Analysis on Pricing of Tents with a Deeper Usage of `ggplot2`

Dominic Tu

10/29/2017

Post1

Introduction

Couple of weeks ago, we have learn about a very useful package `ggplot2`. This package is a plotting system for R, based on the grammar of graphics. `ggplot2` takes the good parts of base and lattice graphics and takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.

However, I believe the one lab and homework assignment is not enought to learn about all the functions available in `ggplot2`. I was to use this post to dive deeper into `ggplot2`, using the other graphing tools that we have not touched in class. We will use these function to further analyze a certain data set, specifically data in tent models.

Throughout my life, I have always been interested in camping. It is a trip my family looks forward to evey year. I though it was a great opportunity to analyze data in the tents similar to the data set we work with in this course. So the analysis should seem familiar, but with a refreshing new set of numbers.

Through this post, you will be able to full immerse yourself into `ggplot2` and finally mastering a package that is extremely helping in data reporting and data visualization. I want help other solidify or improve on their current skills in R. This post will be a learning to tools for other and a project for my our curiousity.

Data Preparation

First we have to load `ggplot2` and download the dataset.

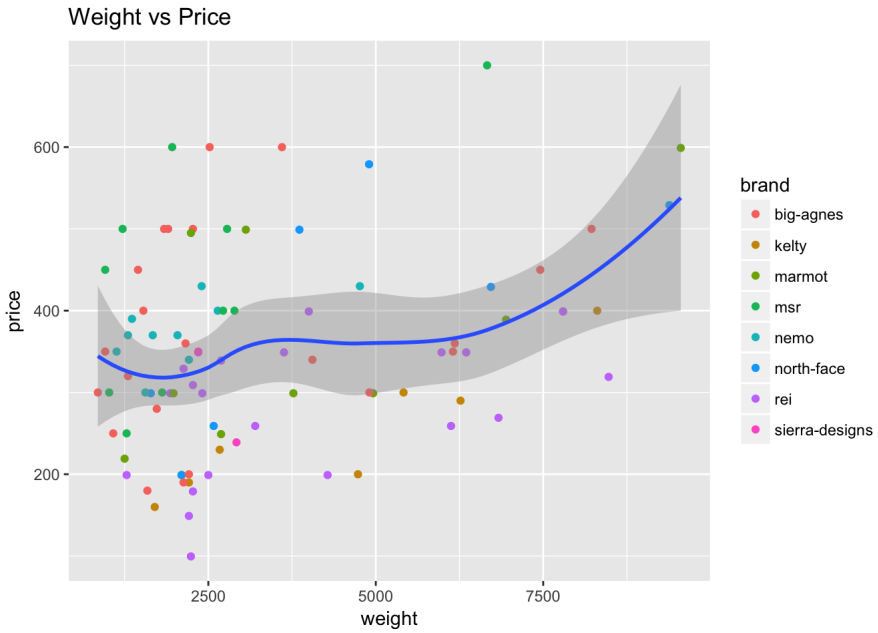
```
#loading package ggpot2
library(ggplot2)

#download the dataset
github <- "https://github.com/ucb-stat133/stat133-fall-2015/raw/master/"
csv <- "data/tents1.csv"
download.file(url = paste0(github, csv), destfile = 'tents1.csv')
dat <- read.csv('tents1.csv', stringsAsFactors = FALSE)
```

Analysis & Correlations

1. Scatter Plot

Scatterplot is frequently used to preform basic data analysis, which makes it easy to see a general correlation between two variables. A scatter plot pairs up values of two quantitative variables in a data set and display them as geometric points inside a Cartesian diagram. In this example, I will plot the values of Weight and Price.

```
ggplot(data = dat, aes(x = weight, y = price)) +
  geom_point(aes(col = brand)) +
  ggtitle("Weight vs Price") +
  geom_smooth(method = loess)
```

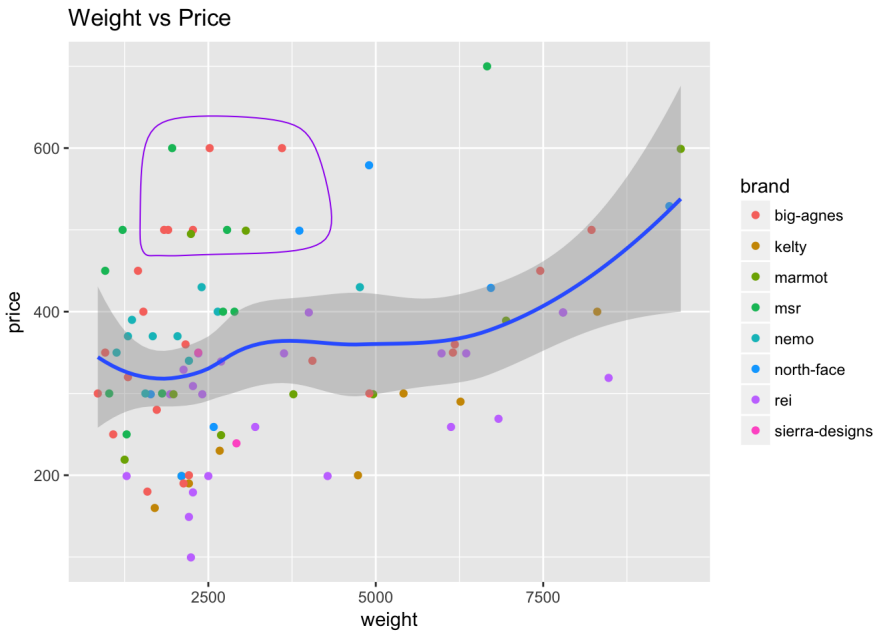


Analysis: There is not a clear or obvious correlation between the weight of a tent and its price. However, it does seem to have a positive relation, especially once the weight is greater than 7000. There seems to be a clear positive correlation to price, which can come from the more features a bigger tent may have.

2. Scatter Plot with Outliers

Notice from the graph above, there are many points far from the loess curve. These are considered outliers because it doesn't follow the gernal trend of a positive correlation. So, we want to point out these points and figure out why those data points are off. Just in luck, there is another package called `ggalt` which have a function called `geom_encircle` that automatically enclose points in a polygon.

```
#download the encircling package
library(ggalt)

#create dataset for the area that is going to be enclosed
gg1 = dat[dat$price >= 490 & dat$price <= 600 & dat$weight >= 1800
          & dat$weight <= 4000, ]

#plor the scatter plot with the enclosed area
ggplot(data = dat, aes(x = weight, y = price)) +
  geom_point(aes(col = brand)) +
  ggtitle("Weight vs Price") +
  geom_smooth(method = loess) +
  geom_encircle(data = gg1, aes(x = weight, y = price), col = "purple")
```

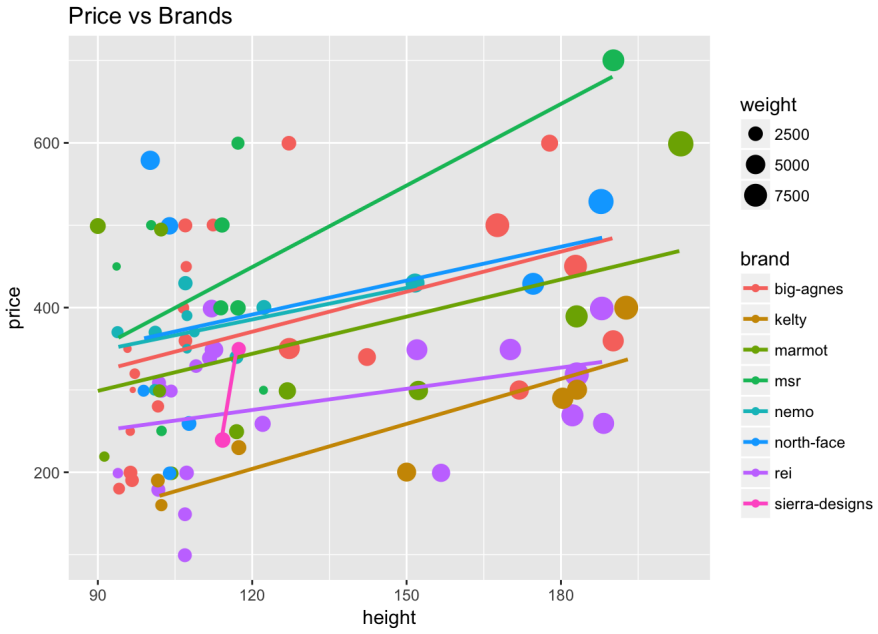


Analysis: The `geom_encircle` is a amazing tool to really point out a certain group of points. With the generally positive correlation, these points enclosed in the circle seems to be outliers to the trend. These tents are light, but still expensive. Maybe that is the feature of these tents, which might be the reason for the higher pricing. The tents might be made of more expensive lighter fabric and materials. Lets look further into the data.

3. Bubble Plot

The bubble plot is similar to a scatter plot. Like the scatter plot, the points are plotted on a xy chart. The x and y are mapped to a quantitative variable. The third variable is mapped by size. Bubble charts are used when you want to compare data points on three quantitative variables. The x and y position represent the magnitude of two of the quantitative variables, and the area of the bubble represents the magnitude of the third quantitative variable. For this example, we are goin to see how the price of each tent compare between brand and its weight. This bubble chart gives us a more interactive visual to understand how variable correlates with each other.

```
ggplot(dat, aes(x = height, y = price)) +
  geom_jitter(aes(col = brand, size = weight)) +
  geom_smooth(aes(col = brand), method = 'lm', se = F) +
  labs(title = "Price vs Brands")
```



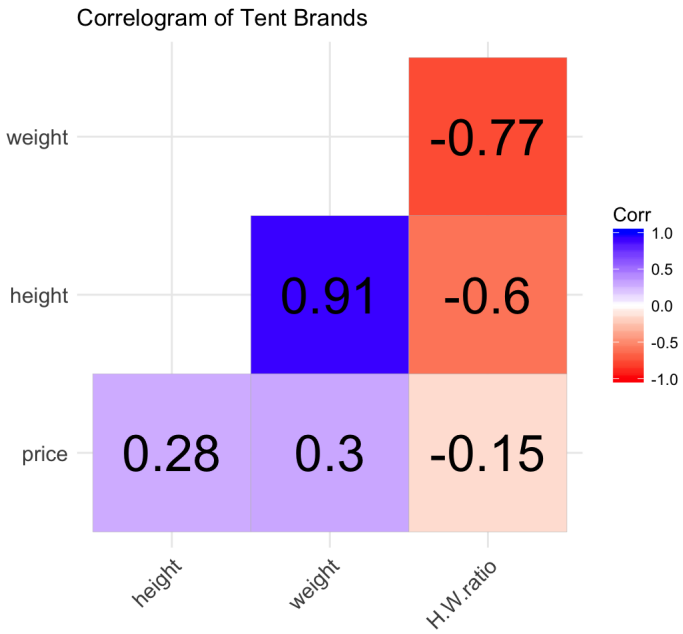
Analysis: From the bubble plot, we can see that that there is a noticeable increase in price as height increases. It is obvious that there is a positive correlation between the two variables. However, with the bubble graph, we can compare with increases between prices and height with brands. We can see which brand prices increases faster, which is brand is more expensive, and which is least expensive. From the regression curves, Sierra-Designs prices increases the most between its product. This is caused from the low number of 2 products. MSR seems to be the most expensive tent by far. Its regression line is above all its competitors and its prices increases considerably quickly between its products. The cheapest brand is Kelty, which has its regression line below all its competitor.

4. Correlogram

Correlation is a powerful statistics that represents the effects of a variable on another variable. We have been looking at the correlation of only a few variables throughout this post. However, there is a great package called `ggcorrplor` that can compute and show the correlation between variables with a single number. Correlograms help us visualize the data in correlation matrices. For example, we will use the function

`ggcorrplot` to find the correlation between the variables.

```
#load the package
library(ggcorrplot)

#create data set for ggcorrplot
dat_corr = data.frame('price'= dat$price,
                      'height' = dat$height,
                      'weight' = dat$weight,
                      'H/W ratio' = dat$height / dat$weight)

#create the correlation plot
ggcorrplot(round(cor(dat_corr), 2),
           type = "lower",
           lab_size = 10,
           method ="square",
           colors = c("red", "white", "blue"),
           title= "Correlogram of Tent Brands",
           lab = TRUE)
```



Analysis: From the correlogram, we are correct to see that there is a positive correlation betweem price and height. However, since there are some product that features its light weight, it significantly decreases the correlation. A correlogram gives a very high-level of our expectation of effects between variables. Moving forward, I will use the `ggcorrplot` to give a brief analysis between variable, which will give me useful insight before I dive deeper in our analysis with all the other functions of `ggplot2`.

Conclusion

This concludes my deeper `ggplot2` tutorial. We dive deeper into scatter plots, bubble plots, and correlogram. `ggplot2` is a very powerful tool for data visualization. This is will help you in any statistical/data analysis. Thank you for tuning into my post!

Reference

- https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
- https://github.com/ucb-stat133/stat133-fall-2015/blob/master/data/tents1.csv
- https://cran.r-project.org/web/packages/ggalt/vignettes/ggalt_examples.html
- https://cran.r-project.org/web/packages/ggalt/ggalt.pdf
- https://www.statmethods.net/advgraphs/correlograms.html
- http://sharpsightlabs.com/blog/bubble-chart-in-r-basic/
- http://rmarkdown.rstudio.com/authoring_basics.html