaning that cases involving private attorneys had greater claim amounts. Marital Status has a weak negative association (-0.19) with Amount, implying that some marital statuses have a modest decrease in claim amounts.



INFERENTIAL STATISTICS AND REGRESSION TESTING

Question 1: *Relationship between the claim payment amount and the age of the claimant?*

- **Dependent Variable**: Amount of the claim payment
- **Independent Variable**: Age of the claimant

```r
# Independent variable: Age of the Claimant
# Convert Amount to millions for better readability
sampled_subset$Amount_million <- sampled_subset$Amount / 1e6

# Identify and remove outliers using IQR
Q1 <- quantile(sampled_subset$Amount_million, 0.25)
Q3 <- quantile(sampled_subset$Amount_million, 0.75)
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR
sampled_subset <- sampled_subset[sampled_subset$Amount_million >= lower_bound & sampled_s
# If there are cases after removing outliers, perform linear regression and generate a sc
if (nrow(sampled_subset) > 0) {
# Perform linear regression on the sampled subset
  model_another <- lm(Amount_million ~ Age, data = sampled_subset)

# Summary of the model
  summary(model_another)
# Scatterplot
  ggplot(sampled_subset, aes(x = Age, y = Amount_million)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    labs(title = "Scatterplot of Amount vs. Age",
         x = "Age of the Claimant",
         y = "Amount of Claim Payment (Millions)")
} else {
  cat("No cases found for the chosen subset after removing outliers.")
}
p_value <- coef(summary(model_another))[, "Pr(>|t|)"][2]
# Display the p-value
cat("P-value for the Age coefficient:", p_value, "\n")
# Check for significance (common threshold is 0.05)
if (p_value < 0.05) {
  cat("The relationship between age and amount is statistically significant.\n")
} else {
  cat("There is no statistically significant relationship between age and amount.\n")
}
```

Code explanation

- Removing Outliers: Removing outliers by using IQR

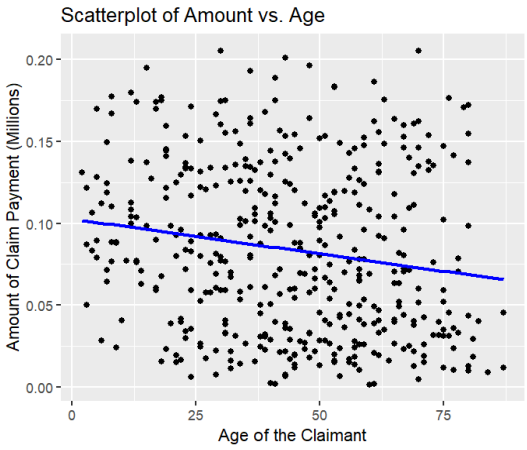
  Calculated the first quartile (Q1), third quartile (Q3), and interquartile range (IQR). Defined lower and upper bounds based on the IQR to identify outliers. Subset the data to include only cases within the determined bounds.

- Linear Regression and Scatterplot:

  Once the outliers are removed perform the linear regression and scatterplot with the sampled data. Generated a scatterplot of Amount vs. Age with a blue regression line.

- Hypothesis Testing and Significance Check:

  Extract the p-value for the Age coefficient from the linear regression model, display the p-value, and check for statistical significance using a common threshold of 0.05.




```
Call:
lm(formula = Amount_million ~ Age, data = sampled_subset)

Residuals:
      Min        1Q    Median        3Q       Max
-0.086222 -0.040496 -0.008408  0.039079  0.132292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1026718  0.0057926  17.725  < 2e-16 ***
Age         -0.0004253  0.0001194  -3.563 0.000407 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05044 on 430 degrees of freedom
Multiple R-squared:  0.02868,   Adjusted R-squared:  0.02642
F-statistic:  12.7 on 1 and 430 DF,  p-value: 0.0004072

P-value for the Age coefficient: 0.0004072384
The relationship between age and amount is statistically significant.
```

- Coefficients: Provides the estimates for the model parameters (intercept and age). Each coefficient has an estimate, standard error, t-value, and p-value. The intercept (0.1026718) represents the estimated claim payment amount when the age is 0. The coefficient for age (-0.0004253) represents the estimated change in claim payment amount for a one-unit increase in age.
- Significance of Coefficients: The p-values associated with each coefficient are used to test the null hypothesis that the true value of the coefficient is zero. For the intercept and age, the p-values are very small (<< 0.05), indicating that both are statistically significant.
- Residuals and Model Fit: Residual standard error: Represents the standard deviation of the residuals, indicating the average magnitude of the prediction errors.
- Multiple R-squared: Measures the proportion of variance in the dependent variable (claim payment amount) explained by the independent variable (age). In this case, it's relatively low at 0.02868, suggesting that age explains only a small portion of the variation in the claim payment amount.
- Adjusted R-squared: Similar to R-squared but adjusted for the number of predictors. It's also low, indicating that the model may not be a great fit.
- Hypothesis Testing: The p-value for the age coefficient is 0.0004072, which is less than the common significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a statistically significant relationship between age and claim payment amount.

- Scatterplot: A scatterplot with a blue regression line is generated to visualize the relationship between claim payment amount and age.

Question 2: *Is there a significant relationship between the claim amount (dependent variable) and the age of the claimant (continuous numerical independent variable) based on their insurance type (categorical independent variable)?*

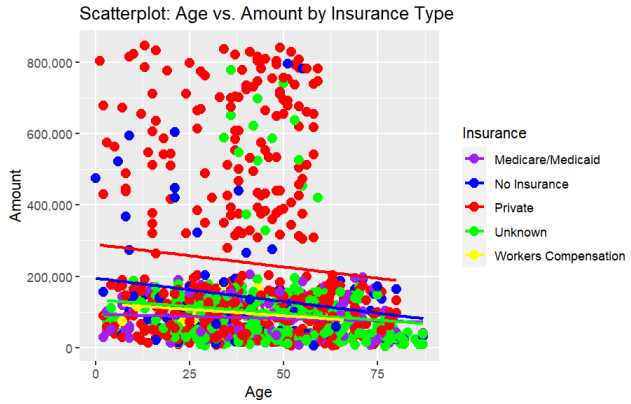
- **Dependent Variable**: Amount of the claim payment
- **Independent Variable**: The age of the claimant based on their insurance type

```r
#Question 2: Is there a significant relationship between the claim amount (dependent variabl
#Dependent Variable:Amount of the claim payment
#Independent Variable:age of the claimant based on their insurance type
quantile_threshold_q4 <- quantile(data$Amount, c(0.01, 0.99))
data_subset_q4 <- subset(data, Amount >= quantile_threshold_q4[1] & Amount <= quantile_thres
ggplot(data_subset_q4_sampled, aes(x = Age, y = Amount, color = Insurance)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE, aes(group = Insurance), show.legend = TRUE) +
  ggtitle("Scatterplot: Age vs. Amount by Insurance Type") +
  scale_y_continuous(labels = scales::comma) +
  scale_color_manual(values = c("Private" = "red", "No Insurance" = "blue", "Unknown" = "gre

# Linear Regression Model
model4_q4 <- lm(Amount ~ Age * Insurance, data = data_subset_q4)
# Hypothesis Testing - Display summary output
summary(model4_q4)
# Extract p-values
p_values <- coef(summary(model4_q4))[, "Pr(>|t|)"]
# Display p-values
cat("P-values for each coefficient:\n")
print(p_values)
# Check for significance (common threshold is 0.05)
significant_terms <- names(p_values[p_values < 0.05])
if (length(significant_terms) > 0) {
  cat("Significant terms:", paste(significant_terms, collapse = ", "), "\n")
} else {
  cat("No statistically significant terms found.\n")
}
```

Question 2 initiates by identifying percentiles of claim amounts, creating a subset based on these percentiles, and then generating a scatterplot to visually inspect the relationship between claim amounts and claimants' ages, color-coded by insurance type. Subsequently, a linear regression model is constructed to quantify this relationship, considering the interaction effect between age and insurance type. The model's summary provides statistical insights into the significance of age, insurance type, and their interaction in predicting claim amounts, offering a comprehensive analysis of the relationships among these variables.

Result:




```
Call:
lm(formula = Amount ~ Age * Insurance, data = data_subset_q4_sampled)

Residuals:
    Min      1Q  Median      3Q     Max
-258737 -106005  -48459   32931  671484

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          93669.4    37982.9   2.466   0.0138 *
Age                                   -248.5      749.9  -0.331   0.7404
InsuranceNo Insurance               101127.3    54091.4   1.870   0.0618 .
InsurancePrivate                    195666.9    43295.4   4.519 6.95e-06 ***
InsuranceUnknown                     40913.6    47807.9   0.856   0.3923
InsuranceWorkers Compensation        30915.9   116972.6   0.264   0.7916
Age:InsuranceNo Insurance            -1052.2     1147.0  -0.917   0.3592
Age:InsurancePrivate                 -1011.4      885.8  -1.142   0.2538
Age:InsuranceUnknown                  -515.0      944.5  -0.545   0.5857
Age:InsuranceWorkers Compensation     -452.7     2768.7  -0.164   0.8701
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 187500 on 990 degrees of freedom
Multiple R-squared:  0.1294,    Adjusted R-squared:  0.1215
F-statistic: 16.36 on 9 and 990 DF,  p-value: < 2.2e-16
```

```
P-values for each coefficient:
> print(p_values)
                  (Intercept)                           Age
                1.782810e-162                  7.033511e-05
          InsuranceNo Insurance                InsurancePrivate
                 1.451757e-76                 1.155126e-314
               InsuranceUnknown     InsuranceWorkers Compensation
                  5.988414e-05                  9.352375e-01
      Age:InsuranceNo Insurance            Age:InsurancePrivate
                  9.442431e-30                  2.661981e-20
           Age:InsuranceUnknown Age:InsuranceWorkers Compensation
                  4.497417e-01                  3.145480e-01
> # Check for significance (common threshold is 0.05)
> significant_terms <- names(p_values[p_values < 0.05])
> if (length(significant_terms) > 0) {
+   cat("Significant terms:", paste(significant_terms, collapse = ", "), "\n")
+ } else {
+   cat("No statistically significant terms found.\n")
+ }
Significant terms: (Intercept), Age, InsuranceNo Insurance, InsurancePrivate, InsuranceUnknown, Ag
e:InsuranceNo Insurance, Age:InsurancePrivate
>
```

The linear regression model examines the relationship between the claim amount (dependent variable) and the age of the claimant, considering different insurance types (categorical independent variable) and their interaction effects.

- Intercept: The intercept is $93,669.4, representing the estimated claim amount when the age is zero, and the insurance type is "Medicare/Medicaid."
- Age: The coefficient for age is -248.5, but it is not statistically significant (p = 0.7404), suggesting that age alone does not have a significant impact on the claim amount.
- Insurance Categories:
  - ❖ No Insurance: The coefficient is positive (101,127.3) but marginally significant (p = 0.0618), suggesting that individuals with no insurance tend to have higher claim amounts.
  - ❖ Private Insurance: The coefficient is significantly positive (195,666.9, p < 0.001), indicating that individuals with private insurance have significantly higher claim amounts compared to the reference category.

- ❖ Unknown Insurance: The coefficient is not statistically significant (p = 0.3923), suggesting that the type of insurance is not a significant predictor of claim amounts.

- Interaction Effects: The coefficients for the interaction terms (e.g., Age:InsurancePrivate) represent how the relationship between age and claim amount changes based on the insurance category. None of the interaction terms are statistically significant.
- Overall Model Fit: The adjusted R-squared is 0.1215, indicating that the model explains about 12.15% of the variance in claim amounts. The p-value for the overall model is highly significant (p < 2.2e-16), suggesting that at least one of the predictors significantly contributes to explaining the variance in claim amounts.

Question 3: Is there a significant relationship between the claim amount (dependent variable) and the age of the claimant (continuous numerical independent variable) based on their marital status (categorical independent variable)?

- **Dependent Variable: Amount of the claim payment**
- **Independent Variable: Age of the claimant based on marital status**

Code

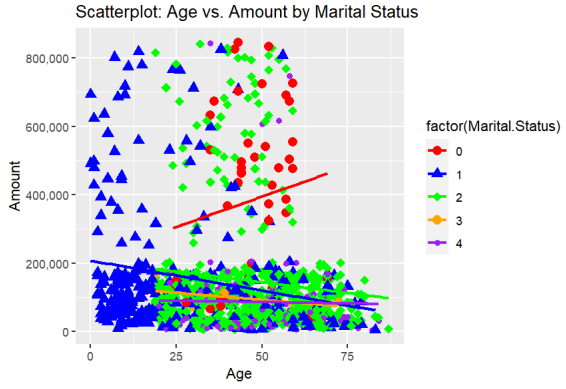
```
#Question 3: Is there a significant relationship between the claim amount (dependent variable) and the age of the claimant (continuo
# Dependent Variable: Amount of the claim payment
# Independent Variable: Age of the claimant based on marital status
# Set seed for reproducibility
set.seed(123)
# Define the quantile threshold for outliers
quantile_threshold_q3 <- quantile(data$Amount, c(0.01, 0.99))
# Create a subset without outliers
data_subset_q3 <- subset(data, Amount >= quantile_threshold_q3[1] & Amount <= quantile_threshold_q3[2])
# Create a random sample for the scatterplot
sample_indices_q3 <- sample(seq_len(nrow(data_subset_q3)), size = 1000)  # Adjust the sample size as needed
data_subset_q3_sampled <- data_subset_q3[sample_indices_q3, ]
# Scatterplot with Regression Lines and Legend
ggplot(data_subset_q3_sampled, aes(x = Age, y = Amount, color = factor(Marital.Status))) +
  geom_point(aes(shape = factor(Marital.Status)), size = 3) +  # Use different shapes for each point based on Marital Status
  geom_smooth(method = "lm", se = FALSE, aes(group = factor(Marital.Status)), show.legend = TRUE) +  # Add regression lines for each
  ggtitle("Scatterplot: Age vs. Amount by Marital Status") +
  scale_y_continuous(labels = scales::comma) +
  scale_color_manual(values = c("0" = "red", "1" = "blue", "2" = "green", "3" = "orange", "4" = "purple")) +
  scale_shape_manual(values = c("0" = 16, "1" = 17, "2" = 18, "3" = 19, "4" = 20))  # Use different shapes for each Marital Status
# Linear Regression
model3_q3_sampled <- lm(Amount ~ Age * factor(Marital.Status), data = data_subset_q3_sampled)
# Hypothesis Testing
summary(model3_q3_sampled)
:
```

```
# Extract p-values for each coefficient
p_values <- coef(summary(model3_q3_sampled))[, "Pr(>|t|)"]

# Display p-values
cat("P-values for each coefficient:\n")
print(p_values)
# Specify the null and alternative hypotheses
cat("\nNull Hypothesis: There is no significant relationship between claim amount and the age
cat("Alternative Hypothesis: There is a significant relationship between claim amount and the
# Check for significance (common threshold is 0.05)
significant_terms <- names(p_values[p_values < 0.05])
if (length(significant_terms) > 0) {
  cat("\nSignificant terms:", paste(significant_terms, collapse = ", "), "\n")
} else {
  cat("\nNo statistically significant terms found.\n")
}
```

The code aims to examine the relationship between claim amounts (the dependent variable) and claimants' ages based on their marital status (the categorical independent variable). It starts by removing outliers from the data and then takes a random sample for the scatterplot. The scatterplot is made up of various shapes and colors to indicate different marital statuses and regression lines are added for each category. After that, a linear regression model is fitted to the sampled data, taking into account the interaction effects of age and married status. The model summary includes coefficients, standard errors, t-values, and p-values, which allow for hypothesis testing.

Output:




```
Call:
lm(formula = Amount ~ Age * factor(Marital.Status), data = data_subset_q3_sampled)

Residuals:
    Min      1Q  Median      3Q     Max
-404625  -90932  -47109   13925  755233

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                       217838     116684   1.867   0.0622 .
Age                                 3540       2421   1.462   0.1441
factor(Marital.Status)1           -10518     117834  -0.089   0.9289
factor(Marital.Status)2           -10768     119067  -0.090   0.9280
factor(Marital.Status)3           -84400     193409  -0.436   0.6627
factor(Marital.Status)4          -124364     128022  -0.971   0.3316
Age:factor(Marital.Status)1        -5292       2466  -2.146   0.0322 *
Age:factor(Marital.Status)2        -4801       2468  -1.945   0.0520 .
Age:factor(Marital.Status)3        -4360       3982  -1.095   0.2738
Age:factor(Marital.Status)4        -3693       2634  -1.402   0.1611
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 173700 on 990 degrees of freedom
Multiple R-squared:  0.1136,    Adjusted R-squared:  0.1055
F-statistic:  14.1 on 9 and 990 DF,  p-value: < 2.2e-16
```

```
P-values for each coefficient:
> print(p_values)
                (Intercept)                         Age      factor(Marital.Status)1
                 0.06220981                  0.14406681                  0.92888898
    factor(Marital.Status)2     factor(Marital.Status)3     factor(Marital.Status)4
                 0.92795984                  0.66265707                  0.33157322
Age:factor(Marital.Status)1 Age:factor(Marital.Status)2 Age:factor(Marital.Status)3
                 0.03215537                  0.05202610                  0.27376307
Age:factor(Marital.Status)4
                 0.16111253
> # Specify the null and alternative hypotheses
> cat("\nNull Hypothesis: There is no significant relationship between claim amount and the age of
the claimant based on marital status.\n")

Null Hypothesis: There is no significant relationship between claim amount and the age of the clai
ant based on marital status.
> cat("Alternative Hypothesis: There is a significant relationship between claim amount and the ag
of the claimant based on marital status.\n")
Alternative Hypothesis: There is a significant relationship between claim amount and the age of th
claimant based on marital status.
> # Check for significance (common threshold is 0.05)
> significant_terms <- names(p_values[p_values < 0.05])
> if (length(significant_terms) > 0) {
+   cat("\nSignificant terms:", paste(significant_terms, collapse = ", "), "\n")
+ } else {
+   cat("\nNo statistically significant terms found.\n")
+ }

Significant terms: Age:factor(Marital.Status)1
```

The code aims to explore the relationship between claim amounts (dependent variable) and the age of claimants based on their marital status (categorical independent variable).

- Residuals and Coefficients: The residuals (differences between observed and predicted values) have a range from -404,625 to 755,233. The coefficients provide estimates for the intercept, age, and the interaction terms between age and each level of marital status.
- Intercept and Age: The intercept is 217,838, representing the estimated claim amount when all other variables are zero. The coefficient for age is 3,540, indicating the expected change in claim amount for each additional year of age.
- Marital Status: The coefficients for marital status levels (1 to 4) represent the estimated difference in claim amounts compared to the reference level (Marital Status 0). None of the marital status coefficients are statistically significant (p-values > 0.05), suggesting no clear evidence of a relationship between marital status and claim amounts.
- The significant interaction term is for Age: Marital.Status1, with a p-value of 0.0322. This suggests that the relationship between age and claim amounts differs for Marital Status 1 compared to the reference level.
- Model Fit: The multiple R-squared value is 0.1136, indicating that the model explains about 11.36% of the variance in claim amounts. The F-statistic tests the overall significance of the model and is highly significant (p-value < 2.2e-16), suggesting that at least one predictor variable has a significant effect on claim amounts.

CONCLUSION

In this analysis of a US medical malpractice dataset, key findings include an average claim amount of $157,484 with a wide range of severity ratings and an average claimant age of 42.7 years. In 66% of instances, private attorneys are involved. Cross-tabulations indicated patterns such as a connection between marital status and attorney representation, as well as gender disparities in severity. Categorical, discrete, and continuous variables were the subject of subset analyses. In particular, inferential statistics and hypothesis testing were used to examine the correlations between claim amounts and characteristics such as age, insurance type, and marital status.

CITATION

1. *medical insurance and malpractice.* (2023, October 9). Kaggle. https://www.kaggle.com/datasets/willianoliveiragibin/mendical-insurance-and-malpratice/data
2. Team, D. (2020, June 26). *Subsetting in R tutorial*. https://www.datacamp.com/tutorial/subsets-in-r
3. DataScienceTutor. (2022, May 13). *How to perform the Kruskal-Wallis test in R? | R-bloggers*. R-bloggers. https://www.r-bloggers.com/2022/05/how-to-perform-the-kruskal-wallis-test-in-r/
4. Turney, S. (2023, June 22). *Pearson Correlation Coefficient (r) | Guide & Examples*. Scribbr. https://www.scribbr.com/statistics/pearson-correlation-coefficient/
5. Datanovia. (2019, December 26). *How To Do Two-Sample T-test in R : Best Tutorial You Will Love - Datanovia*. https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/how-to-do-two-sample-t-test-in-r/