STA 210: Lab 2

Jackson Hubbard

Spetember 10, 2018

Question 1

a

Holiday is whether day is holiday or not (extracted from http://dchr.dc.gov/page/ holiday-schedule)

b

Holiday is an int variable in the dataset. 1 means that it is a holiday and 0 means that it is not a holiday

```
glimpse(bikeshare)
```

```
## Observations: 731
## Variables: 16
## $ instant    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ dteday     <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 20...
## $ season     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ yr         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ mnth       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ holiday    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ...
## $ weekday    <int> 6, 0, 1, 2, 3, 4, 5, 6, 0, 1, 2, 3, 4, 5, 6, 0, 1, ...
## $ workingday <int> 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, ...
## $ weathersit <int> 2, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 1, 2, ...
## $ temp       <dbl> 0.3441670, 0.3634780, 0.1963640, 0.2000000, 0.22695...
## $ atemp      <dbl> 0.3636250, 0.3537390, 0.1894050, 0.2121220, 0.22927...
## $ hum        <dbl> 0.805833, 0.696087, 0.437273, 0.590435, 0.436957, 0...
## $ windspeed  <dbl> 0.1604460, 0.2485390, 0.2483090, 0.1602960, 0.18690...
## $ casual     <int> 331, 131, 120, 108, 82, 88, 148, 68, 54, 41, 43, 25...
## $ registered <int> 654, 670, 1229, 1454, 1518, 1518, 1362, 891, 768, 1...
## $ cnt        <int> 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1...
```

Question 2

```
# creates a new variable in the data set that stores holiday as a factor
# (categorical variable)
# case_when() function in the dplyr package allows us to write several if,
# else if, and else statements in a single statement


bikeshare <- bikeshare %>%
  mutate(holiday.new = as.factor(
    case_when (
    holiday==0 ~ "non-holiday",
    holiday==1 ~ "holiday",
)
)
)
```

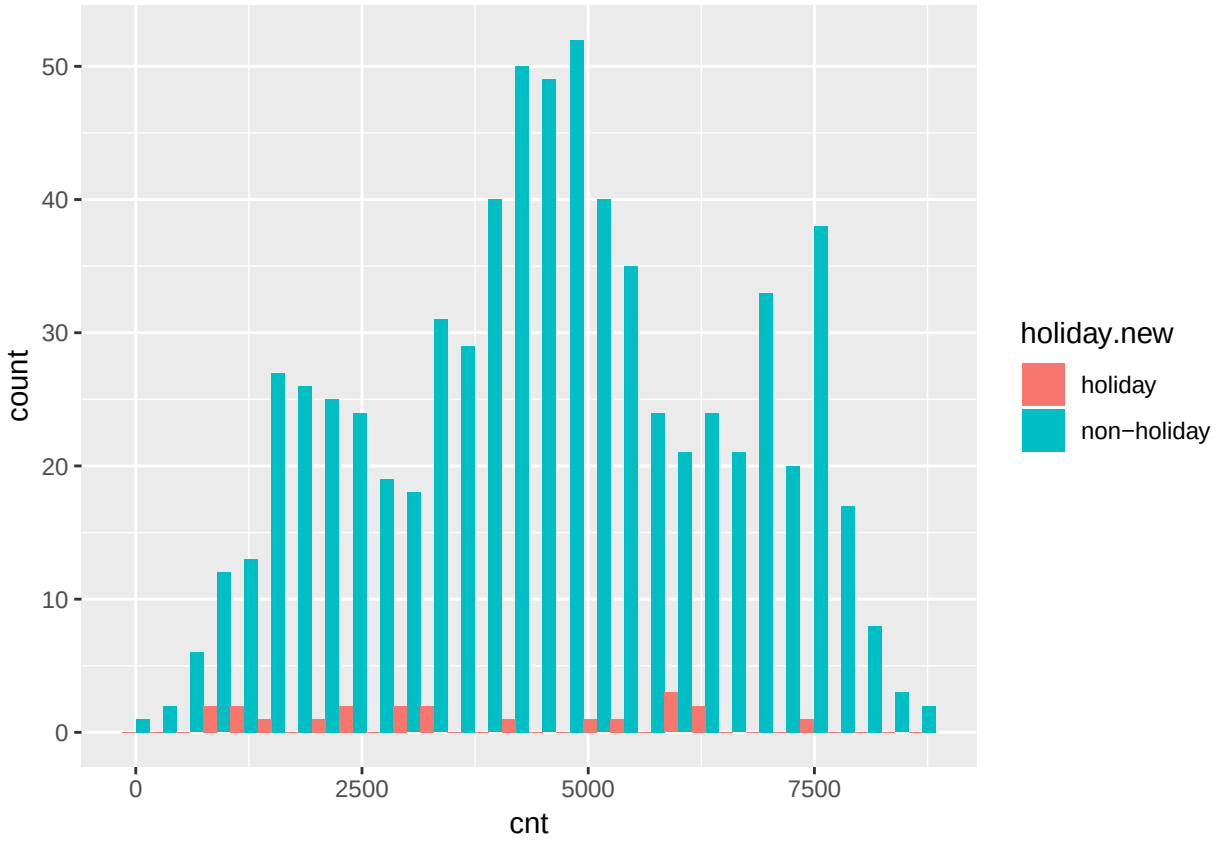
```
glimpse(bikeshare)
```

```
## Observations: 731
## Variables: 17
## $ instant     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
## $ dteday      <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 2...
## $ season      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ yr          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ mnth        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ holiday     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,...
## $ weekday     <int> 6, 0, 1, 2, 3, 4, 5, 6, 0, 1, 2, 3, 4, 5, 6, 0, 1,...
## $ workingday  <int> 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0,...
## $ weathersit  <int> 2, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 1, 2,...
## $ temp        <dbl> 0.3441670, 0.3634780, 0.1963640, 0.2000000, 0.2269...
## $ atemp       <dbl> 0.3636250, 0.3537390, 0.1894050, 0.2121220, 0.2292...
## $ hum         <dbl> 0.805833, 0.696087, 0.437273, 0.590435, 0.436957, ...
## $ windspeed   <dbl> 0.1604460, 0.2485390, 0.2483090, 0.1602960, 0.1869...
## $ casual      <int> 331, 131, 120, 108, 82, 88, 148, 68, 54, 41, 43, 2...
## $ registered  <int> 654, 670, 1229, 1454, 1518, 1518, 1362, 891, 768, ...
## $ cnt         <int> 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, ...
## $ holiday.new <fct> non-holiday, non-holiday, non-holiday, non-holiday...
```

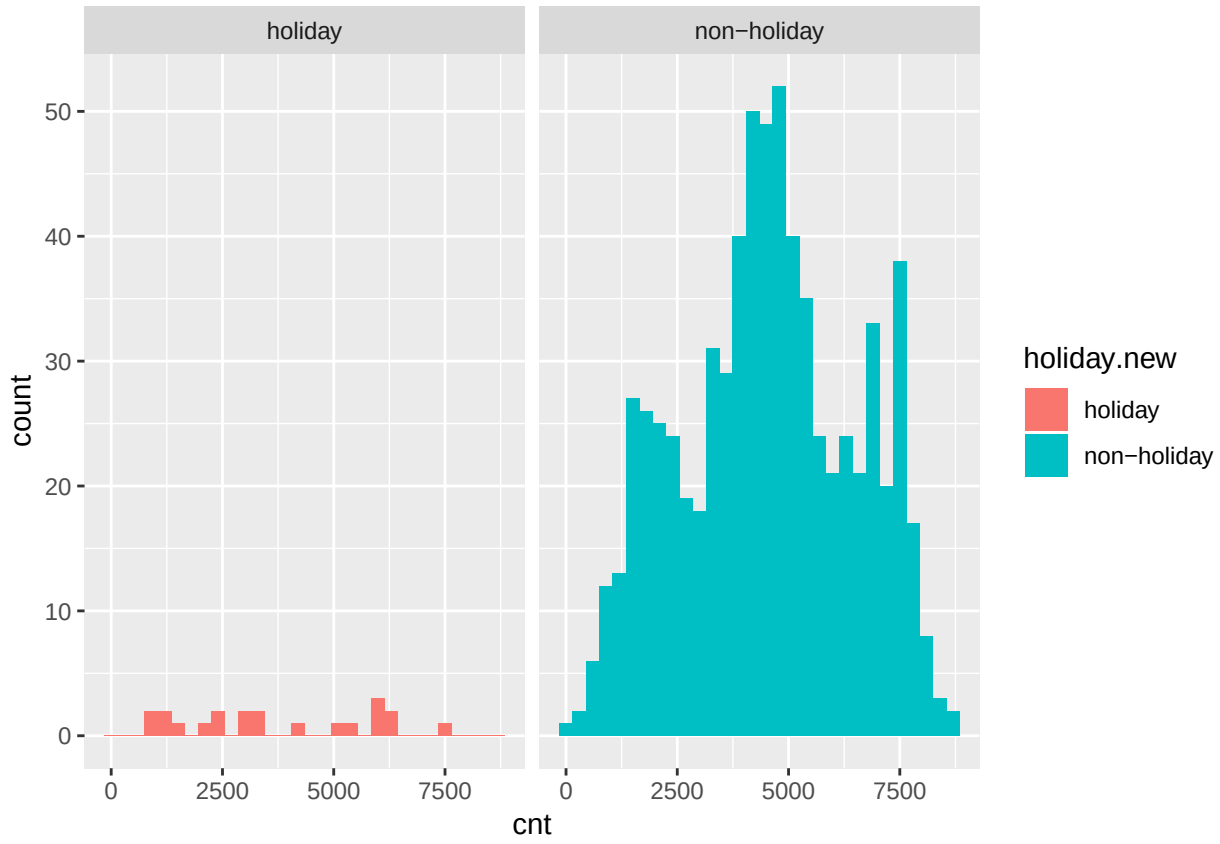
Question 3

a

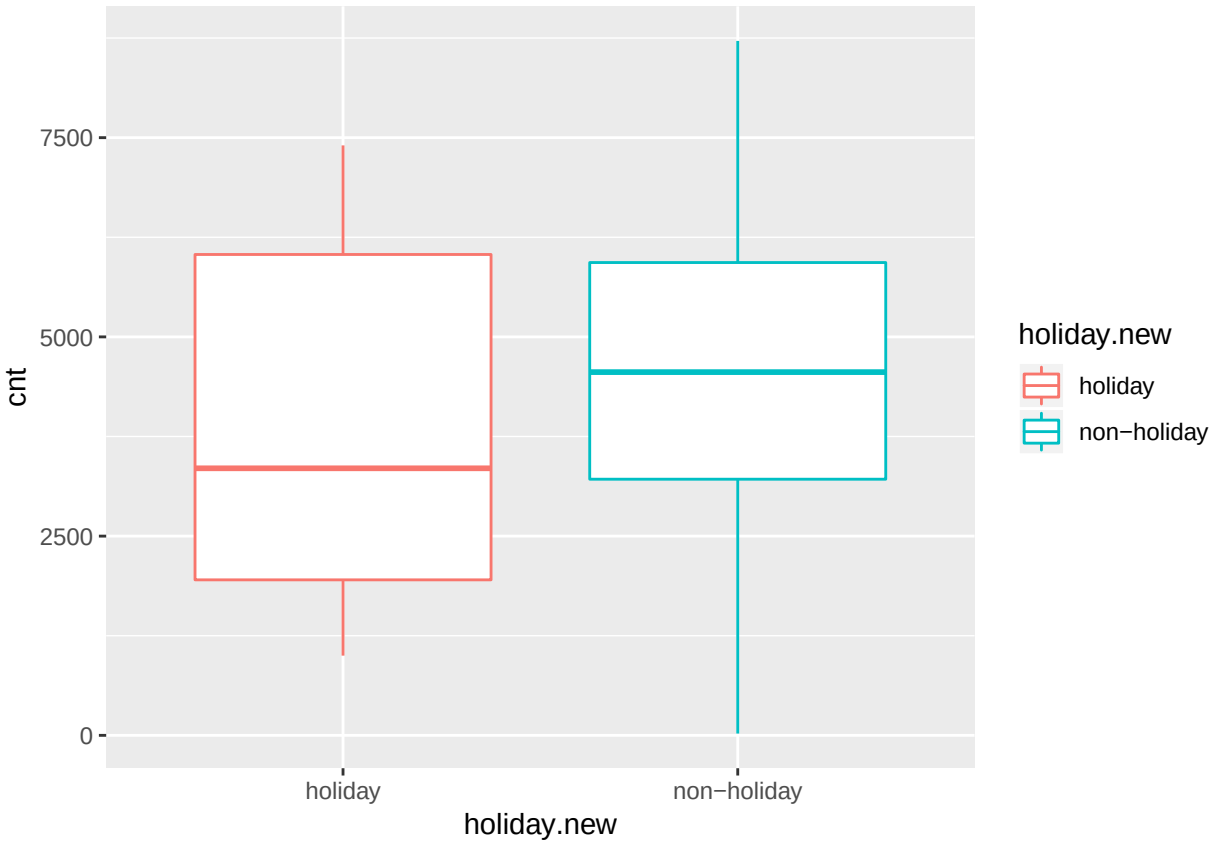
```
# Graph the distributions of cnt for holidays versus non-holidays.

ggplot(bikeshare, aes(cnt, fill= holiday.new)) + geom_histogram(position= "dodge")
```




```
ggplot(bikeshare, aes(cnt, fill= holiday.new)) + geom_histogram() +
  facet_wrap(~holiday.new)
```




```
ggplot(bikeshare, aes(x= holiday.new, y= cnt, colour= holiday.new)) + geom_boxplot()
```



b

```
# Calculate separate summary statistics to describe the distribution of
# cnt for holidays and non-holidays.


bikeshare %>% group_by(holiday.new) %>% summarise(n=n(), mean=mean(cnt),
                                                  median = median(cnt), s=sd(cnt))
```

```
## # A tibble: 2 x 5
##   holiday.new     n  mean median     s
##   <fct>       <int> <dbl>  <dbl> <dbl>
## 1 holiday        21 3735    3351 2103.
## 2 non-holiday   710 4527.   4558 1929.
```

c

For holidays, the mean count is much lower (3735 compared to 4527.1 for non-holidays). Further, there is a larger variation on holidays, as the sample standard deviation is larger for holidays (2103.351) than for non-holidays (1929.014). This can be seen in the histogram and

the boxplot as well. For hoidays, the sample size is very small and the distribution has high varinace, while for non-holidays it is distributed relativally normally.

Question 4

a

H0= no difference in mean number of bike rentals between holidays and non-holidays. HA= there is a difference in mean number of bike rentals between holidays and non-holidays

```
# To use the t.test() function, the data from the two samples must be stored
# separately. The code below creates a new object called holiday that includes
# only the days in the data set that were holidays.
holiday <- bikeshare %>% filter(holiday.new=="holiday")
```

```
# Create an object called non.holiday that includes only the days in the data
# set that were not holidays.
non.holiday <- bikeshare %>% filter(holiday.new=="non-holiday")
```

b

After doing a two sample hypothesis test on the difference of the average number of bike rentals between holidays and non-holidays, the p value of 0.06476 is higher than the threshold of 0.05. This means that there is not sufficient evidence to reject the null hypothesis of there being no difference in number of bike rentals between holidays and non-holidays.

```
# Conduct a two-sample t hypothesis test to determine whether there is a significant
# difference in the average number of bike rentals on days that are holidays vs.
# those that are not.
t.test(holiday$cnt, non.holiday$cnt, alternative="two.sided", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  holiday$cnt and non.holiday$cnt
## t = -1.8497, df = 729, p-value = 0.06476
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1632.81678    48.60833
## sample estimates:
## mean of x mean of y
##  3735.000  4527.104
```

c

After using a t test to get a 99% confidence interval on the difference of average number of bike rentals, we are 99% confident that the difference lies in the interval of (-1898.0486, 313.8402).

```
# Use the t.test() function to calculate a 99% confidence interval to estimate the
# mean difference in bike rentals on holidays vs. non-holidays.
t.test(holiday$cnt, non.holiday$cnt, alternative="two.sided", var.equal=TRUE,
       conf.level = .99)
```

```
##
##  Two Sample t-test
##
## data:  holiday$cnt and non.holiday$cnt
## t = -1.8497, df = 729, p-value = 0.06476
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -1898.0486   313.8402
## sample estimates:
## mean of x mean of y
##  3735.000  4527.104
```

d

The p-value is higher than the threshold of 0.05, so there is not a significant difference in the average number of bike rentals between holidays and non-holidays. Further, we are 99% confident that the difference lies in the interval of (-1898.0486, 313.8402). This interval contains 0, so we cannot say that there is a significant difference.

Question 5

a

H0= There is no difference in the mean number of bike rentals for holidays and non-holidays
HA= At least one mean number of bike rentals is different

```
# Conduct an ANOVA test to determine whether there is a significant
# difference between the average number of bike rentals on holidays and non-holidays.
anova <- aov(cnt ~ holiday.new,data=bikeshare)
tidy(anova) #prints results the ANOVA table in neat format
```

```
## # A tibble: 2 x 6
##   term           df       sumsq    meansq statistic p.value
##   <chr>       <dbl>       <dbl>     <dbl>     <dbl>   <dbl>
```

```
## 1 holiday.new     1   12797494. 12797494.      3.42  0.0648
## 2 Residuals     729 2726737898.  3740381.     NA    NA
```

b

Using ANOVA to determine whether there is a significant difference between the average number of bike rentals on holidays and non-holidays, we get a test statistic of 3.42. This results in a p-value of 0.06475, which is above the threshold of 0.05. This means that there is not sufficient evidence to reject that the mean number of bike rentals for holidays is different than the mean number of bike rentals for non-holidays.

c

Comparing the ANOVA test to the results of the hypothesis test in Question 4, we see that the p value is exactly the same. This makes sense as the ANOVA is just another way to check to see if there is a difference in two parameters. Both tests have high p values. For the hypothesis test, the conclusion is that there is not sufficient evidence to reject the null hypothesis of there being no difference in number of bike rentals between holidays and non-holidays. For ANOVA, the conclusion is that there is not sufficient evidence to reject that the mean number of bike rentals for holidays is different than the mean number of bike rentals for non-holidays. Thus, both tests are saying that there is no significant difference in the average number of bike rentals

Question 6

a

Since we are only comparing two means, the ANOVA test and the two-sample t hypothesis test are the exact same thing. However, if we introduced a third type of variable in addition to holiday and non.holiday and wanted to compare the means across 3+ groups, ANOVA would ahve to be used

b

No, this relationship would not hold because one of the assumptions of t-test is that the variances are equal. If they are not, you must transform the data, which means that comparing it to a two-sample t hypothesis test would be different. Below, you can see by changing var.equal to FALSE in the t.test the results are different.

```
t.test(holiday$cnt, non.holiday$cnt, alternative="two.sided", var.equal=FALSE,
       conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  holiday$cnt and non.holiday$cnt
## t = -1.7047, df = 21.007, p-value = 0.103
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1758.4038   174.1953
## sample estimates:
## mean of x mean of y
##  3735.000  4527.104
```

```
anova2 <- aov(cnt ~ holiday.new,data=bikeshare)
tidy(anova2)
```

```
## # A tibble: 2 x 6
##   term           df       sumsq    meansq statistic p.value
##   <chr>       <dbl>       <dbl>     <dbl>     <dbl>   <dbl>
## 1 holiday.new     1   12797494. 12797494.      3.42  0.0648
## 2 Residuals     729 2726737898.  3740381.     NA    NA
```