Group 4 Report

Richard Davies, Katarina Lundervol, Pakwanja Desiree Twea

2024-09-25

## TidyR data setup

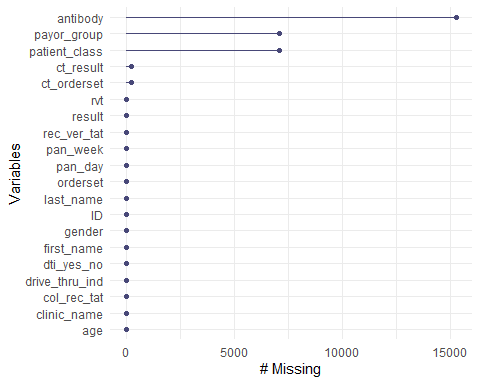
The data is composed of two datasets with other lapping patients (see ID), exam.dataset.txt consists of PCR rdts for COVID19 during the 2020 pandemic, data includes information on patients including clinic, gender, test results week of pandemic, and exam\_joindata.txt containing endpoint titer data from some of the patients (information can be found in the files codebook\_exam\_data.html, exam.descr.md). Data files were joined and processed as per conventions of tidyR for later visulization and analysis.

An overview of the data follows

summary(complete\_data)

## ID age gender first\_name   
## Min. : 1 Min. : 0.00 female:7832 Length:15524   
## 1st Qu.: 2330 1st Qu.: 2.00 male :7692 Class :character   
## Median : 5268 Median : 9.00 Mode :character   
## Mean : 5571 Mean : 14.19   
## 3rd Qu.: 8636 3rd Qu.: 18.00   
## Max. :12346 Max. :138.00   
##   
## last\_name clinic\_name result drive\_thru\_ind  
## Length:15524 clinical lab :7500 invalid : 301 0:7537   
## Class :character emergency dept :3413 negative:14358 1:7987   
## Mode :character oncology day hosp: 533 positive: 865   
## nicu : 294   
## laboratory : 270   
## picu : 261   
## (Other) :3253   
## ct\_result orderset payor\_group   
## Min. :14.05 Min. :0.0000 commercial :3726   
## 1st Qu.:45.00 1st Qu.:0.0000 government :3644   
## Median :45.00 Median :1.0000 unassigned : 733   
## Mean :44.12 Mean :0.6952 self pay : 216   
## 3rd Qu.:45.00 3rd Qu.:1.0000 medical assistance: 84   
## Max. :45.00 Max. :1.0000 (Other) : 34   
## NA's :209 NA's :7087   
## patient\_class pan\_day rec\_ver\_tat   
## inpatient :3438 Min. : 4.00 Min. :-18.600   
## emergency :1378 1st Qu.: 38.00 1st Qu.: 4.000   
## not applicable :1096 Median : 65.00 Median : 5.000   
## outpatient : 973 Mean : 63.21 Mean : 5.639   
## recurring outpatient: 795 3rd Qu.: 87.00 3rd Qu.: 6.200   
## (Other) : 767 Max. :107.00 Max. :218.200   
## NA's :7077   
## col\_rec\_tat rvt pan\_week dti\_yes\_no   
## Min. : 0.00 Length:15524 Min. : 0.5714 Length:15524   
## 1st Qu.: 0.70 Class :character 1st Qu.: 5.4286 Class :character   
## Median : 1.90 Mode :character Median : 9.2857 Mode :character   
## Mean : 7.22 Mean : 9.0298   
## 3rd Qu.: 3.60 3rd Qu.:12.4286   
## Max. :61370.20 Max. :15.2857   
##   
## ct\_orderset antibody   
## Min. : 0.0 Min. : 21.48   
## 1st Qu.: 0.0 1st Qu.: 63.30   
## Median :45.0 Median :103.52   
## Mean :30.6 Mean :109.99   
## 3rd Qu.:45.0 3rd Qu.:156.95   
## Max. :45.0 Max. :199.89   
## NA's :209 NA's :15275

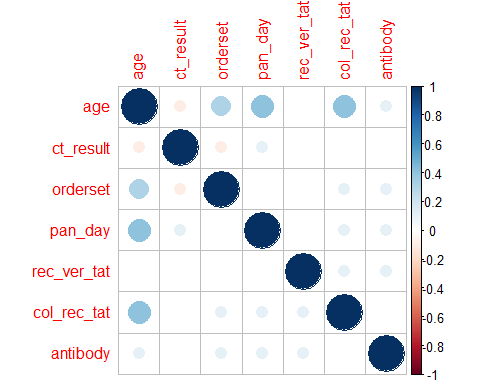
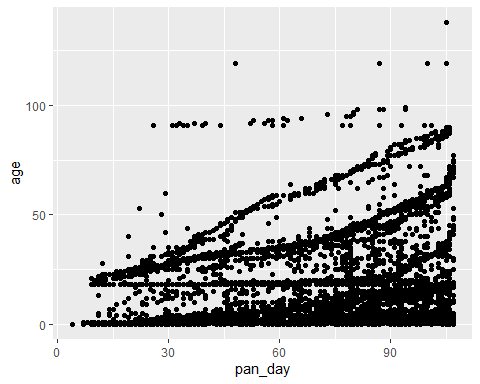
naniar::gg\_miss\_var(complete\_data)



## Descriptive plots

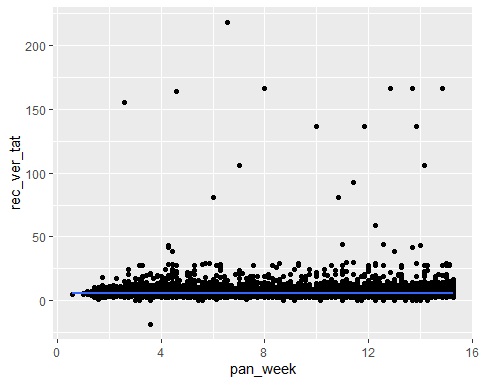
Visulization was done using a number of plots. Significant correlations are observiable with age, and orderset and days since the pandemic. Indicating early in the pandemic individuals seen in the clinic were usually young adults and children

## # A tibble: 7 × 8  
## rowname age ct\_result orderset pan\_day rec\_ver\_tat col\_rec\_tat  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 age 0 4.54e-17 9.34e-151 3.95e-323 0.0666 0.844  
## 2 ct\_result 4.54e- 17 0 3 e- 11 7.86e- 1 0.0829 0.849  
## 3 orderset 9.34e-151 3 e-11 0 7.38e- 1 0.000108 0.158  
## 4 pan\_day 3.95e-323 7.86e- 1 7.38e- 1 0 0.125 0.809  
## 5 rec\_ver\_tat 6.66e- 2 8.29e- 2 1.08e- 4 1.25e- 1 0 0.854  
## 6 col\_rec\_tat 8.44e- 1 8.49e- 1 1.58e- 1 8.09e- 1 0.854 0   
## 7 antibody 4.16e- 2 9.53e- 1 2.89e- 1 4.05e- 2 0.0704 0.623  
## # ℹ 1 more variable: antibody <dbl>

  Suprisingly over the course of the pandemic there was no indication that the wait time from collection of patient sample to PCR test result was reduced.

## # A tibble: 15 × 4  
## pan\_week mean\_value std `median(rec\_ver\_tat)`  
## <dbl> <dbl> <dbl> <dbl>  
## 1 1 5.51 1.57 5.3  
## 2 2 5.27 2.00 4.8  
## 3 3 5.62 5.44 5   
## 4 4 5.50 3.19 5   
## 5 5 5.57 5.07 5   
## 6 6 5.69 3.61 5.1  
## 7 7 5.73 7.81 4.9  
## 8 8 5.66 5.97 4.9  
## 9 9 5.48 2.59 4.9  
## 10 10 5.65 4.68 5   
## 11 11 5.73 4.17 5   
## 12 12 5.57 4.73 4.9  
## 13 13 5.59 5.22 4.9  
## 14 14 5.82 6.88 5   
## 15 15 5.81 5.62 5

## `geom\_smooth()` using formula = 'y ~ x'

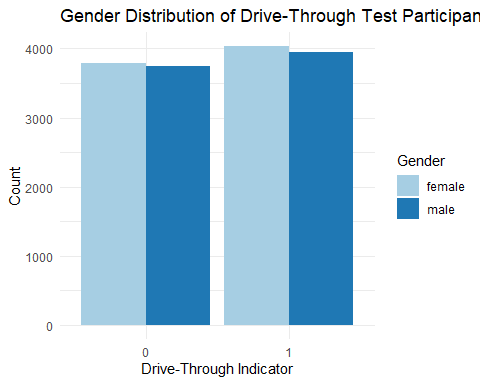


We observe that there are slightly more females than males in the cleaned data set. The bar plot below indicates that more people from both genders took the test at the drive-through than those who did not. However, from observation, the difference between genders is slightly larger among those who took the test at the drive-through compared to those who did not. To examine the relationship between gender and drive-through test participation, we conducted a Chi-squared test. With a p-value of 0.8728, we fail to reject the null hypothesis, suggesting that there is no significant association between gender and drive-through test participation in this data set.

dt\_and\_gender <- table(complete\_data$gender, complete\_data$drive\_thru\_ind)  
# Create dt\_and\_gender as a data frame from the table  
dt\_and\_gender\_df <- as.data.frame(dt\_and\_gender)  
print(chisq.test(dt\_and\_gender))

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: dt\_and\_gender  
## X-squared = 0.02564, df = 1, p-value = 0.8728

# Rename columns for clarity (since table() creates V1, V2, and Freq by default)  
colnames(dt\_and\_gender\_df) <- c("Gender", "Drive\_Through", "Count")  
  
# Now use ggplot2 to plot  
ggplot(dt\_and\_gender\_df, aes(x = Drive\_Through, y = Count, fill = Gender)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Gender Distribution of Drive-Through Test Participants",  
 x = "Drive-Through Indicator",   
 y = "Count") +  
 scale\_fill\_manual(values = c("#a6cee3", "#1f78b4")) + # Custom colors  
 theme\_minimal()



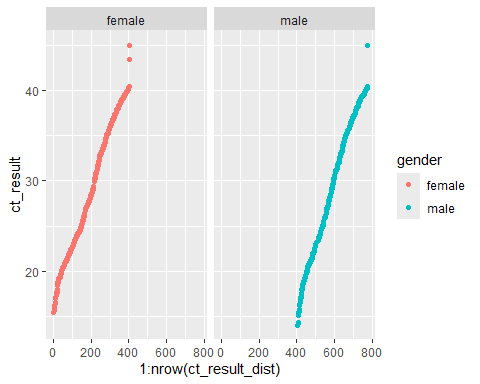
A comparison of cycle threshold values between men and women shows a similar distribution in the number of reaction cycles. However, males showed the lowest cycle threshold value compared to females. Despite this, the majority of threshold values exceeded 40 cycles for both genders. The p-value of 0.272 from the one-way ANOVA test means that we fail to reject the null hypothesis, suggesting that gender does not significantly affect the cycle threshold value.

ct\_result\_cat <- cut(complete\_data$ct\_result, c(10, 20, 30, 40, 50))  
summary(ct\_result\_cat)

## (10,20] (20,30] (30,40] (40,50] NA's   
## 92 346 415 14462 209

ct\_result\_dist <- complete\_data %>%  
 group\_by(gender) %>%  
 count(ct\_result)  
  
ggplot(ct\_result\_dist,   
 aes(  
 x = 1:nrow(ct\_result\_dist),  
 y = ct\_result,  
 group = gender,  
 color = gender)  
) +   
 geom\_point() +  
 facet\_wrap(vars(gender))

## Warning: Removed 2 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



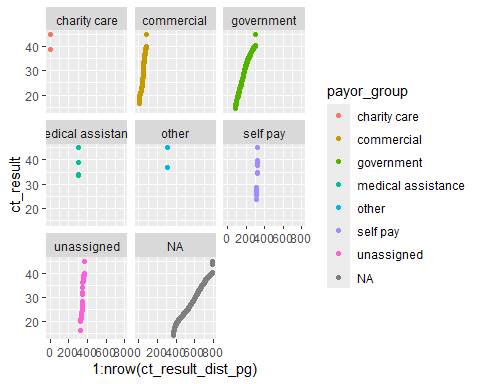
ct\_result\_gender\_anova <- aov(complete\_data$ct\_result ~ complete\_data$gender, data = complete\_data)  
summary(ct\_result\_gender\_anova)

## Df Sum Sq Mean Sq F value Pr(>F)  
## complete\_data$gender 1 19 19.11 1.206 0.272  
## Residuals 15313 242691 15.85   
## 209 observations deleted due to missingness

The distribution of patients by payor group shows that the most common groups were commercial and government, with a significant proportion of patients listed as NA or unassigned. Despite this, a similar pattern was observed between government and commercial patients, while for charity care, medical assistance, other, and self-pay patients the minimum cycle threshold value was higher compared to the former groups. The one-way ANOVA test result is highly significant, indicating that the cycle threshold value is influenced by the payor group.

ct\_result\_dist\_pg <- complete\_data %>%  
 group\_by(payor\_group) %>%  
 count(ct\_result)  
  
ggplot(  
 ct\_result\_dist\_pg,  
 aes(  
 x = 1:nrow(ct\_result\_dist\_pg),  
 y = ct\_result,  
 group = payor\_group,  
 color = payor\_group  
 )  
 ) +  
 geom\_point() +  
 facet\_wrap(vars(payor\_group))

## Warning: Removed 6 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



ct\_result\_pg\_anova <- aov(complete\_data$ct\_result ~ complete\_data$payor\_group, data = complete\_data)  
summary(ct\_result\_pg\_anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## complete\_data$payor\_group 6 672 111.9 9.1 6.03e-10 \*\*\*  
## Residuals 8296 102040 12.3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 7221 observations deleted due to missingness

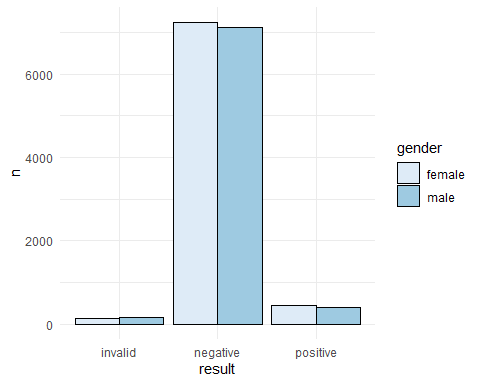
The distribution of results by gender shows a similar pattern for both, with women comprising a slightly larger proportion in each result category except for the invalid category. The chi-square test yielded a p-value of 0.5481, meaning that we fail to reject the null hypothesis, suggesting that gender does not have a significant influence on the test result.

tests\_by\_gender <- complete\_data %>%  
 group\_by(gender) %>%  
 count(result)  
  
tests\_by\_gender

## # A tibble: 6 × 3  
## # Groups: gender [2]  
## gender result n  
## <fct> <fct> <int>  
## 1 female invalid 146  
## 2 female negative 7237  
## 3 female positive 449  
## 4 male invalid 155  
## 5 male negative 7121  
## 6 male positive 416

ggplot(data = tests\_by\_gender, aes(x = result, y = n, fill = gender)) +  
 geom\_bar(stat = "identity", color="black", position=position\_dodge())+  
 theme\_minimal() +  
 scale\_fill\_manual(values=c('#999999','#E69F00')) +  
# Use brewer color palettes  
 scale\_fill\_brewer(palette="Blues")

## Scale for fill is already present.  
## Adding another scale for fill, which will replace the existing scale.



test\_and\_gender <- table(complete\_data$gender, complete\_data$result)  
test\_and\_gender

##   
## invalid negative positive  
## female 146 7237 449  
## male 155 7121 416

#chi-square test  
test\_and\_gender\_df <- as.data.frame(test\_and\_gender)  
print(chisq.test(test\_and\_gender))

##   
## Pearson's Chi-squared test  
##   
## data: test\_and\_gender  
## X-squared = 1.2028, df = 2, p-value = 0.5481

## Statistics

wilcox test indicated that the more tests from a single patient does not associate with increase positive test, in fact the opposite trend was noted.

## # A tibble: 1 × 4  
## statistic p.value method alternative  
## <dbl> <dbl> <chr> <chr>   
## 1 4979785 8.14e-11 Wilcoxon rank sum test with continuity correct… two.sided

While drive-through use was associated with an increase in positive tests

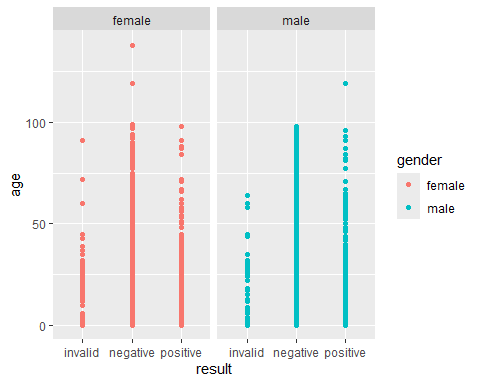
##   
## Pearson's Chi-squared test with simulated p-value (based on 2000  
## replicates)  
##   
## data: positive\_visits$n and positive\_visits$positive  
## X-squared = 44.749, df = NA, p-value = 0.01349

A visual inspection of the distribution suggests that age may be associated with the test result, as there appears to be a positive correlation between age distribution and test outcomes. However, the dataset is skewed toward younger individuals, so the correlation may be due to the age structure. To determine whether this association holds statistically, we performed an ANOVA test. The results showed a significant p-value, indicating that we cannot rule out the possibility of an association between age and test result.

complete\_data$age\_cat <- cut(complete\_data$age,   
 breaks = c(0, 9, 19, 29, 39, 49, 59, 69, 79, 89, 99, Inf),  
 labels = c("0-9", "10-19", "20-29", "30-39",   
 "40-49", "50-59", "60-69", "70-79",   
 "80-89", "90-99", "100+"))  
by(complete\_data$age\_cat, complete\_data$result, summary)

## complete\_data$result: invalid  
## 0-9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89 90-99 100+ NA's   
## 143 60 49 29 8 1 4 2 0 2 0 3   
## ------------------------------------------------------------   
## complete\_data$result: negative  
## 0-9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89 90-99 100+ NA's   
## 7325 3675 935 968 489 370 183 73 56 32 4 248   
## ------------------------------------------------------------   
## complete\_data$result: positive  
## 0-9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89 90-99 100+ NA's   
## 296 278 85 99 22 30 16 4 9 13 1 12

ggplot(complete\_data,   
 aes(  
 x = result,  
 y = age,  
 group = gender,  
 color = gender)  
) +   
 geom\_point() +  
 facet\_wrap(vars(gender))



# ANOVA  
result\_age\_aov <- aov(age ~ result, data = complete\_data)  
result\_age\_aov

## Call:  
## aov(formula = age ~ result, data = complete\_data)  
##   
## Terms:  
## result Residuals  
## Sum of Squares 22781 4189232  
## Deg. of Freedom 2 15521  
##   
## Residual standard error: 16.42886  
## Estimated effects may be unbalanced

summary(result\_age\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## result 2 22781 11391 42.2 <2e-16 \*\*\*  
## Residuals 15521 4189232 270   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The table below shows the distribution of cycle threshold (Ct) results and test outcomes. As seen earlier, the majority of individuals in this dataset have a Ct value between 40 and 50. Interestingly, this range also has the lowest number of positive test results compared to those with lower Ct values. Additionally, the ANOVA test shows a significant result, suggesting a difference in the distribution of Ct values across different result outcome groups.

result\_ct\_table <- complete\_data %>%  
 mutate(ct\_result\_cat = cut(ct\_result, breaks = c(10, 20, 30, 40, 50))) %>%  
 count(result, ct\_result\_cat) %>%  
 pivot\_wider(names\_from = ct\_result\_cat, values\_from = n, values\_fill = 0)  
  
# View the table  
print(result\_ct\_table)

## # A tibble: 3 × 6  
## result `(40,50]` `NA` `(30,40]` `(10,20]` `(20,30]`  
## <fct> <int> <int> <int> <int> <int>  
## 1 invalid 229 72 0 0 0  
## 2 negative 14213 132 13 0 0  
## 3 positive 20 5 402 92 346

ct\_result\_ct\_anova <- aov(complete\_data$ct\_result ~ complete\_data$result, data = complete\_data)  
summary(ct\_result\_ct\_anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## complete\_data$result 2 195741 97870 31905 <2e-16 \*\*\*  
## Residuals 15312 46970 3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 209 observations deleted due to missingness