UNIVERSITÄT BERN

# Applying (Un)Supervised Machine Learning Approaches in Social Science

## *Attitudes, Behavior & Socio-economic Status*

08.11.2024 – David von Holzen

# Content

# Background & Motivation

UNIVERSITÄT BERN

- Applying machine learning approaches in social science

- Unsupervised Learning:
  - Identification of underlying factors that represent groups of related questions, simplifying the data for further analysis or visualization
    - e.g. Principal Component Analysis (PCA) to reduce the dimensionality of survey data

- Supervised Learning:
  - Examination of the relationship between (political) attitudes, behavior and socio-economic factors
    - e.g. predict the political attitudes with socio-economic factors such as educational level

# Dataset: Preparation

UNIVERSITÄT BERN

- Survey data from 2021:
  - questions about attitudes, behavior and socio-economic factors (originally collected for research on carbon footprint of individuals)
- Merging survey data from France (N=2150) and Germany (N=2028)
- Data Preparation:
  - Subsetting:
    - Selection of 12 Variables
  - Cleaning:
    - Removing Missing Values
    - Changing types

# Dataset: Variables Part 1

UNIVERSITÄT BERN

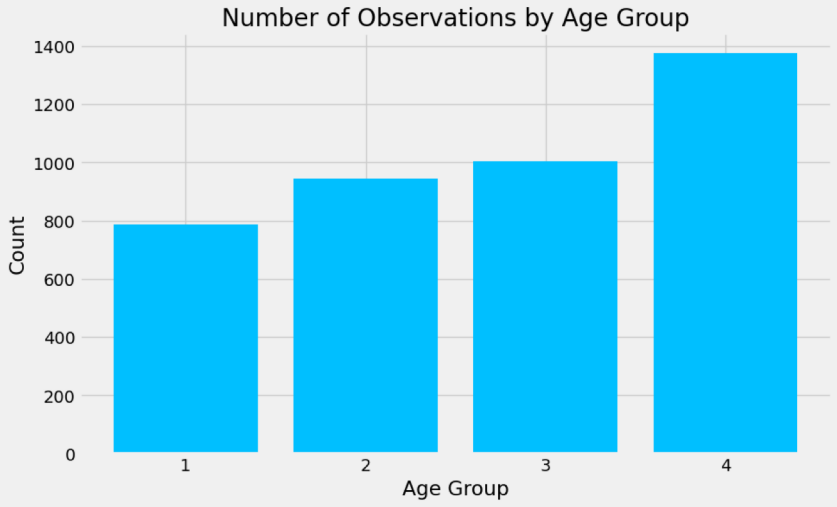
- Socio-Economic Factors:
  - Age Group (1 to 4)
  - Educational Level (1 to 5)
  - Income Level (1 to 5)
- Behavior
  - Airplane Usage (dummy)
  - Nurture (1 to 5)
    - meat, mixed, fish, vegetarian & vegan
  - Resources (1 to 5)
    - Selv-evaluation about the amount of resources used by lifestyle
  - Consumption (1 to 5)
    - Self-evaluation about eco-friendly consumption

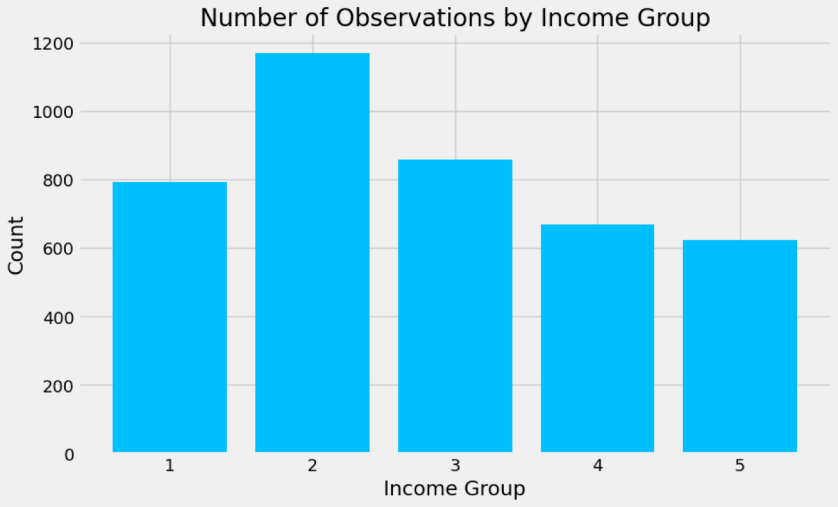
# Dataset: Variables Part 2

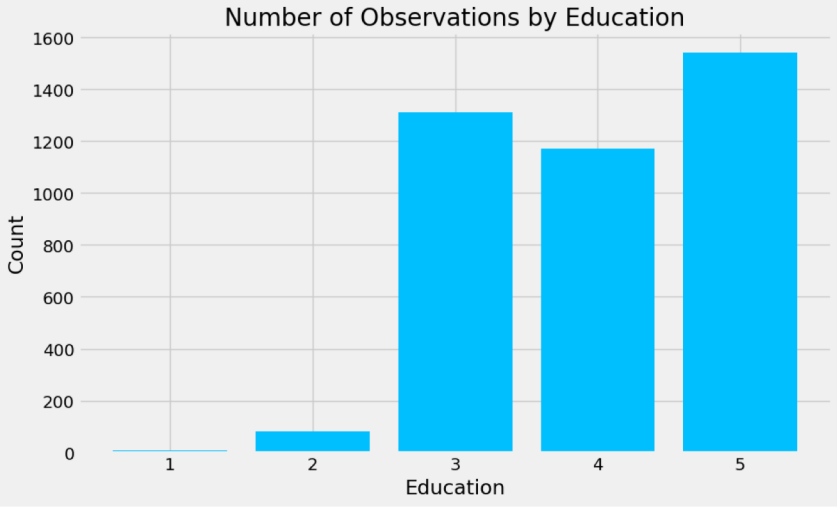
UNIVERSITÄT BERN

- Political Attitude (1 to 5)
  - I support … oriented politics
    - Social
    - Liberal
    - Conservative
    - Environmental
    - National

# Descriptive Statistics: Socio-Economic Factors

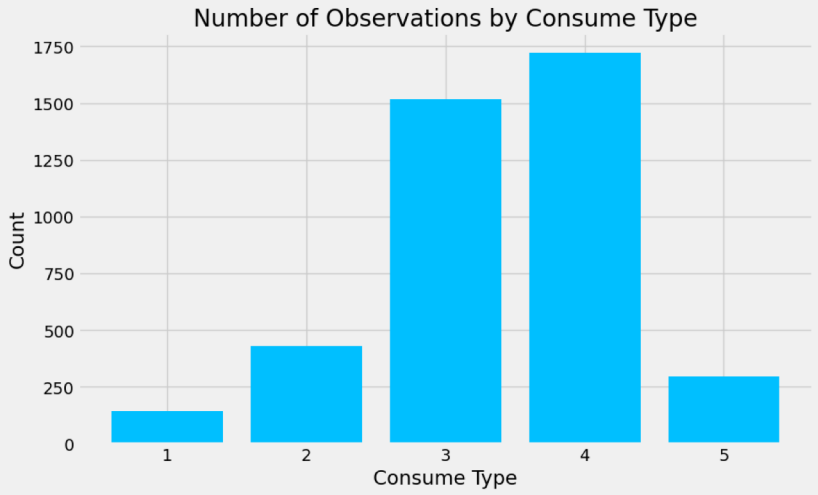


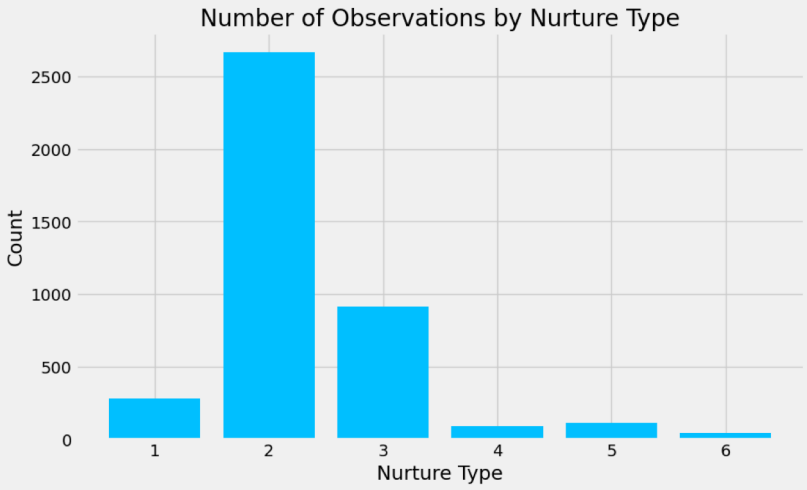



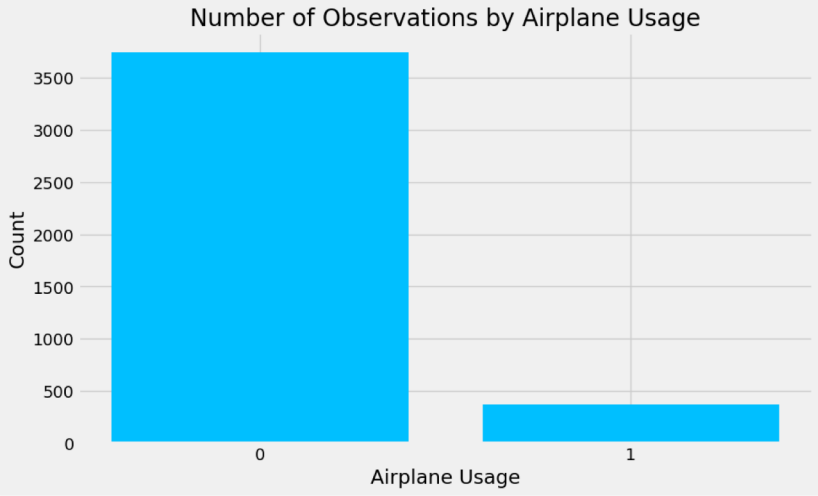



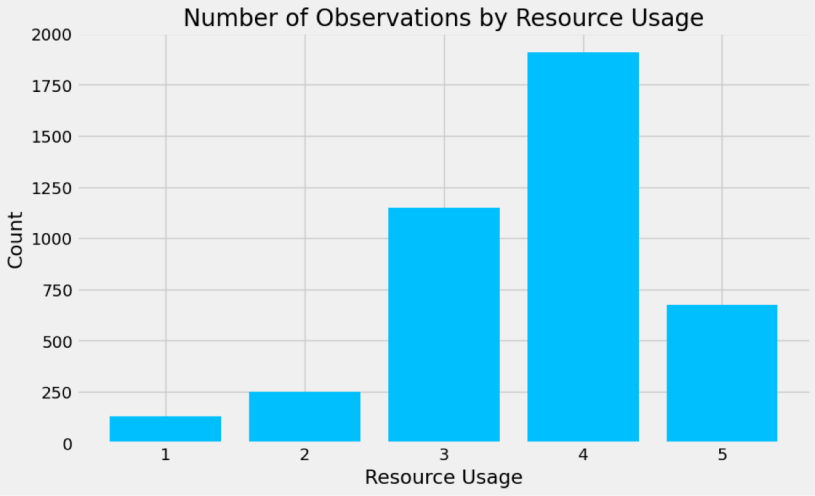
# Descriptive Statistics: Behavior





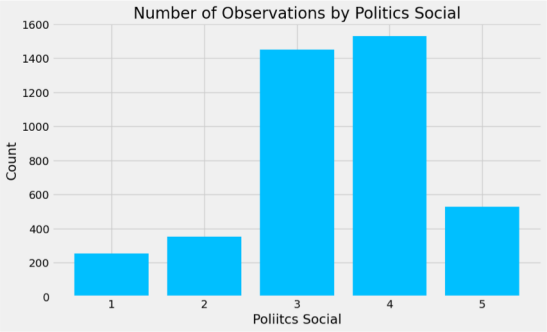



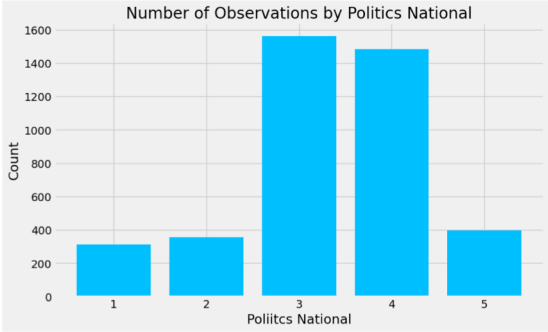



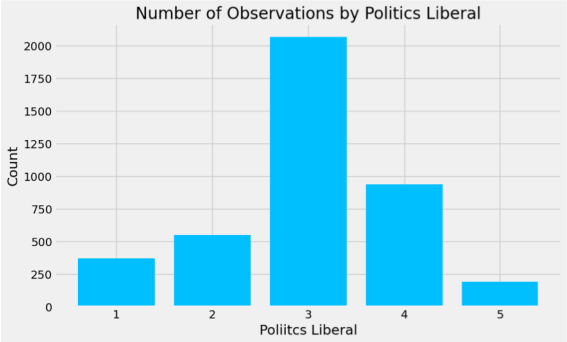



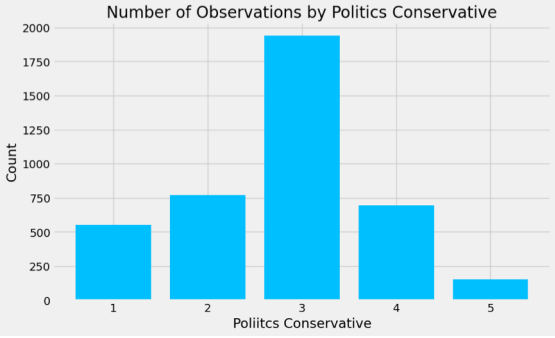
# Descriptive Statistics: Political Attitude

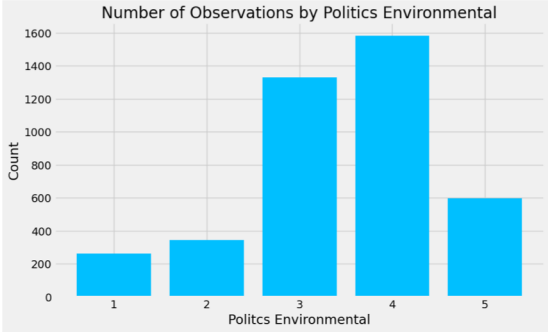




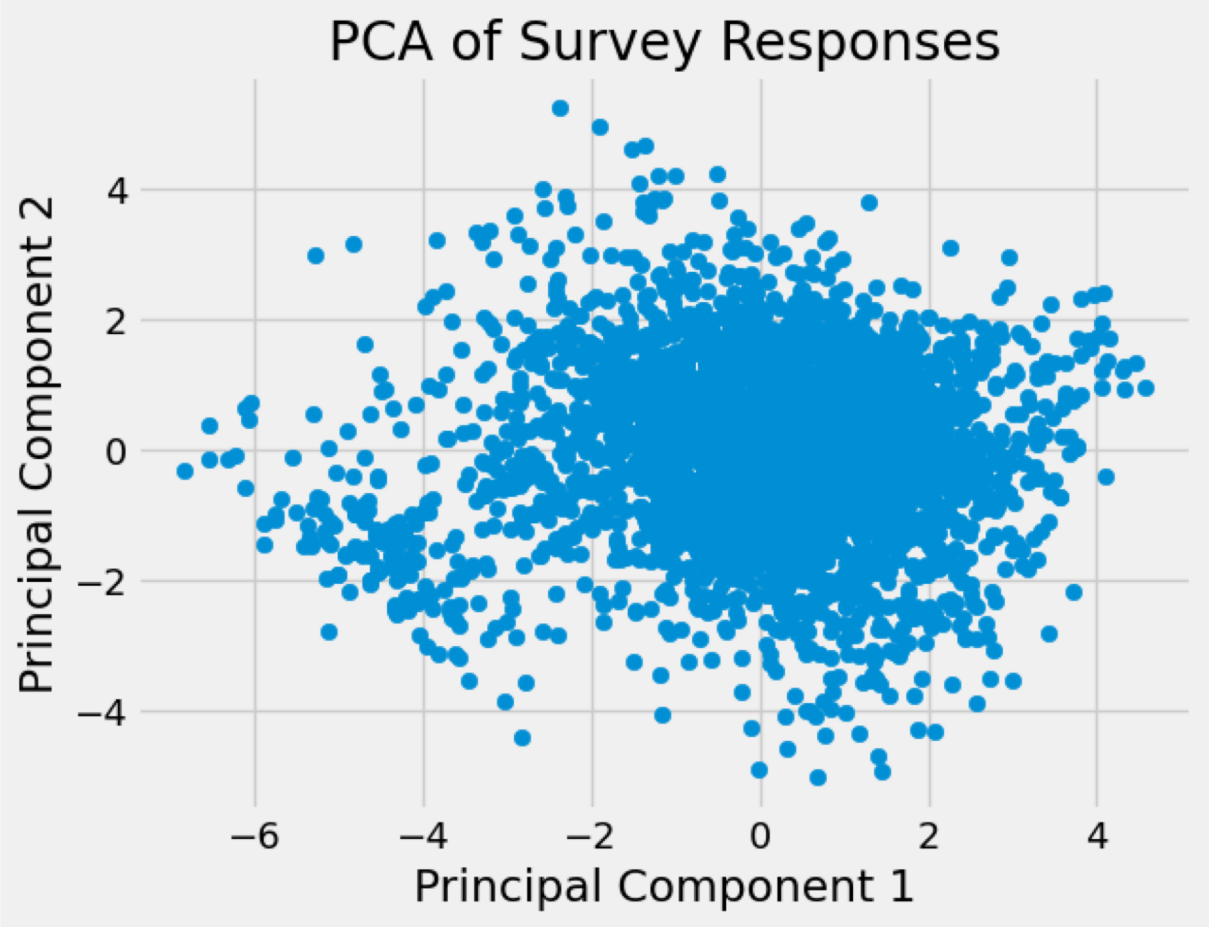
# Part 1: Exploratory Data Analysis

UNIVERSITÄT BERN

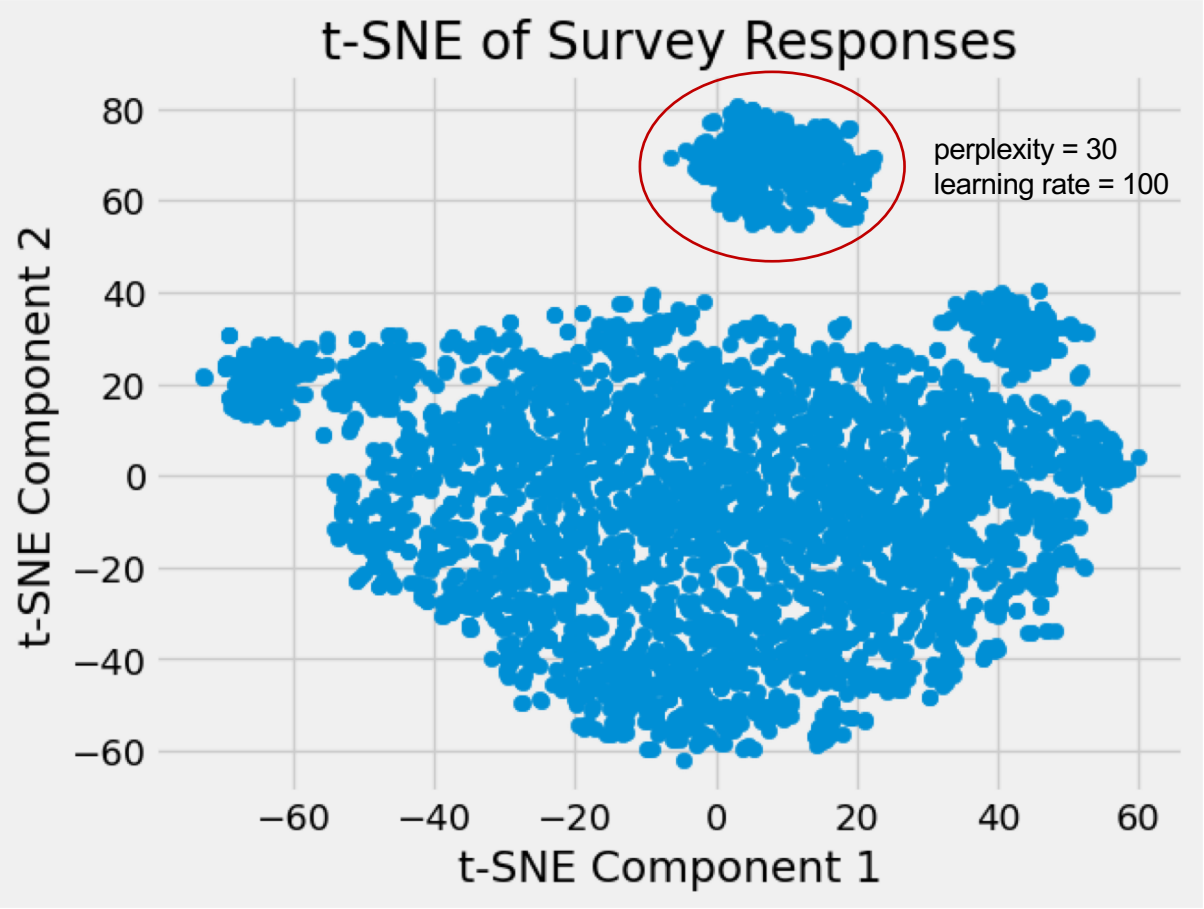
- Purpose:
  - Using an unsupervised approach to take a look at the survey data with more or less randomly selected variables
- Goal:
  - Identifying factors that represent groups of related questions & simplifying the data for further analysis
- Used approaches for the exploratory data analysis:
  - Principal Component Analysis (PCA)
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)

# Exploratory Data Analysis: PCA

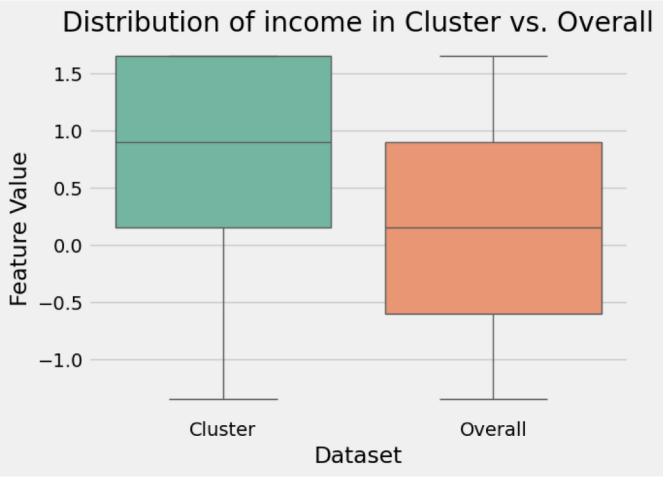


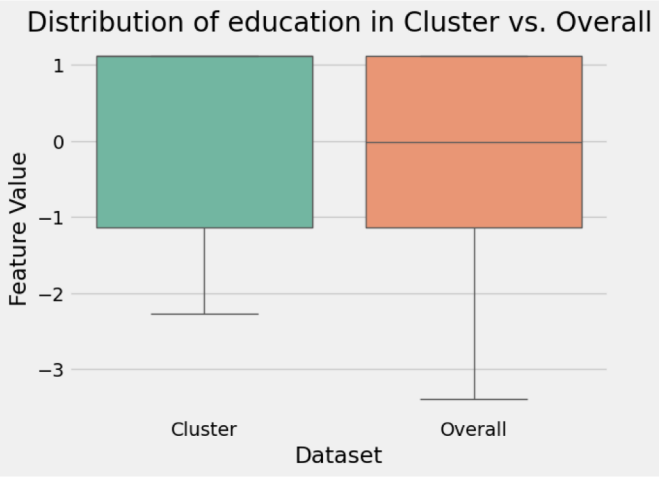


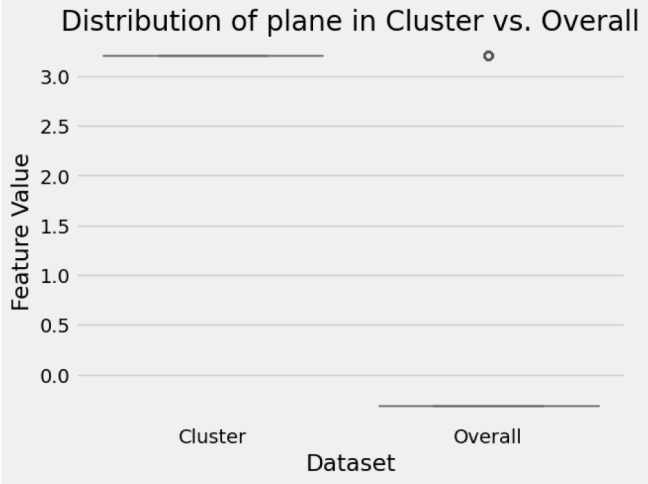
# Exploratory Data Analysis: t-SNE



# Exploratory Data Analysis: Cluster Characteristics



# Model: Airplane Usage by Different Features

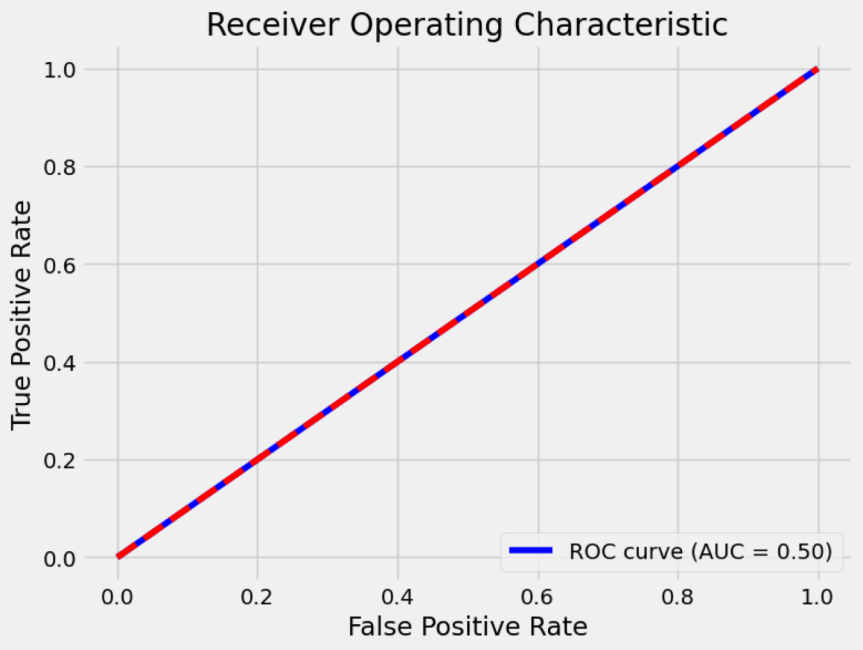
UNIVERSITÄT BERN

| | | |
|---|---|---|
| Dummy Classifier | Confusion Matrix:<br>[[752 0]<br>[ 71 0]] | Classification Report:<br>precision recall f1-score support<br>0 0.91 1.00 0.95 752<br>1 0.00 0.00 0.00 71<br>accuracy 0.91 823 |
| Random Forest Classifier (Balanced) | Confusion Matrix:<br>[[745 7]<br>[ 69 2]] | Classification Report:<br>precision recall f1-score support<br>0 0.92 0.99 0.95 752<br>1 0.22 0.03 0.05 71<br>accuracy 0.91 823 |
| RFC: Education & Income | Confusion Matrix:<br>[[464 288]<br>[ 26 45]] | Classification Report:<br>precision recall f1-score support<br>0 0.95 0.62 0.75 752<br>1 0.14 0.63 0.22 71<br>accuracy 0.62 823 |
| RFC: Income | Confusion Matrix:<br>[[525 227]<br>[ 33 38]] | Classification Report:<br>precision recall f1-score support<br>0 0.94 0.70 0.80 752<br>1 0.14 0.54 0.23 71<br>accuracy 0.68 823 |

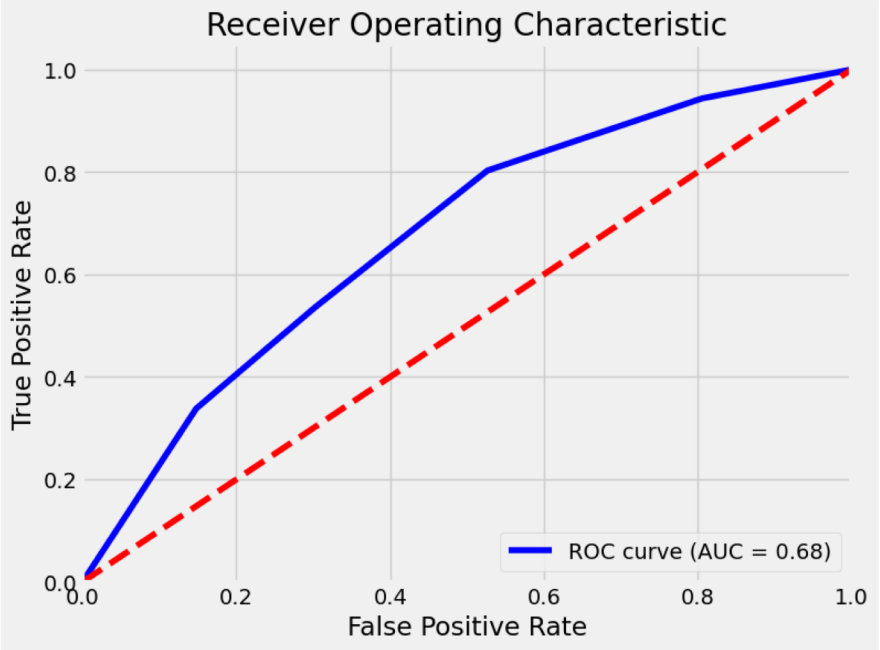
# Model: Airplane Usage by Income

UNIVERSITÄT BERN

## Baseline Model



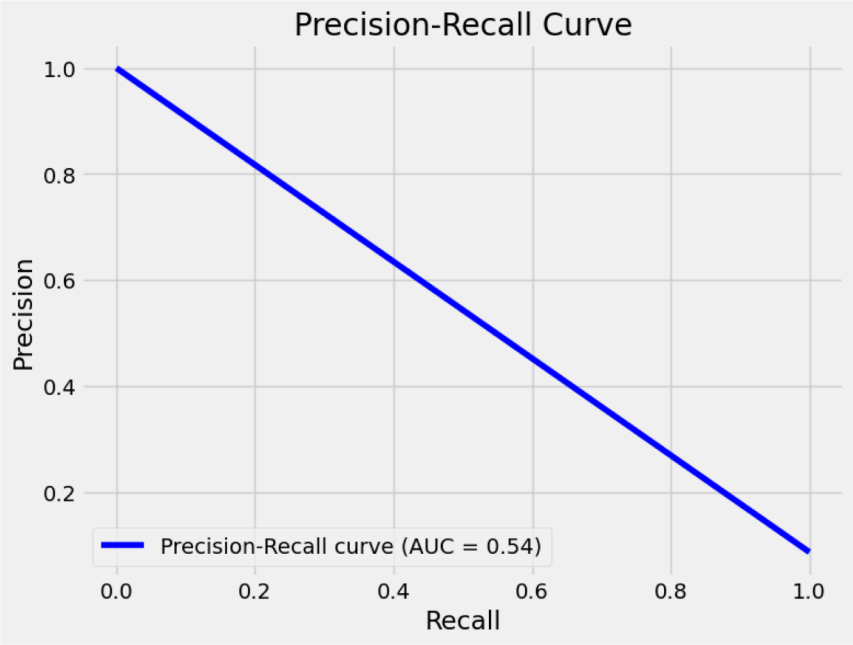

## Final Model



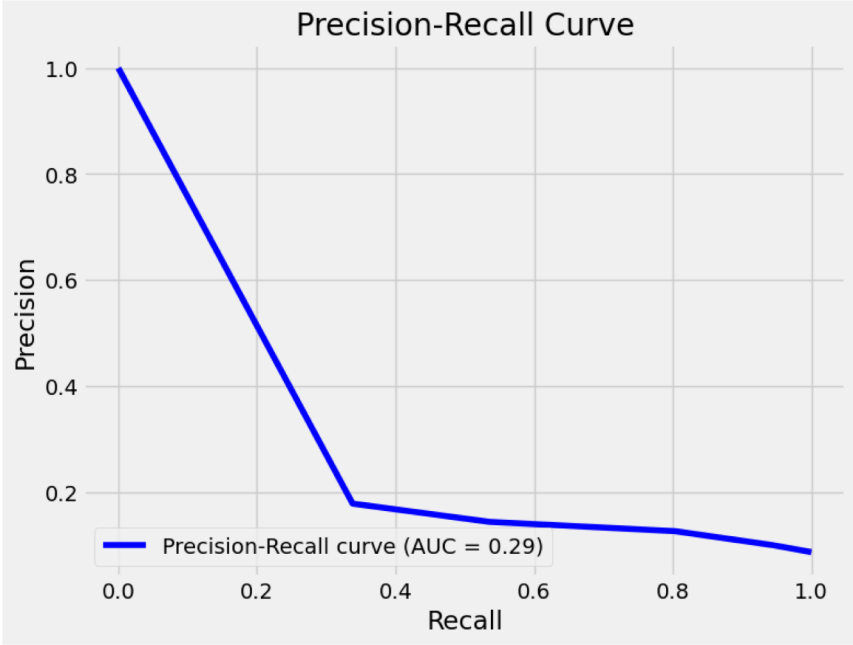
# Model: Airplane Usage by Income

UNIVERSITÄT BERN

## Baseline Model



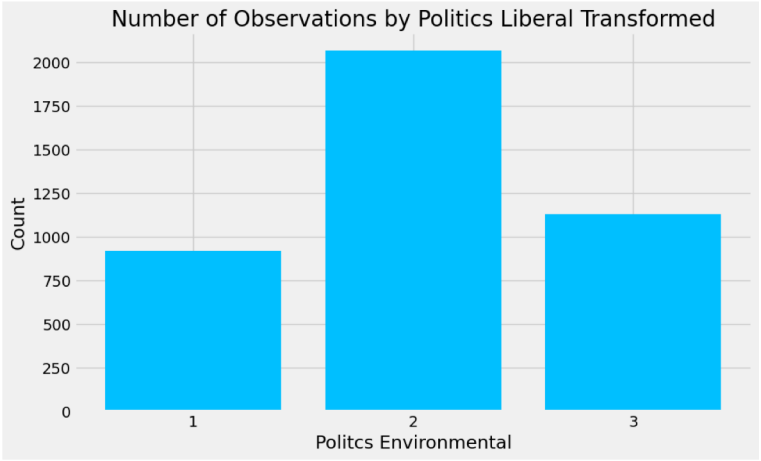

## Final Model



# Part 2: Predictive Models

UNIVERSITÄT BERN

- Try to predict political attitude
  - i.e. political liberalism with age group, educational level and income level
- Classical explanatory variables in political science are socio-economic factors like age, education and income
- Recoding pol_lib → 1 to 3
- Features
  - Age Group
  - Educational Level
  - Income Level
- Label
  - Attitude towards liberal politics



# Model: Liberalism by Age, Education & Income

UNIVERSITÄT BERN

Baseline Model:
Ordinal Logistic Regression

```
Confusion matrix:
[[  0 220   5]
 [  0 488  12]
 [  0 286  17]]
Accuracy: 0.4912451361867704
Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       225
           2       0.49      0.98      0.65       500
           3       0.50      0.06      0.10       303
```

Random Forest Classifier:
all features

```
Feature Importance:
     feature  importance
2     income    0.448230
1  education    0.288195
0  age_group    0.263575
```

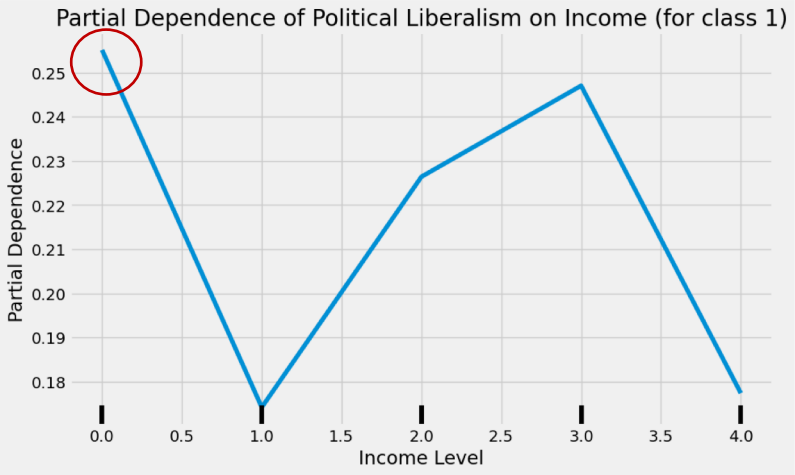
```
Confusion matrix:
[[ 71  92  62]
 [142 238 120]
 [ 67 138  98]]
Accuracy: 0.3959143968871595
Classification Report:
              precision    recall  f1-score   support

           1       0.25      0.32      0.28       225
           2       0.51      0.48      0.49       500
           3       0.35      0.32      0.34       303
```

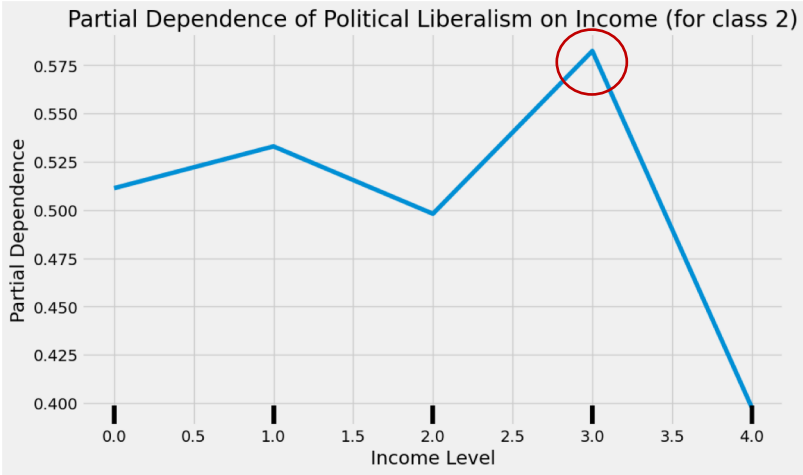
Random Forest Classifier:
Income

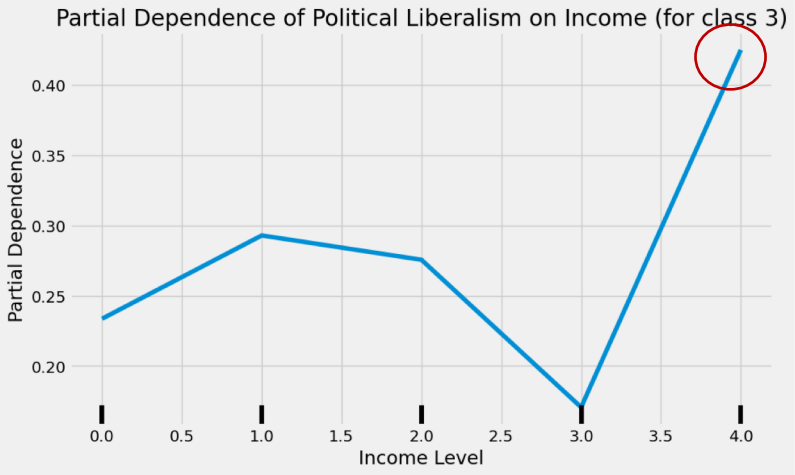
```
Confusion matrix:
[[ 58 105  62]
 [153 211 136]
 [ 74 100 129]]
Accuracy: 0.38715953307393
Classification Report:
              precision    recall  f1-score   support

           1       0.20      0.26      0.23       225
           2       0.51      0.42      0.46       500
           3       0.39      0.43      0.41       303
```

# Partial Dependence of Political Liberalism on Income






No linear relationship between income level and political liberalism → income changes have diminishing or increasing effects at different levels

# Conclusion



- No new revealing insights. But shows how one can find revealing clusters in a survey data set
  - e.g. cluster of individuals with high income and airplane usage
- Further model evaluations necessary
  - e.g. hyperparameter tuning
- Political Science:
  - make predictions for what/who a person will vote for
    - Improving research design by using machine learning tools
- Generally for Social Science:
  - helps with reducing the complexity of survey data
    - PCA/t-SNE: find new relationships in survey data