A3_DSC424_ateboul

Alex Teboul

October 27, 2019

##DSC 324/424 ##Assignment 2 (DUE SUNDAY, October 6th by Midnight)

Problem 2

2) (20 points, Individual, to be turned in with the rest of the homework) Choose a technique that we have covered so far in this course, and try applying that technique to your data. You may choose any of: * a) Model building and Multiple Regression * b) PCA

*** c) CFA**

- d. CCA
- e. CA (correspondence analysis)
- Each member of your group should try a different technique, or the same technique with different aspects of the data.

Data Exploration and Setup

```
#Libraries (not all used at the moment)
library(Hmisc) #Describe Function
library(psych) #Multiple Functions for Statistics and Multivariate Analysis
library(GGally) #ggpairs Function
library(ggplot2) #ggplot2 Functions
library(vioplot) #Violin Plot Function
library(corrplot) #Plot Correlations
library(REdaS) #Bartlett's Test of Sphericity
library(psych) #PCA/FA functions
library(factoextra) #PCA Visualizations
library("FactoMineR") #PCA functions
library(ade4) #PCA Visualizations
library(varhandle)
library(tidyverse)
library("dplyr")
library(car)
#################################################################################################
```

```
#Read in Datasets
housing_data <- read.csv("train.csv")

#Check Sample Size and Number of Variables
dim(housing_data)
```

```
## [1] 1460   81
```

```
#Missing values
sum(is.na(housing_data))
```

```
## [1] 6965
```

```
colSums(is.na(housing_data))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0           259             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0          1369             0             0             0
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##             8             8             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##            37            37            38            37             0
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##            38             0             0             0             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             1             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0           690            81            81
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##            81             0             0            81            81
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0          1453          1179          1406
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

- So Columns that need missing data Treated are:
- LotFrontage, Alley, MasVnrType, MasVnrArea, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Electrical, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature
- Dropped Alley, FireplaceQu, PoolQC, Fence, MiscFeature Columns
- Then dropped the rest of the rows with NA.
- Also need to drop Id because it's just the index and Utilities because it has no variance.
- Also dropped BsmtFinSF1, BsmtFinSF2, BsmtUnfSF because we have TotalBsmtSF
- Also dropped LowQualFinSF to avoid aliasing.
- Also dropped GrLivArea to avoid multicollinearity as it is the sum of X1stFlrSF and X2ndFlrSF .

```
#Show for first 6 rows of data
head(housing_data)
```

```
#Missing Data Fixes
#1 Remove columns with many NAs
housing_data_fix <- subset(housing_data, select = -c(Id, Utilities, Alley, FireplaceQu, PoolQC, Fence, MiscFeature, BsmtFinS
F1, BsmtFinSF2, BsmtUnfSF, LowQualFinSF, X1stFlrSF, X2ndFlrSF))

#2 Remove rows with NA in them
housing_data_fix <- housing_data_fix[complete.cases(housing_data_fix), ]

#Check Sample Size and Number of Variables
dim(housing_data_fix)
```

```
## [1] 1094   68
```

```
#Missing values
sum(is.na(housing_data_fix))
```

```
## [1] 0
```

```
#Any missing?
colSums(is.na(housing_data_fix))
```

```
##    MSSubClass      MSZoning   LotFrontage       LotArea        Street
##             0             0             0             0             0
##      LotShape   LandContour     LotConfig     LandSlope  Neighborhood
##             0             0             0             0             0
##    Condition1    Condition2      BldgType    HouseStyle   OverallQual
##             0             0             0             0             0
##   OverallCond     YearBuilt  YearRemodAdd     RoofStyle      RoofMatl
##             0             0             0             0             0
##   Exterior1st   Exterior2nd    MasVnrType    MasVnrArea     ExterQual
##             0             0             0             0             0
##     ExterCond    Foundation      BsmtQual      BsmtCond  BsmtExposure
##             0             0             0             0             0
##  BsmtFinType1  BsmtFinType2   TotalBsmtSF       Heating     HeatingQC
##             0             0             0             0             0
##    CentralAir    Electrical     GrLivArea  BsmtFullBath  BsmtHalfBath
##             0             0             0             0             0
##      FullBath      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual
##             0             0             0             0             0
##  TotRmsAbvGrd    Functional    Fireplaces    GarageType   GarageYrBlt
##             0             0             0             0             0
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##             0             0             0             0             0
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea       MiscVal        MoSold        YrSold
##             0             0             0             0             0
##      SaleType SaleCondition     SalePrice
##             0             0             0
```

```
#Show for first 6 rows of data
#head(housing_data_fix)

#Column Names
#names(housing_data_fix)
```

- At this point many rows were removed. But we still have 1094 clean data points to use in our analysis.

```
#Describe the data
describe(housing_data_fix)

#Show Structure of Dataset
#str(housing_data_fix)
```

- Later analysis may require scaling/normalizing, but for now we will not change anything.

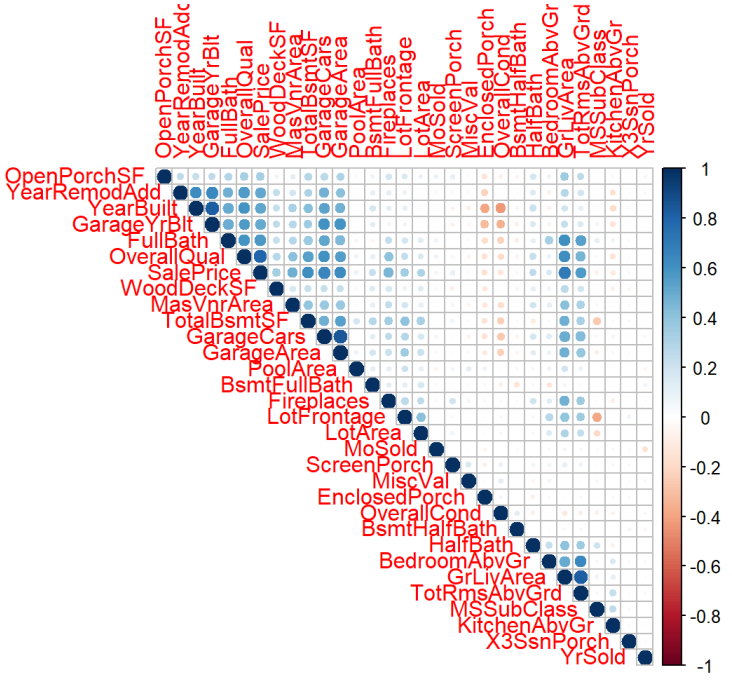
```
#Numeric DataSet
house_nums = select_if(housing_data_fix, is.numeric)
#Factor DataSet
house_factors = select_if(housing_data_fix, is.factor)

#All As Numbers Set
housing_data_all_numeric <- as.data.frame(sapply( housing_data_fix, as.integer ))

#str(house_nums)
#str(house_factors)
#str(housing_data_all_numeric)
```

Begin Factor Analysis - looking at numeric

```
#Evaluate Stability

#Test KMO Sampling Adequacy
library(psych)
KMO(housing_data_all_numeric)
#Overall MSA =  0.86
#This is >=0.5 or 0.6 - fairly good now

#Test Bartlett's Test of Sphericity
library(REdaS)
bart_spher(housing_data_all_numeric)
#p-value < 2.22e-16 (Very Small Number)
#This is significant

#Test for Reliability Analysis using Cronbach's Alpha
#library(psych)
alpha(housing_data_all_numeric, check.keys=TRUE)
#raw_alpha = 0.08
#This should be > 0.7 but it is extremely low. It's only exploratory but still way too low.
```

- Overall MSA = 0.86
- This is >=0.5 or 0.6 - fairly good now
- p-value < 2.22e-16 (Very Small Number)
- This is significant
- raw_alpha = 0.08
- This should be > 0.7 but it is extremely low. It's only exploratory but still way too low.

```
#Check Correlations
housing_cor_mat<-cor(house_nums)
#nut_cor_mat
corrplot(housing_cor_mat, type = "upper", order = "hclust")
```




```
#Most Correlated variables.
library(data.table)
setDT(melt(housing_cor_mat))[order(value)]
```

```
##              Var1          Var2      value
##   1:     YearBuilt   OverallCond -0.4376471
##   2:   OverallCond     YearBuilt -0.4376471
##   3: EnclosedPorch     YearBuilt -0.3995396
##   4:     YearBuilt EnclosedPorch -0.3995396
##   5:   LotFrontage    MSSubClass -0.3894662
##  ---
## 957:      PoolArea      PoolArea  1.0000000
## 958:       MiscVal       MiscVal  1.0000000
## 959:        MoSold        MoSold  1.0000000
## 960:        YrSold        YrSold  1.0000000
## 961:     SalePrice     SalePrice  1.0000000
```

```
#Multicollinearity Check and Quick Linear Model
model_1 <- lm(SalePrice ~ ., house_nums)
m_back <- step(model_1, direction = "backward", trace=FALSE )
#summary(model_1)
summary(m_back)
```

```
##
## Call:
## lm(formula = SalePrice ~ MSSubClass + LotFrontage + LotArea +
##     OverallQual + OverallCond + YearBuilt + MasVnrArea + TotalBsmtSF +
##     GrLivArea + BsmtFullBath + BedroomAbvGr + KitchenAbvGr +
##     TotRmsAbvGrd + Fireplaces + GarageCars + WoodDeckSF + ScreenPorch +
##     PoolArea, data = house_nums)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -435753  -17931   -2281   15145  319394
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.102e+05  1.115e+05  -7.269 6.97e-13 ***
## MSSubClass   -1.872e+02  3.478e+01  -5.383 9.01e-08 ***
## LotFrontage  -1.061e+02  6.040e+01  -1.757 0.079225 .
## LotArea       5.449e-01  1.577e-01   3.454 0.000573 ***
## OverallQual   1.895e+04  1.449e+03  13.077  < 2e-16 ***
## OverallCond   6.730e+03  1.216e+03   5.533 3.95e-08 ***
## YearBuilt     3.736e+02  5.643e+01   6.620 5.64e-11 ***
## MasVnrArea    3.310e+01  6.879e+00   4.812 1.71e-06 ***
## TotalBsmtSF   1.172e+01  4.214e+00   2.782 0.005490 **
## GrLivArea     4.998e+01  5.010e+00   9.977  < 2e-16 ***
## BsmtFullBath  1.405e+04  2.367e+03   5.937 3.92e-09 ***
## BedroomAbvGr -1.040e+04  2.096e+03  -4.965 8.00e-07 ***
## KitchenAbvGr -2.506e+04  7.266e+03  -3.449 0.000584 ***
## TotRmsAbvGrd  5.356e+03  1.507e+03   3.555 0.000395 ***
## Fireplaces    4.540e+03  2.132e+03   2.130 0.033406 *
## GarageCars    1.744e+04  2.428e+03   7.182 1.28e-12 ***
## WoodDeckSF    2.186e+01  9.957e+00   2.195 0.028365 *
## ScreenPorch   5.065e+01  1.997e+01   2.537 0.011337 *
## PoolArea     -5.616e+01  2.930e+01  -1.917 0.055506 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37070 on 1075 degrees of freedom
## Multiple R-squared:  0.8046, Adjusted R-squared:  0.8014
## F-statistic:   246 on 18 and 1075 DF,  p-value: < 2.2e-16
```

```
#alias(model_1)
#vif(model_1)
```

- X1stFlrSF, X2ndFlrSF, and GrLivArea are way over VIF of 10 so heavy multicolinearity here. I'll go back and remove X1stFlrSF and X2ndFlrSF because they sum up to GrLivArea. This removed the multicollinearity.
- Adj R2 is pretty good at 0.80, and given stability of components I can proceed.

```
library(nFactors)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## The following object is masked from 'package:sm':
##
##     muscle
```

```
## Loading required package: boot
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following object is masked from 'package:sm':
##
##     dogs
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

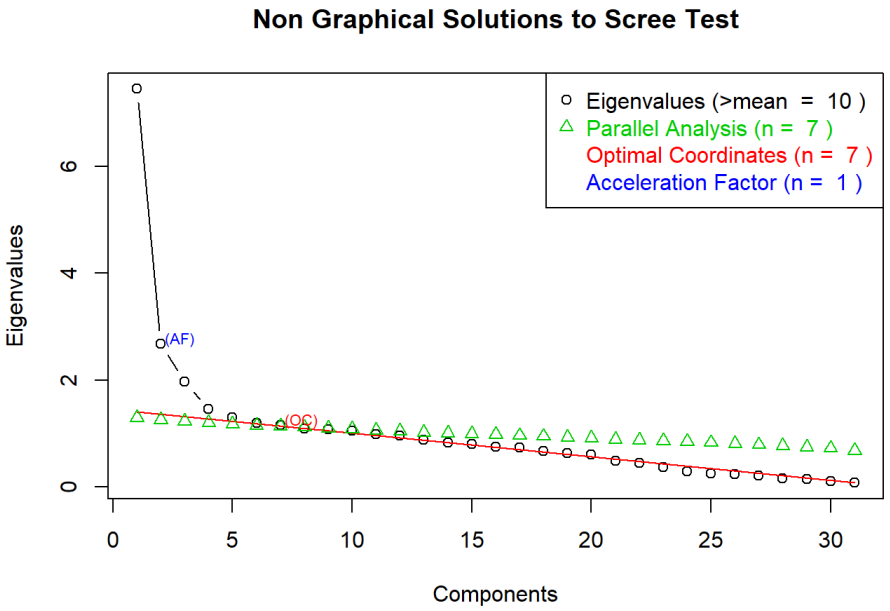
```
## The following object is masked from 'package:survival':
##
##     aml
```

```
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
##
## Attaching package: 'nFactors'
```

```
## The following object is masked from 'package:lattice':
##
##     parallel
```

```
ev <- eigen(cor(house_nums)) # get eigenvalues
ap <- parallel(subject=nrow(house_nums),var=ncol(house_nums), rep=100, cent=.05)
nS <- nScree(x=ev$values, aparallel=ap$eigen$qevpea)
plotnScree(nS)
```




```
#Factor Analysis
CFA2 = factanal(house_nums, 2)
print(CFA2$loadings, cutoff=.4, sort=T)
```

```
##
## Loadings:
##               Factor1 Factor2
## OverallQual    0.676   0.476
## YearBuilt      0.901
## YearRemodAdd   0.699
## FullBath       0.526   0.507
## GarageYrBlt    0.883
## GarageCars     0.661
## GarageArea     0.622
## SalePrice      0.619   0.599
## GrLivArea              0.930
## BedroomAbvGr           0.587
## TotRmsAbvGrd           0.849
## MSSubClass
## LotFrontage            0.415
## LotArea
## OverallCond
## MasVnrArea
## TotalBsmtSF    0.484
## BsmtFullBath
## BsmtHalfBath
## HalfBath
## KitchenAbvGr
## Fireplaces             0.468
## WoodDeckSF
## OpenPorchSF
## EnclosedPorch
## X3SsnPorch
## ScreenPorch
## PoolArea
## MiscVal
## MoSold
## YrSold
##
##                Factor1 Factor2
## SS loadings      5.052   4.197
## Proportion Var   0.163   0.135
## Cumulative Var   0.163   0.298
```

```
summary(CFA2)
```

```
##              Length Class    Mode
## converged      1    -none-   logical
## loadings      62    loadings numeric
## uniquenesses  31    -none-   numeric
## correlation  961    -none-   numeric
## criteria       3    -none-   numeric
## factors        1    -none-   numeric
## dof            1    -none-   numeric
## method         1    -none-   character
## rotmat         4    -none-   numeric
## STATISTIC      1    -none-   numeric
## PVAL           1    -none-   numeric
## n.obs          1    -none-   numeric
## call           3    -none-   call
```

- After looking at results from factor number choices 2 through 20, I can conclude that CFA is not the best approach to understanding our dataset. I even tried taking all the factor datatype data, turning it into numerical and it didn't help any. Cumulative proportion of variance explained by factors even with +15 is under 40%. This is a poor result and leads us to pursue other more appropriate methods of analyzing and grouping our data.
- That said, even from 2 factors we can see the general theme of the dataset - For a higher SalePrice, generally, more is better. This really applies generally across all the datatypes, higher quality, more square footage, more rooms, more features like fireplaces, etc. lead to higher SalePrice.
- Attempts to do FA will likely lead to worse groupings of the data than logically looking over the available fields in the dataset. It may lead to better results to go through and manually create new features that represent certain aspects of the dataset. I will pursue this preprocessing next.
- A significant amount of missing data has also made such analysis more challenging.

Problem 3

3) Paper Review (20 points): An academic paper from a conference or Journal will be posted to the Homework 3 content section of D2L. It contains a usage of Canonical Correlation. Review the paper and evaluate their usage of Canonical Correlation. In particular, address (Multivariate Relationships Between Statistics Anxiety and Motivational Beliefs)

a) How suitable is their data for CC?

- First, their data comes from, 305 college students enrolled in different Turkish Universities.
- A 5-point Likert-type instrument was used, which measured statistics anxiety in 23 statistics anxiety statement items and 28 dealing-with-statistics items. This instrument is referred to as STARS in the paper. Also included is MSLQ, a 7-point Likert-type questionnaire from which the Motivation Scale was used. A demographic questionnaire was also given to the volunteer university students in this study. Basically they get a statistics anxiety set and a motivational beliefs set and are using CCA on those two sets.
- Second, as we discussed in class, CC is an exploratory tool to see if two sets of continuous variables are related. They are considering the likert data as continuous. They also have the theoretical reasoning backing their approach regarding how student anxiety and motivation are understood in the context of statistics anxiety.

  Additionally, they include include the following to support how suitable the data is:

  - Assumption of multivariate normality was evaluated and confirmed with Mahalanobis distance greater than x2(f(2) = 124.84 regarded as a multivariate outlier ($p < .001$) yielding no outliers.
  - Homogeneity of variance assumption met using Box's M with $p>0.005$. They did this to determine whether or not co-variance matricies could be used in their CCA (Baloglu et al., p. 435).
  - Coefficient consistency was also evaluated to determine if their data was suitable for CCA. Cronbach's alpha, means, standard deviations, and more were calculated in reference to the different instruments used.

b) How are they applying CC? What two groups of variables are being correlated? Are they metric, ordinal, nominal?

- They are applying CC to study the relationship between statistics anxiety and motivational beliefs, based on ideas about the interplay of motivation and anxiety in general. The two groups of variables that are being correlated are a Statistics Anxiety Set and a Motivational Beliefs Set. The underlying data comes from the Likert-scale instruments used, so it is Ordinal data that they are treating as continuous for the CCA.

c) What methods do they use to judge the quality of the correlation? Do they evaluate, and how do they evaluate the stability of the components?

- They did look at Cronbach's alpha for the reliability of the scales used and stablity of components was taken into account as well. Wilk's lambda is used to express how well the variates are accounting for probability distributions in the different variables. They don't very clearly look at X-Y variables like in the in-class paper covered, but do still look at variable correlations.

d) How many correlates do they concentrate on in their analysis, and do they attempt to interpret the correlates in terms of the original variables?

- There were 6 in total, the first canonical correlation was 0.62 and accounted for roughyl 39% overlapping variance (Wilk's lambda = .50), the second accounted for 13% overlapping variance at 0.35 (Wilk's lambda = .80), and the last 4 canonical correlations were essentially zero. The two variates combined for 55% of statistics anxiety set variability and 59% for the motivation beliefs set.
- Yes they attempt to interpret the correlates in terms of the original variables. Though they did not give names to the variates, they were just called First Canonical Variate and Second Canoncial Variate. Variables correlated with the statistics anxiety set includedtest/class anxiety, worth of statistics, computational self-concept, fear of statistics instructor, and fear of asking for help. These are pretty obviously connected to anxiety but it's good to show statistically I guess. For the motivational belief set task value, self-efficacy for learning and performance, and intrinsic goal orientation were the top correlations.
- The second canonical variate had variables like worth of statistics connected to anxiety and control of learning beliefs connected to motivation.

Table 2. *Correlations, Standardized Canonical Coefficients, Canonical Correlations, Percentage of Variances, and Redundancies between the Statistics Anxiety and Motivational Beliefs and Their Canonical Variates*

	First Canonical Variate		Second Canonical Variate	
Statistics Anxiety Set	*r*	Coef.	*r*	Coef.
Worth of Statistics	-.73	-.39	.54	.96
Interpretation Anxiety	-.40	.28	-.50	-.43
Test/Class Anxiety	-.82	-.72	-.53	-.54
Computational Self-concept	-.72	-.32	.17	.05
Fear of Asking for Help	-.48	-.08	-.27	.06
Fear of Statistics Instructor	-.52	.05	.04	-.29
Variance Percentage	.40		.15	Total = .55
Redundancy	.15		.02	Total = .17
Motivational Beliefs Set				
Intrinsic Goal Orientation	.57	.18	-.68	-.44
Extrinsic Goal Orientation	-.11	-.19	-.58	-.25
Task Value	.60	.26	-.74	-.66
Control of Learning Beliefs	.19	-.06	-.40	-.08
Self-efficacy for Learning and Performance	.58	.57	-.36	.63
Test Anxiety	-.59	-.69	-.55	-.47
Variance Percentage	.24		.32	Total = .59
Redundancy	.09		.04	Total = .13
r_c	.62		.35	
R_c^2	38.44		12.25	

Coef., standardized canonical coefficients; *r*, canonical loadings (structure coefficients); r_c, canonical correlation.

problem3_image

e) What conclusions does CC allow them to draw?

- They make some expected conclusions that students who recognize statistics as important, useful, and beneficial will experience or display less text anxiety when it comes to learning statistics. Basically if motivated students are less likely to be anxious.
- Also they note that there analsis does not imply any causality, just description of the relationship between motivation and anxiety in statistics.

Problem 4

__ 4) Paper Review (20 points): An academic paper from a conference or Journal will be posted to the Homework 3 content section of D2L. It contains a usage of Canonical Correlation. Review the paper and evaluate their usage of Canonical Correlation. In particular, address (Vacation Benefits and Activities Understanding Chinese Family Travelers)__

a) How suitable is their data for CC?

- The paper authors sought to extend the field's understanding of Chinese family traveler wants and actions at different tourist destinations with the more specific purpose of uncovering relationships between benefits sought and desitination activities. For data collection they used questionnaires, of which 253 were obtained initially, followed by another 53 in a second collection round. So 306 questionnaire responses with items based on 5-point scales with 19 items regarding benefits importanct and 32 items regarding activity participation frequency.
- They also check suitability by checking linearity, multicollinearity, and sample size for CCA.

b) How are they applying CC? What two groups of variables are being correlated? Are they metric, ordinal, nominal?

- After conducting factor analysis, they used CC to assess the relationship between benefits sought and vacation activities - specifically applied towards undestanding Chinese Family travelers behavior. Because they used a questionnaire with Likert-type data as well the variables are ordinal, and for the CC are therefore treated as continuous just like in the last paper.

- That said, four separate analysis were conducted for CC - one for each of the four benefit factors found and the activity items as the other set. So, technically there were 5 groups, 4 benefit groups correlated individually to the 1 activity set.

c) What methods do they use to judge the quality of the correlation? Do they evaluate, and how do they evaluate the stability of the components?

- They check F statistic and for each varaite alpha is 0.05 and redundancy indices are >0.01.
- The don't go into much detail beyond that to prove component stability.

d) How many correlates do they concentrate on in their analysis, and do they attempt to interpret the correlates in terms of the original variables?

- The first canonical variate pair was the only one for each test that accounted for statistically significant portions of the benefits sought/activity participation relationship. And yes they interpret the correlates in terms of the original variables.
- 1 - It showed that there was a relationship between picture/video taking and items from the Communication/Togetherness factor. Specifically, fun with family, respecting family member decisions, finding things in common, and sharing quality time were positively associated with respondents taking pictures and videos.
- 2 - Shared Exploration factor and activity participation basically points to the relationships between cultural experiences like tasting foods and trying new things with visiting historic sites, dining, and also taking pictures.
- 3 - Escape and Relaxation get related to things like kayaking and farm visits, so chinese family travelers who want to escape and relax may be more inclined to participate in outdoor or nature activities.
- 4 - Finally, Experiential Learning for Children is gets correlated to basically activities that may accomplish these things like visiting historical or ecological sights, eating different foods, etc.

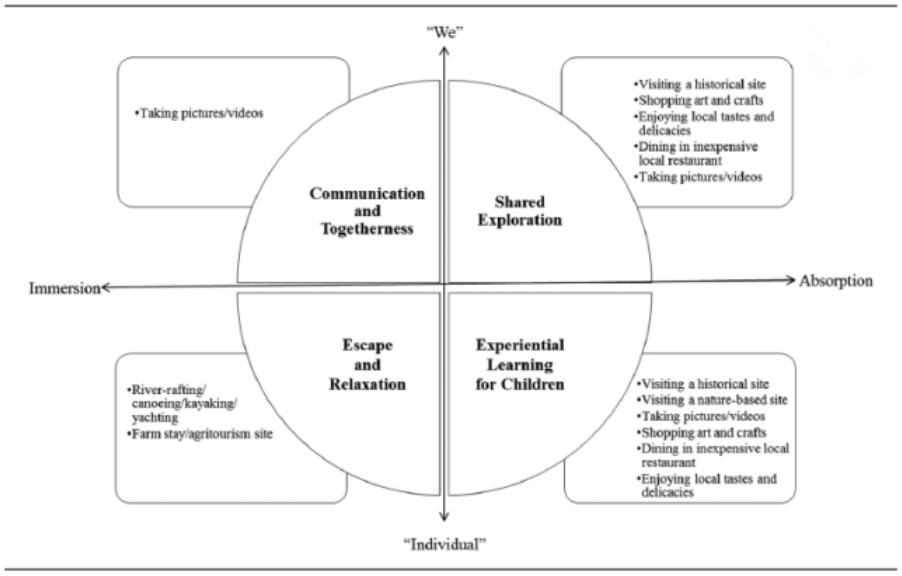
Table 3
Overall Results for Canonical Correlation Analysis

	Eigenvalue	Canonical Correlation	Squared Canonical Correlation	*F* Statistic	Probability
Canonical Function 1	.40451	.53666	.28801	1.39809	< .0001
Tests of Significance	Value	*F* Value	Num *DF*	Den *DF*	Probability
Wilks' lambda	1.31679	1.38537	256	1800.00	< .0001
Pillai's trace	1.66715	1.40828	256	1730.00	< .0001
Hotelling–Lawley trace	.22850	1.39809	256	1712.69	< .0001
Roy's greatest root	.28801				
Canonical Function 2	.41641	.54221	.29399	1.29891	.02
Tests of Significance	Value	*F* Value	Num *DF*	Den *DF*	Probability
Wilks' lambda	.60957	1.26415	128	900.00	.033
Pillai's trace	.77525	1.33550	128	882.00	.012
Hotelling–Lawley trace	.50444	1.29891	128	885.00	.020
Roy's greatest root	.29399				
Canonical Function 3	.28493	.47090	.22174	1.52314	< .0001
Tests of Significance	Value	*F* Value	Num *DF*	Den *DF*	Probability
Wilks' lambda	.71571	1.53225	128	900.00	< .0001
Pillai's trace	.87852	1.51340	128	882.00	< .0001
Hotelling–Lawley trace	.45322	1.52314	128	885.80	< .0001
Roy's greatest root	.22174				
Canonical Function 4	.43832	.55204	.30475	1.87818	< .0001
Tests of Significance	Value	*F* Value	Num *DF*	Den *DF*	Probability
Wilks' lambda	.62996	1.85768	96	675.00	< .0001
Pillai's trace	.82255	1.89928	96	665.00	< .0001
Hotelling–Lawley trace	.48925	1.87818	96	668.45	< .0001
Roy's greatest root	.30475				

problem4_image

e) What conclusions does CC allow them to draw?

- Most notably, there seems to be a strong focus on enriching children's learning and life experiences through travel for Chinese families. The importance of the family unit in Chinese culture plays into the types of vacations and travel they partake in.



problem4E_image

- The above chart provided by the authors summarizes their findings further.
- They also conclude that more research is necessary to understand how different family and travel types play into the greater narrative of Chinese family travel.

Problem 5

__ 5) (20 points): Perform the following Canonical Correlation Analysis on the Young People Survey from Lab 2: PCA/FA. Perform a canonical correlation analysis describing the relationships between the hobbies/interests and music variables using the data under the Lab 2: PCA/FA in the content folder).__

1. Answer the following questions regarding the canonical correlations.

a. Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f., and p-value.

```
#Read in Datasets
responses <- read.csv("responses.csv")
#Check Sample Size and Number of Variables
#dim(responses)
#Show for first 6 rows of data
#head(responses)
#names(responses)
#For All Variables
#sum(is.na(responses))
#571 total missing values
#Treat Missing Values
#Listwise Deletion
responses2 <- na.omit(responses)
#Check new data has no missing data
sum(is.na(responses2))
```

```
## [1] 0
```

```
#Show Structure of Dataset
#str(responses2, list.len=ncol(responses2))

#Show column Numbers
#names(responses2)

responses3 <- responses2[,c(1:73,76,77:107,110:132,134:140,141:144)]

music <- responses2[,1:19]
dim(music)
```

```
## [1] 686  19
```

```
describe(music)
```

```
##                          vars   n mean   sd median trimmed  mad min max
## Music                       1 686 4.76 0.60      5    4.91 0.00   1   5
## Slow.songs.or.fast.songs    2 686 3.29 0.79      3    3.26 0.00   1   5
## Dance                       3 686 3.07 1.19      3    3.09 1.48   1   5
## Folk                        4 686 2.26 1.11      2    2.14 1.48   1   5
## Country                     5 686 2.11 1.07      2    1.97 1.48   1   5
## Classical.music             6 686 2.98 1.24      3    2.98 1.48   1   5
## Musical                     7 686 2.76 1.28      3    2.70 1.48   1   5
## Pop                         8 686 3.44 1.17      4    3.50 1.48   1   5
## Rock                        9 686 3.79 1.15      4    3.91 1.48   1   5
## Metal.or.Hardrock          10 686 2.36 1.40      2    2.20 1.48   1   5
## Punk                       11 686 2.45 1.29      2    2.34 1.48   1   5
## Hiphop..Rap                12 686 2.89 1.35      3    2.86 1.48   1   5
## Reggae..Ska                13 686 2.77 1.21      3    2.73 1.48   1   5
## Swing..Jazz                14 686 2.76 1.26      3    2.70 1.48   1   5
## Rock.n.roll                15 686 3.16 1.21      3    3.20 1.48   1   5
## Alternative                16 686 2.89 1.34      3    2.86 1.48   1   5
## Latino                     17 686 2.81 1.32      3    2.76 1.48   1   5
## Techno..Trance             18 686 2.30 1.31      2    2.15 1.48   1   5
## Opera                      19 686 2.15 1.19      2    2.00 1.48   1   5
##                          range  skew kurtosis   se
## Music                        4 -3.07    10.95 0.02
## Slow.songs.or.fast.songs     4  0.20     0.94 0.03
## Dance                        4 -0.04    -0.86 0.05
## Folk                         4  0.69    -0.20 0.04
## Country                      4  0.83     0.07 0.04
## Classical.music              4  0.09    -0.97 0.05
## Musical                      4  0.22    -0.98 0.05
## Pop                          4 -0.33    -0.77 0.04
## Rock                         4 -0.67    -0.45 0.04
## Metal.or.Hardrock            4  0.63    -0.94 0.05
## Punk                         4  0.45    -0.94 0.05
## Hiphop..Rap                  4  0.03    -1.21 0.05
## Reggae..Ska                  4  0.17    -0.87 0.05
## Swing..Jazz                  4  0.16    -1.01 0.05
## Rock.n.roll                  4 -0.14    -0.87 0.05
## Alternative                  4  0.12    -1.13 0.05
## Latino                       4  0.23    -1.08 0.05
## Techno..Trance               4  0.61    -0.85 0.05
## Opera                        4  0.84    -0.25 0.05
```

```
hobbies_interests <- responses2[,32:63]
dim(hobbies_interests)
```

```
## [1] 686  32
```

```
describe(hobbies_interests)
```

```
##                         vars   n mean   sd median trimmed  mad min max
## History                    1 686 3.23 1.26      3    3.28 1.48   1   5
## Psychology                 2 686 3.14 1.25      3    3.17 1.48   1   5
## Politics                   3 686 2.63 1.29      3    2.53 1.48   1   5
## Mathematics                4 686 2.40 1.35      2    2.25 1.48   1   5
## Physics                    5 686 2.10 1.25      2    1.92 1.48   1   5
## Internet                   6 686 4.19 0.90      4    4.29 1.48   1   5
## PC                         7 686 3.14 1.31      3    3.17 1.48   1   5
## Economy.Management         8 686 2.66 1.37      2    2.58 1.48   1   5
## Biology                    9 686 2.62 1.36      2    2.53 1.48   1   5
## Chemistry                 10 686 2.12 1.36      2    1.90 1.48   1   5
## Reading                   11 686 3.20 1.49      3    3.25 1.48   1   5
## Geography                 12 686 3.11 1.28      3    3.14 1.48   1   5
## Foreign.languages         13 686 3.81 1.12      4    3.94 1.48   1   5
## Medicine                  14 686 2.48 1.33      2    2.35 1.48   1   5
## Law                       15 686 2.22 1.24      2    2.07 1.48   1   5
## Cars                      16 686 2.63 1.41      2    2.54 1.48   1   5
## Art.exhibitions           17 686 2.62 1.32      2    2.52 1.48   1   5
## Religion                  18 686 2.23 1.32      2    2.05 1.48   1   5
## Countryside..outdoors     19 686 3.61 1.23      4    3.73 1.48   1   5
## Dancing                   20 686 2.40 1.43      2    2.25 1.48   1   5
## Musical.instruments       21 686 2.30 1.50      2    2.13 1.48   1   5
## Writing                   22 686 1.87 1.28      1    1.62 0.00   1   5
## Passive.sport             23 686 3.39 1.41      4    3.49 1.48   1   5
## Active.sport              24 686 3.24 1.51      3    3.29 2.97   1   5
## Gardening                 25 686 1.87 1.16      1    1.66 0.00   1   5
## Celebrities               26 686 2.32 1.27      2    2.19 1.48   1   5
## Shopping                  27 686 3.26 1.29      3    3.32 1.48   1   5
## Science.and.technology    28 686 3.27 1.26      3    3.34 1.48   1   5
## Theatre                   29 686 3.02 1.32      3    3.03 1.48   1   5
## Fun.with.friends          30 686 4.55 0.74      5    4.72 0.00   2   5
## Adrenaline.sports         31 686 2.88 1.41      3    2.85 1.48   1   5
## Pets                      32 686 3.32 1.55      4    3.40 1.48   1   5
##                         range  skew kurtosis   se
## History                     4 -0.12    -1.00 0.05
## Psychology                  4 -0.07    -1.03 0.05
## Politics                    4  0.33    -0.97 0.05
## Mathematics                 4  0.53    -0.93 0.05
## Physics                     4  0.90    -0.32 0.05
## Internet                    4 -0.91     0.26 0.03
## PC                          4 -0.10    -1.13 0.05
## Economy.Management          4  0.33    -1.13 0.05
## Biology                     4  0.44    -1.00 0.05
## Chemistry                   4  1.01    -0.29 0.05
## Reading                     4 -0.17    -1.39 0.06
## Geography                   4 -0.07    -1.02 0.05
## Foreign.languages           4 -0.67    -0.36 0.04
## Medicine                    4  0.60    -0.76 0.05
## Law                         4  0.74    -0.49 0.05
## Cars                        4  0.34    -1.19 0.05
## Art.exhibitions             4  0.37    -0.98 0.05
## Religion                    4  0.76    -0.61 0.05
## Countryside..outdoors       4 -0.60    -0.61 0.05
## Dancing                     4  0.60    -1.00 0.05
## Musical.instruments         4  0.72    -1.00 0.06
## Writing                     4  1.27     0.28 0.05
## Passive.sport               4 -0.35    -1.17 0.05
## Active.sport                4 -0.24    -1.37 0.06
## Gardening                   4  1.24     0.59 0.04
## Celebrities                 4  0.57    -0.79 0.05
## Shopping                    4 -0.17    -1.08 0.05
## Science.and.technology      4 -0.19    -0.98 0.05
## Theatre                     4  0.06    -1.12 0.05
## Fun.with.friends            3 -1.63     2.00 0.03
## Adrenaline.sports           4  0.10    -1.27 0.05
## Pets                        4 -0.32    -1.41 0.06
```

```
#library
library(yacca)
c2= cca(hobbies_interests, music)
#function names
ls(c2)
```

```
##  [1] "canvarx"       "canvary"       "chisq"         "corr"
##  [5] "corrsq"        "df"            "xcancom"       "xcanvad"
##  [9] "xcoef"         "xcrosscorr"    "xcrosscorrsq"  "xlab"
## [13] "xrd"           "xstructcorr"   "xstructcorrsq" "xvrd"
## [17] "ycancom"       "ycanvad"       "ycoef"         "ycrosscorr"
## [21] "ycrosscorrsq"  "ylab"          "yrd"           "ystructcorr"
## [25] "ystructcorrsq" "yvrd"
```

```
# Perform a chisquare test on C2
summary(c2)
```

```
Bartlett's Chi-Squared Test:

           rho^2      Chisq  df    Pr(>X)
CV 1  5.3744e-01 1.8432e+03 608 < 2.2e-16 ***
CV 2  3.6760e-01 1.3352e+03 558 < 2.2e-16 ***
CV 3  2.6962e-01 1.0332e+03 510 < 2.2e-16 ***
CV 4  2.2079e-01 8.2612e+02 464 < 2.2e-16 ***
CV 5  1.4870e-01 6.6172e+02 420 4.372e-13 ***
CV 6  1.3372e-01 5.5563e+02 378 6.655e-09 ***
CV 7  1.1784e-01 4.6103e+02 338 9.325e-06 ***
CV 8  1.0243e-01 3.7841e+02 300  0.001419 **
CV 9  9.1794e-02 3.0719e+02 264  0.034801 *
CV 10 8.0726e-02 2.4374e+02 230  0.254903
CV 11 7.4177e-02 1.8827e+02 198  0.678525
CV 12 5.5238e-02 1.3748e+02 168  0.959173
CV 13 3.7437e-02 1.0003e+02 140  0.995633
CV 14 3.5621e-02 7.4889e+01 114  0.998246
CV 15 2.5347e-02 5.0987e+01  90  0.999699
CV 16 2.4613e-02 3.4068e+01  68  0.999811
CV 17 1.1288e-02 1.7645e+01  48  0.999982
CV 18 8.7912e-03 1.0164e+01  30  0.999732
CV 19 6.5712e-03 4.3447e+00  14  0.993010
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

problem5a_image

- The Screenshot above shows the Test Statistic (Chi-Squared was used), the d.f. and p-values for each CV. CV1 is solid.

b. How many significant canonical variates are there?

- There are 9 significant canonical variates based on the Bartlett’s Chi-Squared Test. These were p<0.05.

c. Present the first two canonical correlations (Cancor)?

```
#c = cancor(hobbies_interests, music)
#c
```

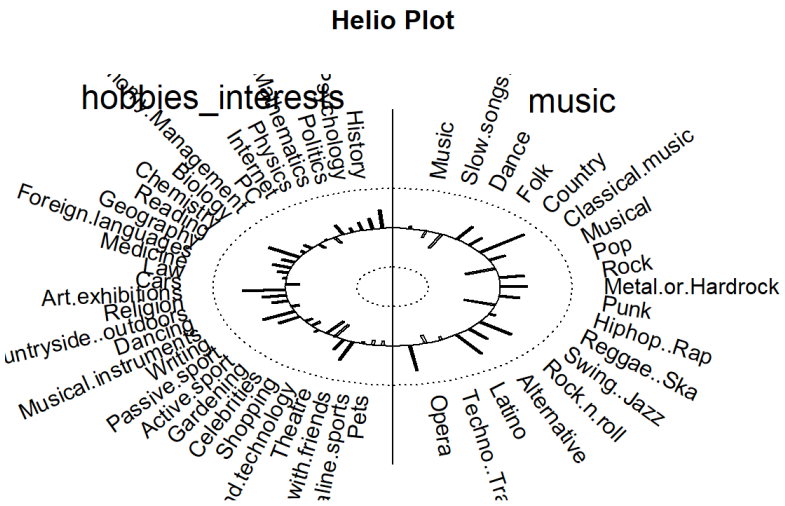
```
Canonical Correlations

      CV 1       CV 2
0.73310410 0.60630173
```

problem5b_image

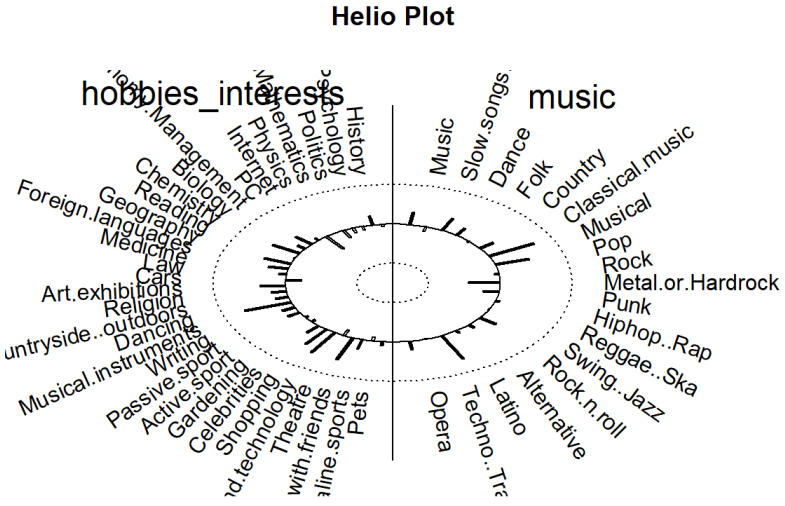
- The screenshot above shows the first two Canonical Correlations. I used the yacca package for this.

d. What can you conclude from the above analyses?

```
helio.plot(c2, cv=1, x.name="hobbies_interests",
           y.name="music")
```




```
helio.plot(c2, cv=2, x.name="hobbies_interests",
           y.name="music")
```



- There appears to be a significant relationship between one's hobbies/interests and one's music preferences as indicated by the strong 0.733 CV1 correlation from part c. This suggests that people with similar hobbies/interests also tend to share certain musical preferences.

2. Answer the following questions regarding the canonical variates.

a. Give the formulae for the first canonical variate for the hobbies/interests and music variables.

Canonical Variate Coefficients:

X Vars:

	CV 1
History	0.172749150
Psychology	0.002125842
Politics	-0.002310111
Mathematics	-0.047144486
Physics	0.120232697
Internet	-0.077974365
PC	-0.041706964
Economy.Management	0.005014450
Biology	-0.019825147
Chemistry	-0.067927310
Reading	0.118641213
Geography	-0.006950079
Foreign.languages	0.077365531
Medicine	0.032505202
Law	0.007091897
Cars	-0.102899297
Art.exhibitions	0.281843295
Religion	0.036877052
Countryside..outdoors	0.033218928
Dancing	-0.188697927
Musical.instruments	0.209544568
Writing	-0.054164761
Passive.sport	-0.006289547
Active.sport	-0.055515407
Gardening	0.024066410
Celebrities	-0.111101445
Shopping	-0.162427967
Science.and.technology	0.211793168
Theatre	0.097878365
Fun.with.friends	0.013056913
Adrenaline.sports	-0.079666650
Pets	-0.050524154

hobbies/interests

- The formula is the sum of the above coefficients linearly. cv1_hobbiesints = 0.173(History) + … - 0.051(Pets)

Y Vars:

	CV 1
Music	-0.01562241
Slow.songs.or.fast.songs	-0.01322202
Dance	-0.06624628
Folk	0.05324425
Country	0.03057841
Classical.music	0.29708481
Musical	0.01336652
Pop	-0.16910909
Rock	0.02995350
Metal.or.Hardrock	0.02201227
Punk	0.02668992
Hiphop..Rap	-0.13743338
Reggae..Ska	0.04113529
Swing..Jazz	0.18372228
Rock.n.roll	-0.05969345
Alternative	0.18844613
Latino	-0.09332992
Techno..Trance	-0.03643442
Opera	0.22067889

music

- The formula is the sum of the above coefficients linearly. cv1_music = -0.016(Music) + … + 0.221(Opera)

b. Give the correlations between the first canonical variate for hobbies/interests and the hobbies/interests variables, and the correlations between the first canonical variate for music and the music variables.

Correlations between the first canonical variate for hobbies/interests and the hobbies/interests variables

Structural Correlations (Loadings):

X Vars:

	CV 1
History	0.46221429
Psychology	0.25979316
Politics	0.20744422
Mathematics	0.10592829
Physics	0.25990474
Internet	-0.16869551
PC	0.02187078
Economy.Management	-0.13229445
Biology	0.13118878
Chemistry	0.07097934
Reading	0.49882401
Geography	0.20030543
Foreign.languages	0.26757344
Medicine	0.14914547
Law	0.04932124
Cars	-0.20389353
Art.exhibitions	0.59570392
Religion	0.33755228
Countryside..outdoors	0.23038461
Dancing	-0.11930085
Musical.instruments	0.49261678
Writing	0.33707600
Passive.sport	-0.11724919
Active.sport	-0.15457963
Gardening	0.08370508
Celebrities	-0.38491574
Shopping	-0.34901957
Science.and.technology	0.29319546
Theatre	0.49997655
Fun.with.friends	-0.04567797
Adrenaline.sports	-0.12035347
Pets	-0.11924072

Correlations between the first canonical variate for music and the music variables

Y Vars:	CV 1
Music	0.05748834
Slow.songs.or.fast.songs	-0.18510350
Dance	-0.38749621
Folk	0.38178244
Country	0.25568618
Classical.music	0.76319684
Musical	0.25721542
Pop	-0.43751181
Rock	0.35701104
Metal.or.Hardrock	0.37875199
Punk	0.29228065
Hiphop..Rap	-0.45282608
Reggae..Ska	0.10692618
Swing..Jazz	0.54369432
Rock.n.roll	0.36417364
Alternative	0.59611787
Latino	-0.07093147
Techno..Trance	-0.20765156
Opera	0.65959492

c. What can you conclude from the above analyses?

- I can conclude that there is a strong correlation between hobbies/interests and music based on the CV1 Canonical Correlation of 0.733.
- Top4 Hobbies/Interests Correlations:
- Art.exhibitions 0.596
- Theatre .499
- Reading .498
- Musical.instruments .492
- These correlations indicated that interests or hobbies in Art Exhbitions, Theatre, Reading, and/or Musical Instruments play a significant role in the makeup of hobbies/interests and how they relate to Music through CV1.

- Top4 Music Correlations:
- Classical.music 0.763
- Opera 0.659
- Alternative 0.596
- Swing..Jazz 0.543
- These correlations indicated that music preferences in Classical Music, Opera, Alternative, and Swing/Jazz play a significant role in the what explains music preferences in our dataset and how they relate to hobbies/interests through CV1.

- Also, the other variables also play a role and can be interpretted via the screenshots in part b. Negative correlations have the opposite meaning.

Extra Credit

EXTRA CREDIT (10 points) Perform a correspondence analysis on countries and time spent traveling data in travels2.xlsx. In this file you are provided with the table for the two sets of categories. In particular perform the following:

a) Create a mosaic plot of the two categorical variables.

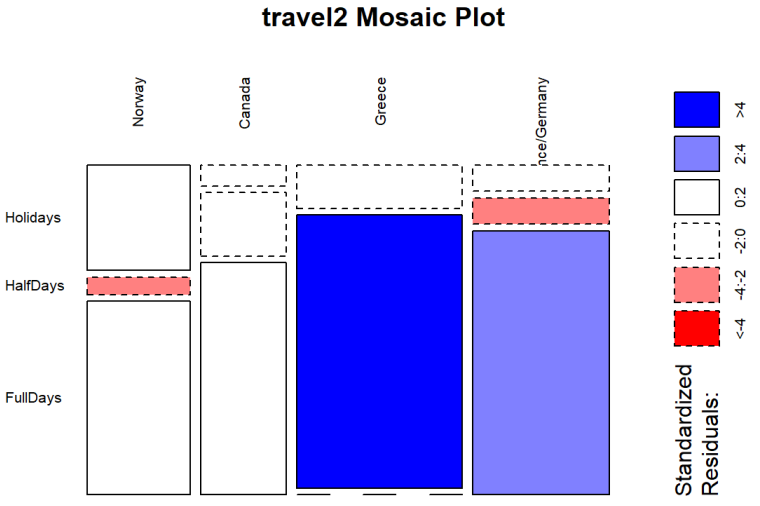
```
# libs
library("FactoMineR")
library("factoextra")
library("graphics")

#Read in Datasets
travels2 <- read.csv("travels2.csv")
head(travels2)
```

```
##         Country Holidays HalfDays FullDays
## 1        Norway        6        1       11
## 2        Canada        1        3       11
## 3        Greece        4       25        0
## 4 FranceGermany        2        2       20
```

```
#create table
nums <- matrix(c(6,1,11,1,3,11,4,25,0,2,2,20), ncol=3,byrow=TRUE)
colnames(nums)<-c("Holidays","HalfDays","FullDays")
rownames(nums)<-c("Norway","Canada","Greece","France/Germany")
nums<-as.table(nums)
nums
```

```
##                Holidays HalfDays FullDays
## Norway                6        1       11
## Canada                1        3       11
## Greece                4       25        0
## France/Germany        2        2       20
```

```
#Mosaic plot
mosaicplot(nums, shade = TRUE, las=2, main = "travel2 Mosaic Plot")
```

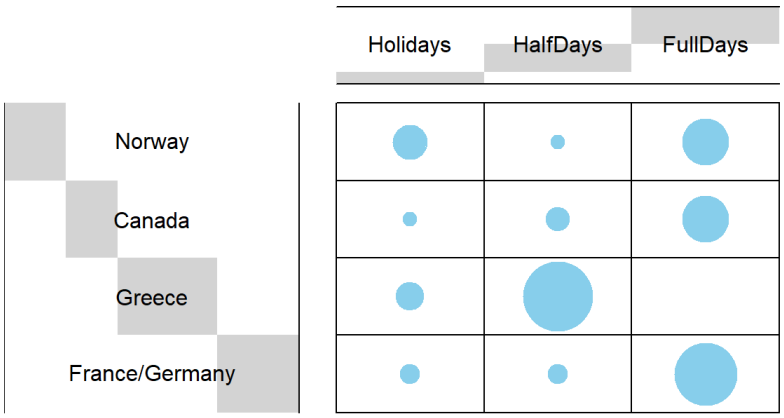



```
#Another Cool Visualization
library("gplots")
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
balloonplot(t(nums), main ="travel2 Balloonplot", xlab ="", ylab="",label = FALSE, show.margins = FALSE)
```



b) Plot the results of the correspondence analysis

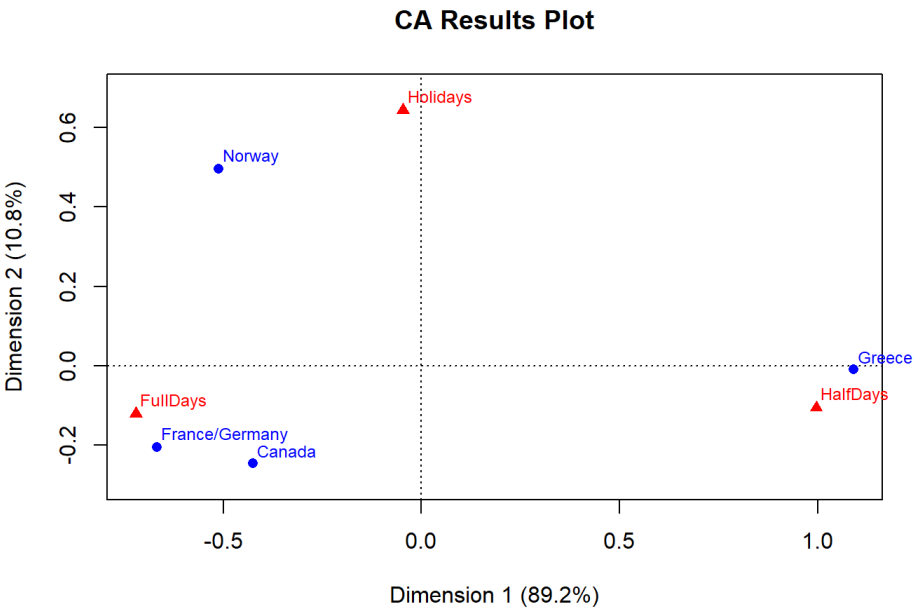
```
library(ca)

#CA Results Plot
fit<-ca(nums)
print(fit)
```

```
##
##  Principal inertias (eigenvalues):
##            1        2
## Value      0.610953 0.073964
## Percentage 89.2%    10.8%
##
##
##  Rows:
##            Norway    Canada    Greece France/Germany
## Mass     0.209302  0.174419  0.337209       0.279070
## ChiDist  0.712953  0.491461  1.089741       0.697971
## Inertia  0.106389  0.042128  0.400448       0.135952
## Dim. 1  -0.655435 -0.543830  1.394129      -0.853102
## Dim. 2   1.823120 -0.906976 -0.034972      -0.758222
##
##
##  Columns:
##          Holidays  HalfDays  FullDays
## Mass     0.151163  0.360465  0.488372
## ChiDist  0.645948  1.001123  0.730443
## Inertia  0.063072  0.361275  0.260570
## Dim. 1  -0.059651  1.273595 -0.921571
## Dim. 2   2.368929 -0.390062 -0.445337
```

```
summary(fit)
```

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.610953  89.2  89.2  **********************
##  2      0.073964  10.8 100.0  ***
##         -------- -----
##  Total: 0.684917 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1  cor ctr    k=2 cor ctr
## 1 | Nrwy |  209 1000  155 | -512  516  90 |  496 484 696 |
## 2 | Cand |  174 1000   62 | -425  748  52 | -247 252 143 |
## 3 | Grec |  337 1000  585 | 1090 1000 655 |  -10   0   0 |
## 4 | FrnG |  279 1000  198 | -667  913 203 | -206  87 160 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | Hldy |  151 1000   92 |  -47   5   1 |  644 995 848 |
## 2 | HlfD |  360 1000  527 |  995 989 585 | -106  11  55 |
## 3 | FllD |  488 1000  380 | -720 973 415 | -121  27  97 |
```

```
plot(fit)
title("CA Results Plot")
```



- The plot above summarizes the results of the correspondance analysis. We see that dimension 1 accounts for 89.2% variance and dimension 2 10.8%, so all is accounted for in the plot. Interesting to note that France/Germany are most disimilar from Greece. Greece is actually opposite the other 3 countries groups.

c) With each country, create a profile for the time spent traveling. Which countries spend least amount traveling? Which countries spend the most time traveling? For each country, draw the scale for that country and demonstrate that time spent traveling on the graph.

- Norway - Most associated with FullDays actually, even on the plot as visualize by the distance from origin and angles. Next most is Holidays for Norway's travel data. It has more Holiday travel than France/Germany and Canada - closer to Greece actually on Holliday travel as seen in distance to origin/angles.
- France/Germany - These countries seem to be more associated with FUllDays travel opposed to HalfDays and Holidays. They are most similar to Canada in this respect and then Norway, then Greece.
- Canada - Canada is most associated with FullDays travel like France/Germany. They take more HalfDay travel than France/Germany and Norway, but significantly less than Greece. In terms of Holiday travel it is lowest.
- Greece - Greece is most associated with HalfDays travel, which is quite unlike the other countries. Most unlike FUllDays in fact it has 0 and it has relatively more Holidays travel than France/Germany and Canada, but less than Norway.

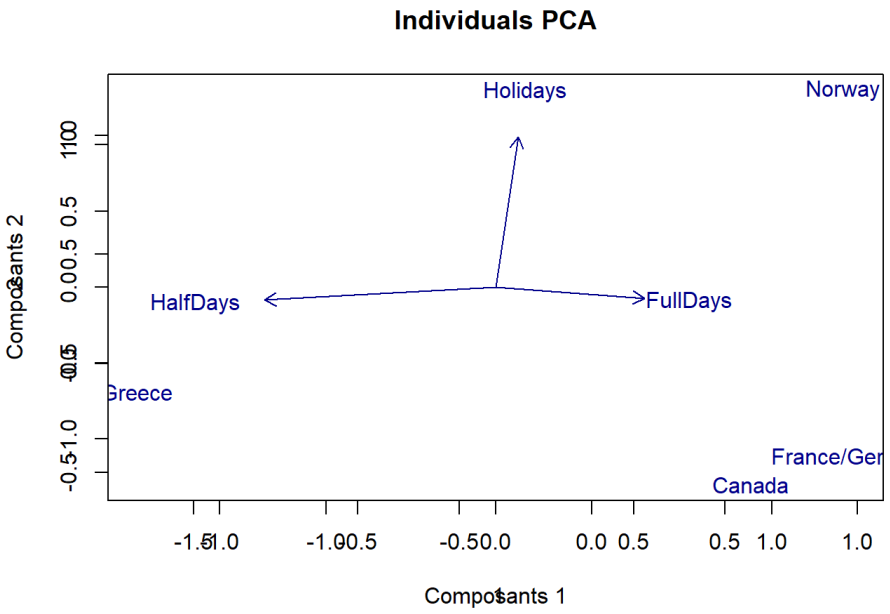
Another Visualization:

```
#Another Visualization
library(amap)
```

```
##
## Attaching package: 'amap'
```

```
## The following object is masked from 'package:psych':
##
##     pca
```

```
ca_map = afc(nums)
plot(ca_map)
```



- Shows drawn on scaling and tells the same story about how travel differs between the 4 country groups.