Exploratory data analysis of the Cervical cancer (Risk factors) Data set

In [1]:

```python
import numpy as np
import pandas as pd
import altair as alt
from sklearn.model_selection import train_test_split, StratifiedKFold

alt.data_transformers.enable('data_server')
alt.renderers.enable('mimetype')
```

Out[1]:

```
RendererRegistry.enable('mimetype')
```

Summary of the data set

The data set was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The data set 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). This data set was sourced from the UCI Machine Learning Repository and can be found here.

The data set was used in Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017, available here.

The data set has 4 different target variables each having a value of 0(tested negative for that specific medical test) or 1(tested positive for that specific medical test). For the purpose of this project, these binary class variables will be combined into a single binary target variable which will be 1(True) if any medical test is positive and 0(False) if no test was positive.

In [2]:

```python
# load dataset into pandas dataframe
cervical_raw = pd.read_csv('../data/risk_factors_cervical_cancer.csv')

# create target variable 'risk'
risk = []
for row in range(len(cervical_raw)):
    risk.append(
        cervical_raw.loc[cervical_raw.index[row], 'Hinselmann'] or
        cervical_raw.loc[cervical_raw.index[row], 'Schiller'] or
        cervical_raw.loc[cervical_raw.index[row], 'Citology'] or
        cervical_raw.loc[cervical_raw.index[row], 'Biopsy']
    )
cervical_modified = cervical_raw.copy()
cervical_modified['risk'] = risk

# drop the previous target variables
cervical_modified = cervical_modified.drop(columns=['Hinselmann', 'Schiller', 'Citology'

# create dataframe with counts of each class
class_counts = pd.DataFrame(cervical_modified['risk'].value_counts()).rename(index={0:'N
                                                                                   1:'R
                                                                          columns={'r
```

```
# set caption for Table 1
class_counts.style.set_caption('Table 1. Counts of observation for each class')
```

Out[2]:

Table 1. Counts of observation for each class

	Target
No risk of cervical cancer	756
Risk of cervical cancer	102

Split data set into training and test splits

before splitting the dataset, we replace all occurences of '?' in the data with `np.nan` so that it is easier to work with the missing values. We also change the data types of columns to match the data stored in them.

In [3]:

```
# replace the ? values with NaN
cervical_clean = cervical_modified.replace('?', np.nan)

# convert columns to relevant data types
for col_name in cervical_clean.columns:
    if cervical_clean[col_name].dtype == 'object':
        cervical_clean[col_name] = cervical_clean[col_name].astype(float)
```

We now split our data so that 80% of the examples are in the training set while 20% are in the test set.

In [4]:

```
# split data into training and test sets
train_df, test_df = train_test_split(cervical_clean, test_size=0.2, random_state=123)
```

In [5]:

```
# create dataframe with counts of each class and for both train and test set
train_class_counts = pd.DataFrame(train_df['risk'].value_counts())
test_class_counts = pd.DataFrame(test_df['risk'].value_counts())

train_test_class_counts = pd.concat([train_class_counts, test_class_counts], axis=1).ren
    index={0:'No risk of cervical cancer',
           1:'Risk of cervical cancer'}
)
train_test_class_counts.columns = ['Train', 'Test']

# set caption for Table 2
train_test_class_counts.style.set_caption('Table 2. Counts of observations for each clas
```

Out[5]:

Table 2. Counts of observations for each class and partition

	Train	Test
No risk of cervical cancer	608	148
Risk of cervical cancer	78	24

There is quite a bit of class imbalance in this dataset. We won't try and use under-sampling or over-sampling to remedy this since our data set is quite small. We will deal with this after the inital model building and tuning phase in the case that the model is performing poorly. We can evaluate whether class imbalance is a major issue based on the confusion matrix (if the False Negative rate is high).

Exploratory analysis on the training set

We plotted the distributions of each explanatory variable in the training data set to see whether or not it will be useful for predicting the target variable.

Most of the numeric features are extremely skewed. This can have a negative impact on the model as machine learning models generally perform better on normalized data. As such, we might experiment with some transformations (eg: log transformation) to try and normalize the data. A bunch of our feature variables have a either all or atleast a significant amount of missing values. These features will likely be omitted from the final model. Taking a look at correlations between certain columns, we can see that some features are almost colinear. This means they can be safely removed as they do not add to model performance. This should reduce complexity in the model as well.

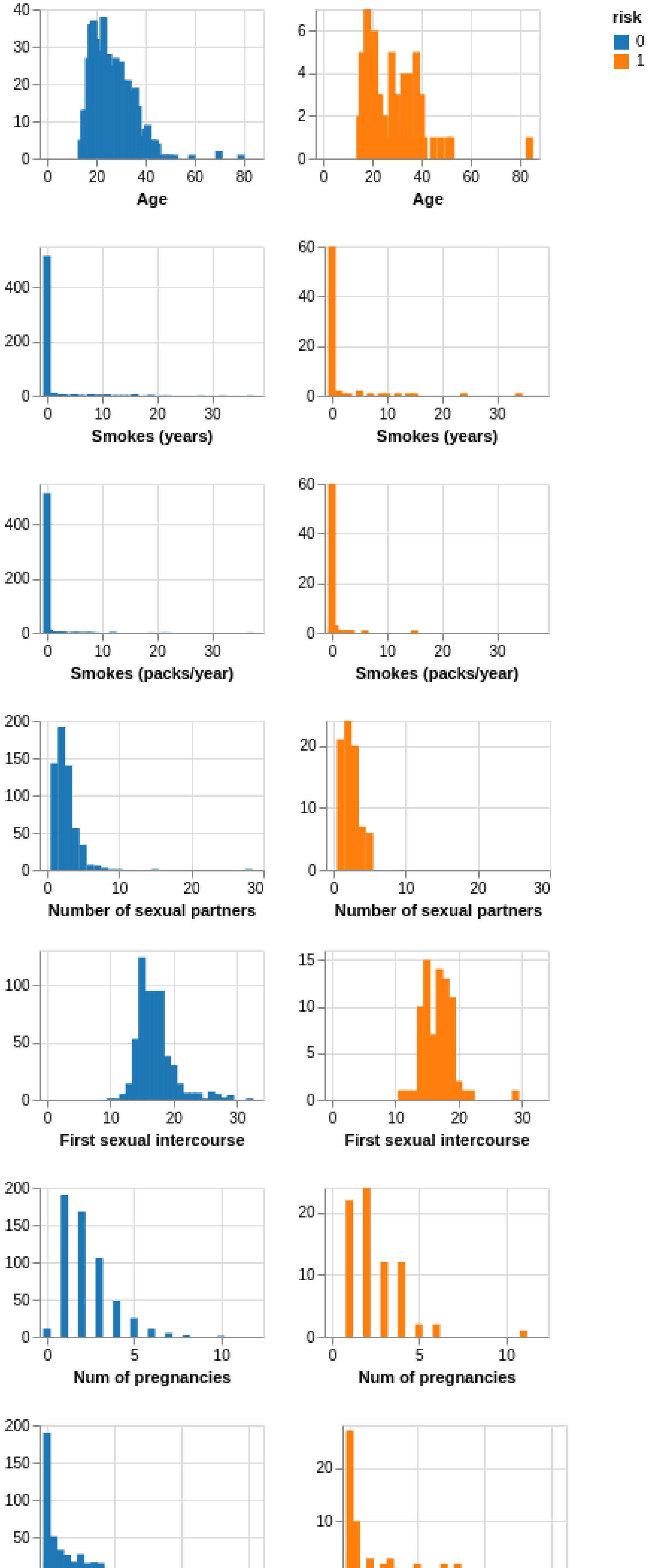
In [6]:

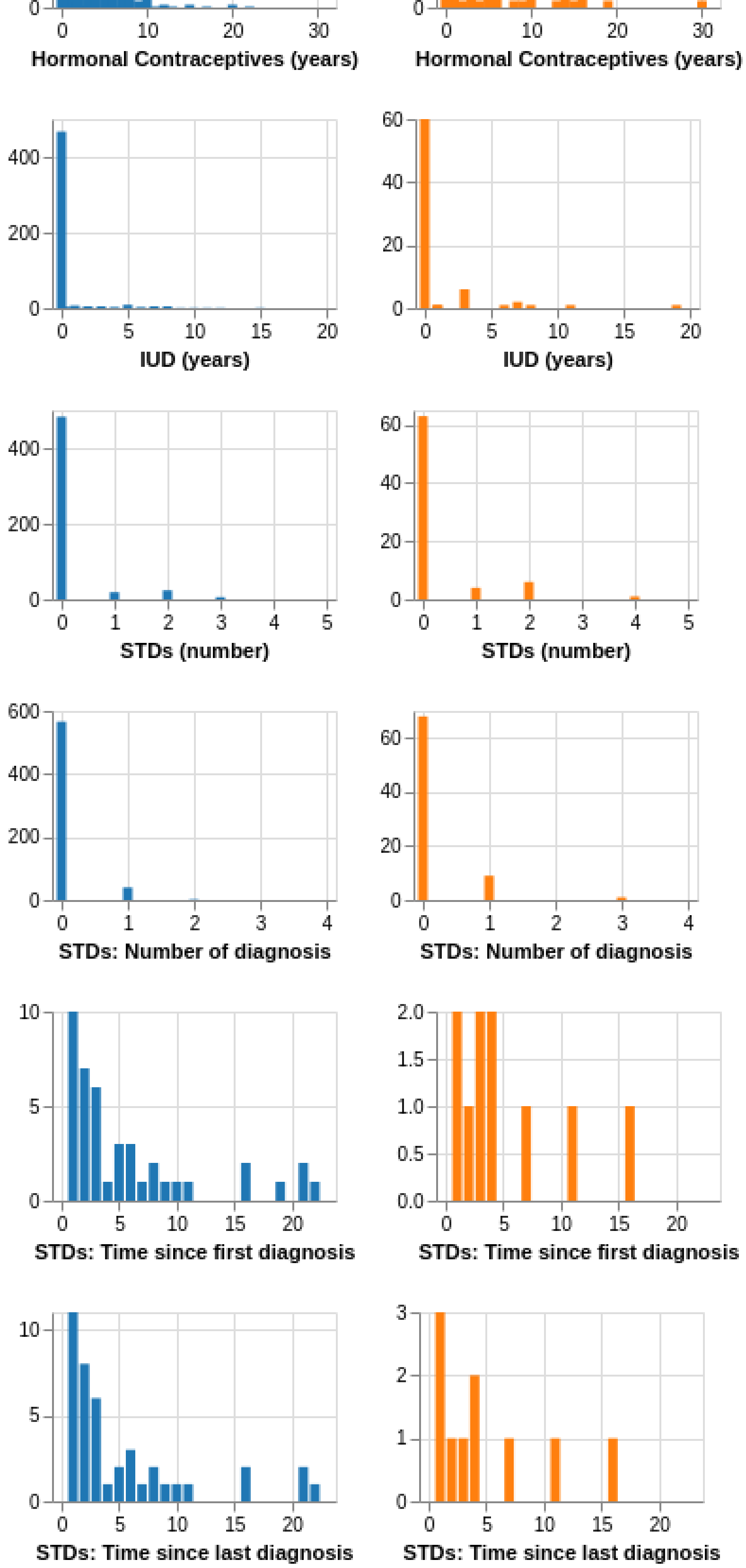
```
def hist( feat = None, feat_list = None, repeat = False):
    if repeat == False:
        chart = alt.Chart( train_df).mark_bar().encode(
            alt.X( 'Age', type='quantitative'),
            alt.Y( 'count()', stack=False, title=''),
            alt.Color( 'risk', type='ordinal', scale=alt.Scale(scheme='category10'))
        ).properties(
            height=100,
            width=150
        ).facet( 'risk', columns = 1)
        return chart
    if repeat == True:
        chart_list_0 = []
        chart_list_1 = []
        chart_list_concat = []
        for feat in feat_list:
            chart_tmp_0 = alt.Chart( train_df.query('risk==0')).mark_bar().encode(
                alt.X( feat, type='quantitative', scale = alt.Scale( domain = ( 0, train
                alt.Y( 'count()', stack=False, title=''),
                alt.Color( 'risk', type='ordinal', scale=alt.Scale(scheme='category10'))
            ).properties(
                height=100,
                width=150
            )
            chart_tmp_1 = alt.Chart( train_df.query('risk==1')).mark_bar().encode(
                alt.X( feat, type='quantitative', scale = alt.Scale( domain = ( 0, train
                alt.Y( 'count()', stack=False, title=''),
                alt.Color( 'risk', type='ordinal', scale=alt.Scale(scheme='category10'))
            ).properties(
                height=100,
                width=150
            )
            chart_list_0.append( chart_tmp_0)
            chart_list_1.append( chart_tmp_1)
            chart_concat = chart_tmp_0 | chart_tmp_1
            chart_list_concat.append( chart_concat)
        return alt.vconcat( *chart_list_concat)
```

In [7]:

```
# create list of binary features
binary_features = ['Smokes', 'Hormonal Contraceptives', 'IUD', 'STDs', 'STDs:condylomato
                   'STDs:cervical condylomatosis', 'STDs:vaginal condylomatosis',
                   'STDs:vulvo-perineal condylomatosis', 'STDs:syphilis', 'STDs:pelvic i
                   'STDs:genital herpes', 'STDs:molluscum contagiosum', 'STDs:AIDS', 'ST
                   'STDs:Hepatitis B', 'STDs:HPV', 'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx'

# create list of numeric features
numeric_features = ['Age', 'Smokes (years)', 'Smokes (packs/year)', 'Number of sexual pa
                    'Num of pregnancies', 'Hormonal Contraceptives (years)', 'IUD (years
                    'STDs (number)', 'STDs: Number of diagnosis', 'STDs: Time since firs
                    'STDs: Time since last diagnosis']

# create charts for binary features
binary_charts = alt.Chart(train_df).mark_bar().encode(
    alt.X(alt.repeat(), type='ordinal'),
    alt.Y('count()'),
    alt.Color('risk', type='ordinal', scale=alt.Scale(scheme='category10'))
).properties(
    height=150,
    width=75
).repeat(
    binary_features,
    columns=4
)

hist(feat_list=numeric_features, repeat=True)
```

Out[7]:

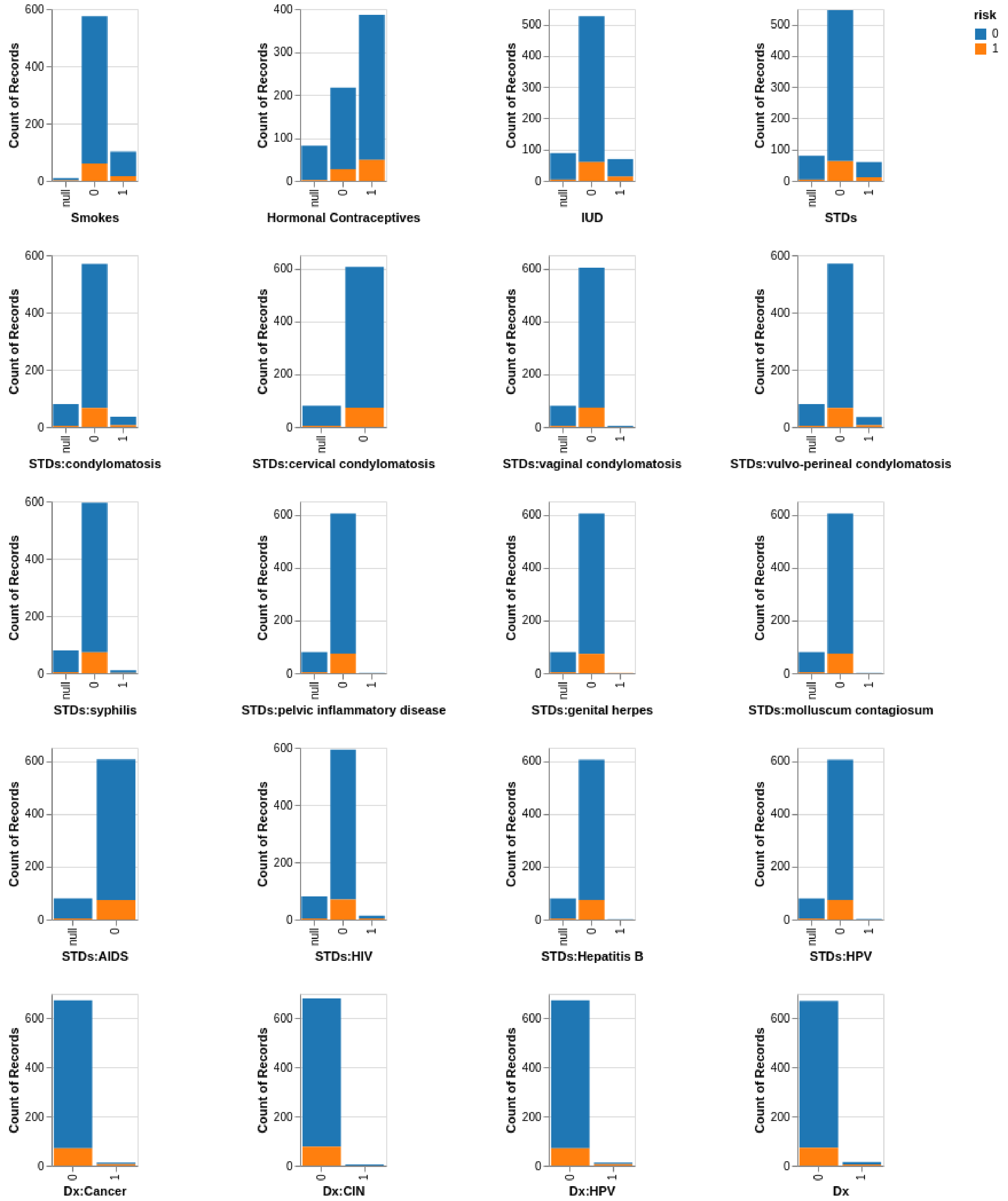

risk
0
1
Age
0
20
40
60
80
0
10
20
30
40
0
2
4
6
Smokes (years)
0
10
20
30
0
200
400
0
20
40
60
Smokes (packs/year)
Number of sexual partners
0
50
100
150
200
First sexual intercourse
15
Num of pregnancies
5
10



In [8]:

```
binary_charts
```

Out[8]:



In [9]:
```
train_df.loc[:,['Smokes', 'Smokes (years)', 'Smokes (packs/year)']].corr('spearman').sty
```

Out[9]:

	Smokes	Smokes (years)	Smokes (packs/year)
Smokes	1.000000	0.995594	0.995591
Smokes (years)	0.995594	1.000000	0.997172
Smokes (packs/year)	0.995591	0.997172	1.000000

In [10]:
```
stds = ['STDs:condylomatosis', 'STDs:vaginal condylomatosis',
        'STDs:vulvo-perineal condylomatosis', 'STDs:syphilis', 'STDs:pelvic inflammatory
        'STDs:genital herpes', 'STDs:molluscum contagiosum', 'STDs:HIV',
        'STDs:Hepatitis B', 'STDs:HPV']
train_df.loc[:, stds].corr('spearman').style.background_gradient()
```

Out[10]:

	STDs:condylomatosis	STDs:vaginal condylomatosis	STDs:vulvo-perineal condylomatosis	STDs:syphilis	STDs:pelvic inflammatory disease	STD
STDs:condylomatosis	1.000000	0.324353	0.985150	0.070412	-0.010217	-
STDs:vaginal condylomatosis	0.324353	1.000000	0.241887	-0.011083	-0.003314	-
STDs:vulvo-perineal condylomatosis	0.985150	0.241887	1.000000	0.072310	-0.010066	-
STDs:syphilis	0.070412	-0.011083	0.072310	1.000000	-0.005528	-
STDs:pelvic inflammatory disease	-0.010217	-0.003314	-0.010066	-0.005528	1.000000	-
STDs:genital herpes	-0.010217	-0.003314	-0.010066	-0.005528	-0.001653	
STDs:molluscum contagiosum	-0.010217	-0.003314	-0.010066	-0.005528	-0.001653	-
STDs:HIV	0.107336	-0.012069	0.109811	0.065182	-0.006020	-
STDs:Hepatitis B	-0.010217	-0.003314	-0.010066	-0.005528	-0.001653	-
STDs:HPV	-0.014461	-0.004691	-0.014247	-0.007824	-0.002339	-

References

Dua, Dheeru, and Casey Graff. 2017. “UCI Machine Learning Repository.” University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

Fernandes, K., Cardoso, J.S., & Fernandes, J.C. (2017). Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. Iberian Conference on Pattern Recognition and Image Analysis. https://www.semanticscholar.org/paper/Transfer-Learning-with-Partial-Observability-to-Fernandes-Cardoso/1c02438ba4dfa775399ba414508e9cd335b69012

Cervical cancer (Risk Factors) Data Set
https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29