GMV Statistical Analysis

0. loading pacakge

```r
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.3

Warning: package 'readr' was built under R version 4.3.3

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(PMCMRplus)
```

```
Warning: package 'PMCMRplus' was built under R version 4.3.3
```

```r
library(stats)
```

1. Income level

```r
data <- data.frame(
  IncomeRange = factor(rep(c("$1 - $60,000 (Low)", "$60,000 - $100,000 (Low-mid)", "$100,001 - $150
  GMVTier = factor(c("< $10,000", "$10,000 to $29,999", "$30,000 to $49,999", "$50,000 to $99,999"
                   levels = c("< $10,000", "$10,000 to $29,999", "$30,000 to $49,999", "$50,000 to
  Restaurants = c(197, 483, 178, 52, 4, 357, 782, 247, 59, 3, 122, 254, 60, 13, 1, 20, 23, 5, 2, 0)
)


reshape_data <- xtabs(Restaurants ~ IncomeRange + GMVTier, data = data)

# step 3: Run the Chi-Square test
chi_results <- chisq.test(reshape_data, simulate.p.value = TRUE, B = 2000)

# Print the results
print(chi_results)
```

```
	Pearson's Chi-squared test with simulated p-value (based on 2000
	replicates)

data:  reshape_data
X-squared = 26.397, df = NA, p-value = 0.01499
```

2. Population Density

```r
  client_data <- matrix(c(
    91, 232, 84, 8, 2,    # Rural
    189, 471, 159, 45, 1, # Rural/Suburban
    450, 900, 275, 91, 4, # Suburban
    93, 190, 66, 17, 1    # Urban
  ), nrow = 4, byrow = TRUE)

  # Define row and column names for clarity
  rownames(client_data) <- c("Rural", "Rural/Suburban", "Suburban", "Urban")
  colnames(client_data) <- c("< $10,000", "$10,000 to $29,999", "$30,000 to $49,999", "> $50,000",

  # Conduct the Chi-squared test
  chi_results <- chisq.test(client_data, simulate.p.value = TRUE, B = 5000)

  # Print the results
  print(chi_results)
```

```
	Pearson's Chi-squared test with simulated p-value (based on 5000
	replicates)
```

```
data:  client_data
X-squared = 21.211, df = NA, p-value = 0.04539
```

3. Region

```r
regions <- c('Northeast', 'Southeast', 'Midwest', 'Southwest', 'West')
icp_sam <- c(15815, 14859, 8512, 6232, 15065)
avg_gmv <- c(20839, 20695, 17556, 13318, 12994)
num_clients <- c(1384, 1151, 609, 115, 120)

regions_numeric <- c(1, 2, 3, 4, 5)  # From Northeast to West

lm_icp <- lm(icp_sam ~ regions_numeric)
lm_gmv <- lm(avg_gmv ~ regions_numeric)
lm_clients <- lm(num_clients ~ regions_numeric)

summary(lm_icp)
```

```
Call:
lm(formula = icp_sam ~ regions_numeric)

Residuals:
    1     2     3     4     5 
 1693  1750 -3585 -4852  4994 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)  
(Intercept)        15135       4966   3.048   0.0555 .
regions_numeric    -1013       1497  -0.676   0.5473  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4735 on 3 degrees of freedom
Multiple R-squared:  0.1323,	Adjusted R-squared:  -0.1569 
F-statistic: 0.4575 on 1 and 3 DF,  p-value: 0.5473
```

```r
summary(lm_gmv)
```

```
Call:
lm(formula = avg_gmv ~ regions_numeric)

Residuals:
```

```
      1       2       3       4       5 
 -854.8  1307.9   475.6 -1455.7   527.0 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      24000.5     1362.7  17.613 0.000399 ***
regions_numeric  -2306.7      410.9  -5.614 0.011171 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1299 on 3 degrees of freedom
Multiple R-squared:  0.9131,    Adjusted R-squared:  0.8841 
F-statistic: 31.52 on 1 and 3 DF,  p-value: 0.01117
```

```r
summary(lm_clients)
```

```
Call:
lm(formula = num_clients ~ regions_numeric)

Residuals:
     1      2      3      4      5 
  -4.6  118.8  -66.8 -204.4  157.0 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)   
(Intercept)      1745.00     176.57   9.883   0.0022 **
regions_numeric  -356.40      53.24  -6.695   0.0068 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 168.4 on 3 degrees of freedom
Multiple R-squared:  0.9373,    Adjusted R-squared:  0.9163 
F-statistic: 44.82 on 1 and 3 DF,  p-value: 0.006799
```

##4. School Zone Rating

```r
data <- matrix(c(253, 497, 128, 35, 2,
                 401, 890, 283, 75, 5,
                 151, 363, 157, 42, 0,
                 18, 41, 16, 8, 1),
               nrow = 4, ncol = 5, byrow = TRUE)

rownames(data) <- c("A (Excellent)", "B (Good)", "C (Average)", "D (Poor)")
colnames(data) <- c("< $10,000", "$10,000 to $29,999", "$30,000 to $49,999", "$50,000 to $99,999",
```

```r
chisq_result <- chisq.test(data, simulate.p.value = TRUE, B = 2000)

print(chisq_result)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  data
X-squared = 36.74, df = NA, p-value = 0.003498
```

5. Cuisine

```r
GMV_data <- matrix(c(
    524, 1445, 521, 136, 8,   # Chinese
    110, 63, 11, 3, 0,        # Japanese
    6, 4, 3, 1, 0,            # Cajun
    18, 22, 6, 3, 0,          # American
    18, 20, 3, 1, 0,          # Korean
    3, 3, 0, 0, 0,            # Thai
    4, 6, 1, 1, 0             # Others
), byrow = TRUE, nrow = 5,
dimnames = list(
    c('< $10,000', '$10,000 to $29,999', '$30,000 to $49,999', '$50,000 to $99,999', '>$100,000'),
    c('Chinese', 'Japanese', 'Cajun', 'American', 'Korean', 'Thai', 'Others')
))

chi_result <- chisq.test(GMV_data, simulate.p.value = TRUE, B=2000)

print(chi_result)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  GMV_data
X-squared = 478.2, df = NA, p-value = 0.0004998
```

```r
expected <- chi_result$expected  # The expected frequencies
residuals <- (GMV_data - expected) / sqrt(expected)  # Standardized residuals

print(residuals)
```

```
                      Chinese  Japanese       Cajun   American    Korean
< $10,000           -0.2248717  1.185311 -0.06556244 -0.8940432 -2.376689
$10,000 to $29,999  2.4902951 -2.935241 -2.28297099  3.7474690  8.984546
$30,000 to $49,999 -3.5539801 -2.668917  2.69815719  1.3328592  3.904642
$50,000 to $99,999  6.0317740 -3.093647 -1.93992643 -1.2527144  6.377917
>$100,000          -1.5040706 -2.446992  1.18182600  6.7807350  3.315076
                         Thai     Others
< $10,000          -0.1472610 -1.7172357
$10,000 to $29,999  1.7890455  0.2492414
$30,000 to $49,999 -1.6317793 11.8103316
$50,000 to $99,999  1.6555885 -0.9141104
>$100,000           0.7574749 -0.5781342
```

6. Price Range

```r
# Input the data with all price ranges
GMV_data_all <- matrix(c(
    396, 1146, 427, 120, 7,      # $ restaurants
    279, 396, 111, 23, 0,        # $$ restaurants
    3, 4, 0, 0, 0,               # $$$ restaurants
    1, 1, 0, 0, 0                # $$$$ restaurants
), nrow = 5, byrow = TRUE,
dimnames = list(
    c('< $10,000', '$10,000 to $29,999', '$30,000 to $49,999', '$50,000 to $99,999', '>$100,000'),
    c('$', '$$', '$$$', '$$$$')
))

# View the data
GMV_data_all
```

```
                     $   $$ $$$ $$$$
< $10,000          396 1146 427  120
$10,000 to $29,999   7  279 396  111
$30,000 to $49,999  23    0   3    4
$50,000 to $99,999   0    0   0    1
>$100,000            1    0   0    0
```

```r
# Perform the Chi-square test
chisq_result_all <- chisq.test(GMV_data_all)
```

```
Warning in chisq.test(GMV_data_all): Chi-squared approximation may be incorrect
```

```r
chisq_result_all
```

```
    Pearson's Chi-squared test

data:  GMV_data_all
X-squared = 514.69, df = 12, p-value < 2.2e-16
```

```
chisq_result_all <- chisq.test(GMV_data_all, simulate.p.value = TRUE, B = 2000)

chisq_result_all
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  GMV_data_all
X-squared = 514.69, df = NA, p-value = 0.0004998
```

7. year in business

```
# Load necessary library
library(MASS)
```

```
Warning: package 'MASS' was built under R version 4.3.3


Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```
# Your data
data <- data.frame(
  FirstSeen = factor(rep(c("Pre-2020", "2020", "2021", "2022", "2023", "Now"), each = 5),
                     levels = c("Now", "2023", "2022", "2021", "2020", "Pre-2020")),
  GMVTier = factor(rep(c("< $10,000", "$10,000 to $29,999", "$30,000 to $49,999", "$50,000 to $99,9
                   levels = c("< $10,000", "$10,000 to $29,999", "$30,000 to $49,999", "$50,000 to
  NumClients = c(332, 1204, 472, 128, 7, 10, 17, 6, 1, 1, 19, 19, 7, 1, 0, 14, 19, 5, 1, 0, 24, 28
)

# Display the first few rows of the data
head(data)
```

```
  FirstSeen            GMVTier NumClients
1  Pre-2020           < $10,000        332
2  Pre-2020 $10,000 to $29,999       1204
3  Pre-2020 $30,000 to $49,999        472
4  Pre-2020 $50,000 to $99,999        128
5  Pre-2020           >$100,000          7
6      2020           < $10,000         10
```

```
# Create a contingency table
table_data <- xtabs(NumClients ~ FirstSeen + GMVTier, data = data)

# Conduct chi-square test
chisq_result_all <- chisq.test(table_data, simulate.p.value = TRUE, B = 2000)

# Display the result
chisq_result_all
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  table_data
X-squared = 85.881, df = NA, p-value = 0.009495
```

8. review by month

```
set.seed(12356)
client_counts <- matrix(c(
    332, 1204, 472, 128, 7,   # Pre-2020
    10, 17, 6, 1, 1,          # 2020
    19, 19, 7, 1, 0,          # 2021
    14, 19, 5, 1, 0,          # 2022
    24, 28, 3, 0, 0,          # 2023
    3, 1, 0, 0, 0             # Now (2025 onwards)
), byrow = TRUE, nrow = 6,
dimnames = list(
    'FirstSeen' = c('Pre-2020', '2020', '2021', '2022', '2023', 'Now'),
    'GMVTier' = c('< $10,000', '$10,000 to $29,999', '$30,000 to $49,999', '$50,000 to $99,999', '>
))

print(client_counts)
```

```
         GMVTier
FirstSeen  < $10,000 $10,000 to $29,999 $30,000 to $49,999 $50,000 to $99,999
  Pre-2020       332               1204                472                128
```

```
  2020           10                 17                 6                 1
  2021           19                 19                 7                 1
  2022           14                 19                 5                 1
  2023           24                 28                 3                 0
  Now             3                  1                 0                 0
          GMVTier
FirstSeen  >$100,000
  Pre-2020          7
  2020              1
  2021              0
  2022              0
  2023              0
  Now               0
```

```r
chi_test_result <- chisq.test(client_counts, simulate.p.value = TRUE, B=2000)

# View the results
print(chi_test_result)
```

```
	Pearson's Chi-squared test with simulated p-value (based on 2000
	replicates)

data:  client_counts
X-squared = 85.881, df = NA, p-value = 0.01199
```

```r
# Creating the contingency table
review_gmv <- matrix(c(
    271, 702, 291, 87, 4,    # 1 - 49 reviews
    204, 427, 127, 31, 3,    # 50 - 299 reviews
    50, 116, 29, 8, 0,       # 300 - 999 reviews
    2, 8, 2, 0, 0            # 1000+ reviews
), byrow = TRUE, nrow = 4,
dimnames = list(
    'MonthlyReviewVolume' = c('1 - 49', '50 - 299', '300 - 999', '1000+'),
    'GMVTier' = c('< $10,000', '$10,000 to $29,999', '$30,000 to $49,999', '$50,000 to $99,999', '>
))

# Viewing the contingency table
print(review_gmv)
```

```
                   GMVTier
MonthlyReviewVolume < $10,000 $10,000 to $29,999 $30,000 to $49,999
          1 - 49          271                702                291
          50 - 299        204                427                127
          300 - 999        50                116                 29
```

```
          1000+              2                  8                  2
                   GMVTier
MonthlyReviewVolume $50,000 to $99,999 >$100,000
          1 - 49                    87          4
          50 - 299                  31          3
          300 - 999                  8          0
          1000+                      0          0
```

```r
set.seed(12356)
# Perform the Chi-square test
chi_result <- chisq.test(review_gmv, simulate.p.value = TRUE, B = 2000)

# View the results
print(chi_result)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  review_gmv
X-squared = 28.316, df = NA, p-value = 0.03898
```

```r
# Assuming you have your 'review_gmv' contingency table already
chi_result <- chisq.test(review_gmv, simulate.p.value = TRUE, B = 2000)

# Now, extract standardized residuals
std_residuals <- chi_result$stdres  # This gets the standardized residuals from your test result

# View the standardized residuals
print(std_residuals)
```

```
                   GMVTier
MonthlyReviewVolume  < $10,000 $10,000 to $29,999 $30,000 to $49,999
          1 - 49    -3.1301429         -1.4008973          3.5441373
          50 - 299   2.8570737          0.5988908         -2.6162977
          300 - 999  0.8300571          1.2226748         -1.7940477
          1000+     -0.4708819          0.9476793         -0.2073494
                   GMVTier
MonthlyReviewVolume $50,000 to $99,999   >$100,000
          1 - 49             2.7249851 -0.01198982
          50 - 299          -2.1817145  0.52344063
          300 - 999         -0.9241680 -0.81248473
          1000+             -0.8244152 -0.18934341
```

```r
# Find and print significantly large residuals
significant_cells <- which(abs(std_residuals) > 2, arr.ind = TRUE)

for (idx in 1:nrow(significant_cells)) {
  cell <- significant_cells[idx, ]
  cat(sprintf("Significant cell at Monthly Review Volume '%s' and GMV Tier '%s': Residual = %.2f\n'
              rownames(std_residuals)[cell[1]],
              colnames(std_residuals)[cell[2]],
              std_residuals[cell[1], cell[2]]))
}
```

```
Significant cell at Monthly Review Volume '1 - 49' and GMV Tier '< $10,000': Residual = -3.13
Significant cell at Monthly Review Volume '50 - 299' and GMV Tier '< $10,000': Residual = 2.86
Significant cell at Monthly Review Volume '1 - 49' and GMV Tier '$30,000 to $49,999': Residual = 3.54
Significant cell at Monthly Review Volume '50 - 299' and GMV Tier '$30,000 to $49,999': Residual = -2
Significant cell at Monthly Review Volume '1 - 49' and GMV Tier '$50,000 to $99,999': Residual = 2.72
Significant cell at Monthly Review Volume '50 - 299' and GMV Tier '$50,000 to $99,999': Residual = -2
```

9. Review - starts

```r
review_gmv <- matrix(c(
    20, 22, 7, 0, 0,    # 5 Stars
    403, 919, 276, 65, 6,    # 4.5 Stars
    210, 484, 207, 65, 2,    # 4 Stars
    57, 143, 59, 15, 0       # <3 Stars
), byrow = TRUE, nrow = 4,
dimnames = list(
    'AverageReviewScore' = c('5 Stars', '4.5 Stars', '4 Stars', '<3 Stars'),
    'GMVTier' = c('< $10,000', '$10,000 to $29,999', '$30,000 to $49,999', '$50,000 to $99,999', '>
))

print(review_gmv)
```

```
                  GMVTier
AverageReviewScore < $10,000 $10,000 to $29,999 $30,000 to $49,999
         5 Stars          20                 22                  7
         4.5 Stars       403                919                276
         4 Stars         210                484                207
         <3 Stars         57                143                 59
                  GMVTier
AverageReviewScore $50,000 to $99,999 >$100,000
         5 Stars                    0         0
         4.5 Stars                 65         6
         4 Stars                   65         2
         <3 Stars                  15         0
```

```
set.seed(12356)
chi_result <- chisq.test(review_gmv)
```

```
Warning in chisq.test(review_gmv): Chi-squared approximation may be incorrect
```

```
print(chi_result)
```

```

	Pearson's Chi-squared test

data:  review_gmv
X-squared = 36.075, df = 12, p-value = 0.0003151
```

```
chi_result_simulated <- chisq.test(review_gmv, simulate.p.value = TRUE, B = 2000)
print(chi_result_simulated)
```

```

	Pearson's Chi-squared test with simulated p-value (based on 2000
	replicates)

data:  review_gmv
X-squared = 36.075, df = NA, p-value = 0.006497
```

10. Vatiality score

```
set.seed(12356)
data <- data.frame(
  VitalityScore = factor(rep(c('0 - 25', '26 - 50', '51 - 75', '76 - 100'), each = 5),
                         levels = c('0 - 25', '26 - 50', '51 - 75', '76 - 100')),
  GMVTier = rep(c('< $10,000', '$10,000 to $29,999', '$30,000 to $49,999', '$50,000 to $99,999', '>
  Clients = c(31, 58, 30, 3, 1, 298, 666, 205, 57, 2, 278, 660, 222, 54, 4, 83, 184, 92, 31, 1)
)

data$GMVTier <- ordered(data$GMVTier, levels = c('< $10,000', '$10,000 to $29,999', '$30,000 to $49

contingency_table <- xtabs(Clients ~ VitalityScore + GMVTier, data = data)
chi_result <- chisq.test(contingency_table)
```

```
Warning in chisq.test(contingency_table): Chi-squared approximation may be
incorrect
```

```r
print(chi_result)
```

```
	Pearson's Chi-squared test

data:  contingency_table
X-squared = 27.006, df = 12, p-value = 0.007711
```

```r
chi_result_simulated <- chisq.test(contingency_table, simulate.p.value = TRUE, B =2000)
print(chi_result_simulated)
```

```
	Pearson's Chi-squared test with simulated p-value (based on 2000
	replicates)

data:  contingency_table
X-squared = 27.006, df = NA, p-value = 0.01249
```

```r
expected_counts <- chi_result$expected
print(expected_counts)
```

```
             GMVTier
VitalityScore < $10,000 $10,000 to $29,999 $30,000 to $49,999
     0 - 25    28.67230           65.15676           22.81318
     26 - 50  286.25676          650.50811          227.76081
     51 - 75  283.92568          645.21081          225.90608
     76 - 100  91.14527          207.12432           72.51993
             GMVTier
VitalityScore $50,000 to $99,999 >$100,000
     0 - 25             6.025338 0.3324324
     26 - 50           60.155405 3.3189189
     51 - 75           59.665541 3.2918919
     76 - 100          19.153716 1.0567568
```

11. reputation score

```r
# Creating the data frame
data <- data.frame(
  ReputationScore = factor(rep(c('<70', '70 - 79', '80 - 90', '90 - 100'), each = 5)),
  GMVTier = rep(c('< $10,000', '$10,000 to $29,999', '$30,000 to $49,999', '$50,000 to $99,999', '>
  NumberOfClients = c(23, 60, 17, 2, 0, 75, 168, 89, 25, 1, 357, 849, 297, 94, 4, 235, 491, 146, 24
)
```

```r
# Creating the contingency table
contingency_table <- xtabs(NumberOfClients ~ ReputationScore + GMVTier, data = data)

set.seed(12356)
# Perform the Chi-square test
chi_result <- chisq.test(contingency_table,simulate.p.value = TRUE, B = 2000)

# View the results
print(chi_result)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  contingency_table
X-squared = 36.347, df = NA, p-value = 0.001499
```

```r
# Assuming chi_result is your Chi-squared test result
std_residuals <- chi_result$residuals  # Obtain the standardized residuals

# Identifying cells with significant contribution
sig_cells <- which(abs(std_residuals) > 1.96, arr.ind = TRUE)  # Using 1.96 for approximately a 95%

# Print out the significant cells and their residuals
if(length(sig_cells) > 0) {
  for(idx in 1:nrow(sig_cells)) {
    cell <- sig_cells[idx, ]
    cat(sprintf("Significant cell: Reputation Score '%s' and GMV Tier '%s' with Residual = %.2f\n"
                rownames(std_residuals)[cell[1]],
                colnames(std_residuals)[cell[2]],
                std_residuals[cell[1], cell[2]]))
  }
} else {
  cat("No cells significantly contribute to the chi-squared statistic beyond the 95% confidence lev
}
```

```
Significant cell: Reputation Score '70 - 79' and GMV Tier '$30,000 to $49,999' with Residual = 2.77
Significant cell: Reputation Score '90 - 100' and GMV Tier '$50,000 to $99,999' with Residual = -3.02
```