Lab 9: T test and ANOVA

Your name and student ID

today's date

Instructions

- Due date: Tuesday, April 13 at 10:00pm PST with 2 hour grace period.
- Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for assignments submitted thereafter.
- This assignment is graded on **correct completion**, all or nothing. You must pass all public tests and submit the assignment for credit.
- Submission process: Follow the submission instructions on the final page. Make sure you do not remove any `\newpage` tags or rename this file, as this will break the submission.

Part 1: T test and NHANES

The NHANES is a large national survey conducted by the CDC. Here we will look at a reduced set of data from the NHANES

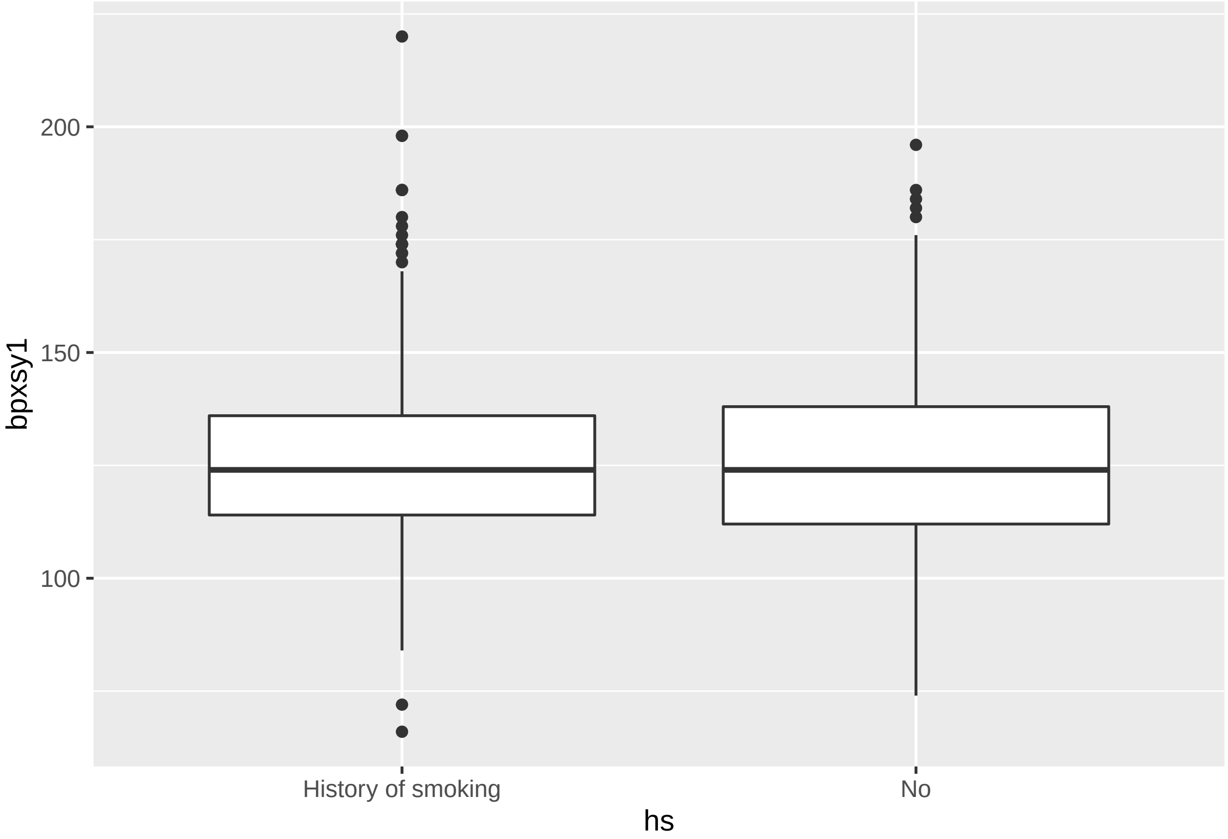
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   ridageyr = col_double(),
##   drinks = col_double(),
##   bmxwt = col_double(),
##   bmxht = col_double(),
##   bmxbmi = col_double(),
##   bpxpls = col_double(),
##   bpxsy1 = col_double(),
##   bpxsy2 = col_double(),
##   bpxdi1 = col_double(),
##   bpxdi2 = col_double(),
##   lbdhdd = col_double(),
##   sleep = col_double(),
##   lbdldl = col_double()
## )
## i Use `spec()` for the full column specifications.
```

1. [1 mark] We are interested in looking at the systolic blood pressure "bpxsy1" by history of smoking "hs" Start by generating an appropriate box plot to look at these data.

```
nhanes
```

```
## # A tibble: 1,178 x 40
##    ridageyr agegroup gender military  born  citizen drinks drinkscat bmxwt bmxht
##       <dbl> <chr>    <chr>  <chr>     <chr> <chr>    <dbl> <chr>     <dbl> <dbl>
##  1       72 65+      Male   History ~ Born~ US cit~      0 0          88.9  175.
##  2       73 65+      Female No        Born~ US cit~      0 0          52    162.
##  3       61 50-64    Female No        Born~ US cit~      2 11-Jan     93.4  162.
##  4       50 50-64    Male   No        No    No           0 0          80.9  185
##  5       57 50-64    Female No        No    US cit~    104 96-364    104    165.
##  6       75 65+      Male   History ~ Born~ US cit~      0 0         112.   170.
##  7       43 35-49    Male   History ~ Born~ US cit~    782 365+       90.2  177.
##  8       54 50-64    Female No        Born~ US cit~    365 365+       79.4  156.
##  9       80 65+      Male   History ~ Born~ US cit~    261 96-364     76.4  176.
## 10       63 50-64    Female No        Born~ US cit~      0 0          60.9  157.
## # ... with 1,168 more rows, and 30 more variables: bmxbmi <dbl>, bmicat <chr>,
## #   bpxpls <dbl>, bpxsy1 <dbl>, bpxsy2 <dbl>, sys1d <chr>, sys2d <chr>,
## #   bpxdi1 <dbl>, bpxdi2 <dbl>, dias1d <chr>, dias2d <chr>, bpcat <chr>,
## #   chest <chr>, fs1 <chr>, fs2 <chr>, fs3 <chr>, lbdhdd <dbl>, hdlcat <chr>,
## #   highhdl <chr>, hi <chr>, asthma <chr>, vwa <chr>, vra <chr>, va <chr>,
## #   aspirin <chr>, sleep <dbl>, is <chr>, hs <chr>, lbdldl <dbl>, highldl <chr>
```

```
plot1 <- ggplot(nhanes, aes(x = hs, y = bpxsy1)) + geom_boxplot() # YOUR CODE HERE
plot1
```



```r
. = ottr::check("tests/p1.R")
```

```
##
## Attaching package: 'testthat'
```

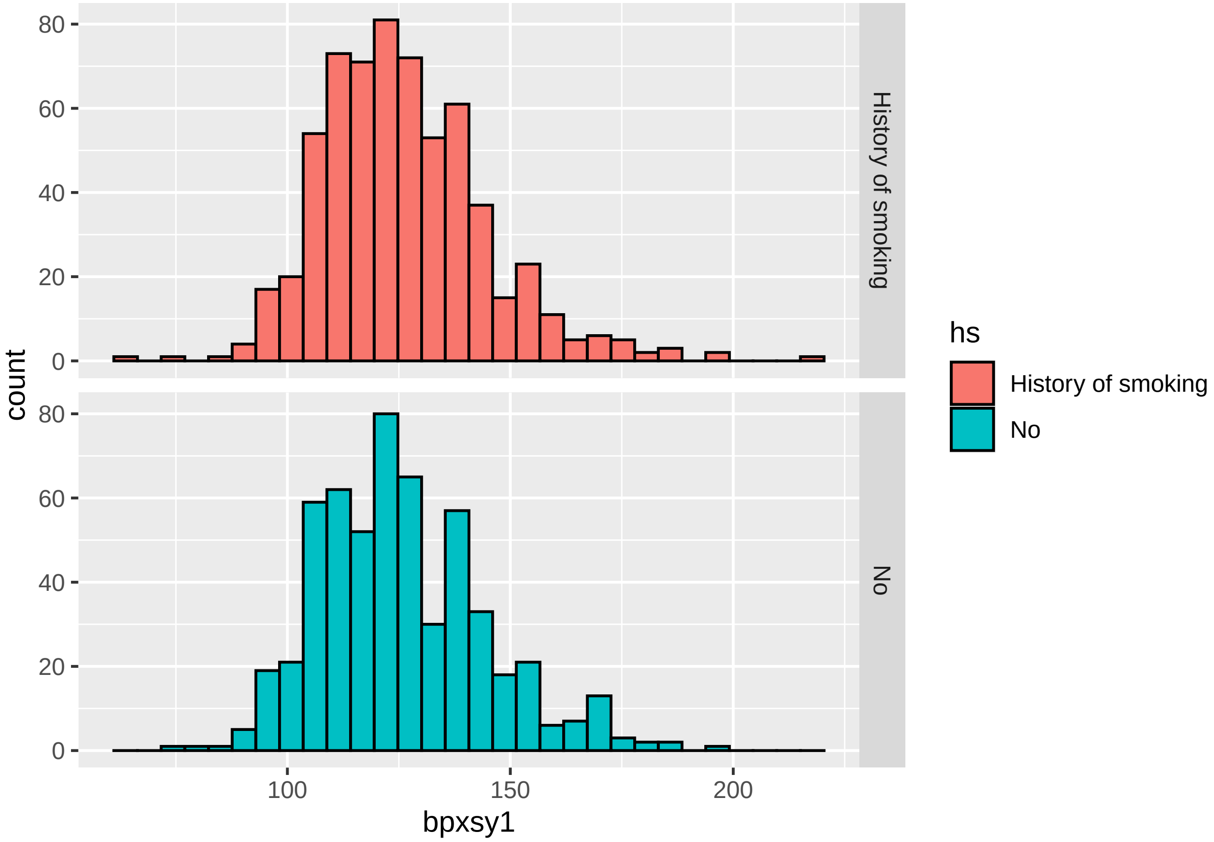
```
## The following object is masked from 'package:dplyr':
##
##     matches
```

```
## [1] "Checking: ggplot defined"
## Test passed
## [1] "Checking: nhanes data used"
## Test passed
## [1] "Checking:history of smoking on x axis"
## Test passed
## [1] "Checking: systolic blood pressure on y axis"
## Test passed
## [1] "Checking: boxplot created"
## Test passed
## All tests passed!
```

2. [1 mark] Now generate a set of faceted histograms that show the same data.

```
plot2 <- ggplot(nhanes, aes(x = bpxsy1)) +
  geom_histogram(aes(fill = hs), col = "black") +
  facet_grid(hs~.)
plot2
```

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



```
. = ottr::check("tests/p2.R")
```

```
## [1] "Checking: ggplot defined"
## Test passed
## [1] "Checking: nhanes data used"
## Test passed
## [1] "Checking: blood pressure on x axis on x axis"
## Test passed
## [1] "Checking: facet used"
## Test passed
## All tests passed!
```

3. [2 marks] Summarize the means and standard deviations of the outcome for each BMI category using dplyr functions. p3 should end up being a dataframe - check this in your environment.

```
# Should be smoking history
p3 <- nhanes %>% group_by(hs) %>% summarise(avg = mean(bpxsy1), std = sd(bpxsy1))
p3
```

```
## # A tibble: 2 x 3
##   hs                    avg   std
##   <chr>               <dbl> <dbl>
## 1 History of smoking  126.   18.6
## 2 No                  126.   18.7
```

```
. = ottr::check("tests/p3.R")
```

```
## [1] "Checking: dataframe created"
## Test passed
## [1] "Checking: group_by used correctly"
## Test passed
## [1] "Checking: summarize - columns for mean and sd created"
## Test passed
## All tests passed!
```

4. [1 mark] Now consider the assumptions that need to be hold in order to run the two-sample t-test. Do they hold here? Why or why not?

The assumptions to run the two-sample t test are that the observations are independent and the shape of each sample is normally distributed with an unknown population mean and sd. The assumptions hold here as we are looking at a large cohort study and we can assume each participant is independent. This may not hold if multiple participants came from the same household and thus affected their smoking history and blood pressure. The shapes of the data are normal and we do not know the population mean and SD.

5. [2 marks] State your null and alternative hypotheses.

$H_0 =$ The mean of blood pressure for those with a history of smoking is the same for those without a smoking history. $\mu_s = \mu_n$.

$H_A =$ The mean of blood pressure for those without a history of smoking is different than those without a smoking history. $\mu_s \neq \mu_n$.

6. [2 marks] Now run the t-test to see if the variability gives us evidence to reject the null hypothesis of no difference between blood pressure means by smoking history.

```
p6 <- t.test(bpxsy1~hs, data=nhanes)
p6
```

```
##
##  Welch Two Sample t-test
##
## data:  bpxsy1 by hs
## t = 0.23094, df = 1161.9, p-value = 0.8174
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.883164  2.385630
## sample estimates:
## mean in group History of smoking                 mean in group No
##                         126.1260                         125.8748
```

```
. = ottr::check("tests/p6.R")
```

```
## [1] "Checking: t.test function used"
## Test passed
## [1] "Checking: value of test statistic to at least 3 decimals"
## Test passed
## All tests passed!
```

7. [2 marks] Use these results to interpret your p-value in context. Following that, decide whether to accept or reject the null hypothesis.

The pvalue for the current test is p-value = 0.8174, meaning there is more than 81% chance to have a more extreme(larger) value for the two-sample t statistic than the observed one. Therefore we don't have enough evidence to reject the null hypothesis.

8. [2 marks] Repeat this analysis without using the t.test function.

First, you will need to get your test statistic:

```r
#this code gives you the number of smokers in the dataset
n_s <- nrow(nhanes %>% filter(hs == 'History of smoking'))
n_s
```

```
## [1] 619
```

```r
#this code gives you the number of non-smokers in the dataset
n_ns <- nrow(nhanes %>% filter(hs == 'No'))
n_ns
```

```
## [1] 559
```

```r
#calculate your test statistic. You can make more objects if you wish.
se <- sqrt((18.56617^2/619) + (18.71515^2/559))
t_stat <- 0.2512/se
t_stat
```

```
## [1] 0.2309112
```

```r
. = ottr::check("tests/p8-1.R")
```

```
## [1] "Checking: value of test statistic to at least 3 decimals"
## Test passed
## All tests passed!
```

Now compare your test statistic to a t-distribution with df = 558 and calculate the p-value. This is an approximation using the smaller of the two sample sizes - 1.

```r
df <- n_ns - 1
df
```

```
## [1] 558
```

```r
p_value <- pt(t_stat, df = df, lower.tail = F) * 2
p_value
```

```
## [1] 0.8174684
```

```r
. = ottr::check("tests/p8-2.R")
```

```
## [1] "Checking: range of p-value"
## Test passed
## [1] "Checking: value of p-value to at least 3 decimals"
## Test passed
## All tests passed!
```

9. [2 marks] Finally, construct a 99% confidence interval for these data. Interpret the interval in context and decide whether or not to reject the null hypothesis based on these data.

```
critical_val <- qt(p = 0.995, df = df)
lowerbound <- 0.2512 - critical_val*se
upperbound <- 0.2512 + critical_val*se
conf_int <- c(lowerbound, upperbound)
conf_int
```

```
## [1] -2.560568  3.062968
```

```
. = ottr::check("tests/p9.R")
```

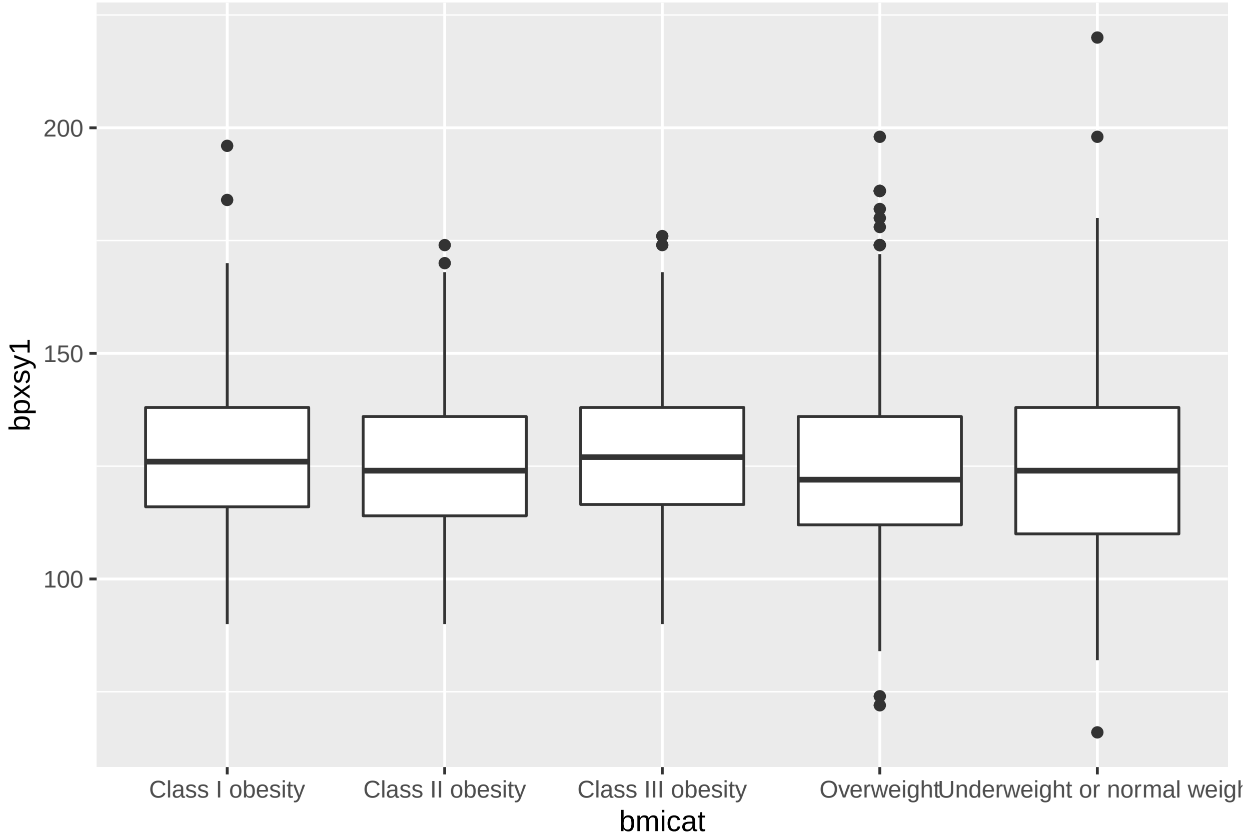
```
## [1] "Checking: order of inputs"
## Test passed
## [1] "Checking: value of lowerbound to at least 3 decimals"
## Test passed
## [1] "Checking: value of upperbound to at least 3 decimals"
## Test passed
## All tests passed!
```

If our assumptions hold and we were to repeat this process 100 times, 99 of them would contain the true value of the difference in population means. This confidence interval is (-2.58, 3.04) which contains 0 and thus we fail to reject the null hypothesis of no difference in mean blood pressure in smokers vs non-smokers.

Part 2: ANOVA

10. [1 mark] We are interested in looking at the systolic blood pressure "bpxsy1" by BMI category "bmicat" Start by generating an appropriate box plot to look at these data.

```
plot10 <- ggplot(data = nhanes, aes(x = bmicat, y = bpxsy1)) + geom_boxplot() # SOLUTION
plot10
```



```
. = ottr::check("tests/p10.R")
```

```
## [1] "Checking: ggplot defined"
## Test passed
## [1] "Checking: nhanes data used"
## Test passed
## [1] "Checking:cmicat on x axis"
## Test passed
## [1] "Checking: systolic blood pressure on y axis"
## Test passed
## [1] "Checking: boxplot created"
## Test passed
## All tests passed!
```

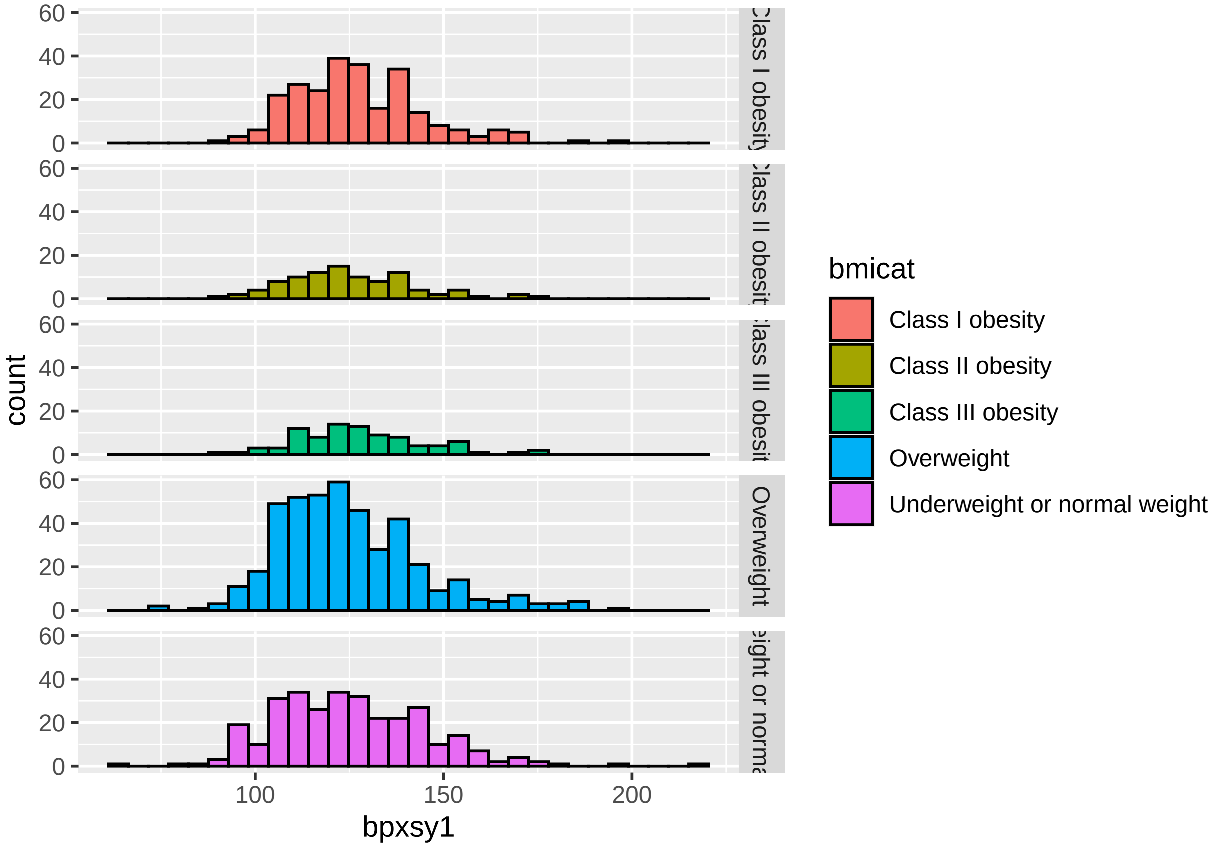
11. [1 mark] Now generate a set of faceted histograms that show the same data. It might be useful to assign a fill color to each category.

```
plot11 <- "Your plot here"
plot11
```

```
## [1] "Your plot here"
```

```
# BEGIN SOLUTION NO PROMPT
plot11 <- ggplot(data = nhanes, aes(x = bpxsy1)) +
  geom_histogram(aes(fill = bmicat), col = "black",
                 position = "stack") + facet_grid(bmicat~.)
plot11
```

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



```
# END SOLUTION
```

```
. = ottr::check("tests/p11.R")
```

```
## [1] "Checking: ggplot defined"
## Test passed
## [1] "Checking: nhanes data used"
## Test passed
## [1] "Checking: blood pressure on x axis on x axis"
## Test passed
## [1] "Checking: facet_wrap or facet_grid used"
## Test passed
## All tests passed!
```

12. [2 marks] Summarize the means and standard deviations of the outcome for each BMI category

```
p12 <- "Your code here"
p12
```

```
## [1] "Your code here"
```

```
# BEGIN SOLUTION NO PROMPT
p12 <- nhanes %>% group_by(bmicat) %>% summarize(mean_bp = mean(bpxsy1),
                                                 sd_bp = sd(bpxsy1))
p12
```

```
## # A tibble: 5 x 3
##   bmicat                       mean_bp sd_bp
##   <chr>                          <dbl> <dbl>
## 1 Class I obesity                 128.  17.0
## 2 Class II obesity                126.  16.9
## 3 Class III obesity               128.  17.0
## 4 Overweight                      125.  19.0
## 5 Underweight or normal weight    125.  20.3
```

```
# END SOLUTION
```

```
. = ottr::check("tests/p12.R")
```

```
## [1] "Checking: dataframe created"
## Test passed
## [1] "Checking: group_by used correctly"
## Test passed
## [1] "Checking: summarize - columns for mean and sd created"
## Test passed
## All tests passed!
```

13. [2 marks] Now run the ANOVA test to see if the variability gives us evidence to reject the null hypothesis of no difference between blood pressure means by BMI category.

```
p13 <- aov(bpxsy1 ~ bmicat, data = nhanes)
tidy(p13) #tidy displays your output. It lives in the `broom` package
```

```
## # A tibble: 2 x 6
##   term         df   sumsq meansq statistic p.value
##   <chr>     <dbl>   <dbl>  <dbl>     <dbl>   <dbl>
## 1 bmicat        4   1651.   413.      1.19   0.314
## 2 Residuals  1173 406837.   347.     NA     NA
```

```
. = ottr::check("tests/p13.R")
```

```
## [1] "Checking: aov function used"
## Test passed
## [1] "Checking: formula syntax"
## Test passed
## [1] "Checking: formula syntax"
## Test passed
## [1] "Checking: values"
## Test passed
## All tests passed!
```

14. [2 marks] Use these results to conduct a Tukey's HSD for these groups. Use the standard error rate of 5%. What conclusion can you draw?

```
p14 <- TukeyHSD(p13) # SOLUTION
tidy(p14)
```

```
## # A tibble: 10 x 7
##    term   contrast              null.value estimate conf.low conf.high adj.p.value
##    <chr>  <chr>                      <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
##  1 bmicat Class II obesity-C~            0  -2.09      -8.19      4.01       0.883
##  2 bmicat Class III obesity--~           0   0.638     -5.61      6.89       0.999
##  3 bmicat Overweight-Class I~            0  -2.60      -6.63      1.43       0.396
##  4 bmicat Underweight or nor~            0  -2.18      -6.51      2.16       0.646
##  5 bmicat Class III obesity--~           0   2.73      -4.74     10.2        0.856
##  6 bmicat Overweight-Class I~            0  -0.510     -6.25      5.23       0.999
##  7 bmicat Underweight or nor~            0  -0.0871    -6.04      5.87       1.00
##  8 bmicat Overweight-Class I~            0  -3.24      -9.13      2.66       0.562
##  9 bmicat Underweight or nor~            0  -2.81      -8.92      3.29       0.716
## 10 bmicat Underweight or nor~            0   0.423     -3.38      4.22       0.998
```

```
. = ottr::check("tests/p14.R")
```

```
## [1] "Checking: tukeyHSD function used"
## Test passed
## [1] "Checking: adjusted p-values"
## Test passed
## All tests passed!
```

All of the pairwise comparisons contain 0 so we can conclude the variability in the means of blood pressure is no difference between bmi categories.

Submission

For assignments in this class, you'll be submitting using the **Terminal** tab in the pane below. In order for the submission to work properly, make sure that:

1. Any image files you add that are needed to knit the file are in the `src` folder and file paths are specified accordingly.
2. You **have not changed the file name** of the assignment.
3. The file is saved (the file name in the tab should be **black**, not red with an asterisk).
4. The file knits properly.

Once you have checked these items, you can proceed to submit your assignment.

1. Click on the **Terminal** tab in the pane below.
2. Copy-paste the following line of code into the terminal and press enter.

```
cd; cd ph142-sp21/lab/lab09; python3 turn_in.py
```

3. Follow the prompts to enter your Gradescope username and password. When entering your password, you won't see anything come up on the screen–don't worry! This is just for security purposes–just keep typing and hit enter.
4. If the submission is successful, you should see "Submission successful!" appear as output.
5. If the submission fails, try to diagnose the issue using the error messages–if you have problems, post on Piazza.

The late policy will be strictly enforced, **no matter the reason**, including submission issues, so be sure to submit early enough to have time to diagnose issues if problems arise.