ACS - Probabilistic Graphical Models (Minutemen DP-pgm)

Evaluation Synthetic Data Creation

Steffen Moritz, Hariolf Merkle, Felix Geyer, Michel Reiffert, Reinhard Tent (DESTATIS)

January 31, 2022

- Executive Summary
- Dataset Considerations
- Privacy and Risk Evaluation
- Utility Evaluation

Executive Summary

We used **minutemen** from the provided python scripts. As you could expect from the **second place submission** of 2021 NIST differential privacy, minutemen scored good in our main privacy metrics. Out of all different methods we tested (`FCS`, `IPSO`, `GAN`, `Simulation`, `Minutemen`) together with the simulation approach it was the **best method in terms of privacy**. Unfortunately, the minutemen method could not produce useful synthetic data based on ACS as well in our case. There is **barely any measure that would indicate a high utility**.

Looking at utility measures is not actually motivating. The `S_pMSE` for tables and for distributions is very high. The Pearson correlation coefficients for binary and (semi-)continuous are in many cases practically zero. The absolute difference in densities and the Bhattacharyya distance show extreme results. There is **no reasonable utility** according to Mlodak's information loss criterion. From a privacy perspective the minutemen looks quite good (also when looking at more detailed metrics).

USE CASE RECOMMENDATIONS

Releasing_to_Public	Testing_Analysis	Education	Testing_Technology
NO	NO	NO	NO

The utility of **minutemen** has some flaws, thus we **don't** think it is a good idea to **release this data to the public**. This could lead to false impressions. Also scientists may be led to false conclusions when using this data for **testing analysis**. We could not imagine our minutemen generated data to be used for **education** or in **technology testing** because of the lost variable dependencies.

Dataset Considerations

When deciding, if data is released to the public it is of utmost importance to define, **which variables** are the most relevant in terms of **privacy and utility**. This process is very **domain and country** specific, since different areas of the world have different privacy legislation and feature specific overall circumstances. This step would require input and discussions with actual domain experts. Since we are foreign to US privacy law, the assumptions made for the Synthetic Data Challenge are basically an **educated guess** from our side. From a utility perspective it is important to know which variables and correlations are **most interesting** for actual users of the created synthetic dataset. Different use cases might require focus on different variables and correlations. We could not single out a most important variable, thus in our utility analysis we decided to focus on the overall utility and not to prioritize a specific variable. We decided to remove the first column of the **ACS** dataset, since it only contains column numbers and hence does not need to be altered by any means. From a privacy perspective it has to be decided, which variables are **confidential** and which are **identifying**. As already mentioned, specifying this depends on multiple factors e.g. regulations or also other public information, that could be used for **de-anonymization**. For our analysis, we made the following assumptions: Of course any information about **income** has to be considered as **confidential**, otherwise publishing income statistics would be a way easier task for NSOs than it actually is. So `INCTOT`, `INCWAGE`, `INCWELFR`, `INCINVST`, `INCEARN` and `POVERTY` are treated as confidential variables. Additionally the times a person is not at home also is an information that encroaches in personal right and might be to the respondents detriment e.g. by burglars. The features HHWT and PERWT are weights that only present information about the way the dataset was created and hence are neither confidential nor identifying. All the other information (like Sex, Age, Race...) contain observable information and hence, in our opinion, are **identifying variables**. # Method Considerations

Privacy and Risk Evaluation

Disclosure Risk (R-Package: synthpop with own Improvements)

Our starting point was the **matching of unique records**, as described in the disclosure risk measures chapter of the starter guide. The synthpop package provides us with an easy-to-use implementation of this method: `replicated.uniques`. However, one downside of just using `replicated.uniques` is that it does **not consider almost exact matches in numeric variables**. Imagine a data set with information about the respondents' income. If there is a matching data point in the synthetic data set for a unique person in the original data set, that only differs by a slight margin, the original function would not identify this as a match. **Our solution** is to borrow the notion of the **p% rule** from **cell suppression methods**, which identifies a data point as critical, if one can guess the original values with **some error of at most p%**. Thus, **our improved risk measure** is able to evaluate disclosure risk in numeric data. Our Uniqueness-Measure for **"almost exact"** matches provides us with the following outputs:

- **Replication Uniques** | Number of unique records in the synthetic data set that replicates unique records in the original data set w.r.t. their quasi-identifying variables. In brackets, the proportion of replicated uniques in the synthetical data set relative to the original data set size is stated.

- **Count Disclosure** | Number of replicated unique records in the synthetical data set that have a real disclosure risk in at least one confidential variable, i.e. there is at least one confidential variable where the record in the synthetical data set is "too close" to the matching unique record in the original data set. We identify two records as "too close" in a variable, if they differ in this variable by at most p%.

- **Percentage Disclosure** | Proportion of the number of replicated unique records in the synthetical data set that have a real disclosure risk in at least one confidential variable relating to the original data set size. For our selected best parametrized solution in this method-category, we got the following results:

Replication.Uniques	Number.Replications	Percentage.Replications
0	0	0

Perceived Disclosure Risk (R-Package: synthpop)

Unique records in the synthetic dataset may be **mistaken for unique records** based on the fact that **only the identifying variables match**. This can lead to problems, even if the associated confidential variables significantly differ from the original record. E.g. people might assume a certain income for a person, because they believe to have identified her from the identifying variables. Even if her real income **is not leaked** (as the confidential variables are different), this assumed (but wrong) information about him **might lead to disadvantages**. The **perceived risk** is measured by matching the unique records among the quasi-identifying variables (compare with non-confidential variables in Section "Dataset Considerations"). We applied the method `replicated.uniques` of the synthpop package. There is no fixed threshold that must not be exceeded in this measure, however, a smaller percentage of unique matches (referred to as Number Replications) is preferred to minimize the perceived disclosure risk. These are the results variables for perceived disclosure risk:

- **Number Uniques** | Number of unique individuals in the original data set.

- **Number Replications** | The number of matching records in the synthetic data set (based only on identifying variables). This is the number of individuals, which might perceived as disclosed (real disclosures would also count into this metric).

- **Percentage Replications** | The calculated percentage of duplicates in the synthetic data. For our selected best parametrized solution in this method-category, we got the following results:

| Metric	Number.Uniques	Number.Replications	Percentage.Replications
Perceived Risk	1035201	0	0

Utility Evaluation

Different utility measures are applied in this section. These utility measures are the basis of utility evaluation for the generated synthetic dataset. The R packages synthpop, sdcMicro and corrplot were used to compute the following metrics. We do not use tests incorporating significance here.

Confidence intervals in large surveys often tend to be extremely small so many slight differences appear to be significant. We do not consider the variable PUMA for our utility evaluation. During the ACS reports, some minor changes in availability regarding plots might occur. This is caused by the application of standardised scripts on different synthetic datasets.
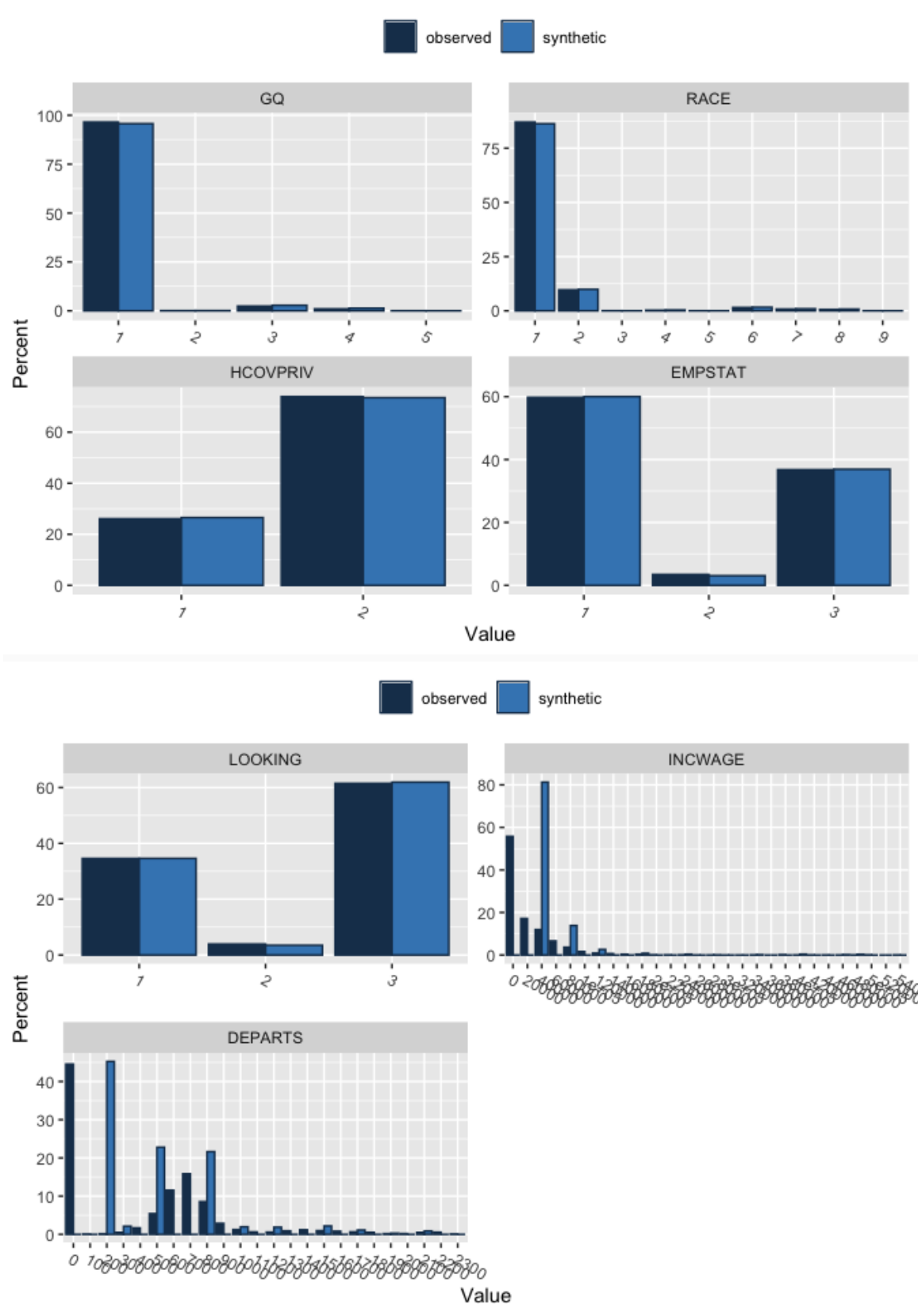
Graphical Comparison for Margins (R-Package: synthpop)

The following histograms provide an ad-hoc overview on the marginal distributions of the original and synthetic dataset. Matching or close distributions are related to a high data utility.
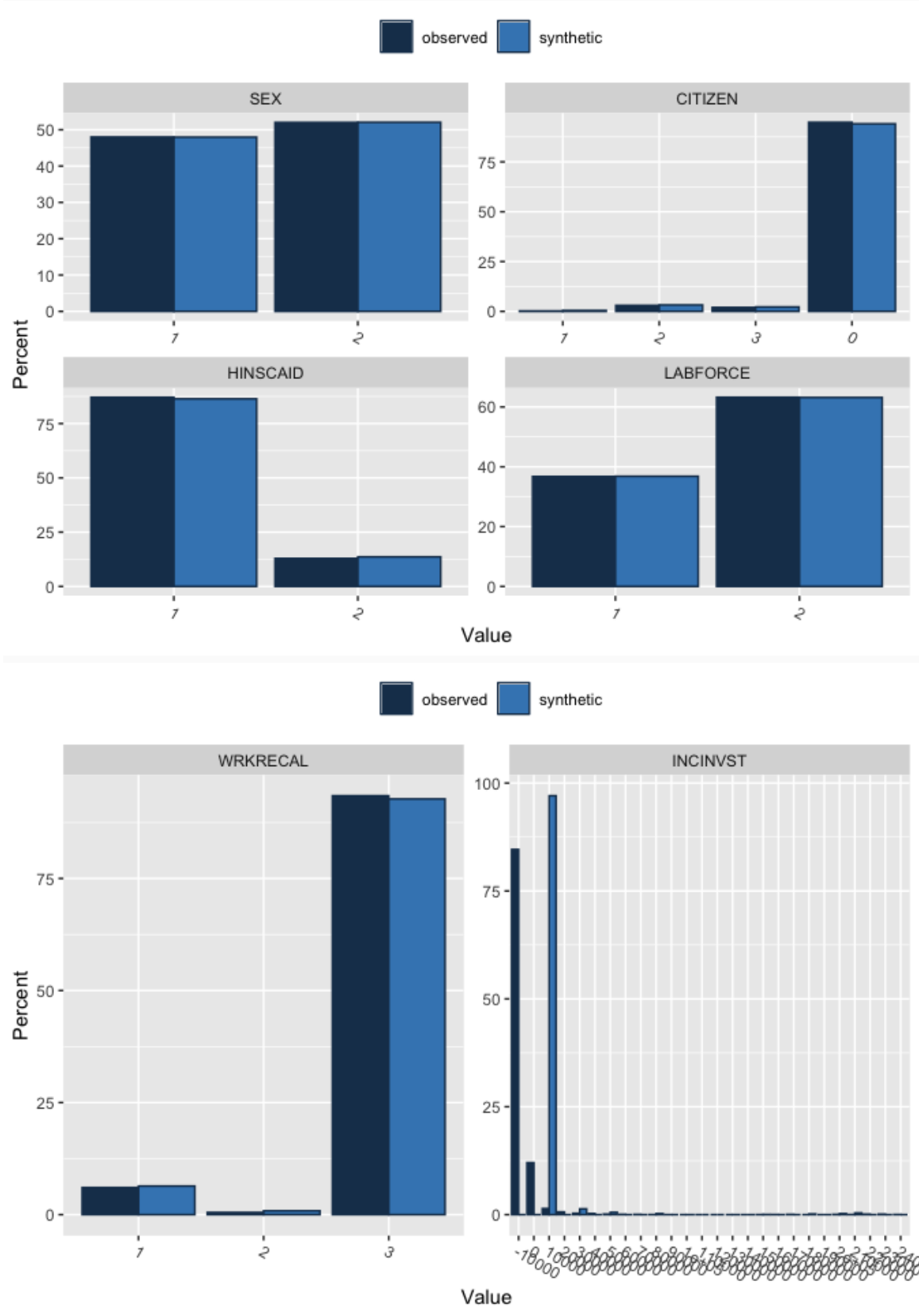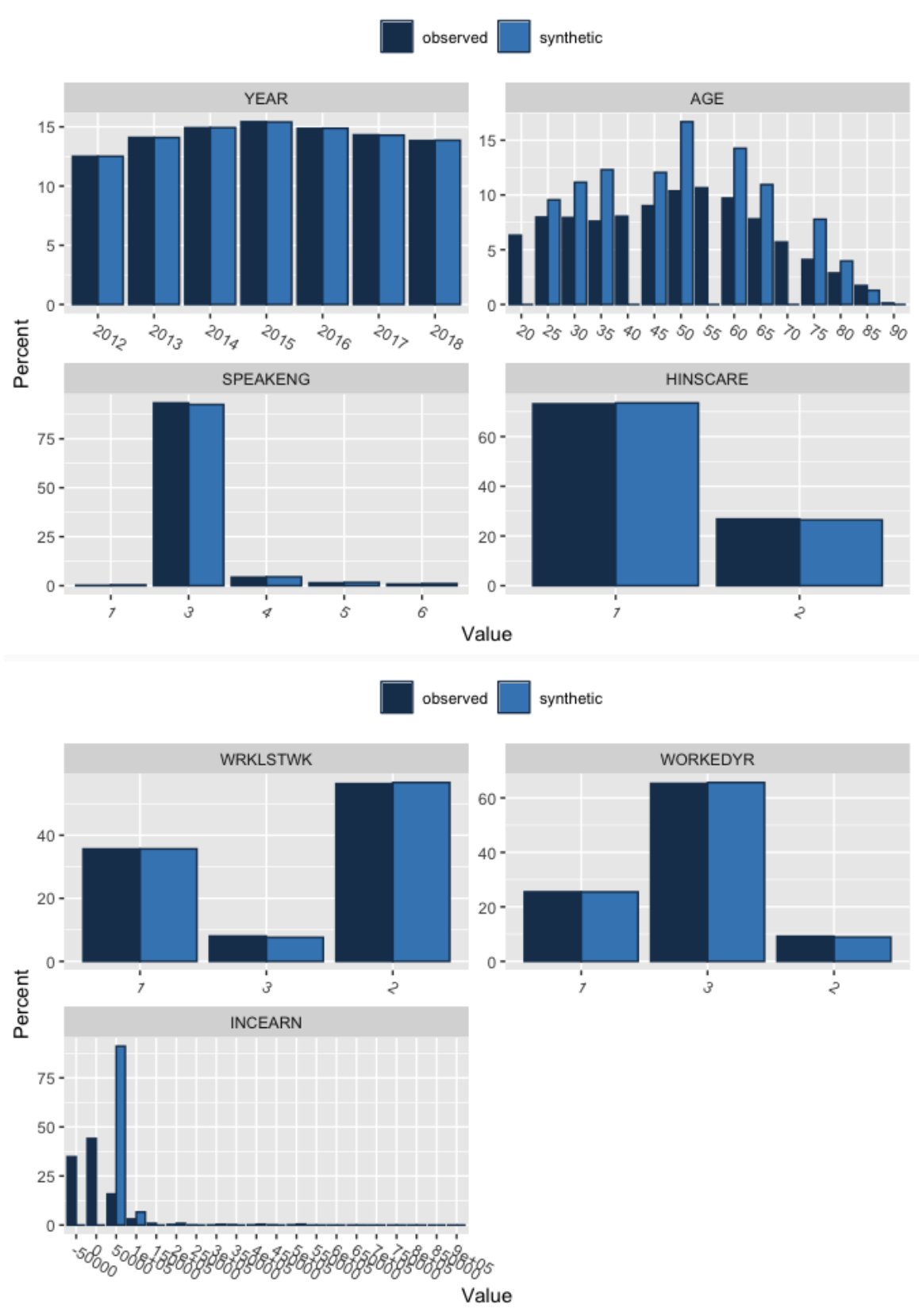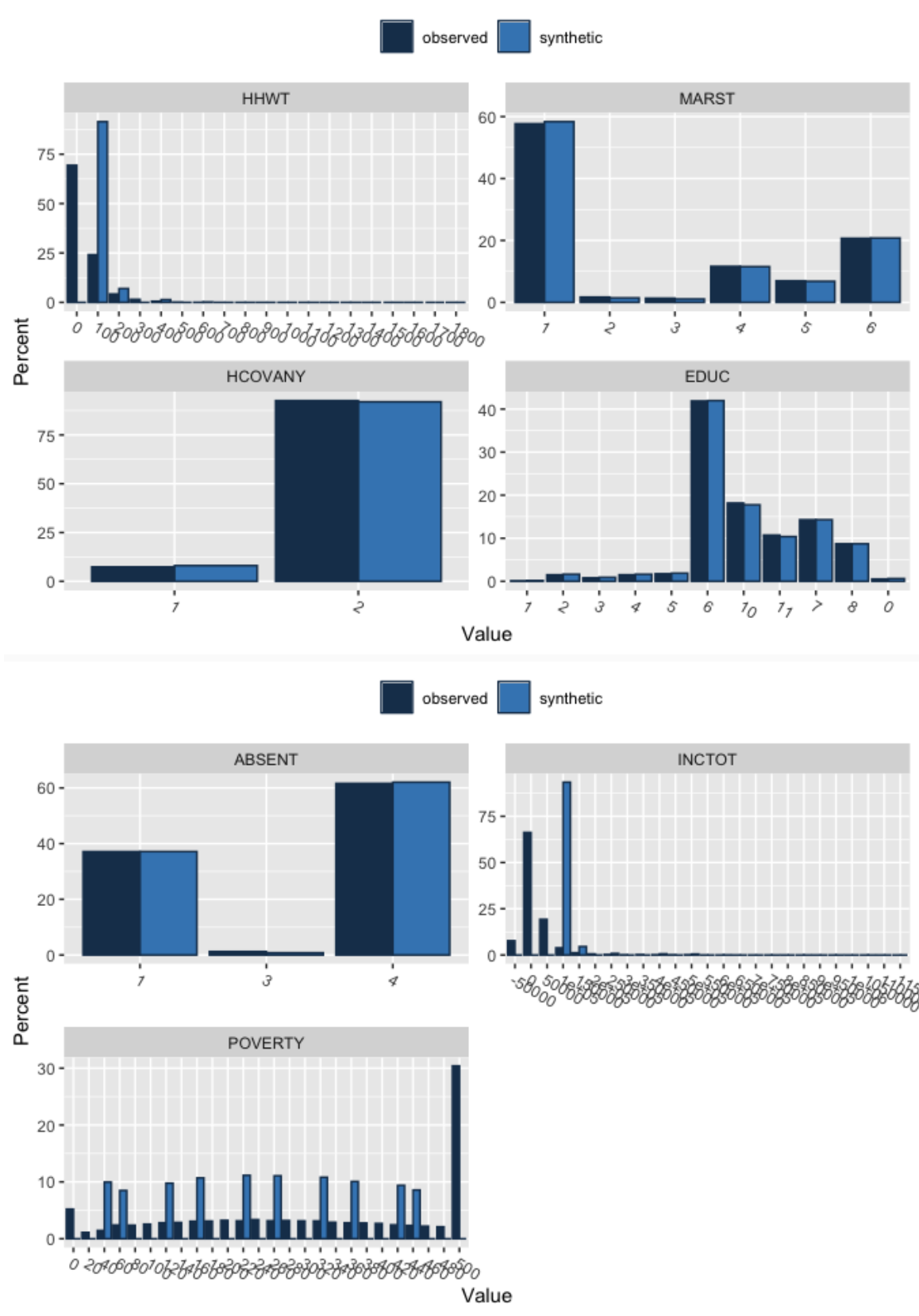
Correlation Plots for Graphical Comparison of Pearson Correlation

Synthetic Datasets should represent the dependencies of the original datasets. The following correlation plots provide an ad-hoc overview on the Pearson correlations of the original and synthetic dataset. The left plot shows the original correlation whereas the right plot provides the correlation based on the synthetic dataset.
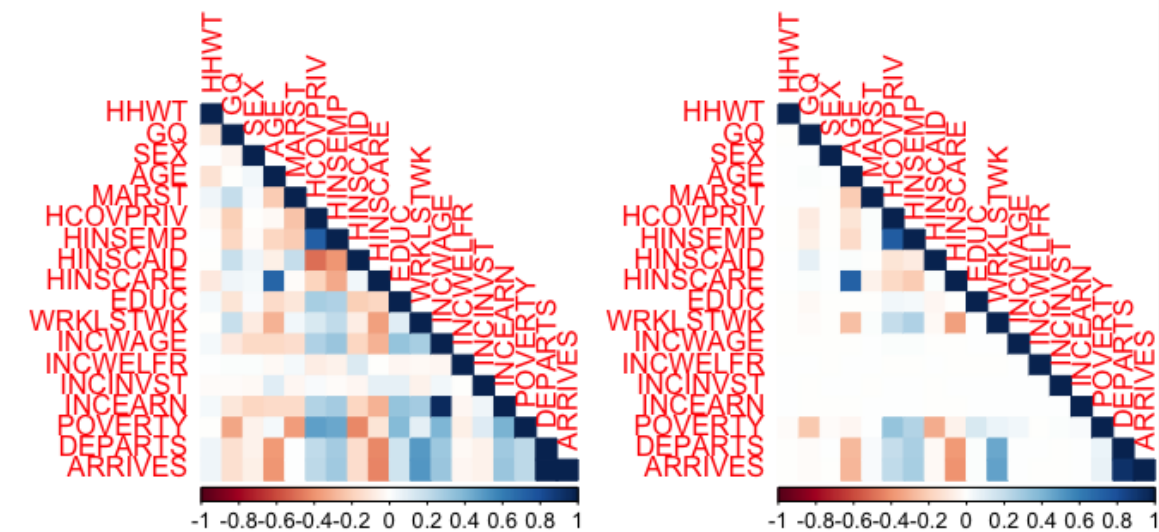
Distributional Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE

Propensity scores are calculated on a combined dataset (original and synthetic). A model (here: CART) tries to identify the synthetic units in the dataset. Since both datasets should be identically structured, the pMSE should equal zero. The S_pMSE (standardised pMSE) should not exceed 10 and for a good fit below 3 according to Raab (2021, https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Raab_AD.pdf)

	pMSE	S_pMSE	df
SEX	0.0000000	5.102196e-01	1
CITIZEN	0.0001703	9.244680e+02	3
HINSCAID	0.0000276	4.492361e+02	1
LABFORCE	0.0000001	2.177926e+00	1
WRKRECAL	0.0001435	1.168322e+03	2
INCINVST	0.2395194	1.950079e+06	2

pMSE	S_pMSE
0.2484669	1626.453

	pMSE	S_pMSE	df
YEAR	0.0000001	3.018637e-01	6
AGE	0.0077622	3.159859e+04	4

	pMSE	S_pMSE	df
SPEAKENG	0.0001604	6.530851e+02	4
HINSCARE	0.0000045	7.382504e+01	1
WRKLSTWK	0.0000150	1.223837e+02	2
WORKEDYR	0.0000086	7.041004e+01	2
INCEARN	0.2106121	1.143151e+06	3

pMSE	S_pMSE
0.2485162	384.5451

	pMSE	S_pMSE	df
HHWT	0.2134683	1158653.3496	3
MARST	0.0000687	223.8566	5
HCOVANY	0.0000342	556.6046	1
EDUC	0.0000639	103.9779	10
ABSENT	0.0001302	1060.1785	2
INCTOT	0.2211937	1200585.2797	3
POVERTY	0.0288527	117453.8627	4

pMSE	S_pMSE
0.25	257.326

	pMSE	S_pMSE	df
GQ	0.0001484	604.0495	4
RACE	0.0001179	239.9770	8
HCOVPRIV	0.0000063	101.7801	1
EMPSTAT	0.0000283	230.0871	2
LOOKING	0.0000340	276.9484	2
INCWAGE	0.1710486	928409.9725	3
DEPARTS	0.1003127	544472.9242	3

pMSE	S_pMSE
0.2487321	608.7562

Two-way Tables Comparison of Synthesised Data (R-Package: synthpop) by (S_)pMSE
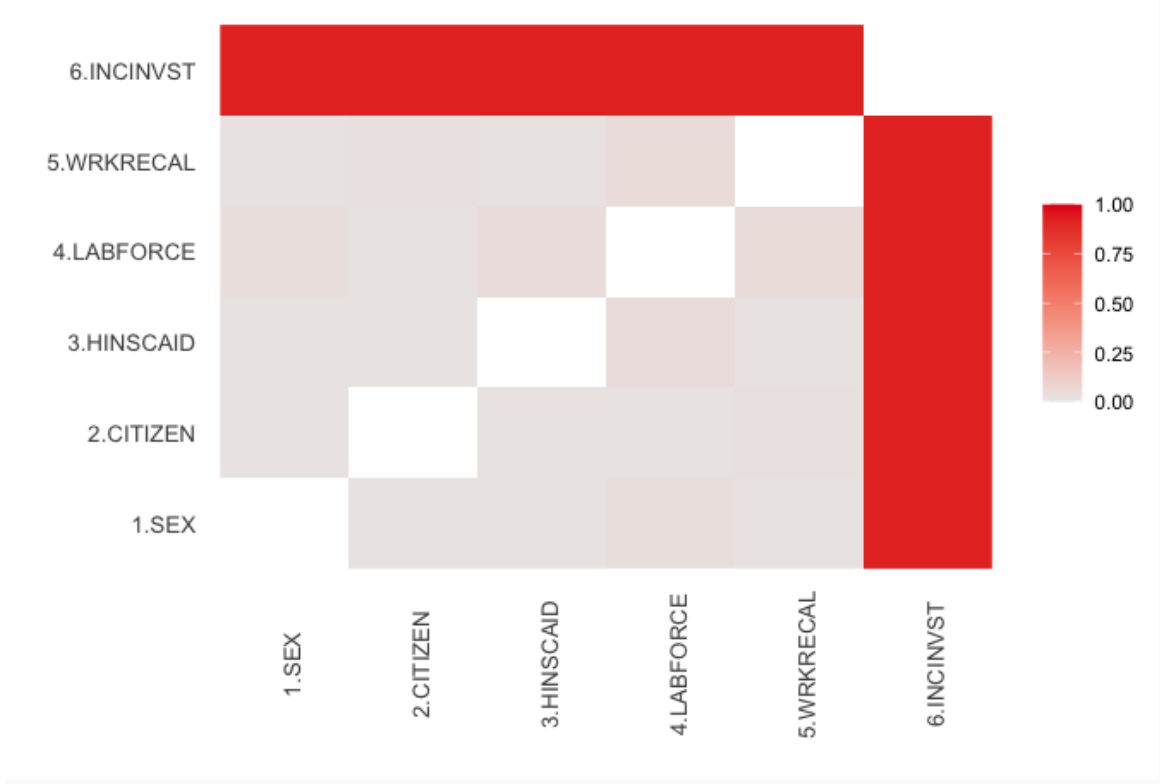
Two-way tables are evaluated based on the original and the synthetic dataset based on S_pMSE (see above). We also present the results for the mean absolute difference in densities (MabsDD) and the Bhattacharyya distance (dBhatt).
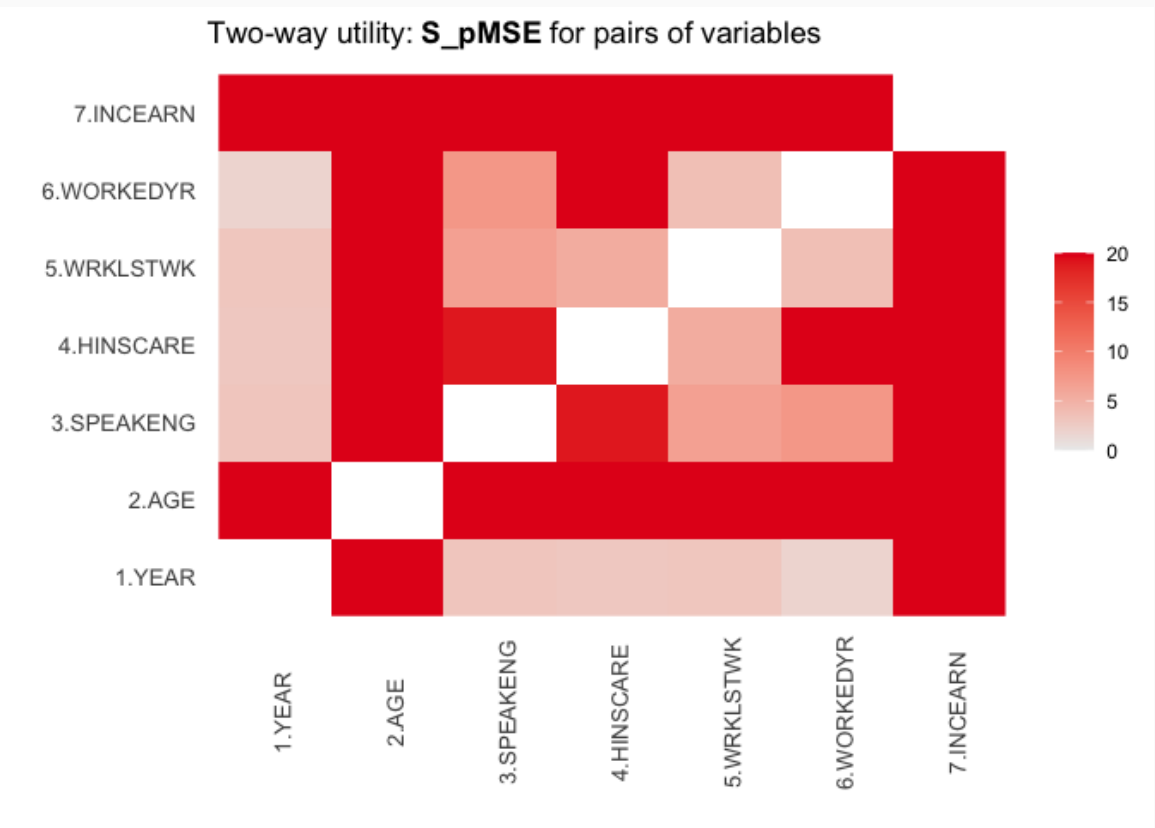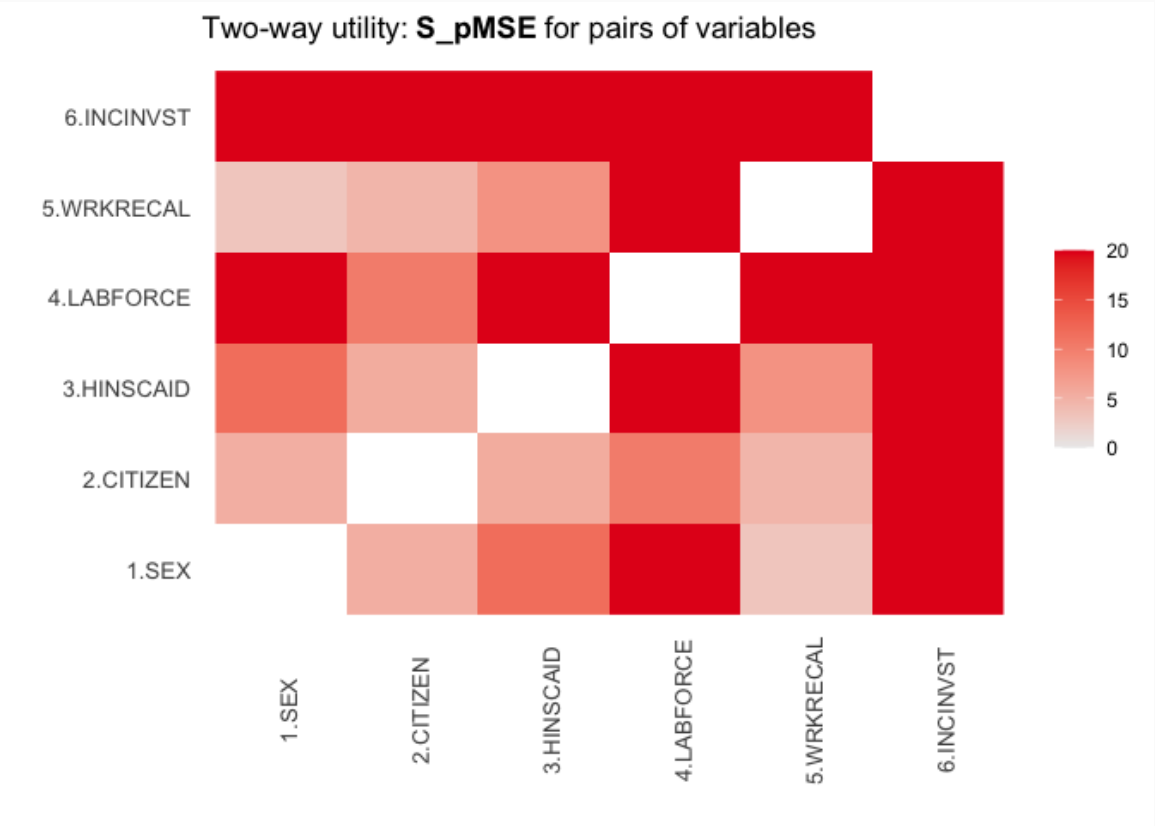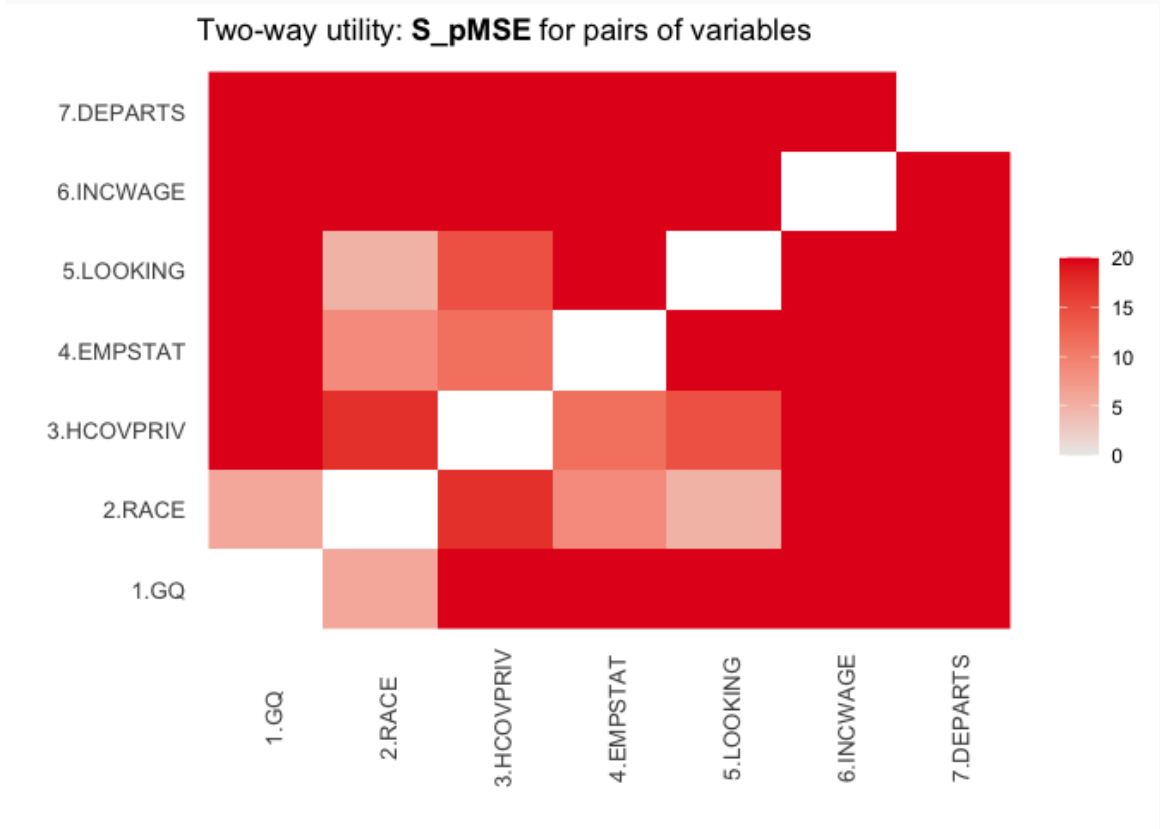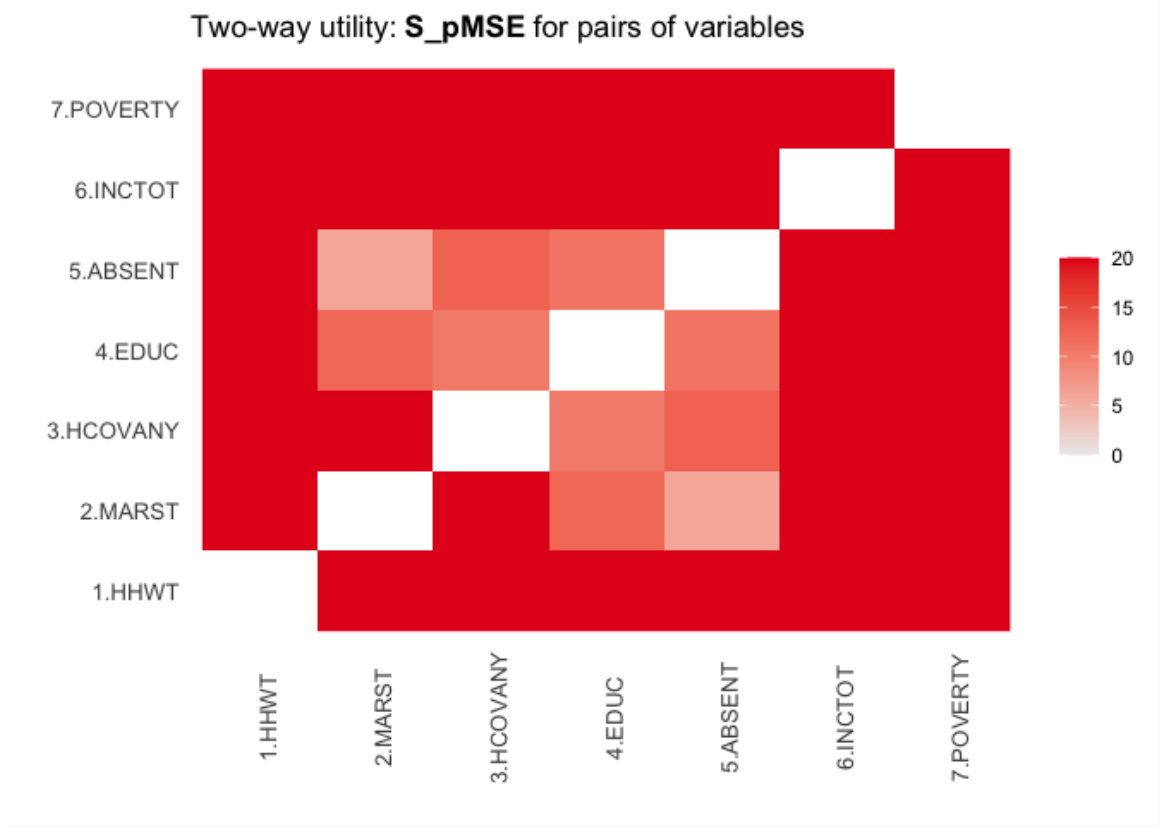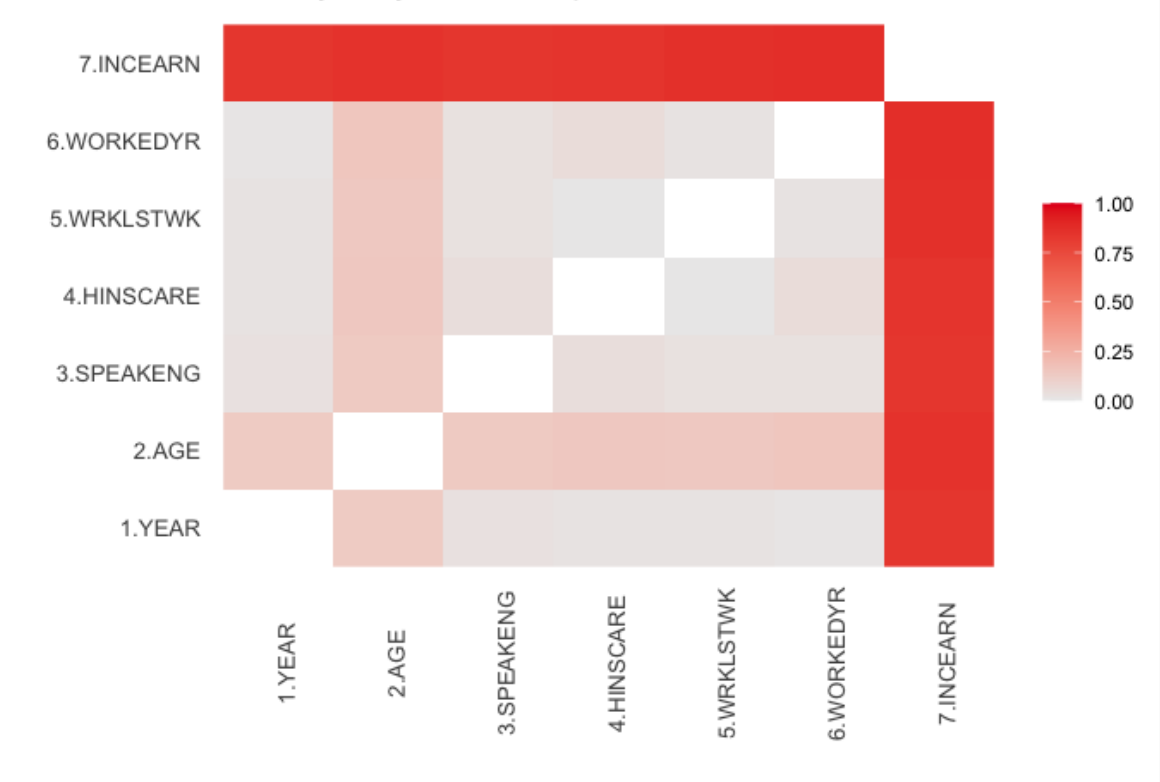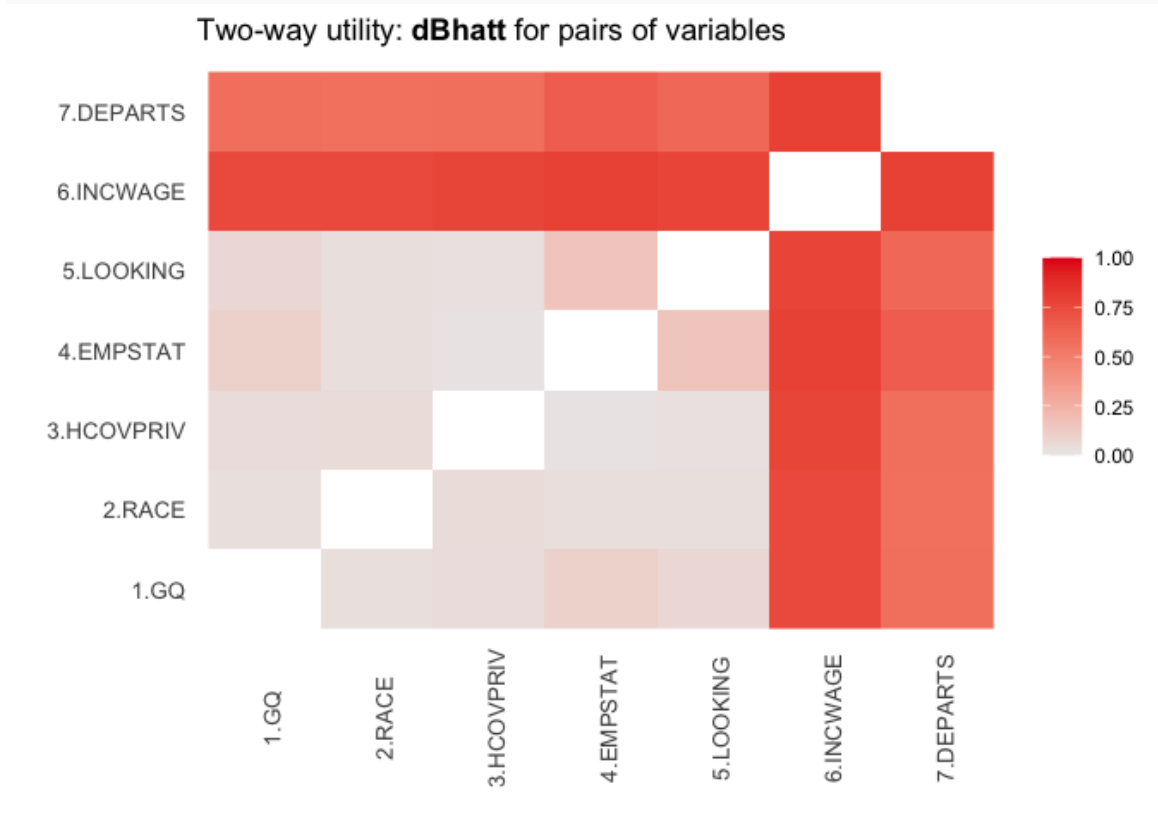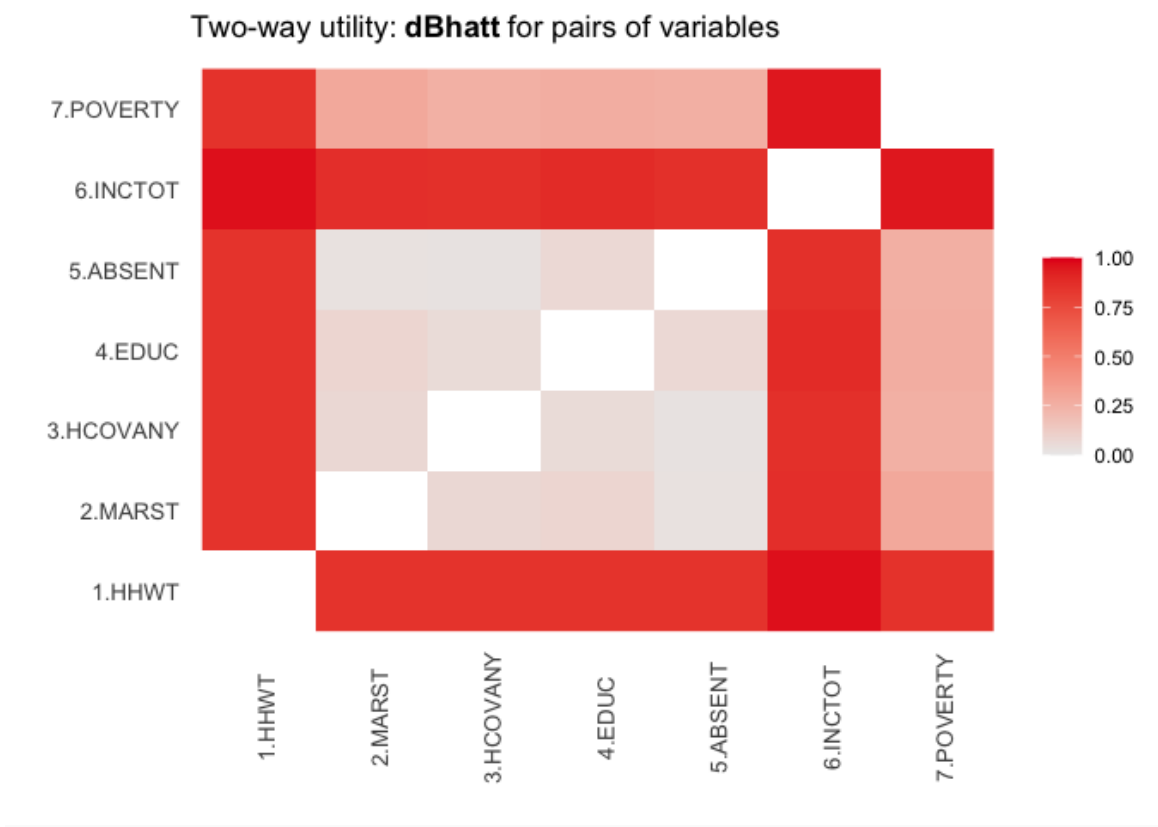

Two-way utility: **S_pMSE** for pairs of variables

Two-way utility: **S_pMSE** for pairs of variables

Two-way utility: **S_pMSE** for pairs of variables

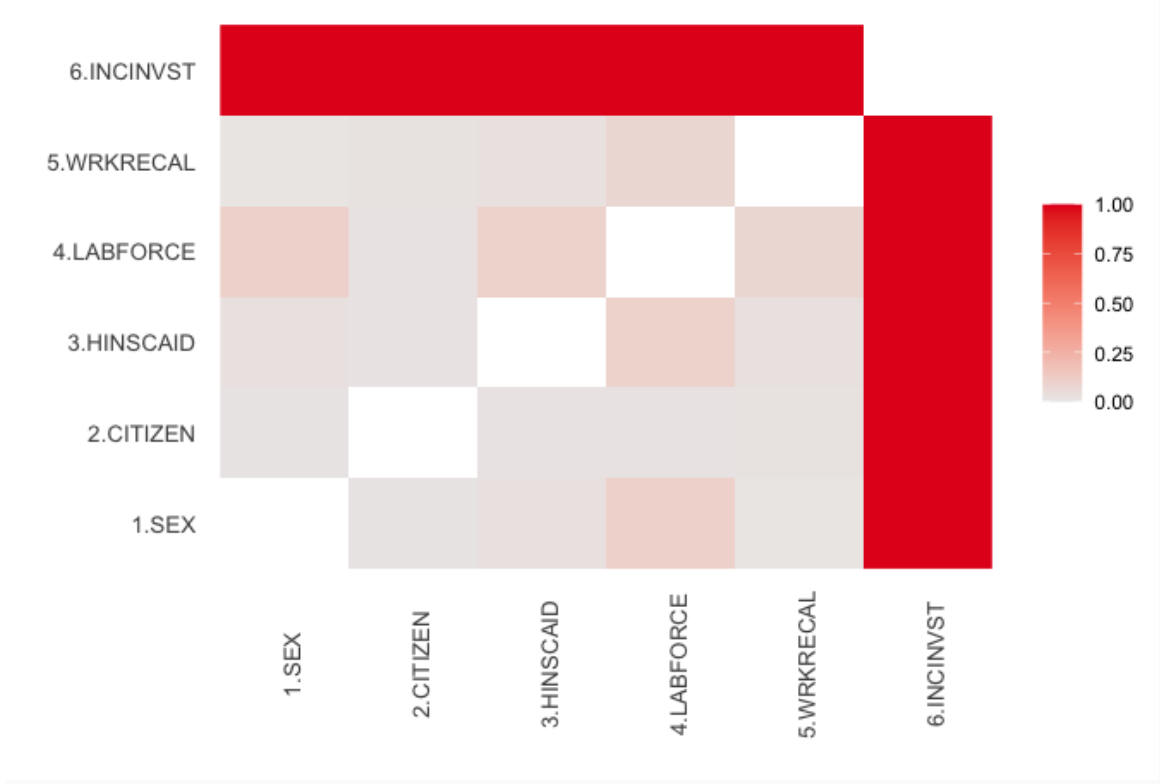## Two-way utility: **S_pMSE** for pairs of variables


```
## NULL
```

Two-way utility: **dBhatt** for pairs of variables

Two-way utility: **dBhatt** for pairs of variables

Two-way utility: **dBhatt** for pairs of variables

Two-way utility: **dBhatt** for pairs of variables


```
## NULL
```

Two-way utility: **MabsDD** for pairs of variables

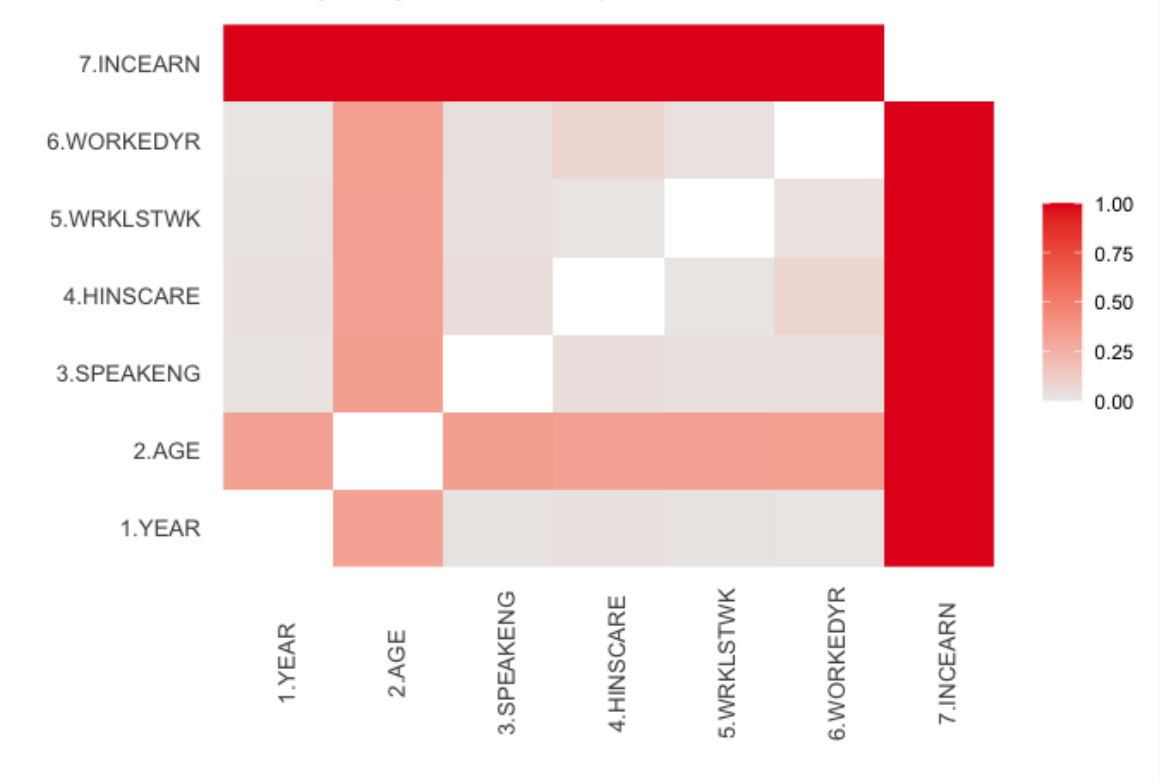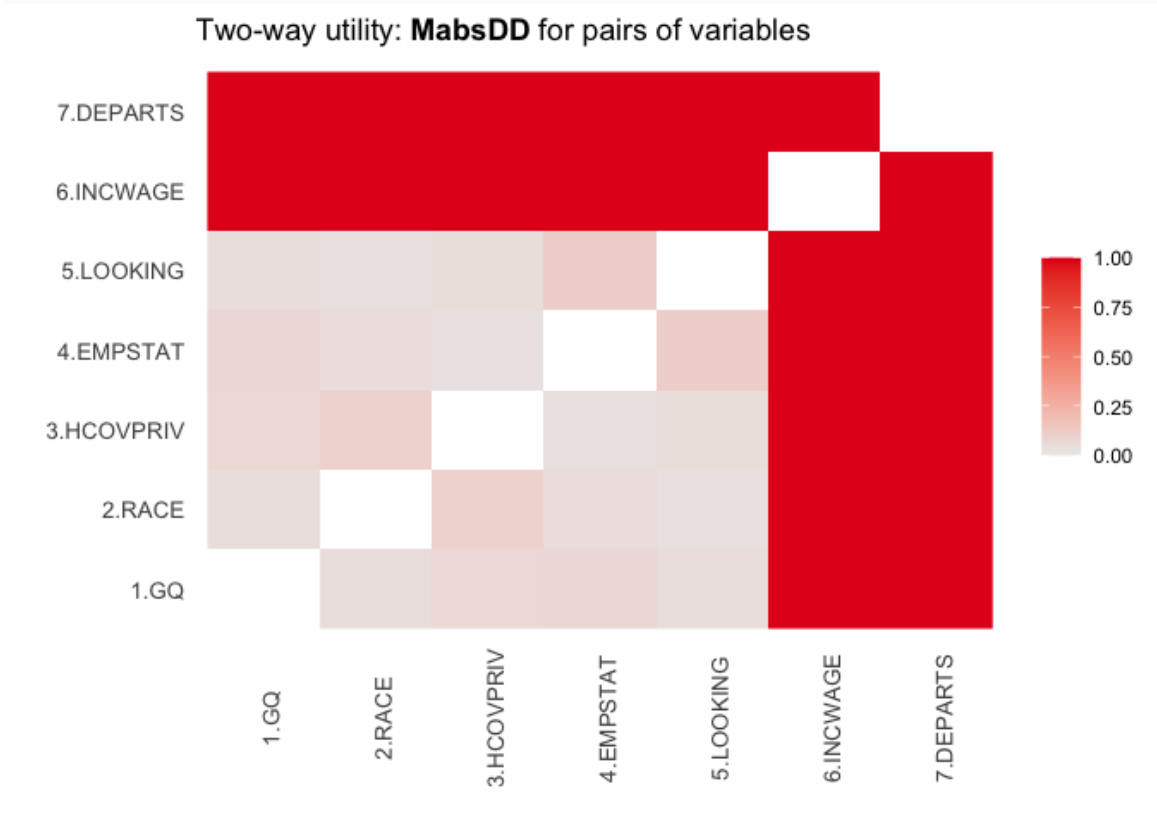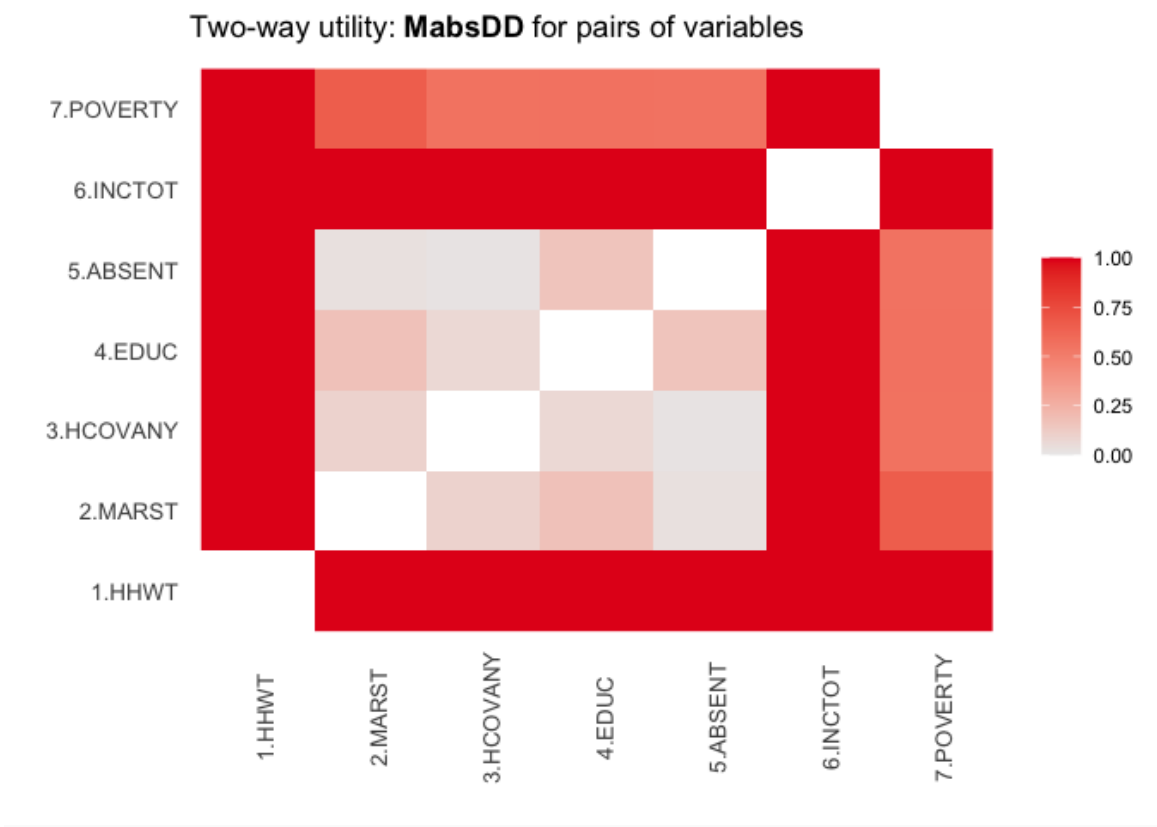## Two-way utility: **MabsDD** for pairs of variables

Two-way utility: **MabsDD** for pairs of variables

Two-way utility: **MabsDD** for pairs of variables


```
## NULL
```

Information Loss Measure Proposed by Andrzej Mlodak (R-Package: sdcMicro)

The value of this information loss criterion is between 0 (no information loss) and 1. It is calculated overall and for each variable.

Information.Loss
0.5775517

Individual Distances for Information Loss:

```
##       YEAR       HHWT         GQ       PERWT        SEX        AGE       MARST
## 0.85665871 0.98917712 0.07539913 0.98600258 0.49894665 0.90964133 0.60246398
##       RACE      HISPAN    CITIZEN    SPEAKENG     HCOVANY   HCOVPRIV    HINSEMP
## 0.23887889 0.06500515 0.10625192 0.13442514 0.14261023 0.38777582 0.47623572
##    HINSCAID   HINSCARE       EDUC     EMPSTAT    EMPSTATD   LABFORCE   WRKLSTWK
## 0.23028798 0.39041018 0.75218236 0.50345287 0.51548387 0.46424501 0.54633273
##     ABSENT     LOOKING    AVAILBLE    WRKRECAL    WORKEDYR     INCTOT    INCWAGE
## 0.47846918 0.49724576 0.18486220 0.12909942 0.49744032 0.99998358 0.99995903
##    INCWELFR    INCINVST     INCEARN     POVERTY     DEPARTS    ARRIVES
## 0.99961495 0.99994939 0.99997459 0.98751727 0.99624002 0.99453489
```