HDDA Tutorial: Getting Started with R : Solutions

Department of Econometrics and Business Statistics, Monash University

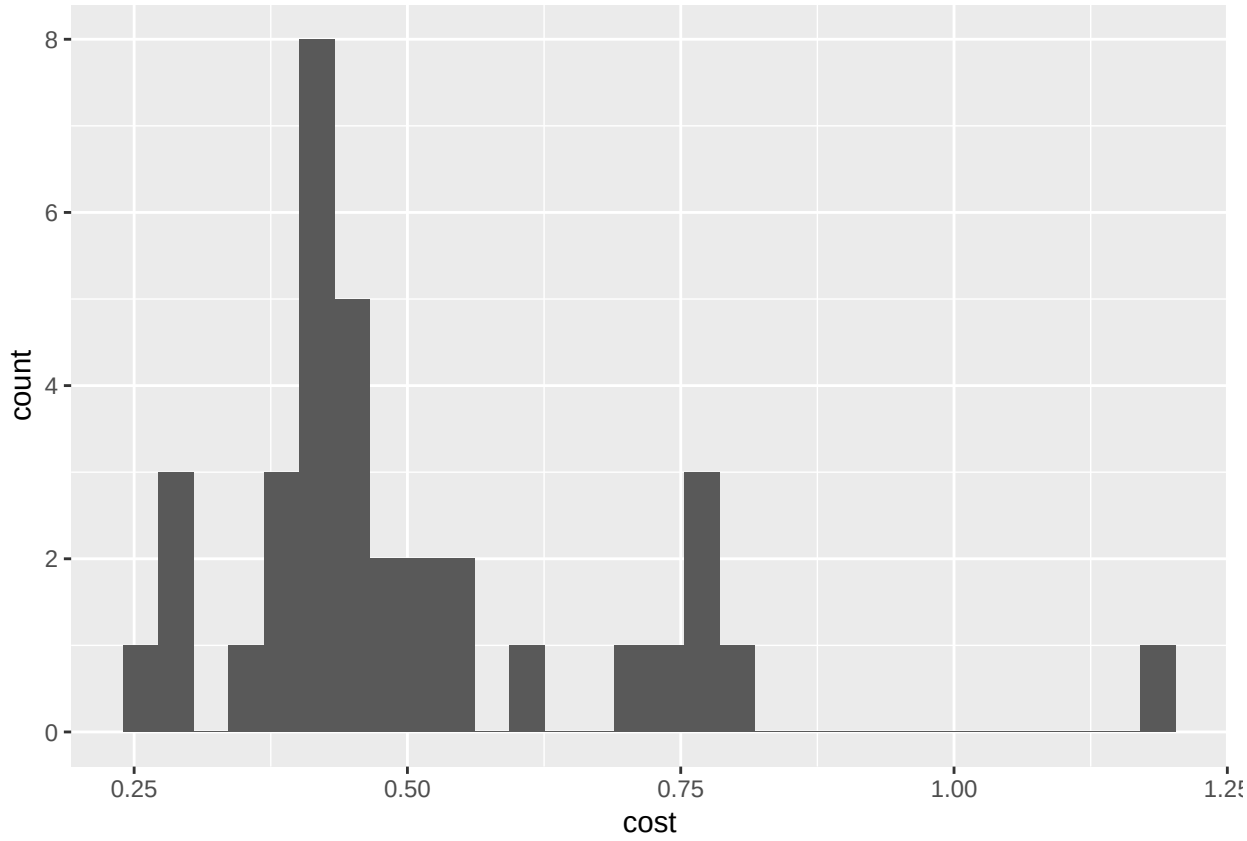
Tutorial 2

The aim of this week's tutorial is to do more preliminary data analysis using R. You will need to use the datasets *Beer.rds* and *comScore.rds* from which can be downloaded from Moodle.

Beer Data

1. Without using the `qplot` function, produce a histogram of the cost per 12 fl. oz. variable.

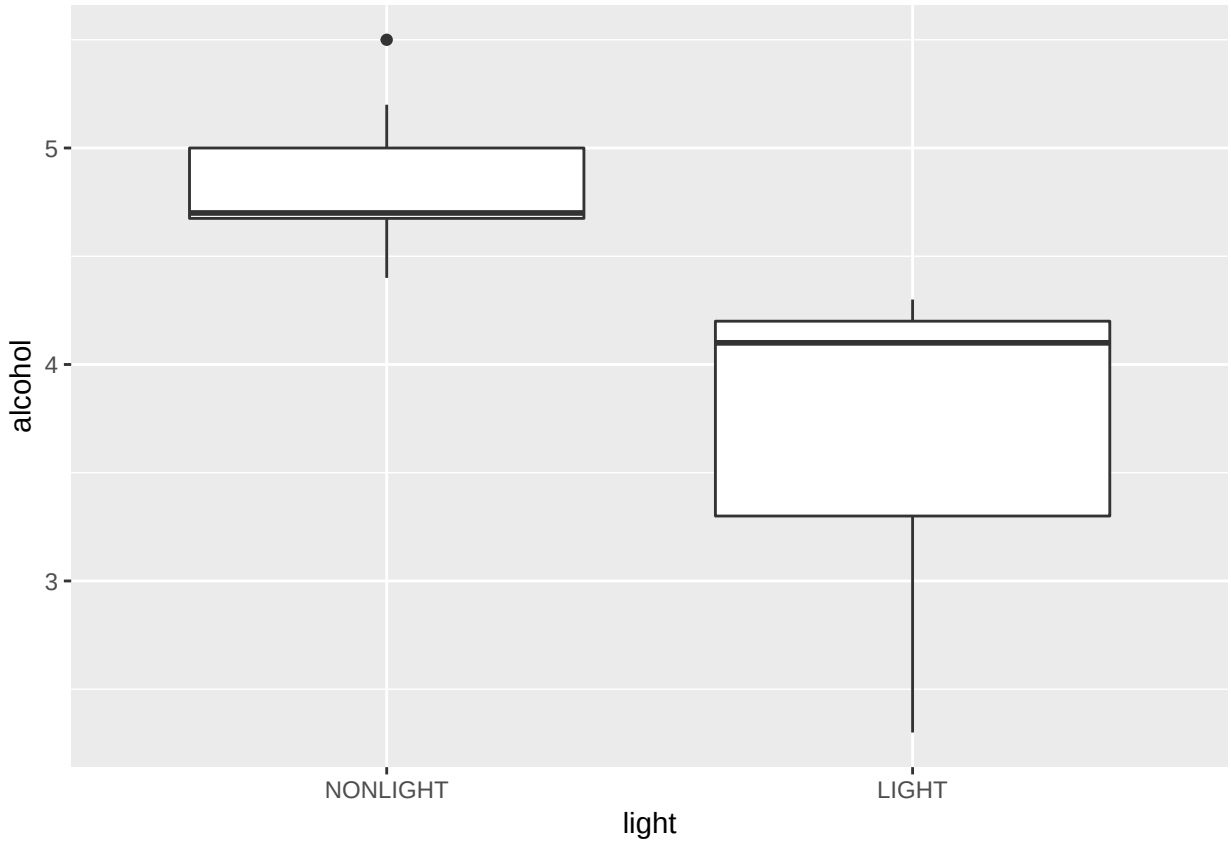
```
#First load required packages
library(dplyr)
library(ggplot2)
#Then load in data
Beer<-readRDS('Beer.rds')
#Histogram of cost
ggplot(Beer,aes(x=cost))+geom_histogram()
```




```
# In the histogram notice the presence of the outlier.
```

2. Without using the `qplot` function, produce boxplots of alcohol content. On the same plot there should be a separate boxplot for light beers and a separate boxplot for nonlight beers.

```r
#Boxplots
ggplot(Beer,aes(x=light,y=alcohol))+geom_boxplot()
```




```r
# Unsuprisingly the light beers have less alcohol content than the
#non-light beers.
```

3. Produce a frequency table of beer rating

```r
#Boxplots
table(Beer$rating)
```

```
##
## VeryGood     Good     Fair
##       11       14       10
```

```r
# There are roughly equal numbers of fair good and very good beers.
```

4. Produce a cross tab of beer rating against light/nonlight

```r
#Boxplots
table(Beer$rating,Beer$light)
```

```
##
##            NONLIGHT LIGHT
##   VeryGood       11     0
##   Good           10     4
##   Fair            7     3
```

```r
# Notice that zero light beers are rated very good.  This suggests
#a relationship between light/non-light and the rating of beer.
```

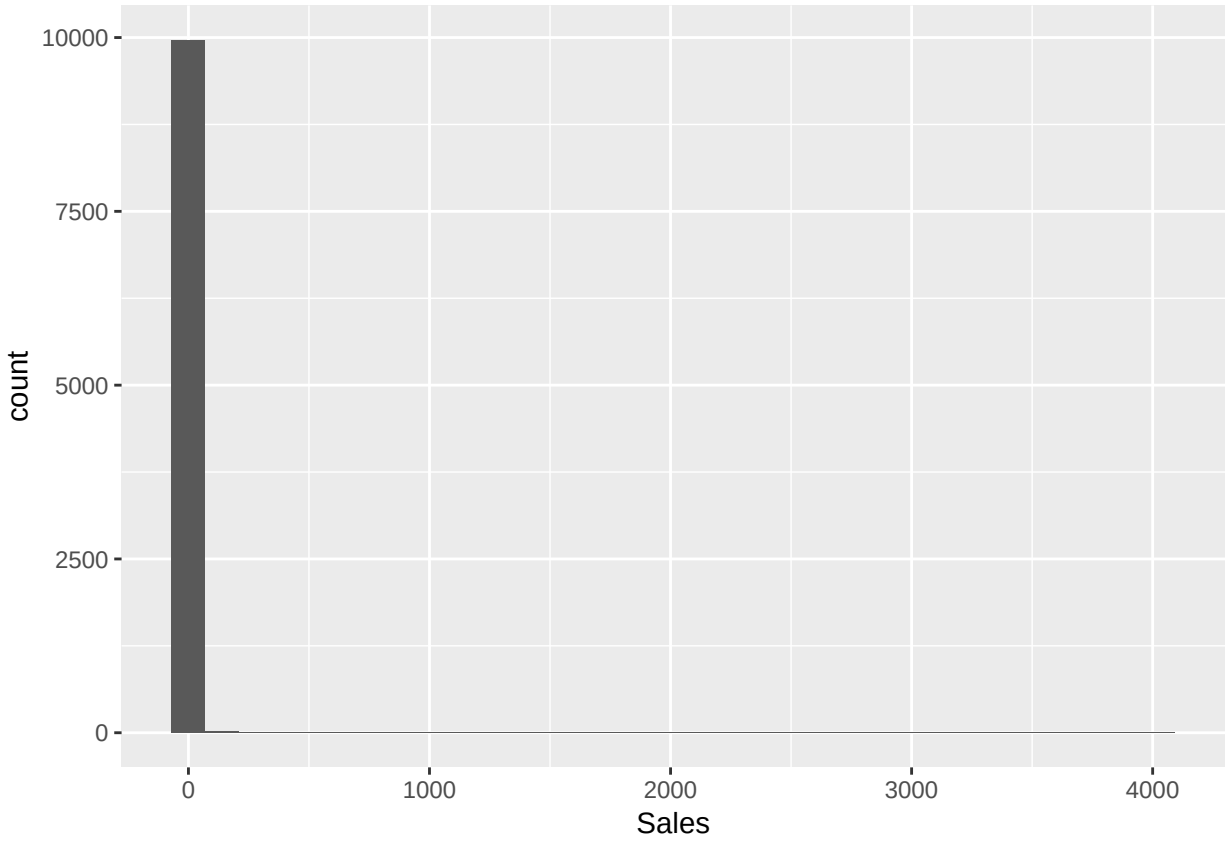
comScore Data

The company comScore records the online behaviour of subscribers. Each observation of the dataset that you have been provided with is a unique visit to the wesite apple.com. Four variables are recorded: **Buy** indicates whether a purchase was made, **Sales** indicates the value of any purchase, **Duration** indicates how much time was spent on apple.com, while **PageViews** indicates how many pages were clicked on under the apple.com domain name in a single visit. An interesting marketing question is whether browsing behaviour (duration and page views) are associated with purchase behaviour (buy and sales).

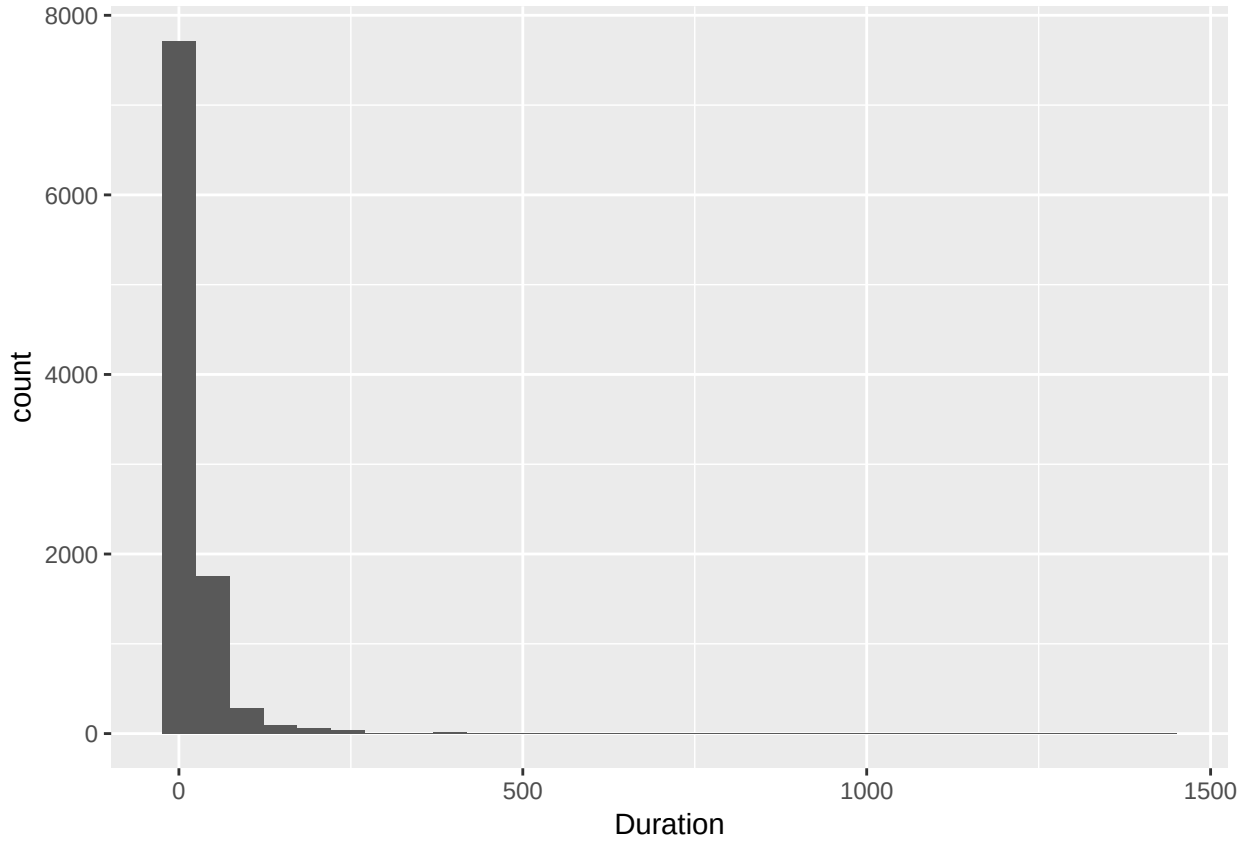
1. Without using the `qplot` function, produce histograms of

- Sales
- Duration
- Page Views

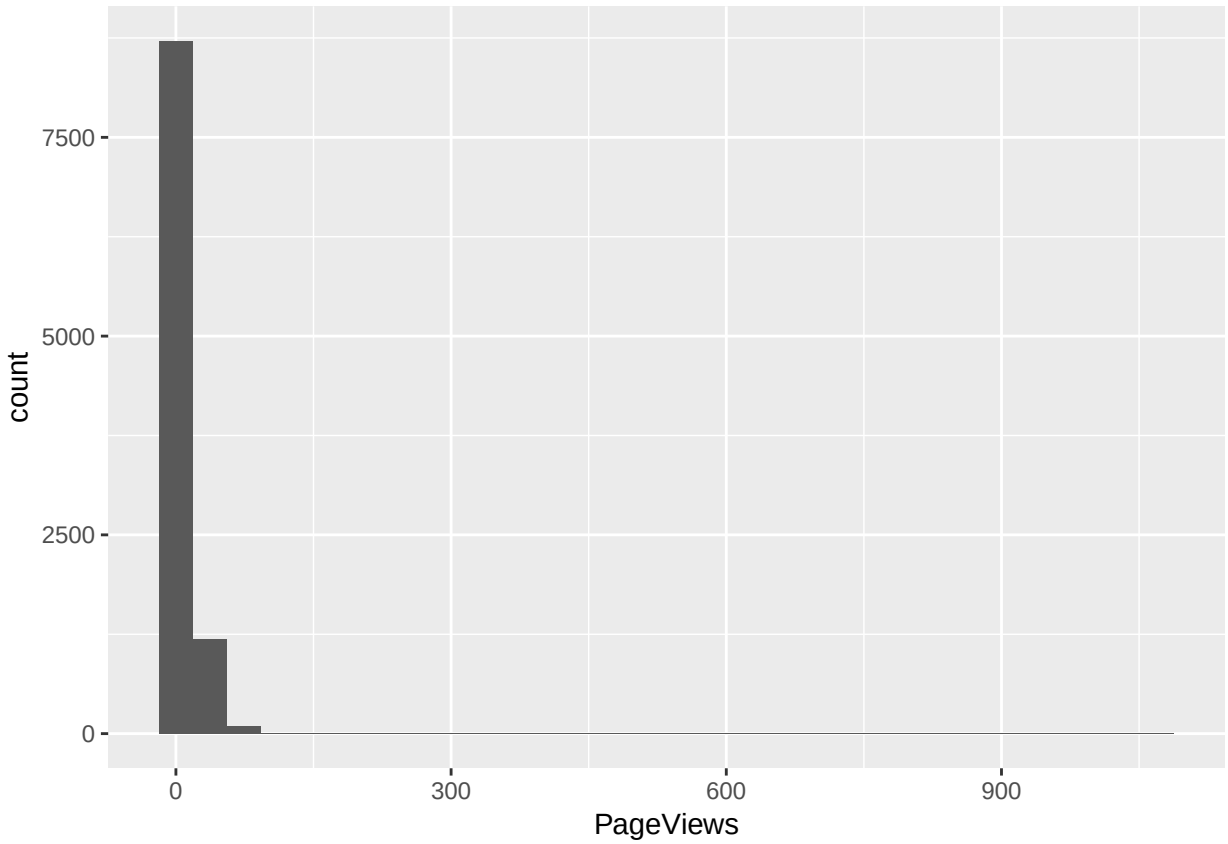
```
#Then load in data
comScore<-readRDS('comScore.rds')
#Histogram of sales
ggplot(comScore,aes(x=Sales))+geom_histogram()
```




```
#Histogram of duration
ggplot(comScore,aes(x=Duration))+geom_histogram()
```




```
#Histogram of page views
ggplot(comScore,aes(x=PageViews))+geom_histogram()
```




```
#The resulting histograms are not particularly informative since
#low values are so prevalent and since the data is very highly
#skewed.
```

2. Produce summary statistics for the comScore data

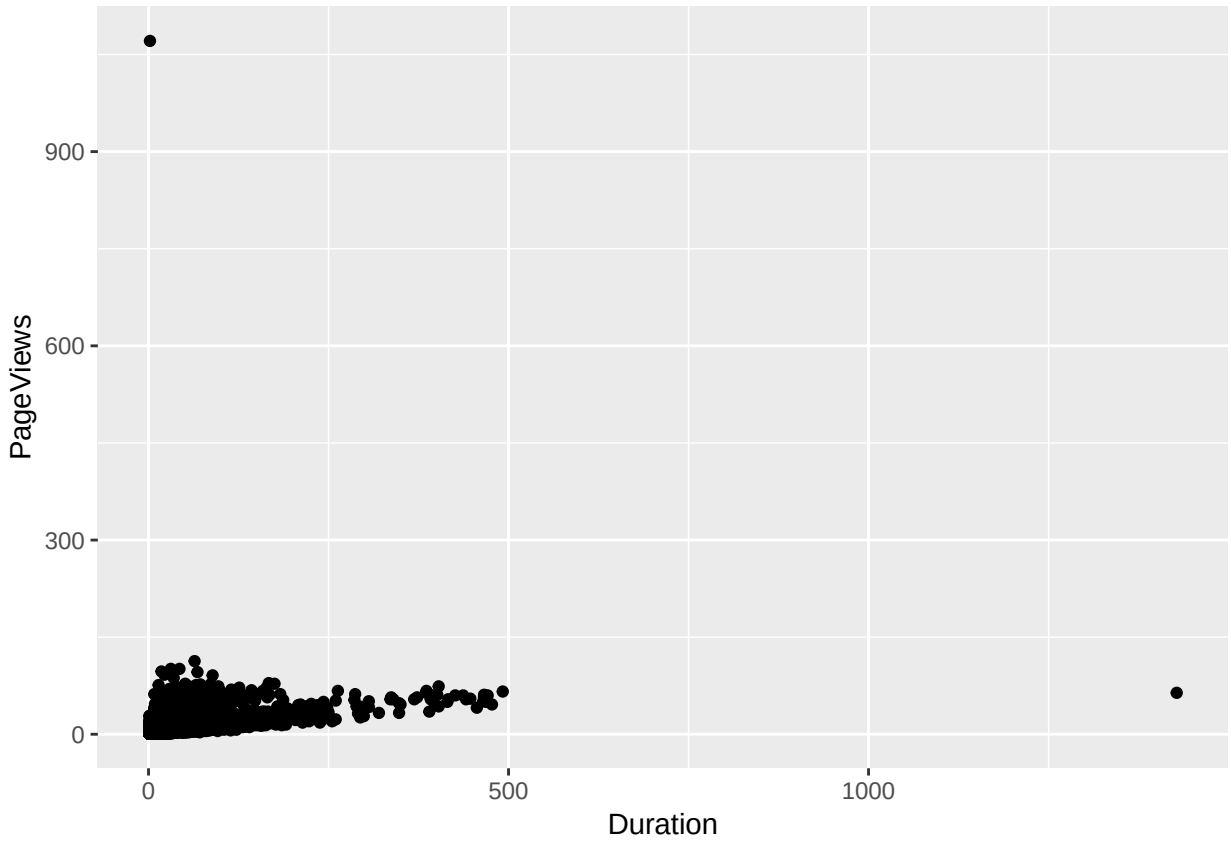
```
summary(comScore)
```

```
##      Sales              Buy          Duration        PageViews
##  Min.   :   0.000   No Buy:7265   Min.   :   1.00   Min.   :   1.00
##  1st Qu.:   0.000   Buy   :2735   1st Qu.:   2.00   1st Qu.:   3.00
##  Median :   0.000                 Median :   8.00   Median :   6.00
##  Mean   :   2.397                 Mean   :  20.42   Mean   :   9.63
##  3rd Qu.:   0.990                 3rd Qu.:  22.00   3rd Qu.:  11.00
##  Max.   :4023.000                 Max.   :1428.00   Max.   :1071.00
```

```
#Notice that most visits do not result in a purchase.  For this
#reason means are higher than medians. Also the median value of
#Sales is 0.  Finally, notice that the maximum values of Sales,
#Duration and Page Views are all very high.
```

3. Without using the `qplot` function, produce a scatter plot of duration against page views

```
ggplot(comScore,aes(x=Duration,y=PageViews))+geom_point()
```




```
# There are two obvious outliers.  The outlier in page views
#implies that 1071 clicks were recorded in 2 minutes.  This may be
#due to an error in how page views are recorded or a bot.  The
#outlier in duration implies that 64 pages were clicked on in over
#23 hours.  Here it is possible that the user simply left the
#computer on.
```

More Advanced

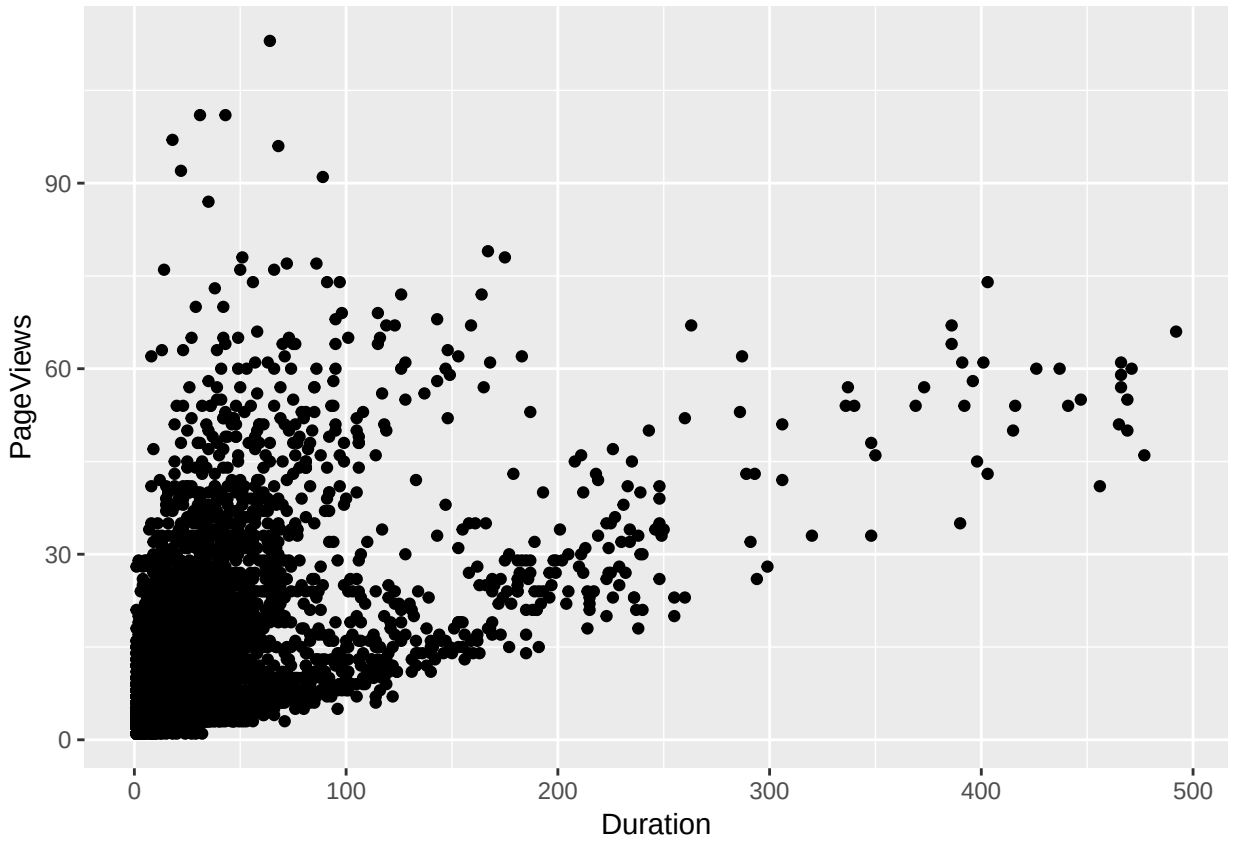
1. Create a new data frame that removes the two outliers using the `filter` function in `dplyr`.

```
# Only select Page Views and Duration
# that are less than maximum values
comScore_no<-filter(comScore,
                    (PageViews<1071)&
                      (Duration<1428))

#Note you must use the and operator. As an additional exercise can
#you think of another way to complete this task?
```

2. Do the scatterplot again with the outliers removed.

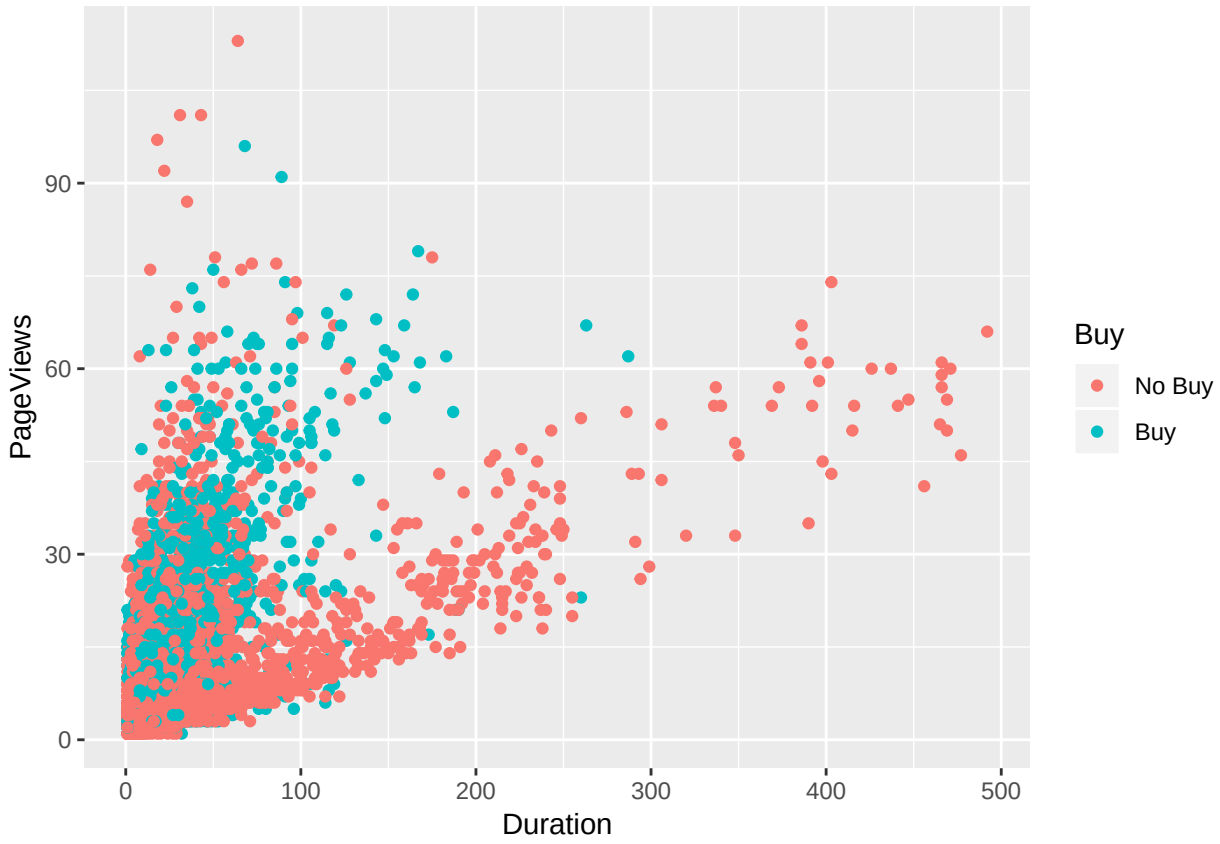
```
# Same as before, simply change the data frame
ggplot(comScore_no,aes(x=Duration,y=PageViews))+geom_point()
```




```
#There is a positive relationship between page views and clicks
#however the points almost appear to come from two different
#distributions, one where the correlation is stronger than the
#other
```

3. Do the scatterplot where the points have a different colour if the observation corresponds to a buy and a different colour if it corresponds to no buy.

```
#Add the color aesthetic
ggplot(comScore_no,aes(x=Duration,y=PageViews,col=Buy))+
  geom_point()
```




```
#Here the group with low correlation between page views and #duration are seen to mostly be non-purchas
#the correlation is higher has a higher incidence of buy.  This
#may suggest that websites that are sticky, i.e. involve more #interaction result in a higher purchase
#involve multiple page views.
```