# Principal Component Analysis – An analysis on risk factors of cervical cancer

Advanced Business Data Analysis

Marion Bonnard
Student – X19161964
Higher Diploma in Data Analytics
National College of Ireland

## I. Introduction

This report is a factor analysis of data collected from a medical study on the risk factors of cervical cancer. By using the Principal Component Analysis technique, the goal is to determine the main components that explain the variance of the sample by reducing the number of variables into dimensions. The outcome expected is to obtain clusters of risk factors that could lead to cervical cancer diagnosis. To reach the result, we first need to inspect the dataset and make sure it is workable, then we want to understand the relationship between each variables; and finally, to explain the sample variance, a Principal Component Analysis has to be performed.

## II. Background

### A. Factor Analysis

A factor analysis is "used to reduce large number of variables into fewer number of factors" (Statistic Solutions, n.d.). What it does is that it selects the most correlated variables and groups them together. There are two types of factor analysis: Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). CFA is a quantitative method used to discover if the data analyzed fit into a hypothetic model (Connelly, 2019). EFA examines the output from the correlations between the variables and highlight groups also called factors. This type of factor analysis uses the Principal Component Analysis technique to define the factors that will explain the variance (Connelly, 2019).

### B. Principal component analysis

Principal component analysis or PCA is a technique used when conducting an EFA. As previously mentioned, this technique is a variable reduction that emphasize the strong patterns in a dataset. (Powell, n.d.) The goal is to reduce a large set of variables into smaller dimensions called principal components to find patterns in the data (Raschka, 2015) . PCA uses the Eigenvalue to indicate the significance of each component and "how much variance is explained by a given factor" (Connelly, 2019).

### C. Cervical cancer study

The chosen dataset has been found in a study from the university of Porto (Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes, 2017). The objective of this study is to prevent cervical cancer using Transfer Learning and other techniques. One of the techniques applied for that study was using the "Cervical cancer (Risk Factors) dataset" from the University Hospital of Caracas in Venezuela where 850 patients answered a survey on demographic, habits, and personal medical records. In the purpose of our analysis, we used the dataset for a different purpose in order to complete a factor analysis.

## III. Data Description

As previously mentioned in the background, the dataset called "Cervical cancer (Risk Factor) Data Set" was gathered at the "Hospital Universitario de Caracas" in Venezuela. It includes patients' demographic information, habits, and historic medical records; due to privacy reasons some values are missing (Center for Machine Learning and Intelligent Systems, 2017). The sample of this multivariate dataset is composed by 858 records and counts 36 variables which 24 are Boolean values and 12 are numeric data. It is recommended by Comrey and Lee (1992) to have a sample over 500 to obtain an adequate outcome (Comrey, A., & Lee, H., 1992). In this case, the sample size selected is considered as "very good" and will provide satisfactory insights.

A data cleansing (R Code in appendix 4) has been performed to be able to work with relevant data and to conduct the Principal Component Analysis. First, the 3622 missing values had to be replaced using RStudio programming by replacing the missing values with the median or the mean of the variables and by 0 where required. For values such as number of years, number of partners using the median is the most appropriate to avoid decimal numbers. The 0 value has been used to replace missing values for variables related to "time since first or last diagnosis" when the patient was never diagnosed of a STD. Then, we had to exclude columns that included Boolean values (24 columns), since the PCA must be performed on numeric values. Finally, the dataset has been exported onto a CSV file that will be the base of the updated dataset. The finalized data description containing the 12 variables can be found on appendix 1.

## IV. Methodology and Calculations

Fabrigar et al. indicate the steps to complete a factor analysis. On the first step, we must decide on the size and the nature of the sample and the type of variable to include. Second step, we must define the type of factor analysis to be used. Third step is to select a "specific procedure" where the data will suit the model. Fourth step, we must define the number of factors to include. Fifth step, a rotating method must be chosen (Fabrigar, L. R. et al., 1999).

The methodology chosen for this report followed partly the steps proposed by Fabrigar et al. After having selected and cleaned the dataset to include relevant variables we chose to use EFA to conduct the analysis with the PCA technique as it is the topic of this report. The updated dataset was used to process a PCA with SPSS software with the selected 12 variables (Figure 1).

When processing the data through SPSS, we selected the option to display a Correlation Matrix (to understand the interrelationships between the variables), the unrotated factor solution and the Scree plot (Figure 2). Contrary to what Fabrigar et al. suggested on fourth step, we did not define the number of factors to extract however we selected Eigenvalues parameter. Factor rotation is a method that helps to interpret clearly what variables are loaded onto a factor (UCLA, n.d.). In our case the oblique rotation called "Direct Oblimin" was used to run the model as it allows to have correlations between factors. (Fabrigar, L. R. et al., 1999) Detailed steps of the process on SPSS are developed in appendix 2.

*Figure 1 – Variables selection on SPSS*

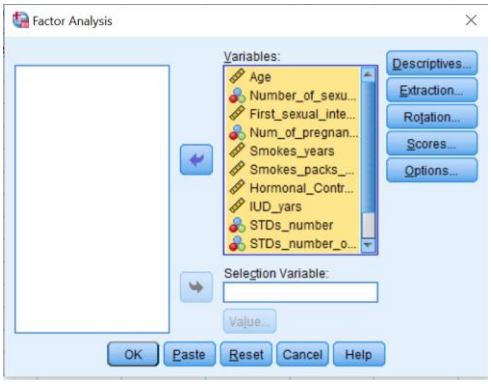
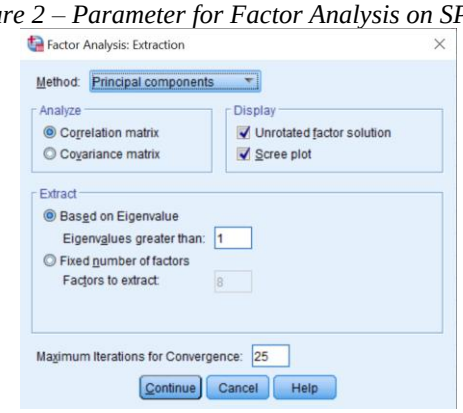



*Figure 2 – Parameter for Factor Analysis on SPSS*




## V. RESULTS

### A. Descriptive Statistics

Starting the analyses with descriptive statistics (Table 1) helps to understand the way the data are distributed when looking at the standard deviation. Moreover, we can observe the mean for each variable; the average age is almost 27 years old; the participants had their first sexual intercourse on average at 17 years old and the sample counts a mean of 2 pregnancies. We can see that not all patients had STD with a mean of 0.15 of number of STD and the smokers would smoke on average 1.2 years.

*Table 1 – Descriptive Statistics*

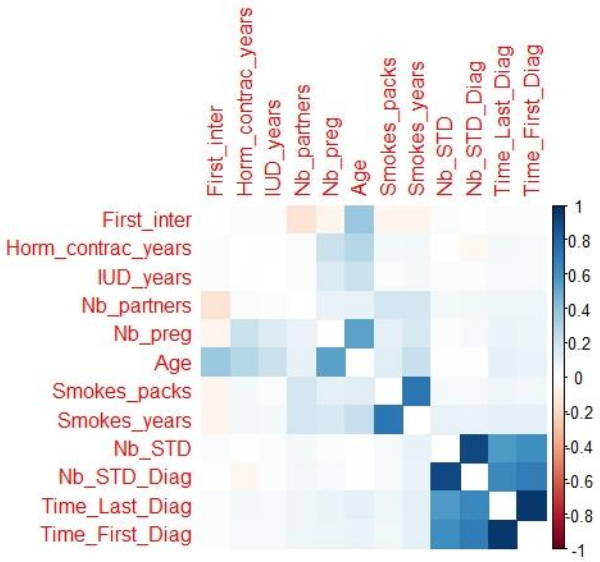
**Descriptive Statistics**

| | Mean | Std. Deviation | Analysis N |
|---|---|---|---|
| Age | 26.8205 | 8.49795 | 858 |
| Number_of_sexual_partners | 2.5117 | 1.64476 | 858 |
| First_sexual_intercourse | 16.9953 | 2.79188 | 858 |
| Num_of_pregnancies | 2.2576 | 1.40098 | 858 |
| Smokes_years | 1.2197 | 4.05788 | 858 |
| Smokes_packs_year | .4531 | 2.20966 | 858 |
| Hormonal_Contraceptive_years | 2.2564 | 3.51908 | 858 |
| IUD_yars | .5148 | 1.80559 | 858 |
| STDs_number | .1550 | .52962 | 858 |
| STDs_number_of_diagnosis | .0874 | .30254 | 858 |
| STDs_time_since_first_diagnosis | .5082 | 2.38833 | 858 |
| STDs_time_since_last_diagnosis | .4814 | 2.29712 | 858 |

### B. Correlations

The Correlation Matrix in appendix 3, highlights a few variables that are correlated. These interrelated variables are "Age" with "First sexual intercourse" and "Number of pregnancies", also the number of years smoking is strongly correlated with the number of packs per year and finally "STDs_number" is strongly and positively correlated with "STD_number_of_diagnosis,"STD_time_since_first_diagnosis" and "STD_time_since_last_diagnosis". The correlation plot shown in figure 3, highlight in dark blue the above-mentioned variables that are positive-strongly intercorrelated. (R Code in appendix 5)

*Figure 3 - Correlation plot*




### C. Communalities

The communalities table (table 2) or common variances, are showing values close to 1, which means that the extracted

factors explain the variance better than individual variable (UCLA, n.d.).

*Table 2 - Communalities*

**Communalities**

| | Initial | Extraction |
|---|---|---|
| Age | 1.000 | .812 |
| Number_of_sexual_partners | 1.000 | .329 |
| First_sexual_intercourse | 1.000 | .833 |
| Num_of_pregnancies | 1.000 | .639 |
| Smokes_years | 1.000 | .840 |
| Smokes_packs_year | 1.000 | .842 |
| Hormonal_Contraceptive_years | 1.000 | .287 |
| IUD_yars | 1.000 | .212 |
| STDs_number | 1.000 | .741 |
| STDs_number_of_diagnosis | 1.000 | .824 |
| STDs_time_since_first_diagnosis | 1.000 | .842 |
| STDs_time_since_last_diagnosis | 1.000 | .799 |

Extraction Method: Principal Component Analysis.

## D. *Total Variance Explained*

The Total Variance Explained in table 3 has generated four components that explain 67% of the variance. The first component has the highest contribution and explains almost a third of the variance with 27%. The Eigenvalues selected in the parameter had to be greater than 1, therefore we can see the Eigenvalues going from 1.1 to 3.2 for the four extracted components. In this case we have reduced the original variables to 4 dimensions.

*Table 3 – Total Variance Explained*

**Total Variance Explained**

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings[a] |
|---|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total |
| 1 | 3.248 | 27.070 | 27.070 | 3.248 | 27.070 | 27.070 | 3.212 |
| 2 | 2.068 | 17.229 | 44.299 | 2.068 | 17.229 | 44.299 | 1.777 |
| 3 | 1.564 | 13.032 | 57.332 | 1.564 | 13.032 | 57.332 | 1.901 |
| 4 | 1.118 | 9.318 | 66.649 | 1.118 | 9.318 | 66.649 | 1.265 |
| 5 | 1.000 | 8.331 | 74.980 | | | | |
| 6 | .881 | 7.342 | 82.323 | | | | |
| 7 | .754 | 6.286 | 88.609 | | | | |
| 8 | .675 | 5.626 | 94.234 | | | | |
| 9 | .304 | 2.534 | 96.768 | | | | |
| 10 | .262 | 2.181 | 98.949 | | | | |
| 11 | .094 | .782 | 99.731 | | | | |
| 12 | .032 | .269 | 100.000 | | | | |

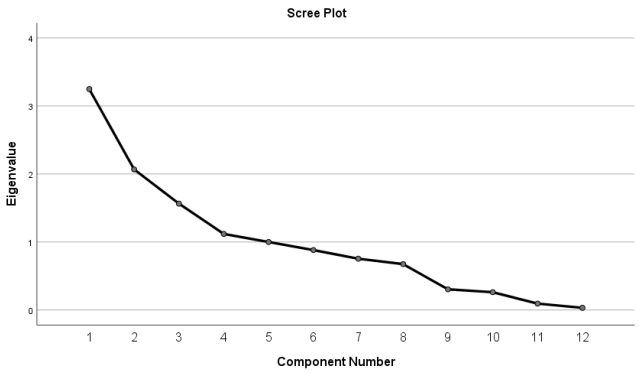
Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

## E. *Scree Plot*

The Scree Plot (Figure 4) is a representation of the Eigenvalue including the component. It confirms that the model has correctly selected four components that have an Eigenvalue higher than 1.

*Figure 4 – Scree Plot*




## F. *Matrices*

The Component Matrix (Table 4) shows the loadings before the rotation, it displays the correlations between the variables and the components. We can already see a trend with a group of STD related data, however the second component shows a moderately high correlation with "age" and years or quantity smoked data but not easily distinguished with other variables, the third component is moderately related to "age" and "age of first sexual intercourse" and the fourth component is strongly linked with the "age of first sexual intercourse".

The Structure Matrix (Table 5) indicates the correlations between variables and factors after the rotation was executed (IBM, n.d.). We can observe a clearer distinction and stronger correlations between variables and components. Especially for component 2 were we can easily distinguish that "age" and "number of pregnancies" are related to the second component. The rotation has also clarified that "Smokes_years" and "Smokes_packs_year" are not related to component 2 anymore, however they are strongly and negatively correlated with component 3. And a stronger correlation between "Age of First sexual intercourse" is now confirming that this variable is related to component 4.

*Table 4 – Component Matrix*

**Component Matrix**

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Age | | .685 | .553 | |
| Number of sexual partners | | | -.318 | -.363 |
| First sexual intercourse | | | .517 | .746 |
| Num of pregnancies | | .598 | .357 | -.359 |
| Smokes years | | .677 | -.512 | |
| Smokes packs year | | .632 | -.579 | |
| Hormonal Contraceptive years | | .359 | .324 | |
| IUD_vars | | | | -.321 |
| STDs number | .835 | | | |
| STDs number of diagnosis | .881 | | | |
| STDs time since first diagnosis | .907 | | | |
| STDs time since last diagnosis | .886 | | | |

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

*Table 5 – Structure Matrix*

**Structure Matrix**

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Age | | .765 | | .408 |
| Number of sexual partners | | | -.312 | -.478 |
| First sexual intercourse | | | | .908 |
| Num of pregnancies | | .794 | | |
| Smokes years | | | -.916 | |
| Smokes packs year | | | -.912 | |
| Hormonal_Contraceptive_years | | .533 | | |
| IUD_vars | | .434 | | |
| STDs number | .857 | | | |
| STDs number of diagnosis | .905 | | | |
| STDs_time_since_first_diagnosis | .916 | | | |
| STDs_time_since_last_diagnosis | .890 | | | |

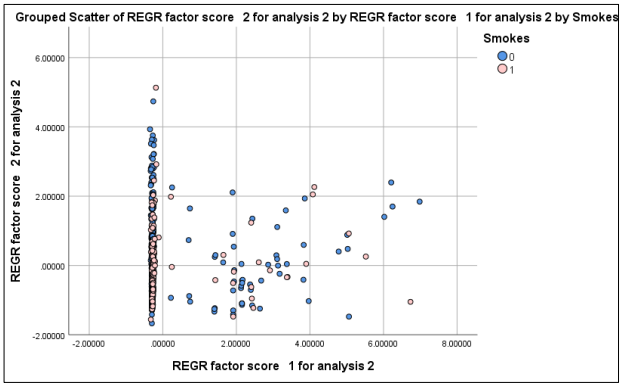
Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

## G. *Scatter plot*

The plot shown in figure 5, shows the data spread according to component 1 on the X axis and component 2 on the Y axis. The binary value used to differentiate the data is whether the person smokes (1) or not (0). As we can see there is a higher number of non-smokers related to component 2. As seen in the Structure Matrix, component 2 includes age and number of pregnancies, therefore it implies that when there are more pregnancies there are less smokers.

*Figure 5 – Scatter plot*




## VI. Discussion and Analysis of Results

The results obtained using Principal Component Analysis technique to reduce the variables into component concluded that 4 principal components explain 67% of the variance. These four components were then extracted to process the rotation that, as mentioned in the methodology, clearly identified the variables linked to each component.
When looking at the components and with the help from the variables linked to them, we can interpret the following meaning for each component:

1. Component 1 → STDs
2. Component 2 → Pregnancies
3. Component 3 → Smoking habits
4. Component 4 → Age of first sexual intercourse

The components have been translated into clusters related to health, habits, demographic and sexual activity. Therefore, we could conclude that these four areas should be closely monitored as they could be considered as risk factor of cervical cancer.

On the other hand, the dataset did not confirm that all of the 858 patients had cervical cancer but for the purpose of the study and due to the name of the dataset title "Cervical cancer (Risk Factors) Data Set", we considered that they all had cervical cancer. Moreover, with the 3622 missing values, the dataset was not complete and by replacing values with the mean or median it probably did not reflect the reality. Furthermore, the strongly negative correlation of number of years smoking did not necessarily mean that the less they smoke, the more risk there is to get cervical cancer; however it reflected that the sample did not have many patients that were smoking for many years and did not consume a lot of packs per year.

For further analysis, it is advised to select an appropriate dataset with complete values that are very well defined and explains clearly what common issue is connecting all the patients. Moreover, it is recommended to use the above 4

clusters for future research on another collection of data in the medical sector.

## VII. Conclusion

To conclude, we have managed to examine a dataset with missing values and did a cleansing to obtain appropriate data. Next, we found some correlations between values thanks to the Correlation Matrix. Then with the PCA method we managed to reduce variables into 4 components that we converted into more generic clusters to provide recommendations for other studies in the medical field.

## VIII. References


[1] Center for Machine Learning and Intelligent Systems. (2017, 03 03). *Marchine Learning Repository*. Retrieved from UCI.edu: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

[2] Comrey, A., & Lee, H. (1992). *A first course in factor analysis.* Hillsdale: Erlbaum. Retrieved from https://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=1390&context=jmasm

[3] Connelly, L. M. (2019, October). *What is Factor Analysis?* Retrieved May 8, 2020, from Ebscohost.com: http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=a9h&AN=139018061&site=eds-live&scope=site

[4] Fabrigar, L. R. et al. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, pp. 272-299.

[5] IBM. (n.d.). *Pattern Matrix and Structure Matrix Definition in SPSS FACTOR output*. Retrieved from IBM.com: https://www.ibm.com/support/pages/pattern-matrix-and-structure-matrix-definition-spss-factor-output

[6] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. (2017, June 20). Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. *Pattern Recognition and Image Analysis*, pp. 243-250.

[7] Powell, V. (n.d.). *Principal Component Analysis*. Retrieved from steosa.io: https://setosa.io/ev/principal-component-analysis/

[8] Raschka, S. (2015, January 27). *Principal Component Analysis*. Retrieved from sebastianraschka.com: https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html

[9] Statistic Solutions. (n.d.). *Factor Analysis*. Retrieved from Satisticssolutions.com: https://www.statisticssolutions.com/factor-analysis-sem-factor-analysis/

[10] UCLA. (n.d.). *PRINCIPAL COMPONENTS (PCA) AND EXPLORATORY FACTOR ANALYSIS (EFA) WITH SPSS*. Retrieved from https://stats.idre.ucla.edu/: https://stats.idre.ucla.edu/spss/seminars/efa-spss/

# IX. APPENDICES

*Appendix 1: Finalised data description of "Cervical cancer (Risk Factor) Data Set"*

| **Variable** | **Age** | **Number_of_sexual_partners** | **First_sexual_intercourse** |
|---|---|---|---|
| Description | Age of the patient | Number of sexual partners | Age of first sexual intercourse |
| **Variable** | **Num_of_pregnancies** | **Smokes_years** | **Smokes_packs_year** |
| Description | Number of pregnancies | Number of years smoking | Number of packs per year |
| **Variable** | **Hormonal_Contraceptive_years** | **IUD_years** | **STDs_number** |
| Description | Number of years using hormonal contraception | Number of years using IUD | Number of STDs |
| **Variable** | **STDs_number_of_diagnosis** | **STDs_time_since_first_diagnosis** | **STDs_time_since_last_diagnosis** |
| Description | Number of STDs diagnosis | Time since first STD diagnosis | Time since las STD diagnosis |

*Appendix 2: Detailed steps on how to execute a PCA on SPSS*

1. Select Analyse – Dimension reduction – Factor…
2. Select the items to be analyse into variables
3. On Descriptive – Select Coefficients in Correlation Matrix
4. On Extraction – Select Correlation Matrix, Screeplot and choose the criteria of factors. In our case we chose Eigenvalues that are greater than 1.
5. Rotation – Select Direct Oblimin Method and display the Rotated solution and the Loading plots
6. Options – Suppress small coefficients below 0.3 (this will improve the visualization when analysing the results)

*Appendix 3: Correlation Matrix*

## Correlation Matrix

| | | Age | Number_of_sexual_partners | First_sexual_intercourse | Num_of_pregnancies | Smokes_years | Smokes_packs_year | Hormonal_Contraceptive_years | IUD_yars | STDs_number | STDs_number_of_diagnosis | STDs_time_since_first_diagnosis | STDs_time_since_last_diagnosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Correlation | Age | 1.000 | .086 | .369 | .526 | .217 | .131 | .277 | .206 | -.001 | -.002 | .085 | .100 |
| | Number_of_sexual_partners | .086 | 1.000 | -.146 | .077 | .178 | .176 | .020 | .006 | .041 | .053 | .058 | .064 |
| | First_sexual_intercourse | .369 | -.146 | 1.000 | -.056 | -.058 | -.056 | .008 | -.025 | .016 | -.013 | .010 | .015 |
| | Num_of_pregnancies | .526 | .077 | -.056 | 1.000 | .175 | .096 | .209 | .144 | .011 | .035 | .073 | .084 |
| | Smokes_years | .217 | .178 | -.058 | .175 | 1.000 | .724 | .049 | .038 | .091 | .082 | .103 | .108 |
| | Smokes_packs_year | .131 | .176 | -.056 | .096 | .724 | 1.000 | .040 | .016 | .032 | .032 | .054 | .056 |
| | Hormonal_Contraceptive_years | .277 | .020 | .008 | .209 | .049 | .040 | 1.000 | .000 | -.010 | -.037 | .028 | .036 |
| | IUD_yars | .206 | .006 | -.025 | .144 | .038 | .016 | .000 | 1.000 | .016 | .008 | .030 | .034 |
| | STDs_number | -.001 | .041 | .016 | .011 | .091 | .032 | -.010 | .016 | 1.000 | .898 | .606 | .565 |
| | STDs_number_of_diagnosis | -.002 | .053 | -.013 | .035 | .082 | .032 | -.037 | .008 | .898 | 1.000 | .691 | .641 |
| | STDs_time_since_first_diagnosis | .085 | .058 | .010 | .073 | .103 | .054 | .028 | .030 | .606 | .691 | 1.000 | .965 |
| | STDs_time_since_last_diagnosis | .100 | .064 | .015 | .084 | .108 | .056 | .036 | .034 | .565 | .641 | .965 | 1.000 |

*Appendix 4 : R Code for data set cleansing*

```
#Cervical cancer risk factor dataset
Cervdata<-read.csv(file="risk_factors_cervical_cancer.csv", header=TRUE, sep=",")
head(Cervdata)
Cervdata [!complete.cases(Cervdata),]
#
#Data cleansing
#Question marks replaced by NA
Cervdata<-read.csv(file="risk_factors_cervical_cancer.csv", header=TRUE, sep=",", na.strings=c("?"))
Cervdata [complete.cases(Cervdata),]
Cervdata[!complete.cases(Cervdata),]
sum(is.na(Cervdata))
colSums(is.na(Cervdata))
rowSums(is.na(Cervdata))

#Replacing NA values
#
#Replacing NA values for variable "number of sexual partners" by median (to obtain integer for a number of people or age)
sum(is.na(Cervdata$Number.of.sexual.partners))
Medianpartner<-median(Cervdata$Number.of.sexual.partners, na.rm=TRUE)
Medianpartner
Cervdata$Number.of.sexual.partners[is.na(Cervdata$Number.of.sexual.partners)]<-Medianpartner
sum(is.na(Cervdata$Number.of.sexual.partners))
print(Cervdata$Number.of.sexual.partners)

#Replacing NA values for variable "first sexual intercourse" by median (to obtain an integer as it's an age)
sum(is.na(Cervdata$First.sexual.intercourse))
Medianfirstime<-median(Cervdata$First.sexual.intercourse, na.rm=TRUE)
Medianfirstime
Cervdata$First.sexual.intercourse[is.na(Cervdata$First.sexual.intercourse)]<-Medianfirstime
sum(is.na(Cervdata$First.sexual.intercourse))
print(Cervdata$First.sexual.intercourse)

#Replacing NA values for variable "Number of pregnancies" by median
sum(is.na(Cervdata$Num.of.pregnancies))
Medianpreg<-median(Cervdata$Num.of.pregnancies, na.rm=TRUE)
Medianpreg
Cervdata$Num.of.pregnancies[is.na(Cervdata$Num.of.pregnancies)]<-Medianpreg
sum(is.na(Cervdata$Num.of.pregnancies))
print(Cervdata$Num.of.pregnancies)

#Replacing NA values for Smokes Year using the mean since the median will provide 0
sum(is.na(Cervdata$Smokes..years.))
Meansmokeyears<-mean(Cervdata$Smokes..years., na.rm=TRUE)
Meansmokeyears
Cervdata$Smokes..years.[is.na(Cervdata$Smokes..years.)]<-Meansmokeyears
sum(is.na(Cervdata$Smokes..years.))

#Replacing NA values for Smokes ¨Packs/Year using the mean since the median will provide 0
sum(is.na(Cervdata$Smokes..packs.year.))
Meanpacks<-mean(Cervdata$Smokes..packs.year., na.rm=TRUE)
Meanpacks
Cervdata$Smokes..packs.year.[is.na(Cervdata$Smokes..packs.year.)]<-Meanpacks
sum(is.na(Cervdata$Smokes..packs.year.))

#Replacing NA values for years under hormonal contraceptives using the mean
sum(is.na(Cervdata$Hormonal.Contraceptives..years.))
Contraceptyears<-mean(Cervdata$Hormonal.Contraceptives..years., na.rm=TRUE)
Contraceptyears
Cervdata$Hormonal.Contraceptives..years.[is.na(Cervdata$Hormonal.Contraceptives..years.)]<-Contraceptyears
```

```
sum(is.na(Cervdata$Smokes..years.))

#Replacing NA values for years with IUD using the mean
sum(is.na(Cervdata$IUD..years.))
Iudyears<-mean(Cervdata$IUD..years., na.rm=TRUE)
Iudyears
Cervdata$IUD..years.[is.na(Cervdata$IUD..years.)]<-Iudyears
sum(is.na(Cervdata$IUD..years.))

#Replacing NA values for Number of STDs with the mean
sum(is.na(Cervdata$STDs..number.))
STDmedian<-median(Cervdata$STDs..number., na.rm=TRUE)
STDmedian
Cervdata$STDs..number.[is.na(Cervdata$STDs..number.)]<-STDmedian
sum(is.na(Cervdata$STDs..number.))

#Replacing NA values for Number of diagnosis using the median
sum(is.na(Cervdata$STDs..Number.of.diagnosis))
Nodiagnosmedian<-median(Cervdata$STDs..Number.of.diagnosis, na.rm=TRUE)
Nodiagnosmedian
Cervdata$STDs..Number.of.diagnosis[is.na(Cervdata$STDs..Number.of.diagnosis)]<-Nodiagnosmedian
sum(is.na(Cervdata$STDs..Number.of.diagnosis))

#Replacing NA values for Time since First Diagnosis using 0
sum(is.na(Cervdata$STDs..Time.since.first.diagnosis))
Cervdata$STDs..Time.since.first.diagnosis[is.na(Cervdata$STDs..Time.since.first.diagnosis)]<-0
sum(is.na(Cervdata$STDs..Time.since.first.diagnosis))

#Replacing NA values for Time since Last Diagnosis using 0
sum(is.na(Cervdata$STDs..Time.since.last.diagnosis))
Cervdata$STDs..Time.since.last.diagnosis[is.na(Cervdata$STDs..Time.since.last.diagnosis)]<-0
sum(is.na(Cervdata$STDs..Time.since.last.diagnosis))



sum(is.na(Cervdata$))


#Exlude several column that are binary or boolean values (Smokes)
Cervdata<-subset(Cervdata, select= -Smokes)
Cervdata<-subset(Cervdata, select= -Hormonal.Contraceptives)
Cervdata<-subset(Cervdata, select= -IUD)
Cervdata<-subset(Cervdata, select= -STDs)
Cervdata<-subset(Cervdata, select= -STDs.condylomatosis)
Cervdata<-subset(Cervdata, select= -STDs.cervical.condylomatosis)
Cervdata<-subset(Cervdata, select= -STDs.vaginal.condylomatosis)
Cervdata<-subset(Cervdata, select= -STDs.vulvo.perineal.condylomatosis)
Cervdata<-subset(Cervdata, select= -STDs.syphilis)
Cervdata<-subset(Cervdata, select= -STDs.pelvic.inflammatory.disease)
Cervdata<-subset(Cervdata, select= -STDs.genital.herpes)
Cervdata<-subset(Cervdata, select= -STDs.molluscum.contagiosum)
Cervdata<-subset(Cervdata, select= -STDs.AIDS)
Cervdata<-subset(Cervdata, select= -STDs.HIV)
Cervdata<-subset(Cervdata, select= -STDs.Hepatitis.B)
Cervdata<-subset(Cervdata, select= -STDs.HPV)
Cervdata<-subset(Cervdata, select= -Dx.Cancer)
Cervdata<-subset(Cervdata, select= -Dx.CIN)
Cervdata<-subset(Cervdata, select= -Dx.HPV)
Cervdata<-subset(Cervdata, select= -Dx)
Cervdata<-subset(Cervdata, select= -Hinselmann)
Cervdata<-subset(Cervdata, select= -Schiller)
Cervdata<-subset(Cervdata, select= -Citology)
```

```
Cervdata<-subset(Cervdata, select= -Biopsy)

head(Cervdata)

#
#generate the correlation matrix
correlationMat<- cor(Cervdata)
correlationMat

write.csv(Cervdata, file="Cervical data", row.names=FALSE)
write.csv(Cervdata, file="Cervicaldatav2.csv")
```

*Appendix 5: R Code for Correlation Plot*

```
Cdata<-read.csv(file="Cervicaldatav2.csv", header=TRUE, sep=",")
Cdata
head(Cdata)

#generate correlation Matrix
CervicalCorMat<-cor(Cdata)
CervicalCorMat

#Create a Matrix
CorMatrix <- as.matrix(Cdata[,c(1:12)])
head(Cdata.data)

# What is the mean of each of the numeric columns?
round(colMeans(Cdata.data),2)

##Rename the colnames because they are too long
cNames <- c("Age","Nb_partners","First_inter",
        "Nb_preg","Smokes_years","Smokes_packs","Horm_contrac_years",
        "IUD_years","Nb_STD","Nb_STD_Diag",
        "Time_First_Diag","Time_Last_Diag")
colnames(CorMatrix) <- cNames

# Create the correlation matrix - awkward to read in R output
M <- round(cor(CorMatrix), 2)
M

# Create corrplot
corrplot(M, diag = FALSE, method="color", order="FPC", tl.srt = 90)
```