Homework3

### 1 阅读部分

* 阅读提供的材料,附加读书笔记

1. "03-Linear Models.pdf"
2. "03-Visualization.pdf"
3. "04-Customizing.pdf"

### 1.1线性模型

回归反映变量间关系，预测变量来预测相应变量。拟合线性模型使用lm。

常用函数：summary、fitte、plot、residuals

线性模型结果参数，coefficients中Intercept是截距，预测变量对应值（如:height）为斜率。Mutiple R-squared为变量间相关关系数值（模型方差解释率）。Pr(>|t|)反映截距是否显著为0的t检验（简单线性模型），p值小于0.01，显著不为0。

公式中常见符号，“+” 分隔预测变量；“:” 预测变量的交互项，如y ~a+b+a:b，表示y与a的关系依赖于b，随b变化的情况；“\*” 所有可能的交互方式，a \* b即a + b + a:b

多项式回归，可以通过I添加二次项或n次项，如:I(height^2)。有时，多项式拟合比线性拟合度更好。

预测变量多个时为多元线性回归，两预测变量关系可cor判断，公式中多个变量间用“+”分隔。

lm只能用于拟合，但是模型是否合适，需要回归诊断。Confint函数能够反映模型参数的置信区间，响应变量变化1%，预测变量在95%的置信区间内变动。

OSL回归的统计假设为：正态性、误差独立性、线性、同方差性

对lm返回对象用plot，产生模型拟合的图形。正态q-q图，若满足正态假设，图上的点应该都在45度线上。残差图-拟合图中，如果因变量与自变量线性相关，残差与预测值应该没有任何关系关联，应该随机分布，没有规律，若图中看出曲线关系，说明应该增加二次项来拟合。位置-尺度图，若满足同方差性，点应该在水平线周围随机分布。残差图-杠杆图能鉴别出离群点、高杠杆值点。

qqplot比plot图生成的正态假设更精确，若所有点都在线附近，且都在置信区间内，说明正态假设符合。Durbin-Watson检验能检验残差是否独立，p值不显著>0.01，说明无相关性。成分残差图能得出因变量与自变量间是否呈非线性关系。

ncvTest检验不显著p>0.01，满足同方差性，违反原假设，会有非水平的曲线。

### 1.2可视化

简单绘图：plot（散点图）、qplot

ggplot2包是强大的绘图工具，qplot函数(x,y,data = mpg)。

* 1. Aesthetic：color、size、shape、alpha（透明度）、size（设置点大小）
  2. Faceting（网格反映子集）:facet\_grid、facet\_wrap
  3. Geoms：smooth（平滑线）、boxplot（箱线图）、bar（连续变量：直方图；离散变量：条形图）、freepoly、density（核密度图）、polygon（多边形）等。
  4. 直方图绘制中，binwidth设置组宽，fill用于填充、color根据变量值绘制不同颜色。
  5. ggsave，将图保存（格式、大小）。

### 1.3定制绘图

ggplot

ggtitle增加标题

coord\_系列设置坐标：coord\_ploar改为极坐标（可以画出饼图）、flip（横竖坐标互换）、fixed（按比例缩小或放大）、coord\_trans

scale\_系列设置颜色、点大小、点形状：scale\_color\_gradient、scale\_size\_area、scale\_shape\_manual、scale\_fill\_brewer。

theme()设置图形元素；ggthemes包能选择主题（图模板）

ylab、xlab添加行、列标签。

theme函数调整lengend的位置，guides更改图例的类型。Scale系列能修改图例的名称、图例标签。不更改原图，修改图例使用scale\_fill\_discrete

extrafont包设置字大小、字体、字颜色。

注意：

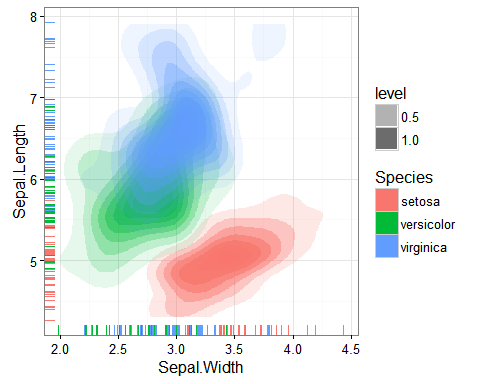
1. 定制化图形为了更清楚反映数据及关系，不等于花里胡哨的美。
2. 好的图形：标题、数值清晰、颜色分明、白背景、坐标轴、图例

问题：

ggplot中position参数被废弃了，不懂如何画出stack、dodge、identity、fill、jitter

### 2 重现以下图形

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



library(ggplot2)

ggplot(data = iris,aes(x=Sepal.Width,y=Sepal.Length,fill = Species)) +

theme\_bw() + #白色背景

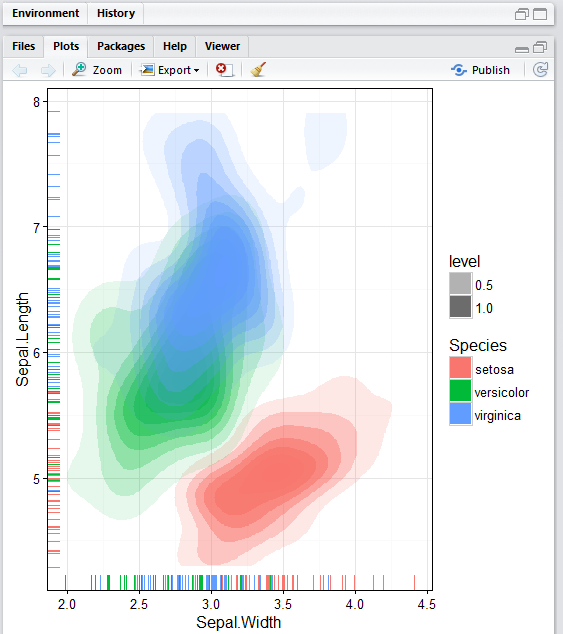
stat\_density2d(aes(alpha = ..level..),geom = "polygon") +#绘图区（二维）

geom\_rug(position = 'jitter',aes(color = Species)) + #边际地毯

theme(panel.border = element\_rect(colour = "black", fill = NA))+ #黑色框

guides(color=F)#删除多余的color标签

绘图结果：



### 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)

#### 要求

1. 注册并参加比赛
2. 参考并运行示例代码(可选)
3. 自己改良或者自己完成预测代码

代码如下：

train <- read.csv("F:/R/train.csv")

test <- read.csv("F:/R/test.csv")

str(test)

C:\Users\sherl\AppData\Roaming\Tencent\Users\1070746603\QQ\WinTemp\RichOle\LSUV~MRVJ`}J%R4M1I4KP[4.png

#test$SalePrice <- rep(NA, 1459)

分为离散和连续变量：

Fac <- sapply(train,is.factor)

trainDis <- train[,Fac]

Fac1 <- sapply(train,is.numeric)

trainCon <- train[,Fac1]

#缺失值处理

trainNew <- na.omit(trainCon)#直接删掉缺失值

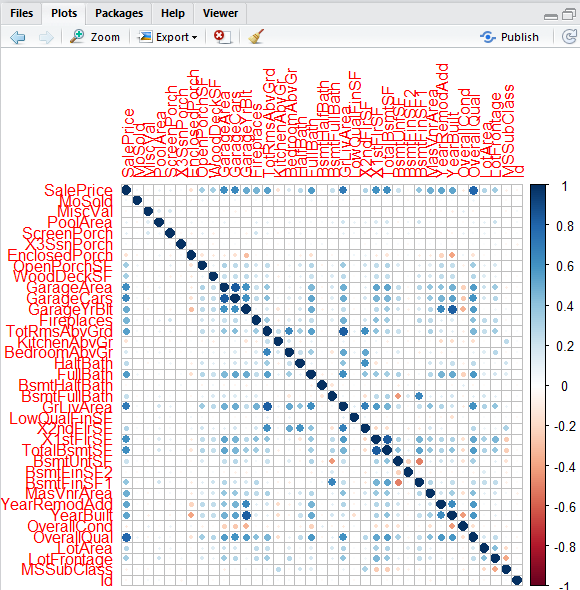
#test不能删，可能只能补全

#特征选择

correlations <- cor(trainNew[,-2])

library("corrplot")

corrplot(correlations)#绘制相关关系矩阵



corMatrix <- correlations[1,which(correlations[1,] > 0.5)]

#找出相关关系大于0.5

corMatrix[order(-corMatrix)[2:5]]

mode <- lm(SalePrice ~ OverallQual + GrLivArea + GarageCars + GarageArea,data = trainNew)

pre <- predict(mode,test)#测试集中的salesprice

res <- cbind(Id = test$Id,SalePrice = pre)

write.csv(res,file = "./res.csv")

结果csv：

