You work for the **Tropicana** brand.  Your state sales manager is putting pressure on the sales team to bring sales numbers in line with her office's projections--and she is telling you that she sees much room for improvement in the area for which you are assigned as data analyst.  If your numbers aren't up by the end of the year, your job will be relocated to Alaska.

Here are the questions she needs you to answer:

1. What is the data quality?  What missing or implausible values did you encounter, and what did you do to correct them?  Explain the thinking steps and show the R code you used to determine this. (3 points)

Data quality means which data identifies that difference or relation between variables, and which help in making decision as per the organization requirement. Data provided of oj.csv was resolving with complete data, but how ever data cannot be trusted completely, so identified that data is almost identical at most the stores. There are 28752 records which belongs to Tropicana brand, in which the identical and implausible data is not identified, and by removing the null values using na.rm=TRUE , we are making the data summary, covariance and correlation between each value to get 95% data exactly matched with the value and to get the results.  
  
NA == NULL by executing this which resolves that there are no missing values and resolves with logical(0) , which resolves with that there no missing values in the table.

|  |
| --- |
| > summary(oj)  store brand week logmove feat price  Min. : 2.00 Min. : NA Min. : 40.0 Min. : 4.159 Min. :0.0000 Min. :0.520  1st Qu.: 53.00 1st Qu.: NA 1st Qu.: 70.0 1st Qu.: 8.490 1st Qu.:0.0000 1st Qu.:1.790  Median : 86.00 Median : NA Median :101.0 Median : 9.034 Median :0.0000 Median :2.170  Mean : 80.88 Mean :NaN Mean :100.5 Mean : 9.168 Mean :0.2373 Mean :2.282  3rd Qu.:111.00 3rd Qu.: NA 3rd Qu.:130.0 3rd Qu.: 9.765 3rd Qu.:0.0000 3rd Qu.:2.730  Max. :137.00 Max. : NA Max. :160.0 Max. :13.482 Max. :1.0000 Max. :3.870  NA's :28947  AGE60 EDUC ETHNIC INCOME HHLARGE  Min. :0.05805 Min. :0.04955 Min. :0.02425 Min. : 9.867 Min. :0.01351  1st Qu.:0.12210 1st Qu.:0.14598 1st Qu.:0.04191 1st Qu.:10.456 1st Qu.:0.09794  Median :0.17065 Median :0.22939 Median :0.07466 Median :10.635 Median :0.11122  Mean :0.17313 Mean :0.22522 Mean :0.15556 Mean :10.617 Mean :0.11560  3rd Qu.:0.21395 3rd Qu.:0.28439 3rd Qu.:0.18776 3rd Qu.:10.797 3rd Qu.:0.13517  Max. :0.30740 Max. :0.52836 Max. :0.99569 Max. :11.236 Max. :0.21635    WORKWOM HVAL150 SSTRDIST SSTRVOL CPDIST5  Min. :0.2445 Min. :0.002509 Min. : 0.1321 Min. :0.4000 Min. :0.7725  1st Qu.:0.3126 1st Qu.:0.123486 1st Qu.: 2.7670 1st Qu.:0.7273 1st Qu.:1.6262  Median :0.3556 Median :0.346154 Median : 4.6507 Median :1.1154 Median :1.9634  Mean :0.3592 Mean :0.343766 Mean : 5.0973 Mean :1.2073 Mean :2.1204  3rd Qu.:0.4023 3rd Qu.:0.528313 3rd Qu.: 6.6506 3rd Qu.:1.5385 3rd Qu.:2.5337  Max. :0.4723 Max. :0.916700 Max. :17.8560 Max. :2.5714 Max. :4.1079    CPWVOL5 brandC  Min. :0.09456 Min. : NA  1st Qu.:0.27167 1st Qu.: NA  Median :0.38323 Median : NA  Mean :0.43891 Mean :NaN  3rd Qu.:0.56024 3rd Qu.: NA  Max. :1.14337 Max. : NA  NA's :28947 |
|  |
| |  | | --- | | > | |

1. In what region and in what stores are your sales the weakest in comparison to your competitors?  What role does **the ethnic composition of store clientele** play?  Explain the thinking steps and show the R code you used to determine this.  Make sure to provide an answer to your question and recommendations for advertising strategies that your state sales manager will understand. (3 points)

Taking region and competitors, store numbers, the sale value is decreasing as on the range of the store is increasing, and for this west region is more than compared to other competitors. In this this come in interpretation that the value resolves with 0.995690759 because of which it resolved with straight line as mention in the below diagram, and resolves with positive value. The R Code is given here that value is having positive value, and having value of 1.04. By ideolizing with pivot table we Tropicana sale is more than compared other sale in other stores. Personal idea of advertisement to store manager that in south region the value is very less than compared to all region, so we can make the to increase promoting of the product so that we can increase sale in south region, and age of the people greater than 60 is having more and income with more than 150 is coming to to the store in remaining stors are coming to the requirement of the product so focusing on the value we are having more promoting irrespective of thinicity so that product demand increases.   
  
  
R Code:

tbl=table(oj$store,oj$ETHNIC)

tbl

chisq.test(tbl)

age = oj$store

tis=oj$SSTRVOL

gest=oj$ETHNIC

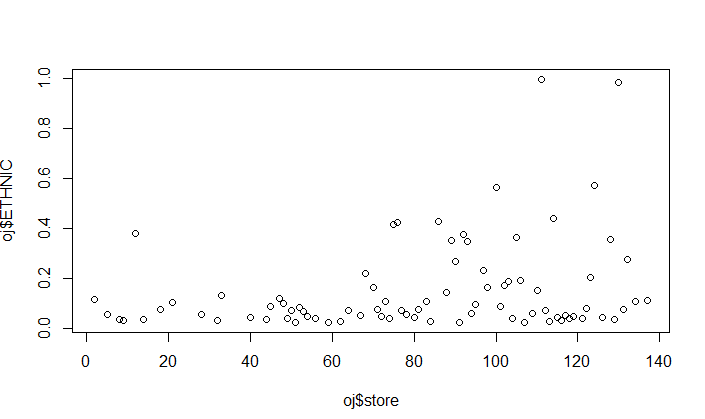
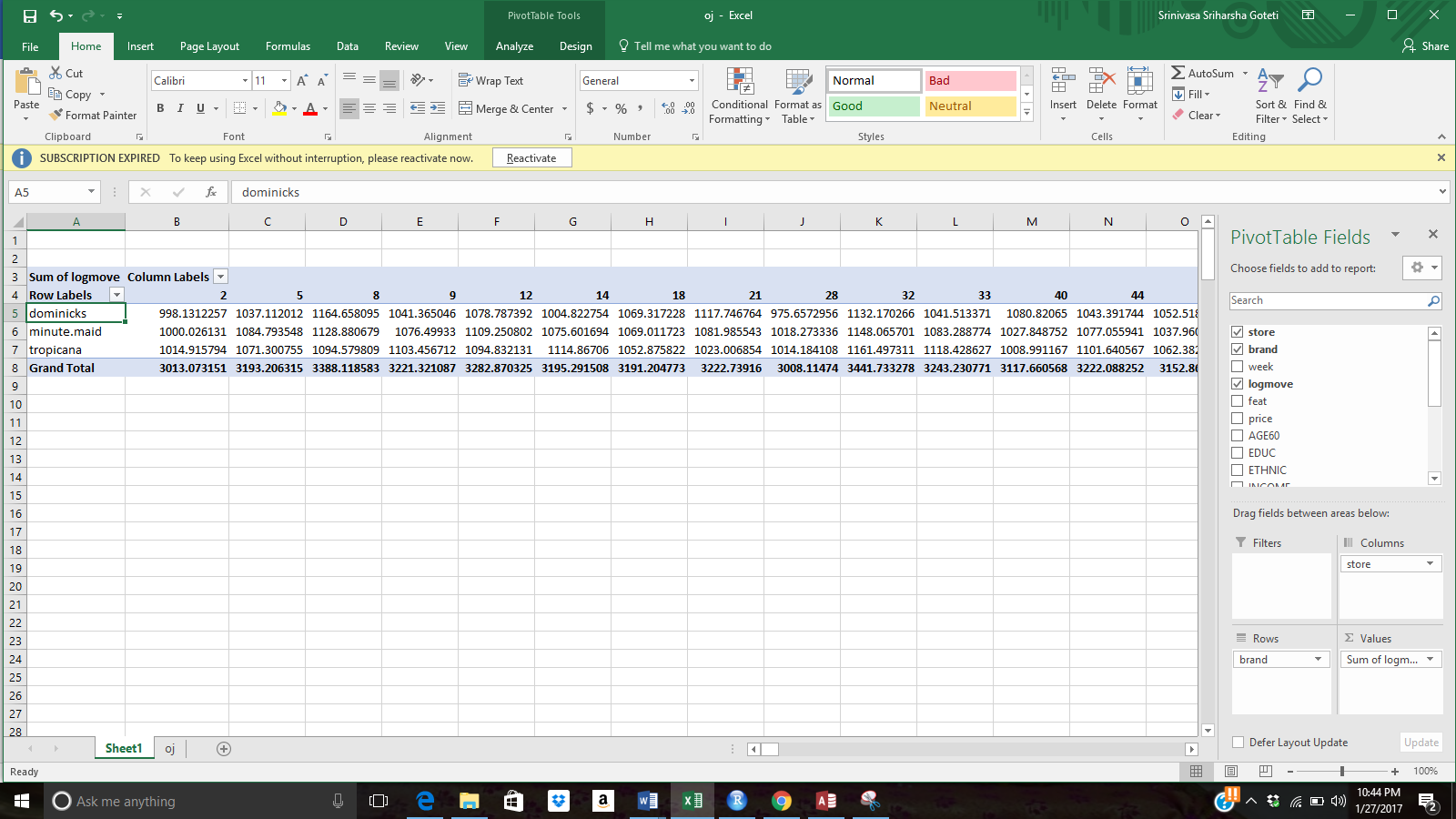
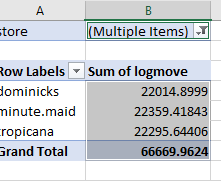
cov(tis,age)

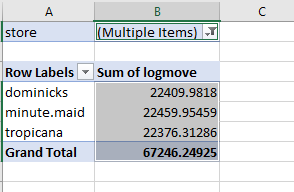
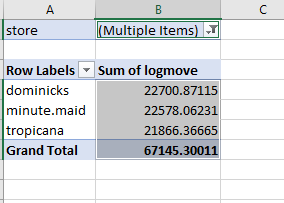
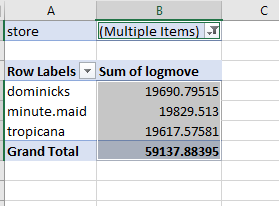
cov(gest,age)  
  
Output:  
  
> cov(tis,age)

[1] 1.041541

cov(gest,age)

[1] 1.857838

  
  
  
  
West region

  
North Region  
  
East Region  
  
South Region

1. In what household demographic (**median income, households over $150k**) do you see the largest potential for sales increase?   Explain the thinking steps and show the R code you used to determine this.  Make sure to provide an answer to your question and recommendations for advertising strategies that your state sales manager will understand. (4 points)

Please include screenshots of your R code and its outcomes.  Remember that the results of your work and the advertising recommendations must be interpreted in language that the state sales manager can understand.

R Code:  
tbl=table(oj$INCOME,oj$HVAL150)

tbl

chisq.test(tbl)

cov(oj$INCOME,oj$HVAL150)

cor(oj$INCOME,oj$HVAL150)

plot(oj$INCOME,oj$HVAL150)

summary(oj$INCOME,oj$HVAL150)  
  
Output :  
  
> tbl=table(oj$INCOME,oj$HVAL150)

> tbl

0.00250941 0.005144033 0.005700872 0.009842116 0.009843473 0.011117974 0.012746711

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 357 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 360

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 357 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.034301485 0.042876284 0.049271137 0.050756901 0.054227156 0.068989386 0.074418605

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.080565371 0.090007348 0.093351946 0.093860685 0.096330275 0.099963249 0.123485737

9.867082871 0 0 0 0 0 0 0

9.966650127 342 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 348 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.134417808 0.135732648 0.144424674 0.148870968 0.15625 0.163882428 0.18957346

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.21005386 0.218654434 0.228675136 0.232741256 0.240180587 0.26618705 0.280446271

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 357 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 357 0

10.1883657 0 345 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 360 0 0 0 0

0.292651757 0.303356282 0.308160505 0.316889001 0.335671343 0.342017119 0.346153846

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.352834113 0.354573672 0.367446809 0.376326075 0.378173027 0.386627907 0.407496492

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 345 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.416315588 0.431708991 0.441323063 0.445140033 0.455995496 0.459405941 0.463887065

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 363 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.475752773 0.478846154 0.491441083 0.491839763 0.496 0.505747126 0.528312901

9.867082871 0 0 0 0 354 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.533949633 0.535883355 0.538346427 0.548122392 0.56065207 0.571315163 0.573626374

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.578125 0.61086249 0.624113475 0.64268521 0.64424253 0.646023263 0.691439828

9.867082871 0 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0 0

0.727728684 0.750769908 0.86073903 0.862559242 0.878804348 0.916699488

9.867082871 0 0 0 0 0 0

9.966650127 0 0 0 0 0 0

9.996659083 0 0 0 0 0 0

10.0365751 0 0 0 0 0 0

10.08897077 0 0 0 0 0 0

10.13828296 0 0 0 0 0 0

10.1406129 0 0 0 0 0 0

10.15342882 0 0 0 0 0 0

10.1883657 0 0 0 0 0 0

10.25895677 0 0 0 0 0 0

10.30811898 0 0 0 0 0 0

10.33410045 0 0 0 0 0 0

[ reached getOption("max.print") -- omitted 71 rows ]

Pearson's Chi-squared test

data: tbl

X-squared = 2373700, df = 6724, p-value < 2.2e-16

Warning message:

In chisq.test(tbl) : Chi-squared approximation may be incorrect

> cov(oj$INCOME,oj$HVAL150)

[1] 0.04314715

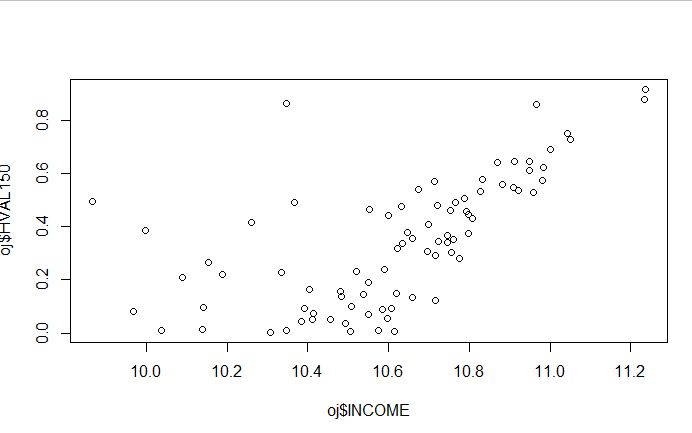
> cor(oj$INCOME,oj$HVAL150)

[1] 0.639397

> plot(oj$INCOME,oj$HVAL150)  
> summary(oj$INCOME,oj$HVAL150)

Min. 1st Qu. Median Mean 3rd Qu. Max.

9.867 10.460 10.640 10.620 10.800 11.240

  
  
  
  
  
  
By the above graph we identified that the value of income is ranging with 10.4 to 11.0 and HVAL 15 is 0.1and till 0.8 value. Summary resolves with median of 10.6 and mean also ranges between the same value. The graph is increasing with in between region, so we can have the store manager to increase promoting not only within range, but make the other region to be focused so that in all region, the values increases. AS in this specified range the sale is already high, so rather having a promotion in this space, we can Have new places to be increased so that sales in all region increases, and all stores would be opened , without closing the other stores, so asking to increase sales by planning good advertisement about the product .