Tips at Restaurants

In this homework we will use the `tips` data set. This data set is part of the `reshape2` package. You can load the data set by executing the command:

```
data(tips, package="reshape2")
```

If you do not have available the package `reshape2`, issue `install.packages('reshape2')` to install it. The information contained in the data is collected by one waiter, who recorded over the course of a season information about each tip he received working in one restaurant. See `?tips` for a description of all of the variables.

Submission instructions: Create a folder named `ds202_hw3`, and name the RMarkDown file `hw3.Rmd` which should include your solutions under this folder. For submission, create a GitHub repository named `ds202_hw3` under your GitHub account, and push both `hw3.Rmd` and the knitted `hw3.html` before the deadline. I will assume you use the same GitHub username as for your HW2 submission. The latest version of your homework appearing on GitHub before the deadline is graded. *It is crucial to follow the exact folder structure and file names*, so that your homework can be reproduced and graded by an automated script. This homework is **due on Feb 19** before class.

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

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

```
names(tips)
```

```
## [1] "total_bill" "tip"        "sex"        "smoker"     "day"
## [6] "time"       "size"
```

```
head(tips)
```

```
##   total_bill  tip    sex smoker day   time size
## 1      16.99 1.01 Female     No Sun Dinner    2
## 2      10.34 1.66   Male     No Sun Dinner    3
## 3      21.01 3.50   Male     No Sun Dinner    3
## 4      23.68 3.31   Male     No Sun Dinner    2
## 5      24.59 3.61 Female     No Sun Dinner    4
## 6      25.29 4.71   Male     No Sun Dinner    4
```

1. How many parties did the waiter serve? Store the number in `numParty` and print.

```
numParty <- length(tips)
print(numParty)
```

```
## [1] 7
```

2. What are the types of variables in the data frame `tips`? Include the code and also explain verbally.

```
str(tips)
```

```
## 'data.frame':    244 obs. of  7 variables:
```

```
##  $ total_bill: num  17 10.3 21 23.7 24.6 ...
##  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
##  $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...
##  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

3. Create a vector named `day` containing the day of visits in `tips`. The factor levels should be ordered from Thursday to Sunday. Print the variable.

```
day <- factor(tips$day)
print(day)
```

```
##   [1] Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun
##  [16] Sun  Sun  Sun  Sun  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat
##  [31] Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sun  Sun  Sun  Sun  Sun
##  [46] Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sat  Sat  Sat  Sat  Sat
##  [61] Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat
##  [76] Sat  Sat  Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur
##  [91] Fri  Fri  Fri  Fri  Fri  Fri  Fri  Fri  Fri  Fri  Fri  Fri  Sat  Sat  Sat  Sat
## [106] Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sun  Sun  Sun  Sun  Sun  Thur Thur Thur Thur
## [121] Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur
## [136] Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur
## [151] Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun
## [166] Sun  Sun  Sun  Sat  Sat  Sat  Sat  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun
## [181] Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Sun  Thur Thur Thur Thur Thur
## [196] Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Thur Sat  Sat  Sat  Sat  Sat
## [211] Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Fri  Fri  Fri  Fri  Fri
## [226] Fri  Fri  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat  Sat
## [241] Sat  Sat  Sat  Thur
## Levels: Fri Sat Sun Thur
```

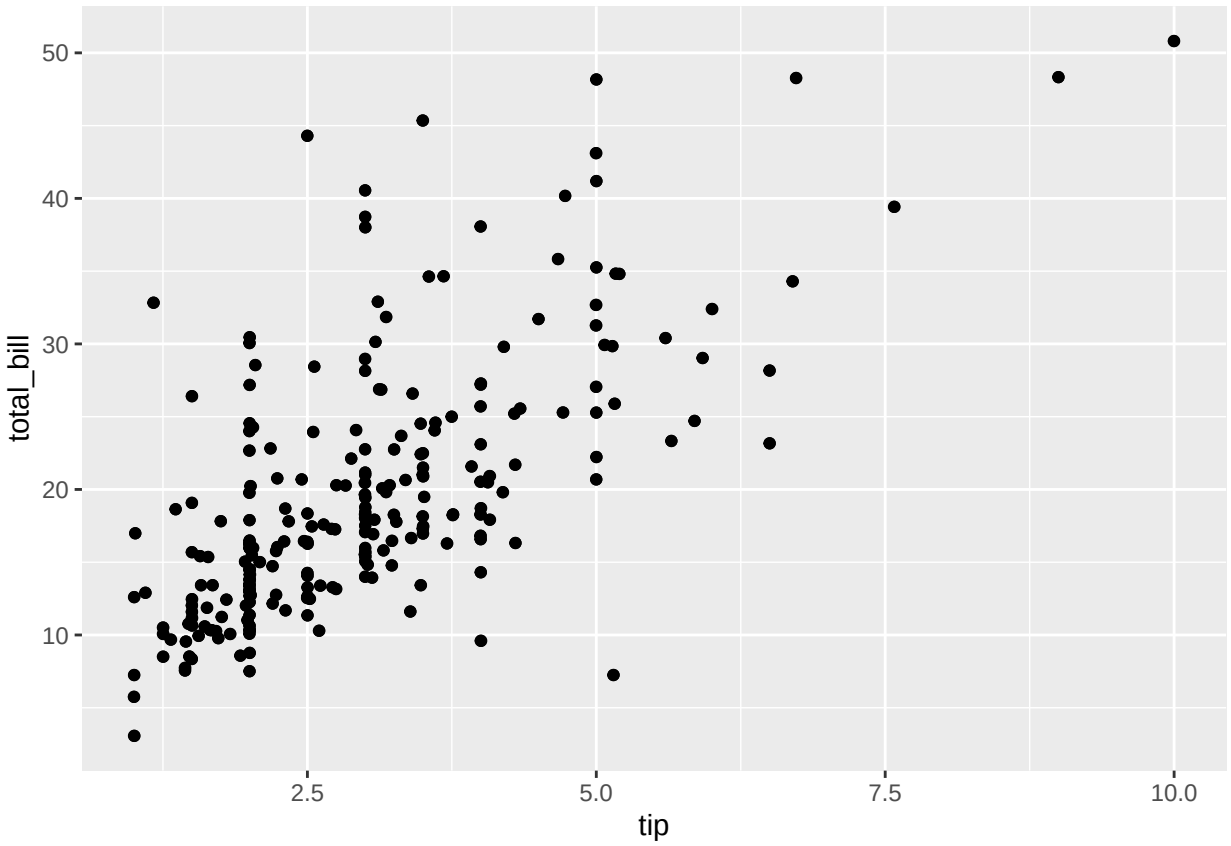
4. Create a data frame named `female5` containing the meal paid by a female payer in a party with size greater than or euqal to 5. Print the data frame.

```
female5 <- tips[tips$sex == 'Female' & tips$size == 5, ]
print(female5)
```

```
##     total_bill  tip    sex smoker day   time size
## 156      29.85 5.14 Female     No Sun Dinner    5
```

5. How does the tipping amount (`tip`) depend on the overall bill (`total_bill`)? Use the `ggplot2` package to make a chart. Describe the relationship in words.

```
ggplot(tips, aes(tip, total_bill)) + geom_point() + geom_jitter()
```



The relationship appears to be positively weak. There is little if any correlation between the tip left by a customer and the total bill.

6. Describe at least two types of anomalies in the previous plot. What do they mean?

There appears to be a couple points which appear to overshoot the typical trend, these points indictate that a larger bill typically does not influence the tip amount. This could imply the fact that an expensive meal is already an expensive meal, there no need to keep on stacking surplus charges for the customer.

7. Introduce a variable `tiprate` into the data set which stands for the rate of tips. What is the average rate for tips? Explain verbally.

```
tips$tiprate = tips$tip / (tips$total_bill - tips$tip)
mean(tips$tiprate)
```

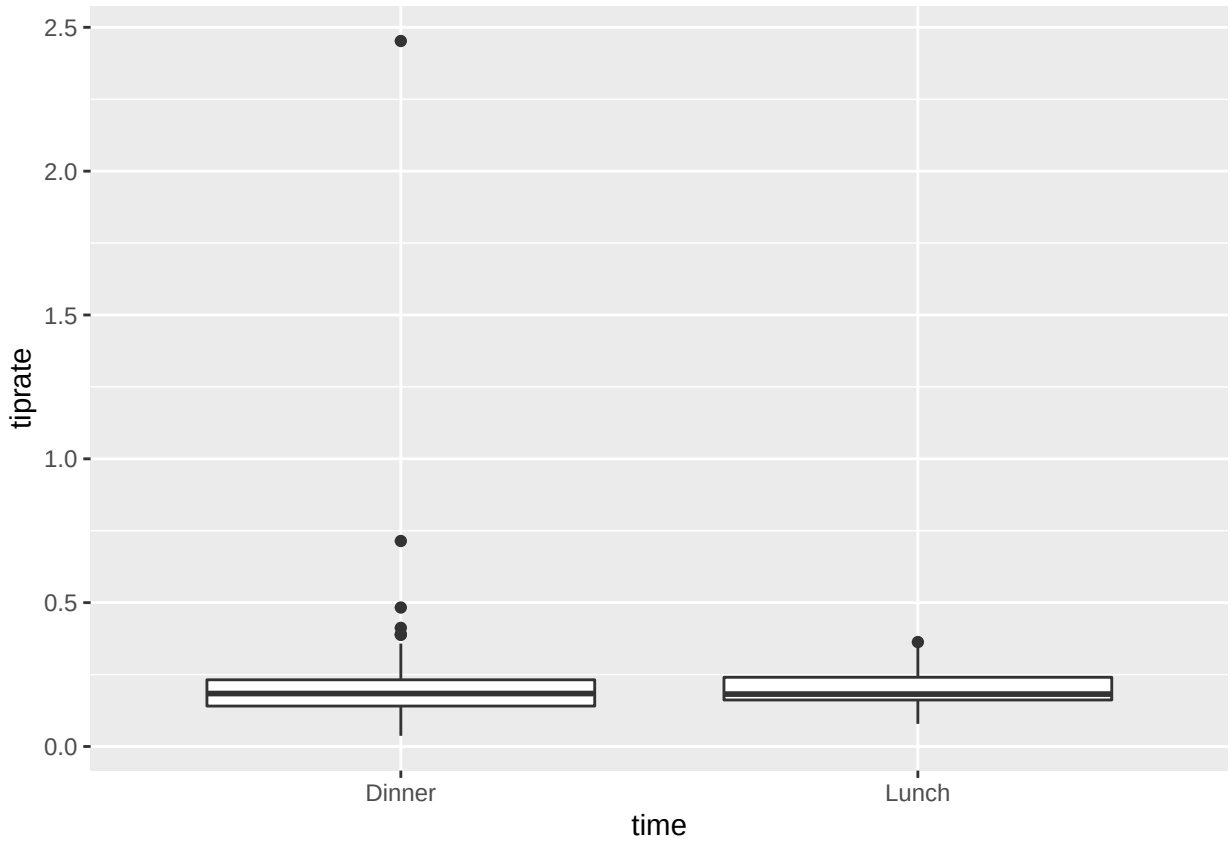
```
## [1] 0.2021235
```

On an Weekend or Thursday, you could expect to be tipped 20.2% on average per meal.

8. Make a boxplot of the tip rate by time. The x-axis should be ordered by lunch and then dinner. Use `ggplot2` to make a chart. Verbally explain the chart.

Raw results (all observations)

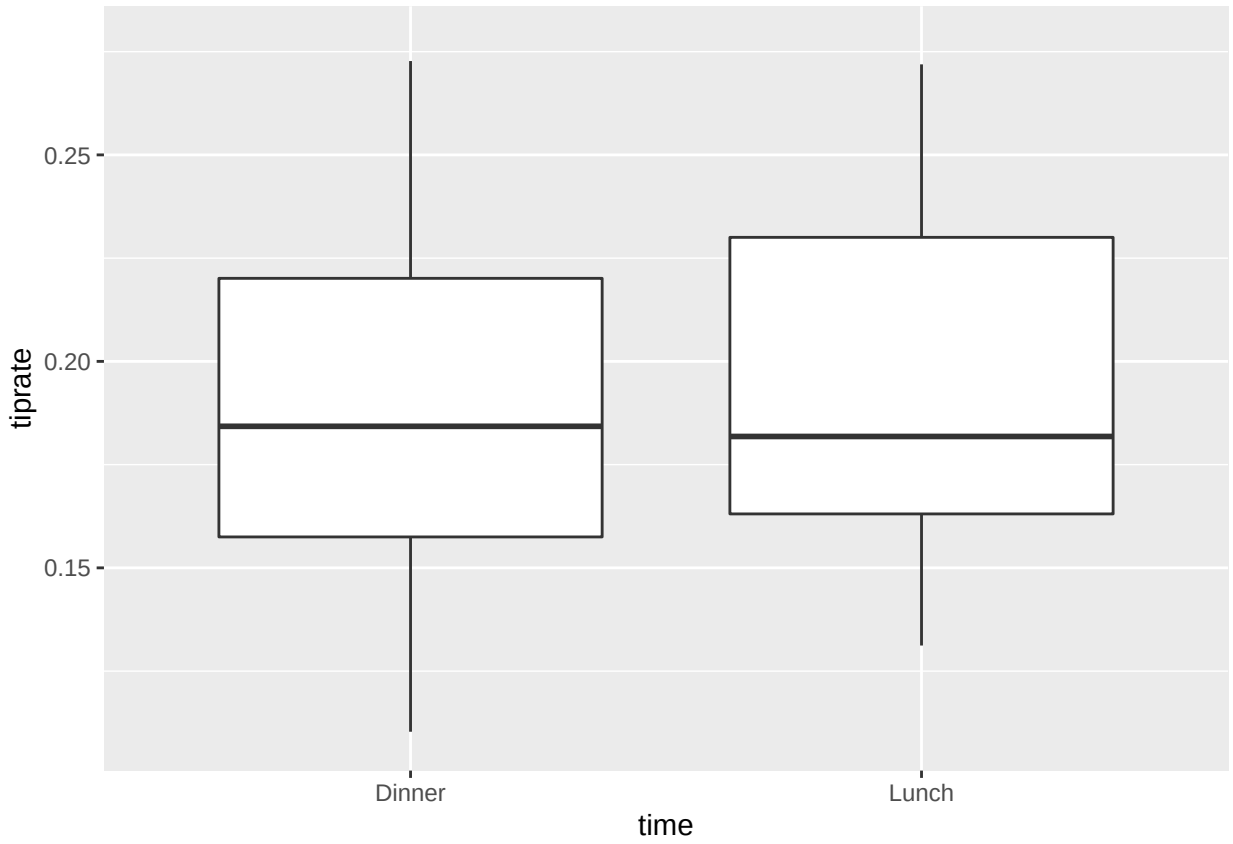
```
tips %>%
  ggplot(aes(time, tiprate)) + geom_boxplot()
```



Normalized results (anomalies omitted)

```
tips %>%
  ggplot(aes(time, tiprate)) + geom_boxplot() + scale_y_continuous(limits = quantile(tips$tiprate, c(0
```

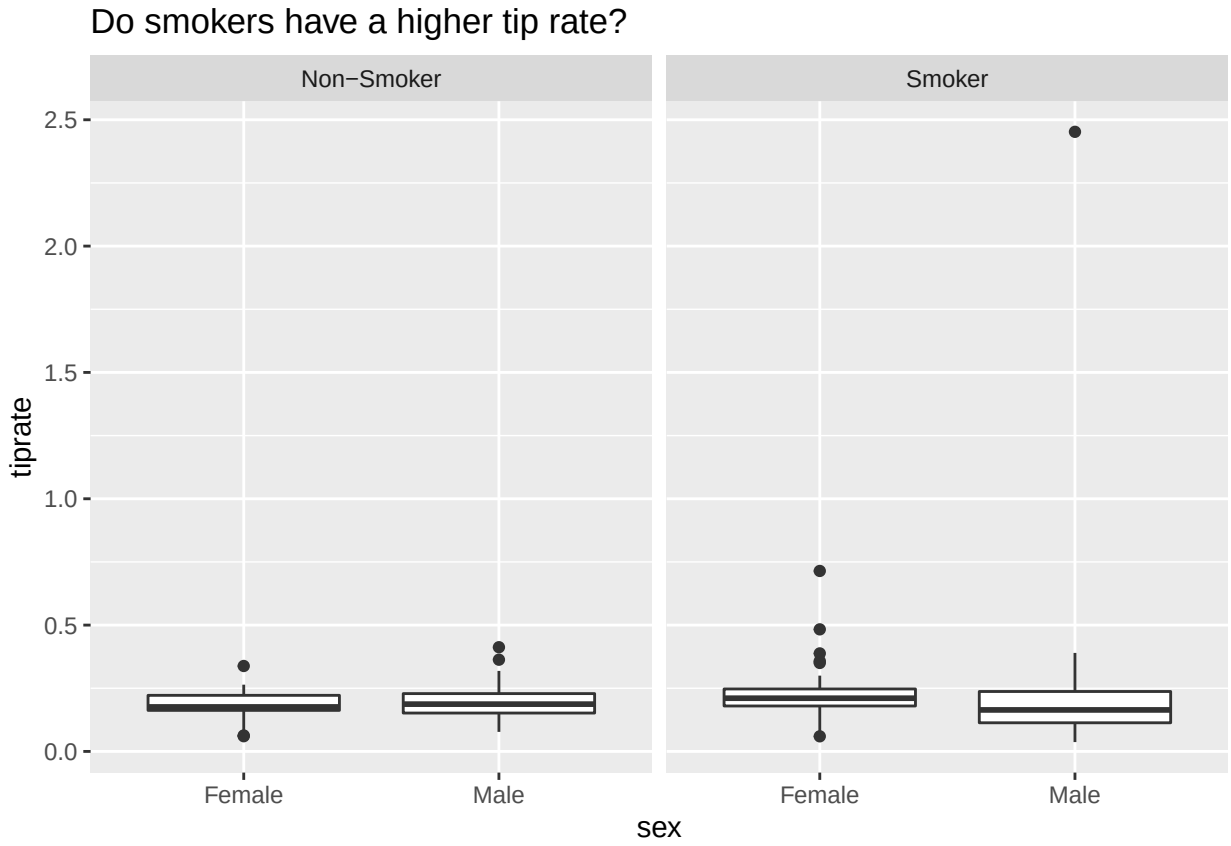
```
## Warning: Removed 50 rows containing non-finite values (stat_boxplot).
```



9. How does smoking behavior and gender of the person who pays impact the relationship between tip and total bill? Find a visualization that incorporates all four variables. Interpret the result.

Raw results (all observations)

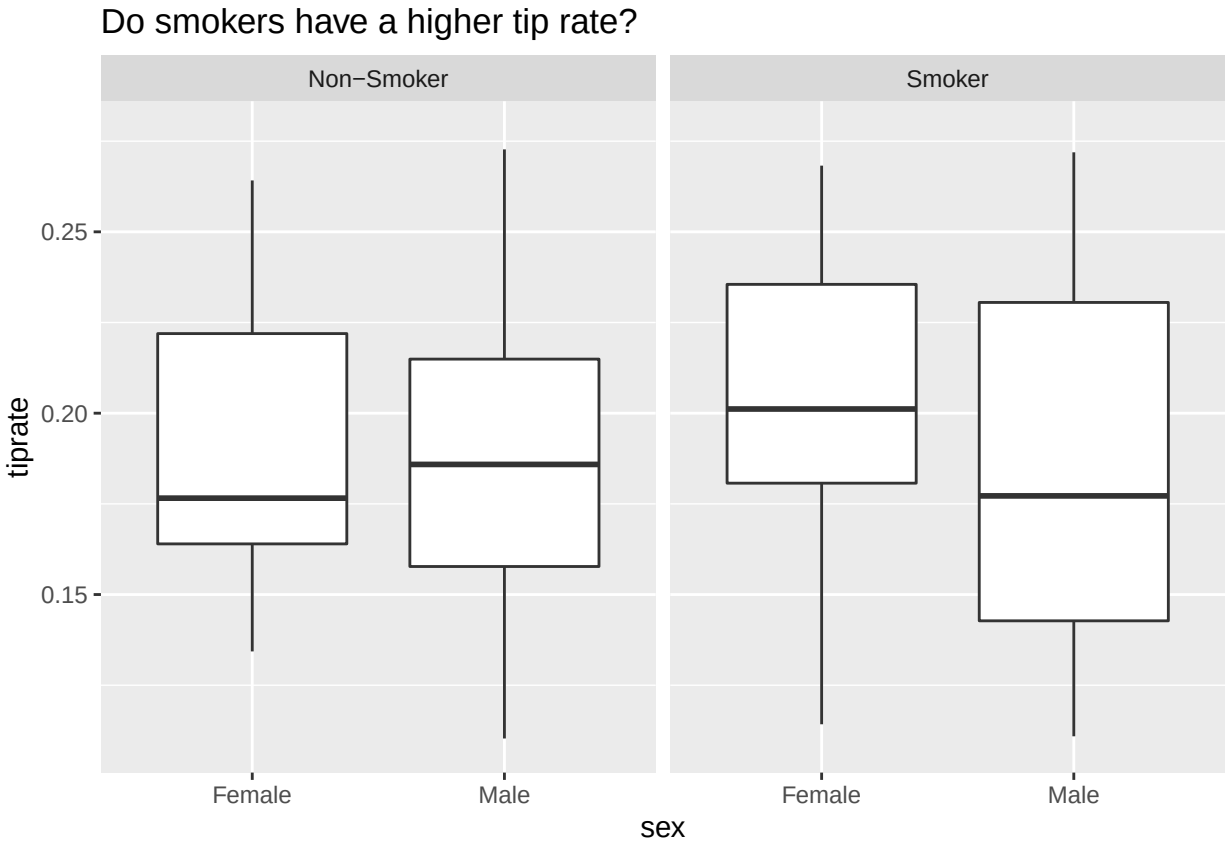
```
tips$smokerLabel <- ifelse(tips$smoker == "Yes", "Smoker","Non-Smoker")

tips %>%
  ggplot(aes(sex, tiprate)) + geom_boxplot() + facet_grid(cols = vars(smokerLabel)) + ggtitle('Do smo
```



Normalized results (anomalies omitted)

```
  tips %>%
    ggplot(aes(sex, tiprate)) + geom_boxplot() + facet_grid(cols = vars(smokerLabel)) + ggtitle('Do smo
```

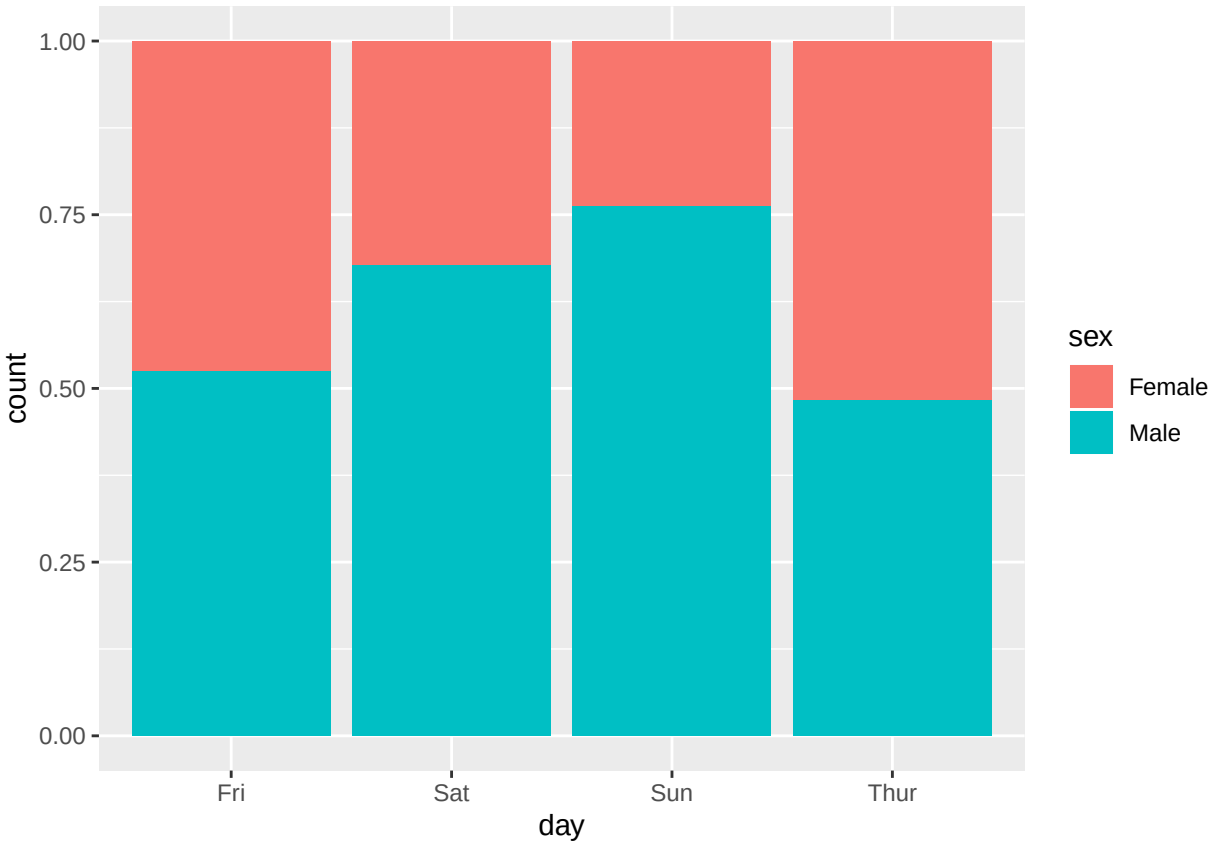
```
## Warning: Removed 50 rows containing non-finite values (stat_boxplot).
```



From the normalized results, it's clear that smoker females tend to overall have a higher tiprate for the weekend and thursday. On the non-smoking side, males typically tend to tip more. However, on the topic of standard deviation, the inter-quartile range of smoker males appears to be much more prominent, this could imply that smoker males are more influenced by many more signals than just smoking compared to smoker females.

10. Use ggplot2 to find a graphical summary of the relationship between day of the week and gender of the person paying the bill. What can you say about this relationship?

```
tips %>%
  ggplot(aes(x = day, fill = sex)) + geom_bar(position = 'fill')
```



It appears for the most part, males appear to pay the most during weekends, on weekdays like Thursday males appear to be less likely to cover a check. Though, this data may be misleading as we don't take into account if the party is fully male or partially male and partially female. This ambiguity has implications such that we only know which party-hosting gender is most frequently visiting the restaurant on four days.

Note: your submission is supposed to be fully reproducible, i.e. the TA and I will 'knit' your submission in RStudio.