870 Advanced Marketing Analytics

EXAM
on R workshop

Assigned: April 1, 2020

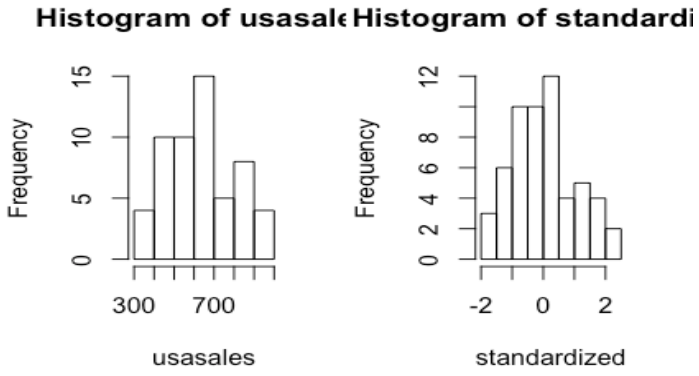
Due: April 10, 2020

Instructor: Varsha S. Kulkarni

Instructions : This is an exam is based on the R workshop lectures in Advanced Marketing Analytics. The aim is to test your knowledge of R for using it in marketing analytics. There are 6 questions for a total of 100 points. You are expected to use R for all the questions and show your codes *and* your output unless specified otherwise. Please submit on time, late submissions will not be graded. You may not discuss with anyone else and the work you turn in must be your own.

1. **[3x4=12 points]** Refer to the regional sales dataset. Do these

   (i) Show the histogram of USA sales. Standardize USA column of sales. Draw the histogram again. Do you see any difference? If so point out.




```
sales= regional sales data
usasales=sales$USA
par(mfrow=c(1,2))
hist(usasales)
standardized=(usasales-mean(usasales))/sd(usasales)
hist(standardized)
The range becomes lower after standardization and the
distribution looks more normal.
```

(ii) Standardize all the 4 columns of the 4 regions, leaving out USA column. (You may only show your code for this and not the output).

```
ne=sales$Northeast
mw=sales$Midwest
s=sales$South
w=sales$West
ne=(ne-mean(ne))/sd(ne)
mw=(mw-mean(mw))/sd(mw)
s=(s-mean(s))/sd(s)
w=(w-mean(w))/sd(w)
```

(iii) Compute the pairwise correlations of regions, e.g. cor(region1, region2),… Which pair is the most positively correlated and which pair is most negatively correlated? Interpret this.
The correlation between south and west regions is highest and positive. It means that the sales in both these regions go up and down simultaneously quite often. There is no pair of regions that is negatively correlated.

```
cor(s,w)
[1] 0.8891592
```

(iv) Give the (count) distribution of USA sales using the parameter "cut_width" as equal to 100.

```
sales %>%
  count(cut_width(USA,100))
```

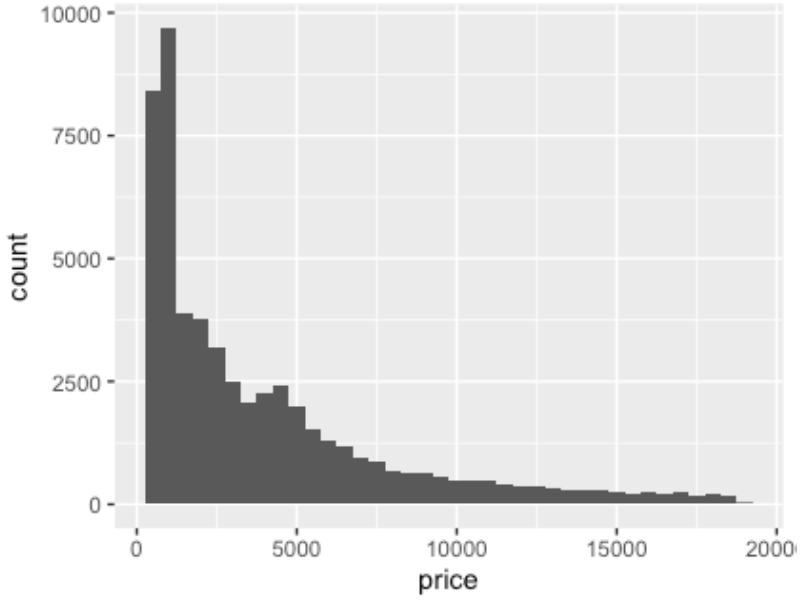
```
# A tibble: 8 x 2
  `cut_width(USA, 100)`     n
  <fct>                 <int>
1 [250,350]                 2
2 (350,450]                 7
3 (450,550]                11
4 (550,650]                12
5 (650,750]                10
6 (750,850]                 5
7 (850,950]                 6
8 (950,1.05e+03]            3
```

2. **[3x8=24 points]** Refer to the dataset in tidyverse – diamonds. Do the following

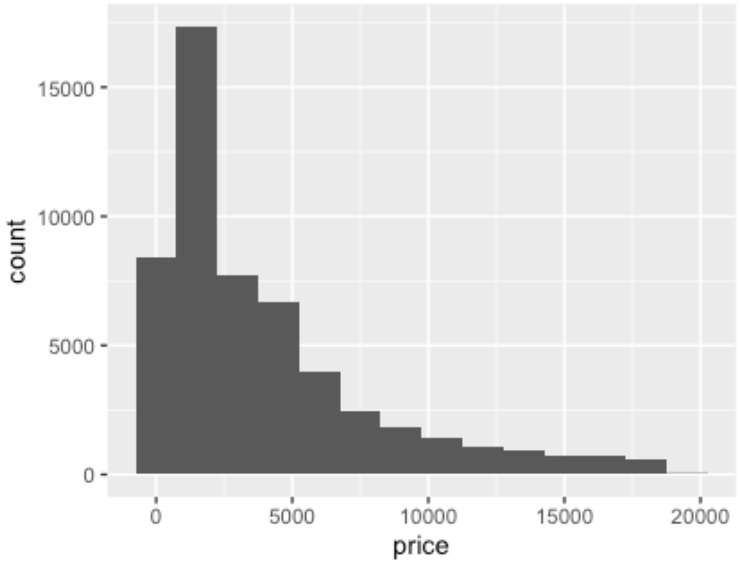
(i) Show histograms of price with "binwidth" of 500 and 1500. What happens as binwidth increases? Describe the skewness in 1-2 sentences.

```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = price), binwidth = 500)
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = price), binwidth = 1500)
```

Binwidth=500:



Binwidth=1500



The frequency distribution of price appears less skewed as bin width is increased, more observations to the right.

(ii) What are the mean, variance, median of price? Is the variance of price greater or lesser than the variance of carat? Which of the two has more dispersion- carat or price?

```
mean(diamonds$price)
[1] 3932.8
> var(diamonds$price)
[1] 15915629
> median(diamonds$price)
[1] 2401

var(diamonds$carat)
[1] 0.2246867
```

Carat shows lower dispersion than price, variance is less.

(iii) What is the average carat value whenever price is above its 33rd percentile value

```
p=diamonds$price
crt=diamonds$carat

a=quantile(p,33/100)
mean(crt[which(p>=a)])
[1] 1.017794
```

(iv) Rank the average prices for all the 5 cut types.

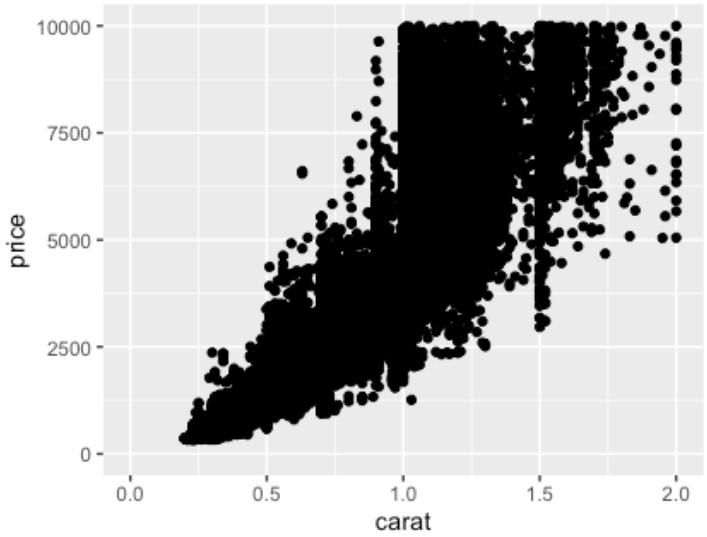
```
f=mean(p[which(diamonds$cut=="Fair")])
g=mean(p[which(diamonds$cut=="Good")])
vg=mean(p[which(diamonds$cut=="Very Good")])
prm=mean(p[which(diamonds$cut=="Premium")])
i=mean(p[which(diamonds$cut=="Ideal")])

> rank(c(f,g,vg,prm,i))
[1] 4 2 3 5 1
```

(v) Refer to a carat versus price plot as given in lecture R script. Now show a different plot by zooming into the carat-price plot. In the lecture R script, we learnt how to plot geom_point using ggplot for carat versus price. Extend that by adding + sign and in the next line put a limit on y coordinate and then a limit of x coordinate like this:

```
ggplot(data = diamonds) +
  geom_point(..) +
    ylim(0,..)+          #give an upper limit of y after 0
    xlim(0,..)          # give an upper limit of x after 0
```

You may choose any upper limits of y and x axis. Show the best possible zoom plot in your opinion.

```
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price))+
  ylim(0,10000)+
  xlim(0,2)
```



(vi) Filter a subset of diamonds of unusual or highly influential observations. Construct a subset named "group" to filter all the price values with a range so that they are all either less than a *or* greater than b. We define
a = average of price – 5 times the standard deviation of price
b= average of price+3.7 times standard deviation of price.
Show the first 10 rows of group.

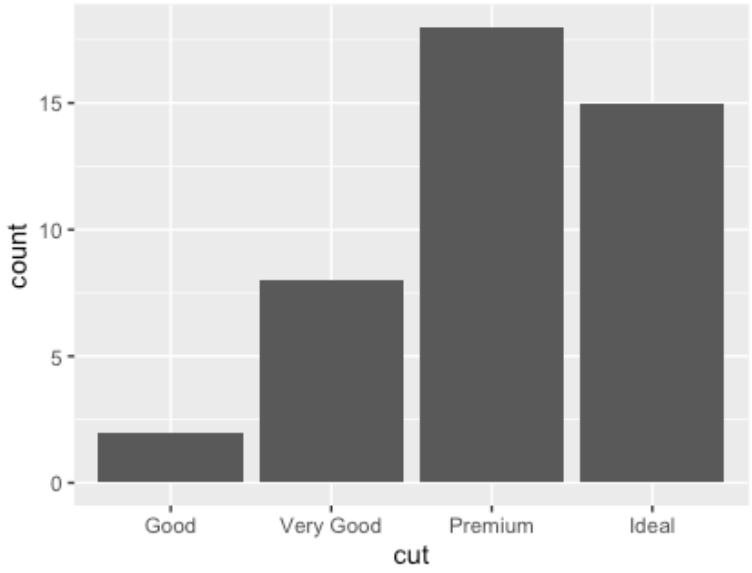
```
group <- diamonds %>%
  filter(price < mean(price)-5*sd(price) | price >
mean(price)+3.7*sd(price))
```

```
> group
# A tibble: 43 x 10
   carat cut       color clarity depth table price     x     y     z
   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
 1  1.28 Ideal     E     IF       60.7  57   18700  7.09  6.99  4.27
 2  2.02 Ideal     G     VS2      62    57   18700  8.1   8.05  5.01
 3  3.51 Premium   J     VS2      62.5  59   18701  9.66  9.63  6.03
 4  2.01 Premium   G     SI2      61.2  57.2 18705  8.08  8.14  4.97
 5  2.22 Premium   J     VS1      60    60   18706  8.49  8.43  5.08
 6  2.07 Good      I     VS2      61.8  61   18707  8.12  8.16  5.03
```

```
 7  2    Very Good E     SI1      60.5  59   18709  8.09  8.14  4.94
 8  3.01 Premium   J     SI2      60.7  59   18710  9.35  9.22  5.64
 9  3.01 Premium   J     SI2      59.7  58   18710  9.41  9.32  5.59
10  2.18 Premium   F     SI1      61.2  60   18717  8.38  8.3   5.1
```

(vii) Show the bar plot of "cut" in the group dataset created above.

```
ggplot(data = group) +
  geom_bar(mapping = aes(x = cut))
```



(viii) For the group dataset, rank the average price values for all the 5 cut types. Observe any difference in the ranking from your result in (iv) in 2-3 sentences.

```
f=mean(p[which(group$cut=="Fair")])
g=mean(p[which(group$cut=="Good")])
vg=mean(p[which(group$cut=="Very Good")])
prm=mean(p[which(group$cut=="Premium")])
i=mean(p[which(group$cut=="Ideal")])

> rank(c(f,g,vg,prm,i))
[1] 5 4 3 1 2
```

There is a huge difference between the rankings obtained in group versus those in diamonds. This is because group dataset consists of the very high price ranges (a<0, so group contains only high prices). The "good" and "ideal" cut types are found more in this range.
The cut type "fair" does not exist in group, premium was highest ranked in diamonds and becomes lowest in group in terms of average price.

The average price rank of "good" increases from 2 to 4. There is a slight decrease in the average price rank of "ideal" whereas "very good" remains same.

3. **[5x3=15 points]** Answer these questions by estimating linear regressions

(i) Estimate the linear model for price ~ carat in the diamonds and group datasets. Point out the difference in interpretation of slope coefficients in the two datasets. Which coefficient or slope is more significant if at all?
The coefficient estimate of carat is positive high and significant for diamonds dataset whereas that of group dataset is smaller, negative and not significant.

```
linmod=lm(price~carat,data=diamonds)
> summary(linmod)
```
Estimate of slope= 7756.43, p-value < 0.05

```
linmod=lm(price~carat,data=group)
> summary(linmod)
```
Estimate of slope= -14.42, p-value > 0.05

(ii) Fill in the blanks: A unit increase in **carat** value in diamonds data set _*increases*_(increases/decreases) the **price** by _*7756.43*_units.

(iii) In airbnb2 dataset, consider the first 50 observations (1:50) of the variables in bold and fill in the blanks. A one unit increase in **accommodates per bedroom** _*increases*_ (increases/decreases) the **price** by _*965.3*_% and _*decreases*_ (increases/decreases) the **overall satisfaction** by _*14.26*_%.

```
a=airbnb2$accommodates
b=airbnb2$bedrooms
ab=a/b
p=airbnb2$price
linmod=lm(p[1:50]~ab[1:50])
summary(linmod)
Call:
lm(formula = p[1:50] ~ ab[1:50])

Residuals:
     Min       1Q   Median       3Q      Max
-132.481  -43.108    5.825   37.478  194.825

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   94.215     14.683   6.416 5.78e-08 ***
ab[1:50]       9.653      4.249   2.272   0.0276 *
---
```

4. **[5x5= 25 points]** The airbnb2 dataset lists the neighborhood ids of places, the prices, number of accommodates and bedrooms, customer satisfaction and number of reviews. Answer these

```
os=airbnb2$overall_satisfaction
br=airbnb2$bedrooms
r=airbnb2$reviews
```

   (i) the predicted price for 0 reviews is =__**78.65**__

```
coef(lm(airbnb2$price ~ r))
```

   (ii) Predicted reviews for 3 bedrooms is = _**4.74**___

```
coef(lm(r ~ br))
7.5346514  -0.9286752*3
```

Average reviews for 3 bedrooms is = _**4.28**_

```
mean(r[which(br==3)])
```

   (iii) Average accommodates for 4 or 5 bedrooms is _**7**_

```
mean(a[which((br==4)|(br==5))])
```

   (iv) Range of overall satisfaction of 2 bedrooms with atleast 10 reviews is _**1**_

```
max(os[which((br==2)&(r>=10))])-
min(os[which((br==2)&(r>=10))])
```

   (v) Give the count distribution of neighborhoods. Identify if there is a neighborhood that is an outlier and give your reason.

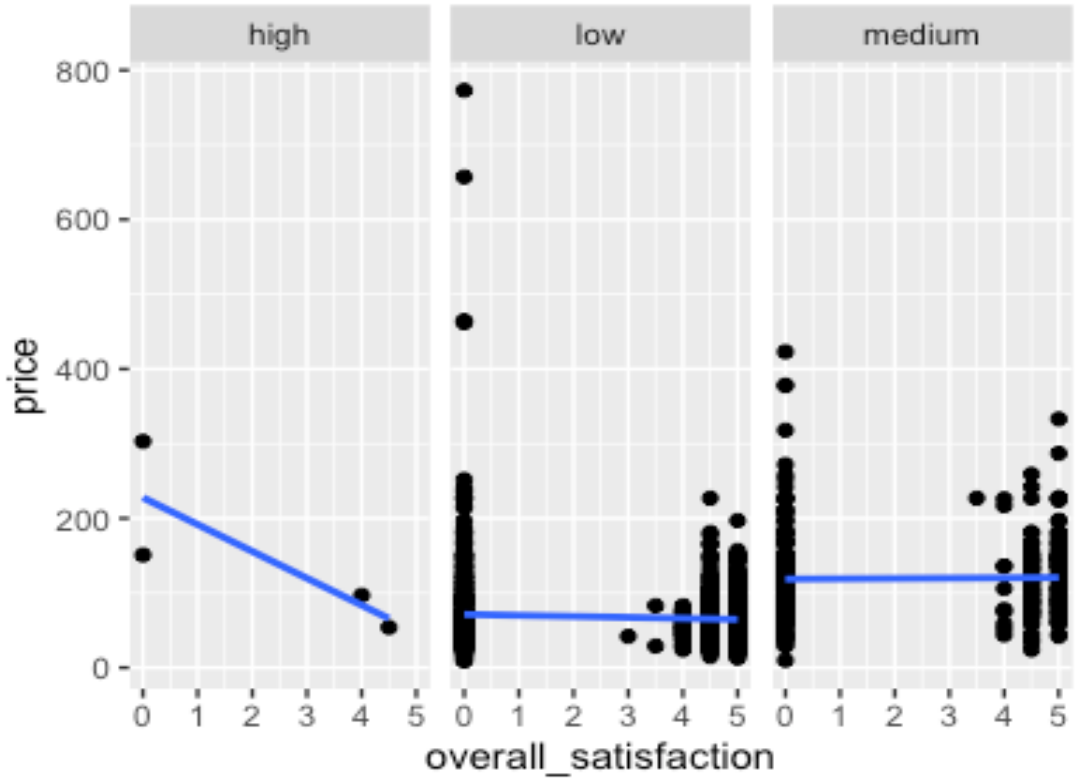
```
airbnb2 %>%
  count(neighborhood)
```

Neighborhood 24 has very high number of observations and is an outlier as are neighborhoods 23 and 12 due to having very less frequency (=1).

5. **[9 points]** Do a simple mutation of airbnb2 dataset by adding another column. This is column is a transformation of "accommodates" column into 3 categories: low, medium, high. Name this column as "class". Construct a vector class by labeling all those entries as "low" whenever accommodates values are less than 5, as "medium" if the values are between 5 and 10, "high" if accommodates values exceed 10. [Use if(condition on accommodates) then class=..]. Add this class column to the airbnb2 dataset. Show the ggplot of overall satisfaction and price with geom_point(), and facet_wrap(~class), and a regression trend line.

```
a=airbnb2$accommodates
class=NULL
for(i in 1:length(a)){
  if(a[i]<5){
    class[i]="low";}
  if((a[i]>=5)&(a[i]<=10)){
    class[i]="medium"}
  if(a[i]>10){
    class[i]="high"}
}
#Alternative method for generating class:
class=NULL
class[which(a<5)]="low"
class[which((a>=5)&(a<=10))]="medium"
class[which(a>10)]="high"

#Add to airbnb2 data set

airbnb2=data.frame(airbnb2,class)



ggplot(data = airbnb2, mapping = aes(x = overall_satisfaction, y =
price)) +
  geom_point() +
  facet_wrap(~class)+
  stat_smooth(method = "lm", se=FALSE)
```

It is clear from the plot that price and overall_satisfaction show no dependence for low and medium number of accommodates but for the high accommodates category, price decreases as overall_satisfaction increases.



6. **[15 points]** Do you think the customers of airbnb2 dataset are consciously satisfied? If so, how much? {To answer this, consider comparing the nature of estimates of linear models of y=consumer overall satisfaction with (i) number of reviews and (ii) price. Interpret the estimations}.

The idea of conscious consumption is to increase the awareness of customers before they buy products so that they make wise and informed decisions about their purchase. This is to avoid waste and to promote sustainable marketing. In the Airbnb example, the reviews are a source of information to the customers about the listing, but some customers would go for other factors such as pricing. If they are conscious, however, they would consider reviews more seriously than price to make a decision to visit and their level of satisfaction would be determined more significantly by that. We assess this by running two regressions to study the impact of two factors- reviews or price

1. Overall satisafction~reviews

```
summary(lm(airbnb2$overall_satisfaction~airbnb2$reviews))
```

```
Call:
lm(formula = airbnb2$overall_satisfaction ~ airbnb2$reviews)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.481  -1.488  -1.488   2.225   3.175

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.487974   0.046382   32.08   <2e-16 ***
airbnb2$reviews 0.112441   0.003599   31.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.011 on 2435 degrees of freedom
Multiple R-squared:  0.2862,     Adjusted R-squared:  0.2859
F-statistic: 976.3 on 1 and 2435 DF,  p-value: < 2.2e-16
```

2. Overall_satisfaction~price

```
summary(lm(airbnb2$overall_satisfaction~airbnb2$price))

Call:
lm(formula = airbnb2$overall_satisfaction ~ airbnb2$price)

Residuals:
   Min     1Q Median     3Q    Max
-2.370 -2.211 -1.978  2.686  3.536

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.3984870  0.0913234  26.264  < 2e-16 ***
airbnb2$price -0.0028054  0.0009999  -2.806  0.00506 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.377 on 2435 degrees of freedom
Multiple R-squared:  0.003222,   Adjusted R-squared:  0.002813
F-statistic: 7.872 on 1 and 2435 DF,  p-value: 0.005061
```

Although both price and reviews are significant variables, the variation explained by reviews is more (Rsquared) and the effect of price is much lesser than that of reviews. One unit increase in price reduces the overall satisfaction by 0.28% whereas when reviews

increase by 1, the overall satisfaction goes up by 11.2%. Therefore they seem to be consciously satisfied.
We can also run a multiple regression model by considering reviews + price and then assess the effect of adding price to the model in (1) there.