多元回归-第一次作业

刘若水 2022312123 信管22

2024-10-08

一、读取文件

library(readxl)  
data<-read\_excel("D:\\mycodelife\\workshop\\course\_task\\Bicycle Inventory.xlsx")

二、数据概述

library(psych)

## Warning: 程辑包'psych'是用R版本4.3.3 来建造的

describe(data)

## vars n mean sd median trimmed mad min max  
## ID 1 24 12.50 7.07 12.50 12.50 8.90 1.00 24.00  
## Product Category\* 2 24 3.25 1.42 3.00 3.30 1.48 1.00 5.00  
## Product Name\* 3 24 12.50 7.07 12.50 12.50 8.90 1.00 24.00  
## Purchase Cost 4 24 232.03 160.50 194.22 227.87 198.30 15.00 490.50  
## Selling Price 5 24 320.04 216.37 273.24 315.00 276.14 25.50 649.95  
## Supplier\* 6 24 3.08 1.35 3.00 3.10 1.48 1.00 5.00  
## Quantity on Hand 7 24 4.42 2.81 4.50 4.35 2.97 0.00 10.00  
## inventory value 8 24 839.72 728.27 482.72 794.55 552.25 0.00 2254.75  
## Percentage 9 24 0.04 0.04 0.02 0.04 0.03 0.00 0.11  
## Cumulative 10 24 0.75 0.27 0.86 0.79 0.20 0.11 1.00  
## range skew kurtosis se  
## ID 23.00 0.00 -1.35 1.44  
## Product Category\* 4.00 -0.25 -1.29 0.29  
## Product Name\* 23.00 0.00 -1.35 1.44  
## Purchase Cost 475.50 0.22 -1.66 32.76  
## Selling Price 624.45 0.21 -1.66 44.17  
## Supplier\* 4.00 -0.14 -1.25 0.28  
## Quantity on Hand 10.00 0.16 -0.96 0.57  
## inventory value 2254.75 0.59 -1.21 148.66  
## Percentage 0.11 0.59 -1.21 0.01  
## Cumulative 0.89 -0.99 -0.31 0.06

注意到，根据表中数据关系，，而selling price没有与其他数据关联，cumulative作为percentage的逐行累加。

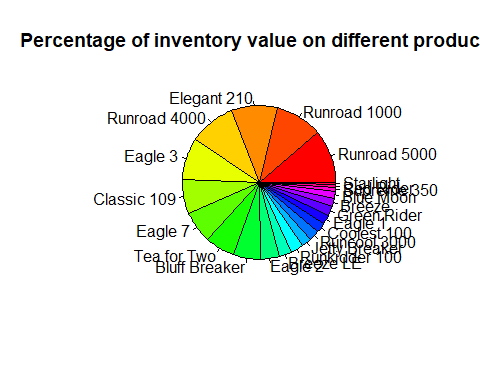
三、描述性统计 （一）对数据进行简单描述 1.绘制扇形图 由于数据中有零值，因此需要处理后才能绘制饼状图

non\_zero\_index <- data$`inventory value` > 0  
inventory\_value\_filtered <- data$`inventory value`[non\_zero\_index]  
inventory\_value\_filtered

## [1] 2254.75 2007.60 1970.64 1954.75 1752.60 1452.43 1358.01 1287.06 1125.00  
## [10] 802.22 549.75 509.50 455.95 429.75 419.94 410.01 381.88 359.80  
## [19] 301.16 150.00 120.00 100.47

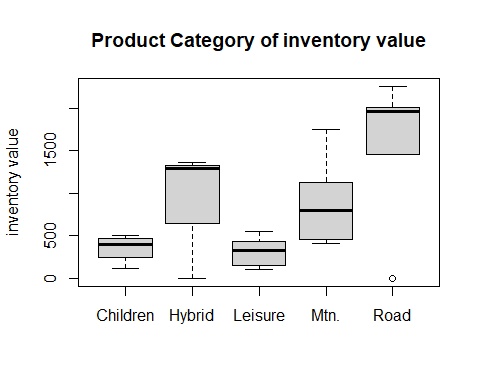
获取非零元素后进行绘制：

pie(inventory\_value\_filtered, labels = data$`Product Name`, main = "Percentage of inventory value on different products", col = rainbow(length(inventory\_value\_filtered)))

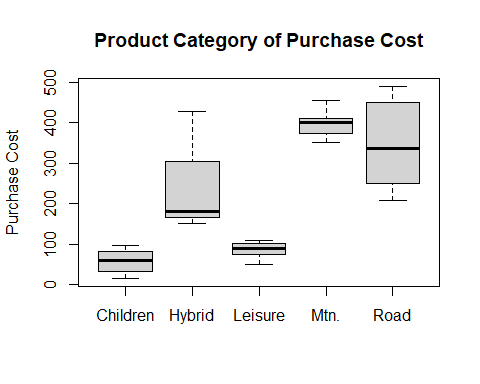


（二）按照product category和supplier进行划分

boxplot(data$'inventory value'~data$'Product Category',   
 main = "Product Category of inventory value",   
 xlab ="",  
 ylab = "inventory value")

 该boxplot反映了不同product category下inventory value的取值

boxplot(data$'Purchase Cost'~data$'Product Category',   
 main = "Product Category of Purchase Cost",   
 xlab ="",  
 ylab = "Purchase Cost")

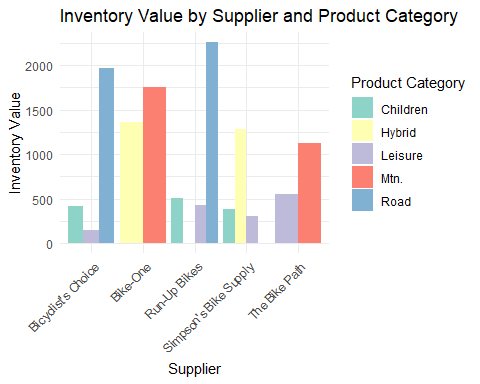
 该boxplot反映了不同product category下purchase cost的取值

library(ggplot2)

##   
## 载入程辑包：'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

ggplot(data, aes(x = Supplier, y = `inventory value`, fill = `Product Category`)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Inventory Value by Supplier and Product Category",  
 x = "Supplier",  
 y = "Inventory Value") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 scale\_fill\_brewer(palette = "Set3")

 该图反映了不同product category,supplier下的inventory value。

四、回归分析

1.研究成本是否和售价高低存在关联

model\_multi <- lm(data$`Selling Price` ~ data$`Purchase Cost`)  
  
summary(model\_multi)

##   
## Call:  
## lm(formula = data$`Selling Price` ~ data$`Purchase Cost`)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.755 -4.430 -0.003 8.493 28.796   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.01597 5.64885 1.419 0.17   
## data$`Purchase Cost` 1.34475 0.02016 66.711 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.52 on 22 degrees of freedom  
## Multiple R-squared: 0.9951, Adjusted R-squared: 0.9949   
## F-statistic: 4450 on 1 and 22 DF, p-value: < 2.2e-16

回归表示： 残差（Residuals）： Min：-37.755，1Q（第一四分位数）：-4.430，Median（中位数）：-0.003，3Q（第三四分位数）：8.493，Max：28.796。 这些残差值代表实际售价和模型预测值之间的差异，范围从大约 -37.755 到 28.796。理想情况下，残差应尽可能接近零，表明模型拟合效果较好。 系数（Coefficients）：

(Intercept) 截距：8.01597，这意味着当购买成本为0时，模型预测的销售价格为8.02（这个值在实际情境下可能没有太大实际意义）。 Purchase Cost 系数：1.34475，表示购买成本每增加1个单位，销售价格预计增加约1.34个单位。 标准误差 (Std. Error)：表示系数的不确定性。 截距的标准误差：5.64885。 购买成本系数的标准误差：0.02016，较小的标准误差表明这个估计比较精确。 t 值：t值为66.711，表示购买成本系数的估计值远离零，表明其对售价的影响非常显著。 p 值 (Pr(>|t|))：购买成本的p值小于2e-16，表明它在统计上非常显著，具有高度相关性。 模型的整体拟合优度：

残差标准误差 (Residual standard error)：15.52，这表示模型的残差有15.52个单位的标准误差，残差越小模型拟合越好。 Multiple R-squared：0.9951，表明购买成本解释了99.51%的销售价格的变异，模型拟合度非常好。 Adjusted R-squared：0.9949，与Multiple R-squared非常接近，这表明模型的拟合度依然较高，即使考虑到自变量的个数。 F统计量 (F-statistic)：4450，且p值 < 2.2e-16，这表明整个模型非常显著。 总结： 这个模型表明，购买成本对销售价格有非常强的正向影响，且这种关系具有高度的统计显著性。 购买成本可以很好地解释销售价格的变化，R平方值接近1，表示模型的拟合度非常高。 由于残差较小，且p值极小，可以认为模型在解释数据时表现得非常好。

2.进行分组回归，不同产品的售价和价格的关系 首先将非数值型变量转化为因子

data$supplier <- as.factor(data$Supplier)  
data$category <- as.factor(data$`Product Category`)

构建回归模型

model <- lm(data$`Selling Price` ~ data$`Purchase Cost` + supplier + category, data = data)  
  
# 查看回归结果  
summary(model)

##   
## Call:  
## lm(formula = data$`Selling Price` ~ data$`Purchase Cost` + supplier +   
## category, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.613 -7.183 1.058 7.559 19.951   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.15665 10.45404 1.163 0.264   
## data$`Purchase Cost` 1.32657 0.05318 24.944 5.29e-13 \*\*\*  
## supplierBike-One -12.42067 15.79818 -0.786 0.445   
## supplierRun-Up Bikes -6.70717 10.49318 -0.639 0.533   
## supplierSimpson's Bike Supply -7.88595 12.09602 -0.652 0.525   
## supplierThe Bike Path -13.14548 14.21130 -0.925 0.371   
## categoryHybrid 19.21894 17.09894 1.124 0.280   
## categoryLeisure 5.01694 10.83773 0.463 0.651   
## categoryMtn. 26.13637 25.87954 1.010 0.330   
## categoryRoad -4.46352 19.63869 -0.227 0.823   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.46 on 14 degrees of freedom  
## Multiple R-squared: 0.9969, Adjusted R-squared: 0.9949   
## F-statistic: 498.9 on 9 and 14 DF, p-value: 5.636e-16

系数解读： (Intercept)：截距（12.15665）表示当成本和所有因子变量的值为零时，售价的估计值。然而，因子变量（如供应商和类别）不可能为零，它们表示的是相对于基准类别的相对影响。

data$Purchase Cost（成本的回归系数）：估计值为1.32657，表示当成本每增加 1 个单位，售价预计增加约 1.33 个单位，并且这个变量显著（p值为 5.29e-13，远小于 0.05）。这意味着成本对售价有强烈的正向影响。

supplier 和 category 的系数：这些系数是相对于基准类别（默认第一个类别）的影响。我们可以看到：

supplier 的 p 值：所有供应商的 p 值都大于 0.05（如 Bike-One, Run-Up Bikes 等），意味着供应商对售价的影响不显著。 category 的 p 值：所有产品类别的 p 值也都大于 0.05，意味着产品类别对售价的影响不显著。 2. 统计显著性： 星号标记（*）****：通常用于标记显著性水平。 data$Purchase Cost 的 p 值非常小（< 0.001），有三个星号（***），说明成本对售价有显著的影响。 其他变量（如供应商和类别）没有显著性标记，说明这些控制变量没有显著影响。 3. R-squared 值： Multiple R-squared: 0.9969 和 Adjusted R-squared: 0.9949：表示模型能够很好地解释数据的变异，99.69% 的售价波动可以由成本、供应商和类别共同解释。这是非常高的 R-squared 值，说明你的模型拟合度很好。 4. F-statistic： F-statistic 值为 498.9，p 值非常小（5.636e-16），表明整个模型在统计上是显著的，至少有一个解释变量对因变量（售价）有显著影响。 结论： 成本对售价有显著的正向影响，随着成本增加，售价也会增加。 供应商和产品类别在这个模型中没有显著影响（p 值大于 0.05）。 模型的解释力很高（R-squared 值接近 1），说明你选取的变量对售价有很强的解释力。