KV6015-DataAnalyticsRAssessmentAnswer

Rayhan Chowdhury Bijoy

2024-12-05

library(readxl)  
library(tibble)  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2)  
library(scales)  
library(tidyr)

# Load the data

# Load the data from the Excel file  
data <- read\_excel("./PreparedData.xlsx")  
  
# Convert the data to a tibble  
data\_tibble <- as\_tibble(data)  
  
# View the tibble (first few rows)  
print(data\_tibble)

# A tibble: 6 × 10  
 Date `Male tests-Conducted` `Male tests-Passes` `Male tests-Pass rate (%)`  
 <chr> <dbl> <dbl> <dbl>  
1 2018/19 46098 20393 44.2  
2 2019/20 53500 23994 44.8  
3 2020/21 18938 8655 45.7  
4 2021/22 83086 36848 44.3  
5 2022/23 116349 52309 45.0  
6 2023/24 173766 77497 44.6  
# ℹ 6 more variables: `Female tests-Conducted` <dbl>,  
# `Female tests-Passes` <dbl>, `Female tests-Pass rate (%)` <dbl>,  
# `Total tests-Conducted` <dbl>, `Total tests-Passes` <dbl>,  
# `Total tests-Pass rate (%)` <dbl>

#———————–1. Hypothesis 01 (Gender-based Focus): Linear Model Test———————–

# Create a 'Period' column based on the 'Date'  
data\_tibble <- data\_tibble %>%  
 mutate(Period = case\_when(  
 Date %in% c("2018/19", "2019/20") ~ "Pre-pandemic",  
 Date %in% c("2020/21", "2021/22") ~ "Pandemic",  
 Date %in% c("2022/23", "2023/24") ~ "Post-pandemic",  
 TRUE ~ "Unknown"  
 )) %>%  
 mutate(Period = factor(Period, levels = c("Pre-pandemic", "Pandemic", "Post-pandemic")))  
  
# Calculate the reductions for Male and Female candidates by period  
data\_tibble <- data\_tibble %>%  
 arrange(Date) %>%  
 group\_by(Period) %>%  
 mutate(  
 Male\_tests\_Reduction = c(NA, diff(`Male tests-Conducted`)),  
 Female\_tests\_Reduction = c(NA, diff(`Female tests-Conducted`))  
 )  
  
# Remove rows with NA values (first row of each period has NA for reduction)  
data\_tibble <- data\_tibble %>%  
 filter(!is.na(Male\_tests\_Reduction), !is.na(Female\_tests\_Reduction))  
  
# Fit the linear model for Male tests reduction  
model\_male <- lm(Male\_tests\_Reduction ~ Period, data = data\_tibble)  
summary(model\_male)

Call:  
lm(formula = Male\_tests\_Reduction ~ Period, data = data\_tibble)  
  
Residuals:  
ALL 3 residuals are 0: no residual degrees of freedom!  
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 7402 NaN NaN NaN  
PeriodPandemic 56746 NaN NaN NaN  
PeriodPost-pandemic 50015 NaN NaN NaN  
  
Residual standard error: NaN on 0 degrees of freedom  
Multiple R-squared: 1, Adjusted R-squared: NaN   
F-statistic: NaN on 2 and 0 DF, p-value: NA

# Fit the linear model for Female tests reduction  
model\_female <- lm(Female\_tests\_Reduction ~ Period, data = data\_tibble)  
summary(model\_female)

Call:  
lm(formula = Female\_tests\_Reduction ~ Period, data = data\_tibble)  
  
Residuals:  
ALL 3 residuals are 0: no residual degrees of freedom!  
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 10051 NaN NaN NaN  
PeriodPandemic 108219 NaN NaN NaN  
PeriodPost-pandemic 63814 NaN NaN NaN  
  
Residual standard error: NaN on 0 degrees of freedom  
Multiple R-squared: 1, Adjusted R-squared: NaN   
F-statistic: NaN on 2 and 0 DF, p-value: NA

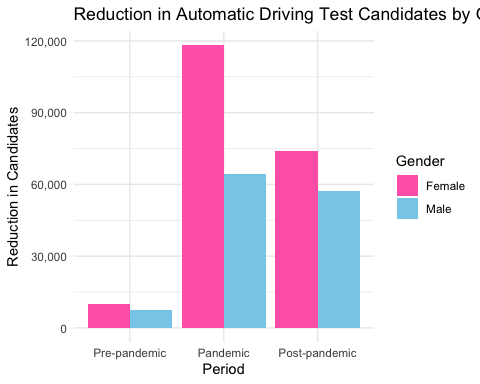
#———————–1. Hypothesis 1 (Gender-based Focus): Chi-Square Test———————–

# Create a contingency table for the Chi-square test  
# Summarize reductions by gender and period  
data\_tibble\_summary <- data\_tibble %>%  
 group\_by(Period) %>%  
 summarise(  
 Male\_Reduction = sum(Male\_tests\_Reduction, na.rm = TRUE),  
 Female\_Reduction = sum(Female\_tests\_Reduction, na.rm = TRUE)  
 )  
  
# Convert reductions into a matrix for the Chi-square test  
reduction\_matrix <- as.matrix(data\_tibble\_summary[, c("Male\_Reduction", "Female\_Reduction")])  
rownames(reduction\_matrix) <- data\_tibble\_summary$Period  
  
# Perform the Chi-square test  
chi\_square\_result <- chisq.test(reduction\_matrix)  
  
# Output the test result  
print(chi\_square\_result)

Pearson's Chi-squared test  
  
data: reduction\_matrix  
X-squared = 2451.2, df = 2, p-value < 2.2e-16

# Visulize the Hypothesis 01 - Chi-Square Data

# Create a dataset for plotting by summarizing Male and Female reductions by period and reshaping the data  
ggplot(data\_tibble %>%  
 group\_by(Period) %>%  
 summarise(  
 Male\_Reduction = sum(Male\_tests\_Reduction, na.rm = TRUE),  
 Female\_Reduction = sum(Female\_tests\_Reduction, na.rm = TRUE)   
 ) %>%   
 pivot\_longer(cols = c(Male\_Reduction, Female\_Reduction), names\_to = "Gender", values\_to = "Reduction") %>%  
 mutate(Gender = recode(Gender, "Male\_Reduction" = "Male", "Female\_Reduction" = "Female")),   
 aes(x = Period, y = Reduction, fill = Gender)) +  
 geom\_bar(stat = "identity", position = "dodge") + # Create a dodged bar plot  
 scale\_fill\_manual(values = c("Male" = "skyblue", "Female" = "hotpink")) + # Correct color mapping  
 labs(  
 title = "Reduction in Automatic Driving Test Candidates by Gender and Period",  
 x = "Period",  
 y = "Reduction in Candidates",  
 fill = "Gender"  
 ) +  
 theme\_minimal() + # Apply a minimal theme  
 scale\_y\_continuous(labels = scales::comma) # Format y-axis with commas for better readability



# Hypothesis 01 Discussion - Overall Summary, Linear Modeal VS Chi-Square

The analysis supports Hypothesis 1, showing significant differences in the reduction of driving test candidates between males and females during the COVID-19 pandemic. The Chi-square test revealed a highly significant difference in reductions across gender and periods (p-value < 2.2e-16), indicating that gender-based impacts were not equal. The linear regression models also indicated reductions in both male and female candidates, but due to limited data and perfect model fit, the Chi-square test was more reliable for confirming the significant difference between genders. Thus, the Chi-square test is the more robust method for assessing gender-based differences in reductions.

#———————–2. Hypothesis 2 (Year-based Focus): Linear Model Test———————–

# Load the data function  
load\_data <- function() {  
 data <- read\_excel("./PreparedData.xlsx")  
 as\_tibble(data)  
}  
  
# Reload the data before Hypothesis 2 analysis  
data\_tibble <- load\_data()  
  
# Hypothesis 2: Year-based Focus Analysis  
# Recreate Period column to ensure consistency  
data\_tibble <- data\_tibble %>%  
 mutate(Period = case\_when(  
 Date %in% c("2018/19", "2019/20") ~ "Pre-pandemic",  
 Date %in% c("2020/21", "2021/22") ~ "Pandemic",  
 Date %in% c("2022/23", "2023/24") ~ "Post-pandemic",  
 TRUE ~ "Unknown"  
 )) %>%  
 mutate(Period = factor(Period, levels = c("Pre-pandemic", "Pandemic", "Post-pandemic")))  
  
# Calculate the reduction in 'Total tests-Conducted' for each period  
data\_tibble <- data\_tibble %>%  
 arrange(Date) %>%  
 group\_by(Period, .drop = FALSE) %>% # Retain unused levels  
 mutate(Test\_Reduction = c(NA, diff(`Total tests-Conducted`))) %>%  
 filter(!is.na(Test\_Reduction)) # Remove rows with NA values  
  
# Check levels after filtering  
print(levels(data\_tibble$Period))

[1] "Pre-pandemic" "Pandemic" "Post-pandemic"

print(table(data\_tibble$Period))

Pre-pandemic Pandemic Post-pandemic   
 1 1 1

# Fit the linear model to examine the reduction in tests by period  
model\_total\_tests <- lm(Test\_Reduction ~ Period, data = data\_tibble)  
  
# Output the summary of the linear model  
summary(model\_total\_tests)

Call:  
lm(formula = Test\_Reduction ~ Period, data = data\_tibble)  
  
Residuals:  
ALL 3 residuals are 0: no residual degrees of freedom!  
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 17463 NaN NaN NaN  
PeriodPandemic 165041 NaN NaN NaN  
PeriodPost-pandemic 113749 NaN NaN NaN  
  
Residual standard error: NaN on 0 degrees of freedom  
Multiple R-squared: 1, Adjusted R-squared: NaN   
F-statistic: NaN on 2 and 0 DF, p-value: NA

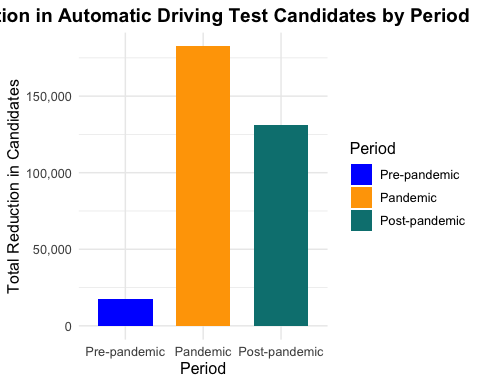
#———————–2. Hypothesis 2 (Year-based Focus): Chi-Square Test———————–

# Reload the data to ensure a clean slate  
data\_tibble <- load\_data()  
  
# Recreate Period column  
data\_tibble <- data\_tibble %>%  
 mutate(Period = case\_when(  
 Date %in% c("2018/19", "2019/20") ~ "Pre-pandemic",  
 Date %in% c("2020/21", "2021/22") ~ "Pandemic",  
 Date %in% c("2022/23", "2023/24") ~ "Post-pandemic",  
 TRUE ~ "Unknown"  
 )) %>%  
 mutate(Period = factor(Period, levels = c("Pre-pandemic", "Pandemic", "Post-pandemic")))  
  
# Calculate the reductions in 'Total tests-Conducted' for each period  
data\_tibble <- data\_tibble %>%  
 arrange(Date) %>%  
 group\_by(Period, .drop = FALSE) %>%  
 mutate(Test\_Reduction = c(NA, diff(`Total tests-Conducted`))) %>%  
 filter(!is.na(Test\_Reduction)) # Remove rows with NA values  
  
# Summarize reductions by period  
data\_tibble\_summary <- data\_tibble %>%  
 group\_by(Period) %>%  
 summarise(Total\_Reduction = sum(Test\_Reduction, na.rm = TRUE))  
  
# Create a contingency table for Chi-Square Test  
reduction\_matrix <- as.matrix(data\_tibble\_summary[, "Total\_Reduction"])  
rownames(reduction\_matrix) <- data\_tibble\_summary$Period  
  
# Perform the Chi-Square Test  
chi\_square\_result <- chisq.test(reduction\_matrix)  
  
# Output the test result  
print(chi\_square\_result)

Chi-squared test for given probabilities  
  
data: reduction\_matrix  
X-squared = 129260, df = 2, p-value < 2.2e-16

# Visulize the Hypothesis 02 - Chi-Square Data

# Visualization of Total Reductions by Period for Hypothesis 2  
ggplot(data\_tibble\_summary, aes(x = Period, y = Total\_Reduction, fill = Period)) +  
 geom\_bar(stat = "identity", width = 0.7) +  
 scale\_fill\_manual(values = c("Pre-pandemic" = "blue", "Pandemic" = "#FFA500", "Post-pandemic" = "#008080")) +  
 labs(  
 title = "Reduction in Automatic Driving Test Candidates by Period",  
 x = "Period",  
 y = "Total Reduction in Candidates",  
 fill = "Period"  
 ) +  
 theme\_minimal() +  
 theme(  
 text = element\_text(size = 12),  
 plot.title = element\_text(hjust = 0.5, face = "bold")  
 ) +  
 scale\_y\_continuous(labels = scales::comma) # Format y-axis with commas for readability



# Hypothesis 02 Discussion - Overall Summary, Linear Modeal VS Chi-Square

The analysis supports Hypothesis 2, showing significant differences in the reduction of driving test candidates across the pre-pandemic, pandemic, and post-pandemic periods. The Chi-Square Test revealed a highly significant difference in reductions across these periods (p-value < 2.2e-16), indicating that the impact of the pandemic on driving test candidates was not consistent with the other periods.

The linear regression model attempted to quantify the reductions but was limited by the small sample size, resulting in perfect model fit and indeterminate p-values due to zero degrees of freedom. While the regression model highlighted noticeable differences in reductions across periods, the Chi-Square Test was more robust in confirming these differences.

Thus, the Chi-Square Test provides stronger evidence supporting Hypothesis 2, emphasizing that the reduction in driving test candidates during the pandemic was significantly different from both the pre-pandemic and post-pandemic periods.