Study over diabetic inpatients readmission rates

Data Mining Course

26th October, 2018

Marc Badia Romero, Aleix Balletbó Gregorio, Bernat Gené Skrabec,
Víctor Giménez Ábalos, Guillem Ferrer Nicolás, Daniel Tarrés Amselem

Outline

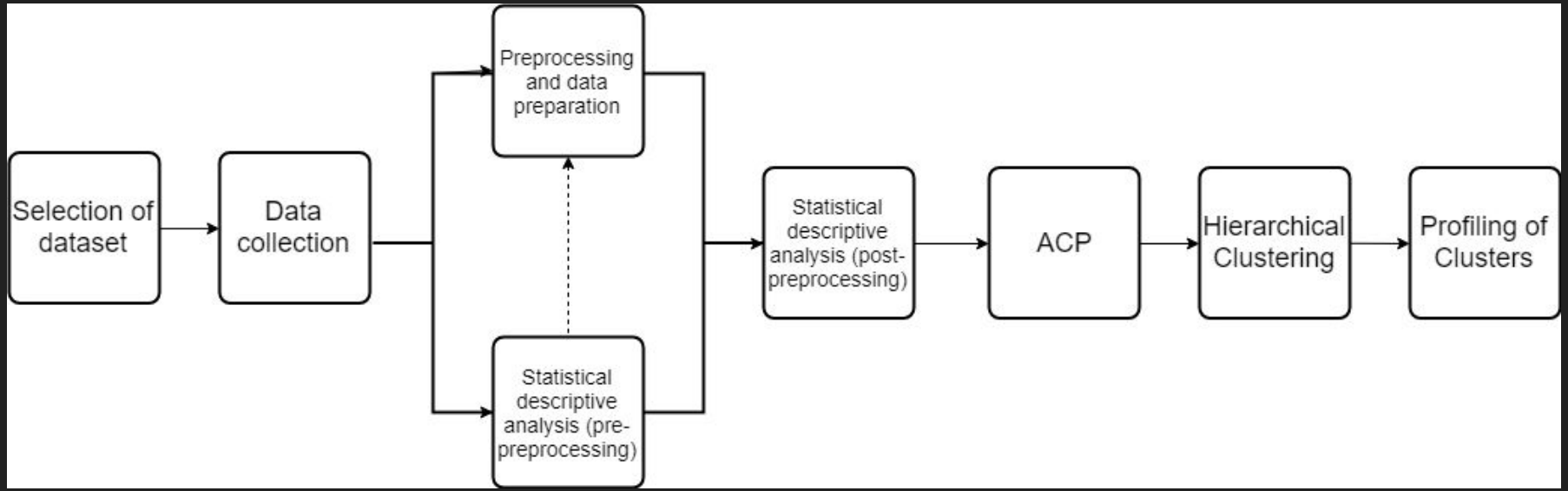
- Goals and topics addressed
- Data mining process schema
- Descriptive analysis
- Preprocessing
- ACP
- Clustering
- Profiling
- Conclusions
- Scheduling

Topics, goals, overview and variables

- Topics addressed
  - Diabetes
  - Readmission rates of diabetes inpatients
- Goals of the work
  - Discover wrong practices in established diabetes protocols
  - Predict readmission rates
- Urls from data sources
  - https://www.kaggle.com/brandao/diabetes/home
- Overview of DB
  - size ≈ 100.000 rows
  - 50 variables
  - Low missing rate (3.79%)
    - Weight (97%)
    - payer_code (40%)
    - medical_speciality (50%)

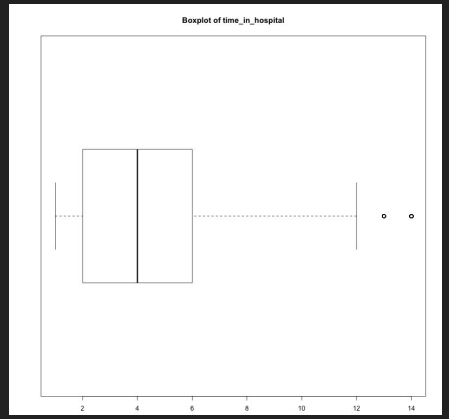
5. Variables analyzed
   - race, gender, age, weight, adm_type_id, disch_id, adm_source_id, time_in_hpt, payer_code, specialty, n_lab_proc, n_proc, n_med, n_outp, n_emerg, n_inp, , diag_1, diag_2, diag_3, n_diag, A1Cresult, metformin, insulin, change, diabetesMed, readmitted, other_meds

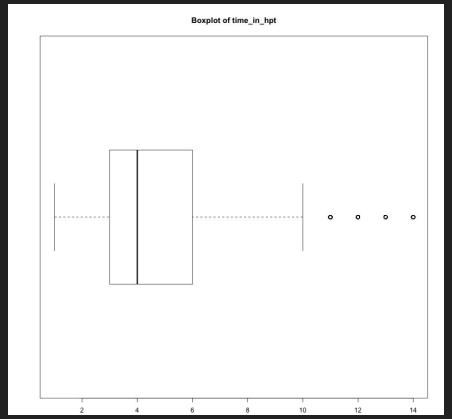
Data mining process schema



Statistical descriptive analysis

Time in Hospital (Numerical)

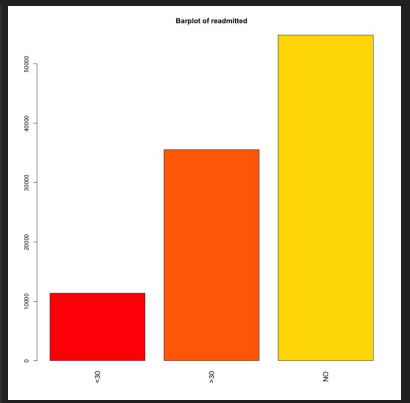


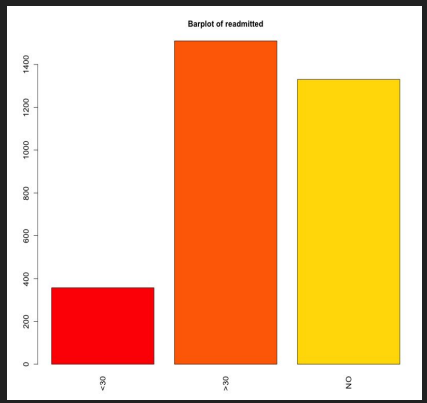


Raw

Preprocessed

Readmitted (Qualitative)





Category	Individuals, Raw	Individuals, Pprcsed.	Percentage, Raw	Percentage, Pprcsed.
<30	11357	357	11.16%	11.17%
>30	35545	1510	34.93%	47.23%
No	54864	1330	54.91%	41.60%

Synthesis of univariate descriptive analysis

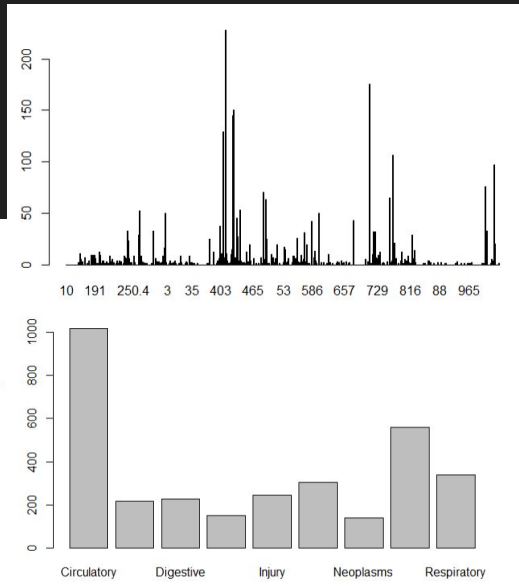
- For qualitative variables:
  - Barplot
  - Proportion
- For numerical variables:
  - Boxplot
  - Mean, Median, SD, Min and Max values
- Subset selection affected the shape of the variables
- After analysing, we can say about our patients:
  - Predominantly old
  - About normal weight for U.S. (80 kg)
  - Predominantly caucasian

Preprocessing

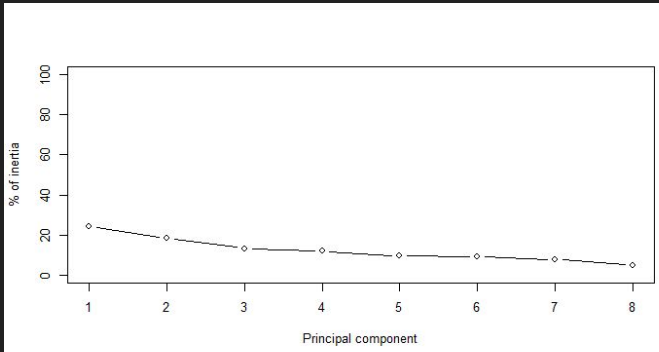
- Input missing values
- Drop every observation with missing weight (95 % of the dataset)
- Reduce variable names

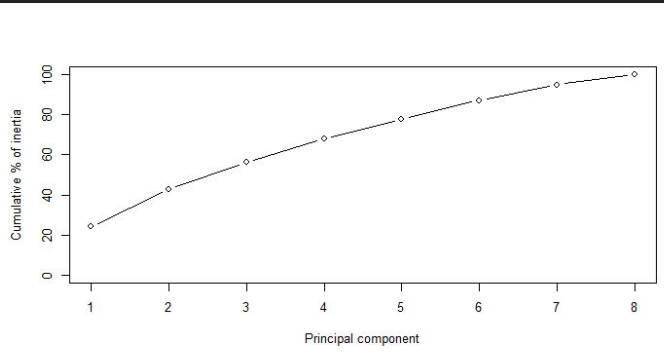
- Collapsing medicines (22 vars to 1)
- Generalizing diagnoses to coarser granularity (specific id to category)

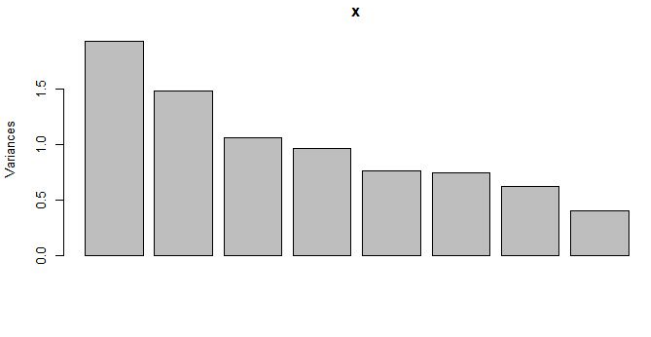
```
Previous data: [1] "encounter_id"            "patient_nbr"             "race"                    "gender"                  "age"
 [6] "weight"                "admission_type_id"       "discharge_disposition_id" "admission_source_id"     "time_in_hospital"
[11] "payer_code"            "medical_specialty"       "num_lab_procedures"       "num_procedures"          "num_medications"
[16] "number_outpatient"     "number_emergency"        "number_inpatient"         "diag_1"                  "diag_2"
[21] "diag_3"                "number_diagnoses"        "max_glu_serum"            "A1Cresult"               "metformin"
[26] "repaglinide"           "nateglinide"             "chlorpropamide"           "glimepiride"             "acetohexamide"
[31] "glipizide"             "glyburide"               "tolbutamide"              "pioglitazone"            "rosiglitazone"
[36] "acarbose"              "miglitol"                "troglitazone"             "tolazamide"              "examide"
[41] "citoglipton"           "insulin"                 "glyburide.metformin"      "glipizide.metformin"     "glimepiride.pioglitazone"
[46] "metformin.rosiglitazone" "metformin.pioglitazone" "change"                  "diabetesMed"             "readmitted"
[51] "other_meds"

New data:
 [1] "encounter_id"          "patient_nbr"             "race"                    "gender"                  "age"
 [6] "weight"                "admission_type_id"       "discharge_disposition_id" "admission_source_id"     "time_in_hospital"
[11] "payer_code"            "medical_specialty"       "num_lab_procedures"       "num_procedures"          "num_medications"
[16] "number_outpatient"     "number_emergency"        "number_inpatient"         "diag_1"                  "diag_2"
[21] "diag_3"                "number_diagnoses"        "A1Cresult"                "metformin"               "insulin"
[26] "change"                "diabetesMed"             "readmitted"               "other_meds"
```

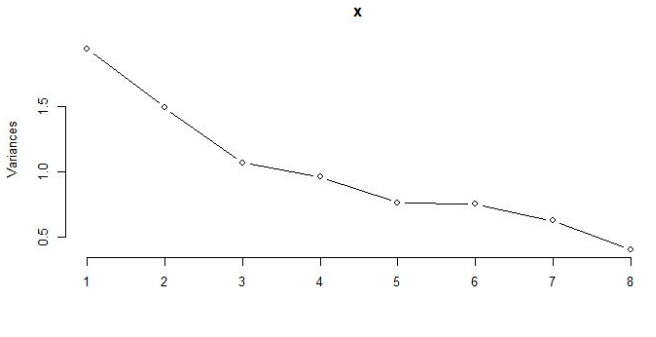


ACP specifications









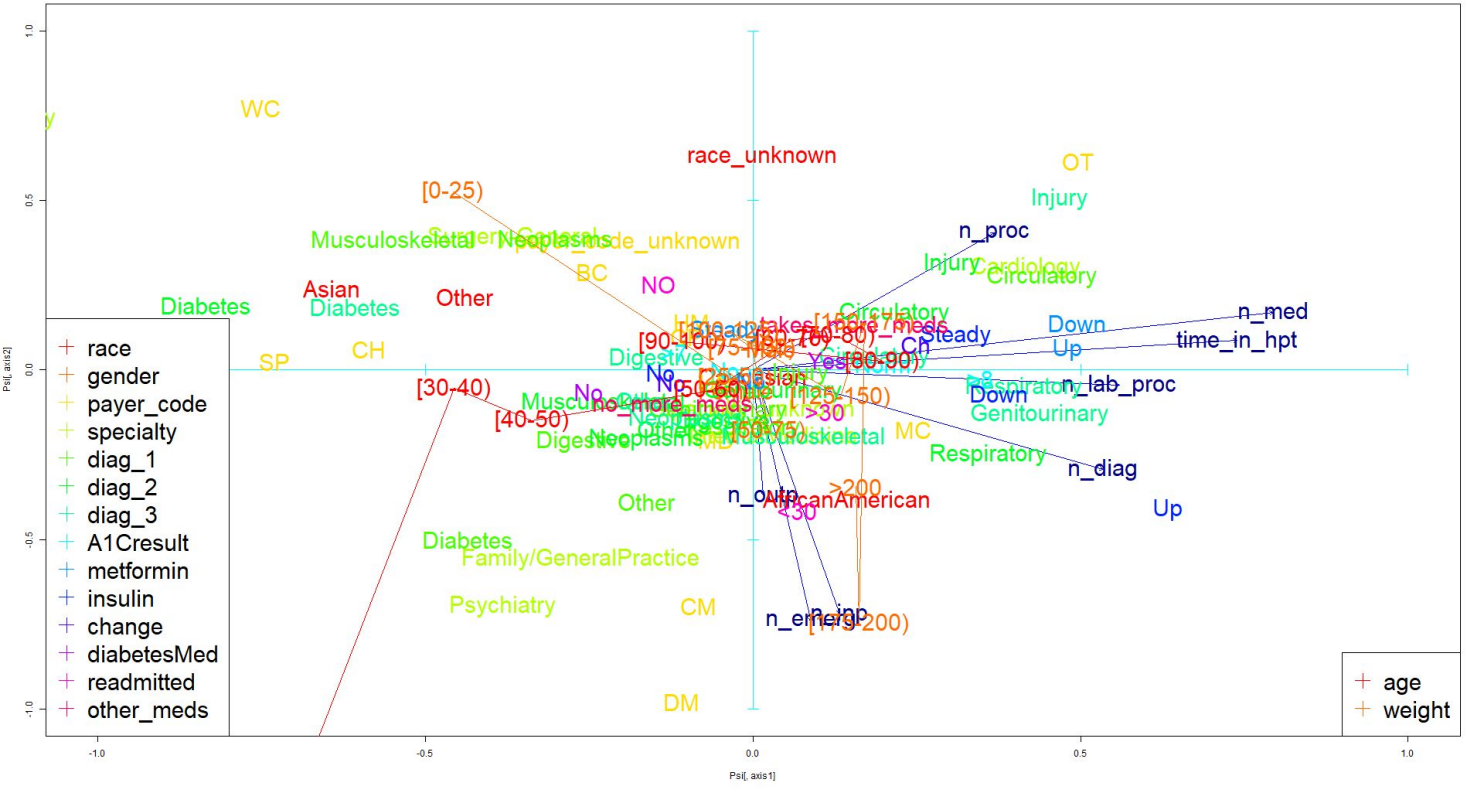
5 vs 6 principal components:

77.72693

87.12418

We will use 6

First ACP plane



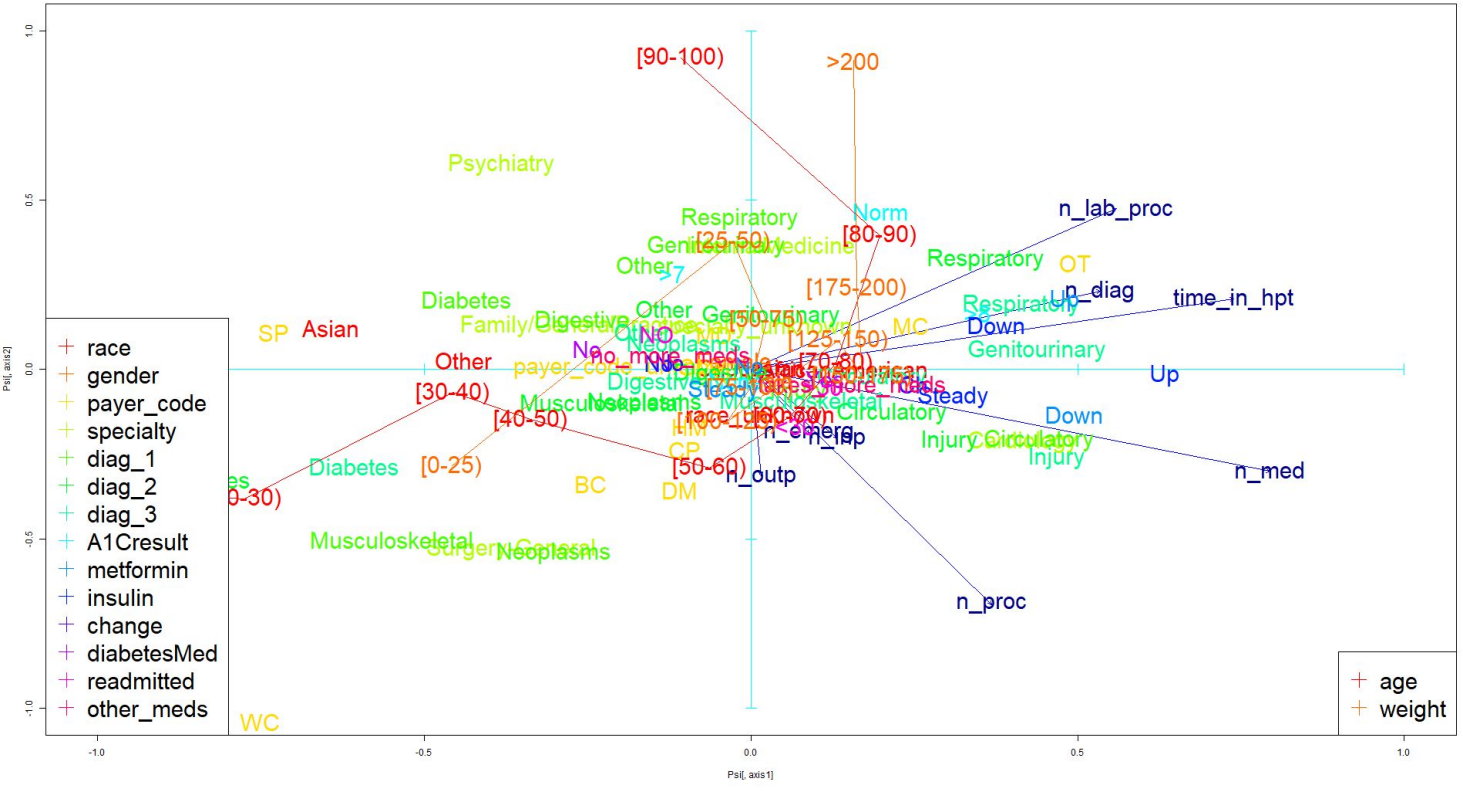
‘Gravity’ vs ‘Reincidence’

Reincidence to the right and down.

Weight linked to readmission.

Rise in insulin for respiratory problems and high number of diagnoses.

Second ACP plane



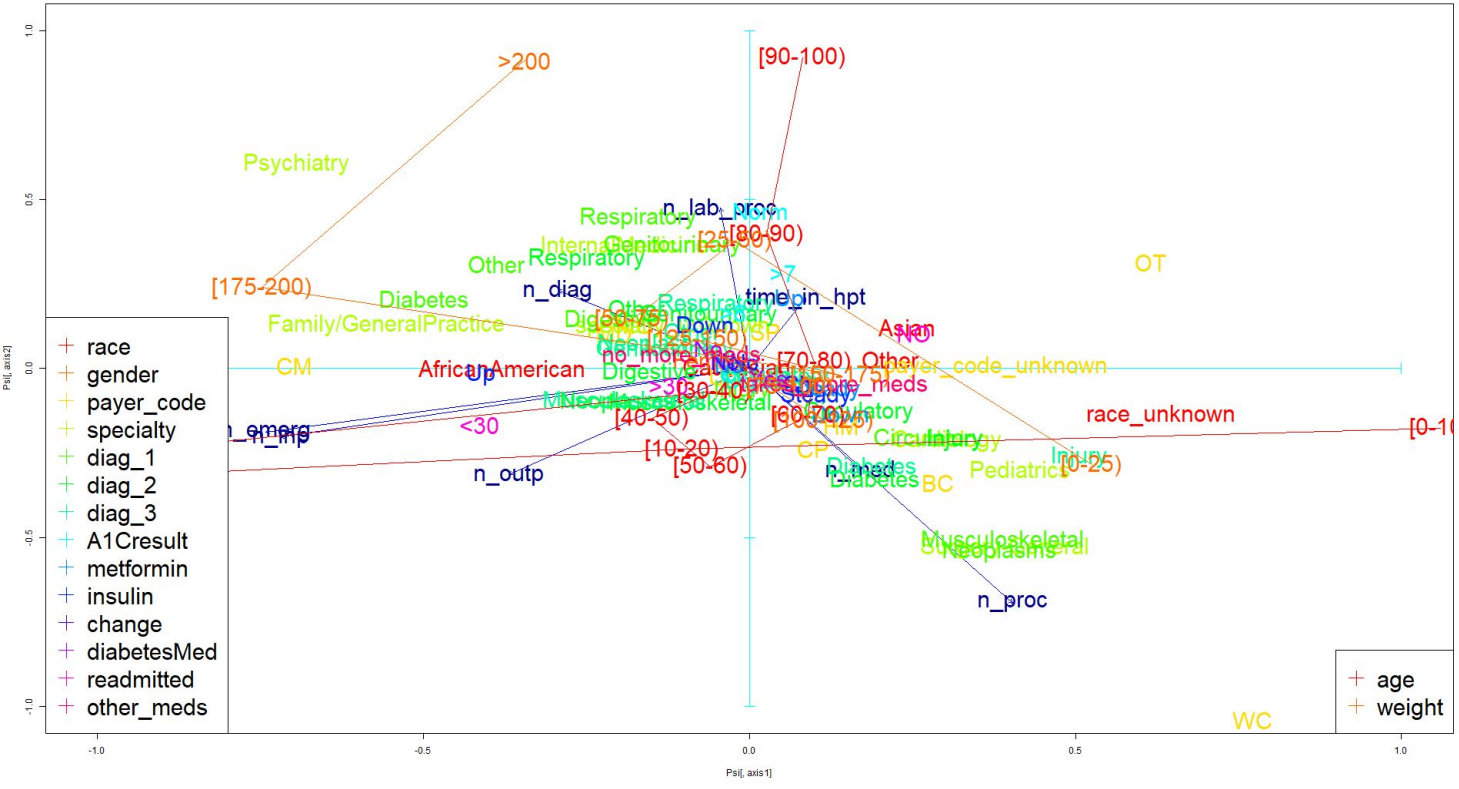
‘Gravity’ vs ‘Riskiness’

Not taking more medications linked to low chance of readmission.

Taking more medications linked to higher chance of readmission.

Rise in metformin and reduction in insulin for respiratory problems and high number of diagnoses.

Third ACP plane



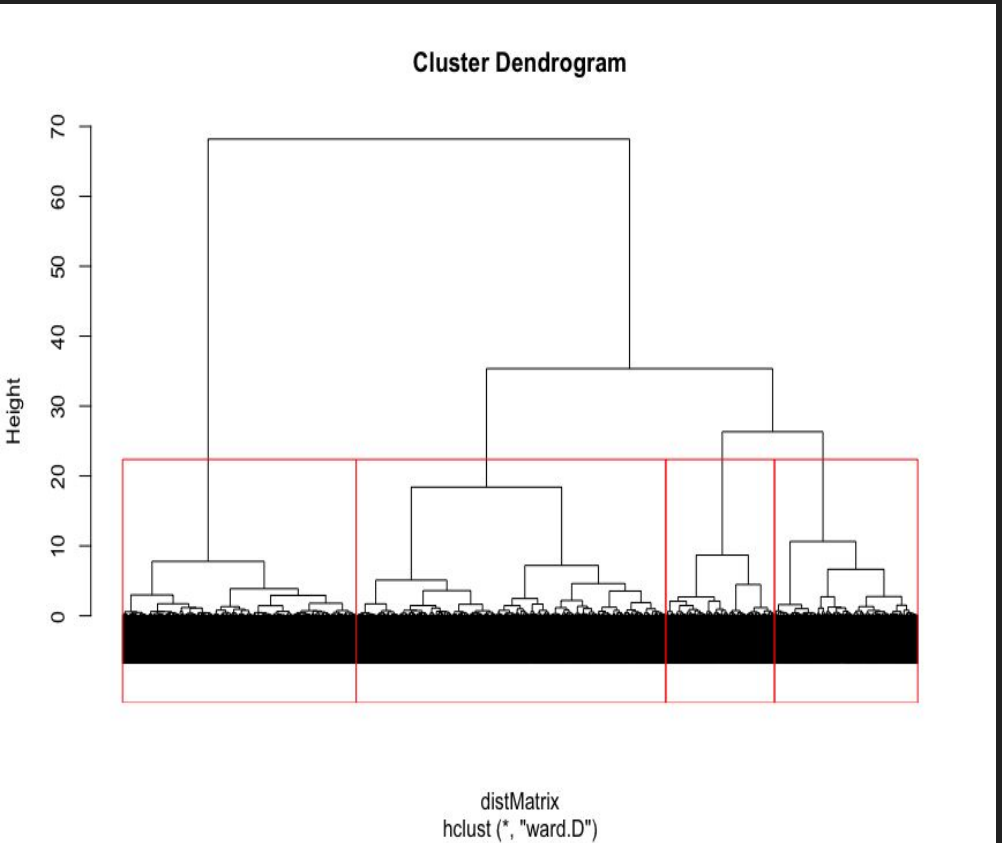
'Reincidence' vs 'Riskiness'

Diagnosis of diabetic problems strongly linked to number of medications and procedures during the encounter.

ACP conclusions

- Having a higher number of visits during the past year is strongly correlated to an earlier readmission.
- Weight linked to readmission.
- Taking more or less medication correlated to being readmitted or not.
- Patients diagnosed with diabetic problems generally take more medication and go through more procedures.

Clustering

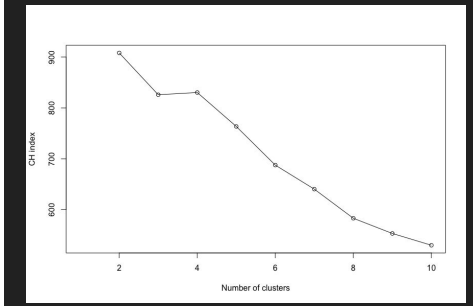


Removed identifiers and response variable

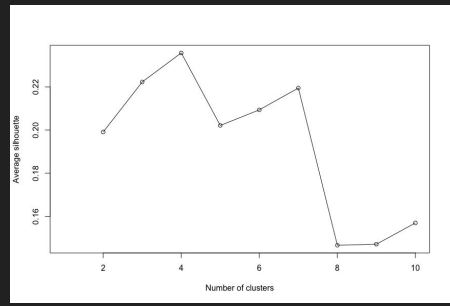
Ascendent hierarchical clustering

Gower's metric

Ward's aggregation



Calinski-Harabasz Index



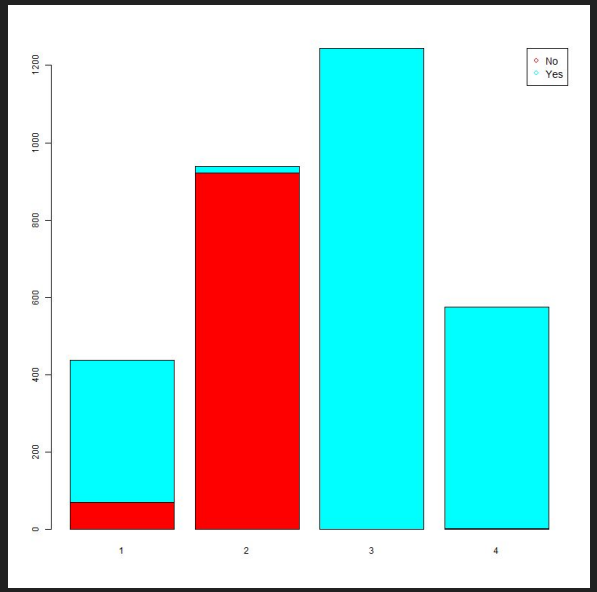
Silhouette

Tools of class interpretation

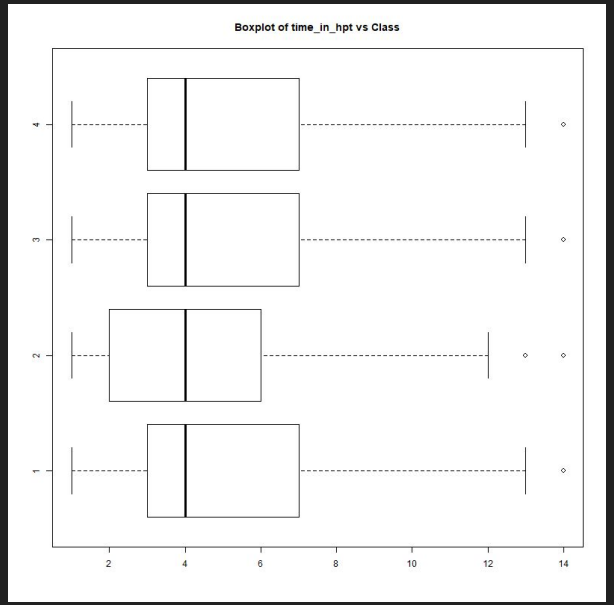
- CPG
- Numerical Information
- For categorical variables:
  - Barplot
  - Pos&Neg
- For numerical variables:
  - Boxplot
  - Barplot

Profiling graphs

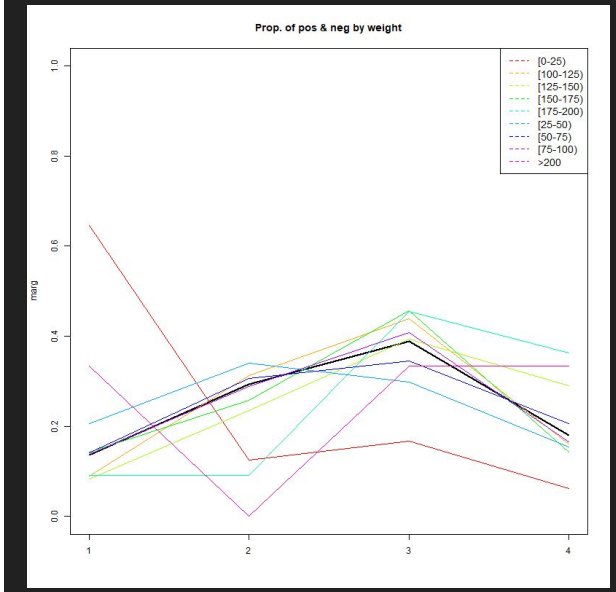
Diabetes medications



Time in hospital

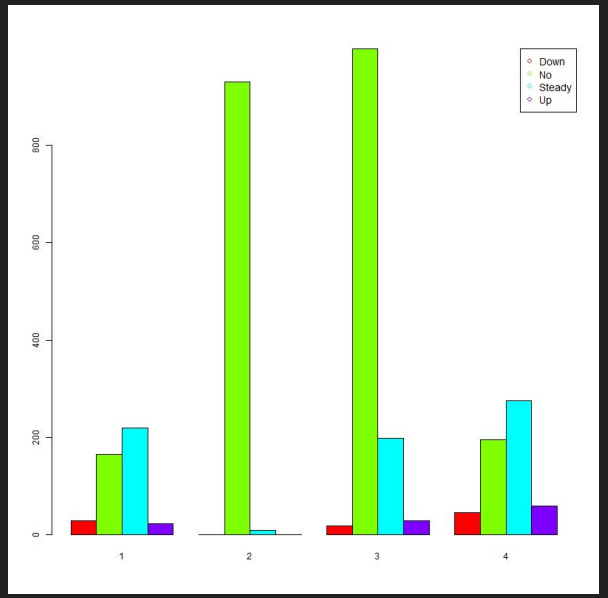


Weight

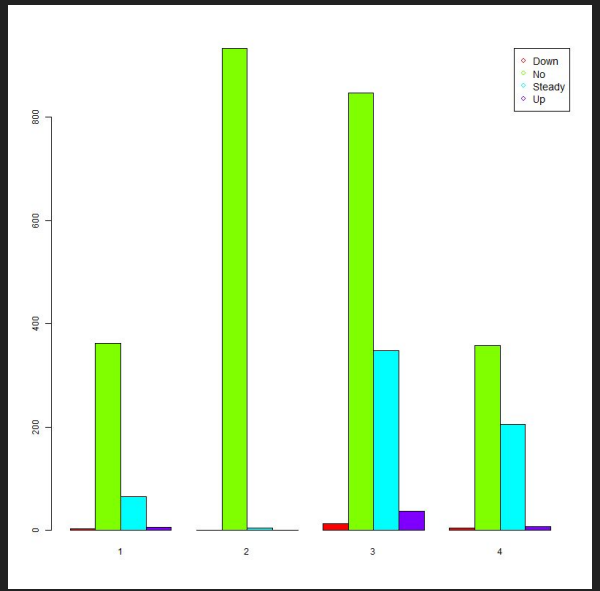


Profiling graphs

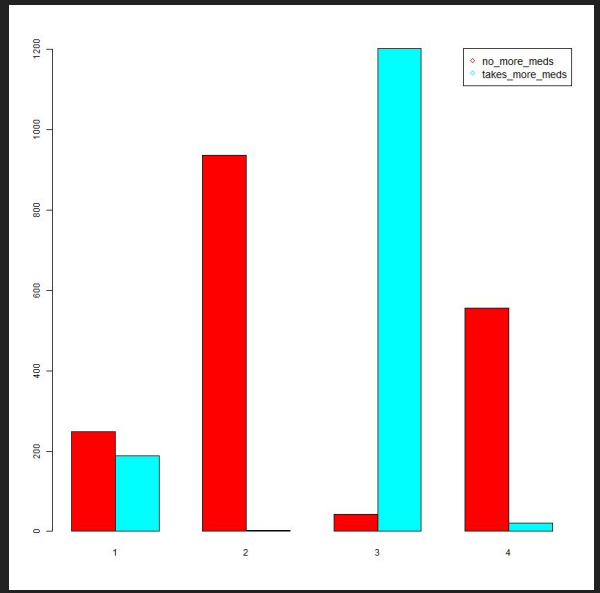
Insulin



Metformin

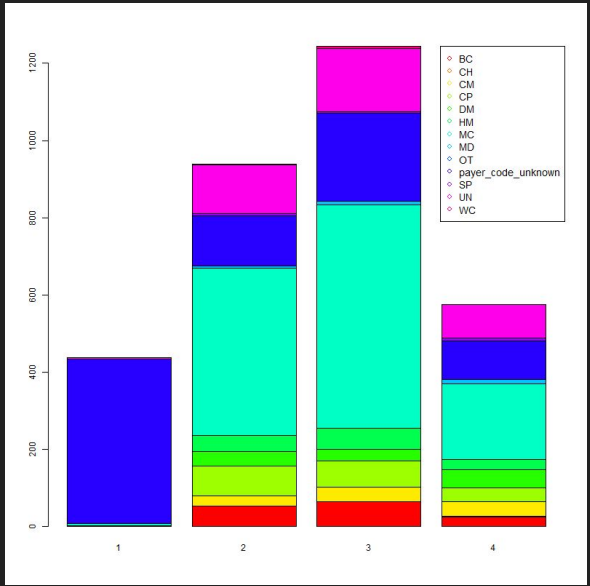


Other

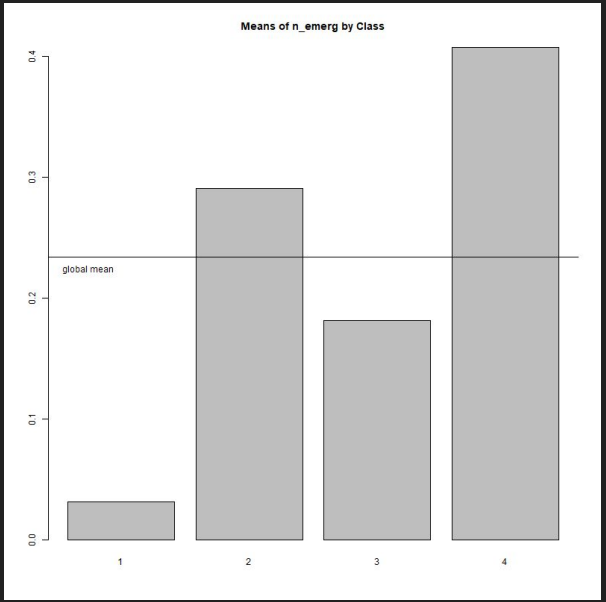


Profiling graphs

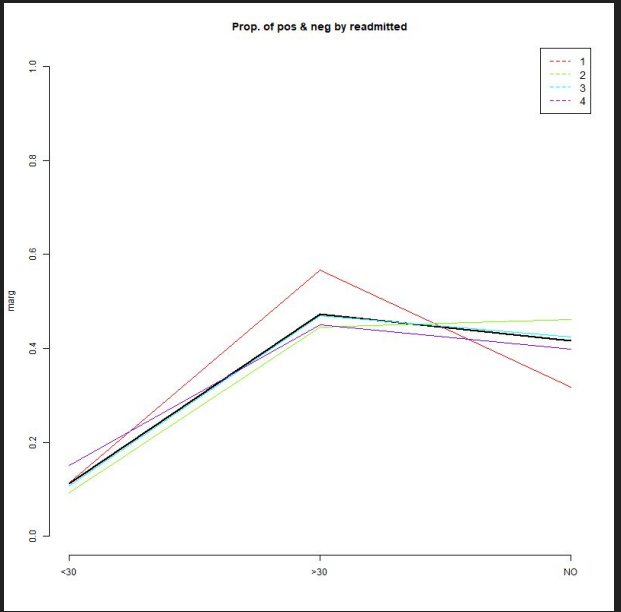
Payer code



Emergency visits



Readmitted



Profiling of Clusters

Cluster 1

- Payer code unknown
- Low nº of visits and diags
- Late readmission
- Low social status

Cluster 2

- No prescribed medication
- Less time in hospital
- Less readmission
- Patients in less severe conditions

Cluster 3

- Biggest cluster
- Predominantly old and overweight
- Heavily medicated
- Type 2 diabetes

Cluster 4

- More prop. of young people
- Medicated, but mostly insulin
- Less procedures, more readmissions
- Type 1 diabetes

Discussion ACP and Clustering

- Various things seen in ACP can be seen in the clusters:
- Cluster 1: Low weight and age linked to readmission.
- Cluster 2: Not taking medication, spending less time in the hospital and low readmission rate
- Cluster 3: Old and overweight people with changes in medication, taking diabetic medication and metformin but less insulin
- Cluster 4: Highest amount of visits, very high ratio of early readmission.

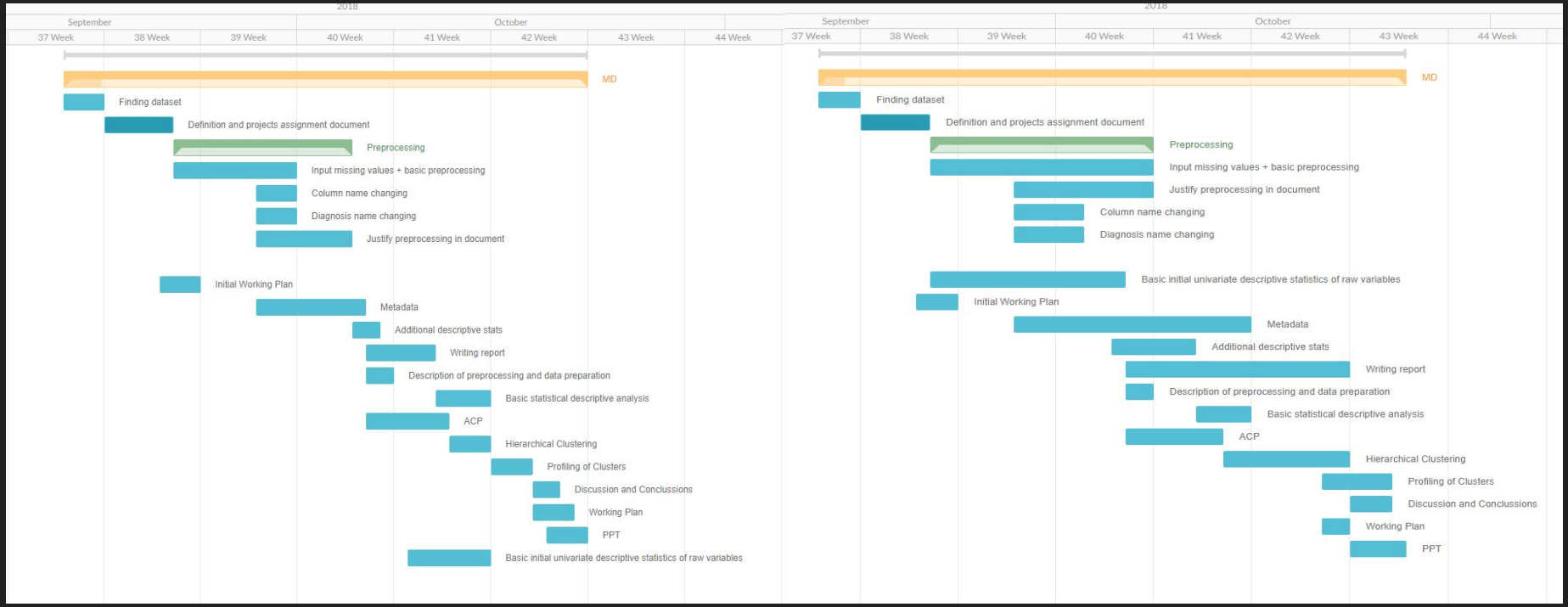
- Some others aren't seen:
- Weight being linked to readmission
- Asian patients not being readmitted
- Slightly old patients having a higher number of visits

Conclusions

- Age and weight not linked to readmission (only in profiling)
- Number of visits, young age and diabetes type 1 heavily linked to early readmission
- Kids and lack of information about payer code linked to later readmission

- Very good synergy between ACP and profiled clusters
- Calinski-Harabasz encouraged as a metric for hierarchical clustering.

Scheduling


2018
September
October
37 Week
38 Week
39 Week
40 Week
41 Week
42 Week
43 Week
44 Week
MD
Finding dataset
Definition and projects assignment document
Preprocessing
Input missing values + basic preprocessing
Column name changing
Diagnosis name changing
Justify preprocessing in document
Initial Working Plan
Metadata
Additional descriptive stats
Writing report
Description of preprocessing and data preparation
Basic statistical descriptive analysis
ACP
Hierarchical Clustering
Profiling of Clusters
Discussion and Conclussions
Working Plan
PPT
Basic initial univariate descriptive statistics of raw variables
2018
September
October
37 Week
38 Week
39 Week
40 Week
41 Week
42 Week
43 Week
44 Week
MD
Finding dataset
Definition and projects assignment document
Preprocessing
Input missing values + basic preprocessing
Justify preprocessing in document
Column name changing
Diagnosis name changing
Basic initial univariate descriptive statistics of raw variables
Initial Working Plan
Metadata
Additional descriptive stats
Writing report
Description of preprocessing and data preparation
Basic statistical descriptive analysis
ACP
Hierarchical Clustering
Profiling of Clusters
Discussion and Conclussions
Working Plan
PPT