Assignment 4

Brian Detweiler

April 11, 2017

1. Visual quantification: To visualize the following quantity, mention what plot you will generate

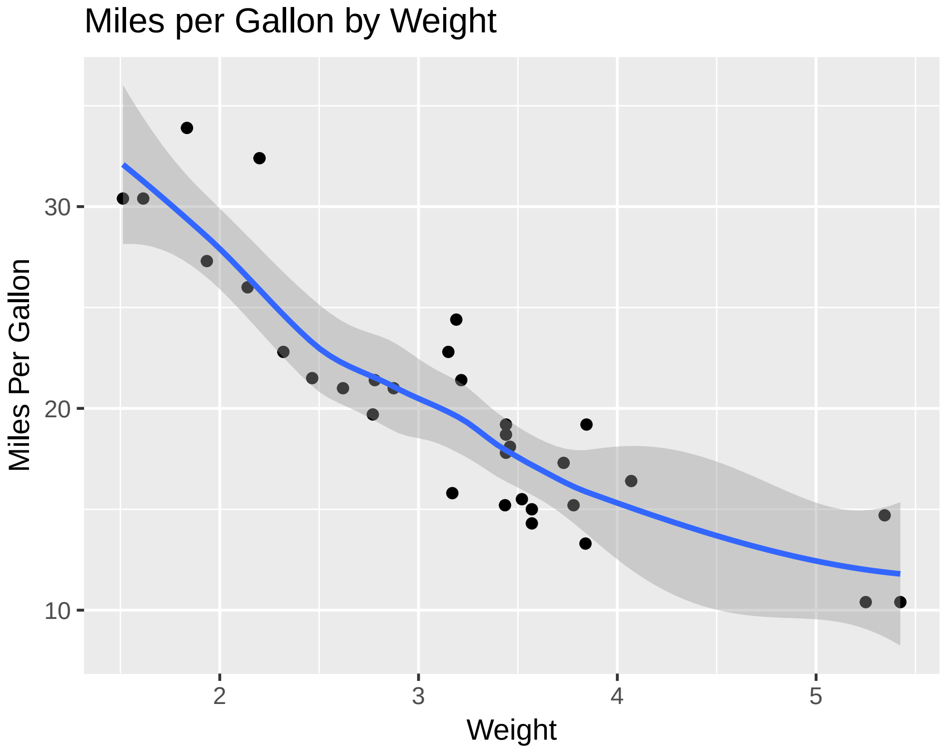
1. Association between two variables when

a) both are numeric

Answer: We would use a scatter plot.

Example:

```r
data(mtcars)

ggplot(mtcars, aes(x=wt, y=mpg)) +
  geom_point() +
  geom_smooth(method="loess") +
  labs(title="Miles per Gallon by Weight", x="Weight", y="Miles Per Gallon")
```

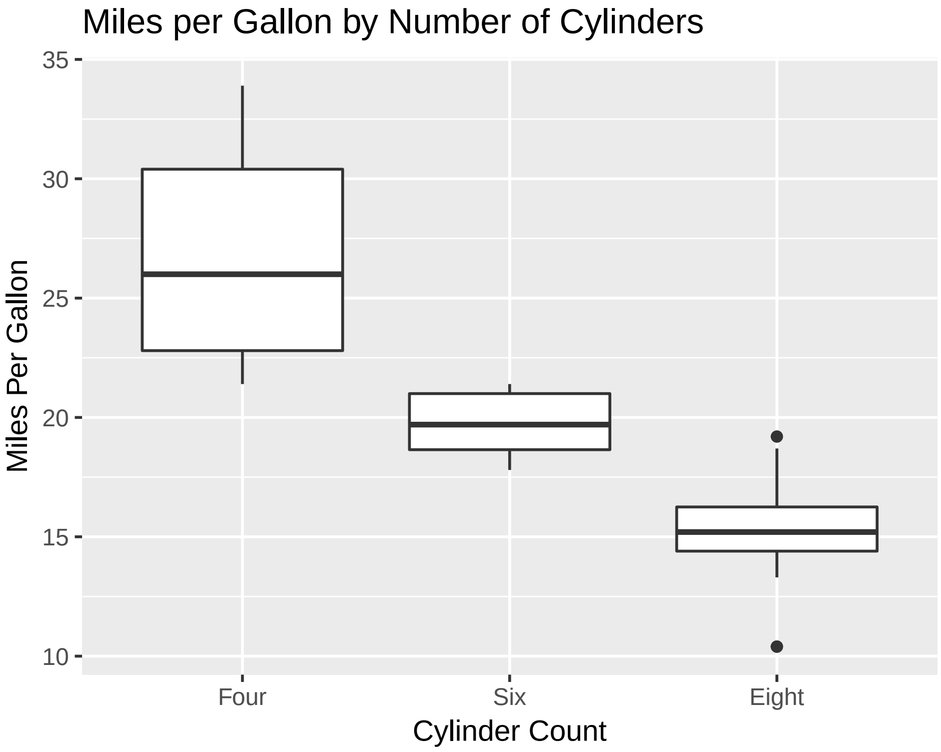


b) one numeric and one categorical

Answer: For this we could use a boxplot.

Example:

```
ggplot(mtcars, aes(x=factor(cyl), y=mpg)) +
  geom_boxplot() +
  labs(title="Miles per Gallon by Number of Cylinders", x="Weight", y="Miles Per Gallon") +
  scale_x_discrete("Cylinder Count", labels=c("Four", "Six", "Eight"))
```



c) both are categorical

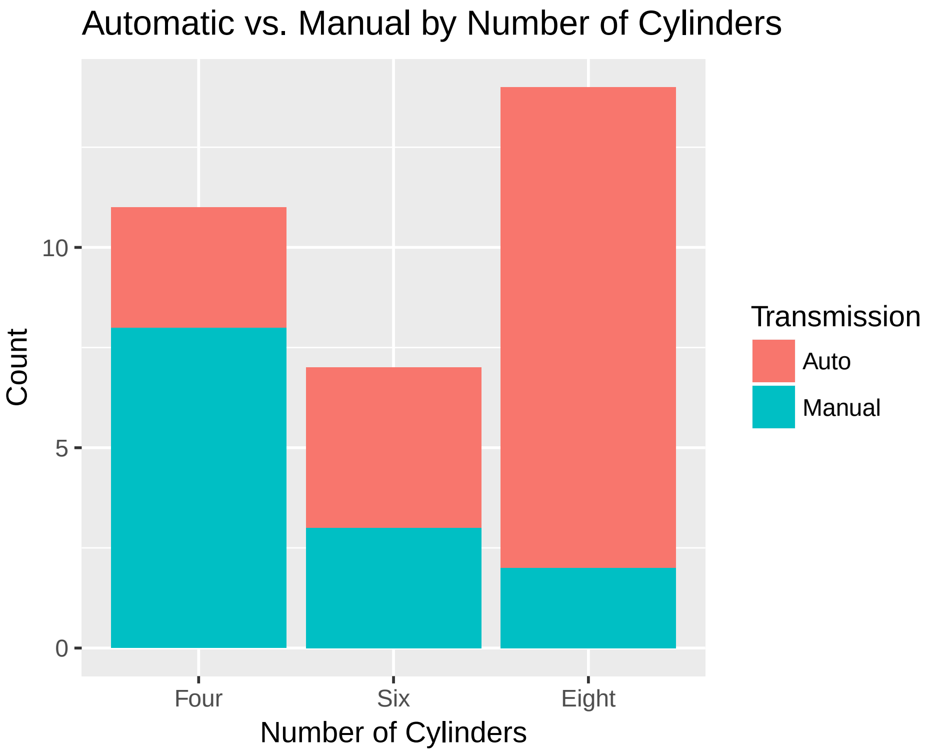
Answer: A stacked barchart will work in this case.

Example:

```
factor(mtcars$am)
```

```
##  [1] 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1
## Levels: 0 1
```

```
ggplot(mtcars, aes(factor(cyl), fill=factor(am))) +
  geom_bar() +
  labs(title="Automatic vs. Manual by Number of Cylinders", x="Cylinders", y="Count") +
  scale_x_discrete("Number of Cylinders", labels=c("Four", "Six", "Eight")) +
  scale_fill_discrete("Transmission",
                      labels=c("Auto", "Manual"))
```

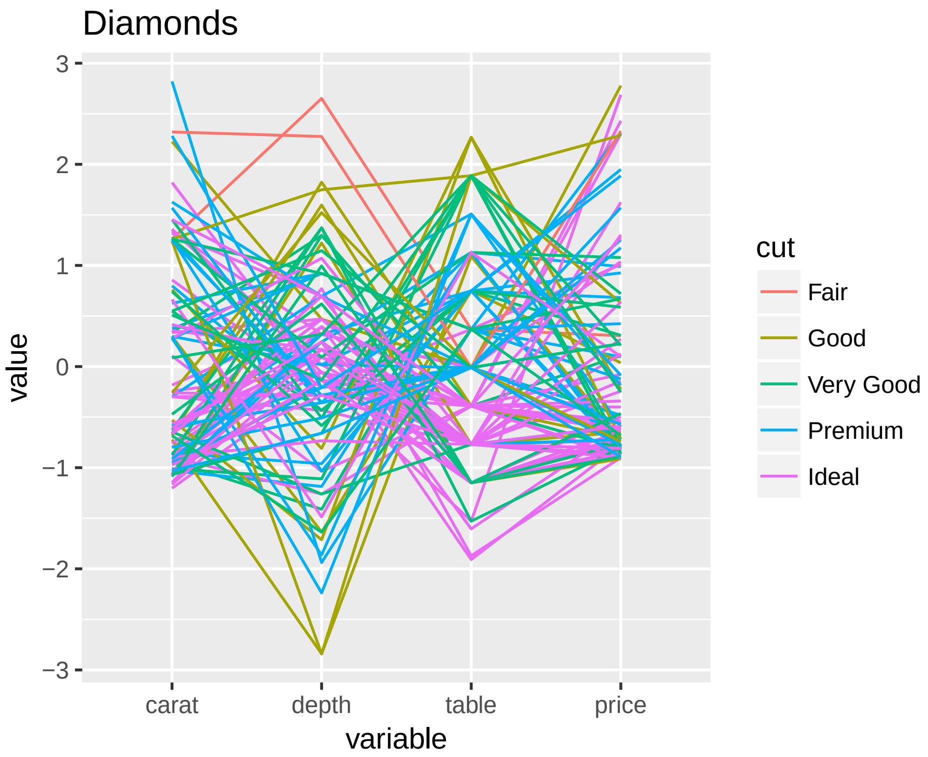


Association of a categorical variable with many numerical variables together

Answer: For this we can use a parallel coordinate plot.

Example:

```
data(diamonds)
diamonds.samp <- diamonds[sample(1:dim(diamonds)[1], 100), ]

ggparcoord(data = diamonds.samp, columns = c(1, 5:7), groupColumn = 'cut') +
  labs(title='Diamonds')
```



Difference of two variables in terms of

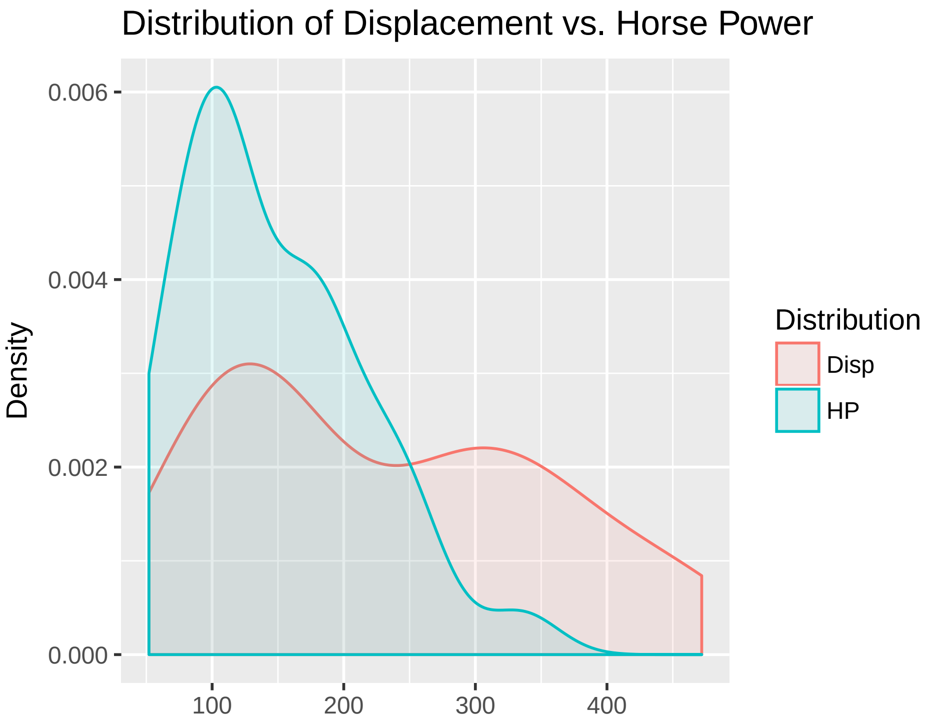
a) spread

Answer: Either a side-by-side boxplot or an overlayed density plot will work here.

Example:

```
df <- rbind(data.frame(x=mtcars$disp, Distribution='Disp'),
            data.frame(x=mtcars$hp, Distribution='HP'))
ggplot(df, aes(x, group=Distribution, col=Distribution, fill=Distribution)) +
  geom_density(position='dodge', alpha=0.1) +
  labs(title='Distribution of Displacement vs. Horse Power', x='', y='Density')
```

```
## Warning: Width not defined. Set with `position_dodge(width = ?)`
```

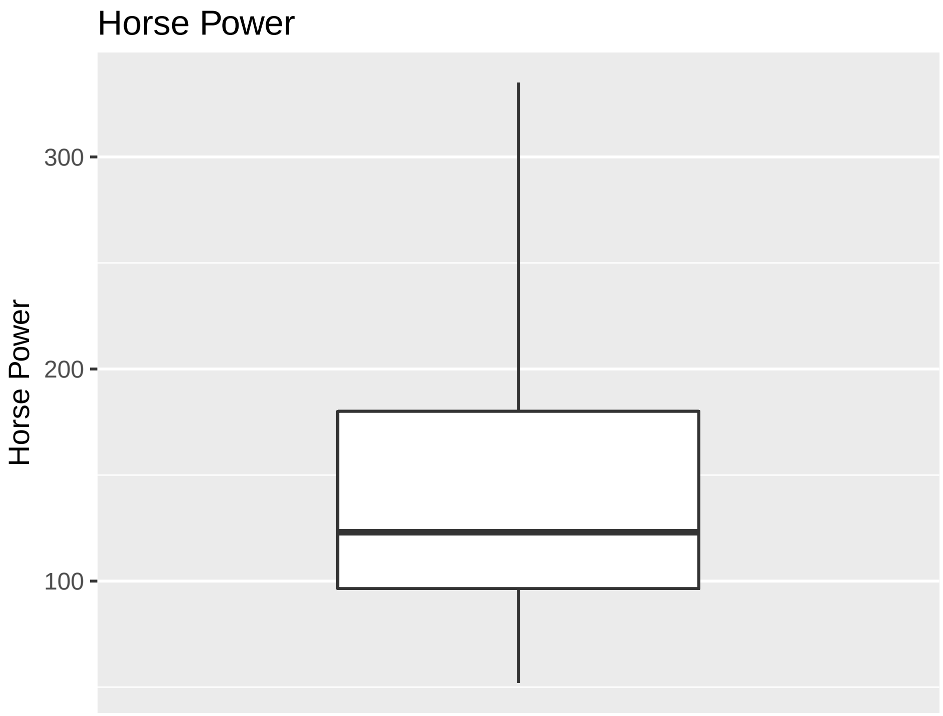


b) center

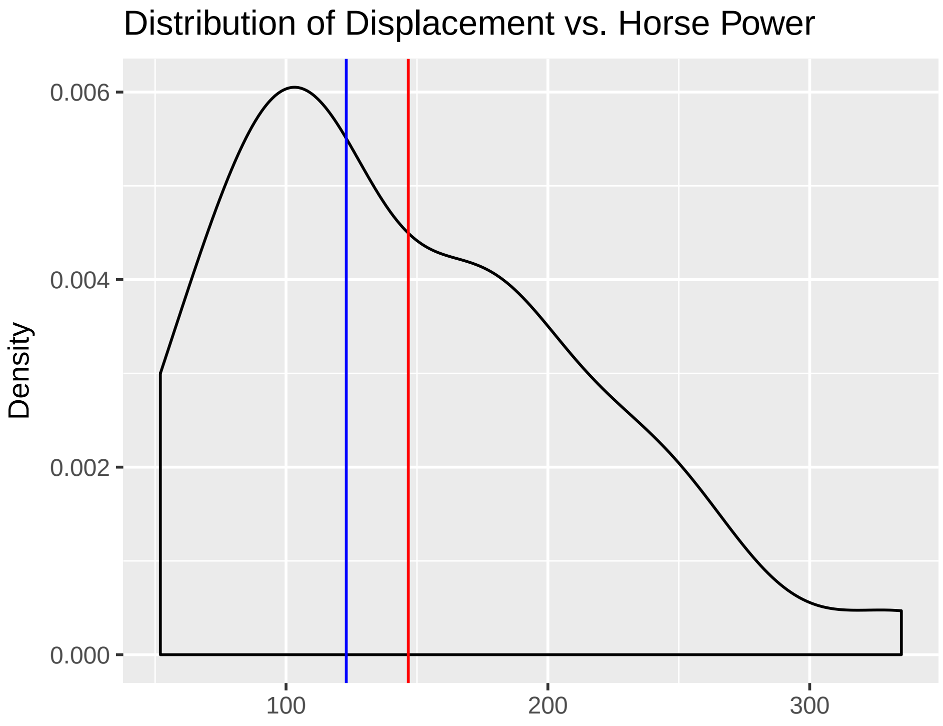
Answer: A boxplot shows the median, or we can add a mean and median to a density plot.

Example: Here, we'll show both options.

```r
hp <- mtcars$hp
lower <- quantile(hp)[[2]]
upper <- quantile(hp)[[4]]
middle <- quantile(hp)[[3]]
ymin <- quantile(hp)[[1]]
ymax <- quantile(hp)[[5]]
ggplot(data.frame(hp), aes(hp, x=0)) +
  geom_boxplot(stat = "identity", aes(lower=lower, upper=upper, middle=middle, ymin=ymin, ymax=ymax)) +
  labs(title="Horse Power", x="", y="Horse Power") +
  scale_x_discrete(labels=(""))
```



```r
ggplot(mtcars, aes(hp, fill=hp)) +
  geom_density() +
  geom_vline(xintercept = mean(hp), col="red") +
  geom_vline(xintercept = median(hp), col="blue") +
  labs(title='Distribution of Displacement vs. Horse Power', x='', y='Density')
```

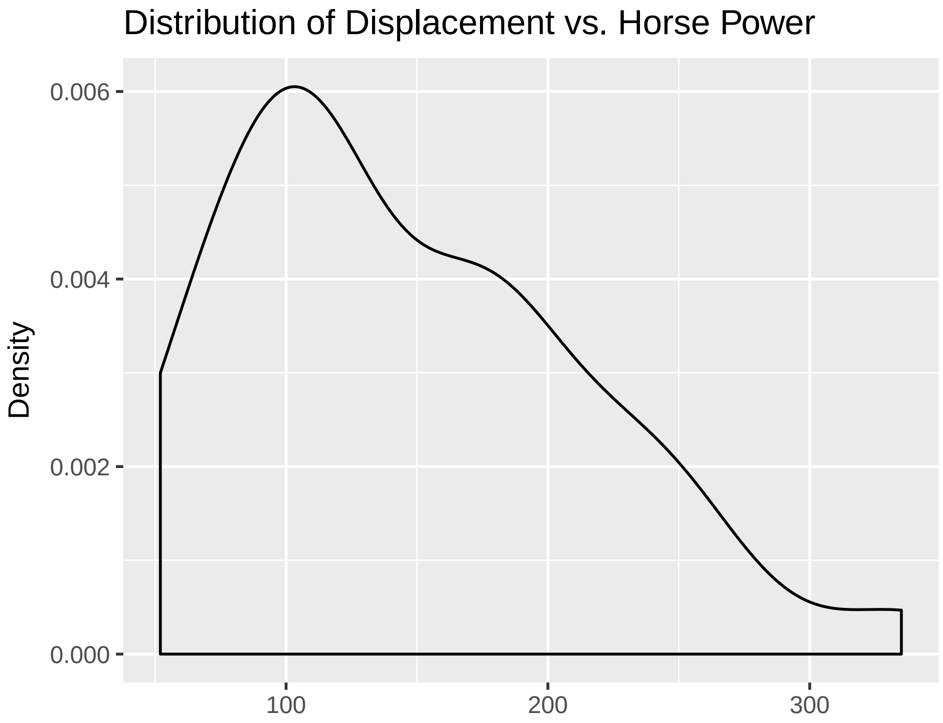


c) overall distribution

Answer: A distribution plot.

Example:

```r
ggplot(mtcars, aes(hp, fill=hp)) +
  geom_density() +
  labs(title='Distribution of Displacement vs. Horse Power', x='', y='Density')
```



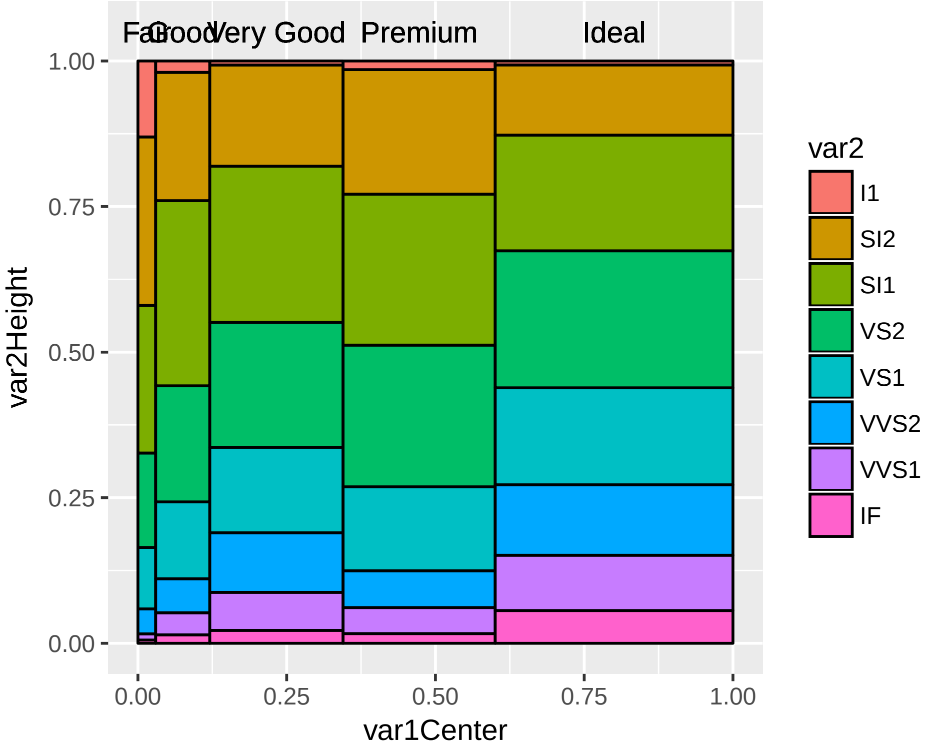
Proportion of categories of a variable

Answer: A mosaic plot can show this.

Example:

```r
# Code from http://stackoverflow.com/questions/19233365/how-to-create-a-marimekko-mosaic-plot-in-ggplot
ggMMplot <- function(var1, var2){
  require(ggplot2)
  levVar1 <- length(levels(var1))
  levVar2 <- length(levels(var2))

  jointTable <- prop.table(table(var1, var2))
  plotData <- as.data.frame(jointTable)
  plotData$marginVar1 <- prop.table(table(var1))
  plotData$var2Height <- plotData$Freq / plotData$marginVar1
  plotData$var1Center <- c(0, cumsum(plotData$marginVar1)[1:levVar1 -1]) +
    plotData$marginVar1 / 2

  ggplot(plotData, aes(var1Center, var2Height)) +
    geom_bar(stat = "identity", aes(width = marginVar1, fill = var2), col = "Black") +
    geom_text(aes(label = as.character(var1), x = var1Center, y = 1.05))
  }

ggMMplot(diamonds$cut, diamonds$clarity)
```

```
## Warning: Ignoring unknown aesthetics: width
```



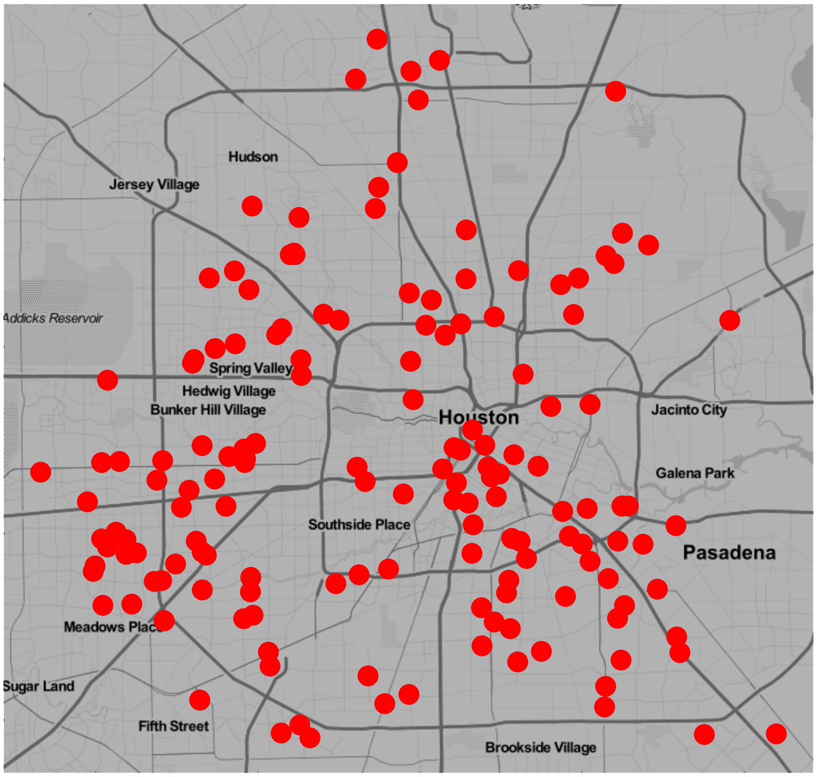
Spatial dependency of a variable

Answer: A map.

Example:

```
murder <- subset(crime, offense == "murder")
qmplot(lon, lat, data = murder, colour = I('red'), size = I(3), darken = .3)
```

```
## Using zoom = 11...

## Map from URL : http://tile.stamen.com/toner-lite/11/479/845.png

## Map from URL : http://tile.stamen.com/toner-lite/11/480/845.png

## Map from URL : http://tile.stamen.com/toner-lite/11/481/845.png

## Map from URL : http://tile.stamen.com/toner-lite/11/482/845.png

## Map from URL : http://tile.stamen.com/toner-lite/11/479/846.png

## Map from URL : http://tile.stamen.com/toner-lite/11/480/846.png

## Map from URL : http://tile.stamen.com/toner-lite/11/481/846.png

## Map from URL : http://tile.stamen.com/toner-lite/11/482/846.png

## Map from URL : http://tile.stamen.com/toner-lite/11/479/847.png
## Map from URL : http://tile.stamen.com/toner-lite/11/480/847.png
## Map from URL : http://tile.stamen.com/toner-lite/11/481/847.png
## Map from URL : http://tile.stamen.com/toner-lite/11/482/847.png
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```



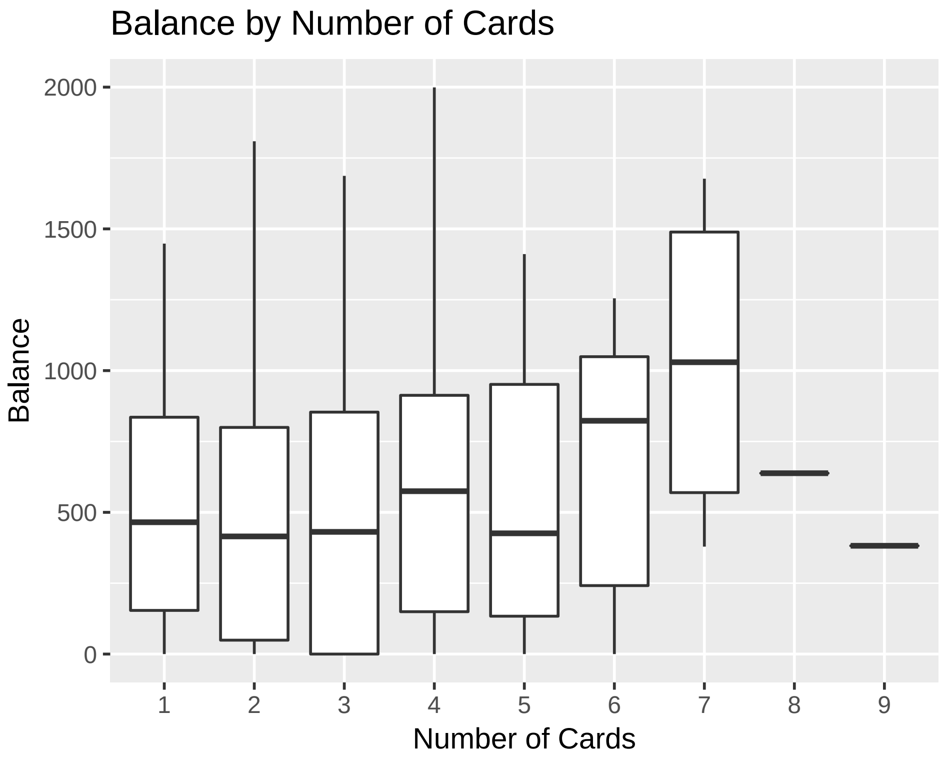
Credit card balances of 400 individuals are given in the data file `Credit.csv`. The goal is to find which explanatory variables are affecting the card balance. Based on this data, answer the following questions. Justify your answer by providing a suitable plot in each case.

```
credit <- read.csv('Credit.csv')
head(credit)
```

```
##   id  Income Limit Rating Cards Age Education Gender Student Married
## 1  1  14.891  3606    283     2  34        11   Male      No     Yes
## 2  2 106.025  6645    483     3  82        15 Female     Yes     Yes
## 3  3 104.593  7075    514     4  71        11   Male      No      No
## 4  4 148.924  9504    681     3  36        11 Female      No      No
## 5  5  55.882  4897    357     2  68        16   Male      No     Yes
## 6  6  80.180  8047    569     4  77        10   Male      No      No
##   Ethnicity Balance
## 1 Caucasian     333
## 2     Asian     903
## 3     Asian     580
## 4     Asian     964
## 5 Caucasian     331
## 6 Caucasian    1151
```

a) Do you think number of cards has anything to do with higher balance?

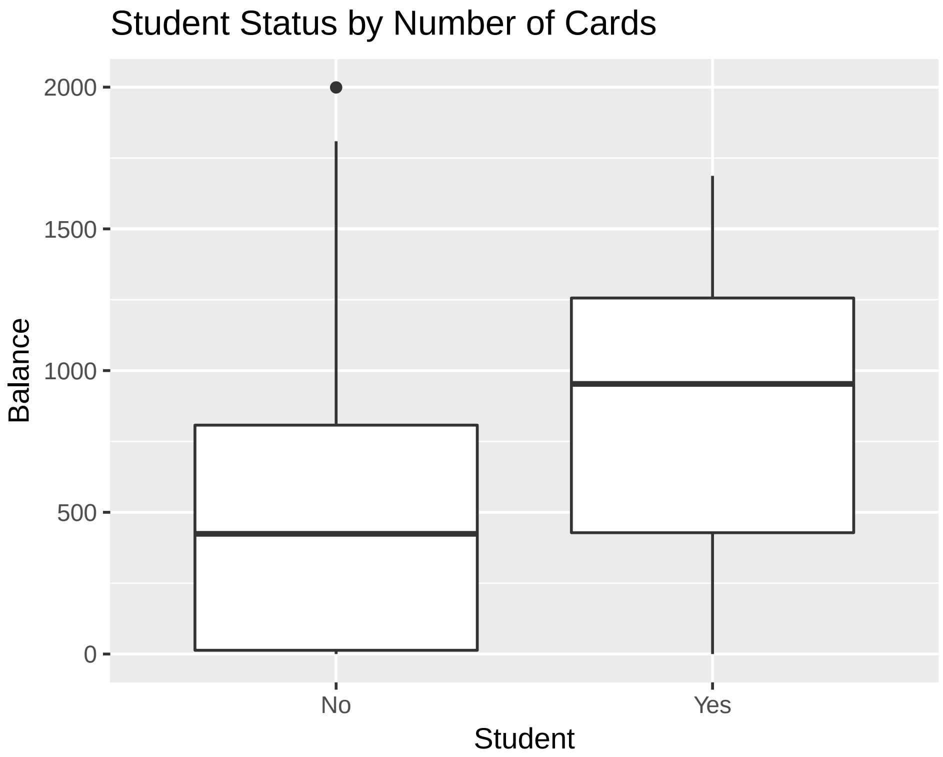
```
ggplot(credit, aes(x=factor(Cards), y=Balance)) +
  geom_boxplot() +
  labs(title="Balance by Number of Cards", x="Number of Cards")
```



There does not appear to be a linear relationship between number of cards and balance. Although people with 6 and 7 cards appear to have the highest balance by median.

b) Do you think that higher balances can be attributed to individual's studentship status?

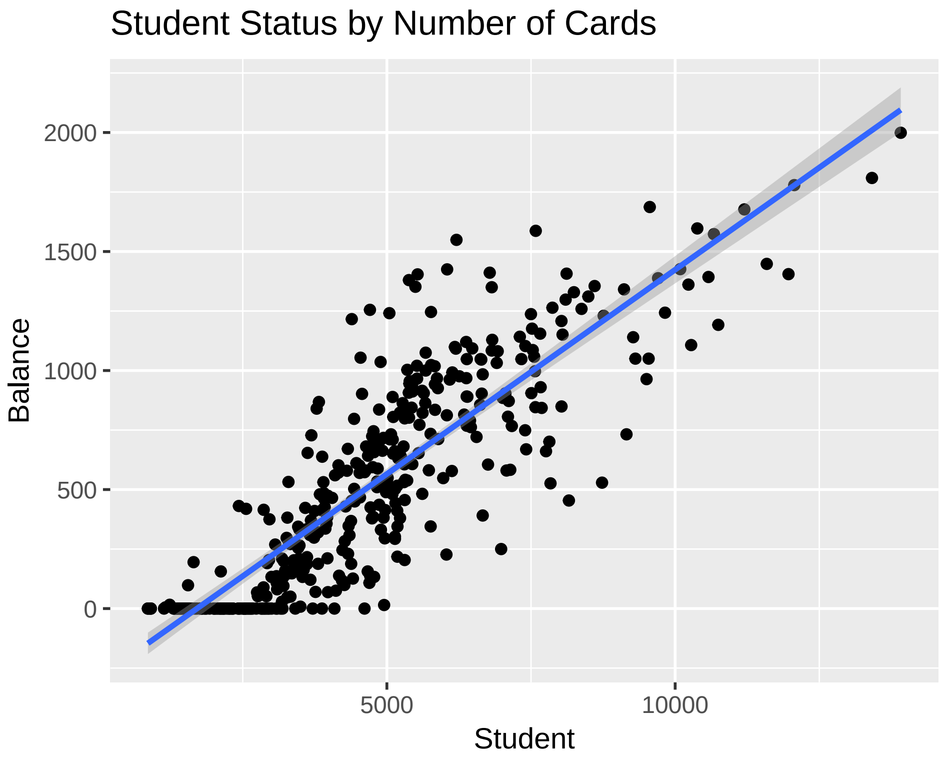
```r
ggplot(credit, aes(x=factor(Student), y=Balance)) +
  geom_boxplot() +
  labs(title="Student Status by Number of Cards", x="Student")
```



There appears to be clear visual evidence that being a student has a positive correlation to credit card balance. As most of us know from first hand experience.

c) Does higher limit causes higher balance?

```
fit <- lm(credit$Balance ~ credit$Limit)
r.sq <- summary(fit)$adj.r.squared

ggplot(credit, aes(x=Limit, y=Balance)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(title="Student Status by Number of Cards", x="Student")
```



With an adjusted R^2 of 0.7418753, there is a strong positive correlation between limit and balance.

d) Is there any interaction effect of Student and limit on Balance?

```
credit$Student <- as.numeric(credit$Student) - 1
fit <- lm(credit$Balance ~ (credit$Student + credit$Limit)^2)
sum.fit <- summary(fit)
p.val <- sum.fit$coefficients[4,4]
```

With a p-value of 0.1830097, there is not enough evidence to say that the interaction effect is significantly different than zero.

e) Is there any interaction effect of Income and Ethnicity on Balance?

```
credit$Ethnicity <- as.numeric(credit$Ethnicity) - 1
fit <- lm(credit$Balance ~ (credit$Income + credit$Ethnicity)^2)
sum.fit <- summary(fit)
eth.p.val <- sum.fit$coefficients[3,4]
inc.eth.p.val <- sum.fit$coefficients[4,4]
```

There doesn't appear to be any effect caused by Ethnicity, with a p-value of 0.124171, and the p-value for the interaction effect is 0.0651074 which is not less than 0.05, and thus we cannot say that the effect is significantly different than zero.

f) Is there any 3-way interaction effect of Income, Ethnicity and Student on Balance?

```r
fit <- lm(credit$Balance ~ (credit$Income + credit$Ethnicity + credit$Student)^3)
sum.fit <- summary(fit)
sum.fit
```

```
##
## Call:
## lm(formula = credit$Balance ~ (credit$Income + credit$Ethnicity +
##     credit$Student)^3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -811.9 -332.3  -54.9  326.9  818.6
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                    120.1304    59.1141   2.032
## credit$Income                                    7.5234     0.9916   7.587
## credit$Ethnicity                                66.9405    39.9113   1.677
## credit$Student                                 613.1709   205.4414   2.985
## credit$Income:credit$Ethnicity                  -1.1116     0.6806  -1.633
## credit$Income:credit$Student                    -2.3873     3.1564  -0.756
## credit$Ethnicity:credit$Student                -95.7977   150.5243  -0.636
## credit$Income:credit$Ethnicity:credit$Student   -0.1989     2.6686  -0.075
##                                                Pr(>|t|)
## (Intercept)                                     0.04281 *
## credit$Income                                  2.41e-13 ***
## credit$Ethnicity                                0.09429 .
## credit$Student                                  0.00302 **
## credit$Income:credit$Ethnicity                  0.10322
## credit$Income:credit$Student                    0.44991
## credit$Ethnicity:credit$Student                 0.52487
## credit$Income:credit$Ethnicity:credit$Student   0.94062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.3 on 392 degrees of freedom
## Multiple R-squared:  0.2884, Adjusted R-squared:  0.2757
## F-statistic: 22.69 on 7 and 392 DF,  p-value: < 2.2e-16
```

```r
eth.p.val <- sum.fit$coefficients[3,4]
inc.eth.stu.p.val <- sum.fit$coefficients[8,4]
```

With a p-value of 0.9406179, there is no three-way interaction effect between Income, Ethnicity, and Student status on the balance.

3. Creating dashboard. For this problem you will start with `dashboard.R` as a template. Modify this template and create a new dashboard according to the instruction below;

a) Inside `sidebarMenu`, create a new `menuItem` called “My state crime”

b) Inside `dashboardBody` `tabItems`, create a new `tabItem` and make the header “My state crime map goes here”

c) Inside the newly created `tabItem` create a `fluidRow` using column instead of box which has following features. For example codes, review the link https://github.com/mamajumder/usa-crime/blob/master/ui.R

Use first 5 columns to create a `wellPanel` where there will be one `selectInput` and one `sliderInput`. `selectInput` should allow to select a specific crime rate to display while `sliderInput` will use a specific year of crime rate. For this use `usaCrimeDat.rds`

Use 6 columns to show the state crime map of USA colored by crime rate.

d) Inside server add a new output of the state crime map so that map can be interactively generated based on crime rate and the year selected.