Homework3

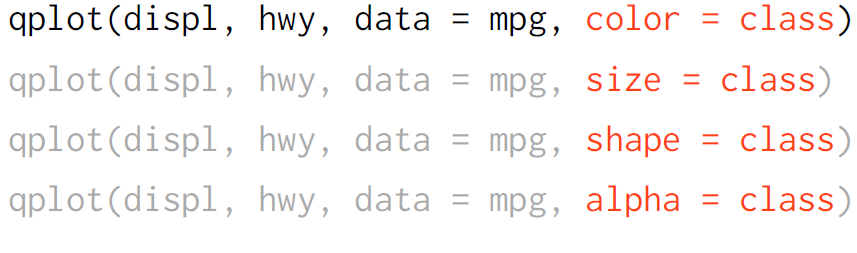
Chiffon

23 October 2016

### 1 阅读部分

* 最好的选择残差平方和最小的线
* y~x
* geom\_smooth将会呈现y~x的回归 method表示使用哪种方式  参数se表示是否显示置信区间
* 在lm()函数中，对于公式y~x 如果想要移除截距可以采用y~x-1,或者y~0+x的方式
* 类型变量：选择一个值作为基线，其他类型系数估计均在此基础上加减
* 在多元线性回归，coef()函数对于每一个元进行系数估计，且每个系数都是在其他元固定的情况下估计出来的
* 辛普森悖论：当你考虑第三个变量的时候，原来两个变量之间的关系可能会改变
* lm(earn~height+race+sex+ed+age,data = wages)等价于lm(earn~.,data = wages)
* 符号"."表示所有，也可以使用".-age"表示除age以外的所有

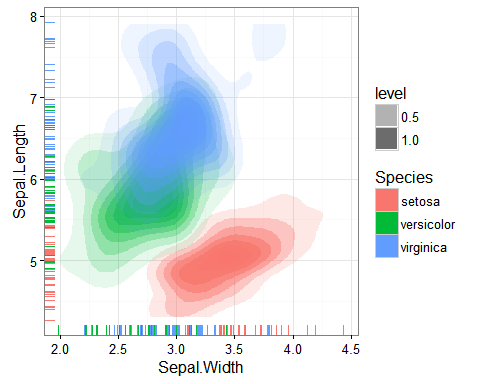
对于数据可视化中可采用四种方式进行展示，分别是颜色，尺寸，形状，透明度



* 离散的可以采用颜色的不同表示，连续的可以采用颜色深浅表示；
* 离散和连续均可以采用尺寸大小表示；
* 针对形状，离散可以采用不同形状表示，而连续就不行；

### 2 重现以下图形

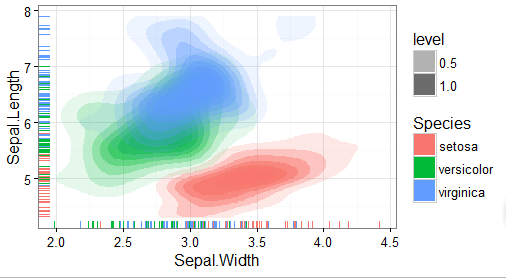
## Warning: package 'ggplot2' was built under R version 3.2.5



m1<-ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length),color=Species)

m2<-m1+stat\_density\_2d(aes(alpha =..level.. ,fill=Species), geom = "polygon")

m3<-m2+geom\_rug(position = "jitter",aes(color=Species))+guides(color=F)

m4<-m3+theme\_bw()

### 3 Kaggle: House Prediction

* 网址: [<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>](https://www.kaggle.com/c/house-prices-advanced-regression-techniques)
* 将数据分成两类，字符型和数值型；读取进入R工作环境当中；

cat\_var <- names(train)[which(sapply(train, is.character))]#字符型组名

numeric\_var <- names(train)[which(sapply(train, is.numeric))]#数值型组名

* 对于缺失值NA的情况，将train和test中的NA如果是数值型的采用0替代，如果是字符型的采用none替代；

例如：

train\_raw$Alley[is.na(train\_raw$Alley)] <- 'None'

test\_raw$Alley[is.na(test\_raw$Alley)] <- 'None'

train\_raw$BsmtFinSF1[is.na(train\_raw$BsmtFinSF1)] <- 0

test\_raw$BsmtFinSF1[is.na(test\_raw$BsmtFinSF1)] <- 0

* 将两种类型的列读入；

train\_cat <- train[,.SD, .SDcols = cat\_var]#新的数值型训练集

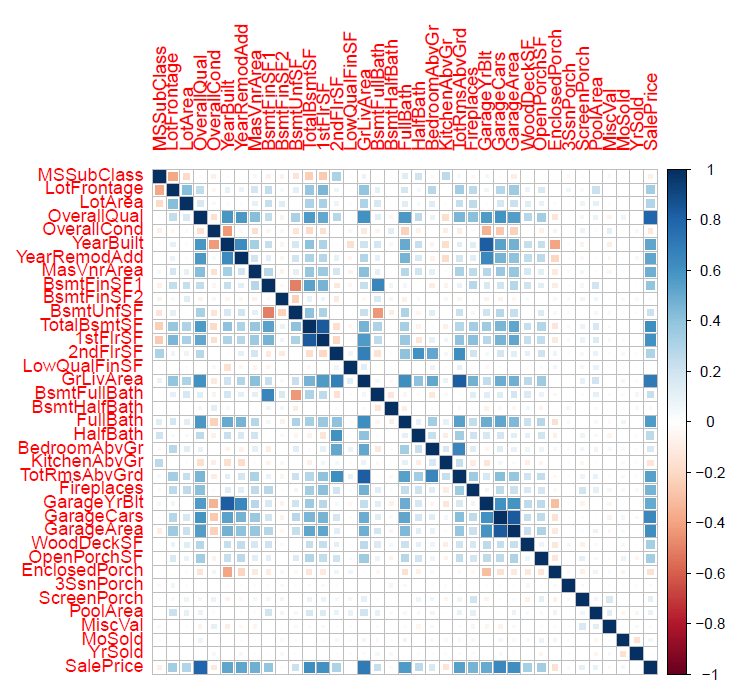
train\_cont <- train[,.SD,.SDcols = numeric\_var]#新的数值型训练集

* 对于数值型的求出每一个之间的相关性

correlations <- cor(na.omit(train\_cont[,-1, with = FALSE]))

* 绘制图形查看每个数值型字段之间的相关性

corrplot(correlations, method="square")



* 筛选出相关性大于0.5或者小于-0.1的列

spcor<-correlations[37,which(correlations[37,1:37]>0.5|correlations[37,1:37]< -0.1)]

spcor

OverallQual OverallCond YearBuilt YearRemodAdd

0.7978807 -0.1243912 0.5253936 0.5212533

TotalBsmtSF 1stFlrSF GrLivArea FullBath

0.6156122 0.6079691 0.7051536 0.5666274

KitchenAbvGr TotRmsAbvGrd GarageYrBlt GarageCars

-0.1404974 0.5470674 0.5047530 0.6470336

GarageArea EnclosedPorch SalePrice

0.6193296 -0.1548432 1.0000000

结果显示存在14个数值型的列与SalesPrice相关性高。

* 采用多元线性回归的方式对SalesPrice和14列做模拟：

将相关性较高的14列数值提取成新的数据框

num\_train<-train[,.SD, .SDcols = c(18,19,20,21,39,44,47,50,53,55,60,62,63,69,81)]

做多元线性模拟

m<-lm(salesPrice~.,data = num\_train)

* 结合测试集，预测数据

predict(m,test[,.SD, .SDcols = c(18,19,20,21,39,44,47,50,53,55,60,62,63,69)])

* 将预测的结果写入CSV文件当中

write.csv(pre,file="output.csv")