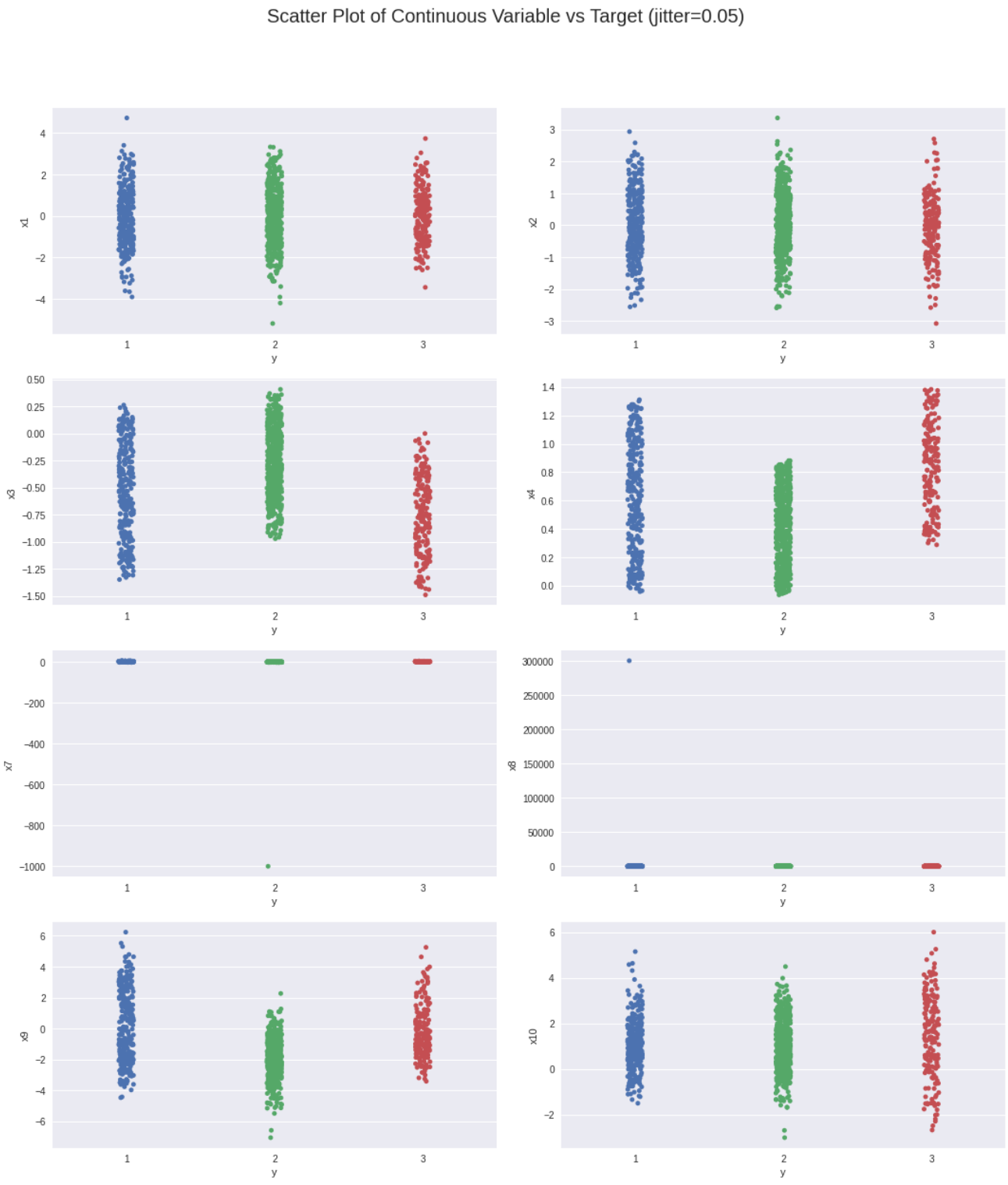
In [7]:

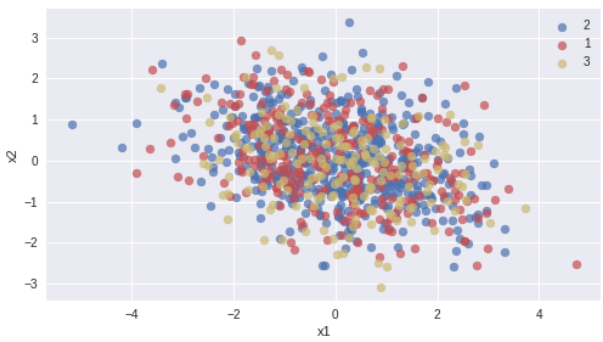
```
import multiprocessing
import os
import random
from datetime import datetime

import pandas as pd
import scipy
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# global parameters :)
train_path = 'train.csv'
test_path = 'eval.csv'
name = f"{datetime.now().strftime('%Y-%m-%d--%H:%M:%S')}"
xes = ['x1', 'x2', 'x3', 'x4', 'x5', 'x6', 'x7', 'x8', 'x9', 'x10']
grade_mapping = {'A': 7, 'B': 6, 'C': 5, 'D': 4, 'E': 3, 'F': 2, 'Fx': 1}
scale = False
scale_indices = xes[:4] + xes[6:]
y_mapping = {"Atsuto": 1, "Bob": 2, "Jörg": 3}
y_mapping_reversed = {v: k for k, v in y_mapping.items()}
true_false_mapping = {True: 1, False: 0, "True": 1, "False": 0}


def preprocess_train(df):
    df.dropna(inplace=True)
    df = df[~df[xes].isin(["?"]).any(axis=1)]
    df.loc[:, 'y'] = df.loc[:, 'y'].map(y_mapping)
    df.loc[:, ['x1', 'x2']] = df.loc[:, ['x1', 'x2']].astype(float)
    return df


def preprocess(df, scaler):
    df.loc[:, 'x6'] = df.loc[:, 'x6'].map(grade_mapping)
    df.loc[:, 'x5'] = df.loc[:, 'x5'].map(true_false_mapping)

    if scaler:
        scaler.transform(df[xes[:4] + xes[6:]])

    return df

if __name__ == '__main__':
    random.seed(72)
    pd.np.random.seed(72)
    scipy.random.seed(72)
    print(os.listdir())
    pd.set_option('display.max_rows', 500)
    pd.set_option('display.max_columns', 500)
    pd.set_option('display.width', 1000)

    train, test = pd.read_csv(train_path, comment="#"), pd.read_csv(test_pat
h)

    # Preprocessing
    train = preprocess_train(train)
    scaler = StandardScaler().fit(train[scale_indices]) if scale else None
    train, test = preprocess(train, scaler), preprocess(test, scaler)

    X, y = train.loc[:, xes], train.loc[:, "y"]
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)

#     import matplotlib.pyplot as plt
#     import seaborn as sns
#
#     sns.set_style("whitegrid");
```

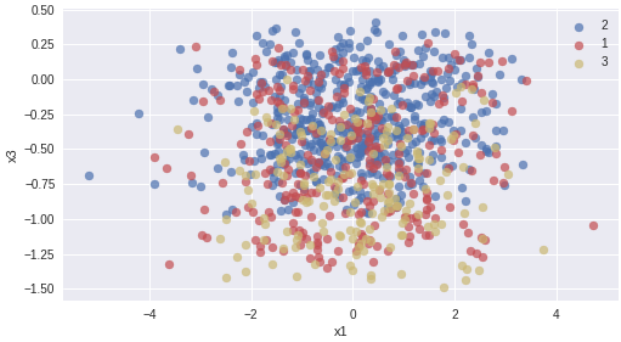
```
#     sns.pairplot(train, hue="y", size=3);
#     plt.show()
    from autoviz.AutoViz_Class import AutoViz_Class
    AV = AutoViz_Class()
    dft = AV.AutoViz(
        "train_clean.csv",
        ",",
        "y",
        train,
        header=0,
        verbose=0,
        lowess=False,
        chart_format="svg",
        max_rows_analyzed=1500000,
        max_cols_analyzed=300,
    )
```

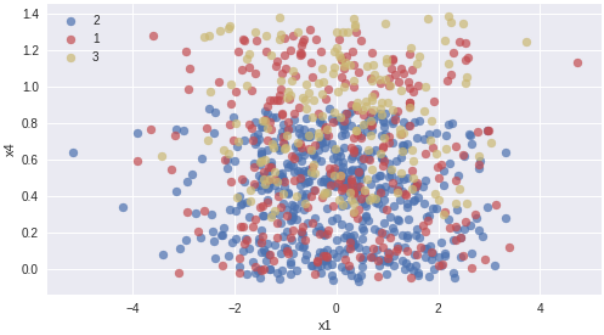
```
['.git', '.ipynb_checkpoints', '.~lock.eval.csv#', '.~lock.train.csv#', 'demo.py', 'eval.csv', 'kth-ml-ch.txt', 'train.csv', 'train_clean.csv', 'Untitled.ipynb']
Shape of your Data Set: (996, 12)
Classifying variables in data set...
    11 Predictors classified...
        This does not include the Target column(s)
    1 variables removed since they were ID or low-information variables
Total Number of Scatter Plots = 36
Could not draw Pivot Charts against Dependent Variable
All plots done
Time to run AutoViz (in seconds) = 5.221
```

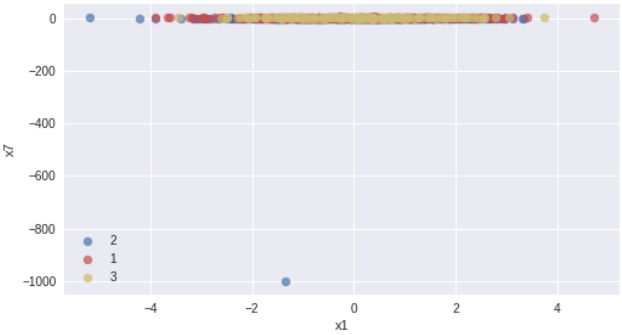


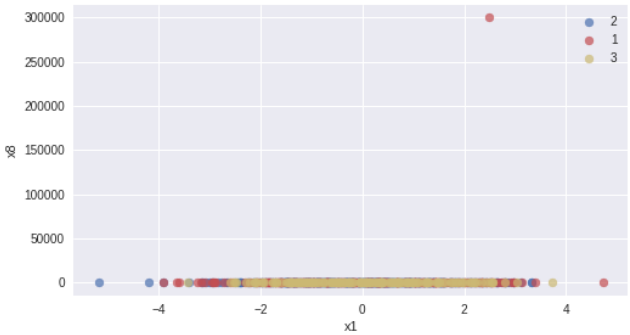
Scatter Plot of each Continuous Variable against Target Variable

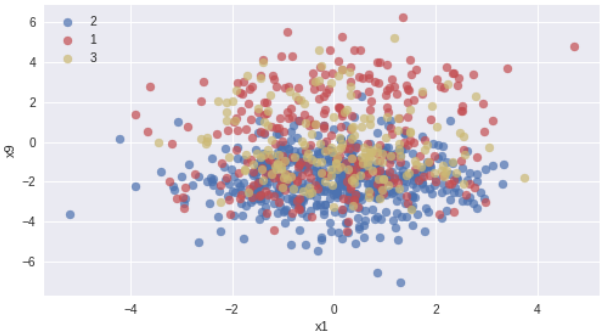


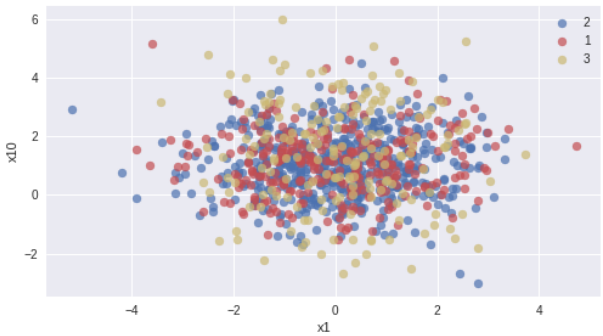



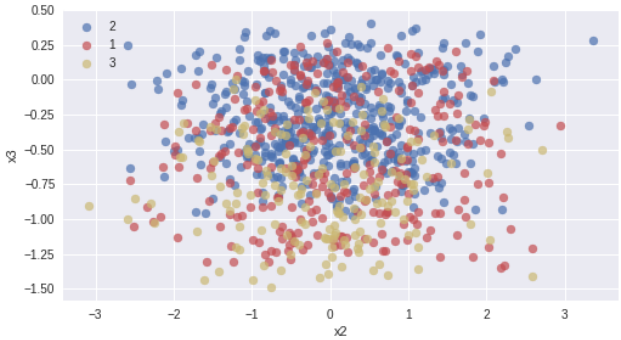



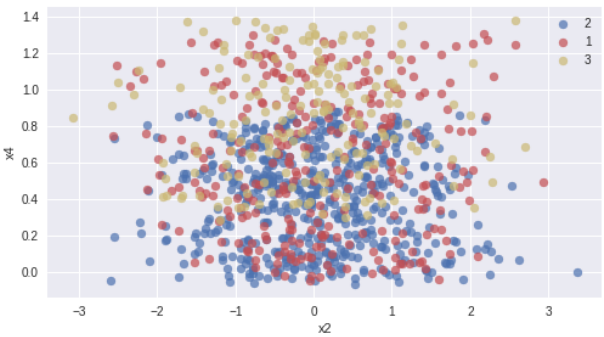



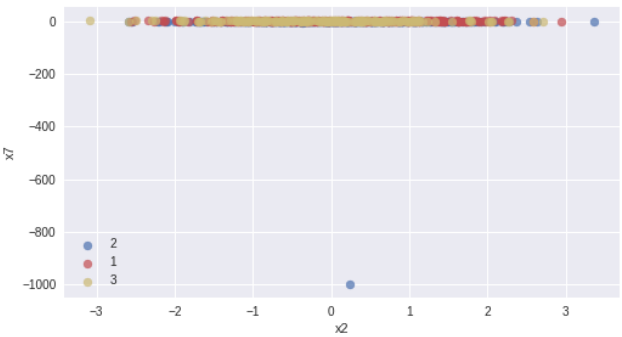



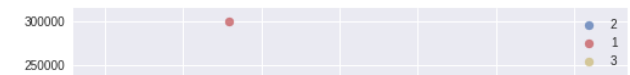



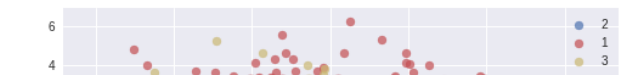



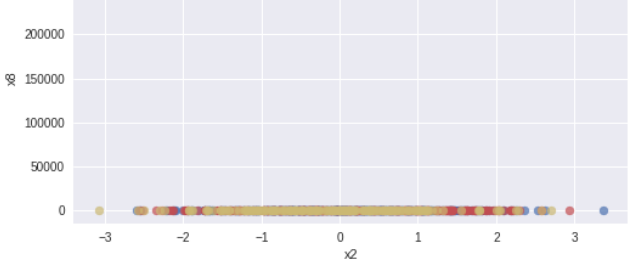



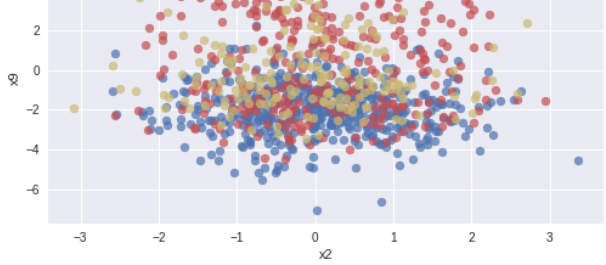



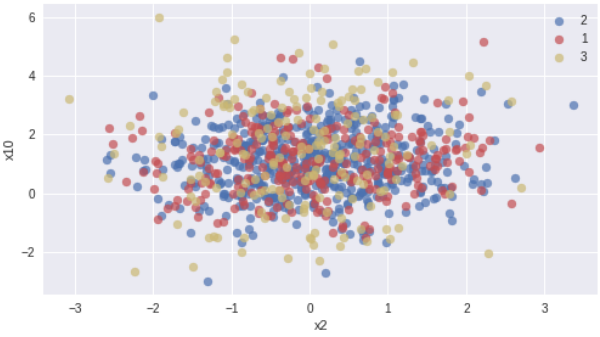



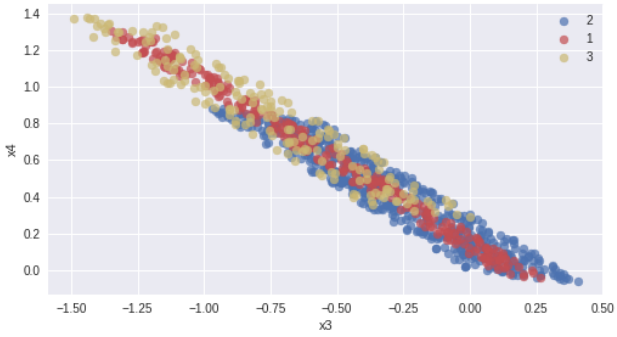



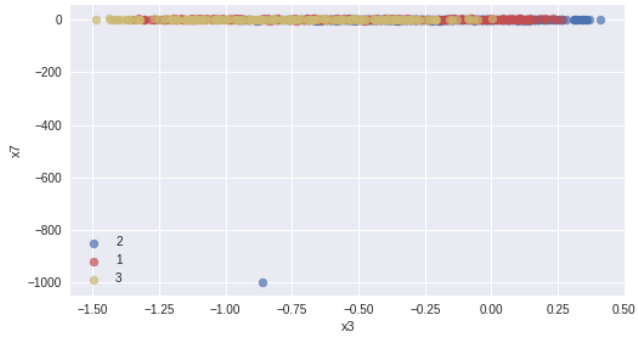



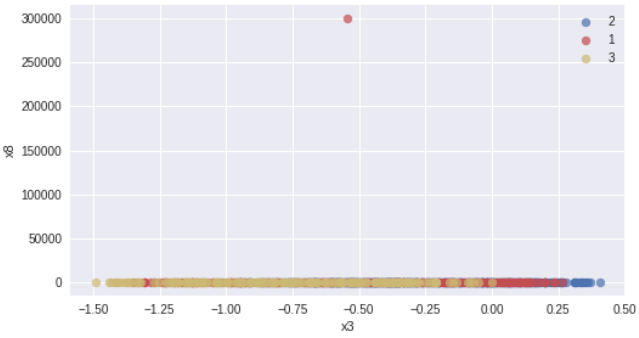

x8
x2


x9
x2

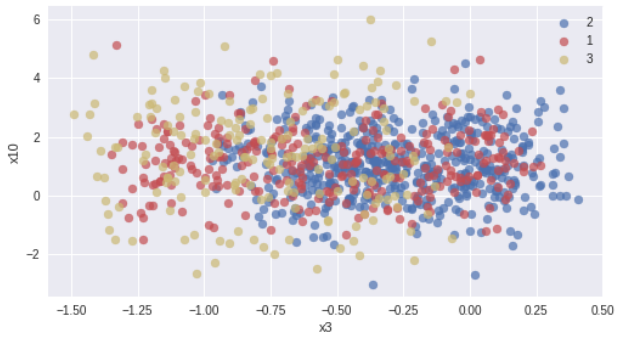

x10
x2
2
1
3

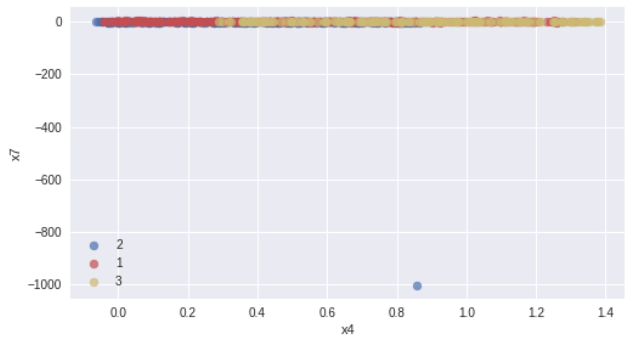

x4
x3
2
1
3

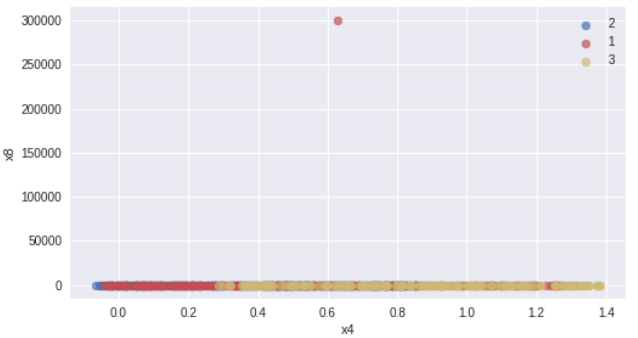

x7
x3
2
1
3

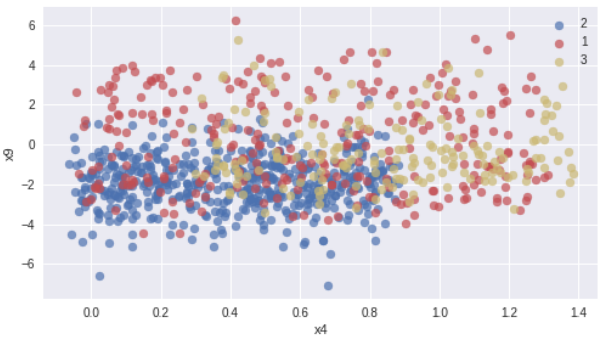

x8
x3
2
1
3

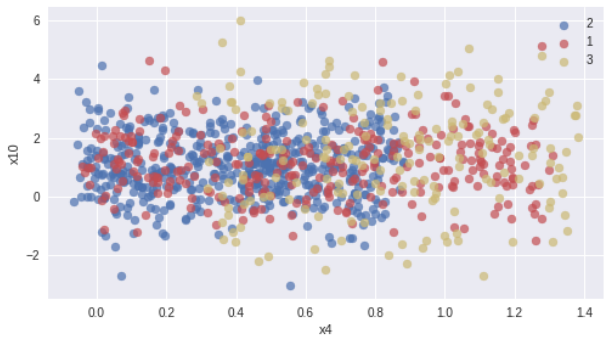

x9
x3
2
1
3

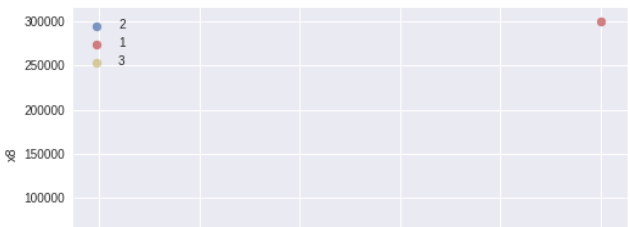

x10
x3
2
1
3

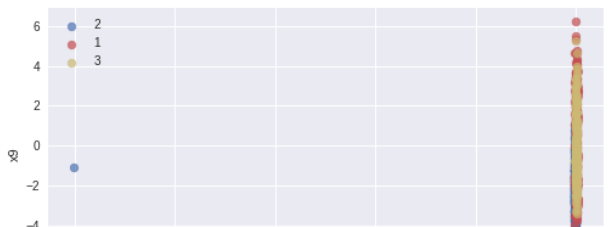

x7
x4
2
1
3

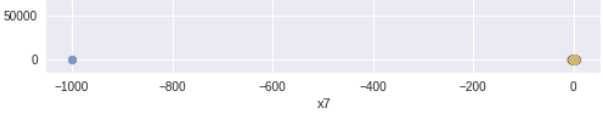

x8
x4
2
1
3

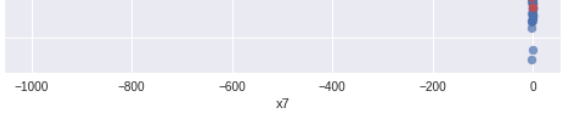

x9
x4
2
1
3

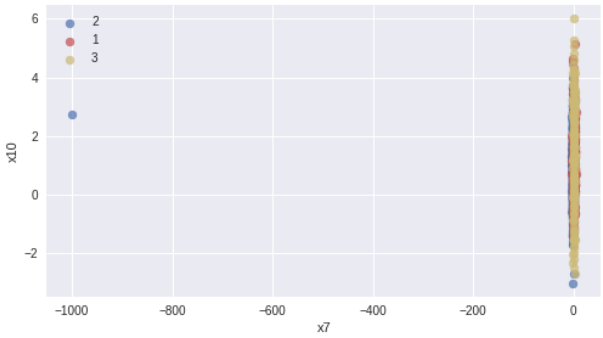

x10
x4
2
1
3

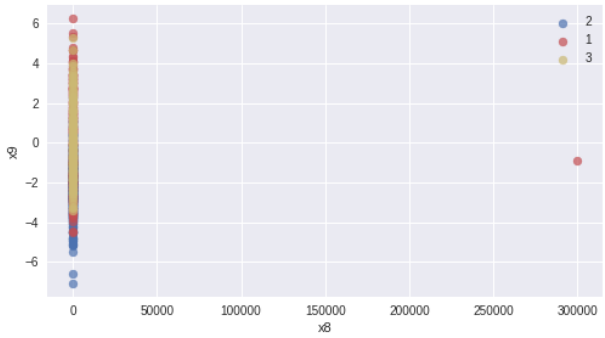

x8
2
1
3

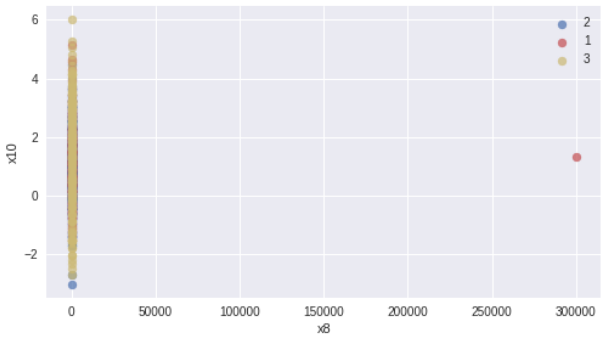

x9
2
1
3

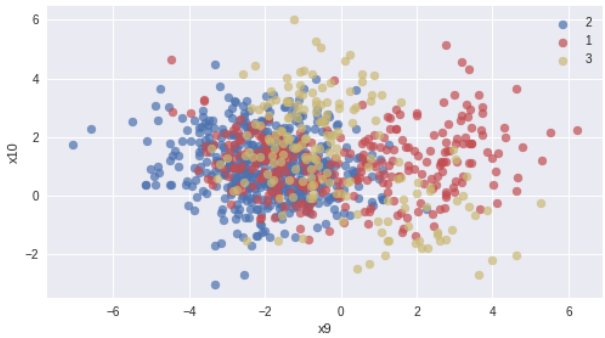

50000
0
−1000
−800
−600
−400
−200
0
x7


−6
−1000
−800
−600
−400
−200
0
x7

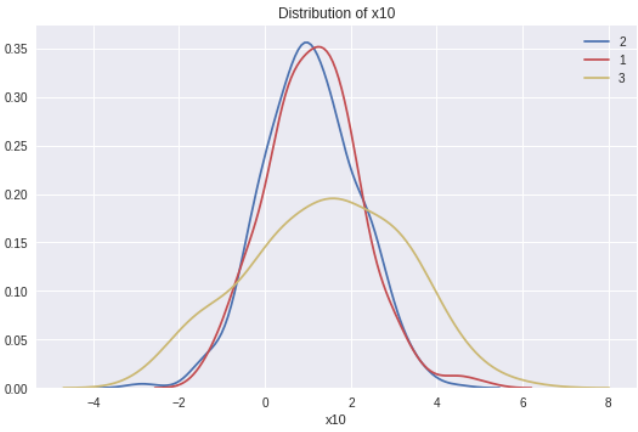

2
1
3
6
4
2
0
−2
x10
−1000
−800
−600
−400
−200
0
x7

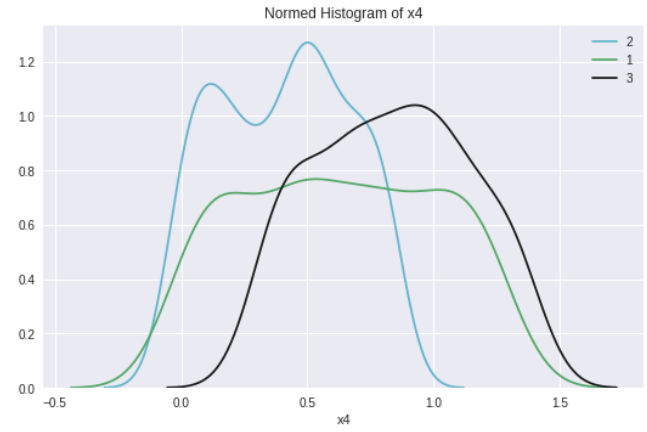

2
1
3
6
4
2
0
−2
−4
−6
x9
0
50000
100000
150000
200000
250000
300000
x8

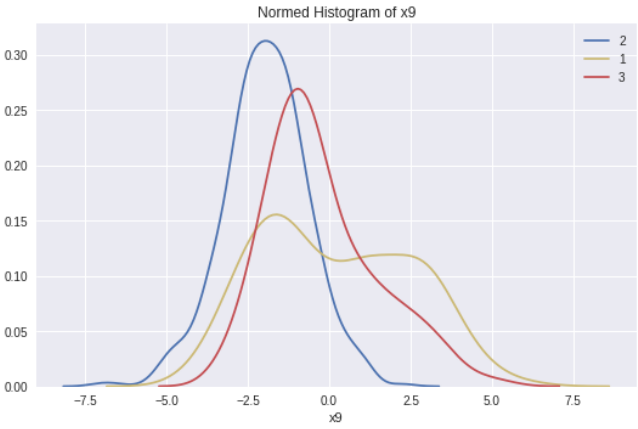

2
1
3
6
4
2
0
−2
x10
0
50000
100000
150000
200000
250000
300000
x8

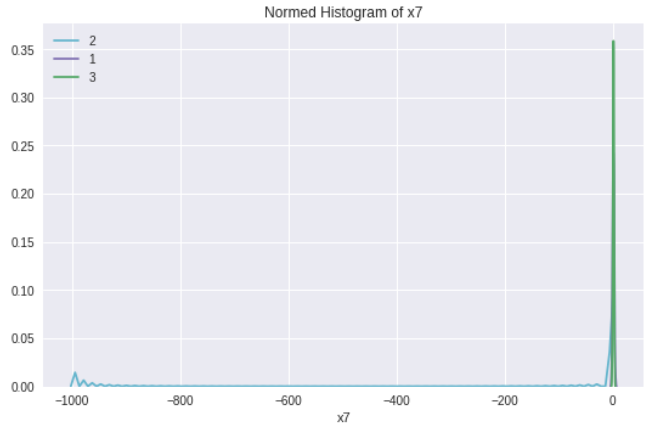

2
1
3
6
4
2
0
−2
x10
−6
−4
−2
0
2
4
6
x9

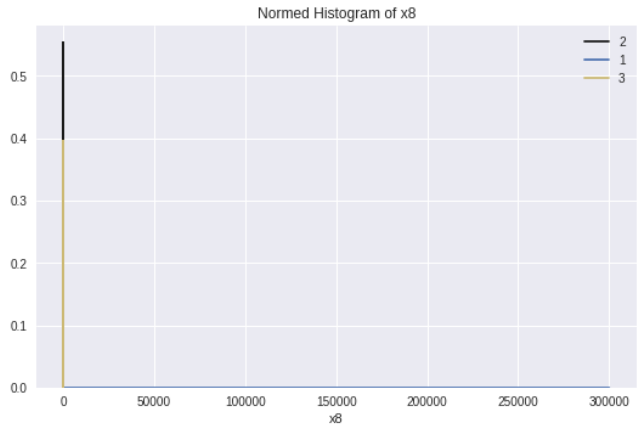
# Histograms (KDE plots) of all Continuous Variables

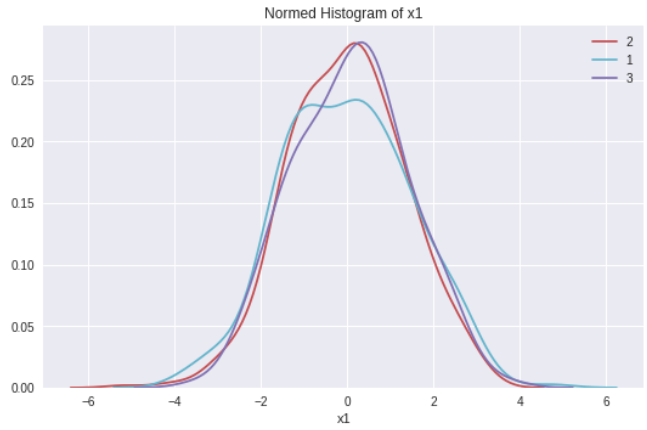


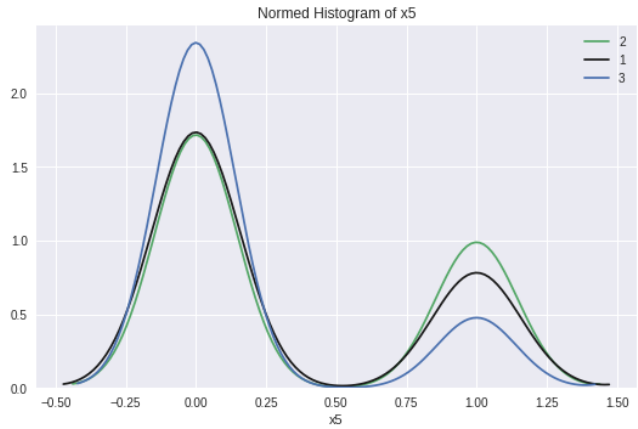



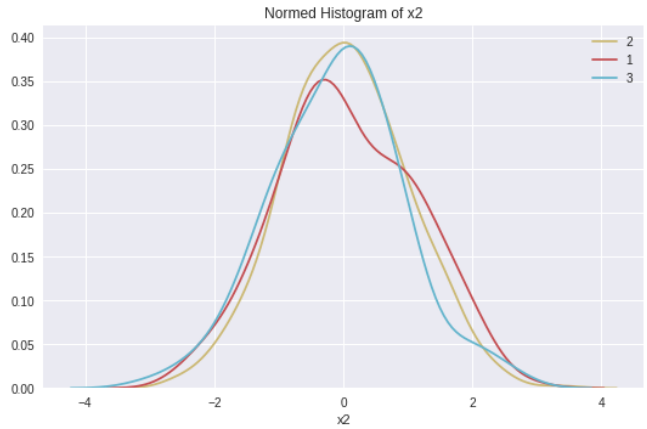



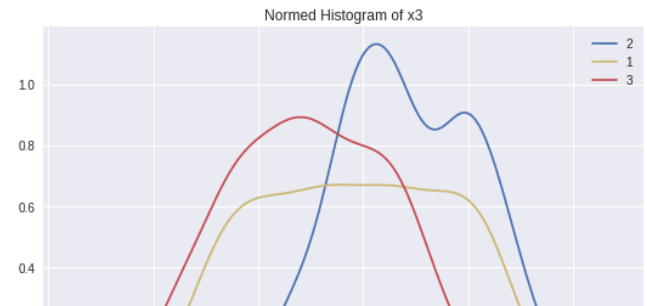



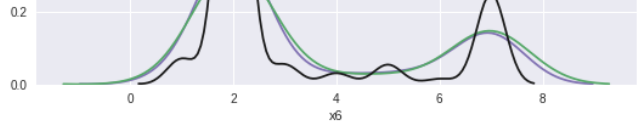



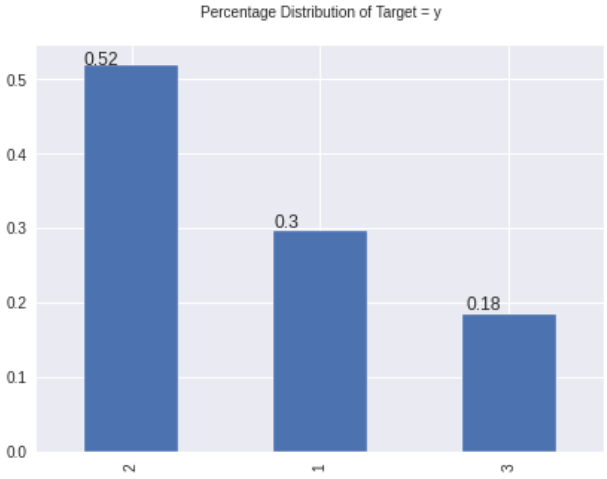


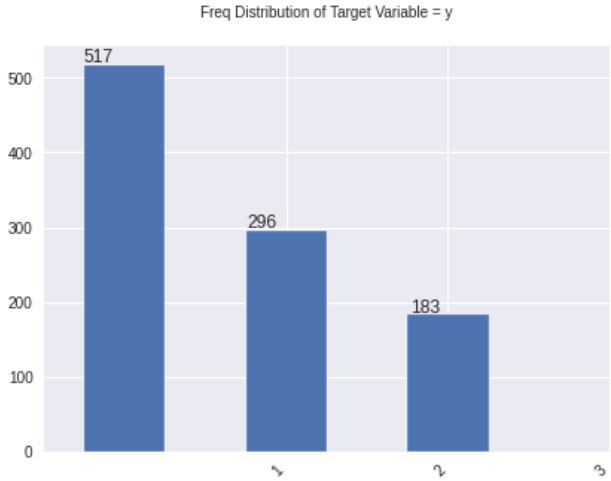




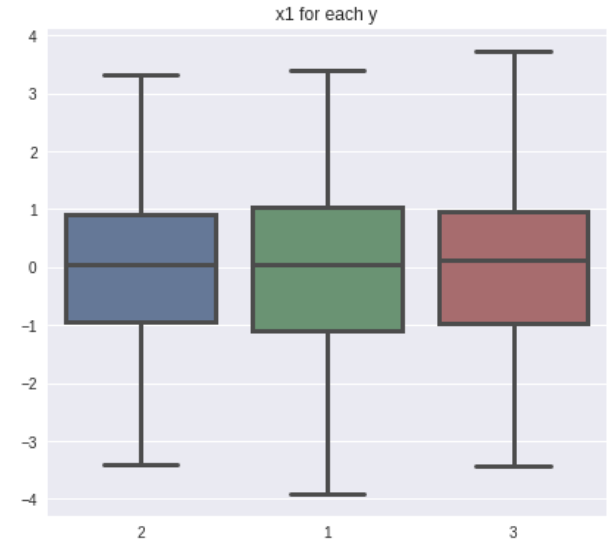


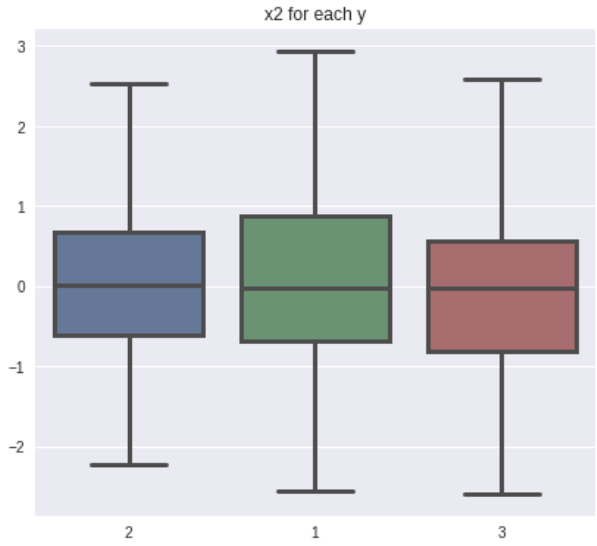
## y : Distribution of Target Variable

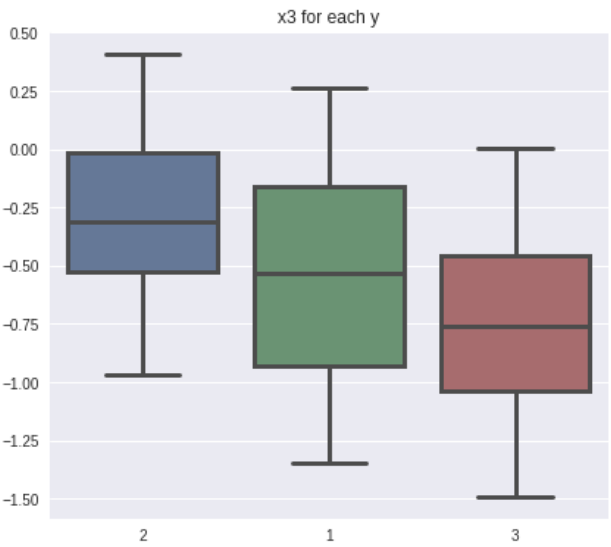


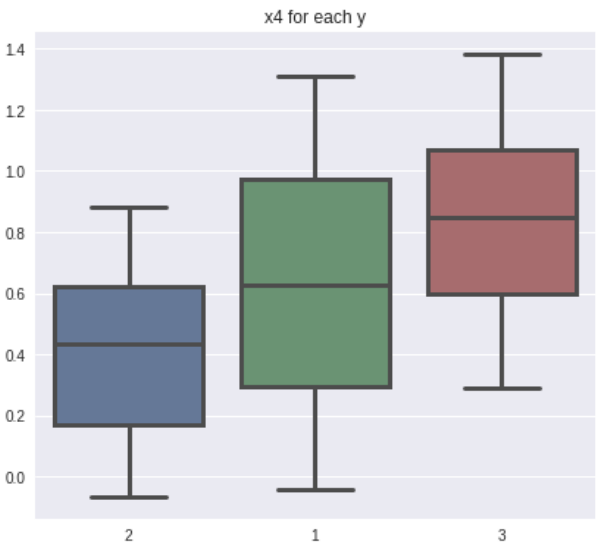



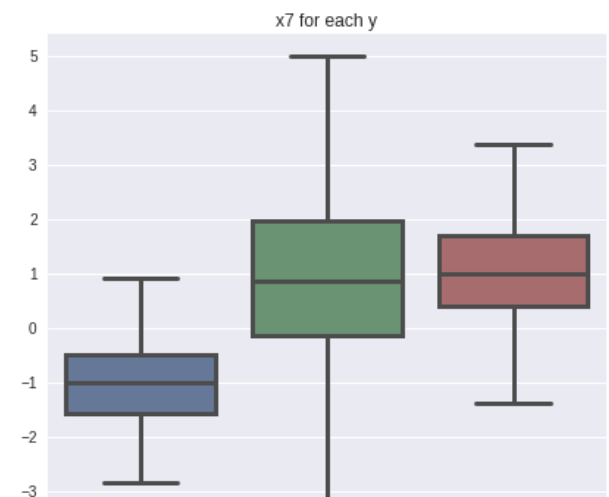
Box Plots without Outliers shown

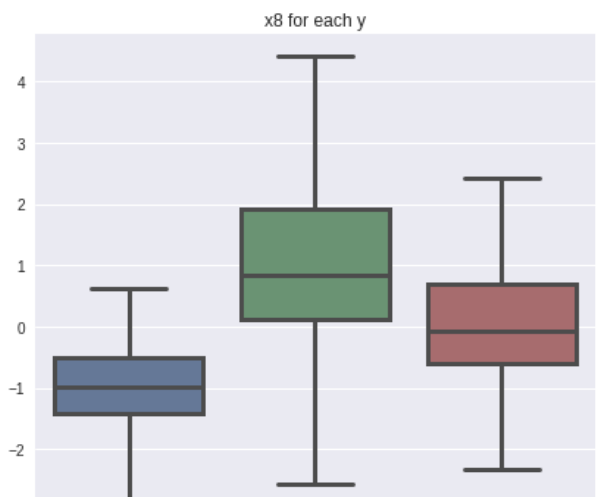


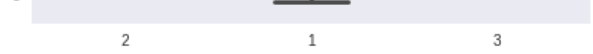



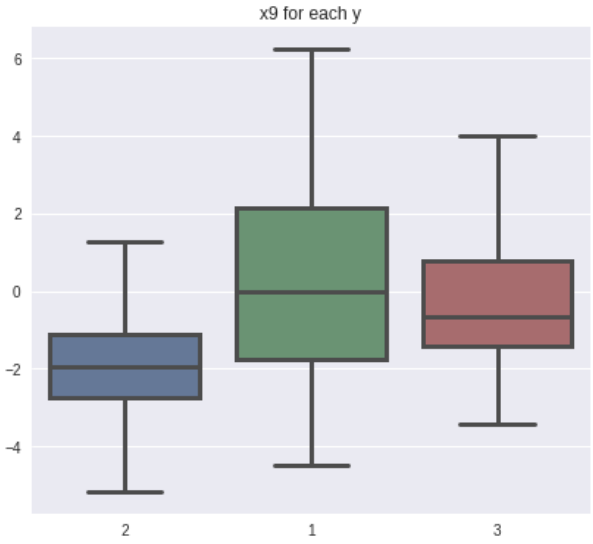



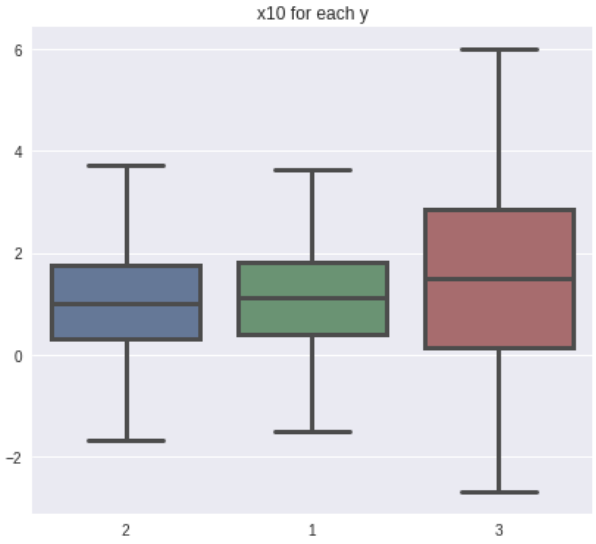



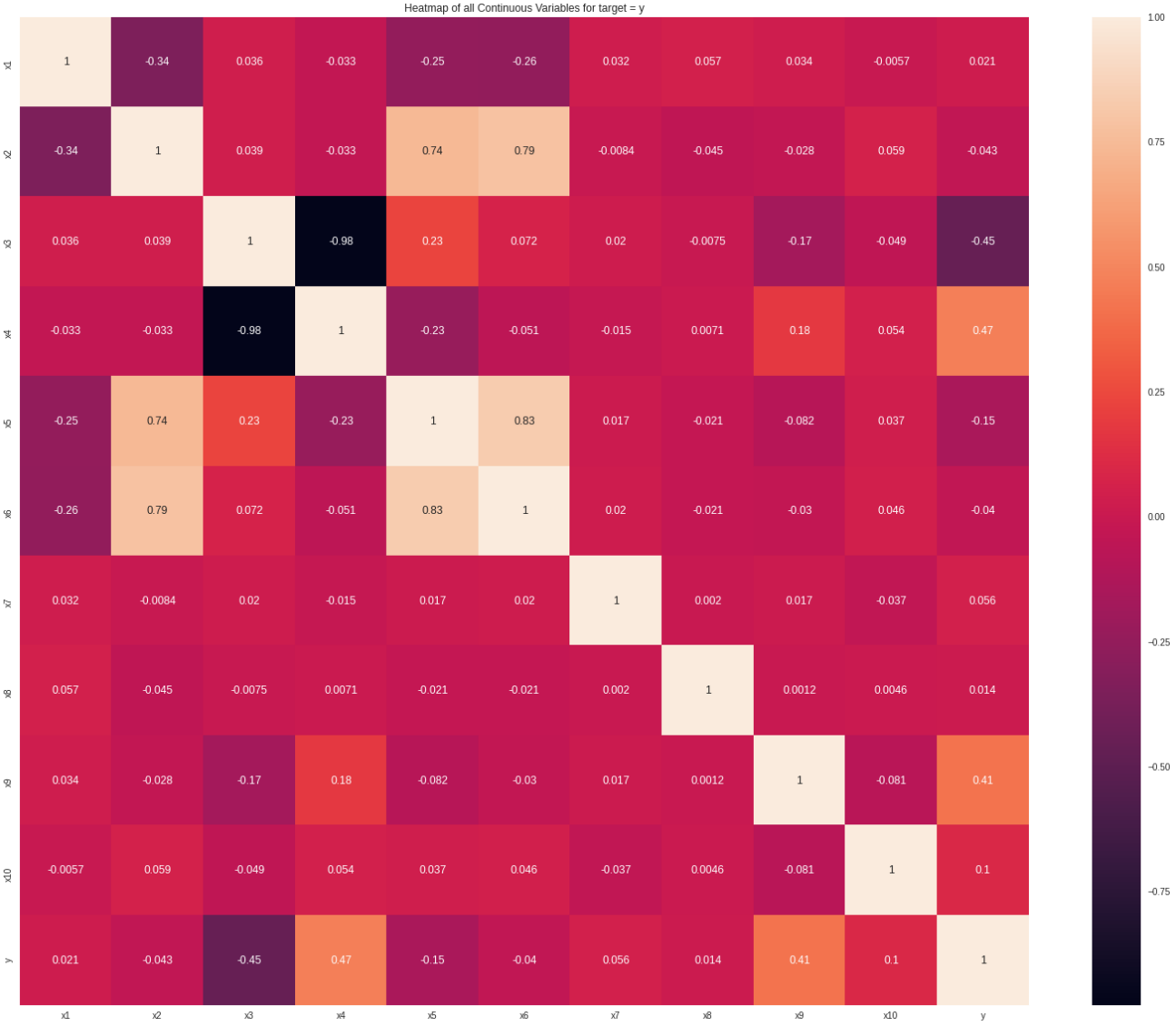




2
1
3

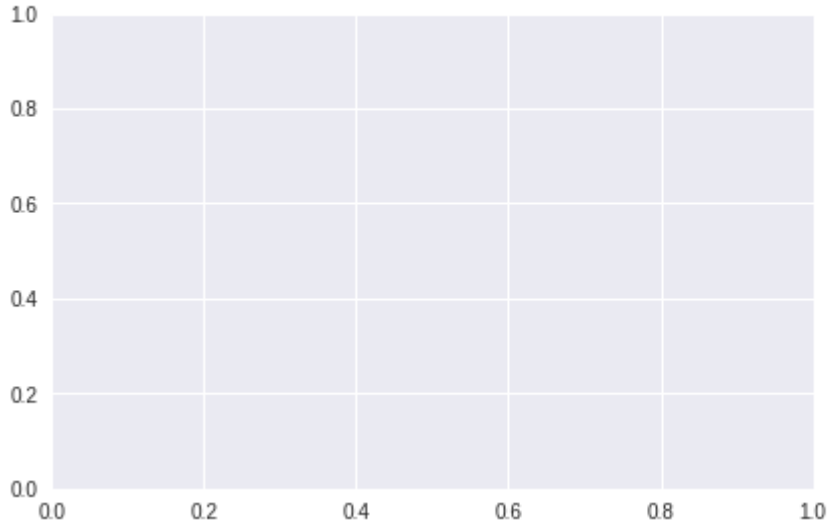

−3
2
1
3


x9 for each y
6
4
2
0
−2
−4
2
1
3

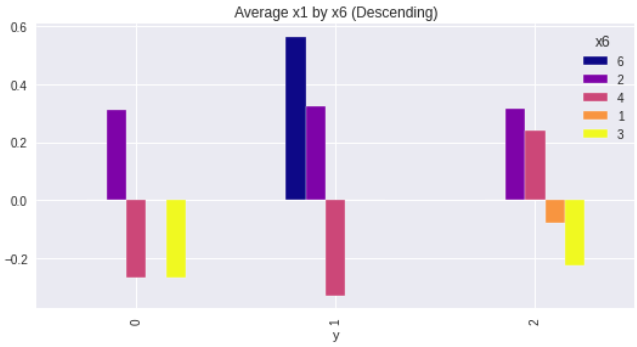

x10 for each y
6
4
2
0
−2
2
1
3

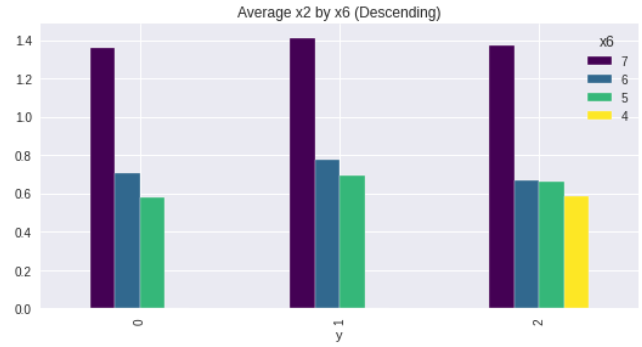

Heatmap of all Continuous Variables for target = y
x1 1 -0.34 0.036 -0.033 -0.25 -0.26 0.032 0.057 0.034 -0.0057 0.021
x2 -0.34 1 0.039 -0.033 0.74 0.79 -0.0084 -0.045 -0.028 0.059 -0.043
x3 0.036 0.039 1 -0.98 0.23 0.072 0.02 -0.0075 -0.17 -0.049 -0.45
x4 -0.033 -0.033 -0.98 1 -0.23 -0.051 -0.015 0.0071 0.18 0.054 0.47
x5 -0.25 0.74 0.23 -0.23 1 0.83 0.017 -0.021 -0.082 0.037 -0.15
x6 -0.26 0.79 0.072 -0.051 0.83 1 0.02 -0.021 -0.03 0.046 -0.04
x7 0.032 -0.0084 0.02 -0.015 0.017 0.02 1 0.002 0.017 -0.037 0.056
x8 0.057 -0.045 -0.0075 0.0071 -0.021 -0.021 0.002 1 0.0012 0.0046 0.014
x9 0.034 -0.028 -0.17 0.18 -0.082 -0.03 0.017 0.0012 1 -0.081 0.41
x10 -0.0057 0.059 -0.049 0.054 0.037 0.046 -0.037 0.0046 -0.081 1 0.1
y 0.021 -0.043 -0.45 0.47 -0.15 -0.04 0.056 0.014 0.41 0.1 1
x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 y
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

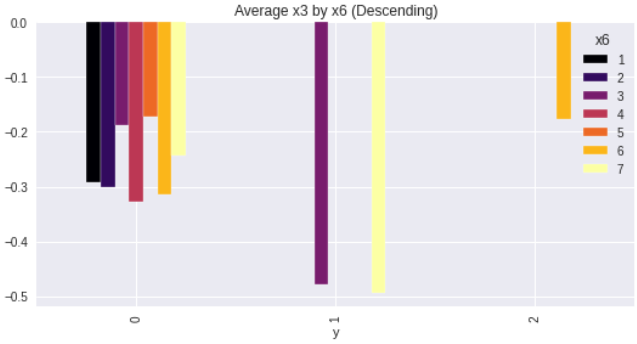
Plots of each Continuous Var by y

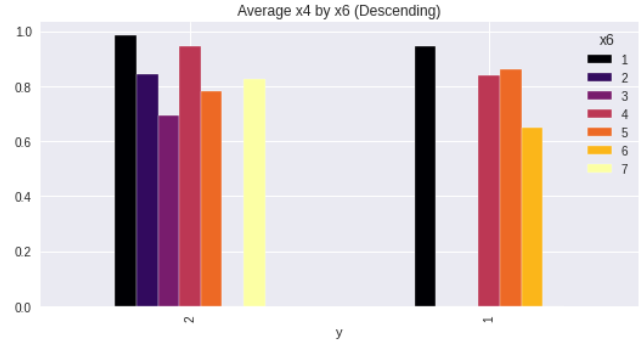


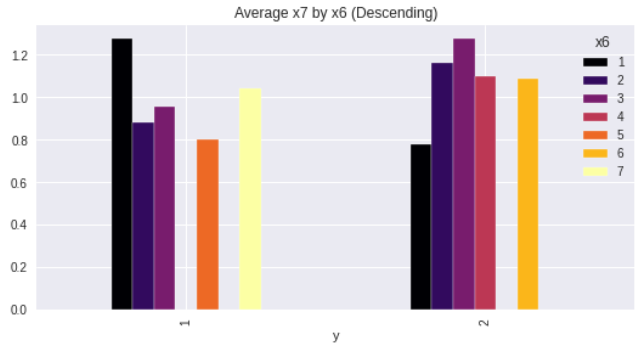
# Bar Plots of Continuous Variables by x6

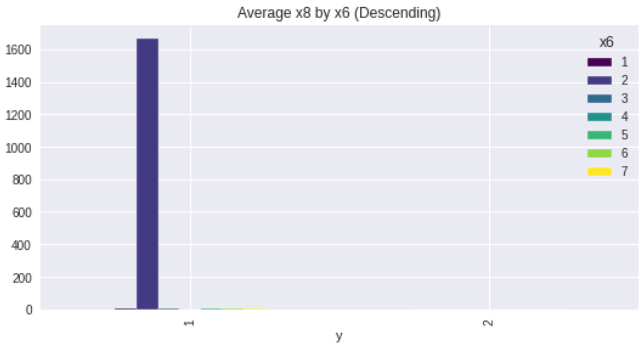


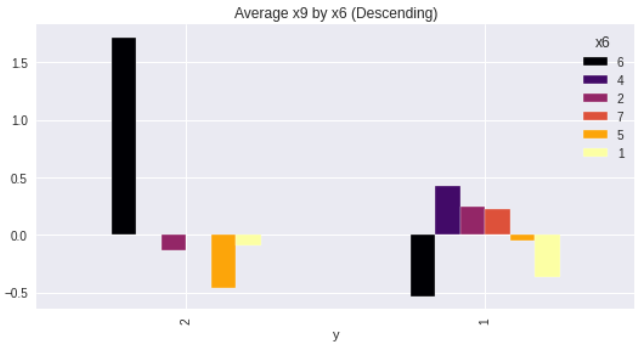



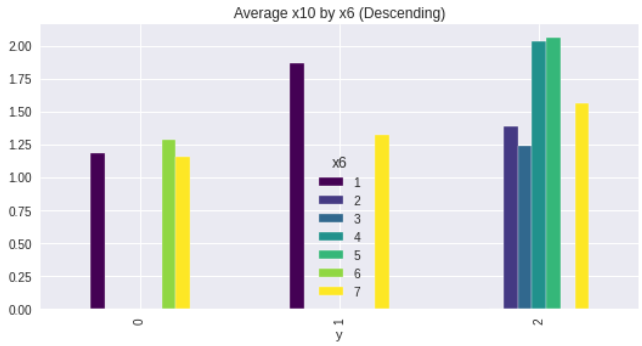



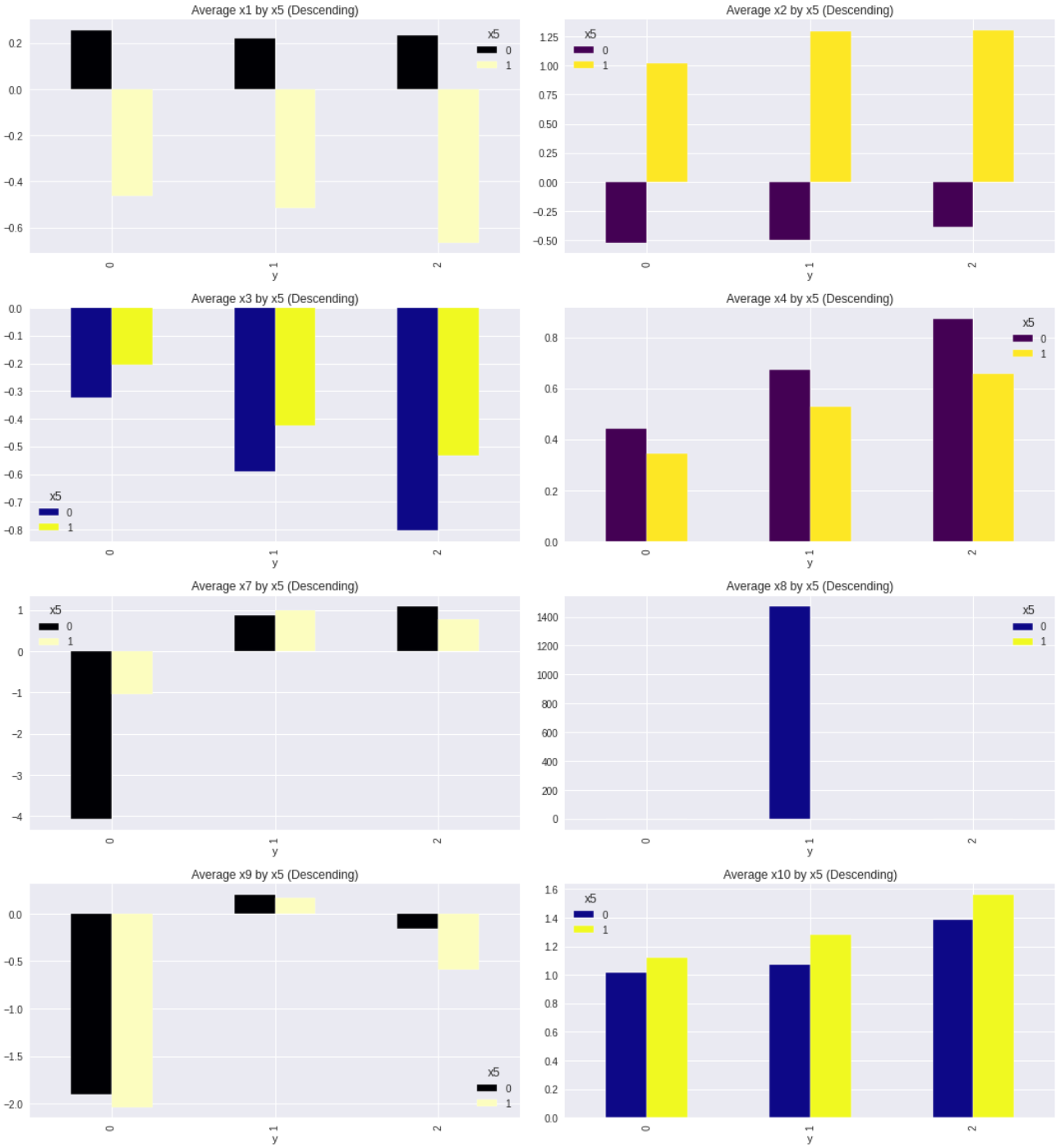



Bar Plots of Continuous Variables by x5



Bar Plots of Continuous Variables by x5