Shaukat Khanum Memorial Hospital

Rabia Shafique

09/21/2012

Table of Contents

1 Data Analysis 3
1.1 MRNO 3
1.2 REGISTRATION_DATE 3
1.3 DOB 3
1.4 FIRST_VISIT_DATE 3
1.5 AGE_AT_PRESENTATION 3
1.6 SEX 3
1.7 TEHSIL 3
1.8 DISTRICT 3
1.9 PROVINCE 4
1.10 COUNTRY 4
1.11 MARITAL_STATUS 4
1.12 OCCUPATION 4
1.13 RELATION 4
1.14 FAMILY_HISTORY_ICD 4
1.15 FAMILY_HISTORY_ICD_DESC 4
1.16 ADDICTION 4
1.17 PREVIOUS_TREATMENT 5
1.18 VALID_BASIS_OF_DIAGNOSIS 5
1.19 DIAGNOSIS_AT_SKM 5
1.20 DEFINITIVE_DIAGNOSIS_INSTITUTE 5
1.21 ICDO_3_DATE 5
1.22 ORGAN 5
1.23 SUBSITE 5
1.24 MORPHOLOGY 5
1.25 GRADE 6
1.26 LATERALITY 6
1.27 TUMOR_SIZE 6
1.28 REGIONAL_LYMPH_NODE_EXAMINED 6
1.29 REGIONAL_LYMPH_NODE_POSITIVE 6
1.30 SURGICAL_MARGINS 6
1.31 CLINICAL_STAGE 7
1.32 COMMENTS_CLNICAL 7
1.33 PATHOLOGICAL_STAGE 7
1.34 COMMENTS_PATHOLOGICAL 7
1.35 SEER_SUMMARY_STAGE 7
1.36 FIRST_TREATMENT_DATE 7
1.37 FIRST_TREATMENT_ICD 7
1.38 FIRST_TREATMENT_ICD_DESC 7
1.39 DATE_OF_DEATH 8
1.40 A_DEATH_CAUSE 8
1.41 CLASS_OF_CASE 8
2 Data Statistics 9
2.1 Basic Statistics 9
2.1.1 Comments 10
2.2 Correlation Matrix 10
2.2.1 Comments on Results 13
2.3 Covariance Matrix 14
2.3.1 Comments 15

3 Frequent item set Mining 16
3.1 Need for mining frequent item sets 16
3.1.1 Minimum Support: .005 16
3.1.2 Min Support: 1.0 23
4 Cluster Analysis 31
4.1 Overall Analysis 31
4.2 Detail Analysis 32
4.2.1 Cluster 1 32
4.2.2 Cluster 2 33
4.2.3 Cluster 3 34
4.2.4 Cluster 4 35
4.2.5 Cluster 5 36
5 References 39
6 Appendix A 40
7 Appendix B 46

1 Data Analysis

The following data analysis of the dataset is for the year 2004-2008. This analysis includes the values of different attributes and the amount of missing values for each attribute.

1.1 MRNO

MRNO stands for Medical Record Number. It's an integer value and is unique for every patient. Every patient must have one MRNO. There are no missing values for this attribute in the data.

1.2 REGISTRATION_DATE

It's a date on which a patient is registered with SKM. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.3 DOB

It's a date on which a patient is born. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.4 FIRST_VISIT_DATE

It's a date on which a patient has first visited to SKM. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.5 AGE_AT_PRESENTATION

This attribute shows the age of the patient. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.6 SEX

This attribute shows the gender of the patient. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.7 TEHSIL

This attribute shows the Tehsil from which the patient belongs to. There are 675 missing values for this attribute in the data. This attribute has a nominal value.

1.8 DISTRICT

This attribute shows the District from which the patient belongs to. There are 675 missing values for this attribute in the data. This attribute has a nominal value.

1.9 PROVINCE

This attribute shows the Province from which the patient belongs to. There are 675 missing values for this attribute in the data. This attribute has a nominal value.

1.10 COUNTRY

This attribute shows the Country from which the patient belongs to. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.11 MARITAL_STATUS

This attribute shows the Marital status of a patient. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.12 OCCUPATION

This attribute shows the occupation of a patient. There are 223 missing values for this attribute in the data. This attribute has a nominal value.

1.13 RELATION

This attribute shows those relations of a patient who have diagnosed cancer in them. There are 53749 missing values for this attribute in the data which means that these patients didn't have anyone in family with cancer. This attribute has a nominal value.

1.14 FAMILY_HISTORY_ICD

This attribute shows the ICD(International code for disease) of those relations of a patient who have diagnosed cancer in them. There are 53700 missing values for this attribute in the data. This attribute has a real value.

1.15 FAMILY_HISTORY_ICD_DESC

This attribute shows the description of the ICD(International code for disease) of those relations of a patient who have diagnosed cancer in them. There are 53700 missing values for this attribute in the data. This attribute has a Nominal value.

1.16 ADDICTION

This attribute shows those addictive substances to which a patient is addicted to. There are 16 missing values for this attribute in the data. This attribute has a nominal value.

1.17 PREVIOUS_TREATMENT

This attribute shows the previous treatments that a patient has taken. There are 194 missing values for this attribute in the data. This attribute has a nominal value.

1.18 VALID_BASIS_OF_DIAGNOSIS

This attribute shows the procedure through which cancer was diagnosed in the patient. There are 29584 missing values for this attribute in the data. This attribute has a nominal value.

1.19 DIAGNOSIS_AT_SKM

This attribute shows that whether cancer of a particular patient was diagnosed in SKM. There are 1495 missing values for this attribute in the data. This attribute has a nominal value.

1.20 DEFINITIVE_DIAGNOSIS_INSTITUTE

This attribute shows that the institute in which a cancer was diagnosed in the patient. There are 35840 missing values for this attribute in the data. This attribute has a nominal value.

1.21 ICDO_3_DATE

This attribute shows the date for the ICDO-3 of a patient. There are 149 missing values for this attribute in the data. This attribute has a nominal value.

1.22 ORGAN

This attribute shows the organ in which a patient has cancer. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.23 SUBSITE

This attribute shows a little detail about the organ in which a patient has cancer. There are no missing values for this attribute in the data. This attribute has a nominal value.

1.24 MORPHOLOGY

The morphology of a cancer refers to the histological classification of the cancer tissue (histopathological type) and a description of the course of development that a tumour is likely to take: benign or malig0t (behaviour). The designation is based on a microscopic diagnosis of morphology by the pathologist (Esteban, Whelan, Laudico & Parkin 1995).[1]

This attribute shows the morphology of the cancer and there are 4 missing values for this attribute. This attribute has a nominal value.

1.25 GRADE

The grade is a system used to classify cancer cells in terms of how abnormal they look under a microscope and how quickly the tumor is likely to grow and spread.[2]

This attribute shows the grade of the cancer of a particular patient. There are 28607 missing values for this attribute. This attribute has a nominal value.

1.26 LATERALITY

This attribute shows on which side of the organ a patient has cancer. There are 2321 missing values for this attribute. This attribute has a nominal value.

1.27 TUMOR_SIZE

This attribute shows the tumor size of the cancer of a particular patient. There are 21276 missing values for this attribute. This attribute has an integer value.

1.28 REGIONAL_LYMPH_NODE_EXAMINED

A regional lymph node is a lymph node that drains lymph from the region around a tumour.

This attribute shows the number of lymph nodes examined in a particular patient. There are 21671 missing values for this attribute. This attribute has an integer value.

1.29 REGIONAL_LYMPH_NODE_POSITIVE

The total number of regional lymph nodes examined by a pathologist and reported as containing tumour.

This attribute shows the number of Regional lymph nodes positive in a particular patient. There are 21681 missing values for this attribute. This attribute has an integer value.

1.30 SURGICAL_MARGINS

Surgical margin in a surgery reports define the visible margin or free edge of "normal" tissue seen by the surgeon with the naked eye. Surgical margin as read in a pathology report define the histological measurement of normal or unaffected tissue surrounding the visible tumour under a microscope on a glass mounted histology section. A "narrow" surgical margin implies that the tumour exists very close to the surgical margin, and a "wide" surgical margin implies the tumour exists far from the cut edge or the surgical margin.

There are 22498 missing values for this attribute. This attribute has integer values.

1.31 CLINICAL_STAGE

Cancer stage refers to the extent or severity of the cancer, based on factors such as the location of the primary tumor, tumor size, number of tumors, and lymph node involvement (spread of cancer into lymph nodes).

There are 16107 missing values for this attribute and has nominal values.

1.32 COMMENTS_CLNICAL

This attribute shows the comments on the value of the previous attribute. There are 50764 missing values for this attribute and takes nominal values only.

1.33 PATHOLOGICAL_STAGE

Pathology is the study and diagnosis of disease through examination of organs, tissues, bodily fluids, and whole bodies (autopsies). Pathology also encompasses the related scientific study of disease processes, called general pathology.[3]

This attribute has 56053 missing values and takes nominal data only.

1.34 COMMENTS_PATHOLOGICAL

This attribute shows the comments on the value of the previous attribute. There are 56476 missing values for this attribute and takes nominal values only.

1.35 SEER_SUMMARY_STAGE

This attribute shows the comments on the value of the stage attribute. There are no missing values for this attribute and takes nominal values only.

1.36 FIRST_TREATMENT_DATE

This attribute shows the date for the first treatment of a particular patient. There are 9576 missing values for this attribute and takes nominal values only.

1.37 FIRST_TREATMENT_ICD

This attribute shows the ICD of a patient on date for his/her first treatment. There are 9546 missing values for this attribute and takes nominal values only.

1.38 FIRST_TREATMENT_ICD_DESC

This attribute shows the ICD description of the previous attribute. There are 9546 missing values for this attribute and takes nominal values only.

1.39 DATE_OF_DEATH

This attribute shows the date on which a particular patient died. There are 55531 missing values for this attribute and takes nominal values only.

1.40 A_DEATH_CAUSE

This attribute shows the cause of death of a particular patient. There are 55561 missing values for this attribute and takes nominal values only.

1.41 CLASS_OF_CASE

This attribute shows the class of cancer for a particular patient. There are 19650 missing values for this attribute and takes nominal values only.

2 Data Statistics

This section shows the results of different statistics applied on the dataset. The results are shown in the tabular form and are commented also.

2.1 Basic Statistics

Following is the analysis of data with respect to the value type, min, max, mean, standard deviation and mode of each attribute.

Attribute	Value Type	Min	Max	Mean	Standard Deviation	Mode
MRNO	Integer			57015.78	10991.03	
Registration_Date	Nominal					19 09 2005 03:04:00
DOB	Nominal					01-Jan-1960
First_Visit_Date	Nominal					16-Jun-2008
Age_At_Presentation	Nominal					50 Years
Sex	Nominal					Female
Tehsil	Nominal					Lahore
District	Nominal					Lahore
Province	Nominal					Punjab
Country	Nominal					Pakistan
Mrital_Status	Nominal					Married
Occupation	Nominal					HouseWife
Relation	Nominal					Mother
Family_History_ICD	Numerical	15	239	175.591	17.48	
Family_History_ICD_Desc	Nominal					neoplasm of breast (female), unspecified site
Addiction	Nominal					No
Previous_Treatment	Nominal					
Valid_Basis_of_Diagnosis	Nominal					Pathology
Diagnosis_At_SKM	Nominal					Yes
Definitive_Diagnosis_Institute	Nominal					Unknown
ICDO_3_Date	Nominal					15-Jun-2006
Organ	Nominal					Breast
Subsite	Nominal					Breast, NOS
Morphology	Nominal					Infiltrating Duct Carcinoma
Grade	Nominal					III
Laterality	Nominal					No
Tumor_Size	Numerical	0	999	531.12	475.36	

Regional_Lymph_Node_Examined	Numerical	0	99	33.52	42.22	
Regional_Lymph_Node_Positive	Numerical	0	99	67.17	44.51	
Surgical_Margins	Nominal	0	6	3.46	2.57	
Clinical_Stage	Nominal					Stage IV Any T Any N M1
Comments_Clinical	Nominal					Unstageable
Pathological_Stage	Nominal					Stage IIA T2 N0 M0
Comments_Pathological	Nominal					T2N2aMx
Seer_Summary_Stage	Nominal					Local
First_Treatment_Date	Nominal					24-Dec-2008
First_Treatment_ICD	Nominal					V58.0
First_Treatment_ICD_Decs	Nominal					Radiotharapy
Date_Of_Death	Nominal					01 01 2010 11:30:00
A_Death_Cause	Nominal					cardiopulmonary arrest
Class_Of_Death	Nominal					Class 1

2.1.1 Comments

- Mostly patients diagnosed at 50 years old and belong to city Lahore.
- Female cancer patients are more than males and cancer organ is breast.
- Mostly diagnosis at Shaukat Khanam hospital.
- Mostly basis of Diagnosis is pathological.
- Min size for tumour is 0 and it goes at maximum 999.
- Mostly class of death cases are of class1.

2.2 Correlation Matrix

Following are the results of correlation of each attribute. Since it is lot of data and it is difficult to show the correlation in the form of matrix, we have shown the correlation as follows.

First Attribute	Second Attribute	Correlation
REGISTRATION_DATE	DOB	0.68
REGISTRATION_DATE	FIRST_VISIT_DATE	1.00
REGISTRATION_DATE	AGE_AT_PRESENTATION	0.04
REGISTRATION_DATE	TEHSIL	0.06
REGISTRATION_DATE	DISTRICT	0.04
REGISTRATION_DATE	COUNTRY	0.04
REGISTRATION_DATE	MARITAL_STATUS	0.03
REGISTRATION_DATE	OCCUPATION	0.14
REGISTRATION_DATE	RELATION	0.04

REGISTRATION_DATE	FAMILY_HISTORY_ICD	0.07
REGISTRATION_DATE	DIAGNOSIS_AT_SKM	0.09
REGISTRATION_DATE	DEFINITIVE_DIAGNOSIS_INSTITUTE	0.51
REGISTRATION_DATE	ICDO_3_DATE	0.92
REGISTRATION_DATE	SUBSITE	0.02
REGISTRATION_DATE	MORPHOLOGY	0.07
REGISTRATION_DATE	LATERALITY	0.01
REGISTRATION_DATE	TUMOR_SIZE	0.02
REGISTRATION_DATE	REGIONAL_LYMPH_NODE_POSITIVE	0.02
REGISTRATION_DATE	CLINICAL_STAGE	0.04
REGISTRATION_DATE	COMMENTS_CLNICAL	0.55
REGISTRATION_DATE	PATHOLOGICAL_STAGE	0.13
REGISTRATION_DATE	COMMENTS_PATHOLOGICAL	0.67
REGISTRATION_DATE	SEER_SUMMARY_STAGE	0.24
REGISTRATION_DATE	FIRST_TREATMENT_DATE	0.93
REGISTRATION_DATE	FIRST_TREATMENT_ICD	0.34
REGISTRATION_DATE	FIRST_TREATMENT_ICD_DESC	0.34
REGISTRATION_DATE	DATE_OF_DEATH	0.99
REGISTRATION_DATE	A_DEATH_CAUSE	0.67
REGISTRATION_DATE	CLASS_OF_CASE	0.05
DOB	FIRST_VISIT_DATE	0.68
DOB	AGE_AT_PRESENTATION	0.04
DOB	TEHSIL	0.03
DOB	DISTRICT	0.02
DOB	COUNTRY	0.04
DOB	MARITAL_STATUS	0.07
DOB	OCCUPATION	0.10
DOB	RELATION	0.06
DOB	FAMILY_HISTORY_ICD	0.05
DOB	FAMILY_HISTORY_ICD_DESC	0.06
DOB	DIAGNOSIS_AT_SKM	0.06
DOB	DEFINITIVE_DIAGNOSIS_INSTITUTE	0.35
DOB	ICDO_3_DATE	0.63
DOB	SUBSITE	0.01
DOB	MORPHOLOGY	0.08
DOB	TUMOR_SIZE	0.04
DOB	REGIONAL_LYMPH_NODE_POSITIVE	0.02
DOB	CLINICAL_STAGE	0.03
DOB	COMMENTS_CLNICAL	0.27
DOB	PATHOLOGICAL_STAGE	0.08
DOB	COMMENTS_PATHOLOGICAL	0.10

DOB	SEER_SUMMARY_STAGE	0.17
DOB	FIRST_TREATMENT_DATE	0.64
DOB	FIRST_TREATMENT_ICD	0.22
DOB	DATE_OF_DEATH	0.74
DOB	A_DEATH_CAUSE	0.43
DOB	CLASS_OF_CASE	0.03
FIRST_VISIT_DATE	AGE_AT_PRESENTATION	0.04
FIRST_VISIT_DATE	TEHSIL	0.06
FIRST_VISIT_DATE	COUNTRY	0.04
FIRST_VISIT_DATE	MARITAL_STATUS	0.03
FIRST_VISIT_DATE	OCCUPATION	0.14
FIRST_VISIT_DATE	RELATION	0.03
FIRST_VISIT_DATE	FAMILY_HISTORY_ICD	0.07
FIRST_VISIT_DATE	DIAGNOSIS_AT_SKM	0.09
FIRST_VISIT_DATE	DEFINITIVE_DIAGNOSIS_INSTITUTE	0.51
FIRST_VISIT_DATE	ICDO_3_DATE	0.93
FIRST_VISIT_DATE	SUBSITE	0.02
FIRST_VISIT_DATE	MORPHOLOGY	0.07
FIRST_VISIT_DATE	COMMENTS_CLNICAL	0.53
FIRST_VISIT_DATE	PATHOLOGICAL_STAGE	0.14
FIRST_VISIT_DATE	COMMENTS_PATHOLOGICAL	0.66
FIRST_VISIT_DATE	SEER_SUMMARY_STAGE	0.24
FIRST_VISIT_DATE	FIRST_TREATMENT_DATE	0.93
FIRST_VISIT_DATE	DATE_OF_DEATH	0.99
FIRST_VISIT_DATE	A_DEATH_CAUSE	0.65
TEHSIL	DISTRICT	0.70
DEFINITIVE_DIAGNOSIS_INSTITUTE	COMMENTS_PATHOLOGICAL	0.27
DEFINITIVE_DIAGNOSIS_INSTITUTE	FIRST_TREATMENT_DATE	0.52
DEFINITIVE_DIAGNOSIS_INSTITUTE	FIRST_TREATMENT_ICD	0.18
DEFINITIVE_DIAGNOSIS_INSTITUTE	DATE_OF_DEATH	0.59
DEFINITIVE_DIAGNOSIS_INSTITUTE	A_DEATH_CAUSE	0.47
ICDO_3_DATE	COMMENTS_CLNICAL	0.37
ICDO_3_DATE	COMMENTS_PATHOLOGICAL	0.49
ICDO_3_DATE	SEER_SUMMARY_STAGE	0.25
ICDO_3_DATE	FIRST_TREATMENT_DATE	0.88
ICDO_3_DATE	FIRST_TREATMENT_ICD	0.30
ICDO_3_DATE	DATE_OF_DEATH	0.92
ICDO_3_DATE	A_DEATH_CAUSE	0.61
MORPHOLOGY	PATHOLOGICAL_STAGE	0.23

GRADE	REGIONAL_LYMPH_NODE_POSITIVE	0.15
GRADE	SURGICAL_MARGINS	0.12
TUMOR_SIZE	REGIONAL_LYMPH_NODE_EXAMINED	0.22
TUMOR_SIZE	REGIONAL_LYMPH_NODE_POSITIVE	0.48
TUMOR_SIZE	SURGICAL_MARGINS	0.50
REGIONAL_LYMPH_NODE_EXAMINED	CLASS_OF_CASE	0.04
REGIONAL_LYMPH_NODE_POSITIVE	SURGICAL_MARGINS	0.73
REGIONAL_LYMPH_NODE_POSITIVE	PATHOLOGICAL_STAGE	0.27
REGIONAL_LYMPH_NODE_POSITIVE	FIRST_TREATMENT_ICD	0.05
SURGICAL_MARGINS	PATHOLOGICAL_STAGE	0.22
COMMENTS_CLNICAL	COMMENTS_PATHOLOGICAL	0.71
COMMENTS_CLNICAL	SEER_SUMMARY_STAGE	0.10
COMMENTS_CLNICAL	FIRST_TREATMENT_DATE	0.52
COMMENTS_CLNICAL	FIRST_TREATMENT_ICD	0.18
COMMENTS_CLNICAL	DATE_OF_DEATH	0.51
COMMENTS_CLNICAL	A_DEATH_CAUSE	0.38
COMMENTS_CLNICAL	CLASS_OF_CASE	0.06
PATHOLOGICAL_STAGE	COMMENTS_PATHOLOGICAL	0.08
PATHOLOGICAL_STAGE	DATE_OF_DEATH	0.22
COMMENTS_PATHOLOGICAL	FIRST_TREATMENT_DATE	0.44
COMMENTS_PATHOLOGICAL	DATE_OF_DEATH	0.61
COMMENTS_PATHOLOGICAL	A_DEATH_CAUSE	0.47
SEER_SUMMARY_STAGE	FIRST_TREATMENT_DATE	0.28
SEER_SUMMARY_STAGE	DATE_OF_DEATH	0.24
SEER_SUMMARY_STAGE	A_DEATH_CAUSE	0.22
FIRST_TREATMENT_DATE	FIRST_TREATMENT_ICD_DESC	0.31
FIRST_TREATMENT_DATE	DATE_OF_DEATH	0.93
FIRST_TREATMENT_DATE	A_DEATH_CAUSE	0.61
FIRST_TREATMENT_ICD	FIRST_TREATMENT_ICD_DESC	1.00
FIRST_TREATMENT_ICD	DATE_OF_DEATH	0.41
DATE_OF_DEATH	A_DEATH_CAUSE	0.64

2.2.1 Comments on Results

- Registration date and date-of-death, First treatment date, death-cause are strongly dependent and correlated attributes.
- ICDO, seer summary stage, occupation, pathological stage, family history ICD, Morphology, country and district with Registration date are positively correlated

whereas surgical margins, province and lymph-node are negatively correlated or independent of registration date.

- Date of death and first treatment date strongly correlated with date of birth of patients.
- Date of death depends upon first visit date.
- Sex strongly correlated with Seer summary stage and first treatment date.
- First treatment date and date of death are dependent strongly.
- Tehsil and District are strongly correlated.
- ICDO 3 Date and Date of death are strongly correlated.
- Grade dependent on REGIONAL_LYMPH_NODE_POSITIVE and surgical margins.
- Marital status correlated with lymph node examined, morphology, grade, laterality, tumour size, lymph node positive and surgical margins.
- Tumour size and surgical margins are strongly correlated.
- Comments clinical and comments pathological are strongly correlated.
- Regional lymph node positive and surgical margins are strongly correlated.
- Addiction is positively correlated with surgical margins, lymph node positive, basis of diagnose, pathological stage, grade, subsite, tumour size, clinical stage and seer summary stage.
- Occupation is positively correlated with ICD-date, regional lymph node positive, first treatment date, date of death, tumour size, surgical margins, pathological stage, addiction, first treatment ICD, subsite, morphology, death cause, grade and class of case.

2.3 Covariance Matrix

Since it is lot of data and it is difficult to show the covariance in the form of matrix, we have shown the covariance as follows. There were a lot of attributes that having zero covariance (non-linear relationship between them or they were independent), we have removed them. This following table is showing only those attributes that have either positive or negative covariance.

First Attribute	Second Attribute	Covariance
REGISTRATION_DATE	DOB	10689453.48
REGISTRATION_DATE	AGE_AT_PRESENTATION	4633.30
REGISTRATION_DATE	SEX	-78.70
REGISTRATION_DATE	COUNTRY	26.39
REGISTRATION_DATE	MARITAL_STATUS	82.40
REGISTRATION_DATE	SUBSITE	4003.95
REGISTRATION_DATE	SEER_SUMMARY_STAGE	2642.15
DOB	AGE_AT_PRESENTATION	3126.08
DOB	SEX	-27.81
DOB	COUNTRY	18.72
DOB	MARITAL_STATUS	119.03
DOB	SUBSITE	1360.43
DOB	SEER_SUMMARY_STAGE	1305.09
AGE_AT_PRESENTATION	SEX	-1.50

AGE_AT_PRESENTATION	COUNTRY	-0.01
AGE_AT_PRESENTATION	MARITAL_STATUS	1.95
AGE_AT_PRESENTATION	SUBSITE	-5.04
AGE_AT_PRESENTATION	SEER_SUMMARY_STAGE	-2.69
SEX	MARITAL_STATUS	0.01
SEX	SUBSITE	-2.83
SEX	SEER_SUMMARY_STAGE	0.11
COUNTRY	SUBSITE	0.20
COUNTRY	SEER_SUMMARY_STAGE	0.01
MARITAL_STATUS	SUBSITE	0.07
MARITAL_STATUS	SEER_SUMMARY_STAGE	-0.04
SUBSITE	SEER_SUMMARY_STAGE	-3.33

2.3.1 Comments

- Registration date and date of birth of patient is strongly related and values of both dates varied together in same direction from their means.
- Registration date have negative covariance or relationship with sex.
- Date of Birth and age at presentation have positive covariance and have strong relationship.
- Registration date have positive covariance with Age at presentation, country, marital status, subsite and seer summary stage.
- Age at presentation have positively related with marital status.
- But negative covariance with sex, country, and subsite and seer summary stage, indicates that *higher* than average values of one age at presentation tend to be paired with *lower* than average values of the other attributes.
- Sex has positive covariance with marital status and seer summary stage but negative covariance with subsite.
- Country is positively related with subsite and seer summary stage.
- Marital status is negatively related with subsite and seer summary stage but positively related with subsite and subsite.
- Subsite has negative covariance with seer summary stage.

3 Frequent item set Mining

3.1 Need for mining frequent item sets

Frequent sets play an essential role in many Data Mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, clusters and many more of which the mining of association rules, is one of the most popular problems. The identification of sets of items, objects products, symptoms, characteristics, and so forth, that often occur together in the given database of year 2004-2008 is given here.

FP-Growth:

3.1.1 Minimum Support: .005

With this min support we got following result where no. of occurances are > 307.
No. of sets: **165,** Total Max size: **4**

1	0.432	ADDICTION = Unknown
1	0.370	ADDICTION = No
1	0.365	CLASS_OF_CASE = Class 1
1	0.330	ORGAN = BREAST
1	0.269	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)
1	0.239	FIRST_TREATMENT_ICD_DESC = Radiotherapy
1	0.239	FIRST_TREATMENT_ICD = V58.0
1	0.232	GRADE = III
1	0.232	FIRST_TREATMENT_ICD_DESC = Chemotherapy
1	0.232	FIRST_TREATMENT_ICD = V58.1
1	0.220	GRADE = II
1	0.219	CLASS_OF_CASE = Class 2
1	0.162	FIRST_TREATMENT_ICD_DESC = Surgery
1	0.162	FIRST_TREATMENT_ICD = V58.49
1	0.128	MORPHOLOGY = Squamous Cell Carcinoma, Nos
1	0.114	SUBSITE = Breast, NOS
1	0.097	MORPHOLOGY = Adenocarcinoma, Nos
1	0.087	CLASS_OF_CASE = Class 3
1	0.087	FIRST_TREATMENT_ICD_DESC = antineoplastic chemotherapy
1	0.087	FIRST_TREATMENT_ICD = V58.11
1	0.086	SUBSITE = Upper-outer quadrant of breast
1	0.070	GRADE = I
1	0.066	FIRST_TREATMENT_ICD_DESC = Palliative Care
1	0.066	FIRST_TREATMENT_ICD = V66.7
1	0.058	FIRST_TREATMENT_ICD_DESC = Hormonal
1	0.058	FIRST_TREATMENT_ICD = 99.24

3.1.1.1 Comments of itemset size 2

- 26352 patients addiction is unknown whereas 22725 patients have no addiction.
- 16521.98 patient's MORPHOLOGY=infilitrating Duct Carcinoma,Nos
- 14,679 patient's first treatment ICD Desc is Radiotherapy.
- 14,679 patient's first treatment ICD is V58.
- 14,249 patient's Grade is III.
- 14,249 patient's First treatment ICD Description is Chemotherapy.
- 13,512 Patient's GRADE=II
- 13,451 patient's class of case is Class 2.
- 9950 patient's first treatment ICD Description is Surgery.
- 7862 patient's MORPHOLOGY=squamous Cell Carcinoma
- 7,002 patient's subsite is Breast.
- 5,957 patient's MORPHOLOGY=Adenocarcinoma.
- 5343 patient's class of case is Class3.

- 5343 patient's first treatment ICD Description is "antineoplastic chemotherapy"
- 4054 patient's first treatment ICD Desc is "Palliative Care".

3	0.079	ADDICTION = Unknown	CLASS_OF_CASE = Class 1	ORGAN = BREAST
3	0.064	ADDICTION = Unknown	CLASS_OF_CASE = Class 1	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)
3	0.160	ADDICTION = Unknown	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)
3	0.087	ADDICTION = Unknown	ORGAN = BREAST	GRADE = III
3	0.066	ADDICTION = Unknown	ORGAN = BREAST	GRADE = II
3	0.073	ADDICTION = Unknown	ORGAN = BREAST	SUBSITE = Breast, NOS
3	0.084	ADDICTION = Unknown	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	GRADE = III
3	0.053	ADDICTION = Unknown	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	GRADE = II
3	0.052	ADDICTION = Unknown	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	SUBSITE = Breast, NOS
3	0.096	ADDICTION = Unknown	FIRST_TREATMENT_ICD_DESC = Radiotherapy	FIRST_TREATMENT_ICD = V58.0
3	0.099	ADDICTION = Unknown	FIRST_TREATMENT_ICD_DESC = Chemotherapy	FIRST_TREATMENT_ICD = V58.1
3	0.077	ADDICTION = Unknown	FIRST_TREATMENT_ICD_DESC = Surgery	FIRST_TREATMENT_ICD = V58.49
3	0.105	ADDICTION = No	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)
3	0.057	ADDICTION = No	ORGAN = BREAST	GRADE = III
3	0.054	ADDICTION = No	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	GRADE = III
3	0.086	ADDICTION = No	FIRST_TREATMENT_ICD_DESC = Radiotherapy	FIRST_TREATMENT_ICD = V58.0
3	0.093	ADDICTION = No	FIRST_TREATMENT_ICD_DESC = Chemotherapy	FIRST_TREATMENT_ICD = V58.1
3	0.063	ADDICTION = No	FIRST_TREATMENT_ICD_DESC = Surgery	FIRST_TREATMENT_ICD = V58.49
3	0.105	CLASS_OF_CASE = Class 1	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)
3	0.065	CLASS_OF_CASE = Class 1	ORGAN = BREAST	GRADE = III
3	0.061	CLASS_OF_CASE = Class 1	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	GRADE = III
3	0.084	CLASS_OF_CASE = Class 1	FIRST_TREATMENT_ICD_DESC = Radiotherapy	FIRST_TREATMENT_ICD = V58.0
3	0.070	CLASS_OF_CASE = Class 1	FIRST_TREATMENT_ICD_DESC = Chemotherapy	FIRST_TREATMENT_ICD = V58.1
3	0.064	CLASS_OF_CASE = Class 1	FIRST_TREATMENT_ICD_DESC = Surgery	FIRST_TREATMENT_ICD = V58.49
3	0.054	CLASS_OF_CASE = Class 1	FIRST_TREATMENT_ICD_DESC = antineoplastic chemoth	FIRST_TREATMENT_ICD = V58.11

3.1.1.2 Comments of itemset size 3

- 9336 patient's class of case is one & addiction cause is not given.
- 12,161 patient's diagnose cancer in Organ "Breast" & no addiction cause available for them.
- 9,827 patient's addiction is unknown having Marphology= Inflitrating Duct Carcinoma.
- 4,852 patient's addiction is unknown, class of case is "Class 1" and organ= "BREAST"
- 5,159 patient's addiction is unknown, Marphology=" infilitrating Duct Carcinoma,Nos" and GRADE=III
- 3,255 patient's addiction is unknown, Marphology=" infilitrating Duct Carcinoma,Nos" and GRADE=II.
- 3,194 patient's addiction is unknown, Marphology=" infilitrating Duct Carcinoma,Nos" and subsite=Breast.
- 5,856 patient's addiction is unknown, First Treatment ICD Description is "Radiotherapy" and first treatment ICD is V58.0
- 6,080 patient's addiction is unknown, First Treatment ICD Description is "Chemotherapy" and first treatment ICD is V58.1
- 4,729 patient's addiction is unknown, First Treatment ICD Description is "Surgery" and first treatment ICD is V58.49
- 5,282 patient's having no addiction, First Treatment ICD Description is "Radiotherapy" and first treatment ICD is V58.0

- 5,712 diagnosed patient's having no addiction, First treatment ICD Description is "chemotherapy" and first treatment ICD is V58.1
- 3,869 patient's having no addiction, First Treatment ICD Description is "Surgery" and first treatment ICD is V58.49
- 3,746 patient's class of case=class1, Marphology=" infilitrating Duct Carcinoma,Nos" and GRADE=III
- 5,159 patient's class of case = class1, First Treatment ICD Description is "Radiotherapy" and first treatment ICD is V58.0
- 4,300 patient's class of case = class1,First Treatment ICD Description is "Chemotherapy" and first treatment ICD is V58.1
- 3,930 patient's class of case = class1,First Treatment ICD Description is "Surgery" and first treatment ICD is V58.49
- 3,316 patient's class of case = class1, First Treatment ICD Description is antineoplastic chemotherapy and first treatment ICD is V58.11

4	0.064	ADDICTION = Unknown	CLASS_OF_CASE = Class 1	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)
4	0.084	ADDICTION = Unknown	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	GRADE = III
4	0.053	ADDICTION = Unknown	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	GRADE = II
4	0.052	ADDICTION = Unknown	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	SUBSITE = Breast, NOS
4	0.054	ADDICTION = No	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	GRADE = III
4	0.061	CLASS_OF_CASE = Class 1	ORGAN = BREAST	MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)	GRADE = III
4	0.061	ORGAN = BREAST	MORPHOLOGY = In	FIRST_TREATMENT_ICD_DESC = Radiotherapy	FIRST_TREATMENT_ICD = V58.0
4	0.058	ORGAN = BREAST	MORPHOLOGY = In	FIRST_TREATMENT_ICD_DESC = Chemotherapy	FIRST_TREATMENT_ICD = V58.1
4	0.066	ORGAN = BREAST	MORPHOLOGY = In	FIRST_TREATMENT_ICD_DESC = Surgery	FIRST_TREATMENT_ICD = V58.49

3.1.1.3 Comments of itemset size 4

- 3930 patient's addiction is unknown, class of case=class1, organ=BREAST, and Marphology=" infilitrating Duct Carcinoma,Nos"
- 5,159 patient's addiction is unknown,organ=BREAST, Marphology=" infilitrating Duct Carcinoma,Nos" and GRADE=III
- 3,255 patient's addiction is unknown,organ=BREAST, and Marphology=" infilitrating Duct Carcinoma,Nos" and GRADE=II
- 3,194 patient's addiction is unknown,organ=BREAST, and Marphology=" infilitrating Duct Carcinoma,Nos" and SUBSITE=Breast
- 3,3166 patients have no addiction, organ=BREAST, Marphology=" infilitrating Duct Carcinoma,Nos" and GRADE=III
- 3,747 patients Class of case= class 1, Organ=Breast, Marphology=" infilitrating Duct Carcinoma,Nos" and GRADE=III
- 3,747 patients Organ=Breast, Marphology=" infilitrating Duct Carcinoma,Nos" ,First treatment ICD DESC= Radiotherapy and First Treatment ICD is V58.0
- 3,562 patient's Organ=Breast, Marphology=" infilitrating Duct Carcinoma,Nos" ,First treatment ICD DESC= Chemotherapy and First Treatment ICD is V58.1
- 4,053 patient's Organ=Breast, Organ=Breast, First treatment ICD DESC= Surgery and First Treatment ICD is V58.49

3.1.1.4 Association Rules

The purpose of the analysis is to find associations between the attributes of diagnosed patients, i.e., to derive association rules that identify the attributes and co-occurrences of different attributes of patients that appear with the greatest (co-)frequencies

Min Confidence: 0.8 or 80%
Min Support: 0.5 %
Best association rules where confidence 100%. For example If X then (likely) Y where X and Y can be single values items, words, etc., or conjunctions of values, items, words, etc.

```
Association Rules
[CLASS_OF_CASE = Class 1, ORGAN = BREAST] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.804)
[ADDICTION = Unknown, ORGAN = BREAST] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.807)
[ADDICTION = Unknown, CLASS_OF_CASE = Class 1, ORGAN = BREAST] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.811)
[ADDICTION = Unknown, ORGAN = BREAST, GRADE = II] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.814)
[ORGAN = BREAST] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.814)
[ORGAN = BREAST, FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.818)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.0] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.818)
[ORGAN = BREAST, FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD = V58.0](confidence: 0.818)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.0] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD_DESC = Radiotherapy](confidence: 0.818)
[ORGAN = BREAST, FIRST_TREATMENT_ICD_DESC = Radiotherapy, FIRST_TREATMENT_ICD = V58.0] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)](confidence: 0.818)
T, FIRST_TREATMENT_ICD_DESC = Chemotherapy] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD = V58.1] (confidence: 0.84
T, FIRST_TREATMENT_ICD = V58.1] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD_DESC = Chemotherapy] (confidence: 0.84
T, FIRST_TREATMENT_ICD_DESC = Chemotherapy, FIRST_TREATMENT_ICD = V58.1] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.84
```

Comments

- 81% confidence if patient's diagnose cancer in Organ Breast and First treatment ICD Desc is radiotherapy then Morphology is infiltrating duct carcinoma.
- 82% confidence if patient's diagnosed cancer organ breast, first treatment ICD Desc Radiotherapy then Marophology is infiltrating duct carcinoma First treatment ICD=V58.0
- 83% confidence if patient organ is breast, first treatment ICD Desc is surgery then Morphology is infiltrating duct carcinoma and first treatment ICD is V58.49.
- 84% confidence if patient diagnosed cancer in organ Breast and First treatment ICD Description is Chemotheropy then morphology is infiltrating duct carcinoma and First treatment ICD=V58.1

```
[SUBSITE = Upper-outer quadrant of breast] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.857)
[SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST, MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.857)
[ORGAN = BREAST, SUBSITE = Upper-outer quadrant of breast] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.857)
[ORGAN = BREAST, CLASS_OF_CASE = Class 2] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.865)
[FIRST_TREATMENT_ICD_DESC = Hormonal] --> [ORGAN = BREAST] (confidence: 0.872)
[FIRST_TREATMENT_ICD = 99.24] --> [ORGAN = BREAST] (confidence: 0.872)
[FIRST_TREATMENT_ICD_DESC = Hormonal] --> [ORGAN = BREAST, FIRST_TREATMENT_ICD = 99.24] (confidence: 0.872)
[FIRST_TREATMENT_ICD = 99.24] --> [ORGAN = BREAST, FIRST_TREATMENT_ICD_DESC = Hormonal] (confidence: 0.872)
[FIRST_TREATMENT_ICD_DESC = Hormonal, FIRST_TREATMENT_ICD = 99.24] --> [ORGAN = BREAST] (confidence: 0.872)
[CLASS_OF_CASE = Class 1, ORGAN = BREAST, GRADE = III] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.934)
[ADDICTION = No, ORGAN = BREAST, GRADE = III] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.935)
[ORGAN = BREAST, GRADE = III] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.944)
[ADDICTION = Unknown, ORGAN = BREAST, GRADE = III] --> [MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] (confidence: 0.957)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), CLASS_OF_CASE = Class 2] --> [ORGAN = BREAST] (confidence: 0.995)
[ADDICTION = Unknown, MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), GRADE = III] --> [ORGAN = BREAST] (confidence: 0.997)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), GRADE = III] --> [ORGAN = BREAST] (confidence: 0.998)
[ADDICTION = Unknown, MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] --> [ORGAN = BREAST] (confidence: 0.998)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [ORGAN = BREAST] (confidence: 0.999)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD = V58.0] --> [ORGAN = BREAST] (confidence: 0.999)
```

Comments:

- 87% confidence if First treatment ICD=99.24 then organ is breast and first treatment ICD Desc=hormonal.
- 93% confidence if class of case is class1, organ=breast, grade=III then morphology is Infiltrating duct carcinoma.
- 99% confidence if morphology=infiltrating duct, first treatment ICD=V58.0 then Organ is breast.

[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] --> [ORGAN = BREAST] (confidence: 0.999)
[SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
[FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [FIRST_TREATMENT_ICD = V58.0] (confidence: 1.000)
[FIRST_TREATMENT_ICD = V58.0] --> [FIRST_TREATMENT_ICD_DESC = Radiotherapy] (confidence: 1.000)
[FIRST_TREATMENT_ICD_DESC = Chemotherapy] --> [FIRST_TREATMENT_ICD = V58.1] (confidence: 1.000)
[FIRST_TREATMENT_ICD = V58.1] --> [FIRST_TREATMENT_ICD_DESC = Chemotherapy] (confidence: 1.000)
[FIRST_TREATMENT_ICD_DESC = Surgery] --> [FIRST_TREATMENT_ICD = V58.49] (confidence: 1.000)
[FIRST_TREATMENT_ICD = V58.49] --> [FIRST_TREATMENT_ICD_DESC = Surgery] (confidence: 1.000)
[FIRST_TREATMENT_ICD_DESC = antineoplastic chemotherapy] --> [FIRST_TREATMENT_ICD = V58.11] (confidence: 1.000)
[FIRST_TREATMENT_ICD = V58.11] --> [FIRST_TREATMENT_ICD_DESC = antineoplastic chemotherapy] (confidence: 1.000)
[FIRST_TREATMENT_ICD_DESC = Palliative Care] --> [FIRST_TREATMENT_ICD = V66.7] (confidence: 1.000)
[FIRST_TREATMENT_ICD = V66.7] --> [FIRST_TREATMENT_ICD_DESC = Palliative Care] (confidence: 1.000)
[FIRST_TREATMENT_ICD_DESC = Hormonal] --> [FIRST_TREATMENT_ICD = 99.24] (confidence: 1.000)
[FIRST_TREATMENT_ICD = 99.24] --> [FIRST_TREATMENT_ICD_DESC = Hormonal] (confidence: 1.000)
[ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[ADDICTION = Unknown, FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [FIRST_TREATMENT_ICD = V58.0] (confidence: 1.000)
[ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.0] --> [FIRST_TREATMENT_ICD_DESC = Radiotherapy] (confidence: 1.000)
[ADDICTION = Unknown, FIRST_TREATMENT_ICD_DESC = Chemotherapy] --> [FIRST_TREATMENT_ICD = V58.1] (confidence: 1.000)
[ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.1] --> [FIRST_TREATMENT_ICD_DESC = Chemotherapy] (confidence: 1.000)
[ADDICTION = Unknown, FIRST_TREATMENT_ICD_DESC = Surgery] --> [FIRST_TREATMENT_ICD = V58.49] (confidence: 1.000)
[ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.49] --> [FIRST_TREATMENT_ICD_DESC = Surgery] (confidence: 1.000)
[ADDICTION = No, MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] --> [ORGAN = BREAST] (confidence: 1.000)
[ADDICTION = No, FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [FIRST_TREATMENT_ICD = V58.0] (confidence: 1.000)
[ADDICTION = No, FIRST_TREATMENT_ICD = V58.0] --> [FIRST_TREATMENT_ICD_DESC = Radiotherapy] (confidence: 1.000)
[ADDICTION = No, FIRST_TREATMENT_ICD_DESC = Chemotherapy] --> [FIRST_TREATMENT_ICD = V58.1] (confidence: 1.000)

```
[ADDICTION = No, FIRST_TREATMENT_ICD = V58.1] --> [FIRST_TREATMENT_ICD_DESC = Chemotherapy] (confidence: 1.000)
[ADDICTION = No, FIRST_TREATMENT_ICD_DESC = Surgery] --> [FIRST_TREATMENT_ICD = V58.49] (confidence: 1.000)
[ADDICTION = No, FIRST_TREATMENT_ICD = V58.49] --> [FIRST_TREATMENT_ICD_DESC = Surgery] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._)] --> [ORGAN = BREAST] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [FIRST_TREATMENT_ICD = V58.0] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, FIRST_TREATMENT_ICD = V58.0] --> [FIRST_TREATMENT_ICD_DESC = Radiotherapy] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, FIRST_TREATMENT_ICD_DESC = Chemotherapy] --> [FIRST_TREATMENT_ICD = V58.1] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, FIRST_TREATMENT_ICD = V58.1] --> [FIRST_TREATMENT_ICD_DESC = Chemotherapy] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, FIRST_TREATMENT_ICD_DESC = Surgery] --> [FIRST_TREATMENT_ICD = V58.49] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, FIRST_TREATMENT_ICD = V58.49] --> [FIRST_TREATMENT_ICD_DESC = Surgery] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, FIRST_TREATMENT_ICD_DESC = antineoplastic chemotherapy] --> [FIRST_TREATMENT_ICD = V58.11] (confidence: 1.000)
[CLASS_OF_CASE = Class 1, FIRST_TREATMENT_ICD = V58.11] --> [FIRST_TREATMENT_ICD_DESC = antineoplastic chemotherapy] (confidence: 1.000)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD_DESC = Chemotherapy] --> [ORGAN = BREAST] (confidence: 1.000)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD = V58.1] --> [ORGAN = BREAST] (confidence: 1.000)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), GRADE = II] --> [ORGAN = BREAST] (confidence: 1.000)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
[ORGAN = BREAST, FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [FIRST_TREATMENT_ICD = V58.0] (confidence: 1.000)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.0] --> [FIRST_TREATMENT_ICD_DESC = Radiotherapy] (confidence: 1.000)
[ORGAN = BREAST, FIRST_TREATMENT_ICD_DESC = Chemotherapy] --> [FIRST_TREATMENT_ICD = V58.1] (confidence: 1.000)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.1] --> [FIRST_TREATMENT_ICD_DESC = Chemotherapy] (confidence: 1.000)
[ORGAN = BREAST, FIRST_TREATMENT_ICD_DESC = Surgery] --> [FIRST_TREATMENT_ICD = V58.49] (confidence: 1.000)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.49] --> [FIRST_TREATMENT_ICD_DESC = Surgery] (confidence: 1.000)
[ORGAN = BREAST, FIRST_TREATMENT_ICD_DESC = Hormonal] --> [FIRST_TREATMENT_ICD = 99.24] (confidence: 1.000)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = 99.24] --> [FIRST_TREATMENT_ICD_DESC = Hormonal] (confidence: 1.000)
[MORPHOLOGY = Infiltrating Duct Carcinoma, Nos (C50._), FIRST_TREATMENT_ICD_DESC = Radiotherapy] --> [FIRST_TREATMENT_ICD = V58.0] (confidence: 1.00
```

Comments:

- These are best association rules as here confidence is 100% and minimum support is 0.5%.
- Left hand side of rules showing if part and right hand side showing then part means the occurances of attributes with relation of left part of rule.

3.1.2 Minimum Support: 1.0

With this min support we got following result where no. of occurrences are in range 1000-50000.

No. of sets: 145
Total Max size: 5

- Support value showing the percentage of item in complete dataset records. And size showing combination of item either occurring alone or with combination of others as at most 5 item size would show results of occurrences of 5 items together.

Size	Support	Item 1
1	0.824	MARITAL_STATUS = Married
1	0.567	SEX = Female
1	0.433	SEX = Male
1	0.432	ADDICTION = Unknown
1	0.370	ADDICTION = No
1	0.330	ORGAN = BREAST
1	0.293	TEHSIL = Lahore
1	0.239	FIRST_TREATMENT_ICD = V58.0
1	0.232	FIRST_TREATMENT_ICD = V58.1
1	0.162	FIRST_TREATMENT_ICD = V58.49
1	0.142	MARITAL_STATUS = Single
1	0.114	SUBSITE = Breast, NOS
1	0.087	FIRST_TREATMENT_ICD = V58.11
1	0.086	SUBSITE = Upper-outer quadrant of breast
1	0.066	FIRST_TREATMENT_ICD = V66.7
1	0.058	FIRST_TREATMENT_ICD = 99.24
1	0.057	ORGAN = LYMPH NODES
1	0.056	ADDICTION = Smoker

3.1.2.1 Comments of itemset size 1

1. 51000 diagnosed patients out of 62000 are married.
2. 8662 patients material status is single.
3. 35000 diagnosed patients are female and 26000 are male.
4. 20,000 patients effected organ is breast.
5. 18000 diagnosed patients belong to tehsil Lahore.
6. 3416 patients got cancer due to addiction of smoking.
7. 14600 diagnosed patients assigned First_treatment_icd=V58.0
8. 14250 patients assigned First_treatment_icd=V58.1
9. 1000 patients assigned First_treatment_icd=V58.49
10. 5344 patients assigned First_treatment_icd=V58.11
11. 4053 patients assigned First_treatment_icd=V66.7
12. 3477 patients effected organ is LYMPH NODES.

2	0.291	MARITAL_STATUS = Married	ADDICTION = No
2	0.239	SEX = Female	ADDICTION = No
2	0.132	SEX = Male	ADDICTION = No
2	0.052	MARITAL_STATUS = Married	FIRST_TREATMENT_ICD = 99.24
2	0.051	SEX = Female	FIRST_TREATMENT_ICD = 99.24
2	0.050	ORGAN = BREAST	FIRST_TREATMENT_ICD = 99.24
2	0.203	MARITAL_STATUS = Married	FIRST_TREATMENT_ICD = V58.0
2	0.133	SEX = Female	FIRST_TREATMENT_ICD = V58.0
2	0.106	SEX = Male	FIRST_TREATMENT_ICD = V58.0
2	0.096	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.0
2	0.086	ADDICTION = No	FIRST_TREATMENT_ICD = V58.0
2	0.074	ORGAN = BREAST	FIRST_TREATMENT_ICD = V58.0
2	0.069	TEHSIL = Lahore	FIRST_TREATMENT_ICD = V58.0
2	0.173	MARITAL_STATUS = Married	FIRST_TREATMENT_ICD = V58.1
2	0.126	SEX = Female	FIRST_TREATMENT_ICD = V58.1
2	0.106	SEX = Male	FIRST_TREATMENT_ICD = V58.1
2	0.099	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.1
2	0.093	ADDICTION = No	FIRST_TREATMENT_ICD = V58.1
2	0.069	ORGAN = BREAST	FIRST_TREATMENT_ICD = V58.1
2	0.064	TEHSIL = Lahore	FIRST_TREATMENT_ICD = V58.1
2	0.063	MARITAL_STATUS = Married	FIRST_TREATMENT_ICD = V58.11
2	0.137	MARITAL_STATUS = Married	FIRST_TREATMENT_ICD = V58.49
2	0.112	SEX = Female	FIRST_TREATMENT_ICD = V58.49
2	0.077	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.49
2	0.063	ADDICTION = No	FIRST_TREATMENT_ICD = V58.49
2	0.079	ORGAN = BREAST	FIRST_TREATMENT_ICD = V58.49
2	0.057	MARITAL_STATUS = Married	FIRST_TREATMENT_ICD = V66.7
2	0.056	SEX = Female	MARITAL_STATUS = Single
2	0.086	SEX = Male	MARITAL_STATUS = Single
2	0.069	ADDICTION = Unknown	MARITAL_STATUS = Single
2	0.066	ADDICTION = No	MARITAL_STATUS = Single
2	0.053	FIRST_TREATMENT_ICD = V:	MARITAL_STATUS = Single
2	0.294	MARITAL_STATUS = Married	ORGAN = BREAST
2	0.327	SEX = Female	ORGAN = BREAST
2	0.198	ADDICTION = Unknown	ORGAN = BREAST
2	0.127	ADDICTION = No	ORGAN = BREAST

3.1.2.2 Comments of itemset size 2

1. 17751 patients having material status married diagnosed cancer without any addiction.
2. 14580 female and 8052 male patients diagnosed cancer without any addiction.
3. 3132 female patients have first treatment ICD= 99.24
4. 3194 married patients have first treatment ICD= 99.24
5. 12,383 married patients have first treatment ICD=V58.0
6. 7,686 female patients have first treatment ICD=V58.1
7. 6,466 male patients have first treatment ICD= V58.1
8. 6,832 female patients have first treatment ICD=V58.49
9. 5,426 patients diagnosed cancer without any addiction and assigned first treatment ICD=V58.0
10. 3,842 patients diagnosed cancer without any addiction and assigned first treatment ICD=V58.49
11. 3,477 married patients assigned first treatment ICD=V66.7
12. 3,416 female diagnosed patients are single.
13. 5,246 male diagnosed patients are single.
14. 200086 female patients diagnosed cancer in organ Breast.

3	0.115	MARITAL_STATUS = Married	ADDICTION = No	ORGAN = BREAST
3	0.090	MARITAL_STATUS = Married	ADDICTION = No	TEHSIL = Lahore
3	0.070	MARITAL_STATUS = Married	ADDICTION = No	FIRST_TREATMENT_ICD = V58.0
3	0.065	MARITAL_STATUS = Married	ADDICTION = No	FIRST_TREATMENT_ICD = V58.1
3	0.052	MARITAL_STATUS = Married	ADDICTION = No	FIRST_TREATMENT_ICD = V58.49
3	0.126	SEX = Female	ADDICTION = No	ORGAN = BREAST
3	0.076	SEX = Female	ADDICTION = No	TEHSIL = Lahore
3	0.057	SEX = Female	ADDICTION = No	FIRST_TREATMENT_ICD = V58.0
3	0.057	SEX = Female	ADDICTION = No	FIRST_TREATMENT_ICD = V58.1
2	0.052	MARITAL_STATUS = Married	ADDICTION = Smoker	
2	0.052	SEX = Male	ADDICTION = Smoker	
2	0.346	MARITAL_STATUS = Married	ADDICTION = Unknown	
2	0.304	SEX = Female	ADDICTION = Unknown	
2	0.128	SEX = Male	ADDICTION = Unknown	
3	0.175	MARITAL_STATUS = Married	ADDICTION = Unknown	ORGAN = BREAST
3	0.111	MARITAL_STATUS = Married	ADDICTION = Unknown	TEHSIL = Lahore
3	0.079	MARITAL_STATUS = Married	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.0
3	0.071	MARITAL_STATUS = Married	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.1
3	0.064	MARITAL_STATUS = Married	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.49
3	0.065	MARITAL_STATUS = Married	ADDICTION = Unknown	SUBSITE = Breast, NOS
3	0.197	SEX = Female	ADDICTION = Unknown	ORGAN = BREAST
3	0.099	SEX = Female	ADDICTION = Unknown	TEHSIL = Lahore
3	0.069	SEX = Female	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.0
3	0.064	SEX = Female	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.1
3	0.062	SEX = Female	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.49
3	0.072	SEX = Female	ADDICTION = Unknown	SUBSITE = Breast, NOS
3	0.077	MARITAL_STATUS = Married	SEX = Female	SUBSITE = Upper-outer quadrant of breast

3.1.2.3 Comments of itemset size 3

- 3172 married patients diagnosed cancer due to addiction of smoking.
- 3172 male diagnosed cancer due to addiction of smoking.
- 4299 married patients got cancer without having any addiction and assigned first treatment ICD=V58.0
- 3194 married patients got cancer without having any addiction and assigned first treatment ICD=V58.49
- 3477 female patients have no addiction and assigned first treatment ICD=V58.1 or ICD=V58.0
- 4697 female married patients diagnosed cancer in upper-outer quadrant of breast.

4	0.061	MARITAL_STATUS = Married	ADDICTION = Unknown	ORGAN = BREAST	TEHSIL = Lahore
4	0.065	MARITAL_STATUS = Married	ADDICTION = Unknown	ORGAN = BREAST	SUBSITE = Breast, NOS
4	0.067	SEX = Female	ADDICTION = Unknown	ORGAN = BREAST	TEHSIL = Lahore
4	0.072	SEX = Female	ADDICTION = Unknown	ORGAN = BREAST	SUBSITE = Breast, NOS

3.1.2.4 Comment of itemset size 4

- 4392 female patients diagnosed cancer in organ breast and subsite=breast, NOS without any addiction.

4	0.174	MARITAL_STATUS = Married	SEX = Female	ADDICTION = Unknown	ORGAN = BREAST
4	0.086	MARITAL_STATUS = Married	SEX = Female	ADDICTION = Unknown	TEHSIL = Lahore
4	0.058	MARITAL_STATUS = Married	SEX = Female	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.0
4	0.053	MARITAL_STATUS = Married	SEX = Female	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.1
4	0.053	MARITAL_STATUS = Married	SEX = Female	ADDICTION = Unknown	FIRST_TREATMENT_ICD = V58.49
4	0.065	MARITAL_STATUS = Married	SEX = Female	ADDICTION = Unknown	SUBSITE = Breast, NOS
4	0.114	MARITAL_STATUS = Married	SEX = Female	ADDICTION = No	ORGAN = BREAST
4	0.065	MARITAL_STATUS = Married	SEX = Female	ADDICTION = No	TEHSIL = Lahore
4	0.100	MARITAL_STATUS = Married	SEX = Female	ORGAN = BREAST	TEHSIL = Lahore
4	0.066	MARITAL_STATUS = Married	SEX = Female	ORGAN = BREAST	FIRST_TREATMENT_ICD = V58.0
4	0.061	MARITAL_STATUS = Married	SEX = Female	ORGAN = BREAST	FIRST_TREATMENT_ICD = V58.1
4	0.070	MARITAL_STATUS = Married	SEX = Female	ORGAN = BREAST	FIRST_TREATMENT_ICD = V58.49
4	0.101	MARITAL_STATUS = Married	SEX = Female	ORGAN = BREAST	SUBSITE = Breast, NOS
4	0.077	MARITAL_STATUS = Married	SEX = Female	ORGAN = BREAST	SUBSITE = Upper-outer quadrant of breast

- 4,729 patients are female married, diagnosed cancer in organ Breast in subsite Upper-outer Quadrant of breast
- 6,142 female married patients diagnosed cancer in Breast belongs to Lahore.

5	0.061	MARITAL_STATUS = Married	SEX = Female	ADDICTION = Unknown	ORGAN = BREAST	TEHSIL = Lahore
5	0.065	MARITAL_STATUS = Married	SEX = Female	ADDICTION = Unknown	ORGAN = BREAST	SUBSITE = Breast, NOS

3.1.2.5 Comment of itemset size 5

- 3,685 female married patients diagnosed organ breast belong to tehsil lahore and addiction cause is unknown.

3.1.2.6 Association Rules

Min. Support 0.95 Minimum confidence 0.9

```
SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
MARITAL_STATUS = Married, SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female] (confidence: 1.000)
MARITAL_STATUS = Married, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
MARITAL_STATUS = Married, SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
SEX = Female, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female, ORGAN = BREAST] (confidence: 1.000)
SEX = Female, SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
ORGAN = BREAST, SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female] (confidence: 1.000)
ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
MARITAL_STATUS = Married, SEX = Female, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
MARITAL_STATUS = Married, SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female, ORGAN = BREAST] (confidence: 1.0
MARITAL_STATUS = Married, SEX = Female, SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.0
MARITAL_STATUS = Married, ORGAN = BREAST, SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female] (confidence: 1.0
MARITAL_STATUS = Married, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
SEX = Female, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
MARITAL_STATUS = Married, SEX = Female, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1
```

Comments:

- These are best association rules as here minimum confidence is 90% and minimum support is 95%.
- Left hand side of rules showing if part and right hand side showing then part means the occurances of attributes with relation of left part of rule.

Minimum Support 0.95
Min. Confidence 0.8

```
Association Rules
[ADDICTION = Unknown] --> [MARITAL_STATUS = Married] (confidence: 0.801)
[TEHSIL = Lahore, FIRST_TREATMENT_ICD = V58.1] --> [MARITAL_STATUS = Married] (confidence: 0.804)
[SEX = Female, FIRST_TREATMENT_ICD = V58.1] --> [MARITAL_STATUS = Married] (confidence: 0.814)
[ADDICTION = No, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.820)
[ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.821)
[MARITAL_STATUS = Married, ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.49] --> [SEX = Female] (confidence: 0.824)
[ADDICTION = No, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.828)
[SEX = Female, ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.1] --> [MARITAL_STATUS = Married] (confidence: 0.829)
[ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.49] --> [MARITAL_STATUS = Married] (confidence: 0.829)
[ADDICTION = No, FIRST_TREATMENT_ICD = V58.49] --> [MARITAL_STATUS = Married] (confidence: 0.830)
[SEX = Female, ADDICTION = No] --> [MARITAL_STATUS = Married] (confidence: 0.839)
[SEX = Female] --> [MARITAL_STATUS = Married] (confidence: 0.842)
[FIRST_TREATMENT_ICD = V58.49] --> [MARITAL_STATUS = Married] (confidence: 0.844)
[SEX = Female, ADDICTION = Unknown] --> [MARITAL_STATUS = Married] (confidence: 0.846)
[SEX = Male, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.848)
[ADDICTION = Unknown, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.849)
[SEX = Female, ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.849)
[FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.850)
[SEX = Female, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.851)
[SEX = Female, ADDICTION = No, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.853)
```

Comments:

- These are best association rules as here minimum confidence is 80% and minimum support is 95%.
- Left hand side of rules showing if part and right hand side showing then part means the occurances of attributes with relation of left part of rule.

```
[SEX = Male, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.856)
[SEX = Female, ADDICTION = Unknown, FIRST_TREATMENT_ICD = V58.49] --> [MARITAL_STATUS = Married] (confidence: 0.857)
[TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.860)
[SEX = Female, FIRST_TREATMENT_ICD = V58.49] --> [MARITAL_STATUS = Married] (confidence: 0.860)
[SEX = Female, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.863)
[FIRST_TREATMENT_ICD = V66.7] --> [MARITAL_STATUS = Married] (confidence: 0.868)
[FIRST_TREATMENT_ICD = 99.24] --> [ORGAN = BREAST] (confidence: 0.872)
[SEX = Female, ADDICTION = Unknown, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.876)
[TEHSIL = Lahore, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.879)
[ADDICTION = Unknown, ORGAN = BREAST] --> [MARITAL_STATUS = Married, SEX = Female] (confidence: 0.879)
[SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married, SEX = Female] (confidence: 0.882)
[SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married, SEX = Female, ORGAN = BREAST] (confidence: 0.882)
[ORGAN = BREAST, SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married, SEX = Female] (confidence: 0.882)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married, SEX = Female] (confidence: 0.882)
[FIRST_TREATMENT_ICD = 99.24] --> [SEX = Female] (confidence: 0.882)
[SEX = Female, ADDICTION = Unknown, ORGAN = BREAST] --> [MARITAL_STATUS = Married] (confidence: 0.884)
[ORGAN = BREAST] --> [MARITAL_STATUS = Married, SEX = Female] (confidence: 0.884)
[ADDICTION = Unknown, ORGAN = BREAST] --> [MARITAL_STATUS = Married] (confidence: 0.884)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.49] --> [MARITAL_STATUS = Married, SEX = Female] (confidence: 0.886)
[SEX = Female, ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.887)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.0] --> [MARITAL_STATUS = Married] (confidence: 0.887)
[SEX = Female, ORGAN = BREAST] --> [MARITAL STATUS = Married] (confidence: 0.890)
```

Comments:

- The above rules association is 85% to 89%.
- Left hand side of rules showing if part and right hand side showing then part means the occurances of attributes with relation of left part of rule.

```
ORGAN = BREAST, SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married] (confidence: 0.895)
SEX = Female, ORGAN = BREAST, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.897)
ORGAN = BREAST, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.898)
ADDICTION = No, ORGAN = BREAST] --> [MARITAL_STATUS = Married, SEX = Female] (confidence: 0.899)
FIRST_TREATMENT_ICD = 99.24] --> [MARITAL_STATUS = Married] (confidence: 0.899)
SEX = Female, ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.1] --> [MARITAL_STATUS = Married] (confidence: 0.900)
ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.1] --> [MARITAL_STATUS = Married] (confidence: 0.900)
SEX = Female, ADDICTION = Unknown, ORGAN = BREAST, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.900)
ADDICTION = Unknown, ORGAN = BREAST, TEHSIL = Lahore] --> [MARITAL_STATUS = Married] (confidence: 0.901)
SEX = Female, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married] (confidence: 0.902)
SEX = Female, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married, ORGAN = BREAST] (confidence: 0.90
SEX = Female, ADDICTION = Unknown, ORGAN = BREAST, SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married] (confidence: 0.90
ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married] (confidence: 0.903)
ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married, ORGAN = BREAST] (confidence: 0.903)
ADDICTION = Unknown, ORGAN = BREAST, SUBSITE = Breast, NOS] --> [MARITAL_STATUS = Married] (confidence: 0.903)
SEX = Female, ADDICTION = No, ORGAN = BREAST] --> [MARITAL_STATUS = Married] (confidence: 0.907)
ADDICTION = No, ORGAN = BREAST] --> [MARITAL_STATUS = Married] (confidence: 0.907)
ADDICTION = Smoker] --> [MARITAL_STATUS = Married] (confidence: 0.928)
ADDICTION = Smoker] --> [SEX = Male] (confidence: 0.939)
MARITAL_STATUS = Married, SUBSITE = Breast, NOS] --> [SEX = Female] (confidence: 0.985)
MARITAL_STATUS = Married, SUBSITE = Breast, NOS] --> [SEX = Female, ORGAN = BREAST] (confidence: 0.985)
MARITAL_STATUS = Married, ORGAN = BREAST, SUBSITE = Breast, NOS] --> [SEX = Female] (confidence: 0.985)
SUBSITE = Breast, NOS] --> [SEX = Female] (confidence: 0.986)
[SUBSITE = Breast, NOS] --> [SEX = Female, ORGAN = BREAST] (confidence: 0.986)
[ORGAN = BREAST, SUBSITE = Breast, NOS] --> [SEX = Female] (confidence: 0.986)
[MARITAL_STATUS = Married, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [SEX = Female] (confidence: 0.988)
[MARITAL_STATUS = Married, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [SEX = Female, ORGAN = BREAST] (confidence: 0.98
[MARITAL_STATUS = Married, ADDICTION = Unknown, ORGAN = BREAST, SUBSITE = Breast, NOS] --> [SEX = Female] (confidence: 0.98
[ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [SEX = Female] (confidence: 0.989)
[ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [SEX = Female, ORGAN = BREAST] (confidence: 0.989)
[ADDICTION = Unknown, ORGAN = BREAST, SUBSITE = Breast, NOS] --> [SEX = Female] (confidence: 0.989)
[MARITAL_STATUS = Married, ADDICTION = No, ORGAN = BREAST] --> [SEX = Female] (confidence: 0.991)
[MARITAL_STATUS = Married, ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.49] --> [SEX = Female] (confidence: 0.991)
[ADDICTION = No, ORGAN = BREAST] --> [SEX = Female] (confidence: 0.991)
[MARITAL_STATUS = Married, ORGAN = BREAST] --> [SEX = Female] (confidence: 0.992)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.49] --> [SEX = Female] (confidence: 0.992)
[ORGAN = BREAST] --> [SEX = Female] (confidence: 0.993)
[MARITAL_STATUS = Married, ADDICTION = Unknown, ORGAN = BREAST, TEHSIL = Lahore] --> [SEX = Female] (confidence: 0.993)
[MARITAL_STATUS = Married, ORGAN = BREAST, TEHSIL = Lahore] --> [SEX = Female] (confidence: 0.993)
[ADDICTION = Unknown, ORGAN = BREAST, TEHSIL = Lahore] --> [SEX = Female] (confidence: 0.994)
[ORGAN = BREAST, TEHSIL = Lahore] --> [SEX = Female] (confidence: 0.994)
[MARITAL_STATUS = Married, ADDICTION = Unknown, ORGAN = BREAST] --> [SEX = Female] (confidence: 0.994)
[MARITAL_STATUS = Married, ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.1] --> [SEX = Female] (confidence: 0.994)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.1] --> [SEX = Female] (confidence: 0.994)
[MARITAL_STATUS = Married, ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.0] --> [SEX = Female] (confidence: 0.994)
[ORGAN = BREAST, FIRST_TREATMENT_ICD = V58.0] --> [SEX = Female] (confidence: 0.994)
```

Comments:

- The above rules association is 89% to 99%.

- Left hand side of rules showing if part and right hand side showing then part means the occurances of attributes with relation of left part of rule.

```
[ADDICTION = Unknown, ORGAN = BREAST] --> [SEX = Female] (confidence: 0.994)
[SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female] (confidence: 1.000)
[SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
[MARITAL_STATUS = Married, SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female] (confidence: 1.000)
[MARITAL_STATUS = Married, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[MARITAL_STATUS = Married, SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
[SEX = Female, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female, ORGAN = BREAST] (confidence: 1.000)
[SEX = Female, SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
[ORGAN = BREAST, SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female] (confidence: 1.000)
[ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[MARITAL_STATUS = Married, SEX = Female, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[MARITAL_STATUS = Married, SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female, ORGAN = BREAST] (confidence: 1.000)
[MARITAL_STATUS = Married, SEX = Female, SUBSITE = Upper-outer quadrant of breast] --> [ORGAN = BREAST] (confidence: 1.000)
[MARITAL_STATUS = Married, ORGAN = BREAST, SUBSITE = Upper-outer quadrant of breast] --> [SEX = Female] (confidence: 1.000)
[MARITAL_STATUS = Married, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[SEX = Female, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.000)
[MARITAL_STATUS = Married, SEX = Female, ADDICTION = Unknown, SUBSITE = Breast, NOS] --> [ORGAN = BREAST] (confidence: 1.00
```

Comments:

- The above rules association is 100% these are best association rules with min support 95% and min confidence 80% .
- Left hand side of rules showing if part and right hand side showing then part means the occurances of attributes with relation of left part of rule.

4 Cluster Analysis

To see the trends, similarities and dissimilarities between difference instances we divided the whole data into five clusters. There were 61426 instances in the complete dataset. After making cluster, the dataset is divided as follows:

Cluster Id	Instances	Percentage
1	17495	28%
2	10640	17%
3	4539	7%
4	12677	21%
5	16075	26%

4.1 Overall Analysis

The following table gives the analysis of each cluster. This table is showing the centroid or most occurring value for each attribute of the cluster.

Attribute	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Age_At_Presentation	46.0775	46.341	40.2575	46.5432	47.5204
Sex	Female	Female	Female	Male	Male
District	Lahore	Lahore	Lahore	Lahore	Lahore
Province	Punjab	Punjab	Punjab	Punjab	Punjab
Country	Pakistan	Pakistan	Pakistan	Pakistan	Pakistan
Mrital_Status	Married	Married	Married	Married	Married
Occupation	Housewife	Housewife	Unknown	Other	Other
Relation	Missing	Missing	Missing	Missing	Missing
Addiction	Unknown	Unknown	Unknown	No	No
Previous_Treatment	No	Surgery	Surgery	No	No
Valid_Basis_of_Diagnosis	Missing	Missing	Pathology	Missing	Pathology
Diagnosis_At_SKM	Yes	No	Yes	No	Yes
Organ	Breast	Breast	Lymph nodes	Lymph Node	Breast
Grade	III	III	Missing	Missing	Missing
Laterality	Right	Left	No	No	No
Tumor_Size	241.1591	527.1914	879.8874	765.3826	550.3291
First_Treatment_ICD	V58.1 (Chemotherapy)	V58.49 (Surgery)	V58.1 (Chemotherapy)	V58.0 (Radiotherapy)	V58.49 (Surgery)
Class_Of_Death	Class 1	Missing	Missing	Class 2	Class 1

4.2 Detail Analysis

In this section we will analyse each cluster by defining the most commonly occurring values for each attribute.

4.2.1 Cluster 1

- **Age:**The mean age of patients of this cluster is 46.47
- **Gender:** 5% of people of this cluster are male and 94% are female.
- **District:** 40% people of this cluster belong to Lahore and 6% from Gujranwala, 5% from Faisalabad and 4% from Sialkot, and 2% from Gujrat, 2% from Sargodha,1% from Islamabad, 2% from Multan, 1% from Rawalpindi and 2% from Peshawar.
- **Province**: 85% patients from Punjab, 9% are from N.W.F.P, 1% from Sindh and 1% from Federal Capital.
- **Country:** 90% (17418) are from Pakistan. Only 77 patients are from Afghanistan that is very low quantity.
- **Marital Status:** 86% are married,8% are single and 4% are widow.
- **Occupation:** 58% patients are housewife,19% are Others and 10% occupation is unknown and 2% are teachers.
- **Relation:** 80% of people didn't have any one in family who ever had cancer. Only 4% people had cancer in their mother and 4% had in their sister.
- **Addiction:** 60% of patient's addiction is unknown and 34% patients' are not addicted to anything.
- **Previous Treatment:** 71**%** didn't have any previous treatment while only 20% are previously treated by surgery, 3% patient's treatment is unknown, 2% by Chemo and 1% by Hormonal.
- **Valid basis of diagnosis:** 41% of patients' cancer was diagnosed due to "Pathology", 4% due to consultant and data about 51% of patients is missing for this attribute.
- **Diagnoses at SKM:** The cancer of 80% patients of this cluster was diagnosed at SKM.
- **ORGAN:**73% of patients got cancer diagnosed in "Breast", 1% in "BONES_JOINTS_AND_ARTICULAR_CARTILAGE_OF_LIMBS", 1% in "CONNECTIVE_SUBCUTANEOUS_AND_OTHER_SOFT_TISSUES", 1% in "CERVIX_UTERI", 1% in "Lymph node", 1% in "BRONCHUS_AND_LUNG", 3% in "Ovary" and 1% in "HEMATOPOIETIC_AND_RETICULOENDOTHELIAL_SYSTEMS"
- **GRADE:** 43% patients have cancer of grade III and 26% have cancer of grade II and 26% patient's grade is missing.
- **Laterality:** 49% of data has "Right" value for this attribute, 31**%** has "left" and 13% has "No".

- **Tumor Size:** The value of centroid for this attribute is 188.549 +/-344.702.
- **First Treatment ICD:** 35% patients have ICD V58.1 (Chemotherapy), 20% have V58.0 (Radiotherapy), 13% have V58.49 (Surgery), 11% have 99.24(Hormonal), 8% have V58.11 and 6% patients' data was missing.
- **Class of Case:** 54% of patients belong to Class 1, 9% belongs to Class 2 and 33% patients' data was missing for this attribute.

4.2.2 Cluster 2

- **Age:**The mean age of patients of this cluster is 46.63
- **Gender:** 11% of people of this cluster are male and 88% are female.
- **District:** 33% people of this cluster belong to Lahore and 5% from Gujranwala, 4% from Faisalabad and 4% from Sialkot, and 2% from Gujrat, 2% from Sargodha, 2% from Multan, 2% from Islamabad, 3% from Rawalpindi and 3% from Peshawar.
- **Province**: 78% patients from Punjab, 12% are from N.W.F.P, 2% from Sindh , 1% from Baloshistan and 2% from Federal Capital.
- **Country:** 90% (10573) are from Pakistan. Only 59 patients are from Afghanistan and 2 from kenya.
- **Marital Status:** 87% are married,7% are single and 3% are widow.
- **Occupation:** 52% patients are housewife,16% are Others,17% occupation is unknown, 2% teacher and 1% student.
- **Relation:** 80% of people didn't have any one in family who ever had cancer. Only 4% people had cancer in their mother and 4% had in their sister.
- **Addiction:** 66% of patient's addiction is unknown, 21% patients' are not addicted to anything and 1% is smokers.
- **Previous Treatment:** 17**%** didn't have any previous treatment while only 64% are previously treated by surgery, 3% patient's treatment is unknown, 7% by Chemo, 2% by Hormonal and 4% by radiation.
- **Valid basis of diagnosis:** 34% of patients' cancer was diagnosed due to "Pathology", 15% due to consultant and data about 48% of patients is missing for this attribute.
- **Diagnoses at SKM:** The cancer of 16% patients of this cluster was diagnosed at SKM.
- **ORGAN:**61% of patients got cancer diagnosed in "Breast", 1% in "HEMATOPOIETIC_AND_RETICULOENDOTHELIAL_SYSTEMS ", 1% in " PROSTATE_GLAND", 1% in "THYROID GLAND", 1% in "Bladder", 2% in "CERVEX UTERI", 1% in " CONNECTIVE_SUBCUTANEOUS_AND_OTHER_SOFT_TISSUES", 1% in "Brain", 1% in " BONES_JOINTS_AND_ARTICULAR_CARTILAGE_OF_LIMBS",1% in "Colon", 1% in "Rectum", 1% in "Corpus Uteri" and 5% in "Ovary"

- **GRADE:** 26% patients have cancer of grade III and 27% have cancer of grade II,5% have of grade I and 39% patient's grade is missing.
- **Laterality:** 22% of data has "Right" value for this attribute, 48**%** has "left" and 18% has "No".
- **Tumor Size:** The value of centroid for this attribute is 583.08.
- **First Treatment ICD:** 7% patients have ICD V58.1 (Chemotherapy), 18% have V58.0 (Radiotherapy), 32% have V58.49 (Surgery), 9% have 99.24(Hormonal), 4% have V58.11 (antineoplastic_chemotherapy), 2% have V66.7 (Palliative_Care) and 24% patients' data was missing.
- **Class of Case:** 3% of patients belong to Class 1, 26% belongs to Class 2, 20% belongs to Class 3 and 49% patients' data was missing for this attribute.

4.2.3 Cluster 3

- **Age:**The mean age of patients of this cluster is 36.45
- **Gender:**77% of people of this cluster are male and 22% are female.
- **District:** 5% people of this cluster belong to Lahore and 3% from Gujranwala, 2% from Faisalabad and 2% from Sialkot, and 1% from Gujrat, 2% from Sargodha, 1% from Multan, 1% from Islamabad, 5% from Swat, 3% from Rawalpindi, 4% from Mardan, 2% from Noshera and 19% from Peshawar.
- **Province**: 35% patients from Punjab, 51% are from N.W.F.P, 2% from Sindh, 2% from Baloshistan, 2% from FATA and 1% from Federal Capital.
- **Country:** 98% are from Pakistan. Only 1% are from Afghanistan.
- **Marital Status:** 64% are married and 35% are single.
- **Occupation:** 7% patients are housewife, 21% are Others, 28% occupation is unknown, 6% labourer, 2% farmer, 2% retired, 1% teacher, 1% businessman, 1% shopkeeper and 10% student.
- **Relation:** 93% of people didn't have any one in family who ever had cancer. Only 1% people had cancer in their brothers.
- **Addiction:** 60% of patient's addiction is unknown, 17% patients' are not addicted to anything, 6% are addicted to Naswar, 2% are Ex.Smoker, 2% are chronic-Smoker and 5% is smokers.
- **Previous Treatment:** 71**%** didn't have any previous treatment while only 16% are previously treated by surgery, 4% patient's treatment is unknown, 5% by Chemo and 2% by radiation.
- **Valid basis of diagnosis:** 48% of patients' cancer was diagnosed due to "Pathology", 8% due to consultant and data about 41% of patients is missing for this attribute.

- **Diagnoses at SKM:** The cancer of 58% patients of this cluster was diagnosed at SKM.
- **ORGAN:**31% of patients got cancer diagnosed in “Lymph node”,2% in “ ESOPHAGUS”, 1% in “ PROSTATE_GLAND”, 1% in “THYROID GLAND”, 1% in “Bladder”, 1% in “CERVEX UTERI”, 2% in “ CONNECTIVE_SUBCUTANEOUS_AND_OTHER_SOFT_TISSUES”, 2% in “Stomach”, 1% in “ BONES_JOINTS_AND_ARTICULAR_CARTILAGE_OF_LIMBS”,1% in “Colon”, 3% in “Brain”, 1% in “Corpus Uteri”, 3% in “Rectum”, 1% in “Bronchus and Lungs” and 1% in “Colon”
- **GRADE:** 6% patients have cancer of grade III and 9% have cancer of grade II,5% have of grade I, 1% have of grade IV and 76% patient’s grade is missing.
- **Laterality:** 6% of data has “Right” value for this attribute, 6**%** has “left” and 80% has “No”.
- **Tumor Size:** The value of centroid for this attribute is 909.74.
- **First Treatment ICD:** 71% patients have ICD V58.1 (Chemotherapy), 8% have V58.0 (Radiotherapy), 5% have V58.11 (antineoplastic_chemotherapy), 2% have V66.7(Palliative_Care) and 10% patients’ data was missing.
- **Class of Case:** 33% of patients belong to Class 1, 17% belongs to Class 2, 7% belongs to Class 3 and 40% patients’ data was missing for this attribute.

4.2.4 Cluster 4

- **Age:**The mean age of patients of this cluster is 46.5
- **Gender:**71% of people of this cluster are male and 28% are female.
- **District:** 29% people of this cluster belong to Lahore and 4% from Gujranwala, 3% from Faisalabad and 3% from Sialkot, and 2% from Gujrat, 2% from Sargodha, 1% from Multan, 1% from Islamabad, 5% from Swat, 3% from Rawalpindi, 1% from Mardan and 2% from Peshawar.
- **Province**: 74% patients from Punjab, 14% are from N.W.F.P, 3% from Sindh, 2% from Baloshistan, 1% from FATA and 1% from Federal Capital.
- **Country:** 98% are from Pakistan. Only 1% are from Afghanistan.
- **Marital Status:** 81% are married, 16% are single and 1% are widow.
- **Occupation:** 11% patients are housewife, 41% are Others, 9% occupation is unknown, 5% labourer, 1% driver, 2% farmer, 4% retired, 1% teacher, 2% businessman, 1% shopkeeper and 4% student.

- **Relation:** 91% of people didn't have any one in family who ever had cancer. Only 1% people had cancer in their brothers, 1% in father and 1% in mother.
- **Addiction:** 25% of patient's addiction is unknown, 40% patients' are not addicted to anything, 5% are addicted to Naswar, 3% are Ex.Smoker, 3% are chronic-Smoker, 3% is bettle leaf and 10% is smokers.
- **Previous Treatment:** 65**%** didn't have any previous treatment while only 23% are previously treated by surgery, 1% by Hormonal, 2% patient's treatment is unknown, 4% by Chemo and 2% by radiation.
- **Valid basis of diagnosis:** 26% of patients' cancer was diagnosed due to "Pathology", 15% due to consultant and data about 55% of patients is missing for this attribute.
- **Diagnoses at SKM:** The cancer of 14% patients of this cluster was diagnosed at SKM.
- **ORGAN:**8% of patients got cancer diagnosed in "Lymph node",2% in "Breast", 2% in " Tongue", 14% in " HEMATOPOIETIC_AND_RETICULOENDOTHELIAL_SYSTEMS", 5% in " PROSTATE_GLAND", 1% in "THYROID GLAND", 3% in "Bladder", 1% in "CERVEX UTERI", 2% in " CONNECTIVE_SUBCUTANEOUS_AND_OTHER_SOFT_TISSUES", 2% in "Stomach", 3% in " BONES_JOINTS_AND_ARTICULAR_CARTILAGE_OF_LIMBS",1% in "Colon", 3% in "Esophagus", 4% in "Brain", 1% in "Corpus Uteri", 5% in "Rectum", 3% in "Bronchus and Lungs", 5% in "Larynx"
- **GRADE:** 13% patients have cancer of grade III and 19% have cancer of grade II, 11% have of grade I, 1% have of grade IV and 53% patient's grade is missing.
- **Laterality:** 9% of data has "Right" value for this attribute, 7**%** has "left" and 77% has "No".
- **Tumor Size:** The value of centroid for this attribute is 760.8
- **First Treatment ICD:** 13% patients have ICD V58.1 (Chemotherapy), 52% have V58.0 (Radiotherapy), 9% have V58.11 (antineoplastic_chemotherapy), 1% have 99.24(Hormonal), 9% have V66.7(Palliative_Care) and 12% patients' data was missing.
- **Class of Case:** 8% of patients belong to Class 1, 58% belongs to Class 2, 12% belongs to Class 3 and 20% patients' data was missing for this attribute.

4.2.5 Cluster 5

- **Age:**The mean age of patients of this cluster is 47.87
- **Gender:**73% of people of this cluster are male and 26% are female.

- **District:** 35% people of this cluster belong to Lahore and 6% from Gujranwala, 5% from Faisalabad and 3% from Sialkot, and 2% from Gujrat, 2% from Sargodha, 2% from Multan, 1% from Islamabad, 5% from Swat, 1% from Rawalpindi, 1% from Mardan and 2% from Peshawar.
- **Province**: 81% patients from Punjab, 11% are from N.W.F.P, 1% from Sindh, 1% from Baloshistan, 0% from FATA and 1% from Federal Capital.
- **Country:** 98% are from Pakistan. Only 1% are from Afghanistan.
- **Marital Status:** 80% are married, 17% are single and 1% are widow.
- **Occupation:** 11% patients are housewife, 41% are Others, 9% occupation is unknown, 4% labourer, 1% driver, 2% farmer, 5% retired, 1% teacher, 3% businessman, 1% shopkeeper and 4% student.
- **Relation:** 90% of people didn't have any one in family who ever had cancer. Only 1% people had cancer in their brothers, 1% in father, 1% in sister and 1% in mother.
- **Addiction:** 18% of patient's addiction is unknown, 47% patients' are not addicted to anything, 3% are addicted to Naswar, 5% are Ex.Smoker, 4% are chronic-Smoker, 2% is bettle leaf and 9% is smokers.
- **Previous Treatment:** 76**%** didn't have any previous treatment while only 14% are previously treated by surgery, 1% by Hormonal, 3% patient's treatment is unknown, 2% by Chemo and 1% by radiation.
- **Valid basis of diagnosis:** 45% of patients' cancer was diagnosed due to "Pathology", 9% due to consultant and data about 40% of patients is missing for this attribute.
- **Diagnoses at SKM:** The cancer of 86% patients of this cluster was diagnosed at SKM.
- **ORGAN:**4% of patients got cancer diagnosed in "Lymph node",13% in " LIVER_AND_INTRAHEPATIC_BILE_DUCTS", 3% in "Breast", 4% in " Tongue", 4% in " HEMATOPOIETIC_AND_RETICULOENDOTHELIAL_SYSTEMS", 3% in " PROSTATE_GLAND", 1% in "THYROID GLAND", 3% in "Bladder", 1% in "CERVEX UTERI", 2% in " CONNECTIVE_SUBCUTANEOUS_AND_OTHER_SOFT_TISSUES", 3% in "Stomach", 2% in " BONES_JOINTS_AND_ARTICULAR_CARTILAGE_OF_LIMBS",3% in "Colon", 2% in "Esophagus", 2% in "Brain", 1% in "Corpus Uteri", 4% in "Rectum", 5% in "Bronchus and Lungs", 2% in "Larynx"
- **GRADE:** 11% patients have cancer of grade III and 18% have cancer of grade II, 8% have of grade I, 1% have of grade IV and 59% patient's grade is missing.
- **Laterality:** 9% of data has "Right" value for this attribute, 10**%** has "left" and 72% has "No".
- **Tumor Size:** The value of centroid for this attribute is 571.89

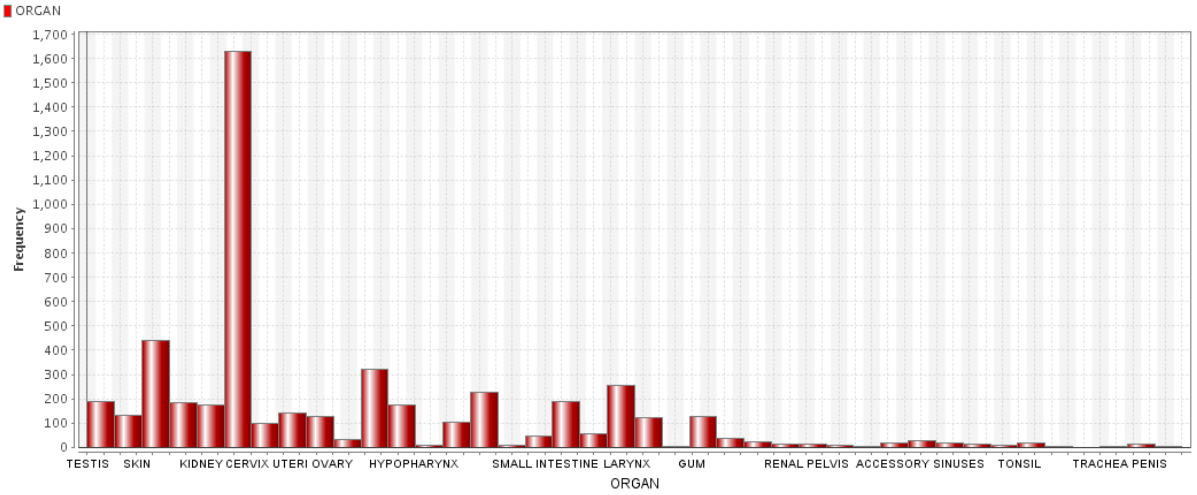
- **First Treatment ICD:** 13% patients have ICD V58.1 (Chemotherapy), 13% have V58.0 (Radiotherapy), 24% have V58.39(Surgery),11% have V58.11 (antineoplastic_chemotherapy), 1% have 99.24(Hormonal), 12% have V66.7(Palliative_Care) and 23% patients' data was missing.
- **Class of Case:** 61% of patients belong to Class 1, 5% belongs to Class 2, 4% belongs to Class 3, 2% belongs to Class 0 and 25% patients' data was missing for this attribute.

5 References

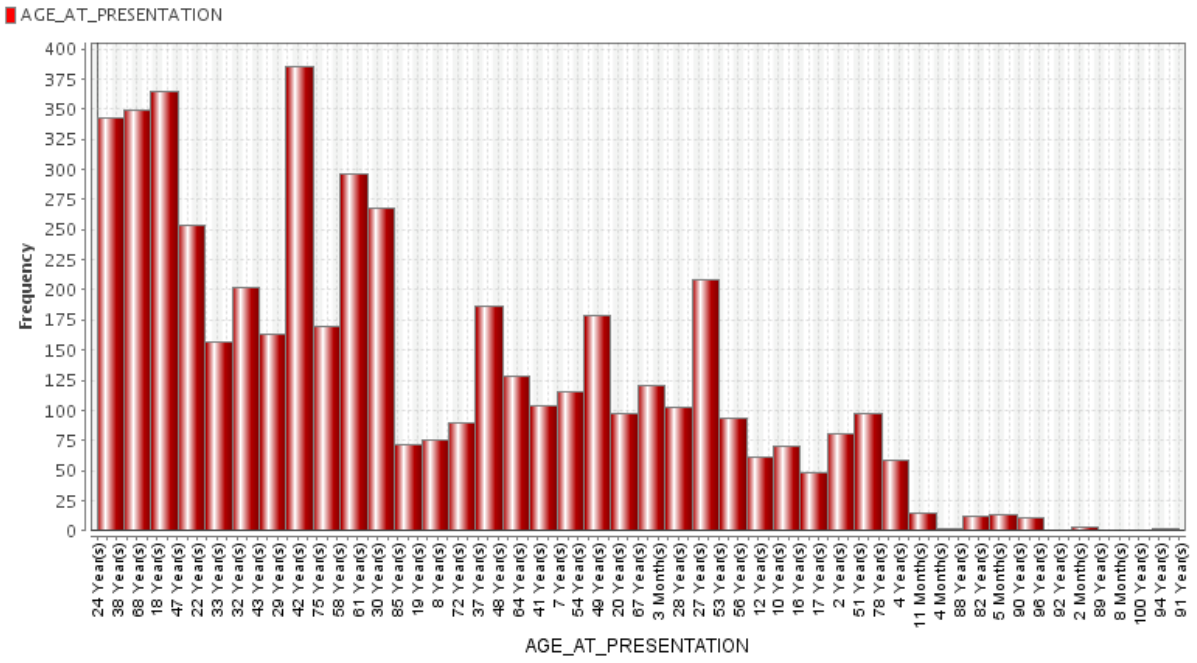
http://meteor.aihw.gov.au/content/index.phtml/itemId/269647/
http://www.cancer.gov/cancertopics/factsheet/Detection/tumor-grade
http://en.wikipedia.org/wiki/Pathology/
http://meteor.aihw.gov.au/content/index.phtml/itemId/289198/
http://en.wikipedia.org/wiki/Surgical_margin/

6 Appendix A

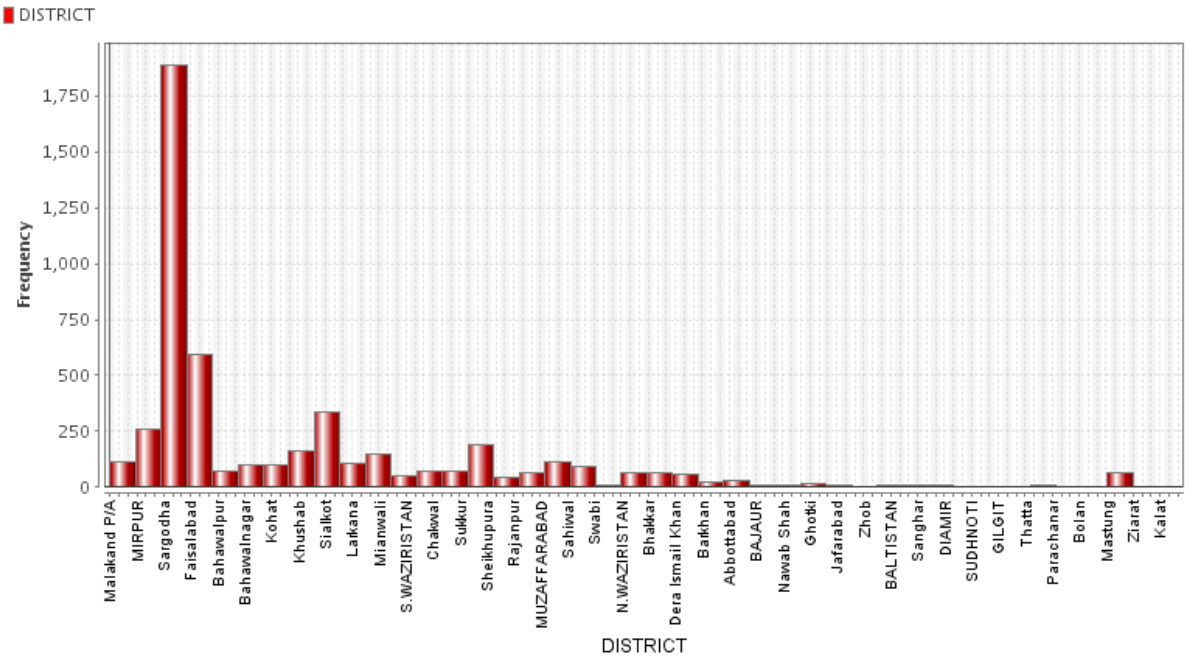
The following histograms shows the important facts about each attribute of the dataset.



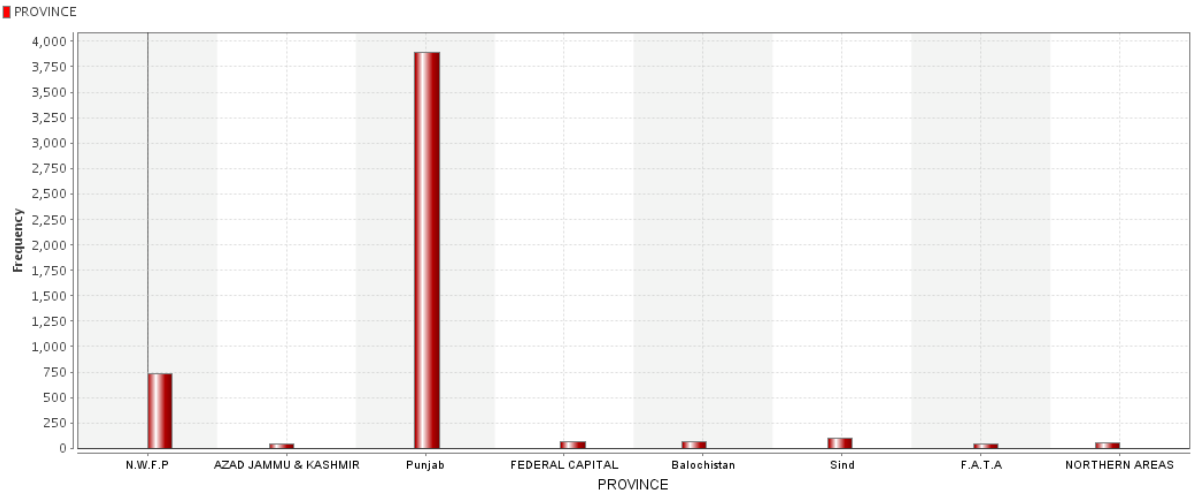
The above graphs shows that the most common cancer diagnosed is the organ “Cervix”.



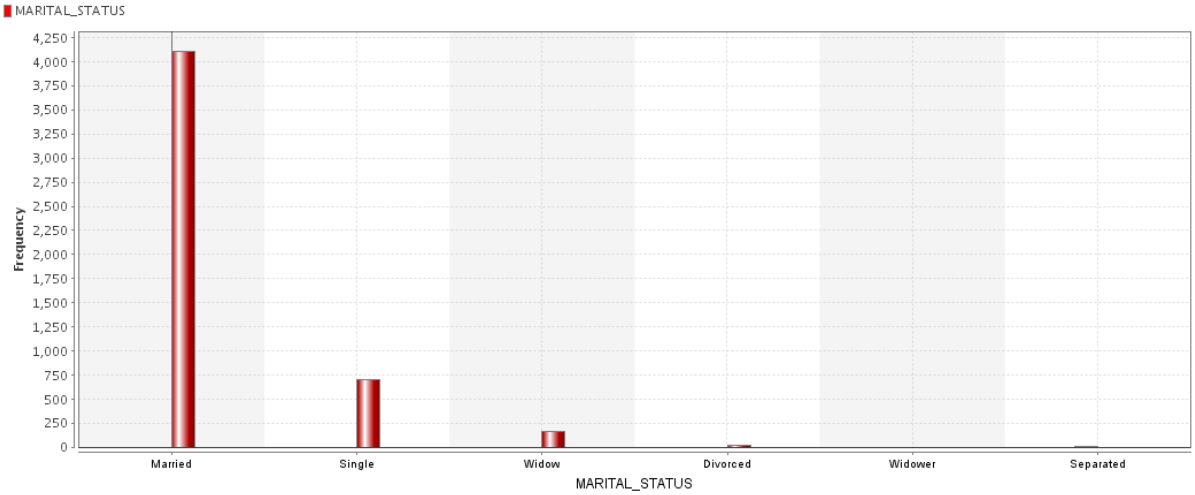
Mostly cancer is diagnosed in people having age 42 years.



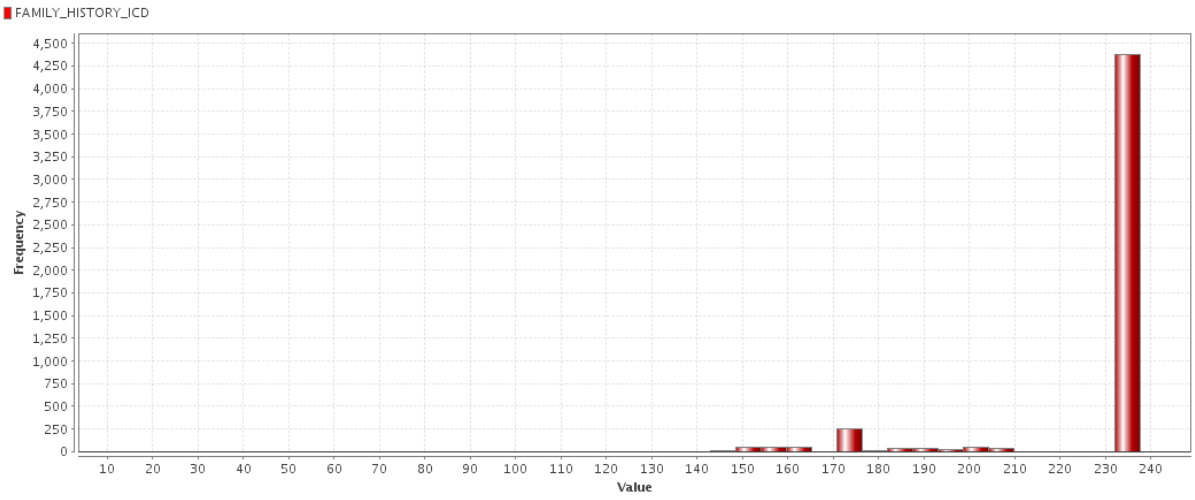
The people of Lahore are mostly affected by Cancer by given statistics.



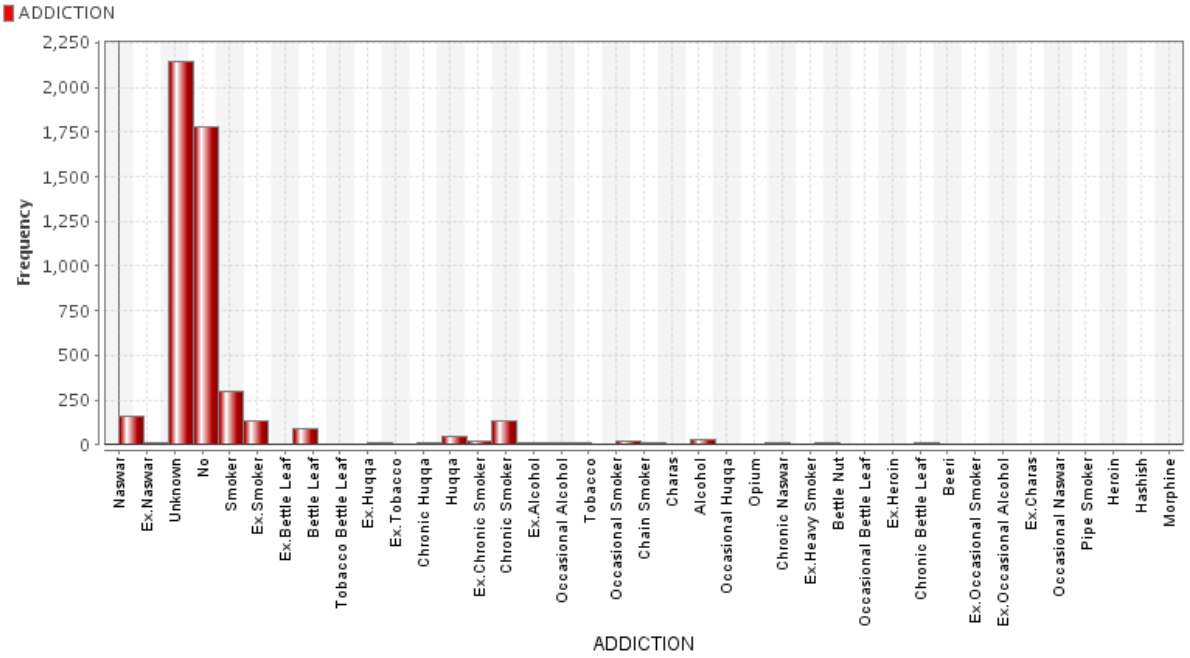
Province of Punjab has the most number of cancer patients.



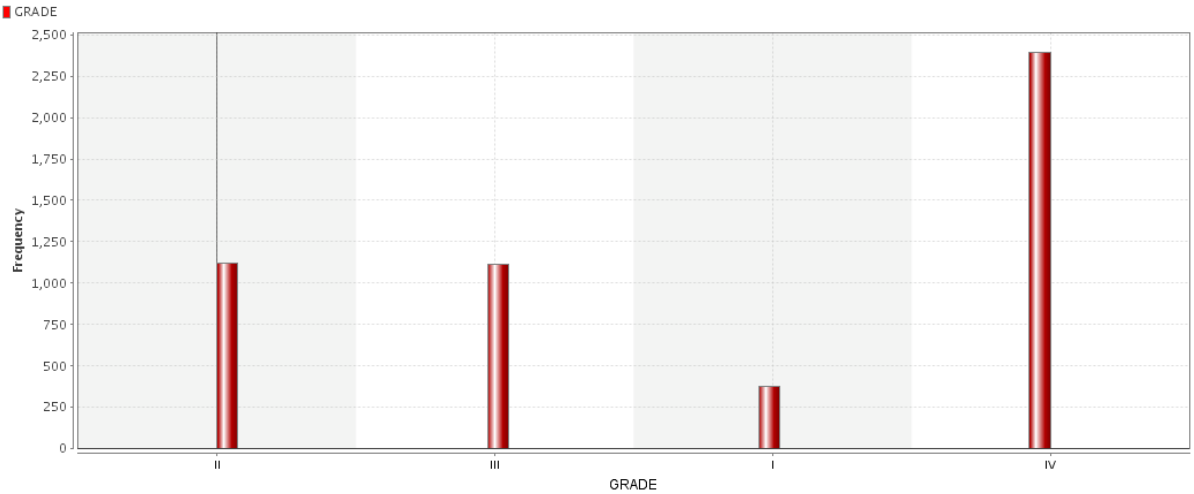
The above histogram displays that Marriad people are the frequent patient of cancer.



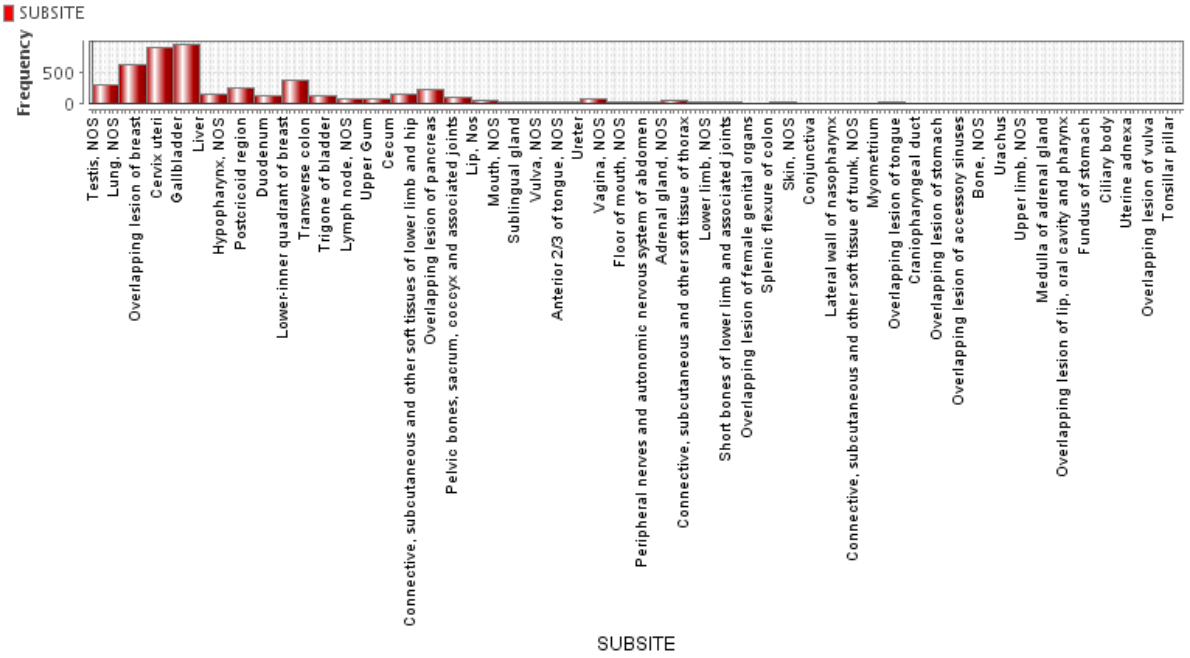
Usually the people who have ralatives having ICD 230-240 are mostly affected by the cancer



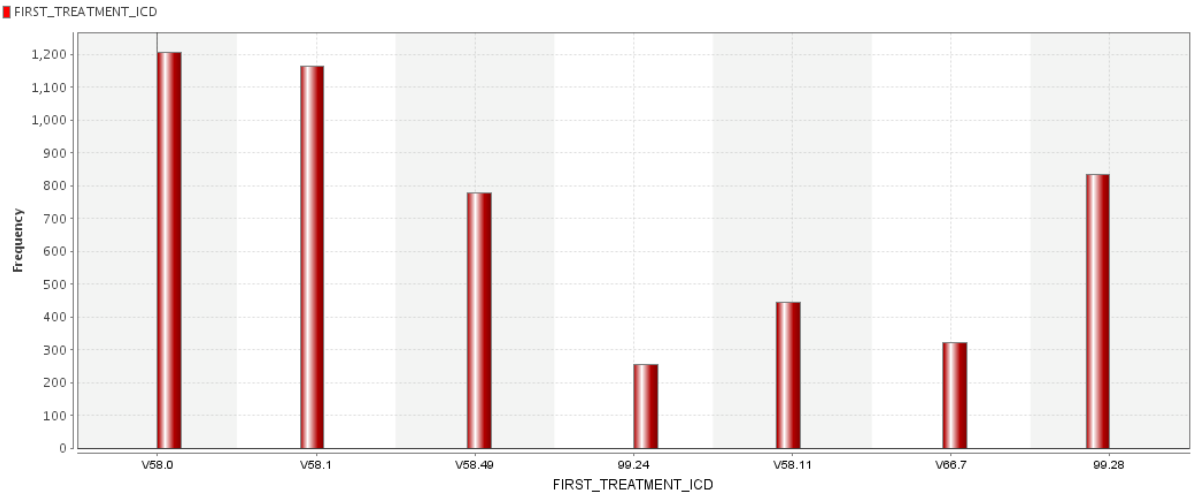
Mostly the cancer patients donot have any Addiction.



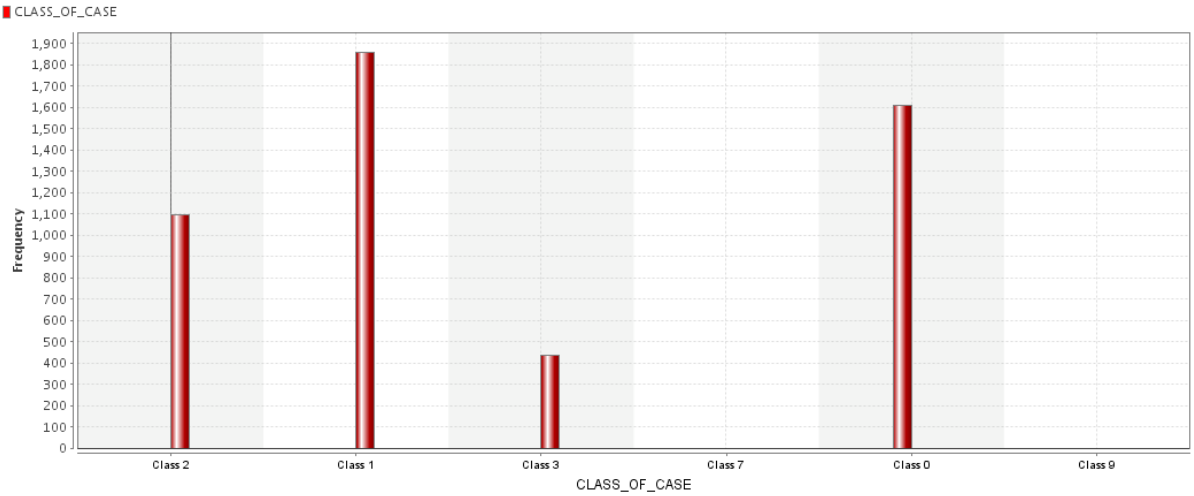
The most common cancer diagnosed is of grade IV.



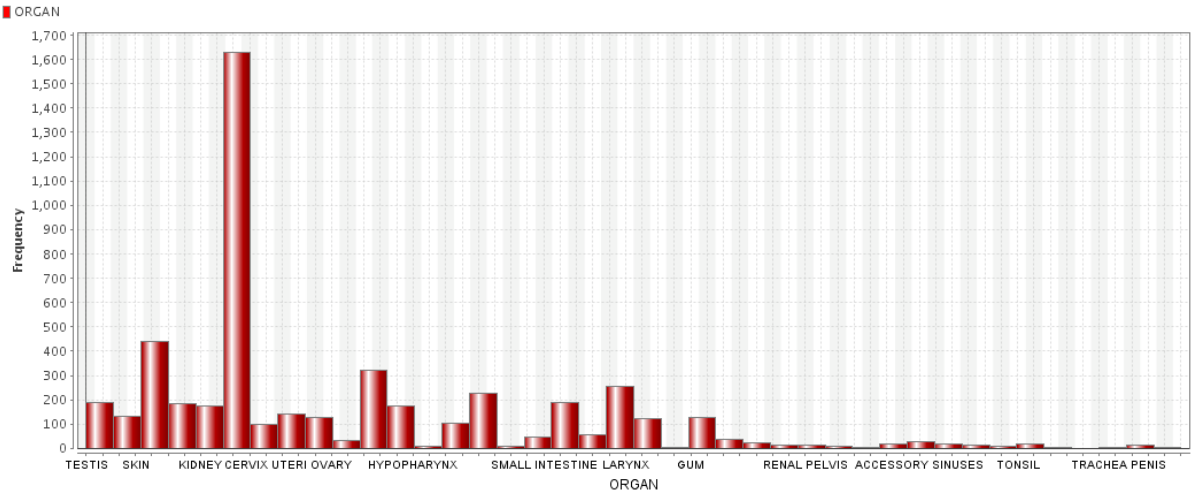
Mostly cancer is detected in Cervix uteri and Gallblader.



The most common ICD detected is V58.0



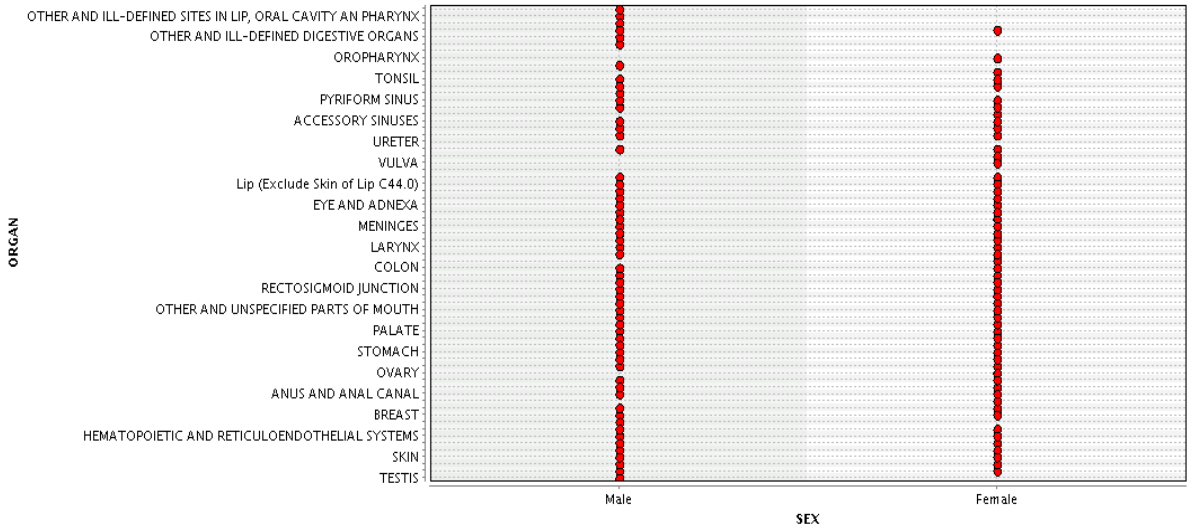
Mostly the people have cancer of Class1.



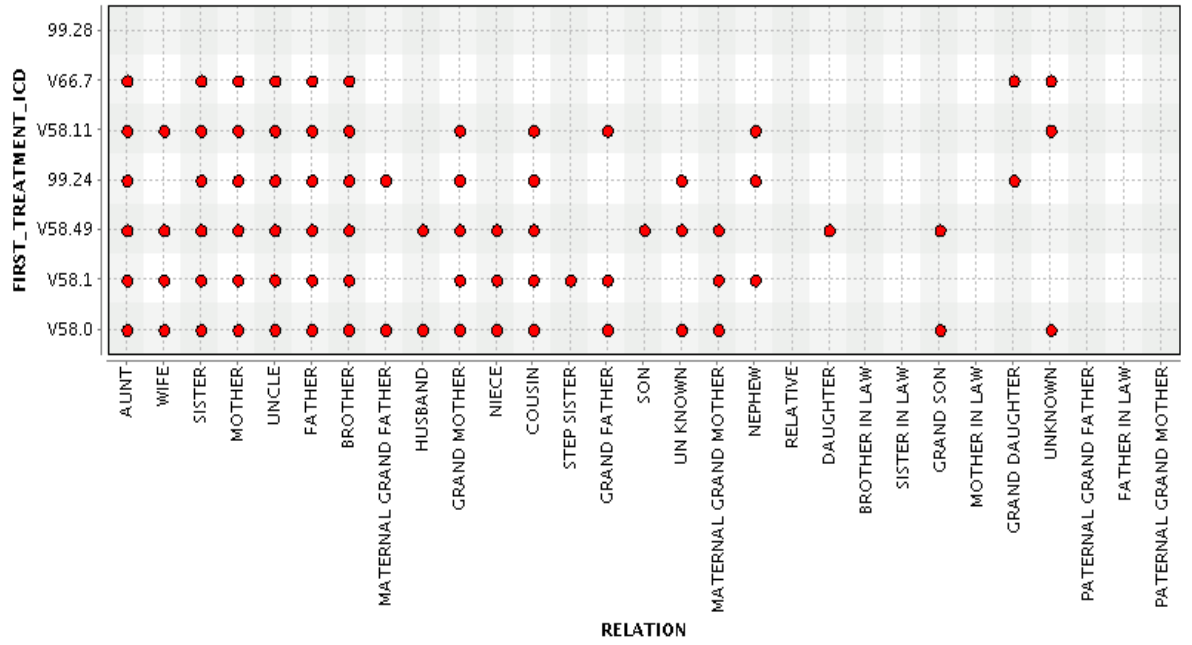
The most common type of cancer detected is Breast Cancer.

7 Appendix B

Scatter Plots for visualization of data

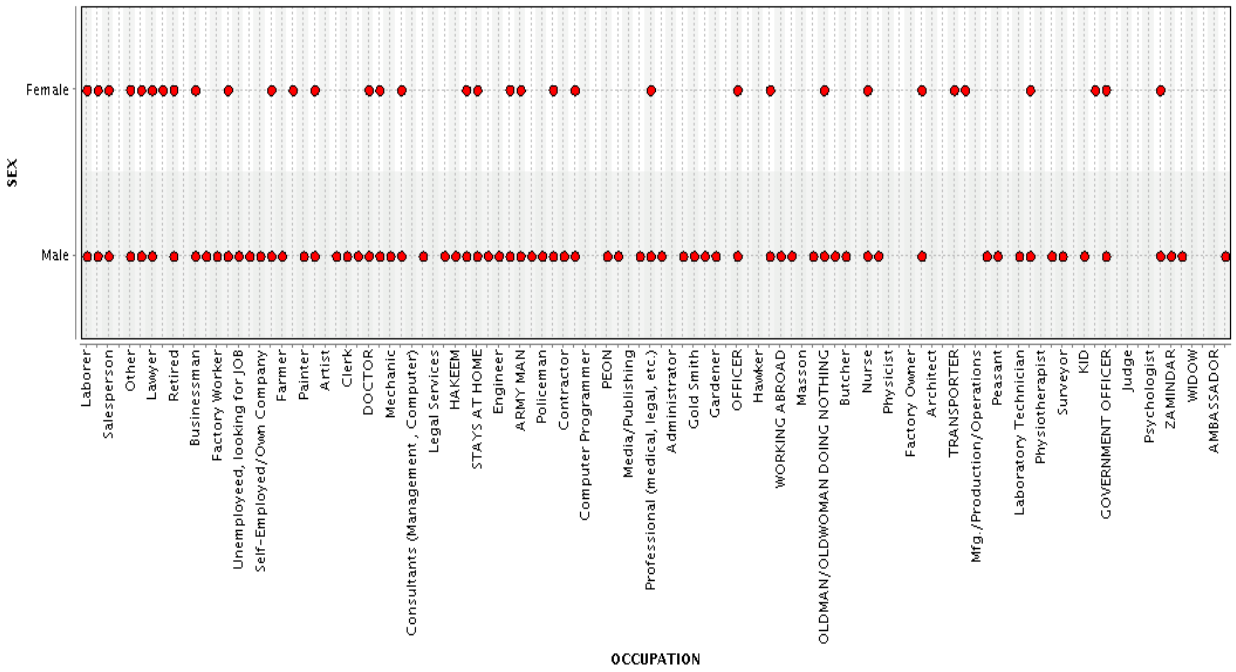


This graph shows that **females** don't have any cancer diagnosed for the "ill-defined digestive organs" and "ill-defined sites in lip, oral cavity an pharynx" and mostly diagnose in these organs; breast, ovary and anal canal.
And **males** don't have cancer diagnosed in vulva, or pharynx and ovary. Whereas mostly diagnosed at Testis and colon.

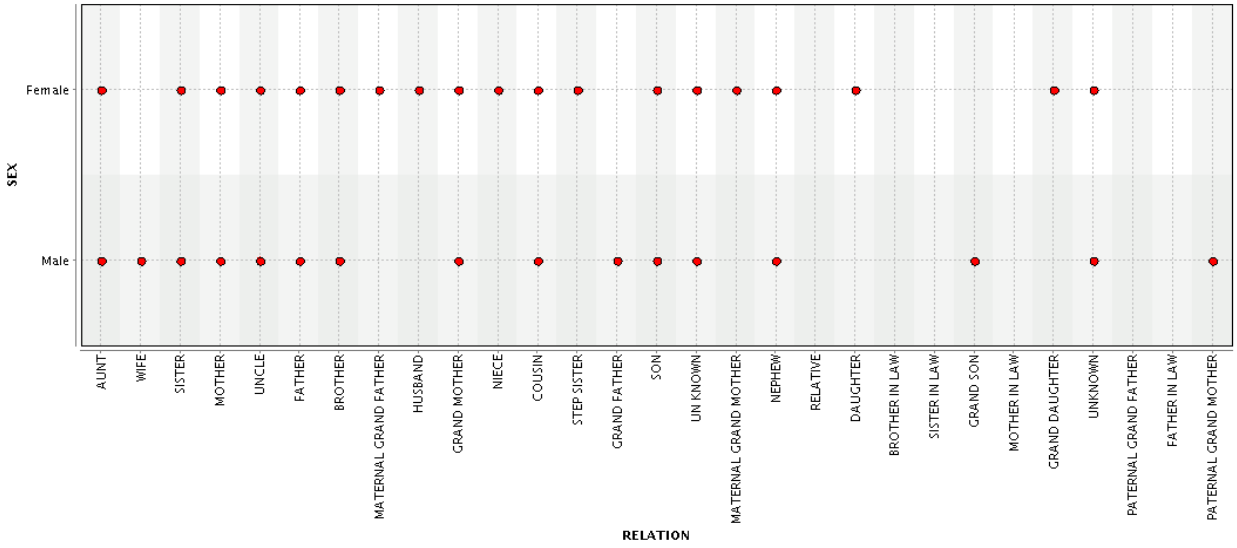


This graph shows the ICD numbers that are assigned according to specific relations.
The symbol * showing the presence of ICD allot due to relation showing in that row. Blank cells showing no ICD against that relation .

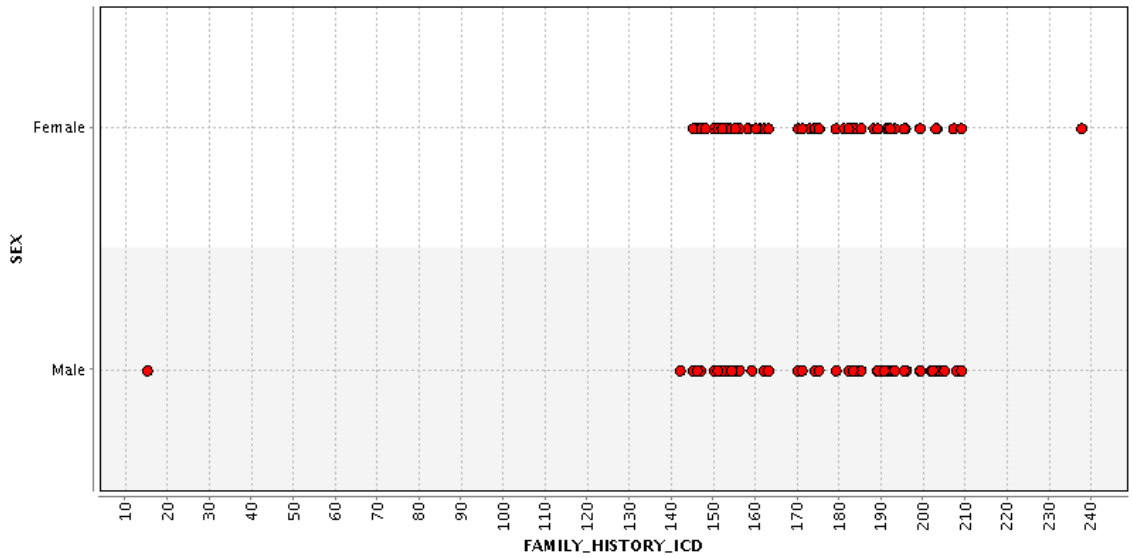
Relation	FIRST-TREATMENT-ICD						
	V58.0	V58.1	V58.49	99.24	V58.11	V66.7	99.28
Aunt	*	*	*	*	*	*	
Wife	*	*	*		*		
Sister	*	*	*	*	*	*	
Mother	*	*	*	*	*	*	
Uncle	*	*	*	*	*	*	
Father	*	*	*	*	*	*	
Brother	*	*	*	*	*	*	
Maternal Grand father	*			*			
Husband	*		*				
Grand Mother	*	*	*	*	*		
Niece	*	*	*				
Cousin	*	*	*	*	*		
Step sister		*					
Grand Father	*	*					
Son			*				
Maternal Grand Father	*	*	*				
Nephew		*		*	*		
Daughter			*				
Brother in Law							
Sister in Law							
Grand Son	*		*				
Mother in law							
Grand Daughter				*		*	
Paternal Grand Father							
Father-in-law							
Paternal Grand Mother							



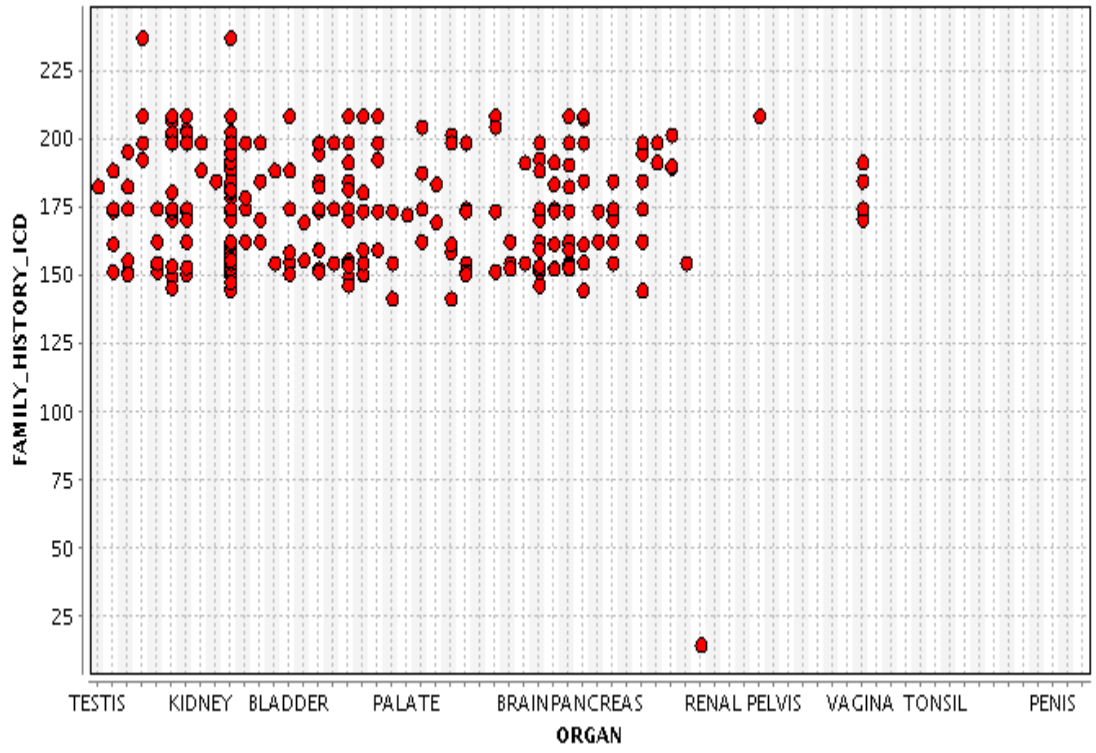
There is more span for males as they adopt different occupation whereas females work in limited fields that's why more variations for males than females.



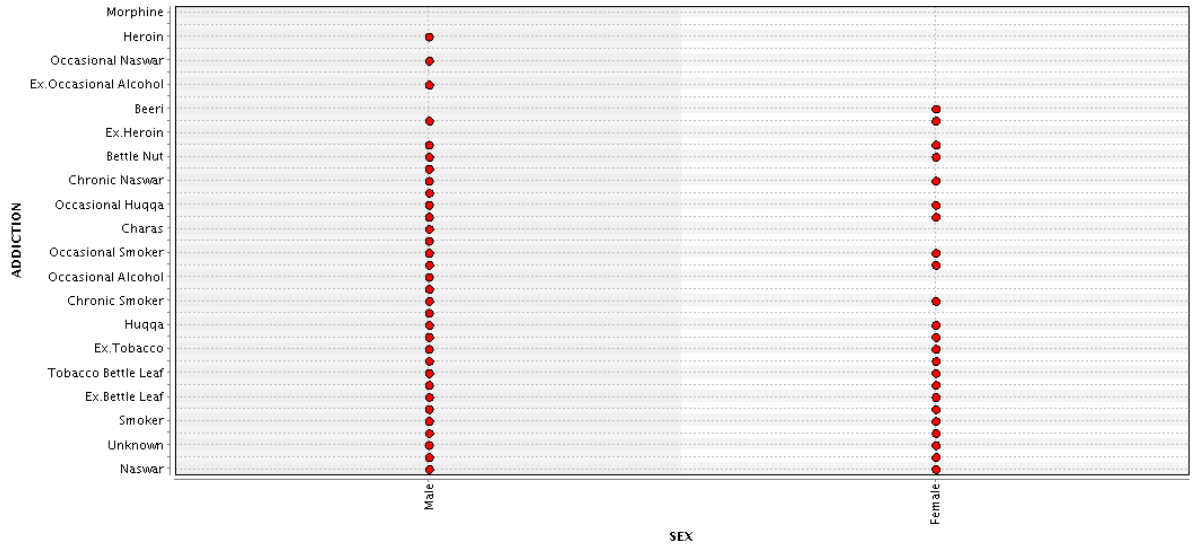
There is no cancer diagnosed for males if the relations are marital grand father, niece, step sister, daughter, sister and brother in law, mother in law, grand father and father-in-law. Similarly there is no cancer diagnose for females if relations are grand father, mother in law, sister in law, grand son, mother-in-law and father-in-law.



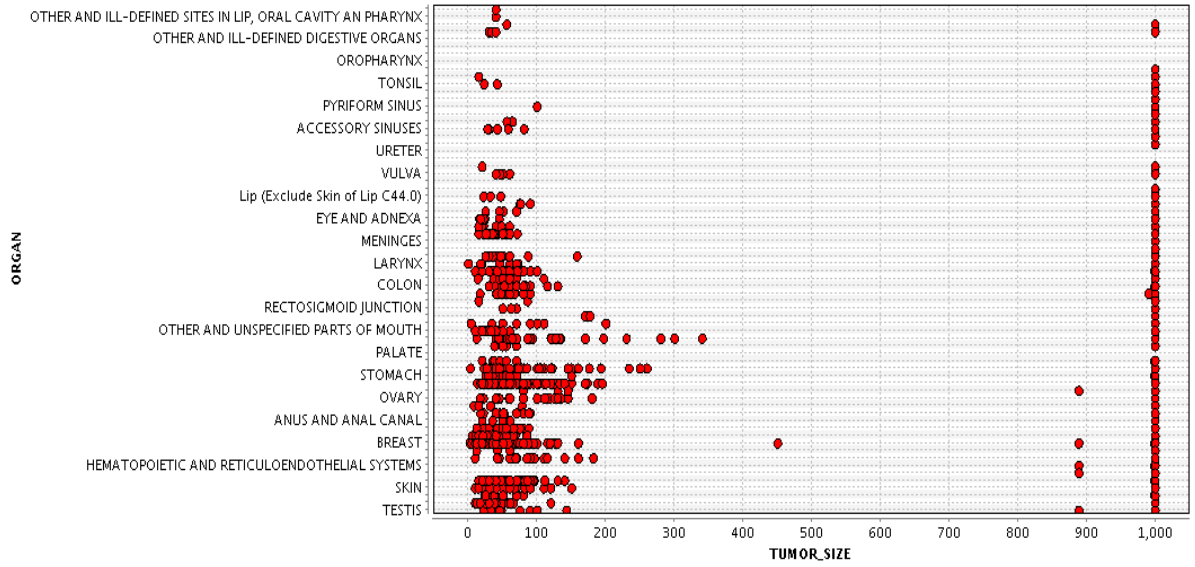
In males cancer is diagnosed due to family relation ICD mostly in the range of 140 – 210 and also at ICD 10.And for females its mostly in the range 145-208.9 and also at 237.5.



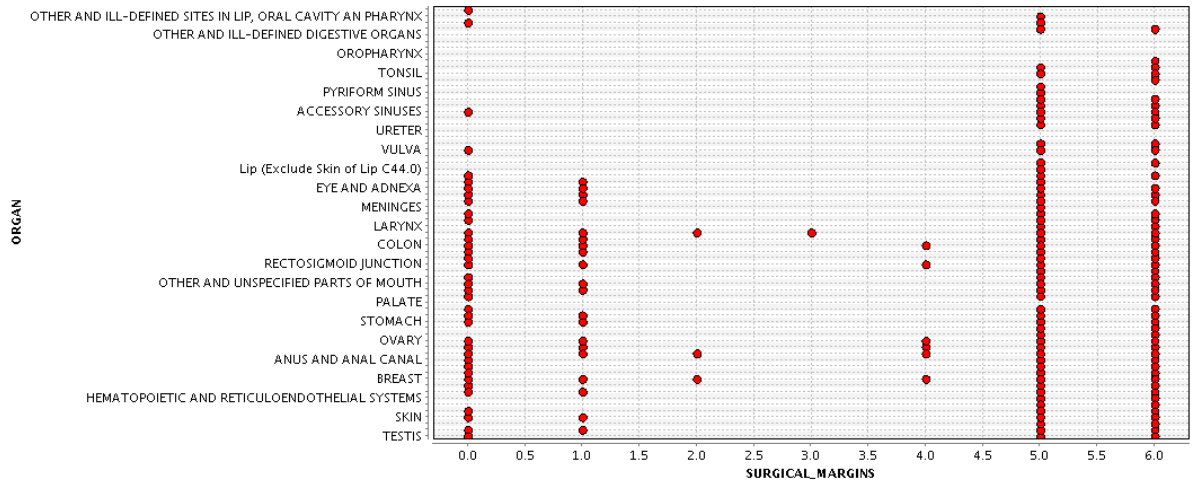
The above graph showing if Family History ICD is 15 then Effected Organ is Renal Pelvis, and from range 140-220 Testis,Kidney,Bladder,Palate,Brainpancreas and Vagina tonsil organ would effect.



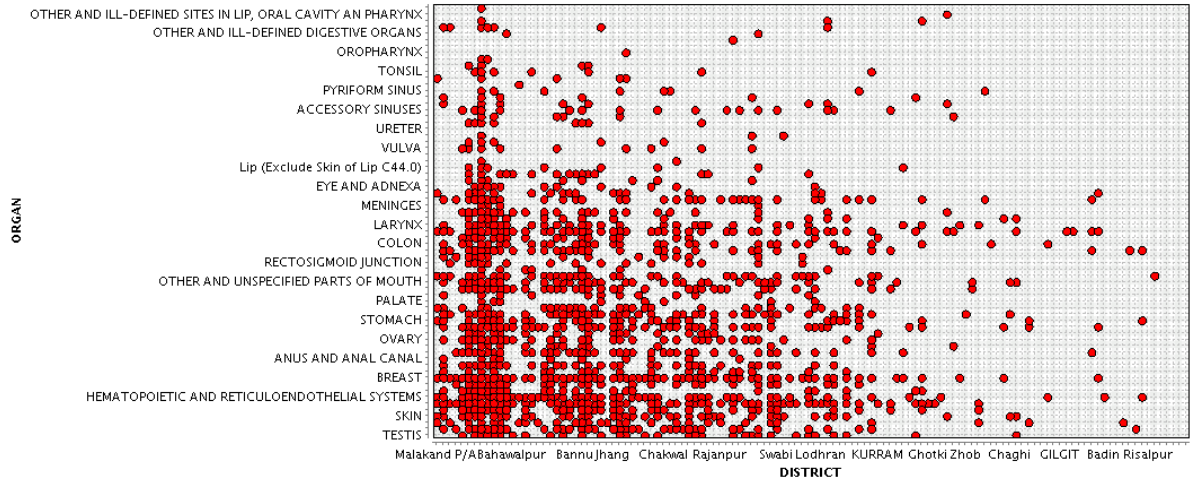
The above graph showing different addictions for both males and females. Males are more addicted while females are less addicted to addictive substances.



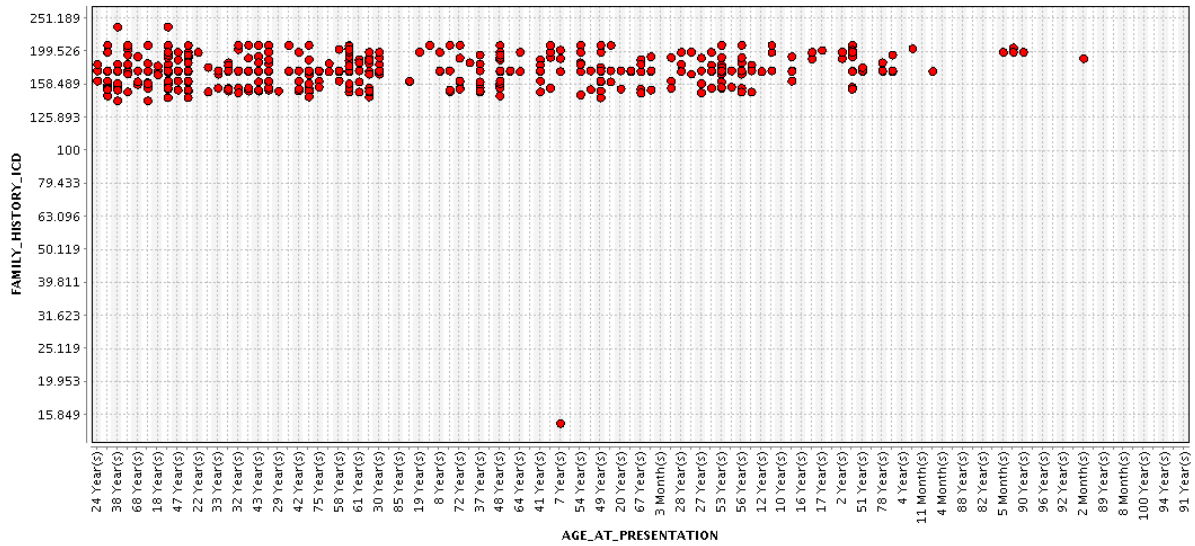
The above graph shoes that mostly all the tumors are of size 1000 when they are diagnosed but for testis,skin,breast,anal,ovary,stomach,colon,larynx and Vulva, tumor of other sizes are also diagnosed.



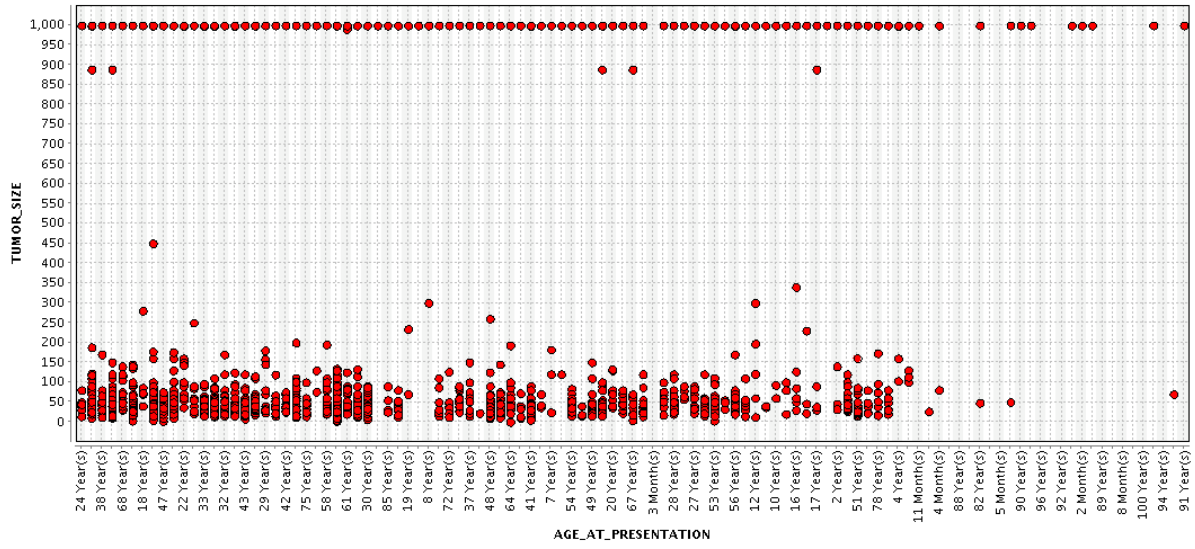
The above graph shows the sugical margin for each organ. Commonly for all organs except Orophynx and Uterer have Surgical margins of 5 and 6, which means that these cancers effect their surronded regions as well.



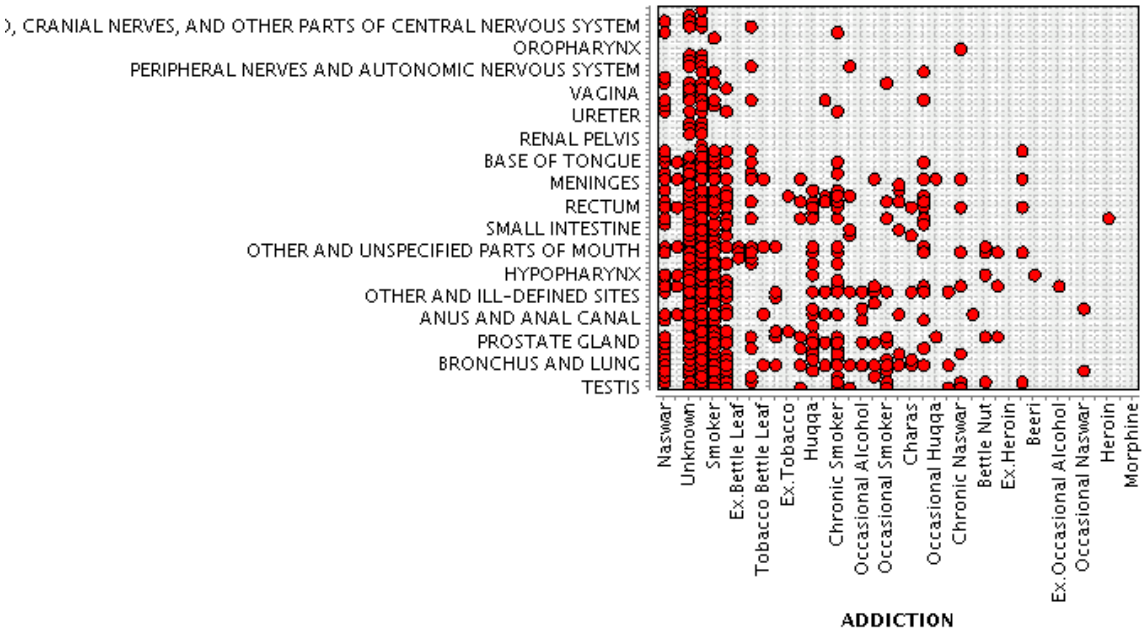
The above graph shows that almost all kind of cancers are diagnosed for the districts like Malakand and Bahawalpur. This might be due to the environment of these districts. While for Gilgit the only cancer detected are “Colon” and “Hematopoetic and Reticuloendothelial System”.



The above graph shows that if the Family_History_ICD is 15.849 then the only age in which cancer was diagnosed is 7 Years while almost in every age and Family_History_ICD of 158.409 to 199.526, the cancer is diagnosed.

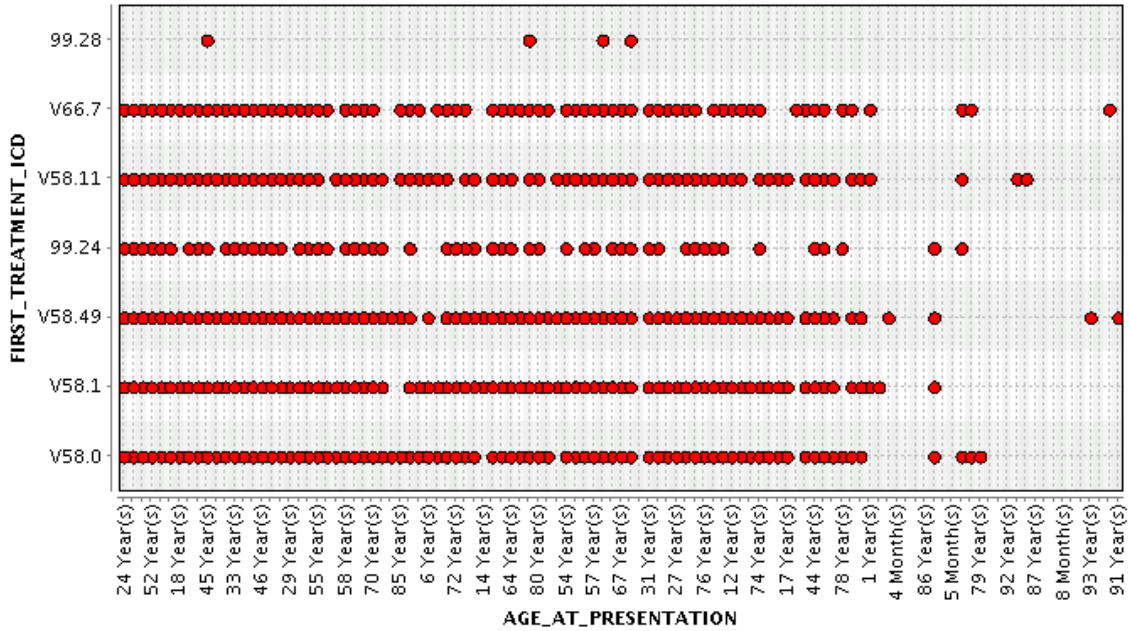


This graph shows the different tumour sizes for different ages. As we can see that 1000 Tumour size is diagnosed in almost all ages.



This graph shows the effects of addicted materials on different organs some important are showing below:-

Addictive Materials	**Organs**
Naswar	Testis,Skin,Prostrate grand, Anus and anul canal,unspecified parts of mouth,Rectum,Meninges, base of tounge, pyriform sinus and ill-defined sites in lip, oral cavity an pharynx
Smoking	Testis, Prostrate grand, Anus and anul, Meninges, base of tounge,hypopharynx,small intestine.rectum, peripheral nerves , autonomic nervous system and other parts of nervous system
Tobacco battle leaf	Bronchus and lung, Anus and anal canal, unspecified parts of mouth and Meninges
Huqqa	Bronchus and lung, unspecified parts of mouth, Anus and anal canal, hypopharynx and Rectum
Beer	Hypopharynx
Herion	Rectum



This graph shows the ICD(International code for disease) allocation among different ages.

Like for;

- 24,52,18,33,46,29,55,58,70 years old persons ICDs are v58.0,v 58.1,v58.49,v58.11,v66.7
- 85 year old patients don't have v58.0,v58.49,v58.11,v66.7
- 45,57,67 years old patients normally have ICD 99.28 with others too.
- 91,92 years old have V58.11 ICD.
- 5-8 Months patients have V58.0