Project of stat 28

Yudong Zhang

April 18, 2018

1. DATA ENTRY

```
# read data
combinedData <- read.csv("combinedData.csv")
# subset
combinedData <- subset(combinedData, combinedData$DRG.Definition %in%
    c("192 - CHRONIC OBSTRUCTIVE PULMONARY DISEASE W/O CC/MCC",
        "293 - HEART FAILURE & SHOCK W/O CC/MCC", "536 - FRACTURES OF HIP & PELVIS W/O MCC",
        "638 - DIABETES W CC"))
# Make Urban and regions factors
combinedData$Urban <- factor(combinedData$Urban)
combinedData$regions <- factor(combinedData$regions)
# Create PatientPays and PctPatientPays
combinedData$PatientPays <- combinedData$Average.Total.Payments -
    combinedData$Average.Medicare.Payments
combinedData$PctPatientPays <- combinedData$PatientPays/combinedData$Average.Total.Payments
# Create a factor variable urbanByRegions
combinedData$urbanByRegions <- combinedData$Urban:combinedData$regions
# Apply droplevels
combinedData$urbanByRegions <- droplevels(combinedData$urbanByRegions)
# summary
summary(combinedData)
```

```
##  Provider.State
##  CA     : 596
##  TX     : 592
##  FL     : 519
##  NY     : 459
##  PA     : 399
##  IL     : 389
##  (Other):4947
##                                                DRG.Definition
##  192 - CHRONIC OBSTRUCTIVE PULMONARY DISEASE W/O CC/MCC:2593
##  293 - HEART FAILURE & SHOCK W/O CC/MCC                :2443
##  638 - DIABETES W CC                                   :1762
##  536 - FRACTURES OF HIP & PELVIS W/O MCC               :1103
##  039 - EXTRACRANIAL PROCEDURES W/O CC/MCC              :   0
##  057 - DEGENERATIVE NERVOUS SYSTEM DISORDERS W/O MCC   :   0
##  (Other)                                               :   0
##   Provider.Id                        Provider.Name       Provider.City
##  Min.   : 10001   GOOD SAMARITAN HOSPITAL :  29   CHICAGO     :  81
##  1st Qu.:110186   MERCY MEDICAL CENTER    :  20   BALTIMORE   :  42
##  Median :250069   ST JOSEPH HOSPITAL      :  20   BROOKLYN    :  41
##  Mean   :257006   ST JOSEPH MEDICAL CENTER:  20   HOUSTON     :  41
##  3rd Qu.:390016   MERCY HOSPITAL          :  19   PHILADELPHIA:  40
##  Max.   :670071   ST FRANCIS HOSPITAL     :  15   SPRINGFIELD :  37
##                   (Other)                 :7778   (Other)     :7619
##  Total.Discharges Average.Covered.Charges Average.Total.Payments
```

```
##  Min.   : 11.00   Min.   :  3134          Min.   : 3144
##  1st Qu.: 17.00   1st Qu.: 11352          1st Qu.: 4212
##  Median : 25.00   Median : 15544          Median : 4711
##  Mean   : 33.47   Mean   : 18368          Mean   : 5072
##  3rd Qu.: 41.00   3rd Qu.: 22070          3rd Qu.: 5532
##  Max.   :326.00   Max.   :130690          Max.   :19512
##
##  Average.Medicare.Payments Provider.Zip.Code      regions      Urban
##  Min.   : 2182             Min.   : 1040     midwest  :1842   0   :2948
##  1st Qu.: 3255             1st Qu.:27565     northeast:1469   1   : 104
##  Median : 3723             Median :44112     south    :3357   2   :1827
##  Mean   : 4093             Mean   :47812     west     :1233   3   :   6
##  3rd Qu.: 4517             3rd Qu.:72342                      5   :2520
##  Max.   :18613             Max.   :99701                      NA's: 496
##
##   PatientPays      PctPatientPays       urbanByRegions
##  Min.   :  261.2   Min.   :0.03898   0:south    :1239
##  1st Qu.:  770.2   1st Qu.:0.15297   2:south    :1010
##  Median :  898.4   Median :0.19230   5:south    : 784
##  Mean   :  979.3   Mean   :0.19949   0:midwest  : 663
##  3rd Qu.: 1059.3   3rd Qu.:0.23569   5:northeast: 625
##  Max.   :10676.8   Max.   :0.74215   (Other)    :3084
##                                      NA's       : 496
```

2. BASIC SUMMARIES

```r
# summary of PatientPays and PctPatientPays
summary(combinedData$PatientPays)
```

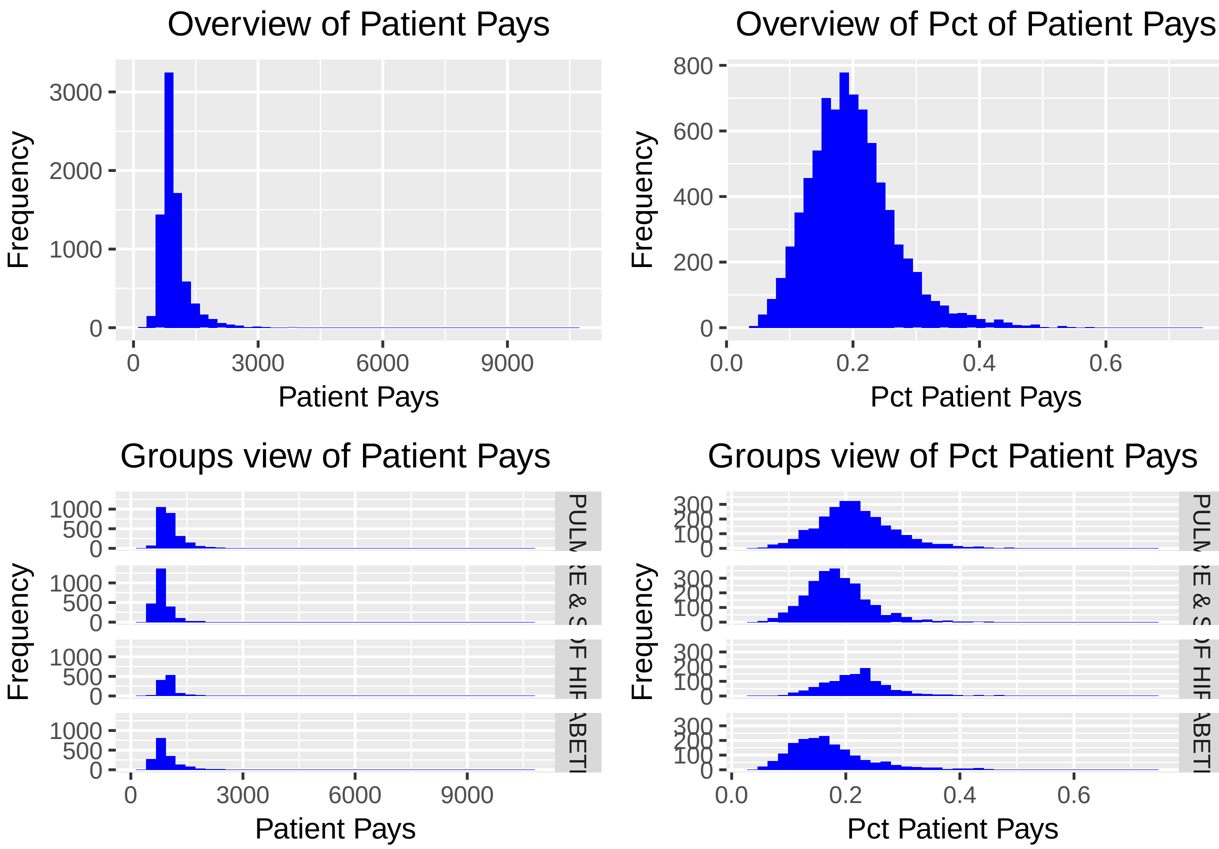
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   261.2   770.2   898.4   979.3  1059.3 10676.8
```

```r
summary(combinedData$PctPatientPays)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03898 0.15297 0.19230 0.19949 0.23569 0.74215
```

```r
# histograms of PatientPays and PctPatientPays
pay_hist <- ggplot(combinedData, aes(x = PatientPays))
pctpay_hist <- ggplot(combinedData, aes(x = PctPatientPays))

p1 <- pay_hist + geom_histogram(bins = 50, fill = "blue") + ggtitle("Overview of Patient Pays") +
    xlab("Patient Pays") + ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5))

p2 <- pctpay_hist + geom_histogram(bins = 50, fill = "blue") +
    ggtitle("Overview of Pct of Patient Pays") + xlab("Pct Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5))

# Groups view of PatientPays and PctPatientPays
p3 <- pay_hist + geom_histogram(bins = 40, fill = "blue") + ggtitle("Groups view of Patient Pays") +
    xlab("Patient Pays") + ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5)) +
    facet_grid(DRG.Definition ~ .)

p4 <- pctpay_hist + geom_histogram(bins = 40, fill = "blue") +
    ggtitle("Groups view of Pct Patient Pays") + xlab("Pct Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5)) +
```

```
    facet_grid(DRG.Definition ~ .)

grid.arrange(p1, p2, p3, p4, ncol = 2, nrow = 2)
```



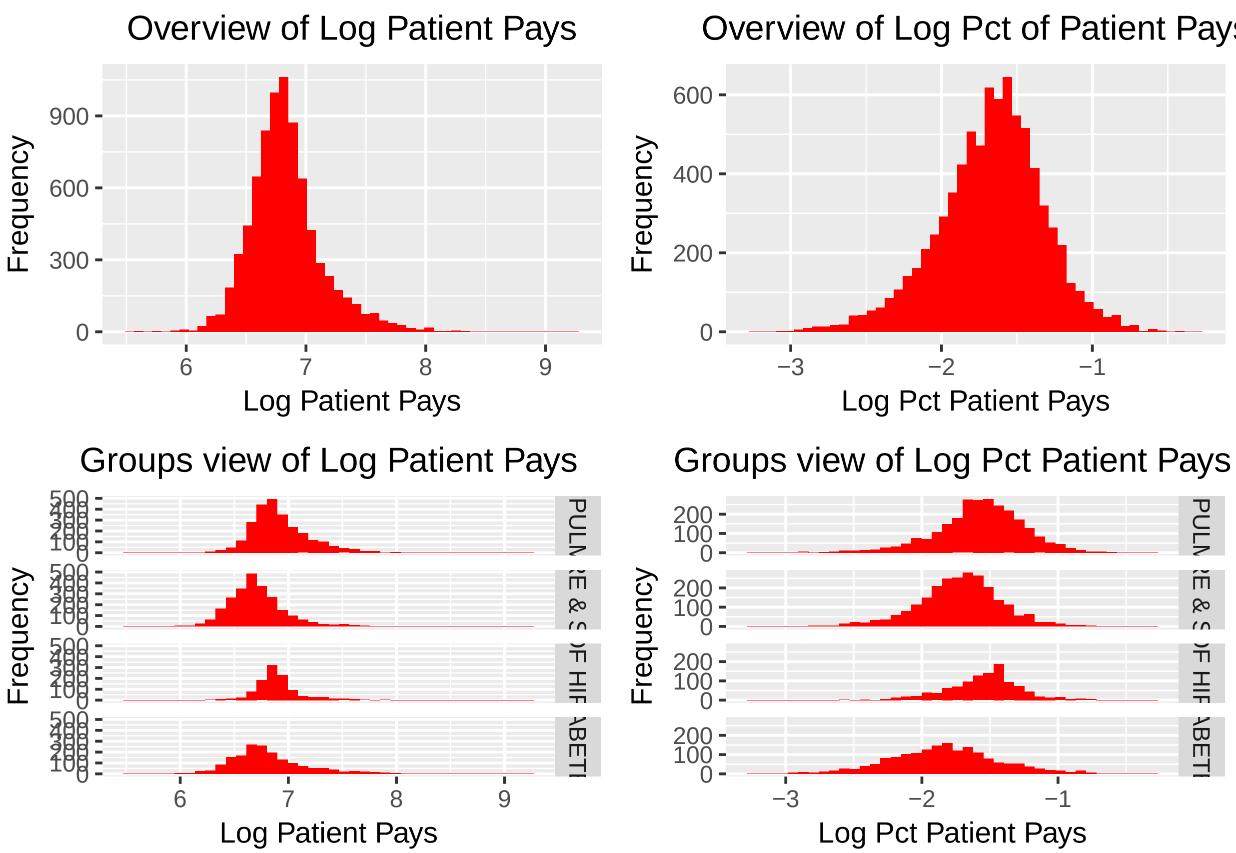
>

Comments: From the summaries we can see that in `PctPatientPays` and `PatientPays`, mean is larger than median; and from the plots (no matter the total view or separate view), there are right tails in the plots, thus both distributions are right-skewed. which means there're some outliers in large value range. To solve this, it would be helpful to apply logarithm or square-root transformation to the data.

Now we apply transformation to the data.

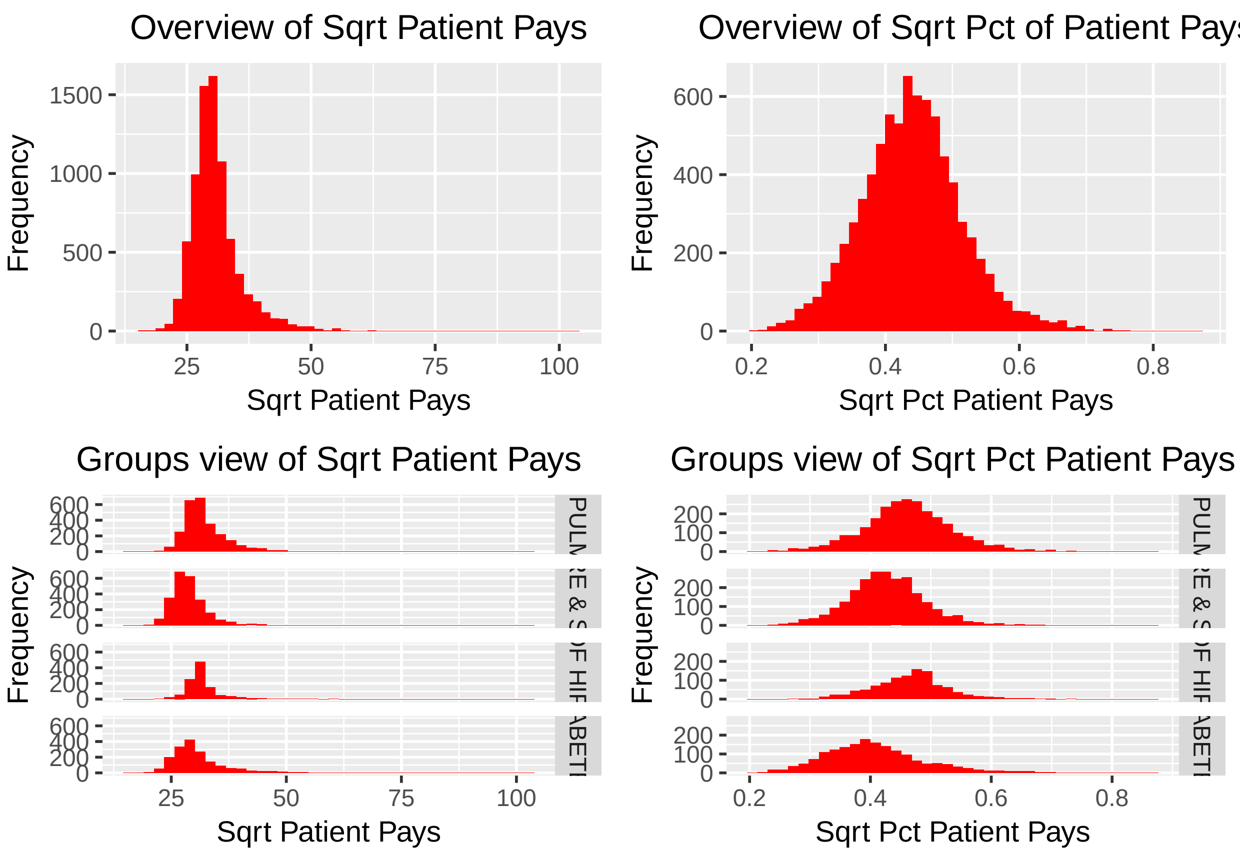
```
# Log transformation
logpay_hist <- ggplot(combinedData, aes(x = log(PatientPays)))
logpctpay_hist <- ggplot(combinedData, aes(x = log(PctPatientPays)))

p1 <- logpay_hist + geom_histogram(bins = 50, fill = "red") +
    ggtitle("Overview of Log Patient Pays") + xlab("Log Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5))

p2 <- logpctpay_hist + geom_histogram(bins = 50, fill = "red") +
    ggtitle("Overview of Log Pct of Patient Pays") + xlab("Log Pct Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5))

# Groups view of PatientPays and PctPatientPays
p3 <- logpay_hist + geom_histogram(bins = 40, fill = "red") +
    ggtitle("Groups view of Log Patient Pays") + xlab("Log Patient Pays") +
```

```
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5)) +
    facet_grid(DRG.Definition ~ .)

p4 <- logpctpay_hist + geom_histogram(bins = 40, fill = "red") +
    ggtitle("Groups view of Log Pct Patient Pays") + xlab("Log Pct Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5)) +
    facet_grid(DRG.Definition ~ .)

grid.arrange(p1, p2, p3, p4, ncol = 2, nrow = 2)
```




```
# sqrt transformation
sqrtpay_hist <- ggplot(combinedData, aes(x = sqrt(PatientPays)))
sqrtpctpay_hist <- ggplot(combinedData, aes(x = sqrt(PctPatientPays)))

p1 <- sqrtpay_hist + geom_histogram(bins = 50, fill = "red") +
    ggtitle("Overview of Sqrt Patient Pays") + xlab("Sqrt Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5))

p2 <- sqrtpctpay_hist + geom_histogram(bins = 50, fill = "red") +
    ggtitle("Overview of Sqrt Pct of Patient Pays") + xlab("Sqrt Pct Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5))

# Groups view of PatientPays and PctPatientPays
p3 <- sqrtpay_hist + geom_histogram(bins = 40, fill = "red") +
    ggtitle("Groups view of Sqrt Patient Pays") + xlab("Sqrt Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5)) +
    facet_grid(DRG.Definition ~ .)
```

```r
p4 <- sqrtpctpay_hist + geom_histogram(bins = 40, fill = "red") +
    ggtitle("Groups view of Sqrt Pct Patient Pays") + xlab("Sqrt Pct Patient Pays") +
    ylab("Frequency") + theme(plot.title = element_text(hjust = 0.5)) +
    facet_grid(DRG.Definition ~ .)

grid.arrange(p1, p2, p3, p4, ncol = 2, nrow = 2)
```



Comments: we can see that no matter using logarithm or square-root transformation, the distribution looks more normal and centered.

```r
# cross-tabulation/contingency table of Urban and regions
cross_table <- with(combinedData, table(Urban, regions))
cross_table
```

```
##      regions
## Urban midwest northeast south west
##     0     663       535  1239  511
##     1      13        13    75    3
##     2     464       199  1010  154
##     3       0         2     4    0
##     5     612       625   784  499
```

```r
summary(combinedData$urbanByRegions)
```

```
##   0:midwest 0:northeast     0:south      0:west   1:midwest 1:northeast
##         663         535        1239         511          13          13
```

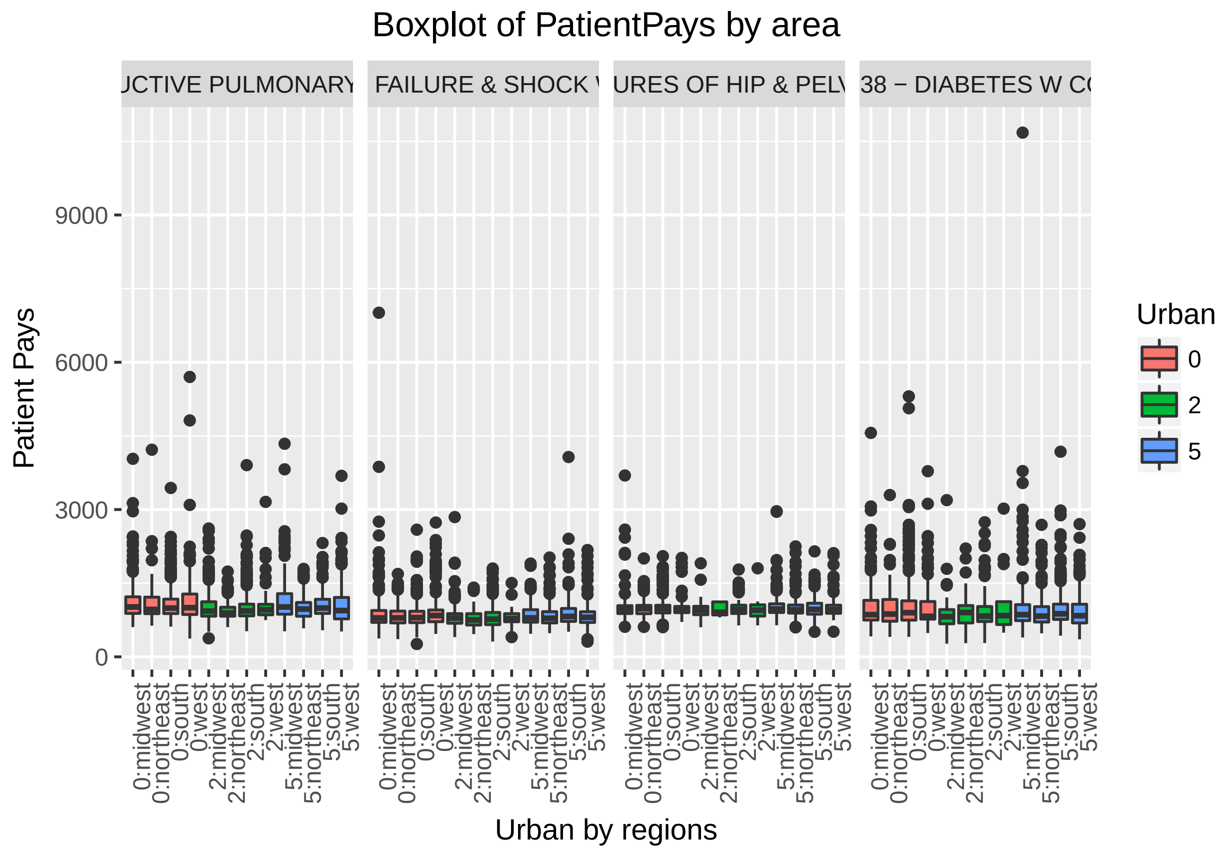
```
##     1:south      1:west   2:midwest 2:northeast     2:south      2:west
##          75           3         464         199        1010         154
## 3:northeast     3:south   5:midwest 5:northeast     5:south      5:west
##           2           4         612         625         784         499
##        NA's
##         496
```

Comments: The usage of urbanByRegions is that it groups the data by both `Urban` and `regions`, so we can explore the relationship and patterns between the numbers and both `Urban` and `regions` rather than a single variable. We can also fetch the number of certain region and certain urban label conveniently.

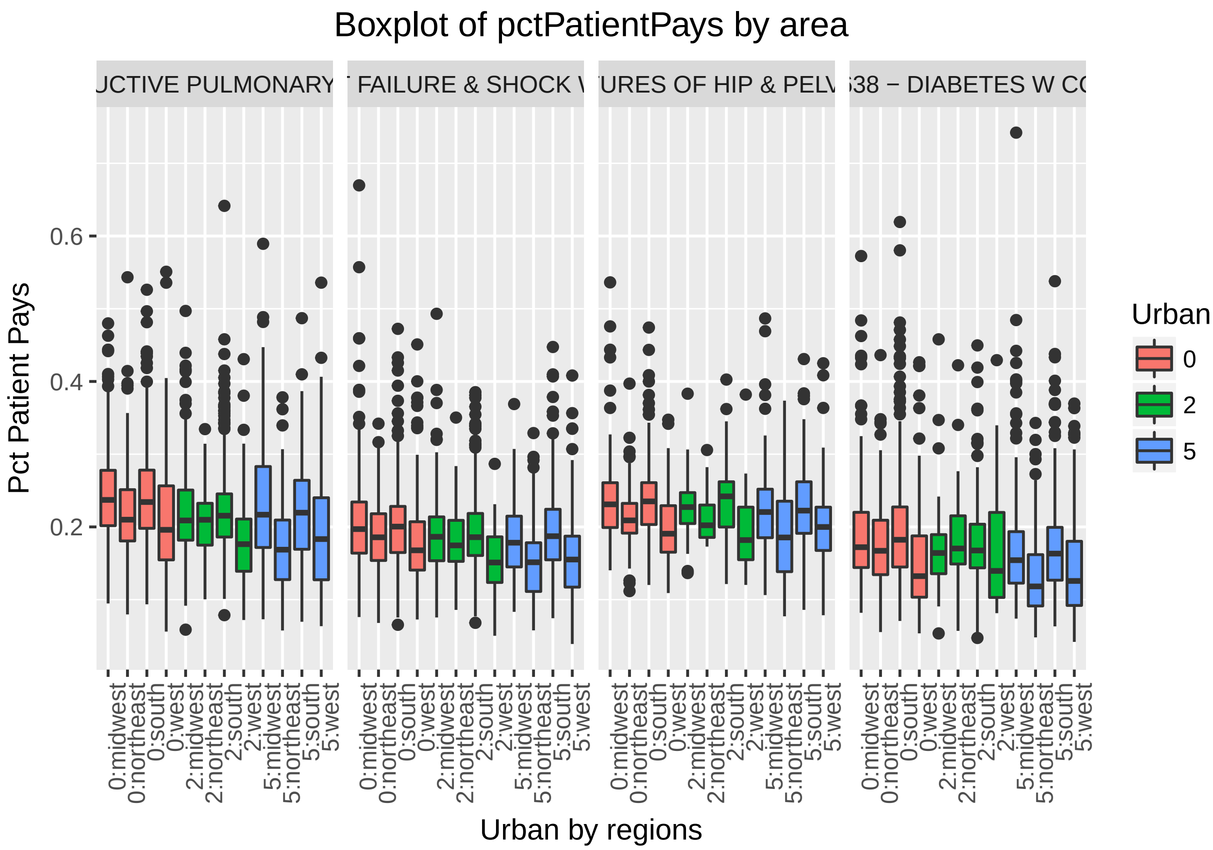
Comments: we can see in the table that the majority of data gathered in urban label 0, 2 and 5, while there're very few data in label 1 and 3, even not any data with label 4. Our conclusion could have high bias if we conclude based on those data (in label 1, 3 or 4), that's the problem.

3. EXPLORATION OF DISTRIBUTIONS IN GROUPS

```
# Create 1-2 useful visualizations
MajorData <- subset(combinedData, combinedData$Urban %in% c(0,
    2, 5))

group_pay <- ggplot(MajorData, aes(x = urbanByRegions, y = PatientPays,
    fill = Urban))
group_pctpay <- ggplot(MajorData, aes(x = urbanByRegions, y = PctPatientPays,
    fill = Urban))

p1 <- group_pay + geom_boxplot() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Patient Pays") + ggtitle("Boxplot of PatientPays by area") +
    facet_grid(. ~ DRG.Definition)

p2 <- group_pctpay + geom_boxplot() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Pct Patient Pays") + ggtitle("Boxplot of pctPatientPays by area")
    facet_grid(. ~ DRG.Definition)

p1
```

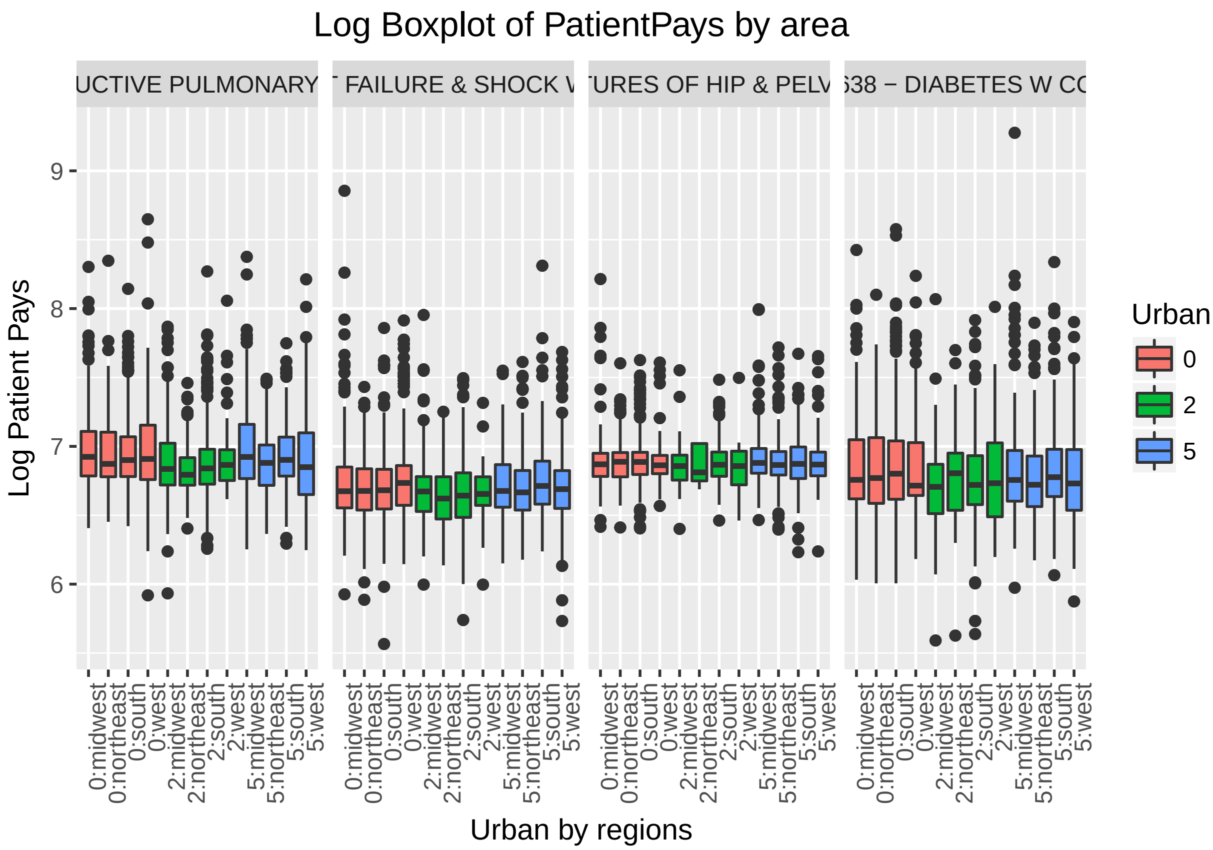

Boxplot of PatientPays by area
UCTIVE PULMONARY
FAILURE & SHOCK \
URES OF HIP & PELV
38 − DIABETES W CO
Patient Pays
9000
6000
3000
0
0:midwest
0:northeast
0:south
0:west
2:midwest
2:northeast
2:south
2:west
5:midwest
5:northeast
5:south
5:west
Urban by regions
Urban
0
2
5

```
p2
```



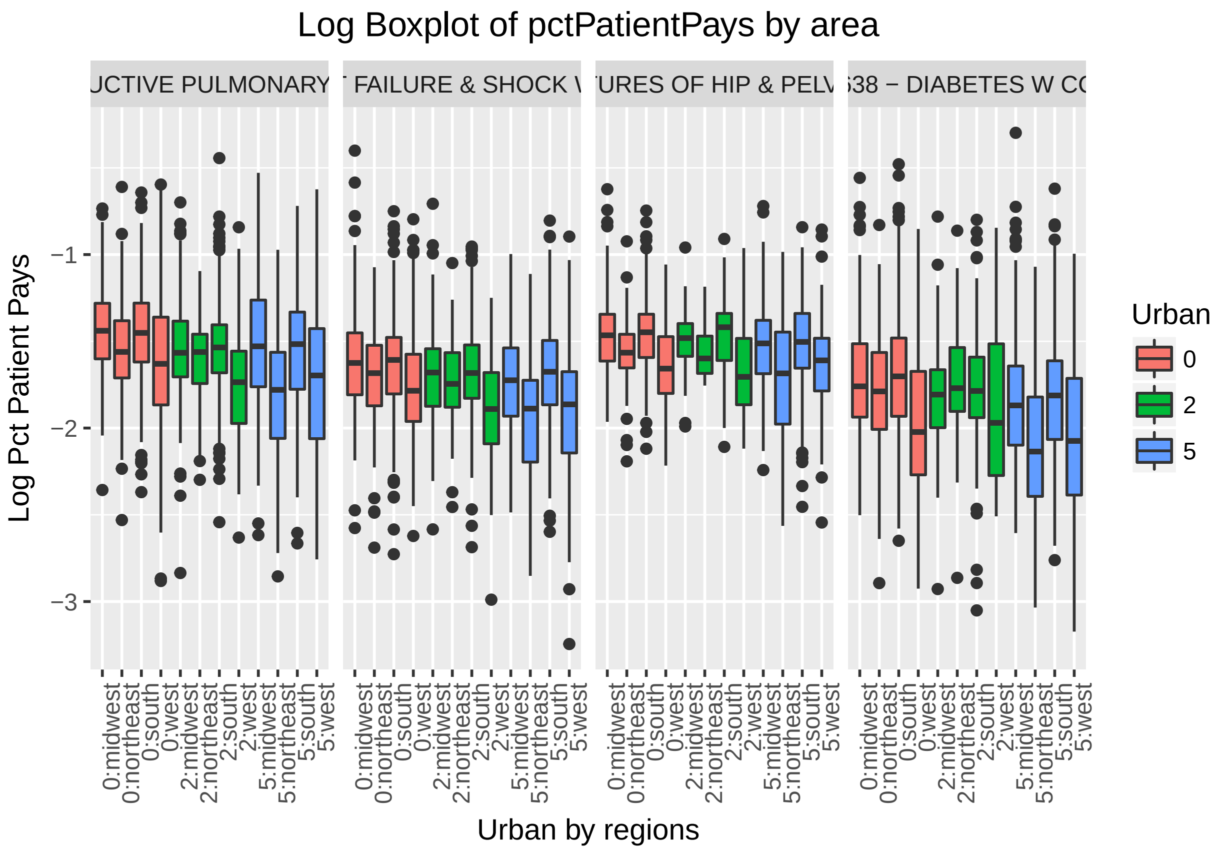
Comments: From the two plots above, we can see that the distribution is still skewed for nearly every entry in both PatientPays and PctPatientPays, most large values are regarded as outliers. Therefore we need to make a logarithm transformation before making conclusion.

```
# Log Transformation
log_group_pay <- ggplot(MajorData, aes(x = urbanByRegions, y = log(PatientPays),
    fill = Urban))
log_group_pctpay <- ggplot(MajorData, aes(x = urbanByRegions,
    y = log(PctPatientPays), fill = Urban))

p3 <- log_group_pay + geom_boxplot() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Patient Pays") + ggtitle("Log Boxplot of PatientPays by area")
    facet_grid(. ~ DRG.Definition)

p4 <- log_group_pctpay + geom_boxplot() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Pct Patient Pays") +
    ggtitle("Log Boxplot of pctPatientPays by area") + facet_grid(. ~
    DRG.Definition)

p3
```



Log Boxplot of PatientPays by area
UCTIVE PULMONARY
FAILURE & SHOCK W
URES OF HIP & PELV
638 − DIABETES W CC
Log Patient Pays
Urban by regions
Urban
0
2
5


```
p4
```

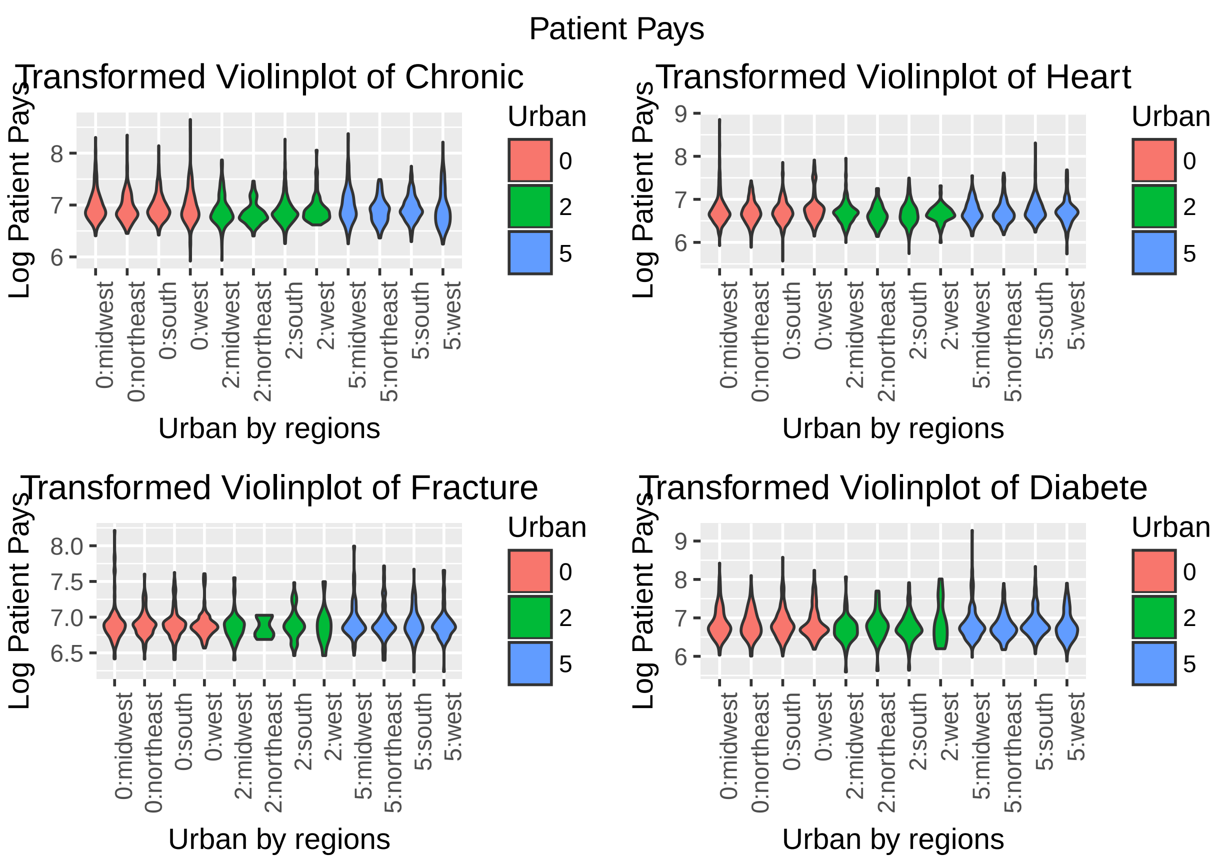


Comment: The plots looks better, but there're still some outliers in every entry. We use the transformed data for further opertation.

```
# data transformation
MajorData$PatientPays <- log(MajorData$PatientPays)
MajorData$PctPatientPays <- log(MajorData$PctPatientPays)
```

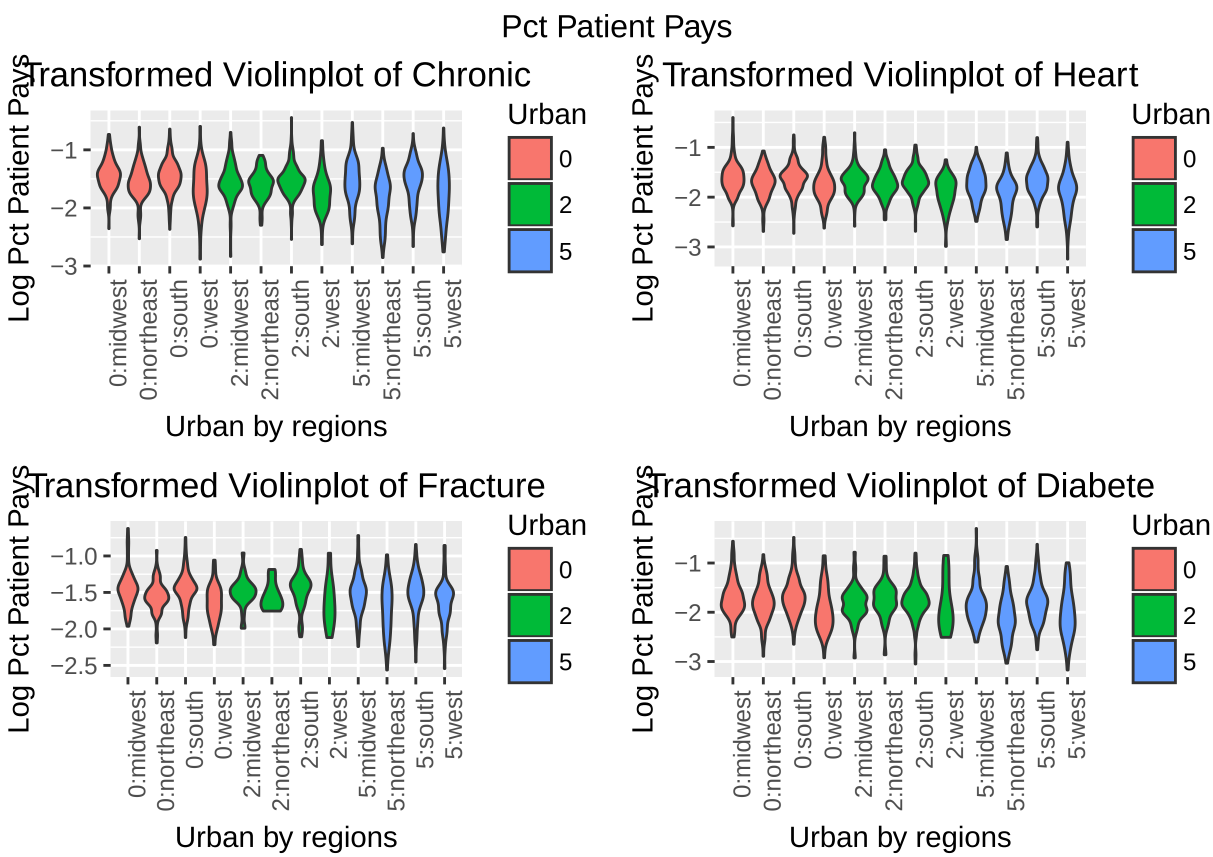
```
# Use transformed data for violin plot visualization
diagData_chronic <- subset(MajorData, MajorData$DRG.Definition ==
    "192 - CHRONIC OBSTRUCTIVE PULMONARY DISEASE W/O CC/MCC")
diagData_heart <- subset(MajorData, MajorData$DRG.Definition ==
    "293 - HEART FAILURE & SHOCK W/O CC/MCC")
diagData_fracture <- subset(MajorData, MajorData$DRG.Definition ==
    "536 - FRACTURES OF HIP & PELVIS W/O MCC")
diagData_diabete <- subset(MajorData, MajorData$DRG.Definition ==
    "638 - DIABETES W CC")

# Patient Pays
chronic_group <- ggplot(diagData_chronic, aes(x = urbanByRegions,
    y = PatientPays, fill = Urban))
p1 <- chronic_group + geom_violin() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Patient Pays") + ggtitle("Transformed Violinplot of Chronic")

heart_group <- ggplot(diagData_heart, aes(x = urbanByRegions,
    y = PatientPays, fill = Urban))
```

```r
p2 <- heart_group + geom_violin() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Patient Pays") + ggtitle("Transformed Violinplot of Heart")

fracture_group <- ggplot(diagData_fracture, aes(x = urbanByRegions,
    y = PatientPays, fill = Urban))
p3 <- fracture_group + geom_violin() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Patient Pays") + ggtitle("Transformed Violinplot of Fracture")

diabete_group <- ggplot(diagData_diabete, aes(x = urbanByRegions,
    y = PatientPays, fill = Urban))
p4 <- diabete_group + geom_violin() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Patient Pays") + ggtitle("Transformed Violinplot of Diabete")

grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2, top = "Patient Pays")
```




```r
# Percentage of Patient Pays
chronic_group <- ggplot(diagData_chronic, aes(x = urbanByRegions,
    y = PctPatientPays, fill = Urban))
p1 <- chronic_group + geom_violin() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Pct Patient Pays") +
    ggtitle("Transformed Violinplot of Chronic")

heart_group <- ggplot(diagData_heart, aes(x = urbanByRegions,
```

```
    y = PctPatientPays, fill = Urban))
p2 <- heart_group + geom_violin() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Pct Patient Pays") +
    ggtitle("Transformed Violinplot of Heart")

fracture_group <- ggplot(diagData_fracture, aes(x = urbanByRegions,
    y = PctPatientPays, fill = Urban))
p3 <- fracture_group + geom_violin() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Pct Patient Pays") +
    ggtitle("Transformed Violinplot of Fracture")

diabete_group <- ggplot(diagData_diabete, aes(x = urbanByRegions,
    y = PctPatientPays, fill = Urban))
p4 <- diabete_group + geom_violin() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 1), plot.title = element_text(hjust = 0.5)) +
    xlab("Urban by regions") + ylab("Log Pct Patient Pays") +
    ggtitle("Transformed Violinplot of Diabete")

grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2, top = "Pct Patient Pays")
```



Comments: For visualization part, I applied boxplot and violinplot on logarithically transformed PatientPays and PctPatientPays data. After transformation, the distribution of all entries become more normal and centered. Both kinds of plots tell us how the majority of data in each entry is distributed.

From the vionlinplots we can tell that the mean of the patient payment in "Chronic" and "Heart" case are similar (within itself), however, the data of some entries in urbanByRegion are more normally centered, while the rest looks more uniform. While in "Fracture" and "Diabetes" cases, both the mean and whole distribution are quite different within the group itself.

4. PERFORM INFERENCE TO EVALUATE THE DIFFERENCE BETWEEN THE GROUPS > To do inference between groups, we need to perform tests on them.

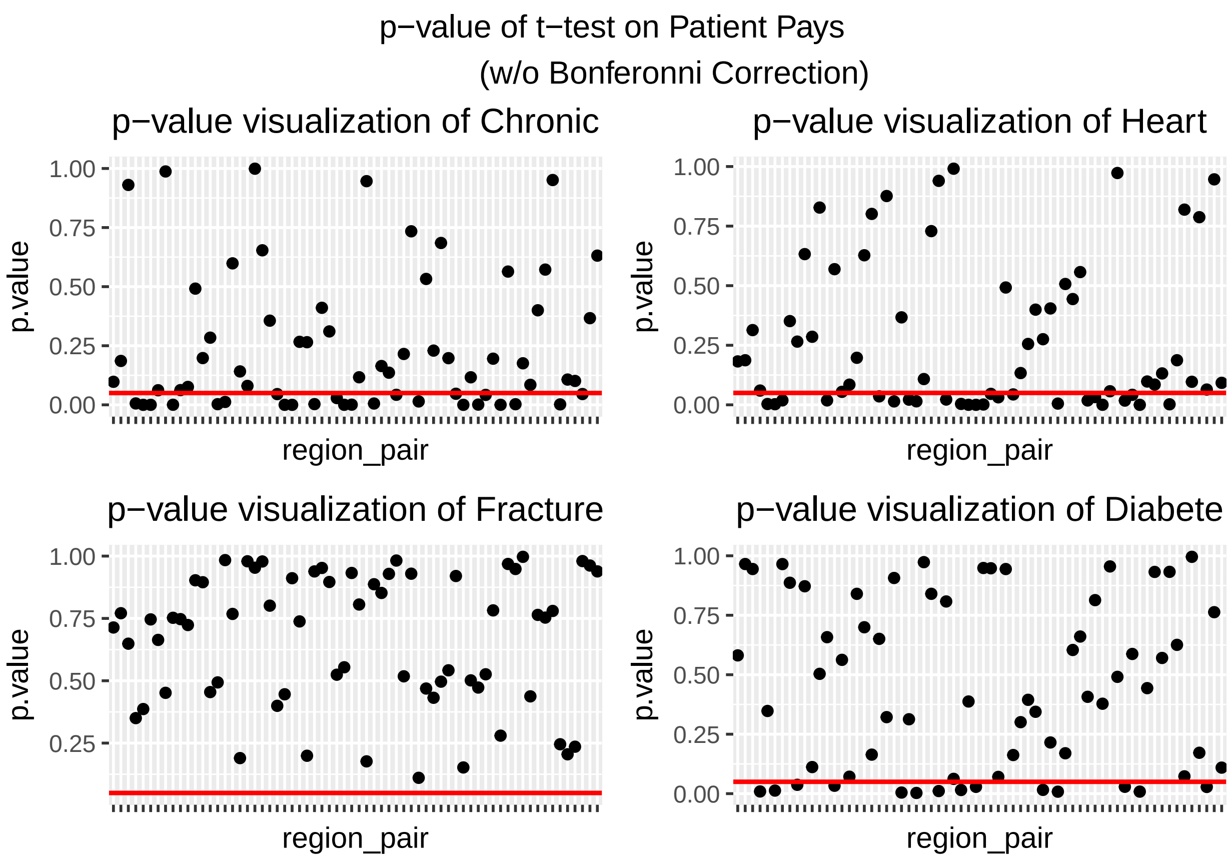
```
# t-tests on Patient Pays data
diagData_all <- list(diagData_chronic, diagData_heart, diagData_fracture,
    diagData_diabete)
# names(diagData_all) <- c('Chronic', 'Heart', 'Fracture',
# 'Diabete')
diag_names <- c("Chronic", "Heart", "Fracture", "Diabete")

ttest_func <- function(x) {
    data1 <- subset(sub_data$PatientPays, sub_data$urbanByRegions ==
        x[1])
    data2 <- subset(sub_data$PatientPays, sub_data$urbanByRegions ==
        x[2])
    pvalue <- t.test(data1, data2)$p.value
    pvalue
}

Bonferonni_ttest_func <- function(x) {
    data1 <- subset(sub_data$PatientPays, sub_data$urbanByRegions ==
        x[1])
    data2 <- subset(sub_data$PatientPays, sub_data$urbanByRegions ==
        x[2])
    pvalue <- t.test(data1, data2)$p.value
    pvalue <- pvalue * dim(UBR.pairs)[1]  # correction
    pvalue
}

CI_func <- function(x) {
    data1 <- subset(sub_data$PatientPays, sub_data$urbanByRegions ==
        x[1])
    data2 <- subset(sub_data$PatientPays, sub_data$urbanByRegions ==
        x[2])
    CI <- t.test(data1, data2, conf.level = 1 - 0.05/dim(UBR.pairs)[1])$conf.int
    low <- CI[[1]]
    high <- CI[[2]]
    estimation <- mean(data1) - mean(data2)
    return(c(low = low, high = high, estimate = estimation))
}

plot_pv <- list()
plot_adpv <- list()
store_adjusted_df <- list()
```

```r
plot_CI <- list()
for (i in 1:4) {
    sub_data <- droplevels(data.frame(diagData_all[i]))
    urbanByRegion_pairs <- combn(levels(sub_data$urbanByRegions),
        2)
    UBR.pairs <- data.frame(t(urbanByRegion_pairs))

    p.values <- apply(urbanByRegion_pairs, 2, ttest_func)
    adp.values <- apply(urbanByRegion_pairs, 2, Bonferonni_ttest_func)
    CIs <- apply(urbanByRegion_pairs, 2, CI_func)
    UBR.pairs$p.value <- p.values
    UBR.pairs$adp.value <- adp.values
    UBR.pairs$low <- CIs[1, ]
    UBR.pairs$high <- CIs[2, ]
    UBR.pairs$estimate <- CIs[3, ]
    UBR.pairs$region_pair <- factor(paste(UBR.pairs$X1, "&",
        UBR.pairs$X2))

    pairs_num <- length(UBR.pairs$p.value)
    print(paste("(Patient Pays) The region-pairs (", diag_names[i],
        ") that are significantly different in mean value are:"))
    for (j in 1:pairs_num) {
        UBR.pairs$p.value[j] <- min(UBR.pairs$p.value[j], 1)  # upper bound is 1.0
        UBR.pairs$adp.value[j] <- min(UBR.pairs$adp.value[j],
            1)
        if (UBR.pairs$adp.value[j] < 0.05) {
            print(paste(UBR.pairs$region_pair[[j]], ", adjusted p-value= ",
                round(UBR.pairs$adp.value[[j]], 6)))
        }
    }
    cat("\n")

    plot_pv[[i]] <- ggplot(UBR.pairs, aes(x = region_pair, y = p.value)) +
        geom_point() + theme(axis.text.x = element_blank(), plot.title = element_text(hjust = 0.5)) +
        ggtitle(paste("p-value visualization of", diag_names[i])) +
        geom_hline(yintercept = 0.05, color = "red", lwd = 0.8)
    plot_adpv[[i]] <- ggplot(UBR.pairs, aes(x = region_pair,
        y = adp.value)) + geom_point() + theme(axis.text.x = element_blank(),
        plot.title = element_text(hjust = 0.5)) + ggtitle(paste("adjusted p-value visualization of",
        diag_names[i])) + geom_hline(yintercept = 0.05, color = "blue",
        lwd = 0.8)
    plot_CI[[i]] <- ggplot(UBR.pairs, aes(x = region_pair, y = estimate)) +
        geom_point() + geom_errorbar(aes(ymax = high, ymin = low)) +
        theme(axis.text.x = element_blank(), plot.title = element_text(hjust = 0.5)) +
        ggtitle(paste("CI visualization of", diag_names[i]))
}
```

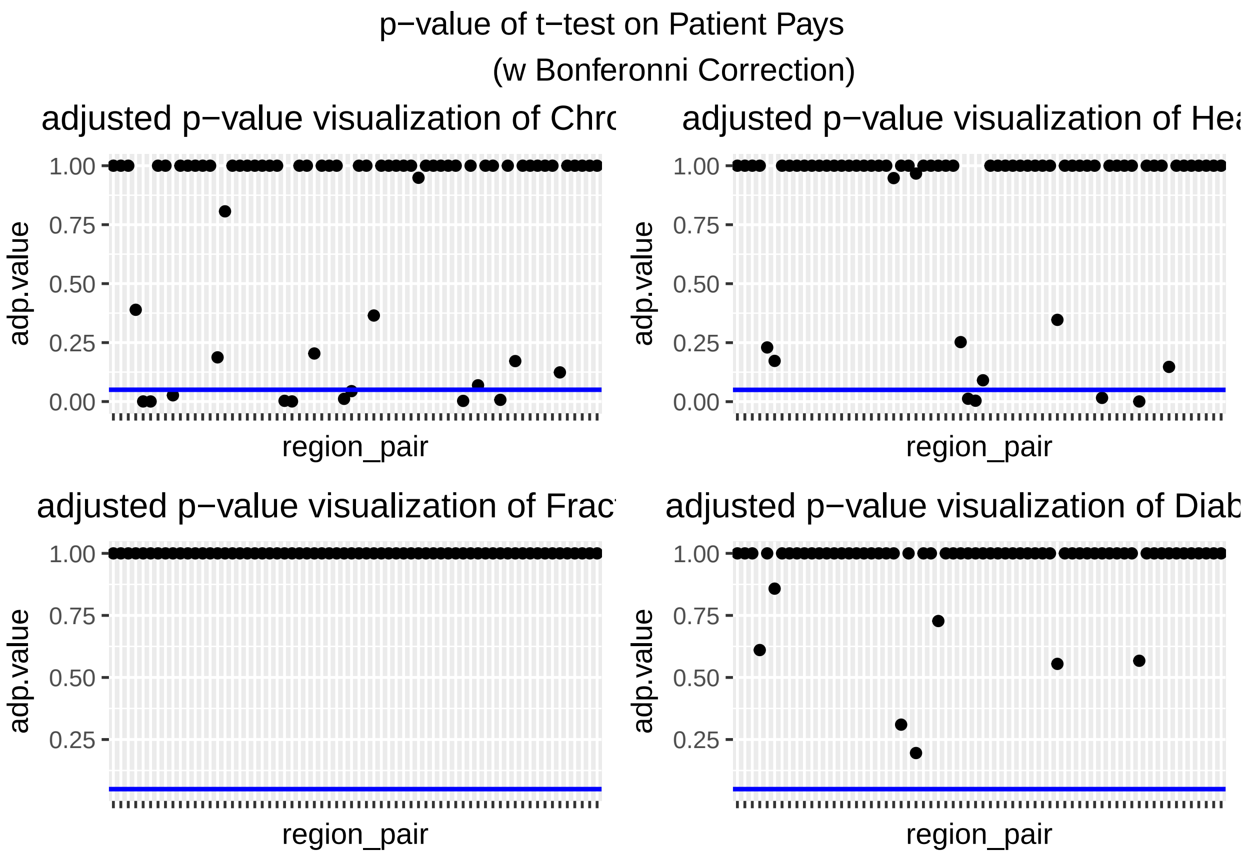
```
## [1] "(Patient Pays) The region-pairs ( Chronic ) that are significantly different in mean value are:
## [1] "0:midwest & 2:northeast , adjusted p-value=  0.000443"
## [1] "0:midwest & 2:south , adjusted p-value=  0.000345"
## [1] "0:midwest & 5:northeast , adjusted p-value=  0.026438"
## [1] "0:south & 2:northeast , adjusted p-value=  0.00314"
## [1] "0:south & 2:south , adjusted p-value=  0.00057"
## [1] "0:west & 2:northeast , adjusted p-value=  0.011443"
```

```
## [1] "0:west & 2:south , adjusted p-value=  0.044272"
## [1] "2:northeast & 5:midwest , adjusted p-value=  0.002784"
## [1] "2:south & 5:midwest , adjusted p-value=  0.007245"
##
## [1] "(Patient Pays) The region-pairs ( Heart ) that are significantly different in mean value are:"
## [1] "0:west & 2:northeast , adjusted p-value=  0.01231"
## [1] "0:west & 2:south , adjusted p-value=  0.00399"
## [1] "2:northeast & 5:south , adjusted p-value=  0.015969"
## [1] "2:south & 5:south , adjusted p-value=  0.000765"
##
## [1] "(Patient Pays) The region-pairs ( Fracture ) that are significantly different in mean value are
##
## [1] "(Patient Pays) The region-pairs ( Diabete ) that are significantly different in mean value are:
```

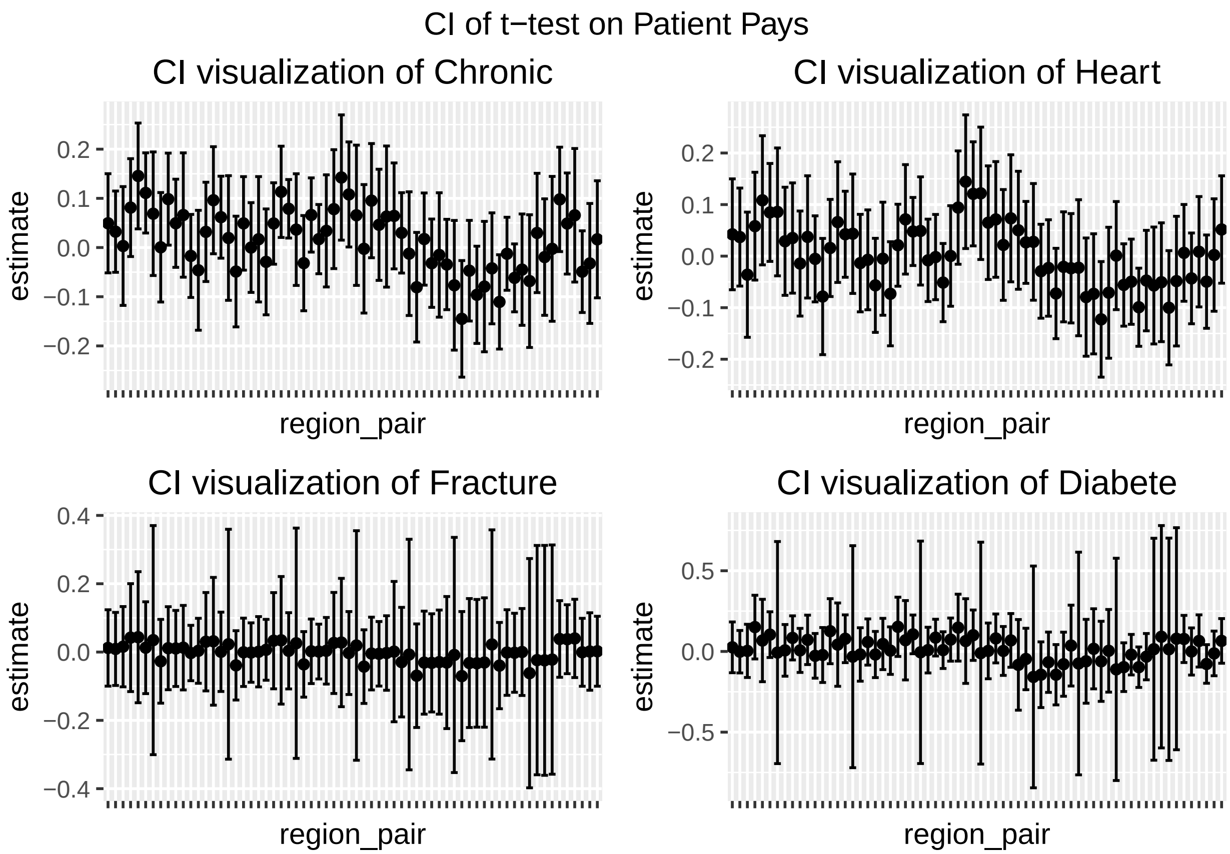
```r
# plot original p-values
grid.arrange(plot_pv[[1]], plot_pv[[2]], plot_pv[[3]], plot_pv[[4]],
    nrow = 2, ncol = 2, top = "p-value of t-test on Patient Pays
             (w/o Bonferonni Correction)")
```




```r
# plot adjusted p-values
grid.arrange(plot_adpv[[1]], plot_adpv[[2]], plot_adpv[[3]],
    plot_adpv[[4]], nrow = 2, ncol = 2, top = "p-value of t-test on Patient Pays
             (w Bonferonni Correction)")
```




```r
# plot CI
grid.arrange(plot_CI[[1]], plot_CI[[2]], plot_CI[[3]], plot_CI[[4]],
    nrow = 2, ncol = 2, top = "CI of t-test on Patient Pays")
```



>

Comments: For Patient Pays data, we can see in the first graph that before Bonferonni correction, there're more pairs shows siginificant difference, but after correction, only a few pairs were left to show siginificant difference.

> Comments: Now we focus on adjusted p-values. As for certain diagnoses, in "Chronic" and "Heart" cases, there're some pairs that are siginificantly different in mean value and "Chronic" has more than "Heart"; while for "Fracture" and "Diabete" cases, no pairs are siginificantly different, which means that we can't reject any Null hypothesis in pairs of these two cases. We can also see accurate number of above described pairs in the printed pairs.

```r
# t-tests on Pct Patient Pays data
ttest_func <- function(x) {
    data1 <- subset(sub_data$PctPatientPays, sub_data$urbanByRegions ==
        x[1])
    data2 <- subset(sub_data$PctPatientPays, sub_data$urbanByRegions ==
        x[2])
    pvalue <- t.test(data1, data2)$p.value
    pvalue
}

Bonferonni_ttest_func <- function(x) {
    data1 <- subset(sub_data$PctPatientPays, sub_data$urbanByRegions ==
        x[1])
    data2 <- subset(sub_data$PctPatientPays, sub_data$urbanByRegions ==
        x[2])
    pvalue <- t.test(data1, data2)$p.value
    pvalue <- pvalue * dim(UBR.pairs)[1]  # correction
```

```
    pvalue
}

CI_func <- function(x) {
    data1 <- subset(sub_data$PctPatientPays, sub_data$urbanByRegions ==
        x[1])
    data2 <- subset(sub_data$PctPatientPays, sub_data$urbanByRegions ==
        x[2])
    CI <- t.test(data1, data2, conf.level = 1 - 0.05/dim(UBR.pairs)[1])$conf.int
    low <- CI[[1]]
    high <- CI[[2]]
    estimation <- mean(data1) - mean(data2)
    return(c(low = low, high = high, estimate = estimation))
}

# store_adjusted_df <- list()
plot_pv <- list()
plot_adpv <- list()
plot_CI <- list()
for (i in 1:4) {
    sub_data <- droplevels(data.frame(diagData_all[i]))
    urbanByRegion_pairs <- combn(levels(sub_data$urbanByRegions),
        2)
    UBR.pairs <- data.frame(t(urbanByRegion_pairs))
    # perform tests
    p.values <- apply(urbanByRegion_pairs, 2, ttest_func)
    adp.values <- apply(urbanByRegion_pairs, 2, Bonferonni_ttest_func)
    CIs <- apply(urbanByRegion_pairs, 2, CI_func)
    UBR.pairs$p.value <- p.values
    UBR.pairs$adp.value <- adp.values
    UBR.pairs$low <- CIs[1, ]
    UBR.pairs$high <- CIs[2, ]
    UBR.pairs$estimate <- CIs[3, ]
    UBR.pairs$region_pair <- factor(paste(UBR.pairs$X1, "&",
        UBR.pairs$X2))

    pairs_num <- length(UBR.pairs$p.value)
    # setting upper bound for p-value
    print(paste("(Pct Patient Pays) The region-pairs (", diag_names[i],
        ") that are significantly different in mean value are:"))
    for (j in 1:pairs_num) {
        UBR.pairs$p.value[j] <- min(UBR.pairs$p.value[j], 1)  # upper bound is 1.0
        UBR.pairs$adp.value[j] <- min(UBR.pairs$adp.value[j],
            1)
        if (UBR.pairs$adp.value[j] < 0.05) {
            print(paste(UBR.pairs$region_pair[[j]], ", adjusted p-value= ",
                round(UBR.pairs$adp.value[[j]], 6)))
        }
    }
    cat("\n")
    # original p-value plot
    plot_pv[[i]] <- ggplot(UBR.pairs, aes(x = region_pair, y = p.value)) +
        geom_point() + theme(axis.text.x = element_blank(), plot.title = element_text(hjust = 0.5)) +
```

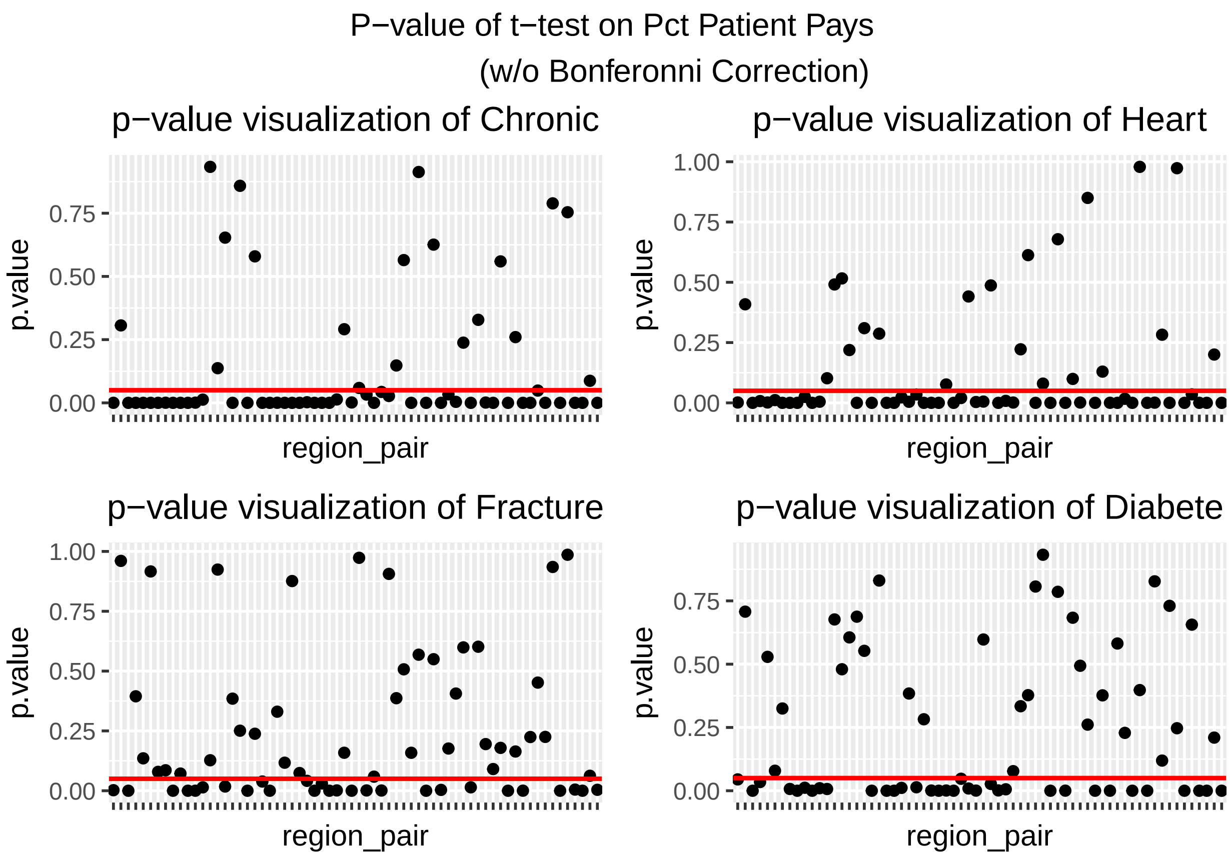
```
        ggtitle(paste("p-value visualization of", diag_names[i])) +
        geom_hline(yintercept = 0.05, color = "red", lwd = 0.8)
    # Bonferonni-adjusted p-value plot
    plot_adpv[[i]] <- ggplot(UBR.pairs, aes(x = region_pair,
        y = adp.value)) + geom_point() + theme(axis.text.x = element_blank(),
        plot.title = element_text(hjust = 0.5)) + ggtitle(paste("adjusted p-value visualization of",
        diag_names[i])) + geom_hline(yintercept = 0.05, color = "blue",
        lwd = 0.8)
    plot_CI[[i]] <- ggplot(UBR.pairs, aes(x = region_pair, y = estimate)) +
        geom_point() + geom_errorbar(aes(ymax = high, ymin = low)) +
        theme(axis.text.x = element_blank(), plot.title = element_text(hjust = 0.5)) +
        ggtitle(paste("CI visualization of", diag_names[i]))
}
```

```
## [1] "(Pct Patient Pays) The region-pairs ( Chronic ) that are significantly different in mean value a
## [1] "0:midwest & 0:northeast , adjusted p-value=  0.008145"
## [1] "0:midwest & 0:west , adjusted p-value=  2e-06"
## [1] "0:midwest & 2:midwest , adjusted p-value=  0.003104"
## [1] "0:midwest & 2:northeast , adjusted p-value=  9.1e-05"
## [1] "0:midwest & 2:south , adjusted p-value=  0.000361"
## [1] "0:midwest & 2:west , adjusted p-value=  0"
## [1] "0:midwest & 5:midwest , adjusted p-value=  0.029791"
## [1] "0:midwest & 5:northeast , adjusted p-value=  0"
## [1] "0:midwest & 5:south , adjusted p-value=  0.000773"
## [1] "0:midwest & 5:west , adjusted p-value=  0"
## [1] "0:northeast & 2:west , adjusted p-value=  0.002273"
## [1] "0:northeast & 5:northeast , adjusted p-value=  0"
## [1] "0:northeast & 5:west , adjusted p-value=  0.000286"
## [1] "0:south & 0:west , adjusted p-value=  1e-05"
## [1] "0:south & 2:midwest , adjusted p-value=  0.020512"
## [1] "0:south & 2:northeast , adjusted p-value=  0.000578"
## [1] "0:south & 2:south , adjusted p-value=  0.001522"
## [1] "0:south & 2:west , adjusted p-value=  0"
## [1] "0:south & 5:northeast , adjusted p-value=  0"
## [1] "0:south & 5:south , adjusted p-value=  0.005048"
## [1] "0:south & 5:west , adjusted p-value=  0"
## [1] "0:west & 5:northeast , adjusted p-value=  0.000935"
## [1] "2:midwest & 2:west , adjusted p-value=  0.002295"
## [1] "2:midwest & 5:northeast , adjusted p-value=  0"
## [1] "2:midwest & 5:west , adjusted p-value=  0.000275"
## [1] "2:northeast & 5:northeast , adjusted p-value=  1e-06"
## [1] "2:south & 2:west , adjusted p-value=  0.000169"
## [1] "2:south & 5:northeast , adjusted p-value=  0"
## [1] "2:south & 5:west , adjusted p-value=  8e-06"
## [1] "2:west & 5:midwest , adjusted p-value=  0.010123"
## [1] "2:west & 5:south , adjusted p-value=  0.011177"
## [1] "5:midwest & 5:northeast , adjusted p-value=  0"
## [1] "5:midwest & 5:west , adjusted p-value=  0.001935"
## [1] "5:northeast & 5:south , adjusted p-value=  0"
## [1] "5:south & 5:west , adjusted p-value=  0.00184"
##
## [1] "(Pct Patient Pays) The region-pairs ( Heart ) that are significantly different in mean value ar
## [1] "0:midwest & 0:west , adjusted p-value=  0.000562"
## [1] "0:midwest & 2:west , adjusted p-value=  1e-06"
```

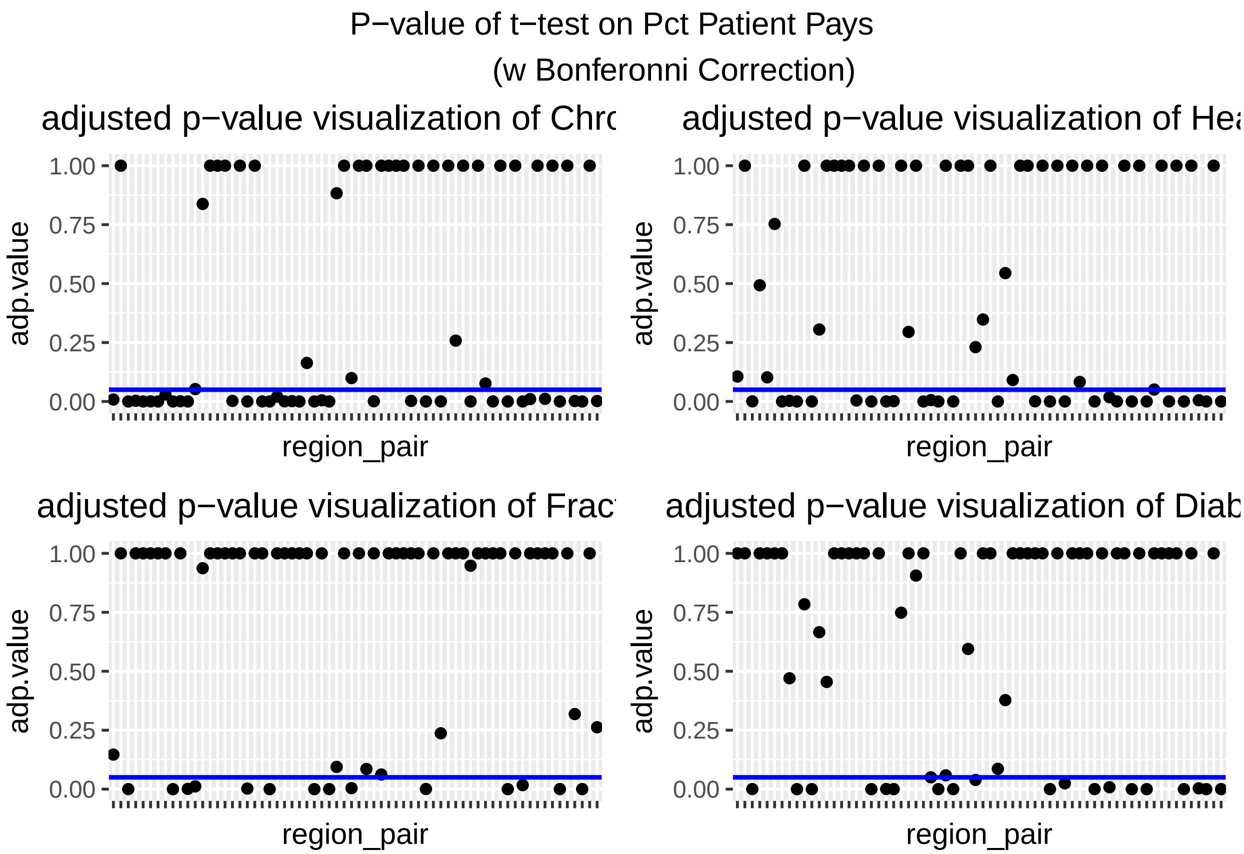
```
## [1] "0:midwest & 5:midwest , adjusted p-value=  0.002609"
## [1] "0:midwest & 5:northeast , adjusted p-value=  0"
## [1] "0:midwest & 5:west , adjusted p-value=  0"
## [1] "0:northeast & 2:west , adjusted p-value=  0.004248"
## [1] "0:northeast & 5:northeast , adjusted p-value=  0"
## [1] "0:northeast & 5:west , adjusted p-value=  0.000142"
## [1] "0:south & 0:west , adjusted p-value=  0.0013"
## [1] "0:south & 2:west , adjusted p-value=  2e-06"
## [1] "0:south & 5:midwest , adjusted p-value=  0.005789"
## [1] "0:south & 5:northeast , adjusted p-value=  0"
## [1] "0:south & 5:west , adjusted p-value=  0"
## [1] "0:west & 5:northeast , adjusted p-value=  2.7e-05"
## [1] "2:midwest & 2:west , adjusted p-value=  0.000573"
## [1] "2:midwest & 5:northeast , adjusted p-value=  0"
## [1] "2:midwest & 5:west , adjusted p-value=  5e-06"
## [1] "2:northeast & 5:northeast , adjusted p-value=  9e-06"
## [1] "2:northeast & 5:west , adjusted p-value=  0.018078"
## [1] "2:south & 2:west , adjusted p-value=  0.000113"
## [1] "2:south & 5:northeast , adjusted p-value=  0"
## [1] "2:south & 5:west , adjusted p-value=  0"
## [1] "2:west & 5:south , adjusted p-value=  0.000233"
## [1] "5:midwest & 5:northeast , adjusted p-value=  0"
## [1] "5:midwest & 5:west , adjusted p-value=  0.005278"
## [1] "5:northeast & 5:south , adjusted p-value=  0"
## [1] "5:south & 5:west , adjusted p-value=  1e-06"
##
## [1] "(Pct Patient Pays) The region-pairs ( Fracture ) that are significantly different in mean value
## [1] "0:midwest & 0:west , adjusted p-value=  8.6e-05"
## [1] "0:midwest & 5:northeast , adjusted p-value=  0"
## [1] "0:midwest & 5:west , adjusted p-value=  0.001126"
## [1] "0:northeast & 0:south , adjusted p-value=  0.011741"
## [1] "0:northeast & 5:northeast , adjusted p-value=  0.002311"
## [1] "0:south & 0:west , adjusted p-value=  2e-06"
## [1] "0:south & 5:northeast , adjusted p-value=  0"
## [1] "0:south & 5:west , adjusted p-value=  0.000117"
## [1] "0:west & 2:south , adjusted p-value=  0.004211"
## [1] "2:midwest & 5:northeast , adjusted p-value=  0.000534"
## [1] "2:south & 5:northeast , adjusted p-value=  1.1e-05"
## [1] "2:south & 5:west , adjusted p-value=  0.01699"
## [1] "5:midwest & 5:northeast , adjusted p-value=  0.000166"
## [1] "5:northeast & 5:south , adjusted p-value=  9.8e-05"
##
## [1] "(Pct Patient Pays) The region-pairs ( Diabete ) that are significantly different in mean value a
## [1] "0:midwest & 0:west , adjusted p-value=  0.000239"
## [1] "0:midwest & 5:northeast , adjusted p-value=  0"
## [1] "0:midwest & 5:west , adjusted p-value=  0"
## [1] "0:northeast & 5:northeast , adjusted p-value=  0"
## [1] "0:northeast & 5:west , adjusted p-value=  0.000507"
## [1] "0:south & 0:west , adjusted p-value=  8e-06"
## [1] "0:south & 5:northeast , adjusted p-value=  0"
## [1] "0:south & 5:west , adjusted p-value=  0"
## [1] "0:west & 2:south , adjusted p-value=  0.039"
## [1] "2:midwest & 5:northeast , adjusted p-value=  5.8e-05"
## [1] "2:midwest & 5:west , adjusted p-value=  0.024636"
```

```
## [1] "2:northeast & 5:northeast , adjusted p-value=  0.00012"
## [1] "2:northeast & 5:west , adjusted p-value=  0.007965"
## [1] "2:south & 5:northeast , adjusted p-value=  0"
## [1] "2:south & 5:west , adjusted p-value=  7e-06"
## [1] "5:midwest & 5:northeast , adjusted p-value=  0"
## [1] "5:midwest & 5:west , adjusted p-value=  0.00347"
## [1] "5:northeast & 5:south , adjusted p-value=  0"
## [1] "5:south & 5:west , adjusted p-value=  0.000187"
```

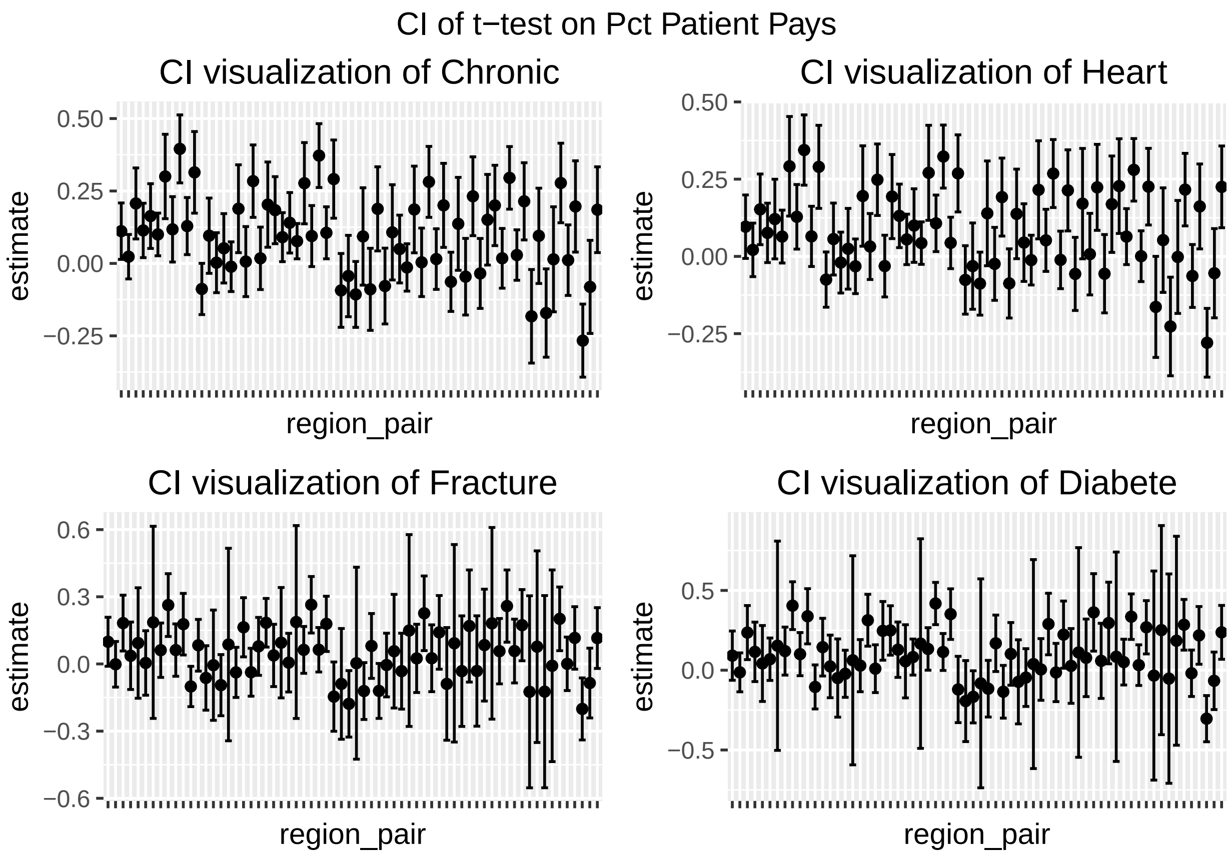
```r
grid.arrange(plot_pv[[1]], plot_pv[[2]], plot_pv[[3]], plot_pv[[4]],
    nrow = 2, ncol = 2, top = "P-value of t-test on Pct Patient Pays
             (w/o Bonferonni Correction)")
```




```r
grid.arrange(plot_adpv[[1]], plot_adpv[[2]], plot_adpv[[3]],
    plot_adpv[[4]], nrow = 2, ncol = 2, top = "P-value of t-test on Pct Patient Pays
             (w Bonferonni Correction)")
```




```r
grid.arrange(plot_CI[[1]], plot_CI[[2]], plot_CI[[3]], plot_CI[[4]],
    nrow = 2, ncol = 2, top = "CI of t-test on Pct Patient Pays")
```



Comments: Results of Pct Patient Pays data are different with that of Patient Pays. we can see in the first graph that before Bonferonni correction, there're a large number of pairs shows siginificant difference, even after correction, there're still many pairs that show show siginificant difference. So we can conclude with Bonferonni correction on our t-test results that there're more significant difference between groups in Patient Pays data than that in Pct Patient Pays data.

Comments: Now we focus on adjusted p-values. All four diagnose cases have many pairs that are significantly different (on which we can reject Null hypothesis). From printed siginificant;y different pairs, we see the precise numbers of those pairs.

Comments: we can also find that the CIs are of different width, some are much large. I guess it might be because of the lack of data (according to the central limit theorem that the variance of sample distributon should decrease with the sample number decreasing, thus the CIs should be narrow if there're enough data for certain pairs)

5. COMMENT ON THE OVERALL ANALYSIS

Comments: When I was doing this project, I find that after subseting the data of certain diagnoses that we want, our dataset is really small, which could result in the problem of lack of data when we want to make some conclusion, we could mistakenly "find" some relationships that could be caused by randomness. And in the following process, I did encounter that problem. When I want to perform some tests and make some inference, I found that the data in some cases are quite few, so that we can't make a solid assumption on them (such as normal distribution with central limit theorem). Even I have drawn some conclusions out of those data, it could not be solid.

Comments: Also, our project was focusing on the patient payment, and we need to figure out the relationship with the urban area and regions. However, the dataset is collected in different states and places, which could result in high varience because of different situations in different places in real-world, thus some relationships could be concealed and it would be harder for us to discover them.

Comments: I used t-test to make inference, and I think it would be better to use both t-test and permutation test to do double check. Comparing the results of those two tests we may also varify or demonstrate whether our inference is valid and whether out conclusion is solid.

Comments: To find more useful information, we can further use PCA on the data that we're interested in and maybe do some clustering to see if there're some indicative information.

Comments: After drawing conclusions, I didn't dig too deep into its meaning with the real world relationship. So if I need to improve it, this could be a direction for me.