IJCRR
Section: Healthcare
Sci. Journal Impact
Factor: 6.1 (2018)
ICV: 90.90 (2018)

Copyright@IJCRR

Exploratory Data Analysis and ETL with SAS on Hadoop Eco-system with Cervical Cancer Dataset

Cheng Xiaotian, Vinesh Thiruchelvam, Daniel Mago Vistro

School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia.

ABSTRACT

Objective: The main objective of this project is to explore and analyse a secondary dataset which collected from “Hospital Universitario de Caracas” in Caracas, Venezuela.

Methods: The dataset comprises 858 patients’ information relating to demographic information and medical history data. There is a large number of records which are left with blank, which might be intentionally avoided by the patient due to privacy considerations. SAS Studio is utilized in data exploration and data pre-processing. Data cleaning and data transformation are conducted basing on the knowledge gathered in the process of data exploration. Afterwards, the dataset was exported from SAS Studio and uploaded to Hadoop Hortonworks platform for analysing purpose. Lastly, five hypotheses have been explored with the visualization tool of Tableau.

Key Words: Data Management, SAS, Hadoop, Cervical cancer, Tableau, Healthcare

INTRODUCTION

According to the stats from the World Health Organization, cervical cancer is recognized as the fourth most common cancer-causing woman mortality¹. There are many effective methods to prevent and detect cervical cancer in early stage, such as cytology-based cervical cancer screening, visual inspection with Acetic Acid Application (VIA) as an alternative in developing countries, HPV test in cervical screening². However, due to limited infrastructure, lack of skilled medical professionals in place and low public awareness of the risk of this disease, it is still causing high mortality rate in developing countries such as the countries in Africa, Latin American countries and parts of Asia.

There are many factors which will lead to high chance of cervical cancer infection, including but not limited to: (1) Human Papilloma Virus (HPV), 99.7% of the cervical cancer infection is linked with HPV infection (Chichareon, 1998). (2) sexual partner numbers (3) individual immunity (4) count of birth deliveries (5) tobacco and (6) birth controlling device usage².

In this report, will use one of the secondary dataset collected from a Latin American country, “Hospital Universitario de Caracas” in Caracas, Venezuela. There are 858 records in the

dataset, which comprises some demographic information, and historical medical records. There is a large number of records which are left with blank, which might be intentionally avoided by the patient due to privacy considerations. First, will study about the privacy and security issue in the healthcare industry and some techniques to preserve data security in Section 2. Then, data is imported into SAS for data exploration and pre-processing, this will be talked about in Section 3, 4 and 5. Next, to leverage big data computational power, the dataset is pumped into Hadoop and stored as Optimized Row Columnar (ORC) file, this is shown in Section 6. In Section 7, the dataset from Hadoop will be linked with the visualization tool, Tableau to analysis five hypotheses. Finally, in Section 8, there is a discussion and conclusion part.

RELATED WORKS

Privacy and security concern in healthcare

The digitalization of healthcare data is becoming a prevalent approach for major hospitals and clinics. There are evident benefits such as reduction of cost, and rapid delivery of healthcare services (physicians are not bond to location any more), etc. In the meantime, it has also arisen public

Corresponding Author:

Cheng Xiaotian, School of Computing, Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia.
Email: cxtalex@foxmail.com

ISSN: 2231-2196 (Print)

ISSN: 0975-5241 (Online)

Received: 11.06.2020

Revised: 09.08.2020

Accepted: 23.09.2020

Published: 06.10.2020

concern in regards to privacy and security issue. According to a survey conducted in the US (2013), 69% of the patients hold privacy concern about their medical record since healthcare providers deployed digital record³. In response to this, they will either give faulty information or avoid giving some of the information. This will jeopardize the veracity of the source of data and further impact prediction modelling accuracy.

The issue of privacy and security in healthcare has been of concern dated back since the 20th century. In 1996, the US has issued the Health Insurance Portability and Accountability Act (HIPAA), to safeguard the medical record privacy with the power of federal law⁴. It requires health entities to set up proper administrative and technical protections to ensure data confidentiality and security. Besides, it also permits the individual to require full disclosure of its personal healthcare information from healthcare entities. Besides this, there are many other acts published in other regions and countries, for example, the Data Protection Directive in the EU. Despite all these, even though sometimes data are removed with identifiers (de-identified), it is still possible to locate individual unique information with some external dataset to assist⁵. In case of sensitive information like healthcare data is leaked out, it will cause a huge loss to enterprises as well as individuals.

On the other hand, with the digitalization of medical information, medical data can be transferred and shared with other professionals easily, this has facilitated the improvement of the medical industry. This has brought benefit to the patient for better diagnose and stimulated the breakthroughs in the medical field.

In consideration of this, this is a dilemma to find a balance between the protection of privacy and digitalization in the healthcare industry.

Recent Development in Healthcare

(1) Cloud Computing

Due to the nature of healthcare-related data, it is the cross-boundary area between computer technology, biology and biomedical science. Hence, datasets of the healthcare field are usually characterized as vast volume, various type (including structured and unstructured). To better assist with this, big data and cloud computing is advocated as a solution⁶. Cloud computing is enjoying its inherited advantages such as the extendable capacity of storage and fast processing, reduced cost for maintenance, convenience for data sharing with different research centres, etc. However, the adoption of cloud is always accompanied by the concern in privacy and security.

(2) Mobile Device Applications

The use of smartphone and other mobile devices are getting ubiquitous. This allows the hospital to provide continuous monitoring over to the patients and can diagnose and delivery interventions remotely. At the same time, it enables doctors to share condition and graphical information about the patient with other specialists from whom they can communicate, and exchange ideas then can get informed with the optimal treatment choice across the world. This is an effective method for driving improvement in the medical field. The research did in 2013 shows that more than half of the healthcare professionals are working either from a smartphone or tablet. However, despite all the merits it conveys, 69% of the patients are showing privacy and security concern about utilizing mobile devices for healthcare information delivery.

Approaches for privacy-preserving

The rapid development and application of new technologies in healthcare surely improved the healthcare services performance, but potentially it also increased the threat of privacy and security breaches. To prevent improper usage of this sensitive data, the healthcare entity could deploy some physical and electronic methods to enhance the security level.

(1) Encryption

Encryption by definition is to transform the plain understandable text into ciphertext, which cannot be understood by people without authorization. Encryption is based on symmetric and asymmetric algorithms. For the symmetric approach, a secret key is utilized to generate cipher text and vice versa, when decrypting cipher text back to plain text. While for the asymmetric approach, a pair of security key is generated. The decryption key is kept in secret and the public key will be sent to whoever needs the authorization to access data.

(2) Authorization control

Authorization control is referring to controlling users' access to the availability of the data. Under the different authorization, the user will be granted different access authorize and they will be under the governess when retrieve or modify information in the database. By this way, this technique could protect the integrity and confidentiality of data source. There are many technologies currently been pervasively used:

- a) Biological method: by scanning a fingerprint, iris, facial expression, or recognition of voice, signature etc.
- b) Mandatory Access Control (MAC): This controls the information disclosure extent user can access by designating security roles. It will classify the sensitivity level of the information and classify authorization of the subject, then assign authorization to the subject. This is a highly restricted and confidential way, which is being used for maintaining military information as well.

- c) Password: This is a simple and most widely used approach.

(3) Data Masking

Data masking is also called de-identifying, which means to remove the individual unique identifiers such as name, or security number from the dataset. A most common technique is K-anonymity. Which means when releasing a record, it must have at least (k-1) other records whose values are indistinct even with external resource⁸. By doing this, k-anonymity makes it difficult to recognize the identity of the individuals in the dataset. However, a drawback of this technique is it will face the problem when anonymizing high dimensional dataset³.

(4) Laws and Regulations

Healthcare providers are under compulsory to conform to rules and regulations. With the revolution of transiting paper medical record to paperless electronic medical record, this regulatory compliance has become a major focus from the public³. There are many laws and regulations in place, such as the Health Insurance Portability and Accountability Act (HIPAA) of the US and the constitution published by Brazil in 1988. Failure to conform to these regulations will lead to civil and criminal penalties.

Initial Data Exploration

Type of Each Attribute

According to the metadata, there are 36 attributes in the dataset, types of each attribute are identified as follows (Table 1):

Table 1: Types of each attribute

No	Variables	Attributes Types
1	Age	Ratio
2	Number of sexual partners	Ratio
3	First sexual intercourse	Ratio
4	Number of pregnancies	Ratio
5	Smokes	Nominal
6	Smokes (years)	Ratio
7	Smokes (packs/year)	Ratio
8	Hormonal Contraceptives	Nominal
9	Hormonal Contraceptives(years)	Ratio
10	IUD	Nominal
11	IUD (years)	Ratio
12	STDs	Nominal
13	STDs (number)	Ratio
14	STDs: condylomatosis	Nominal
15	STDs: cervical condylomatosis	Nominal

16	STDs: vaginal condylomatosis	Nominal
17	STDs: vulvo-perineal condylomatosis	Nominal
18	STDs: syphilis	Nominal
19	STDs: pelvic inflammatory diseases	Nominal
20	STDs: genetical herpes	Nominal
21	STDs: molluscum contagiosum	Nominal
22	STDs: AIDS	Nominal
23	STDs: HIV	Nominal
24	STDs: Hepatitis B	Nominal
25	STDs: HPV	Nominal
26	STDs: Number of diagnoses	Ratio
27	STDs: Time since the first diagnosis	Ratio
28	STDs: Time since the last diagnosis	Ratio
29	Dx: Cancer	Nominal
30	Dx: CIN	Nominal
31	Dx: HPV	Nominal
32	DX	Nominal
33	Hinselmann	Nominal
34	Schiller	Nominal
35	Citology	Nominal
36	Biopsy	Nominal

Dataset Understanding

Importing Dataset

Upload the dataset onto SAS server, then import data into Library with codes as below. The imported dataset is named as "cervical_cancer", Fig 1. Output data, as well as table of contents, are shown in Fig 2 and Fig 3. The variables are shown in Fig 4.

PROC MEANS

PROC MEANS procedure is a descriptive statistic analyzing tool for understanding data summarization. By default, the output report shows the number of observations, mean, standard deviation, min and max of the group of values. As in Fig.5, it can be further extended to get the output of other statistical analysis values⁹.

From below Fig.7, we can tell the sample size, central tendency (mean, median), range (min, max), variety (standard deviation), dispersion (1st quartile and 3rd quartile), and also the number of missing value. This gives a rough view of the data distribution and representative value, however, for nominal attributes it does not make too much sense to calculate the mean, median, etc. Will use PROC FREQ and PROC UNIVARIATE to further understand the frequency and dispersion for categorical variables later. Another finding is,

there is a large amount of incomplete value for information related to the historical medical record, which will impact model performance. Hence, missing values are expected to be handled before analyzing phase.

PROC FREQ

PROC FREQ procedure provides an analysis function to understand the frequency, percentage, cumulative frequency, as well as cumulative percentage of each variable. It can also be utilized to stratify data and plot cross-table between two variables. Below is the code for getting PROC FREQ output.

PROC UNIVARIATE

PROC UNIVARIATE procedure provides descriptive statistic measures for understanding numeric data. Comparing with PROC MEANS, UNIVARIATE can also indicate skewness and kurtosis, frequency table, and also extreme values. It can also be used together with the VAR statement and plot statistic summary. Below is the code for getting UNIVARIATE output for all variables.

Summarization of Each Variable

1). Age

As can see from Fig.6, total observation number of *Age* is 858, there is no missing value in *Age*. Mean, Median and mode are provided as 26.82, 25, and 23 respectively. Slightly positive skewness and positive kurtosis tendency are detected. Standard deviation (8.50) is less than the mean (26.82). Interquartile range ($Q3 - Q1$) is 12, hence the upper boundary for a suspected outlier is 50 ($Q3 + 1.5 \times IQR$), and lower boundary for suspected outlier is 2 ($Q1 - 1.5 \times IQR$).

Fig.7 depicted distribution histogram of *Age*. The most frequent value appears around 22.5, and the distribution is showing slightly right-skewed and many data dispersed along X-axis.

Fig.8 is the boxplot of *Age*. The box and whisker plotted out the range (bottom and top whisker), Interquartile range (between 2nd and 4th whisker), as well as median (3rd whisker) and mean (diamond). Besides, outliers are also plotted as dots above the upper whisker.

The findings of visualization graphs match up with what can tell from the descriptive output.

PROC MEANS Procedure											
Variable	N	Mean	Median	Std Dev	Minimum	Maximum	Range	Lower 95%	Upper 95%	Lower Quantile	Upper Quantile
Age	858	26.8205128	26.8205128	8.49794807	13.0000000	84.0000000	71.0000000	26.2519244	27.3899323	20.0000000	32.0000000
Number of sexual partners	832	2.52764423	2.0000000	1.66776048	0.0000000	27.0000000	27.0000000	2.4141554	2.6411331	2.0000000	3.0000000
Smokes	846	0.1696261	0.0000000	0.3826796	0.0000000	1.0000000	1.0000000	0.1217754	0.1833889	0.0000000	0.0000000
Smokes (packs/year)	846	1.2197214	0.0000000	4.0890169	0.0000000	37.0000000	37.0000000	0.5436243	1.4958195	0.0000000	0.0000000
Smokes (packs/year)	846	0.4311440	0.0000000	2.2660980	0.0000000	37.0000000	37.0000000	0.3327996	0.6034883	0.0000000	0.0000000
Humoral Contraceptives (years)	750	0.6413333	1.0000000	0.4796292	0.0000000	1.0000000	1.0000000	0.6898933	0.9797564	0.0000000	1.0000000
Humoral Contraceptives (years)	750	2.2644192	0.0000000	3.7642535	0.0000000	30.0000000	30.0000000	1.9885840	2.5242548	0.0000000	3.0000000
STDs	753	0.1696261	0.0000000	0.3826796	0.0000000	1.0000000	1.0000000	0.1217754	0.1833889	0.0000000	0.0000000
STDs (number)	753	0.1786268	0.0000000	0.6109268	0.0000000	4.0000000	4.0000000	0.1342177	0.2168319	0.0000000	0.0000000
STDs condylomatosis	753	0.0684229	0.0000000	0.2347162	0.0000000	1.0000000	1.0000000	0.0416413	0.0752246	0.0000000	0.0000000
STDs cervical condylomatosis	753	0.0053121	0.0000000	0.0727395	0.0000000	1.0000000	1.0000000	0.00108366	0.0105158	0.0000000	0.0000000
STDs vaginal condylomatosis	753	0.0053121	0.0000000	0.0727395	0.0000000	1.0000000	1.0000000	0.00108366	0.0105158	0.0000000	0.0000000
STDs vulvo-perineal condylomatosis	753	0.0053121	0.0000000	0.0727395	0.0000000	1.0000000	1.0000000	0.00108366	0.0105158	0.0000000	0.0000000
STDs syphilis	753	0.0239644	0.0000000	0.1528626	0.0000000	1.0000000	1.0000000	0.0129693	0.0340395	0.0000000	0.0000000
STDs pelvic inflammatory disease	753	0.0013200	0.0000000	0.0364420	0.0000000	1.0000000	1.0000000	-0.0127900	0.0039351	0.0000000	0.0000000
STDs genital herpes	753	0.0013200	0.0000000	0.0364420	0.0000000	1.0000000	1.0000000	-0.0127900	0.0039351	0.0000000	0.0000000
STDs molluscum contagiosum	753	0.0013200	0.0000000	0.0364420	0.0000000	1.0000000	1.0000000	-0.0127900	0.0039351	0.0000000	0.0000000
STDs AIDS	753	0.0013200	0.0000000	0.0364420	0.0000000	1.0000000	1.0000000	-0.0127900	0.0039351	0.0000000	0.0000000
STDs HIV	753	0.0239644	0.0000000	0.1528626	0.0000000	1.0000000	1.0000000	0.0129693	0.0340395	0.0000000	0.0000000
STDs Hepatitis B	753	0.0013200	0.0000000	0.0364420	0.0000000	1.0000000	1.0000000	-0.0127900	0.0039351	0.0000000	0.0000000
STDs HPV	753	0.0026580	0.0000000	0.0515025	0.0000000	1.0000000	1.0000000	-0.0102085	0.0053495	0.0000000	0.0000000
Number of diagnosis	858	0.0514236	0.0000000	0.3025447	0.0000000	3.0000000	3.0000000	0.0071401	0.1768551	0.0000000	0.0000000
Dx Cancer	858	0.0209790	0.0000000	0.1433876	0.0000000	1.0000000	1.0000000	0.0113704	0.0305876	0.0000000	0.0000000
Dx CIN	858	0.0148856	0.0000000	0.1019392	0.0000000	1.0000000	1.0000000	0.0036859	0.0173251	0.0000000	0.0000000
Dx HPV	858	0.0209790	0.0000000	0.1433876	0.0000000	1.0000000	1.0000000	0.0113704	0.0305876	0.0000000	0.0000000
Dx	858	0.0279729	0.0000000	0.1658080	0.0000000	1.0000000	1.0000000	0.0159187	0.0392714	0.0000000	0.0000000
Histemann	858	0.0407925	0.0000000	0.1976246	0.0000000	1.0000000	1.0000000	0.0273303	0.0640548	0.0000000	0.0000000
Schiller	858	0.0824211	0.0000000	0.3898923	0.0000000	1.0000000	1.0000000	0.0514234	0.1056468	0.0000000	0.0000000
Citology	858	0.0812821	0.0000000	0.2207011	0.0000000	1.0000000	1.0000000	0.0364836	0.0660705	0.0000000	0.0000000
Biopsy	858	0.0841028	0.0000000	0.2495784	0.0000000	1.0000000	1.0000000	0.0476807	0.0959245	0.0000000	0.0000000

Figure 5: Output of PROC MEANS Procedure.

The UNIVARIATE Procedure Variable: Age			
Moments			
N	858	Sum Weights	858
Mean	26.8205128	Sum Observations	23012
Std Deviation	8.49794807	Variance	72.2151213
Skewness	1.39427877	Kurtosis	4.77857515
Uncorrected SS	679082	Corrected SS	61888.359
Coeff Variation	31.68451	Std Error Mean	0.29011516

Basic Statistical Measures			
Location		Variability	
Mean	26.82051	Std Deviation	8.49795
Median	25.00000	Variance	72.21512
Mode	23.00000	Range	71.00000
		Interquartile Range	12.00000

Quantiles (Definition 5)	
Level	Quantile
100% Max	84
99%	50
95%	41
90%	37
75% Q3	32
50% Median	25
25% Q1	20
10%	18
5%	16
1%	15
0% Min	13

Figure 6: Descriptive Statistic Values.

Figure 7: Histogram of Age.

Figure 8: Boxplot of Age.

2) Number of sexual partners

By applying PROC UNIVARIATE and PROC MEANS, below outputs of Fig.9 and Fig.10 are acquired. From Fig.9, descriptive statistics values are computed. The mean, median and mode are roughly same, 2.53, 2.0, and 2.0 respectively. The distribution is showing positive tendency (skewness =

5.46), however, the Kurtosis has far exceeded normal value, which indicates the distribution curve is intensively centralized around the mean. Due to this, standard deviation and variance is relatively small (std = 1.67, variance = 2.78). The IQR is 1.0, with 1st quartile equals to 2 and 3rd quartile equals to 3, making high and low limitation for outlier identifier being 4.5 and 0.5. Hence, around 5 percentage of the value can be regarded as an outlier. Besides, 26 missing values detected.

The UNIVARIATE Procedure Variable: Number_of_sexual_partners			
Moments			
N	832	Sum Weights	832
Mean	2.52764423	Sum Observations	2103
Std Deviation	1.66776048	Variance	2.78142501
Skewness	5.45464861	Kurtosis	69.2047544
Uncorrected SS	7627	Corrected SS	2311.36418
Coeff Variation	65.9808234	Std Error Mean	0.05781919

Basic Statistical Measures			
Location		Variability	
Mean	2.527644	Std Deviation	1.66776
Median	2.000000	Variance	2.78143
Mode	2.000000	Range	27.00000
		Interquartile Range	1.00000

The MEANS Procedure	
Analysis Variable : Number_of_sexual_partners	
N Miss	26

Figure 9: Descriptive Statistics of Number of sexual partners.

Figure 10: Histogram for Number of sexual partners.

Fig.10 is the Histogram for *Number of sexual partners*. As can tell from the graph, the distribution curve is showing a thin tail, and most of the values are tightly distributed surrounding the mean. Some outliers can be identified on the right side of the X-axis.

Figure 11: Q-Q Plot of Number of sexual partners.

Fig.11 is the Quantile-Quantile plot for *Number of sexual partners*. As can tell from the graph upper right part of the data value has deviated from the normal reference line. This matches the deduction from descriptive statistics that there are some outliers exists at the upper end of *Number of sexual partners* variation range.

3) Smokes

Smokes	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	722	85.44	722	85.44
1	123	14.56	845	100.00
Frequency Missing = 13				

```

PROC FREQ data=CERVICAL_CANCER;
table Smokes;
run;

```

Figure 12: Frequency Distribution of Smokes.

Since *Smokes* is a binary variable indicating YES/NO, it makes no meaning to analyzing central tendency. Hence, PROC FREQ is applied to understand frequency distribution. As seen in Fig.12, 85.44 percentage of the observations are 0, and 14.56 percentage is 1. There are 13 missing values in the dataset. The histogram in Fig13 & Fig.14 is visualizing the same trend.

Figure 13: Histogram of Smokes.

Figure 14: Histogram of Smokes (years).

4) Smokes (years)

Variable: Smokes__years__			
Moments			
N	845	Sum Weights	845
Mean	1.21972141	Sum Observations	1030.66459
Std Deviation	4.08901694	Variance	16.7200595
Skewness	4.4654839	Kurtosis	23.7684182
Uncorrected SS	15368.8539	Corrected SS	14111.7302
Coeff Variation	335.241875	Std Error Mean	0.14066646
Basic Statistical Measures			
Location		Variability	
Mean	1.219721	Std Deviation	4.08902
Median	0.000000	Variance	16.72006
Mode	0.000000	Range	37.00000
		Interquartile Range	0

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	13	1.52	100.00

Figure 15: Descriptive Statistics of Smokes (years).

Fig.15 shows the descriptive stats of *Smoke (years)*. There are 13 missing values in this attribute. Amongst the 845 responded records, more than three quarter answered 0 to this attribute. Hence there is a high tendency of skewness and kurtosis. The mean is 1.22 and median, mode are 0. IQR is 0, which indicates most of the value points are tightly distributed around the median. Fig.19 is showing the distribution histogram, which indicates the same trend.

5) Smokes (packs/year)

From Fig 16 details of descriptive statistics data, the mean/median/mode all-around 0, the standard deviation is 2.22 and range is 37. The skewness and kurtosis are all far exceeding normal level, which means the distribution curve is highly positively skewed and tightly centralized surrounding the mean. There are 13 incomplete values.

Plotting the histogram as seen in Fig.17, right skewness and thin kurtosis are noticed from the distribution curve. Besides, some outliers also noticed.

Figure 16: Descriptive Statistics for Smokes (packs/year).

Figure 17: Histogram for Smokes (packs/year).

6) STDs

Figure 18: Frequency for STDs.

The frequency graph and frequency distribution for *STDs* can be seen from Fig.18. As can tell from the histogram, more observations have replied 0 comparing with 1. To be concrete, 10.49 percentage of the values are with 1 for *STDs*, while 89.51 percentage with 0. There are 105 missing values detected in *STDs*.

7) STDs (number)

As can tell from Fig.19 and Fig.20, the mean, mode and median all closely around 0. For this attribute, there are a lot of missing values involved. And most of the observations (roughly 90%) are giving 0 as feedback. The range is 4, and the maximum value appears in 4.

Figure 19: Descriptive Statistics for STDs (number).

Figure 20: Histogram for STDs (number).

8) STDs: syphilis

Fig.21 depicted the frequency distribution table as well as distribution histogram for *STDs_syphilis*. 2.39 percentage of the observations responded with 1 and rest are noted with 0. There are 105 missing values in this attribute.

Figure 21: Frequency for STDs: syphilis.

9) STDs: Number of diagnoses

Figure 22: Frequency for STDs: Number of diagnoses.

As can tell from Fig.22, highest value for *STDs: Number of diagnoses* appears in 3 and the majority responded with 0. No missing value for this attribute.

10) STDs: Time since the last diagnosis

Fig.23 shows the distribution of the values for *STDs: Time since the last diagnosis*. As can tell from the histogram, 787 of the observations did not provide respond for this attribute. Among the values provided, the highest value resides in 9 (which appeared one time) and the lowest value is 1 (which appeared 28 times). Due to 91.72 per cent of the value is missing, for this attribute, it requires either data cleaning or this attribute shall be ignored, otherwise, the predictive model will be highly biased.

Figure 23: Frequency for STDs: Time since last diagnosis.

11) Dx: Cancer, CIN, HPV, Dx

The FREQ Procedure					The FREQ Procedure				
Dx_Cancer	Frequency	Percent	Cumulative Frequency	Cumulative Percent	Dx_CIN	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	840	97.90	840	97.90	0	849	98.95	849	98.95
1	18	2.10	858	100.00	1	9	1.05	858	100.00

The FREQ Procedure					The FREQ Procedure				
Dx_HPV	Frequency	Percent	Cumulative Frequency	Cumulative Percent	Dx	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	840	97.90	840	97.90	0	834	97.20	834	97.20
1	18	2.10	858	100.00	1	24	2.80	858	100.00

Figure 24: Frequency Distribution for Dx: Cancer, CIN, HPV, Dx.

Dx: Cancer, CIN, HPV and Dx will be recognized as the target variable, which indicates whether or not the patient is diagnosed with cancer. As can see from the distribution Fig

24, around 2 percentage of the collected sample data is diagnosed with either kind of cancer. There are some missing values noticed as well.

DISCUSSION

There are 858 observations in the target dataset, of which 2.1 percentage is infected with Cancer, 1.05 percentage with

CIN, 2.1 percentage with HPV and another 2.8 percentage with another kind of cancer. However, there are a large number of missing values in the dataset, which will largely affect the modelling performance. The incomplete values are mainly seen in columns requiring more private information (No missing value detected in an attribute such as *Age*). To achieve better model accuracy, Data Pre-processing and transformation is necessary.

1. Data Cleaning

Table 2: Variables identified with quality issue need to be pre-processed

No	Variables	Attributes Types	incomplete	inconsistent	outlier
1	Age	Ratio			
2	Number of sexual partners	Ratio	Y		
3	First sexual intercourse	Ratio	Y		
4	Number of pregnancies	Ratio	Y		
5	Smokes	Nominal	Y		
6	Smokes (years)	Ratio	Y	Y	
7	Smokes (packs/year)	Ratio	Y	Y	
8	Hormonal Contraceptives	Nominal	Y		
9	Hormonal Contraceptives (years)	Ratio	Y	Y	
10	IUD	Nominal	Y		
11	IUD (years)	Ratio	Y	Y	
12	STDs	Nominal	Y		
13	STDs (number)	Ratio	Y		
14	STDs: condylomatosis	Nominal	Y		
15	STDs: cervical condylomatosis	Nominal	Y		
16	STDs: vaginal condylomatosis	Nominal	Y		
17	STDs: vulvo-perineal condylomatosis	Nominal	Y		
18	STDs: syphilis	Nominal	Y		
19	STDs: pelvic inflammatory diseases	Nominal	Y		
20	STDs: genetical herpes	Nominal	Y		
21	STDs: molluscum contagiosum	Nominal	Y		
22	STDs: AIDS	Nominal	Y		
23	STDs: HIV	Nominal	Y		
24	STDs: Hepatitis B	Nominal	Y		
25	STDs: HPV	Nominal	Y		
26	STDs: Number of diagnosis	Ratio			
27	STDs: Time since the first diagnosis	Ratio	Y		Y
28	STDs: Time since the last diagnosis	Ratio	Y		Y
29	Dx: Cancer	Nominal			
30	Dx: CIN	Nominal			
31	Dx: HPV	Nominal			
32	DX	Nominal			
33	Hinselmann	Nominal			
34	Schiller	Nominal			
35	Citology	Nominal			
36	Biopsy	Nominal			

There are mainly three types of data quality issues which requires for cleansing, including missing data (incomplete data), outlier (noisy data), inconsistent data. As per the work performed in data exploration, attributes with data quality issues are identified as indicated. Since data quality will have direct impact on model performance, each type of issue will be handled respectively before loading the data for further analysis. (See in Table 2)

Incomplete Data

This dataset contained a large volume of missing data. Commonly, there are many reasons which would lead to data incompleteness: 1). Malfunction of the machine or human operational mistake 2). Due to privacy concern, the interviewee did not provide the information. For this dataset, the latter one is considered as the main cause.

There are many feasible approaches to handle missing values. In regards to this dataset, the median value will be filled as a substitute.

Below is the code for processing data incompleteness:

1.1 Inconsistent Data

Table 3: Attributes with inconsistent issue

No	Variables	Attributes Types	Inconsistent
1	Smokes (years)	Ratio	Y
2	Smokes (packs/year)	Ratio	Y
3	Hormonal Contraceptives(years)	Ratio	Y
4	IUD (years)	Ratio	Y

Inconsistent data is referring data which does not fit into the current format. This is mainly due to data integration from the heterogenous data source. Different format data does not consistent with each other and cause to data redundancy and data inaccuracy. In this dataset, some attributes for counting of years is supposed to be an integer, however, in the given dataset is shown with unreasonable precision. To handle this, will round the values in these attributes and get an integer. Processing code as shown in Fig.25.

```
/*inconsistent data*/
data cervical_cancer2;
set cervical_cancer3;
Smokes_years_ = round(Smokes_years_,1);
Smokes_packs_year_ = round(Smokes_packs_year_,1);
Hormonal_Contraceptives_years_ = round(Hormonal_Contraceptives_years_,1);
IUD_years_ = round(IUD_years_,1);
```

Figure 25: Code for adjusting precision for inconsistent attributes.

Outlier

Table 4: Attributes with suspect outliers

No	Variables	Attributes Types	Outlier
1	STDs: Time since the first diagnosis	Ratio	Y
2	STDs: Time since the last diagnosis	Ratio	Y

The outlier is referring to the numeric value in the dataset which far deviates from other values. In this dataset, below attributes *STDs: Time since first diagnosis* and *STDs: Time since the last diagnosis* is suspected with outliers, (Table 4) because the years provided in some of the observations are equal to or greater than the *Age* value. The responder purposely gave the wrong answer or there is some malfunction in the data collection process. Through inspection, only two records are detected with this kind of situation, hence will remove it. The total dataset size is now 856 instead of 858. (Fig.26)

```
/*outlier*/
data cervical_cancer3;
set cervical_cancer2;
if STDs_Time_since_first_diagnosis >= age then delete;
if STDs_Time_since_last_diagnosis >= age then delete;
```

Figure 26: Code for handling outline.

Dataset Transformation

For better modelling performance, enhanced visualization effect and improved computational power, data transformation is required. There are mainly two objectives in data transformation, one is to generate an intermediate variable for analysing, the other is for fixing missing value and skewness problem. There are many transformation techniques, in this dataset, two main approaches will be applied: 1). Generalization 2). Feature scaling. Table 5 is the summarization table for variables which will be transformed, and also which type of transformation will be applied.

Table 5: Attributes to be transformed and the transformation method

No	Variables	Attributes Types	Transformation Method
1	Age	Ratio	Generalization
2	Number of sexual partners	Ratio	Generalization
3	First sexual intercourse	Ratio	Generalization

Generalization

This is also called discretization, to discretize values into ordinal groups. Basing on the extreme value and number of

instances, bin number and bin width can be decided, then each of the value is pumped into specific groups.

(1) Age

Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
13	1	0.12	1	0.12
14	5	0.58	6	0.70
15	21	2.45	27	3.15
16	23	2.68	50	5.83
17	35	4.08	85	9.91
18	50	5.83	135	15.73
19	44	5.13	179	20.86
20	45	5.24	224	26.11
21	46	5.36	270	31.47
22	30	3.50	300	34.97
23	54	6.29	354	41.26
24	39	4.55	393	45.80
25	39	4.55	432	50.35
26	38	4.43	470	54.78
27	33	3.85	503	58.62

Figure 27: Frequency Table for Age.

Fig.27 is the frequency table for *Age*. There are 858 data entries, the value range is $84 - 13 = 71$. Considering the data range and dataset size, choose to discretize with 10 bins, each bin with roughly 86 instances. Fig.28 shows the code for discretization as well as output variable after transformation:

```
data CERVICAL_CANCER_transform;
set CERVICAL_CANCER;
if age le 17 then age_transformed =1;
else if age le 19 then age_transformed =2;
else if age le 21 then age_transformed = 3;
else if age le 23 then age_transformed = 4;
else if age le 25 then age_transformed = 5;
else if age le 28 then age_transformed = 6;
else if age le 30 then age_transformed = 7;
else if age le 34 then age_transformed = 8;
else if age le 37 then age_transformed = 9;
else age_transformed = 10;
```

	Age	age_transformed
1	15	1
2	17	1
3	17	1
4	17	1
5	17	1
6	17	1
7	17	1
8	14	1
9	17	1
10	17	1

Figure 28: Code for Transformation and Age after transformation.

(2) Perform the same to the rest variables. Detailed codes can be found in the appendix. Fig.29 shows the output after transformation.

	age_transformed	Number_of_sexual_p_transformed	First_sexual_inter_transform
1	2	3	1
2	1	1	1
3	8	1	1
4	10	3	2
5	10	3	3
6	10	3	3
7	10	3	2
8	6	1	3
9	10	1	3
10	10	3	1
11	10	3	3

Figure 29: Output variables after transformation.

1.2 Feature scaling

Feature scaling techniques are for scale down the range of the variable, so latter one when applying machine learning algorithm on it, computational power can be enhanced. Besides, some algorithms such as neural network, SVM can only work with a number between zero and one, scale down the feature and get it normalized is a necessary step. (Table 6)

Table 6: Attributes to be transformed with feature scaling

No	Variables	Attributes Types	Transformation Method
1	Number of pregnancies	Ratio	Feature Scaling
2	IUD (years)	Ratio	Feature Scaling
3	STDs (number)	Ratio	Feature Scaling
4	STDs: Number of diagnosis	Ratio	Feature Scaling

The techniques applied here will be normalization. The final normalized value would be the current value be subtracted with a minimum value of the entire column and divided by the subtraction of maximum and minimum value:

Equation (1):

$$\text{normalized value} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

(1) Number of pregnancies

For the *number of pregnancies*, firstly use PROC FREQ get the maximum and minimum value, which is 0 and 8, so each value will be subtracted with 0 and divided by 8. Processing code as follows (Fig.30):

```
data CERVICAL_CANCER_transform;
set CERVICAL_CANCER_transform;
Num_of_pregnancies_transform = (Num_of_pregnancies - 0)/8;
```

Figure 30: Normalization for Number of pregnancies.

(2) Rest of the variables are handled in the same method. Output after transformation as shown below (Fig.31):

	STDs__number_transform	Num_of_pregnancies_transform	IUD_years_transform
1	1	0.25	0
2	0.75	0.25	0
3	0.75	0.25	0
4	0.75	0.125	0
5	0.75	0.5	0
6	0.75	0.25	0
7	0.75	0.25	0
8	0.5	-	-
9	0.5	0.125	0
10	0.5	0.375	0
11	0.5	0.125	0

Figure 31: Normalization output.

DISCUSSION

Section 4 and 5 are dealing with data cleaning and data transformation. Both of which are necessary steps of data pre-processing before data can be loaded later to other analysing tools. In this task, SAS studio is employed for initial data exploration (understanding spread and trend), data cleaning as well as data transformation. In the process of pre-processing, the original data file has been modified and new attributes are generated as the secondary analysing variable. The output after data cleaning is named as “X_cleanup” and this has been used as the input for data transformation. The final output from data transformation is named as “X_transformation” (see in Fig.32). This file will be downloaded and exported for analysing afterwards.

Figure 32: Output dataset from data pre-processing.

Apache Hadoop

Hortonworks Data Platform (HDP) is an open-sourced framework providing platform for Apache Hadoop distributed architecture. It is mainly used for handling data-at-rest. While Hortonworks DataFlow (HDF) provides an end-to-end data collection, curates and analysis platform for handling data in real-time. Hortonworks Sandbox is an easy access desktop portal for utilizing HDP and HDF. It is based on a single node VM and is integrated with solutions such as HDFS, YARN and HIVE¹⁰ For analysing the cervical data, will upload the dataset onto Hadoop through HDP sandbox, and access manipulate via HQL of HIVE. Below Fig.33 shows, the dashboard of sandbox after Linux VM machine is started.

Figure 33: Hortonwork Sandbox.

Next, change to HIVE view and create a new database named “dm”. Code is shown as in Fig.34 and can see from the left menu side, the new database is created. After this, upload transformed dataset from local onto HIVE by entering into “Upload Table” option. As see in Fig.35, after specifying a local directory, few changes need to be applied: (1) In File Type option, check the box identifying there is subject title included in the first row of the dataset. (2) Store the dataset as Optimised Row Columnar file format (ORC) to get more efficient performance. (3) Change the database from default to “dm”. Now, the dataset has already been uploaded from local to HIVE.

Next step, to utilize the ORC file, command in Fig.36 is executed and the dataset is now stored as ORC file and named as “cervical_cancer_orc”.

Figure 34: Creating a new Database.

Figure 35: Uploading local dataset.

Figure 36: Code for creating ORC table.

Data Mining Hypothesis and Analysis

To visualize the dataset, and further study the relationships between each attribute will load the dataset into Tableau. There are several different approaches to perform data loading. For this assignment, will use Open Database Connectivity (ODBC) as the bridge to connect Tableau application to the Database Management System (DBMS). ODBC permits applications with multiple concurrent connections with different databases in a consistent way¹¹. The following steps show the details for configuring ODBC to connect to HDP sandbox and load data from Hadoop to Visualization tool.

Set up the ODBC connection

Figure 37: Config for ODBC.

As shown in Fig.37, firstly, call up ODBC for 64bits from windows system, then config for Hortonworks Hive DSN by providing host portal and username. Click to test the connection to ensure the connection is successfully established.

Connecting HDP to Tableau

Start-up Tableau, choose to load data from Hadoop Server (See in Fig.38). Provide Hortonwork host portal address and username, then to sign-in.

Figure 38: Connect Tableau to Hadoop.

Next thing is to extract data from Hadoop server and get the table named “cervical_cancer_orc”. It can be accessed by setting schema as “dm” first, then extract table by the search for table name (as seen in Fig.39).

Hypotheses and Analysis

(1) Hypotheses

Basing on the business understanding from Literature Review and data understanding from data exploration and analysis in SAS and HIVE, below hypotheses are set up, and visualization techniques are applied.

Hypothesis 1: There is a positive correlation between the consumption of cigarette and cancer infection rate.

Hypothesis 2: The use of Intrauterine device (IUD) will reduce the chance for cervical cancer infection rate.

Hypothesis 3: Infection of HPV will cause a high risk of cervical cancer infection.

Hypothesis 4: With the rise of the number of sexual partners, the chance of getting cervical cancer will also increase.

Hypothesis 5: There is an ageing period most risky for females to be infected with cervical cancer.

(2) Analysis with visualization

a) Cigarette consumption and cancer infection ratio

As can tell from Fig.40, whether the patient is diagnosed with cancer is set as a dimension (0,1). Then average cigarette consumption years is visualized in the y-axis. As can tell for D_x : *Cancer* and D_x : *HPV*, patient-reported with disease infection showing higher cigarette consumption period. This trend does not apply for D_x and D_x *CIN* records with value for 1. Suspect this is due to limitation of record volume, hence plot Fig.41 showing distribution chart. As can tell, total records for D_x positive response is 24 and D_x *CIN* positive the response is 9, therefore analysis result is highly biased and cannot be trusted.

Figure 39: Extract Table from Hadoop.

In light of the above, further study into the relationship between cigarette and cancer infection rate basing on the data of *Dx Cancer* and *Dx HPV*. Will generate an intermediate variable: *packs total consumption* by multiple *smoke years* and *smoke_packs_year*. And then get the average cigarette consumption for infected patients and uninfected patients respectively (As see in Fig.42). Total cigarette consumption amount for *Dx HPV* (1) records is higher than that of *Dx HPV* (0). Same goes for *Dx Cancer*.

From smoker and non-smoker perspective to analysis whether smoking habit will simulate the occurrence of cancer, below Fig.43 is plotted. As can tell from the graph, the infection rate has raised from 2.04% to 2.46% between non-smoker and smoker.

In conclusion, whether the patient is a smoker, how many years she smokes, and the amount of cigarette consumption all intensively related to cervical cancer infection rate. Hypothesis 1 is proved correct.

Figure 40: Visualization between cigarette consumption years and cancer infection.

Figure 41: Dataset distribution by Disease Category.

Figure 42: Cigarette consumption vs cancer infection.

Figure 43: Infection rate among smoker vs non-smoker.

b) IUD usage and cancer infection rate

Fig.44 shows an overview of IUD usage. There are 856 observations, amongst which 117 responders did not provide the answer for IUD usage, 656 gave 0 and 83 of them gave 1. Next, the infection rate of HPV is showcased in Fig.45. The infection rate of IUD user is higher than that of not IUD user, this might due to the record size as well. Hypothesis 2 is rejected.

Figure 44: IUD record distribution.

Figure 45: Infection rate among different groups.

c) The infection rate of HPV and infection rate of cancer

Fig.46 shows distribution of HPV infects and non-infected records. Totally 18 HPV (1) records comparing with 838 HPV (0) records. Further, delve into the cancer infectious rate within each group. As see in Fig.47 HPV positive result is highly possible to be linked with Cancer positive result, the infection rate increased from 0.24% to 88.89%.

Figure 46: HPV infects vs non-infect records.

Figure 47: Cancer infection rate among HPV patient and non-HPV patient.

d) Number of sexual partners and cancer infection rate

Fig.48 shows the distribution of the *number of sexual partners* amongst the responders. As can tell the majority answered 1 to 3, and there some other records sparsely distributed along X-axis. After this, will use transformed variable *number_of_sexual_partner_transform* as dimension. A roughly equal number of records are put into each bin. Then the infection rate is visualized as in Fig.49. There is an upward trend from Group one to Group three, raised from 1.46% to 1.69% to 2.82%. This indicates with the number of sexual partners increasing, the risk of getting infected with cervical cancer is also increasing.

Figure 48: Number of Sexual Partners Distribution.

Figure 49: Infection rate among different groups

e) Cancer infection rate against age distribution

Fig.50 shows the distribution of *Age*. The majority of the responders are aged between 15 and 37. Next, will employ transformed variable *Age_transform*, which has roughly the same bin width among the 10 age bins, to analysis infection rate within age groups. As can summarize from Fig.51 infection rate is relatively low among Age group 1 to 3 (equal or younger than age 21), then raised a bit among Age group 4 to 7 (equal or younger than age 30) and reaches a peak time among age group 8 to 10 (beyond 30). This indicates the risk of cancer infectious rate is positively related to age increase. In consideration of this, a woman should be more cautious about cervical cancer with age growing.

Figure 50: Age Distribution.

Figure 51: Infection rate among different age groups.

CONCLUSION

Through this work, roughly more than 80% of the time is spent on data exploration and pre-processing. Data cleaning and transformation steps are necessary since during the work of data exploration, it has shown there are some data quality issues such as missing data, and noisy data. There are several issues detected with this dataset:

Firstly, the size of the dataset is not large enough, which might result in biased prediction result. For example, when studying about the relation of cigarette consumption and cancer infection rate, the data pattern is showing adverse in comparison with the initial hypothesis and related literature understanding, which is caused by the sparsity of records.

Secondly, the incompleteness of this dataset. For many attributes such as STD status and IUD usage situation, there is a large amount of missing value. This is most probably due to privacy and security concern from the patient.

Thirdly, the veracity of the dataset. There are some attributes such as *STDs: Time since first diagnosis* and *STDs: Time since the last diagnosis* are identified with some unreasonable values. This might be caused by a human mistake in the data collection process or the responder intentionally gave the wrong answer in consideration of privacy security.

Due to abovementioned reasons, in data analysing phase, the result sometimes conflicts with reasoning deduction. In consideration of this, further study in this field shall fix the above issues, then prediction accuracy can be improved.

Since some attributes of this dataset are relating with quite sensitive information, in the fear of information leakage or out of embarrassment some of the responders gave a false answer or avoided answering. This highly jeopardized the genuineness of the data and affected the accuracy of prediction. To safeguard the privacy of patients and conform to legal regulatory, healthcare centres are expected to apply data-preserving techniques before publishing or sharing the data.

As understand from literature, several factors will cause a high risk of cervical cancer. Basing on the theory, five hypotheses were set up, including factors of, cigarette consumption, IUD usage, HPV infection status, number of sexual partners and age group. For four of the hypotheses, visualization from data matches the assumption, except for IUD usage, which is suspected to be caused by limited record volume. Proved insights can be drawn from this study are: cigarette consumption amount is positively related with cancer infection rate; For HPV infected patients, there is a higher chance for them to get infected with cervical cancer; The number of sexual partners will impact the risk of cervical cancer infection risk; Woman with elder age group is under higher infection risk of cervical cancer. A future study could collect more data relating to IUD factor and further study the relationship between cancer and IUD.

ACKNOWLEDGEMENTS

The authors also wish to express gratitude to the management of Asia Pacific University of Technology & Innovation (APU) for their support.

Conflict of Interest

The authors involved in the current study does not declare any competing conflict of interest.

Funding and Sponsorship

No fund or sponsorship in any form was obtained from any organization for carrying out this research work.

REFERENCE

1. S. Emma, "World Cancer Day 2017: how to prevent cervical cancer cases around the globe - Cancer Research UK - Science blog," *Cancer Research 2017*, 2017. [Online]. Available: <https://scienceblog.cancerresearchuk.org/2017/02/08/world-cancer-day-2017-how-to-prevent-cervical-cancer-cases-around-the-globe/>. [Accessed: 10-Aug-2018].
2. M. Griffiths, "Report of a WHO consultation : Cervical Screening in developing countries," *World Heal. Organ.*, p. 984, 2002, DOI: 10.4103/0019-509X.64704.
3. Wong, A.N, and Jayabalan, M., 2018, Data Preparation Techniques for Cervical Cancer Screening. *Journal of Advanced Research in Dynamical and Control Systems* 10(12).
4. Jayabalan, M. and O'Daniel, T., 2016. Access control and privilege management in an electronic health record: a systematic literature review. *Journal of medical systems*, 40(12), p.261.
5. B. L. Filkins *et al.*, "Privacy and security in the era of digital health: What should translational researchers know and do about it?," *Am. J. Transl. Res.*, vol. 8, no. 3, pp. 1560–1580, 2016.
6. E. S. Dove *et al.*, "Genomic cloud computing: Legal and ethical points to consider," *Eur. J. Hum. Genet.*, vol. 23, no. 10, pp. 1271–1278, 2015, DOI: 10.1038/ejhg.2014.196.
7. S. Thiebes, G. Kleiber, and A. Sunyaev, "Cancer Genomics Research in the Cloud: A Taxonomy of Genome Data Sets," no. October, pp. 0–20, 2017.
8. L. Sweeney, "Achieving fc-Anonymity Privacy Protection Using Generalization and Suppression," *Int. J. Uncertain.*, vol. 10, no. 5, pp. 571–588, 2002, sDOI: 10.1142/S021848850200165X.
9. SAS, "SAS Elementary Statistics Procedures," *Base SAS® 9.4 Procedures Guide*, 2018. [Online]. Available: <http://documentation.sas.com/?docsetId=proc&docsetTarget=p0v0y1on1hbxukn0zqgsp5ky8hc0.htm&docsetVersion=9.4&locale=en>.
10. Hortonworks, "Download Apache Hadoop Sandbox on Virtual Machine or Azure | Hortonworks," *Hortonworks*, 2018. [Online]. Available: https://hortonworks.com/products/sandbox/?utm_source=google&utm_medium=cpc&utm_campaign=hw_sandbox&gclid=Cj0KCQjwzK_bBRD-DARIsAFQF7zPkj-ILYHRmv6Z7AoJ8XLgYawesDUz48M-Ki70HrMd4f4F58mzdk38oaAjj2EALw_wcB. [Accessed: 09-Aug-2018].
11. IBM, "IBM Knowledge Center - Open Database Connectivity (ODBC)," *IBM*, 2018. [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SS6SGM_4.1.1/com.ibm.cobol.aix.doc/PGandLR/concepts/cpodb02.html. [Accessed: 09-Aug-2018].