Final Project

```r
data = read_csv("Sleep_Analysis.csv")
dim(data)
```

```
[1] 46 11
```

```r
names(data)
```

```
 [1] "Age"                "Gender"             "meals/day"          "physical illness"
 [5] "screen time"        "bluelight filter"   "sleep direction"    "exercise"
 [9] "smoke/drink"        "beverage"           "sleep time"
```

```r
head(data)
```

```
# A tibble: 6 x 11
    Age Gender `meals/day` `physical illness` `screen time` `bluelight filter`
  <dbl> <chr>  <chr>       <chr>              <chr>         <chr>
1    22 Male   two         no                 2hrs          yes
2    22 Female three       no                 3-4 hrs       no
3    23 Male   three       no                 3-4 hrs       no
4    23 Female two         no                 1-2 hrs       no
5    22 Male   three       no                 more than 5   yes
6    22 Male   two         no                 2-3 hrs       yes
# i 5 more variables: `sleep direction` <chr>, exercise <chr>,
#   `smoke/drink` <chr>, beverage <chr>, `sleep time` <dbl>
```

There are a couple of rows that appear to be close-duplicates (only difference is sleep time: one is 2 hours (the only non-binned value), and the other is 2-3 hours (which includes the non-binned value)). In addition, they both have sleep duration values of 6.7575, a strange value– which also equals the sample mean of the dataset. Therefore, I have decided to exclude these 2 rows in my analysis.

```
colnames(data) = make.names(colnames(data))
data = data %>%
  mutate(
    physical.illness = if_else(physical.illness=='no',0,1),
    bluelight.filter = ifelse(bluelight.filter=='no',0,1),
    smoke.drink = ifelse(smoke.drink=='no',0,1),
    meals.day = case_when(
      meals.day=='one'~1,
      meals.day=='two'~2,
      meals.day=='three'~3,
      meals.day=='four'~4,
      meals.day=='five'~5,
      meals.day=='more than 5'~6,
      .default=0
    ),
    screen.time = case_when(
      screen.time=="2hrs"~2,
      screen.time=="3-4 hrs"~3.5,
      screen.time=="1-2 hrs"~1.5,
      screen.time=="2-3 hrs"~2.5,
      screen.time=="4-5 hrs"~4.5,
      screen.time=="more than 5"~5.5,
      screen.time=="0-1 hrs"~0.5,
      .default=0
    )
  )
```

```
filter_data = function(df) {
  df %>% dplyr::filter(round(sleep.time,4) != 6.7575)
}
data_cleaned = filter_data(data)
data_cleaned
```

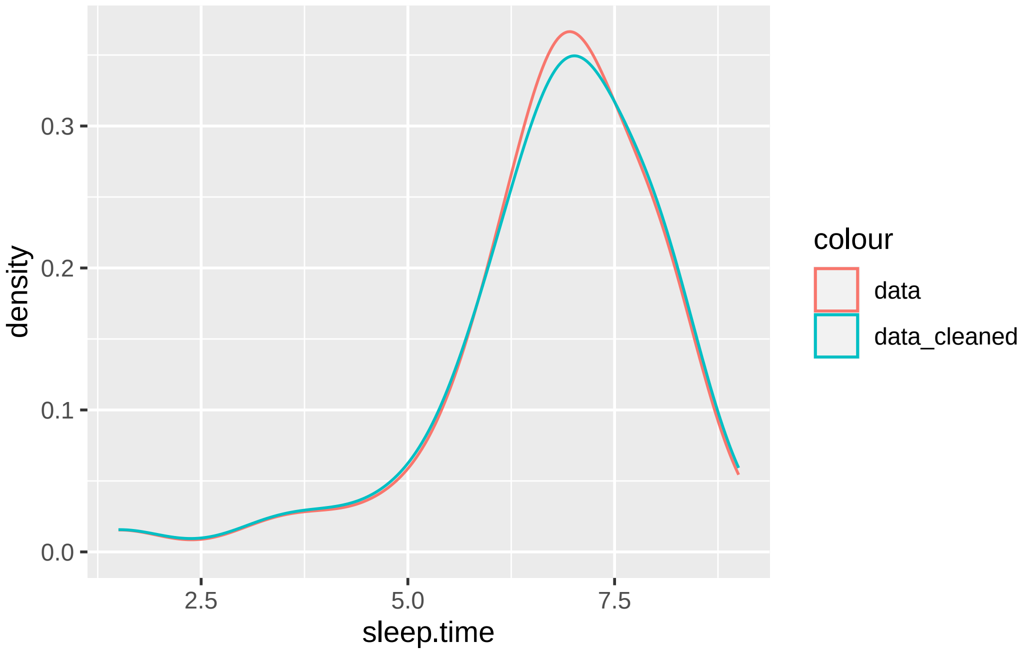
```
# A tibble: 44 x 11
     Age Gender meals.day physical.illness screen.time bluelight.filter
```

```
   <dbl> <chr>      <dbl>            <dbl>       <dbl>            <dbl>
 1    22 Female         3                0         3.5                0
 2    23 Male           3                0         3.5                0
 3    23 Female         2                0         1.5                0
 4    22 Male           3                0         5.5                1
 5    22 Male           4                0         1.5                1
 6    24 Female         3                1         4.5                0
 7    24 Male           4                0         2.5                1
 8    23 Female         3                0         5.5                1
 9    28 Female         3                0         0.5                1
10    59 Male           2                0         0.5                0
# i 34 more rows
# i 5 more variables: sleep.direction <chr>, exercise <chr>, smoke.drink <dbl>,
#   beverage <chr>, sleep.time <dbl>
```

```r
plot_density = function(df, group, is_filtered) {
  title = paste0(
    "Density by ",
    str_to_title(group)
  )
  subtitle = ifelse(is_filtered, "Removing Non-Rounded Values", "Full Data")
  return(
    df %>%
      ggplot() +
        geom_density(aes_string(x="sleep.time",color=group)) +
        labs(x="Sleep Duration", y="Density") +
        ggtitle(label=title, subtitle=subtitle)
  )
}

plot_and_test = function(df, group) {
  cleaned_df = filter_data(df)

  plots = list()
  plots$orig = plot_density(df, group, F)
  plots$clean = plot_density(cleaned_df, group, T)
  print(
    plots$orig +
      plots$clean +
      plot_layout(ncol = 2, nrow = 1)
  )
  formula_string = paste0("sleep.time~", group)
  formula = as.formula(formula_string)
  print("###Cleaned Data###")
  print(t.test(formula, data=cleaned_df))
  print(wilcox.test(formula, data=df))
  print("###Full Data###")
  print(t.test(formula, data=df))
  print(wilcox.test(formula, data=df))
}
```

Mean Sleep Duration

```
ggplot() +
  geom_density(data=data,aes(x=sleep.time, color="data")) +
  geom_density(data=data_cleaned,aes(x=sleep.time, color="data_cleaned"))
```



Cleaned

```r
# parametric
med=median(data_cleaned$sleep.time)
t.test(data_cleaned$sleep.time, mu=med)
```

```

    One Sample t-test

data:  data_cleaned$sleep.time
t = -1.1657, df = 43, p-value = 0.2502
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.337962 7.177038
sample estimates:
mean of x
   6.7575
```

```r
# non-parametric (sign test)
above = data_cleaned %>%
  mutate(sign = sleep.time > med) %>%
  filter(sign == 1) %>%
  nrow()
binom.test(above,data_cleaned %>% nrow(),0.5)
```

```

    Exact binomial test

data:  above and data_cleaned %>% nrow()
number of successes = 13, number of trials = 44, p-value = 0.00956
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1676440 0.4520218
sample estimates:
probability of success
             0.2954545
```

Non-cleaned

```
# parametric
med=median(data$sleep.time)
t.test(data_cleaned$sleep.time, mu=med)
```

```

    One Sample t-test

data:  data_cleaned$sleep.time
t = -1.1657, df = 43, p-value = 0.2502
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.337962 7.177038
sample estimates:
mean of x
   6.7575
```

```
# non-parametric (sign test)
above = data %>%
  mutate(sign = sleep.time > med) %>%
  filter(sign == 1) %>%
  nrow()
binom.test(above,data %>% nrow(),0.5)
```

```

    Exact binomial test

data:  above and data %>% nrow()
number of successes = 13, number of trials = 46, p-value = 0.004534
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1598667 0.4346041
sample estimates:
probability of success
             0.2826087
```

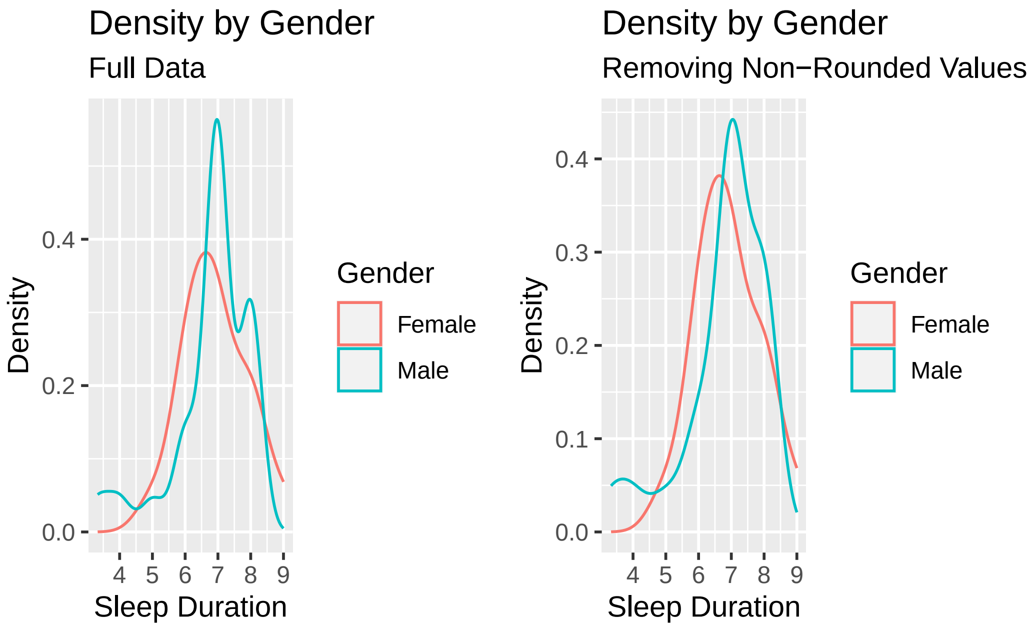
Compare male/female

```
data_by_gender = data %>%
  dplyr::filter(Gender %in% c("Male", "Female"))

plot_and_test(data_by_gender, "Gender")
```

```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.

[1] "###Cleaned Data###"

    Welch Two Sample t-test

data:  sleep.time by Gender
t = 0.32284, df = 39.828, p-value = 0.7485
alternative hypothesis: true difference in means between group Female and group Male is not
95 percent confidence interval:
 -0.585311  0.807811
sample estimates:
mean in group Female   mean in group Male 
            6.916667             6.805417 

Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
compute exact p-value with ties


    Wilcoxon rank sum test with continuity correction

data:  sleep.time by Gender
W = 222, p-value = 0.7762
alternative hypothesis: true location shift is not equal to 0

[1] "###Full Data###"

    Welch Two Sample t-test

data:  sleep.time by Gender
t = 0.34806, df = 40.53, p-value = 0.7296
```

```
alternative hypothesis: true difference in means between group Female and group Male is not
95 percent confidence interval:
 -0.5521898  0.7820616
sample estimates:
mean in group Female   mean in group Male
            6.916667             6.801731
```

```
Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
compute exact p-value with ties
```



```
    Wilcoxon rank sum test with continuity correction

data:  sleep.time by Gender
W = 222, p-value = 0.7762
alternative hypothesis: true location shift is not equal to 0
```

Direction of sleep

Data claims southwest is best. As we only have 4 directions, we will test for south/west vs north/east

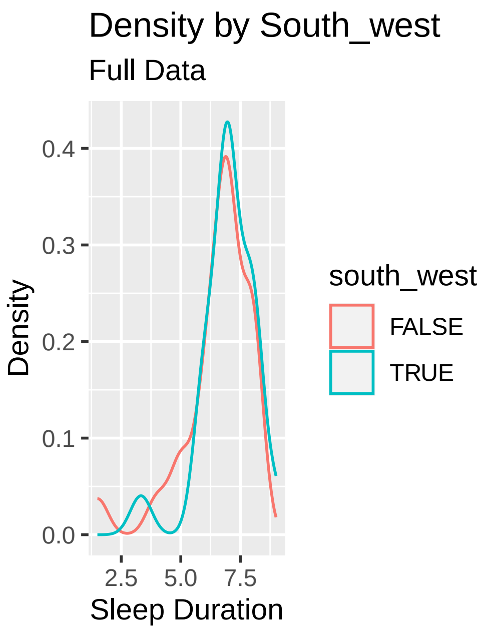
```
data_direction = data %>%
  mutate(
    south_west=sleep.direction %in% c("south","west")
  )
data_direction
```

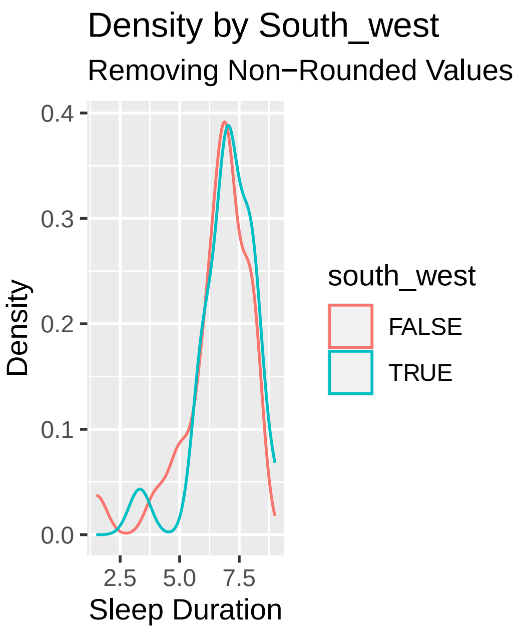
```
# A tibble: 46 x 12
     Age Gender meals.day physical.illness screen.time bluelight.filter
   <dbl> <chr>      <dbl>            <dbl>       <dbl>            <dbl>
 1    22 Male           2                0         2                  1
 2    22 Female         3                0         3.5                0
 3    23 Male           3                0         3.5                0
 4    23 Female         2                0         1.5                0
 5    22 Male           3                0         5.5                1
 6    22 Male           2                0         2.5                1
 7    22 Male           4                0         1.5                1
 8    24 Female         3                1         4.5                0
 9    24 Male           4                0         2.5                1
10    23 Female         3                0         5.5                1
# i 36 more rows
# i 6 more variables: sleep.direction <chr>, exercise <chr>, smoke.drink <dbl>,
#   beverage <chr>, sleep.time <dbl>, south_west <lgl>
```

```
plot_and_test(data_direction, "south_west")
```

```
[1] "###Cleaned Data###"

    Welch Two Sample t-test

data:  sleep.time by south_west
t = -1.164, df = 41.945, p-value = 0.251
alternative hypothesis: true difference in means between group FALSE and group TRUE is not e
95 percent confidence interval:
 -1.298084  0.348417
sample estimates:
mean in group FALSE  mean in group TRUE 
           6.541667            7.016500 

Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
compute exact p-value with ties
```

```
    Wilcoxon rank sum test with continuity correction

data:  sleep.time by south_west
W = 223, p-value = 0.3574
alternative hypothesis: true location shift is not equal to 0

[1] "###Full Data###"

    Welch Two Sample t-test

data:  sleep.time by south_west
t = -1.1506, df = 42.642, p-value = 0.2563
alternative hypothesis: true difference in means between group FALSE and group TRUE is not ec
95 percent confidence interval:
 -1.2424660  0.3398902
sample estimates:
mean in group FALSE  mean in group TRUE
           6.541667            6.992955

Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
compute exact p-value with ties
```





```
    Wilcoxon rank sum test with continuity correction

data:  sleep.time by south_west
W = 223, p-value = 0.3574
alternative hypothesis: true location shift is not equal to 0
```