Introduction to tidyverse

What's tidyverse?

Tidyverse is a collection of R packages designed for data science. They all share common grammar and data structures.

The core packages are:

- **ggplot2**
- **dplyr**
- tidyr
- readr
- purrr
- tibble
- **stringr**
- forcats

Installing and Loading tidyverse

First install tidyverse: `install.packages("tidyverse")`

Then load tidyverse: `library(tidyverse)`

Loading the Data

First install the Data Package: `install.packages("titanic")`

Then load the Data: `library(titanic)`

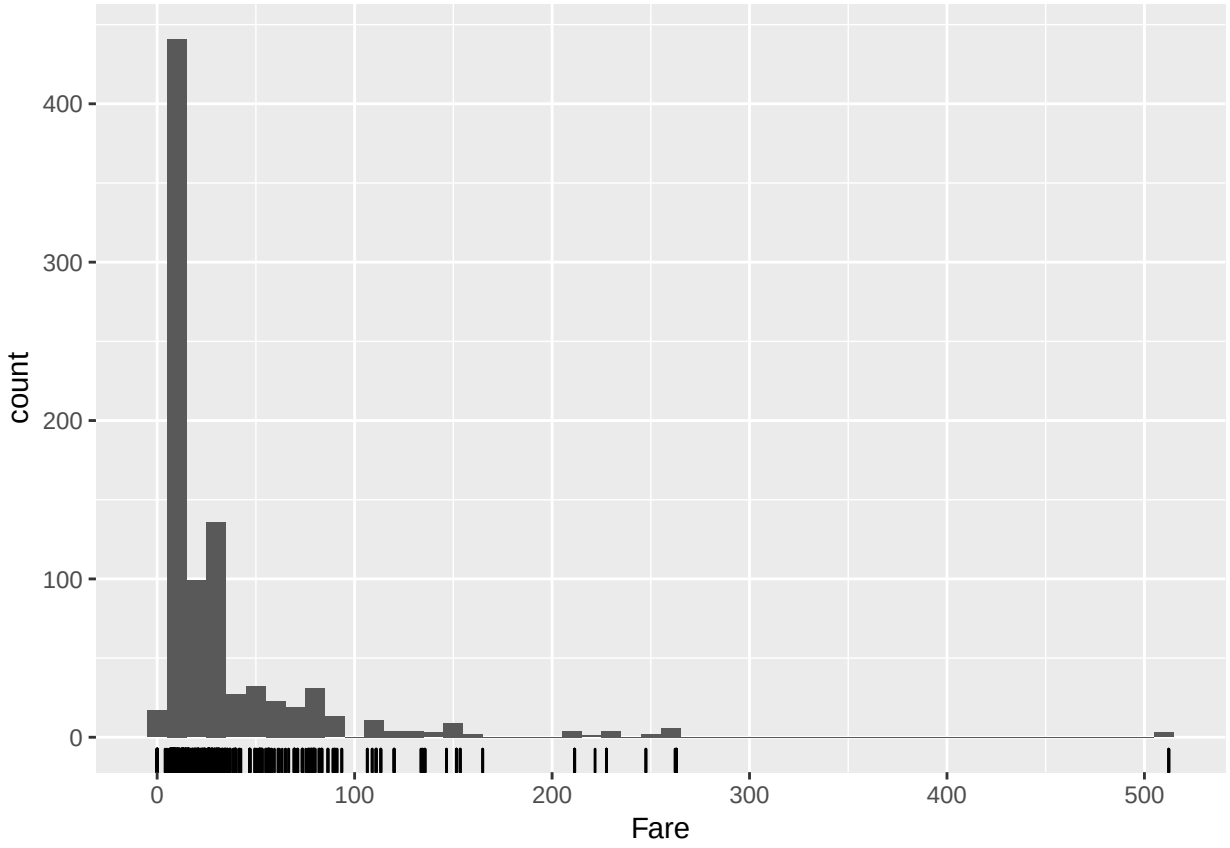
Inspecting the Data

```
str(titanic_train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

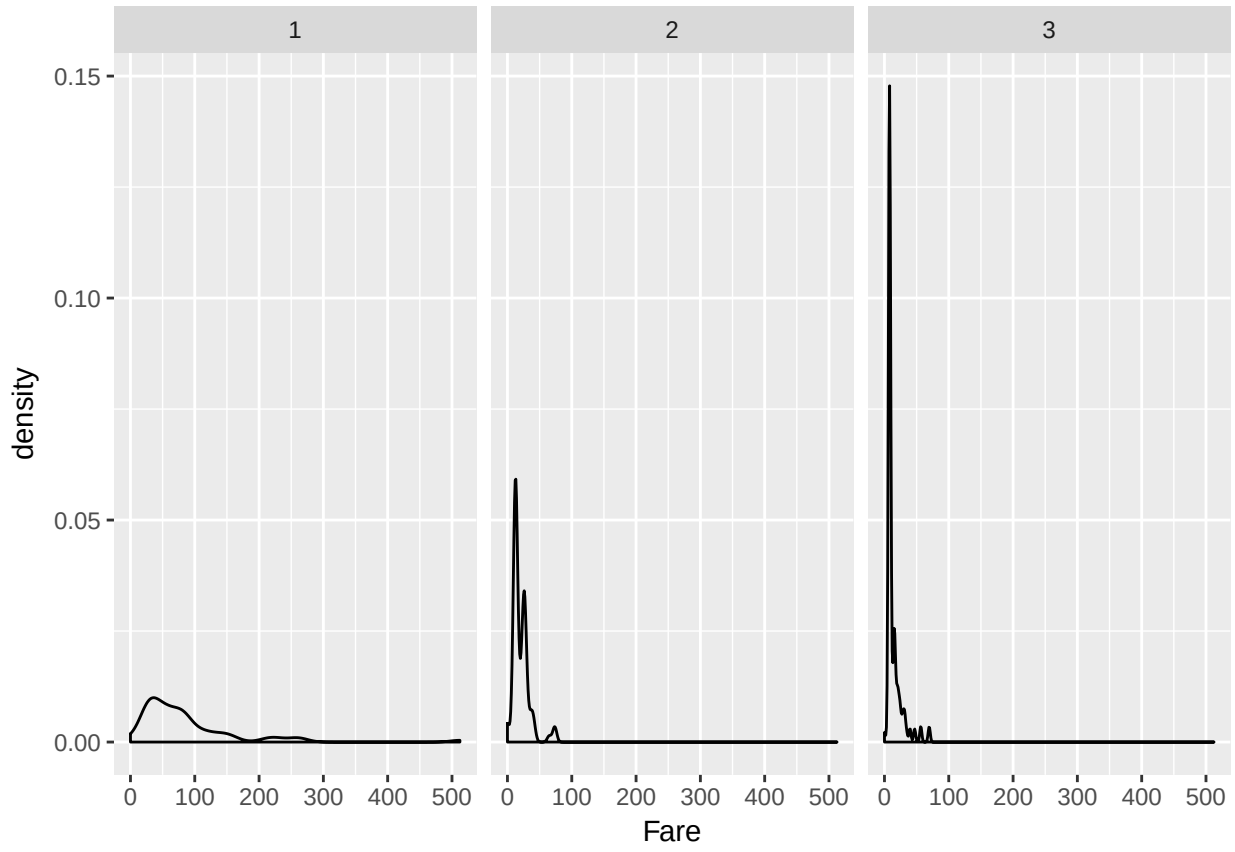
A Question You Might Ask: How much did people pay?

```
ggplot(titanic_train, aes(Fare)) + geom_histogram(binwidth = 10) + geom_rug()
```



What about for the different classes?

```
ggplot(titanic_train, aes(Fare)) + geom_density() + facet_wrap(~ Pclass)
```



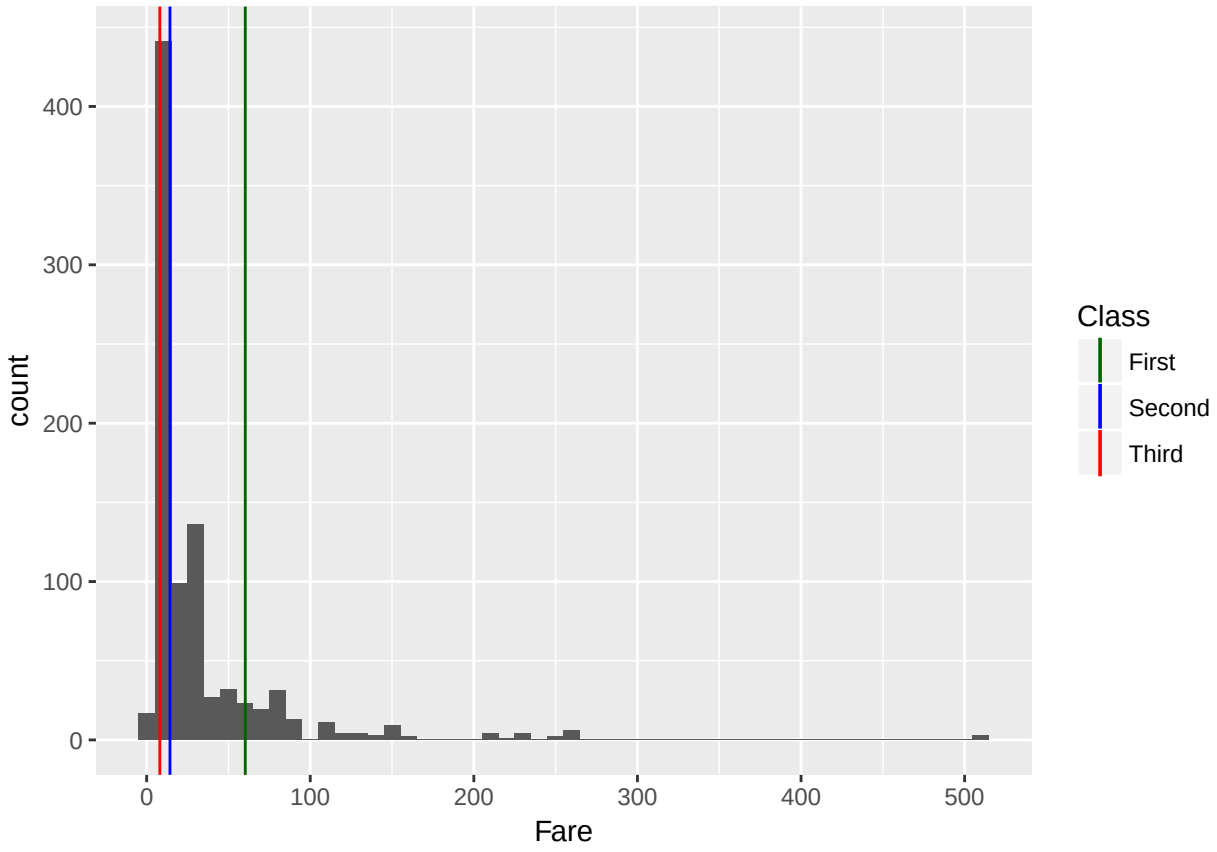
Let's go back and see how the medians compare

```
first_class_mean = median(select(filter(titanic_train, Pclass == 1), Fare)[[1]])
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

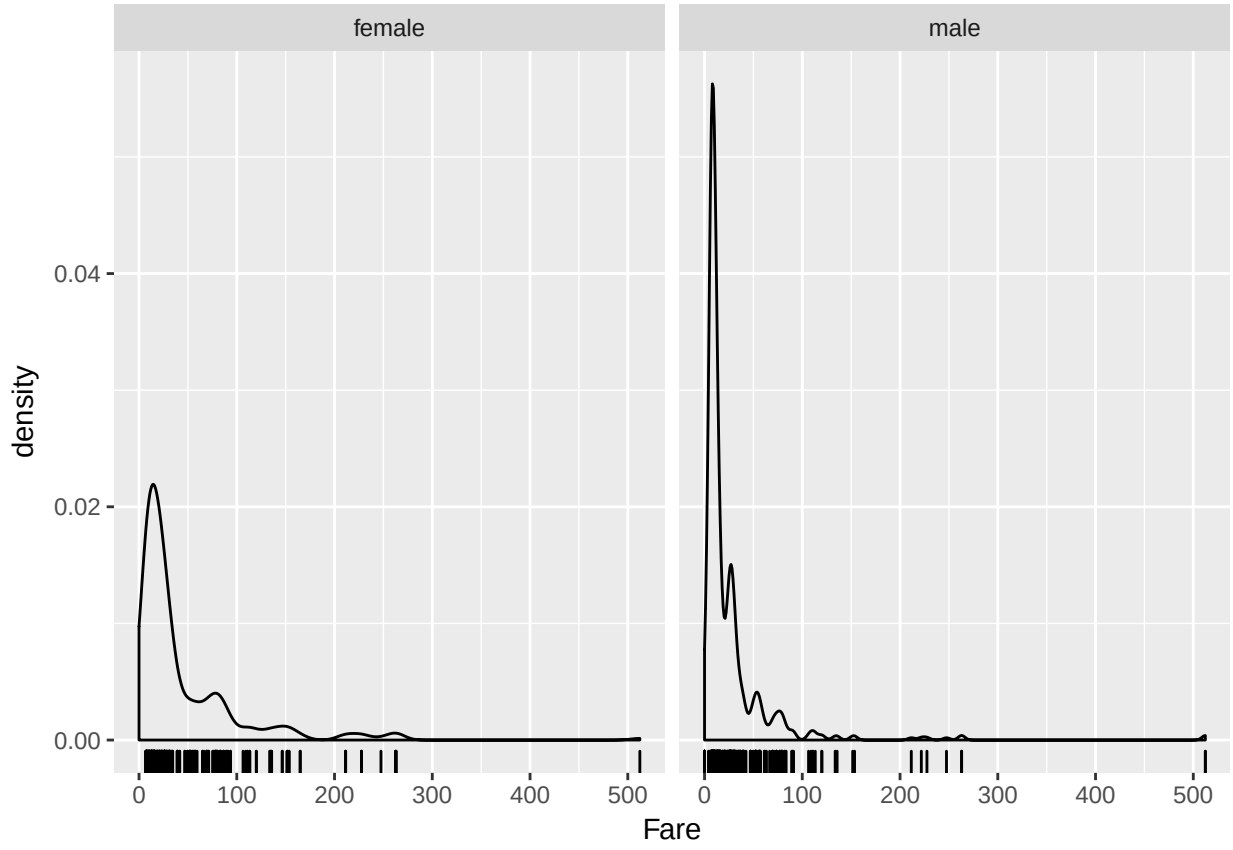
```
second_class_mean = median(select(filter(titanic_train, Pclass == 2), Fare)[[1]])
third_class_mean = median(select(filter(titanic_train, Pclass == 3), Fare)[[1]])

ggplot(titanic_train, aes(Fare)) + geom_histogram(binwidth = 10) + geom_vline(aes(xintercept = first_cla
scale_color_manual(name="Class", values = c("First" = "darkgreen", "Second" = "blue", "Third" = "red"))
```



YOUR TURN: Create a density plot of `Fare` with a rug plot for each `Sex`

```
ggplot(titanic_train, aes(Fare)) + geom_density() + facet_wrap(~ Sex) + geom_rug()
```



Another Question You Might Ask: Is there a relationship between a passenger's title and their survival rate?

```
names = select(titanic_train, "Name")[[1]]
mr = str_detect(names, "Mr.")
mrs = str_detect(names, "Mrs.")
miss = str_detect(names, "Miss.")
master = str_detect(names, "Master")
leftovers = filter(titanic_train, !mr & !mrs & !miss & !master)$Name
leftovers
```

```
##  [1] "Uruchurtu, Don. Manuel E"
##  [2] "Byles, Rev. Thomas Roussel Davids"
##  [3] "Bateman, Rev. Robert James"
##  [4] "Minahan, Dr. William Edward"
##  [5] "Carter, Rev. Ernest Courtenay"
##  [6] "Moraweck, Dr. Ernest"
##  [7] "Aubart, Mme. Leontine Pauline"
##  [8] "Pain, Dr. Alfred"
##  [9] "Reynaldo, Ms. Encarnacion"
## [10] "Peuchen, Major. Arthur Godfrey"
## [11] "Butt, Major. Archibald Willingham"
## [12] "Kirkland, Rev. Charles Leonard"
## [13] "Stahelin-Maeglin, Dr. Max"
## [14] "Sagesser, Mlle. Emma"
```

```
## [15] "Simonius-Blumer, Col. Oberst Alfons"
## [16] "Frauenthal, Dr. Henry William"
## [17] "Weir, Col. John"
## [18] "Crosby, Capt. Edward Gifford"
## [19] "Rothes, the Countess. of (Lucy Noel Martha Dyer-Edwards)"
## [20] "Brewe, Dr. Arthur Jackson"
## [21] "Leader, Dr. Alice (Farnham)"
## [22] "Reuchlin, Jonkheer. John George"
## [23] "Harper, Rev. John"
## [24] "Montvila, Rev. Juozas"
```

```
mr_survival = sum(select(filter(titanic_train, mr), "Survived")[[1]])/length(select(filter(titanic_trai
mrs_survival = sum(select(filter(titanic_train, mrs), "Survived")[[1]])/length(select(filter(titanic_tra
miss_survival = sum(select(filter(titanic_train, miss), "Survived")[[1]])/length(select(filter(titanic_
master_survival = sum(select(filter(titanic_train, master), "Survived")[[1]])/length(select(filter(titar
```

```
ggplot(data.frame("Title" = c("Mr.", "Mrs.", "Miss", "Master"), "prop" = c(mr_survival, mrs_survival, mi
```