R Notebook

Source: https://gist.github.com/andrewheiss/faadbc3e737547961d6e16e2c2512867

```
devtools::install_github("seankross/lego")
```

```
library(tidyverse)
library(lego)  # https://github.com/seankross/lego
library(broom)
library(glue)
library(pander)

# Shrink and clean this dataset
lego_clean <- legosets %>%
  filter(Year > 2011) %>%
  mutate(Boxed = ifelse(Packaging == "Box", "Boxed", "Not boxed"))
```

```
# Just as a toy example, check if there are differences in prices between boxed
# and non-boxed LEGO sets in 2012-2015
lego_tests <- lego_clean %>%
  group_by(Year) %>%
  nest() %>%
  # This runs a t-test
  mutate(t_test = data %>% map(~ t.test(USD_MSRP ~ Boxed, data = .)),
         # This converts the t-test to a tibble
         t_test_tidy = t_test %>% map(tidy))
lego_tests
```

```
#> # A tibble: 4 x 4
#>    Year data               t_test      t_test_tidy
#>   <int> <list>             <list>      <list>
#> 1  2015 <tibble [417 x 14]> <S3: htest> <tibble [1 x 10]>
#> 2  2014 <tibble [434 x 14]> <S3: htest> <tibble [1 x 10]>
#> 3  2013 <tibble [409 x 14]> <S3: htest> <tibble [1 x 10]>
#> 4  2012 <tibble [417 x 14]> <S3: htest> <tibble [1 x 10]>
```

```
# unnest the tibble list-column
lego_tests %>% unnest(data)
```

```
#> # A tibble: 1,677 x 15
#>     Year Item_Number Name  Theme Subtheme Pieces Minifigures Image_URL
#>    <int> <chr>       <chr> <chr> <chr>     <int>       <int> <chr>
#>  1  2015 10246       Dete~ Adva~ Modular~   2262           6 http://i~
#>  2  2015 10247       Ferr~ Adva~ Fairgro~   2464          10 http://i~
#>  3  2015 10248       Ferr~ Adva~ Vehicles   1158          NA http://i~
#>  4  2015 10249       Toy ~ Adva~ Winter ~    898          NA http://i~
#>  5  2015 10581       Ducks Duplo Forest ~     13           1 http://i~
#>  6  2015 10582       Anim~ Duplo Forest ~     39           2 http://i~
#>  7  2015 10583       Fish~ Duplo Forest ~     32           2 http://i~
#>  8  2015 10584       Fore~ Duplo Forest ~    105           3 http://i~
#>  9  2015 10585       Mom ~ Duplo ""           13           2 http://i~
#> 10  2015 10586       Ice ~ Duplo ""           11           2 http://i~
#> # ... with 1,667 more rows, and 7 more variables: GBP_MSRP <dbl>,
#> #   USD_MSRP <dbl>, CAD_MSRP <dbl>, EUR_MSRP <dbl>, Packaging <chr>,
```

```
#> #   Availability <chr>, Boxed <chr>
```

```
# unnest the tibble list-column
# Unnest so that we can verify that there's a row of test results for each year
lego_tests %>% unnest(t_test_tidy)
```

```
#> # A tibble: 4 x 13
#>    Year data  t_test estimate estimate1 estimate2 statistic  p.value
#>   <int> <lis> <list>    <dbl>     <dbl>     <dbl>     <dbl>    <dbl>
#> 1  2015 <tib~ <S3: ~     26.4      43.2     16.8       6.30 1.33e- 9
#> 2  2014 <tib~ <S3: ~     32.9      42.2      9.34     11.7  1.93e-26
#> 3  2013 <tib~ <S3: ~     19.5      45.7     26.2       2.92 3.75e- 3
#> 4  2012 <tib~ <S3: ~     33.1      40.5      7.37     11.9  4.45e-26
#> # ... with 5 more variables: parameter <dbl>, conf.low <dbl>,
#> #   conf.high <dbl>, method <chr>, alternative <chr>
```

```
# Now we can use pandoc to make a nice table of these results
#
# Little helper function for formatting the results of a t-test:
#
# diff
# (upper, lower)
# p = p_value
#
# The "\\ \n" sequence will add a Markdown line break
format_t_test <- function(x) {
  glue("{diff}\\ \n({upper}, {lower})\\ \np = {p_value}",
       diff = round(x$estimate, 2),
       upper = round(x$conf.high, 2),
       lower = round(x$conf.low, 2),
       p_value = round(x$p.value, 3))
}
```

```
# Make the table!
lego_tests %>%
  # Format the tidied t-test into a fancy character string with format_t_test()
  mutate(clean_results = t_test_tidy %>% map_chr(format_t_test)) %>%
  # Only choose two columns. Add a fancy two-line header with LaTeX in it, just for fun
  select(Year, `Difference in average price\\ \n$\\Delta_{\\text{Boxed} - \\text{Not boxed}}$` = clean_
  pandoc.table(split.table = Inf,  # Make sure the table isn't split horizontally
               justify = "lc",  # Align the columns left and centered
               style = "multiline",  # Allow for multiple lines in cells
               # Keep any existing line breaks (like the \\ \n) we created
               keep.line.breaks = TRUE,
               # Add a caption
               caption = "Difference in means, 95% confidence interval, and p-value shown")
```

```
#>
#> -------------------------------------
#> Year    Difference in average price\
#>            $\Delta_{\text{Boxed} -
#>              \text{Not boxed}}$
#> ------ ------------------------------
#> 2015              26.39\
#>               (34.64, 18.14)\
#>                     p = 0
```

```
#>
#> 2014              32.91\
#>               (38.46, 27.35)\
#>                    p = 0
#>
#> 2013              19.46\
#>                (32.58, 6.34)\
#>                  p = 0.004
#>
#> 2012              33.12\
#>               (38.59, 27.65)\
#>                    p = 0
#> ---------------------------------------
#>
#> Table: Difference in means, 95% confidence interval, and p-value shown
```