ANOVA Practice Solutions

```
library(tidyverse)
```

```
## -- Attaching packages ----------- tidyverse 1.2.1 --

## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1. Training Program

First I'll import the `TrainingProgram` dataset. I'll `attach` it so its easy to refer to the columns, then I'll look at the first few rows.

```
TrainingProgram <- read_csv("TrainingProgram.csv")
```

```
## Parsed with column specification:
## cols(
##   Scores = col_double(),
##   Exam = col_character()
## )
```

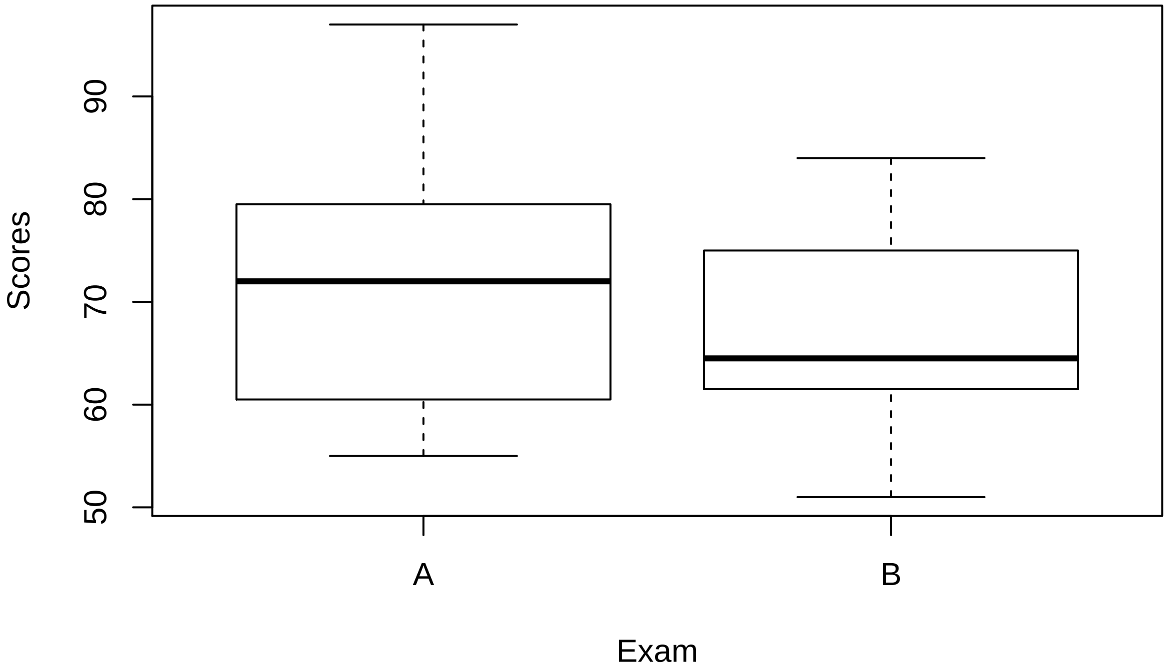
```
attach(TrainingProgram)
head(TrainingProgram)
```

```
## # A tibble: 6 x 2
##   Scores Exam
##    <dbl> <chr>
## 1     66 A
## 2     74 A
## 3     59 B
## 4     62 B
## 5     82 A
## 6     75 A
```

Independence The people were randomly assigned to two training programs, so the samples are independent.

Normality We can use boxplots to help us decide if the distributions seem approximately symmetric.

```r
boxplot(Scores ~ Exam)
```



Both seem reasonably symmetric with no outliers, so we can assume they are approximately normal.

Standard Deviations remember the `tapply` command

```r
tapply(Scores, Exam, sd)
```

```
##         A         B
## 11.671725  9.099222
```

these are fine.

Now run the test.

```r
myAnalysis <- aov(Scores ~ Exam)
summary(myAnalysis)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Exam         1    144   144.5   1.319   0.26
## Residuals   30   3285   109.5
```

The p-value is rather large, so we do not have evidence that the mean scores are different in the two groups.

Aisle location

Import the dataset, `attach` it, look at the first few rows.

```r
AisleLocation <- read_csv("AisleLocation.csv")
```

```
## Parsed with column specification:
## cols(
##   Time = col_double(),
##   Location = col_character()
## )
```

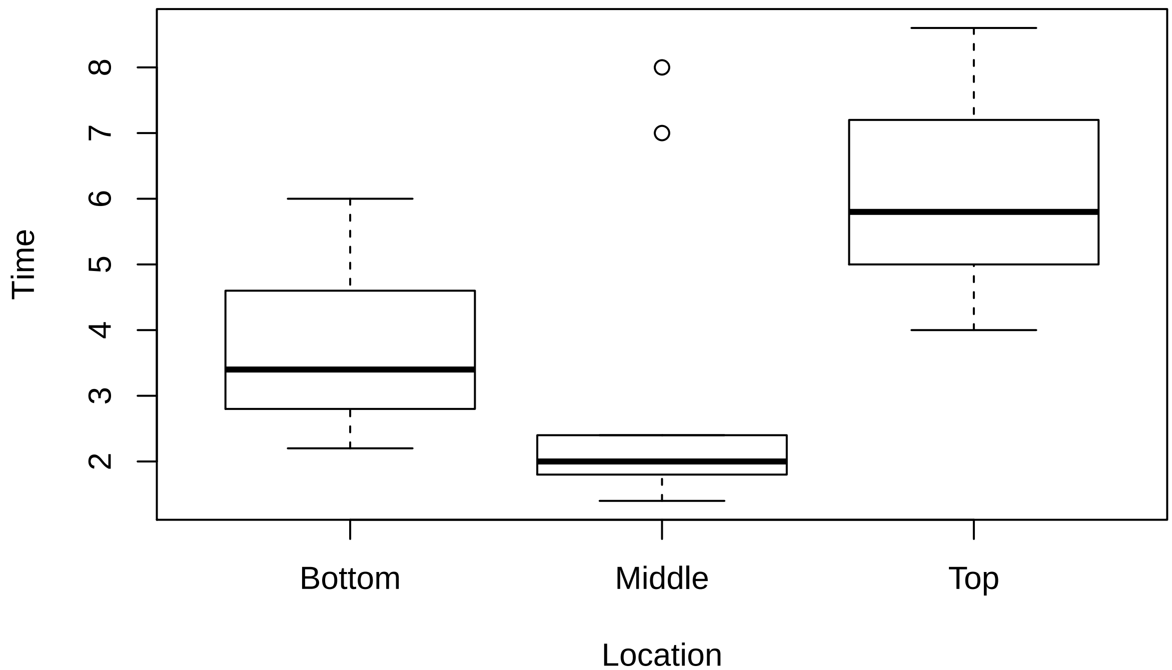
```r
attach(AisleLocation)
head(AisleLocation)
```

```
## # A tibble: 6 x 2
##    Time Location
##   <dbl> <chr>
## 1   8.6 Top
## 2   7.2 Top
## 3   6.2 Top
## 4   5.4 Top
## 5   5   Top
## 6   4   Top
```

Independence Seems a reasonable assumption since subjects were assigned randomly.

Normality Let's check boxplots.

```r
boxplot(Time ~ Location)
```



The `Middle` category does not look at all normal. It is not symmetric and has some outliers. I don't think the normality assumption is met, so we cannot use the ANOVA test.

I would recommend we collect more data with more subjects if possible.

New Packaging

Import the dataset, `attach` it, look at the first few rows.

```r
Packaging <- read_csv("Packaging.csv")
```

```
## Parsed with column specification:
## cols(
##   Rating = col_double(),
##   Design = col_character()
## )
```

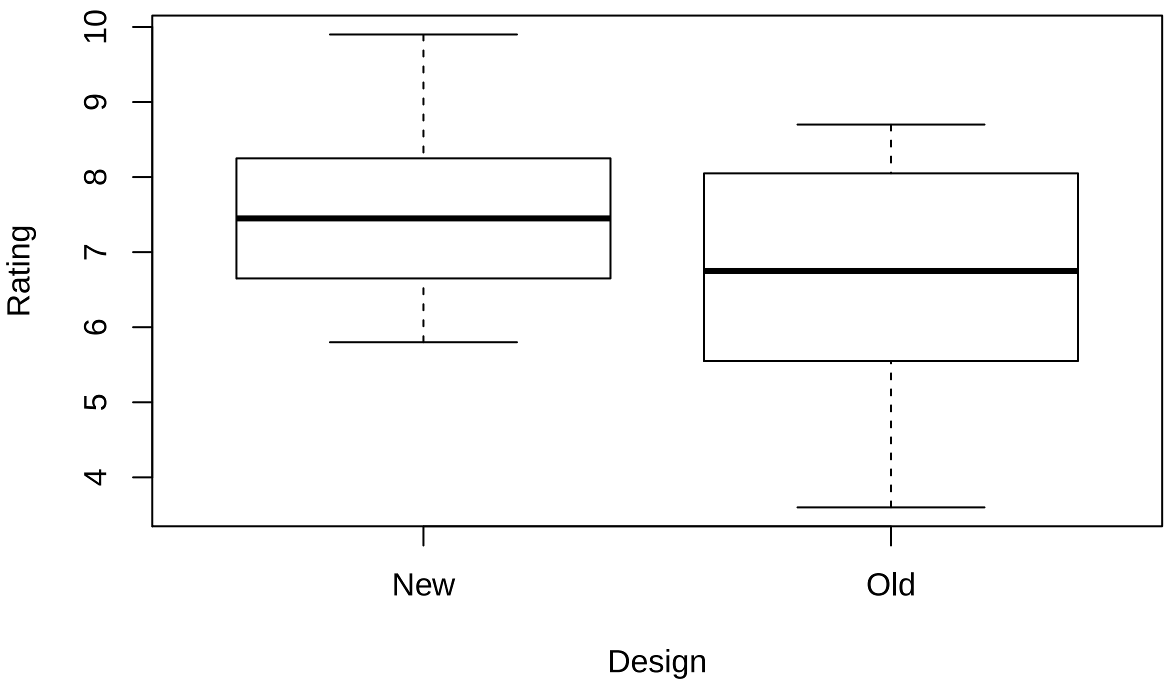
```r
attach(Packaging)
head(Packaging)
```

```
## # A tibble: 6 x 2
##   Rating Design
##    <dbl> <chr>
## 1    6.7 New
## 2    8.6 New
## 3    6.2 New
## 4    7.6 New
## 5    6.1 New
## 6    5.8 New
```

Independence Seems a reasonable assumption since subjects were assigned randomly.

Normality Let's check boxplots.

```r
boxplot(Rating ~ Design)
```



`Old` might have a slight skew, but both are reasonably symmetric, so we are ok with the normality assumption.

Standard Deviations

```r
tapply(Rating, Design, sd)
```

```
##      New      Old
## 1.157129 1.504686
```

These are fine.

Now run the test

```r
packageAnalysis <- aov(Rating ~ Design)
summary(packageAnalysis)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Design       1   8.56   8.556   4.749 0.0356 *
```

```
## Residuals   38  68.46   1.802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05, so we have some moderate evidence that there is a difference in the mean values between the two package designs.