CERVICAL CANCER RISK FACTORS MILESTONE REPORT

1. Problem Statement:

Cervical cancer is the third most common cancer in women worldwide, affecting over 500,000 women and resulting in approximately 275,000 deaths every year.

Cervical cancer can be prevented through early administration of the HPV vaccine and regular pap smear screenings, which indicate the presence of precancerous cells. It is also sometimes curable by the removal of the early-stage cancerous tissue that is identified through pap smears. Screening and early treatment can lead to potential cures in about 95% of women at risk for cervical cancer.

Numerous studies of the epidemiology of cervical cancer have shown strong associations with religious, marital and sexual patterns. Although it is well established that women with multiple partners and early ages at first intercourse are at high risk, less is known about how these factors interact or how risk is affected by specific sexual characteristics. Recent studies indicate that number of steady partners and frequent intercourse at early ages may further enhance risk, supporting hypotheses regarding a vulnerable period of the cervix and a need for repeated exposure to an infectious agent. It is now widely accepted that HPV is the major infectious etiological agent, but whether other infectious agents play supportive or interactive roles is unclear. Other speculative risk factors for cervical cancer include cigarette smoking, oral contraceptive usage and certain nutritional deficiencies, but again it is not clear whether these factors operate independently from HPV.

2. Who Might Care:

Medical staff, Females, Female patients and Scholars

3. Description of the Data Set:

The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values). Data set has 36 features and 858 data points. Since target variable (‘Dx:Cancer’) consists of 18 positive samples and 840 negatives, the data set is extremely **imbalanced.**

Features (Integers):

Age, Number of sexual partners, First sexual intercourse (age), Num of pregnancies, Hormonal Contraceptives (years), IUD (years), STDs (number), STDs: Number of diagnosis, STDs: Time since first diagnosis, STDs: Time since last diagnosis

Features (Booleans):

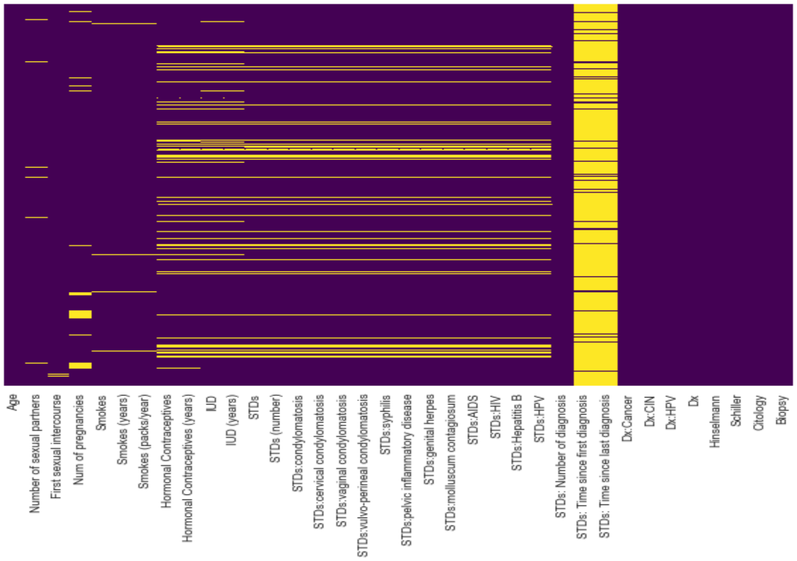
Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives, IUD, STDs, STDs:condylomatosis, STDs:cervical condylomatosis, STDs:vaginal condylomatosis, STDs:vulvo-perineal condylomatosis, STDs:syphilis, STDs:pelvic inflammatory disease, STDs:genital herpes, STDs:molluscum contagiosum, STDs:AIDS, STDs:HIV, STDs:Hepatitis B, STDs:HPV, Dx:Cancer, Dx:CIN, Dx:HPV, Dx, Hinselmann, Schiller, Cytology, Biopsy

26 out of 36 features have missing values in the data set. Missing values of each feature and the respective percentages are written below:

	Missing Values	% of Total Values
STDs: Time since last diagnosis	787	91.7
STDs: Time since first diagnosis	787	91.7
IUD	117	13.6
IUD (years)	117	13.6
Hormonal Contraceptives	108	12.6
Hormonal Contraceptives (years)	108	12.6
STDs:vulvo-perineal condylomatosis	105	12.2
STDs:HPV	105	12.2
STDs:Hepatitis B	105	12.2
STDs:HIV	105	12.2
STDs:AIDS	105	12.2
STDs:molluscum contagiosum	105	12.2
STDs:genital herpes	105	12.2
STDs:pelvic inflammatory disease	105	12.2
STDs:syphilis	105	12.2
STDs:cervical condylomatosis	105	12.2
STDs:vaginal condylomatosis	105	12.2
STDs:condylomatosis	105	12.2
STDs (number)	105	12.2
STDs	105	12.2
Num of pregnancies	56	6.5
Number of sexual partners	26	3.0
Smokes (packs/year)	13	1.5
Smokes (years)	13	1.5
Smokes	13	1.5
First sexual intercourse	7	0.8

4. Data Wrangling:

Since ‘STDs: Time since last diagnosis’ and ‘STDs: Time since first diagnosis’ features have more than %91 percent missing values we dropped these two features.



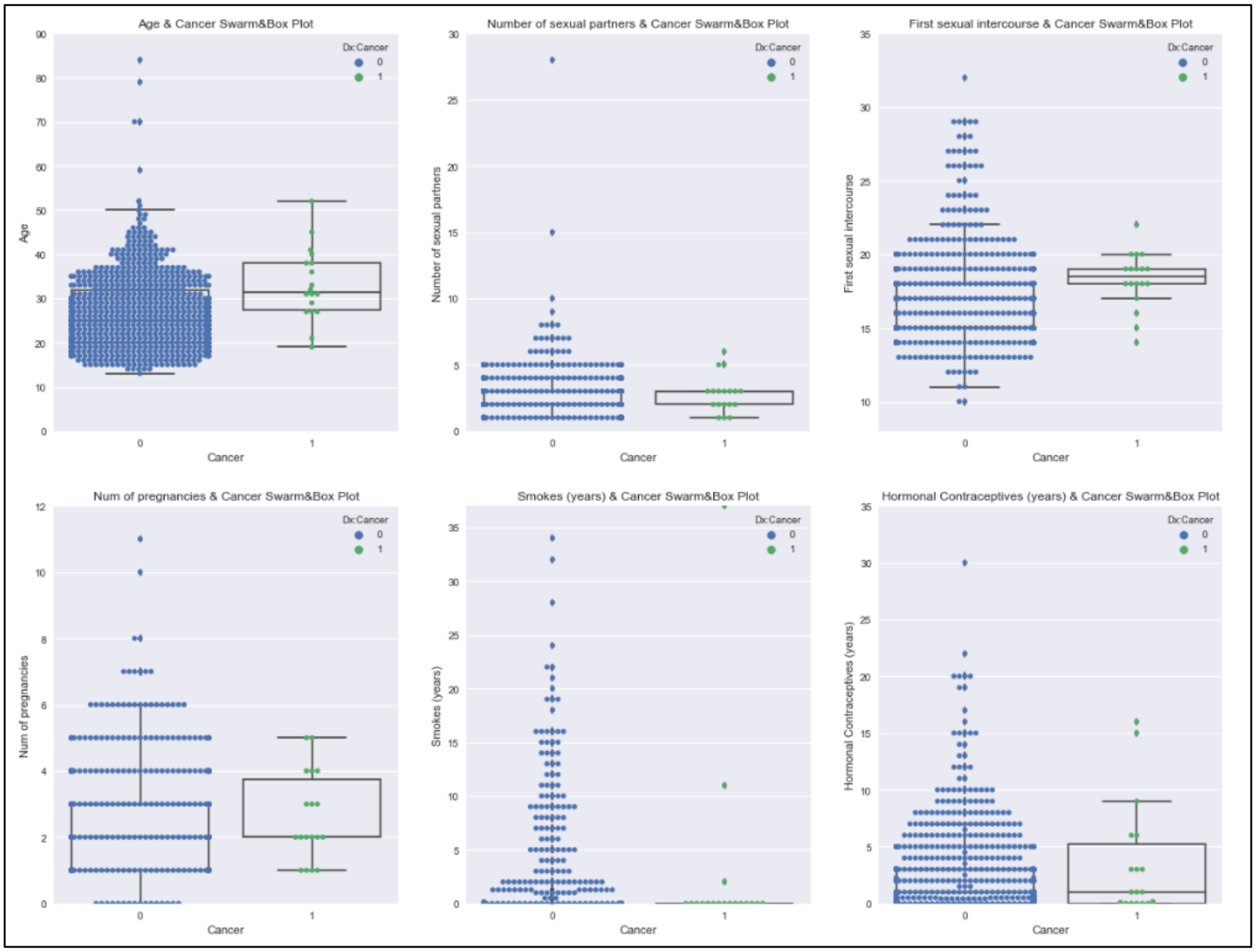
For the rest numeric features which had missing values were applied mean statistical method. But categorical features which had missing values were applied pd.get_dummies() function to create dummy variables for all categorical values including the missing value (‘NaN’).

Before doing that, we converted the string type of values to categorical ones and then applied the function. After concatenating the new data set consisted of dummy features to the main data set, we dropped the features from which we produced the dummy ones from the main data set.

The wrangled data set had 64 features and 848 data points and all the features consisted of numeric values.

At the end the cleaned data set was saved as ‘Cervical_Cancer_Risk_Cleaned.csv’ and uploaded to the Github.

5. Data Story Telling:



a. Cancer diagnosed patient's age are cumulated between 27 to 42. One patient younger than 20 got cancer. There is no outlier from cancer patient's age. Cancer patient's median age is higher than non-cancers.

b. Cancer diagnosed patient's number of sexual partners are cumulated between 1 to 5. Most of the patients have had either 5 or less partners. It seems the correlation between number of sexual partner and cancer is not strong.

c. Cancer diagnosed patient's first sexual intercourses are cumulated between 17 to 20. There is outlier even at 10. It seems the correlation between early age sexual intercourse and cancer is not strong. Cancer patient's median first sexual intercourse age is higher than non-cancer ones.

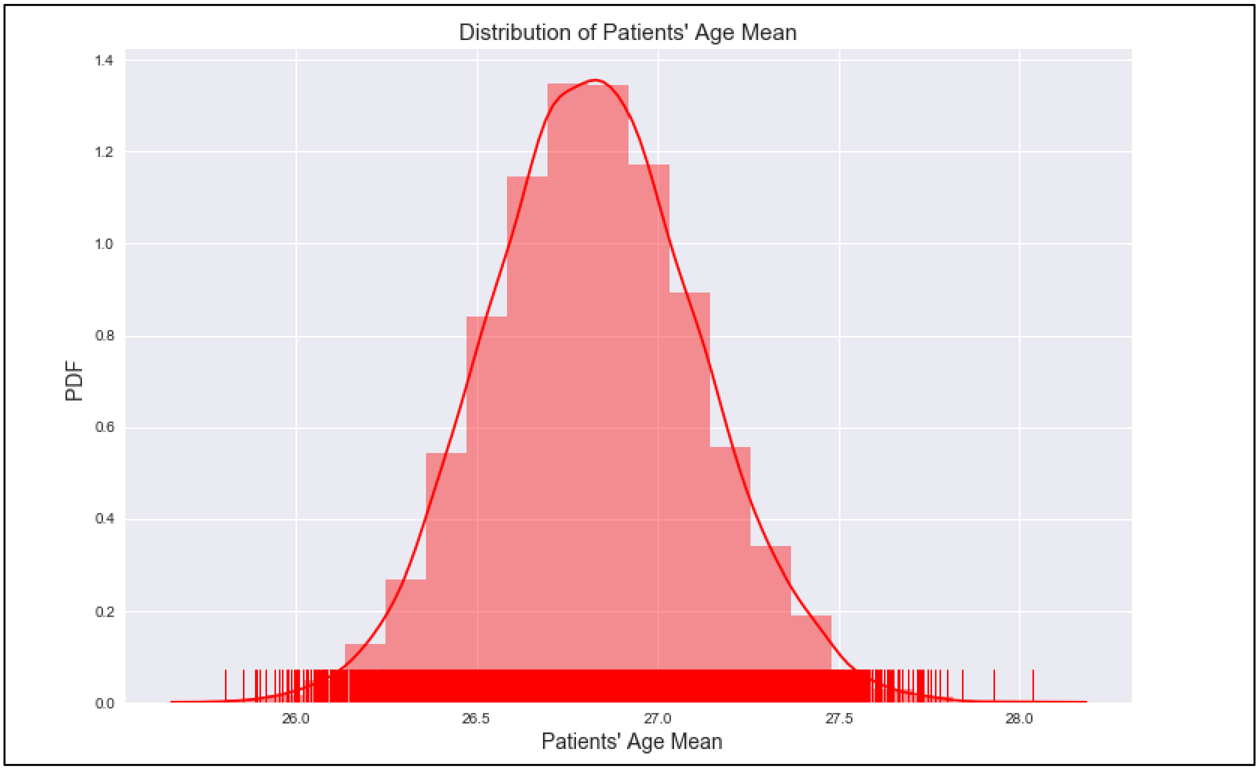
d. Based on the 4th plot, it seems that there is no strong correlation between number of pregnancies and cancer. Cancer patient's median number of pregnancies is higher than non-cancer.

e. It seems that there is no strong correlation between smokes(year) and cancer. Most of the patients are not cancer smoke more than 3 years as well.

f. It seems that there is no strong correlation between hormonal contraceptives and cancer. Most of the non-cancer patients use hormonal contraceptives.

5. Initial Findings from Exploratory Analysis:

In the data set the age of the patients are normally distributed. We applied 5 hypothesis, one of them is written below with the results.



First Hypothesis Application:

Null Hypothesis: The true mean of the ages is 27.26.

Result: Since p-value is 0 or less than 0.05, we reject the claim that the mean of patient's age is 27.26 in favor of the alternative hypothesis that the mean of patient's age differs from 27.26. Based on the 95% confidence interval, ages between 26.6 and 27.9 are considered normal.

6. Next Steps:

Brief Description:

I will build a base line Logistic Regression modelling to predict the Cancer patients accurately based on the results of the model, I will apply different model(s) to increase the prediction accuracy.