question2

DAY

packages

```r
library(tidyverse) #for data cleaning
library(car) #for MANOVA and Levene's test
library(rstatix) #tidy stats
library(ggpubr) #for creating some plots
library(ltm) #for biserial correlation
library(GGally) #for creating some plots
library(patchwork) #for combining figures
```

read in the data

```r
car_raw <- read_csv("car_data.csv")
```

preview the data

```r
head(car_raw)
```

```
# A tibble: 6 x 4
  v_gender CarColor Blink_or_Honk RespTimeSeconds
     <dbl>    <dbl>         <dbl>           <dbl>
1        0        5             0           11.6
2        0        4             1           11.3
3        0        5             1           10.2
4        0        5             0           11.0
5        1        5             1            8.70
6        0        4             1           10.6
```

summary statistics

```r
summary(car_raw)
```

```
    v_gender          CarColor      Blink_or_Honk    RespTimeSeconds
 Min.   :0.0000   Min.   :1.000   Min.   :0.0000   Min.   : 1.958
 1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 5.601
 Median :1.0000   Median :3.000   Median :0.0000   Median : 6.716
 Mean   :0.5782   Mean   :3.005   Mean   :0.2512   Mean   : 6.809
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:0.7500   3rd Qu.: 8.049
 Max.   :1.0000   Max.   :5.000   Max.   :1.0000   Max.   :11.563
```

are there any NA or Missing values? – no

```r
sum(is.na(car_raw))
```

```
[1] 0
```

```r
anyNA(car_raw)
```

```
[1] FALSE
```

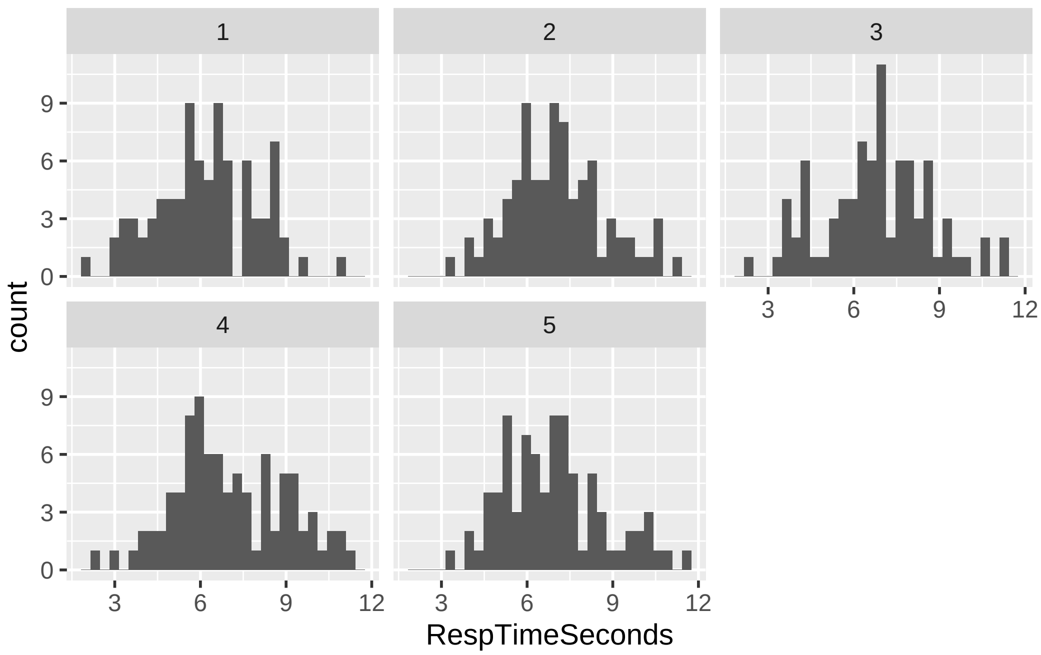
converting variables to appropriate formats

```r
## converting variables to appropriate formats
## the below code converts the 3 variables from numeric to factors

car_raw <- car_raw %>%
  mutate(v_gender = factor(v_gender, levels = c(0,1), ordered = TRUE),
         Blink_or_Honk = factor(Blink_or_Honk, levels = c(0,1), ordered = TRUE),
         CarColor = factor(CarColor, levels = 1:5, ordered = TRUE))
```

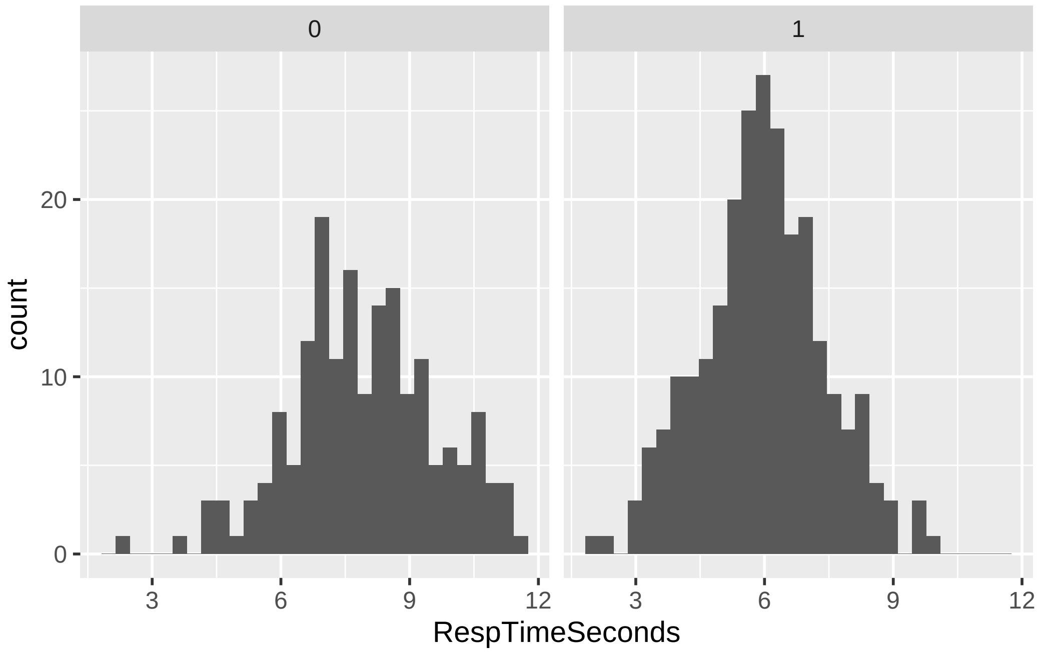
let's look at the distributions of the variables

Response Time (in seconds) – appears normally distributed

```
# Response Time
# appears normally distributed

#Car Color
ggplot(car_raw, aes(x = RespTimeSeconds)) +
  geom_histogram() +
  facet_wrap(~CarColor)
```

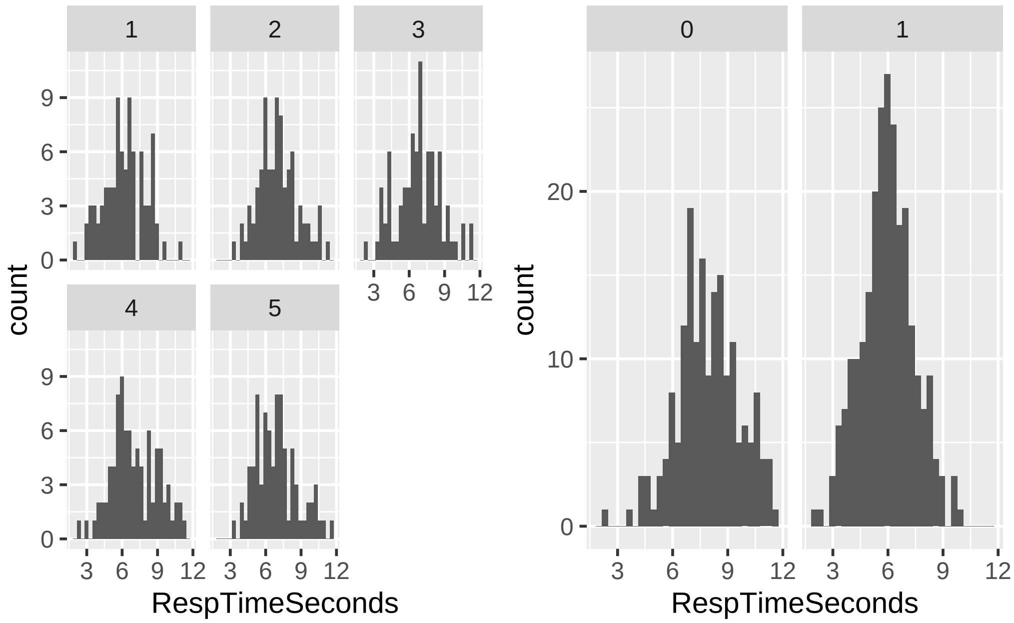



```
#Gender of Driver
ggplot(car_raw, aes(x = RespTimeSeconds)) +
  geom_histogram() +
  facet_wrap(~v_gender)
```




```
#combine the figures
p1 <- ggplot(car_raw, aes(x = RespTimeSeconds)) +
  geom_histogram() +
  facet_wrap(~CarColor)

p2 <- ggplot(car_raw, aes(x = RespTimeSeconds)) +
  geom_histogram() +
  facet_wrap(~v_gender)

p1 + p2
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

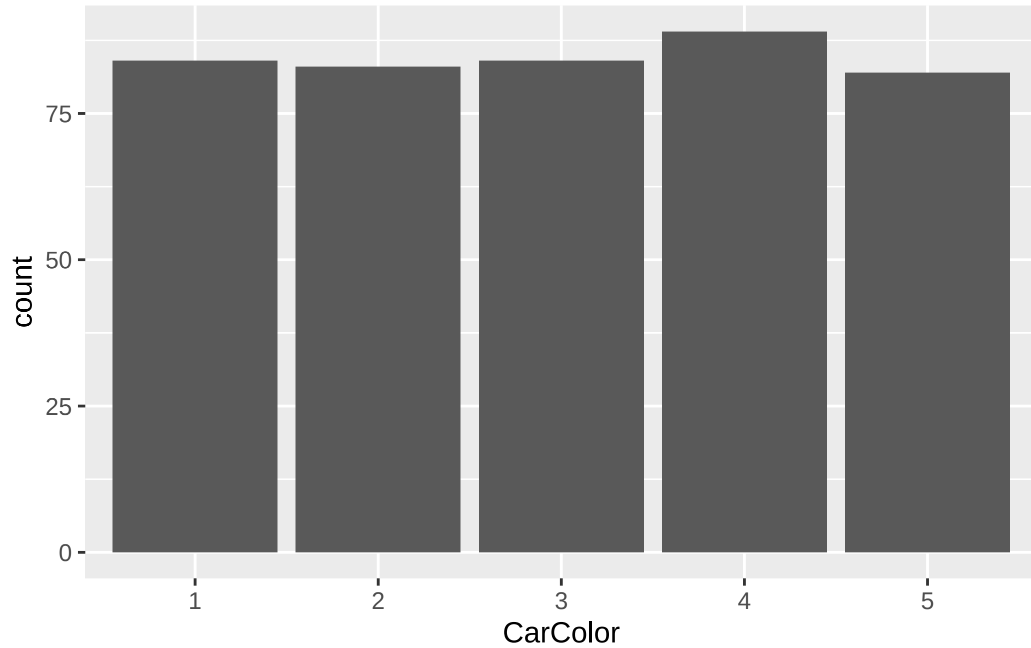


Car Color – approximately equal group sizes

```r
## Car Color
## group sizes look equal
table(car_raw$CarColor)
```

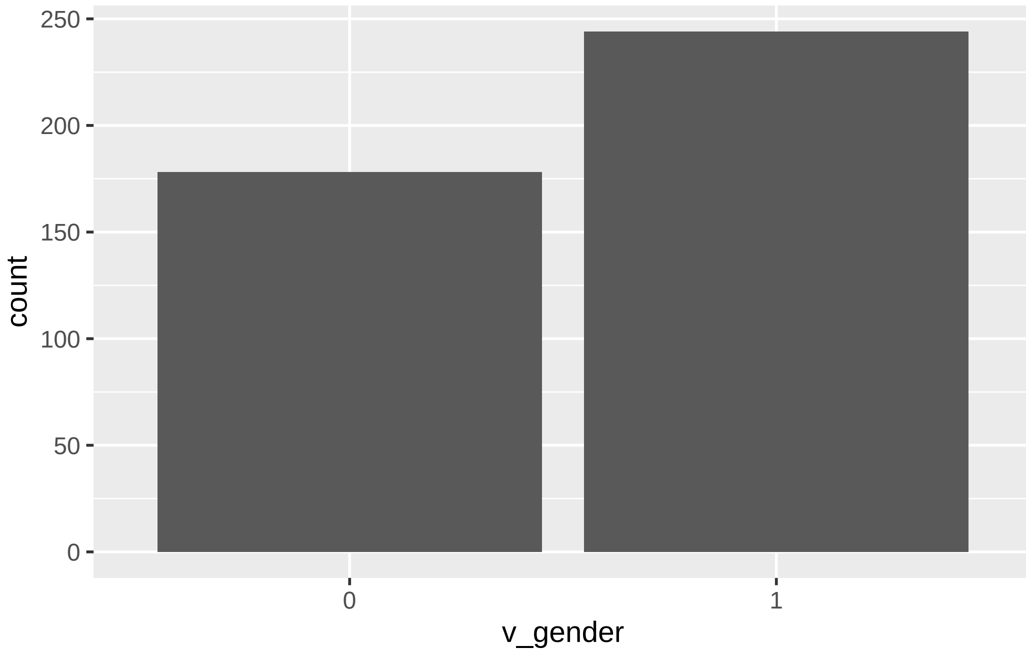
```
 1  2  3  4  5 
84 83 84 89 82 
```

```r
##let's visualize that
ggplot(car_raw, aes(x = CarColor)) +
  geom_bar()
```



Gender of Driver – 178 female, 244 male

```
ggplot(car_raw, aes(x = v_gender)) +
  geom_bar()
```




```
table(car_raw$v_gender)
```

```
  0   1
178 244
```

```
# 0 == female
# 1 == male
```

List of MANOVA assumptions:

1. Adequate sample size
2. Independence of the observations
3. Absense of univariate or multivariate outliers
4. Univariate and Multivariate normality
5. Absence of multicollinearity
6. Linearity between outcome variables
7. Homogeneity of variances
8. Homogeneity of variance-covariance matrices

Let's test each assumption below

Assumption 1 / Adequate Sample Size : `Satisfied`

```
car_raw %>%
  group_by(CarColor) %>%
  summarise(N = n())
```

```
# A tibble: 5 x 2
  CarColor     N
  <ord>    <int>
1 1           84
2 2           83
3 3           84
4 4           89
5 5           82
```

```
car_raw %>%
  group_by(v_gender) %>%
  summarise(N = n())
```

```
# A tibble: 2 x 2
  v_gender     N
  <ord>    <int>
1 0          178
2 1          244
```

Assumption 1 states that number of observations in each group should be greater than the number of outcome variables. This assumption is clearly met in the data (see above tables).

Assumption 2 / Independence of Observations : `Satisfied`

Each row is an independent observation, so this assumption is satisfied.

Assumption 3 / Absense of outliers : `Satisfied for Response Time DV`

Test for univariate outliers : Box Plot Method

Response Time Variable

```
car_raw %>%
  group_by(CarColor) %>%
  identify_outliers(RespTimeSeconds)
```

```
# A tibble: 1 x 6
  CarColor v_gender Blink_or_Honk RespTimeSeconds is.outlier is.extreme
  <ord>    <ord>    <ord>                   <dbl> <lgl>      <lgl>
1 5        0        0                        11.6 TRUE       FALSE
```

This univariate outlier test identified one outlier, but there results show that it is considered "not extreme." As a result, I think it is fine to leave this "outlier" in the dataset. We cannot test for univariate outliers in the other DV (Driver's Reaction of Blinking or Honking) because it is a binary/categorical variable.

Test for multivariate outliers : Mahalanobis Distance

```
car_raw %>%
 group_by(CarColor) %>%
 mahalanobis_distance() %>%
 filter(is.outlier == TRUE) %>%
  as.data.frame()
```

```
[1] RespTimeSeconds mahal.dist      is.outlier
<0 rows> (or 0-length row.names)
```

According to the Mahalanobis Distance test, there appears to be no multivariate outliers in this dataset.

Assumption 4 / Univariate and Multivariate Normality : `Satisfied for Response Time DV`

Shapiro-Wilks Test for Univariate Normality

Grouped by Car Color

```
car_raw %>%
  group_by(CarColor) %>%
  shapiro_test(RespTimeSeconds)
```

```
# A tibble: 5 x 4
  CarColor variable        statistic      p
  <ord>    <chr>               <dbl>  <dbl>
1 1        RespTimeSeconds     0.991 0.849
2 2        RespTimeSeconds     0.983 0.363
3 3        RespTimeSeconds     0.989 0.705
4 4        RespTimeSeconds     0.981 0.230
5 5        RespTimeSeconds     0.973 0.0798
```

Grouped by gender of Driver

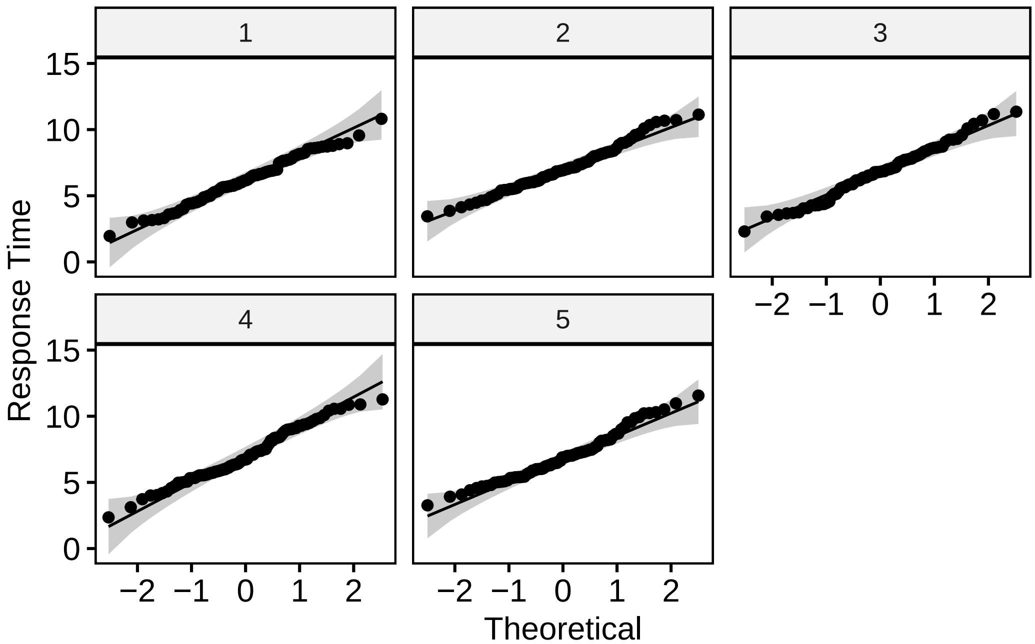
```r
car_raw %>%
  group_by(v_gender) %>%
  shapiro_test(RespTimeSeconds)
```

```
# A tibble: 2 x 4
  v_gender variable        statistic     p
  <ord>    <chr>               <dbl> <dbl>
1 0        RespTimeSeconds     0.990 0.236
2 1        RespTimeSeconds     0.996 0.773
```

According to the Shapiro-Wilks Test, Response Time is normally distributed for each group ($p > 0.05$ for each group). We cannot run the Shapiro-Wilks test on the other DV (Driver's Reaction of Blinking or Honking) because it is a binary/categorical variable.

Here are some QQ plots to visualize the univariate normality of the Response Time variable:

```r
ggqqplot(car_raw, "RespTimeSeconds", facet.by = "CarColor",
         ylab = "Response Time")
```



Shapiro test for multivariate normality

```
mshapiro_test(car_raw$RespTimeSeconds)
```

```
# A tibble: 1 x 2
  statistic p.value
      <dbl>   <dbl>
1     0.993  0.0509
```

The p value of the shapiro test is slightly greater than 0.05, which indicates that it is not significant, and we can assume multivariate normality. We cannot perform this test on the other dependent variable (Driver's reaction of blink or honk) because it is a binary/categorical variable.

Assumption 5 / Absense of Multicollinearity : `Satisfied`

We have two outcome variables: Response Time and Driver's Reaction (Blink or Honk). Since Driver's Reaction is a binary/categorical variable and Response Time is a continuous variable, we will need to conduct a point-biserial correlation to evaluate multicollinearity.

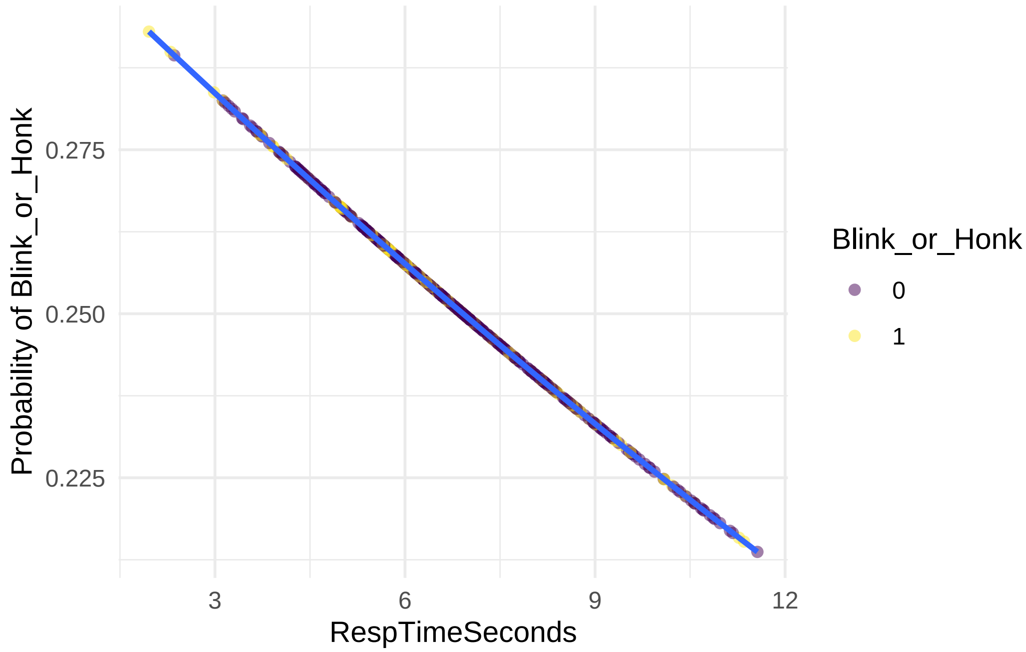
```r
biserial.cor(car_raw$RespTimeSeconds, car_raw$Blink_or_Honk, use = "all.obs")
```

```
[1] 0.03467538
```

From the biserial correlation, we can see that the two dependent variables are only slightly positively correlated, so we do not have multicollinearity of DVs in this dataset.

Assumption 6 / Linearity between outcome variables: `Satisfied`

```r
# Fit logistic regression model
logit_model <- glm(Blink_or_Honk ~ RespTimeSeconds, data = car_raw, family = "binomial")

# Predict probabilities
probabilities <- predict(logit_model, type = "response")

# Create a data frame for plotting
plot_data <- data.frame(RespTimeSeconds = car_raw$RespTimeSeconds,
                        probability = probabilities,
                        Blink_or_Honk = car_raw$Blink_or_Honk)

# Plotting the logistic regression curve
ggplot(plot_data, aes(x = RespTimeSeconds, y = probability)) +
  geom_point(aes(color = factor(Blink_or_Honk)), alpha = 0.5) +  # Add points colored by B
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +  # Ad
  labs(x = "RespTimeSeconds", y = "Probability of Blink_or_Honk", color = "Blink_or_Honk")
  theme_minimal()
```



We will have to use logistic regression to look at the linear relationship between Response Time and the log odds of the outcome, the driver's response (blink or honk). It appears linear, so this assumption is satisfied.

Assumption 7 / Homogeneity of Variances: `Not able to conduct this test`

The Levene's test of equality of variances assumes continuous dependent variables, so we are not able to assess the homogeneity of variances in this case because we have one continuous dependent variable (Response Time) and one binary/categorical dependent variable (Blink or Honk response).

Assumption 8 / Homogeneity of Covariances: `Not able to conduct this test`

We are also not able to conduct Box's M test because it is designed for continuous variables and depends on the variance of those continuous variables. ## MANOVA

```
## MANCOVA model
model <- manova(cbind(RespTimeSeconds, Blink_or_Honk) ~ CarColor + v_gender, data = car_ra

summary(model)
```

```
           Df   Pillai approx F num Df den Df    Pr(>F)    
CarColor    4 0.075298    4.069      8    832 9.149e-05 ***
v_gender    1 0.296780   87.571      2    415 < 2.2e-16 ***
Residuals 416                                              
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#m2 <- lm(cbind(RespTimeSeconds,Blink_or_Honk) ~ CarColor + v_gender,  data = car_raw)
#coef(m2)
```

Overall MANOVA Test:

For CarColor, Pillai's trace is 0.075298 with an approximate F statistic of 4.069. For v_gender, Pillai's trace is 0.296780 with an approximate F statistic of 87.571.

In summary, the MANOVA results indicate that both CarColor and v_gender have significant effects on the dependent variables. However, MANOVA may not be appropriate here because we have a binary dependent variable.

Trying SEM instead

```r
library(lavaan)
```

```
This is lavaan 0.6-17
lavaan is FREE software! Please report any bugs.
```

```r
# Specify the SEM model
model <- "
    # Measurement model for CarColor
    latent_car_color =~ CarColor

    # Structural model
    Blink_or_Honk ~ latent_car_color + v_gender
    RespTimeSeconds ~ latent_car_color + v_gender
"

# Fit the SEM model
fit <- sem(model, data = car_raw)
```

```
Warning in lav_data_full(data = data, group = group, cluster = cluster, :
lavaan WARNING: exogenous variable(s) declared as ordered in data: v_gender
```

```
Warning in lav_model_vcov(lavmodel = lavmodel, lavsamplestats = lavsamplestats, : lavaan WARN
    Could not compute standard errors! The information matrix could
    not be inverted. This may be a symptom that the model is not
    identified.
```

```
Warning in lav_test_satorra_bentler(lavobject = NULL, lavsamplestats = lavsamplestats, : lava
```

```
summary(fit)
```

```
lavaan 0.6.17 ended normally after 25 iterations

  Estimator                                         DWLS
  Optimization method                             NLMINB
  Number of model parameters                          13

  Number of observations                             422

Model Test User Model:
                                              Standard      Scaled
  Test Statistic                                 0.032          NA
  Degrees of freedom                                 0           0

Parameter Estimates:

  Parameterization                               Delta
  Standard errors                           Robust.sem
  Information                                 Expected
  Information saturated (h1) model        Unstructured

Latent Variables:
                      Estimate  Std.Err  z-value  P(>|z|)
  latent_car_color =~
    CarColor             1.000

Regressions:
                    Estimate  Std.Err  z-value  P(>|z|)
  Blink_or_Honk ~
    latent_car_clr    -0.386       NA
```

```
    v_gender           0.208       NA
  RespTimeSeconds ~
    latent_car_clr     0.406       NA
    v_gender          -1.954       NA

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
 .Blink_or_Honk ~~
   .RespTimeSecnds    0.104       NA

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .RespTimeSecnds    9.893       NA

Thresholds:
                   Estimate  Std.Err  z-value  P(>|z|)
    CarColor|t1      -0.816       NA
    CarColor|t2      -0.235       NA
    CarColor|t3       0.269       NA
    CarColor|t4       0.891       NA
    Blink_r_Hnk|t1    1.002       NA

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .CarColor          0.418
   .Blink_or_Honk     0.913
   .RespTimeSecnds    2.319       NA
    latent_car_clr    0.582       NA
```

The model did not converge properly, leading to “NA” values in the parameter estimates. Since we are not super familiar with SEM, it is difficult to troubleshoot here.