Data Analysis

Tuo

4/5/2022

# Set up

# 载入包裹  
library(tidyverse)

## Warning in file(con, "r"): cannot open file '/var/db/timezone/zoneinfo/  
## +VERSION': No such file or directory

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.4 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 2.0.1 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readxl)

# Load data

# 读取数据  
price <- read\_excel("price.xlsx")  
View(price)

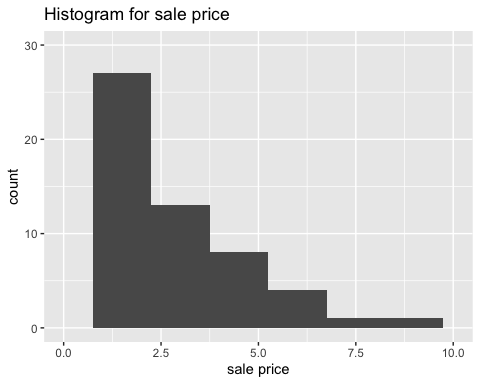
# Data Transformation

# 更改数据列名称  
colnames(price) <- c("names","price","category","sales","warehouse")  
price$category <- as.character(price$category)

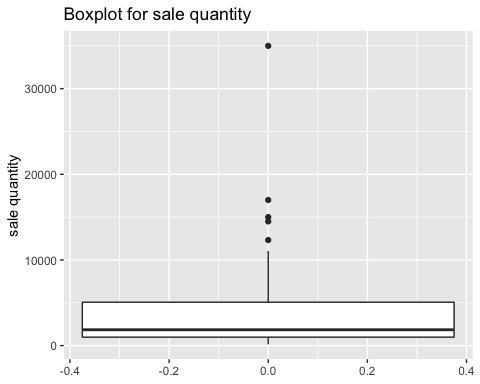
# Plots for one variable

# 可以从图看出，大部分销售价格在0-3.5之间， 销售价格在7.5以上的产品比较罕见  
ggplot(data = price, aes(x=price)) +   
 geom\_histogram(bins = 25, binwidth = 1.5)+  
 ylim(0,30)+  
 xlim(0,10)+   
 labs(x = "sale price",  
 y = "count",  
 title = "Histogram for sale price",  
 )

## Warning: Removed 2 rows containing missing values (geom\_bar).



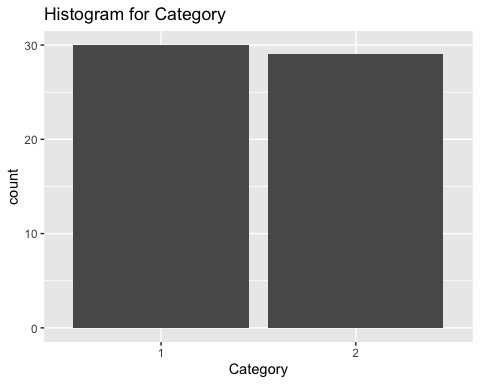
# 可以从图看出，大部分销售数量在5000以下，只有5个商贩的数量在10000以上。其中有一个商贩的销售数量在30000之上。  
ggplot(data = price, aes(y=sales)) +   
 geom\_boxplot()+  
 labs(y = "sale quantity",  
 title = "Boxplot for sale quantity",  
 )



#大部分货品是从发货地（东）发出的  
ggplot(data = price, aes(x=warehouse)) +   
 geom\_bar()+  
 labs(x = "Warehouse Location(East and West)",  
 y = "count",  
 title = "Histogram for Warehouse Location",  
 )

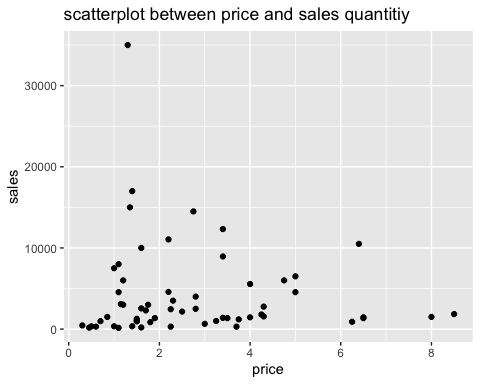


# 对于商品的分类1和2 占比相当，趋于相同  
ggplot(data = price, aes(x=category)) +   
 geom\_bar()+  
 labs(x = "Category",  
 y = "count",  
 title = "Histogram for Category",  
 )



# Plots for two variables

# 可以从散点出看出，大部分商品的销售价格在4以下，销售数量在10000以下。  
ggplot(data = price, aes(x=price, y=sales)) +   
 geom\_point()+  
 labs(title = 'scatterplot between price and sales quantitiy')



# OLS model

model1 <- lm(sales ~ price, data = price)  
summary(model1)

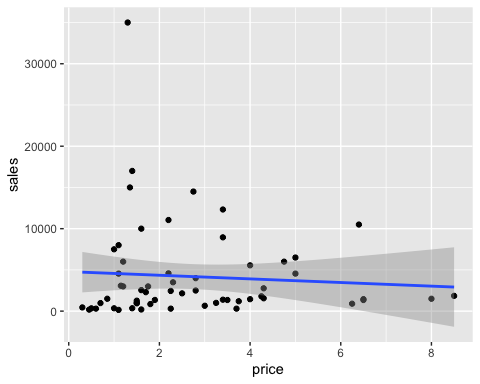
##   
## Call:  
## lm(formula = sales ~ price, data = price)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4511 -3328 -1911 1167 30496   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4789.7 1322.0 3.623 0.000621 \*\*\*  
## price -219.8 396.5 -0.554 0.581556   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5808 on 57 degrees of freedom  
## Multiple R-squared: 0.005361, Adjusted R-squared: -0.01209   
## F-statistic: 0.3072 on 1 and 57 DF, p-value: 0.5816

confint(model1)

## 2.5 % 97.5 %  
## (Intercept) 2142.453 7436.9209  
## price -1013.789 574.2283

ggplot(data=price, aes(x=price, y=sales))+  
 geom\_point()+  
 geom\_smooth(method=lm)

## `geom\_smooth()` using formula 'y ~ x'



# 模型结果显示p值=0.5816，模型的统计意义不显著（需要小于0.05可以认为显著）。销售价格和销售数量没有明显的回归关系  
# 虽然结果不显著，仍可以将这个模型概括为 销售数量 = 4789.7 - 219.8 \* 销售价格（只有参考意义）  
# 销售价格的增高会导致销售数量的减少  
# 建议将异常值（销售量大于10000的观察点）去除，再做后续分析

# OLS model 2

price\_1 <- price %>% filter(sales <10000)   
model2 <- lm(log(sales) ~ log(price), data = price\_1)  
summary(model2)

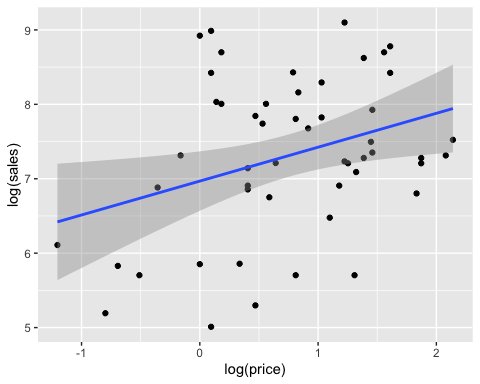
##   
## Call:  
## lm(formula = log(sales) ~ log(price), data = price\_1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.0014 -0.6001 -0.1321 0.7977 1.9751   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.9687 0.1989 35.044 <2e-16 \*\*\*  
## log(price) 0.4554 0.1852 2.459 0.0175 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.012 on 49 degrees of freedom  
## Multiple R-squared: 0.1098, Adjusted R-squared: 0.09166   
## F-statistic: 6.046 on 1 and 49 DF, p-value: 0.01752

confint(model2)

## 2.5 % 97.5 %  
## (Intercept) 6.56904330 7.368262  
## log(price) 0.08321189 0.827686

ggplot(data=price\_1, aes(x=log(price), y=log(sales)))+  
 geom\_point()+  
 geom\_smooth(method=lm)

## `geom\_smooth()` using formula 'y ~ x'



# 将销售数量高于10000的数据切除，  
# 再将销售价格和销售数量对数函数化，用两者来做回归模型  
# 模型结果显示p值=0.01752，模型的统计意义显著。log（销售价格）和log（销售数量）有明显的回归关系  
# 模型结果统计意义显著，但R平方0.1098表示模型效果欠佳  
# 可以将这个模型概括为 log销售数量 = 6.9687 - 0.4554 \* log销售价格  
# log销售价格的增高会导致log销售数量的增加  
# 建议尝试其他模型做后续分析