Resultaten & Conclusion

Jorick Baron

2022-10-03

Table 1: Codebook

Name	Fullname	Description	Type	Unit
sample_id	Sample ID	Unique string identifying each subject	string	NA
patient_cohort	Patient's Cohort	Cohort 1, previously used samples; Cohort 2, newly added samples	string	NA
sample_origin	Sample Origin	BPTB: Barts Pancreas Tissue Bank; ESP: Spanish National Cancer Research Centre; LIV: Liverpool University;UCL: University College	string	NA
age	Age	Age in years	int	years
sex	Sex	M = male, F = female	char	NA
diagnosis	Diagnosis (1=Control, 2=Benign, 3=PDAC)	1 = control, 2 = benign hepatobiliary disease; 3 = Pancreatic ductal adenocarcinoma, i.e. pancreatic cancer	int	NA
stage	Stage	For those with pancratic cancer, what stage was it? One of IA, IB, IIA, IIIB, III, IV	string	NA
benign_sample_diagnosis	Benign Samples Diagnosis	For those with a benign, non-cancerous diagnosis, what was the diagnosis?	string	NA
plasma_CA19_9	Plasma CA19-9 U/ml	Blood plasma levels of CA 19–9 monoclonal antibody that is often elevated in patients with pancreatic cancer.	float	plasma units/milliliter
creatinine	Creatinine mg/ml	Urinary biomarker of kidney function	float	mg/ml
LYVE1	LYVE1 ng/ml	Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis	float	ng/ml
REG1B	REG1B ng/ml	Urinary levels of a protein that may be associated with pancreas regeneration.	float	ng/ml
TFF1	TFF1 ng/ml	Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract	float	ng/ml
REG1A	REG1A ng/ml	Urinary levels of a protein that may be associated with pancreas regeneration.	float	ng/ml

Results

Codebook

To aid in the understanding of the data a codebook explaining every variable in the data is displayed bellow.

Loading The Data

For this project the data from https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer was used. Upon loading the data the first few data entries where checked using the head function, this revealed many NAs.

```
##   sample_id patient_cohort sample_origin age sex diagnosis stage
## 1        S1        Cohort1          BPTB  33   F         1  <NA>
## 2       S10        Cohort1          BPTB  81   F         1  <NA>
## 3      S100        Cohort2          BPTB  51   M         1  <NA>
## 4      S101        Cohort2          BPTB  61   M         1  <NA>
## 5      S102        Cohort2          BPTB  62   M         1  <NA>
## 6      S103        Cohort2          BPTB  53   M         1  <NA>
##   benign_sample_diagnosis plasma_CA19_9 creatinine      LYVE1     REG1B
## 1                    <NA>          11.7    1.83222 0.89321920  52.94884
## 2                    <NA>            NA    0.97266 2.03758500  94.46703
## 3                    <NA>           7.0    0.78039 0.14558890 102.36600
## 4                    <NA>           8.0    0.70122 0.00280488  60.57900
## 5                    <NA>           9.0    0.21489 0.00085956  65.54000
## 6                    <NA>            NA    0.84825 0.00339300  62.12600
##       TFF1    REG1A
## 1 654.2822 1262.000
## 2 209.4882  228.407
## 3 461.1410       NA
## 4 142.9500       NA
## 5  41.0880       NA
## 6  59.7930       NA
```

The ones from the stage column can be explained because only patients with cancer can have a stage. And those from benign_sample_diagnosis are there because there has to be a non-cancerous diagnosis. Reading in the source the NAs in the colums plasma_CA19_9 and REG1A are there because they were not measured in every sample, thus they contain NAs. From this the conclusion is that the data is intact and further processing can be applied.

Exploring & Analysing

When exploring the data for processing a boxplot to investigate the distribution of every numeric variable was created.

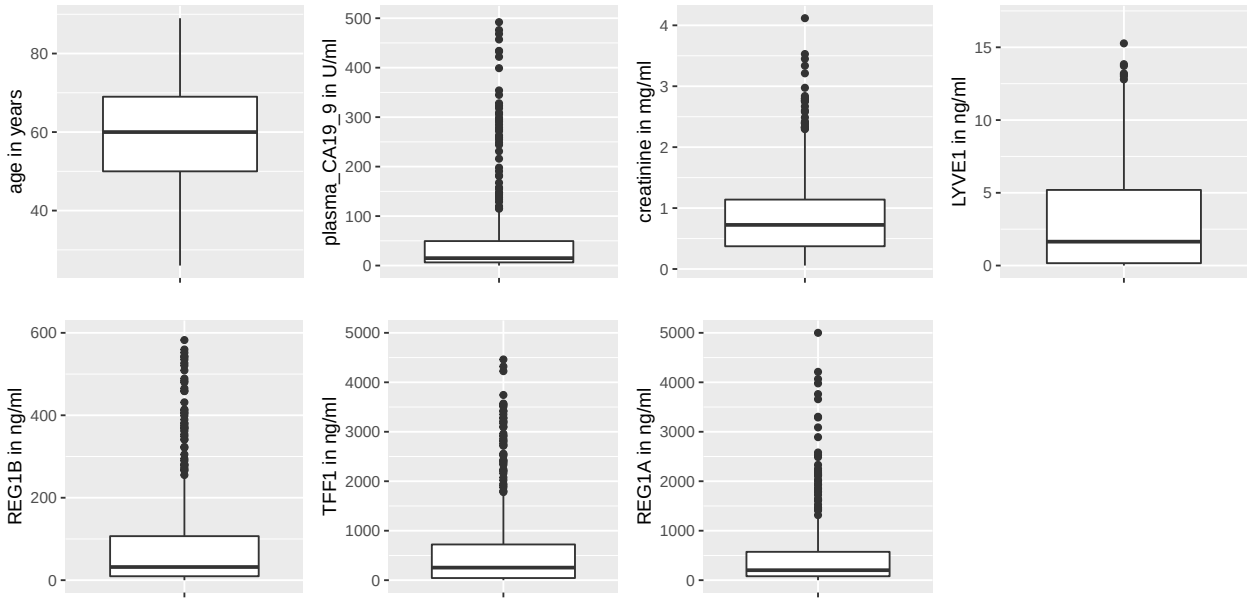

Figure 1: A boxplot displaying the distribution of all numeric values in the data

In this figure it became clear that the distributions of most variables seem highly skewed, to investigate this another boxplot was made to see if this was due to the influence of the label

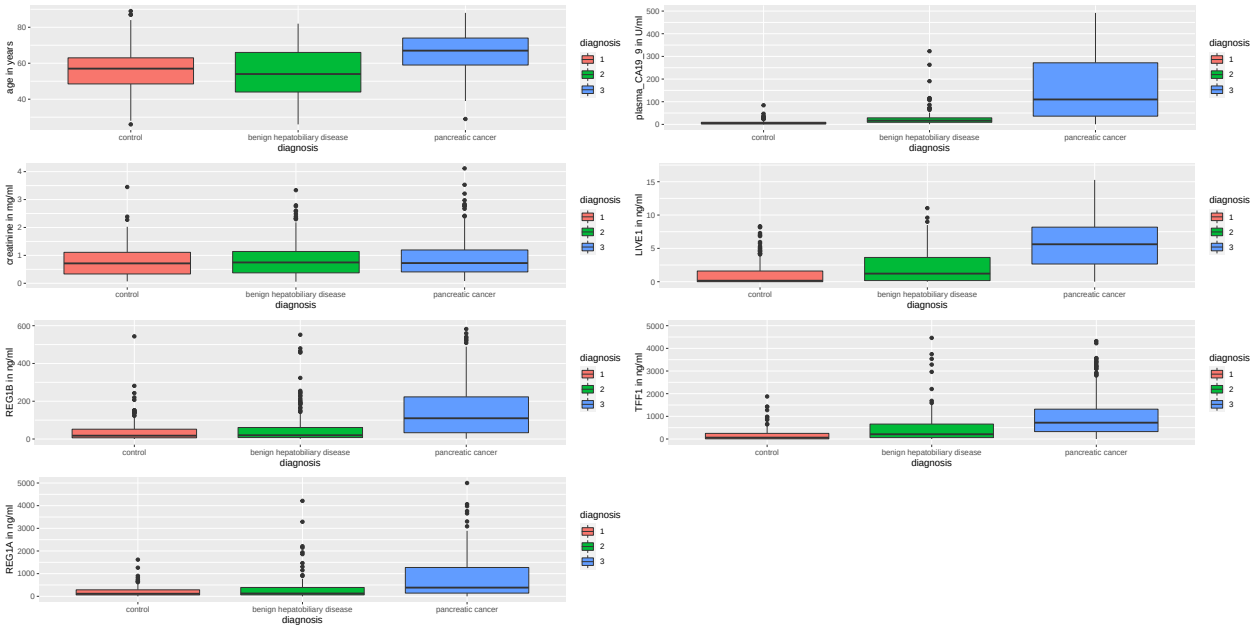

Figure 2: boxplots with added dimension (diagnosis)

It seems something else is at play to cause this perceived skewness. A statistical test using the skewness function in R can be used to test if the data is skewed and in what direction.

```
## [1] "skewness of age: -0.215731155227186"

## [1] "skewness of plasma_CA19_9: 7.95038202376304"

## [1] "skewness of creatinine: 1.45896535920294"

## [1] "skewness of LYVE1: 1.38693339066177"

## [1] "skewness of REG1B: 3.31699245780334"

## [1] "skewness of TFF1: 5.13210348006469"

## [1] "skewness of REG1A: 4.42540378216825"
```

This test proves that all but age is significantly skewed, since everything greater than a 1 or smaller than -1 is significant. To combat this skewness the data can be log transformed, scaling it logarithmically makes sense when the data can not be negative and seems to have expectational differences.

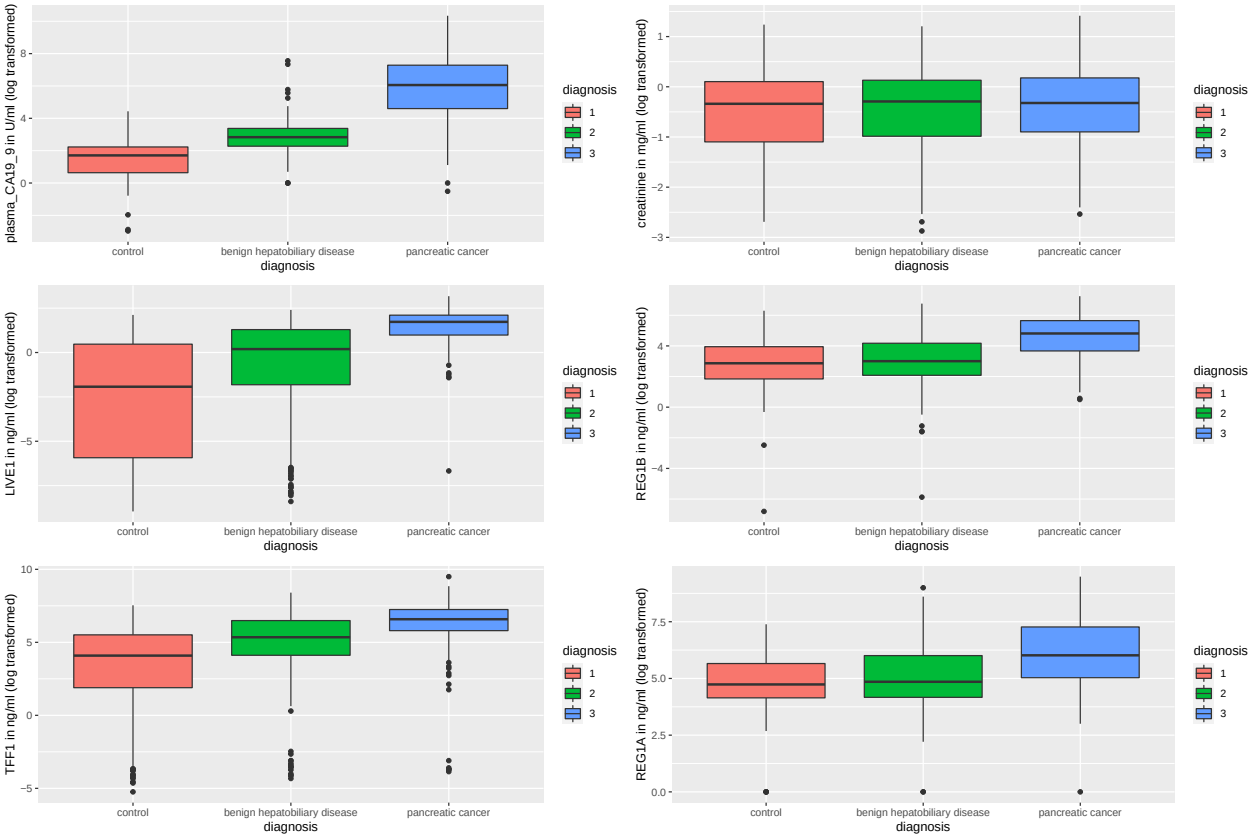

Figure 3: boxplot of log transformed data

looking at the data now it seems much less skewed however after a data transformation the normalcy should be tested this can be done using both a qq_plot and a shapiro.test.

```
##
##  Shapiro-Wilk normality test
##
## data:  trans$plasma_CA19_9
## W = NaN, p-value = NA
```

```
##
##  Shapiro-Wilk normality test
##
## data:  trans$creatinine
## W = 0.98211, p-value = 1.254e-06
```

```
##
##  Shapiro-Wilk normality test
##
## data:  trans$LYVE1
## W = 0.81496, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  trans$REG1B
## W = 0.9695, p-value = 9.888e-10
```

```
##
##  Shapiro-Wilk normality test
##
## data:  trans$TFF1
## W = 0.82765, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  trans$REG1A
## W = 0.97007, p-value = 5.564e-06
```

These variables are obviously not normally distributed due to none of them being above a p-value of 0.05, meaning that the null-hypothesis of the data being normaly distributed is rejected.

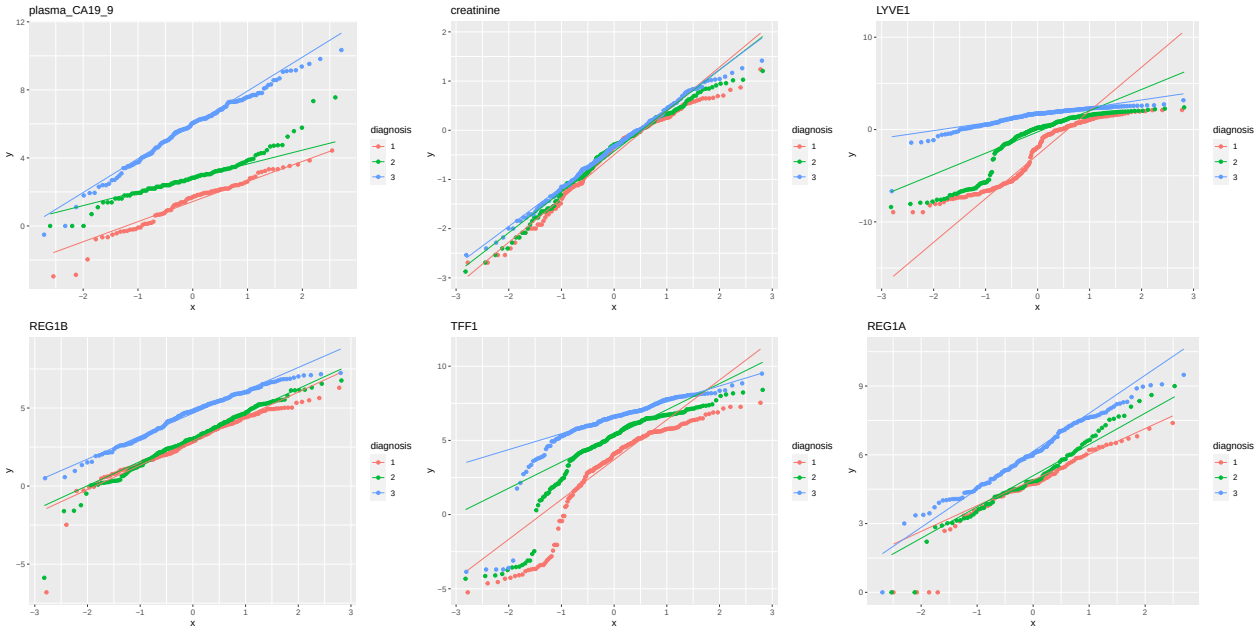

Figure 4: qqplots displaying normalcy

The final analysis is to explore if any variables are correlated this can be displayed using a heat-map.

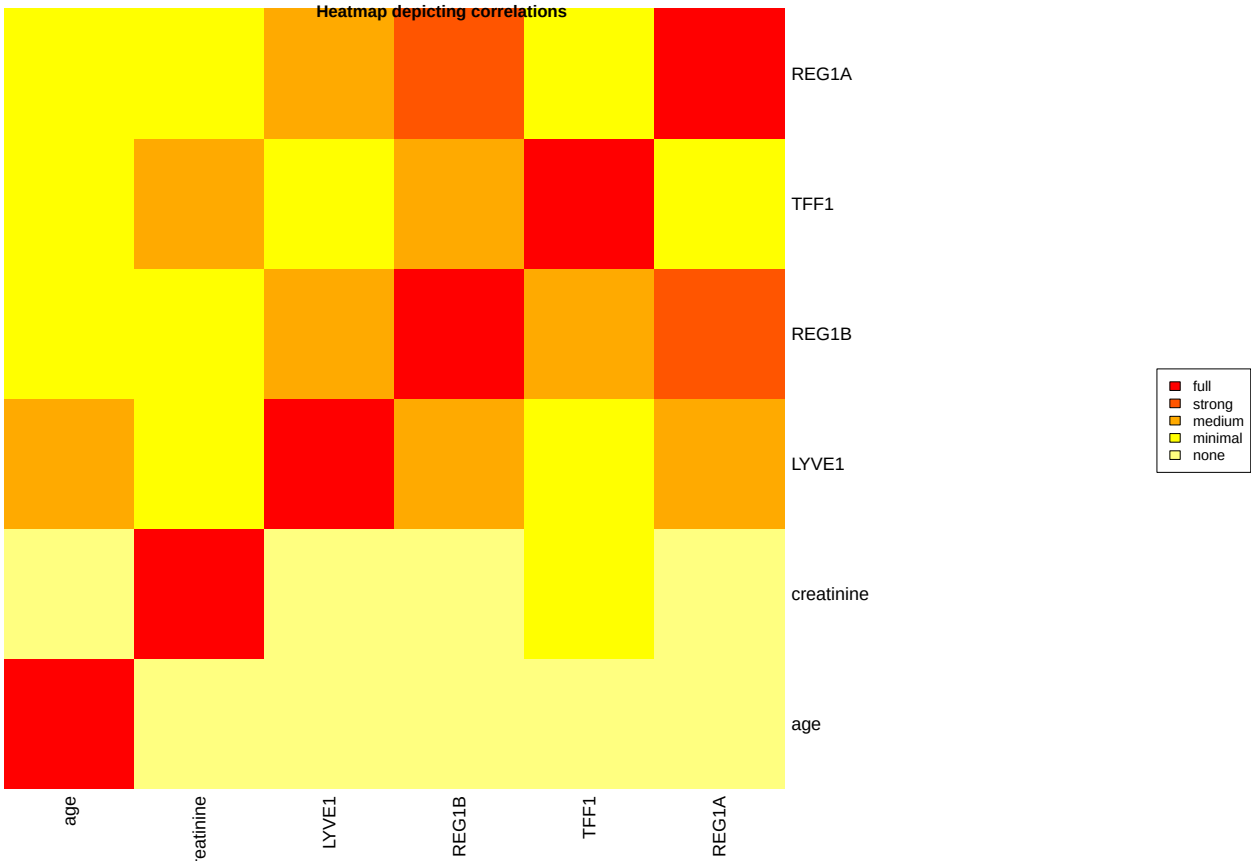

Figure 5: heatmap displaying correlation of values

REG1 A and B seem moderately correlated (0.7641084), otherwise no real strong correlation is observed.

Conclusion

Final Results

The final data set has had all numeric values except age log-transformed to reduce the skew but this left the data unnormal and also columns: sample_id, patient_cohort, sample_origin, stage and benign_sample_diagnosis dropped because them not being suitable for machine learning, furthermore collum diagnosis has been replaced with has_cancer because the research is focused to detect cancer. And plasma_CA19_9 has been dropped because in this research we focus on data that can be gathered from a urine sample. The data should however be useful and relaiable for machine learning.